R\_project

PS/2021/078\_M.C.MADUSANKA

2025-02-02

# Clear environment and load libraries  
rm(list=ls())  
library(Hmisc)

## Warning: package 'Hmisc' was built under R version 4.4.1

##   
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':  
##   
## format.pval, units

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.4.1

# Load data  
data <- read.csv("C:\\Users\\M-C-S\\Downloads\\COVID19\_line\_list\_data.csv")  
describe(data)

## data   
##   
## 27 Variables 1085 Observations  
## --------------------------------------------------------------------------------  
## id   
## n missing distinct Info Mean pMedian Gmd .05   
## 1085 0 1085 1 543 543 362 55.2   
## .10 .25 .50 .75 .90 .95   
## 109.4 272.0 543.0 814.0 976.6 1030.8   
##   
## lowest : 1 2 3 4 5, highest: 1081 1082 1083 1084 1085  
## --------------------------------------------------------------------------------  
## case\_in\_country   
## n missing distinct Info Mean pMedian Gmd .05   
## 888 197 197 1 48.84 38 54.99 2.00   
## .10 .25 .50 .75 .90 .95   
## 4.00 11.00 28.00 67.25 110.30 153.65   
##   
## lowest : 1 2 3 4 5, highest: 365 443 875 925 1443  
## --------------------------------------------------------------------------------  
## reporting.date   
## n missing distinct   
## 1084 1 43   
##   
## lowest : 02/01/20 02/02/20 02/03/20 02/04/20 02/05/20   
## highest: 2/24/2020 2/25/2020 2/26/2020 2/27/2020 2/28/2020  
## --------------------------------------------------------------------------------  
## summary   
## n missing distinct   
## 1080 5 967   
##   
## lowest : confirmed COVID-19 pneumonia patient No.11 in Tianjin: female, 55, symptom onset on 01/23/2020, hospitalized on 01/23/2020, confirmed on 01/26/2020 confirmed COVID-19 pneumonia patient No.12 in Tianjin: female, 79, symptom onset on 01/24/2020, hospitalized on 01/24/2020, confirmed on 01/26/2020 confirmed COVID-19 pneumonia patient No.13 in Tianjin: female, 19, symptom onset on 01/19/2020, hospitalized on 01/20/2020, confirmed on 01/26/2020 confirmed COVID-19 pneumonia patient No.14 in Tianjin: male, 71, Wuhan resident, visited Malaysia from 01/19/2020 to 01/25/2020, arrived in Tianjin on 01/25/2020, symptom onset on 01/25/2020, hospitalized on 01/25/2020, confirmed on 01/26/2020 confirmed imported COVID-19 pneumonia patient in Gansu: female, 20, lives in Wuhan, arrived in Gansu on 01/18/2020, symptom onset on 01/19/2020, visit clinic on 01/24/2020, hospitalized on 01/24/2020.   
## highest: new recovered imported COVID-19 pneumonia patient in Beijing: female, returned to Beijing from Wuhan on 01/08/2020, symptom onset afterwards, recovered on 01/24/2020. new recovered imported COVID-19 pneumonia patient in Beijing: male, returned to Beijing from Wuhan on 01/08/2020, symptom onset afterwards, recovered on 01/25/2020. Second confirmed imported COVID-19 pneumonia patient in Guangxi: male, 46, in contact with individuals from Wuhan before symptom onset. symptom onset on 01/20/2020. Second confirmed imported COVID-19 pneumonia patient in Liaoning: male, 40, works in Wuhan, visit Fushun, Liaoning on 01/12/2020, symptom onset on 01/14/2020, visit clinic in Fushun Dalian on 01/19/2020. Second confirmed imported COVID-19 pneumonia patient in Sichuan: male, 57, Wuhan resident, visited Sichuan on 01/15/2020, symptom onset on 01/16/2020 and hospitalized.   
## --------------------------------------------------------------------------------  
## location   
## n missing distinct   
## 1085 0 156   
##   
## lowest : Afghanistan Aichi Prefecture Alappuzha Algeria Amiens   
## highest: Yunnan Zabaikalsky Zaragoza Zhejiang Zhuhai   
## --------------------------------------------------------------------------------  
## country   
## n missing distinct   
## 1085 0 38   
##   
## lowest : Afghanistan Algeria Australia Austria Bahrain   
## highest: Thailand UAE UK USA Vietnam   
## --------------------------------------------------------------------------------  
## gender   
## n missing distinct   
## 902 183 2   
##   
## Value female male  
## Frequency 382 520  
## Proportion 0.424 0.576  
## --------------------------------------------------------------------------------  
## age   
## n missing distinct Info Mean pMedian Gmd .05   
## 843 242 85 0.999 49.48 50 20.79 22.0   
## .10 .25 .50 .75 .90 .95   
## 25.0 35.0 51.0 64.0 75.0 78.9   
##   
## lowest : 0.25 0.5 1 2 4 , highest: 86 87 89 91 96   
## --------------------------------------------------------------------------------  
## symptom\_onset   
## n missing distinct   
## 563 522 70   
##   
## lowest : 01/02/20 01/03/20 01/04/20 01/05/20 01/06/20   
## highest: 2/22/2020 2/23/2020 2/24/2020 2/25/2020 2/26/2020  
## --------------------------------------------------------------------------------  
## If\_onset\_approximated   
## n missing distinct Info Sum Mean   
## 560 525 2 0.123 24 0.04286   
##   
## --------------------------------------------------------------------------------  
## hosp\_visit\_date   
## n missing distinct   
## 507 578 60   
##   
## lowest : 01/01/20 01/03/20 01/05/20 01/06/20 01/08/20   
## highest: 2/24/2020 2/25/2020 2/26/2020 2/27/2020 2/28/2020  
## --------------------------------------------------------------------------------  
## exposure\_start   
## n missing distinct   
## 128 957 39   
##   
## lowest : 01/03/20 01/06/20 01/08/20 01/09/20 01/10/20   
## highest: 2/15/2020 2/17/2020 2/19/2020 2/20/2020 2/21/2020  
## --------------------------------------------------------------------------------  
## exposure\_end   
## n missing distinct   
## 341 744 52   
##   
## lowest : 01/02/20 01/03/20 01/04/20 01/05/20 01/06/20   
## highest: 2/21/2020 2/22/2020 2/23/2020 2/24/2020 2/25/2020  
## --------------------------------------------------------------------------------  
## visiting.Wuhan   
## n missing distinct Info Sum Mean   
## 1085 0 2 0.437 192 0.177   
##   
## --------------------------------------------------------------------------------  
## from.Wuhan   
## n missing distinct Info Sum Mean   
## 1081 4 2 0.37 156 0.1443   
##   
## --------------------------------------------------------------------------------  
## death   
## n missing distinct   
## 1085 0 14   
##   
## 0 (1022, 0.942), 02/01/20 (1, 0.001), 1 (42, 0.039), 2/13/2020 (1, 0.001),  
## 2/14/2020 (1, 0.001), 2/19/2020 (2, 0.002), 2/21/2020 (2, 0.002), 2/22/2020 (1,  
## 0.001), 2/23/2020 (4, 0.004), 2/24/2020 (1, 0.001), 2/25/2020 (2, 0.002),  
## 2/26/2020 (3, 0.003), 2/27/2020 (2, 0.002), 2/28/2020 (1, 0.001)  
## --------------------------------------------------------------------------------  
## recovered   
## n missing distinct   
## 1085 0 32   
##   
## lowest : 0 02/02/20 02/04/20 02/05/20 02/06/20   
## highest: 2/24/2020 2/25/2020 2/26/2020 2/27/2020 2/28/2020  
## --------------------------------------------------------------------------------  
## symptom   
## n missing distinct   
## 270 815 108   
##   
## lowest : chest discomfort chills cold, fever, pneumonia cough cough with sputum   
## highest: throat pain, chills throat pain, fever tired vomiting, cough, fever, sore throat vomiting, diarrhea, fever, cough   
## --------------------------------------------------------------------------------  
## source   
## n missing distinct   
## 1085 0 85   
##   
## lowest : ABC ABC News Al Arabiya Aljazeera Bangkok Post   
## highest: Yonnhap News Agency 人民日报 人民日报官方微博 央视新闻 新浪   
## --------------------------------------------------------------------------------  
## link   
## n missing distinct   
## 1085 0 490   
##   
## lowest : http://behdasht.gov.ir/news/%DA%A9%D8%B1%D9%88%D9%86%D8%A7+%D9%88%DB%8C%D8%B1%D9%88%D8%B3/199807/%D8%AF%D8%B1+%D8%B1%D9%88%D8%B2%D9%87%D8%A7%DB%8C+%DA%AF%D8%B0%D8%B4%D8%AA%D9%87+735+%D8%A8%DB%8C%D9%85%D8%A7%D8%B1+%D8%A8%D8%A7+%D8%B9%D9%84%D8%A7%D8%A6%D9%85+%D8%B4%D8%A8%D9%87+%D8%A2%D9%86%D9%81%D9%84%D9%88%D8%A2%D9%86%D8%B2%D8%A7+%D8%AF%D8%B1+%DA%A9%D8%B4%D9%88%D8%B1+%D8%A8%D8%B3%D8%AA%D8%B1%DB%8C+%D8%B4%D8%AF%D9%86%D8%AF+%D8%A8%D8%B1+%D8%A7%D8%B3%D8%A7%D8%B3+%D8%A2%D8%AE%D8%B1%DB%8C%D9%86+%D9%86%D8%AA%D8%A7%DB%8C%D8%AC+%D8%A2%D8%B2%D9%85%D8%A7%DB%8C%D8%B4+%D9%87%D8%A7+%D8%A7%D8%A8%D8%AA%D9%84%D8%A7%DB%8C+13+%D9%85%D9%88%D8%B1%D8%AF+%D8%AF%DB%8C%DA%AF%D8%B1+%D8%A8%D9%87+%DA%A9%D9%88%D9%88%DB%8C%D8%AF19+%D9%82%D8%B7%D8%B9%DB%8C+%D8%A8%D9%87+%D9%86%D8%B8%D8%B1+%D9%85%DB%8C+%D8%B1%D8%B3%D8%AF http://english.alarabiya.net/en/News/gulf/2020/02/25/Number-of-Kuwait-coronavirus-cases-rises-to-eight-KUNA.html http://sxwjw.shaanxi.gov.cn/art/2020/1/27/art\_9\_67483.html http://wjw.beijing.gov.cn/xwzx\_20031/wnxw/202001/t20200121\_1620353.html http://wjw.sz.gov.cn/wzx/202001/t20200120\_18987787.htm   
## highest: https://www3.nhk.or.jp/nhkworld/en/news/20200116\_23/ https://www3.nhk.or.jp/nhkworld/en/news/20200124\_14/ https://www3.nhk.or.jp/nhkworld/en/news/20200126\_31/ https://www3.nhk.or.jp/nhkworld/en/news/20200130\_02/ https://www3.nhk.or.jp/nhkworld/en/news/20200131\_01/   
## --------------------------------------------------------------------------------  
##   
## Variables with all observations missing:  
##   
## [1] X X.1 X.2 X.3 X.4 X.5 X.6

# Clean up death column  
data$death\_dummy <- as.integer(data$death != 0)  
death\_rate <- sum(data$death\_dummy) / nrow(data)  
print(paste("Death rate:", death\_rate))

## [1] "Death rate: 0.0580645161290323"

# Handle missing values  
data <- data[complete.cases(data$age, data$gender), ]  
  
# Analyze age  
dead <- subset(data, death\_dummy == 1)  
alive <- subset(data, death\_dummy == 0)  
mean\_age\_dead <- mean(dead$age, na.rm = TRUE)  
mean\_age\_alive <- mean(alive$age, na.rm = TRUE)  
print(paste("Mean age (dead):", mean\_age\_dead))

## [1] "Mean age (dead): 68.5862068965517"

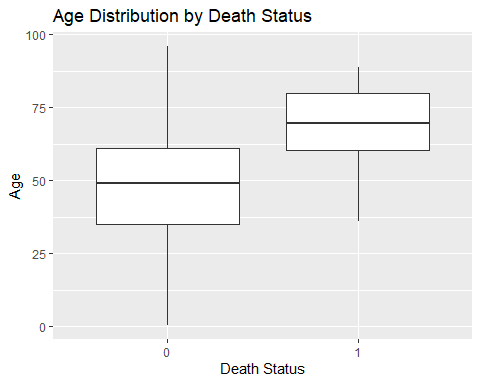
print(paste("Mean age (alive):", mean\_age\_alive))

## [1] "Mean age (alive): 48.3318122555411"

# T-test for age  
t\_test\_age <- t.test(alive$age, dead$age, alternative = "two.sided", conf.level = 0.99)  
print(t\_test\_age)

##   
## Welch Two Sample t-test  
##   
## data: alive$age and dead$age  
## t = -10.707, df = 72.093, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 99 percent confidence interval:  
## -25.25953 -15.24926  
## sample estimates:  
## mean of x mean of y   
## 48.33181 68.58621

# Visualize age distribution  
ggplot(data, aes(x = factor(death\_dummy), y = age)) +  
 geom\_boxplot() +  
 labs(x = "Death Status", y = "Age", title = "Age Distribution by Death Status")



# Analyze gender  
men <- subset(data, gender == "male")  
women <- subset(data, gender == "female")  
death\_rate\_men <- mean(men$death\_dummy, na.rm = TRUE)  
death\_rate\_women <- mean(women$death\_dummy, na.rm = TRUE)  
print(paste("Death rate (men):", death\_rate\_men))

## [1] "Death rate (men): 0.092436974789916"

print(paste("Death rate (women):", death\_rate\_women))

## [1] "Death rate (women): 0.0401146131805158"

# T-test for gender  
t\_test\_gender <- t.test(men$death\_dummy, women$death\_dummy, alternative = "two.sided", conf.level = 0.99)  
print(t\_test\_gender)

##   
## Welch Two Sample t-test  
##   
## data: men$death\_dummy and women$death\_dummy  
## t = 3.0871, df = 818.24, p-value = 0.00209  
## alternative hypothesis: true difference in means is not equal to 0  
## 99 percent confidence interval:  
## 0.008562968 0.096081756  
## sample estimates:  
## mean of x mean of y   
## 0.09243697 0.04011461

# Confidence interval for gender difference  
prop\_test\_gender <- prop.test(x = c(sum(men$death\_dummy), sum(women$death\_dummy)), n = c(nrow(men), nrow(women)), conf.level = 0.99)  
print(prop\_test\_gender)

##   
## 2-sample test for equality of proportions with continuity correction  
##   
## data: c(sum(men$death\_dummy), sum(women$death\_dummy)) out of c(nrow(men), nrow(women))  
## X-squared = 7.6526, df = 1, p-value = 0.005669  
## alternative hypothesis: two.sided  
## 99 percent confidence interval:  
## 0.006234264 0.098410459  
## sample estimates:  
## prop 1 prop 2   
## 0.09243697 0.04011461