Idi Amin da Silva, Student Number:10363300

Programming For Big Data

CA05

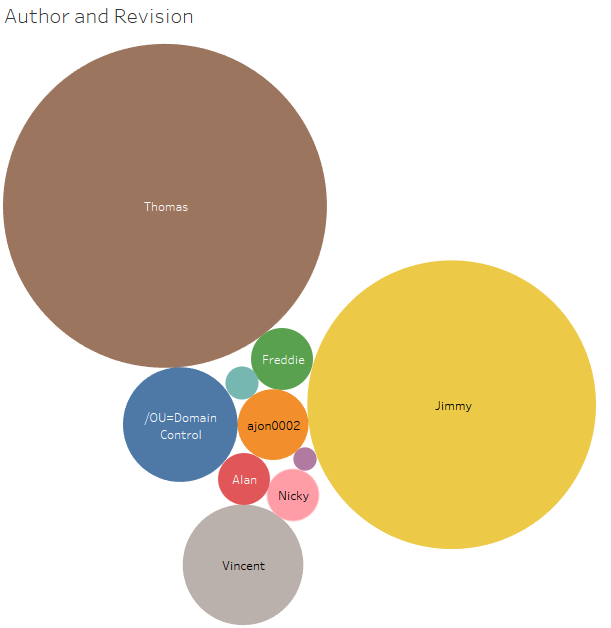
In this present Continuous Assessment (CA05) I have cleaned the data set in the excel spread sheet that reduced to the total of 422 columns and 14 rows.

I have decided to use two software/tools (Tableau and RStudio) to analyse and withdraw a valuable conclusion about the data set.

1st. **Tableau**  is used to provide the data visualization through the graphs that will allowed to compare different attributes in the data set.

2nd. **RStudio** is also used to provide the statistical Analysis of data set by using different parameters that will allowed us to measure, compare and quantify some attributes of our data set.

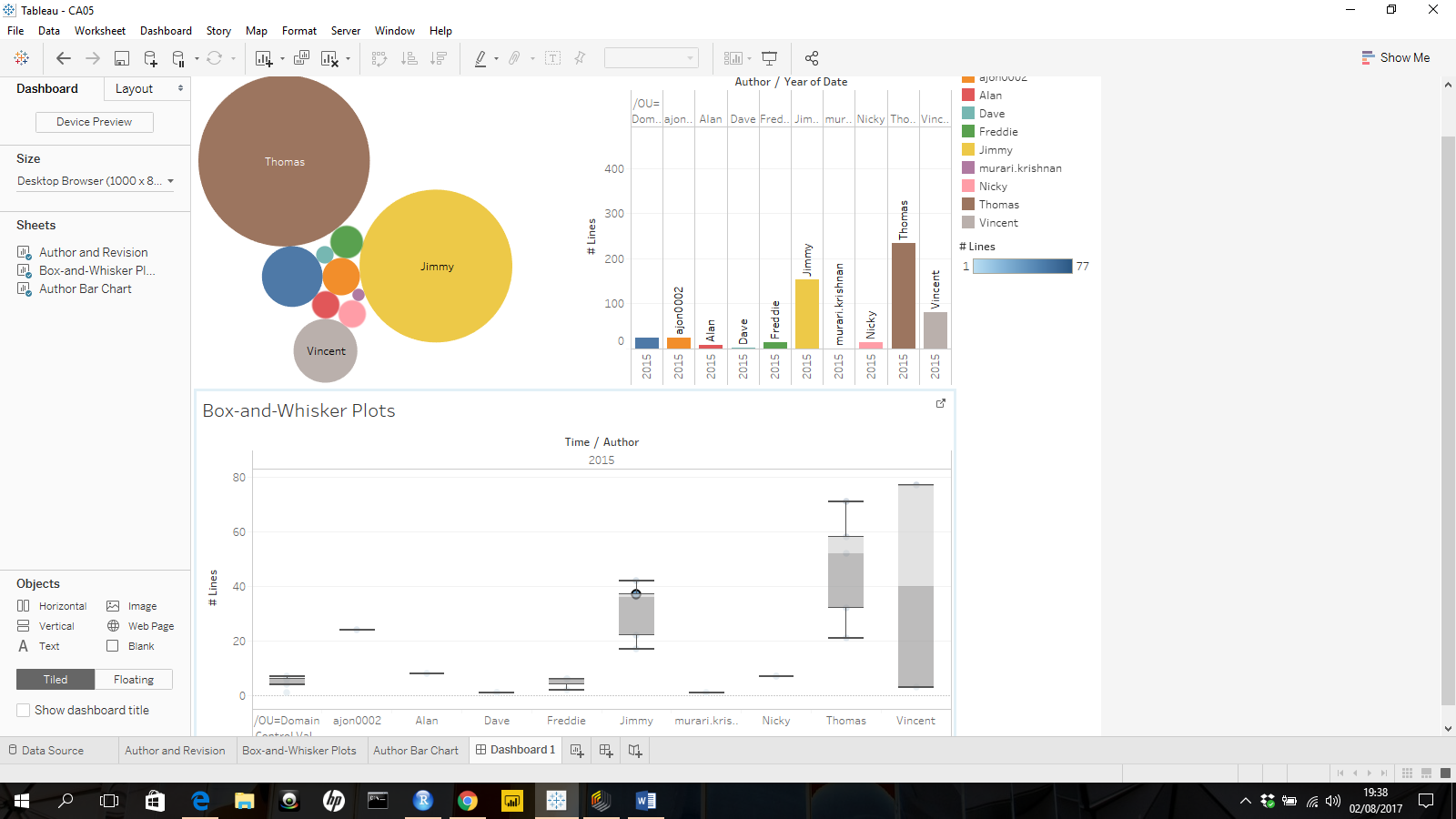
Analyse of the Author and the Revision



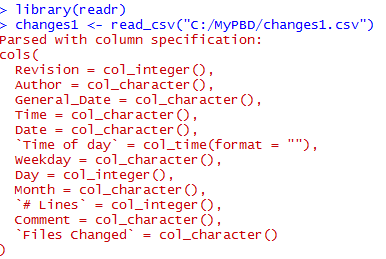
From the picture above we can see clearly that the Author Thomas is leading in term of the revision with 290,256,683 in the second-place Jimmy with 230,520,957 in the last position Murari.Krishnan with 1,511,531 revision.

Therefore, by the simple visualization of the graph we can conclude that Author Thomas has the highest revision among the others Authors.

The dash board will congregate some graphs representation in a single sheet for analysis.



**Statistics Analysis**



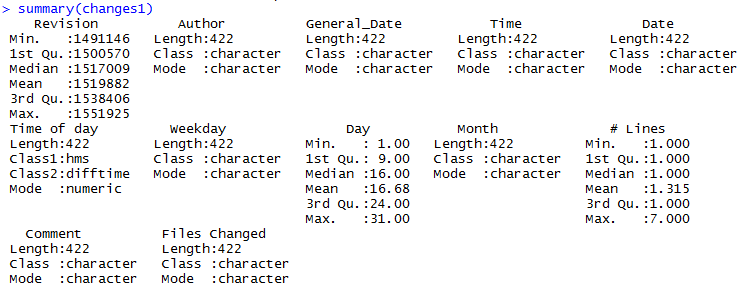


Table1.1

**Analysis of the measures of central tendency**.

For instance, let analyse the **Revision:**

From the table 1.1 we have the expected value/arithmetic mean of revision is 1,519,882 (i.e. =1,519,882) this value represents the average value of revision, we add all individual revision per Author and divide for the total number of the revision.

The first quartile (1st Qu=1,500,570) represent 25% of total revision.

The 3rd quartile (3rd Qu= 1,538,406) represent 75% of total revision.

The range will be difference between the maximum and minimum value (Range=Maximum Value – Minimum Value = 1,551,925 – 1,491,146=60,779).

Multiple Linear Regression

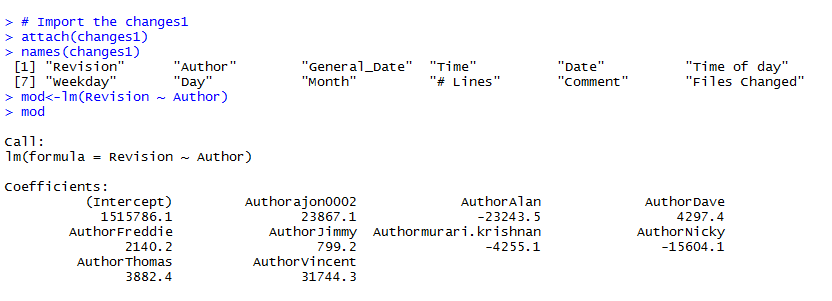


Table 1.2

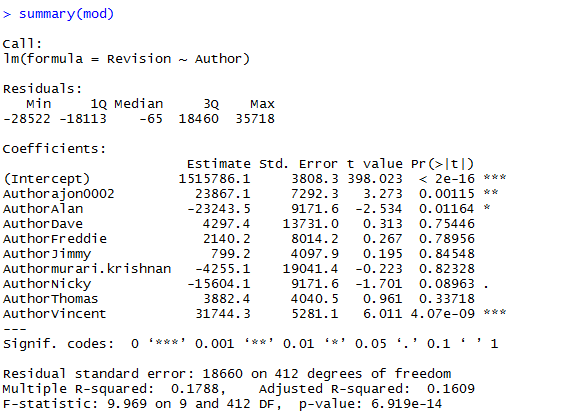


Table1.3

In the multiple linear regression, I have used the Author as Independent/Explanatory Variables to predict the value of Dependent/Response Variable.

From the table1.3, I have Coefficient of the Multiple Determination (Multiple R-Square= 0.1788=17.88%) this means that 17.88% of the variation in the revision can be explained, also the coefficient of the multiple determination is the ration between the Regression Sum of the Square(RSS) and Total Sum of Square(TSS).

The p-value = 6.919e-14=6.919x is the probability of obtaining a test statistic equal to or more extreme that the result obtained from the sample data, given that the null hypothesis H0 is true.