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)

Analysis of the categorical variables from shows that bike rental rates are likely to be higher in

- Demand is higher in the summer and fall seasons, especially when the weather is clear and pleasant.
- Rental rates are higher during the months of March, May, June, July, September, and October.
- Saturdays and Fridays have the highest demand for bike rentals.
- Bike rental counts have increased from 2018 to 2019.

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)

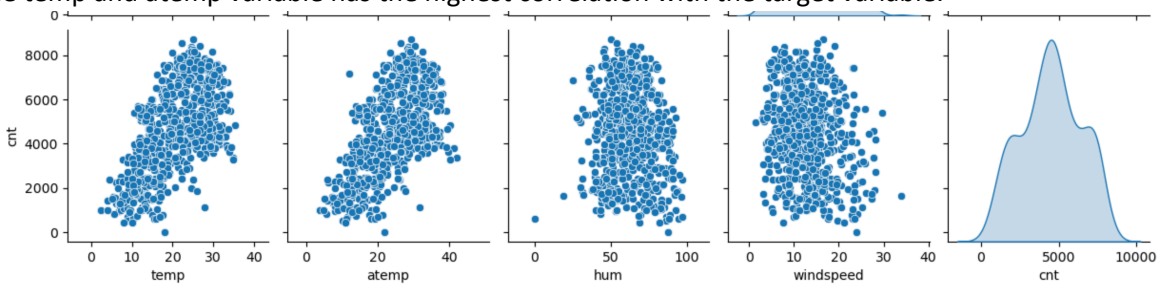
drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. If we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

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)

The temp and atemp variable 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)

Five assumptions of Linear regression are :

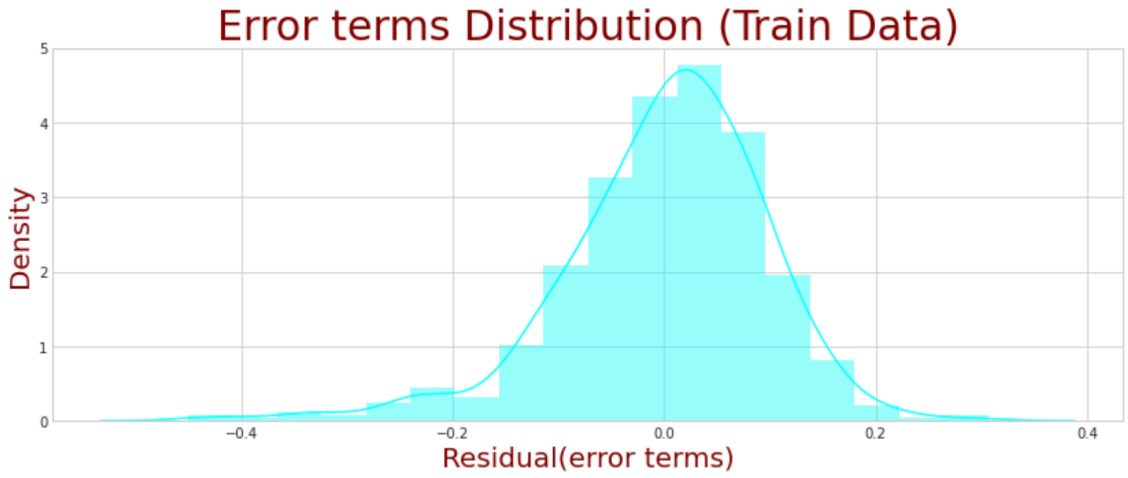
1. Normality : Error terms are normally distributed (around mean = 0) i.e, for any fixed value of X, Y is normally distributed.

