Cellphone App Usage Analysis

James Lutes

DSC 140S

12/01/24

In the past 30 years cellphones have grown rapidly in popularity and usability, from simple flip phones to now mobile computers in the palm of your hand. I set out to learn more on how phones are used by different people. The data set I used was “mobile_usage_behavioral_analysis” which is a Comma-Separated Values (CSV) file that contains user cellphone usage behaviors, such as the amount of time spent on various types of apps such as social media, games, productivity apps and the amount of screen time that is spent on these apps. The data frame contains ten columns of data with the various types of app, the location of the phone user, their age, and gender. The data was analyzed to find interesting details between the different data points, such as the difference between gaming app usage and productivity app usage between males and females, and the difference in screen time for different ages.

To analyze the data, Microsoft Excel and Python were used to process the CSV (Comma-Separated Values) to find relationship in the data. The first evaluation made with the data was finding out the statistics data of the average, max, min, range and standard deviation of the numerical datasets. The statistical data is recorded in Table 1. This information is used to compare later discoveries.

Table 1 Statistical Analysis

	Age	Total App Usage Hours	Daily Screen Time Hours	Number of Apps Used	Social Media Usage Hours	Productivity App Usage Hours	Gaming App Usage Hours
Average	38.745	6.41	7.70	16.65	2.46	2.50	2.48
Max	59	11.97	14.00	29.00	4.99	5.00	5.00
Min	18	1.00	1.01	3.00	0.00	0.00	0.01
Range	41	10.97	12.99	26.00	4.99	5.00	4.99
Standard Deviation	12.18	3.13	3.71	7.62	1.44	1.44	1.45

Following the statistical analysis in excel, the CSV was imported into Python to analyze and build various graphics to give a better idea how the data is related. To analyze data in python a good practice is to first start with preprocessing of data, such as determining the types of data contained in a data frame. The line of code "phone_data.info()" was ran in python to find the column names, the amount of each type of data and the overall size of the data frame as displayed in Figure 1.

Figure 1 Data Frame Information

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 10 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   User_ID                   1000 non-null   int64
 1   Age                       1000 non-null   int64
 2   Gender                    1000 non-null   object
 3   Total_App_Usage_Hours     1000 non-null   float64
 4   Daily_Screen_Time_Hours   1000 non-null   float64
 5   Number_of_Apps_Used       1000 non-null   int64
 6   Social_Media_Usage_Hours  1000 non-null   float64
 7   Productivity_App_Usage_Hours  1000 non-null   float64
 8   Gaming_App_Usage_Hours    1000 non-null   float64
 9   Location                  1000 non-null   object
dtypes: float64(5), int64(3), object(2)
memory usage: 78.3+ KB
```

In this data frame as shown in Figure 2 there are two columns with categorical data represented by “object”, and there are eight columns containing numerical data which is represented by “int64” and “float64”, which means its either an integer or a string of numbers respectively. The range index is 1,000 indicating that there are 1,000 rows of data to be analyzed. Every column contains 1,000 non-null counts meaning every row of data has some sort of usable data point contained in it. Therefore, there is not much preprocessing to do in order to perform the analytical tests to answer my questions.

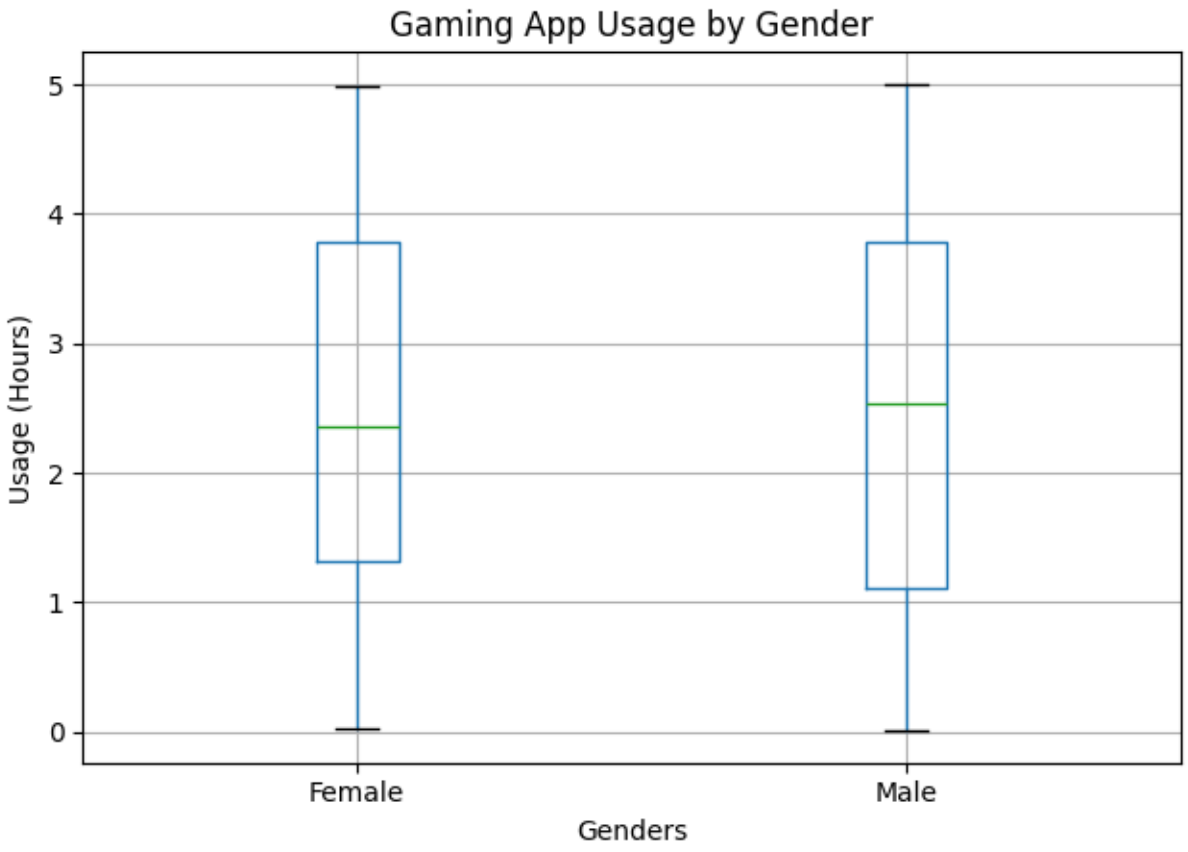
The first question that was analyzed was the difference in hours of game app usage between male and female users. This was done by filtering the data by male and female users. Then using the function “.describe” on each of the filtered data sets the statistical values were calculated and recorded in Table 2. Along with the normal statistical data the number of counts is recorded. Which is the amount of data points per gender that were used in the analysis. This shows that the data frame being analyzed is split roughly 52/48 between males and female cellphone users.

Using Python and its useful ability of creating graphics, a boxplot was made to help visualize this information as seen in Figure 2 “Gaming App Usage by Gender”. As seen in both Table 2 and Figure 2 the hourly cellphone usage per day to play games is roughly the same. The difference in the mean value was just 0.03 hours or 1.2 minutes, with females spending a small amount of time more than males playing games.

Table 2 Gaming Statistics

	Male	Female
Count	517.00	483.00
Mean	2.46	2.49
STD	1.47	1.43
Min	0.01	0.02
25%	1.11	1.32
50%	2.54	2.36
75%	3.79	3.78
Max	5.00	4.99

Figure 2

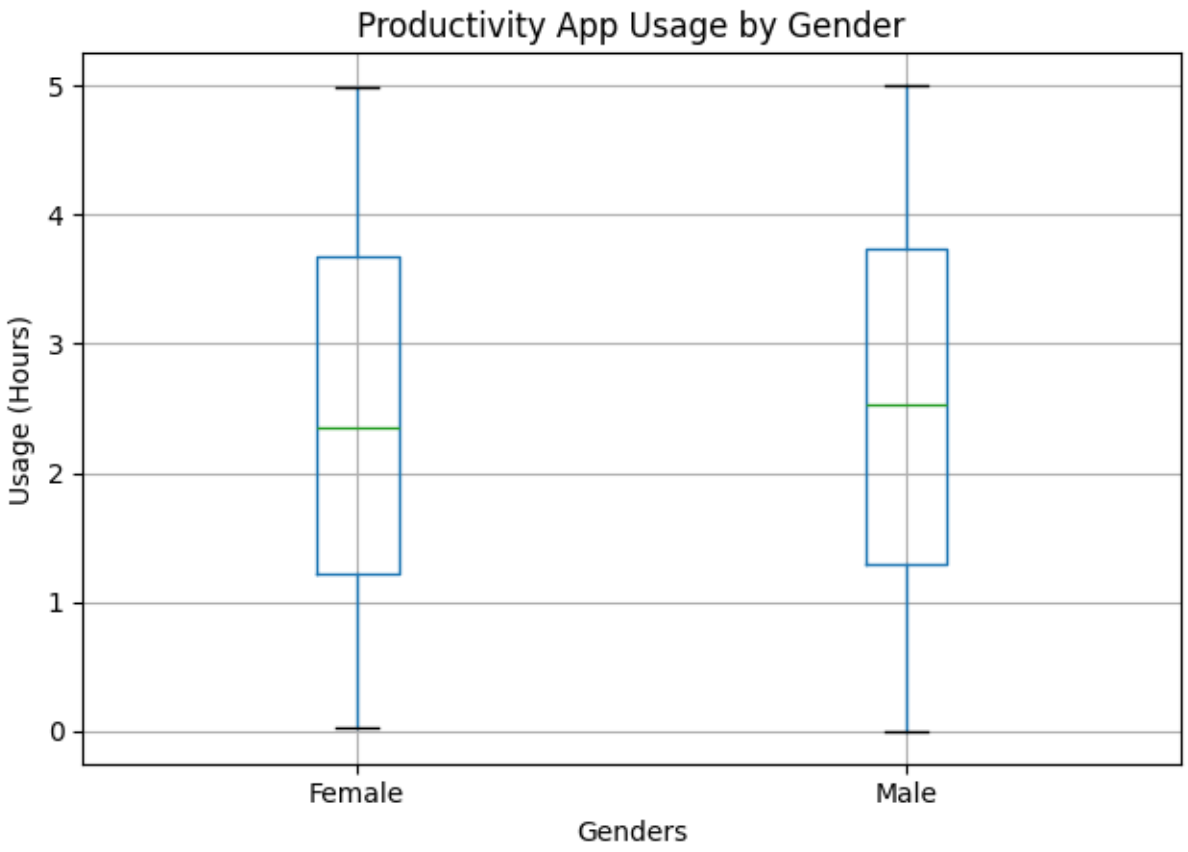


Using the same process of analysis, the difference in use of productivity apps was then analyzed per gender.

