DATA 602 Project: Exploring the City of Calgary’s Public Trees Dataset

Harneet Cheema, Michael Ellsworth

October 17, 2019

Introduction

Trees within municipalities have proven benefits. By improving the aesthetics of communities and filtering the air we breathe, trees bring a variety of quantifiable advantages to cities that are not typically studied.

In Calgary, there are approximately 7 million trees within city limits. What makes this statistic impressive is that Calgary’s arid, prairie type climate is not necessarily conducive to growing trees, let alone 7 million of them. This becomes a difficult challenge for the local government because in large part, trees are a publicly owned asset and the cost and resources associated with the maintenance of those trees fall on the municipality. Additionally, understanding the value of investing resources into trees is not particularly intuitive for citizens.

As pressure on the City to demonstrate fiscal restraint continues to increase, the budget for maintaining public trees is unlikely to grow. The purpose of this project is to take a data-driven approach to analyzing the City of Calgary’s public tree data to make recommendations that may assist in improved cost and resource allocation that allows the City to maintain or improve the condition of their public trees. This analysis may also assist in the future development of public trees in order to enhance the appropriate diversity of the tree population within Calgary.

The City of Calgary has a publicly available data set that provides a variety of information on Calgary’s trees including name, genus, mature size and condition. This is a structured data set with recorded information on almost 500,000 trees. This information was collected by trained Urban Forestry staff who visit each tree and populate the information accordingly.

Overall, the purpose of our analysis will be to statistically analyze publicly available tree information, derive meaningful trends and provide actionable insights. Findings will be used to make recommendations that could to assist the City with their public tree resource allocation.

Dataset

As mentioned previously, we will be using the Public Trees data set from the Open Calgary data portal. This data set was created on January 18th, 2018 and is updated on a weekly basis.

In this data set, there are over 496,000 observations and 20 variables. Each observation represents a single tree and each variable represents information about that specific tree. Each observation may have some missing information indicated in this data set by “NA” which will be excluded in our analysis.

The data is provided by the City of Calgary and contains information licensed under the Open Government License – City of Calgary. A link to the license agreement is provided in the References section. This license agreement grants us permission to use this data for our analysis.

Packages

The packages that will be used for this statistical analysis include the following:

- dplyr
- ggplot2
- readr
- stringr
- mosaic

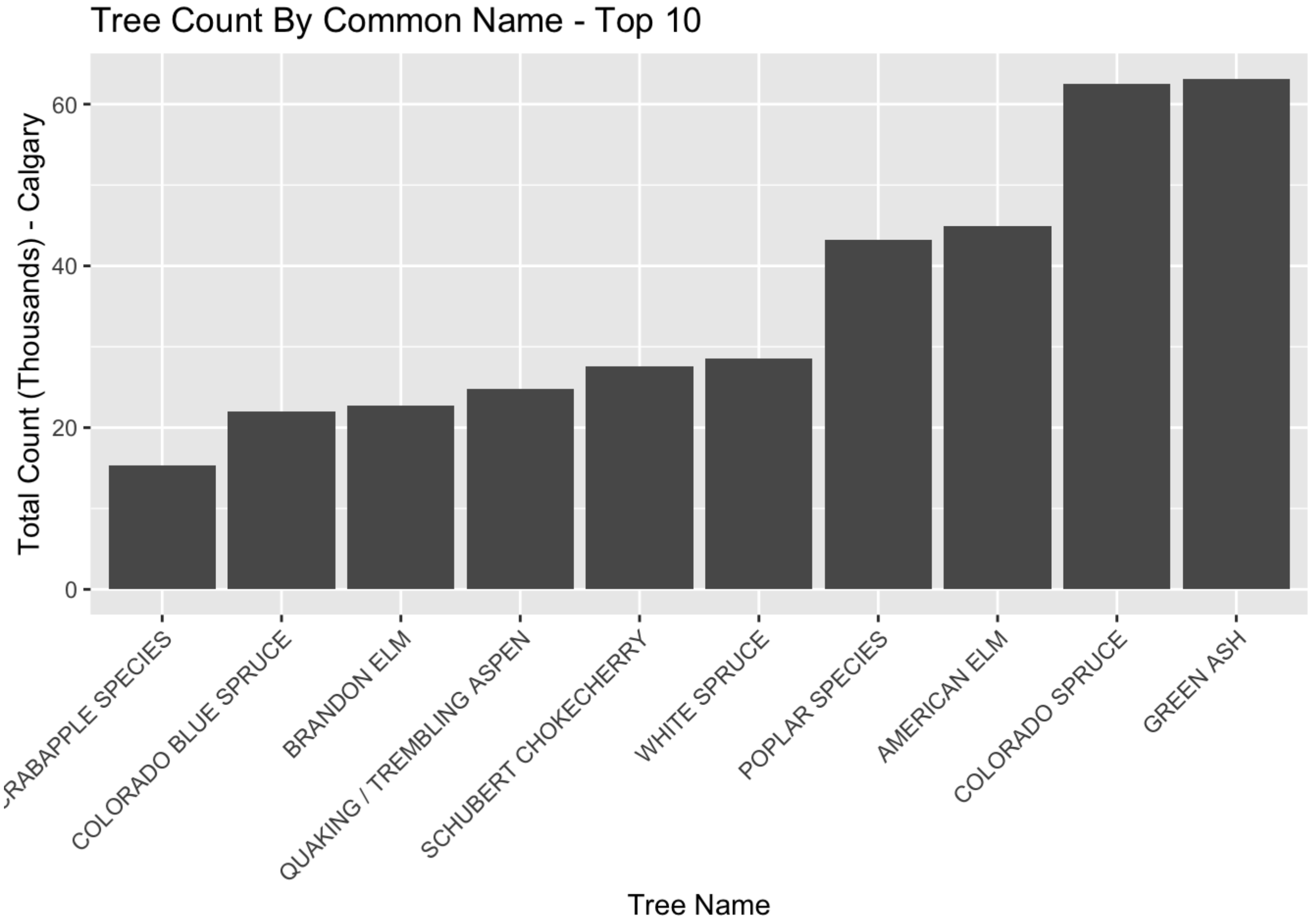
Statistical Analysis

In order to investigate how different variables affect tree condition, our group will estimate the difference between two population means using conventional techniques. This estimation of the difference between two population means will be used to compare the condition of trees of different tree names, Genus type and mature size. Ultimately, the goal of this statistical analysis will be to determine the variables that impact tree condition positively so that future tree development can focus on incorporating the variables that keep trees in good condition.

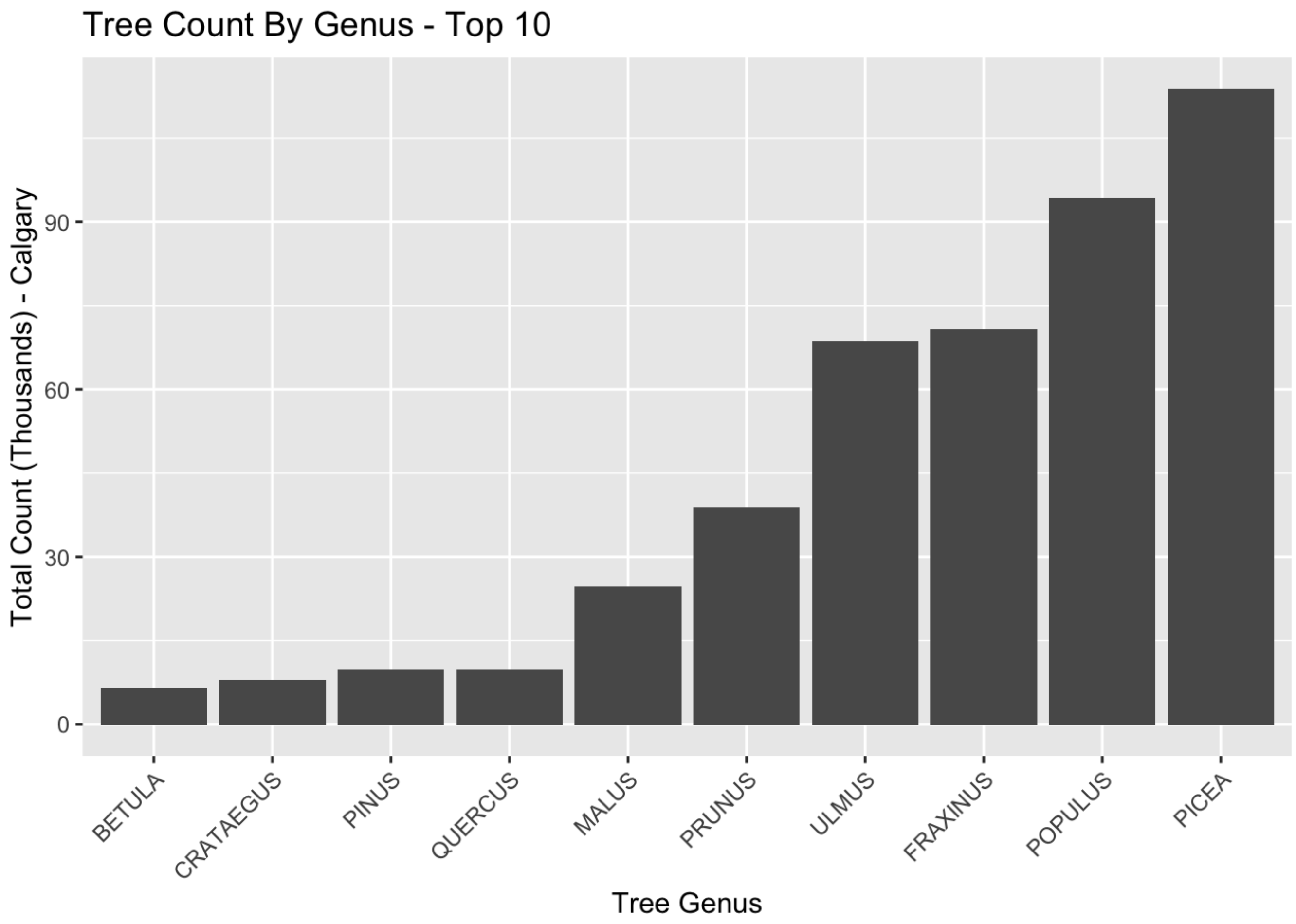
Additionally, we will attempt to estimate tree proportions in developing and developed communities by common name, genus type and mature size. This will be completed using bootstrap intervals. The goal of this statistical analysis will be to observe trends in tree planting preferences in newer communities.

Explore the data

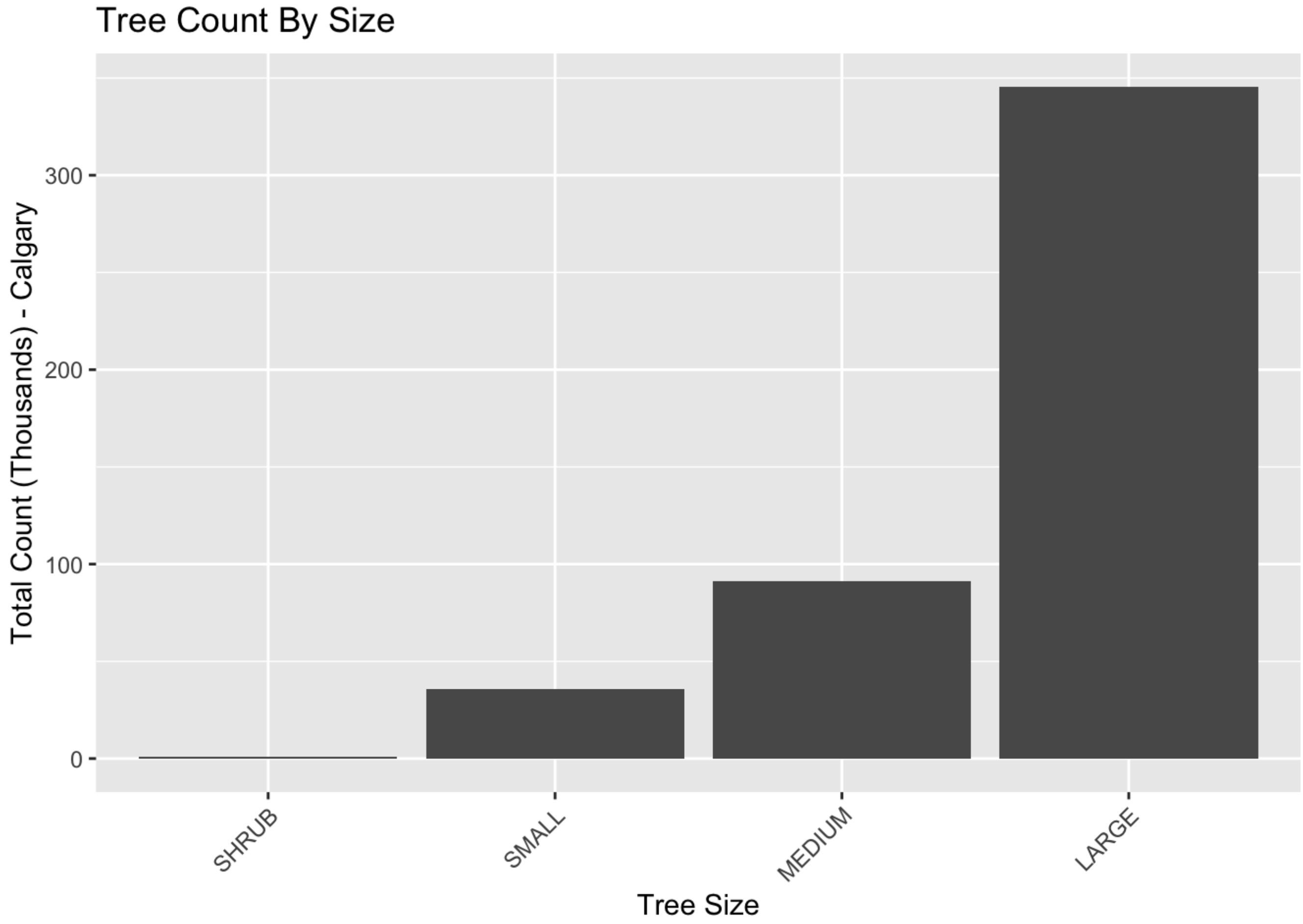
The first portion of this analysis focused on exploring the data available in the Public Trees data set. Since our focus is on comparing condition of different tree names, genus type and mature size, these features in the data set were explored.



Based on the above plot, we determined that Green Ash was the most common tree in Calgary by common name. This type of tree will be used in the statistical analysis.



Based on the above plot, we determined that Picea was the most common tree in Calgary by Genus type. This type of tree will also be used in the statistical analysis.



Based on the above plot, we determined that Large trees were most common in Calgary followed by Medium and Small sized trees. To capture the ends of the spectrum, we will be comparing Large and Small sized trees in our statistical analysis.

Difference of Tree Condition Means - Green Ash and Other

In the first conventional confidence interval estimation technique, the group was looking to estimate if there was a difference in the mean Tree Condition of Green Ash trees (most popular name) and Other tree types.

The statistical hypothesis below is based on the initial that most popular tree in Calgary would have a better tree condition than other trees in the city. Before completing the analysis, our assumption was based on the idea that since Green Ash is the most popular tree in Calgary it must have higher condition maintenance rating that other type of trees in the city.

Please see the statistical hypothesis below:

$$\text{Stastical Hypothesis}$$

$$\begin{aligned} H_0 &: \quad \mu_{\text{Tree Condition}(GreenAsh)} \leq \mu_{\text{Tree Condition}(Other)} \\ H_A &: \quad \mu_{\text{Tree Condition}(GreenAsh)} > \mu_{\text{Tree Condition}(Other)} \end{aligned}$$

After setting up the hypothesis above, we started the analysis process. A t-test was used to find the confidence interval, p-value and means of Green Ash and Other trees. See the results below from the t test.

```
## 
##  Welch Two Sample t-test
## 
## data:  TREE_CONDITION_RATING_PERC by COMMON_NAME
## t = -5.7177, df = 94963, p-value = 1
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.3120178        Inf
## sample estimates:
## mean in group GREEN ASH     mean in group OTHER 
##                60.65806                60.90037
```

Following were the findings from the analysis above:

- Difference between mean values of Green Ash and other tree types was marginal.
- P-value was equal to 1.
- Since this was greater than 0.05, we failed to reject our null hypothesis
- This revealed than mean tree condition of Green Ash trees was lower than other tree types

Difference of Tree Condition Means - Picea and Other

Conventional confidence interval estimation was used again to estimate mean difference between most common genus type (Picea) and Other Genus types. The motivation behind this analysis was to determine why this tree genus type was the most common in Calgary.

Statistical hypothesis below assumes that most Picea genus type is the most common type of tree in Calgary because it has a higher condition rating than other type of trees. Since it has higher condition maintenance rating it is easier to maintain which results in it being the most popular tree type in Calgary.

Please see the statistical hypothesis below:

$$\text{Stastical Hypothesis}$$

$$\begin{aligned} H_0 &: \quad \mu_{\text{Tree Condition}(Picea)} \leq \mu_{\text{Tree Condition}(Other)} \\ H_A &: \quad \mu_{\text{Tree Condition}(Picea)} > \mu_{\text{Tree Condition}(Other)} \end{aligned}$$

T test revealed the following results upon completion of the analysis:

```
## 
##  Welch Two Sample t-test
## 
## data:  TREE_CONDITION_RATING_PERC by GENUS
## t = -41.57, df = 197400, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -1.521122
## sample estimates:
## mean in group OTHER mean in group PICEA 
##            60.48765            62.07144
```

Following were the findings from the analysis above:

- Difference between mean values of Picea and other genus type was evident
- P-Value was determined to be very close to 0
- Since P-value was very small we reject our null hypothesis in favour of alternative hypothesis
- This test revealed than the mean tree condition of Picea was indeed greater than mean tree condition of other tree type. Therefore, our initial assumptions were found to be valid through this test.

Difference of Tree Condition Means - Large and Small trees

Conventional confidence interval estimation method was utilized below to estimate mean difference between large size and small size trees in Calgary. Motivation behind this analysis was to compare average tree condition between large and small trees and determine if the city should be planning certain size of trees more than the others.

Statistical hypothesis below assumed that due to size difference larger trees will have lower condition maintenance than the smaller trees due to poor weather conditions.

Please see the statistical hypothesis below:

$$\text{Stastical Hypothesis}$$

$$\begin{aligned} \text{H}_0: &\quad \mu_{\text{Tree Condition}(Large)} \geq \mu_{\text{Tree Condition}(Small)} \\ \text{H}_A: &\quad \mu_{\text{Tree Condition}(Large)} < \mu_{\text{Tree Condition}(Small)} \end{aligned}$$

T test revealed the following results upon completion of the analysis:

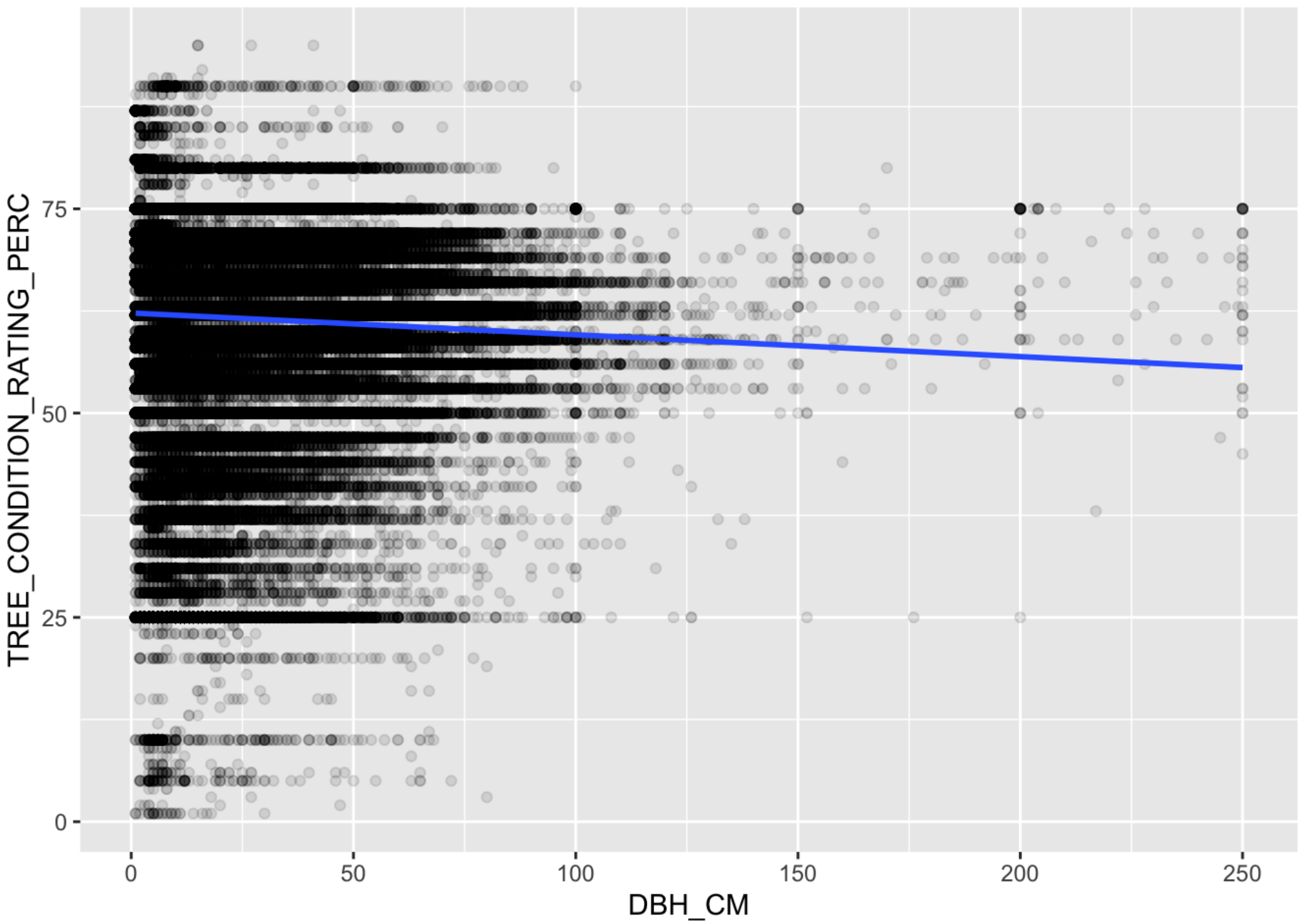
```
## 
##  Welch Two Sample t-test
## 
## data:  TREE_CONDITION_RATING_PERC by MATURE_SIZE
## t = -1.8853, df = 40574, p-value = 0.0297
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf -0.01564818
## sample estimates:
## mean in group LARGE mean in group SMALL 
##            60.94202            61.06476 
```

Following were the findings from the analysis above:

- Difference between mean tree conditions large and small trees was marginal
- The P-value was determined to be 0.0297
- Since p-value was less than 0.05 we reject our null hypothesis in favour of alternative hypothesis
- Therefore, our initial assumption was found to be valid as test revealed that average tree condition of smaller trees was indeed greater than larger trees

By plotting the tree condition data versus the diameter of the trunk, we also observed a negative linear trend as trunk diameter increased, which helps confirm our conclusion that small trees had a higher average tree condition. With that being said, the coefficient of determination is fairly small with only 0.2% of the data being explained by this negative linear trend.

Tree Condition vs Size (Diameter of Trunk)



```
## (Intercept)      DBH_CM
## 62.25473402 -0.02672251
```

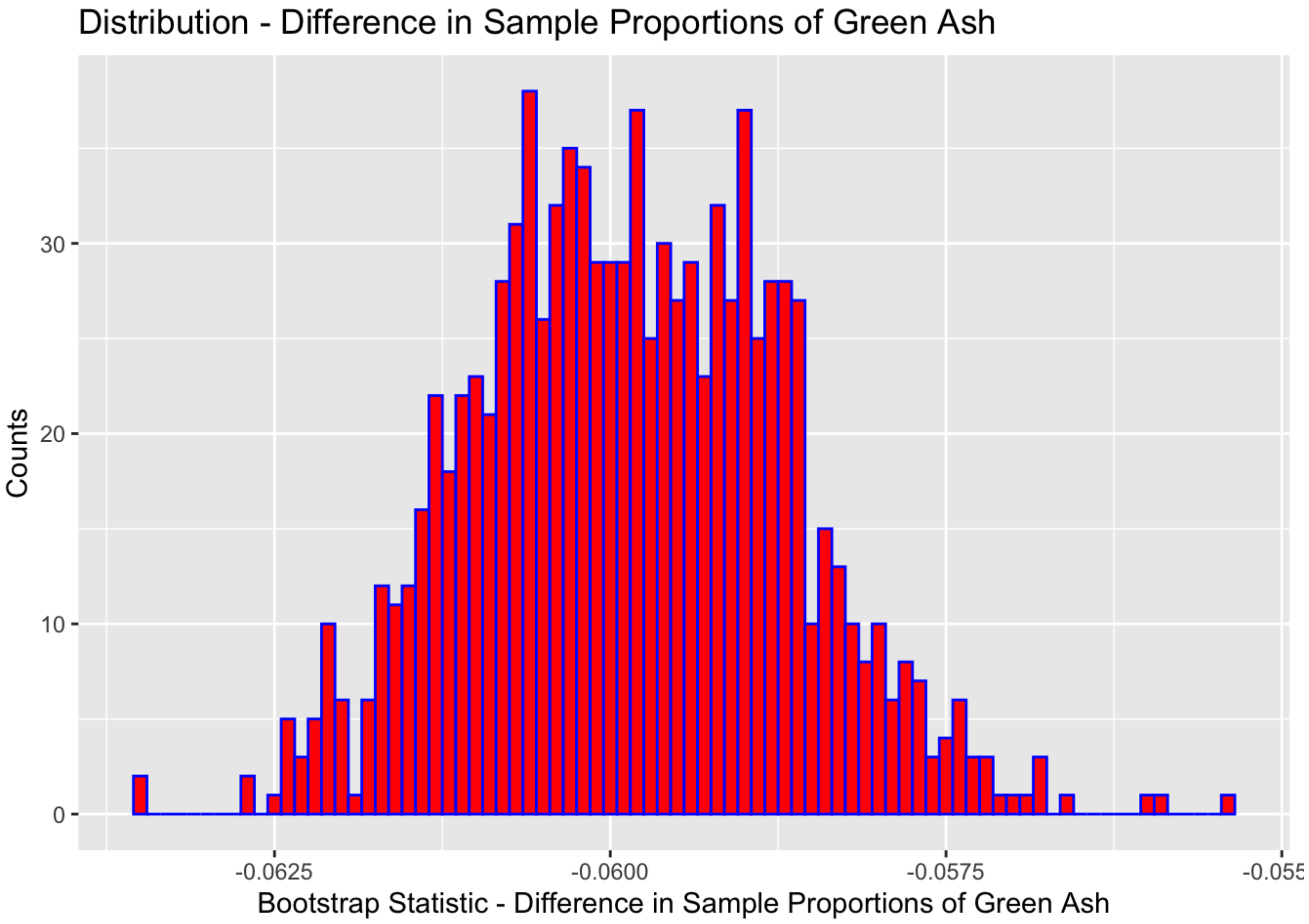
```
## [1] 0.002649334
```

Green Ash Proportion

Based on the previous statistical analysis comparing the average tree condition of Green Ash versus other types of trees, we determined with confidence that Green Ash had on average a worse tree condition than the average tree condition of other trees combined. Next, we would like to determine if the City of Calgary is aware of this condition and is planting fewer Green Ash trees in developing communities. In order to complete this statistical analysis, we utilized a proportional comparison using bootstrap intervals. In this analysis, we took the difference of the proportion of Green Ash in developing and developed communities.

$$p_{\text{Green Ash - Developing}} - p_{\text{Green Ash - Developed}}$$

Green Ash Proportion Differences



From this analysis, we determined that there was a clear difference between the Green Ash proportion in developing and developed communities. From the data, it was clear that fewer Green Ash trees are being planted in developing communities. Although there is no available data in this data set to suggest that the city is planting fewer Green Ash because the average tree condition is relatively poorer than other trees, it is plausible that this is the reason.

min	Q1	median	Q3	max	mean	sd	n	missing
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
-0.06352749	-0.06067582	-0.05990438	-0.05902454	-0.05537354	-0.0598547	0.001159649	1000	0

1 row

	quantile	p
	<dbl>	<dbl>
2.5%	-0.06209885	0.025
97.5%	-0.05754226	0.975

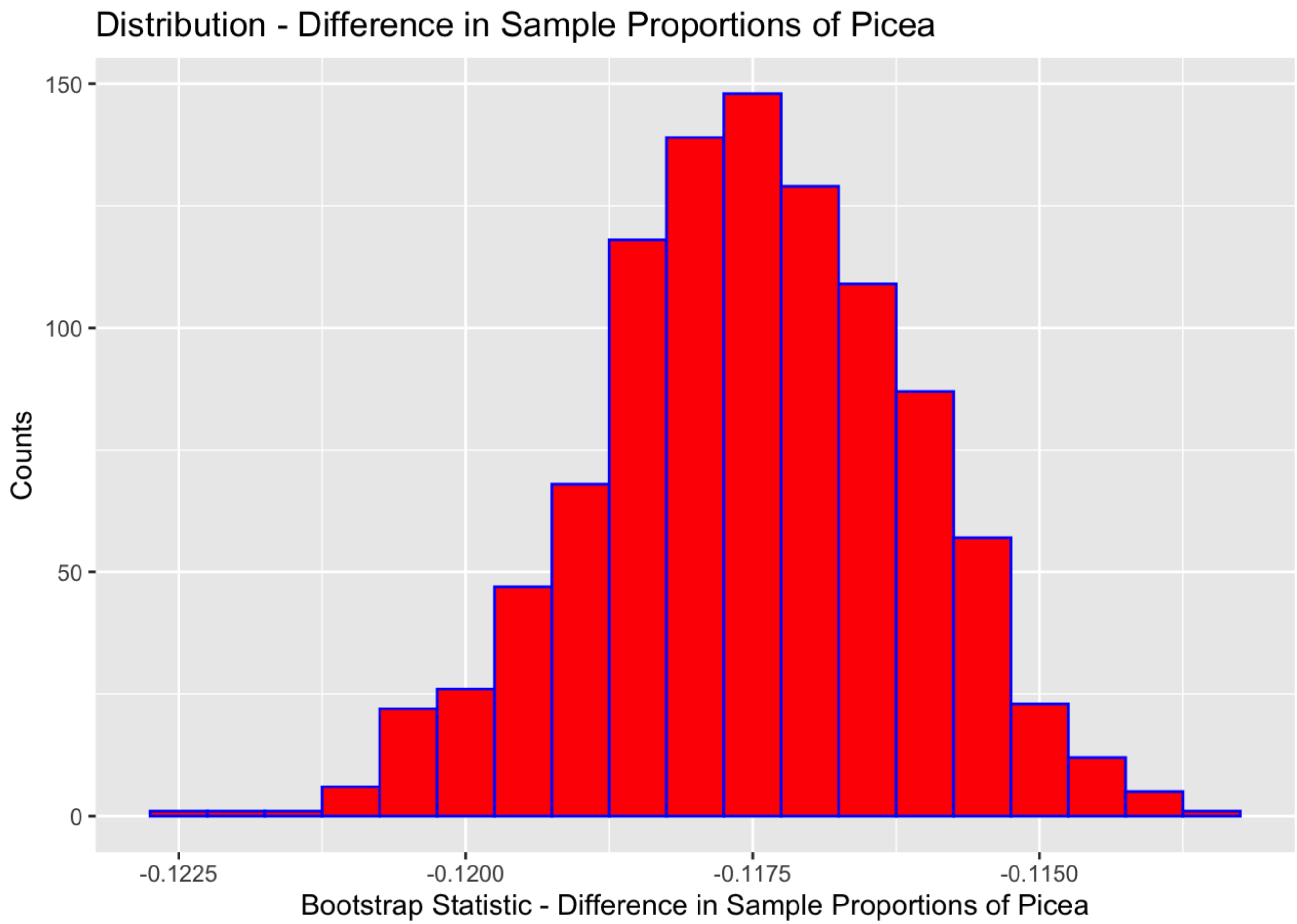
2 rows

Picea Proportion

As there are quite a few different types of names in the Calgary Public Trees data set, we decided that some of the small variations in the names might affect the quality in the data and average conditions may be affected as a result. Based on our data exploration, Genus type appeared to have smaller variation and would present a better statistical analysis.

Based on the previous statistical analysis comparing the average tree condition of Picea versus other Genus types, we determined with confidence that Picea had on average a better tree condition than the average tree condition of other Genus types combined. Next, we would like to determine if the City of Calgary is aware of this condition and is planting more Picea trees in developing communities. In order to complete this statistical analysis, we utilized a proportional comparison using bootstrap intervals. In this analysis, we took the difference of the proportion of Picea in developing and developed communities.

$$p_{\text{Picea - Developing}} - p_{\text{Picea - Developed}}$$



As evidenced on the chart above which displays the differences in proportions of small trees in developed and developing communities, we determined that there was a clear difference between the Picea proportion in developing and developed communities. It was also clear that fewer Picea trees are being planted in developing communities. As Picea has a higher average tree condition relative to other trees in Calgary, it might be better for the City of Calgary to plant more Picea trees in order to take advantage of the benefits of having a better condition of tree.

min	Q1	median	Q3	max	mean	sd	n	missing
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
-0.1225274	-0.1183815	-0.1174794	-0.116523	-0.1137228	-0.1175141	0.001369306	1000	0

1 row

	quantile	p
	<dbl>	<dbl>
2.5%	-0.1203249	0.025
97.5%	-0.1149000	0.975

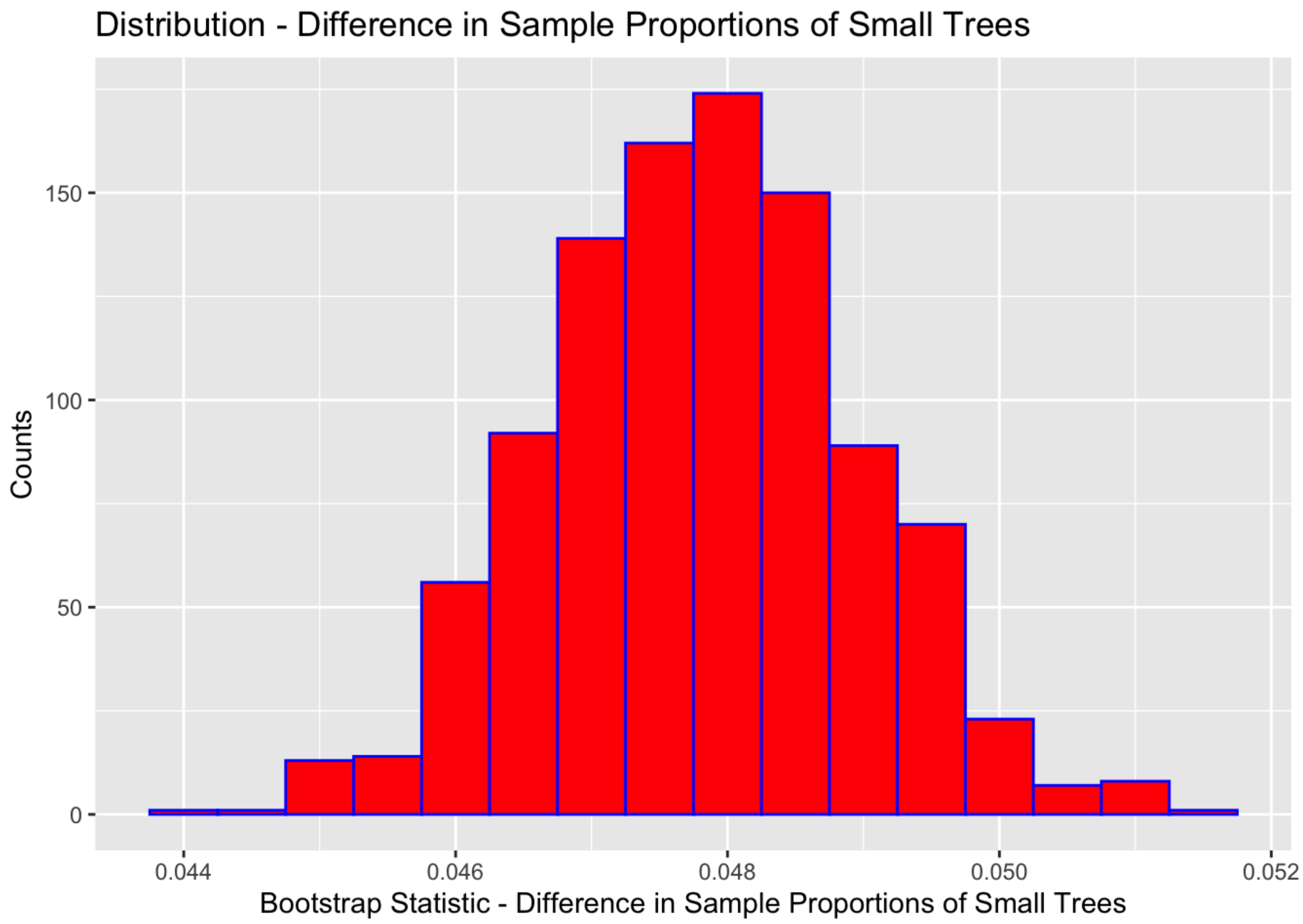
2 rows

Mature Size Proportion

The last portion of the proportional analyses was comparing the proportions of small trees in developing and developed communities. As small trees had a better relative tree condition than larger trees, we isolated this analysis on small trees to see if more small trees were being planted in developing communities.

$$p_{\text{Small - Developing}} - p_{\text{Small - Developed}}$$

Small Proportion Differences



Based on the chart above, it was clear that there was a difference between these two proportions and that more small trees are being planted in developing communities. In order to complete this analysis, it is recommended that future analysis also incorporate trends in medium sized trees.

min <dbl>	Q1 <dbl>	median <dbl>	Q3 <dbl>	max <dbl>	mean <dbl>	sd <dbl>	n <int>	missing <int>
0.04375849	0.04703465	0.04780613	0.04855278	0.05153706	0.04780516	0.001133147	1000	0

1 row

quantile <dbl>	p <dbl>

	quantile	p
	<dbl>	<dbl>
2.5%	0.04571255	0.025
97.5%	0.04993447	0.975

2 rows

Conclusions

The statistical analysis outlined in this report gave us an understanding of tree condition over a variety of variables including common name, genus type and mature size. Additionally, we were able to determine tree planting trends by comparing tree proportions in developing and developed communities.

Of all trees in Calgary, the most popular tree based on common name, the Green Ash, had a lower average tree condition than the average tree condition of all other trees combined. The Green Ash, potentially due its average tree condition, had a lower proportion in developing communities compared with developed communities, suggesting that the City is moving away from planting these trees. Based on our analysis, this is a positive trend and one we would recommend continue.

Based on Genus type, Picea had a higher average tree condition than the average tree condition of all other trees combined. Although Picea would be a recommended Genus type for future tree planting in Calgary, this was not the trend that we saw from the proportions of Picea in developing and developed communities.

Lastly, we found that small trees had a higher average tree condition than large trees and were being planted more frequently in developing communities compared with developed communities. We believe based on our analysis that this is a good trend.

References

1 City of Calgary (2019) Public Trees [Online]. Available at: https://data.calgary.ca/Environment/Public-Trees/tfs4-3wwa (https://data.calgary.ca/Environment/Public-Trees/tfs4-3wwa) (Accessed: 28 September 2019) Community Points

2 City of Calgary (2019) Community Points [Online]. Available at: https://data.calgary.ca/Base-Maps/Community-Points/j9ps-fyst (https://data.calgary.ca/Base-Maps/Community-Points/j9ps-fyst) (Accessed: 28 September 2019)