Covid Data

Hao Earm

Simple Data Analysis on Covid Dataset

Data Source: Kaggle https://www.kaggle.com/datasets/sudalairajkumar/novel-corona-virus-2019-dataset/versions/25

```
rm(list=ls())
data <- read.csv("COVID19_line_list_data.csv")
```

```
summary(data)
```

```
##        id       case_in_country   reporting.date        X
##  Min.   :   1   Min.   :   1.00   Length:1085        Mode:logical
##  1st Qu.: 272   1st Qu.:  11.00   Class :character   NA's:1085
##  Median : 543   Median :  28.00   Mode  :character
##  Mean   : 543   Mean   :  48.84
##  3rd Qu.: 814   3rd Qu.:  67.25
##  Max.   :1085   Max.   :1443.00
##                 NA's   :197
##    summary             location            country             gender
##  Length:1085        Length:1085        Length:1085        Length:1085
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##       age        symptom_onset      If_onset_approximated hosp_visit_date
##  Min.   : 0.25   Length:1085        Min.   :0.0000        Length:1085
##  1st Qu.:35.00   Class :character   1st Qu.:0.0000        Class :character
##  Median :51.00   Mode  :character   Median :0.0000        Mode  :character
##  Mean   :49.48                      Mean   :0.0429
##  3rd Qu.:64.00                      3rd Qu.:0.0000
##  Max.   :96.00                      Max.   :1.0000
##  NA's   :242                        NA's   :525
##  exposure_start     exposure_end       visiting.Wuhan    from.Wuhan
##  Length:1085        Length:1085        Min.   :0.000   Min.   :0.0000
##  Class :character   Class :character   1st Qu.:0.000   1st Qu.:0.0000
##  Mode  :character   Mode  :character   Median :0.000   Median :0.0000
##                                        Mean   :0.177   Mean   :0.1443
##                                        3rd Qu.:0.000   3rd Qu.:0.0000
##                                        Max.   :1.000   Max.   :1.0000
##                                                        NA's   :4
##     death              recovered            symptom              source
##  Length:1085        Length:1085        Length:1085        Length:1085
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##      link             X.1            X.2            X.3            X.4
##  Length:1085        Mode:logical   Mode:logical   Mode:logical   Mode:logical  
##  Class :character   NA's:1085      NA's:1085      NA's:1085      NA's:1085     
##  Mode  :character                                                              
##                                                                                
##                                                                                
##                                                                                
##                                                                                
##    X.5            X.6         
##  Mode:logical   Mode:logical  
##  NA's:1085      NA's:1085     
##                               
##                               
##                               
##                               
##                               
```

Cleaning up Data in $Death 14 Distinct values in $death Deaths are recorded as (0,1), but some rows have the date recorded instead

```
library(Hmisc)
#Cleaned up death col.
data$death_new <- as.integer(data$death != 0)

# Calculating Deathrate
sum(data$death_new) / nrow(data)
```

```
## [1] 0.05806452
```

Testing a possible claim:

Claim: Older people are more likely to die from Covid

```
dead = subset(data, death_new == 1)
alive = subset(data,death_new == 0)
# Calculating mean age to support claim,NA exists in age col
mean(dead$age, na.rm = TRUE)
```

```
## [1] 68.58621
```

```
mean(alive$age, na.rm = TRUE)
```

```
## [1] 48.07229
```

68.58621 and 48.07229 Is this statistically significant to support the claim?

```
# Using t.test , two-sided, and a confidence level of 0.95
t.test(alive$age, dead$age, alternative='two.sided', conf.level = 0.95)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  alive$age and dead$age
## t = -10.839, df = 72.234, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -24.28669 -16.74114
## sample estimates:
## mean of x mean of y 
##  48.07229  68.58621
```

From Student’s t-test p-value < 2.2e-16 If p < 0.05, null hypothesis is rejected with this p-value close to 0, we can reject the null hypothesis and conclude that the claim is statistically significant

Testing another possible claim

Gender has no effect on deaths from covid

```
men = subset(data, gender == "male")
women = subset(data,gender == "female")

# Calculating mean age to support claim , NA exists in age col
mean(men$death_new, na.rm = TRUE)
```

```
## [1] 0.08461538
```

```
mean(women$death_new, na.rm = TRUE)
```

```
## [1] 0.03664921
```

0.08461538 and 0.03664921 Is this statistically significant to support the claim? Using t.test , two-sided, and a confidence level of 0.95

```
t.test(men$death_new, women$death_new, alternative='two.sided', conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  men$death_new and women$death_new
## t = 3.084, df = 894.06, p-value = 0.002105
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.01744083 0.07849151
## sample estimates:
##  mean of x  mean of y
## 0.08461538 0.03664921
```

p-value of 0.002105,< 0.05 at 95% confidence level. Reject null hypothesis,
Men have higher death rates than compared to women for covid in this dataset is statistically significant