Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer:

1. Bike on season – 3(fall) has high booking and demand
2. Company product usage has increased in 2019 compare to 2018
3. JUNE, JULY, AUG, Sep, OCT the bike demand increases
4. Bike usage on weekday wed, thu, fri are little high comparing to other days
5. WeatherSit bike are mostly used during Cloudy Days (Clear, Few clouds, partly cloudy)

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer:

1. It's always necessary to create n-1 dummy variable for categorical variable greater.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer:

1. Temp and atemp has high correlation with 0.63
2. Registered and cnt also has high correlation with 0.945
3. A good correlation is Temp and Cnt with 0.627

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer:

1. Linearity relationship between response and predictor variable
2. Remove multicollinearity by compare Rsquare and VIF
3. Error Distribution using Residual Analysis to check normally disturbed

---

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer:

1. Atemp
2. Season_Winter
3. Month_jan

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

- Linear Regression is a Supervised Learning Algorithm used for predicting the continues values from a dataset and establish a Liner relationship between Independent variable and Dependent variable using straight line Equation(y=mx+c)
- The Simple Linear Regression with One independent variable is represented as

  $Y = \beta0+\beta1X+ E$
- Where

  X = One Independent Variable

  Y = Dependent Variable

  β0+β1 = Intercept and Coefficient

  E = Error Term

*The Goal is to find the Bit Fit Line that minimizes the error

$SSE=\sum(Yi-Yi\hat{})2$

Where:

Yi = Actual Values

Yi^ = Predicted values

- Calculate the Rsquare score that helps you fit the best line on linear regression model

---

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Anscombe's quartet – Understanding the importance of Data Visualization

It's a statistic such as mean, variance, correlation, regression that can mislead if data is not visualized properly

Even if the dataset have identical statistical property their distribution and relationship variables can be completely different in nature.

There are four Datasets in Anscombe's Quartet

1.Mean of X and Y

Data Follow a linear trend and a simple regression line to fit

2.Variance of x and y

Data follows a curve relationship

3. Correlation Coefficient

One outlier at x & y is strongly influencing the regression line

4. Linear Regression equation

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Pearson Correlation Coefficient is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables.

Most commonly used in correlation metric

Its range is from -1 to 1

Do not depend on measurement scale

Measures linear association between two variables

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Scaling is the process of converting numerical data in to a similar range

It's used to improve the performance of machine learning models

Scaling

- Improves model Performance
- Reduces Bias
- Ensure Fair Comparisons
- Distance based Algorithms

MIN-Max Scaling

- Transform data into fixed rages (-1 to 1)
- No robust to outliers
- minMax = x-min(c)/max(c)-min(c)

Standardization

- Best for Linear Models, PCA, logistic regression
- Transform data to mean = 0 and Standard Deviation = 1
- SS = X-mean(c)/SD(c)

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

VIF measure the multicollinearity in regression model. A high VIF means a variable is highly correlated with other variable

If Predictor is perfectly correlated with other predictors the denominator in the VIF forula becomes zero and leads the VIF to Infinite

---

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

Quantile-Quantile Plot in Linear regression is a graphical tool used to compare the distribution of a dataset to a theoretical distribution.

Typically the normal distribution in linear regression

It helps to check is errors follow a normal distribution in linear regression

---