IST 687 Introduction to Visualization

Corey Jackson

2021-02-17 14:47:45

Today's Agenda

- Announcements
- Exam logistics
- Wrapping up week 5 - Connecting using different data sources
- Week 6 - Introduction to data visualization
- Breakout (Lab & Project Updates)
- Homework 6 Tips
- Next week's agenda

Announcements

- Office Hours: Wed. 6-7pm EDT and by appointment
- Upcoming Schedule:
  - Week 7: Working with map data (caution working ahead)
  - Week 8: Linear modeling & Mid-term
    - 30 minute Live Session in Week 8 (48 hour window to complete the exam. **Due: Friday, March, 5th at 9:30 EST**)
    - Practice exam available Friday, February 26th via Syllabus Link under Week 8
- Team Process Agreement (**Due: Monday August 22nd 11:59 pm EDT**). Submit via SLACK

Exam Logistics

Exam Logistics

- **Format**
  - Closed book/notes/R
  - 1 hour time limit (no pausing)
- **Materials covered**: Weeks 1-8
- **Question types**
  - Given code what is the expected output: 2
  - Write code to perform: 10
  - Open-ended questions: 9

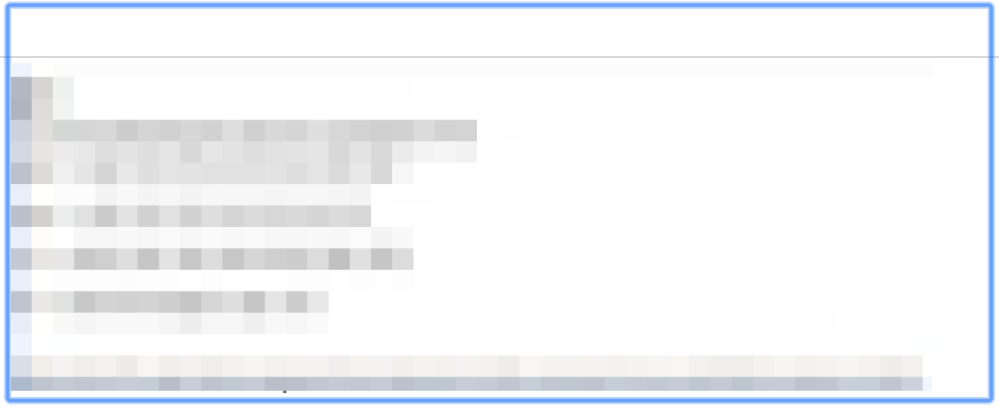
Question distribution

Week	# Questions
2 - Using R to manipulate data.	8
3 - Descriptive Statistics & Functions	5
4 - Inferential statistics	4
6 - Introduction to visualization	1
7 - Working with map data	1
8 - Linear modeling	2

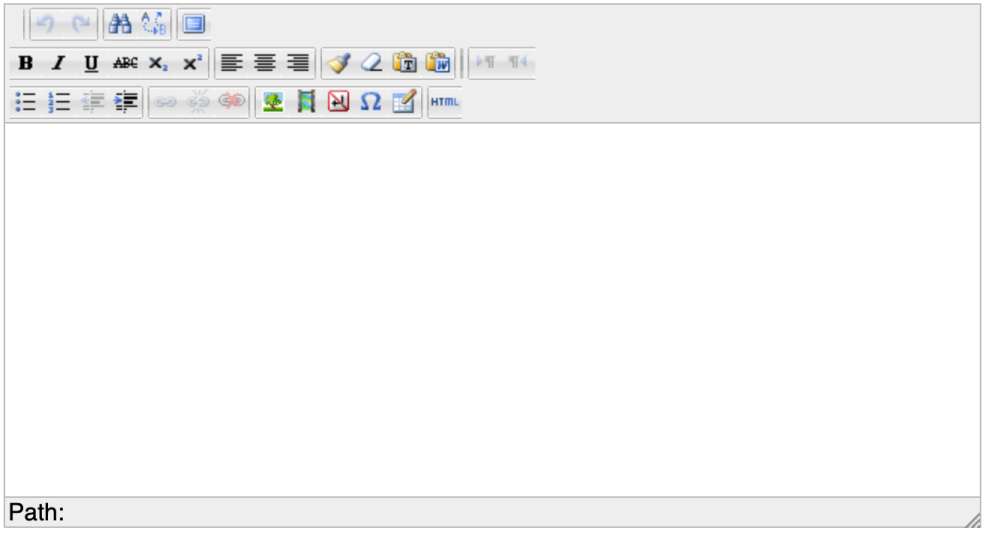
Exam office hours: Wednesday, August 19th (Zoom from 5-6 pm EDT) or post questions on SLACK

