Kelvin Carvalho Prado
Railson Martins da Mata
Ana Caroline Coutinho Carvalho

Análise da base de dados de turistas para o território brasileiro

Trabalho apresentado ao curso de Engenharia de Computação – IFMG – Campus Bambuí como requisito para obtenção de créditos no componente curricular Tópicos Especiais.

BAMBUÍ
2020

Professor(a): Jéssica

Análise da base de dados de turistas para o território brasileiro

BAMBUÍ
2020

1 INTRODUÇÃO

O turismo pode ser definido pelo movimento temporário de pessoas para destinos fora da sua localidade, sejam pequenas ou grandes viagens. As atividades realizadas nesses lugares causam satisfação às suas necessidades. Esse trabalho visa analisar as chegadas internacionais ao Brasil, nos anos de 2010 à 2019. Esse artigo irá descrever a análise da base de dados da Chegada de Turistas Internacionais ao Brasil, disponibilizada pelo site do Ministério do Turismo.

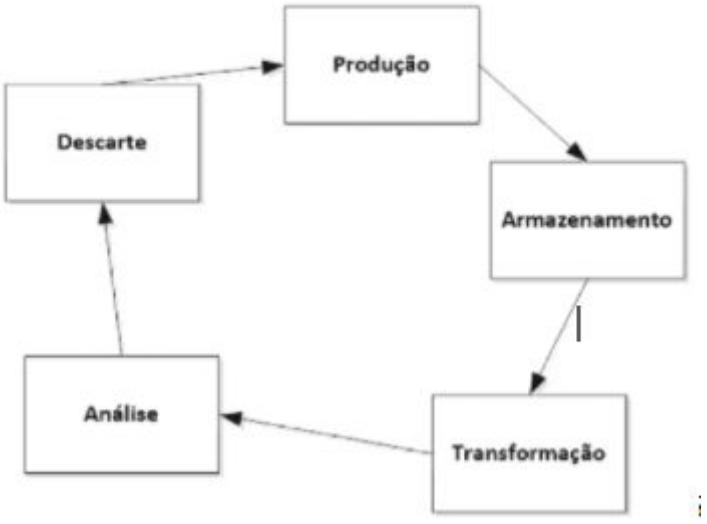
Dentre os tópicos disponibilizados pela base de dados do governo essa foi a mais interessante. Aplicando vários tipos de classificadores e visando identificar algum padrão de comportamento ou características dos turistas recém chegados ao Brasil. Segundo Coordenação-Geral de Estudos e Pesquisas (2016) a base de dados é formada por registros administrativos de migração coletados nos postos de fronteira e entregues ao Ministério do Turismo pelo Departamento da Polícia Federal.

Esses dados já estão de acordo com o marco teórico das Recomendações Internacionais, cujo foco é a padronização. Os atributos usados são: Continente, país de residência, UF, via de acesso, ano de chegada, mês de chegada e número de chegadas. Sendo esse arquivo no formato .csv (COORDENAÇÃO-GERAL DE ESTUDOS E PESQUISAS, 2016).

Para entender a *Data Science* é preciso saber o que é informação, dado e conhecimento. “O dado consiste em fatos coletados e armazenados. Informação é um dado analisado e com algum significado. O conhecimento é a informação interpretada, entendida e aplicada para um fim.” (AMARAL, 2016, p.03).

A *Data Science* é uma ciência nova, que é mal compreendida e controversa. Tratando como processos, modelos e tecnologias que estudam os dados durante todo o seu ciclo de vida, da produção ao descarte, Figura 1. Normalmente a *Data Science* é associada de forma equivocada apenas aos processos de análises dos dados, onde com o uso de estatísticas, *machine learning* ou a simples aplicação de um filtro se produz informação e conhecimento.

Figura 1: Ciclo de vida da *Data Science*



Fonte: Amaral (2016)

Big Data está associada a grandes volumes dados, sua definição formal é dada por um conjunto de 3 à 5 "Vs". Tendo como objetivo cada uma delas:

1. Volume - é uma grande quantidade de dados gerados a cada segundo, como por exemplo, troca de e-mails, transações bancárias, interações em redes sociais, registro de chamadas e tráfego de dados em linhas telefônicas.
2. Velocidade - é a velocidade com que os dados são criados. A análise desse grande volume de dados deve ser realizada com a mesma rapidez, para que esses dados sejam trabalhados, atualizados e expandidos com eficácia.
3. Variedade - Os dados podem aparecer de várias formas, em imagens, vídeos, áudios, documentos, planilhas, etc. É essencial entender essa variedade e identificar como devem ser analisados e armazenados.
4. Veracidade - Um dos pontos mais importantes de qualquer informação é que ela seja verdadeira. Com as análises e estatísticas de grandes volumes de dados é possível compensar as informações incorretas.
5. Valor - Quanto maior a riqueza de dados, o mais importante, é que empresas entrem no negócio do *Big Data*, lembrando dos custos e benefícios e tentar agregar valor ao que se está fazendo.

