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? (3 marks)

Ans. Following are the analysis of the categorical variables from the dataset on their effect on the dependent variable 'cnt' :

1. The categorical variable 'season' had an overall positive effect on the dependent variable 'cnt' as it had correlation of 0.4. Most of the bookings happened in season of Summer and Fall.

2. The variable 'yr' also had positive effect on 'cnt' with their correlation of 0.57. It was observed that there were more number of bookings done in 2019 than that in the year 2018.

3. 'Holiday' on the other side had little bit of negative effect on 'cnt' with there correlation of -0.069, though it is not huge but it was observed that number of bookings happening on non-holidays was significantly larger than that on holidays.

4. 'Weekday' had little to nothing effect on 'cnt' with their correlation being at 0.068, it was observed that number of bookings happening were almost similar for each weekday.

5. 'Workingday' also had little to nothing effect on 'cnt' with their correlation at 0.063.

6. 'Weathersit' had negative effect on 'cnt' with their correlation being at -0.3. It was observed that people were booking more on Clear and Mist weather and less on Snowy weather.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

Ans. It is important use drop_first = True during creation of dummy variable because, the rest of the variables could easily define the dropped variable.
Example: Let us consider we have a categorical variable 'weather' having values ['clear','rain','snow']. So their dummy variable would be:

Clear : 1,0,0
Rain : 0,1,0
Snow : 0,0,1

So when we use drop_first=True the newly created column clear will be dropped but we could easily get the value for clear weather. If the Rain and Snow column have values 1,0 respectively it would indicate that the weather is Rain and if their values are 0,1 it would indicate the weather is Snow. But in case of Clear weather both the values of Rain and Snow will be zero indicating that weather is not Rain and Snow.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans. Looking at the pair-plot for the numerical variables, atemp and temp had the highest correlation with the target variable 'cnt'. This is after dropping columns casual and registered.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans. To check the assumption of Normality of error a distribution plot is used to show that the mean of the error term in at Zero
To check the Linearity Relationship pair-plot is used to show linear relationship of the variable with that of target variable
To check the assumption of Homoscedasticity a scatter plot is used to show that their is no visible pattern in the residual values.
And finally to check that there is no multicolinearity within the dependent variables VIF and Correlation Heatmaps are used.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

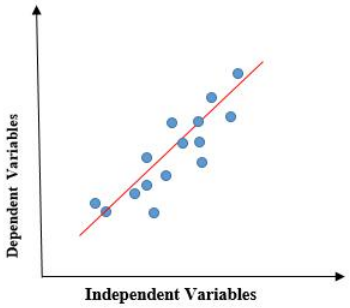
Ans. Based on the final model the top 3 features contributing significantly towards the explanation in the demand of the shared bikes are 'yr', 'atemp' and 'hum'.
'yr' and 'atemp' had positive coefficient values of 0.224247 and 0.425314 respectively indicating increase in one unit of each feature would lead to increase in the 0.224247 and 0.425314 unit of the target variable assuming all the other variables stays constant.
The variable 'hum' had negative coefficient value of -0.251472 indicating increase in one unit in 'hum' will lead to decrease in 0.251472 units of 'cnt' assuming all the other variables remains constant.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Ans. Linear regression is a supervised machine learning algorithm in which it shows a line that passes through all data points on the target predictor in a way that the distance between the points and line is minimum.

Linear Regression is simple statistical algorithm which shows the relationship between continuous data points. In short it shows the linear relationship between independent variable/variables (X) and dependent variable (Y). If the there is a single independent variable than such linear regression is called Simple Linear Regression and if there are multiple independent variable such linear regression is call Multiple Linear Regression. The linear relationship between independent variable and dependent variable is defined by the slope-intercept formula (y = mx +c) where m is the slope coefficient and c in the intercept.



The above graph presents the linear relationship between the dependent variable and

independent variables. So when the value of independent variable increases, the value of dependent variable increases. The red line is known as the best fit line.

In case of multiple independent variables the formula can be written as

y = a0 + a1*x1 + a2*x2 + a3*x3 + … + e,

Where: a0 is constant, a1,a2,a3… are the coefficient, x1,x2,x3… are the independent variable and e is error.

To find the best fit line, a cost function is used called Mean Squared Error (MSE), which is the average of squared error occurred between the predicted values and actual values. The use of the cost function is to find the best fit line by the means of minimizing the error between predicted values and actual values.

We check the linear regression's best fit line performance we used the metric R-squared. R-squared determines the goodness of fit, i.e. it measures the relationship between the dependent and independent variables. If the R-square value is high it indicates that there is less difference in the predicted values and the actual values and hence it represents good fit of the line.

There are few assumptions in Linear Regression that we have keep in mind and check for it before finalizing the linear regression model.

2. Explain the Anscombe's quartet in detail. (3 marks)

Ans. Anscombe's quartet comprises of four datasets that have different appearance when plotted but nearly identical statistical properties. Each of the dataset consists of eleven data points (x,y). It was developed by statistician Francis Anscombe to show both the importance of plotting data before analyzing and effect of the outliers on the statistical properties.

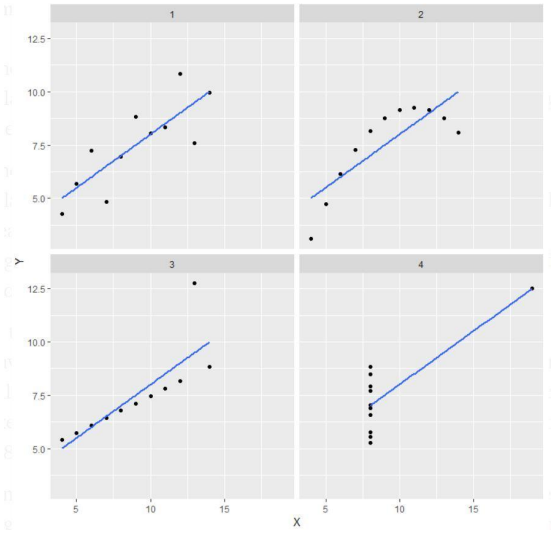
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

On analyzing the above datasets we can find the mean, standard deviation and correlation of all the datasets are identical.

Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

But the problem arises when we the the scatterplot of the datasets.



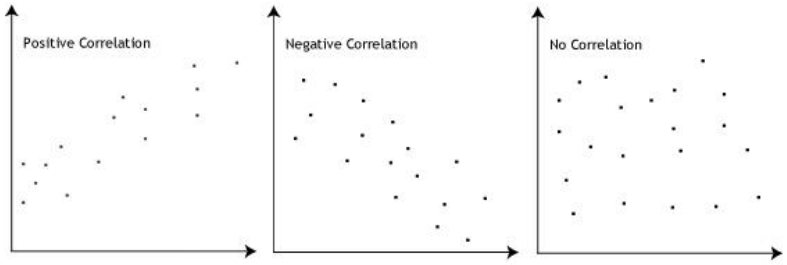
As we can clearly see that each dataset have different scatterplot but the regression line remains same. One first plot we can see that there is linear relationship between x and y but on the second plot (top,right) we see that there is non-linear relationship between the x and y. The third one has almost perfect linear relationship between x and y with just one outlier. And the last one shows an example how one high data point is enough to produce a high correlation coefficient.

The Anscombe's quartet is used to illustrate the importance of looking at the data by plotting before starting the analysis.

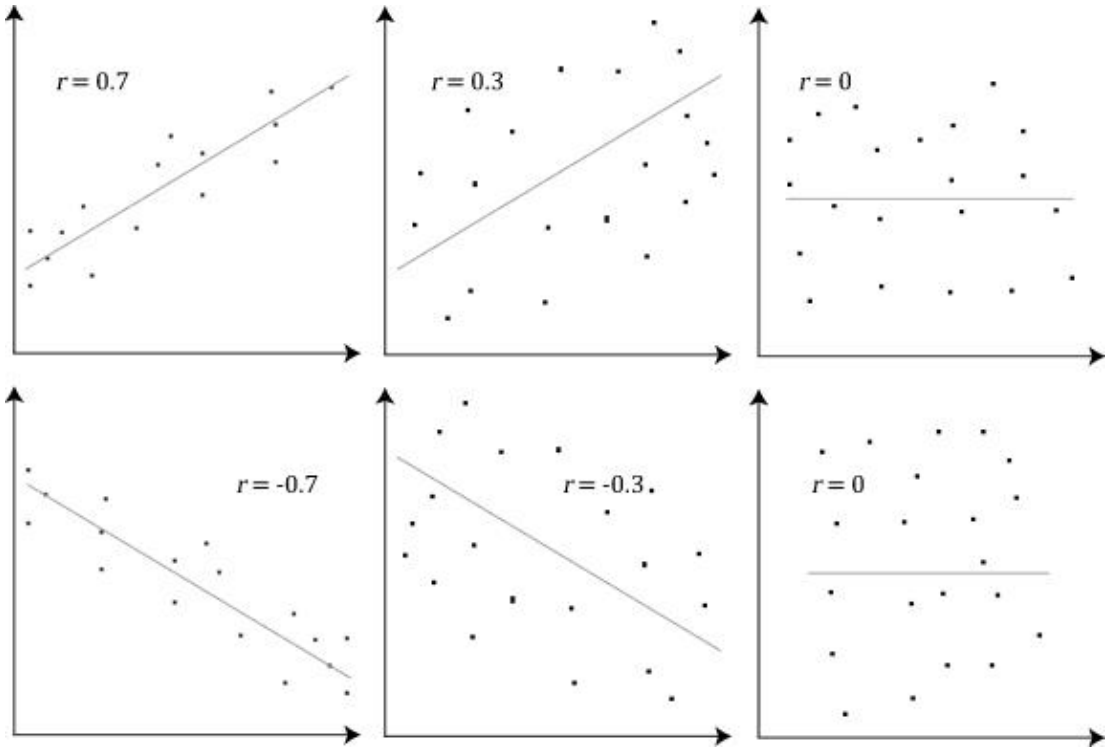
3. What is Pearson's R? (3 marks)

Ans. The Pearson's product-moment correlation coefficient or the Pearson's R is the measure of the strength of a linear relationship between two variables and it is denoted by 'r'. Pearson's R attempts to draw best fit line the data points of two variables and the coefficient,r indicates how far away all these data points are to the best fit line.

The Pearson's r can take a range of values from -1 to +1. A value of 0 indicates that there is no relationship between the two variables and a value greater than 0 indicates a positive relationship, that means increase in one variable increase the value of the other variable. If the value less than 0,it indicates that there is a negative relationship between the variables and as the value of variable increases the value of the other decreases.



The stronger the relationship of the two variables, the closer the Pearson's r, will be to either +1 or -1 depending on whether the relationship is positive or negative, respectively. Achieving a value of +1 or -1 means that all your data points are included on the best fit line – there are no data points that show any variation away from this line. Values for r between +1 and -1 (for example, r = 0.8 or -0.4) indicate that there is variation around the line of best fit. The closer the value of r to 0 the greater the variation around the line of best fit. Different relationships and their correlation coefficients are shown in the diagram below:



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans. Scaling is technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying values. If scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Normalized Scaling or Min-Max Scaling is used to transform features to be on a similar scale. The new data points is calculated as:

X_new = (X - X_min) / (X_max - X_min)

This scales the range to [0,1]. Normalization is used when there are no outliers as it cannot cope up with them. Usually we will scale age and not incomes because only a few people have high incomes but the age is close to uniform.

Standardized Scaling is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.

X_new = (X - mean) / std

Standardization is useful in cases where the data follows a Gaussian Distribution. In general it brings the mean of the new data points to zero, with standard deviation of 1.

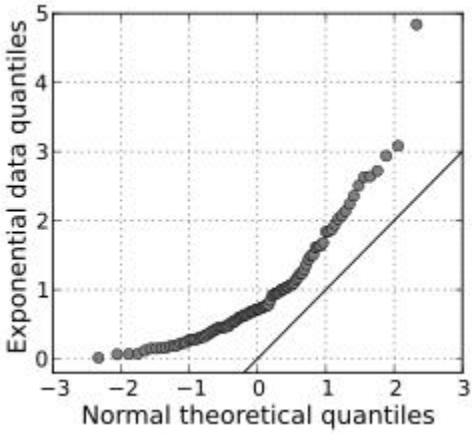
Difference between min-max scaling and standardized scaling is that in min max scaling the values range from [0,1] while in standard scaling the value ranges from [-1,1]. Min max scaling is affected by the outliers while in standardized scaling the the values are not affected by the outliers. We use min-max scaling when the distribution of the data is not known while we use standardized scaling when the data follows Normal Distribution.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans. If the VIF value is infinite it indicates that there is a perfect correlation between two independent variable.
In such cases the value of R2 will be equal to 1 so the value of 1 / (1 - R2) will be infinity.
To overcome this problem we drop one of the variable from the dataset which is causing multicollinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans. Q-Q plots (Quantile-Quantile plots) are plots that compare two quantiles. A quantile is a percentage of the population in which specific values fall below it. The median, for example, is a quantile where 50% of the data falls below it and 50% of the data falls above it. Q Q plots are used to determine whether two sets of data are from the same distribution. On the Q Q plot, a 45 degree angle is drawn; if the two data sets are from the same distribution, the points will fall on that reference line.



The points in the Q–Q plot will roughly lie on the line y = x if the two distributions being compared are similar. The points in the Q–Q plot will roughly lie on a line if the distributions are linearly related, but not necessarily on the line y = x.