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?

Answer:

1. Season: 3:fall has highest demand for rental bikes
2. I see that demand for next year has grown
3. Demand is continuously growing each month till June. September month has highest demand. After September, demand is decreasing
4. When there is a holiday, demand has decreased.
5. Weekday is not giving clear picture abount demand.
6. The clear weathershit has highest demand
7. During September, bike sharing is more. During the year end and beginning, it is less, this could be due to extreme weather conditions.

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

Answer: it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

Answer: Temperature(Temp)

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

Answer: 1. Error terms are normally distributed with mean zero.
2. There is a linear relationship between X and Y
3. There is no multi-collinearity between predictor variables.

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

Answer: [+]Temperature (temp)

[+]Weather Situation 3 (weathersit_3)

[+]Year (yr) (2 marks)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:[+] Linear Regression is a Supervised Machine Learning algorithm.

[+]Linear Regression is a linear approach to find relationship between predictors and a response variable.

[+]Used to make predictions when the outcome is a continuous variable.

[+]Linear Regression also helps us to find the features (i.e. predictors) that contribute most significantly in predicting the outcome value.

[+]Linear Regression is the most commonly used predictive model, known for its ease of use and interpretation.

Equation of a line

- The most popular form of line representation in algebra is the “slope-intercept” form, written as

$$Y = m*X + c$$

m is the slope of the line and is calculated as the ratio of change in Y-value (rise) and the corresponding change in X-value (run)

c is the y-intercept. i.e. the Y-value at X=0 (where line intersects Y-axis)

- Slope is also called Gradient of a straight line and shows how steep a straight line is.

• Machine Learning borrows the concept of linear regression from statistics to make predictions for a numeric output variable.
• A Simple Linear Regression maps a relationship between an input predictor variable (X) and an output response variable (Y) as:
$Y = \beta_0 + \beta_1 * X$
Here, β_0 is the constant intercept value that describes the y value when x is zero.
And β_1 is the called the coefficient of the X variable, and describes the amount of change in Y variable when X changes by 1 unit.
• Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. More specifically, our goal is to estimate values of parameters $\beta_{0,}$ β_1 to calculate predicted value of Y for a given X.
• We denote the predicted values with a circumflex or a “hat” on the corresponding variable notation. For examples, estimated values of $\beta_{0,}$ β_1 and Y would be ($\hat{\beta}_{0,}$ $\hat{\beta}_1$, $\hat{Y}$ respectively. So, we write predictions as:
$\hat{Y} = (\hat{\beta}_0 + \hat{\beta}_1 * X$

Using linear regression algorithm our goal would be to find a line that minimizes the sum of error for all the observed data points.

2. Explain the Anscombe’s quartet in detail.

Answer: Anscombe's Quartet shows that multiple data sets with many similar statistical properties like variance values, means, and even linear regression can not accurately portray data in its native form and can still be vastly different from one another when graphed.

3. What is Pearson’s R?

Pearson’s correlation (also called Pearson’s *R*) is a correlation coefficient commonly used in linear regression.

Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.

One of the most commonly used formulas is Pearson's correlation coefficient formula.

What is Pearson Correlation?

Correlation between sets of data is a measure of how well they are related. The most common measure of correlation in stats is the Pearson Correlation. The full name is the Pearson Product Moment Correlation (PPMC). It shows the linear relationship between two sets of data.

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

Answer: Scaling is necessary to get the all the columns in a data set to get same level of magnitude or unit.

Normalized scaling is between 0 to 1. It is useful when there are no outliers.

standardized is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score. Mean and standard deviation is used for scaling. It is not bounded to a certain range and much less affected by outliers.

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

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

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. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.