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)

The categorical variable in the dataset were workingday,season,weathersit,holiday,mnth,yr and weekday.These were visualized using a boxplot .These variables had the following effect on our dependant variable:-

1. **Season –** From the graph spring season had least booking for the bike whereas fall had maximum value of bookings and summer and winter had intermediate value of bookings.
2. **Weathersit** – When weather is Clear and Mist & cloudy then there are high chances of booking the bike but if there is light snow its less number of bookings but if weather is bad like heavy no bookings are taken place so weather sit might impact a lot ’.
3. **Holiday -** rentals reduced during holiday.
4. **Month -** September saw highest no of rentals while January saw least. This observation is on par with the observation made in weather sit. The weather situation in January is usually harsh .
5. **Year**- The number of rentals in 2019 was more than 2018

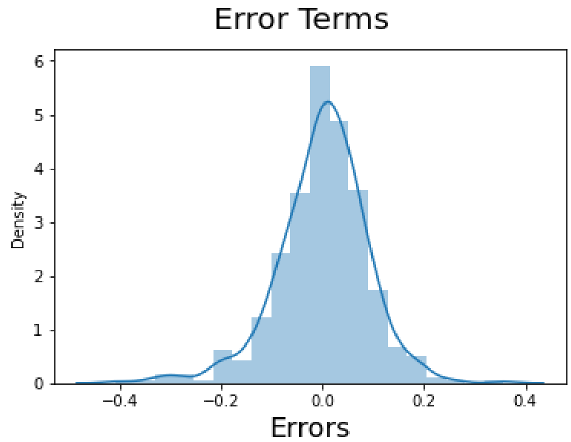
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

If we don't drop the first column then the dummy variables will be highly correlated. This may affect some models and the effect is so stronger when the cardinality is smaller. if we have all dummy variables into the model then it leads to Multicollinearity betweenthe dummy variables To reduce the Multicollinearity we need to drop one column

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

“temp” and “atemp” are the two numerical variables that are highly correlated with the target variable (cnt)

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



Residual Analysis of the train data will validate the assumption of linear regression

If the error terms are normally distributed and centered around 0 (which is in fact, one of the major assumptions of linear regression)

This can be done by validate this assumption about residuals by plotting a distplot of residuals and see if residuals are following normal distribution or not.The above diagram shows that the residuals are distributed about mean = 0 and normally distributed

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

The top 3 features are

1.**temp** - coefficient: 0.464

2. **year**(yr) - coefficient : 0.2311

3.**weathersit_Light Snow & Rain** - coefficient - 0.245

General Subjective Questions

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

Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values. Linear Regression is the most basic form of regression analysis.Regression is the most commonly used predictive analysis model.
Linear regression is based on the popular equation **"y = mx + c".**

It assumes that there is a linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x). In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable.

Regression is performed when the dependent variable is of continuous data type and Predictors or independent variables could be of any data type like continuous, nominal/categorical etc. Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error.

In regression, the output/dependent 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 the dependent variable is predicted using only **one** independent variable

2. **Multiple Linear Regression :MLR i**s used when the dependent variable is predicted using multiple independent variables.

The equation for MLR will be:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots ,$$

β1 = coefficient for X1 variable

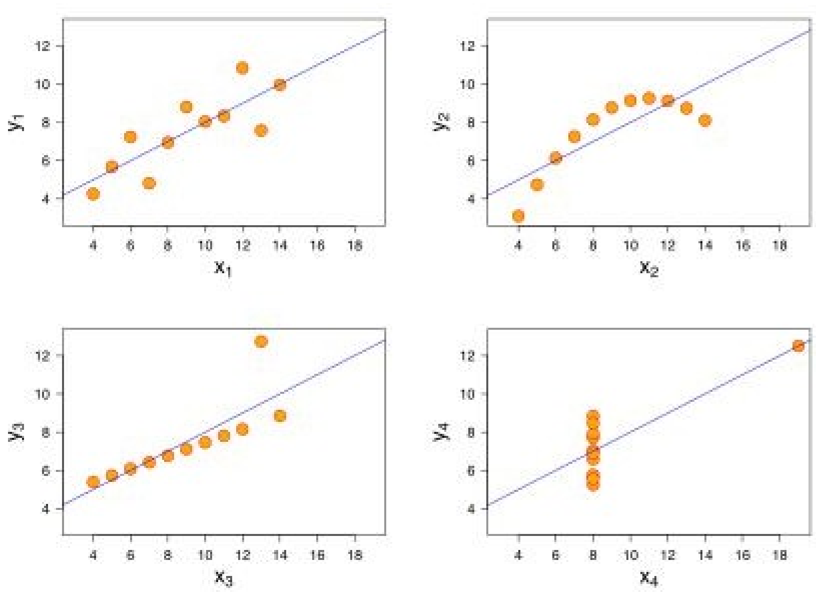
β2 = coefficient for X2 variable

β3 = coefficient for X3 variable and so on…

β0 is the intercept (constant term).

2. Explain Anscombe's quartet in detail.

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 data before analyzing it and the effect of outliers and other influential observations on



statistical properties

- The first scatter plot (top left) appears to be a simple linear relationship.
- The second graph (top right) is not distributed normally; while there is a relation between them, it's not linear.
- In the third graph (bottom left), 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 (bottom right) 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? (3 marks)

Pearson's r is a numerical summary of the strength of the linear association between the variables. It 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 slope
r = -1 means the data is perfectly linear with a 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)

Feature **scaling** is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data preprocessing 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 doesnot follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors 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? (3 marks)

VIF - the variance inflation factor -The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity. (VIF) $= \frac{1}{(1-R\_i^2)}$. If there is perfect correlation, then VIF = infinity. Where R-1 is the R-square value of that 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 gives VIF = 1/0 which results in "infinity"

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

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. It is used to compare the shapes of distributions. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. if
both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.
The q-q plot is used to answer the following questions:

1.Do two data sets come from populations with a common distribution?
2.Do two data sets have common location and scale?
3.Do two data sets have similar distributional shapes?
4.Do two data sets have similar tail behavior?