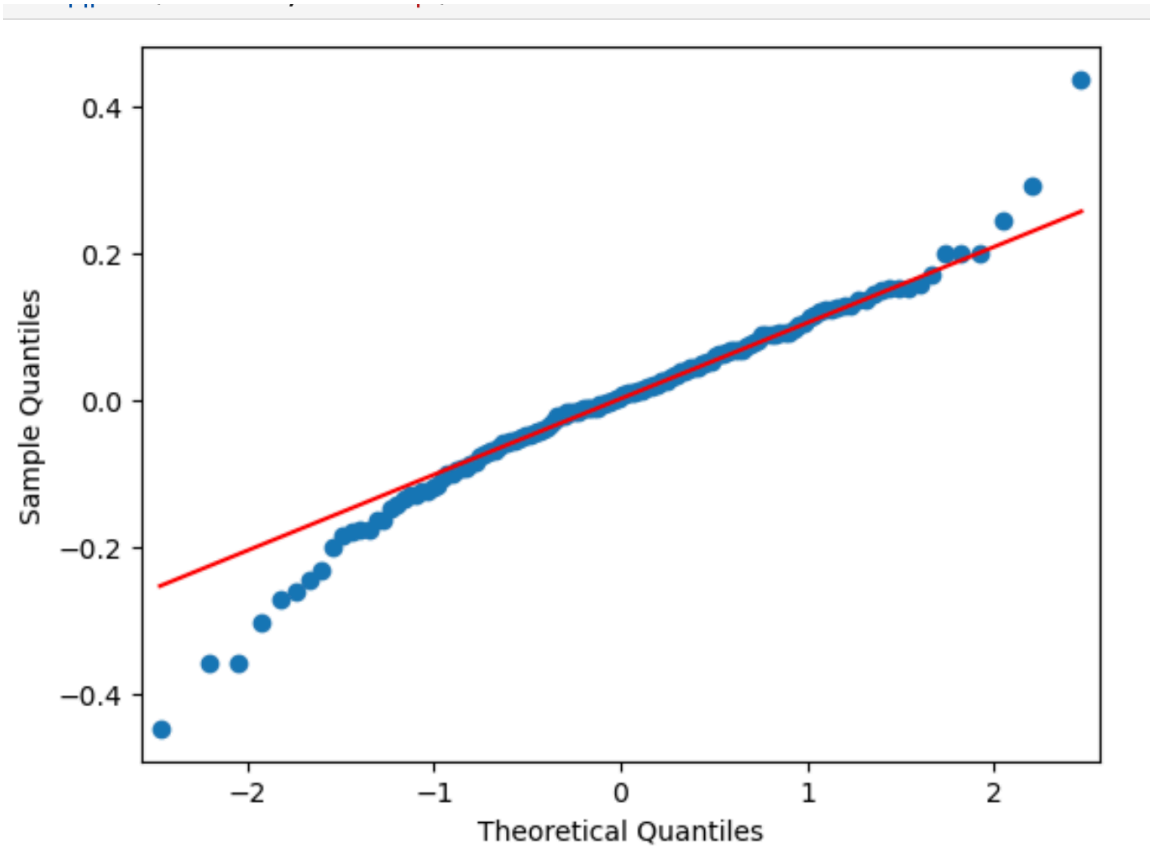
2. Linearity: The relationship between X and the mean of Y is linear.

