SUBJECTIVE QUESTIONS

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?

Ans:

- In case of season variable, we could find that the demand was high during the season “fall” and was least for season “spring”
- In case of year variable,2019 has higher count of users than 2018
- August month has highest count of users
- Count of users is less during the holidays

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

Ans:

- It is used as it helps in reducing extra column produced during the creation of dummy variables.
- It also reduces the correlation created among dummy variables.

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

Ans: The numerical variable “temp” has the highest correlation

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

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

Ans:

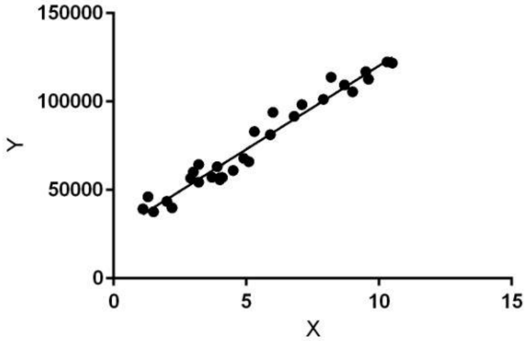
Based on final model top three features contributing significantly towards explaining the demand are:

- Temperature
- weathersit : Light Snow, Light Rain + Mist & Cloudy
- Year

General Subjective Questions

1. Explain the linear regression algorithm in detail

   Linear regression is a machine learning algorithm based on supervised learning. Run a regression task. Regression models target predictors based on independent variables. It is mainly used to find relationships between variables and predictions. Different regression models differ based on the type of relationship between dependent and independent variables considered and the number of independent variables used.



   Linear regression performs the task of predicting the value of a dependent variable (y) based on a specified independent variable (x). Therefore, this regression technique finds a linear relationship between x (input) and y (output). Hence the name linear regression. In the diagram above, X (input) is an individual's work experience and Y (output) is an individual's salary. The regression line is the best fit line for the model.

2. Explain the Anscombe's quartet in detail.

   Anscombe's quartet can be defined as a set of four data sets that are nearly identical in simple descriptive statistics, but the data sets have some idiosyncrasies that fool the regression model during construction. They have very different distributions and look different when plotted on a scatterplot. It was created to illustrate the importance of plotting graphs prior to analysis and modeling, and the impact of other observations on statistical properties. We have these four data set plots with nearly identical statistical observations that provide the same statistical information, including variance. Average of all x,y points for all four data sets.

The four datasets can be described as:

1. **Dataset 1:** this **fits** the linear regression model pretty well.

2. **Dataset 2:** this **could not fit** linear regression model on the data quite well as the data is non-linear.

3. **Dataset 3:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

4. **Dataset 4:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

3. What is Pearson's R?

The Pearson correlation coefficient (r) is the most common way to measure linear correlation. It is a number between -1 and 1 that measures the strength and direction of the relationship between two variables. The Pearson correlation coefficient is a descriptive statistic that summarizes the properties of a dataset. In particular, it describes the strength and direction of linear relationships between two quantitative variables. Interpretation of relationship strength (a.k.a. effect size) varies across disciplines, but the following table provides a general rule of thumb.

Pearson correlation coefficient (*r*) value	**Strength**	**Direction**
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables.

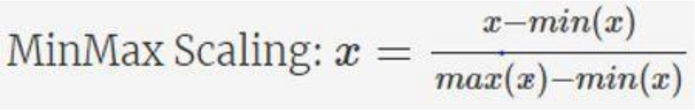
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

This is a data preprocessing step applied to the independent variables to normalize the data within a certain range. It also helps to speed up computation of algorithms. In most cases, collected datasets contain features that vary widely in size, units, and extent. Without scaling, the algorithm only considers the size and not the unit, which is incorrect modeling. To solve this problem, all variables should be scaled to have the same size.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling:

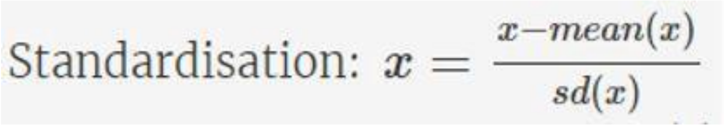
- It brings all of the data in the range of 0 and
  1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.



$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**).



$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

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

If there is perfect correlation, VIF = infinity. This shows perfect correlation between the two independent variables. A perfect correlation would have R2 = 1 and 1/(1-R2) infinity. To fix this problem, one of the variables responsible for this full multicollinearity needs to be removed from the dataset. An infinite VIF value indicates that the corresponding variable can be exactly represented by a linear combination of other variables (also denoting an infinite VIF).

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

A Q-Q plot (quantile - quantile plot) is two quantiles plotted against each other. A quantile is the percentage of a given value below that quantile. For example, the median is the quantile where 50% of the data are below that point and 50% above it. The purpose of the Q Q chart is to find out if two data sets come from the same distribution. A 45 degree angle is plotted in the Q Q plot. Points are placed on this baseline if the two datasets come from a common distribution.

If the two distributions being compared are similar, the points on the Q–Q plot lie approximately on the y = x line. If the distributions are linearly related, the points on the Q–Q plot lie approximately on a straight line, but not necessarily on the y = x line. Q-Q plots can also be used as a graphical means of estimating parameters for a family of site-scale distributions. A Q-Q chart is used to compare the shape of distributions and graphically shows how similar or different properties such as location, scale, and skewness are in two distributions.