(it is anticipated that multidimensional statistical methods will inform about important relationships between the target variable and the others.)

Materials

A random sample containing 5000 registers extracted from the green taxi archive

Initial pre-treatment

First points:
Variables originally considered as "quantitative" but corresponding to qualitative concepts have to be recoded as "factors".

Error detection:
For each variable, errors and inconsistencies have to be detected and recoded as missing (using first the summary function and then, if the variable is quantitative, a boxplot which allows for detecting outliers. These outliers are either correct values or erroneous values.)

Sometimes, there is no doubt (for example, a negative Total Amount is an error) but in other cases it is difficult to conclude (a very high tip is an error or not). In this latter case, make a decision but explaining the rule that support it. In the case of the latitude and longitude, we should detect the trips taking place outside of New-York (neither pick-up nor dropoff points located in New-York); these trips have to be counted and deleted (these data are said to be "outside the scope").
However, we keep the trips with null latitude and longitude (these values are missing). 4.

Profiling

• Numeric Target (Total Amount)
• Factors: (Final Decision, Non-cash TipisGiven). Discretized into five levels Total Amount

We should then create a report on Data Quality and Profiling, including: an outline (partial, at this moment) and a short description of the strategy and methods, properly commented

Concerning the individuals:
We should create a new variable which is "missing values count" and describe it in the same way as the other quantitative variables. Further, when studying the multidimensional methods, we will see how to use this variable to detect its relationship with the others. · Comment the results, mainly how many individuals present more missing values than the other ones.

We also should report, if any, the eliminated "out of scope" individuals

Concerning profiling results:
A short comment about what the profiling highlighted. In particular, what was expected and what is "discovered".
We should pay special attention to compare which information profiling "Total amount" as a quantitative variable versus profiling "Total amount" as a qualitative variable has given.

PCA

The main objectives are to understand the relationships between the variable Total_amount and the other variables.

Supplementary coding
We should recode the pickup and dropoff variables into factore.
It could be interesting to create the variable "actual velocity during the trip" and also the variable computing the part of the amount due to the total distance of the trip.

Analysis of the variables influencing the total-amount through PCA
A PCA applied to these variables will provide a synthesis of them. The other variables, available in the file, have to be used as illustrative.

Clustering
A trip typology is obtained by clustering the trips from their principal coordinates.
An appropriate number of principal componenets has to be selected in the clustering step.
The clusters will be described from all the variables.

First conclusions are drawn on the results obtained through both complementary methods (ACP + clustering)

Performing the analysis
You apply the strategy that you have chosen and obtain the results. The interpretation of these results will constitute the main part of the report.

Results report
- concerning the factorial methods: succession of eigenvalues and inertia ratios, number of retained factors (justified), “names” of the factors that summarize their interpretation representation of the results though clear and worked graphics.
- clustering results: number of clusters, justification, interpretation
- commenting the complementarity of the results of factorial methods and of classification,
- results on diet and diet patterns

Global conclusions
Synthesis of the results, critical comment on the initial data, conclusions but also new hypotheses that arise, new studies that have to be performed.