Big data é um conjunto de dados maior e mais complexo, especialmente de novas fontes de dados. Esses conjuntos de dados são tão volumosos que o software tradicional de processamento de dados simplesmente não consegue gerenciá-los. No entanto, esses grandes volumes de dados podem ser usados para resolver problemas de negócios que você não conseguiria resolver antes.

Segundo (Amaral, 2016) para o mundo empresarial, *Big Data* traz boas e más notícias, a boa notícia é que o *Big Data* oferece muitas oportunidades. Estas oportunidades virão de duas formas: vantagem competitiva ou criação de produtos e/ou serviços orientados a dados.

Já a má notícia são as empresas que não souberem usar *Big Data* vão desaparecer, engolidas pelas concorrentes que serão mais eficientes, com menos custos, com produtos com mais qualidade e clientes mais satisfeitos.

2 METODOLOGIA

2.1 Organização, limpeza e preparação dos dados

No estudo, foram utilizadas nove *datasets* referentes ao mesmo tema. Os *datasets* variam dos anos 2010 à 2019, além da biblioteca pandas (para manipulação dos dados) e a matplotlib para geração dos gráficos e a linguagem de programação *Python*.

Em primeiro momento, foram selecionadas as bases de dados realizando sua leitura por uma função disponibilizada pelo pandas, entretanto algumas bases de dados estavam com o formato diferente do que a biblioteca pandas exige para se trabalhar, formato esse denominado UTF-8, logo, ao pressionar CTRL + S, no editor de texto *Sublime*, o mesmo mostrava qual o formato a base de dados se encontrava, em seguida teve a renomeação dos atributos da base de dados para os listados a seguir:

- 'cod continente':'Ordem continente';
- 'cod pais': 'Ordem país';
- 'ano':'Ano';
- 'cod via':'Ordem via de acesso';
- 'cod mes':'Ordem mês';
- 'cod uf':'Ordem UF';
- 'Via':'Via de acesso'

Foi feito isso, pois posteriormente iria ser feito a concatenação de todas as bases de dados, sem isso, teria atributos repetidos, aumentando a dimensão da base de dados, e assim prejudicando o desempenho de qualquer algoritmo subsequente a ela aplicada, além disso, após a concatenação, um atributo com o nome *Unnamed*: 12, que em tese não fazia parte daquela base de dados, foi removido, após realizado às devidas avaliações no mesmo, como o descobrimento de que não tinham valores em suas linhas.

Com esse processo de limpeza e preparação dos dados, utilizando técnicas vistas durante o aprendizado. Foi usado a IDE do *Visual Studio Code*, com a instalação das referidas bibliotecas.

2.2 Descrição dos procedimentos realizados e análise sobre os resultados obtidos

Como citado por Santana (2019) em sua publicação, a melhor maneira de entender um problema de *Data Science* é realizando perguntas, neste trabalho, foram realizadas 3 perguntas, para que os resultados fossem analisados, sendo elas:

- Quais meses houveram mais visitas ao Brasil?
- Quais os estados mais visitados no Brasil?
- Quais os tipos de acesso mais foram usados?

Com isso em mente, foi desenvolvido às técnicas para a elaboração dos gráficos e assim realizar suas análises. Para a geração dos gráficos, foram usados métodos disponibilizados pela biblioteca pandas, que ajuda muito o desenvolvedor, o Código 1 a seguir, mostra a extração da quantidade de pessoas que chegaram na região da amazônia para a referida base de dados.

```
1- Amazonas = dados[(dados['UF'] == 'Amazonas')]
2- Amazonas = Amazonas['Chegadas']
3- Amazonas = Amazonas.sum(axis=0)
```

