STATISTICS PROJECTS USING RSTUDIO: GETTING STUDENTS ACTIVELY INVOLVED IN LEARNING

Marsha Davis
Eastern Connecticut State University
Mathematical Sciences Department
83 Windham Street, Willimantic, CT 06226
davisma@easternct.edu

Background

Eastern Connecticut State University is a public liberal arts university. Up until Fall 2018 we offered a solid, but "one-size-fits-all," Mathematics major. However, Fall 2018 we implemented a substantial revision to the Mathematics major that allowed students to better tailor their major to their goals upon graduation. We now offer distinctive B.A. and B.S. degrees in Mathematics, and students choose among the following concentrations:

- B.A in Mathematics (44 credits) with Concentrations in:
  - Mathematics for Teaching
  - Mathematical Structures and Applications
- B.S. in Mathematics (53 credits) with Concentrations in:
  - Actuarial Science
  - Data Science
  - Mathematical Structures and Applications

There is a common program core to the revised Mathematics major that includes an introductory statistics course, *Applied Probability and Statistics* (MAT 315), designed for students with strong mathematics backgrounds (calculus required!). While most of the students who enroll in MAT 315 are Mathematics majors, we consistently have a few Computer Science, Economics, and Biology majors each semester.

MAT 315 is a writing-intensive course in which students work on a data-based research project throughout the semester that culminates with a technical report. The dataset for this project comes from the *Monitoring the Future (MTF) Study*, which for the 2018 survey data involved over 14,000 participants. Students use RStudio to analyze the data relevant to their project topic.

The focus of this paper will be to twofold:

- To demonstrate how R and RStudio are introduced to students
- To provide examples of two projects used to support students' active learning

Jupyter Notebook: Introduction to R

Tools being developed to support collaboration among data scientists can also be an important addition to the statistics classroom. These tools can be an educational aid in

which students can develop programming skills in R while at the same time apply statistical techniques to tease out information from real data. One such tool is the Jupyter Notebook, an open-source, interactive notebook (developed specifically for data scientists) that contains both Markdown (text) cells and Code cells. While the name Jupyter is a loose acronym for Julia, Python, and R, Jupyter Notebook supports over 40 programming languages. At Eastern, Jupyter Notebook is installed in computer classrooms/labs with two kernels, R and Python 3. As shown in Figure 1, students choose the appropriate kernel from a pull-down menu.

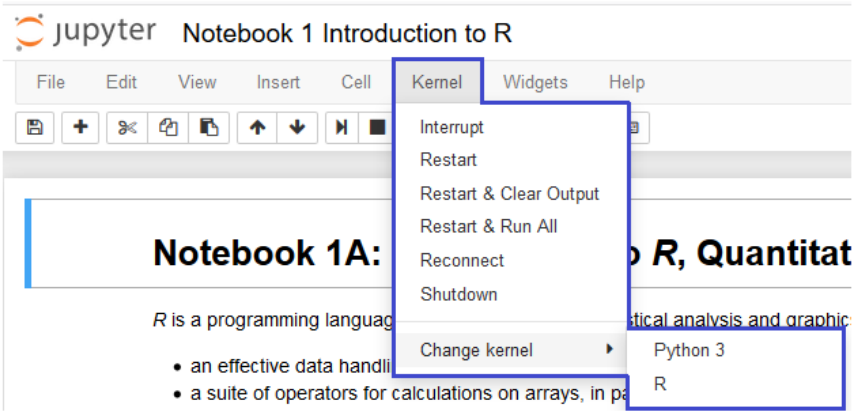


Figure 1. Selecting between Python and R.

This semester several students requested to take MAT 315 online and thus had to access Jupyter Notebook on their own computers. That can be accomplished by going to https://jupyter.org/. Students have two choices for accessing Jupyter Notebook: they can click "Try it in your browser" or "Install the Notebook." My students chose the "Try it in your browser" option, after which they had to select the programming language, in this case R. (See Figure 2.) Students then had to upload the assignment *Jupyter Notebook: Introduction to R* (File Type: IPYNB).

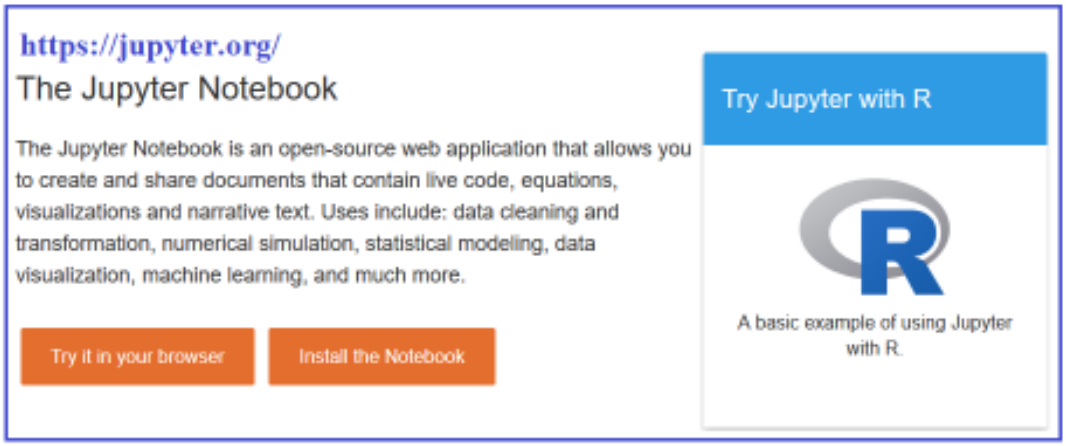


Figure 2. Jupyter Notebook on student computers.

Figure 3 shows a portion of *Jupyter Notebook: Introduction to R* that students complete as their first assignment using R. Notice that the Markdown cell contains instruction. The code cell serves as an assignment cell. The instructor, in a comments section (preceded by #), describes the problem to be solved. The student enters the answer in the same cell. Because it is a code cell, students can execute the cell as many times as they need in order to check their output and, if necessary, correct their code. For the sample code cell in Figure 3, students simply click Run to execute the code in the cell and observe what happens. Once executed, the output appears below the code cell. The instructor has great flexibility in creating an assignment cell. The instructor could provide a portion of the code needed for an analysis and have students complete the code, or provide the code needed to analyze one variable in a dataset and ask students to adapt the code to analyze another variable, or have students write the complete code to solve a problem.

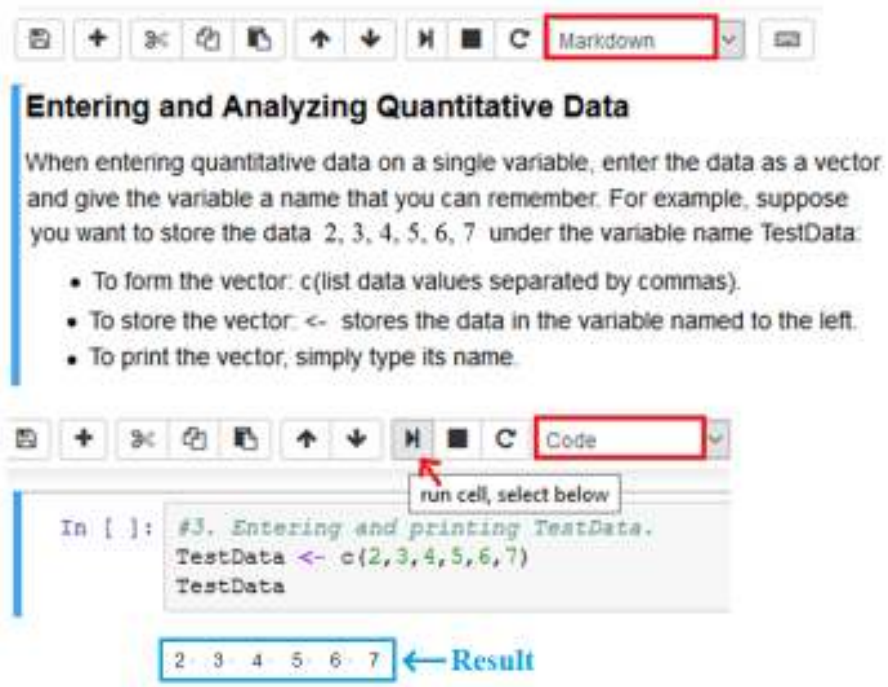


Figure 3. Sample Markdown cell and Code cell with output after execution.

On the first day of class, I involve students in gathering data that I use during the semester for a variety of activities. I start with a class survey. One of the questions on my survey asks students to estimate the outside temperature. Responses to that question were used in *Jupyter Notebook: Introduction to R*. The estimated temperatures were entered into a variable named Temp and then students used R's *stem* command to create a stem-and-leaf display. Figure 4 shows the code and the output. Even from this first assignment, I encourage students to critique and interpret any graphical displays that they create. In this case, the stem on the stemplot contains only even numbers (Figure 4(a)). Hence, 2|6

might represent the number 26 or 36 or 8|0 might represent 80 or 90. Students are asked to expand the stem (Figure 4(b)), which resolves this problem and also shows an extreme outlier. With outliers, students should be encouraged to ask if there is any explanation. In this case, a student was not from the U.S. and estimated the outside temperature in degrees Celsius. (I intentionally did not specify the units for the temperature in the survey question and have encountered this situation in several of the class survey datasets.) This gives students an opportunity to edit the *Temp* data, redraw the stem plot, and analyze the estimated temperatures.

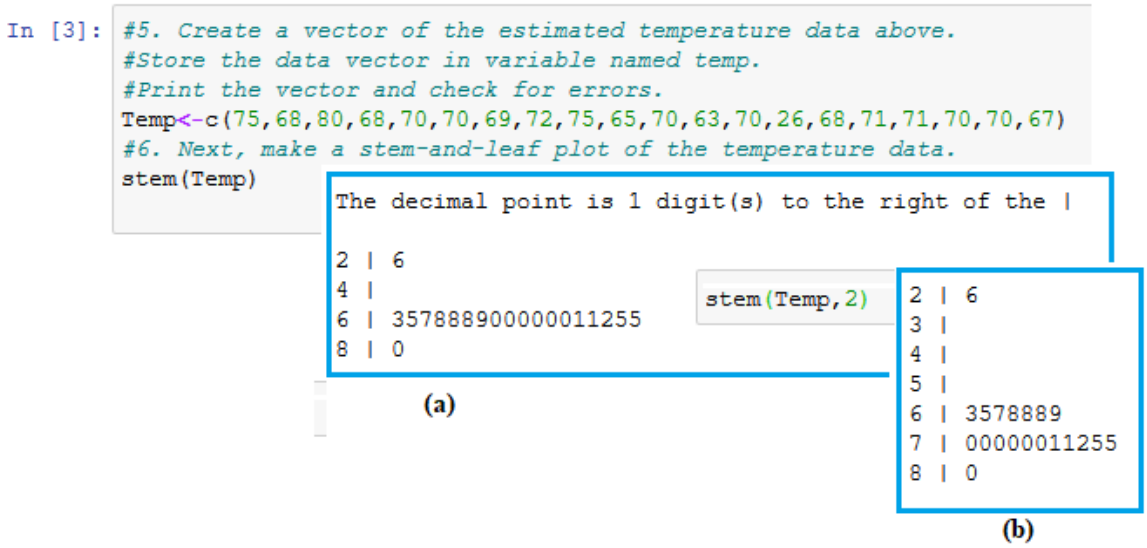


Figure 4. Analyzing temperature estimates.

Jupyter Notebook: Introduction to R covered arithmetic operations, built-in R functions, entering a data vector (both quantitative and categorical data), editing data values, creating stemplots, histograms, and boxplots (for quantitative data) and making tables, pie charts, and bar charts (for categorical data).

Sample Projects Using RStudio

After completing *Jupyter Notebook: Introduction to R*, students were given instructions for installing R and RStudio on their own computers. RStudio is an integrated development environment (IDE) for R, which helps students (as well as researchers) use R more effectively. (Given the current situation with COVID-19, students are required to install R and RStudio on their personal computers.) Here are the instructions students were given:

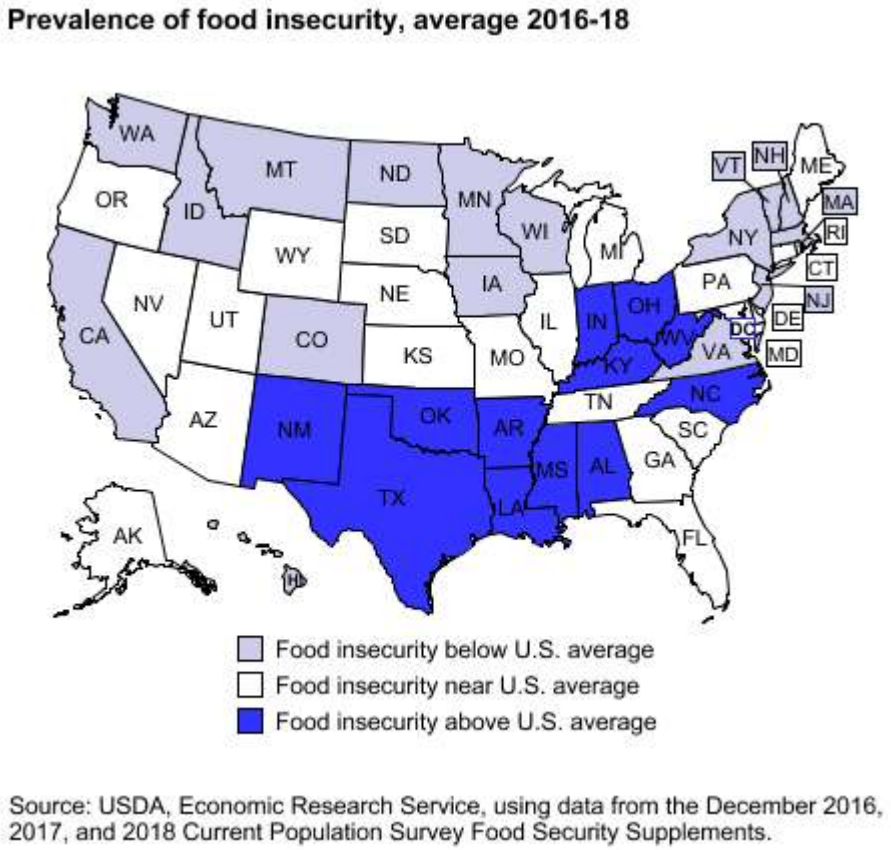
- Go to https://cran.r-project.org/ download for base R (for Linux, Mac, or Windows). Download and install R. Check that it is functional.
- Go to https://rstudio.com/products/rstudio/download/. You will be asked to choose your version. You want RStudio Desktop (Free!). Download the appropriate version for your computer and install.

- For anyone who might be struggling, view the YouTube video *Getting Started with R and RStudio*: https://www.youtube.com/watch?v=lVKMsaWju8w.

At this point students have some basic background in R and are ready to take on more substantial projects. The descriptions of two projects using RStudio are included in this paper: Food Insecurity and Monitoring the Future.

Food Insecurity

Food insecurity has been a sizable problem in many states across the country even before the COVID pandemic struck. On many campuses, including Eastern, a segment of the student population has experienced food insecurity. (Eastern has established a food bank and has instituted a "pay it forward" program where patrons of the Student Center can pay for an extra meal when purchasing a meal for themselves.) This project is based on U.S. Census Bureau data extracted from the Current Population Survey, March Supplement, Food Security Supplement. At the time this project was created, data were extracted using the DataFerrett tool, which was decommissioned in June 2020. In an effort to make data more accessible, you should now go to https://data.census.gov/ or to https://www.census.gov/data/api.html, the U.S. Census API. Figure 5 shows the extent of this problem across the United States.



Source: USDA, Economic Research Service, using data from the December 2016, 2017, and 2018 Current Population Survey Food Security Supplements.

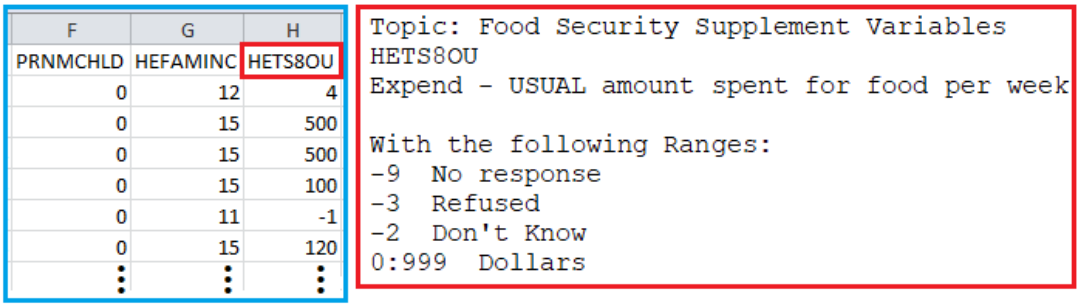
Figure 5. Food insecurity in the United States.

This project covered the following topics: creating a data frame, importing data, missing values, and normal quantile plots. In addition, students got to practice using commands that they learned from the *Jupyter Notebook: Introduction to R*, but this time in the new environment of RStudio. The dataset consists of 13 variables listed in Table 1 from the Food Security Survey.

Table 1. Variables extracted for use in Food Security Lab.

Variable	Description
PESEX	Sex
PRTAGE	Age
PEMLR	Employment status
GEREG	Geographic region
PRHRUSL	Hours usually worked weekly
PRNMCHLD	Number of own children < 18
HEFAMINC	Household total family income in past 12 months
HETS9OU	Usual amount spent on food per week
HESSM3	Food bought didn't last (in past 30 days)
HESSM4	Couldn't afford balanced meals (in past 30 days)
HESCM3	Frequency got food from food pantry (in past 30 days)
HESC4	Ate meals at soup kitchen (in past 12 months)
HRPOOR	Household income relative to 185% poverty

Students are given the raw data together with a codebook that provides details for each variable. An excerpt from the Food Security data with a codebook entry for one of the variables appears in Figure 6.



F	G	H
PRNMCHLD	HEFAMINC	HETS8OU
0	12	4
0	15	500
0	15	500
0	15	100
0	11	-1
0	15	120
⋮	⋮	⋮

```
Topic: Food Security Supplement Variables
HETS8OU
Expend - USUAL amount spent for food per week

With the following Ranges:
-9  No response
-3  Refused
-2  Don't Know
0:999  Dollars
```

Figure 6. Food Security Data Excerpts.

Figure 7 shows a screenshot of RStudio during a statistical analysis session. At this point in the analysis, the data have been imported. We see from the Environment panel (upper right) that the FoodSurvey data frame contains 13 variables and 126,065 cases. After labeling the numeric outcomes for geographic region (GEREG), 1 = Northeast, 2 = Midwest, 3 = South, and 4 = West, students write code for creating a frequency table, transforming the table into percentages, and representing the table with a pie chart. These

commands can be run individually or highlighted and run as a group (similar to running a code cell in Jupyter Notebook).

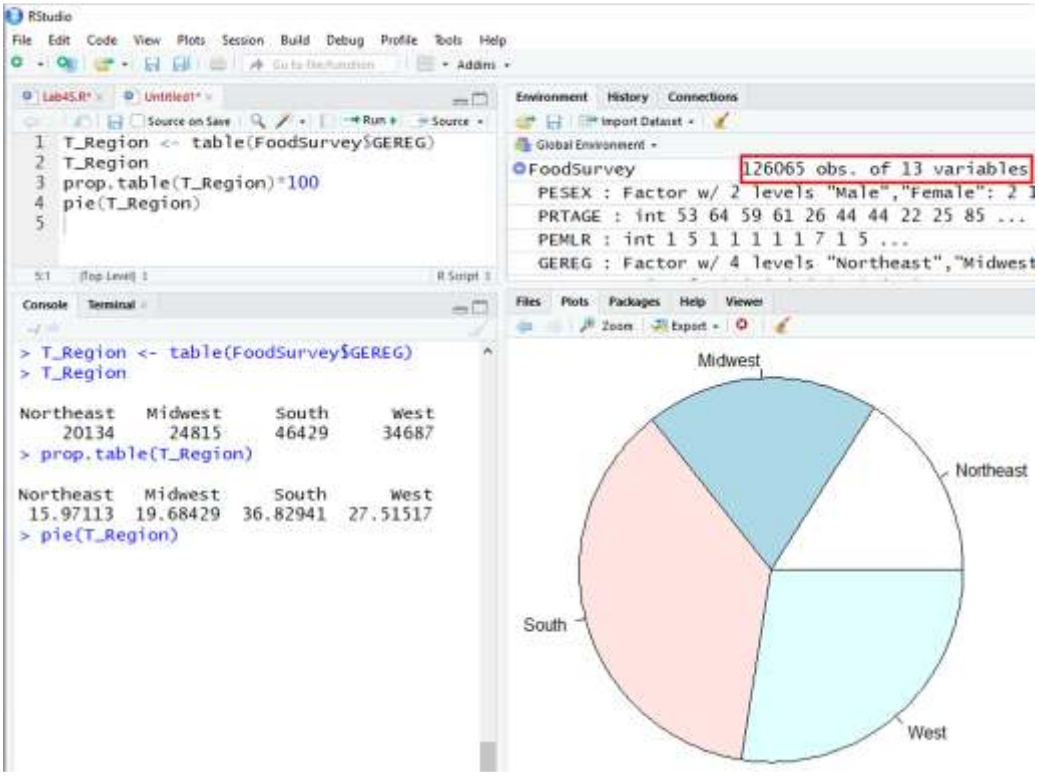


Figure 7. RStudio showing analysis from FoodSurvey data.

Again, the goal of this assignment is not simply to use R to make a table and a pie chart, but to analyze the results. One question that students should ask is why would the sample consist of around 16% of participants from the Northeast compared to more than double that percentage, around 37%, from the South? Wouldn't it make more sense to have the percentages roughly equal? To understand the reason for the unequal frequencies students need to dig a bit deeper. The data imported for Figure 7 were from 2017 (which was the most recent year for which data were available at the start of this student project). Table 2 shows the estimated population for each of these regions for 2017. Notice that the sample percentages for each region are relatively close to the population percentages.

Table 2. Estimated population and percentage by region for 2017.
From: www.census.gov>popclock>data_tables

Region	Population	Percentage
Northeast	56,059,240	17.2%
Midwest	68,126,781	21.0%
West	77,257,329	23.8%
South	123,542,189	38.0%

Next students analyze data from the quantitative variable, HETS8OU, the usual amount spent on food per week. (See Figure 6.) Students start by applying the summary command. The results appear in Figure 8.

```
summary(FoodSurvey$HETS8OU)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   -1.0    -1.0    80.0   103.4   160.0   500.0
```

Figure 8. Results from summary of HETS8OU

Again, students need to ask: Does this result make sense? Here the -1 is the code for a missing value and not a usual amount spent on food per week. In this course, this is the students' first encounter with missing values—something generally not encountered in textbook datasets. If students return to the codebook (see Figure 6), they will find three missing value codes: -9 = No response, -3 = Refused, -2 = Don't know, which have all been recoded as -1 for missing value. To indicate to R that -1 is a missing value, students run the following command:

FoodSurvey$HETS8OU[FoodSurvey$HETS8OU < 0] = NA

When they re-execute the summary command they get the results shown in Figure 9.

```
summary(FoodSurvey$HETS8OU)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  0.0    80.0   130.0   153.5   200.0   500.0   40870
```

Figure 9. Results from summary of HETS8OU (after declaring -1 = NA)

Going back to the original size of the sample, we find that 85,915 of the survey participants answered this question, while 40,870 did not. In examining this summary, students should notice a sizable difference between the mean and median. To get a sense of why this difference exists, we need to bring in some graphic displays. Figure 10 shows a histogram, boxplot and normal quantile plot of these data. From the histogram, the shape of the data is right skewed, with most of the data falling between $0 and $200. Based on the boxplot, there are outliers above $400 spent on food per week. The normal quantile plot clearly indicates that it is not reasonable to assume that these data were drawn from a normal distribution. (Sometimes students will identify any data that appears to have a mount shape as normal data without paying attention to whether or not the data are roughly symmetric.

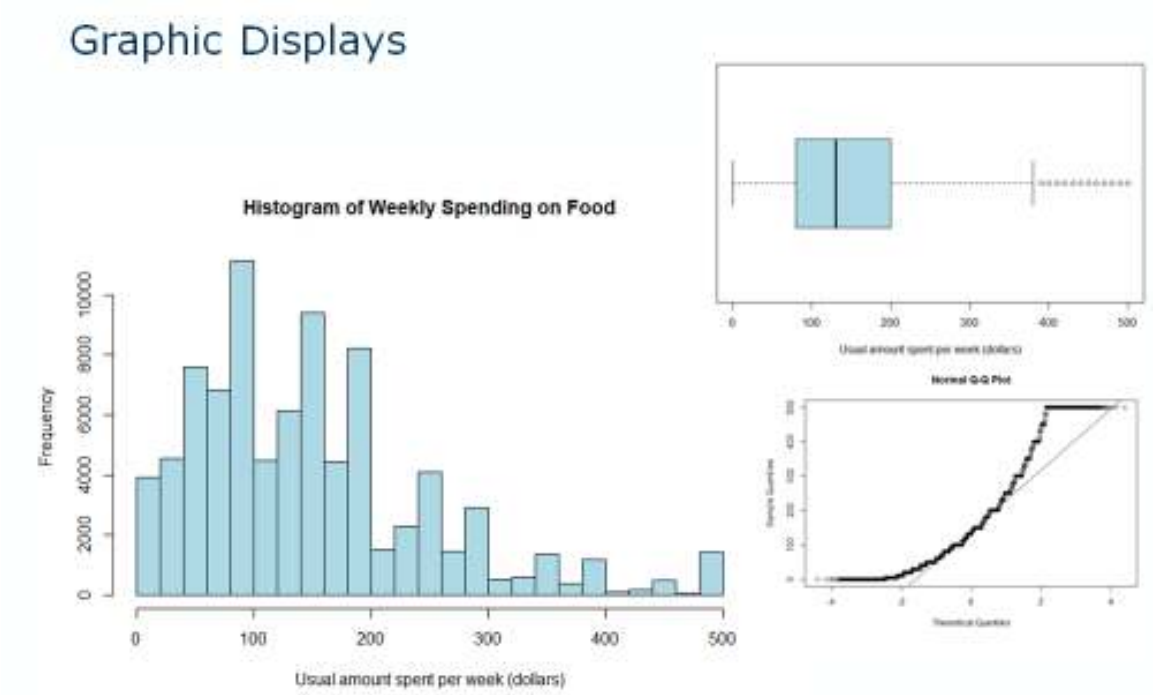


Figure 10. Graphic displays of data from HETS8OU.

Sometimes you get a better understanding of the distribution of a variable by tapping into its relationship with another variable. In the case of the usual amount spent on food per week, it might be a good idea to examine this variable's (HETS8OU) relationship with the number of children under 18 (PRNMCHLD). We can apply the summary command to HETS8OU broken down by number of children under 18. Here is the basic command structure when there are no children under 18:

summary(FoodSurvey$HETS8OU[FoodSurvey$PRNMCHLD == 0])

Now change 0 in the command above to 1, 2, 3, 4, 5, and 6 to complete Table 3.

Table 3. Mean and median amount spent weekly on food by number of children.

Number Children < 18	Median	Mean
0	130	147.0
1	150	165.1
2	160	184.4
3	180	195.2
4	200	211.5
5	200	222.3
6	180	193.9

From Table 3 we observe that as the number of children increases, both the mean and median amount spent weekly on food increases, until you get to a certain point; at five children the median stops increasing and at six children the mean decreases. The graphic display in Figure 11 shows comparative boxplots of the weekly amount spent on food broken down by the number of children under 18. Notice that the boxplots start to fall apart after about six children. That is because so few families have seven or more children. So, students need to think about how to best display these data. Should they cut off the display at the six children mark? Should they combine into one category seven or more children. These questions get at the "art of statistics." What graphic display most clearly tells the story of these data? That's a question for students to decide.

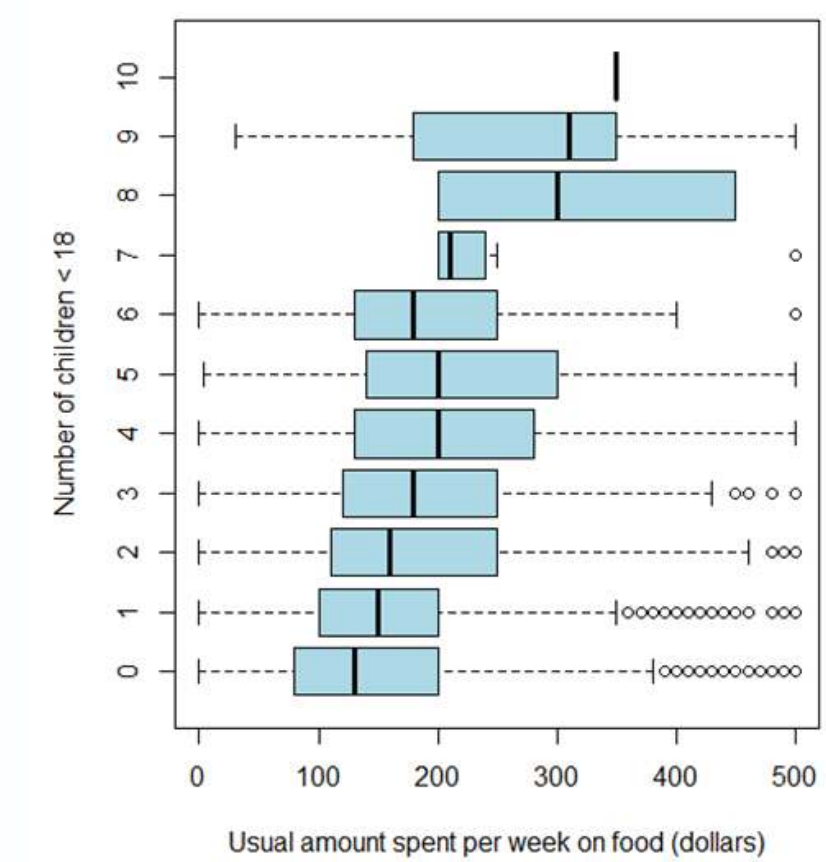


Figure 11. Comparative boxplots for weekly amount spent on food.

Up to this point, we have not yet discussed questions related to food insecurity. Next, we work with data on the variable HESSM3, which contains responses to the prompt: Food didn't last (in the past 30 days) and PESEX. What we are particularly interested in is whether the response distributions for food lasting differ for males and females. Output from the *table* and *prop.table* commands appears in Table 4. A chi-square test indicates a significant association between these two variables ($\chi^2 = 9$, p =0.003). From Table 4(b) notice that a higher percentage of females (52.13%) responded yes to running out of food compared to males (49.30%). **If students are paying attention to these percentages**

alarm bells should be going off! Do you really think that half of the country has run out of food in the past 30 days (even in times of COVID this seems really high)? More investigation is needed to understand this situation. From the frequency table (Table 4(a)), we find that only 11,520 answered both of the survey questions for HESSM3 and PESEX. Yet, there were 126,065 participants in the survey.

Table 4. (a) Cross-tabulation table of HESSM3 and PESEX; (b) Row percentages

Frequency

	Yes	No
Male	2617	2688
Female	3240	2975

(a)

Row percent

	Yes	No
Male	49.30	50.70
Female	52.13	47.87

(b)

On further investigation (found from the Department of Agriculture!), not all participants were asked the questions about food insecurity. Participants were screened for this question based on their family income. The variable HRPOOR classified participants as above or below the 185% poverty line. Only the participants who were classified as below 185% of the poverty line were asked the questions related to food insecurity.

Table 5. Table of HRPOOR
(participants above or below 185% of the poverty line).

Missing	Below	Above
35,982	26,265	62,825

Monitoring the Future (MTF) Project

A sizable percentage of Eastern's Mathematics majors plan on careers as secondary school teachers. The Monitoring the Future (MTF) Project gives these Mathematics majors insights into their future students' attitudes, behaviors, and values. In addition, the MTF project provides all students in the class with an opportunity to apply what they have learned about analyzing categorical data to a research project that culminates with a technical report. The data are from *Monitoring the Future: A Continuing Study of American Youth* which has administered annual surveys since 1975. (For more information see http://monitoringthefuture.org/.) MTF data are released to the public about a year and a half after they have been collected. The sample data used for this paper are from the 12th-grade survey from 2016, which can be accessed at https://www.icpsr.umich.edu/icpsrweb/NAHDAP/studies/36798. Students were given the Core Data in its Comma Delimited form (saved as a .csv file) and the corresponding Codebook for the 2016 12th-grade survey. Figure 12 provides a partial view of what students saw when they opened the file MTF2016.csv containing the Core Data from the 2016 12th-grade survey.

	A	B	C	D	E	F	G	H	I	J
1	CASEID	V1	V3	RESPONDI	ARCHIVE_	V13	V16	V17	V2101	V2102
2	1	2016	1	10001	1.30696	3	0	0	3	1
3	2	2016	1	10002	1.453475	3	0	0	1	1
4	3	2016	1	10003	1.399638	3	0	0	2	1
5	4	2016	1	10004	1.416772	3	0	0	1	1
6	5	2016	1	10005	1.515468	3	0	0	4	3
7	6	2016	1	10006	1.416125	3	0	0	1	1

Figure 12. Partial view of MTF2016 data.

Using the data in this format gives students experience naming variables and coding data. Once imported into R, students discovered that this dataset consisted of 12,600 cases and 175 variables. The variables in the Core Data can be classified into the following categories:

- Survey/Dataset Information
- Geographic
- Demographic/Respondent Characteristics
- Attitudes and Beliefs
- Aspirations
- Employment/Income
- Recreation
- Driving/Substance Abuse
- Substance Use

For this project, students worked in teams to select a topic, extract the data relevant to their chosen topic, analyze the data, and report their findings in a technical report. Students used the Codebook to find the survey questions and the labels that were associate with the coded numeric responses.

As an example, consider the following two survey questions: (1) What is your sex? and (2) How intelligent do you think you are compared with others your age? From the 2016 Codebook, we find these questions correspond to variables V2150 and V2174. The question, variable name, coded responses, and frequencies/percentages are shown in Figures 12 and 13.

From Figure 13, we note that the one-variable analysis on V2150, Rs Sex, indicates that respondents to this question were fairly evenly split between males and females; however, 9.9% of participants choose not to identify their sex. So, again students have to deal with missing values (this time coded as -9) in their analyses of MTF data.

V2150: 166C03 :Rs SEX

Item Number: 00030

What is your sex?

1="Male" 2="Female"

Value	Label	Unweighted Frequency	%
1	MALE:(1)	5701	45.2 %
2	FEMALE:(2)	5646	44.8 %
	Missing Data		
-9	MISSING:(-9)	1253	9.9 %
	Total	**12,600**	**100%**

Based upon 11,347 valid cases out of 12,600 total cases.

Figure 13. Codebook information on variable V2150.

From Figure 14, we note that a majority of participants rated their intelligence in the above-average categories.

V2174: 166C17 :RT SF INTELL>AVG

Item Number: 00420

How intelligent do you think you are compared with others your age?

1="Far Below Average" 2="Below Average" 3="Slightly Below Average" 4="Average" 5="Slightly Above Average" 6="Above Average" 7="Far Above Average"

Value	Label	Unweighted Frequency	%
1	FAR BELOW:(1)	197	1.6 %
2	BELOW AVG:(2)	203	1.6 %
3	SLIGHT BELOW:(3)	519	4.1 %
4	AVERAGE:(4)	3456	27.4 %
5	SLIGHT ABOVE:(5)	2980	23.7 %
6	ABOVE AVG:(6)	3182	25.3 %
7	FAR ABOVE:(7)	1038	8.2 %
	Missing Data		
-9	MISSING:(-9)	1025	8.1 %
	Total	**12,600**	**100%**

Based upon 11,575 valid cases out of 12,600 total cases.

Figure 14. Codebook information on variable V2174.

After analyzing Sex and Intelligence Rating (Intel) separately, the natural question to ask is whether there is an association between the two variables. After labeling the numeric data values (and combining the seven Intelligence Rating categories into three categories: Below, Average, and Above), the following code was used to create the two-way frequency table shown in Table 6:

```
Table_Sex_Intel <- table(Sex, Intel)
Table_Sex_Intel_Margins <- addmargins(Table_Sex_Intel)
Table_Sex_Intel_Margins
```

Table 6. Two-way frequency table of Sex and Intel.

	Intel			
Sex	Below	Average	Above	Sum
Male	346	1390	3773	5509
Female	489	1865	3128	5482
Sum	835	3255	6901	10991

A chi-square test ($\chi^2 = 200$, df = 2, p $\approx$ 0.000) indicates a significant relationship between Sex and Intel. To further explore the relationship between these two variables, Figure 15 shows a bar chart of the conditional distributions of Intelligence Rating for males and for females.

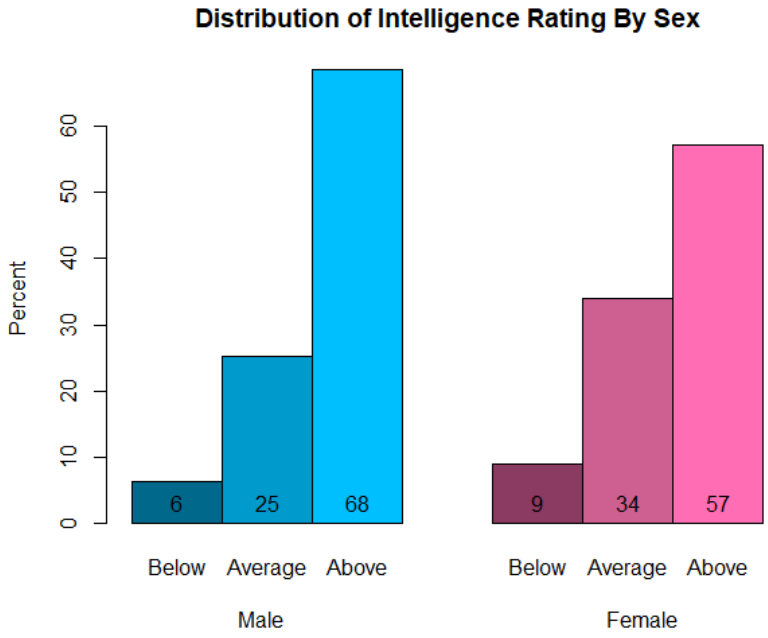


Figure 15. Conditional distribution of Intel for each Sex.

From the bar chart, we discover that a majority of both male and female students rated their intelligence as above average. However, the percentage of males who rated their intelligence as above average exceeded the percentage for females by 11%. The percentage of females who rated their intelligence as average exceeded the percentage for males by 9%.

The sample analysis for the variables Sex and Intelligence Rating gives some flavor of the types of analyses that students completed for their technical reports. While not included in this discussion, students were also required to create a times series for one of their variables to track how the percentages had changed over time.

Conclusion

Using Jupyter notebooks proved to be an efficient way to introduce R. Students learned by reading instructions in the Markdown cells followed by running code, adapting code, and writing code in the code cells. After completing the Notebook, students had opportunities through the projects to become researchers using their R skills to extract information from data and report their findings. My students reported that they understood the statistical techniques better when they got to apply them through projects.

There are many sources for data. Here are a few examples:

- Government data
- Monitoring the Future data
- Sports data (my students love sports data):
    - NBA: www.basketball-reference.com
    - MLB: www.baseball-reference.com
    - NHL: www.hockey-reference.com
    - Soccer: https://fbref.com/en
- Kaggle (Data Science Community): www.kaggle.com