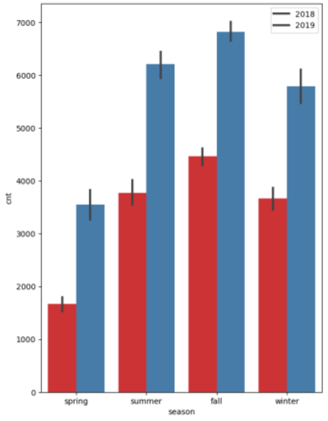
Linear Regression Assignment – Subjective

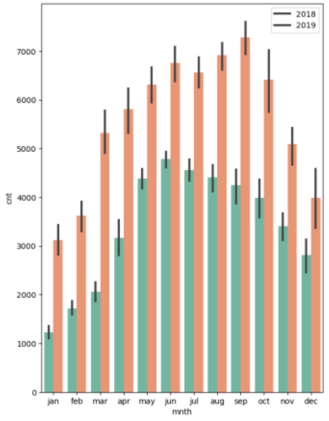
-MANJIRI P. GAJMAL

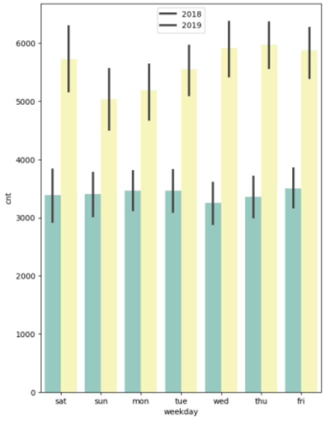
C57

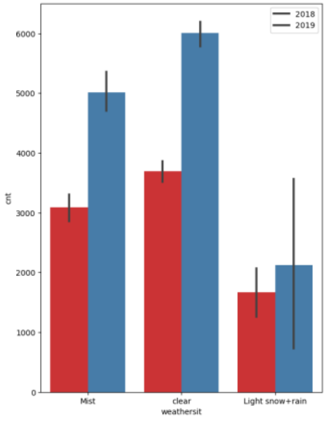
Assignment-based Subjective Questions

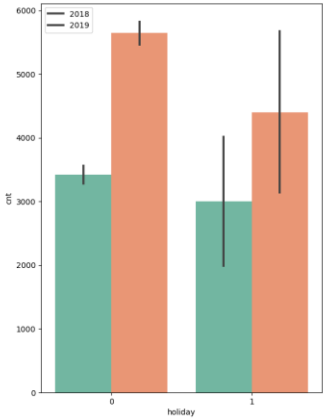
1.From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

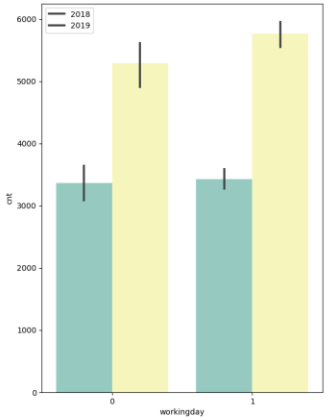


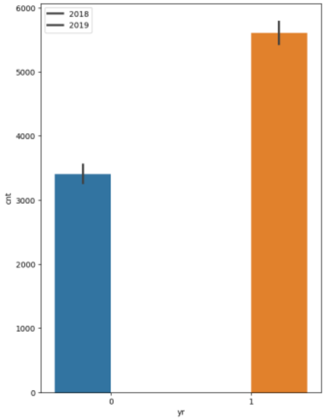












1) **Season**- It can be seen that fall season has more followed by summer, winter and least with spring. It can also be seen that booking has drastically increased from 2018 to 2019

2) **Mnth** - It can be seen more booking are done in month of May, June, July, august, September and October. It's been seen that booking trend shows increase in booking at thar start of the year but by the end of the year it shows decrease, this trend is same in 2018 and 2019 though there is increase in booking from 2018 to 2019 in each month.

3) **Weekday**- Thu, Fir, Sat and Sun have a greater number of bookings as compared to the start of the week.

