Vancouver street trees analysis with Tidyverse

Mboya Daisy

2024-10-04

1. Introduction

This report presents an analysis of Vancouver's street trees using the `tidyverse` collection of R packages. The primary objectives include data cleaning, transformation, visualization, and statistical analysis using `dplyr`, `tidyr`, and `ggplot2`.

Learning Objectives:

- Familiarization with the tidyverse environment
- Data cleaning and basic analysis
- Plotting and mapping aesthetics using `ggplot2`

2. Dataset Description

The dataset is sourced from Vancouver City Open Data Portal and contains information on trees located on public streets. The key attributes include:

- Tree ID
- Tree species (Latin and common name)
- Diameter at breast height (in inches)
- Height class (integer scale from 0-10)
- Street and neighborhood name

3. Required Software and Libraries

```r
# Load necessary packages
library(tidyverse)
```

```
Warning: package 'tidyverse' was built under R version 4.3.3
```

```
Warning: package 'ggplot2' was built under R version 4.3.3
```

```
Warning: package 'tibble' was built under R version 4.3.3
```

```
Warning: package 'tidyr' was built under R version 4.3.3
```

```
Warning: package 'readr' was built under R version 4.3.3
```

```
Warning: package 'purrr' was built under R version 4.3.3
```

```
Warning: package 'dplyr' was built under R version 4.3.3
```

```
Warning: package 'stringr' was built under R version 4.3.3
```

```
Warning: package 'forcats' was built under R version 4.3.3
```

```
Warning: package 'lubridate' was built under R version 4.3.3
```

```
-- 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 1.9.4     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(dplyr)
library(janitor)
```

```
Warning: package 'janitor' was built under R version 4.3.3
```

```
Attaching package: 'janitor'

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

    chisq.test, fisher.test
```

```r
library(ggplot2)
library(ggrepel)
```

```
Warning: package 'ggrepel' was built under R version 4.3.3
```

4. Data Cleaning and Transformation

4.1. Checking and Reshaping `trees_count.csv`

```r
# Load the data
trees_count <- data.frame(read_csv("trees_count.csv", show_col_types = FALSE))
# Reshape the data to a tidy format
trees_count <- pivot_longer(data = trees_count, cols = 2:ncol(trees_count),
                            names_to = 'neighbourhoods', values_to = 'count')
```

4.2. Checking and Reshaping `trees_height_diameter.csv`

```r
# Load the data
trees_h_d <- data.frame(read_csv("trees_height_diam.csv", show_col_types = FALSE))
# Reshape to wide format
trees_h_d <- pivot_wider(data = trees_h_d, names_from = 'attribute',
                         values_from = 'value')
```

4.3. Cleaning `street_trees.csv` Column Names

```r
# Load the data
street_trees <- data.frame(read_csv("street_trees.csv", show_col_types = FALSE))
# Rename columns for consistency
street_trees <- dplyr::rename(street_trees,
```

```
                              tree_id = Tree.ID,
                              street = Street.Name,
                              neighbourhood = Neighbourh,
                              species = SpeciesName,
                              common_name = CommonName,
                              height_rank = hrank,
                              diameter_in = Diameter,
                              year_planted = YearPlanted)
```

4.4. Handling Data Types and Unit Conversion

```
# Convert diameter from inches to cm
street_trees <- mutate(street_trees, diameter_cm = diameter_in * 2.54)
# Convert height rank to ordered factor
height_rank_unique <- unique(street_trees$height_rank)
height_rank_ordered <- sort(height_rank_unique)
street_trees$height_rank <- factor(street_trees$height_rank,
                                   levels = height_rank_ordered, ordered = TRUE)
```

4.5. Removing Unrealistic Tree Diameters

```
# Identify trees with unrealistic diameters
num_zero_diameter <- nrow(filter(street_trees,
                                 diameter_cm == 0))
num_large_diameter <- nrow(filter(street_trees,
                                  diameter_cm > 300))
# Filter to remove unrealistic values
trees_clean <- filter(street_trees, diameter_cm != 0 & diameter_cm <= 300)
```

5. Data Summarization

5.1. Overall Summary of Tree Diameters

```
# Compute summary statistics
trees_summary_all <- summarize(trees_clean,diameter_mean = mean(diameter_cm),
                               diameter_min = min(diameter_cm),
```

```
                              diameter_max = max(diameter_cm),
                              diameter_sd = sd(diameter_cm),n_obs = n())
```

5.2. Tree Summary by Species

```
# Grouped summary by species
trees_summary_sp <- trees_clean %>%
  group_by(common_name) %>%
  summarize(diameter_mean = mean(diameter_cm),
            diameter_min = min(diameter_cm),
            diameter_max = max(diameter_cm),
            diameter_sd = sd(diameter_cm),
            n_obs = n())
```

5.3. Top 5 Most Common Species

```
# Identify the top 5 species
top_5_sp <- trees_clean %>%
  count(common_name) %>%
  arrange(desc(n)) %>%
  slice_head(n = 5)
```

6. Discussion and Insights

1. The most common tree species in Vancouver is **Kwanzan Flowering Cherry**, followed by **Pissard Plum**.
2. **South Cambie** has the highest tree density, but it covers a small area, making it a local hotspot for street trees.
3. Neighborhoods with high tree densities include **Renfrew-Collingwood, Hastings-Sunrise, and Dunbar-Southlands**, each having over 100 trees per km of street.
4. **Downtown** has the smallest tree diameters on average, likely due to urban constraints and maintenance practices.
5. Some species, such as **Pyramidal European Hornbeam**, tend to have smaller tree diameters, while **Norway Maple** has trees with a broader range of diameters.

7. Conclusion

This analysis demonstrates how `tidyverse` can effectively process and analyze urban forestry data. Key findings include:

- **Tree diversity and distribution**: Vancouver's street trees are varied, with some neighborhoods having a higher density than others.
- **Data cleaning is crucial**: Identifying and removing incorrect data (such as zero-diameter trees) is necessary for accurate analysis.
- **Urban planning implications**: Understanding tree distribution helps in urban greening efforts and policy decisions.

Further analysis could involve spatial mapping of tree distributions, investigating the impact of tree age on diameter, and evaluating environmental benefits such as carbon sequestration and shade provision.

8. References

- Vancouver Open Data Portal: https://opendata.vancouver.ca/
- R Documentation: https://cran.r-project.org/web/packages/tidyverse/