# Data Journalism - A Quick & Practical Guide

Rob Wells - thanks to Yihui Xie

2021-06-02

# Contents

# Data Journalism - A Quick & Practical Guide

UNIVERSITY OF ARKANSAS | School of Journalism and Strategic Media

**Rob Wells, Ph.D.**
rswells@uark.edu
Twitter: rwells1961

---

**Inspiration:**

Mapping in Tableau

Please review this mapping tutorialbasic-mapping-transcriptSample WorkbookVideos: https://www.tableau.com/learn/tutorials/on-demand/getting-started-mapping?playlist=230855

Build a Map COVID-19 Positivity in Arkansas CountiesData: County data for one day.

–Import the countyonlytoday.csv data as text into Tableau–Check the geographic role of county is assigned to county.–Go to Sheet 1:a) drag Longitude to columnsb) Latitude to rowsc) County Name to the pane–A map of Arkansas appears.d) In top Menu: MAP | Edit Locations | State / Province | Fixed - Select the great state of Arkansas–The map now has all counties represented as dotse) In Marks card: Change Automatic to Mapf) drag Positive to the pane - you now have a map of positive rates by county.Video below with all of these steps:

Dual Mapping - Bubble Maps

Tutorialhttps://onlinehelp.tableau.com/current/pro/desktop/en-us/maps_dualaxis.htmlUsing countyonlytoday.csv

New sheet, begin map: Drag counties to mapFix the missing counties: Map | Edit Locations | Fixed | ArkansasMarks Card | MapDrag Deaths map

Here's where it is tricky:

Click on Longitude pill in Columns. Press Command key. Drag to Right. Release mouse–Creates two Longitude pills and two maps–Marks Card Now Has Controls for Two MapsMarks Card Has Two Maps. Lower Map, Drag off color pill. Marks Card, switch to Circle.Drag Deaths to Size. Enlarge the CirclesDrag Deaths to Labels.

In Columns, Select Down Arrow on Longitude | Dual Axis–Maps are combined Drag County to Marks Card | TooltipEdit Tooltips so data displays properly

**Teams**

```
For the first class, please have Teams installed so we can do some exercises.
Teams is free through your university Office365 account. Download the Teams App throu
```

https://its.uark.edu/communication-collaboration/office365/office365-desktop-apps.php

**Teams Videos**

```
Microsoft Teams allows us to easily share information through the class or in discre
```

```
Chat in Teams
```

https://www.microsoft.com/en-us/videoplayer/embed/RE4rLgJ?pid=ocpVideo5-innerdiv-oneplayer&postJsllMsg=true&maskLevel=20&market=en-us

```
Create a post
```

https://www.microsoft.com/en-us/videoplayer/embed/RE2BIrO?pid=ocpVideo0-innerdiv-oneplayer&postJsllMsg=true&maskLevel=20&market=en-us

```
How to tag a person in Teams
```

https://www.microsoft.com/en-us/videoplayer/embed/RWkJ9C?pid=ocpVideo0-innerdiv-oneplayer&postJsllMsg=true&maskLevel=20&market=en-us

# Chapter 1

# Data Journalism - A Quick & Practical Guide

## 1.1 Outline - Introduction - Table of Contents

Each chapter would cover 5 to 10 pages. Much of the material will consist of original content I have prepared using open government datasets, such as the Arkansas Department of Health Covid-19 data or the FBI's crime statistics. I have been trained about the Fair Use issues and do not anticipate copyright issues.

Images will be screenshots I generate from my lessons. The chapters will include links to other web resources, such as the Tableau Public frequently asked questions. The chapters will include links to original video content I produced for my classes and that is hosted on video.uark.edu.

## 1.2 Table of Contents

## 1.3 Introduction to data analysis

–Introduction to data analysis, including math for journalists

```
Basics of Data Analysis
Numbers in the Newsroom
Excel Exercise: Transit Data and Calculating a Rate
Review: Mac OSX Basics
```

–Organizing your workflow

```
Best practices in data management
Organizational tips for files
Data documentation skills
```

–Data cleaning

```
Filtering
Reading Data Dictionaries
Data Cleaning Exercises
```

–Data Visualization

```
Principles of Data Visualization
Cleveland McGill Scale
Important Resources for Surveying the Data Visualization Options
Color Choices
Build a Cover Image Using Canva or InDesign or Powerpoint
Higher Resolution Graphics in Tableau
```

–Writing About Data

```
Writing Style Notes
Common Errors - Math
AP Style with Numbers
```

–Excel Bootcamp

```
Review four corners
How to Filter in Excel
Basics and Sorting in Excel
Practice Rates and Ratios
Excel formulas
Pivot Tables
Countif Function
```

## 1.4 Section 2: Data Analysis Tools

–Basic Tableau

```
Downloading instructions for Tableau
Getting started tutorial with video
Building a basic COVID data chart with video and transcript
Using filters and calculations with video and transcript
Tutorial on Tableau calculations with video
Proper formatting of a filter bar in Tableau, video
Links to additional Tableau Tutorials
```

–Tableau - Maps

```
Mapping tutorial & sample dataset
Build a Map COVID-19 Positivity in Arkansas Counties
Video
Dual Mapping - Bubble Maps
```

–Basic Flourish

```
Videos to Get Started
Beginning Documents
Flourish Design Tips
More Resources
Adam Marton Cheat Sheet
Flourish newsrooms plan
Flourish - Stories
Examples from Fall Class
Basic Map
Flourish Links
```

–Basic Datawrapper

```
First Datawrapper Chart
Tutorials
Adam Marton Datawrapper Training
Automatic chart updates
Maps in Datawrapper
```

## 1.5 Section 3: R for data journalists

–Introduction to R

```
Install R and R Studio.
Basic tutorial on R
```

```
https://profrobwells.github.io/Guest_Lectures/Intro_To_R/R1_Intro-to-R.html
Reading
Reproducible research Repetitive tasks in modern newsrooms
Popular R Libraries
Data Types and R
Reference: Logical Operators in R
Packages
Important Reference Materials
R and R Studio CHECK REPETITION
```

–R data visualization

```
Basic GGPLOT
Using color
Formatting PNG for export
Scatterplots
Histograms
Box plots
Line graphs
ggplot cookbook 4-26-20.rmd
```

–R data cleaning

```
Data cleaning exercises using SF police data
Clean names, Process dates
top_n: table with just the top five counties' crime rate
Grouping by Disposition
Filters
A more complex filter
String manipulation
Rename specific strings: str_replace_all
Using a lookup table to replace all the values
gsub - delete space
convert all text to lowercase
Make into html table
Make bubble chart
```

R Markdown to distribute findings to Stanford, Feb 2020 https://profrobwells.github.io/HomelessSP2020/SF_311_Calls_UofA.html Homeless Children, Feb 25 https://profrobwells.github.io/HomelessSP2020/Homeless_Children_Feb_25_2020.html

–Cookbook of common tasks

## 1.6 Section 4: Publishing your work

–WordPress

```
Using WordPress
Access back end
Embedding interactive data
Embed Flourish in WordPress
Embed Tableau in WordPress
Building the Web Page
```

–GitHub for beginners

## 1.7 Appendix

–Teams and Slack

## 1.8 NOTES

Lectures Investigative Reporters and Editors Inc., Society for Advancing Business Editing and Writing and similar national journalism organizations

### 1.8.1 Instructions from OER

project outline and project completion timeline. Please include examples or links to content you will use (if known), drafts of any completed chapters or other supporting documentation. Please also include anticipated needs for support from the OER team such as assistance searching for resources, instructional design support or assistance creating ancillary materials or graphics

You can label chapter and section titles using `{#label}` after them, e.g., we can reference Chapter **??**. If you do not manually label them, there will be automatic labels anyway, e.g., Chapter **??**.

# Chapter 2

# Introduction to Data Analysis

Sections in this Module

Basics of Data Analysis

Numbers in the Newsroom

Excel Exercise: Transit Data and Calculating a Rate

Review: Mac OSX Basics

Basics of Data Analysis

TransparencyReliability: How sure are we that we got the right answer? That we've done everything correctly?Replicability: If we had to do it all again, would we get the same answer? If someone else did it, would they?Transparency: If our results are challenged, can we show exactly what we've done to defend it?–Matt WaiteData Analysis

— Review methodology with one or more other data people— Check results to other available comparable data— Ensure all record counts are consistent across stages— Check averages — Examine outputs to ensure logical consistency (do things that should add up to 100% add up to 100%?)— Recheck all coding line by line if possible or in aggregate if not — Re-read all programs/scripts— Re-run entire analysis from scratch— Check each number against analysis or source material prior to publication— Recheck each number against analysis or source material on each draft Credit: Daniel Lathrop. Dallas Morning News

AP Stylebook Entry on Data Journalism

Data sources used in stories should be vetted for integrity and validity. When evaluating a data set, consider the following questions:–What is the original source for the data? How reliable is it? Can we get answers to questions about

it?– Is this the most current version of the data set? How often is the data updated? How many years of data have been collected?–Why was the data collected? Was it for purposes of advocacy? Might that affect the data's reliability or completeness? Does the data make intuitive sense? Are there anomalies (outliers, blank values, different types of data in the same field) that would invalidate the analysis?–What rules and regulations affect the gathering (and interpretation) of the data?–Is there an alternative source for comparison? Does the data for a parallel industry, organization or region look similar? If not, what could explain the discrepancy?–Is there a data dictionary or record layout document for the data set? This document would describe the fields, the types of data they contain and details such as the meaning of codes in the data and how missing data is indicated. If the data collectors used a data entry form, is the form available to review? For example, if the data entry was performed by inspectors, is it possible to see the form they used to collect the data and any directions they received about how to enter the data?Data and the results of analysis must be represented accurately in stories and visualizations. Any limitations of the data must also be conveyed. If one point in the analysis is drawn from a subset of the data or a different data set altogether, explain why this was done.Use statistics that include a meaningful base for comparison (per capita, per dollar). Data should reflect the appropriate population for the topic: for example, use voting-age population as a base for stories on demographic voting patterns. Avoid percentage and percent change comparisons from a small base. Rankings should include raw numbers to provide a sense of relative importance.When comparing dollar amounts across time, be sure to adjust for inflation. When using averages (that is, adding together a group of numbers and dividing the sum by the quantity of numbers in the group), be wary of extreme, outlier values that may unfairly skew the result. It may be better to use the median (the middle number among all the numbers being considered) if there is a large difference between the average (mean) and the median.Correlations should not be treated as a causal relationship. Where possible, control for outside factors that may be affecting both variables in the correlation. Use round numbers where possible, particularly to avoid a false appearance of precision. Be clear about limitations of sample size in reporting on data sets. See the polls and surveys section for more specific guidance on margin of error.Try not to include too many numbers in a single sentence or paragraph.

A refresher on AP Stylebook on numbers

How to Lie With Statisticshttps://www.datasciencecentral.com/profiles/blogs/how-to-lie-with-visualizations-statistics-causation-vs

Sheffo, Catherine. "How to Avoid 10 Common Mistakes in Data Reporting." American Press Institute (blog), August 9, 2016. https://www.americanpressinstitute.org/publicatio reporting-common-mistakes/

Writing Assignment

Numbers in the Newsroom

Sarah Cohen, Math Diva

Sarah Cohen's "Numbers in the Newsroom" is a classic in journalism numeracy. She is a Pulitzer-winning journalist at The Washington Post, a former Duke University professor, a data journalist at The New York Times., now a professor at Arizona State University. That's why we read her book.

- Limit yourself to 8- 12 digits, including dates such as 2012, in a single paragraph.–This allows us to stress the most important numbers

–Simplify your story using rates, ratios or percentages. "One in four" = ratio or rate. "Forty percent" = ratio or rate. 235 deaths per 100,000 is another. See pg. 11

*Memorize some common numbers on your beat: Population of Fayetteville. Population of Arkansas. Population of the U.S. Per capita income Arkansas and U.S.

*Round off! Unless you're dealing with really small numbers, decimal points may not be meaningful. "I'm a big fan of rounding," Cohen said.* To make a very small number more understandable, divide it into 1. For example, .0081 is the proportion of the U.S. population who die every year. 1/.0081 translates to 1 in every 124 Americans die each year.* If you have a story filled with numbers – and not people — it needs to be really, really short.

- Portion of whole – For example, at the time of the Million Man March in 1995, a turnout of 1 million black men would have represented 1/12th of all the black men in the country at the time.

