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? (3 marks)

Ans:

- Bike rentals are highest in the fall season and least in the spring season
- 2019 demand was higher than 2018 demand, this means there is scope for business to grow in the coming years
- The median of demand numbers when there isn't a holiday is higher than the when it is a holiday, this could be attributed to the fact that people might use bike rentals for office commute
- Not a lot of difference in the weekdays when the bike is hired
- Bike rental demand is the highest for weather 1 which is Clear, Few clouds, Partly cloudy, Partly cloudy
- From the box plots of Months against demand, we see the fall months of Aug, Sep, Oct have the highest demand

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

Ans:

Drop_first = True will drop the reference variable from the list of dummy variables. This is done to get m-1 dummy variables from m leveled categorical variables. This is useful to reduce the dependency of calculating again all factors like p value, VIF etc for a new variable which can be easily denoted by other dummy variables combination. This will then reduce the additional correlation between numerical variables

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans:

Atemp has the highest correlation with cnt (0.63)

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans:

We checked the errors are normally distributed and have a mean sum =0 by plotting a distplot using seaborn library.

We also checked whether errors are not related. We also plotted a heatmap of all numerical variables which helped us understand the correlations between the numerical variables. These came out to be pretty less so hence with confidence we can say we don't have multi-collinearity within the predictors

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans:

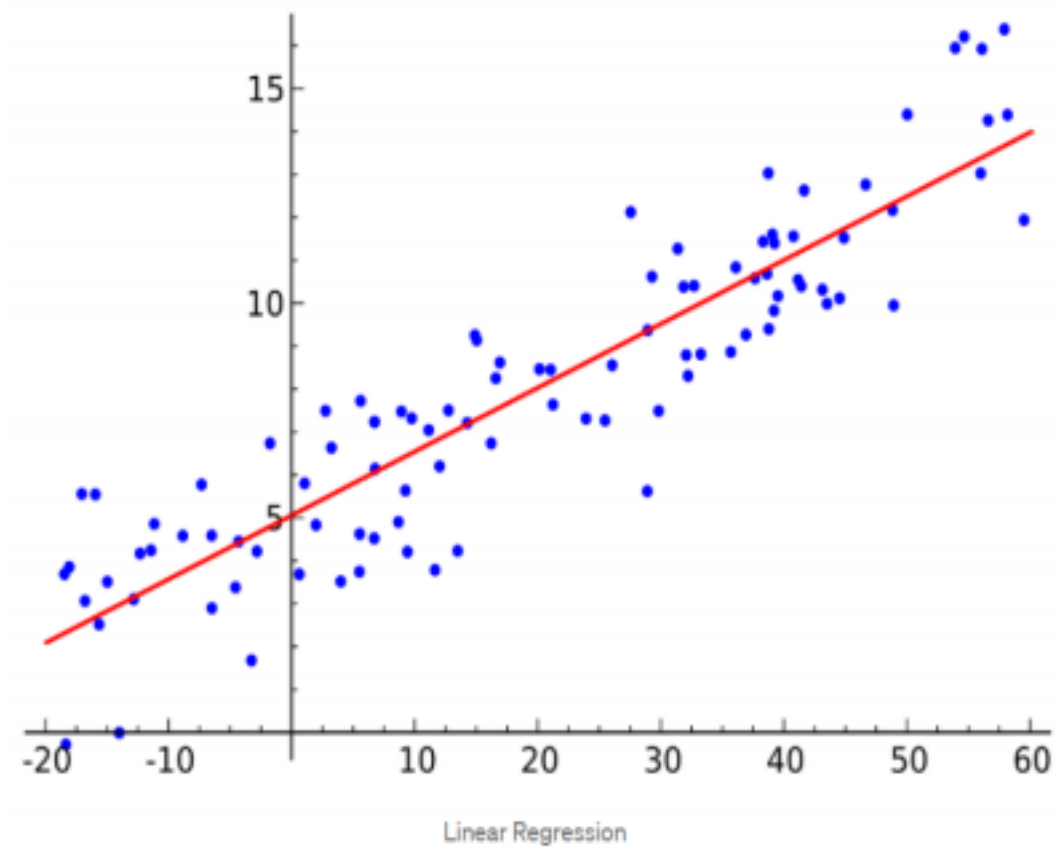
Based on my model- Temp, yr_2019 (dummy variable for yr with 2019) and weathersit_weather3 (dummy variable with 3 as weathersit value) impact the most on the model

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Ans:

Regression is used to study the relationship between 2 variables (generally dependent and independent). A linear regression is used when we know the relationship between these variables is linear in nature. We use this concept to devise an algorithm for machine learning which is used to predict the values of certain independent variables provided the relationship remains linear all through. Now when we have 1 independent variable then its called simple linear regression but when we have multiple independent variables which drive the dependent variable then we call the algo a multiple linear regression. We use a train test split to split the model and then try to fit the OLS algorithm(Ordinary least square algorithm) over the train set and then validate the model on the test set. OLS method is basically those optimized parameters for each predictor used where the sum of squares of error between predicted and actual is the least.



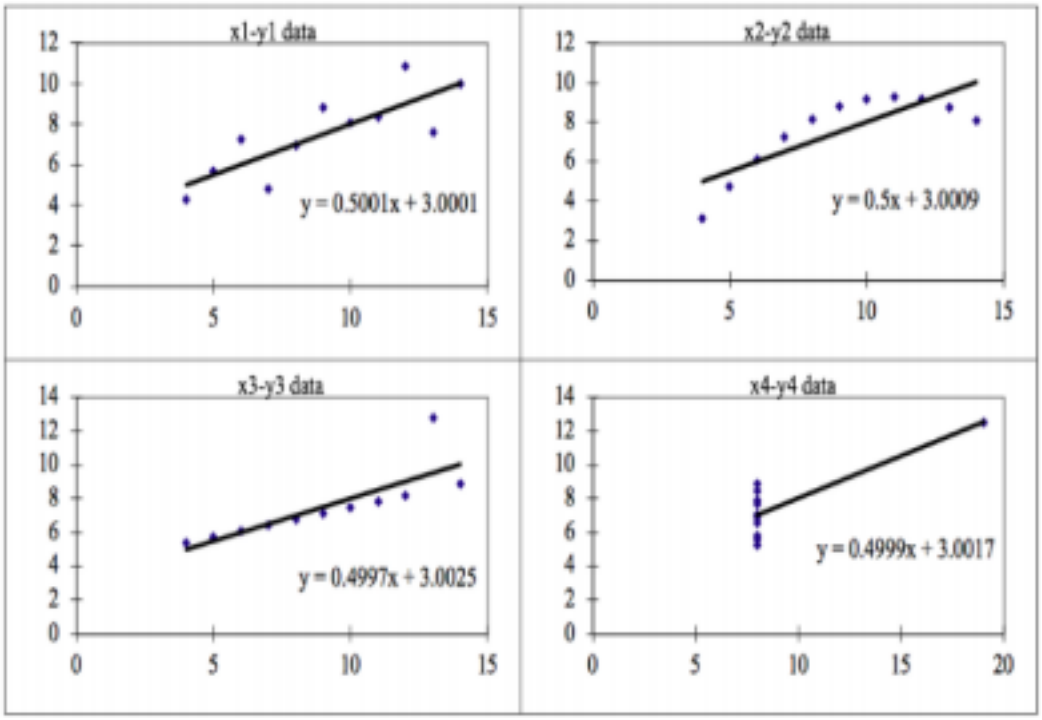
Linear Regression

2. Explain the Anscombe's quartet in detail. (3 marks)

Ans:

Anscombe's Quartet is a group of 4 datasets which are identical but still have differences. Even though their statistical measures such as mean, variance etc will be same there would substantial differences if we plot them onto the graph visually. If we try to plot linear regression model on these data sets we might get same coefficients for all these datasets.

This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.



3. What is Pearson's R? (3 marks)

Ans:

Correlation coefficients are used to measure how strong a relationship is between two variables. Pearson's R is a way to denote correlation between 2 variables numerically. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down and are inversely proportional with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson's correlation coefficient varies between -1 and +1 where: r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction) and r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions) r = 0 means there is no linear association

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans:

Feature scaling (also known as data normalization) is the method used to standardize the range of features of data. Since, the range of values of data may vary widely, it becomes a necessary step in data pre-processing while using machine learning algorithms.

If feature scaling is not done, then a machine learning algorithm tends to weigh greater values as higher and consider smaller values as the lower values, regardless of the unit of the values.

In scaling (also called min-max scaling), you transform the data such that the features are within a specific range e.g. [0, 1]

There are 2 ways to scale the numerical variables:

A) Standardization is scaling technique where the values are centered around on the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

B) Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution. To normalize your data, you need to import the Min-Max Scalar from the sklearn library

- Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization. To standardize our data, you need to import the Standard Scalar from the sklearn library

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans:

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to multi-collinearity amongst the predictors. The higher the value, the greater the correlation of the variable with other variables. Values of more than 4 or 5 are sometimes regarded

as being moderate to high, with values of 10 or more being regarded as very high. These numbers are just rules of thumb; in some contexts a VIF of 2 could be a great problem.

It is calculated by taking the ratio of the variance of all a given model's betas divide by the variance of a single beta if it were fit alone.

If there is a perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

(3 marks)

Ans:

The quantile-quantile (Q-Q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. It plots quantiles for both the datasets and checks if both are similar. This is done post sorting the data for both datasets

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences.

A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions