1 Introduction

The goal of this project is to analyze our Taxi dataset using multiple correspondence analysis using the FactorMineR library. This is similar to principal component analysis except it performs better for categorical data. We are mainly interested in finding the relationship between the *Total_amount* variable and a series of categorical variables we will investigate later.

2 Data Description

We are working with the Taxi Dataset from before. A Proposition from my last investigation involved dropping all erroneous rows before feature generation began. At this point we have 4566 rows. There are only two erroneous rows still in the dataset which have now been removed so we have 4564 rows. We also drop the missing values leaving 4433 rows.

3 Data cleaning

A script has been included from previous work on PCA to clean up the data.

4 Data Preparation

4.1 Discretization Of Variables

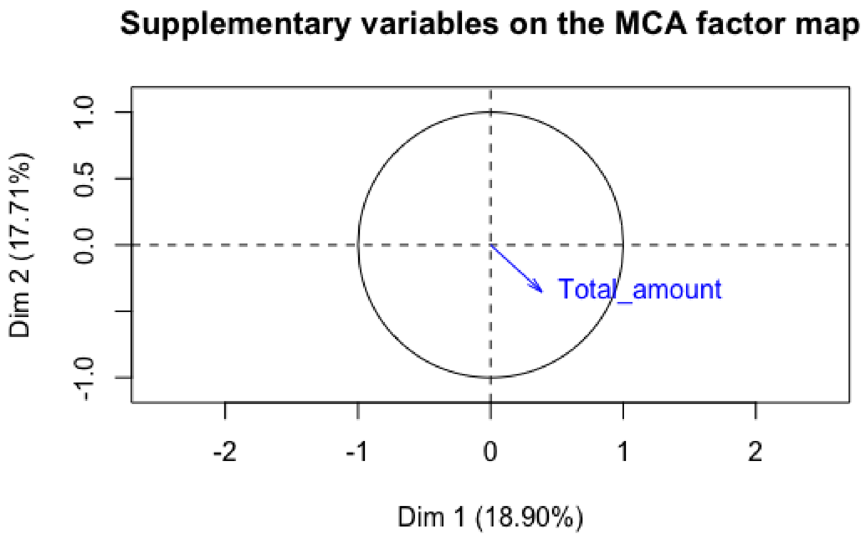
Next we need to convert The following into Discrete Variables:

1. Total_amount
2. Trip_distance
3. Fare_amount

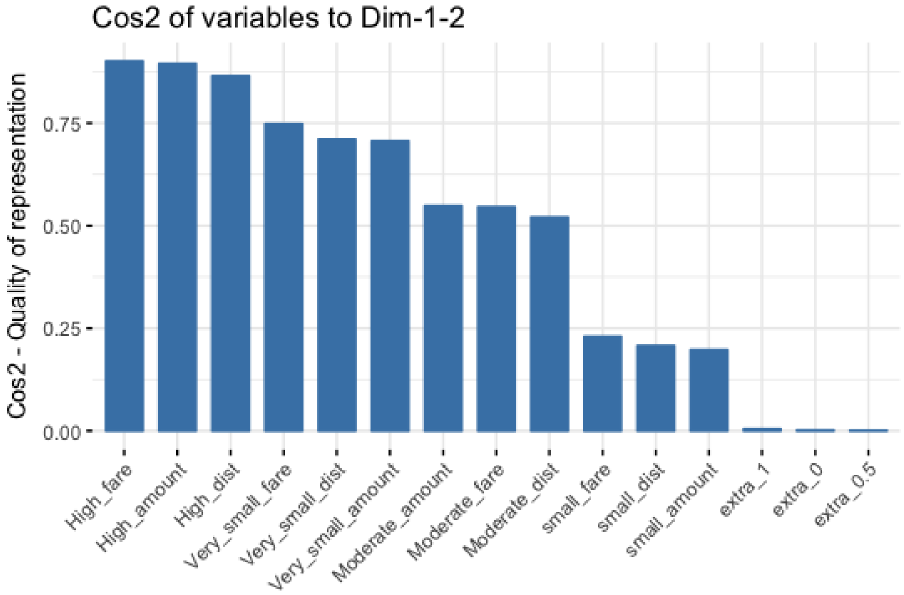
We rename the quartiles for the above variables are set to Very Small, Small, Moderate, High.

The following need to be made into factors:

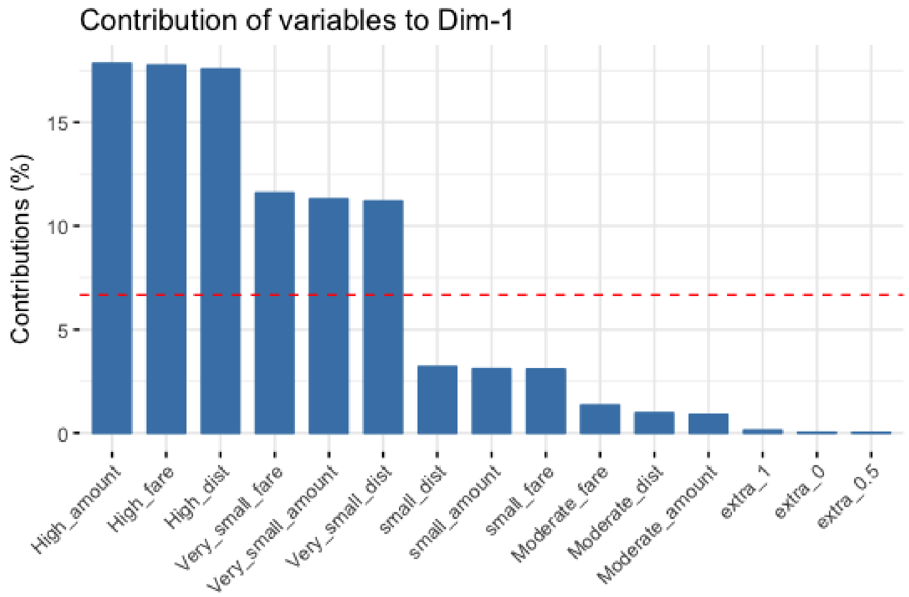
1. Extra
2. MTA_tax
3. Tolls_amount
4. improvement_surcharge

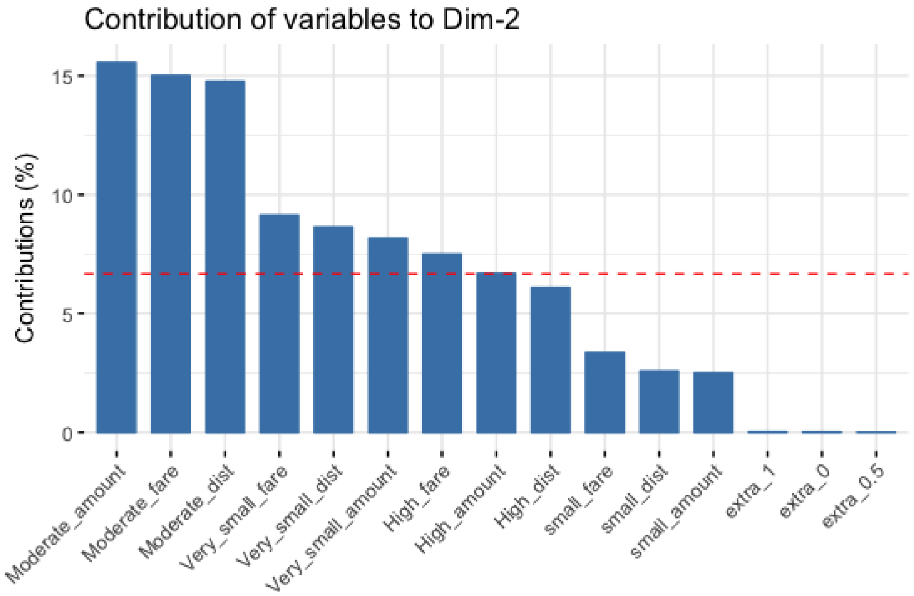

Supplementary variables on the MCA factor map

The below graphs show the most important and well represented values in our dataset which correlate with fare amounts.

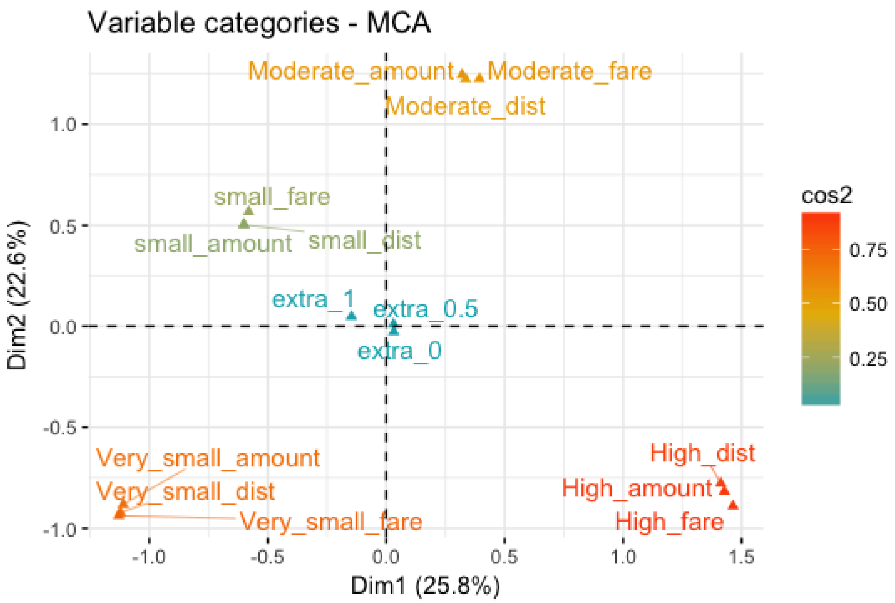

Cos2 of variables to Dim-1-2

5 Contribution of Dimensions





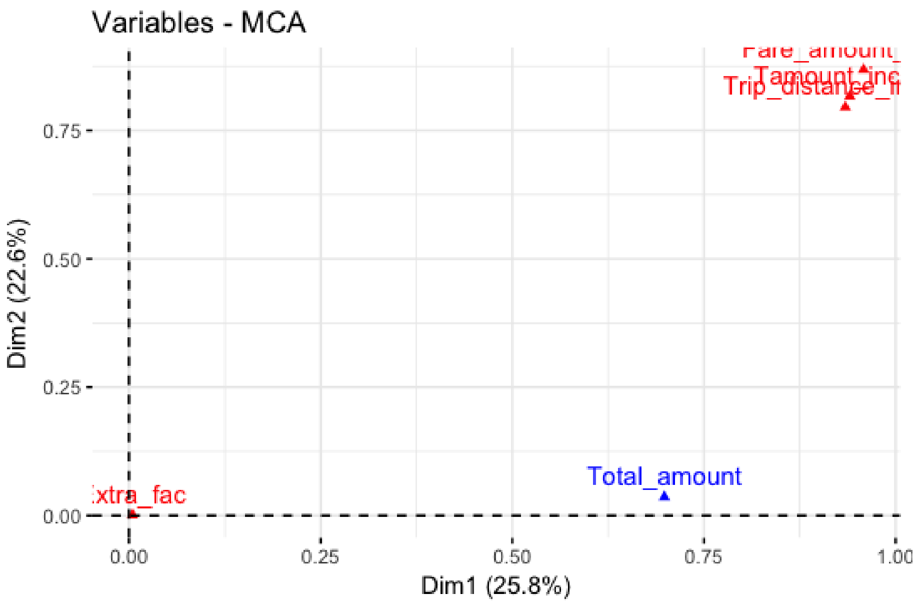
If a variable category is well represented by two dimensions, the sum of the cos2 is closed to one. For some of the row items, more than 2 dimensions are required to perfectly represent the data.



