

Introduction to R

Rachel Lesniak

February 18, 2019

What is data science?

"Data science is an exciting discipline that allows you to turn raw data into understanding, insight, and knowledge."

1. Capture data
2. Maintain data
3. Process data
4. Analyze data
5. Communicate data

Engineers do the first two steps (capture and maintain). **Don't worry about those.**

Statisticians do the third step (process). **Don't worry about this.**

Analysts do steps four and five (analyze and communicate). *You are already doing this!*

What is code/programming?

- Data science heavily focuses on code.
- If you type in directions into your computer, you are coding.
 - Yes, even in Excel!
- If you are writing multiple steps of code, you are programming.
- People use these interchangeably - you can too.

Why code or program?

- Code is text
 - copy and paste!!
- Code is read-able
 - read your code days, weeks, months later
 - check your work/someone else's work
 - understand unfamiliar processes
- Code is share-able
 - put it on GitHub, someone can use your work
 - learn from others' work
- Code is open
 - FREE (which means inclusive)

Why on earth would we use this for planning?

- You are around more data than you think
- Speed up repetitive processes
 - I literally received 112 Excel files this week
- Spreadsheets have limitations
- Not all data lives in spreadsheets

Pep Talk

- This is a skill you can tackle.
- You do not need to be a genius to get it.
- You do not need to struggle on your own.
- Today, let's have the confidence of a mediocre white man.

What is R?

- Doesn't stand for anything
- Language
- "Environment" or system
- First built for statistics
- Install packages to extend R

Why use R over other options?

- Easy to install (unlike Python)
- Easy to install packages (unlike Python)
- Easy to keep updated (unlike Python)
- Language is easy to read
- Great at building charts and graphs
- **R has a fantastic community for women.**

Installing R

- We're downloading the R system
- Download here: <https://cloud.r-project.org/>
- Follow the defaults

Installing R Studio Desktop

- We're downloading software that makes R easy to use.
- Download here:
<https://www.rstudio.com/products/rstudio/download/>

R Studio Layout

The screenshot shows the R Studio interface with four callouts pointing to different panels:

- 1) Script**: Points to the Script editor on the left, which contains R code for loading data, creating a scatterplot, and adding fit lines.
- 2) Console**: Points to the Console panel at the bottom left, which shows the output of the R code, including data summaries and the execution of plotting functions.
- 3) Workspace**: Points to the Workspace panel on the right, which shows the objects loaded in memory, including the 'google' dataset and the 'regline' object.
- 4) Results**: Points to the Plots panel on the right, which displays a scatterplot titled 'Interest in Data Visualization Searches by Percent of Population with College Degrees'. The plot shows a positive correlation between the percentage of the population with college degrees and the number of searches for data visualization. A red line represents a linear regression fit, and a blue line represents a lowess smoother.

From <http://dev1.ed-projects.nyu.edu/statistics/r-studio/>

Install tidyverse

- Packages are installed from CRAN (where you downloaded R)
- Documentation is available on CRAN and GitHub
 - The GitHub documentation is almost always better
 - Might be called "vignettes" or "READMEs"
- **tidyverse** is a package of multiple packages
- we are using **readr** and **dplyr** today

```
install.packages("tidyverse")
```

Load tidyverse

```
library(tidyverse)
```

```
## -- Attaching packages -----
```

```
## v ggplot2 3.1.0      v purrr  0.3.0
## v tibble  2.0.1      v dplyr  0.7.8
## v tidyr   0.8.2      v stringr 1.4.0
## v readr   1.3.1      v forcats 0.3.0
```

```
## -- Conflicts ----- tidy
```

```
## x dplyr::filter() masks stats::filter()
```

```
## x dplyr::lag()     masks stats::lag()
```

- You can also use `library()` to see all of your installed packages

Find our working directory

```
getwd()
```

```
## [1] "C:/Downloads/women-of-urp-master/women-of-urp-master/intro-t
```

- Working directory can (and should) be changed with each project
- We're not going to get into this today
- Learn more [here](#) later

Download today's data

- List of Certified Business Enterprises, or contracting businesses that get preferred opportunities from the DC Gov
- Download from GitHub repo
- Originally from OpenDataDC
- Save where your working directory is

Import data

- We'll use the `read_csv()` function
- A function has a name and arguments
- `?` in front of a function opens a help window that explains the function and its arguments
 - I use this constantly.
- `<-` means "assign" or "create". The object name on the left becomes whatever is on the right side

```
?read_csv
```

```
raw_data <- read_csv("Certified_Business_Enterprise.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   OBJECTID = col_double(),
##   EXPIRATIONDATE = col_datetime(format = ""),
##   GIS_LAST_MOD_DTTM = col_datetime(format = ""),
##   WARD = col_double(),
##   STARTDATE = col_datetime(format = ""),
##   PROPOSALPOINTS = col_double(),
##   OTHERCERTIFICATIONS = col_logical(),
##   DATEESTABLISHED = col_datetime(format = ""),
##   BIDPRICEREDUCTION = col_double(),
```

16/28

Look at data, example 1

- Many ways to look at data
- Column Names
- Column Types
- Organization

raw_data

```
## # A tibble: 1,729 x 27
##   OBJECTID BUSINESSNAME PRINCIPALOWNER EXPIRATIONDATE
##   <dbl> <chr>          <chr>          <dtm>
## 1    1001 Goldblatt M~ Thorn Pozen;D~ 2021-05-14 00:00:00
## 2    1002 Goldin & St~ Brian Matting~ 2021-03-12 00:00:00
## 3    1003 Key Global ~ Cindy Quiroz   2020-12-08 00:00:00
## 4    1004 Keystone Pl~ Carlos Perdomo 2021-04-30 00:00:00
## 5    1005 KeyUrban     Dahn Warner     2021-04-05 00:00:00
## 6    1006 KGO, LLC      William R Ken~ 2019-12-21 00:00:00
## 7    1007 KGP Design ~ William Galla~ 2019-05-27 00:00:00
## 8    1008 MindFinders~ Tim Booker      2019-04-03 00:00:00
## 9    1009 Princess P ~ Kyree Clarke   2021-03-30 00:00:00
## 10   1010 PRISM INTER~ Deon Ford       2021-04-02 00:00:00
## # ... with 1,719 more rows, and 23 more variables:
## #   GIS_LAST_MOD_DTTM <dtm>, WEBSITE <chr>, WARD <dbl>, STARTDAT
## #   PROPOSALPOINTS <dbl>, PHONE <chr>, OTHERCERTIFICATIONS <lgl>,
## #   ORGANIZATIONTYPE <chr>, FAX <chr>, EMAIL <chr>,
## #   DATEESTABLISHED <dtm>, CONTACTNAME <chr>, CERTIFICATIONNUMBE
## #   BUSINESSDESC <chr>, BIDPRICEREDUCTION <dbl>, ADDRESS <chr>, S
```

17/28

Look at data, example 2

```
head(raw_data)
```

```
## # A tibble: 6 x 27
##   OBJECTID BUSINESSNAME PRINCIPALOWNER EXPIRATIONDATE
##   <dbl> <chr>          <chr>          <dtm>
## 1    1001 Goldblatt M~ Thorn Pozen;D~ 2021-05-14 00:00:00
## 2    1002 Goldin & St~ Brian Matting~ 2021-03-12 00:00:00
## 3    1003 Key Global ~ Cindy Quiroz  2020-12-08 00:00:00
## 4    1004 Keystone Pl~ Carlos Perdomo 2021-04-30 00:00:00
## 5    1005 KeyUrban    Dahn Warner    2021-04-05 00:00:00
## 6    1006 KGO, LLC    William R Ken~ 2019-12-21 00:00:00
## # ... with 23 more variables: GIS_LAST_MOD_DTTM <dtm>, WEBSITE <
## #   WARD <dbl>, STARTDATE <dtm>, PROPOSALPOINTS <dbl>, PHONE <ch
## #   OTHERCERTIFICATIONS <lgl>, ORGANIZATIONTYPE <chr>, FAX <chr>,
## #   EMAIL <chr>, DATEESTABLISHED <dtm>, CONTACTNAME <chr>,
## #   CERTIFICATIONNUMBER <chr>, BUSINESSDESC <chr>,
## #   BIDPRICEREDUCTION <dbl>, ADDRESS <chr>, SBE <lgl>, CATEGORIES
## #   MAR_ID <dbl>, XCOORD <dbl>, YCOORD <dbl>, LATITUDE <dbl>,
## #   LONGITUDE <dbl>
```

