FuyawYang_EDA-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)

1.2 Introduction

1.2.1 Question(s) of interests

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

1. Is there a correlation between diameter and height of the trees?
2. How does the diameter 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.

2 Import libraries

Let's import the libraries nessecary for our EDA

[1]:
```
import altair as alt
import pandas as pd
import numpy as np
import os
from vega_datasets import data

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

[1]:
```
DataTransformerRegistry.enable('data_server')
```

2.0.1 Read in the dataframe from the url provided to us

[2]:
```
# 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
```

[2]:

	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]
```

2.0.2 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'
'std_street'
"street_side_name"
'civic_number'
"tree_id"
"on_street_block"
'cultivar_name'
'date_planted'

[3]:
```
#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
```

[3]:

```
            on_street   species_name        neighbourhood_name  diameter  \
0          W 20TH AV    PLATANOIDES                Riley Park      28.5
1          W 18TH AV     CALLERYANA             Arbutus-Ridge       6.0
2            ROSS ST          NIGRA                    Sunset      12.0
3           DOMAN ST      AMERICANA                 Killarney      11.0
4     EAST BOULEVARD  HIPPOCASTANUM                Shaughnessy      15.5
…               …            …                         …       …
4995       E 53RD AV      SERRULATA       Victoria-Fraserview      17.0
4996       E 32ND AV             XX  Kensington-Cedar Cottage       3.0
4997       DAWSON ST     TULIPIFERA                 Killarney       3.5
4998       E 13TH AV    INVOLUCRATA            Mount Pleasant       5.5
4999     CULLODEN ST      CAMPESTRE  Kensington-Cedar Cottage       3.0

        genus_name assigned plant_area curb                  common_name  \
0             ACER        N         15    Y                 NORWAY MAPLE
1            PYRUS        N          7    Y             CHANTICLEER PEAR
2            PINUS        N          7    Y                AUSTRIAN PINE
3         FRAXINUS        N          7    Y          AUTUMN APPLAUSE ASH
4         AESCULUS        Y          N    Y         COMMON HORSECHESTNUT
…              …      …         …   …                            …
4995        PRUNUS        N          9    Y     KWANZAN FLOWERING CHERRY
4996        CORNUS        N         10    N  EDDIES WHITE WONDER DOGWOOD
4997  LIRIODENDRON        N          7    Y             ARNOLD TULIPTREE
4998       DAVIDIA        N          5    Y    DOVE OR HANDKERCHIEF TREE
4999          ACER        N          8    Y               RED SHINE MAPLE

      height_range_id root_barrier   latitude   longitude
0                   4            N  49.252711 -123.106323
1                   2            N  49.256350 -123.158709
2                   4            N  49.213486 -123.083254
3                   4            N  49.220839 -123.036721
4                   4            N  49.238514 -123.154958
…               …          …         …         …
4995                2            N  49.221161 -123.061023
4996                1            N  49.241544 -123.070644
4997                2            N  49.224511 -123.048723
4998                1            Y  49.259208 -123.096905
4999                1            N  49.243772 -123.078967

[5000 rows x 13 columns]
```

Now let's see the Dataframe' 's info

[4]:

```
#describe dataframe

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 13 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   on_street           5000 non-null   object
 1   species_name        5000 non-null   object
 2   neighbourhood_name  5000 non-null   object
 3   diameter            5000 non-null   float64
 4   genus_name          5000 non-null   object
 5   assigned            5000 non-null   object
 6   plant_area          4950 non-null   object
 7   curb                5000 non-null   object
 8   common_name         5000 non-null   object
 9   height_range_id     5000 non-null   int64
 10  root_barrier        5000 non-null   object
 11  latitude            5000 non-null   float64
 12  longitude           5000 non-null   float64
dtypes: float64(3), int64(1), object(9)
memory usage: 507.9+ KB
```

[4]:

	diameter	height_range_id	latitude	longitude
count	5000.000000	5000.00000	5000.000000	5000.000000
mean	12.340888	2.73440	49.247349	-123.107128
std	9.266600	1.56957	0.021251	0.049137
min	0.000000	0.00000	49.202783	-123.220560
25%	4.000000	2.00000	49.230152	-123.144178
50%	10.000000	2.00000	49.247981	-123.105861
75%	18.000000	4.00000	49.263275	-123.063484
max	71.000000	9.00000	49.293930	-123.023311

3 Question 1

3.1

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

To find out we are going to chart a graph with diameter(inches) on the x axis and height(using height range id as it is 10ft per unit) on the y axis.

[5]:
```
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
```

[5]: alt.Chart(…)

We can see the realtionship on the chart above that it is linear but it isnt very definitive so lets see if it'll show a more linear result if we take the mean of both height and diameter. With the code below, we can see that it is a linear progression as diameter increases height increases so for the rest of the EDA we can focus on using one of them as they have a positive relationship.

[6]:
```
# create linear regression line
regression = scatter.transform_regression(
    'diameter', 'height_range_id', method='poly', order=1
).mark_line(color='red')
regression
```

[6]: alt.Chart(…)

[7]:
```
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
```

[7]: alt.LayerChart(…)

There is! As height increases diameter also increases around all the species of trees provided

4 Question 2

4.1 Do different genus of tree have different median diameter?

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

[8]:
```
# 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
```

[8]:

	on_street	species_name	neighbourhood_name	diameter	genus_name	assigned	plant_area	curb	common_name	height_range_id	root_barrier	latitude	longitude
0	W 20TH AV	PLATANOIDES	Riley Park	28.5	ACER	N	15	Y	NORWAY MAPLE	4	N	49.252711	-123.106323
3	DOMAN ST	AMERICANA	Killarney	11.0	FRAXINUS	N	7	Y	AUTUMN APPLAUSE ASH	4	N	49.220839	-123.036721
6	NASSAU DRIVE	CAMPESTRE	Victoria-Fraserview	12.0	ACER	N	15	Y	HEDGE MAPLE	3	N	49.217522	-123.071311
8	W PENDER ST	PALUSTRIS	Downtown	8.0	QUERCUS	N	C	Y	PIN OAK	1	N	49.281303	-123.108253
11	W 45TH AV	CERASIFERA	Kerrisdale	4.5	PRUNUS	N	8	Y	NIGHT PURPLE LEAF PLUM	2	N	49.230925	-123.156131
…	…	…	…	…	…	…	…	…	…	…	…	…	…
4991	WALES ST	AMERICANA	Renfrew-Collingwood	19.0	TILIA	N	7	Y	BASSWOOD	5	N	49.236139	-123.051816
4992	E 53RD AV	SERRULATA	Victoria-Fraserview	20.0	PRUNUS	N	9	Y	KWANZAN FLOWERING CHERRY	2	N	49.221161	-123.060833
4994	ASH ST	TRUNCATUM	Marpole	3.0	ACER	N	NaN	Y	PACIFIC SUNSET MAPLE	1	N	49.216851	-123.120103
4995	E 53RD AV	SERRULATA	Victoria-Fraserview	17.0	PRUNUS	N	9	Y	KWANZAN FLOWERING CHERRY	2	N	49.221161	-123.061023
4999	CULLODEN ST	CAMPESTRE	Kensington-Cedar Cottage	3.0	ACER	N	8	Y	RED SHINE MAPLE	1	N	49.243772	-123.078967

```
[2962 rows x 13 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

[9]:
```
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
)
chart
```

[9]:
```
alt.Chart(…)
```

Then we can Facet the graph above to each area and see if the diameter of the trees stay linear in every area

[10]:
```
chart_facet =chart.facet(column='genus_name', columns=5).
 ↪resolve_scale(y='independent')
chart_facet
```

[10]:
```
alt.FacetChart(…)
```

With the facted graphs we can see that they do not have the same distribution of diameter. and if we want to go on detail,we can plot a boxplot to figure out the median diameter and more.

[11]:
```
# Create a boxplot to analyze diameter distribution for the selected species
filtered_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
)

filtered_chart
```

[11]:
```
alt.Chart(…)
```

From the boxplot we can see that Fraxinus has smallest diameter of trees.Then Prunus and Tilia that has the same median,while diameter of Prunus is more spread apart compared to Tilia and

finally with the highest median diameter is Quercus.

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!

[12]:
```
# create a dropdown selection tool for common_name
genus_name_select = alt.binding_select(options=list(filtered_df['genus_name'].
 ↪unique()), name='Genus Name')
genus_name_selector = alt.selection_single(fields=['genus_name'],␣
 ↪bind=genus_name_select)

# create a dropdown selection tool for neighbourhood_name
neighbourhood_name_select = alt.
 ↪binding_select(options=list(trees_df['neighbourhood_name'].unique()),␣
 ↪name='Neighbourhood')
neighbourhood_name_selector = alt.
 ↪selection_single(fields=['neighbourhood_name'],␣
 ↪bind=neighbourhood_name_select)
```

Let's focus on the neighbourhood on the first chart.With this graph we can see the average diameter distribution of all the top 5 trees in a selected area

[13]:
```
genus_name_chart = chart.encode(opacity=alt.
 ↪condition(neighbourhood_name_selector, alt.value(1), alt.value(0))).
 ↪add_selection(neighbourhood_name_selector)
genus_name_chart
```

[13]:
```
alt.Chart(…)
```

After fiddling around the dropdown,we can see that the distribution of the diameter of trees are quite similar accross areas but its not quite clear.

We will need a mmmore suitable graph that includes all the trees without separating them into their own genus' so we can produce a more clearer picture of the distribution of diameter in different areas

[14]:
```
# Create a scatter plot to correlate location planted with tree diameter
scatter_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
```

```
)

scatter_plot
```

[14]: alt.Chart(…)

In this scatterplot, the distribution of diameter among all the trees across the neighbourhoods are about even so location does not affect the growth rate of trees in Vancouver

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

[15]:
```
scatter_plot = alt.Chart(filtered_df).mark_circle(
).encode(
    x=alt.X('genus_name:N', title='Genus Name'),
    y=alt.Y('diameter:Q', title='Diameter'),
    color=alt.Color('root_barrier:N', scale=alt.Scale(scheme='darkred',␣
 ↪reverse=True), legend=alt.Legend(title='Root Barrier')),
    tooltip=['genus_name', 'root_barrier', 'diameter']
).properties(
    title='Correlation between Genus Name, Root Barrier, and Diameter',
    width=600,
    height=400
).interactive()


scatter_plot
```

[15]: alt.Chart(…)

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:

[16]:
```
bar_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 Assigned with Tree Diameter',
    width=400,
    height=300
)
bar_plot
```

[16]: `alt.Chart(…)`

With this it is very clear that root barriers almost halves their growth rate

7 Conclusion

7.1 Question 1 : Is there a correlation between diameter and height of the trees?

~ The combined graph show that there is a positive relationship between diameter and height.
As diameter increases,height also increases

7.2 Question 2 : Do different genus of tree have different median diameter ?

~ We can see that in the faceted and boxplot graphs that the distribution of diameter accross t

7.3 Question 3 : Does the location planted have an impact on tree growth?

~ we plotted out a scattermap of the top 5 trees around vancouver and are able to see that it

7.4 Question 4 : Do root barriers affect growth in this case diameter of trees

~ Yes The bar plot paints a very clear picture that root barriers almost halves the growth rate

8 References

Trees DataFrame : https://opendata.vancouver.ca/explore/dataset/street-trees/information/?disjunctive.species_name&disjunctive.common_name&disjunctive.height_range_id&disjunct
Vancouver Map Help : https://cdn-uploads.piazza.com/paste/klvia6r082u1jy/f31721ac4272704ae2ef201335cd6142
Data Visualisation Modules : https://canvas.ubc.ca/courses/114341/modules