Exam interface

Midterm Quiz



Answer:



QUIZ REPORTS

- Info
- Overview & Regrade
- Manual Grading
- Item Analysis
- **Preview**
- Quiz Settings
- User Overrides

Go to Gradebook to publish scores to students &raquo

1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21				

Finish Exam

Time left **0:59:45**

Start a new preview

Week 5 - Connecting using different data sources

- Importing data from different sources e.g., JSON
- Querying data frames using SQL and R functions
- Data munging

Week 5 - Connecting using different data sources

Book (CH 11) and asynchronous topics covered

- Coding using more "complex" queries on data frames
  - `sqldf()` allows SQL queries in R
  - `tapply()` format/arguments is `tapply(summary variable, group variable, function)`
  - `which()` returns the position of the elements in a logical vector
- Importing non-tabular data
  - RJSONIO gets JSON data and places it in a list using `fromJSON()`
  - R packages for other data formats: Data import tutorials

Week 5 - Connecting using different data sources

- Conflicting package functions with (RJSONIO and jsonlite)
  - To detach packages: `detach("package:jsonlite")`
  - Try namespace calls e.g., `RJSONIO::fromJSON()`
  - Reconciling duplicated functions from packages

Week 5 - Connecting using different data sources

- Data munging JSON to R readable
  - `unlist()` takes a list and returns a simple vector
  - `matrix()` takes a vector and coerces a matrix
  - `data.frame()` takes x and coerces a dataframe

Week 5 - Connecting using different data sources

- Working with NAs
  - remove records containing NAs `dataframe[complete.cases(dataframe), ]` or `na.omit(dataframe)`
  - replace with mean of column `airquality$Ozone[is.na(airquality$Ozone)] <- mean(airquality$Ozone, na.rm = TRUE)`
  - ignore in computation `mean(vector, na.rm=TRUE)`

Week 5 - Connecting using different data sources

- Data transformations
  - `TRIM()` - removing spaces " SUNDAY" vs. “SUNDAY” in SQL queries
    ```
    sun_acc <- sqldf("select count(DAY_OF_WEEK) from
    df where TRIM(DAY_OF_WEEK) = 'SUNDAY'")
    ```
  - `gsub()` - replacing characters `gsub(" ", "", x)`

Week 6 - Introduction to visualization

Week 6 - Introduction to visualization

Book (CH 12) and asynchronous topics covered

- Creating visualization with `ggplot()`. See: https://ggplot2.tidyverse.org/
- Principle components of ggplot - data, aesthetics, geometry
  - data are the data you're working with, aesthetics are the x and y variables they also control color, the size or the shape of points, the height of bars , and geometry is the type of graph.

Week 6 - Introduction to visualization

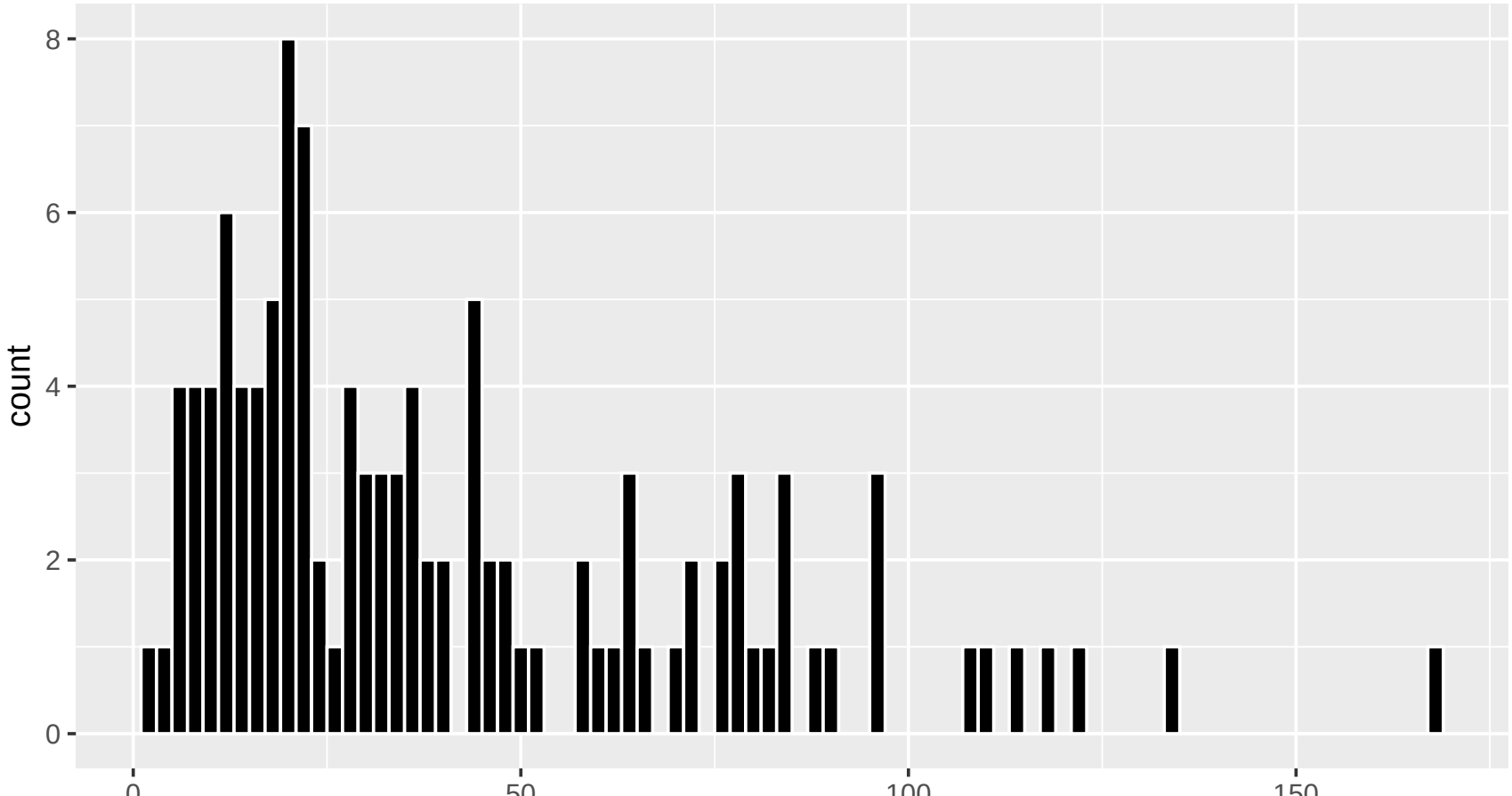
```
##   Ozone Solar.R Wind Temp Month Day
## 1    41     190  7.4   67     5   1
## 2    36     118  8.0   72     5   2
## 3    12     149 12.6   74     5   3
```

If we had data on `airquality` and were asked to create a histogram of Ozone. Load `ggplot2` using `library(ggplot)`, use the geom_histogram function contained inside the ggplot2 package.

