Bike Sharing Dataset

Linear Regression Subjective Questions

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?

- Bookings are in order of season 3,2,4,1 - with median booking >5000 for season 3, season definitely has impact on bookings
- Months: 5-10 have comparatively higher booking than the rest, mnth can be a relevant variable for analysis
- Weathersit - there is a visible affect on bookings, 1 having highest no of bookings; relevant variable
- Holiday - 75th quartile is similar for both cases; maybe a relevant variable
- Weekday - Not much difference between individual days, little chances of being a relevant variable
- yr - Not a relevant variable

Assignment-based Subjective Questions

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

- It avoids the dummy variable trap, which means where the dummy variables are highly correlated, leading to multicollinearity in regression model.
- It avoids multicollinearity by dropping the first category of each category feature and prevents the creation of redundant dummy variables.
- It also helps in reduce the dimensions, which can make the model simpler and more interpretable.

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

- Temp and atemp has the highest correlation with the target variable as it shows a strong linear relationship indicating a high positive correlation.

Assignment-based Subjective Questions

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

- Variables with high p-value is supposed to be removed from the training set.
- If p-value is high, it can also affect the VIF (Variance Inflation Factor), it is important to remove the high p-value variables from the training set so that can be able to balance the VIF of the model.

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

- Temp: With coefficient 0.5173, p-value 0.000, this feature has the positive coefficient.
- Yr: With coefficient 0.2326, p-value 0.000, this positive coefficient suggest that the demand for shared bikes has increased over the years.
- Season_4: With coefficient 0.1371, p-value 0.000, this also has a high positive coefficient, indicating the demand of shared bikes higher in this season.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

- Linear regression is a fundamental statistical and machine learning technique used to model the relationship between a dependent variable and one or more independent variables. The primary goal of linear regression is to predict the value of the dependent variable based on the values of the independent variables.
- It can be completed with the help of Data collection, EDA, pre-processing, model specification, parameter estimation, model fitting, model evaluating and prediction.
- After all these steps, OLS (Ordinary Least Squares) is the most common method for estimating the parameters of a linear regression model. It minimizes the sum of the squared differences between the observed and predicted values of the dependent variable.

General Subjective Questions

2. Explain the Anscombe’s quartet in detail.

- Anscombe's quartet is a collection of four datasets that have nearly identical simple descriptive statistics, yet appear very different when graphed. These datasets were constructed by the statistician Francis Anscombe to demonstrate the importance of graphical data analysis and the effect of outliers and the structure of data on statistical properties.
- It gives the importance of Visualization, sensitivity to Outliers, Model Appropriateness.

3. What is Pearson’s R?

- Pearson's R, also known as Pearson's correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is one of the most commonly used correlation coefficients in statistics.
- It is helpful to create assumptions for the relationship between the variables is linear, the difference between the observed and predicted value is constant and both variables are approximately normally distributed.

General Subjective Questions

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

- Scaling is a pre-processing technique used in data analysis and machine learning to adjust the range of independent variables or features of data. It is essential because many machine learning algorithms are sensitive to the scales of the input features. If features have different scales, algorithms may perform poorly or produce biased results. Scaling helps to ensure that each feature contributes equally to the model, improving performance and convergence speed.
- It improves convergence in Optimization Algorithms, reduces bias in Distance-Based Algorithms, enhances Model Interpretability and prevents Numerical Instability.
- Difference between normalized scaling and standardized scaling is:
  - ❖ Normalization rescales data to a fixed range typically [0,1] meanwhile, Standardization rescales data to have a mean of 0 and standard deviation of 1.
  - ❖ Normalization compresses or stretches the data based on the min and max values, which may not maintain the original distribution, meanwhile, Standardization centres the data around the mean, preserving the original distribution shape but scaling it to a standard normal distribution.
  - ❖ Normalization is sensitive to outliers, as min and max values can be significantly affected, meanwhile, Standardization is less sensitive to outliers compared to normalization but can still be influenced by them.

General Subjective Questions

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

- The Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in regression analysis. It quantifies how much the variance of a regression coefficient is inflated due to collinearity with other predictors.
- The value of VIF becomes infinite when Ri2 = 1. This occurs when there is perfect multicollinearity, meaning that the predictor Xi can be perfectly explained as a linear combination of the other predictors. In such cases, the denominator 1−Ri2 becomes zero, leading to an infinite VIF.

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

- A Q-Q plot, or Quantile-Quantile plot, is a graphical tool used to assess if a dataset follows a particular theoretical distribution, typically the normal distribution. It plots the quantiles of the data against the quantiles of the theoretical distribution. If the data follows the theoretical distribution, the points will approximately lie on a straight line.
- In linear regression, a Q-Q plot is primarily used to check the assumption of normality of the residuals. The assumption that the residuals (errors) are normally distributed is crucial for making valid inferences about the regression coefficients and for the accuracy of confidence intervals and hypothesis tests.
- It helps in checking the normality of residuals, identifying outliers and model diagnostics.