BSOS

Cloud and Classroom Technology for Data Science

COLLEGE OF
BEHAVIORAL &
SOCIAL SCIENCES

TLTC Experiential Learning Grant

Project Title:

Cloud-based Active Learning for BSOS Statistics and Data Science Courses

- Program-level grant
- Aimed at improving data science education within BSOS

Grant Personnel

Dr. Brian Kim, BSOS/JPSM, Co-Director of Social Data Science major

Dr. Sarah Croco, GVPT, Director of Honors Global Communities LLP

Dr. Candace Turitto, GVPT, Director of Applied Political Analytics

Dr. Jesse Klein, INFO

Dr. Madeline Brown, ANTH

Dr. Taylor Oshan, GEOG

Dr. Zubin Jelveh, CCJS

Graduate Assistants:

Ujjayini Das

Joe Hoskisson

Motivation for the Project

- Data Science is growing in all fields, not just Computer Science or Statistics.
- Needs to be incorporated into social science curricula
- Technology can facilitate incorporating data science into existing courses

Goals for the Project

- Develop a **cloud-computing environment** so that students can get practice programming without needing to install anything
- Develop **modular tutorials** that can be used in many different classes to teach basic concepts
- Build up a **data repository** of social science datasets which can be used in classes to put the concepts in context

BSOS

JupyterHub and HyFlex for Data Science

COLLEGE OF
BEHAVIORAL &
SOCIAL SCIENCES

Social Data Science major (SDSC)

- Officially launched Fall 2022
- Joint major between INFO and BSOS
- Students learn about data science as well as a social science track
- Focus on application and human context

See https://sdsc.umd.edu for more information!

BSOS 233

- Core class for SDSC major
- Designed to be an introductory course on using Python for Data Science

Challenges:

- Need to make programming accessible
- Students from a variety of backgrounds
- Scalability for growth

Addressing the Challenges

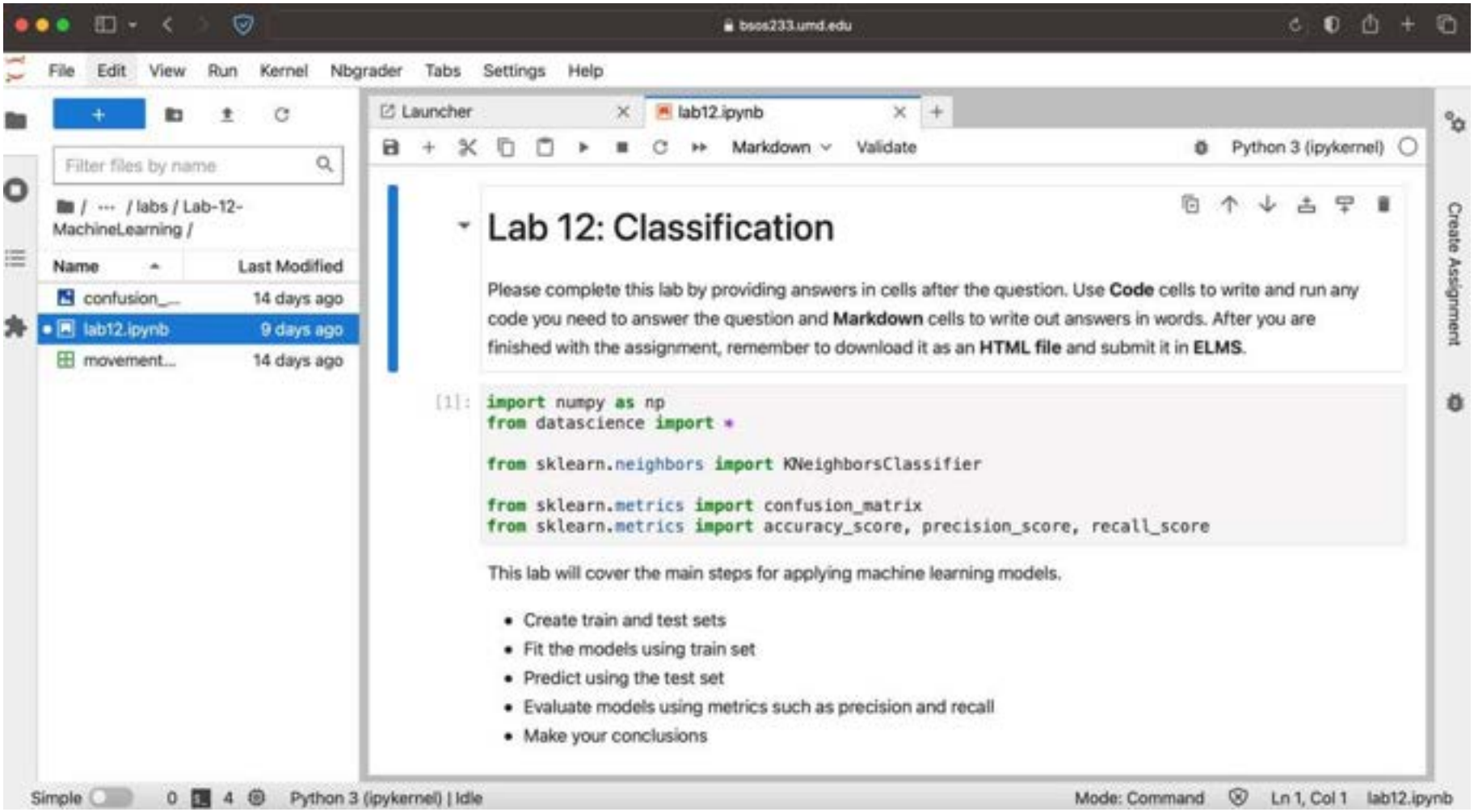
- **JupyterHub Cloud-Computing Environment** to make the programming as accessible as possible.

- **HyFlex (Hybrid-Flexible)** format to make learning in the classroom as accessible as possible.

JupyterHub

- Hosted on BSOS OACS servers
- Cloud-based platform
- Everything is done through the browser -- no need to install anything
- Students work completely within the environment, with all work saved in the cloud
- Submit assignments in ELMS by downloading an HTML file

JupyterHub


bsos233.umd.edu
File Edit View Run Kernel Nbgrader Tabs Settings Help
Filter files by name
/ ··· / labs / Lab-12-MachineLearning /
Name
Last Modified
confusion_...
14 days ago
lab12.ipynb
9 days ago
movement...
14 days ago
Launcher
lab12.ipynb
Markdown
Validate
Python 3 (ipykernel)
Create Assignment
Lab 12: Classification
Please complete this lab by providing answers in cells after the question. Use Code cells to write and run any code you need to answer the question and Markdown cells to write out answers in words. After you are finished with the assignment, remember to download it as an HTML file and submit it in ELMS.
[1]:
import numpy as np
from datascience import *
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score, precision_score, recall_score
This lab will cover the main steps for applying machine learning models.
Create train and test sets
Fit the models using train set
Predict using the test set
Evaluate models using metrics such as precision and recall
Make your conclusions
Simple
0 4
Python 3 (ipykernel) | Idle
Mode: Command
Ln 1, Col 1
lab12.ipynb

Jupyter Notebooks

- Combine executable code and narrative text together

Lab 12: Classification

Please complete this lab by providing answers in cells after the question. Use **Code** cells to write and run any code you need to answer the question and **Markdown** cells to write out answers in words. After you are finished with the assignment, remember to download it as an **HTML file** and submit it in **ELMS**.

1]:
```
import numpy as np
from datascience import *

from sklearn.neighbors import KNeighborsClassifier

from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score, precision_score, recall_score
```

This lab will cover the main steps for applying machine learning models.

- Create train and test sets
- Fit the models using train set
- Predict using the test set
- Evaluate models using metrics such as precision and recall
- Make your conclusions

Technical Details

- Accessible through https://bsos233.umd.edu
- Account needs to be created and approved manually, working on implementing campus authentication
- Assignments, labs, other documents distributed via link on ELMS (pulls from GitHub repository behind the scenes)

