Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer - By doing the analysis of categorical variables from the dataset, I was able to find the following inferences on the dependant variable (Count):

1. Fall season seems to have more renting of bike. And the booking count is drastically increased from 2018 to 2019. May be due to awareness of environmental issues.
2. Overall spread in the month plot is reflection of season plot as fall months seem to have higher median.
3. More people prefer to rent bike on working days and during holiday people prefer to stay home.
4. Clear weather or few clouds weather attracted more booking and booking increased from 2018 to 2019.
5. Overall, the year 2019 have done more business as compared to year 2018.

Q2. Why is it important to use drop_first=True during dummy variable creation?

Answer - During dummy variables the attribute drop_first = True is very important to use because as it helps in reducing the extra column and make the model less complex. And hence it reduces the correlations created among dummy variables.

For the attribute drop_first: bool, the default value is False, which implies whether to get k1 dummies out of k categorical levels by removing the first level.

For example:

Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If value A is 1 then value of B & C is 0, if value B is 1 then value of A & C is 0. Therefore if the value of A & B is 0 then definitely it would be C. So we don't need three variables.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer - Temp and atemp variable are more corelated with target variable cnt.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer - I validate the assumption of Linear Regression based on the below parameter are as follows:

1. Normality of error terms
2. Error terms should be normally distributed
3. Multicollinearity check
4. There should be insignificant multicollinearity among variables.
5. Homoscedasticity
6. There should be no visible pattern in residual values.
7. Linear relationship validation
8. Linearity should be visible between actual value and predicted value

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

Answer - Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes:

1. Temp
2. September
3. Winter

General Subjective Questions

Q.1 Explain the linear regression algorithm in detail.

Answer - Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of data based on some variables. In the case of linear regression as the name suggests, linear that means the two variables which are on the x-axis and y-axis should be linearly correlated. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

Where a and b given by the formulas:

$$b(slobe) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

Here, x and y are two variables on the regression line.

b = Slope of the line

a = y-intercept of the line

x = Independent variable from dataset

y = Dependent variable from dataset

Q.2 Explain the Anscombe's quartet in detail.

Answer - Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points

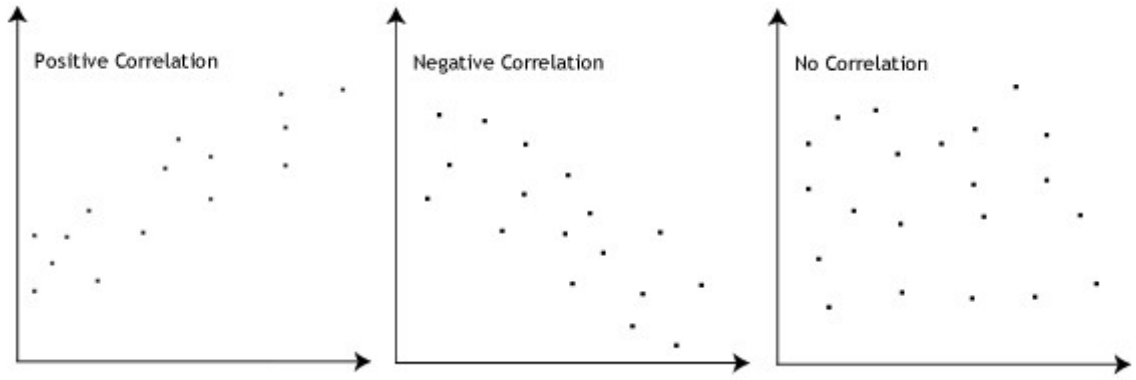
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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

The summary statistics show that the means and the variances were identical for x and y across the groups:

1. Mean of x is 9 and mean of y is 7.50 for each dataset
2. Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
3. The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

Q.3 What is Pearson's R?

Answer - In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



Pearson r Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Here,

- =correlation coefficient
- =values of the x-variable in a sample
- =mean of the values of the x-variable
- =values of the y-variable in a sample
- =mean of the values of the y-variable

Q.4 What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answers - Scaling is a 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 magnitudes or values or units. If feature 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. It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Q.5 You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer - If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Answer - Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.