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?

`Inferences:`

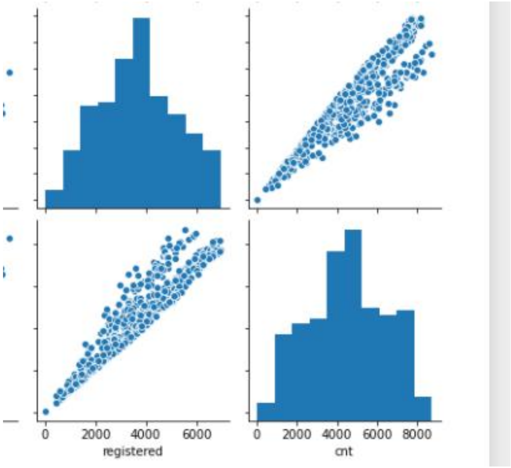
- Bike rentals are greater in fall season.
- There is a clear growth in bike rentals from inaugural year (2018) to the next year.
- There is little change in rentals on whatever day it was, the count is almost same on all days. And there is no parity between working day and non-working day, since the rentals are same.

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

Ans). It is important to give drop_first = True, because if we have **'n'** levels for a categorical variable, then 'n - 1' dummy variables is sufficient to represent 'n' levels.

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

Ans). Registered Column has the highest correlation with target variable.



P.T.O

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

Ans). With the help of p-values and VIF (variance inflation factor).

- We need to see if any p-values within the limits (i.e., p-value < 0.05)
- We need to see if VIF is within the limits (i.e., VIF < 5)

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

Ans). Temperature, season, weather situation are the top three factors contribute to the demand of shared bikes.

General Subjective Questions

1). Explain the linear regression algorithm in detail.

Steps we follow:

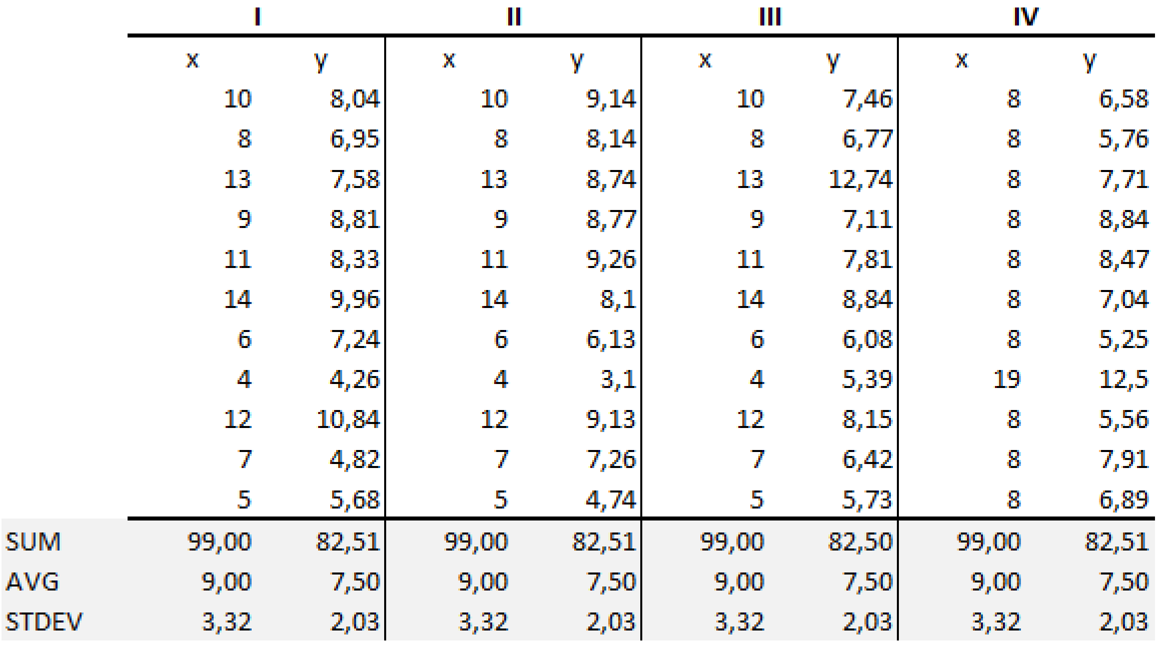
- Reading, understanding and visualizing the data.
- Preparing the data for modelling (train-test-split)
- Training the data.
- Residual Analysis
- Predictions and evaluation the test set

Reading, understanding and visualizing the data: Read the data and visualize the categorical variables with the help of boxplots and numerical variables with pair-plots. Drop the columns with high correlation.

2. Explain the Anscombe's quartet in detail.

Ans).

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 s tatistics.

