CENTRO UNIVERSITÁRIO DO ESTADO DO PARÁ
ARGO ESCOLA DE NEGÓCIO E TECNOLOGIA
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

Caio Abdon da Silva

ANÁLISE DE DADOS DO MERCADO IMOBILIARIO UTILIZANDO *PYTHON* PARA ENCONTRAR ESTATÍSTICAS DESCRITIVAS:

uma proposta para contornar o *Big Data*

Belém
2019

Caio Abdon da Silva

ANÁLISE DE DADOS DO MERCADO IMOBILIARIO UTILIZANDO *PYTHON* PARA ENCONTRAR ESTATÍSTICAS DESCRITIVAS:

uma proposta para contornar o *Big Data*

Trabalho de Curso na modalidade monografia, apresentado como requisito parcial para a obtenção do grau de bacharelado em Ciência da Computação na Escola de Negócio Tecnologia e Inovação – ARGO, sob orientação do Professor MSc. Carlos Benedito Pereira da Paixão.

Belém

2019

Dados Internacionais de Catalogação-na-publicação (CIP)
Biblioteca do Cesupa, Belém – PA

Silva, Caio Abdon da.

Análise de dados do mercado imobiliário utilizando python para encontrar estatísticas descritivas: uma proposta para contornar o Big Data / Caio Abdon da Silva; orientador Carlos Benedito Pereira da Paixão. – 2019.

Trabalho de Conclusão de Curso (Graduação) – Centro Universitário do Estado do Pará, Ciência da Computação, Belém, 2019.

Tecnologia da informação. 2. Exploração de dados. 3. Gestão da informação. 4. Comunicação (Inovações tecnológicas). I. Paixão, Carlos Benedito Pereira da, *orient*. II. Título.

CDD 23ª ed. 004

Caio Abdon da Silva

ANÁLISE DE DADOS DO MERCADO IMOBILIARIO UTILIZANDO *PYTHON* PARA ENCONTRAR ESTATÍSTICAS DESCRITIVAS:

uma proposta para contornar o *Big Data*

Trabalho de Curso na modalidade monografia, apresentado como requisito parcial para a obtenção do grau de bacharelado em Ciência da Computação na Escola de Negócio Tecnologia e Inovação – ARGO, sob orientação do Professor MSc. Carlos Benedito Pereira da Paixão.

Aprovado em: ____ de _____ de 2019.

Banca Examinadora:

_____

Prof. Carlos Benedito Paixão – CESUPA (orientador)

_____

Profa. Polyana Santos Fonseca Nascimento – CESUPA

_____

Profa. Alessandra Natasha Alcantara Barreiros Baganha – CESUPA

Dedico este trabalho a plataforma ALURA que foi o impulsionamento para me ajudar a compreender esta área que resolvi desenvolver este trabalho, com sua comunidade e seus cursos pude ter o primeiro contato e me aprimorar nesta área, tirando dúvidas e sempre aprimorando seus cursos para me tornar um profissional mais capacitado.

AGRADECIMENTOS

Agradeço a todas as pessoas que me acompanharam ao longo desta jornada acadêmica.

À minha família.

Agradeço aos meus amigos.

Agradeço meus colegas de classe.

Agradeço minha namorada Carol que sempre me ajudou em momentos de dificuldade.

Agradeço a nova coordenação do CESUPA que trouxe nova vida nova ao curso.

Agradeço ao meu orientador tanto na parte estatística como em me ajudar a entender o mundo pós faculdade e a não ter tanto medo do mesmo.

Todos juntos foram essenciais.

RESUMO

Com a diversificação da internet nos últimos anos, houve um aumento na quantidade de dados produzidos e logo nasceu a dificuldade de analisar os mesmos. Com isso surgiu um novo fenômeno no mundo da computação que foi o *Big data*. O objetivo deste trabalho é trazer uma alternativa de contornar esta situação por meio da programação, será abordado como este problema dificulta a análise de dados e será mostrada uma solução para o mesmo. As metodologias utilizadas para realizar a análise dos dados foram *python* e suas bibliotecas, juntamente com estatísticas descritivas. Em seguida será falado sobre a comparação entre as ferramentas de análise contra a programação. E ao final deste trabalho, será mostrado como utilizando esse método se conseguiu melhorar a análise de dados.

Palavras-Chave: Dados. Análise. *Big Data*. Estatística.

ABSTRACT

With the diversification of the Internet in recent years, there has been an increase in the amount of data produced and soon the difficulty arose to analyze them. With this came a new phenomenon in the world of computing that was the Big Data. The objective of this work is to bring an alternative to circumvent this situation through programming, will be addressed as this problem hampers data analysis and will be shown a solution to it. The methodologies used to perform the data analysis were python and its libraries, along with descriptive statistics. Next we will talk about the comparison between the analysis tools against programming. And at the end of this work, it will be shown how to use this method if it was able to improve the data analysis.

Keywords: Data. Analysis. Big Data. Statistics

SUMÁRIO

1 INTRODUÇÃO

O aumento do fluxo de dados devido a diversificação das tecnologias de informação nos últimos anos trouxe um problema para donos de estabelecimentos que registram seus dados utilizando computadores: O *Big Data*. Por isso, surge a necessidade de explicar de que forma seria possível processar referida quantidade de dados de maneira rápida e eficiente.

> O mundo gera, diariamente, 2,5 quintilhões de bytes (sendo 1 quintilhão igual a 10 elevado à 18ª potência). As mais diversas ações diárias da sociedade (de manifestações de usuários nas redes sociais a registros corporativos e movimentações financeiras) tornaram-se dados valiosos paras as empresas, que podem utilizá-los para conhecerem melhor seus clientes (EKIMA, 2018).

O *Big Data* é um problema que surgiu devido à grande diversificação da internet, fazendo com que dados fossem gerados de qualquer plataforma, de qualquer lugar e em volumes imensos, desta forma dificultando a análise e o entendimento dos dados gerados (INTEL, 2013).

O impacto que o *Big Data* trouxe para o mercado imobiliário foi grande, com o popularização da internet e aplicativos que coletam informações dos usuários, o processamento dos mesmos se tornou difícil, como diz Dohan (2018): “A Tegra Incorporadora, antiga *Brookfield*©, identificou que mais de 95% dos seus clientes iniciam o processo de busca por um imóvel através de canais digitais”.

Portanto, a partir da numerosa quantidade de informações a serem processadas no mundo atual, surge uma necessidade, quase que substancial, de mecanismos que sejam capazes de processar com eficiência.

O entendimento das informações geradas por uma empresa causa grande impacto na tomada de decisões das pessoas que a administram. Por isso, sua análise é de suma importância para o crescimento empresarial e sucesso da mesma (EKIMA, 2018). Assim, em um cenário dinâmico como o atual, ter acesso as mudanças do mercado antes do rival é a chave para a sobrevivência no mundo corporativo.

Ocorre que, chegamos a um ponto em que as ferramentas de análise estatística presentes no mercado não estão conseguindo suprir a demanda dos donos de estabelecimentos que diariamente produzem uma quantidade absurda de

informações, devido a suas limitações quanto a quantidade de dados que podem ser processados, uma vez que os mesmos são analisados de forma manual, exigindo tempo e esforço humano desnecessário.

Nesse sentido, existem técnicas e ferramentas no mundo da computação que se apresentam como formas de analisar os dados automaticamente, como exemplo, a *structed query language* (sql). Porém, quando falamos em analisar dados em tempo real, essa solução já não fica tão atraente, como diz Melo (2018): “criar um banco, criar as tabelas, carregar os dados, e só então explorar os dados leva tempo… é melhor fazer esse tipo de análise com pandas”.

Primeiramente, será abordado o conceito de *Big Data*, analisando a problemática envolvida na sua existência, para posteriormente, demonstrar as necessidades nascidas deste entrave. Assim, após demonstrar a ineficiência dos instrumentos de análise existentes, será apresentado como a programação pode soluciona tal problema. De igual forma, será exposta a montagem de novos ambientes e a preparação destes com bibliotecas (PANDAS, SEABORN, MATPLOTLIB), bem como, apresentar *prints* de tela do software e de seus resultados, concluindo com a indicação de como aprimorar esse trabalho no futuro.

1.1 PROBLEMÁTICA

1.1.1 Ineficiência das ferramentas atuais

O motivo que levou ao desenvolvimento deste trabalho foi a ineficácia dos até então conhecidos softwares de análise, como Excel©, devido ao seu limites nas quantidades de dados que suportam armazenar, pois se tornaram pequenos, frente a quantidade de informações que são geradas diariamente e também demonstram demora na análise dos dados, já que a mesma é feita de forma manual e como explica Patil (2018): “*In Excel, once you exceed 10,000 rows, it starts to slow down — considerably*”.

No que tange as ferramentas de programação, observa-se que não sofrem dessas limitações, pois a quantidade de dados não se restringe como no excel©. Pois, com a utilização da programação é permitido obter um poder computacional maior, sem considerar que o código pode ser compartilhado e reutilizado, fazendo assim, com que análises futuras sejam facilitadas.

As ferramentas já utilizadas, como o *sql*, se tornam ineficazes em certas situações, principalmente, quando se trata de uma análise em tempo real. Assim, essas ferramentas, que são usadas na maioria das vezes pelas empresas, normalmente, não são a melhor opção devido ao seu funcionamento, ainda mais quando aliamos a uma possível predição e análise em tempo real (GALDINO; NATANAEL, 2019).

1.1.2 Big Data

O grande aumento na popularidade dos *smartphones*, em junção com a portabilidade da internet, fez com que a quantidade de dados gerados fosse grande o suficiente para dificultar a análise dos mesmos (GALDINO; NATANAEL, 2019). Com isso mostra-se a necessidade de novas formas de analisar os dados bem como profissionais capacitados no ramo.

Portanto, com o *Big Data*, surge um problema recente devido a essa alta conectividade, os servidores de todas as áreas recebem dados de todos os dispositivos conectados à internet e com isso hoje já são produzidos *zetabytes* (10^{21}) *bytes*) de informações.

Outro grande fator que faz com que a produção de dados aumente exponencialmente, são as nuvens. Nuvens nada mais são que um serviço oferecido por uma empresa onde se pode armazenar e enviar dados de qualquer lugar do mundo com acesso à internet, estes serviços armazenam uma quantidade enorme de informação o que aumentou mais o problema como dizem Fagundes, Macedo e Freund (2017): “O acesso e o uso destas tecnologias fizeram com que a quantidade de dados aumentasse de uma forma contínua e a uma velocidade sem precedentes”.

Assim, soma-se o fato de que para uma empresa é indubitável que a análise dos dados obtidos no decorrer do expediente, sejam suas vendas, a produtividade dos seus funcionários, as baixas no seu estoque, dentre outros elementos, como diz Neoway (2019): a não adaptação da empresa a este novo paradigma pode resultar na perda da inteligência de negócios.

1.1.3 Crescimento dos Dados no Mercado Imobiliário

Antigamente imóveis eram vendidos através de corretores de imóveis, anúncios em jornais, feirões e etc... Estes eram meios eficientes de divulgar um local que estava a venda e atrair possíveis compradores. Porém, segundo Dohan (2018) o mercado mudou, e empresas ainda presas neste modelo tendem a ficar para trás

Como no mundo atual altamente conectado, as pessoas utilizam de aplicativos, internet ou outras formas digitais como forma de procurar por anúncios imobiliários, e devido a esta massiva quantidade de dados gerados ocorreu de precisar de uma nova forma de processa-los (DINO, 2018).

O motivo disso é que muitas empresas deste mercado viraram *Data Driven,* o que significa que todas as suas ações são tomadas com base nos dados coletados, “Romeo Busarrelo, diretor de marketing da Tecnisa, disse que hoje, 100% do que fazem passa por dados” (DOHAN, 2018).

1.2 JUSTIFICATIVA

Ao observar as pessoas que cuidam da gestão de uma empresa sempre com urgência para obterem uma rápida e eficiente forma de estudo sobre os dados obtidos em determinado período e sobre certo assunto, foi o motivo que levou o desenvolvimento deste trabalho, com o intuito de resolver o problema que as mesmas enfrentam.

Uma forma de contornar esse problema seria usando programação para a coleta e processamento de dados, por meio da qual uma máquina separada apenas acessaria o banco de dados de produção para obtenção de dados e processaria os mesmos em memória sem atrapalhar as demais funcionalidades do banco.

Paralelamente a pesquisa sobre o que poderia ser feito quanto ao processamento destes dados, foi descoberta a área de *Data Science*, que é uma área da computação totalmente dedicada ao trabalho com dados. Portanto, observando referida necessidade, foi desenvolvido este *software* para ajuda-las a obter as estatísticas desejadas o mais rápido o possível. Corroborando com tal entendimento, dispõe Patenate (2018):

> Confiabilidade e agilidade são atributos que agregam assertividade às decisões estratégicas. Esses quesitos só são alcançados quando um negócio sabe como fazer análise de dados e consegue usufruir das respostas obtidas para compreender e explorar bem todos os cenários.

1.3 OBJETIVOS

1.3.1 Objetivos Gerais

Analisar dados do mercado imobiliário com a linguagem *python*.

1.3.2 Objetivos Específicos

- Examinar os dados, entregando as estatísticas descritiva;
- Utilizar a linguagem *python* e suas bibliotecas para realizar a análise e mostrar os dados obtidos em forma de gráfico;
- Comparar os dados obtidos por programação com ferramentas já existentes.

1.4 METODOLOGIA/PROCEDIMENTOS METODOLOGICOS

Após realizar uma pesquisa, encontrei várias fontes que falavam sobre o problema da quantidade massiva de dados sendo produzida nos dias atuais e a dificuldade de processá-los. Com essa informação em mãos deparei-me com uma forma nova de processamento de dados que soluciona o problema.

Além disso, foram utilizadas as estatísticas descritivas para poder explicar o que os dados analisados representam em relação a atual situação do negócio, depois foi utilizada a representação destes dados na forma de gráficos, pois apresenta-se como um modo de representar os resultados obtidos de uma forma mais agradável e de fácil compreensão.

Para o desenvolvimento do *software* foi utilizada a linguagem *python* e suas bibliotecas (*pandas, seaborn, matplotlib*). Aliado a estas, fez-se uso das equações estatísticas (moda, media, mediana, desvio padrão, variância, curtose, assimetria, erro padrão, quartis e decis) para se retirar *insights* dos dados.

1.5 ESTRUTURA DO TRABALHO

Após a introdução, o restante do trabalho foi dividido em 5 partes, onde foram relatados todos os métodos estatísticos aplicados, assim como, a plotagem dos gráficos que melhor se adaptam a situação, além da apresentação dos motivos que levaram ao desenvolvimento deste software, bem como, suas vantagens.

No primeiro capítulo tratou-se da introdução do trabalho, no qual abordou-se os problemas que geraram a quantidade massiva de dados produzida diariamente, e também, as dificuldades que as ferramentas atuais de análise de dados enfrentam para processá-los.

No segundo capítulo foi abordado o referencial teórico, estuda-se de forma mais aprofundada os métodos estatísticos que são utilizados, assim como suas desvantagens e vantagens.

No terceiro capítulo foi exposto como instalar o software e fazê-lo funcionar, fazendo o download da plataforma, instalando pacotes adicionais, removendo pacotes não desejados e criando novos ambiente para criar um ambiente apenas com os pacotes e ferramentas que forem necessários para não modificar o diretório base.

No quarto capítulo dissertou-se sobre as considerações finais quanto ao desenvolvimento deste trabalho, assim como, apresentou-se uma perspectiva de aprimorá-lo no futuro, para que fique ainda mais completo.

2 FUNDAMENTAÇÃO TEÓRICA/REVISÃO BIBLIOGRÁFICA

A estatística é uma ciência utilizada desde a antiguidade, já que no Antigo Egito (entre 3200 a.C. e 2200 a.C.) era usada para registrar colheitas e cheias do Rio Nilo (CALVO a*pud* DIEHL, SOUZA, DOMINGOS, 2007).

O ramo da estatística é responsável pelos processos de coletar, interpretar e apresentar os dados. Por conseguinte, com a utilização desta ciência, pode-se analisar os dados coletados e trazer informações úteis sobre eles.

2.1 ESTATÍSTICA DESCRITIVA

A estatística subdivide-se em três áreas (descritiva, inferencial e probabilistica) e a área que será utilizada para este trabalho é a descritiva, que segundo Guedes, *et al* (2019), tem o objetivo de sintetizar uma série de valores, tudo no intuito de ter uma visão dos mesmos e representá-los em três formas, sendo essas: gráficos, tabelas e medidas descritivas.

2.1.1 Média

A média é utilizada para encontrar um valor médio dentro de uma população de dados e, para se encontrar esse valor, deve-se realizar a soma de todos os elementos observados e por fim dividir esse resultado pela quantidade dos mesmos.

Apesar de ser ótima para achar o ponto de equilíbrio entre os dados Guedes *et al*. (2017), ela sofre influência de valores altos, os quais afetam a análise dos dados utilizando esta medida.

A fórmula (1) para se achar a média é $\bar{X} = \frac{\sum_{i=1}^{N} xi}{N}$, na qual xi representa um valor genérico da observação, sendo o N o número de observações.

2.1.2 Moda

A moda, diferente das outras medidas, analisa a quantidade de aparição de um determinado valor que está dentro de um conjunto de dados, não sofrendo influência de valores discrepantes. Porém, nem sempre um conjunto de dados terá uma moda,

pois se este não tiver valores repetidos, será considerado amodal. Em contrapartida, há casos em que um conjunto possui mais de uma moda.

A fórmula (2) para encontrar a moda é dada por $M_o = l + c \frac{\Delta_1}{\Delta_1 + \Delta_2}$, entende-se que, "$l$" representa o limite inferior da classe modal e "c" é a maior frequência na amostra. No que diz respeito a Δ_1 e Δ_2 , os mesmos são representados por Δ_1 é igual a frequência da classe modal subtraída pela frequência da classe anterior e Δ_2 é igual a frequência da modal subtraída pela frequência da classe posterior.

2.1.3 Mediana

A mediana é uma medida de tendência central que não se deixa ser afetada por valores exorbitantes no conjunto de dados. Ela consiste na ordenação dos dados que serão analisados e logo após, divididos em metade inferior e metade superior, pelo que por fim é calculado o valor central, o que faz com que dessa forma ela não sofra alterações por valores discrepantes na base de dados.

A fórmula (3) para se encontrar a mediana é dado por $M_d = l + c \frac{E_{md} - F_{ant}}{f_{md}}$, onde l representa o limite inferior da classe mediana, c é a menor frequência na amostra, E_{md} representa a frequência total dividida por dois, f_{ant}frequência acumulada até a classe mediana anterior e f_{md}representa a frequência simples da classe mediana.

2.1.4 Variância e Desvio Padrão

A variância é a medida estatística que define a distância dos valores com relação à média. É calculada a partir do quadrado da média das distâncias dos valores com relação ao valor esperado (média). Assim, devido ao valor que é encontrado pela variância ser dado de forma quadrática, é usado o desvio padrão, o qual é representado pela raiz quadrada da variância com a finalidade de se ter uma melhor compreensão sobre o valor obtido.

A fórmula (4) representa a variância e é descrita por $S^2 = \frac{\sum_{i=1}^{N} (xi - \bar{x})^2}{N}$, ao obtermos o valor da variância, utilizamos a fórmula (5) para encontrar o desvio padrão $s = \sqrt{S^2}$.

2.1.5 Coeficiente de Variação

Trata-se de uma medida utilizada para verificar a distância dos valores da distribuição em relação à média, em porcentagem, dizendo assim, se os dados se encontram muito dispersos ou não, (MARTINS; FONSECA, 1996).

Caso o coeficiente de variação corresponda a menos de 15%, tem-se uma dispersão baixa. Se corresponder de 15% a 30%, uma dispersão média. E de 30% para cima, tem-se uma dispersão dos dados muito elevada.

A fórmula (6) que serve para achar o coeficiente de variação é $CV = \frac{\sigma}{\bar{x}}\ x\ 100$, sendo σ o desvio padrão e $\bar{x}$ a média da distribuição.

2.1.6 Quartis

Os quartis são em certa forma similares à mediana. Porém, a diferença entre esses dois métodos estatísticos é que a mediana divide a distribuição em duas partes iguais, enquanto os quartis a dividem em quatro (TOLEDO; OVALLE, 2013).

Para a utilização do mesmo, é mister ordenar a distribuição em ordem crescente. Em seguida pode-se achar o primeiro quartil representado por Q_1 por meio da seguinte fórmula (7), $Q_1 = \frac{n}{4}$, em qual n é a quantidade de elementos na distribuição.

O segundo quartil também pode ser chamado de mediana, já que a forma para se achar o segundo quartil e a mesma para se achar a mediana. Achasse-o por meio da fórmula (8), $Q_2 = \frac{n}{2}$.

O terceiro quartil é identificado pelos dados não pertencentes a Q_1 e Q_2, o restante da distribuição. A sua fórmula (9) é dada por $Q_3 = \frac{3_n}{4}$.

2.1.7 Decil

Os decis são medidas separatrizes que separam os dados ordenados da distribuição em dez partes, o primeiro decil D_1 representa 10% dos menores valores da distribuição e pode ser achado através da formula (8) $D_1 = \frac{n}{10}$. Para achar o segundo decil, o qual representa 20% dos valores, usa-se a fórmula (10) $D_2 = \frac{2n}{10}$ e assim, segue-se até que se encontre os dez decis.

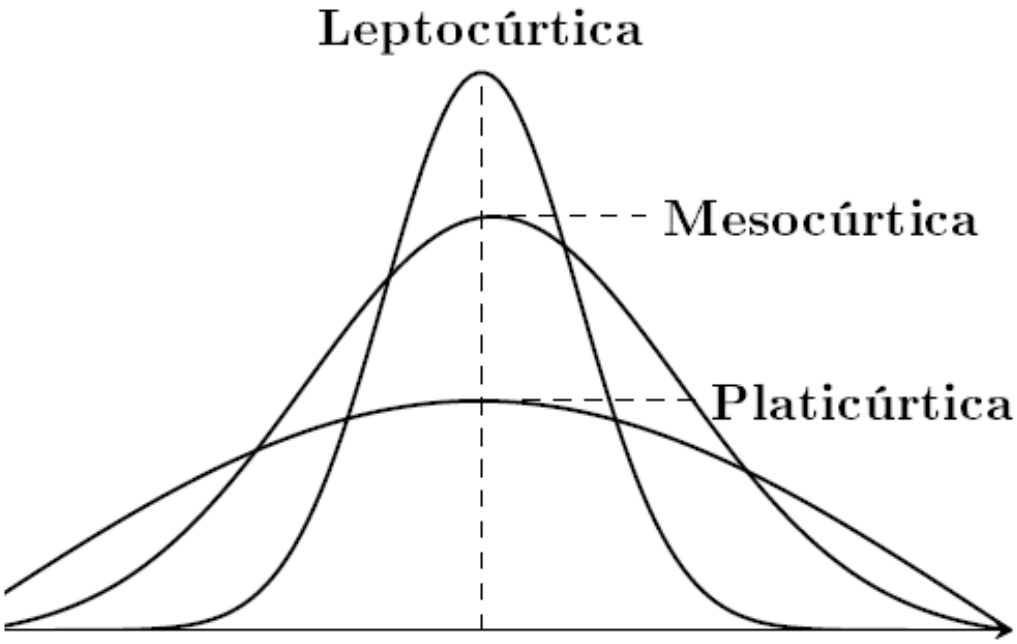
2.1.8 Curtose

Curtose é a visualização na curva gráfica da distribuição e pode ser classificada em 3 versões: Quando k > 0 a mesma é Leptocúrvica, quando k = 0; Mesocúrtica, quando os dados estão mais próximos da média, no entanto não tão próxima das medidas de tendência central; ou Platicúrtica, quando k < 0.

Nesse sentido, observa-se que a curtose nos diz a dispersão dos dados, s bem como, se eles estão próximos um do outro ou distantes.

Para achar a curtose podemos utilizar a fórmula (11) $k = \frac{(Q_3 - Q_1)}{2(D_{90} - D_{10})}$

Figura 1 – Ilustração do gráfico de Curtose



Fonte: biostatistics-uem.github.io (2019).

2.1.9 Assimetria

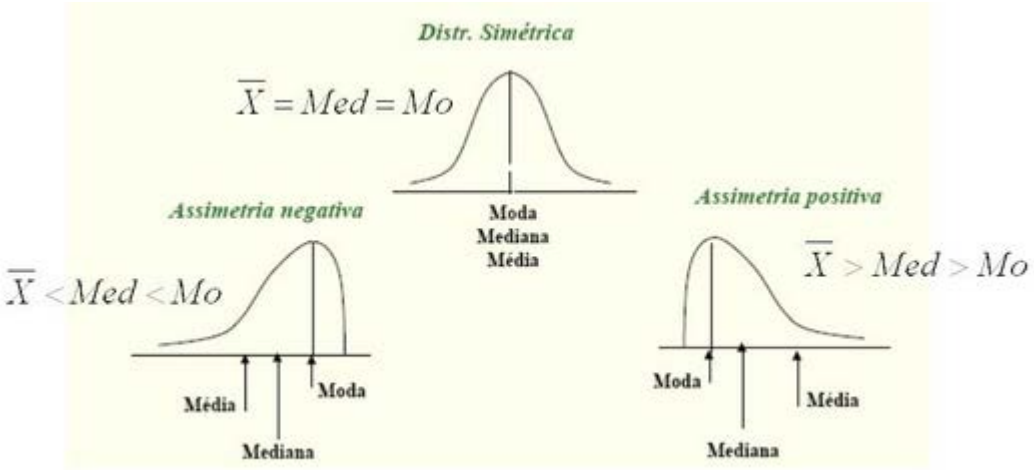
De acordo com Fonseca (*apud* Alexandre, 2011), assimetria é a medida do grau de afastamento de uma distribuição de sua medida central, geralmente se comparando média, moda e mediana.

Uma distribuição é considerada simétrica quando sua média, moda e mediana coincidem, logo, mostrando que a amostra possui uma distribuição normal. Caso tal amostra seja negativa, a média é menor que a mediana, que por sua vez é menor que

a moda, e o gráfico tem uma curva a esquerda. Por outro lado, caso ela seja positiva, sua moda é menor que a mediana, que por sua vez, é menor que a média e no gráfico a mesma tem uma curva a direita.

A fórmula (12) para se calcular a assimetria é $A_S = \frac{\bar{x} - M_o}{s}$, onde $\bar{x}$ representa a média, M_o a moda e s representa o desvio padrão. Isto posto, na figura 2 é possível ver o gráfico da assimetria.

Figura 2 – Ilustração do gráfico de assimetria



Fonte: biostatistics-uem.github.io (2019).

2.1.10 Tabelas

Devido ao uso de computadores para armazenamento de informação nos dias atuais, é muito difícil compreendê-los. Por isso as tabelas são muito utilizadas para nos permitir organizá-los e entendê-los (GUEDES, *et al*, 2019).

As Tabelas permitem, além de visualizar os dados de forma mais simples, identificar padrões através delas, pois pega um conjunto de dados dispersos e os organiza de uma maneira simples, facilitando na hora da análise exploratória.

Observa-se na figura 3 uma tabela estatística:

Figura 3 – Ilustração de uma tabela

Tabela 1 - Estatística descritiva para as variáveis fósforo – P (mg dm^{-3}), potássio – K ($mmol_c$ dm^{-3}), capacidade de troca de cátions – CTC ($mmol_c$ dm^{-3}) saturação de bases – V%, necessidade de fósforo – NP (kg ha^{-1}), necessidade de potássio – NK (kg ha^{-1}) e necessidade de calagem – NC (t ha^{-1}), de amostras coletadas na profundidade de 0,0-0,2 m, em amostragem de solo após o corte da cana-planta.

Estatística descritiva	P	K	CTC	V%	NP	NK	NC
Média	3,90	2,18	54,46	52,58	174,60	88,80	0,53
Mediana	4,00	2,20	52,00	53,05	178,00	82,00	0,43
[1]DP	1,45	0,98	21,47	10,95	17,26	28,15	0,30
Variância	1,60	0,97	464,50	119,92	298,00	792,49	0,25
[2]CV (%)	37,32	44,95	39,42	20,82	9,88	31,70	56,60
Assimetria	1,56	0,11	0,69	0,12	-2,94	0,13	0,63
Curtose	4,06	-0,06	-0,73	-0,30	6,85	0,29	-0,82
[3]d	0,24	0,09[ns]	0,16	0,08[ns]	0,33	0,29	0,14

[1]DP = desvio padrão; [2]CV = coeficiente de variação; [3]d = teste de normalidade, [ns]não significativo pelo teste de Kolmogorov-Smirnov.

Fonte: Scielo (2010).

2.1.11 Erro Padrão

O erro padrão é uma medida de variação utilizada para obter um intervalo de confiança para a média da distribuição analisada, retornando um nível de significância, que quanto menor for, menos dispersos são os valores da média da distribuição.

A fórmula para se achar o erro padrão é dado por $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, sendo σ o desvio padrão, e $\sqrt{n}$ a raiz quadrada do tamanho da amostra.

2.2 GRÁFICOS

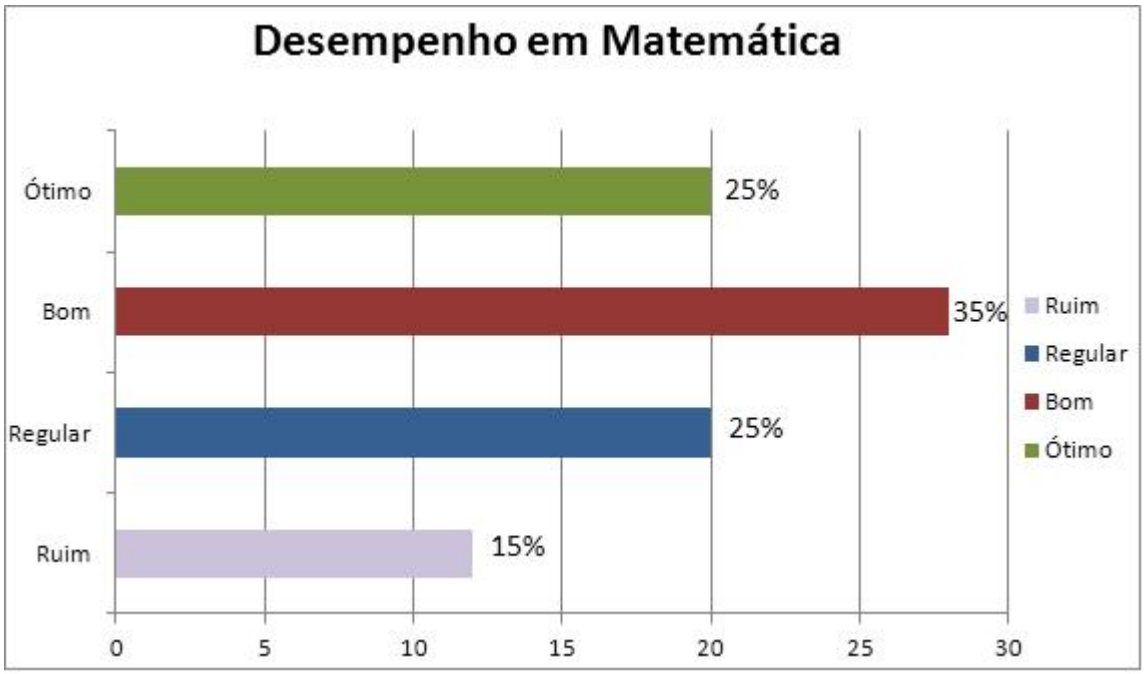
Gráficos são muito utilizados na estatística para realizar a representação de dados de uma forma visual e de fácil entendimento para todos as pessoas que o observarem. Apesar de não mostrar tanta informação quanto uma tabela, a utilização tomou uma grande importância na visualização, pois conseguimos entender o que está acontecendo com determinada análise de forma clara e objetiva.

Gráficos facilitam muito o entendimento dos dados, mas nem todos os gráficos servem para todas as ocasiões. Segundo Guedes, *et al* (2019, p. 17): “Todo gráfico, em sua versão final deve primar pela simplicidade, clareza e veracidade nas informações. Para atingir tal objetivo, a construção de um gráfico exige muito trabalho e cuidados”.

2.2.1 Gráfico de Barra

É um gráfico representado por barras, no qual o eixo y representa a variável a ser mostrada e no eixo X a quantidade que está variável se apresenta na amostra. É mais recomendável para quando as variáveis possuem descrição longa.

Figura 4 – Gráfico de Barra

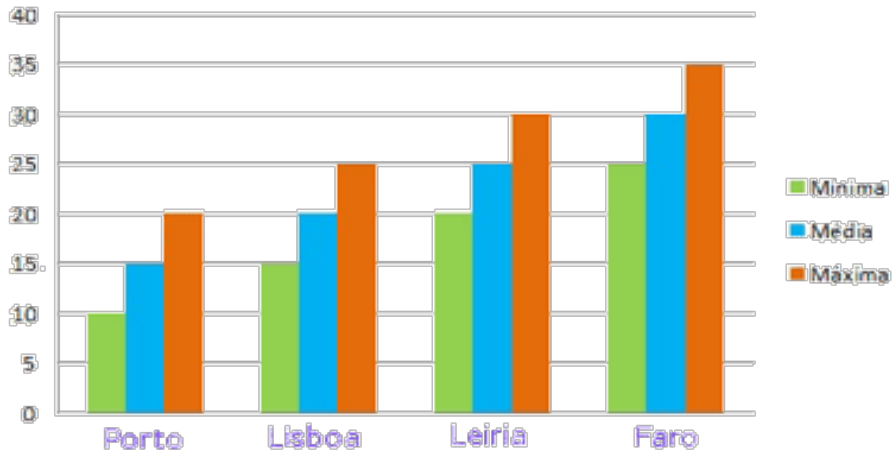


Brasilescola (2019)

2.2.2 Gráfico de Coluna

Similar ao gráfico de barra, mas neste caso as barras são feitas no eixo das abscissas e a quantidade que a variável aparece no eixo y, são indicados para representar variáveis cuja descrição seja breve. O número de barras neste gráfico não deve superar 12, pois atrapalha sua visualização.

Figura 5 – Gráfico de coluna

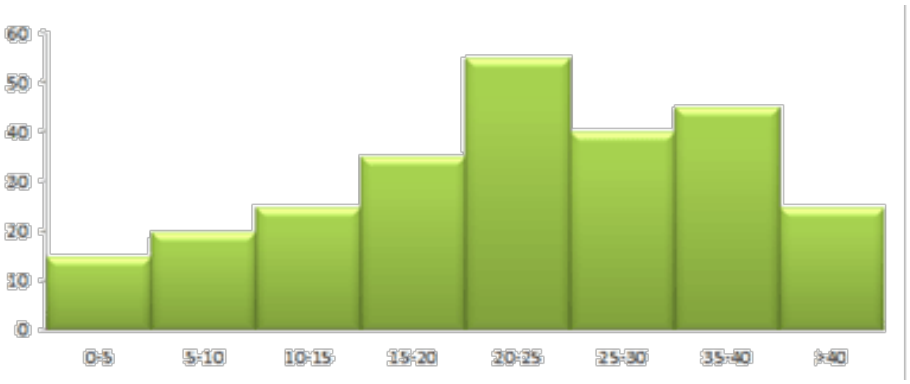


Nunes (2019)

2.2.3 Gráfico Histograma

Diferentemente do gráfico de barra e do de coluna, utilizamos o histograma quando possuímos uma base de dados com valores distintos e de caráter continuo ou discreto. As colunas não têm espaçamento entre ela, fato que permite a exibição de mais dados.

Figura 6 – Gráfico histograma

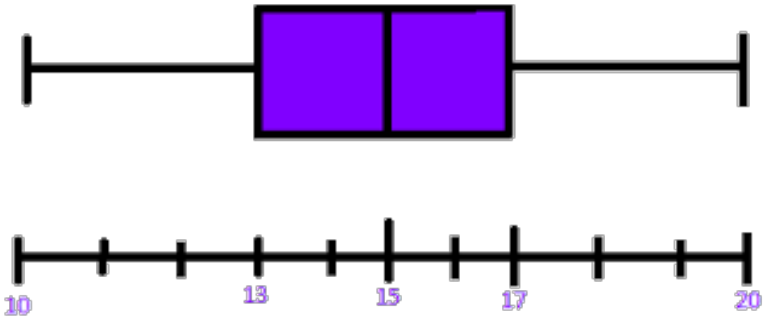


Nunes (2019)

2.2.4 Gráfico Boxplot

O gráfico de *boxplot* é um gráfico feito com base em 5 medidas estatísticas: valor mínimo, valor máximo, mediana, primeiro quartil e terceiro quartil. Não tem muito utilização em relatórios empresariais, porém, no mercado financeiro é muito utilizado para representar o andamento do mesmo. Já que ele mostra a máxima e a mínima, fazendo com que se possa ter uma boa análise sobre o mercado.

Figura 7 – Gráfico boxplot



Nunes (2019)

2.3 FERRAMENTAS DE DESENVOLVIMENTO

2.3.1 Python

Python teve sua primeira versão lançada em fevereiro de 1991 e foi criada por Guido van Rossum. O principal motivo para o próprio desenvolver a nova linguagem era porque ele achava que muito tempo era perdido devido a sintaxe da linguagem C. Apesar de existir outras opções como o *Shell*, muitas funções não funcionavam devido à restrição do mesmo (MAGNUN, 2014).

Tendo esse problema em mente, Rossum começou o projeto de desenvolvimento da linguagem *python,* com uma sintaxe simples e com todas as funcionalidades não contidas no *shell*, unindo desse modo, “o melhor dos dois mundos”. Sendo assim, nos dias atuais referida linguagem está no top do *ranking* das mais utilizadas. Segundo Silva (*apud* DIAS, 2018, p.9): “o python foi criado com o

intuito de pesquisa em Ciência da Computação". Nesse sentido, A figura 8 mostra a popularidade da linguagem em 2019.

Veja-se:

Figura 8 – Ranking das linguagens de programação

Jan 2019	Jan 2018	Change	Programming Language	Ratings	Change
1	1		Java	16.904%	+2.69%
2	2		C	13.337%	+2.30%
3	4	↑	Python	8.294%	+3.62%
4	3	↓	C++	8.158%	+2.55%
5	7	↑	Visual Basic .NET	6.459%	+3.20%
6	6		JavaScript	3.302%	-0.16%
7	5	↓	C#	3.284%	-0.47%
8	9	↑	PHP	2.680%	+0.15%
9	-	⇈	SQL	2.277%	+2.28%
10	16	⇈	Objective-C	1.781%	-0.08%
11	18	⇈	MATLAB	1.502%	-0.15%
12	8	⇊	R	1.331%	-1.22%
13	10	↓	Perl	1.225%	-1.19%
14	15	↑	Assembly language	1.196%	-0.86%
15	12	↓	Swift	1.187%	-1.19%
16	19	↑	Go	1.115%	-0.45%
17	13	⇊	Delphi/Object Pascal	1.100%	-1.28%

Fonte: TIOBE (2019).

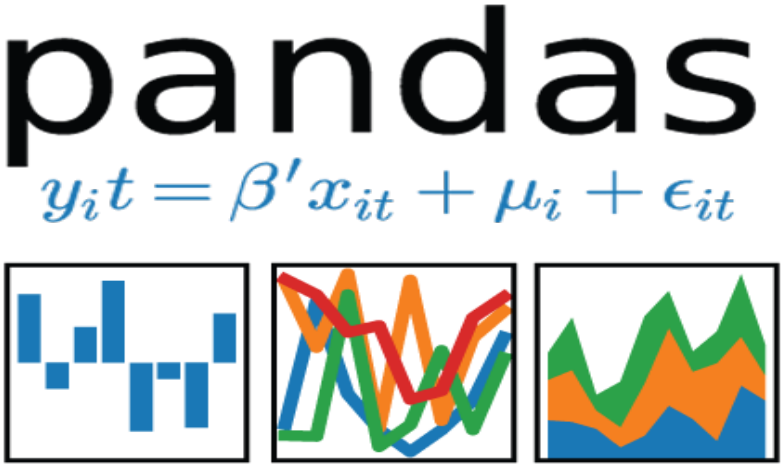
Devido ao grande aumento de popularidade, muitas ferramentas surgiram para esta linguagem uma muito famosa para a análise de dados que é chamada de pandas.

2.3.2 Pandas

Pandas é uma biblioteca escrita em *python*, criada para trabalhar com dados em forma de tabela e explorá-los de forma rápida e simples. Seus inúmeros métodos facilitam muito na hora de calcular estatísticas e até mesmo na plotagem de gráficos, funcionalidades que já vem na biblioteca por padrão.

Tem sido considerada uma verdadeira ferramenta para se contornar a grande quantidade de dados, "Data is unavoidably messy in real world. And Pandas is *seriously* a game changer when it comes to cleaning, transforming, manipulating and analyzing data. In simple terms, Pandas helps to clean the mess" (LEE, 2018).

Figura 9 – Logo da biblioteca PANDAS



Fonte: SIMONLINDGREN (2019, *online*).

2.3.3 Anaconda-Navigator

A plataforma *Anaconda-Navigator* foi criada com o intuito de reunir tudo que um cientista de dados necessita, em vista da mesma possuir um ambiente em nuvem, onde podemos realizar o *upload* de nossos ambientes para que fiquem seguros.

A plataforma está ganhando muito espaço devido a sua facilidade e praticidade em organizar pacotes necessário e mantê-los atualizados ou estáticos para que não afetem um projeto em andamento, por apresentar criação fácil de diversos ambientes e pacotes para análise científica de dados (VIDAL, 2016).

Figura 10 – Logo ferramenta Anaconda-Navigator

Fonte: anaconda.org (2019, *online*).

2.3.4 Seaborn

Seaborn é uma biblioteca baseada em outra biblioteca, a matplotlib, ela é de código aberto e sua principal diferença é que a mesma vem com a proposta de produzir gráficos de alto nível, mais elegantes e fazendo com que o usuário tenha um melhor entendimento sobre os dados analisados. “Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics” (WASKON, 2018).

2.3.5 Jupyter Notebook

Jupyter notebooks é uma aplicação que roda no navegador web, o mesmo contém um esquema de listas encadeadas nomeada de células, aonde é possível executar trechos de código separadamente, o que facilita na experimentação quando se lida com dados. (INGARGIOLA, 2015).

3 IMPLEMENTAÇÂO DO SOFTWARE

A implementação deste *software* ocorreu em várias etapas, desde a escolha do sistema operacional (SO), o download e configuração do *anaconda-navigator*, até a configuração dos pacotes e organizações dos mesmos. Será abordado as ferramentas que já estão inseridas por padrão e quais iremos precisar bem como *prints* de resultados obtidos pela análise exploratória.

3.1 INSTALAÇÃO

O *anaconda-navigator* tem duas versões, uma mais leve e outra mais pesada. Porém, a versão mais leve não tem todas as ferramentas e recursos visuais que sua versão completa possui. Tendo este fato em consideração, foi optado pela versão completa do *anaconda-navigator*.

Para iniciar a utilização da plataforma *anaconda-navigator*, primeiro é necessário ir no site do *anaconda-navigator* e realizar o download do arquivo e após isso executa-lo para que comece a instalação na máquina.

A plataforma já vem com um aparato de ferramentas como *jupyter notebooks*, *spyder*, *orange* e etc... Como foi utilizado apenas o *jupyter notebooks*, as outras ferramentas não serão abordadas.

3.2 INSTALANDO PACOTES

A plataforma vem com várias bibliotecas do *python*, mesmo assim, uma biblioteca ou outra podem não vir por padrão, e para facilitar o trabalho, a plataforma permite uma rápida e simples forma de adicionar as bibliotecas via linha de comando no Linux, através do comando *conda install* -c ‘nome da biblioteca’. Fazendo com que assim seja possível deixar nosso ambiente o mais completo possível e com tudo que for necessário, para que cada novo projeto possa “herdar” os mesmos pacotes sem precisar adicionar as bibliotecas sempre que um novo projeto for criado.

3.3 CRIANDO AMBIENTES

Caso não seja necessário todos os pacotes que já vêm por padrão, podemos criar outro ambiente com a versão do *python* que se adequar mais as necessidades do projeto e também as bibliotecas necessárias assim como suas respectivas versões, desta forma a plataforma nos disponibiliza um controle muito bom sobre o que podemos ter e separar nossos projetos em diferentes ambientes.

3.4 REALIZANDO A ANÁLISE DE DADOS

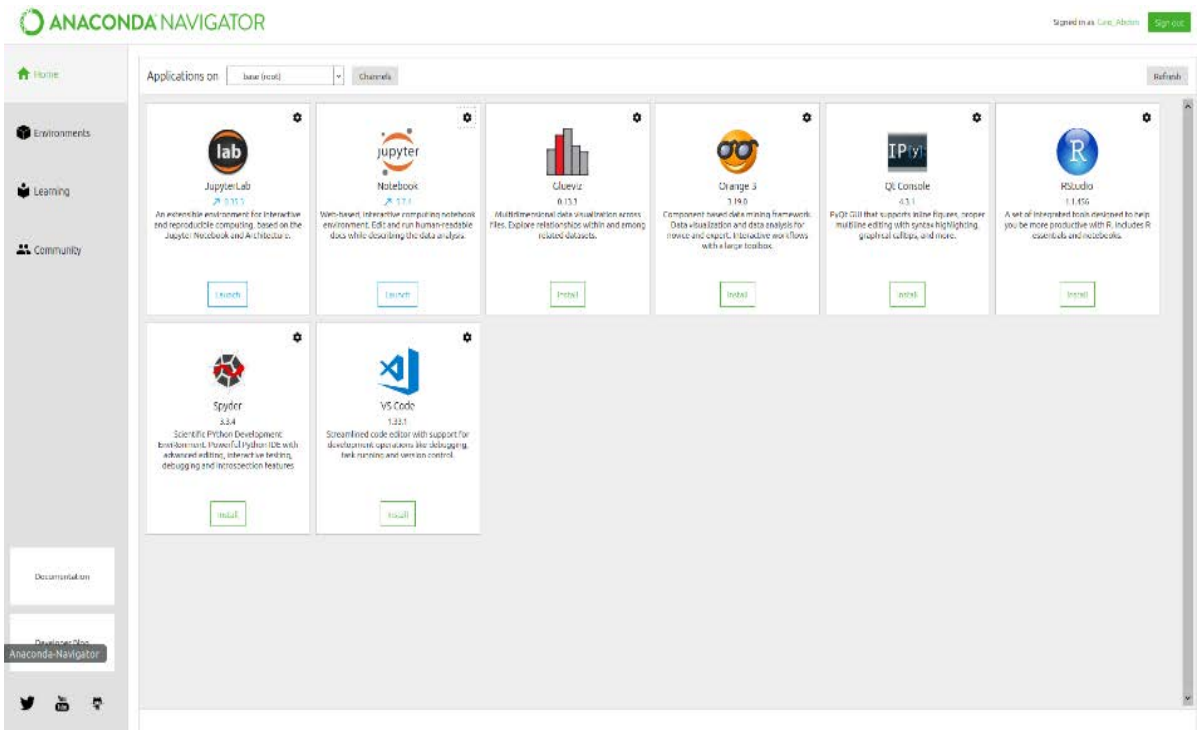
Após todo o procedimento de instalação e preparação do ambiente, dá início a análise de dados, o processo em que a base de dados escolhida passa por várias etapas. Pré-processamento, quando fazemos uma verificação da base de dados para checar se todos os valores estão inseridos e se nem um deles está faltando. Análise exploratória, quando verificamos os valores através de métodos estatísticos a procura de valores discrepantes e depois verificar se estes estão dentro ou fora do escopo aceitável. E por último à análise de fato, o estágio em que plotamos gráficos representando *insigths* sobre os dados, e estes podem ser usados pelos administradores para a tomada de decisão de uma empresa.

4 FUNCIONAMENTO DO PROGRAMA

4.1 RESULTADOS

Nesta seção será retratado como o programa funciona, mostrando os resultados gerados através do mesmo, bem como o *anaconda-navigator* e suas ferramentas.

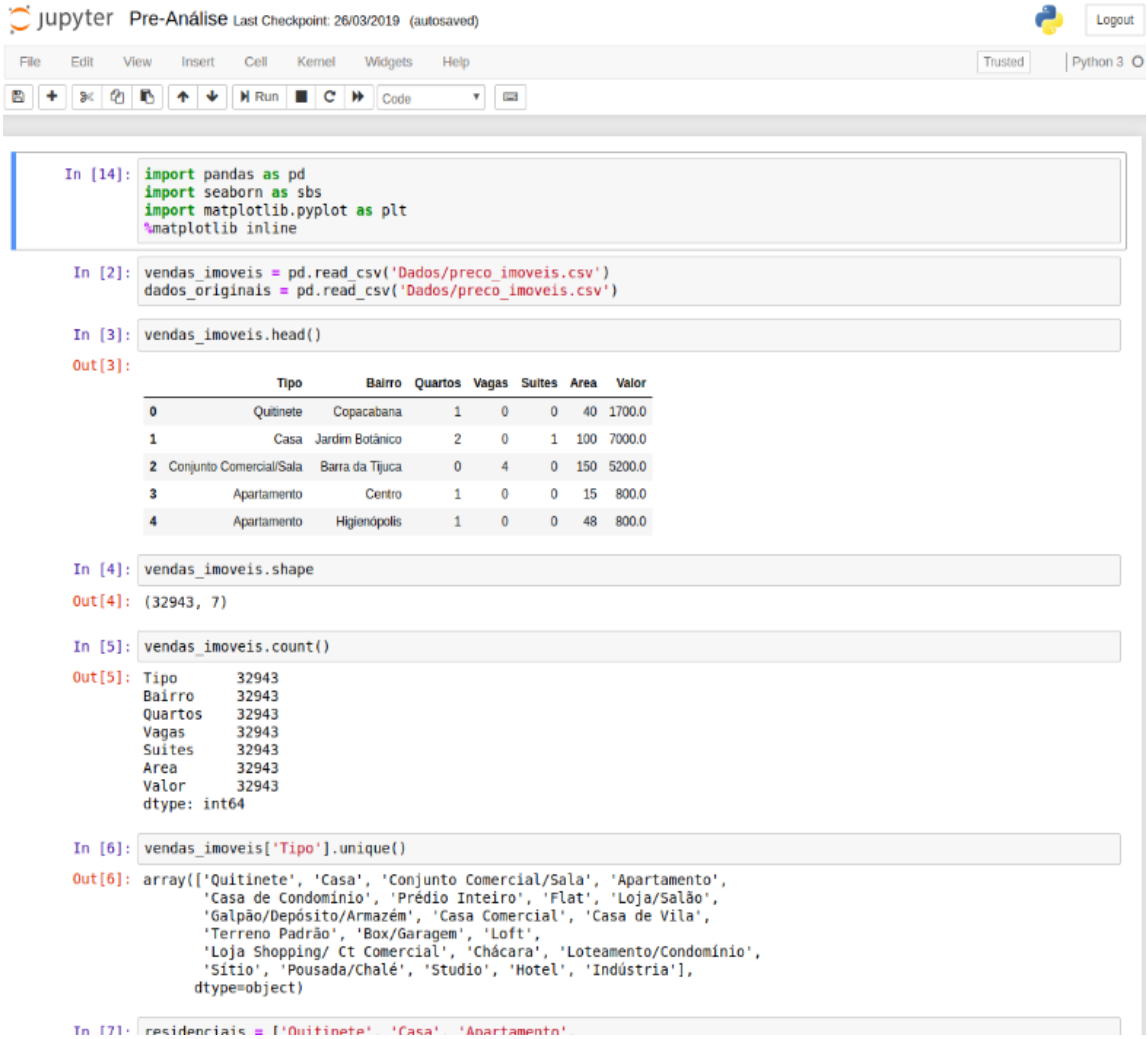
Figura 11 – Anaconda-navigator



Fonte: Autor (2019).

Na figura 11 pode-se verificar como é a interface do *anaconda-navigator*, aonde o mesmo mostra as ferramentas disponíveis, bem como as que tem de ser instaladas e as que podemos instalar caso seja necessário.

Figura 12 - Interface do programa



Fonte: Autor (2019).

A imagem 12 mostra como é o programa, foi escolhido por utilizar a interface da própria plataforma. Pois, o intuito é focar nas funcionalidades e no código, para mostrar suas vantagens e facilidade de uso.

Na figura 12 ocorre o processo de pré-análise, onde se verifica se existem valores faltantes na base de dados e se está tudo pronto para prosseguir para a análise exploratória de fato.

Figura 13 – imóveis residenciais e estatísticas

```
In [3]: residenciais.head()
```

Out[3]:

	Tipo	Bairro	Quartos	Vagas	Suites	Area	Valor
0	Quitinete	Copacabana	1	0	0	40	1700.0
1	Casa	Jardim Botânico	2	0	1	100	7000.0
2	Apartamento	Centro	1	0	0	15	800.0
3	Apartamento	Higienópolis	1	0	0	48	800.0
4	Apartamento	Vista Alegre	3	1	0	70	1200.0

```
In [4]: residenciais['Tipo'].unique()
```

```
Out[4]: array(['Quitinete', 'Casa', 'Apartamento', 'Casa de Condomínio', 'Flat',
       'Casa de Vila', 'Loft'], dtype=object)
```

```
In [5]: analise_casas = residenciais.query("Tipo == 'Casa'")
        analise_quitinete = residenciais.query("Tipo == 'Quitinete'")
        analise_apartamento = residenciais.query("Tipo == 'Apartamento'")
        analise_casa_de_condomínio = residenciais.query("Tipo == 'Casa de Condomínio'")
        analise_flat = residenciais.query("Tipo == 'Flat'")
        analise_casa_de_vila = residenciais.query("Tipo == 'Casa de Vila'")
        analise_loft = residenciais.query("Tipo == 'Loft'")
```

```
In [6]: analise_casas['Valor'].describe()
```

```
Out[6]: count      965.000000
        mean      6793.454922
        std       8955.421677
        min        400.000000
        25%       1100.000000
        50%       2200.000000
        75%       9800.000000
        max      60000.000000
        Name: Valor, dtype: float64
```

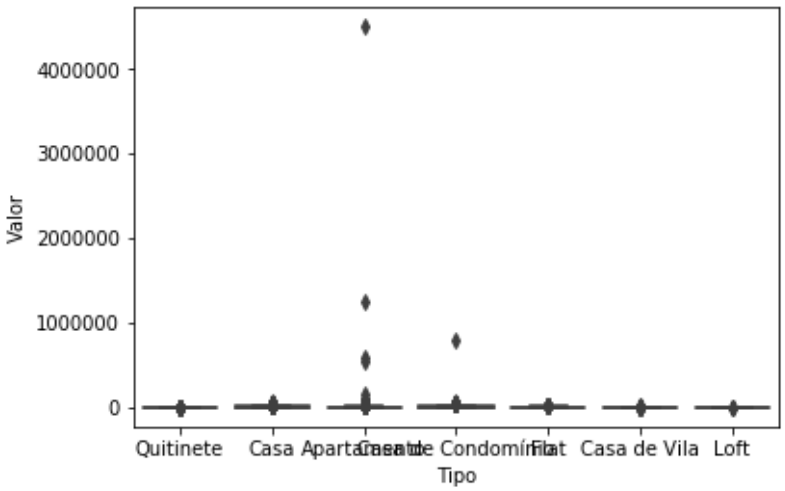
Fonte: Autor (2019).

A figura 13 mostra a separação dos imóveis residenciais dos imóveis comerciais. Após isso, os imóveis residenciais foram unidos e encontradas as estatísticas descritivas sobre os mesmos, expõe-se que com apenas uma linha de código “analise_casas[‘Valor].describe()’, já se obtém um resumo estatístico.

Esta funcionalidade ajuda o mercado imobiliário no quesito que, é possível isolar certos setores e os analisar de forma independente, trazendo uma análise mais profunda dos dados.

Figura 14 – *boxplot dos valores*

```
In [6]: sbs.boxplot(x="Tipo",y="Valor",data=residenciais)
Out[6]: <matplotlib.axes._subplots.AxesSubplot at 0x7f1609725b38>
```

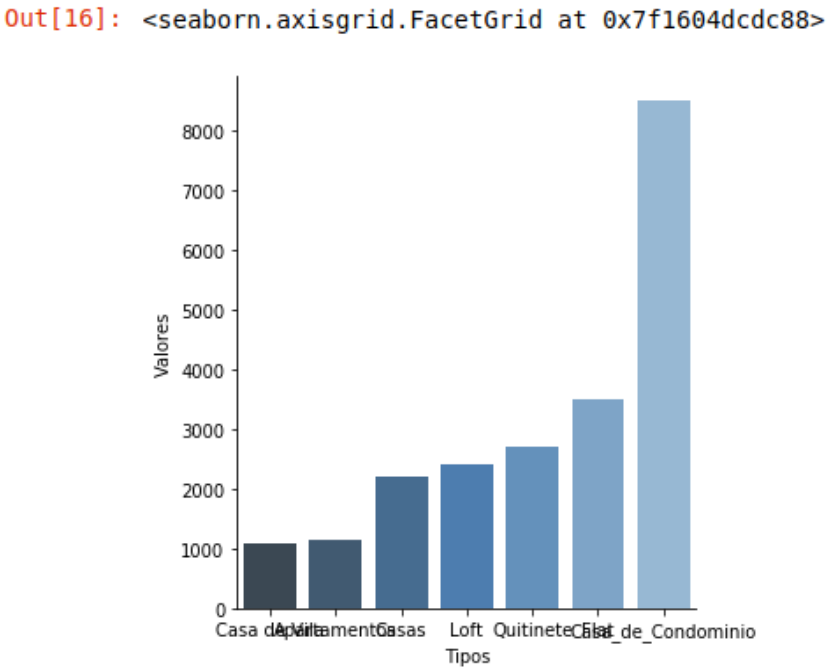


Fonte: Autor (2019).

A figura 14 mostra um gráfico de *boxplot* sobre os imóveis residências separados anteriormente. E como visto, o gráfico mostra que existem valores muito discrepantes dos demais, principalmente nos apartamentos. Com esta informação é possível ver que não podemos usar a média. Pois como foi falado no capitulo 2.1.1, a mesma é influenciada por tais valores.

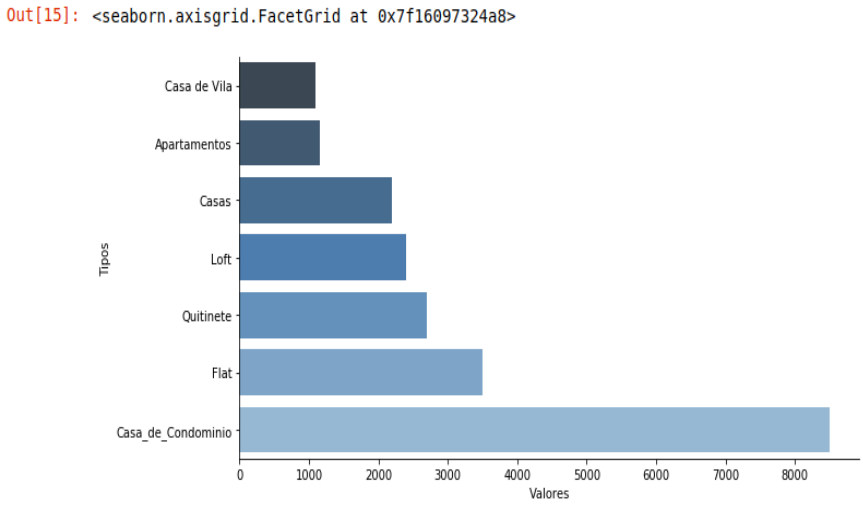
Para que se tenha uma análise mais precisa, far-se-á uso da mediana, que não sofre influência destes valores discrepantes, (como foi citado no capitulo 2.1.3), e utilizando desta medida estatística, são gerados gráficos sobre o valor de cada imóvel.

Figura 15 – Gráfico Errado da Mediana dos Valores



Fonte: Autor (2019).

Figura 16 – Gráfico Certo da Mediana dos Valores



Fonte: Autor (2019).

A figura 15 mostra um gráfico que foi utilizado de maneira indevida devido a forma que os dados foram organizados, e na figura 16 podemos ver os gráficos sendo aplicados da maneira correta, aonde suas informações são bem legíveis e é possível ter um boa leitura sobre os mesmos.

Com estas imagens é testemunhável a importância da boa utilização dos gráficos. Todavia que, a má utilização dos mesmos pode acarretar em um prejuízo enorme no entendimento das informações.

É possível analisar também a partir do gráfico da figura 16, o quanto cada setor de imóvel arrecadou, desta fora fazendo com que as empresas imobiliárias possam focar esforços para equilibrarem o lucro gerado por cada setor.

Figura 17 – Metro quadrado

Out[58]:

	Tipo	Bairro	Quartos	Vagas	Suites	Area	Valor	Preco M2
0	Quitinete	Copacabana	1	0	0	40	1700.0	42.50
1	Casa	Jardim Botânico	2	0	1	100	7000.0	70.00
2	Conjunto Comercial/Sala	Barra da Tijuca	0	4	0	150	5200.0	34.67
3	Apartamento	Centro	1	0	0	15	800.0	53.33
4	Apartamento	Higienópolis	1	0	0	48	800.0	16.67

Fonte: Autor (2019).

Na figura 17, é retratado como podemos utilizar os dados disponíveis para achar mais informações, ao dividir o valor do imóvel pela área do mesmo, podemos encontrar o valor do m², agregando mais informação a nossa base de dados.

Figura 18 – Resumo estatístico

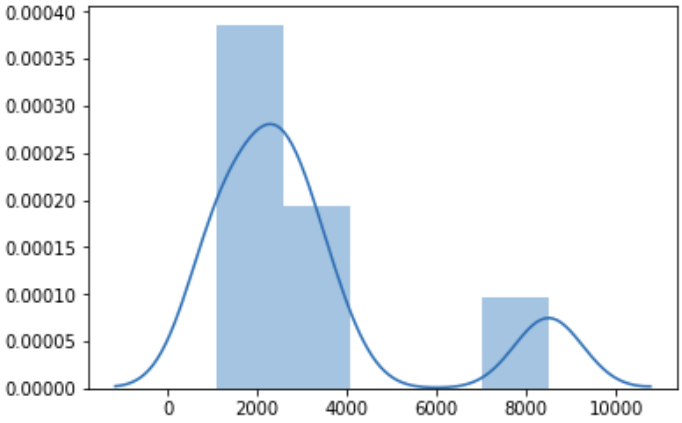
	Estatisticas	Valores
0	Media	1246.84
1	Mediana	1150.00
2	Moda	1000.00
3	Variancia	314189.35
4	Desvio_Padrao	560.53
5	Quartil 1	900.00
6	Quartil 3	1500.00
7	Decil 1	700.00
8	Decil 9	1875.00
9	Coeficiente de Variacao	44.96
10	Assimetria	15.53
11	Curtose	2.63
12	Erro Padrao	19.39
13	Quantidade	836.00

Fonte: Autor (2019).

Na figura 18, é possível ver um resumo estatístico completo dos dados das vendas de quitinetes, todo o resumo foi realizado com a utilização de um método criado no *python*. Desta forma, em futuras análises, não é preciso selecionar novamente as colunas, pois o software procura a variável pelo nome e realiza a criação da tabela de forma automatizada.

Através da análise destes dados, é possível se analisar cada setor individualmente, fazendo assim, com que seja possível verificar o setor em questão está com uma boa taxa de lucro ou se precisa de ajustes.

Figura 19 - Histograma



Fonte: Autor (2019).

Na figura 19 foi plotado um histograma referente aos dados analisados na figura 15. Podemos ver através deste gráfico e também pelo resumo estatístico, que a amostra é leptocúrtica, o que diz que os dados são muito distantes da média.

Ao analisar o mesmo gráfico, a assimetria diz que sua moda é maior que a mediana que é maior que a média e que a mesma é negativa. E ao analisar o seu valor na figura 15, a mesma possui valores acima da média.

É possível ver com todos estes gráficos e tabelas, o grande impacto que esta análise pode ter n mercado imobiliário, e como ela pode mudar o rumo de decisões da empresa. Por isso, tem-se tornado tão crucial neste ramo empresarial.

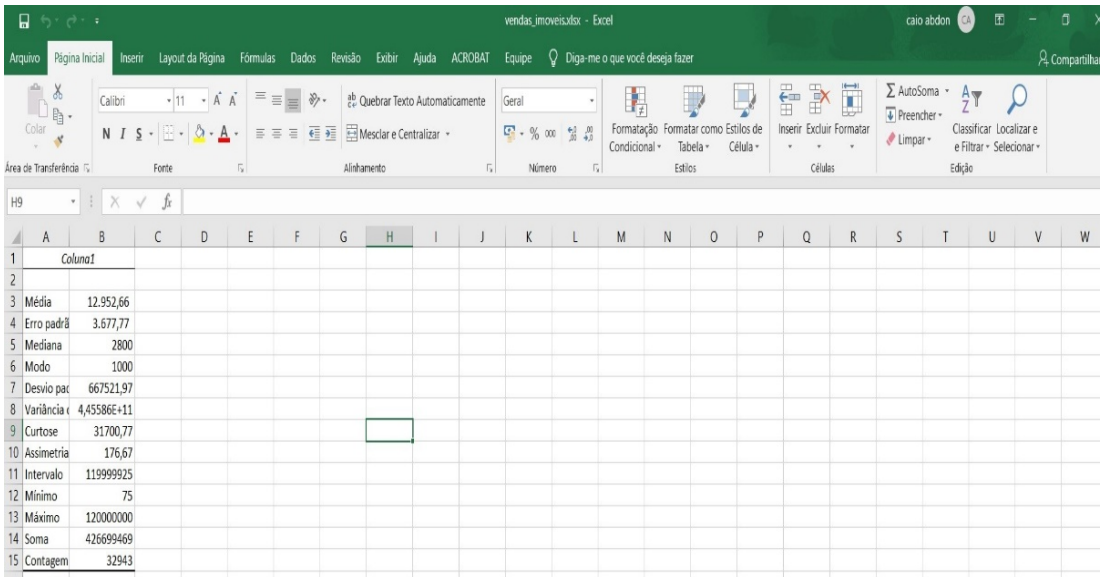
4.2 COMPARAÇÃO ENTRE PLATAFORMAS

Figura 20

Estatisticas	Valores
Media	12952.66
Mediana	2800.00
Moda	1000.00
Variancia	4.455856e+11
Desvio_Padrao	667521.97
Quartil 1	1500.00
Quartil 3	6500.00
Decil 1	900.00
Decil 9	15000.00
Coeficiente de Variacao	5153.55
Assimetria	31700.77
Curtose	176.67
Erro Padrao	3677.77
Quantidade	32943.00

Fonte: Autor (2019)

Figura 21

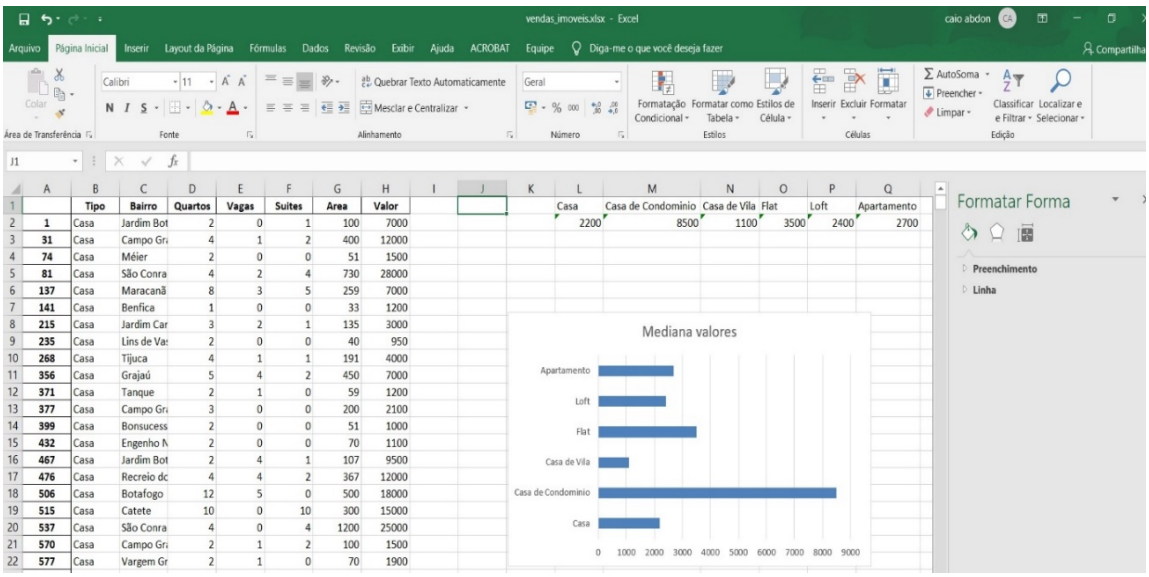


Fonte: Autor (2019)

A figura 20 e a figura 21 representam um resumo estatístico do valor de todos os imóveis juntos, o primeiro sendo utilizando programação e o segundo no excel©, é notável que os resultados são os mesmos. No entanto, a diferença destes dois é que a velocidade deste resumo no excel© levou entre 5 a 7 segundos, enquanto que com a programação foi instantâneo.

Ao finalizar o desenvolvimento do método, não é preciso reescrevê-lo em análises futuras, basta apenas carregar nosso *dataframe* com as novas informações a serem analisadas. Isso, acelera futuras análises e automatiza o processo.

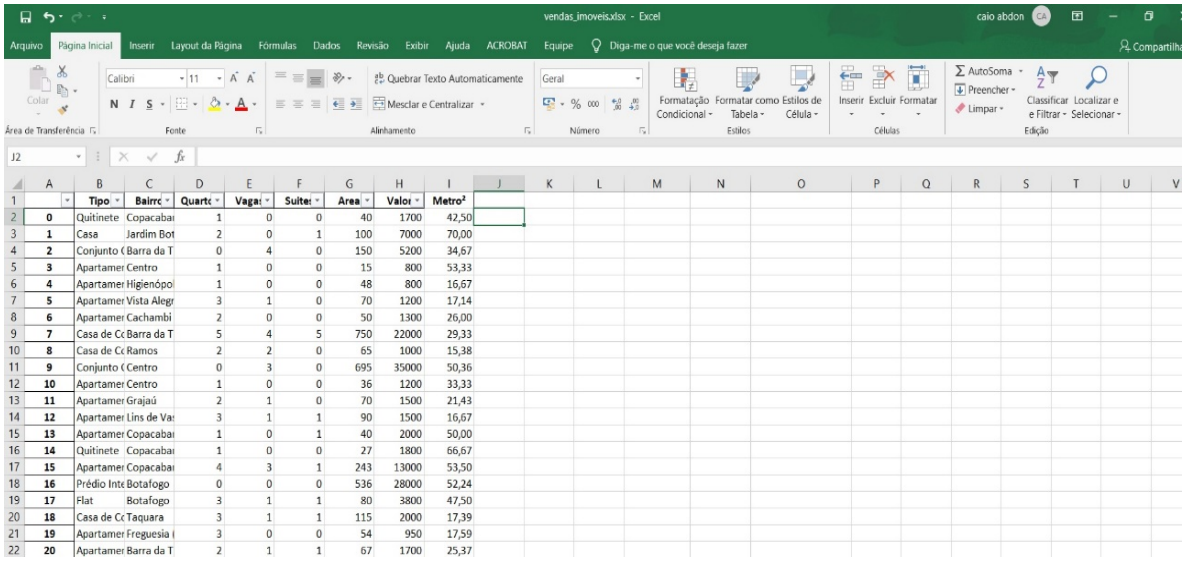
Figura 22 – gráfico mediana



Fonte: Autor (2019)

Na figura 22 é possível observar um gráfico das medianas dos imóveis imobiliários, o mesmo que representa a figura 16. Apesar dos resultados serem os mesmos, o esforço que levou para se separar o *dataframe* no excel© fez com que está tarefa levasse muito tempo, e em caso de análises futuras, o mesmo processo teria de ser feito novamente, o que com a programação não seria necessário.

Figura 23 Metro quadrado



		Tipo	Bairro	Quarto	Vaga	Suite	Area	Valor	Metro²
2	0	Quitinete	Copacaba	1	0	0	40	1700	42,50
3	1	Casa	Jardim Bot	2	0	1	100	7000	70,00
4	2	Conjunto (	Barra da T	0	4	0	150	5200	34,67
5	3	Apartamer	Centro	1	0	0	15	800	53,33
6	4	Apartamer	Higienópol	1	0	0	48	800	16,67
7	5	Apartamer	Vista Alegr	3	1	0	70	1200	17,14
8	6	Apartamer	Cachambi	2	0	0	50	1300	26,00
9	7	Casa de Cc	Barra da T	5	4	5	750	22000	29,33
10	8	Casa de Cc	Ramos	2	2	0	65	1000	15,38
11	9	Conjunto (	Centro	0	3	0	695	35000	50,36
12	10	Apartamer	Centro	1	0	0	36	1200	33,33
13	11	Apartamer	Grajaú	2	1	0	70	1500	21,43
14	12	Apartamer	Lins de Va	3	1	1	90	1500	16,67
15	13	Apartamer	Copacaba	1	0	1	40	2000	50,00
16	14	Quitinete	Copacaba	1	0	0	27	1800	66,67
17	15	Apartamer	Copacaba	4	3	1	243	13000	53,50
18	16	Prédio Inte	Botafogo	0	0	0	536	28000	52,24
19	17	Flat	Botafogo	3	1	1	80	3800	47,50
20	18	Casa de Cc	Taquara	3	1	1	115	2000	17,39
21	19	Apartamer	Freguesia	3	0	0	54	950	17,59
22	20	Apartamer	Barra da T	2	1	1	67	1700	25,37

Fonte: Autor (2019)

Na figura 23, observasse que é possível achar novas informações a partir de informações já existentes. Por exemplo, foi achado o valor do metro quadrado dos imóveis, cruzando os dados do valor e da área, nesta análise o excel© não teve dificuldades, perdendo somente no quesito automatização.

Quanto a comparação com o banco de dados, os dois possuem velocidades parecidas e automatização quando se trata de análise de dados. Porém, o grande diferencial é que no banco de dados, a análise afeta o funcionamento de todo o sistema, o que torna a análise inviável. Com a programação separamos esta carga de trabalho, fazendo assim que a análise possa ser feita sem que a mesma afete as outras funcionalidades de um sistema

5 CONSIDERAÇÕES FINAIS

O *Big Data* é um problema que surgiu a pouco tempo, e dificultou muito a exploração de dados através das formas convencionais que tínhamos, e por isso que trouxe esta proposta de análise através da programação.

Tendo em vista está problemática, foi a implementado este software, como uma alternativa para resolver este problema que a cada dia se agrava, fazendo com que a análise seja realizada de forma mais eficiente e precisa.

Com as vantagens que o software apresenta, pode-se contornar a situação atual da falta de forma simples e eficiente, sem afetar a performance do banco de dados da aplicação para realizar a análise de dados.

5.1 DIFICULDADES ENCONTRADAS

Como este problema é relativamente novo no mercado, existe uma carência muito grande quanto a materiais falando sobre o assunto, além do desenvolvimento do *software* ser um grande desafio devido a necessidade de entender os dados e não somente processá-los, o que por várias vezes fez com que reiniciar o projeto fosse necessário.

5.2 TRABALHOS FUTUROS

Para trabalhos futuros pretende-se ampliar as funcionalidades já apresentadas aplicando-se *machine-learning* para explorar estatísticas mais avançadas e fazer a predição de vendas, o que melhoraria ainda a análise.

Além de *machine-learning*, podem ser implementadas ferramentas como o *apache spark* para processamento distribuído e paralelo com intuito de dividir a tarefa entre computadores para agilizar o processo, e eliminar as limitações de hardware.

Com a grande quantidade de dados geradas, os gráficos mais comuns talvez não sejam suficientes, gostaria de ampliar esta visualização com técnicas gráficas mais avançadas, como por exemplo, mapa de calor, gráfico de radar, gráfico de área e etc...

6 REFERÊNCIAS

ALEXANDRE. **Assimetria e Curtose.** *Online*. Disponível em: http://alexandreprofessor.blogspot.com/p/assimetria-e-curtose.html. Acesso em: 25/03/2019.

BIOSTATISTICS-UEM.GITHUB.IO. **Análise Descritiva.** *Online.* Disponível em: https://biostatistics-uem.github.io/Bio/descritiva.html. Acesso em: 17/06/2019

BRASILESCOLA. **Gráficos.** *Online.* Disponível em: https://brasilescola.uol.com.br/matematica/graficos.htm. Acesso em: 17/06/2019

DIAS. K. **Aparato para estudo sobre ciência de dados**. Escola de gestão de negócios, 2018.

DIEHL, C.; SOUZA, M.; DOMINGOS, L. **O Uso da Estatística Descritiva na Pesquisa em Custos: Análise do XIV Congresso Brasileiro de custos**. **Contexto**. Porto Alegre, v. 7, n. 12, 2º semestre 2007. Disponível em: https://seer.ufrgs.br/ConTexto/article/view/11157. Acesso em: 31/05/2019.

DINO. **Transformação Digital Chega ao Mercado Imobiliário e Influencia a Gestão do Processo de Vendas**. *Online*. 31/08/2018. Disponível em: https://exame.abril.com.br/negocios/dino/transformao-digital-chega-ao-mercado-imobiliario-e-influencia-a-gestao-do-processo-de-vendas/. Acesso em: 11/05/2019.

DOHAN, R. **Como as Novas Tecnologias Influenciaram o Mercado Imobiliário**. *Online*. 13/08/2018. Disponível em: http://room33.com.br/blog/2018/08/13/novas-tecnologias-mercado-imobiliario/. Acesso em: 11/05/2019.

EKIMA. ***Big Data*: tudo que você sempre quis saber sobre o tema!**. *Online*. 2018. Disponível em: http://www.bigdatabusiness.com.br/tudo-sobre-big-data/. Acesso em: 08/04/2019.

EXAME. **8 cursos para quem quer atuar como cientista de dados**. *Online*. Disponível em: https://exame.abril.com.br/tecnologia/8-cursos-para-quem-quer-atuar-como-cientista-de-dados/. Acesso em: 26/11/2018.

FAGUNDES, P.; MACEDO, D.; FREUND, G.; **A Produção Científica Sobre Qualidade de Dados em Big Data: Um Estudo na Base de Dados Web of Science**. RDBCI: Revista Digital De Biblioteconomia E Ciência Da Informação, 16(1), p. 194-210, 2017. Disponível em: https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/8650412. Acesso em: 31/05/2019.

FONSECA, J.; MARTINS, G. **Curso de Estatística.** São Paulo: ATLAS S.A., 1996.

INGARGIOLA, A.; **What is the Jupyter Notebook?.** *Online.* Disponível em: https://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/what_is_jupyter.html. Acesso em 03/06/2019

INTEL. **Guia de Planejamento:** Saiba mais sobre o Big Data. Intel Corporation, *online*, 2013. Disponível em: https://www.intel.com.br/content/dam/www/public/lar/br/pt/documents/articles/90318386-1-por.pdf. Acesso em: 31/05/2019.

LEE, A. **Why and How to Use Pandas with Large Data**. *Online.* 2018. Disponível em: https://towardsdatascience.com/why-and-how-to-use-pandas-with-large-data-9594dda2ea4c. Acesso em: 15/04/2019.

GALDINO, N. **Big Data: Ferramentas e Aplicabilidades**. Instituição de Ensino Superior Sant'Ana, 2019. Disponível em: https://www.aedb.br/seget/arquivos/artigos16/472427.pdf. Acesso em: 31/05/2019.

GUEDES, T.; ACORSI, C.; MARTINS, A.; JANEIRO, V. **Aprender Fazendo Estatística.** São Paulo: Universidade de São Paulo, 2019. Disponível em: http://www.each.usp.br/rvicente/Guedes_etal_Estatistica_Descritiva.pdf. Acesso em: 31/05/2019.

GONÇALVES, R.; MARIA, B. **Data Science: Ciência Orientada a Dados**. Universidade Estadual de Londrina (UEL). Inf. Inf., Londrina, v. 21, n. 2, p. 1 – 3, maio/ago., 2016. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/27929/20119. Acesso em: 31/05/2019.

MAGNUN. **A história do python**. *Online*. 08/10/2014. Disponível em: http://mindbending.org/pt/a-historia-do-python. Acesso em 23/01/2019.

MELO, F. **O erro mais caro da sua vida é não usar python ou sql pra arrumar toda aquela bagunça**. *Online*. 2018. Disponível em: https://universodosdados.com/2018/12/12/data-wrangling-python-ou-sql/. Acesso em: 08/04/2019.

NEOWAY. **Data Driven**: inteligência de negócios através do Big Data para todos. *Online*. 28/03/2019. Disponível em: https://www.neoway.com.br/data-driven-inteligencia-de-negocios-atraves-do-big-data-para-todos/. Acesso em: 31/05/2019.

NUNES, V. **Tipos de Gráficos Estatísticos.** *Online*. Disponível em: https://www.matematica.pt/util/resumos/tipos-graficos-estatisticos.php. Acesso em: 17/06/2019

PATENATE, M. **A importância da análise de dados para um negócio**. *Online*. 2018. Disponível em: https://www.escolaedti.com.br/a-importancia-da-analise-de-dados-para-um-negocio/. Acesso em: 23/11/2018.

PATIL, N. **Why Pandas is a Better Data Analysis Tool Than Excel.** *Online*. 2018. Disponível em: https://www.cbtnuggets.com/blog/2018/10/why-pandas-is-a-better-data-analysis-tool-than-excel/. Acesso em: 21/03/2019.

VIDAL,V. **Controle de pouso de veículo quadrotor auxiliado por Visão Computacional**. Trabalho de Conclusão de Curso de graduação de Engenharia Elétrica - Habilitação em Robótica e Automação Industrial da Universidade Federal de Juiz de Fora. 77 fls. 2016. Disponível em: http://www.ufjf.br/eletrica_automacao/files/2013/11/TCC-templateUFJF.pdf. Acesso em: 21/03/2019.

SCIELO. **Aplicação de adubo e corretivo após o corte da cana-planta utilizando técnicas geoestatísticas.** *Online.* Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-84782008000400011. Acesso em 22/03/2019

TIOBE. **TIOBE Index for June 2019.** *Online.* Disponível em: https://www.tiobe.com/tiobe-index/. Acesso em 17/06/2019

TOLEDO, G.; OVALLE, I. **Estatística Básica.** São Paulo: Atlas S.A, 2013

WASKOM, M. **Seaborn: Statistical Data Visualization**. *Online*. 2018. Disponível em: https://seaborn.pydata.org/. Acesso em: 16/04/2019.