Introduction to EDA
(Exploratory Data Analysis)

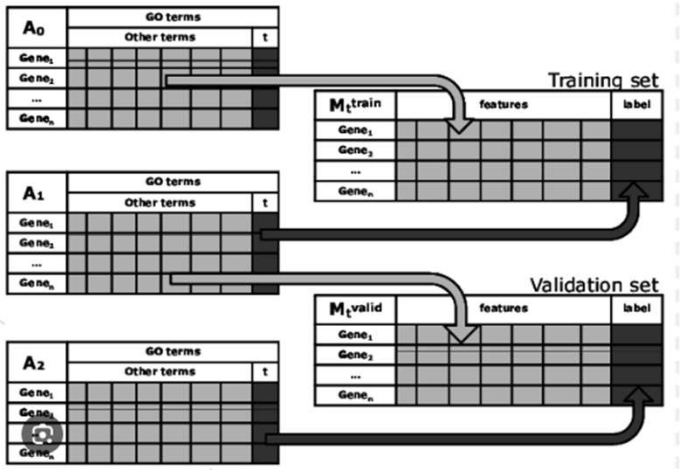
RECORD LESSON

Agenda

- The importance of Exploratory Data Analysis
- Data Cleaning
- Data Processing and the Concept of ETL
- Data Visualization

What is EDA?

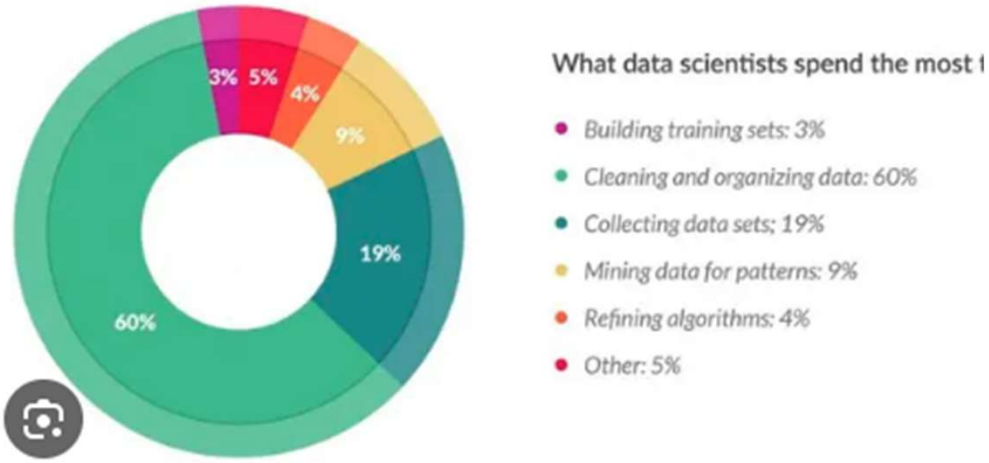
Exploratory Data Analysis (EDA) is a crucial step in the data analysis process, which involves examining the main characteristics of a dataset, often visually, before making any assumptions or building statistical models. This approach helps in uncovering patterns, spotting anomalies, testing hypotheses, and checking assumptions through summary statistics and graphical representations.



- How many rows?
- What is the avg of “Total”? Does it make sense?
- Is there any data missing in variable “C”?

Data Cleaning

Assume the data is dirty - but what does this mean?
You will be faced with datasets of inconsistent data, that you will need to fix - aka, clean - yourselves.