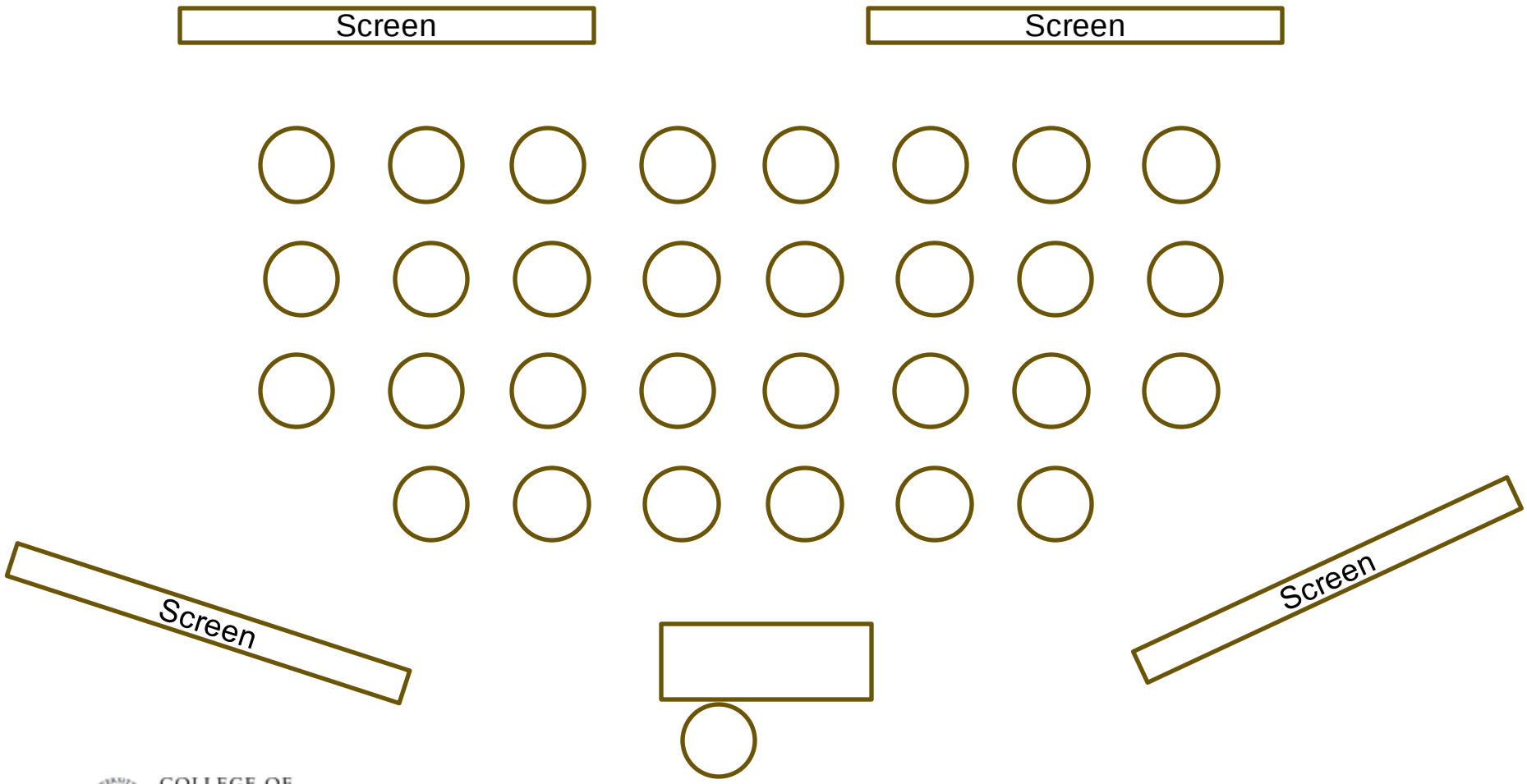
HyFlex

Motivation: Students may learn better through different modes of learning

- In-person synchronous
- Online synchronous
- Online asynchronous

HyFlex (Hybrid-Flexible) uses all of these methods of learning and lets the student choose whichever mode they like at any point throughout the semester.

Room Layout



HyFlex

HyFlex Lessons Learned

- Many students participated **even if they did not turn cameras on** in Zoom.
- Students **took advantage of the flexibility** to choose when to come in person and when to attend online.
- After around Spring Break, **the vast majority of students who attended class did so online**.

BSOS

Modular Interactive Tutorials for R

COLLEGE OF BEHAVIORAL & SOCIAL SCIENCES

Motivation

- Data science is taking over the world; why should we stay behind?
- Programming is hard!! There should be an easier way to deal with it; isn't it a good time to find out how?
- Sounds like a lot of work; where to begin?

Let's try to figure out!

What is the plan?

- Build a set of modular data science tutorials in R, which use real-world data.
- Host them on BSOS cloud servers (with the help of OACS) to easily use as assignments with back-end completion tracking.
- Use them as refresher material in more advanced courses to make sure students meet the requisite level of programming.

What is the benefit here?

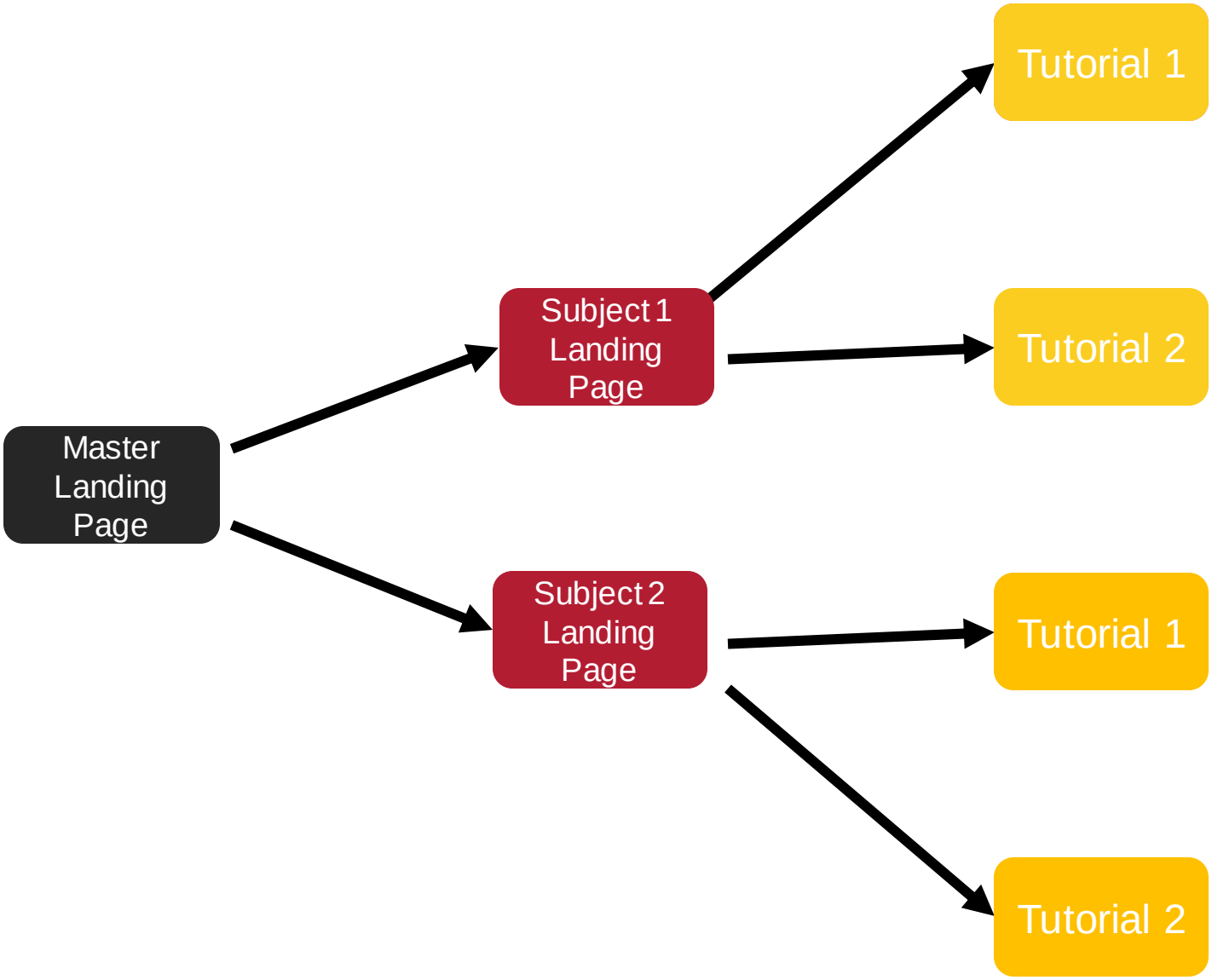
- No need to install software in personal computers, hence less burden for students to get started!
- More streamlined, improved accessibility
- Step-by-step guide on chosen topics based on different levels of courses
- Interactive tutorials, hands on experience in coding
- Exercises with sequential difficulty levels to try out
- Repeated attempts allowed to encourage true learning

Tools up our sleeves!

