Università degli Studi di Milano
MSc in Data Science for Economics LM-DATA

Obesity: An exploration of the Influential Factors (Supervised)
Unveiling Tourist Personas (Unsupervised)

Mina Beric
11055A

ACADEMIC YEAR 2023-2024

Contents:

1. Supervised learning:

2. Unsupervised learning:

Supervised learning:

Obesity: An Exploration of the Influential Factors

1.1. Abstract:

Overweight and obesity, according to the World Health Organization (WHO), can be defined as the excessive accumulation of fat in different parts of the body[1], and is recognized as an important public health problem as it is related to various diseases, and even morbidity and mortality [2].
Early and accurate identification of obesity is crucial for developing personalized interventions and optimizing healthcare strategies. Although BMI (Body Mass Index) is the most widely used estimation of obesity, there are other factors that can contribute to gaining weight such as lifestyle factors.

This study presents an approach for **classifying individuals as obese or non-obese,** as well as finding the features that would be most relevant in training this model.

I leveraged a comprehensive dataset encompassing anthropometric measurements, lifestyle factors, and demographics of a group of individuals. Various supervised learning techniques, including logistic linear regression, Lasso, decision trees, Random forest and Adaptive boosting are evaluated for their effectiveness in obesity classification.

The aim is to assess the accuracy, sensitivity, specificity, and overall performance of the developed classifiers. The results aim to demonstrate the efficacy of a data-driven approach in achieving high accuracy for **binary obesity classification and** the findings highlights the crucial role that both physical activity and healthy eating habits play in tackling the obesity crisis.

[1] World Health Organization (WHO), Obesity and overweight, 2021. URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight.

[2] H. Rosen, "Is Obesity A Disease or A Behavior Abnormality? Did the AMA Get It Right?". Missouri medicine, 111(2014): 104–108.

1.2 The Dataset

The survey data is stored in a table with 2111 rows, where each row represents a unique respondent. The table has 17 columns, each corresponding to a specific survey question.
The respondents were originally 458 people from Mexico, Peru, and Colombia. Later, Fabio Mendoza Palechor and Alexis de la Hoz Manotas used an oversampling technique (SMOTE) to create a larger dataset, were 77% of the data points are fictitious [3].

The survey and the dataset were created and elaborated with the intent of applying supervised learning techniques, in order to predict the respondent's obesity level. For the purpose of my analysis, I will change the target variable Obesity **(Obesity_level,** chr) into a binary one (**ObeseYes, ObeseNo**).

The questions and the possible answers are reported below, all the variables were renamed from the original dataset to be more interpretable:

- What is your gender? (**Gender**, chr):
    - Male ("Male")
    - Female ("Female")

- What is your age? (**Age**, num)

- What is your height in metres? (**Height**, num)

- What is your weight in kilograms? (**Weight**, num)

- Has a family member suffered or suffers from overweight? (**family_history_with_overweight**, chr):
    - Yes ("yes")
    - No ("no")

- Do you eat high caloric food frequently? (**High_cal_food_freq**, chr):
    - Yes ("yes")
    - No ("no")

- Do you usually eat vegetables in your meals? **Veg_in_meals**, num):
    - Never (1)
    - Sometimes (2)
    - Always (3)
- How many meals do you have daily? (**N_main_meal_daily**, num):
    - Between 1 and 2 (1)
    - Three (2)
    - More than three (3)

[3] F. Mendoza Palechor, A. de la Hoz Manotas, Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico, Data in Brief, Volume 25, 2019

- Do you eat any food between meals? (**Snacks**, chr):
  - No ("no")
  - Sometimes ("Sometimes")
  - Frequently ("Frequently)
  - Always ("Always")

- Do you smoke? (**SMOKE**, chr):
  - Yes ("yes")
  - No ("no")

- How much water do you drink daily? (**Water_daily_l**, num):
  - Less than a litre (1)
  - Between 1 and 2 litres (2)
  - More than two litres (3)

- Do you monitor the calories you eat daily? (**Calory_monitor**, chr):
  - Yes ("yes")
  - No ("no")

- How often do you have physical activity? **(Physical_activity_days**, num):
  - I do not have (0)
  - 1 or 2 days (1)
  - 3 or 4 days (2)
  - 4 or 5 days (3)

- How much time do you do you use technological devices such as cell phone, video games, television, computer and others? (**Tech_use_hours**, num):
  - 0-2 hours (0)
  - 3-5 hours (1)
  - More than 5 hours (2)

- How often do you drink alcohol? (**Alcohol_consump**, chr):
  - I do not drink ("no")
  - Sometimes ("Sometimes")
  - Frequently ("Frequently")
  - Always ("Always")

- Which transportation do you usually use? (**Transportation**, chr):
  - Automobile ("Automobile")
  - Motorbike ("Motorbike")
  - Bike ("Bike")
  - Public Transportation ("Public_Transportation")
  - Walking ("Walking")

The dataset presents a set of mixed variables with 9 being categorical and 8 being numerical.

For the purpose of the analysis the categorical were converted into factors while the numerical were converted into integers because some records exhibit an unexpected number of decimal places, the potential causes could be Data entry Errors or measurement inaccuracy.

Fortunately, no null values were present while 24 duplicates were identified.
*To maintain data integrity and given the potential for inconsistencies in synthetically generated data, I employed data cleaning techniques by removing the duplicates ent*Figure 2: Gender distrib. by age group 1*ries from the dataset to ensure the accuracy and robustness of the analysis.*

1.3 EDA (Explanatory Data Analysis):

Firstly, I want to see the distribution of the population of the survey based on Age and Gender

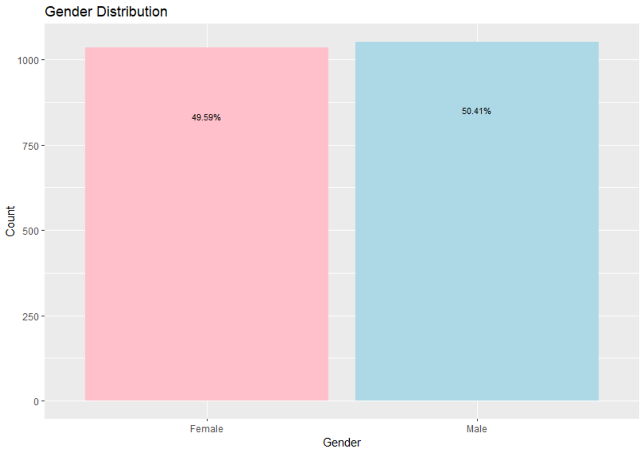


Figure 1: Gender distribution

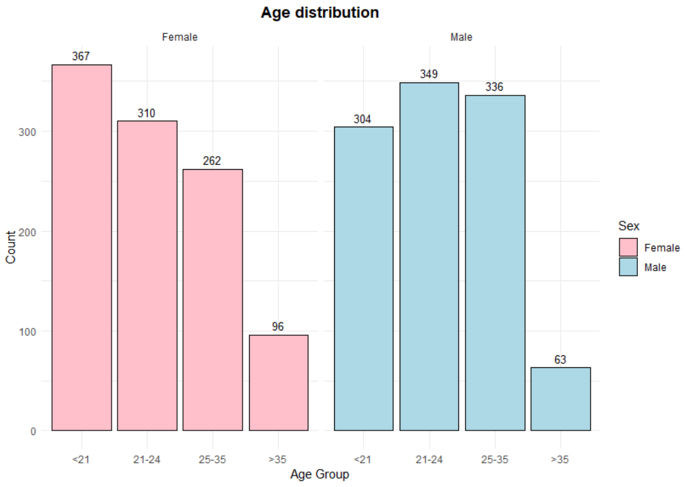


Figure 2: Gender distribution by age group

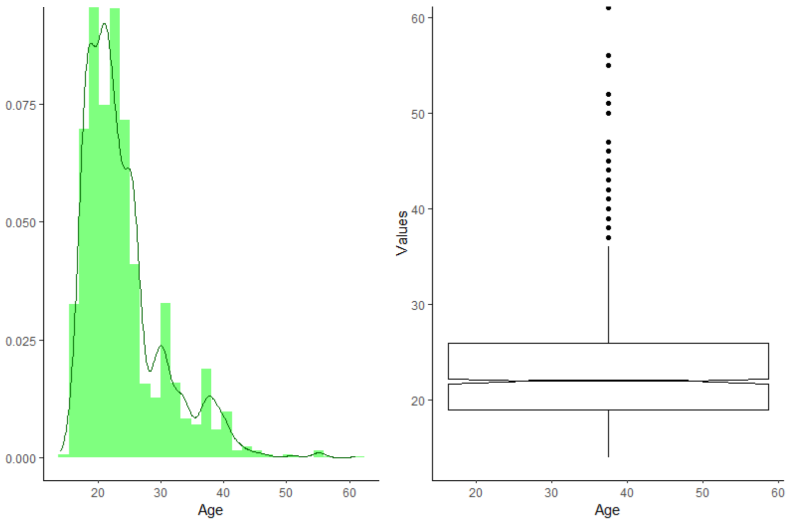


Figure 3: Age histogram and boxplot

The gender distribution seems very balanced but the same cannot be said about the age distribution where there is a high concentration of 18 -26 years old. Because of that, we can identify in the boxplot some outliers (later estimated by the IQR criterion to be 147). Since 77% of the data is synthetically generated using SMOTE, it's likely the age distribution may not perfectly reflect a natural population. Removing outliers based on the original sample (458) might not be suitable for the oversampled data and may introduce bias, by unintentionally removing valid data points that were generated to account for underrepresented groups (Mexico, Peru, Colombia) in the original sample. The conclusion is that the sample is not very representative of the overall population, and this must be kept in mind later.

With respect to the target variable, Obesity_level, the distribution seems very much balanced.

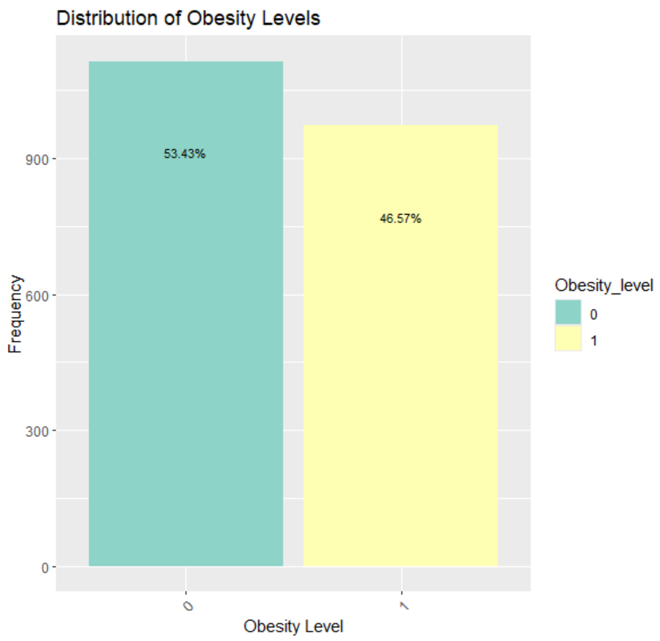


Figure 4

To understand how various factors might influence Obesity, we will now explore its relationship with other features. We'll start by examining categorical variables.

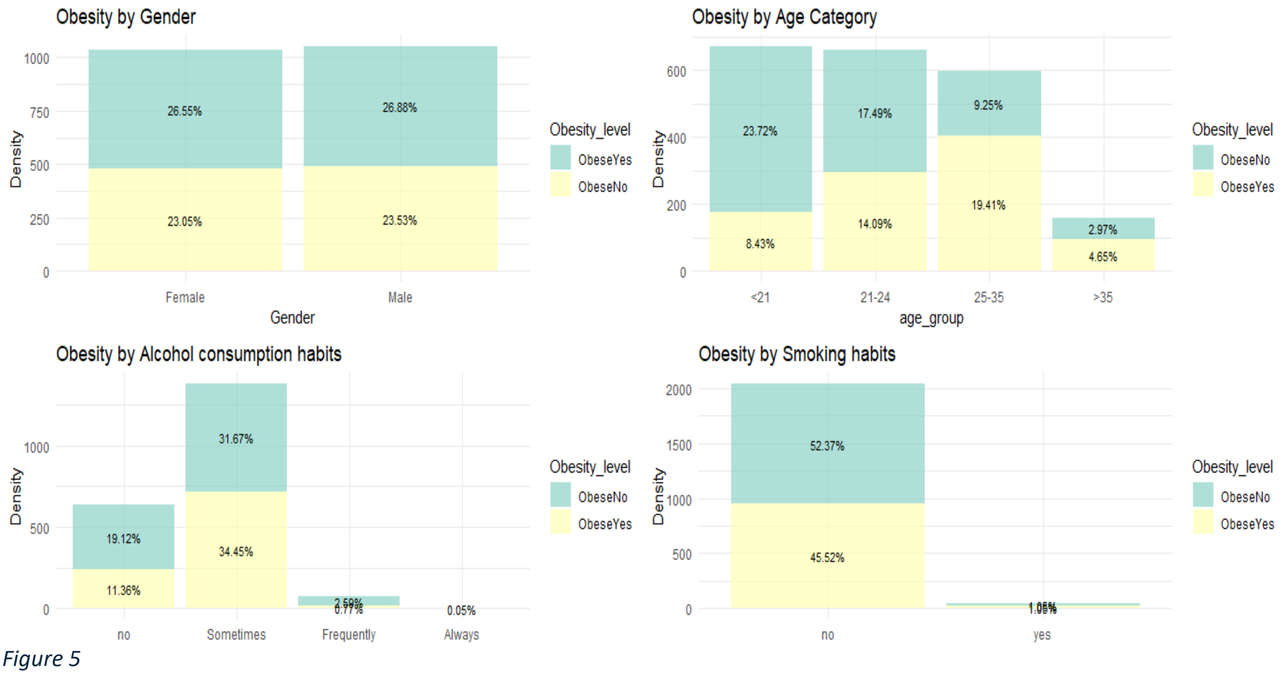


Figure 5

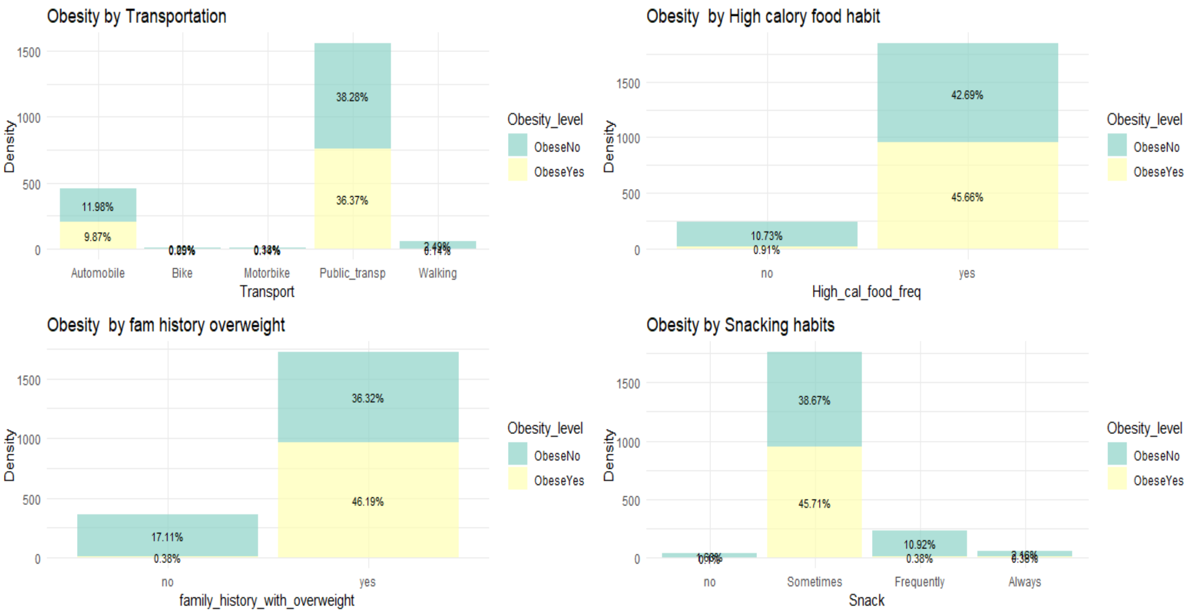


Figure 6

By taking into account the disproportion of the distribution of the features, Obesity seems to be more present in young adults (23-35 age group), that eat caloric food and have a snacking habit between meals, with a less frequent consumption of alcohol and with at least 1 family member that is suffering or has suffered from Obesity.

Next, we'll delve into some interesting aspects of some of the numerical features:

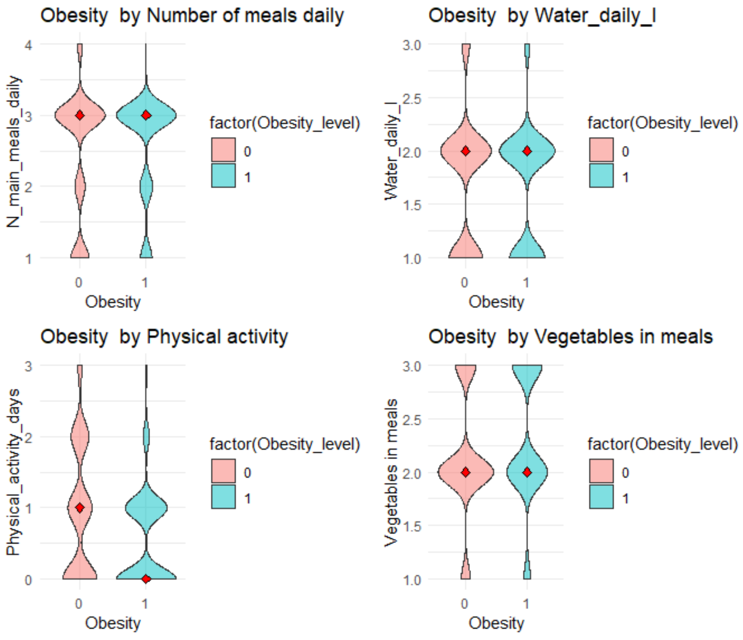


Figure 7: Violin plots of Obesity level vs numeric variables

This visualization allows us to observe potential differences in the spread and central tendency of the variables as follows.

Most of the respondents do little to no physical activity, preferring a more sedentary lifestyle, occasionally eat vegetables during their meals that tend to be more than three daily.

These were just a few insights to see if some variables could be relevant for the analysis carried on later.

By inspecting the correlation, it would help me determine if there are any variables I should drop because they're too correlated with the target variable, obesity.

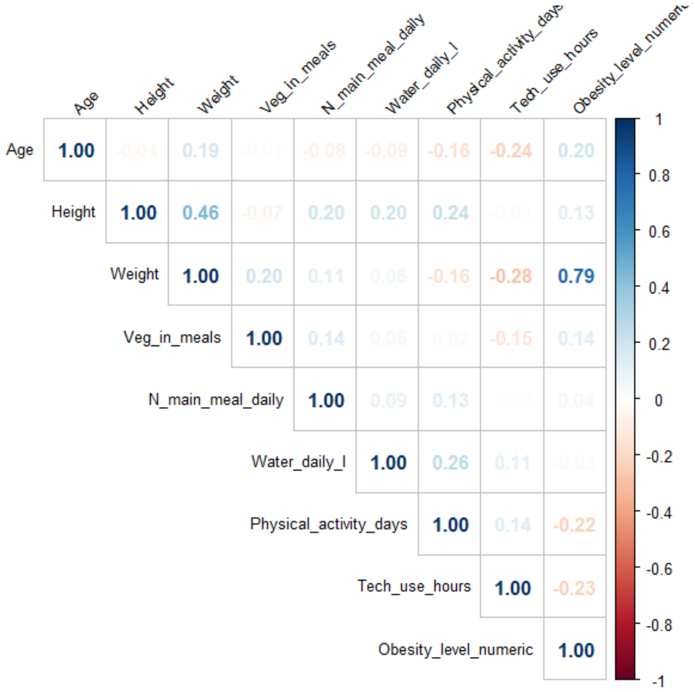


Figure 8: Correlation matrix

The weight variable is highly correlated with obesity (0.79). This makes sense, as weight is an integral factor in determining a person's BMI and so if someone is obese. Therefore, the weight variable is dropped from my analysis.

Age, while positively correlated, has a weaker relationship compared to weight.

Physical activity and technology use show negative correlations, indicating that higher physical activity and lower technology usage might be associated with lower obesity levels, though the correlations are not very strong.

Surprisingly, contrary to my assumptioThe other variables (Height, Veg_in_meals, N_main_meal_daily, and Water_daily_l) have correlations close to zero with obesity level, suggesting they might not be as influential in determining obesity in this dataset. For sure, some deeper analysis is needed.

The distribution of the numerical variables can be summarised by the following boxplots:

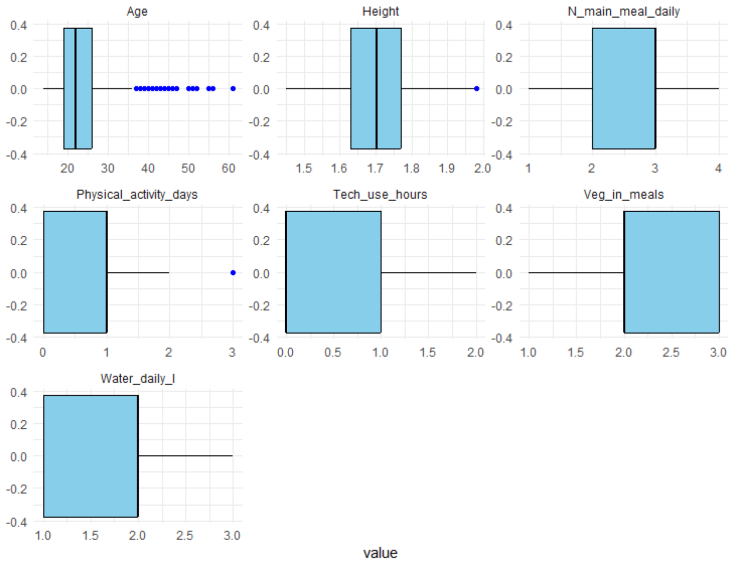


Figure 9: Boxplot numerical features

As stated before, there are few outliers, but they do not seem problematic, this can be acknowledge also by looking at the histograms.

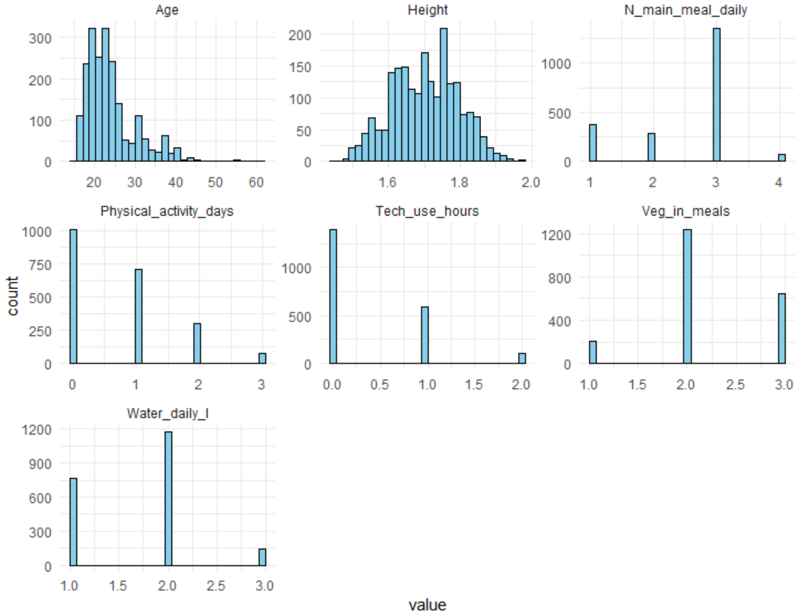


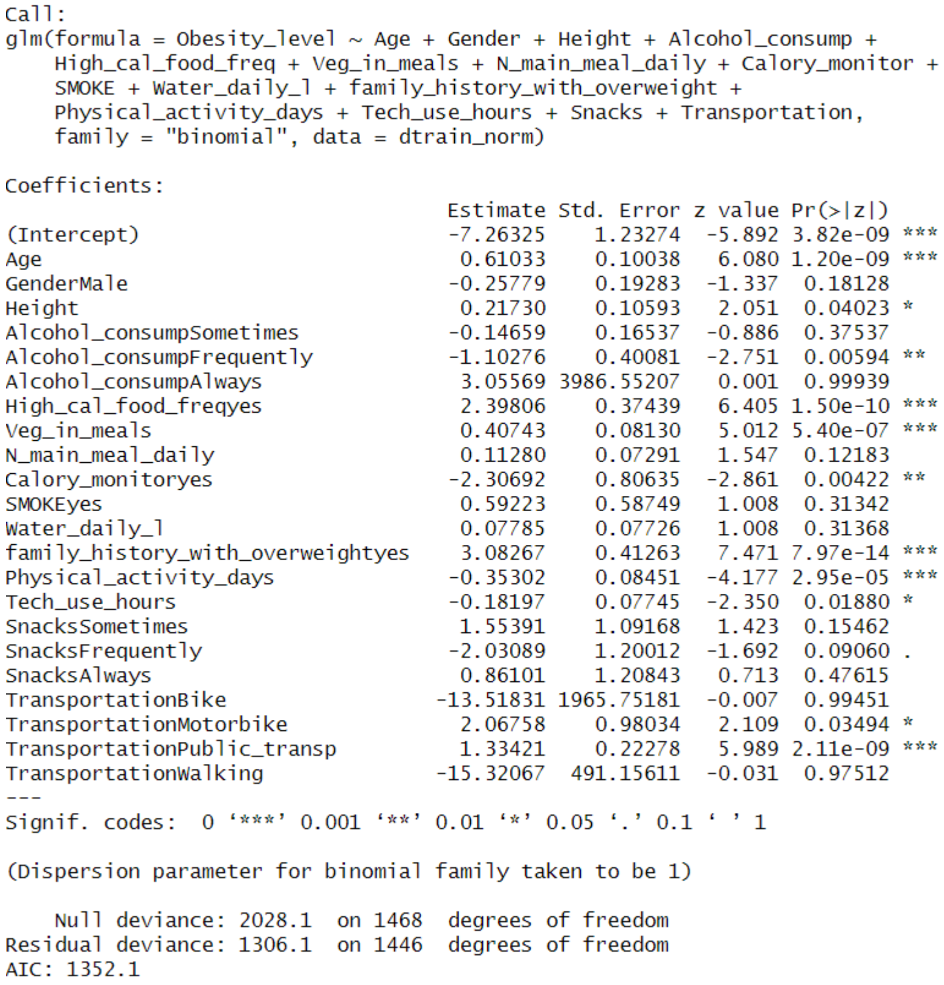
Figure 10: Histogram plots numerical features

1.4 Models

As said in the abstract, my goal was to see if Obesity condition could be predicted by lifestyle factors, and demographics of a group of individuals using statistical learning methods. Before implementing any of the models, the dataset was normalized because the ranges of the dataset features are not the same. For each model, a brief explanation will be provided along with comments on the obtained results.

1.4.1. Logistic regression

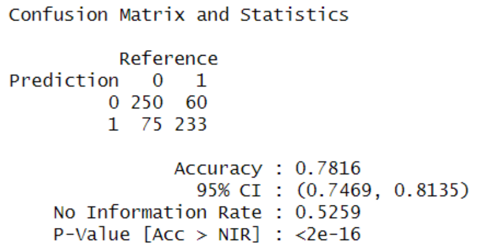
Dealing with a binary classification problem, the first supervised model that comes to mind in terms of simplicity and efficiency is the logistic model. The dataset was split in training set, containing 70% of the data, and a test set with the remaining 30%.



```
Call:
glm(formula = Obesity_level ~ Age + Gender + Height + Alcohol_consump + 
    High_cal_food_freq + Veg_in_meals + N_main_meal_daily + Calory_monitor + 
    SMOKE + Water_daily_l + family_history_with_overweight + 
    Physical_activity_days + Tech_use_hours + Snacks + Transportation, 
    family = "binomial", data = dtrain_norm)

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)    
(Intercept)                     -7.26325    1.23274  -5.892 3.82e-09 ***
Age                              0.61033    0.10038   6.080 1.20e-09 ***
GenderMale                      -0.25779    0.19283  -1.337  0.18128    
Height                           0.21730    0.10593   2.051  0.04023 *  
Alcohol_consumpSometimes        -0.14659    0.16537  -0.886  0.37537    
Alcohol_consumpFrequently       -1.10276    0.40081  -2.751  0.00594 ** 
Alcohol_consumpAlways            3.05569 3986.55207   0.001  0.99939    
High_cal_food_freqyes            2.39806    0.37439   6.405 1.50e-10 ***
Veg_in_meals                     0.40743    0.08130   5.012 5.40e-07 ***
N_main_meal_daily                0.11280    0.07291   1.547  0.12183    
Calory_monitoryes               -2.30692    0.80635  -2.861  0.00422 ** 
SMOKEyes                         0.59223    0.58749   1.008  0.31342    
Water_daily_l                    0.07785    0.07726   1.008  0.31368    
family_history_with_overweightyes  3.08267    0.41263   7.471 7.97e-14 ***
Physical_activity_days          -0.35302    0.08451  -4.177 2.95e-05 ***
Tech_use_hours                  -0.18197    0.07745  -2.350  0.01880 *  
SnacksSometimes                  1.55391    1.09168   1.423  0.15462    
SnacksFrequently                -2.03089    1.20012  -1.692  0.09060 .  
SnacksAlways                     0.86101    1.20843   0.713  0.47615    
TransportationBike             -13.51831 1965.75181  -0.007  0.99451    
TransportationMotorbike          2.06758    0.98034   2.109  0.03494 *  
TransportationPublic_transp      1.33421    0.22278   5.989 2.11e-09 ***
TransportationWalking          -15.32067  491.15611  -0.031  0.97512    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2028.1  on 1468  degrees of freedom
Residual deviance: 1306.1  on 1446  degrees of freedom
AIC: 1352.1
```

Figure 11: Logistic coefficients

Age (p < 0.001) positively correlates with the outcome, indicating increasing age raises odds. High-calorie food frequency, Family history of overweight (p < 0.001) strongly associates with the outcome, while Physical activity days (p < 0.001) inversely relate to the outcome.
One surprise was that Certain transportation modes (motorbike, p = 0.035; public transport, p < 0.001) influence the outcome.



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 250  60
         1  75 233

               Accuracy : 0.7816
                 95% CI : (0.7469, 0.8135)
    No Information Rate : 0.5259
    P-Value [Acc > NIR] : <2e-16
```

Figure 12: Confusion Matrix and accuracy

The logistic regression achieves a 78.16% accuracy. It is crucial to look at the ROC (Receiver Operating Characteristic) curve, which is displaying the sensitivity (along the y axis) and the specificity (1-specificity along the x axis) at the same time.

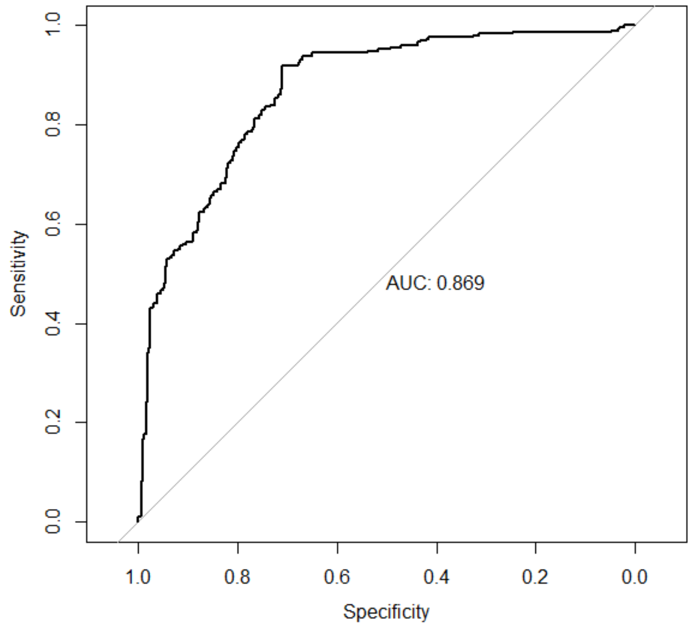


Figure 13: ROC curve and AUC

The Area Under the ROC Curve, called AUC, is also displayed. The higher the AUC (which varies between 0-1), the better the model. If the AUC is 0.5 it means that the model is no better than a random guessing. It is safe to say, given a 0.869 value, the logistic model has a decent power of discrimination.

1.4.2. The Lasso Model

To enhance the logistic regression model's performance, I then opted to employ the Lasso. This model was chosen for its ability to potentially select a subset of variables while discarding others, thereby achieving dimensionality reduction, and retaining only the crucial variables. Initially, I conducted cross-validation to tune the lambda parameter.

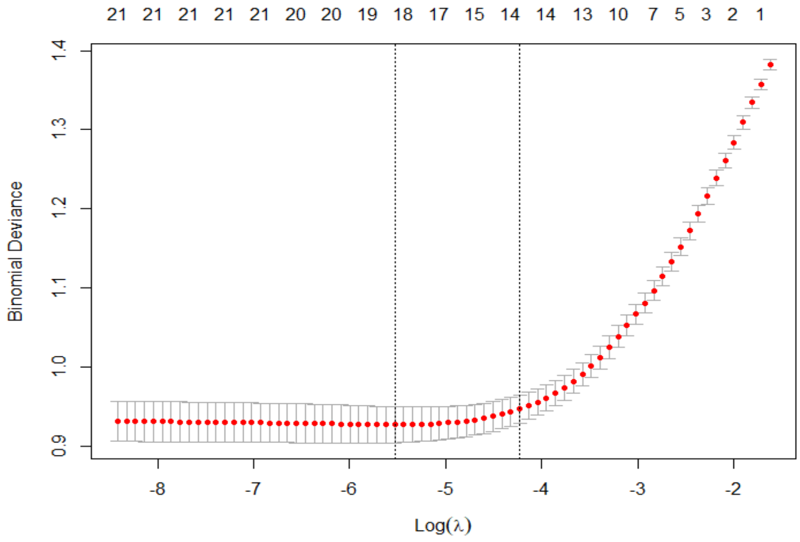
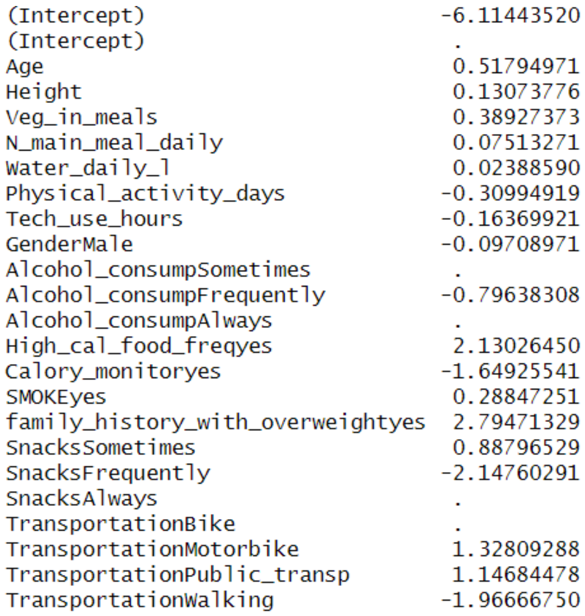


Figure 14: Tuning Lambda

After determining the optimal lambda value, I run the model on the training set and subsequently evaluated its performance on the test set.

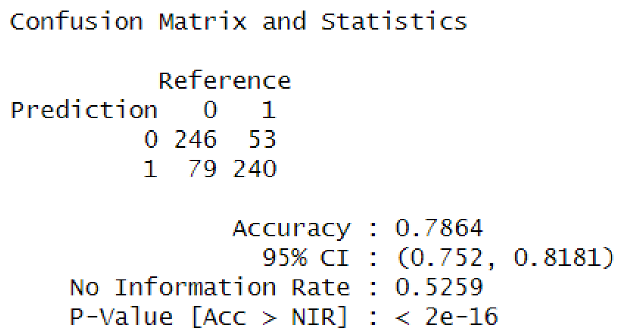
It's noteworthy that the Lasso model employs a penalty term that shrinks some variable coefficients to zero. These zeroed coefficients indicate that certain variables were deemed less influential or redundant in predicting obesity levels by the Lasso model. In my analysis the model has effectively shrunken to 0 the following variables.



```
(Intercept)                        -6.11443520
(Intercept)                          .
Age                                 0.51794971
Height                              0.13073776
Veg_in_meals                        0.38927373
N_main_meal_daily                   0.07513271
Water_daily_l                       0.02388590
Physical_activity_days             -0.30994919
Tech_use_hours                     -0.16369921
GenderMale                         -0.09708971
Alcohol_consumpSometimes             .
Alcohol_consumpFrequently          -0.79638308
Alcohol_consumpAlways                .
High_cal_food_freqyes               2.13026450
Calory_monitoryes                  -1.64925541
SMOKEyes                            0.28847251
family_history_with_overweightyes   2.79471329
SnacksSometimes                     0.88796529
SnacksFrequently                   -2.14760291
SnacksAlways                         .
TransportationBike                   .
TransportationMotorbike             1.32809288
TransportationPublic_transp         1.14684478
TransportationWalking              -1.96666750
```

Figure 15: Lasso Coefficients

To assess the model's effectiveness, I’ve generated a confusion matrix to examine the outcomes.



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 246  53
         1  79 240

               Accuracy : 0.7864
                 95% CI : (0.752, 0.8181)
    No Information Rate : 0.5259
    P-Value [Acc > NIR] : < 2e-16
```

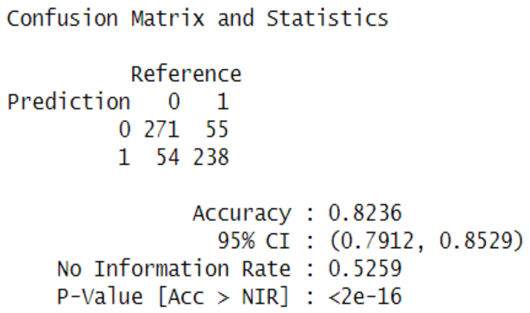
Figure 16: Confusion matrix and accuracy

The model demonstrates an accuracy of 78.64%, that is slightly higher than the one of the logistic model, but I've obtained a more simpler model, with less variables.

1.4.3. Generalized additive model:

The next model that will be implemented is Generalized additive model. Generalized additive models (GAMs) provide a general framework for generalized extending a standard linear model by allowing non-linear functions of each additive model of the variables, while maintaining additivity[4]. GAMs are more useful when the relationship between predictors and the target is nonlinear and can't be easily captured by linear models like logistic regression. Flexibility and non-linearity accommodated by GAMs may capture more complex relationships between predictors and the outcome, leading to enhanced predictive performance. By allowing for non-linear relationships between predictors and the outcome, I wanted to find out if GAMs can better accommodate the diverse range of factors contributing to obesity classification.

I run the model on the training set and subsequently evaluated its performance on the test set.



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 271  55
         1  54 238

               Accuracy : 0.8236
                 95% CI : (0.7912, 0.8529)
    No Information Rate : 0.5259
    P-Value [Acc > NIR] : <2e-16
```

Figure 17:Confusion matrix and accuracy

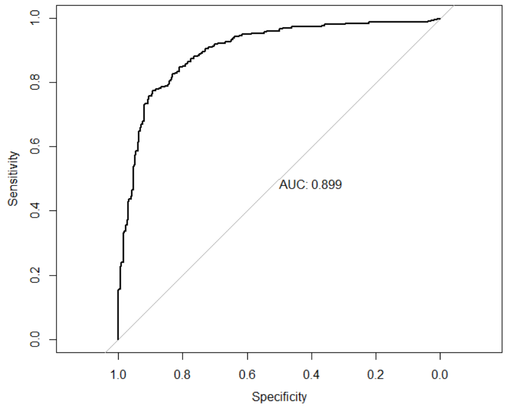


Figure 18: ROC curve and AUC

As we can see the accuracy reached is higher, 82.36% with an AUC value of 0.899 value, slightly higher than the two previous models. Overall, the GAM model's superior accuracy indicates its potential as a powerful tool for capturing intricate patterns in the data but interpreting it can be more complex.

[4] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (1st ed.) [PDF]. Springer.

While logistic regression provides the straightforward interpretations of coefficients as log-odds ratios, and Lasso regression offers variable selection and coefficient shrinkage, GAMs introduce additional complexity. For that reason, the next model was implemented.

1.4.4. Decision Tree:

As I've said, Decision Tree was then implemented especially for its main advantage of being easy to interpret and visualize.
To build the decision tree, I initially developed a large tree and subsequently pruned it until reaching a specific threshold for the Cost of Pruning (CP). This approach ensured that the resulting model avoids overfitting and maintains controllable error levels.

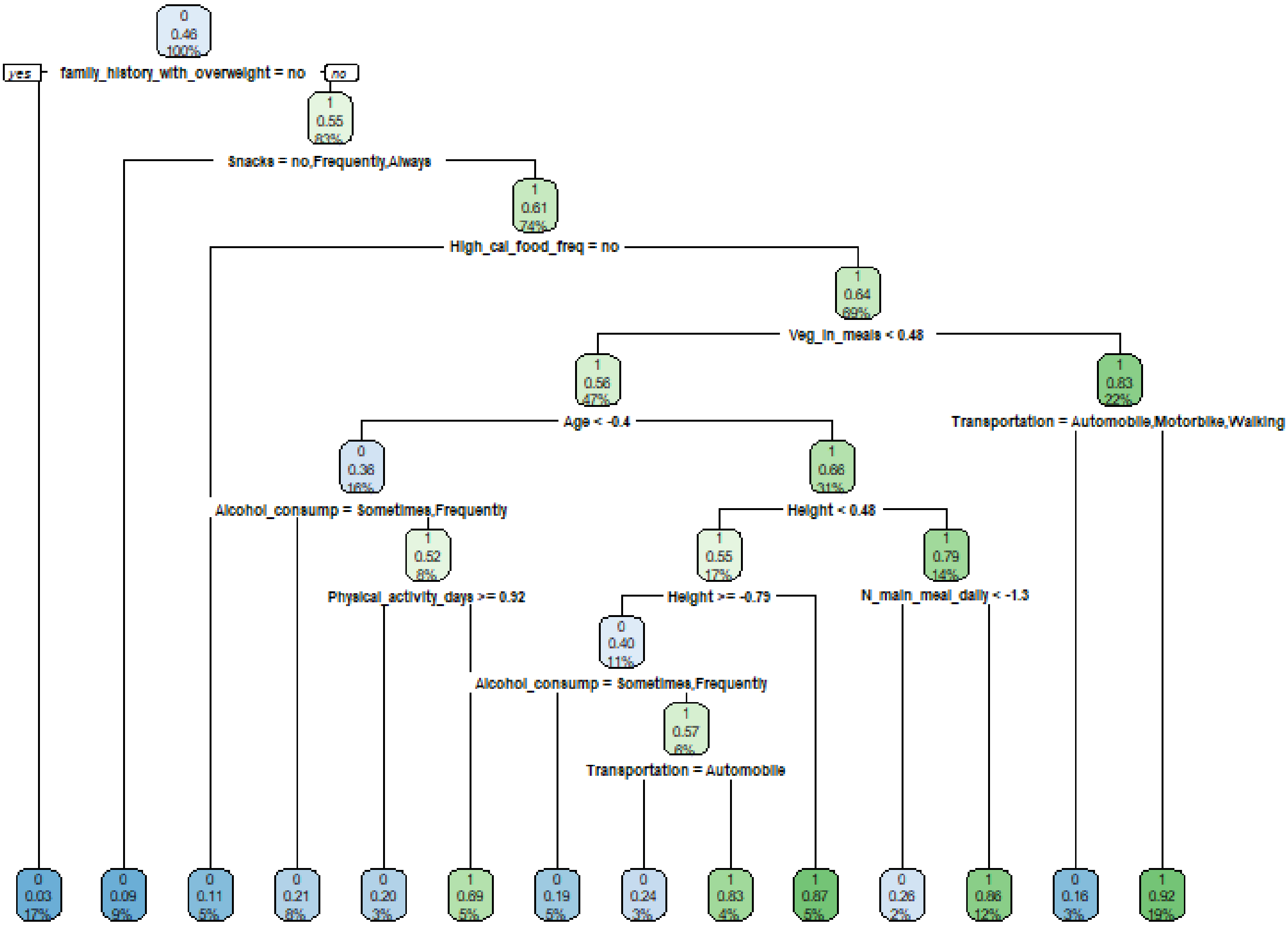
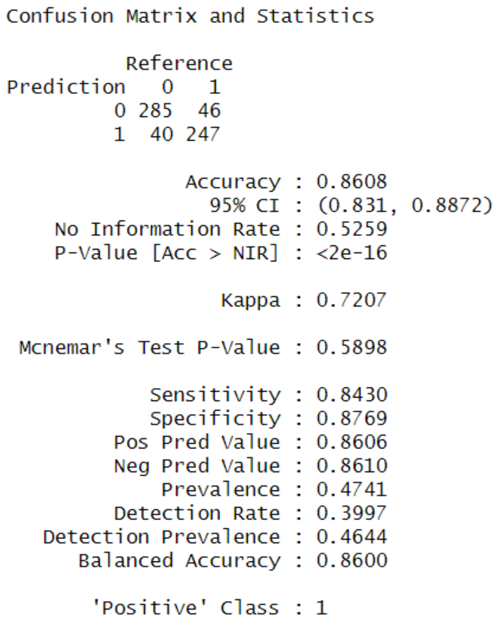


Figure 19: Tree plot

In the tree plot the ideal number of splits appears to be 13 with family history with overweight being the best discriminant variable, followed by eating habits variable like Snacks, High Calory food consumption and Vegetable in the meals.

I then print the confusion matrix.



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 285  46
         1  40 247

               Accuracy : 0.8608
                 95% CI : (0.831, 0.8872)
    No Information Rate : 0.5259
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7207

 Mcnemar's Test P-Value : 0.5898

            Sensitivity : 0.8430
            Specificity : 0.8769
         Pos Pred Value : 0.8606
         Neg Pred Value : 0.8610
             Prevalence : 0.4741
         Detection Rate : 0.3997
   Detection Prevalence : 0.4644
      Balanced Accuracy : 0.8600

       'Positive' Class : 1
```

Figure 20: Confusion matrix and accuracy 1

The tree classifier is able to achieve 86.08% accuracy on the test set. Sensitivity, representing the true positive rate (TPR = TP / (TP + FN)), measures the model's ability to identify truly obese individuals. Here, the model accurately identifies 84.30% of individuals who are genuinely obese. Specificity, denoting the true negative rate (TNR = TN / (TN + FP)), gauges the model's proficiency in identifying non-obese individuals correctly. The model achieves a specificity of 87.69%, correctly classifying the majority of non-obese individuals.

1.4.5.Random Forest

Knowing that Random Forest usually performs better than decision tree models, due to its ability to combine multiple trees enhancing precision, I decided to implement it as my next model. Initially, I fine-tuned the 'mtry' hyperparameter, determining the number of variables randomly selected at each node split. Subsequently, I've tuned the optimal number of trees, plotting the Out-of-Bag (OOB) error, and identifying the point of minimal error. I can plot this quantity as a function of the number of trees and even to identify with different colours the errors for value “0”, “1” and the general one.

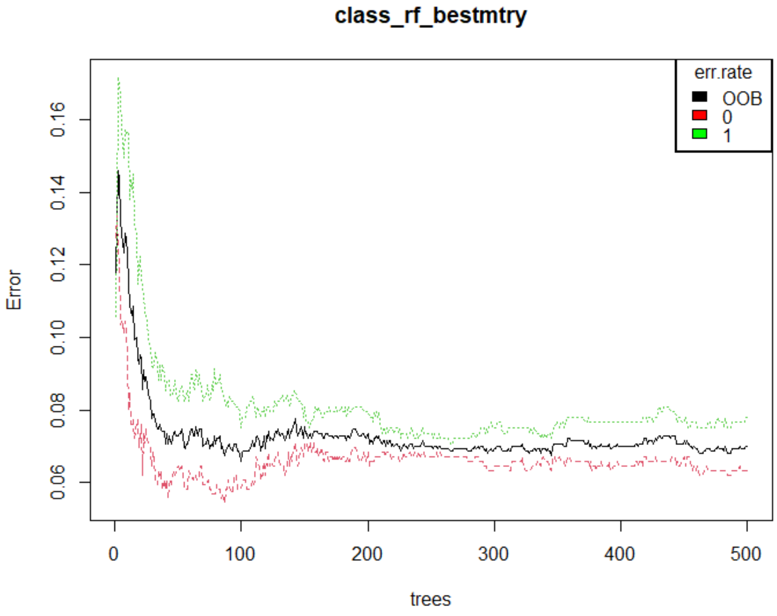
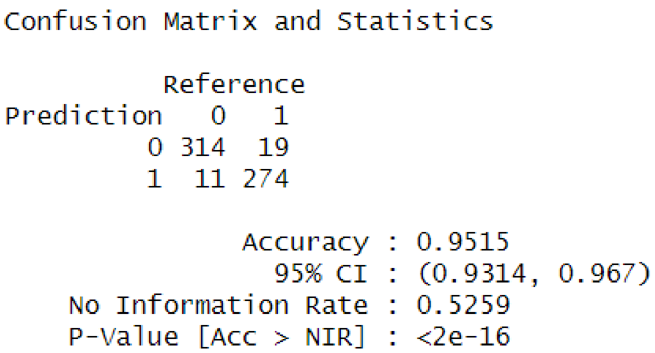


Figure 21: OOB error and number of trees

The graph shows the Random Forest estimated test error as a function of the number of trees, and even displaying with different colours the errors: the green representing the Obese class (1) the red representing the Non-Obese class (0) and the black one representing and the general error.

I've then built the model with the tuned hyperparameters, and I've evaluated the performance with the confusion matrix.



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 314  19
         1  11 274
                                          
               Accuracy : 0.9515          
                 95% CI : (0.9314, 0.967) 
    No Information Rate : 0.5259          
    P-Value [Acc > NIR] : <2e-16          
```

Figure 22: Confusion matrix and accuracy

Among all the models experimented with so far, the accuracy of this random forest model stands out as the highest, reaching 95.15%.
An additional advantage of employing a random forest model is its capability to measure the importance of each variable and quantify the increase in error if that feature were excluded.
The *Mean Decrease Accuracy* plot illustrates the extent of accuracy reduction achieved by removing each variable from the model. Meanwhile, the *mean decrease in Gini coefficient* serves as an indicator of each variable's contribution to the homogeneity of the nodes and leaves within the resulting Random Forest. The higher the value of mean decrease accuracy or mean decrease Gini score, the higher the importance of the variable in the model.

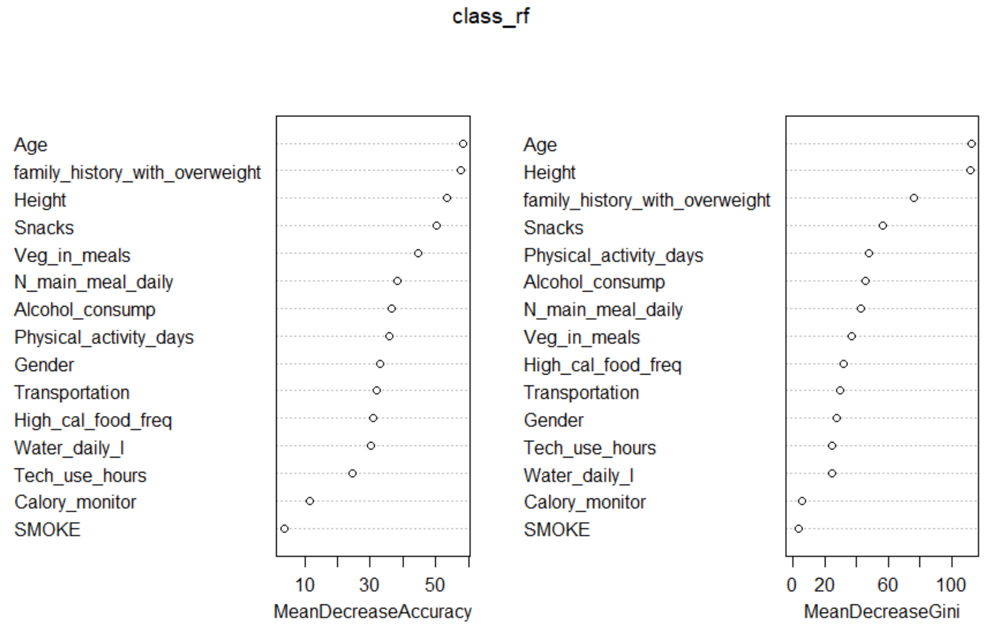


Figure 23: Importance of the variables in the model

As expected, Age and Family history represent the most important variables but a difference with respect to the other models is the importance given by the Height variable that logically play a pivotal role in determining BMI and therefore Obesity. Also, Alcohol consumption that in the other models was not classified as significant, here it is placed in a higher relevance scale, reflecting the reality of Alcohol consumption being a reason for weight gain.

1.5. Conclusions and Key points

After implementing and testing all the models, we can say that:

- The Lasso model builds upon logistic regression by incorporating regularization, resulting in a slightly improved accuracy of 78.64%. It effectively selects a subset of relevant variables, thereby reducing model complexity and enhancing interpretability. However, it may overlook interactions between variables due to its linear nature.
- The GAM outperforms both logistic regression and the Lasso model. By allowing for nonlinear relationships between predictors and outcomes. However, interpreting GAMs was challenging due to their flexibility and nonlinearity,
- The Random Forest model outperforms all others in predicting whether someone is obese or not, boasting an accuracy of 95.15%.
- Across all the models examined, a consistent finding emerges: the family environment and history significantly influence the likelihood of developing the illness. This observation reflects the reality that one's upbringing profoundly shapes dietary choices (such as meal frequency, consumption of high-calorie foods, and snacking habits) and lifestyle behaviours (including physical activity levels).
- It is important to consider that the original data were collected from individuals not representative of the overall population as already said. It could be interesting to conduct the

same kind of analysis in other parts of the world, and comparing the results, since eating habits and physical activities may be different from region to region.

2. Unsupervised learning

Unveiling Tourist Personas

2.1. Abstract

In the dynamic landscape of the tourist market, understanding customer preferences and behaviours is paramount for businesses to tailor their offerings effectively. User reviews serve as a rich source of information, offering valuable insights into customer experiences, preferences, and satisfaction levels. Recognizing the significance of these reviews, my project is focused on uncovering hidden patterns within them.

The question that I've posed to myself was: *Does a reviewer's affinity for certain attraction types in turn show affinity for others?*

By analysing patterns in user ratings across diverse attraction types, the aim is to identify clusters of traveller's personalities, with similar preferences and characteristics.
This exploration may reveal valuable insights, such as whether certain attraction types attract a dedicated following or if there are hidden gems that appeal to a niche audience.

PCA is the first unsupervised model that will be applied. Principal component analysis is a technique for dimensionality reduction while retaining as much of the original information as possible, but it is useful also for visualization purposes.
K-means clustering will be implemented next. The algorithm groups similar observations together and uncover underlying patterns within the data, which is indeed the purpose of the analysis.

2.2. Dataset

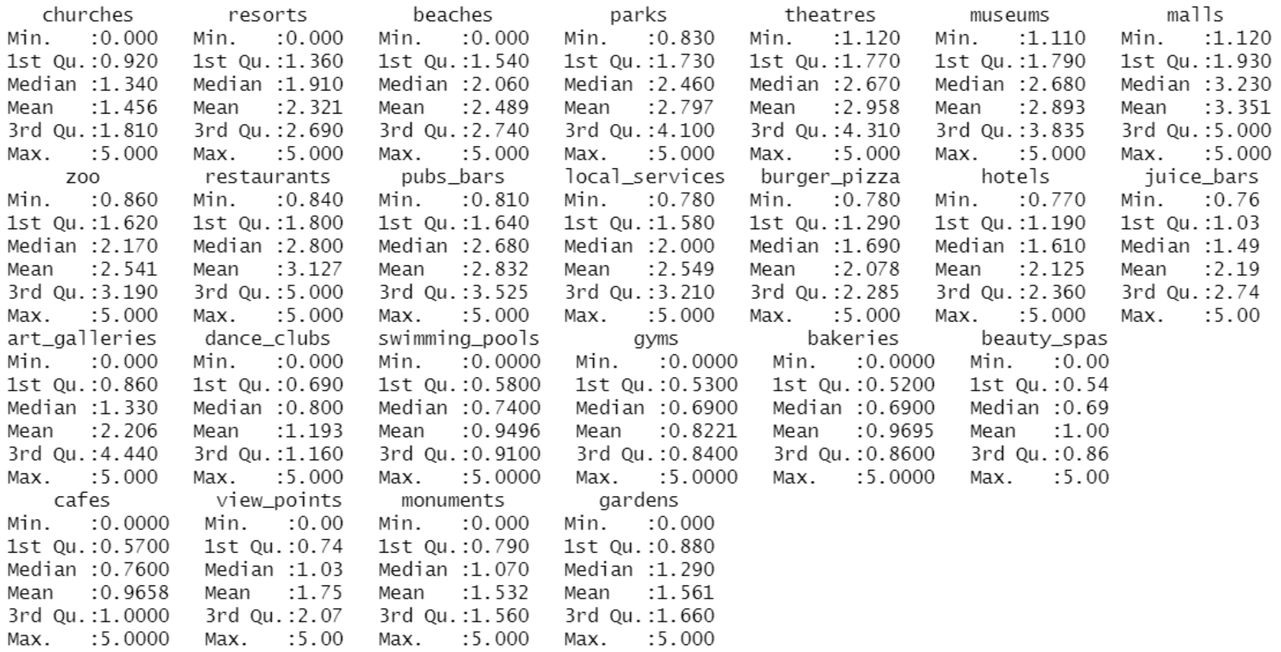
The dataset "Travel Reviews Ratings" was taken from UC Irvine Machine learning Repository and has 5456 rows and 26 columns.
The data set is populated by capturing user ratings from Google reviews. Reviews on attractions from 24 categories across Europe are considered. Google user rating ranges from 1 to 5 and average user rating per category is calculated[5].
For the purpose of the analysis, the first variable (User, chr) and the last variable (X, *NA*) were drop ped and the remaining variables (24), were renamed to be more interpretable as follows:

- churches
- resorts
- beaches

[5] https://archive.ics.uci.edu/dataset/485/tarvel+review+ratings

- parks
- theatres
- museums
- malls
- zoo
- restaurants
- pubs_bars
- local_services
- burger_pizza
- hotels
- juice_bars
- art_galleries
- dance_clubs
- swimming_pools
- gyms
- bakeries
- beauty_spascafes
- view_points
- monuments
- gardens

All are of type *dbl*, except local_services that was of type *chr* so I've quickly converted it into numerical. I've then identified NA values and duplicates dropping the corresponding rows.
The summary of my variables is as follows:



```
    churches        resorts         beaches           parks          theatres        museums           malls      
 Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.830   Min.   :1.120   Min.   :1.110   Min.   :1.120  
 1st Qu.:0.920   1st Qu.:1.360   1st Qu.:1.540   1st Qu.:1.730   1st Qu.:1.770   1st Qu.:1.790   1st Qu.:1.930  
 Median :1.340   Median :1.910   Median :2.060   Median :2.460   Median :2.670   Median :2.680   Median :3.230  
 Mean   :1.456   Mean   :2.321   Mean   :2.489   Mean   :2.797   Mean   :2.958   Mean   :2.893   Mean   :3.351  
 3rd Qu.:1.810   3rd Qu.:2.690   3rd Qu.:2.740   3rd Qu.:4.100   3rd Qu.:4.310   3rd Qu.:3.835   3rd Qu.:5.000  
 Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000  
      zoo         restaurants      pubs_bars     local_services   burger_pizza       hotels         juice_bars   
 Min.   :0.860   Min.   :0.840   Min.   :0.810   Min.   :0.780   Min.   :0.780   Min.   :0.770   Min.   :0.76  
 1st Qu.:1.620   1st Qu.:1.800   1st Qu.:1.640   1st Qu.:1.580   1st Qu.:1.290   1st Qu.:1.190   1st Qu.:1.03  
 Median :2.170   Median :2.800   Median :2.680   Median :2.000   Median :1.690   Median :1.610   Median :1.49  
 Mean   :2.541   Mean   :3.127   Mean   :2.832   Mean   :2.549   Mean   :2.078   Mean   :2.125   Mean   :2.19  
 3rd Qu.:3.190   3rd Qu.:5.000   3rd Qu.:3.525   3rd Qu.:3.210   3rd Qu.:2.285   3rd Qu.:2.360   3rd Qu.:2.74  
 Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.00  
 art_galleries    dance_clubs    swimming_pools        gyms            bakeries       beauty_spas   
 Min.   :0.000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00  
 1st Qu.:0.860   1st Qu.:0.690   1st Qu.:0.5800   1st Qu.:0.5300   1st Qu.:0.5200   1st Qu.:0.54  
 Median :1.330   Median :0.800   Median :0.7400   Median :0.6900   Median :0.6900   Median :0.69  
 Mean   :2.206   Mean   :1.193   Mean   :0.9496   Mean   :0.8221   Mean   :0.9695   Mean   :1.00  
 3rd Qu.:4.440   3rd Qu.:1.160   3rd Qu.:0.9100   3rd Qu.:0.8400   3rd Qu.:0.8600   3rd Qu.:0.86  
 Max.   :5.000   Max.   :5.000   Max.   :5.0000   Max.   :5.0000   Max.   :5.0000   Max.   :5.00  
     cafes          view_points      monuments        gardens     
 Min.   :0.0000   Min.   :0.00   Min.   :0.000   Min.   :0.000  
 1st Qu.:0.5700   1st Qu.:0.74   1st Qu.:0.790   1st Qu.:0.880  
 Median :0.7600   Median :1.03   Median :1.070   Median :1.290  
 Mean   :0.9658   Mean   :1.75   Mean   :1.532   Mean   :1.561  
 3rd Qu.:1.0000   3rd Qu.:2.07   3rd Qu.:1.560   3rd Qu.:1.660  
 Max.   :5.0000   Max.   :5.00   Max.   :5.000   Max.   :5.000  
```

Figure 24: Summary statistics

2.3. Explanatory Data Analysis:

Before anything, an observation that I want to make is that all the 24 features tend to follow a similar distribution like the two displayed below, with higher frequency when the ratings on average are close to 0 or close to 5:

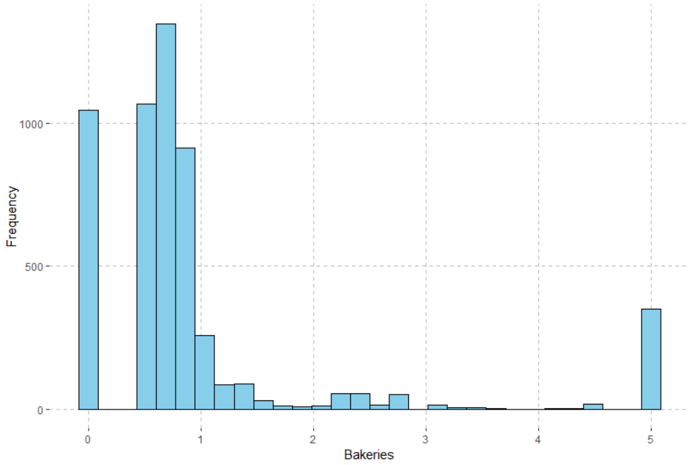


Figure 25

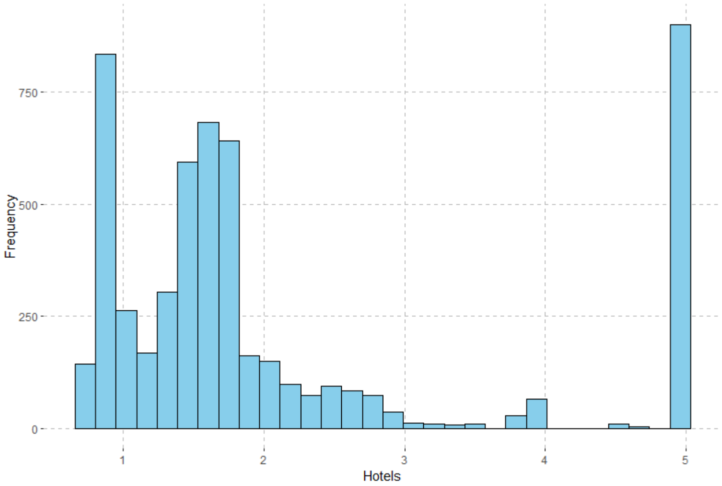


Figure 26

I've tried also to Consider the ratings collectively, "melting" together all the categories and examining the overall trend in people's rating choices, independently of the specific object of rating. The distribution was as follows:

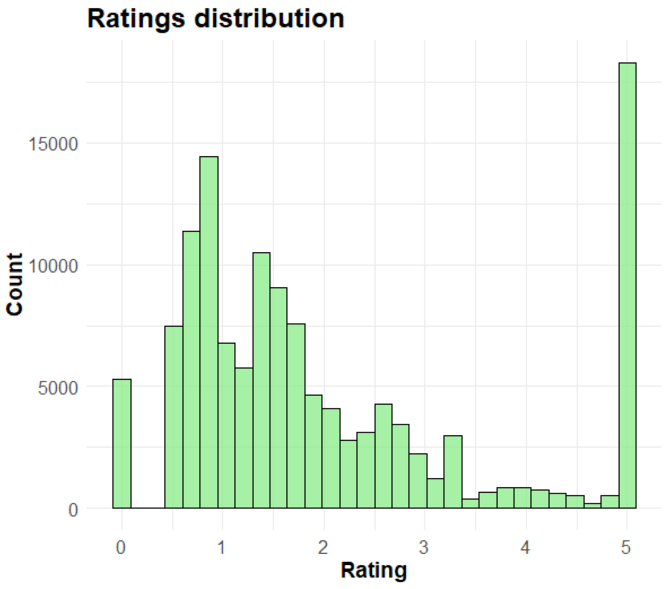
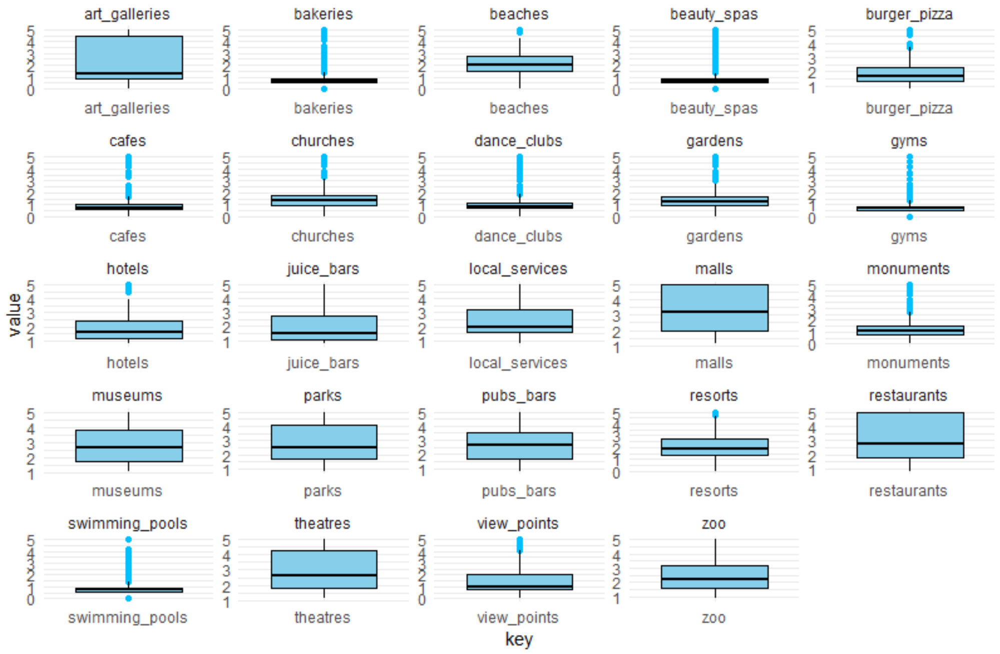


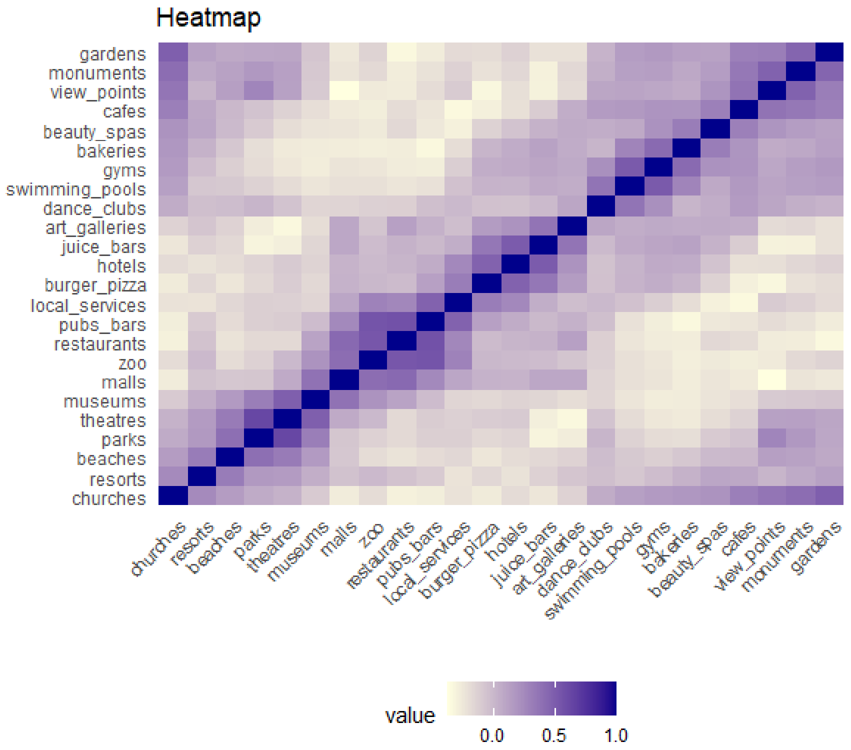
Figure 27 1

As expected, there are some outliers in 2 points: 0 (about 5000) and 5(about 17500).

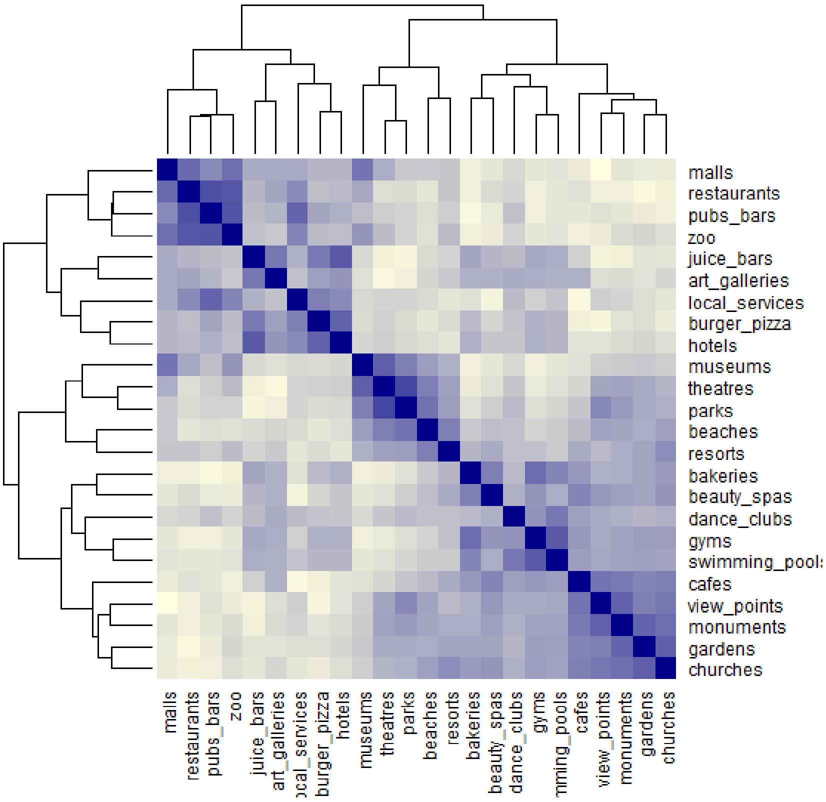
I expect that outliers will be depicted when plotting the boxplot, with the following result:



In my opinion this represents true reviews, people usually tend to rate more when they have a bad or a very good experience, so I've expected to have higher frequencies in these 2 "extreme" points of the rating range. Because of that, outliers will not be removed and in order to avoid them interfering later on k-means clustering, the dataset will be scaled.



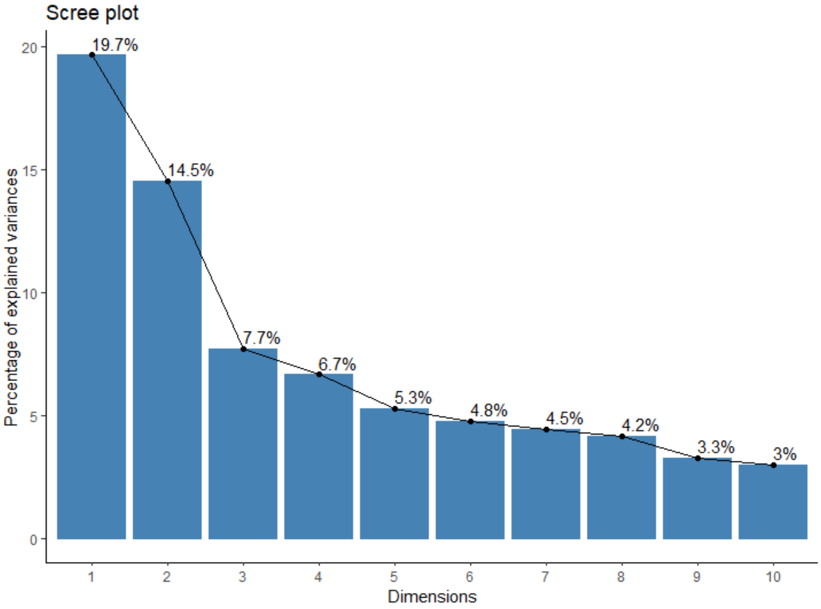
From the Heat Map, it is quite clear that some attraction types of groups have particularly strong correlations. For instance, Viewpoints and Monuments (0.4714), Zoos and Pubs/Bars (0.5520), Gyms and Swimming Pools (0.5147). These strong correlations indicate a potentially high degree of association between these attraction types, suggesting that they might cater to similar visitor preferences or demographics.



Just for visualization's sake, I've also plotted the cluster Map, making me even more sure that the data is actually divided into groups. Unfortunately, this is shown to be too fragmented.

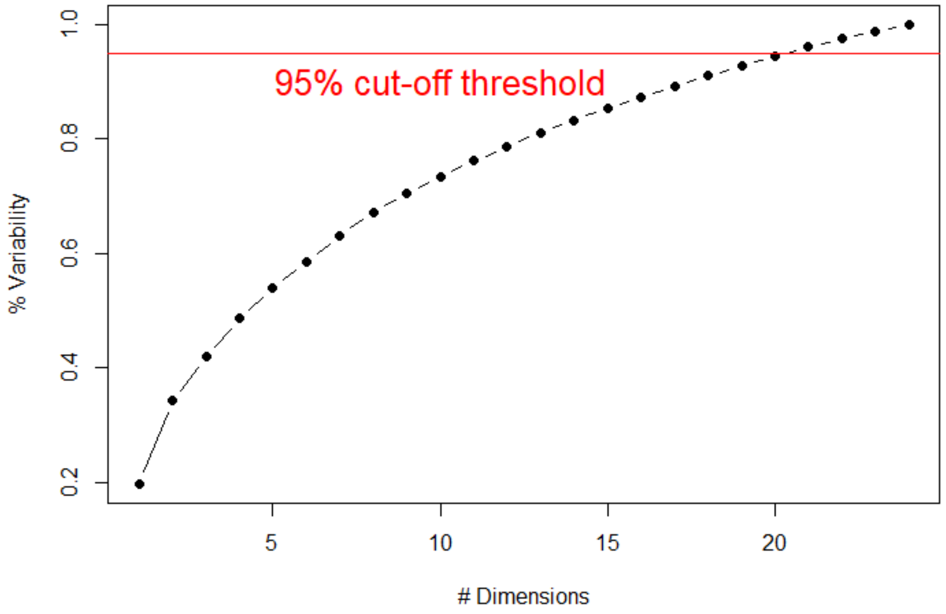
2.4. Principal Component Analysis

Given the fact that I'm dealing with a lot of variables, Principal Component Analysis was the first unsupervised learning technique that I've implemented.
As stated before, PCA is a useful method for dimensionality reduction, but it can also bring insights even for visualization.



The reduction in dimensionality has the expensive cost to reduce the variability explained.
It is a good practice to look for the 'elbow' in the scree Plot (Figure number), which represent the point from which adding more variables will lower the raising in the explained variability, meaning where it becomes disadvantageous to add dimensions.

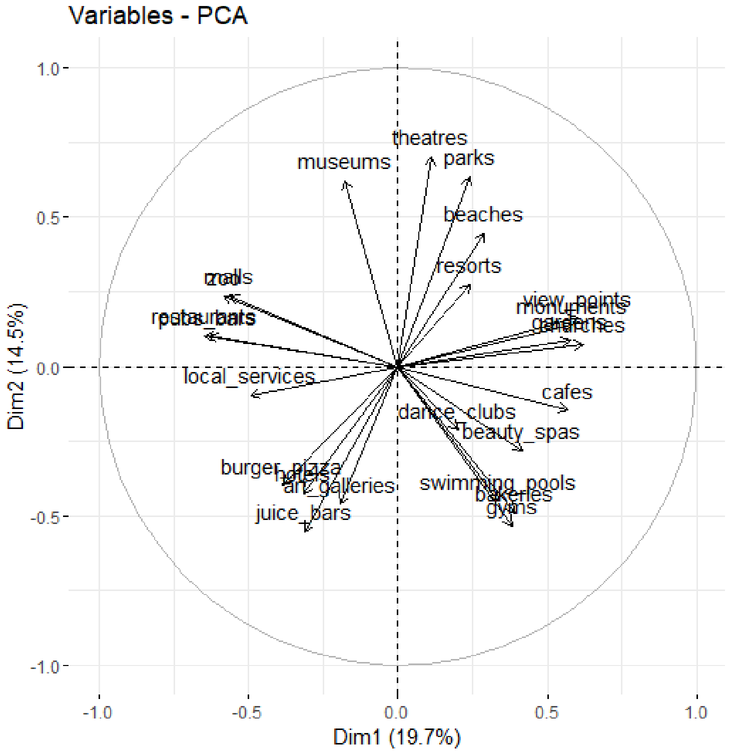
Contrary to what I've hoped, the first 2 dimensions explain only the 34.2% of the variability, meaning that the 2-dimensional best representations of the dataset will allow only to visualize 34.2% of the spread of the data.

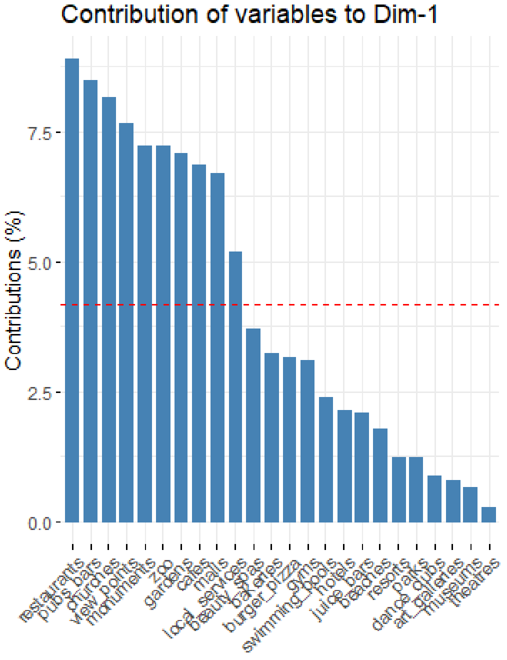


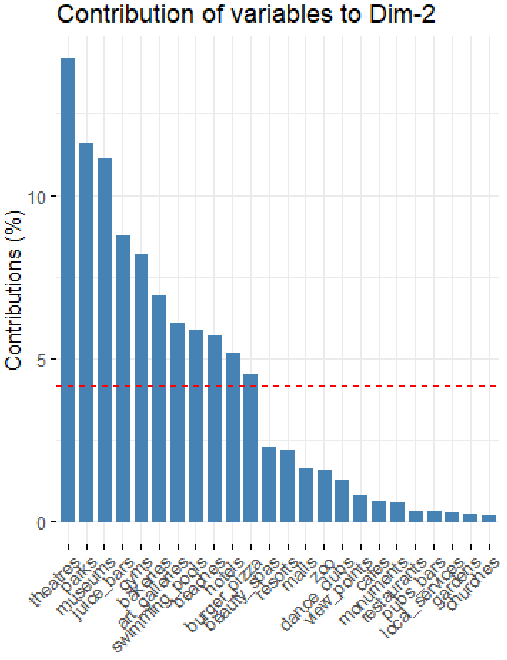
To get 95% of variance explained I need 21 principal components, which is not the dimensionality reduction that I've hoped for.

Through PCA, I'm able also to visualize the relationship between the variables, this is known as correlation plot. To interpret it:

- Positively correlated variables are grouped together. It is interesting that burger_pizza and art galleries are close to each other, or even bakeries with gyms.
- Negatively correlated variables are positioned on opposite sides of the plot origin (opposed quadrants). It is quite strange that monuments and art galleries are at the opposite sides



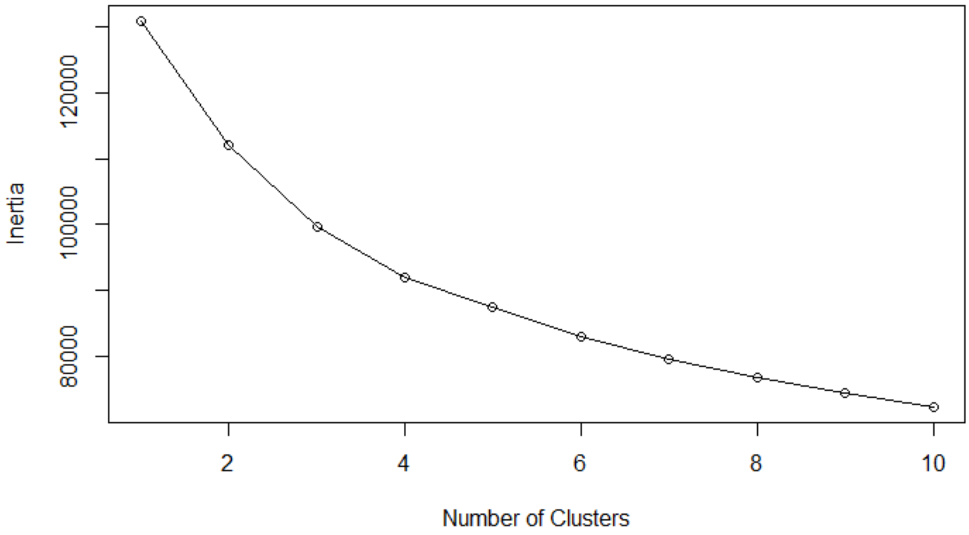




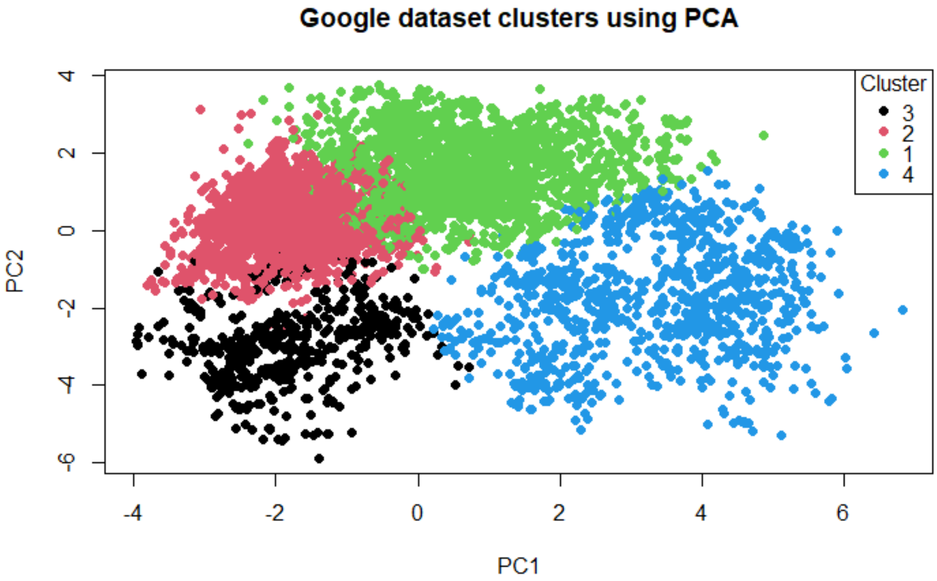
2.5. K-means Clustering

For the final goal of the analysis, which is identifying clusters of traveller's personalities, I've implemented K-mean clustering. It is a technique that allows to group observations into k clusters based on internal similarity within clusters and external dissimilarity between clusters. Distances between observations are measured using the squared Euclidean distance, which should be small within the same cluster and large between different clusters.

To choose an optimal K Clusters, I've decided to analyse inertia (sum of squared distances) and check where the "elbow" is. As we can see below, the optimal is at k=4.



I then run the model and plot the results.



Lastly, I’ve looped through the data and split the points based on their clusters.
Based on the average ratings on the categories within each cluster, I’ve identified the following four traveller’s personalities:

Cluster 1: ***The Cultural Explorer***

This cluster seems to be interested in cultural and recreational activities such as visiting museums, theatres, parks, and monuments. They also seem to prefer natural attractions like beaches and gardens. Negative ratings for malls, restaurants, and pubs/bars indicate a lower preference for commercial and nightlife activities.

Cluster 2: ***The Metropolitan Lover***
This cluster appears to be more inclined towards commercial and dining experiences, with high ratings for restaurants, pubs/bars, and malls. They may also enjoy activities like visiting zoos and swimming pools. Negative ratings for churches, resorts, and beaches suggest a lower interest in cultural and beach-related activities.

Cluster 3: ***The Nightlife Animal***
This cluster shows strong positive ratings for burger/pizza joints, suggesting a preference for fast food. They also seem to enjoy nightlife activities such as pubs/bars and dance clubs.
Negative ratings for churches, resorts, and beaches indicate a lower interest in cultural and beach-related activities similar to Cluster 2.

Cluster 4: ***The Wellness Seeker***
This cluster seems to be interested in fitness and wellness-related activities, with high ratings for gyms, swimming pools, and beauty spas. They may prefer cafes and bakeries for leisure activities.
Negative ratings for cultural attractions like museums and theatres suggest a lower interest in such activities compared to the other clusters.

2.6. Conclusions and Key points:

- Principal Component Analysis (PCA) was employed as the initial unsupervised learning technique for dimensionality reduction and visualization. However, the analysis revealed that the first two principal components explained only a modest portion of the variability in the data. Nonetheless, PCA facilitated the visualization of relationships between variables, aiding in the identification of correlated attraction types.

- To answer my initial question, “*Does a reviewer's affinity for certain attraction types in turn show affinity for others*?” I’ve employed K-means clustering.
  I’ve obtained 4 stable clusters of travellers, identifying 4 different traveller persona based on attraction preferences (average ratings across the categories).
  So to answer the question, yes, affinity for one attraction type is associated with an affinity (or lack thereof) for others.

3. Appendix

This project was realized using R and all the code can be found on my GitHub: https://github.com/MinaBeric