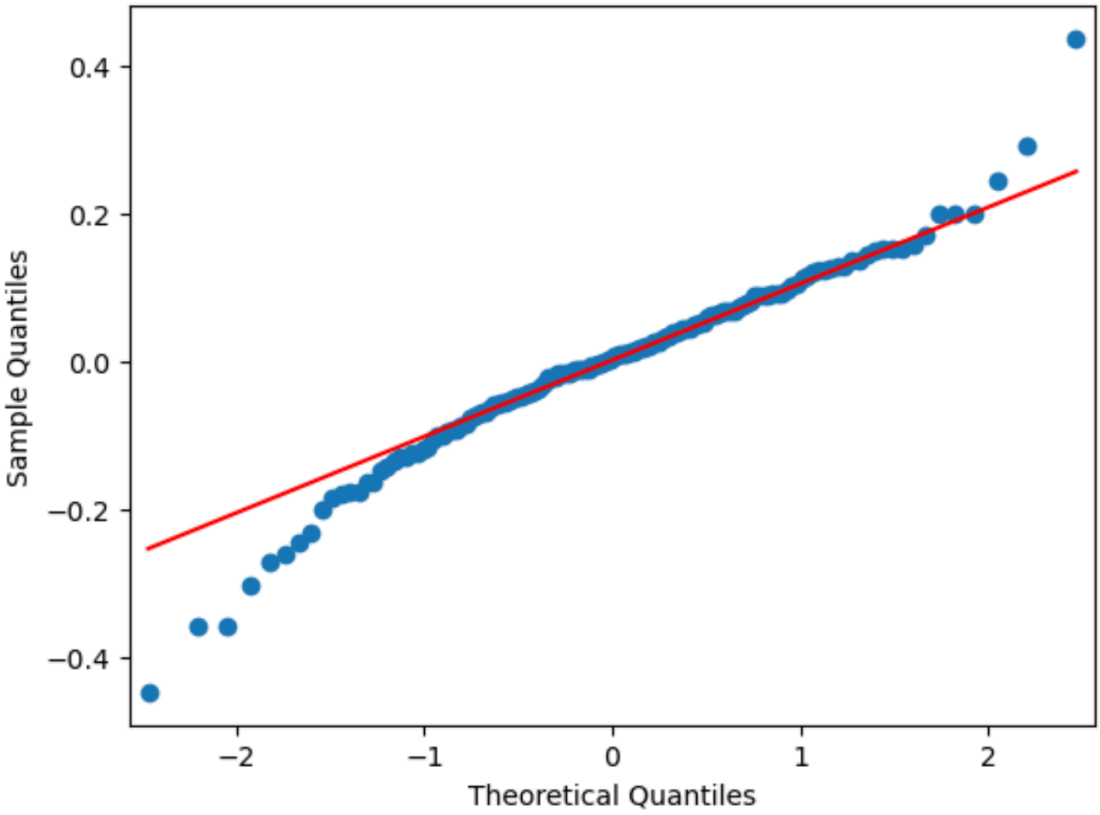
3. No or little multicollinearity : Observations are independent of each other.

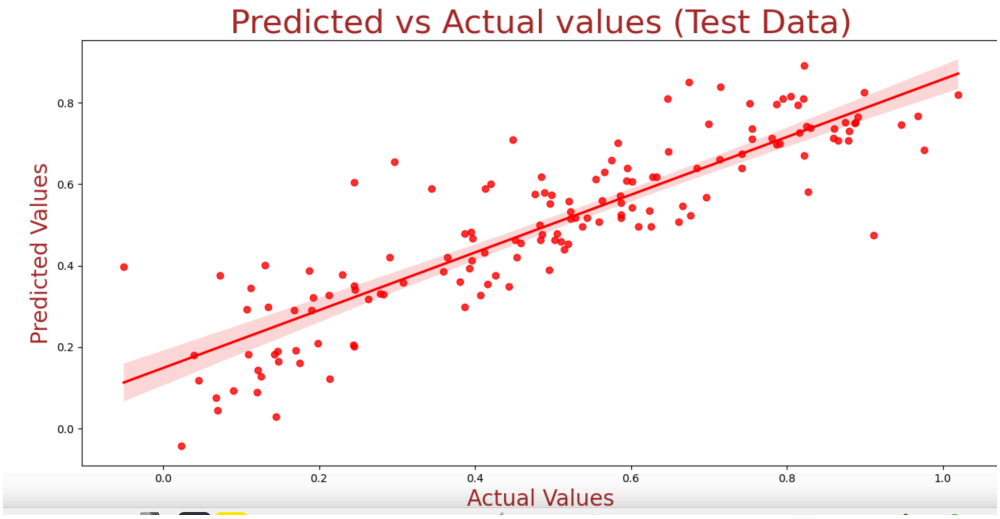
4. Homoscedasticity : The variance of residual is the same for any value of X.

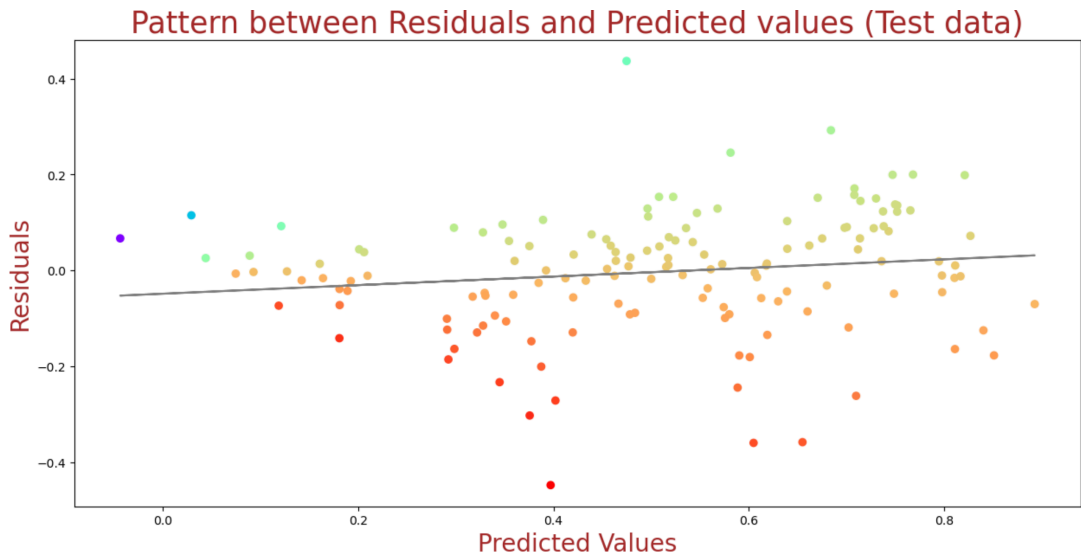
5. No pattern or auto-correlation : Error terms(residuals) are independent of predicted value.






Sample Quantiles
Theoretical Quantiles
0.4
0.2
0.0
−0.2
−0.4
−2
−1
0
1
2


Predicted vs Actual values (Test Data)
Predicted Values
Actual Values
0.0
0.2
0.4
0.6
0.8
1.0



This clearly reflects as we don't see any pattern here as expected from our model.
There is no specific relation between the residuals and predicted values

---

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)

- Temperature (temp): Temperature has the most significant positive impact on bike rental demand. As the temperature rises, the demand for bike rentals increases significantly.
- Year (yr): The year has a positive impact on demand. Over time, there has been an increasing trend in bike rentals.
- Winter Season (winter): The winter season is associated with increased demand, due to recreational winter activities and holidays.

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 fundamental supervised machine learning algorithm used for modeling the relationship between a dependent variable (also known as the target or outcome) and one or more independent variables (predictors or features) by fitting a linear equation to the observed data.
- The goal of linear regression is to find the best-fitting linear relationship that can be used for prediction and understanding the dependencies between variables. Linear regression explained in detail:
-

1. Simple Linear Regression: Simple linear regression is used when there is only one independent variable that affects the dependent variable. The linear equation for simple linear regression is: $Y = \beta 0 + \beta 1X + \varepsilon$

---

2. Multiple Linear Regression :Multiple linear regression extends the concept to situations where there are multiple independent variables. The linear equation for multiple linear regression is $Y=\beta0+\beta1X1+\beta2X2+...+\beta pXp+\varepsilon$

3. Fitting the Model : The goal in linear regression is to estimate the values of $\beta0$, $\beta1$, $\beta2,...,\beta p$ that minimize the sum of squared errors between the predicted values (Y) and the actual values of the dependent variable in the training data.This is typically done using the method of least squares, which minimizes the sum of the squared residuals (the differences between the predicted and actual values).

4. Assumptions:
- The relationship between the independent variables and the dependent variable is linear.
- The error terms (ε) are independent of each other.
- The variance of the error terms is constant across all levels of the independent variables.
- The error terms follow a normal distribution.
- Independent variables are not highly correlated with each other.

