**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)

Ans: The categorical variables are season, mnth, yr, weekday, working day and weathersit. They a major effect on the dependent variable 'cnt'.

1. Why is it important to use **drop\_first=True** during dummy variable creation?

Ans: Lets say there are 4 variables (Spring, Summer, Fall, Winter) With drop\_first=True we might drop only 'Spring', and now your model has three dummies (Summer, Fall, Winter). The coefficient for each of these dummies represents the effect of that season relative to Spring, simplifying interpretation and also it creates an unnecessary redundancies

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

Ans: Temp & atemp has the highest correlation

A screenshot of a graph

Description automatically generated

1. How did you validate the assumptions of Linear Regression after building the model on the

training set? (3 marks)

Ans: The assumptions of Linear Regression after building the model on the training set is done by Normality of error terms, Multicollinearity check, Linear relationship validation, Homoscedasticity, Residuals

1. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes?

Ans: The top 3 features contributing significantly towards explaining the demand of the shared bikes are temperature, year and season

**General Subjective Questions**

1. Explain the linear regression algorithm in detail.

Ans: Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The simplest form is simple linear regression, which models the relationship between two variables by fitting a linear equation (y = mx + c) to the data, where y represents the dependent variable, m is the slope of the line, x is the independent variable, and c is the y-intercept. In multiple linear regression, the model involves more than one independent variable and the equation is y = b0 + b1x1 + b2x2 + ... + bnxn, where b0 is the intercept and b1, b2, ..., bn are the coefficients of the independent variables x1, x2, ..., xn. The objective of the linear regression algorithm is to find the values of the coefficients that minimize the difference between the predicted values and the actual values — this difference is quantified using a cost function such as the mean squared error. The fitting process involves solving for these coefficients that minimize the cost function, often using techniques like the gradient descent algorithm or ordinary least squares estimation.

1. Explain the Anscombe’s quartet in detail.

Ans: Anscombe's quartet is a set of four datasets, each consisting of eleven (x,y) points, that were constructed by the statistician Francis Anscombe in 1973. The remarkable feature of these datasets is that they have nearly identical simple statistical properties, such as mean, variance, correlation, and linear regression line, yet when graphed, they have very different distributions and appear very distinct from one another. The quartet serves as a powerful illustration of the importance of graphing data before analyzing it, as summary statistics can be misleading. It also demonstrates the effect of outliers on statistical properties and the limitations of basic statistical measures in fully describing a dataset. Anscombe's quartet is frequently used as an example in statistics education to emphasize the importance of visualizing data and not relying solely on summary statistics.

1. What is Pearson’s R?

Ans: Pearson's R, also known as the Pearson product-moment correlation coefficient, is a measure of the linear correlation between two variables X and Y. It is a value between -1 and +1, where a value of +1 represents a perfect positive linear relationship, a value of -1 represents a perfect negative linear relationship, and a value of 0 represents no linear relationship. Pearson's R is calculated by dividing the covariance of the two variables by the product of their standard deviations. It is a widely used measure in statistics to determine the strength and direction of a linear relationship between two continuous variables, although it does not necessarily imply causation. Pearson's R is sensitive to outliers and assumes that the data follows a normal distribution and that the relationship between the variables is linear.

1. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans: Scaling is a data preprocessing technique used to standardize the range of features in a dataset. It is performed to ensure that all features contribute equally to the analysis and to avoid features with larger ranges dominating those with smaller ranges. Scaling is crucial when using machine learning algorithms that are sensitive to the scale of the input features, such as k-nearest neighbors, support vector machines, and logistic regression.

The two common methods of scaling are:

1. Normalized scaling (also known as min-max scaling): This method rescales the features to a fixed range, typically between 0 and 1. The formula for normalized scaling is: (x - min(x)) / (max(x) - min(x)), where x is the feature value. This method preserves the original distribution of the data.
2. Standardized scaling (also known as z-score normalization): This method transforms the features to have a mean of 0 and a standard deviation of 1. The formula for standardized scaling is: (x - mean(x)) / std(x), where x is the feature value. This method does not bound the values to a specific range but is less sensitive to outliers compared to min-max scaling.
3. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans: The Variance Inflation Factor (VIF) is a measure of multicollinearity, which occurs when there is a high correlation between independent variables in a regression model. VIF quantifies the severity of multicollinearity by providing an index that measures how much the variance of an estimated regression coefficient is increased because of collinearity.

The VIF value becomes infinite when there is perfect multicollinearity, meaning that one independent variable is a perfect linear combination of the other independent variables. In other words, when an independent variable can be perfectly predicted from the other independent variables, the VIF for that variable will be infinite.

This happens because, in the formula for calculating VIF (VIF = 1 / (1 - R^2)), the denominator (1 - R^2) becomes zero when R^2 equals 1. R^2 represents the coefficient of determination, which measures the proportion of variance in the dependent variable that is predictable from the independent variable. When R^2 is 1, it means that all of the variance in the dependent variable can be explained by the independent variable, indicating perfect multicollinearity.

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

(3 marks)

Ans: A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess if a set of data plausibly came from a specified distribution, typically the normal distribution. In the context of linear regression, Q-Q plots are used to check the assumption of normality of the residuals (the differences between the observed and predicted values). The plot compares the quantiles of the residuals against the quantiles of a normal distribution. If the points in the Q-Q plot lie approximately along a straight line, it suggests that the residuals are normally distributed, which is a key assumption of linear regression. Departures from linearity in the Q-Q plot indicate that the residuals are not normally distributed, which may require further investigation or transformation of the variables. The importance of checking the normality assumption using a Q-Q plot lies in the fact that non-normality of residuals can affect the validity of hypothesis tests, confidence intervals, and the overall interpretation of the regression results.