4) **Weather Situation** - Clear weather has more booking followed by Mist, light snow + rain, as other variable it also shows increase in booking from 2018 to 2019.

5) **Holiday** - on holiday there is decrease in booking.

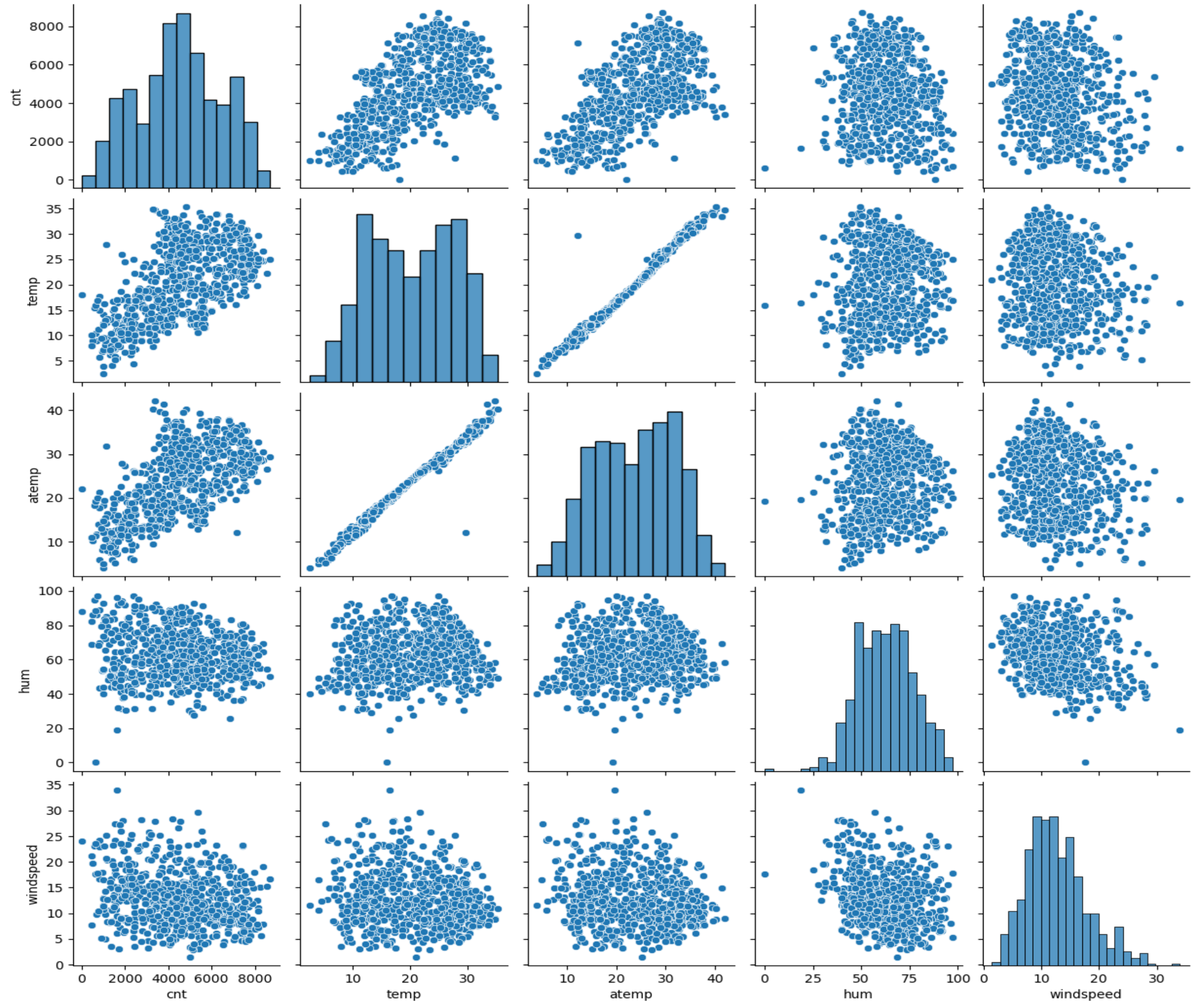
6) **Working Day** - there is no much difference in working day or non-working day.

