# Quantifying the World

## MSDS 7333

*Dr. Robert Slater*

*rslater@smu.edu*

World Changers
Shaped Here

SMU®

# Agenda

- Introductions
- Syllabus
- Expectations
- Imputation

# Your Professor

## Robert Slater

- Doctorate (Physics) 2001- Michigan State University
  - Experimental Physics (I build/break stuff and I know things)
  - 10 years as a research engineer building semiconductors and hard drive read/write heads
- Switched Fields in 2010 to Data
- Current Field: Natural Language Processing
- Previous Work
  - Image Recognition
  - Sensitive Data

# Contact Info

- rslater@smu.edu
- 972-837-5072
  - Text (Identify who you are, please!)
- Office Hours TBD
  - Right now!
  - Also by appointment
- Remember I have a day job, too!
  - Don’t expect responses between 9-5
  - Goal is within 24 hours follow up
    - Remember that **email is not instantaneous**!!
    - Assume it takes me 2-3 hours to even see your message

# Around the Horn Introductions

- Who are you?
- What do you do?
- What do you want to learn in this class?
- Why is R superior to Python?

# That last question was a plant

- I am here to challenge you to learn more and THINK about what you are doing
- My job is to challenge your assumptions and put you on firm footing to make your own decisions
- Although I do prefer Python over R, I recognize that is my personal choice

# Syllabus + Expectations

- Work submitted by deadline (no late work)
  - Life does intervene
  - Professors are more forgiving knowing ahead of time
  - Your professor hates surprises
  - 5 points a day for late work (If prearranged, limit of 10 points lost)
- Participate in Class
  - Be active
  - I AM WATCHING
    - I will try to encourage your participation
    - Also aware not everyone is a ‘talker’

# Grading

- First semester everyone panicked when I made them do their own work
- Group work is accepted
- I won’t be nice with grading as (theoretically) you should have 100% participation and 100% viewing the course
  - Seriously, show up to class and read the material
  - 2DS takes attendance (which I use)
  - 2DS records if you viewed a page (which I use)

# Grading

- Seriously, show up to class and read the material
- 2DS takes attendance (which I use)
- 2DS records if you viewed a page (which I use)

# Homework Format

Package

Adobe Acrobat Document

# Homework tips

- Despite what the doc says (I am not re-writing it), you may submit as a group.
- Keep the report serious.  Things like the following will cost you points:
  - *“The results are interesting”*
  - *“I struggled with code”*
  - *“This assignment was a waste”  (note that this is when submitted as text in an assignment.  If you do feel an assignment is a waste, bring it up to me--just not in your writeup)*

# Writeups

- Don’t include code from the text in your write/method. Put it in an appendix.
- If you come up with something new/modified, THEN put it in your method
- Make your charts readable
- Presentation and neatness count
- Titles/captions/figures—details, details, details!
- Group submission—make sure your name is included!!

# Writeups

- I am not an English teacher
- I have been a boss
- Write clearly
- Use proper equations ($x^2$ vs x^2)
- Caption your figures
- Answer the question/problem posed
- Data should be summarized in charts/graphs
- Include your code as an appendix

# DO NOT

- CODE BLAST me (submit a bunch of code with no comments or writeup)
- Ignore my beautiful formatting documents.  Figures need captions. Graphs need titles. Presentation is a lot in the world.
- Wait until the last minute and then ask for an extension
- Seriously. We are adults—the carnage of ER bills for this class is very suspicious.

# Final Grade Policy Aims

| Final Grade | Final Score |
|---|---|
| A | 90 - 100 |
| B | 80 - 89 |
| C | 70 - 79 |
| D | 60 - 69 |
| F | < 60 |

*Don't be here; you are in graduate school*

- Don't panic if you flop an individual assignment. I give out 50s and 60s when you don't do the work or just do a bad job.
- One bad grade won't  kill you.
- Generally, **Final** grades of C or less means you did REALLY bad—but it is just FEEDBACK.
- Try and keep your unit assignments average above 70 and all will be well.