Week 6 - Introduction to visualization

```
library(ggplot2)
ggplot(airquality, #data
    aes(x=Ozone)) + #aesthetics
geom_histogram(color="white", fill="black") # geom and more
```

Note the "+" needs to be included for adding other layers



Breakouts - Lab 6 and Project Updates (60 minutes)

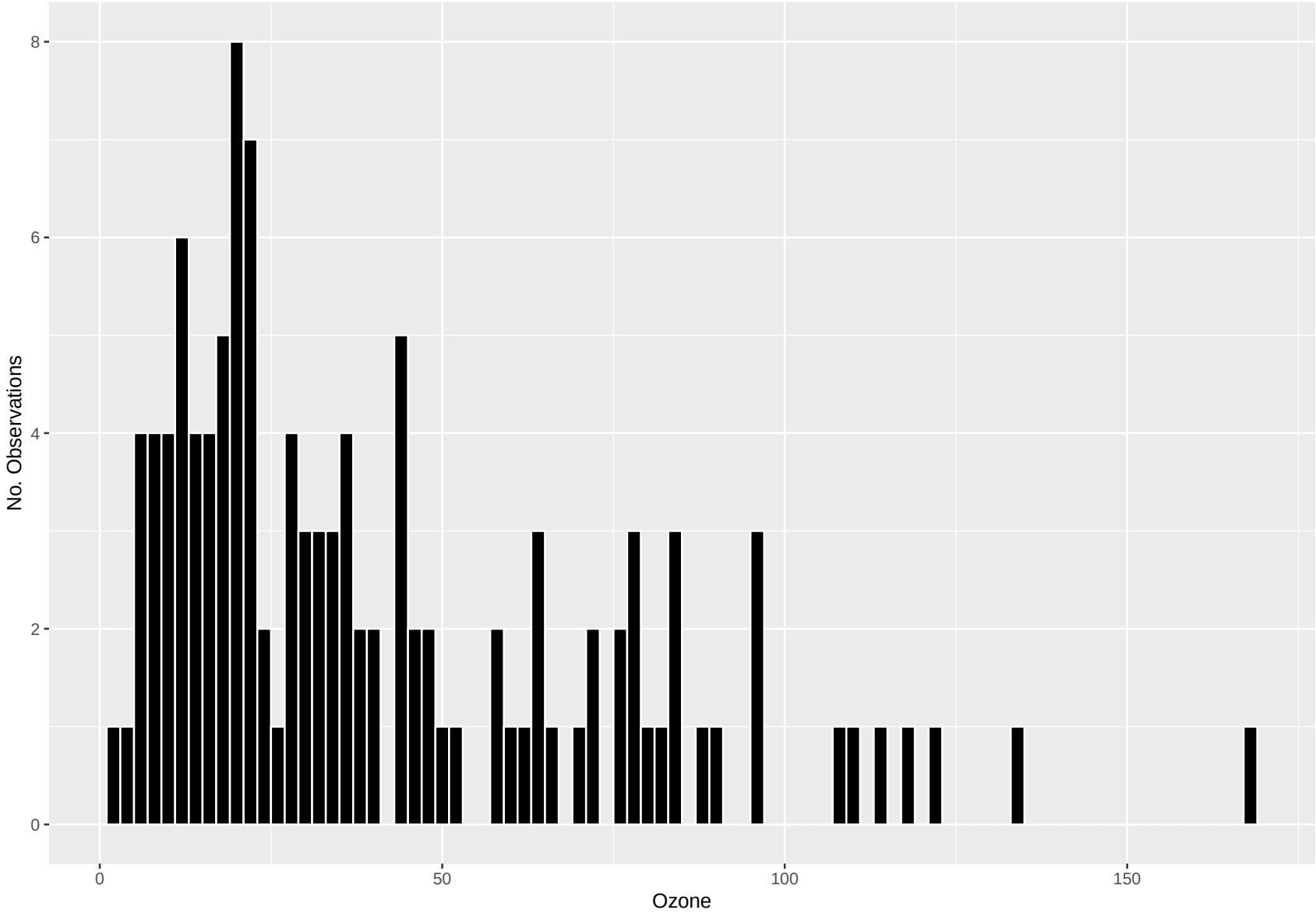
Breakouts - Lab 6

Steps for completing lab 6

- Install ggplot2: `install.packages("ggplot2")`
- Read through the assignment taking note of the required visualizations
- Check examples of example output on SLACK #lab channel as reference
- Work through creating visualizations of different geometry and learn about various aesthetics e.g., to change the labels you can add `xlab("Ozone")` and `ylab("No. Observations")`. Use: https://ggplot2.tidyverse.org/ to find the appropirate geoms and aesthetics

```
ggplot(airquality, aes(x=Ozone)) +
geom_histogram(color="white", fill="black") +
ylab("No. Observations") + xlab("Ozone")
```

