Exploratory Data Analysis

Explore data with the aim of extracting useful and actionable information from it.EDA is arguably the most important and revelatory step in any kind of data analysis.
On a high Level below steps are used for EDA

1. Data sourcing
2. Data cleaning
3. Univariate analysis
4. Bivariate analysis
5. Derived metrics

Data Sourcing

To solve a business problem using analytics, you need to have historical data. Data is the key — the better the data, the more insights you can get out of it.

Typically, data comes from various sources and your first job as a data analyst is to procure the data from them. In this session, you will learn about various sources of data and how to source data from public and private sources. The broad agenda for this session is as follows:

Public Data

A large amount of data collected by the government or other public agencies is made public for the purposes of research. Such data sets do not require special permission for access and are therefore called public data.*Public Data isn't always relevant*.

- Public Data Catalogues -Awesome Public Datasets (US Specific)
- Government Data Catalogues https://data.gov.in
- Census Data - https://censusindia.gov.in
- Google Search - we can search online also

Private Data

Private data is that which is sensitive to organisations and is thus not available in the public domain. Banking, telecom, retail, and media are some of the key private sectors that rely heavily on data to make decisions.

Most of the time client cannot provide you the data for analysis outside organisation reach ,hence a Non Disclosure Agreement is signed between the Company and the Analyst.

Client Data isn't easy to get.

A large number of organisations seek to leverage data analytics to make crucial decisions. As organisations become customer-centric, they utilise insights from data to enhance customer experience, while also optimising their daily processes.

- **Banking** - Uses data to make credit decision
- **Telecom** - Uses data to optimise plans for customers and predict customer churn
- **HR** - Helps identify and predicts employee behaviour.
- **Retail** - Drive decisions such as pricing and stocking, Market Basket Analysis is used to analyse what product go well with the another another one .For eg If I have Bread and Pulses go well.
- **Media** The media industry uses data extensively to target viewers better.
- **Advertisers** - Use data to identify the best avenues for targeting customers, while journalists use data visualisation to aid information.

It is recommended that you keep the following data sources handy when looking for data sets.

- GitHub: Awesome Public Datasets
- Open Government Data (OGD) Platform India
- GitHub: DataMeet

Data sourcing is only the first step in the process. After being sourced, the data needs to be cleaned before you can analyse it. In the next session, you will learn how to clean the data for analysis.

Data Cleaning

Once you have procured the data, the next step is to clean it to get rid of **data quality** issues.

There are various types of quality issues when it comes to data, and that's why data cleaning is one of the most time-consuming steps of data analysis. For example, there could be formatting errors (e.g. rows and columns are ill-formatted, unclearly named etc.), missing values, repeated rows, spelling inconsistencies etc. These issues could make it difficult to analyse data and could lead to errors or irrelevant results. Thus, these issues need to be corrected before data is analysed.

https://www.dropbox.com/s/voh6zf3nqf91wdo/Data%2BCleaning%2B_%2BChecklist.xlsx?dl=0

Fix rows and columns

Data quality issue	Examples	How to resolve
Incorrect rows	Header rows, footer rows	Delete
Summary rows	Total, subtotal rows	Delete
Extra rows	Column numbers, indicators, blank rows	Delete
Missing Column Names	Column names as blanks, NA, XX etc.	Add the column names
Inconsistent column names	X1, X2,C4 which give no information about the column	Add column names that give some information about the data
Unnecessary columns	Unidentified columns, irrelevant columns, blank columns	Delete
Columns containing Multiple data values	E.g. address columns containing city, state, country	Split columns into components
No Unique Identifier	E.g. Multiple cities with same name in a column	Combine columns to create unique identifiers e.g. combine City with the State
Misaligned columns	Shifted columns	Align these columns

Fix missing values

Data quality issue	Examples	How to resolve

Disguised Missing values	blank strings, "NA", "XX", "999" etc	Set values as missing values ,blank string or zero is not always NA
Significant number of Missing values in a row/column		Delete rows, columns
Partial missing values	Missing time zone, century etc	Fill the missing values with the correct value

Good methods add information, bad methods exaggerate information.

In case you can add information from reliable external sources, you should use it to replace missing values. But often, it is better to let missing values be and continue with the analysis rather than extrapolate the available information

Let us summarise how to deal with missing values:

- Set values as missing values: Identify values that indicate missing data, and yet are not recognised by the software as such, e.g treat blank strings, "NA", "XX", "999", etc. as missing.
- Adding is good, exaggerating is bad: You should try to get information from reliable external sources as much as possible, but if you can't, then it is better to keep missing values as such rather than exaggerating the existing rows/columns.
- Delete rows, columns: Rows could be deleted if the number of missing values are insignificant in number, as this would not impact the analysis. Columns could be removed if the missing values are quite significant in number.
- Fill partial missing values using business judgement: Missing time zone, century, etc. These values are easily identifiable.

https://pandas.pydata.org/pandas-docs/stable/user_guide/missing_data.html

Standardise values

Data quality issue	Examples	How to resolve
Non-standard units	Convert lbs to kgs, miles/hr to km/hr	Standardise the observations so all of them have the same consistent units

Values with varying Scales	A column containing marks in subjects, with some subject marks out of 50 and others out of 100	Make the scale common. E.g. a percentage scale
Over-precision	4.5312341 kgs, 9.323252 meters	Standardise precision for better presentation of data. 4.5312341 kgs couldbe presented as 4.53 kgs
Remove outliers	Abnormally High and Low values	Correct if by mistake else Remove
Extra characters	Common prefix/suffix, leading/trailing/multiple spaces	Remove the extra characters
Different cases of same words	Uppercase, lowercase, Title Case, Sentence case, etc	Standadise the case/bring to a common case
Non-standard formats	23/10/16 to 2016/10/20 “Modi, Narendra" to “Narendra Modi"	Correct the format/Standardise format for better readability in R

Scaling ensures that the values have a common scale, which makes analysis easier. E.g. let's take a data set containing the grades of students studying at different universities. Some of the universities give grades on a scale of 4, while others give grades on a scale of 10. Therefore, you cannot assume that a GPA of 3 on a scale of 4 is equal to a GPA of 3 on a scale of 10, even though they are same quantitatively. Thus, for the purpose of analysis, these values need to be brought to a common scale, such as the percentage scale.

Fix invalid values

Data quality issue	Examples	How to resolve
Encoding Issues	CP1252 instead of UTF-8	Encode unicode properly

Incorrect data types	Number stored as a string: "12,300"	Convert to Correct data type
Date stored as a string: "2013-Aug"		
String stored as a number: PIN Code "110001" stored as 110001		
Correct values not in list	Non-existent country, PIN code	Delete the invalid values, treat as Missing
Wrong structure	Phone number with over 10 digits	
Correct values beyond range	Temperature less than -273° C (0° K)	
Validate internal rules	Gross sales > Net sales	
Date of delivery > Date of ordering		
If Title is "Mr" then Gender is "M"		

If you have an invalid value problem, and you do not know what accurate values could replace the invalid values, it is recommended to treat these values as missing. E.g. in the case of a string “tr8ml” in a Contact column, it is recommended to remove the invalid value and treat it as a missing value.

Let’s summarise what you learnt about fixing invalid values. You could use this as a checklist for future data cleaning exercises.

- **Encode unicode properly**: In case the data is being read as junk characters, try to change encoding, E.g. CP1252 instead of UTF-8.
- **Convert incorrect data types**: Correct the incorrect data types to the correct data types for ease of analysis. E.g. if numeric values are stored as strings, it would not be possible to calculate metrics such as mean, median, etc. Some of the common data type corrections are — string to number: "12,300" to “12300”; string to date: "2013-Aug" to “2013/08”; number to string: “PIN Code 110001” to "110001"; etc.
- **Correct values that go beyond** range: If some of the values are beyond logical range, e.g. temperature less than -273° C (0° K), you would need to correct them

as required. A close look would help you check if there is scope for correction, or if the value needs to be removed.

- **Correct values not in the list**: Remove values that don't belong to a list. E.g. In a data set containing blood groups of individuals, strings "E" or "F" are invalid values and can be removed.
- **Correct wrong structure**: Values that don't follow a defined structure can be removed. E.g. In a data set containing pin codes of Indian cities, a pin code of 12 digits would be an invalid value and needs to be removed. Similarly, a phone number of 12 digits would be an invalid value.
- **Validate internal rules**: If there are internal rules such as a date of a product's delivery must definitely be after the date of the order, they should be correct and consistent.

Filter data

After you have fixed the missing values, standardised the existing values, and fixed the invalid values, you would get to the last stage of data cleaning. Though you have a largely accurate data set by now, you might not need the entire data set for your analysis. It is important to understand what you need to infer from the data and then choose the relevant parts of the data set for your analysis. Thus, you need to filter the data to get what you need for your analysis.

Data quality issue	Examples	How to resolve
Duplicate data	Identical rows, rows where some columns are identical	Deduplicate Data/ Remove duplicated data
Extra/Unnecessary rows	Rows that are not required in the analysis. E.g if observations before or after a particular date only are required for analysis, other rows become unnecessary	Filter rows to keep only the relevant data.
Columns not relevant to analysis	Columns that are not needed for analysis e.g. Personal Detail columns such as Address, phone column in a dataset for	Filter columns- Pick columns relevant to analysis
Dispersed data	Parts of data required for analysis stored	Bring the data

	in different files or part of different datasets	together, Group by required keys, aggregate the rest