**CSCI.4923 Capstone in Interprofessional Informatics**

**Methodology and Data**

Sachin Karki

Texas Woman’s University

Denton, Texas

Author Note

Sachin Karki, Texas Woman’s University, Mathematics & Computer Science

**Selection and Interpretation of Dataset**

Selecting a dataset was a challenging part of this research. While many resources and academic portals were utilized to search, locate, and retrieve a dataset, it was initially a daunting task since the right and precise dataset took a lot of searching. Coming across a research paper about the cross-sectional study between smartphone usage and interpersonal behaviors was relieving as it had the right dataset that was appropriate for our study. The raw dataset used for the ***Depression, anxiety, and smartphone addiction in university students - A cross-sectional study*** by *Jocelyne Matar Boumoslehm and Doris Jaalouk* would be used moving forward to test the hypothesis and answer our research question with necessary analysis.

To correctly address our question from the retrieved dataset only a relevant and streamlined set of variables are needed which is where the filtering and sorting out of the variables took place. As the raw data had 26 columns of variables, most of those variables were meant for a higher level of analytical overview on smartphone usages. Only the variables that were previously used by the authors for their study to compute the Smartphone Addiction Index (SPAI) would be used in our study as well. In addition to this, our research will effectively make use of variables that were less utilized or had minimal contribution in finding the results from the prior study. Those less utilized variables do have some potential to generate more insights that have not been explored or mentioned in the prior study. This research will draw out new conclusions and insights from the dataset. Our research will primarily focus on exploration and analysis of these variables to answer the questions which we will discuss and dissect below.

**Preparation of Dataset**

First, after retrieving the dataset it was obvious that the dataset was in a raw state. The number of observations was large (n=688), each observation had 26 relating variables/columns. The dataset needed modification and adjustment. To be clear of what variables/columns are needed to be filtered out we must define questions that we are trying to answer with the dataset. The goal of our research is to answer the following questions statistically and visually in whichever way is more feasible.

1. What interpersonal behaviors candidates practiced that were susceptible to excessive smartphone usage?
2. Is there any relationship between different personality types and excessive smartphone usage?
3. What activities of smartphones contributed to excessive usage in different personality types?
4. Which personality type is more vulnerable to the effects of excessive smartphone usage?

The only way to answer our questions with this relatively big dataset was to trim it down and streamline it in a way that looks easy as well as methodical for our analysis. First, splitting the dataset was done. The main dataset was split into three other datasets where each dataset had variables relating to the activity and interpersonal habits of candidates (Activity.xlsx), effects that were induced after excessive smartphone usage on both personality types (Effects.xlsx) and the scale of smartphone addiction respectively (Scores.xlsx). Each dataset was created by trimming different variables from the main dataset. For an instance, for the *Activity.xlsx* dataset variables like “reading news, calling friends, entertainment, studying, etc.” were used. Similarly, variables like “decreased sleep time, not being able to stop worry” for *Effects.xlsx*. And lastly variables like “anxiety and depression scores” for the *Scores.xlsx*.

**Description of Method**

To extract meaningful results out of the newly created three datasets, various methods were utilized. Each dataset had a purpose of its own to answer our research questions, so each dataset was utilized with specific methods that served as a base for the different type of analysis. To keep it simple and precise, two versions of analysis were done (i.e. statistical and visual) and then each dataset were allocated specifically to use either of these methods to answer our questions.

1. ***Statistical Method (using R in RStudio and RMarkdown)***

For the statistical part, the *activity.xlsx and effects.xlsx* datasets were utilized to draw series of conclusions. Various methods of analysis like Spearmen’s correlation, Backward Akaike Information Criterion (AIC), chi-square, Cohen’s d, linear and multinomial regression were used to see and check the relationship and association between smartphone activity, personality type, and score of effects from smartphone usage. Here’s a breakdown of all the procedures.

1. *Addressing Missing Data:* To begin with, the main dataset had consistent number of missing data from different variables. These missing data remained intact following the creation of our datasets as well. Those missing data are simply a result of no feedback from the participating candidates. In the prior research (from the main dataset), the missing data were treated as outliers to glean statistical results. Although outliers can create a discrepancy in data analysis, a proper way of identifying and treating the missing value can improve the result of the analysis. To identify such missing data, the R library of *ggplot* was utilized. Running the procedure helped in identifying a total of 55 and 47 candidates with missing data from every column of the *activity.xlsx and effects.xlsx* datasets respectively. Now that we identify the total number of missing candidates, we can be certain that they can cause a bias in the result of our analysis. The variables with maximum count of missing data must be excluded from the analysis since they are likely to have less or no linear relationship with any other variables. In addition to this, candidates having missing data (i.e. 55 and 47 from *activity.xlsx* and *effects.xlsx* datasets respectively)would simply be discarded as they entail little to no contribution in ensuring accuracy of the test. In our case variables like “Study\_Purposes and Rlgn\_Prctces” must be taken out of the analysis. This can be done using a *ggplot* package and a procedure built within R.
2. *Data Sorting for Categorical Values*: The dataset is mostly based on categorical values and the number of analyses can be limited since our data model is semantically numeric i.e. 1 or 0 (yes or no). To sort our categorical data, *Cohen’s d* was used to stratify the outcome of different variables based on two personality types on both datasets. The first Cohen’s d was used to measure the mean of every activity on both personality types and the second Cohen’s d was used to measure the mean of effects of smartphone usage among both personality types. As a reminder, the significance level of 0.05 was used as a reference for cross-comparison of the mean from each of the tests from both datasets.
3. *Variable Selection*: After sorting the categorical values, a backward AIC was used on both datasets to reduce the error rate. From the computed results, variables with reduced AIC were selected for the chi-square test. This would define what variables would be suitable for cross-referencing and cross comparing with other independent variables from both activity and effects datasets.
4. *Multinomial and Linear Regression*: After the selection of variables, a multinomial regression on both datasets was done which was followed by linear regression. Using a multinomial regression was necessary to predict the placement of categorical values on dependent variables from multiple independent variables. For an instance, for the first dataset variables like “ExcessveSmrtPhn\_Use” were tested alongside activity variables like “texting, entertainment, call friends”. This multinomial regression was done on four levels with different variables to streamline the series of conclusions. Once this was done a final linear regression was performed to extract more precise results on independent variables. From this analysis one of the interesting insights that came to our conclusion was the variable “smoking” or the candidates who admit smoking were prone to excessive smartphone usage. This also briefly answers one of the questions our research is trying to answer.
5. ***Visualization Method (Using Tableau and R)***

The purpose of visualization was to extract a visual summary for readers who are not familiar with or rather find statistical results a bit intricate. Visualization is also a great way to graphically summarize key details. For the visual part, the dataset relating to a score of smartphone addiction *(Scores.xlsx)* was used in Tableau to make a visual comparison of different effects on both personality types. To do so, the variables were converted to their right state i.e. dimensions or measures. The data type of these variables was also rightfully defined to string and numeric type. Similar processes and approaches were taken for other datasets that would help to visualize the growing number of cellphone users in the world and so on.

The statistical methods mentioned above in R have also generated crucial visuals that can be used to visually summarize the statistical results.

**Illustration**

For all the steps that were involved in preparing and exploring the dataset here are some illustration of selected statistical and visual methods.

1. Bar graph of missing data in Activity.xlsx dataset (From R)

We can see from the graph illustrated above that, the total count of missing candidates from each variable is 55.

Sample Code:

1. Bar graph of missing data in Effects.xlsx dataset (From R)

We can see from the graph illustrated above that, the total count of missing candidates from each variable is 47.

Sample Code:

1. Snapshot of summary on depression score on both personality types.

We can see in the bar graph above that personality type B has the high score on the variable “Depression \_score” from the SPAI scale. This hints that personality type A is less susceptible to effects of excessive smartphone usage.

**Summary**

All the observations seen, and the methods discussed here within this paper are a fraction of the results and outcome from the analysis of our three created datasets. There certainly would be more depth and interpretation to these observations once the test and methods are performed again to accurately assess our analysis. These sample illustrations and procedures are a mere depiction of what could be done with the datasets.

**References**

Boumosleh, J. M., & Jaalouk, D. (2017). Depression, anxiety, and smartphone addiction in university students- A cross sectional study. PLOS ONE 12(8): e0182239. <https://doi.org/10.1371/journal.pone.0182239>

Boumosleh, J. M., & Jaalouk, D. (2017). *Raw data for all participants*. (Version 1). PLOS ONE. <https://doi.org/10.1371/journal.pone.0182239.s001>

Arumugam, N., Selvanayagam, S., & Sathiyasenan, S. T. (2020). The Effects of Smartphone Usage on University Students. *International Journal of Academic Research in Progressive Education and Development*, 9(3), 170–183. : <http://dx.doi.org/10.6007/IJARPED/v9-i3/7960>