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:

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)

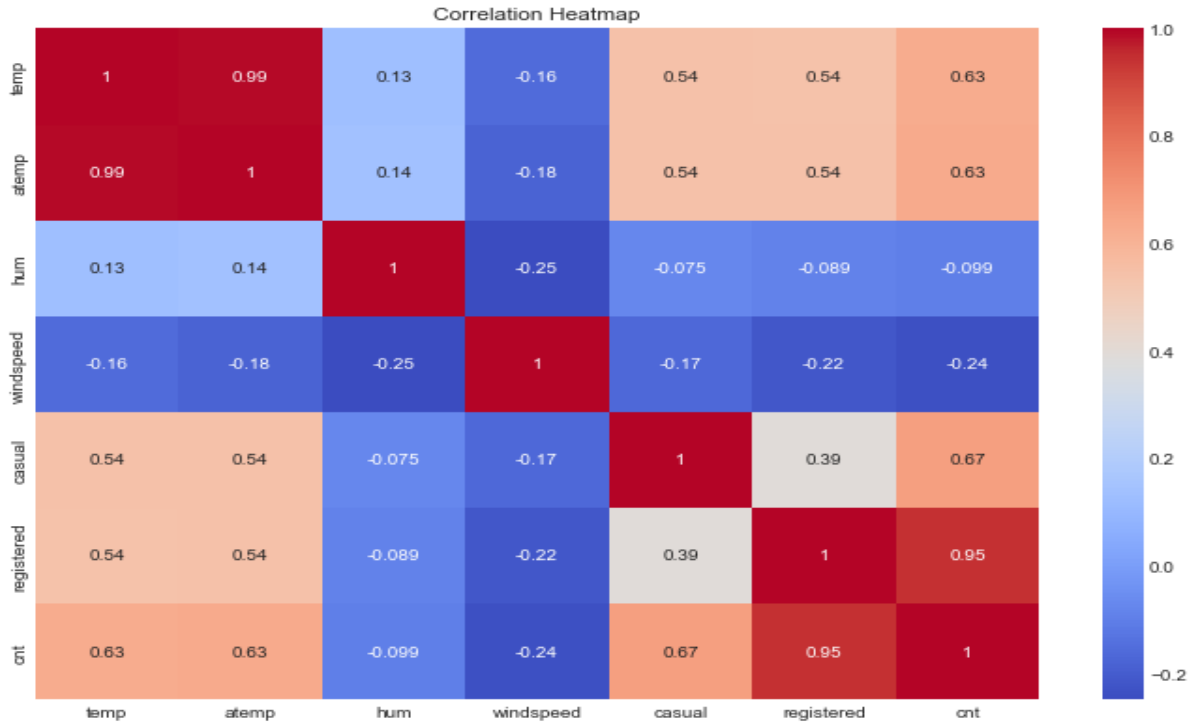
- drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. If we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

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)

- “temp” is the variable which has the highest correlation with target variable i.e. 0.63
- The casual and registered variables are actually part of the target variable as values of these columns sum up to get the target variable, hence ignoring the correlation of these 2 variables.
- “atemp” is the derived parameter from temp, humidity and windspeed, hence not considering it as it is eliminated in the model preparation

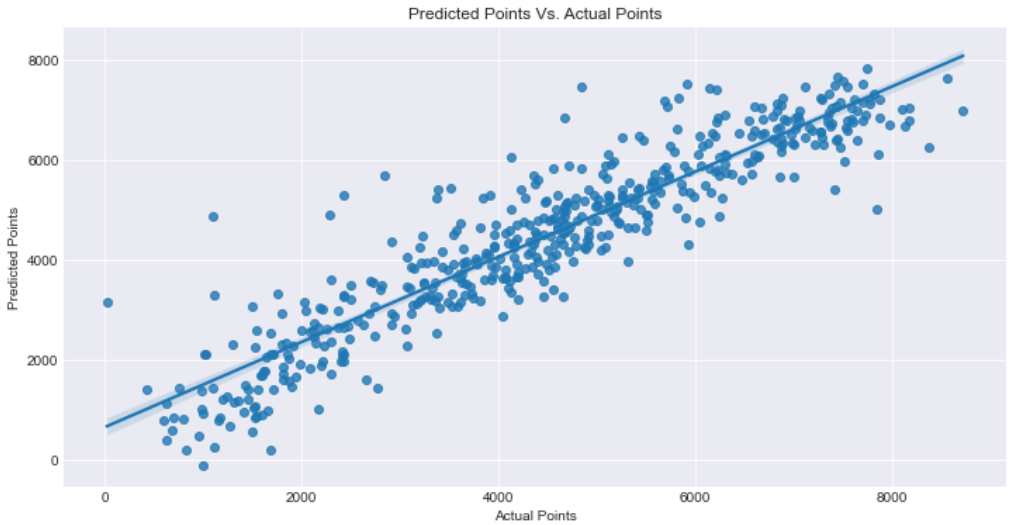


Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

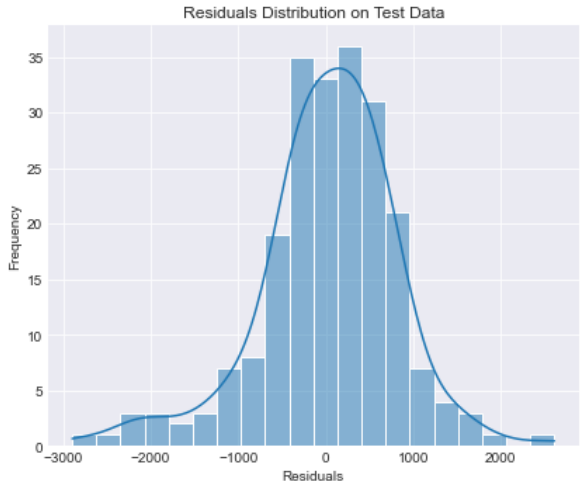
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

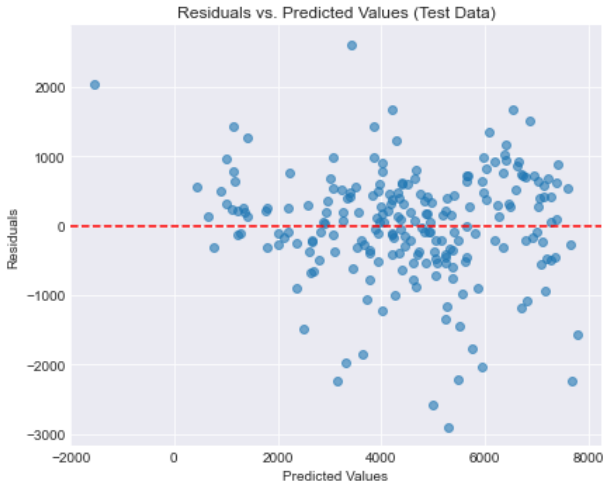
1. Linear relationship between independent and dependent variables – The linearity is validated by looking at the points distributed symmetrically around the diagonal line of the actual vs predicted plot as shown in the below figure.



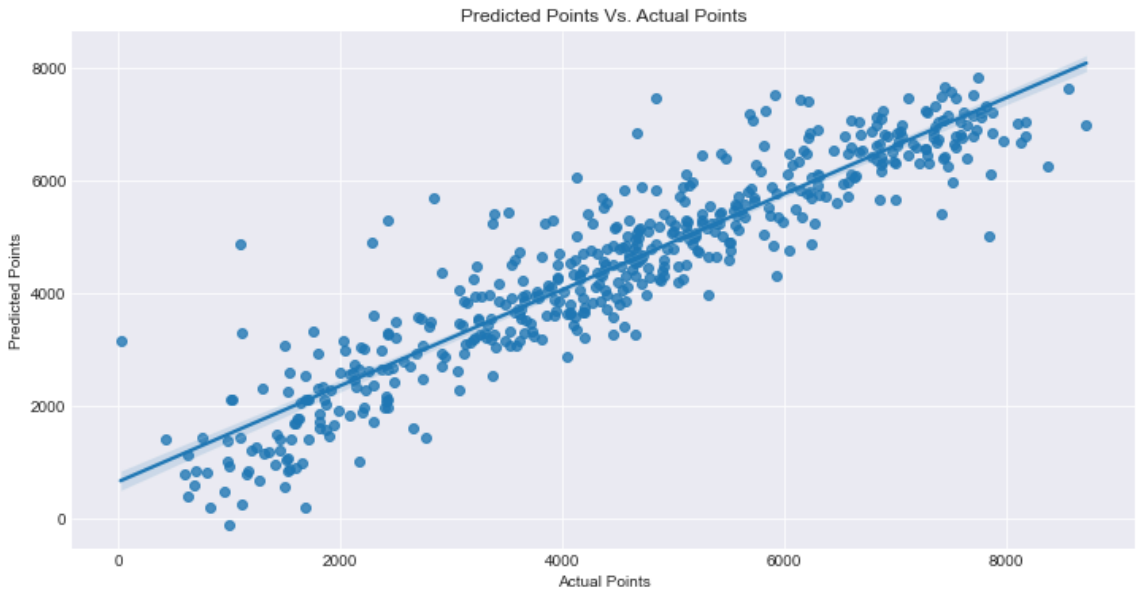
2. Error terms are independent of each other – We can see there is no specific Pattern observed in the Error Terms with respect to Prediction, hence we can say Error terms are independent of each other
3. Error terms are normally distributed: Histogram and distribution plot helps to understand

the normal distribution of error terms along with the mean of 0. The figure below clearly depicts the same.





4. Error terms have constant variance (homoscedasticity): We can see Error Terms have approximately a Constant Variance, hence it follows the Assumption of Homoscedasticity.



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 is playing the crucial role in the demand of shared bikes.

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.
- There are 2 types of linear regression algorithms
    - Simple Linear Regression – Single independent variable is used. ▪ $Y = \beta 0 + \beta 1X$ is the line equation used for SLR.
    - Multiple Linear Regression – Multiple independent variables are used. ▪ $Y = \beta 0 + \beta 1X1 + \cdots + \beta pXp +$ ∈ is the line equation for MLR.
    - $\beta 0 = value\ of\ the\ Y\ when\ X = 0\ (Y\ intercept)$
    - $\beta 1, \beta 2, \ldots, \beta p = Slope\ or\ the\ gradient.$

- Cost functions – The cost functions helps to identify the best possible values for the $\beta 0$, $\beta 1, \beta 2, \ldots, \beta p$ which helps to predict the probability of the target variable. The minimization approach is used to reduce the cost functions to get the best fitting line to predict the dependent variable.
- While finding the best fit line we encounter that there are errors while mapping the actual values to the line. These errors are nothing but the residuals. To minimize the error squares OLS (Ordinary least square) is used
    - o $ei = yi - ypred$ is provides the error for each of the data point.
    - OLS is used to minimize the total e2 which is called as Residual sum of squares.
    - • Ordinary Lease Squares method is used to minimize Residual Sum of Squares and estimate beta coefficients.

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 consists of four datasets that share identical statistical properties (mean, variance, correlation, and regression line) but exhibit different distributions when graphed. Here's a concise overview:

- Dataset I: Linear relationship; points form a straight line
- Dataset II: Quadratic relationship; points form a U-shape
- Dataset III: Strong linear correlation but with one significant outlier that affects the regression.
- Dataset IV: Vertical clustering with a single outlier that distorts the linearity
    - Identical Statistics: All datasets have similar means, variances, and correlations.
    - Importance of Visualization: Graphs reveal different relationships that statistics

alone can't show

- Outlier Impact: Outliers can heavily influence regression results and correlation, leading to misleading conclusions.

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)

- The Pearson correlation coefficient (PCC)[a] is a correlation coefficient that 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.
- $r = n(\Sigma x * y) - (\Sigma x) * (\Sigma y) \sqrt{[n\Sigma x\ 2 - (\Sigma x)\ 2] * [n\Sigma y\ 2 - (\Sigma y)\ 2]}$

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.
