21st of October 2022

Data mining for Airbnb New York

Presented by Xavier Marti Llull, Mario Font Blanc, Ramon Ribas Domingo, Ricard Guixaró Trancho, David Daniel Streuli

Outline of talk

Metadata

Preprocessing

Basic statistical descriptive analysis

PCA analysis

Clustering

Conclusions

Basic structure of the data

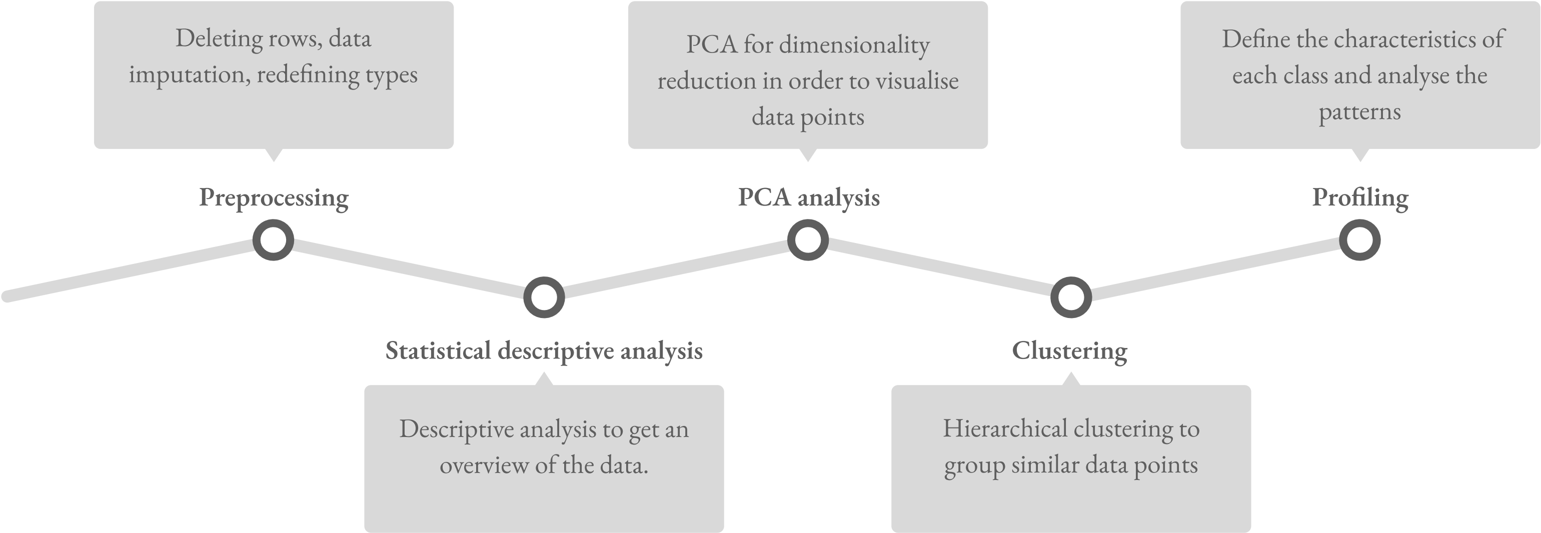
Number of records	5000
Number of variables	21
Number of numerical variables	13
Number of binary variables	2
Number of date variables	1
Number of qualitative variables	5

https://www.kaggle.com/datasets/arianazmoudeh/airbnbopendata.

Goals

1. Preprocessing the data and prepare it for further analysis
2. Visualising the data using different data mining techniques
3. Find relationships between the different data points and dimensions

Complete data mining process



Preprocessing

Deleting rows, data imputation, redefining types

Statistical descriptive analysis

Descriptive analysis to get an overview of the data.

PCA analysis

PCA for dimensionality reduction in order to visualise data points

Clustering

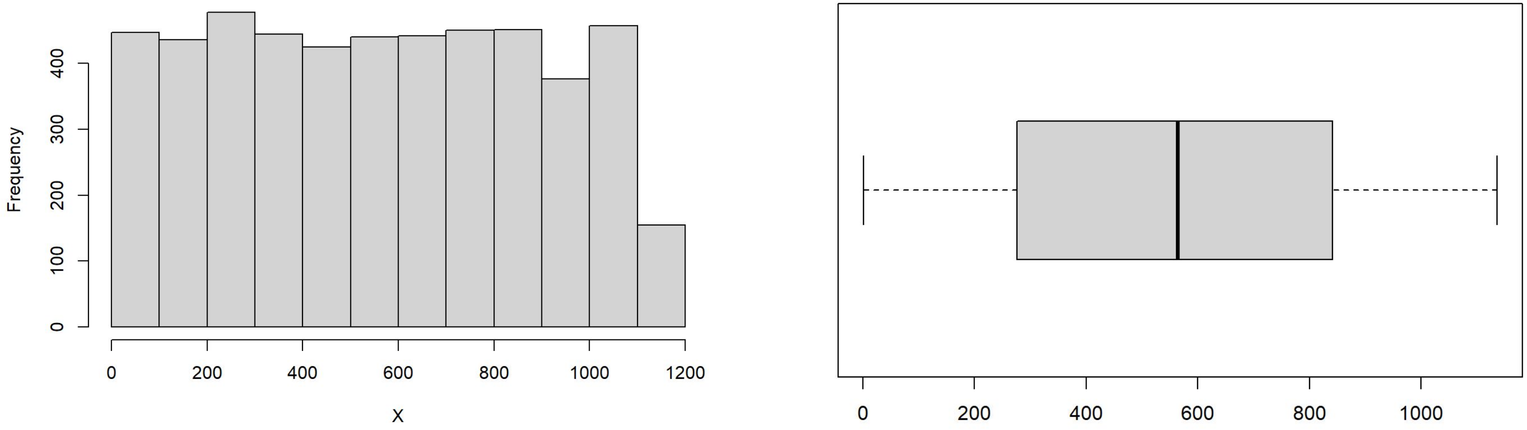
Hierarchical clustering to group similar data points

Profiling

Define the characteristics of each class and analyse the patterns

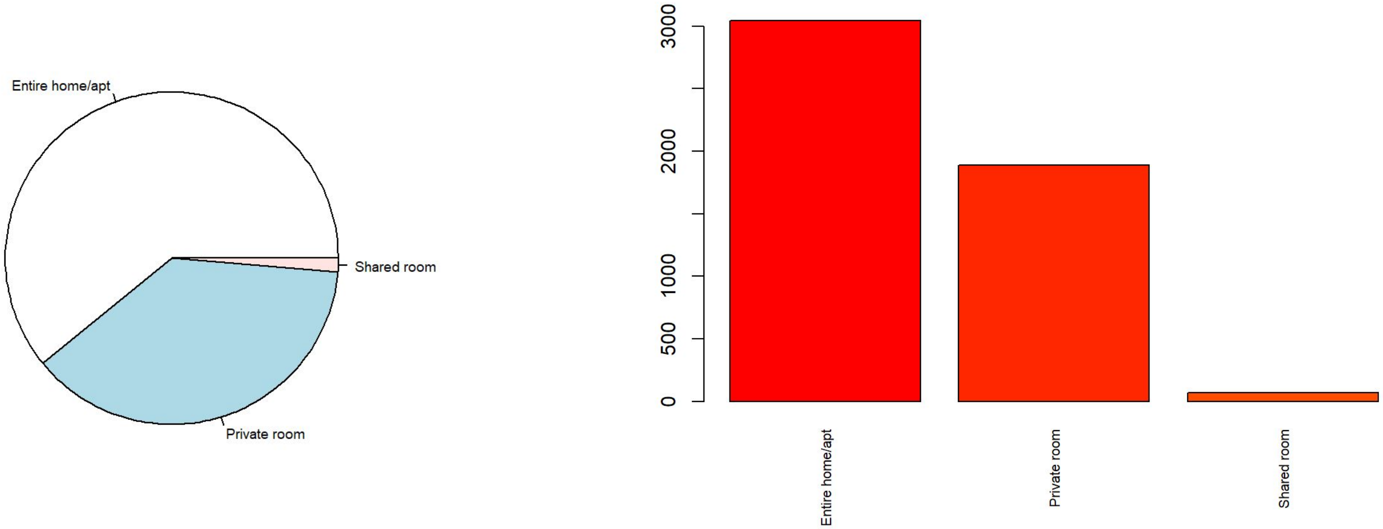
Basic statistical descriptive analysis

In the following slides we present various descriptive charts of our data



Histogram and box plot of ***price***

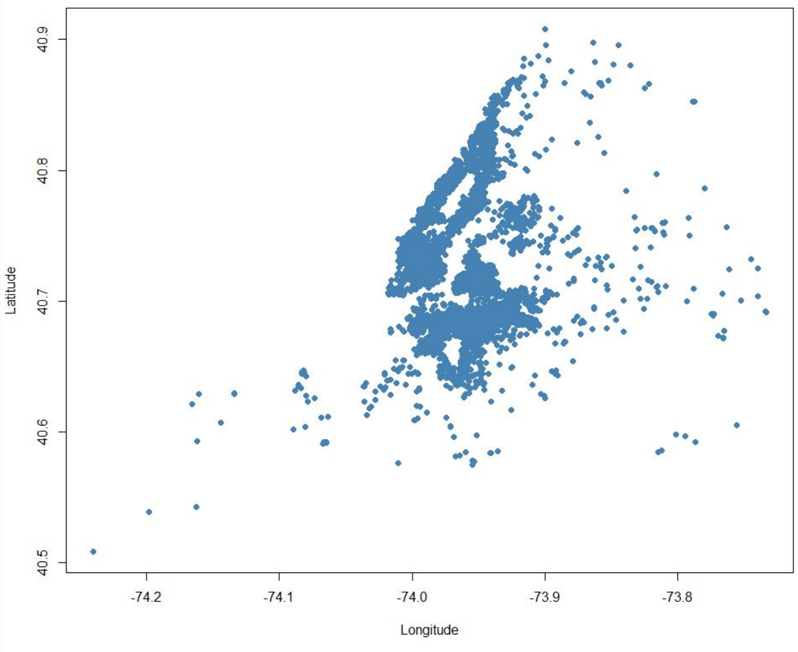
Minimum value	$1
Maximum value	$1136
Mean	$563.1



Entire home/apt	61%
Private room	38%
Shared room	1%

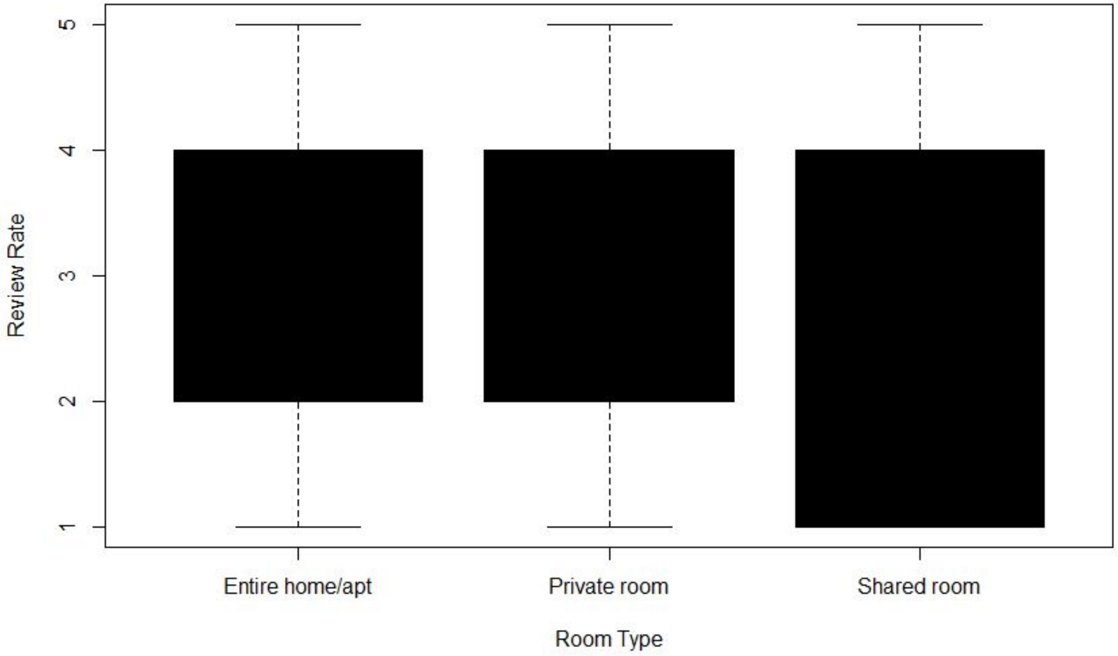
Pie-chart and bar plot of ***room.type***

Bivariate descriptive analysis

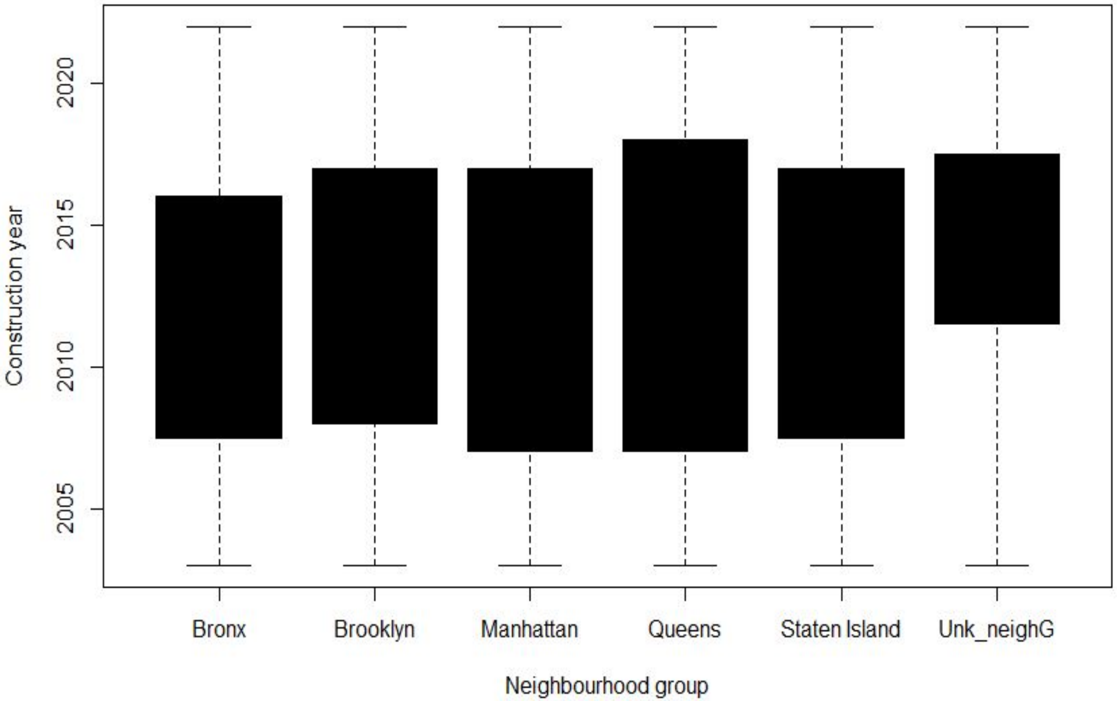


Latitude **vs** Longitude

Review rate **vs** Room type



Construction year **vs** Neighbourhood group



Preprocessing

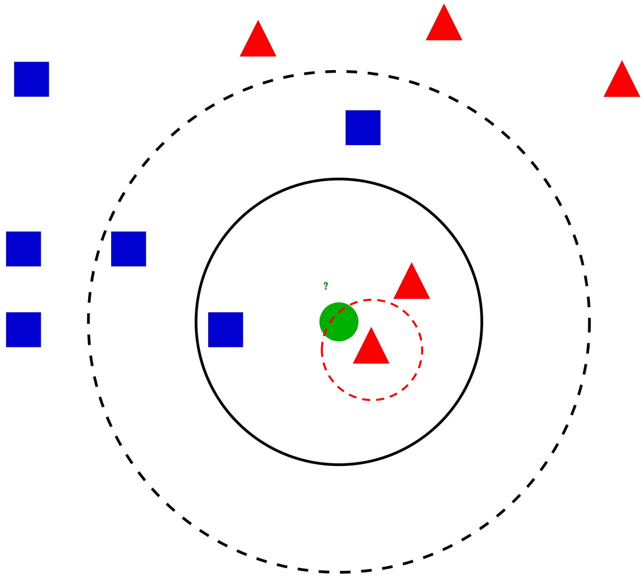
1. Deletion of some rows for efficiency reasons
2. Redefining the type of some variables
3. Data imputation using KNN
4. Outlier detection

KNN method

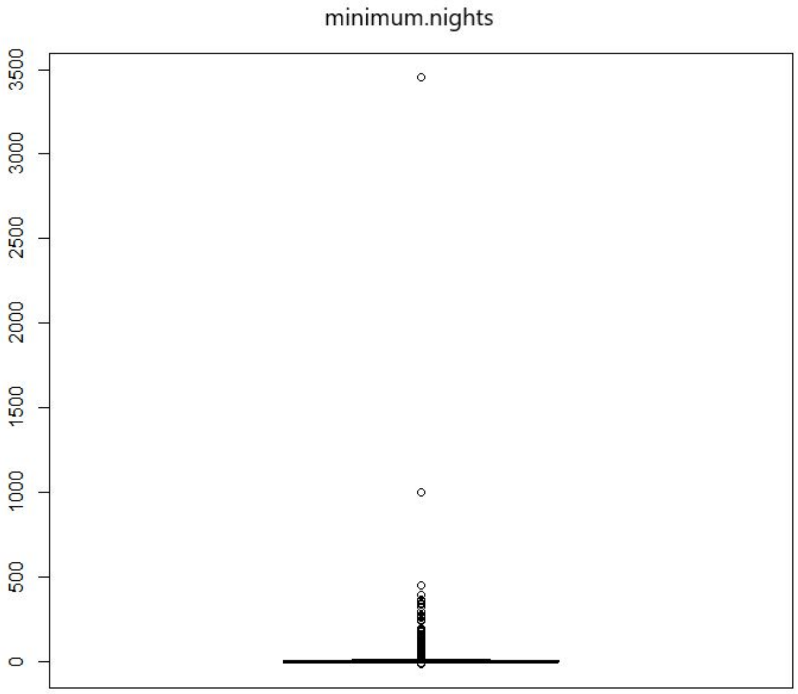
1 nearest neighbour for data imputation

data point with missing values for some columns

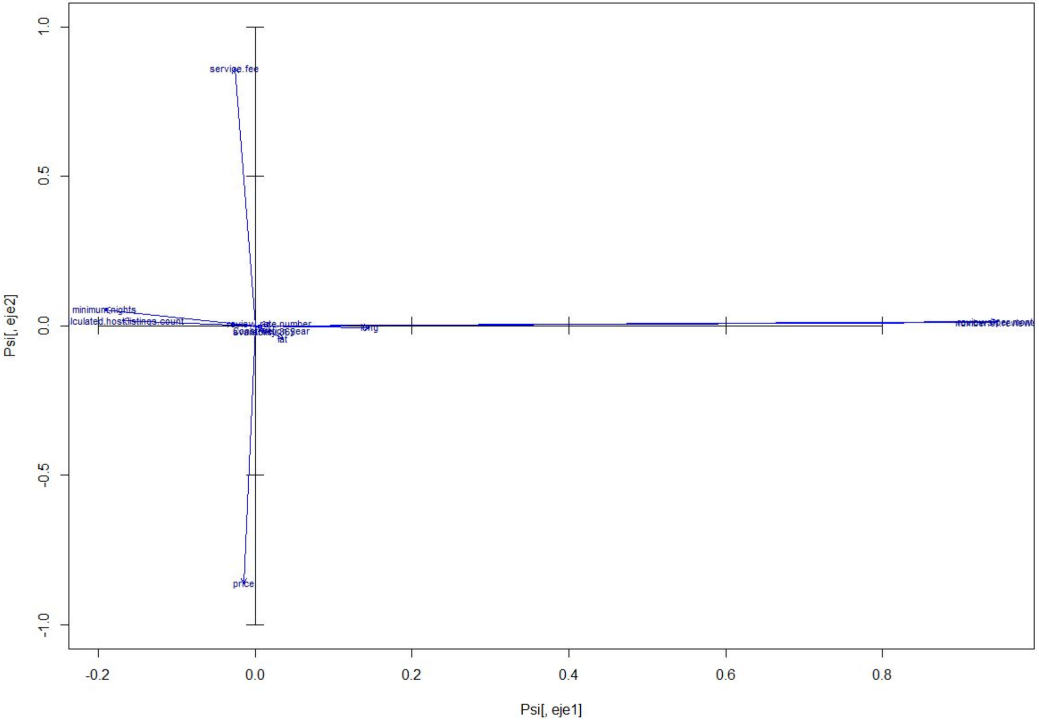
nearest neighbour



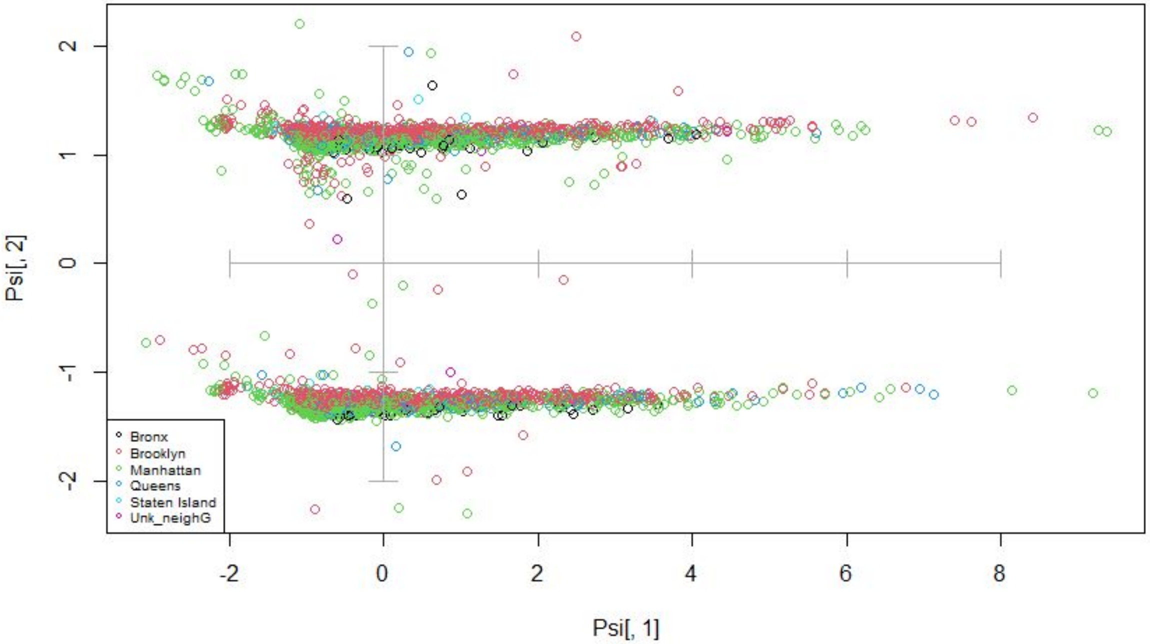
Outlier detection



PCA analysis

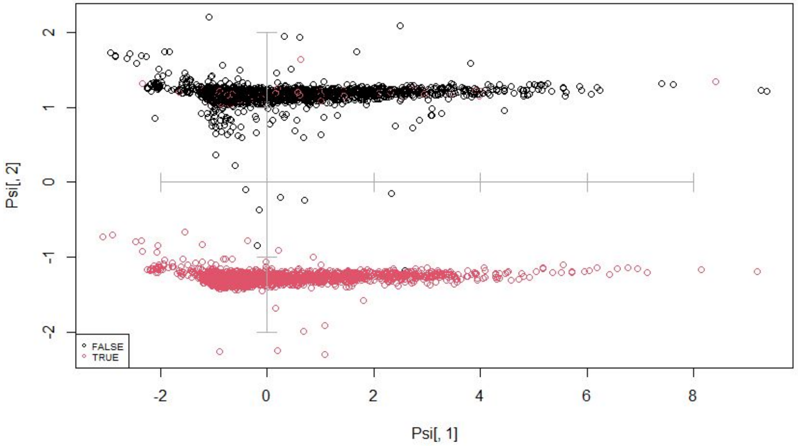


PCA analysis

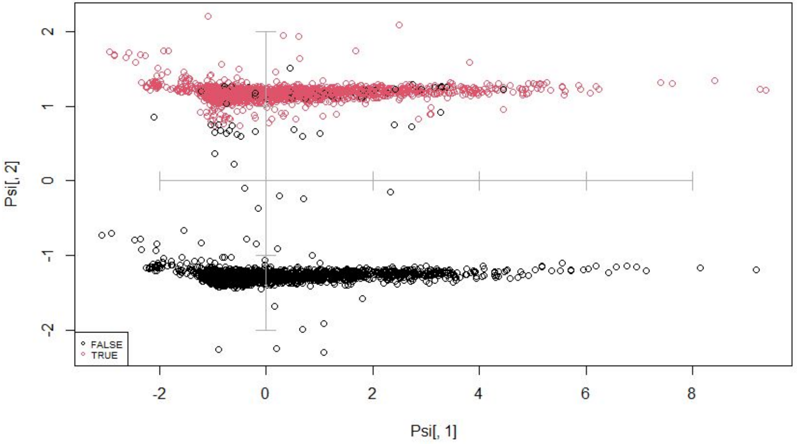


neighbourhood.group

PCA analysis

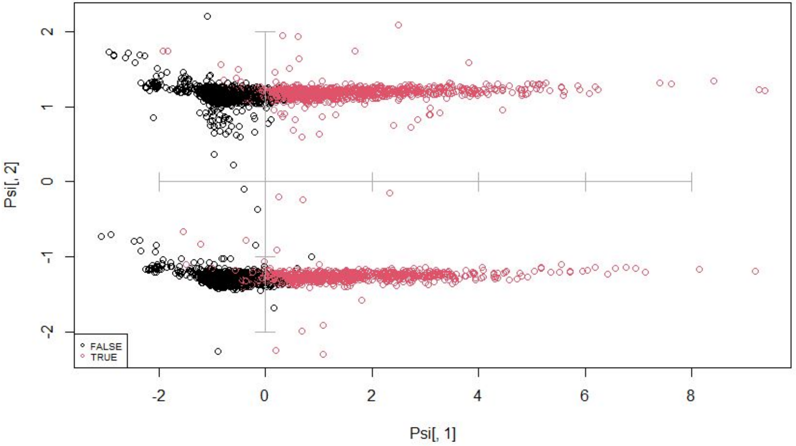


above.mean.price

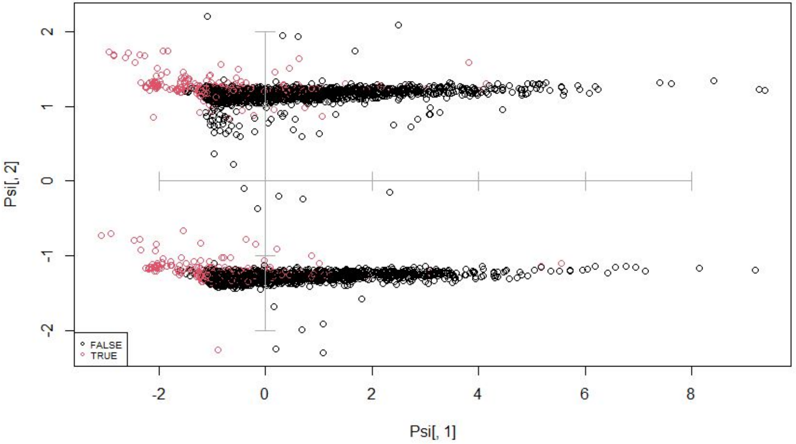


above.mean.service.fee

PCA analysis

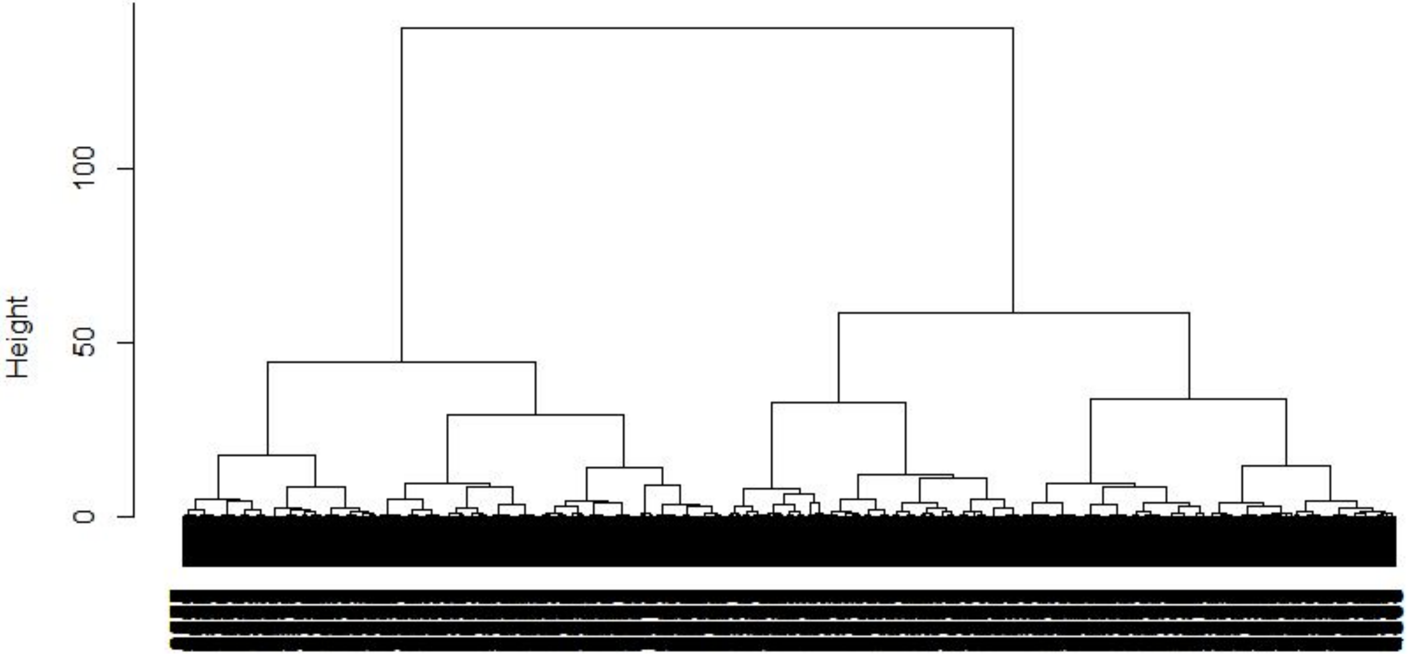


above.mean.reviews.per.month

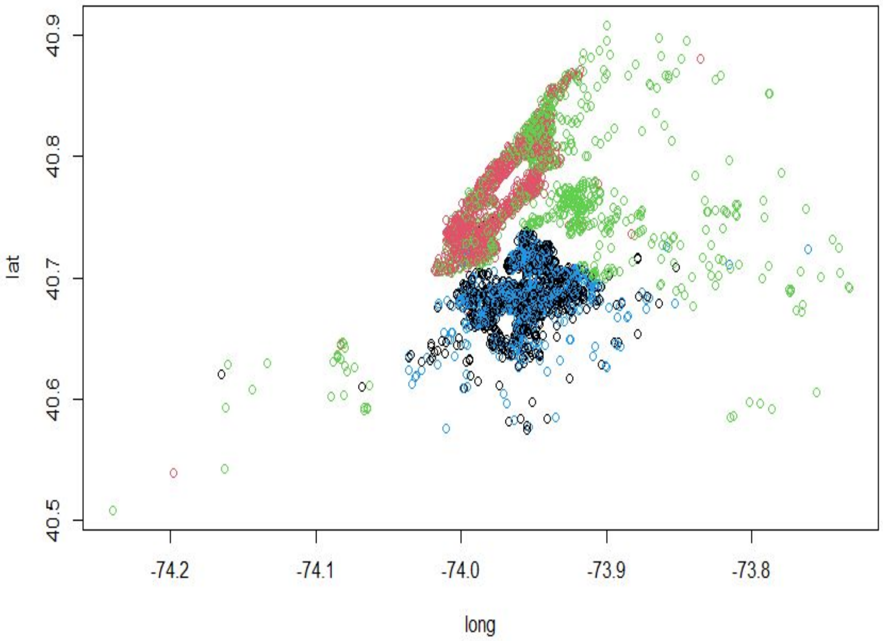


above.mean.minimum.nights

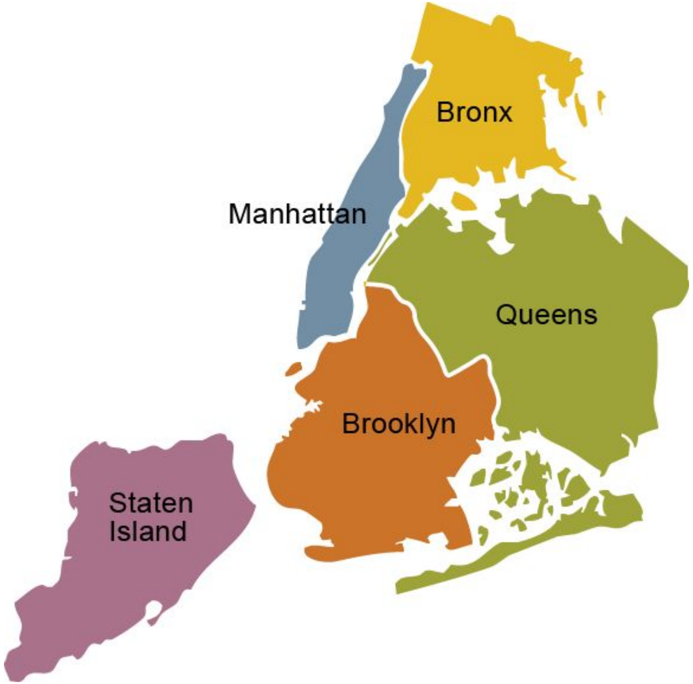
Clustering



Locational membership

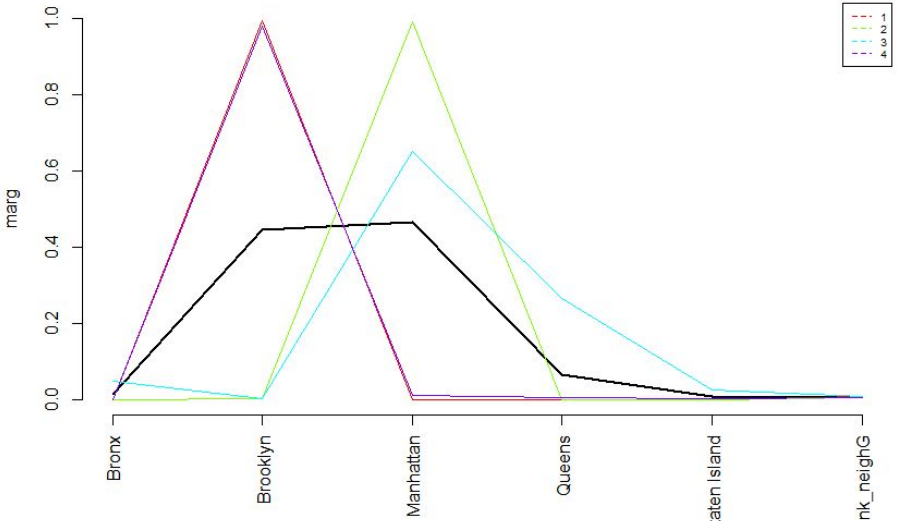


Actual map of New York

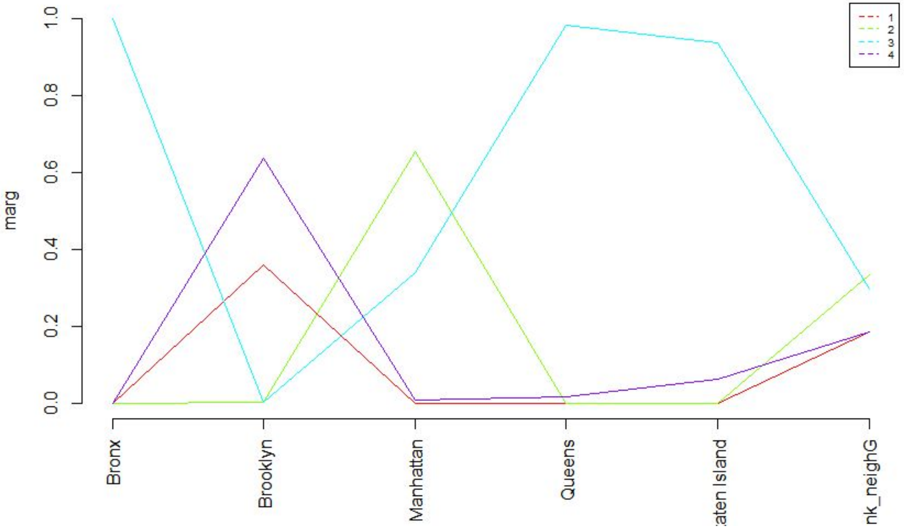


Profiling variables

neighbourhood.group

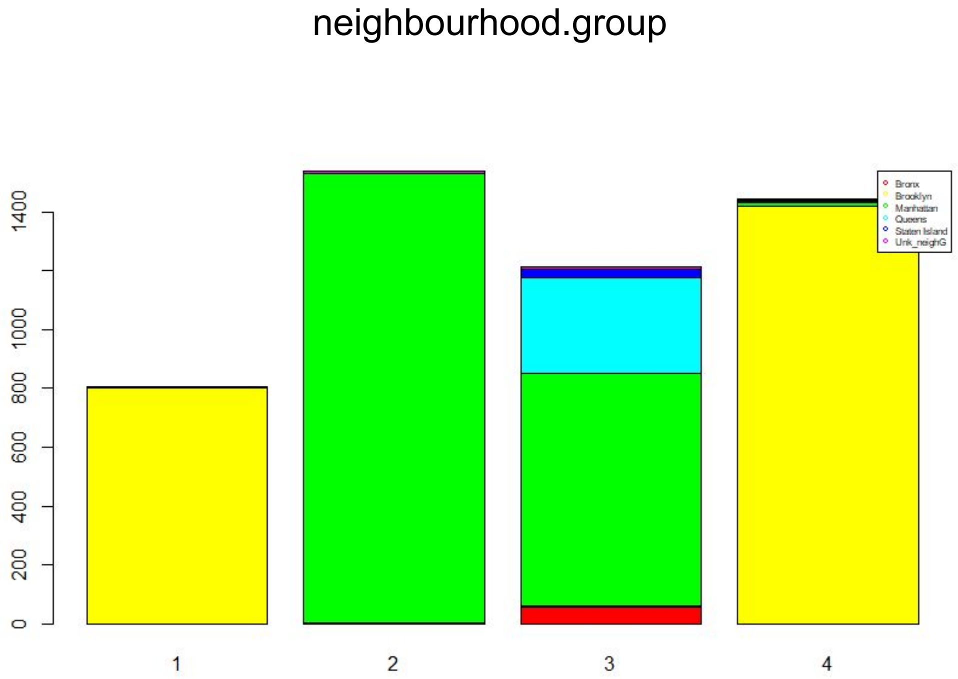


Distribution by class



P value of classes

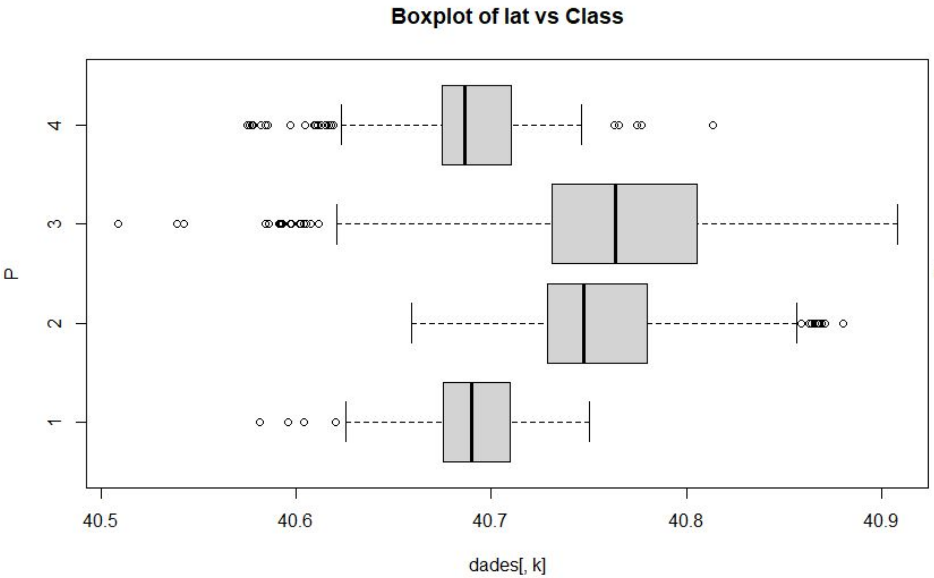
Profiling variables

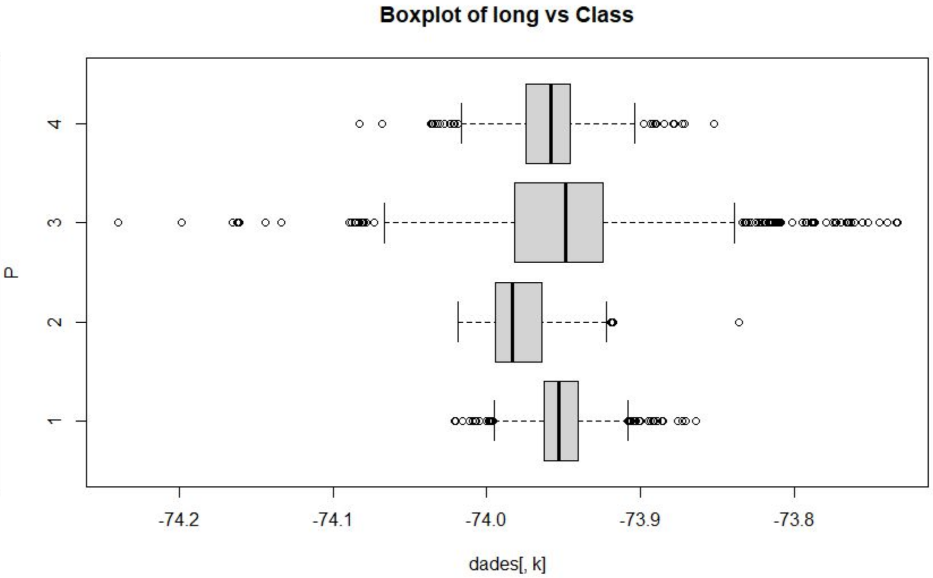


Class histogram

Profiling variables

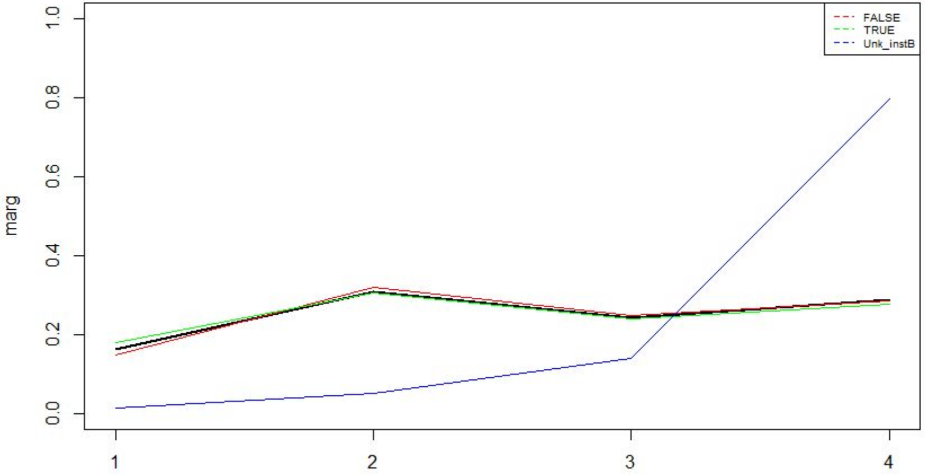
lat and long



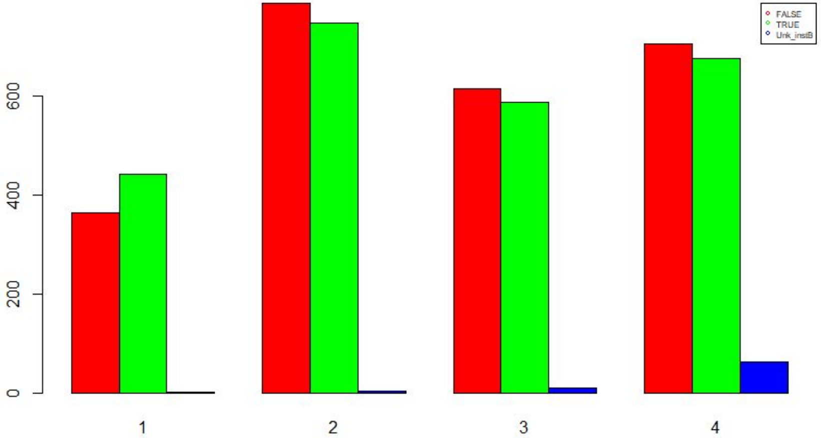


Profiling variables

Instant.bookable



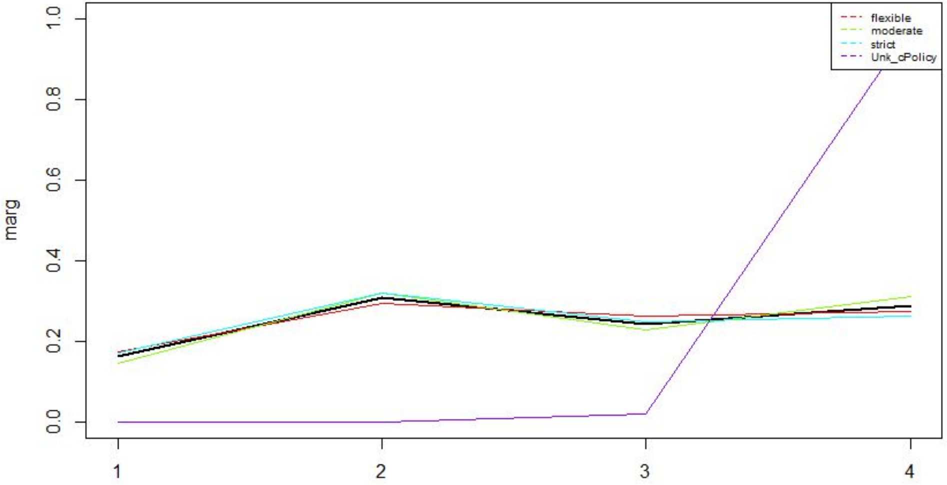
P value of classes



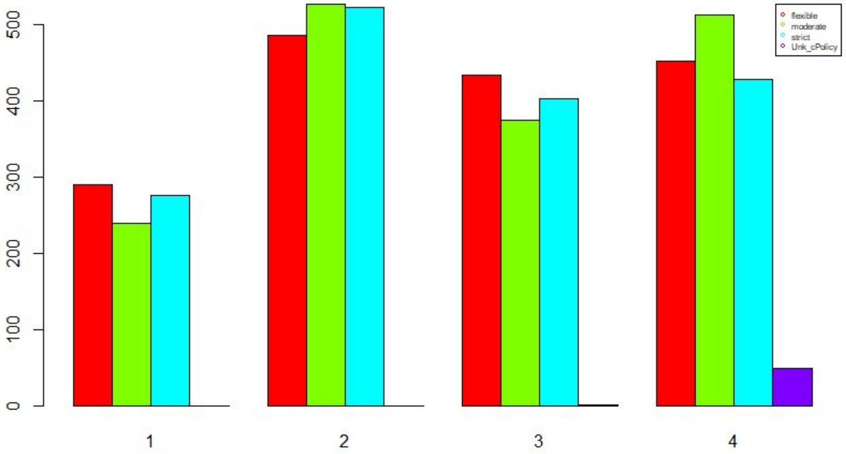
Class histogram

Profiling variables

cancellation.policy



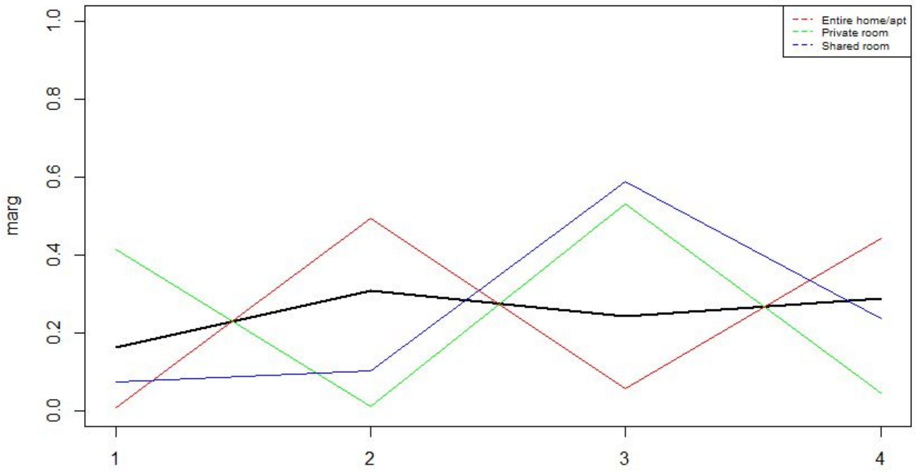
P value of classes



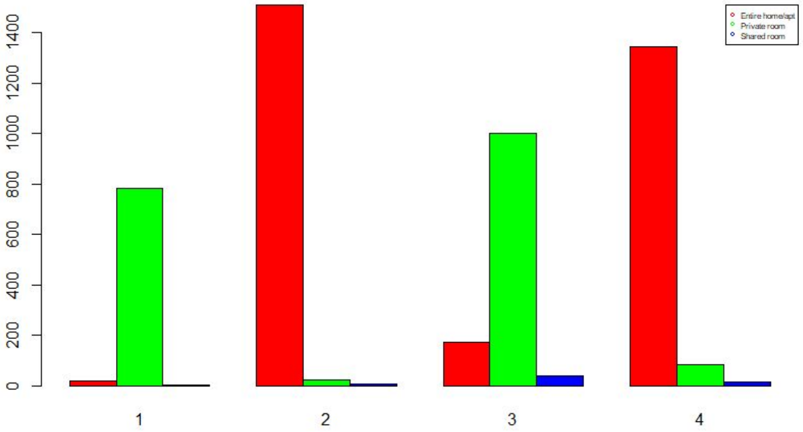
Class histogram

Profiling variables

room.type



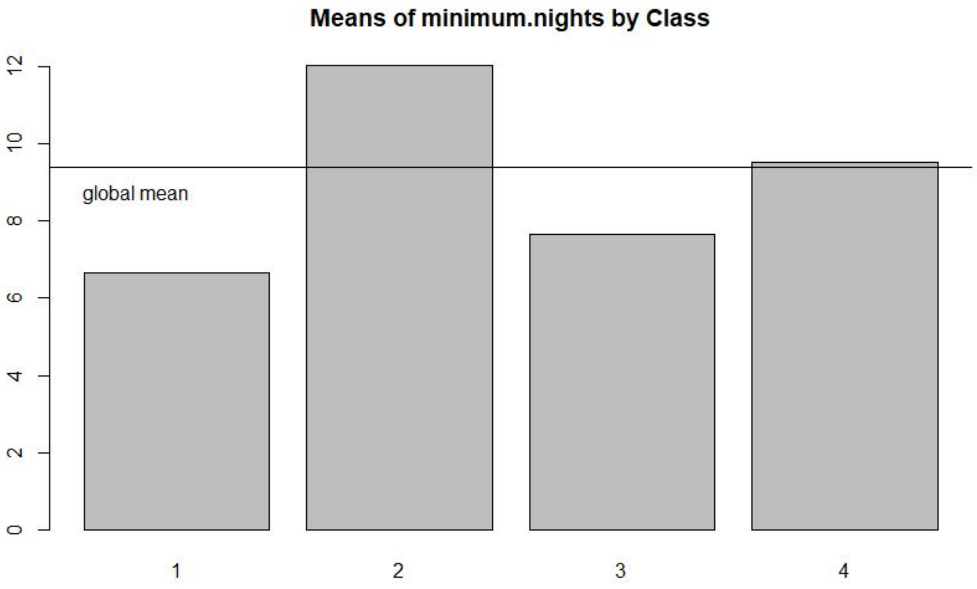
P value of classes



Class histogram

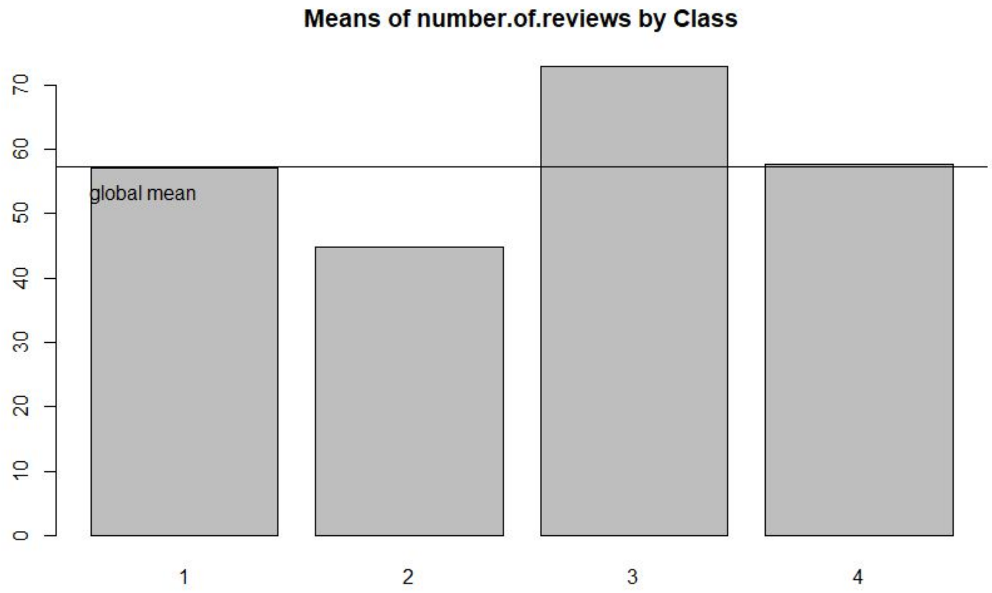
Profiling variables

minimum.nights



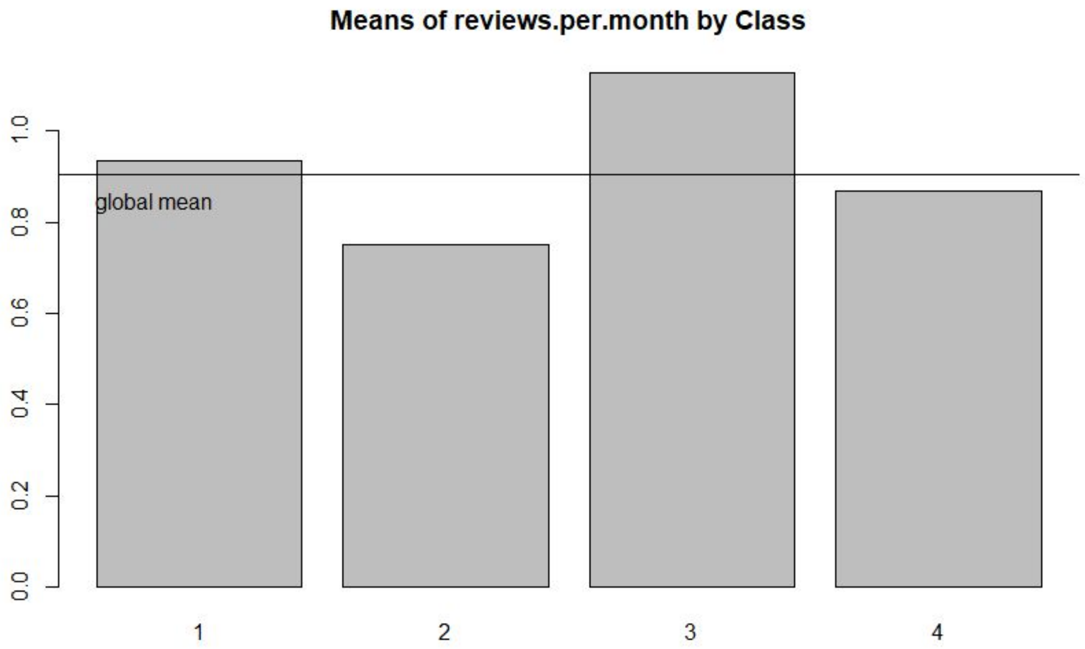
Profiling variables

number.of.reviews



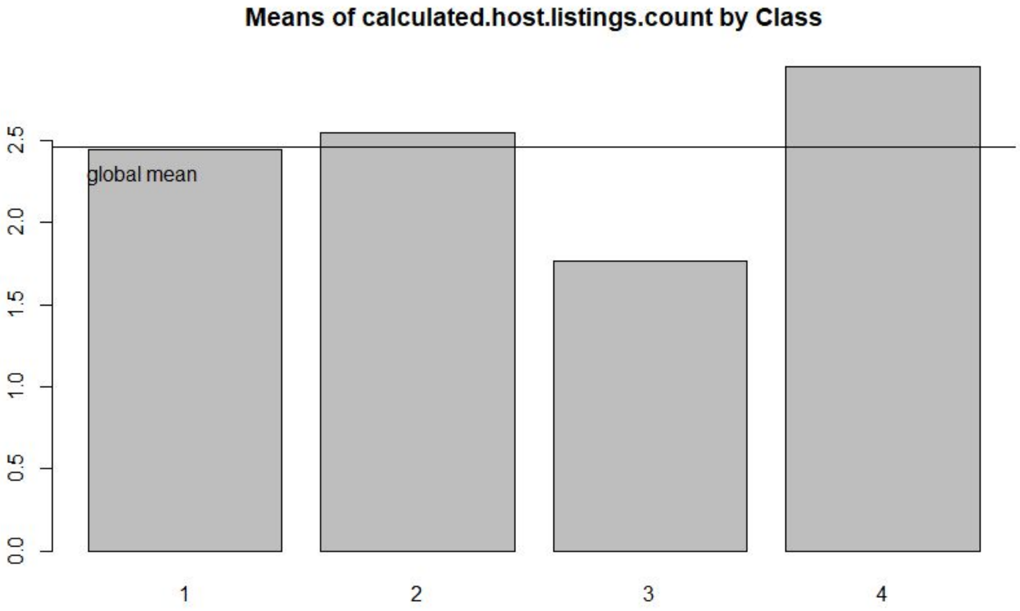
Profiling variables

reviews.per.month



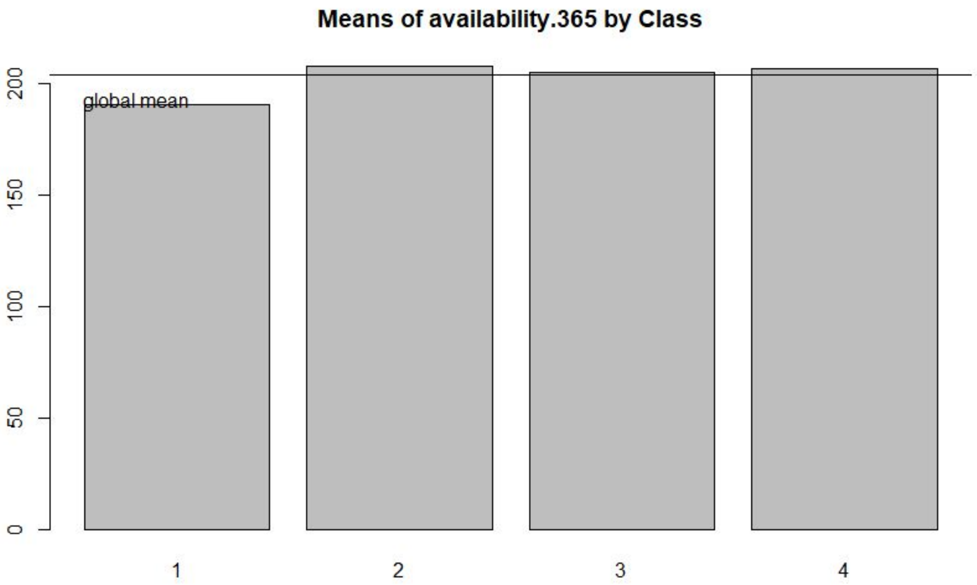
Profiling variables

calculated.host.listings.count



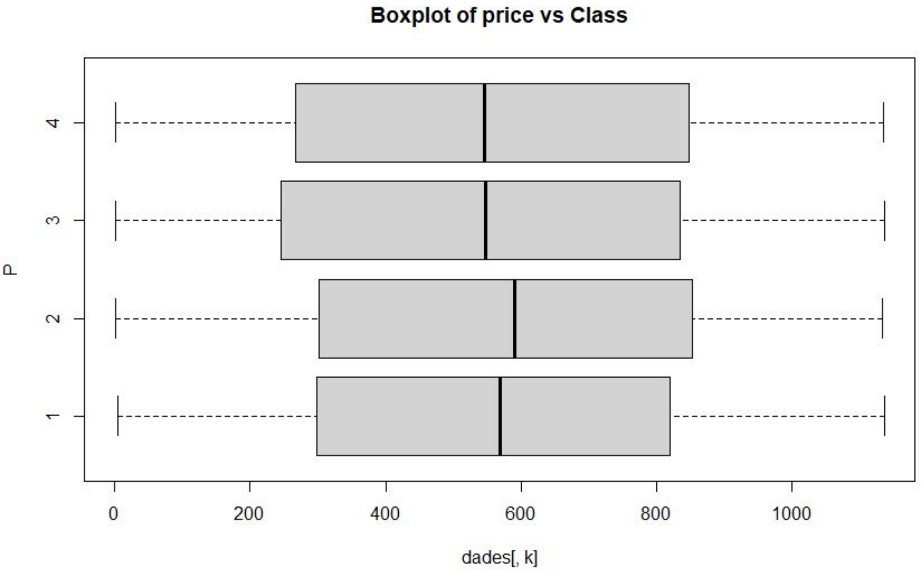
Profiling variables

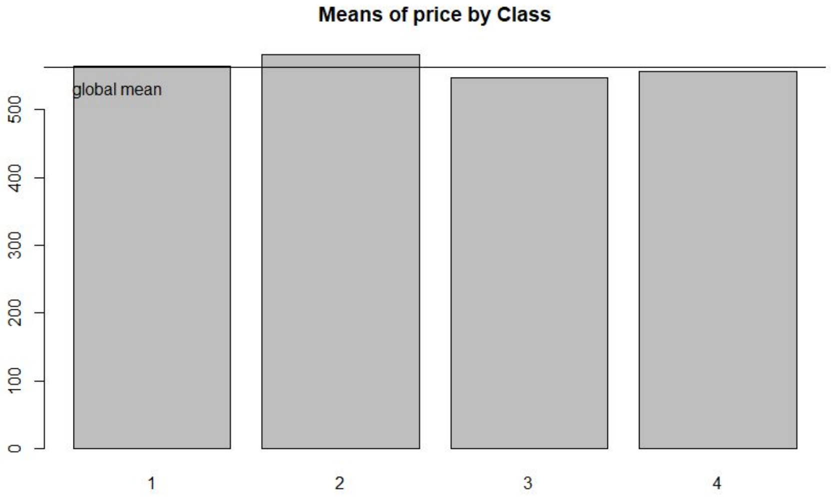
availability.365



Profiling variables

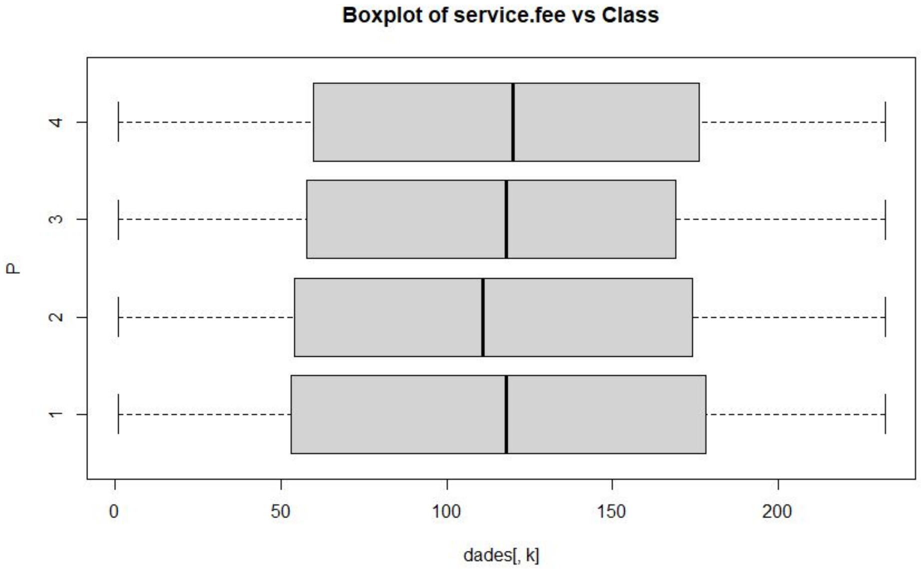
price


Boxplot of price vs Class

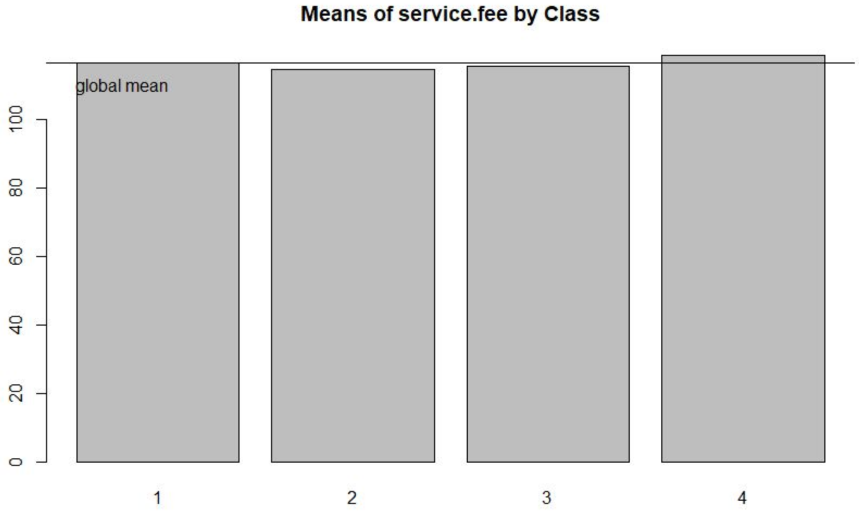

Means of price by Class

Profiling variables

service.fee



Boxplot of service.fee vs Class



Means of service.fee by Class

Cluster Description

Cluster 1

- Brooklyn
- Private rooms
- Low minimum nights
- High reviews per month
- Instant bookability

Cluster 2

- Manhattan
- Full apartments
- High minimum nights
- Low reviews per month

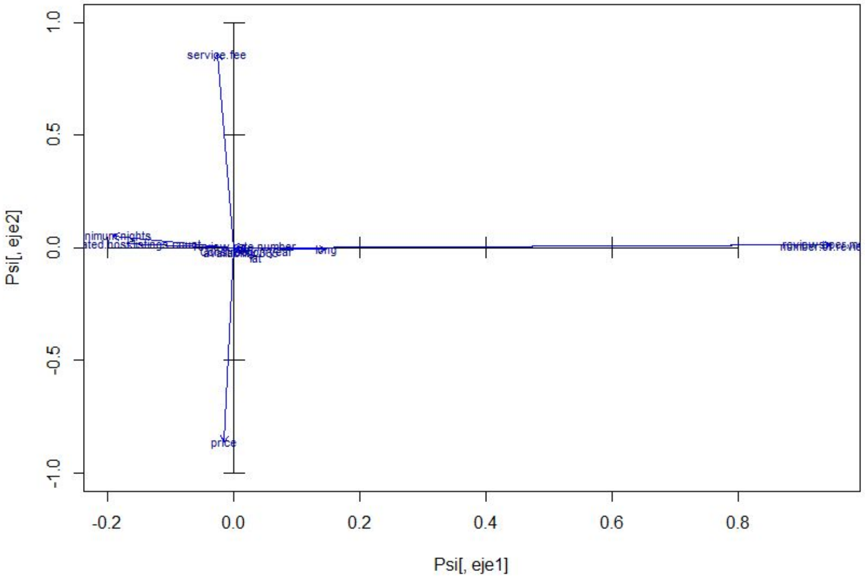
Cluster 3

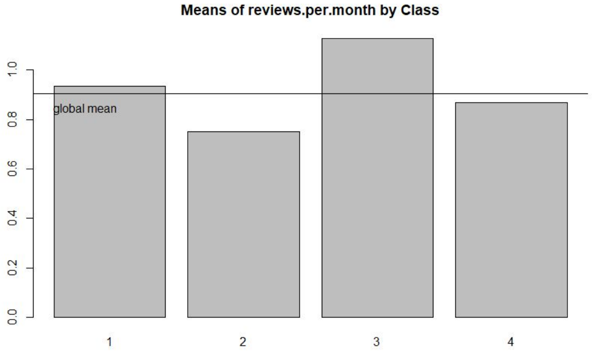
- Location diversity
- Has most shared rooms
- Mostly private rooms
- Low minimum nights
- High reviews per month
- Cheapest

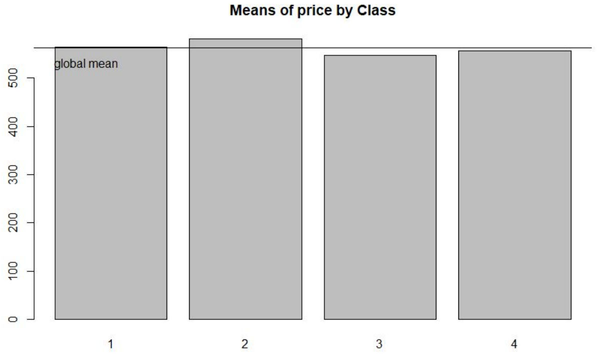
Cluster 4

- Brooklyn
- Full apartments
- High minimum nights
- Low reviews per month
- Highest hosts listings
- Unknown data

Comparison of conclusions between PCA and clustering





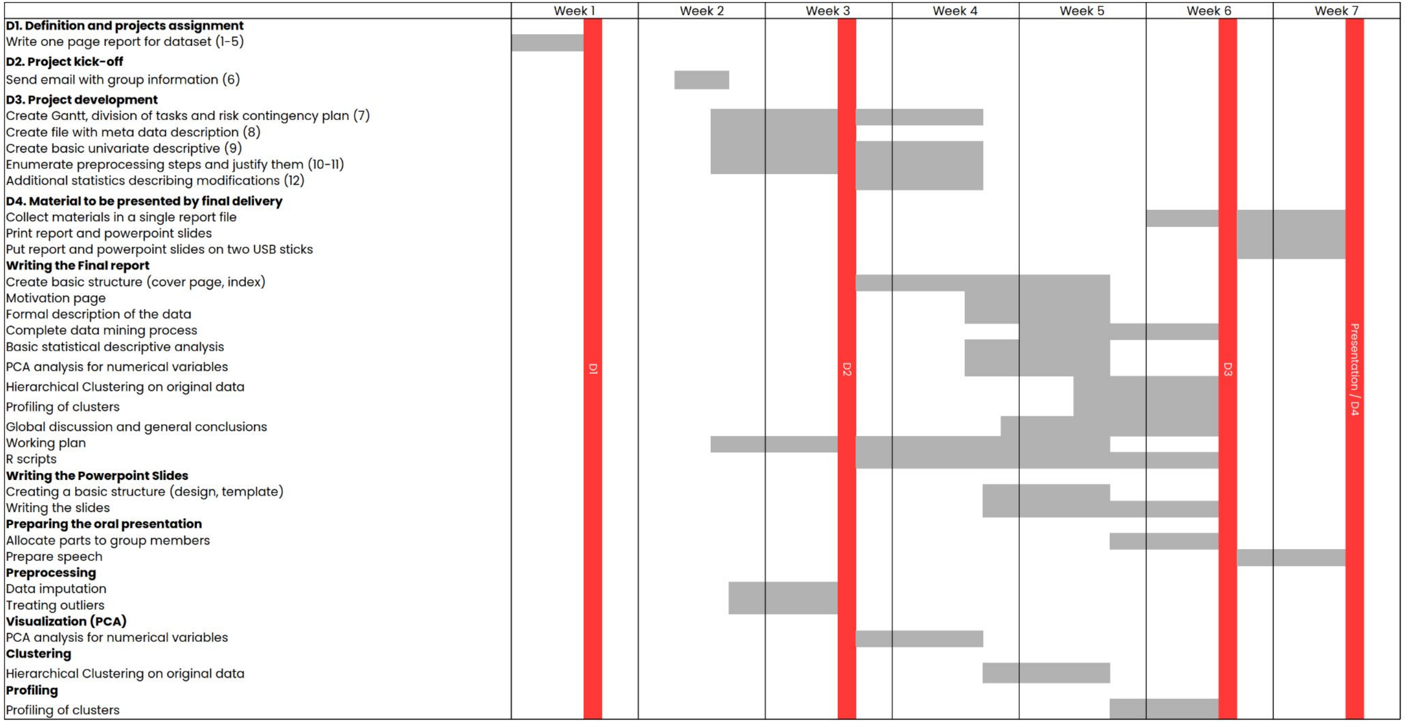


Conclusions

Good overview of tourist accommodation data in New York

Could be used for market research in tourist sector or city planning

Original and final scheduling



D1. Definition and projects assignment
Write one page report for dataset (1-5)
D2. Project kick-off
Send email with group information (6)
D3. Project development
Create Gantt, division of tasks and risk contingency plan (7)
Create file with meta data description (8)
Create basic univariate descriptive (9)
Enumerate preprocessing steps and justify them (10-11)
Additional statistics describing modifications (12)
D4. Material to be presented by final delivery
Collect materials in a single report file
Print report and powerpoint slides
Put report and powerpoint slides on two USB sticks
Writing the Final report
Create basic structure (cover page, index)
Motivation page
Formal description of the data
Complete data mining process
Basic statistical descriptive analysis
PCA analysis for numerical variables
Hierarchical Clustering on original data
Profiling of clusters
Global discussion and general conclusions
Working plan
R scripts
Writing the Powerpoint Slides
Creating a basic structure (design, template)
Writing the slides
Preparing the oral presentation
Allocate parts to group members
Prepare speech
Preprocessing
Data imputation
Treating outliers
Visualization (PCA)
PCA analysis for numerical variables
Clustering
Hierarchical Clustering on original data
Profiling
Profiling of clusters