P1 - Car sales in US states

A data-driven analysis

Alex Beauchamp - Otman Ezzayat Maid - Joan Gómez Català - Matija Jakovac - Álvaro Monclús Muñoz

Delivery date : 24th of March 2025

Format of presentation

1. Introduction
2. Data and preprocessing
3. Analysis, PCA, clustering
4. Profiling and interpretations
5. Final conclusions

1. Introduction

Presentation of dataset

- Origin: Kaggle.com
- Used car sales across the US
- 8128 records
- Key variables:
  - City and state
  - Selling price
  - Number of owners
  - Brand name
  - Km driven

Variable	Short Name	Modalities	Short Mod Name	Meaning	Type	Units	Missing Code	Measuring Procedure	Range	Role
Sales_ID	Id			ID of the sale	Quali					Explanatory
name	Name			Name of the car	Quali					Explanatory
year	Year			Year of fabrication	Num	years			[1994, 2020]	Explanatory
selling_price	Price			Price it was sold at	Num	euros			[330, 71753]	Explanatory
km_driven	Km			Km driven	Num	km			[1000, 577414]	Explanatory
Region	Region			Region where it was sold	Quali					Explanatory
State or Province	State			State or province where it was sold	Quali					Explanatory
City	City			City where it was sold	Quali					Explanatory
fuel	Fuel			Type of fuel it uses	Quali					Explanatory
seller_type	Seller			Type of seller	Quali					Explanatory
transmission	Trans			Type of transmission	Boolean					Explanatory
		Manual	M							
		Automatic	A							
owner	Owner			Number of owner	Quali					Explanatory
mileage	Mileage			Mileage	Num	mpg			[9, 33.44]	Explanatory
engine	Engine			Engine displacement	Num	cm3			[624, 3604]	Explanatory
max power	MaxPower			Maximum power of the engine	Num	hp			[32.80, 282.00]	Explanatory
seats	Seats			Number of seats	Num	seats			[4, 10]	Explanatory
sold	Sold			Sold status	Boolean					Explanatory
		Yes	Y							
		No	N							

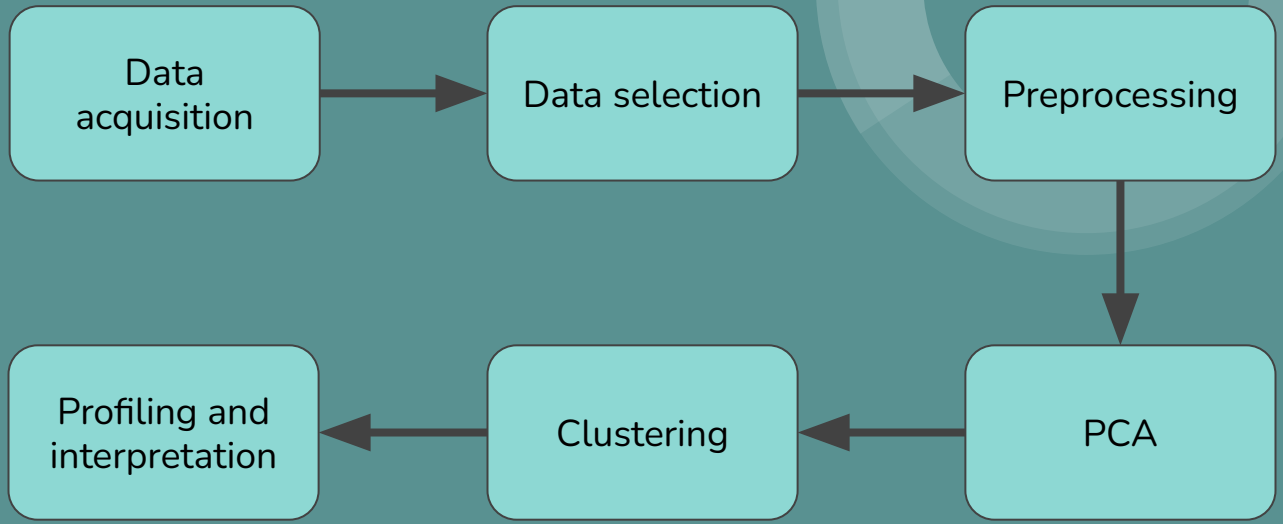
Kaggle URL : https://www.kaggle.com/datasets/shubham1kumar/usedcar-data

Primary objectives

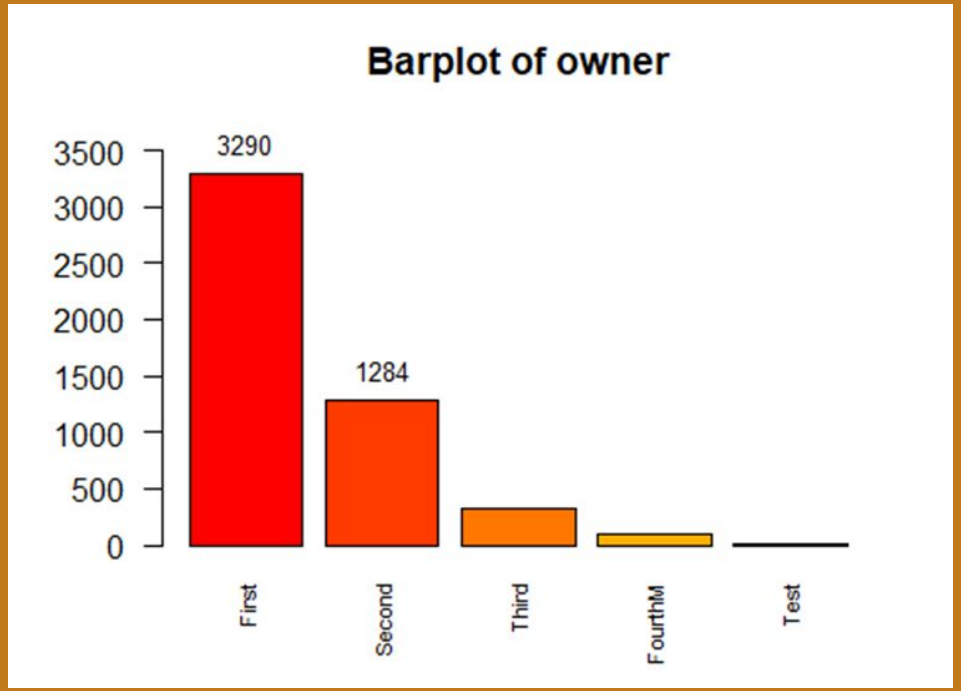
- Understanding pricing factors
- Predicting sales likelihood
- Consumer and seller behaviour
- Regional market variations

2. Data and preprocessing

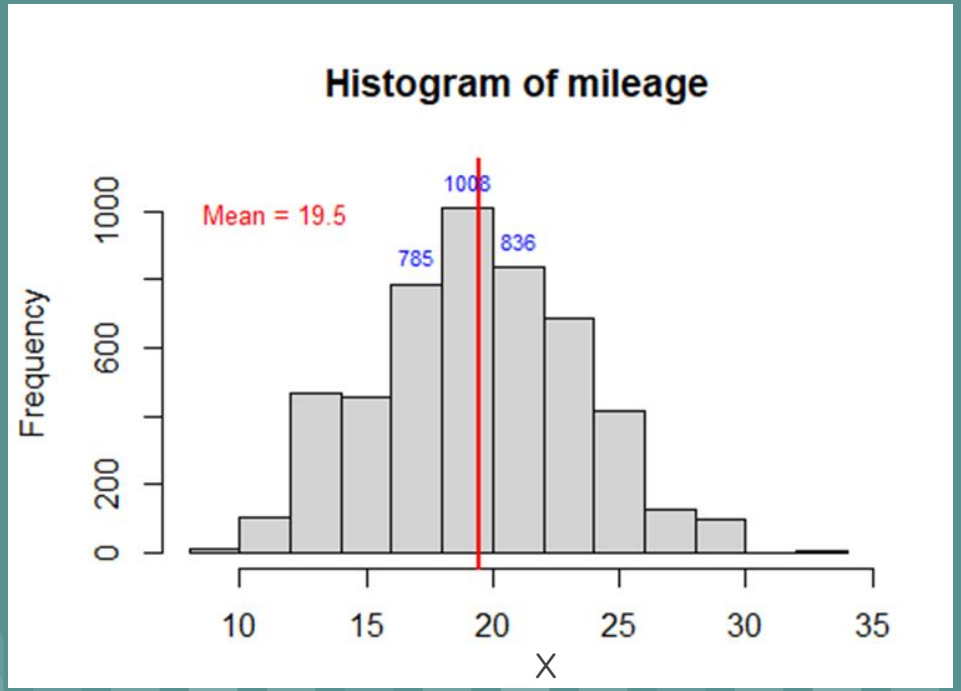
Data mining process

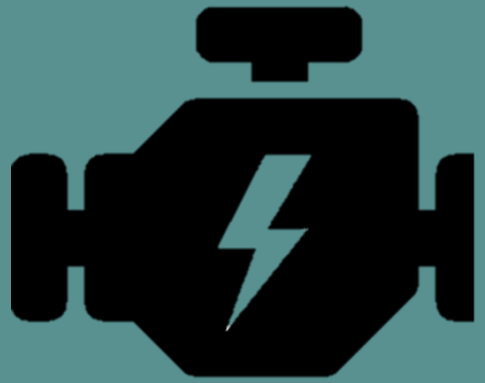


Qualitative descriptive - Owner



Numerical descriptive - Mileage



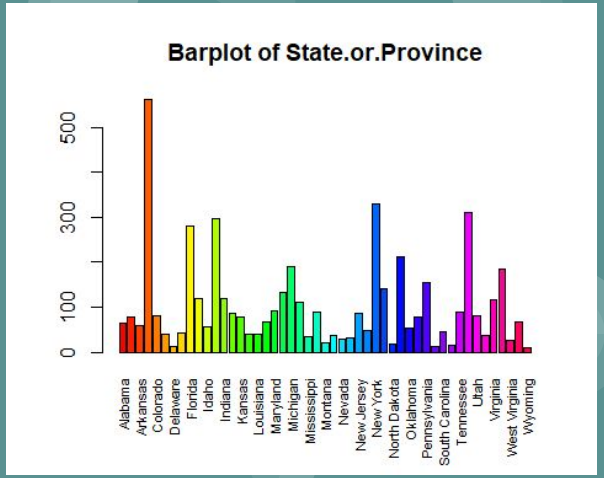


Synthesis of univariate descriptive

- Broad information about qualities of cars
- NUMERICAL VARIABLES:
  - Large ranges of values
  - Represented with histograms and boxplots
  - Big amount of extreme values (outliers)
- CATEGORICAL VARIABLES:
  - Short ranges of values (except Brand Name/State)
  - Represented with histograms and pie plots

Issues with descriptive

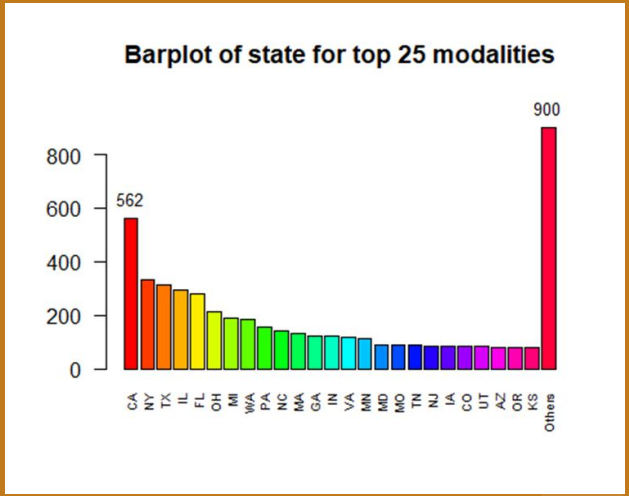
- **Outliers**
- **Number of modalities**
- **Uninteresting graphs and variables**



This is what it looked like with all the modalities included

Preprocessing steps

- **Shortening names**
- **Managing outliers**
- **Removal of uninteresting variables**

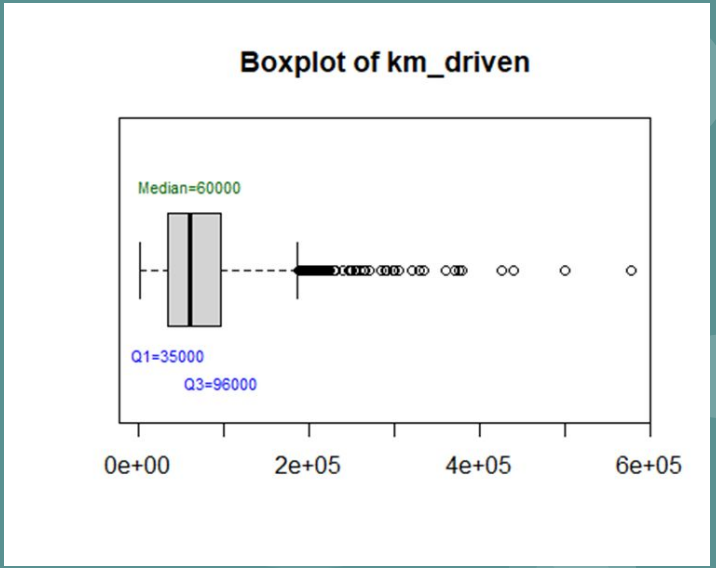


This is with state codes instead of names, and with joining the rest into “Others”

km_driven

How realistic is a value close to a world record?

More importantly, how much does it break the analysis?

