Universidade Federal de Pernambuco
Centro de Informática

Graduação em Ciência da Computação

Uma Ferramenta para a Melhoria da Qualidade de Dados Estruturados da Web

Lucas Nunes de Souza

Trabalho de Graduação

Recife
Julho de 2017

Universidade Federal de Pernambuco
Centro de Informática

Lucas Nunes de Souza

Uma Ferramenta para a Melhoria da Qualidade de Dados Estruturados da Web

Trabalho apresentado ao Programa de Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: *Prof. Luciano de Andrade Barbosa*

Recife
Julho de 2017

Agradecimentos

Gostaria de agradecer a minha família por todo o suporte durante esses anos e que me permitiram focar na minha formação.

Ao professor Luciano de Andrade Barbosa, que tornou esse trabalho possível. Agradeço pela oportunidade e disponibilidade em momentos de dúvida.

Agradeço a todos os amigos da universidade que sempre estavam dispostos a ajudar, principalmente em tempos de prova.

E a todos os professores e funcionários que fazem com que o Centro de Informática seja esse centro de excelência. Foi um prazer estudar nesse centro e eu aprendi muito.

Resumo

A Internet contém uma grande quantidade de dados que podem ser usados para diversos propósitos, por exemplo, para modelagem de dados, análises estatísticas e previsões. No entanto, dados obtidos da web podem conter vários problemas como campos faltando, duplicatas ou informações erradas. Esses problemas podem surgir de algum erro nas fontes de dados ou no método de obtenção dos dados. Detectar e tratar esses problemas pode exigir muito trabalho e geralmente requer o uso de diferentes ferramentas e técnicas estatísticas. O principal objetivo desse trabalho é criar uma solução para ajudar o usuário nesses processos, de forma que ele possa utilizar, de forma integrada, diferentes métodos estatísticos e de visualização para ajudá-lo no processo de melhoria da qualidade de dados.

Palavras-chave: recuperação de informação, qualidade de dados, web

Abstract

The Internet contains a large amount of data that can be used for various purposes, for example, for data modeling, statistical analysis and forecasting. However, data obtained from the web may contain various problems such as missing fields, duplicates, or wrong information. These problems may arise from some error in the data sources or in the method of obtaining the data. Detecting and treating these problems can require a lot of work and usually needs the use of different statistical tools and techniques. The main goal of this work is to create a solution to help users in these processes, so that they can use different statistical and visualization methods in an integrated way to help them in the process of data quality improvement.

Keywords: information retrieval, data quality, web

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

Na era da internet em que vivemos, entretenimento e informações são predominantemente digitais e a quantidade de dados que são gerados cada dia é enorme. Podemos encontrar todo tipo de informação na Web, desde informações e preços de produtos até dados e artigos científicos. Esses dados podem ser usados, por exemplo, para tarefas de previsões e tomadas de decisão [1].

Com o aumento de automação na forma de adquirir os dados, a preocupação com a qualidade dos dados aumenta. Normalmente é necessário que os dados estejam no correto estado e representem fielmente o mundo real [2]. A presença de dados incorretos ou inconsistentes podem distorcer significativamente resultado de análises, potencialmente negando os benefícios de usar métodos e algoritmos de previsão e análise de dados [3].

No contexto da Web, estamos interessados em dados estruturados, aqueles que possuem atributos e valores de um determinado domínio. Por exemplo, um produto numa loja online pode ter atributos preço, nome, estado, etc. Esses tipos de dados são de mais fácil manipulação e análise. Dados extraídos de sites com informação estruturada podem conter problemas como: valores faltando, dados duplicados, erro de formatação, entre outros. As duas principais fontes de erros de dados obtidos da web são:

- Fonte dos dados: Problemas podem surgir tanto na entrada manual de dados, campos errados e faltando, ou na geração automática de páginas, erros de formatação e valores duplicados. Por exemplo, um site de imóveis pode conter valores faltando como vagas ou suítes, informações normalmente não obrigatórias.

- Método extração: Métodos de extração automatizados podem ter problemas para extrair informação devido à diferença de estrutura que cada página pode ter. Até em um único domínio, não há garantia que os dados vão estar no mesmo formato.

Nas Figuras 1.1 e 1.2, temos exemplos desses tipos de problema. A primeira é um exemplo de site que tem informações difíceis de serem extraídas por um extrator genérico. O problema se encontra nas informações de quartos e banheiros, que embora estejam presentes, são representadas por uma imagem, o que torna a extração difícil. Na segunda imagem o problema é do valor disponível no site, o apartamento em questão contém a informação de 75 vagas de garagem e pode ser considerado um outlier em relação a outros com perfil parecido.

Para poder lidar com esses problemas, diversos métodos têm sido propostos na área de qualidade de dados e limpeza de dados. O processo de melhoria dos dados possui alguns desafios. Métodos diferentes podem ter resultados diferentes ou até mesmo um mesmo método pode ser melhor para um determinado conjunto de dados que outro. O processo de limpeza também pode criar novos erros [4]. Esse processo pode ser interativo, no qual o usuário pode

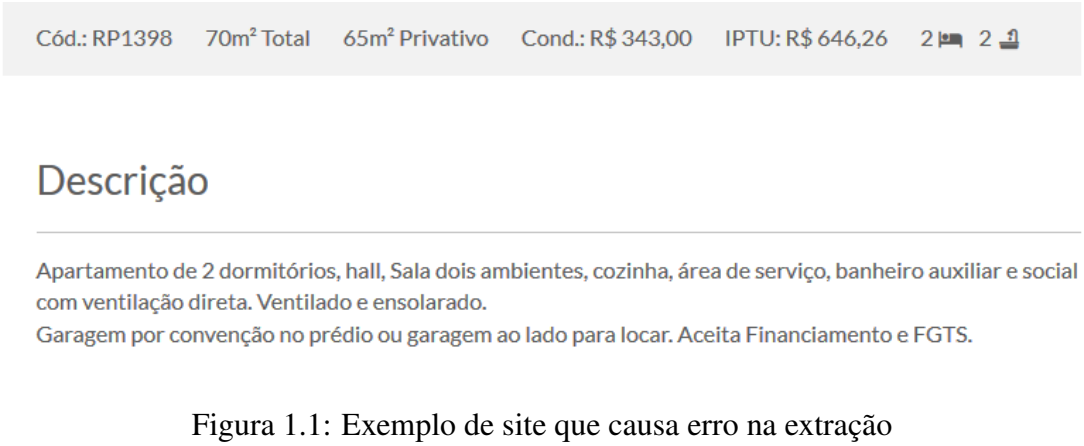


Figura 1.1: Exemplo de site que causa erro na extração

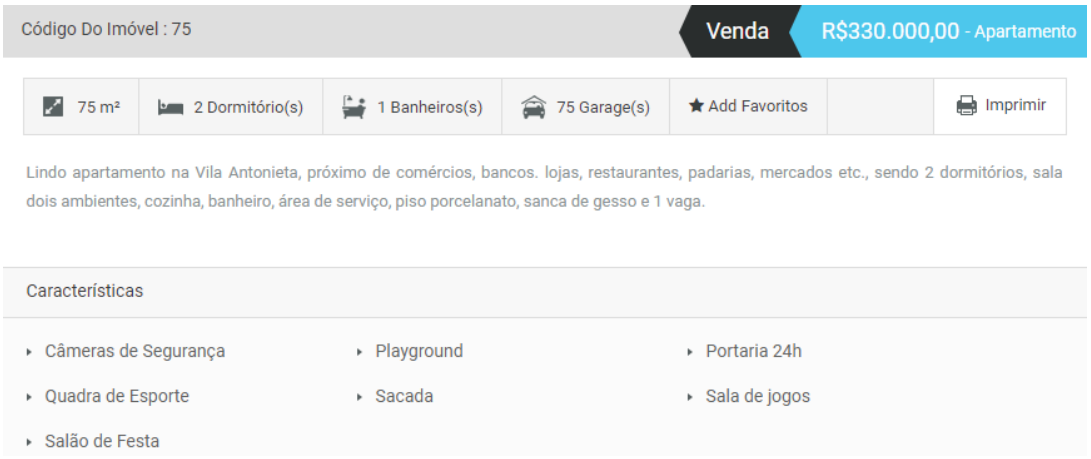


Figura 1.2: Exemplo de site com outlier nas informações

analisar cada etapa e decidir qual o melhor resultado [5]. Desse modo, pretende-se criar nesse trabalho uma ferramenta de suporte à melhoria na qualidade de dados estruturados que auxilie o tratamento e estudo de dados adquiridos da web. Para isso usamos técnicas estatísticas para tratar dos problemas mais comuns encontrados nos dados enquanto o usuário pode visualizar e analisar informações sobre os dados durante o processo.

A principal motivação para a criação desse trabalho é a falta de ferramentas específicas que ajudem a detectar problemas como outliers e usem métodos estatísticos. Esses tipos de operações ainda são normalmente realizadas manualmente através de uma planilha ou software do tipo. Ou quando é automatizado, ainda é necessário criar um script em linguagens como python e R, que requer algum conhecimento extra.

As ferramentas para limpeza de dados atuais focam em tratar problemas focados em integridade de dados, oferecendo algoritmos de matching para detectar dados duplicados ou fazendo transformações nos dados para um determinado formato, como transformar datas, temperaturas ou velocidade.

Por isso esse trabalho propõe uma ferramenta capaz de auxiliar usuário a aplicar técnicas estatísticas em seu conjunto de dados sem necessitar algum conhecimento extra como uma

linguagem ou framework específico. O usuário também pode visualizar seus dados através gráficos ou informações numéricas e textuais, possibilitando o acompanhamento e análise dos dados enquanto aplica as técnicas. A ferramenta é uma aplicação web, pelo fato de ser a forma mais fácil de se compartilhar e é independente de dispositivo.

Este trabalho é organizado da maneira seguinte. O capítulo 2 contém alguns fundamentos básicos para o auxílio da compreensão deste trabalho e a alguns trabalhos relacionados à limpeza de dados. No capítulo 3, a ferramenta proposta é apresentada, explicando suas funcionalidades e os métodos utilizados. Em seguida, o capítulo 4 contém casos de uso mostrando a usos da ferramenta com um conjunto de dados. Por fim, no capítulo 5, são feitas considerações finais sobre o trabalho e trabalhos futuros.

CAPÍTULO 2

Fundamentos

Neste capítulo, é explicado fundamentos básicos para o trabalho, abordando conceitos como o que são outliers e técnicas quantitativas. Em seguida é feita uma análise nos trabalhos relacionados à limpeza de dados e soluções existentes.

2.1 Outliers

Em um conjunto de dados, um outlier é uma observação se difere bastante de outros valores. Esses valores extremos podem prejudicar análises dos dados e causar efeitos indesejáveis. Por exemplo, considere um conjunto dados sobre uma determinada população de uma região. Esses dados apresentam medidas como idade, peso e altura. Informações como essa podem ser usadas para fazer análises como a taxa de obesidade ou a média de idade. Um outlier poderia ser um valor muito extremo nesse conjunto, como pessoas com altura de 10 metros ou 300 anos de idade. Esses valores são incorretos nesse conjunto e podem alterar valores como a média, tornando uma análise errada. Embora a definição de valor anormal pode ser diferente dependendo do tipo de dado ou do contexto, existem métodos estatísticos para identificar valores que tem grande chance de serem outliers comparados com outros do mesmo conjunto. A seguir damos mais detalhes sobre esses métodos:

2.1.1 Métodos Univariados

2.1.1.1 SD method

O método clássico de detectar outliers é usar o método de desvio padrão. É definido como:

2 SD Method: x ± 2 SD
3 SD Method: x ± 3 SD,

onde x é a média da amostra x e SD é o desvio padrão da amostra. As observações fora desses intervalos podem ser consideradas outliers.[6]

2.1.1.2 MAD_e method

O método MAD_e usa a mediana e a MAD (Median Absolute Deviation) para detectar outliers. É um método básico robusto que não é afetado pela presença de valores extremos no conjunto

de dados. É uma abordagem muito semelhante ao método SD, a principal diferença é o uso da mediana e MAD_e no lugar da média e desvio padrão. É definido como:

2 MAD_e Method: Median ± 2 MAD_e
3 MAD_e Method: Median ± 3 MAD_e,

onde $MAD_e = 1.483 \times MAD$ para dados de tamanho grande com distribuição normal.

Este método não é indevidamente afetado por valores extremos. Mesmo que poucas observações fazem a distribuição dos dados assimétrica, o intervalo é raramente inflado, já que utiliza a mediana, diferente do SD method.[6]

2.1.1.3 Tukey's method

O método de Tukey de 1997, de criar um boxplot, é uma ferramenta bem conhecida para exibir informações de conjunto de dados univariados, como a mediana, primeiro quartil, terceiro quartil e valores extremos. É menos sensível que outros métodos que usam a média e desvio padrão, já que usa quartis, que são resistentes a valores extremos enquanto a média e desvio padrão podem ser afetados.

- O IQR (Inter Quartile Range) é a distância entre o primeiro quartil Q1 e o terceiro quartil Q3.
- As cercas internas são localizadas a uma distância de 1,5 IQR abaixo de Q1 e acima de Q3
- As cercas externas são localizadas a distância de 3 IQR abaixo de Q1 e acima de Q3
- Os valores que estão entre a cerca interna e externa são possíveis outliers. Os valores que estão fora da cerca externa são prováveis outliers e são os valores que são considerados outliers pelo programa.

Como o método de Tukey leva em consideração os quartis para delimitar a região com outliers, tem a vantagem sobre métodos que usam a média e desvio padrão[6]. A Figura 2.1 mostra um exemplo gráfico de um boxplot com outliers nas extremidades.

2.1.2 Métodos Bivariados

2.1.2.1 Bivariate boxplot

Método criado para construir generalizações bivariadas do boxplot. Utiliza o que é chamado de quel, uma generalização de elipses, que é mais apropriada para representar dados assimétricos. Como o boxplot, essa versão bivariada é baseada nos dados e não em sua distribuição e a região mais próxima ao centro contém cerca de 50% dos dados[7]. Na figura 2.2 é possível ver que uma grande quantidade dos pontos estão mais perto do centro.

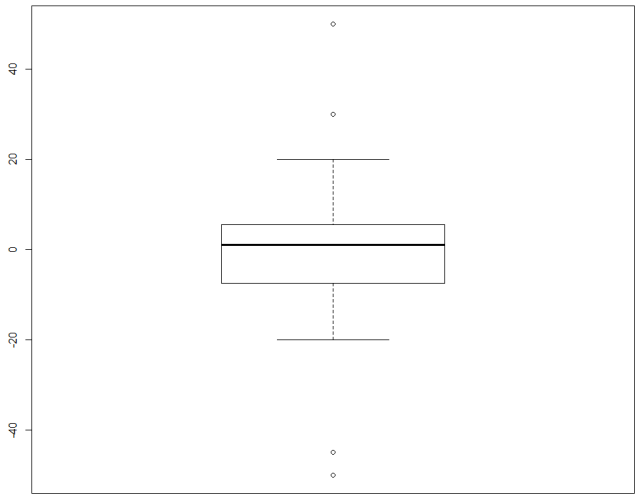


Figura 2.1: Exemplo de um Boxplot

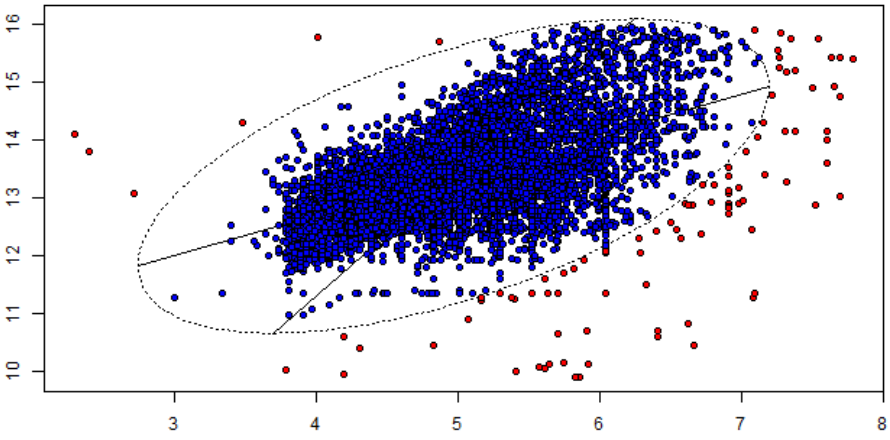


Figura 2.2: Exemplo de um Bivariate Boxplot: Os pontos vermelhos são outliers

2.1.2.2 Bagplot

Bagplot é uma das generalizações bivariadas do boxplot. Os principais componentes são o bag, fence e loop. O bag contém 50% dos pontos, fence separa os inliers dos outliers e o loop contém os pontos fora da bag mas dentro da fence. O gráfico resultante é chamado de bagplot e contém 2 regiões poligonais representando o bag e loop, os pontos do lado de fora são considerados outliers[8]. A figura 2.3 deixa mais claro a separação das regiões.

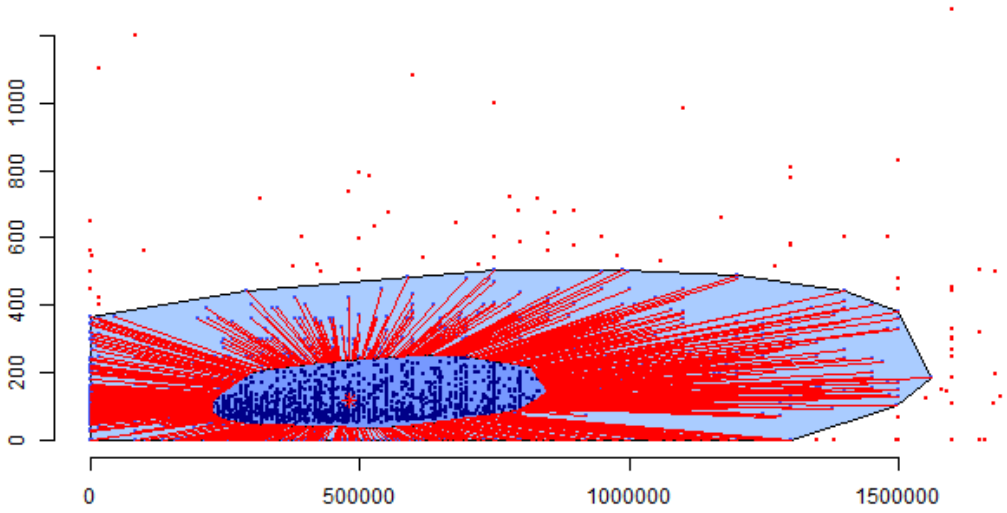


Figura 2.3: Exemplo de um Bagplot: Os pontos vermelhos fora da região azul (loop) são outliers

2.2 Técnicas Quantitativas

Existem diversas técnicas para analisar e tratar de dados com problemas. Elas são divididas em técnicas qualitativas e quantitativas.Técnicas qualitativas são técnicas que estão normalmente ligadas a conjuntos de dados que contêm algum tipo de relação ou dependência. Por exemplo, em um banco de dados de uma empresa existem vários tipos de funcionários que podem se relacionar com diferentes tipos de clientes. Essas técnicas envolvem restrições de integridades, regras e detecção de padrões. No mesmo banco do exemplo anterior podemos ter as seguintes regras: “todo funcionário tem que possuir um código” e “dois funcionários não podem possuir o mesmo código”, e uma violação dessa regra pode ser considerada um dado incorreto. Por outro lado técnicas quantitativas trabalham normalmente com dados que podem ser representados numericamente e utilizam técnicas estatísticas para detectar erros como outliers. [1]

2.3 Trabalhos Relacionados

Limpeza de dados é uma etapa importante para maioria das bases de dados, principalmente de dados extraídos da web. Existem vários trabalhos e pesquisas relacionados a resolver problemas encontrados nesses conjuntos de dados.

Em [3], temos estudos de erros em banco de dados e técnicas quantitativas para correção dos mesmos, mostrando como detectar outliers tanto univariados, quanto multivariados. Além disso há um estudo sobre o design da visualização de dados. A importância de levar em consideração todos os métodos em conjunto ao invés de forma separada é vista em [4]. O estudo mostra que a aplicação de certos métodos podem esconder outros, como tratar valores faltando através

de imputação pode mascarar duplicatas, e que detecção e limpeza são operações que têm mais sucesso em conjunto.

CrowdCleaner [9] tenta criar uma solução inovadora através de crowdsourcing. Ele tenta superar as dificuldades de limpar e manter os dados em bom estado através de um serviço web em que qualquer pessoa pode colaborar, tornando o processo mais fácil.

Atualmente, existem várias soluções comerciais e produtos para limpeza de dados. A maioria das ferramentas contém funções audição e transformação. O primeiro tipo são usadas pelos usuários para detectar problemas e discrepâncias no seu conjunto de dados. O segundo tipo são funcionalidades que são usadas após a análise dos dados para fazer transformações e transformá-los para um formato adequado. Alguns produtos e ferramentas encontrados no mercado são:

- Winpure Clean & Match[1] é um software para windows desktop para corrigir, padronizar e remover duplicatas de conjunto de dados. Pode importar dados de diversos tipos, como listas, planilhas ou bancos de dados. Os dados podem ser visualizados em tabelas e filtrados. O principal foco do programa é utilizar algoritmos e técnicas fuzzy para identificar dados duplicados e incorretos.

- DataCleaner[2] é uma ferramenta de *profiling*, análise dinâmica, de dados. Ela contém detecção de padrões, campos faltando, valores e outras características.

- Drake[3] é uma ferramenta baseada em texto que ajuda a executar ações sobre dados e suas dependências. É necessário que seja definido quais operações, inputs e outputs e Drake trata de agilizar e resolver dependências.

Podemos analisar que a maioria das ferramentas atuais se dedicam a corrigir e monitorar bancos de dados e realizar transformações e detectar dados duplicados. Diferente das ferramentas acima, o foco deste trabalho é na remoção de erros e outliers utilizando técnicas estatísticas.

Podemos analisar que a maioria das ferramentas atuais se dedicam a corrigir e monitorar bancos de dados e realizar transformações e detectar dados duplicados. Diferente das ferramentas acima, o foco deste trabalho é na remoção de erros e outliers utilizando técnicas estatísticas.

[1] http://www.winpure.com/cleanmatch2.html
[2] https://datacleaner.org/
[3] https://github.com/Factual/drake

CAPÍTULO 3

Ferramenta

O principal objetivo da ferramenta é, dado um conjunto de dados, ajudar o usuário a limpar e corrigi-los de forma simples. O programa possui as seguintes funções:

- Aceitar um conjunto de dados em um formato comum e estruturado.
- Fornecer formas de visualizar e analisar o estado atual dos dados
- Ajudar o usuário a aplicar técnicas e métodos estatísticos para melhoria dos dados.

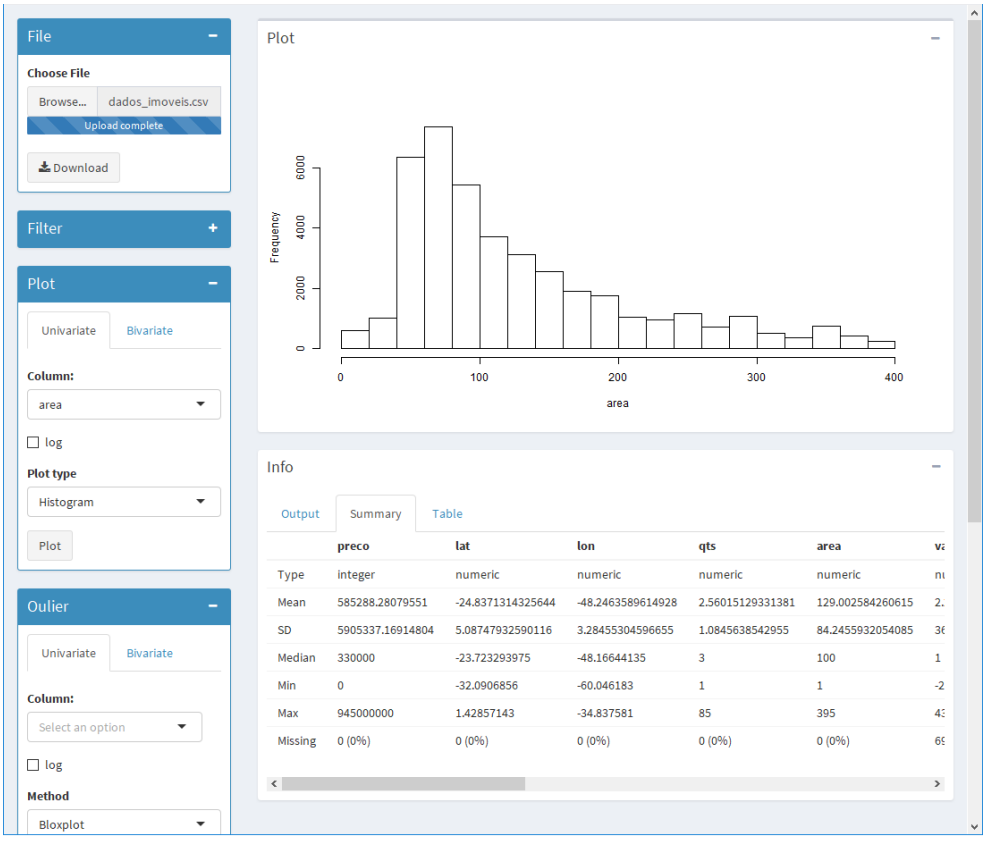


Figura 3.1: Interface da aplicação

3.1 Implementação

Para a implementação da ferramenta foi escolhida a linguagem R, que é uma linguagem para computação estatística e visualização de gráficos. Além disso, possui o benefício de conter uma grande quantidade de pacotes com métodos estatísticos. As funções e a lógica da computação que são usadas na aplicação foram todas implementadas em R. Por isso, alguns termos da linguagem são usados‘no programa. Os tipos dos atributos são definidos como *integer* para valores inteiros, *numeric* para decimais e *factor* para valores discretos. Além desses, temos NA (Not Avaliable) para valores ausentes.

Como é uma aplicação Web, a ferramenta é dividida em servidor e interface. O servidor fica responsável por carregar os dados do usuário e realizar as computações. Todas as operações e estados são realizadas nele, enquanto a interface tem apenas a função de interação com o usuário e solicitar as operações ao servidor. Para implementação foi utilizado o pacote shiny[1]. Com ele foi possível implementar o servidor chamando funções diretamente em R e a interface através de widgets, html, css e javascript fornecidos pelo shiny.

3.2 Entrada e Saída

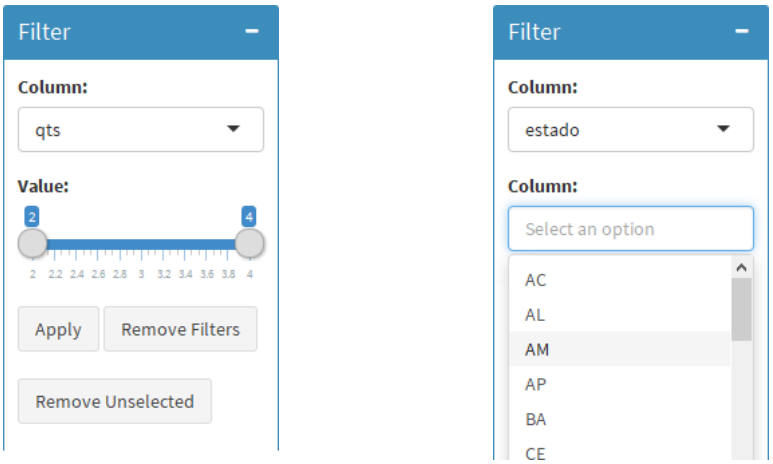
O programa aceita arquivos no formato csv (comma-separeted values) como entrada. A primeira linha determina o título das colunas e são usados para as operações em que o usuário tem que escolher. Depois de fazer todas as operações que quiser, o usuário pode então salvar os seus dados limpos no mesmo formato. Essas opções podem ser vistas na Figura 3.1 na caixa File.

3.3 Filtros

Dependendo do conjunto de dados, pode ser necessário apenas selecionar um subconjunto para trabalhar, seja porque alguns dados não sejam considerados relevantes pelo usuário ou porque eles precisam de um contexto separado. Como exemplo nos dados de imóveis, preço dos imóveis normalmente variam de uma cidade ou estado para outro e pode fazer com alguns valores que pareçam ser extremos ou errados, estejam corretos em um subconjunto. Para isso, a ferramenta disponibiliza de um mecanismo de filtro, onde o usuário pode selecionar uma coluna e um valor ou intervalo para selecionar apenas esse subconjunto. (Imagem do menu de filtros na Figura 3.2)

Os dados provenientes da web contêm vários valores ausentes ou entradas duplicadas. Então é fornecido a opção de filtrar essas entradas. As operações são divididas em operações de linha e operações de coluna. As operações de linha são remover NAs (Valores Ausentes), que remove qualquer linha que tenha valor NA e remover duplicatas, que remove duas linha idênticas. A operação com coluna disponível é remover coluna.

[1] https://shiny.rstudio.com/



(a) Filtro para atributo contínuo (b) Filtro para atributo discreto

Figura 3.2: Opções de filtros para atributos discretos e contínuos

3.4 Visualização

O principal método de visualização é através de gráficos. Um exemplo de histograma gerado pela ferrameta é apresentado na Figura 3.3. O usuário pode escolher entre métodos de plot univariados e bivariados.

Métodos Univariados:

- Histogram
- Boxplot
- Scatterplot (Usando índice e valor)

Métodos Bivariados:

- Bivariate Boxplot
- Bagplot
- Scatterplot

3.5 Info

A sua função é fornecer informação extra sobre os dados que não são obtidas através dos plots. É separado em três abas: Summary, Output e Table.

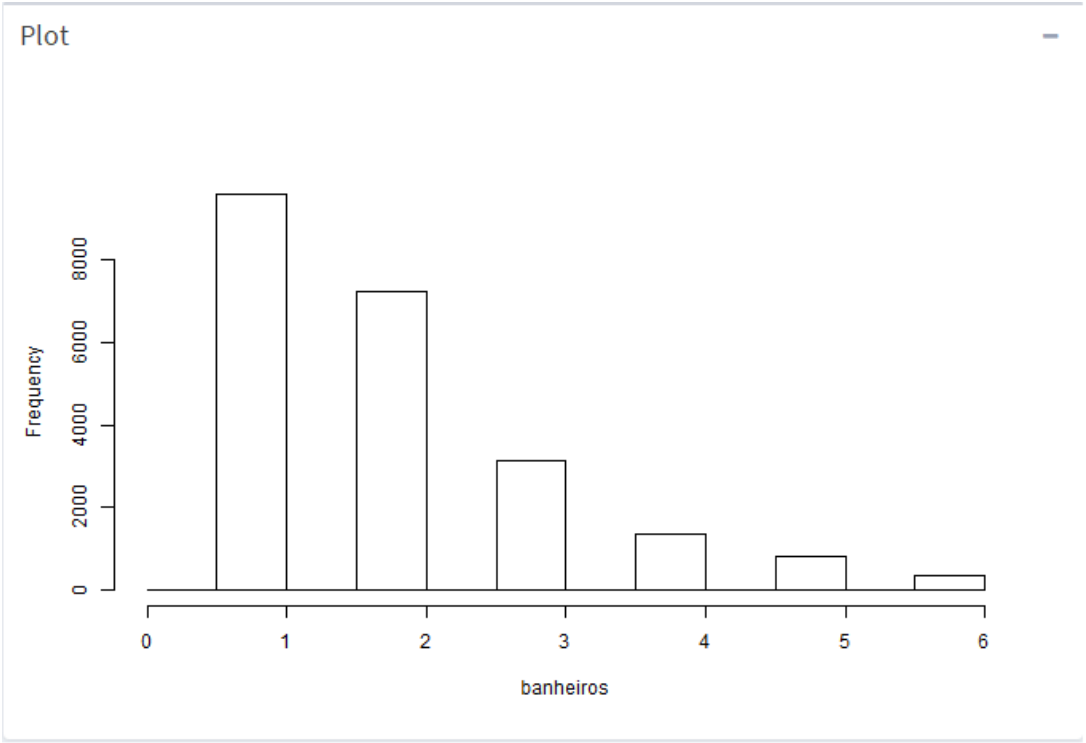


Figura 3.3: Exemplo de um histograma plotado pela ferramenta

3.5.1 Summary

O summary exibe informações sobre as colunas do conjunto de dados. Informações como número de valores faltando, média, mediana, são importantes para tomar decisões e analisar o estados dos dados. Por exemplo, para uma coluna que contém um número de valores ausentes muito grande, pode ser decidido se é melhor removê-la completamente ao invés de remover as linhas que contém NAs, que faria com muitos dados fossem perdidos.

3.5.2 Output

O output fornece informações sobre as operações feitas pelo usuário, como linhas ou colunas deletadas, informações sobre filtros aplicados, quantidade de outliers detectados e removidos (Figura 3.4). Serve como um log de ações e ao mesmo tempo um feedback para o usuário.

3.5.3 Table

Uma outra forma de o usuário visualizar seus dados. Na tabela é possível reordenar os dados de acordo com alguma coluna ou pesquisar por resultados. É uma forma mais fácil de visualizar linhas separadas e comparar todas as variáveis ao mesmo tempo.

Figura 3.4: Aba Output: Possui informações sobre o resultado das ações feitas pelo usuário

3.6 Detecção e Remoção de Outliers

A ferramenta possui uma seção para detecção e remoção de outliers. Assim como a visualização o menu é dividido em univariado e bivariado e o usuário pode escolher as variáveis usadas.

A método mais adequado pode mudar dependendo do conjunto de dados e sua distribuição, por isso, além da opção de remover outliers, o programa disponibiliza uma função de apenas detectá-los. Assim é possível ver a quantidade de outliers encontrados, intervalo dos valores válidos e outras informações dependentes do método.

Alguns métodos de detecção de outliers também possuem um gráfico para visualização como o boxplot e o bagplot e a visualização é importante ao se tratar de outliers. Então quando usada a opção de detectar, um gráfico apropriado é mostrado ao usuário com os dados atuais e quando removidos, um gráfico com os dados após a operação.

Como diferentes tipos de dados possuem diferentes distribuições, a ferramenta também disponibiliza a opção de transformar o dados usados nas funções de plot e de outliers com a função log. Uma das vantagens dessa função é que ela afeta a distribuição, tornando mais fácil trabalhar com dados muito assimétricos.

Na Figura 3.5 podemos ver a lista de métodos que podem ser escolhidos e abaixo aopção de aplicar o log a uma coluna. Os métodos implementados na ferramente são para uma variável (univariado) e para duas variáveis (bivariados) e são listados a seguir:

Métodos Univariados:

- Boxplot
- SD Method (2SD)
- MAD_e Method (2MAD_e)

Métodos Bivariados:

- Bivariate Boxplot
- Bagplot

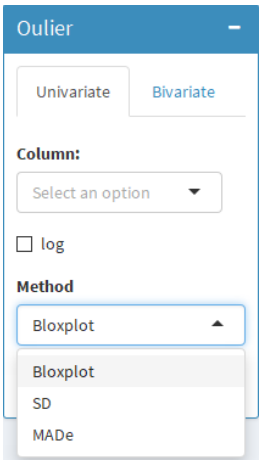


Figura 3.5: Caixa de controle para operações envolvendo outliers

CAPÍTULO 4

Casos de Uso

4.1 Conjunto de dados

Os dados utilizados para testar uma ferramenta contém informações sobre venda de imóveis no Brasil. O conjunto é gerado a partir de páginas web por um extrator e contém problema de dados duplicados, faltando e outliers. Por isso é ideal para testar as funcionalidades da aplicação. Ele contém as informações sobre preço do imóvel, latitude, longitude, número de quartos, área, vagas, subúrbio, distrito, cidade, estado e site.

4.2 Análise Geral

O primeiro passo é carregar os seus dados a serem trabalhados. Os dados de imóveis estão no formato csv e podem ser lidos normalmente. Após carregar os nossos dados, podemos obter informações sobre cada coluna na aba Summary. Na Figura 4.1 temos uma parte do sumário, e partir dela já é possível ver alguns erros:

Info

Output | Summary | Table

qts	area	vagas	banheiros	suites
numeric	numeric	numeric	integer	integer
4.61938672922252	500.11516565982	2.34105369807497	2.03660776913057	1.60285414016262
357.742934654465	3528.37997178314	31.328537282751	1.51055561673899	1.08470985061765
2	110	1	2	1
-4	0	-2	0	-4
61986	580000	4300	78	11
7103 (10.64%)	14640 (21.92%)	17433 (26.1%)	36844 (55.17%)	48704 (72.93%)

Figura 4.1: Sumário gerado do conjunto de dados de imóveis

- As colunas banheiros e suítes contém um número muito baixo de dados, com mais de 50% valores em branco para os banheiros e 70% para suítes.
- O valor mínimo indica que alguns dados possuem valores menores que zero, e o máximo contém valores muito extremos como número de quartos com valor 61986. O efeito pode ser notado no valor do desvio padrão que é de 357.74 ou na diferença da média e a mediana na área. Como os dados representam informações do mundo real, esses valores podem ser considerados erros.

4.3 Filtrar os Dados

Tendo encontrado esses erros iniciais, podemos diminuí-los filtrando as variáveis. Primeiro, vamos remover as colunas que têm poucos valores. Normalmente, é melhor manter uma coluna e apenas remover as linhas, mas no caso da ausência de mais de 50%, decidimos removê-las. Em seguida filtramos as colunas numéricas, mantendo apenas valores maiores que 0. Filtrando com um intervalo remove, além dos valores fora do intervalo, valores inexistentes.

Em seguida, vamos visualizar gráficos para obter informações como a distribuição das variáveis. Porém, como os dados são de imóveis de todo o país, não faz sentido comparar todos os dados juntos. O preço dos imóveis podem variar muito dependendo da localização. Por este motivo, utilizaremos apenas um subconjunto que consiste de dados apenas do estado de São Paulo para visualizar e aplicar os métodos de detecção de outliers.

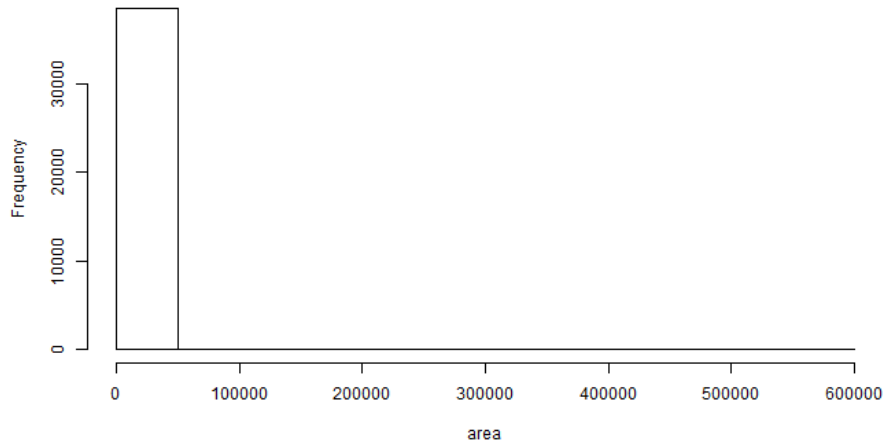
4.4 Visualizar os Dados

Com o subconjunto das cidades de São Paulo selecionado através da função de filtros, vamos analisar alguns atributos através de gráficos. Com o histograma podemos ter uma ideia da distribuição dos dados. Como mostrado na Figura 4.2a, os dados de área são bastante assimétricos, contendo praticamente todos os valores no lado esquerdo, problema que é causado por poucos valores extremos do lado direito. Podemos aplicar a função log para mudarmos distribuição e pode ser visualizado na Figura 4.2b que o resultado é mais simétrico e próximo a uma distribuição normal.

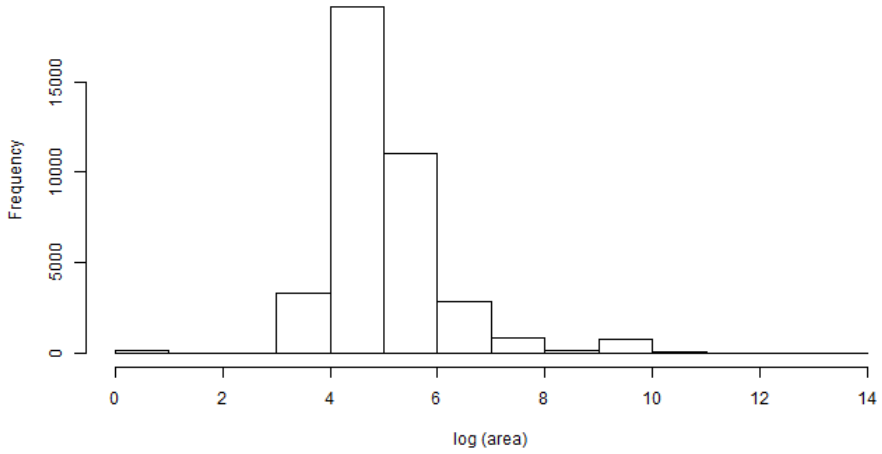
4.5 Remover Outliers

O próximo passo é detectar outliers nos dados. Utilizamos a ferramenta para selecionar e aplica métodos de detecção de outliers. Começamos com outliers univariados utilizando a mesma variável area, já que vimos nos gráficos anteriores.

O resultado da operação pode ser visto na tabela 4.1. Intervalo se refere ao intervalo dos valores que não são outliers e variável é a valor usado para calcular o intervalo, onde cada cada método usa uma diferente (IQR, SD, MAD_e). Pelos resultados podemos ver que o método



(a) Sem aplicar log



(b) Aplicando log

Figura 4.2: Histograma do atributo area

do desvio padrão (SD) é muito afetado pelos valores extremos e acaba aceitando valores em um intervalo entre -10363.02 e 12522.48. Entre o boxplot e o MAD_e, o segundo possui um intervalo menor e consequentemente considera mais valores como outliers

Método	Nº de outliers	Intervalo	Variável
Boxplot	2249	[-262.5, 637.5]	IQR: 225
SD Method	666	[-10363.02, 12522.48]	SD: 5721.37
MAD_e Method	3538	[-91.28, 371.28]	MAD_e: 115.64

Tabela 4.1: Tabela com o resultado da detecção de outliers da area

Em seguida, testamos os mesmos métodos com o log da área (Figura 4.2). A diferença mais notável é no método do desvio padrão. O log diminui consideravelmente os valores extremos do

lado direito, transformando a distribuição em uma mais parecida com uma distribuição normal. Por isso, o intervalo e a quantidade de outliers encontrados é semelhante ao método do boxplot. O método MAD_e continua com o menor intervalo.

Método	Nº de outliers	Intervalo	Variável
Boxplot	1083	[2.23, 7.78]	IQR: 1.38
SD Method	1053	[2.42, 8.03]	SD: 1.40
MAD_e Method	1750	[2.92, 6.9]	MAD_e: 1.00

Tabela 4.2: Tabela com o resultado da detecção de outliers do log da area

Removemos então os outliers com o método boxplot e aplicando o log. A Figura 4.3 compara o sumário da área antes de depois da remoção. Notamos a diferença nos valores como desvio padrão e média e a diminuição do intervalo de minímo e máximo. E a Figura 4.4 contém os novos histogramas mostrando que a distribuição está menos assimétrica.

area	area
numeric	numeric
1079.73128764108	244.700367734568
5721.37828154227	331.207675401718
140	130
1	9.926
580000	2400
0 (0%)	0 (0%)

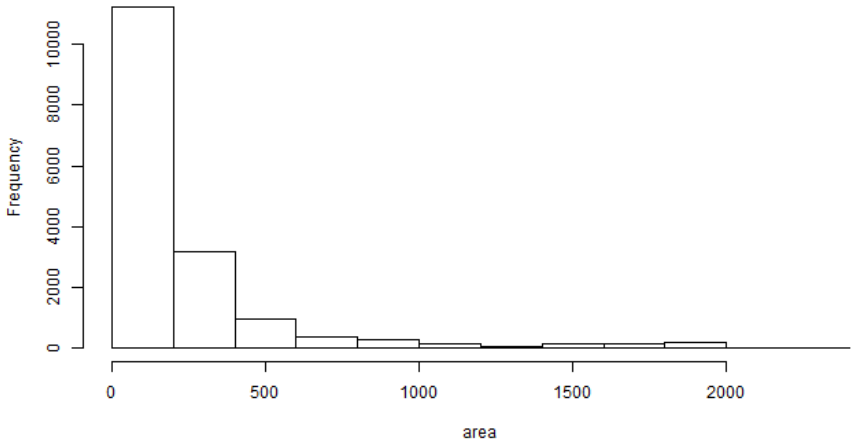
(a) Antes (b) Depois

Figura 4.3: Comparação entre o sumário do atributo área antes e depois da remoção de outliers: As linhas indicam tipo, média, desvio padrão, mediana, mínimo, máximo e valores faltando

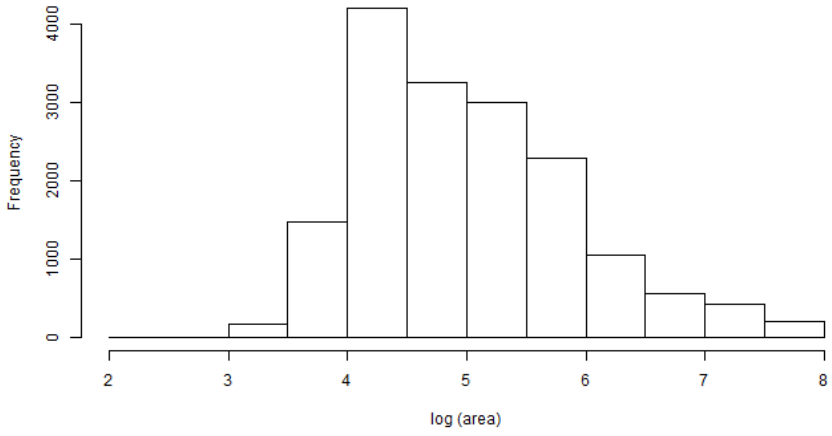
Para o caso bivariado, escolhemos as variáveis área e quartos. Essas váriáveis são normalmente diretamente proporcionais em imóveis, então se existir um caso de área muito grande com poucos quartos, é provavelmente um outlier. O método bagblot encontrou 37 outliers enquanto o bivariate boxplot encontrou 737. Essa diferença pode ser vista nos gráficos (Figura **??**). Vamos usar o bivariate boxplot para remover os outliers.

