Exploratory Data Analysis

Exploratory Data Analysis

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:

Private Data
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.

On the other hand, 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.

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.

Though data cleaning is often done in a somewhat haphazard way and it is too difficult to define a 'single structured process', we will study data cleaning in the following steps:

- Fix rows and columns
- Fix missing values
- Standardise values
- Fix invalid values
- Filter data

Exploratory Data Analysis

Checklist for Fixing Rows

Delete summary rows: Total, Subtotal rows

Delete incorrect rows: Header rows, Footer rows

Delete extra rows: Column number, indicators, Blank rows, Page No.

Checklist for Fixing Columns

Merge columns for creating unique identifiers if needed: E.g. Merge State, City into Full address

Split columns for more data: Split address to get State and City to analyse each separately

Add column names: Add column names if missing

Rename columns consistently: Abbreviations, encoded columns

Delete columns: Delete unnecessary columns

Align misaligned columns: Dataset may have shifted columns

Exploratory Data Analysis

Standardising Values

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.

One of the concepts that surely caught your attention is outliers. Removing outliers is an important step in data cleaning. An outlier may disproportionately affect the results of your analysis. This may lead to faulty interpretations. It is also important to understand that there is no fixed definition of an outlier. It is left up to the judgment of the analyst to decide the criteria on which data would be categorised as abnormal or an outlier. We will look into one such method in the next session.

Exploratory Data Analysis

Invalid Values

When standardising values, you do not really pay attention to the validity of the actual values of the variables. This is what we will discuss now as you learn to fix invalid values.

A data set can contain invalid values in various forms. Some of the values could be truly invalid, e.g. a string “tr8ml” in a variable containing mobile numbers would make no sense and hence would be better removed. Similarly, a height of 11 ft would be an invalid value in a set containing heights of children.

On the other hand, some invalid values can be corrected. E.g. a numeric value with a data type of string could be converted to its original numeric type. Issues might arise due to python misinterpreting the encoding of a file, thus showing junk characters where there were valid characters. This could be corrected by correctly specifying the encoding or converting the data set to the accurate format before importing.

Exploratory Data Analysis

Filtering 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.

Deduplicate data: Remove identical rows, remove rows where some columns are identical
Filter rows: Filter by segment, filter by date period to get only the rows relevant to the analysis
Filter columns: Pick columns relevant to the analysis
Aggregate data: Group by required keys, aggregate the rest

Exploratory Data Analysis

Univariate Analysis

As the term “univariate” suggests, we deal with analysing variables one at a time. It is important to separately understand each variable before moving on to analysing multiple variables together.

The broad agenda is as follows:

Metadata description
Data distribution plots
Summary metrics

Data Description

Given a data set, the first step is to understand what it contains. Information about a data set can be gained simply by looking at its metadata. Metadata, in simple terms, is the data that describes the each variable in detail. Information such as the size of the data set, how and when the data set was created, what the rows and variables represent, etc. are captured in metadata.

Exploratory Data Analysis

Types of Variables

You learnt the difference between **ordered** and **unordered categorical variables** -
Ordered ones have some kind of ordering. Some examples are

Salary = High-Medium-low
Month = Jan-Feb-Mar etc.

Unordered ones do not have the notion of high-low, more-less etc. Example:

Type of loan taken by a person = home, personal, auto etc.
Organisation of a person = Sales, marketing, HR etc.

Apart from the two types of categorical variables, the other most common type is **quantitative variables**. These are simply numeric variables which can be added up, multiplied, divided etc. For example, salary, number of bank accounts, runs scored by a batsman, the mileage of a car etc.

So far, we have discussed the following types of variables:

Categorical variables

Unordered
Ordered

Quantitative / numeric variables

Exploratory Data Analysis

It is important to note that **rank-frequency plots** enable you to extract meaning even from seemingly trivial **unordered categorical variables** such as country, name of an artist, name of a github user etc.

The objective here is not to put excessive focus on power laws or rank-frequency plots, but rather to understand that non-trivial analysis is possible even on unordered categorical variables, and that plots can help you out in that process.

Why plotting on a log-log scale helps

The objective of using a log scale is to make the plot readable by changing the scale. For example, the first ranked item had a frequency of 29000, the second ranked had 3500, the seventh had 700 and most others had very low frequencies such as 100, 80, 21 etc. The range of frequencies is too large to fit on the plot.

Plotting on a log scale compresses the values to a smaller scale which makes the plot easy to read.

This happens because log(x) is a much smaller number than x. For example, log(10) = 1, log(100) = 2, log(1000) = 3 and so on. Thus, log(29000) is now approx. 4.5, log(3500) is approx. 3.5 and so on. What was earlier varying from 29000 to 1 is now compressed between 4.5 and 0, making the values easier to read on a plot.

Exploratory Data Analysis

To summarise,

Plots are immensely helpful in identifying hidden patterns in the data
It is possible to extract meaningful insights from unordered categorical variables using rank-frequency plots
Rank-frequency plots of unordered categorical variables, when plotted on a log-log scale, typically result in a power law distribution

Exploratory Data Analysis

Quantitative Variables - Univariate Analysis

Mean and median are single values that broadly give a representation of the entire data, it is very important to understand when to use these metrics to avoid doing inaccurate analysis.

While mean gives an average of all the values, median gives a typical value that could be used to represent the entire group. As a simple rule of thumb, always question someone if someone uses the mean, since median is almost always a better measure of ‘representativeness’.

Standard deviation and interquartile difference are both used to represent the spread of the data.

Interquartile difference is a much better metric than standard deviation if there are outliers in the data. This is because the standard deviation will be influenced by outliers while the interquartile difference will simply ignore them

Exploratory Data Analysis

Outliers

Roughly speaking, outliers are abnormally large or small values. There is no one fixed rule in deciding outliers. However, you would have noticed that going from the 99th to the 100th percentile, there is an approximately 25x increase in the number of shares.

You may have also noted that in the lower quartiles, there is only about 5% increase in the number of shares per percentile. This increases to about 10% per percentile in the higher quartiles and 20% beyond the 95th percentile. In this case, some articles in the top quartiles are clearly outliers.

To classify some articles as outliers, you may need to consult your client or the business to understand the reasons behind abnormal values. In some cases, they are justifiable, whereas in others they cannot be explained and are thus labelled as outliers.

For example, let’s say that you decide to label all the articles beyond the **95th percentile** as outliers since you observe a roughly 20% increase in the number of shares beyond that. Thus, the last article you include will be the one at the 95th percentile. You remove every article beyond that point from the data set.

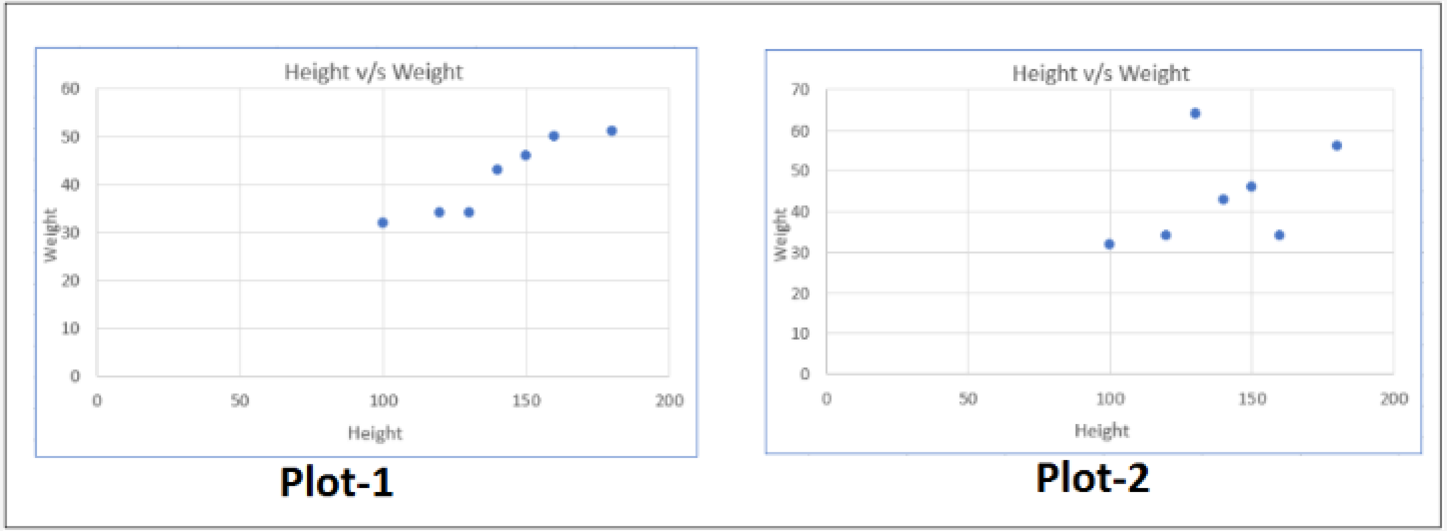
Exploratory Data Analysis

To summarise, **correlation** is a number between **-1 and 1** which quantifies the extent to which two variables 'correlate' with each other.
If **one increases** as the **other increases**, the correlation is **positive**
If **one decreases** as the **other increases**, the correlation is **negative**
If one **stays constant** as the **other varies**, the correlation is **zero**

In general, a positive correlation means that two variables will increase together and decrease together, e.g. an increase in rain is accompanied by an increase in humidity. A negative correlation means that if one variable increases the other decreases, e.g. in some cases, as the price of a commodity decreases its demand increases.

A perfect positive correlation means that the correlation coefficient is exactly 1. This implies that as one variable moves, either up or down, the other one moves in the same direction. A perfect negative correlation means that two variables move in opposite directions, while a zero correlation implies no relationship at all.

Exploratory Data Analysis



Plot-1

Plot-2

Fig-1 Correlation Plots

Exploratory Data Analysis

What are Derived Metrics?
Sometimes, you would not get the most valuable insights by analysing the data available to you. You often need to create new variables using the existing ones to get meaningful insights.

New variables could be created based on your business understanding or they can be suggested by your clients. Let's understand how business understanding plays an important role in deriving new variables

Broadly, there are three different types of derived metrics:

1. Type-driven metrics
2. Business-driven metrics
3. Data-driven metrics

Exploratory Data Analysis

Type-Driven Metrics

These metrics can be derived by understanding the variable’s typology. You have already learnt one simple way of classifying variables/attributes — **categorical (ordered, unordered)** and **quantitative or numeric**. Similarly, there are various other ways of classification, one of which is Steven's typology.

Steven’s typology classifies variables into four types — nominal, ordinal, interval and ratio:

Nominal variables: Categorical variables, where the categories **differ only by their names**; there is **no order** among categories, e.g. colour (red, blue, green), gender (male, female), department (HR, analytics, sales)

These are the most basic form of categorical variables

Ordinal variables: Categories follow a certain **order**, but the **mathematical difference between categories is not meaningful**, e.g. education level (primary school, high school, college), height (high, medium, low), performance (bad, good, excellent), etc.

Ordinal variables are **nominal as well**

Interval variables: Categories follow a certain order, and the **mathematical difference between categories is meaningful** but division or multiplication is not, e.g. temperature in degrees celsius (the difference between 40 and 30 degrees C is meaningful, but 30 degrees x 40 degrees is not), dates (the difference between two dates is the number of days between them, but 25th May / 5th June is meaningless), etc.

Interval variables are **both nominal and ordinal**

Ratio variables: Apart from the mathematical difference, the ratio (division/multiplication) is possible, e.g. sales in dollars ($100 is twice $50), marks of students (50 is half of 100), etc.

Exploratory Data Analysis

Data-driven metrics can be created based on the variables present in the existing data set.

For example, if you have two variables in your data set such as "weight" and "height" which shows a high correlation. So, instead of analysing "**weight**" and "**height**" variables separately, you can think of deriving a new metric "Body Mass Index **(BMI)**".

Once you get the BMI, you can easily categorise people based on their fitness, e.g. a BMI below 18.5 should be considered as an underweight category, while BMI above 30.0 is considered as obese, by standard norms. This is how data-driven metrics can help you discover hidden patterns out of the data.