Breakouts - Lab 6 (40 minutes)



Week 6 Homework tips

Week 6 Homework tips

- Converting between wide and long data formats (Step 3)
- Extracting and combining columns from an existing dataframes
- Working with dates in R

Week 6 Homework tips: Wide vs. long data formats

```
##   country year avgtemp
## 1  Sweden 1994      10
## 2 Denmark 1994       7
## 3  Norway 1994       8
## 4  Sweden 1995       6
## 5 Denmark 1995       7
## 6  Norway 1995       3
```

```
##   country avgtemp.1994 avgtemp.1995 avgtemp.1996
## 1  Sweden           10            6            8
## 2 Denmark            7            7            5
## 3  Norway            8            3           11
```

Week 6 Homework tips: Converting between long and wide data

Converting wide to long using `melt()` (in the `reshape2` package.)

```
##   country avgtemp.1994 avgtemp.1995 avgtemp.1996
## 1  Sweden           10            6            8
## 2 Denmark            7            7            5
## 3  Norway            8            3           11
```

```
country_longl <- melt(country_wide, id=c("country"))
```

```
##   country     variable value
## 1  Sweden avgtemp.1994    10
## 2 Denmark avgtemp.1994     7
## 3  Norway avgtemp.1994     8
## 4  Sweden avgtemp.1995     6
```

Week 6 Homework tips: Converting between long and wide data

Converting long to wide using `dcast()`

```
##    country year avgtemp
## 1   Sweden 1994      10
## 2 Denmark 1994       7
## 3   Norway 1994       8
## 4   Sweden 1995       6
## 5 Denmark 1995       7
```

```
country_widel <- dcast(country_long, country ~ year)
```

```
##    country 1994 1995 1996
## 1   Sweden   10    6    8
## 2 Denmark    7    7    5
## 3   Norway    8    3   11
```

Week 6 Homework tips: Extracting and combining columns from an existing dataframe

```
##   Ozone Solar.R Wind Temp Month Day
## 1    41     190  7.4   67     5   1
## 2    36     118  8.0   72     5   2
## 3    12     149 12.6   74     5   3
```

```
aq1 <- data.frame(airquality$Ozone, airquality$Solar.R)
```

```
##   airquality.Ozone airquality.Solar.R
## 1               41                190
## 2               36                118
## 3               12                149
```

Week 6 Homework tips: Working with dates in R

```
##   Month Day
## 1     5   1
## 2     5   2
## 3     5   3
```

We need to create a date that could be interpreted by R. We can use `paste()` to combine elements

```
sessiondates$Date <- paste(sessiondates$Month, +
sessiondates$Day, 2018, sep="/")
```

Week 6 Homework tips: Working with dates in R

```
##   Month Day      Date
## 1     5   1 5/1/2018
## 2     5   2 5/2/2018
```

... and then convert to an R date object

```
## 'data.frame':    153 obs. of  3 variables:
##  $ Month: int  5 5 5 5 5 5 5 5 5 5 ...
##  $ Day  : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Date : chr  "5/1/2018" "5/2/2018" "5/3/2018" "5/4/201
```

```
## NULL
```

Week 6 Homework tips: Working with dates in R

Convert the date character to an R readable date using `as.Date()`

```
sessiondates$Date <- as.Date(sessiondates$Date, +
"%m/%d/%Y")
```

```
## 'data.frame':    153 obs. of  3 variables:
##  $ Month: int  5 5 5 5 5 5 5 5 5 5 ...
##  $ Day  : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Date : Date, format: "2018-05-01" "2018-05-02" ...
```

```
## NULL
```

Next week

- Asynchronous Materials
  - Week 7: Working with map data
  - Submit HW/Lab 6 Monday
  - Review supplemental visualization links in syllabus
  - Bookmark resources for doing data science: Awesome R and Awesome Machine Learning
- Live Session
  - Lab 7
  - Exam logistics