EXPLORE AI
ACADEMY

Exploratory Data Analysis

2304PTDS

October 2023

Objectives

By the end of this session you should be able to:

- Recall the basics of EDA - what it is and why we do it
- Explain the concepts of univariate and multivariate analysis
- Practice steps to conduct simple EDA on a dataset
- Compare and contrast the EDA tools/methods used

1 / **Recap: What is EDA?**

2 / **Why do we use EDA?**

3 / **Data Cleaning**

4 / **EDA Approach**

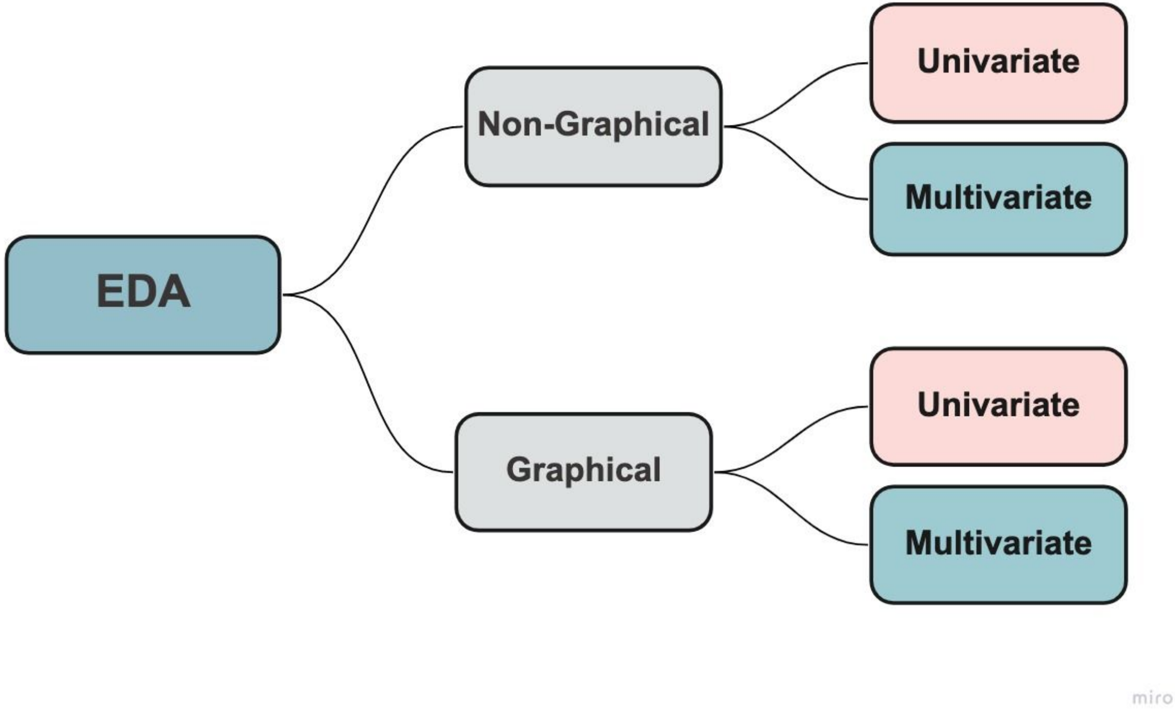
5 / **To the notebook...**

“Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.”

- John Tukey

What is Exploratory Data Analysis?

Let’s recap what we know



1 / **Recap: What is EDA?**

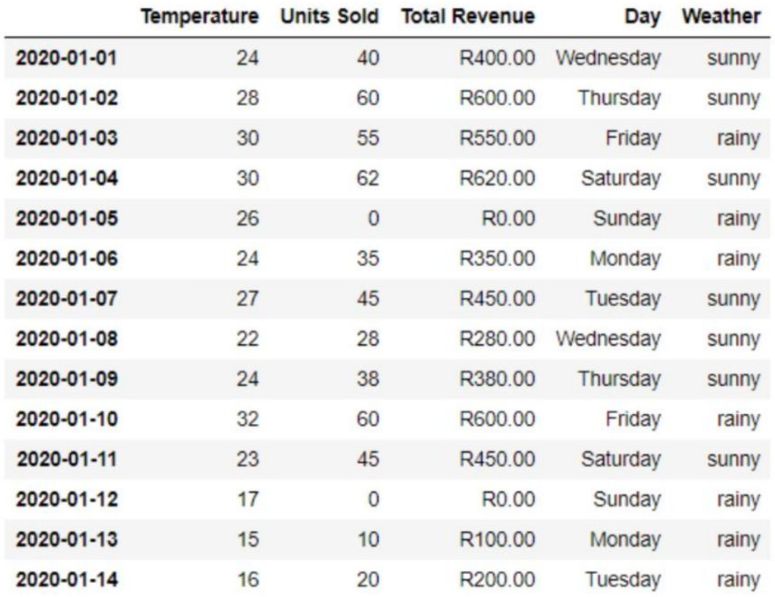
2 / **Why do we use EDA?**

3 / **Data Cleaning**

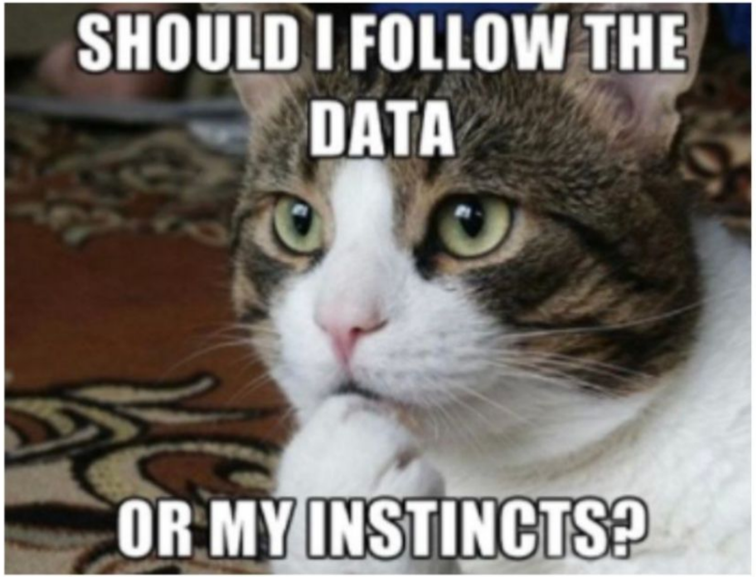
4 / **EDA Approach**

5 / **To the notebook…**

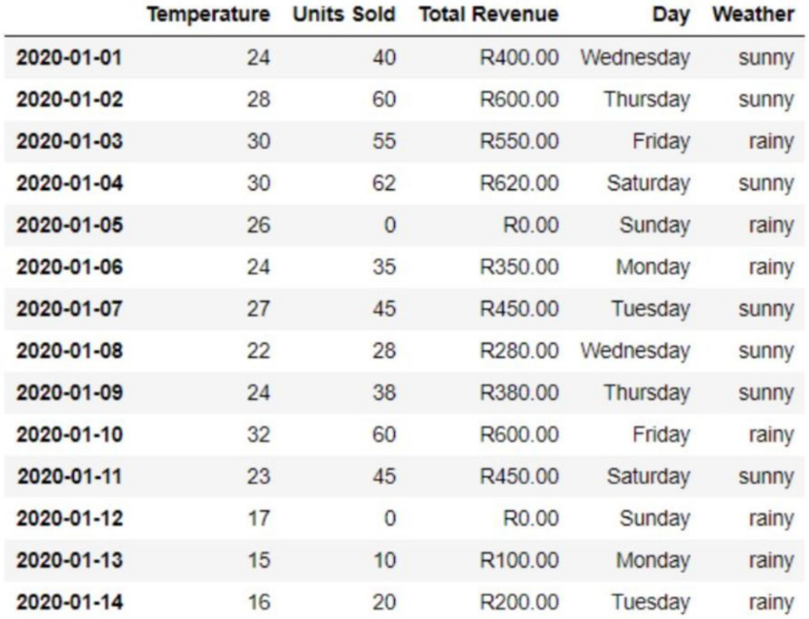
Why do we do Exploratory Data Analysis?



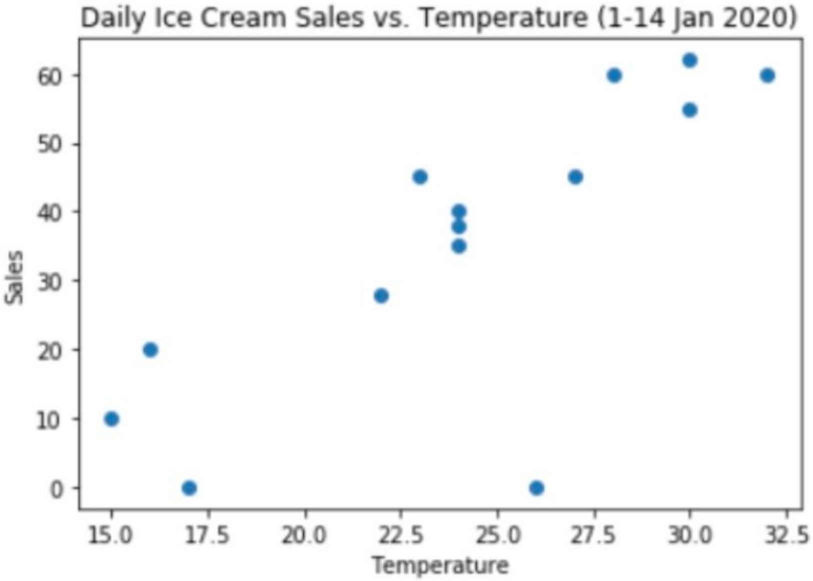
	Temperature	Units Sold	Total Revenue	Day	Weather
2020-01-01	24	40	R400.00	Wednesday	sunny
2020-01-02	28	60	R600.00	Thursday	sunny
2020-01-03	30	55	R550.00	Friday	rainy
2020-01-04	30	62	R620.00	Saturday	sunny
2020-01-05	26	0	R0.00	Sunday	rainy
2020-01-06	24	35	R350.00	Monday	rainy
2020-01-07	27	45	R450.00	Tuesday	sunny
2020-01-08	22	28	R280.00	Wednesday	sunny
2020-01-09	24	38	R380.00	Thursday	sunny
2020-01-10	32	60	R600.00	Friday	rainy
2020-01-11	23	45	R450.00	Saturday	sunny
2020-01-12	17	0	R0.00	Sunday	rainy
2020-01-13	15	10	R100.00	Monday	rainy
2020-01-14	16	20	R200.00	Tuesday	rainy



Why do we do Exploratory Data Analysis?

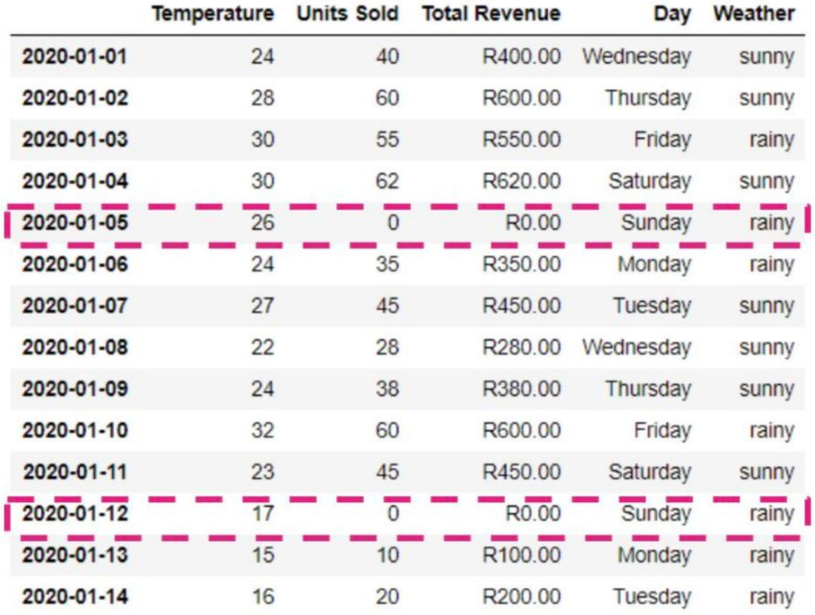


	Temperature	Units Sold	Total Revenue	Day	Weather
2020-01-01	24	40	R400.00	Wednesday	sunny
2020-01-02	28	60	R600.00	Thursday	sunny
2020-01-03	30	55	R550.00	Friday	rainy
2020-01-04	30	62	R620.00	Saturday	sunny
2020-01-05	26	0	R0.00	Sunday	rainy
2020-01-06	24	35	R350.00	Monday	rainy
2020-01-07	27	45	R450.00	Tuesday	sunny
2020-01-08	22	28	R280.00	Wednesday	sunny
2020-01-09	24	38	R380.00	Thursday	sunny
2020-01-10	32	60	R600.00	Friday	rainy
2020-01-11	23	45	R450.00	Saturday	sunny
2020-01-12	17	0	R0.00	Sunday	rainy
2020-01-13	15	10	R100.00	Monday	rainy
2020-01-14	16	20	R200.00	Tuesday	rainy

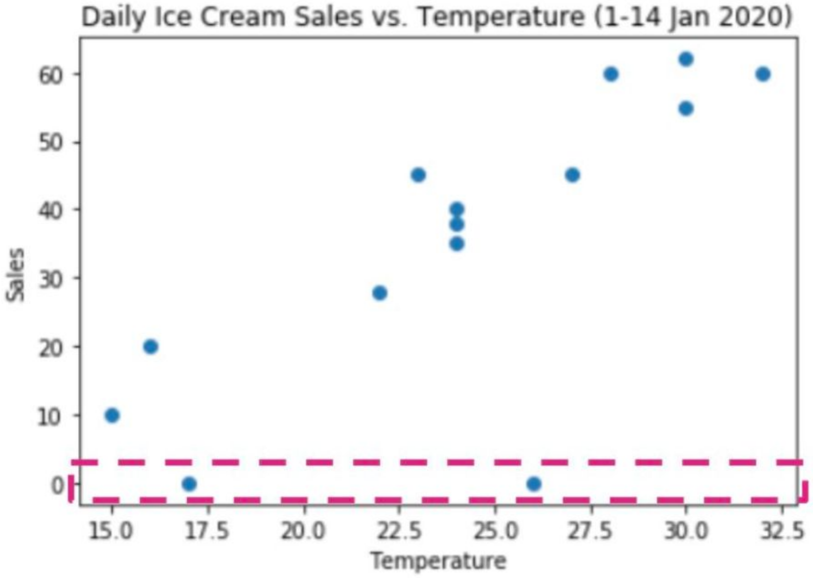


Daily Ice Cream Sales vs. Temperature (1-14 Jan 2020)

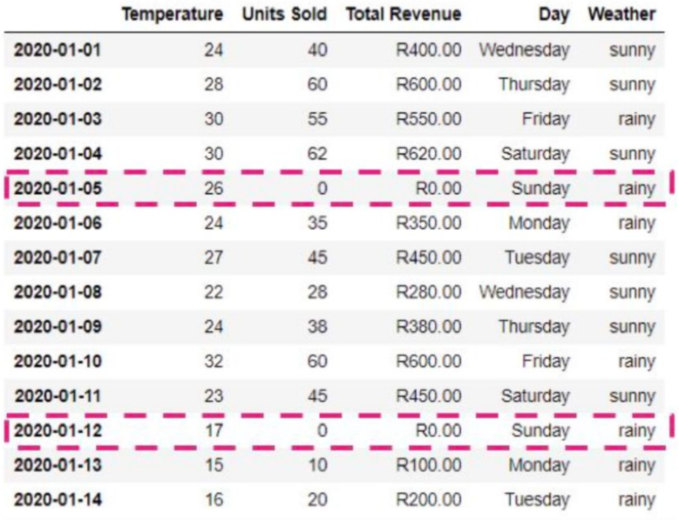
Why do we do Exploratory Data Analysis?



