final_project

May 30, 2023

1 Exploration of Vancouver Trees

Author : Muntakim Rahman UBC Student Number : 71065221

1.1 Introduction

This notebook will be conducting an analysis for the Vancouver Trees dataset located in the `small_unique_vancouver.csv` file.

1.2 Import Packages

[1]:
```
import numpy as np
import pandas as pd
import altair as alt
import datetime as dt

from tree_functions import *
```

[2]:
```
vancouver_df = pd.read_csv('small_unique_vancouver.csv', index_col = 0)
display(vancouver_df.head())
```

```
            std_street        on_street   species_name neighbourhood_name  \
10747        W 20TH AV        W 20TH AV    PLATANOIDES         Riley Park
12573        W 18TH AV        W 18TH AV     CALLERYANA      Arbutus-Ridge
29676          ROSS ST          ROSS ST          NIGRA             Sunset
8856          DOMAN ST         DOMAN ST      AMERICANA          Killarney
21098   EAST BOULEVARD   EAST BOULEVARD  HIPPOCASTANUM        Shaughnessy

      date_planted  diameter street_side_name genus_name assigned  \
10747   2000-02-23      28.5             EVEN       ACER        N
12573   1992-02-04       6.0              ODD     PYRUS        N
29676          NaN      12.0              ODD      PINUS        N
8856    1999-11-12      11.0             EVEN   FRAXINUS        N
21098          NaN      15.5              ODD   AESCULUS        Y

       civic_number plant_area curb  tree_id         common_name  \
10747            66         15    Y    21421        NORWAY MAPLE
12573          2323          7    Y   129645    CHANTICLEER PEAR
```

```
29676          7855          7     Y    154675         AUSTRIAN PINE
8856           6938          7     Y    180803    AUTUMN APPLAUSE ASH
21098          5295          N     Y     74364   COMMON HORSECHESTNUT

       height_range_id  on_street_block    cultivar_name root_barrier  \
10747                4                0              NaN            N
12573                2             2300       CHANTICLEER            N
29676                4             7800              NaN            N
8856                 4             6900  AUTUMN APPLAUSE            N
21098                4             5200              NaN            N

        latitude   longitude
10747  49.252711 -123.106323
12573  49.256350 -123.158709
29676  49.213486 -123.083254
8856   49.220839 -123.036721
21098  49.238514 -123.154958
```

[3]:
```python
print(f''''There are {len(vancouver_df)} entries in the dataset.''')
```

```
'There are 5000 entries in the dataset.
```

1.2.1 Observe Outputs

Let's start by getting an understanding of the data sparsity (i.e. *NULL* values), as well as the column distributions.

[4]:
```python
display(vancouver_df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 5000 entries, 10747 to 7450
Data columns (total 20 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   std_street         5000 non-null   object
 1   on_street          5000 non-null   object
 2   species_name       5000 non-null   object
 3   neighbourhood_name 5000 non-null   object
 4   date_planted       2363 non-null   object
 5   diameter           5000 non-null   float64
 6   street_side_name   5000 non-null   object
 7   genus_name         5000 non-null   object
 8   assigned           5000 non-null   object
 9   civic_number       5000 non-null   int64
 10  plant_area         4950 non-null   object
 11  curb               5000 non-null   object
 12  tree_id            5000 non-null   int64
 13  common_name        5000 non-null   object
 14  height_range_id    5000 non-null   int64
```

```
 15   on_street_block     5000 non-null   int64
 16   cultivar_name       2658 non-null   object
 17   root_barrier        5000 non-null   object
 18   latitude            5000 non-null   float64
 19   longitude           5000 non-null   float64
dtypes: float64(3), int64(4), object(13)
memory usage: 820.3+ KB
```

```
None
```

Data Sparsity There are *NULL* occurrences in the `date_planted`, `plant_area`, `cultivar_name` columns. Let's keep these for now to visualize the data in the entries without *NULL* values.

Non-Numeric Data

[5]:
```
objects_df = vancouver_df.describe(include = 'object').T
display(objects_df)
```

	count	unique	top	freq
std_street	5000	603	CAMBIE ST	52
on_street	5000	607	CAMBIE ST	49
species_name	5000	171	SERRULATA	463
neighbourhood_name	5000	22	Renfrew-Collingwood	384
date_planted	2363	1599	2004-02-16	7
street_side_name	5000	4	ODD	2554
genus_name	5000	67	ACER	1218
assigned	5000	2	N	4564
plant_area	4950	38	10	736
curb	5000	2	Y	4593
common_name	5000	361	KWANZAN FLOWERING CHERRY	383
cultivar_name	2658	176	KWANZAN	383
root_barrier	5000	2	N	4679

Observing the data stored as objects, there seem to be variation in distinct values for given columns.

The `std_street` and `on_street` column have greater than *600* distinct values and would not be good candidates for the *EDA*.

Looking at the `date_planted` column, it seems that there are only *1599* distinct values in the entire dataset. This would entail repeated dates across the entries, which is rather interesting.

The `curb` and `root_barrier` columns are binary in nature and should be one-hot encoded in our final analysis.

Numeric Data

[6]:
```
numeric_df = vancouver_df.describe(include = np.number).T
display(numeric_df)
```

```
                  count        mean         std        min  \
diameter         5000.0   12.340888    9.266600   0.000000
```

civic_number	5000.0	2975.707600	2078.580429	2.000000
tree_id	5000.0	128682.584600	75412.260406	36.000000
height_range_id	5000.0	2.734400	1.569570	0.000000
on_street_block	5000.0	2960.227000	2086.861052	0.000000
latitude	5000.0	49.247349	0.021251	49.202783
longitude	5000.0	-123.107128	0.049137	-123.220560

	25%	50%	75%	max
diameter	4.000000	10.000000	18.000000	71.000000
civic_number	1300.500000	2639.000000	4123.000000	9113.000000
tree_id	61321.500000	130130.500000	191332.000000	270750.000000
height_range_id	2.000000	2.000000	4.000000	9.000000
on_street_block	1300.000000	2600.000000	4100.000000	9100.000000
latitude	49.230152	49.247981	49.263275	49.293930
longitude	-123.144178	-123.105861	-123.063484	-123.023311

Observing the data stored as type `np.number`, there seem to be differences in `std deviation` for given columns.

Based on the `std deviation` of *75412.260406*, the `tree_id` column probably includes data for a unique identifier. We can use this to identify our trees, but it doesn't serve much other use for our *EDA*.

There is a very large `std deviation` for the `civic_number` column, with the `min` value being *2* and the `max` being *9113*. There is similar behavior in the `on_street_block` column, which very similar `mean`, `min`, and `max` values to `civic_number`. I'm not particularly interested in these columns, but we can visualize the correlation.

The `height_range_id` column has a `mean` value, as well as a `25th` and `50th` percentile ~*2* which is interesting. I'd like to see the distribution of this column.

The `latitude` and `longitude` column have a `std deviation` less than *0.1*, which would entail most trees being in the same vicinity. We can try using this data to see where trees are densely concentrated.

1.3 Questions of Interest

We want to explore this dataset to understand :

- What trees are commonly found in Vancouver?
- Where are trees located in Vancouver?
- How big are these trees?
- When were these trees planted?

1.3.1 Columns of Interest

We are going to be visualizing the data in the following columns :

- `genus_name`
- `latitude`
- `longitude`
- `neighbourhood_name`

- height_range_id
- diameter
- date_planted

[7]:
```
vancouver_df = vancouver_df[
    [
        'latitude', 'longitude', 'neighbourhood_name',
        'genus_name',
        'height_range_id', 'diameter', 'plant_area',
        'date_planted'
    ]
]
```

[8]:
```
display(vancouver_df.head())
display(vancouver_df.tail())
```

```
        latitude   longitude neighbourhood_name genus_name  height_range_id  \
10747  49.252711 -123.106323         Riley Park       ACER                4
12573  49.256350 -123.158709      Arbutus-Ridge      PYRUS                2
29676  49.213486 -123.083254             Sunset      PINUS                4
8856   49.220839 -123.036721          Killarney   FRAXINUS                4
21098  49.238514 -123.154958        Shaughnessy   AESCULUS                4

       diameter plant_area date_planted
10747      28.5         15   2000-02-23
12573       6.0          7   1992-02-04
29676      12.0          7          NaN
8856       11.0          7   1999-11-12
21098      15.5          N          NaN
```

```
        latitude   longitude        neighbourhood_name    genus_name  \
6132   49.221161 -123.061023       Victoria-Fraserview        PRUNUS
5642   49.241544 -123.070644  Kensington-Cedar Cottage        CORNUS
8777   49.224511 -123.048723                 Killarney  LIRIODENDRON
23489  49.259208 -123.096905            Mount Pleasant       DAVIDIA
7450   49.243772 -123.078967  Kensington-Cedar Cottage          ACER

       height_range_id  diameter plant_area date_planted
6132                 2      17.0          9          NaN
5642                 1       3.0         10   2014-01-14
8777                 2       3.5          7   2002-04-15
23489                1       5.5          5   2003-12-02
7450                 1       3.0          8          NaN
```

1.3.2 Data Transformation

Prior to visualizing the dataset, we will be assigning the `decade_planted` column to provide more meaning to the time periods in which trees were planted. This will also enable us to implement a `decade_planted` filter to our visualizations.

[9]:
```
vancouver_df = vancouver_df.assign(
    decade_planted = vancouver_df['date_planted'].apply(
        lambda x : f'''{(dt.datetime.strptime(x, '%Y-%m-%d').year // 10) *␣
↪10}s''' if x == x else np.nan
    )
)

display(vancouver_df.head())
display(vancouver_df.tail())
```

```
        latitude   longitude neighbourhood_name genus_name  height_range_id  \
10747  49.252711 -123.106323         Riley Park       ACER                4
12573  49.256350 -123.158709      Arbutus-Ridge      PYRUS                2
29676  49.213486 -123.083254             Sunset      PINUS                4
8856   49.220839 -123.036721          Killarney   FRAXINUS                4
21098  49.238514 -123.154958        Shaughnessy   AESCULUS                4

       diameter plant_area date_planted decade_planted
10747      28.5         15   2000-02-23          2000s
12573       6.0          7   1992-02-04          1990s
29676      12.0          7          NaN            NaN
8856       11.0          7   1999-11-12          1990s
21098      15.5          N          NaN            NaN
        latitude   longitude         neighbourhood_name    genus_name  \
6132   49.221161 -123.061023         Victoria-Fraserview        PRUNUS
5642   49.241544 -123.070644  Kensington-Cedar Cottage        CORNUS
8777   49.224511 -123.048723                  Killarney  LIRIODENDRON
23489  49.259208 -123.096905             Mount Pleasant       DAVIDIA
7450   49.243772 -123.078967  Kensington-Cedar Cottage          ACER

       height_range_id  diameter plant_area date_planted decade_planted
6132                 2      17.0          9          NaN            NaN
5642                 1       3.0         10   2014-01-14          2010s
8777                 2       3.5          7   2002-04-15          2000s
23489                1       5.5          5   2003-12-02          2000s
7450                 1       3.0          8          NaN            NaN
```

1.4 Analysis

1.4.1 Q1 : What trees are commonly found in Vancouver?

Let's plot the count of each `genus_name` to visualize the most and least common trees within the city. Let's trim down the `genus_name` visualized to include the *10* most and *10* least common trees.

[10]:
```
genera_df = vancouver_df['genus_name'].value_counts() \
.sort_values(ascending = False) \
.to_frame() \
.assign(total_trees = len(vancouver_df))
```

```
genera_df = genera_df.reset_index()
genera_df.columns = ['genus_name', 'number_of_trees', 'total_trees']

display(genera_df.head())
```

	genus_name	number_of_trees	total_trees
0	ACER	1218	5000
1	PRUNUS	1050	5000
2	FRAXINUS	238	5000
3	TILIA	238	5000
4	QUERCUS	218	5000

[11]:
```
fig_num = 1
```

[12]:
```
most_common_plot, fig_num = get_genera_plot(
    effective_df = vancouver_df,
    subtitle = 'Most Common Vancouver Tree Genera',
    most_common = True,
    fig_num = fig_num
)
```

[13]:
```
least_common_plot, _ = get_genera_plot(
    effective_df = vancouver_df,
    subtitle = 'Least Common Vancouver Tree Genera',
    most_common = False
)
```

