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)

   Unlike numerical variables, Categorical variables require special handling. Categorical variables are converted into series of variables, we would create k-1 new variables where k is the number of levels of the categorical variables.

   In the Bike sharing example our inference is as below

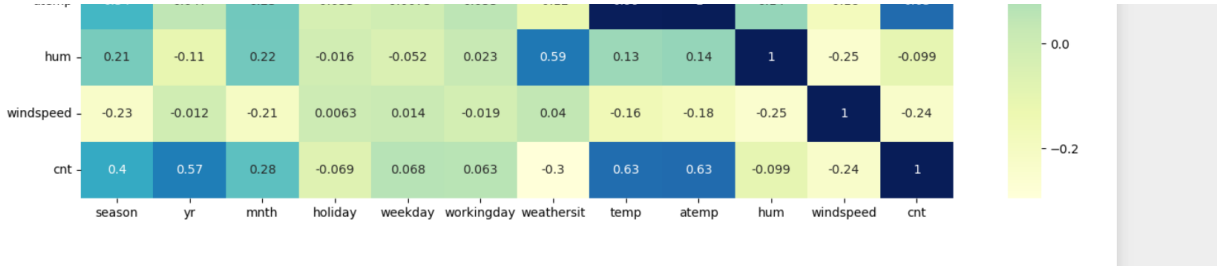
   - Identified 'season','yr','mnth','holiday','weekday','workingday','weathersit' as categorical variables.
   - It's noticed out of four seasons fall has high number of count of rental bikes **'cnt'**
   - From the boxplot we can conclude that year 2019 has more bike rentals.
   - Rentals have increase from March till October
   - Day of the week has no effect on bike rentals, the rentals remain the same.
   - We can also notice on clear weather bike rentals increase.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
   We would create k-1 new variables where k is the number of levels of the categorical variables, If we don't apply drop_first=True, we end up creating k variables instead of k-1.

   One can always achieve the n variation in n-1 combinations.
   This will reduce multi collinearity between dummy variables.

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

   Considering **'cnt'** is the target, From pairplot and after dropping the **['instant','dteday','casual','registered'] as irrelavant**, It's evident the variable temp and atemp has the highest correlation to the target variable with .63 .



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

The assumptions are a formal check to ensure that the linear model built gives us the best results

The assumptions are as follows

1: Multi collinearity:

We have decided to pickup the variables with VIF values < 5, any variables with higher VIF can be dropped, and variables with high correlation can be merged to create one variable.

2: Residuals must be normally distributed.

Residual should be a normal distribution, meaning the error distribution mean should be equal to 0.

3: Linear Assumption:

The relation between dependent variables and the independent variables should be linear in fashion. All the independent variables have p values < 0.05 which means all the independent variables are significant.

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

Based on the final model, The top 3 features are

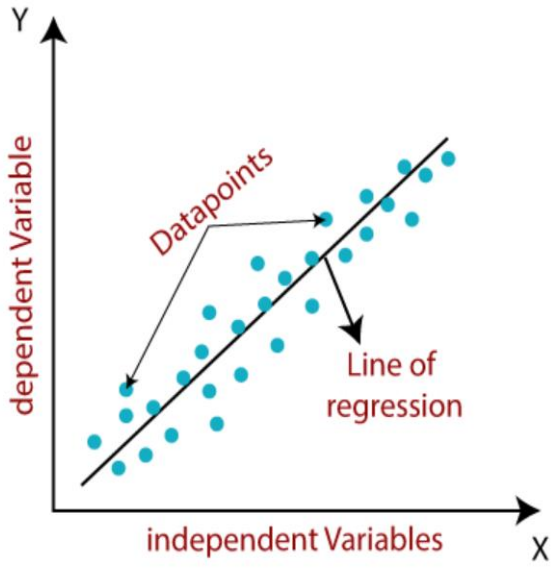
- Temp: has a coefficient of 0.5754, meaning increase in temp will increase bike share cnt
- Yr: has a coefficient of 0.2561, meaning increase in year will increase bike share cnt, as one can notice 2019 had higher bike share cnt
- `Light Snow`: has coefficient of -0.2638, negatively co-related, increase in snow will de crease the bike share cnt

General Subjective Questions

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

Linear regression explains the relationship between dependent and independent variables using the most relevant straight line.

