Univariate Analysis

# **Introduction to Univariate Analysis**

Now that you have a clear idea about the process of cleaning a dataset, the next step in Exploratory Data Analysis (EDA) is to learn univariate analysis. It deals with analysing a single column/variable. Let's hear our expert Rahim, as he explains the concept of univariate analysis.

Play Video

1569002

The major takeaway from the video above is that univariate analysis is nothing new to us; you have performed this step on numerical variables while handling missing values and outliers.

## In this session

You will learn about univariate analysis, which broadly is of the following four types:

* Categorical unordered univariate analysis
* Categorical ordered univariate analysis
* Statistics on the numerical variable
* Numerical variable univariate analysis

## Guidelines for In-Module Questions

The in-video and in-content questions for this module are not graded. Note that graded questions are given on a separate page labelled 'Graded Questions' at the end of each session. The graded questions in these sessions will adhere to the following guidelines:

|  | First Attempt Marks | Second Attempt Marks |
| --- | --- | --- |
| Question with 2 Attempts | 10 | 5 |
| Question with 1 Attempt | 10 | 0 |

## People you will hear from in this module

**Subject Matter Expert**

[Mirza Rahim Baig](https://www.linkedin.com/in/rahim-baig)

Analytics Lead, Flipkart

Flipkart is one of the leading e-commerce companies in India. It started with selling books and has now expanded its business to almost every product category, including consumer electronics, fashion and lifestyle products. Rahim is currently the Analytics Lead at Flipkart. He holds a graduate degree from BITS Pilani, a premier educational institute in India.

**Subject Matter Expert**

[S Anand](https://www.linkedin.com/in/sanand0)

CEO, Gramener

Gramener is one of the most prominent data analytics and visualisation companies in India. Anand, currently the CEO, was previously the Chief Data Scientist at Gramener and also has extensive experience in management consulting and equity research.

# **Categorical Unordered Univariate Analysis**

Univariate analysis involves the analysis of a single variable at a time. The concept of univariate analysis is divided into **ordered** and **unordered** category of variables. In this segment, you will learn how to conduct univariate analysis on **categorical unordered variables.**

Unordered data is the type of data that does not have measurable terms such as high-low, more-less, fail-pass, etc. Example:

* The type of loan taken by an individual (home loan, personal loan, auto loan, etc.) does not have any ordered notion. They are just different types of loans.
* Departments of an organisation - Sales, Marketing, HR - are different departments in an organization, with no measurable attribute attached to any term.

Unordered variables also called **Nominal** variables.

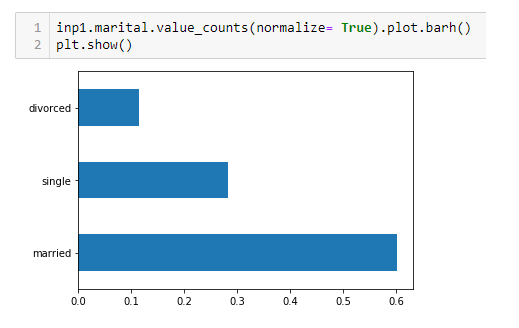
An unordered variable, primarily, is a categorical variable that has no defined order. Let's consider the example of the **job** column in the Bank Marketing dataset. There, '**job**' is divided into many sub-categories like technician, blue-collar, services, management, etc. There is no weight or measure given to any value in the '**job**' column.

In the following video, you will understand how univariate analysis is performed on unordered variables using the Bank Marketing campaign dataset.

Play Video

1569002

You can see from the above video that married people have been contacted the most by the bank. This can be visualised in Python using the following graph.



In bivariate analysis, when variables like marital status, job and education will be plotted with response variables, then you will be in a position to decide which categories in respective columns have the highest chances of a positive reply.

In the next segment, you will learn about categorical ordered univariate analysis.

Question 1/2

Mandatory

#### **Unordered Categorical Variables**

Which of the following options are unordered categorical variables? Choose all correct options.

The blood group of any person, such as A, B, O or AB.

✓ Correct

Feedback:

*Blood group does not have any order in them. they are just typical category of blood group.*

Gear of a vehicle: 1st, 2nd, 3rd or 4th gear.

Gender of a person, like male or female.

✓ Correct

Feedback:

*Gender is an unordered category, as a person can be male or female; there is no hierarchy in genders.*

Various stages of cancer.

Your answer is Correct.

Question 2/2

Mandatory

#### **Unordered Categorical Variables**

Which of the two job categories are the least and the most contacted by the bank respectively?

Student and management professional

Housemaid and blue-collar employee

Student and housemaid

Student and blue-collar employee

✓ Correct

Feedback:

*Write a code to find the information about the job variable, and plot the bar graph of the job variable. Once the code is done, then you will get the correct answer.*

inp1.job.value\_counts(normalize= **True**).plot.barh()

plt.show()

Your answer is Correct.

# **Categorical Ordered Univariate Analysis**

Ordered variables are those variables that follow a natural rank of order. Some examples of categorical ordered variables from the Bank Marketing dataset are:

* Age group: <30, 30-40, 40-50 and so on
* Month: Jan, Feb, Mar, etc.
* Education: primary, secondary and so on

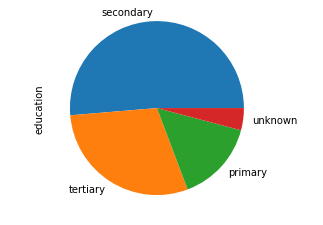
There are other ordered variables in the Bank Marketing data set as well. Let's perform order univariate analysis on that dataset. At the very beginning of this video, you will see the job variable histogram, which you have already covered in the previous segments.

Play Video

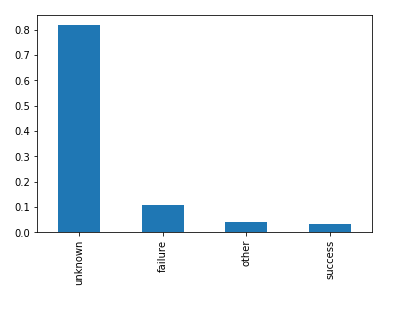
1569002

Let’s summarise the major takeaways from the above video:

* You have seen that *education*, *poutcome* and *response* are the ordered categorical variables.
* The bank has primarily contacted those customers who have completed their secondary education. You can observe that in the pie chart below:



* For the majority of the customers, the previous campaign has not been conducted. Refer to the bar graph below to understand more about the '*poutcome'* variable. As you can see, 'unknown' has the major share within the '*poutcome*' variable.



**Transition of a numerical variable into an ordered categorical variable**

Let’s consider a very interesting example of your school life. Suppose you have a dataset containing the marks of all the students in the 'Science' subject, and you are one of the students in that group. These marks can be considered as categorical if you divide the total marks into different categories like High, Medium, Average, Below Average, Poor. From this analysis, you can determine your ranking in the class and also find out how many students got more marks than you and how far away your score is from the **mean** or the **average** score.

The important thing to note here is that your marks are a numerical variable, which you have then categorised into 'high marks' and 'low marks'. This is an approach that you will need to adopt in the future, and you will learn more about this approach in the next segment on numerical variable analysis.

In the next video, you will understand the basics of statistics and its applications in real-life examples.

Question 1/2

Mandatory

#### **Ordered Categorical Variables**

You have already worked with ordered categorical variables before - there is a certain order or notion of 'high-low', 'before-after', etc., among the categories. For example, the days of the week (Monday comes before Tuesday), grades of students (A is better than B), number of overs bowled by a bowler (3, 4, 9), etc.

Which of the following are other examples of ordered categorical variables? Choose all correct options.

Dates in a year, e.g., Jan 2, Mar 15, etc.

✓ Correct

Feedback:

*Dates are ordered - each day comes before or after a certain day.*

Star rating of a restaurant on Zomato on a scale of 1-5.

✓ Correct

Feedback:

*A rating of 5 is better than 4, 3, 2, 1. Thus, the ratings follow an order.*

Type of education board - ICSE, CBSE, Maharashtra board, etc.

Country names

Your answer is Correct.

Question 2/2

Mandatory

#### **Ordered Categorical Variables**

Suppose you are given a data set containing the salary details of the citizens of a tier-2 city. The salaries are not provided in a numerical format but is given qualitatively, for instance, higher salary, above-average salary, average salary, below average salary and no income. Based on this data, how can you determine the standard of living of the citizens of this city? What steps would you follow in order to dissect this information? Write your answer in the comment box below.

Word Count 10Word Limit 5 - 500

Suggested Answer

First, you should plot the histogram of each category of salary, or you can find the count of each salary group in the data set. Once you have done this, you can predict which category has the highest count, and based on that information, you can predict the standard of living of the people in this city. Suppose you find that the maximum count is for the below-average salary category and the second highest is for the average salary category; thus, you can conclude that the standard of living of the city’s residents is not very high.

# **Statistics on Numerical Features**

You have seen how to conduct univariate analysis on categorical variables. Now, let's look at quantitative or numeric variables.

Numeric variables can be continuous like height, temperature, weight, etc. Numerical variables can also be discrete like the number of items bought by a customer in a store, the number of people in a city, the number of 'heads' you get when flipping three coins.

In this segment, our expert Anand will take you through various statistical metrics such as mean, median, mode and standard deviation.

Let’s now learn how to analyse quantitative variables.

Play Video

1569002

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

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

Let’s now look at some other descriptive statistics such as mode, interquartile distance, standard deviation, etc.

Play Video

1569002

Both standard deviation and interquartile difference are used to represent the spread of the data.

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

You also saw how box plots are used to understand the spread of data.

Question 1/1

Mandatory

#### **Statistics**

The table below shows the marks scored (out of 100) in a course exam:

| Quantiles | Min | 10% | 25% | 50% | 75% | 90% | Max |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Final Marks | 10 | 48 | 55 | 66 | 78 | 87 | 93 |

Which of the following statements is FALSE?

About 1/4 of the class received a score of 55 or less.

✕ Incorrect

Feedback:

*The 25th percentile represents the 1/4 of the total strength of the class, and 25th percentile of the data is 55. Hence, this statement is true.*

About 3/4 of the class received a score of 78 or less.

✕ Incorrect

Feedback:

*The 75th percentile represents the 3/4 of the total strength of the class, and 75th percentile of the data is 78. Hence, this statement is true.*

About 50% of the class received grades between 55 and 78.

✕ Incorrect

Feedback:

*The range from the 25th to the 75th percentile contains 50% of the data points, and the 25th percentile and 75th percentile of this data set are 55 and 78, respectively. Hence, this statement is true.*

The median of class marks is more than 66.

✓ CorrectYou missed this!

Feedback:

*Median is simply the 50th percentile of the data; hence, the median of the data is 66, and not more than 66.*

About 1/30 of the class received a score of 48 or less.

✓ CorrectYou missed this!

Feedback:

*About 10% of the class received a score of 48 or less.*

Your answer is Correct.

# **Graded Questions**

Suppose you are working as an analyst in an e-commerce company and you have been given the two data sets which consist of the details of women's apparels that have been sold during the last year.

* **Attribute DataSet:** This data set contains the different features of women's apparels.
* **Dress Sales:** This data set contains the number of sales for a particular dress ID on a certain date.

Let’s look into the first data set.

You can download the attribute data set from the link below:

Dress Attribute Dataset

Download

In this data set, there is a total of 13 features regarding the dresses.

* **Dress\_ID:** This represents the ID number of a particular dress. This is a unique identification number for different dresses.
* **Style:** This represents the style of a particular dress according to the occasion, like a party, a vintage event, etc.
* **Price:** Each dress ID can fall in a particular price bucket, which can be Low, Medium and High.
* **Rating:** This is the average rating given by the customers for a particular dress ID.
* **Size:** The size column represents the majority of the size bought by the customers for that particular dress ID in the previous sale.
* **Season:** This depicts the season in which particular dress is suitable, for example, summer, winter, etc.
* **Neckline:** This contains the type of neck in the dress, like V-neck, round-neck, etc.
* **SleeveLength:** This represents the type of sleeve of the dresses. Half sleeves, full sleeves, cap sleeves, etc.
* **Material:** This contains information regarding which material the dress has been made of, like cotton, nylon, polyester, silk, etc.
* **FabricType:** This contains information regarding the type of fabric of the dress, for example, chiffon, broadcloth, jersey, etc.
* **Decoration:** This represents the kind of decoration around the dress, like ruffles, bow, embroidery, etc.
* **PatternType:** This represents the type of pattern a particular dress has. Pattern may be solid colours, geometric designs, printed or patchwork.
* **Recommendation:** This is the target variable. 'Recommendation' is based on the features and sales of the dress in the previous year. This is either 1 (means yes) or 0 (means no). This represents whether a particular dress is suitable for sale to the customers or not.

Now, come to the second data set:

You can download the Dress Sales data set from the link below:

Dress Sales Dataset

Download

This particular dataset represents the number of sales for a particular dress ID on a certain date, where columns represent the dates on which a particular dress ID was sold.

Now, based on the above two datasets, you are expected to perform the EDA and draw useful insights. Based on EDA, answer the graded questions of this module.

You have gone through univariate analysis with an example of the Bank Marketing dataset in the previous segments. Now, let’s answer the following questions based on the steps that you have learnt in this session.

You are provided with a blank Jupyter notebook with all the comments to perform the operations.

Question 1/5

Mandatory

#### **Categorical Unordered Univariate Analysis (Multiple Correct Question)**

In the ‘Attribute Dataset’, there is a column named ‘Style’, which contains the different style categories of women's apparel. Certain categories whose total sales is less than 50,000 across all the seasons are grouped under a single category named ‘Others’.

Which of the following categories under the ‘Style’ column can be grouped in the ‘Others’ category? Perform the grouping operation in the notebook for further analysis.

(note: this can have multiple correct answers, select all which fulfil the requirement)

Flare, Fashion

✓ CorrectYou missed this!

Feedback:

*Use the groupby and sum commands to find the categories in the 'Style' column that have sales values less than 50,000 across all the seasons. You can refer to the following code to perform such operations:*

*Convert the Style categories:*

total\_collection=inp0.groupby(['Style'])['Spring','Winter','Autumn', 'Summer'].sum()

total\_collection.sum(axis = **1**)

*Find the categories that have sales values less than 50,000 across all the seasons.*

total\_collection[total\_collection.sum(axis= **1**)<**50000**]

*Group the Style categories that have sales less than 50,000 across all the seasons under ‘Others’ .*

inp0.replace(total\_collection[total\_collection.sum(axis=**1**)<**50000**].index, 'Others', inplace=**True**)

Novelty, Bohemian

✕ Incorrect

Feedback:

*Use the groupby and sum commands to find the categories in the 'Style' column that have sales values less than 50,000 across all the seasons. Bohemian has a sales value more than 50,000 across all the seasons.*

OL, Fashion, Work

Novelty, Fashion, Flare

✓ Correct

Feedback:

*Use the groupby and sum commands to find the categories in 'Style' column that have sales values less than 50,000 across all the seasons.*

*Convert the Style categories:*

total\_collection=inp0.groupby(['Style'])['Spring','Winter','Autumn', 'Summer'].sum()

total\_collection.sum(axis = **1**)

*Find the categories that have sales values less than 50,000 across all the seasons.*

total\_collection[total\_collection.sum(axis = **1**)<**50000**]

*Group the Style categories that have sales less than 50,000 across all the seasons under ‘Others’ .*

inp0.replace(total\_collection[total\_collection.sum(axis=**1**)<**50000**].index, 'Others', inplace=**True**)

Your answer is Wrong.

Question 2/5

Mandatory

#### **Categorical Unordered Univariate Analysis**

What is the respective percentage of the 'Cute' and 'Others' categories in the 'Style' column in the 'Attribute Dataset'?

46%, 5%

✕ Incorrect

Feedback:

*Once you have grouped Novelty, Fashion, OL and Flare under 'Others', you need to calculate the percentage of each category in the 'Style' column using the code in Python.*

9%, 2.1%

✓ Correct

Feedback:

*The Python code is as follows:*

print(inp0.Style.value\_counts(normalize=**True**))

inp0.Style.value\_counts(normalize=**True**).plot.barh()

plt.show()

2.1%, 5%

13.8%, 9%

Your answer is Wrong.

Question 3/5

Mandatory

#### **Categorical Ordered Univariate Analysis**

Which of the following is an unordered variable in the 'Attribute Dataset'?

Style

✓ Correct

Feedback:

*The Style category contains the type of styles available, and it is an unordered type.*

Price

Season

Size

Your answer is Correct.

Question 4/5

Mandatory

#### **Numerical Variable Univariate Analysis**

What is the approximate difference between the maximum value and the 75th percentile in the 'Autumn' column?

Approximately 54,000

Approximately 55,000

Approximately 52,000

✓ Correct

Feedback:

*Correct. This is the difference between the maximum and the 75th percentile.*

*Plot a box plot and describe these variables using the following code:*

inp0.Autumn.describe()

sns.boxplot(inp0.Autumn)

plt.show()

Approximately 50,000

Your answer is Correct.

Question 5/5

Mandatory

#### **Numerical Variable Univariate Analysis**

Which of the following seasons shows the highest difference between the maximum value and the 99th quantile of sales?

Winter

Summer

Spring

Autumn

✓ Correct

Feedback:

*Correct. The difference between the maximum value and the 99th percentile of sales in Autumn is 31,398, which is the highest on the list. You can use the following code to find the maximum value and the 99th percentile of sales in Autumn:*

inp0.Autumn.quantile([**0.99**, **1**])

Your answer is Correct.

# **Summary**

Univariate analysis is the analysis of a single variable at a time. This particular variable can be ordered or unordered, or it may be a numerical variable. So, based on the types of variables, the whole understanding of univariate analysis is divided into the following parts:

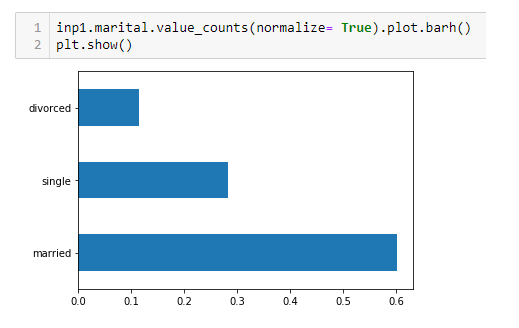
* **Categorical unordered univariate analysis**: Unordered variables are those variables that do not contain any notion of ordering, for example, increasing or decreasing order. These are just various types of any category. The examples can be job types, marital status, blood groups, etc.

* **Categorical ordered univariate analysis**: Ordered variables are those that have some kind of ordering in them, like high-low, fail-success, yes or no. Examples can be education level, salary group like high or low, gradings in any exam, etc.

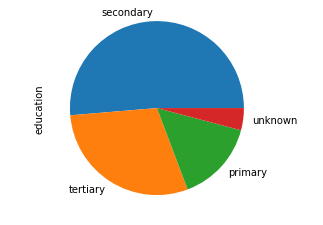
* **Numerical variable univariate analysis**: Numerical variables can be classified into continuous and discrete type. To analyse numerical variables, you need to have an understanding of statistic metrics like mean, median, mode, quantiles, and box plots, etc. It is important to understand that numerical variable univariate analysis is nothing but what we have done earlier, i.e., the treatment of missing values and handling outliers. The crux of univariate analysis lies in the single variable analysis, which is covered in the process of cleaning the dataset.
* **The transition of a numerical variable into a categorical variable:** This is an important aspect that you need to think about before performing univariate analysis. Sometimes, it is essential to just convert numerical variables into categorical ones, through a process which is called 'binning'.

Let's summarise univariate analysis on the Bank Marketing Campaign dataset.

* You have seen that there is a variable called "**marital**" in the Bank Marketing dataset. This is **categorical unordered variable**. You have seen that the bank has contacted mostly married people, as can be seen in the image below.



* There is a variable called "**education**" in the Bank Marketing dataset. This is a **categorical ordered variable** because there is ordering of education levels, like primary, secondary and tertiary education. You have seen that the bank has mostly contacted people who have completed secondary education, as can be seen in the image below.



* You have already performed univariate analysis on numerical variables in the process of missing values treatment and handling outliers. You have seen that there are no outliers in the "**age**" variable, as the values of age like 80 or 90 are also genuine values. There are higher values in 'balance' and 'salary' variables, which can be treated as outliers. Hence, it can be avoided while performing the analysis.

Hence, univariate analysis is nothing but an analysis of one particular variable at a time. It is important to look at each and every variable and perform analysis on it.