

Relatório 5 - Prática: Estatística p/ Aprendizado de Máquina (I)

João Pedro Gomes

1. Introdução

O objetivo do card 5 Prática: Estatística p/ Aprendizado de Máquina (I) é de assistir 13 vídeos aulas e após isso realizar uma atividade prática com os conhecimentos obtidos em cada uma e depois realizar este relatório pra explicar oque foi feito e oque foi aprendido.

2. Desenvolvimento

Esse card tem 13 videos diferentes pra assistir, então irei explicando um por um

1) Types of data:

Ele começa explicando que existem vários tipos de dados diferentes e eles são,

Númericos, Categóricos e ordinais, segue o insight visual sobre eles

Numericos		Categoricos	Ordinal
Discreto	Continuo	Onde você divide dados em categorias, tipo na CNH, carteira do tipo A, B, C e etc mas a A não é melhor que a B são só categorias sem ordem	Aqui já pode ter representação matemática ou não, porque segue uma forma lógica e hierarquica, tipo uma nota pra um filme se um filme é nota 1 ele é ruim mas se for um nota 4 é um filme bom sem números daria pra exemplificar com ta manho de camisas, PPÉ menor que M e por assim vai.
Um número inteiro que não tem quebrado tipo 5 ou 10 mas nunca 5,4 ou 10,2	Um número quebrado que pode ter infinitas possibilidades tipo 3,261823723612312 ou 3,3 pra simplificar		

2) Mean, Median, Mode:

O segundo fala de média moda e mediana, onde a média é a soma de todos os numero de um conjunto e dividida pela quantidade de elementos somados, a mediana é o número em um conjunto que divide ele ao meio se for um conjunto ímpar de números ele tem a mediana correta mas se for par tem que pegar os dois numero do meio somar e dividir por 2 para obter a mediana do conjunto e a moda é o número que mais aparece no conjunto. E se o conjunto não tem um número com mais freequencia doque outros esse conjunto é amodal, não tem moda.

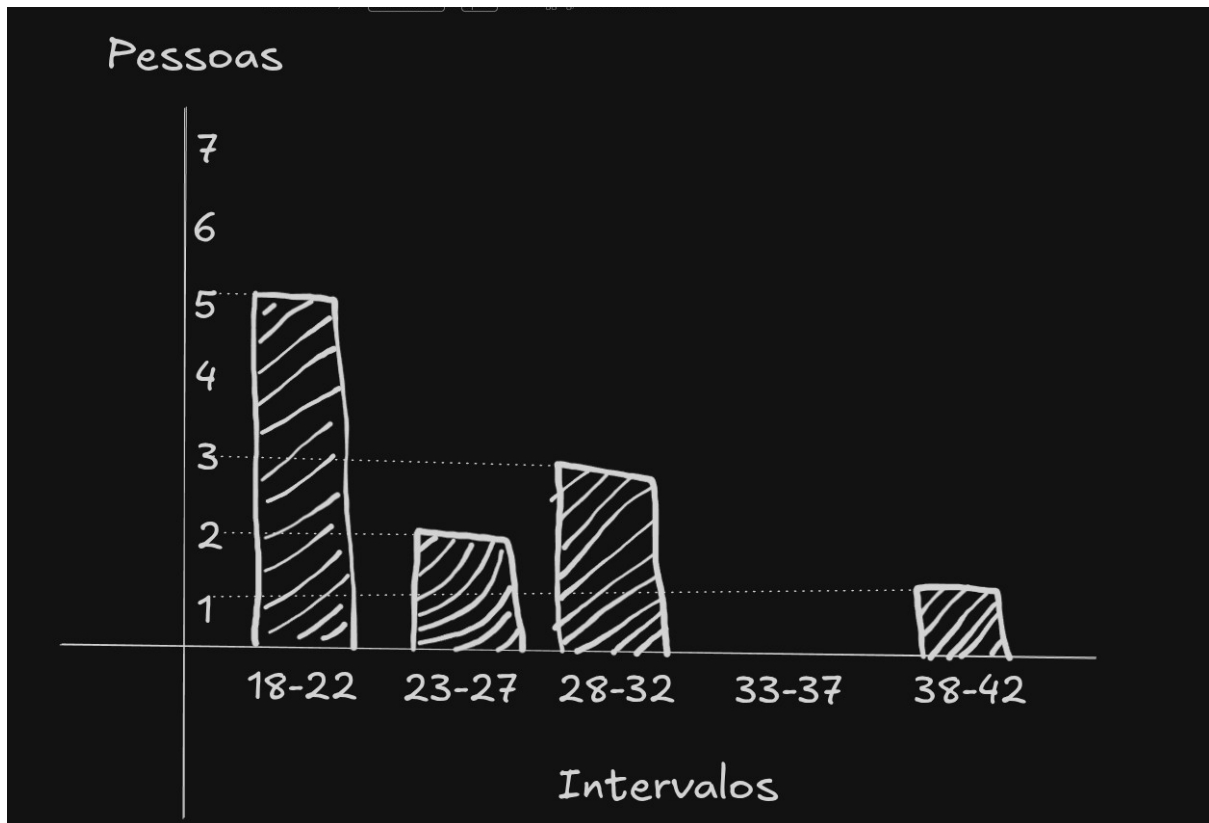
C = 9, 10, 20, 32, 9		
Média	Mediana	Moda
<p>Para a média a gente soma todos os elementos e divide pela quantidade dos mesmos</p> $\bar{x} = \frac{80}{5} = 16$ <p>A média de C é 16</p>	<p>A mediana é o número que divide o conjunto no meio no nosso exemplo é o 20</p>	<p>É o número que mais aparece no conjunto aqui é o 9.</p>

