Merna Mostafa mam1192 (#07)
Tejaswi Tripathi tt507 (#07)
CS 210 Final Project Writeup
https://github.com/tejaswitripathi/parkscores-with-population
9 December 2024

Background

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?

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 (Richter 2023). 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 a 2023 study, 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 (Silverman et al 2024). Consequently, 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 a 2010 study, as well as plenty of research since then, physical activity is generally the best way to limit screen time (Carlson et al 2010). Unlike self-imposed limits, which simply remove a positive stimulus (phone use), physical activity introduces an alternative positive stimulus, potentially making it a more impactful treatment. This study aims to explore whether accessibility to physical activity opportunities, such as parks, influences health app usage, a proxy for physical activity.

Our data are primarily sourced from the 2024 US Park Score Index Rankings, as of March 2024. This database ranks cities based on a number of factors, such as park accessibility, investment, acreage, and park amenities. The health app usage data is not publicly available. Therefore, we generate synthetic data (approximately 200,000 records of user data) for the 100 cities in the 2024 Park Score Rankings, giving us 2,000 records per city. This synthetic dataset is not as comprehensive or accurate as we were hoping, and future research should utilize better data.

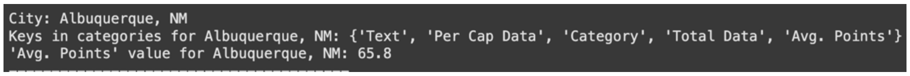
We believe that answering this question will help us determine what we can do as a society to collectively limit screen time and, in turn, depression among users. If we find a strong relationship, it's possible that increasing a consumer's accessibility to parks by investing more in parks, which are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

Method

1. Data Preprocessing & Cleaning

Our first data are a database consisting of the 2024 Park Score Rankings for the top 100 most populated US cities. Each city (entry) in the database has 6 categories – Access, Acreage, Amenities, Equity, Investment, and Total. For each category, the database stores a number of points, a ranking, a description, and a null column.

We load this database from a JSON file to a Python dictionary. To ensure that the database has been correctly processed into the dictionary, we print each entry:



```
City: Albuquerque, NM
Keys in categories for Albuquerque, NM: {'Text', 'Per Cap Data', 'Category', 'Total Data', 'Avg. Points'}
'Avg. Points' value for Albuquerque, NM: 65.8
```

We then utilize MySQL to store the dictionary as a database 'parkscores.db':

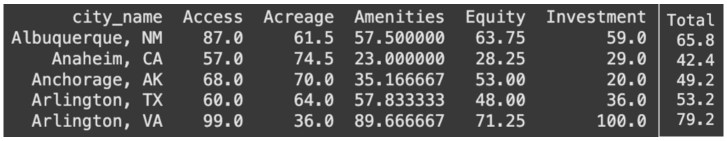
```
Data successfully stored in parkscores.db
```

Next, we perform exploratory data analysis and data cleaning to establish usability. To clean this data, we remove all columns within each category except for the number of points since that is the only relevant one to our analysis. We ensure that all numeric values are floats by removing all quotes and any other symbols from the points entries.

We continue using MySQL queries to create 6 separate tables for each category, where each entry in a given table is a city, and the column corresponds to the category. The data entered into the table is the points awarded for the given category.

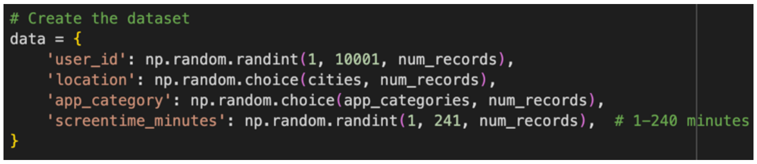
```
Database updated: 'Total Data' column removed, and separate tables created for each category.
```

Finally, we store each table into a single Pandas dataframe called 'df_parkscores' using the "groupby" and "pivot" operations. The Park Scores data have been preprocessed.

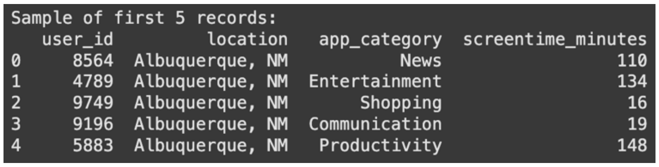


```
        city_name  Access  Acreage  Amenities  Equity  Investment  Total
Albuquerque, NM    87.0     61.5  57.500000   63.75        59.0   65.8
    Anaheim, CA    57.0     74.5  23.000000   28.25        29.0   42.4
  Anchorage, AK    68.0     70.0  35.166667   53.00        20.0   49.2
  Arlington, TX    60.0     64.0  57.833333   48.00        36.0   53.2
  Arlington, VA    99.0     36.0  89.666667   71.25       100.0   79.2
```

Our second data are synthetic data containing user screen time for the 100 cities in the Park Scores data. We start by defining app categories: "Social Media", "Entertainment", "Productivity", "Gaming", "Education", "Health & Fitness", "Shopping", "Communication", "News", "Travel". We set our random seed to 42 and implement 10,000 unique users, and we generate our dataset:



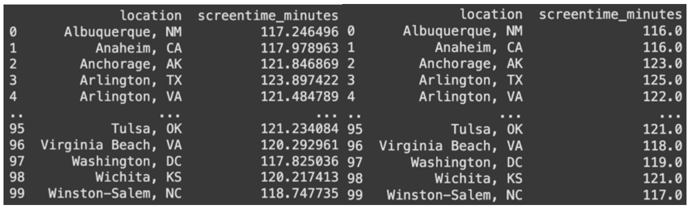
```
# Create the dataset
data = {
    'user_id': np.random.randint(1, 10001, num_records),
    'location': np.random.choice(cities, num_records),
    'app_category': np.random.choice(app_categories, num_records),
    'screentime_minutes': np.random.randint(1, 241, num_records),  # 1-240 minutes
}
```



```
Sample of first 5 records:
   user_id          location   app_category  screentime_minutes
0     8564  Albuquerque, NM           News                 110
1     4789  Albuquerque, NM  Entertainment                 134
2     9749  Albuquerque, NM       Shopping                  16
3     9196  Albuquerque, NM  Communication                  19
4     5883  Albuquerque, NM   Productivity                 148
```

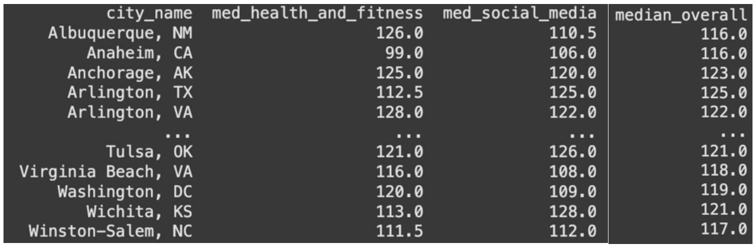
We perform an exploratory data analysis to determine how to use the dataset. We have 200,000 total records across 100 locations, giving us 2,000 records per location. Since the data were generated randomly, we can assume normality, and we verify using descriptive statistics:

print(df_screentime.groupby("location")["screentime_minutes"].mean().reset_index())
print(df_screentime.groupby("location")["screentime_minutes"].median().reset_index())



```
              location  screentime_minutes              location  screentime_minutes
0      Albuquerque, NM          117.246496 0      Albuquerque, NM               116.0
1          Anaheim, CA          117.978963 1          Anaheim, CA               116.0
2        Anchorage, AK          121.846869 2        Anchorage, AK               123.0
3        Arlington, TX          123.897422 3        Arlington, TX               125.0
4        Arlington, VA          121.484789 4        Arlington, VA               122.0
..                 ...                 ... ..                 ...                 ...
95           Tulsa, OK          121.234084 95           Tulsa, OK               121.0
96  Virginia Beach, VA          120.292961 96  Virginia Beach, VA               118.0
97      Washington, DC          117.825036 97      Washington, DC               119.0
98         Wichita, KS          120.217413 98         Wichita, KS               121.0
99   Winston-Salem, NC          118.747735 99   Winston-Salem, NC               117.0
```

Based on these results, we can proceed with this dataset. Our goal is to compare screen time to the points values, so we use median screen time. In addition to using total median screen time for each city, we will use median social media screen time and median health & fitness screen time. In recent years, users have been finding more ways to use their devices for positive mental health activities; aside from determining whether a better park score is correlated with reduced overall screen time, we will determine its correlation with social media screen time (sedentary activity) and health & fitness screen time (non-sedentary activity).

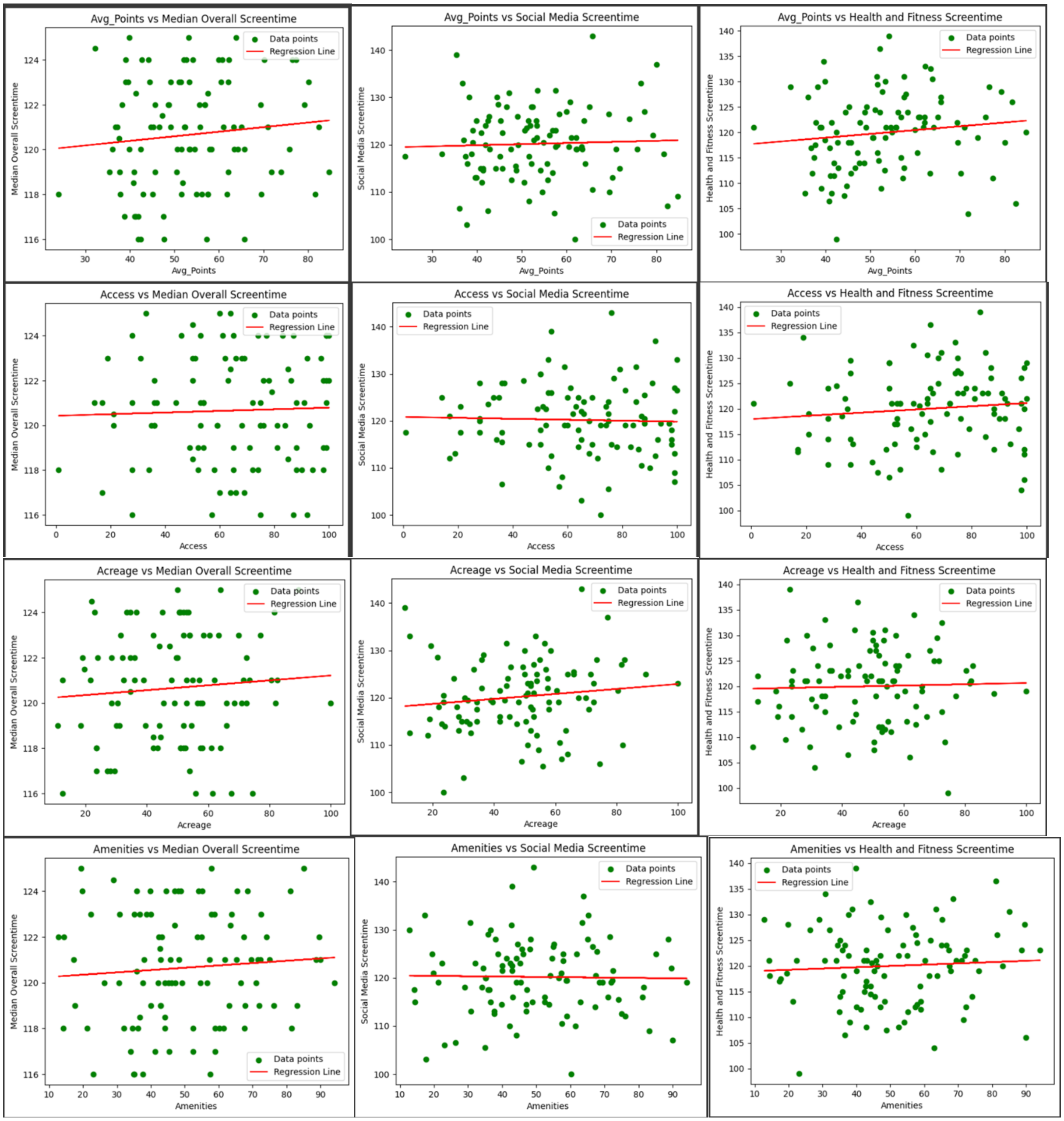


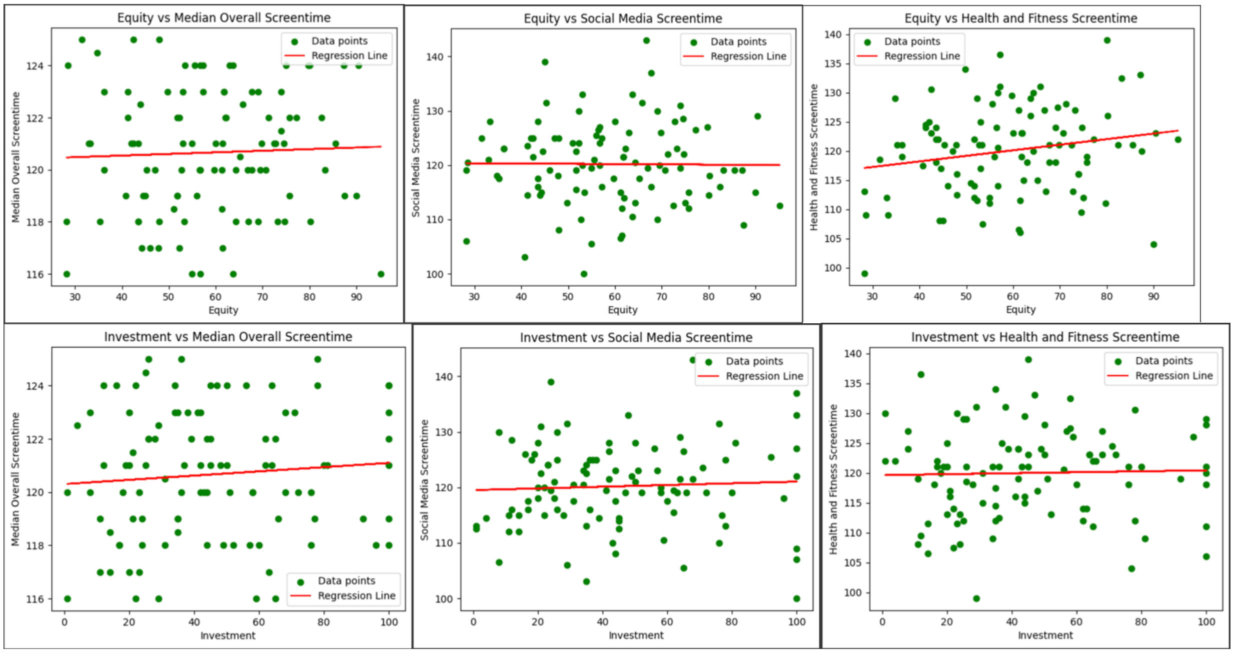
```
           city_name  med_health_and_fitness  med_social_media  median_overall
     Albuquerque, NM                   126.0             110.5           116.0
         Anaheim, CA                    99.0             106.0           116.0
       Anchorage, AK                   125.0             120.0           123.0
       Arlington, TX                   112.5             125.0           125.0
       Arlington, VA                   128.0             122.0           122.0
                 ...                     ...               ...             ...
           Tulsa, OK                   121.0             126.0           121.0
  Virginia Beach, VA                   116.0             108.0           118.0
      Washington, DC                   120.0             109.0           119.0
         Wichita, KS                   113.0             128.0           121.0
   Winston-Salem, NC                   111.5             112.0           117.0
```

Finally, we merged both datasets for our data analysis using the “merge” operation on city name: merged_df = pd.merge(df_parkscores, df_screentime, on='city_name', how='inner')

2. Exploratory Data Analyses

We generated 18 scatterplots using matplotlib.pyplot with Pearson’s correlation coefficient to describe each plot, as well as a regression line. Each scatterplot compares one category from the Park Score rankings to one category from the median screen times.

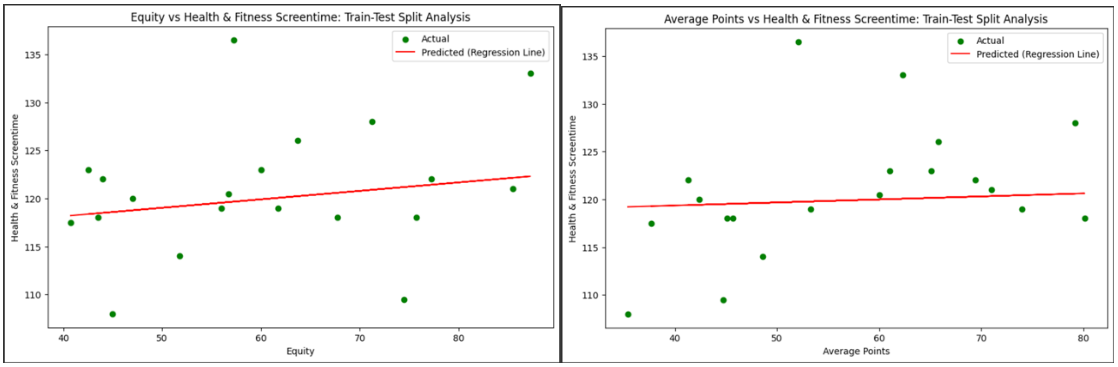




The scatterplots demonstrate weak correlations with an absolute maximum r-score of 0.20. The most promising correlations are Avg_Points vs Health & Fitness (r = 0.13) and Equity vs Health & Fitness (r = 0.20).

3. Linear Regression Model

We implemented Scikit-Learn train-test split on the two most promising correlations for regression analysis. For Equity vs Health & Fitness, we observed a MSE of 41.19 with $R^2 = 0.05$, and a regression slope of 0.09. For Avg Points vs Health & Fitness, we observed a MSE of 41.98 with $R^2 = 0.03$, and a regression slope of 0.03. We provided visualizations of each regression analysis:



Results & Discussion

From our research, we observe little to no correlation between park score and screen time. This outcome suggests that our initial hypothesis—greater accessibility to parks correlates with decreased screen time—may not hold, at least within the scope of this dataset and methodology.

However, although the r-score of 0.20 for Equity vs Health App Usage is minimal, it leads to some interesting implications. According to the Park Score website, they define equity based on 4 equally weighted measures:

1. On a per person basis, ratio of nearby public park space between neighborhoods of color and white neighborhoods;
2. On a per person basis, ratio of nearby public park space between low-income neighborhoods and high-income neighborhoods;
3. Percentage of people of color living within a 10-minute (half-mile) walk of a public park;
4. Percentage of low-income households living within a 10-minute (half-mile) walk of a public park.

Based on this definition, if equity is indeed positively correlated with health app usage, and assuming increased health app usage is indicative of increased physical activity, then people generally tend to be more physically active if parks are more accessible to low-income neighborhoods. Future research should examine this relationship more thoroughly using more sophisticated data collection methodology. Because of our inconclusive r-score, these speculations are also inconclusive. Moreover, with an r-score that low, as well as our randomly generated data, it's possible that more runs of the same program will generate completely different results.

Given what data were accessible to us, the glaring limitation to this study is that the health app usage data were likely inaccurate to reality since they were randomly generated. Despite this, our study lays a foundation for future research. Should a relationship between park accessibility and physical activity be identified with more comprehensive data, the implications could be significant. Cities with underfunded park systems could benefit from targeted investments, which may in turn improve public health by promoting physical activity. For example, increased funding for parks could reduce sedentary behavior, enhance community well-being, and potentially mitigate health issues such as obesity and depression.

For further research, it would make sense to perform a survey on either physical activity per US city or screen time data per US city. While screen time is worth considering, mobile usage has become somewhat intertwined with physical activity with the addition of health apps, so measuring physical activity directly can offer a more representative reading of the health of a city populace. Researchers can also consider obesity rates for each US city as a substitute for physical activity, which is arguably even more accurate, since people may misreport their physical activity levels. Based on the results of those observations, the same analysis can be performed as was done in this study, and more actionable insights can be produced.

Bibliography

Carlson, S.A., Fulton, J.E., Lee, S.M., Foley, J.T., Heitzler, C., Huhman, M. (2010). Influence of Limit-Setting and Participation in Physical Activity on Youth Screen Time. *Pediatrics*, 126 (1).

Richter, F. (2023). Digital Detox? How Americans Try to Limit Their Screen Time [Digital image]. Retrieved December 04, 2024, from https://www.statista.com/chart/30968/measures-taken-to-manage-screen-time/

Silverman, J., Srna, S., Etkin, J., (2024). Can Time Limits Increase Time Spent?.

Statista Survey. (2017). Percentage of U.S. adults who would be willing to use an app to measure health metrics as of 2017, by community [Graph]. In *Statista*.