Rates and Ratios

Numbers in the Newsroom: Rates and Ratios

Class exercise: Cohen: Think in ratios – construct a ratio on the poverty beat. Memorize common numbers on the beat:

Use the Census Poverty Data:%C2%A0;US Ark Counties Poverty ACS_16_5YR_DP03_with_ann-1w6iwss

–In the spirit of "memorizing numbers on your beat," find three statistics about poverty in this dataset

–Construct a rate or ratio about the number of households earning $15,000 to $24,999 for the U.S., Arkansas, and the counties with the highest and lowest percentages in this category. Remember – "percents are Fractions. Fractions are percents"

Excel Exercise: Transit Data and Calculating a Rate

Basic Excel: http://www.interhacktives.com/2015/11/02/quick-tips-excel-google-sheets/

This exercise involves calculating train rate fatalities.Click here for the instructions:Exercise4Click here for the data:transitNotes:–Create data dictionary, backup, do four corners test–Be very careful about copying different block of data to a new sheet: mixups–Copy labels over and then delete them just to be sure all is aligned–Class walkthrough with 2008 - 2009 derailments–Be very specific about the headers: Total Derailments 2009, Vehicle Revenue Miles–Word Wrap for headers–We are constructing two derail rates, one in 2009 and another in 2008.–Results are 0? Wait, check the decimal tool–Results to two decimals. Rarely more than that–Copy of acronym definitions to data dictionaryExercise #1:–Calculate derailment rates for 2008-2013, determine the average rate, which agency had the highest average rate?Exercise #2:–Calculate the rate of fatalities (excluding suicides) by total miles (vehicle revenue miles)–Copy all of the Total Heavy Rail Fatality Sum, excluding suicides and all of the Vehicle Revenue Miles (VRM)–Create rates for each year, then average them Which city has the highest rate of fatalities (excluding suicides) over the last six years and where does Chicago rank? Exercise #3:Over the six years, did Chicago transit have more derailments than other major city transit systems? Is it getting better or worse? Which year was the worst for all major transit in terms of fatalities (excluding suicides)? How many suicides happened at CTA in 2013? What questions should I ask the DOT data clerks regarding the data? What other data might be useful to mine after this story runs?Resources: Excel Formulas in NICAR Coursepack

transit

Relative Risk

“Black applicants are denied mortgages at twice the rate of whites with similar incomes.”

If 20 smokers per thousand contract cancer, and yet non-smokers have a cancer rate of only 10 per thousand, the relative risk of smoking is 2.

“More than” or “less than” = compute difference between the smokers, an extra step

Example: Relative RiskFiguring Rates – Numbers in NewsroomMathCrib-Doig

Excel Exercises

Click here for: Basics and Sorting in ExcelClick here for: CityBudget.xls

Click here for the data: UrbanPopClick here for assignment: Exercise #1 Answer these questions:Sorting–Which urban agglomeration was the largest in 1950?–Which is expected to be the largest in 2030? Percentage ChangeFormula: (New number-Old Number)/Old Number * 100 and use % symbolCreate columnWhat is difference.—copy forumulaWhat is percentage change—copy

formula Percentage Change–Which had greatest rate of change between 1950-2015?–Are any urban areas expected to lose population from 2010 to 2030?–If so, how many and which one is expected to lose the most?–Which United States urban area is expected to have the largest percent increase from 2015 to 2030?

Refresher on Mac OSX operating systemHere is a short video course that you can skim through and get up to speed on how to use the Apple operating system, OSX.https://www.linkedin.com/learning/macos-mojave-essential-training/understand-macos-the-foundation-of-working-with-a-mac?u=50849081I would hammer through the following as soon as possible.Chs. 1, 3 are importantChapter 2: Finder will be crucial.Ch. 5 on downloading from the web is importantCh. 4, 13 should be skimmedChs 6-11 aren't important for our class

# Chapter 3

# Organize Your Data

This module addresses:–Best practices in data management–Organizational tips for files–Data documentation skills

Staying organized is a key problem for beginning data students.You can't find files. You have duplicate files and struggle to find the latest version.Your data software fails because it can't find your files.You can't remember where you got the source data or what the headers mean.You waste hours with this stuff when you really should be reporting.

I want to put an end to this nightmare. These organizational tools below are essential.

Storage

Organize Your Data: Finder

Finder, not always up for the job

1. Sort by grid, by date.–This allows you to see the latest version of your files.

2. Path name.–Follow this convention: Description of File With Some Detail, Date. If you are editing something, put your initials at the end.–i.e.: Covid_Master_File_Jan_11_2021-rsw

3. Copying File Paths from the Mac Finder.Navigate to the file or folder you wish to copy.Right-click (or Control+Click, or a Two-Finger click on trackpads) on the file or folder in the Mac Finder While in the right-click menu, hold down the OPTION key to reveal the "Copy (item name) as Pathname" option, it replaces the standard Copy optionOnce selected, the file or folders path is now in the clipboard, ready to be pasted anywhere

Data Diary

Data Dictionary

Data Diary ExamplesThe following material was posted on NICAR-L, a listserv for data journalists. There are some great examples of how the pros use data diaries / data dictionaries in their workflow.1) GeoffThis is a great question, and I'm finding as I think through my response that it's helpful to remind myself of good practices.I use Jupyter notebooks for when I'm doing analysis or exploration in Python or SQL and R Markdown for when I'm doing it in R. However, I would stress that any data diary you keep and keep in a detailed way that is useful to you and others later, regardless of format, is better than the one you don't.https://github.com/newsapps/public-notebooks/blob/master/Shooting%20victims%20by%20block.ipynb is an example of a representative but not great notebook for a small data task.A few things that I try (but don't always succeed) to do:– Link to the source data, summary reports and codebooks near the top of my notebook. This is both a convenience to me, because I refer to these often, and especially to others who may not have seen those things before.– Put a high level summary of why I'm interested in the data and what I'm trying to find at the top of the notebook. This keeps me focused as I'm doing my exploration and also is helpful for others who might be skimming.– Keep a parking lot of questions (or potential concerns about validity or cleanliness of data) near the top of the notebook. That way I can quickly capture things I think about as I'm exploring or analyzing the data, while still staying focused.– Near the end of my day (or the first thing the next morning), do a quick pass over a notebook I worked on during the day. Do my notes still make sense? Are they as clear as they could be? If not, try to clean them up. If I don't have time at the moment, I at least leave a "TODO" note to flag the section as needing some love.– Share the notebook with someone else as early as possible, even if you're still in-progress. This is the most helpful way to know if I'm capturing your process with enough granularity. Or maybe I'm getting too granular. If so, is there a way to summarize process and findings at the top of a section?– If using code, don't give a play-by-play of the code in text. Instead, describe what I'm trying to find out, why it's important and why I'm taking a particular approach. Also note any assumptions my code is making.Hopefully this is helpful.Best,Geoff2) Christian McDonaldOh, do I have feelings about this one…I keep a data diary for myself that has everything from notes about public information requests, notes about where I got data, descriptions of what I did, sql queries and all kinds of things. I sometimes also make a data report that is really RESULTS of what I learned, as opposed to how I got there in the data diary. The data report is more for other reporters, editors and maybe sources, but the diary is for me, so less formal.These days I'm trying to script more of my work using Jupyter Notebooks, which then tends to be a mix of the two. It has info about where the data came from and the code that made the result. Sometimes it is written for future me, sometimes for the public. Generally, I'll still keep a personal data diary just for my future self, 'cause I can't remember what I did yesterday much less last week.Data diaries I tend to write in markdown

files on my machine so code doesn't get wigged with curly-quote translations. Data reports are typically Google Docs or Jupyter Notebooks on Github.

# Chapter 4

# Data Cleaning

Basic Population - Race Census Data Download Instructions

Census: data.census.gov

https://data.census.gov

1) Advanced Search–Topics | Geography | Years | Surveys | Codes2) Topics | Race and Ethnicity | White–Note that the "White" filter displays below3) Geography | County | Arkansas | All Counties in Arkansas–Note that the "All counties in Arkansas" filter displays4) Search!5) Select Table Named RACEAmerican Community SurveyTotal PopulationTableID: B020016) Switch to 2016: ACS 5-Year Estimates Detailed Tables7) Customize Table. Download. Make Sure to Download 2016: ACS 5-Year Estimates Detailed Tables

Clean Census Data1) Create Data Dictionary2) Duplicate Sheet3) Four corners select and copy4) New Sheet. Paste Special | Transpose–the races are now the rows–Filter by Estimate: Contains Estimate, Delete5) Edit Headers: White, Black, Hispanic6) Check totals - do they add up?7) Two races including Some other race. Two races excluding Some other race, and three or more races (delete)8) Save and Load to Tableau9) Build a Arkansas Population Map by Race

Income by Race

https://data.census.govAdvanced SearchFilters | GeographyCounties | Arkansas | All countiesFilters | Topics | Income and PovertyFilters | Topics | Race and EthnicityFilters | Years | 2016Filters | Text Search in Find a Filter: "Income" | Select "Income (Households, Families, Individuals)Search

Download White Only, Black Only, Hispanic or Latino HouseholderYour tables will say this:HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN 2016 INFLATION-ADJUSTED DOLLARS) (WHITE ALONE HOUSEHOLDER)

Survey/Program: American Community Survey Product: 2016: ACS 1-Year Estimates Detailed Tables

Tables: B19001A, B19001B, B19001I Download - select .csv

TableauClean Data as described in previous lessonCombine the three tables in Tableau linking to the income as a common field.Create a chart

PAST TUTORIALS IN AMERICAN FACT FINDER.

NEED TO REVISE

https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml

Household income data for counties and state and national. Gender and demographics of low-wage workersAmerican FactFinderhttps://factfinder.census.gov/faces/nav/jsf/pages/inde
Search | Show Me AllTopics | People | PovertyGeographies | County | Arkansas | All Counties Within ArkansasSelect Table S1701, Poverty Status in the Past 12 MonthsModify Table—Select top Filter—Total and Percent Below Poverty Level—Select second Filter—Keep Estimate, do not check margin of ErrorDownload

30:00 shows how to use the fact finderhttps://www.census.gov/data/training-workshops/recorded-webinars/measuring-america.html

Selected Economic Characteristics DP03 2012-2016 American Community Survey 5-Year Estimates

Standard Data Cleaning

Data is: ACS_16_5YR_DP03 DP03 SELECTED ECONOMIC CHARACTERISTICS 2012-2016 American Community Survey 5-Year EstimatesCopy main data sheet and call copy wages below $25kDelete all data fields except headers and these columnsHC01_VC74 Estimate; INCOME AND BENEFITS (IN 2016 INFLATION-ADJUSTED DOLLARS) – Total householdsHC01_VC75 Estimate; INCOME AND BENEFITS (IN 2016 INFLATION-ADJUSTED DOLLARS) – Total households – Less than $10,000 HC03_VC75 Percent; INCOME AND BENEFITS (IN 2016 INFLATION-ADJUSTED DOLLARS) – Total households – Less than $10,000 HC01_VC76 Estimate; INCOME AND BENEFITS (IN 2016 INFLATION-ADJUSTED DOLLARS) – Total households – $10,000 to $14,999 HC03_VC76 Percent; INCOME AND BENEFITS (IN 2016 INFLATION-ADJUSTED DOLLARS) – Total households – $10,000 to $14,999 HC01_VC77 Estimate; INCOME AND BENEFITS (IN 2016 INFLATION-ADJUSTED DOLLARS) – Total households – $15,000 to $24,999 HC03_VC77 Percent; INCOME AND BENEFITS (IN 2016 INFLATION-ADJUSTED DOLLARS) – Total households – $15,000 to $24,999 HC01_VC85 Estimate; INCOME AND BENEFITS (IN 2016 INFLATION-ADJUSTED DOLLARS) – Total households – Median household income (dollars)–Rotate header rows, wrap text.Shrink verbiage from Estimate; INCOME AND BENEFITS (IN 2016 INFLATION-ADJUSTED DOLLARS) – Total households to “Total households”Total households %Total households

Total households – >$10k %Total households – >$10k Total households – $10kto $14,999 %Total households – $10kto $14,999 Total households – $15,000 to $24,999 %Total households – $15,000 to 24, 999$Median household income$– Specify Arkansas-stateThen find/replace to eliminate "County, Arkansas" from geography labels.Create Total Under $25 column.Add Total households – >$10k + Total households – $10k to $14,999 + Total households – $15,000 to $24,999Create % Under $25k column (total Under $25k / total households)Copy formulas downCheck mathWhen satisfied, copy and paste valuesMore on Data Cleaning Census spreadsheets –Download the view and the data versions of large spreadsheets. One to guide you. the other to do the work.–Merge / unmerge cells–Find-Replace— =CONCATENATE(B3, B4).

Census Demographic data

Household income data for counties and state and national. Gender and demographics of low-wage workersAmerican FactFinderhttps://factfinder.census.gov/faces/nav/jsf/pages/index.xhtmlAdva Search | Show Me AllTopics | People | Poverty | Poverty (added to Your Selections)Geographies | County | Arkansas | All Counties Within ArkansasGeographies | United StatesGeographies | ArkansasSelect Table S1701, Poverty Status in the Past 12 MonthsModify Table—Select top Filter—Total and Percent Below Poverty Level—Select second Filter—Keep Estimate, do not check margin of ErrorDownload—Use the Data-Download Again—View the Data—Excel spreadsheetQuestions about categories and definitions:See "Table Notes" to far right on factfinder website after you've generated a table.https://www2.census.gov/programs-surveys/acs/tech_docs/subject_definitions/2016_ACSSubjectDefinitions.pdfRead: "Poverty Status in the Past 12 Months""Poverty Status of Households"Definitions. working Poor–Poverty thresholds:The actual poverty thresholds vary with the makeup of the family. In 2015, the weighted average poverty threshold for a family of four was $24,257; for a family of nine or more people, the threshold was $49,177; and for one person (see Unrelated individuals), it was $12,082. Poverty thresholds are updated each year to reflect changes in the Consumer Price Index for All Urban Consumers (CPI-U). Thresholds do not vary geographically. (For more information, see "Income and poverty in the United States: 2015.")https://www.bls.gov/opub/reports/working-poor/2015/home.htm#unrelatedindividualWeighted Average PovertyThresholds in 2015 by Size ofFamily(Dollars)One person 12,082Two people 15,391Three people 18,871Four people 24,257Five people 28,741Six people 32,542Seven people 36,998Eight people 41,029Nine people or more 49,177Source: U.S. Census Bureau.https://www.census.gov/content/dam/Census/library/publications/2016/demo/p60-256.pdfhttps://www.census.gov/data/tables/time-series/demo/income-poverty/historical-poverty-thresholds.html

–Download the view and the data versions of large spreadsheets. One to guide you. the other to do the work.

–Merge / unmerge cells

–Find-Replace

— =CONCATENATE(B3, B4).

Cleaned and download 2011-2015 estimates with detailed poverty metricsArk Counties full income search 5-10-17 ACS_15_5YR_DP03

Students assigned geographical location for Census data.

Questions:

–Number and Percentage of Minimum Wage Households?

–Compare to National, State Averages

–Produce basic Tableau chart

# Chapter 5

# Data Visualization

Principles of Data Visualization“Our limited brains are incapable of grasping reality in all its glorious complexity.”What you design is never exactly what your audience ends up interpreting so reducing the chances for misinterpretation becomes crucial. Cleveland McGill Scale

Cleveland McGill Scale for Data Visualization

https://www.youtube.com/watch?v=XGPkdOczRtk&feature=youtu.be

Important Resources for Surveying the Data Visualization Options Dataviz Catalog http://www.datavizcatalogue.com/ FT Visual Vocabulary https://github.com/ft-interactive/chart-doctor/blob/master/visual-vocabulary/Visual-vocabulary.pdf For the weekly memo: Select two examples from the Dataviz Catalog and FT Visual Vocabulary that you find interesting or useful. Include a screenshot of the chart in your memo and describe how it could apply for our project.

Graphics Comments from Jon Schleuss, Los Angeles TimesAnd the colors. What does “red” mean when it’s used? And what about using too many colors. At the Times we really only have two or three colors: basic default, a highlight color and a negative color. We break from convention, but keeping it simple helps. I figure now I’ll show them how I approach chart building from start to finish.Also, I see a desire to combine different data into the same chart. So there’s a left, bottom and right axis. But that’s a bit confusing to the reader, especially when things have the same values (percentages vs. percentages instead of percentages vs. hard counts).I think my big takeaways are that most of these charts should be flipped on their sides. That’s because when we sort data largest to smallest (nearly everyone did) we then think of it as time passing if it’s a column chart (bars situated left to right). And that there’s a downward marching trend. Best to flip a lot of these on their side. Build choropleth maps

A Comment on Color Choices A reader made an excellent point that the red

shading of the active cases map was misleading since 'red zone' is a specific concept in the White House task force reports. Our shading does not match the red zone definition of the task force report and most readers would expect that it would. I swapped out the shading for blue-green until we figure out the calculations for active cases per capita on that White House task force scale. It just goes to show you how color choices on graphics are major communication issues. Reader Comment: **?** Have you considered using a standard for which counties are red? The map gives a false impression that Arkansas counties are not "red zones" for CoVID when they are. Replying to **?** and **?>**

Build a Cover Image Using Canva or InDesign or Powerpoint https://www.canva.com/

Rachell Sanchez-Smith used Canva for a simple animation.

Higher Resolution Graphics in Tableau https://www.dataplusscience.com/HighResolution.html
Higher Resolution Photos: at least 72 DPI. Practically, should be higher. –500k or more is a safe bet –Cropping reduces file size. Grainy Guidance: https://cft.vanderbilt.edu/wp-content/uploads/sites/59/Image_resolutions.pdf
Higher Resolution Graphics in Tableau https://www.dataplusscience.com/HighResolution.html

# Chapter 6

# Writing About Data

Writing Style Notes Writing style notesDon't use this "respectively" construction. It confuses the readers and leads to data errors.Other private trade schools also made the top 10 list, such as Philander Smith in Little Rock and Bryan University in Rogers, with increases of 81 percent and 74 percent respectively. .ITT Technical Institute, the University of phoenix, Philander Smith College and Bryan University are all classified as private schools and are the top four schools with graduated student loan debt in the state. Their debt has increased 106 percent, 102 percent, 81 percent and 74 percent respectively since 2012.

AP Style with Numbers

AP Numerals EntryDownload

# Chapter 7

# Excel Bootcamp

PIVOT TABLES

Basic introduction to pivot tables

Calculations and Pivot Tables

COUNTIF Function for Tabulating Text

https://youtu.be/5NdImlnWmsI

# Chapter 8

# Basic Tableau

This module addresses:–Downloading instructions for Tableau–Getting started tutorial with video–Building a basic COVID data chart with video and transcript–Using filters and calculations with video and transcript–Tutorial on Tableau calculations with video–Proper formatting of a filter bar in Tableau, videoLinks to additional Tableau Tutorials

Dashboards and Embedding Tableau Public in WordPress

Maps in Tableau

Tableau Download and License

Introduction to Tableau

Build Your First Tableau Chart - Deaths by Counties

This chart was filtered to show counties with deaths 25 and greater

Formatting a Graphic

Filters and Calculations

More on Tableau, Basic Excel Skills: FBI Data

Tableau Calculations

How to Tweak A Filter to Display Properly

Additional Tableau Tutorials

Dashboards and Embedding Tableau Public in WordPress

Maps in Tableau -

Dual Axis

MATERIAL ON THIS PAGE USES NON-COVID DATA EXAMPLES

Histogram

This is a detailed walk-through of how to add a cumulative distribution to a histogram in Tableau. Build a histogram with % Female Households in PovertyEasiest method: click on% Female Households in Poverty and choose the histogram from Show MeFind out where this is happening: drag Geography to the pane. Now hover your mouse over the blocks and you will see the counties. Duplicate the CNT(Profit) pill on Rows by Ctrl+drag the pill next to itself on the Rows shelfThis gives you a new marks card – now you can create 2 different types of mark for a single data field. On the new marks card "CNT(Profit)(2)", change the mark type to a lineChange the color as well, if desiredRight click on the duplicated field on the Rows shelf and make it a dual axisApply a Quick Table Calculation of Running Total to the duplicated field[[If students find this histogram to be uninteresting, add a filter on Profit to ignore outliers (maybe go from -500 to +500). Then right-click Profit(bin) dimension and Edit to make the bin size smaller, maybe 25.]]Create a Reference Line for PovertyStep 1 – Build the ViewDrag % Female Households – Children 5 Years to the Rows shelf.Drag Geography to the Columns shelf.Step 2 – Create ParametersRight-click in the Data pane and then select Create Parameter.Name the parameter "Arkansas Average".Under Data Type select Integer.Under Current Value, set to 55.8Under Allowable values select All.Click OK.Step 3: Create the calculated fieldSelect Analysis > Create Calculated Field.Name the calculated field "Reference Line".In the formula field, enter the following formula:IF[% Female Households – Children 5 years younger]=[Arkansas Average] THEN [Arkansas Average] ENDClick OK.Step 4 – Use the calculated field as a Parameter ControlDrag the "Reference Line "calculated field to Details. This is the box below Color in the Marks CardClick the arrow to change the measure from SUM to Minimum.In the view, right-click on the Y axis and select Add Reference Line.In the Value drop down menu, select Minimum(Reference Line).In the Label drop-down menu, select Value.Click OK.Visualize the income distribution in a histogram% of Female Households – Children 5 Years and Younger.–Drag to Columns–Show me: Select Histogram–Distribution of the poverty level.

Bins and GroupsCreate Bins: https://onlinehelp.tableau.com/current/pro/desktop/en-us/calculations_bins.htmOther options besides bins: Use parameters to organize data:%C2%A0;https://onlinehelp.tableau.com/current/pro/desktop/en-us/parameters_create.htm Use sets to organize data:%C2%A0;https://onlinehelp.tableau.com/current/pro/desktop/en-us/sortgroup_sets_create.htm#UseCreate Bins: https://onlinehelp.tableau.com/current/pro/desktop/en-us/calculations_bins.htmFormat Filters – See video belowUse parameters to organize data:%C2%A0https://onlinehelp.tableau.com/current/pro/desktop/en-us/parameters_create.htmUse sets to organize data:%C2%A0https://onlinehelp.tableau.com/current/pro/desktop/en-us/sortgroup_sets_create.htm#Use

##Need to Edit below, may duplicate the earlier section **Track #1: Tableau**

**Arkansascovid.com Data**

# Chapter 9

# Flourish

Flourish

# Chapter 10

# WordPress

Looks like this:

Building the Web PageGutenberg has blocks to display text, graphics, video etc.Everyone has to build their own page in DataReporting site of WordPress.Tasks:Create a Post.Upload your graphic from Assignment #1 and your story. Format so it doesn't look uglyClick "Student Work" for category.

# Chapter 11

# Datawrapper

# Chapter 12

# R Introduction

Beginner's guide to R: Introduction

https://www.computerworld.com/article/2497143/business-intelligence/business-intelligence-beginner-s-guide-to-r-introduction.html

RStudio IDE Easy Tricks You Might've Missed

https://rviews.rstudio.com/2016/11/11/easy-tricks-you-mightve-missed/

How Do I?

https://smach.github.io/R4JournalismBook/HowDoI.html

Packages

https://smach.github.io/R4JournalismBook/packages.html

Download R and RStudio

http://www.machlis.com/R4Journalists/download-r-and-rstudio.html

You can download the most recent version of R at https://www.r-project.org/, which is the home of R (formally known as the R Project for Statistical Computing). The R-project home page usually includes information about the latest versions of R. Don't be put off by the sometimes odd nicknames for R versions, such as "Very, Very Secure Dishes" and "Bug in Your Hair" – the software is much more useful than you might assume from the nicknames. (The whimsical version names come from various Peanuts cartoons.)

There should also be a prominent link to download R. Click that download option and you should be taken to CRAN, the Comprehensive R Archive Network, and a list of CRAN servers, called mirrors, around the world. Pick a server and choose the precompiled binary distribution for your operating system. Once the file finishes downloading, install it like any other software program - run the .exe for Windows or .pkg for Mac.

You should be fine accepting all the Mac defaults. On Windows, you'll need to decide whether you want the 32- or 64-bit R version. (Unless you've got a pretty old system, chances are you'll want 64-bit.)

This is all you need to start running R, but I strongly recommend also installing RStudio, a free platform designed to make it easier and more enjoyable to create and run R code. Head to RStudio.com and under products, look for RStudio and then RStudio Desktop (not Server), and download the free Open Source Edition version for your operating system. This, too, installs like a typical software program

R NOTES THAT NEED TO BE CLEANED UP

Today's set of tasks:

March 29 class-2lrqupg

Data

US Ark Counties Poverty ACS_16_5YR_DP03-Jan 24-y46vv7

Race Poverty Set

ACS_16_5YR_S1701_with_ann-28nvp6q

File Header Definitions

ACS_16_5YR_S1701_metadata-1on99rr

R4JournalismBook

https://github.com/smach/R4JournalismBook

https://smach.github.io/R4JournalismBook/booklinks.html

Functions

https://smach.github.io/R4JournalismBook/functions.html

[et_pb_section bb_built="1"][et_pb_row][et_pb_column type="4_4"][et_pb_text admin_label="Key Resources for R" _builder_version="3.17.2"]

You will be using these materials

a lot during the course of the semester.

Machlis, Sharon. Practical R for Mass Communication and Journalism. Chapman & Hall/CRC, 2018. http://www.machlis.com/R4Journalists/.

RStudio IDE Easy Tricks You Might've Missed https://rviews.rstudio.com/2016/11/11/easy-tricks-you-mightve-missed/

How Do I? https://smach.github.io/R4JournalismBook/HowDoI.html

Functions: https://smach.github.io/R4JournalismBook/functions.html

Packages: https://smach.github.io/R4JournalismBook/packages.html

[/et_pb_text][/et_pb_column][/et_pb_row][/et_pb_section]

**R and R Studio**

Install R and R Studio.

This is free and open source software. It is not large and doesn't tax the memory a lot.
R runs on Windows, Mac and Linux, but this course is designed for the Mac version.
If you use Windows, there may be variations in the lessons and instructions. Please see me f

Installing R is a two-step process:
1) Install R, the actual program
2) Install RStudio, a common interface

1) Download the most recent version of R for Mac:

https://mirrors.nics.utk.edu/cran/bin/macosx/R-4.0.2.pkg

--If you have a Windows computer, go to:

https://mirrors.nics.utk.edu/cran/bin/windows/base/R-4.0.2-win.exe

Accept all of the default settings for Mac.

2) Install RStudio, the interface we use to manage and create R code. Download the open

https://rstudio.com/products/rstudio/download/#download

Good instructions for installing R

http://www.machlis.com/R4Journalists/download-r-and-rstudio.html

Good overview of the program

https://docs.google.com/presentation/d/1O0eFLypJLP-PAC63Ghq2QURAnhFo6Dxc7nGt4y_
l90s/edit#slide=id.p

# Chapter 13

# R Data Viz

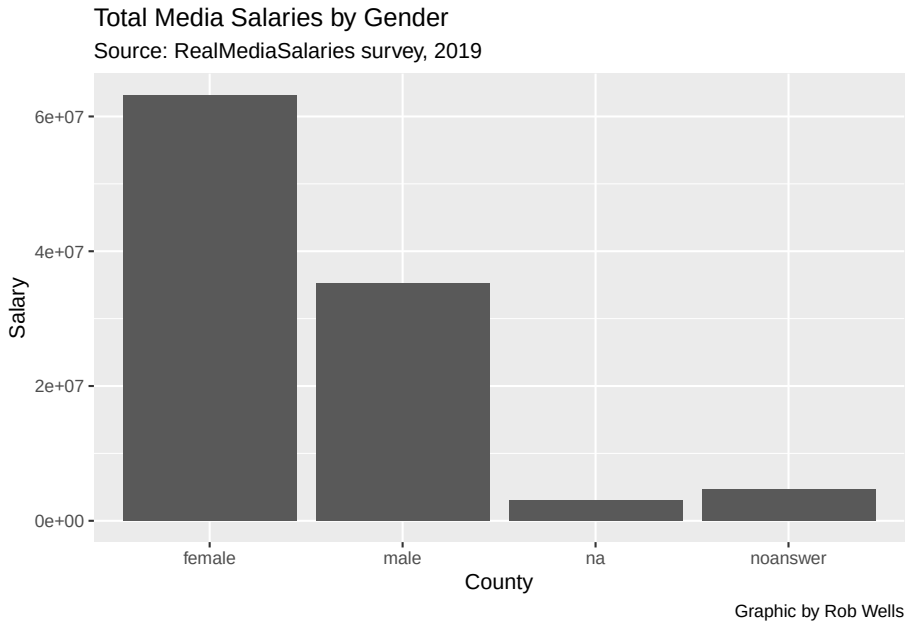
# Chapter 14

# Part 6: Chart By Gender

```
MediaBucks <- rio::import("https://github.com/profrobwells/Guest_Lectures/blob/master/Intro_To_R/
```

```
MediaBucks %>% ggplot(aes(y = Salary, x=Gender)) +
  geom_bar(stat = "identity") +
  labs(title = "Total Media Salaries by Gender",
       subtitle = "Source: RealMediaSalaries survey, 2019 ",
       caption = "Graphic by Rob Wells",
       x="County",
       y="Salary")
```

```
## Warning: Removed 4 rows containing missing values (position_stack).
```



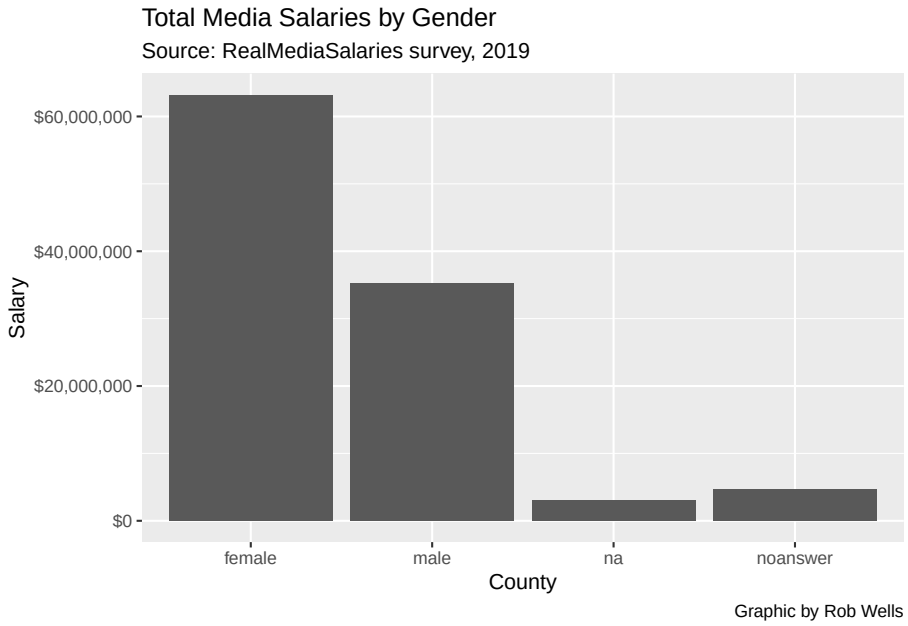

This needs some work

- **Get rid of scientific notation**

```
options("scipen"=100, "digits"=4)
```

```
MediaBucks %>% ggplot(aes(y = Salary, x=Gender)) +
  geom_bar(stat = "identity") +
   scale_y_continuous(labels = scales::dollar) +
  labs(title = "Total Media Salaries by Gender",
       subtitle = "Source: RealMediaSalaries survey, 2019 ",
       caption = "Graphic by Rob Wells",
       x="County",
       y="Salary")
```

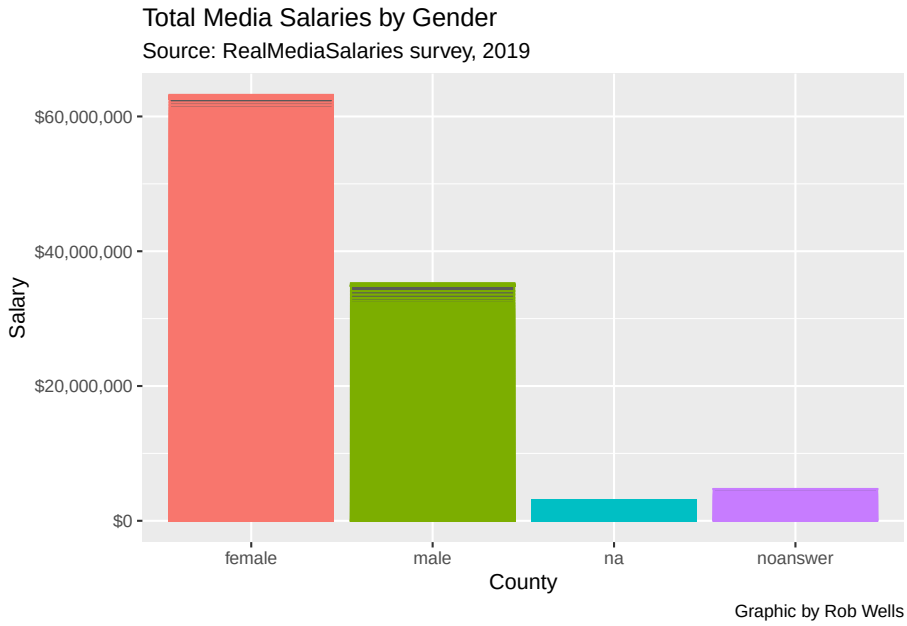
```
## Warning: Removed 4 rows containing missing values (position_stack).
```




- **Add Color**

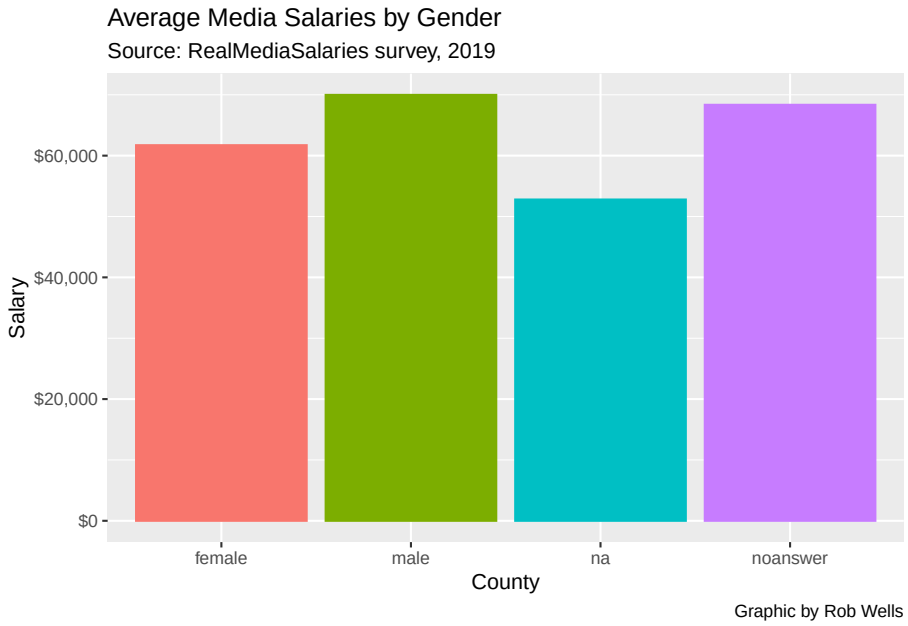
```
MediaBucks %>% ggplot(aes(y = Salary, x=Gender, color = Gender)) +
  geom_bar(stat = "identity") +
  theme(legend.position = "none") +
  #coord_flip() +     #this makes it a horizontal bar chart instead of vertical
  scale_y_continuous(labels = scales::dollar) +
  labs(title = "Total Media Salaries by Gender",
       subtitle = "Source: RealMediaSalaries survey, 2019 ",
       caption = "Graphic by Rob Wells",
       x="County",
       y="Salary")
```

```
## Warning: Removed 4 rows containing missing values (position_stack).
```




- **Average salaries is the story!**

```
MediaBucks %>%
  select(Gender, Salary) %>%
  group_by(Gender) %>%
  summarize(mean = mean(Salary, na.rm=TRUE))  %>%
  ggplot(aes(y = mean, x=Gender, color = Gender, fill=Gender)) +
  geom_bar(stat = "identity") +
  theme(legend.position = "none") +
  scale_y_continuous(labels = scales::dollar) +
  labs(title = "Average Media Salaries by Gender",
       subtitle = "Source: RealMediaSalaries survey, 2019 ",
       caption = "Graphic by Rob Wells",
       x="County",
       y="Salary")
```




- **Export: Lower right, Export as .png file**

#Machlis - Basic Data Visualization #Wickham - MPG charts #Updated Feb 4 2019

#load software - Select NO when asked to restart install.packages("ggplot2") install.packages("dplyr") install.packages("usethis") install.packages("forcats")

#call software into memory library(ggplot2) library(dplyr) library(usethis) library(forcats)

#Basic demo #You will run the commands from the Console below demo(topic="graphics")

library(dplyr)

#Tutorial #Import Data, Create Dataframe, Rename Columns snowdata <- rio::import("data/BostonChicagoNYCSnowfalls.csv") bostonsnow <- select(snowdata, Winter, Boston) names(bostonsnow)[2] <- "TotalSnow"

#Doing the same thing but with pipe function bostonsnow2 <- select(snowdata, Winter, Boston) %>% rename(TotalSnow = Boston)

#Doing the same thing but more efficiently bostonsnow3 <- select(snowdata, Winter, TotalSnow = Boston)

#Basic graphs plot(bostonsnow$TotalSnow)

hist(bostonsnow$TotalSnow) boxplot(bostonsnow$TotalSnow) barplot(bostonsnow$TotalSnow) barplot(sort(bostonsnow$TotalSnow, decreasing = TRUE))

#qplot qplot(data=bostonsnow, y = TotalSnow) qplot(y = bostonsnow$TotalSnow)

#basic ggplot2 - boxplot ggplot(data=snowdata) + geom_boxplot(aes(x = "Boston", y = Boston))

#dual box plots ggplot(data=snowdata) + geom_boxplot(aes(x = "Boston", y = Boston)) + geom_boxplot(aes(x = "Chicago", y = Chicago))

#bring in snowdata tidy snowdata_tidy <- rio::import("data/snowdata_tidy.csv")

#view a tidy table View(snowdata_tidy)

#Boxplot with ggplot ggplot(snowdata_tidy, aes(x = City, y = TotalSnow)) + geom_boxplot()

#Line graphs ggplot(snowdata_tidy, aes(x = Winter, y = TotalSnow, group = City)) + geom_line()

#ggplot with colors and points

ggplot(snowdata_tidy, aes(x = Winter, y = TotalSnow, group = City, color = City)) + geom_line()

#ggplot with colors and points ggplot(snowdata_tidy, aes(x = Winter, y = TotalSnow, group = City, color = City)) + geom_line() + geom_point()

#Filtered for two years, 1999 and 2000 snowdata_tidy21 <- filter(snowdata_tidy, Winter >= "1999-2000") ggplot(snowdata_tidy21, aes(x = Winter, y = TotalSnow, group = City, color = City)) + geom_line() + geom_point()

#Barplots ggplot(data = snowdata_tidy21, aes(x = Winter, y = TotalSnow, group = City, color = City)) + geom_col()

#Not so ugly bars ggplot(data = snowdata_tidy21, aes(x = Winter, y = TotalSnow, group = City, fill = City)) + geom_col(position = "dodge")

#————————————————————————————-# #Build a chart - Snow #————————————————————————————-#

library(ggplot2) SnowChartBoston <- ggplot(bostonsnow, aes(x = reorder(Winter, TotalSnow), y = TotalSnow)) + geom_bar(stat = "identity") + coord_flip() + labs(title = "Snow", subtitle = "lots of it", caption = "Graphic by Rob Wells", x="Years", y="snow in inches") plot(SnowChartBoston)

#Notes from R for Data Scientists - Wickham #https://r4ds.had.co.nz/

#Feb. 2 2019 install.packages('tidyverse') install.packages(c("nycflights13", "gapminder", "Lahman"))

#If we want to make it clear what package an object comes from, we'll use the package name followed by two colons, like dplyr::mutate(), or #nycflights13::flights. This is also valid R code.

library(tidyverse)

#Do cars with big engines use more fuel than cars with small engines? #displ, a car's engine size, in litres. #hwy, a car's fuel efficiency on the highway, in miles per gallon (mpg). #A car with a low fuel efficiency consumes more fuel than a car with a high fuel efficiency when they travel the same distance. #To learn more about mpg, open its help page by running ?mpg.

mpg

mpg <- as_data_frame(mpg) View(mpg)

ncol(mpg)

#Create a ggplot ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy))

#3.2.3 A graphing template #Let's turn this code into a reusable template for making graphs with ggplot2. #To make a graph, replace the bracketed sections in the code below with a dataset, a geom function, or a collection of mappings. #ggplot(data = ) + # (mapping = aes())

ggplot(data = mpg)

#using color to distinguish class ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy, color = manufacturer, size=hwy))

str(mpg)

#MPG categorical = manufacturer model cyl trans drv fl class #continuous = disply cty hwy #Map a continuous variable to color, size, and shape. #How do these aesthetics behave differently for categorical vs. continuous variables? #Answer - they do not come up in discrete blocks by on a spectrum range

#Color by manufacturer ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy, color =manufacturer))

#Color and Size ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy, color =manufacturer, size=manufacturer))

#adding size ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy, size = class, color = class))

#What does the stroke aesthetic do? What shapes does it work with? (Hint: use ?geom_point) ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy, color = hwy))

#What happens if you map an aesthetic to something other than a variable name, #like aes(colour = displ < 5)? Note, you'll also need to specify x and y.

ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy, color = displ < 3))

#It gives you a true false by color

#Map two variables #Example #https://stackoverflow.com/questions/3777174/plotting-two-variables-as-lines-using-ggplot2-on-the-same-graph

# Chapter 15

test_data <- data.frame( var0 = 100 + c(0, cumsum(runif(49, -20, 20))), var1 = 150 + c(0, cumsum(runif(49, -10, 10))), date = seq(as.Date("2002-01-01"), by="1 month", length.out=100) )

ggplot(test_data, aes(date)) + geom_line(aes(y = var0, colour = "var0")) + geom_line(aes(y = var1, colour = "var1"))

#Apply to mpg ggplot(data=mpg, aes(hwy)) + geom_point(aes(y = displ, colour = "displ")) + geom_point(aes(y = cyl, colour = "cyl"))

#Exercise: #Exercise with ArkCo_Income_2017 #1. Create a plot chart with the top 10 counties with the greatest percentage of low-income population #Your answer should look like this https://bit.ly/2BgPmyo

#2. Create a plot chart with the top 10 counties with the greatest percentage of upper-income population

## 15.1 # R Data Cleaning

title: "SFPD Calls for Service_Feb 23,2020" author: "Rob Wells" date: "2/23/2020"

---

### 15.1.1 Reporting on Homelessness: Data Analysis for Journalists

### 15.1.2 Jour 405v, Jour 5003, Spring 2020

# Chapter 16

# Analysis of San Francisco Police Calls for Service Data

- **Here is the original dataset: 3,048,797 records**

https://data.sfgov.org/Public-Safety/Police-Department-Calls-for-Service/hz9m-tj6z/data

- **This tutorial uses a subset of this data**

  The Calls for Service were filtered as follows: CONTAINS homeless, 915, 919, 920: Downloaded 157,237 records 3/31/16 to 11/30/2019. This is 5.1% of all calls in the broader database. File renamed to: SF_311_Jan29.xlsx

---

# Chapter 17

# Part 1: Quick Start

```
library(tidyverse)
library(janitor)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

Reload Data

```
SF <- rio::import("https://github.com/profrobwells/HomelessSP2020/blob/master/Data/SF_311_Jan29.x
```

- **Clean names, Process dates**

```
SF <- janitor::clean_names(SF)
#Process dates
SF$call_date2 <- ymd(SF$call_date)
SF$year <- year(SF$call_date2)
```

- **Process dates**

```
Days <- SF %>%
  count(call_date2) %>%
  group_by(call_date2) %>%
  arrange(desc(n))
```

- **Types of Crimes**

```
Types <- SF %>% count(original_crime_type_name) %>%
  group_by(original_crime_type_name) %>%
  arrange(desc(n))
```

- **Calls by Year**

```
Years <- SF %>%
  count(year) %>%
  group_by(year) %>%
  arrange(desc(year))
```

- **Actions Taken**

```
Action <- SF %>%
  count(disposition) %>%
  arrange(desc(n))
```

# Chapter 18

# Part 2: Cleaning & Analysis

- **Question**: How many rows? Columns? Supply a list of the column names

nrow(SF) [1] 157237 > ncol(SF) [1] 14 Process dates, check file types

```
str(SF)
```

```
## 'data.frame':    157237 obs. of  16 variables:
##  $ crime_id                : num  190040497 190200644 190213959 190271011 190141952 ...
##  $ original_crime_type_name: chr  "919" "Homeless Complaint" "Homeless Complaint" "915" ...
##  $ report_date             : POSIXct, format: "2019-01-04" "2019-01-20" ...
##  $ call_date               : POSIXct, format: "2019-01-04" "2019-01-20" ...
##  $ offense_date            : POSIXct, format: "2019-01-04" "2019-01-20" ...
##  $ call_time               : POSIXct, format: "1899-12-31 06:58:00" "1899-12-31 06:19:00" ...
##  $ call_date_time          : POSIXct, format: "2019-01-04 06:58:00" "2019-01-20 06:19:00" ...
##  $ disposition             : chr  "HAN" "HAN" "ADV" "HAN" ...
##  $ address                 : chr  "400 Block Of Jones St" "8th And Market" "Mission St/24th St
##  $ city                    : chr  "San Francisco" NA "San Francisco" "San Francisco" ...
##  $ state                   : chr  "CA" "CA" "CA" "CA" ...
##  $ agency_id               : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ address_type            : chr  "Premise Address" "Geo-Override" "Intersection" "Intersectio
##  $ common_location         : chr  NA NA NA NA ...
##  $ call_date2              : Date, format: "2019-01-04" "2019-01-20" ...
##  $ year                    : num  2019 2019 2019 2019 2019 ...
```

Examine how we have created a new date and year column and how they are formatted differently than the rest We can now perform date and year calculations Create Days Table

- **Question**: Using the summary() function, describe the minimum, maximum, median and mean of calls in the Days table

```
summary(Days)
```

```
##    call_date2                 n
##  Min.   :2016-03-31   Min.   : 10
##  1st Qu.:2017-02-28   1st Qu.: 86
##  Median :2018-01-29   Median :119
##  Mean   :2018-01-29   Mean   :117
##  3rd Qu.:2018-12-30   3rd Qu.:148
##  Max.   :2019-11-30   Max.   :232
```

Between March 31, 2016 and Nov. 30, 2019, San Francisco residents placed **an average 117 calls** to police complaining about homeless people.

- **Question**: Which day had the most calls? Which day had the least?

```
Days %>%
  filter(n == 232)
```

```
## # A tibble: 1 x 2
## # Groups:   call_date2 [1]
##   call_date2     n
##   <date>     <int>
## 1 2019-08-15   232
```

```
Days %>%
  filter(n == 10)
```

```
## # A tibble: 1 x 2
## # Groups:   call_date2 [1]
##   call_date2     n
##   <date>     <int>
## 1 2016-03-31    10
```

Examine the types of events

```
Types <- SF %>% count(original_crime_type_name) %>%
  group_by(original_crime_type_name) %>%
  arrange(desc(n))
```

- **Question**: What are the top five complaints in this data and provide the number of complaints

```
Types <- SF %>% count(original_crime_type_name) %>%
  group_by(original_crime_type_name) %>%
  arrange(desc(n))
```

Create separate table with just the top five counties' crime rate: dplyr has a "top_n" function that i find handy

```
Types <- SF %>%
  count(original_crime_type_name) %>%
  top_n(5, n) %>%
  arrange(desc(n))
```

Export a table into a spreadsheet (csv is a comma separated file)

```
write.csv(Days,"Days.csv")
```

Build a table totalling the number of complaints by year

```
Years <- SF %>%
  count(year) %>%
  group_by(year) %>%
  arrange(desc(year))
```

- **EXERCISE: Grouping by Disposition**

Look at the Radio Codes spreadsheet under dispositions

https://data.sfgov.org/api/views/hz9m-tj6z/files/b60ee24c-ae7e-4f0b-a8d5-8f4bd29bf1de?download=true&filename=Radio%20Codes%202016.xlsx

Total by disposition

```
Action <- SF %>%
  count(disposition) %>%
  arrange(desc(n))
```

Ceate a table with serious infractions described in disposition

Example: Here's a table filtering the dispositions column to show "no disposition" or "gone on arrival"

```
Nothing <- SF %>%
  filter(disposition == "ND" | disposition == "GOA")
```

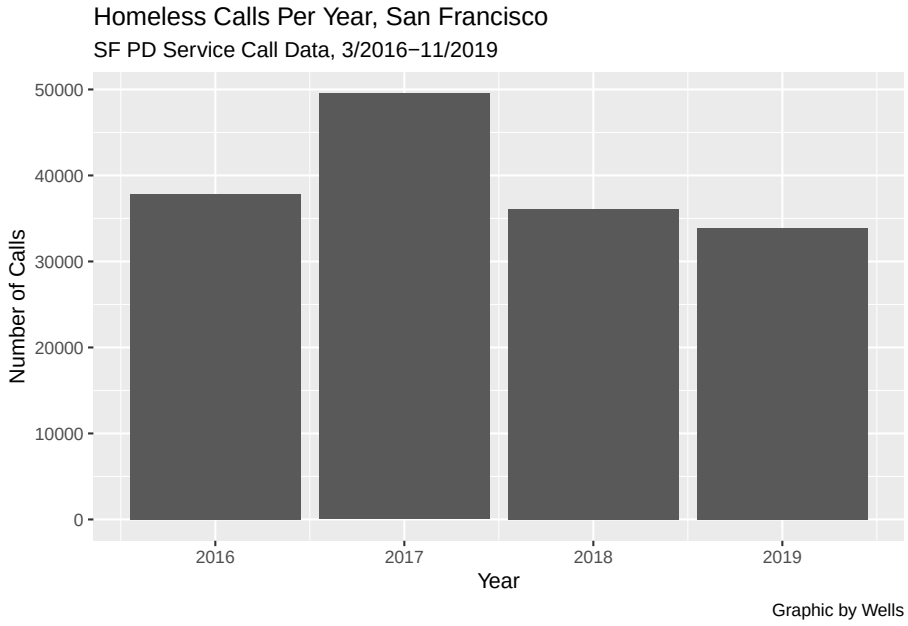
- **Question**: Create a table with the serious actions including citations and arrests police took in the dispositions

Arrest, Cited, Criminal Activation, SF Fire Dept Medical Staff engaged

```
Busted <- SF %>%
  filter(disposition == "ARR" | disposition == "CIT" | disposition == "CRM" | disposit
  count(disposition) %>%
  arrange(desc(n))
```

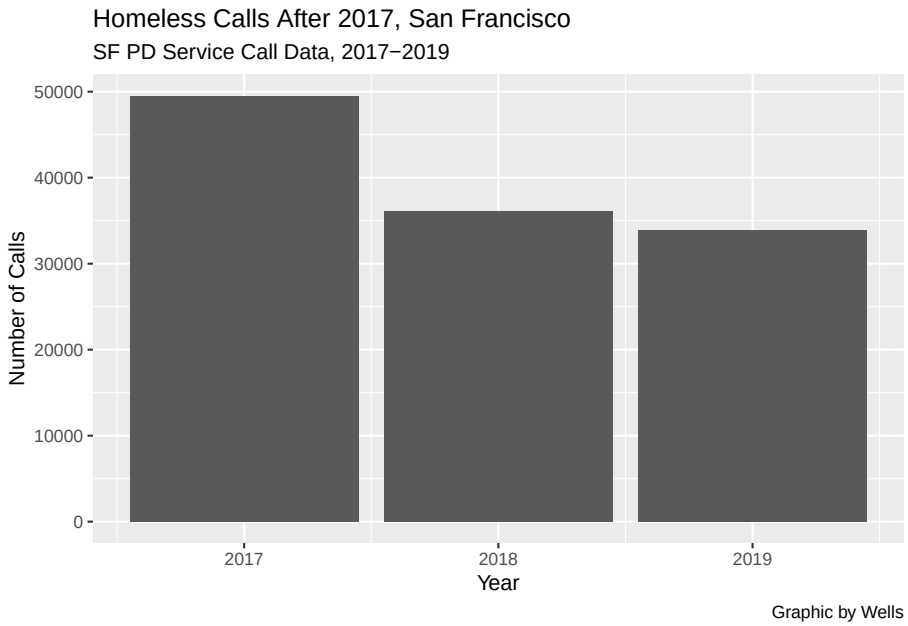
- **EXERCISE** - A Basic chart of the crime data

```
ggplot(Years, aes(x = year, y = n)) +
  geom_bar(stat = "identity") +
  #coord_flip() +    #this makes it a horizontal bar chart instead of vertical
  labs(title = "Homeless Calls Per Year, San Francisco",
       subtitle = "SF PD Service Call Data, 3/2016-11/2019",
       caption = "Graphic by Wells",
       y="Number of Calls",
       x="Year")
```



A chart using a dplyr filtering language

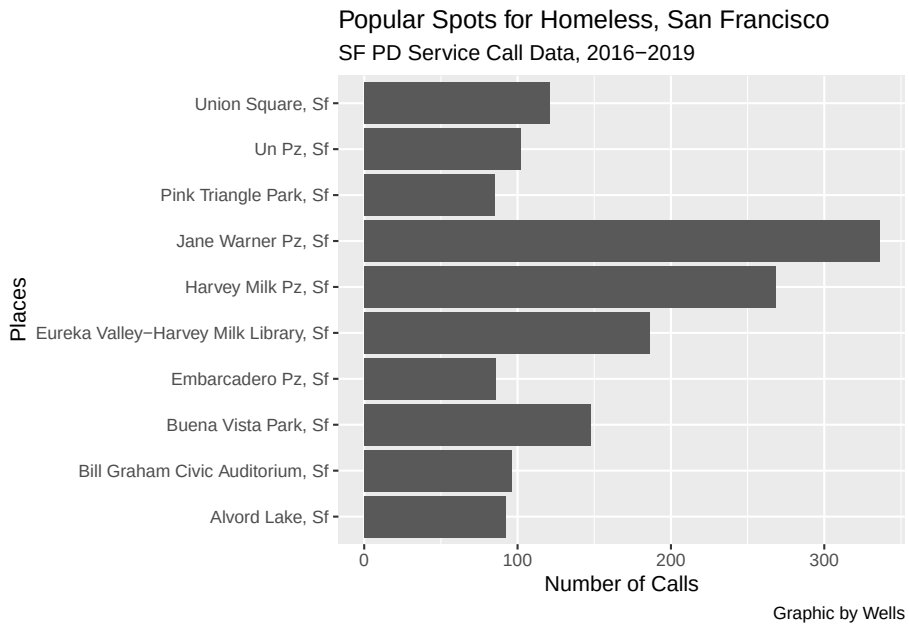
```
Years %>%
  filter(year >= 2017) %>%
  ggplot(aes(x = year, y = n)) +
  geom_bar(stat = "identity") +
  #coord_flip() +    #this makes it a horizontal bar chart instead of vertical
  labs(title = "Homeless Calls After 2017, San Francisco",
       subtitle = "SF PD Service Call Data, 2017-2019",
       caption = "Graphic by Wells",
       y="Number of Calls",
       x="Year")
```




A more complex filter

```
SF %>%
  filter(!is.na(common_location)) %>%
  count(common_location) %>%
  top_n(10, n) %>%
  ggplot(aes(x = common_location, y = n)) +
  geom_bar(stat = "identity") +
  coord_flip() +    #this makes it a horizontal bar chart instead of vertical
  labs(title = "Popular Spots for Homeless, San Francisco",
       subtitle = "SF PD Service Call Data, 2016-2019",
       caption = "Graphic by Wells",
```

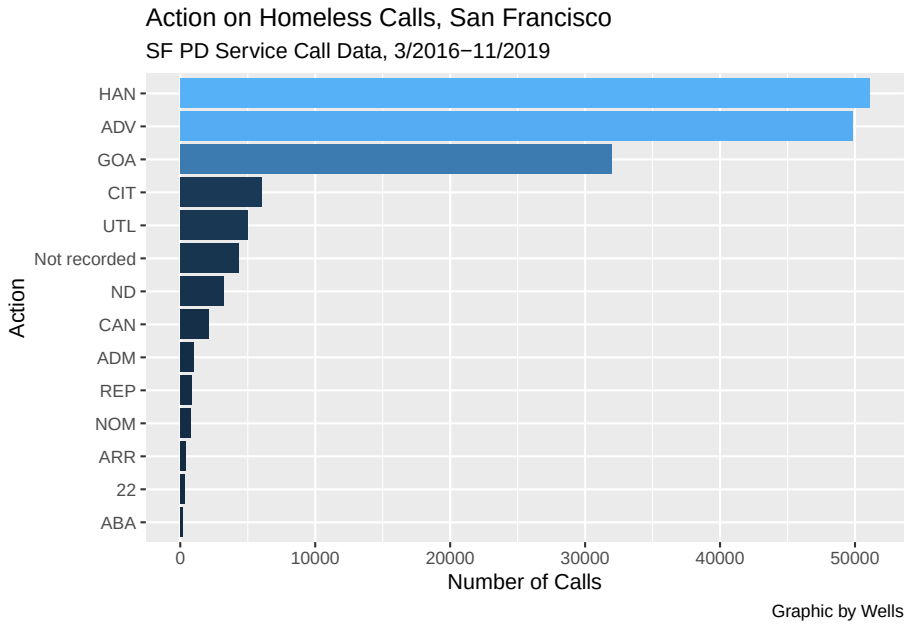
```
       y="Number of Calls",
       x="Places")
```




- **Question**: Chart the total dispositions.

Filter for at least 100 actions. Add color, export image to Blackboard.

```
Action %>%
  filter(n > 100) %>%
  ggplot(aes(x = reorder(disposition, n), y = n, fill=n)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  coord_flip() +    #this makes it a horizontal bar chart instead of vertical
  labs(title = "Action on Homeless Calls, San Francisco",
       subtitle = "SF PD Service Call Data, 3/2016-11/2019",
       caption = "Graphic by Wells",
       y="Number of Calls",
       x="Action")
```


Action on Homeless Calls, San Francisco
SF PD Service Call Data, 3/2016−11/2019
Action
HAN
ADV
GOA
CIT
UTL
Not recorded
ND
CAN
ADM
REP
NOM
ARR
22
ABA
0
10000
20000
30000
40000
50000
Number of Calls
Graphic by Wells

# Chapter 19

# Part 3: Cleaning Dispositions

Making our charts less ugly

The disposition column is in cop-speak. We need to clean it up

Step #1: Duplicate the column you want to mess with

```
SF$disposition1 <- SF$disposition
```

Rename specific strings. Example:

str_replace_all(test.vector, pattern=fixed('-'), replacement=fixed(':') )

Details on string manipulation:

https://dereksonderegger.github.io/570L/13-string-manipulation.html

We can do this to replace ABA with "Abated"

```
SF$disposition1 <- str_replace_all(SF$disposition1, pattern=fixed('ABA'), replacement=fixed('Abat
#Again with ADM
SF$disposition1 <- str_replace_all(SF$disposition1, pattern=fixed('ADM'), replacement=fixed('Admo
```

We can do that 19 times. OR….

Look at this example using a lookup table to replace all the values
https://stackoverflow.com/questions/50615116/renaming-character-variables-in-a-column-in-data-frame-r

Build a table to translate the Cop Speak to English:

```
dispo_lkup <- c(ABA="Abated", ADM="Admonish", ADV="Advised", ARR="Arrest", CAN="Cancel
                CIT="Cited", CRM="Criminal", GOA="Gone", HAN="Handled", NCR="No_Crimina
                NOM="No_Merit", PAS="PlaceSecure", REP="Report", SFD="Medical", UTL="Un
#22="Cancel" was handled differently because it is a numeric value: '22'="Cancel"
#This scans "disposition", finds ABA and replaces with Abated, finds ARR, replaces wit
SF$disposition1 <- as.character(dispo_lkup[SF$disposition])
```

Rerun Action with disposition1

```
Action <- SF %>%
  count(disposition1) %>%
  arrange(desc(n))
```

Compare our renamed variables to the original disposition

```
Action <- SF %>%
  count(disposition1, disposition) %>%
  arrange(desc(n))
```

We have codes not listed on the sheet

NA Not recorded 4339

Get rid of the space

```
SF$disposition <- gsub("Not recorded", "Not_Recorded", SF$disposition)
```

Add to the list

```
dispo_lkup <- c(ABA="Abated", ADM="Admonish", ADV="Advised", ARR="Arrest", CAN="Cancel
                CIT="Cited", CRM="Criminal", GOA="Gone", HAN="Handled", NCR="No_Crimina
                NOM="No_Merit", PAS="PlaceSecure", REP="Report", SFD="Medical", UTL="Un
                VAS="Vehicle_Secure", '22'="Cancel", Not_Recorded="NotRecorded")
```

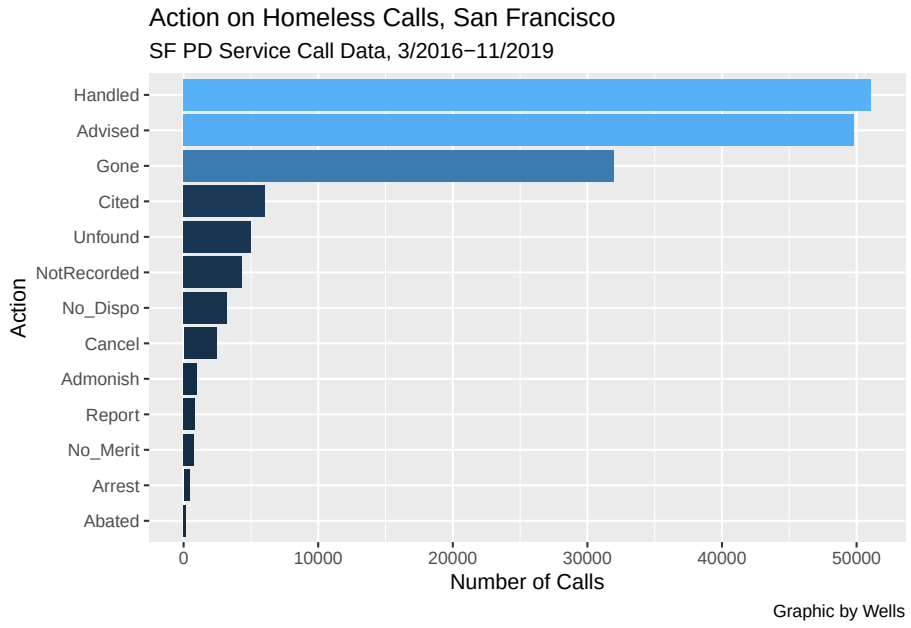
Rerun

# Chapter 20

```
Action <- SF %>%
  count(disposition1) %>%
  arrange(desc(n))
```

Chart Dispositions

```
Action %>%
  filter(n > 100) %>%
  ggplot(aes(x = reorder(disposition1, n), y = n, fill=n)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  coord_flip() +    #this makes it a horizontal bar chart instead of vertical
  labs(title = "Action on Homeless Calls, San Francisco",
       subtitle = "SF PD Service Call Data, 3/2016-11/2019",
       caption = "Graphic by Wells",
       y="Number of Calls",
       x="Action")
```




- **Parse out police codes from narrative: original_crime_type_name** Look at the Types table: some columns have one code, some have two. 919 2879 915 Sleeper 290

Some are separated by a slash 915/919 161

We need to unpack that - **Cleaning Sequence**

```
#convert all text to lowercase
SF$crime1 <- tolower(SF$original_crime_type_name)
#Replace / with a space
SF$crime1 <- gsub("/", " ", SF$crime1)
#Replace '
SF$crime1 <- gsub("'", "", SF$crime1)
#fix space in homeless complaint
SF$crime1 <- gsub("homeless complaint", "homeless_complaint", SF$crime1)
#split data into two columns
SF <- separate(data = SF, col = crime1, into = c("crime2", "crime3", "crime4"), sep = '
```

Look at the categories now

```
Types2 <- SF %>% count(crime2) %>%
  group_by(crime2) %>%
  arrange(desc(n))
```

- **Question** Take the top 10 crime categories from Type2 Relabel them from the numeric radio codes into English

Using the technique earlier in “Build a table to translate the Cop Speak to English”
Relabel the offenses

```
clean <- c(homeless_complaint="homeless_complaint", '915'="homeless_call", '919'="sit_lying", '92
           poss="poss", aggressive="aggressive", '811'="intoxicated")
```

```
SF$crime2 <- as.character(clean[SF$crime2])
```

Look at the categories now

```
Types2 <- SF %>% count(crime2) %>%
  group_by(crime2) %>%
  arrange(desc(n))
```

- **Question:** Make a chart from your cleaned data

Basic chart but with a messed up x axis

```
Types2 %>%
  ggplot(aes(x = crime2, y = n, fill=n)) +
  geom_bar(stat = "identity") +
  coord_flip() +    #this makes it a horizontal bar chart instead of vertical
  labs(title = "Top 10 Homeless Complaints, San Francisco",
       subtitle = "SF PD Service Call Data, 3/2016-11/2019",
       caption = "Graphic by Wells",
       y="Number of Calls",
       x="Complaint")
```

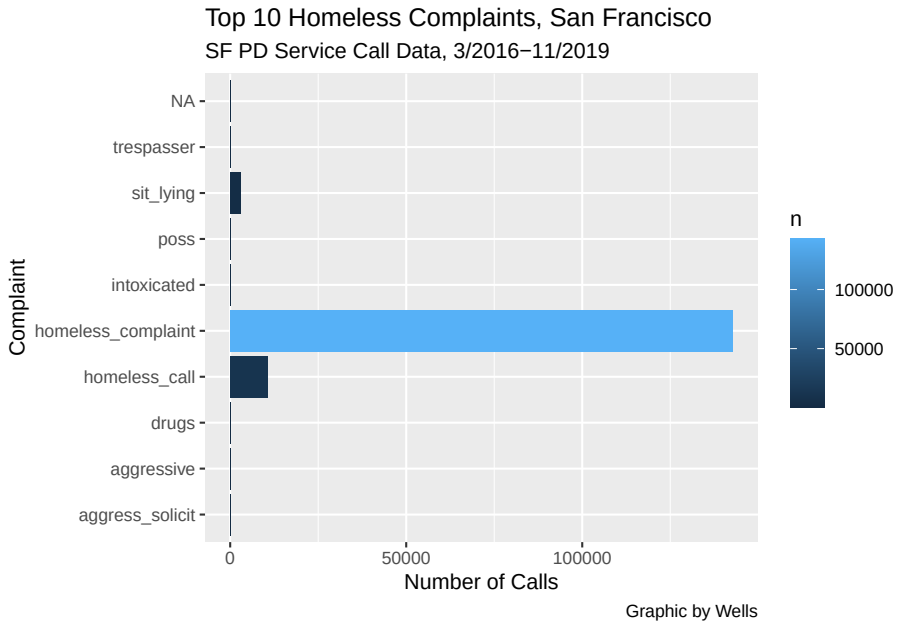
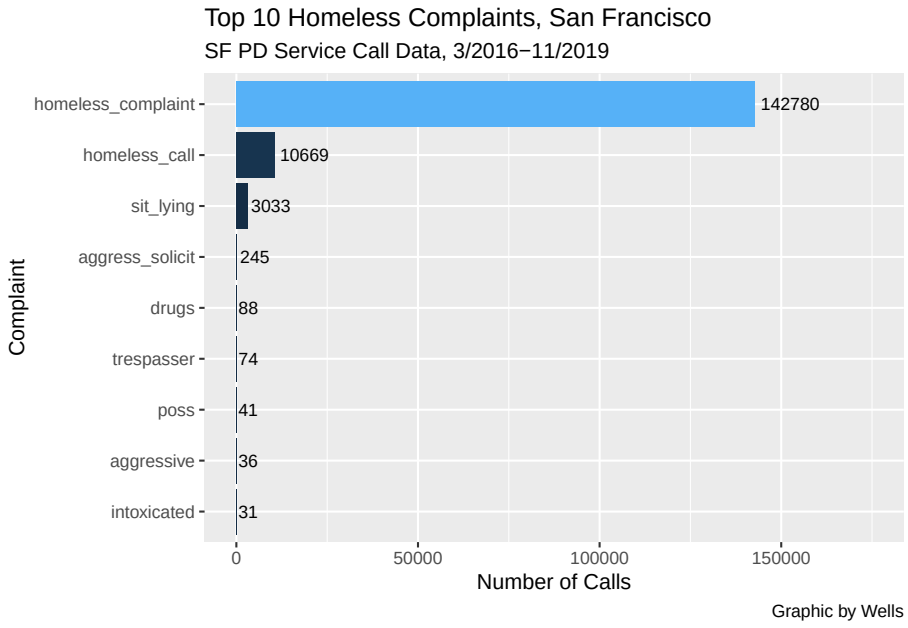



Chart with a fixed x axis scale; No values filtered out; Labels added to bars

```
Types2 %>%
  filter(!is.na(crime2)) %>%
  #filter(crime2!=" ") %>%  - a crude alternative to previous line!
  ggplot(aes(x = reorder(crime2, n), y = n, fill=n)) + #reorder sorts the bars
  geom_bar(stat = "identity", show.legend = FALSE) +
  geom_text(aes(label = n), hjust = -.1, size = 3) +
  scale_y_continuous(limits=c(0, 175000)) + #fixes scientific notation
  coord_flip() +    #this makes it a horizontal bar chart instead of vertical
  labs(title = "Top 10 Homeless Complaints, San Francisco",
       subtitle = "SF PD Service Call Data, 3/2016-11/2019",
       caption = "Graphic by Wells",
       y="Number of Calls",
       x="Complaint")
```


Top 10 Homeless Complaints, San Francisco
SF PD Service Call Data, 3/2016−11/2019
homeless_complaint
homeless_call
sit_lying
aggress_solicit
drugs
trespasser
poss
aggressive
intoxicated
142780
10669
3033
245
88
74
41
36
31
Complaint
0
50000
100000
150000
Number of Calls
Graphic by Wells

# Chapter 21

# Part 4: Using Mutate, Pct Calcs

mutate - Create new column(s) in the data, or change existing column(s).

mutate() adds new variables and preserves existing

Example: mtcars <- as.data.frame(mtcars) View(mtcars)

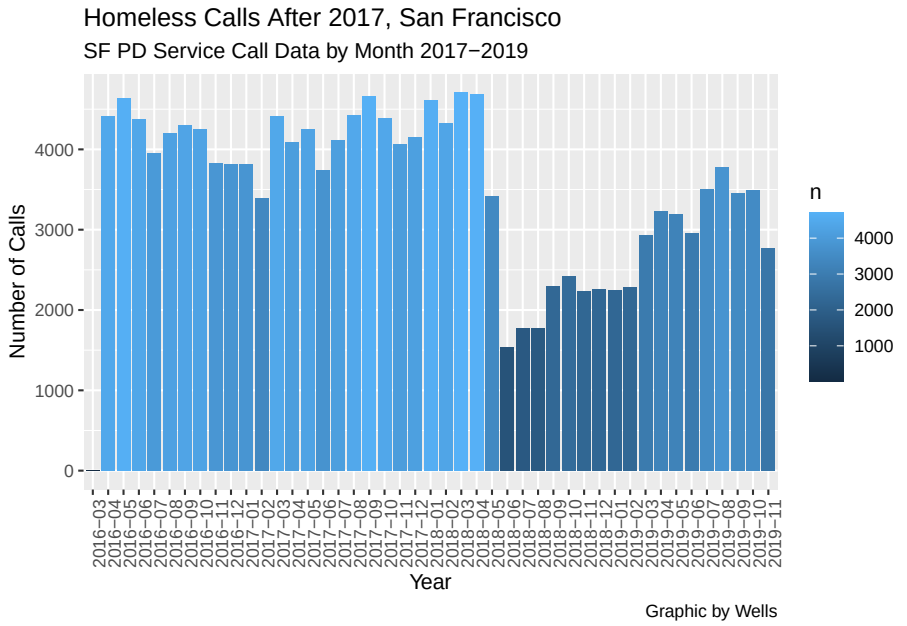
mtcars2 <- mtcars %>% as_tibble() %>% mutate( cyl2 = cyl * 2, cyl4 = cyl2 * 2 )

Process dates using lubidate

```
SF <- SF %>%
  mutate(yearmo = format(call_date, "%Y-%m"))
```

Chart the number of calls by year and month

```
SF %>%
  count(yearmo) %>%
  group_by(yearmo) %>%
  ggplot(aes(x = yearmo, y = n, fill=n)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle=90)) +
  #Changes angle of x axis labels
  #coord_flip() +    #this makes it a horizontal bar chart instead of vertical
  labs(title = "Homeless Calls After 2017, San Francisco",
       subtitle = "SF PD Service Call Data by Month 2017-2019",
       caption = "Graphic by Wells",
       y="Number of Calls",
       x="Year")
```




Percentage change per month

```
PCT_CHG_CALLS <- SF %>%
  select(original_crime_type_name, disposition, address, call_date2, yearmo) %>%
  count(yearmo) %>%
  mutate(difference = (n-lag(n))) %>%
  mutate(pct_change = (difference/abs(lag(n)))*100)
```

- **Use grepl to search and tabulate**

grep and grepl: see ??grep

http://www.endmemo.com/program/R/grepl.php

Cleaning Sequence

```
#convert all text to lowercase
SF$crime1 <- tolower(SF$original_crime_type_name)
```

- **Search for term, rename, put in new column called "cleaned"**

```
x915 <- SF %>%
  filter(grepl ("915", original_crime_type_name)) %>%
  mutate(cleaned = "homeless_complaint")
```

```
x919 <- SF %>%
  filter(grepl ("919", original_crime_type_name)) %>%
  mutate(cleaned = "sitting_lying")
xsleep <- SF %>%
  filter(grepl ("sleep", original_crime_type_name)) %>%
  mutate(cleaned = "sleep")
xaggr <- SF %>%
  filter(grepl ("aggr", original_crime_type_name)) %>%
  mutate(cleaned = "aggressive")
xdrug <- SF %>%
  filter(grepl ("drug", original_crime_type_name)) %>%
  mutate(cleaned = "drug")
xhomeless <- SF %>%
  filter(grepl ("homeless_complaint", crime2)) %>%
  mutate(cleaned = "homeless_complaint")
#Moe, Brooke's Work:
xnoise <- SF %>%
  filter(grepl ("415", original_crime_type_name)) %>%
  mutate(cleaned = "noise")
xposs <- SF %>%
  filter(grepl ("poss", original_crime_type_name)) %>%
  mutate(cleaned = "possession")
xtrespasser <- SF %>%
  filter(grepl ("601", original_crime_type_name)) %>%
  mutate(cleaned = "trespasser")
xsolicit <- SF %>%
  filter(grepl ("920", original_crime_type_name)) %>%
  mutate(cleaned = "solicit")
xinterview <- SF %>%
  filter(grepl ("909", original_crime_type_name)) %>%
  mutate(cleaned = "interview")
xtent <- SF %>%
  filter(grepl ("tent", crime1)) %>%
  mutate(cleaned="tent")
xdog <- SF %>%
  filter(grepl ("dog", crime1)) %>%
  mutate(cleaned="dog")
xchopshop <- SF %>%
  filter(grepl ("chop shop", crime1)) %>%
  mutate(cleaned="chopshop")
xpanhandling <- SF %>%
  filter(grepl ("panhandling", crime1)) %>%
  mutate(cleaned="panhandling")
xmusic <- SF %>%
  filter(grepl ("music", crime1)) %>%
```

```
  mutate(cleaned="music")
```

Create new dataframe using rbind

```
new_total <- rbind(xhomeless, x915, x919, xaggr, xdrug, xsleep, xnoise, xposs, xtrespas
                   xchopshop, xpanhandling, xmusic)
```

Count it up!

```
Total_Calls_Master <- new_total %>%
  count(cleaned) %>%
  arrange(desc(n))
#rename columns
colnames(Total_Calls_Master)[1:2] <- c("Complaints", "Number")
#export
write_csv(Total_Calls_Master, "Total_Calls_Master.csv")
```

Make into html table

```
#install.packages("kableExtra")
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
#This makes html tables called "kables"
Total_Calls_Master %>%
  kable() %>%
  kable_styling("striped")
```

Export from Viewer as .png

- **Task: Tabulate complaints by day of the week**

https://github.com/profrobwells/Data-Analysis-Class-Jour-405v-5003/blob/master/Readings/dealing-with-dates.pdf

| Complaints | Number |
|---|---|
| homeless_complaint | 153895 |
| sitting_lying | 3282 |
| solicit | 262 |
| trespasser | 100 |
| sleep | 77 |
| interview | 75 |
| drug | 52 |
| noise | 27 |
| aggressive | 18 |
| dog | 12 |
| tent | 9 |
| music | 8 |
| chopshop | 7 |
| panhandling | 5 |
| possession | 5 |

```
SF <- SF %>%
  mutate(weekday = wday(call_date, label=TRUE, abbr=FALSE))
```

Build a summary table with the days of the week with the greatest number of calls. Create a graphic. Then build a table to see if the complaints vary by day

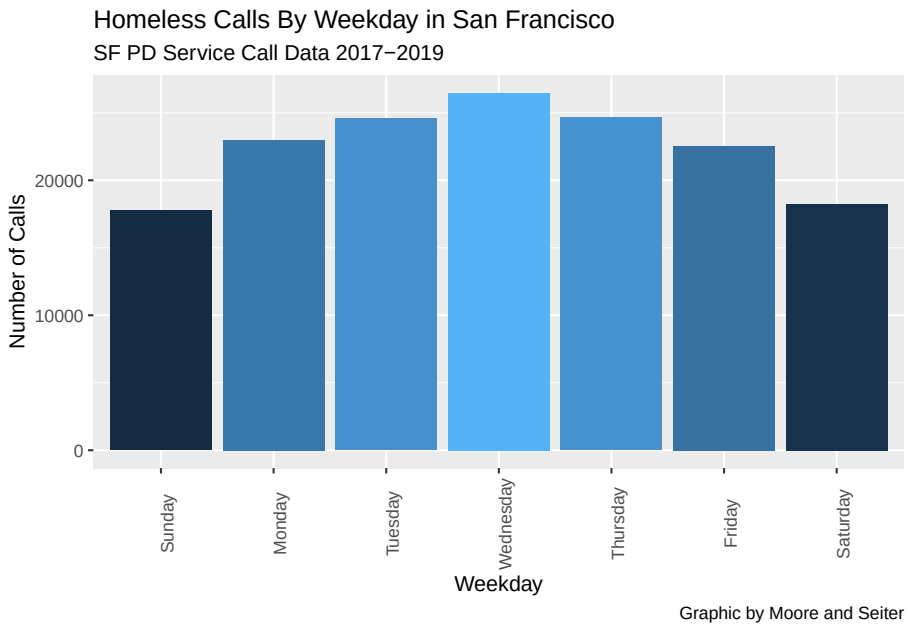
Below from Matthew Moore, Katy Seiter, Wells edited

```
SF <- SF %>%
  mutate(weekday = wday(call_date, label=TRUE, abbr=FALSE))
Weekday_Count <- SF %>%
  select(weekday, crime_id) %>%
  count(weekday) %>%
  arrange(desc(n))
```

Graphic of calls by weekdays

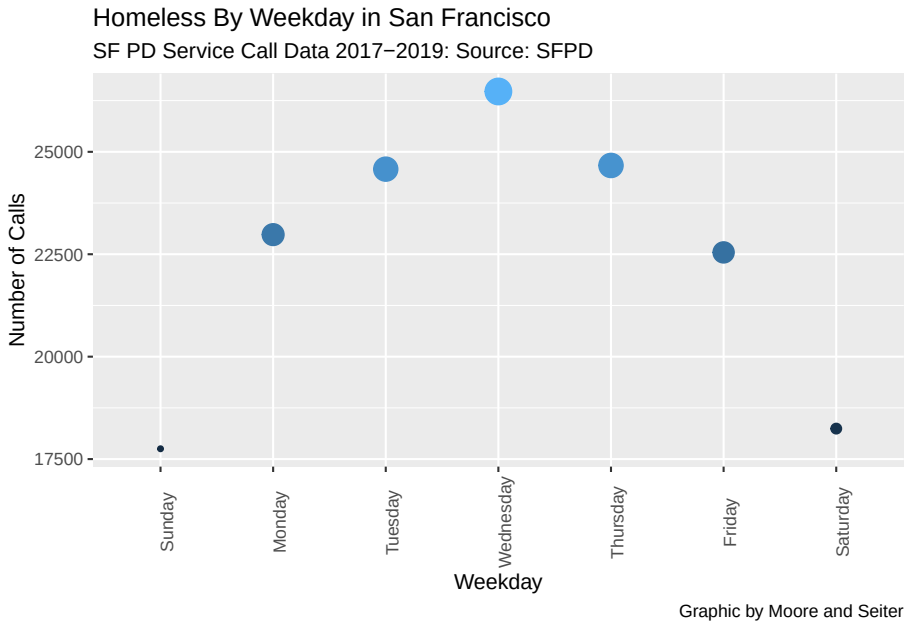
```
Weekday_Count %>%
  ggplot(aes(x = weekday, y = n, fill=n)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  theme(axis.text.x = element_text(angle=90)) +
  #Changes angle of x axis labels
  #coord_flip() +    #this makes it a horizontal bar chart instead of vertical
  labs(title = "Homeless Calls By Weekday in San Francisco",
       subtitle = "SF PD Service Call Data 2017-2019",
       caption = "Graphic by Moore and Seiter",
```

```
       y="Number of Calls",
       x="Weekday")
```




Create a Bubble graphic

```
ggplot(data = Weekday_Count) +
  geom_point(mapping = aes(x = weekday, y = n, size = n, color = n), show.legend = FALS
  theme(axis.text.x = element_text(angle=90)) +
  labs(title = "Homeless By Weekday in San Francisco",
       subtitle = "SF PD Service Call Data 2017-2019: Source: SFPD",
       caption = "Graphic by Moore and Seiter",
       y="Number of Calls",
       x="Weekday")
```




Improved bubble chart

```
# ggplot(Weekday_Count, aes(x = weekday, y = n)) +
#   xlab("Weekday") +
#   ylab("Number of Calls") +
#   theme_minimal(base_size = 12, base_family = "Georgia") +
#   geom_point(aes(size = n, color = n), alpha = 0.7, show.legend = FALSE) +
#   scale_size_area(guide = FALSE, max_size = 15) +
#   labs(title = "Homeless By Weekday in San Francisco",
#        subtitle = "SF PD Service Call Data 2017-2019: Source: SFPD",
#        caption = "Graphic by Moore and Seiter")
```

- **Task #3: Calls vs Dispositions**

What calls resulted in arrests? What calls resulted in citations?

```
Action2 <- SF %>%
  select(crime_id, original_crime_type_name, disposition)
```

We need to pair the crime type and disposition and then count them

From Michael Adkison:

```
callsarrest <- Action2 %>%
  filter(grepl("ARR", disposition)) %>%
  mutate(cleaned = "Arrest")
```

To quickly format into percents, load formattable

```
#install.packages("formattable")
library(formattable)
callsarrest2 <- callsarrest %>%
  arrange(original_crime_type_name, disposition) %>%
  count(original_crime_type_name) %>%
#mutate(PctTotal = (n/441)) %>%
  arrange(desc(n))
colnames(callsarrest2)[1:2] <- c("Complaints", "Arrests")
```

Build a table to translate the Cop Speak to English:

```
clean <- c('Homeless Complaint'="homeless_complaint", homeless_complaint="homeless_com
           '919'="Sit_lying", '920'="Aggress_solicit", '915s'="homeless_complaint", '9
           drugs="drugs", '601'="trespasser", poss="poss", aggressive="aggressive", '8
           'Drugs / 915'="Drugs", 'Drugs/915'="Drugs")
```

This scans "disposition", finds ABA and replaces with Abated, finds ARR, replaces with Arrest, etc callsarrest2$Complaints <- as.character(clean[callsarrest2$Complaints])

```
callsarrest3 <- callsarrest2 %>%
  select(Complaints, Arrests) %>%
  group_by(Complaints) %>%
  summarise(total = sum(Arrests)) %>%
  mutate(PctTotal = (total/441)) %>%
  arrange(desc(total))
colnames(callsarrest3)[2] <- "Arrests"
callsarrest3$PctTotal <- percent(callsarrest3$PctTotal)
#This makes kables
callsarrest3 %>%
  kable() %>%
  kable_styling("striped")
```

| Complaints | Arrests | PctTotal |
|---|---|---|
| Homeless Complaint | 412 | 93.42% |
| 915 | 19 | 4.31% |
| 919 | 4 | 0.91% |
| 920 | 2 | 0.45% |
| 601 / 915 | 1 | 0.23% |
| 915 / 909 | 1 | 0.23% |
| Drugs / 915 | 1 | 0.23% |
| Drugs/915 | 1 | 0.23% |

# Chapter 22

# Part 5: Trends over time

- **Question**: What were the common days for arrests?

```
SF %>%
  select(weekday, crime_id, disposition) %>%
  filter(grepl("ARR", disposition)) %>%
  count(weekday)
```

```
##     weekday  n
## 1    Sunday 47
## 2    Monday 63
## 3   Tuesday 79
## 4 Wednesday 77
## 5  Thursday 74
## 6    Friday 56
## 7  Saturday 45
```

Make bubble chart

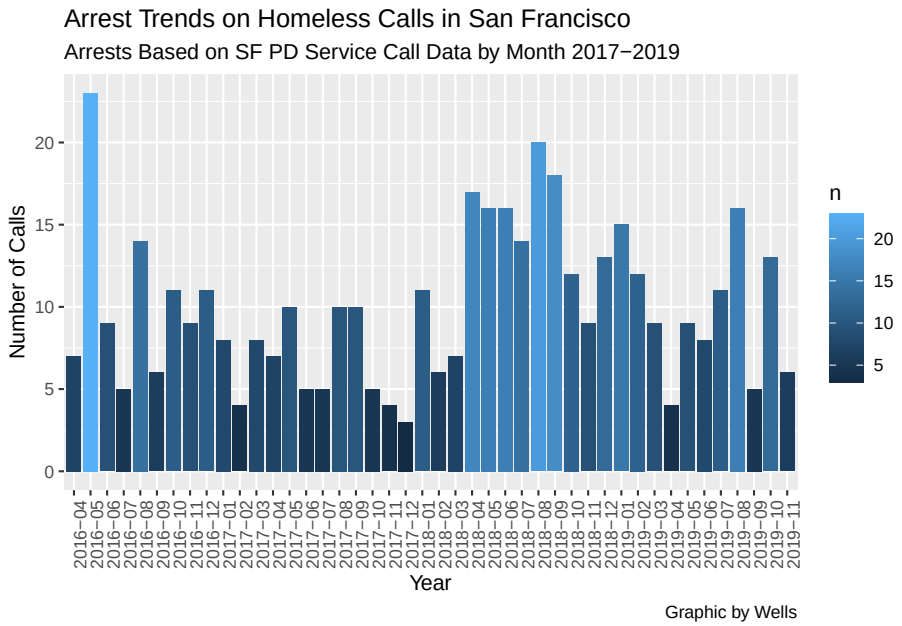
```
# SF %>%
#   select(weekday, crime_id, disposition) %>%
#   filter(grepl("ARR", disposition)) %>%
#   count(weekday) %>%
#   ggplot(aes(x = weekday, y = n)) +
#   xlab("Weekday") +
#   ylab("Arrests") +
#   theme_minimal(base_size = 12, base_family = "Georgia") +
#   geom_point(aes(size = n, color = n), alpha = 0.7, show.legend = FALSE) +
#   scale_size_area(guide = FALSE, max_size = 15) +
```

```
#   labs(title = "Homeless Arrests By Weekday in San Francisco",
#        subtitle = "SF PD Service Call Data 2017-2019: Source: SFPD",
#        caption = "Graphic by Wells")
```

- **Question**: What is the trend for arrests over the time period?

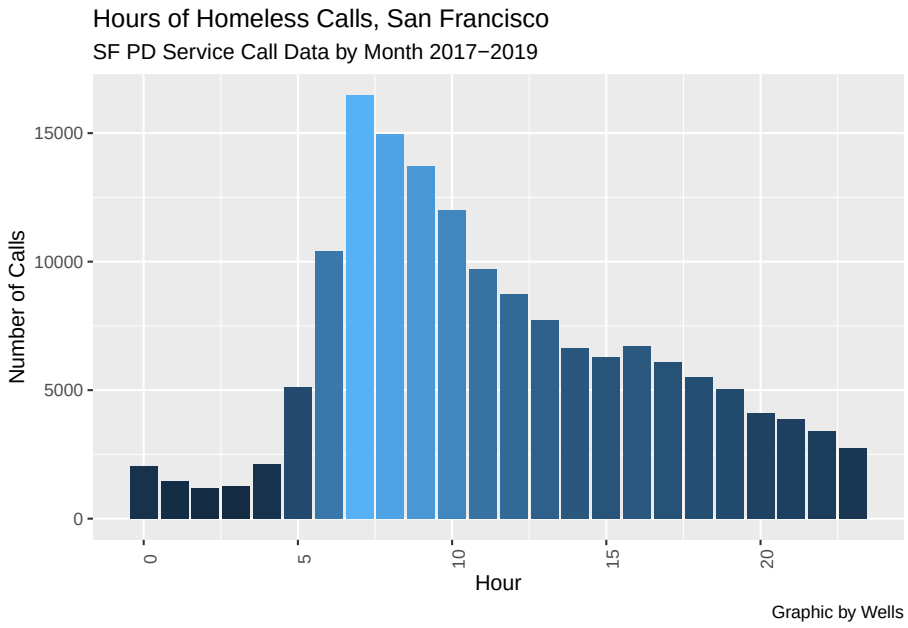
```
SF %>%
  filter(grepl("ARR", disposition)) %>%
  count(yearmo) %>%
  group_by(yearmo) %>%
  ggplot(aes(x = yearmo, y = n, fill=n)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  geom_smooth(method = lm, se=FALSE, color = "red") +
  theme(axis.text.x = element_text(angle=90)) +
  #Changes angle of x axis labels
  #coord_flip() +    #this makes it a horizontal bar chart instead of vertical
  labs(title = "Arrest Trends on Homeless Calls in San Francisco",
       subtitle = "Arrests Based on SF PD Service Call Data by Month 2017-2019",
       caption = "Graphic by Wells",
       y="Number of Calls",
       x="Year")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



- **Question**: What are the hours most likely for complaints?

```
#format to hours
SF$hour <- hour(SF$call_date_time)
SF %>%
  count(hour) %>%
  group_by(hour) %>%
  ggplot(aes(x = hour, y = n, fill=n)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  theme(axis.text.x = element_text(angle=90)) +
  #Changes angle of x axis labels
  #coord_flip() +    #this makes it a horizontal bar chart instead of vertical
  labs(title = "Hours of Homeless Calls, San Francisco",
       subtitle = "SF PD Service Call Data by Month 2017-2019",
       caption = "Graphic by Wells",
       y="Number of Calls",
       x="Hour")
```




- **Question**: Examine some of the charting options on this tutorial and adapt them to this data using any chart you want # https://paldhous.github.io/wcsj/2017/

# Chapter 23

# –30–

# Bibliography