7) **Year** - From 2018 to 2019 there is drastic increase in booking for bikes, which is good indication in terms of business perspective.

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

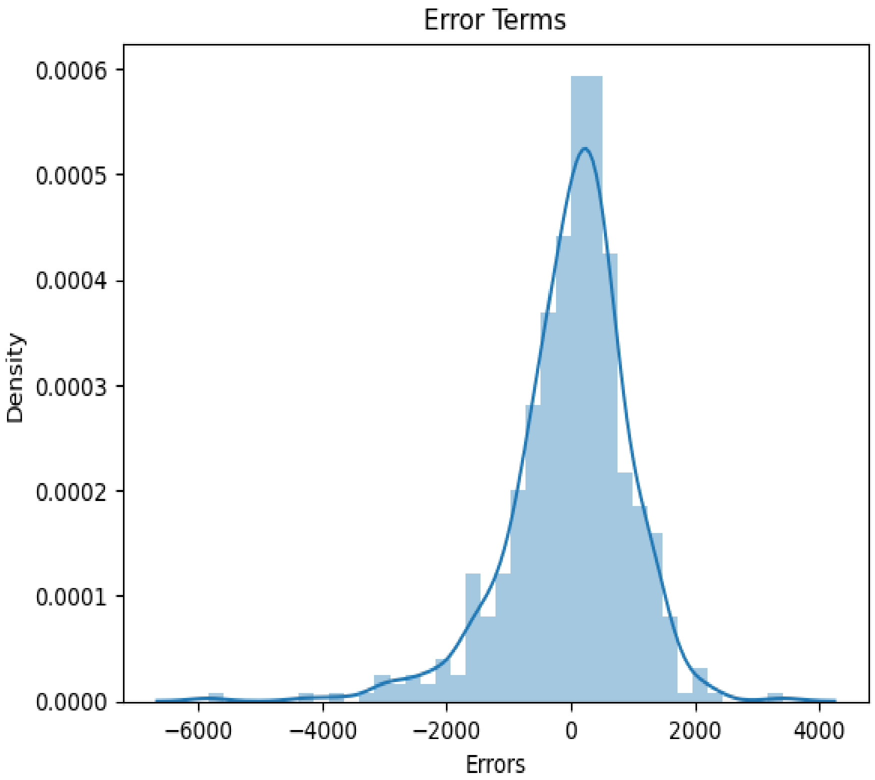
- In regression analysis, a dummy variable takes a binary value(0,1) to indicate presence or absence of some categorical effect .
- drop_first=True is important, as it helps in reducing in extra column created during dummy variable creation. Hence it reduces correlation among the dummy variables.
- For eg if we have 3 types of values in Categorical column and we are creating dummy variable for that column. If one variable is not A and B, then it C. So we do not need 3rd variable to identify variable C.

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

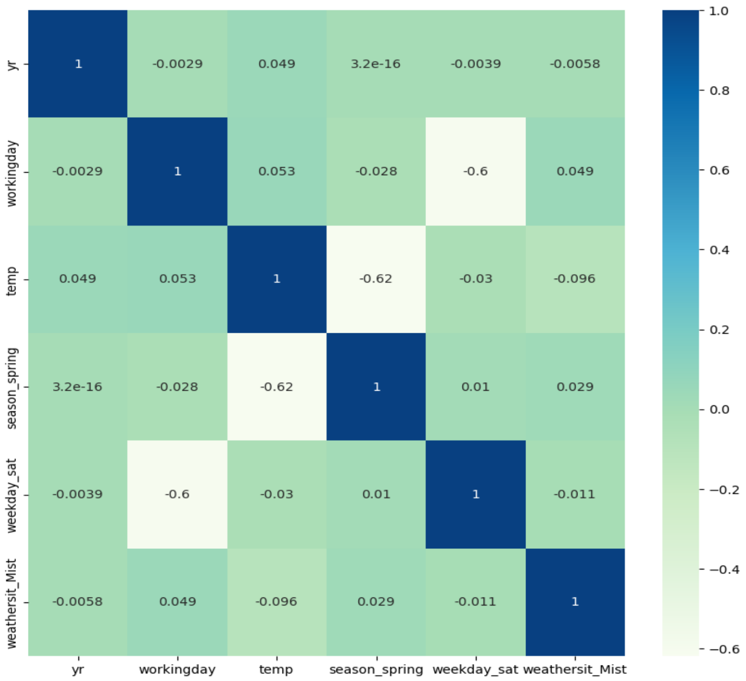


As seen in the figure ‘temp’ and ‘atemp’ are the two numerical variables which are highly correlated with the target variable (cnt).

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

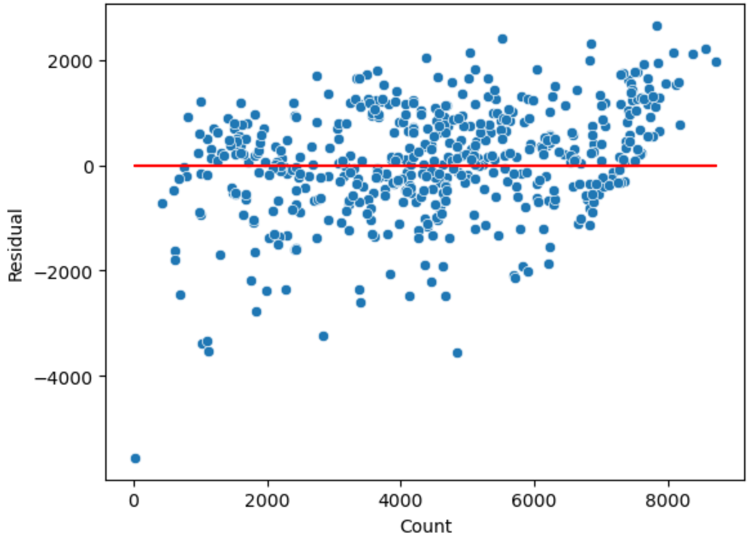


Residual Analysis- As seen in fig residual are scattered around mean = 0



	yr	workingday	temp	season_spring	weekday_sat	weathersit_Mist
yr	1	-0.0029	0.049	3.2e-16	-0.0039	-0.0058
workingday	-0.0029	1	0.053	-0.028	-0.6	0.049
temp	0.049	0.053	1	-0.62	-0.03	-0.096
season_spring	3.2e-16	-0.028	-0.62	1	0.01	0.029
weekday_sat	-0.0039	-0.6	-0.03	0.01	1	-0.011
weathersit_Mist	-0.0058	0.049	-0.096	0.029	-0.011	1

Multicollinearity Check –
Insignificant multicollinearity is seen among variables



Homoscedasticity–
No visible pattern observed from above plot for residuals.

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

Following are top 3 features contributing significantly towards explaining the demand of the shared bikes –

