

# 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 contains values 0 and 0.626386. The second column contains values 1.52325 and null. An arrow points from the text "axis=1" to the second column, indicating that axis=1 aggregates along the columns. Another arrow points from the text "axis=0" to the first column, indicating that axis=0 aggregates along the rows.

- 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'

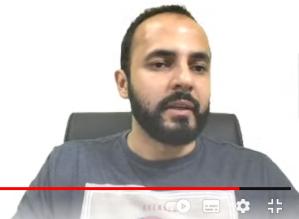
**3. Crie uma nova coluna chamada: "condition\_type"**

- 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?**

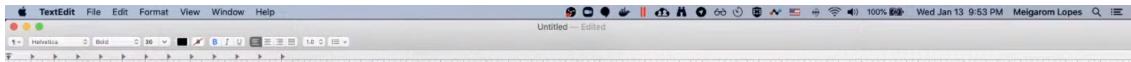
**9. Qual a data mais antiga de renovação de um imóvel?**

**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**





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 para água” ?

13. Quantos imóveis estão com a condição igual a “good” e são “new\_house”? 

I  
14. Qual o valor do imóvel mais caro do tipo “studio” ?

15. Quantos imóveis do tipo “apartment” foram reformados em 2015 ?

16. Qual o maior número de quartos que um imóveis do tipo “house” possui ?

17. Quantos imóveis “new\_house” foram reformados no ano de 2014?

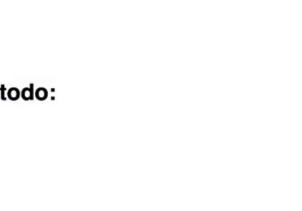
18. Selecione as colunas: “id”, “date”, “price”, “floors”, “zipcode” pelo método:

10.1. Direto pelo nome das colunas.

10.2. Pelos Índices.

10.3. Pelos Índices das linhas e o nome das colunas

10.4. Índices Booleanos

15. Quantos imóveis do tipo “apartment” foram reformados em 2015 ? 

16. Qual o maior número de quartos que um imóveis do tipo “house” possui ?

17. Quantos imóveis “new\_house” foram reformados no ano de 2014?

18. Selecione as colunas: “id”, “date”, “price”, “floors”, “zipcode” pelo método:

10.1. Direto pelo nome das colunas.

10.2. Pelos Índices.

10.3. Pelos Índices das linhas e o nome das colunas

10.4. Índices Booleanos

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” 