Vancouver Tree Height Geography Analysis

DSCI 522, Group 33

2024-11-23

Vancouver Tree Height Geography Analysis

Summary

Our group was interested in answering the question, "**Do tree heights vary significantly from neighborhood to neighborhood in Vancouver?**"

We aim to analyze the relationship between tree height distribution and neighborhoods. Specifically, we focus on tree height ranges and their counts in various neighborhoods to determine if tree height is influenced by location.

Introduction

Urban trees are essential to creating livable cities, offering ecological, aesthetic, and health benefits. They improve air quality, provide shade, support biodiversity, and enhance the overall urban environment. In Vancouver, street trees play a vital role in the city's commitment to sustainability. However, the characteristics of these trees, such as their heights, can vary significantly across neighborhoods due to factors like local urban planning, soil quality, and maintenance practices. Understanding these patterns is key to equitable urban forestry management and informed decision-making.

This project explores the relationship between **tree height distribution** and **neighborhoods** in Vancouver. Using data from the City of Vancouver Open Data Portal

To address this, we analyze the dataset through a combination of:

1. **Exploratory Data Analysis (EDA):** We use contingency tables and visualizations (heatmaps) to identify patterns in tree height distributions across neighborhoods.
2. **Statistical Testing:** A Chi-squared test of independence is performed to determine if the observed variations in tree height distributions are statistically significant.

By uncovering these patterns, this analysis contributes to urban forestry strategies that aim to distribute greenery benefits equitably across neighborhoods in Vancouver. The findings could help guide future decisions in tree planting, maintenance, and sustainable urban planning.

Methods and Results

Loading Required Packages

```r
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```r
library(janitor)
```

```
Attaching package: 'janitor'

The following objects are masked from 'package:stats':

    chisq.test, fisher.test
```

```r
library(ggplot2)
library(knitr)
```

Loading the Data

```r
trees <- read_csv2("data/street-trees.csv")
```

```
i Using "','" as decimal and "'.'" as grouping mark. Use `read_delim()` for more control.
```

```
Rows: 184003 Columns: 21
-- Column specification --------------------------------------------------------
Delimiter: ";"
chr  (15): STD_STREET, GENUS_NAME, SPECIES_NAME, CULTIVAR_NAME, COMMON_NAME,...
dbl   (4): TREE_ID, CIVIC_NUMBER, ON_STREET_BLOCK, HEIGHT_RANGE_ID
num   (1): DIAMETER
date  (1): DATE_PLANTED

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
head(trees)
```

```
# A tibble: 6 x 21
  TREE_ID CIVIC_NUMBER STD_STREET GENUS_NAME SPECIES_NAME   CULTIVAR_NAME
    <dbl>        <dbl> <chr>      <chr>      <chr>          <chr>
1    5313         2645 W 6TH AV   ACER       PSEUDOPLATANUS <NA>
2    5315         2648 W 6TH AV   ACER       RUBRUM         RED SUNSET
3    5321         2689 W 6TH AV   ACER       PSEUDOPLATANUS <NA>
4    5324         2710 W 6TH AV   ACER       PSEUDOPLATANUS <NA>
5    5327         2725 W 6TH AV   ACER       PSEUDOPLATANUS <NA>
6    5331         2727 W 6TH AV   ACER       PSEUDOPLATANUS <NA>
# i 15 more variables: COMMON_NAME <chr>, ASSIGNED <chr>, ROOT_BARRIER <chr>,
#   PLANT_AREA <chr>, ON_STREET_BLOCK <dbl>, ON_STREET <chr>,
#   NEIGHBOURHOOD_NAME <chr>, STREET_SIDE_NAME <chr>, HEIGHT_RANGE_ID <dbl>,
#   HEIGHT_RANGE <chr>, DIAMETER <dbl>, CURB <chr>, DATE_PLANTED <date>,
#   Geom <chr>, geo_point_2d <chr>
```

Cleaning the Data

In our analysis, we are primarily interested in the NEIGHBOURHOOD_NAME, HEIGHT_RANGE, and HEIGHT_RANGE_ID columns, so it is crucial to ensure that there are no missing values in these columns to maintain the accuracy and reliability of our statistical results. Upon checking, we confirmed that none of these three columns contain any missing data.

Additionally, we examined the entire dataset for duplicate records using the unique identifier TREE_ID and confirmed that there are no duplicate rows. This step ensures that no records are inadvertently counted multiple times.

```r
# str(trees) # maybe we don't use this
# Check for missing data in NEIGHBOURHOOD_NAME, HEIGHT_RANGE, HEIGHT_RANGE_ID columns
missing_data_check <- trees |>
  summarise(
    NEIGHBOURHOOD_NAME_missing = sum(is.na(NEIGHBOURHOOD_NAME)),
    HEIGHT_RANGE_missing = sum(is.na(HEIGHT_RANGE)),
    HEIGHT_RANGE_ID_missing = sum(is.na(HEIGHT_RANGE_ID))
  )
print(missing_data_check) # no missing data in the columns of interest
```

```
# A tibble: 1 x 3
  NEIGHBOURHOOD_NAME_missing HEIGHT_RANGE_missing HEIGHT_RANGE_ID_missing
                       <int>                <int>                   <int>
1                          0                    0                       0
```

```r
# Check for duplicates in data set
duplicate_count <- trees |>
  select(TREE_ID) |>
  duplicated() |>
  sum()
print(paste("Number of duplicate records:", duplicate_count)) # no duplicated records
```

```
[1] "Number of duplicate records: 0"
```

Exploratory Data Analysis

The columns of interest are:

1. `NEIGHBOURHOOD_NAME` (explanatory/treatment variable) - A string representing the neighbourhood the tree is in
2. Tree height data (the outcome/response variable) is represented in two columns, in different formats:
    1. `HEIGHT_RANGE` - a string representing tree heights (categorical levels) in buckets of 10ft, e.g. `0' - 10'`, `10' - 20'`, etc.
    2. `HEIGHT_RANGE_ID` - a numeric column (integers) with levels corresponding to the above strings

First, we should examine the levels of the two available versions of the response categorical variable (height), because we may be able to choose one that will simplify our subsequent analysis.

Based on the data, it looks like there should be a one-to-one correspondence between levels of `HEIGHT_RANGE` and `HEIGHT_RANGE_ID`, but we should confirm this. To do this, we can look at all unique combinations of the two variables. If they properly correspond (there are no issues with the data), we should see only one row for each. We will print the results using `kable()` (from the `knitr` package).

```
unique_combinations <- trees |>
  distinct(HEIGHT_RANGE, HEIGHT_RANGE_ID) |>
  arrange(HEIGHT_RANGE_ID)

unique_combinations |>
  kable(caption = "Mapping between tree height category names and their IDs.")
```

Table 1: Mapping between tree height category names and their IDs.

HEIGHT_RANGE	HEIGHT_RANGE_ID
0' - 10'	0
10' - 20'	1
20' - 30'	2
30' - 40'	3
40' - 50'	4
50' - 60'	5
60' - 70'	6
70' - 80'	7
80' - 90'	8
90' - 100'	9
> 100'	10

It looks like there is a proper correspondence between the levels of these two variables. Since the levels of `HEIGHT_RANGE_ID` are simpler and shorter, we will use this variable. This should make our plots easier to display.

Now we can select only the treatment and response variable columns and convert them to type `factor`, which will allow us to build a contingency table.

```
trees_subset <- trees |>
  select(NEIGHBOURHOOD_NAME, HEIGHT_RANGE_ID) |>
  mutate(across(everything(), as.factor))
```

```r
head(trees_subset)
```

```
# A tibble: 6 x 2
  NEIGHBOURHOOD_NAME HEIGHT_RANGE_ID
  <fct>              <fct>
1 KITSILANO          6
2 KITSILANO          2
3 KITSILANO          6
4 KITSILANO          4
5 KITSILANO          4
6 KITSILANO          3
```

The new `trees_subset` dataframe contains one row per tree, with only the columns `NEIGHBOURHOOD_NAME` and `HEIGHT_RANGE_ID`. We can use this along with the `tabyl()` function from the `janitor` package to generate a contingency table. We will then print the contingency table using `kable()`.

```r
cont_table <- trees_subset |>
  tabyl(NEIGHBOURHOOD_NAME, HEIGHT_RANGE_ID)

cont_table |>
  kable(caption = "Contingency table showing counts of trees in various levels of the tree he
```

Table 2: Contingency table showing counts of trees in various levels of the tree height variable per levels of the neighbourhood variable. For a mapping of tree height category names, refer to Table 1.

NEIGHBOURHOOD_NAME	0	1	2	3	4	5	6	7	8	9	10
ARBUTUS RIDGE	9	1544	1373	1301	712	324	273	182	98	34	25
DOWNTOWN	212	1888	2448	1770	1859	760	516	408	314	272	563
DUNBAR-SOUTHLANDS	11	2340	2444	1444	1129	980	772	481	207	53	41
FAIRVIEW	13	998	1036	831	1042	421	236	81	33	15	7
GRANDVIEW-WOODLAND	12	1929	1700	1275	1118	421	226	84	45	25	0
HASTINGS-SUNRISE	36	3890	3139	2250	1998	966	497	282	109	42	61
KENSINGTON-CEDAR COTTAGE	30	3459	3201	2067	2262	880	465	188	142	56	90
KERRISDALE	37	2062	2290	1243	1130	1019	621	421	260	151	130

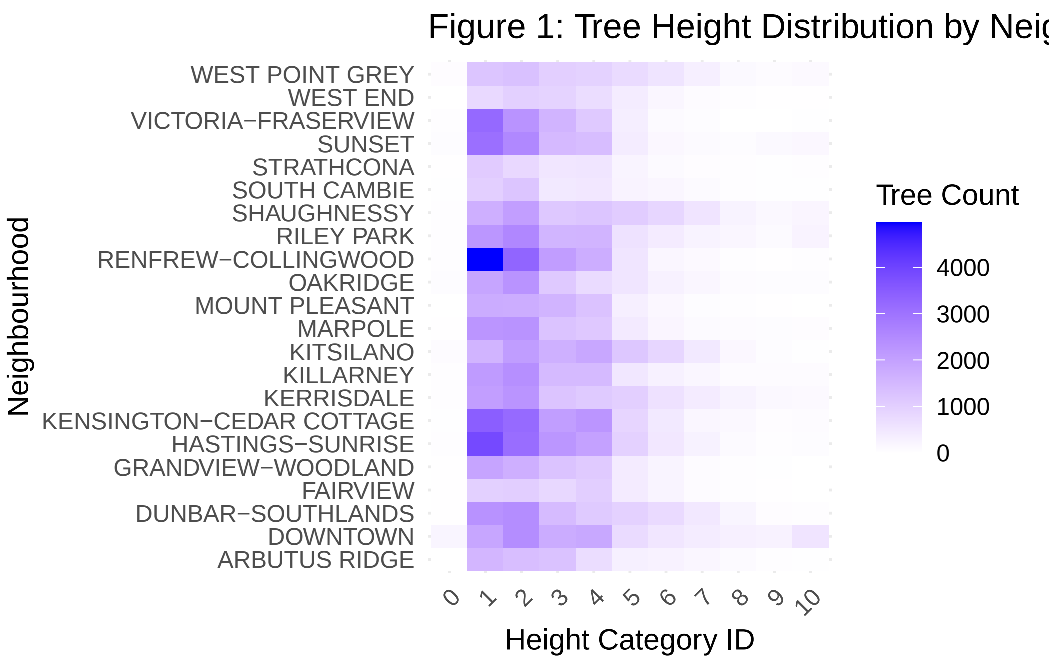
NEIGHBOURHOOD_NAME	0	1	2	3	4	5	6	7	8	9	10
KILLARNEY	42	2130	2392	1473	1469	500	300	180	79	88	86
KITSILANO	94	1599	2094	1679	1883	1193	856	467	163	70	28
MARPOLE	13	2257	2298	1264	1155	437	201	115	90	72	53
MOUNT PLEASANT	3	1776	1738	1590	1287	345	158	70	38	15	0
OAKRIDGE	67	1918	2306	1139	760	557	293	187	81	66	62
RENFREW-COLLINGWOOD	38	4956	3312	2086	1753	555	194	125	51	15	32
RILEY PARK	42	2238	2561	1578	1596	607	424	256	186	112	242
SHAUGHNESSY	48	1698	2060	1171	1231	1085	869	565	253	151	201
SOUTH CAMBIE	22	1006	1236	469	498	242	192	98	41	19	31
STRATHCONA	11	1089	812	515	549	232	106	55	32	27	38
SUNSET	58	3096	2552	1475	1406	395	164	102	66	133	164
VICTORIA-FRASERVIEW	34	3228	2314	1599	1147	359	115	73	10	7	28
WEST END	7	802	992	923	713	404	187	90	44	16	11
WEST POINT GREY	53	1232	1324	1004	955	753	575	342	136	91	128

In order to better understand patterns in this data, we can visualize the above contingency table as a two-dimensional histogram (a.k.a heat map):

```r
heatmap_data <- cont_table |>
  pivot_longer(
    cols = -NEIGHBOURHOOD_NAME,
    names_to = "HEIGHT_RANGE_ID",
    values_to = "Tree_Count"
  ) |>
  mutate(
    HEIGHT_RANGE_ID = factor(HEIGHT_RANGE_ID,
                             levels = as.character(0:10))
  )

ggplot(heatmap_data, aes(x = HEIGHT_RANGE_ID,
                         y = NEIGHBOURHOOD_NAME,
                         fill = Tree_Count)) +
  geom_tile() +
  scale_fill_gradient(low = "white",
                      high = "blue") +
  labs(
```

```
    title = "Figure 1: Tree Height Distribution by Neighbourhood",
    x = "Height Category ID",
    y = "Neighbourhood",
    fill = "Tree Count"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



From the plot, it looks like there may be some differences in tree heights across neighbourhoods. In particular:

- RENFREW-COLLINGWOOD has, by a large margin, the most trees in the 10’ - 20’ height category (ID 1)
- VICTORIA-FRASERVIEW, SUNSET, RENFREW-COLLINGWOOD, KENSINGTON-CEDAR COTTAGE, and HASTINGS-SUNRISE seem to have more trees in the lower height categories (1-4) than other neighbourhoods.
- SHAUGHNESSY and DOWNTOWN seem to have the tallest trees.

Statistical Test

Although there are some visually identifiable patterns, we can only determine whether there are statistically significant differences in tree heights per neighbourhood using an appropriate test.

Choosing a Test and Significance Level

Because we are dealing with two categorical variables, each with multiple levels, a Chi-squared test of Independence/Homogeneity is appropriate.

The Chi-squared test makes the following assumptions:

1. The observations are independent.
2. The expected counts are large enough (greater than 5 is typical).

First, the height measurement of a particular tree does not depend on that of another, so we can assume independence. Second, almost every intersection of the contingency table has counts much larger than 5, with many in the hundreds or thousands.

Because neither of the test's two key assumptions appear to be violated, it is safe to proceed with a Chi-squared test of independence.

Finally, we will choose the standard significance level of $\alpha = 0.05$ as our threshold for determining statistical significance.

Performing a Chi-squared Test of Independence

We will perform the test, store the results in `chisq_results`, and display them.

```r
chisq_results <- chisq.test(cont_table, correct = FALSE)

chisq_results
```

```

	Pearson's Chi-squared test

data:  cont_table
X-squared = 15339, df = 210, p-value < 2.2e-16
```

The Chi-squared test yielded a statistically significant result, with a test statistic of $X^2 = 15339$ and $p < 2.2 \times 10^{-16}$, which is less than our predefined significance level of 0.05.

Discussion

After conducting the analysis, we conclude that tree heights vary significantly across neighborhoods. The Chi-squared test of independence result shows the p-value approximately equal to zero thus is less than our predefined significance level of 0.05. Therefore, we can reject null hypothesis that the two categorical variables are independent (there is no association). This means that there is a statistically significant association between neighborhood and tree height.

This reflects differences in tree density in local urban planning. The EDA plot reveals that the RENFREW-COLLINGWOOD neighborhood stands out with the darkest color, indicating the highest concentration of trees, particularly among the shortest height ranges. Other neighborhoods such as VICTORIA-FRASERVIEW, SUNSET, RENFREW-COLLINGWOOD, KENSINGTON-CEDAR COTTAGE, and HASTINGS-SUNRISE tend to have a greater concentration of trees in the lower height ranges (1-4) than other areas. On the other hand, SHAUGHNESSY and DOWNTOWN are notable for having the tallest trees.

Across neighborhoods, the most common tree height range appears to be between 1 and 3 units (height range ID referring to the specific tree height), indicating that the urban tree population is predominantly composed of younger or medium-sized trees. Taller trees (height range 6-10 units) are scarce or entirely absent in most neighborhoods, which may reflect the challenges posed by urban constraints such as limited space, infrastructure interference, or deliberate pruning practices to control growth.

Tree management patterns vary greatly across neighborhoods. Urban areas like DOWNTOWN and FAIRVIEW exhibit consistently light colors across the height ranges, suggesting limited green spaces for tree planting. Surprisingly, even non-urban neighborhoods such as DUNBAR-SOUTHLANDS and ARBUTUS RIDGE show lighter colors, indicating potential land availability that could be utilized for planting new trees. This points to opportunities for the government or community organizations to prioritize these areas for greening initiatives.

Ecologically, neighborhoods with a higher density of trees, such as RENFREW-COLLINGWOOD, enjoy significant environmental benefits, including improved air quality, better urban cooling effects, and enhanced biodiversity. However, the dominance of trees in the 1-3 height range also suggests that many of these trees are young and might require additional care to ensure healthy growth into taller, mature trees.

Note that we do not have the evidence to state these insights and patterns are statstically significant, as the Chi-squared test does not specify which levels are significantly different.

Overall, the findings emphasize the need for targeted greening initiatives, particularly in urban neighborhoods like DOWNTOWN and FAIRVIEW, where tree coverage is limited. Additionally, fostering the growth of taller trees is crucial across all neighborhoods to promote long-term environmental sustainability. Encouraging the planting and care of taller tree species can help balance the urban ecosystem and create more resilient green spaces in the face of growing urbanization.

References

- City of Vancouver Open Data Portal: *Street Trees Dataset*. Available at: https://opendata.vancouver.ca/explore/dataset/street-trees/information/?disjunctive.species_name&disjunctive.common_name&disjunctive.on_street&disjunctive.neighbourhood_name
- Wickham, H., & Grolemund, G. (2016). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media.
- Janitor R Package Documentation. Available at: https://cran.r-project.org/web/packages/janitor/janitor.pdf
- ggplot2 R Package Documentation. Available at: https://ggplot2.tidyverse.org/
- Vancouver Urban Forestry Strategy. (2018). City of Vancouver. Available at: https://vancouver.ca/parks-recreation-culture/urban-forestry-strategy.aspx