Proving a Theory using Statistics and Data Visualization

This is a tribute to Dr. Ignaz Semmelweis who is considered as the pioneer of hand washing. During mid 1800s when 1 out of 6 women were dying during giving birth, Dr. Semmelweis observed a pattern why it was happening. Doctors at hospitals had a routine of performing autopsy at the morgue and returning to labor room to treat mothers giving birth without washing their hands. It started to cause a novel sickness in mothers at that time which was referred as child-bed fever. After implementing a hand washing routine by Dr. Semmelweis, cases of child-bed fever deaths reduced from 1/6 to 1/50 to almost practically zero. He was criticized back then by the doctors and their association and later had to lose his job and leave the city. He had no tools to prove his theory but luckily, we have sophisticated statistical tools today to prove his results statistically. I have tried to give him a tribute by proving his results statistically significant and showing a graph before and after hand washing has been implemented to depict how greatly cases went down.

Loading Tidyverse library

```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
## v purrr   0.3.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Importing data set into R

```
yearly <- read_csv("C:/MS ENM/Datasets/Dr. Semmelweis and the Discovery of Handwashing/datasets/yearly_
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   year = col_double(),
##   births = col_double(),
##   deaths = col_double(),
##   clinic = col_character()
## )
```

Printing out data frame

```
yearly
```

```
## # A tibble: 12 x 4
##     year births deaths clinic
##    <dbl>  <dbl>  <dbl> <chr>
##  1  1841   3036    237 clinic 1
##  2  1842   3287    518 clinic 1
##  3  1843   3060    274 clinic 1
##  4  1844   3157    260 clinic 1
##  5  1845   3492    241 clinic 1
##  6  1846   4010    459 clinic 1
##  7  1841   2442     86 clinic 2
##  8  1842   2659    202 clinic 2
##  9  1843   2739    164 clinic 2
## 10  1844   2956     68 clinic 2
## 11  1845   3241     66 clinic 2
## 12  1846   3754    105 clinic 2
```

Adding a new column that can be used in the analysis.

```
yearly <- yearly %>% mutate (proportion_deaths = deaths / births)
yearly
```

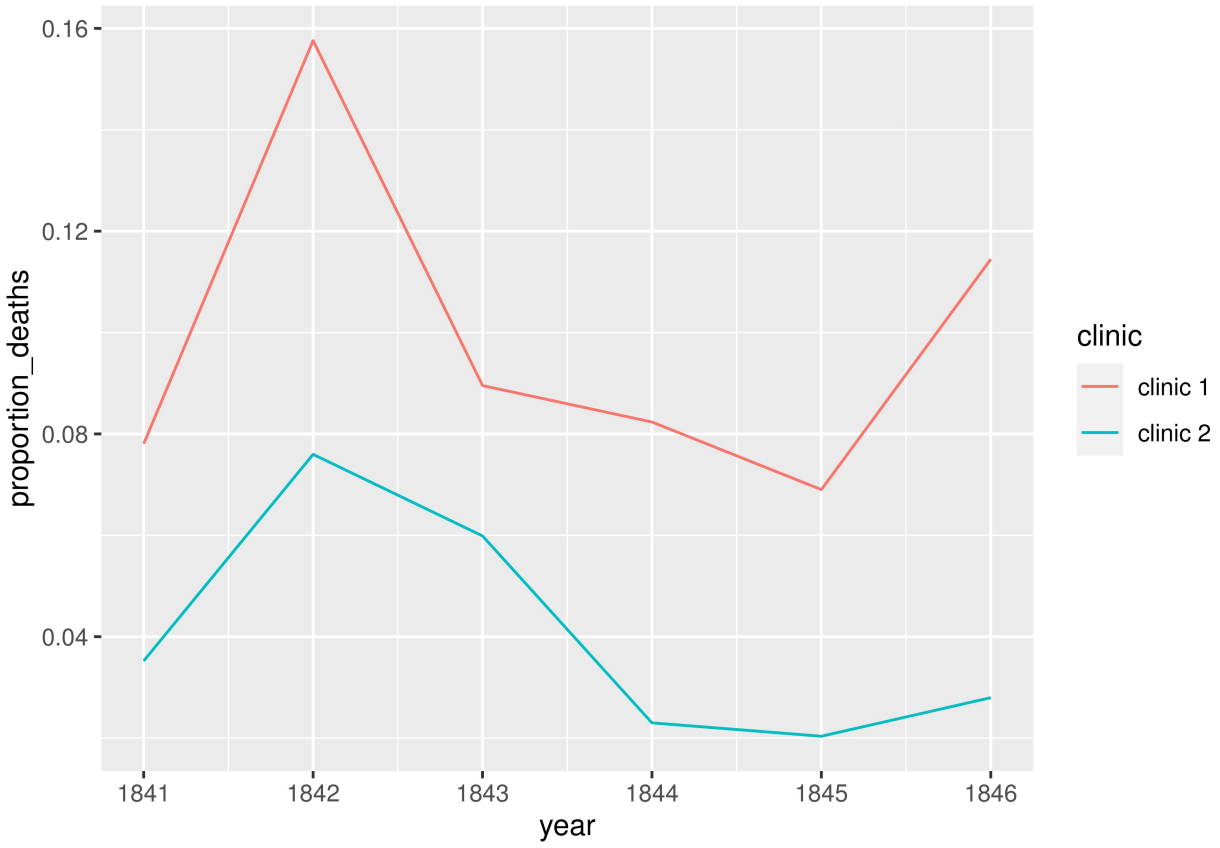
```
## # A tibble: 12 x 5
##     year births deaths clinic   proportion_deaths
##    <dbl>  <dbl>  <dbl> <chr>                <dbl>
##  1  1841   3036    237 clinic 1            0.0781
##  2  1842   3287    518 clinic 1            0.158
##  3  1843   3060    274 clinic 1            0.0895
##  4  1844   3157    260 clinic 1            0.0824
##  5  1845   3492    241 clinic 1            0.0690
##  6  1846   4010    459 clinic 1            0.114
##  7  1841   2442     86 clinic 2            0.0352
##  8  1842   2659    202 clinic 2            0.0760
##  9  1843   2739    164 clinic 2            0.0599
## 10  1844   2956     68 clinic 2            0.0230
## 11  1845   3241     66 clinic 2            0.0204
## 12  1846   3754    105 clinic 2            0.0280
```

Adjusting the size of the plot

```
options(repr.plot.width=7, repr.plot.height=4)
```

Difference in proportion of deaths between two clinics

```
ggplot(yearly, aes(year, proportion_deaths, color=clinic)) + geom_line()
```



Importing another data set

