Merna Mostafa mam1192
Tejaswi Tripathi tt507
CS 210 Final Project Proposal

- **Project Definition: What problem are you solving? What strategic aspects are involved? How does your project relate to the lectures/papers we discussed?**

With this project, we seek to assess the correlation between the accessibility of demonstrated screen time limiters (in this case, physical activity in the form of parks) and user screen time. We can also determine some other interesting relationships: does increased accessibility to parks reduce screen time spent overall, or does it reduce time spent on a specific app class but leaves the others unchanged? We will need to find datasets on user data with classifications on apps, as well as location data. Additionally, we will need to find data on park scores for each US city to determine accessibility to physical activity. This project combines all aspects of data management, from sourcing to cleaning to analysis.

- Novelty and Importance: Why is your project important? Why are you excited about it? What are some existing issues in current data management practices? Are there any prior/related works? Provide a brief summary.

In the past decade, there have been many studies demonstrating the positive link between screen time and stress and anxiety, as well as depression. According to Deloitte's 2023 Connected Consumer survey, there is a real desire for the consumer to limit their screen time. However, it is unclear how much people actually adhere to the limitations they set on themselves. This leads to plenty of bias in the results; additionally, according to this article by Techno Sapiens, people tend to overstate how effective self-imposed screen time limits set through device features actually are, and only 12% of people use those limits anyways. Because of all this, a user's adherence to screen time limitations is a confounding variable in measuring the effectiveness of such limitations. However, there still is a way to determine the effectiveness of screen time limits on actual screen time. According to this 2010 study, as well as plenty of research since then, physical activity is generally the best way to limit screen time. The next question to ask, as well as the question we seek to answer, is whether or not a user's *accessibility* to physical activity contributes to limiting screen time. Although the device features are arguably more accessible than physical activity would ever be yet demonstrate little effects on screen time, it would psychologically make more sense for physical activity to have a stronger effect. The device features simply remove a positive stimulus (phone usage), but physical activity replaces a positive stimulus with another.

We believe that answering this question will help us determine what we can do as a society to collectively limit screen time and, in turn, depression among users. If we find a positive relationship, it's possible that increasing a consumer's accessibility to parks by investing more in parks, which are hubs for physical activity, would have a measurable impact on their well-being by reducing the amount of time spent on their phone.

- Plan: Be specific and succinct.

- To carry out this task, we will combine two datasets:

    - The first dataset consists of user data with screen time, percentage of screen time spent on particular app classes (gaming, social media, and productivity), and user location (US city); this dataset comes from Kaggle and is licensed for community usage.
    - The second dataset consists of park score rankings for the top 100 most populated US cities: this dataset also contains park accessibility, median park size, parkland as a percentage of city area, and park amenities for each city. This second dataset comes from the Trust for Public Land website.
- We will group each user data in the first dataset by city location, then calculate the median for each relevant column therein.
- Once all data is cleaned, we will make various plots, one for each column in the park dataset, plotting park score against the median user screen time data, and calculate relevant correlation coefficients.
- Finally, for the strongest correlations we find, we will split the data into training and testing to perform regression analysis.
    - Ultimately, we seek to determine which feature(s) best predicts screen time data.

We will use SQL here to group the user data by city , join the datasets and clean them by filtering out cities where park scores or median screen time values are missing or invalid. Then we will extract the relevant columns in plotting and calculating the correlation coefficients.

- What models/techniques/algorithms do you plan to use or develop?
We plan to make use of Pandas for the primary data cleaning. We will remove outliers, handle missing values, and filter by relevant columns.

We will make use of the SQL JOIN command clause. This will help us to merge the two datasets on city location easing the process to compute medians for screen time per city.

We will also perform exploratory analysis by using the correlation matrices and visualizations (scatter plots) to see patterns between screen time and park metrics

We will use SciPy for train-test split and regression analysis for park metrics that show a strong correlation, since it will help to predict screen time based on park accessibility.

Implementation steps:

First we will load the data sets into a SQL database in order to manage and combine them by city.

Then we will clean the data filtering it for consistency and removing any outliers.

Then we will use exploratory analysis to see which park feature(s) relate the most to screen time.

Then we will build the model using correlation and regression analysis to predict the relationships between park accessibility and screen time.

We will test the model with data and see its accuracy and error rates.

Then present the findings in graphs or tables to show what features have the strongest relationship with decreased screen time.

This will answer our question of does increased accessibility to parks reduce screen time spent overall, or does it reduce time spent on a specific app class but leaves the others unchanged?

Testing success:

Strong relationship between park access and screen time - meaning that cities with more parks , larger parks , closer parks , etc have lower average screen time.

An accurate model explaining or predicting screen time based on park accessibility - this means that the model will help us predict how much average screen time one might have in a city based on park accessibility.

Datasets to consider:

- Mobile device usage, with app classification (social media, gaming, productivity) and location (US city) (2024)
    - https://www.kaggle.com/datasets/bhadramohit/smartphone-usage-and-behavioral-dataset
- City walkability and urban connectedness (walk score and bike score specifically)
    - https://www.kaggle.com/datasets/vellis1/us-cities-urban-connectivity
- US Park score for top 100 most populated cities (2024)
    - Includes median park size, parkland as a percentage of city area, park accessibility, park amenities, overall score
    - https://parkserve.tpl.org/downloads/historic/2024_ParkScoreRank.pdf

Useful studies to consider:

- https://publications.aap.org/pediatrics/article/126/1/e89/68258/Influence-of-Limit-Setting-and-Participation-in?autologincheck=redirected
    - Consistent restrictions / limits in electronics usage in youth generally leads to less screen time (only applies to young ppl, like kids)
    - Physical activity is the best way to reduce screen time in kids
- https://www.statista.com/chart/30968/measures-taken-to-manage-screen-time/
    - There is a desire for limited screen time, but only 12% of those surveyed used screen time limits set up with device features. Additionally, there is questionable adherence to limitations and restrictions.