Lab Mr. Clean

Abhi Thanvi, Paul Holaway

June 23rd, 2022

Contents

Lab Mr. Clean

Lab Mr. Clean

Welcome

Just like learning a new spoken language, you will not learn the language without practice. Labs are an important part of this course. Collaboration on labs is **extremely encouraged**. If you find yourself stuck for more than a few minutes, ask a neighbor or course staff for help. When you are giving help to your neighbor, explain the **idea and approach** to the problem without sharing the answer itself so they can figure it out on their own. This will be better for them and for you. For them because it will stick more and they will have a better understanding of the concept. For you because if you can explain it to other students, that means you understand it better too.

The Idea of this Lab

The idea behind this lab is to allow us to think about the different aspects of Data Cleansing and Experimental Design, and why they are relevant to becoming a data scientist. We want you to understand that becoming a Data Scientist is more than becoming a coder, but also a problem solver who can think critically about what to do with data. Therefore, this lab is designed to be more of answering questions and reflecting with your peers and not coding in R.

"It is as important to ask the right questions as it is to give right answers" - Woke Abhi

Problem 1: Mr. Clean has some questions for you...

Hey Guys and Girls! This is Mr. Clean! I have written some of the questions I had for you to answer regarding the importance of cleaning. While I clean grease stains, you should be familiar with cleaning data! Feel free to ask your friends if you get stuck, and always reflect upon your answers.

Question 1: What are some of the ways I can "clean" data? Make sure to give 2-3 examples relevant to Data Science and elaborate on why or how it helps!

Answer: Here are some ways to clean data: - You can rename columns (this helps by clarifying data for later use) - You can remove data (this gets rid of unnecessary data) - You can select specific columns to work with (this helps you focus on relevant data)

Question 2: If I have a data set about the Apple iPod Sales from the year 2000 to year 2015. I noticed that a column called Sales in Dollars, but it has a bunch of empty spots in the beginning. I think it's because iPod did not begin selling in the year 2000. What can we do to solve this discrepancy or "emptiness"? Can we fill it with something? This is something to prepare you for next lecture, so all ideas are welcome!

Hint: Make sure to think about what you fill affects your data and to think if that affect is valid

Answer: We shouldn't fill it with zeros, since this can introduce skew in the data. We could remove all values from 2000 in order to have unskewed and filled data.

Question 3: Why do we filter things out of our data set? Isn't it bad to drop row/columns or is there a way we can make a copy of the original data set with variables or something?

Hint: You are technically answering a two part question here

Answer: We filter out data in order to keep only relevant information without introducing skew or bias. It isn't always bad to remove rows or columns, since they can sometimes store unimportant or unnecessary information.

Question 4: Renaming Columns. We want you rename one of the columns. Import the library needed and hello data set (a.k.a `hello.csv`) and print the first 10 rows. You will notice that the Name column has only first names. So let's change that. **Change the Name column to First_Name in Hello data set you imported**

Answer:

```
#Import library
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#Import Dataset
Hello = hello_csv <- read.csv("~/Desktop/DPI 2022/hello.csv", stringsAsFactors=TRUE)
#Change the Name column to First Name
Hello = Hello %>% rename(FirstName = "Name")
head(Hello, 10)
```

```
##    FirstName                        Major      Year   Phone
## 1   Mathilde Community Health & Chemistry    Senior  iPhone
## 2       Luke                   Stats & CS Sophomore Android
## 3     Johnny                        ETMAS    Junior  iPhone
## 4     miller            stat/data science  Freshman  iPhone
## 5        Tri                      CS+Math    Junior Android
## 6     Dhruva         Computer Engineering  Freshman  iPhone
## 7    Jeffrey          Information Science  Freshman  iPhone
## 8      Josue                     Business    Senior  iPhone
## 9     Marcel          Information Science    Junior  iPhone
## 10    Odalys          Information Science    Junior  iPhone
##                    Computer Straw Shoe.Size Pets Hot.Dog
## 1   Windows-based computer     1       9.0    0     Yes
## 2   Windows-based computer     1       9.5    0      No
## 3  Mac OS X-based computer     1      10.0    2      No
## 4  Mac OS X-based computer     1      10.0    2      No
## 5   Windows-based computer     2      10.0    0     Yes
## 6   Windows-based computer     1      10.0    0      No
## 7   Windows-based computer     1      10.5    1      No
## 8  Mac OS X-based computer     1      10.0    0      No
## 9   Windows-based computer     1      11.0    1      No
## 10 Mac OS X-based computer     1       9.5    1     Yes
##                             Streaming Prior.Programming Season
## 1                             Netflix                No   Fall
## 2                     Twitch, Youtube               Yes Spring
## 3                    Netflix, Youtube                No   Fall
## 4                    Netflix, HBO Max               Yes Summer
## 5                     Twitch, Youtube               Yes   Fall
## 6                    Netflix, Youtube               Yes Summer
## 7    Hulu, Netflix, HBO Max, Youtube               Yes Spring
## 8             Hulu, Netflix, Youtube                No Summer
## 9  Netflix, HBO Max, Twitch, Youtube               Yes Summer
## 10 Netflix, HBO Max, Twitch, Youtube               Yes Summer
##    Statistics.Courses Programming.Courses Study.Hours Siblings Sleep Shoes
## 1                   3                   0         3.0        1   8.0    10
## 2                   4                   2         4.0        1   6.5     4
## 3                   1                   0         2.0        1   7.0    10
## 4                   0                   2         2.0        2   6.0     6
## 5                   3                   6         3.0        2   6.0     1
## 6                   1                   3         2.5        1   6.0     5
## 7                   0                   4         2.0        1   8.0     8
## 8                   2                   1         3.0        6   7.0    15
## 9                   4                   4         4.0        1   7.0     6
## 10                  1                   1         5.0        1   7.0     8
##    Texts Personality Zodiac.Sign
## 1      7   Introvert Sagittarius
## 2     15   Introvert      Gemini
## 3      4   Introvert         Leo
## 4     12   Extrovert Sagittarius
## 5      0   Introvert       Libra
## 6     17   Extrovert   Capricorn
## 7     50   Introvert      Cancer
## 8      8   Introvert       Virgo
## 9      6   Introvert      Pisces
## 10     6   Extrovert      Gemini
```

Problem 2: Design Questions

Question 1: A group of researchers wants to study the effect of music at different volumes on the reaction times of drivers. They recruit 500 volunteers. They assign each subject a number from 1 to 500 by using a random number generator to assign the first 250 subjects to take the driving test at one music level. The remaining 250 subjects take the test with second music level.

What type of experiment is this and why is it that?

A.) Clustering B.) Favoritism C.) Completely Randomized D.) Winner Takes All

Answer: C - Completely Randomized

Question 2: Ask a question. For example, "What is the best time to go workout?" or "What is the best Starbucks drink?". Find inspirations from these examples, and come up with your own question. Then, explain what sort of data would you collect and how would you collect it. This is the last technical question of the lab, so feel free to spend some time on this!

Hint: We haven't covered sampling, but answer should be intuitive. Instructors and Friends are here to help! Ask questions, have fun, and be creative!

Answer: My question would be "What is your favorite season?" I would stratify my data by sorting the people I ask into categories (for example, organized by age), in order to compare responses of people based on age. There would only be four options (four seasons) and I would collect it online so I can assign it to lots of people to get many responses in order to have less potential bias or skew in my observational data.

Feedback

Hey this is Abhi! As this first week comes to an end, I would like to know whether you are liking the course or you hate your summer because of us (hopefully not!).

Please give some feedback of what you like about the course and what you would like to change about this course! We will try our best to make this the best course and have the best time as much we can! Have a great weekend and Paul and I will see you on the other side :)

Feedback: I like this course a lot! One thing I'd like to have more of is more talk about mathematical aspects of statistical analysis (for example, using R to do regression analysis) or maybe using R to create graphics.

Submission

Once you have finished your lab…

1. Go to the top left and click `File` and `Save`.
2. Click on the `Knit` button to convert this file to a PDF.
3. Submit **BOTH** the `.Rmd` file and `.pdf` file to Blackboard by 11:59 PM tonight.