**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)**

**Ans :**  i. box plots shows that Wednesday- Saturday there is more booking

ii. three is more booking during 2019 than 2018

iii . Spring has lower booking

**i**v. clear and misty weather has more booking

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

It avoids the extra column creation during dummy variable creation. So lesser correlation

**3. Looking at the pair-plot among the numerical variables, which one has the highest correlation**

**with the target variable? (1 mark)**

Temp variable has good correlation with target

**4. How did you validate the assumptions of Linear Regression after building the model on the**

**training set? (3 marks)**

Validation is based on multicollinearity, Linear relationship, homoscedacity and residuals

**5. Based on the final model, which are the top 3 features contributing significantly towards**

**explaining the demand of the shared bikes? (2 marks)**

temp,spring and summer

**General Subjective Questions**

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

It is a statistical model that analyses the relationship between a dependent variable with given set of independent variables. 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

It is represented as y= C+mx

Y is the dependent variable we are trying to predict.

X is the independent variable we are using to make predictions.

M tanO/ slope or gradient

C is a constant, known as the Y-intercept.

Types of LR based on the gradient

1. +ve LR
2. –Ve LR

Types of LR based on number of variable

1. simple LR-model with 1 independent variable

2. Multiple LR- model with multiple independent variable

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

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.It was developed by Francis Anscombe. It will help in explain the portance of graphing data before analyzing it and the effect of outliers on statistical properties.

To explain it further, the mean, sum and average of data set may be same. But when we plot the graphs of these variables/data sets the graphs will show different layout.

1. **What is Pearson’s R? (3 marks)**

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative. 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

**4. What is scaling? Why is scaling performed? What is the difference between normalized scaling**

**and standardized scaling? (3 marks)**

Scaling is the Pre-Processing steps of data, which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Collected data may set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Differences

1.

In Normalisation Minimum and maximum value of features are used for scaling

In Standardisation Mean and standard deviation is used for scaling.

2.

Normalisation is used when features are of different scales.

Standardisation is used when we want to ensure zero mean and unit standard deviation.

3

Normalisation Scales values between [0, 1] or [-1, 1].

Standardisation is not bounded to a certain range.

4.

Normalisation is really affected by outliers.

Standardisation is much less affected by outliers.

5.Scikit-Learn

MinMaxScaler transformer for Normalization.

StandardScaler transformer for standardization.

6.

Normalisation squishes the n-dimensional data into an n-dimensional unit hypercube.

standardization translates the data to the mean vector of original data to the origin and squishes or expands.

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

**(3 marks)**

In a perfect correlation the VIF is 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.

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

**(3 mark**

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

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.