Statistical learning project

Presented by Mina Beric
MSc in Data Science for Economics

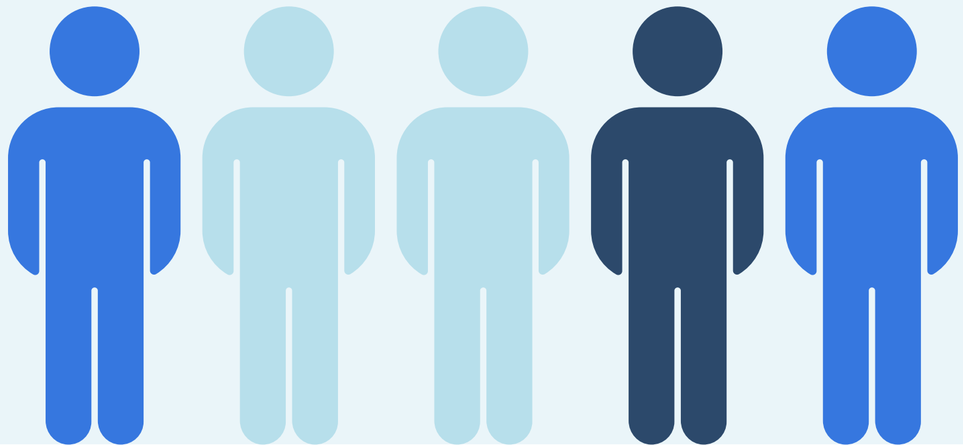
Content

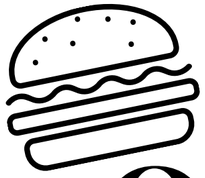
Obesity: An exploration of the different factors

Supervised project

Unveiling Tourist Personas

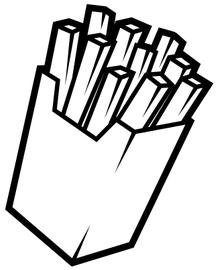
Unsupervised project





Obesity

An exploration of the influential factors



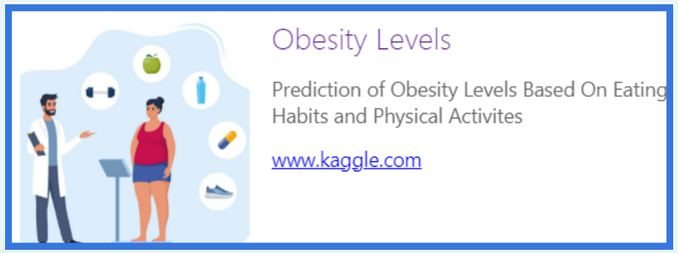
Index

The goal

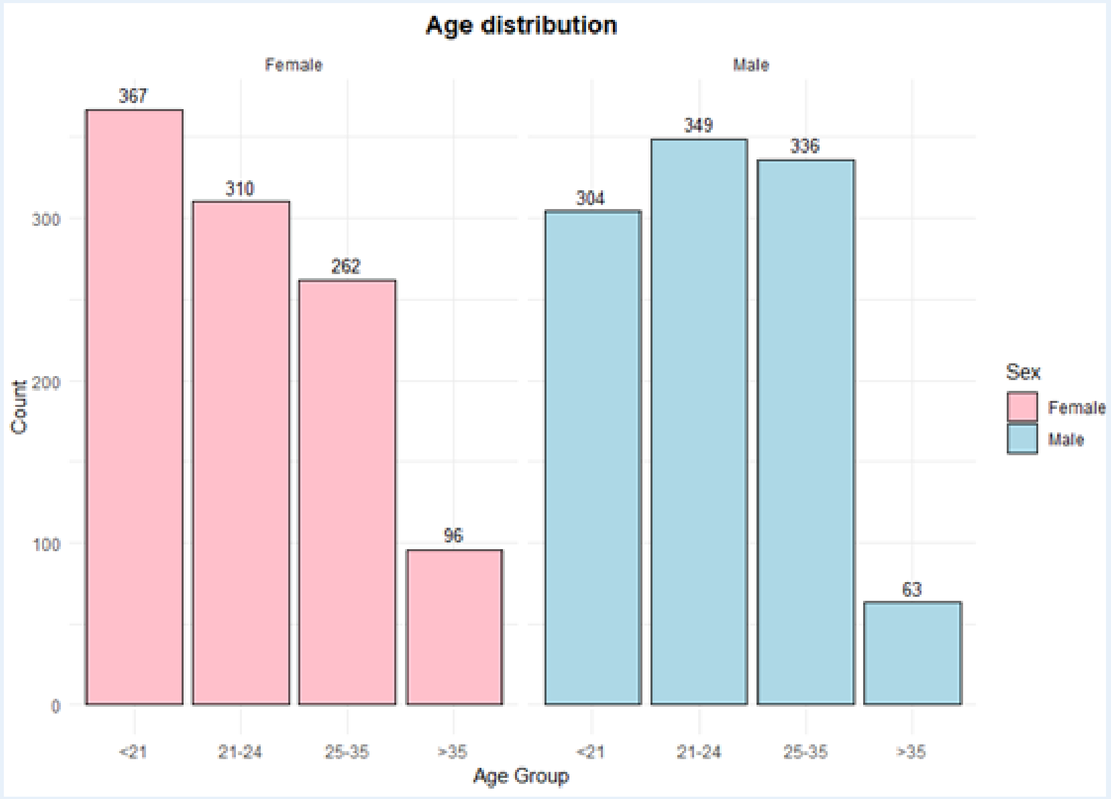
Exploring the Impact of Dietary Choices and Lifestyle Behaviors on Obesity in Individuals

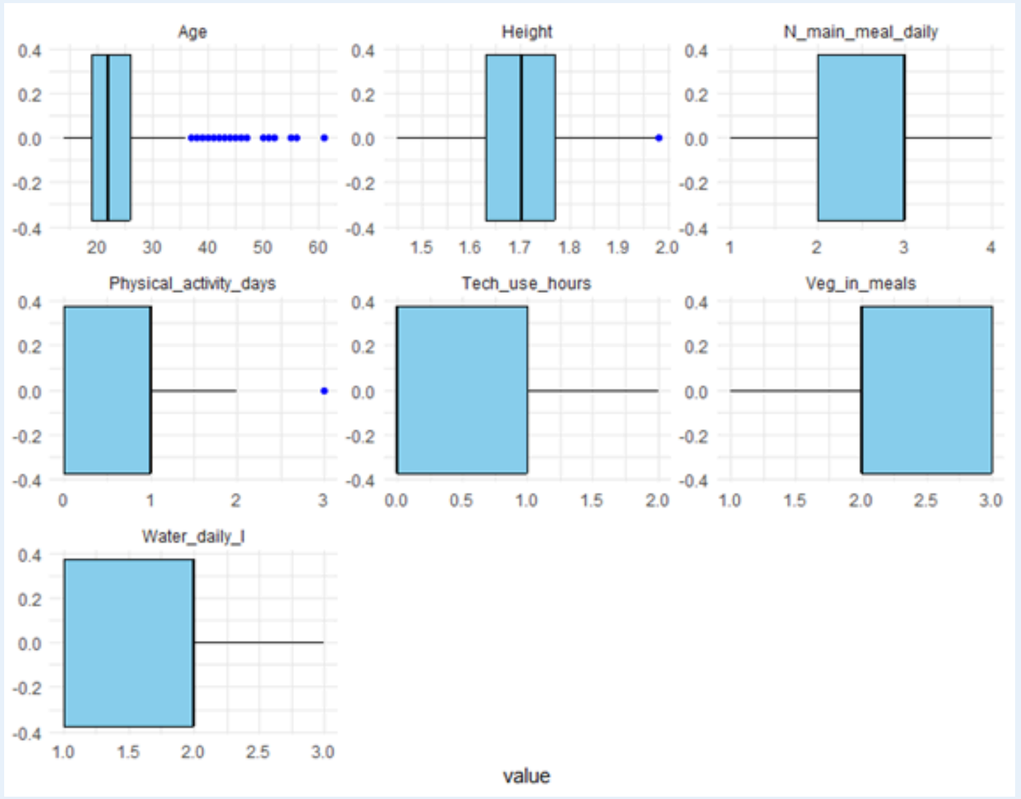
The Dataset

- Originally, the respondents were 458 individuals from Mexico, Peru, and Colombia, but later expanded using an oversampling technique, resulting in **77% fictitious data points**.
- It is formed by **2111 individuals**, responding to **17 questions (variables)** covering various aspects such as gender, age, dietary habits and lifestyle choices with responses that range from simple “Yes” or “No”, to multiple options (both categorical and numerical).
- The target variable "Obesity_level" was **transformed into a binary** variable **"ObeseYes" or "ObeseNo"** for the goal of the analysis.
- Categorical variables were converted into factors, while numerical ones were converted into integers
- The dataset had **no null values** but contained 24 duplicates, which were removed to ensure data integrity and analysis robustness.
- Lastly, before implementing the models, **dataset was normalized** because the ranges of the dataset features are not the 10 same.

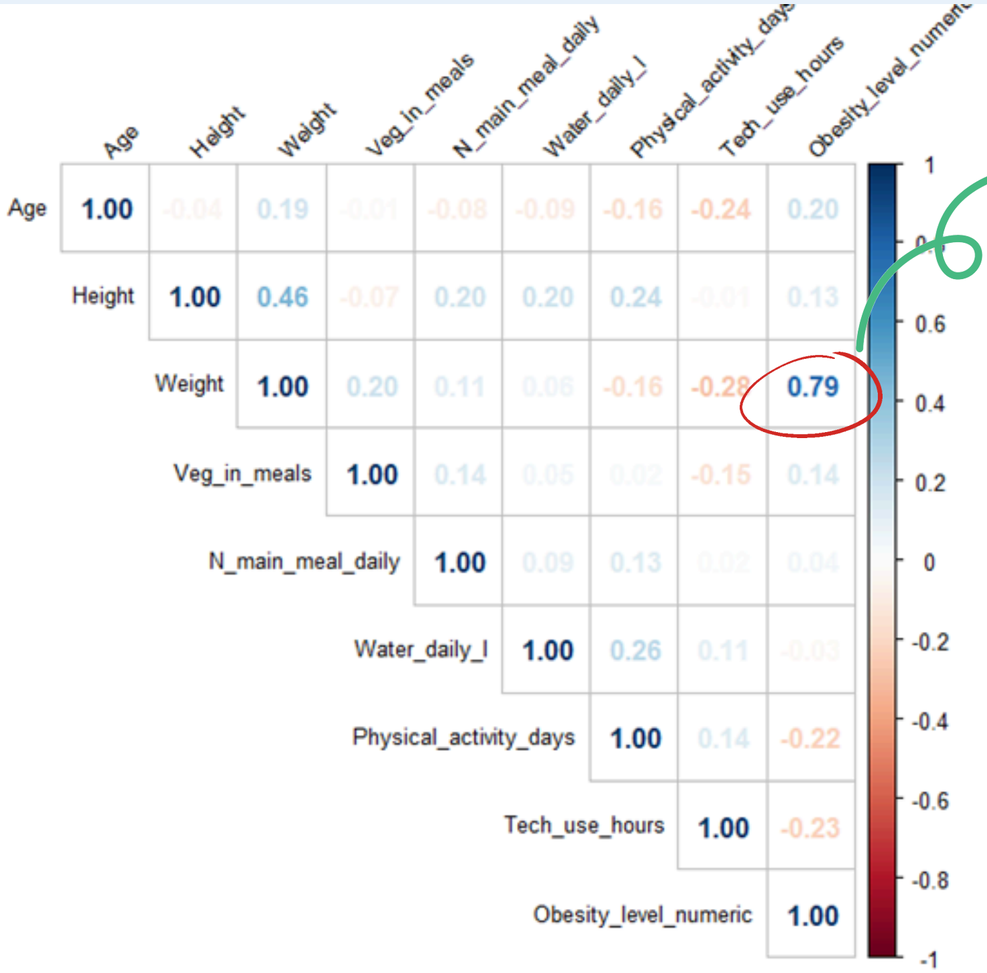


EDA


Age distribution
Female
Male
367
310
262
96
304
349
336
63
Count
0
100
200
300
<21
21-24
25-35
>35
Age Group
Sex
Female
Male


Age
Height
N_main_meal_daily
Physical_activity_days
Tech_use_hours
Veg_in_meals
Water_daily_l
value

EDA



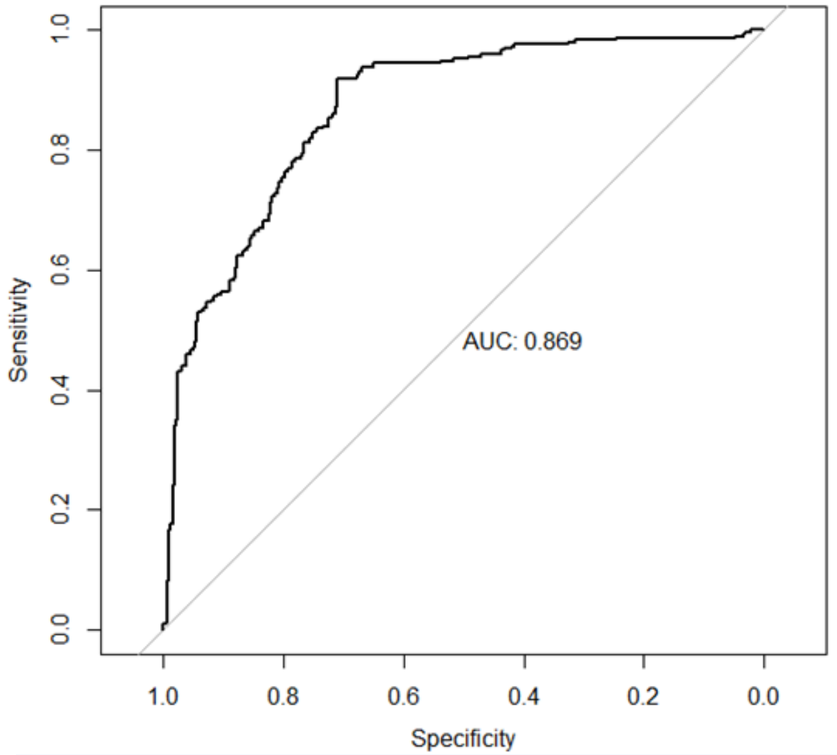
Dropped from the dataset

Logistic regression

```
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
```

Logistic regression

```
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

                  Kappa : 0.563

 Mcnemar's Test P-Value : 0.2282

            Sensitivity : 0.7952
            Specificity : 0.7692
         Pos Pred Value : 0.7565
         Neg Pred Value : 0.8065
             Prevalence : 0.4741
         Detection Rate : 0.3770
   Detection Prevalence : 0.4984
      Balanced Accuracy : 0.7822
```

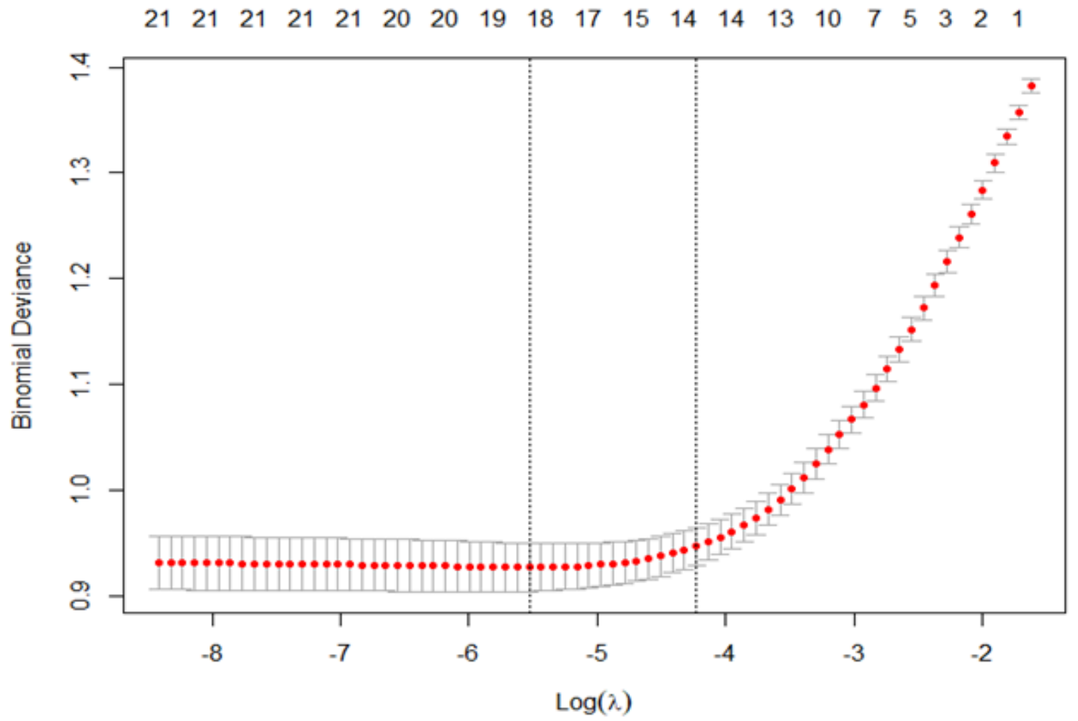


78.16% Accuracy reached

Decent power of discrimination

Lasso Model

Tuning Lambda




```
> print(cv.lasso$lambda.min)
[1] 0.002730536
```

Lasso Model

(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

4 variables shrunk to zero

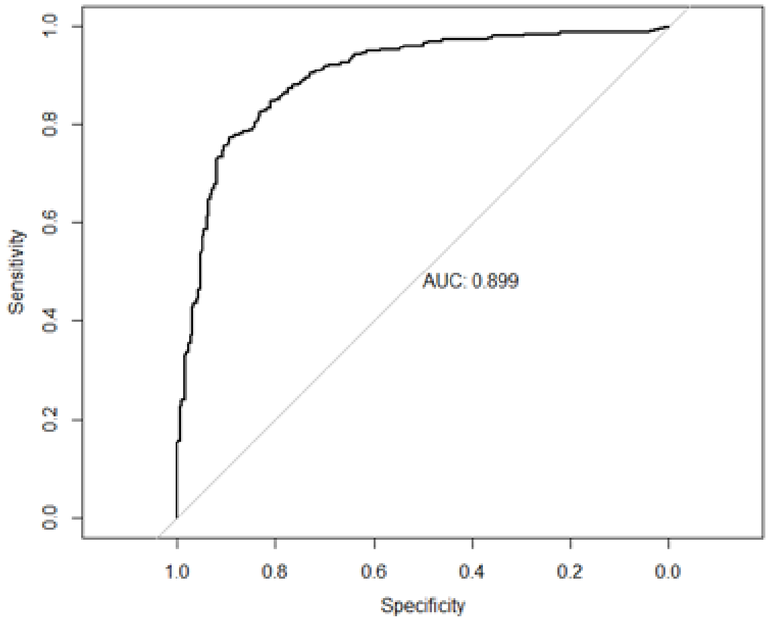
```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 247  54
         1  78 239

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

                  Kappa : 0.5734

 Mcnemar's Test P-Value : 0.0453

            Sensitivity : 0.8157
            Specificity : 0.7600
         Pos Pred Value : 0.7539
         Neg Pred Value : 0.8206
             Prevalence : 0.4741
         Detection Rate : 0.3867
   Detection Prevalence : 0.5129
      Balanced Accuracy : 0.7878

       'Positive' Class : 1
```

78.64% Accuracy reached

Generalized additive model

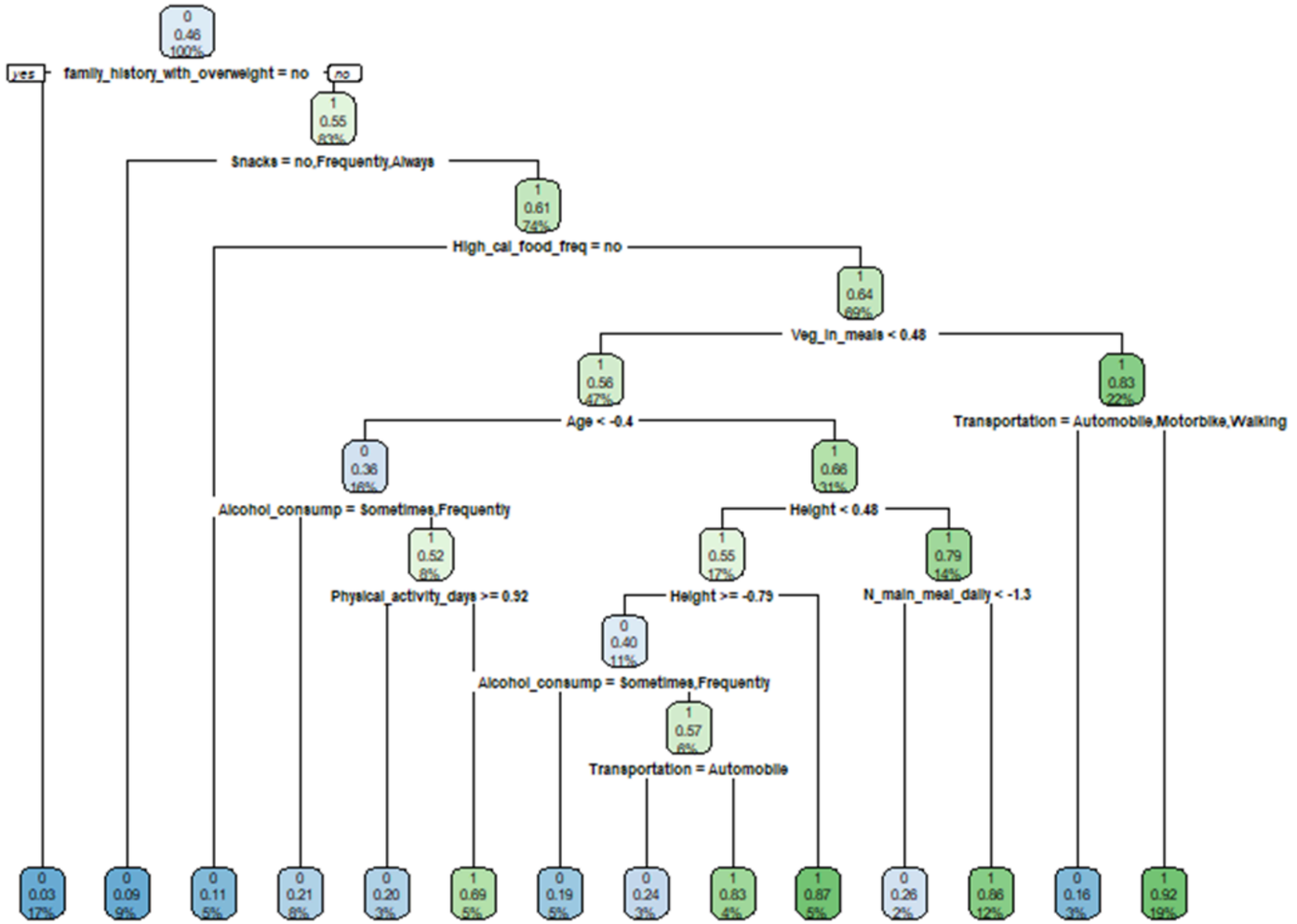
```
Confusion Matrix and Statistics

          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
```

82.36% Accuracy reached



But GAM coefficients are more challenging to interprete due to the inclusion of smoothed terms.

Decision Tree

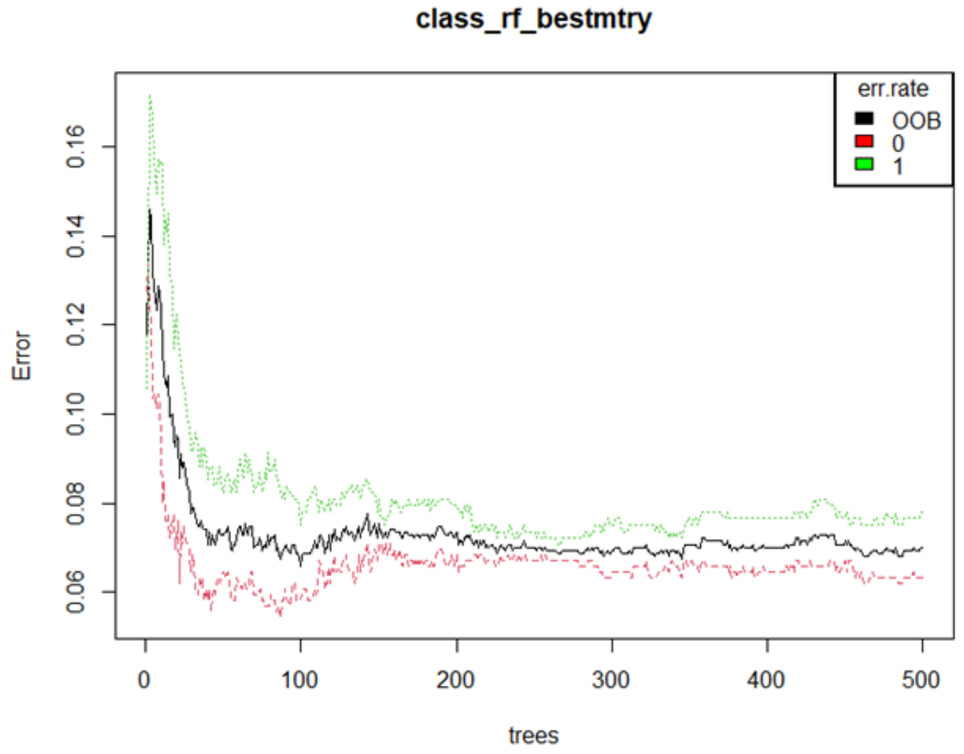


Decision Tree

```
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
```

86.08% Accuracy reached

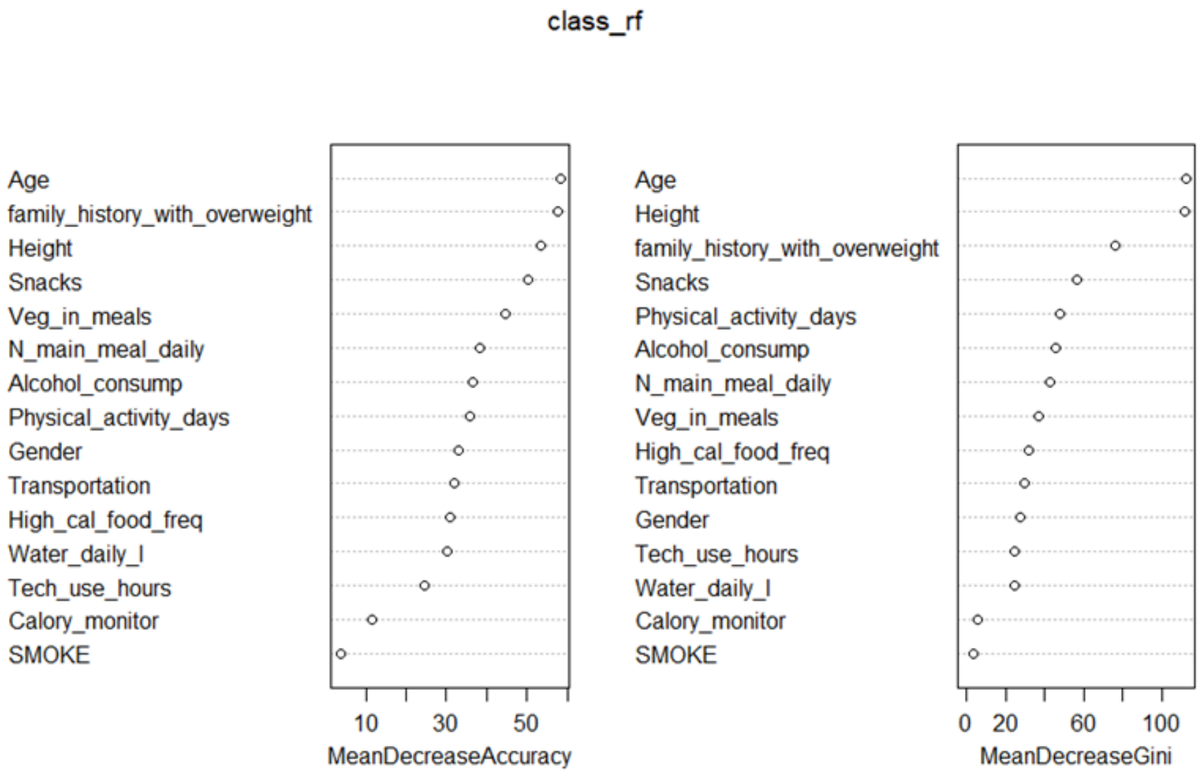
Random Forest




```
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

                  Kappa : 0.9025

 Mcnemar's Test P-Value : 0.2012

            Sensitivity : 0.9352
            Specificity : 0.9662
         Pos Pred Value : 0.9614
         Neg Pred Value : 0.9429
             Prevalence : 0.4741
         Detection Rate : 0.4434
   Detection Prevalence : 0.4612
      Balanced Accuracy : 0.9507

       'Positive' Class : 1
```

95.15% Accuracy reached

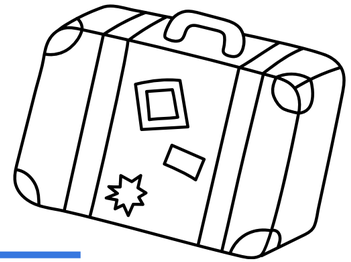
Random Forest



The **Mean Decrease Accuracy** plot illustrates the extent of accuracy reduction achieved by removing each variable from the model.
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.

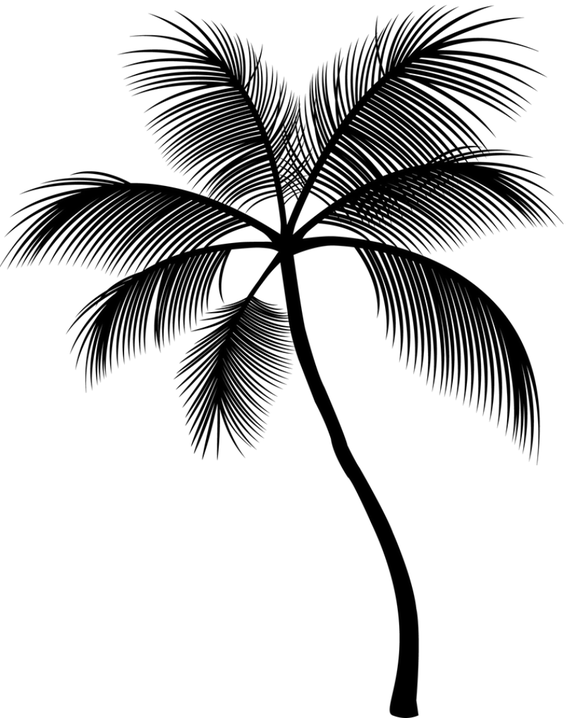
Conclusions

- 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).



Unveiling

Tourist Personas



Index

Does a reviewer's affinity for certain attraction types in turn show affinity for others?

The goal

Identify clusters of traveller’s personalities, with similar preferences and characteristics.

The Dataset

- The original dataset was sourced from UC Irvine Machine Learning Repository, comprises **5456 rows** and **26 columns**.
- Each variable corresponds to a **specific category of attraction**, ranging from landmarks like monuments and gardens to amenities like gyms and beauty spas and with **values of rating ranging from 1 to 5**
- The dataset underwent preprocessing where the first and last variables were dropped, resulting in 24 remaining columns.
- Among the 24 variables, most are of type "dbl" (numeric), except for "local_services," which was converted from character to numerical.
- Column names were renamed for interpretability, while **NA values** and **duplicate entries** were identified and removed to ensure data integrity.

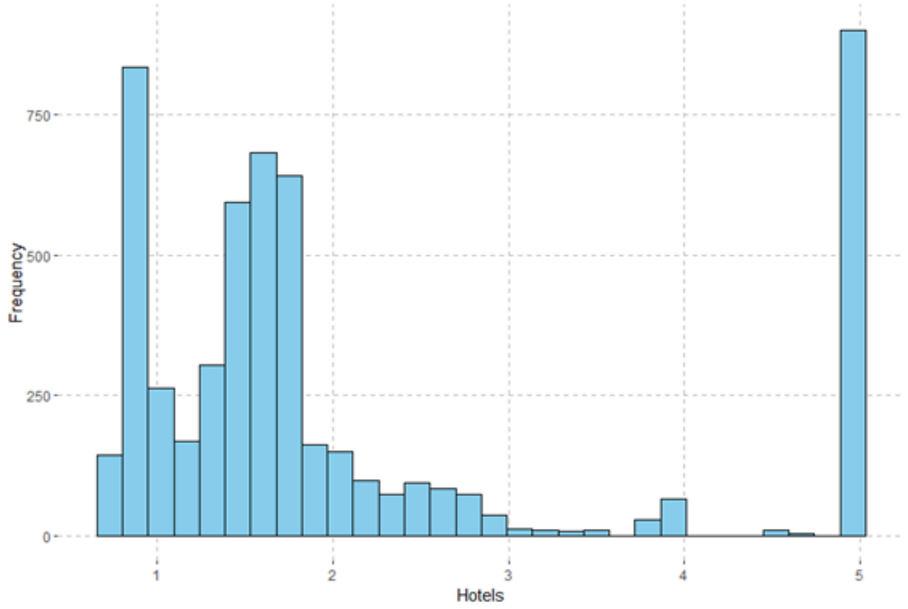
Summary statistics

```
    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  
```

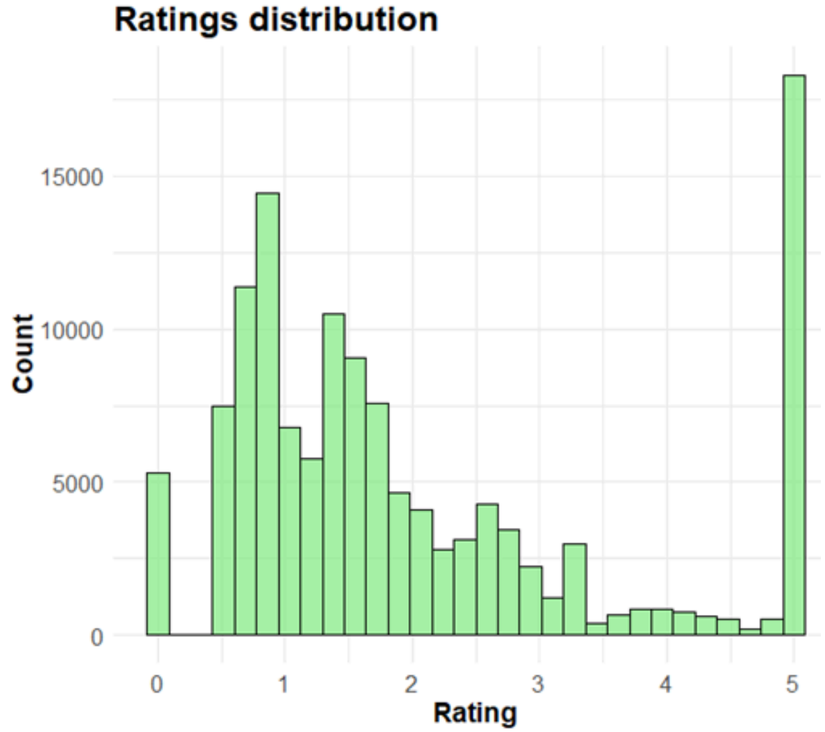
EDA

So, Outliers?

All the 24 variables follow this kind of distribution, with higher frequencies when ratings on average are close to 0 or close to 5



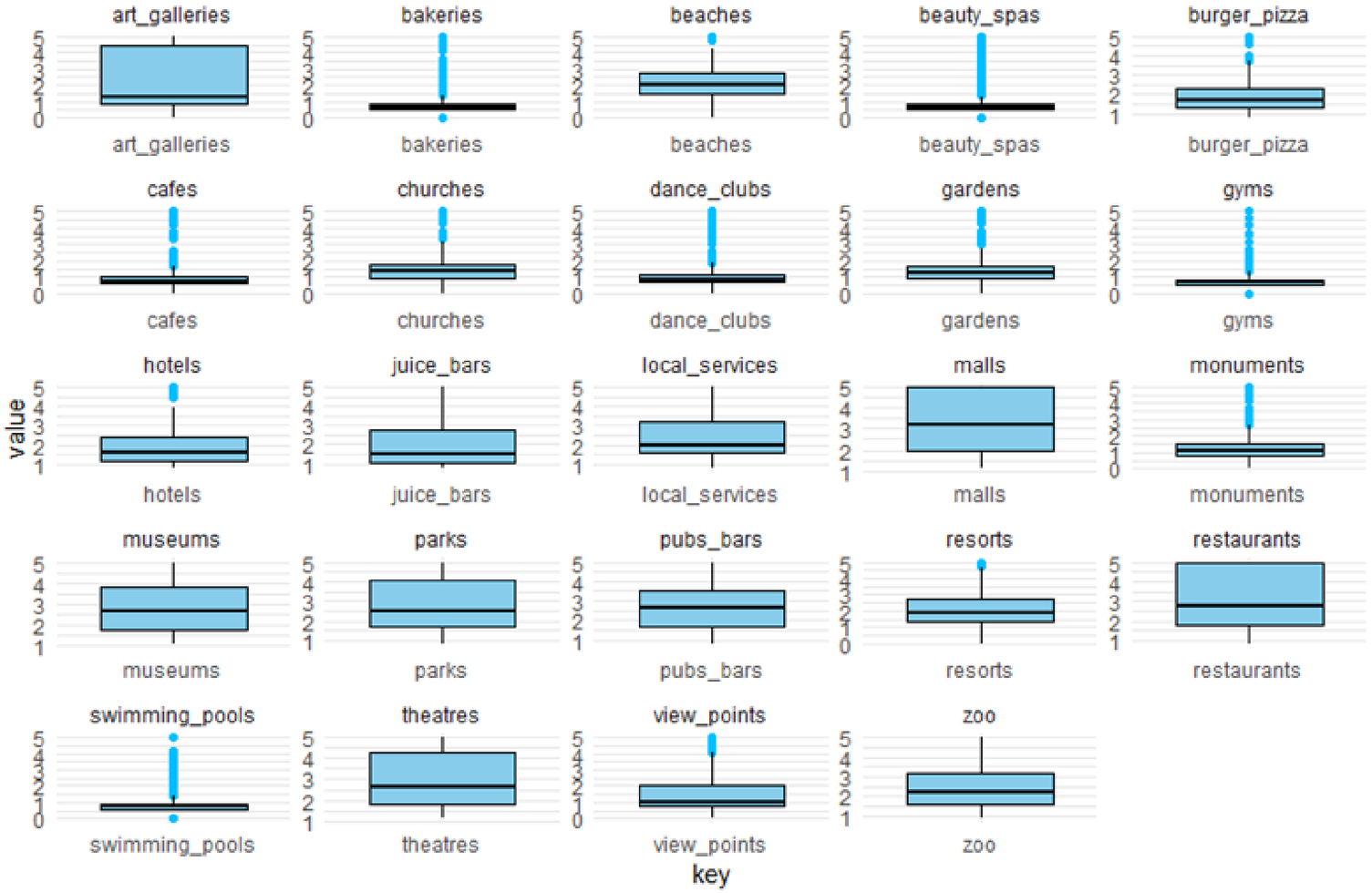
Same behaviour



Overall trend in people's rating choices, indipendently of the specific object of rating

EDA

Yes...



