Data Visualization

Presentation of the data in a graphical or pictorial form is known as Data visualization, which not only makes the data easy to understand but also shows the hidden patterns of the data. Especially it plays a vital role when the data is great in number i.e., Time-Series. It is the representation of raw data visually for useful information which can easily be translated by the reader. Data Visualization makes the facts more convenient in terms of explanation. Summary Statistics of the data are useful but mostly there is hidden detail within the data which cannot be identified by the summary so relying on just the summary is very dangerous as it can lead to incorrect solutions while on the other hand visualization of the data provide as much information which leads to the veracious side of the data. As per the academic view, the visualization is considered as a relationship between the data and the elements used in graphics e.g., Plots, lines, and charts. The roots of the data visualization are coming from statistics therefore mostly it is considered as descriptive statistics.

As living in the era of Big Data, data visualization is an essential tool as without there is no meaning of billions of trillions of rows of data produced by different gadgets every minute. As suitable and formed visualization, precise the story of Big data by removing the noise and not use data to get the informatic insights only which plays important role in big decisions. However, just visualizing the data is not enough it is meaningful when done right and it is said right when it fulfills its purpose. Generally, we divide the data visualization int to two types: exploration and explanation besides this visualization also has many other types including 2D Area, Temporal, and the like.

Because of all these data visualizations is a very active field of research. Although a lot many ways and visualization are already been discovered for different types of data but still there is a large space for more research. Principles rules and the enhancement in the current visualization are required. With the increasing speed of Big data, we also need to find other ways of visualizations to gather more efficient and useful information from it. Some common examples of the visualization are: Bar chart, Pie Chart, Scatterplot, Radar chart and such visualization of which some of them are discussed below

Radar Chart

Fig 1: is a Radar chart also known as spider or web chart. It is a multi-dimensional type of chart used to visualize one or more variables in which each variable is independent of its axis while the center of the figure is working as the joint point of all the variables. Precisely Radar chart is the combination of multiple spokes generally named radii to represent each variable. The length of the spoke tells the magnitude of that specific variable against the remaining variables. Also, it has some limitations too of which the most one is that sometimes it will be very difficult to compare the magnitudes of different variables because of the same length of the spokes. But it will be handled using some preprocessing techniques of the data or with the modeling of the data in a better way. As in this, we have the data of visitors to the different stores for the year 2019. As we have 40 different stores and have to visualize them in such a way that which store has more visitors and which has less relatively. So, we consider the stores as the variables and the radii as the number of customers (for the year 2019)

Radar Chart of Average of Customer of Stores

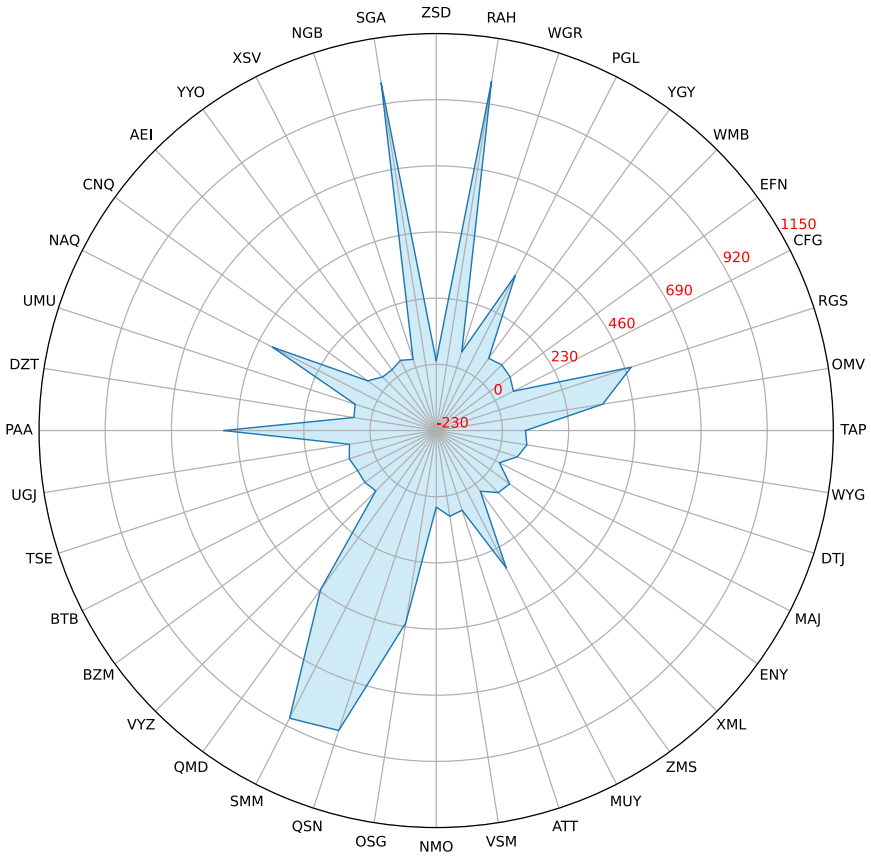

Figure 1: A Radar Chart Which Shows the Average Customers in each Store.

Fig 1: is the Radar Chart of the number of customers who visited each store for the year 2019 as the sum of the total number of customers of each store is very high so we take the mean of the sum which give us some relative values. In the chart, the names of the stores are used as the variables and the mean of the sum of the customers used as the spokes. The magnitude of the spokes shows which store has more customers and which has less. As from the chart, SGA and RAH have the largest spoke so from this figure we consider SGA and RAH as the most visited store by the customer for the year 2019. From the figure, you realize that each subplot is subdivided into 230 points, and for this figure, we took -230 as the central point of all the variables. As most the store has the spoke magnitude of zero or near to zero which not only makes the clutter but also makes no sense. As mentioned above that this problem will be handled using different techniques so we set the center point as negative so that reader can easily conclude that most of the store has the spoke magnitude equal to zero

Grouped Bar Chart

The bar chart as shown in fig 2, is one of the simplest and easy to understand. It is very effective when we deal with one variable or two on the same side i.e., x-axis or y-axis and their value on the other side. Because of its simplicity, it allows readers, to gather the required information easily. Generally, the x-axis is used to compare the data set while the y-axis for the change over time of that data set and the height or length of the bars show the frequency of that particular variable. Also, its overall shape has no meaning as the same data having different arrangements show the different types of bar charts. So, whenever a bar chart is used, reader should not try to gather information from its shape. As in our case, we have the data of expenditures on the marketing and overheads so we selected bar chart for it because of its simplicity everyone can easily infer that which store has more expenditure in terms of marketing and overheads throughout 2019.

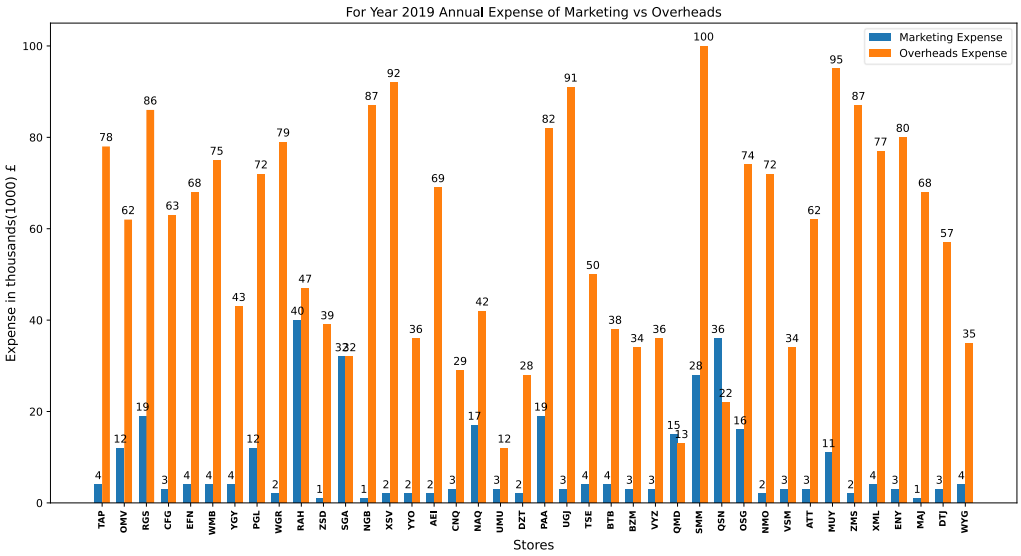

Figure 2: A Grouped Bar Chart Which Shows the Annual Expense of Marketing vs Overheads.

As in fig 2, we visualize two quantities on the y-axis of the bar chart which changing over one variable on the x-axis. For this, we use a bar chart having a different color for each quantity. In our data set, we have two types of expenditure against every store first one is “Marketing” (Expenses on the marketing of the stores) and the other one is “Overhead” (includes the other expenses of the store i.e., bills, salaries, and such things) so we use these quantities on the y-axis and the expenses on the x-axis for the visualization. As the value of expenses is very high so we normalize them by 1000. The length of the bars shows the value of the expenses by each store throughout 2019. In the chart, the colors we use for the two different quantities are “Blue” and “Orange” which show “Marketing” and “Overheads” respectively. Precisely we infer many things from this visualization but the most important and the most noticeable thing is that every store invests more in overheads as compared to the marketing and “QSM” is one of the expensive stores in terms of Overhead while “RAH” in terms of marketing.

Correlation Matrix

As shown in fig:3 this matrix is known as a correlation matrix. It is used to show the correlation coefficient between the different variables of the data. The best thing about is the matrix is that shows all

the results on the same matrix i.e., if you find a correlation of A with B, C, and D then on the same matrix it shows all the results. Each correlation between two variables can be represented by each cell. The main purpose of this matrix is the summarization of the data or you can say advance analysis of the input. The main point while creating the matrix is the selection of variables from the data of which you want to find relations as most people make a mistake by finding the correlation between all the variables of which mostly are of no use. Traditionally the matrix is formed by showing the same variables on the rows and the columns as shown in fig:3. As in our case, we have many important variables i.e., Staff and Size or Marketing and Overhead which need to be correlated so this visualization is the best choice for this type of data.

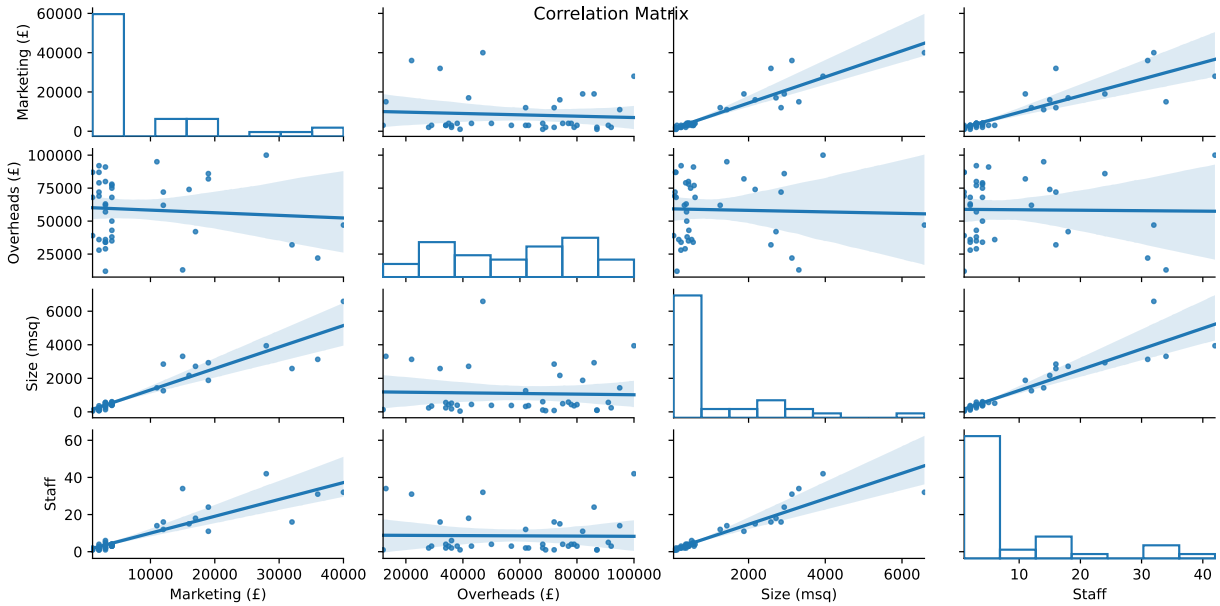

Figure 3: A Correlation Matrix

As in Fig:3, we visualize the correlation matrix using the four variables i.e., Marketing, Overhead, Staff, and Size. All the variables are written on the rows and columns respectively. From the visualization is visible that which variables are highly correlated and which are not. As if we select the correlation cell of “Marketing” and “Overhead” it is very clear that their co-relation is not strong is they have a straight line on the x-axis which at the end show some decline. While on the other hand if see the “Size” and “Staff” cell it is very much visible that they have a clear positive correlation in the upward direction which means both these variables are high co-related. From the fig:3 it is also very clear that “Overheads” is the only variable that lies the low and negative correlation with the other variables as compared to the other variables. This visualization is the most suitable one for this type of data as from this we will very clear which variables have the relations and which have not for the further visualizations.

Scatter Plot

Fig:4 shows the diagram which is known as “Scatter Plot”. It is typically used to visualize two variables of data with the use of Cartesian Coordinates. For the visualization of more than two variables color, shape and size are the additional features of this diagram. As with the variations of these you can also display as many as 5 variables on the single scatter plot easily. In this plot data is visualized as a

collection point, determining from the x-axis and y-axis respectively. Scatterplot is used to visualization different types of correlation between the variables i.e., “Staff” and “Size” on the x and y-axis respectively show some relation between them. Same as the correlation chart it also shows the type of relation i.e., positive, negative, or no relation. For the data like ours’s this is also a very good choice as from the correlation matrix we concluded that many variables have a strong relation between them but what happened when we add more variables to their relation.

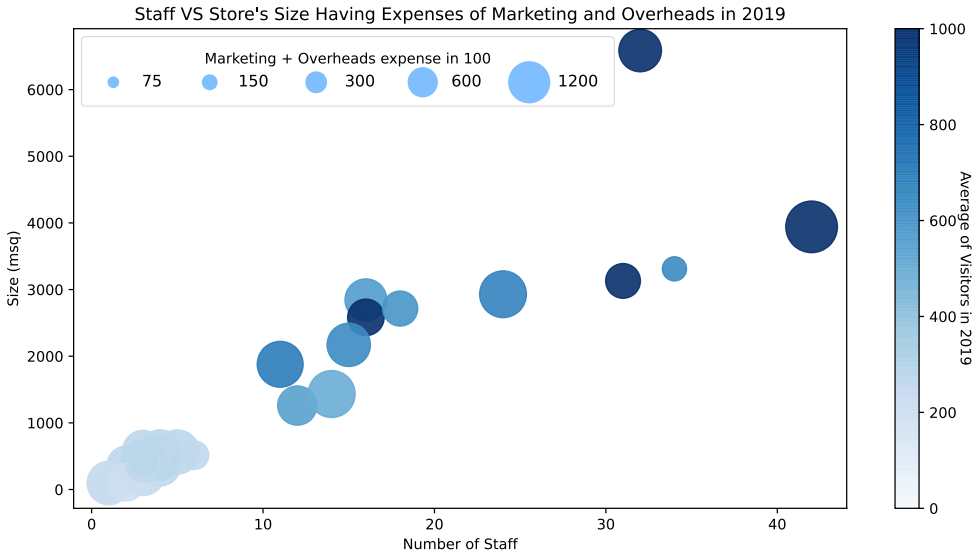

Figure 4: A Scatter Plot Which Shows Staff VS Store's Size Having Expenses of Marketing and Overheads in 2019

