Data@ANZ Program

Exploratory Data Analysis with Excel by Gustavo Montenegro

The beginning

Our Dataset has 12043 rows (without the headings row) and 23 columns. There are not duplicate records, we have 100 recurring customers and the dataset need to be cleaned.

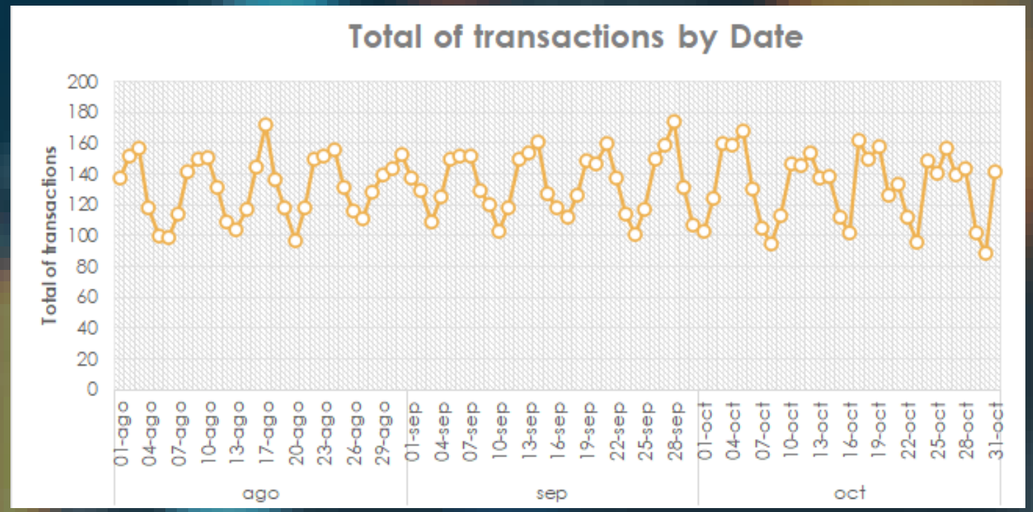
The date column need a format change, because it is unavailable with the dafault format.

The 'bpay_biller_code' column has two different patterns of data, one of them is numeric and the other is text. I think its a data entry mistake.

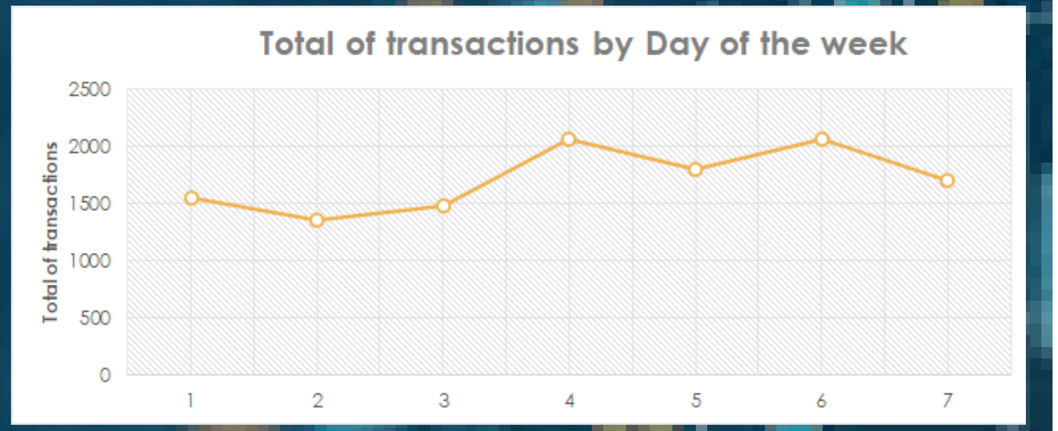
Many columns have NaN values, the cells are empty, but they are not neccesary for this task.

Interesting Findings

If we make a chart of total transactions by date, we will notice that there is a pattern each week. We can prove it with another chart that contains all the transactions by day of the week.



The first day of the week is sunday, on monday there is the lowest volume of transactions, while on wednesday and friday we have the maximum.

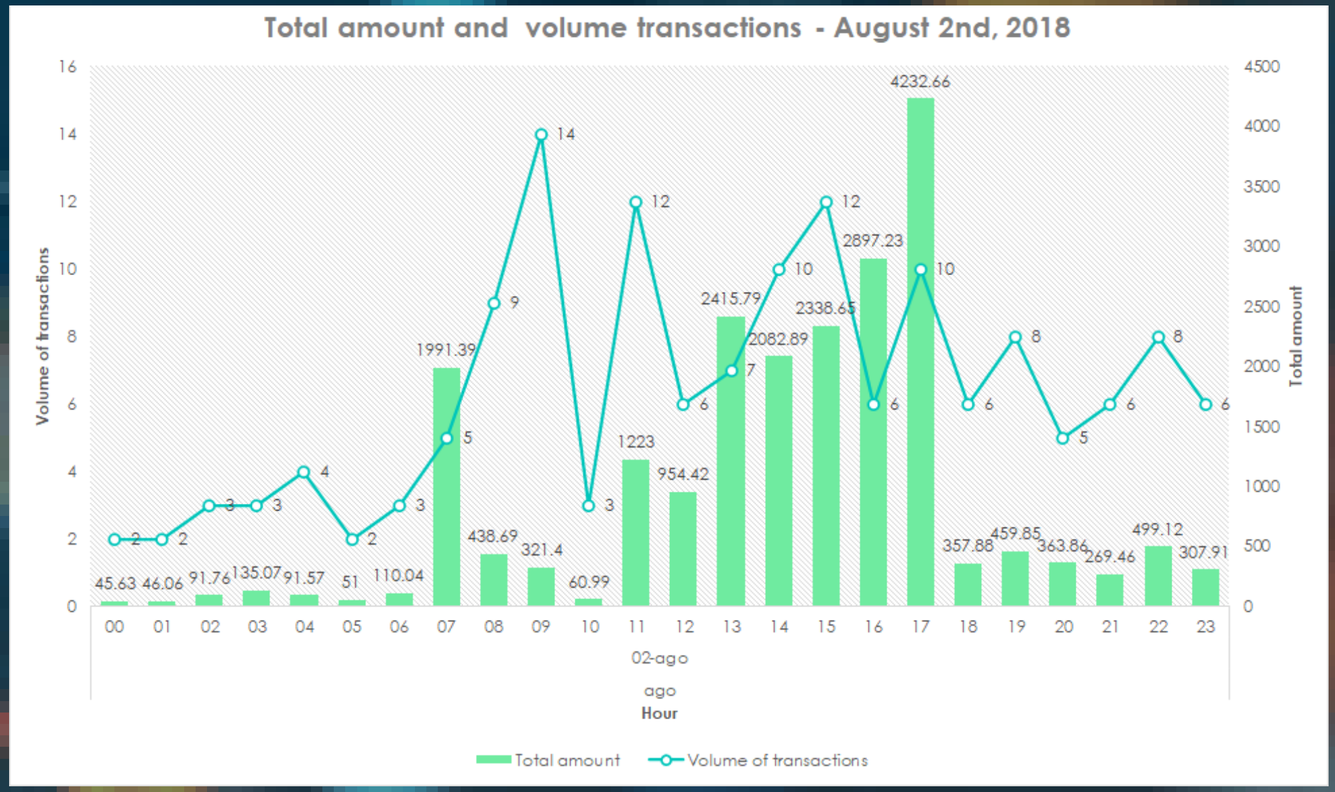


The relevant data	Average of amount	Average of transactions (rounded)
Minimum (Tuesday)	373.22	97.14
Maximum (Wednesday)	195.22	147.36
Maximum (Friday)	249.35	159.46

The task

The challenge is to create a visualization with data segmentation. It includes the volume of transactions, the total amount, date and time of an average day or week. The time is grouped by hours.

Chart with outliers



How do the outliers impact the chart?

If we use all the time points in order to make a visualization, we can notice that the average of transactions by time is near to one. If we want to calculate the outliers by the interquartile range (In this case its zero), we will have that all the numbers except one are outliers.

Chart with all the time points

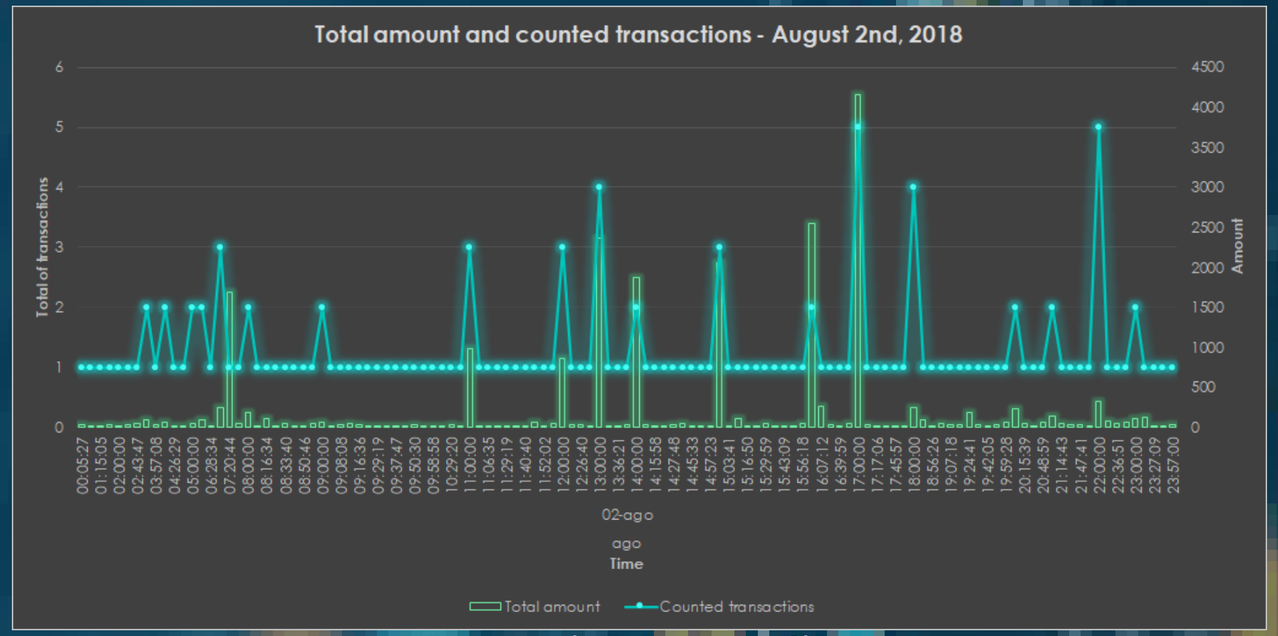
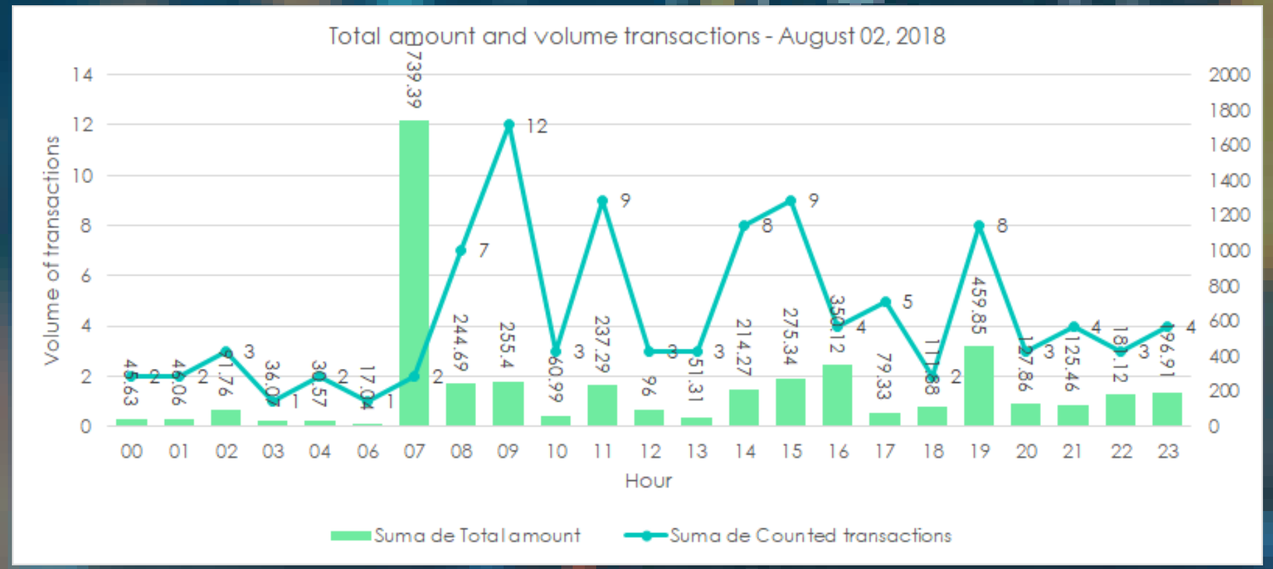


Chart without outliers



The transaction outliers modify the total amount and the trasanction volume. In the first chart (with outliers) we have that the highest value of transaction volume is 14 while in the second chart (without outliers) is 12. Of course, the interquartile range is an easy method in order to find outliers, there are many other methods that can help us to identify these points better.

Challenge: Total amount by coordinates