Homework 1

Yiyu Zhou

September, 2023

A little bit about myself

a. Why am I interested in taking this course?

   After years of programming, I want to switch my gear a little bit and discover some new things that are very closely related to what I've been doing but which I have never done before, that is Data Science.

b. Have I ever used R before this course? If so, to what extent?

   I have never used R before, and all I have heard about R is that it's part of the GNU project.

c. Have I ever used another programming language before? How would I rate my skill level?

   I have used C, RISC-V Assembly, Java, C#, Python, JavaScript, Lisp (Emacs Lisp), Vimscript/VimL, Bash and Lua. I have extensive knowledge with C and GNU toolchain.

d. What major am I? What do I plan on doing when you graduate from USF?

   I'm a Computer Science major. I plan to graduate from USF at 2026.

e. Things that I hope to do with R this semester.

   I hope to learn how to analyze data using R this semester.

f. Something interesting I want to share.

   I was telling you earlier about this project called glow, it gives you the ability to render Markdown in the terminal emulator. It is very cool because before you commit a change to a Markdown file, you can preview the visual effects of it right in the terminal.

Chapter 1 of *Doing Data Science*: What is Data Science?

a. From reading the article, what in your opinion is the biggest issue facing the field of "data science," and what do you think can be done to help fix this issue?

   From reading the article, I think the biggest issue facing the field of "data science" is the lack of understanding on the subject. Due to people not fully knowing what data science is, there is also a lack of respect for the people working within the industry of data science. Many people believe it to be a new concept but Big Data and the field of data science have been around for a very long time and are built upon everything that has come before with work from statisticians, computer scientists, mathematicians to name a few. I think more education on the history and field of data science would be very helpful in expanding the repect and understanding for all the work being done.

b. What, in your opinion, is the biggest difference between traditional statistics and data science? How does technology play a role in this difference?

   In my opinion, the biggest difference between traditional statistics and data science is how we handle the data. For statisticians, their main goal is hypothesis testing on data whereas for data scientists, they are more focused on finding patterns and building models from data. Data science is also an interdisciplinary study combining computer science, statistics, communication, data visualization and much more.

c. What, in your opinion, is the biggest difference between data scientists in academia and industry?

   In academia, there are no data scientists. At least none who refer to themselves as such within academia. Data scientists in academia focus on advancing research whether that be through publishing papers, securing research funding, or presenting their knowledge. The definition of research within data science is a gray area and must be defined more formally for the idea of a data scientist in academia to progress. In industry, a data scientist uses data to provide actionable business value whether that be through increasing revenue, reducing costs, or improving the user experience. Data scientists in industry often measure their success through the tangible impact their work has on a commpany or consumer.

d. What kind of data scientist do you want to be? What skills are you most interested in developing?

   I want to be a data scientist who focuses on security of systems. I am most interested in strengthening my skills with the R language and better understanding how to use data to make impactful technologies for users.

e. Based on this article, a major obstacle in understanding data science in the first place is understanding what data science actually is. From what you've read, and what you understand so far, how would **you** define the term "data science"?

   I would define data science as a blend of technology and mathematics. It combines the hypothesizing of data and programming to utilize the data. The data is used for many different purposes ranging from decision making to developing new technologies and data scientists play a huge role within all of it as well as computer scientists and mathematicians. Data science can be applied in nearly every industry.

Calculations

Here are some examples of calculations in R.

a. assignment

```
x <- 3
y <- 4
```

b. ln$(x + y)$

```
log(x + y)
```

```
## [1] 1.94591
```

c. $\log_{10}(\frac{xy}{2})$

```
log(x * y / 2)
```

```
## [1] 1.791759
```

d. $2x^{\frac{1}{3}} + y^{\frac{1}{4}}$

```
2 * x ^ (1 / 3) + y ^ (1 / 4)
```

```
## [1] 4.298713
```

e. $10^{x-y} + e^{xy}$

```
e <- exp(1)
10 ^ (x - y) + e ^ (x * y)
```

```
## [1] 162754.9
```

What does `M-K` do in RStudio?

In RStudio, pressing `M-K` (`Alt + Shift + K`) brings up the `Keyboard Shortcut Quick Reference`. Another way to access RStudio keyboard shortcuts is with the menu entry: `Tools > Keyboard Shortcuts Help`.

What is "Big Data"?

Big Data refers to extremely large data sets that are beyond the capability of traditional databases and data processing systems to capture, store, manage, and analyze. These data sets can be structured (like databases), semi-structured (like JSON or XML files), or unstructured (like text documents or images).