Assignment-based Subjective Questions

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

- The year 2019 saw a lot more bike renting than the year before i.e. 2018. This could be attributed to the fact that due to COVID, its possible that the traffic was very less and many people did not wish to be cooped up at home.
- The frequency of renting of bikes was at it's highest during the Summer and Fall.
- Most of the people were willing to ride bikes when the weather we either clear or misty. This could be because clear weather conditions would be ideal for cycling, and that people could prefer a safer and easier mode of transportation during misty weather due to reduced visibility.
- More people rented bikes during a holiday. This could probably be due to the fact that they had more free time
- Renting of bikes was at its highest during the months of May, June, July, August and September.
- This is relatable to the weather seasons, considering the May, June and July usually constitute the Summer months, while August and September usually constitute the Months of Fall.

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

- One of the reasons is computational efficiency as fewer features can lead to optimization which can improve the speed of the model
- Another reason is simplicity. By dropping one dummy variable, we can simplify the model as the dropped variable can be inferred from the other variables. This makes interpretation of the categorical values easier
- By dropping a dummy variable, we can reduce complexity which in turn, leads to the prevention of overfitting

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

- Based on the pairplot, the variable 'temp' has 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?

- We plotted the histogram of the error terms and found that the distribution was a normal distribution centered around 0. This is one of the major assumptions of Linear Regression.

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

- The top 3 features that contributed significantly to explaining the demand of the shared bikes are

    - temp : representing temperature in Celcius
    - weathersit : representing the weather situation
        - The categorical variable `3` which represents `Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds` played a major role here
    - season : representing the seasons in a year
        - The `summer` and `winter` season were heavily involved in influencing the dependent variable

General Subjective Questions

Q. Explain the linear regression algorithm in detail.

- Linear Regression is a Machine Learning Algorithm that is used to find the relationship between a dependent variable and one or more independent variables
    - Dependent variable is the variable that the model is trying to predict
    - Independent variables are the variables that have the potential to influence the variable that is to be predicted.
- We basically use it to predict values of continuous variables.
- This Algorithm is used to find a linear relationship between the subject variables, i.e. the dependent variable and the independent variables.
- The intention is to fit a straight line through the datapoints in order to make predictions that are as close to the real world values as possible.
- The line usually follows the formula y = mx + c
    - y being the predicted variable
    - m being the slope of te line
    - c being the intercept
- From a multiple linear regression perspective the equation ends up being $y=\beta0+\beta1x1+\beta2x2+...+\beta nxn+\epsilon$ where x1, x2…xn represent the independent variables, β0, β1 and βn representing the co-efficients and ϵ being the error term
- One of the important steps is to ensure that we find the best fit line that explains the data as well as possible by determining the coefficients that form the so called best fit line
    - This will involve minimizing the sum of the squares of the differences between the real world variable value and the predicted variable. This method is called the Least Squared Method.
- These are the following assumptions we make when executing linear regression
    - Linearity: The relationship between the dependent variable and the independent variable is linear
    - The Observations made are independent of each other
    - The Residuals i.e. the difference between the real values and the predicted values are normally distributed
    - The Residuals have constant variance
- With Multiple Linear Regression, we select variables that can have business impact, but we also ensure that they do not contribute to multicollinearity by ensuring that they are

as independent as possible from each other. This is determined by the Correlation Matrix and the Variance Inflation Factor

- We then evaluate the model by checking the value of R^2 . The higher the value of R^2 the more effective is the model
- However, we use adjusted R^2 value in the case of Multiple Linear Regression in order to penalize having highly correlated variables in the set of selected features.

Q. Explain the Anscombe's quartet in detail.

- Anscombe's quartet is a set of four distinct datasets that have nearly identical statistics, but end up being completely different when represented through a graph.
- It demonstrates the effect of graphing data, the effect of outliers on the data and other influences that can affect the properties of data.
- Details of the four datasets that are involved:
    - Dataset I : Appears to be in a linear relationship and the datapoints seem to be in line with the assumptions of linear regression
    - Dataset II : Demonstrates a clear non linear relationship.
    - Dataset III : Shows a linear relationship similar to Dataset I, but outliers have an influence on how the linear line equation is set up.
    - Dataset IV : Has one X-value with multiple Y values, except for an outlier that affects the regression line.
- Anscombe's quartet underlines the importance of Data Visualization as we can infer their true behaviour when all the properties are the same

Q. What is Pearson's R?

- It is the measure of strength of the relationship between two variables which are continuous in nature.
- It ranges from -1 to +1 where :-
    - +1 is perfect positive correlation
    - -1 is perfect negative correlation
    - 0 means that there is no relationship between the variables
- A positive correlation means that the variables are directly proportional, meaning that an increase in one variable will lead to an increase in the other variable
- A negative correlation means that the variables are inversely proportional, meaning that an increase in one variable will lead to a decrease in the other variable and vice versa.
- The assumptions are that the relationship between the two variables is linear and that the variables are normally distributed.
- The problems with this measure is that
    - It is influenced by outliers
    - It works well only for linear relationships.
    - A high correlation does not mean that it will cause an increase or decrease in the variable (causation)

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

- Scaling is the process of transforming numerical variables so that they share a common scale.
- With this, we ensure that the model is not skewed towards variables with huge average values.
- It makes sure that no information is list as well.
- Distance based algorithms benefit from this very much.
- Benefits of scaling include Faster algorithm performance and improvement of accuracy in some of the models.
- Normalized scaling is sensitive to outliers as the minimum and maximum values are used for scaling, but that is not the case with standardized scaling
- Normalized scaling binds all values within a fixed range which is usually between 0 and 1. Standardized scaling does no such thing.
- Normalized scaling does not depend on standard deviation Standardized scaling transforms values to units of standard deviation away from the mean value.

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

- This usually happens when one variable is an exact linear combination of other variables. This is an example of perfect multicollinearity
- It will also occur if there is no variability in a particular predictor variable, i.e. the values are the same across all instances. Since VIF depends on variance, it leads to a divide by 0 scenario which is obviously going to be infinite.
- VIF becomes infinite if the information provided by one variable is completely redundant, i.e. there is perfect linear correlation between the variables.

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

- It is a tool used to compare two probability distributions by comparing their quantiles against each other.
- It is used to check if a particular dataset follows a normal distribution.
- It is useful from a linear regression perspective as we will need to see if the distribution of errors found when predicting values follows a normal distribution or not.