upGrad
Campus

Course : Exploratory Data Analysis

Lecture On : Overview of EDA

Agendas

- We will learn the concepts of Exploratory Data Analysis or EDA.
- Topics to be covered in this session:
    - Data Sourcing
    - Data Cleaning
    - Univariate Analysis
    - Segmented Univariate
    - Bivariate Analysis
    - Derived Metrics

What is Exploratory Data Analysis(EDA)?

In simple terms, it means trying to understand the given data much better, so that we can make some sense out of it.

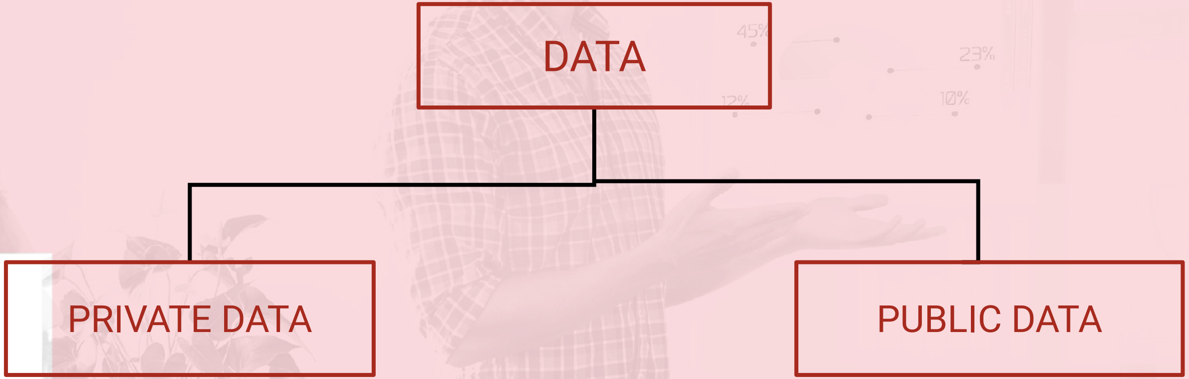
EDA is arguably the most important and revelatory step in any kind of data analysis.

In statistics, **exploratory data analysis** is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

DATA SOURCING

Data Sourcing is the process of finding and loading the data into our system.

Depending upon its source, data can be classified into two types:



DATA

PRIVATE DATA

PUBLIC DATA

PRIVATE DATA

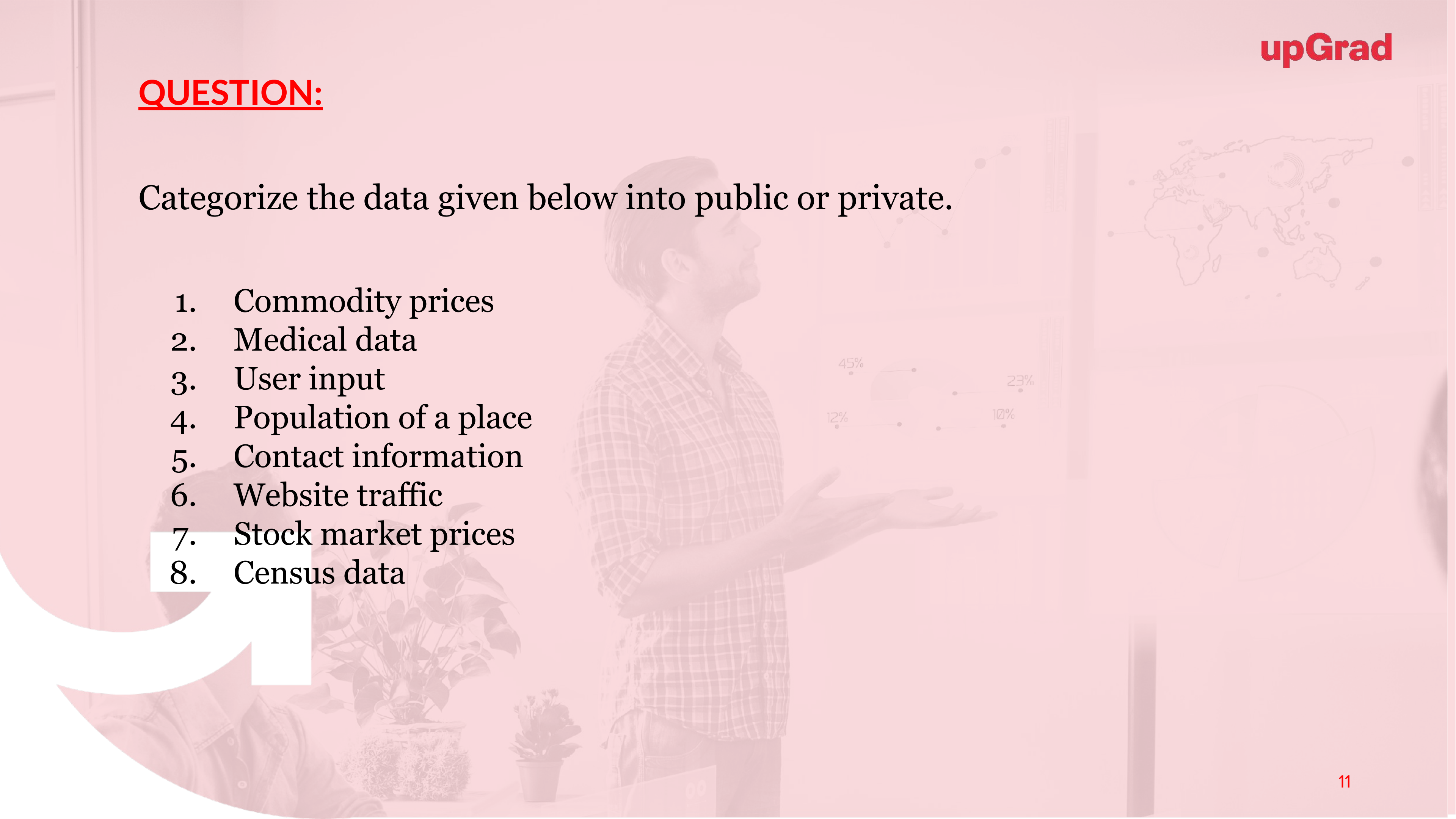
Private data is given by private organizations. There are some security and privacy concerns attached to it. This type of data is used for mainly organizations internal analysis.

PUBLIC DATA

This type of Data is available to everyone. We can find this in government websites and public organizations etc. Anyone can access this data, we do not need any special permissions or approval.

Who uses data?

Banking, Telecommunications, HR Industries, Advertisers, Media and Retail Industries



QUESTION:

Categorize the data given below into public or private.

1. Commodity prices
2. Medical data
3. User input
4. Population of a place
5. Contact information
6. Website traffic
7. Stock market prices
8. Census data

SOLUTION:

Categorize the data given below into public or private.

1. Commodity prices - ***public***
2. Medical data - ***private***
3. User input - ***private***
4. Population of a place - ***public***
5. Contact information - ***private***
6. Website traffic - ***private***
7. Stock market prices - ***public***
8. Census data - ***public***

DATA CLEANING

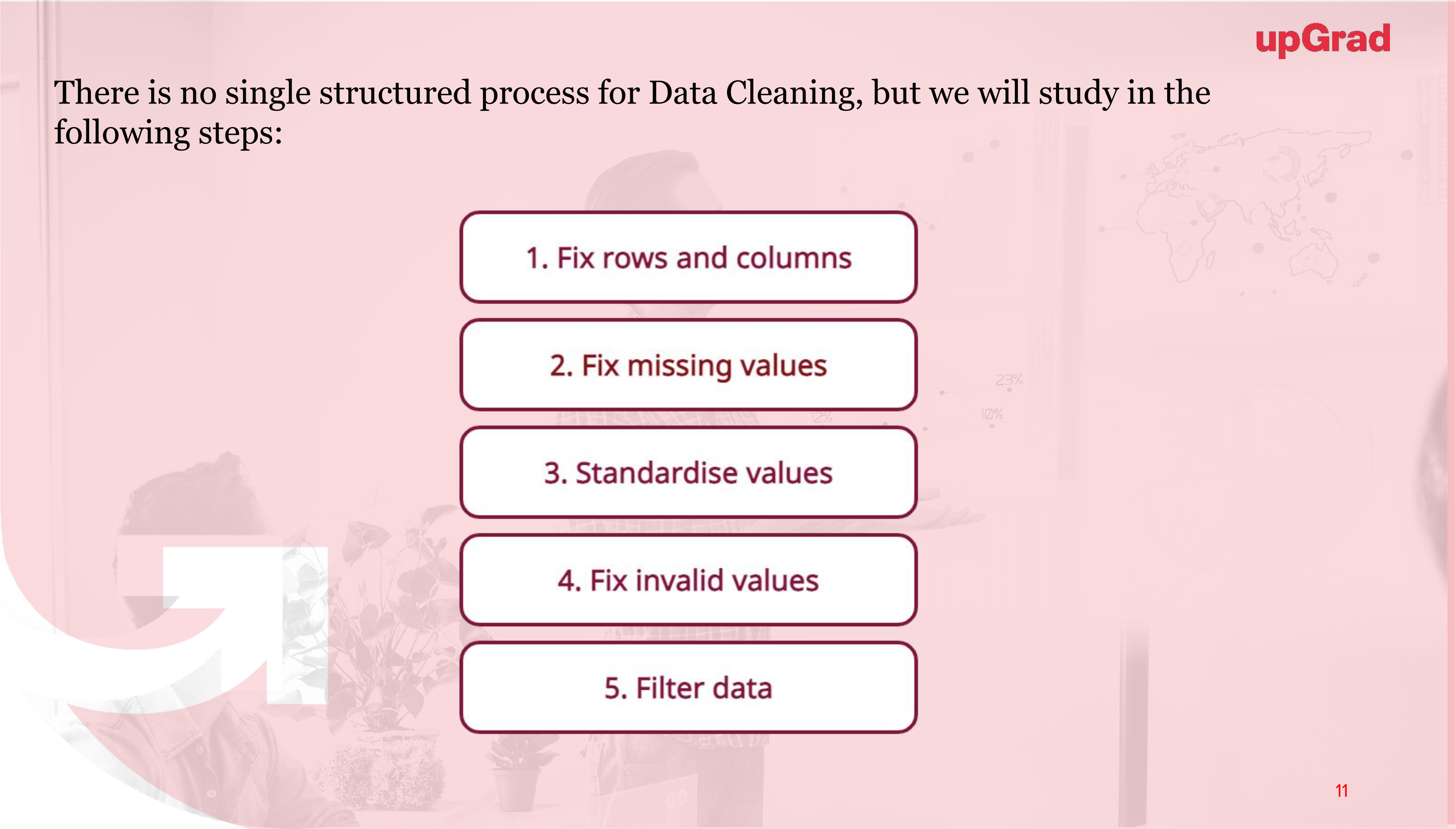
After completing the Data Sourcing, the next step in the process of EDA is **Data Cleaning**. It is very important to get rid of the irregularities and clean the data after sourcing it into our system.

There could be different kinds of errors in the data like *formatting errors, 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.

There is no single structured process for Data Cleaning, but we will study in the following steps:

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



1. Fixing rows and 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 number*

Checklist for fixing columns

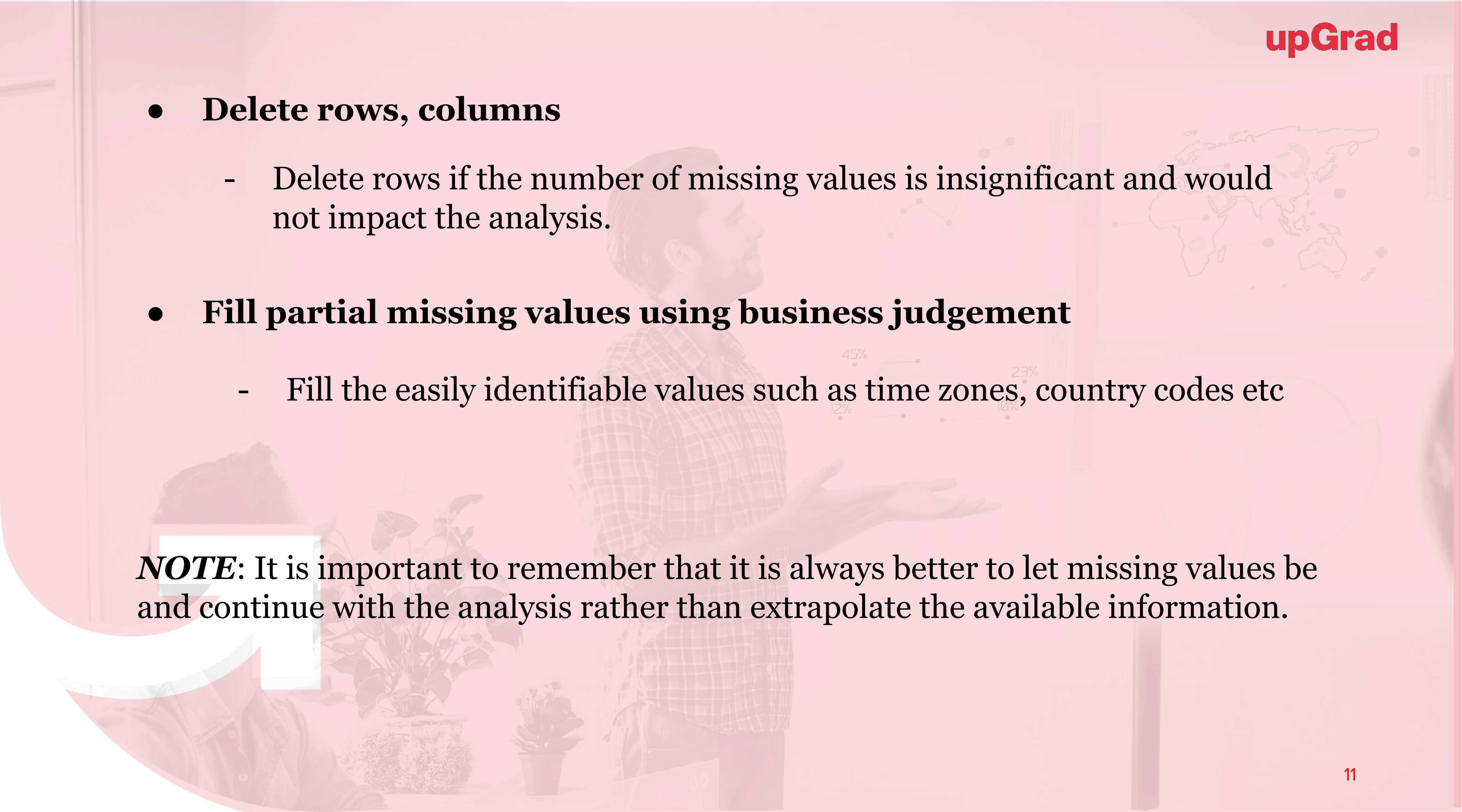
- **Merge columns for creating unique identifiers if needed**: *e.g., merge state and city details* into the full address.
- **Split columns for more data**: Eg: *Split the address to get state and city details so that you can analyse each separately.*
- **Add missing column names**.
- **Rename columns consistently, with abbreviations and encoded columns.**
- **Delete unnecessary columns.**
- **Align misaligned columns**: *The data set might have shifted columns.*

2. Missing Values

If there are missing values in the Dataset before doing any analysis, we need to handle those missing values.

How to deal with missing values:

- **Set values as missing values**
    - Identify the values that indicate missing data (NA, 99) and replace such data with a blank cell
- **Adding is good, exaggerating is bad**
    - If you can not find information from a reliable external source, it is better to keep the missing values as such



- **Delete rows, columns**
  - Delete rows if the number of missing values is insignificant and would not impact the analysis.

- **Fill partial missing values using business judgement**
  - Fill the easily identifiable values such as time zones, country codes etc

NOTE: It is important to remember that it is always better to let missing values be and continue with the analysis rather than extrapolate the available information.

3. Standardise Values

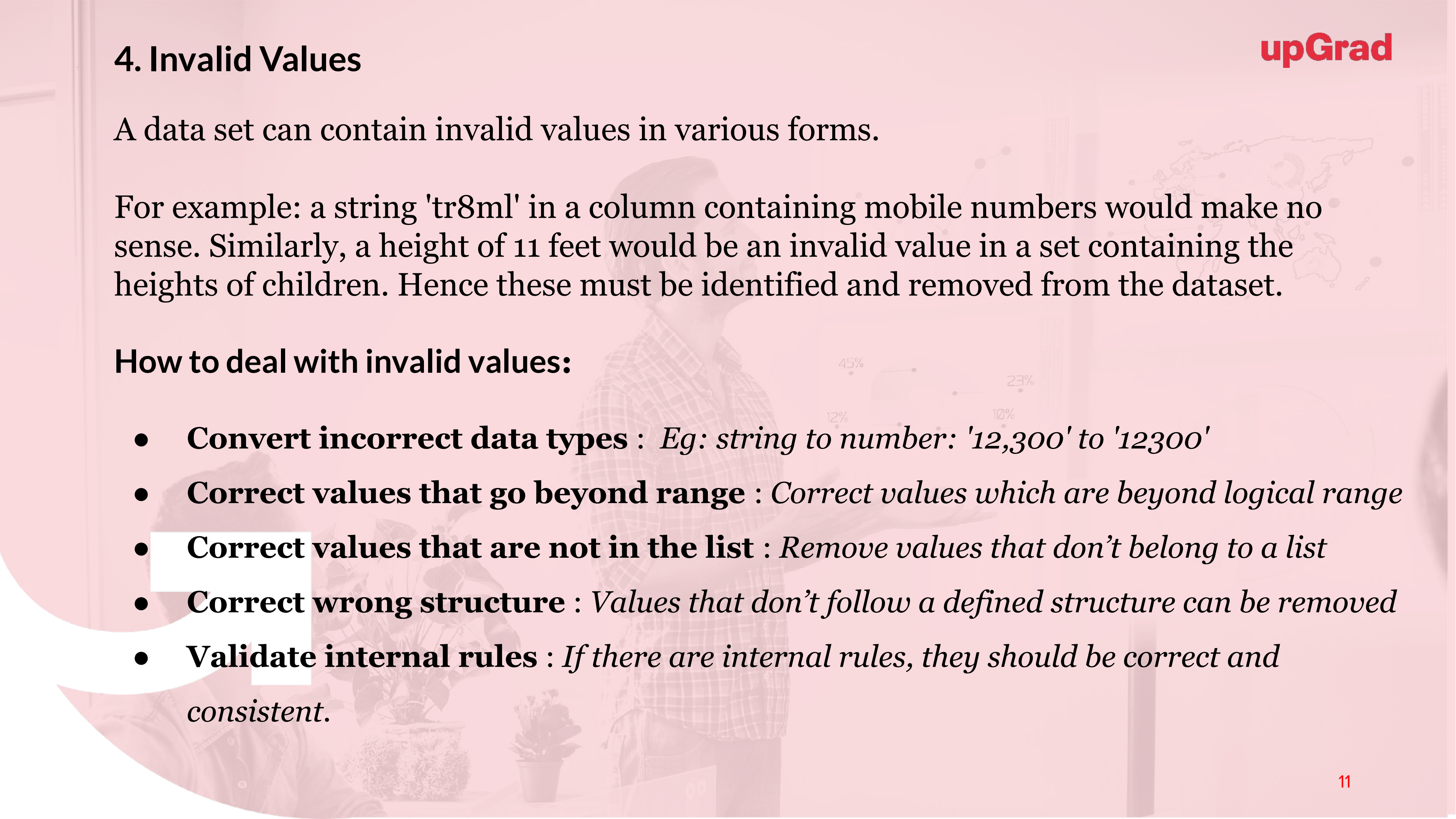
To perform data analysis on a set of values, we have to make sure the values in the same column should be on the same scale.

For example, if the data contains the values of the top speed of different companies' cars, then the whole column should be either in meters/sec scale or miles/sec scale.

Things to keep in mind for standardising variables:

- **Standardise units**
- **Scale values if required**
- **Standardise precision**
- **Remove outliers**

***NOTE*:** An outlier may disproportionately affect the results of your analysis. This may lead to faulty interpretations. One must be careful not to let an outlier affect their analysis.



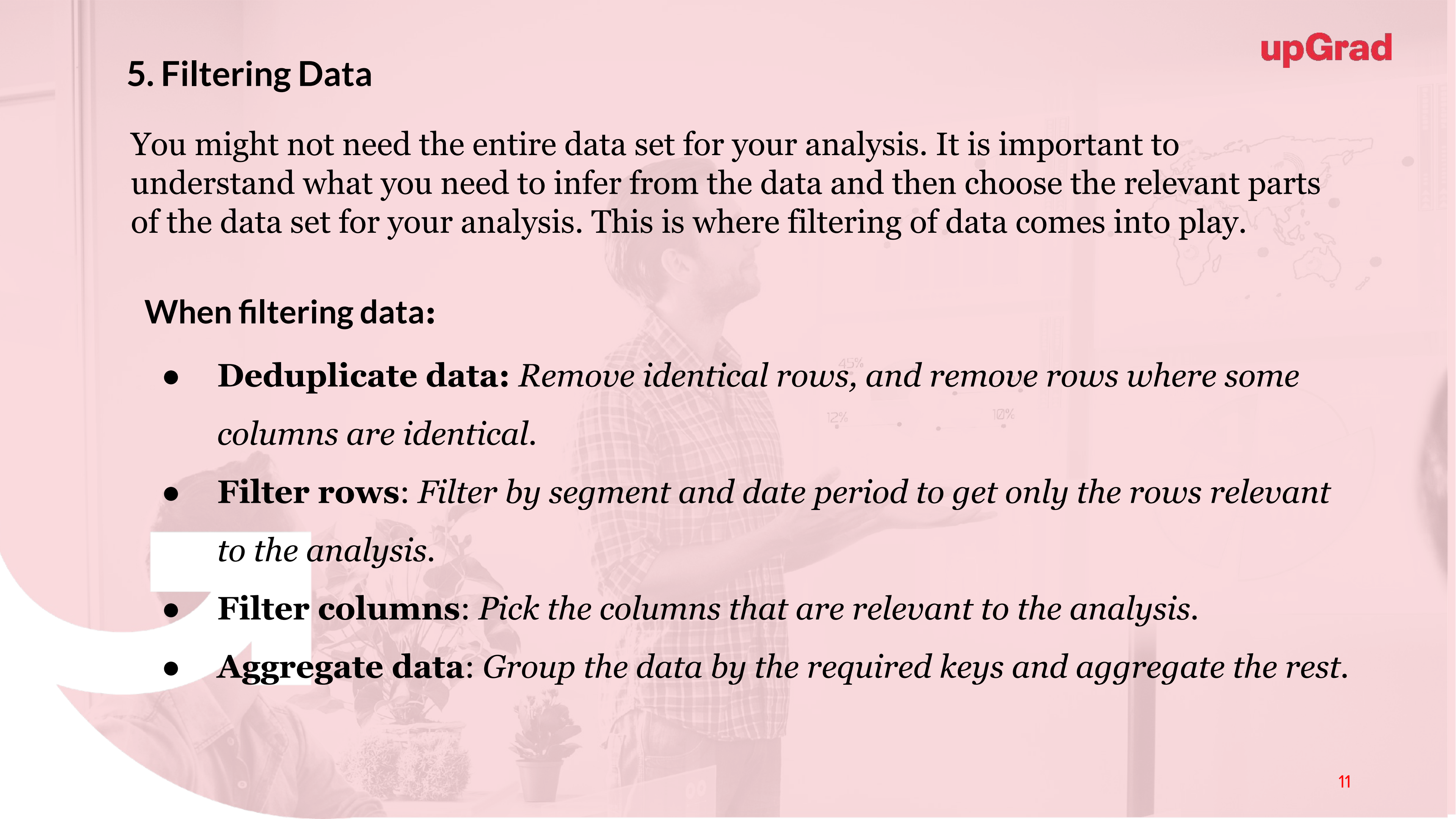
4. Invalid Values

A data set can contain invalid values in various forms.

For example: a string 'tr8ml' in a column containing mobile numbers would make no sense. Similarly, a height of 11 feet would be an invalid value in a set containing the heights of children. Hence these must be identified and removed from the dataset.

How to deal with invalid values:

- **Convert incorrect data types** : *Eg: string to number: '12,300' to '12300'*
- **Correct values that go beyond range** : *Correct values which are beyond logical range*
- **Correct values that are not in the list** : *Remove values that don't belong to a list*
- **Correct wrong structure** : *Values that don't follow a defined structure can be removed*
- **Validate internal rules** : *If there are internal rules, they should be correct and consistent.*



5. Filtering Data

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. This is where filtering of data comes into play.

When filtering data:

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

QUESTION:

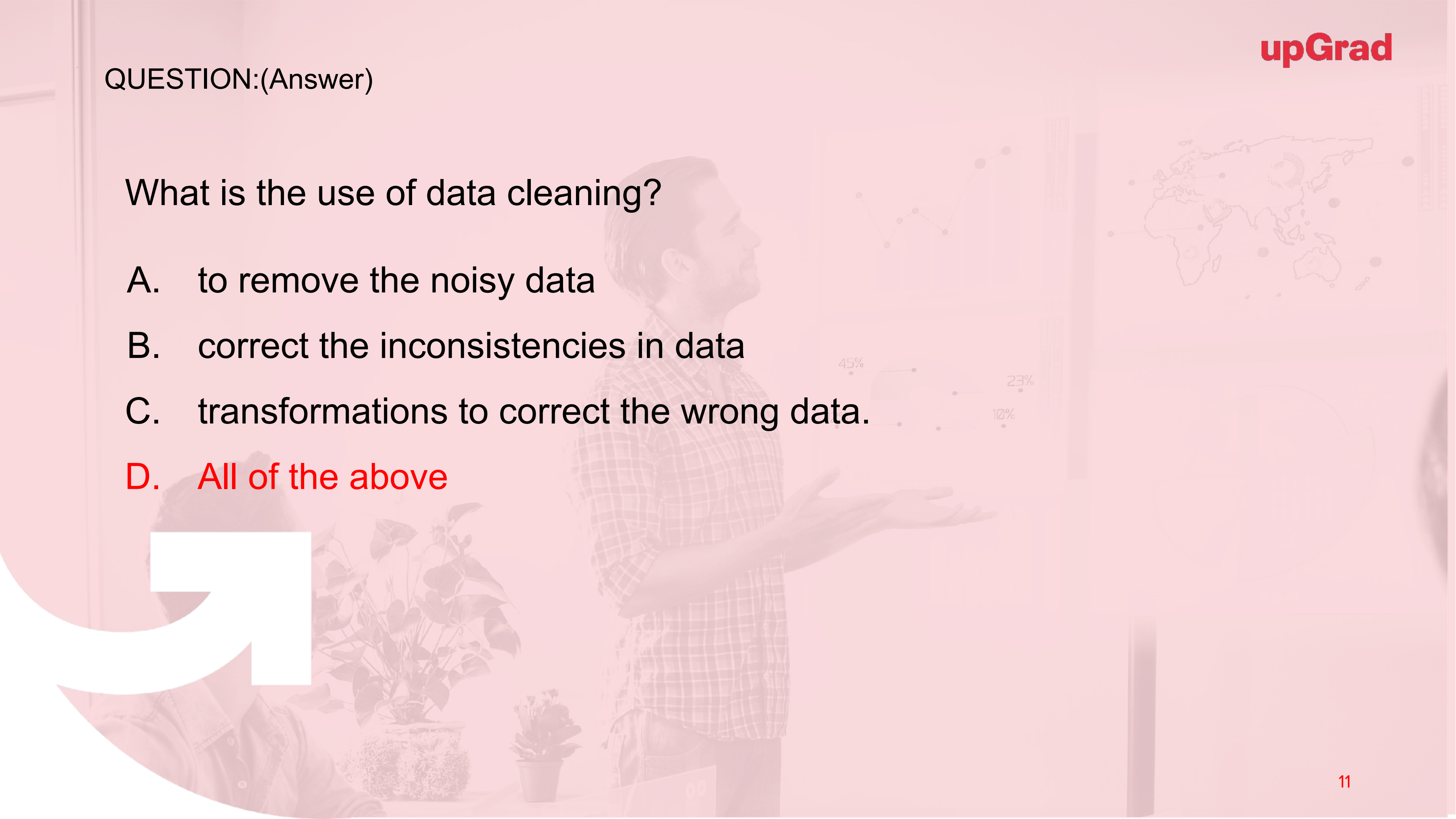
What is the use of data cleaning?

A. to remove the noisy data

B. correct the inconsistencies in data

C. transformations to correct the wrong data.

D. All of the above



QUESTION:(Answer)

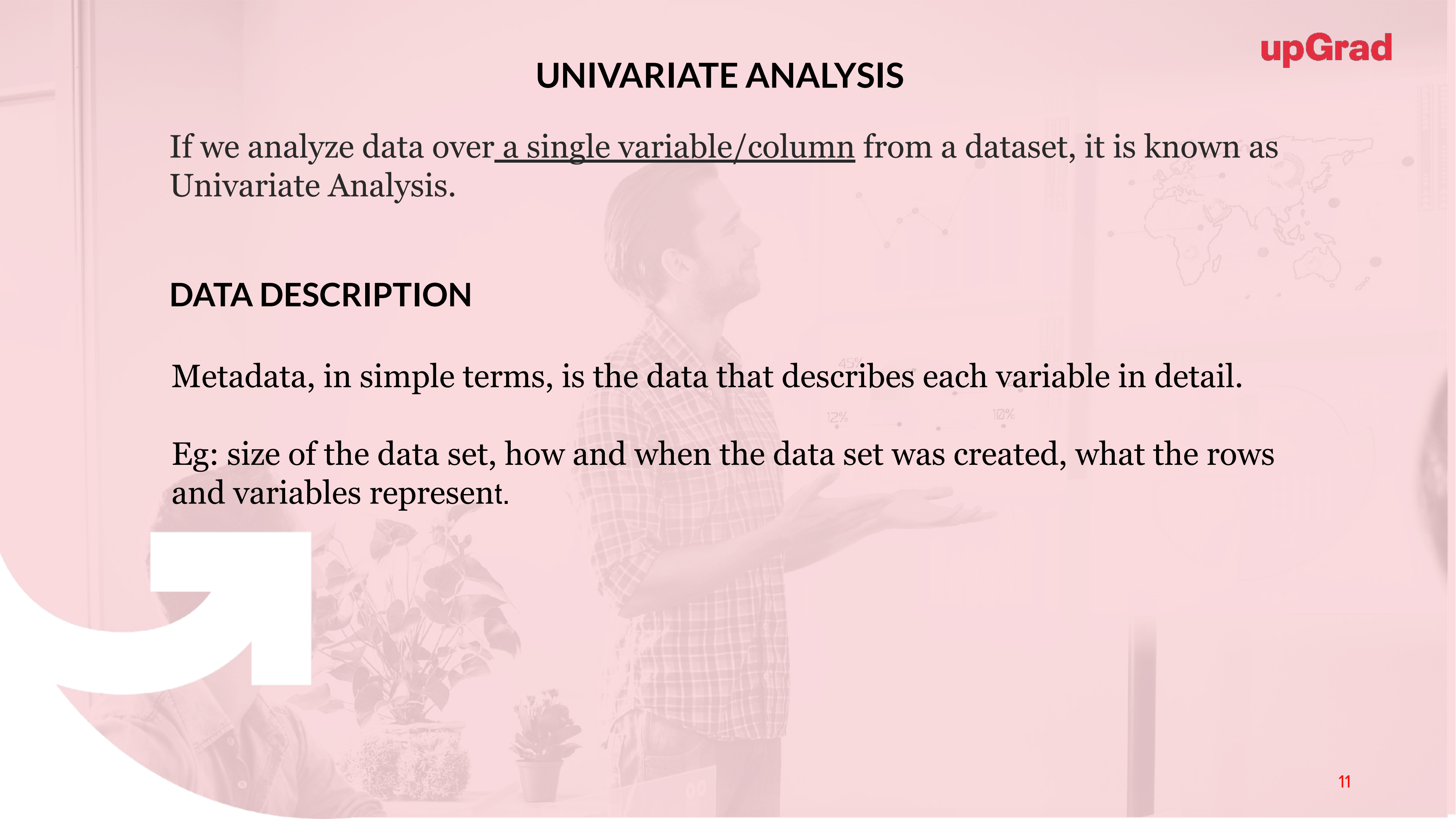
What is the use of data cleaning?

A. to remove the noisy data

B. correct the inconsistencies in data

C. transformations to correct the wrong data.

D. All of the above



UNIVARIATE ANALYSIS

If we analyze data over a single variable/column from a dataset, it is known as Univariate Analysis.

DATA DESCRIPTION

Metadata, in simple terms, is the data that describes each variable in detail.

Eg: size of the data set, how and when the data set was created, what the rows and variables represent.

Types of variables

1. Categorical Variables

- Unordered

  An unordered variable is a categorical variable that has no defined order.
  Eg: Type of loan taken by a person = home, personal, auto, etc.

- Ordered

  Ordered variables are those variables that have a natural rank of order.
  Eg: Salary = High-Medium-low

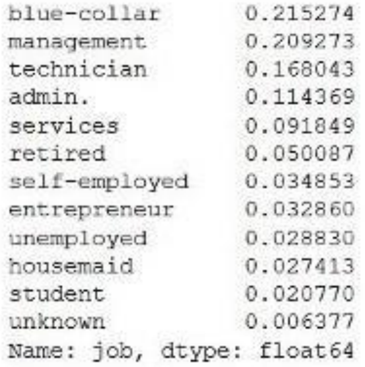
2. Quantitative/ Numeric variables

These are simply numeric variables, which can be added up, multiplied, divided, etc.
Eg: salary, number of bank accounts, runs scored by a batsman, the mileage of a car, etc.

Unordered Categorical Variables

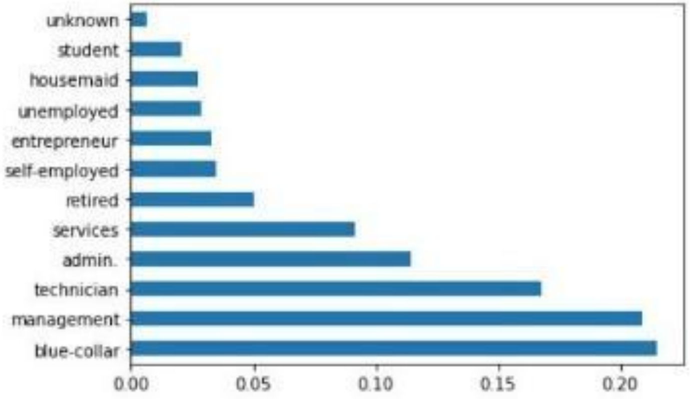
- Plots are immensely helpful in identifying hidden patterns in data.
- It is possible to extract meaningful insights from unordered categorical variables using rank-frequency plots.
- Rank-frequency plots of unordered categorical variables, when plotted on a log-log scale, typically result in a power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

Eg:



```
blue-collar      0.215274
management       0.209273
technician       0.168043
admin.           0.114369
services         0.091849
retired          0.050087
self-employed    0.034853
entrepreneur     0.032860
unemployed       0.028830
housemaid        0.027413
student          0.020770
unknown          0.006377
Name: job, dtype: float64
```

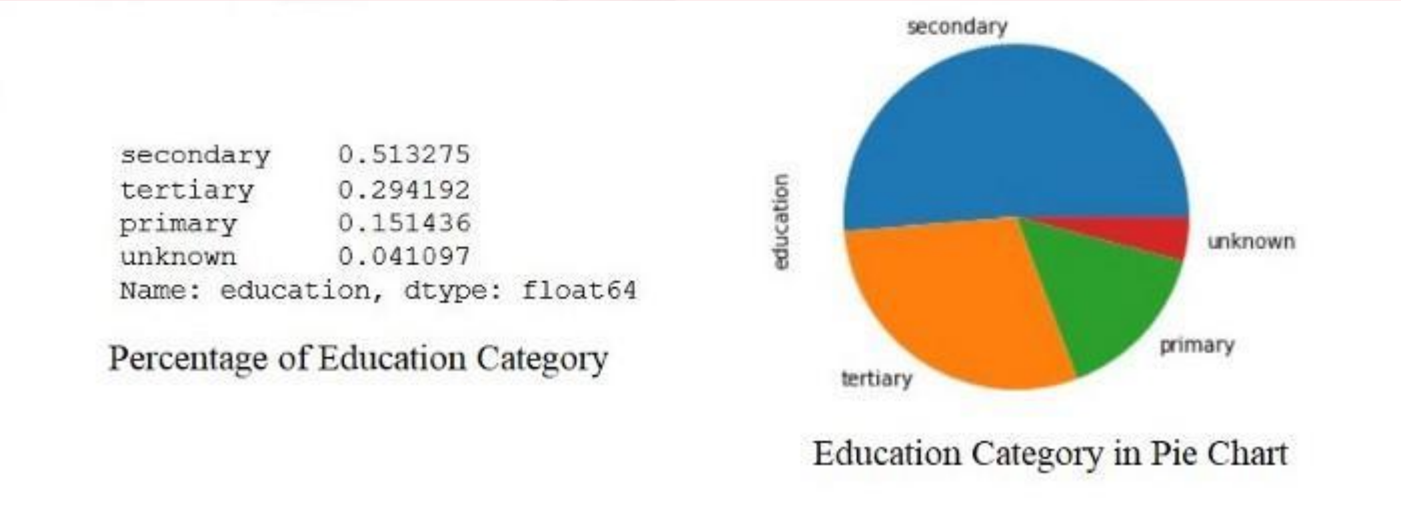
Percentage of Job Categories



Bar Plot of Job Column

Ordered Categorical Variables

- Derive insights from ordered categorical variables through univariate analysis
  Eg:



```
secondary    0.513275
tertiary     0.294192
primary      0.151436
unknown      0.041097
Name: education, dtype: float64
```

Percentage of Education Category

Education Category in Pie Chart

Quantitative Variables

Mean: This is the sum of all the data values, divided by the total number of sample values.

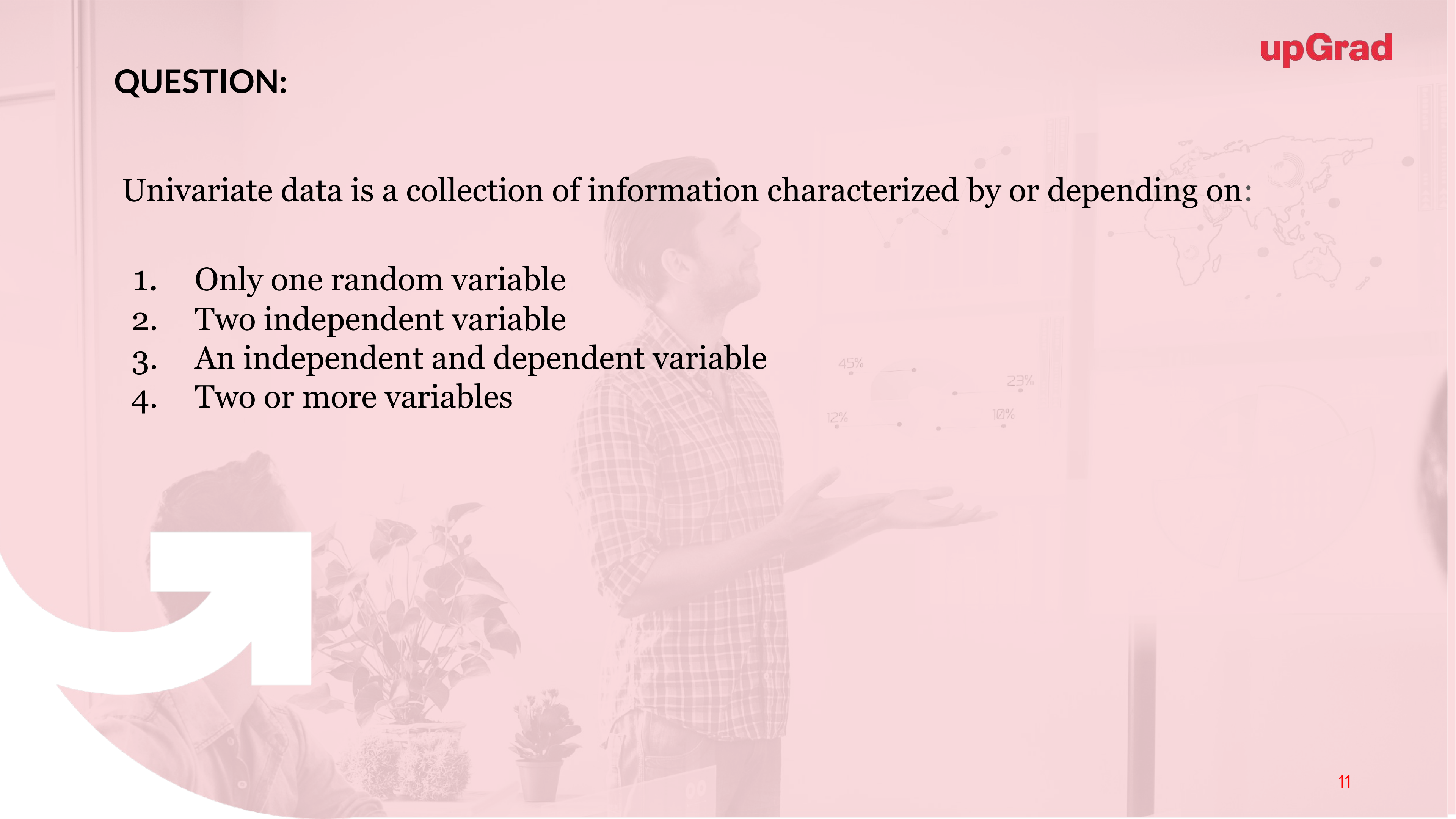
Mode: In your sample data, the value that has the highest frequency is the mode.

Median: If you arrange the sample data in ascending order of frequency, from left to right, the value in the middle is called the median.

Standard Deviation: Measure of the amount of variation or dispersion of a set of values.

Interquartile Difference: The IQR of a set of values is calculated as the difference between the upper and lower quartiles.

NOTE: The 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, whereas the interquartile difference will simply ignore them.



QUESTION:

Univariate data is a collection of information characterized by or depending on:

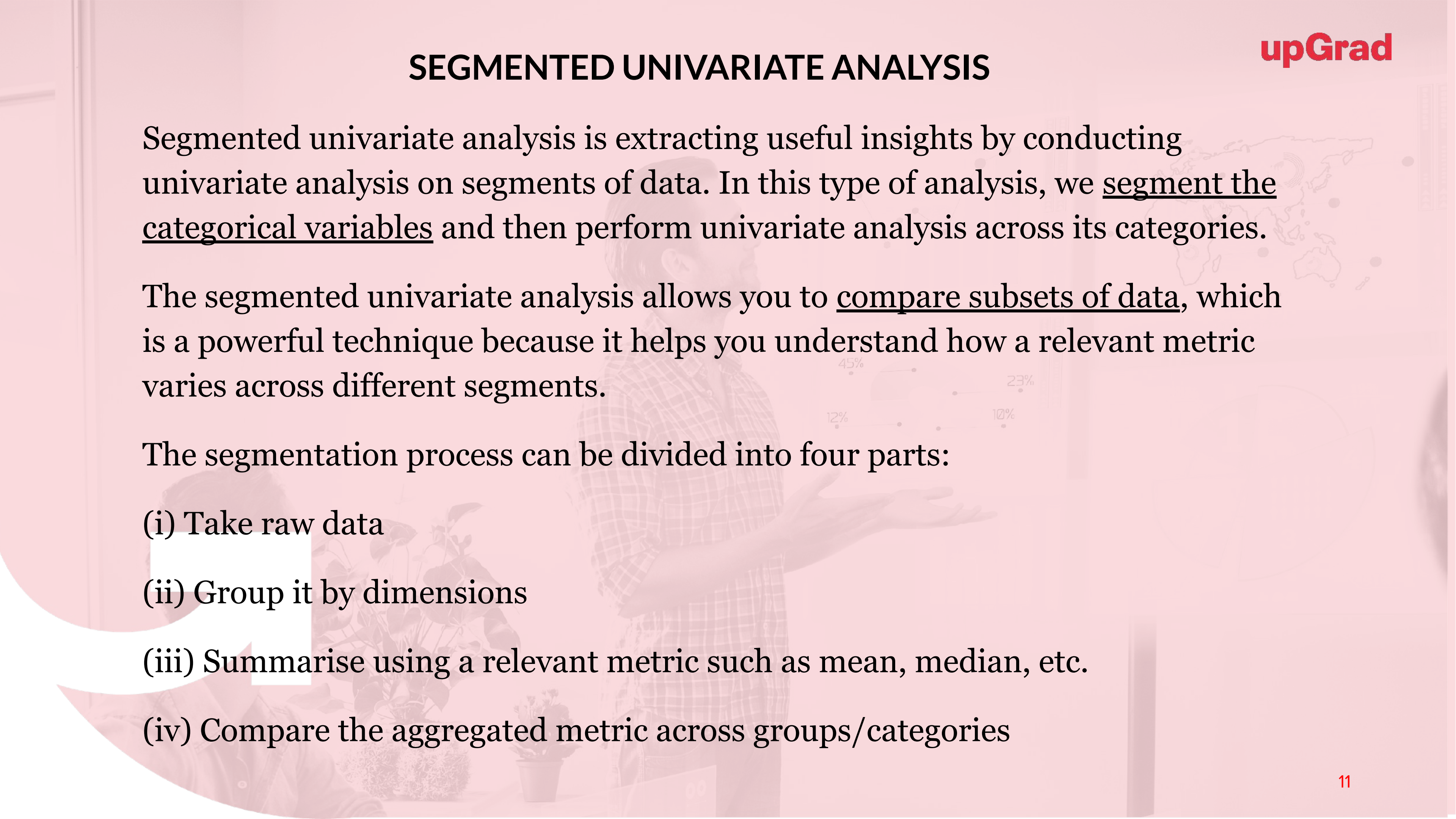
1. Only one random variable
2. Two independent variable
3. An independent and dependent variable
4. Two or more variables

SOLUTION:

Univariate data is a collection of information characterized by or depending on:

1. **Only one random variable**
2. Two independent variable
3. An independent and dependent variable
4. Two or more variables

SEGMENTED UNIVARIATE ANALYSIS



Segmented univariate analysis is extracting useful insights by conducting univariate analysis on segments of data. In this type of analysis, we segment the categorical variables and then perform univariate analysis across its categories.

The segmented univariate analysis allows you to compare subsets of data, which is a powerful technique because it helps you understand how a relevant metric varies across different segments.

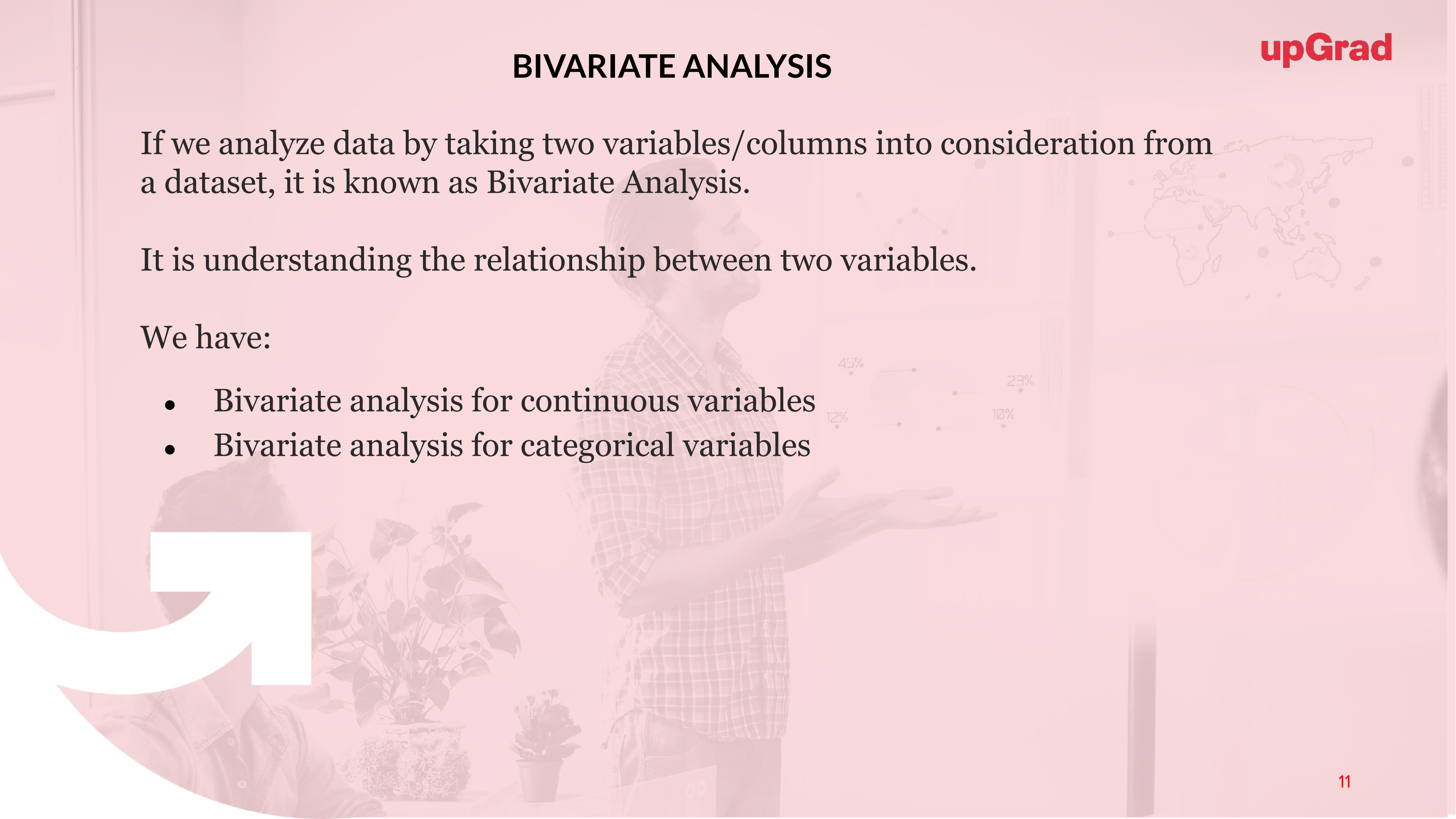
The segmentation process can be divided into four parts:

(i) Take raw data

(ii) Group it by dimensions

(iii) Summarise using a relevant metric such as mean, median, etc.

(iv) Compare the aggregated metric across groups/categories



BIVARIATE ANALYSIS

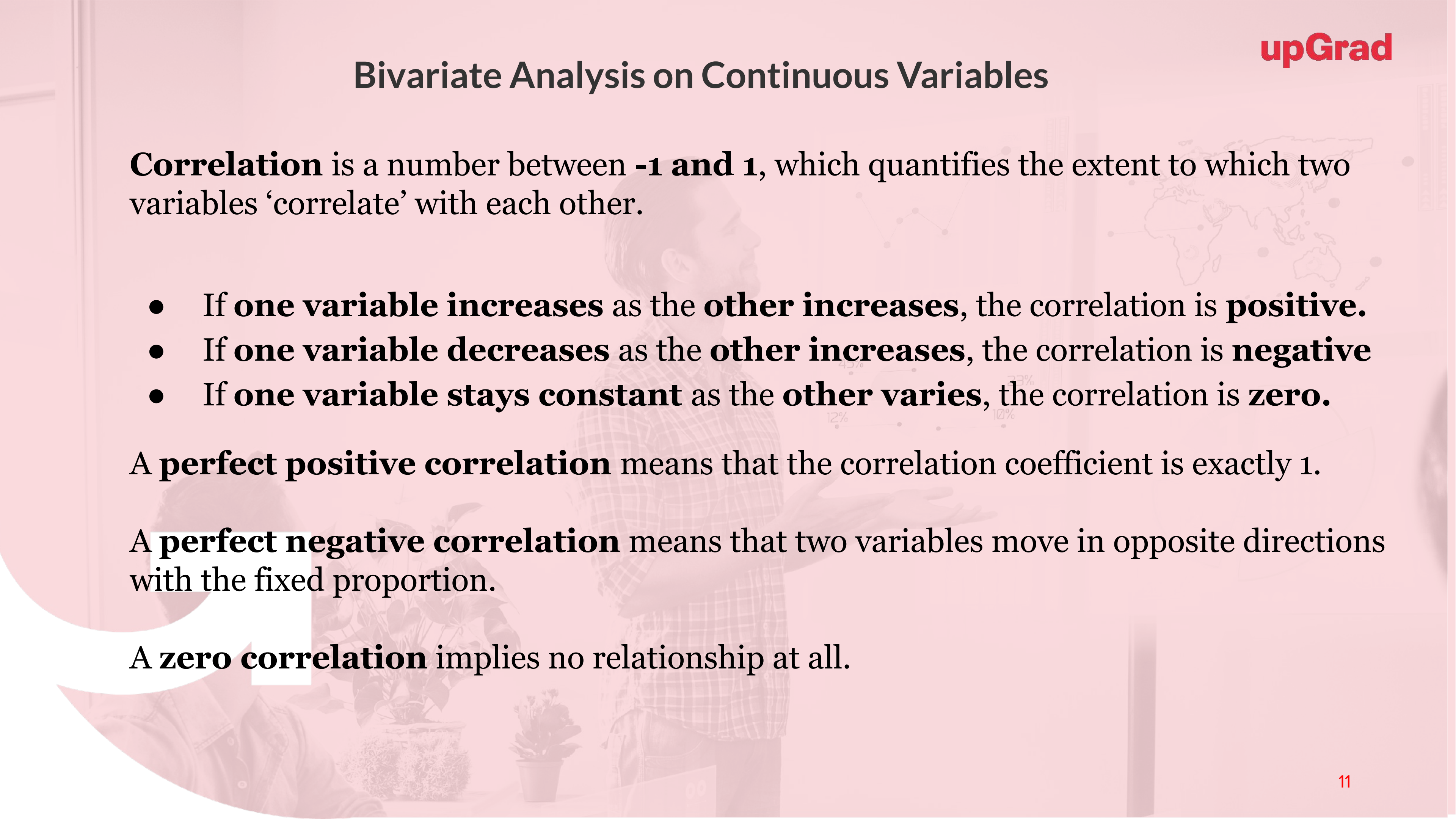
If we analyze data by taking two variables/columns into consideration from a dataset, it is known as Bivariate Analysis.

It is understanding the relationship between two variables.

We have:

- Bivariate analysis for continuous variables
- Bivariate analysis for categorical variables

Bivariate Analysis on Continuous Variables



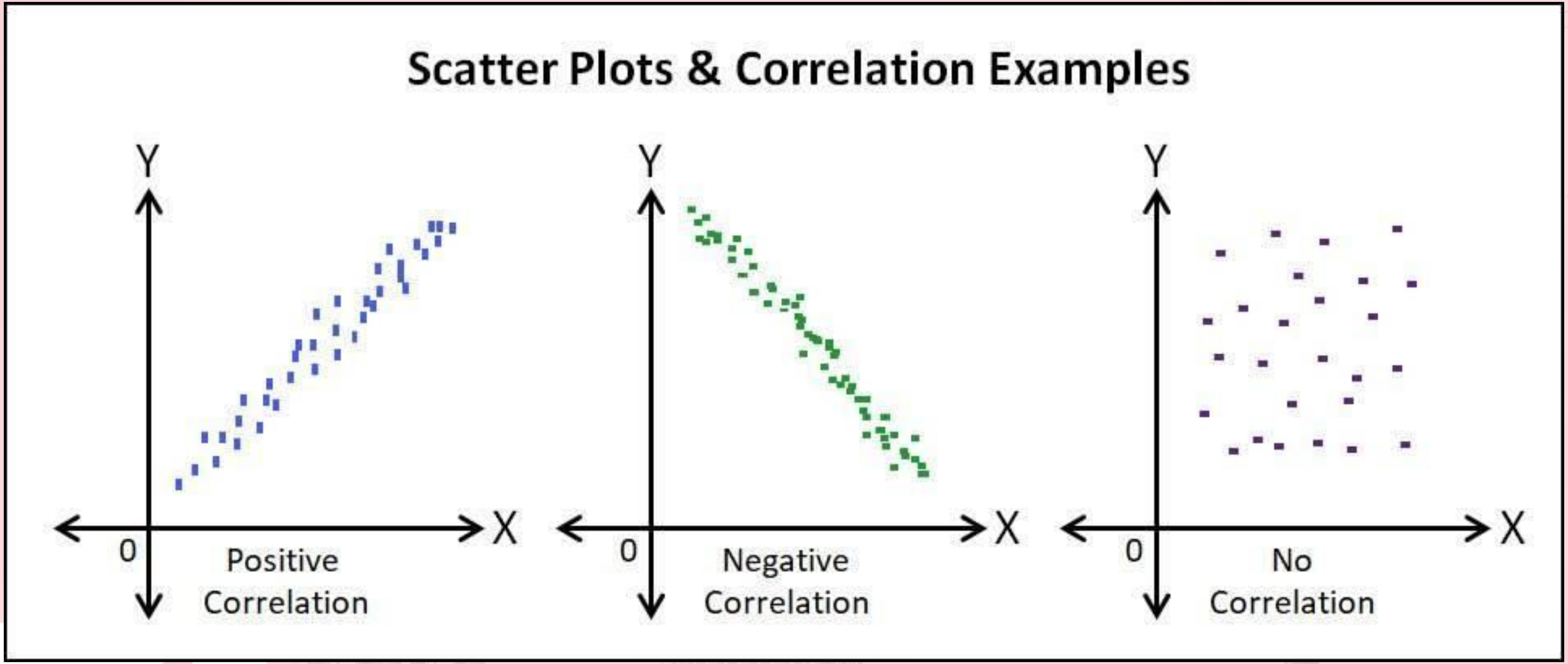
Correlation is a number between **-1 and 1**, which quantifies the extent to which two variables ‘correlate’ with each other.

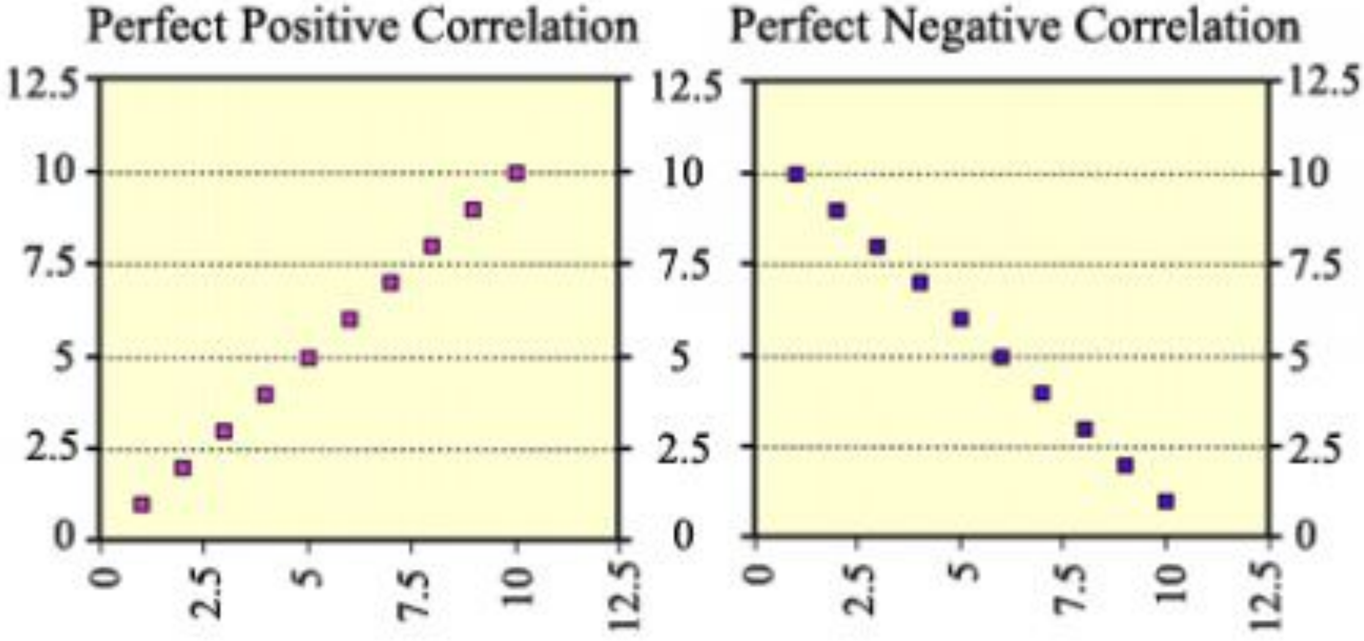
- If **one variable increases** as the **other increases**, the correlation is **positive.**
- If **one variable decreases** as the **other increases**, the correlation is **negative**
- If **one variable stays constant** as the **other varies**, the correlation is **zero.**

A **perfect positive correlation** means that the correlation coefficient is exactly 1.

A **perfect negative correlation** means that two variables move in opposite directions with the fixed proportion.

A **zero correlation** implies no relationship at all.


Scatter Plots & Correlation Examples
Y
X
0
Positive Correlation
Y
X
0
Negative Correlation
Y
X
0
No Correlation


Perfect Positive Correlation
Perfect Negative Correlation
12.5
10
7.5
5
2.5
0
0
2.5
5
7.5
10
12.5

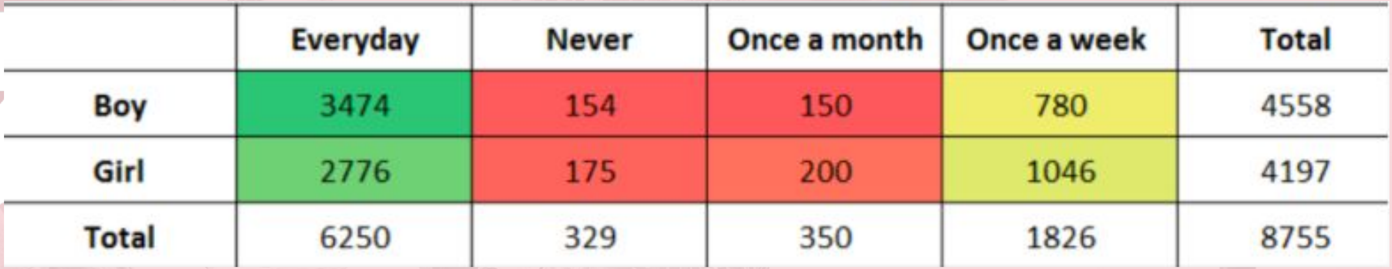
Bivariate Analysis on categorical variables

This type of analysis is used to quantify and understand the relationship between pairs of categorical and continuous variables.

There are two fundamental aspects of analysing categorical variables:

1. **To see the distribution of two categorical variables**

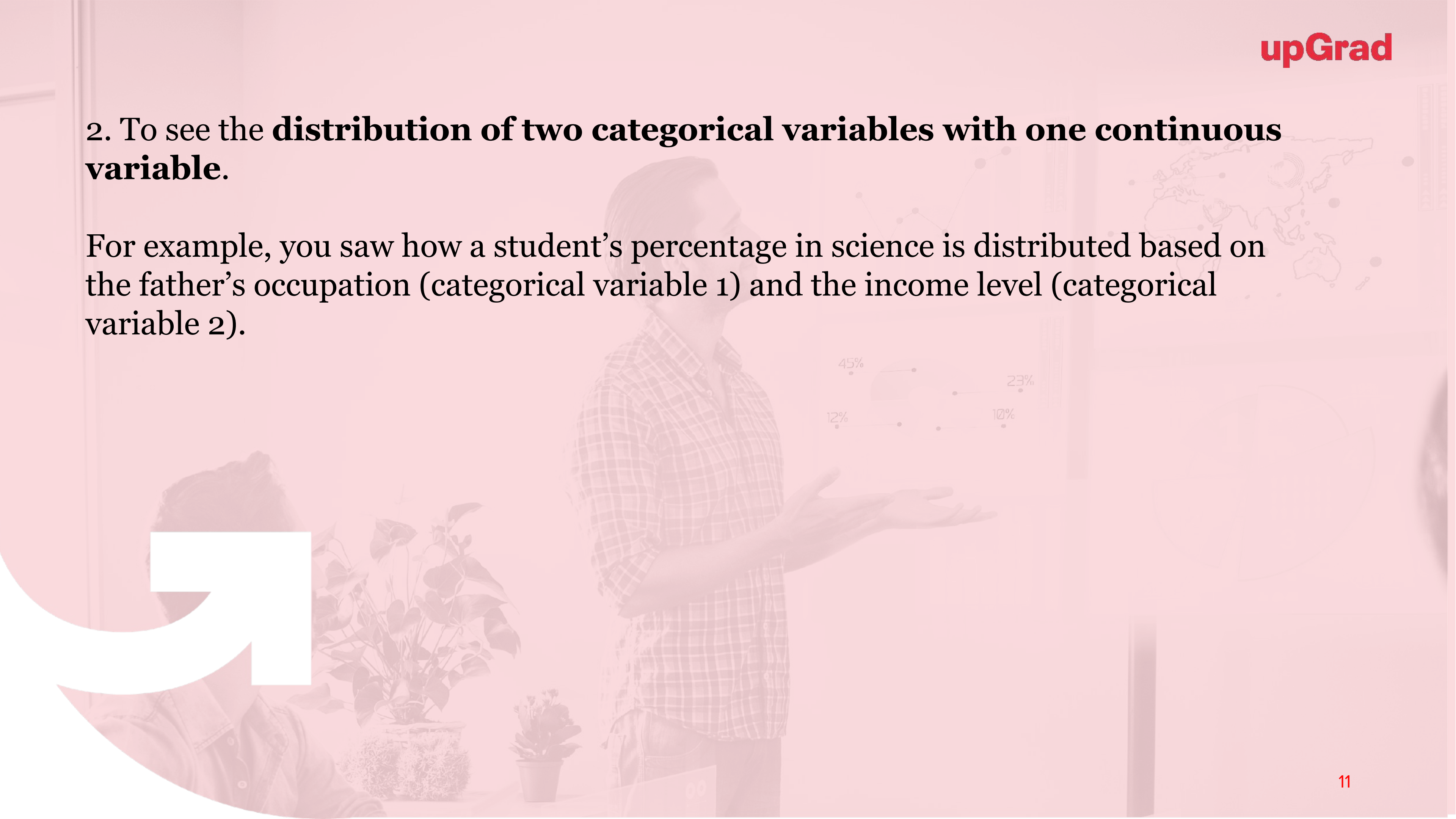
   For example, if you want to compare the number of boys and girls who play games, you can make a ‘cross table’.



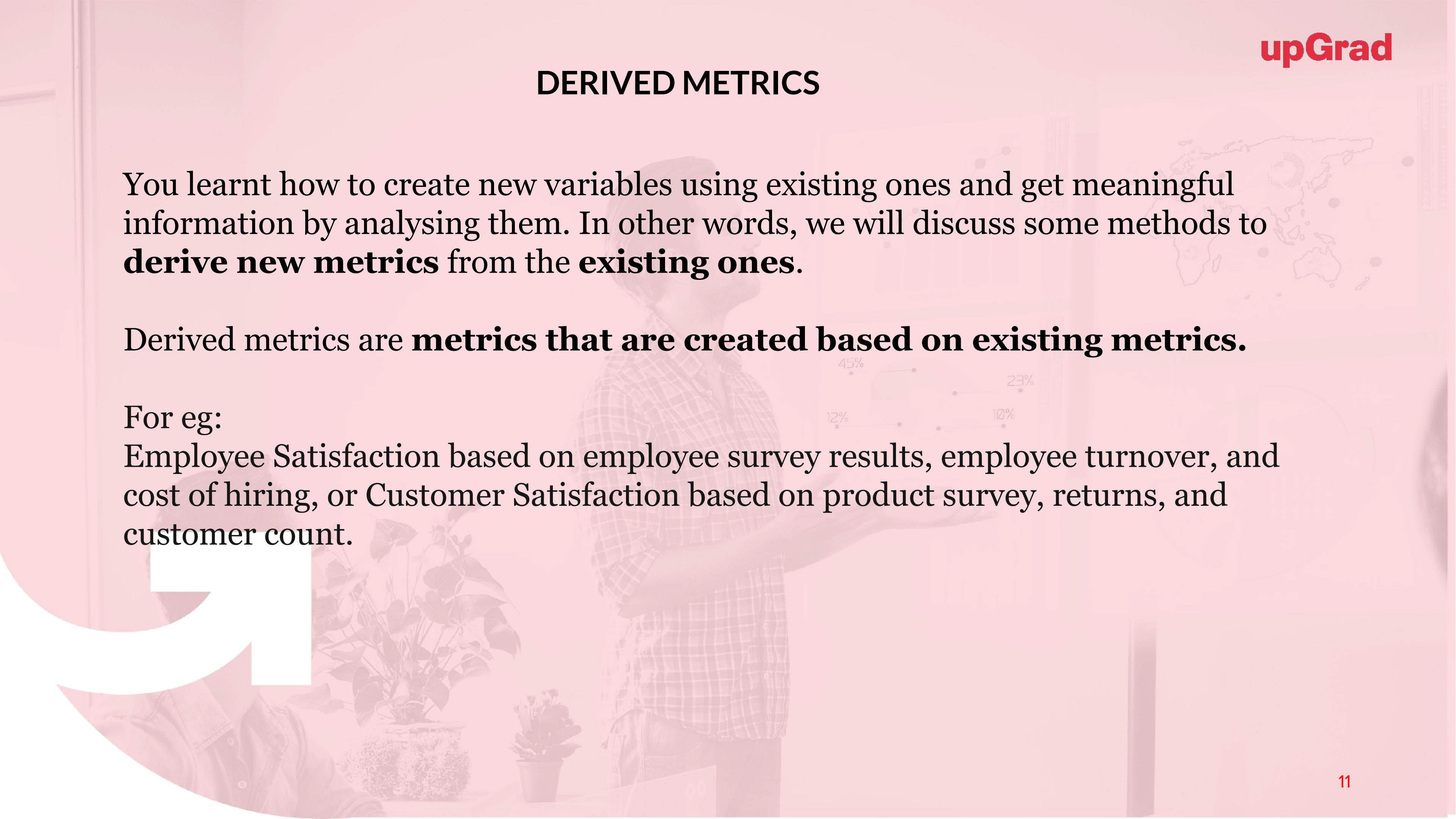
	Everyday	Never	Once a month	Once a week	Total
Boy	3474	154	150	780	4558
Girl	2776	175	200	1046	4197
Total	6250	329	350	1826	8755

2. To see the **distribution of two categorical variables with one continuous variable**.

For example, you saw how a student’s percentage in science is distributed based on the father’s occupation (categorical variable 1) and the income level (categorical variable 2).



DERIVED METRICS



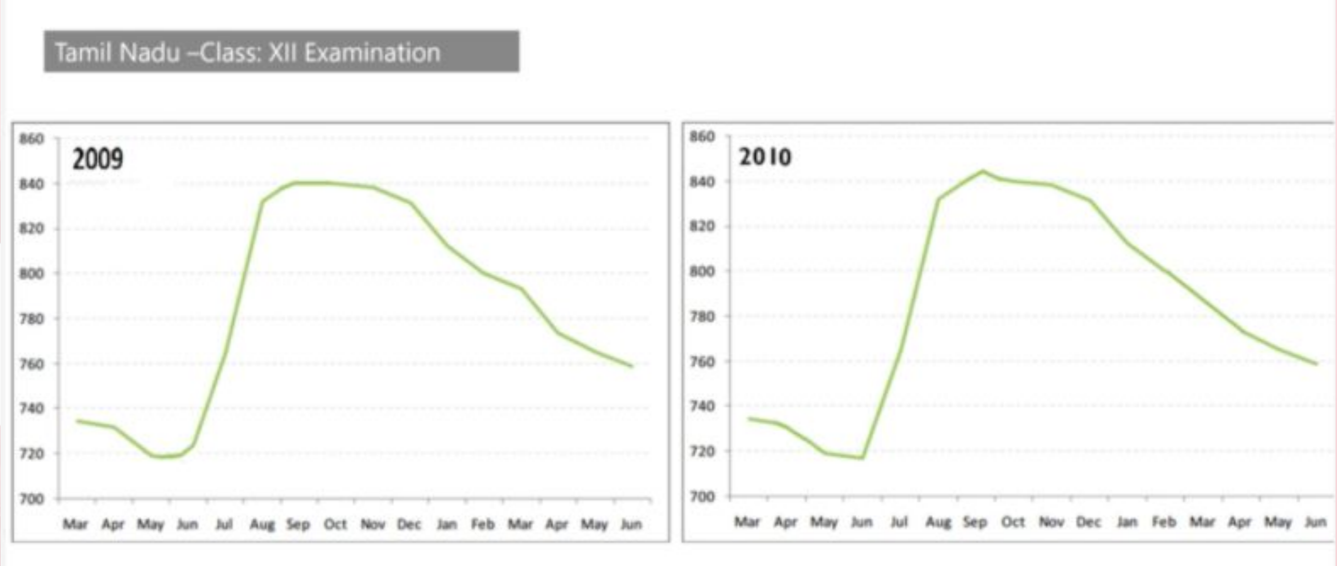
You learnt how to create new variables using existing ones and get meaningful information by analysing them. In other words, we will discuss some methods to **derive new metrics** from the **existing ones**.

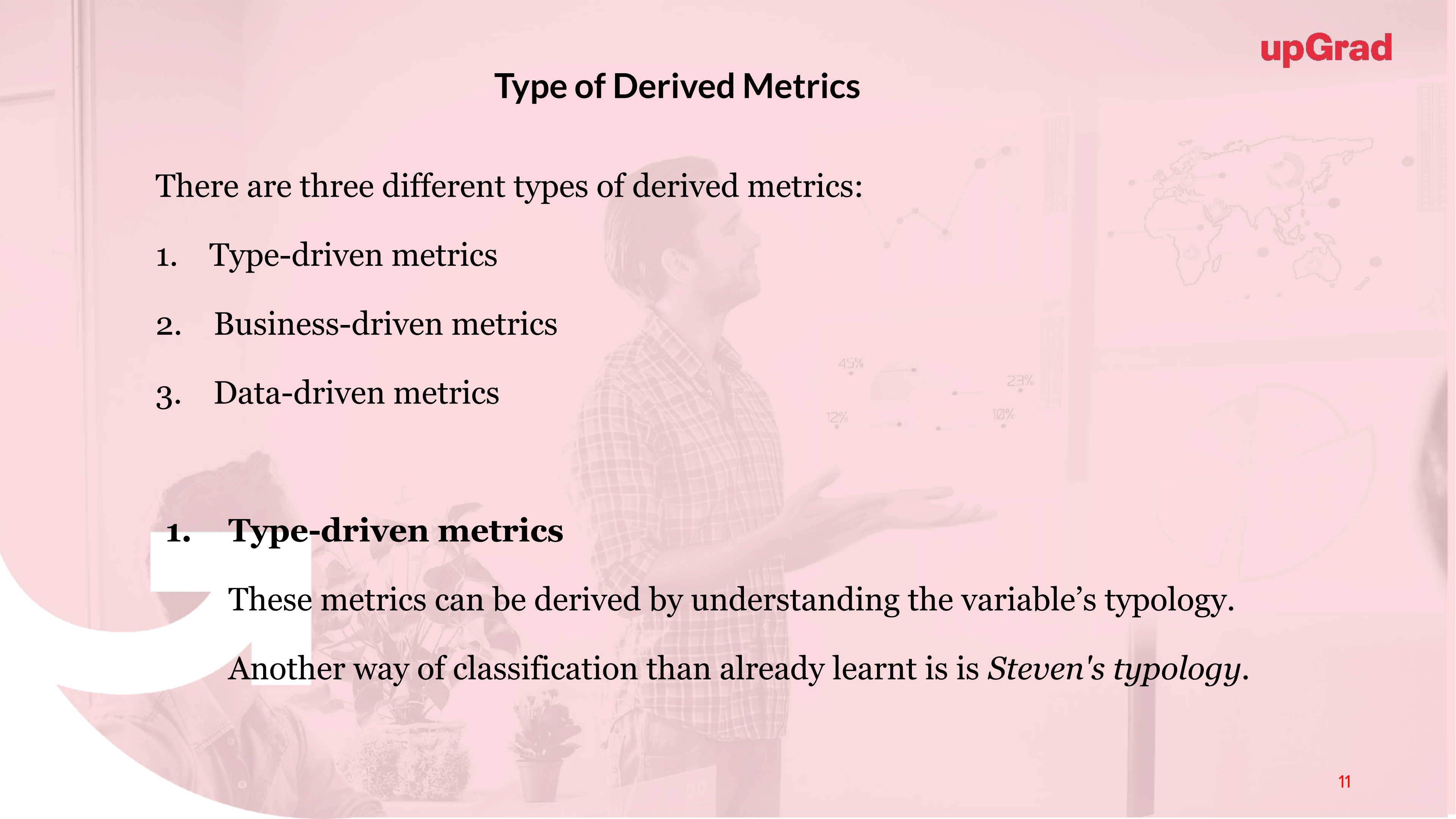
Derived metrics are **metrics that are created based on existing metrics.**

For eg:
Employee Satisfaction based on employee survey results, employee turnover, and cost of hiring, or Customer Satisfaction based on product survey, returns, and customer count.

Example:

Plotting the marks against the ‘month of birth’ (derived variable), it was observed that the children who were born after June would have missed the cutoff by a few days and gotten admission at the age of 5. The ones born after June were intellectually and emotionally more mature than their peers because of their older age, resulting in better performance.





upGrad

Type of Derived Metrics

There are three different types of derived metrics:

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

1. Type-driven metrics

These metrics can be derived by understanding the variable’s typology.

Another way of classification than already learnt is is *Steven's typology*.

Steven's Topology

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.
  Eg: colour (red, blue, green), gender (male, female), department (HR, analytics, sales)

- **Ordinal variables**: Categories follow a certain **order**, but the **mathematical difference between categories is not meaningful**
  Eg: educational 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.
  Eg: temperature in degrees celsius (the difference between 40 and 30 degrees Celsius is meaningful, but 30 degrees x 40 degrees is not).

  Interval variables are **both nominal and ordinal**.

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

  Ratio variables are **nominal, ordinal and interval type**.

2. **Business-driven metrics**

A Business Metric is a quantifiable measure that is used to track and assess the status of a specific business process

Eg: A metric may monitor website traffic for a company

3. **Data-driven metrics**

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

Eg: if you have two variables height and weight we can derive a new metric called BMI.

upGrad

#LifeKoKaroLift

Thank You!