Thandeka Chaka
Project 2

The summary analysis for the Selected Dataset for clustering
Source of the data:https://www.kaggle.com/giovamata/airlinedelaycauses

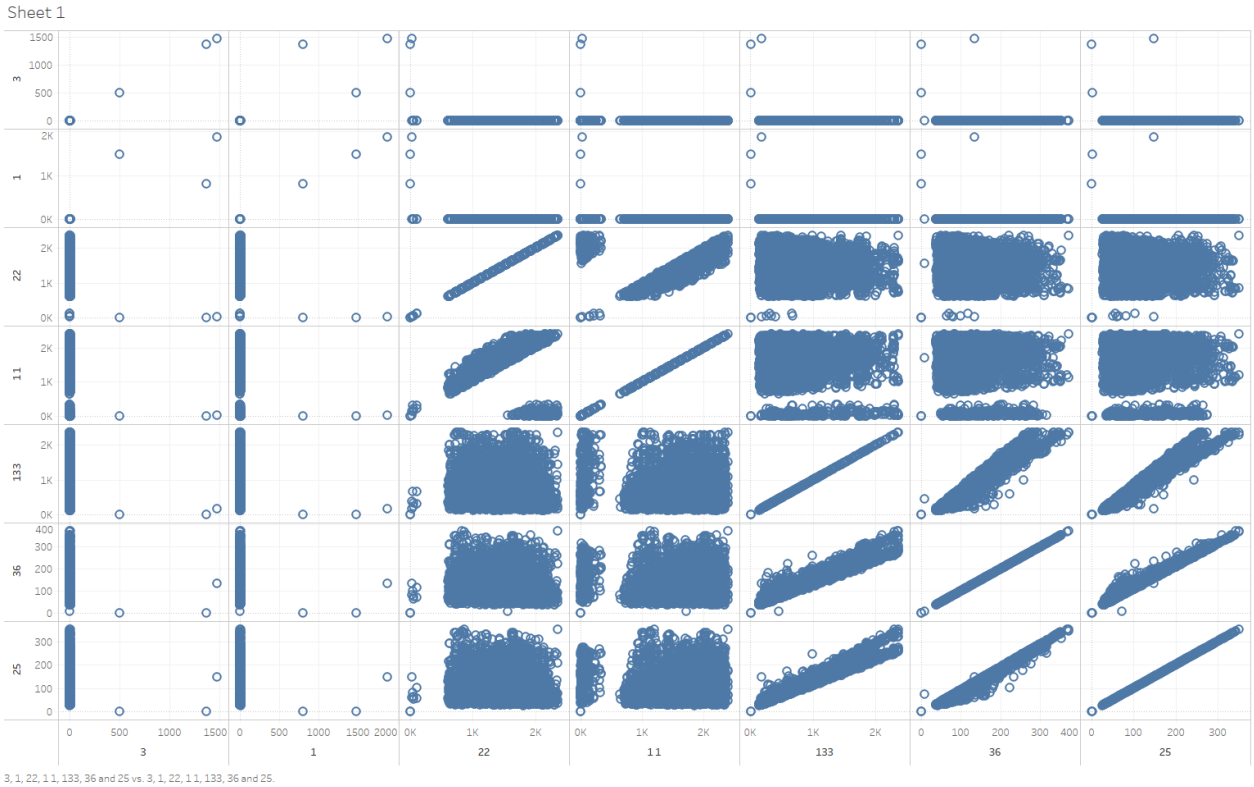
The categorical variables were used as identifiers for example the origins and destinations of flights for conclusive study of the travel analysis. For the missing data for quantitative variables that had too many missing data , I removed them during cleaning inorder to reduce data bias.

I did not feature scale any of my data as the ranges were quite sensible with no data overpowering another per column. Each row represents each flight that left a different or same airport location to another place.. It also contains the day, month, time for both departure and arrival, the distance travelled and the actual time elapsed. It also contains the airtime for each flight and the unique categorical identifiers for each flight like the flight number and tail number.

The reason why I chose the nine dimensions was firstly to reduce the possibility of maxing out the solver rows. However most importantly I wanted to be able for later clustering purposes of the test data I did not include in this project to help predict the different flights that were delayed around certain times(in terms of date, time,month and day of week) and the different distance travelled to see if the length of a distance contributed to the delay of flights.

This data is from the year 2008 during the wall street market crash so I thought it was interesting to evaluate the travel patterns around that time in some of the US airports etc. the data source is data.world.com.

Thandeka Chaka
Project 2



The above data for flights is the monthly Air travel consumer reports that summarises the statistics and raw data of air travel at that time. This data focuses on the domestic air travel within the US.

Since I did not have a precise hypothesis while looking at the data conclusion drawn from the data clustering were a little bit inconclusive. However, from the attempted clustering that I did, most flights were delayed between the months April, June and July.This can be due to the estimated 20% decline in global stock market especially around mid 2008 when this data was recorded. I also found it interesting that there wasn't that much travel during December which is highly a festive season that one might have anticipated flight cancellation .