4.6 Baixar os Dados Limpos

Após todas as operações podemos baixar o arquivo no mesmo formato que foi carregado e ele está pronto para ser utilizado. Por exemplo para fazer previsões e análise de preços sem os



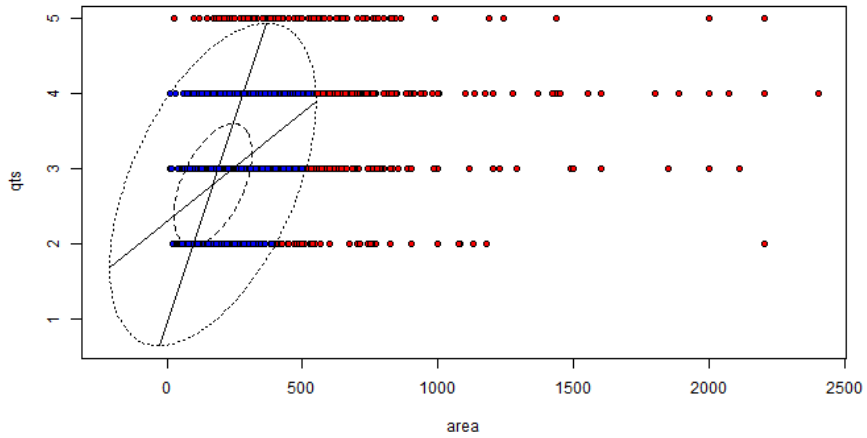
(a) Sem aplicar log



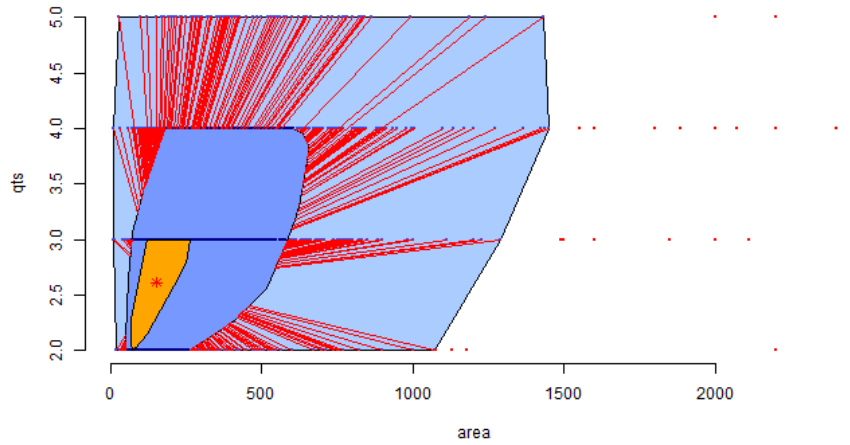
(b) Aplicando log

Figura 4.4: Histograma do atributo area após a remoção de outliers

dados errados que existiam no começo.



(a) Bivariate Boxplot



(b) Boxplot

Figura 4.5: Detecção de outliers bivariados

CAPÍTULO 5

Conclusão

Este trabalho propôs uma ferramenta para limpeza de dados. Foram utilizadas tecnologias web para criar uma aplicação independente de plataforma e que pode ser usada em qualquer dispositivo. O sistema apresenta uma forma interativa de tratar os dados, onde o usuário pode escolher quais métodos são utilizados e pode visualizar vários tipos de informação e gráficos ao mesmo tempo. As técnicas estatísticas fornecidas são técnicas de detecção de outliers para uma ou duas variáveis, como o boxplot e o bagplot. Embora o trabalho foi focado em dados da web, a ferramenta funciona para outros tipos de dados que tenham os mesmos tipos de problema.

Uma proposta de trabalho futuro seria expandir as funcionalidades da ferramenta como adicionar técnicas de imputação para substituir valores não existentes. O desenvolvimento foi focado em resolver o problema de outliers nos dados que vieram da web, que é apenas um dos problemas que ocorrem nos dados. Novos métodos podem ser adicionados para resolver outros tipos de problemas.

A construção da interface não teve nenhum teste ou método de design iterativo com o usuário. A melhoria da interface e testes de usabilidade é outro trabalho futuro que pode ser feito.

Referências Bibliográficas

[1] X. Chu, I. F. Ilyas, S. Krishnan, and J. Wang, "Data cleaning: Overview and emerging challenges," *Proceedings of the 2016 International Conference on Management of Data - SIGMOD '16*, 2016.

[2] M. v. d. L. Edwin de Jonge, "An introduction to data cleaning with r," *Proceedings of the 2016 International Conference on Management of Data - SIGMOD '16*, 2013.

[3] J. M. Hellerstein, "Quantitative data cleaning for large databases," *United Nations Economic Commission for Europe (UNECE)*, February 2008.

[4] D. S. Laure Berti-Équille, Tamraparni Dasu, "Discovery of complex glitch patterns: A novel approach to quantitative data cleaning," *Data Engineering (ICDE), 2011 IEEE 27th International Conference on Data Engineering*, May 2011.

[5] T. Dasu, "Data glitches: Monsters in your data," *Handbook of Data Quality*, p. 163–178, 2013.

[6] S. Seo, "A review and comparison of methods for detecting outliers in univariate data sets," *Master of Science thesis. University of Pittsburg*, 2006.

[7] K. M. Goldberg and B. Iglewicz, "Bivariate extensions of the boxplot," *Technometrics*, vol. 34, p. 1182–1185, August 1992.

[8] P. J. Rousseeuw, I. Ruts, and J. W. Tukey, "The bagplot: A bivariate boxplot," *The American Statistician*, vol. 53, p. 1182–1185, November 1999.

[9] Y. Tong, C. C. Cao, C. J. Zhang, Y. Li, and L. Chen, "Crowdcleaner: Data cleaning for multi-version data on the web via crowdsourcing," *ICDE*, p. 1182–1185, 2014.