Table 3 Productivity Statistics

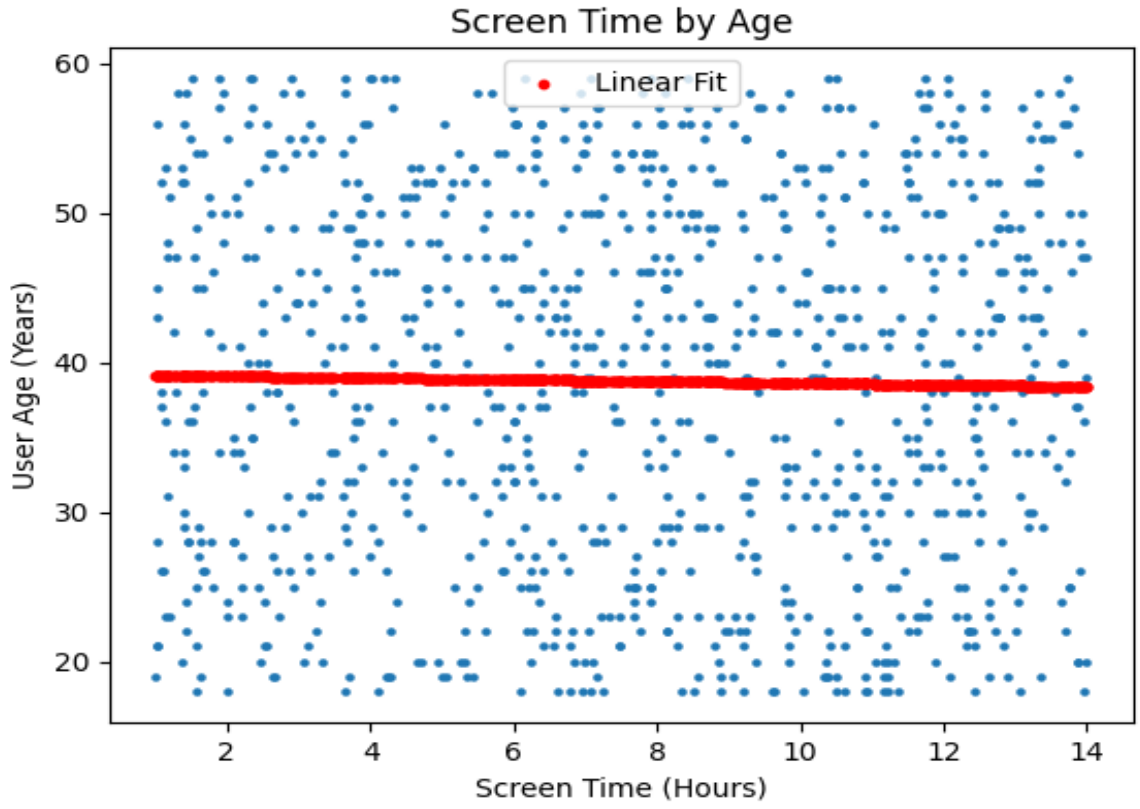
	Male	Female
Count	517.00	483.00
Mean	2.52	2.47
STD	1.45	1.44
Min	0.00	0.03
25%	1.29	1.22
50%	2.53	2.35
75%	3.74	3.68
Max	5.00	4.99

Figure 3



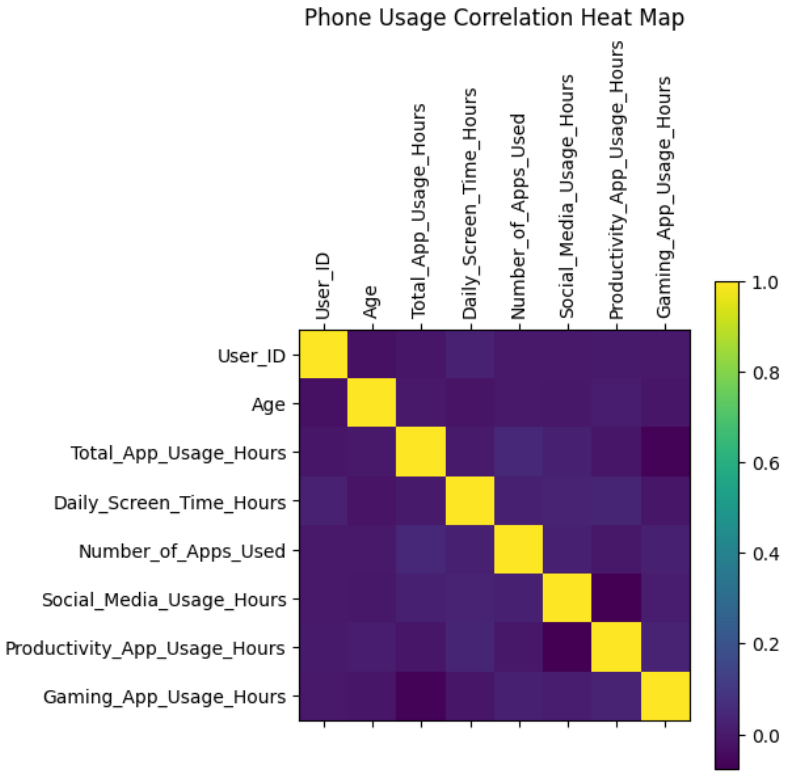
Again, the analysis revealed that the two genders had a relatively close spread of data when determining which one used productivity apps the most. Males tended to use productivity slightly higher with a mean usage of 2.52 hours, which is only 0.05 hours or 3 minutes higher than female usage. Indicating that gender doesn't really play much of a roll when it comes to different app usage.

Since the gender of the phone user showed that there wasn't much of a change I decided to analyze if a user's age would determine the amount of screen time that was spent each day. From Table one the range of users is 18-59 years old. To analyze this relationship a linear regression model was created and graphed on a scatter plot of user age vs total screen time. Linear regression is a statistical machine learning tool to model the relationship between a dependent and independent variables.[1]



The correlation between the user's age and the total amount of screen time ended up being -0.02. A correlation score ranges between 1 and -1. The closer the correlation score is to positive one indicates a strong direct relationship with the data, where if one data point were to increase so would the other. As the correlation score goes towards negative one it indicates a strong inverse relationship in the data. A correlation scores close to zero results in data that does not have any real relationship between them. Therefore, the correlation score of -0.02 shows that there is no real relationship between the age of the phone user and the total amount of screen time that is spent by them.

Following this revelation, I wanted to see if there were any numerical columns that shared a large correlation between them. In order to do this a heat map of correlation scores was made in Python to quickly visualize it.

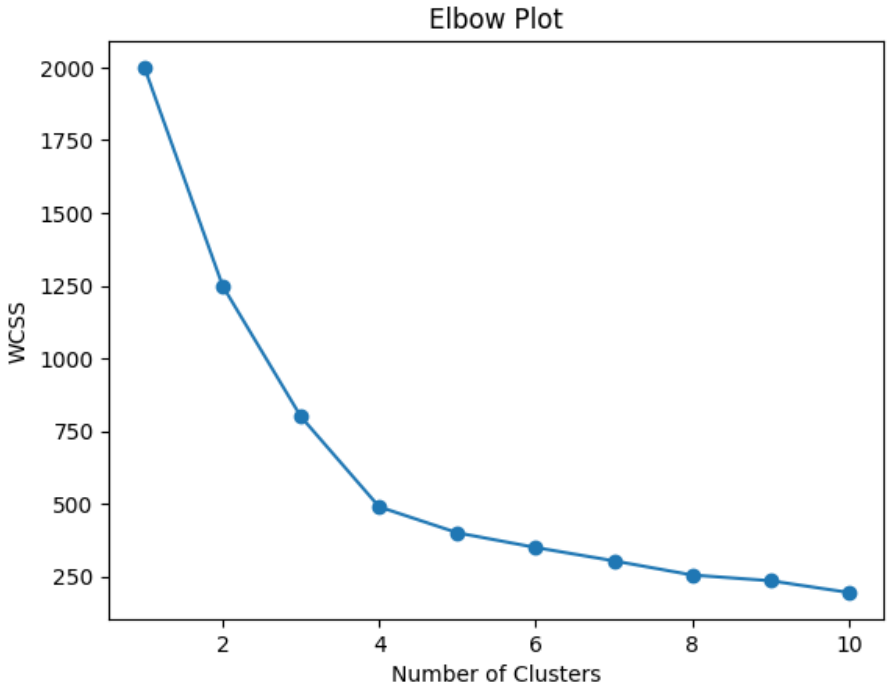


The heat map shows that there isn't much of a correlation in any of the numerical data. The heat bar on the right side of the heat map shows the color coordination with the numerical value. Most of the correlations are close to zero, and the only values that are a value of one is when a data column corresponds with itself.

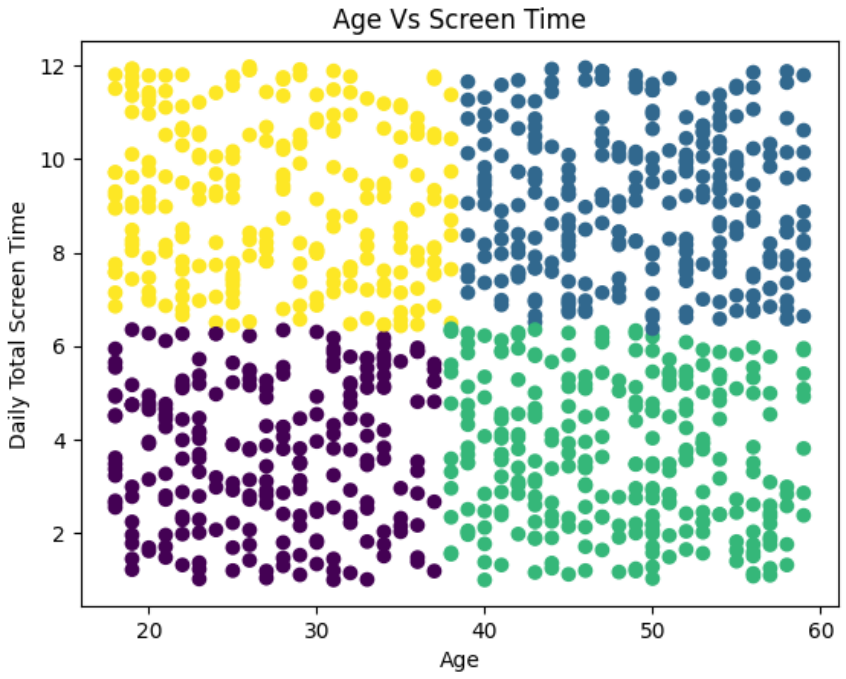
Now that it was determined that there is no correlation between the numerical data, the categorical data needed to be analyzed to determine if it had a correlation. With categorical data, instead of using a Pearson correlation score a chi squared test is performed to determine the p-value for the relationship. By creating a cross tab of the two columns and running a chi squared test the p-value resulted in 0.55, Which is considerably high for a p-value. Indicating that the categorical also does not show any real correlation between them.

Even though the numerical data and categorical data appear to not have any correlation between them further data analysis could reveal some type of pattern. Through the use of Kmeans the data will be further analyzed. Kmeans is an unsupervised machine learning algorithm that groups data that has similar data point into clusters to find relationships.

While performing a Kmeans analysis the optimal number of clusters is calculated and can be graphically presented with the use of an elbow plot, when the line transitions from a large slope to a shallower slope indicates where the clusters are most compact.



The sharp change in slope that occurs at 4 indicates that 4 is the optimal number of clusters to analyze this data. Therefore, a kmeans scatter plot was made to group this data into the four groups.



The Kmeans plot shows the four separate groups but there is no discernable information that can be derived from this scatter plot to answer the question if a change in age results in a change in the total screen time usage per day.

Ultimately this data set does not have sufficient data to derive any significant revelations. Maybe if more data was collected with a broader age range and throughout the entirety of the United States there would be enough data to answer the questions I was searching for.

Citation

1. https://www.geeksforgeeks.org/ml-linear-regression/#what-is-linear-regression
2. https://www.kaggle.com/code/yts2024/mobile-usage-behavior