Data Cleaning and Documentation

1.1 Data Source - Online News Popularity (Lichman, 2013)

Our dataset summarizes features about articles published by Mashable, which is an online news sharing platform. The dataset has 61 columns and 39,644 rows. We obtained this dataset from the archives section of the Machine Learning Repository of University of California, Irvine.

The data set contains details about articles posted on Mashable over a period of two years. Some of the attributes of the dataset includes things such as Day Published, Number of Shares, Category of the Article (Entertainment, Tech, World etc.) and Number of Words in the Title. Mashable donated this dataset for public use with the goal of identifying characteristics of an article that make it popular. The findings of the analysis of this dataset can then be used to predict the popularity of an article before it is published. This dataset can be used by digital media firms as well as individual bloggers to maximize sharing of their published articles. The dataset is complete in its own and we are not using any other data set along with it.

1.2 Licensing

Usage of this dataset is not restricted. However, to comply with UCI's citation policy, we need to provide the information of any assistance received using this repository if we publish our project online. An APA reference format is suggested as follows:

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Our dataset also requires additional citation as:

K. Fernandes, P. Vinagre and P. Cortez. A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. Proceedings of the 17th EPIA 2015 - Portuguese Conference on Artificial Intelligence, September, Coimbra, Portugal.

1.3 Metadata

Our dataset, the content of which mainly focused on area of business, was donated by Mashable on May 31, 2015. The attribute characteristics are integer and real, and it doesn't contain any missing values.

Associated tasks involved in this dataset include classification and regression. The authors also used assessment methods, specifically Random Forest classifier and rolling windows, to estimate the relative performance value, which is helpful for us understanding the data set as we could study their methods.

The detailed attribute information of our dataset is provided at the following website (Lichman, 2015): https://archive.ics.uci.edu/ml/datasets/Online%20News%20Popularity

1.4 Issues Encountered and Steps Taken to Remediate Data

Fortunately, our data set was relatively clean and structured and we did not have issues like missing values or inconsistent data. However, we did encounter a few issues while going through the data:

1. The metadata provided for the data set was insufficient for some of the columns. We had to look at external sources and existing analyses on this data set to understand the meaning of such columns and what they represent.
   For Example: One column header says *LDA_00* and the column description in the metadata says "Closeness to LDA topic 0". There is no mention of what LDA represents and what is topic 0.
2. The range of values a particular variable can take is not specified. While the range is implied in the case of quite a few variables, having a range of values for some of the variables can be used for comparing actual values with the possible values and determining if our dataset is skewed towards one extreme. We have assumed the possible min and max values of such variables as the minimum and maximum values they take in our data set.
3. Some machine learning and statistical methods or models involved with this data set are helpful for data analysis, however, they are also very new for us to learn and comprehend. For example, the LDA topic mentioned above stands for *Latent Dirichlet Allocation*, a statistical model used to discover relationships between topics. Even though it brings more insights into our dataset, these particular statistical method are higher than our level of understanding at this stage.

1.5 Steps for Data Cleaning

Since there are no missing values, we do not need to delete rows or assume values for any missing data. However, since this dataset is very large we need to subset it according to our research questions.

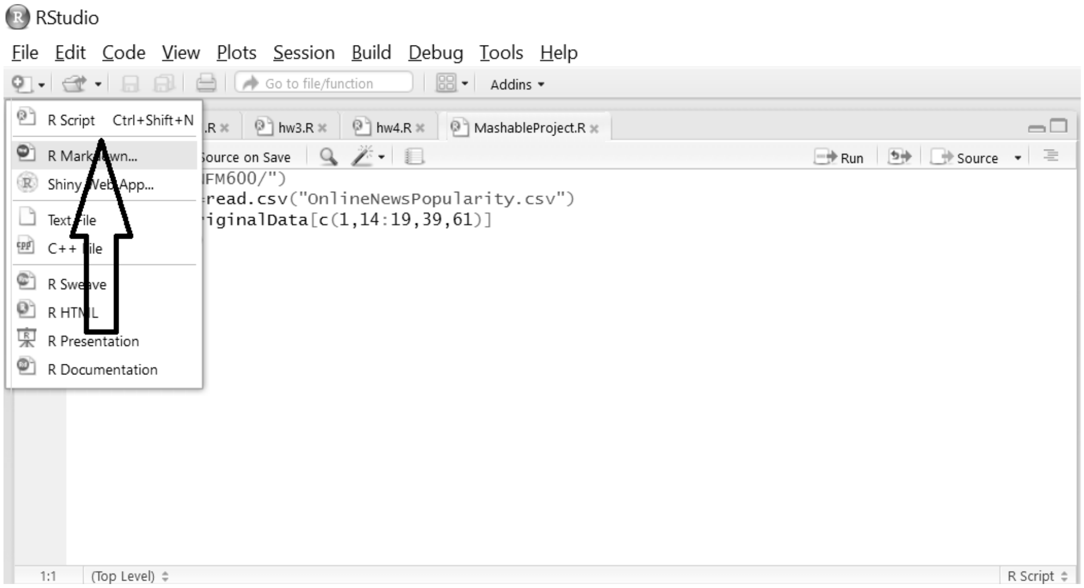
Following are the steps done to clean and format the dataset according to our needs:

1. We create subsets of the dataset, choosing only the attributes required for each research question. We used RStudio to create the subsets:

a) Make sure you have R and RStudio installed on your computer

b) Open RStudio on your system

c) Create a new R script by clicking on the small + icon below the File menu (Figure 1)



```
FM600/")
read.csv("OnlineNewsPopularity.csv")
iginalData[c(1,14:19,39,61)]
```

Figure 1

d) In the new R script file, write the following code:

```
setwd("c:\\INFM600/")
OriginalData=read.csv("OnlineNewsPopularity.csv")
NewData <- OriginalData[c(1,14:19,39,61)]
head(NewData)
```

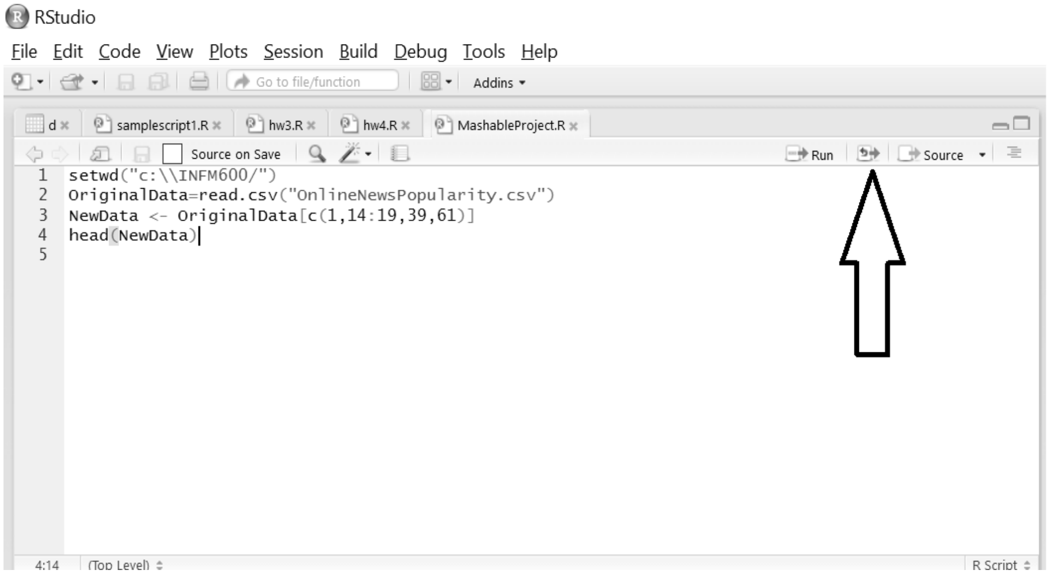
e) The setwd command is used for setting the working directory; enter the path to the directory where your data set is saved.

f) Our data set file is titled “OnlineNewsPopularity”. The read.csv command is used to read the file into R.

g) A subset of the dataset is created by specifying the column numbers needed in the subset in line 3 of the code

h) The head function is used to display the first few rows of data to ensure that the correct columns have been selected

i) You can run all these lines by clicking on the small “repeat and run” icon next to the “Run button”. (Figure 2)



```
setwd("c:\\INFM600/")
OriginalData=read.csv("OnlineNewsPopularity.csv")
NewData <- OriginalData[c(1,14:19,39,61)]
head(NewData)
```

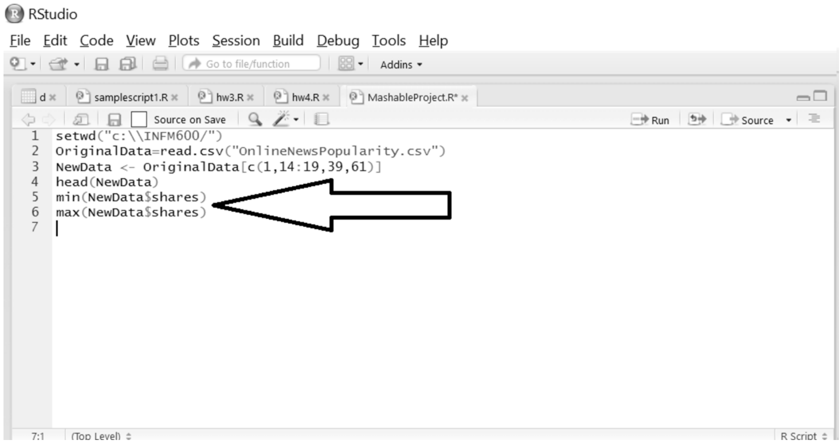
Figure 2

j) All subsets were created in the same way as this one except the column numbers were changed corresponding to each research question.

2. For all numerical variables in each of our subsets, we calculated their minimum and maximum values to get a sense of the range of values these variables can take. The minimum and maximum values can be calculated in RStudio easily as follows:

a) Assuming that the steps mentioned before are followed, a subset named “NewData” should be available for use.

b) min() and max() functions can be used to calculate these values. The argument required for these functions is: *“NewData$VariableName”* (Figure 3)



```r
setwd("c:\\INFM600/")
OriginalData=read.csv("OnlineNewsPopularity.csv")
NewData <- OriginalData[c(1,14:19,39,61)]
head(NewData)
min(NewData$shares)
max(NewData$shares)
```

Figure 3

Word Count- 1023

References

1. Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
2. K. Fernandes, P. Vinagre and P. Cortez. A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. Proceedings of the 17th EPIA 2015 - Portuguese Conference on Artificial Intelligence, September, Coimbra, Portugal
3. Lichman, M. (2015, May 31). UCI machine learning repository: Online news popularity data set. Retrieved November 2, 2016, from https://archive.ics.uci.edu/ml/datasets/Online%20News%20Popularity