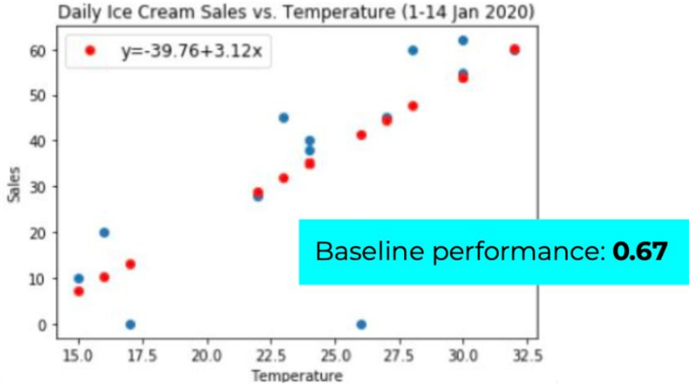
	Temperature	Units Sold	Total Revenue	Day	Weather
2020-01-01	24	40	R400.00	Wednesday	sunny
2020-01-02	28	60	R600.00	Thursday	sunny
2020-01-03	30	55	R550.00	Friday	rainy
2020-01-04	30	62	R620.00	Saturday	sunny
2020-01-05	26	0	R0.00	Sunday	rainy
2020-01-06	24	35	R350.00	Monday	rainy
2020-01-07	27	45	R450.00	Tuesday	sunny
2020-01-08	22	28	R280.00	Wednesday	sunny
2020-01-09	24	38	R380.00	Thursday	sunny
2020-01-10	32	60	R600.00	Friday	rainy
2020-01-11	23	45	R450.00	Saturday	sunny
2020-01-12	17	0	R0.00	Sunday	rainy
2020-01-13	15	10	R100.00	Monday	rainy
2020-01-14	16	20	R200.00	Tuesday	rainy



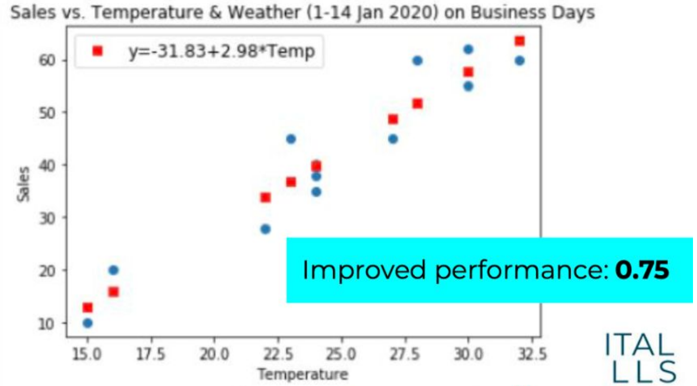
Why do we do Exploratory Data Analysis?



	Temperature	Units Sold	Total Revenue	Day	Weather
2020-01-01	24	40	R400.00	Wednesday	sunny
2020-01-02	28	60	R600.00	Thursday	sunny
2020-01-03	30	55	R550.00	Friday	rainy
2020-01-04	30	62	R620.00	Saturday	sunny
2020-01-05	26	0	R0.00	Sunday	rainy
2020-01-06	24	35	R350.00	Monday	rainy
2020-01-07	27	45	R450.00	Tuesday	sunny
2020-01-08	22	28	R280.00	Wednesday	sunny
2020-01-09	24	38	R380.00	Thursday	sunny
2020-01-10	32	60	R600.00	Friday	rainy
2020-01-11	23	45	R450.00	Saturday	sunny
2020-01-12	17	0	R0.00	Sunday	rainy
2020-01-13	15	10	R100.00	Monday	rainy
2020-01-14	16	20	R200.00	Tuesday	rainy



Baseline performance: **0.67**



Improved performance: **0.75**

Why do we do Exploratory Data Analysis?

Let’s recap what we know

- Exploratory data analysis will allow you to:
  - Discover patterns
  - Detect anomalies (outliers)
  - Form hypotheses based on our understanding of the dataset
  - Take actions - question, clean, transform!

1 / **Recap: What is EDA?**

2 / **Why do we use EDA?**

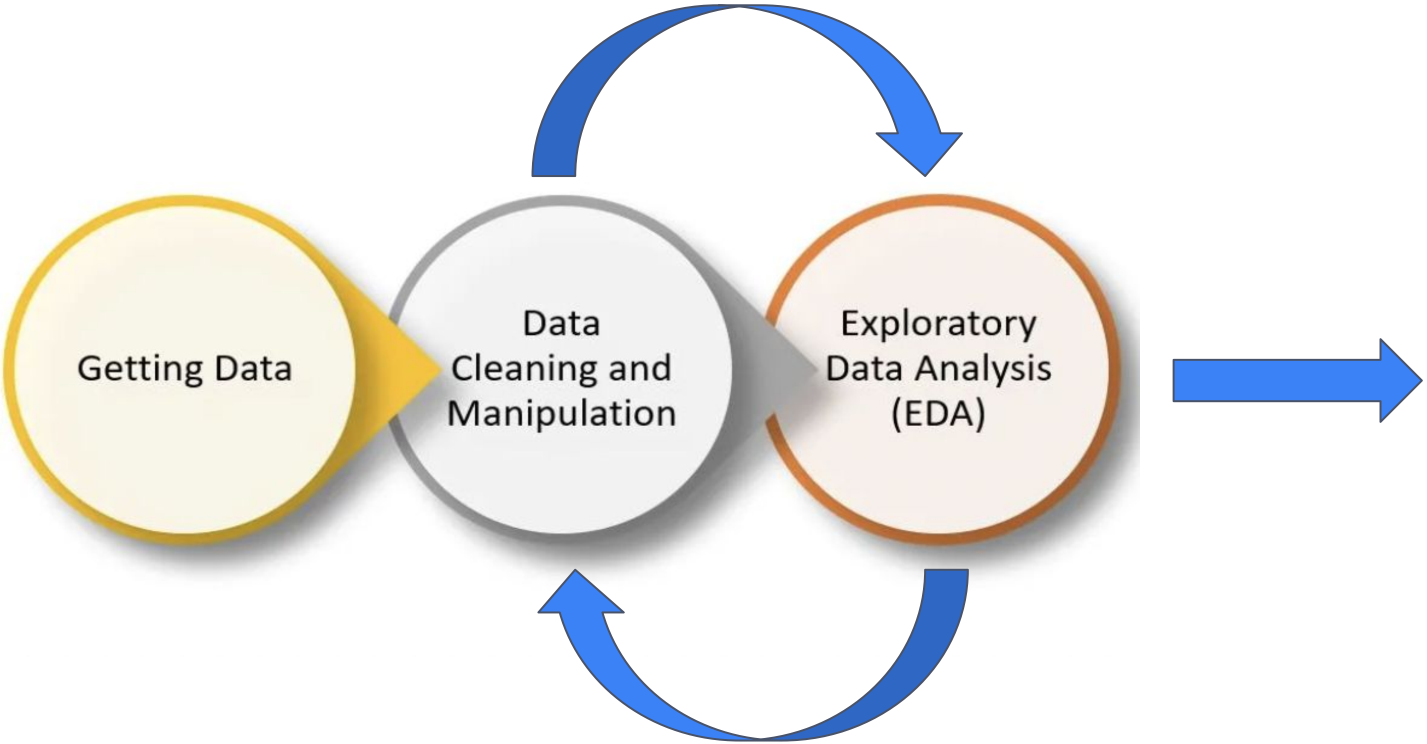
3 / **Data Cleaning**

4 / **EDA Approach**

5 / **To the notebook…**

Data Cleaning

How does EDA relate?


Getting Data
Data Cleaning and Manipulation
Exploratory Data Analysis (EDA)

Data Cleaning

How does EDA relate?

- Detection of:
  - Missing values
  - Outliers
  - Data type mismatches
  - Etc...
- Some of these are readily visible, some only arise when we start looking into the data!

1 / **Recap: What is EDA?**

2 / **Why do we use EDA?**

3 / **Data Cleaning**

4 / **EDA Approach**

5 / **To the notebook…**

EDA Approach

Remember, EDA is not just checking a box - we need to truly understand our data

- Understanding your data
  - Visualise - univariate and multivariate distributions and relationships
  - Understand our features and our target variable
- Missing values
  - Are missing values random, or based on a pattern?
  - Can we impute these values? Remove them?
- Outlier / anomaly detection
  - Are these artificial or natural occurrences?
  - Do we remove them? If so, how?
- Feature engineering/selection
  - Can we create anything else useful from what we have?
  - Do any formats need to change?

To the notebook...