O Item 1 do código anterior, faz uma busca na base dados, retornando todos as linhas que possuem a UF Amazonas, com isso, como era necessário apenas saber o número de pessoas que chegaram naquela região, foi realizado uma filtragem, como mostra o Item 2, retornando assim uma tabela, que mostrava o id e o número de pessoas que chegaram, entretanto, o id não era necessário, logo, foi realizado a codificação do Item 3, onde na tabela selecionava apenas o eixo que representa a quantidade de pessoas que chegaram ao Brasil, realizando sua soma, com a função sum, retornando um *float* com o valor total. Fazendo esse procedimento para todos os estados que estavam na base de dados, foi possível gerar os gráficos como mostra o Código 2.

```
1- estados = ['Amaz' ,'Ba' ,'Ce', 'DF' ,'MG',
   'OUF', 'Pará', 'Per' ,'Paraná',
   'Rio', 'RioNorte' ,'RioSul',
   'San', 'SP', 'MGS', 'Acr', 'Amap',
   'Ro']
2- valores = [Amazonas, Bahia, Ceara, DF, MG, OUF, Para, Pernambuco, Parana, Rio, RioNorte, RioSul, SantaCatarina, SaoPaulo, MGS, Acre, Amapa, Roraima]
3- plotar.plot(estados,valores)
4- plotar.show()
```

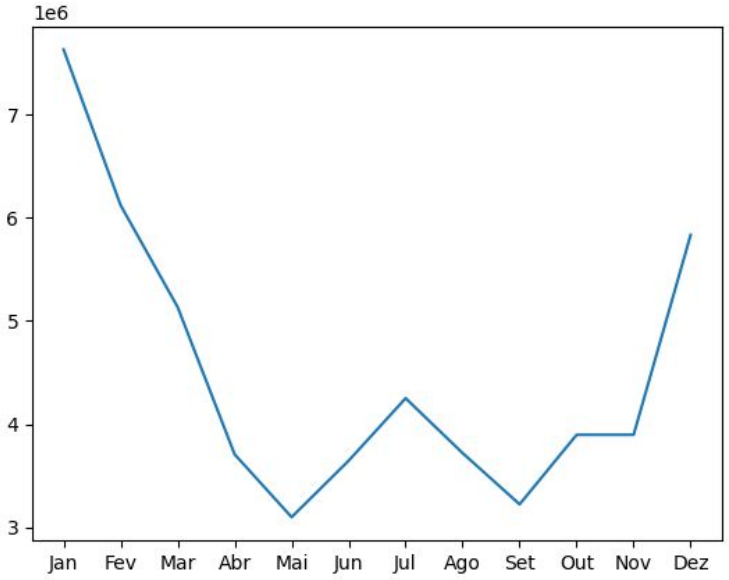
Para geração dos gráficos, como mostrado no Código 2, no Item 1, tem-se uma lista, contendo todos os nomes que apareceram no eixo X do gráfico, já para o eixo Y, se tem o Item 2, também uma lista, possuindo todos os valores da quantidade total de pessoas que vieram para determinado estado, no Item 3, a função plot, cria o gráfico, entretanto, para a efetivação dessa plotagem, e sua visualização na tela é usado a função *show()*. Assim como resultados desses procedimentos, foi-se possível a visualização dos dados de forma detalhada.

3 RESULTADOS

3.1 Geração dos gráficos e análise dos dados

Com o que foi dito e feito no tópico anterior, foram realizados as plotagens dos gráficos para responder as 3 perguntas feitas conforme citado por Santana (2019), nas Figuras 1, 2, 3, podem-ser vistos os gráficos gerados pela biblioteca matplotlib.

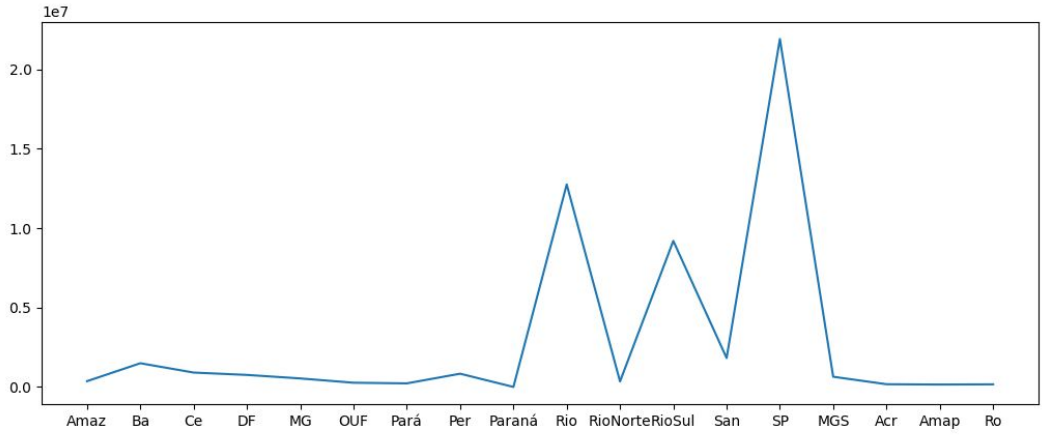
Figura 1: Quantidade de pessoas em relação aos meses



Fonte - Próprios autores (2020)

Olhando o comportamento do gráfico anterior, pode se perceber que entre os anos de 2010 a 2019, o mês que os turistas gostaram mais de vir ao Brasil foi o de janeiro, por ser um período de férias passando de 7 milhões de turistas, com o pior mês sendo o de maio com um pouco mais de 3 milhões de pessoas.