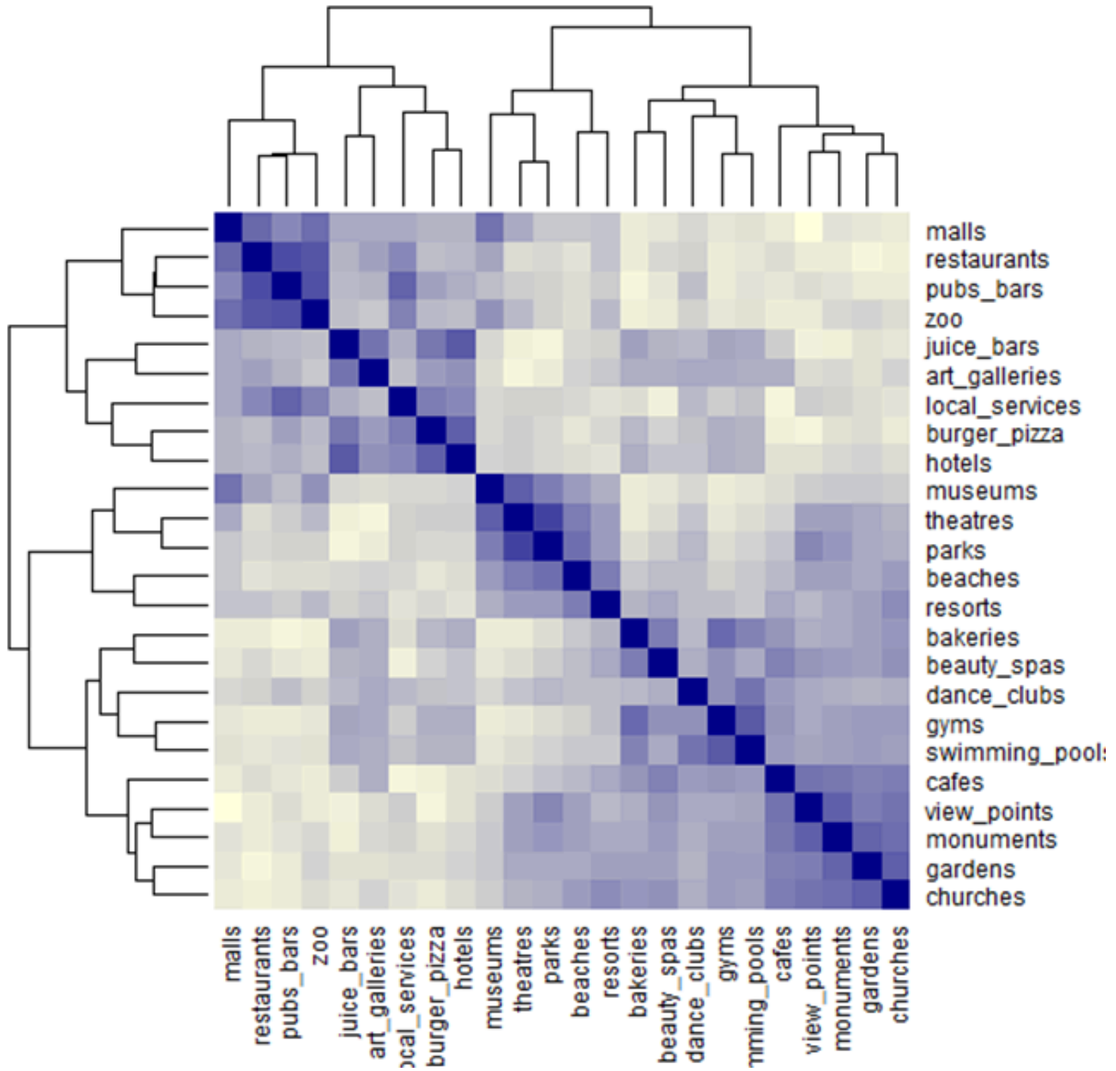
But

This checks with reality (people rate more when they have really bad or really good experiences, So I've delt with this by scaling)

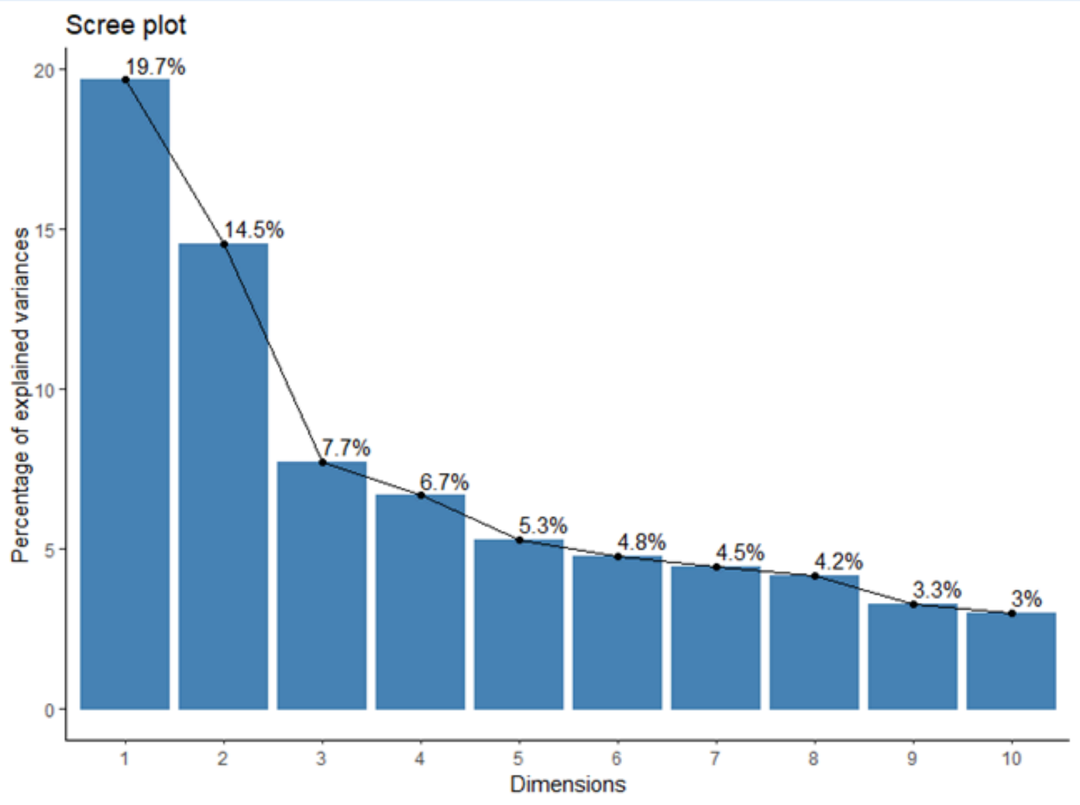
EDA

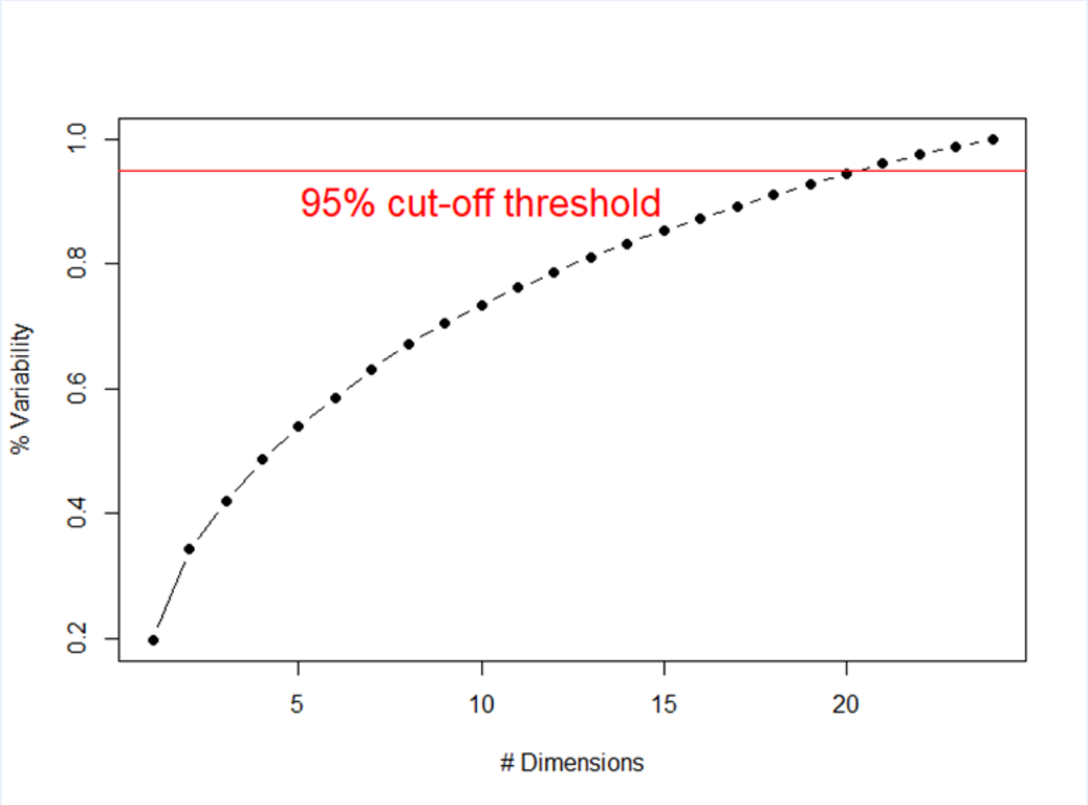
Presence of strong correlations indicating a potentially high degree of association between these attraction types, suggesting that they might cater to similar visitor preferences.

Cluster map conferms that data is “divided” into groups (unfortunately too fragmented here).



PCA

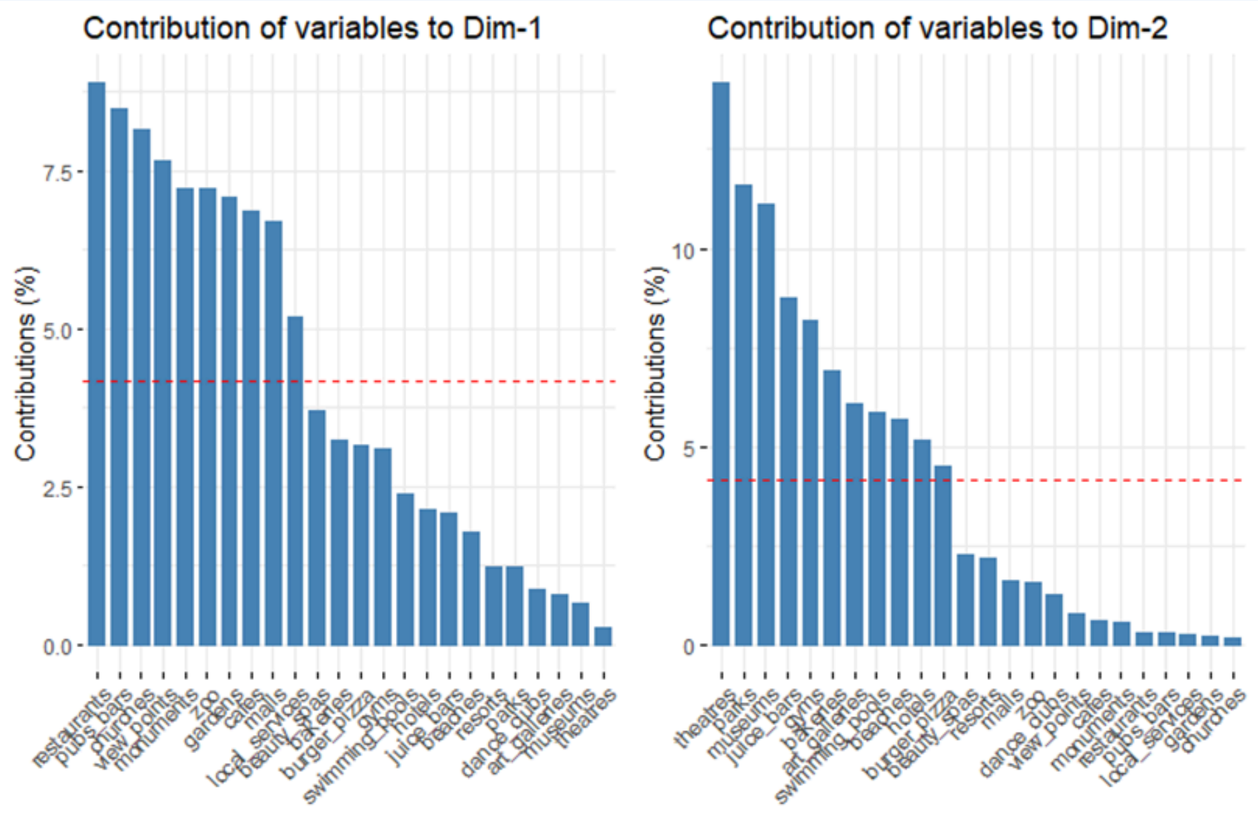


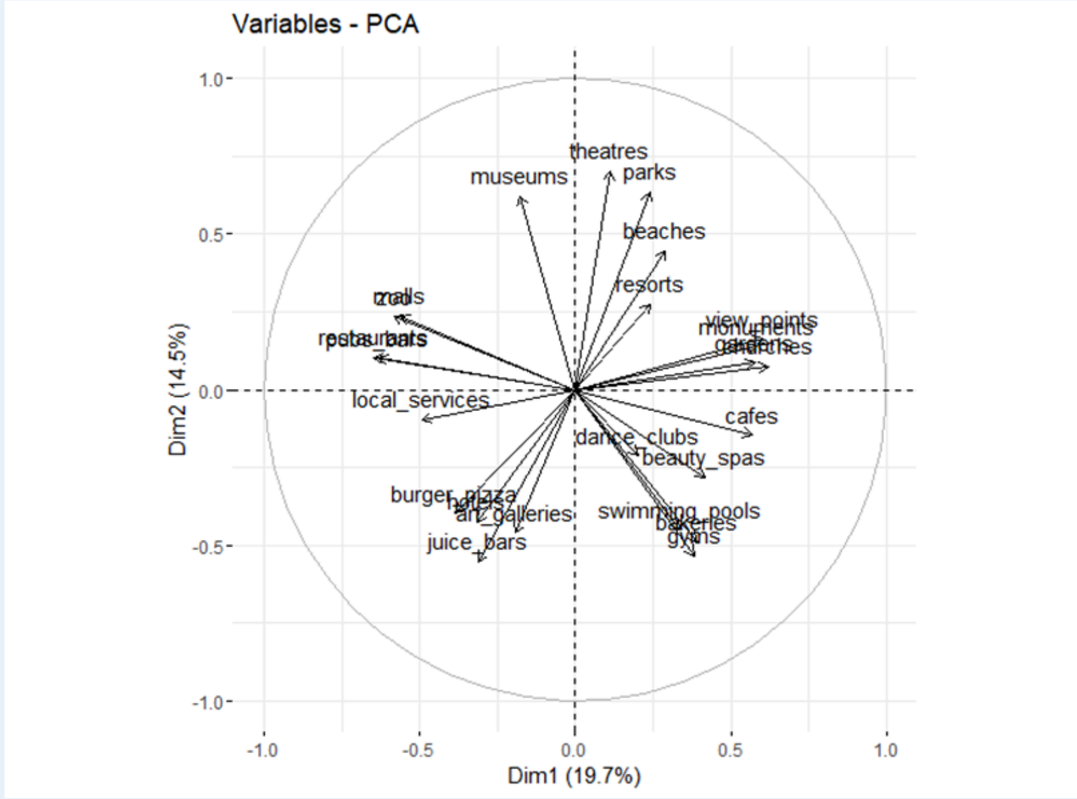


The first 2 dimensions explain **only 34.2%** of the variability

not the dimensionality reduction that I've hoped for.

