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?
    - The demand has increased in the fall season while there is also positive demand in summer season.
    - The demand is increased in 2019 than in 2018 which is showing that the demand is increasing yearly.
    - The bar plot of month vs cnt showing that the demand is more between the month of april and September which aligns with the season bar plot.
    - There is more demand in working day rather than holiday which shows that job goers are one of the major demanders.
    - There is high demand when the weather is moderate and good.

2. Why is it important to use drop first=True during dummy variable creation?

- As per theory if there are n dummy variables then (n-1) variables are enough to explain the all n variables, so it is important to drop first=True also it reduces the complexity and storage of the data set.

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

- Looking at the pair plot temp 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?

- By plotting distplot we can conclude that the error terms as normally distributed or not.
- By plotting scatter plot of error terms we can see is there any pattern forming between the error terms or is there any other cluster forming or not.

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

- Based on the model top 3 features are temp, season and weather.

1.Explain the linear regression algorithm in detail.

- Process of estimating linear relationship between variables
  - Focus: Relationship between dependent variable and independent variable.
- Explains change in dependent variable with change in predictor variable.
  - Simple Linear regression:- Changing only one variable at a time.
  - Multiple linear regression:- Changing multiple variable at a time.
- Uses:- Forecasting and prediction.
- Shows Correlation not causation.
- A form of a parametric regression.
- Linear regression guarantees interpolation of data and extrapolation of data.

2. Explain the linear regression algorithm in detail.

- Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

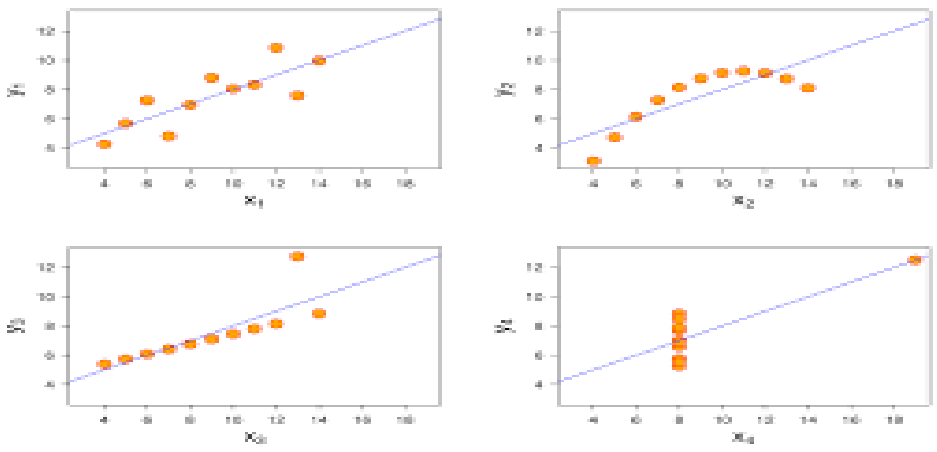
	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

The summary statistics show that the means and the variances were identical for x and y across the groups:

• Mean of x is 9 and mean of y is 7.50 for each dataset.

• Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset

• The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



Dataset I appears to have clean and well-fitting linear models. • Dataset II is not distributed normally. • In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier. • Dataset IV shows that one outlier is enough to produce a high correlation coefficient. This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

3. What is Pearson's R?

- 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. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.
- The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

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

- Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values. Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

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

- If there is 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.
- When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this 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.

- The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.
- Use of Q-Q plot: A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. 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.

- Importance of Q-Q plot: 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. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests