Data Mining On Second-hand Cars internet sales:

Manuel Esteban ARIAS

1. Introduction:

The Vehicles database, containing a register of 426 800 cars offered to be sold across the 50 statutes of the USA through e-commerce webpages. As such, the dataset contains information about the car features such as the brand, the model, the cylinders, the status, the color, the year of release, the mileage and, of course, the price. Aswell, the database presents context information such as where is the car offer taking place (the state and the region of the current owner) , the coordinates of this place, the website the announcement was published, the link of the announcement and the main description included in the annonce.

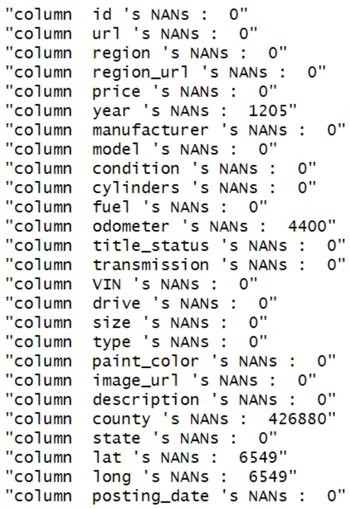
This database is a result of a titanic work scrapping from several websites. This has created a set of information that presents several gaps when the announcements on the internet are incomplete or wrongly filled. As a result of this, the data contains several outliers that, given their nature (multiples of 10 , or mono digits) are suspected to be just an expression of missing data. This data will be cleaned as it is presented below.

Finally, we would like to mention that some of the information contained in the dataset, while precious, can not be processed with data mining algorithms but may need other procedures such as NLP that escapes from the scope of this project. As so, as it will be mentioned later, text variables as description wont be considered.

2. Cleaning the data:

The original inputs contain 426 880 observations described in 26 features.

a. Non available values analysis:

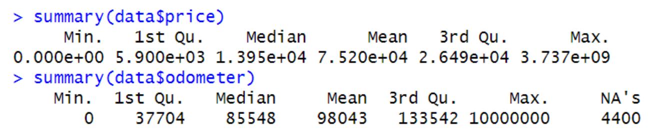


```
"column  id 's NANs :  0"
"column  url 's NANs :  0"
"column  region 's NANs :  0"
"column  region_url 's NANs :  0"
"column  price 's NANs :  0"
"column  year 's NANs :  1205"
"column  manufacturer 's NANs :  0"
"column  model 's NANs :  0"
"column  condition 's NANs :  0"
"column  cylinders 's NANs :  0"
"column  fuel 's NANs :  0"
"column  odometer 's NANs :  4400"
"column  title_status 's NANs :  0"
"column  transmission 's NANs :  0"
"column  VIN 's NANs :  0"
"column  drive 's NANs :  0"
"column  size 's NANs :  0"
"column  type 's NANs :  0"
"column  paint_color 's NANs :  0"
"column  image_url 's NANs :  0"
"column  description 's NANs :  0"
"column  county 's NANs :  426880"
"column  state 's NANs :  0"
"column  lat 's NANs :  6549"
"column  long 's NANs :  6549"
"column  posting_date 's NANs :  0"
```

While extensive, the database does not present complete information of all the observations. We can see, in the count of NANs in the left side of the page, a notorious gap of information in the columns Year, Odometer, County, latitude and longitude. It becomes clear the variable County (completely empty), should be removed.In the same way, the columns posting_date, description, image-url, VIN region_url and url will display unique variables for each observation and, as so they will not contribute any information to the analysis, and, therefore, shall be removed.As well, given that the goal of this study is not to perform spatial analysis, the coordinates (lat and log) may not be necessary, so they can be removed as well. Once removed, it was noticed that observations with NAN within these 3 columns were not related to the absence data in year and odometer, which held the same amount of missing data. This can be seen verifying the distributions

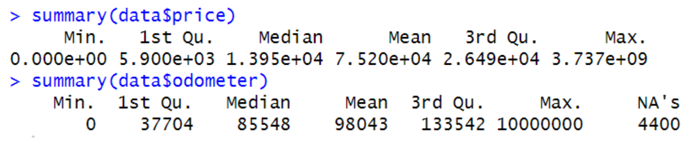
If the the rest of the NAN values are omitted only 3 % of the data is lost, retaining 414 863 rows. When this 5605 Rows are observed, it can be seen that most of their categorical variables are As so, holding only complete observations we may see that the distribution of the variables would take the following measures:

<table>
<tr><td>426 880 observations described in 26 features</td><td>426 880 observations described in 19 features(suppressing Lat, Long, County)</td></tr>
<tr><td>



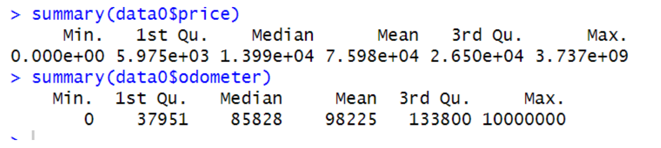
```
> summary(data$price)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.000e+00 5.900e+03 1.395e+04 7.520e+04 2.649e+04 3.737e+09
> summary(data$odometer)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
       0    37704    85548    98043   133542 10000000     4400
```

</td><td>



```
> summary(data$price)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.000e+00 5.900e+03 1.395e+04 7.520e+04 2.649e+04 3.737e+09
> summary(data$odometer)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
       0    37704    85548    98043   133542 10000000     4400
```

