Education Data Science Summit

Ryan Estrellado

4/30/2021

This document is meant to accompany my talk at the Education Data Science Summit. It uses publicly available data. I hope it helps you!

Ryan Estrellado

linktr.ee/ry_estrellado

What is an R Markdown Document?

The authors of **R Markdown: The Definitive Guide** write that R Markdown empowers data scientists to "interweave narratives with code in a document . . ."

Think of it as a Word document that holds not only story, but also code and graphs.

Things to Watch For

Look for how writing, code, and plots appear in a single document.

Look for the parts of good data analysis using code:

- Identify the Question
- Import the Data
- Prepare the Data
- Pick a method to answer the question
- Visualize the data
- Analyze the Data
- Report the findings

Load Packages

Packages are add-ons to R that are built by software developers and communit members. They extend the capability of base R.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(here)
```

```
## here() starts at /Users/restrellado/OneDrive - San Diego County Superintendent of Schools/2017-2018/1
```

```
library(janitor)
```

```
##
## Attaching package: 'janitor'
```

```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

Identify the Question

How many schools do districts have on average?

Import Data

Before you work with data in R, you have to import the data.

You can get this data from the California Department of Education's website:

```
enroll <- read_tsv(here::here("19-20-enrollment.txt"))
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   CDS_CODE = col_character(),
##   COUNTY = col_character(),
##   DISTRICT = col_character(),
##   SCHOOL = col_character(),
##   GENDER = col_character()
## )
## i Use `spec()` for the full column specifications.
```

Here's one way working in R is different from working in a spreadsheet: you aren't dragging and dropping things around in a spreadsheet. Instead, you're using code to submit instructions to R.

Sometimes you just want to see the dataset. Here's a way to do that:

```
# See the data using View()
# View(enroll)
```

Question: Let's look at a few variables. Do these variables contain unique values? In other words, does Castle Rock school appear only once?

Prepare Data

Most publicly available datasets aren't prepared for the analysis you want to do. You have to prepare the dataset before you work with it.

- We want the average number of schools in a district
- To get that, we need to be able to count the number of schools in each district
- And to do that, we need each school in each district to only appear once
- Last, we only want districts in San Diego county

```
sd <- enroll %>%
  filter(COUNTY == "San Diego") %>%
  distinct(DISTRICT, SCHOOL) # Each school should only appear once

#View(sd)
```

Question: How many schools does the Alpine district have?

Count the schools in each district:

```
school_count <- sd %>%
  count(DISTRICT, sort = TRUE)

school_count
```

```
## # A tibble: 50 x 2
##    DISTRICT                    n
##    <chr>                   <int>
##  1 San Diego Unified         223
##  2 Chula Vista Elementary     50
##  3 Poway Unified              40
##  4 Vista Unified              34
##  5 Sweetwater Union High      32
##  6 Cajon Valley Union         31
##  7 Escondido Union            26
##  8 Oceanside Unified          26
##  9 La Mesa-Spring Valley      25
## 10 San Marcos Unified         22
## # ... with 40 more rows
```

Pick a Method to Answer the Question

Compute the average number of schools in a district. Compute the standard deviation of the number of schools in a district.

Visualize Data

Make the dataset easier to visualize by filtering for the top ten:

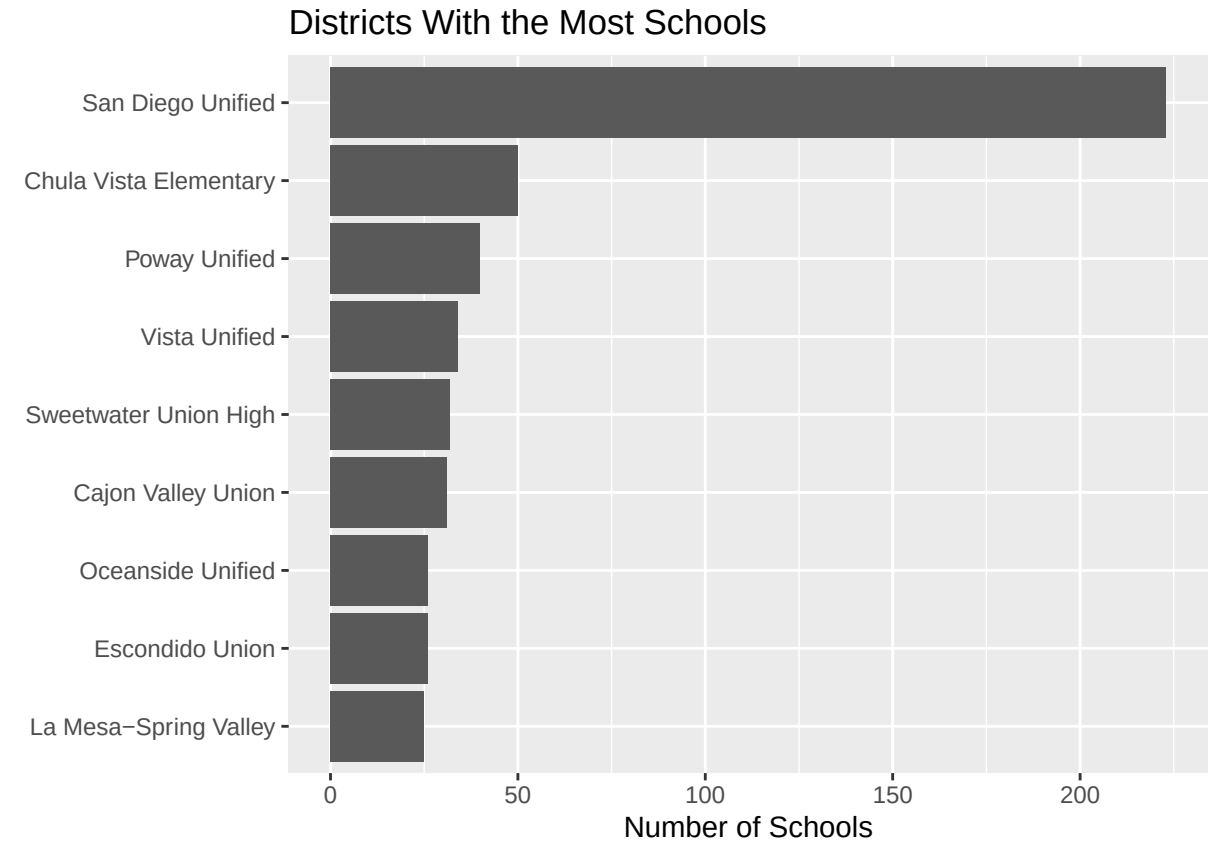
```
top_10 <- school_count %>%
  filter(min_rank(desc(n)) < 10)

top_10
```

```
## # A tibble: 9 x 2
##   DISTRICT                    n
##   <chr>                   <int>
## 1 San Diego Unified         223
## 2 Chula Vista Elementary     50
## 3 Poway Unified              40
## 4 Vista Unified              34
## 5 Sweetwater Union High      32
## 6 Cajon Valley Union         31
## 7 Escondido Union            26
## 8 Oceanside Unified          26
```

```
## 9 La Mesa-Spring Valley     25
```

Visualize the top ten districts by number of schools. As I do this, watch how I talk and build the visualization layer by layer based on what I want to see.

```
ggplot(data = top_10, aes(x = reorder(DISTRICT, n), y = n)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Districts With the Most Schools",
       x = "",
       y = "Number of Schools")
```



Analyze the Data

Refresh our memory of the dataset:

```
#View(school_count)
```

. . . and let's compute the mean and standard deviation.

```
school_count %>%
  summarise(avg_schools = mean(n), standard_dev = sd(n), median_schools = median(n))
```

```
## # A tibble: 1 x 3
##   avg_schools standard_dev median_schools
##         <dbl>        <dbl>          <dbl>
## 1        15.9         31.9              9
```

Report Your Findings

Districts in San Diego have about 16 schools in them, on average.

We should interpret this with caution, because there's a lot of variation in school counts in San Diego County. San Diego Unified has far and away the highest school count at 223 schools. For the more statistics savvy, that's 7 times the standard deviation—a very large difference from the county average.