UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS QUIXADÁ

TECNÓLOGO EM REDES DE COMPUTADORES

FÁBIO CORREIA LOPES

UMA FERRAMENTA PARA ANÁLISE DE DADOS DO PROGRAMA DE BICICLETAS COMPARTILHADAS BICICLETAR

QUIXADÁ – CEARÁ

2017

FÁBIO CORREIA LOPES

UMA FERRAMENTA PARA ANÁLISE DE DADOS DO PROGRAMA DE BICICLETAS COMPARTILHADAS BICICLETAR

Monografia apresentada no curso de Redes de Computadores da Universidade Federal do Ceará, como requisito parcial à obtenção do título de tecnólogo em Redes de Computadores. Área de concentração: Computação.

Orientador: Prof. Dr. Paulo Antonio Leal Rego

QUIXADÁ – CEARÁ

2017

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

L852f Lopes, Fábio Correia.
Uma ferramenta para análise de dados do programa de bicicletas compartilhadas bicicletar / Fábio Correia Lopes. – 2017.
50 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Campus de Quixadá, Curso de Redes de Computadores, Quixadá, 2017.
Orientação: Prof. Dr. Paulo Antonio Leal Rego.

1. Mobilidade Urbana. 2. Bicicletas. 3. Compartilhamento-Bicicletas-Transporte. I. Título.
CDD 004.6

FÁBIO CORREIA LOPES

UMA FERRAMENTA PARA ANÁLISE DE DADOS DO PROGRAMA DE BICICLETAS COMPARTILHADAS BICICLETAR

Monografia apresentada no curso de Redes de Computadores da Universidade Federal do Ceará, como requisito parcial à obtenção do título de tecnólogo em Redes de Computadores. Área de concentração: Computação.

Aprovada em: __/__/__

BANCA EXAMINADORA

Prof. Dr. Paulo Antonio Leal Rego (Orientador)
Universidade Federal do Ceará – UFC

Profª Dra. Ticiana Linhares Coelho da Silva
Universidade Federal do Ceará - UFC

Prof. Me. Regis Pires Magalhães
Universidade Federal do Ceará - UFC

À Deus.

AGRADECIMENTOS

Agradeço à minha mãe, Liduina, e ao meu pai, Aurélio, pela boa educação que sempre me foi dada, pelas motivações para que eu não desistisse dos meus objetivos e por sempre estarem ao meu lado.

Agradeço ao Prof. Dr. Paulo Antonio Leal Rego, por ter me dado a ideia sobre este projeto, pela paciência, e pela excelente orientação que permitiu a conclusão deste trabalho.

Agradeço aos professores Regis Pires Magalhães e Ticiana Linhares Coelho da Silva, pela disponibilidade em participar da banca desse trabalho e pelas excelentes colaborações e sugestões.

Agradeço ao professor Antônio Anselmo, por ter me iniciado na área da TI.

Agradeço a minha professora de português Tatiana Vieira de Lima, pela ajuda importantíssima que me deu.

Agradeço aos professores Paulo Rego, Marcos Dantas, Regis Pires, Arthur Callado, Neto Feitosa, Aragão, Rafael Braga, Lívia Almada, Paulo Henrique, Bruno Góis, Alisson Barbosa, Lavor, Bárbara Sampaio, Helder, João Marcelo, Michel Sales pelas aulas e pelos conselhos.

Agradeço aos meus amigos, Dieinison Jack, Dian Ferreia e João Vitor por terem me aguentado esses últimos dois semestres e pelas alegrias compartilhadas.

Agradeço aos meus amigos, Ana Lucia, Renan Alves e Walafi Ferreira, pela união, companheirismo que tivemos uns com os outros durante todo o curso.

Agradeço de forma especial à Juliana Castro pela paciência que teve comigo, por sempre estar ao meu lado me auxiliando nesta jornada, fazendo com que eu focasse nos estudos, saindo das distrações.

Agradeço aos "caba"de Ocara, pelas companheirismo e brincadeiras de todos as vezes no ônibus.

Agradeço aos meus amigos, e amigas, Jocélio, Ueliton Sousa, Randel Souza, Calebe Tavares, Wesla Nogueira, Kelly Domingos, Andreza Cristina, Juninho Rodrigues, Elton Celestino, Sara Cibelle, Nilmara Caetano, Valdeir Maia e Crislanio Macedo que colaboraram direta ou indiretamente para minha graduação.

A todos que direta ou indiretamente fizeram parte da minha formação.

"Daqui vinte anos você estará mais decepcionado pelas coisas que você não fez do que pelas coisas que você fez. Portanto livre-se das bolinas. Navegue longe dos portos seguros. Pegue os ventos da aventura em suas velas. Explore. Sonhe. Descubra."

(Mark Twain)

RESUMO

A mobilidade urbana tem sido prejudicada pelo crescente aumento dos veículos, provocando insatisfação nas pessoas, a partir da análise dos meios de transportes utilizados pela população, é possível alcançar soluções e ainda conhecimentos sobre a mobilidade urbana. Este trabalho, propõe uma ferramenta para fazer análise dos dados do programa de bicicletas compartilhadas Bicicletar da cidade de Fortaleza-CE. Foi feito um *script* para obter os dados do site oficial do programa, bem como o tratamento deles para obter respostas de 17 perguntas. As perguntas foram definidas pensando nos usuários e administradores do Bicicletar, pois as informações que são disponibilizadas no site oficial e aplicativo são poucas e não possuem estatísticas de uso das estações, mostrando apenas o estado das estações no momento. Com a ferramenta, foi possível verificar quais estações são mais e menos utilizadas, bem como que horário determinada estação tem bicicletas disponíveis.

Palavras-chave: Mobilidade Urbana. Bicicleta. Sistema de Compartilhamento de Bicicletas.

ABSTRACT

Urban mobility has been hampered by the growing number of vehicles, causing people's dissatisfaction by analyzing the types of transport used by the population, it is possible to reach solutions and knowledge about urban mobility. This work proposes a tool to analyze the data of the bike program shared by the city of Fortaleza-CE. A script was created to get the data from the official website of the program as well as the treatment of them to get answers to 17 questions. The questions were defined with the users and administrators of Bicicletar, since the information that is available in the official website and application are few and do not have statistics of use of the stations, showing only the state of the stations at the moment. With the tool, it was possible to check which stations are more and less used, as well as which particular season has bicycles available.

Keywords: Urban Mobility. Bicycle. Bike Sharing System

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

BBS	Bike Sharing System
GPS	Global Positioning System
SGBD	Sistema de Gerenciamento de Banco de Dados
SQL	Structured Query Language
KDD	Knowledge Discovery in Databases
API	Application Programming Interface
HTTP	Hypertext Transfer Protocol
XML	Extensible Markup Language
JSON	JavaScript Object Notation
AWS	Amazon Web Service
CSV	Comma-Separated Values
PHP	Hypertext Preprocessor
AJAX	Asynchronous JavaScript and XML

SUMÁRIO

1 INTRODUÇÃO

A mobilidade urbana nas grandes cidades tem sido afetada pelo crescente tráfego de veículos, causando atrasos, estresse e insatisfação nas pessoas (JÚNIOR et al., 2016). Para solucionar alguns desses problemas, é preciso realizar uma análise mais aprofundada sobre como as pessoas utilizam os transportes, sejam eles públicos ou privados.

Nos últimos anos, um dos sistemas de mobilidade urbana que está, cada vez mais, ganhando popularidade é o sistema de transporte de bicicletas compartilhadas. Esses sistemas contribuem de forma satisfatória para o meio ambiente e para as pessoas (MONCAYO-MARTíNEZ; RAMIREZ-NAFARRATE, 2016). Para Liu et al. (2015), o sistema de bicicletas compartilhadas é uma opção de transporte inovadora e sustentável com vários benefícios, trazendo uma solução amigável ao meio ambiente. Para os ciclistas, utilizar bicicletas é conveniente, acessível e saudável, além de mais eficiente que alguns transportes públicos em cidades urbanas congestionadas. Ainda é possível evitar atrasos e fazer desvios no tráfego. Segundo Chen et al. (2013), é um desafio fundamental construir sistemas de transporte eficientes e sustentáveis que acolham o crescente aumento da população nas cidades.

Para Purnama et al. (2015), é de fundamental importância entender a mobilidade urbana, já que possibilitaria o avanço em diversas aplicações, como sistemas de transporte inteligente, planejamento urbano, serviços baseados em localização e em outras áreas socieconômicas. O *Bike Sharing System (BSS)* com detecção de origem e destino, que registram a hora da partida e chegada, estão entre os sistemas de transporte mais promissores à análise. Mesmo sem o rastreamento por GPS(*Global Positioning System*), o sistema ainda pode obter informações úteis de forma satisfatória, como caraterísticas espaço temporal da mobilidade individual.

No presente trabalho, assim como em (PURNAMA et al., 2015) e (MONCAYO-MARTíNEZ; RAMIREZ-NAFARRATE, 2016), o foco é dado a sistemas de bicicletas compartilhadas, mais especificamente o programa Bicicletar[1], da cidade de Fortaleza, que atualmente conta com 69 estações espalhadas por vários bairros da capital cearense. Esse sistema oferece uma opção de transporte sustentável, não poluente e, além disso, é uma solução de transporte de pequeno trajeto, que facilita o deslocamento de pessoas em centros urbanos. Este trabalho tem como objetivo desenvolver uma ferramenta para fazer análise dos dados do programa Bicicletar. Dados como número de bicicletas disponíveis, número de vagas, status e

[1] http://www.bicicletar.com.br/

localização de cada estação foram obtidos no site oficial do programa e a ferramenta é capaz de responder uma lista de perguntas, dentre elas: quais estações são mais utilizadas, quais os dias e horários que uma determinada estação é mais usada e em que dias e horários determinada estação geralmente está vazia.

1.1 Objetivos

A seguir, é apresentado os objetivos do trabalho.

1.1.1 Objetivo Geral

O objetivo deste trabalho é desenvolver uma ferramenta para fazer a análise dos dados do programa de bicicletas compartilhadas Bicicletar, da cidade de Fortaleza, Ceará.

1.1.2 Objetivos específicos

a) Fazer um levantamento das técnicas de coleta de dados.
b) Coletar a base de dados para o programa Bicicletar.
c) Definir perguntas a serem respondidas com a ferramenta.
d) Definir arquitetura e implementar a ferramenta.

Os demais capítulos desta monografia estão organizadas da seguinte forma: O capítulo 2 apresenta a fundamentação teórica. Os trabalhos relacionados são apresentados no capítulo 3, enquanto o capítulo 4 apresenta os procedimentos metodológicos, no capítulo 5 a ferramenta desenvolvida e o capítulo 6 as conclusões obtidas.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, é apresentada a fundamentação teórica, onde são descritos os principais conceitos relacionados ao trabalho. Dentre eles: mobilidade urbana, Bicicletar, descoberta de Conhecimento em Banco de Dados (*Knowledge Discovery in Databases*) e técnicas de coleta de dados.

2.1 Mobilidade Urbana

De acordo com Prata e Sanches (2016), mobilidade urbana é "[...] tudo aquilo acerca do deslocamento de pessoas dentro do perímetro urbano". Após a I Guerra Mundial, a urbanização ganhou força, objetivando fazer com que as zonas habitacionais fossem dotadas de diversos centros e outras facilidades. Além disso, a urbanização ainda causou congestionamentos e poluição, com o aumento do transporte individual motorizado.

Atualmente, as comunicações urbanas têm sido facilitadas significativamente pelas aplicações de *smartphones*, como por exemplo, fornecendo dados de localização em tempo real, para serviços de mapas. Algumas dessas aplicações mais famosas são *Google Maps*[1] e *Waze*[2]. O *Google maps*, além de fornecer instruções passo a passo de como um usuário obter caminhos para algum lugar, também mostra condições de congestionamento de tráfego e tempo estimado de chegada. Já o *Waze*, além de exibir as condições de congestionamento de tráfego , também funciona como uma rede social para os motoristas, onde eles fornecem informações em tempo real como atolamentos, buracos e acidentes em determinados locais, exibindo-os pelo mapa (WANG et al., 2017).