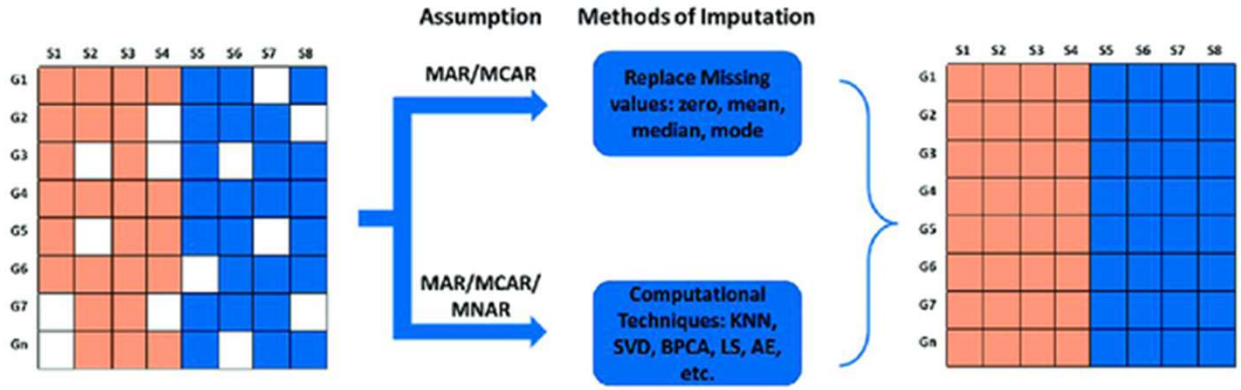
1. San Francisco, CA, USA
2. San Franciso, California, USA (typo)
3. san francisco, california, usa (case issues)
4. SF, California, USA (abbreviation)
5. San Francisco CA USA (missing commas)
6. S. Francisco, California, USA (shortening)
7. San Francisco, Calif., USA (different abbreviation)
8. San Francisco - California, USA (alternative delimiter)
9. san francisco, ca, usa (case and abbreviation issues)
10. San Fran, CA, USA (nickname)
11. SAN FRANCISCO, CALIFORNIA, USA (all caps)
12. San Francisco California, USA (missing comma)
13. San Fransisco, California, USA (typo)
14. Saint Francisco, California, USA (formal/incorrect name)
15. SFO, California, USA (airport code used as city name)
16. San Francisco, Cali, USA (informal abbreviation)
17. San Fran., Cal., USA (abbreviated with periods)
18. San Francisco Calif USA (missing punctuation)
19. San Francisco, Cal., US (country abbreviation inconsistency)
20. San Francsco, California, US (typo and country abbreviation inconsistency)

Data Injection

Cleaning data is no just about typos - often you will also having missing data that is central to your analysis.
"Something" is often better than nothing. But which "something?

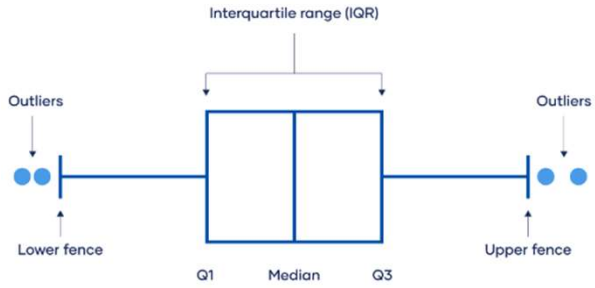
Common imputation techniques:

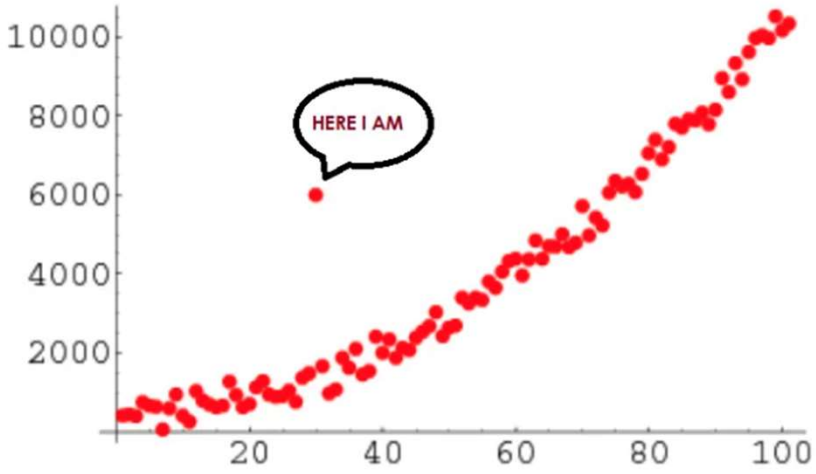
- *Mean*
- *Median*
- *Mode*
- *Zero*
- *and many more*



Outliers

The concept of an outlier is a point that is valid - it exists in our data and in reality - but it is not **representative of our dataset.**





Outliers

What should we do about Outliers?

There is no Golden Rule – Outliers are still valid situations in our reality

Outliers

What should we do about Outliers?

Sales of a shop

vs

Wealth of Millenials

Data processing - Filtering

Filtering is one of the main parts of EDA - it reflects our choice over the analysis that we want to do and which data we want to include.

