# FuyawYang_FinalProject-2

May 29, 2023

## 1 Trees of Vancouver

### 1.1 Foreword

Author: Daniel Fu Yaw Yang Date 19/5/2023

We are going to do an exploratory data analysis for a subset of Vancouver Trees Data ( Vancouver Street Trees)

## 2 Introduction

For this project, we will be using a subset of the Vancouver Street TreesLinks. data set. We are provided with a smaller data set with 5,000 rows. https://raw.githubusercontent.com/UBC-MDS/data_viz_wrangled/main/data/Trees_data_sets/small_unique_vancouver.csvLinks. The data were obtained from The city of Vancouver's Open Data Portal and follows an Open Government License – VancouverLinks to an external site and has been wrangled and cleaned and then was generated randomly

The dataset has undergone wrangling and cleaning processes to ensure its quality and usability. Additionally, to preserve privacy and confidentiality, the dataset has been randomly generated, while still maintaining the overall characteristics and attributes of the original data.

By leveraging this curated dataset, we aim to explore and analyze various aspects of the tree population in Vancouver. The dataset contains a wealth of information about each tree, including attributes such as diameter, height range, genus name, and neighbourhood location. These attributes provide a comprehensive view of the trees' physical characteristics, species diversity, and spatial distribution.

Throughout this analysis, we will utilize interactive visualizations to effectively present the findings and engage with the data. These visualizations will include scatter plots, bar plots, and box plots, among others, enabling us to examine relationships, patterns, and trends within the tree population.

Through this project, we seek to gain insights into the urban forest in Vancouver, understand the factors influencing tree growth and distribution, and highlight the importance of trees in enhancing the city's environmental quality. Let's embark on this journey of exploring the fascinating world of Vancouver's street trees using this curated dataset!

#### 2.0.1 Question(s) of interests

1. Is there a correlation between diameter and height of the trees?
2. How does the growth rate vary across different species ?

3. Does the location planted have an impact on tree growth?
4. Do root barriers affect growth in this case diameter of trees.

# 3 Import libraries

Let's import the libraries needed for this project

[96]:
```
import altair as alt
import pandas as pd
import numpy as np
import os
from vega_datasets import data
from altair import datum

#This line of code is so it shows u in html
alt.data_transformers.enable("data_server")
```

[96]:
```
DataTransformerRegistry.enable('data_server')
```

### 3.0.1 Read in the dataframe from the url provided to us

[97]:
```
# Read data from url

url="https://raw.githubusercontent.com/UBC-MDS/data_viz_wrangled/main/data/
 ↪Trees_data_sets/small_unique_vancouver.csv"
trees_df=pd.read_csv(url)

trees_df
```

[97]:

| | Unnamed: 0 | std_street | on_street | species_name | \ |
|---|---|---|---|---|---|
| 0 | 10747 | W 20TH AV | W 20TH AV | PLATANOIDES | |
| 1 | 12573 | W 18TH AV | W 18TH AV | CALLERYANA | |
| 2 | 29676 | ROSS ST | ROSS ST | NIGRA | |
| 3 | 8856 | DOMAN ST | DOMAN ST | AMERICANA | |
| 4 | 21098 | EAST BOULEVARD | EAST BOULEVARD | HIPPOCASTANUM | |
| … | … | … | … | … | |
| 4995 | 6132 | E 53RD AV | E 53RD AV | SERRULATA | |
| 4996 | 5642 | E 32ND AV | E 32ND AV | XX | |
| 4997 | 8777 | DAWSON ST | DAWSON ST | TULIPIFERA | |
| 4998 | 23489 | E 13TH AV | E 13TH AV | INVOLUCRATA | |
| 4999 | 7450 | CULLODEN ST | CULLODEN ST | CAMPESTRE | |

| | neighbourhood_name | date_planted | diameter | street_side_name | \ |
|---|---|---|---|---|---|
| 0 | Riley Park | 2000-02-23 | 28.5 | EVEN | |
| 1 | Arbutus-Ridge | 1992-02-04 | 6.0 | ODD | |
| 2 | Sunset | NaN | 12.0 | ODD | |
| 3 | Killarney | 1999-11-12 | 11.0 | EVEN | |

```
4                  Shaughnessy          NaN      15.5              ODD
…                          …          …       …                 …
4995       Victoria-Fraserview          NaN      17.0             EVEN
4996  Kensington-Cedar Cottage   2014-01-14       3.0             EVEN
4997                 Killarney   2002-04-15       3.5             EVEN
4998            Mount Pleasant   2003-12-02       5.5             EVEN
4999  Kensington-Cedar Cottage          NaN       3.0              ODD

        genus_name assigned  …  plant_area curb tree_id  \
0             ACER        N  …          15    Y   21421
1            PYRUS        N  …           7    Y  129645
2            PINUS        N  …           7    Y  154675
3         FRAXINUS        N  …           7    Y  180803
4         AESCULUS        Y  …           N    Y   74364
…              …      …  …          …  …     …
4995        PRUNUS        N  …           9    Y   47059
4996        CORNUS        N  …          10    N  247874
4997  LIRIODENDRON        N  …           7    Y  192642
4998       DAVIDIA        N  …           5    Y  202500
4999          ACER        N  …           8    Y  259433

                     common_name height_range_id  on_street_block  \
0                    NORWAY MAPLE               4                0
1                CHANTICLEER PEAR               2             2300
2                   AUSTRIAN PINE               4             7800
3             AUTUMN APPLAUSE ASH               4             6900
4            COMMON HORSECHESTNUT               4             5200
…                             …               …                …
4995     KWANZAN FLOWERING CHERRY               2             2200
4996  EDDIES WHITE WONDER DOGWOOD               1             1700
4997              ARNOLD TULIPTREE               2             6500
4998    DOVE OR HANDKERCHIEF TREE                1              300
4999               RED SHINE MAPLE               1             4500

             cultivar_name root_barrier   latitude   longitude
0                      NaN            N  49.252711 -123.106323
1              CHANTICLEER            N  49.256350 -123.158709
2                      NaN            N  49.213486 -123.083254
3          AUTUMN APPLAUSE            N  49.220839 -123.036721
4                      NaN            N  49.238514 -123.154958
…                      …          …        …         …
4995               KWANZAN            N  49.221161 -123.061023
4996  EDDIE'S WHITE WONDER            N  49.241544 -123.070644
4997                ARNOLD            N  49.224511 -123.048723
4998                   NaN            Y  49.259208 -123.096905
4999             RED SHINE            N  49.243772 -123.078967
```

[5000 rows x 21 columns]

# 4 Data Description

### 4.0.1 Cleaning Dataframe

Now that we have the Dataframe, we can see that there are alot of columns and they might be ir
Hence we will be dropping them so we can see a cleaner Dataframe.We will be dropping these colu

'Unnamed: 0' - Doesnt provide any useful data
'std_street' - location will not be used in this project
"street_side_name" - location will not be used in this project
'civic_number' - We will be using genus name instead
"tree_id" - id doesnt specify type of tree so it is not relevant
"on_street_block" - location will not be used in this project
'cultivar_name' - Not relevant to project
'date_planted' - It wil be interesting to use this but it has too much NaN values therefor
'on_street' - location will not be used in this project
'assigned' - not sure what this provides therefore dropped as well
'plant_area' - we will be using root barrier instead
'curb' - not relevant
'common_name' - genus name is prefered
'latitude' - location will not be used in this project
'longitude' - location will not be used in this project

[98]:
```
#drop irrelevant columns
trees_df= trees_df.drop(columns=['Unnamed: 0',␣
↪'std_street',"street_side_name",'civic_number',"tree_id","on_street_block",'cultivar_name',
trees_df
```

[98]:

| | species_name | neighbourhood_name | diameter | genus_name | \ |
|---|---|---|---|---|---|
| 0 | PLATANOIDES | Riley Park | 28.5 | ACER | |
| 1 | CALLERYANA | Arbutus-Ridge | 6.0 | PYRUS | |
| 2 | NIGRA | Sunset | 12.0 | PINUS | |
| 3 | AMERICANA | Killarney | 11.0 | FRAXINUS | |
| 4 | HIPPOCASTANUM | Shaughnessy | 15.5 | AESCULUS | |
| … | … | … | … | … | |
| 4995 | SERRULATA | Victoria-Fraserview | 17.0 | PRUNUS | |
| 4996 | XX | Kensington-Cedar Cottage | 3.0 | CORNUS | |
| 4997 | TULIPIFERA | Killarney | 3.5 | LIRIODENDRON | |
| 4998 | INVOLUCRATA | Mount Pleasant | 5.5 | DAVIDIA | |
| 4999 | CAMPESTRE | Kensington-Cedar Cottage | 3.0 | ACER | |

| | height_range_id | root_barrier |
|---|---|---|
| 0 | 4 | N |
| 1 | 2 | N |
| 2 | 4 | N |

```
3                    4             N
4                    4             N
…                  …             …
4995                 2             N
4996                 1             N
4997                 2             N
4998                 1             Y
4999                 1             N

[5000 rows x 6 columns]
```

Now let's see the Dataframe' 's info

[99]:
```
#describe dataframe

trees_df.info()
print("\n")
trees_df.describe()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 6 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   species_name        5000 non-null   object
 1   neighbourhood_name  5000 non-null   object
 2   diameter            5000 non-null   float64
 3   genus_name          5000 non-null   object
 4   height_range_id     5000 non-null   int64
 5   root_barrier        5000 non-null   object
dtypes: float64(1), int64(1), object(4)
memory usage: 234.5+ KB
```

[99]:

| | diameter | height_range_id |
|---|---|---|
| count | 5000.000000 | 5000.00000 |
| mean | 12.340888 | 2.73440 |
| std | 9.266600 | 1.56957 |
| min | 0.000000 | 0.00000 |
| 25% | 4.000000 | 2.00000 |
| 50% | 10.000000 | 2.00000 |
| 75% | 18.000000 | 4.00000 |
| max | 71.000000 | 9.00000 |

There are no more missing or NaN values in our Dataframe Now that we have a clean dataframe with all the relevant data,we can proceed to our reports!

The trees_df dataset provides valuable information about trees in Vancouver. The dataset is

sourced from a URL, which contains comprehensive data related to tree characteristics, such as diameter, height range, genus name, and neighbourhood location. This dataset is particularly useful for studying the correlation between tree attributes and their spatial distribution.

The trees_df dataset offers a wide range of variables that can be explored to gain insights into the tree population. It includes attributes such as the diameter of trees, which provides an indication of their size and growth. The height range, expressed in multiples of 10 feet, allows for further examination of the vertical dimensions of the trees. Additionally, the genus name provides the species classification, offering insights into the different types of trees present in the area.

With the trees_df dataset and the accompanying visualizations, we can delve into the characteristics of the tree population, study their spatial distribution, and gain insights into the factors influencing their growth and presence. Let's now explore the visualizations and dive into the fascinating world of trees of Vancouver!

### 4.1 Is there a relationship between diameter and height of the trees?

[100]:
```
scatter = alt.Chart(trees_df).mark_circle().encode(
    x=alt.X('diameter', title='Diameter (in)'),
    y=alt.Y('height_range_id',title ='Height(*10 ft)'),
    color=alt.Color('genus_name:N', legend=None),
    tooltip=['genus_name:N', 'diameter:Q', 'height_range_id:Q',␣
 ↪'neighbourhood_name:N']
).properties(
    width=700,
    height=500
)
scatter

# create linear regression line
regression = scatter.transform_regression(
    'diameter', 'height_range_id', method='poly', order=1
).mark_line(color='red')
regression

brush = alt.selection_interval(encodings=['x', 'y'])
height_diameter_chart = (scatter + regression).add_selection(brush).properties(
    title='Relationship between tree height and diameter',
    width=600,
    height=400
)

# Apply opacity based on brush selection
height_diameter_chart = height_diameter_chart.encode(
    opacity=alt.condition(brush, alt.value(0.8), alt.value(0.1))
)

height_diameter_chart
```

[100]:
```
alt.LayerChart(…)
```

One important aspect of studying tree populations is understanding their growth patterns. In this analysis, we will explore the relationship between tree height and diameter to gain insights into the growth characteristics of different tree species.

Upon examining the scatter plot and linear regression line, we observe a clear positive correlation between tree height and diameter. The linear relationship indicates that as the height of a tree increases, so does its diameter. This relationship holds true across various tree species represented in the dataset.

The presence of a linear trend suggests that diameter and height can be a reliable indicator of tree growth. The mean line further reinforces this observation, indicating the average growth pattern of trees in terms of height and diameter. This finding aligns with our understanding of tree biology, as height and diameter are fundamental measures of a tree's size and development.

By focusing on either height or diameter as a growth factor, we can gain valuable insights into the overall growth patterns of different tree species. Further analysis can be conducted to examine the growth rates of specific tree species or identify any variations in growth patterns among different neighborhoods or regions in Vancouver.

Understanding the growth dynamics of trees is essential for urban planning, ecological studies, and maintaining a healthy urban forest. By leveraging the relationships between height and diameter, we can gain a deeper understanding of tree growth and contribute to effective tree management practices and conservation efforts.

### 4.2 How does the growth rate vary across different species?

The Dataframe is has a lot of genus of trees so lets filter it out to the top 5 most trees planted around Vancouver

[121]:
```
# Calculate the count of each species
genus_count = trees_df['genus_name'].value_counts()

# Select the top three species
top_genus = genus_count.head(5).index.tolist()

# Filter the dataframe for the top three species
filtered_df = trees_df[trees_df['genus_name'].isin(top_genus)]
filtered_df
```

[121]:

| | species_name | neighbourhood_name | diameter | genus_name | \ |
|---|---|---|---|---|---|
| 0 | PLATANOIDES | Riley Park | 28.5 | ACER | |
| 3 | AMERICANA | Killarney | 11.0 | FRAXINUS | |
| 6 | CAMPESTRE | Victoria-Fraserview | 12.0 | ACER | |
| 8 | PALUSTRIS | Downtown | 8.0 | QUERCUS | |
| 11 | CERASIFERA | Kerrisdale | 4.5 | PRUNUS | |
| … | … | … | … | … | |
| 4991 | AMERICANA | Renfrew-Collingwood | 19.0 | TILIA | |
| 4992 | SERRULATA | Victoria-Fraserview | 20.0 | PRUNUS | |

```
4994    TRUNCATUM                   Marpole       3.0       ACER
4995    SERRULATA       Victoria-Fraserview      17.0     PRUNUS
4999    CAMPESTRE  Kensington-Cedar Cottage       3.0       ACER

      height_range_id root_barrier
0                   4            N
3                   4            N
6                   3            N
8                   1            N
11                  2            N
…                 …            …
4991                5            N
4992                2            N
4994                1            N
4995                2            N
4999                1            N

[2962 rows x 6 columns]
```

Now that we have a filtered dataframe, we can chart a circle point graph of the diameter of the 5 tree genus' in vancouver

[122]:
```
genus_names = filtered_df['genus_name'].unique().tolist()

dropdown = alt.binding_select(options=genus_names)
selection = alt.selection_single(fields=['genus_name'], bind=dropdown,␣
 ↪name='Genus')

top5_diameter_chart = alt.Chart(filtered_df).mark_circle().encode(
    x=alt.X('diameter', title='Diameter (cm)'),
    y=alt.Y('height_range_id', sort='-x', title ='Count'),
    color=alt.Color('genus_name', title='Tree Species'),
    tooltip=['genus_name', 'diameter', 'height_range_id', 'neighbourhood_name']
).properties(
    title='Diameter Distribution for Top 5 Species',
    width=600,
    height=400
).add_selection(selection).transform_filter(selection)
top5_diameter_chart
```

[122]:
```
alt.Chart(…)
```

[123]:
```
# Create a boxplot to analyze diameter distribution for the selected species
median_diameter_chart = alt.Chart(filtered_df).mark_boxplot().encode(
    x='genus_name:N',
    y='diameter:Q',
    color='genus_name:N'
```

```
).properties(
    title='Diameter Distribution for Top 5 Species',
    width=600,
    height=400
)

median_diameter_chart
```

[123]: alt.Chart(…)

To investigate whether different tree genera exhibit variations in median diameter, we can perform a comparative analysis of the median diameter across different genera. By examining the central tendency of diameter measurements for each genus, we can identify potential differences in size among tree species.

First we used a scatterplot to show the distribution of diameter for the top 5 genus of trees and can see that they are about the same but it isnt definitive so median diameter will be used moving forward.

Next let's visualize the distribution of diameters for different tree genera using a box plot. The box plot provides a concise summary of the distribution by displaying the median, quartiles, and potential outliers for each genus. This visual representation will allow us to compare the medians of different genera and assess any variations.

By analyzing the box plot, we can compare the median diameters among different tree genera. A significant difference in median diameters across genera would suggest variations in tree size and growth patterns. Additionally, we can also observe the spread of the data and identify any potential outliers that might indicate exceptional cases within specific genera. Our results shows that different genus do have different median diameter hence they have a different growth rate

# 5 Question 3

## 5.1 Does the location planted have an impact on tree growth?

Now we are going to see if different locations produces healthier hence wider trees!

[124]:
```
# Create a scatter plot to correlate location planted with tree diameter
location_plot = alt.Chart(filtered_df).mark_circle().encode(
    x=alt.X('neighbourhood_name:O', title='Neighbourhood'),
    y=alt.Y('diameter:Q', title='Diameter'),
    color=alt.Color('diameter:Q', scale=alt.Scale(scheme='viridis'),␣
 ↪legend=None)
).properties(
    title='Correlation between Location Planted and Tree Diameter',
    width=600,
    height=400
)

location_plot
```

[124]: alt.Chart(…)

To investigate whether the location where trees are planted has an impact on tree growth, we can examine the correlation between the location planted (neighbourhood) and the diameter of the trees.

By analyzing the scatter plot, we can look for any noticeable patterns or trends in the distribution of tree diameters across different neighbourhoods. If there is a correlation between the location planted and tree diameter, we might observe clusters of data points with similar diameter values in specific neighbourhoods.

From location_plot we can see that the clusters among the neighbourhood are about the same across so in Vancouver the location doesnt really affect growth rate of trees.

# 6 Question 4

## 6.1 Do root barriers affect growth in this case diameter of Trees

Now we will see if a root barrier affects the growth of the trees

Eventhough the color scheme isnt the greatest even with the graph all zoomed in, we can vaguely tell that trees with root barriers has lesser growth and it makes sense as it impedes their space to grow but if we want a clear and simple graph we can just go for a bar plot as shown below:

[125]:
```
root_barrier_plot = alt.Chart(filtered_df).mark_bar().encode(
    x=alt.X('root_barrier:N', title='Root Barrier'),
    y=alt.Y('diameter:Q', title='Diameter'),
    color=alt.Color('root_barrier:N', legend=None)
).properties(
    title='Comparison of Root Barrier Assigned with Tree Diameter',
    width=400,
    height=300
)
root_barrier_plot
```

[125]: alt.Chart(…)

To explore the potential impact of root barriers on tree growth, specifically the diameter of trees, we can create a bar plot that compares the presence or absence of root barriers with the tree diameter.

By examining the bar plot, we can observe the differences in the average or total diameter of trees based on the presence or absence of root barriers. There is a noticeable disparity in the diameter between trees with root barriers and those without, it suggests that root barriers might have an impact on tree growth, specifically in terms of diameter in that it almost halves their growth rate compared to tres without root barriers.

# 7 Discussion

Based on the combined graph of diameter and height, we observed a positive relationship between these two variables. As the diameter increases, the height of the trees also tends to increase.

The linear regression line further supports this correlation. This suggests that there is a strong association between diameter and height in tree growth.

To assess the variation in growth rates across different tree species, we focused on the top five species in the dataset. Using faceted and box plot graphs, we analyzed the distribution of tree diameter among these species. The graphs reveal distinct differences in diameter distributions, suggesting that each species exhibits its own unique growth characteristics. Consequently, growth rates can vary significantly among different tree species.

We investigated the impact of the location planted on tree growth by creating a scatter plot map representing the top five tree species across Vancouver. However, our analysis did not indicate a significant correlation between the location planted and tree growth. This suggests that factors other than location play a more substantial role in influencing the growth rate of trees in Vancouver.

By analyzing the bar plot comparing root barrier presence and tree diameter, we can clearly see that root barriers have a noticeable impact on tree growth. The bar plot indicates that trees without root barriers exhibit significantly higher diameter compared to those with root barriers. This finding suggests that root barriers have a limiting effect on tree growth, particularly in terms of diameter. The presence of root barriers appears to reduce the growth rate of trees, resulting in smaller diameters.

In conclusion, our analysis has provided valuable insights into the growth factors and patterns of trees. We observed a positive correlation between diameter and height, indicating that these factors are interrelated in tree growth. Additionally, we identified variations in growth rates among different tree species, highlighting the importance of species-specific growth considerations. While the location planted does not seem to have a direct impact on tree growth, the presence of root barriers significantly affects tree diameter, leading to reduced growth rates. These findings contribute to our understanding of tree growth dynamics and can inform decision-making processes related to urban forestry and tree management.

## 8 DashBoard

[113]:
```
brush = alt.selection_single(encodings=['x'], on='click', empty='none',␣
 ↪fields=['root_barrier'])

# Create a bar plot to compare root assigned with tree diameter
root_barrier_plot = alt.Chart(filtered_df).mark_bar().encode(
    x=alt.X('root_barrier:N', title='Root Barrier'),
    y=alt.Y('diameter:Q', title='Diameter'),
    color=alt.Color('root_barrier:N', legend=None),
).properties(
    title='Comparison of Root Barrier Assigned with Tree Diameter',
    width=400,
    height=300
).add_selection(brush)



# Create a scatter plot to correlate location planted with tree diameter
```

```
location_plot = alt.Chart(filtered_df).mark_circle().encode(
    x=alt.X('neighbourhood_name:O', title='Neighbourhood'),
    y=alt.Y('diameter:Q', title='Diameter'),
    color=alt.Color('diameter:Q', scale=alt.Scale(scheme='viridis'),␣
 ↪legend=None),
    opacity=alt.condition(brush, alt.value(0), alt.value(1))
).properties(
    title='Correlation between Location Planted and Tree Diameter',
    width=700,
    height=500
).add_selection(brush)
```

[114]:
```
genus_names = filtered_df['genus_name'].unique().tolist()

dropdown = alt.binding_select(options=genus_names)
selection = alt.selection_single(fields=['genus_name'], bind=dropdown,␣
 ↪name='Genus')

# Scatter plot with opacity based on selection
top5_diameter_chart = alt.Chart(filtered_df).mark_circle().encode(
    x=alt.X('diameter', title='Diameter (cm)'),
    y=alt.Y('height_range_id', sort='-x', title ='Count'),
    color=alt.condition(selection, alt.Color('genus_name:N', title='Tree␣
 ↪Species'), alt.value('lightgray')),
    tooltip=['genus_name', 'diameter', 'height_range_id', 'neighbourhood_name']
).properties(
    title='Diameter Distribution for Top 5 Species',
    width=700,
    height=500
).add_selection(selection)

# Boxplot with selection interaction
median_diameter_chart = alt.Chart(filtered_df).mark_boxplot().encode(
    x='genus_name:N',
    y='diameter:Q',
    color=alt.condition(selection, alt.Color('genus_name:N', legend=None), alt.
 ↪value('lightgray'))
).properties(
    title='Diameter Distribution for Top 5 Species',
    width=600,
    height=400
).add_selection(selection)
```

[115]:
```
(location_plot | root_barrier_plot )  & (top5_diameter_chart  |␣
 ↪median_diameter_chart)
```

[115]:
```
alt.VConcatChart(…)
```

# 9 References


- Trees DataFrame : https://opendata.vancouver.ca/explore/dataset/street-trees/information/?disjunctive.species_name&disjunctive.common_name&disjunctive.height_range_id&di
- Vancouver Map Help : https://cdn-uploads.piazza.com/paste/klvia6r082u1jy/f31721ac4272704ae2ef201335c
- Data Visualisation Modules : https://canvas.ubc.ca/courses/114341/modules