Segundo Soares (2015), acreditou-se "[..] que o automóvel respondia a esta necessidade de acessibilidade tanto para os residentes das cidades como para os habitantes das zonas não urbanas". Necessidades que o meio urbano oferece, dentre elas cultura, comércio, formação, serviços, atividades sociais e políticas. Porém o automóvel passou a ter efeito nocivo aos centros urbanos, causando aborrecimentos às pessoas pelas horas perdidas em engarrafamentos. Tornou-se necessária a redução dos automóveis, para a própria manutenção da mobilidade dos automóveis. Ainda segundo o autor, a bicicleta tem surgido nos últimos anos, como alternativa de meio de transporte nos centros urbanos, trazendo facilidades como redução direta dos congestionamentos, redução de estacionamentos, melhoramento geral da qualidade de

[1] https://www.google.com.br/maps
[2] https://www.waze.com/pt-BR

vida da cidade reduzindo a poluição como exemplo e custos de manutenção como limpezas. Além disso, impactando significativamente na redução dos automóveis nas cidades.

É de fundamental importância entender a mobilidade urbana, possibilitando avanços em planejamento urbano, sistema de transporte inteligente, serviços baseados em localização entre outros. Estudos de mobilidade urbana já foram feitos através de dados como celular, meios de comunicação social, dados de táxi, rastreamentos baseados em GPS e BBS (PURNAMA et al., 2015).

2.1.1 Bicicletar

O Bicicletar é um projeto de bicicletas compartilhadas da cidade de Fortaleza. O projeto é da Prefeitura Municipal de Fortaleza, mas operado pela empresa Serttel com apoio da Unimed Fortaleza. Atualmente o programa conta com 69 estações inteligentes espalhadas pela capital cearense. Além disso existe também o Mini Bicicletar, o mesmo projeto só que este focado para crianças, atualmente consta com 4 estações. A maioria das estações contém 12 espaços para disposição das bicicletas, tanto o Mini Bicicletar como o Bicicletar.

Para que seja possível a comunicação dos clientes, ao retirar ou repor uma bicicleta, as estações são conectadas a uma central de operações utilizando tecnologia *wireless*. As pessoas que desejam utilizar as bicicletas do Bicicletar precisam se cadastrar, adquirir passes e liberar as bicicletas através do aplicativo oficial do programa. Além disso, também é possível utilizar o sistema sem fazer cadastro através de um passe diário, que pode ser obtido ao telefonar para o número oficial do Bicicletar.

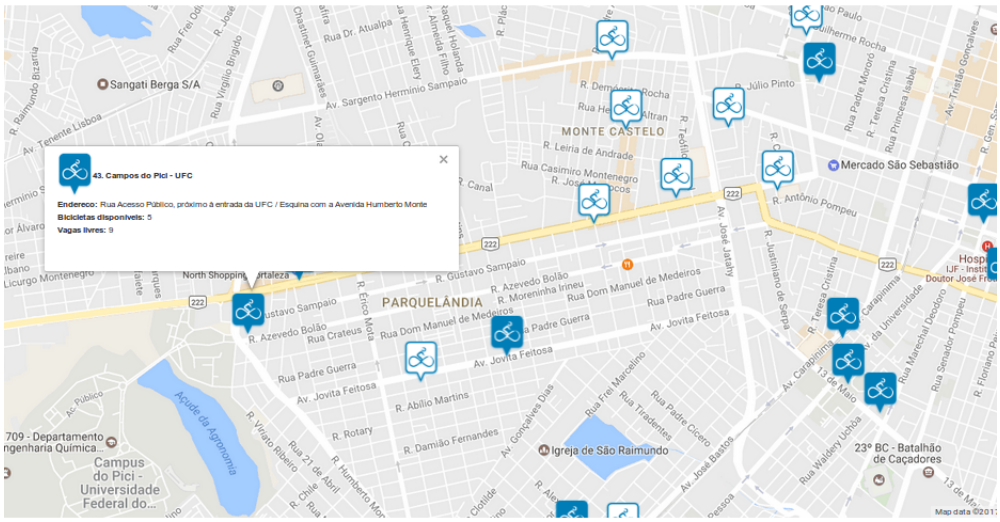
De acordo com o site oficial do Bicicletar (2017), o programa tem como objetivos:

- Introduzir a bicicleta como modalidade de transporte público saudável e não poluente.
- Combater o sedentarismo da população e promover a prática de hábitos saudáveis.
- Reduzir os engarrafamentos e a poluição ambiental nas áreas centrais das cidades.
- Promover a humanização do ambiente urbano e a responsabilidade social das pessoas.

O site oficial contém o mapa das estações, que pode ser observado na Figura 1. No mapa, são exibidas as estações, informações dos estados das estações, a localização da estação, a quantidade de bicicletas e vagas disponíveis para os usuários em tempo real. Os estados das estações são: em operação, em implementação ou manutenção, todas as vagas ocupadas, nenhuma bicicleta disponível e offline, mostrados na Figura 2. Na Figura 1, como exemplo, está selecionada a Estação 43 (Campus do Pici - UFC), o estado da estação está Em Operação, com 5

bicicletas disponíveis e 9 vagas livres.

Figura 1 – Mapa de estações Bicicletar



Fonte: Bicicletar (2017)

Figura 2 – Estados das Estações

Fonte: Bicicletar (2017)

No Brasil, existem projetos como esse espalhados por várias capitais, alguns deles são: Bike Sampa na cidade de São Paulo-SP[3], Bike Rio[4] na cidade de Rio de Janeiro-RJ, Bike PE[5] em Pernambuco, Bike Brasília[6] na capital do Brasil, GynDebike[7] em Goiânia-GO, Bike Vitória[8] em Vitoria-ES e Caju Bike[9] em Aracaju-SE.

[3] http://www.mobilicidade.com.br/bikesampa.asp
[4] https://www.mobilicidade.com.br/bikerio.asp
[5] http://www.bikepe.com/
[6] http://www.bikebrasilia.com/
[7] http://www.debikegoiania.com/
[8] http://www.bikevitoria.com/home.aspx
[9] http://www.cajubike.com/home.aspx

2.2 Terminologia de Banco de Dados

Segundo Elmasri e Navathe (2010), banco de dados é "uma coleção de dados relacionados. Por dados, nos referimos a fatos conhecidos que podem ser registrados e que têm significado implícito". No trabalho de Cabeça et al. (2009), são definidos alguns termos como:

> Sistema Gerenciador de Banco de Dados (SGBD) é um conjunto de programas que permite ao usuário executar operações como criar e manter bancos de dados. A junção de um sistema gerenciador de banco de dados com um ou mais bancos de dados constitui os chamados sistemas de banco de dados. Uma aplicação ou aplicação de banco de dados consiste em um ou mais programas que interagem com sistemas de banco de dados. Os bancos de dados relacionais representam os dados como um conjunto de relações. Uma relação é uma tabela na qual cada linha representa dados sobre uma entidade particular e cada coluna representa um aspecto particular dos dados.

Existem vários SGBDs atualmente, dentre eles são: *MySQL*[10], *SQL Server*[11] *FireBird*[12], *MongoDB*[13] e *PostgreSQL*[14]. Mas para este estudo foi utilizado o *PostgreSQL*, pois ele é gratuito, fácil instalação, multiplataforma (funciona em vários sistemas operacionais) e funciona muito bem com grandes quantidade de dados.

Spoto et al. (2000) afirmam que os Sistemas Gerenciadores de Banco de Dados Relacionais, "usam comandos de SQL (*Structured Query Language*) para manipular os dados armazenados na relação da base de dado".

2.2.1 Linguagem de Banco de Dados Relacional - SQL

Originalmente, o SQL foi projetado e implementado na IBM Research como interface para um sistema de banco de dados relacional experimental chamado SYSTEM R. SQL é agora o idioma padrão para SGBDs relacionais comerciais. O SQL possui declarações para definições, consultas e atualizações de dados. Também usa os termos formais relação, tupla e atributo para denotar tabela, linha e coluna, respectivamente.

A linguagem SQL foi de fundamental importância para este trabalho, pois ela possibilitou a realização das consultas a base de dados, de forma a auxiliar nas respostas da lista de perguntas.

[10] https://www.mysql.com/
[11] https://www.microsoft.com/pt-br/sql-server/sql-server-downloads
[12] https://firebirdsql.org/
[13] https://www.mongodb.com/
[14] https://www.postgresql.org/

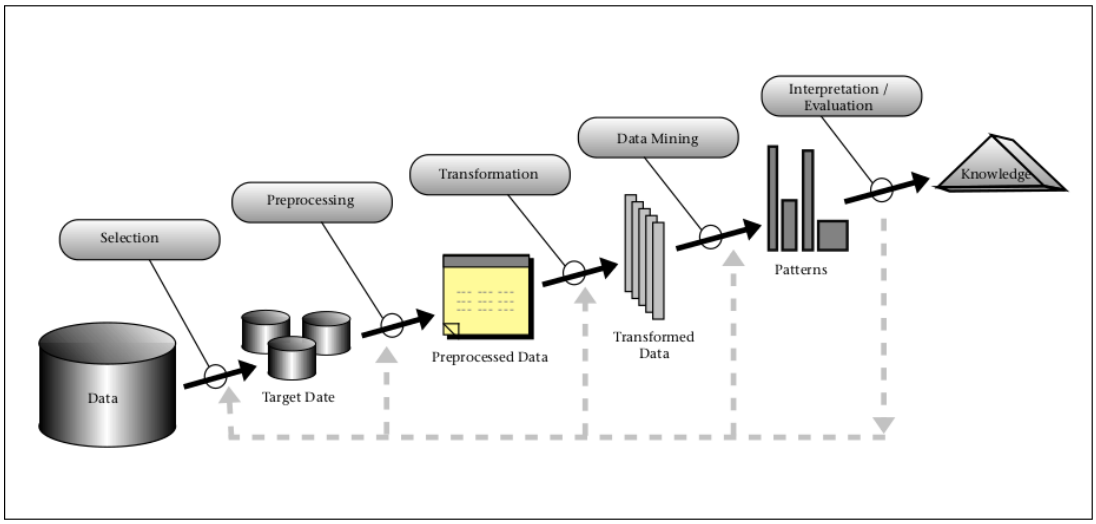
2.3 Descoberta de Conhecimento em Banco de Dados (*Knowledge Discovery in Databases*)

Para Fayyad, Piatetsky-Shapiro e Smyth (1996) KDD (*Knowledge Discovery in Databases*) é o "processo, não trivial, de extração de informações implícitas, previamente desconhecidas e potencialmente úteis, a partir dos dados armazenados em um banco de dados". Na Figura 3 são mostrados os processos para obter conhecimento a partir de dados. Fayyad, Piatetsky-Shapiro e Smyth (1996) descrevem esses processos como:

1. Desenvolver conhecimento sobre o domínio da aplicação e identificar o objetivo do processo KDD.
2. Selecionar o conjunto de dados sobre o qual a descoberta deve ser realizada.
3. Limpeza e pré-processamento de dados. Essas operações incluem a remoção de ruído (erros ou valores estranhos), se necessário, coletando as informações necessárias para modelar ou representar o ruído, decidindo estratégias para lidar com os campos de dados em falta.
4. Redução e projeção de dados: encontrar recursos úteis para representar os dados de acordo com o objetivo da tarefa. Com métodos de redução da dimensionalidade ou de transformação, o número efetivo de variáveis em consideração pode ser reduzido, ou podem ser encontradas representações invariantes para os dados.
5. Análise exploratória e seleção de modelo e hipótese: escolhendo os algoritmos de *data mining* e os métodos de seleção a serem usados para pesquisar padrões de dados. Este processo inclui decidir quais modelos e parâmetros podem ser apropriados (por exemplo, os modelos de dados categóricos são diferentes dos modelos de vetores sobre os reais) e combinando um método particular de mineração de dados com os critérios gerais do processo KDD.
6. Mineração de dados: busca por padrões de interesse em uma determinada forma de representação ou um conjunto de tais representações, incluindo regras de classificação ou árvores, regressão e agrupamento.
7. Interpretando padrões extraídos. Esta etapa também pode envolver a visualização dos padrões e modelos extraídos ou a visualização dos dados dados os modelos extraídos.
8. Conhecimento descoberto: usando o conhecimento diretamente, incorporando o conhecimento em outro sistema para ações futuras, ou simplesmente documentando-o e

denunciando as partes interessadas. Esse processo também inclui verificar e resolver possíveis conflitos com o conhecimento acreditado anteriormente (ou extraído).

Figura 3 – Visão geral das etapas do processo KDD



Fonte: Fayyad, Piatetsky-Shapiro e Smyth (1996)

Neste estudo não segue todos os passos do KDD, pois alguns não são necessários para o objetivo do trabalho, porém são utilizados os processos de seleção, pré-processamento e transformação.

2.3.1 Técnicas de Coleta de Dados

Segundo Benevenuto, Almeida e Silva (2011), existem várias formas de obter dados da *web*, dentre elas, utilizando API (*Application Programming Interface*), dados de aplicação e *crawler*. Uma API, no contexto de desenvolvimento *web* é um conjunto de tipos de requisições HTTP (*Hypertext Transfer Protocol*) com as suas respectivas definições de resposta. APIs são boas para coletas de dados, pois podem oferecer dados estruturados em formatos como XML (*Extensible Markup Language*) ou JSON (*JavaScript Object Notation*).

Segundo Silva e Loureiro (2015), em dados de aplicação, os dados são obtidos através da criação de aplicações na própria plataforma onde se pretende obter os dados. Algumas dessas plataformas são Facebook[15], Instragram[16] e Runkeeper[17]. Através dessas aplicações, é possível obter os dados dos usuários que as utilizam, mas apenas se os usuários permitirem o compartilhamento desses dados. Para Santos (2010), *crawlers* "são robôs que percorrem a

[15] https://www.facebook.com/
[16] https://www.instagram.com/
[17] https://runkeeper.com/

web de uma maneira específica, salvando as páginas visitadas de acordo com as necessidades do usuário". Como dizem Silva e Loureiro (2015), nem todos os dados disponíveis na *web*, estão disponibilizados através de API's oficiais dos próprios desenvolvedores da plataforma, que é o caso do Bicicletar e uma alternativa que os autores mostram são os *web crawlers*. Porém existe outra forma chamada de *scraping*, que é uma técnica de software utilizada para extrair informações de sites, transformando os dados não estruturados em dados estruturados (SUNDARAMOORTHY; DURGA; NAGADARSHINI, 2017). A coleta de dados através de *scraping* depende de como os dados estão estruturados nas páginas *web* por exemplo, em que *tags* HTML estão os dados que quer obter).

Como o Bicicletar não possibilita a criação de aplicações na sua plataforma, nem fornece APIs para acesso aos dados, este trabalho utilizou a técnica *scraping* para obter os dados diretamente do site oficial do programa, pois os *web crawlers* desempenham a função de ir em página por página (link por link) ou a partir de uma lista de links obtendo informações, enquanto *scraping* é mais especifico sobre o que deseja obter de uma determinada página *web*.

3 TRABALHOS RELACIONADOS

Em (JÚNIOR et al., 2016), foram analisados dados de corridas de táxi solicitadas a partir da aplicação para *smartphones WayTaxi* na cidade de Belo Horizonte-MG. O estudo baseia-se em uma metodologia de caracterização e foi dividido em caracterização temporal e caracterização espacial. Os autores constataram que, nos momentos de pico de solicitações, a demanda não foi atendida, gerando cancelamentos dos usuários. Além disso, a maioria das solicitações foram realizadas na região centro-sul da cidade e finalizadas em um bairro nobre ou em lugares que possibilitam a saída da cidade.

No trabalho de Purnama et al. (2015), a mobilidade urbana foi analisada através de um conjunto de dados de bicicletas compartilhadas de Londres, que continha cerca de 2.961.183 viagens com 566.888 usuários e 569 estações. Com o objetivo de caracterizar e prever a mobilidade urbana, o estudo focou na identificação de usuários altamente previsíveis, revelando suas características de mobilidade e prevendo seus movimentos. O conjunto de dados tinha informações importantes como a identificação do usuário, identificação da estação de retirada e retorno de bicicletas, hora do inicio e fim de viagem, duração em minutos da viagem e a localização geográfica da estação. Foi feita uma distribuição de usuários pelo seu nível de previsibilidade e caracterização de padrões de mobilidade entre os grupos de usuários, através de análises temporais e espaciais.

Com esses dados, foi possível descobrir que existia forte distinção entre as características dos usuários em relação à mobilidade espaço-temporal. Pela manhã, constatou-se que os usuários andam em maior velocidade e em maiores distâncias que em outros momentos e que a próxima localização do usuário depende principalmente de sua localização atual. Além disso, notou-se que em um contexto de previsão, os usuários registrados são altamente previsíveis, mas usuários com alto índice de viagens são ligeiramente superiores na questão de previsão, sendo eles registrados ou não.

Moncayo-Martínez e Ramirez-Nafarrate (2016) analisaram a mobilidade urbana com base no sistema de bicicletas compartilhadas na Cidade do México e o conjunto de dados foi obtido através de um site de dados abertos, com cerca de 18 milhões de viagens entre 444 estações, do período de janeiro de 2010 a janeiro de 2015. Esse conjunto era dividido por meses, cada mês com informações em 10 colunas. Algumas dessas informações eram: ID da viagem, sexo do usuário, ID da bicicleta, estação de partida/chegada, data de partida e hora de partida. Os autores obtiveram o tempo médio dos usuários ao usar as bicicletas e foi possível classificar as

estações como estações de suprimento ou de demanda, onde estações de suprimento são aquelas estações em que existe mais saídas de viagens do que chegadas, enquanto, nas estacões de demanda, existe um índice maior de chegadas de viagens. Além disso, os autores identificaram que muitas viagens começam e terminam na mesma estação. A Tabela 1, traz as principais características dos estudos apresentados.

Tabela 1 – Trabalhos Relacionados

ESTUDO	OBJETO ANALISADO	FONTE DE DADOS	OBJETIVO DO ESTUDO
(JÚNIOR et al., 2016)	Táxi	Aplicativo WayTaxi.	Descobrir razões para estado do trafico e possíveis soluções.
(PURNAMA et al., 2015)	Sistema de Bicicletas Compartilhadas	Dados Abertos	Analisar o grau de aleatoriedade e previsibilidade dos usuários do Sistema de Bicicletas Compartilhadas
(MONCAYO-MARTíNEZ; RAMIREZ-NAFARRATE,,2016)	Sistema de Bicicletas Compartilhadas	Dados Abertos	Analisar a demanda do Sistema de Bicicletas Compartilhadas
Presente Estudo	Sistema de Bicicletas Compartilhadas	Site Oficial	Desenvolver uma ferramenta para fazer a análise do Sistema de Bicicletas Compartilhadas

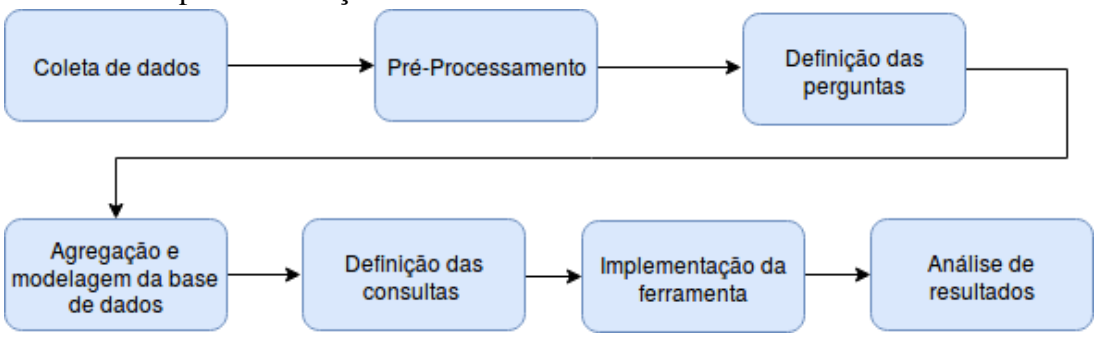
O presente estudo, assim como os trabalhos mencionados, também analisou dados de programas voltados para a mobilidade urbana, mas diferente de Júnior et al. (2016), que analisava dados de corridas de táxis, este trabalho analisou os dados de um sistema de bicicletas compartilhadas da cidade de Fortaleza - Ceará. Diferente de (MONCAYO-MARTíNEZ; RAMIREZ-NAFARRATE, 2016) e (PURNAMA et al., 2015), que também focaram em bicicletas compartilhadas, nossa solução não utilizará uma base de dados pronta ou fornecida pelos administradores dos sistemas. Este trabalho coletou sua base de dados utilizando a técnica *scraping*, que obteve as informações diretamente do site oficial do Bicicletar[1]. Além disso, não teve as informações dos usuários e identificação da bicicleta, o que limitou o escopo das perguntas. Porém, as informações que foram geradas com a ferramenta auxilia o usuário final e os administradores do programa Bicicletar.

[1] http://www.bicicletar.com.br/

4 PROCEDIMENTOS METODOLÓGICOS

Neste capítulo, são apresentados os procedimentos metodológicos que foram realizados para a conclusão deste trabalho. A Figura 4 apresenta a ordem dos processos que foram utilizados para a execução.

Figura 4 – Passos para a execução do trabalho



Fonte: O próprio autor (2017)

4.1 Coleta de dados

Nesta etapa, foi utilizado o processo de seleção do KDD. Para a coleta dos dados do programa Bicicletar, uma técnica chamada *scraping* foi utilizada, sendo a mesma executada em um servidor com processador Intel(R) Xeon(R) CPU E5-2676, 990MB de memória RAM e sistema operacional Ubuntu 16.04.2 LTS na nuvem da *Amazon Web Service* (AWS). Com essa técnica, o *script* desenvolvido foi executado através do *crontab*, um serviço do Unix que gerencia comandos para serem executados em períodos predeterminados. O *script* foi executado a cada minuto (sendo que dentro do *script*, a coleta era executada a cada 30 segundos, extraindo os dados do site oficial do programa e armazenando em arquivos CSV (*Comma-Separated Values*) com informações importantes das estações, tais como número de bicicletas disponíveis, número de vagas, status (ativa,inativo, operação e manutenção) e localização de cada estação. Para este estudo, foram coletados dados de 01 de abril de 2017 às 00:00 horas até 30 de junho de 2017 às 23:59 horas, totalizando 4,1 GB (Gigabyte) de dados.

4.2 Pré-Processamento

Nesta etapa, foi utilizado o pré-processamento do KDD. Foram removidos 1318 arquivos CSV que estavam em branco de 261702 arquivos baixados. Um dos possíveis motivos para que os arquivos estivessem em branco é a indisponibilidade do site oficial do programa

Bicicletar.

4.3 Definição das perguntas

Nesta etapa, foi definida uma lista de perguntas que deveriam ser respondidas utilizando a ferramenta proposta. Tais perguntas foram pensadas para auxiliar os usuários do programa Bicicletar, bem como seus administradores, a entenderem o comportamento do sistema com o tempo.

Abaixo estão listadas as perguntas escolhidas para implementação na ferramenta:

1. Quais as estações mais e menos utilizadas?
2. Dado um dia da semana, quais as estações mais e menos utilizadas?
3. Quais estações são mais utilizadas em um determinado dia/horário?
4. Quais os dias da semana em que uma determinada estação é mais usada?
5. Dado um dia da semana e uma estação, em qual horário ela é mais usada?
6. Dado um dia da semana e uma estação, quais horários ela costuma estar com bicicletas disponíveis?
7. Dado um dia da semana e uma estação, qual a média de tempo em que a estação tem bicicletas disponíveis?
8. Dado um dia da semana e um horário, quais estações costumam estar com bicicletas disponíveis?
9. Qual a probabilidade de uma estação ter bicicleta em um determinado dia/horário?
10. Qual a probabilidade de uma estação não ter bicicleta em um determinado dia/horário?
11. Qual estação seria escolhida como melhor opção, para uma pessoa que deseja uma bicicleta em uma determinada área, dia/horário?
12. Qual a média de tempo que uma estação fica sem bicicletas dado um dia da semana?
13. Quais as estações que tiveram mais e menos inatividades?
14. Quais dias da semana uma estação costuma ficar inativa?
15. Qual a média de tempo que uma estação ficar inativa dado um dia da semana?
16. Quais estações ficaram em manutenção?
17. Qual a probabilidade de pegar uma bicicleta em uma estação em um determinado dia e horário e conseguir devolver a bicicleta em outra/mesma estação em outro horário?

4.4 Agregação e modelagem da base de dados

Esta etapa utilizou o processo de transformação do KDD. A agregação foi feita utilizando a linguagem de programação *Python* junto com as bibliotecas *pandas*[1] e *numpy*[2], que são ferramentas poderosas para análise de dados. O *pandas* foi utilizado para criação de *DataFrames*, que são estruturas de dados com duas dimensões, com colunas e linhas semelhante a estrutura de tabelas de um banco de dados. O *numpy* foi utilizado para realizar os cálculos de média, desvio padrão dos dados do *DataFrame*.

A agregação dos dados foi feita em intervalos de tempo, de 15 em 15 minutos, obtendo as informações de: média, mínimo, máximo e desvio padrão das bicicletas disponíveis assim como das vagas disponíveis. Também foi obtida a porcentagem de tempo em que as estações estavam ativas e em manutenção. Para a resolução de algumas perguntas foram definidas duas métricas: uma que conta quantas bicicletas foram retidas e outra quantas foram devolvidas. Além disso, foram armazenados a quantidade de arquivos que foram usados para cada agrupamento, bem como outro campo com o valor da data de inicio do agrupamento. Depois da agregação, os dados foram salvos em um SGBD *PostgreSQL*[3] através de uma função do próprio *pandas*, que automaticamente identifica os tipos dos dados do *DataFrame* e cria os atributos da tabela conforme os tipos de dados identificados. O diagrama do banco de dados está no Apêndice A.

4.5 Definição das consultas

Nesta etapa, foram definidas quais seriam as consultas SQL, utilizadas para obter as respostas das perguntas propostas. Essas consultas são explicadas na Seção 5.2 e estão em anexo no Apêndice B.

4.6 Implementação da ferramenta

Com os passos anteriores concluídos, a implementação da ferramenta tem como função desenvolver uma forma para que os dados salvos no banco de dados gerem conhecimento para os usuários. Neste trabalho, foi criada uma aplicação *web*, que utiliza os dados agregados

[1] https://pandas.pydata.org/
[2] http://www.numpy.org/
[3] https://www.postgresql.org/

para apresentar gráficos, feitos em *JavaScript*.

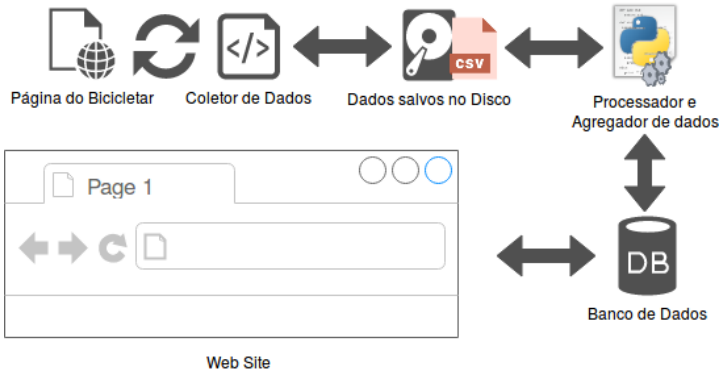
5 FERRAMENTA PROPOSTA

Este capítulo apresenta a arquitetura e componentes da ferramenta desenvolvida, e discute as consultas criadas para responder as perguntas listadas na Seção 4.3.

5.1 Visão Geral da Arquitetura da Ferramenta

A Figura 5 apresenta a arquitetura da ferramenta proposta. O Coletor de Dados é responsável por obter as informações das estações do programa Bicicletar através do site oficial. O Processador de Dados é utilizado para escolher apenas os dados importantes, bem como fazer a inserção de novos atributos. Em seguida, o Agregador realiza o agrupamento dos dados de 15 em 15 minutos, e também insere novos dados. O Banco de Dados é responsável pelo armazenamento dos dados e também por processar e responder as consultas. O *web site*, responsável pela execução das consultas e exibição dos resultados, foi desenvolvido utilizando as seguintes tecnologias: PHP (Hypertext Preprocessor), HTML, BootStrap, Jquery, Canvasjs, Chartjs e Google Maps JavaScript API. Através de requisições Ajax (Asynchronous JavaScript e XML), são selecionadas as consultas adequadas para cada pergunta, e utilizando o PHP para realizar tais consultas, obtém-se como resposta dados no formato JSON. O *web site* traz duas funcionalidades extras, que não estão no escopo das perguntas: um mapa de calor para a avaliar a taxa de utilização das estações no tempo e um gráfico em linha para apresentar a variação da quantidade de bicicletas disponíveis no tempo. Ambos os gráficos mostram dados de todos os dias do período de coletado.

Figura 5 – Ferramenta Proposta



Fonte: O próprio autor (2017)

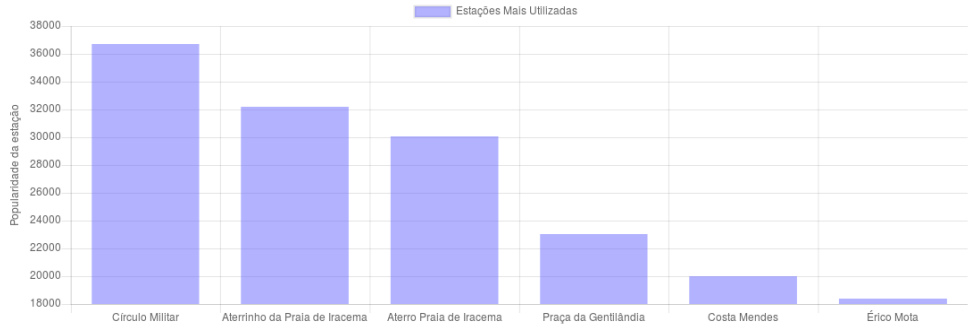
5.2 Resultados

Para todas as questões propostas, as resoluções foram feitas levando em conta os dados do período coletado, de 01 de abril a 30 de junho de 2017. A seguir, é apresentado como a ferramenta responde as perguntas:

1. Quais estações foram mais e menos utilizadas?

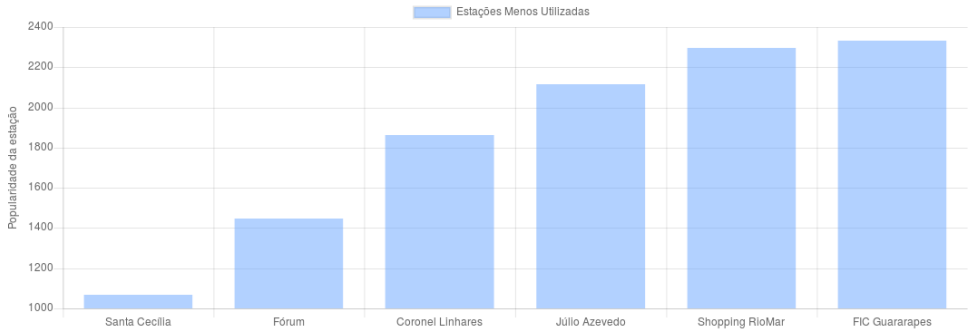
   Para responder a pergunta, foram utilizadas duas métricas mencionadas anteriormente - a que conta quantas bicicletas foram retiradas e a que conta quantas bicicletas foram devolvidas na estação. Somando os dois valores, é definido uma terceira métrica que foi chamada de popularidade da estação, que representa quantas operações de retirada e devolução foram executadas em uma dada estação. Assim, para esta consulta foi calculada a popularidade da estação agrupando os dados pelos ID's das estações. Assim, conforme o gráfico da Figura 6, a estação do Circulo Militar possui uma maior utilização, seguida das estação Aterrinho da Praia de Iracema, Aterro Praia de Iracema, Praça da Gentilândia, Costa Mendes e Érico Mota, a maioria são estações próximas à praia, o que pode ser um dos motivos que as fazem possuir uma maior movimentação de bicicletas. Por outro lado, as estações Santa Cecília, Fórum, Coronel Linhares, Júlio Azevedo, Shopping RioMar e FIC Guararapes obtiveram as menores popularidades, como pode-se observar na Figura 7.

Figura 6 – Estações mais utilizadas



Fonte: O próprio autor (2017)

Figura 7 – Estações menos utilizadas

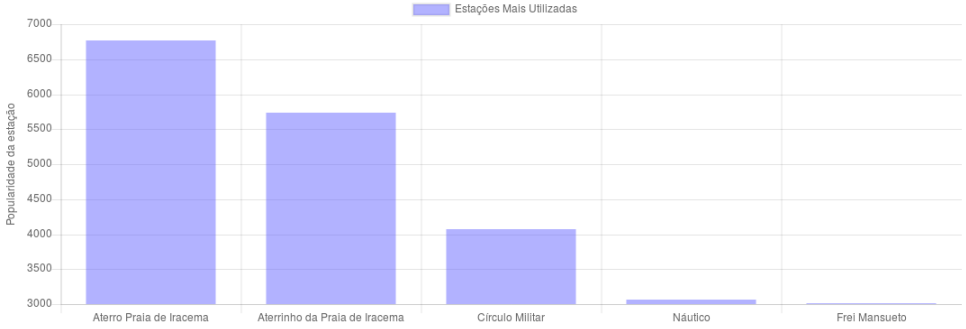


Fonte: O próprio autor (2017)

2. Dado um dia da semana, quais as estações mais e menos utilizadas?

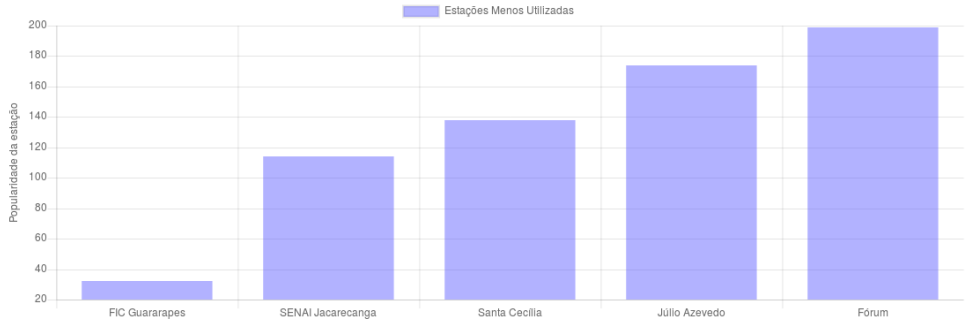
   Nesta pergunta, foi utilizada a mesma métrica de popularidade da estação, porém agrupando os dados pelo dia da semana. Na Figura 8, são mostradas as estações com mais popularidade tomando o dia da semana domingo: Aterro Praia de Iracema, Aterrinho da Praia de Iracema, Círculo Militar, Náutico e Frei Mansueto. Por outro lado, como mostra a Figura 9, as estações menos utilizadas são FIC Guararapes, SENAI Jacarecanga, Santa Cecília, Júlio Azevedo e Fórum. Existe sempre uma estação próxima à praia sendo mais utilizada em qualquer dia da semana e no domingo por exemplo existe 4 estações.

Figura 8 – Estações mais utilizadas dado dia da semana domingo



Fonte: O próprio autor (2017)

Figura 9 – Estações menos utilizadas dado dia da semana domingo

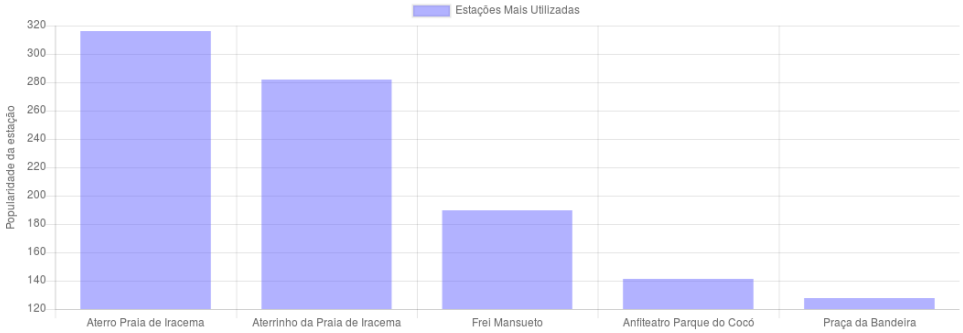


Fonte: O próprio autor (2017)

3. Quais estações são mais utilizadas em um determinado dia/horário?

   Da mesma forma como foi respondida a questão anterior, porém agrupando os dados por uma determinada hora. Como ilustrado na Figura 10, as estações mais utilizadas no domingo às 8 horas são Aterro Praia de Iracema, Aterrinho Praia de Iracema, Frei Mansueto, Anfiteatro Parque do Cocó e Praça da Bandeira.

Figura 10 – Estações mais utilizadas dado dia da semana domingo às 8 horas

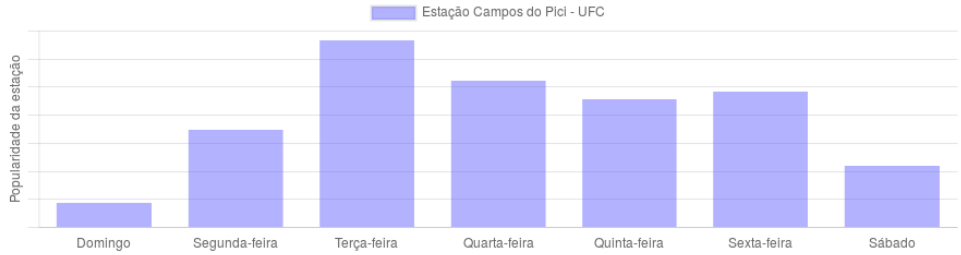


Fonte: O próprio autor (2017)

4. Quais os dias da semana em que uma determinada estação é mais usada?

   Para esta pergunta, foi utilizada a métrica de popularidade já mencionada, mas os dados foram agrupados pelo dia da semana. Na Figura 11, é mostrada a estação Campus do Pici - UFC, observando que os dias que tem menos utilização são os sábados e os domingos pois não são dias letivos, levando à conclusão da utilização por parte dos alunos, enquanto terça-feira é o dia da semana que mais tem utilização da estação.

Figura 11 – Popularidade semanal da estação Campus do Pici - UFC

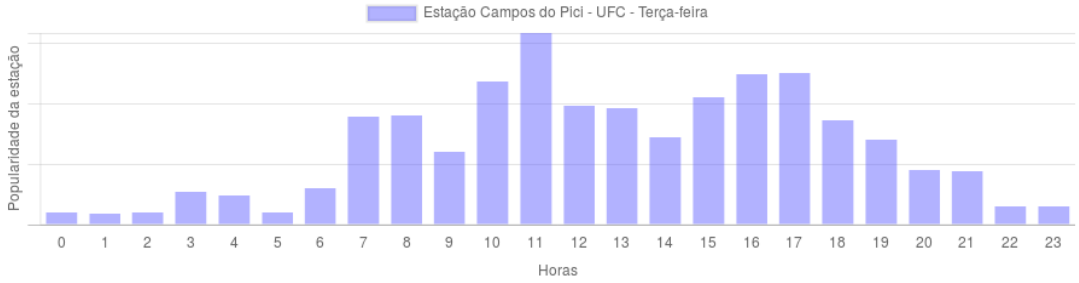


Fonte: O próprio autor (2017)

5. Dado um dia da semana e uma estação, em qual horário ela é mais usada?

   Foi utilizada a mesma ideia da questão 4, porém agrupando os dados por horas. Os gráficos gerados para esta resposta usam o valor máximo (pico) de utilização daquela estação levando em conta todos os dias da semana. Como visto na Figura 12, a estação Campus do Pici - UFC, teve o seu pico às 11 horas em dias de terça-feira.

Figura 12 – Popularidade das horas do dia da semana terça da estação Campus do Pici - UFC

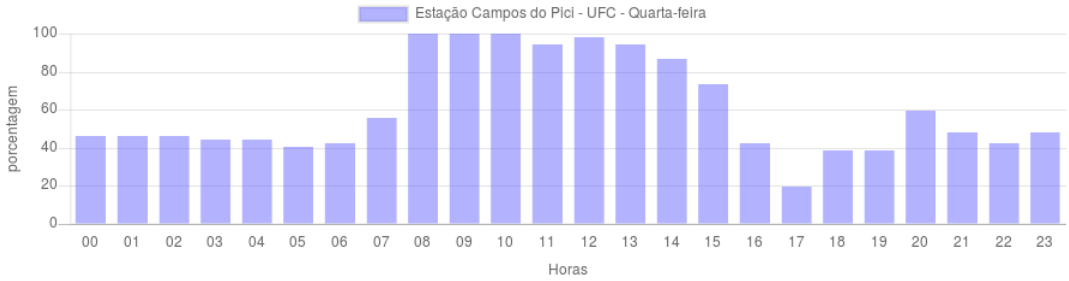


Fonte: O próprio autor (2017)

6. Dado um dia da semana e uma estação, quais horários ela costuma estar com bicicletas disponíveis?

   Nesta pergunta, para obter a resposta, foram feita duas consultas. Na primeira, a partir de um dia da semana, seleciona-se as horas e médias de bicicletas disponíveis, criando um novo atributo que é baseado na média de bicicletas (se a média for maior ou igual a 1, seu valor é definido como 1, se não é definido como 0). Na segunda consulta, é feito um agrupamento por hora, calculando a média dos valores do novo atributo criado e multiplicando por 100, obtendo assim a porcentagem de bicicletas disponíveis para cada hora. Na Figura 13, pode ser visto que a estação Campus do Pici - UFC, na quarta-feita, como de se esperar entre as 7 e 8 horas tem aumento da disponibilidade de bicicletas pois são as horas que os alunos estão chegando inicio das aulas, 16 e 17 horas tem uma baixa na disponibilidade pois são as horas de saídas dos alunos, termino das aulas.

Figura 13 – Porcentagem de disponibilidade nas horas da estação Campus do Pici - UFC na quarta-feira

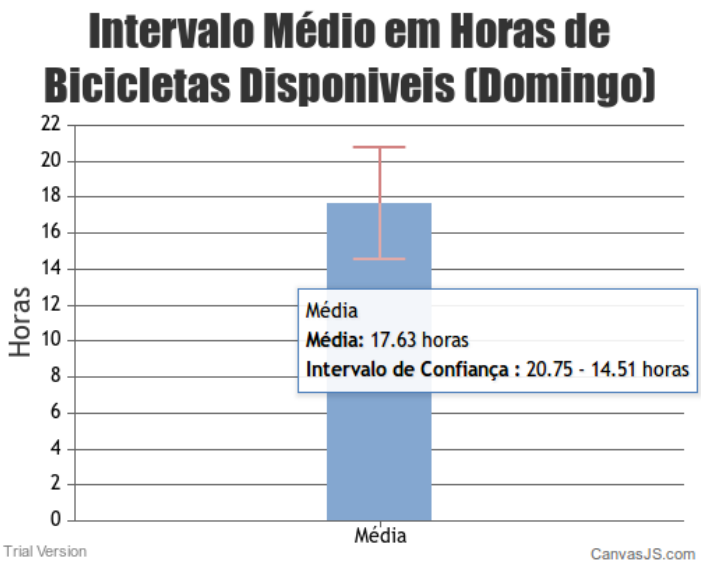


Fonte: O próprio autor (2017)

7. Dado um dia da semana e uma estação, qual a média de tempo em que a estação tem bicicletas disponíveis?

   Para a resposta desta pergunta, foram feita três consultas. Na primeira, dado um dia da semana e uma estação, são selecionados os dias e a média de bicicletas disponíveis e é criado um novo atributo. Para cada tupla do novo atributo, é definido o valor 1 se a média de bicicletas for maior ou igual a 1 e 0 se a média de bicicletas for menor que 1. Na segunda consulta, é obtida a média do novo atributo agrupando os dados pelos dias, e multiplica-se por 24 para obter a disponibilidade em horas. Na terceira consulta é feito o cálculo do intervalo de confiança com 95% de confiabilidade, obtendo assim a média, limite superior e inferior. A Figura 14 mostra o tempo médio que a estação Campus do Pici - UFC tem bicicletas disponíveis, levando em conta o dia da semana domingo - média de 17,63 horas, com intervalo de confiança de 20,75 a 14,51 horas. Como era de se esperar o domingo é o dia de menor utilização da estação, então a estação tem alta media de disponibilidade neste dia da semana.

Figura 14 – Intervalo médio de bicicletas disponíveis da estação Campus do Pici - UFC no domingo

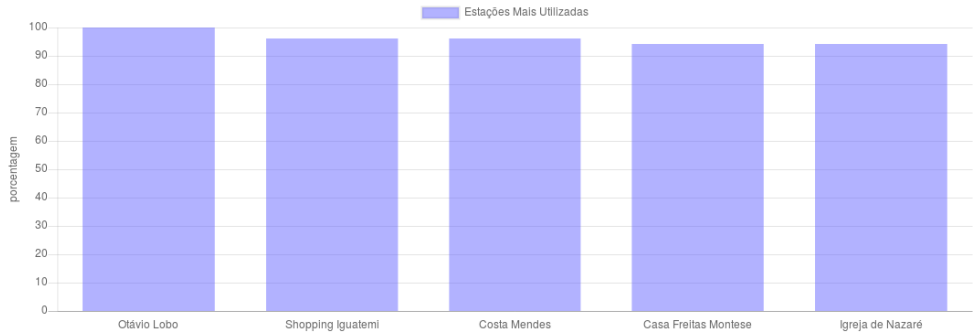


Fonte: O próprio autor (2017)

8. Dado um dia da semana e um horário, quais estações costumam estar com bicicletas disponíveis?

   Esta pergunta foi respondida com duas consultas, selecionado a média de bicicletas a partir de uma hora e um dia da semana, como na questão anterior, foi criado o novo atributo dando o valor 1 se a média de bicicletas for maior ou igual a 1 e 0 se a média de bicicletas for menor que 1. Na segunda consulta é obtido a média do novo atributo multiplicando por 100, obtendo assim a média em porcentagem. Na Figura 15 é mostrado as estações que tem uma porcentagem elevada de bicicletas disponíveis que são Otávio Lobo, Shopping Iguatemi, Costa Mendes, Casa Freitas Montese e Igreja de Nazaré, sendo o dia da semana quarta-feira as 18 horas.

Figura 15 – Estações com porcentagem de disponibilidade de bicicletas na quarta-feira as 18 horas

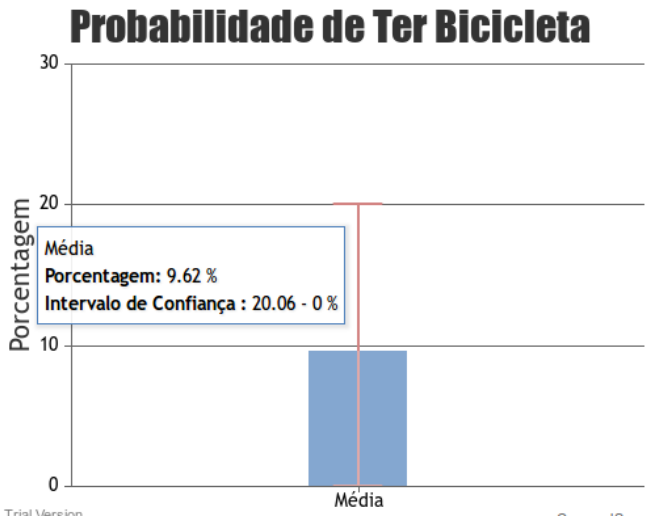


Fonte: O próprio autor (2017)

9. Qual a probabilidade de uma estação ter bicicleta em um determinado dia/horário?

   Esta pergunta foi respondida de forma semelhante à pergunta anterior, porém uma determinada estação foi selecionada. Depois disso, é feito o cálculo do intervalo de confiança com 95% das médias de todos os dias da semana em uma determinada hora retornados. Como visto na Figura 16, a estação Campus do Pici - UFC, na segunda-feira de 17 horas a 18 horas, obteve uma média de 9,62% de chance de ter bicicletas disponíveis, uma média muito baixa pois como já descrito antes, 17 horas é o horário de termino das aulas.

Figura 16 – Probabilidade de ter bicicletas disponíveis da estação Campus do Pici - UFC às 17-18 horas

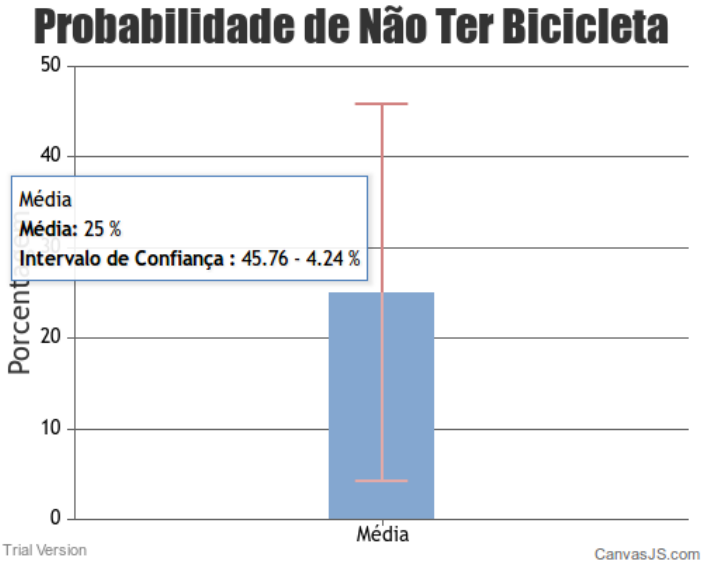


Fonte: O próprio autor (2017)

10. Qual a probabilidade de uma estação não ter bicicleta em um determinado dia/horário?

    Esta pergunta foi respondida semelhante à pergunta anterior, porém o novo atributo criado foi definido como 1 quando a média de bicicletas era menor que 1, e 0 quando a média era maior ou igual a 1. Depois disso, é feito o cálculo do intervalo de confiança com 95%. A Figura 17 mostra a porcentagem e o intervalo de confiança de não ter bicicletas disponíveis, levando em conta o dia da semana domingo às 10-11 horas, da estação Campus do Pici - UFC. A média obtida foi de 25%. Uma média baixa, pois como já comprovado domingo não é dia de muita utilização.

Figura 17 – Probabilidade de não ter bicicletas disponíveis da estação Campus do Pici - UFC domingo às 10-11 horas

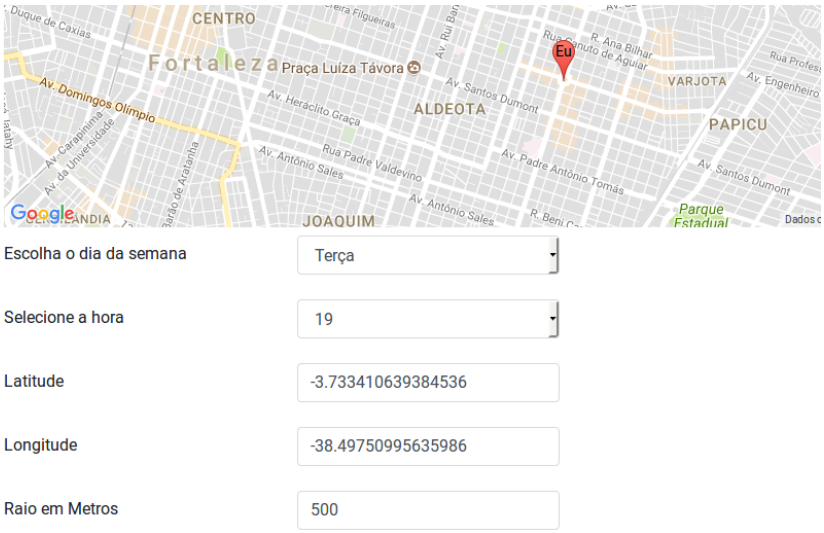


Fonte: O próprio autor (2017)

11. Qual estação seria escolhida como melhor opção, para uma pessoa que deseja uma bicicleta em uma determinada área, dia/horário?

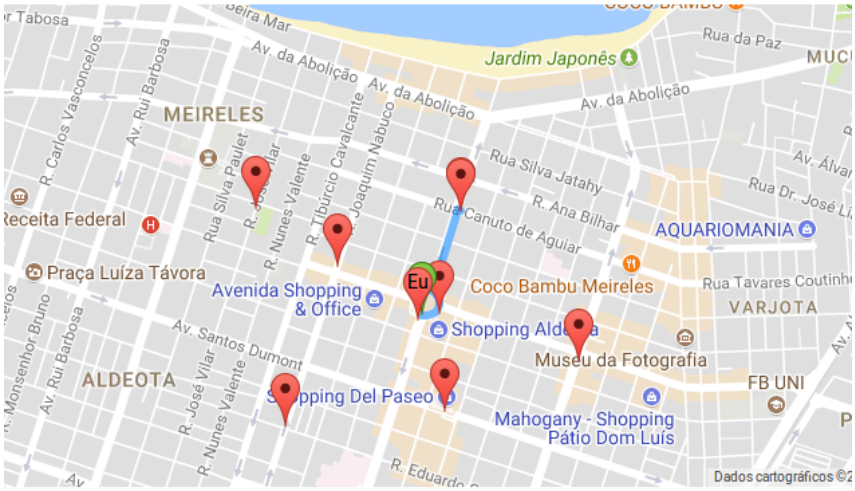
    Para esta pergunta, dado a latitude, longitude e um raio, retorna-se quais estações estão dentro desse perímetro, para isto é utilizado o modulo *earthdistance* do *PostgreSQL* que trabalha com dados espaciais, depois é feita a mesma consulta da pergunta 9 para todas as estações retornadas. Na Figura 18, são mostrados os parâmetros que foram utilizados para demostração da resposta. Dado o dia da semana terça-feira às 19 horas, com latitude, longitude mostrados na Figura 18 e raio de 500 metros, foram retornadas as estações mostradas na Figura 19. A Figura 20 mostra a tabelas das estações retornadas junto com seus respectivos endereços e porcentagem de disponibilidade. Nesta consulta, a melhor estação foi a Círculo Militar com 98,08% de disponibilidade de bicicletas. A ferramenta também apresenta a melhor rota para a estação selecionada através da própria API do *google*, como apresentado na Figura 19. Ao clicar em qualquer estação, sua rota é calculada e apresentada.

Figura 18 – Parâmetros da consulta



Fonte: O próprio autor (2017)

Figura 19 – Estações retornadas com o caminho para a estação com maior porcentagem de disponibilidade



Fonte: O próprio autor (2017)

Figura 20 – Tabela das estações retornadas

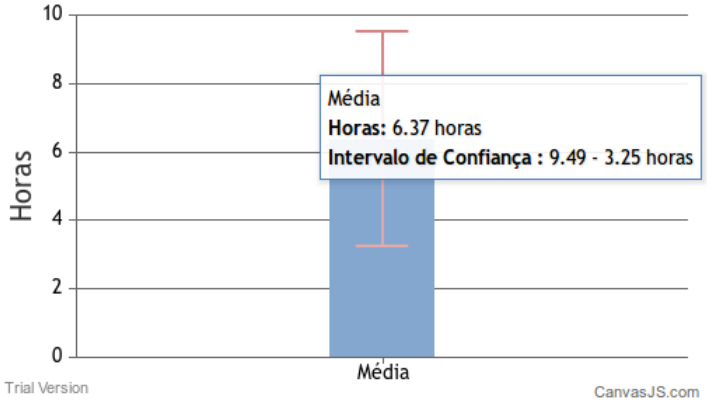
Estação	Endereço	Porcentagem de Bicicletas Disponíveis
Circulo Militar	Rua Canuto de Aguiar, 712B / Esquina Rua Desembargador Moreira	98.08%
Campo do América	Rua José Vilar, 540 / Esquina Rua Tenente Benévolo	59.62%
Joaquim Nabuco	Rua Joaquim Nabuco, 730 / Esquina Rua Dom Luís	55.77%
Torres Câmara	Rua Joaquim Nabuco, em frente ao Edifício Embratel / Esquina Rua Torres Câmara	42.31%
Livraria Cultura	Rua Senador Virgílio Távora, lateral do número 1010 (Shopping Varanda) / Esquina Rua Dom Luís	40.38%
Praça Portugal	Praça Portugal, na calçada em frente a canteiro do Shopping Aldeota / Esquina Avenida Dom Luís	36.54%
Shopping Del Paseo	Avenida Santos Dumont, 3131 (Shopping del Paseo) / Esquina Rua Leonardo Mota	34.62%

Fonte: O próprio autor (2017)

12. Qual a média de tempo que uma estação fica sem bicicletas dado um dia da semana?

    Esta questão é o oposto da questão 7. Assim, muda-se apenas a definição do novo atributo criado, que agora é definido como 1 quando a média de bicicletas for menor que 1, e 0 quando a média for maior ou igual a 1. A Figura 21 mostra o intervalo médio de indisponibilidade de bicicletas da estação Campus do Pici - UFC no domingo, onde obteve-se uma média 6,37 horas. Média baixa, reforçando ainda mais o que já foi descrito anteriormente.

Figura 21 – Intervalo médio de indisponibilidade de bicicletas da estação Campus do Pici - UFC no domingo



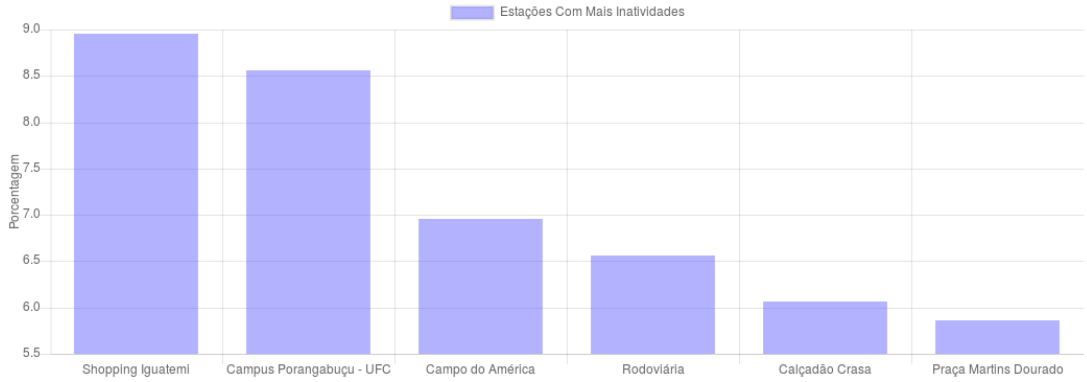
Fonte: O próprio autor (2017)

13. Quais as estações que tiveram mais e menos inatividades?

    Para esta consulta, foi selecionado o atributo ativo, que contém a porcentagem de tempo que a estação está ativa. Assim, para saber a porcentagem de inatividade da estação, é feito o cálculo 1, menos o valor da porcentagem da estação ativa. Depois que é feito este cálculo, é calculada a média desses valores obtidos, multiplicando por 100, para obter o valor em

porcentagem e agrupando os dados pelos IDs das estações. Na Figura 22, são mostradas as estações com as maiores porcentagens de inatividades, sendo elas: Shopping Iguatemi, Campus Porangabuçu - UFC, Campo do América, Rodoviária, Calçadão Crasa e Praça Martins Dourado. Enquanto a estação Esplanada Montese obteve a menor porcentagem de inatividade, seguidas pelas estações Francisco Matos, Igreja Redonda, Extra Aguanambi, Faculdade Lourenço Filho e Igreja de Fátima mostradas na Figura 23.

Figura 22 – Estações com mais inatividades



Fonte: O próprio autor (2017)

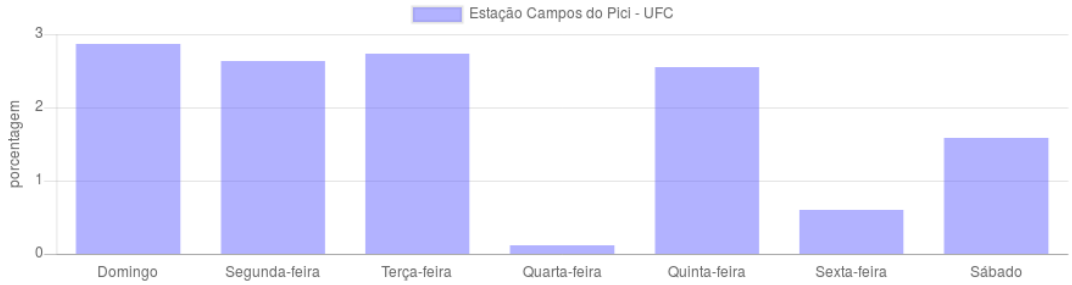
Figura 23 – Estações com menos inatividades

Fonte: O próprio autor (2017)

14. Quais dias da semana uma estação costuma ficar inativa?

    Para esta pergunta, é feito o cálculo de inatividade descrito na questão anterior, com esses valores obtidos do cálculo, é feita a média desses valores, multiplicando por 100, para obter o valor em porcentagem e agrupando os dados pelos dias da semana. Na Figura 24 é apresentada a porcentagem semanal de inatividade da estação Campus do Pici - UFC, domingo, segunda-feira,terça-feira e quarta-feira são os dias da semana com maiores porcentagens de inatividade, porém este valor não chega a 3% de inatividade.

Figura 24 – Porcentagem semanal de inatividade da estação Campus do Pici - UFC

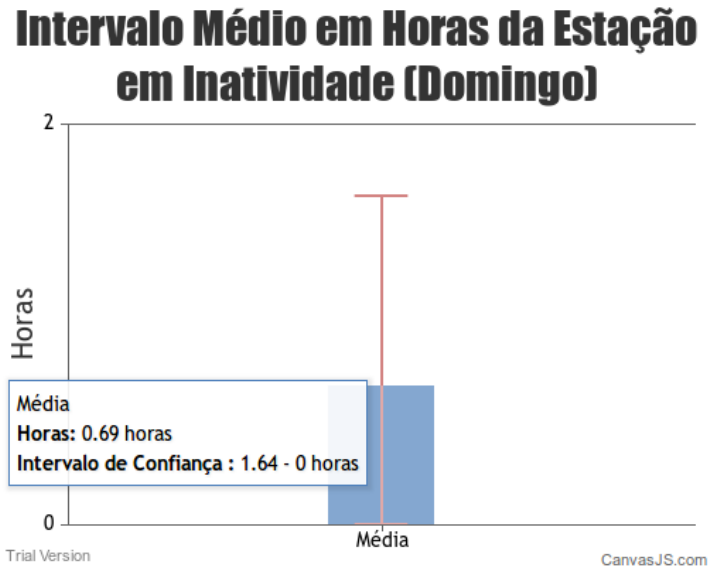


Fonte: O próprio autor (2017)

15. Qual a média de tempo que uma estação fica inativa dado um dia da semana?

    Esta consulta é semelhante à anterior, porém os dados são selecionados a partir de um dia da semana e é feito o agrupamento dos dados pelos dias. Depois de obtida a média de inatividade dos dias, o valor é multiplicado por 24, conseguindo assim a representação em horas. Em seguida, é feito o cálculo do intervalo de confiança com 95%. Na Figura 25, é apresentado o intervalo médio em horas de inatividade no domingo da estação Campus do Pici - UFC. A média resultante não chega a 1 hora.

Figura 25 – Intervalo médio em horas de inatividade da estação Campus do Pici - UFC no domingo



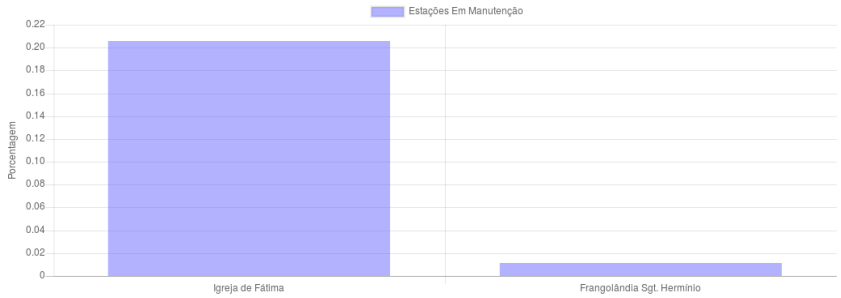
Fonte: O próprio autor (2017)

16. Quais estações ficaram em manutenção?

    Para responder esta pergunta, foram utilizadas duas consultas. A primeira foi utilizada para descobrir quais estações ficaram em manutenção e a segunda para calcular a porcentagem

de tempo que elas ficaram em manutenção em relação ao tempo de operação. Para descobrir as estações em manutenção, foi verificado se o atributo *emoperacao* era diferente de 1, sendo que 1 representa 100%, depois foi feito o cálculo para descobrir a porcentagem de manutenção, que foi 1 menos o valor de *emoperacao*. Com o resultado do cálculo, os dados foram agrupados pelos ID's das estações, tirando a média do valor calculado e multiplicando por 100 obtendo assim a porcentagem. Na Figura 26, são mostradas as estações que ficaram em manutenção: Igreja de Fátima e Frangolândia Sgt. Hermínio.

Figura 26 – Estações em manutenção

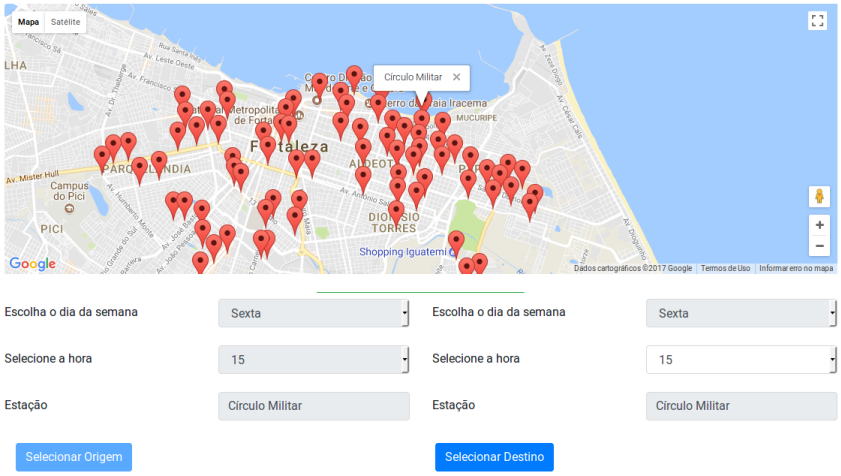


Fonte: O próprio autor (2017)

17. Qual a probabilidade de pegar uma bicicleta em uma estação em um determinado dia e horário e conseguir devolver a bicicleta em outra/mesma estação em outro horário?
    Esta questão foi divida em dois passos. O primeiro passo é descobrir a porcentagem de disponibilidade de bicicletas dado um dia da semana e horário da estação de origem, e o segundo passo é descobrir a porcentagem de ter vaga para bicicletas na estação de destino dado dia da semana e horário.
    No primeiro passo, a consulta foi semelhante à pergunta 6, porém selecionando uma hora específica. Com isso, é retornado o valor em porcentagem de bicicleta disponível naquele dia e horário. No segundo passo, a consulta também é semelhante à pergunta 6, mas os dados selecionados não são as médias de bicicletas e sim as médias de vagas. O restante da consulta é a mesma, porém selecionando uma hora especifica, com isso é obtido a porcentagem de ter vaga dado o dia da semana e horário. Depois, na página *web*, é feita a multiplicação dessas duas porcentagens e o valor resultante divido por 100 obtendo a porcentagem do evento acontecer. Na Figura 27, são mostrados os parâmetros de uma consulta, dado a estação de origem Círculo Militar, dia da semana sexta às 15 horas e a própria estação de destino dia da semana sexta às 15 horas. A probabilidade de pegar uma bicicleta e deixar na estação conforme os parâmetros descritos, resultou em uma

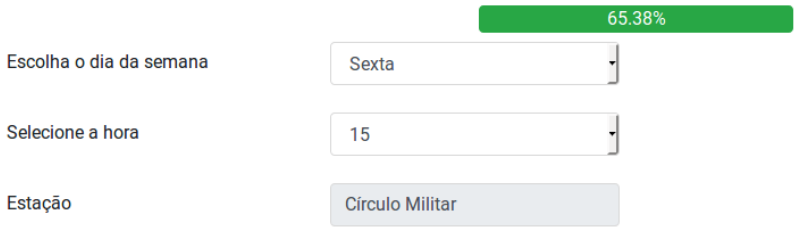
probabilidade de 65,38%, como mostrado na Figura 28.

Figura 27 – Parâmetros da consulta



Fonte: O próprio autor (2017)

Figura 28 – Resultado da consulta: Sexta, 15 horas, Círculo Militar



Fonte: O próprio autor (2017)

6 CONCLUSÃO

Com o aumento do transporte individual vem o aumento significativo dos congestionamentos, que por sua vez trás atrasos, insatisfação e perdas financeiras. A bicicleta entra como um meio alternativo, opção de transporte sustentável, não poluente e, além disso, é uma escolha de transporte de pequeno trajeto, que facilita o deslocamento das pessoas. É de total importância entender a mobilidade urbana, com ela é possível trazer benefícios tanto para o espaço urbano como para a locomoção das pessoas.

Neste trabalho, foi criada uma ferramenta para fazer análise dos dados do programa de bicicletas compartilhadas Bicicletar da cidade de Fortaleza-CE. Um *script* foi implementado para obter os dados do site oficial do programa, bem como para fazer o pré-processamento destes. Houve a realização de agregação dos dados coletados, salvando-os em um banco de dados. Também foram criadas consultas para responder as perguntas de uma lista de perguntas criada pensando no interesse de usuários e administradores do programa Bicicletar. Para apresentar as perguntas e suas respostas, foi desenvolvido um *web site*. Como apresentado no Capítulo 5, a ferramenta proposta responde questões que o site oficial do Bicicletar não responde e as respostas são mostradas de forma clara por gráficos para os usuários.

Com o trabalho concluído, conseguimos alcançar os objetivos propostos, o backup do banco de dados, os *scripts* de coleta, processamento e os arquivos de criação do *web site* estão disponíveis no GitHub[1] e o *web site* pode ser acessado pelo link http://54.207.71.5/site-tcc/.

Não foi realizada a avaliação da ferramenta, pois seria mais eficiente se fosse realizada com os usuários do programa Bicicletar. Além disso, inicialmente pensou-se em criar um *Data Warehouse* (armazém de dados), mas não foi necessário, pois não foram coletados muitos dados nos 4 meses de coletas, e, além disso, apenas uma fonte de dados foi utilizada (site do Bicicletar), o que simplificou o trabalho. Porém, fica como trabalho futuro a criação de um *Data Warehouse*, pois ele suporta a crescente inserção de dados, facilitando assim as consultas em uma base de dados com tamanho maior. Outro trabalho futuro interessante é a criação de uma aplicação para dispositivos móveis, com as mesmas funcionalidades da nossa *web*. Por fim, utilizar algoritmos de aprendizado de máquina para identificar padrões no uso das estações também parece um trabalho futuro interessante.

[1] https://github.com/fabtrompet/analisededadosbicicletar

REFERÊNCIAS

BENEVENUTO, F.; ALMEIDA, J. M.; SILVA, A. S. Explorando redes sociais online: Da coleta e análise de grandes bases de dados às aplicações. **Porto Alegre: Sociedade Brasileira de Computação**, 2011.

BICICLETAR. **Bicicletar - Bicicletas compartilhadas de Fortaleza**. [S.l.], 2017. Bicicletar. Disponível em: <http://www.bicicletar.com.br/>. Acesso em: 10 abr. 2017.

CABEÇA, A. G. et al. Análise de mutantes em aplicações sql de banco de dados. **Universidade Estadual de Campinas (UNICAMP)**, 2009.

CHEN, B.; PINELLI, F.; SINN, M.; BOTEA, A.; CALABRESE, F. Uncertainty in urban mobility: Predicting waiting times for shared bicycles and parking lots. In: **16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)**. [S.l.: s.n.], 2013. p. 53–58. ISSN 2153-0009.

ELMASRI, R.; NAVATHE, S. B. **Fundamentals of database systems**. [S.l.]: Pearson, 2010.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37, 1996.

JÚNIOR, A. M. S.; SOUSA, M. L.; XAVIER, F. Z.; XAVIER, W. Z.; ALMEIDA, J. M.; ZIVIANI, A.; RANGEL, F.; AVILA, C.; MARQUES-NETO, H. T. Caracterizaç ao do serviço de táxi a partir de corridas solicitadas por um aplicativo de smartphone. **XXXIV Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC '16)**, 2016.

LIU, J.; LI, Q.; QU, M.; CHEN, W.; YANG, J.; XIONG, H.; ZHONG, H.; FU, Y. Station site optimization in bike sharing systems. In: **2015 IEEE International Conference on Data Mining**. [S.l.: s.n.], 2015. p. 883–888. ISSN 1550-4786.

MONCAYO-MARTíNEZ, L. A.; RAMIREZ-NAFARRATE, A. Visualization of the mobility patterns in the bike-sharing transport systems in mexico city. In: **2016 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)**. [S.l.: s.n.], 2016. p. 1851–1855.

PRATA, P.; SANCHES, R. Mobilidade urbana no municÍpio de presidente prudente. **ETIC - ENCONTRO DE INICIAÇÃO CIENTÍFICA - ISSN 21-76-8498**, v. 12, n. 12, 2016. Disponível em: <http://intertemas.toledoprudente.edu.br/revista/index.php/ETIC/article/view/5416/5148>. Acesso em: 22 maio 2017.

PURNAMA, I. B. I.; BERGMANN, N.; JURDAK, R.; ZHAO, K. Characterising and predicting urban mobility dynamics by mining bike sharing system data. In: **2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom)**. [S.l.: s.n.], 2015. p. 159–167.

SANTOS, L. M. Protótipo para mineração de opinião em redes sociais: estudo de casos selecionados usando o twitter. **Universidade Federal de Lavras (UFLA)**, 2010.

SILVA, T. H.; LOUREIRO, A. Computaçao urbana: Técnicas para o estudo de sociedades com redes de sensoriamento participativo. **Anais da XXXIV JAI**, v. 8329, p. 68–122, 2015.

SOARES, R. D. G. Bicicleta e mobilidade urbana: Modismo ou solução sustentável para o transporte na cidade de são paulo. **Centro de Estudos Latino-Americanos sobre Cultura e Comunicação (CELACC)**, 2015. Disponível em: <http://www.usp.br/celacc/?q=celacc-tcc/819/detalhe>. Acesso em: 14 abr. 2017.

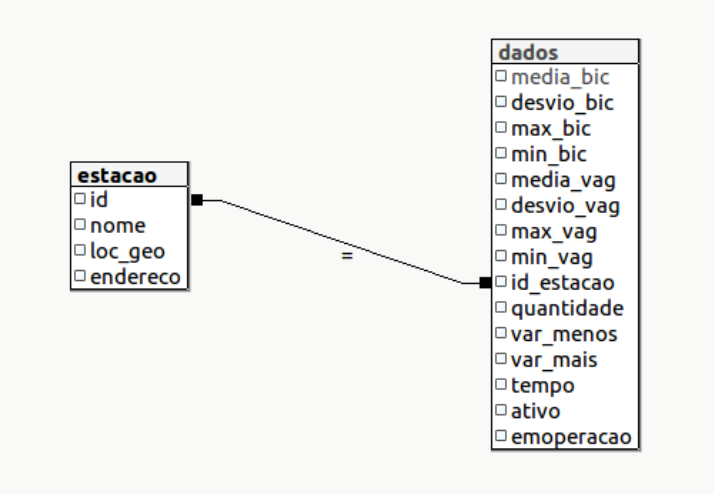
SPOTO, E. S. et al. Teste estrutural de programas de aplicação de banco de dados relacional. **Universidade Estadual de Campinas (UNICAMP)**, 2000.

SUNDARAMOORTHY, K.; DURGA, R.; NAGADARSHINI, S. Newsone—an aggregation system for news using web scraping method. In: IEEE. **Technical Advancements in Computers and Communications (ICTACC), 2017 International Conference on**. [S.l.], 2017. p. 136–140.

WANG, X.; DING, L.; WANG, Q.; XIE, J.; WANG, T.; TIAN, X.; GUAN, Y.; WANG, X. A picture is worth a thousand words: Share your real-time view on the road. **IEEE Transactions on Vehicular Technology**, v. 66, n. 4, p. 2902–2914, April 2017. ISSN 0018-9545.

APÊNDICE A – TABELAS DO BANCO DE DADOS

Figura 29 – Diagrama do Banco de Dados



Fonte: O próprio autor (2017)

APÊNDICE B – CONSULTAS UTILIZADOS PARA RESPONDER AS PERGUNTAS PROPOSTAS

Código-fonte 1 – Quais estações mais e menos utilizadas?

```
SELECT id_estacao, nome, sum(var_menos) + sum(var_mais) AS
   total FROM dados, estacao WHERE id = id_estacao GROUP BY
    id_estacao,nome ORDER BY total (DESC/ASC) LIMIT 6
```

Código-fonte 2 – Dado um dia da semana, quais as estações mais e menos utilizadas?

```
SELECT id_estacao, nome,sum(var_mais + var_menos) AS total
   FROM dados,estacao WHERE EXTRACT(dow FROM tempo) = (0-6)
    AND id_estacao = id GROUP BY id_estacao,nome ORDER BY
   total (DESC/ASC) LIMIT 5
```

Código-fonte 3 – Quais estações são mais utilizadas em um determinado dia/horário?

```
SELECT id_estacao,nome,sum(var_mais + var_menos) AS total
   FROM dados,estacao WHERE EXTRACT(dow FROM tempo) = (0-6)
    AND EXTRACT(hour FROM tempo) = (0-23) AND id_estacao =
   id GROUP BY id_estacao,nome ORDER BY total DESC LIMIT 5
```

Código-fonte 4 – Quais os dias da semana em que uma determinada estação é mais usada?

```
SELECT sum(var_menos) + sum(var_mais) AS total, EXTRACT(dow
    FROM tempo) AS dia_semana FROM dados WHERE id_estacao =
    (1-80) GROUP BY id_estacao,dia_semana ORDER BY
   dia_semana
```

Código-fonte 5 – Dado um dia da semana e uma estação, em qual horário ela é mais usada?

```
SELECT sum(total) AS total, hora FROM (SELECT sum(var_menos
) + sum(var_mais) AS total, EXTRACT(hour FROM tempo) AS
hora,tempo, EXTRACT(dow FROM tempo) AS dia FROM dados
WHERE id_estacao = (1-80) AND EXTRACT(dow FROM tempo) =
(0-23) GROUP BY id_estacao,tempo ORDER BY hora DESC) grp
 GROUP BY hora ORDER BY hora
```

Código-fonte 6 – Dado um dia da semana e uma estação, quais horários ela costuma estar com bicicletas disponíveis?

```
SELECT AVG(total) * 100,hora FROM (SELECT media_bic,to_char
(tempo,  HH24 ) AS hora, CASE WHEN media_bic>=1 THEN 1
WHEN media_bic<1 THEN 0 END AS total FROM dados WHERE
id_estacao = (1-80) AND EXTRACT(dow FROM tempo) = (0-23)
 ORDER BY hora DESC) grp GROUP BY hora ORDER BY hora
```

Código-fonte 7 – Dado um dia da semana e uma estação, qual a média de tempo em que a estação tem bicicletas disponíveis?

```
SELECT ( 1.96 * (stddev_samp(total)/ (sqrt(COUNT(total)))))
 AS error,AVG(total) AS media FROM (SELECT AVG(total)
*24,dias FROM (SELECT media_bic,to_char(tempo,  YYYY-MM-
DD ) AS dias, CASE WHEN media_bic>=1 THEN 1 WHEN
media_bic<1 THEN 0 END AS total FROM dados WHERE
id_estacao = (1-80) AND EXTRACT(dow FROM tempo) = (0-6)
) grp GROUP BY dias) gpr2
```

Código-fonte 8 – Dado um dia da semana e um horário, quais estações costumam estar com bicicletas disponíveis?

```
SELECT id_estacao,nome,AVG(total) * 100  FROM (SELECT
id_estacao,nome,media_bic,to_char(tempo,  HH24 ) AS hora
, CASE WHEN media_bic>=1 THEN 1 WHEN media_bic<1 THEN 0
```

```
END AS total FROM dados, estacao WHERE id = id_estacao
AND EXTRACT(dow FROM tempo) = (0-6) AND to_char(tempo,
HH24 ) =  (01-23)  ORDER BY hora DESC) grp GROUP BY nome
,id_estacao ORDER BY total DESC LIMIT 5
```

Código-fonte 9 – Qual a probabilidade de uma estação ter bicicleta em um determinado dia/horário?

```
1 SELECT (1.96 * (stddev_samp(total)/ (sqrt(COUNT(total))))
  AS error, AVG(total) AS media FROM (SELECT AVG(total)
  *100, dia FROM (SELECT to_char(tempo,  YYYY-MM-DD ) AS
  dia,media_bic, CASE WHEN media_bic>=1 THEN 1 WHEN
  media_bic<1 THEN 0 END AS total FROM dados WHERE
  id_estacao = (1-80) AND EXTRACT(dow FROM tempo) = (0-6)
  AND to_char(tempo,  HH24 ) = ( 01-23 )) grp GROUP BY dia
  ) grp2
```

Código-fonte 10 – Qual a probabilidade de uma estação não ter bicicleta em um determinado dia/horário?

```
1 SELECT (1.96 * (stddev_samp(total)/ (sqrt(COUNT(total))))
  AS error, AVG(total) AS media FROM (SELECT AVG(total)
  *100, dia FROM (SELECT to_char(tempo,  YYYY-MM-DD ) AS
  dia,media_bic, CASE WHEN media_bic>=1 THEN 0 WHEN
  media_bic<1 THEN 1 END AS total FROM dados WHERE
  id_estacao = (1-80) AND EXTRACT(dow FROM tempo) = (0-6)
  AND to_char(tempo,  HH24 ) =  01-23 ) grp GROUP BY dia)
  grp2
```

Código-fonte 11 – Qual estação seria escolhida como melhor opção, para uma pessoa que deseja uma bicicleta em uma determinada área, dia/horário?

```
SELECT lat,lng,nome,endereco,id, AVG(total)*100 FROM(SELECT
    loc_geo[0] AS lat, loc_geo[1] AS lng,id,nome,endereco,
   media_bic,to_char(tempo, HH24 ) AS hora, CASE WHEN
   media_bic>=1 THEN 1 WHEN media_bic<1 THEN 0 END AS total
    FROM dados,estacao WHERE id_estacao = id AND EXTRACT(
   dow FROM tempo) = (0-6) AND to_char(tempo, HH24 ) =
   (01-23) AND earth_box(ll_to_earth(loc_geo[0],loc_geo
   [1]), (raio em metros)) @> ll_to_earth((latitude), (
   longitude))) grp GROUP BY nome,endereco,lat,lng,id ORDER
    BY total DESC
```

Código-fonte 12 – Qual a média de tempo que uma estação fica sem bicicletas dado um dia da semana?

```
SELECT (1.96 * (stddev_samp(total)/ (sqrt(COUNT(total)))))
   AS error, AVG(total) AS media FROM (SELECT AVG(total)
   *24, dias FROM (SELECT media_bic,to_char(tempo, YYYY-MM
   -DD ) AS dias, CASE WHEN media_bic>=1 THEN 0 WHEN
   media_bic<1 THEN 1 END AS total FROM dados WHERE
   id_estacao = (1-80) AND EXTRACT(dow FROM tempo) = (0-6)
   ) grp GROUP BY dias) grp
```

Código-fonte 13 – Quais AS estações que tiveram mais e menos inatividades?

```
SELECT id_estacao,nome,AVG((1-ativo))*100 AS total FROM
   dados,estacao WHERE id = id_estacao GROUP BY id_estacao,
   nome ORDER BY total (DESC/ASC) LIMIT 6
```

Código-fonte 14 – Quais dias da semana que uma estação costuma ficar inativa?

```
SELECT EXTRACT(dow FROM tempo) AS dia_semana,AVG((1-ativo))
   *100 AS total FROM dados WHERE id_estacao = (1-80) GROUP
    BY dia_semana ORDER BY dia_semana
```

Código-fonte 15 – Qual a média de tempo que uma estação fica inativa dado um dia da semana?

```
SELECT AVG(media) AS media, (1.96 * (stddev_samp(media)/ (
   sqrt(COUNT(media))))) AS error FROM (SELECT AVG(ativo)
   *24 AS media FROM (SELECT (1-ativo) AS ativo,to_char(
   tempo,  YYYY-MM-DD ) AS dias FROM dados WHERE EXTRACT(
   dow FROM tempo) = (0-6) AND id_estacao = (1-80)) grp
   GROUP BY dias) grp2
```

Código-fonte 16 – Qual a probabilidade de pegar uma bicicleta em uma estação em um determinado dia e horário e conseguir devolver a bicicleta em outra/mesma estação em outro horário?

```
#primeira consulta porcentagem disponibilidade
SELECT AVG(teste)*100 AS total FROM (SELECT media_vag,CASE
   WHEN media_vag>=1 THEN 1 WHEN media_vag<1 THEN 0 END AS
   teste FROM dados WHERE id_estacao = (1-80) AND EXTRACT(
   dow FROM tempo) = (0-6) AND to_char(tempo,  HH24 ) =
   01-23 ) grp
#segunda consulta porcentagem vaga
SELECT AVG(teste)*100 AS total FROM (SELECT media_vag,CASE
   WHEN media_vag>=1 THEN 1 WHEN media_vag<1 THEN 0 END AS
   teste FROM dados WHERE id_estacao = (1-80) AND EXTRACT(
   dow FROM tempo) = (0) AND to_char(tempo,  HH24 ) =  0-23
    ) grp
```