**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 variables have less impact rather then continuous variable have more impact on the model score
- Before fitting to the model we converted categorical columns into one hot encoding (dummpy variable) which creates n columns where n is categorical classes.
- If we see in notebook we find that casual, registered and temp variables are continuous variables and they are impacting more that 96% on model accuracy

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

- It is helps in reducing the extra column which we created while masking dummy variables. It reduces the correlation created among dummy variables.
- Dummpy variable which creates n -1 columns where n is categorical classes and we do drop_first = True.

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

- Temp and registered variable looking highest correlation with target variable cnt.

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

- Assumption 1 — Independence of observations(Remove highly correlated variables from the dataframre i'e temp and atemp both are +ve correlated, so removed the atemp from dataset)
- Assumption 2 — No Hidden or Missing Variables (there are no missing values in dataset)
- Assumption 3 — Linear relationship (taking linear(continous) variables first to the model to check the relationship)
- Assumption 4 — Normality of the residuals (Checking the residual sum of the model across all the features )

- Assumption 5 — No or little Multicollinearity (Remove multicollinearity variables from dataset)
- Assumption 6 — Homoscedasticity (After training the dataset check the Homoscedasticity behaviour of model)
- Assumption 7 — All independent variables are uncorrelated with the error term(Checked all the variables have same error distibution and model is not bias)

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

- We can see the temp variable have the highest impact on the model, according to that we can estimate bike sharing hikes bases on high temp or low temp
- Season also impact the major role on the model accuracy, because according to the season people like to rent the bike.
- Weekends and holidays are one more useful aspect for taking bike renting

**General Subjective Questions**

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

    - Linear regression is a method for determining the best linear relationship between two variables $X$ and $Y$. If variables $X$ and $Y$ are uncorrelated, it is pointless embarking upon linear regression. However, if a reasonable degree of correlation exists between $X$ and $Y$ then linear regression may be a useful means to describe the relationship between the two variables.
    - The straight line relating $X$ and $Y$ is $y = mx + c$, where $m$ and $c$ are the gradient and constant values (to be determined) defining the straight line.

Thus, $y(x_i)$ - $y_i$ is the difference between the line and data point $i$ . Taking all the data points, we seek values of $m$ and $c$ that minimize the squared difference $SD$.

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

    - Anscombe's comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points.

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

    - the Pearson correlation coefficient (PCC, pronounced /ˈpɪərsən/) — also known as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), the bivariate correlation, or colloquially simply as the correlation coefficient is a measure of linear correlation between two sets of data.

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

    - Scaling is also know as min-max sacler . It is youes to bring the all the variables range in 0 and 1.
    - Standradization changes the values by their Z scores. It brings all of the data into a standard scale, normal distribution which has means zero and standard deviation one.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
    - An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
    - Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution