Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
   → There are seven categorical variables in the dataset namely season, year , months ,holiday ,weekday ,working day , weather situation. The target variable or the dependent variable is “cnt” . We used boxplots to visualise the categorical variables.

   1.The cnt increased in summer and made its peak in fall and then decreased.

   2.If we look at the months they explain the same story as the seasons.

   3.The cnt increased in 2019.

   4.People used the service mostly when the weather situation is Clear, Few clouds, Partly cloudy, Partly cloudy.

   5.There is no data related to Heavy Snow.

   6.All the days have medians in the same range for the weekday.

   7.Majority of bookings are on working days

2. Why is it important to use drop_first=True during dummy variable creation?
   → We created dummy variables for weather situation , season , months , weekdays .
   It is important to use drop_first=True , because it helps in reducing the extra column created during dummy variable creation.
   Example:

Furnished	Semi-Furnished	Unfurnished
1	0	0
0	1	0
0	0	1

Suppose we drop Furnished , so in order to represent Furnished we can say that when Semi-Furnished and Unfurnished is 0 , it means Furnished

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   →From the heatmap it is clear that ‘registered’ is highly correlated with target variable.Then ‘casual’ is highly correlated with target variable.But if we look closely the summation of casual and registered is cnt . So they will have high correlation.
   Temp,atemp variables are highly correlated with cnt variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   →In order to use linear regression , there must be some linear relationship between target and predictor variables . We checked this using pairplot and found that linear relationship exists.
   After we built the model , we did the residual analysis. Residual is the difference between the actual value and predicted value. The assumption of normal distribution of error terms with mean zero was found correct with the histogram we plotted.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   →Based on the final model we found that these variables were significant in explaining the demand:
   1.temp
   2.year
   3.weather situation

General Subjective Questions

1. Explain the linear regression algorithm in detail.
   →Linear regression algorithm is one of the most important ML algortihms.This algorithm is based on Supervised Learning .The output from this algorithm is a continuous one.

Linear regression algorithm is used to predict an output variable y with the help of one or more independent variables / predictor variables.

The general form of Linear regression is given by:

y=Beta0 + Beta1*x

where , Beta0= value of y when x =0 i.e intercept
Beta1 is slope

Similarly, for multiple linear regression we have many independent variables which will each have a coefficient with them .

y=Beta0 + Beta1*x1 + Beta2*x2 +…. + Betan*xn

Interpretation:
Change in Y with change in a variable when other variables are held constant.

The coefficients of variables are calculated by minimizing the square of the sum of residuals.
Residual=Actual Value – Predicted value

Linear Regression Assumptions:

1.There is a linear relationship between X and y.
2.Error terms are normally distributed with mean zero.
3.Error terms are independent of each other.
4.Error terms have constant variance.

Hypothesis Testing:

The null hypothesis is that
H0 : Beta1=0 i.e the coefficients are insignificant
The alternate hypothesis is that
H0 : Beta1!=0 i.e the coefficients are significant

Steps in building the model:

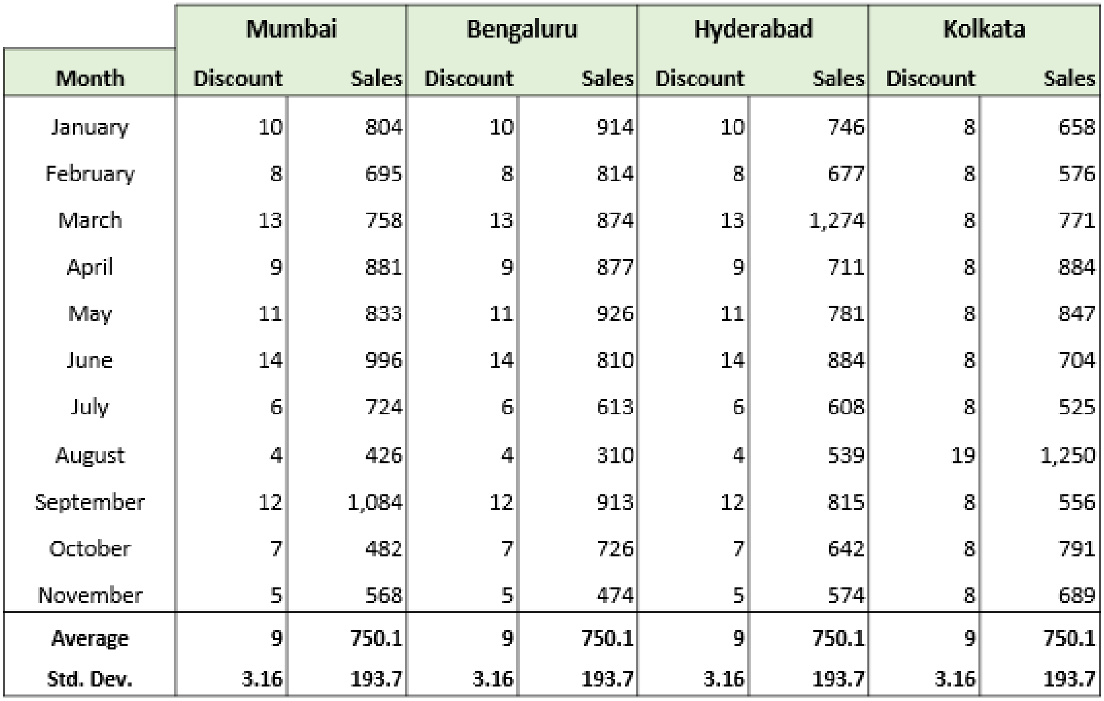
1.Reading the data set and performing analysis .
2.Checking datasets for null values,checking data types of variables and converting them if necessary.
3.Visualising both numerical and categorical data and getting inference.
4.Plotting heatmaps to get the correlation between the variables.
5.Use one hot encoding for the categorical variables.
6.Concatenating the data frames.
7.Splitting the data into train set and test set.
8.Scaling the variables for ease of interpretation
9.Building model using libraries like scikitlearn and statsmodels
10.If the number of predictor variables are more in number we use Recursive feature elimination to find the top variables
11.Using variance inflation factors to reduce multicollinearity
12.Based on p values and VIF remove variables one by one and build models. In our case we stopped when the pvalues of variables was less than 0.05 and VIF less than 5.

When p value of the variables is equal to zero it means the that value of coefficients of respective variables is significant and we can reject null hypothesis

13.Plotting histogram of error terms to check the assumption that they are normally distributed with mean zero
14.Make predictions on test set using the model.The r2 score of test model must within +-5 range of values of r2 score of train values

2. Explain the Anscombe's quartet in detail.
→The Anscombe's quartet consists of four data sets that have nearly identical descriptive statistics, yet have very different distributions and appear very different when graphed . This was constructed by Francis Anscombe .Each data set consists of 11 points (x,y).
It was constructed in order to prove the the importance of graphing data before analyzing and the effect of outliers on statistical properties.

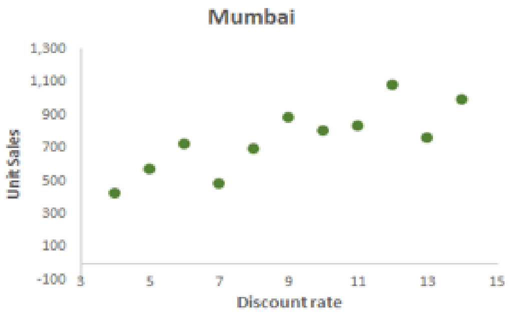
Example of Sales vs discount:

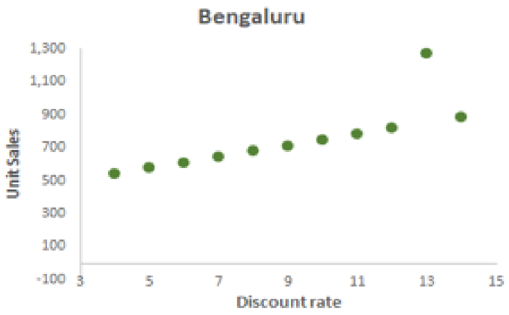


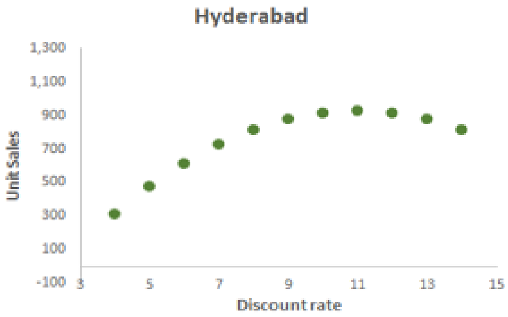
	Mumbai		Bengaluru		Hyderabad		Kolkata	
Month	Discount	Sales	Discount	Sales	Discount	Sales	Discount	Sales
January	10	804	10	914	10	746	8	658
February	8	695	8	814	8	677	8	576
March	13	758	13	874	13	1,274	8	771
April	9	881	9	877	9	711	8	884
May	11	833	11	926	11	781	8	847
June	14	996	14	810	14	884	8	704
July	6	724	6	613	6	608	8	525
August	4	426	4	310	4	539	19	1,250
September	12	1,084	12	913	12	815	8	556
October	7	482	7	726	7	642	8	791
November	5	568	5	474	5	574	8	689
Average	**9**	**750.1**	**9**	**750.1**	**9**	**750.1**	**9**	**750.1**
Std. Dev.	**3.16**	**193.7**	**3.16**	**193.7**	**3.16**	**193.7**	**3.16**	**193.7**

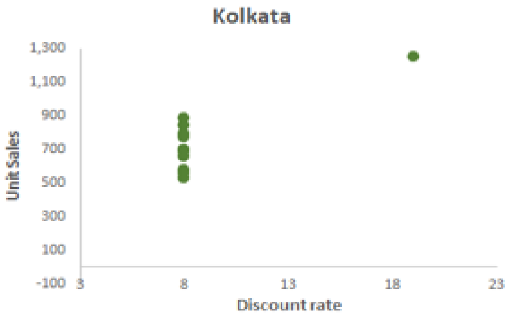
The std dev. and average for both discount and sales is same for all the cities .

Lets look at the graph plots:









From the above images it is clear that the graphs can explain a lot of things and give insights rather than just drawing insights from numerical values.For Mumbai even though the discount rates kept on increasing , there were some points were the sales dropped.For Bengaluru , the sales increased slightly with increase in discount rate.For Hyderabad the sales first increased and then decreased with increase in discount.For Kolkata even though the discount rate was same , the sales increased .

3. What is Pearson’s R?
   → Pearson’s R is the correlation coefficient between two variables.The values lies between -1 to +1.
   +1 means there is a strong correlation between both the variables i.e if one increases the other increases too whereas -1 indicates that if one variable increases the other decreases. Correlation doesn’t imply causation.
   Pearson's correlation coefficient is covariance of the two variables divided by the product of their standard deviations.
   Example of positive correlation: As the temperature goes up, icecream sales tend to grow up.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   →Scaling is a data preparation step which is applied to numerical variables .In a dataset , the numerical variables can have different magnitude and units. When we use this data for machine learning models without scaling , the algorithm doesn’t take into account the units of these variables. Hence , in order to avoid this , we use scaling so that all the numeric variables are at the same scale.
   It helps in easy interpretation and doesn’t affect t-statistics,F-statistics, p-values, R squared.
   It just affects the coefficients.

   We can use the below methods:

   1.Min Max Scaling : It helps to bring the values in range of 0 to 1.
   We can use from sklearn.preprocessing import MinMax scaler.
   Then we can make a object and then use that object to scale the data.

fit_transform is used on train set whereas transform is used on test set.

x= x- min(x) / max(x)-min(x)

2.Standardization:

It is similar to calculating zscores. It converts data into standard normal distribution which has mean zero and standard deviation 1.

Z= x-mean/sigma

We can use sklearn.preprocessing.scale for this type of scaling.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   →VIF is given by :

   VIF= 1/ 1-Ri square

   When the value of Ri square is 1 then VIF will become infinity.It indicates there is a perfect correlation between the independent variables.
   In order to avoid this , we can drop one of the correlated variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   →Q-Q plot stands for Quantile-Quantile plots .These are the plots of two quantiles against each other. The purpose of Q-Q plot is to find out if two sets of data come from the same distribution.
   If the two distributions being compared are similar , the points in the Q-Q plot will approximately lie on the line y=x.