Assignment-based Subjective Questions

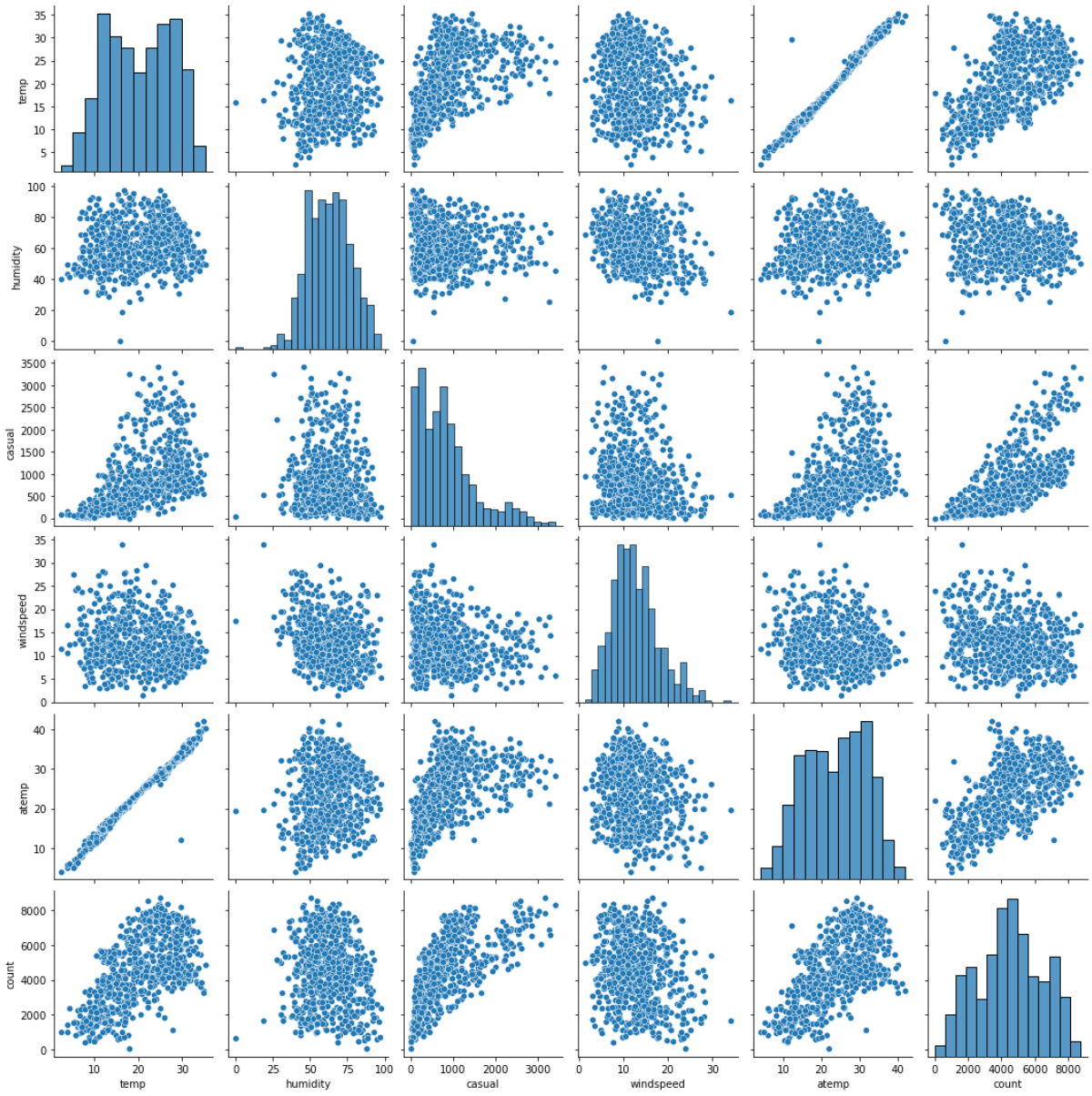
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

   Season- The Barplot says that Season fall has more number of bike rentals ,followed by Summer, Winter , Spring. During Spring season the bike rentals are very less.
   Year – The Barplot says that the bike rentals are more in year 2019 than bike rentals in year 2018.
   Month- The Barplot says that bike rentals are more during September month and less during January month. And between May till October bike rental counts are more than 5000 count.
   Weekdays- The Barplot between bike rentals and count says that bike rentals are more during weekdays and less during Sundays, from this it is clear that people are more interest during working days.
   Weathersit- The Barplot says that people are less interested in bike rentals during Heavy rain/ Ice Pallets/Thunderstorm/Mist/Snow/Fog. And more interested during Clear/Partial Cloudy days.

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

   Drop_first=True is important because it helps in reducing the extra column created during dummy variable creation and it is also important because it reduces the Correlations created during dummy variable creation. Due to which redundant will be reduced during dummy variable creation.

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



Temp and Atemp is correlated with the target variable Cnt.

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

Validated the assumptions of Linear Regression based on below points:

- Multicolinearity Check
- Normality Error terms
- Linear relationship validation

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

   Top three features contributing are:
   - Temperature
   - Season (Summer)
   - Month (September)

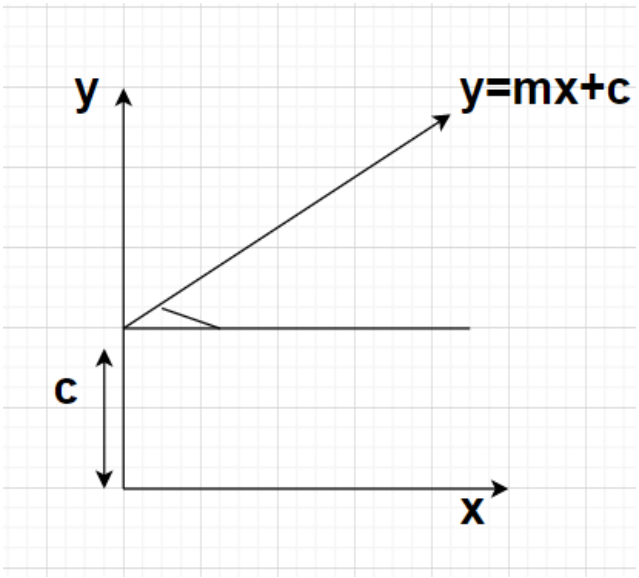
General Subjective Question

1.**Explain the linear regression algorithm in detail.**

Linear Regression is a Machine Learning algorithm based on Supervised Machine Learning Model. Linear regression is applicable for continuous target variable. And this is mainly used for predictive purposes.

"y=mx+c" is the standard equation for representing linear regression

Where, c= intercept line, x= independent variable, Y= dependent variable, m= signifies how strong the relation between x and y variables.



There are two types of Linear Regression, they are:-

- Simple Linear Regression
- Multiple Linear Regression

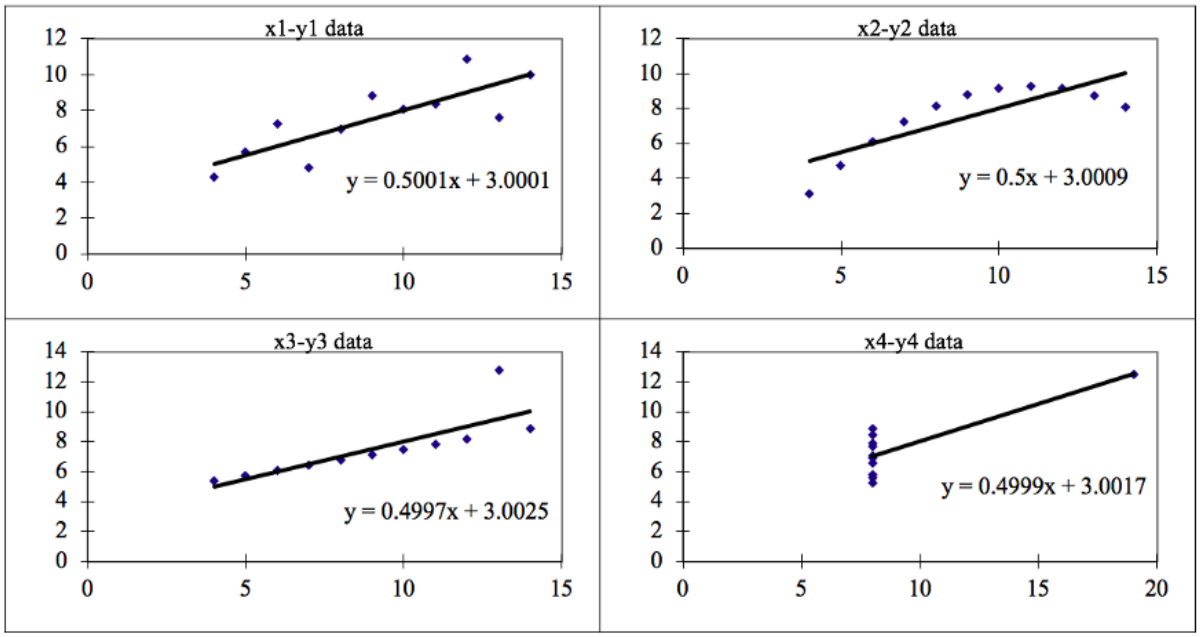
Simple Linear Regression is used when the dependent variable is predicted using only one independent variable. SLR in short.

Multiple Linear Regression is used when the dependent variable is predicted using multiple independent variables. MLR in short.

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. Moreover they have very different distributions and appear differently when plotted on scatter plots.

It was constructed by Francis Anscombe to illustrate the importance of plotting the graphs before analysing the model and effect of other observations on statistical properties.



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

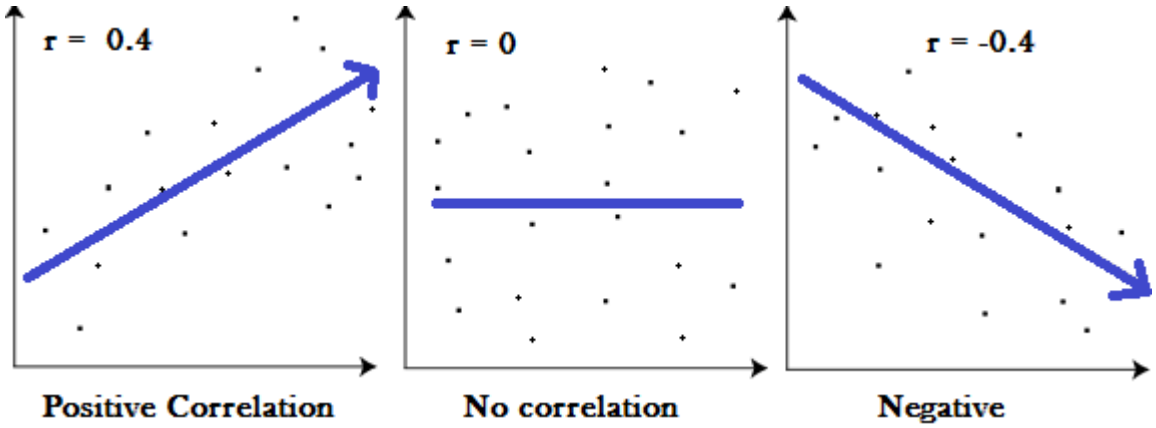
3. What is Pearson's R?

The Pearson correlation coefficient that is also known as Pearson product-moment correlation coefficient or Bivariate correlation. 'r' is a measure to determine the relationship between two quantitative variables and the degree to which the two variables coincide with one another. Numerical value lies in between -1 and +1. Based on Pearson's value there are 3 types, they are: Positive Correlation, Negative Correlation, No Correlation

Positive Correlation: Relationship between two variables in which both move in same direction.

Negative Correlation: Relationship between two variables in which increase in one variable will decrease in another variable.

No Correlation: No relationship between the two variables.



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

   Feature Scaling is a method used to Normalize or Standarize the range of independent variables of data. It is mainly performed during Data preprocessing stage to deal with varying values. If feature scaling is not done then machine learning model will tends to higher values assuming it as more weight and consider lower values as lower weight , irrespective of the units of values.

   Differences between Normalization and Standarization Scaling:

   Normalization:- Minimum and Maximum values are used for scaling and they are scaled between [0,1] or [-1,1]. Outliers are treated much better here. Scikit learn provides a transformer called MinMaxScaler for Normalisation.

   Standarization:- Mean and Standard Deviation is used for scaling. And it is not bounded to a certain range. Outliers are not much treated compared to normalization. Scikit learn provides a transformer called StandardScaler for Standarization.

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

   Variance Inflation Factor in short VIF , where this gives how much the variance of the coefficient estimate is being estimated by collinearity. VIF=(1/1-R^2).

   When VIF is infinite then it shows that there is a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-1) tends to infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

   Q-Q Plots also knows as Quantile-Quantile plots. 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. And also, it helps to determine if two data sets come from populations with a common distribution.

   Importnace of Q-Q plot is that :

   It helps in understanding if two data sets are from populations with a common distribution.

   It helps in understand that two datasets have common location and scale.

   It helps in understand that two datasets have similar distributional shapes.

   It helps in understand that two datasets have similar tail behavior.