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)

- *People rent bikes most often during fall and summer; they prefer it least during spring.*
- *Bikes have high demand when the skies are clear or partly cloudy & least wanted during light or heavy shower/snow.*
- *Demand for bikes has increased considerably from the year 2018 to 2019.*
- *People use bikes most often from June to September, peaking at September.*
- *Average count of bikes are similar across weekdays with increased variations on Saturday & Wednesday.*
- *People prefer to use bikes slightly higher during working days.*

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

- *It's easy to identify the dropped column just by looking at the values of other dummy encoded columns. All zeroes indicate it's the dropped column.*
- *Dropping this redundant column created during dummy variable creation further reduces the correlations among dummy variables.*
- *E.g. if we have 3 types of values in Categorical column, if it's 0 for A & B we could discern that it's C.*

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

- *Temp & atemp (i.e. feels-like temperature) share the highest correlation with count of rental bikes.*

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

- ***Linearity**: The relationship between X and Y must be linear. A pairplot between count and predictor variables showed they shared a linear relationship.*
- ***Normality of errors**: The residuals must be approximately normally distributed with a mean of 0. A distplot drawn using y-actual & y-predicted revealed a normal distribution curve.*
- ***Independence of errors**: There is no relationship between the residuals and the Y-predicted. A residplot drawn against y-actual & y-predicted showed no discernible pattern.*
- ***Equal variances (homoscedastic)**: There is no relationship between the residuals and the Y-predicted. A residplot drawn against y-actual & y-predicted showed no discernible pattern.*

- *Multiple regression assumes that the independent variables are not highly correlated with each other. This assumption is tested using Variance Inflation Factor (VIF) values.*

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

- *atemp (or feels-like temperature): Increase in temperature increased the demand for bikes.*
- *Year: Every year goes by there are more users for rental bikes.*
- *Winter: People want to use bikes mostly during winter.*

General Subjective Questions

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

- *Linear regression is the method of finding the best linear relationship within the independent variables and dependent variables.*
- *The algorithm uses the best fitting line to map the association between independent variables with dependent variable.*
    - *The best-fit line is found by minimising the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot.*
    - *The strength of the linear regression model can be assessed using 2 metrics:*
        - *R2 or Coefficient of Determination*
        - *Residual Standard Error (RSE)*
- *There are 2 types of linear regression algorithms*
    - *Simple Linear Regression – Single independent variable is used.*
        - $Y = \beta 0 + \beta 1X$ *is the line equation used for SLR.*
    - *Multiple Linear Regression – Multiple independent variables are used.*
        - $Y = \beta 0 + \beta 1X1 + \cdots + \beta pXp + \in$ *is the line equation for MLR.*
    - $\beta 0$ = *value of the Y when X = 0 (Y intercept)*
    - $\beta 1, \beta 2, \ldots, \beta p$ = *Slope or the gradient*

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

- ***Anscombe's Quartet*** *can be defined as a group of four data sets which are* ***nearly identical in simple descriptive statistics****, but there are some peculiarities in the dataset that* ***fools the regression model*** *if built. They have very different distributions and* ***appear differently*** *when plotted on scatter plots.*
- *It stresses on* ***the importance*** *of* ***plotting the graphs*** *before analyzing and model building, and the effect of other* ***observations on statistical properties****.*

- *There are these four data set plots which have nearly **same statistical observations**, which provides same statistical information that involves **variance**, and **mean** of all x,y points in all four datasets.*
- *This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the **data with linear relationships** and is incapable of handling any other kind of datasets.*

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

The Pearson's R (also known as Pearson's correlation coefficients) measures the strength between the different variables and the relation with each other. The Pearon's R returns values between -1 and 1. The interpretation of the coefficients are :

- *-1 coefficient indicates strong inversely proportional relationship.*
- *0 coefficient indicates no relationship.*
- *1 coefficient indicates strong proportional relationship.*

- *For the Pearson r correlation, both variables should be normally distributed.*
- *There should be no significant outliers.*
- *Each variable should be continuous i.e. interval or ratios for example weight, time, height, age etc. If one or both of the variables are ordinal in measurement, then a Spearman correlation could be conducted instead.*
- *The two variables have a linear relationship. Scatter plots will help you tell whether the variables have a linear relationship.*
- *The observations are paired observations.*
- *Homoscedasticity is a situation in which the error term (that is, the "noise" or random disturbance in the relationship between the independent variables and the dependent variable) is the same across all values of the independent variables.*

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

- What - The scaling is the data preparation step for regression model. The scaling normalizes these varied datatypes to a particular data range.
- Why – Most of the times the feature data is collected at public domains where the interpretation of variables and units of those variables are kept open collect as much as possible. This results in to the high variance in units and ranges of data. If scaling is not done on these data sets, then the chances of processing the data without the appropriate unit conversion are high. Also the higher the range then higher the

possibility that the coefficients are impaired to compare the dependent variable variance. The scaling only affects the coefficients. The prediction and precision of prediction stays unaffected after scaling.

- Normalization/Min-Max scaling – The Min max scaling normalizes the data within the range of 0 and 1. The Min max scaling helps to normalize the outliers as well.
- Standardization converges all the data points into a standard normal distribution where mean is 0 and standard deviation is 1.

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

The VIF formula clearly signifies when the VIF will be infinite. If the R2 is 1 then the VIF is infinite. The
reason for R2 to be 1 is that there is a perfect correlation between 2 independent variables.

$VIF = 1/1\text{-}R^2$

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

Q-Q plots are the quantile-quantile plots. It is a graphical tool to assess the 2 data sets are from common distribution. The theoretical distributions could be of type normal, exponential or uniform.
The Q-Q plots are useful in the linear regression to identify the train data set and test data set are from the populations with same distributions. This is another method to check the normal distribution of the data sets in a straight line with patterns explained below:

Interpretations

- Similar distribution: If all the data points of quantile are lying around the straight line at an angle of 45 degree from x-axis.
- Y values < X values: If y-values quantiles are lower than x-values quantiles.
- X values < Y values: If x-values quantiles are lower than y-values quantiles.
- Different distributions – If all the data points are lying away from the straight line.

Advantages

- Distribution aspects like loc, scale shifts, symmetry changes and the outliers all can be identified from the single plot.
- The plot has a provision to mention the sample size as well.