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)

- We can observe seasons, weather, yr and mnth variables has some good insights
- Summer and fall has more rentals compared to fall and spring
- When weather is clear more rentals are observed, low rentals during snow
- During summer months we see rise in rentals.
- 2019 has a rise in rentals compared to 2018
- working day and weekday does not provide useful info on rentals

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

- Setting drop_first as True will transform k distinct categories to k-1 columns, since k-1 columns are enough to explain k categories.
- For example, in bike rentals data we have seasons column with 4 categories, they are spring, summer, fall, winter. After encoding if we remove spring, we still know it is spring if all other categories are zero.

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

- Temperature has the highest linear correlation with dependant variable

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

- Observe the distribution plot of train error, it should be normally distributed with mean at zero.
- Pairplots with numeric variables and correlation matrix tell us if there is a linear relation between variables and independent variable.
- Plotting error and the dependant variable to check if there is pattern to show the error terms are dependent on each other.

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

- **Temperature**, if temperature increases the bike rentals increase.
- **Snow**, reduces the number of bike rentals
- Higher **windspeed** means low bike rentals.

General Subjective Questions

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

- Linear regression Is a supervised ML algorithm, this models find the linear relationship between the independent variables and the dependent variables. This algorithm can be applied only on the regression problems.
- This model fits the best fit line, the equation of the line is used to predict the dependent variable y, where y = mx +c where x is the independent variable and c is the intercept which is zero if the line passes through the origin, for simple linear regression.
- The best line is fit by reducing the squared error (y_actual – y_predicted)^2

- For multiple linear regression the equation y = m1x1+ m2x2 + m3x3+ …….. +c, the change of y for every unit change of x1 is m1 considering all other variables does not change.

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

- Anscombe's quarter is when we have identical statistical description of different datasets but when plotted they are completely different.
- This tells us the importance of visualising the data before modelling. Since linear regression model is useless when applied on non-linear data.

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

- Pearson's R measures the strength of linear relationship between two variables.
- This value lies between -1 and 1
- -1 to 0 means negative correlation, when one increases the other decreases. 0 to 1 mean positive correlation, when one increases the other variable increases. The number also indicates how strong the relationship is.

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

- Scaling is transforming the numeric data to fit within a range.
- Scaling is required because each variable will have its own range and units which cannot be interpreted by the ML algorithm. It also helps the model to converge faster.
- Normalised scaling is done using min and max values, mostly lies between o and 1. This method is affected by outliers hence, not recommended for variables with outliers. Calculated by (x-Xmin)/(Xmax- Xmin)
- Standardised scaling is done using mean and standard deviation. This expands or squishes around mean if standard deviation is 1, this tells us the deviation of data point from the mean. This is not affected by outliers. Calculated by x- mean/std.

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

- This happens when there is a perfect match between the variables or perfect multi-collinearity. That means that some variables are able to perfectly predict this variable with infinity VIF.
- This leads to R2=1 and 1/(1-R2) as infinity

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 quantile -quantile plot, which plots the quantiles of two different data sets to see if both of them belong to same distribution or not. One quantile from first data set on x axis and the other one on y-axis.
- In linear regression, Q-Q plot helps us to check for the assumption if the error terms are normally distributed, if they follow a straight line at 45 degree inclination then the data sets are normally distributed.