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?**

- Variables like ‘holiday’, ‘windspeed’, ‘weathersit variables taking values 2 & 3’ have negative impact on the dependent variable
- Variables like ‘season column values’ have positive impact on the dependent variable
- Variable ‘yr’ has positive relationship which means in comparison to 2018, in 2019 values of the dependent variable has increased

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

It’s important; otherwise we will introduce redundant information in the dataset. This will also result in high correlation between dummy variables resulting in the coefficient prediction unstable.

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

From my final model it’s variable ‘yr’. However considering all variables it would be ‘atemp’.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

- Validated the normal distribution criteria for the residuals by making a distribution plot for errors.
- For existence of autocorrelation I checked the Durbin-Watson value from the summary table.

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

- Weathersit
- Situation
- yr

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a supervised machine learning method that is used to find a linear equation that best describes the correlation of the explanatory variables with the dependent variable. This is achieved by fitting a line to the data using least squares. The line tries to minimize the sum of the squares of the residuals. The residual is the distance between the line and the actual value of the explanatory variable. Finding the line of best fit is an iterative process.

The following is an example of a resulting linear regression equation:

$y = b_0 + b_1*x_1 + b_2*x_2 + ...$

In the example above, y is the dependent variable, and x_1, x_2, and so on, are the explanatory variables. The coefficients (b_1, b_2, and so on) explain the correlation of the explanatory variables with the dependent variable. The sign of the coefficients (+/-) designates whether the variable is positively or negatively correlated. b_0 is the intercept that indicates the value of the dependent variable assuming all explanatory variables are 0.

A linear regression model helps in predicting the value of a dependent variable, and it can also help explain how accurate the prediction is. This is denoted by the R-squared and p-value values. The R-squared value indicates how much of the variation in the dependent variable can be explained by the explanatory variable and the p-value explains how reliable that explanation is. The R-squared values range between 0 and 1. A value of 0.8 means that the explanatory variable can explain 80 percent of the variation in the observed values of the dependent variable. A value of 1 means that a perfect prediction can be made, which is rare in practice. A value of 0 means the explanatory variable doesn't help at all in predicting the dependent variable. Using a p-value, we can test whether the explanatory variable's effect on the dependent variable is significantly different from 0.

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet is the modal example to demonstrate the importance of data visualization which was developed by the statistician Francis Anscombe in 1973 to signify importance of plotting data before we analyze it and build our model.

This comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (*x*, *y*) points.

These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another.

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

- **Data Set 1:** fits the linear regression model pretty well.
- **Data Set 2:** cannot fit the linear regression model because the data is non-linear.
- **Data Set 3:** shows the outliers involved in the data set, which cannot be handled by the linear regression model.
- **Data Set 4:** shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

3. What is Pearson's R?

In statistics, the **Pearson correlation coefficient** (**PCC**) is a correlation coefficient that measures linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationships or correlations

The Pearson correlation coefficient is a good choice when all of the following are true:

- **Both variables are quantitative:** We need to use a different method if either of the variables is qualitative.
- **The variables are normally distributed:** We can create a histogram of each variable to verify whether the distributions are approximately normal. It's not a problem if the variables are a little non-normal.
- **The data have no outliers:** Outliers are observations that don't follow the same patterns as the rest of the data. A scatterplot is one way to check for outliers—look for points that are far away from the others.
- **The relationship is linear:** "Linear" means that the relationship between the two variables can be described reasonably well by a straight line. We can use a scatterplot to check whether the relationship between two variables is linear.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is a data processing technique. Using this we change the *range* of your data.

We use this when we want to transform our data so that it fits within a specific scale, like 0-100 or 0-1. By scaling our variables, we can help compare different variables on equal footing.

Standardization involves transforming the features such that they have a mean of zero and a standard deviation of one. This is done by subtracting the mean and dividing by the standard deviation of each feature. On the other hand, normalization scales the features to a fixed range, usually [0, 1]

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

If there is perfect correlation, then VIF = infinity. That means there is a perfect correlation exists between variables resulting $R^2 = 1$.

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

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

The advantages of the q-q plot are:

1. The sample sizes do not need to be equal.
2. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line