Analyzing Paris Tree Data

Presented by: Octave Antoni

Last Updated: September 12th, 2022

Table of contents

Introduction

- Project based on an open-source dataset of Paris trees (downloadable at opendata.paris.fr)
- Analyze the data ➔ Univariate and bivariate analysis
- **Main objective** ➔ Help optimize the city tree maintenance
- Analysis of tree data, **highlighting key figures**
- Bivariate analyses ➔ Measuring correlation
- Creating tree clusters ➔ Maintenance areas

Dataset

Dataset

- Dataset = 1 csv table, several columns :
  - Domain
  - Arrondissement (district)
  - Address
  - Tree name
  - Tree type
  - Tree species
  - Tree variety
  - Circumference (cm)
  - Height (m)
  - Development phase
  - Outstanding / Not outstanding (Boolean)
  - Coordinates (latitude/longitude)

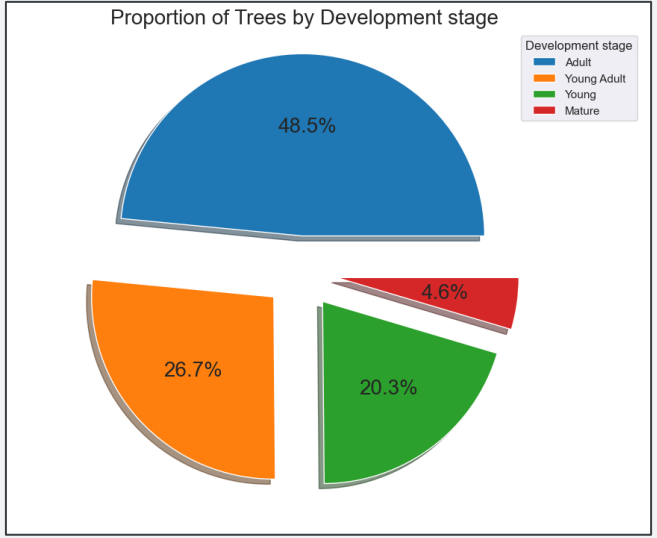
Data cleaning

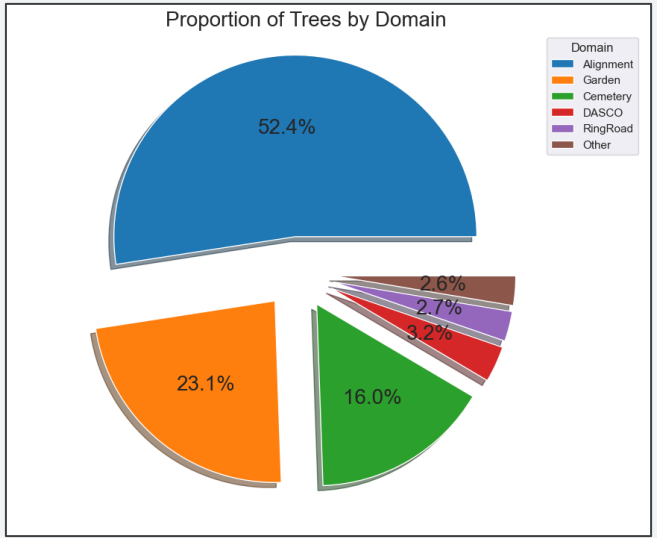
- Removing empty / single value columns
- Removing illogical values for circumference and height (based on maximal values in France)
- Renaming coordinate columns into latitude/longitude
- Data imputation (median) for NaN numeric values
- Analysis of categorical variables and replacement of mistypes based on calculation of Levenshtein distance

Key figures

Repartition of trees

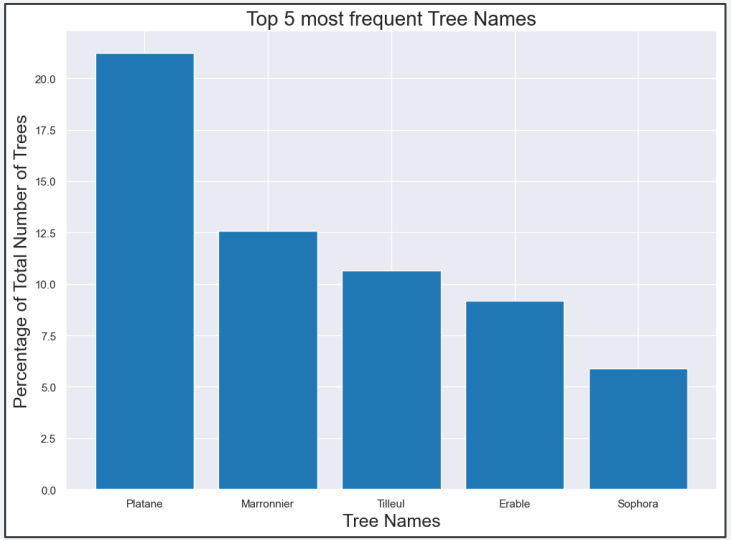
- **48.5%** of trees **are classified as Adult**, with **less than 5% Mature** trees
- The domain of most referenced trees is **Alignment** (52,4%)
- **Garden and Cemetary** represent respectively 23% and 16% of the dataset
- Only **0.1%** of referenced trees have been categorized as **Outstanding**

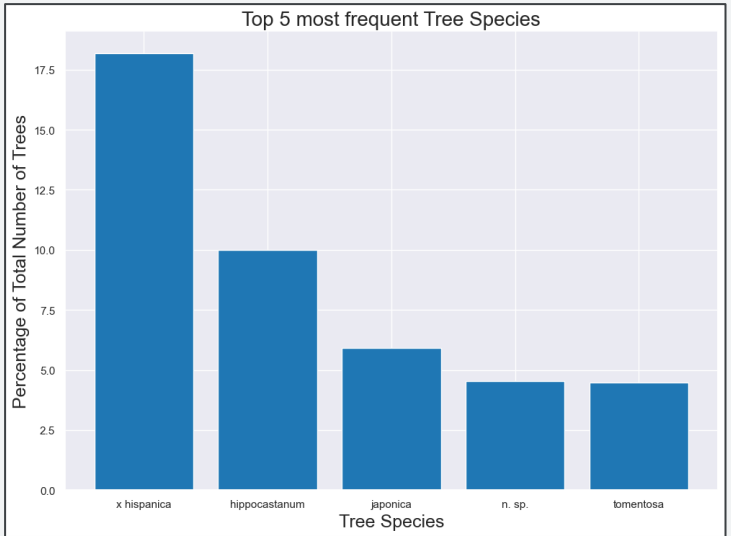




Most frequent tree names, species and variety

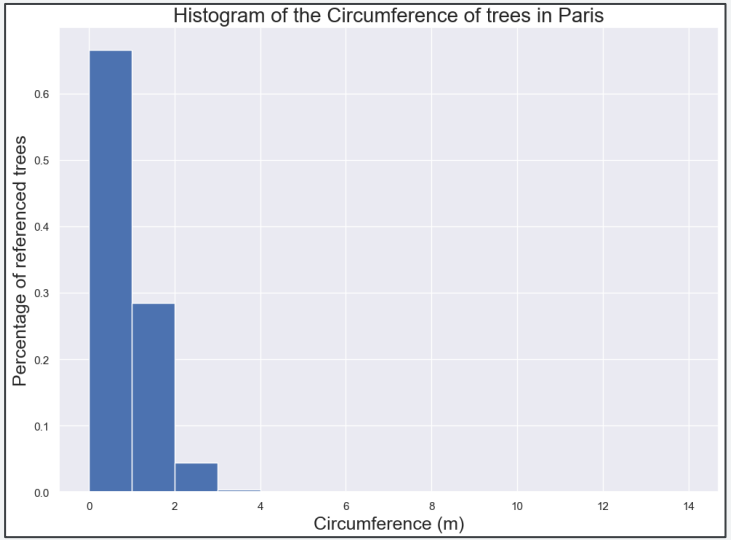
- **More than 20**% of the studied trees are **Plane trees**
- 2 other types of trees are over 10% : Chestnut (Maronnier – 12.5%) and Lime tree (Tilleul -10.5%)
- **The 2 most prevalent species** are X Hispanica at 18% and hippocastanum at 9.9%.
- The **varieties are much more spread** with the top variety Baumannii at only 2.3%

