Data Cleanup/GitHub Assignment

Cong Wang

May 20, 2017

```
#Load libraries
library(stringr)
library(stringi)
library(zoo)
```

Step 1

- Populate the missing values in the Area variable with an appropriate values (Birmingham, Coventry, Dudley, Sandwell, Solihull, Walsall or Wolverhampton).

```
dirtydata <- read.csv("C:/Users/Emma/Desktop/Analytics 2/Team Assignment/dirty_data.csv", header = T, na
cleandata <- transform(dirtydata, Area = na.locf(Area))
```

Step 2

- Remove special characters, padding (the white space before and after the text) from Street 1 and Street 2 variables.

```
cleandata$Street <- gsub(pattern = "[^[:blank:][:alnum:]+,&/\\-]", replacement = " ", cleandata$Street)
cleandata$Street.2 <- gsub(pattern = "[^[:blank:][:alnum:]+,&/\\-]", replacement = " ", cleandata$Stree

cleandata$Street <- gsub("\\s+", " ", str_trim(cleandata$Street))
cleandata$Street.2 <- gsub("\\s+", " ", str_trim(cleandata$Street.2))
```

Step 3

- Make sure the first letters of street names are capitalized and the street denominations are following the same standard (for example, all streets are indicated as “str.”, avenues as “ave.”, etc.

```
cleandata$Street <- str_to_title(cleandata$Street)
cleandata$Street.2 <- str_to_title(cleandata$Street.2)
# Street to St
cleandata$Street <- str_replace_all(cleandata$Street,"Street", "St")
cleandata$Street.2 <- str_replace_all(cleandata$Street.2,"Street", "St")
# Road/Raod/Roads to Rd
cleandata$Street <-str_replace_all(cleandata$Street,"Road", "Rd")
cleandata$Street.2 <- str_replace_all(cleandata$Street.2,"Road", "Rd")

cleandata$Street <-str_replace_all(cleandata$Street,"Raod", "Rd")
cleandata$Street.2 <- str_replace_all(cleandata$Street.2,"Raod", "Rd")

cleandata$Street <- str_replace_all(cleandata$Street,"Roads", "Rd")
cleandata$Street.2 <- str_replace_all(cleandata$Street.2,"Roads", "Rd")
# Lane to Ln
cleandata$Street <- str_replace_all(cleandata$Street,"Lane", "Ln")
```

```r
cleandata$Street.2 <- str_replace_all(cleandata$Street.2,"Lane", "Ln")
# Avenue to Ave
cleandata$Street <- str_replace_all(cleandata$Street,"Avenue", "Ave")
cleandata$Street.2 <- str_replace_all(cleandata$Street.2,"Avenue", "Ave")

cleandata$Street <- gsub(pattern = " ,", replacement = ",", cleandata$Street)
cleandata$Street.2 <- gsub(pattern = " ,", replacement = ",", cleandata$Street.2)
```

Step 4

- If the value in Street 2 duplicates the value in Street 1, remove the value in Street 2.

```r
cleandata$Street.2 <- ifelse(cleandata$Street.2==cleandata$Street,"",cleandata$Street.2)
```

Step 5

- Remove the "Strange HTML column".

```r
clean_data <- subset(cleandata[,-5])
write.csv(clean_data, file = "clean_data.csv", row.names = FALSE, na = "")
```

Step 6

- Export the data set

```r
write.csv(clean_data, file = "clean_data.csv", row.names = FALSE, na = "")
```