Final Report

Stats 20 - Fall 2016

Due: Wednesday, December 7, 2016 by 9:00PM

Introduction

The purpose of this report is give students an opportunity to apply the `R` skills learned during the course to create a polished report document. The ability to create polished documents is an important skill for future coursework as well as for future employment.

The skills that will needed to complete the final report include:

- Reading in data of different types from different sources.
- Cleaning, manipulating and joining data for analysis and plotting.
- Create and interpret statistical graphics.
- Perform statistical tests and interpret the results.
- Create functions to perform tasks.
- Use loops, if/else and other control structures.

Guidelines

To receive full credit, students must write a single, reproducible R Markdown document which can be compiled by the lecturer to create an HTML file. Each document must rely solely on `R` code and the data provided to create the final report document. Importing figures or using other data analysis sources is not permitted.

Since each report `.Rmd` file must be reproducible, meaning the will compile each `.Rmd` document to create the resulting output, please use the following folder/file conventions:

- Create a new folder on your computer to store the report and data files.
  - Place your .Rmd file into this folder.
- Create a new subfolder named `data` to store the data files.
  - Place the data files into this subfolder
- Do not alter the data nor the names of the data files.
  - Do not use other software to convert non-`R` data into other file formats. Read the data, as given, into `R` using appropriate `R` functions.

A generic example of how to arrange your files/folders is given below:

```
/final-report-folder/
    /data/
        data-file-01.ext
        data-file-02.ext
        ...
    final-report-document.Rmd
```

Students that submit documents which require alterations by the lecturer in order to compile will be penalized.

Style

The final report should look and read like a polished report. `R` code should not appear in the document except in instances where you are asked to include it and only relevant outputs should be printed to the compiled

document (such as the outputs from statistical tests, for example). Plots should be sized appropriately and should include descriptive titles, proper labels, etc. Numerical summaries should be included in tables (via the `kable()` function). Your document should include an appropriate title, your name and appropriate section headings.

Reports that include excessive/irrelevant outputs, lack explanation of the steps needed to achieve outputs, lack interpretation, include poorly formatted plots, etc. will be penalized.

Seeking help

You may seek help from other people (such as your lecturer, TA or classmates) but each report should include only your own work. Reports that are thought to share code/writing will be forwarded to the Dean for investigation.

Assignment

The assignment, and the report that students write, should be broken up into the following sections. Within each section, students should perform the required tasks, explain their methods and reasoning and interpret any outputs that are included in the report.

Introduction:

- Briefly mention the data files you've been given and talk about how you use the data to answer the questions in the **Analysis** section.
  - Write about why the questions you're answering in the **Analysis** section might be of interest to others.

Data:

NOTE: You should read the data files directly into `R` as received. Don't use other software to change the file formats.

- **Describe the individual data files**. Topics you might consider addressing include the original sources of the data, what are the observations of each dataset and, broadly, what sort of variables are present in the data that describes each observation, what are the dimensions of the original data and what were the dimensions after you were done cleaning the data, etc.
- **Clean the data**. Drop unneeded variables (variables not used in your analysis), change missing values in the datasets to `NA`s, change date/datetime variables to date/POSIX class variables, ensure numeric values are indeed numeric and not factors/characters, etc.. In your report, describe the steps taken to clean the data.
  - From the `edmunds.dta` file, feel free to drop the variables `bodytype`, `trim`, `intcolor`, `fabric_intcolor`, `fuel`, `engine`, `transmission` and `new_used`.
  - From the `irs-la-zip.xls` file, you'll want to keep, at a minimum, the following variables: `N1`, `MARS1`, `MARS2`, `MARS4`, `NUMDEP`, `A00100`
  - You might find it useful to give these variables more descriptive names when you're analysing your data.
- **Summarize**. Answer the following questions (or perform the directed tasks) by using plots, tables of numerical summaries and/or frequency tables. Be sure to interpret/explain the outputs, plots, tables, etc.

  - In the `edmunds.dta`: (1) Describe the distribution of when leads were submitted by plotting the distribution of the months for the `lead_date` variable (2) Describe the distribution of `model_year` (3) Which `makes` were the most/least popular (4) what was they typical `msrp` and describe how much it varied.
  - In the `irs-la-zip.xls` data, describe the variables: `N1`, `MARS1`, `MARS2`, `MARS4`, `NUMDEP`, `A00100` and how they relate to eachother.
  - In the `la-cell-towers.csv` data: (1) Which zip codes have the most/least number of cell towers (2) What are the names of the cities that are located in the zip code with the most cell towers?

- **Join your data files**.
  - Append the information in `irs-la-zip.xls` to the data in `edmunds.dta`.
  - Calculate the zipcode with the most cell towers and use these zip codes to semi_join the combined edmunds/irs data. How many observations were left in the Edmunds/IRS data after semi-joining?

- **Bonus task**: Include a map that shows the location of cell towers.

Analysis

- **t-test**. Choose two numerical variables and perform a t-test to see if there's a significant difference between the mean values of each variable.
  - Write why you chose to compare these variables, what you hope to learn by looking at the difference in means and why using a difference in means t-test is appropriate.
  - Include the output of your t-test in your document.
- **Linear regression**. Create a linear model to answers the question *How are the number of dependents and the number of returns that are filed as "single" related?*
  - Include a plot of the variables with the line of best fit overlaid.
  - Display the summary output of the linear model in the document.
- **Custom functions**. Create a function that takes a vector of zip codes as inputs and outputs a table that describes, for each zip code, the number of cell towers in that zip code, the total number of filed tax returns (MARS1, MARS2 & MARS4) and the number of car leads that came from that zip code.
  - Include your function code in your document and include relevant comments in your code.

Results & conclusions.

- Describe the results of t-test. Interpret the output and explain what the result means in the context of the question that motivated using the t-test.
- Interpret the results of your linear model. Specifically, interpret what the slope coefficient means and how strongly correlated the variables are. Should we be surprised by the relationship between the number of dependents and the number of single filers? Is there a lurking variable that might be impacting both variables?
- Demonstrate that the function you created in the analysis portion works.