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)
Answer: There are 7 categorical variables - Year, Holiday, Workingday, Season, Month, Weathersit, Weekday - in my final model.
Based on the values of coefficients, here are my findings

a) Bike demand was higher in 2019 than in 2018, implying that the demand grew in the year 2019 since people were more aware
b) Demand for bikes is higher in the later months of the year, which is from Aug-Oct. Since November and December are holiday months, people don't seem to prefer renting bikes on holidays.
c) Demand for bikes is high in fall compared to the other seasons. It is possible that people don't rent them in summer due to the heat, winter due to the cold.
d) People rent more on non-holidays than on holidays, indicating that people are using these bikes for office commute

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
Answer: In order to have an efficient model, the number of features must be relatively low. Let us say a categorical variable has 3 different values (a,b,c). These 3 values can be represented with just 2 values. If the values of b and c are false (0) then it implies that a is true(1). This reduces the number of variables by 1 for each categorical variable. To generalize, a categorical variable with n values can be represented by n-1 dummy variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
Answer: Temperature since the coefficien is 0.51

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
Answer: By verifying that the error terms are normally distributed

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
Answer:

a) Temp - highest positive correlation
b) Weathersit_3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) - highest negative correlation
c) Year - second highest positive correlation

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Answer:

Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. The goal of linear regression is to find the best fit line that can predict the value of the dependent variable based on the independent variable(s).

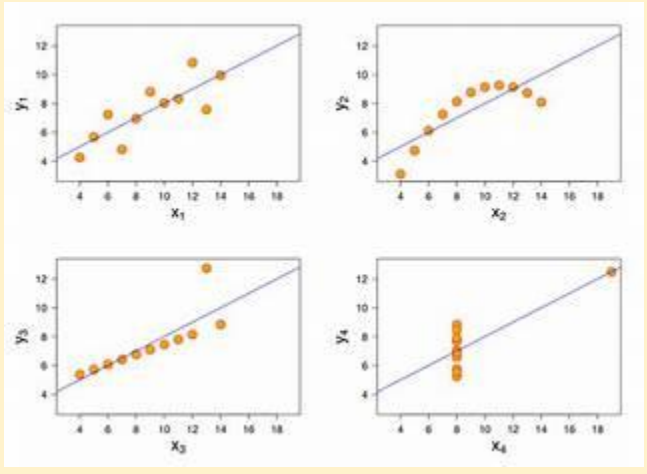
The least squares method is used to draw the best fit line. This method involves drawing a line through all the plotted data points in a way that it minimizes the distance to all of the data points. The distance is called “residuals” or “errors”.

Linear regression can be classified into two types:

a) Simple Linear Regression - Simple linear regression involves only one independent variable. The equation of the line is represented as y = mx + c, where y is the dependent variable, x is the independent variable, m is the slope of the line, and c is the y-intercept
b) Multiple Linear Regression - Multiple linear regression varies from Simple linear regression in that it involves more than one independent variable. In multiple linear regression, the equation of the line is represented as y = b0 + b1x1 + b2x2 + ... + bnxn, where y is the dependent variable, x1, x2, …, xn are the independent variables, and b0, b1, b2, …, bn are the coefficients. The least squares method is used to estimate the values of the coefficients in both simple and multiple linear regression. The coefficients are estimated in such a way that the sum of the squared residuals is minimized. We typically use a recursive feature elimination, in case the number of features is high (over 20). Then cut it down by using VIF and manually eliminating one feature by feature.

2. Explain the Anscombe's quartet in detail. (3 marks)

Answer:
Anscombe's quartet is a group of four datasets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. These datasets are shown below.



Each dataset consists of eleven (x, y) points. The quartet was constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough".

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3. What is Pearson's R? (3 marks)

Answer:

Pearson's R is a correlation coefficient that measures the strength and direction of the linear relationship between two quantitative variables. It is a number between -1 and 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation.

The formula for Pearson's R is

r = (nΣxy - ΣxΣy) / sqrt((nΣx^2 - (Σx)^2)(nΣy^2 - (Σy)^2))

Here, x and y are the two variables, n is the number of observations, Σ represents the sum of the values, and sqrt represents the square root function [2].

Pearson's R is widely used in statistics to determine the strength and direction of the relationship between two variables. It is also used to test hypotheses about the relationship between two variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scalingand standardized scaling? (3 marks)
Answer: Scaling refers to the process of bringing numeric variables to one scale. This is done for two reasons

1) To mitigate the effect of a numeric variable with high values or low values on the coefficients.
   For example: if one of the numeric variables such as salary takes values from 100K to 1M, while another numeric variable such as tenure in a company takes values from 1-10 (years), the salary would have a disproportionately high effect on the target variable.
2) To speed up the optimization process. Optimization algorithms, such as gradient descent, converge faster when the input variables are on the same scale.

There are two ways to perform scaling.

1) Normalization - it scales the data to be between zero and one
2) Standardization - it scales the data such that the mean is zero and standard deviation is one

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
Answer: A VIF value of infinity indicates that there is perfect multicollinearity between the independent variables in the model. This means that one or more of the independent variables can be expressed as a linear combination of the other independent variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
Answer: A Q-Q plot (quantile-quantile plot) is a graphical tool used to assess if a set of data plausibly came from some theoretical distribution such as a normal, exponential, or uniform distribution. It is also used to determine if two data sets come from populations with a common distribution.

In linear regression, Q-Q plots are used to check the validity of the assumption that the residuals are normally distributed. The residuals are the differences between the observed values and the predicted values of the dependent variable.

A Q-Q plot is a plot of the quantiles of the first dataset against the quantiles of the second dataset. If the residuals are normally distributed, the points on the Q-Q plot will lie approximately on a straight line at an angle of 45 degrees from the x-axis.

Q-Q plots are important in linear regression because they help to identify non-normality in the residuals. Non-normality can lead to biased estimates of the regression coefficients and incorrect hypothesis tests.

In conclusion, Q-Q plots are a useful tool in linear regression to check the validity of the assumption that the residuals are normally distributed. They help to identify non-normality in the residuals, which can lead to biased estimates of the regression coefficients and incorrect hypothesis tests.