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)

-There are 7 categorical variable Season : The top most season for riders is fall and followed by summer and winter. Season is definatly good predictor of independent variable year: The number of bookings was high in 2019 than 2018 Month: Most of the bookings are done from May to Oct. This shows month has some trend in bookings Holiday: Bookings are mostly done on weekends or Holidays Weekday: Bookings are mostly on all of the weekdays weathersit: The bookings are quite more on weather A ie cloudy partly cloudy inshort when weather is nice.

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

Value	Indicator Variable
Gender	Female
Male	0
Female	1

If a columns has n unique values then we need n-1 dummy variables because it is understood the last combination would be for that 1 variable .

Get_dummies method makes n dummy variables for n values but since we need only n-1 hence we drop one dummy variable

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 have correlation of 0.99

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

By plotting the histogram of residuals/ error terms if the plot is normally distributed then we can say the assumptions made are valid

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

Top 3 predictors of the model 4 are

Temp : A unit increase in temp increase the bike hire number by 0.432 yr_1(2019) : A unit increase in yr increase the bike hire number by 0.2333 weathersit_c : A unit increase in weathersit_c decrease the bike hire number by 0.3120

General Subjective Questions

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

Linear regression is an algorithm that builds various models to predict target variable(y) dependent on other variables (X). The target variable should be numerical and continuous.

The regression model builds a relationship between X and y

Y = Bo + B1x

Where Bo is intercept and B1 is coefficient of x

If the independent variables are more then equation would

Y = Bo + B1X1 + B2X2….. BnXn

After finding the values Bo and B1 we get the best fit line. Now the model predicts the value so that the difference between the predicted and true value is minimum.

Hence B0 and B1 keep on updating till we get the minimum difference

RSS (Residual sum square) = ∑(yi – B0 -B1Xi)^2 where i = 0 to n

R^2 = 1 – RSS/TSS

TSS – Total variance in y

RSS should be minimum R^2 should be maximum

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

Anscombe's quartet has 4 datasets which have statistical properties that is mean variance R2 and Best fit line but on plotting all the 4 data sets in graph. The graphs are totally different hence it emphasizes on graphing the data before analysing and also the effects of outliers on statistical properties

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

Pearson's R measures the strength of linear relationship between two variables

Pearson's R is always between -1 to 1

When r =1 then that is perfect linear relationship

When r=-1 then there is perfect negative linear relationship

The formula to calculate Pearson's R =

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

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

Scaling is done to bring all the independent variables in particular range so that one significant number will not impact the model because of its larger magnitude. It helps to speed up the calculation.

Scaling is performed to lower the effect of higher magnitudes because scaling only considers the magnitude not the unit.

Scaling only changes the coefficients not the other parameters like P values , F statistics, R squared etc

Min-Max scaling is x = x – min(x)/max(x)-min(x)

Used when no outliers are there and does not follow gaussian distribution and this makes values in range of 0 to 1.

Standarised scaling = (x – mean(x))/sd(x)

Standard scaling centers the values around mean with a unit standard deviation.

This is usually done when data is gaussian distributed and outliers are there.

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

VIF is Variance inflation factor which measures the multicollinearity between the dependent variables.

So the larger value of VIF indicates the larger collinearity. So if VIF = infinite then it means it is the perfect correlation as

VIF = 1/(1-R^2)

For perfect R^2 is 1 hence VIF is 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 is quantiles quantiles plot ie normal quantiles on x axis and Data quantiles on Y- axis.

On x-axis we plot the normal distribution and on y the data Quantile means percentage of data falls under that quantile.QQ plot is used to check if data is normally distributed. After plotting. If the points form a straight line with an angle of 45 degree then data is normally distributed otherwise not

It helps in finding out if error terms are normally distributed.