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)

From our boxplots of categorical variables, we can infer the following

1) With box plot of season, we can see relatively the usage of bikes is less in the season represented by 1 which is spring as well highest in season 3 which is fall
2) With box plot of weathersit, we can see the usage is very in situation 3 which says there is light snow or rain.
3) With box plot of mnth, we can see the usage being less during the early part of the year or the later end of the year coinciding with the holiday season

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

During dummy variable creation, we use drop_first=True. This changes the number of dummy variables from n to n-1 and thereby giving efficiency during the calculations. For our case, though this is miniscule, when the data is huge, this will be a big efficiency value, which becomes more pronounced when we redo the iterations with different variables.

From our example, we can see that a variable such as season has values spring, summer, fall & winter. These can be represented by 100, 010, 001 with n-1 variable counts and the last value 000 where none is having 1 providing the last categorical variable, which is implied.

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

From the pair-plot of the numerical variables, we can see that the highest correlation is between atemp and temp variables

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

The assumptions of Linear Regression are validated as below

1) Error terms is plotted using distplot to check if it follows the normal distributed which in our case, we find it is true
2) Error terms should have constant variance or homoscedastic. This is seen by plotting a scatter plot between predicted values and the target variable values provided in training set which provides truth to our assumption
3) Error terms are independent of each other. This can be seen by plotting residuals with target variable values from the training set.

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

The top 3 features are:

1) Temp
2) Weathersit_LightRain
3) Yr

General Subjective Questions

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

Linear Regression Algorithm is a basic regression algorithm used in machine learning to check if there is a linear relationship between independent and dependent variables. This is widely used for prediction and projection models.

There are 2 distinct variations namely Simple Linear Regression and Multiple Linear Regression which basically are models derived by using only one variable at a time against target variable or using multiple independent variables against target variable respectively.

The model assumes the below statements:

1) There is a linear relationship between X (independent variable) & y (dependent variable
2) Error terms are normally distributed with mean zero
3) Error terms are independent of each other
4) Error terms have constant variance (homoscedasticity)

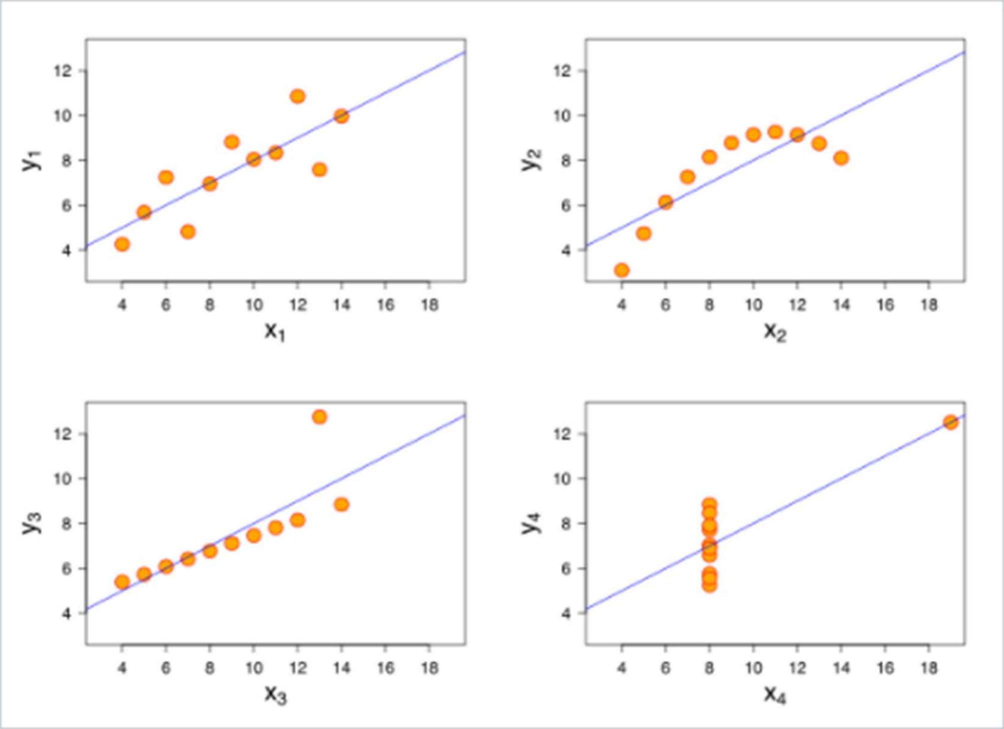
Apart from above, for multiple linear regression models, we need to achieve optimal feature selection with minimal correlation between independent variables. For arriving at this, we use different criteria such as adjusted R-squared which penalises based on the number of variables used, VIF or Variance Inflation Factor which provides a numerical way of gathering the pairwise correlations with combination of other variables

Again since we are dealing with multiple variables, we need to scale the variables so that all variables are brought in the same range as well as converting the categorical variables using encoding like one hot encoding.

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

Anscombe’s quartet is a set of 4 data sets which have identical statistics but are very different graphs when plotted. This tells us the importance of analysing the data visually before starting with the model creation. This is profound in case of linear regression, if we develop a model with significant outliers or those data which follow a pattern.

An example depiction of the Anscombe’s quartet is given below:



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

Image credit: https://en.wikipedia.org/wiki/Anscombe%27s_quartet

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

Pearson's R is a statistic used to measure the linear correlation between two sets of data. It is also known by PPMCC (Pearson Product Moment correlation coefficient) or bivariate correlation. It is the ratio of covariance of two variables and the product of their standard deviations. Its value lies within -1 and 1.

Ref:

https://www.analyticssteps.com/blogs/pearsons-correlation-coefficient-r-in-statistics

https://en.wikipedia.org/wiki/Pearson_correlation_coefficient#:~:text=In%20statistics%2C%20the%20Pearson%20correlation,between%20two%20sets%20of%20data

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

Scaling is the process of converting the data to a range which is easily interpretable.

Scaling is performed for the following reasons:

1) All the data is converted to the same range so that model is correctly interpreting the data in case of multiple linear regression
2) This provides a way for faster convergence for gradient descent methods

There are 2 important ways we can scale which is given below:

Normalized Scaling – Also called minmax scaling which is used to bring all the data points within range of 0 and 1. Formula is x = (x – min(x))/(max(x) – min(x))

Standardized Scaling – This is used to bring the data into a standard normal distribution with mean of 0 and standard deviation of 1. This means that we are scaling it into data which is centred around mean of 0. Formula is x = (x – μ) / σ

We can use sklearn preprocessing library with the function for doing the scaling.

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

In our assignment case, we have observed that the VIF for few variables were infinite. We can see that in that case, the variables showing infinite VIF are derived dummy variables from a singular categorical variable. This implies that there is a huge correlation between these variables which is mathematically shown by VIF as infinite.

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

Q-Q plot stands for Quantile-Quantile plot which is a plot that can be used to check if the data from two different sets come from the same distributions or not. This is usually used when we get training and test data sets separately and we can use this Q-Q plot to confirm if they come from the populations with the same distributions.

From the plot we can observe the below:

1) If all the point of quantiles lie around the 45° line, then they can be said to be from the same distribution
2) We can check if the scale from the 2 data sets are similar or not
3) We can check for outliers from the graph as well