PM 592
Regression Analysis for Public Health Data Science

Week 1

Introductory Topics

Introductory Topics

1. Course Logistics
2. Goals of Data Analysis
3. Study Types
4. Variable Types
5. An Example: The Children's Health Study
6. Choosing A Statistical Package
7. Exploratory Data Analysis (EDA)

Lecture Objectives

- ➢ Provide a motivation for learning regression analysis.
- ➢ Review types of studies and data formats.
- ➢ Learn to manage and analyze data.
- ➢ Become acquainted with R.
- ➢ Learn to examine and interpret output from R.

Syllabus Overview

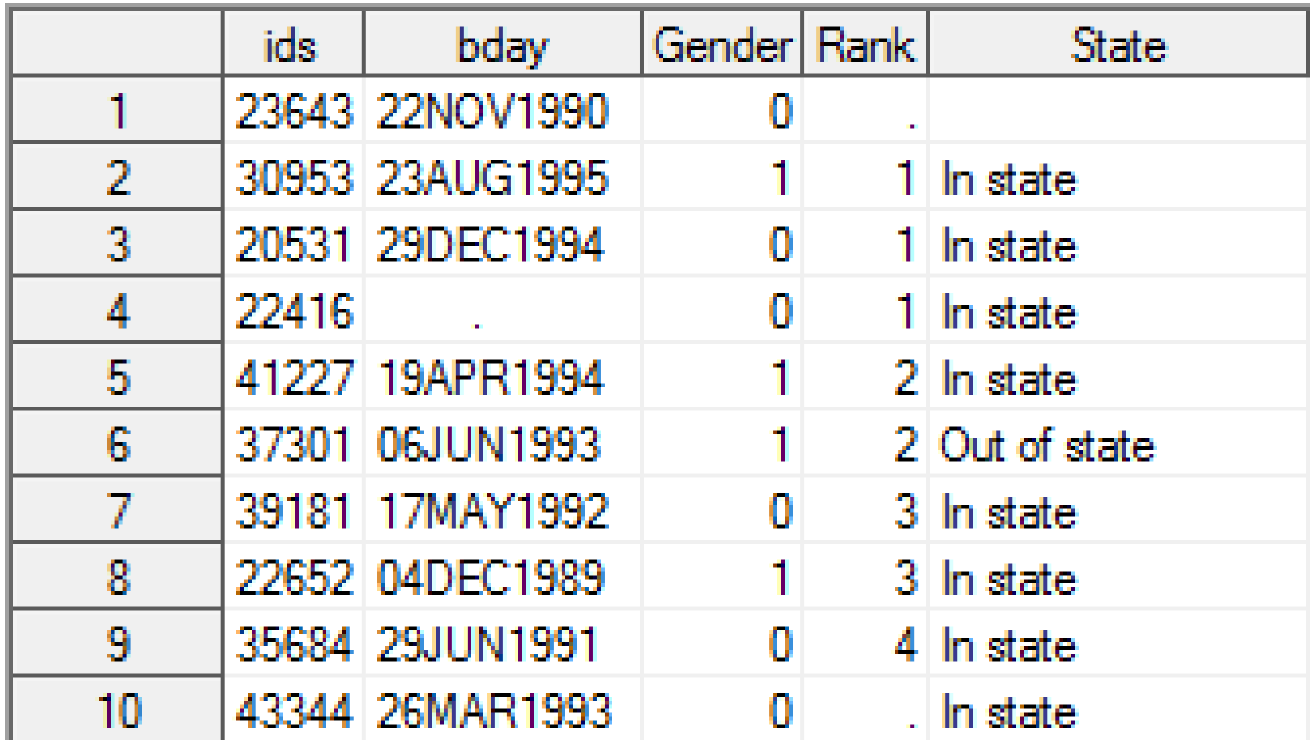
2. Basic Goal of Data Analysis

The goal of data analysis is to **summarize the signal in the noise.**

In other words, we want to **reduce myriad information** into concise descriptions that will make sense to us and our audience.

2. Basic Goal of Data Analysis

Suppose a team conducts a study and gets the following data on participants:



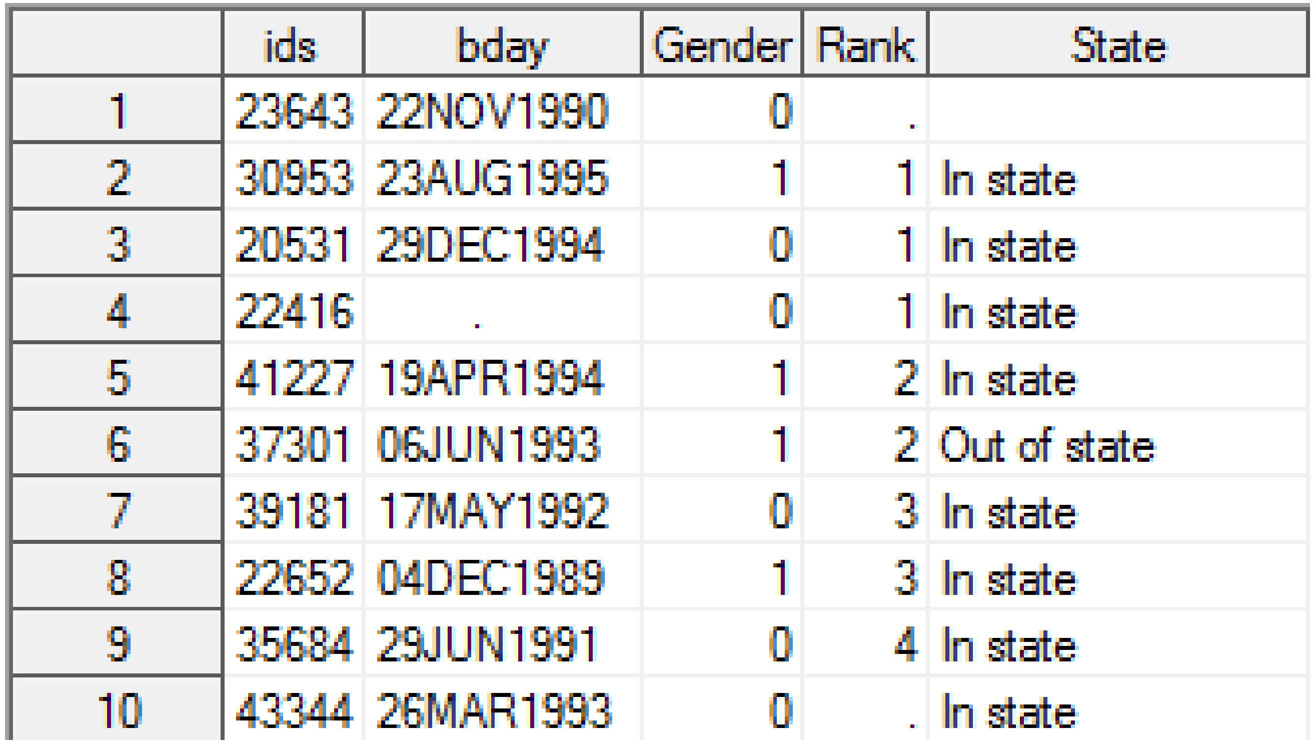
	ids	bday	Gender	Rank	State
1	23643	22NOV1990	0	.	
2	30953	23AUG1995	1	1	In state
3	20531	29DEC1994	0	1	In state
4	22416	.	0	1	In state
5	41227	19APR1994	1	2	In state
6	37301	06JUN1993	1	2	Out of state
7	39181	17MAY1992	0	3	In state
8	22652	04DEC1989	1	3	In state
9	35684	29JUN1991	0	4	In state
10	43344	26MAR1993	0	.	In state

2. Basic Goal of Data Analysis

Is pasting the data set into the publication informative?

Is there another way we can present this data?

How can we reduce this data set down into a way that it can be synthesized easily?



	ids	bday	Gender	Rank	State
1	23643	22NOV1990	0	.	
2	30953	23AUG1995	1	1	In state
3	20531	29DEC1994	0	1	In state
4	22416	.	0	1	In state
5	41227	19APR1994	1	2	In state
6	37301	06JUN1993	1	2	Out of state
7	39181	17MAY1992	0	3	In state
8	22652	04DEC1989	1	3	In state
9	35684	29JUN1991	0	4	In state
10	43344	26MAR1993	0	.	In state

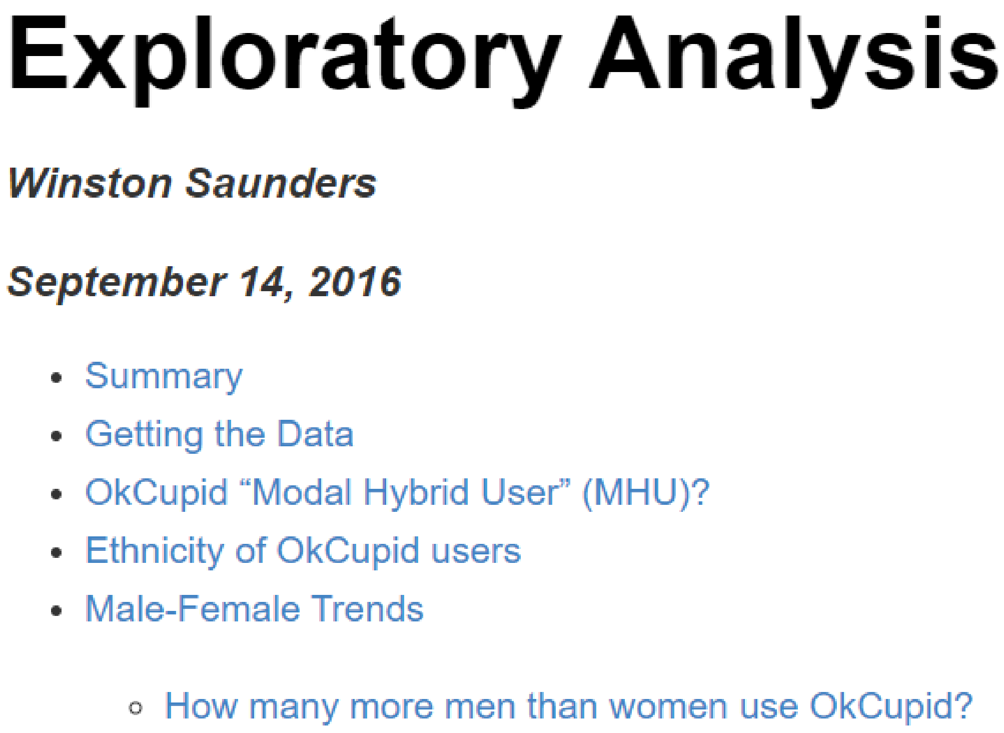
❑ Name 4 different pieces of “easily-digestible” information we can get from this table.

2. Basic Goal of Data Analysis

Example

Okcupid released de-identified data from 2012 for several of its users in San Francisco. Some people have used this data as an analysis exercise. Consider the following website:

https://rstudio-pubs-static.s3.amazonaws.com/209370_b62220c849b946088b463fdbec935848.html



Exploratory Analysis of OkCupid dataset

Winston Saunders

September 14, 2016

- Summary
- Getting the Data
- OkCupid “Modal Hybrid User” (MHU)?
- Ethnicity of OkCupid users
- Male-Female Trends
    - How many more men than women use OkCupid?

2. Basic Goal of Data Analysis

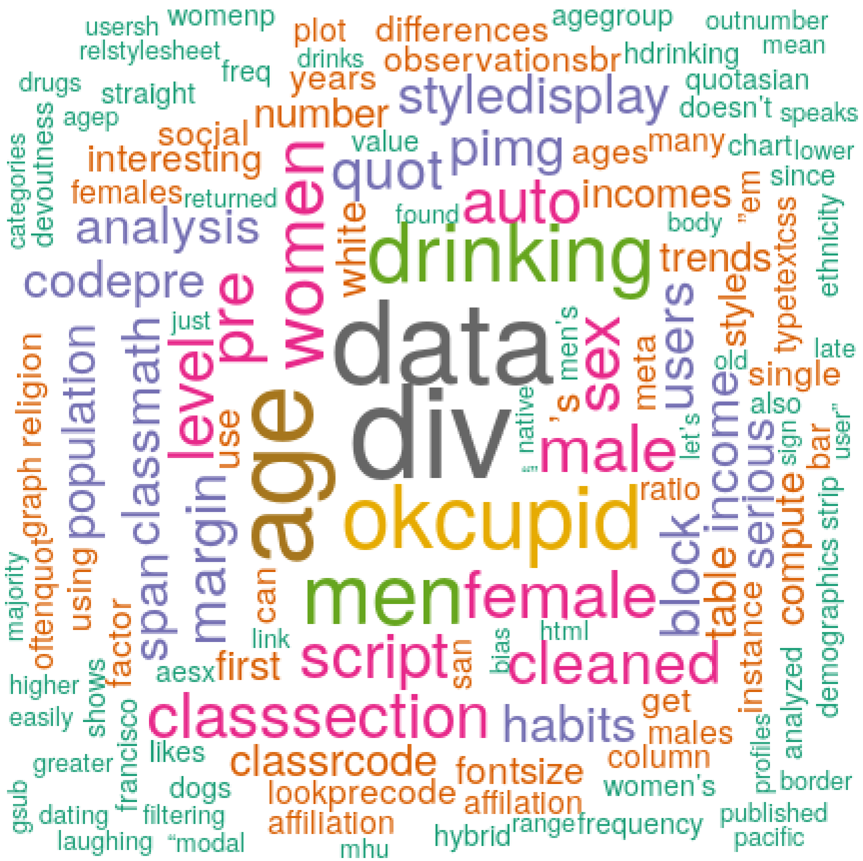
Example

Let’s put ourselves in a “summary” mindset:

How could we summarize the information in this particular webpage for someone else?

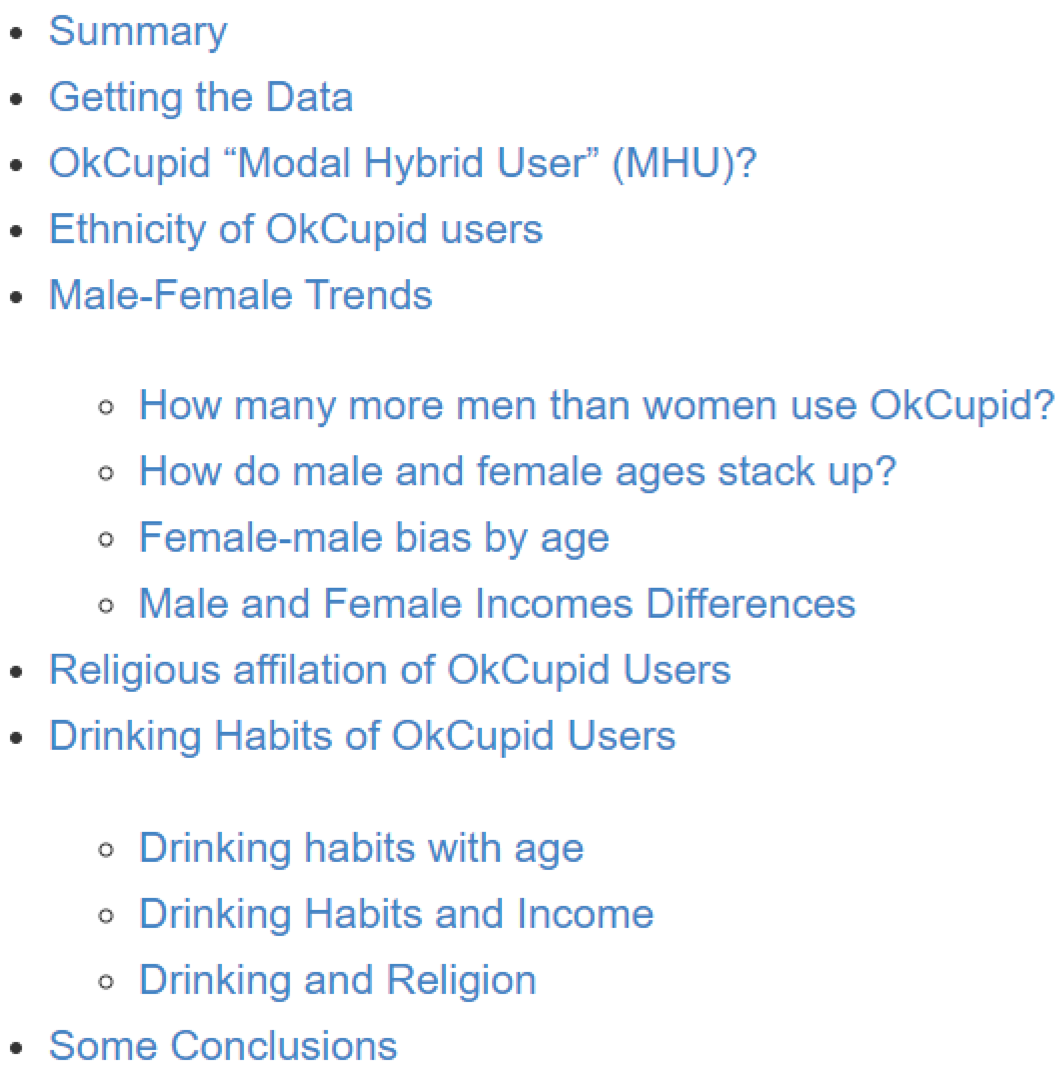
2. Basic Goal of Data Analysis

Option 1. We could present a word cloud of the words that frequently appear on the webpage.



Note: see this article for some benefits and drawbacks of using word clouds
https://www.keatext.ai/en/blog/artificial-intelligence/3-strengths-and-3-weaknesses-of-word-clouds/

Option 2. We could present an outline of the content.



- Summary
- Getting the Data
- OkCupid “Modal Hybrid User” (MHU)?
- Ethnicity of OkCupid users
- Male-Female Trends
    - How many more men than women use OkCupid?
    - How do male and female ages stack up?
    - Female-male bias by age
    - Male and Female Incomes Differences
- Religious affilation of OkCupid Users
- Drinking Habits of OkCupid Users
    - Drinking habits with age
    - Drinking Habits and Income
    - Drinking and Religion
- Some Conclusions

2. Basic Goal of Data Analysis

What are the similarities and differences of these two methods?

Which is more effective?

These are both ways of summarizing the information on the webpage. However, the data analyst clearly must think of the best way to present this information.

2. Basic Goal of Data Analysis

Some common goals of data analysis include:

A. Describe/explore data

B. Test hypotheses

C. Build models (regression)

D. Estimate parameters

E. Summarize data by interpreting models

Some common goals of data analysis include:

A. Describe/explore data

- Graphically display data
- Condense large datasets down into summary statistics (means, medians, etc.)
- Means, variances, frequencies, etc.

B. Test hypotheses

C. Build models (regression)

D. Estimate parameters

E. Summarize data by interpreting models

2. Basic Goal of Data Analysis

Some common goals of data analysis include:

A. Describe/explore data

B. Test hypotheses

- Single well-defined *a priori* hypothesis
- Multiple hypotheses
- Translate a research question / hypothesis into a statistical question

C. Build models (regression)

D. Estimate parameters

E. Summarize data by interpreting models

Some common goals of data analysis include:

A. Describe/explore data

B. Test hypotheses

C. Build models (regression)

- Choose the correct model for the given data (linear regression, logistic regression)
- Assess model fit and evaluate model assumptions
- Add model complexity
- Determine whether a model is for "estimation" or "prediction"

D. Estimate parameters

E. Summarize data by interpreting models

Some common goals of data analysis include:

A. Describe/explore data

B. Test hypotheses

C. Build models (regression)

D. Estimate parameters

- Point estimates
- Uncertainty estimates: standard error, confidence intervals, etc.

E. Summarize data by interpreting models

2. Basic Goal of Data Analysis

Some common goals of data analysis include:

A. Describe/explore data

B. Test hypotheses

C. Build models (regression)

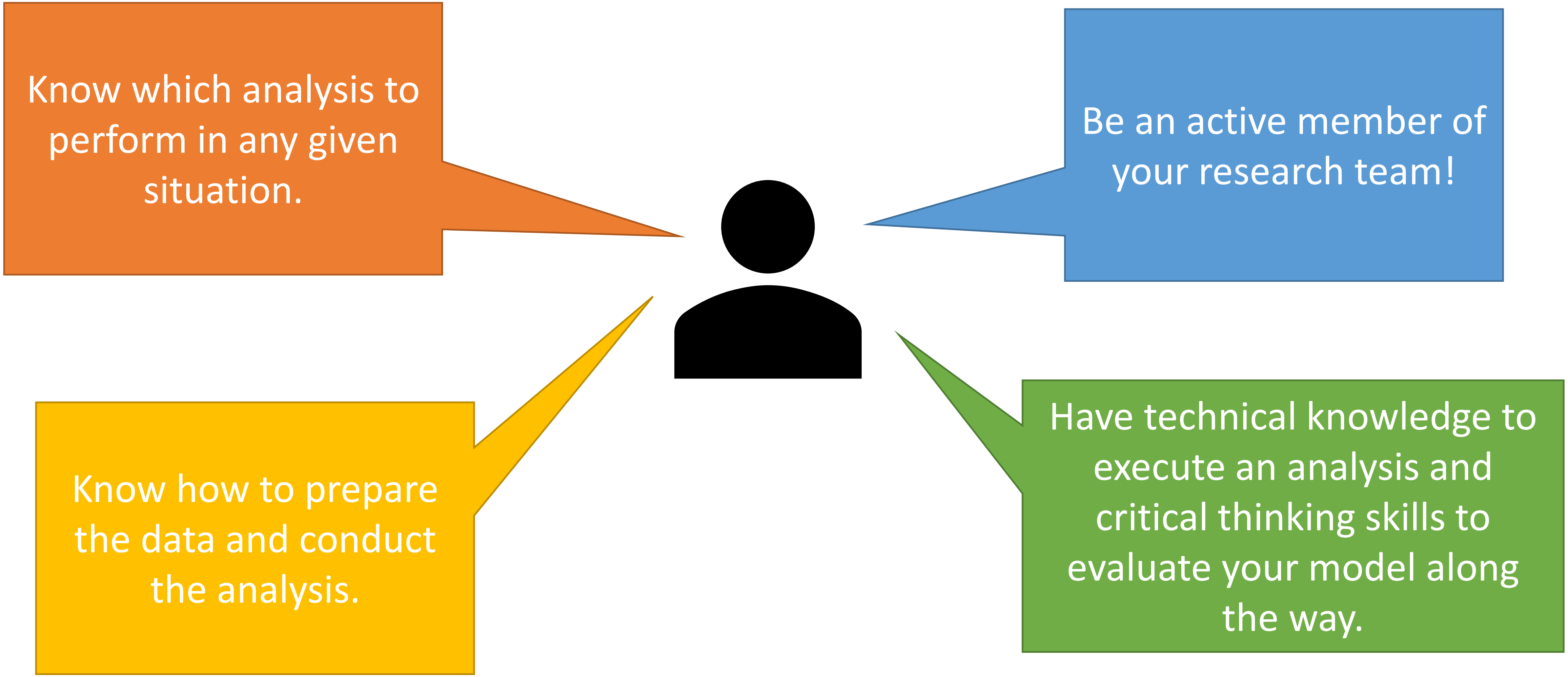
D. Estimate parameters

E. Summarize data by interpreting models

- Describe your methods and results
- Display results in tables and figures
- Discuss strengths and limitations of your analysis and the study

2. Basic Goal of Data Analysis

What skills are needed for a good data analyst?



Know which analysis to perform in any given situation.

Be an active member of your research team!

Know how to prepare the data and conduct the analysis.

Have technical knowledge to execute an analysis and critical thinking skills to evaluate your model along the way.

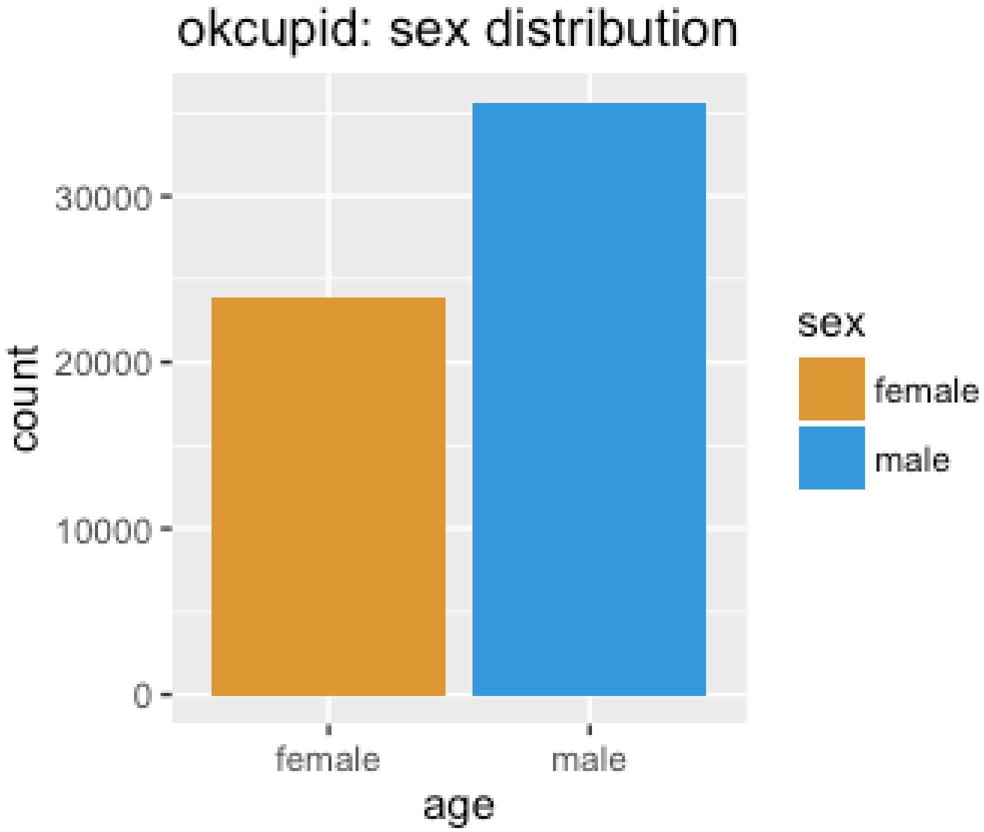
2. Basic Goal of Data Analysis

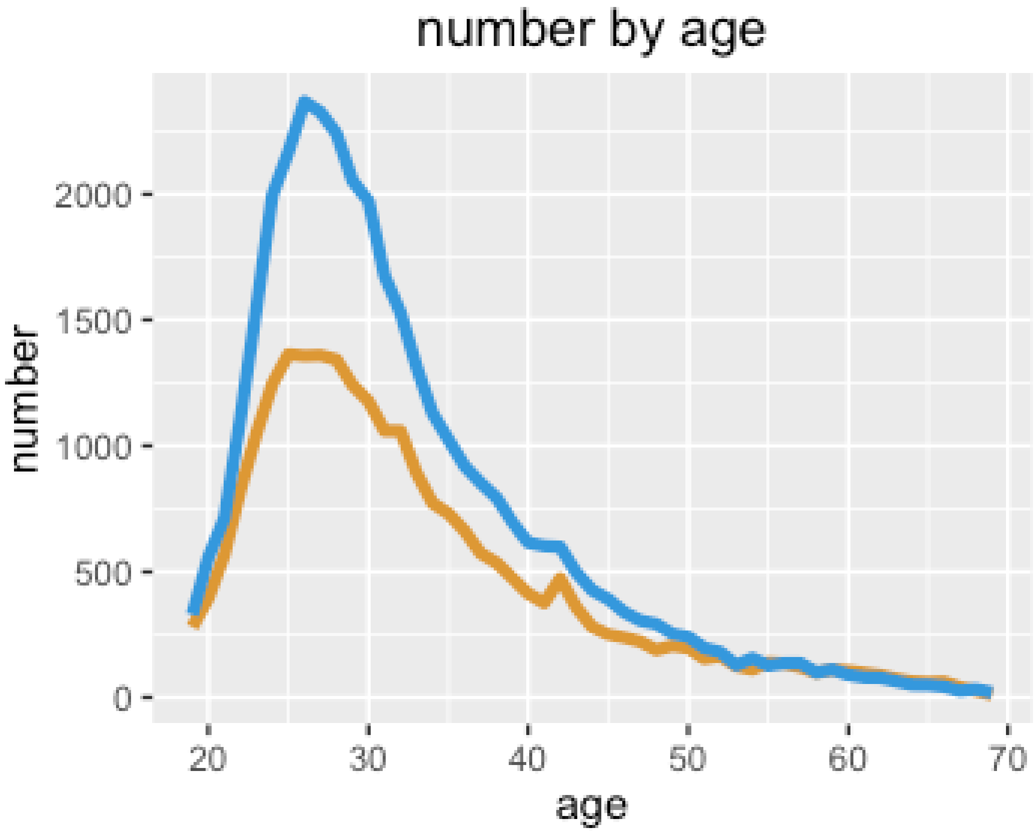
Which would make for a better data analyst?

Summarizes data blindly and presents all the output from a particular analysis.

Has specific questions in mind, tries to create a "story" with the analysis.

What is the story being told here?





Recap

- There are many possible goals of data analysis
- Knowing the purpose of the analysis can help you produce more meaningful results

Test Yourself

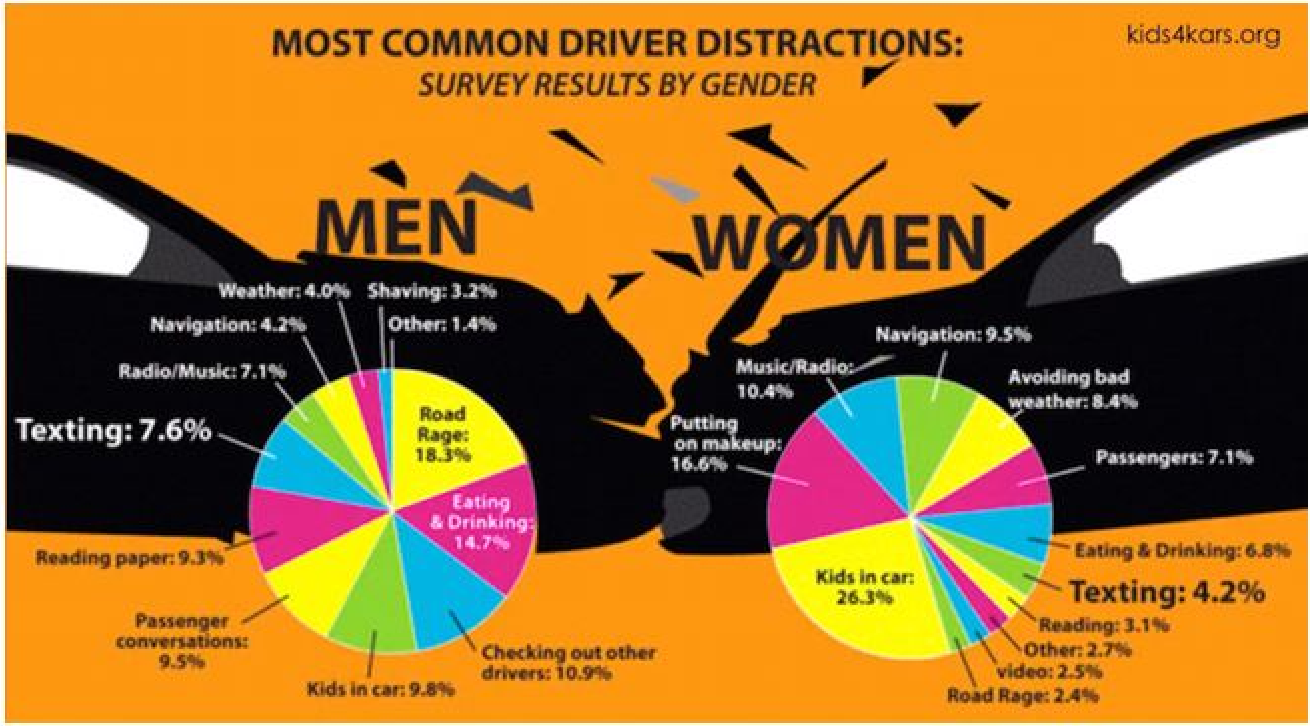
Which of the following is NOT part of the story being told by this graphic?

a. Men are distracted while driving just as much as females are distracted while driving.

b. Women are typically more distracted by kids in the car, compared to men.

c. Men are typically more distracted by road rage, compared to women.

d. Men are typically more distracted by checking out other drivers, compared to women.



2. Basic Goal of Data Analysis

Test Yourself

Which of the following is NOT part of the story being told by this graphic?

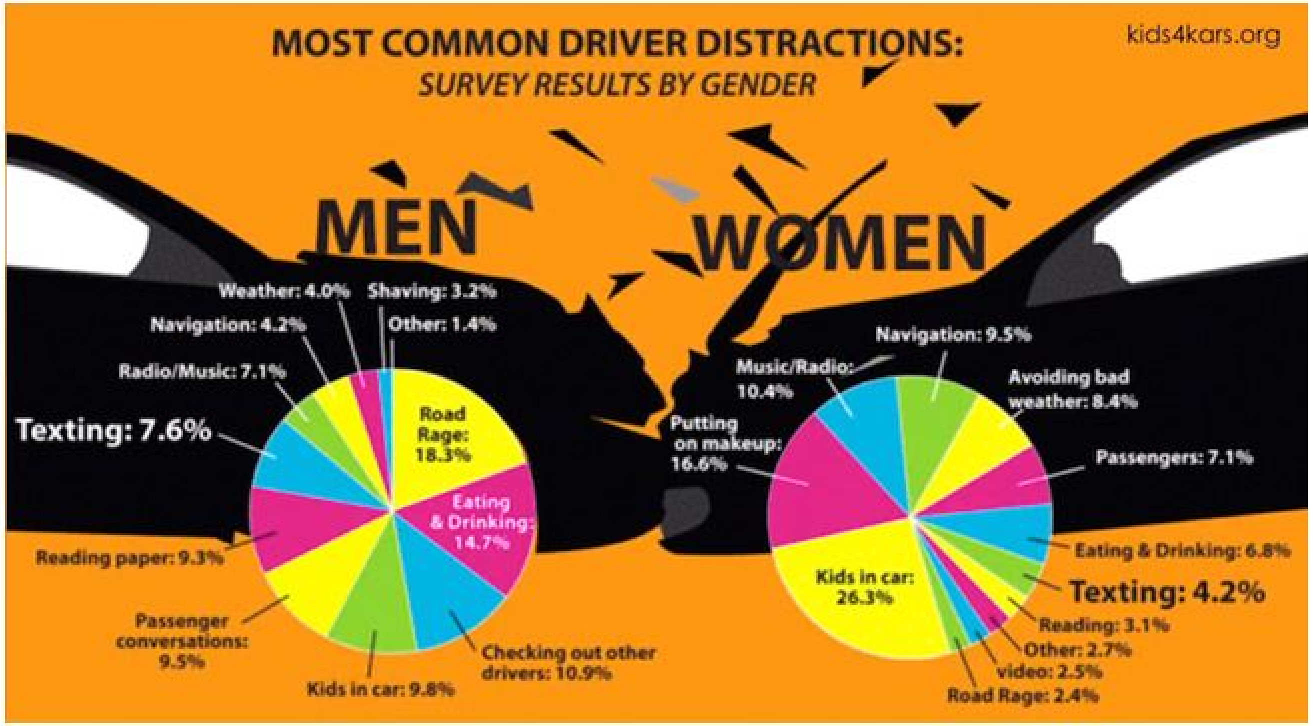
a. Men are distracted while driving just as much as females are distracted while driving.

b. Women are typically more distracted by kids in the car, compared to men.

c. Men are typically more distracted by road rage, compared to women.

d. Men are typically more distracted by checking out other drivers, compared to women.

The pie charts show us the *percentage* of distractions within each gender, but there is no information comparing the frequency of distractions between males and females.



Kids in car causes 26.3% of distractions for females, compared to 9.8% for males. Road rage causes 18.3% of distractions for males, compared to 2.4% for females. Checking out other drivers is not listed as a distraction for females, but causes 10.9% of distractions for males.

3. Study Types

Data doesn’t live alone in a vacuum. The **type of study** that was conducted is important as it affects:

- How the data should be analyzed
- The conclusions that can be drawn from the data analysis
- The generalizability of the findings

In general, the more rigor goes into determining the treatment groups, the stronger the study type.

Example

How could we test to see whether in-person learning is more effective than online learning?

3. Study Types

Let’s discuss the main types of studies:

A. Experimental

B. Quasi-experimental

C. Observational

3. Study Types

A. Experimental

- Subjects are randomly assigned to exposure categories
- A **clinical trial** is a special case of an experimental study, where subjects are randomly assigned to treatment group and followed forward to determine outcome
- **A/B testing** is another special case of an experimental study. Typically, users of a website (or some other user interface) are randomly presented with one of two versions of the site (A vs. B). The investigator compares some metric (e.g., did the user click through to enroll in health insurance?) between the two versions.

B. Quasi-experimental

C. Observational

3. Study Types

A. Experimental

B. Quasi-experimental

- Subjects are identified by exposure levels, but without randomization
- In a **cohort study**, subjects are identified by exposure status and then followed forward in time to determine outcome
- In a **case-control** study, subjects are identified by outcome status (e.g., presence of disease, death, etc.) and then exposure status is determined retrospectively

C. Observational

3. Study Types

A. Experimental

B. Quasi-experimental

C. Observational

- There is no randomization or assignment of subject by exposure levels
- In a **cross-sectional** study, individuals’ exposure and outcome status are assessed simultaneously (i.e., no randomization, and exposure/outcome status is not used as a selection criteria)

Recap

- There are generally 3 different types of studies, and the strength of your findings depends on the study type

➢Given a scenario of a study, be able to identify the study type

Test Yourself

A study was designed to examine the effectiveness of a new medication. Fifty volunteer subjects from the area around USC were randomly assigned to either the vaccine or a placebo. What type of study is this?

a) Experimental

b) Quasi-Experimental: Case-Control

c) Quasi-Experimental: Cohort

d) Observational: Cross-Sectional

Test Yourself

A study was designed to examine the effectiveness of a new medication. Fifty volunteer subjects from the area around USC were randomly assigned to either the vaccine or a placebo. What type of study is this?

a) Experimental

b) Quasi-Experimental: Case-Control

c) Quasi-Experimental: Cohort

d) Observational: Cross-Sectional

This is an experimental study because participants were randomly assigned to treatment condition (exposure).

Test Yourself

500 residents of Los Angeles were surveyed about their attitudes toward the monkeypox vaccine and were asked their age and gender. Researchers collected these surveys between July and September 2022. What type of study is this?

a) Experimental

b) Quasi-Experimental: Case-Control

c) Quasi-Experimental: Cohort

d) Observational: Cross-Sectional

Test Yourself

500 residents of Los Angeles were surveyed about their attitudes toward the monkeypox vaccine and were asked their age and gender. Researchers collected these surveys between July and September 2022. What type of study is this?

a) Experimental

b) Quasi-Experimental: Case-Control

c) Quasi-Experimental: Cohort

d) Observational: Cross-Sectional

This is a cross-sectional study because all information was assessed simultaneously.

Test Yourself

40 schools were randomized into either 1) a mental health intervention or 2) standard practice. Outcomes were assessed at 1 and 2 years later. What type of study is this?

a) Experimental

b) Quasi-Experimental: Case-Control

c) Quasi-Experimental: Cohort

d) Observational: Cross-Sectional

Test Yourself

40 schools were randomized into either 1) a mental health intervention or 2) standard practice. Outcomes were assessed at 1 and 2 years later. What type of study is this?

a) Experimental

b) Quasi-Experimental: Case-Control

c) Quasi-Experimental: Cohort

d) Observational: Cross-Sectional

This is an experimental study because participants were randomly assigned to treatment condition (exposure).

Test Yourself

200 individuals who drink coffee daily, and 200 individuals who do not drink coffee daily, were followed for 5 years to assess the development of depression and anxiety disorder. What type of study is this?

a) Experimental

b) Quasi-Experimental: Case-Control

c) Quasi-Experimental: Cohort

d) Observational: Cross-Sectional

Test Yourself

200 individuals who drink coffee daily, and 200 individuals who do not drink coffee daily, were followed for 5 years to assess the development of depression and anxiety disorder. What type of study is this?

a) Experimental

b) Quasi-Experimental: Case-Control

c) Quasi-Experimental: Cohort

d) Observational: Cross-Sectional

This is a cohort study because exposure status was ascertained first, and then individuals were followed to assess outcome.

4. Types of Variables

The type of analysis that can be performed depends on the type of variables you wish to include.

A. Continuous

B. Discrete

4. Types of Variables

A. Continuous

- Any variable such that between any two potentially observable values, there exists another potentially observable value.
- Accuracy of the observed values is limited by the measurement device
- *E.g., age, height, cholesterol, lung function*

B. Discrete

4. Types of Variables

A. Continuous

B. Discrete

- Any variable that is not continuous
- **Ordered numeric** variables have numerical, non-continuous values (e.g., number of deaths, number of school absences)
- **Ordered categories** are non-numeric, but have ordered values (e.g., socioeconomic status as high, middle, or low)
- **Unordered categories** have no inherent order (e.g., race, study site)
- **Binary** variables have just two categories (e.g., sex, mortality status)

4. Types of Variables

A. Continuous

B. Discrete

- Any variable that is not continuous
- **Ordered numeric** variables have numerical, non-continuous values (e.g., number of deaths, number of school absences)
- **Ordered categories** are non-numeric, but have ordered values (e.g., socioeconomic status as high, middle, or low)
- **Unordered categories** have no inherent order (e.g., race, study site)
- **Binary** variables have just two categories (e.g., sex, mortality status)

Ordered variables are sometimes called “ordinal.”

Binary variables are also known as “dichotomous.”

Unordered categories are sometimes referred to as “nominal.”

A typical dataset

This is the WCGS data set from the textbook.

- Each **row** is an observation. In this example each row reflects a person. However, rows could also represent other units such as cars, organizations, assays, schools, etc.
- Each **column** is a variable.

```
# A tibble: 3,154 x 6
      age   bmi  chol   sbp arcus dibpat
    <dbl> <dbl> <dbl> <dbl> <int> <chr>
 1     50  31.3   249   132     1 Type A
 2     51  25.3   194   120     0 Type A
 3     59  28.7   258   158     1 Type A
 4     51  22.1   173   126     1 Type A
 5     44  22.3   214   126     0 Type A
 6     47  27.1   206   116     0 Type A
 7     40  23.2   190   122     0 Type A
 8     41  23.0   212   130     0 Type A
 9     50  27.2   130   112     1 Type A
10     43  28.4   233   120     0 Type A
# … with 3,144 more rows
```

4. Types of Variables

Continuous Variables

- age
- bmi
- chol
- sbp

Note that chol and sbp are technically discrete (ordered numeric) variables. In many cases, we can approximate these as continuous when the number of possible values is large.

