

# DataAnalysisPython

## ▼ Semana 1

- **Sempre precisamos planejar a solução antes de começar a trabalhar no problema**
- **Nessa semana vamos ter um problema de negócio específico e vamos trabalhar nele:**
  - ▼ **Descrição desse primeiro projeto como um todo:**

Link da pagina → <https://sejaumdatacientist.com/os-5-projetos-de-data-science-que-fara-o-recrutador-olhar-para-voce/>

## Projeto Número 01: O Projeto de Insights

Nesse post, eu vou falar especificamente sobre o Projeto do tipo Insights.

O objetivo do projeto de Insights é recomendar soluções para o negócio através de Insights gerados por uma ótima Análise Exploratória de Dados.

O Projeto de Insight cobre 5 passos do roadmap de resolução de problemas em Data Science, sendo esses: A Questão de Negócio, O Entendimento do Negócio, A Coleta de Dados, A Limpeza de Dados e A Exploração de Dados.

Se você cumprir esses 5 passos, você criará um solução para um problema de negócio. Observe que nenhum passo fala sobre usar algoritmos de Machine Learning, isso é proposital, porque eu quero mostrar pra você que Machine Learning é apenas uma ferramenta de Data Science e que vários problema que as empresas enfrentam, podem ser resolvidas com uma ótima análise exploratória de dados.

E para te ajudar a cumprir esses 5 passos, eu vou criar um desafio fictício de uma empresa imaginária para simular um contexto real em problema de negócio. Para um Projeto de Insights, eu sugiro que você crie uma solução para a empresa **House Rocket Company**

***Disclaimer: O Contexto a seguir, é completamente fictício, a empresa, o contexto, o CEO, as perguntas de negócio existem somente na minha imaginação.***

## Contexto do Desafio

A **House Rocket** é uma plataforma digital que tem como modelo de negócio, a compra e a venda de imóveis usando tecnologia.

Você é um Data Scientist contratado pela empresa para ajudar a encontrar as melhores oportunidades de negócio no mercado de imóveis. O CEO da House Rocket gostaria de maximizar a receita da empresa encontrando boas oportunidades de negócio.

Sua principal estratégia é comprar boas casas em ótimas localizações com preços baixos e depois revendê-las posteriormente à preços mais altos. Quanto maior a diferença entre a compra e a venda, maior o lucro da empresa e portanto maior sua receita.

Entretanto, as casas possuem muitos atributos que as tornam mais ou menos atrativas aos compradores e vendedores e a localização e o período do ano também podem influenciar os preços.

Portanto, seu trabalho como Data Scientist é responder as seguinte perguntas:

- 1. Quais casas o CEO da House Rocket deveria comprar e por qual preço de compra?**
- 2. Uma vez a casa em posse da empresa, qual o melhor momento para vendê-las e qual seria o preço da venda?**
- 3. A House Rocket deveria fazer uma reforma para aumentar o preço da venda? Quais seriam as sugestões de mudanças? Qual o incremento no preço dado por cada opção de reforma?**

## Os Dados do Desafio

O conjunto de dados que representam o contexto está disponível na plataforma do Kaggle. Esse é o link: <https://www.kaggle.com/harlfoxem/housesalesprediction>

Esse conjunto de dados contém casas vendidas entre Maio de 2014 e Maio de 2015. Você usará esses dados para desenvolver sua solução.

## Como solucionar esse desafio?

Não se assuste com o problema, respire fundo, mantenha a mente clara e limpa e então, comece a pensar de forma estruturada em alternativas para responder à essas perguntas.

Eu vou deixar aqui um roteiro para você se orientar, ele pode ser modificado da forma que você preferir ou simplesmente ignorado. Provavelmente, você já tem um roteiro de resolução melhor para abordar esse desafio.

Dicas preciosas para você começar: Tenha calma, não tenha medo de criar suposições e considerações, faça um passo de cada vez, não se prenda muito na parte técnica e foque em responder as perguntas,

todas suas ações devem te deixar um passo mais próximo da solução final. Sempre pense: “Se eu fizer isso, me ajuda a chegar mais próximo da resposta?” Se a resposta for Sim, faça, se não, tome outra ação.

E o mais importante, tenha paciência, criar uma solução leva tempo, as respostas não ficam prontas do dia pra noite, assuma uma postura resiliente e nunca desista, afinal você quer ser um Data Scientist e ganhar um ótimo salário, não quer?

## Roteiro Sugerido para a Resolução.

Esse é o roteiro de resolução do desafio que eu sugiro

### 1. Identifique a causa raiz.

- Porque o CEO fez essas perguntas? Se você fosse ele, porque você perguntaria isso? Quer aumentar receita? A empresa está indo bem?
- Anote essas causas.

### 2. Colete os dados ( Os dados estão no link acima )

### 3. Aplique uma limpeza nos dados.

- Entenda as variáveis disponíveis, possíveis valores faltantes, faça uma estatística descritiva para entender as características dos dados.

### 4. Levante Hipóteses sobre o Comportamento do Negócio.

- Casas com garagens são mais caras? Porque?
- Casas com muitos quartos são mais caras? Porque? A partir de quantos quartos o preço aumenta? Qual o incremento de preço por cada quarto adicionado?

- As casas mais caras estão no centro? Qual a região? Existe alguma coisa na região que tem correlação com valor de venda da casa? Shoppings? Montanhas? Pessoas Famosas?

#### **5. Faça uma ótima Análise Exploratória de Dados.**

- Quais hipóteses são falsas e quais são verdadeiras?
- Quais as correlações entre as variáveis e a variável resposta?

#### **6. Escreve os Insights que você encontro.**

#### **7. Escreve possíveis soluções para o problema do CEO.**

## **O Ferramental da Solução**

Usa as ferramentas que você se sente mais confortável para desenvolver a solução. Você pode usar tanto Python quanto R e qualquer IDE de sua preferência Jupyter Notebook, Spyder, VS Code, entre outros.

Você pode usar o Google Colab também, caso você não tenha um computador razoável, ou caso queira testar essa incrível ferramenta do Google.

Aproveite esse projeto para melhorar sua velocidade na manipulação de dados com linguagens de programação. Alcance um nível, no qual você consiga escrever códigos rapidamente, sem ficar olhando no Stackoverflow a cada linha código.

## **Vá em Frente!**

Não existe caminho fácil, de curto prazo em nenhum profissão, muito menos em Data Science, mas existe o caminho certo. E o caminho certo é adquirir experiência através do desenvolvimento de projetos para mostrar sua capacidade.

Volto a repetir, os projetos do seu portfólio precisam demonstrar que você é tão capaz de resolver desafios

de negócio quantos os Data Scientists que já atuam profissionalmente nas empresas.

Quando você conseguir solucionar esse desafio, escreva um artigo, explicando toda sua linha de raciocínio, o contexto do problema, todas as considerações assumidas, as hipóteses validadas e as rejeitadas e as suas sugestões para solucionar o problema da House Rocket Company.

Se quiser publicar aqui no blog, me manda um msg no LinkedIn ([@meigarom](#)) ou no Instagram ([@meigarom.datascience](#)).  
Publicarei seu trabalho com o maior prazer do mundo.

▼ **Vamos começar as anotações da live 1 para resolver os problemas:**

▼ **O problema de negocio**

- Empresa: House Rocket
- O que a empresa faz : House rocket plataforma de compras e vendas de imoveis
- Qual o problema dela : O CEO quer maximizar o lucro de sua empresa encontrando bons negócios para fazer(imoveis com valor abaixo da média que devem ser vendidos mais caro de forma máxima)
- A principal estratégia hoje : usar fontes externas para comprar imóveis(falar com pessoas, procurar em sites como olx,...), só que ele acha que isso é muito ineficiente
- As perguntas dele para saber como filtrar melhor os imóveis para comprar :
  1. quantas casas estão disponíveis para compra

2. quantos atributos as casas possuem? (num de quartos, num de garagens, m2, vista pro mar,...)
3. quais sao os atributos?
4. qual a casa mais cara do portifolio(maior valor)?
5. qual a casa com maior num de quartos

### ▼ **Planejamento da solução**

#### ▼ Planejamento do Produto final

1. O que vou entregar? (planilha, texto, email, Modelos de ML,...)

R→Texto com respostas

2. Como vai ser a entrega? (como vai ser a planilha, quais os tipos de grafico, como vai ser seu storytelling )

Exemplo:

1. quantas casas estao disponiveis para compra

R→ 2300 imoveis para compra

2. quantos atributos as casas possuem? (num de quartos, num de garagens, m2, vista pro mar,...)

R→ 10 atributos

#### ▼ Planejamento do processo

1. Onde esta as informaçoes que vou trabalhar? (Excel, Banco de dados, API, manual,...)

R→ <https://www.kaggle.com/harlfoxem/housesalesprediction>

2. Como vou coletar essas informações? (SQL, Python, Streamlit,...)

R→ Apertar o botão para baixar(Download)

3. Responder as perguntas?

1. quantas casas estão disponíveis para compra

R→ Contar o número de linhas do conjunto de dados

2. quantos atributos as casas possuem? (num de quartos, num de garagens, m<sup>2</sup>, vista pro mar,...)

R→Contar o número de colunas do conjunto de dados

3. quais são os atributos?

R→Mostrar o nome das colunas(de forma automática)

4. qual a casa mais cara do portifólio(maior valor)?

R→Ordenar as linhas pela coluna de preço(atributos)

5. qual a casa com maior num de quartos?

R→Ordenar as linhas pela coluna de num de quartos(ordenar os atributos)

#### ▼ Planejamento das ferramentas

1. Quais ferramentas eu posso usar?

-Excel

• +

- Facil de usar
- Barato
- Muito usado pelos times nao tecnicos
- -
- Poder de processamento limitado(1 milhao)

#### ▼ Disclaimer

Disclaimer→ o excel trouxe nos ate aqui hoje, nao  
menosprezar pois é uma ferramenta poderosa,  
Problema que vivemos em uma era de BIG DATA entao 1  
milhao de linhas apenas é muito pouco para os  
problemas de data analysis que enfrentamos nas  
empresas hoje.

como é gerado os dados?

Dados = pessoas + produto(a interação delas)

Exemplo: Cliente moda + website = dados de  
navegação (view, add to cart,...), esses clientes geram  
dados que nos podemos analizar e tomar decisões para  
maximizar o lucro, eficiência da nossa empresa, isso  
que é a era big data e isso que o data scientist vai  
resolver de problema

profissional de big data→ processa dados em grande  
volume, e gera soluções a partir disso

-Linguagem de programação

- Desenvolvidas para criar softwares
- É escalável, da para aplicar e processar muito mais os  
insights que vc tiver nela

-

#### ▼ O que é Python

▼ linguagem de prog:

1. O que é linguagem de programação?

É um tipo de linguagem usada pelo homem para desempenhar comunicação com a máquina, pois essa não reconhece a linguagem normal do ser humano. Atualmente é possível encontrar diversos tipos de linguagem de programação, sendo as principais: Java, C, C++, C#, Php, Delphi, entre outras. ...

▼ Exemplos de comandos em python

Ideia: Selecionar duas colunas

Comando: `data[[X, Y]]`

Ideia: Ordenar as linhas do conjunto de dados pela coluna Z

Comando: `data.sort_values( Z )`

Ideia: Ordena as linhas do conjunto de dados pela coluna Z de forma crescente

Comando: `data.sort_values( Z, ascending=True )`

▼ Para responder as perguntas do CEO, usaria os seguintes comandos:

- Contar o numero de linhas do conjunto de dados:

R→

-Contar o numero de colunas do conjunto de dados

R→

-Mostrar o nome das colunas(de forma automatica)

R→

-Ordenar as linhas pela coluna de preço(atributos)

R→

-Ordenar as linhas pela coluna de num de quartos(ordenar os atributos)

R→

▼ Escrevendo Primeiros codigos:

- Vamos de pycharm nesse começo, ide da JetBrains(jetbrains te amo)
- Sempre ter na sua cabeça:
  1. O que preciso fazer
  2. qual função faz pra mim
  3. qual biblioteca ela está armazenada

▼ Exercício Semana 1



## - Novas Perguntas do CEO para você:

1. Quantas casas estão disponíveis para compra?
2. Quantos atributos as casas possuem?
3. Quais são os atributos das casas?
4. Qual a casa mais cara ( casa com o maior valor de venda )?
5. Qual a casa com o maior número de quartos?
  
6. Qual a soma total de quartos do conjunto de dados?
7. Quantas casas possuem 2 banheiros?
8. Qual o preço médio de todas as casas no conjunto de dados?
9. Qual o preço médio de casas com 2 banheiros?
- ~~10. Qual o preço mínimo entre as casas com 3 quartos?~~

### ▼ Semana 2

- Nessa aula vamos seguir o seguinte fluxo
  1. Novas Perguntas de Negócio
  2. Planejamento da Solução
  3. Tipos de Variáveis
  4. Manipulação de Variáveis
  5. Exercícios Práticos

**(ctrl+/' comenta as linhas selecionadas)**

### ▼ Novas Perguntas de Negócio

#### ▼ Recapitulando o que queremos ao final

<https://medium.com/@meigarom/os-5-projetos-de-data-science-que-fará-o-recrutador-olhar-para-você-c32c67c17cc9#:~:text=Os%205%20tipos%20de%20Projetos%20de%20Data%20Science.&text=Eu%20acredito%20que%20os%205,e%20Projetos%20de%20Data%20Science.>

- Empresa: House Rocket
- O que a empresa faz : House rocket plataforma de compras e vendas de imoveis
- Qual o problema dela : O CEO quer maximizar o lucro de sua empresa encontrando bons negocios para fazer(imoveis com valor abaixo da media que de para vender mais caro de forma maxima)

#### ▼ Novas Perguntas

1. Qual a data do imovel mais antigo do portifolio
2. Quantos imoveis possuem o maximo e andares
3. Criar uma classificacao para os imoveis, separando-os em alto e baixo padrao, se acordo om o preço(acima de 540.000 é alto padrao)
4. Relatorio ordenado pelo preço e contendo as seguintes informaçoes:
  1. id do imovel
  2. data que o imovel ficou disponivel para compra
  3. numero de quartos
  4. tamanho total do terreno
  5. preço
  6. classificacao do imovel(alto e baixo padrao)
  7. Mapa indicando onde as casas tao localizadas geograficamente

#### ▼ Planejamento da Soluçao(mais importante para nao se perder)

- Precisamos de 3 coisas para fazer um planejamento rapido e eficiente:
  1. Produto Final(o que vou entregar: Planilha, Grafico, Email, Modelo de ML,...)
  2. Ferramenta(qual ferramenta eu vou usar?[se a empresa ja tem essa ferramenta, se vai ter que comprar uma nova, usar um open source,...])
  3. Processo(Como vou fazer?)

▼ Planejamento

1. Produto Final(o que vou entregar: Planilha, Grafico, Email, Modelo de ML,...)

-Email + 2 anexos:

-Email:

Texto: Perguntas | Respostas

-Anexo:

-Um relatorio em .csv

-A foto de um mapa em html

2. Ferramenta(qual ferramenta eu vou usar?[se a empresa ja tem essa ferramenta, se vai ter que comprar uma nova, usar um open source,...])

-Python 3.8.0

-PyCharm

3. Processo(Como vou fazer?)

1. Qual a data do imovel mais antigo do portifolio

- Ordenar conjunto de dados pela menor data
- printar o primeiro elemento

2. Quantos imoveis possuem o maximo e andares

- encontrar os numeros de andares e determinar o maior

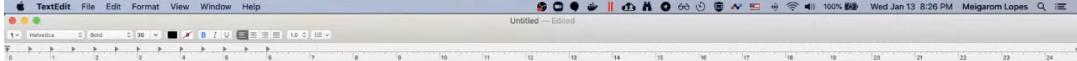
- Contar quantos imoveis tem aquele numero de andar
3. Criar uma classificacao para os imoveis, separando-os em alto e baixo padrao, se acordo om o preço(acima de 540.000 é alto padrao)
- Criar uma nova coluna no conjunto de dados chamado standard(standard e high\_standard )
    - para cada linha do DataFrame vou comparar a coluna "price"
      - Se "price" for maior que 540.000 escrevo na coluna "standard" high\_standard, se nao escrevo low\_standard
4. Relatorio ordenado pelo preço e contendo as seguintes informações:
1. id do imovel
  2. data que o imovel ficou disponivel para compra
  3. numero de quartos
  4. tamanho total do terreno
  5. preço
  6. classficacao do imovel(alto e baixo padrao)
    - ate aqui é só mandar pra ele as colunas desejadas no pedido
    - ou deletar as colunas não desejadas pelo CEO
7. Mapa indicando onde as casas estão localizadas geograficamente
- Procurar uma biblioteca em Python que armazena uma função que desenha mapa

- Aprender a usar a função que desenha mapa

▼ Tipos de variáveis

▼ Boas práticas de nome

- Nesse tópico ele dá uma geral sobre como devemos dar nome a variáveis, só que nesse ponto não julgo necessário repetir isso



**3. Os tipos das variáveis:**

- Caixa armazenadora ( espaço de memória )
- Precisa ter um **NOME** e um **TIPO**
- **Boas práticas para o nome:**
  - expressar a responsabilidade da variável
  - seguir estilo “Kamel Case” e “Snake case”
    - “Kamel Case”: HousePrice,
    - “Snake Case”: house\_price, house price, house,



- Esse é o print do que ele escreveu

▼ Tipos de variáveis em Python



```
- "Snake Case": house_price
```

- **Tipos de variáveis em Python**

- Numérica ( Inteiro, Float ) - Inteiro: Valor sem vírgula, Float: com vírgula. ( 4, 3.8 )
- Categórica ( characters, strings ) "o" "m" - "meigarom"
- Dates ( date, timestamp ) - Date: Ano-Mes-Dia, Timestamp: Ano-Mes-Dia H:M:S

- **Identificar os tipos das variáveis:**

- Comando "dtypes"



- o "dtype" é para saber qual tipo da variável
- Cara, Sempre antes de uma análise de dados a primeira coisa que devemos fazer é VER QUAIS SÃO OS TIPOS DAS VARIÁVEIS DE CADA COLUNA

## ▼ Manipulação de Variáveis

- Criar( colunas de variáveis e novas linhas )
- Deletar( colunas de variáveis e novas linhas )
- Selecionar:
  - 4 formas de Selecionar dados:
    1. Direto pelo nome das colunas
    2. Pelos índices das colunas
    3. Pelos índices das linhas e pelo nome das colunas
    4. Pelos índices booleanos ( True ou False )

The screenshot shows a web browser window with multiple tabs open. The active tab is a Stack Overflow question titled "python - What does axis in pandas mean?". The question has 22 answers. One answer, by user numeratus, includes a diagram to explain the difference between axis=0 and axis=1.

