Lab 1 - Pittsburgh Trees

Trees cared for and managed by the City of Pittsburgh Department of Public Works Forestry Division.

Source: City of Pittsburgh

Setup

In [75]:

```
# Install
# install.packages("dplyr")
# install.packages("tidyverse")
# install.packages("dslabs")
# install.packages("vtreat")
```

In [3]:

```
# Libraries
library(dplyr)
library(tidyverse)
library(dslabs)
library(vtreat)
```

Read in Data

In [5]:

```
trees_raw <- read_csv('../datasets/pittsburgh_trees.csv', col_types = cols(.

head(trees_raw)
```

```
Warning message:
“One or more parsing issues, call `problems()` on your data frame for
details, e.g.:
  dat <- vroom(...)
  problems(dat)”
```

_id	id	address_number	street	common_name	scientific_name	he
<dbl>	<dbl>	<dbl>	<chr>	<chr>	<chr>	<
1	754166088	7428	MONTICELLO ST	Stump	Stump	
2	1946899269	220	BALVER AVE	Linden: Littleleaf	Tilia cordata	
3	1431517397	2822	SIDNEY ST	Maple: Red	Acer rubrum	
4	994063598	608	SUISMON ST	Maple: Freeman	Acer x freemanii	
5	1591838573	1135	N NEGLEY AVE	Maple: Norway	Acer platanoides	
6	1333224197	5550	BRYANT ST	Oak: Pin	Quercus palustris	

In [6]:

```
colnames(trees_raw)
```

```
'_id' · 'id' · 'address_number' · 'street' · 'common_name' · 'scientific_name' · 'height' ·
'width' · 'growth_space_length' · 'growth_space_width' · 'growth_space_type' ·
'diameter_base_height' · 'stems' · 'overhead_utilities' · 'land_use' · 'condition' ·
'stormwater_benefits_dollar_value' · 'stormwater_benefits_runoff_elim' ·
'property_value_benefits_dollarvalue' · 'property_value_benefits_leaf_surface_area' ·
'energy_benefits_electricity_dollar_value' · 'energy_benefits_gas_dollar_value' ·
'air_quality_benfits_o3dep_dollar_value' · 'air_quality_benfits_o3dep_lbs' ·
'air_quality_benfits_vocavd_dollar_value' · 'air_quality_benfits_vocavd_lbs' ·
'air_quality_benfits_no2dep_dollar_value' · 'air_quality_benfits_no2dep_lbs' ·
'air_quality_benfits_no2avd_dollar_value' · 'air_quality_benfits_no2avd_lbs' ·
'air_quality_benfits_so2dep_dollar_value' · 'air_quality_benfits_so2dep_lbs' ·
'air_quality_benfits_so2avd_dollar_value' · 'air_quality_benfits_so2avd_lbs' ·
'air_quality_benfits_pm10depdollar_value' · 'air_quality_benfits_pm10dep_lbs' ·
'air_quality_benfits_pm10avd_dollar_value' · 'air_quality_benfits_pm10avd_lbs' ·
'air_quality_benfits_total_dollar_value' · 'air_quality_benfits_total_lbs' ·
'co2_benefits_dollar_value' · 'co2_benefits_sequestered_lbs' ·
'co2_benefits_sequestered_value' · 'co2_benefits_avoided_lbs' ·
'co2_benefits_avoided_value' · 'co2_benefits_decomp_lbs' · 'co2_benefits_maint_lbs' ·
'co2_benefits_totalco2_lbs' · 'overall_benefits_dollar_value' · 'neighborhood' ·
'council_district' · 'ward' · 'tract' · 'public_works_division' · 'pli_division' · 'police_zone' ·
'fire_zone' · 'latitude' · 'longitude'
```

In [7]:

```
problems(trees_raw)
```

A tibble: 8 × 5

row	col	expected	actual	file
<int>	<int>	<chr>	<chr>	<chr>
45296	3	a double	1200 Diana	/Users/isaacbraun/personal/data-analytics/datasets/pittsburgh_trees.csv
45310	3	a double	1402 w north ave	/Users/isaacbraun/personal/data-analytics/datasets/pittsburgh_trees.csv
45324	3	a double	18 sprain st	/Users/isaacbraun/personal/data-analytics/datasets/pittsburgh_trees.csv
45325	3	a double	18 sprain st	/Users/isaacbraun/personal/data-analytics/datasets/pittsburgh_trees.csv
45326	3	a double	502 Foreland	/Users/isaacbraun/personal/data-analytics/datasets/pittsburgh_trees.csv
45327	3	a double	502 Foreland st	/Users/isaacbraun/personal/data-analytics/datasets/pittsburgh_trees.csv
45335	3	a double	345 dalton ave	/Users/isaacbraun/personal/data-analytics/datasets/pittsburgh_trees.csv
45706	3	a double	499 N LANG AVE	/Users/isaacbraun/personal/data-analytics/datasets/pittsburgh_trees.csv

In [8]:

```
# New df with limited columns
trees <- trees_raw %>% select('id', 'common_name', 'height', 'width', 'growt
                              'growth_space_type', 'diameter_base_height', '
                              'condition', 'stormwater_benefits_dollar_value
```

In [9]:

```
head(trees)
```

id	common_name	height	width	growth_space_length	growth_space_width
<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
754166088	Stump	0	0	10	2
1946899269	Linden: Littleleaf	0	0	99	99
1431517397	Maple: Red	22	6	6	3
994063598	Maple: Freeman	25	10	3	3
1591838573	Maple: Norway	52	13	99	99
1333224197	Oak: Pin	45	18	35	3

Summarize #0 - Counts of Species

In [26]:

```
trees %>%
    count(common_name) %>%
    arrange(desc(n)) %>%
    head(10)
```

A tibble: 10 × 2

common_name	n
<chr>	<int>
Maple: Norway	3717
Maple: Red	3422
London planetree	3238
Pear: Callery	2969
Vacant Site Small	2419
Linden: Littleleaf	2413
Honeylocust: Thornless	2019
Oak: Pin	1672
Crabapple: Flowering	1310
Ginkgo	1218

Summarize #1 - Group By

Grouping the trees by Common Name to find average height/width/stems.

In [11]:

```
species_averages <- trees %>%
  group_by(common_name) %>%
  summarize(height_avg = mean(height), width_avg = mean(width), stems_avg =
  arrange(desc(height_avg))

head(species_averages)
```

A tibble: 6 × 4

common_name	height_avg	width_avg	stems_avg
<chr>	<dbl>	<dbl>	<dbl>
Cottonwood: Eastern	52.80000	14.700000	1.900000
Butternut	45.83333	9.833333	1.000000
Poplar: White	40.00000	10.000000	2.000000
Willow: Black	39.60000	8.800000	1.000000
Hickory: Bitternut	36.42857	9.000000	1.142857
Maple: Silver	36.01852	NA	1.418981

Summary #2 - Summary

Getting the summaries for Eastern Cottonwood and English Walnut

In [12]:

```
cottonwood <- trees %>%
    filter(common_name == "Cottonwood: Eastern") %>%
    select(-id, -common_name, -overhead_utilities, -land_use, -condition, -p

walnut_english <- trees %>%
    filter(common_name == "Walnut: English") %>%
    select(-id, -common_name, -overhead_utilities, -land_use, -condition, -p

cottonwood %>% summary()
walnut_english %>% summary()
```

```
    height          width       growth_space_length growth_space_width
Min.   :22.00   Min.   : 1.00   Min.   : 5.0        Min.   : 2.0
1st Qu.:34.50   1st Qu.: 7.25   1st Qu.:99.0        1st Qu.: 3.0
Median :60.00   Median :10.00   Median :99.0        Median :99.0
Mean   :52.80   Mean   :14.70   Mean   :80.2        Mean   :60.5
3rd Qu.:68.75   3rd Qu.:23.75   3rd Qu.:99.0        3rd Qu.:99.0
Max.   :80.00   Max.   :35.00   Max.   :99.0        Max.   :99.0
growth_space_type  diameter_base_height     stems
Length:10          Min.   : 3.00        Min.   :1.00
Class :character   1st Qu.:10.00        1st Qu.:1.00
Mode  :character   Median :19.50        Median :1.00
                   Mean   :18.70        Mean   :1.90
                   3rd Qu.:27.75        3rd Qu.:1.75
                   Max.   :32.00        Max.   :6.00
stormwater_benefits_dollar_value property_value_benefits_dollarvalue
Min.   : 1.552                   Min.   : 56.15
1st Qu.: 7.262                   1st Qu.: 80.25
Median :21.456                   Median :111.37
Mean   :20.171                   Mean   : 97.19
3rd Qu.:32.080                   3rd Qu.:114.99
Max.   :37.794                   Max.   :116.93
neighborhood
Length:10
Class :character
Mode  :character
```

```
    height         width      growth_space_length growth_space_width
Min.   :25.0   Min.   :6.00   Min.   :99          Min.   :99
1st Qu.:25.0   1st Qu.:6.75   1st Qu.:99          1st Qu.:99
Median :27.5   Median :7.50   Median :99          Median :99
Mean   :27.5   Mean   :7.25   Mean   :99          Mean   :99
3rd Qu.:30.0   3rd Qu.:8.00   3rd Qu.:99          3rd Qu.:99
Max.   :30.0   Max.   :8.00   Max.   :99          Max.   :99
growth_space_type  diameter_base_height     stems
Length:4           Min.   :7            Min.   :1.0
Class :character   1st Qu.:7            1st Qu.:1.0
Mode  :character   Median :7            Median :1.5
                   Mean   :7            Mean   :1.5
                   3rd Qu.:7            3rd Qu.:2.0
                   Max.   :7            Max.   :2.0
stormwater_benefits_dollar_value property_value_benefits_dollarvalue
Min.   :3.566                    Min.   :76.08
1st Qu.:4.471                    1st Qu.:76.08
Median :4.773                    Median :76.08
Mean   :4.595                    Mean   :76.08
3rd Qu.:4.896                    3rd Qu.:76.08
Max.   :5.267                    Max.   :76.08
neighborhood
Length:4
Class :character
Mode  :character
```

Summary #3 - Arrange

Arrange by Growth Space Length and then by Growth Space Width. May be useful to find trees that have the most room to grow, etc.

In [13]:

```
growth_space <- trees %>%
    arrange(desc(growth_space_length), desc(growth_space_width))

head(growth_space)
```

id	common_name	height	width	growth_space_length	growth_space_width
<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
928337304	Hornbeam: American	23	6	188	27
1858158720	Hornbeam: American	23	6	188	23
1092102844	Hornbeam: American	21	6	188	23
95131321	Stump	0	0	175	3
154608906	Maple: Norway	47	14	135	3
1372689231	Oak: Pin	45	8	130	3

Mutuate: extend with calculated column

Calculate area of available growth space.

In [14]:

```
trees <- mutate(trees, growth_space_area = growth_space_length * growth_spac

head(trees) %>% select(common_name, growth_space_length, growth_space_width,
```

A tibble: 6 × 5

common_name	growth_space_length	growth_space_width	growth_space_area	growth_
<chr>	<dbl>	<dbl>	<dbl>	
Stump	10	2	20	
Linden: Littleleaf	99	99	9801	Open or
Maple: Red	6	3	18	
Maple: Freeman	3	3	9	
Maple: Norway	99	99	9801	Open or
Oak: Pin	35	3	105	

Clean with vtreat

In [30]:
```
varlist <- colnames(trees)

treated <- design_missingness_treatment(trees, varlist = varlist)
training_prepared <- prepare(treated, trees)
```

In [31]:
```
colnames(training_prepared)
head(training_prepared)
```

'id' · 'common_name' · 'height' · 'height_isBAD' · 'width' · 'width_isBAD' ·
'growth_space_length' · 'growth_space_length_isBAD' · 'growth_space_width' ·
'growth_space_width_isBAD' · 'growth_space_type' · 'diameter_base_height' ·
'diameter_base_height_isBAD' · 'stems' · 'stems_isBAD' · 'overhead_utilities' · 'land_use' ·
'condition' · 'stormwater_benefits_dollar_value' ·
'stormwater_benefits_dollar_value_isBAD' · 'property_value_benefits_dollarvalue' ·
'property_value_benefits_dollarvalue_isBAD' · 'neighborhood' · 'police_zone' ·
'police_zone_isBAD' · 'fire_zone' · 'growth_space_area' · 'growth_space_area_isBAD'

id	common_name	height	height_isBAD	width	width_isBAD	growth_space_
<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	
754166088	Stump	0	0	0	0	
1946899269	Linden: Littleleaf	0	0	0	0	
1431517397	Maple: Red	22	0	6	0	
994063598	Maple: Freeman	25	0	10	0	
1591838573	Maple: Norway	52	0	13	0	
1333224197	Oak: Pin	45	0	18	0	

In [17]:

```
# Check NA replacements
height_missing <- which(is.na(trees$height))

trees[height_missing, c('common_name', 'height', 'width', 'growth_space_area
```

A tibble: 4374 × 4

common_name	height	width	growth_space_area
<chr>	<dbl>	<dbl>	<dbl>
Maple: Norway	NA	NA	NA
Vacant Site Not Suitable	NA	NA	NA
Vacant Site Small	NA	NA	NA
Vacant Site Not Suitable	NA	NA	NA
Vacant Site Not Suitable	NA	NA	NA
Maple: Red	NA	NA	NA
Vacant Site Medium	NA	NA	NA
Vacant Site Medium	NA	NA	NA
Vacant Site Small	NA	NA	NA
London planetree	NA	NA	NA
Vacant Site Small	NA	NA	NA
Vacant Site Not Suitable	NA	NA	NA
Vacant Site Not Suitable	NA	NA	NA
Vacant Site Not Suitable	NA	NA	NA
Vacant Site Not Suitable	NA	NA	NA
Vacant Site Not Suitable	NA	NA	NA
Vacant Site Not Suitable	NA	NA	NA
Vacant Site Not Suitable	NA	NA	NA
Vacant Site Not Suitable	NA	NA	NA
Vacant Site Not Suitable	NA	NA	NA
Vacant Site Not Suitable	NA	NA	NA
Vacant Site Not Suitable	NA	NA	NA
Vacant Site Not Suitable	NA	NA	NA
Vacant Site Not Suitable	NA	NA	NA
Vacant Site Not Suitable	NA	NA	NA
Vacant Site Not Suitable	NA	NA	NA
Vacant Site Not Suitable	NA	NA	NA
Vacant Site Not Suitable	NA	NA	NA
Vacant Site Not Suitable	NA	NA	NA
Vacant Site Not Suitable	NA	NA	NA

common_name	height	width	growth_space_area
<chr>	<dbl>	<dbl>	<dbl>
⋮	⋮	⋮	⋮
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA
Ash: Green	NA	NA	NA
NA	NA	NA	NA
Linden: Littleleaf	NA	NA	NA
Linden: Littleleaf	NA	NA	NA
Maple: Norway	NA	NA	NA
Maple: Sugar	NA	NA	NA
Maple: Sugar	NA	NA	NA
Maple: Sugar	NA	NA	NA
Maple: Sugar	NA	NA	NA

Plot: Property Value Benefits Distribution by Land Use

In [37]:

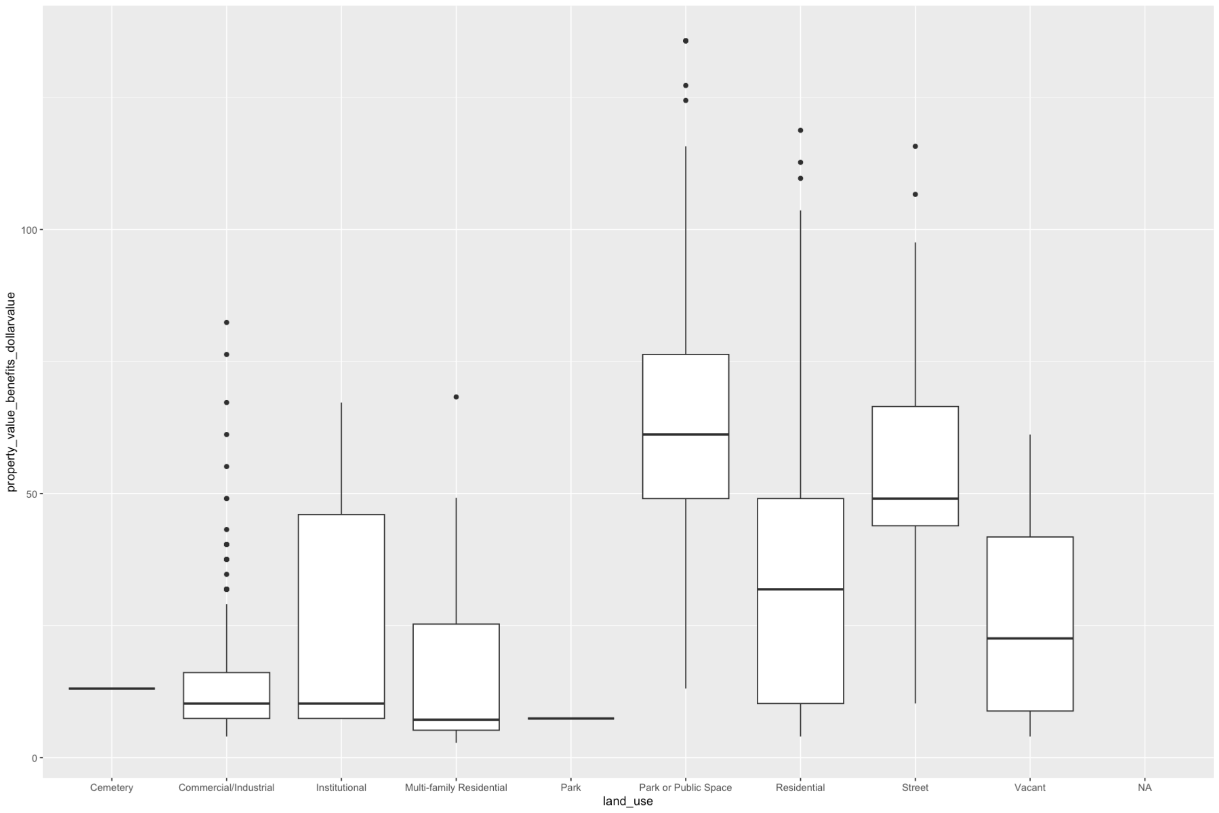
```
red_maple <- trees %>% filter(common_name == "Ginkgo" & growth_space_area <

# Make Land Use a Factor
red_maple$land_use <- as.factor(red_maple$land_use)
head(red_maple)

# Increase plot size
options(repr.plot.width=15, repr.plot.height=10)
# Create Box Plot
ggplot(red_maple, aes(x = land_use, y = property_value_benefits_dollarvalue)
```

id	common_name	height	width	growth_space_length	growth_space_width
<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1492169209	Ginkgo	40	7	2	2
272719655	Ginkgo	14	6	12	4
523208226	Ginkgo	15	6	10	4
1049586714	Ginkgo	12	6	10	4
734654174	Ginkgo	0	0	8	4
311777126	Ginkgo	7	0	10	3

```
Warning message:
“Removed 13 rows containing non-finite outside the scale range
(`stat_boxplot()`).”
```


property_value_benefits_dollarvalue
100
50
0
Cemetery
Commercial/Industrial
Institutional
Multi-family Residential
Park
Park or Public Space
Residential
Street
Vacant
NA
land_use

```
In [ ]:
```