Assignment-based Subjective Questions

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

I can see that the following:

- Bike rental is growing year after year. The rental of the bikes in 2019 is clearly higher than 2018.
- Overall, it does not look like holidays or weekends making a difference but if you segment it by season, you can see the number of bike rentals during weekdays in winter is higher than weekends. We can see that the uncertainty of the number of bike rentals during winter and summer holidays is less compared to the rest of the seasons, we can be more certain that the number of bike rentals will be low during winter holidays and high during summer holidays
- No clear pattern regarding the effect of the weekday
- Rentals are low when the weather is rainy, and it is higher when the weather is clear.
- Winter is a low business season for the bike rentals while Summer is the highest but late Spring and Early fall are also a good time for bike sharing business due to high bike rentals.

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

Because we do not need all the columns, for example if we have a variable which has 3 dummy variables, if two of the dummy variables values are Zero then we can be sure the third dummy variable is 1, it also helps reducing the number of the columns, it also helps in reducing the correlation between dummy variables.

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

Temperature (temp) and feeling temperature (atemp) appears to be the two variables with the highest correlation among the continues variables.

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

I have used Variance Inflation Factor (VIF) to validate the assumption of Multicollinearity and dropped any feature with VIF > 5.

I have plotted the distribution of the errors (Residuals) and checked if they follow the normal distribution to validate the homoscedasticity of the residuals.

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

1- Temperature: As the temperature increase as the bike rental increases
2- Weather: If the weather is Rainy or not, bike rental increases when the weather is not rainy.
3- Year: The growth of rental bike over years is the third important feature

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Y = β0 + β1 x X1 + … + βn x Xn

This is the equation of the linear regression; the process of the algorithm is to find the best values of the βs which can provide the best fit of the regression line to the data points. This is going to be a cost function (minimization problem) where the goal is to minimize the difference between the predicted value and the actual value (sum (y_pred – y)^2 / n), this is called the Mean square Error (MSE) where we deduct predicted value from actual value and square them and sum them all and divide them by the total number of data points. The function changes the values of the βs until it reaches the minimum value for MSE. This process is called Gradient Descent, it starts with some values for the βs and then change these values iteratively to reach the minimum MSE.

2. Explain the Anscombe's quartet in detail.

It is four datasets that can have the same summary statistics, but they appear differently when it is graphed. It illustrates the importance of graphing the data before stating the analysis so you can discover any anomalies in the data such as outliers, linearity of the data and diversity of the data. Plotting these datasets in a scatter plot. One of these datasets can be well explained by a linear regression line, the second one looks nonlinear, and it is better to fit it to nonlinear regression model, the third one shows an outlier which should be removed or imputed before fitting it to a linear model, the fourth one, shows outliers and most of the data points with the same X value except one.

3. What is Pearson's R?

It is a measurement of the strength of the relationship between two variables and how much they are associated with each other. It can help in determine the magnitude and the direction of the relationship between two variables, the value of r varies between -1 and 1, as the value of r gets closer to the extreme values (1, -1) as the association between the two variables, and as the values get closer to Zero as weak the association between the two variables. Any values above Zero means that the association is positive and both values increase as the other value increase., and if it is below Zero the association is negative, and this means that each value increase as the other value decrease.

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

Scaling is part of the data preparation before fitting it to the model. It transforms the data to unified data range through normalization or standardization of the data, it can help on dealing with outliers and explain the importance of the features after fitting the model as well speed up the calculation of the algorithm.

As explained above the data features can vary in terms of magnitude and range, by performing scaling we can be sure that all variables can have the same level of magnitude by having the same range. Scaling affects only the coefficients of the model, and it does not affect any other parameters.

Normalized scaling is working in scaling the data to be between 0 and 1 using MinMax scaling, while standardization replace the values of the data by the Z score of each point and it ranges between -3 and 3.

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

Since VIF – 1/(1-r2), then this case occurs when this feature is highly correlated with another feature where r2 is equal to 1. In this case you need to drop this variable and run the VIF process again

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

Q_Q plot is a scatter plot used to compare two probability distributions by plotting their quantiles against each other. If they are equal, then the points are going to form a linear line of (y= x.

It is used to find if the distribution of the variable follows normal distribution or not. In x axes we should have the theoretical quantiles (mean =0 and SD =1), and in the y axes we can have the normal data quantiles, as more the data points fit the straight line of (y = x) as more the data is normally distributed. This can be helpful on understanding the distribution of the data and if it is normal or not. We can use the Q-Q plot to test if the data follow other distribution types as well.