How valuable is your data?

Prof. Jonathan Davis

Due by Friday, October 8th at Noon

1. Average Class Size

The data file `EconEnrollment.csv` includes enrollment data for every economics course at the University of Oregon from 2014Q1 to 2020Q3. A description of the variables in the data is in the table below.

Name	Description
`Term`	Quarter Code
`course number`	Number of course
`coursename`	Descriptive Name
`instructor`	Instructor
`enrollment`	Number of students in course
`level`	Course level (0 - Intro, 1 - Intermediate, 2 - Masters, 3 - PhD)

In this problem, we will explore a couple of R's "tidyverse" packages while analyzing this data.

1. Install and load the 'readr' and 'dplyr' packages.

```
# You can type your R code here.
library(readr)
library(dplyr)
```

2. Set your working directory to the folder on your computer where you downloaded "Enrollment-ByTerm.csv" using the "setwd()" function. Load the data using the read_csv function from the 'readr' package. Give the dataframe a logical name, like "enroll".

```
# Erase eval=FALSE. You will need to do this for every block of code.
# I had to put it there to show but not run the code.
# It is an option that tells Markdown to not run the code.

# Complete the commands below
setwd("C:/Users/campb/OneDrive/econometrics/HWone")
getwd()
```

```
## [1] "C:/Users/campb/OneDrive/econometrics/HWone"
```

```
enroll <- read_csv("C:\\Users\\campb\\OneDrive\\econometrics\\HWone\\EconEnrollment (1).csv")
```

'dplyr' is a popular package for cleaning data in R. You can use it to apply a sequence of functions to a dataframe. It returns the data with all of the functions applied in sequence. Dplyr uses "pipes" which are written as "%>%".

For example, if you want to filter the data to only see courses taught by me, you can use the following code:

```
DavisCourses <- enroll %>% filter(instructor=="Davis, Jon")
DavisCourses
```

```
## # A tibble: 8 x 6
##     Term coursenumber coursename                    instructor enrollment level
##    <dbl>        <dbl> <chr>                         <chr>           <dbl> <dbl>
## 1 202001          423 Econometrics                  Davis, Jon         NA     2
## 2 201903          607 Experimental Econ             Davis, Jon         13     3
## 3 201901          311 Inter Micro Theory            Davis, Jon         75     1
## 4 201901          423 Econometrics                  Davis, Jon         25     2
## 5 201802          428 Behav and Exp Econ            Davis, Jon         80     2
## 6 201802          607 Applied Behavioral Economics,~ Davis, Jon          4     3
## 7 201801          311 Inter Micro Theory            Davis, Jon         83     1
## 8 201801          311 Inter Micro Theory            Davis, Jon         74     1
```

This shows that I have taught 8 classes in the dataset. My smallest class had 4 students and my largest class had 83 students. Note that the enrollment for this term is listed as NA.

Let's drop the current term to make things easier.

```
enroll <- enroll %>% filter(Term<202001)
```

3. In the above example, "filter" is an example of a verb being applied to the data. The power of dplyr is in all of the verbs that can be used! Fill in the missing pieces of the following code to find the average class size in each term. In what terms were class sizes biggest and smallest?

```
byTerm <- enroll %>%
  group_by(Term) %>%
  summarize(avg = mean(enrollment)) %>%
  arrange(avg)

byTerm
```

```
## # A tibble: 18 x 2
##      Term   avg
##     <dbl> <dbl>
##  1 201901  59.7
##  2 201801  64.7
##  3 201902  69.1
##  4 201903  70.3
##  5 201803  71.3
##  6 201802  78.2
##  7 201701  80.9
##  8 201601  81.3
##  9 201703  82.1
## 10 201603  84.8
## 11 201402  88.1
```

```
## 12 201602  88.3
## 13 201503  91.4
## 14 201502  91.4
## 15 201702  94.6
## 16 201401  95.6
## 17 201501  97.7
## 18 201403  98.1
```

```
#Class size was smallest during term 201901 and biggest during term 201403
```

4. What was the average class size in the economics department over this time period? Hint: You can base your code off of the code above.

```
avgclass <- enroll %>%
  summarize(averag = mean(enrollment)) %>%
  arrange(averag)
avgclass
```

```
## # A tibble: 1 x 1
##   averag
##    <dbl>
## 1   82.1
```

```
#average class size was 82.107 students
```

5. What was the average class size in the economics department by level of the course over this time period? Hint: Use the group_by verb.

```
byLevel <- enroll %>%
  group_by(level) %>%
  summarize(classavg = mean(enrollment)) %>%
    arrange(classavg)

byLevel
```

```
## # A tibble: 4 x 2
##   level classavg
##   <dbl>    <dbl>
## 1     3     12.4
## 2     2     53.8
## 3     1     75.9
## 4     0    204.
```

```
#A verage class size for level 3 is 12.41, level 2 is 53.82, level 1 is 75.88, and level 0 is 204,44>
```

6. Now, calculate the average class size by level using the weighted.mean() instead of the mean() function. Weight the mean by each classes enrollment. Interpret what this weighted average tells us. Would a prospective student prefer knowing the weighted or unweighted average? Wht about a prospective faculty hire?

```r
weightedlevel <- enroll %>%
  group_by(level) %>%
  summarize(acsl = weighted.mean(enrollment, enrollment)) %>%
  arrange(acsl)

weightedlevel
```

```
## # A tibble: 4 x 2
##   level  acsl
##   <dbl> <dbl>
## 1     3  14.3
## 2     2  65.5
## 3     1  87.0
## 4     0 251.
```

```r
# The weighted mean causes the class levels with the higher enrollment to be even higher. I don't see h
```

2. Predicting Criminality

This question will be based on the paper "Automated Inference on Criminality using Face Images" by Xialoin Wu and Xi Zhang. The paper is posted on Canvas.

This paper attracted a lot of attention and stirred controversy when it was first posted in 2016. See for example this Vice [article](. Most of the media about the article focused on the ethics of predicting criminality. This question will help you assess how concerned you should be about the future of predicting criminality with only data on faces.

2.a What is the authors' research question?

They wished to see how accurate AI technology could be at inferring criminality based on a photo of an individuals face.

2.b What do the authors' find? How accurate are their predictions?

They found that criminals faces have a significantly larger variety than law abiding citizens faces. Based on this finding it could mean AI could possibly find the potential for criminal behavior based on if an individuals face fit in with the general law abiding public, or had a larger variety. The authors say that the "variation among criminal faces is significantly greater than that of non-criminal faces". They later say that the CNN classifier had a 89.51% accuracy which is significant.

2.c As a student in a graduate economics course, would you describe their methods as accessible or inaccessible?

The methods were quite inaccessible, I had trouble following along with many terms and methods they discussed. This would lead me to believe that the general public would have an equally difficult time in understanding some of the methods used in this paper. Their different classifiers were difficult to understand along with the discussions of what parts of the face were significant and the techniques they used to ensure the photos were controlled were especially difficult to understand.

2.d How do the authors collect their data? Are the photos of the criminals and non-criminals comparable?

They collected data of the law-abiding citizens ID photos from the internet and got criminal ID photos from wanted suspect ads and from city police departments. They used ID photos for both so they are comparable.

2.e Look at Figure 10. What jumps out to you about the main difference between the "average" criminal's face compared to the "average" non-criminal's face?

The faces of the non-criminals appear to me to be more “friendly”, but I am not sure if that is biased from me reading the caption and knowing which faces were criminal. The criminal facial features appear to be a bit more spread out.

2.f True/False/Uncertain. You need to understand the methodology of a paper to assess whether it’s conclusions are plausible. Justify your answer.

I would say false as you as an individual do not necessarily need to understand the methodology to assume the conclusions are plausible. You would want the researchers to use the best methods possible and in certain sectors these are likely to be highly advanced that the average citizen would not understand. As long as there are trusted individuals that are outside the realm of the paper that understand the methodology and can confirm it is valid, the general public does not need to understand.