Look at data, example 3

```
glimpse(raw_data)
```

```
## Observations: 1,729
## Variables: 27
## $ OBJECTID      <dbl> 1001, 1002, 1003, 1004, 1005, 1006, 1
## $ BUSINESSNAME  <chr> "Goldblatt Martin Pozen LLP", "Goldin
## $ PRINCIPALOWNER <chr> "Thorn Pozen;David Goldblatt;Thomas M
## $ EXPIRATIONDATE <dtm> 2021-05-14, 2021-03-12, 2020-12-08,
## $ GIS_LAST_MOD_DTTM <dtm> 2019-02-08 05:02:22, 2019-02-08 05:0
## $ WEBSITE       <chr> "www.gmpllp.com", "www.goldinandstaff
## $ WARD          <dbl> 2, 5, 2, 8, 4, 2, 2, 2, 8, 2, 7, 1, 2
## $ STARTDATE     <dtm> 2018-05-14, 2018-02-07, 2017-12-08,
## $ PROPOSALPOINTS <dbl> 12, 7, 9, 12, 12, 7, 12, 12, 12, 12,
## $ PHONE         <chr> "2027959999", "2028822600", "20277092
## $ OTHERCERTIFICATIONS <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, N
## $ ORGANIZATIONTYPE <chr> "Partnership", "Corporation", "Corpor
## $ FAX           <chr> "2027959192", "2028825330", NA, "2028
## $ EMAIL         <chr> "dgoldblatt@gmpllp.com", "bmattingly@
## $ DATEESTABLISHED <dtm> 2013-02-07, 1992-02-01, 2015-05-27,
## $ CONTACTNAME   <chr> "David Goldblatt", "Brian Mattingly",
## $ CERTIFICATIONNUMBER <chr> "LSZR71856052021", "LSZ49524032021",
## $ BUSINESSDESC   <chr> "Golblatt Martin Pozen LLP provides t
## $ BIDPRICEREDUCTION <dbl> 0.12, NA, NA, 0.12, NA, NA, NA, NA, 0
## $ ADDRESS       <chr> "1625 K STREET NW WASHINGTON DC 20006
## $ SBE           <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, T
## $ CATEGORIES     <chr> "Local Business Enterprise (LBE);Sma
## $ MAR_ID         <dbl> 242032, 287305, 279142, 53665, 224880
## $ XCOORD         <dbl> 396755.1, 402769.5, 397164.6, 402063.
```

Analyze data

- Sample analysis time!
- Goal: count how many CBEs are in Ward 8 by year established and export
- Uses functions from **dplyr**
- It is okay if we don't get `through all of these
- Saved on GitHub for later

Filter

- Select just Ward 8
- Uses two equal signs
- `%>%` is called a pipe and is part of **tidyverse**
 - Read it as "and then"

```
sample_data <- raw_data %>%
  filter(WARD == 8)
```

```
head(sample_data)
```

```
## # A tibble: 6 x 27
##   OBJECTID BUSINESSNAME PRINCIPALOWNER EXPIRATIONDATE
##   <dbl> <chr>          <chr>          <dtm>
## 1    1004 Keystone Pl~ Carlos Perdomo 2021-04-30 00:00:00
## 2    1009 Princess P ~ Kyree Clarke   2021-03-30 00:00:00
## 3    1016 Jahphut      Darryl Roberts 2021-11-30 00:00:00
## 4    1020 Gotta Go No~ Frederick Hil~ 2021-07-22 00:00:00
## 5    1029 MLG Truckin~ Timothy Goodw~ 2021-10-25 00:00:00
## 6    1030 MMP Enterpr~ Alvin Butler;~ 2021-04-24 00:00:00
## # ... with 23 more variables: GIS_LAST_MOD_DTTM <dtm>, WEBSITE <chr>,
## #   WARD <dbl>, STARTDATE <dtm>, PROPOSALPOINTS <dbl>, PHONE <chr>,
## #   OTHERCERTIFICATIONS <lgl>, ORGANIZATIONTYPE <chr>, FAX <chr>,
## #   EMAIL <chr>, DATEESTABLISHED <dtm>, CONTACTNAME <chr>,
## #   CERTIFICATIONNUMBER <chr>, BUSINESSDESC <chr>,
## #   BIDPRICEREDUCTION <dbl>, ADDRESS <chr>, SBE <lgl>, CATEGORIES
## #   MAR_ID <dbl>, XCOORD <dbl>, YCOORD <dbl>, LATITUDE <dbl>,
## #   LONGITUDE <dbl>
```

