Shaina Mehta
7CSE 4Y
A2305219268

Assignment 2.1

Q1) Calculate the following Mean, Median, Mode, and Standard Deviation for Column A.

A	B	C
77	02	174
86	10	145
78	03	155
79	06	188
70	07	167

Ans) Sort the data in ascending order w.r.t. column A

A	B	C
70	07	167
77	02	174
78	03	155
79	06	188
86	10	145

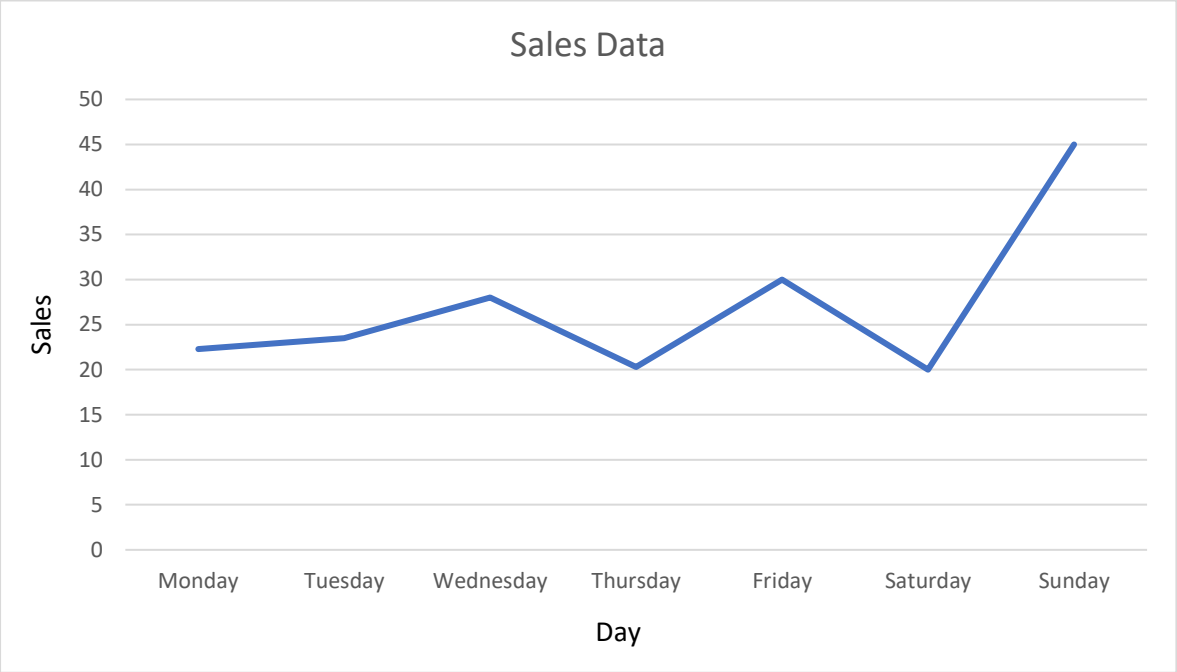
(a) Mean = (77+86+78+79+70)/5 = 78
(b) Median = (n+1)/2th value in column A = 3rd Value in col A = 78
(c) Mode = (3*Mean) – (2*Median) = (3*78) – (2*78) = 78
(d) Standard Deviation = $\sqrt{\frac{\sum_{i=1}^{n}(x_i - Mean)^2}{n-1}}$ = 5.7008

Q2) Plot the line graph for the given data.

Day	Mon	Tue	Wed	Thurs	Fri	Sat	Sun
Sales	22.3	25.5	28	20.3	30	20	45

Ans) The line graph for the given data is given below as:

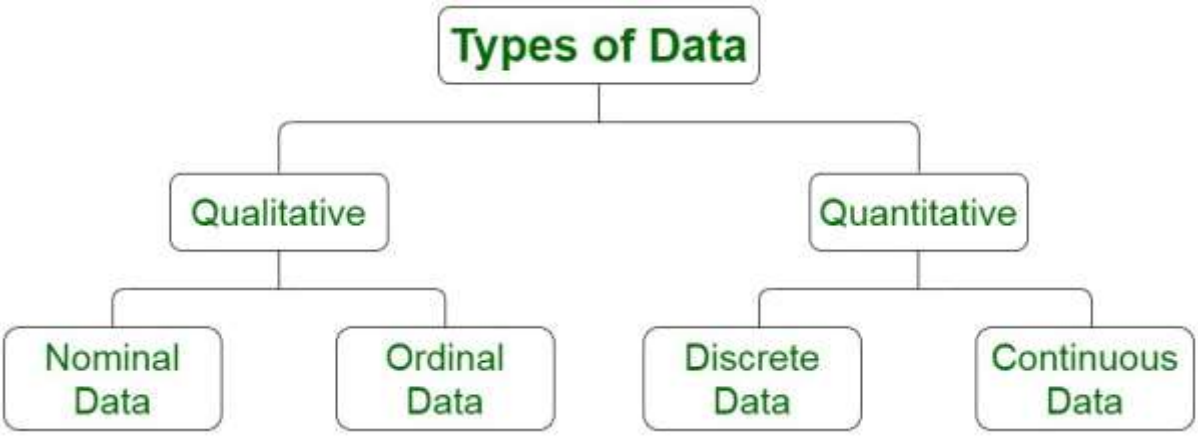
Shaina Mehta
7CSE 4Y
A2305219268



Q3) What do you mean by Data? Explain various types of data in detail.

Ans) Data can be defined as a systematic record of a particular quantity. It is the different values of that quantity represented together in a set. It is a collection of facts and figures to be used for a specific purpose such as a survey or analysis. When arranged in an organized form, can be called information.

Types of Data: There are the following types of known forms of data:



1. Qualitative Data

Qualitative data is used to represent some characteristics or attributes of the data. The facts and figures depicted by the qualitative data cannot be computed. These properties reflect observable attributes. These are non-numerical in nature. The qualitative data characteristics are exploratory on a larger end than being conclusive in nature. For instance, data on attributes such as honesty, loyalty, wisdom, and creativity for a set of persons defined can be considered as qualitative data.

Examples:

- Attitudes of people to a political system.
- Music and art
- Intelligence
- Beauty of a person

(a) Nominal Data

Nominal data is a sub-category belonging to one of the types of qualitative information. Also known as the nominal scale, it is used to label the variables without providing the numerical value for them. Nominal data attributes can't either be ordered or measured. The nominal data can be both qualitative and quantitative in nature. For instance, some of the nominal data attributes are letters, symbols or gender, etc.

The examination of the nominal data is based on the usage of the grouping method. This method is based on the principle of the grouping of data into different categories. This is followed by the calculation of the frequency or the percentage of the data. The visualization of this data is done using the pie charts.

Examples:

- Gender (Women, Men)
- Eye color (Blue, Green, Brown)
- Hair color (Blonde, Brown, Brunette, Red, etc.)
- Marital status (Married, Single)
- Religion (Muslim, Hindu, Christian)

(b) Ordinal Data

Ordinal data/variable is the specific type of data that follows a natural order. The difference between the data values is not determined in the case of nominal data. For instance, ordinal data variable is mostly found in surveys, economics, questionnaires, and finance operations.

The examination of the nominal data is based on the usage of visualization tools. The visualization of this data is done using the bar chart. The ordinal data can be expressed in the form of tables which have each row corresponding to the distinct category.

Examples:

- Feedback is recorded in the form of ratings from 1-10.
- Education level: elementary school, high school, college.
- Economic status: low, medium, and high.
- Letter grades: A, B, C, and etc.
- Customer level of satisfaction: very satisfied, satisfied, neutral, dissatisfied, very dissatisfied.

2. Quantitative Data

Quantitative data can be measured and is not just observable. The measurement of data is numerically recorded and represented. Calculations and interpretations can then be performed on the obtained results. Numerical data is indicated by quantitative data. For instance, data can be recorded about how many users found a product satisfactory in terms of the collected rating, and therefore, an overall product review can be generated.

Examples:

- Daily temperature
- Price
- Weights
- Income

Shaina Mehta
7CSE 4Y
A2305219268

(a) Discrete Data

Discrete data refers to the data values which can only attain certain specific values. Discrete data can't attain a range of values. Discrete data can be represented using bar charts. For instance, ratings of a product made by the users can only be in discrete numbers.

Examples:

- The number of students in a class,
- The number of chips in a bag,
- The number of stars in the sky

(b) Continuous Data

Continuous Data can contain values between a certain range that is within the highest and lowest values. The corresponding difference between the highest and lowest value of these intervals can be termed as the range of data. Continuous data can be tabulated in what is called a frequency distribution. The frequency distribution table can be computed for the range type of data. It can also be depicted using histograms. For example, the heights of the students in the class can be largely varying in nature, therefore, they can be divided into ranges to summarise the data.

Examples:

- Height and weight of a student,
- Daily temperature recordings of a place
- Wind speed measurement

Q4) How will you deal with the data which is incomplete and noisy?

Ans) The methods for handling incomplete and noisy data is given below as:

(a) Handling Missing Values/ Incomplete Data:

1. Standard values like "Not Available" or "NA" can be used to replace the missing values.
2. Missing values can also be filled manually but it is not recommended when that dataset is big.
3. The attribute's mean value can be used to replace the missing value when the data is normally distributed wherein in the case of non-normal distribution median value of the attribute can be used.
4. While using regression or decision tree algorithms the missing value can be replaced by the most probable value.

(b) Handling Noisy Data:

1. **Binning:** This method is to smooth or handle noisy data. First, the data is sorted then and then the sorted values are separated and stored in the form of bins. There are three methods for smoothing data in the bin. Smoothing by bin mean method: In this method, the values in the bin are replaced by the mean value of the bin; Smoothing by bin median: In this method, the values in the bin are replaced by the median value; Smoothing by bin boundary: In this method, the using minimum and maximum values of the bin values are taken and the values are replaced by the closest boundary value.

2. **Regression:** This is used to smooth the data and will help to handle data when unnecessary data is present. For the analysis, purpose regression helps to decide the variable which is suitable for our analysis.
3. **Clustering:** This is used for finding the outliers and also in grouping the data. Clustering is generally used in unsupervised learning.

Q5) What does Standardisation mean? (Explain with respect to z-score).

Ans) Standardization or Z-Score Normalization is the transformation of features by subtracting from the mean and dividing by standard deviation. This is often called a Z-score.

X_new = (X - mean)/Std

Standardization can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Geometrically speaking, it translates the data to the mean vector of original data to the origin and squishes or expands the points if std is 1 respectively. We can see that we are just changing mean and standard deviation to a standard normal distribution which is still normal thus the shape of the distribution is not affected.

Standardization does not get affected by outliers because there is no predefined range of transformed features.

Q6) How will you visualize the data?

Ans) Data visualization can be done in the following ways:

1. Indicator: If you need to display one or two numeric values such as a number, gauge or ticker, use the Indicators visualization. You can add additional titles and a color-coded indicator icon, such as a green up arrow or a red down arrow to represent the value in the clearest way.