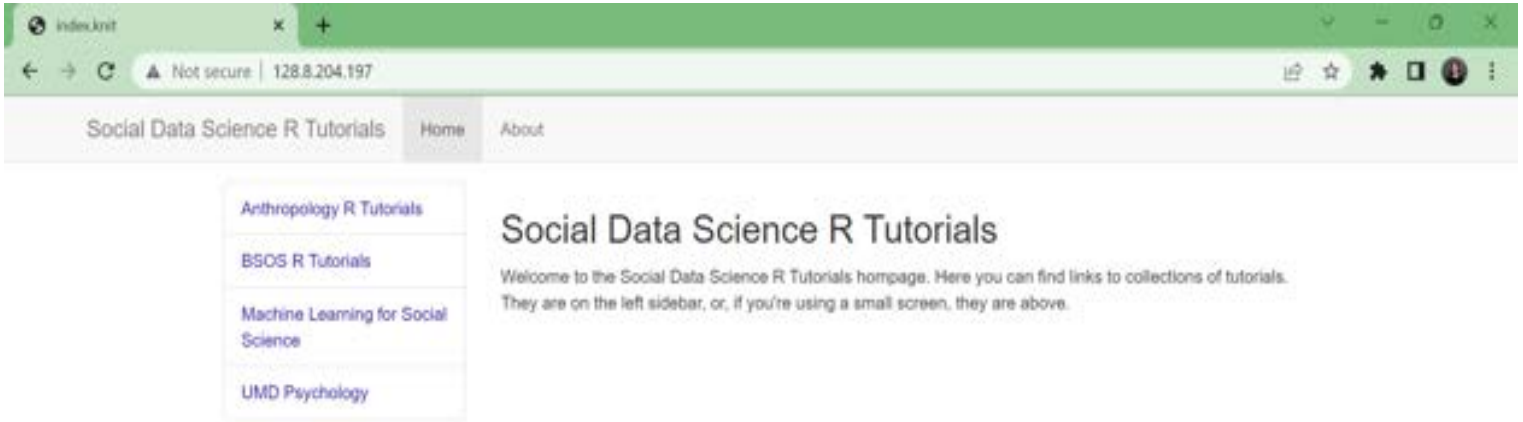
- **Software used**
  - R Markdown
- **Main Packages**
  - learnr
  - gradethis
- **Hosting Server**
  - R Shiny

