College Student Stress

Group 25: Gayatri Dutt, Veda Gadhiya, Jared Gutierrez, Mark Hashimoto, Nafisa Nikhath, Swetha Prabakaran

1.

The Problem

tl;dr - students are stressed. Very stressed.

61% of students

Feel overwhelming anxiety

40% of students

Too depressed to function

32% of students

say stress is biggest factor affecting performance

Source: American College Health Association
https://www.acha.org/documents/ncha/NCHA-II_Fall_2018_Reference_Group_Executive_Summary.pdf

Students are more stressed than ever

Increasing Demands:

- **Academic** stress
- **Social** pressures
- **Family** demands
- **Financial** demands
- Adolescent Psychological **Development**
- Future **Planning**
  - Ex. Internship and job search

Negative Impacts:

- Decreased **Academic** Performance
- **Burnout**
- Increased anxiety
- Higher susceptibility to **mental illnesses**
- Increased rates of physical illnesses
  - Ex: **autoimmune conditions** triggered by cortisol-triggered inflammation

2.

The Approach

“**Internet-based** clinical tools may be helpful in providing treatment to students who are **less inclined** to pursue services on campus”

“It’s **essential** that students know when it’s **time to reach out** for help”

Randy Auerbach (Columbia University Psychology Professor) and
Dr. Sherry Benton (current VP of the Society of Counseling Psychology of the APA, former director of the University of Florida counseling center)
https://www.cnbc.com/2018/10/04/4-ways-to-be-proactive-about-your-mental-health-in-college.html

Primary Goal

Help students **manage stress** by **identifying** potentially dangerous **stress levels** and provide **recommendations** on how to **balance/optimize** stress and performance.

Priorities

- Predict stress levels based on current habits
- Advise students on when to seek professional help
  - Encourage timid students to access services
  - Prevent crisis episodes by proactively seeking help
- Educate students on lifestyle changes
- Prevent mental or physical health deterioration

Stretch Goals

- Predict academic performance based on stress
    - Help answer the question of if the stress is “worth it”
- Predict expected stress during ‘optimal’ performance

Desired End Product

Web application where students enter **current habits and mental state**, and receive **recommendations** (and eventually, predictions).

Data Set

- Dartmouth College StudentLife dataset
- Chosen because:
  - Publicly **available**
  - **Reliable** source
  - Good **starting point** (high caliber university, high stress major)
  - **12 GB** data
    - Information on different **aspects of student life** including:
      - Activities/Lifestyle *(ex. physical activity, conversations, sleep)*
      - Academics *(ex. grades, deadlines, piazza)*
      - Personality *(ex. personal outlook, loneliness)*
      - Large variety of responses
- **Observations** = Students
- **Features** = Student life aspects

Data Collection Methodology

- Researchers:
  - From Dartmouth College
  - Contributions from researchers at University of Texas - Austin and Northeastern University
- Computer science students
- Spring semester at Dartmouth College
- Collected via:
  - App/sensors on android phones
  - Surveys administered before and after the semester
- Student identities removed and replaced by user id numbers (UID)

Data Cleaning

Problems:

- File/Data Structure
  - **Decentralized**, data distributed in many directories
  - Data stored in various **types** (csv, json)
  - **Temporal** differences - hourly vs daily vs semester
  - **Inconsistencies** in row sizes for different users, responses for EMA are up to

Techniques:

- Merge
  - Create **functions** to merge by UID
  - Convert to **uniform** time
  - Creation of aggregator functions

Data Cleaning

Problems:

- Mixed Data Types
  - **Strings** in survey responses
    - Ex. Pre/Post, Agree/Disagree
  - **App/Sensor** data
    - Ex. Yes/No, # of people, hrs sleep, mins activity
  - **Positive vs negative** aspects of stress
  - Text response values, non categorical

Techniques:

- Translate Data → Numeric Scale
  - **Survey** Data
    - 1-5 scale
  - **App/Sensor** data
    - Create keys/bins
  - **Categorize** each question on if it suggests a **+ or -** outlook
    - Inverted the numeric values if negative

Data Cleaning

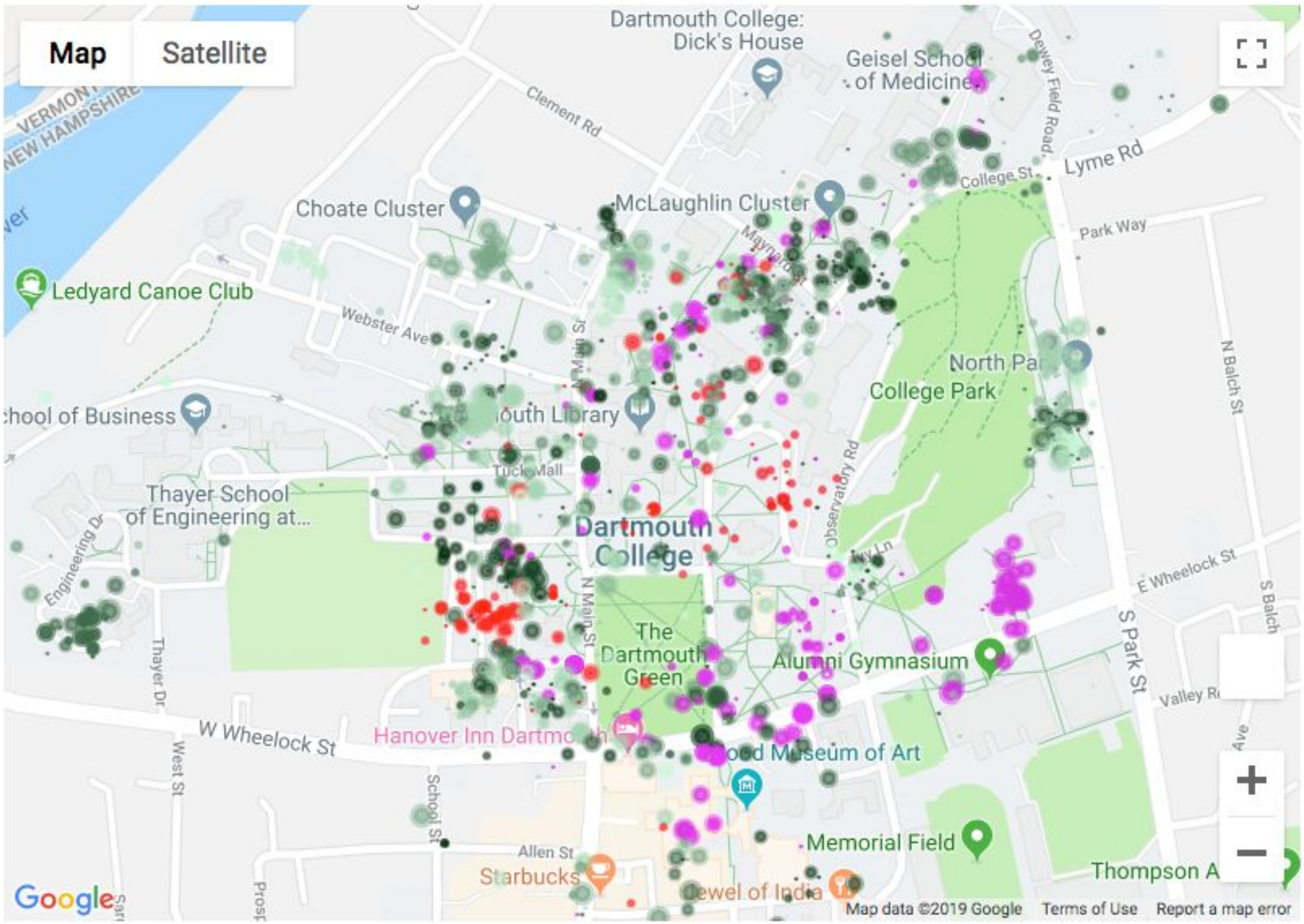
Problems:

- **Missing data**
  - Missing different **UIDs** (observations) in various files
    - Trade-off between more features and more observations
  - Missing certain **feature values** in files
    - Missing values of different types
    - Drop or figure out appropriate way to replace

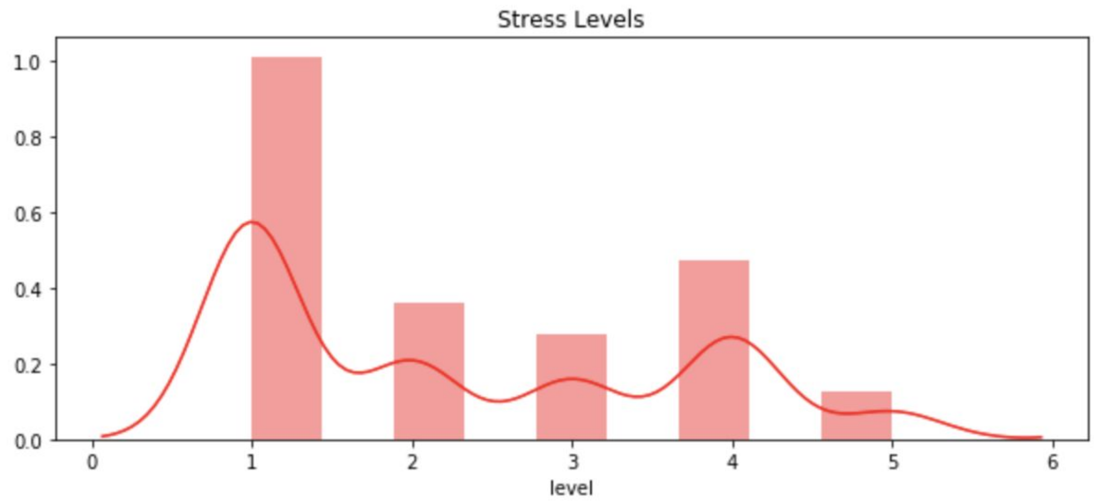
Techniques:

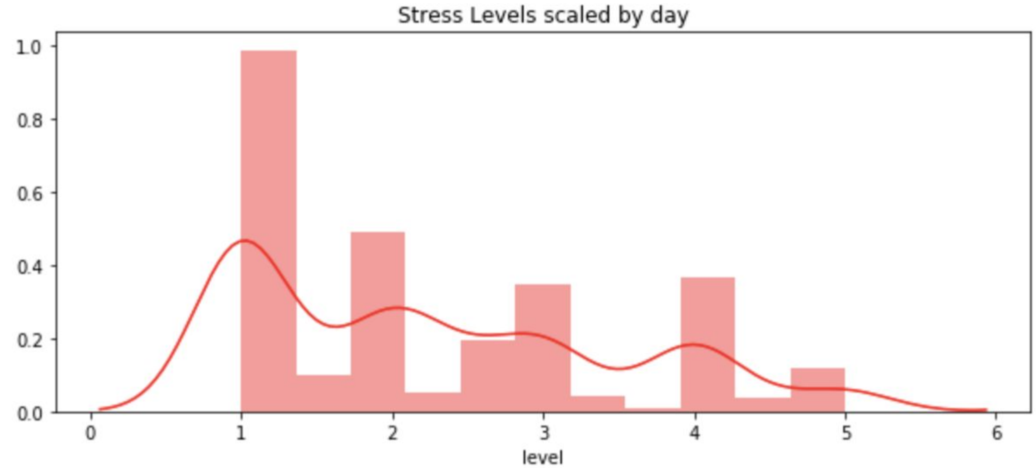
- **Drop**
  - UIDs missing from multiple files
  - Columns with too many NaNs (threshold)
- **Fill** remaining NaNs with various methods
  - Ex. Agree/Disagree 1-5 scale (Big 5)
    - Fill NaN with 3 (neutral)
  - Ex. Sleep (PSQI)
    - Fill NaN with median

EDA - Locations Stress Reported

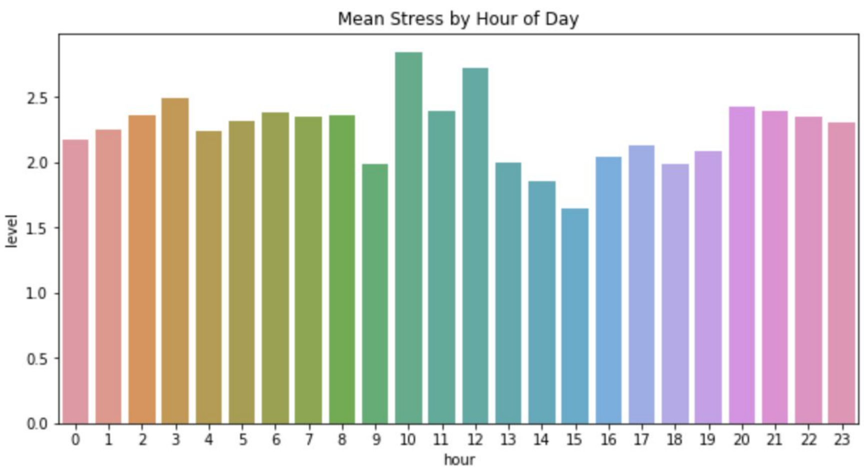


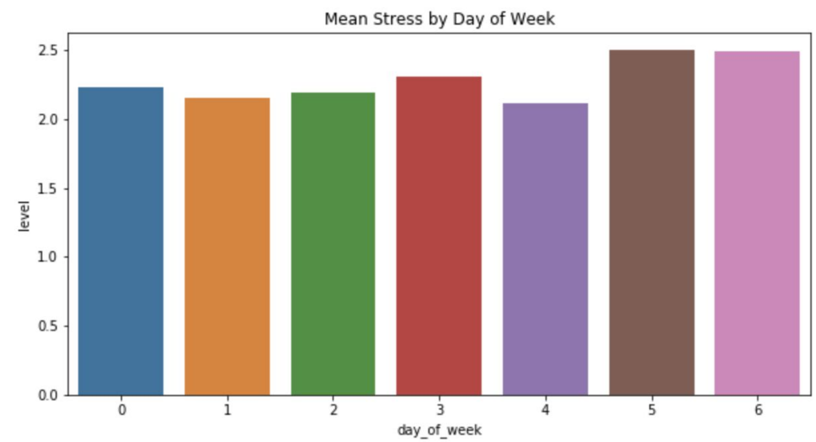
EDA - Stress Levels

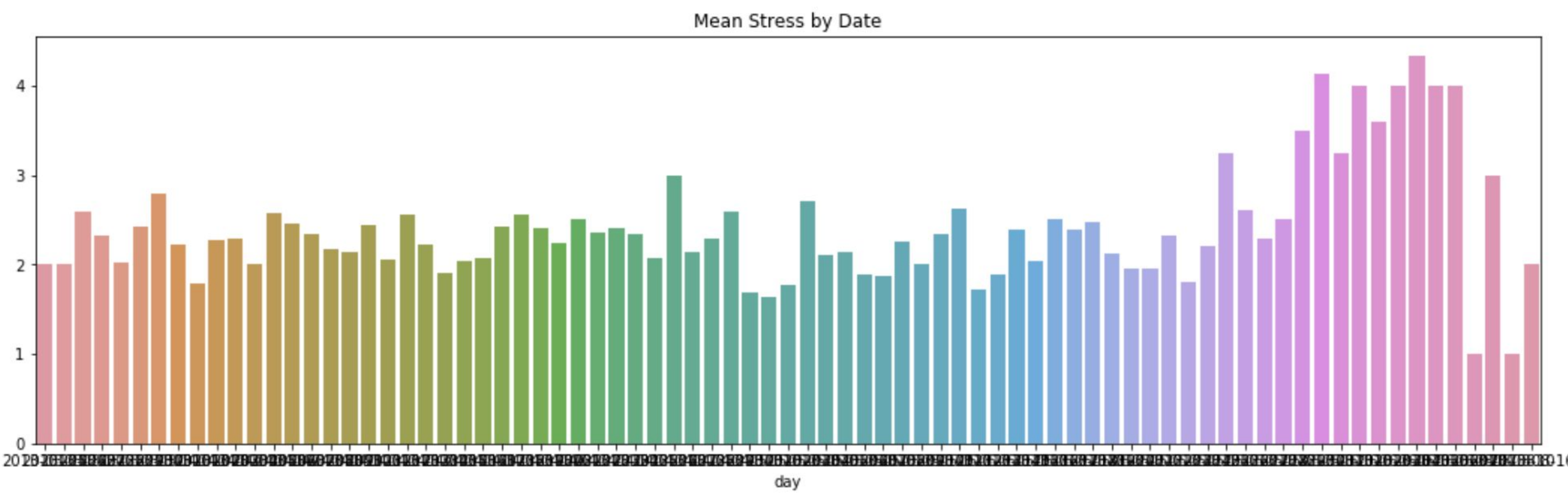




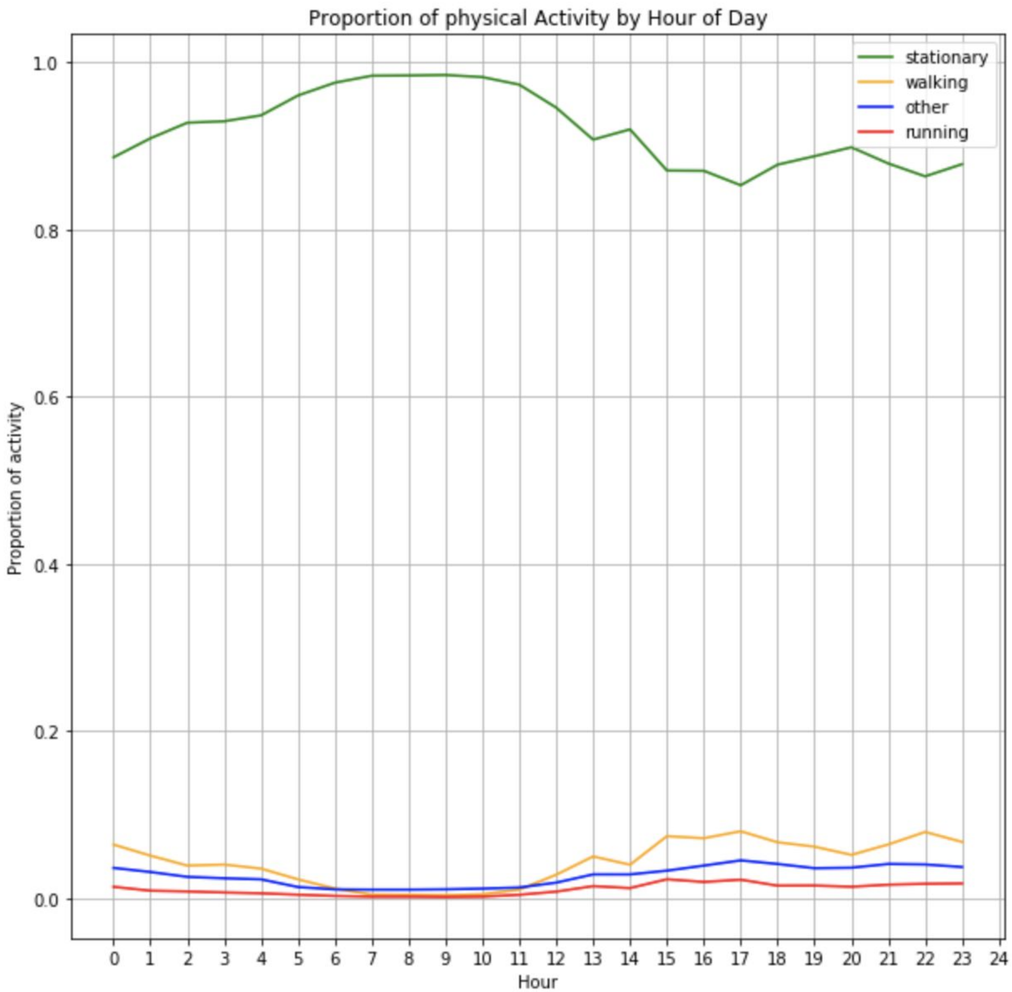
EDA - Stress Levels



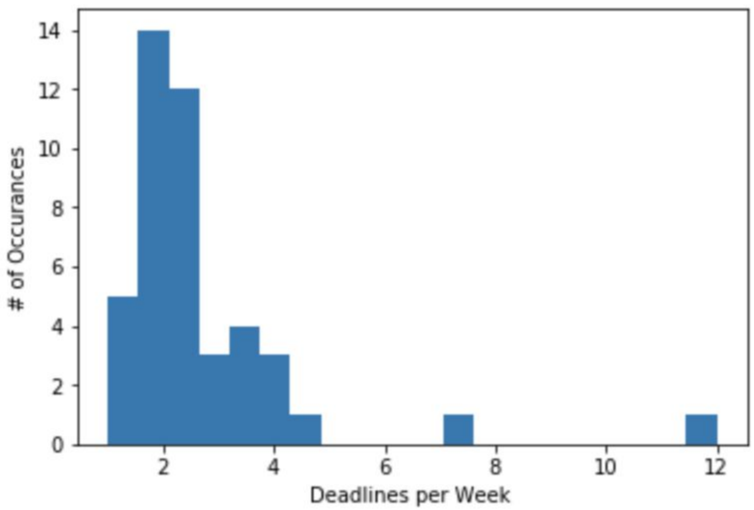


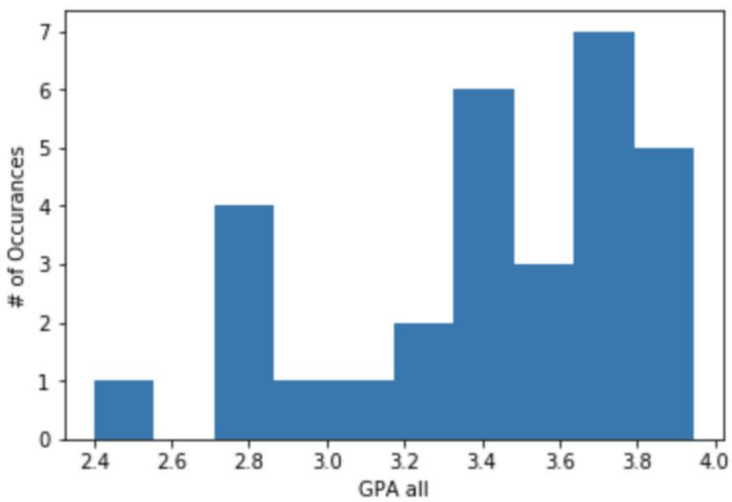


EDA - Activity Levels

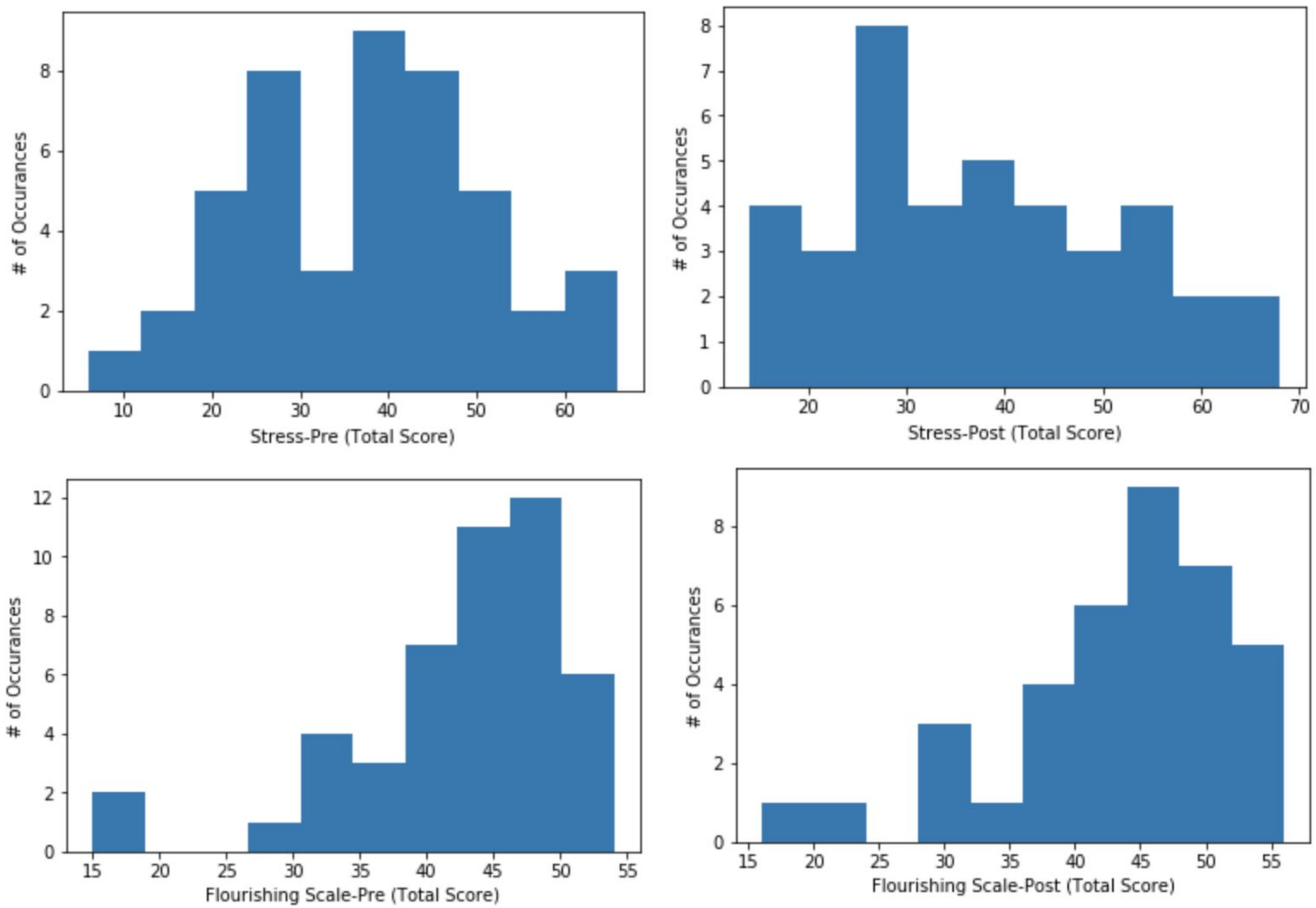


EDA - Some Academics





EDA - Some Pre/Post Survey Results



EDA - Takeaways

- Students have a similar outlook when comparing their pre and post **self-perceived success**.
    - **Sensor and App** data can help make up for this
- Recorded levels of stress are highly **concentrated** in the middle
- Best type of questions that can identify mood are those that are extreme and more factual
- Interquartile range of **deadlines** is 5-30
    - upper end of deadlines shows a stronger correlation to stress levels

Models

- Experimented with different models
  - Linear Regression
  - Logistic Regression
- Feature Reduction
  - First built models that used all features
  - Progressively experimented by removing a feature or condensing a feature and seeing how it affected accuracy
    - Lasso Regression
    - Feature Reduction
    - K-Cross Validation
- Built models to predict stress levels and classify students into different stress buckets

Feature Correlation

- By Analyzing student's responses we were able to develop a correlation chart highlighting how responses to a particular question affected another mood.

- This allowed us to see what the most influential moods are and which questions best reflect student's outlook

Feature Correlation

		flour		lonely		Interested		Distressed		Upset		Strong		Guilty
	type	post	pre	post	pre	post	pre	post	pre	post	pre	post	pre	post	pre
	type														
flour	post	1	0.549	-0.467	-0.423	0.124	0.365	-0.347	-0.0189	-0.31	0.0645	0.165	0.00643	0.128	-0.0512
	pre	0.549	1	-0.527	-0.424	0.4	0.269	-0.282	-0.0697	-0.369	0.109	0.248	0.0998	-0.133	0.153
lonely	post	-0.467	-0.527	1	0.802	-0.298	-0.298	0.501	0.0667	0.132	-0.088	-0.274	-0.147	0.0213	-0.0867
	pre	-0.423	-0.424	0.802	1	-0.273	-0.0694	0.469	0.163	0.157	0.00804	-0.231	-0.124	-0.0109	-0.105
Interested	post	0.124	0.4	-0.298	-0.273	1	0.239	-0.377	-0.031	-0.403	0.058	0.42	-0.152	-0.00408	0.336
	pre	0.365	0.269	-0.298	-0.0694	0.239	1	-0.193	0.0803	-0.104	-0.000731	0.0217	0.432	0.0466	-0.0342
Distressed	post	-0.347	-0.282	0.501	0.469	-0.377	-0.193	1	0.287	0.68	0.0894	-0.276	-0.0433	0.0908	-0.326
	pre	-0.0189	-0.0697	0.0667	0.163	-0.031	0.0803	0.287	1	0.27	0.741	-0.061	-0.0326	0.132	0.194
Upset	post	-0.31	-0.369	0.132	0.157	-0.403	-0.104	0.68	0.27	1	-2.54e-17	-0.283	0.0645	0.0663	-0.177
	pre	0.0645	0.109	-0.088	0.00804	0.058	-0.000731	0.0894	0.741	-2.54e-17	1	0.0694	-0.103	0.102	0.169
Strong	post	0.165	0.248	-0.274	-0.231	0.42	0.0217	-0.276	-0.061	-0.283	0.0694	1	0.133	0.00826	0.15
	pre	0.00643	0.0998	-0.147	-0.124	-0.152	0.432	-0.0433	-0.0326	0.0645	-0.103	0.133	1	0.24	-0.0891
Guilty	post	0.128	-0.133	0.0213	-0.0109	-0.00408	0.0466	0.0908	0.132	0.0663	0.102	0.00826	0.24	1	0.195
	pre	-0.0512	0.153	-0.0867	-0.105	0.336	-0.0342	-0.326	0.194	-0.177	0.169	0.15	-0.0891	0.195	1
Scared	post	-0.067	-0.417	0.0975	0.176	-0.305	-0.0224	0.198	0.294	0.357	0.143	-0.101	0.0973	0.388	0.0544
	pre	-0.129	-0.0111	-0.156	-0.0867	0.252	0.168	-0.103	0.356	0.0377	0.463	0.112	0.0888	0.0163	0.405
Hostile	post	-0.311	-0.554	0.422	0.361	-0.239	-0.222	0.393	0.323	0.442	0.19	0.0708	0.258	0.0526	-0.251
	pre	0.118	0.134	-0.00203	-0.0695	0.186	0.256	-0.0156	0.435	3.11e-17	0.47	0.141	0.362	0.231	0.453
Enthusiastic	post	0.0445	0.0266	-0.146	0.0663	0.459	0.305	-0.166	0.0379	0.0154	-0.0312	0.251	0.216	-0.151	0.0668
	pre	0.229	0.229	-0.366	-0.178	-0.0016	0.469	-0.136	-0.098	0.106	-0.0954	0.277	0.523	-0.138	-0.068
	post	0.244	0.206	-0.173	0.0871	0.533	0.25	-0.256	0.141	-0.199	0.0417	0.448	0.118	-0.0963	0.192

Feature Importance Chart

LASSO

	feature	weight
0	vr_12_ment	-0.817026
1	flour	-0.073805
2	vr_12_phys	0.039093
3	Conscientiousness	-0.735498
4	Active	-0.275755
5	Alert	-0.956694
6	Neuroticism	0.413749
7	Upset	-0.000000
8	Nervous	0.000000
9	phq_total	0.144005
10	Hostile	0.798189

Correlation Matrix

type	level_0	type	post
0	vr_12_ment	post	-0.765781
1	Interested	post	-0.520546
2	Active	post	-0.507939
3	Alert	post	-0.434042
4	flour	pre	-0.411800
5	Conscientiousness	post	-0.387363
6	Attentive	post	-0.386443
7	vr_12_ment	pre	-0.380732
8	vr_12_phys	post	-0.351661
9	flour	post	-0.349798

type	level_0	type	post
0	Hostile	post	0.431651
1	Neuroticism	pre	0.461548
2	Distressed	post	0.476272
3	Nervous	post	0.478903
4	Upset	post	0.541151
5	Neuroticism	post	0.581124
6	phq_total	post	0.632501
7	stress	pre	0.634861
8	phq_total	pre	0.699325

Model Performance

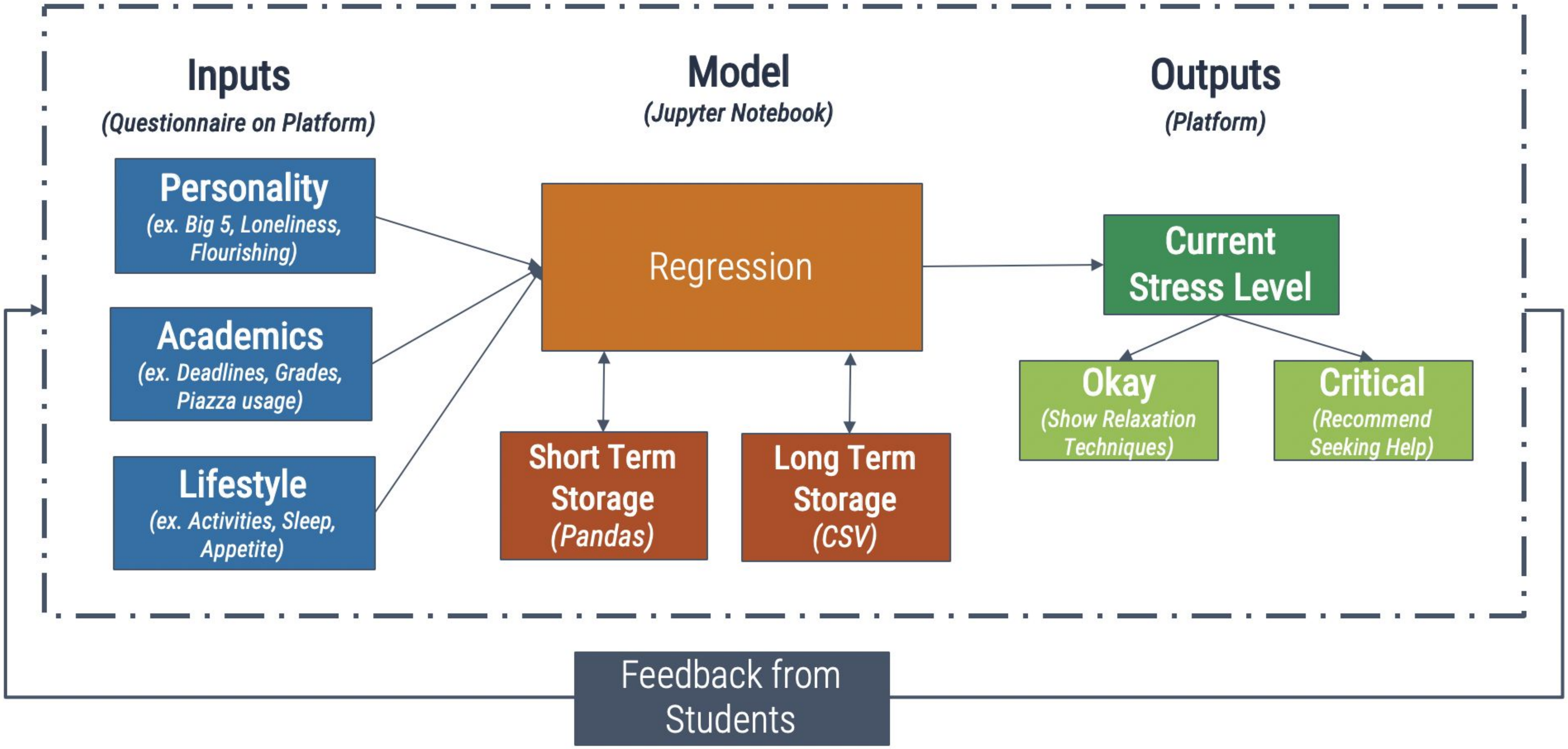
Regression	Type	Hyperparameter	MSE	MAE
Linear Regression	All Survey		19.035	3.602
		Lasso Alpha = 0.004	15.792	3.713
	Feature Reduction		15.877	2.985
		Lasso Alpha = 0.17	14.420	2.779

Model	Type	Hyperparameter	Accuracy	Notes
Logistic Regression	All Survey	C = 1	72%	No predictions were over 1 off of the true value Test: 0.2 Train: 0.8
	Feature Reduction	C = 10	73%	Predictions were skewed toward the middle of Test: 0.2 Train: 0.8

3.

Architecture

System Overview



4.

User Interface

User Interface

Try it out live!

https://dataxfinalproject.bubbleapps.io/version-test?debug_mode=true

5.

Learning Path and Future

Learning Path

- Finding Data
  - Gaining access
    - American College Health Association – National College Health Assessment (ACHA-NCHA)
  - Reliable source, relevance, feature/response rich
    - Kaggle Ideal Student Life Survey (Singapore students, more observations but less features/no performance data and only binary responses)
  - Trade-offs: Features/ Observations/Response types
- Cleaning Data
  - Tried both Kaggle and Dartmouth sets
  - Scope
    - Initially limited scope to Dartmouth survey data only due to difficulties in cleaning but added sensor and app data back in
  - Merging, translating, dealing with NaNs, how to process
  - Trade-offs: Features vs Observations vs Quality

Learning Path

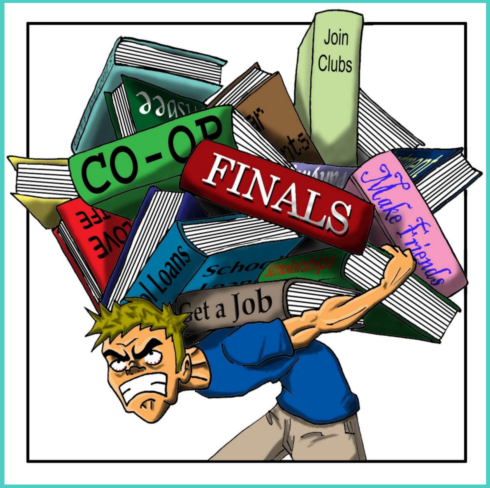
- EDA
  - Understanding the data and useable features
  - Plotting
- Building Models
  - Stress is abstract/subjective, difficult assign numeric/categorize
  - No clear cut path
- Model Improvement
  - Metrics and tuning parameters - trial and error
- Creating a UX Interface prototype
  - Figuring out platform
  - What to include

Future Exploration

- UX Interface
  - Create and Link Interface to Model
- Add Feedback from Students
- Try Other Models
- Attempt Reach Goals
- Add more data
  - Continuing study at University of Texas - Austin and Northeastern University

Thank You!

Questions?



6.

Appendix

Links

- GitHub
  - https://github.com/mkhash/Modelling-Student-Stress
- Dataset
  - https://studentlife.cs.dartmouth.edu/dataset.html?fbclid=IwAR3EvHteBs3EaG4XjXqtPA6j6T5Z2saA4Ml0DLZScpQB90Wm_fhUmQznJ0Y#sec:data_dir:survey_dir