KLASA
By Rajut Indonesia

Module 3

Data Analytics 1.0

Objectives

- How to start exploratory data analysis
- How to present insight through data storytelling
- Important skill for data analysts

HOW TO START
EXPLORATORY DATA ANALYSIS

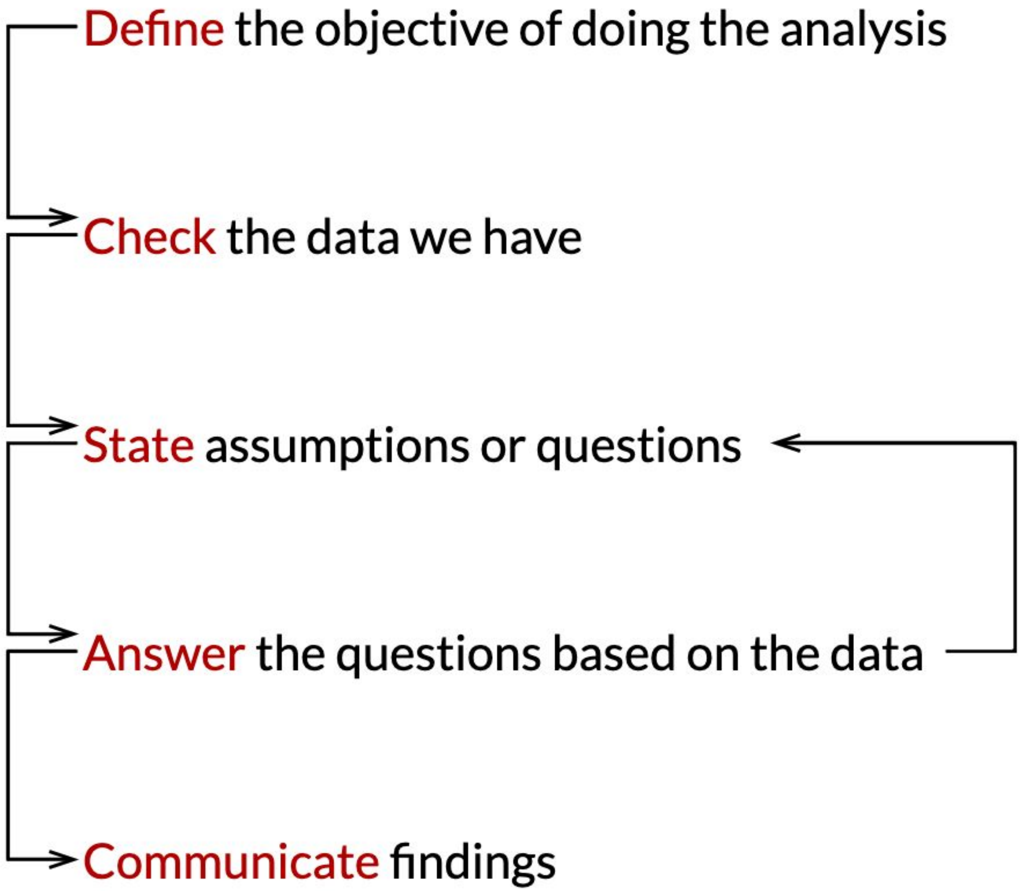
EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is an activity where we explore data to gain a deep understanding of

- **The properties of the data**
  schema, data types, statistical properties, etc
- **The quality of the data**
  missing values, inconsistent data types, etc
- **The relationship between variables**

EDA can be used in business analytics to answer business question and give insights or recommendation which will be used for strategic decision making

EDA WORKFLOW



Define the objective of doing the analysis

- Objective usually comes from the stakeholders
- Usually in a form of open question (e.g : how can we improve sales, why is there 50% drop of revenue this month, etc)

Check the data we have

- Identify areas that we can explore
- Indicate whether we track enough data or not

State assumptions or questions

- Can be from stakeholders or analyst
- Limited based on the data we have
- Usually in a form of specific questions (e.g : does giving vouchers leads to sales improvement, how many users ended their subscription this month, etc)
- Will be performing univariate or bivariate analysis

Answer the questions based on the data

Communicate findings

- Give conclusion and recommendation

CALIFORNIA HOUSING

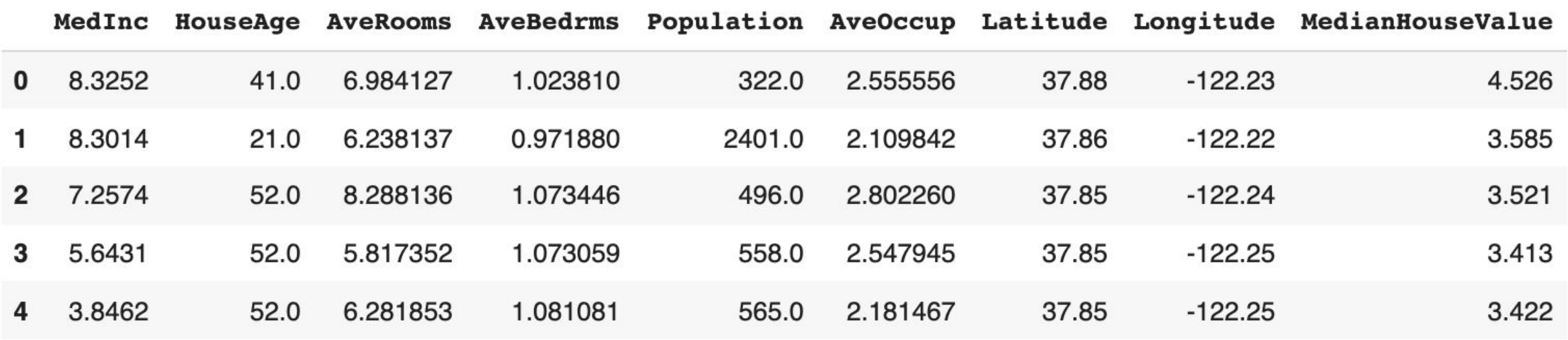
Defining the objective of doing the analysis

Let’s do an EDA exercise using California Housing Dataset and define the objective as :

What are factors that influence the house pricing in California?

CALIFORNIA HOUSING

Check the data we have



	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude	MedianHouseValue
0	8.3252	41.0	6.984127	1.023810	322.0	2.555556	37.88	-122.23	4.526
1	8.3014	21.0	6.238137	0.971880	2401.0	2.109842	37.86	-122.22	3.585
2	7.2574	52.0	8.288136	1.073446	496.0	2.802260	37.85	-122.24	3.521
3	5.6431	52.0	5.817352	1.073059	558.0	2.547945	37.85	-122.25	3.413
4	3.8462	52.0	6.281853	1.081081	565.0	2.181467	37.85	-122.25	3.422

Data Description

- This dataset was derived from the 1990 U.S. census, using one row per census block group
- A block group is the smallest geographical unit for which the U.S. Census Bureau publishes sample data
- A block group typically has a population of 600 to 3,000 people

- **Number of Instances** : 20640
- **Number of Attributes** : 8 numeric, predictive attributes and the target

- **Attribute Information:**

MedInc	median income in block
HouseAge	median house age in block
AveRooms	average number of rooms
AveBedrms	average number of bedrooms
Population	block population
AveOccup	average house occupancy
Latitude	house block latitude
Longitude	house block longitude
MedianHouseValue	median house value

CALIFORNIA HOUSING

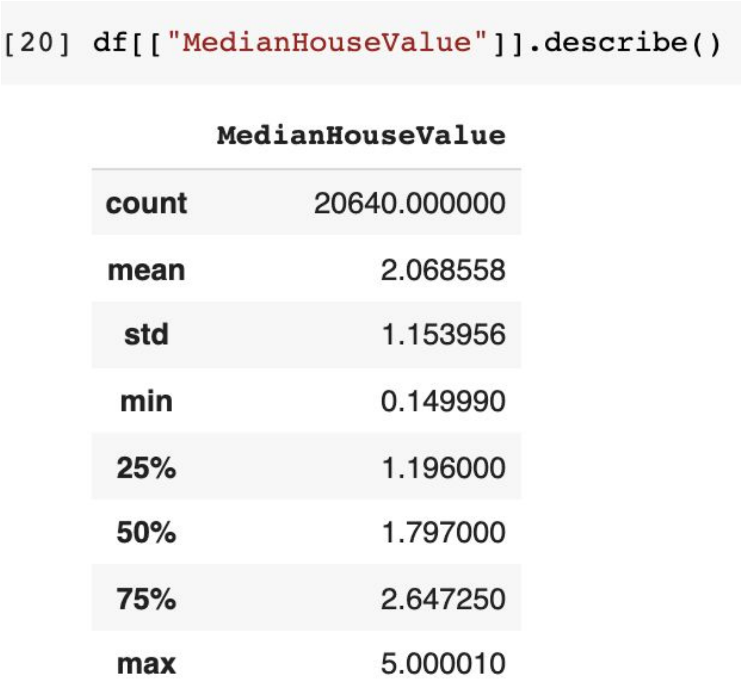
State assumptions or questions

1. What is the distribution of the house pricing looks like?
2. The older the house age, the lower the price
3. The more populated the house block, the more expensive the price is
4. Is there certain areas with more expensive price than the other?

**1 is univariate analysis, 2-4 is bivariate analysis*

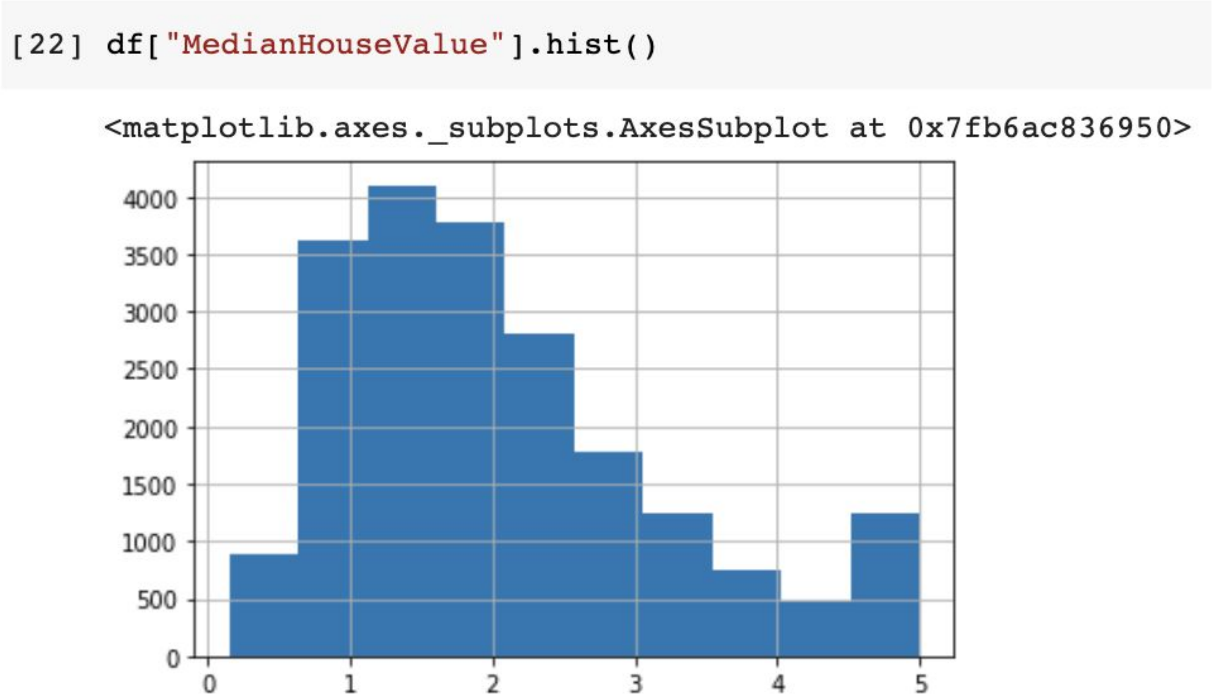
CALIFORNIA HOUSING

1. What is the distribution of the house pricing looks like?



```
[20] df[["MedianHouseValue"]].describe()
```

	MedianHouseValue
count	20640.000000
mean	2.068558
std	1.153956
min	0.149990
25%	1.196000
50%	1.797000
75%	2.647250
max	5.000010



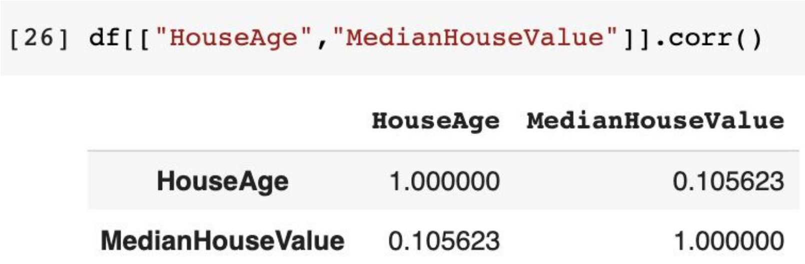
```
[22] df["MedianHouseValue"].hist()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fb6ac836950>
```

- What we try to do here is **univariate analysis** where we only look at the price variable
- There are several things to measure data distribution, we can **present several stats** like mean, median, percentiles, etc. or we can use **visualization**

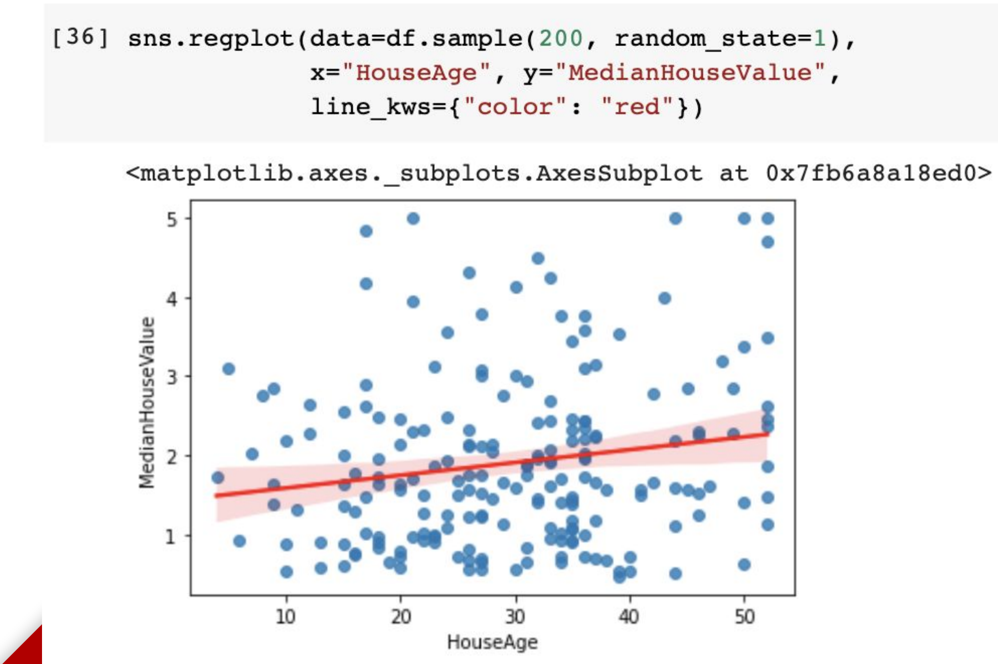
CALIFORNIA HOUSING

2. The older the house age, the lower the price



```
[26] df[["HouseAge","MedianHouseValue"]].corr()
```

	HouseAge	MedianHouseValue
HouseAge	1.000000	0.105623
MedianHouseValue	0.105623	1.000000



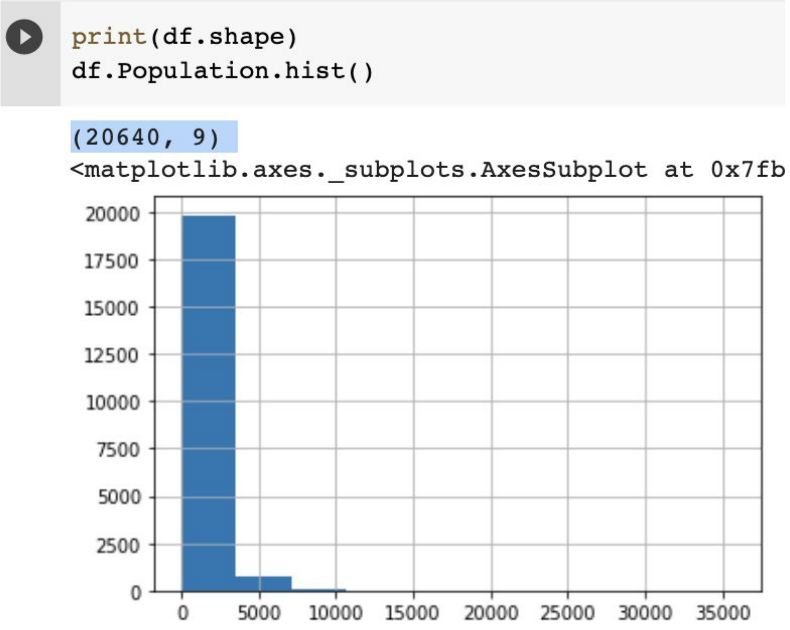
```
[36] sns.regplot(data=df.sample(200, random_state=1),
                 x="HouseAge", y="MedianHouseValue",
                 line_kws={"color": "red"})
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fb6a8a18ed0>
```

- What we try to do here is comparing bivariate analysis
- Here, we assume a linear relationship between the two variables
- We can use corr() function to measure the linear correlation
- We can also use visualization to show it

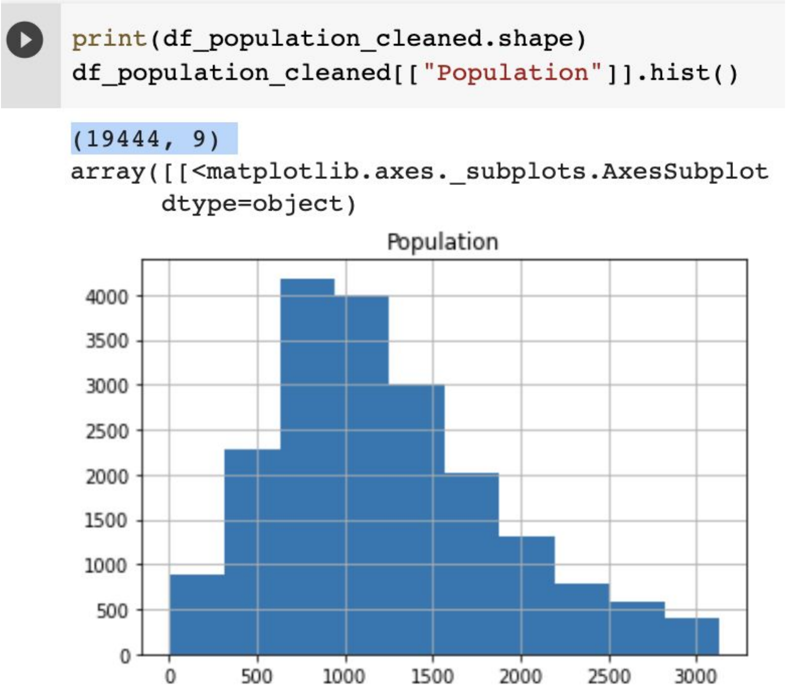
CALIFORNIA HOUSING

3. The more populated the house block, the more expensive the price is



```
print(df.shape)
df.Population.hist()
```

```
(20640, 9)
<matplotlib.axes._subplots.AxesSubplot at 0x7fb
```



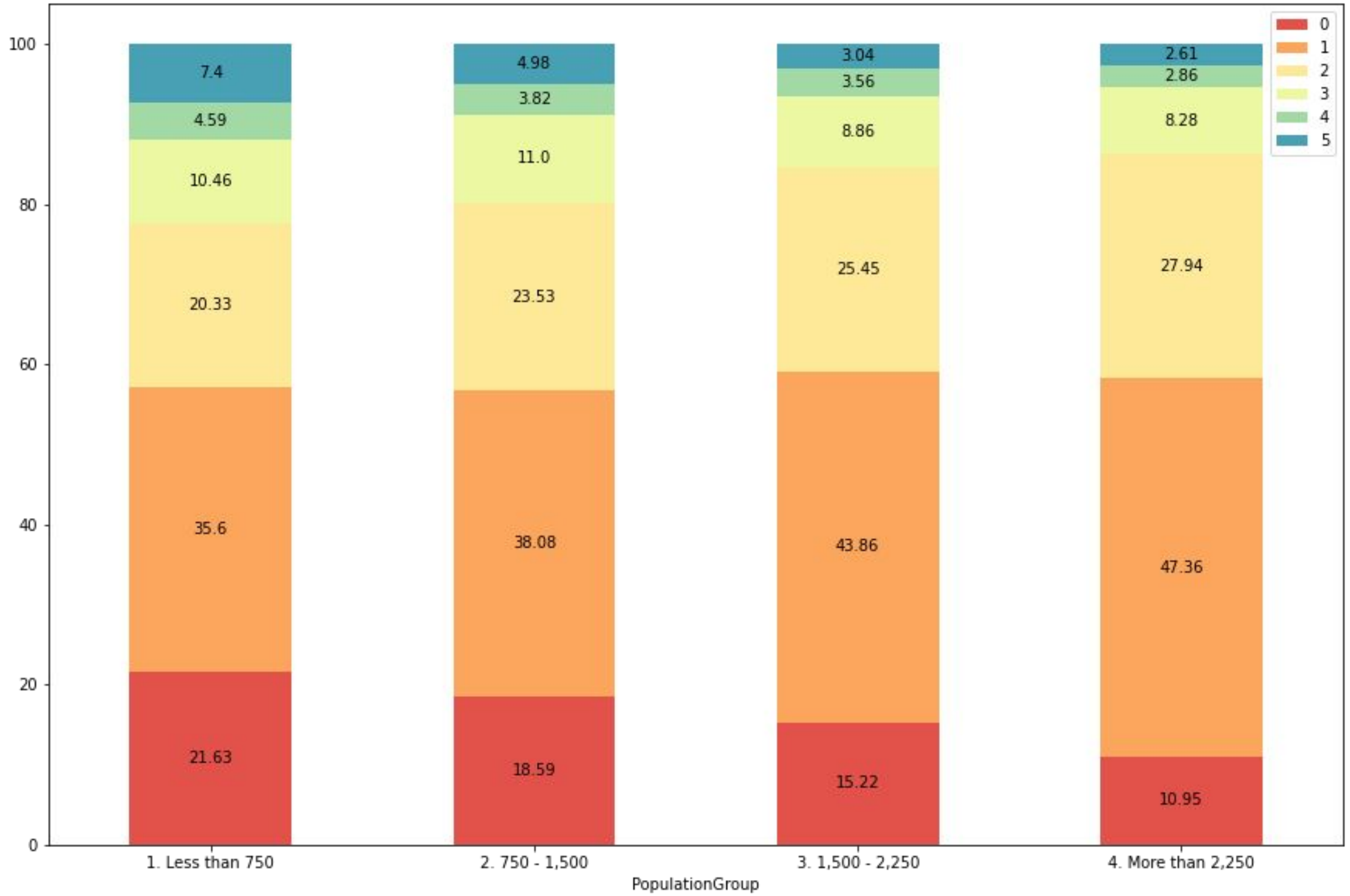
```
print(df_population_cleaned.shape)
df_population_cleaned[["Population"]].hist()
```

```
(19444, 9)
array([[<matplotlib.axes._subplots.AxesSubplot
        dtype=object)
```

- Before we analyse the relationship between number of population and house price, notice that the distribution of population data is very skewed. This indicates that the occurrences of outliers
- We need to do necessary cleanups to remove the outliers. This resulting a removal of ~1,200 data points (5.8%)

CALIFORNIA HOUSING

3. The more populated the house block, the more expensive the price is

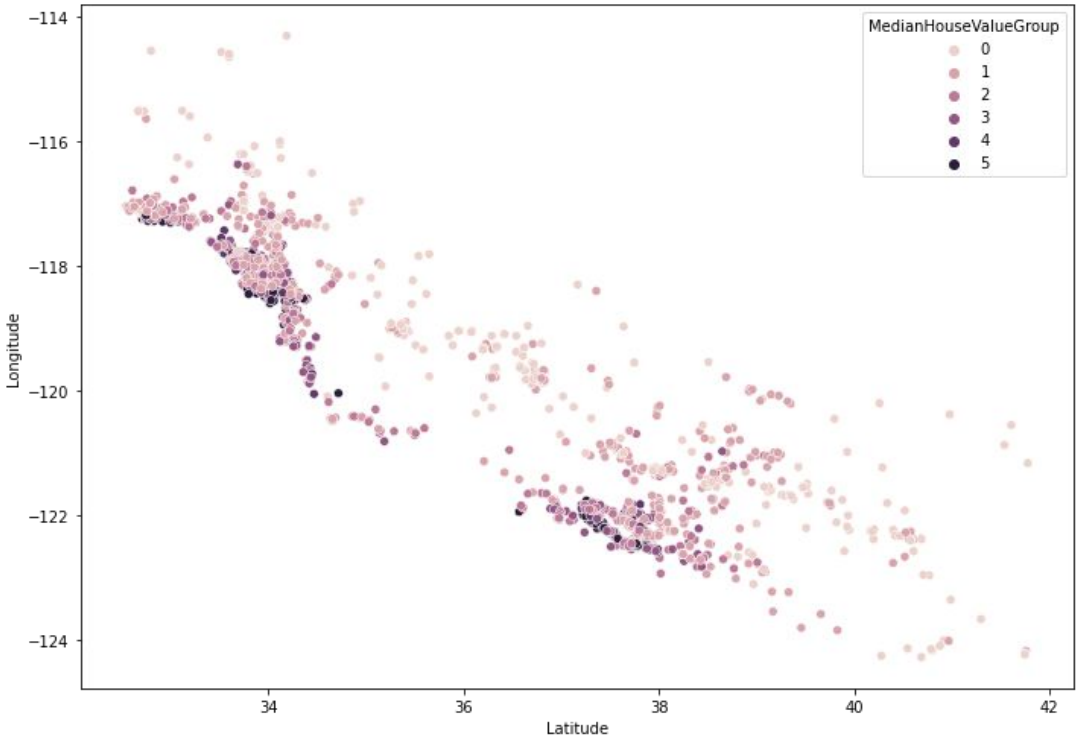


- In this analysis we convert number of population and house price data into categorical type

- In this examples. the chart shows the percentage of house block in each Population Group based on the house price