Dichotomous Variables

- arcus

Unordered Categorical Variables

- dibpat

```
# A tibble: 3,154 x 6
     age   bmi  chol   sbp arcus dibpat
   <dbl> <dbl> <dbl> <dbl> <int> <chr>
 1    50  31.3   249   132     1 Type A
 2    51  25.3   194   120     0 Type A
 3    59  28.7   258   158     1 Type A
 4    51  22.1   173   126     1 Type A
 5    44  22.3   214   126     0 Type A
 6    47  27.1   206   116     0 Type A
 7    40  23.2   190   122     0 Type A
 8    41  23.0   212   130     0 Type A
 9    50  27.2   130   112     1 Type A
10    43  28.4   233   120     0 Type A
# … with 3,144 more rows
```

Recap

- Continuous variables can (in theory) take on any value
- Discrete variables are categorical and further divided into ordered numeric (integer), ordered categorical, unordered categorical, and binary

➢Given a variable, identify its scale of measurement

Test Yourself

For the following questions assessing participants' iced tea consumption, please state the variable type.

How many cups of iced tea do you drink per day?

a) Continuous

b) Discrete: Ordinal

c) Discrete: Nominal

d) Discrete: Binary

Test Yourself

For the following questions assessing participants' iced tea consumption, please state the variable type.

How many cups of iced tea do you drink per day?

a) Continuous

b) Discrete: Ordinal

c) Discrete: Nominal

d) Discrete: Binary

The distinct number of cups of iced tea per day is discrete: ordinal, though we could certainly approximate it as continuous in data analysis.

Test Yourself

For the following questions assessing participants' iced tea consumption, please state the variable type.

Do you drink iced tea on a regular basis?

a) Continuous

b) Discrete: Ordinal

c) Discrete: Nominal

d) Discrete: Binary

Test Yourself

For the following questions assessing participants' iced tea consumption, please state the variable type.

Do you drink iced tea on a regular basis?

a) Continuous

b) Discrete: Ordinal

c) Discrete: Nominal

d) Discrete: Binary

There are only two response options (yes/no) so this is a binary variable.

Test Yourself

For the following questions assessing participants' iced tea consumption, please state the variable type.

What is the volume of iced tea you drink per day?

a) Continuous

b) Discrete: Ordinal

c) Discrete: Nominal

d) Discrete: Binary

Test Yourself

For the following questions assessing participants' iced tea consumption, please state the variable type.

What is the volume of iced tea you drink per day?

a) Continuous

b) Discrete: Ordinal

c) Discrete: Nominal

d) Discrete: Binary

In theory, the volume of iced tea could take on any value (e.g., 17.9854 ounces).

The **Children’s Health Study** is a large longitudinal cohort study of the long-term effects of air pollution in southern California residents.

Why are we talking about this study?

- It has large sample size, good study design, and has been published in high impact journals
- Several faculty at USC are investigators on the study (Peters, Avol, Berhane, Eckel, Franklin, Fruin, Gauderman, Gilliland, Lurmann, Kuenzli, McConnell, Thomas, and more)
- There is a good chance you will encounter this data set at another time in your master’s program

Motivation

It is well-established that air pollution causes acute (short-term) effects

- Increased physician visits
- Short-term lung function changes
- Acute symptoms in asthmatics and other susceptible subgroups

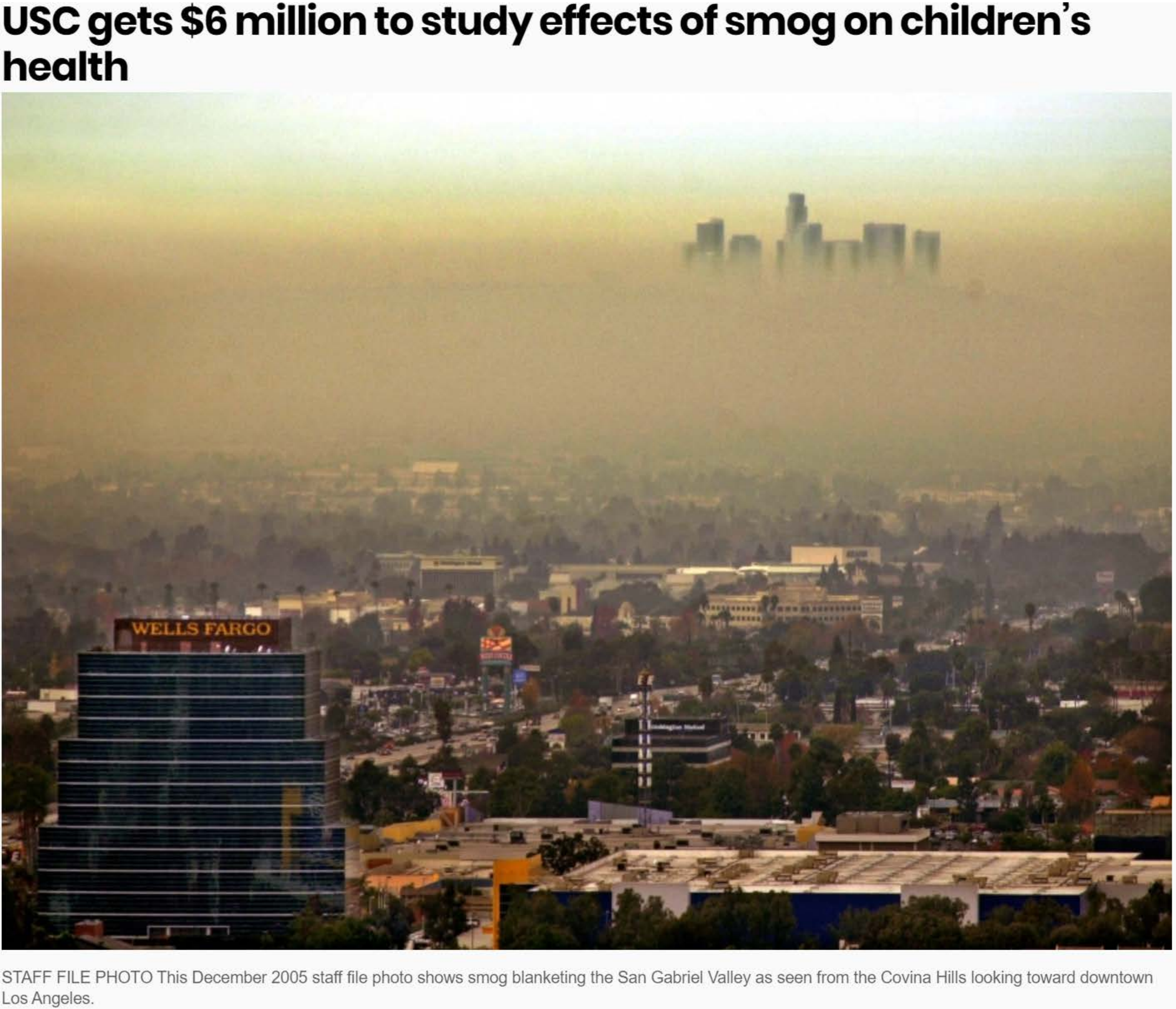
Primary Study Question

Does breathing air pollution cause long-term health effects in children?

5. An Example: The Children's Health Study

The answer to this question isn't that simple!

Air pollution in Southern California is both **regional** and **local**.



USC gets $6 million to study effects of smog on children's health

STAFF FILE PHOTO This December 2005 staff file photo shows smog blanketing the San Gabriel Valley as seen from the Covina Hills looking toward downtown Los Angeles.

5. An Example: The Children's Health Study

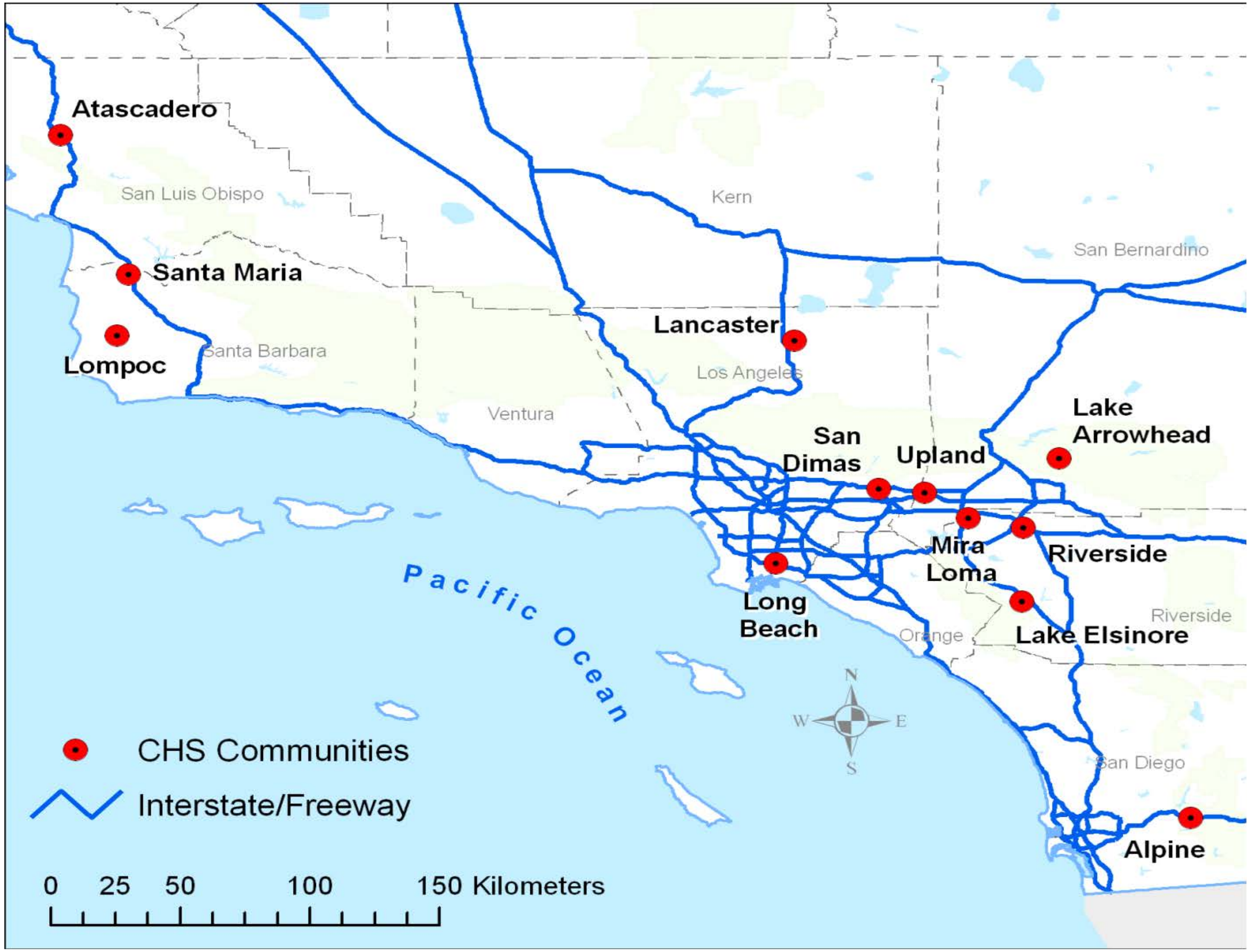
There are also many ways to **address the effects** of air pollution.

- Does it lead to delayed or completely inhibited lung function?
- Is it associated with chronic respiratory symptoms?
- Does it cause more school absences?
- Will it lead to onset of asthma?

Each of these research questions would be a different analysis, with a different model (or set of models), different variables, and possibly different analytic methods!

About the Data

- Participants are from communities with a wide range of air pollution exposures
- 12,000 children have been followed up annually
- Measures of community pollution and lung function measures
- Much more to the data!



6. Choosing a Statistical Package

There are many statistical packages for regression analysis: SAS, Stata, R, SPSS, Splus, JMP, MINITAB, etc.

I’m under the impression that most people in the Biostatistics division use SAS or R, with a contingent of Stata and SPSS users.

You can take programming classes in R or SAS through the department.

We will use R for this class.

R is Free

R is freeware. That’s pretty good in itself.

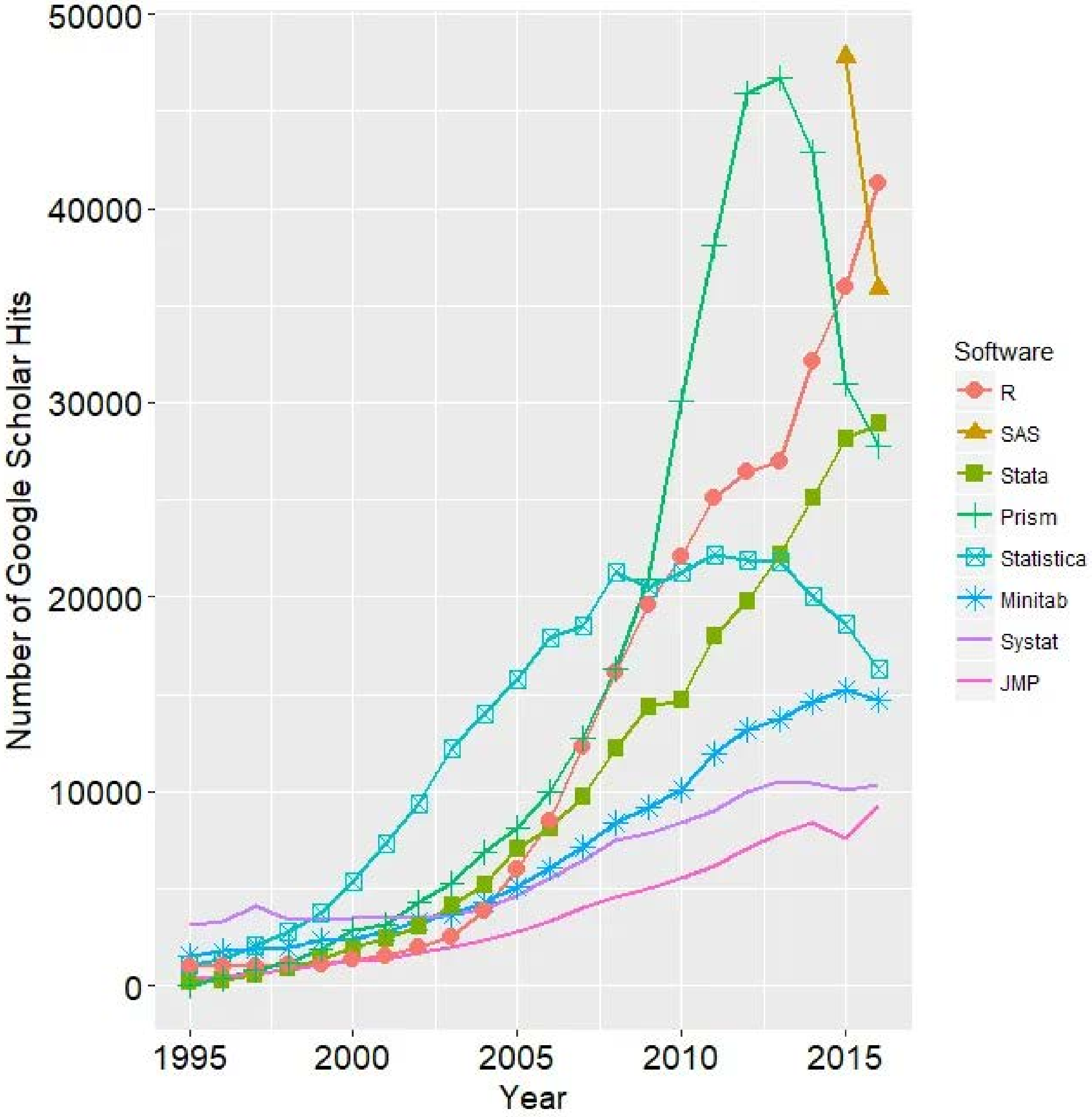
R is Versatile

Because it is open-source, users are able to create custom packages that can be loaded and used. This makes it great for specialty analyses such as:

- Spatial statistics
- Social network analysis
- Latent variable analysis
- Analyses with complex and large data sets
- Integration with high performance computing clusters

R is Becoming More Popular

The popularity of published articles is increasing recently.

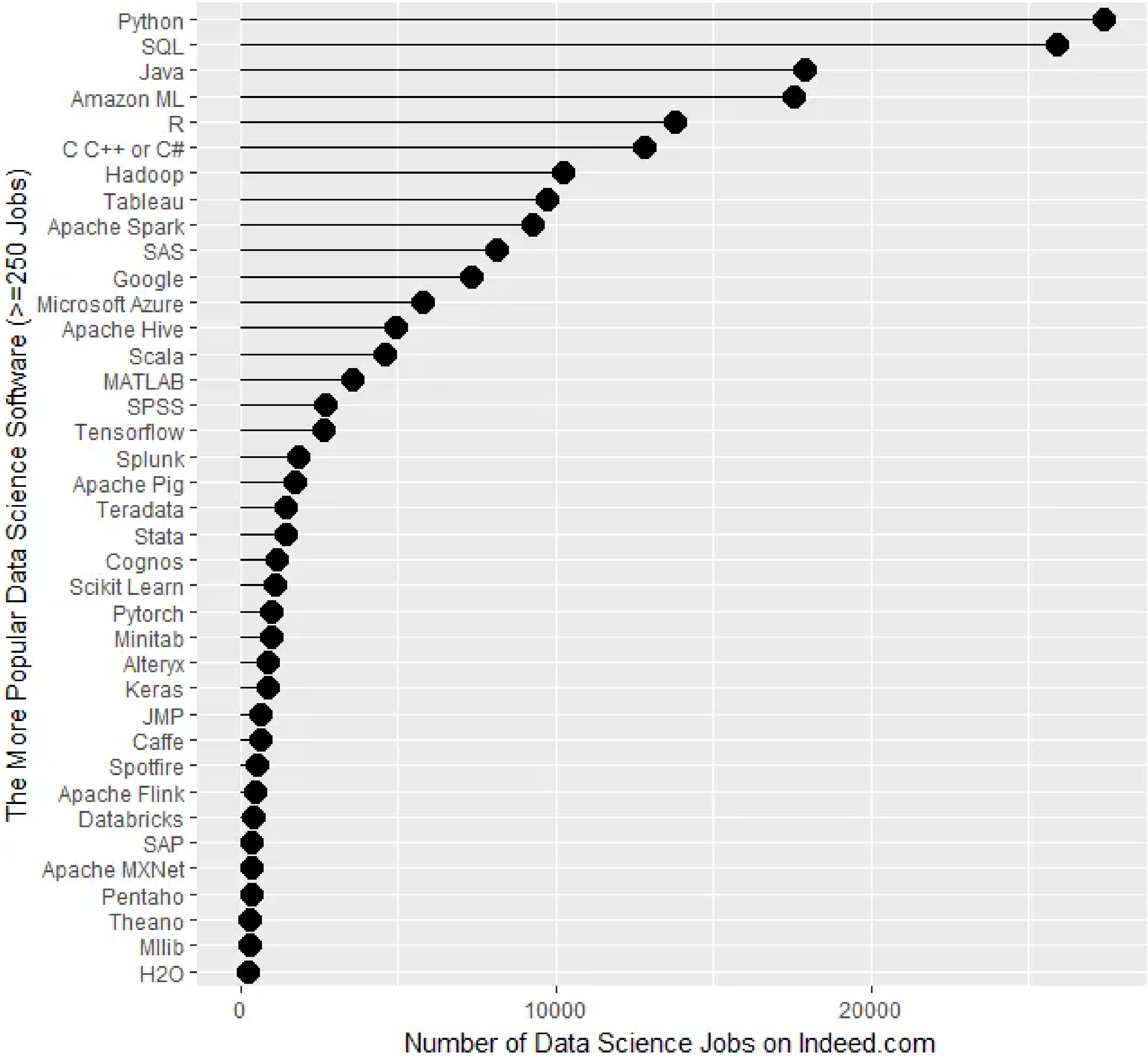


http://r4stats.com/articles/popularity/

R is Becoming More Popular

And R is one of the most common software with statistical capabilities among data scientists.

http://r4stats.com/articles/popularity/



Additional Resources

Our textbook is quite good for regression methods, but uses Stata. Another free textbook that uses R is available at https://leanpub.com/openintro-statistics.

Another great resource is the UCLA Statistical Consulting website https://stats.idre.ucla.edu/r/.

Exploratory data analysis is an important first step when performing statistical analyses.

We want to obtain:

- Summary statistics
  - Central tendency (mean, median, mode)
  - Percentiles (quartiles, quintiles, deciles)
  - Measures of variability (variance, standard deviation, range)
- Graphical displays
  - Histogram
  - Boxplot

Why do we do this?

- To become familiar with the data
- To detect errors in the dataset
- To begin assessing assumptions for model fitting
- To look at missingness in data

You should always explore your data before you proceed to confirmatory data analysis.

Example

Vittinghoff: The western collaborative group study (WCGS) was a large epidemiological study designed to investigate the association between the “type A” behavior pattern and coronary heart disease (CHD) (Rosenman et al. 1964).

7. Exploratory Analysis

wcgs is a “tibble,” which is a type of data set

```
> wcgs
# A tibble: 3,154 x 22
     age arcus behpat   bmi chd69  chol   dbp dibpat height    id lnsbp lnwght
   <dbl> <dbl> <chr>  <dbl> <chr> <dbl> <dbl> <chr>   <dbl> <dbl> <dbl>  <dbl>
 1    50     1 A1      31.3 No      249    90 Type A     67  2343  4.88   5.30
 2    51     0 A1      25.3 No      194    74 Type A     73  3656  4.79   5.26
 3    59     1 A1      28.7 No      258    94 Type A     70  3526  5.06   5.30
 4    51     1 A1      22.1 No      173    80 Type A     69 22057  4.84   5.01
 5    44     0 A1      22.3 No      214    80 Type A     71 12927  4.84   5.08
 6    47     0 A1      27.1 No      206    76 Type A     64 16029  4.75   5.06
 7    40     0 A1      23.2 No      190    78 Type A     70  3894  4.80   5.09
 8    41     0 A1      23.0 No      212    84 Type A     70 11389  4.87   5.08
 9    50     1 A1      27.2 No      130    70 Type A     71 12681  4.72   5.27
10    43     0 A1      28.4 No      233    80 Type A     68 10005  4.79   5.23
# … with 3,144 more rows, and 10 more variables: ncigs <dbl>, sbp <dbl>,
#   smoke <chr>, t1 <dbl>, time169 <dbl>, typchd69 <dbl>, uni <dbl>, weight <dbl>,
#   wghtcat <chr>, agec <chr>
```

<dbl> refers to “double” which is a way of storing a numeric variable

<chr> is “character” which is a type of string/word variable

Other variable types can include <fct> which is a factor/categorical variable, or <int> which is an integer variable, and more.

https://swcarpentry.github.io/r-novice-inflammation/13-supp-data-structures/

7. Exploratory Analysis

```
> wcgs %>%
+   select(age, bmi, chol, sbp) %>%
+   skim()
── Data Summary ────────────────────────
                           Values
Name                       Piped data
Number of rows             3154
Number of columns          4
_______________________
Column type frequency:
  numeric                  4
________________________
Group variables            None

── Variable type: numeric ──────────────────────────────────────────────
  skim_variable n_missing complete_rate  mean    sd    p0   p25   p50   p75  p100 hist
1 age                   0         1      46.3  5.52  39    42    45    50    59   ▇▅▃▃▂
2 bmi                   0         1      24.5  2.57  11.2  23.0  24.4  25.8  38.9 ▁▃▇▁▁
3 chol                 12         0.996 226.  43.4  103   197.  223   253   645   ▅▇▁▁▁
4 sbp                   0         1     129.  15.1   98   120   126   136   230   ▇▇▁▁▁
```

The data set is piped to the next line for further operations

Of all the variables, “age,” “bmi,” “chol,” and “sbp” are selected.

Then we run the “skim” function, which summarizes data.

7. Exploratory Analysis

```
> wcgs %>%
+   select(age, bmi, chol, sbp, dibpat) %>%
+   group_by(dibpat) %>%
+   skim()
── Data Summary ────────────────────────
                           Values
Name                       Piped data
Number of rows             3154
Number of columns          5
_______________________
Column type frequency:
  numeric                  4
________________________
Group variables            dibpat

── Variable type: numeric
  skim_variable dibpat n_missing complete_rate  mean    sd    p0   p25   p50   p75  p100 hist
1 age           Type A         0         1       46.8  5.70  39    42    46    51    59   ▇▅▅▃▂
2 age           Type B         0         1       45.8  5.30  39    41    45    49    59   ▇▅▃▂▁
3 bmi           Type A         0         1       24.6  2.60  11.2  23.0  24.4  25.8  37.2 ▁▁▇▂▁
4 bmi           Type B         0         1       24.4  2.53  15.7  22.8  24.4  25.8  38.9 ▁▇▅▁▁
5 chol          Type A         5         0.997 229.   44.3  103   200   226   256   645   ▅▇▁▁▁
6 chol          Type B         7         0.996 224.   42.4  110   195   221   251   400   ▂▇▆▁▁
7 sbp           Type A         0         1     130.   15.7  100   120   128   138   212   ▇▇▂▁▁
8 sbp           Type B         0         1     127.   14.4   98   118   124   136   230   ▇▇▁▁▁
```

The data set will be separated into two groups by "dibpat" (personality type).

Since we are using R, there are many ways to perform descriptive statistics. Here’s an example using Base R (no package needs to be installed).

```
> summary(wcgs[, c("age", "bmi", "chol", "sbp")])
      age             bmi             chol             sbp       
 Min.   :39.00   Min.   :11.19   Min.   :103.0   Min.   : 98.0  
 1st Qu.:42.00   1st Qu.:22.96   1st Qu.:197.2   1st Qu.:120.0  
 Median :45.00   Median :24.39   Median :223.0   Median :126.0  
 Mean   :46.28   Mean   :24.52   Mean   :226.4   Mean   :128.6  
 3rd Qu.:50.00   3rd Qu.:25.84   3rd Qu.:253.0   3rd Qu.:136.0  
 Max.   :59.00   Max.   :38.95   Max.   :645.0   Max.   :230.0  
                                 NA's   :12                     
```

Tips

- Find the package that suits **you** best
- Packages typically have **vignettes** for use; take advantage of them (e.g., https://cran.r-project.org/web/packages/skimr/vignettes/skimr.html)
- Packages will be **updated** and **new packages** will be released; keep looking

What if we wanted to create a table of behavior pattern by heart disease? There are a couple ways we can do this.

```
> table(wcgs$chd69, wcgs$dibpat)

      Type A Type B
  No    1411   1486
  Yes    178     79
```

```
> wcgs %>% count(chd69, dibpat)
# A tibble: 4 x 3
  chd69 dibpat     n
  <chr> <chr>  <int>
1 No    Type A  1411
2 No    Type B  1486
3 Yes   Type A   178
4 Yes   Type B    79
```

Some useful analyses for **one continuous variable**.

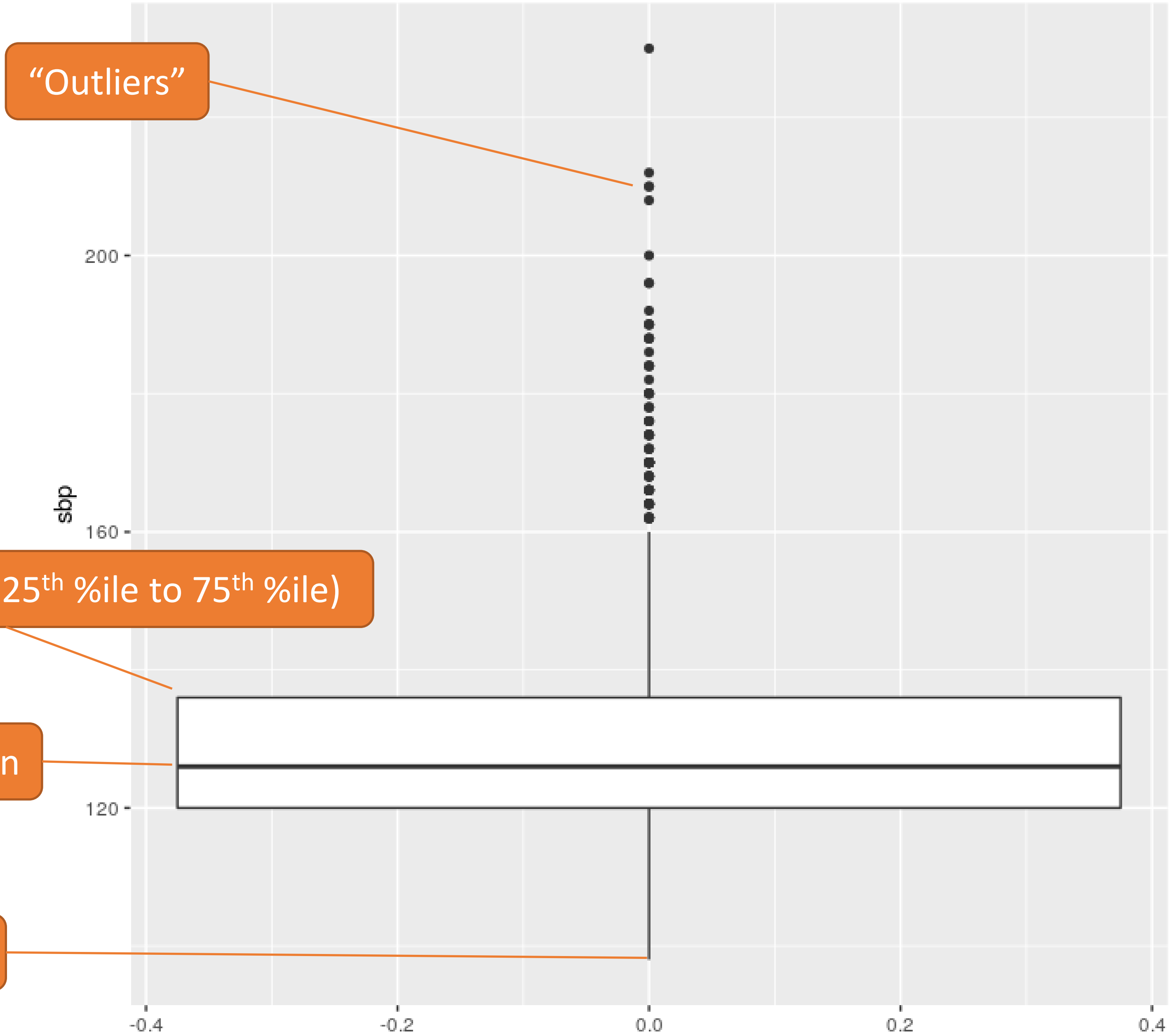
- Percentiles
- Central tendency (mean, median, mode)
- Variability or dispersion (range, variance, SD, IQR)

7. Exploratory Analysis

Some useful analyses for **one continuous variable**.

- Boxplots
  - Can detect outliers
  - Provides summary of information

```
wcgs %>%
ggplot(aes(y=sbp)) +
geom_boxplot()
```



“Outliers”

Interquartile Range (25th %ile to 75th %ile)

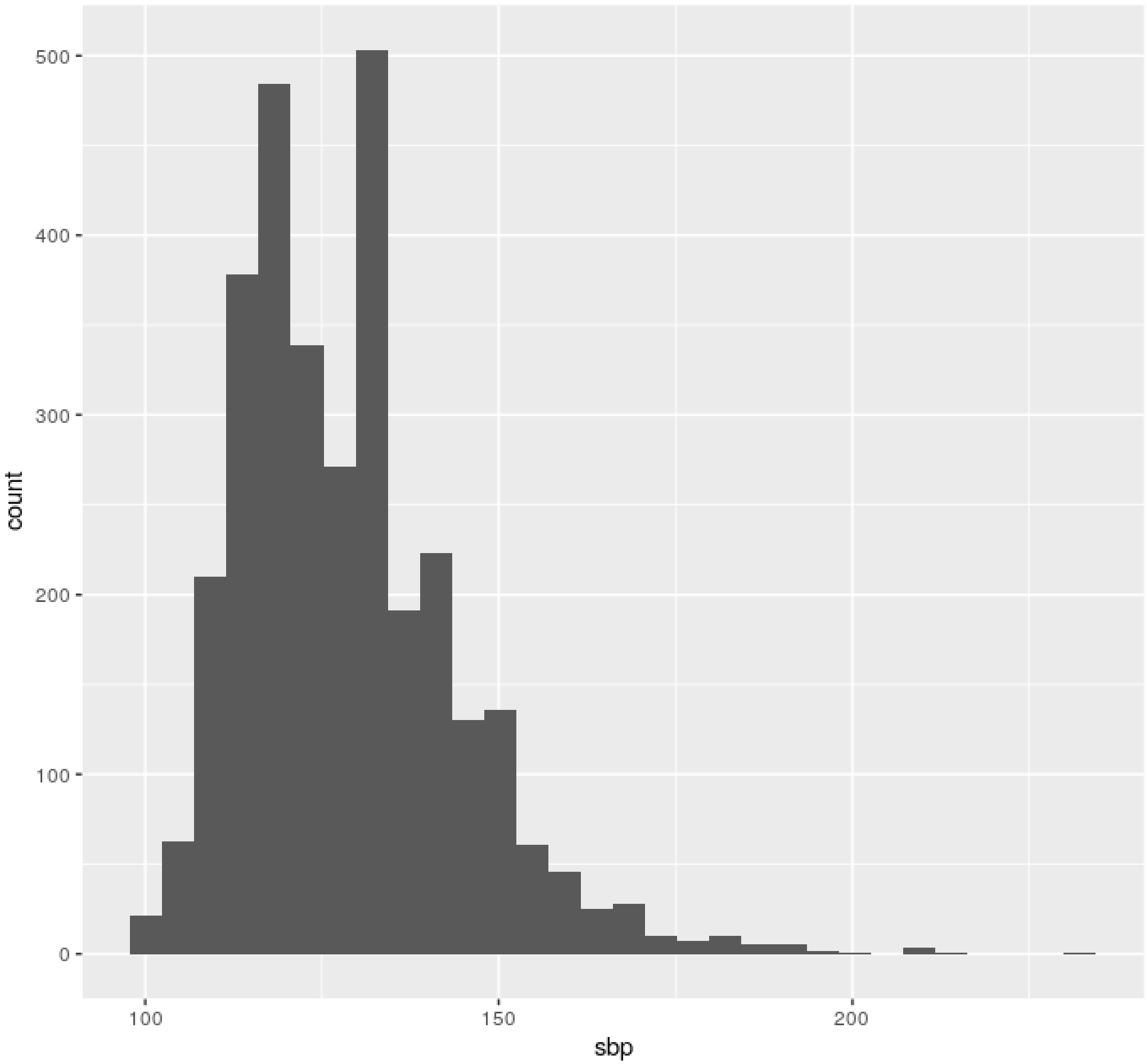
Median

1.5 IQR

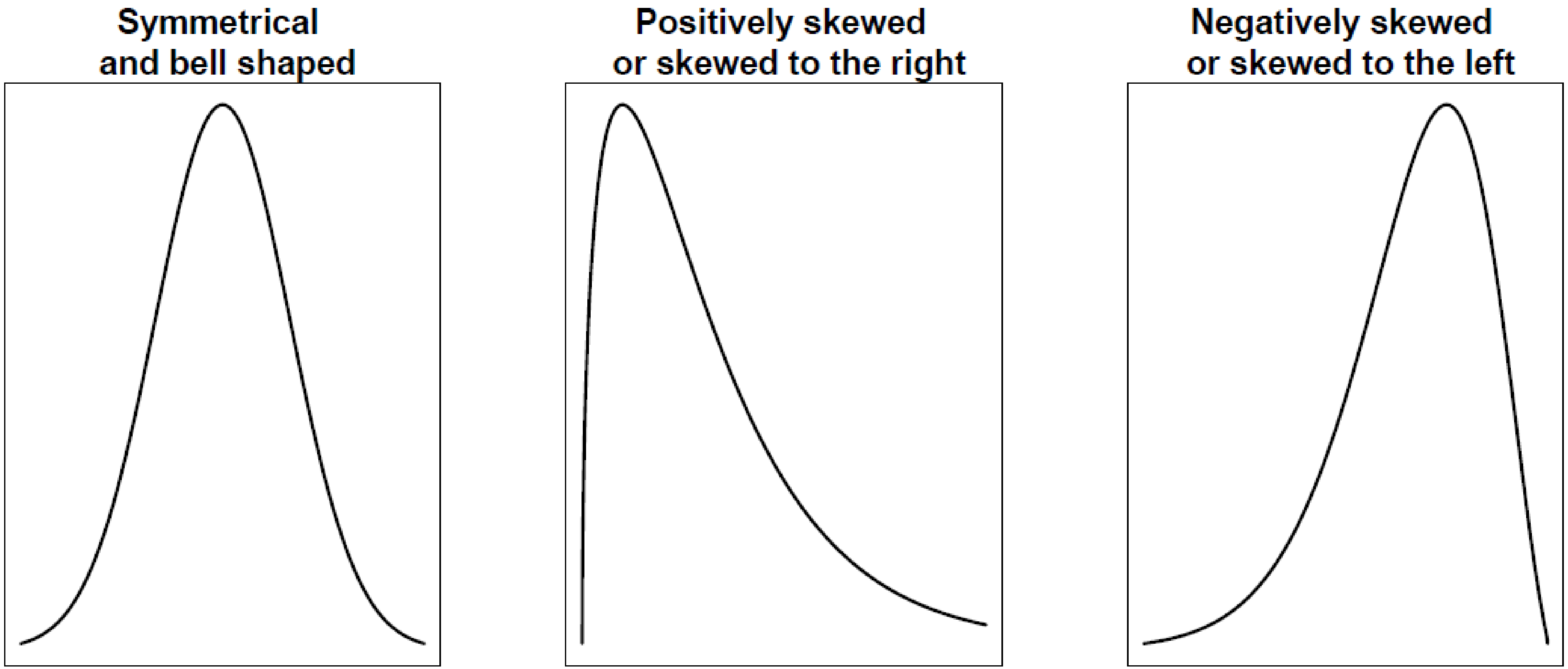
Some useful analyses for **one continuous variable**.

- Histograms
  - More information about the shape of the distribution

```
wcgs %>%
ggplot(aes(x=sbp)) +
geom_histogram()
```



Common shapes of variable distributions



Symmetrical and bell shaped

Positively skewed or skewed to the right

Negatively skewed or skewed to the left

Some useful analyses for **one categorical variable**.

- Frequency tables
- Frequency bar charts

Some useful analyses for **one categorical variable**.

- Frequency tables

```
> wcgs %>% count(wghtcat)
# A tibble: 4 x 2
  wghtcat      n
  <chr>    <int>
1 < 140      232
2 > 200      213
3 140-170   1538
4 170-200   1171
```

These are two distinct ways of arriving at the same information.

```
> with(wcgs, table(wghtcat))
wghtcat
  < 140   > 200 140-170 170-200 
    232     213    1538    1171 
```

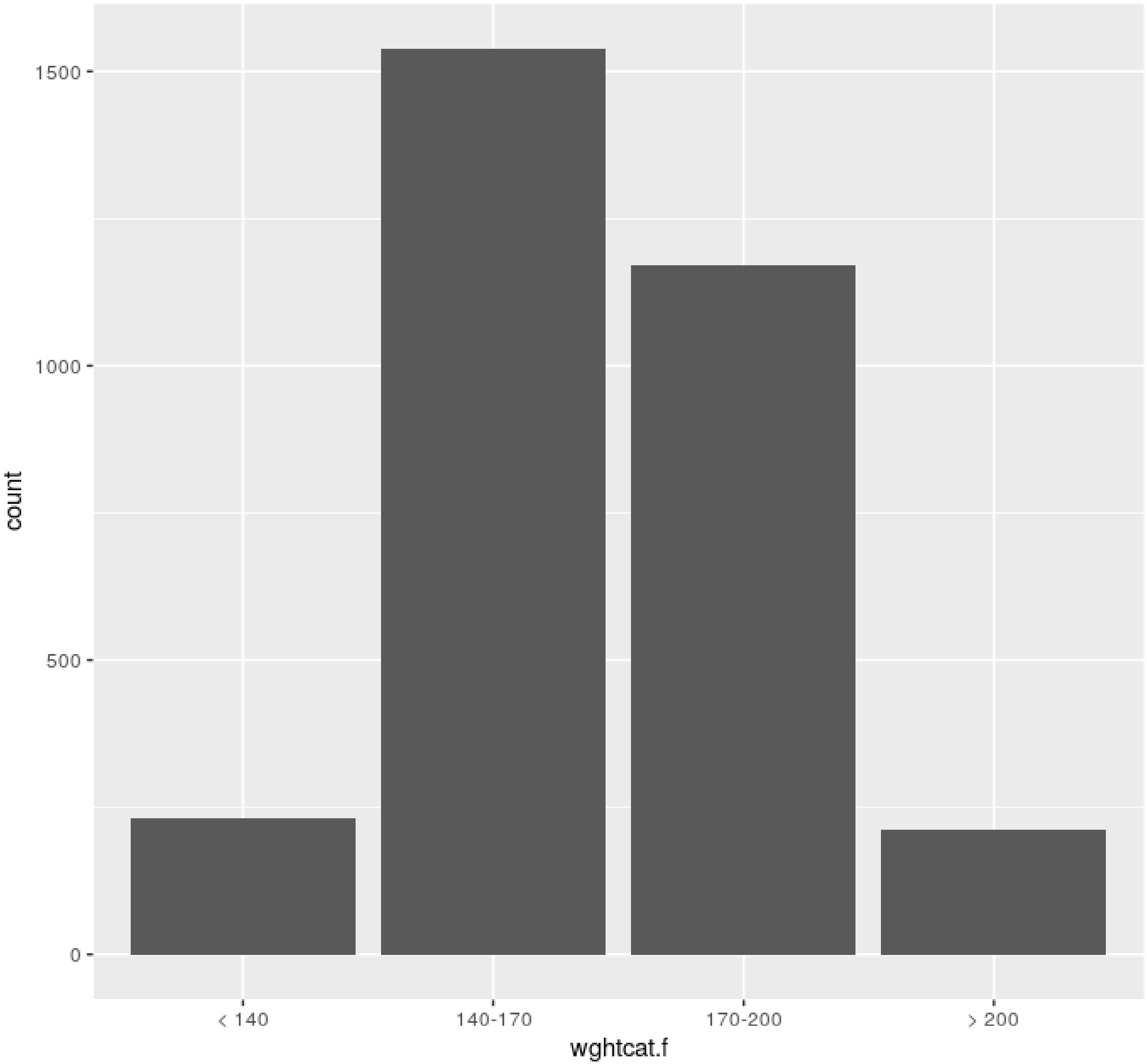
How could we visualize this information?

Some useful analyses for **one categorical variable**.

- Frequency tables

```
> wcgs %>% count(wghtcat)
# A tibble: 4 x 2
  wghtcat     n
  <chr>   <int>
1 < 140     232
2 > 200     213
3 140-170  1538
4 170-200  1171
```

```
> with(wcgs, table(wghtcat))
wghtcat
  < 140   > 200 140-170 170-200 
    232     213    1538    1171 
```



7. Exploratory Analysis

Some useful analyses for **one categorical variable**.

- Proportions (dichotomous variables)

```
> table(wcgs$dibpat)

Type A Type B
  1589   1565


> table(wcgs$dibpat) %>%
       prop.table()

   Type A    Type B
0.5038047 0.4961953
```

Some useful analyses for **one categorical variable**.

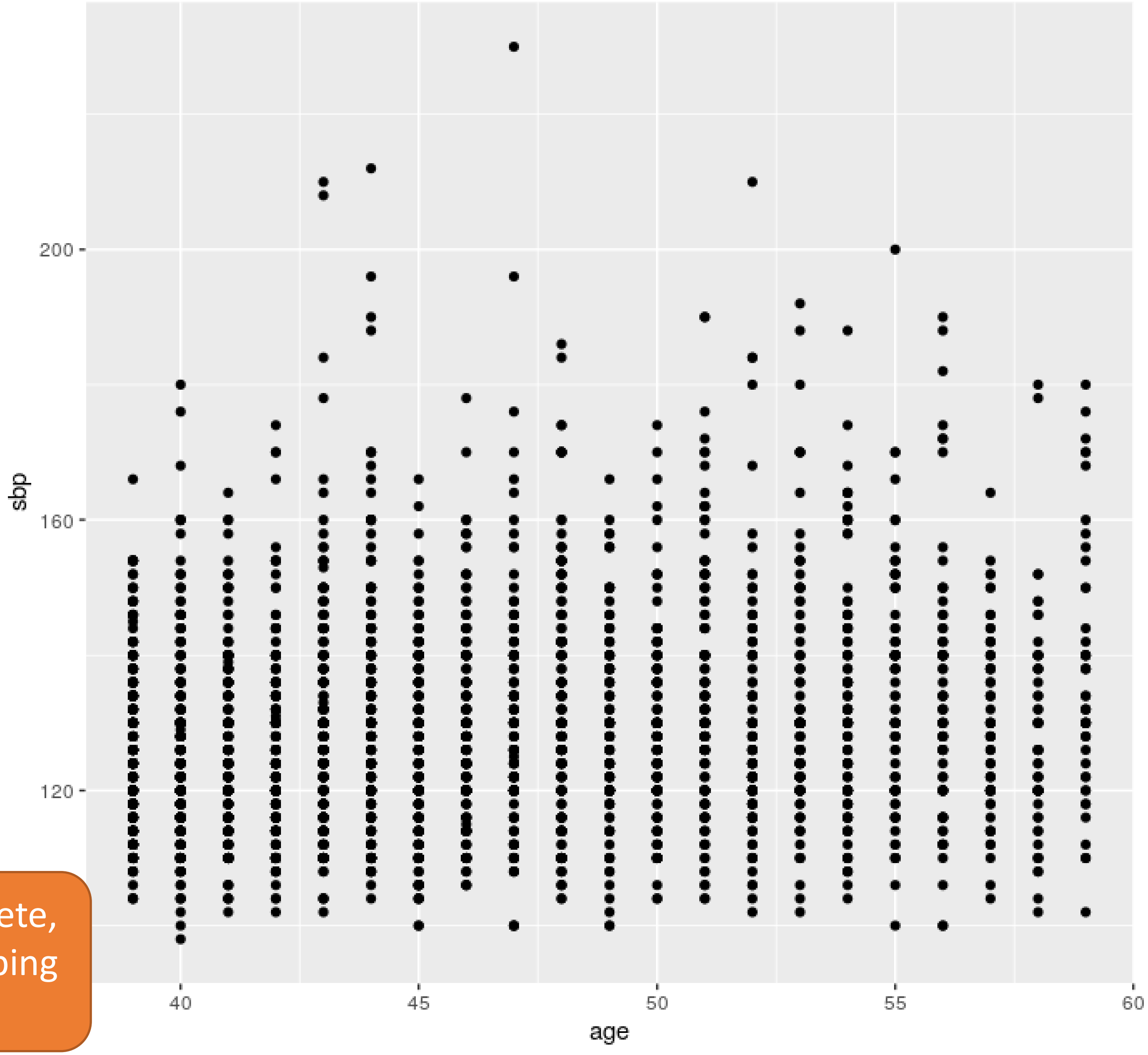
- Proportions (dichotomous variables)

```
> wcgs %>%
+   group_by(dibpat) %>%
+   summarise(n = n()) %>%
+   mutate(pct = n / sum(n))
`summarise()` ungrouping output (override with `.groups` argument)
# A tibble: 2 x 3
  dibpat     n   pct
  <chr>  <int> <dbl>
1 Type A  1589 0.504
2 Type B  1565 0.496
```

Some useful analyses for **two continuous variables**.

- Scatterplots

```
wcgs %>%
ggplot(aes(x = age, y = sbp)) +
geom_point()
```

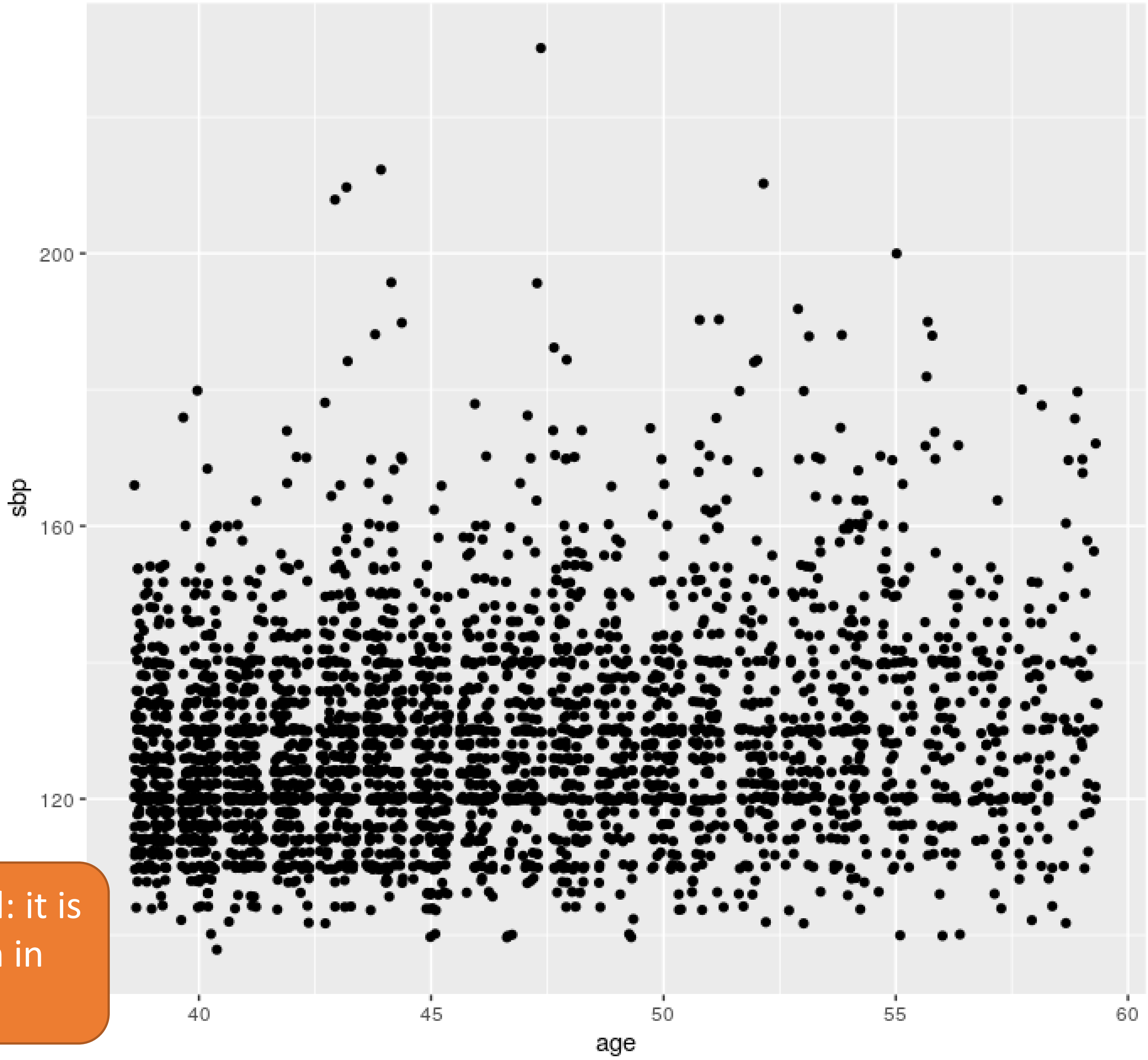


Problem: because age is discrete, there may be several overlapping age values.

Some useful analyses for **two continuous variables**.

- Scatterplots (jittered)

```
wcgs %>%
ggplot(aes(x = age, y = sbp)) +
geom_jitter()
```



Keep in mind that these values are jittered: it is not the actual data but a representation in order to better visualize the data.

Some useful analyses for **one continuous, one categorical variable**.

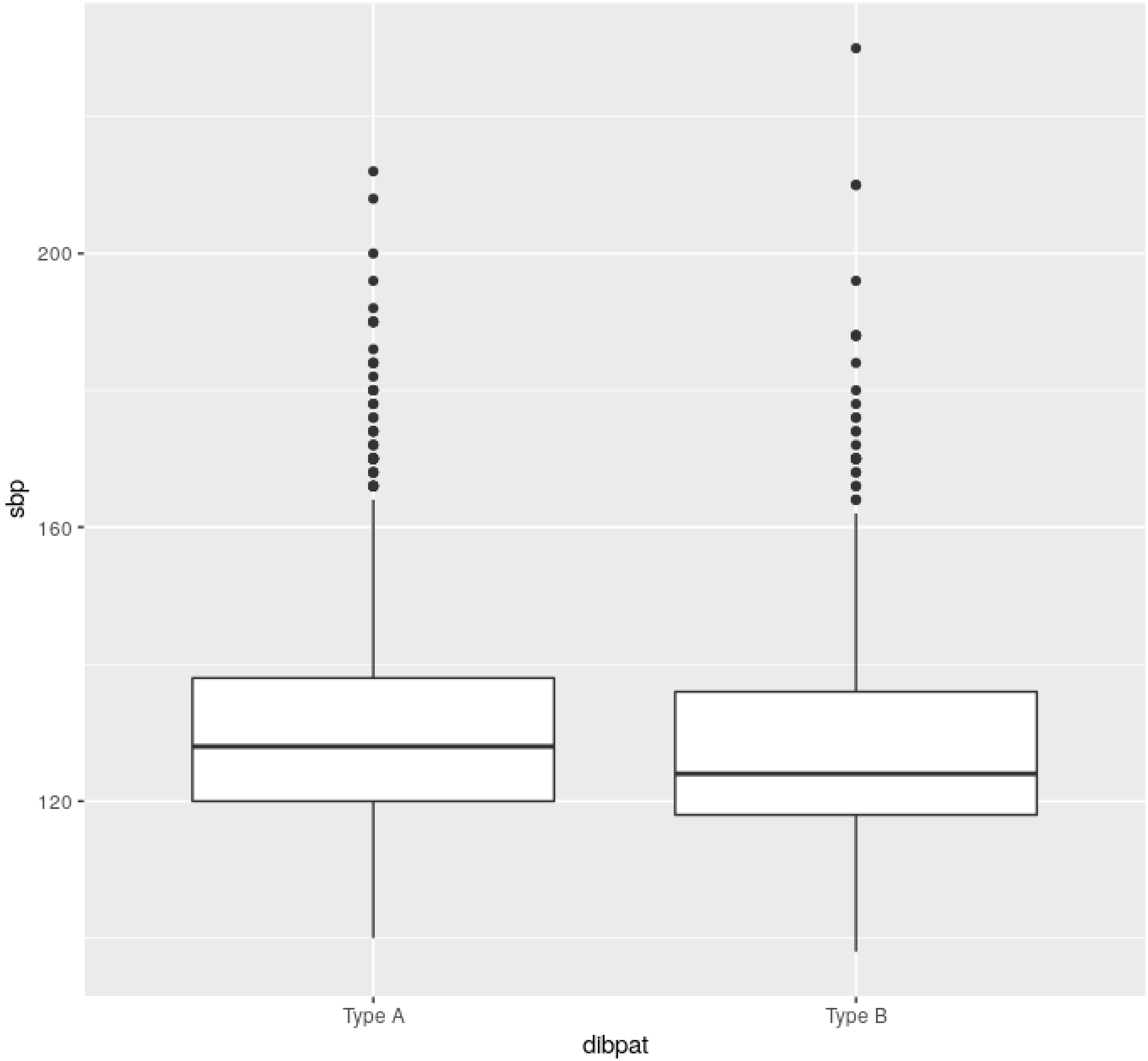
- Means (and SD) by group
- Boxplots by group

```
> wcgs %>%
+   group_by(wghtcat) %>%
+   summarise(sbp=mean(sbp, na.rm=TRUE),
+             age=mean(age, na.rm=TRUE))
`summarise()` ungrouping output (override with `.groups` argument)
# A tibble: 4 x 3
  wghtcat   sbp   age
  <chr>   <dbl> <dbl>
1 < 140    123.  46.9
2 > 200    138.  46.0
3 140-170  126.  46.3
4 170-200  131.  46.2
```

Some useful analyses for **one continuous, one categorical variable**.

- Means (and SD) by group
- Boxplots by group

```
wcgs %>%
ggplot(aes(x=dibpat, y=sbp)) +
geom_boxplot()
```



Some useful analyses for **two categorical variables**.

- Frequency tables

```
> library(gmodels)
> CrossTable(wcgs$wghtcat, wcgs$dibpat, prop.t=F, prop.chisq=F)


   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|-------------------------|


Total Observations in Table:  3154


             | wcgs$dibpat
wcgs$wghtcat |    Type A |    Type B | Row Total |
-------------|-----------|-----------|-----------|
       < 140 |       120 |       112 |       232 |
             |     0.517 |     0.483 |     0.074 |
             |     0.076 |     0.072 |           |
-------------|-----------|-----------|-----------|
       > 200 |       120 |        93 |       213 |
             |     0.563 |     0.437 |     0.068 |
             |     0.076 |     0.059 |           |
-------------|-----------|-----------|-----------|
     140-170 |       737 |       801 |      1538 |
             |     0.479 |     0.521 |     0.488 |
             |     0.464 |     0.512 |           |
-------------|-----------|-----------|-----------|
     170-200 |       612 |       559 |      1171 |
             |     0.523 |     0.477 |     0.371 |
             |     0.385 |     0.357 |           |
-------------|-----------|-----------|-----------|
Column Total |      1589 |      1565 |      3154 |
             |     0.504 |     0.496 |           |
-------------|-----------|-----------|-----------|
```

Recap

- It is always a good idea to perform exploratory data analysis with your data to "get to know" it better
- The type of numerical and visual analyses you perform will depend on the type of variable and whether you are looking at one or multiple variables

➢Perform the correct exploratory method given a particular variable type

Test Yourself

What information could we present in order to…

a) Describe the typical SAT score of incoming USC undergraduate students?

b) Visualize the distribution of ages of students entering the MS program?

c) Describe the association between political party and presidential vote choice?

d) Visualize the number of individuals who picked each flavor at Baskin Robbins as their favorite?

Test Yourself

What information could we present in order to…

a) Describe the typical SAT score of incoming USC undergraduate students?
We would use a measure of central tendency, such as the median or mode, to describe the "typical" values.

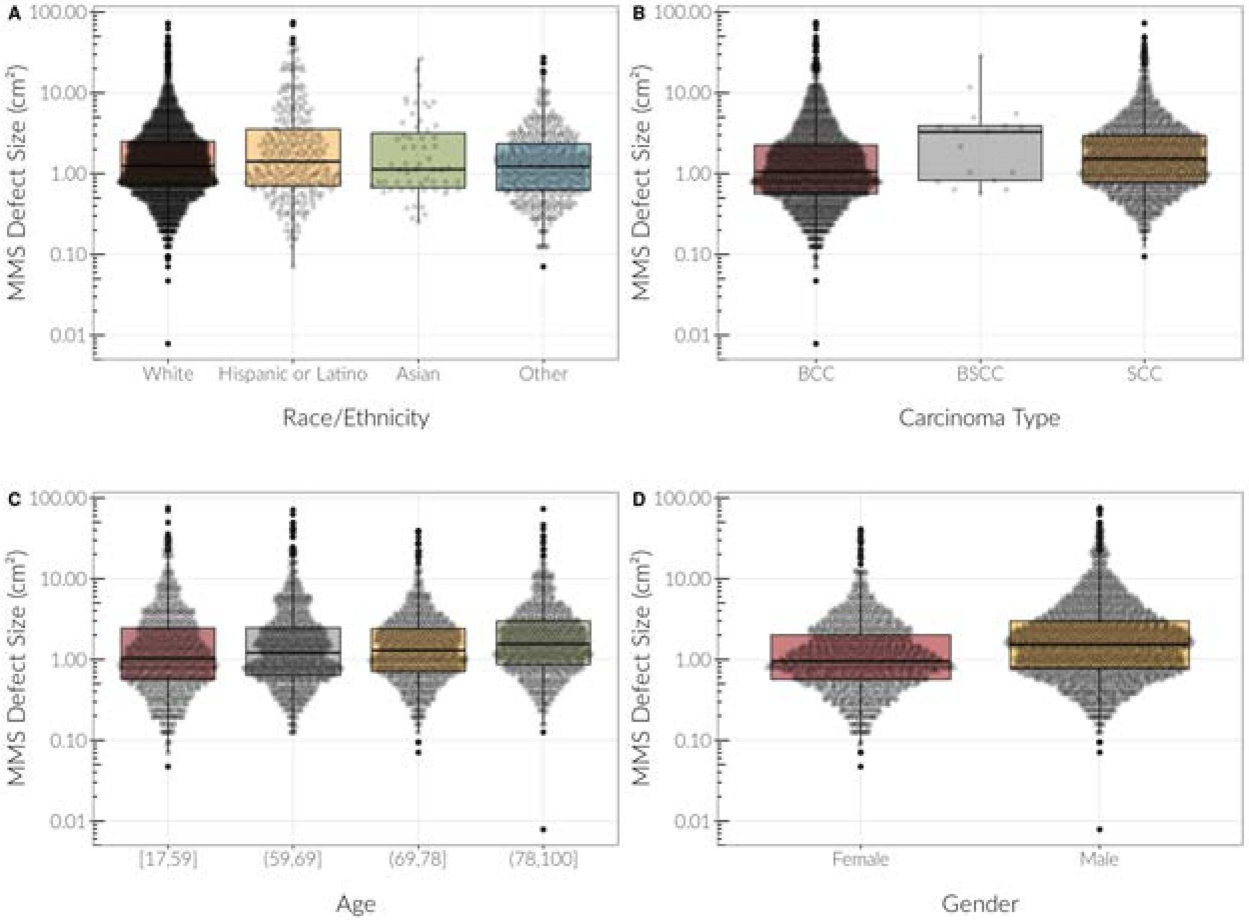
b) Visualize the distribution of ages of students entering the MS program?
To visualize a distribution we could use a histogram or boxplot.

c) Describe the association between political party and presidential vote choice?
We could produce a frequency table of this information.

d) Visualize the number of individuals who picked each flavor at Baskin Robbins as their favorite?
We could visualize this with a bar chart.

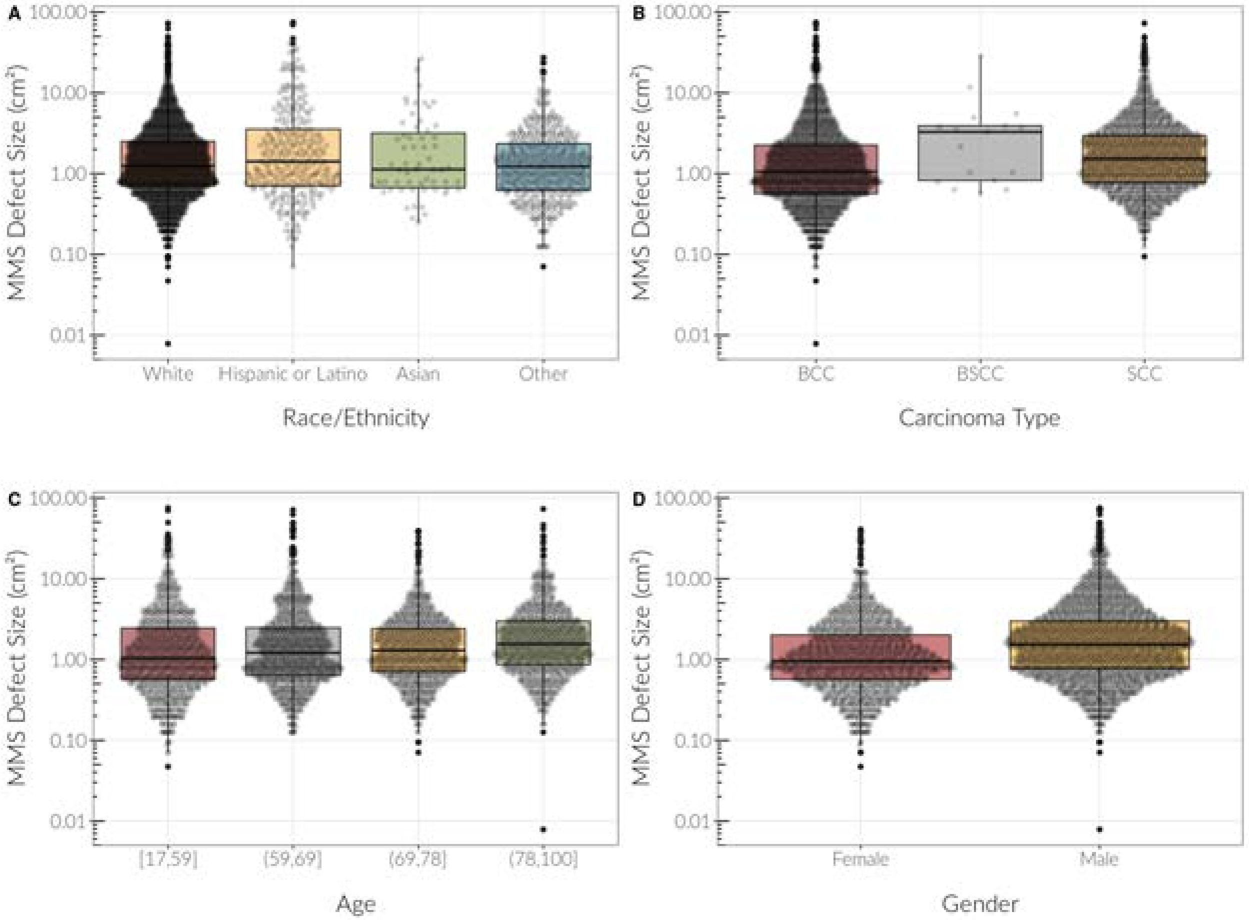
Test Yourself

Which “geom_” was used to create this?



Test Yourself

Which “geom_” was used to create this? geom_boxplot



Test Yourself

You want to examine the relationship between duration of exercise (continuous) and heart rate (continuous). How could you explore this relationship?

a) geom_bar

b) geom_boxplot

c) geom_histogram

d) geom_jitter

e) geom_point

Test Yourself

You want to examine the relationship between duration of exercise (continuous) and heart rate (discrete). How could you explore this relationship?

a) geom_bar

b) geom_boxplot

c) geom_histogram

[✓ d) geom_jitter]

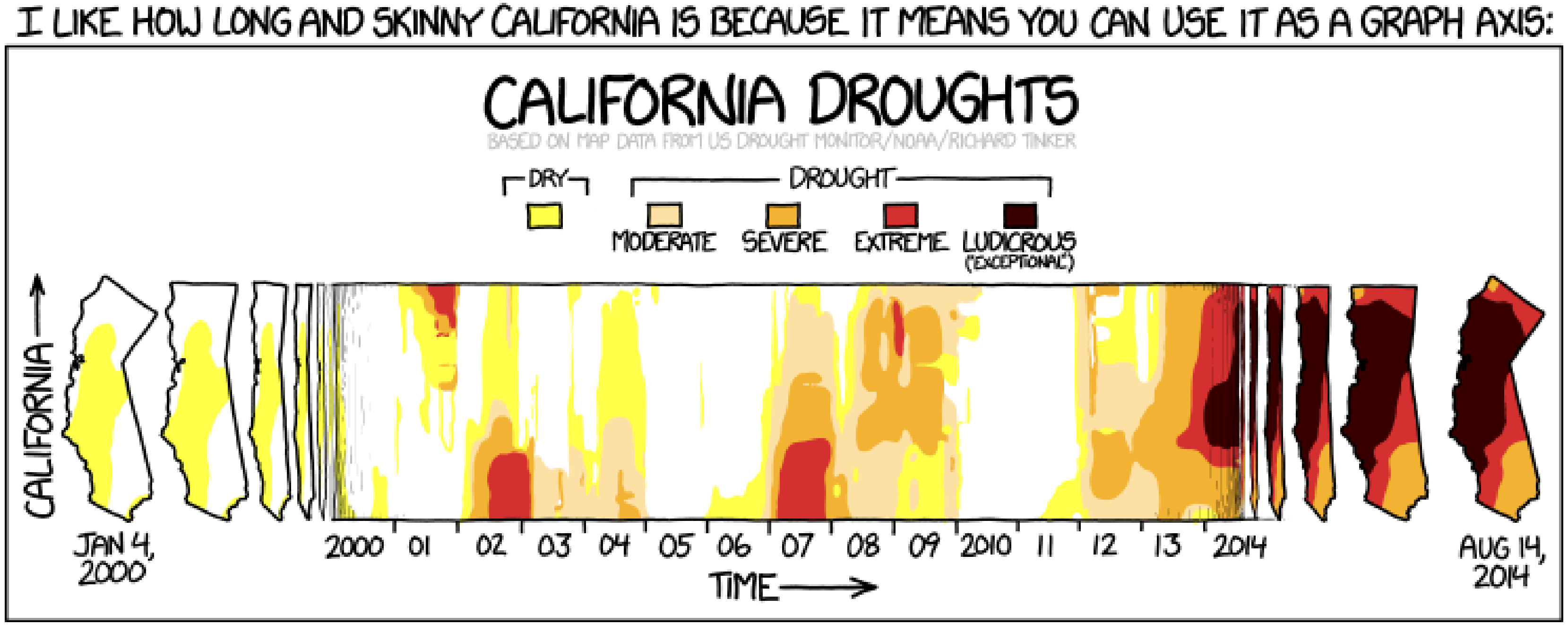
e) geom_point

Since heart rate is typically recorded to the nearest integer, using geom_point may cause some points to overlap on each other. It may be better to use geom_jitter to help see overlapping points.

8. Recap

Always explore your data and seek the most effective way to display interesting patterns.

Don’t be afraid to be creative; of all software, R gives you the most versatility to accomplish what you’d like.



http://www.latimes.com/science/la-me-g-california-drought-map-htmlstory.html

Additional Reading

- Gelman, Andrew, Cristian Pasarica, and Rahul Dodhia. "Let's practice what we preach: turning tables into graphs." *The American Statistician* 56.2 (2002): 121-130.
- Behrens, John T. "Principles and procedures of exploratory data analysis." Psychological Methods 2.2 (1997): 131.
- An Introduction to R (useful as a reference manual) https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf
- Getting Started with R (useful to help orient yourself to R if you are unfamiliar) https://moderndive.com/1-getting-started.html
- Questions in A/B testing https://towardsdatascience.com/6-questions-to-understand-a-b-testing-8d0ed05e5cc4
- The Art of Storytelling in Analytics and Data Science https://www.analyticsvidhya.com/blog/2020/05/art-storytelling-analytics-data-science/

Packages & Functions

- `base::summary` (generic function, will make a descriptive stats summary on a data set)
- `base::factor` (encodes a vector as a factor)
- `base::table` (create a 1- or 2-way contingency table)
- `readxl::read_xls` (read Excel files)
- `dplyr::select` (select variables)
- `dplyr::group_by` (group data by a categorical variable)
- `dplyr::count` (counts unique values of one or more variables)
- `dplyr::mutate` (create or replace variables in the data set)
- `dplyr::summarise` (create a new data set with summary statistics for each group of observations)
- `skimr::skim` (descriptive statistics)
- `ggplot2::ggplot` (creates a ggplot graphics object)
- `ggplot2::geom_boxplot` (boxplot)
- `ggplot2::geom_point` (scatterplot)
- `ggplot2::geom_jitter` (jittered scatterplot)
- `gmodels::CrossTable` (cross-tabulation of categorical variables)