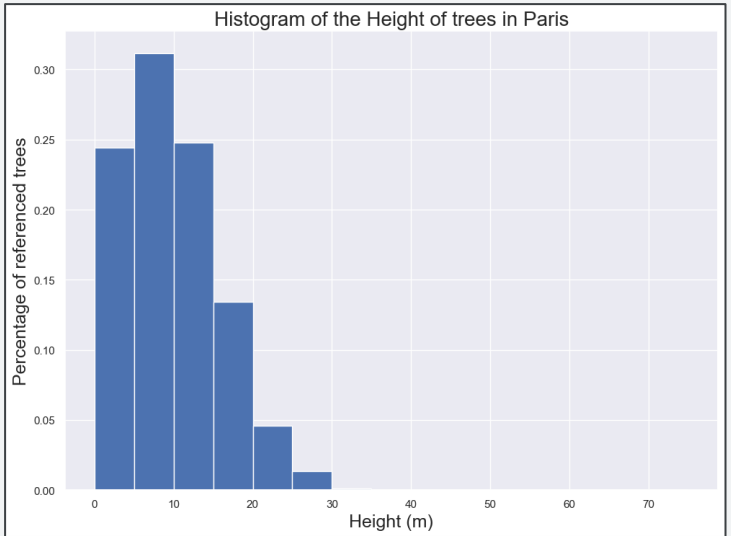




Distribution of Height and Circumference (1/2)

- **More than 65% of the trees** have a **circumference below 1m** (100cm)
- Several outliers in tree circumference: **0.10% of trees with a circumference over 4m** (max = 13.6m)
- Several outliers in tree height : only **0.11% of trees with a height over 30m** (max = 66m)



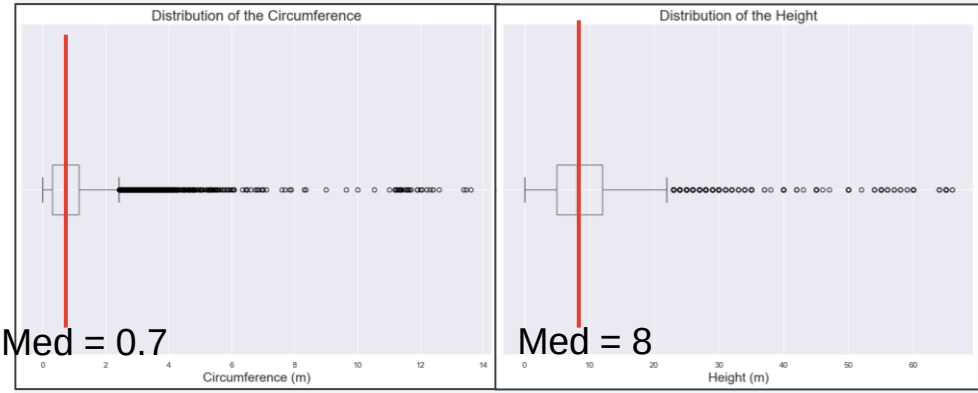


Distribution of Height and Circumference (2/2)

- There's ~ a **tenfold difference** in height and circumference mean/median values
- Height **has a very high variance (28.23)**
- **The standard deviation remains relatively low for both variables** (0.61 / 5.31)
- Both variables are **skewed right,** with **circumference having the higher skewness**
- **Circumference has a very high kurtosis at 25**, while height has only a kurtosis of 5.

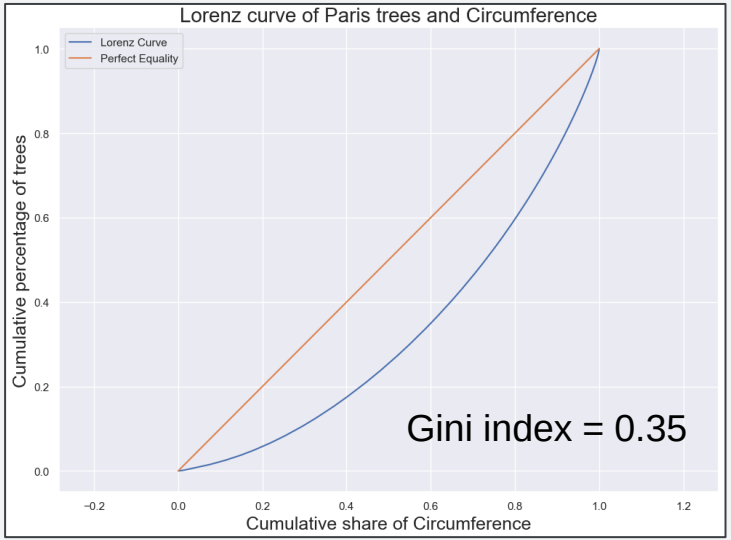
```
Circumference Mean: 0.8
Circumference Median: 0.7
Circumference Mode: 0.0
Circumference Variance: 0.42
Circumference Standard deviation: 0.65
Circumference Skewness: 1.98
Circumference Kurtosis: 18.11
```

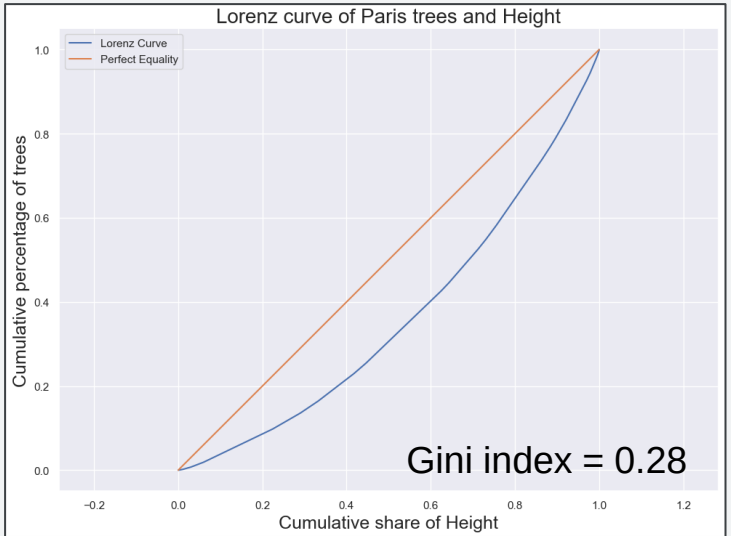
```
Height Mean: 8.36
Height Median: 8.0
Height Mode: 0.0
Height Variance: 39.68
Height Standard deviation: 6.3
Height Skewness: 0.75
Height Kurtosis: 2.15
```



Concentration of height and circumference

- The concentration of circumference is more inequal than that of height
- **Gini index : 0.35 for circumference / 0.28 for height**
- The Gini indexes are below 0.4 ➔ **distribution is adequately distributed**

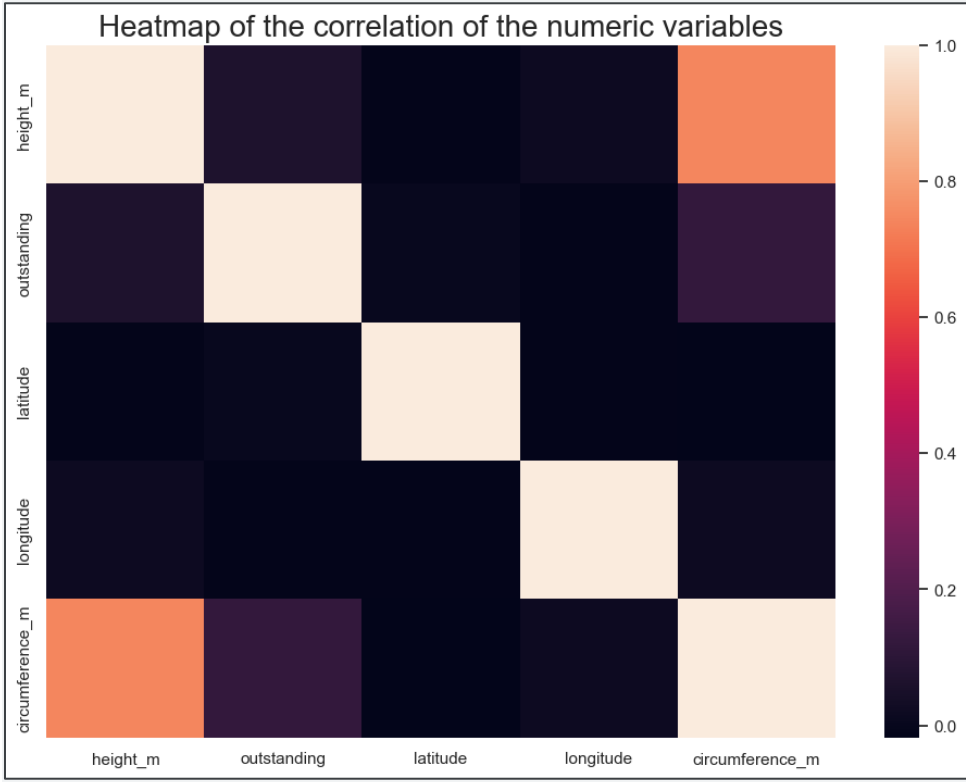




Correlation Analysis

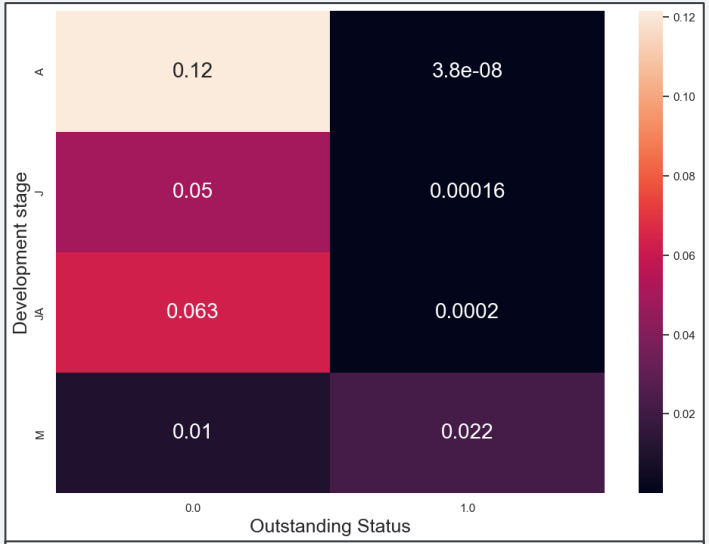
Correlation of numeric variables

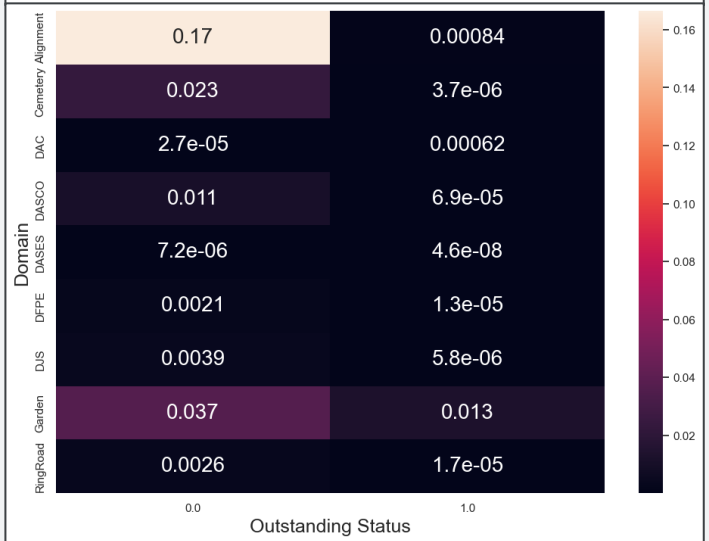
- **Strong correlation** between height and circumference (R = 0.73)
- **No significant correlation between the other numeric variables**