In fig:4 we visualize a scatter plot having “Number of Staff” on the axis while “Size” on the y-axis, plus we also add two more features using the size and color. As color represents the Average number of visitors and the size shows the “Sum of Marketing and Overhead Expense”. This plot shows many things about our data. As if we talk about the color, we realize that points near to zero have the light color means a smaller number of “Average Customers” but almost all of them lies the same size of the circle as of the points having the large number of daily customers which concluded that although these points have a smaller number of staff and small size of store, they lie the same expenses as the large stores but with the smaller number of daily customers which is a total loss. It has visible that most of the stores follow the same positive relationship between “Staff” and “Size” but if you see there are two points which are considered as the outliers as both of them have different behavior as compared to the other. So, this visualization is the most informative as compared to the others because it shows the relation between four variables of the data.

Bar Chart

Fig:5 shows the visualization of a simple bar chart. As mentioned earlier, it is the easiest and simplest visualization of all the other. This visualization is very interactive and shows the insight even

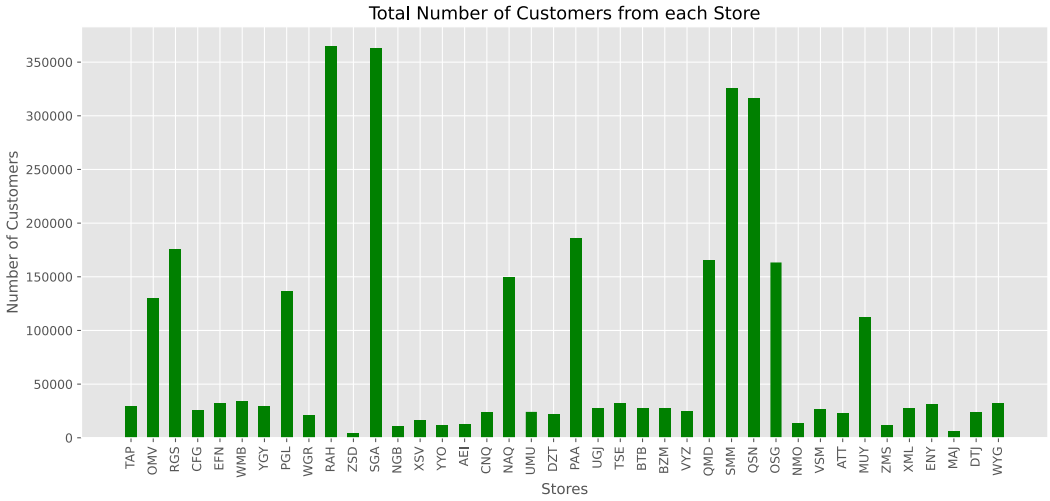

Figure 5: A Bar Chart Which Shows the Total Number of Customers in Each Store

looking once. Mostly used for categorical data but you can also use it for numerical data with some different angles. Readers prefer the bar chart visualizations because of its easy to decoding feature as in this chart you can have some variable on one axis and its value on the other. Besides all its also have the disadvantage which is related to its shape as it shows the data in the form of bars and if you change the input in terms of sequence then it changes its shapes and most people try to gather information from its shape which is very wrong or lead to the wrong perception. In our case, we have many variables to be visualized on this chart but for the simple bar chart, we choose “Number Of customers” on the y-axis and “Stores “on the x-axis which aggregately shows a very good insight.

As shown in fig:5 we use a simple bar chart having a variable name as “Stores” and “Number of Customers”. The main purpose of this visualization is to get information about the customers who visited the specific stores throughout 2019. As per the chart, we easily identify which store is the most visited one and which is the least. As mentions it’s the simplest and the easiest to be understood to from this reader can easily get the insights which of course lead toward the wise decision i.e., if someone wants to invest in some store then this chart is very useful as anyone can easily see that which store has the more customers attraction of the all. Besides this based on this chart, many store owners also make a wise decision for the marketing purpose by looking at the strategies used by the store which has the most customers attraction. Precisely the simplest and most useful visualization among the others.

Stacked Bar Chart

Fig:6 shows the “Stacked Bar Chart”, one of the finest and the most informative types of the bar chart. As mentioned earlier that simple visualizations of bar chart are the simplest and the easiest ones and the reader can easily get insight from it but very limited information due to the contribution of only one variable so for this purpose, this type was introduced which not only the simplest one but also the most

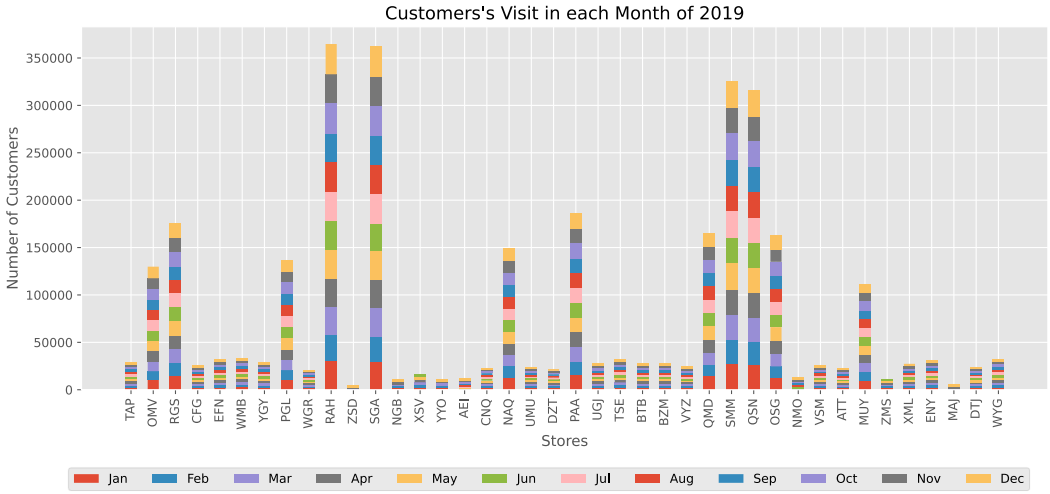

Figure 6: A Stacked Bar Chart Which Shows Customer's Visit in Each Store

informative one as we can easily visualize two or three variables on it. As in our case, we visualize "Date", Stores, and "Number of Customers" on a single bar chart.

As shown in Fig:6 we use three variables which show the Customer Visit in each Month of 2019. In this visualization, we divide the customer visit over the date or you can say monthly basis. If you see we use the x-axis for the "Stores" and the y-axis for the "Number of Customers" while the stacked bar with different colors for the months. Precisely each color represents the number of customers who visited that store over that month. This visualization will help the owner to check that which are the hot selling month and which are not and the circumstances depending on the months due to which customers visited their store i.e. Weather, Season and such things.

Violon and Box plot

Fig:7 shows the comparison between violin and box plots. A violin plot is a way to plot the numerical data. Box plot and violin plot are the same the some of the additional features in the violin plot i.e. Probability density and kernel density. Generally, Violin plots are used when there is a need for more and deep detail of the data as it includes all the information of the box plot with the density measures. Box plot is also known as whisker plot and it works on the quartiles. typically for this, we find the quartiles of the datapoint and plot it. The box plot is mostly used to identify the outlier as in the box plot any reader can easily identify that which data point is not the actual part of the outlier as compared to the other. In our case, each plot helps a lot in many aspects especially while finding the outliers. Also as mentioned the Violon plot shows the densities so it will also help in the finding of insight regarding the densities.