- yr
- Working day
- Temp

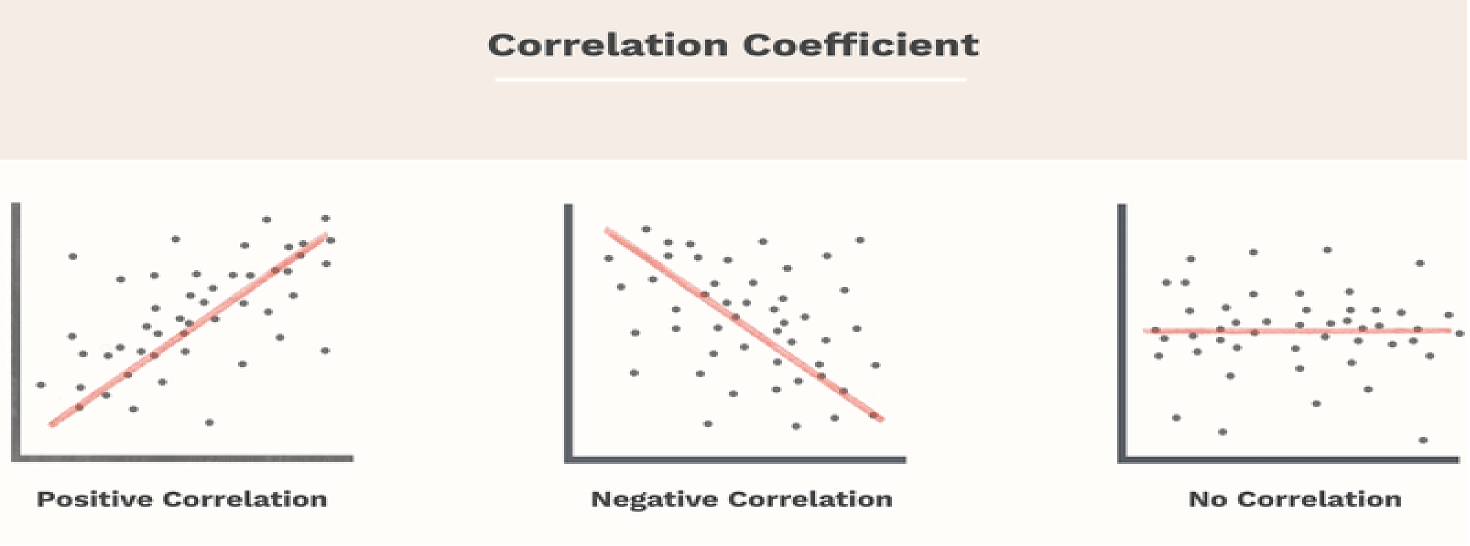
General Subjective Questions

1. Explain the linear regression algorithm in detail.

- Machine Learning is a subfield of Artificial Intelligence that focuses on the creation of algorithms and statistical models that can learn from and predict data. Linear regression is a supervised machine-learning technique.
- Linear regression is a sort of supervised machine learning technique that computes the linear connection between one or more independent features and a dependent variable. When the number of independent features is one, it is referred to as univariate linear regression; when there are more than one feature, it is referred to as multivariate linear regression.
- . Linear regression is based on the popular equation –

"y = mx + c“

- It assumes that there is a linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x).



Linear regression shows correlation coefficient as shown in fig

Regression is broadly divided into simple linear regression and multiple linear regression.

1. Simple Linear Regression : SLR is used when the dependent variable is predicted using only one independent variable.
2. Multiple Linear Regression : MLR is used when the dependent variable is predicted using multiple independent variables.

The equation for MLR will be:

yi = βo+ β1x1+ β2x2+…+ βpxp

β1 = coefficient for X1 variable

β2 = coefficient for X2 variable

β3 = coefficient for X3 variable and so on…

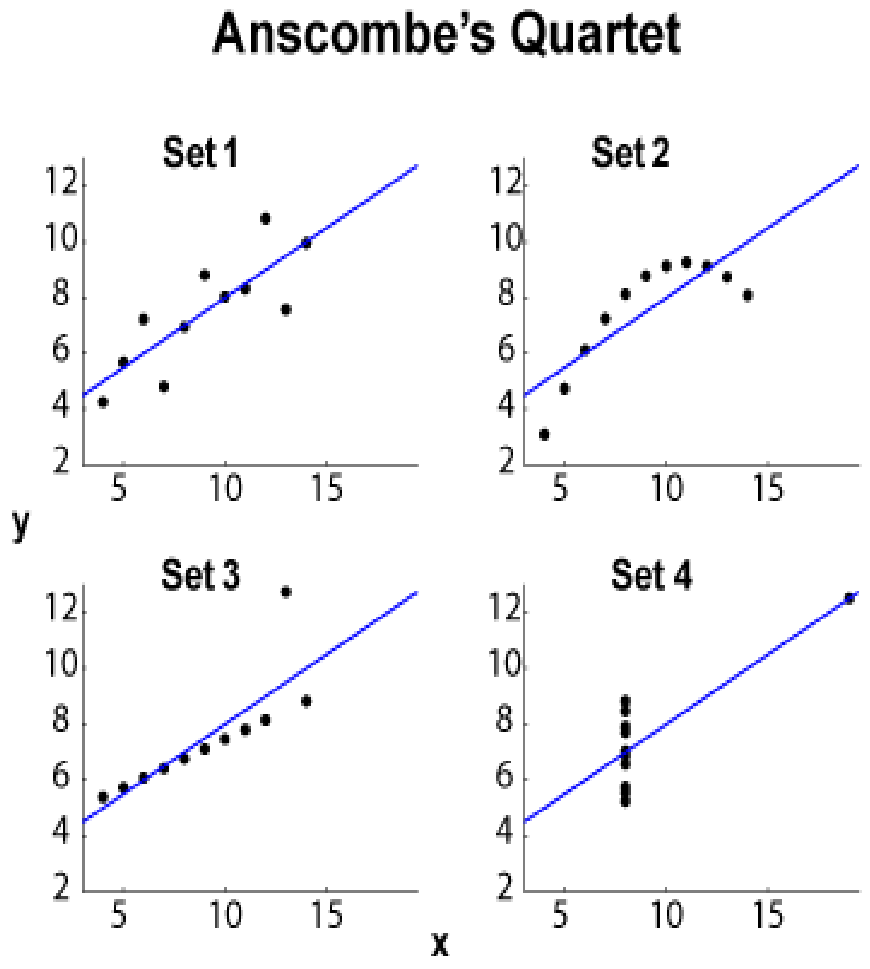
β0 is the intercept (constant term).

2. Explain the Anscombe’s quartet in detail.

Anscombe's quartet teaches us the value of seeing data before constructing models using different approaches. This implies that in order to see the distribution of the samples and identify the many types of anomalies found in the data, the data features (outliers, diversity, linear separability, etc.) must be plotted. Furthermore, no other type of data set can be handled by linear regression; it can only be regarded as a fit for data having linear connections.

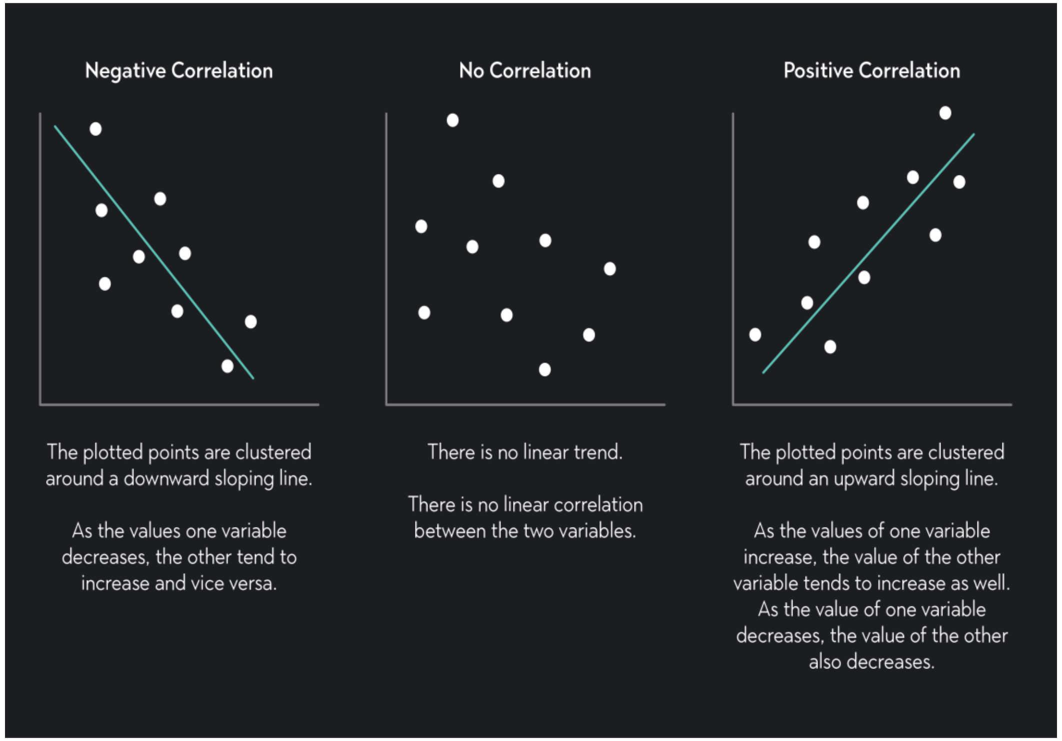
The four datasets in Anscombe's Quartet share the following statistical properties:

1. Mean (x and y): Approximately 9 for both x and y in each dataset.
2. Variance (x and y): Approximately 11 for both x and y in each dataset.
3. Correlation (between x and y): Approximately 0.816 in each dataset.
4. Linear Regression Line (y = a + b*x): All four datasets have nearly the same regression line parameters.



1. The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two correlated variables, where y could be modelled as gaussian with mean linearly dependent on x.
2. For the second graph (top right), while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
3. In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier, which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
4. Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

3. What is Pearson’s R?



Negative Correlation

The plotted points are clustered around a downward sloping line.

As the values one variable decreases, the other tend to increase and vice versa.

No Correlation

There is no linear trend.

There is no linear correlation between the two variables.

Positive Correlation

The plotted points are clustered around an upward sloping line.

As the values of one variable increase, the value of the other variable tends to increase as well. As the value of one variable decreases, the value of the other also decreases.

- Pearson’s R is a numerical representation of the strength of the linear relationship between the variables. The correlation coefficient will be positive if the variables tend to rise and fall together. The correlation coefficient will be negative if the variables tend to go up and down in opposite directions, with low values of one variable correlated with high values of the other.
- The Pearson correlation coefficient, r, can take a range of values from +1 to -1

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

- Scaling is a pre-processing technique in data that is used on independent variables to normalize the data within a specific range. It also facilitates algorithmic computation performance increases.
- The majority of the time, the gathered data set includes attributes with widely dissimilar magnitudes, range and units. The algorithm only considers magnitude if scaling is not performed, and not units, leading to inaccurate modeling. We must scale in order to address this problem and bring all the factors to the same degree of significance. It is crucial to remember that scaling only has an impact on the coefficients and not the other parameters such as R-squared, p-values, t-statistic, F-statistic, etc.

1. Normalization/Min-Max Scaling:

   It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

2. Standardization Scaling:

   Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

- sklearn.preprocessing.scale helps to implement standardization in python.

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

The variance inflation factor (VIF) measures how much collinearity has increased the variance of the coefficient estimate.

(VIF) is the same as 1/(1-Ri2).

If there is perfect correlation, VIF = infinite.

Where R-1 is the R-square value of the independent variable being examined to evaluate how well it is explained by other independent variables.

If one independent variable can perfectly explain another independent variable, it has perfect correlation and an R-squared value of one. As a result, VIF = 1/(1-1) yields VIF = 1/0, which is "infinity".

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

A probability plot, or Q-Q plot, is a graphical technique that compares two probability distributions by showing their quantiles against one another.

An effective graphical tool for determining if a set of data may have originated from a normal, exponential, or uniform distribution is the quantile-quantile (Q-Q) plot.

Finding the similarity or dissimilarity between two distributions may also be done using the QQ plot.

The QQ plot should be more linear if they are quite comparable. Scatter plots are the finest means of testing the linearity assumption. Second, every variable in the linear regression analysis must be multivariate normal. A Q-Q-Plot or histogram are the most effective tools for verifying this premise.