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?*

- More rentals on fall and summer with slightly lesser rentals on winter and spring
- More rentals were taken in the year 2019 compared to 2018
- More rentals happening on holidays
- More rentals happening on Clear day, followed by Cloudy and Light Snow days and none on days of Heavy Rain
- More rentals were taken from month April to Oct with peaks in Sep and Oct

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

For a column that can take n values, it takes only n-1 number of columns to represent the data filled with dummies. i.e, for a column that stores 5 seasons Summer, Monsoon, Winter, Spring, Autumn, it takes only 4 columns, say Monsoon, Winter, Spring, Autumn. '1' in Monsoon and rest of the others '0' means it is Monsoon. '1' in Winter and rest of the others '0' means it is Winter. '0' in all the columns automatically means it is Summer. Also, when a categorical variable is represented by all its dummy variables (without n-1 applied), it can lead to R^2=1 and hence infinite VIF. So, this one field should be dropped.

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

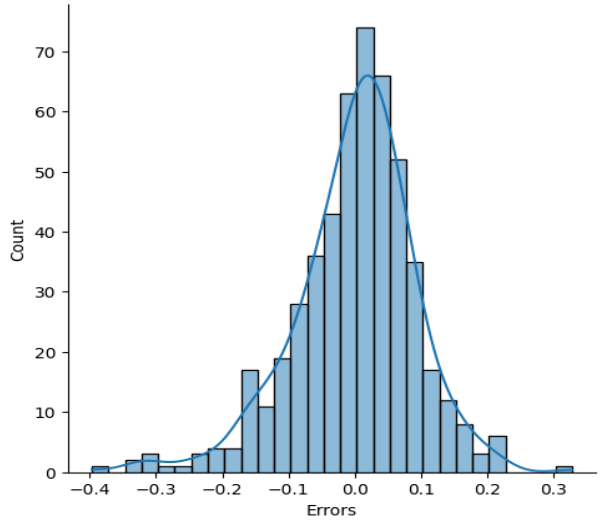
From the pair plot, it is obvious that 'temp' has the highest correlation with the target variable 'cnt'

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

Linearity: From the pair plot, it is evident that there exists a linear relation between 'cnt' and other numerical variables, eg 'temp'
Independence: Apart from very few columns, most of them are independent. There is no multiple correlation. This is evident from the heat map.

Homoscedasticity: the variance of the residuals is the same for any value of X. From the plot, we can see that it is not scattered or triangular but linear.

Normality: The Shape of the Errors plot looks normally distributed. The distribution is Centred around 0



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

Temperature, humidity and windspeed are the 3 features that has the highest correlation with the dependent variable 'cnt' with the coefficients + 0.5666, - 0.2864, - 0.2014 respectively. While temp has the positive correlation, humidity and windspeed has negative correlation.

General Subjective Questions

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

Linear Regression is a technique used to predict the value of interest based on the historical data. It takes historical data which comprises of the values of the variables that depends on the other variable(s) and the independent variables. In other words, the dataset should have a target variable that depends on at least one other predictor variable. If the preconditions or the assumptions of the Linear regression are respected, then the algorithm can be used to build the model that explains how the target variable changes with the the change in the predictor variable. These assumptions are

Homoscedasticity: (the variance of the residuals(errors) are uniform across the range of X)

Normality: The error distribution is normal and is centred around 0

Independent: The predictor variables do not multi-correlate with other predictor variables

Linearity: The target variable and the predictor variables are linear. In other words, the target variable can be explained by a linear equation of predictor variables.