Filtering is done by rules, not by hand

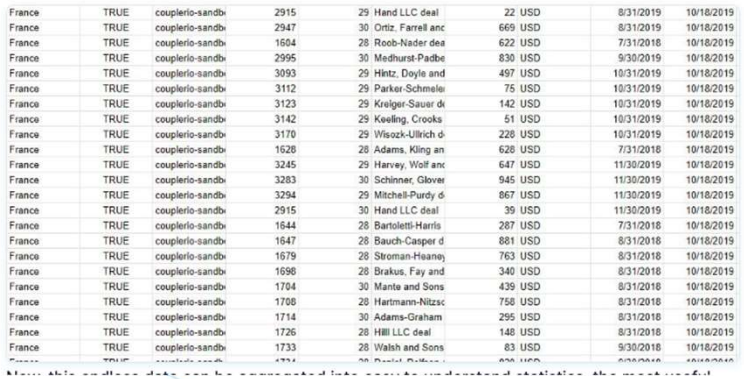
- Filter on a subset of elements
- Filter on a condition (a > b) or a combination of condictions
- Filter on an issue (remove Nas or invalid rows)

Filtering is done on the rows

				Brand Ratings					
ID	Age	Gender	Preferred cola	Coca-Cola	Diet Coke	Coke Zero	Pepsi	Diet Pepsi	Pepsi Max
1	25 to 29	Female	Pepsi Max	2	5	2	3	1	4
2	45 to 49	Male	Pepsi Max	5	1	5	5	3	4
3	25 to 29	Female	Diet Coke	5	4	2	3	1	1
4	25 to 29	Female	Coca-Cola	4	2	2	2	2	2
5	55 to 64	Female	Diet Coke	3	4	3	3	4	2
6	55 to 64	Female	Diet Pepsi	3	3	3	3	4	4
7	50 to 54	Female	Coke Zero	2	3	5	2	2	2
8	35 to 39	Female	Coca-Cola	4	2	5	3	2	5
9	65 or more	Male	Diet Pepsi	5	5	3	5	5	3
10	45 to 49	Female	Coke Zero	4	4	4	5	5	3
11	45 to 49	Male	Coca-Cola	4	1	1	4	1	1
12	55 to 64	Male	Coca-Cola	5	2	2	5	2	2
13	55 to 64	Male	Coca-Cola	5	2	2	3	2	2
14	30 to 34	Male	Pepsi Max	3	2	5	3	3	5
15	65 or more	Female	Diet Pepsi	2	4	2	5	4	2

Data processing - Aggregations

The importance of aggregating data cannot be overstated. Every time technology advances, humanity is able to collect and process a high and more detailed volume of data – but the human brain still needs to understand high level, aggregated data to have a full picture.



France	TRUE	couplerio-sandb	2915	29	Hand LLC deal	22	USD	8/31/2019	10/18/2019
France	TRUE	couplerio-sandb	2947	30	Ortiz, Farrell anc	669	USD	8/31/2019	10/18/2019
France	TRUE	couplerio-sandb	1604	28	Roob-Nader dea	622	USD	7/31/2018	10/18/2019
France	TRUE	couplerio-sandb	2995	30	Medhurst-Padbe	830	USD	9/30/2019	10/18/2019
France	TRUE	couplerio-sandb	3093	29	Hintz, Doyle and	497	USD	10/31/2019	10/18/2019
France	TRUE	couplerio-sandb	3112	29	Parker-Schmele	75	USD	10/31/2019	10/18/2019
France	TRUE	couplerio-sandb	3123	29	Kreiger-Sauer de	142	USD	10/31/2019	10/18/2019
France	TRUE	couplerio-sandb	3142	29	Keeling, Crooks	51	USD	10/31/2019	10/18/2019
France	TRUE	couplerio-sandb	3170	29	Wisozk-Ullrich d	228	USD	10/31/2019	10/18/2019
France	TRUE	couplerio-sandb	1628	28	Adams, Kling an	628	USD	7/31/2018	10/18/2019
France	TRUE	couplerio-sandb	3245	29	Harvey, Wolf and	647	USD	11/30/2019	10/18/2019
France	TRUE	couplerio-sandb	3283	30	Schinner, Glover	945	USD	11/30/2019	10/18/2019
France	TRUE	couplerio-sandb	3294	29	Mitchell-Purdy d	867	USD	11/30/2019	10/18/2019
France	TRUE	couplerio-sandb	2915	30	Hand LLC deal	39	USD	11/30/2019	10/18/2019
France	TRUE	couplerio-sandb	1644	28	Bartoletti-Harris	287	USD	7/31/2018	10/18/2019
France	TRUE	couplerio-sandb	1647	28	Bauch-Casper d	881	USD	8/31/2018	10/18/2019
France	TRUE	couplerio-sandb	1679	28	Stroman-Heaney	763	USD	8/31/2018	10/18/2019
France	TRUE	couplerio-sandb	1698	28	Brakus, Fay and	340	USD	8/31/2018	10/18/2019
France	TRUE	couplerio-sandb	1704	30	Mante and Sons	439	USD	8/31/2018	10/18/2019
France	TRUE	couplerio-sandb	1708	28	Hartmann-Nitzsc	758	USD	8/31/2018	10/18/2019
France	TRUE	couplerio-sandb	1714	30	Adams-Graham	295	USD	8/31/2018	10/18/2019
France	TRUE	couplerio-sandb	1726	28	Hilll LLC deal	148	USD	8/31/2018	10/18/2019
France	TRUE	couplerio-sandb	1733	28	Walsh and Sons	83	USD	9/30/2018	10/18/2019

Country	Conversion rate	Total revenue
Australia	32.01%	$42,851.00
Canada	28.57%	$38,630.00
Denmark	28.47%	$34,078.00
France	27.89%	$39,561.00
Germany	30.10%	$43,460.00
Netherlands	28.09%	$45,102.00
Ukraine	28.04%	$31,025.00
United Kingdom	33.44%	$52,149.00
United States	29.50%	$40,088.00
United Arab Emirates	27.84%	$38,934.00

Total deals	Total revenue	Average deal life time (days)
299	$45,102.00	50

Data processing – Transformation

During our data flow, transformations will be necessary – and they start in our exploratory analysis.

Transformations / Formulas: often we will need to alter what the data looks like

Date of birth	becomes	Age
02/07/1988		35

Data processing – Transformation

During our data flow, transformations will be necessary – and they start in our exploratory analysis.

Transformations / Formulas: often we will need to alter what the data looks like.

Statistical Analysis: normalization, standardization

Wealth Rage	becomes	Wealth Rage normalized
30$-13B$		0-1

Data Visualization

The outcome of all Data Analysis it for a user make an informed decision – that information needs to be shared with the user in some sort of visual matter – often through charts.

The way we visualize that data is, in itself, a science: Data Visualization

Data Visualization

As we study how to visualize data, there are several areas of DataViz to be studied.

What can be displayed:
DIMENSIONS:

Data Attributes
Examples: Geo-Location (City, Country), Categorical Breakdown
(Usually) strings METRICS:

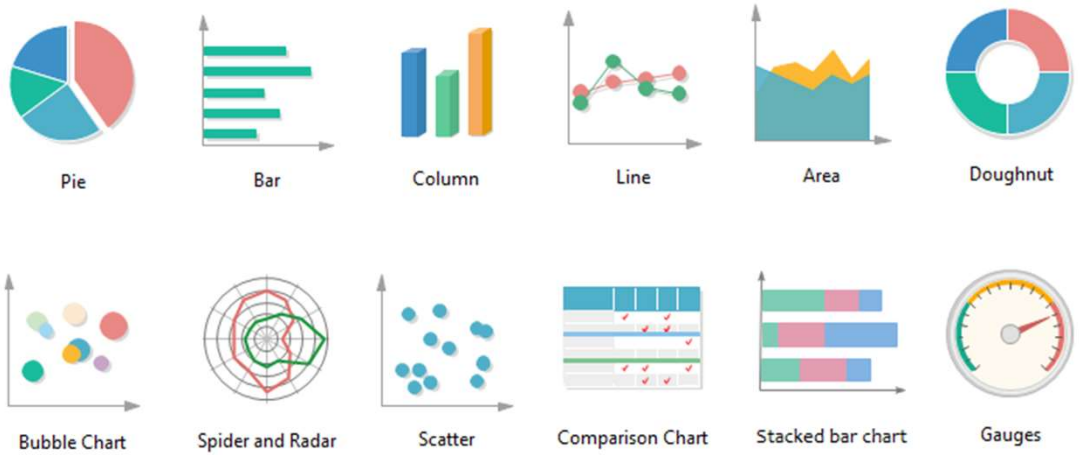
METRICS

Measurable values that describe your data
Examples: Sales, Profit, dedicated KPIs (what are KPIs??)

Data Visualization

As we study how to visualize data, there are several areas of DataViz to be studied.

The different ways it can be displayed



Thank you

Let’s get started!