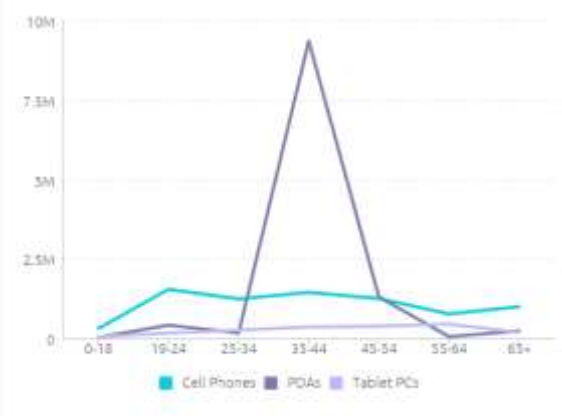
Gauge Indicator

Numeric Indicator

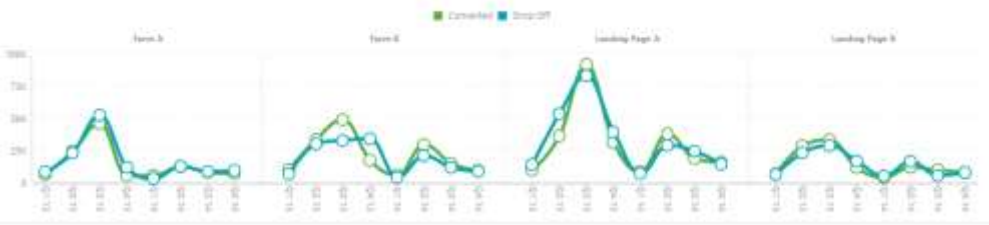
2. Line chart: The line chart is a popular chart because it works well for many business cases, including to:

- Compare data over time to view trends (Example: analyze sales revenue for the past year)

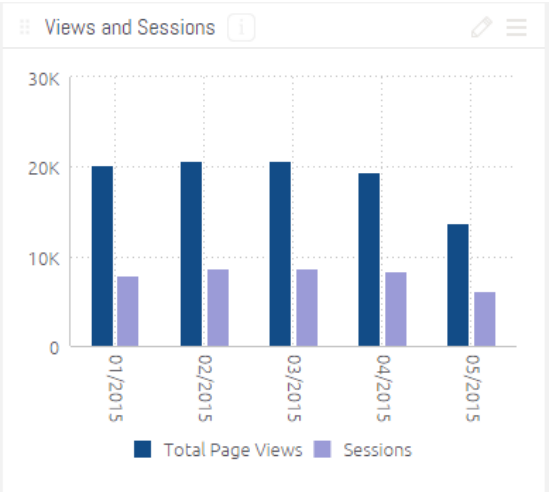
Shaina Mehta
7CSE 4Y
A2305219268



- Compare changes over the same period of time for more than one group or category(Example: analyze expenditures of different business units for the past year).Here you just need to simply add a “break by” category.

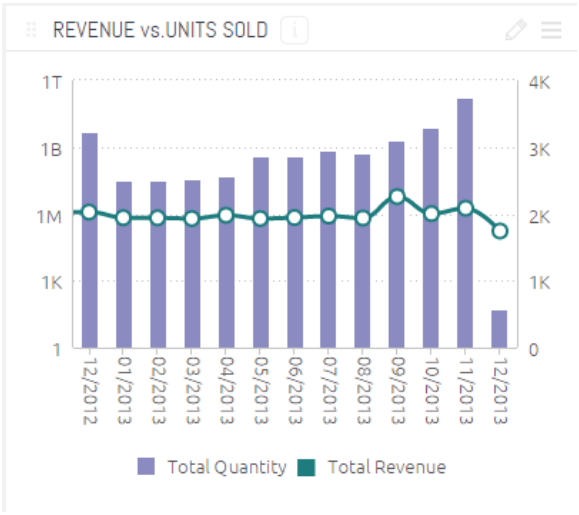


3. Column chart: The column chart is best used for comparing items and comparing data over time. The column chart can include multiple values on both the X and Y axis, as well as a breakdown by categories displayed on the Y axis.

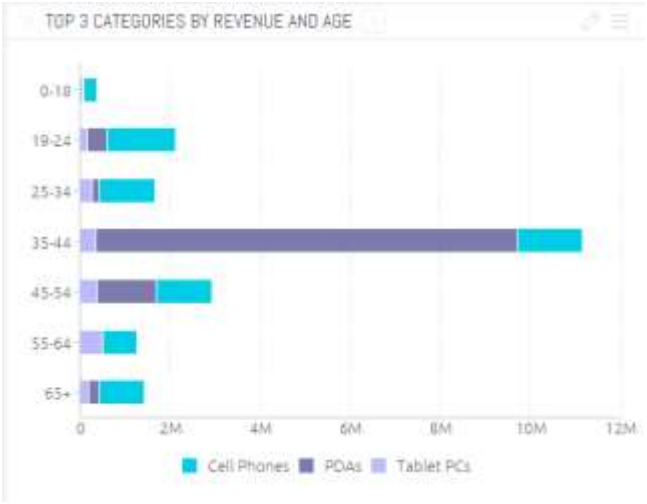


To highlight peaks and trends, you can also combine the column chart with a line chart.

Shaina Mehta
7CSE 4Y
A2305219268

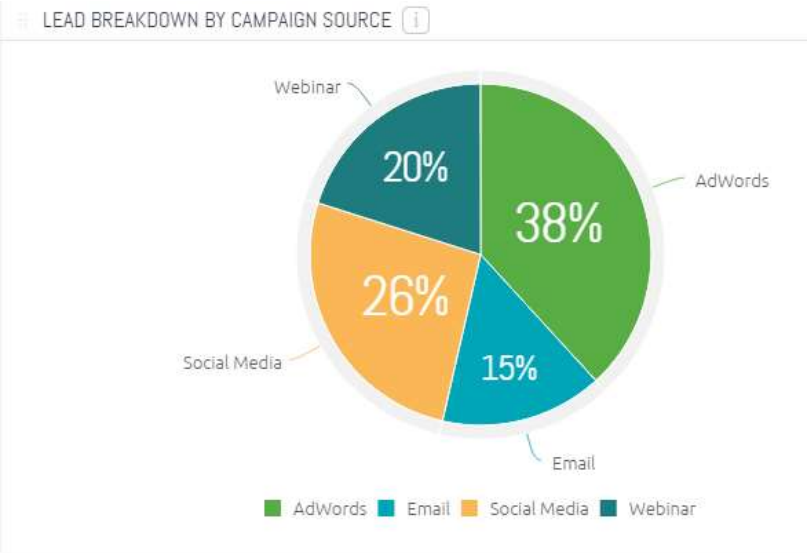


4. Bar chart: Use the bar chart to compare many items. The bar chart typically presents categories or items displayed along the Y axis, with their values displayed on the X axis. You can also break up the values by another category or group.

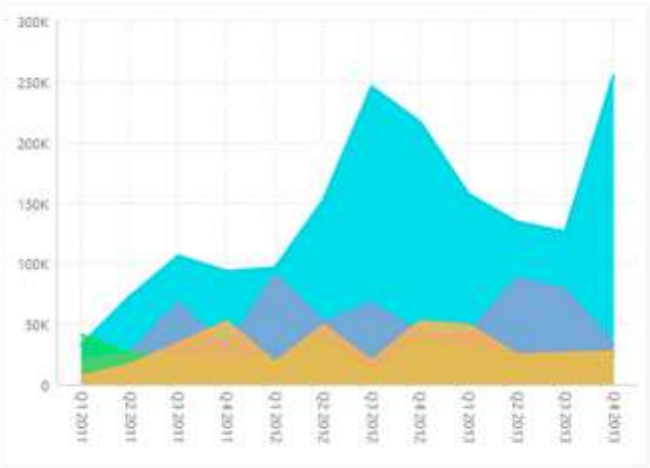


Shaina Mehta
7CSE 4Y
A2305219268

5. Pie chart: The pie chart is best when you are aiming to display proportional data, and/or percentages. Since the pie chart represents the size relationship between the parts and the entire entity, the parts need to sum to a meaningful whole. Pie charts should only display around six categories or fewer.



6. Area chart: Though an area chart may seem similar to a line chart, the areas under each line are filled in (colored), and it is therefore possible to display them as stacked for better comparison. Use an area chart if you are looking to display absolute or relative (stacked) values over a time period.

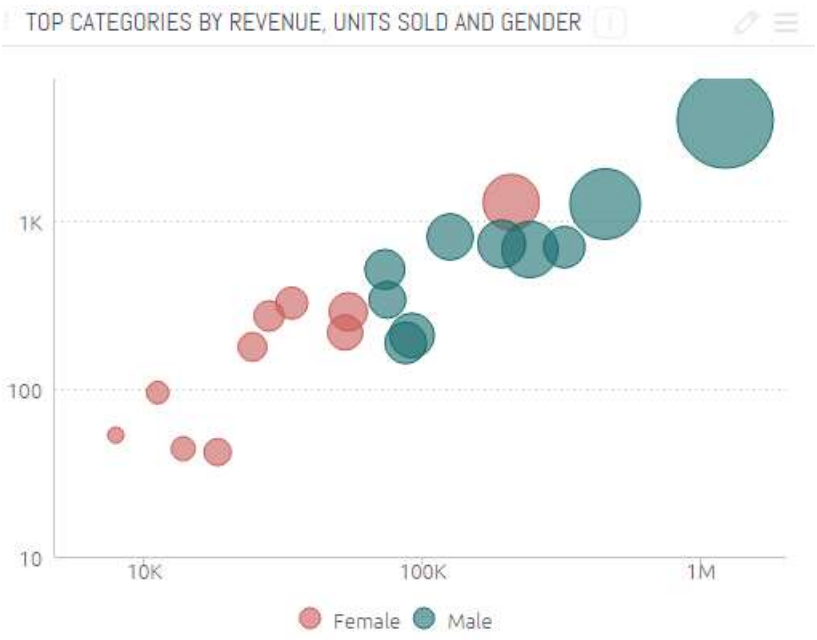


Shaina Mehta
7CSE 4Y
A2305219268

7. Pivot table: Pivot tables are one of the most simple and useful ways to visualize data. You can quickly summarize and analyze large amounts of data and use additional features such as color formatting and data bars to enhance the visual aspects.

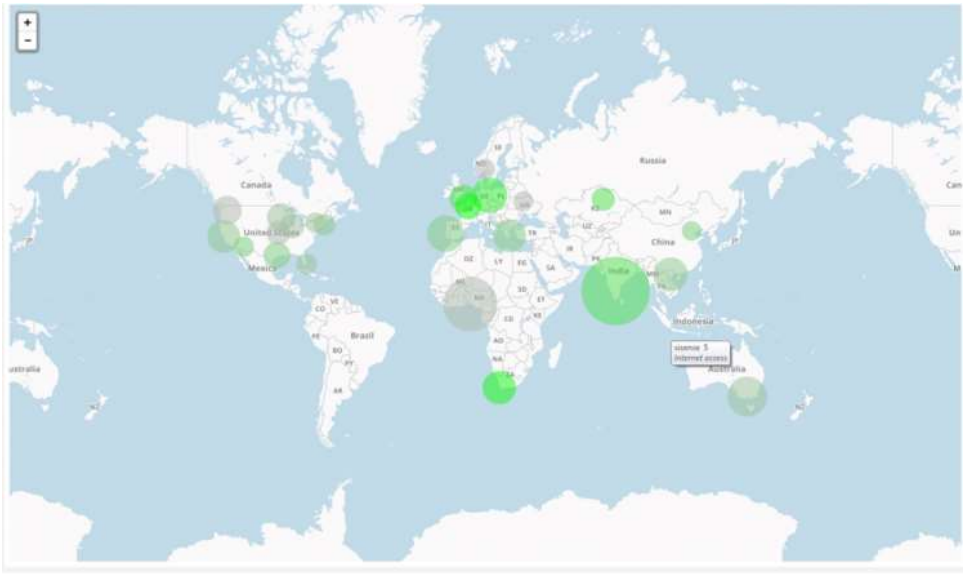
Age Range	Condition	Total Quantity
0-18	New	1,209
	Refurbished	292
	Unspecified	360
	Used	1,913
19-24	New	2,544
	Refurbished	592
	Unspecified	569
	Used	3,790
25-34	New	5,950
	Refurbished	1,401
	Unspecified	1,407
	Used	8,831

8. Scatter chart: The best scenario to use the popular scatter chart is when you are trying to display the distribution and relationship of two variables. The circles on the chart represent the categories being compared (circle color), and the size or numeric data (indicated by the circle size). A good example is if you are trying to compare revenue and units sold by gender.



Shaina Mehta
7CSE 4Y
A2305219268

9. Scatter map / Area map: A scatter map helps viewers visualize geographical data across a region as data points on a map. You can visualize up to two sets of numeric data using circle color and size to represent the value of your data.



10. Treemap: The treemap is a multi-dimensional widget that displays hierarchical data in the form of nested rectangles. You can use this type of chart in different scenarios, for example, instead of a column chart when you want to compare many categories and sub-categories.