**Diagram Explanation:**

```

+-----+-----+
|       | A     | B   |
+-----+-----+
| 0    | 0.626386 | 1.52325 |
+-----+-----+
|       |           |
|       | axis=0      |
+-----+

```

The diagram shows a 2x2 matrix with columns labeled A and B. The first column has a value of 0.626386 and is associated with axis=0. The second column has a value of 1.52325 and is associated with axis=1. Arrows point from the labels "axis=0" and "axis=1" to their respective columns.

- no material a gente usa o axis mas n entende direito, ja aqui ta uma explicacao legal do stackoverflow

## ▼ Exercícios Práticos(Aula)

- tem as perguntas e os códigos comentados no repositório

## ▼ Exercícios Práticos("Casa")



**1. Crie uma nova coluna chamada: "house\_age"**

- Se o valor da coluna "date" for maior que 2014-01-01 => 'new\_house'
- Se o valor da coluna "date" for menor que 2014-01-01 => 'old\_house'

**2. Crie uma nova coluna chamada: "dormitory\_type"**

- Se o valor da coluna "bedrooms" for igual à 1 => 'studio'
- Se o valor da coluna "bedrooms" for igual a 2 => 'apartament'
- Se o valor da coluna "bedrooms" for maior que 2 => 'house'

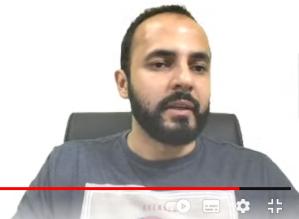
**3. Crie uma nova coluna chamada: "condition\_type"**

- Se o valor da coluna "condition" for menor ou igual à 2 => 'bad'
- Se o valor da coluna "condition" for igual à 3 ou 4 => 'regular'
- Se o valor da coluna "condition" for igual à 5 => 'good'

**4. Modifique o TIPO a Coluna "condition" para STRING**

**5. Delete as colunas: "sqft\_living15" e "sqft\_lot15"**

**6. Modifique o TIPO a Coluna "yr\_build" para DATE**



- Se o valor da coluna "condition" for menor ou igual à 2 => 'bad'
- Se o valor da coluna "condition" for igual à 3 ou 4 => 'regular'
- Se o valor da coluna "condition" for igual à 5 => 'good'

**4. Modifique o TIPO a Coluna "condition" para STRING**

**5. Delete as colunas: "sqft\_living15" e "sqft\_lot15"**

**6. Modifique o TIPO a Coluna "yr\_build" para DATE**

**7. Modifique o TIPO a Coluna "yr\_renovated" para DATE**

**8. Qual a data mais antiga de construção de um imóvel?**

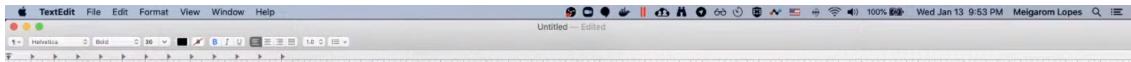
**9. Qual a data mais antiga de renovação de um imóvel?**

**10. Quantos imóveis tem 2 andares?**

**11. Quantos imóveis estão com a condição igual a "regular" ?**

**12. Quantos imóveis estão com a condição igual a "bad" e possuem "vista p**





19. Salve um arquivo .csv com somente as colunas do item 10.
20. Modifique a cor dos pontos no mapa de “pink” para “verde-escuro”



## ▼ Semana 3

- Planejamento dessa aula
  1. Novas perguntas de negocio

2. Planejamento da solução
3. Ferramentas para criar códigos em Python
4. Estrutura de Dados
5. Transformação de Dados
6. Respondendo as perguntas do CEO
7. Exercícios Práticos

▼ Novas Perguntas de negócio

▼ Recapitulando o desafio

<https://medium.com/@meigarom/os-5-projetos-de-data-science-que-fará-o-recrutador-olhar-para-você-c32c67c17cc9>

- Empresa : House Rocket
- Modelo de negócio : comprar imóveis por um preço mais baixo que tenha alto potencial de revenda(lucro = venda - compra)
- Qual desafio: Encontrar bons negócios, ou seja, encontrar casas com um preço baixo e com alto potencial de revenda

▼ Novas perguntas de negócio do CEO

1. Qual o número de imóveis por ano de construção?
2. Qual o menor número de quartos por ano de construção dos imóveis?( se tem uma tendência de aumentar o número de quartos por ano em média )
3. Qual o preço de compra mais alto por cada número de quarto?( se tem relação entre o número de quartos e o preço do imóvel )
4. qual a soma de todos os preços de compra por número de quartos?( se talvez os imóveis de 1 quarto são mais valorizados em média que os de 3 ou 4 por exemplo )
5. Qual a soma de todos os preços de compra por número de quartos e banheiros?( mesma ideia de cima só com a variável banheiro a

mais )

6. Qual o tamanho medio das salas dos imoveis por ano de construcao?( se os tamanhos de sala foram aumentando ao longo dos anos )
7. Qual o tamanho mediano(mediana) das salas dos imoveis por ano de construcao?( vendo se a conclusao que tiramos na pergunta acima se confirma, qual o termo centralizado [esse é um conceito de estatistica, vamos ver em estatistica esploratoria 1] )
8. Qual o desvio-padro do tamanho das salas dos imoveis por ano de construcao( o quao meus dados estao espalhados da media )?
9. Como é o crescimento medio de preços de compras dos imoveis, por ano, por dia, pela semana do ano?
10. Eu gostaria de olhar no mapa e conseguir identificar as casas com maior preço.

▼ Planejamento da solução

▼ Produto Final:

▼ explicaçao

- O que vou entregar?
- Ex: planilha, grafico, modelo de ML,...
- Email + 2 anexos:
  - Email: as respostas das perguntas
    - Pergunta | Resposta
  - Anexo 1 : Um dashborard com 3 graficos
  - Anexo 2 : Um mapa no formato html.

▼ Ferramenta:

▼ explicaçao

- O que vou usar?

- ou seja, discutir com o time de dados como vamos atacar o problema, com que ferramentas nos vamos fazer isso
- big data por ex ⇒ spark, é melhor que o pandas para grande volume de dados
- Python 3.8.0
- Jupyter NoteBook

▼ Processo( como fazer? ):

1. Qual o numero de imoveis por ano de construção?
  - Contar o numero de ids por ano de construção
2. Qual o menor numero de quartos por ano de construção dos imoveis?( se tem uma tendência de aumentar o numero de quartos por ano em media )
  - Filtrar todos os imoveis por ano de construção e selecionar o menor numero de quartos
3. Qual o preço de compra mais alto por cada numero de quarto?( se tem relação entre o numero de quartos e o preço do imovel )
  - Filtrar todos os imoveis por numero de quartos e selecionar o de maior preço
4. qual a soma de todos os preços de compra por numero de quartos?( se talvez os imoveis de 1 quarto são mais valorizados em média que os de 3 ou 4 por exemplo )
  - Filtrar todos os imoveis por numero de quartos e somar todos os preços
5. Qual a soma de todos os preços de compra por numero de quartos e banheiros?( mesma ideia de cima só que com a variável banheiro a mais )
  - Filtrar todos os imoveis por numero de quartos e por numero de banheiros e realizar a soma de todos os preços
6. Qual o tamanho medio das salas dos imoveis por ano de construção?( se os tamanhos de sala foram aumentando ao

longo dos anos )

- Filtrar todos os imoveis por ano de construção e fazer a media do tamanho de salas
7. Qual o tamanho mediano(mediana) das salas dos imoveis por ano de construção? ( vendo se a conclusao que tiramos na pergunta acima se confirma, qual o termo centralizado [esse é um conceito de estatistica, vamos ver em estatistica esploratoria 1] )
- Filtrar todos os imoveis por ano de construção e fazer a mediana do tamanho de salas
8. Qual o desvio-padrão do tamanho das salas dos imoveis por ano de construção( o quao meus dados estao espalhados da media )?
- Filtrar todos os imoveis por ano de construção e fazer a desvio-padrão do tamanho de salas
9. Como é o crescimento medio de preços de compras dos imoveis, por ano, por dia, pela semana do ano?
- Filtrar todos os imoveis por ano e fazer um grafico onde o eixo x eu tenho o ano e o eixo y eu tenha a media do preço por ano( repito isso pro dia e semana do ano )
  - Estudar uma biblioteca que tenha alguma função que desenhe um grafico de linhas
10. Eu gostaria de olhar no mapa e conseguir identificar as casas com maior preço.
- Modificar o mapa da entrega da aula anterior com que os pontos tenham o tamanho dependente do preço

▼ Ferramentas para criar códigos em Python

▼ IDEs( Interface Development Environment )

- Pycharm
- VSCode
- Spyder
- JupyterLab

▼ Notebooks( + recentes )

- Jupyter Notebook

▼ Vantagens e Desvantagens de um ou outro

▼ IDEs

- IDE'S é sempre necessario executar TODOS os comandos para codigo de maquina, todas aasa vezes que quiser executar seu script

▼ Notebook

- Notebook voce pode "rodar" comandos independentemente um dos outros
  - Organiza melhor seus codigos
  - A Analise Exploratoria de Dados fica mais facil e mais ornanizada
  - StoryTelling é facilitado demais pelo jupyter notebook( tem como usar esse formato de historia )

▼ Anaconda

- Anaconda é um ambiente de desenvolvimento que gerencia biblioteca de python e R para ambientes
- Massa que tambem que tem umas bibliotecas de I.A para a gente usar mais na frente depoois dessa parte basica de dados

▼ Estrutura de Dados

- As 4 Estruturas de dados mais usadas em Python são:
  - ▼ Listas ( Proxima semana so )
  - ▼ Dicionarios
    - É uma tabela de espalhamento, uma estrutura de dados que armazena informaçao me uma estrutura de chave-valor
    - Todos os dados armazenados no dicionario precisam ter uma chave
    - Precisam de um nome
    - Nao aceita valores duplicados, mesmo valor
  - ▼ Dicionario sintaxe( sempre é string a chave )
    - Declaracao:

```
d = {'chave' : valor, 'chave2' : valor2, 'chave3' : valor3 ,
      'chave4' : valor4}
```

`skirt = {'size' : 'M', 'price' : 139.90, 'color' : 'black'}` ⇒ isso é um dicionario que vai ser armazenado dentro da variavel skirt

`skirt = { 'size' : 'M', 'price' : 139.90, 'color' :
 ['black','red','green'], date : 2020-01-01' }` ⇒ pode ser uma lista os valores de uma chave, ou ate mesmo outros dicionarios

- Acessar os valores do dicionario:
  - `skirt['size']` → retorna o valor 'M'
  - `skirt['color']` → retorna a lista ['black','red','green']
  - `skirt['color'][0]` → retorna 'black'
- Como criar um dicionario vazio:

```
skirt = {}
```

- Adicionar novos dados dentro de um dicionario:  

```
skirt['category'] = 'bottom'
```
- Nao aceita valores duplicados:  

```
skirt['size'] = 'M' → esse valor nao vai dar erro, mas como no size ja tem o valor 'M' entao ele simplesmente nao adiciona nada
```

## ▼ Tuples ( Proxima semana so )

## ▼ DataFrames

- Armazenam dados na forma tabular com nomes nas linhas e nas colunas
- precisam de nomes

### ▼ DataFrame

- Como Criar um dataframe vazio

```
data = pd.DataFrame()
```

- Como popular um dataframe vazio:

- Atraves de um dicionário

```
data = {'size' : ['P','M','G'], 'price' : [139.90, 59.90, 29.90],  
       'color' : ['black','red','blue'] } → se esse dicionario for com  
       um objetivo de popular um dataframe entao ele vai ter  
       que ter a mesma quantidade de valores em todas as  
       chaves, a nao ser que passemos um parametro que  
       vamos ver mais na frente que povoar as rows que nao  
       tem nada de entrada
```

```
data = pd.DataFrame( data ) → as chaves do dicionario  
viraram as colunas e os valores viraram as rows( linhas )
```

## ▼ Transformação de Dados

### ▼ Agrupamento

- Sequencia de 3 tarefas: Split, Apply, Combine ( separa, aplica, combina )
- é uma operação de agrupamento o .groupby( ele separa, aplica nosso filtro e combina no resultado que gera pra gente )
- Agrupar certinho é 80% de manipulação de dados na verdade, então entendendo groupby, já é muito caminhão andado

### ▼ Operações matemáticas

- Com dados agrupados, podemos realizar operações matemáticas:
  - Exemplos:
    - Contagem
    - Mínimos
    - Máximos
    - Soma
    - Média
    - Mediana
    - Desvio-padrão

## ▼ Respondendo as perguntas do CEO

1. Qual o número de imóveis por ano de construção?
  - Contar o número de IDs por ano de construção
2. Qual o menor número de quartos por ano de construção dos imóveis?(  
se tem uma tendência de aumentar o número de quartos por ano em média )

- Filtrar todos os imoveis por ano de construcao e selecionar o menor numero de quartos
3. Qual o preço de compra mais alto por cada numero de quarto?( se tem relaçao entre o numero de quartos e o preço do imovel )
- Filtrar todos os imoveis por numeor de quartos e selecionar o de maior preço
4. qual a soma de todos os preços de compra por numero de quartos?( se talvez os imoveis de 1 quarto sao mais valorizados em media que os de 3 ou 4 por exemplo )
- Filtrar todos imoveis por numero de quartos e somar todos os preços
5. Qual a soma de todos os preços de compra por numero de quartos e banheiros?( mesma ideia de cima so que com a variavel baneiro a mais )
- Filtrar todos os imoveis por num de quartos e por numero de banheiros e realizo a soma de todos os preços
6. Qual o tamanho medio das salas dos imoveis por ano de construcao?( se os tamanhos de sala foram aumentando ao longo dos anos )
- Filtrar todos os imoveis por ano de contruçao e fazer a media do tamanho de salas
7. Qual o tamanho mediano(mediana) das salas dos imoveis por ano de construcao?( vendo se a conclusao que tiramos na pergunta acima se confirma, qual o termo centralizado [esse é um conceito de estatistica, vamos ver em estatistica esploratoria 1] )
- Filtrar todos os imoveis por ano de contruçao e fazer a mediana do tamanho de salas
8. Qual o desvio-padroao do tamanho das salas dos imoveis por ano de construcao( o quao meus dados estao espalhados da media )?
- Filtrar todos os imoveis por ano de contruçao e fazer a desvio-padroao do tamanho de salas

9. Como é o crescimento medio de preços de compras dos imoveis, por ano, por dia, pela semana do ano?

- Filtrar todos os imoveis por ano e fazer um grafico onde o eixo x eu tenho o ano e o eixo y eu tenha a media do preço por ano( repito isso pro dia e semana do ano )
- Estudar uma biblioteca que tenha alguma função que desenhe um grafico de linhas



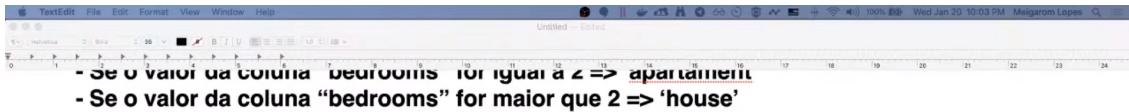
## ▼ Exercícios Práticos



### Novas perguntas do CEO para você:

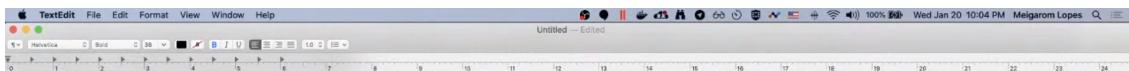
1. Crie uma nova coluna chamada: "dormitory\_type"
  - Se o valor da coluna "bedrooms" for igual à 1 => 'studio'
  - Se o valor da coluna "bedrooms" for igual a 2 => 'apartament'
  - Se o valor da coluna "bedrooms" for maior que 2 => 'house'
2. Faça um gráfico de barras que represente a soma dos preços pelo número de quartos.
3. Faça um gráfico de linhas que represente a média dos preços pelo ano construção dos imóveis.





- Se o valor da coluna "bedrooms" for igual a 2 => apartamento  
- Se o valor da coluna "bedrooms" for maior que 2 => 'house'

2. Faça um gráfico de barras que represente a soma dos preços pelo número de quartos.
3. Faça um gráfico de linhas que represente a média dos preços pelo ano construção dos imóveis.
4. Faça um gráfico de barras que represente a média dos preços pelo tipo dos dormitórios.
5. Faça um gráfico de linha que mostre a evolução da média dos preços pelo ano da reforma dos imóveis, a partir do ano de 1930.
6. Faça um tabela que mostre a média dos preços por ano de construção e tipo de dormitórios dos imóveis.



