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)

- Fall is having high range of users who rented the bikes.
- 2019 to 2020 is having high count of users that rented a vehicle compared to past year.
- June, July, September and October are having high median and total number of users are also high in September and October.
- If it is holiday there are more rentals booked but coming to median it is high in working days.
- On an average Thursday & Friday are having greater median compared to all other
- days, Friday has more range of users.
- Weather is bad at that time the users decreased and increased when it is moderate and
- good climate.

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

- Dummies are used to represent subgroups present in the column of your dataset
- Use dummies as 0 & 1 on the use case
- Used in the data manipulation

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

- Temperature - the target variable ‘count’ with 0.63 correlation value

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

- no multicollinearity in the dataset
- As the value of independent variable increases the dependent variable should increase.
- If the value of independent variable decreases dependent variable should also decrease.
- Homoscedasticity: It means that the residuals should have constant variance irrespective of dependent variable throughput the dataset.

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

- season_winter 1153.9938
- mnth_sep 1002.4979
- temp 4771.8712

General Subjective Questions

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

- Linear Regression – ML algorithm used to predict the value of a variable based on the value of another variables. / A simple linear regression model attempts to explain the relationship between a dependent variable and an independent one using a straight line.
- The variable you want to predict is called the dependent variable (output variables).
- The variable you are using to predict the other variable's value is called the independent variable (predictor variable). / regressorA regressor is also referred to as: An explanatory

variable / An independent variable / A manipulated variable / A predictor variable / A feature

- 2 types of linear regression models: Simple Linear Regression, Multiple Linear Regression
- Simple Linear Regression – only 1 independent variable
- Multiple Linear Regression – more than 1 independent variable
- Equation of a Straight Line: y = mx + c
- 2 parameters m and c
- m signifies how strong is the relation b/w x and y
- positive slope = direct relationship, negative slope = inverse relationship
- c signifies value of y when x=0
- The slope of any straight line can be calculated by $(y_2 - y_1)/(x_2 - x_1)$, where (x_1, y_1) and (x_2, y_2) are any two points through which the given line passes.
- Standard Notation in Regression: y = β0 + β1 x
- β_0 is the intercept, β_1 is the slope
- In regression, a best fit line is a line that fits the given scatter plot in the best way.

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

- Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed.
- It can be defined as group of four datasets which are identical in descriptive statistics.
- When they were plotted on scatter plots the distributions are very strangely different.
- Francis Anscombe constructed this quartet in 1973 to demonstrate the importance of plotting the graphs.

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

- Pearson's Correlation Coefficient: It is the test statistics that measures the statistical relationship, or association, between two continuous variables.
- It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance.
- It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.
- 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)

- Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.
- If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

- Differences

Normalized scaling	Standardized scaling
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1]	It is not bounded to a certain range
It is really affected by outliers.	It is much less affected by outliers.

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

(3 marks)

- If there is perfect correlation, then VIF = 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 marks)

- The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.
- Use of Q-Q plot: A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value.
- Importance of Q-Q plot: When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. I