Ashvini Chandra

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: I have plotted the categorical variables with the target variables on boxplot and has inferred following effect on target:

- Season : 3: fall has the highest demand for rental bikes.
- I see that demand for next year has grown.
- Demand is continuously growing each month till June, September month has highest demand. After September , demand is decreasing.
- When there is a holiday, demand has reduced.
- Weekday is not giving a clear picture about demand.
- The clear weathersit has highest demand.

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

Ans : Drop_first = true is important to use , as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

If we do not drop one of the dummy variables created from a categorical variable then it becomes redundant with dataset as we will have constant variable(intercept) which will create multicollinearity issue.

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

The feature temp has highest correlation. It is very well linearly related with the target cnt.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
I have checked the following assumptions:

Error terms are normally distributed with mean 0.
Error Terms do not follow any pattern.
Multicollinearity check using VIF(s).
Linearity Check.

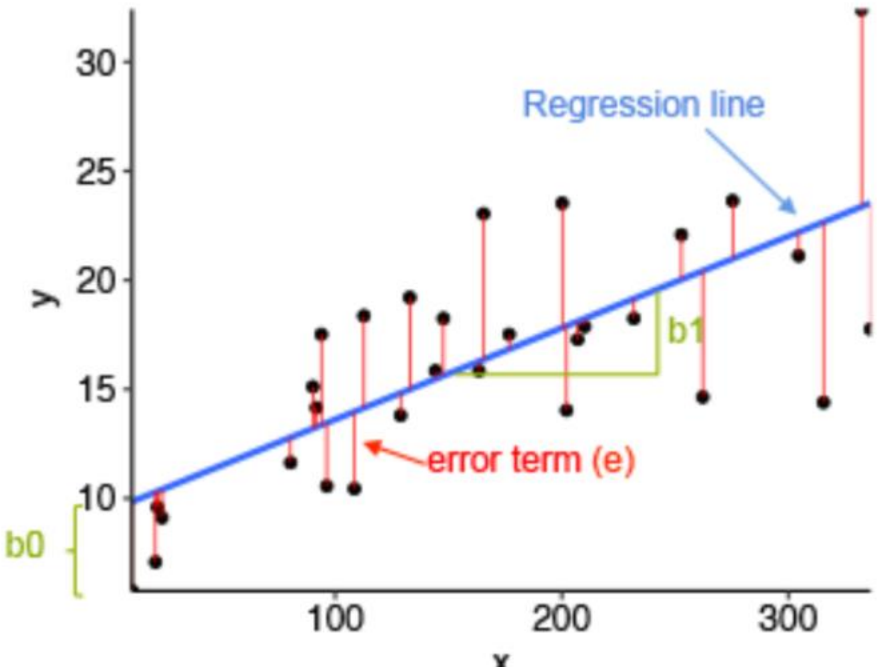
5 Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes

Features holiday, temp and season hum are highly related with target column, so these are top contributing features in model building.

General Subjective Questions

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

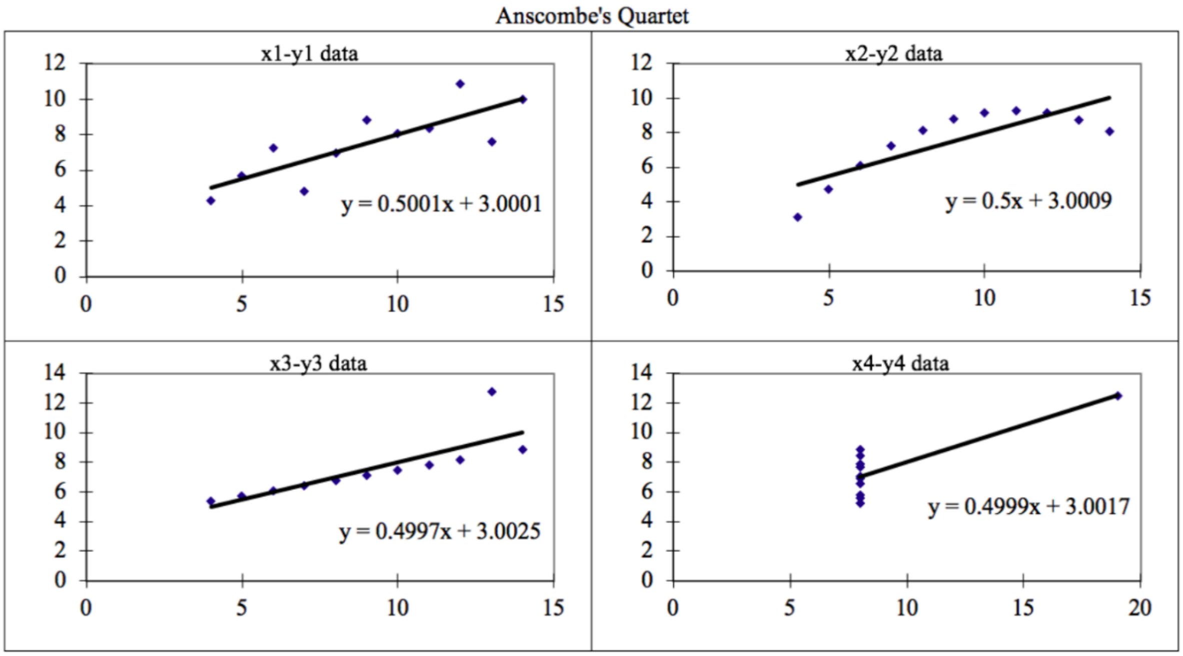
   Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.



Here x and y are two variables on the regression line,
b1 = slope of the line.
b0 = y-intercept of the line.
 = Independent variable from dataset.
Y= Dependent variable from dataset.

2. Explain the Anscombe’s quartet in detail.

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x, y) points.



The four datasets can be described as:
Dataset 1 : this fits the linear regression model pretty well.
Dataset 2: this could not fit linear regression model on the data quire well as the data is non-linear.
Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model.
Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model.

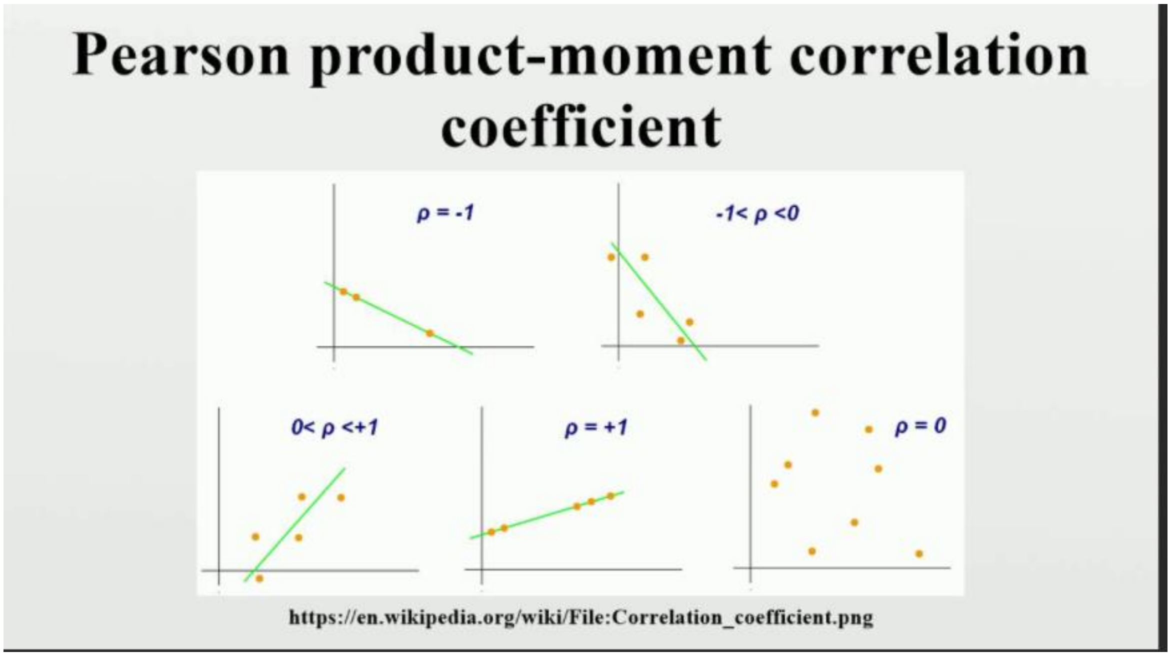
3 What is Pearson’s R? (3 marks)

Ans – Pearson’s r Is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive.

