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?

Ans:

1. Highest number of bookings are happening in fall(season 3). This can be used as a predictor
2. Bookings tend to happen more when there is clear weather
3. More bookings happening in months 5,6,7,8 & 9 with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.

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

Ans: drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column.

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

   Ans: temp and atemp are the two numerical variable which have high correlation with the target variable. They have 0.63 positive correlation with cnt.

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

Ans: I validated the assumptions of linear regression by plotting a histogram of the error terms of actual y values and predicted y values. The histogram shows that the error terms are normally distributed.

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

Ans: A coefficient value of '0.5636' indicated that a unit increase in temp variable increases the bike hire numbers by 0.5636 units. Weather Situation 3 (weathersit_3) - A coefficient value of '-0.3070' indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.3070 units. Year (yr) - A coefficient value of '0.2308' indicated that a unit increase in yr variable increases the bike hire numbers by 0.2308 units. So, it's suggested to consider these variables utmost importance while planning, to achive maximum Booking

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans:

Linear regression is a type of supervised machine learning algorithm that is used for the predication of numeric values. Linear regression is the most basic form of regression analysis. Regression is most commonly used predictive analysis model.

Linear regression is based on the popular equation “**y=mx+c**”.

It assumes that there is linear relationship between dependant variable(y) and predictor variable(s)/independent variables X. In regression, We calculate the best fit line which describes the relationship between independent variable and dependant variable.

Regression is performed when dependant variable is of continuous type. And predictor variables can be of any type like continuous, categorical etc. Regression method tries to find the best fit line which shows the relationship between the dependant and independent variable with least error.

I regression, the output/dependant variable is the function of an independent variable and the coefficient and the error term.

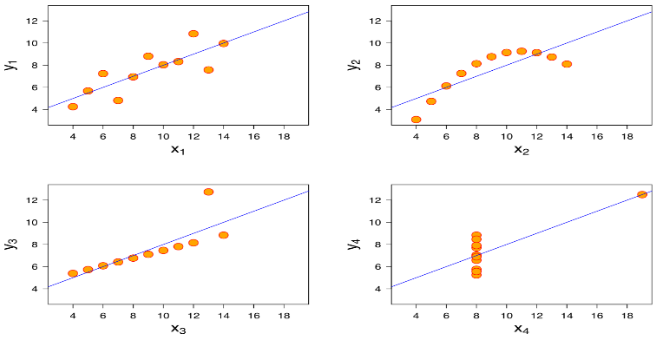
Regression is broadly divided into simple linear regression and multiple linear regression.

1. Simple Linear Regression: SLR is used when dependant variable can be predicted using one independent variable.
2. Multiple Linear Regression: MLR is used when the dependant variable is predicted using multiple independent variables.

2. Explain Anscombe’s quartet in detail.

ANS:

Anscombe’s quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to emphasize both the importance of graphing dat before analysing it and the effect of outliers influential observations on statistical properties.



- The first scatter plot appears to be a simple linear relationship.
- The second graph(top right) is not distributed normally distributed, While there is a relation between them is not linear.
- In the third graph the distribution is linear, but should have a different regression line. The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

3.What is pearson's R?

Ans:

Pearsons R is a numerical summary of the strength of the linear association between the variables. It's value ranges between -1 to +1. It shows the linear relationship between two sets of data. In simple terms, it tells us can we draw a line graph to represent the data?

r=1 means the data is perfectly linear with a positive slopw.

r=-1 means the data is perfectly linear with negative slope.

r=0 means there is no linear association.

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

Ans:

Feature scalling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data processing stage to deal with varying values in the dataset. 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, irrespective of the units of the values.

- Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbours and neural networks.
- Standardization on the other hand can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, Unlike normalization, Standardization does not have a bounding range. SO, even if you have outliers in your data they will not be affected by standardization.

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

Ans: VIF- the variable inflation factor- the VIF gives how much the variance of the coefficient estimate is being inflated by collinearity(VIF)= 1/(1-R_1^2). If there is perfect correlation, then VIF = Infinity. Where R-1 is the R-square value of thet independent variable which we want to check how well this independent variable is explained well by other independent variables. IF that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it’s R-SQUARED VALUE WILL BE EQUAL TO 1. So , VIF=1/(1-1) which is 1/0 resulting in infinity.

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

Ans: A q-q plot is a plot of the quantities of the first data set against the quantities of the second data set. It is used to compare the shapes of distribution. A Q-Q plot is a scattered plot created by plotting two sets of quantities against one another. If both sets of quantities came from the same distribution, we should use the points forming a line that’s roughly straight.

The Q-Q plot is used to answer the following questions:

- Do two data sets come from populations with a common distribution?
- Do tow data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar behaviour?