Association of categorical variables

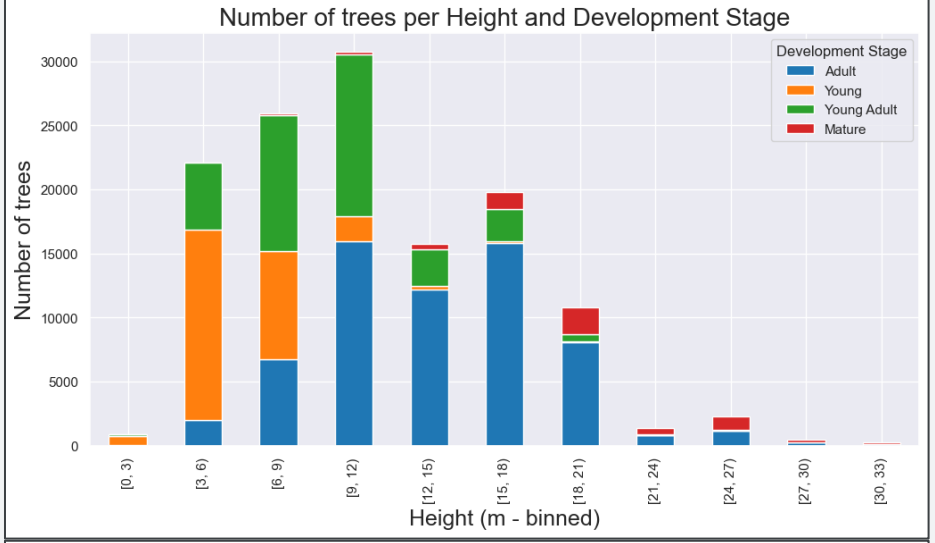
- **Strong association** between Outstanding status and Development stage (Cramer's V = 0.57)
- **Moderate association** between Outstanding status and Domain (Cramer's V = 0.45)
- **Moderate association** between Domain and Development Stage (Cramer's V = 0.35)