# Relax

- I grade tough on assignments to distinguish between A and B students
- I used to be totally chill and then my students walked all over me for deadlines
- If you have a long term issue, discuss it with me ASAP.
  I have had students go through major surgery, loss of a parent in a foreign country
  - There are narrow exceptions to late work—you'll know if you have one and I will absolutely work with you to get around personal issues

# But Professor,…

- The online content isn't up to date
    - Welcome to the real world
- Units 13 & 14 are broken and we won't be covering them, do I still have to view the videos?
    - No
- I hate Python/R/~~SAS~~/Giraffes (YEAH, I KILLED SAS!)
    - You're gonna have a bad time then
- I'm not a strong coder
    - Get better
    - Come to office hours

# But Professor,… (part 2)

- I hate Breakout exercises
  - Then you are gonna have a bad time
- I didn’t realize you changed x….
  - That’s why you come to class
  - That’s why I record class
  - That’s why I post on the wall

*I am sad I have to include these slides, but it is necessary*
*(these things have all happened)*

# Expectations (Continued)

- Ask if you need help (Its my job!)
  - Struggle for a while. If you're stuck for an hour, call the Prof!
- 1 Additional Instructor
  - Maybe better fit
  - Better time
  - Different Methods
- Be Aggressive and Curious
  - Got an error message?  Google it…..
- Collaborate
  - No copying!

# Expectations (NO, I Am NOT joking)

- ***You are a professional***
  - Use Google
  - Use Stack Overflow
  - Look at GitHub
  - LOOK AT YOUR ERROR MESSAGES
    - The computer is trying to tell you something!
  - Find a better way
    - Find a package that does what you want
    - Understand the problem
    - Brute force also works at times

# First, a word about Math

- Data Science is a conglomerate discipline
  - Watch the notation – everyone is different
  - Example – Least squares:

| | |
|---|---|
| $y = \beta x = \varepsilon$ | vague |
| $y = x^T \beta + \varepsilon$ | bad math |
| $y_1 = \beta_j X_{ij} + \varepsilon$ | index hell (but most clear) |

# Just make it clear:

$$\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \end{bmatrix} = (\beta_0 \quad \beta_1 \quad \beta_2 \quad \beta_3) \begin{bmatrix} a_0 & a_1 & a_2 & a_3 \\ b_0 & b_1 & b_2 & b_3 \\ c_0 & c_1 & c_2 & c_3 \\ d_0 & d_1 & d_2 & d_3 \end{bmatrix} + \varepsilon$$

$$X_1 = \begin{bmatrix} a_1 \\ b_1 \\ c_1 \\ d_1 \end{bmatrix}$$

# Week 1 : Pirate Week

RRRRRRRRRRRRRRRRRRRRRR!

# R review

- Counting starts at 1
- Assignment

```
x <- 5 # preferred
x = 5  # top level only
```

```
median(x=1:10)
x
## Error: object 'x' not found

median(x<-1:10)
x
## [1]123456789 10
```

```
if(x=0)1 else x
Error:syntax error
```

# Chain assignments

```
x = y = 5
x <- y <- 5
x = y <- 5    #makes me sad
x <- y = 5    #error
```

‘<-’ is higher precedence than ‘=‘

So the last line reads like

```
x <- ‘y = 5’
```

# DataFrames

- Counting starts at 1:
- By Index

```
df[1,1]
```

- By Label

```
df['1','MPG']
```

- By column '$'

```
df$MPG
```

Will not get errors if you 'go over'

```
df['xxyz','dgsgsd']
NULL
```

# R tends to be function oriented

```
colMeans(test) # every column
colSums(test)
```

- Pass vectors for subsets:

```
colMeans(test[,c(‘MPG’,’ACCEL’)]
```

- Does not like NA

```
colMeans(test[,c('MPG','SIZE')],na.rm = TRUE)
```

# Dataframe Features

```
colNames(test)
rowNames(test)
colNames(test)<-c(’Mike’, ‘Rick’, ‘Steve’, ‘Mark’ ,’Bob’ , ‘Lee’)
```

- If you go ‘over’ the length, an error is thrown. If you are ‘under’ length, your list is appended with NA to fit

# And or Or (Or and And)

| # element wise and

|| # Or

& # element wise and

&& # and

|| and && are preferred for flow control

If you compare vectors with && or || only the first element is used!!

# Selection in R

- Very Similar to Python

```
test[test$MPG>20,'Auto']
test[test['MPG']>20,'Auto’]
```

- Multiple Selections

```
test[test$MPG > 30 | test$MPG < 17],c('Auto’, ‘MPG’)]
```

# Breakouts

- Group 1: Fords with MPG less than 20
- Group 2: Average HP for MPG between 25 and 30
- Group 3: Avg Weight for Chevy cars with ENG_TYPE 1
- Group 4: Average Weight for ENG_TYPE 0 and HP more than 100

- As usual, I threw some curves

# Packages

- R has a TON of packages
- Many work with dataframes
- If they are not official they are ILLEGAL

- JUST KIDDING

# Functions

CURLY BRACKETS: WOOT WOOT!

Indentation optional (but nice)

R does read line by line, so be careful

```
pythonSucks<-function(x){
If (x>2){
  print('look at me indenting')
}
else{
Print('or  not')}
}
```

World Changers
Shaped Here

SMU®