$y = mx + c$

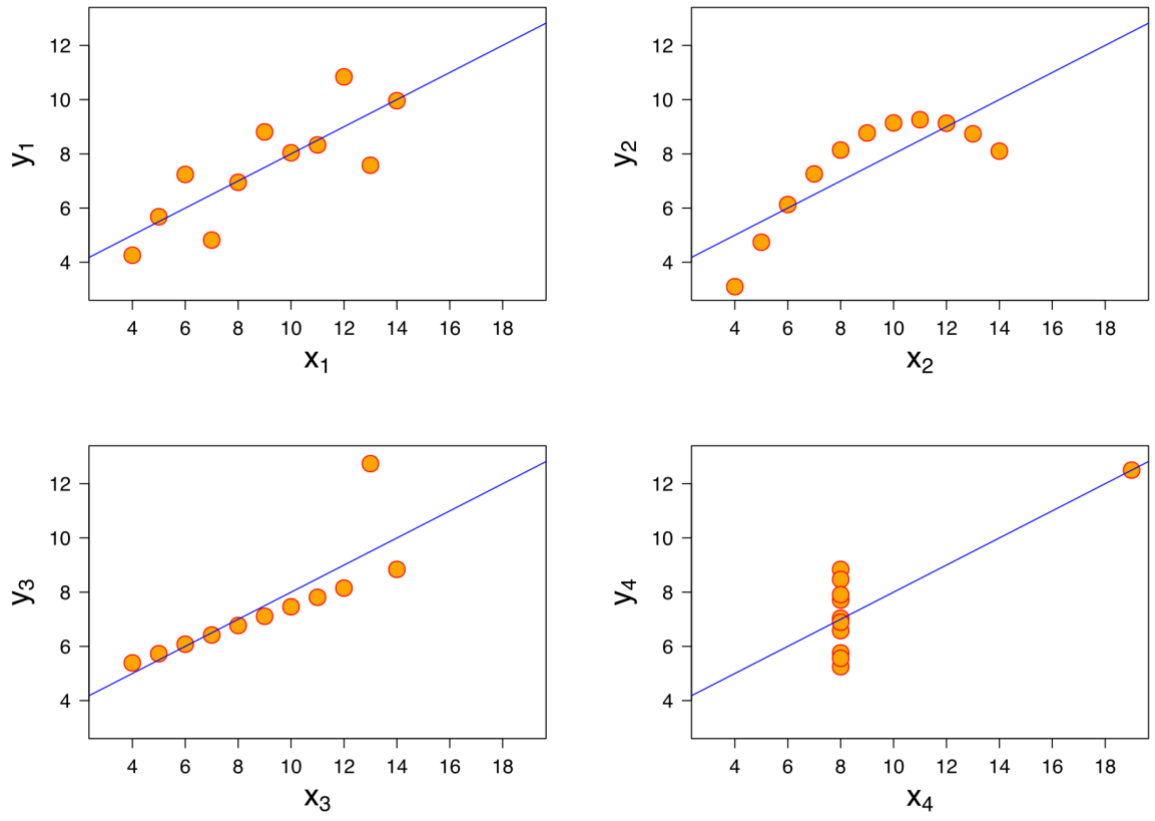
To apply this technique, the following steps are followed:

1) Explore and understand the historical data. This helps in preliminary understanding of the target and the predictor variables.

2) Prepare the data by treating null and duplicate values, converting boolean columns to binary columns, fill them with dummies and rescale the numerical columns that does not match in scale with other columns.
3) Split the data into 70-30 (or 80-20) training and test sets.
4) Build the model (find the coefficients and interrupts) iteratively till a satisfactory p-values are achieved for the coefficients and F-stat (for the fit).
5) Apply this model to predict the dependent variable(y_predict) for the test data.
6) Compare this predictor values with the test values(y_test) and see how well they match (Adjusted R^2)

2. *Explain the Anscombe's quartet in detail.*

This is an set of four data sets which are constructed to disprove one notion among statisticians – "numerical calculations are exact, but graphs are rough"