	R2	p.value
Tamount_inclass	0.940182014	0.000000e+00
Trip_distance_inclass	0.934443162	0.000000e+00
Fare_amount_inclass	0.958139867	0.000000e+00
Extra_fac	0.004601991	3.654608e-05

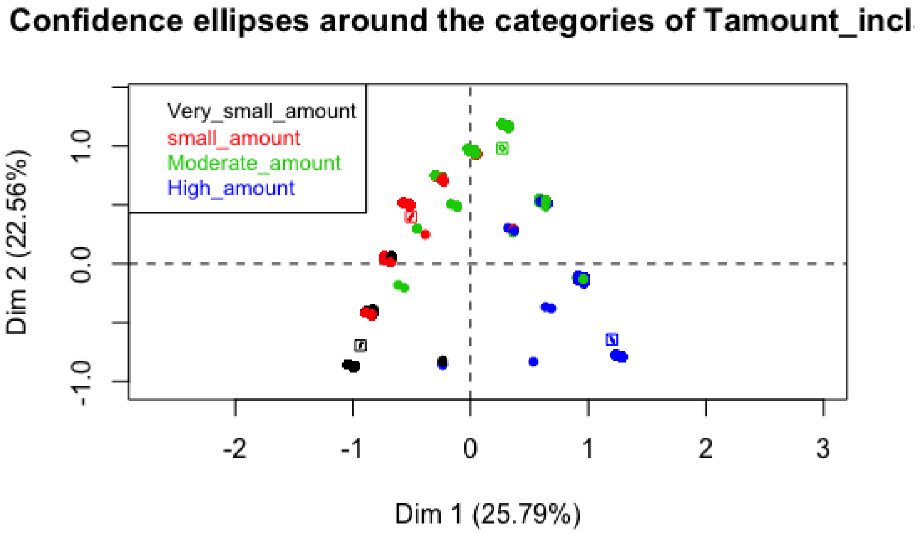
The strongest weighting dimensions are Dim 1 and Dim 2. There are weighted as follows:

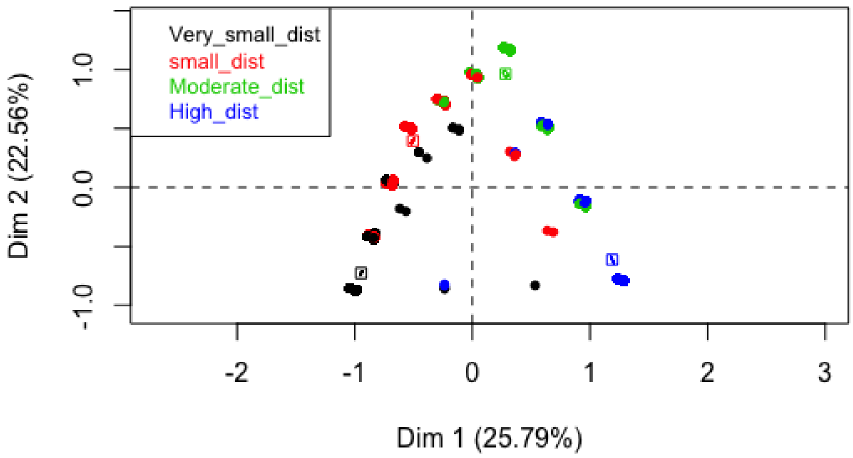
	R2	p.value	
	Dim.1	Dim.2	Dim.3
Tamount_inclass	0.940	0.817	0.623
Trip_distance_inclass	0.934	0.796	0.641
Fare_amount_inclass	0.958	0.869	0.769
Extra_fac	0.005	0.001	0.000