3) Using mean, median, mode in python:

Aqui ele aborda a parte prática no python usando numpy, matplotlib.pyplot e o scipy, vou focar nessas duas ultimas bibliotecas pois a de numpy eu já expliquei em relatórios passados.

A parte prática dela é bem curta, no pyplot ele ensinar a fazer histogramas que é um grafico que separa dados por intervalos e mostra a quantidade de dados nesses intervalos, exemplo:

Vamos pegar a idade dos fãs dos Beatles no spotify, 18, 19, 20, 21, 22, 25, 27, 30, 32, 31 e 40, agora vamos separar os intervalos de 18 – 22, 23 – 27, 28 – 32, 33 – 37, 38 – 42. Ai jogamos pra um gráfico assim:



Vamos colocar as idades sendo a variavel Idade, no python usando a biblioteca você deveria colocar `plt.hist(idade, a quantidade de intervalos)` e iria fazer o histograma.

E com o scipy ele ensina a pegar o numero com mais frequencia e quantas vezes ele aparece com a funcao `mode()`, supondo que temos um array `[20, 30, 30, 32, 30, 32]`, a função `mode` pega o 30 e fala que ele aparece 3 vezes.

4) Variation and Standart Deviation:

Nesse ele fala de variancia e desvio padrão.

A variancia mede se os dados estão muito espalhados ou não, tipo, temos 2 grupos no primeiro grupo todas as pessoas 20 anos variancia 0 todos tem a mesma idade, no segundo um tem 90 anos e o outro 10 outro tem 2 meses de idade, varia demais e é isso que ela busca calcular.

Pra calcular a variancia pega a média dos dados e depois subtrai cada dado por ela e eleva ao quadrado e faça a média deles também e a vârianca é essa.

Existem a variancia populacional e a amostral, a populacional é quando você tem todos os dados do conjunto para fazer o calculo ai você divide por N , se você tem só uma amostra dos dados o calculo muda para a divisão de $N - 1$, vamos supor que você quer a variancia de peso dos brasileiros, você não consegue ter todos os dados de todos os brasileiros ai você trabalha com a amostra, o $n - 1$ é uma correção pra aumentar a variancia pois uma mostra os dados tendem a ter uma variancia menor.

O desvio padrão é a raiz quadrada da variancia que joga a variancia pra mesma unidade de dados, por exemplo, a variancia é 25 o DP é 5 é uma forma mais simples de ver.

No python ele não ensina muitas coisas novas, ele mexe com numpy e pyplot, e ensina que pra pegar variancia e dp de um conjunto de dados basta usar `.std()` para dp e `.var()` para variancia.

5) Probability density function: probability mass function:

Nesse ele fala sobre a Função de densidade probabilidade e Função de massa de probabilidade que são:

Função de densidade probabilidade: Ela dá a probabilidade de um dado estar entre um intervalo. Tipo se você quer saber a probabilidade de alguém ter exatamente 1,732432432 metros isso é praticamente impossível aí entra essa função.

Função de massa de probabilidade: Essa já é usada pra números inteiros, não tem um meio termo, quantidade de filhos, números ao lançar um dado, etc..., ela já dá uma probabilidade exata.

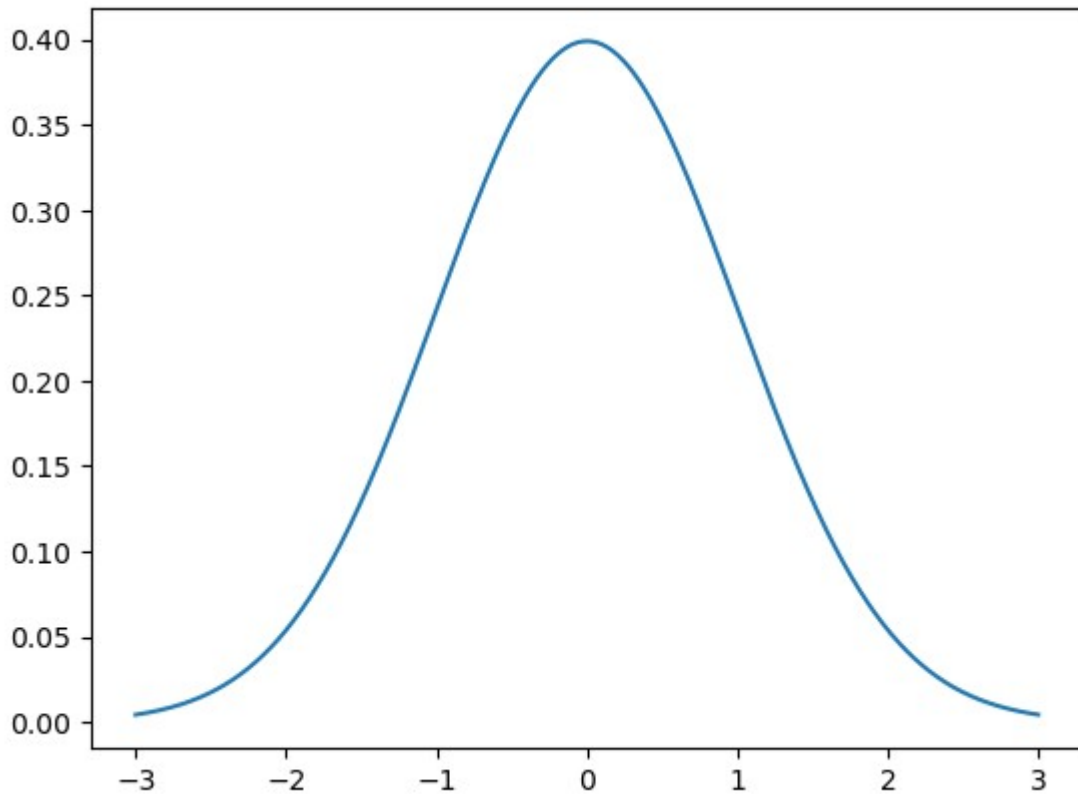
6) Common data distributions(Normal, Binomial, Poisson, etc):

Nesse ele vai pro python dar exemplos das distribuições de probabilidades:

Ele começa falando da distribuição uniforme, os dados tem a mesma chance de aparecer, no vídeo ele faz 100000 número entre -10 e 10 e cada um tem a mesma

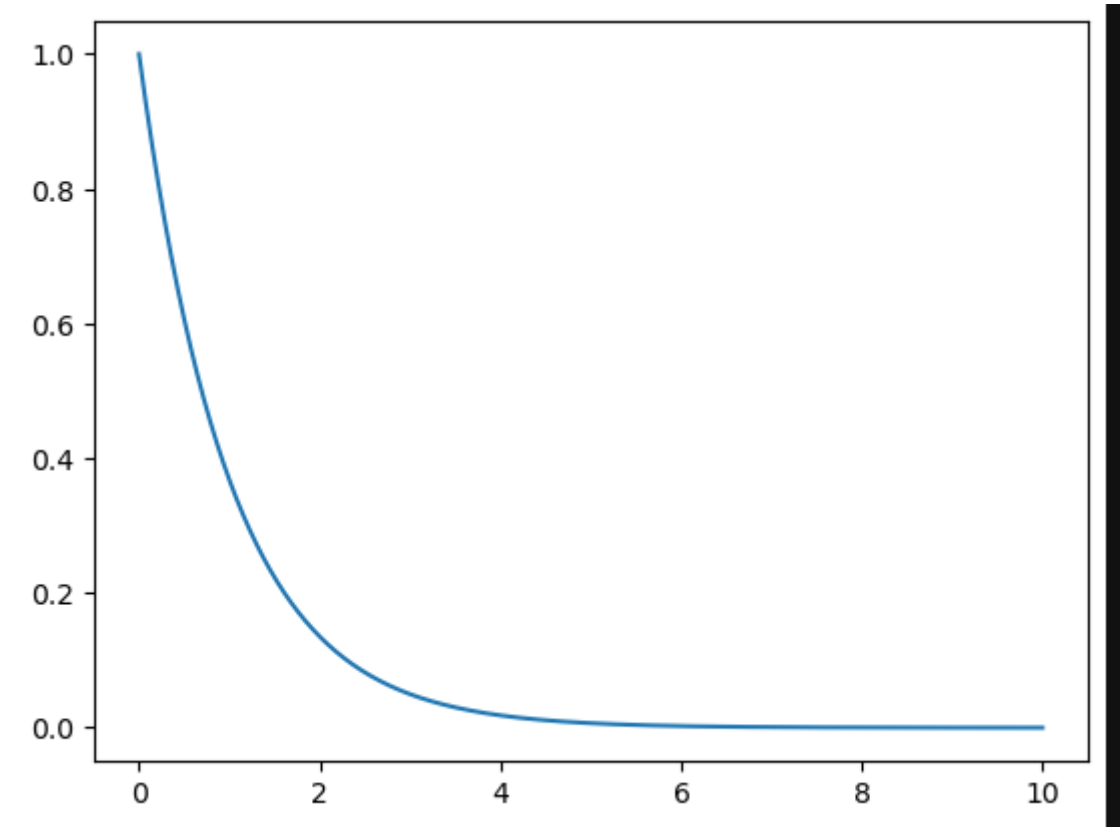
chance de aparecer, e joga pra um histograma onde as barras do histograma ficam todas quase da mesma altura.

A distribuição normal graficamente sempre terá um formato de sino, pois a chance de um numero aparecer fica na média, e quanto mais longe da média menos a chance, no video ele faz vários numeros de -3 a 3 pulando de 0.001 em 0.001 e usa o comando `plt.plot()` e joga pra um grafico de distribuição normal que fica assim:



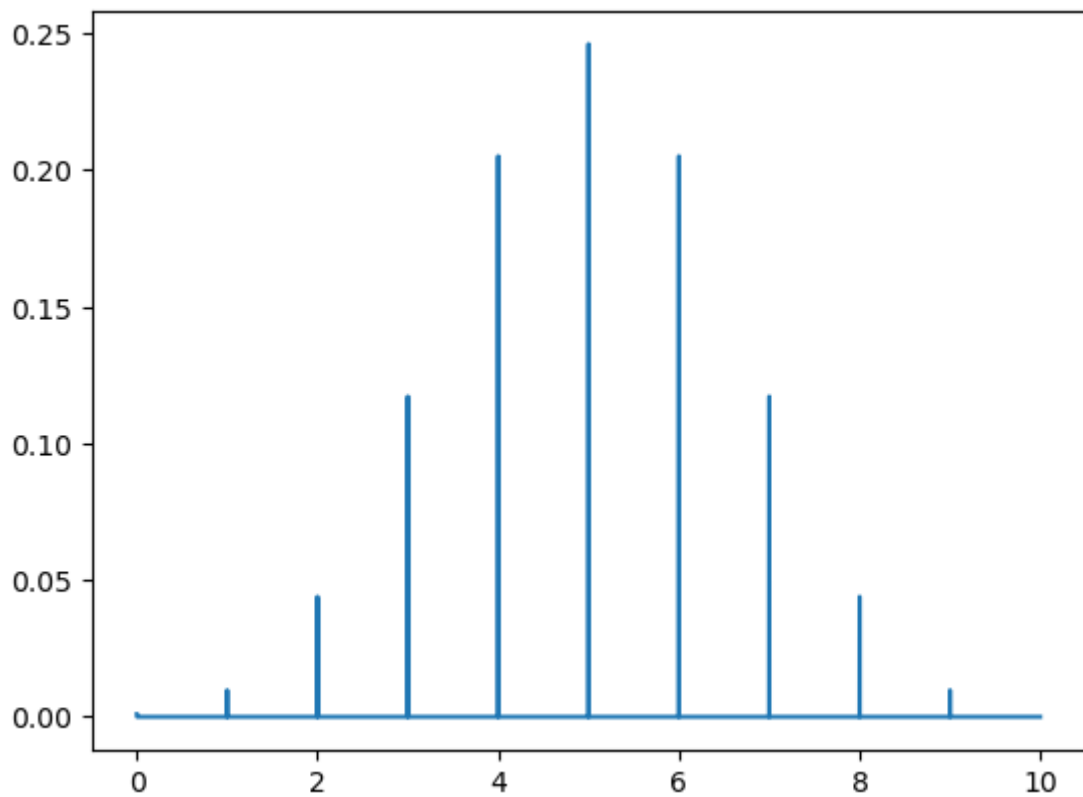
Pra exponencial se impor o `expon` do `scipy` para usar, ele pula de 0.001 em 0.001 de novo mas agora de 0 ate 10 e depois usa o `plt.plot()` e coloca graficamente numa funcao exponencial.

Uma função exponencial parece um L arredondado, quanto maior o numero menos a probabilidade dele aparecer fazendo ela ter esse formato.



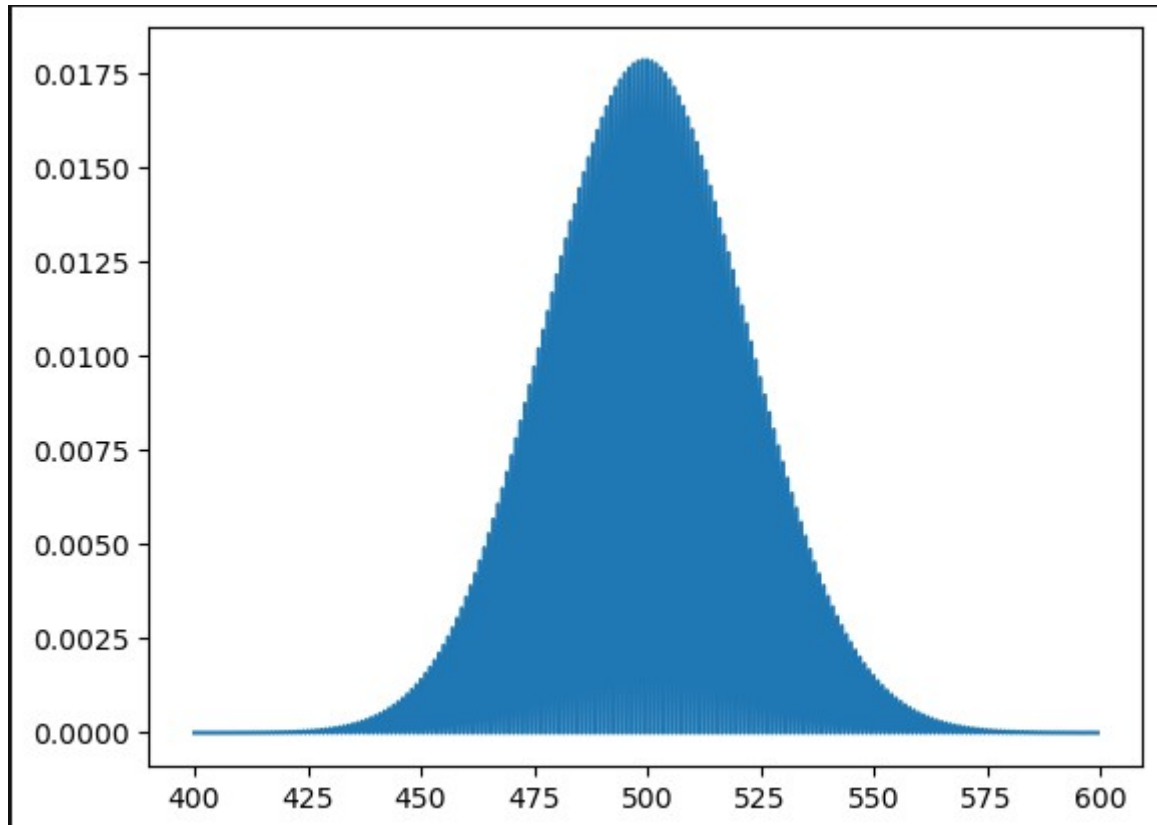
A binomial é usada pra resultados binário, 0 ou 1, sim ou não, então a chance de dar “sucesso” em algo é 50%, ela trabalha com números discretos nunca continuo, no código ele faz a predição de algo dar certo 10 vezes, vamos usar de exemplo que ‘dar

certo' é tirar cara em uma moeda jogando ela 10 vezes, a chance de você tirar 0 caras nas 10 vezes é quase impossível igual você tirar 10 caras nas 10 vezes também é quase impossível, mas tirar 5 caras é mais provável, por isso o formato do gráfico.



A poisson é a chance de algo acontecer dado as vezes que ela já aconteceu antes em uma situação por exemplo se um site recebe 500 visitas por dia qual a chance dele receber 550 amanhã.

No código ele define o numero medio esperado o μ como 500, cria valorer entre 400 e 600 e depois coloca em um gráfico de poisson que mostra as chances.



7) Percentiles and Momentos:

Fala sobre percentil e momentos.

Percentil é qual a porcentagem está abaixo desse valor, vamos supor que você é um dos 10 alunos de uma turma e você tirou 5 de 10, e a professora fala q seu percentil é de 90%, você foi melhor que 90% da turma na prova.

Os momentos explicam, onde os dados ficam, se estão muito espalhados, se estão tortos, se têm valores estranhos.

O primeiro momento é a média, ela pode enganar pois pode ter 5 valores, 4 deles serem baixos mas 1 valor pode ser extremamente grande e modificar a média geral.

O segundo é a variancia, que ve se eles estão muito espalhados ou seja, longe da média, ou bem juntos perto da média,

O terceiro é o skew ou assimetria, diz pra que o grafico vai, se tem muita gente ganhando pouco ou pouca gente ganhando muito, se o grafico for pra direita a assimetria é positiva.

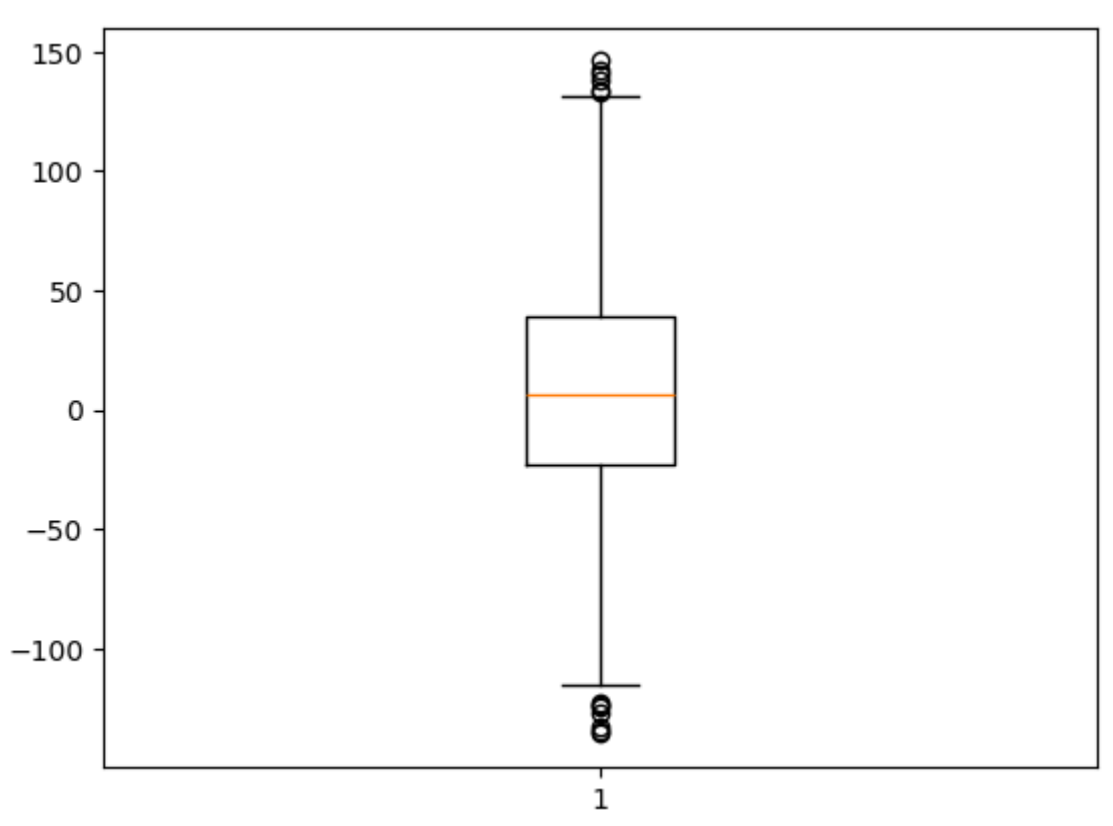
O quarto momento é a kurtosis ou curtoses que mede se a onda do gráfico é estreita ou larga comparada a uma distribuição normal.

8) Pyplot

O pyplot desenha graficos em python, todos os gráficos acima são feitos com ele, na aula 8 ele aprofunda muito mais o uso dessa biblioteca.

Com o pyplot tem como fazer 2 linhas dentro de um gráfico só, alterar formato e cor das linhas, fazer gráficos de pizza, barras, dispersão, gráficos de distribuições de probabilidades (como exemplo a cima deste relatório) e o boxplot.

Exemplo de bloxpot:

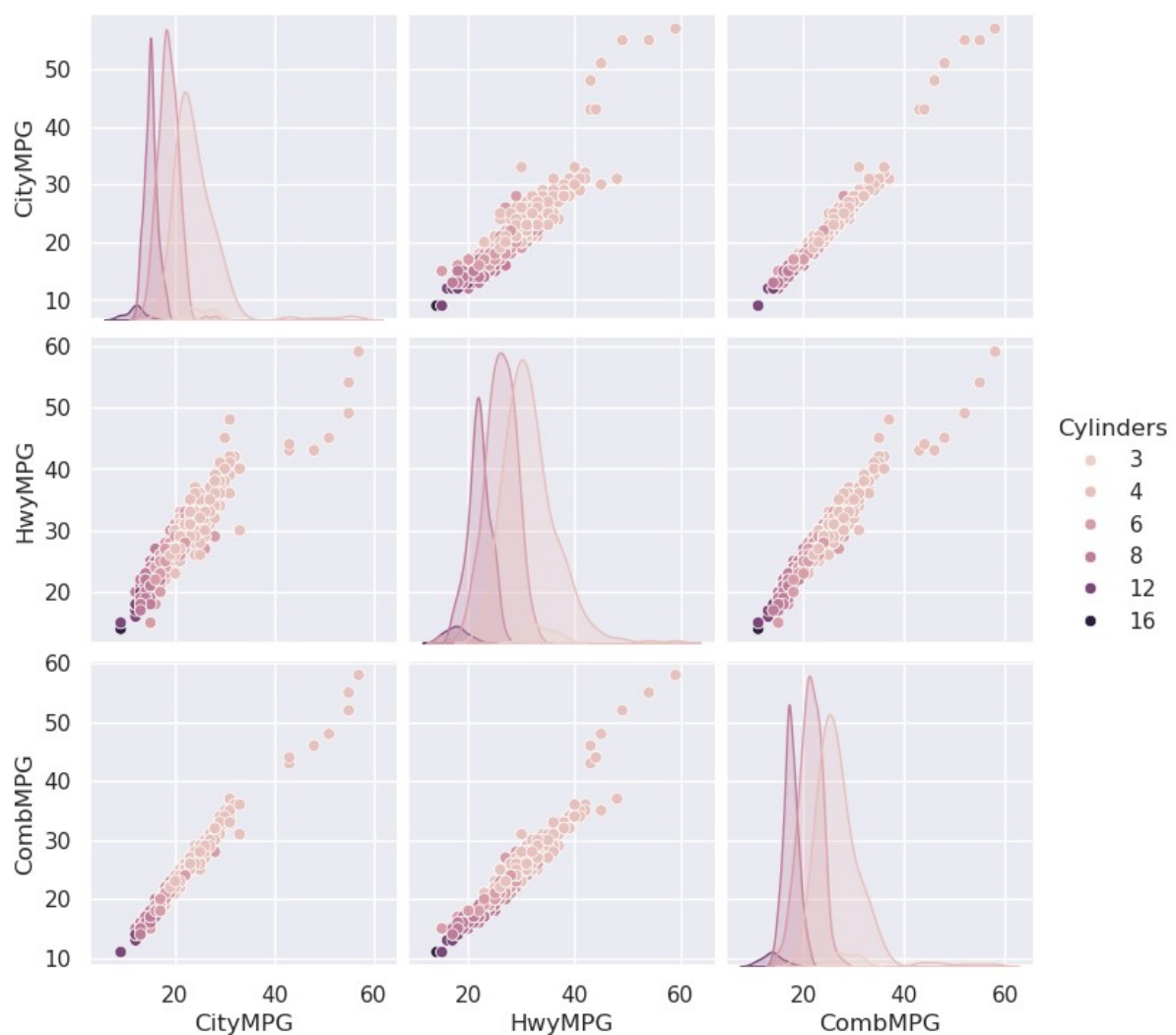


Esse é um dos gráficos um pouco mais chatos de entender de primeira, mas irei explicar, a caixa no meio representa 50% dos dados do conjunto, a linha inferior da caixa é chamada de Q1 é o primeiro quartil ela representa que 25% dos dados estão abaixo desse valor, a linha laranja no meio chamada de Q2 é a mediana a que separa os dados no meio e a linha superior da caixa Q3 é o terceiro quartil indica que 75% dos dados estão abaixo desse valor.

As linhas que saem de cima e de baixo do quadrado são os whiskers ela contam até onde os dados normais vão tudo que fica fora deles são outliers, dados fora do padrão que são essas bolinhas.

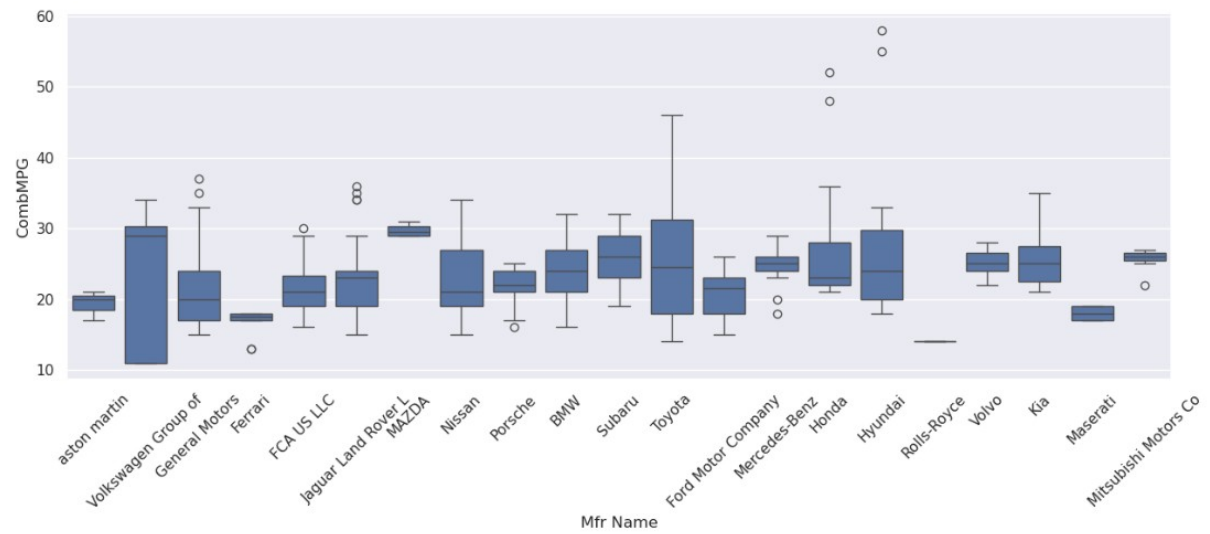
9) Seaborn

Com o seaborn os gráficos são muito mais bonitos visualmente e mais fáceis de escrever o código pra fazer, tem como fazer somente um gráfico, tem como fazer uma matriz de gráficos com vários tipos como exemplo:



Os dados do gráfico foram tirados de um csv que o professor manda ler, tem gráficos de dispersão, gráfico com linha de regressão, boxplot, gráfico com mapa de calor e gráficos de densidade.

Um exemplo de um boxplot:

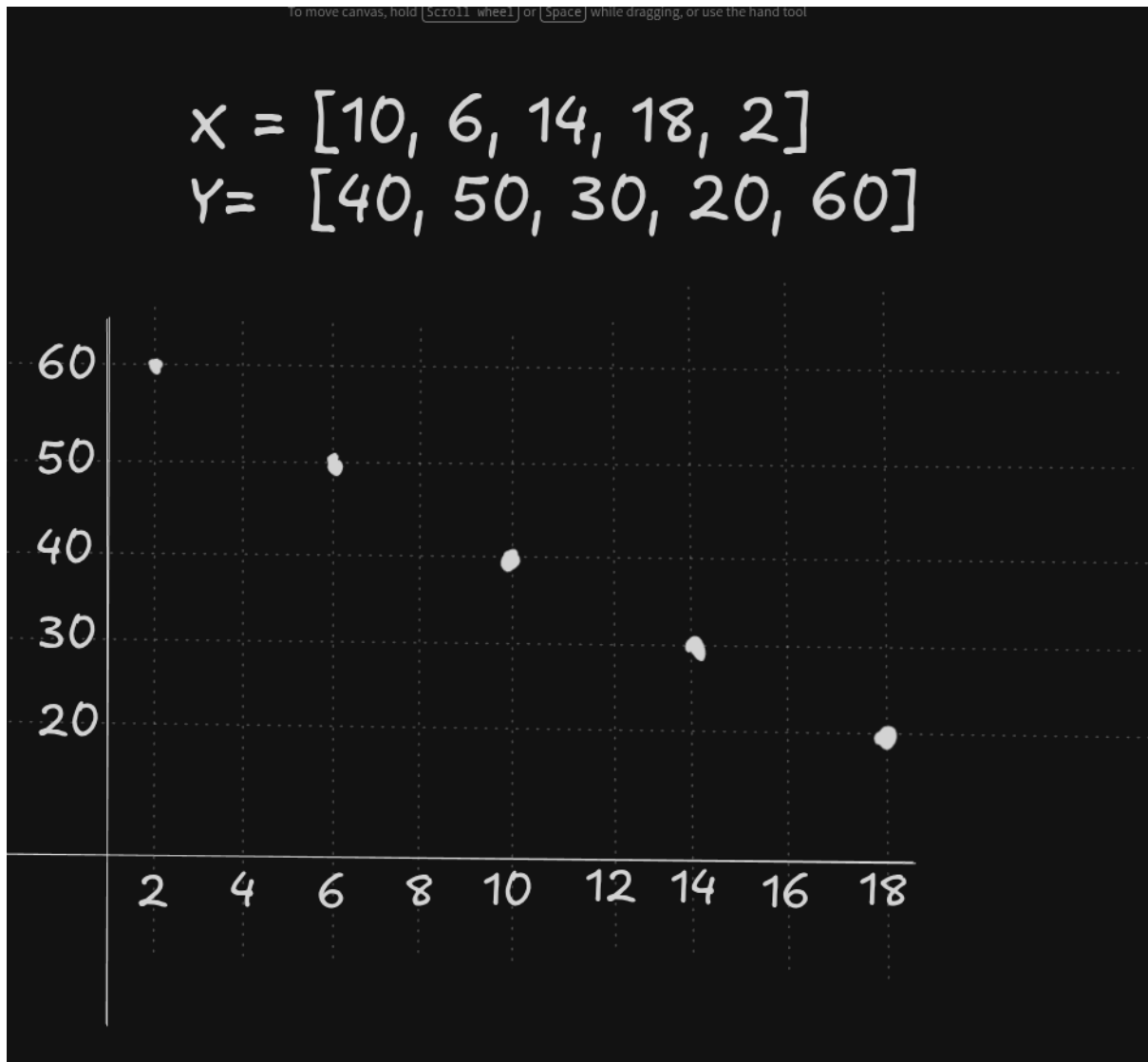


10) Covariance and Correlation:

Fala sobre correlação e covariância.

Correlção analisa dados e se pergunta, quando uma coisa muda a outra muda junto?, por exemplo se x aumenta y aumenta tambem?

Um exemplo visual:



3. Conclusão

A Conclusão deve ser o resumo mais essencial de tudo, em algo direto e final. Um parágrafo único de 5-10 linhas. Reforce o seu entendimento final sobre a tecnologia ou o assunto da aula.

Exemplo Prático:

"O Docker resolve o problema de dependências de ambiente, garantindo que o software funcione de forma idêntica em qualquer máquina. A principal conclusão é que o uso de imagens leves e a separação correta dos dados são cruciais para a eficiência e o desempenho em ambientes containerizados."

Motivos para Devolução do Card

Devolução Imediata:

- **Fora do Padrão Exemplificado:** O relatório não segue o formato obrigatório das três seções (Introdução, Desenvolvimento, Conclusão) ou ignora as regras de formatação (uso de tabelas, exclusão de código em prints, etc.).
- **Relatório Simples ou Falta de Insights do Aluno:** O conteúdo é superficial, foca apenas na execução de comandos e não demonstra a profundidade de análise e aprendizado esperada.
- **Falta de Conceitos Importantes:** O relatório não aborda ou falha em aplicar corretamente algum conceito importante apresentado no material da aula.
- **Não ter 1 Insight Visual Próprio:** Falta o insight visual obrigatório que deve ser de autoria própria.
- **Confuso ou Má Formatação:** A estrutura do texto é confusa ou a formatação prejudica a leitura.
- **Entrega fora do padrão especificado:** <número do card> - <nome do participante>.pdf

Desligamento do Bootcamp:

- **Plágio:** Qualquer trecho de texto ou código plagiado de terceiros sem a devida referência.
- **Uso de IA para Criação do Texto:** Utilização de ferramentas de Inteligência Artificial para gerar o conteúdo principal do relatório (texto).
- **Uso de IA para Criação dos Insights Visuais:** Utilização de ferramentas de Inteligência Artificial para gerar fluxogramas, diagramas ou mapas mentais.