Pearson’s r measures the strength of the lieanr relationship between two variables.

Peason’s r always between – 1and 1.

If data lie on a perfect straight line with negative slop then r = -1.



Pearson product-moment correlation coefficient

https://en.wikipedia.org/wiki/File:Correlation_coefficient.png

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

Ans – Scaling is a method to normalize the range of independent variables. It is performed to bring all the independent variables on a same scale in regression. If scaling is not done, then regression algorithm will consider greater values as higher and smaller values as lower values.

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.

Ex- Weight of a device = 500 grams and weight of another device is 5kg.In this example machine learning algorithm will consider 500 as greater value which is not the case.And I will do wrong prediction.

Machine learning algorithm works on numbers and not units. So, before regression on a dataset it is a necessary step to perform.

Scaling can be performed in two ways: Normalization : It scales a variable in range 0 and 1.
Standardization : It transforms data to have a mean of 0 and standard deviation of 1.

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

Ans – When there is a perfect relationship then VIF = infinity whereas if all the independent variables are orthogonal then to each other than VIF = 1.0. Means if a variable is expressed exactly by a linear combination of other variable then it is said that VIF is infinite.

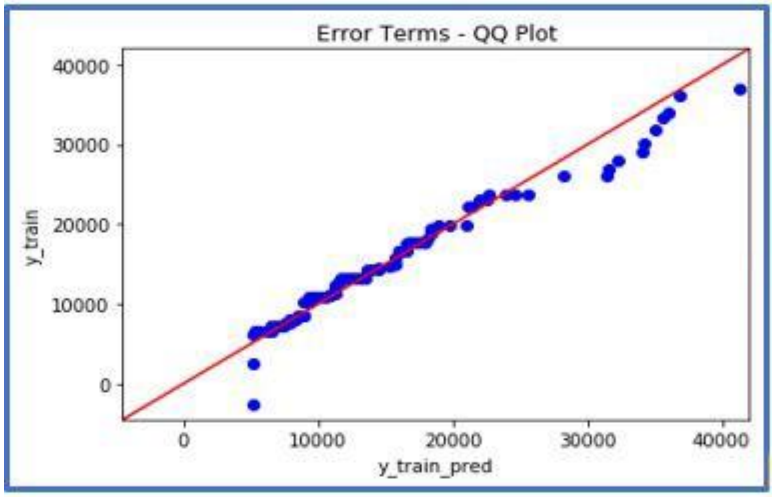
6 What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.
This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

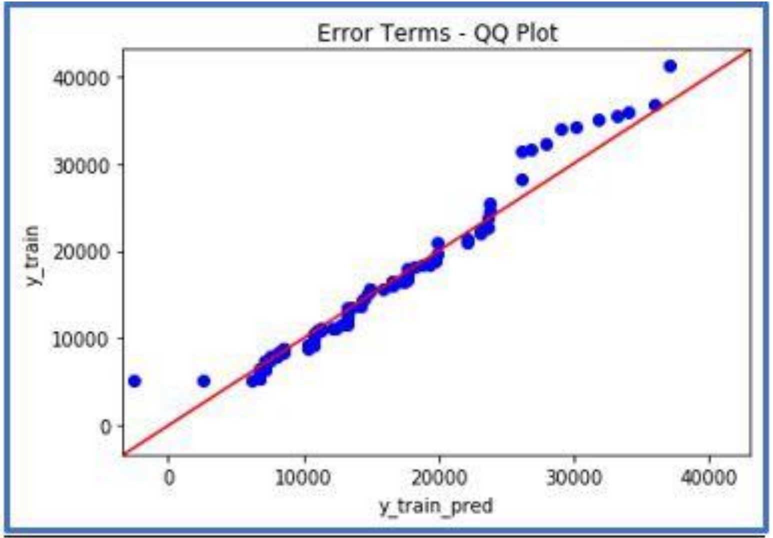
Few advantages:
a) It can be used with sample sizes also
b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:
If two data sets —
i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behavior

Interpretation:
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
Below are the possible interpretations for two data sets.
a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis

Python:

statsmodels.api provide qqplot and qqplot_2samples to plot Q-Q graph for single and two different data sets respectively.