**Data-science Project Documentation**

When working with huge real world data sets, we can find them messy and that’s why we first go through a process called Data wrangling, to organize and clean up the data and make it more suitable for creating regression models.

**Data Wrangling steps:**

1. We examine our datasets and profile it
2. Assess the Datasets quality
3. Clean the data(Handle missing values, convert data types, remove duplicates, … etc)

In this project we will be using 3 datasets.

The first dataset we are using is “mobile\_usage\_behavioral\_analysis” from Kaggle. This dataset provides insight on the usage habits of 1000 users in the United States.

After examining the data set here are some initial notes:

* The total\_app\_usage\_hours are off (They should be equal to: productivity\_app\_usage\_ hours + gaming\_app\_usage\_hours)
* The total\_app\_usage\_hours is sometimes greater than the daily\_screen\_time\_hours. It should be less than daily\_screen\_time\_hours.
* These issues could be due to the data being manually or synthetically made up, or it could also be because of logging errors

Next Using the Pandas library to analyze the data these are more notes:

* Checked the data type for each column
* Found the mean, minimum, maximum, standard deviation, and the quartiles
* Checked for Null values in the dataset (there were none)
* Checked for duplicates (there were no duplicate rows)
* I ran a check for outliers, and surprisingly found none.
* Created a correlation matrix to see how strong the relationships between each of the variables are.

So far it seems that most of the people in the sample are around 38-39 years old, with a standard deviation of 12.186. It also seems that most people have a daily screen time of 7 – 8 hours.

The second Dataset we are using ‘user\_behavior\_dataset’ from Kaggle as well. Explores the user behavior for 700 participants and explores variables such as app usage time, screen time, battery drainage, and data usage.

After examining the dataset here are some initial notes:

* We find this dataset unlike the previous seems to be mostly consistent with its values
* The app\_usage\_time column is in minutes, while the screen\_time is in hours.
* Screen\_On\_time is > app\_usage\_time as it should be
* Batter\_drainage and data\_usage seem to be both correlated to app\_usage\_time and screen\_time.
* Most of this dataset are android users

After using the pandas library built in functions to further analyze this dataset here are some more notes:

* App\_usage\_time has a mean of 271.12, and a standard deviation of 177.19
* Screen\_on\_time has a mean of 5.27, and a standard deviation of 3.06
* The mean, Std, min, max, and the quartiles for the remaining variables have also been found.
* Checked for null values (none were found)
* Checked for duplicates (none were found)
* After creating histograms with 100 bins for various variables, it seems that we have diverse range of ages, from 18 to 59.
* From the histogram App\_usage\_time seems to be mostly skewed to the left most users are moderate users. Also applies to screen\_on\_time.
* Looking at the correlation matrix and the heat map, most variables except for the user\_ID, and age seem to be strongly correlated to each other.
* The IQR method was used to find and flag outliers (none were found).

The third dataset we are using “AffectOfMobilePhonesOnStudents” from github , investigates the Affect of mobile phones on students. With a sample size of 99 students. It takes into consideration variables like:

* Age
* Gender
* Mobile phone use for education
* Mobile phone activities
* Educational apps
* Daily usage
* Attention span
* Performance impact
* Health risks
* Usage symptoms
* Symptom frequency
* Health rating …..

After examining the Dataset here are some initial notes:

* Most students are android users, have 4-6 hours of screen time, and sometimes use the phone for educational purposes.
* Most students find it useful for studying, using educational videos.
* Most students agree it impacts their performance, and distracts them while studying.
* Most students get usage symptoms like headcahes, anxiety and stress.
* Most students have a good health rating

After using the pandas libarary to further analyze te dataset here are some findings:

* Obtained the datatype for each variable all of them are objects
* There are six null values found in these rows:
  + Helpful for studying
  + Educational Apps
  + Usage distraction
  + Useful features
  + Health risks
  + Usage symptoms
* No duplicate values found
* The Dataset seems to be consistent

Then we impute the missing six values, using the most frequent values for each column.

The Fourth dataset ‘mobile\_addiction.csv’ also from Kaggle. Investigates the addiction level of participants with a sample size of 13,600. With variables like daily\_screen\_time, app\_sessions, social\_media\_usage, gaming\_time, notifications, night\_usage, age, work\_study\_hours, stress\_level, apps\_installed, Addicted.

After taking an initial look on the dataset here are some notes:

* Daily\_screen\_time increases as app\_sessions increases.
* Stress\_levels increase as social\_media\_usage increases.
* Subjects who have minimal social\_media\_usage are usually not addicted.
* Subjects who are younger tend to have higher social\_media\_time.

After analyzing the file using pandas library here are some more notes:

* All columns have int data type except for the addicted column it is an object
* There are no null values
* There are no duplicates
* Created graphs to get a better understanding of the dataset
* Created a heatmap and correlation matrix to see the correlations between variables
* Tested for outliers using z-test method (271 outliers were found)
* Removed all outliers (dataset size is now 13,331)