Linear regression equation can express as below



$y = \beta_0 + \beta_1 x + \varepsilon$

where y = Dependent variable
X = Independent variable
β_0 = Intercept of the line
β_1 = slope of the line
ε= random error

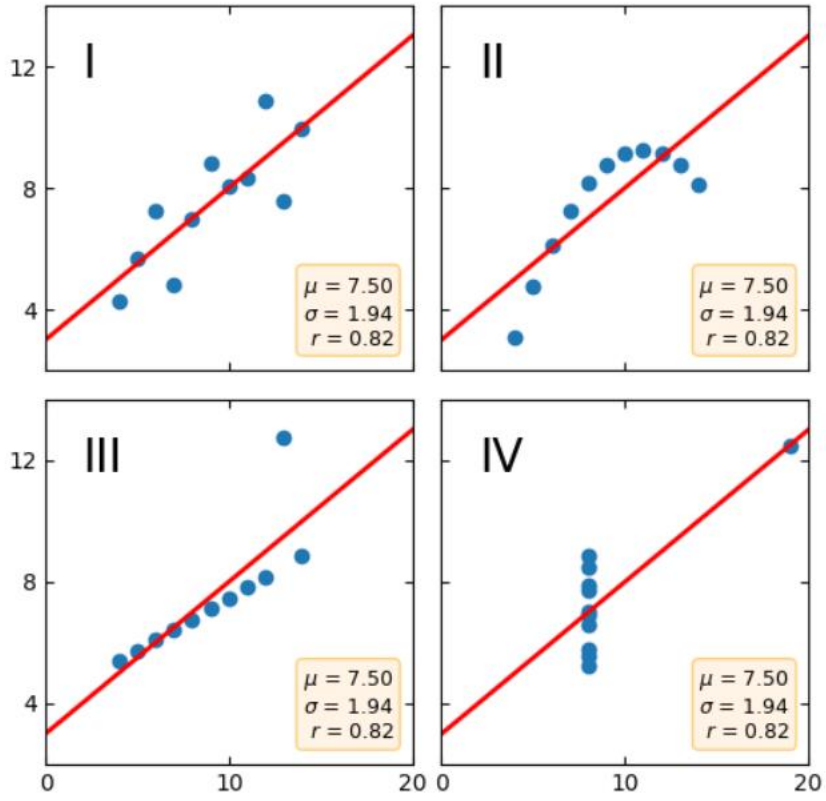
Assumption for Linear Regression Model

There are set of assumption to be proved this linear model is the best fitting one

1. There should be linear relationship between dependent and independent variables.
2. Independence: The observations in the dataset are independent of each other. This means that the value of the dependent variable for one observation does not depend on the value of the dependent variable for another observation.
3. The errors in the model are normally distributed.
4. Multi collinearity: highly correlated variables should be reduced removed or reduced to one variable.

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

Anscombe's quartet shows that multiple data sets with similar statistical properties can still be vastly different from one another when graphed.

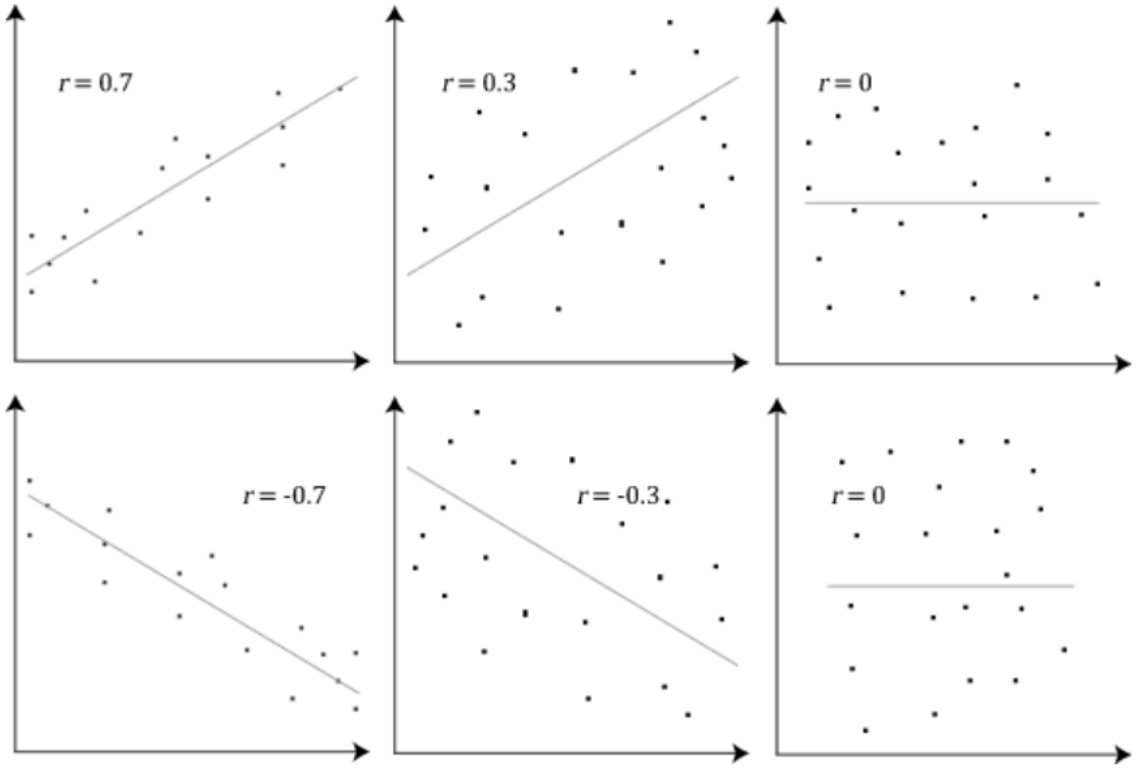


- The first graph has almost a well-fitting linear model
- The second graph the variables are not distributed normally
- The third graph is linear, but the regression can be thrown off with an outlier
- Fourth graph shows one outlier is enough to produce higher correlation coefficient.

The Anscombe's quartet emphasizes the importance of visualization of the data.

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

   Pearson's R or Pearson's correlation coefficient is the statistics that measures the statistical relationship between two continuous variables. It is the ration between the covariance of two variables and the product of their standard deviation. It's a normalized measurement, the result will always be between -1 and 1.



- A value greater than 0 indicates a Positive correlation, meaning increase in the value of a variable will means increase in the other
- A value lesser than 0 indicates a Negative correlation, meaning increase in the value of a variable will means decrease in the other
- A value equal to 0 no association between the variables

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

   Scaling a data processing step, where we transform the large values of data to fit in a specific range. It is the idea of bringing all the variables into the same level of magnitude.

   Why scaling is performed:
   Provided the values of variables are closer to each other chances of algorithms getting trained data will evaluated faster and trained well.
   If the values in any given data set vary a lot or far from each other, Scaling is used to generalize the data points so the distance between them is lowered. We need to scale for two reasons

   1. Ease of interpretation
   2. 2. Faster convergence for gradient descent methods

normalized scaling and standardized scaling:

Normalized Scaling:
Normalization id a method of rescaling the data between the range of 0 to 1 using the min and max values in the data, so that the data points can become closed to each other.

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardized scaling:
Here the scaling is achieved in such a way that the mean is zero and standard deviation is 1. The data points are rescaled by ensuring after scaling the data points are in curve shape.

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

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

   VIF (variance inflation factor) is a measure of amount of collinearity in regression analysis. As the name suggests, the variance is inflated with existence of multicollinearity. **If there is a perfect co-relation, then the value is infinity**.

$$\mathrm{VIF_i} = \frac{1}{1 - R_i^2}$$

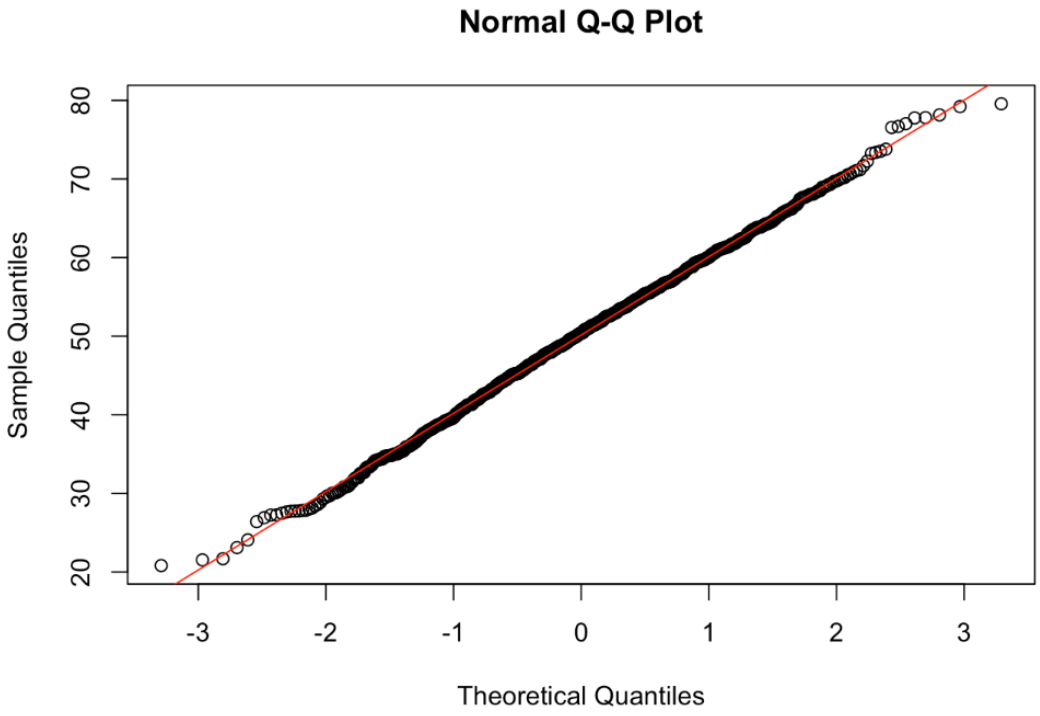
   An infinite value of VIF for an independent variable indicated that it can be perfectly predicted by the other variables, this happens when the R2 value reaches 1.

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

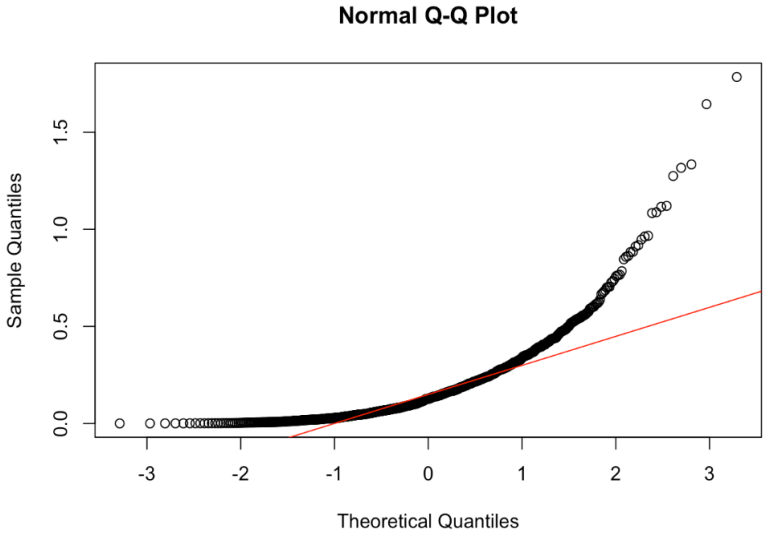
   Q-Q plot (Quantile-Quantile plot) tool is used show if the two data sets come from same distribution. Plotting the first data set along x-axis and second data set along y-axis is how the plot is constructed.

   Importance and Interpretation of Q-Q plot

   - If the data set is normally distributed, then the Q-Q plot will show a straight line



- When the data is more at the ends and less data at the middle, it's referred as heavy tailed.
- A normal distribution of data if one notices a deviation out of straight line means the data is skewed



  - Right skewed is also referred as positive skew
  - Left skewed is also referred as negative skew