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.

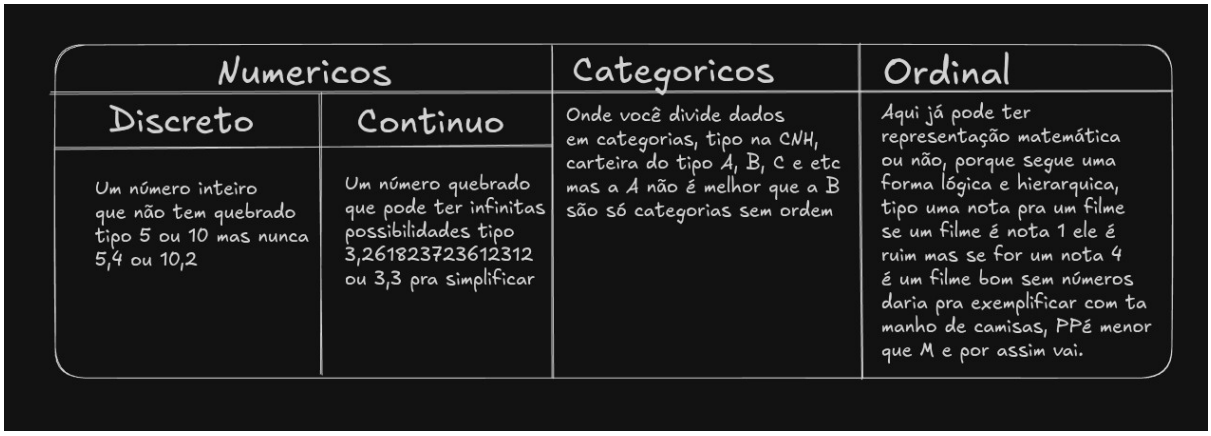
2. Desenvolvimento

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

1) Types of data:

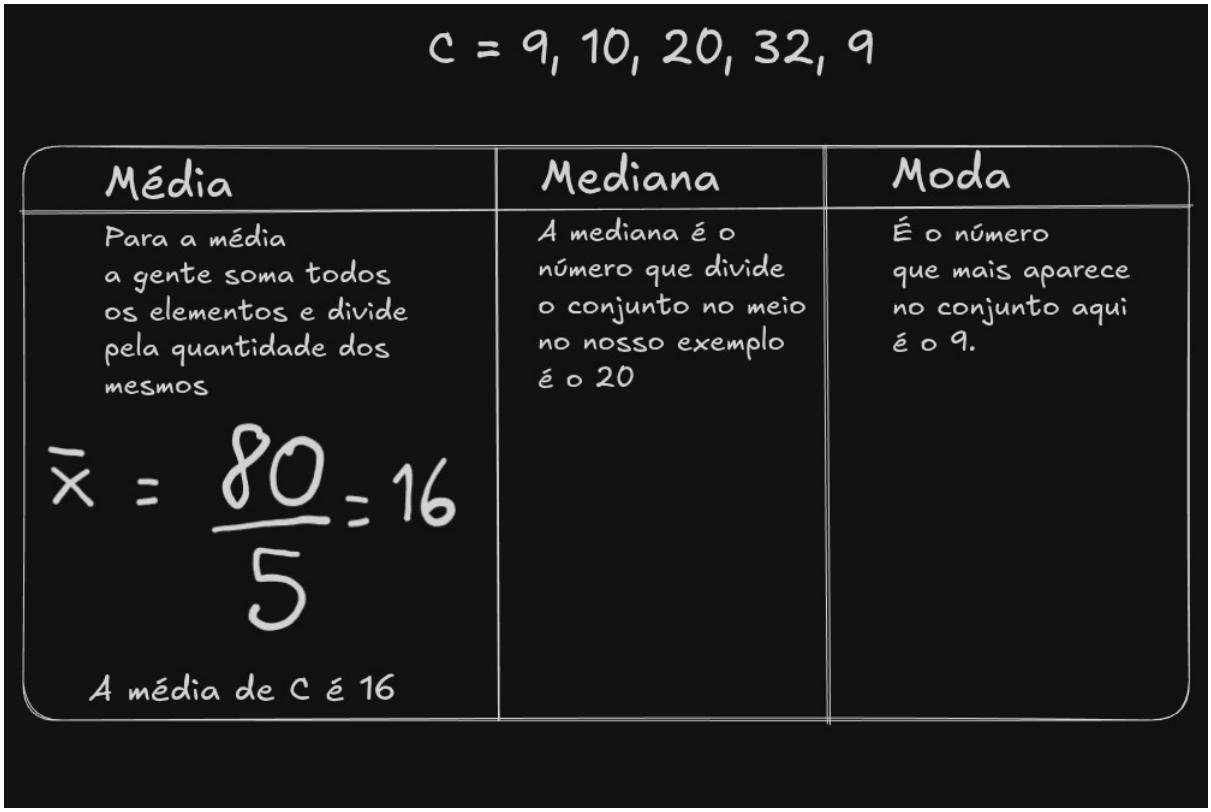
Ele começa explicando que existem vários tipos de dados diferentes e eles são,

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



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.

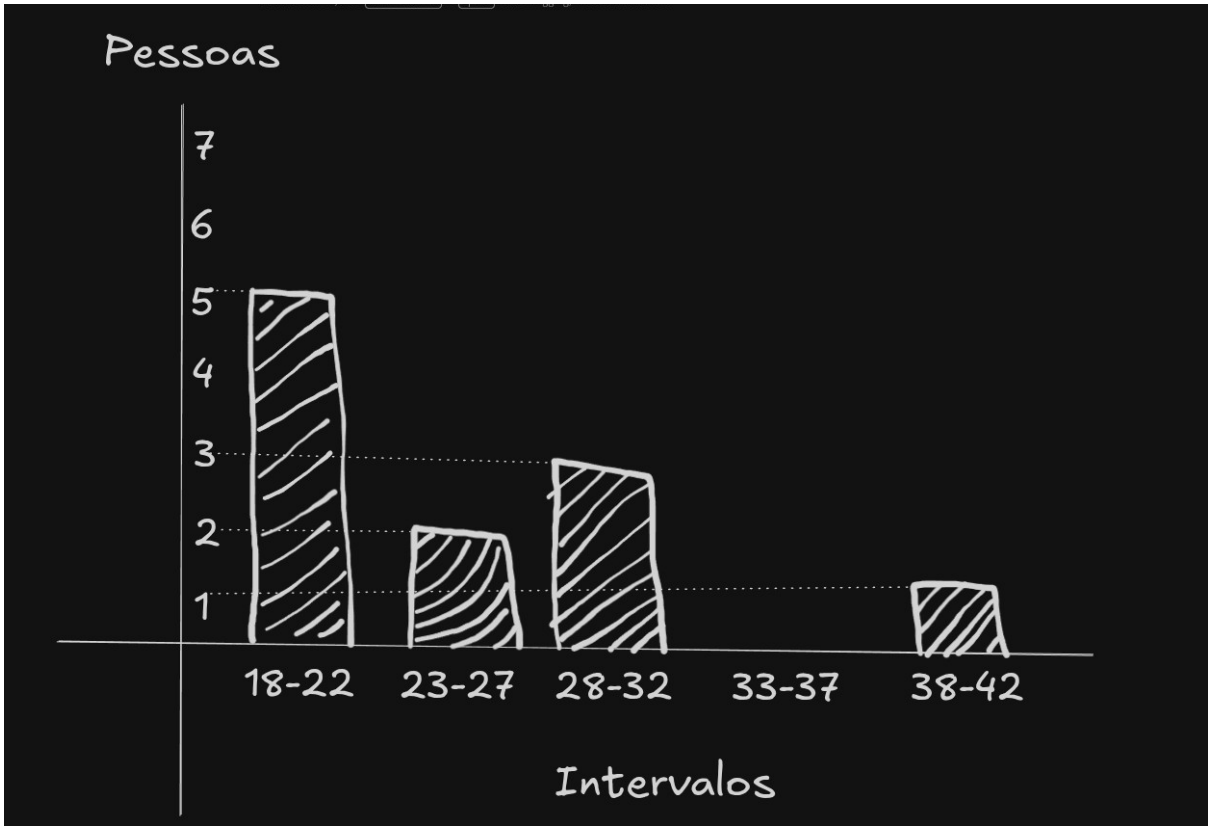


3) Using mean, median, mode in pyhton:

Aqui ele aborda a parte prática no pthon usando numpy, matplotlib.pyplot e o scipy, vou focar nessas duas ultimas bibliotecas pois a de numpy eu já expliquei em relátorios 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 vâriancia 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 trabalho com a amostra, o n – 1 é um 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 da a probabilidade de um dado estar entre um intervalo. Tipo se você quer saber a probabilidade de alguém ter exatamente 1,732432432 metros isso e praticamente impossivel ai entra essa função.

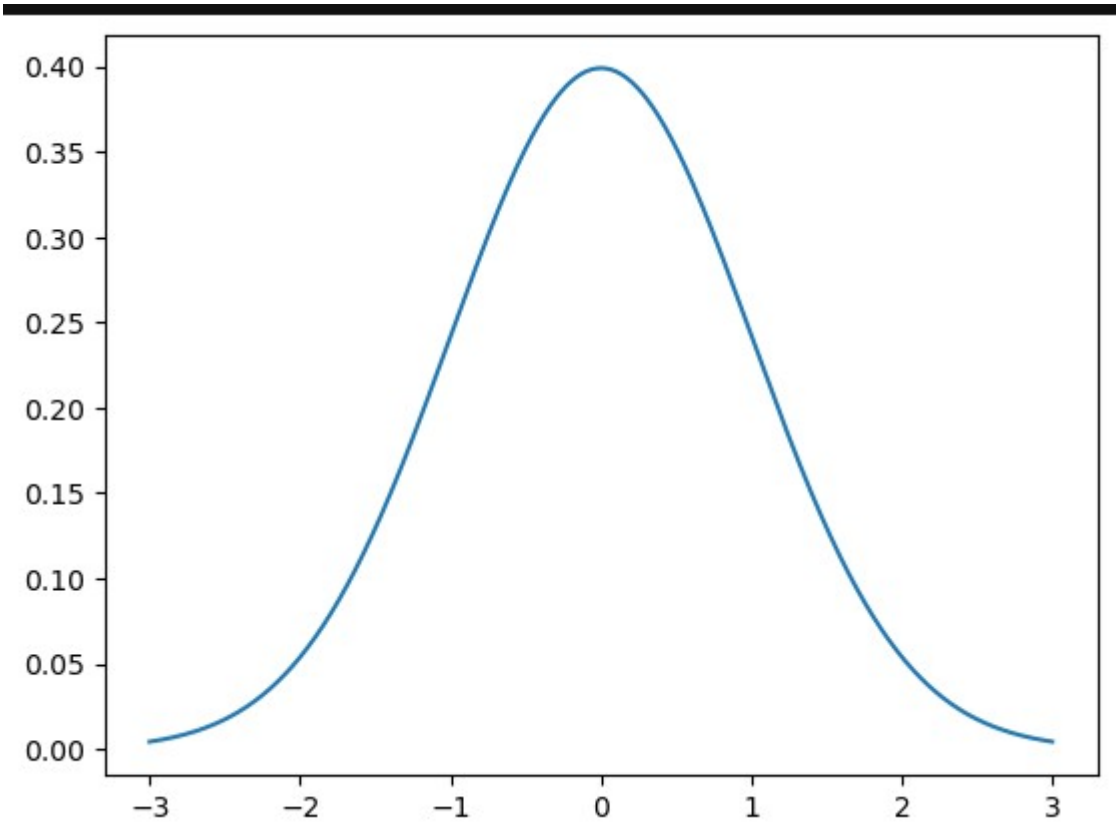
Função de massa de probabilidade: Essa já é usado pra numeros inteiros, não tem um meio termo, quantidade de filhos, numeros ao lançar um dado, etc…, ela já da uma probabilidade exata.

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

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

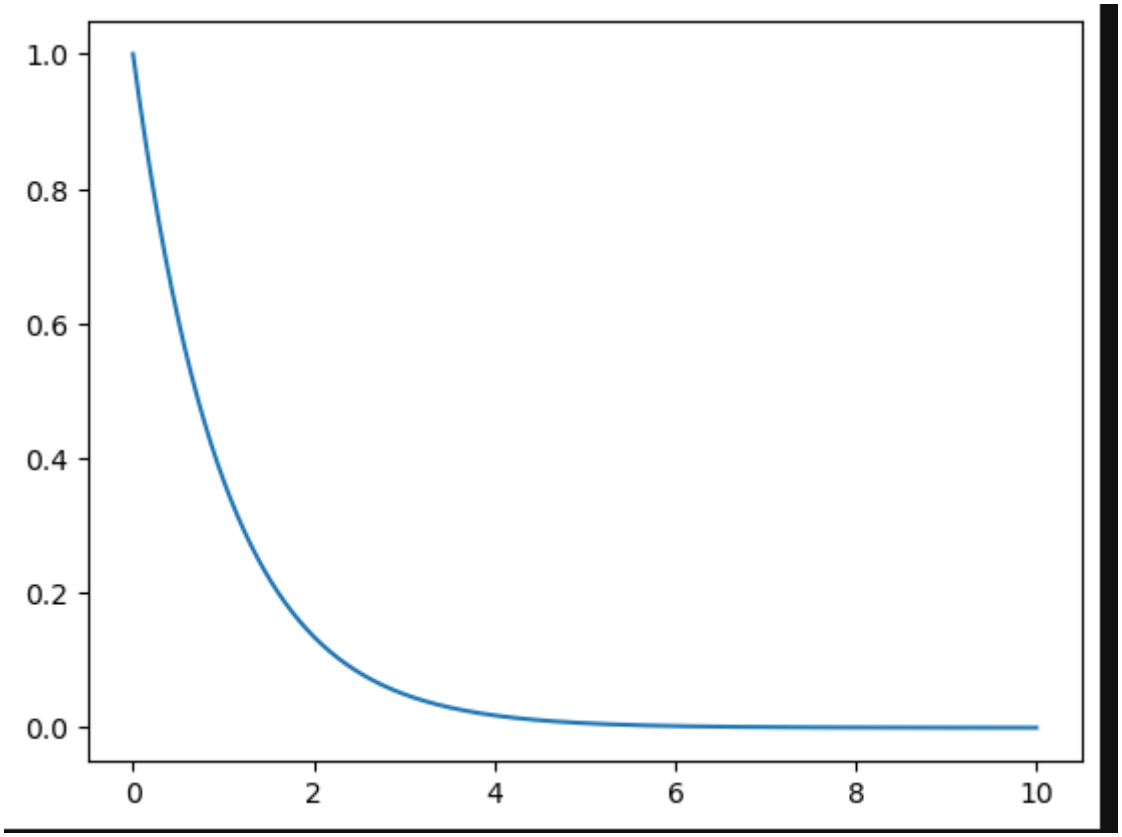
Ele começa falando da distribuição uniforme, os dados tem a mesma chance de aparecer, no video ele faz 100000 numero entre -10 e 10 e cada um tem a mesma chance de aparecer, e joga pra um histograma onde as barras do histograma ficam todas quase da mesma altura.