6. Faça um tabela que mostre a média dos preços por ano de construção e tipo de dormitórios dos imóveis.

7. Crie um Dashboard com os gráficos das questões 02, 03, 04 ( Dashboard: 1 Linha e 2 colunas )
8. Crie um Dashboard com os gráficos das perguntas 02, 04 ( Dashboard: 2 colunas )
9. Crie um Dashboard com os gráficos das perguntas 03, 05 ( Dashboard: 2 Linhas )
10. Faça um gráfico com o tamanho dos pontos sendo igual ao tamanho da sala de estar



- Dashboards são apresentações baseadas em dados, que contêm gráficos, informações,..., que são relevantes para uma apresentação de uma(s) ideia(s)

## ▼ Semana 4

- Planejamento dessa semana:

1. Recapilutando
2. Novas Perguntas de Negocio
3. Planejamento da soluçao
4. Estruturas de Dados - Listas
5. Estruturas de Controle - Condicionais
6. Estruturas de Controle - Laços

## ▼ Recapitulando

The screenshot shows a Mac OS X desktop with a TextEdit window open. The window title is "Untitled — Edited". The content of the window includes:

- 4. Estruturas de Dados - Listas**
- 5. Estruturas de Controle - Condicionais**
- 6. Estruturas de Controle - Laços.**

**1. Recapitulando**

Aula 01: Começando com o python  
Aula 02: Extração e Manipulação de Dados  
Aula 03: Transformação de Dados  
Aula 04: Estruturas de Controle

Aula 05:  
Aula 06:  
Aula 07:  
Aula 08:

— Próximo: 20% seguir estudando para Analista de Dados ( Python + SQL + Storytelling + ETL + Banco de Dados )  
— Próximo: 30% seguir estudando para Cientista de Dados ( Machine Learning + Estatística + Eng.Software )  
— Próximo: 50% vão desistir

3 Recados:

Recado 01: 50% de curso de Python do ZERO ao DS

Below the TextEdit window, there is a video player window showing a man with a beard and dark hair, wearing a black t-shirt, sitting in a black office chair. He appears to be speaking or presenting.

## ▼ Novas perguntas de Negócio

### ▼ Desafio Relembmando

<https://medium.com/@meigarom/os-5-projetos-de-data-science-que-fará-o-recrutador-olhar-para-você-c32c67c17cc9>

- Empresa : House Rocket

- Modelo de negocio : comprar imoveis por um preço mais baixo que tenha alto potencial de revenda(lucro = venda - compra)
- Qual desafio: Encontrar bons negócios, ou seja, encontrar casas com um preço baixo e com alto potencial de revenda

▼ Novas perguntas

1. Quantidade de imoveis por nivel?

- nivel 0 : Preço entre 0 e 321.950
- nivel 1 : Preço entre 321.950 e 450.000
- nivel 2 : Preço entre 450.000 e 645.000
- nivel 3 : Preço acima de 645.000

2. Adicione as seguintes informações ao imovel:

- O nome da rua
- O numero do imovel
- O nome do Bairro
- O nome da Cidade
- O nome do estado

3. Adicione o nivel do imovel no mapa como uma cor

4. Adicione o preço do imovel como o tamanho do ponto no mapa

5. Adicione opções de filtros para eu fazer minhas proprias analises:

- Eu Quero escolher visualizar imoveis com vista pra agua ou não
- Eu Quero escolher filtrar imoveis ate certo valor de preco

6. Adicionar opções de filtros no ultimo dashboard enviado:

- Eu quero visualizar somente valores a partir de uma data disponível para comprar.

▼ Planejamento da solução

1. Produto Final(o que vou entregar: Planilha, Grafico, Email, Modelo de ML,...)

- Email + 3 anexos:
  - Email : Perguntas | Respostas
  - Anexo 1 : Um arquivo .csv com novas informações requisitadas
  - Anexo 2 : um mapa com filtros requisitados
  - Anexo 3: um dashboard com filtros requisitados

2. Ferramenta(qual ferramenta eu vou usar?[se a empresa já tem essa ferramenta, se vai ter que comprar uma nova, usar um open source,...])

- Python 3.8.0
- Jupyter Notebook

3. Processo(Como vou fazer?)

1. Quantidade de imóveis por nível?

- nível 0 : Preço entre 0 e 321.950
- nível 1 : Preço entre 321.950 e 450.000
- nível 2 : Preço entre 450.000 e 645.000
- nível 3 : Preço acima de 645.000
  - Como fazer :

- Popular essa nova coluna vazia chamada nível com os valor condicionados aos intervalos
2. Adicione as seguintes informações ao imovel:
- O nome da rua
  - O numero do imovel
  - O nome do Bairro
  - O nome da Cidade
  - O nome do estado
  - Como fazer :
    - Onde tem essas informações? :
      - Tem no banco de dados da empresa?
      - Tem essas informações em uma API?
      - Dentro de uma pasta do meu colega de trabalho chamado Legilson?
      - ...
    - Qual dados eu tenho base que eu consiga fazer a conexao
    - Como anexar esse dado e como anexa-lo no conjunto de dados original
3. Adicione o nível do imovel no mapa como uma cor
- Adicionar a coluna "nível" como uma cor do mapa.

4. Adicione o preço do imovel como o tamanho do ponto no mapa

- Adicionar a coluna "preço" como um tamanho de ponto no mapa

5. Adicione opções de filtros para eu fazer minhas proprias analyses:

- Eu Quero escolher visualizar imoveis com vista pra agua ou nao
- Eu Quero escolher filtrar imoveis ate certo valor de preco

- Como fazer :

- Encontrar uma biblioteca que tenha funções que permitam colocar filtros no jupyter Notebooks
- Entender o funcionamento da biblioteca e adicionar no jupyter notebook

6. Adicionar opções de filtros no ultimo dashboard enviado:

- Eu quero visualizar somente valores a partir de uma data disponivel para comprar.

- Como fazer :

- Encontrar uma biblioteca que tenha funções que permitam colocar filtros no jupyter Notebooks

- Entender o funcionamento da biblioteca eadicionar no jupyter notebook

▼ As estruturas de Dados em python

- Listas
- Tuples - Falar em uma aula mais na frente ( SQL + Python )( quando a gente usa SQL e requere alguma coisa ele gera um tuple de saida )
- Dicionarios( aula 3 )
- DataFrames ( aula 3 )

▼ Listas

- Possui valores, tanto numericos quanto categoricos
- Os valores são mapeados por posição

▼ Estruturas de Controle - Condicionais

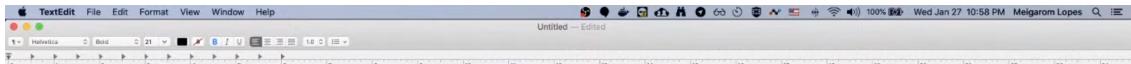
- Condicional permite selecionar linhas e colunas ate que a condiçao seja satisfeita.
- Selecionar linhas e colunas combinando condições.

***O conceito de API nada mais é do que uma forma de comunicação entre sistemas. Elas permitem a integração entre dois sistemas, em que um deles fornece informações e serviços que podem ser utilizados pelo outro, sem a necessidade de o sistema que consome a API conhecer detalhes de implementação do software.***

▼ Estruturas de Controle - Laços

.json ⇒ dicionario do formato ( 'chave': valor )

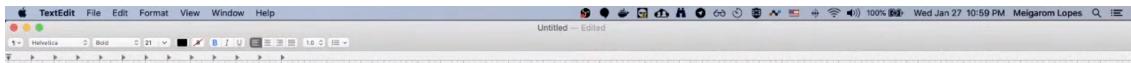
▼ Exercicios atividade



## 7. Exercícios:

Novas perguntas do CEO para você:

1. Qual a média do preço de compra dos imóveis por "Nível"?
  - Nível 0 -> Preço entre R\$ 0 e R\$ 321.950
  - Nível 1 -> Preço entre R\$ 321.950 e R\$ 450.000
  - Nível 2 -> Preço entre R\$ 450.000 e R\$ 645.000
  - Nível 3 -> Acima de R\$ 645.000
2. Qual a média do tamanho da sala de estar dos imóveis por "Size" ?
  - Size 0 -> Tamanho entre 0 e 1427 sqft
  - Size 1 -> Tamanho entre 1427 e 1910 sqft
  - Size 2 -> Tamanho entre 1910 e 2550 sqft
  - Size 3 -> Tamanho acima de 2550 sqft
3. Adicione as seguinte informações ao conjunto de dados original:
  - Place ID: Identificação da localização
  - OSM Type: Open Street Map type
  - Country: Nome do País
  - Country Code: Código do País
4. Adicione os seguinte filtros no Mapa:
  - Tamanho mínimo da área da sala de estar.
  - Número mínimo de banheiros.
  - Valor Máximo do Preço.
  - Tamanho máximo da área do porão.
  - Filtro das Condições do Imóvel.
  - Filtro por Ano de Construção.
5. Adicione os seguinte filtros no Dashboard:
  - Filtro por data disponível para compra.
  - Filtro por ano de renovação.
  - Filtro se possui vista para a água ou não.



## 7. Exercícios:

Novas perguntas do CEO para você:

1. Qual a média do preço de compra dos imóveis por "Nível"?
  - Nível 0 -> Preço entre R\$ 0 e R\$ 321.950
  - Nível 1 -> Preço entre R\$ 321.950 e R\$ 450.000
  - Nível 2 -> Preço entre R\$ 450.000 e R\$ 645.000
  - Nível 3 -> Acima de R\$ 645.000
2. Qual a média do tamanho da sala de estar dos imóveis por "Size" ?
  - Size 0 -> Tamanho entre 0 e 1427 sqft
  - Size 1 -> Tamanho entre 1427 e 1910 sqft
  - Size 2 -> Tamanho entre 1910 e 2550 sqft
  - Size 3 -> Tamanho acima de 2550 sqft
3. Adicione as seguinte informações ao conjunto de dados original:
  - Place ID: Identificação da localização
  - OSM Type: Open Street Map type
  - Country: Nome do País
  - Country Code: Código do País
4. Adicione os seguinte filtros no Mapa:
  - Tamanho mínimo da área da sala de estar.
  - Número mínimo de banheiros.
  - Valor Máximo do Preço.
  - Tamanho máximo da área do porão.
  - Filtro das Condições do Imóvel.
  - Filtro por Ano de Construção.
5. Adicione os seguinte filtros no Dashboard:
  - Filtro por data disponível para compra.
  - Filtro por ano de renovação.
  - Filtro se possui vista para a água ou não.



## ▼ Semana 5

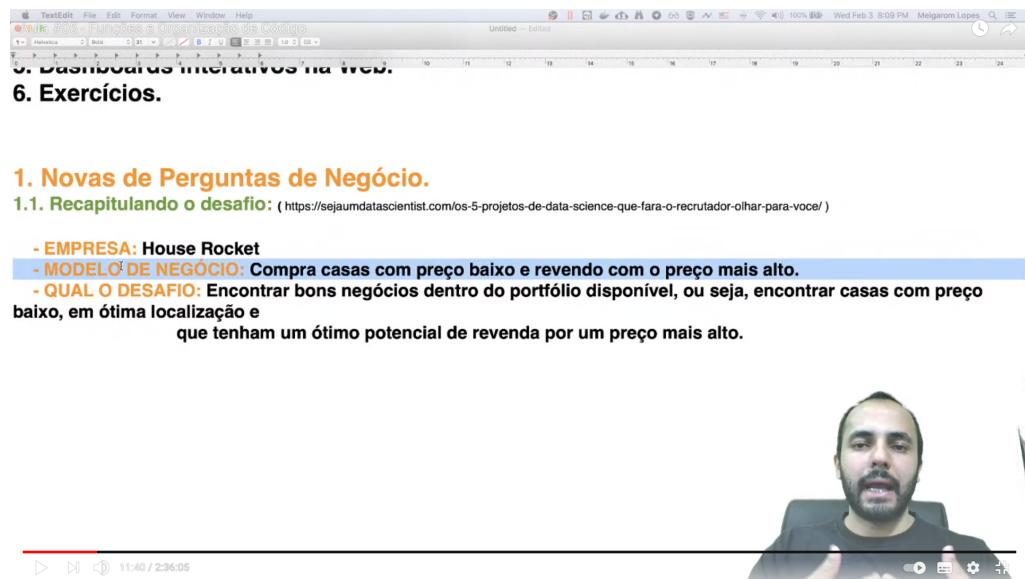
- Planejamento dessa semana

### 1. Novas Perguntas de Negocio

2. Planejamento da Solução
3. A diferença entre um código amador e um profissional
4. O que são funções e por que são úteis
5. Dashboards interativos na Web
6. Exercícios Atividade

▼ Novas perguntas de Negócio

▼ Recapitulando



1. Novas de Perguntas de Negócio.

1.1. Recapitulando o desafio: (<https://sejaumdatascientist.com/os-5-projetos-de-data-science-que-fara-o-recrutador-olhar-para-voce/>)

- **EMPRESA:** House Rocket
- **MODELO DE NEGÓCIO:** Compra casas com preço baixo e revenda com o preço mais alto.
- **QUAL O DESAFIO:** Encontrar bons negócios dentro do portfólio disponível, ou seja, encontrar casas com preço baixo, em ótima localização e que tenham um ótimo potencial de revenda por um preço mais alto.

▼ Novas Perguntas de Negócio

1. Não consegui usar o mapa que você me enviou por email
2. Preciso acessar o mapa e o dashboard do meu celular

▼ Planejamento da solução

1. Produto final( o que vou entregar efetivamente? )
  - Um Link( URL ) <https://www.houserocket.dashboard.com.br>
  - Informações importantes

- Mapa interativo
  - dashboard interativo
2. Ferramenta que vamos usar
    - Python 3.8.0
    - Jupyter notebook
    - Pycharm
  3. Processo
    1. Organizar meus codigos em funções
    2. Pesquisar uma biblioteca com funções para criar mapas que podem ser acessados pelo Browser

▼ A diferença de um código amador para um profissional

1. Profissional Junior x Profissional Senior
  1. Junior:
    - Faz código macarrônico ( código macarrão, não tem nenhum ponto de começo e fim, o código é todo um fluxo só ).
    - Faz código só pra ele, ninguém mais entende o código dele
    - Não usa as estruturas de dados corretamente e acha que só existe dataframe.
    - Não usa funções.
    - Não organiza o código de forma lógica e simples ( a lógica só na cabeça dele )
  2. Senior:
    - Faz código modular.
    - Faz código para seu time e para a sua empresa.

- Sabe exatamente quando usar as estruturas de dados.
- Usar funções para modularizar e escalar seu código, mas com sabedoria.
- Seus códigos parecem livros( fáceis de serem lidos, interpretados, reutilizáveis )
- Sabe usar multiThread para resolver problema, que é uma técnica parecida com a ideia de hashtable, ele vai separar tudo em várias listas e resolver todas e depois juntar, o que fica bem mais rápido
- Usa ETL como padrão de código, pois fica bem mais legível( ver código no repositório [anot aula 5] )

▼ O que são funções e pq são úteis

- É mostrado no código uma estrutura ETL que serve muito para organização do nosso código, encapsulando tudo em funções

▼ Dashboards interativos na Web

- Streamlit
- Serve para deixarmos alguma coisa em produção, seja o nosso dashboard ou o nosso mapa por ex
- Vamos ver melhor na aula que vem

▼ Exercícios Atividade

- não tem exercícios essa semana
- Recado → sempre focar em resolver problemas e não em fazer um dashboard mais bonito, no final o que interessa são os insights para tomada de decisões

▼ Semana 6

- Planejamento dessa semana

1. Requisição do CEO
2. Planejamento da solução
3. Criação de Dashboards Web

▼ Requisição do CEO

- Gostaria de chegar na minha mesa e ter um lugar único onde eu possa observar o portifólio da House Rocket. Nesse portifólio, eu tenho interesse:
  1. Filtros dos imóveis por um ou várias regiões
  2. Escolher um ou mais variáveis para visualizar
  3. Observar o número total de imóveis, a média de preço, a medida da sala de estar e também a média do preço por metro quadrado em cada um dos códigos postais
  4. Analisar cada uma das colunas de um modo mais descrito
  5. Um mapa com densidade de portifólio por região e também por densidade de preço
  6. Checar a variação anual de preço
  7. Checar a variação diária do preço
  8. Conferir a distribuição por:
    - preço
    - número de quartos
    - número de banheiros
    - número de andares
    - Vista para água ou não
- Ou seja, o CEO quer um lugar que ele possa ver tudo isso, um dashboard (conjunto de gráficos [cada gráfico compõe por uma métrica e um resultado] que tem um propósito comum)

- Uniao de dashboards é chamado Cubo, ou seja, varias tabelas e dashboards baseados em varios tipos de entidade que todos apontam para um objetivo especfico

▼ Planejamento da soluçao

1. Produto final

- Um link para acessar o dashboard

2. Ferramentas de uso

- Python 3.8.0

- PyCharm

- Importante falar que temos duas fases na entrega de alguma coisa:

1. Analise exploratoria dos dados

2. Modelo de produção

- Na primeira a gente usa geralemnte o jupyter notebook, pois ele é mais facil visualizar as coisas separadamente, mas na segunda nos geralmente temos que incrementar nosso codigo em um projeto ja existente que nao aceita jupyter(pois é um html,...), entao precisamos na verdade criar um arquivo .py que vai unificar toda nossa analise e gerar os resultados

3. Processo

1. Filtros dos imoveis por um ou varias regioes:

- Objetivo: Visualizar imoveis por codigo postal
- Ação do usuario: Digitar um ou mais codigos desejados
- A visualizaçao( a mensagem que eu quero passar ): uma tabela com todos os atributos e filtrada por codigo postal

- Explicação rápida (ajuda no storytelling):
  - Tabela ⇒ Exploração e Comparação
  - Linhas ⇒ Crescimento ou Decaimento
  - Barras ⇒ Comparar proporções
  - Pizza ⇒ Até duas comparações é legal trabalhar com pizza
  - Tem outros tipos de graficos que vamos ver mais a frente, mas por hora é só, vemos em estatística isso

2. Escolher um ou mais variáveis para visualizar:

- Objetivo: Visualizar as características do imóvel
- Ação do usuário: Digita as características desejadas
- A visualização (a mensagem que eu quero passar): Uma tabela com todos os atributos selecionados

3. Observar o número total de imóveis, a média de preço, a medida da sala de estar e também a média do preço por metro quadrado em cada um dos códigos postais:

- Objetivo: visualizar as médias de algumas métricas por região
- Ação do usuário: digita as métricas desejadas
- A visualização (a mensagem que eu quero passar): uma tabela com todos os atributos selecionados

4. Analisar cada uma das colunas de um modo mais descrito (média, mediana, ..., estatística de primeira ordem):

- Objetivo: Visualizar métricas descritivas de cada um dos atributos escolhidos
- Ação do usuário: digitar as métricas desejadas

- A visualizaçao( a mensagem que eu quero passar ): uma tabela com metricas descritivas por atributo

5. Um mapa com densidade de portifolio por regiao e tambem por densidade de preço:

- Objetivo: Visualizar a densidade do portifolio no mapa( número de imoveis por regiao )( e tambem por preço naquela regiao )
- Ação do usuario: Nenhuma ação
- A visualizaçao( a mensagem que eu quero passar ): Um mapa com densidade de imoveis por regiao

6. Checar a variaçao anula de preço:

- Objetivo: Observar variações anuais de preços
- Ação do usuario: Filtra os dados pelo ano
- A visualizaçao( a mensagem que eu quero passar ): Um grafico de linha com os anos em x e preços medios em y

7. Checar a variaçao diaria do preço:

- Objetivo: Observar variações diárias de preços
- Ação do usuario: Filtra os dados por dia
- A visualizaçao( a mensagem que eu quero passar ): Um grafico de linha com os dias em x e preços medios em y

8. Conferir a distribuiçao por:

- preço
- numero de quartos
- numero de banheiros
- numero de andares
- Vista para agua ou nao
- Objetivo: Observar a concentraçao dos imoveis por preço, quartos, banheiros e andares

- Ação do usuário: Filtro de preço, quarto, banheiro, andar
- A visualização( a mensagem que eu quero passar ): Um histograma com cada atributo definido

#### ▼ Criação de Dashboards Web

- Diferença do merge e do concat em dataframes:
  - O merge garante que um valor de uma coluna que escolhemos para dar merge seja exatamente igual a outra, enquanto isso no concat a ordem pode ser diferente de zipcodes por ex e ele concatenar errado

#### ▼ Semana 7

- Planejamento dessa semana
  1. Recapitulando as requisições do CEO
  2. Repassando pelo planejamento do dashboard
  3. Finalização do dashboard web

#### ▼ Recapitulando as requisições

1. Checar a variação anual de preço
2. Checar a variação diária do preço
3. Conferir a distribuição por:
  - preço
  - número de quartos
  - número de banheiros
  - número de andares
  - Vista para água ou não
4. Ter autonomia para fazer as minhas próprias análises, através de filtros

## ▼ Planejamento da solução

### 1. Produto final

- Um link para acessar o dashboard

### 2. Ferramentas de uso

- Python 3.8.0
- PyCharm
- Importante falar que temos duas fases na entrega de alguma coisa:
  1. Analise exploratoria dos dados
  2. Modelo de produçao

### 3. Processo( Como fazer? ):

#### 1. Checar a variaçao anual de preço:

- Objetivo: Observar variações anuais de preços
- Ação do usuário: Filtra os dados pelo ano
- A visualização( a mensagem que eu quero passar ): Um gráfico de linha com os anos em x e preços médios em y

#### 2. Checar a variaçao diaria do preço:

- Objetivo: Observar variações diárias de preços
- Ação do usuário: Filtra os dados por dia
- A visualização( a mensagem que eu quero passar ): Um gráfico de linha com os dias em x e preços médios em y

#### 3. Conferir a distribuiçao por:

- preço
- numero de quartos
- numero de banheiros
- numero de andares

- Vista para agua ou nao
    - Objetivo: Observar a concentração dos imóveis por preço, quartos, banheiros e andares
    - Ação do usuário: Filtro de preço, quarto, banheiro, andar
    - A visualização( a mensagem que eu quero passar ): Um histograma com cada atributo definido
  - Site massa pra ver streamlit's mt bonitos ⇒ <https://awesome-streamlit.org>
  - Importante fazer essas ressalvas:
    - Streamlit é uma ferramenta importantíssima para mostrar o que foi feito com as bibliotecas em Python
    - Várias empresas têm vários streamlit's para acompanhar os diferentes ramos dela, oq eu estavam fazendo nesse curso é um que talvez seria o streamlit que o CEO viria, um geral de como está a empresa e seu lucro,....
    - Ja com esse conhecimento da para ser Freelancer e fazer uns bicos de fazer análises para pessoas e ate fazer um serviço mensal que a pessoa paga pra você ficar mantendo aquele conjunto de dados do negócio dela
    - Ou seja, STREAMLIT É sinônimo de dinheiro
    - O conceito pode parecer complicado, mas na verdade é bem simples. Essa espécie de “garimpo” da internet envolve extrair informações relevantes de determinado site para depois serem analisadas. Esses dados serão usados para aprimorar a tomada de decisões com maior chance de acerto e sucesso
    - Então, fazemos um web scraping e vendemos o resultado para aquele cliente
- ▼ Finalização do dashboard web
- Fizemos o deploy pelo heroku nessa aula ⇒ <https://house-rocket-dashanalysis.herokuapp.com/>

- Temos como usar outro cloud, vai de cada um
- em uma situação real, não usamos o banco de dados junto da aplicação

## ▼ Semana 8

- Planejamento da Semana:
  - As etapas de um projeto de Ciência de Dados
  - Transformando o Python do Zero ao DS em um projeto de Portifólio
  - Tarefa de casa
  - Recapitulando o que foi aprendido até agora
  - Comunidade DS

## ▼ As etapas de um projeto de Ciência de Dados

1. Questão de negócios → O problema que surge e que podemos resolver
2. Entendimento do negócio →
  - Parar um tempo e analisar tudo daquele problema e levantar o real problema, a que as pessoas não são muito boas para deixar realmente lógico seu problema
  - Planejar como vai ser a possível solução diante o que você sabe das tecnologias que você acha que vai usar
3. Coleta de Dados → Coletar os dados usando SQL, ...
4. Limpeza de Dados → Limpar os dados que geralmente vem com informações a mais do que realmente queremos (por ex: filtrar linhas e colunas para trabalhar com aquilo que queremos para o problema)
5. Exploração de dados →
  - Criar hipóteses de negócio
  - Conseguir ver qual o impacto daqueles dados em uma modelagem de machine learning
  - Nessa fase temos como principal coisa ver de novo o que vamos fazer, qual modelo de machine learning vamos usar e se até é

necessario usar algum algoritmo

6. Modelagem de dados →

- Modelar os dados para os algoritmos de machine learning consigam tirar alguma conclusao daqueles dados

7. Aplicaçao dos algorimos de ML →

- Aplicar o algoritmo

8. Avaliaçao da performance dos modelos →

- Analisar, se possivel com o time de negocio, se ja é satisfatorio aqueles resultados, podendo usar aqueles modelos que criamos, e podemos passar para a produçao

9. Publicaçao( deixar em produçao ) →

- Deixar em cloud o nosso projeto para que as pessoas que precisam dele possam ver, por exemplo no heroku

▼ O Projeto do tipo insgihts:

▼ Passo a passo:

1. Questao de negocio

- Pergunta que temos que responder

2. Entendimento do negocio

- Pegar as perguntas e Quebrar em 3 etapas:

    1. Produto final

    2. Ferramentas que preciso pra fazer esse produto

    3. Qual o processo que imagino pra resolver esse problema

3. Coleta de Dados

- Nesse curso nao tivemos problemas nessa parte, pois so foi pegar os dados do kaggle

4. Limpeza de Dados

- Tirar todas as colunas e linhas que nao nos servem

## 5. Exploração de Dados

- Gerar graficos, mapas de distribuição, tabelas

### ▼ Objetivo do projeto do tipo Insights

1. Objetivo : Gerar insight através da análise e manipulação dos dados para auxiliar a tomada de decisão pelo time de negócios

2. Etapas:

1. Questão de Negócio:

- Quais são os imóveis que a House Rocket deveria comprar e por qual preço
- Uma vez comprado, qual o melhor momento (do ano) para vendê-lo e por qual preço?