PCA

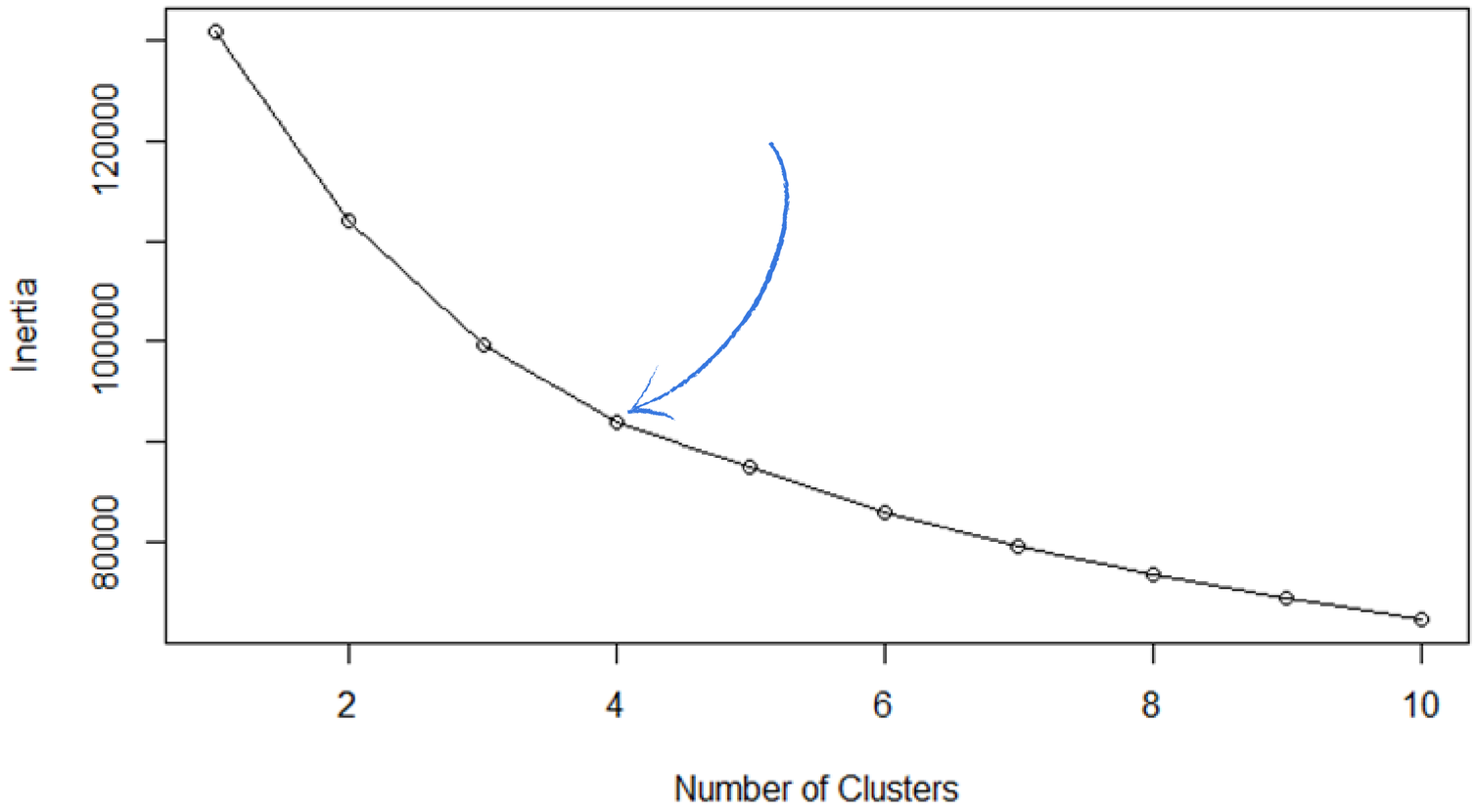




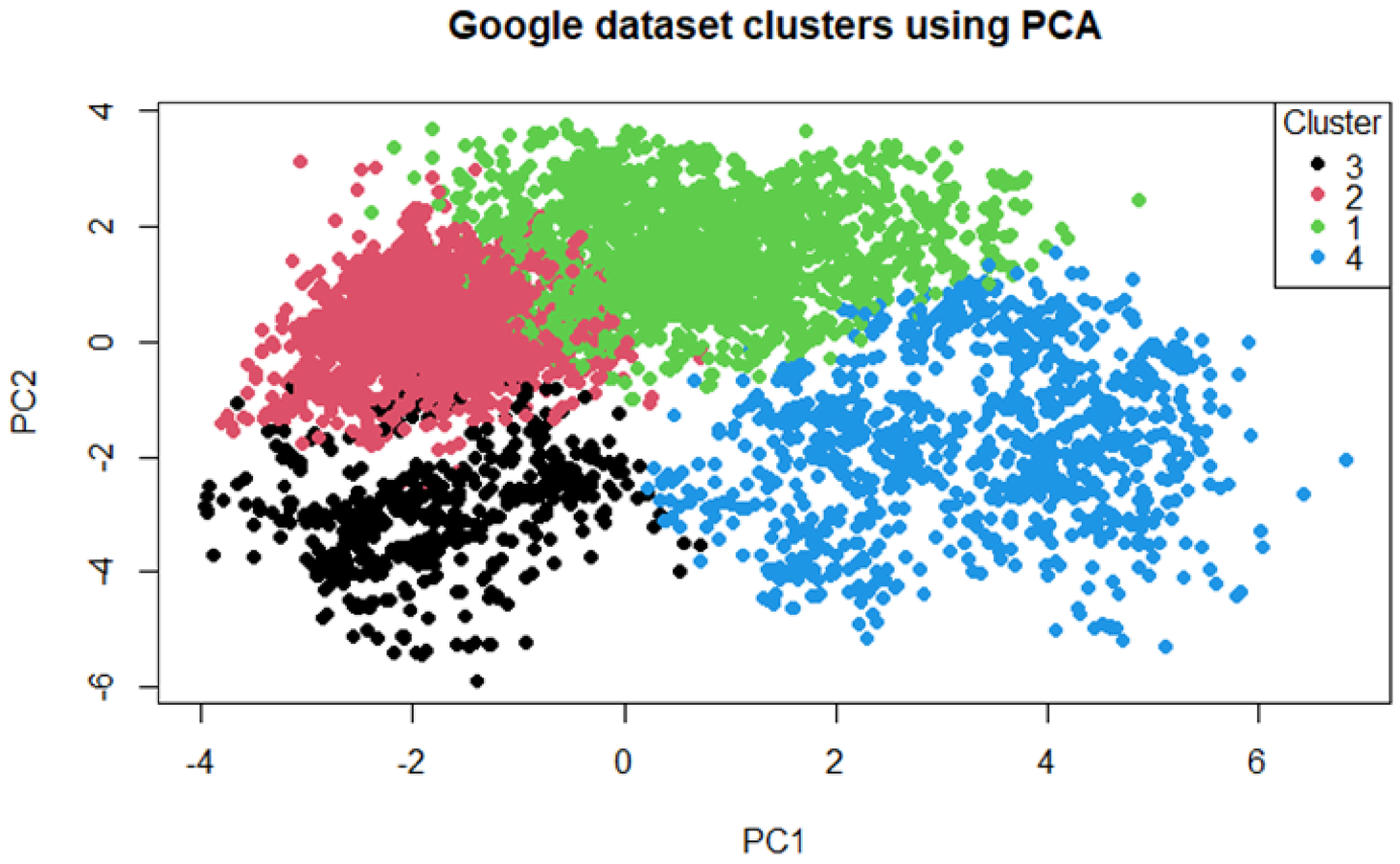
Interesting how bakeries and gyms are so close!

K-Means Clustering

Elbow method to find optimal K

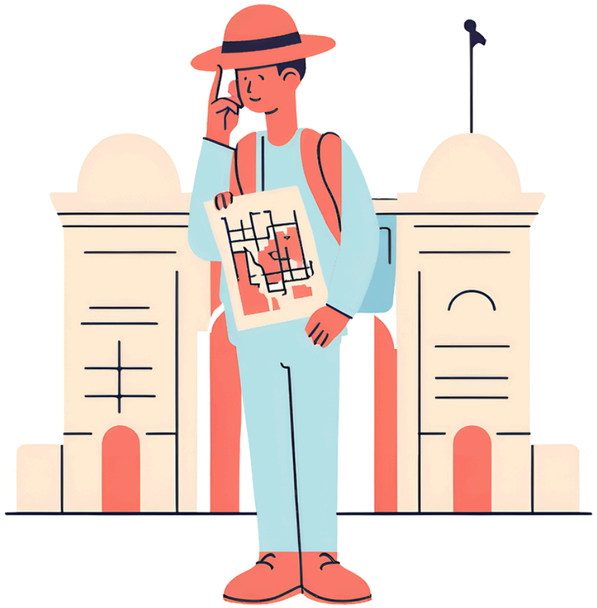


K-Means Clustering



Traveler’s Personas

Cluster 1



The Cultural Explorer

Cluster 2

The Metropolitan Lover

Traveler’s Personas

Cluster 3

The Nightlife Animal

Cluster 4

The Wellness Seeker

Conclusions

- PCA was utilized initially for dimensionality reduction and visualization, revealing that the first two principal components explained only a moderate portion of the data's variability.
- 4 stable clusters of travelers were identified, each representing distinct traveler personas based on their attraction preferences, as reflected in their average ratings across categories.
- *Does a reviewer's affinity for certain attraction types in turn show affinity for others?* The results confirmed that preferences for certain attraction types are indeed associated with preferences (or disinterest) for others

Thanks for the attention