**Assignment-based Subjective Questions**

**Q-1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

Ans-1: “workingday” variable has the highest VIF value among all the categorical variables. Followed by seasons and month.

Graphical user interface, application

Description automatically generated

**Q-2: Why is it important to use drop\_first=True during dummy variable creation?**

Ans-2: drop\_first=True helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Text

Description automatically generated

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

Ans-3: temp and atemp

Chart, histogram

Description automatically generated

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

Ans-4: Post “Residual Analysis”, the Error terms are centered towards 0 value.

Chart

Description automatically generated

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

Ans-5:

* snow+rain – high negative correlation
* temp & yr high positive correlation

Table

Description automatically generated

**General Subjective Questions**

**Q-1: Explain the linear regression algorithm in detail.**

Ans-1: Linear Regression is one of the machine learning algorithms where the result is predicted using known parameters which are correlated with the output. It is used to predict values within a continuous range rather than trying to classify them into categories. The known parameters are used to make a continuous and constant slope which is used to predict the unknown or the result.

A regression problem is when the resulting variable contains a real or a continuous value. It tries to draw the line of best fit from the data gathered from a number of points.

Chart, scatter chart

Description automatically generated

The linear regression model can be represented by the following equation:

Text

Description automatically generated with low confidence

Linear Regression can then be used to draw a trend line which can then be used to confirm or deny the relationship between attributes. If the test is done over a long time duration, extensive data can be collected and the result can be evaluated more accurately.

Linear Regression’s power lies in its simplicity, which means that it can be used to solve problems across various fields. At first, the data collected from the observations need to be collected and plotted along a line. If the difference between the predicted value and the result is almost the same, we can use linear regression for the problem.

**Q-2: Explain the Anscombe’s quartet in detail.**

Ans-2: Visualization may not be as precise as statistics, but it provides a unique view onto data that can make it much easier to discover interesting structures than numerical methods. Visualization also provides the context necessary to make better choices and to be more careful when fitting models. Anscombe’s Quartet is a case in point, showing that four datasets that have identical statistical properties can indeed be very different.

In 1973, Francis J. Anscombe published a paper titled, Graphs in Statistical Analysis. The idea of using graphical methods had been established relatively recently by John Tukey, but there was evidently still a lot of skepticism. Anscombe first lists some notions that textbooks were “indoctrinating” people with, like the idea that “numerical calculations are exact, but graphs are rough.”

He then presents a table of numbers. It contains four distinct datasets (hence the name Anscombe’s Quartet), each with statistical properties that are essentially identical: the mean of the x values is 9.0, mean of y values is 7.5, they all have nearly identical variances, correlations, and regression lines (to at least two decimal places).

**Q-3: What is Pearson’s R?**

Ans-3: The Pearson correlation coefficient (PCC), in statistics, also referred to as Pearson's R, or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

**Q-4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

Ans-4: Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

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

Ans-5: If VIF = Infinity, it means that a perfect correlation between two independent variables.

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

Ans-6: Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution.