Fall 2021

Stat 515
Final Project

Guidance for Tasks 1a and 1b, Oral Presentation, and Written Report

1. Objectives

The final project has several objectives:

- To use data visualization software taught in class to support data analysis in one or more of the following ways:
    - Exploratory data analysis, in which you create graphs or maps indicating possible relationships among variables and then perform additional analysis using methods taught in class;
    - Visual data checks, in which you use data visualizations to examine data quality or to check successful completion of data processing procedures, such as displays of missing-data patterns and graphs of recoded variables and calculated summary statistics;
    - Visual data-analysis methods, in which a data analysis method requires the examination of a data visualization, such as the examination of residual plots to determine if model assumptions are satisfied; and
    - Visualization of analysis results, in which you create graphs /or maps to display the results of data analysis methods taught in class
- To gain experience in developing R scripts that correctly perform multiple data analysis methods taught in class,
- To gain experience in communicating about data visualizations, data analysis procedures, and data analysis results

2. Project components

The final project contains the following components:

2.1. Data set. You are responsible for obtaining a data set for your project. Appendix A lists possible sources, including both primary and secondary sources. Some secondary sources include additional information about their data sets, such as descriptions of variables, processing for cleaning the data, and successful analyses of the data sets. For some data sets, readily available R code or Tableau files exist that visualize the data or analyze the data set using an analysis method listed in Section 2.5. You cannot use such data sets for the STAT 515 final project to create the same visualizations or perform the same data analyses present in the associated R code or Tableau files[1].

[1] For example, for a Kaggle data set, check the "Code" and "Discussions" posts. If any posts contain R code for visualization or data analysis using methods in Section 2.5, that may rule out your being able to use the data set.

The data set should contain a minimum of six variables, and if the data set were to be reshaped into "long format," the result would have a minimum of 100 records. The minimum of six variables, however, can contain meaningful "recodes" of existing variables. For example, if an agricultural-crops data set included variables for

1. *Year* (30 levels),
2. *State* (5 levels – Illinois, Indiana, Iowa, Minnesota, & Nebraska),
3. *Crops* (2 levels – corn or soybeans),
4. *Acreage,* and
5. *Yield*,

then adding *Production = Acreage* x *Yield* can be counted as an additional variable toward the minimum of six variables. If this crops data set were to be reshaped into "long format," it would contain

(# years) x (# states) x (# crops) = 30 x 5 x 2 = 300 records,

so it satisfies the requirement of containing a minimum of 100 long-format records.

The project data set cannot contain classified or proprietary data and cannot be a data set analyzed in any of the instructional materials for this course. (Appendix B lists data sets analyzed in STAT 515 course materials for weeks 10 through 12.) If in order to use a data set, you must agree to <u>not</u> share the data or associated analysis results or associated processing code, then you cannot use this data set for the final project. The size of the data set should not exceed 1 gigabyte.

2.2. <u>Additional data preparation</u> is optional. Because of the time-consuming nature of data preparation, you should attempt to find a data set that requires a minimum amount of data preparation. If it appears the amount of time needed to prepare an initial .csv file will exceed three hours, you should contact the instructor to discuss project options.

2.3. <u>Formulation of research question(s)</u>. It is a requirement that you formulate at least one research question that subsequent data analyses will attempt to answer. One approach is to first use data visualization software taught in class to perform exploratory data analysis to identify possible relationships among variables in the data set. Following this, you formulate one or more research questions, which will drive your subsequent data analyses.

For example, an exploratory data analysis of tipping rates for taxi rides may indicate higher rates for short trips, certain times of day, and particular zip codes. Some possible research questions that you could formulate are the following:

- Can you develop a model that contains a small number of predictors that reliably predicts tipping rates? What are the must-have predictors in this model?
- Can a model be developed to predict "stiffing"—that is, when the driver receives no tip? What are the needed predictors? Are they the same or different from those for the tipping rates model?
- How accurate are these models? Can taxi drivers use these models to decide when is the best time to be on duty (even though such models will not have any predictors about trip lengths, origins, and destinations)?

