Chelsea Evanich

Dr. Butler

DSC-140

06 December 2024

Breakdown of Cell Phone Use in Today’s Society

One raging topic in society today is the use of cell phones. Cell phones have completely changed the way we live, by affecting communication, games played, education, finding answers, etc. Further, the factor that social media plays in everyday life is also huge. To learn more about cell phones and the way that they are used, an analysis of the dataset created by Bhadra Mohit which was posted on *Kaggle* was constricted. This dataset with ten columns and 1000 rows allows us to look for connections in the columns, age, gender, total app usage hours, daily screen time hours, number of apps used, social media usage hours, productivity app usage hours, gaming app usage hours, and location (Mohit). Gender and Location are classified as categorical data which means they are words, while the rest of the columns are considered continuous data. This data consists of either integers (numbers with no decimals) which are the age and number of apps used columns, and the rest of the columns are floats (numbers with decimals). This dataset has been used to find the answers to the following questions. First, is gender compared to different columns of smartphone usage statistically significant for further testing? Second, does the younger generation use their phones significantly more than older users? Last, how accurately can an age group be predicted through the use of machine learning?

A screenshot of a calculator

Description automatically generatedA screenshot of a computer

Description automatically generatedTo begin it is important to gain an understanding of the data that is being worked with. To do this, a chart was created to easily view the count, mean or average, standard deviation (std), minimum (min), first quartile (25%), second quartile (median), third quartile (75%), and maximum (max) of all the numeric columns. Important things to gather from this chart (Figure 1) include the count of data for all columns are 1000, which means every column in every row of the data is filled. In addition, the mean and 50% (median) of every column are very close to each other, therefore there are not any outliers that would be skewing the data. In addition, there are no negative numbers for any of the columns, which is good as person could not be negative three years old or spend negative four hours on the phone. In other words, the data makes sense for the represented column. Lastly, standard deviation is an important value to look at. The standard deviation is a measure that shows how spread out the data is compared to the mean. Ignoring the User ID column because that column does not hold value to this dataset, the standard deviation is all relatively low meaning the data spread is all very close to the mean value.

Figure

A screenshot of a phone

Description automatically generatedA screenshot of a phone

Description automatically generatedAfter analyzing the numeric data, the next step is to see if the distribution of categorical data was even. This means, ensure there is fair representation of both location and gender. To easily accomplish this, the website *GeeksforGeeks* was used as a reference to make a frequency chart. A frequency chart shows you how often the category shows up in the data (GeeksforGeeks). The location frequency table (Figure 2) shows that Chicago users made up 192 rows of the data, Huston made up 181, Los Angeles 185, etc. Although these frequencies are not exactly equal, they still are fair representation for all locations. Same goes for the gender frequency table (Figure 3). Although males have slightly more users than females, overall, the representation in the data is good. However, good representation is never a guarantee. For example, if the table showed male had 900 users and females had 100 users, the females may not be accurately represented because the sample size is so small. It is important to see the representation in the data you are working with.

Figure 2

Figure 3

Next, is to look for connections between the numerical columns in the dataset. To be able to search for connections easily, use a correlation heatmap. To do this, a new library called Seaborn was imported to simplify the steps of the heatmap (GeeksforGeeks). A correlation heatmap does Pearson R evaluations on all the numeric columns in a data. To find strong correlation, look for scores close to 1 or -1. In other words, if the value was 0.98, then as one column increased the other would also increase, as 0.98 is a very strong positive correlation score. To read the heatmap, the legend on the left shows that boxes that are red have the highest A screenshot of a computer screen

Description automatically generatedcorrelation, and dark blue boxes have the lowest correlation. The heatmap (Figure 4) is entirely blue, besides the red diagonal, which is comparing the column to the same column, which means this data is not very correlated. When first reading through the column names, it seems there was a good chance for some of them to have high correlation. Especially columns like Daily Screen Time Hours and Total App Usage Hours or Total App Usage Hours compared to the category hours of Social Media, Productivity, and Gaming. Something surprising about this data set A diagram of a graph

Description automatically generated with medium confidenceis that the closest a score gets to 1 or -1 is -0.08 with Social Media Usage Hours and Productivity App Usage Hours. Although this is the highest correlation score in the data set, it is still an extremely weak correlation. To see just how weak this correlation is as the most two correlated columns in the data, a scatter plot makes a great visual. As displayed in the scatter plot (Figure 5) there is no real correlation between the hours spent on productivity apps and social media. If they were correlated the blue datapoints would be close to the orange line of regression. The line of regression is a prediction of where the points would fall based on each other. The more correlated the data is, the better the line of prediction will be. This dataset is a great example of how a lot of real data is actually not very directly correlated. This dataset with many very similar and comparable columns has no real correlation. As the graph shows the two most correlated columns in the dataset and those really have no true correlation.

Figure 4

Figure 5

A graph of blue squares

Description automatically generatedNext before looking deeper into connections, it is important to look at the statistically significance between gender and the other columns. To do this several Chi-Squared tests to search for a p-value that was less than 0.05 were performed. A p-value that is less than 0.05 means that that the data comparison is statistically significant. However, after coming gender to every single other column the lowest p-value found was 0.11 between gender and gaming. To look deeper into this connection, the mean values of gaming hours of males and females were found. Females have a slightly higher mean of 2.49 with males almost the same at 2.5. Then I looked at the median values of each and males had a higher median of 2.54 with females at 2.36. The next step is to make sure there are no outliers in the data affected the minimum and maximum values, however both the minimum and maximum values of gaming hours by males and females were within 0.01 of the opposed gender. The bar chart (Figure 6) makes it very easy to see that there is no difference in the time spent on gaming apps by males and females. Although some people may speculate males spend more time on games, on average males and females spend the same amount of time on gaming apps.

Figure 6

A graph of a comparison of screen time

Description automatically generatedNext, a deeper look into the concept that the younger generation spends more time on their phones tan older people was taken. Often it is heard that the older generation complains about younger people being “addicted” to their phones. An analysis was done to see if the dataset supports or rejects this claim. To do this, the dataset was split into three different sub datasets. The original dataset had a range of 18-59 years old. The young dataset considers people 25 and under, the middle age group ranges from 26 to 40, and the old 41 years and above. Then charts of boxplots were used to see how each age group compared to the other. The first comparison was just overall screen time, seen in Figure 7. While these boxplots to differ, they have many similarities. All three have a range from 1 to 14. The surprise to me in the data was the oldest group have higher median than the middle age group. One explanation for this could potentially be because some of the oldest group may be retired and therefore have more time to spend on their phones. The median of the oldest group is 7.9 while the median of the young group is 8.4. Which is equal to approximately a half hour difference in the screen time used. Although, this data technically does support the suspicion of the younger group using their phone more, it is not by as drastic of a margin as often made out to be. Since each generation uses about the same amount of screen time, more charts were made to see where each group spends the screen time. Simply, create the same chart as before but comparing the hours spent on social media by each group, this is Figure 8. It would be expected that the youngest generation would have the highest use of social media A graph of a number of people with numbers

Description automatically generated with medium confidenceas it is a newer advancement, and this chart supports that expectation. The age group of 18-25 has the highest use on social media at a median of 2.66 hours. To me, the interesting part of the graph comes from the boxplot of the oldest generation. Although by looking at the graph it appears the middle and oldest groups have the same median, the oldest group is actually slightly higher. Further the median for the oldest generation is surprisingly close to the young group, at a median of 2.46 compared to the 2.66 median of the young group. Next, is the hours spent on productivity apps. These are apps whose purpose is to improve their user’s A graph of a row of rectangular objects

Description automatically generated with medium confidenceefficiency. Some examples of these apps are, Todoist, Google Calendar, Otter.ai, and Slack. In this category, the youngest group of people uses these apps the least as seen in Figure 9. With the middle and oldest groups being almost the same. One thing to note with the bar chart of the oldest range is the size of the box. The box of the oldest group is larger than the other two. This is because the standard deviation of productivity apps used by the oldest group is greater than the young and middle age groups. Having a larger standard deviation means that the data is more spread out which is why the box for the oldest group covers a greater range. The last category to look at is the time spent on gaming appA graph of a graph showing a number of different sizes of objects

Description automatically generated with medium confidences. In fact, each group spends almost the same amount of time on gaming apps, seen in Figure 10. Surprisingly, the group that spends the least amount of time on gaming apps is the youngest group. This could be reasoned by time spent on video games on a console is not being considered, which is very popular thing for high school and college students especially. After comparing the age groups and where they spend their time on cell phones, it is all very close. The younger generation spends on average approximately a half hour more time on their phones than the oldest generation does. By looking at where the time is spent by age, it was found that the oldest and middle age groups have about the same time distribution at about 2.5 hours spent in each of the categories, social media, productivity, and gaming apps. The younger group is last in both times spent on gaming apps and productivity but does spend the most time on social media.

Figure 7

Figure 8

Figure 9

Figure 10

The last part of the analysis of the data is to create a machine learning algorithm that can predict age range that a person is in based on the data. Machine learning is a branch in data science that takes data and creates an algorithm that can predict something else about the data. In this example, the columns used are Social Media Usage Hours, Gaming App Usage Hours, and Daily Screen Time Hours to try to predict the age group of the user. However, different than before, only two age groups are used instead of three. The groups will be split at, if the user is 29 years or younger they will be assigned to group one, or if they are 30 years or older they are in group two. The algorithm took the inputs of the columns listed above and predicted if the user was a member of group one or two. To do this use K-Nearest Neighbors to make predictions. A test size of 20% and ten neighbors. This means that 20% of the data will be used to for testing, and 80% will be used for training the algorithm. Setting the neighbors equal to ten, tells the algorithm to base the A chart of a color chart

Description automatically generated with medium confidenceprediction off of the ten closest data points to the point it is predicting. After the algorithm was trained, it was then tested for its accuracy and it predicted with 67.5% accuracy. To see where the algorithm was messing up, a confusion matrix is very helpful. This matrix (Figure 11) shows what the algorithm predicted correctly and incorrectly. From the matrix it can be read that fourteen users in group two were predicted as group one and 51 users in group one were predicted as group two. This matrix makes it very easy to see that majority of the incorrect predictions come from predicting true group one users as group two. When looking back at the boxplot charts above, (Figures 7, 8, 10) age groups two and three are very close to one another. In that case, age group two was people started at 26. However, these people 26-29 that were in group two above were very close in comparison to the oldest group, but they are actually placed in group one for the algorithm. This age group is probably majority of the people who are true group one but are being predicted as group two.

Figure 11

In conclusion, data is not as connected as one might assume it may be, and common generalizations are not always the truth. First, some people may speculate that males and females use different categories of apps more than the other gender, however there is no statistical significance supporting this claim. After using contingency tables and the Chi-Squared test, there was not a single column that was statistically significant with the gender of the user. Which means that there is no true difference in the dataset on how males and females use their phones. Next, this data both supports and rejects the claim that younger people are more addicted to their phones than older people. In some ways, the data does support this as after splitting the data, the youngest group averages about thirty more minutes of screen time compared to the second which was the oldest generation. However, the results were not nearly as drastic as often made out to be. Lastly, by using machine learning to attempt to predict the age range a user fall in based on their screen time, social media usage, and productivity app usage an algorithm predicted an age group with almost 70% of accuracy. With most of the error coming from predicting group one users as group two. Again, this dataset is a great example of how real data is not as correlation as it may be originally thought, and data may not support the generalizations that are heard over and over.

Works Cited

GeeksforGeeks. (2024, September 5). *How to create a Seaborn Correlation heatmap in python?*. GeeksforGeeks. https://www.geeksforgeeks.org/how-to-create-a-seaborn-correlation-heatmap-in-python/

GeeksforGeeks. (2022, February 20). *How to create frequency tables in python?*. GeeksforGeeks. https://www.geeksforgeeks.org/how-to-create-frequency-tables-in-python/

Mohit, B. (2024, October 23). *Smartphone usage and behavioral dataset*. Kaggle. https://www.kaggle.com/datasets/bhadramohit/smartphone-usage-and-behavioral-dataset