**IN3063 Coursework Task 2**

Study several factors that affect the performance and interpretation of a simple model such as Linear Regression analysis

**Outliers:**

* Description
  + Outliers are the points that does not follow the same pattern as most other points and fall far from the rest. Outliers can have a massive impact on linear expression because it can change the line direction. Outliers can be caused due to an experimental error. These outliers are usually excluded from Linear Regression.
  + Outliers lowers the quality of the data.

* How they affect the estimation of the coefficients:
  + An outlier drastically increases the range of the results.
  + A single outlier can change the y-direction of the linear expression.
  + In Figure 1 below, outlier with result ‘800000’ can be seen in the graph. The outlier makes half of the graph to be useless and the linear regression is unclear.
  + In Figure 2, the outlier with result ‘800000’ has been dropped and the whole graph has changed. Linear Regression has became more accurate and the whole graph is used unlike Figure 1.

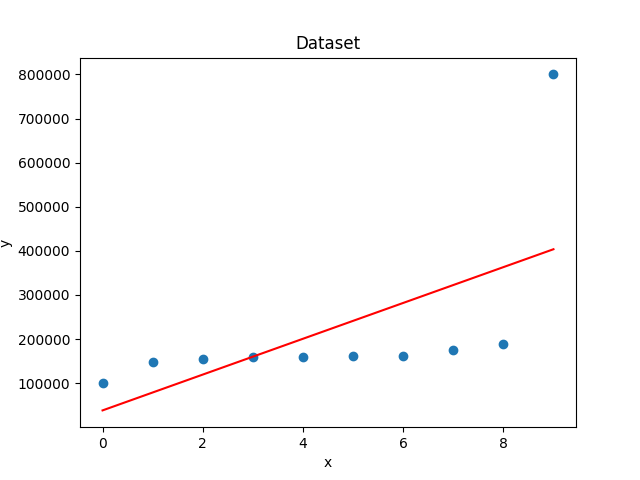


Figure 1

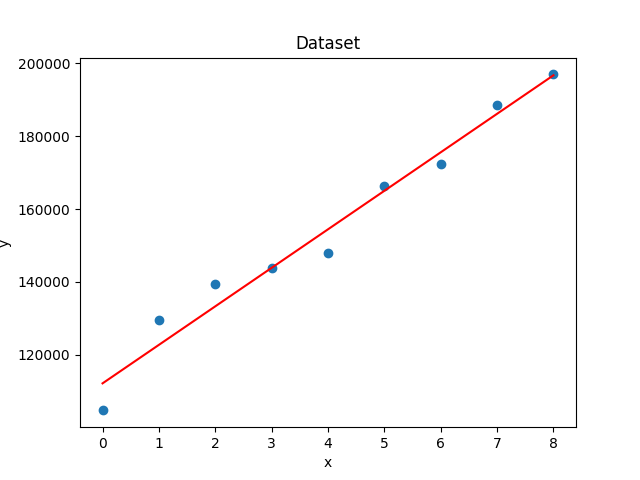


Figure 2

* How can we detect them and remove them:
  + One of the ways to detect and remove outliers are by comparing the observations to interquartile range. Observations that are higher than 150% of interquartile range can be box plotted as outlier and can be removed.
  + Another way to find outliers is by calculating the Z-score. The number of standard deviations by which the value of an observation exceeds the mean of what is being observed is known as the Z-score. Once Z-score is calculated, any data that is too far from zero will be treated as outlier.
  + Z-Score has been calculated using ‘scriby’ library which can be seen on Figure 3. I have used the same data I used for Figures 1 and 2 for Figure 3. Z-score for last value of y is far from 0 which means that we can detect that as an outlier.

Text

Description automatically generated

Figure 3

* How do they affect the normalization of covariates:

**Normalization of covariates**:

* Is the probability distribution of all covariates xi the same: