Lab 2 - Reshaping Data

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

Source: City of Pittsburgh

Setup & Read Data

In [2]:

```
library(lubridate)
library(dplyr)
library(tidyverse)
library(dslabs)
library(data.table)
```

In [3]:

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

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	height	width	growth_space_length
<dbl>	<dbl>	<dbl>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>
1	754166088	7428	MONTICELLO ST	Stump	Stump	0	0	10
2	1946899269	220	BALVER AVE	Linden: Littleleaf	Tilia cordata	0	0	99
3	1431517397	2822	SIDNEY ST	Maple: Red	Acer rubrum	22	6	6
4	994063598	608	SUISMON ST	Maple: Freeman	Acer x freemanii	25	10	3
5	1591838573	1135	N NEGLEY AVE	Maple: Norway	Acer platanoides	52	13	99
6	1333224197	5550	BRYANT ST	Oak: Pin	Quercus palustris	45	18	35

In [15]:

```r
# New df with limited columns
trees <- trees_raw %>% select('id', 'common_name', 'height', 'width', 'growth_space_length', 'growth_space_
                              'growth_space_type', 'diameter_base_height', 'stems', 'overhead_utilities',
                              'condition', 'stormwater_benefits_dollar_value', 'property_value_benefits_do
head(trees)
```

id	common_name	height	width	growth_space_length	growth_space_width	growth_space_type	diameter_bas
<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	
754166088	Stump	0	0	10	2	Well or Pit	
1946899269	Linden: Littleleaf	0	0	99	99	Open or Unrestricted	
1431517397	Maple: Red	22	6	6	3	Well or Pit	
994063598	Maple: Freeman	25	10	3	3	Well or Pit	
1591838573	Maple: Norway	52	13	99	99	Open or Unrestricted	
1333224197	Oak: Pin	45	18	35	3	Tree Lawn or Parkway	

Create New Dataframe by Subsetting

End result is a table with trees that are limited in their growth space width.

In [16]:

```
rows_filter <- trees$growth_space_width <= 20
columns_filter <- c('id', 'height', 'width', 'growth_space_width', 'growth_space_type')

limited_width <- trees[rows_filter, columns_filter]
head(limited_width)
```

A tibble: 6 × 5

id	height	width	growth_space_width	growth_space_type
<dbl>	<dbl>	<dbl>	<dbl>	<chr>
754166088	0	0	2	Well or Pit
1431517397	22	6	3	Well or Pit
994063598	25	10	3	Well or Pit
1333224197	45	18	3	Tree Lawn or Parkway
239290336	8	4	3	Tree Lawn or Parkway
1233652274	27	10	3	Tree Lawn or Parkway

Remove Rows with Missing Values

Removed a little over 5000 rows

In [17]:

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

head(trees[height_missing, c('common_name', 'height', 'width')])
summary(trees)
```

A tibble: 6 × 3

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

```
      id            common_name           height           width
Min.   :5.960e+03   Length:45709       Min.   :  0.00   Min.   : 0.000
1st Qu.:5.356e+08   Class :character   1st Qu.:  9.00   1st Qu.: 2.000
Median :1.073e+09   Mode  :character   Median : 20.00   Median : 6.000
Mean   :1.074e+09                      Mean   : 22.16   Mean   : 6.991
3rd Qu.:1.613e+09                      3rd Qu.: 35.00   3rd Qu.:10.000
Max.   :2.147e+09                      Max.   :158.00   Max.   :65.000
                                       NA's   :4374     NA's   :4409
growth_space_length growth_space_width growth_space_type  diameter_base_height
Min.   :  0.00      Min.   : 0.00      Length:45709       Min.   : 0.00
1st Qu.:  3.00      1st Qu.: 2.00      Class :character   1st Qu.: 4.00
Median : 20.00      Median : 3.00      Mode  :character   Median :10.00
Mean   : 48.87      Mean   :26.53                         Mean   :12.85
3rd Qu.: 99.00      3rd Qu.:25.00                         3rd Qu.:19.00
Max.   :188.00      Max.   :99.00                         Max.   :66.00
NA's   :4194        NA's   :4192                          NA's   :4329
    stems         overhead_utilities   land_use          condition
Min.   :  0.000   Length:45709       Length:45709       Length:45709
1st Qu.:  1.000   Class :character   Class :character   Class :character
Median :  1.000   Mode  :character   Mode  :character   Mode  :character
Mean   :  1.039
3rd Qu.:  1.000
Max.   :211.000
NA's   :2
stormwater_benefits_dollar_value property_value_benefits_dollarvalue
Min.   : 0.000                   Min.   : -1.537
1st Qu.: 1.878                   1st Qu.: 28.174
Median : 5.888                   Median : 51.112
Mean   :10.101                   Mean   : 53.936
3rd Qu.:13.947                   3rd Qu.: 73.112
Max.   :85.320                   Max.   :344.668
NA's   :5665                     NA's   :5665
neighborhood          latitude       longitude
Length:45709       Min.   :40.36   Min.   :-80.09
Class :character   1st Qu.:40.43   1st Qu.:-80.00
Mode  :character   Median :40.45   Median :-79.95
                   Mean   :40.45   Mean   :-79.96
                   3rd Qu.:40.46   3rd Qu.:-79.92
                   Max.   :40.50   Max.   :-79.87
                   NA's   :251     NA's   :251
```

In [18]:

```
cleaned_trees <- trees[complete.cases(trees), , drop = FALSE]
head(cleaned_trees)
summary(cleaned_trees)
```

id	common_name	height	width	growth_space_length	growth_space_width	growth_space_type	diameter_bas
<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	
1946899269	Linden: Littleleaf	0	0	99	99	Open or Unrestricted	
1431517397	Maple: Red	22	6	6	3	Well or Pit	
994063598	Maple: Freeman	25	10	3	3	Well or Pit	
1591838573	Maple: Norway	52	13	99	99	Open or Unrestricted	
1333224197	Oak: Pin	45	18	35	3	Tree Lawn or Parkway	
239290336	Dogwood: Corneliancherry	8	4	99	3	Tree Lawn or Parkway	

```
      id             common_name            height           width
Min.   :5.960e+03   Length:39945       Min.   :  0.00   Min.   : 0.000
1st Qu.:5.375e+08   Class :character   1st Qu.: 10.00   1st Qu.: 3.000
Median :1.075e+09   Mode  :character   Median : 20.00   Median : 6.000
Mean   :1.075e+09                      Mean   : 22.86   Mean   : 7.215
3rd Qu.:1.612e+09                      3rd Qu.: 35.00   3rd Qu.:10.000
Max.   :2.147e+09                      Max.   :158.00   Max.   :65.000
growth_space_length growth_space_width growth_space_type  diameter_base_height
Min.   :  0.00      Min.   : 0.00      Length:39945       Min.   : 0.00
1st Qu.:  3.00      1st Qu.: 2.00      Class :character   1st Qu.: 4.00
Median : 20.00      Median : 3.00      Mode  :character   Median :10.00
Mean   : 49.03      Mean   :26.63                         Mean   :12.87
3rd Qu.: 99.00      3rd Qu.:30.00                         3rd Qu.:19.00
Max.   :188.00      Max.   :99.00                         Max.   :66.00
    stems         overhead_utilities   land_use          condition
Min.   :  0.000   Length:39945       Length:39945       Length:39945
1st Qu.:  1.000   Class :character   Class :character   Class :character
Median :  1.000   Mode  :character   Mode  :character   Mode  :character
Mean   :  1.128
3rd Qu.:  1.000
Max.   :211.000
stormwater_benefits_dollar_value property_value_benefits_dollarvalue
Min.   : 0.000                   Min.   : -1.537
1st Qu.: 1.876                   1st Qu.: 28.174
Median : 5.888                   Median : 51.112
Mean   :10.104                   Mean   : 53.917
3rd Qu.:13.947                   3rd Qu.: 73.112
Max.   :85.320                   Max.   :344.668
neighborhood          latitude       longitude
Length:39945       Min.   :40.36   Min.   :-80.09
Class :character   1st Qu.:40.43   1st Qu.:-80.00
Mode  :character   Median :40.45   Median :-79.95
                   Mean   :40.45   Mean   :-79.96
                   3rd Qu.:40.46   3rd Qu.:-79.92
                   Max.   :40.50   Max.   :-79.87
```

Add Two Columns

- geo_point: GeoJSON point format [longitude, latitude]
- property_value_per_height: the property value benefit of the tree divided by its height.

In [21]:

```
# Make geoJSON point function
makeGEO = function(longitude, latitude) {
    paste('[', longitude, ',', latitude, ']', sep = '')
}
```

In [23]:

```
trees_geo <- mutate(trees, geo_point = makeGEO(longitude, latitude))
trees_geo_prop_height <- mutate(trees_geo, property_value_per_height = property_value_benefits_dollarvalue
head(trees_geo_prop_height)
```

id	common_name	height	width	growth_space_length	growth_space_width	growth_space_type	diameter_bas
<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	
754166088	Stump	0	0	10	2	Well or Pit	
1946899269	Linden: Littleleaf	0	0	99	99	Open or Unrestricted	
1431517397	Maple: Red	22	6	6	3	Well or Pit	
994063598	Maple: Freeman	25	10	3	3	Well or Pit	
1591838573	Maple: Norway	52	13	99	99	Open or Unrestricted	
1333224197	Oak: Pin	45	18	35	3	Tree Lawn or Parkway	

Create New Dataframe and Combine

Created a new dataframe with two more row entries and combined with original dataset.

In [27]:

```
new_entries <- wrapr::build_frame(
    "id", "common_name", "height", "width", "growth_space_length", "growth_space_width", "growth_space_type
    "stems", "overhead_utilities", "land_use", "condition", "stormwater_benefits_dollar_value", "property_v
    1449000842, "Maple: Red", 20, 8, 10, 4, "Well or Pit", 8, 1, "Yes", "Residential", "Good", 7.245601, 48
    1000224821, "Oak: White", 28, 12, 6, 2, "Tree Lawn or Parkway", 18, 0, "No", "Commercial/Industrial", '
```

In [35]:

```
trees_bound <- rbind(trees, new_entries)
# checking if entry was added. Filtering with same id as first entry in manually created data frame
trees_bound[trees_bound$id == 1449000842,]
```

id	common_name	height	width	growth_space_length	growth_space_width	growth_space_type	diameter_bas
<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	
1449000842	Maple: Red	20	8	10	4	Well or Pit	

Pivot Wider

In [43]:

```
new_wide_data <- trees %>%
  pivot_wider(names_from =land_use, values_from = common_name)

head(select(new_wide_data, id, "Vacant":"Cemetery"))
```

A tibble: 6 × 11

id	Vacant	Residential	Commercial/Industrial	Institutional	Park	Multi-family Residential	Transportation	Utility	Golf Course
<dbl>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>
754166088	Stump	NA	NA	NA	NA	NA	NA	NA	NA
1946899269	NA	Linden: Littleleaf	NA	NA	NA	NA	NA	NA	NA
1431517397	NA	NA	Maple: Red	NA	NA	NA	NA	NA	NA
994063598	NA	Maple: Freeman	NA	NA	NA	NA	NA	NA	NA
1591838573	NA	Maple: Norway	NA	NA	NA	NA	NA	NA	NA
1333224197	NA	Oak: Pin	NA	NA	NA	NA	NA	NA	NA

Lab 1 & 2 Conclusions

I explored data on the public trees in the City of Pittsburgh in these two labs. Only a subset of the columns were used for these initial experiments, as there were too many to explore quickly. The initial summary experiments showed that the number "99" may be used as an indicator for unlimited growth width and height, as even though there were values higher than that, they were very few, and 99 appeared many times. This finding may show outliers or inconsistencies in how the data was collected. Related, I found an obvious but clear correlation between a growth space width or height of 99 and a growth space type of "Open or Unrestricted," giving more evidence that "99" was used as a placeholder for "Unlimited" growth space.

From plotting property value benefit versus land use, I found that trees in parks and on streets held more property value than any other land use. This plot was limited to one type of tree, "Ginko," so more experiments are needed to confirm this thesis. However, with many records of "Ginko" trees, the thesis would likely hold for other species. In both lab experiments, it was clear that many missing values would need to be appropriately cleaned up. Some records can be deleted, as they do not hold a tree and are labeled as "not suitable," but others would take more care as some trees have missing heights and widths. Lastly, I found this dataset fascinating and believe it could be very valuable for the City of Pittsburgh to find the most monetarily valuable or beneficial tree sites and for other cities in similar areas to explore which species or location types may benefit them.

In [ ]: