Explore and Summarize Data

Project Overview

In this project, you will use R and apply exploratory data analysis techniques to explore relationships in one variable to multiple variables and to explore a selected data set for distributions, outliers, and anomalies.

Prepare for this project with: Data Analysis with R.

Note

If you have successfully completed the project for the Data Analysis with R course in the past (which entails having graduated from the course and having access to your course certificate), simply email us at dataanalyst-project@udacity.com with your passing evaluation and we'll give you credit for this project.

What do I need to install?

In order to complete the project, you will need to install R. You can download and install R from theComprehensive R Archive Network (CRAN).

After installing R, you will need to download and install R Studio. Choose the appropriate installation for your operating system.

Finally, you will need to install a few packages. We recommend opening R Studio and installing the following packages using the command line.

```
install.packages("ggplot2", dependencies = T)
install.packages("knitr", dependencies = T)
install.packages("dplyr", dependencies = T)
```

For more information on installing R packages, please refer to Installing R Packages on R Bloggers.

Why this Project?

Exploratory Data Analysis (EDA) is the numerical and graphical examination of data characteristics and relationships before formal, rigorous statistical analyses are applied.

EDA can lead to insights, which may uncover to other questions, and eventually predictive models. It also is an important “line of defense” against bad data and is an opportunity to notice that your assumptions or intuitions about a data set are violated.

What will I learn?

After completing the project, you will:

- Understand the distribution of a variable and to check for anomalies and outliers
- Learn how to quantify and visualize individual variables within a data set by using appropriate plots such as scatter plots, histograms, bar charts, and box plots

- Explore variables to identify the most important variables and relationships within a data set before building predictive models; calculate correlations, and investigate conditional means
- Learn powerful methods and visualizations for examining relationships among multiple variables, such as reshaping data frames and using aesthetics like color and shape to uncover more information

Why is this Important to my Career?

"If you are looking for a career where your services will be in high demand, you should find something where you provide a scarce, complementary service to something that is getting ubiquitous and cheap. So what's getting ubiquitous and cheap? Data. And what is complementary to data? Analysis"

— Hal Varian, UC Berkeley, Chief Economist at Google

How do I Complete this Project?

This project is connected to the Data Analysis With R course, but depending on your background knowledge of exploratory data analysis, you may not need to take the whole class to complete this project.

Introduction

For the final project, you will conduct your own exploratory data analysis and create an RMD file that explores the variables, structure, patterns, oddities, and underlying relationships of a data set of your choice.

The analysis should be almost like a stream-of-consciousness as you ask questions, create visualizations, and explore your data.

This project is open-ended in that we are not looking for one right answer. As John Tukey stated, "The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data." We want you to ask interesting questions about data and give you a chance to explore. We will provide some options of data sets to explore; however, you may choose to explore an entirely different data set. You should be aware that finding your own data set and cleaning that data set into a form that can be read into R can take considerable time and effort. This can add as much as a day, a week, or even months to your project so only adventure to find and clean a data set if you are truly prepared with programming and data wrangling skills.

Now, on to the details!

Step One - Choose your Data Set

First, you will choose a data set from the Data Set Options document. You should choose a data set based on your prior experiences in programming and working with data. The data set you choose will not increase or decrease your chances of passing the final project. In general, tidy data sets are easier to work with since each variable is a column and each row is an observation; there's no data cleaning or wrangling involved. We offer guidance below for choosing your data set. Time estimates include reading all of the project instructions and rubric, conducting the analysis, and submitting the final project.

Step Two - Get Organized

Eventually you'll want to submit your project (and share it with friends, family, and employers). Get organized before you begin. We recommend creating a single folder on your desktop that will eventually contain:

1. The RMD file that contains the analysis, final plots and summary, and reflection (in that order)
2. The HTML file that will be knitted from your RMD file

3. The data set you used (which you will only submit if you found your own data set)

Step Three - Explore your Data

This is the fun part. Start exploring your data! Keep track of your thoughts as you go (in an RMD file). Please refer to the Example Project that we have provided. Your report should look similar!

Step Four - Document your Analysis

You will want to document your exploration and analysis in an RMD file which you will submit. That file should be formatted in markdown and should contain (in order):

1. A stream-of-consciousness analysis and exploration of the data.
2. a. Headings and text should organize your thoughts and reflect your analysis as you explored the data.
3. b. Plots in this analysis do not need to be polished with labels, units, and titles; these plots are exploratory (quick and dirty). They should, however, be of the appropriate type and effectively convey the information you glean from them.
4. c. You can iterate on a plot in the same R chunk, but you don't need to show every plot iteration in your analysis.
5. A section at the end called "Final Plots and Summary"
6. You will select three plots from your analysis to polish and share in this section. The three plots should show different trends and should be polished with appropriate labels, units, and titles (see the Project Rubric for more information).
7. A final section called "Reflection"
8. This should contain a few sentences about your struggles, successes, and ideas for future exploration on the data set (see the Project Rubric for more information).

Step Five - Knit your RMD file

Your knitted RMD file should not be one long chunk of R code. It should contain text and plots interspersed throughout. The goal is to give the person reading the file insight into what you were thinking as you explored your data.

Step Six - Document your Data (if you chose your own data set)

The data set you submit (only if you chose your own) should include a text file, like those in the R documentation (e.g. ?diamonds) that describes the source of your data and an explanation of the variables in the data set (definition of any variables, units, levels of categorical variables, and the data generating process, such as how data was collected if possible).

Project Template File

Please download the project template file to get started on your analysis.

Formatting Notes

We want you to submit a readable RMD file. To help you prepare your project, please look over the following notes.

1. The knitted HTML output should be readable. Be sure to review your knitted HTML file and check that the code and plots appear correct.
2.
3. Comments for R code in a RMD or R-Markdown file are included inside of r blocks by using a hash or pound symbol.
4.
5. ```{r}
   library(ggplot2)
   # This is an example of a comment that is not actual code.

```

6.
7. In a RMD or R-Markdown file, use of the hash or pound symbol (#) outside of r blocks of code creates an H1 header.
8. # THIS IS AN H1 HEADER
9. *You won't see the hash symbol in front of the text above once you knit the HTML file. SeeMarkdown Syntax for additional help with Markdown formatting.*
10.
11. Check that all your plots can be viewed and that they are sized appropriately for the output, which is the knitted HTML file.

## Evaluation

Use the Project Rubric to review your project. If you are happy with your submission, then you are ready to submit! If you see room for improvement in any category in which you do not meet specifications, keep working!

Your project will be evaluated by a Udacity reviewer according to the same Project Rubric. Your project must "meet specifications" or "exceed specifications" in each category in order for your submission to pass.

## Submission

Ready to submit your project? Go back to the portal, click on the project, and follow the instructions to submit!

- You can either send us a GitHub link of the files or upload a compressed directory (zip file).
```

- Inside the zip folder include a text file with a list of Web sites, books, forums, blog posts, GitHub repositories etc that you referred to or used in this submission (Add N/A if you did not use such resources).

It can take us up to 2 weeks to grade the project so keep checking back for updates.

If you are having any problems submitting your project or wish to check on the status of your submission, please email us at dataanalyst-project@udacity.com.

What to include in your submission?

1. The RMD file containing the analysis (final plots and summary, and reflection)
2. the HTML file knitted from the RMD file using the knitr package
3. the original data set and source if you used your own data rather than one recommended by Udacity (Note: do not submit a data set if you used one that Udacity recommended)
4. A list of Web sites, books, forums, blog posts, github repositories, etc. that you referred to or used in creating your submission (add N/A if you did not use any such resources).

Example Project

Your final project will be an analysis in which you analyze the variables and relationships within a data set. Your final project should look similar to this Example Project. It should follow the same structure with an Analysis section, a Final Plots section, and a Reflection section. We will provide a template for you to use with these sections already included.

Take at least 10 minutes to review the example project to get a sense of what you will need to do before starting your project.

More Udacious Projects

These projects go above and beyond the course material, and we hope they serve as inspiration to work with data that interests you and to ask interesting questions.

Climatology of Atlantic Hurricanes by Dean D. Churchill

Geography of American Musicians by Stefan Zapf

Common Problems with Project Submissions

To help you succeed, we recommend comparing your project submission against the following list of common problems. Your project must avoid these common problems in order to pass.

- Data processing or transformations (creating a categorical variable) should be included in the RMD file and the final knitted HTML output.
- Reflection section is not included.
- Reflection section is not the last section in the RMD file.
- Final Plots section is not included.
- Final Plots section is not at the end of the RMD file before the Reflection section.
- Final Plots section does not contain three plots.
- One or more plots in the Final Plots section do not reveal a finding or pattern in the data set.
- Final Plots are not polished and are missing titles or units.
- Inappropriate plots are chosen for data in the Final Plots section.

Creating Effective Plots

The Final Plots section in your RMD file should contain three polished plots that give insight into the data set that you investigated.

Each plot should also contain a caption or description about what the plot shows.

In determining whether or not you have three strong plots, please consult the document, Creating Effective Plots. The document covers four common problems that project evaluators have encountered in the past and how those plots can be improved.