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?

The categorical variables taken into consideration are season, yr, mnth, workingday, holiday, weekday and weathersit. The effect on the dependent variable i.e. cnt is as follows:

1. **Season:** We can see that there are less number of users in Spring season and highest is in the Fall Season. Where as summer and winter have almost similar distribution.
2. **yr:** Here we can see that more number of users have used the rental bike in 2019 than in 2018.
3. **mnth:** More number of rentals was done in the month of September, then followed by March and so on. Where as the least number of bike rentals were made in the month of January.
4. **workingday:** We can see that the median line is higher in workingday rather than on weekends or holidays.
5. **holiday:** We can see here that the rentals were less in holiday than on non-holiday.
6. **weekday:** Saturday has seen more number of bike rentals where as Monday has seen less number of bike rentals.
7. **weathersit:** There are no bike rentals done in 4 i.e. Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog, where as more rentals were done in 1 i.e. Clear, Few clouds, Partly cloudy, Partly cloudy.

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

It is important that we use drop_first = True so that it reduces one extra column which would be created since the desired way is that it creates n-1 dummy variables. To avoid an extra step of dropping a variable after the dummy variable creation it is always recommended to use drop_first = True.

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

atemp and temp have the highest correlation among the numerical variables with the target variable(cnt).

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

1. After plotting the Error terms it was seen that they were Normally Distributed and we can see that most of the Error was distributed around 0. If the error terms are normally distributed then we can say that the Assumptions are true.
2. We can also see that there are no multicollinearity as the VIF of the variables are less than 5.
3. As well as the p-value is less than 0.05 which means that the model is statistically significant.

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

1. yr: 0.232940
2. temp: 0.585357
3. light snow: -0.252410

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression is a supervised machine learning algorithm, the output to be predicted is a continuous variable which can be such as score of a student, house prices and so on. Regression is mostly used in forecasting or predictive modeling.

There are some assumptions of linear regression such as: the target variable and input variable are linearly dependent, linear relationship between X and Y, Error terms are normally distributed, Error terms are independent of each other, Error terms have constant variance and should not have multicollinearity.

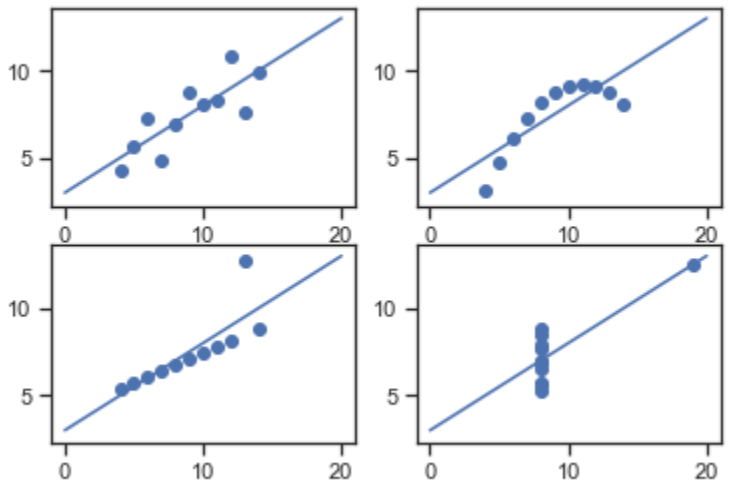
The equation of linear regression is: **$y = mx + c$**, where m is the slope and c is the intercept.

There are two types of linear regression:

1. **Simple Linear Regression**: In this we only have one independent variable and one target variable. The formula of SLR is: **$y = \beta_0 + \beta_1 X$**
2. **Multiple Linear Regression:** In this we have n-Number of independent variable but only one target variable, the variables can be eliminated by using RFE (Recursive Feature Elimination) or by manually seeing VIF and p-value. The formula of MLR is: y = **$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots\ldots + \beta_n X_n$.**

2. Explain the Anscombe's quartet in detail.

According to Francis Frank Anscombe, it comprises of 4 datasets which almost same properties that of a Linear Regression, but when you proceed to graph them up for visualizing, they are seemed to be totally different. Frank showed the importance of Visualizing the data before proceeding to apply various techniques on them which would make us create a better machine learning model.



3. What is Pearson's R?

Pearson's R also known as Person Correlation Coefficient. It is used to see how sets of data are correlated to each other. It helps in showing the linear relationship between two sets of data. The value of Pearson's R would like between -1 and 1.

1. Where -1 would mean that the data is linear with negative slope.
2. And 1 would mean that the data is linear with positive slope.
3. Whereas 0 would mean that there is no linear relationship between the sets of the data.

The formula to calculate the Pearson's R is:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2(y_i - \overline{y})^2}}$$

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 to be performed after the train and test split. In this the data is scaled down to values between 0 and 1 from large continuous variable.
Scaling is to be done so that it does not train the model with respect to its high continuous values that would affect the p-value, VIF, R-Squared, Adjusted R-Squared and so on. If the units of all the features are not scaled down to the same range, then that would be a bad model.

1. Normalized Scaling: Normalized scaling can also be called as MinMax Scaler, it is used to bring the values down to between 0 and 1. This can be used by importing sklearn.preprocessing. It basically finds the min and max of the column and does the calculation, its formula is:

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

2. Standardized Scaling: This technique replaces the values by their Z scores. It scales its data into Normal Distribution which has its mean as 0 and standard deviation as 1. Its formula is:

$$x = \frac{x - \text{mean}(x)}{\text{sd}(x)}$$

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

If a variable turns out to have perfect correlation with another variable where both are independent then that would result the R^2 to be 1. And as the formula goes:

$$VIF = \frac{1}{(1 - R^2)}$$

That would put the R^2 as 1 and the denominator would turn to be 0 and as we know that 1 divided by 0 would turn out to be infinite. That would also mean that the one independent variable is able to explain the other independent variable very well. So it is always advised to drop such variable.

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

Q-Q plot is called as Quantile-Quantile Plot where we check the two sets of data and see if they come from the same distribution such as Normal Distribution, or other distribution. We can use Q-Q plot to check if the test and train data come from the same dataset with similar distribution. It can be used to check following:

1. Two sets of data come from populations with similar distributions
2. The two sets of data have common location and scale.
3. The two sets of data have similar distributional shapes.
4. The two sets of data have similar tail behavior.