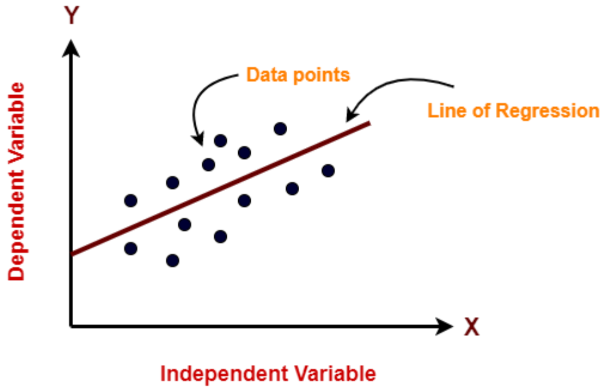
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
    - Summer is the season with high demand for bicycles. May be the summer vacation month, the climate is warm and crowded with tourists.
    - In spring, people have less need to use bicycles
    - Particularly in winter, the rate of bicycle use fluctuates greatly
    - There's no data on Heavy Rain + Ice Pallets + Thunderstorm + Mist so I can't study and analyze it. Event in nice, cloudy weather there is a higher percentage of bike riders than on a light rainy day.
2. Why is it important to use **drop_first=True** during dummy variable creation?
    - drop_first = True is important to use, as it helps to reduce the extra column created during dummy variable creation.
    - Reduces the data size and it reduces the correlations made between dummy variables.
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation
   with the target variable?
    - The 'temp' variable has the highest correlation with the target variable.
4. How did you validate the assumptions of Linear Regression after building the model on the training set?
    - Verify whether the error term is normally distributed or not.
    - Check the multicollinearity of the features.
    - Build a relationship validation model.
    - Residual features to find the best model.
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
    - Temp
    - Winter
    - Jun

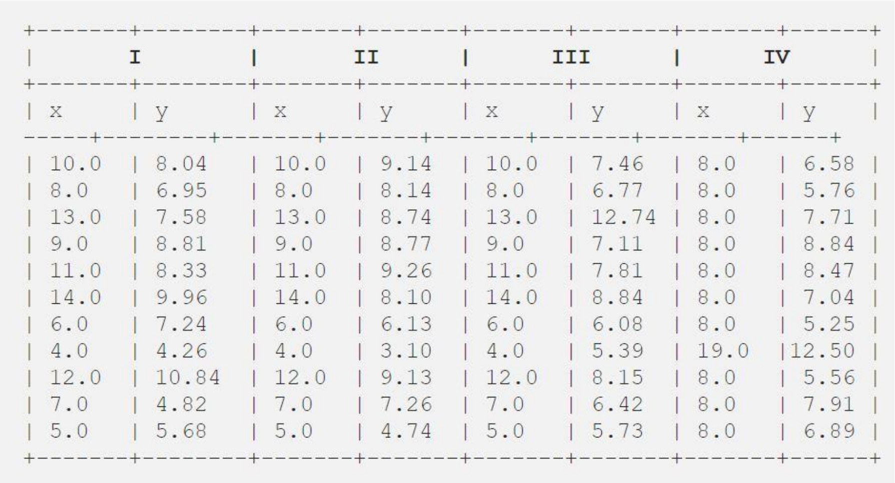
General Subjective Questions

1. Explain the linear regression algorithm in detail?
    - Linear regression is an algorithm in the field of machine learning, using supervised data learning. (labeled). The algorithm works on a dependent variable (target) based on the given variables independently. Algorithm to find a linear relationship between dependent and independent variables



[source]

2. Explain the Anscombe's quartet in detail.?
    - ADD is a quartet of data with simple statistical properties that are similar but very different when plotted on a graph.
    - Each dataset consists of eleven (x, y) points. They were formulated in 1973 by statistician Francis Anscombe to demonstrate both the importance of charting data before analyzing it and the influence of outliers on statistical properties.

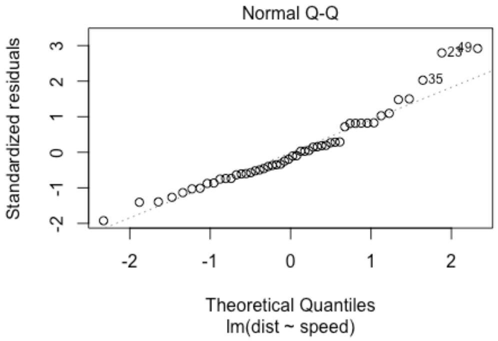


I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

[source]

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 closer to 1, the stronger the positive correlation, and the closer to -1, the stronger the negative correlation.
    - r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
    - 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
    - r > 0 < 5 means there is a weak association
    - r > 5 < 8 means there is a moderate association
    - r > 8 means there is a strong association

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling
   and standardized scaling?
    - Scaling is to normalize data to a certain type to avoid data loss, to fully exploit data.
    - Min-Max Scaling is used to transform features into a similar scale. New points are calculated as: X_new = (X - X_min) / (X_max - X_min)
    - This scales the range to [0, 1] or sometimes [-1, 1]. Geometrically speaking, the transformation reduces the n-dimensional data to an n-dimensional unit superblock. Normalization is useful when there are no exceptions because it cannot deal with them. Usually, we will divide by age, not income because only some people have high income but the age is nearly equal.
    - Normalization or Z-score normalization is the transformation of features by subtracting the mean and dividing it by the standard deviation. This is commonly known as the Z-score : X_new = (X - mean) / Std
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
    - If all the independent variables are orthogonal to each other, then VIF = 1.0. if VIF = infinity means that the values of VIF of the variables have a very large correlation (perfect correlation)
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.?
    - Histogram (Q-Q), is a graphical tool to help us evaluate whether a data set makes sense to come from some theoretical distribution such as a Normal, Exponential, or Homogeneous distribution. In addition, it helps to determine if two data sets come from populations with a common distribution.



        - [source]
    - Suppose you have some observations and you want to check if they come from a normal distribution. You can normalize them (central variance and mean mean 1) and then 'percent match' with a normal normal distribution. You can then plot your score against a perfect fit percentile.