General Subjective Questions

1) Explain the linear regression algorithm in detail.

Linear Regression is a supervised machine learning algorithm used for regression tasks. It models the relationship between a dependent variable (output) and one or more independent variables (inputs). The goal is to find a linear relationship that predicts the output based on the inputs. The algorithm aims to determine the coefficients (intercept and slope) in the equation $\cdot y=a_1+a_2\cdot x$, where a_1 is the intercept, a_2 is the coefficient of x, x is the input, and y is the output.

During training, the algorithm adjusts the coefficients using a Cost Function, often the Root Mean Squared Error (RMSE), which measures the difference between predicted and true output values. The best-fit line minimizes this error. Once trained, the model can predict output values for new input data.

2) Explain the Anscombe’s quartet in detail.

Anscombe’s Quartet consists of four datasets designed by Francis Anscombe to highlight the importance of visualizing data before analysis and modeling. Despite having nearly identical simple descriptive statistics, these datasets exhibit distinct distributions and patterns when plotted. The quartet emphasizes the limitations of relying solely on statistical summaries.

The four datasets share similar statistical properties, including variance and mean of both x and y values. However, they differ significantly when visualized on scatter plots. This underscores the need to inspect data visually before applying analytical techniques or building models. The datasets are:

1. Dataset 1: Well-suited for linear regression modeling.
2. Dataset 2: Unsuitable for linear regression due to its non-linear nature.
3. Dataset 3: Contains outliers that challenge linear regression modeling.
4. Dataset 4: Also contains outliers, illustrating the impact on linear regression.

Anscombe’s Quartet serves as a cautionary example, highlighting that diverse patterns and anomalies in data may not be evident through summary statistics alone, reinforcing the importance of exploratory data analysis.

3) What is Pearson’s R?

Pearson's correlation coefficient (Pearson's r) quantifies the strength of the linear relationship between two variables. It ranges from -1 to +1, where:

- $r=1$indicates a perfect positive linear relationship (both variables move in the same direction),
- $r=-1$indicates a perfect negative linear relationship (both variables move in opposite directions),
- $r=0$ indicates no linear association.

The coefficient's magnitude signifies the strength of the association:

- $0<|r|<0.5$suggests a weak association,
- $0.5<=|r|<0.8$suggests a moderate association,
- $||r|\geq 0.8$ suggests a strong association.

Pearson's r is sensitive to linear relationships and may not capture non-linear associations. It is a valuable tool for summarizing linear dependence but should be complemented with visual inspection of data for a comprehensive understanding.

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

Scaling is a critical step in data preprocessing, specifically applied to independent variables to standardize their values within a defined range. This process not only contributes to the normalization of data but also accelerates algorithmic calculations.

In many cases, datasets exhibit features with significant variations in magnitudes, units, and ranges. Without proper scaling, algorithms may prioritize magnitude over units, leading to inaccuracies in modeling. Scaling addresses this issue by bringing all variables to a uniform magnitude level.

It's crucial to understand that scaling exclusively influences coefficients and does not impact other parameters such as t-statistic, F-statistic, p-values, R-squared, among others.

There are two common scaling techniques:

1. **Normalization/Min-Max Scaling:**

- This method scales data to the range of 0 to 1.

2. **Standardization Scaling:**Standardization transforms values into Z scores, placing the entire dataset into a standard normal distribution with a mean (□μ) of zero and a standard deviation (□σ) of one.

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

This occurs because the VIF serves as a metric for quantifying the degree of correlation between a particular predictor and the other predictors within a model, primarily utilized for identifying collinearity or multicollinearity.

In cases of perfect correlation, the VIF equals infinity. Elevated VIF values signify the challenging or impossible task of accurately assessing the individual contribution of predictors to a model. Specifically, if VIF is 4, it implies that the variance of the model coefficient is inflated by a factor of 4 due to multicollinearity. This inflation leads to a proportional increase, by a factor of 2, in the standard error of the coefficient. The standard error, crucial for determining confidence intervals of model coefficients, may become substantial in the presence of multicollinearity. Consequently, enlarged standard errors may result in wider confidence intervals, potentially rendering the model coefficient non-significant.

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

A Q-Q plot, short for Quantile-Quantile plot, compares two sets of quantiles by plotting them against each other. Quantiles represent fractions where specific values fall below that point, such as the median where 50% of the data lies above and 50% below. The Q-Q plot is a valuable tool for assessing if two datasets share a common distribution, utilizing a reference line at a 45-degree angle. If the datasets stem from the same distribution, the points will align with this line.

In linear regression, the Q-Q plot serves as a graphical method to determine the distributional similarity between two datasets. The plot juxtaposes the quantiles of one dataset against those of another. The slope of the Q-Q plot provides insights into the relative step sizes within the data. For instance, in a dataset with N observations, each step covers 1/(N-1) of the data. A steep slope indicates that, in a specific segment of the data, observations are more dispersed than expected under a normal distribution.

Identifying such deviations in the Q-Q plot can be indicative of outliers or unusual patterns in the data. This graphical technique enhances our ability to discern distributional disparities and supports robust model diagnostics in linear regression.

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?

There were 6 categorical variables in the dataset.

1)Season-Most of bookings happened in season 3 ,followed by season 2 and 4.This indicates season can be a good predictor

2)Year-This was not of much relevance

3) weathersit-Most of bookings happened in weather sit1,followed by. 2 and 3.This can be a good predictor

4) holiday-Most of bookings happened when it was holiday

5) weekday-Not much difference in different weekday bookings

6) workingday-Not much difference

7)mnth-most of bookings happened in mnth 5,6,7,8,9,10.It is a good predictor

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

The inclusion of `drop_first=True` holds significance during dummy variable creation as it efficiently mitigates the generation of an extra column. This practice proves instrumental in diminishing correlations that might arise among the dummy variables.

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

Temp and atemp with 0.99 correlation

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

Low vif between predictors and low p values

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

Top 3 features are
Temperature-coeff value 0.7263.A unit increase in it will increase count by 0.7263
Windspeed :coeff value -0.1257 .A unit increase will decrease booking by 0.1257
Weathersit_3:coeff value -0.2938 1257 .A unit increase will decrease booking count by 2938