LLCU255 Data and Literary study

Individual Assignment 2

Shenshun Yao 260709204

15/09/2020

Abstract

In this assignment, our goal is to get comfortable working with tables in R based on Novel150 data set.

First, we set working directory, load table and print the name of each column.

```
setwd("~/Desktop/Fall_2020/LLCU255_IntrotoTextMining/Individual Assignment 2")
a<-read.csv("txtlab_Novel150_English.csv")
colnames(a)
```

```
## [1] "filename" "id"       "language" "date"     "author"   "title"    "gender"
## [8] "person"   "length"
```

Problem 1

How many documents are there? How many variables?

Solution:

```
# To get the number of documents, count the number of rows using nrow()
nrow(a)
```

```
## [1] 150
```

There are 150 documents(rows) in total, each document has 9 features(columns), so we have 1350 variables.

Problem 2

How many arthurs are there?

Solution:

```
# unique() returns a vector, data frame or array with duplicate elements/rows remove
length(unique(a$author) )
```

```
## [1] 98
```

There are 98 authors.

Problem 3

Name a novel other than one by Jane Austen that you've read in this data.

Solution:

```
# Create a subset of Novel150 data set without the novels written by Jane Austen
a_1<-subset(a, a$author!="Austen,Jane")
# Select a random element in the column"title"
sample(a_1$title, 1)
```

```
## [1] "Pembroke"
```

A random selected answer was printed.

Problem 4

What is the ratio of first-person to third person novels in our sample?

Solution:

```
# Count the number of first-person and third-person novels respectively.
numberofFirstperson<-length(a$person[a$person == "first"])
numberofThirdperson<-length(a$person[a$person == "third"])
# Then divide them to get the ratio
ratio<-numberofFirstperson/numberofThirdperson
ratio
```

```
## [1] 0.4018692
```

The ratio is 0.4018692.

Problem 5

What is the avg length of first and third person novels? Does the difference seem large to you?

Solution:

```
# firstperson<-subset(a, a$person == "first")
# firstperson_avg<-mean(firstperson$length)
# firstperson_avg
# thirdperson<-subset(a, a$person == "third")
# thirdperson_avg<-mean(thirdperson$length)
# thirdperson_avg
# A faster way to do these calculations
tapply(a$length, a$person, mean)
```

```
##    first    third
## 119562.1 124718.6
```

I think the difference does not seam large to me.

Problem 6

What is the average length of 1P novels by women? How does this compare to 3P novels by women?

Solution:

```
female<-a[a$gender == "female",]
tapply(female$length, female$person, mean)
```

```
##    first    third
## 102254.5 133519.6
```

The average length of 1P novels by women is 102254.5 which is less than 3P novels by women(133519.6).

Problem 7

List the avg. length of novels by decade. What do you observe? Figure out how to plot for extra credit (insert screen shot here).

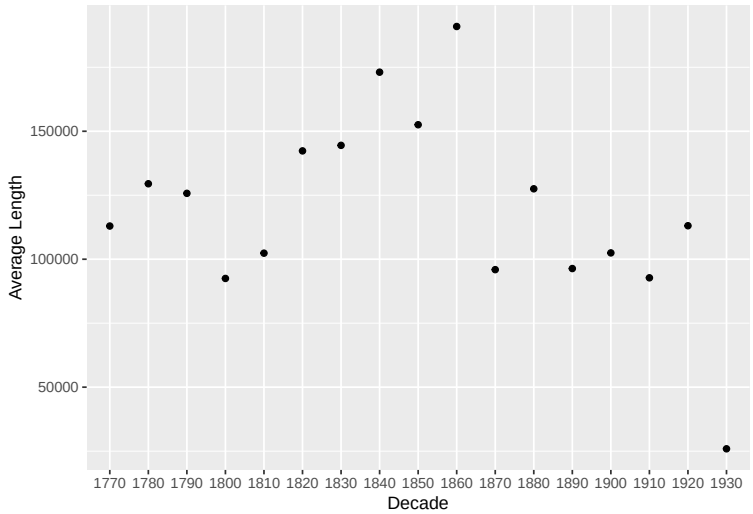
Solution:

```
# Convert the date column to a column of strings to utilize the substring function
a$decade<-as.character(a$date)
# Transform the 4th digit to a 0
substring(a$decade, 4, 4) <- "0"
# Convert back to integers
a$decade<-as.numeric(a$decade)
# Treat decades as factors in order to see what the avg. length is by decade.
dec.length<-tapply(a$length, as.factor(a$decade), mean)
dec.length
```

```
##      1770      1780      1790      1800      1810      1820      1830      1840
## 112967.67 129483.67 125740.00  92484.33 102366.78 142328.29 144493.40 173089.22
##      1850      1860      1870      1880      1890      1900      1910      1920
## 152546.93 190936.00  95921.80 127511.82  96354.56 102468.94  92732.33 113082.88
##      1930
##  25916.00
```

Already listed the avg. length of novels by decade. I observed that the avg.length of novels by decade was fluctuate and reached the peak in 1860. I used ggplot the visualized the result.

```
#scatterplot
dec.length.df<-as.data.frame(dec.length)
ggplot(aes(x = row.names(dec.length.df), y = dec.length), data = dec.length.df)+
  geom_point()+
  labs(x="Decade", y="Average Length")
```



Another way to visualize this is through the use of boxplots. These allow you to see the range of values for each decade which will give you a better sense of those periods that are particularly different.

```
ggplot(a, aes(x=factor(decade), y=length)) +
  geom_boxplot()+
  labs(x="Decade", y="Length")
```