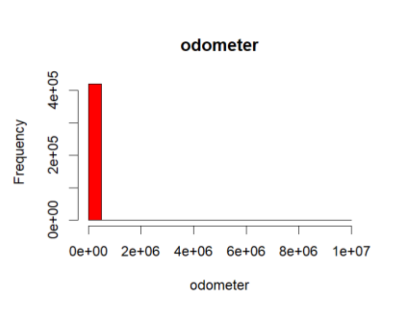
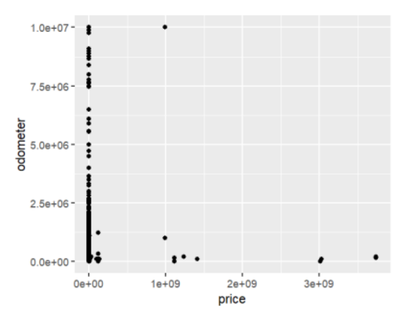
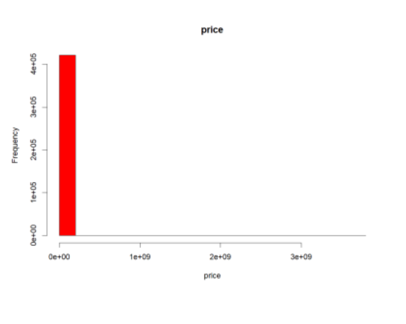
</td></tr>
<tr><td>421 344 observations described in 16 features, no NANs</td><td></td></tr>
<tr><td>



```
> summary(data0$price)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.000e+00 5.975e+03 1.399e+04 7.598e+04 2.650e+04 3.737e+09
> summary(data0$odometer)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
       0    37951    85828    98225   133800 10000000
```

</td><td></td></tr>
</table>

The remaining data contains several categorical features about the car that we cannot impute artificially with any precision and that, therefore will be set as “unknown” instead. For the Numerical ones, the distribution suggests a very important frequency of values of odometer and price set as 0. As see below

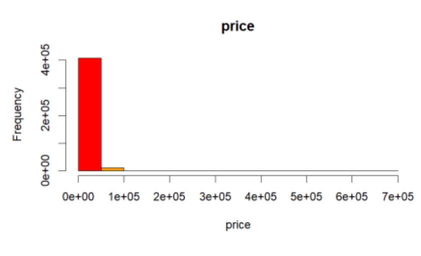
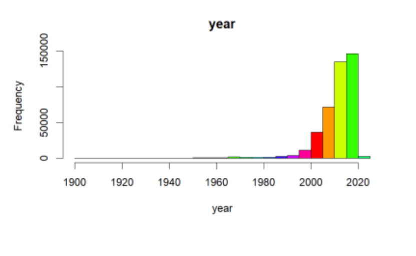
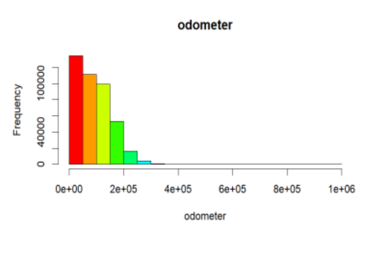
		
Rows when odometer == 0 1943 (0,4% of the data)		Rows when prices = =0 30759 (7% of data)

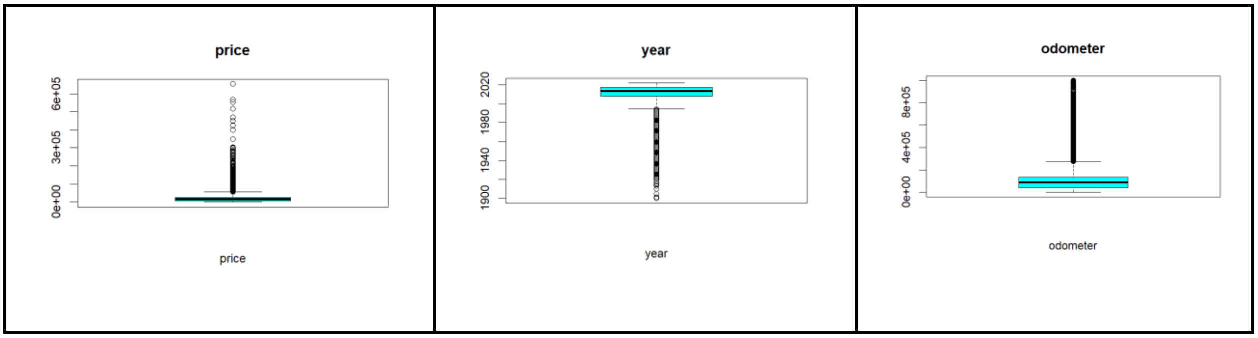
As so, it is proposed to remove the values superior to USD 750 000 for the price, for considering these values excessively high for a car of “standards” brands such as Chevrolet ford or jeep, that often have prices close to the average. In the odometer were also identified inconsistencies in the extent in which the values like 0 and 1 are unlikely to be real data, nevertheless they are still possible. In contrast, several rows (639) were identified having values 999999 and 1000000, which is suspicious, and may suggest that the formulary filler just inserted a big number given the ignorance of the real data. As such, they were removed.

Overall, 98,48% of the observations were kept, resulting in a database with 420 414 observations, 16 variables (3 numerics, 13 categorical) . The empty values were replaced by “unknown”.

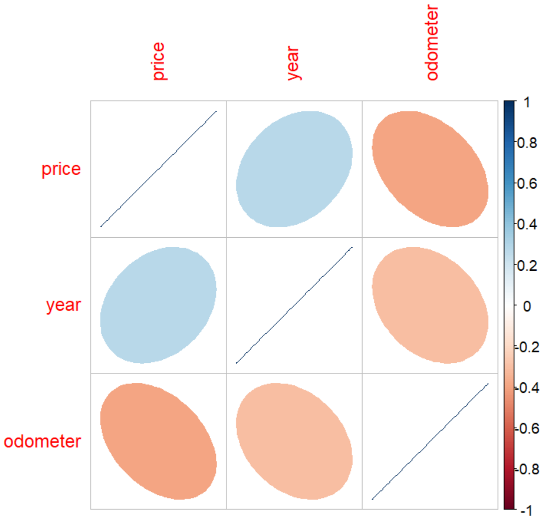
3. Exploratory analysis:

An exploratory analysis for the numeric variables can be done using histograms and boxplots. For the variable Price, it can be seen that most of the prices are concentrated in the first deciles of the distribution while an important amount of observations have prices that can be considered as atypical. This is to say that, while the mean is about USD 17 492, every price higher than USD 57 268 was considered as atypic. The variable year presentes a flatter distribution that is concentrated in the last 30 years. Even when the inventory included old cars when they were selling the year before 1994, suggested as outliers, the mean held in 2011 and the median in 2013. Finally, the odometer presented a softer distribution that decreases in frequency as the value rises,and the mean was 92 187 mi and the median of 85 578 mi. In this case, the outliers can be considered after 276 312 mi.

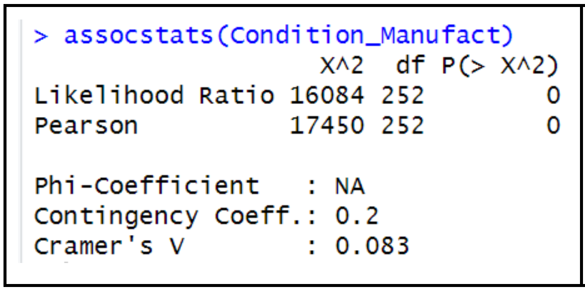
		



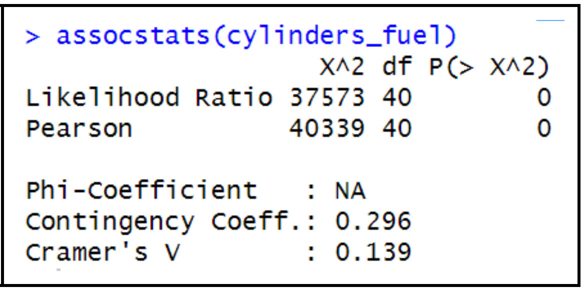
Finally, the correlation between them suggests that the price and the year are positively correlated (0,2753), which makes sense when considering that the time pass decreases the value of the vehicles. As well, it suggests a negative correlation of odometer with prices and year (- 0,3904 and -0,3028 respectively) suggesting the price of the car decreased when the car has a higher mileage, which is sensible. However, the negative correlation with the year is suspicious given that age implies more use time and as well more mileage, which is counterintuitive. This unexpected result may be suggesting that the data is not accurate and as such that the owners of old cars are not reporting the date as they should. This variable should not be considered in the analysis.



Regarding the qualitative variables, while all of them may give very important information, We would like to analyze two relationships in particular: In the first one, the ones between condition and Manufacturer. We would expect that the most expensive cars (those of the best brands)are better entertaining, and as such are in better condition than those of mass production brands. However it is interesting to see that indeed, they are not very correlated as the Cramer's V of 0,083 suggest. The second one, the relationship between cylinders and fuel kind, pretended to test the notion that a highest cylindrage necessarily implied the preference for certain kinds of fuels (expected for fossil, for example). In this case a Cramer's V of 0,13 suggests a relationship that, even when not high, is enough for starting a relationship between these variables.



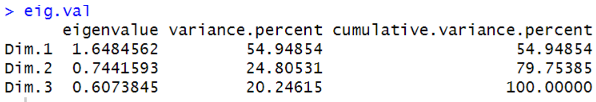
```
> assocstats(Condition_Manufact)
                    X^2  df P(> X^2)
Likelihood Ratio 16084 252        0
Pearson          17450 252        0

Phi-Coefficient   : NA 
Contingency Coeff.: 0.2 
Cramer's V        : 0.083 
```



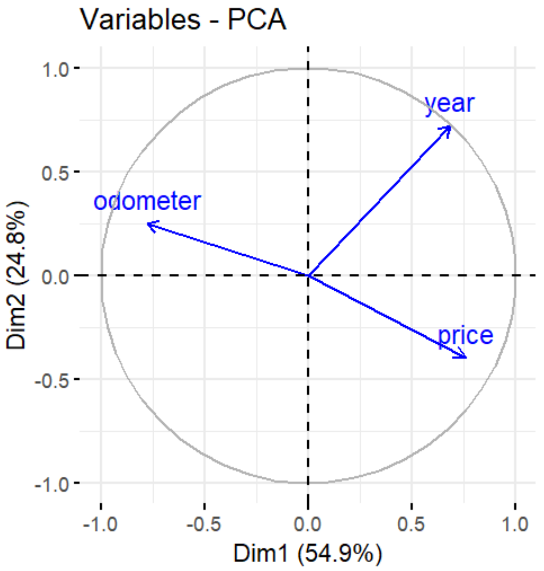
```
> assocstats(cylinders_fuel)
                   X^2 df P(> X^2)
Likelihood Ratio 37573 40        0
Pearson          40339 40        0

Phi-Coefficient   : NA 
Contingency Coeff.: 0.296 
Cramer's V        : 0.139 
```

4. Unsupervised analysis:

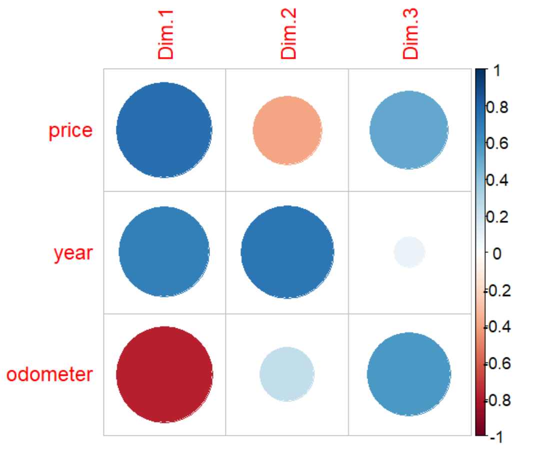
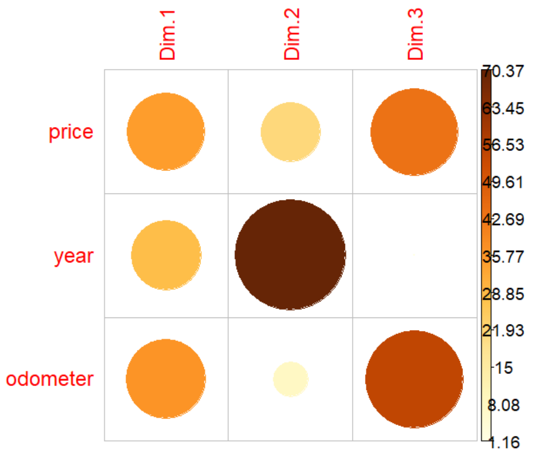
In order to get deep into the data, it is processed to perform two unsupervised analyses. The first one, a **Principal Component Analysis** (PCA) will permit the compilation of the data from the numerical variables and confirm or deny the trends identified in the exploratory analysis. As seen below, this permits to trade 3 dimensions out of three variables in which the half of the explained variance is contained in the first one.



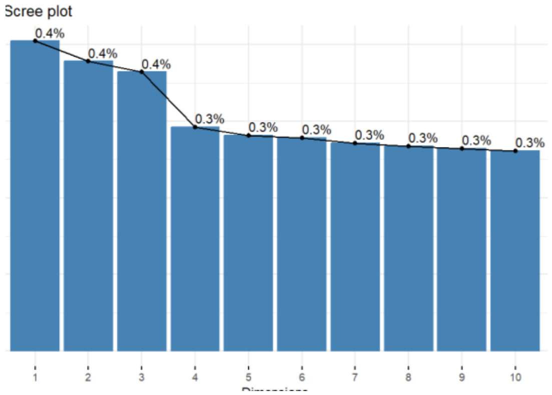
```
> eig.val
      eigenvalue variance.percent cumulative.variance.percent
Dim.1  1.6484562         54.94854                    54.94854
Dim.2  0.7441593         24.80531                    79.75385
Dim.3  0.6073845         20.24615                   100.00000
```



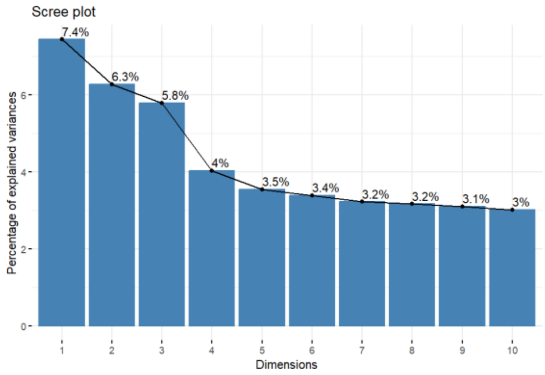
When the dimensions are explored, it can be seen that the dimension 1, while strongly correlated with all the three variables, is strongly representing the variables odometer and price. However these two, are going in opposite senses presenting a negative correlation between them, as was originally expected. In the case of the year, this variable is strongly correlated with the dimensions 2, and presents the highest contribution as well. Nevertheless, its perpendicularity with the vectors of Price and odometer suggest an absence of correlation between them. This suggests that the price and the mileage may be independent from the year of the car's selling date. This, in line with what was found for the exploratory analysis.

