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: <Your answer for Question 1 goes below this line> (Do not edit)

Based on my analysis, following can be inferred about the effect of dependant variables on the dependent variable :

1->From the box plots its clear that season, yr, mnth, holiday and weathersit categorical columns vary vividly in terms of cnt(dependant variables) of bike share users.

2->Whereas workingday and weekday shows no such different impact for count.

3->Most number ok bikes i.e. more than 5,500 are rented in the fall season

4->Year 2019 saw a steep increase in the number of bike rents from almost 3300 in 2018 to 5700 in 2019.

5->Most bikes are rented in the month of May to October. May to October month covers most of the summer and fall seasons in USA. Thus this impact can be easily seen in season barplot as well, as there is significant count of bike rentals in summer and fall season as compared to winter and spring.

6->The bike rental count is mostly equally distributed on all days of the week except for Sunday, which shows significantly low count.

7->Bike reantal count is low on Holidays.

8->It can be seen from weathersit barplot that people usually prefer to rent bikes while the weather is clear. Little misty weather makes the preferrances for bike rental a little less, whereas even a light snow discourages people for bile rental.

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: <Your answer for Question 2 goes below this line> (Do not edit)

1->We use drop_first=True during dummy variable creation in order to avoid the multicollinearity scenario.

2->If we kept all the dummy variables, one variable could easily be predicted by the other.

3-> For Example let us say we have a season column having values of spring, summer, fall, and winter. Thus 4 dummy variables can be created out of this column i.e. spring, summer, fall, and winter.

spring	summer	fall	winter
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

There are only 4 possibilities for these 4 columns.

Thus spring can easily be eliminated such that if summer, fall, and winter have 0 values means the season is spring

---

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: <Your answer for Question 3 goes below this line> (Do not edit)

Among the numerical variables, 'atemp' has the highest correlation with the target variable.

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: <Your answer for Question 4 goes below this line> (Do not edit)

I validated the following assumptions of Linear Regression after building the model on the training set :1

1->Normality of Error Terms: Error terms should be normally distributed

2->Variance Inflation Factor (VIF): VIF value should not be greater than 10

3->Multicollinearity: It should be insignificant.

4->Linearity: The relationship between the independent variables and the dependent variable should be linear.

5->Homoscedasticity: There should not be any visible patterns amongst residuals.

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: <Your answer for Question 5 goes below this line> (Do not edit)

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

1->atemp

2->Sep

3->2019

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)

<Your answer for Question 6 goes here>

A statistical technique called linear regression fits a linear equation to observed data in order to model the relationship between one or more independent variables (predictors) and a dependent variable (outcome). The linear regression algorithm is broken down in detail as:

1->Model Specification

Linear Regression model expression is :

$Y=\beta 0+\beta 1X1+\beta 2X2+\ldots+\beta nXn+\epsilon$

Where:

- Y is the dependent variable.
- β0 is the intercept (the expected value of Y when all X are 0).
- β1,β2,…,βn are the coefficients of the independent variables X1,X2,…,Xn.
- ϵ is the error term, representing the difference between the observed and predicted values.

2-> Assumptions:

A number of presumptions underlie linear regression:

- Linearity: The relationship between the independent and dependent variables is linear.
- Independence: Observations are independent of each other.
- Homoscedasticity: Constant variance of errors across all levels of the independent variables.
- Normality of Errors: The errors are normally distributed.
- No multicollinearity: Independent variables are not highly correlated with each other.

3-> Fit the model:

To fit the model, we need to estimate the coefficients (β). The most common method is Ordinary Least Squares (OLS), which minimizes the sum of the squared differences between the observed values and the values predicted by the model.

Steps in OLS:

1. Calculate the residuals: For each observation, calculate the difference between the observed value and the predicted value.
2. Minimize the sum of squared residuals: The goal is to find coefficients that minimize:

$S=\sum(Yi-Y\hat{}i)\hat{}2$

Where $Y\hat{}i$ is the predicted value for the ith observation.

3. Use calculus: Set the partial derivatives of the sum of squared residuals with respect to each coefficient to zero to find the optimal coefficients.

4->Evaluate mode:

After fitting the model, it's essential to evaluate its performance using various metrics:

- **R-squared (R^2)**: Measures the proportion of the variance in the dependent variable that is predictable from the independent variables. Values range from 0 to 1, with higher values indicating a better fit.
- **Adjusted R-squared**: Adjusts R2R^2R2 for the number of predictors in the model, useful for comparing models with different numbers of predictors.
- **Mean Absolute Error (MAE)**: The average absolute differences between predicted and actual values.
- **Root Mean Squared Error (RMSE)**: The square root of the average of the squared differences between predicted and actual values.

5->Making Predictions

Once the model is fitted, you can use it to make predictions for new data by substituting the values of the independent variables into the regression equation.
6->Check the assumptions
After fitting the model, it's crucial to check the assumptions. Common diagnostic methods include:

- Residual plots: To check for linearity and homoscedasticity.
- Q-Q plots: To assess normality of the residuals.
- Variance Inflation Factor (VIF): To detect multicollinearity.

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)

<Your answer for Question 7 goes here>

Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

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)

<Your answer for Question 8 goes here>

1-> The Pearson coefficient is a type of correlation coefficient that represents the relationship between two variables that are measured on the same interval or ratio scale.
2->The Pearson coefficient is a measure of the strength of the association between two continuous variables.
3->To find the Pearson coefficient, also referred to as the Pearson correlation coefficient or the Pearson product-moment correlation coefficient, the two variables are placed on a scatter plot.
4->The variables are denoted as X and Y.
5->There must be some linearity for the coefficient to be calculated; a scatter plot not depicting any resemblance to a linear relationship will be useless.
6->The closer the resemblance to a straight line of the scatter plot, the higher the strength of association.

7->Numerically, the Pearson coefficient is represented the same way as a correlation coefficient that is used in linear regression, ranging from -1 to +1.
8->A value of +1 is the result of a perfect positive relationship between two or more variables.
9->Positive correlations indicate that both variables move in the same direction. Conversely, a value of -1 represents a perfect negative relationship.
10->Negative correlations indicate that as one variable increases, the other decreases; they are inversely related.
11->A zero indicates no correlation.2

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)

<Your answer for Question 9 goes here>
1-> It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
2-> Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.
3->Normalized scaling brings all of the data in the range of 0 and 1.
sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.
3->Standardized scaling : Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).
sklearn.preprocessing.scale helps to implement standardization in python.
One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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)

<Your answer for Question 10 goes here>

1->The VIF can be infinite when there is perfect multicollinearity, meaning that the regressor is equal to a linear combination of other regressors:

2->When calculating the VIF for one independent variable using all the other independent variables, the VIF will be infinite if the R^2 value is 1.

3->This can happen when one of the independent variables is strongly correlated with many of the other independent variables.

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)

<Your answer for Question 11 goes here>

1->What is Q-Q plot

Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against. For example, if you are testing if the distribution of age of employees in your team is normally distributed, you are comparing the quantiles of your team members' age vs quantile from a normally distributed curve. If two quantiles are sampled from the same distribution, they should roughly fall in a straight line.
Since this is a visual tool for comparison, results can also be quite subjective nonetheless useful in the understanding underlying distribution of a variable(s)

2->Uses:

Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian distribution, uniform distribution, exponential distribution or even a Pareto distribution. You can tell the type of distribution using the power of the Q-Q plot just by looking at it.

3->Importance

1. Assessing Normality

2. Identifying Deviations from Normality

3. Model Assumption Validation

4. Detecting Outliers

5. Comparing Distributions

6. Informing Further Analysis