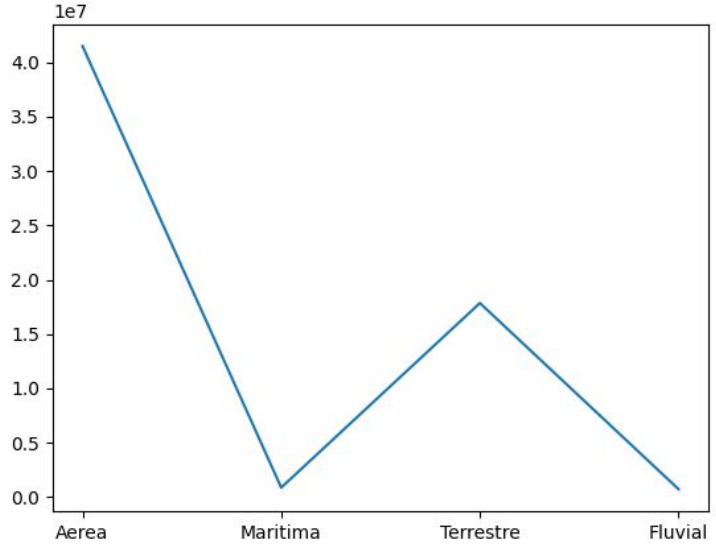
Figura 2: Quantidade de pessoas em relação aos estados



Fonte: Próprios autores (2020)

Olhando o comportamento do gráfico anterior, o estado que os turistas mais gostam de visitar é o de São Paulo com mais de 20 milhões de visitas, no referido período, Rio de Janeiro com um pouco menos de 15 milhões de visitas sequido do Rio Grande do Sul, os outros estados não são tão visitados como esses, um exemplo é o de Minas Gerais com menos de 2 milhões de visitas.

Figura 3 - Quantidade de pessoas em relação a via de acesso



Fonte: Próprios autores (2020)

Como mostrado na Figura 3, a forma de acesso mais usada pelos turistas é a Aérea com mais de 40 milhões de pessoas a usando, seguido pela terrestre com aproximadamente 20 milhões.

Com esses resultados em mãos, empresas de hotelaria, e de meios de transportes, possuem dados significativos em mãos, como por exemplo, o melhor estado para se construir novos hotéis é São Paulo e Rio de Janeiro, em detrimento a Minas Gerais, como um dos piores, para essa área. Para empresas de transporte a área que pode dar mais lucratividade é a Aérea, por ter uma quantidade de turistas que usam essa via de acesso enorme.

4 REFERÊNCIAS

AMARAL, Fernando. Introdução à Ciência de Dados: mineração de dados e big data. Alta Books Editora, 2016. Disponível em:<https://cutt.ly/ggkVrxj> Acesso em: 19 out 2020.

COORDENAÇÃO-GERAL DE ESTUDOS E PESQUISAS. **Chegada de Turistas Internacionais**. [*S. l.*], 17 jun. 2016. Disponível em: http://dados.turismo.gov.br/index.php/chegada-de-turistas-internacionais. Acesso em: 19 out. 2020.

GRUS, J**. Data Science do Zero: Primeiras Regras com o Python**. Rio de Janeiro: Alta Books, 2016.

REDAÇÃO. **Big Data: os cinco Vs que todo mundo deveria saber**. Canal Tech. Disponivel em: <https://canaltech.com.br/big-data/Big-Data-os-cinco-Vs-que-todo-mundo-deveria-saber/>. Acesso em: 19 Outubro 2020.

SANTANA, FILIPE. **Guia passo a passo de como um projeto de Data Science é desenvolvido**. [*S. l.*], 14 jul. 2019. Disponível em: https://minerandodados.com.br/guia-passo-a-passo-de-como-um-projeto-de-data-science-e-desenvolvido/. Acesso em: 19 out. 2020.

TAKE, TAKE. **Site da Take**, 18 Janeiro 2017. Disponível em: <https://cutt.ly/1gkC6oy>. Acesso em: 19 out 2020.