```
monthly <- read_csv("C:\\MS ENM\\Datasets\\Dr. Semmelweis and the Discovery of Handwashing\\datasets\\m
```

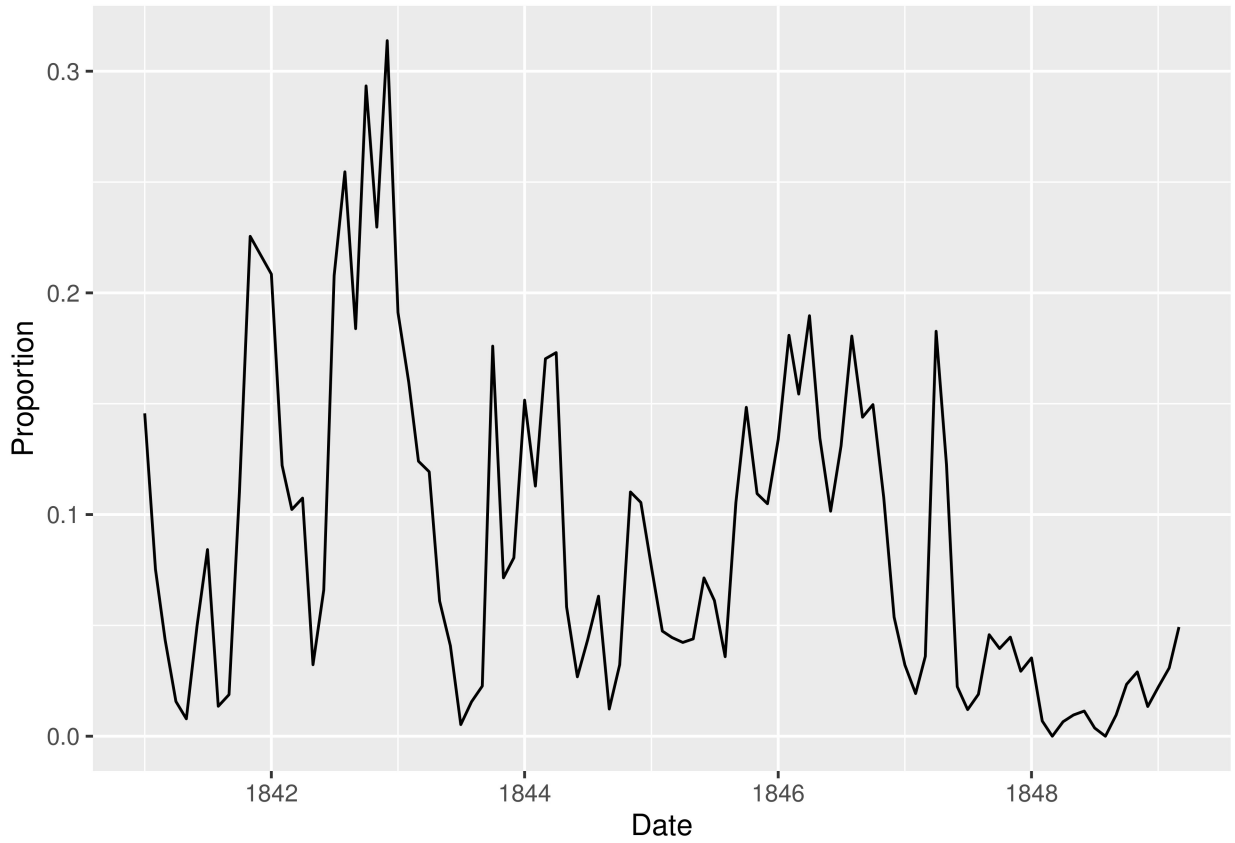
```
##
## -- Column specification --------------------------------------------------------
## cols(
##   date = col_date(format = ""),
##   births = col_double(),
##   deaths = col_double()
## )
```

Adding a new column

```
monthly <- monthly %>% mutate(proportion_deaths = deaths / births)
```

Time series graph of Proportion Deaths

```
ggplot(monthly, aes(date, proportion_deaths)) + geom_line() + xlab("Date") + ylab("Proportion")
```



This is the data when handwashing officially started

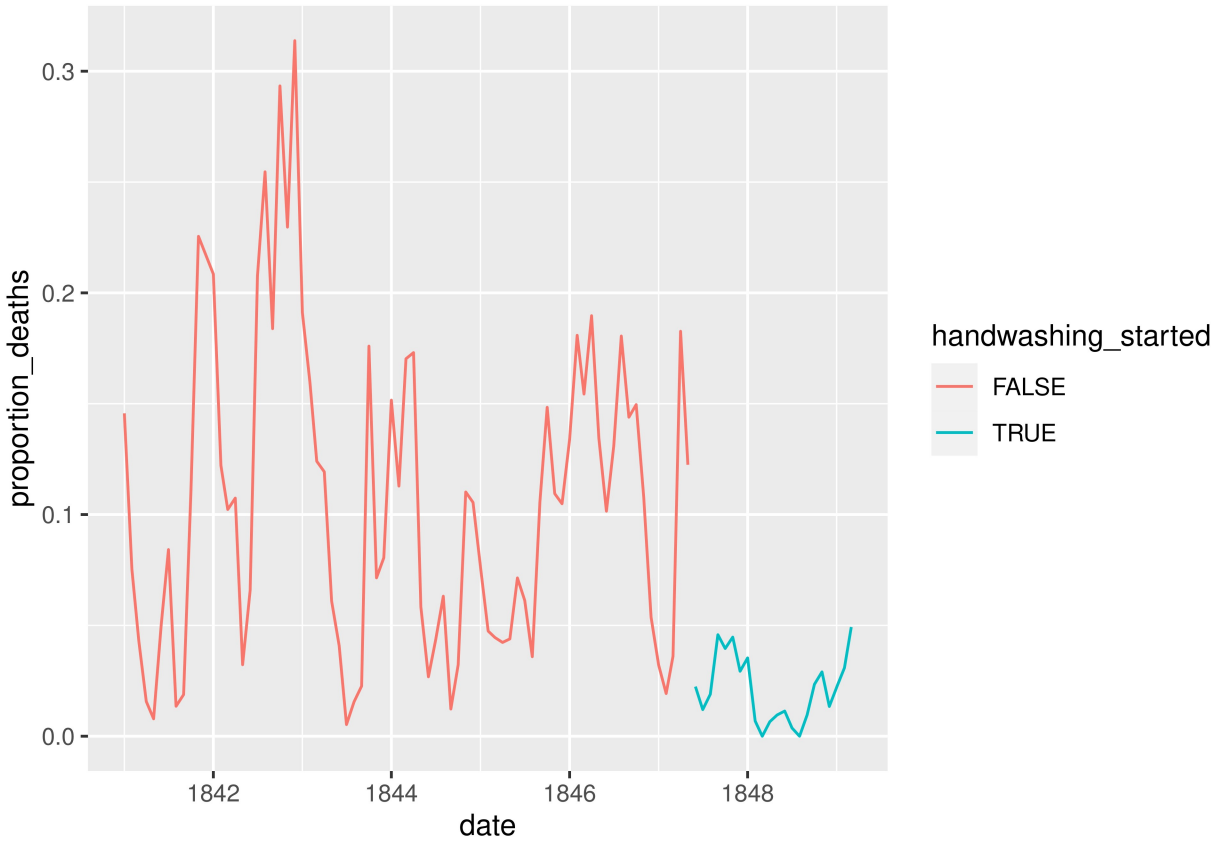
```
handwashing_start = as.Date('1847-06-01')
```

Creating a new column of dates after handwashing initiated

```
monthly <- monthly %>%
  mutate(handwashing_started = date >= handwashing_start)
```

Time series plot of Proportion deaths split by before and after hand washing started

```
ggplot(monthly, aes(x = date, y = proportion_deaths, color = handwashing_started)) +
  geom_line()
```




```
monthly_summary <- monthly %>% group_by(handwashing_started) %>% summarise(mean_proportion_deaths=mean(
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

Printing out monthly summary

```
monthly_summary
```

```
## # A tibble: 2 x 2
##   handwashing_started mean_proportion_deaths
##   <lgl>                                <dbl>
## 1 FALSE                               0.105
## 2 TRUE                                0.0211
```

Performing t-test to find out statistical difference in proportion deaths after hand washing started

```
test_result <- t.test( proportion_deaths ~ handwashing_started, data = monthly)
test_result
```

```
##
##  Welch Two Sample t-test
##
## data:  proportion_deaths by handwashing_started
```

```
## t = 9.6101, df = 92.435, p-value = 1.445e-15
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.06660662 0.10130659
## sample estimates:
## mean in group FALSE  mean in group TRUE
##          0.10504998          0.02109338
```

Result: There is huge reduction in proportion deaths after hand washing started, hence proved by statistical output of t-test

```
doctors_should_wash_their_hands <- TRUE
```