Name – Rishi Shah

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

Here are the inferences that can be derived from categorical variables.

- Season - Cycles are rented more often during summer and fall
- Year - The business has increased in 2019
- Months - Consistent with season, more cycles are rented during summer and fall months.
- Working day/Holiday/Weekday - No significant impact on working day vs holiday vs. weekdays when you look at cnt as the dependent variable. However, for casual users, a holiday results in more business while for a registered user, a working day results in more business.
- Weather - Better weather results in more renting of cycles

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

For k values of a categorical variable, we will need only k-1 variables since these k-1 variables can uniquely determine the k values. Hence, we have drop_first= True to drop the first variable leaving back just k-1 variables.

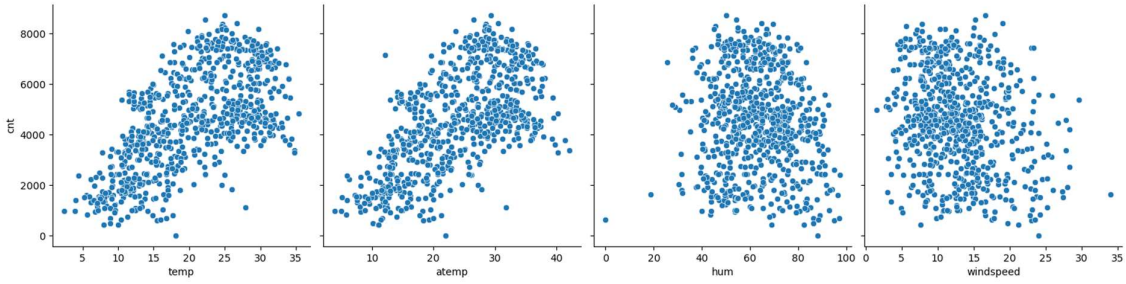
Q. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Actual temperature (temp) and feels like temperature (atemp) have the highest correlation with the target variable.

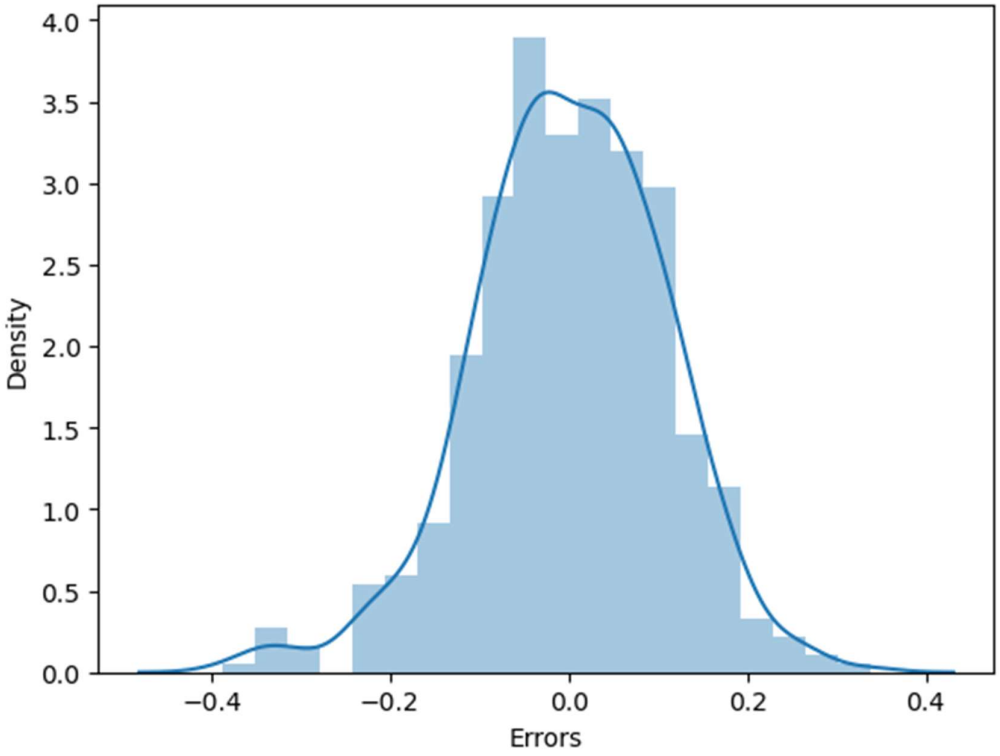
Q. How did you validate the assumptions of Linear Regression after building the model on the training set?

Here are the 4 assumptions and the evidence.

