Module:Summary

Exploratory Data Analysis

Exploratory data analysis is the first and foremost step to analyse any kind of data. Rather than a specific set of procedures, EDA is an approach, or a philosophy, which seeks to explore the most important and often hidden patterns in a data set. In EDA, we explore the data and try to come up with a hypothesis about it which we can later test using hypothesis testing. Statisticians use it to take a bird's eye view of the data and try to make some sense of it.

In this **module**, we will cover the following topics:

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 to come up with actionable insights. 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:

1. **Public Data**
2. **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. Public data is available on the internet on various platforms. A lot of data sets are available for direct analysis, whereas some of the data have to be manually extracted and converted into a format that is fit for analysis.

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. 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.

Data Cleaning

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 merged), 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:

1. **Fix rows and columns**
2. **Fix missing values**
3. **Standardise values**
4. **Fix invalid values**
5. **Filter data**

Fix rows and columns

1. Fix Rows Examples:

FIX ROWS	EXAMPLES
Delete incorrect rows	Unnecessary header rows, footer rows
Delete summary rows	Total, subtotal rows
Delete extra rows	Column number indicator rows, blank rows

Figure 1:Fix Rows

2. Fix Columns Examples:

FIX COLUMNS	EXAMPLES
Add column names if missing	Missing header row
Rename columns consistently	Abbreviations, encoded columns
Delete unnecessary columns	Unidentified, irrelevant columns
Split columns for more data	Split http://host:port/path into [Host, Port, Path]
Merge columns for identifiers	Firstname, Lastname -> NameState, District -> FullDistrict
Align misaligned columns	Shifted columns

Figure 2:Fix Columns

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

Fix 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.

Standardising Values

Checklist:

Standardise units: Ensure all observations under a variable have a common and consistent unit, e.g. convert lbs to kgs, miles/hr to km/hr, etc.

Scale values if required: Make sure the observations under a variable have a common scale

Standardise precision for better presentation of data, e.g. 4.5312341 kgs to 4.53 kgs.

Remove outliers: Remove high and low values that would disproportionately affect the results of your analysis.

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 R 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.

- **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.

Filtering Data

- **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

Univariate Analysis

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

The agenda of univariate analysis is to understand:

- Metadata description
- Data distribution plots
- Summary metrics

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**.

Types of Variables

- **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.

Distribution plots reveal interesting insights about the data. You can observe various visible patterns in the plots and try to understand how they came to be.

Summary metrics are used to obtain a quantitative summary of the data. Not all metrics can be used everywhere. Thus, it is important to understand the data and then choose what metric to use to summarise the data.

Segmented Univariate Analysis

The broad agenda of "Segmented Univariate Analysis" is as follows:

- Basis of segmentation
- Comparison of averages
- Comparison of other metrics

Basis of segmentation:

The entire **segmentation process** can be divided into four parts:

- Take raw data
- Group by dimensions
- Summarise using a relevant metric such as mean, median, etc.
- Compare the aggregated metric across groups/categories

Comparison of Averages

```
Don't blindly believe in the averages of the buckets – you need to observe the distribution of
each bucket closely and ask yourself if the difference in means is significant enough to draw a
conclusion. If the difference in means is small, you may not be able to draw inferences. In suc
h cases, a technique called hypothesis testing is used to ascertain whether the difference in m
eans is significant or due to randomness.“ Don't worry if you do not get the concept of hypothe
sis correctly, It will be dealt separately in hypothesis module.
```

Bivariate Analysis

- Bivariate analysis on continuous variables

Correlation is a metric to find the relationship between the variables . It 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.

- Bivariate analysis on categorical variables

The categorical bivariate analysis is essentially an extension of the **segmented univariate** analysis to another categorical variable. In univariate analysis, you compare metrics such as 'mean of X' across various segments of a categorical variable, e.g. mean marks of a student are higher for 'degree and above' than other levels of the mother's education; or the median income of educated parents is higher than that of uneducated ones, etc.

In the categorical bivariate analysis, you extend this comparison to other categorical variables and ask — is this true for all categories of another variable, say, men and women? Take another categorical variable, such as state, and ask — is the median income of educated parents higher than that of uneducated ones in all states? Thus, you are drilling down into another categorical variable and getting closer to the true patterns in the data. In fact, you may also go to the next level and ask — is the median income of educated parents higher than that of uneducated ones (variable 1) in all states (variable 2) for all age groups (variable 3)? This is what you may call 'trivariate analysis', and though it gives you a more granular version of the truth, it gets a bit complex to make sense of and explain to others (and hence it is not usually done in EDA). Thus, remember that doing only conducting segmented univariate analysis may deceive you into thinking that a certain phenomenon is true without asking the question — is it true for all sub-populations or is it true only when you aggregate information across the entire population?

So in general, there are two fundamental aspects of analysing categorical variables:

1. To see the distribution of two categorical variables. For example, if you want to compare the number of boys and girls who play games, you can make a 'cross table' as given below:

	Everyday	Never	Once a month	Once a week	Total
Boy	3474	154	150	780	4558
Girl	2776	175	200	1046	4197
Total	6250	329	350	1826	8755

Figure 3:Cross Table

From this table, firstly, you can compare boys and girls across a fixed level of 'play games', e.g. a higher number of boys play games every day than girls, a higher number of girls never play games than boys, etc. And secondly, you can compare the levels of 'play games' across a fixed value of gender, e.g. most boys play every day and very few play once a month or never.

2. To see the distribution of two categorical variables with one continuous variable.For example, you saw that how a student's percentage in science is distributed based on the father's occupation (categorical variable 1) and the poverty level (categorical variable 2).

Derived Metrics

There are three different types of derived metrics::

- **Type-driven metrics**
- **Business-driven metrics**
- **Data-driven metrics**

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, e.g. temperature in degrees celsius (the difference between 40 and 30 degrees C is meaningful), dates (the difference between two dates is the number of days between them), 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.

  Ratio variables are nominal, ordinal and interval type

Understanding types of variables enables you to derive new metrics of types different from the same column. For example, age in years is a ratio attributes, but you can convert it into an ordinal type by binning it into categories such as children (< 13 years), teenagers (13-19 years), young adults (20-25 years), etc. This enables you to ask questions, e.g. do teenagers do X better than children, are young adults more likely to do X than the other two types, etc. Here, X is an action you are interested in finding.

Business Driven Metrics:

It is derived from the existing variables but it requires the domain expertise. Driving metrics from the business perspective is not an easy task. Without understanding the domain correctly, deriving insights becomes difficult and prone to errors.

Data Driven Metrics:

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.

Exploratory data analysis helps you, as a data analyst, to look beyond the data. It is a never ending process — the more you explore the data, the more insights you get. Almost 80% of the time, you would spend your time as a data analyst understanding the data and solving various business problems through EDA. If you understand EDA properly, then half the battle is won.

So far in this module, you have learnt the five most crucial topics for any kind of analysis. They are as follows:

- Understanding domain
- Understanding data and preparing it for analysis
- Univariate analysis and segmented univariate analysis
- Bivariate analysis
- Deriving new metrics from the existing data