Project Presentation

Data Management

Florian Deroua and Eve Schmitz Schair

TABLE OF CONTENTS

01 INTRODUCTION

02 WEBSCRAPING

03 VISUALIZATION

04 MACHINE LEARNING

05 CONCLUSION

01
INTRODUCTION

Motivation

Question: can we predict the expected price for adult tickets for attractions in a specific city based on the type of attraction and historical pricing data ?

To respond to the question, we used a website with all the attractions of Wallonia and Brussel.

02

WEBSCRAPING

DATA

Name	Type	City	Price adult	Price child
La boverie	Art	Liège	5€	0€

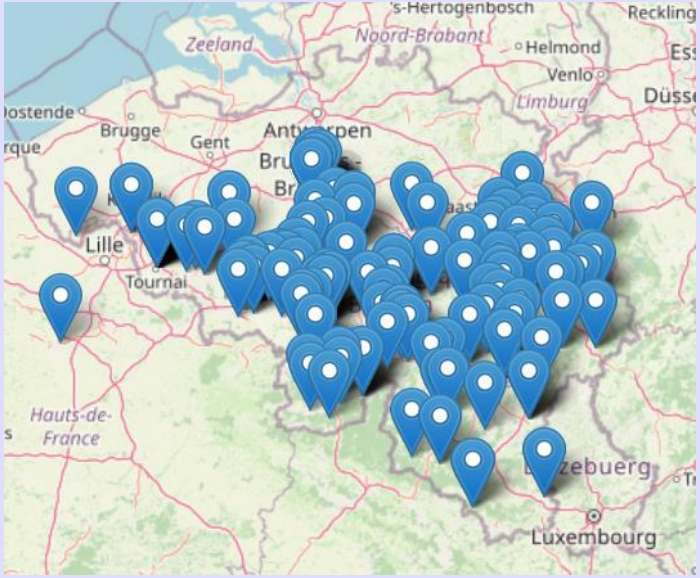
Columns: 5
Lines: 289

We made a list that stores the information about attractions extracted from the website. Each code is for a different type of attraction. Each attraction's information is stored as a dictionary containing the name of the attraction, the type of attraction, the city where the attraction is and the price for each group of persons.

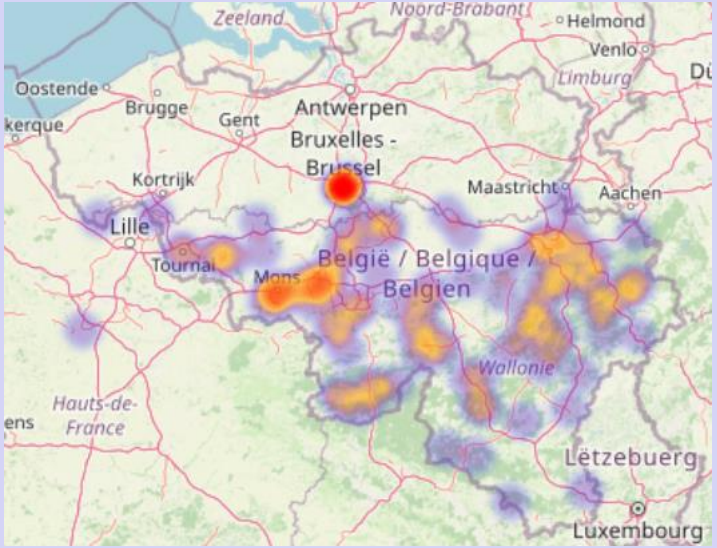
03

VISUALIZATION

MAP



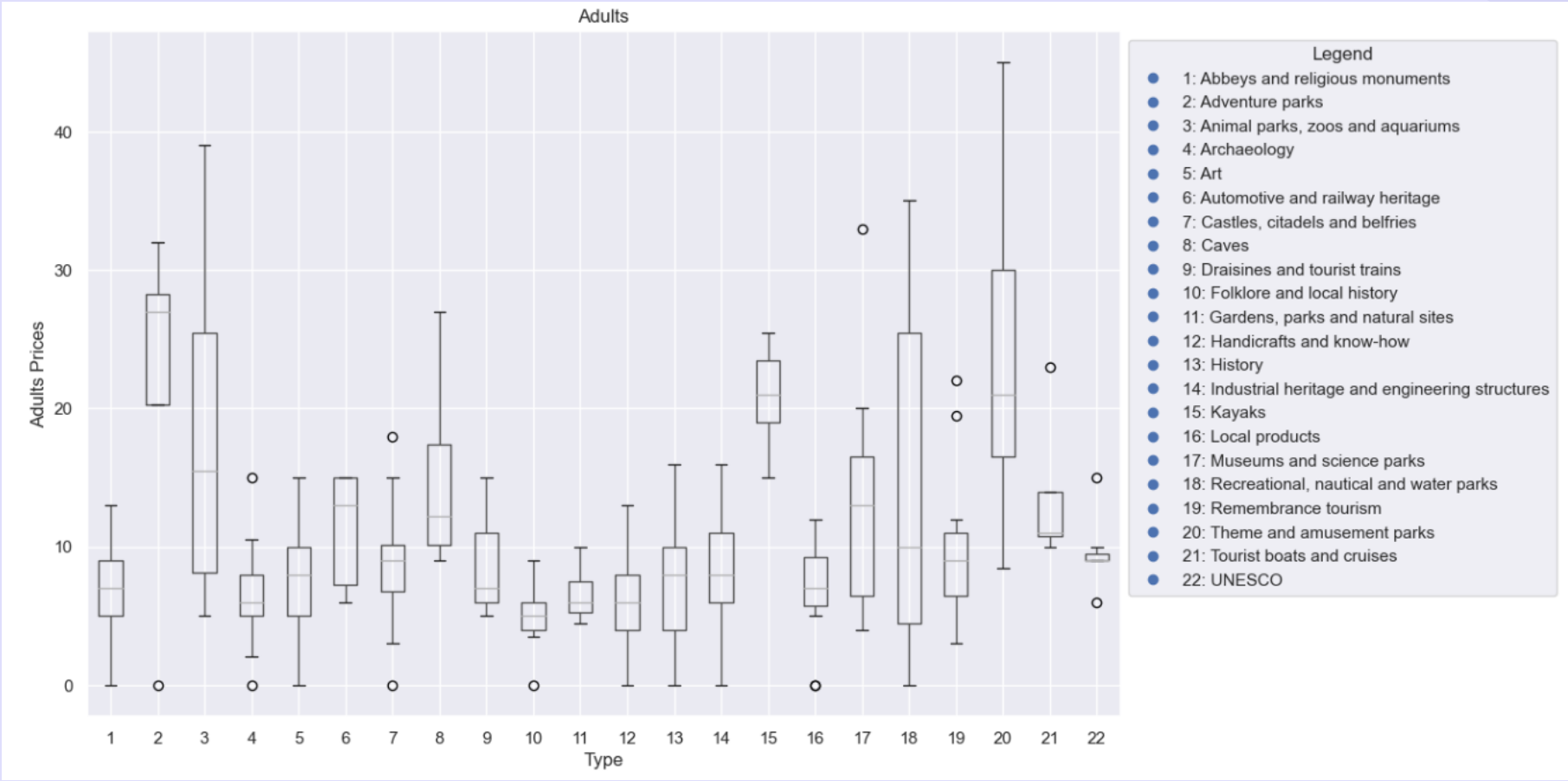
We observe also that the highest density of attractions is in Brussels. We can also see that in the province of Luxembourg, there are fewer attractions than in other provinces. We could conclude that attractions are often near a big city than in the middle of small villages.



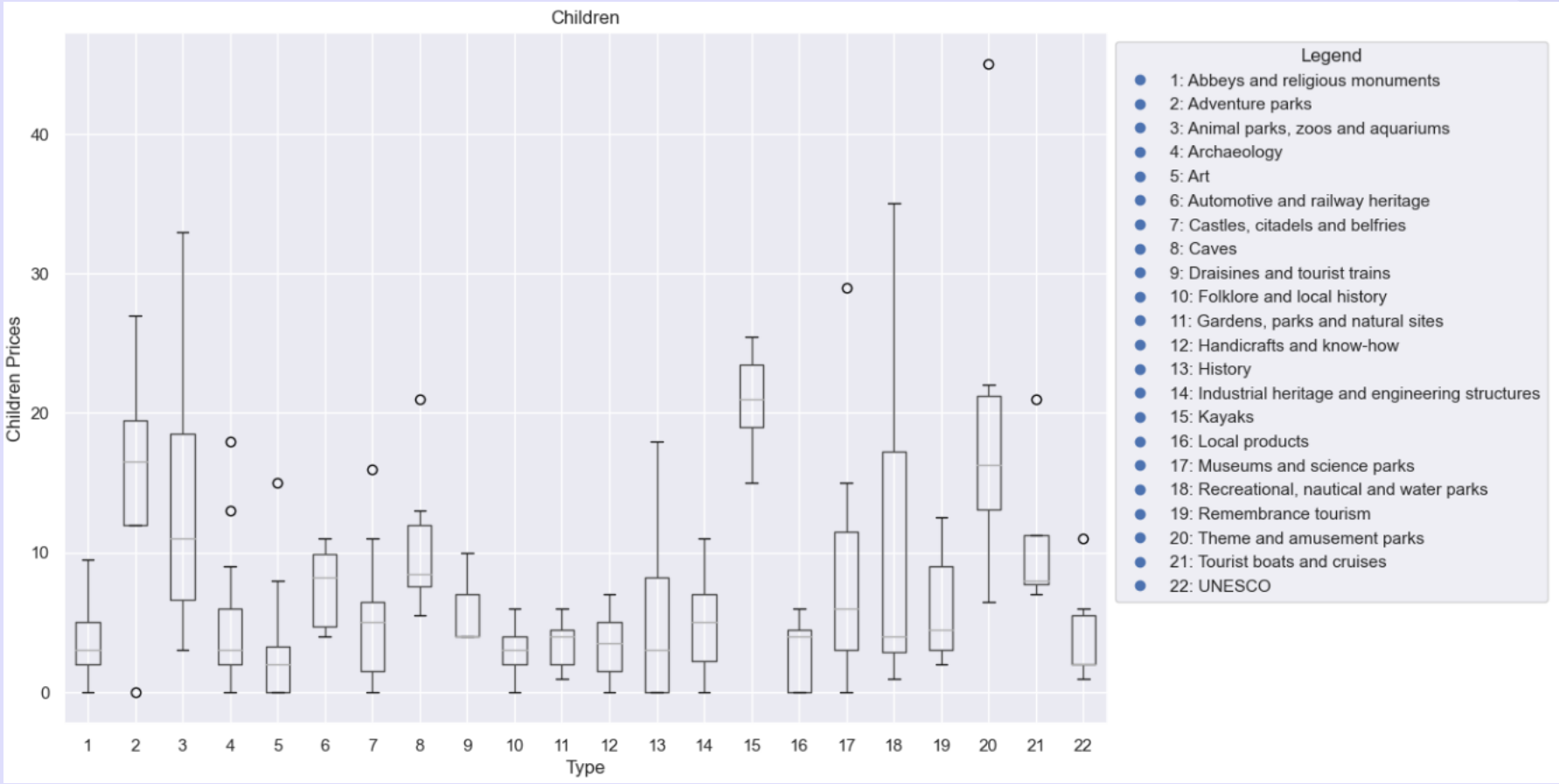
BAR CHART

We can observe that the prices for children are very often lower than the prices of adults.