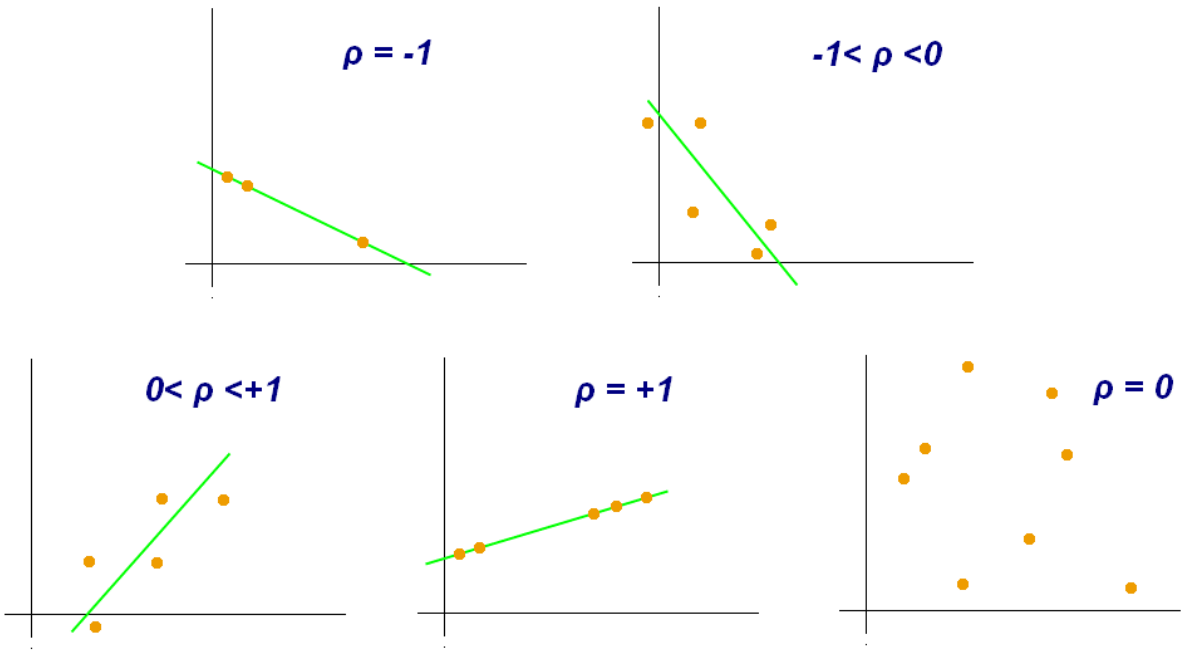
The four datasets (x1-y1, x2-y2…) were constructed in such a way that all their statistical properties(mean X, mean y, sample variance etc.,) are all identical. Even their R^2 is same. But when they are plotted, it looks like the one in the figure; hence disproving the notion that graphs and plots are rough and insight-less and only numerical representation carry exactness.

3. *What is Pearson's R?*

Pearson's R or the Pearson correlation coefficient is a descriptive statistic; it summarizes the characteristics of a dataset. It describes the strength and direction of the linear relationship between two numerical variables. The formula is as follows:

$$\frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2}\sqrt{\sum(y_i - \overline{y})^2}}$$

When comparing two attributes and calculating a score ranging from -1 to +1, a high score(+1 or -1) indicates a high similarity and near zero indicates no correlation.



It also is an inferential statistic; it can be used to test the statistical hypothesis. We can test if there is a significant relationship between two variables.

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

It is a technique used while preparing the data for regression model. The coefficients calculated for the predictors acts as the weightage or significance of that predictor variable. If the predictor variables in the datasets are not in scale, i.e, x1 is in 1000s, x2 is in 10s and x3 is in millions even though the effect each of them, independently have on the target variable is same, the coefficients the model yields will be disproportionate and out of scale. This leads us to believe unrealistic weightages. Scaling the variables brings all the involving variables to a same scale. There are it can be done. One is called Normalizing and the other is called Standardization.

Normalized scaling also known as Min-Max scaling. Data is transformed into a range between 0 and 1 by normalization, which involves dividing a vector by its length.. It assumes that there are no distortion in the data and the dataset is at a standard scale. This technique is affected by outliers.

Standardization on the other hand uses the mean and standard deviation to rescale values to fit a distribution between 0 and 1. Standardization is divided by the standard deviation after the mean has been subtracted.. It assumes the data to be in Gaussian distribution. This technique is immune to outliers.

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

VIF stands for Variance Inflation Factor. It is the statistical measure of multicollinearity in a regression model. Multicollinearity is a state in which multiple independent variables in the regression model correlate with each other.

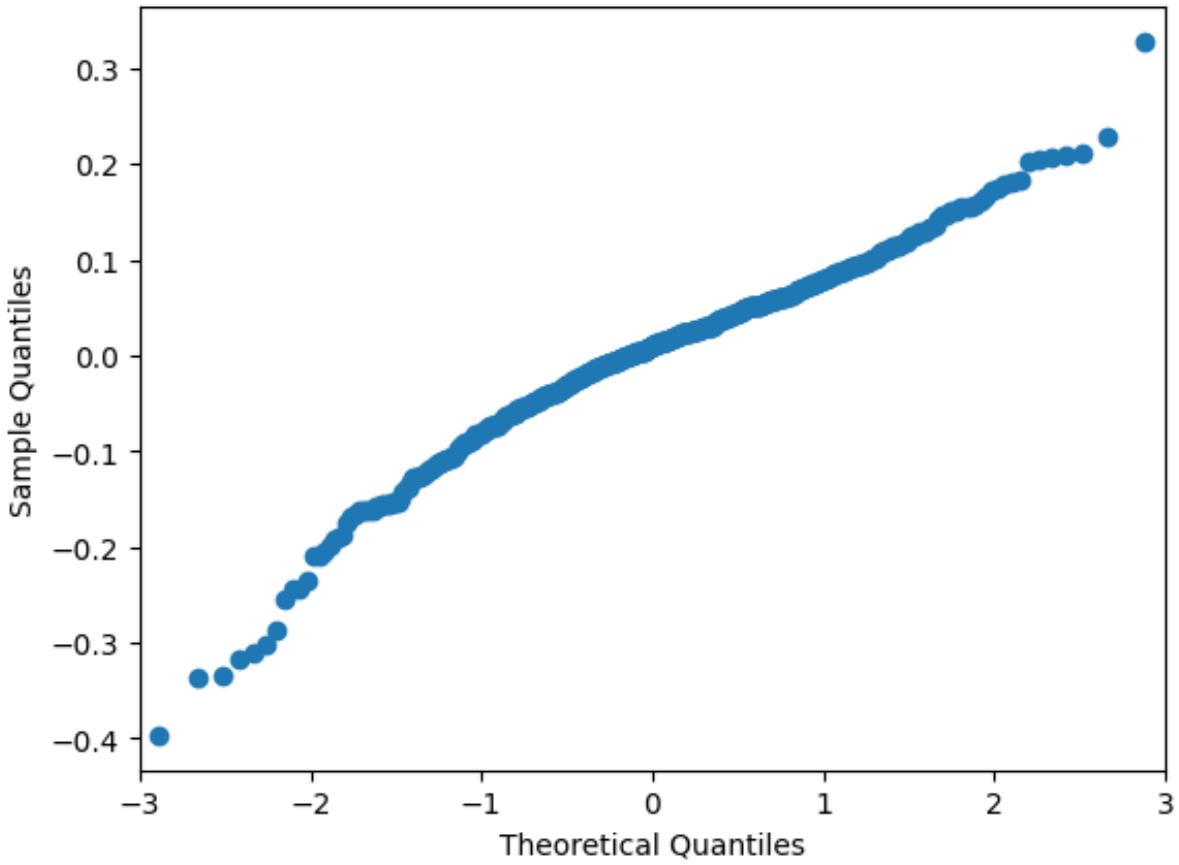
$VIF = 1/ (1 – R^2)$

The above formula is used to calculate VIF. R^2 in the equation refers to SSR, the sum of squared residuals. It ranges from 0-1 where 1 being the highest. VIF reaches maximum with R^2 is at the lowest, and VIF reaches infinity when R^2 is equal to 1. A case when two or more variables are exact duplicates or proportions of each other or if a categorical variable is represented with all its dummy values (dropFirst=False).

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

Q-Q plot is a scattered plot which is created by plotting two sets of the quantiles against one another. If we have data to statistical assume if it is normally distributed, a Q-Q plot can be used. Though it is not an air-tight approach to check for normal/exponential distribution, it helps in quick visual check.

The following plot used in the bike-rental exercise to verify the assumption for the Linear Regression model for Normal distribution of error terms.



As, we can see from the graph, the residuals are linear, and it can be verified that the error terms are normally distributed.