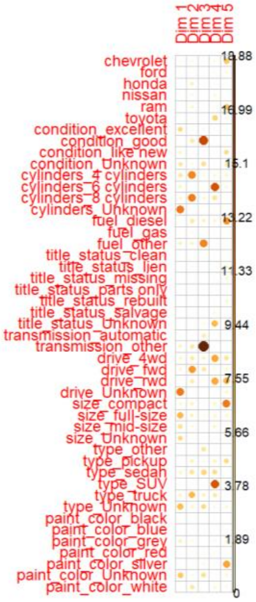
Correlation with dimensions	Contribution to dimensions
	

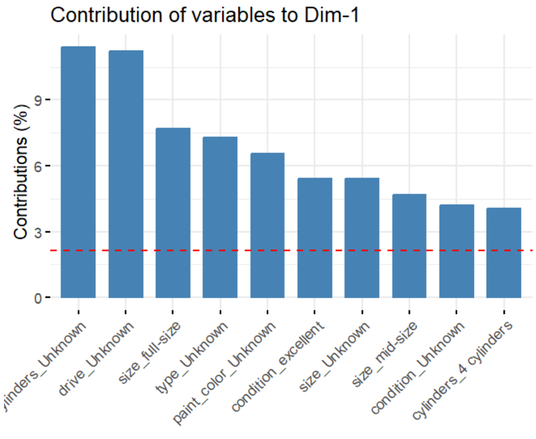
The second one, proposed for the categorical variables, is a **Multiple correspondence analysis (MCA).** While useful, the MCA requires enormous memory sizes and, as so, the analysis was run with a simple of 1000 individuals. Said this, 10 dimensions were estimated with dimension 1,2,and 4 explaining around 0.4% of the explained variance and with the rest of them starting from 0,3% of the variance explained. As it can be seen, the mca is not creating descriptive dimensions, and at the same time that the correlations between the variables are weak.



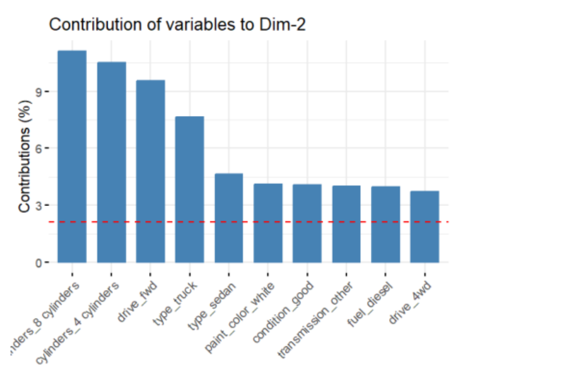
Seeking to discover the correlations, it is proposed to deploy an MCA exclusively for the car features and remove the contextual ones (Region and State). As well, it is proposed to omit the variable Model, whose internal wide variety may be creating a noise that does not contribute to explain the data. The new calculations show that the first dimension explains 7,4% of the inertia, while the second and second and first present 6,3% and 5,8% of it. The distribution may suggest that the first 4 components (4% of the variance in the 4th) may be the most useful for understanding the relationships between the variables.

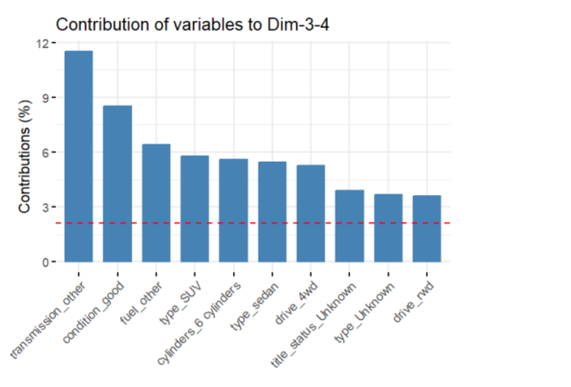




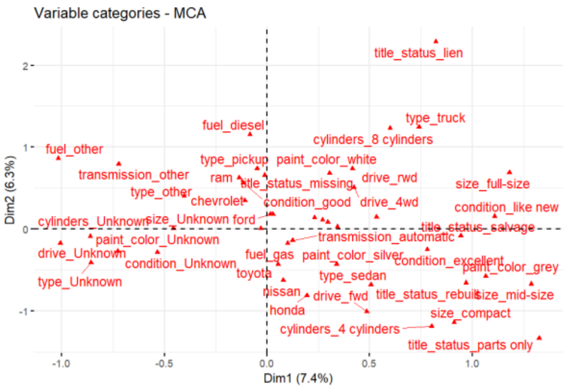


When the contributions to each dimension are analyzed, the Unknown variables take the lead for the first dimensions given their important amount. As so, this first dimension represents mainly unknown data in each of the variables. In the second dimension, in contrast, the cars of 4 and 8 cylinders start to gain protagonist, as well as the cart with drive mode “fwd” and the truck type, what may suggest a plausible correlation given that the truck often need the higher cylindrage possible (8 cylinders). In the third dimension the condition “good”, the alternative fuels and transmission (“other”) present the biggest contribution. In this sense, this dimension may be representing the cars that have uncommon features but, because of their variety, represent a non despicable share of the market. Finally, in the fourth dimension, the 6 cylinders, the drive modes 2wd and rwd, and the SUV type were the most representative as well as the Unknown statustype”.For this case, the correlation is not clear.   In a few words: Dim 1 describes the variance of unknown data, Dim 2 talks mainly about the variance of trucks and Dims 3 and 4 about variance of non conventional cars.Nevertheless one conclusion can be drawn: The cylinders and the drive mode play a major role explaining the variance among the cars.