Arrange

- Arrange A-Z or smallest to largest
- ``arrange(desc(BUSINESSNAME))` would be descending

```
sample_data_1 <- sample_data %>%
  arrange(BUSINESSNAME)
```

```
head(sample_data_1)
```

```
## # A tibble: 6 x 27
##   OBJECTID BUSINESSNAME PRINCIPALOWNER EXPIRATIONDATE
##   <dbl> <chr>          <chr>          <dtm>
## 1      42 24-7 Distri~ Derrick Wood    2021-01-08 00:00:00
## 2     825 a-always en~ Bobby Bullock~ 2022-01-24 00:00:00
## 3      45 A Cut Above~ Wayne Agnew     2020-03-15 00:00:00
## 4      48 A&C Constru~ Alicia Araujo   2020-09-21 00:00:00
## 5      62 Air Vent Cl~ Earl Alston     2020-07-13 00:00:00
## 6    1323 AJK Enterpr~ Antonio Korne~ 2019-02-28 00:00:00
## # ... with 23 more variables: GIS_LAST_MOD_DTTM <dtm>, WEBSITE <
## #   WARD <dbl>, STARTDATE <dtm>, PROPOSALPOINTS <dbl>, PHONE <ch
## #   OTHERCERTIFICATIONS <lgl>, ORGANIZATIONTYPE <chr>, FAX <chr>,
## #   EMAIL <chr>, DATEESTABLISHED <dtm>, CONTACTNAME <chr>,
## #   CERTIFICATIONNUMBER <chr>, BUSINESSDESC <chr>,
## #   BIDPRICEREDUCTION <dbl>, ADDRESS <chr>, SBE <lgl>, CATEGORIES
## #   MAR_ID <dbl>, XCOORD <dbl>, YCOORD <dbl>, LATITUDE <dbl>,
## #   LONGITUDE <dbl>
```

Select

- Select only the columns we want

```
sample_data_2 <- sample_data_1 %>%
  select(BUSINESSNAME, PRINCIPALOWNER, DATEESTABLISHED)
```

```
head(sample_data_2)
```

```
## # A tibble: 6 x 3
##   BUSINESSNAME          PRINCIPALOWNER          DATEESTAB
##   <chr>                <chr>                <dtm>
## 1 24-7 District Volt, In~ Derrick Wood          2017-03-2
## 2 a-always enterprises i~ Bobby Bullock;Sharon Bullock~ 1997-06-0
## 3 A Cut Above General Co~ Wayne Agnew          2015-02-0
## 4 A&C Construction Compa~ Alicia Araujo         2009-06-1
## 5 Air Vent Cleaning Serv~ Earl Alston           2014-03-1
## 6 AJK Enterprise, LLC.    Antonio Kornegay       2006-06-2
```

Mutate

- Add a new column
- `substr()` lets you pick which characters you want (you don't have to understand this today)

```
sample_data_3 <- sample_data_2 %>%
  mutate(year = substr(DATEESTABLISHED, 1, 4))
```

```
head(sample_data_3)
```

```
## # A tibble: 6 x 4
##   BUSINESSNAME      PRINCIPALOWNER      DATEESTABLISHED
##   <chr>              <chr>              <dtm>
## 1 24-7 District Volt, ~ Derrick Wood      2017-03-27 00:0
## 2 a-always enterprises~ Bobby Bullock;Sharon Bul~ 1997-06-09 00:0
## 3 A Cut Above General ~ Wayne Agnew      2015-02-06 00:0
## 4 A&C Construction Com~ Alicia Araujo      2009-06-16 00:0
## 5 Air Vent Cleaning Se~ Earl Alston      2014-03-10 00:0
## 6 AJK Enterprise, LLC. Antonio Kornegay      2006-06-22 00:0
```


Group By and Summarize

- Group By creates groups of rows with the same values
- Summarize does whatever summary you want within the group

```
sample_data_4 <- sample_data_3 %>%  
  group_by(year) %>%  
  summarize(n())
```

```
head(sample_data_4)
```

```
## # A tibble: 6 x 2  
##   year `n()`  
##   <chr> <int>  
## 1 1964     1  
## 2 1978     2  
## 3 1979     2  
## 4 1980     1  
## 5 1983     1  
## 6 1985     3
```

Export data

- Similar to `read_csv()`
- First tell it which data frame you want
- Then tell it the name of the file

```
write_csv(sample_data_4, "CBE_by_year.csv")
```

Learn More:

- R for Data Science book (free)
 - <https://r4ds.had.co.nz/>
- R for Reproducible Scientific Analysis (free)
 - <https://swcarpentry.github.io/r-novice-gapminder/>
- Data Camp (subscription)
- #rstats on Twitter
- Watch screencasts on YouTube
- Cheatsheets from RStudio (free reference)
 - <https://www.rstudio.com/resources/cheatsheets/>

More packages to try out

- `readxl` - import Excel files
- `ggplot2` - create charts/graphs
- `tidycensus` - quickly download Census data (love this)
- `sf` - spatial analysis
- `tigris` - quickly download TIGER shapefiles