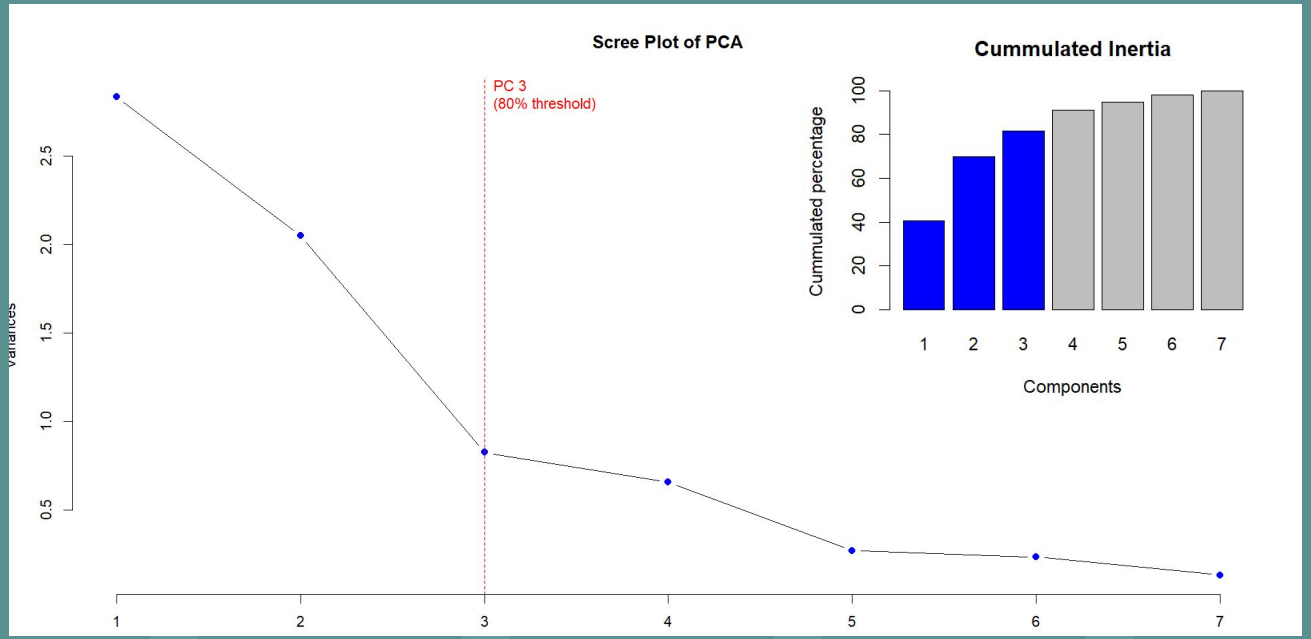


3. Analysis, PCA, clustering

PCA specifications, scree plot

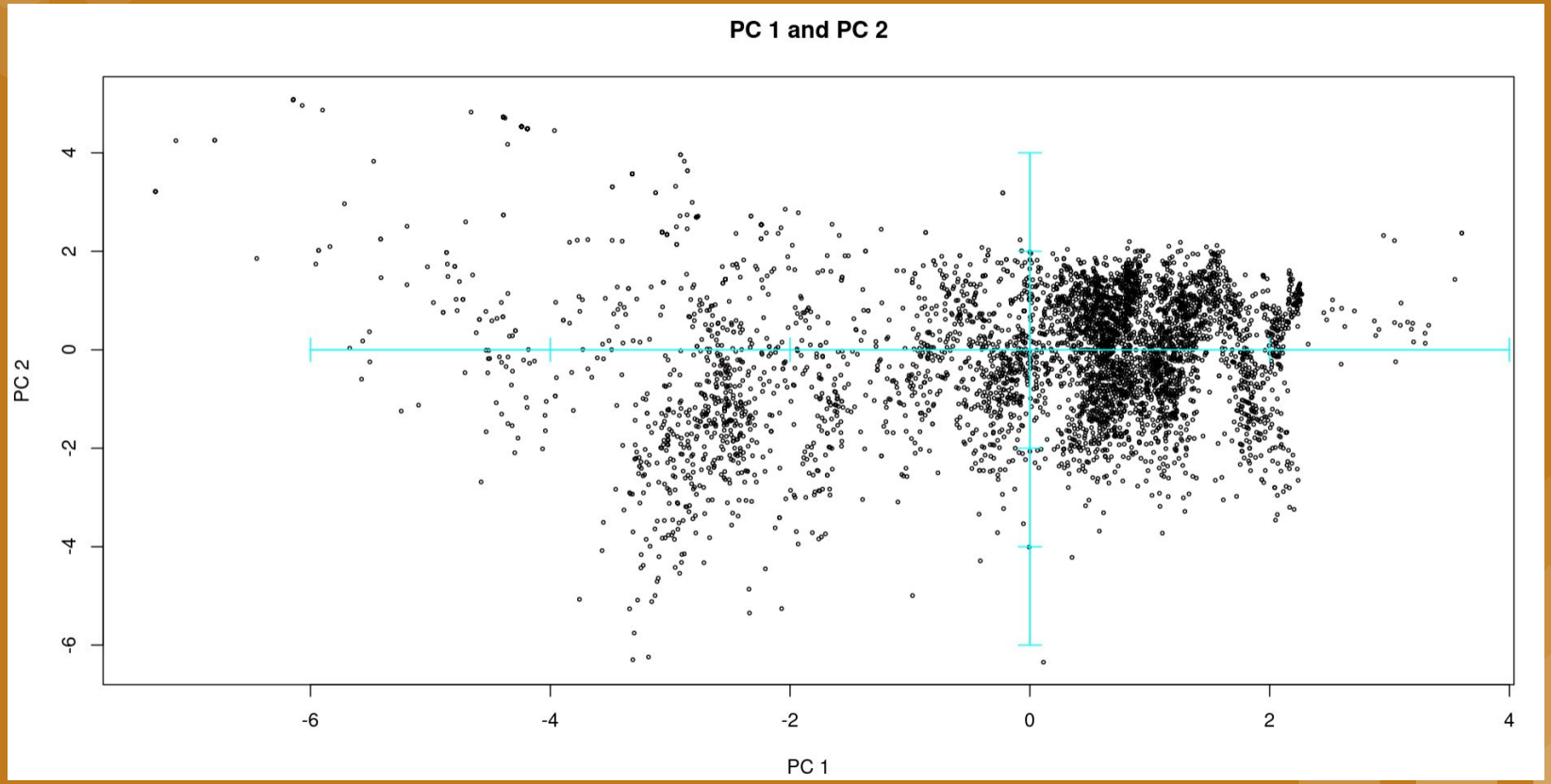
Specifications

- 3 principal components
- 81.5% of total variance
- Dimensionality reduction
- Retaining meaningful information

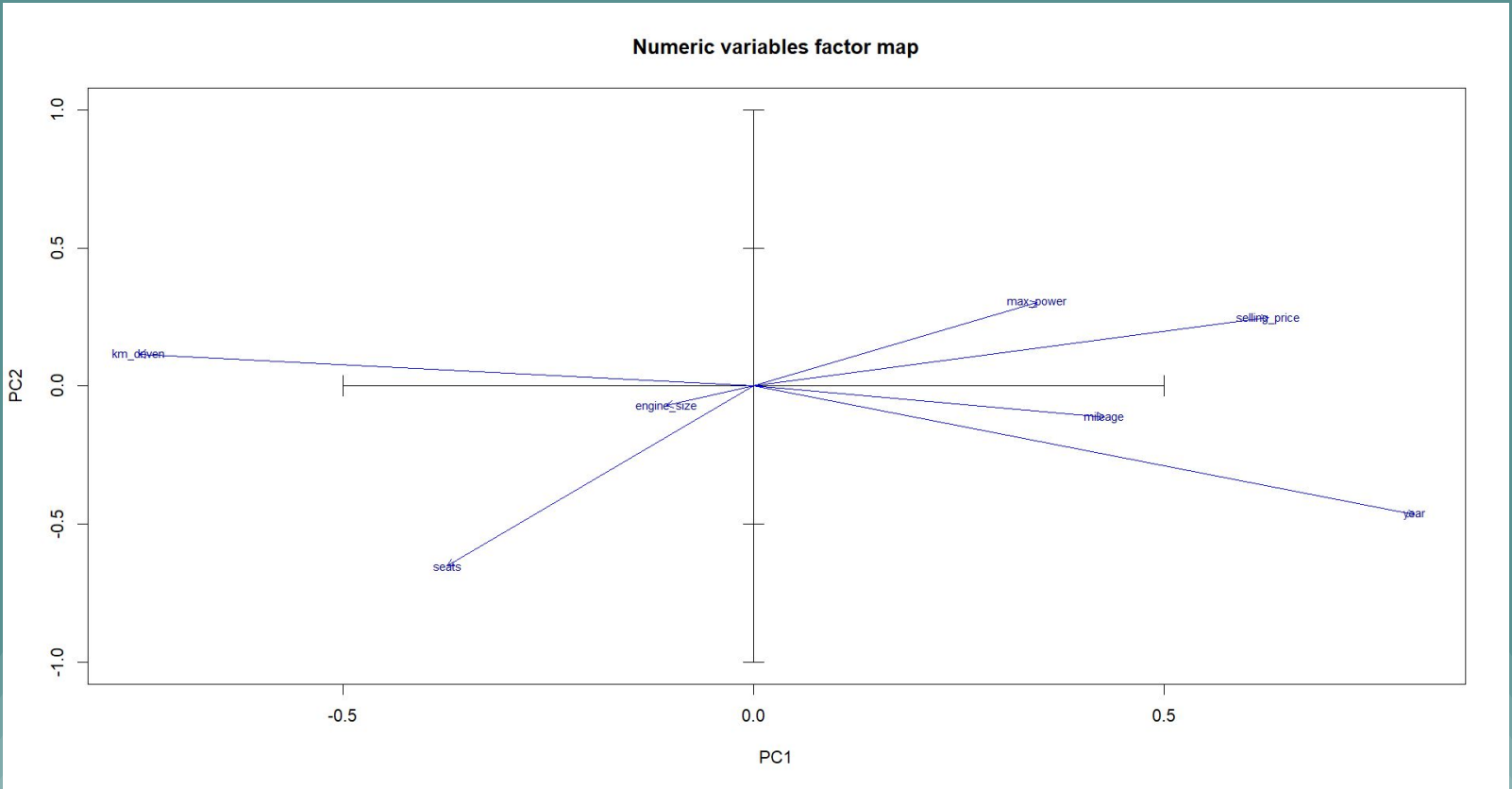


X

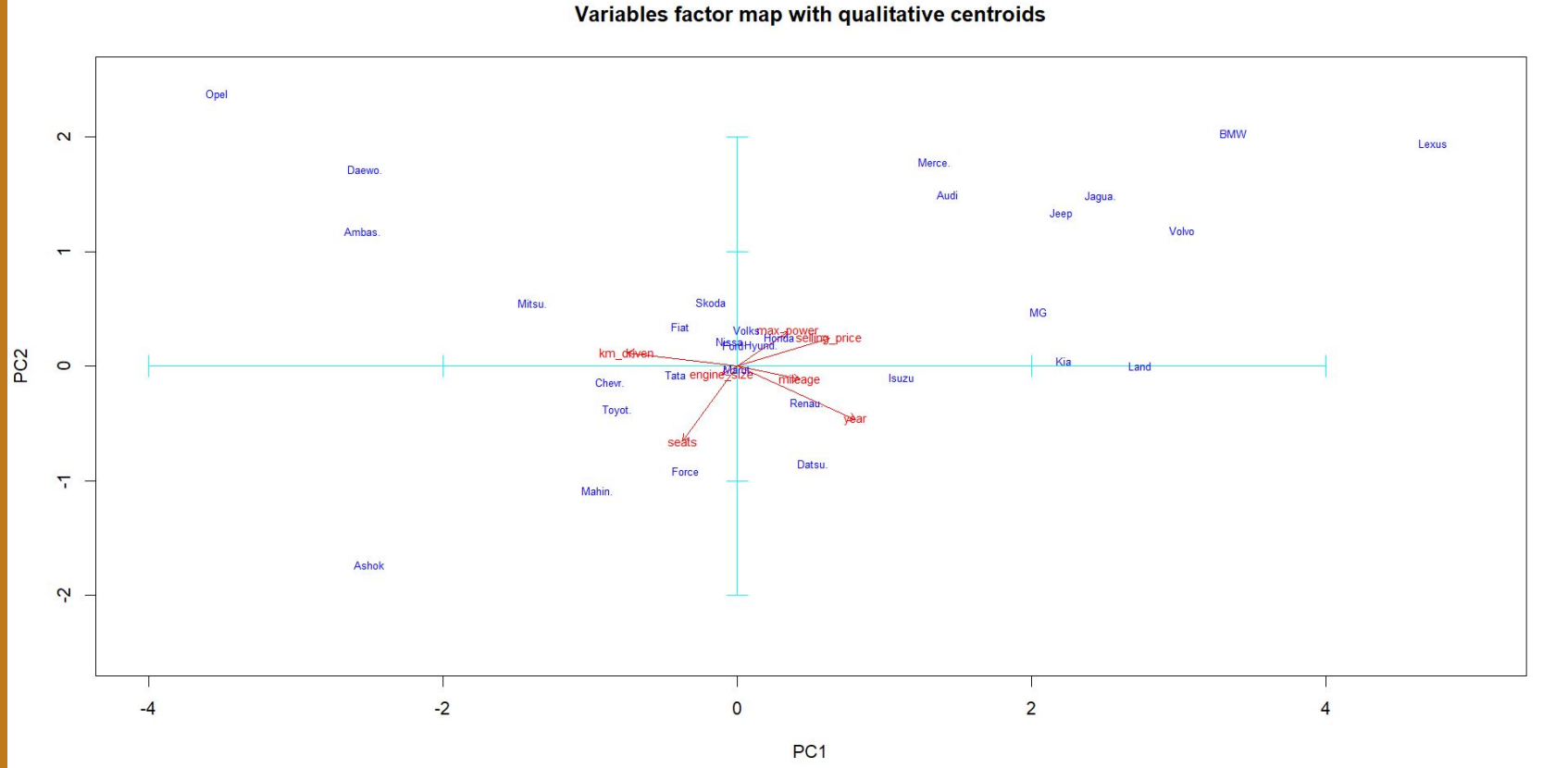
Individual projection for PCA



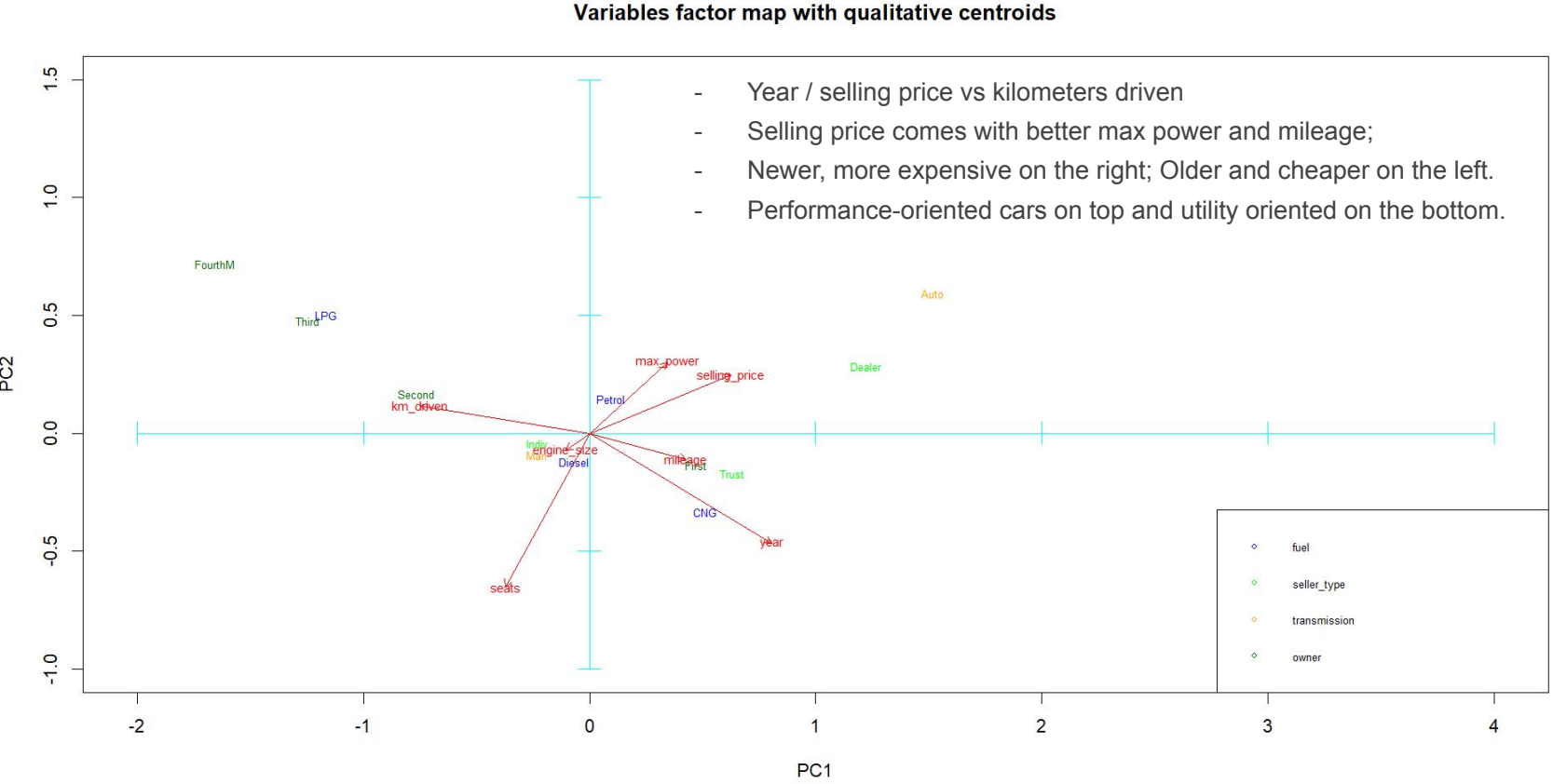
First factorial plane for PCA



Key PCA conclusions

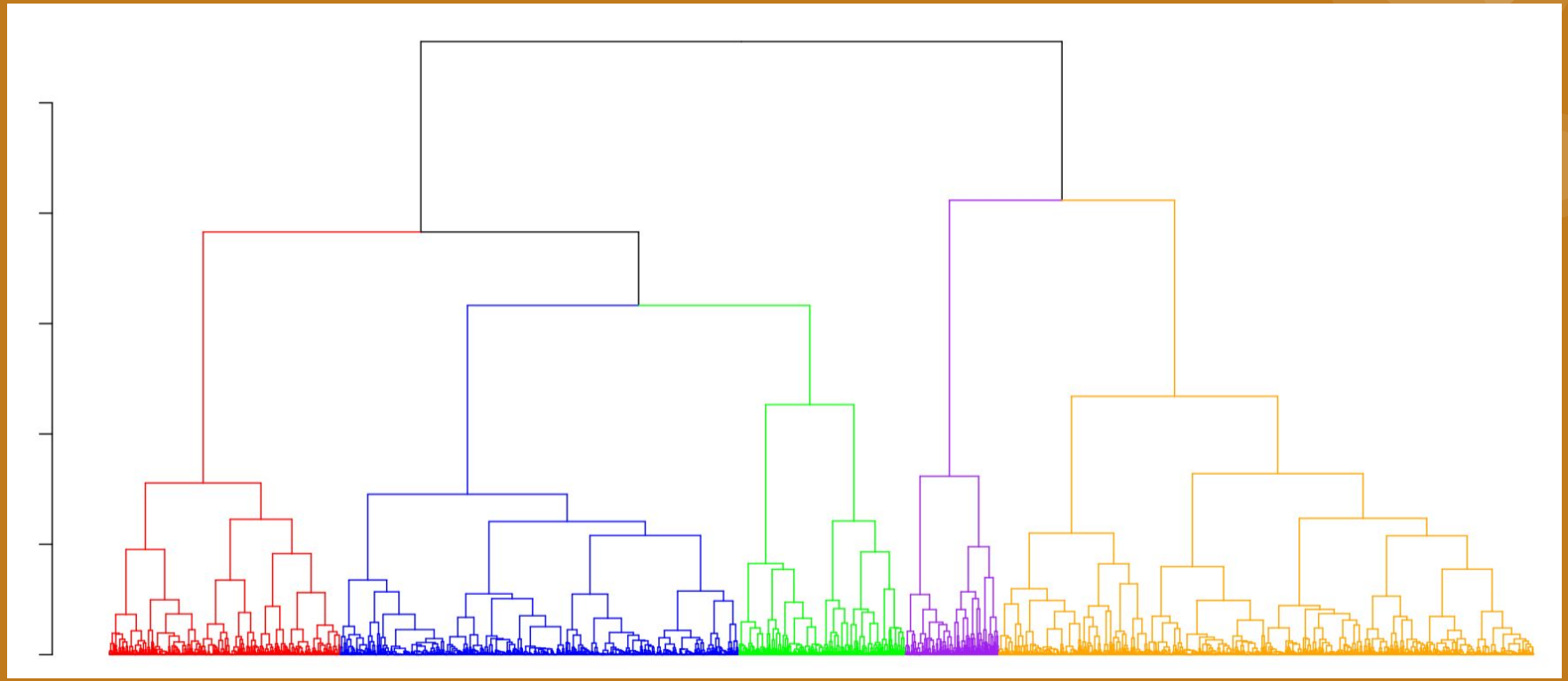


Key PCA conclusions



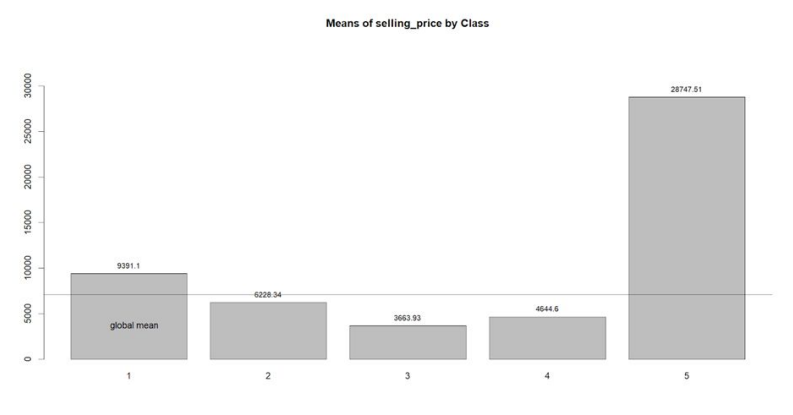
- Year / selling price vs kilometers driven
- Selling price comes with better max power and mileage;
- Newer, more expensive on the right; Older and cheaper on the left.
- Performance-oriented cars on top and utility oriented on the bottom.

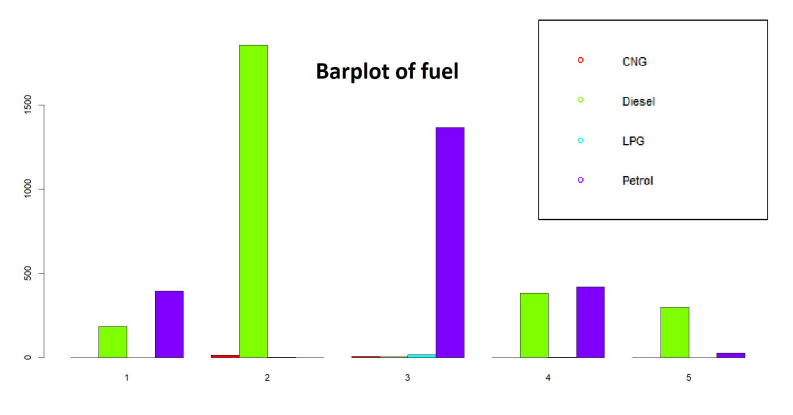
Clustering process



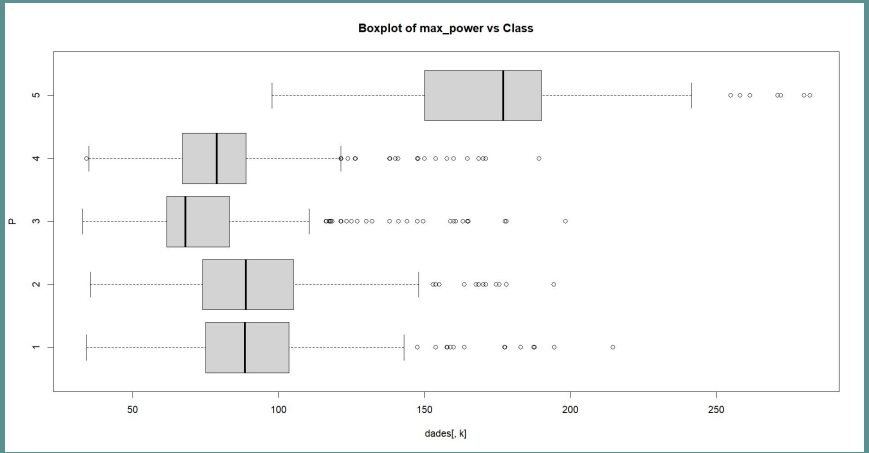
Class interpretation tools

Barplots





Boxplots



4. Profiling and interpretations

Profiling table

	1	2	3	4	5
Engine Size	Average	/	Small	/	Big
Fuel	/	Diesel	Petrol	Half petrol/ half diesel	/
Kilometers driven	Low	Largest	/	Average	Low
Max power	/	/	Lowest	/	Largest
Seats	/	Most	/	/	/
Mileage	/	Largest range	/	/	Lowest
Owner	Only first	/	/	/	/
Selling price	Expensive	/	Cheap	/	Overly expensive
Selling type	Dealer	Individual	Individual	Individual	/
Transmission	/	Manual	Manual	Manual	Automatic
Sold	/	No	No	Yes	/
Year	New	/	Old	/	New

Final class profiling

- Cluster 1 : More expensive (excluding supercars), newer and nicer for upper middle class.
- Cluster 2 : Utility-focused cars with loads of kilometers driven and generally more seats. These cars have seen multiple family vacations.
- Cluster 3 : Cheapest cars on the market, though in too poor of a condition to sell well (low power, small engine, manual transmission).
- Cluster 4 : Average in every way, which makes them easier to sell.
- Cluster 5 : Supercars, bought and sold by the very rich

PCA vs clustering

Consistent information

- PCA shows correlation between price and max power → Cluster 1/5
- PCA shows clear distinction between older, cheaper cars and newer, more expensive cars → cluster 3 vs cluster 1/5
- PCA shows clear separation between utility and power → Cluster 2/3 vs 1/5

Contradicting / non-consistent information

- PCA had cheaper, older car ←→ More kilometers driven, but not really represented in any cluster (cluster 3 is not that much cheaper / older than others)
- PCA had more seats ←→ Lower max power, but not represented in any clusters

5. Final conclusions

Final conclusions

- **Observed trends:** Most sales come from individual sellers and manual transmission cars.
- **Variable relationships:** Inverse correlation between year/price and kilometers driven
  - A seller should not give high prices to old cars with a lot of kms
- **Market segmentation (Clustering)**: 5 distinct car groups identified:
  - From compact budget cars to luxury vehicles.
  - One seller strategy could be to get a position in one of these groups

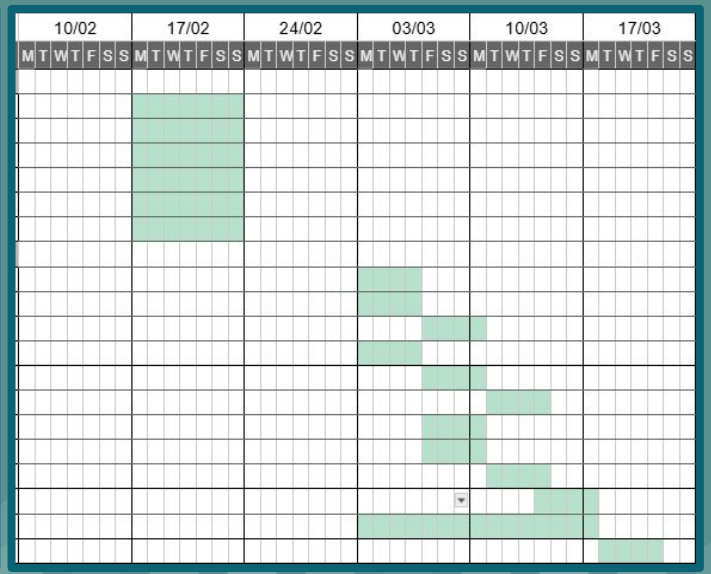
Return on objectives

- Understanding pricing factors (yes - PCA)
- Predicting sales likelihood (yes - descriptive)
- Consumer and seller behaviour (yes - profiling)
- Regional market variations (no)

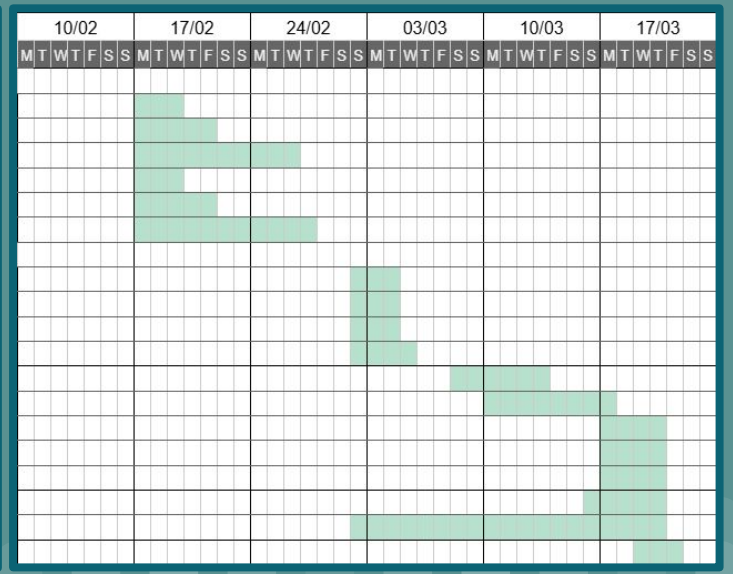
Scheduling

Initial Gantt Diagram

TASK
D3
Initial working plan(Gantt, Task division, Risk Plan)
Metadata file
Basic initial univariate descriptive
Enumeratation of steps of the preprocessing
Justification of decisions for preprocessing
Additional descriptive statistics of variables
D4
Motivation of the work and general description
Cover + index + Data Source presentation
Formal description of Data structure and metadata
Complete Data Mining process performed
Detailed description of Preprocessing and data preparation
Basic statistical descriptive analysis
PCA analysis for numerical variables
Hierarchical Clustering on original data
Profiling of clusters
Global discussion and general conclusions
Working plan
Making the ppt



Final Gantt Diagram



Questions

FIB

UPC