5. Interpretation:

- Linear regression allows for the interpretation of the coefficients $\beta0,\beta1,\beta2,...,\beta p$. For example, $\beta1$ represents the change in the dependent variable Y for a one-unit change in X1, holding all other variables constant.

6. Evaluation:
- To assess the performance of a linear regression model, various metrics can be used, such as Mean Squared Error (MSE), R-squared (R2), and others, to measure how well the model fits the data and makes predictions.

7. Predictions:
- Once the model is trained, it can be used to make predictions on new or unseen data by plugging in the values of the independent variables into the linear equation.

8. Extensions:
- Linear regression has several extensions and variations, including ridge regression, lasso regression, and polynomial regression, which address issues like multicollinearity and allow for more flexible modeling.
- Linear regression is a powerful and interpretable tool commonly used for tasks such as predicting house prices, analyzing the impact of variables on an outcome, and understanding relationships in data. However, it has its limitations, and its effectiveness depends on the assumptions being met and the nature of the data being modeled.

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 is a famous example in statistics and data visualization created by Francis Anscombe in 1973 created by Francis Anscombe in 1973 to demonstrate that summary statistics alone may not provide a complete understanding of the data.

- It consists of four datasets, each containing 11 data points, with two numerical variables (X and Y). What makes Anscombe's quartet remarkable is that these four datasets have identical simple descriptive statistics (such as means, variances, and correlation coefficients) for X and Y, yet they have vastly different graphical representations and relationships between the variables.
- Its emphasis on the importance of data visualization and graphical exploration thus revealing hidden patterns, outliers, and relationships that might not be apparent from summary statistics alone.
- This quartet is often used in statistics and data science education to emphasize the importance of visualizing data before drawing conclusions or making decisions based on statistical summaries.

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)

- It measures linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.
- An absolute value of exactly 1 implies that a linear equation describes the relationship. between X and Y perfectly, with all data points lying on a line. The correlation sign is determined by the regression slope: a value of +1 implies that all data points lie on a line for which Y increases as X increases, and vice versa for −1. A value of 0 implies that there is no linear dependency between the variables.

the formula for ρ can also be written as

$$\rho_{X,Y} = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sqrt{\mathbb{E}[X^2] - (\mathbb{E}[X])^2}\,\sqrt{\mathbb{E}[Y^2] - (\mathbb{E}[Y])^2}}.$$

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)

- Scaling is a technique performed in pre-processing during building a machine learning model to standardize the independent feature variables in the dataset in a fixed range.
- The dataset could have several features which are highly ranging between high magnitudes and units. If there is no scaling performed on this data, it leads to incorrect modelling as there will be some mismatch in the units of all the features involved in the model.
- The difference between normalization and standardization is that while normalization brings all

the data points in a range between 0 and 1, standardization replaces the values with their Z scores.

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)

Reasons why the VIF might become infinite:

- Perfect multicollinearity: Perfect multicollinearity occurs when two or more independent variables in a regression model are perfectly correlated with each other. In this case, one variable can be exactly predicted by a linear combination of the others, leading to an R^2 value of 1 and an infinite VIF.
- Perfect multicollinearity: Even if multicollinearity is not perfect but very high, the R^2 value can still approach 1, resulting in a very large VIF.
- Too small a sample size: When you have a small sample size relative to the number of independent variables in the model, it can lead to unstable estimates and high VIF values, including infinity.

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)

- The quantile-quantile plot is a graphical method for determining whether two samples of data came from the same population or not. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
- The importance of a Q-Q plot in linear regression lies in its ability to assess the normality assumption of residuals visually and quantitatively.
- If the Q-Q plot shows a straight line, it provides evidence that the residuals are normally distributed, which is one of the key assumptions of linear regression.
- On the other hand, if the Q-Q plot shows significant deviations from a straight line, it suggests that the normality assumption may not hold, and you may need to consider transformations or other methods to address outliers in your data before drawing conclusions from your regression analysis.