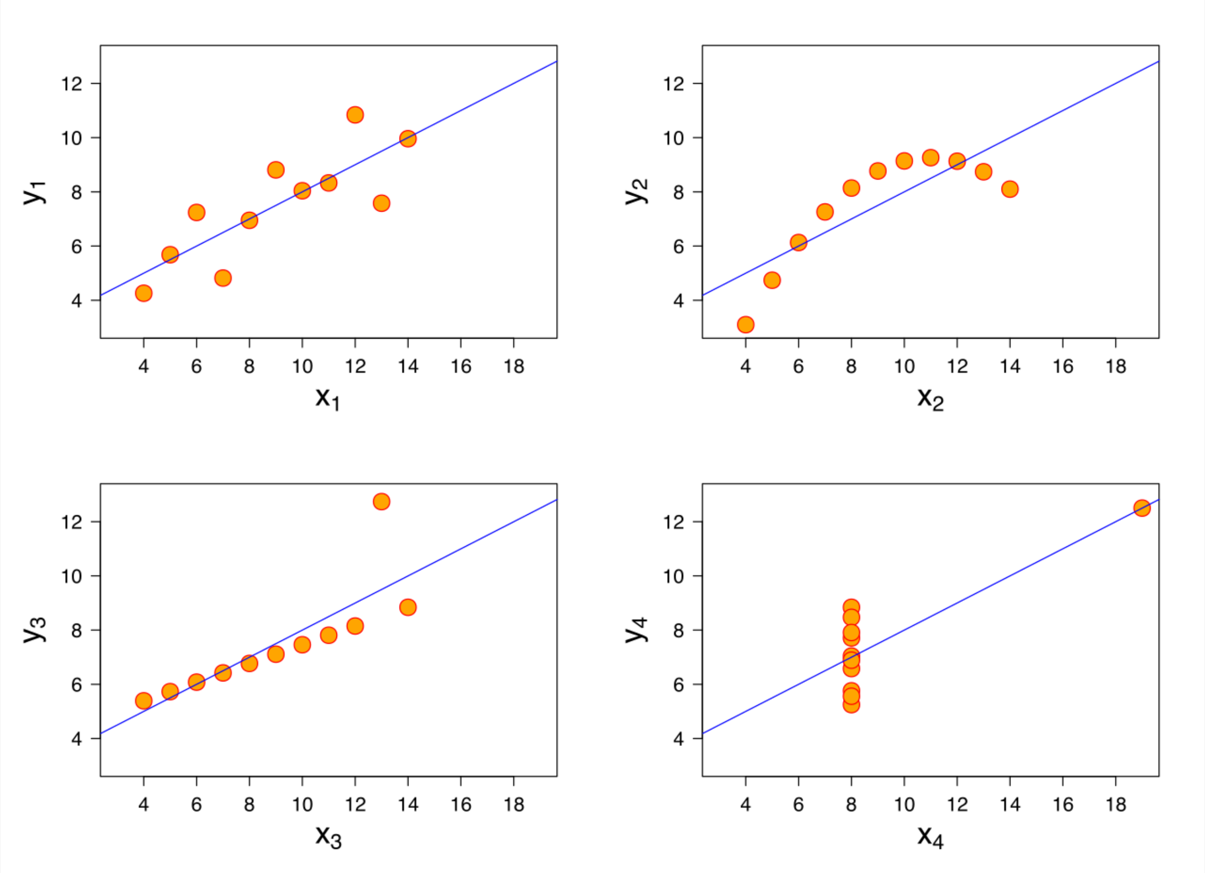


	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	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 Correlation Coefficient

Correlation is a technique for investigating the relationship between two quantitative, continuous variables, for example, age and blood pressure. Pearson's correlation coefficient (r) is a **measure of the strength of the association** between the two variables.

The first step in studying the relationship between two continuous variables is to draw a scatter plot of the variables to check for linearity. The correlation coefficient should not be calculated if the relationship is not linear. For correlation only purposes, it does not really matter on which axis the variables are plotted. However, conventionally, the independent (or explanatory) variable is plotted on the x-axis (horizontally) and the dependent (or response) variable is plotted on the y-axis (vertically).

The nearer the scatter of points is to a straight line, the higher the strength of association between the variables. Also, it does not matter what measurement units are used.

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

What is scaling:

Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step.

Why is scaling performed:

Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, many classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.

Another reason why feature scaling is applied is that gradient descent converges much faster with feature scaling than without it.[1]

It's also important to apply feature scaling if regularization is used as part of the loss function (so that coefficients are penalized appropriately).

Difference between normalized scaling and standardized scaling:

- Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.
- Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

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

   VIF is infinite if there is perfect correlation.

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

   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.