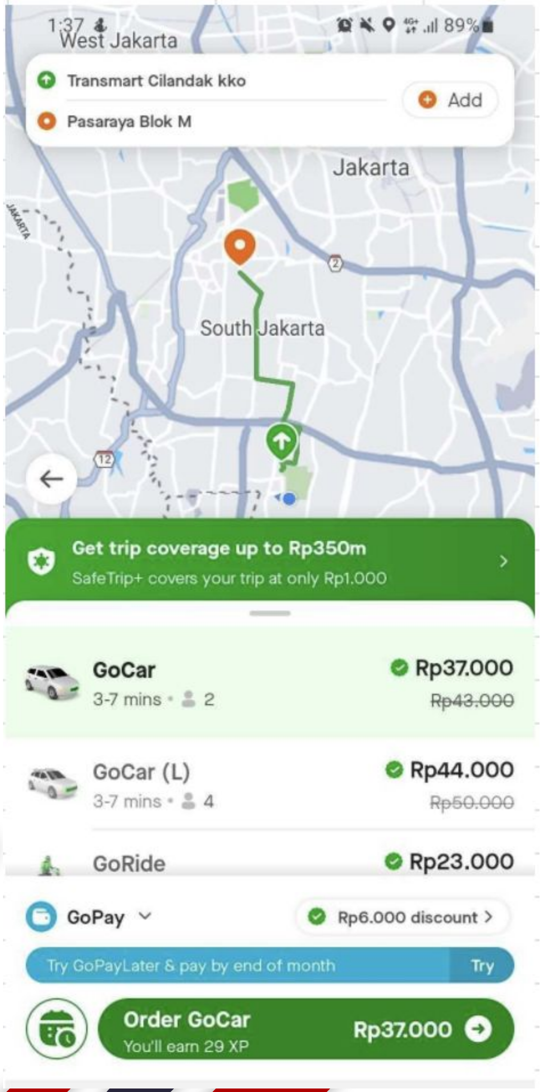
CALIFORNIA HOUSING

4. Is there certain areas with more expensive price than the other?



- We can use a geographical features to plot the data in the map

QUIZ



Transmart Cilandak kko

Pasaraya Blok M

Get trip coverage up to Rp350m
SafeTrip+ covers your trip at only Rp1.000

GoCar 3-7 mins • 2 Rp37.000 ~~Rp43.000~~

GoCar (L) 3-7 mins • 4 Rp44.000 ~~Rp50.000~~

GoRide Rp23.000

Rp6.000 discount

Try GoPayLater & pay by end of month

Order GoCar
You'll earn 29 XP
Rp37.000

Before users order a service, they will be shown the estimate screens like the one in the picture on the left. Unfortunately, from 100% users that land on this page, only 50% of them are through to confirm a booking.

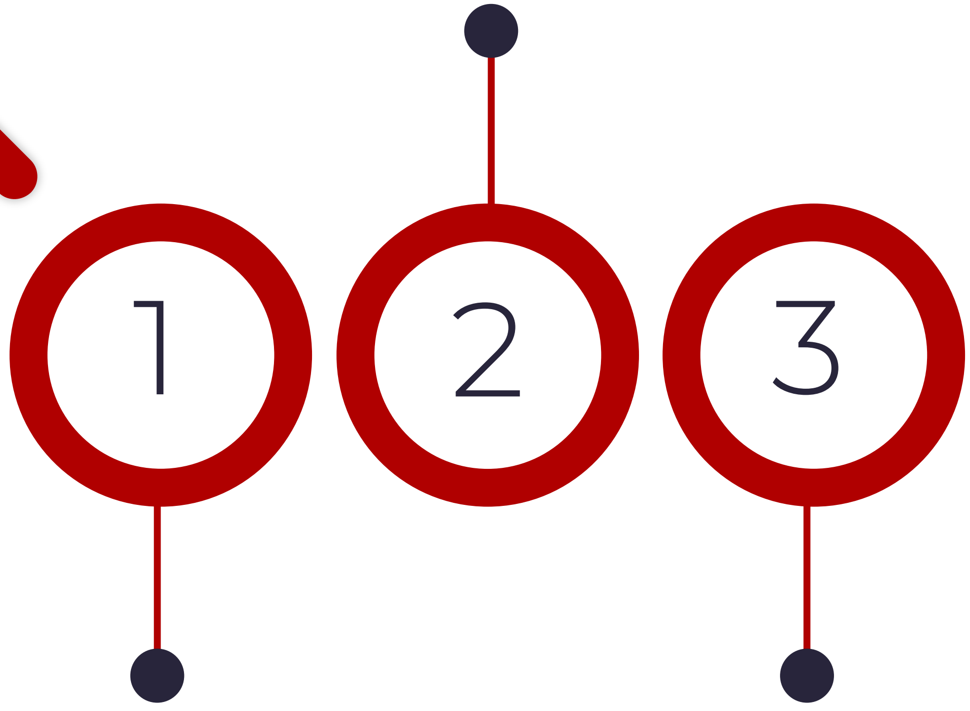
With assumption that we can get all the data we want, list down 5 assumptions and or questions from the defined objective :

Why are users dropping off in the estimate screen?

1. Change to competitor
2. Expensive price per KM
3. Driver too far away

HOW TO PRESENT INSIGHT THROUGH DATA STORYTELLING

PRESENTING INSIGHTS



1. Empathize with your audience
2. Build effective visualization
3. Create a flowing story

Empathize with your audience

- Who am I communicating to?
- What do I want my audience to know or do?
- How can I use data to help make my point?

Build effective visualization

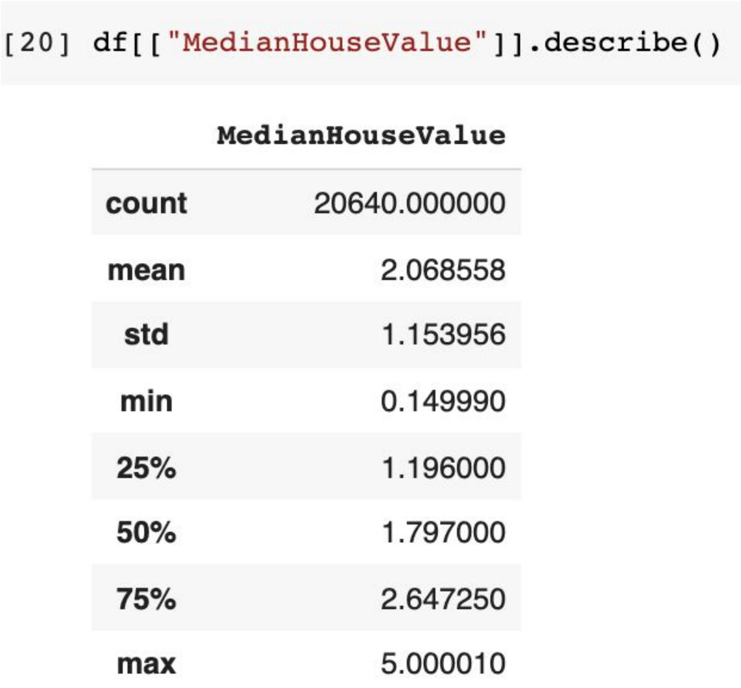
- Choose visualization that is simple but able to deliver the information clearly
- Eliminate the unnecessary elements
- Draw Attention Where You Want It

Create a flowing story

- Start from a clear background
- Continuous and chronological storyline
- Utilize the power of repetition to help your stories stick
- Conclude with a call to action
- Seeking a fresh perspective to ensure that your story comes across clearly in your communication

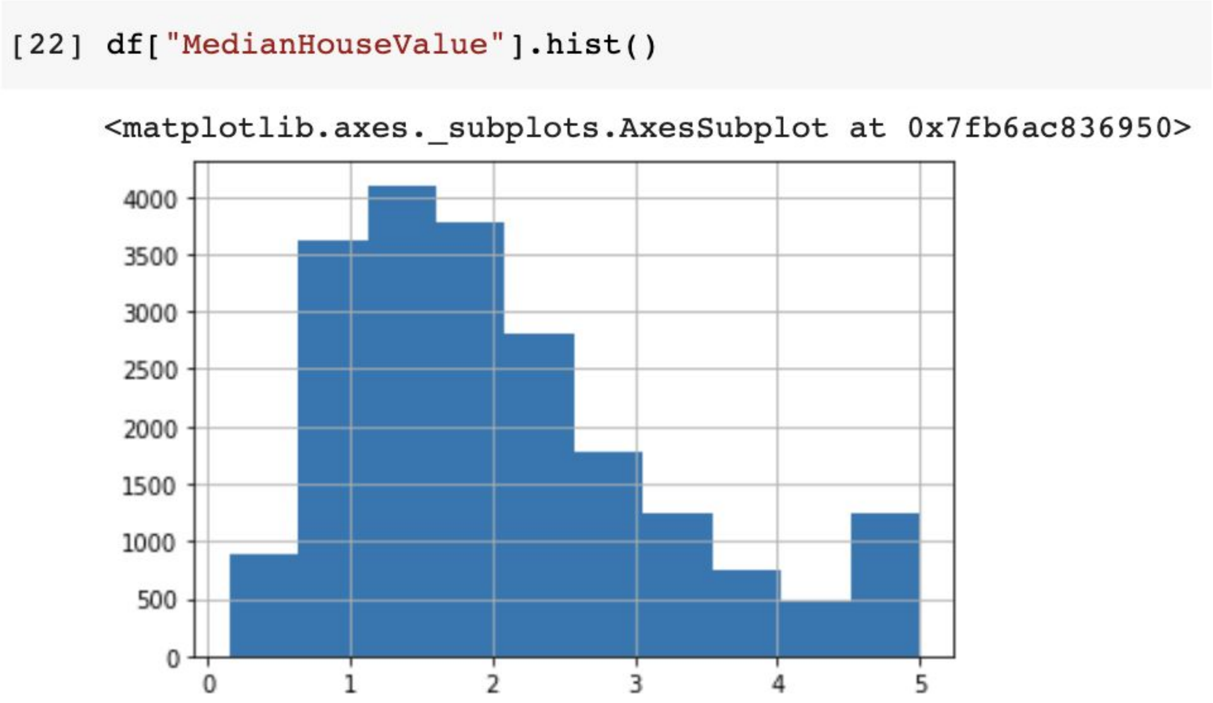
CALIFORNIA HOUSING

1. What is the distribution of the house pricing looks like?



```
[20] df[["MedianHouseValue"]].describe()
```

	MedianHouseValue
count	20640.000000
mean	2.068558
std	1.153956
min	0.149990
25%	1.196000
50%	1.797000
75%	2.647250
max	5.000010



```
[22] df["MedianHouseValue"].hist()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fb6ac836950>
```

Instead of,

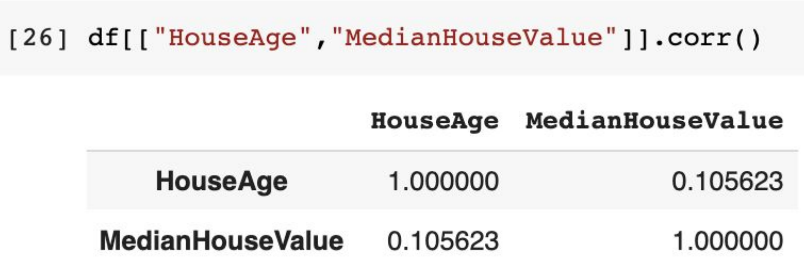
There are **20K datapoints**, the **average price is 2.06**, with **standard deviation 1.15**. The Q1, Q2, Q3 respectively are **1.19, 1.79, 2.65**

We can use,

The price is ranged from **0.14 to 5**. The distribution is right skewed which indicates more houses have a price **below the overall average** (2.07)

CALIFORNIA HOUSING

2. The older the house age, the lower the price

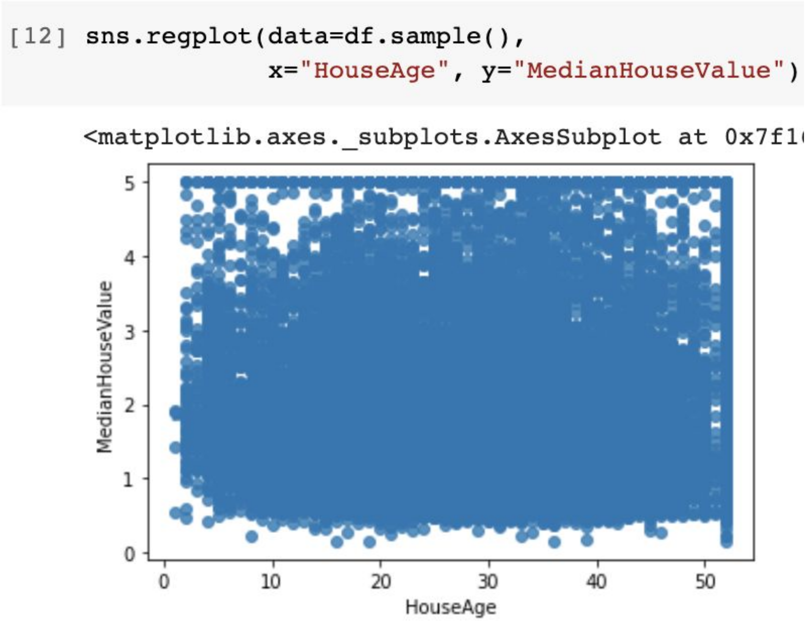


```
[26] df[["HouseAge","MedianHouseValue"]].corr()
```

	HouseAge	MedianHouseValue
HouseAge	1.000000	0.105623
MedianHouseValue	0.105623	1.000000

- The correlation value **0.11** indicates that there is no linear relationship
- Based on the visualization of sample data, we also don't see the scatterplot has a negative linear trend

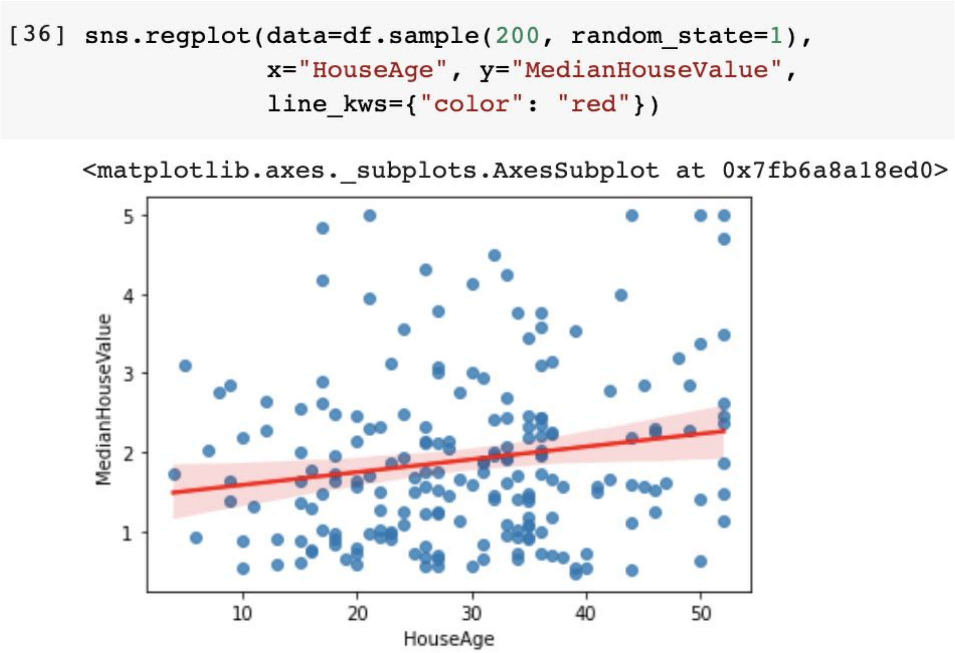
Instead of,



```
[12] sns.regplot(data=df.sample(),
                 x="HouseAge", y="MedianHouseValue")
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f1
```

We can use,

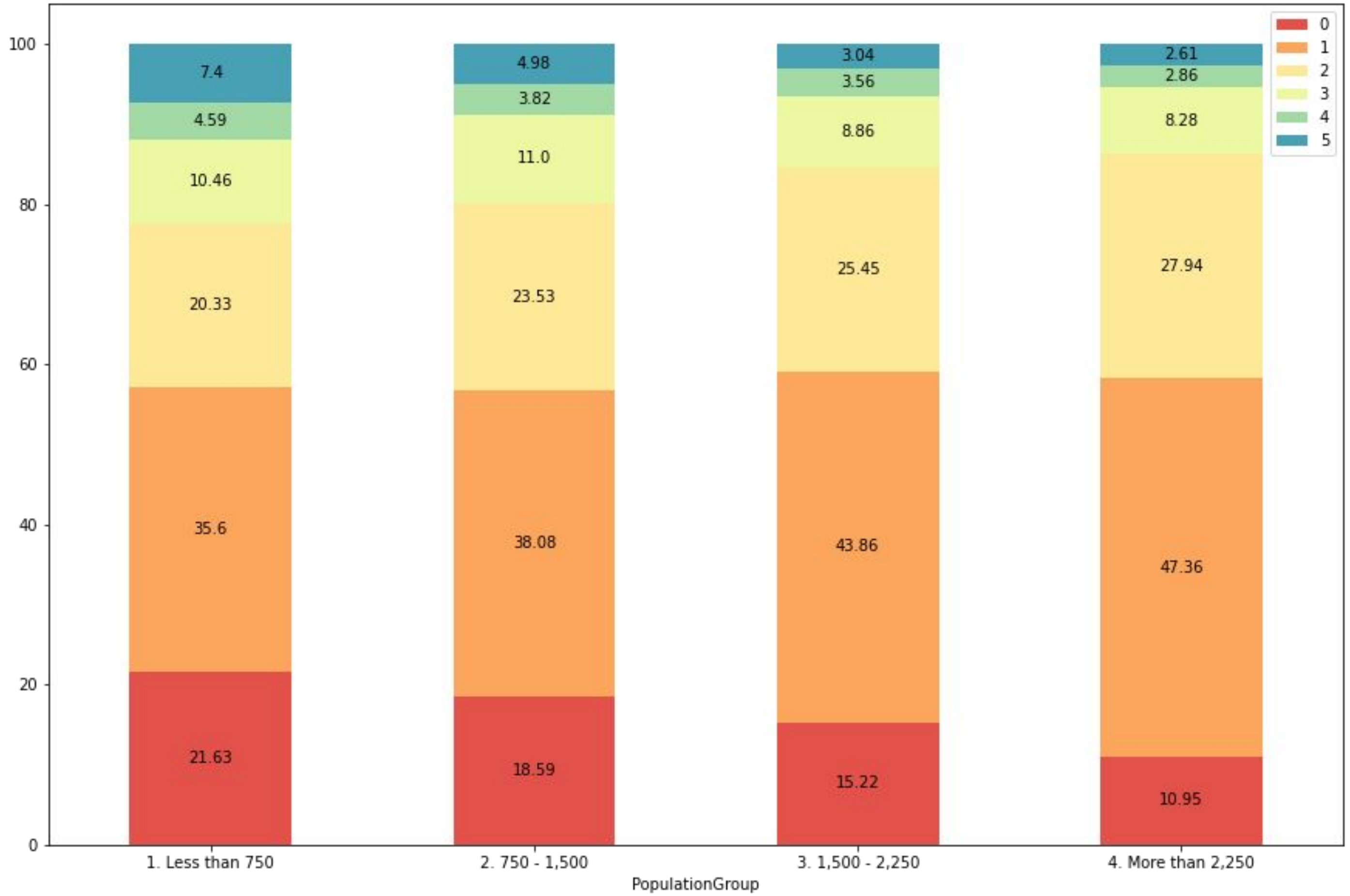


```
[36] sns.regplot(data=df.sample(200, random_state=1),
                 x="HouseAge", y="MedianHouseValue",
                 line_kws={"color": "red"})
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fb6a8a18ed0>
```

CALIFORNIA HOUSING

3. The more populated the house block, the more expensive the price is

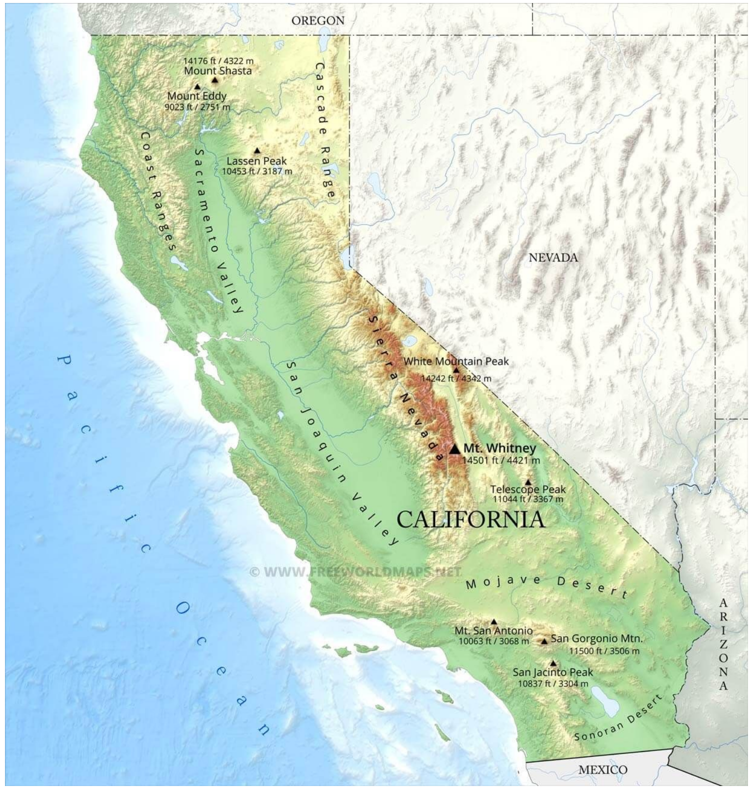


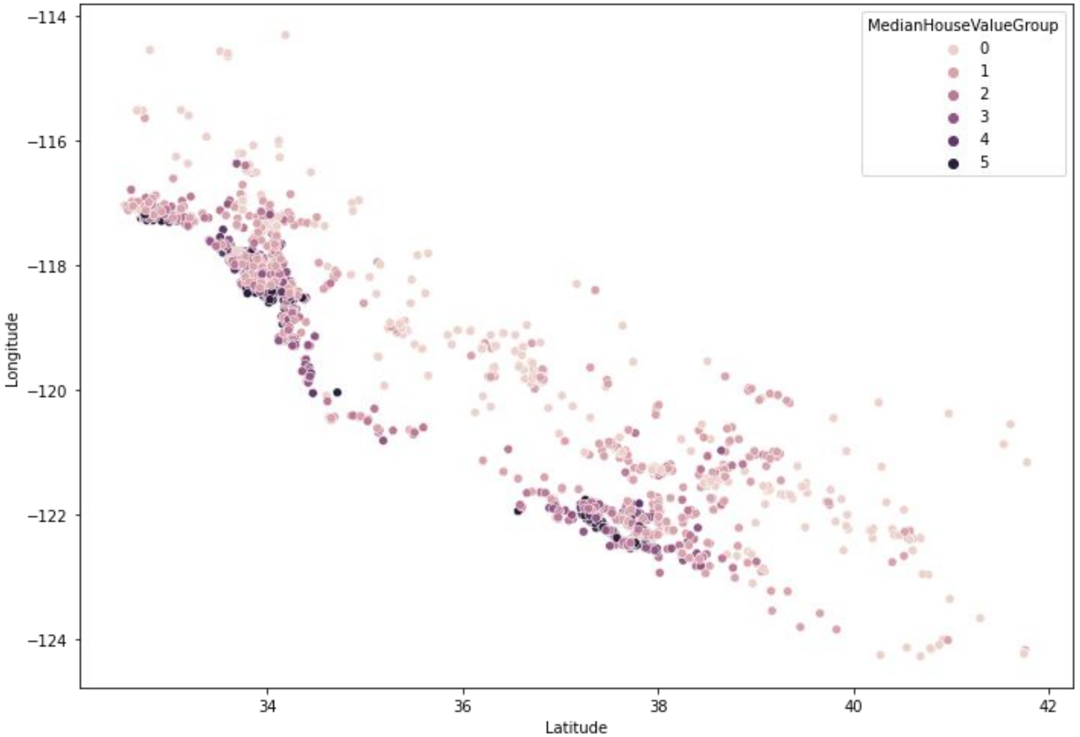
- The percentage of houses with price less than 1 in the least populated area is twice higher than the one in most populated area

- However, it is the same case for houses whose price is 4 or more

CALIFORNIA HOUSING

4. Is there certain areas with more expensive price than the other?





Instead of,

- Based on the plot we can see that more expensive houses are the one with **easier access to the beach**

We can use,

- Based on the plot we can see that more expensive houses are the one that is located more **towards the coast**

SUMMARY

- The data was derived from **California 1990 U.S. census** which consists of about **20K block groups**
- The price is ranged from **0.14 to 5**. The distribution is right skewed which indicates more houses have a price **below the overall average** (2.07)
- One might suggest that the older the house age, the lower the price, but we **can't proof a strong linear relationship** (correlation : 0.11)
- Although the percentage of houses with price less than 1 in the least populated area is **twice higher** than the one in most populated area, the **same case also happens** for houses with price is 4 or more
- Based on the geographical aspect, houses with high price (4-5) is more distributed in the area **towards the coast**
- Based on the analysis we might say that one of the factor that highly influence the price of the house in California is the **location**

IMPORTANT SKILLS FOR DATA ANALYST

IMPORTANT SKILLS

Curious

- ✓ Driven to find what causes some events to happen

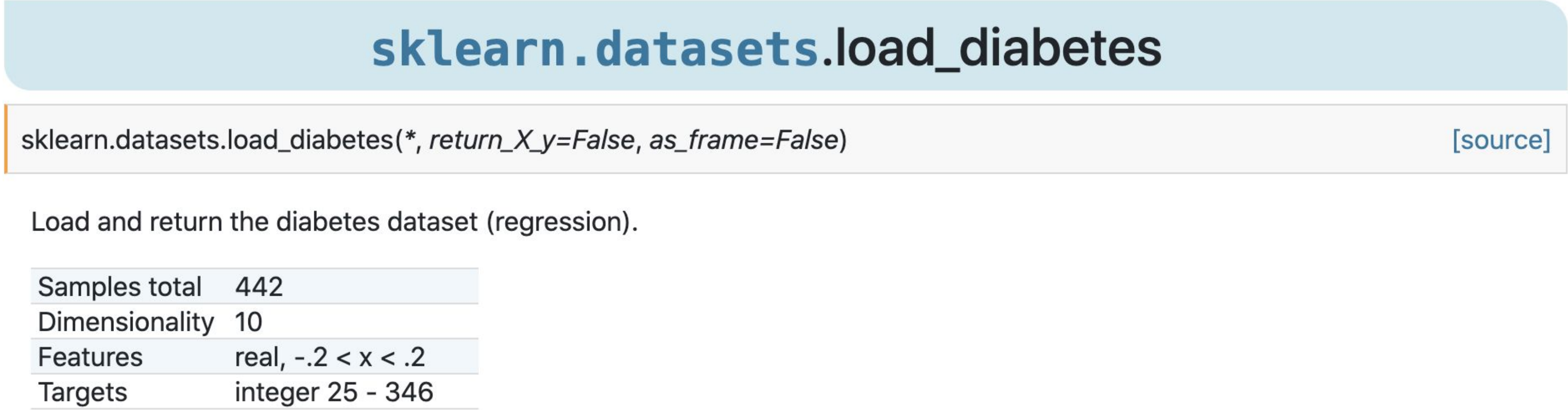
Creative

- ✓ Ability to look at a problem from several angles
- ✓ Ability to break down a problem into several sub-problems

Fast

- ✓ Enabling fast decision making
- ✓ What differs from Data Scientist

TASK



sklearn.datasets.load_diabetes

```
sklearn.datasets.load_diabetes(*, return_X_y=False, as_frame=False)
```

[source]

Load and return the diabetes dataset (regression).

Samples total	442
Dimensionality	10
Features	real, -.2 < x < .2
Targets	integer 25 - 346

Using Sklearn Diabetes data, create an EDA that answers the defined objective :

What are the factors that contribute to the disease progression of a diabetes patient?

Note on the assignment :

- Provide 5 questions and or assumptions
- Use some data visualization
- Provide conclusion and or recommendation based on the analysis

THANKYOU

“Without data, you're just another person with an opinion”

- W. Edwards Deming

✉ yahyaerucakra@gmail.com

🌐 linkedin.com/in/erucakra