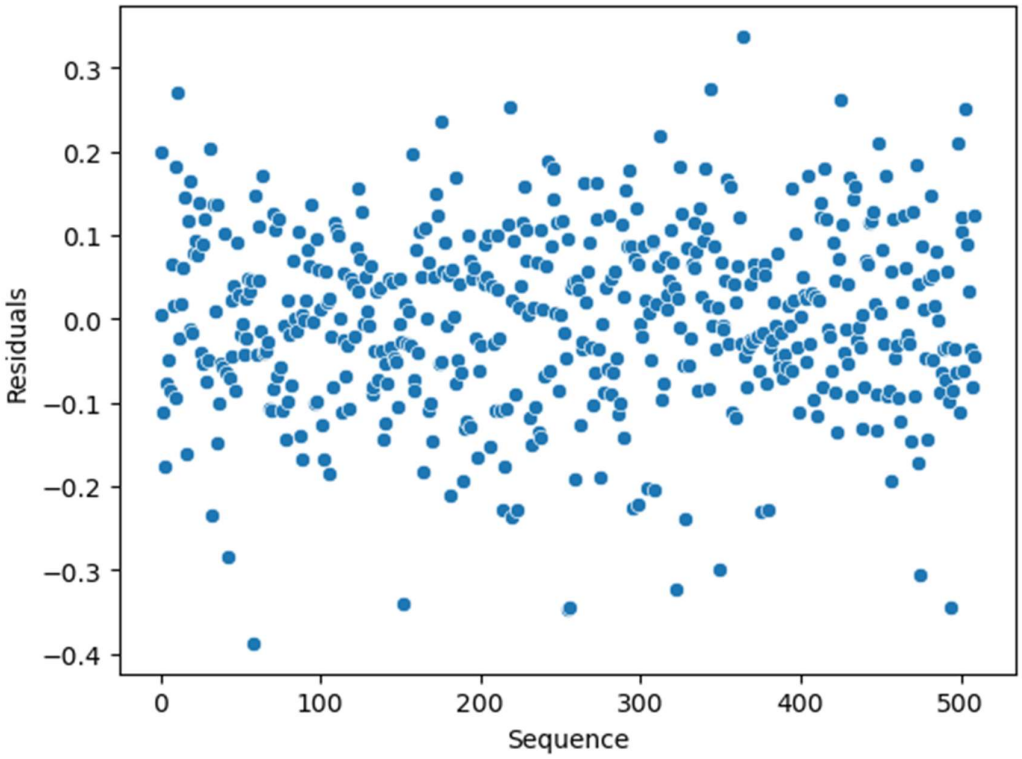
1. There is a linear relationship between X and Y – There is clearly a linear relationship between temperature and cycle count.



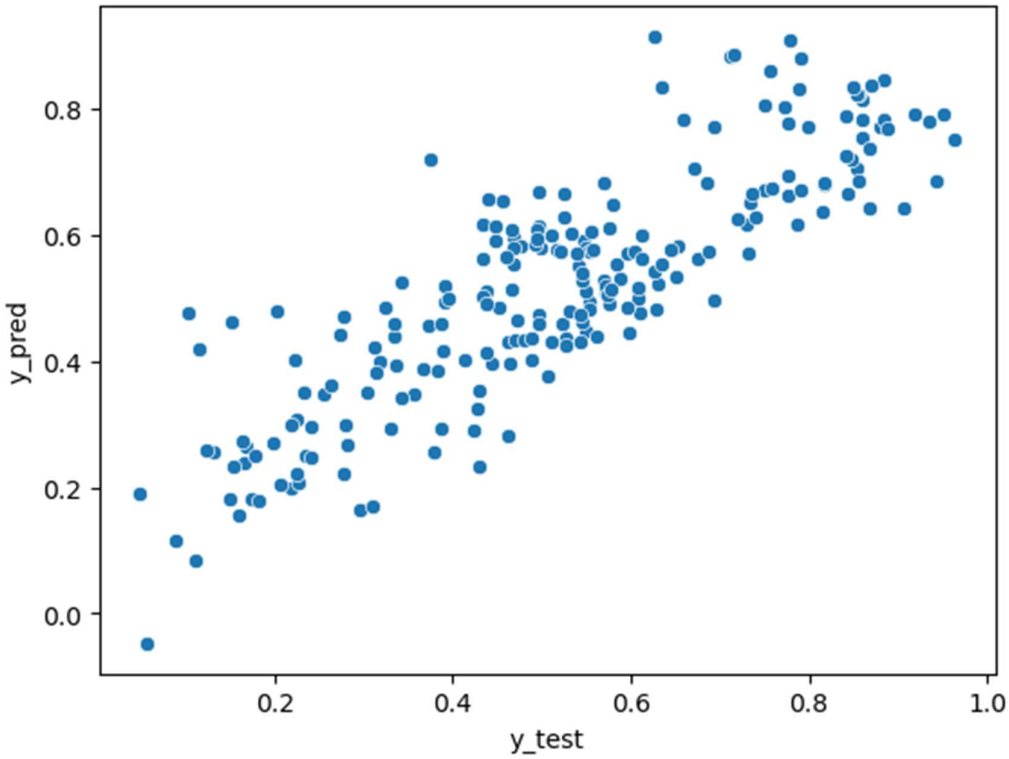
2. Error terms are normally distributed with mean zero – A histogram of the residuals was created.



3. Error terms are independent of each other – This was checked by running a scatter plot against the error terms and each value. See diagram below.



4. Error terms have constant variance – There was no heteroscedasticity observed in the plot between y_pred and y_test (actuals).



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

1. Actual temperature or feels like temperature are the biggest determining factors for cycle renting.
2. Weather conditions (snow and windspeed) are the other factors.
3. For fitting the model, yr is important condition but this is assuming company getting more popular year to year.
4. Working day is important variable but cannot be determined for overall user count. Upon further investigating with casual and registered users, it is an extremely important variable. It has negative correlation with casual users and positive correlation with registered users.

Q. Explain the linear regression algorithm in detail.

Linear regression algorithm should have the following steps.

1. In the provided data, there will be one target variable and other independent variables.
2. The target variable typically should have a linear relationship with independent variables.
3. The goal of linear regression algorithm is to determine the coefficient for independent variables and whether independent variables are significant in determining the value of the target variable. For example, does area of a house determine the price of the house and if area increases by 100 square feet, what is the increase in the price of the house. The relationship

between area and price of the house is determined by the coefficient calculated by the linear regression algorithm.

4. Perform data exploratory analysis on the data to ensure there is no null data, to understand the distribution of data and to determine whether there is correlation between various independent variables.
5. Prepare the data for regression by creating dummy variables for categorical data.
6. Split the data into train and test data. Typically apply the 70-30 or 80-20 rule to split the data into train and test data.
7. Scale the continuous variables of train data using min-max scaling or standardized scaling.
8. Separate the target variable.
9. Run the least squares regression algorithm.
10. Determine the significant variables by checking the p-value and VIF.
11. The significant variables can be determined by incrementally adding the variables or by considering all variables and eliminating the non-significant ones.
12. Plot a distribution graph to check whether the error terms on the training dataset are normally distributed with mean = 0.
13. Plot a scatter graph of the error terms to ensure that they are independent of each other.
14. Prepare the test data by scaling similar to the train data. However, the scaling should be done based on statistics of the train data.
15. Once the significant variables are determined, retrieve the coefficients of each variable and test the predicted value of the target variable and actual value on the test data set.
16. Plot a graph of the predicted target variable and actual target variable to ensure that the relationship has homoskedasticity.

Q. Explain the Anscombe's quartet in detail.

As per Wikipedia, Anscombe's quartet consists of 4 data sets with different distributions but identical descriptive statistics. The first and third plots represented a linear relationship but the regression line is different. The second plot has a non-linear relationship while in the case of the fourth plot, most of the points have same value for x variable except one point. Basically, the point of Anscombe's quartet is that descriptive statistics should be accompanied by exploratory data analysis when analysing the data. No conclusion should be drawn purely on descriptive statistics.

Q. What is Pearson's R?

Pearson's R is the correlation coefficient. It determines the relationship between 2 variables. The values are of Pearson's R are between -1 and 1. As an example, If the correlation coefficient is 0.9, it implies that if one variable increases by 1 unit, the other variable will increase by 0.9 units. It is a linear relationship. In case of negative correlation, if a variable increases, the other variable will decrease.

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

Feature scaling is required when there are variables with different ranges of values. In that case the model takes more time to converge to optimal coefficients and the coefficients could be weird for example one of them may have a value less than 1 while other could have a value of 10,000. Feature scaling converts each variable into a standard range like all values between 0 and 1.

Standardized Scaling – The variable is scaled by subtracting the mean of the dataset from the value of variable and dividing the result by standard deviation. The new dataset will have mean as 0 and standard deviation as 1.

Normalized Scaling – It is also called MinMax scaling and it is calculated by subtracting the min value of the dataset from the value of the variable and dividing the result by difference between max value of dataset and min value of the dataset.

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

When the correlation between 2 variables is 1 the VIF is infinite. The formula for VIF is (1/1-Ri2). If Ri2 is 1 then VIF is infinite.

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

As per Wikipedia, a Q-Q plot is quantile-quantile plot and is used to compare the quantiles of 2 distributions with one being observed data while other being a theoretical distribution. In case of linear regression, a Q-Q plot can be used to check if residuals are normally distributed by plotting observed data on y-axis and expected quantiles on x axis. If the points lie on a 45 degree line, it means that the reference distribution is followed otherwise observed data could be skewed.