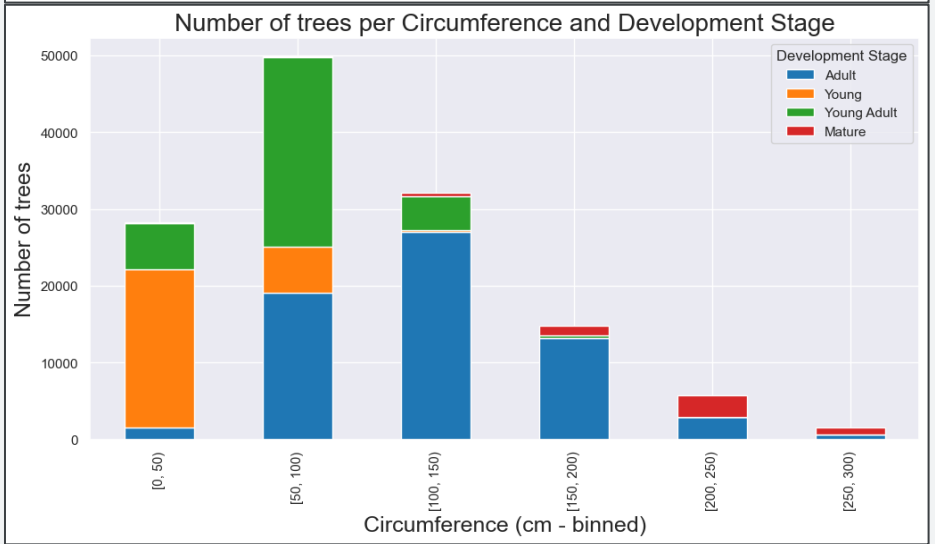




Item category and age

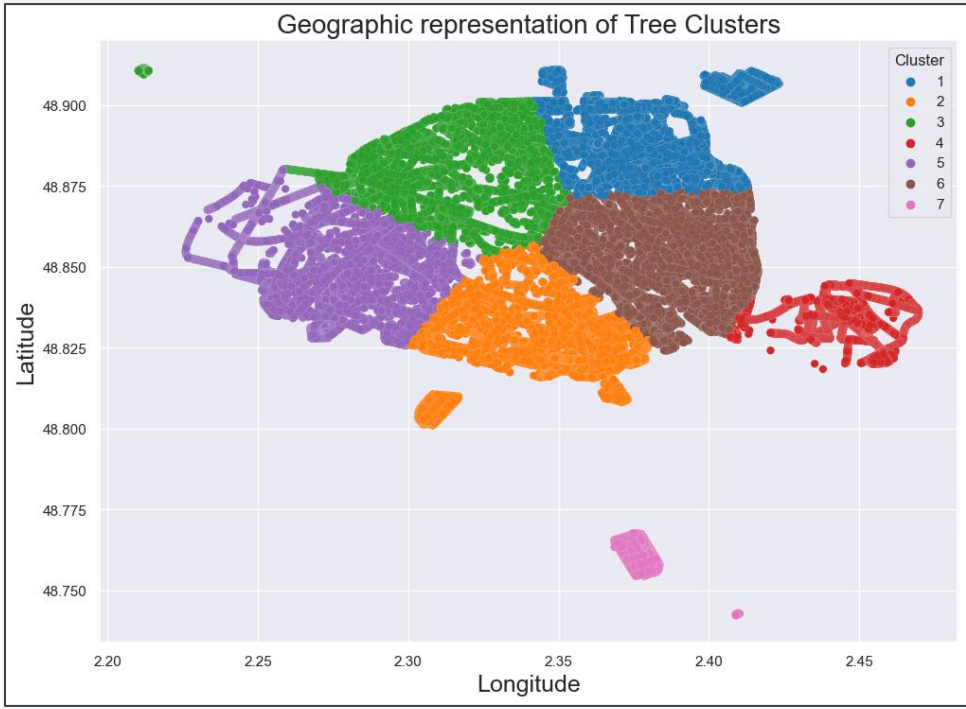
- **Strong correlation correlation** between Circumference and Development stage (eta² = 0.51)
- **Moderate correlation correlation** between Height and Development stage (eta² = 0.38)
- **Other variables are not correlated** (eta² < 0.03)
- Analysis of the distribution of height and circumference shows **young / young adult trees smaller and less wide than Adult and Mature trees**.





Creating coordinate clusters

- Applied **Kmeans clustering algorithm** to divide the trees in coordinate clusters
- Applied elbow method and calculated Calinski Harabasz scores to compute **n_clusters = 7**
- Visualization shows that clusters are **nicely separated**
- **➔ Could be used to divide maintenance crew areas of responsibility**



Conclusion

Conclusion

- **More than 40%** of trees in Paris are either **plane trees**, **chestnut trees** or **lime trees**.
- **Half of the trees** are categorized as **adult**
- **More than 50%** of the trees are in an **Alignment**
- **55%** of the trees have a **height below 10m** and **65%** of the trees have a **circumference below 1m**
- Height and Circumference are **strongly correlated** to each other and to Development Stage
- **7 clusters** have been created to **optimize the division of maintenance crew areas of responsibility**

Questions?