In Fig:7 as you see we use the "Marketing" and "Overhead" expenses to compare both of the plots. From the visualization, it will be seen that Although both of the plots tend to show very useful information violin plot has an important factor of density which the box plot lacks. As shown in the violin plot that we do not identify the outliers but also, we check that where most of the data points reside. while if we talk about the box plot then we came to know that the box plot is very good in outliers and if we see the "Marketing"

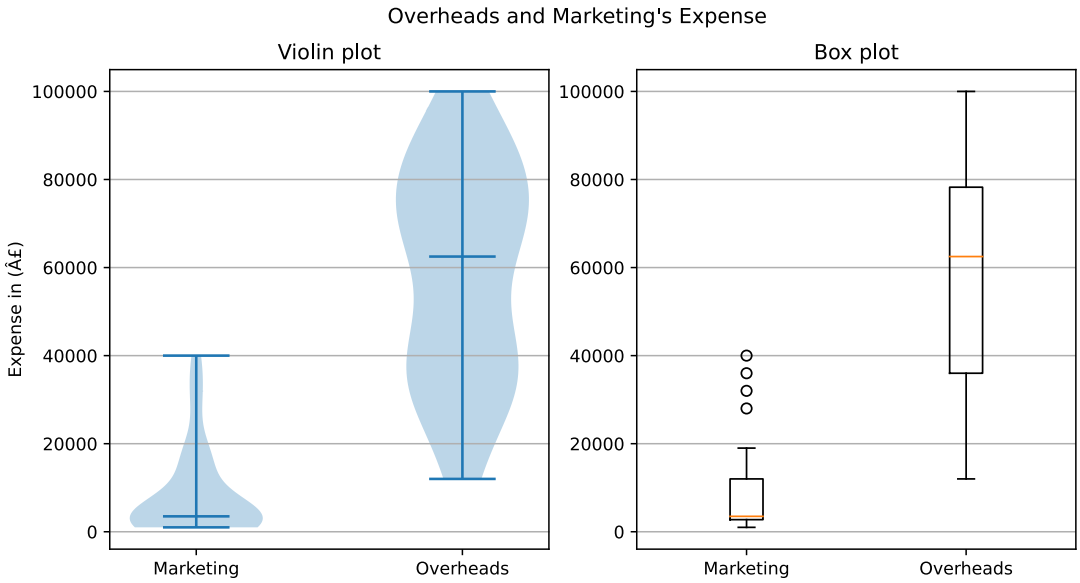

Figure 7: This Box Plot Vs Violine Plot Shows Overheads and Marketing's Expense

the portion of the box plot where outliers are identified then we realize that both the plot have their strengths and weakness and especially for the numerical data these both works very well

Histogram

Histogram as shown in fig:8 is one of the finest and easiest visualizations same as the bar chart. It is very optimal when data is in numerical form.

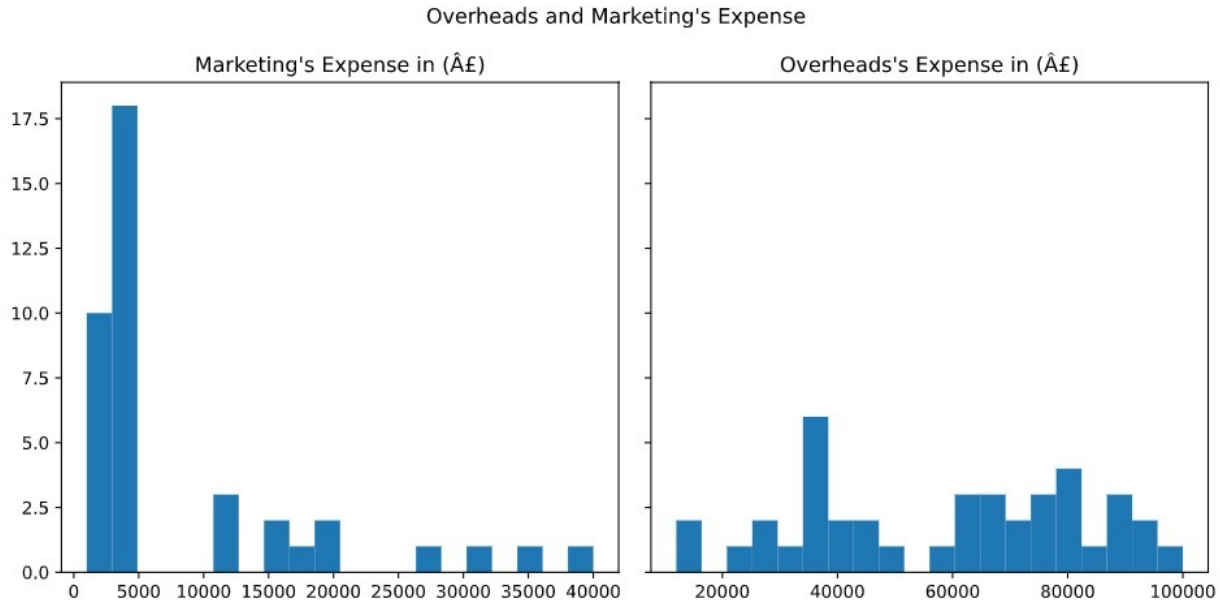

Figure 8: A Histogram Which Shows the Distribution of Overheads and Marketing's Expense

Histogram bins are very common, we usually divide the data into small bins while dealing with the histogram. If the bins are of the same size, then a rectangle is erected with the frequency. In our case, this helps a lot as most of our data is numerical and can be easily preprocessed for this. The best thing about this plot is that it also identifies the skewness in the data i.e., normal distribution, right or left skewness.

As in fig, we see the comparison of “Marketing” and “Overheads” using the histogram and as per its specialty, this plot clearly shows which data is skewed and which is not. if we see, the data points of the “Marketing” have high frequency on the left side which make it “Right Skewed” while on the other hand if we see the “Overhead” has two “Bell Shaped” distribution having no skewness in the data.

. Critical Review

In our visualization, we tried our best to demonstrate every point so that readers will easily get useful information from it by just looking at it. We mostly focused on the simplest and the easiest visualizations i.e., Bar chart and its types. But the thing we missed in our visualization is the factor of the statistics. In most of the charts and graphs, we just focused on the visualization formed with the comparison of two or more variables. According to my review if some of the visualizations show the pure stats i.e., Z-score, T-score, stand deviation of the data then these will be more useful and more informative as with the use of currently applied visualization we only gather the information of the comparison but what if we need to know about the specific data or the variable of the data i.e., strong or the weak point of the variable, stats of the variable and such things. Also, the distribution of the data is not suitable for the simple visualizations as we have 40 stores and mostly, I faced the problem of cluttering due to which of course any reader can miss a lot of useful information. Precisely with the use of some basic and some advanced visualizations I tried to draw each side of the data but still due to some hidden factors I think I lacked some information that should be visualized or which will be visualized using some advanced techniques.

Summary of the Data Points

A lot many conclusions have been made from the data points of the Store data but the most important are those which are dramatical. If you see the scatterplot visualization it clearly states that although most of the stores have less staff and size with the smaller number of visited customers still, they avail the same expenses as compared to the stores having a large number of staff , size, and customer visits. Another clear conclusion is that every store has employees according to the size of the office as the stores having large size possess a large number of staff and vice-versa except the two outliers. Besides all the datapoint “Overhead” is the only one which has no relation or weak relation with every other point if we see the correlation matrix then we came to know that “Overhead” has only weak or negative relation that the reason that it behaves oddly compared to the other variables. If we see the other visualization then we concluded that to get insight from this type of data we should compare at least two to three variables as shown in Bar charts where we see the trends of the customer to each store with the distribution of months. Also, one other form of the bar chart shows the trend of the customers towards the specific store from which at least store owners concluded that weakness and strengths of their stores which attracts the customers. Precisely after the different visualizations, we concluded that we can get

very useful insights in terms of business, investment, visit, and such things if we utilize the data with little care which is very clear in our visualizations.