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?

The categorical variables in the dataset are of three kinds. First, we have features derived from the date including year, month, and weekday. Second, contextually derived such as holiday, working day, and lastly derived from weather including season and weather situation.

- On year-over-year comparison, the total customer across the season, months, and weekdays is significantly higher in 2019 compared to 2018.
- There is strong seasonality effect on the customer opting for the BoomBikes rental services. Customer activity in the Spring is low but with positive trend. The trend continues into the summer and remain stable in the fall. In the winter, because of shape drop in temperature and adverse weather conditions, the customer activity drops. The sharp drop in the late 2019 might also related to emergence of Covid-19.
- More consumers are opting for the BoomBikes rental services over the weekend. In the 2019, there is an exception, relatively more consumers also active on Monday. The distribution is tighter in the weekend, including Friday.
- The effect of working day on the consumers using the rental service is not significant (trend plot and t-test)
- As expected, in the conducive weather conditions, where it is clear, more consumers are opting for the services. Consumer activity is low when there is light rain, and it is consistent across the two years. Note that the dataset does not have records for heavy rains.
- From the workday trend, it also observed that we can treat the variable as an ordinal features and scale while modeling. This operation will be done along with one-hot encoding Because of linear trend starts with Tuesday, it is better to encode Tuesday as the start of the week rather than Monday.

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

In the one-hot encoding (OHE), the unique labels in the categorical feature are transformed into a matrix where each column represents the presence of the label with 0 and 1 (where 1 means the label is present). So, n-unique labels will be transformed into n-column dataframe with binary representation.

Interpretation and multicollinearity are the two issues we have to work with if one of the label is not dropped. Let's take an example, A = ['a', 'b', 'c', 'a']

Without dropping the first column, the OHE results in

	label_a	label_b	label_c
0	1	0	0
1	0	1	0
2	0	0	1

3	1	0	0

With dropping the first column, OHE results in

	label_b	label_c
0	0	0
1	1	0
2	0	1
3	0	0

When label_b and label_c are 0s it is implied that label_a is present.

With dropping the first column, we avoid mutlicollinerity in the regression analysis, prevents perfect correlation between two level of a categorical variables (creates unstable model fit – correlations and thus predictions). Also reduces dimensionality.

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

Among the numberical variables (hum, temp, atemp, and windspeed), atmospheric temperature and feel temperature have the highest correlation (0.63) with the total customer count (cnt) variables.

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

Plotted distribution of residuals, error between ground truth and predicted total customer count, and observed that they are centered around zero. There exists no heterogeneity in the residual across the train and test samples.

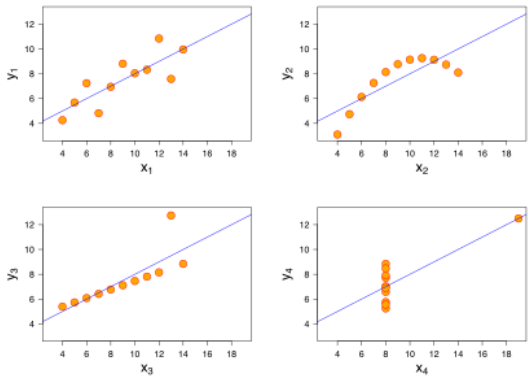
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

After ranking the features on the magnitude of their coefficients, the top 3 are
1) calendar year with coefficient +1.0253
2) season_4 (winter) with coefficient - 0.89
3) weather situation with coefficient -0.5439

General Subjective Questions

1. Explain the linear regression algorithm in detail
Linear regression is a supervised machine learning algorithm. It models the relationship between continuous dependent variable (target) and one or more independent variables (features) by fitting a linear equation. The best fit line is the one with that minimizes the sum of the square difference between the predicted and the actual values. The model adjusts the coefficients of the independent variables until best fit line is found. These coefficients help to understand the corresponding variables' important on the explaining the variance in target variable.

2. Explain the Anscombe's quartet in detail.



Ref: https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Anscombe's quartet consists of 11 data points in 4 different configurations. All the four-configurations exhibits nearly statistics with different underlying pattern as shown in the above image (adopted from Wikipedia). This objective of the datasets is to demonstrate the limitation of relaying solely on summary statistics such as mean, variance and correlation, and emphasis on performing exploratory analysis to understand the patterns and later with model evaluation.

3. What is Pearson's R?
Pearson's R measures correlation between two variables. It ranges from -1 to 1. When two variables move in perfect sync in the same direction, then there exists a perfect correlation and the value will be 1. In the variables moves in opposite direction but in sync then there exists a perfectly negative correlated and the value will be -1. When the value is 0, that means, the two variables are not correlated and movement of one variable has not influence on the other.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Scaling is a preprocessing technique in data analysis and machine learning. It transforms the values of a variable to a specific range. Scaling is performed to ensure all the independent variables are within the sample range and helps model to converge to the solution faster. Scaling is not critical in tree-based method, but with regression and neural networks it is essential. It also helps in interpreting the results including model coefficients.

In the normalized scaling, the values are bounded between 0 and 1 (by defaults), and also shift to -1 and 1. All the values from the original space will be transformed within the bound.
X_normalized = (X - X_min) / (X_max - X_min)

In the standard scaling, the values are scaled between 1-standard deviation and centered 0. The upper and lower bound can go beyond 1 and -1 respectively. Outliers can be easily identified with the scaling.
X_standardized = (X - X_mean) / X_std

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
When there is a perfect multicellularity we can observe very large or infinite VIF. If their exists perfect linear relationship between the two features, we expect one to have infinite VIF and removing it from the analysis helps the model. We observe the same when we including all the label from one-hot-encoding.

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

A Quantile-Quantile (Q-Q) plot is a graphical tool to assess where a variable follows a particular distribution, such as normal. In the linear regression, we use Q-Q plot as a diagnostics tool to investigate if the residuals follow normal distribution.