Data analysis converts data into information. Research questions drive data analysis and determine the quality of the resulting information. If research questions are narrow in scope, they may not produce useful information. For example, in 2008, Google developed a model based on search data to predict seasonal flu. The research question was, "Can Google searches be used to predict seasonal flu?" The model worked well until 2013, when it failed due to a change in the data processing. The research question "Can Google searches be used to predict seasonal flu?" was too narrow in scope because it did not provide information about the model's key inputs. Had this information been available, it might have been possible to avoid the model's failure in 2013.

Note in the taxi-tipping example above, there are several research questions. They ask if it is possible to predict tipping and stiffing rates. They also ask about key predictor variables and prediction accuracy.

2.4. Data visualization. A significant part of the final project is to use software taught in class to create data visualizations to support data analysis. (See the first bullet point in Section 1.) As mentioned in Section 2.3, you can use data visualizations for exploratory data analysis before identifying one or more research questions. However, performing data visualization before you formulate the research question(s) is not a requirement. If you do not use data visualization for exploratory data analysis, you should use the data visualization software taught in class to support data analysis in other ways. For example: visual data checks, visual data-analysis methods, or displaying data-analysis results.

2.5. Data analysis. Following the formulation of one or more research questions, you will develop R scripts that use two or more of the following analysis methods to answer the research question(s):

- Linear regression (including best-subset and stepwise selection methods),
- Logistic regression,
- Poisson regression,
- Lasso regression (linear, logistic, or Poisson),
- Classification tree,
- Regression tree,
- Random forest (for a binary response variable or for a continuous response variable (a.k.a., regression forest)),
- Clustering

Use the R packages taught in class to perform data analysis. No credit will be awarded for using R packages not taught in class or data analysis methods not included in the above list. Analyzing a subset of the variables on the data set is permitted, but for lasso regression and random forests, you must analyze a minimum of six variables.

2.6. Tasks 1a and 1b. See Section 4 below.

2.7. Oral presentation in class. See Section 5.

2.8. Written paper. See Section 6.

3. Teaming option

Your work using data visualization software, analyzing data, and preparing a written paper are individual efforts. You are permitted to team with one or two other individuals in the class to select and prepare a data set, formulate research questions, divide up the preparation of an oral presentation, and provide cognitive-testing feedback about created data visualizations. (You can indicate to others in the class your interest in teaming by posting to the "Final-project in-search-of team member(s)" discussion board.)

If you team with one or two other individuals, each person on the team prepares different data visualizations. Also, each person on the team performs data analysis—either for different research questions or by using different analysis methods (each person still has to use two or more of the analysis methods listed in Section 2.5.)

If you team with one or two other individuals, the written paper you submit describes the specific visualizations you created and analyses you performed, any associated challenges, and what you believe are the conclusions from your work. Collaboration is permitted in preparing the section of the written paper describing the data set and any additional preparation work. This section would appear in the written paper submitted by each person on the team.

4. Tasks 1a (Stake a claim) and Task 1b (data set description and research questions)

You are responsible for finding a data set for use in the final project. Different individuals (or different teams of size 2 or 3) cannot use the same data set. Post to the Blackboard discussion board "Final-project stake a claim" a citation to your selected data set. Once one individual (or a team of size 2 or 3) has staked a claim on a data set, another individual (or team of size 2 or 3) cannot select it. You will receive one grading point if you stake your claim by the due date indicated in Section 8.

Task 1b consists of your submitting to Blackboard a Word/PDF file describing the selected final-project data set (source, size, unit(s) of analysis, and types of variables) and the associated (potential) research question(s). The length of this description should be 250 words or less. Table 1, below, contains some examples of types of data and associated research questions analyzed by STAT 515 students in the past.

Unfortunately, to answer a particular research question, it is possible to select a data set that is not well suited for performing an analysis that answers the research question. For example, assume your research question was about what variables predict the price of a new home. A data set about new home sales from the U.S. Census Bureau would not be the best data set to analyze. Census data contains information about the prices of new homes, plus geographical and housing characteristics variables. However, Census data does not include information about interest rates, which significantly influence the demand for new homes and thus the price of new homes.

The purpose of Task 1b is to provide you early in your project work the opportunity to have the instructor review the characteristics of your selected data set and associated research questions. If necessary, the instructor will raise concerns about whether the data set or research questions may need to be modified to provide you with a satisfactory project experience. In addition to the

review provided by Task 1b, your using email to ask the instructor about the suitability of a data set is another way to avoid selecting a data set that may not be well suited for use in your final project.

Logistics: For **Task 1a**, you should enter information about your selected data set in the "Final-project Stake-a-claim" discussion board by the due date specified in Section 8. For **Task 1b**, you should submit to Blackboard a Word or PDF document containing the Task 1b information by the due date specified in Section 8. There will be a Task 1b assignment item for submitting Task 1b. For grading purposes, each person on teams of size 2 or 3 should post to the Stake-a-claim discussion board and use the Task 1b submission item, providing the required information and also identifying the other individuals on the team.

Table 1. Examples of previous-semester final projects

Data	Research questions
Ratings of coffee beans from different suppliers	Can a model be built that predicts coffee bean ratings? What are the key variables? Was there a bad crop year resulting in lower quality coffee beans?
Arrests for criminal activity in New York City	Can a model using geographical and demographic variables reliably predict criminal arrests? Can this model be modified to include temporal variables? How accurate are these models?
Reviews of different brands of beer	What is the effect of alcohol by volume (abv) on the overall review? Do people always like beers with a higher abv? Are beers that taste good more popular?
Border crossings into the United States	Is there a seasonal trend for border crossings in personal vehicles that predicts traffic volume at particular border crossings in Maine? For crossings at the U.S.-Canadian border, is there a relationship between the volume of pedestrian crossings and the number of crossings in buses, trucks, and personal vehicles?
Masters Tournament golf scores	Does a player's score tend to rise or fall throughout a tournament? How have overall scores changed over the years? How have low scores (i.e., 1st place, 2nd place, and 3rd place) changed over the years?
Pet licenses in Seattle, plus supplementary data from the U.S. Census Bureau	Can zip-code level data about population size, population demographics, and the number of pet-related businesses be used to predict zip codes where pets are more likely to be found? Is it possible to predict the most popular pet species (e.g., dog, cat, pig, horse, etc.) using zip-code level demographic data and other descriptive variables?
Men's and women's tennis tournaments	Can a model that contains a small number of variables reliably predict the winner of a tennis match? If the model does not contain variables about the opponent? How accurate are these models?

5. The oral presentation in class

The purpose of the oral presentation in class is to present information about the project data set, research question(s), data visualization outcomes, data analysis approach, and (initial) results. Use the cognitive principles and guidelines discussed in class to design the data visualizations. (See "Section 3: Cognitive-based principles and guidelines for designing data visualizations" in the mid-term project document.)

The maximum length of each project presentation will be 15 minutes. The required minimum length for a presentation will be as follows:

- If working alone, you need to present at least 5 minutes,
- Teams of size 2 need to present at least 7 minutes, and
- Teams of size 3 need to present at least 9 minutes.

Because other class members will view your presentation, it is permitted to request cognitive-testing feedback about data visualizations in your presentation, similar to the procedure used for the midterm project. It is also permitted for you to respond to requests for cognitive-testing feedback from other class members. Neither requests for feedback or providing feedback will be graded for the final project, however. Also, restrict your requests for feedback to cognitive testing—that is, questions about the reviewer's visual and mental processing when viewing a data visualization. Do not post or respond to questions such as "What is a better way to do this?" or "How should I analyze my data?"

For teams of size two, each team member must speak at least 40 percent of the total time; and for teams of size three, each team member must speak at least 30 percent of the time.

Normally the oral presentation will receive the total number of grading points allowed (see Section 7). However. if you do not deliver an oral presentation or its length is less than the specified minimum time or longer than 15 minutes, if it is in bad taste, if it blatantly goes against class guidelines, or does not discuss any completed data visualization or data analysis work, grading points (perhaps all) will be deducted from the total number of possible points

The following is a suggested outline for the oral presentation:

- Title slide, including your name(s)
- Data set
- Source / description / context / background
- Data tidying and transformations (if any)
- Exploratory visualizations and research questions
- Data analysis and associated data visualizations
- (Initial) results
- Conclusions / challenges / further analysis

Logistics: Post your presentation slides by the date and time shown in Section 8. The title of your discussion board thread should be a short phrase (maximum of five words) describing your project.

If you are a team member of size two or three, then for grading purposes, each team member should create a thread containing the presentation slides and have the same thread title.

6. The written paper

The length of the written paper is not to exceed nine pages (single-spaced, one-inch margins, 10-point font, or larger), though it can be shorter than this if it contains all required materials and discussions. Going over the target maximum size will result in losing points. The paper can include appendixes that do not count against the nine-page maximum. The paper may discuss only a subset of a large complex data set. However, focusing on a subset of records or variables should not analyze only a trivial amount of data, such as analyzing the data for only one state in a data set containing data for all states in the U.S.

The written paper will be graded. The grading is based on quality and level-of-effort factors, described below. The written paper describes the data set, the research question(s), analyses that answer the research question(s), and associated data visualizations. All data visualizations in the main part of the paper should be presentation-ready (containing plot titles, meaningful axes titles, and appropriate context information) and should implement cognitive principles and guidelines taught in class. You can include draft versions of graphs in an appendix if they support the discussion of the data processing workflow, but this is not a requirement.

Except for data visualizations you create that have no relevance to your analysis, the visualizations you discuss in your paper should be included in your paper. It is acceptable to include a screenshot of an interactive visualization result, but the paper cannot discuss other interactive visualization results that are not included in the paper.

Each graph or map you include in your paper should have a figure number in the plot title or centered below the plot. Figure numbers should be consecutive, starting with 1. Multi-panel plots can have a single figure number or a figure number for each panel.

The paper discusses what data patterns, if any, you see in each data visualization. **There should be at least one comment made about every plot, map, table, or included model output.** The comment might be there is not much of a pattern.

While the oral presentation materials may have bulleted items, the written paper should consist of complete sentences and well-structured paragraphs. The quality of the writing should meet the requirements of academic writing for a proceedings or peer-reviewed articles. Every resource used needs to be both cited and referenced (use APA format[2]). Citations appear in the body of your paper and point the reader to a list of references appearing in the References section at the end of your paper. The following is an example of text containing a citation:

> "I used function coefplot() in R package coefplot (Lander, 2021) to plot the estimated coefficients and the associated 95 percent confidence intervals."

And, the following is the corresponding reference in the paper's References section:

> "Jared P. Lander (2021). coefplot: Plots Coefficients from Fitted Models. R package version 1.2.7. https://CRAN.R-project.org/package=coefplot "

[2] See here for details about APA-style reference lists

Each reference in the References section specifies the author(s), date, title, and source of the referenced material. For materials available on the web, a complete URL is provided so readers can paste it into their browser to access the referenced material.

Failing to cite and reference books, documents, R, R packages and codes will result in losing points. Copying pasted text (without quoting, citing, and referencing) is considered plagiarizing, which is using the exact words, opinions, or factual information from another person without giving the person credit[3]. You give credit through enclosing material in quotation marks, adding parenthetical citations, and providing references. For paraphrased material, you must also provide citations and references. A simple listing of books or articles is not sufficient. It is permissible for you to include text from documents you have prepared outside of STAT 515, but such text must be enclosed in quotes and appropriately cited and referenced.

The talk may mention any special efforts that were required. The paper can also briefly mention this and provide supporting evidence in appendices that do not count against the paper's page limit. The oral presentation should include at least one slide explaining any challenges, and the written paper should also contain at least one paragraph about any challenges.

If you team with one or two other individuals, the written paper you submit describes the visualization and analysis work you performed, any associated challenges, and what you believe are the conclusions from your work. Collaboration is permitted in preparing the section of the written paper describing the data set and any additional preparation work. This section would appear in the written paper submitted by each person on the team.

Once you have completed the first draft, check the spelling and grammar with available tools, such as those in Word or Grammarly. After completing the first draft, it is not an honor-code violation for this class to get help in revising the paper from the University Writing Center. It is also okay to have a native English speaker read the paper and comment.

The following is a suggested outline for the written report:

1. Project description
2. Data set
   - 2.1. Source / description / context / background
   - 2.2. Data inspection and preparation of analysis data set
3. Exploratory analysis and research question(s)
4. Data analysis
   - 4.1. Methods and software used
   - 4.2. Results
5. Conclusions / challenges / further analysis

References

Logistics: Upload a Word or PDF document containing your written paper to Blackboard by the due date indicated in Section 8. Also, attach associated data files, R scripts, Compile Report outputs, and Tableau .twbx files. The instructor needs to be able to replicate your visualizations and data analyses by reading in the attached data set and running the associated

[3] See here for ways to avoid plagiarism.

program(s). Though the attached program(s) create visualizations and perform data analyses, to receive credit, you must include the visualizations and discussions of the data analyses in the submitted Word document or PDF. If the paper discusses R code, the discussed code needs to be included in the Word document or PDF. Similarly, if the paper discusses complex Tableau specifications, one or more associated screenshots should be included in the Word document or PDF. For work done by teams, each team member should submit a written paper, data sets, and programs.

The final project's written paper is due on the last day of class, so a late submission will be handled by following University policies for not taking a course's final exam.

7. Grading

7.1. Task 1a --Stake a claim (1 point). See Section 4. You will receive one grading point if you stake your claim by the due date indicated in Section 8.

7.2. Task 1b –Data set description and (potential) research questions (3 points). See Section 4 for content and Section 8 for due dates. In 250 words or less, describe:

- Data source(s)
- Number of variables
- Types of variables with respect to information content; e.g., demographic data, economic data, health data, sports data, etc
- What variables will be response variables
- Number of records
- What is the unit of analysis; i.e., the entity associated with the individual records when in "tidy" format--states, counties, patients, soccer players, etc
- Research questions (If exploratory data analysis is still in progress, then potential research questions.)

Grading criteria include: submitted by due date, completeness, and quality of description. Use complete sentences instead of bullet points. You can use a table to describe the types of variables. Your score will be decreased for numerous misspelled words or grammatical errors.

7.3. Oral presentation in class (6 points). See Section 5. Normally, your oral presentation will receive the total number of grading points allowed (i.e., 6 points). However, if the oral presentation is submitted late or not submitted at all, if it is shorter than the length specified in Section 5 or longer than 15 minutes, if it is in bad taste, if it blatantly goes against class principles and guidelines, or does not discuss any completed data visualization or data analysis work, grading points (perhaps all) will be deducted.

7.4. Written report (60 points). See Section 6. Fifty percent of the score for the final project will be associated with data visualization, and the remaining fifty percent will be associated with data analysis. The grading criteria that will be used to obtain sub-scores for data visualization and data analysis are the following:

- **Level of effort is recognized in many areas:**

- Selection of non-trivial research question(s)
- Data gathering and preparation of data for analysis/graphics
- Preparation of a variety of data visualizations
- Attention-to-appearance details
- Attention to comparability issues
- Improvements resulting from using non-default settings
- Completeness of analyses
- Comparisons between multiple data analysis methods
- Consideration of alternative models
- Consistency across graphical elements of multiple graphs
- Thorough checking of data quality and correctness of calculations and text in titles, labels, and captions
- Providing all specified deliverables

➢ **Use of concepts and software taught in class**
- Data visualization guidelines
- Enable accurate comparisons
- Simplify appearance
- Provide or increase context to support interpretation or to facilitate hypothesis generation
- Attract and engage the reader/analyst
- Making data "tidy"
- Reshaping data
- Development of user-defined functions to reduce repetitive cutting and pasting of code
- Development of statistical models
- Removing unimportant variables from a model
- Examination of model residuals
- Hypothesis testing
- Measures of accuracy

➢ **Usefulness of results**
- Data visualizations reveal patterns of interest and support the data analyses
- Data analyses answer the research questions(s)

➢ **Quality of description**
- Describe the data set – source, size, type of variables, additional data prep
- Goals for data visualizations and data analyses
- Patterns revealed (or not) by the data visualizations
- Clarity of the research question(s)
- Reason(s) for selecting a particular analysis method
- Description(s) of developed statistical model(s)
- Checks on model assumptions
- Implications of calculated p-values
- Comparison of results from multiple models

- Include one or more concluding remarks
- Include citations in the text and references[4] in the References section to resources used
- Complete sentences and well-structured paragraphs instead of bullets
- Graphs and maps include a figure number in the plot title or centered beneath the plot
- Absence of spelling errors, incorrect grammar, and plagiarism[5]

8. Deliverables and Due dates

Deliverable	Due Date
Task 1a: add entry in the "Final project **stake-a-claim**" discussion board (see Section 4)	Sunday, November 14, 11:59 p.m.
Task 1b. submit Word/PDF file describing **data set and research questions** (see Section 4)	Monday, November 15, 11:59 p.m.
Oral presentation (see Section 5)	Wednesday, November 17 and December 1: -Slides due by 4:00 p.m. -Presentations starting at 7:20 p.m.
Written paper (See Section 6)	Sunday, December 5, 11:59 p.m.

[4] See here for details about APA-style refernce lists.
[5] See here for ways to avoid plagiarism

Appendix A
Some Web Sources for Data Sets

1. General Comments

One approach to finding a project data set is to use primary sources. Sections 2 and 3 below follow this approach. Section 2 contains links to 11 agencies of the U.S. federal government. In some cases, these links are to a specific type of data available from the agency. In other cases, the link is to a home page for the agency, from which it is necessary to navigate to different pages to find data sets that can be downloaded. Some agencies, like the U.S. Census Bureau, have helplines or email addresses to send questions to. However, many agencies do not have such resources other than the information they provide on their website. Section 3 contains links to two international organizations that have available data sets. The links in Sections 2 and 3 were compiled for use by students in various GMU statistics courses, so some of these links may not be that helpful in finding a data set for the STAT 515 final project.

Because it can sometimes be time-consuming to find a data set using primary sources, another approach is to use secondary sources. Section 4 below contains links to secondary sources for which it is not necessary to create an account to download data. Section 4 also includes some links to web pages that discuss primary sources satisfying particular criteria. Section 5 lists some secondary sources for which it is necessary to create an account to obtain free additional information or to download free data.

Some secondary sources contain additional information about their listed data sets, such as descriptions of variables, processing to clean the data, and successful analyses of the data sets. If there exists readily available R code that creates visualizations for a data set or analyzes it using an analysis method that can be carried out for the STAT 515 final project, then this data set cannot be used to create the same visualizations or to perform the same data analyses.

2. U.S. federal government agencies that are primary sources of data

a) Department of Education
National Assessment of Educational Progress
State-level education data.

b) National Cancer Institute
State Cancer Profiles

c) Center for Disease Control.
CDC Wonder,
COVID tracking project
https://covidtracking.com/about-data/federal-resources

COVID19 data from National Center for Health Statistics:
https://www.cdc.gov/nchs/covid19/index.htm

d) Department Labor Statistics: has many times series
http://beta.bls.gov/maps/cew/us
Provides access to state and county quarterly Census of Employment and Wages data.

http://www.bls.gov/data/

e) Bureau of Census
data.census.gov
Economic Indicators

e) Environmental Protection Agency
Data Finder

e) Justice Department
National Archive of Crime Justice Data

f) National Science Foundation: National Center for Science and Engineering Statistics
National Center for Science and Engineering Statistics (NCSES)

g) Department of Agriculture
National Agricultural Statistics Service

Economic Research Services
Food and expenditure, nutrition, obesity, and other statistics.

h) National Oceanic and Atmosphere Administration
National Climatic Data Center

National Snow and Ice Data Center
Note that some files are in HDF5 format. There is a capability in R to read such data, but using MatLab may be better. One student reported difficulty getting data into shape using R. There were missing data and other issues to address.

3. International organizations that are primary sources of data

a) World Health Organization

World Heath Observer
WHO Data Access

b) World Bank

World Bank Data

4. Secondary data sources -- not necessary to create an account

a) Machine Learning

UCI Machine learning Repository Many data sets here.

b) War

Correlates of War

c) U.S. Federal Government

Data.Gov

d) USDA Economic Research Service

Atlas of Rural America – U.S. county-level demographic and economic data from multiple data sources. Includes various USDA classifications of U.S. counties.

e) Microsoft

Revolution Analytics blog
More data sets cited by Microsoft

f) Urban Institute

https://datacatalog.urban.org/search/type/dataset

g) 2016 Forbes article about 33 primary data sources

https://www.forbes.com/sites/bernardmarr/2016/02/12/big-data-35-brilliant-and-free-data-sources-for-2016/?c=0&s=trending#dbe02f6b54db

h) 2017 blog article by Marisa Krystian about 50 primary data sources

https://infogram.com/blog/free-data-sources/

i) GMU library

https://infoguides.gmu.edu/find-data

Contains links "Find Data Sets", "Browse Data for Projects", and "Data for Practice & Projects".

j) Google dataset explorer

https://toolbox.google.com/datasetsearch

Can search for data sets using a search term. Some of the search results contain a link to Google Scholar entries for articles that have cited the data set.

5. Secondary data sources – need to create an account to download data

a) Kaggle

https://www.kaggle.com/

Check a data set's "Code" and "Discussions" posts. If any posts contain R code for visualization or data analysis using methods in Section 2.5, that may rule out your being able to use the data set.

b) IPUMS

https://www.ipums.org/

Data sets are available from government agencies (both U.S. and foreign countries), among other data sets. IPUMS contains U.S. decennial census data back to 1790. The value added by IPUMS is that variable names are harmonized across all its data sets. Also, its website is of more recent vintage than websites of some of its primary sources, so it is often easier to download data from IPUMS instead of from the primary source. All data is free, but you must create an account to download data.

c) "Data is Plural" newsletter by Jeremy Singer-Vine jsvine@gmail.com

If you send an email to Jeremy (jsvine@gmail.com), he will put you on the mailing list for his email newsletter "Data is Plural," which describes available data sets of all kinds. He sends out a new edition several times a month, and each edition contains a link to a google docs spreadsheet containing links to the over 1200 data sets he has described in past editions. The following is a link to an archive of the newsletters:

http://tinyletter.com/data-is-plural/archive

and the following is a link to a spreadsheet with all the info:

https://docs.google.com/spreadsheets/d/1wZhPLMCHKJvwOkP4juclhjFgqIY8fQFMemwKL2c64vk/edit#gid=0

d) Pew Research Center
https://www.pewresearch.org/download-datasets/

Appendix B
Data Sets Analyzed in STAT 515 Course Materials for Weeks 10-12

Week	Source	Description
10	ISRL package	Hitters data set College data set
	Agridat package	Davidian.soybean data set
	elect2016_county_ covariates.csv	County-level results and covariates for U.S. 2016 presidential election
11	ISLR package	Carseats data set Boston housing data set
	Ecdat package	Housing data set (for Windsor Canada)
	Columbia University	Speed dating data set
	Dean De Cook	Ames, Iowa, housing data set
12	J. Lander's website	Wine data set
	World Bank (WB)	Created using WB API –see Chapter 25 of R for Everyone (2nd edition)
	Airbnb	Airbnb listings for Washington D.C,