https://pkgs.rstudio.com/learnr/articles/exercises.html

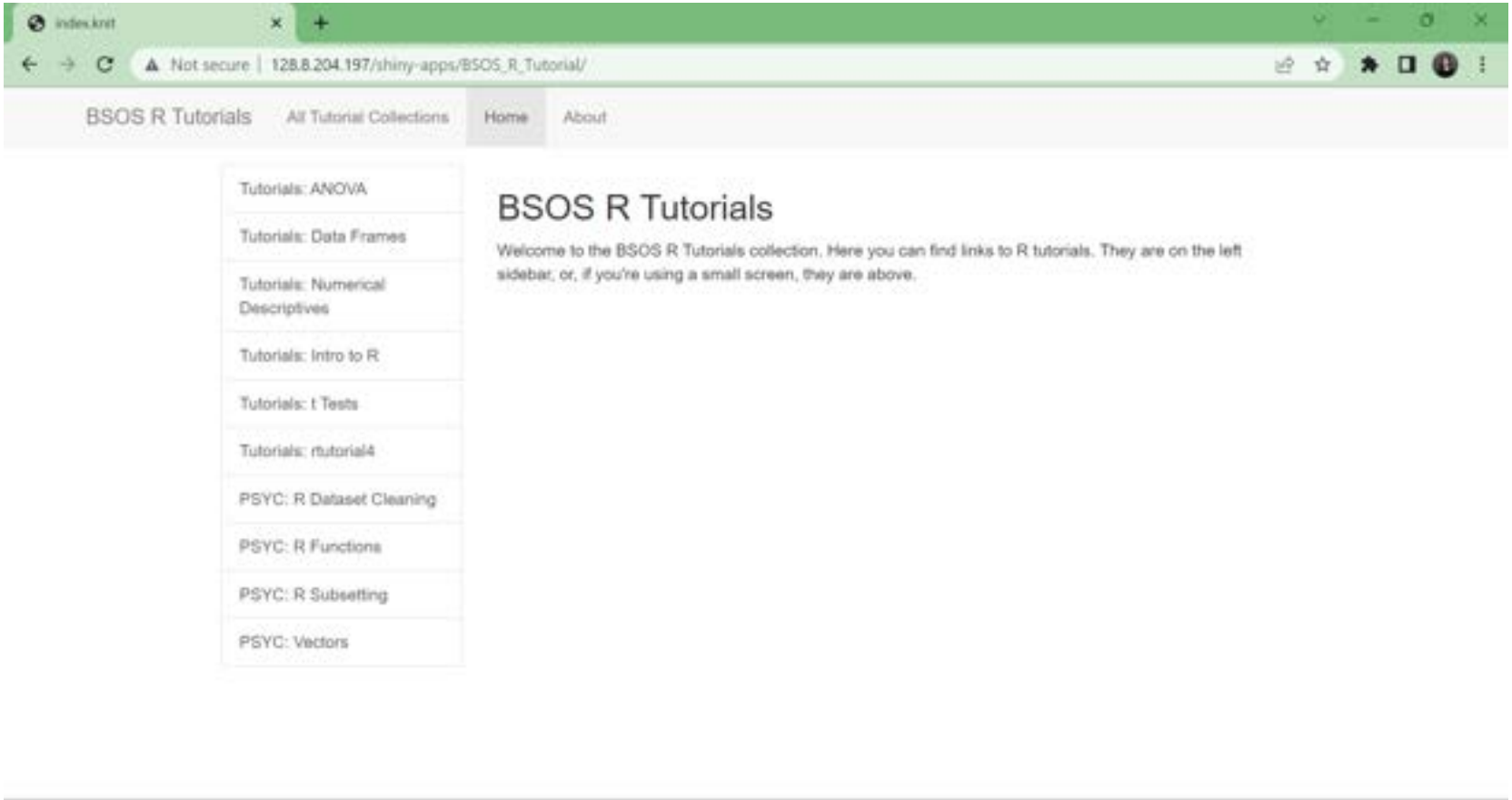
User Interface



Look and Feel (1)



Look and Feel (2)



Look and Feel (3)

Subsetting in R

Not secure | 128.8.204.197/shiny-apps/BSOS_R_Tutorial/PSYC_Tutorials/R%20Subsetting/

Subsetting in R

Section One : Extracting Rows and Columns

Section Two : Extracting & Building Vectors / Dataframes

Section Three : Subsetting

Start Over

Section One : Extracting Rows and Columns

In this module, we will learn some basic data set operations.

- We will start this lesson by using the class dataset we used in Lesson 3.
- Let's start by using the csv class data file.

Opening/loading Data set

```
dat.org <- read.csv("../data_psych/STAT200_Data_Update_4.8.csv")
dat.work <- dat.org

#Note: Make sure you are using readxl instead of readxl
#You will need to redo this part to open the dataset in your folder
```

Revision of Related Functions

In this section we will talk about a few exploratory functions that can help you explore the data set more. Carefully look at the following code chunk to understand what each function does.

```
#Name of the columns
colnames(dat.work) #See all the variable names in dataset
```

```
## [1] "Participant"    "Scores"         "Sleep"          "Motivation"
## [5] "Note_Type_Name" "Stat.Anxiety"   "Missing_Data"
```

Limitations

- Students do not get to see the RStudio environment
- Hides file path, can be difficult for novice students to understand the mechanism
- Cumbersome to program in some functionalities like a stopping rule for attempts or providing feedback properly if someone gets something wrong

BSOS

Data Curation Fellowship

COLLEGE OF
BEHAVIORAL &
SOCIAL SCIENCES

Data Curation Fellows Selection Process

- Had nearly 40 people apply!
- Applications were primarily from BSOS and Journalism, but we also had some CS students.
- We were happy to see such interest from students at different points in their college careers.

