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?**
   - **Season**- Bike's demand peaks at fall season and lowest in spring.
   - **Holiday** - Bike's demand almost remains same if its working day or holiday
   - **Weather**- Bike's demand peaks if weather is Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist and lowest in Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
   - **Month**- Bike's demand peaks and remains same from May to October
   - **Weekdays**- Bike's demand almost remains same across weekdays
   - **Year**- Bike's demand is higher in year 2019 compared to 2018.

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

Drop_first=true helps in deleting extra column while creating dummy variables as data can be interpreted with n-1 columns. This help in reducing columns with out loosing the information. This also helps in reducing multicollinearity.

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

   Among numerical variable, atemp and temp has highest correlation of 0.63

4) **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   Create a scatter plot of the features versus the target to validate the assumption that there is a linear relationship between the two.

5) **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   - Temp (temp)
   - Year(yr)
   - Weather situation – WS3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)

General Subjective Questions

1) **Explain the linear regression algorithm in detail.**

Linear regression

Regression tries to model the relationship between independent variables and dependent variables. Dependent variable is also known as predictor variable.

Types:

a. Simple Linear Regression
   This helps in building model with one independent variable
b. Multiple Linear Regression
   This helps in building model with more than one independent variable

Details

- Linear regression model tries to identify best fit line. It is represented by following equation

  Y = mx + c

  C: Intercept

  M: Slope

  X: predictor variable/independent variable

- Simple Linear Regression model attempts to explain the relationship between dependent and an independent variable using straight line
- Standard linear regression equation can also be written as $y = \beta_0 + \beta_1 x$
- Best Fit Line- Line which fits the scatter-plot in the best way
- Best Fit line can be identified by reducing cost function
- Residual can be identified by (Residual = Measured Value - Predicted Value)
- Residual is actually error
- Multiple linear regression is needed when one variable might not be sufficient to create good model and make accurate predictions
- MLR can be represented by equation Y = $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$
- Coefficients still obtained by minimizing sum of squared error

2) Explain the Anscombe's quartet in detail.

It is a group of four datasets that provide useful caution against applying individual statistical methods to data without first graphing them. They have identical statistical properties, but they look total different when graphed.

We have to be careful to as statistics (mean, variance, lines of best fit , correlation coefficient) are same but data is all very different when plotted graphically

3) What is Pearson's R?

Pearson's R

- It measures the linear correlation between X and Y
- It has value between -1 and +1
- +1 – It represents total linear correlation
- 0 – represents no linear correlation
- -1 – represents negative linear regression

$$\rho X, Y = \frac{Cov(X, Y)}{\sigma_x \sigma_y}$$

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

Scaling

Scaling data is the process of increasing or decreasing the magnitude according to fixed ration. Scaling helps in handling different units of dependent variables and bring these variable to common standard. It also helps in reducing the computation expense.

Normalized scaling and standardized scaling

Standardized scaling

- In standardization, we center the data and divide by standard deviation
- X(std) = (X – X bar)/(Sx)
- It helps to compare different variable measurements

Normalization

- It ranges from 0 to 1
- It helps in maintaining range between 0 and 1
- Also called MinMax Scaling
    - X(Norm) = $\frac{X - \min(X)}{(\max(X) - \min(X))}$

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

VIF Formula
VIF = 1 – 1/R^2

Description
As R^2 approach 1, VIF will become infinity. It means model perfectly fits and describe the behavior.

We can remove the feature having VIF greater than 10 and if VIF is between 5 and 10=, we have to check and give a thought about elimination.

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

Q-Q plot

- Quantile-Quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.

- It helps to determine if two data sets come from populations with a common distribution.

Advantages

- It can be used with sample size

Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.