**Sales Prediction using Machine Learning**

So here are few steps guidance on how to predict sales using machine learning



Introduction:

So here you will get a complete analysis of a Sales Prediction model. Here I will cover all major insights that we used to be see in our ML model or projects, we will do the complete analysis from data visualization to model building and finding

the key observations from the analysis, that will help us to predict the results.

**So, following are the required steps:**

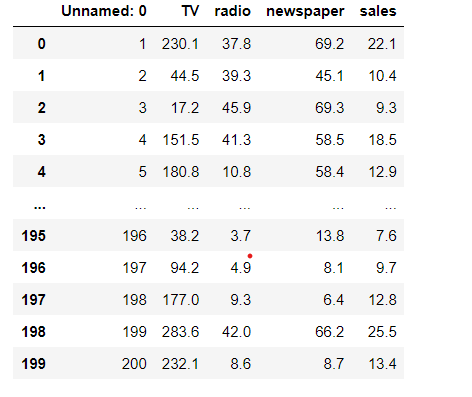
**1.Problem definition**

* So, the problem is that when a company enters a market the distribution strategy and channels it uses are keys to its success in the market
* Because an effective distribution strategy under efficient supply- chain management open doors for attaining competitive advantage and strong brand equity in the market it is a component of the marketing mix that I cannot be ignored.
* The distribution strategy and the channel design have to be right the first time.
* This case study of sales channels say includes the detailed study of TV, radio and newspaper channel.
* Thus predicts the total sales generated from all the sales channel.
* So here we have clarified what is the problem.

**2. Data Analysis**

Now we are just analyzing our sales dataset, and let’s see what

insights we gain from it. So, I am sharing a picture of the dataset, I have used for the analysis.



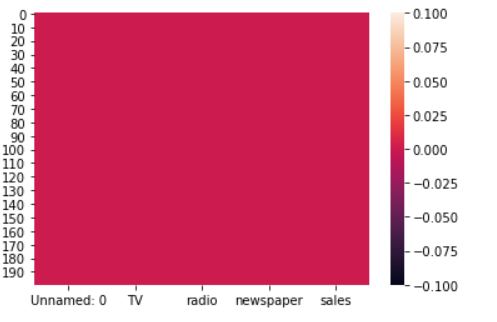
* We can see that we have 200 rows and 5 columns in the dataset.
* We have 5 columns Unnamed :0, TV, radio, newspaper and sales.
* We check the properties like shape, unique, dtypes etc.
* We have sales column as our target and is continuous in nature, thus it is a Linear Regression problem.
* Now I will do the further analysis according to our problem type, which is linear regression is a type of Supervised learning.

**3.EDA Concluding Remarks**

Now we will see the complete EDA of the dataset. By doing following steps

1. **Check missing values**
2. **Statistical Summary**
3. **Univariate Analysis**
4. **Bivariate Analysis**
5. **Multivariate Analysis**
6. **To check Skewness**

Let’s discuss the above steps in detail:

1. We can see whether null values are present in the dataset or not, we also use heatmap like I have done

We can see that no null values are present in the dataset.

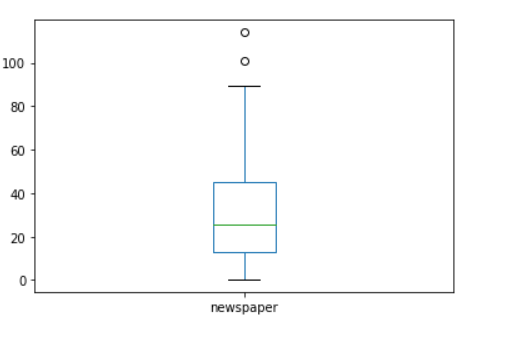
**2)** Statistical summary gives information about the mean, median, std, min, max etc.

key observations:

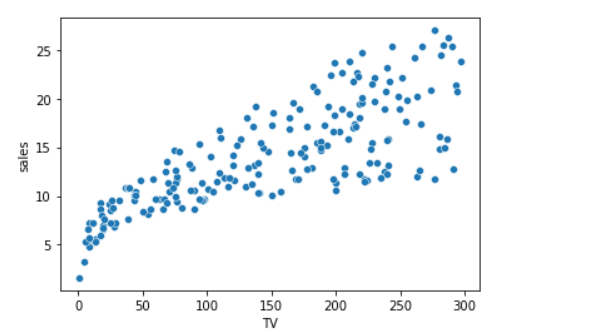
1 The difference in mean and median is almost similar.

2 There is small difference in 75% percentile and max in columns named Unnamed: 0, TV and radio which shows that no outliers are present in it.

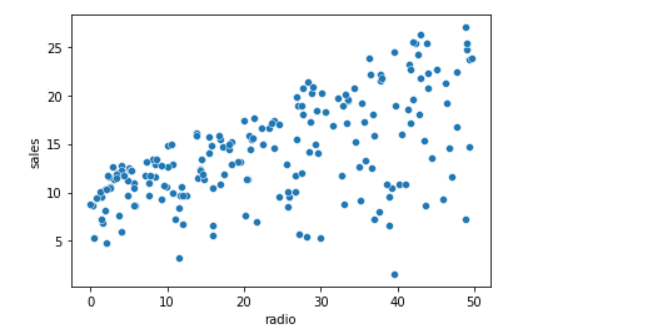
3 we can see the difference in 75% percentile and max in newspaper column which shows that few outliers are present in it.

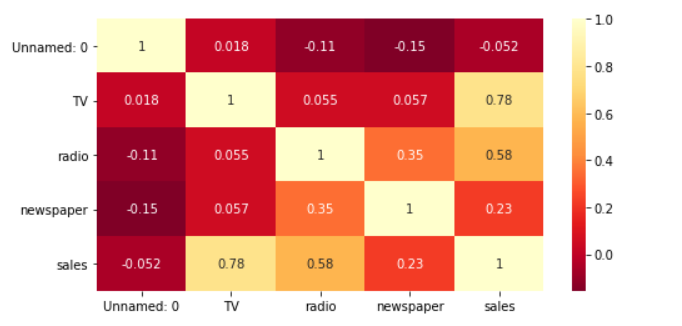
**3)** In univariate analysis I have plotted boxplots, from these plots we can see the mean, median, max, min and we can also see whether outliers are present or not, as we can see in below picture, we can see that outliers are present in newspapercolumn.

**4)** Now in Bivariate analysis I have used Scatter plot to see the relation of each column with the sales column.

we can see the scatter plot of TV column with the sales. Here we can see the positive relation between the sales and TV, as the TV advertisement increases sales also increases.

We can see also see the plot of radio column also, here we can see that the sales increases with the increase in radio advertisement but sometimes sale is decreased too.



**5)** In multivariate analysis we can check the correlation matrix of the dataset, which clearly shows us the relation of each column with other column, here is the correlation matrix.

Observations:

Light shades are highly correlated.

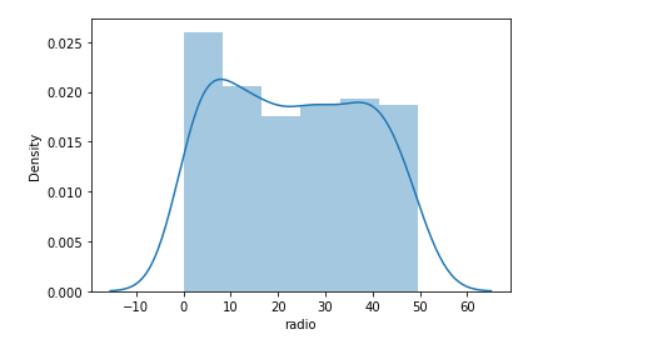
Sales are highly positively correlated with Tv column.

Sales are highly negatively correlated with Unnamed:0 column.

Sales are slightly positively correlated with radio column.

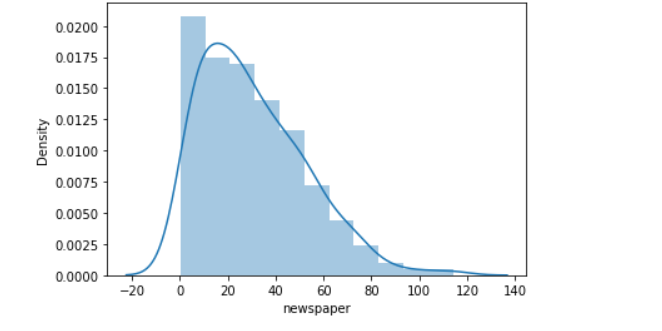
Sales are negatively correlated with newspaper column.

**6)** Now I have checked the skewness of each column by plotting density plot, to check whether the curve is normally distributed or not, how much skewness is present? We have skewness in many columns let’s see in radio column



Here we can see that curve is not normally distributed, thus we can see the skewness clearly.

We can also see the skewness of newspaper column in the below picture.



Here we can also see that the curve is slightly skewed from the right.

**4.Pre-processing pipeline**

we have many steps included in pre- processing like Data cleaning, Data reduction, Data integration etc. Let's discuss each of them in detail the steps we have done in our project

according to our requirement.

1. We have drooped the rows which are negatively correlated, here we have dropped the Unnamed: 0 column from the dataset due to its highly negative correlation with the sales column. This column impacts our data negatively, thus we dropped that column.
2. Now we are removing the outliers present in our data, we have two methods to remove outliers one is zscore and other is by using IQR method here I have used zscore method to remove all the outliers present in our dataset, I have very less outliers in my dataset thus only two rows are deleted from the dataset as outliers.
3. Here we have seen above that no null values are present in our dataset, so here is no need to handle missing values, if there were any missing data then we have to treat it with suitable method, but here no need.
4. Next is the we check whether any column is present in string format or not, if any column is present in string format, then we have to change it in integer format by applying Encoding technique, we have two methods in encoding one is One hot encoding and other is Label Encoding, but here in our dataset we have no need to apply encoding technique because all the columns are already present in integer format. So, let’s move on to the next step.
5. So now we are removing skewness from our dataset, as we have seen that skewness is present in all the columns except the sales column, to remove the skewness we have separated the target variable and the independent variable from the dataset. As we know that skewness between –0.5 to 0.5 is acceptable but more than it is not, so skewness except this range should be treated by using suitable method. So, here I have used cube Root method to handle negative skewness and square root method to handle positive skewness, till now we have treated the skewness by suitable method and removed skewness from our dataset.
6. Next point which comes is the feature engineering but here in this dataset we do not need it, because we don’t have special characters, etc. So here we have no need to apply Feature Engineering at all.
7. The last thing which I can see is the **standardization** technique, we use this technique to scale our data. We have two methods to scale our data first one is standard scaler and the second one is min max scaler. We use these techniques only when there is huge difference between the ranges of any 2 columns, that’s why we use scaling. **Standard scaler** is used when data is normally distributed, its changes the mean=0, std=1 and the value ranges between –3 to +3. **Min –Max scaler** is used when data is not normally distributed this method is also called normalization, it changes the data with mean=0, std=1 but range is 0 to 1. In our dataset we have seen there is no huge difference in ranges of the columns, thus here we have no need to apply standardization technique on our dataset.
8. The last which I can see is the PCA this technique is used only when we have, we large numbers of columns and it's difficult to manage them all, but here in our dataset we have only 5 columns thus here we have no need to use this technique.

**5.Building Machine Learning Models**

Author Name: Neha