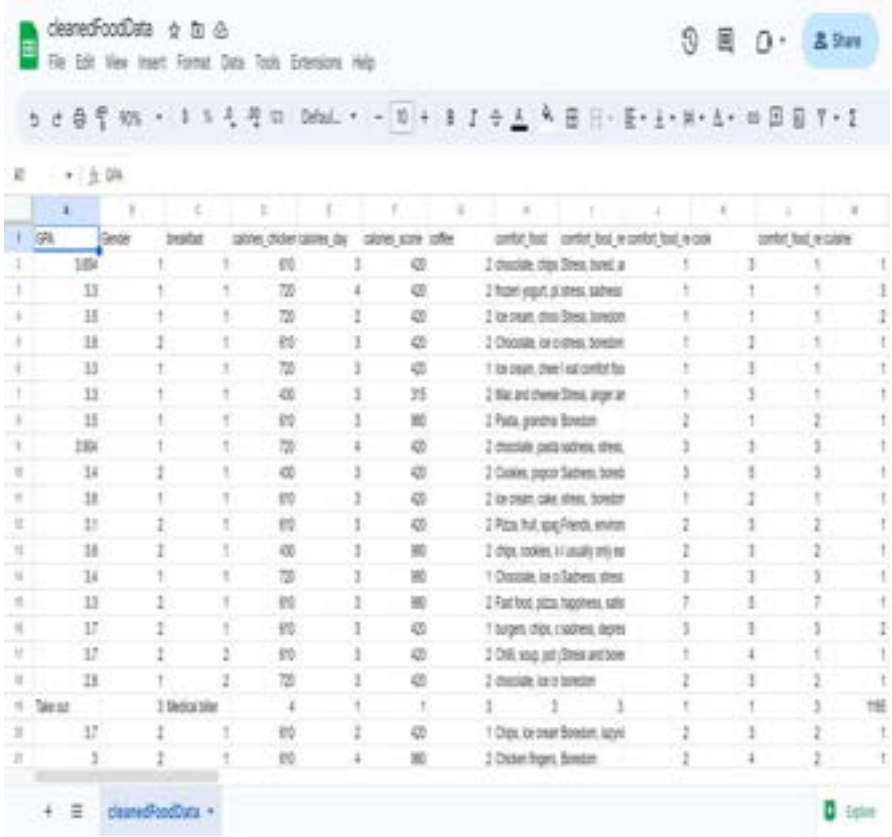
Training the Fellows

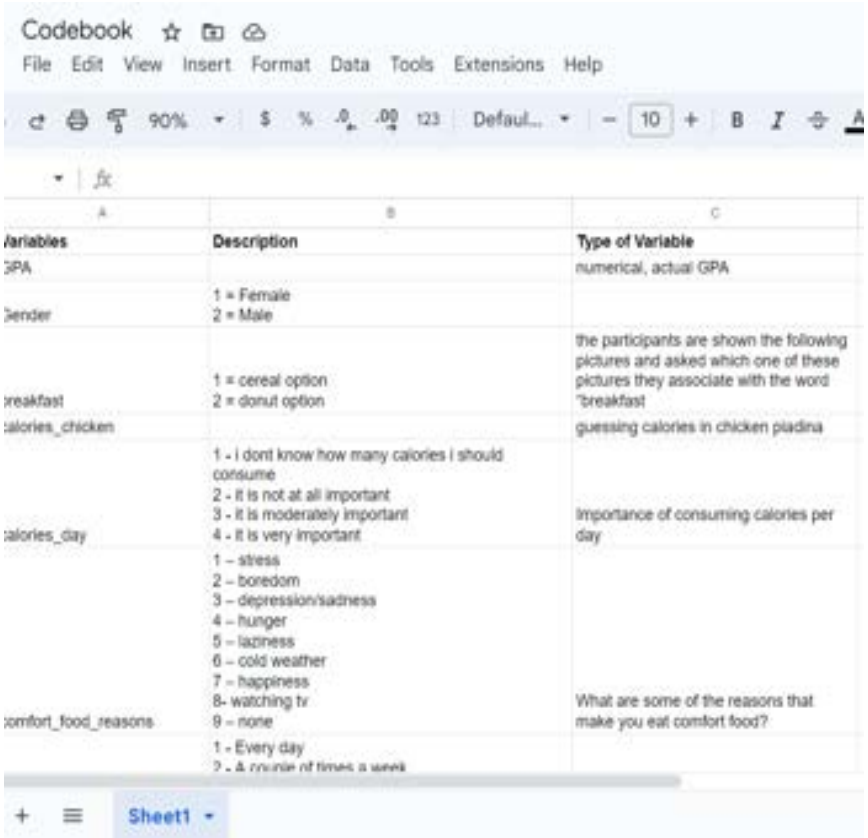
- 3 training sessions
  - 2 in person, one online
- Topics included:
  - Attribution
    - Data documentation
    - Codebooks
    - Methodology
  - Data cleaning
    - Missing data
    - Transforming variables
  - Practice with everyone cleaning the same dataset

Special Qualities of Our Program

- Done with an eye towards reproducible science
- Emphasis on DEI data and talking about what this means.
- Getting student curators to think like teachers as they choose dataset and clean them.
  - Nature of the variables
  - What types of analyses could these be used for?

Some Glimpses of Work of the Data Curation Fellows (1): Data on food consumption





Variables	Description	Type of Variable
GPA		numerical, actual GPA
Gender	1 = Female 2 = Male	
breakfast	1 = cereal option 2 = donut option	the participants are shown the following pictures and asked which one of these pictures they associate with the word "breakfast
calories_chicken		guessing calories in chicken piadina
calories_day	1 - i dont know how many calories i should consume 2 - it is not at all important 3 - it is moderately important 4 - it is very important	Importance of consuming calories per day
comfort_food_reasons	1 – stress 2 – boredom 3 – depression/sadness 4 – hunger 5 – laziness 6 – cold weather 7 – happiness 8- watching tv 9 – none	What are some of the reasons that make you eat comfort food?
	1 - Every day 2 - A couple of times a week	

Some Glimpses of Work of the Data Curation Fellows (2): Data from Twitter

codebook

Variables	Type	Type (additional)	Description
id	numeric	continuous	id of user account
user_name	text/string		account name
user_location	text/string		scraped location data (self-reported)
user_description	text/string		account bio information
user_created	date		date of account creation
user_followers	numeric	continuous	number of account followers
user_friends	numeric	continuous	number of account friends
user_favorites	numeric	continuous	number of favorites
user_verified	numeric	binary	binary value of whether account is verified (0 = FALSE; 1 = TRUE)
date	date		date of tweet
text/string	text/string		text/string of tweet
hastag.1	text/string		hastag text/string
hastag.2	text/string		hastag text/string
hastag.3	text/string		hastag text/string
hastag.4	text/string		hastag text/string
hastag.5	text/string		hastag text/string
hastag.6	text/string		hastag text/string
hastag.7	text/string		hastag text/string
hastag.8	text/string		hastag text/string
source	text/string		devise used to post tweet

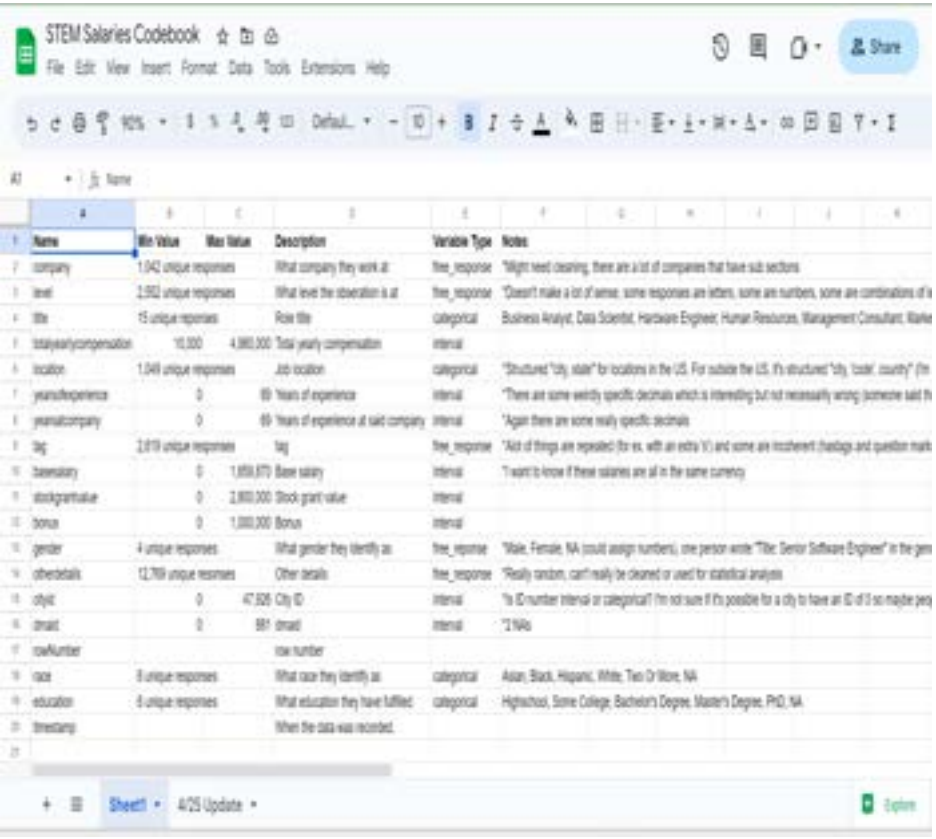
Sheet1

nonvoters_cleaned

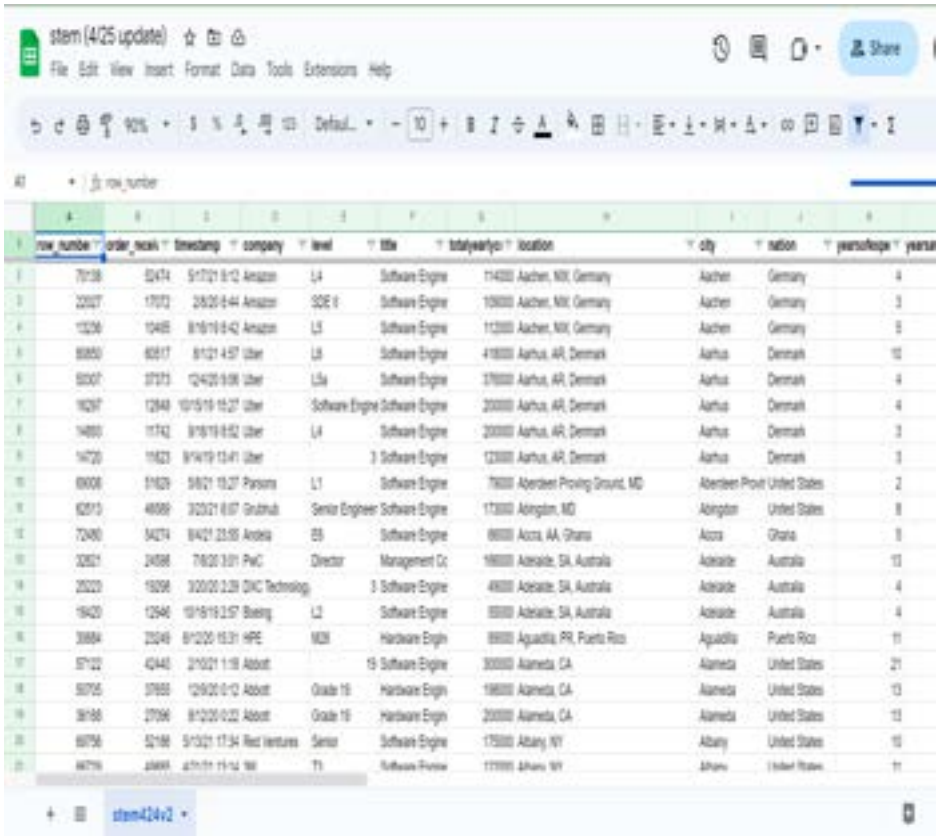
RespId	weight	Q1	Q2_1	Q2_2	Q2_3	Q2_4	Q2_5	Q2
470001	0.7516	1	1	1	2	4	1	
470002	1.0267	1	1	2	2	3	1	
470003	1.0844	1	1	1	2	2	1	
470007	0.6817	1	1	1	1	3	1	
480008	0.991	1	1	1	-1	1	1	
480009	1.0591	1	3	2	3	4	1	
480010	1.1512	1	1	1	2	3	1	
470008	1.0174	1	1	1	2	2	1	
470010	0.8184	1	1	1	1	3	1	
470011	1.1653	1	1	1	2	1	1	
470013	0.9517	1	1	2	2	2	2	
470016	1.374	1	1	1	1	4	1	
480020	0.8438	1	3	4	2	2	4	
470017	1.2745	1	1	1	3	3	1	
470018	0.9742	1	1	1	1	1	1	
470019	0.9893	1	1	1	1	1	1	
470022	0.8473	1	1	1	1	1	1	

nonvoters_cleaned

Some Glimpses of Work of the Data Curation Fellows (3): Salaries for people in STEM fields



Name	Min Value	Max Value	Description	Variable Type	Notes
company	1,042 unique responses		What company they work at	free_response	"Might need cleaning, there are a lot of companies that have sub sections
level	2,902 unique responses		What level the observation is at	free_response	"Doesn't make a lot of sense; some responses are letters, some are numbers, some are combinations of a
title	15 unique responses		Role title	categorical	Business Analyst, Data Scientist, Hardware Engineer, Human Resources, Management Consultant, Marke
totalyearlycompensation	10,000	4,980,000	Total yearly compensation	interval	
location	1,049 unique responses		Job location	categorical	"Structured "city, state" for locations in the US. For outside the US, it's structured "city, 'state', country" (I'm
yearsofexperience	0	69	Years of experience	interval	"There are some weirdly specific decimals which is interesting but not necessarily wrong (someone said th
yearsatcompany	0	69	Years of experience at said company	interval	"Again there are some really specific decimals
tag	2,819 unique responses		tag	free_response	"Alot of things are repeated (for ex. with an extra 'X') and some are incoherent (hastags and question marks
basesalary	0	1,850,870	Base salary	interval	"I want to know if these salaries are all in the same currency
stockgrantvalue	0	2,800,000	Stock grant value	interval	
bonus	0	1,000,000	Bonus	interval	
gender	4 unique responses		What gender they identify as	free_response	"Male, Female, NA (could assign numbers), one person wrote "Title: Senior Software Engineer" in the gen
otherdetails	12,769 unique responses		Other details	free_response	"Really random, can't really be cleaned or used for statistical analysis
cityid	0	47,926	City ID	interval	"Is ID number interval or categorical? I'm not sure if it's possible for a city to have an ID of 0 so maybe pep
timeid	0	991	timeid	interval	"2 NAs
rowNumber			row number		
race	8 unique responses		What race they identify as	categorical	Asian, Black, Hispanic, White, Two Or More, NA
education	6 unique responses		What education they have fulfilled	categorical	Highschool, Some College, Bachelor's Degree, Master's Degree, PhD, NA
timestamp			When the data was recorded		



row_number	order_num	timestamp	company	level	title	totalyearly	location	city	nation	yearsofexpe	yersa
70138	52474	5/17/21 9:12	Amazon	L4	Software Engine	114000	Aachen, NW, Germany	Aachen	Germany	4	
22007	17072	2/8/20 6:44	Amazon	SDE II	Software Engine	108000	Aachen, NW, Germany	Aachen	Germany	3	
13236	10485	8/16/19 8:42	Amazon	L5	Software Engine	112000	Aachen, NW, Germany	Aachen	Germany	5	
85850	60517	8/1/21 4:57	Uber	L8	Software Engine	418000	Aarhus, AR, Denmark	Aarhus	Denmark	10	
52007	37173	12/4/20 9:08	Uber	L5a	Software Engine	370000	Aarhus, AR, Denmark	Aarhus	Denmark	4	
16297	12844	10/15/19 15:27	Uber	Software Engine	Software Engine	200000	Aarhus, AR, Denmark	Aarhus	Denmark	4	
14800	11742	9/18/19 8:52	Uber	L4	Software Engine	200000	Aarhus, AR, Denmark	Aarhus	Denmark	3	
14720	11623	9/14/19 13:41	Uber		3 Software Engine	120000	Aarhus, AR, Denmark	Aarhus	Denmark	3	
69008	51629	5/8/21 15:27	Parsons	L1	Software Engine	78000	Aberdeen Proving Ground, MD	Aberdeen Provi	United States	2	
62513	46589	3/23/21 8:07	Grubhub	Senior Engineer	Software Engine	173000	Abingdon, MD	Abingdon	United States	8	
72480	54274	8/4/21 23:55	Andela	E6	Software Engine	66000	Accra, AA, Ghana	Accra	Ghana	8	
32621	24598	7/8/20 3:01	PwC	Director	Management Co	168000	Adelaide, SA, Australia	Adelaide	Australia	13	
25223	19298	3/20/20 2:28	DXC Technology		3 Software Engine	49000	Adelaide, SA, Australia	Adelaide	Australia	4	
15420	12546	10/16/19 2:57	Boeing	L2	Software Engine	85000	Adelaide, SA, Australia	Adelaide	Australia	4	
33684	25249	6/12/20 15:31	HPE	M28	Hardware Engin	89000	Aguadilla, PR, Puerto Rico	Aguadilla	Puerto Rico	11	
57122	42446	2/1/21 1:18	Abbott		15 Software Engine	300000	Alameda, CA	Alameda	United States	21	
50705	37855	12/9/20 0:12	Abbott	Grade 16	Hardware Engin	196000	Alameda, CA	Alameda	United States	13	
39168	27096	8/12/20 0:22	Abbott	Grade 16	Hardware Engin	200000	Alameda, CA	Alameda	United States	13	
68756	52168	5/13/21 17:34	Red Ventures	Senior	Software Engine	175000	Albany, NY	Albany	United States	10	
66776	49669	4/21/21 15:14	3M	T1	Software Engine	177000	Albany, NY	Albany	United States	11	

THANK YOU!

Dr. Brian Kim
kimbrian@umd.edu

Ujjayini Das
ujstat@umd.edu

Joe Hoskisson
jhoskiss@umd.edu

Dr. Sarah Croco
scroco@umd.edu