[14]:
```
base_genera_plot = (most_common_plot | least_common_plot)

genera_plot = base_genera_plot \
.configure_legend(
  orient = 'right',
  titleFontSize = 15,
  labelFontSize = 12
).configure_axis(
    labelFontSize = 10, titleFontSize = 15
).configure_mark(
  stroke = 'black',
  strokeOpacity = 1,
  strokeWidth = 0.8
).configure_axis(
    labelFontSize = 10, titleFontSize = 15
).configure_title(
    fontSize = 25
)
display(genera_plot)
```

alt.HConcatChart(…)

From `Figure 1` :

- The *10* most common tree genera amount to ~*74.96%* of the total trees in Vancouver.
  - It's noticable that greater than *45%* of trees are either *Acer* or *Prunus*.
- The *10* least common tree genera amount to ~*0.22%* of the total trees in Vancouver.
  - The *9* least common tree genera have only *1* tree in Vancouver, while the *10th* least common tree genera has *2* trees in Vancouver.
- We might want to look into what other features these trees tend to share.

Q2 : Where are trees located in Vancouver? Let's bin the `latitude` and `longitude` coordinates in a heatmap to visualize the tree density within a given area.

[15]:
```
base_coordinates_plot = alt.Chart(
    vancouver_df,
    title = alt.TitleParams(
        text = f'Figure {fig_num} : Location of Vancouver Trees',
        subtitle = ['Latitude and Longitude Heatmap'],
        anchor = 'start', fontSize = 25, subtitleFontSize = 20
    )
).mark_bar().encode(
    x = alt.X('latitude:Q', title = 'Latitude', bin = alt.Bin(maxbins = 15)),
    y = alt.Y('longitude:Q', title = 'Longitude', bin = alt.Bin(maxbins = 15)),
    color = alt.Color(
        'count():Q', scale = alt.Scale(scheme = 'viridis', reverse = True),
        legend = alt.Legend(
            title = 'Number of Trees',
            titleFontSize = 14, labelFontSize = 12
        ),
    ),
    tooltip = [alt.Tooltip('count():Q', title = 'Number of Trees')]
).properties(
    width = 600, height = 500
)

coordinates_plot = base_coordinates_plot \
.configure_mark(
  stroke = 'black',
  strokeOpacity = 1,
  strokeWidth = 1.25,
).configure_axis(
    labelFontSize = 15, titleFontSize = 17.5
).configure_title(
    fontSize = 25
)

fig_num += 1
```

```
display(coordinates_plot)
```

alt.Chart(…)

From `Figure 2` :

- We observe that trees are generally distributed rather evenly.
- It's noticable that there are *230* trees within *49.250* <= `latitude` <= *29.260* and *-123.120* <= `longitude` <= *-123.100*.
- There are fewer trees located around the edges of the map, with the exception of *49.220* <= `latitude` <= *49.290* and *-123.040* <= `longitude` <= *-123.020*.

1.4.2 Q3 : What Sizes are Vancouver Trees? Is there a Relationship Between Diameter and Height Range ID?

Let's plot the `diameter` and `height_range_id` columns to visualize the relationship between the two properties. This might act as a proxy for determining whether trees occupying a greater area also tend to be taller.

[16]:
```
base_sizes_heatmap = alt.Chart(
    vancouver_df
).mark_circle().encode(
    x = alt.X(f'diameter:Q', title = 'Diameter', bin = alt.Bin(maxbins = 15)),
    y = alt.Y(f'height_range_id:Q', title = 'Height Range Id', bin = alt.
 ↪Bin(maxbins = 15)),
    color = alt.Color(
        'count():Q', scale = alt.Scale(
            scheme = 'viridis', reverse = True,
        ),
        legend = alt.Legend(
            title = 'Number of Trees',
            titleFontSize = 14, labelFontSize = 12,
            orient = 'right', direction = 'vertical'
        ),
    ),
    size = alt.Size('count():Q'),
    tooltip = [alt.Tooltip('count():Q', title = 'Number of Trees')]
).properties(
  title = alt.TitleParams(
    text = f'Figure {fig_num} : Vancouver Tree Size Dimensions',
    subtitle = ['Relationship Between Diameter and Height Range ID'],
    anchor = 'start', fontSize = 25, subtitleFontSize = 20
  ), width = 600, height = 500
)

sizes_heatmap = base_sizes_heatmap \
.configure_mark(
  stroke = 'black',
  strokeOpacity = 1,
```

```
  strokeWidth = 0.5
).configure_axis(
    labelFontSize = 15, titleFontSize = 17.5
)

fig_num += 1
display(sizes_heatmap)
```

alt.Chart(…)

From `Figure 3` :

- Trees tend towards lower `diameter` values and *5* <= `height_range_id` <= *10*.
  - There is a slight positive relationship between `diameter` and `height_range_id` where *0* <= `diameter` <= *25* and *1.0* <= `height_range_id` <= *5.0*.
    * More trees seem to tend towards having feature values in the lower bins of this domain.
  - There are *959* trees with *0* <= `diameter` <= *5* and *1.0* <= `height_range_id` <= *2.0*.

1.4.3 Q4 : What neighborhoods have the largest trees? What about the smallest trees?

Let's look at the breakdown of this data for both `diameter` and `height_range_id` by `neighborhood_name`.

[17]:
```
base_neighbourhoods_plot = alt.Chart(
    vancouver_df,
    width = 300, height = 350
).mark_boxplot().encode(
    x = alt.X(alt.repeat(), type = 'quantitative'),
    y = alt.Y('neighbourhood_name:N', title = 'Neighbourhoods'),
).repeat(
    ['diameter', 'height_range_id'],
    columns = 3
).properties(
    title = alt.TitleParams(
        text = f'Figure {fig_num} : Vancouver Tree Sizes in Different␣
↪Neighbourhoods',
        subtitle = ['Diameters and Height Range ID Distributions'],
        anchor = 'start', fontSize = 25, subtitleFontSize = 20
    )
)

neighbourhoods_plot = base_neighbourhoods_plot \
.configure_mark(
  stroke = 'black',
  strokeOpacity = 1,
  strokeWidth = 0.5
).configure_axis(
```

```
    labelFontSize = 12, titleFontSize = 15
).configure_title(
    fontSize = 10
)

fig_num += 1
display(neighbourhoods_plot)
```

alt.RepeatChart(…)

From `Figure 4` :

- For `diameter` :
  - Most regions tend to have *~10 <= median* `diameter` *<= ~15*
  - All regions have a *75th* percentile `diameter` <= *24*.
  - The *Downtown* region has a lower range of *4* <= `diameter` <= *10* between *25th* and *75th* percentiles.
- For `height_range_id` :
  - There seem to be *2* buckets of `neighbourhood_names` :
    * Have a higher *75th* percentile >= *4*
    * Have a lower *75th* percentile *~3*.

1.4.4 Q5 : How did tree sizes change by decade?

[18]:
```
base_decades_plot = alt.Chart(
    vancouver_df,
    width = 300, height = 350
).mark_area(opacity = 0.5).encode(
    x = alt.X(alt.repeat(), type = 'quantitative', bin = alt.Bin(maxbins = 15)),
    y = alt.Y('count():Q', title = 'Number of Trees', stack = None),
    color = alt.Color(
        'decade_planted:O', scale = alt.Scale(
            scheme = 'tableau10', reverse = False,
        ),
        legend = alt.Legend(
            title = 'Decade Planted',
            titleFontSize = 14, labelFontSize = 12
        ),
    )
).repeat(
    ['diameter', 'height_range_id'], columns = 2
).properties(
    title = alt.TitleParams(
        text = f'Figure {fig_num} : Vancouver Tree Sizes in Different Decades',
        subtitle = ['Diameters and Height Range ID Distributions'],
        anchor = 'start', fontSize = 25, subtitleFontSize = 20
    )
)
```

```
decades_plot = base_decades_plot \
.configure_mark(
  stroke = 'black',
  strokeOpacity = 1,
  strokeWidth = 0.8
).configure_axis(
    labelFontSize = 12, titleFontSize = 15
).configure_title(
    fontSize = 10
)

fig_num += 1
display(decades_plot)
```

alt.RepeatChart(…)

From `Figure 5` :

- There are a great number of unidentified `date_planted` and `decade_planted` values.
  - These trees appear to have distributions with higher values for `diameter` and `height_range_id`.
- For the identified `decade_planted` values, it seems that :
  - The *2000s* and *2010s* have a great number of trees with lower values for `diameter` and `height_range_id`.
  - The *1980s* and *1990s* have a similar number of trees with relatively higher values for `diameter` and `height_range_id`.

2 Further Questions

I would like to explore the data in these charts when filtered for criteria including :

- `neighbourhood_names` with the most trees
- most common `genus_names`
- `decade_planted` across the dataset

A few questions start to emerge when looking at data for the columns we've considered for size, as well as trends over time.

- Do trees of the same `genus_name` have similar numerical features?
- Do trees with the same `neighbourhood_name` tend to have the same `genus_names`?
- Where are more trees being planted over time?
- Has the tree density by `latitude` and `longitude` changed over time?

2.1 Interactive Dashboard

Let's create a dashboard from the visuals above in order to start investigating these questions.

[19]:
```
neighbourhoods_select = alt.selection_single(
    fields = ['neighbourhood_name'],
```

```
    bind = {
        'neighbourhood_name' : alt.binding_select(
            name = 'Neighbourhoods',
            options = list(
                vancouver_df.groupby('neighbourhood_name')['genus_name'] \
                    .agg('count').sort_values(ascending = False) \
                    .reset_index()['neighbourhood_name']
            )[:10]
        )
    }
)
```

[20]:
```
decades_select = alt.selection_single(
    fields = ['decade_planted'],
    bind = {
        'decade_planted' : alt.binding_radio(
            name = 'Decades',
            options = sorted([decade for decade in␣
↪vancouver_df['decade_planted'].unique() if decade == decade])
        )
    }
)
```

[21]:
```
genus_select = alt.selection_multi(fields=['genus_name'])
```

[22]:
```
coordinates_plot = base_coordinates_plot \
.add_selection(
    decades_select
).add_selection(
    neighbourhoods_select
).add_selection(
    genus_select
).transform_filter(
    decades_select
).transform_filter(
    neighbourhoods_select
).transform_filter(
    genus_select
)

coordinates_plot.title.text = coordinates_plot.title.text.split(' : ', 1)[-1]
```

[23]:
```
most_common_plot = most_common_plot.encode(
    opacity = alt.condition(genus_select, alt.value(1), alt.value(0.2)),
    tooltip = [
        alt.Tooltip('count():Q', title = 'Number of Trees')
    ]
```

```
).properties(
    width = 600, height = 500
).add_selection(
    genus_select
).transform_filter(
    decades_select
).transform_filter(
    neighbourhoods_select
)

most_common_plot.title.text = 'Most Common Vancouver Tree Genera'
most_common_plot.title.subtitle = 'Number of Trees Planted'
```

[24]:

```
sizes_heatmap = base_sizes_heatmap \
.properties(
    width = 350, height = 350
).transform_filter(
    neighbourhoods_select
).transform_filter(
    genus_select
)

sizes_heatmap.title.text = sizes_heatmap.title.text.split(' : ', 1)[-1]
```

[25]:

```
decades_plot = base_decades_plot \
.transform_filter(
    decades_select
).transform_filter(
    neighbourhoods_select
).transform_filter(
    genus_select
)

decades_plot.title.text = decades_plot.title.text.split(' : ', 1)[-1]
```

[26]:

```
trees_dashboard = (
  (coordinates_plot | most_common_plot).resolve_scale(
    color = 'independent', size = 'independent'
  ) & (sizes_heatmap | decades_plot).resolve_scale(
    color = 'independent', size = 'independent'
  )
).configure_mark(
  stroke = 'black', strokeOpacity = 1, strokeWidth = 0.5
).configure_axis(
    labelFontSize = 15, titleFontSize = 17.5
).properties(
    title = alt.TitleParams(
```

```
        text = f'Trees in Vancouver Metropolitan Area', fontSize = 65, anchor =␣
↪'middle'
    )
)

display(trees_dashboard)
```

alt.VConcatChart(…)

2.2 Visualizations

2.2.1 Choices

In order to visualize the density of data points, we have used both a traditional **heatmap** and **circle plot** of varying colors and sizes. The *viridis* color scheme enables the clear distinction of changes in density, complemented by a legend.

We have also used a **histogram** to visualize the `number_of_trees` per `genus_name` in the effective dataset. This acts as a simple means of displaying the breakdown of tree genera, while also acting as a dashboard filter.

To compare the `diameter` and `height_range_id` distributions, we have layered the data by `decade_planted` in a translucent **area chart**. Since we are implementing color to highlight the nominal `decade_planted` feature here, we are using the *tableau10* color scheme.

2.2.2 Potential Improvements

When arranged in a dashboard, the **heatmap** and **circle plot** are not aligned. This may be visually jarring and could be improved through ensuring consistency of height and width.

Another improvement would be to ensure that axes are fixed for `decade_planted` filter selections in the **area chart**. This filter can be used to help remove unnecessary `decade_planted` data. This would cause easier visual transitions on the **area chart** as we look at effective `decade_planted` values. Consequently, the **area chart** would also be more accomodating to users with visual deficiencies.

3 Discussion

We have intentionally decided to visualize the charts in a dashboard prior to our final discussion. This enables us to answer the subsequent questions which arose from our initial analysis.

3.1 Summary

- Across all `neighbourhood_name`, `decade_planted`, and `genus_name` values, we observe that :
  - The fields `diameter` and `height_range_id` have a relatively stronger positive correlation where there are lower values for both features.
  - There are a great number of unidentified `date_planted` and `decade_planted` values.
    * These trees have distributions with higher values for `diameter` and `height_range_id`.

    - * For identified `decade_planted` values, it seems the `diameter` and `height_range_id` are decreasing post *2000s*.
  - – Trees are generally distributed rather evenly across the `latitude` and `longitude` coordinates, however the density is greater in the center of the map.
  - – The *10* most common tree genera amount to ~*74.96%* of the total trees in Vancouver.
    - * It's noticable that greater than *45%* of trees are either *Acer* or *Prunus*.

3.1.1 Tree Genera

- Considering the dominant *Acer* and *Prunus* `genus_name` :
  - – The *Acer* trees tend to have lower `diameter` and `height_range_id` values and exhibit a positive relationship similar to the entire dataset.
    - * The *Acer* trees are clearly exhibit decreasing `diameter` and `height_range_id` values over time.
  - – The *Prunus* trees have a similar relationship, but there is a cluster of *10* <= `diameter` <= *20* and *2.0* <= >`height_range_id` <= *3.5*.
    - * There are a great deal of unidentified `date_planted` and `decade_planted` values for *Prunus* trees.

3.1.2 Neighbourhoods

- The most common `genus_name` are relatively present in all of the high-density neighbourhoods.
  - – Either *Acer* or *Prunus* are the most common `genus_name`.
  - – *Dunbar-Southlands*, *Sunset*, *Victoria-Fraserview*, and *Marpole* have a greater number of *Prunus* trees than *Acer* trees.
    - * There are ~*2* times as many *Prunus* trees than *Acer* trees in *Victoria-Fraserview*.

3.1.3 Location

- In the *1990s*, there were *25* trees planted at the top edge of the map where *49.230* <= `latitude` <= *49.240* and *-123.040* <= `longitude` <= *-123.020*.
- There was another cluster of *71* trees planted in the *1990s* where *49.210* <= `latitude` <= *49.230* and *-123.140* <= `longitude` <= *-123.100*.
  - – In the *2000s*, there were ~*130* trees planted where *49.210* <= `latitude` <= *49.220* and *-123.120* <= `longitude` <= *-123.060*.
  - – In the *2010s*, more trees were planted where *49.210* <= `latitude` <= *49.220* and -*123.100* <= `longitude` <= *-123.080*.
  - – Trees are consistently planted in this general vicinity.

3.2 Unanswered Questions

In order to understand where trees are being planted over time, it would make sense to visualize the **time-series** data of `number_of_trees` and compare this for different `neighbourhood_name` values. Another question which arises from the dashboard is whether `neighbourhood_name` or `latitude/longitude` coordinates are the better indicator for tree location. These could be better explored in a subsequent analysis. We could visualize the data in a map of Vancouver to get a clear understanding.

3.3 References

The data were obtained from The city of Vancouver's Open Data Portal and follows an Open Government Licence – Vancouver.

These additional resources provide the theory and code segments for the *Analysis Report* in this notebook :

- Data Visualization
- Machine Learning Final Project
- Python for Data Science Final Project