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:

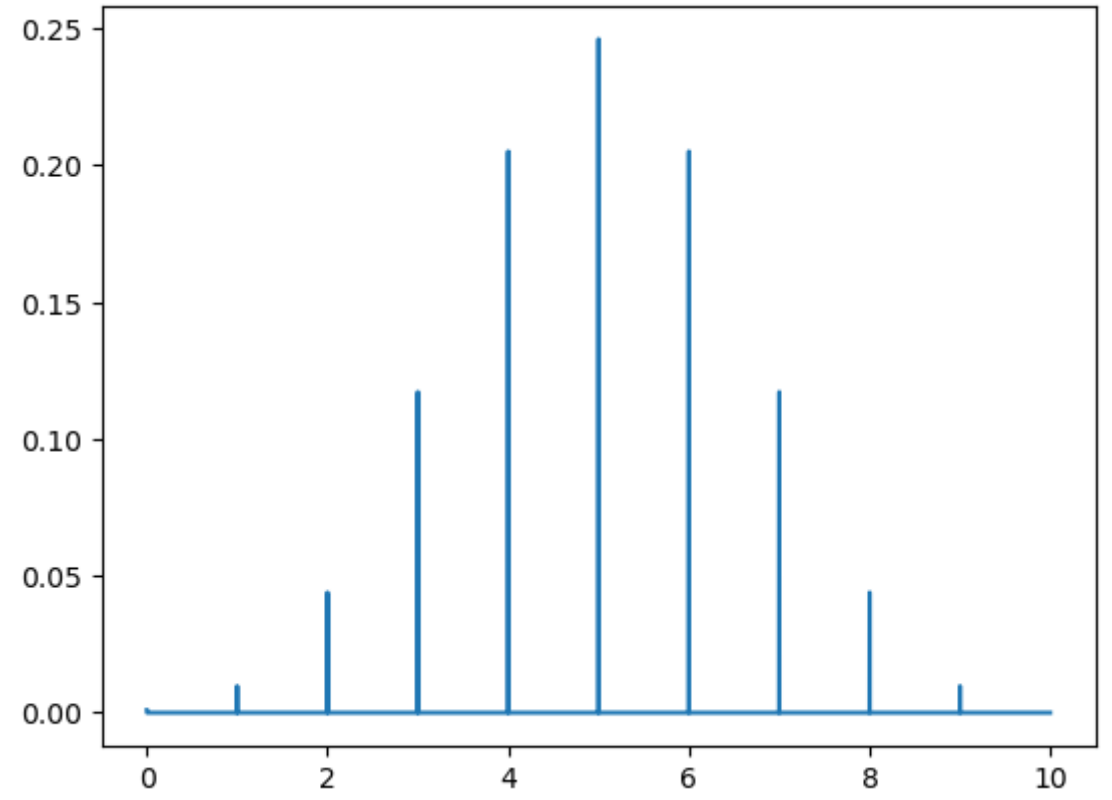


Pra exponencial se impor o expon do scipy para usar, ele pula de 0.001 em 0.001 de novo mas agora de 0 ate 10 e depois usa o plt.plot() e coloca graficamente numa funcao exponencial.

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

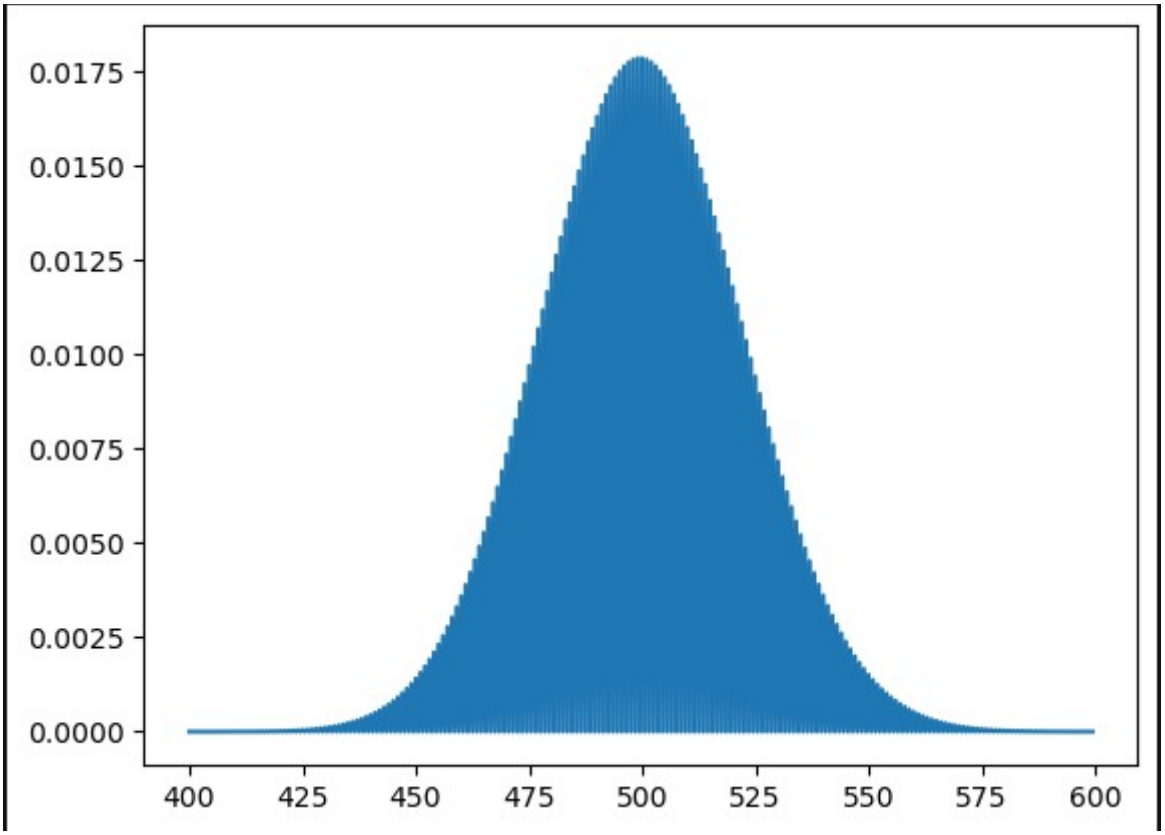


A binomial é usada pra resultados binário, 0 ou 1, sim ou não, então a chance de dar “sucesso” em algo é 50%, ela trabalha com números discretos nunca continuo, no código ele faz a predição de algo dar certo 10 vezes, vamos usar de exemplo que ‘dar certo’ é tirar cara em uma moeda jogando ela 10 vezes, a chance de você tirar 0 caras nas 10 vezes é quase impossível igual você tirar 10 caras nas 10 vezes também é quase impossivel, 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 amanha.

No código ele define o numero medio esperado o mu 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 e 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.

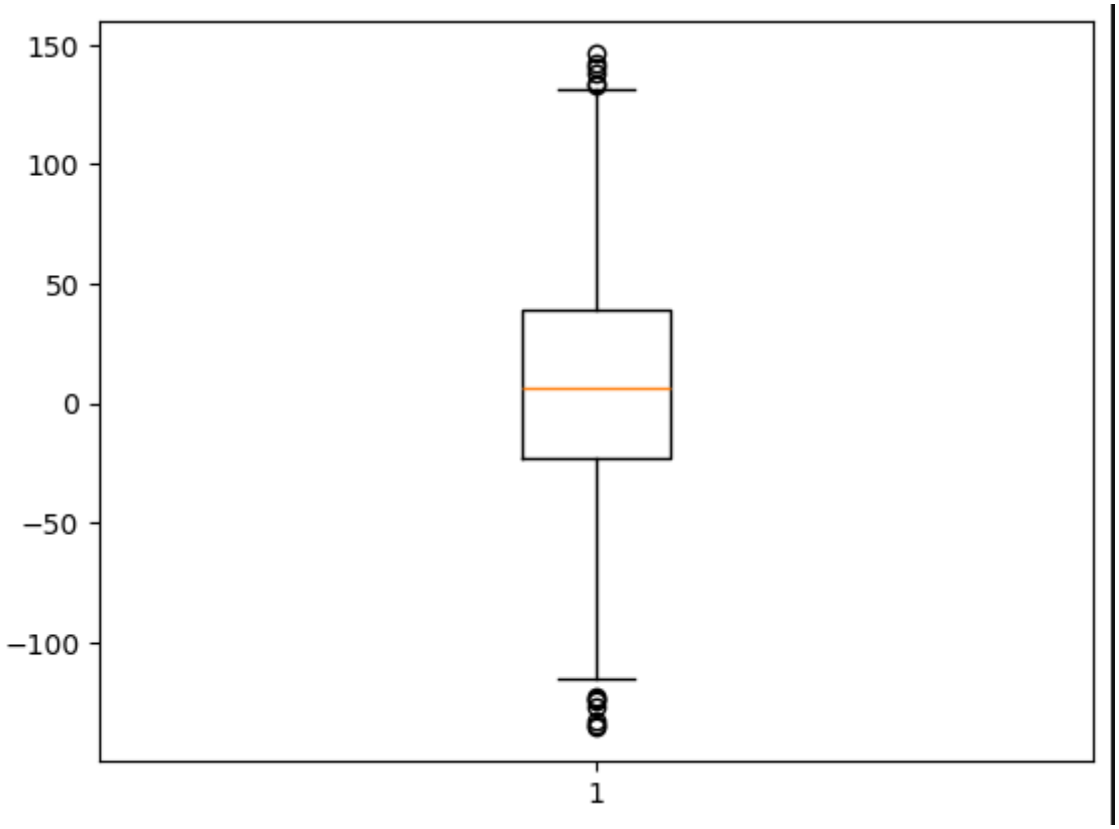
O quarto momento é a kurtosis ou curtoses que mede se a onda do gráfico é estreita ou larga comparada a uma distribuição normal.

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:

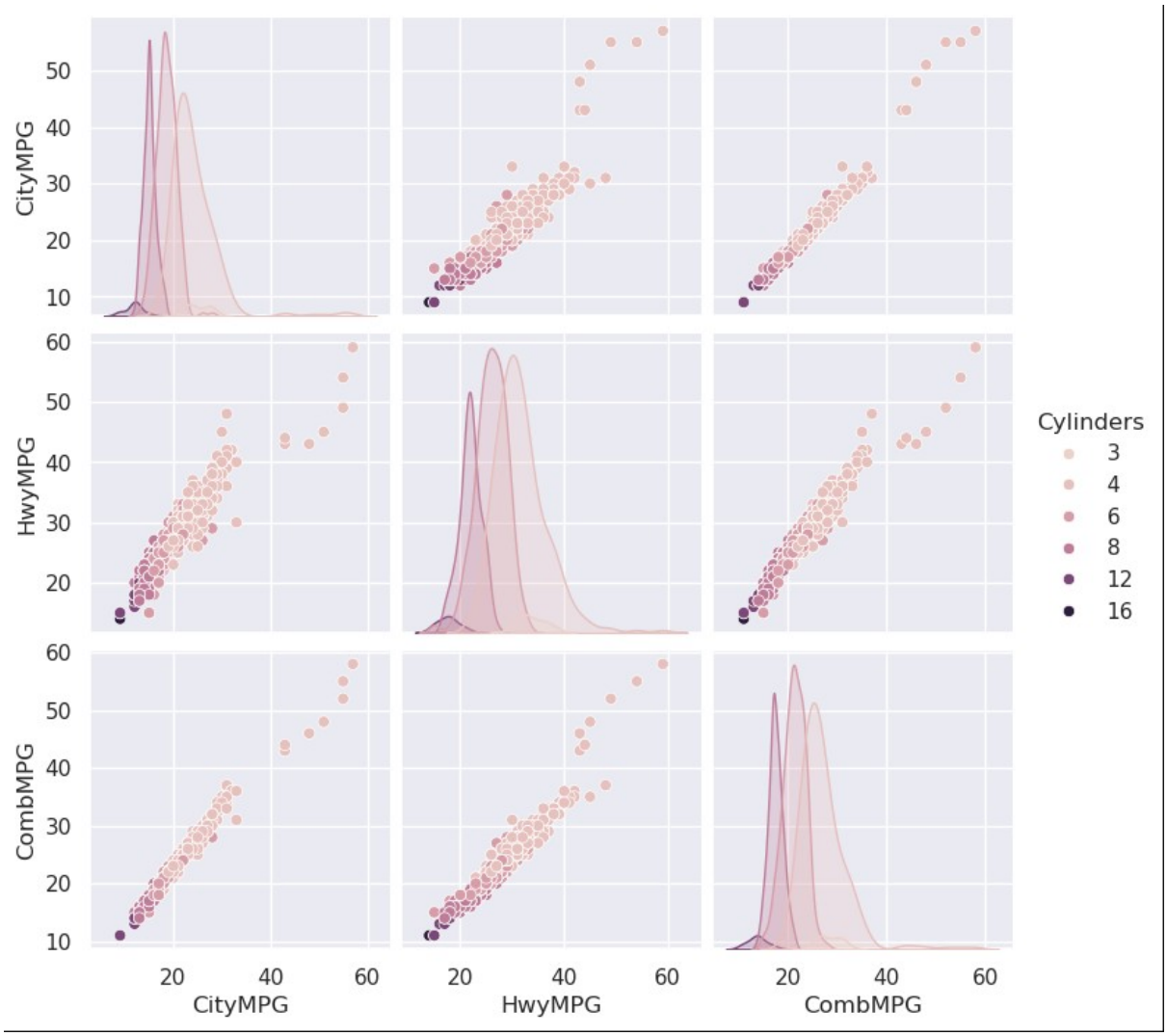


Esse é um dos gráficos um pouco mais chatos de entender de primeira, mas irei explicar, a caixa no meio representa 50% dos dados do conjunto, a linha inferior da caixa é chamada de Q1 é o primeiro quartil ela representa que 25% dos dados estão abaixo desse valor, a linha laranja no meio chamada de Q2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

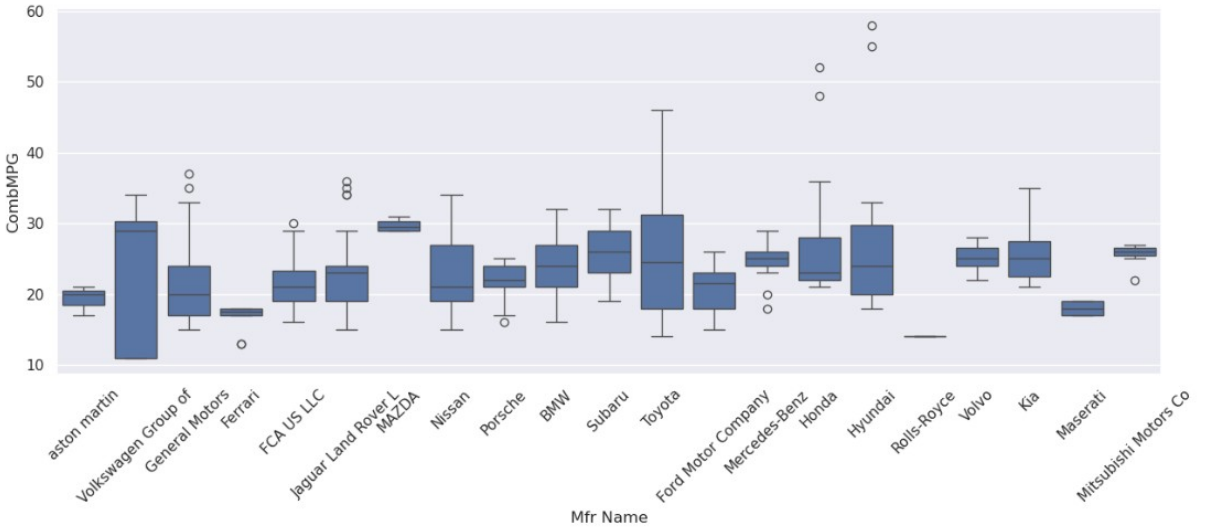
9) Seaborn

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 regrssão, boxplot, gráfico com mapa de calor e gráficos de densidade.

Um exemplo de um boxplot:

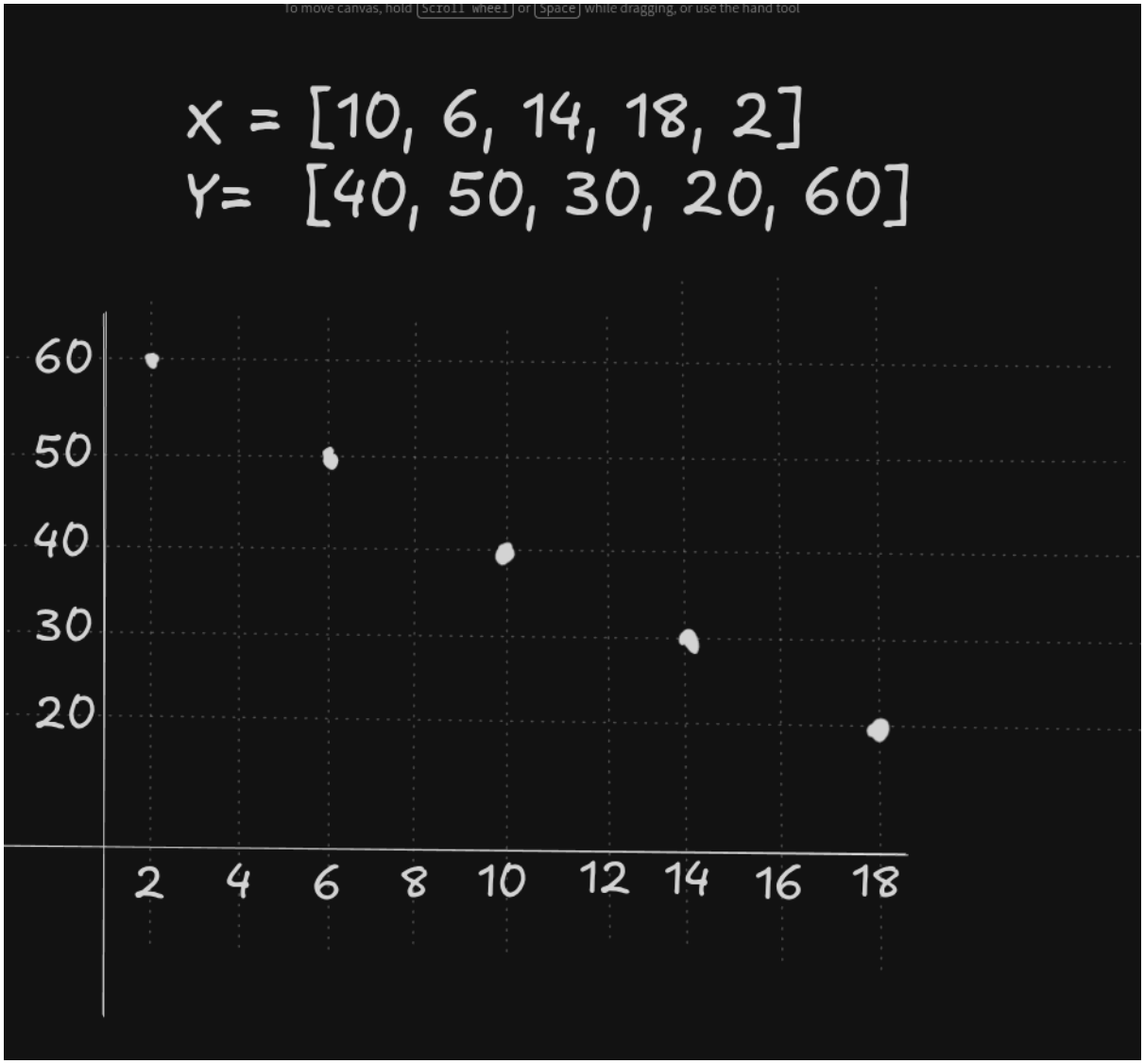


10) Covariance and Correlation:

Fala sobre correlação e covarancia.

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

Um exemplo visual:



A correlação normaliza a covariancia e verifica se tem uma relação perfeita ou uma reação vaseada nos números -1, 0 e 1, o 0 fala que não tem relação, o 1 tem relação e é perfeita positivamente e o -1 é perfeita negativamente.

11) Conditional Probability:

Aqui ele fala sobre a probabildiade condicional, que é a probabilidade de um evento acontecer sabendo que outro evento já aconteceu, tipo, um usuario comprou um livro de Duna qual a probabilidade dele comprar um livro de star wars sabendo que já comprou o de duna.

12) é só resolução de exercicios

13) Bayes's Theorem:

Atualiza as chances de algo ser verdade com base em um dado novo encontrado, exemplo do video, existe um teste com 99% de precisão pra saber se a pessoa usa

droga ou não, e 0.3% da população usa drogas já 99% não serio o suficiente, se a pessoa usa drogas o teste da positivo em 99% dos casos e se a pessoa não usa o teste da um falso positivo em 2% dos casos, a população é de 10.000 pessoas, a soma dos positivos reais e falso positivos da 229, ai o teorema entra, se uma pessoa testou positivo qual é a chance real dela usar drogas, ai você pega o 30 e divide por 229 que da 13%.

Pratica:

Eu baixei um csv do site Kaggle com os dados de músicas da Taylor Swift e fiz vários gráficos diferentes a partir deles.

3. Conclusão

O card ajudou a fortalecer conhecimentos passados mas apresentou muita coisa nova, as aulas foram bem facéis mesmo sendo em inglês eu compreendi tudo, o conteúdo de probabilidade é muito parecido com oque vi na matéria da faculdade então já vim sabendo de várias coisas e a estatística é uma área muito prática de estudar e o python ajudou muito nessa parte do aprendizado. No geral foi um conteúdo muito bem ensinado e de entendimento fácil.