Curso: Engenharia de Computação
Aluno: Alberto da Silva Félix
Disciplina: Probabilidade e Estatística Aplicada à Computação
Professor: Paulo Ribeiro Lins Júnior

Projeto 1 Estatística

Campina Grande, PB, 07/03/2017

Resumo

Este relatório contém os resultados e interpretações feitas a partir da análise estatística dos dados sobre buscas no google relacionadas a gripe no Brasil, nos estados do Ceará, Minas Gerais, Paraná e Rio de Janeiro, e os valores nacionais no intervalo de 22 de Janeiro de 2006 à 09 de Agosto de 2015.

Toda análise foi desenvolvida utilizando as bibliotecas da linguagem de programação Python na versão 3.5.

Sumário

Introdução

Medidas Resumo

Gráficos

Conclusão

Referências

1. Introdução

Em linhas gerais esse projeto tem os seguintes propósitos: aplicação prática do assunto de estatística descritiva que foi ministrado em sala aula, tendo em vista a análise de um número muito grande de dados utilizando as ferramentas computacionais para realização dos cálculos e plotagem dos gráficos.

2. Medidas Resumo

a. Medidas de posição

Localidade	**Média**	**Moda**	**Mediana**
Ceará	161,83	122 148	153
Minas Gerais	218,73	128	208
Paraná	196,71	181	183
Rio de Janeiro	209,10	260	204
Brasil	199,41	149 193 196	192

Tabela com os valores da medidas de posição para cada localidade

As medidas de posição, como o nome está indicando, são medidas que informam sobre a posição típica dos dados.

A média é calculada somando as n observações e dividindo por n, ela é influenciada por valores discrepantes(outliers), assim ela não é recomendada para a análise de uma grande quantidade de dados, vemos em nossa tabela a diferença entre a média e a mediana que é o número central de buscas, a média é maior em todos os casos, não nos informando com precisão a centralidade dos valores.

A moda é(são) o(s) valor(es) que mais se repetem no conjunto dos dados, vemos que na análise para o Brasil obtemos 3 modas, logo ele é trimodal.

A mediana é o valor que deixa 50% das observações acima e 50% abaixo dela, por exemplo para o estado de Minas Gerais, 50% dos valores estão acima de 208 e os outros 50% ficam abaixo desse valor, logo o 208 é o valor central de buscas relacionadas a gripe para esse estado.

b. Medidas de dispersão

Localidade	Amplitude	Desvio Médio Absoluto	Variância	Desvio Padrão
Ceará	258	39,65	2388,36	48,87
Minas Gerais	377	63,87	6268,96	79,18
Paraná	494	61,36	6209,70	78,80
Rio de Janeiro	281	49,16	3522,18	59,35
Brasil	343	52,75	4326,28	65,77

Tabela com os valores das medidas de dispersão para cada localidade

As medidas de dispersão medem a variabilidade do dados analisados.

A amplitude é a diferença entre o maior e o menor valor da classe, vemos que para o estado do Paraná temos a maior amplitude calculada com um diferença de 494 pesquisas entre o dia que foi realizada menos pesquisas e o dia que realizaram mais pesquisas.

O desvio médio absoluto nos informam sobre a dispersão dos dados, em nossa tabela vemos que o estado com a maior dispersão entre as pesquisas é o de Minas Gerais, com 63,87. Logo esse estado possui mais valores discrepantes que os demais.

A variância é a média dos desvios quadráticos, por isso ela possui valores tão grandes, por ela é que calculamos o desvio padrão que segue o mesmo princípio, quanto maior for o desvio padrão é maior a dispersão entre os dados, como vemos em nossa tabela, existe dispersão em todas as localidades analisadas.

c. Separatrizes

Localidade	1º Quartil	2º Quartil	3º Quartil
Ceará	122	153	192
Minas Gerais	157	208	264
Paraná	138,50	183	239
Rio de Janeiro	161	204	251,50
Brasil	150	192	239

Tabela com valores dos quartis para cada localidade

Temos que o 1ºQuartil separa os dados em 25% abaixo e 75% acima do valor do quartil, e o 3ºQuartil separa 25% acima e 75% abaixo, daí concluímos que entre o 1º e o 3º Quartil estão 50% dos dados da pesquisa, o segundo quartil é o valor da mediana calculada anteriormente.

Podemos dizer que 50% da quantidade de acessos por dia no google sobre a gripe no Brasil está entre 150 e 239 acessos, para os valores nacionais.

3. Gráficos

a. Histograma

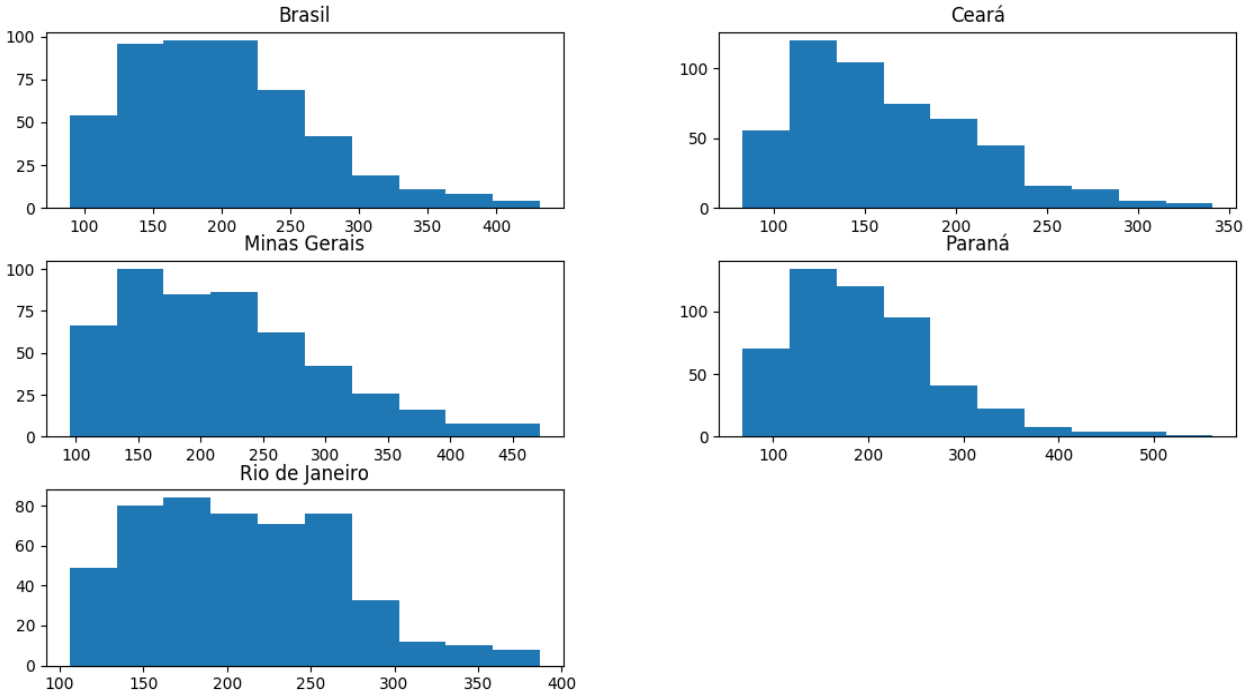

Figura 1. Histograma para cada localidade

Os dados que retiramos do histograma são os seguintes:
A largura da barra equivale a amplitude de cada classe; a altura da barra equivale a frequência das classes e o centro da barra é onde se encontra o ponto médio de cada uma das classes.

Observando os nossos gráficos, concluímos que o estado do Rio de Janeiro apresenta uma uniformidade maior para os dados abaixo da mediana que é 204, já para os valores acima da mesma a variação é maior. Os demais gráficos seguem um padrão semelhante.

b. BoxPlot

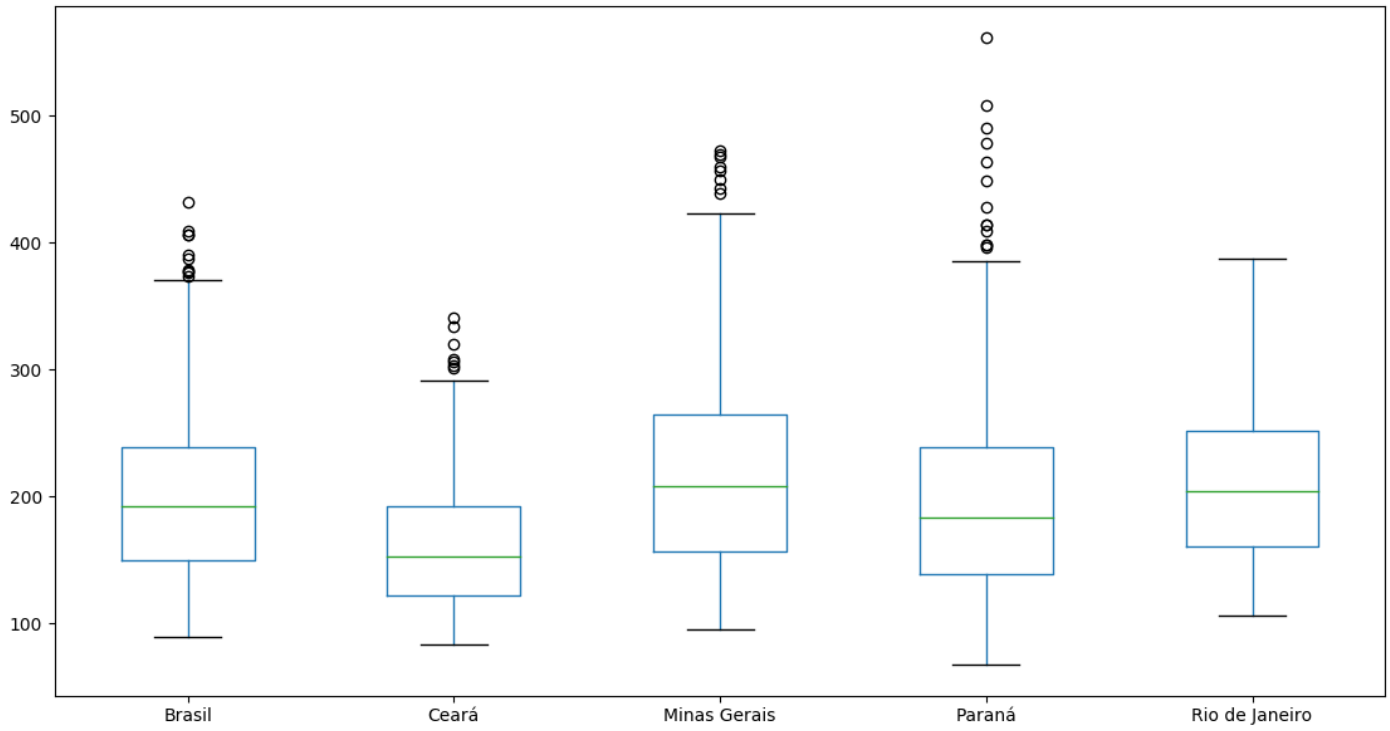

Figura 2. Boxplot para cada localidade

O boxplot é construído a partir dos valores dos quartis, ilustra os principais aspectos da distribuição e é também muito útil na comparação de distribuições.

Dos nossos boxplots vemos que apenas o estado do Rio de Janeiro não existem valores discrepantes(outliers), ou seja, todos os dados da pesquisa para o Rio de Janeiro estão entre as juntas que são os valores de mínimo e máximo do conjunto de dados formado pelos valores não discrepantes. Todos os demais locais estudados neste projeto possuem outliers superiores, isto é, acima da junta superior.

4. Conclusão

Ao término deste projeto alguns pontos merecem destaque, o primeiro é que foi muito proveitoso para o aprendizado da Estatística Descritiva, pois foi necessário a revisão dos conceitos para descrição do projeto. O segundo ponto de destaque foi a utilização das bibliotecas python para os cálculos e para a plotagem dos gráficos, onde o manuseio dos dados ficam bem mais simples e fáceis de serem manipulados. Em terceiro lugar está a experiência de unir a teoria a prática, facilitando o aprendizado do conteúdo proposto.

5. Referências

[1] http://pandas.pydata.org/pandas-docs/stable/
[2] http://matplotlib.org/index.html