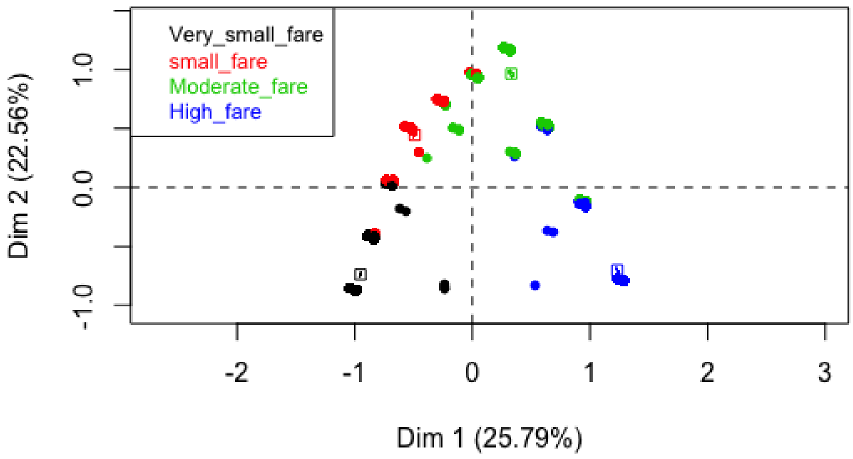
As you can see the Tax Amount, Trip distance and Fare Amount have large weightings in Dim 1 and Dim 2. Extra hardly correlates with these dimensions at all.

6 Ellipse plots

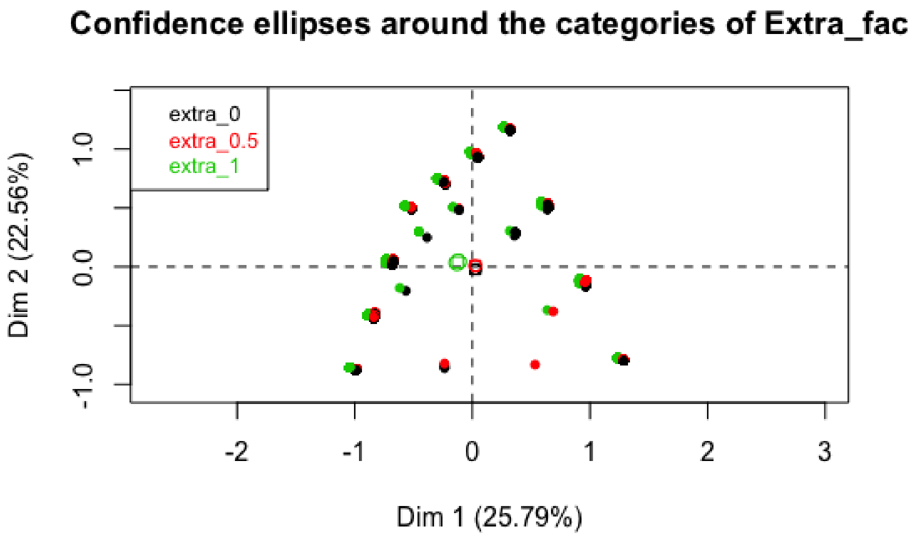




onfidence ellipses around the categories of Trip_distance_ir



onfidence ellipses around the categories of Fare_amount_in



7 Analysis

All graphs with the exception of the Extra graph seem to show a very clear clustering, while there is some overlap this is expected from the small set of possible outcomes.

The graph of Cos2 of the Variables in Dim 1 and 2 shows very clearly that High fare,amount and distance are all well represented by the dimensions chosen. Very small fares, amounts and distances are also well represented by these dimensions.

Dim 1 represents the High and Very small values well, meaning that there will be a large disparity in the horizontal axis of our plots. Dim 2 however represents Moderate values the most, so they should stand out in the vertical axis.

This again leaves Extra, as well as the small values as more difficult to distinguish, however they are still visually noticeable.