2. Entendimento do Negócio:

- Produto final( o que realmente vou entregar? R: Planilha, Modelo ML, Email ):
  - 2 Relatórios:
    - Relatório com as sugestões de compra de apartamento por valor recomendado
    - Relatório com as sugestões de venda de um imóvel por um valor recomendado
  - Qual ferramenta vou usar( o que preciso para realizar esse projeto? A empresa precisa comprar algo? ) :
    - Python 3.8.0
    - PyCharm
    - Jupyter Notebook
  - Processo( quais são os passos necessários para alcançar meu objetivo? )( ai entra o pensamento analítico →

capacidade de pegar um problema aberto e transformar em tarefas que possam ser executaveis ):

1. Quais sao os imoveis que a House Rocket deveria comprar e por qual preço

- Plano 1:
  - Coletar os dados do site do Kaggle
  - Agrupar os dados por regiao( zipcode )
  - Dentro de cada regiao, eu vou encontrar a mediana do preço dos imóveis
  - Vou sugerir que os imoveis que estao abaixo do preço mediano de cada regiao e que estejam em boas condições sejam comprados
- Exemplo:
  - Imovel codigo | Regiao | Preço do imovel | Preço da mediana | Condição | Status
  - 10330 | 302349 | 450000 | 500000 | 3 | compra
  - 10335 | 302349 | 750000 | 500000 | 3 | nao compra
  - 10345 | 302349 | 150000 | 500000 | 1 | nao compra

10345	302349   R\$ 150,000,00   R\$ 500.000,00   1   Não Compra
-------	---

```
df['price_median'] = data[['zipcode', 'price']].groupby('zipcode').median().reset_index()
df2 = pd.merge( data, df, on='zipcode', how='inner')

for i in range( len( df2 ) ):
    if ( df2.loc[i, 'price'] < df2['price_median'] ) & ( df2['condition'] >= 2 ):
        df2['status'] = 'compra'
    else:
        df2['status'] = 'não compra'
```

2. Uma vez o imóvel comprado, qual o melhor momento para vendê-lo e por qual preço



2. Uma vez comprado, qual o melhor momento ( do ano ) para vende-lo e por qual preço?

- Plano 1:
  - Como os dados ja tratados e organizados.
  - Agrupar os imoveis por regiao( zipcode ) e por sazonalidade( Summer, winter )
    - Hipotese: imoveis no inverno voa venda e mais barato
  - Dentro de cada regiao e sazonalidade, eu vou calcula a mediana de preço
  - Condições de venda :
    - Se o preço da compra for maior que a mediana da regiao + sazonalidade :
      - Preço da venda sera igual ao preço da compra + 10%
    - Se o preço da compra for menor que a mediana da regiao + sazonalidade :
      - Preço da venda sera igual ao preço da compra + 30%
- Exemplo:
  - Imovel codigo | Regiao | Temporada | Preço da mediana | Preço da Compra | Preço da venda | Lucro
  - 10330 | 302349 | Verao | 800000 | 450000 | 450000 + 30% | ???
  - 10330 | 302349 | Verao | 800000 | 400000 | 450000 + 10% | ???
  - Todos esses numeros a gente emcontra com algoritmos de machine learning, os otimos de compra e venda das coisas

### 3. Coleta e limpeza dos dados:

- Coletar os dados do site do kaggle
- Limpeza dos dados
  - Remover datas erradas
  - Remover outliers devido à erro do sistema( os que fogem mt do padrao )( isso é um assunto extenso, pois nao é todo valor que foge do padrao que é um outlier, ja que em black friday por exemplo as vendas explodem, e isso faz parte do fenomeno )( Isso nos descobrimos se é um outlier ou ao com machine learning geralmente )

### 4. Exploração de dados( EDA ):

- Descobrir Insights para o time de negocio.
- Explorar os dados para identificar o impacto dos atributos nos algoritmos de ML. ( Qual algoritmo vai performar melhor naquele conjunto de dados )
- O que sao Insights para o negocio ?
  - Insights sao descobertas, atraves dos dados, que sao inesperadas pelas pessoas
  - Insights precisa ser açãoavel( no dia seguinte ja da para fazer algo concreto em relação a sua analise ), caso contrario, ele é apenas uma curiosidade
  - informações ja conhecidas e nao esperadas sao apenas curiosidades
  - Tomar muito cuidado com isso, pois nao somos pagos para falar de curiosidades, somos pagos para trazer insights e possíveis ações concretas para as pessoas tomar com base nas nossas análises
- Exemplo:

- Durante o periodo de natal, vende-se mais casas que na pascoa( Descoberta apenas )
- Imoveis com porao sao maiores que imoveis sem porao( Descoberta apenas )
- Imoveis com porao sao 40% mais caros que outros imoveis( Acionavel, da para tomar uma acao baseado nesse Insight )( Acao possivel: Comprar uma casa sem porao e fazer um porao nela )
- Como fazer para cirar as hipoteses
  - Todas as hipoteses de negocio precisa ter 3 caracteristicas
    1. Precisa ser uma AFIRMAÇÃO
    2. Precisa ser uma comparação entre 2 variaveis
    3. Voce precisa definir um valor base( uma porcentagem por ex )
  - Responder essas 5 hipoteses e criar mais 5

H1: Imoveis que possuem vista para agua, sao 30% mais caros, na média. ( Nao pode ser pergunta, pq a resposta vai ser sim ou nao, e nao um sim e nao mais um pq dessa resposta )

H2: Imoveis com data de construção menor que 1955, sao 50% mais baratos, na media( comparando data com preço )

H3: Imoveis sem porao possuem area total( sqft\_lot ), sao 40% maiores do que os imoveis com porao

H4: O crescimento do preço dos imoveis YoY( Year over Year )( jan 2019 com jan 2020 por ex) é de 10%

H5: Imoveis com 3 banheiros tem um crescimento medio de MoM( month over month )( jan com fev com mar com abril ,... ) de 15%.

H6:....

H7:...

H8:...

▼ Transformando o Python do Zero ao DS em um projeto de Portifolio:

▼ O que é obrigatorio em um projeto de portifolio:( isso vai explicado no readme do github )

1. Questao de negocio:

- O que você quer resolver?

2. Premissas do negocio:

- Metricas que foram alteradas para Acontecer o projeto( casas ficam cheias de neve no inverno )
- O que considerei como Outliers no meu projeto( valores acima de ... sao considerados erros humanos )
- O que levo de consideraçao para iniciar o projeto, o escopo

3. Planejamento da soluçao:

- Como acho que vou resolver o problema no codigo

4. Os 5 principais insights de negocio:

- Insights comprovados e acionaveis a aprtir da analise dos dados

5. Resultados financeiros para o negocio:

- O lucro mais o menos que a empresa vai ter se cumprir com nossa analise e aplica-la

6. Conclusao:

- Seu objetivo inicial foi alcançado?

## 7. Proximos passos;

- O que pode ser melhorado no meu projeto?
- Que coisas o implementariam?

### ▼ Tarefa de Casa

1. Criar visualizações para responder cada uma das 10 hipóteses de negocio
2. Construir um tabela com as recomendações de compra ou não compra
3. Construir uma tabela com recomendações de venda com acréscimo de 10 ou 30 %
4. Fornecer as hipóteses e as tabelas no Streamlit
5. Transformar o projeto do curso de Python do ZERO ao DS em um projeto de portifólio
6. Salvar códigos dentro do github
7. Escrever o README com os requisitos obrigatórios para um portifólio de projetos

### ▼ Recapitulando o que aprendemos até agora

- Aula 1:
  - O que é um comando, interpretador, baixando IDE
- Aula 2:
  - Como selecionar linhas e filtrar colunas
- Aula 3:
  - Criar colunas novas, fazer combinação de colunas para gerar outra
- Aula 4:
  - Como funciona e usa estrutura de controles
- Aula 5:

- Como deixar o código mais modular, padrão de código ETL
- Aula 6:
  - Tipos de visualização de dados, quais são as melhores opções para cada cenário
- Aula 7:
  - Dando deploy em um projeto usando Streamlit
- Aula 8:
  - O que é e como transformar nosso projeto em Projeto de portfólio
  - O novo problema de negócios
  - Extração de dados usando BeautifulSoup
- Curso DS ao DEV:
  - Resolução de Exercícios do Python do ZERO ao ds
  - Instalação de Ferramentas
  - O novo problema de negócios
  - Desenvolvimento do pensamento analítico( pegar problema abstrato e transformar em soluções concretas )( pegar problemas maiores me menores e usar dados pra isso )
  - Extração de dado do HTML usando BeautifulSoup
  - Extração de dados do HTML de modo assíncrono( usar vários códigos rodando ao mesmo tempo para extrair dados e juntar depois tudo )
  - Banco de Dados e SQL com Python e salvar os dados em uma tabela
  - Ambiente virtual em Python( jeito que programamos de verdade em uma empresa )
  - Agendamento automático de tarefas( como fazer ficar automático essa extração acima e salvar no banco de dados )

- Primeira Rodada de Perguntas de negocio do CEO e analise
  - Extraçao de dados do HMTL via Selenium( pois tem uma particularidade de alguns sites quando vamos pegar os dados HTML dele, o selenium resolve isso pra gente )
  - Segunda rodada de perguntas de negocio do CEO e analise
  - Storytelling
- 
- Usar metodo ciclico( repetir estrutura focada na solução )é importante para reforçar o que aprendemos e ver resultados para não desanimar daquilo que estamos vendo