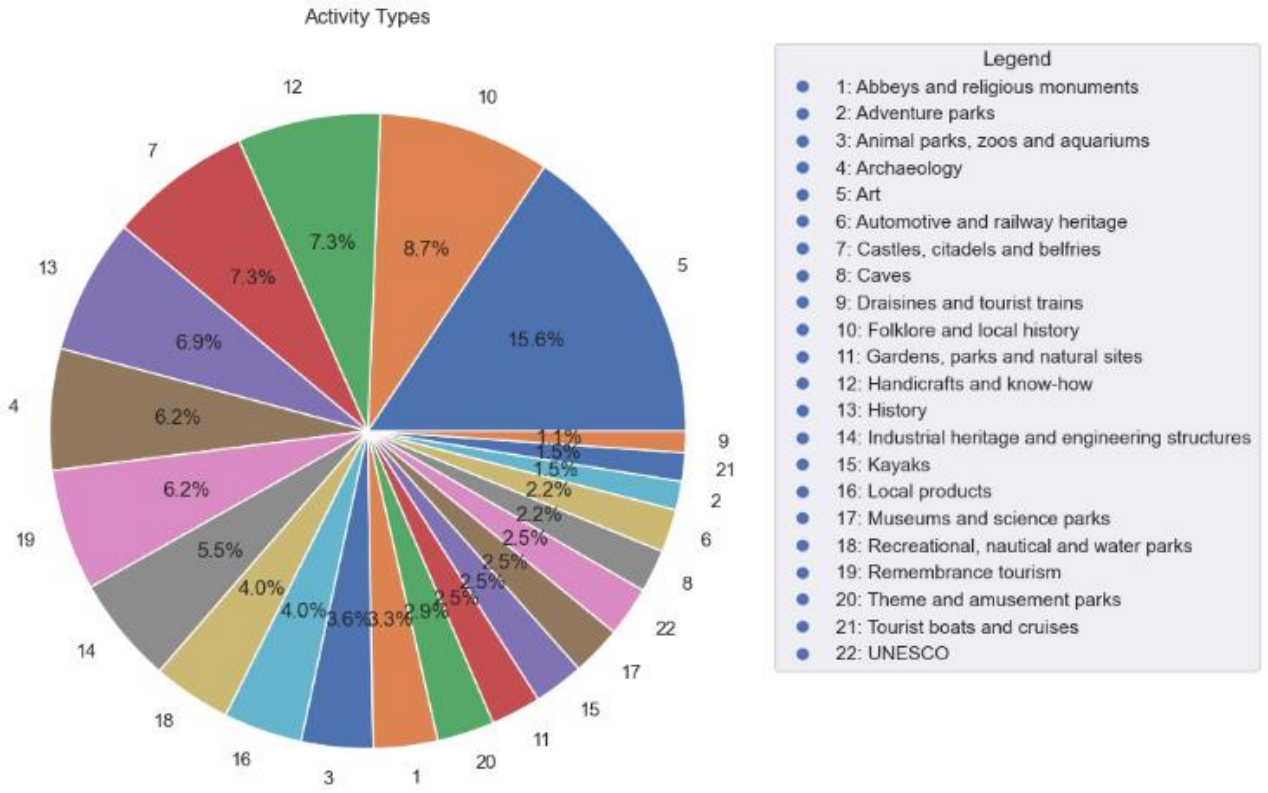
BOXPLOT GROUPED BY TYPE



BOXPLOT GROUPED BY TYPE



PIE CHART


Activity Types
12
10
7
5
13
4
9
21
2
6
19
8
22
14
17
18
15
16
11
3
1
20
7.3%
8.7%
7.3%
15.6%
6.9%
6.2%
1.1%
1.5%
1.5%
2.2%
6.2%
2.2%
2.5%
2.5%
5.5%
2.5%
2.5%
4.0%
4.0%
3.6%
3.3%
2.9%
Legend
1: Abbeys and religious monuments
2: Adventure parks
3: Animal parks, zoos and aquariums
4: Archaeology
5: Art
6: Automotive and railway heritage
7: Castles, citadels and belfries
8: Caves
9: Draisines and tourist trains
10: Folklore and local history
11: Gardens, parks and natural sites
12: Handicrafts and know-how
13: History
14: Industrial heritage and engineering structures
15: Kayaks
16: Local products
17: Museums and science parks
18: Recreational, nautical and water parks
19: Remembrance tourism
20: Theme and amusement parks
21: Tourist boats and cruises
22: UNESCO

04

MACHINE LEARNING

LINEAR REGRESSION

R-squared: 0.338
Test RMSE: 7.148
Train RMSE: 2.617

MODEL REGRESSION 1

$$\text{Adult Price} = \beta_0 + \beta_1 * \text{Activity Type} + \beta_2 * \text{City} + \varepsilon$$

Where:

- Adult Price represents the dependent variable, which is the price for adults.
- Activity Type and City are the independent variables.
- β_0, β_1, and β_2 are the coefficients or parameters that need to be estimated.
- ε represents the error term, which accounts for unexplained variation in the dependent variable.

LINEAR REGRESSION

R-squared: 0.749
Test RMSE: 4.401
Train RMSE: 1.918

MODEL REGRESSION 2

$$\text{Adult Price} = \beta_0 + \beta_1 * \text{Activity Type} + \beta_2 * \text{City} + \beta_3 * \text{Children Price} + \varepsilon$$

Where:

- Adult Price represents the dependent variable, which is the price for adults.
- Activity Type, City, and Children price are the independent variables.
- β_0, β_1, β_2 and β_3 are the coefficients or parameters that need to be estimated.
- ε represents the error term, which accounts for unexplained variation in the dependent variable.

DECISION TREE REGRESSION

R-squared: 0.789
Test RMSE: 4.034
Train RMSE: 0.783

MODEL REGRESSION 3

Adult Price = f(Activity Type, City, Children Price) + ε

f represents the decision tree regression model, which learns how to make predictions based on the values of the independent variables. The ε term represents the error or residual, which captures the unexplained variation in the dependent variable.

In summary, the second model outperforms the first model in terms of train RMSE, test RMSE, and R-squared. Therefore, the second model is likely to be a better choice for predicting the target variable based on the given results. However, we could try in further research to add more data to the dataframe or to test other models to have better predictions.

05 CONCLUSION

THANK YOU

Questions ?