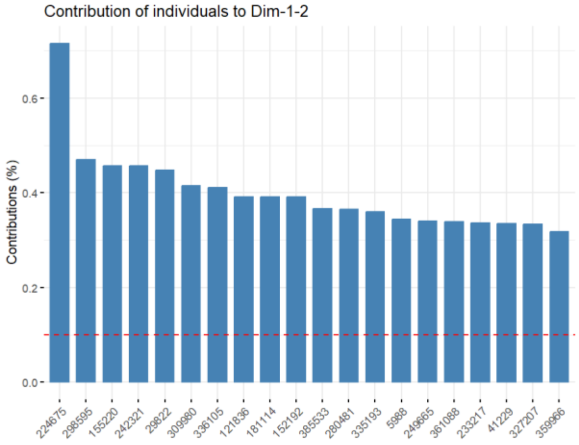




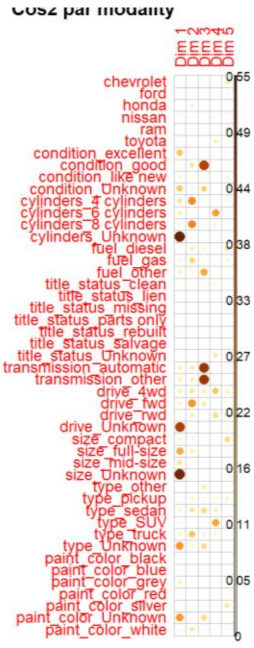
In a bi-dimensional space, it can be seen that , as expected, most of the unknown categories are allocated along the first dimension axis (horizontal one). However, plotting the features in this way permits to see that the mass-production manufacturer classes are mostly situated close to the origins, signaling impartiality across the dimensions and implying as that a wide variety of car types offer by the same brand, as toyota, nissan, ford, or chevrolet.

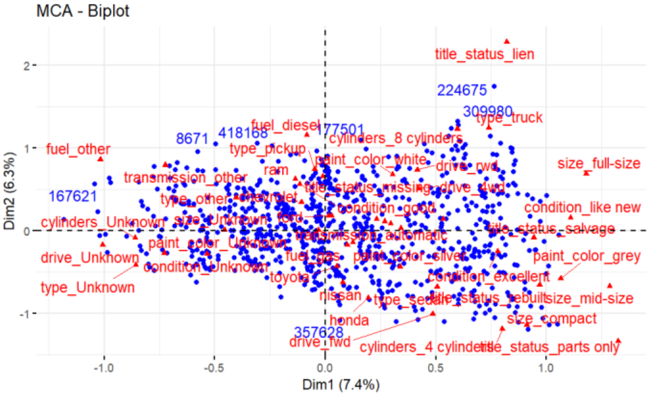


If the individuals are checked in detail, it can be seen that the individual that participates the most in the dimensions 1 and 2 is a car sold in Minnesota with a price of USD 32 995. This Ford f-150, as Truck as expected, was released in 2014, was in an excellent condition, with 8 cylinders and gas fuels. The second with the biggest contribution was sold in Toledo, Ohio, being as well a Ford, this time an edge model, gas fuel, automatic transmission and SUV type. As such, we can see that the dimension represented efficiently the individual observations in the database.



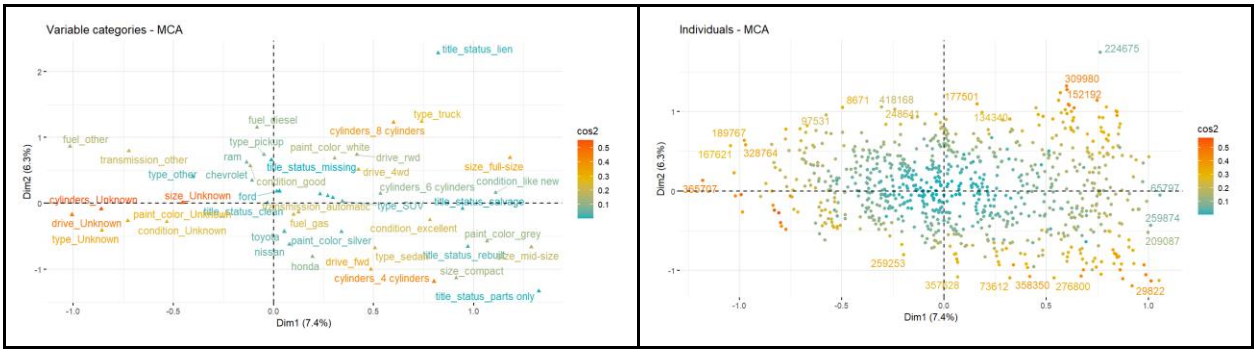
The biplot, a tool for visualizing the modalities and the individual observations, results messy in a case with so many modalities like this one. Nevertheless it is interesting to see that the colors are concentrated in the first and second quadrants of the plane, suggesting a positive correlation with the first dimensions (and a negative with the unknown colors). We can see so farr see that while colors like gray are not popular in the US, the blue, black and red are not only more frequent but are offered for cars with similar features. This starts to suggest that the “quality” of the cars are often associated with specific colors.





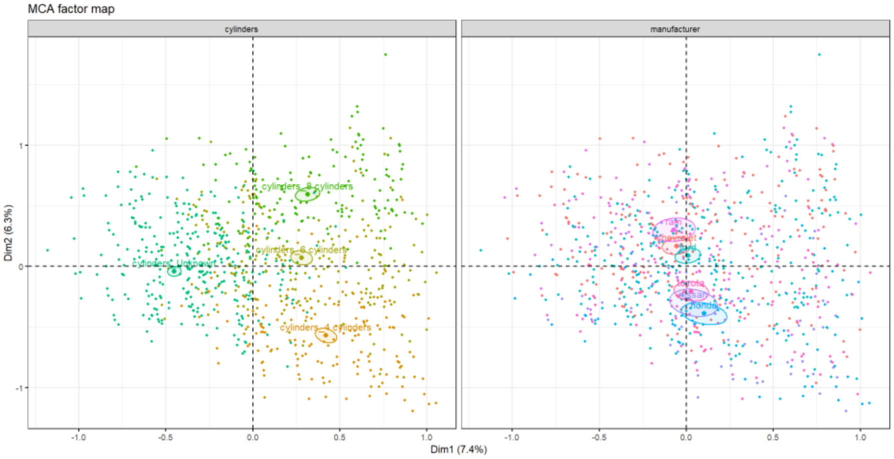
Additional information for the analysis is provided by the Cos2, which illustrate the degree of “quality” of the data representation, this is to say, not just “how much of their variance is explained” but in which degree are the variables associated to each dimension.
When examining this factor, it can be seen that the unknown variables are definitively associated in the first dimension and as such, that the first dimension explains the variance and the “unknown” modality occurrence. The Dimension 2, does not present big changes, suggesting as well that high quality in the representation of the mentioned categories: 4 and 8 cylinders, fwd drive mode, and truck type. In dimension 3, while good condition and other transmission remain as well variance-explained and well quality-represented, the automatic transmission won relevance as the best quality-represented. Finally, the 4th dimension held the same pattern by exposing a good representation and variance explanation for Suv type and 6 cylinders.

The information provided by the Cos2 can be solved applied to the different categories. We can see, for example, that the Unknown modalities are the best associated with the dimensions 1 and 2 summing more than 0,5. In contrast, while contribution a lot to explain the variance, the car brands as Ford, Toyota, Nissan, or Honda, presente low levels of association with the dimensions, what suggest that while the brands influence technical features, like clylindrage or the drive mode, they don't for the kind of car. This is to say, being a Ford or a Honda, does not influence the car to be a truck or a SUV, for example.
If we verify the individuals it will be seen that often, the more affiliated they are with both first and second dimensions, the less effectively associated they are with the dimensions. This means that while the dimensions explain changes between the cars, they are not efficient for predicting their features. This trade off between associability and variance explanation seems to be presente in all the databases.



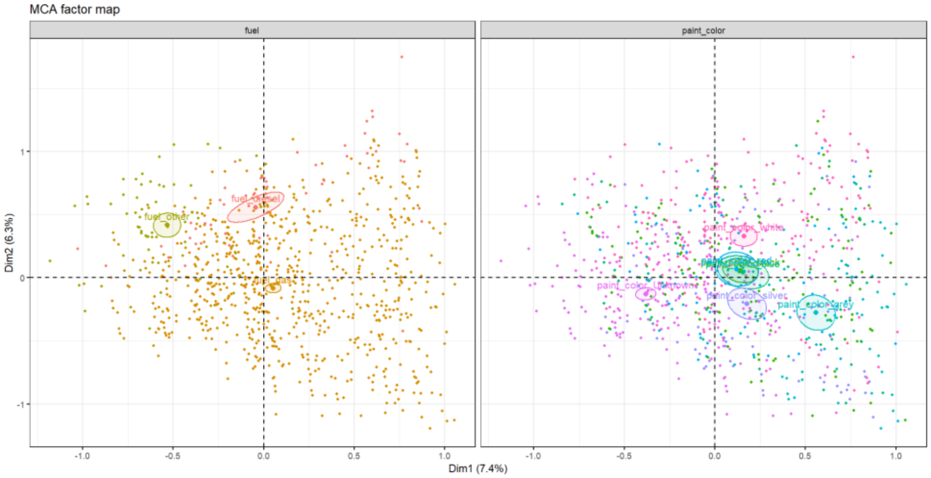
When mapping the variables, it can be seen that the cylinders, for example, can be grouped by "high" and "low clylindrage. We can see how the individuals with 8 and 6 cylinders are allocated in the positive range of the second dimension and how 4 cylinders are mostly in the lower dimension . To this extent, this dimension is useful for predicting this variable.

The map shows that manufacturers, indeed, are mostly explained through the second dimension. Within it, it is possible to see how they can be classified in two big clusters that define their way of variating in the features. In The first group, Ram, Chevorlet, and Ford seems to present similar variance dynamics (and as presumed they may be offering similar car types with equivalent features), in the other Toyota, Nissan and Honda, all of them Japones, seems to have simile variance dynamics as well, what whan be expected given similar design constraints originally thought for the japones industry and market they were subjected to.



When the fuel is analyzed, the symmetry observed for the other variables is broken. In this case the diesel and the gas, while related to the same dimension (the second one) , go in different senses, as competitors. The other fuel, in contrast, is more represented, even when not strongly, in the first dimension. As an additional remark, It is interesting to see that most of the cars sold in the country work with gas fuel and increasingly with other fuels.

In the case of the color of the cars, as signaled before, the color centroids converge close to the origin of the plane. They seem to be equally represented in both dimensions 1 and 2(closer to this last one). The most popular colors, like white, silver, red, explain their variance similarly across the dimensions, suggesting that they are presented in the same kind of cars and even in the same kind of brands.In contrast, the unknown, one again are represented isolated from the rest and closer to the dimension 1.



5. Supervised analysis:

The code we used performs an **OLS** (ordinary least squares) regression on a mixed dataset containing both numerical and categorical data. OLS regression is a common linear regression technique used to model the relationship between a dependent variable and one or more independent variables. It works by finding the line of best fit through the data that minimizes the sum of the squared residuals between the predicted values and the actual values.
In our case, the lm function is used to perform OLS regression on the training data. The target variable is regressed on all the predictor variables using the formula target_variable ~ ., which specifies that all the predictor variables should be included in the model. For the predictor variable we used, based on prior analyses on which variables are more discriminating: *"year", "manufacturer", "condition", "type", "odometer"*.

The resulting model is then used to make predictions on the testing data (30% of a random sample of the total of the data) using the predict function.

To evaluate the accuracy of the predictions, the code uses the root mean squared error (RMSE) metric from the caret package. RMSE is a common metric used to measure the difference between predicted and actual values. It works by taking the square root of the mean of the squared differences between the predicted values and the actual values. By using RMSE, we can measure the accuracy of our predictions in the same units as the target variable. A smaller RMSE indicates better predictive accuracy.

Multiple R-squared	Adjusted R-squared	p-value
0.3185	0.3183	< 2.2e-16

Some potential interruptions we encountered based on the summary and RMSE values:

- The summary output shows that some of the predictor variables are not statistically significant (i.e. have p-values above 0.05) for example : (*manufacturerlincoln, manufacturermorgan, typemini-van*). In many cases, including additional predictor variables can improve the accuracy of the model, even if they are not individually significant. Additionally, the significance of a variable can depend on the other variables in the model, so a variable that appears insignificant on its own may become significant when combined with other variables.
- The RMSE value is slightly large, the model may be performing poorly. A large RMSE indicates that the model is not accurately predicting the target variable. This could be due to a variety of factors, such as a poor choice of predictor variables, the presence of outliers in the data, or the use of an inappropriate regression technique. Some ways we suggest to improve the model, such as exploring different combinations of predictor variables, removing outliers from the data, or trying a different regression technique.
- Multiple R-squared: This value measures the proportion of variance in the target variable that is explained by the predictor variables in the model. In this case, the multiple R-squared is 0.3185, which means that the predictor variables explain about 31.85% of the variance in the target variable.
- Adjusted R-squared: This value is similar to the multiple R-squared, but takes into account the number of predictor variables in the model. It adjusts the multiple R-squared to penalize the inclusion of unnecessary variables that do not improve the model's predictive power. In this case, the adjusted R-squared is 0.3183, which is very close to the multiple R-squared, indicating that the inclusion of additional variables did not significantly impact the model's predictive power.
- p-value: This value indicates the probability of observing an F-statistic as large as the one computed by chance, assuming that the null hypothesis (that the model has no predictive power) is true. In this case, the p-value is less than 2.2e-16, which is very small, indicating that the model has significant predictive power and the null hypothesis can be rejected.

Overall, the output suggests that the model is a good fit for the data and has significant predictive power. However, the residual standard error suggests that there is still room for improvement in the model, and additional analysis may be needed to identify ways to improve the accuracy of the predictions, this may explain why the RMSE value is marginally high.

6. Conclusion

The different analysis exposed in this article permitted us to identify clear trends in the American car market in both the commercial and industrial side. From the commercial, the exploratory analysis shows an average price of USD 17 492 and a median of USD 13 990, holding most of them in the first deciles of the distribution. This suggests a market full of “popular” cars with a few overpriced cars. As the multifactorial analysis was going to show, these cars often are branded as Ford, Nissan, Toyota,Honda and similar, which are the leaders in the American Market. These cars are often released in the last 30 years. However, it is hard to identify in which extent they are used or not. The variable odometer, while still correlated to the price and to the year, presents gaps and trends that may be counterintuitive and as so, it is presumed that the car offerers were often inserting wrong information in the field in the car selling websites. A PCA is permitted to identify as well that the odometer and price variable are inversely correlated (they are represented similarly in the first and second principal components), while the year of release seems not to be a factor that influences these variables.

For computation problems, the MCA was not performed for a variable that we considered as important: State, and this analysis may be done in further research, however, the car features. The findings start by the fact that the features are barely correlated, when the information is collapsed in dimensions, any of the resulting components was able to explain more than 7,5% of the variance. The unknown values took a main role in the first of them while the rest of the components were describing specific types of cars. The second one, for example, typified the trucks and the third one the “unconventional cars” (thos with “other” kinds of fuels and less common brands).

The brands, showed to present 2 may kind of dynamics: in one group the japanese brands as Nissa,Toyota and Honda, had similar elections for the fuel election, the cylindrage and the type while the second group, including Ford and Chavrolet, had its own choices in this same variables. This analysis permitted evidence that, within the cluster trends, the variables color, manufacturer and cylinders, for example, were mainly associated with the same dimension and as so are correlated. What is interesting to see is that this correlation presented to some extent a trade off in the variance explanation. It was noticed that in the cases in which the variables or the modalities were strongly represented in a dimension, their effective association with it (the cos2) was low. In summary, while the variables share variability, they are not strongly associated between them.

For the supervised learning part, we performed a linear regression analysis using Ordinary Least Squares (OLS) to predict a target variable based on a set of predictor variables. The data was split into training and testing sets, and we trained the model on the training set before evaluating its performance on the testing set. We used the Root Mean Squared Error (RMSE) as the evaluation metric, the model's performance was measured using the R-squared statistic, which indicated that the predictor variables explained about 31.85% of the variance in the target variable. The F-statistic was also highly significant, indicating that the model had significant predictive power. Overall, the analysis suggests that the OLS regression model is a useful tool for predicting the target variable based on the available predictor variables. However, further analysis may be needed to identify ways to improve the accuracy of the predictions, and to determine the potential impact of additional variables or factors that were not included in the analysis.