Assignment-based Subjective

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

On the basis of my analysis of the categorical Variables from the dataset by using box plot , I can infer that :

- There is increase in demand of rental bikes in year 2018, 2019
- There is incremental rise in demand for rental bikes from Jan to Jun and as winter arrives there we can see decline in demands as in cold people love to stay at home and enjoy
- Demand is high during Fall season,and least in Spring season which is because of best weather condition .
- During holidays, there is high demand which is because people are free and they have time to explore
- Weekday is same, although there is little bit more demand on Wednesday
- And when weather is clear, there are high demand as it is safe and enjoyable to roam in clear weather

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

While creating dummy variable for the categorical variable, drop_first = True is used to

remove extra column which create while creating dummies variable ,

for example: weekdays_df = pd.get_dummies(Bike_df['weekday'],drop_first = True)

if we do not use drop_first= True, then it will create 7 columns, which data can be compensated by using 6 columns only.

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

temp and atemp has highest correlation with target variable

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

I have validate the assumptions of linear Regression after building the model on training set :

- Normality of Error terms
  Errors should be normally distributed
- Multicollinearity check
  There should be insignificant multicollinearity
- Independence of residuals
  There should no auto correlation

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

- Temp
- Winter
- Yr

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**
   Linear regression is a statistical machine learning algorithm used for predictive modeling, especially in the context of regression problems, which analyse linear relationship of dependent with the given set of independent variables.
   Mathematically, Linear regression is presented by below equation:
   Y= mx+c
   Where,
   Y = Target Variable
   M = slope of line
   X = independent Variable
   C = constant

   Furthermore,
   Linear relationship cane be positive or negative
   And Linear Regression can be of two types:
   - Simple Linear Regression
   - Multiple Linear Regression

Simple Linear Regression:

- Linear regression aims to model the relationship between a dependent variable (target) and one or more independent variables (features). The goal is to find the linear equation that best represents this relationship.

Which can be represented by equation Y= mx+c

Multiple Linear Regression:

- Linear regression can extend to multiple independent variables, known as multiple linear regression. The equation becomes : y = b0 +b1*x1 +b2*x2…….bn*xn where b0 is the intercept and b1,b2,b3….bn are the coefficients.

- Linear regression assumes linearity, independence, homoscedasticity, and normality of residuals.

- Linear regression is widely used for tasks like predicting house prices, stock prices, and other continuous variables. Understanding the relationship between variables is crucial for effective predictive modeling.

2. **Explain the Anscombe's quartet in detail.**
   Anscombe's Quartet is a set of four datasets, each consisting of 11 (x, y) pairs, created by the statistician Francis Anscombe. Despite having different distributions, these datasets share nearly identical summary statistics. The significance lies in illustrating the importance of data visualization and the limitations of relying solely on summary statistics.
   Characteristics:
   The four datasets appear similar when examining basic statistical properties (mean, variance, correlation, and regression coefficients).
   When graphically visualized, the datasets reveal distinct patterns:
   - Dataset I: Linear relationship.
   - Dataset II: Non-linear, but still well-fitted to a linear model.
   - Dataset III: Strongly influenced by an outlier.
   - Dataset IV: No correlation but driven by an influential outlier.

   Implications:

   Anscombe's Quartet highlights the danger of relying solely on summary statistics without visualizing the data. Different datasets with diverse underlying structures can produce the same summary statistics.

3. **What is Pearson's R?**
   Pearson's R, or the Pearson correlation coefficient (PCC), is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It assesses how well a straight line can describe the relationship between the variables. The coefficient ranges from -1 to 1:
   - **Positive correlation(r>0) :** Indicates a direct linear relationship; as one variable increases, the other tends to increase.
   - **Negative correlation(r<0):** Suggests an inverse linear relationship; as one variable increases, the other tends to decrease.
   - **No Correlation(r=0) :** Implies no linear relationship between the variables.

The formula for Pearson's R involves covariance and standard deviations of the variables. It is widely used in statistics to analyze and interpret the association between two quantitative variables.

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

Scaling is a data preprocessing technique applied to independent variables, aiming to bring them within a specific range. It ensures that variables with different scales

contribute equally to the analysis, preventing one dominant variable from influencing the model disproportionately.

Scaling is performed due to :

- Equal Contribution : Variables with larger scales might dominate the modeling process.
- Algorithm Sensitivity: Some machine learning algorithms are sensitive to the scale of input features.

There are 2 ways to do Scaling :

1.Normalized Scaling :

- Range : Typically scales data between 0 and 1.
- Formula: $X_{\text{normalized}}=X_{\max}-X_{\min}X-X_{\min}$

2. Standardized Scaling :

- Range : Scales data to have a mean of 0 and a standard deviation of 1.
- Formula : $X_{\text{standardized}}=\text{std}(X)X-\text{mean}(X)$

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

The occurrence of infinite Variance Inflation Factor (VIF) values indicates perfect multicollinearity among the independent variables in a regression model. Perfect multicollinearity happens when one or more independent variables can be precisely predicted using a linear combination of others. In this scenario, the VIF becomes undefined or infinite because it involves dividing by zero

This situation typically arises when:

- One variable is a constant multiple of another
- A variable is omitted from the model.

Perfect multicollinearity impedes the estimation of regression coefficients, as the information provided by the correlated variables becomes redundant. To address this issue, it is essential to identify and address the root cause, which may involve removing one of the correlated variables or addressing data issues.

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

A Quantile-Quantile (Q-Q) plot is a graphical tool used in statistics to assess whether a dataset follows a particular theoretical distribution. In linear regression, a Q-Q plot is commonly employed to evaluate the normality assumption of residuals. The plot compares the quantiles of the observed residuals against the quantiles expected from a normal distribution.

Use and Importance:

- **Normality Assessment :** Q-Q plots help analysts and data scientists visually inspect if the residuals of a linear regression model are normally distributed.

Deviations from a straight line in the Q-Q plot may indicate departures from normality.

- **Model Assumption Checking :** Linear regression models often assume that residuals are normally distributed. Confirming this assumption ensures the reliability of statistical inferences drawn from the model.
- **Identifying Outliers:** Outliers or extreme values in the tails of the Q-Q plot can be indicative of issues such as heavy tails or skewness in the residuals.
- **Model Validation:** Assessing normality through Q-Q plots is part of the model validation process, ensuring that the chosen statistical model is appropriate for the data.