# COVID-19 ANALYSIS

Final Project - Predictive Analytics

# Meet the Team

Nimeelitha Akkiraju

Bennie Amani

Keerthi Bojja

Pranidhi Prabhat

# Project Objective

- Time Series Analysis – Forecasting the confirmed, death and survival numbers
- Is there a Country level & Age level effect on the Death Rate ?
- Do Male/Female recover/die more than the other gender after contracting the disease ?
- Which ages are more prone to death/recover faster after getting the disease ?

# About the Data

- Shape of the data : 3397, 21
- See NA Values : -->

```
data_p.isnull().sum()
```

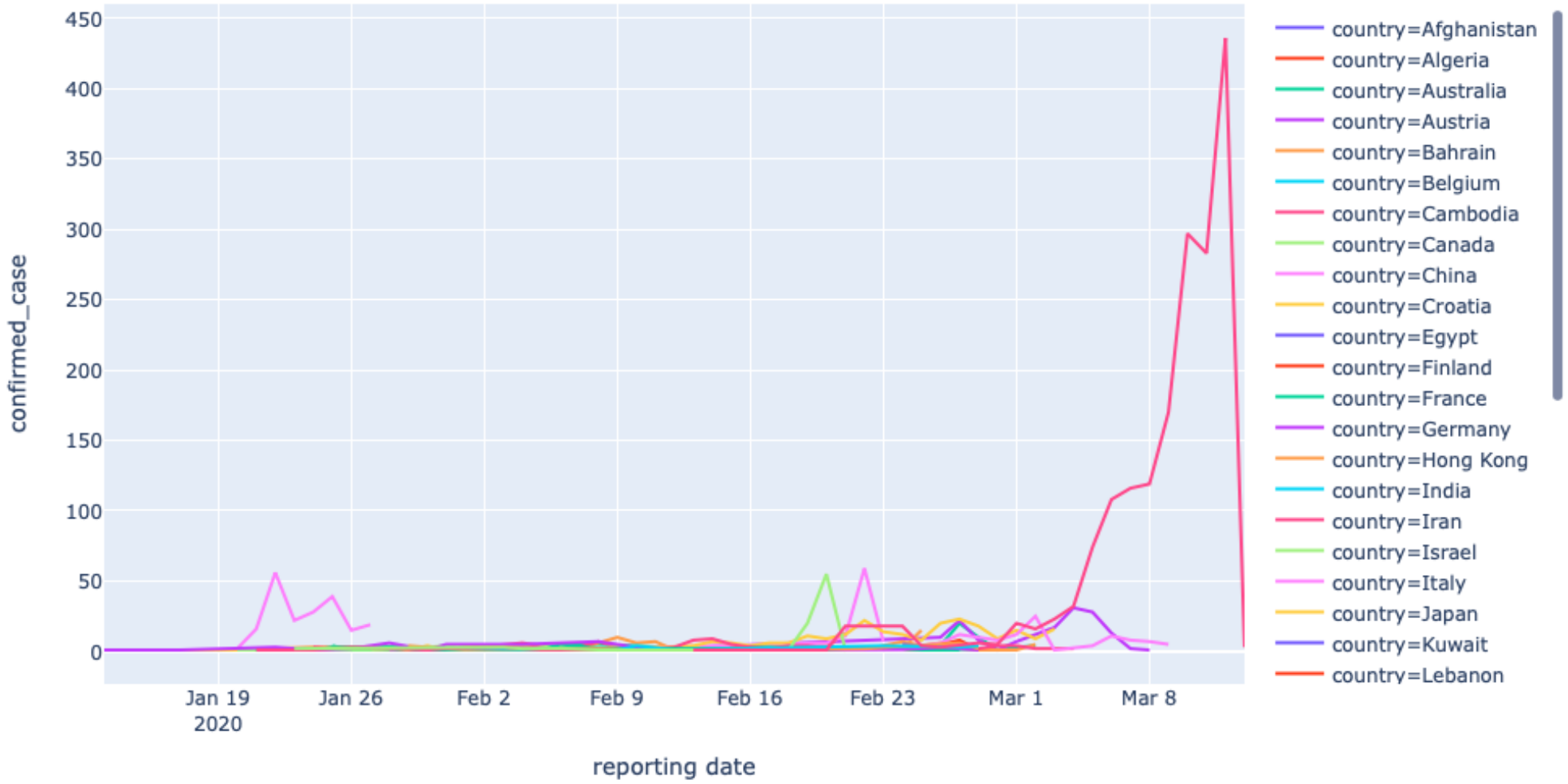
```
id                          0
case_in_country           197
reporting date              1
summary                     5
location                  230
country                   163
gender                   1712
age                      1890
symptom_onset            2713
If_onset_approximated    2742
hosp_visit_date          2777
international_traveler      0
domestic_traveler           0
exposure_start           3203
exposure_end             2909
traveler                 2737
visiting Wuhan           1791
from Wuhan               1795
death                    1778
recovered                1791
symptom                     0
confirmed_case              0
dtype: int64
```

# Exploratory Data Analysis

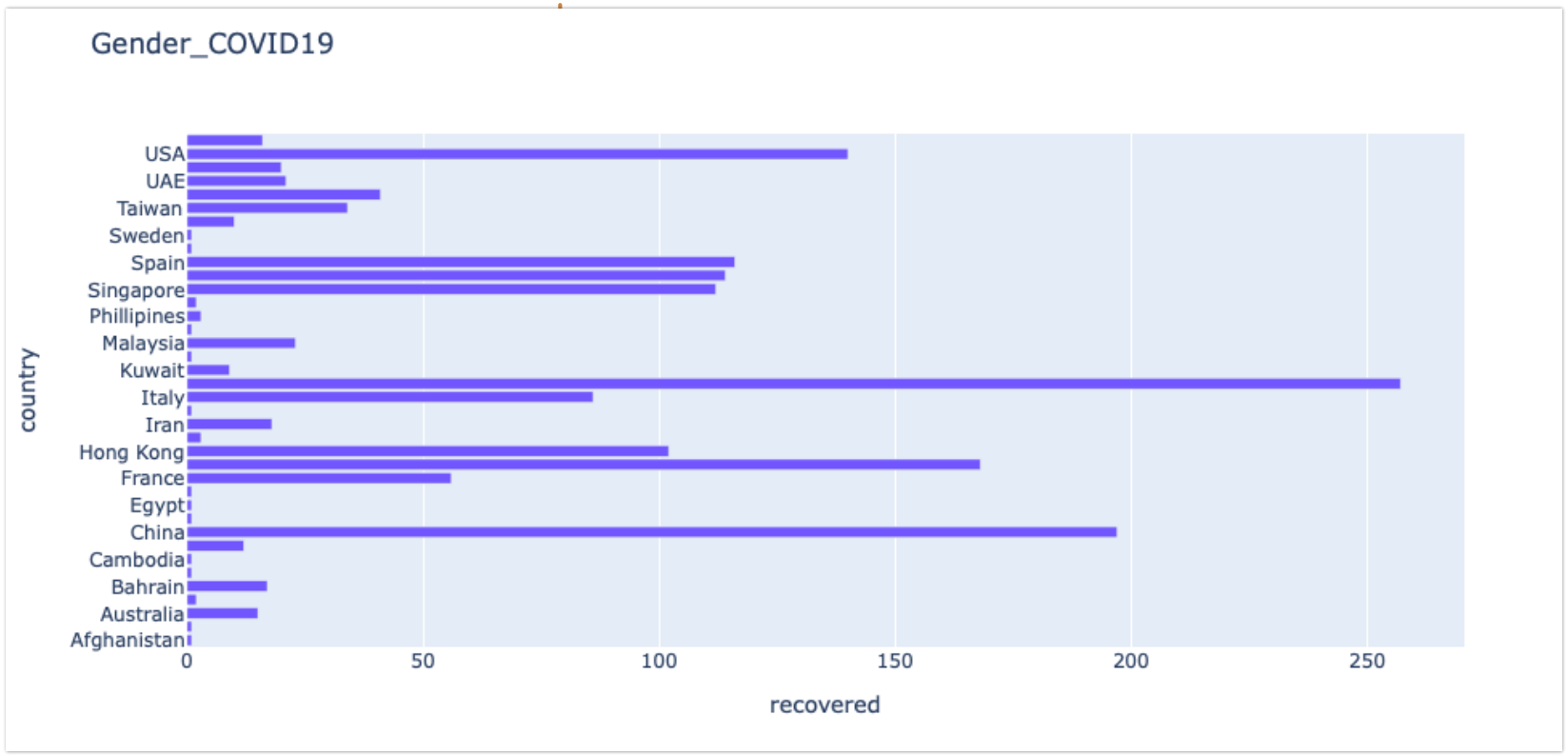
- Country – wise Numbers

| | country | confirmed_case |
|---|---|---|
| 0 | Afghanistan | 1 |
| 1 | Algeria | 1 |
| 2 | Australia | 15 |
| 3 | Austria | 2 |
| 4 | Bahrain | 17 |
| 5 | Belgium | 1 |
| 6 | Cambodia | 1 |
| 7 | Canada | 12 |
| 8 | China | 197 |
| 9 | Croatia | 1 |
| 10 | Egypt | 1 |
| 11 | Finland | 1 |
| 12 | France | 56 |
| 13 | Germany | 168 |
| 14 | Hong Kong | 102 |
| 15 | India | 3 |
| 16 | Iran | 18 |
| 17 | Israel | 1 |
| 18 | Italy | 86 |
| 19 | Japan | 257 |
| 20 | Kuwait | 9 |
| 21 | Lebanon | 1 |
| 22 | Malaysia | 23 |
| 23 | Nepal | 1 |
| 24 | Phillipines | 3 |
| 25 | Russia | 2 |
| 26 | Singapore | 112 |
| 27 | South Korea | 114 |
| 28 | Spain | 116 |
| 29 | Sri Lanka | 1 |
| 30 | Sweden | 1 |
| 31 | Switzerland | 10 |
| 32 | Taiwan | 34 |
| 33 | Thailand | 41 |
| 34 | UAE | 21 |
| 35 | UK | 20 |
| 36 | USA | 1768 |
| 37 | Vietnam | 16 |

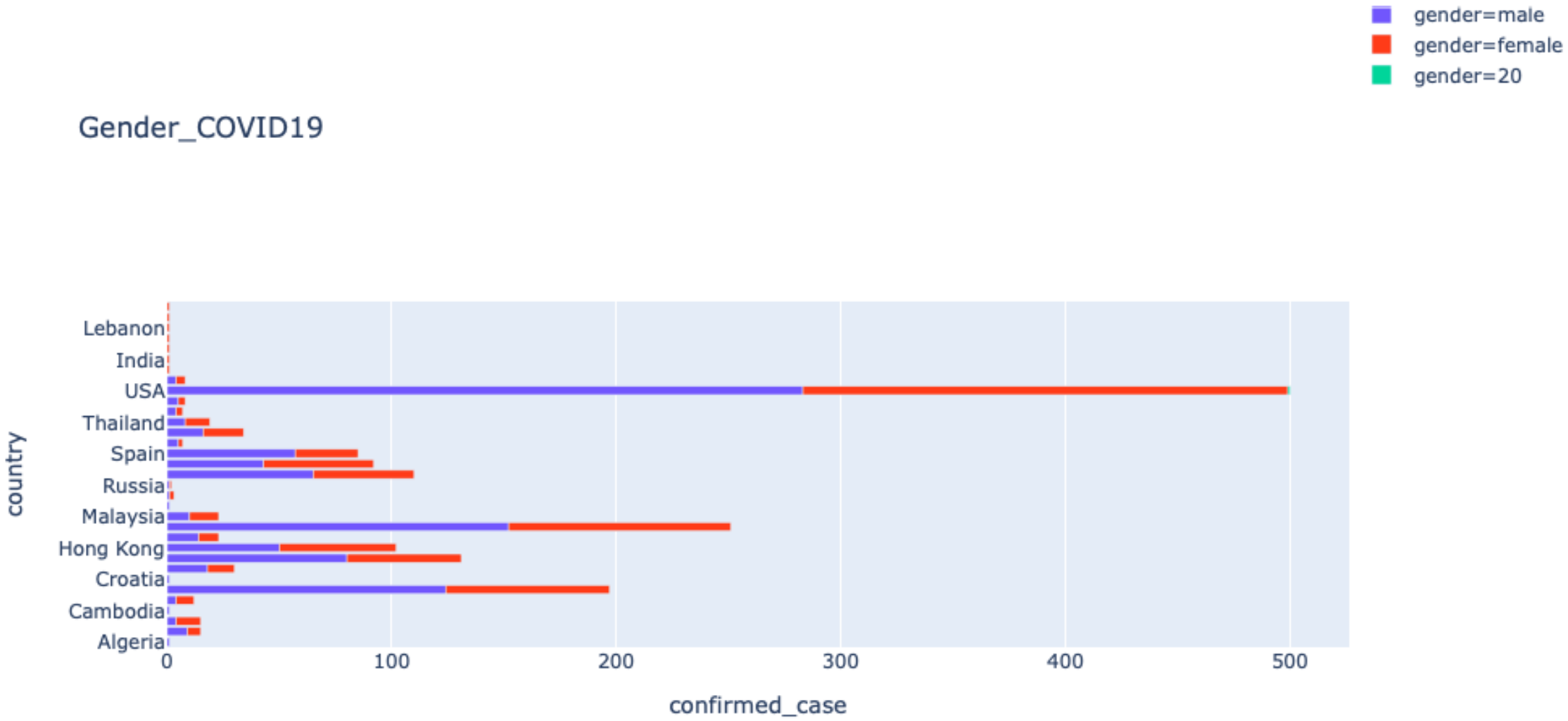
# Country Wise EDA



# Country Wise Survival Numbers

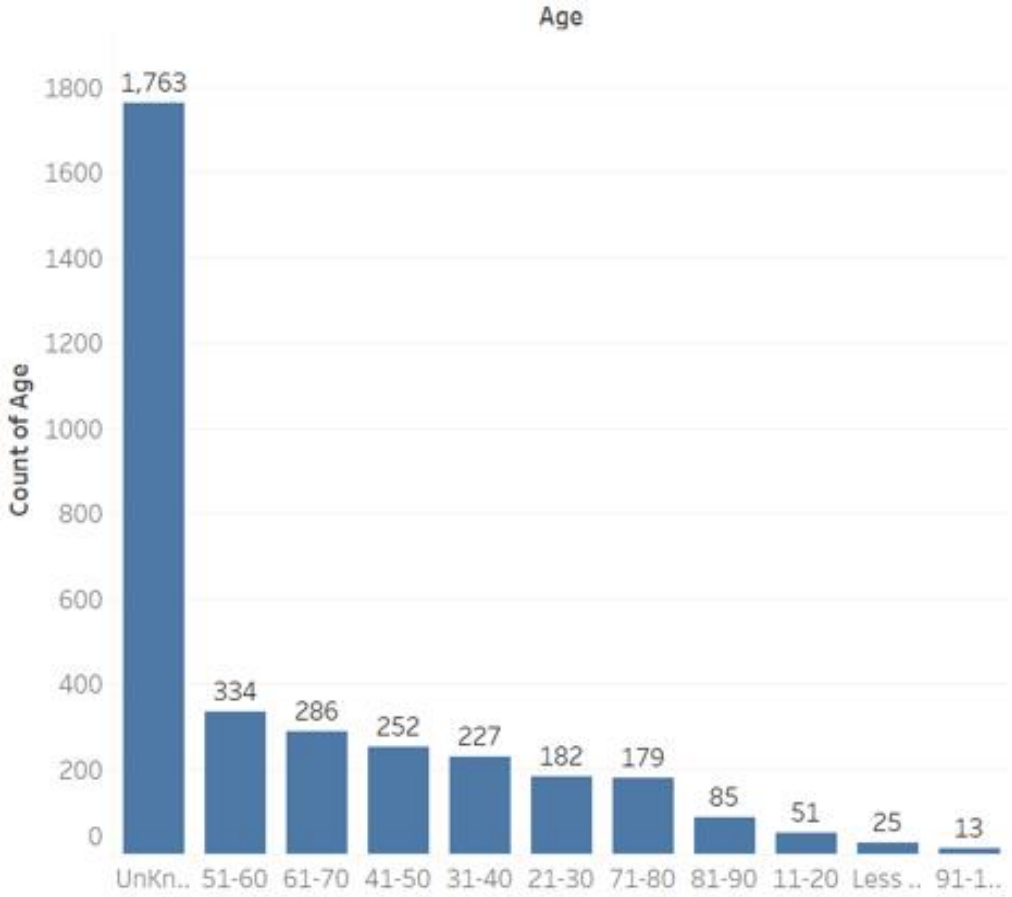


# Gender Wise EDA

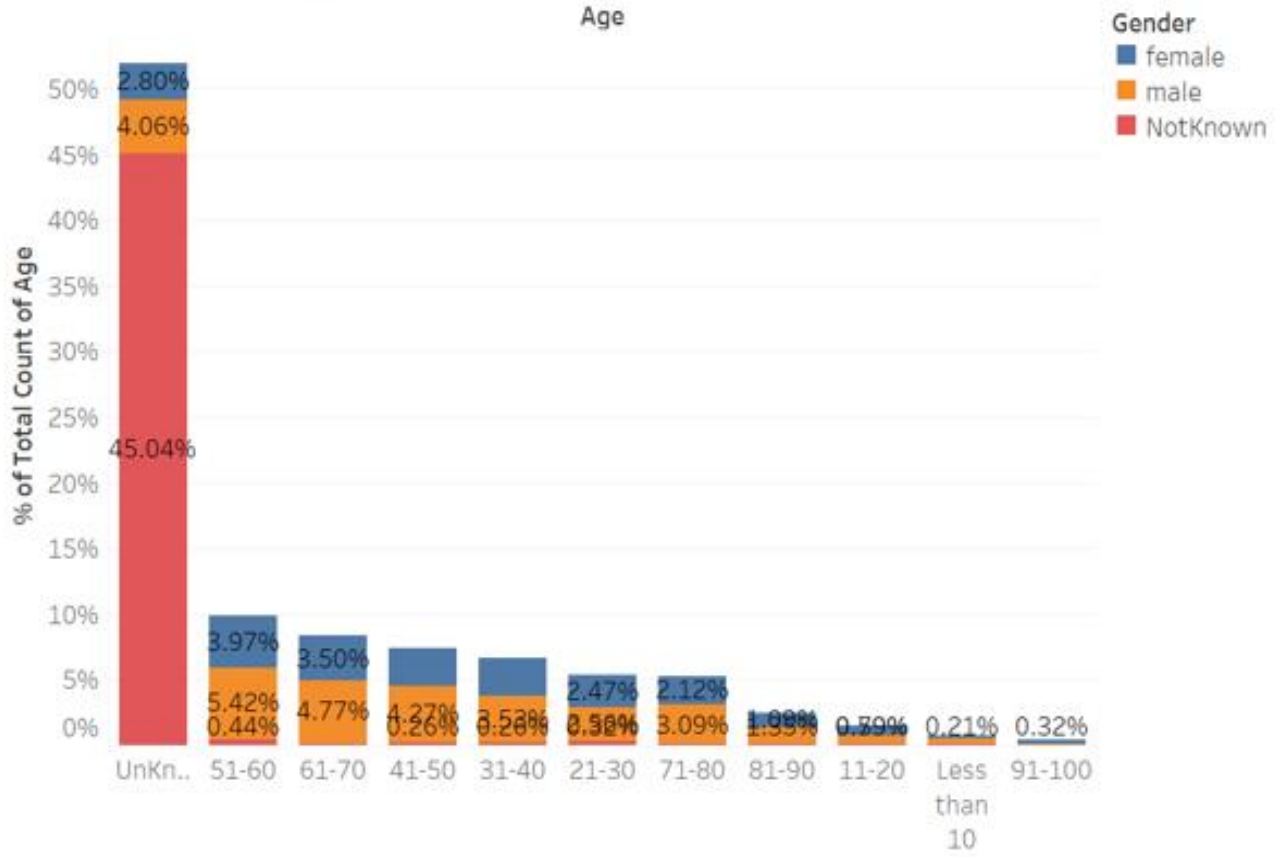


# Age analysis

COnfirmed cases Age-wise



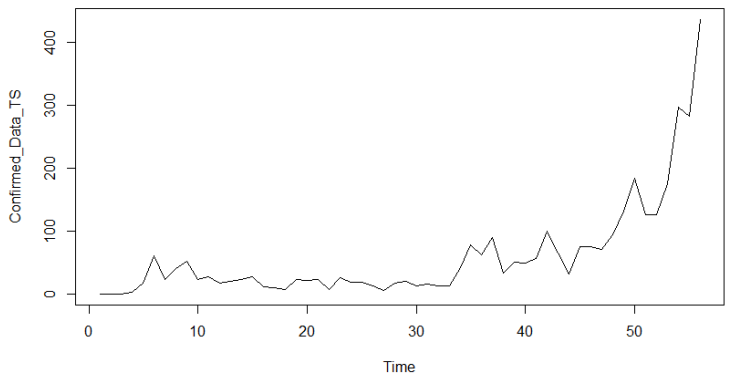

Age-Gender split

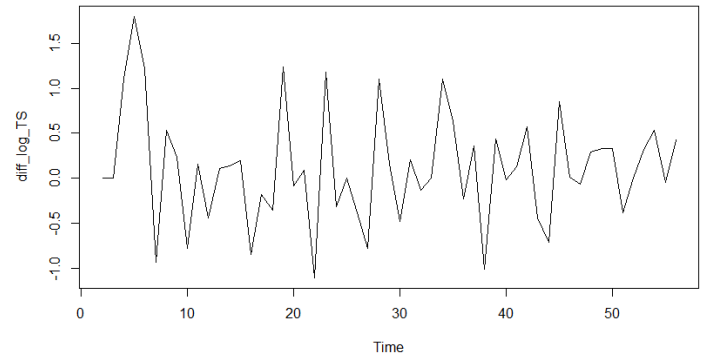


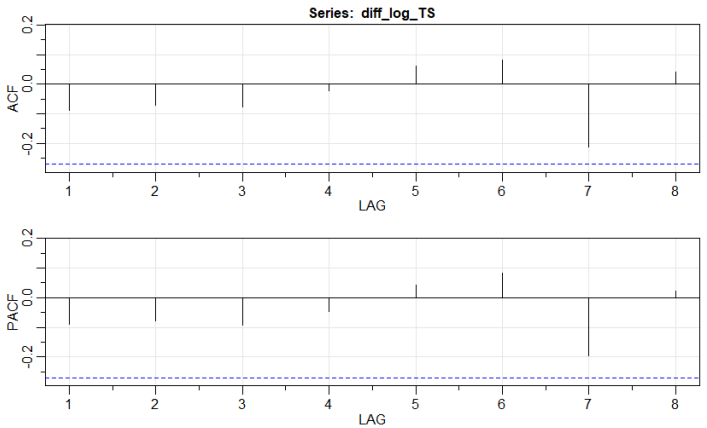
# Time Series Analysis

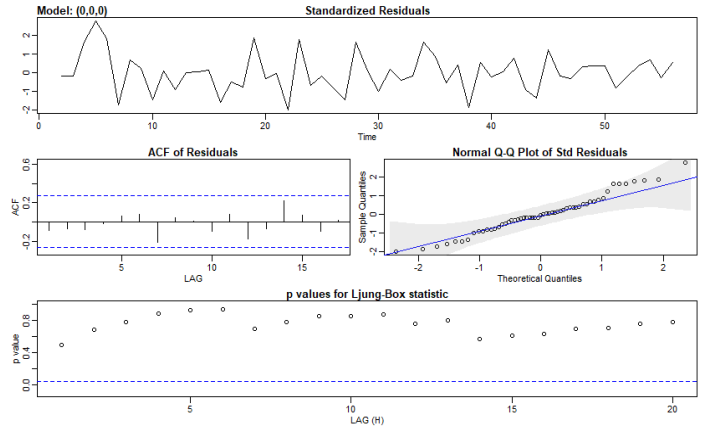
# Time Series Forecasting

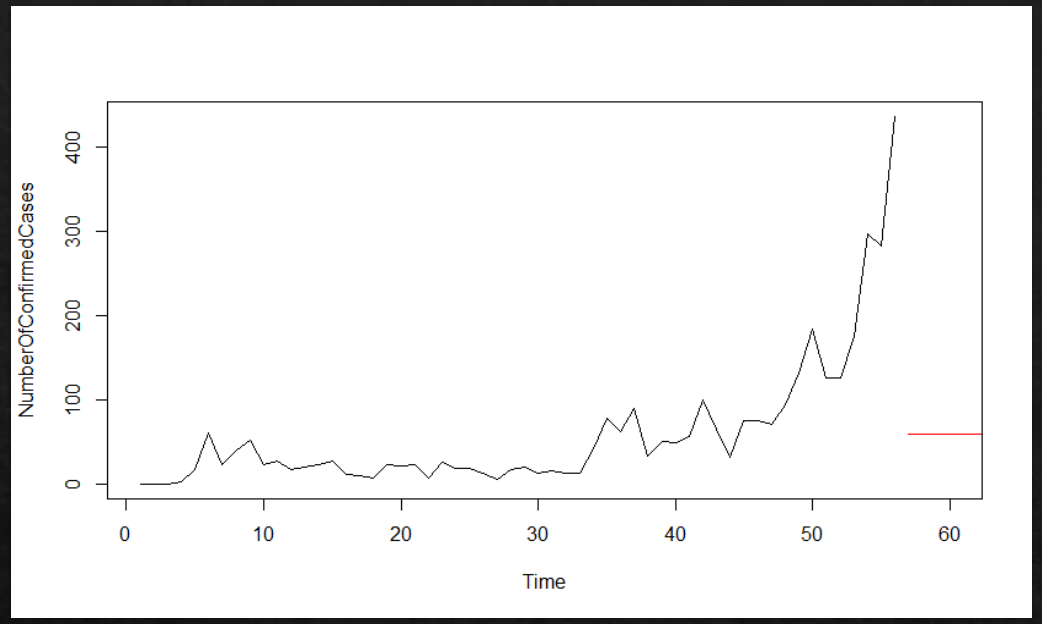
- Stationary time series
- ACF/PACF plots
- Number of Confirmed Cases
  - ARMA(0,0)
- Number of Deaths
  - ARMA(0,0)
- Number of Recoveries
  - ARMA(1,1)



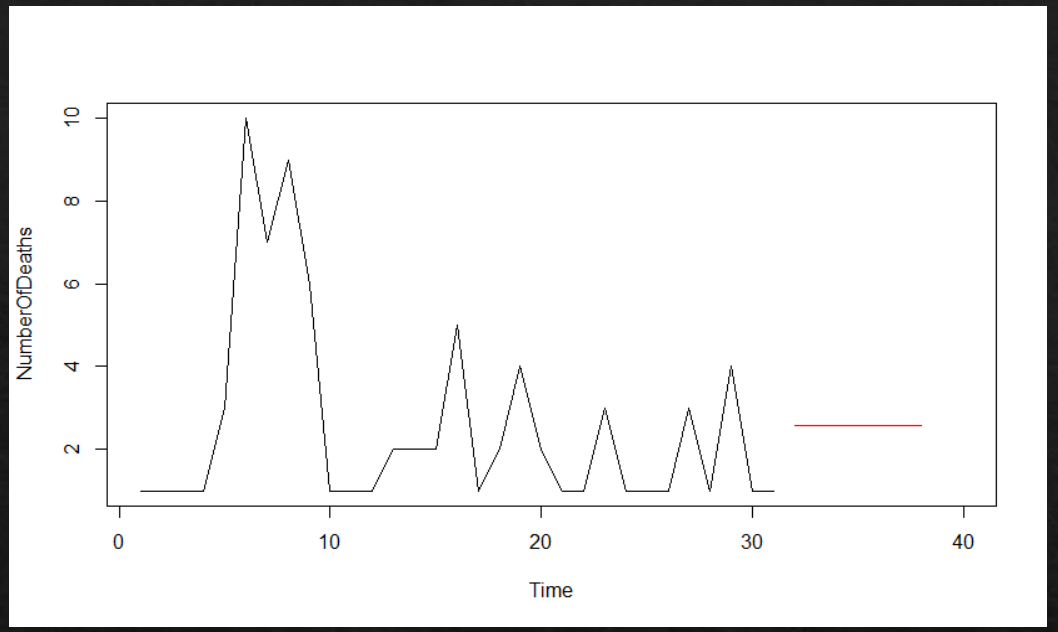



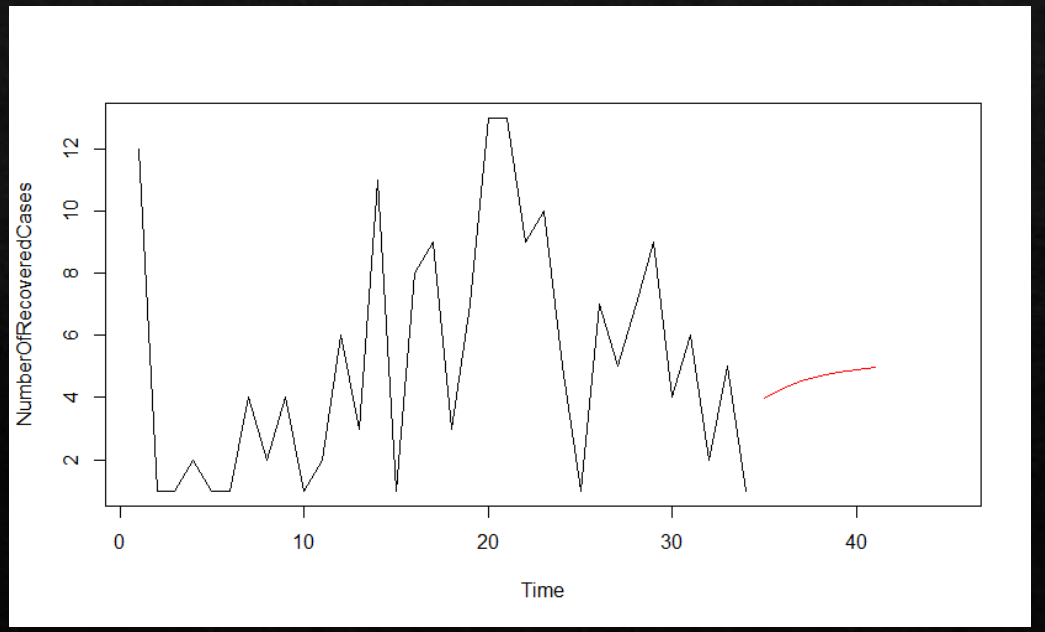






Number of Confirmed Cases




Number of Deaths




Number of Recoveries

# Fixed/Random Effect

# Fixed/Random Effect Approach

- Methodology
- Country-level effect
- Age-level effect

# Data Pre-processing

- Death column
- Filter required data
- Selecting relevant x variables

# Fixed/Random Effect

- X variables:
  - gender
  - Visiting.wuhan
  - From.wuhan
  - Country
- Methodology:
  - LmerTest package
  - Lmer model

```
> lm_country = lm(death~gender+visiting.Wuhan+from.Wuhan+factor(country)-1, data=data)
> summary(lm_country)

Call:
lm(formula = death ~ gender + visiting.Wuhan + from.Wuhan + factor(country) -
    1, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.40347 -0.06273 -0.01889  0.00000  1.08292

Coefficients: (1 not defined because of singularities)
                          Estimate Std. Error t value Pr(>|t|)
genderfemale              6.273e-02  2.112e-02   2.970  0.00303 **
gendermale                6.305e-02  2.031e-02   3.104  0.00194 **
genderNotKnown            5.146e-14  1.922e-01   0.000  1.00000
visiting.Wuhan           -1.011e-01  1.833e-02  -5.512 4.15e-08 ***
from.Wuhan                1.093e-01  2.048e-02   5.338 1.08e-07 ***
factor(country)Algeria   -4.327e-14  2.718e-01   0.000  1.00000
factor(country)Australia -1.790e-02  1.991e-01  -0.090  0.92838
factor(country)Austria   -5.156e-14  2.354e-01   0.000  1.00000
factor(country)Bahrain   -5.240e-14  1.977e-01   0.000  1.00000
factor(country)Belgium    1.011e-01  2.724e-01   0.371  0.71070
factor(country)Cambodia  -1.093e-01  2.725e-01  -0.401  0.68832
factor(country)Canada     2.388e-02  2.002e-01   0.119  0.90509
factor(country)China      1.361e-01  2.671e-02   5.098 3.87e-07 ***
factor(country)Croatia   -3.816e-14  2.718e-01   0.000  1.00000
factor(country)Egypt     -5.387e-14  2.718e-01   0.000  1.00000
factor(country)Finland   -1.093e-01  2.725e-01  -0.401  0.68832
factor(country)France    -2.946e-02  3.229e-02  -0.913  0.36165
factor(country)Germany    1.203e-03  1.927e-01   0.006  0.99502
factor(country)Hong Kong  1.813e-02  1.931e-01   0.094  0.92522
factor(country)India      1.011e-01  2.227e-01   0.454  0.65000
factor(country)Iran       2.222e-01  1.974e-01   1.126  0.26055
factor(country)Israel    -5.563e-14  2.718e-01   0.000  1.00000
factor(country)Italy      2.678e-02  1.933e-01   0.139  0.88982
factor(country)Japan     -4.384e-02  2.294e-02  -1.911  0.05620 .
factor(country)Kuwait    -5.161e-14  2.026e-01   0.000  1.00000
factor(country)Lebanon   -5.460e-14  2.718e-01   0.000  1.00000
factor(country)Malaysia  -8.255e-02  4.506e-02  -1.832  0.06712 .
factor(country)Nepal     -1.724e-01  1.942e-01  -0.888  0.37484
factor(country)Phillipines  2.941e-01  2.226e-01   1.322  0.18651
factor(country)Russia     1.011e-01  2.361e-01   0.428  0.66867
factor(country)Singapore -6.092e-02  2.674e-02  -2.278  0.02284 *
factor(country)South Korea  2.211e-02  2.657e-02   0.832  0.40554
```

```
> summary(lmer(death ~ age + visiting.Wuhan + from.Wuhan + (1|country)-1, data=data))
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: death ~ age + visiting.Wuhan + from.Wuhan + (1 | country) - 1
   Data: data

REML criterion at convergence: -710.1

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.0296 -0.2751 -0.0786  0.0066  5.7001

Random effects:
 Groups   Name        Variance Std.Dev.
 country  (Intercept) 0.003087 0.05556
 Residual             0.034883 0.18677
Number of obs: 1559, groups:  country, 38

Fixed effects:
                Estimate Std. Error         df t value Pr(>|t|)
age20-Nov      -0.02602    0.04458  371.18807  -0.584  0.55972
age21-30       -0.01083    0.02738   62.10364  -0.396  0.69383
age31-40       -0.01204    0.02615   51.67795  -0.461  0.64707
age41-50        0.03245    0.02490   42.35962   1.303  0.19950
age51-60        0.02596    0.02384   36.14122   1.089  0.28336
age61-70        0.08473    0.02405   37.23546   3.523  0.00115 **
age71-80        0.08061    0.02632   52.80970   3.063  0.00345 **
age81-90        0.23256    0.03330  131.39506   6.983 1.29e-10 ***
```

```
> rand(model_country)
ANOVA-like table for random-effects: Single term deletions

Model:
death ~ age + visiting.Wuhan + from.Wuhan + (1 | country)
                npar logLik     AIC    LRT Df Pr(>Chisq)
<none>            15 355.06 -680.12
(1 | country)     14 312.07 -596.14 85.981  1  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

# Country-level effect

```
> lm_age = lm(death~gender+visiting.Wuhan+from.Wuhan+factor(age)-1, data=data)
> summary(lm_age)

Call:
lm(formula = death ~ gender + visiting.Wuhan + from.Wuhan + factor(age) -
    1, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.39619 -0.05041 -0.00866 -0.00866  1.01504

Coefficients: (1 not defined because of singularities)
                       Estimate Std. Error t value Pr(>|t|)
genderfemale           -0.005917   0.042033  -0.141   0.8881
gendermale              0.004079   0.042219   0.097   0.9230
genderNotKnown          0.008662   0.007834   1.106   0.2691
visiting.Wuhan         -0.023698   0.015477  -1.531   0.1259
from.Wuhan              0.172400   0.016835  10.241  < 2e-16 ***
factor(age)21-30       -0.011438   0.045950  -0.249   0.8034
factor(age)31-40       -0.001018   0.045071  -0.023   0.9820
factor(age)41-50        0.041195   0.044454   0.927   0.3542
factor(age)51-60        0.015249   0.043807   0.348   0.7278
factor(age)61-70        0.080526   0.044048   1.828   0.0677 .
factor(age)71-80        0.068818   0.045387   1.516   0.1297
factor(age)81-90        0.229711   0.049970   4.597 4.64e-06 ***
factor(age)91-100       0.174867   0.089514   1.954   0.0509 .
factor(age)Less than 10 0.055828   0.069803   0.800   0.4240
factor(age)UnKnown             NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> model_age = lmer(death ~ gender+ visiting.Wuhan + from.Wuhan + (1|age)-1+(1|countr
y), data=data)
> summary(model_age)
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: death ~ gender + visiting.Wuhan + from.Wuhan + (1 | age) - 1 +
    (1 | country)
   Data: data

REML criterion at convergence: -719.7

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.9230 -0.2840 -0.0987  0.0019  5.7005

Random effects:
 Groups   Name        Variance Std.Dev.
 country  (Intercept) 0.003052 0.05525
 age      (Intercept) 0.005830 0.07635
 Residual             0.034918 0.18686
Number of obs: 1559, groups:  country, 38; age, 11

Fixed effects:
                Estimate Std. Error         df t value Pr(>|t|)
genderfemale     0.06111    0.03315   18.15238   1.844   0.0816 .
gendermale       0.06049    0.03307   17.88423   1.829   0.0841 .
genderNotKnown   0.02594    0.07835    6.98427   0.331   0.7503
visiting.Wuhan  -0.08367    0.01700 1326.49605  -4.922 9.63e-07 ***
from.Wuhan       0.10385    0.01897 1312.95196   5.473 5.28e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> rand(model_age)
ANOVA-like table for random-effects: Single term deletions

Model:
death ~ gender + visiting.Wuhan + from.Wuhan + (1 | age) + (1 |
    country) - 1
             npar logLik     AIC    LRT Df Pr(>Chisq)
<none>          8 359.86 -703.72
(1 | age)       7 332.27 -650.54 55.179  1    1.1e-13 ***
(1 | country)   7 317.97 -621.93 83.788  1    < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

# Age-level effect

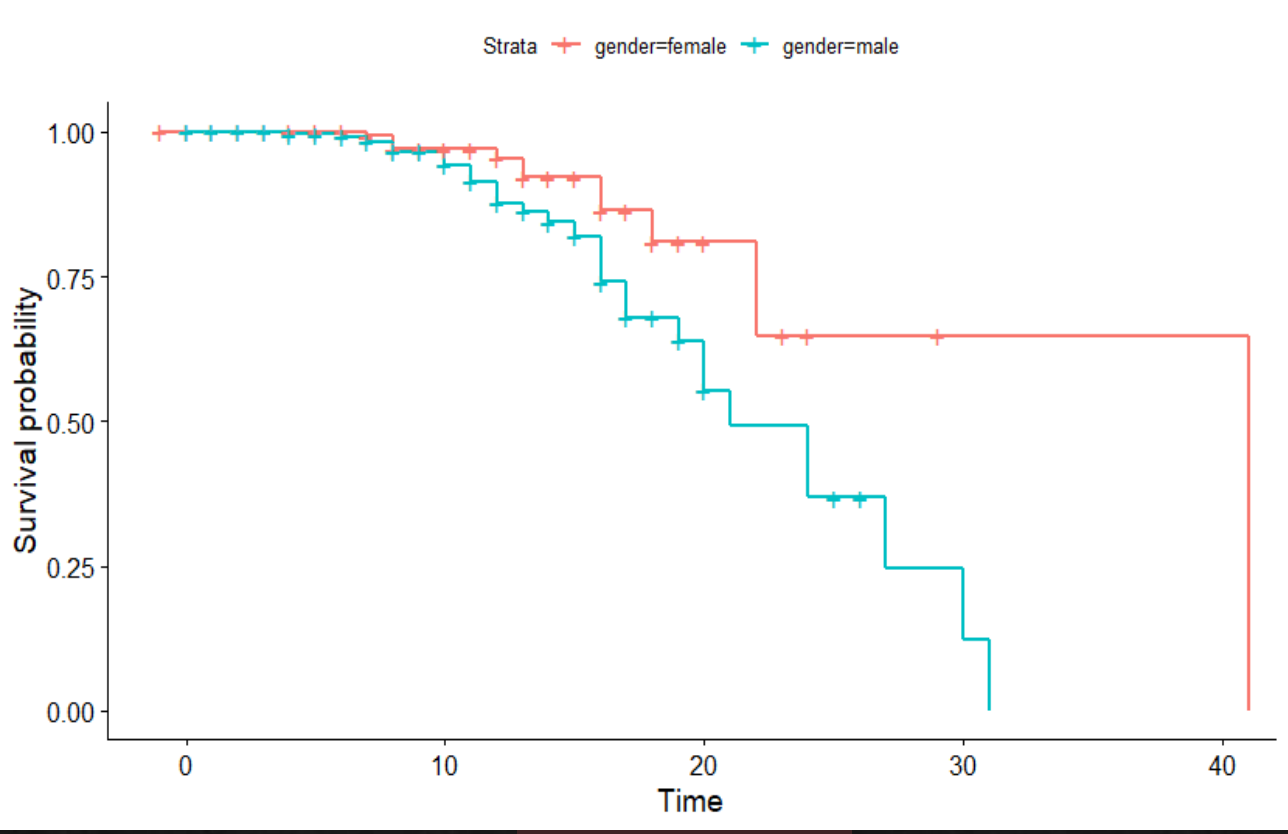
# Survival Analysis

# Survival Analysis Approach

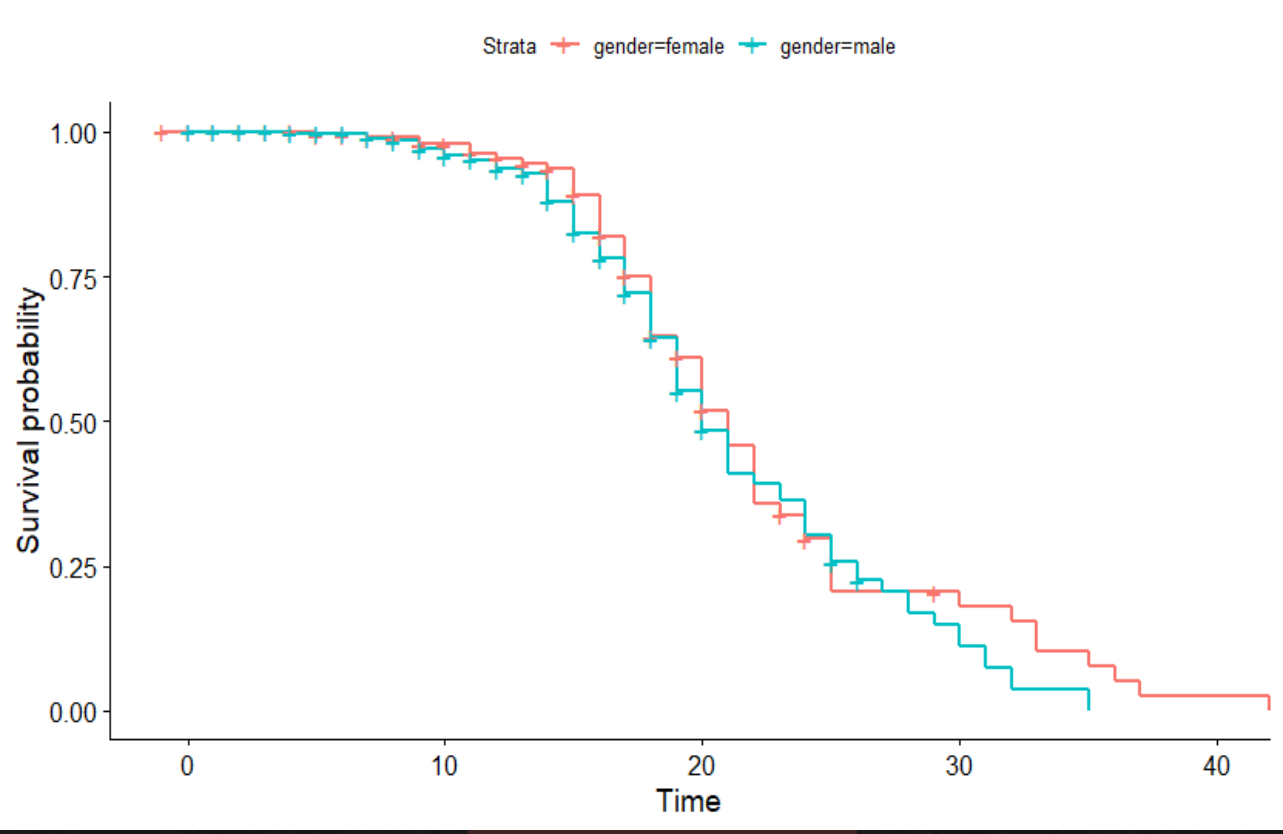
- Gender | Age Bin
  - **Trial 1 :** We first ran the analysis on segregated recovery and death data as we wanted to consider one event at a time.
  - **Trial 2:** We then considered one data set where we just went ahead and assumed the other event time to be infinity(we used the value 120 in this case).

# Data Pre-processing

- Survival/Recovery Time: Symptom onset date – recovery / death date
- Censored Data:
  - Trial1: Non Death Cases & Non Recovery Cases
  - Trial2: Included Death/Recovered (Infinite) Cases
  - Study End Date:
    - Reported Date
- Male / Female gender changed to 1 & 2
- Age bins grouped to 1 –5
- <20 – 1 ; 20-40 – 2; 40-60 – 3; 60-80 –4; 80+ -5



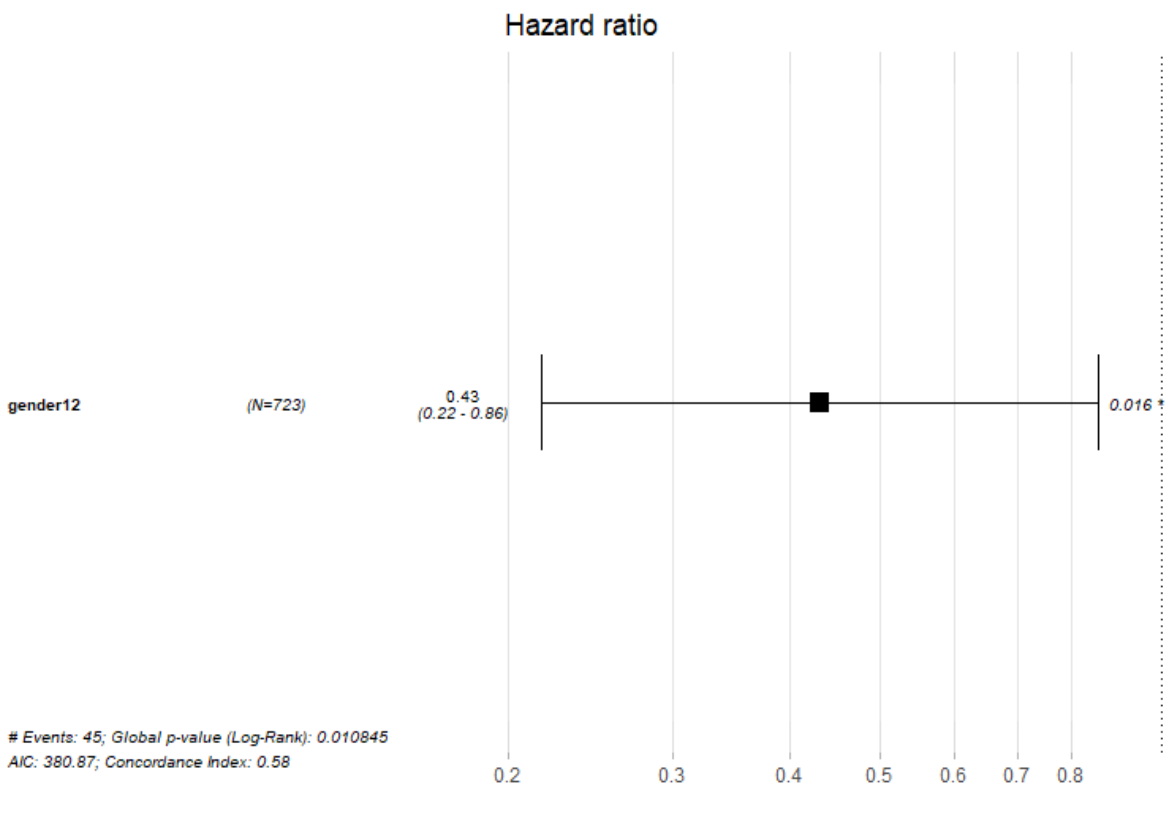

Event : Death



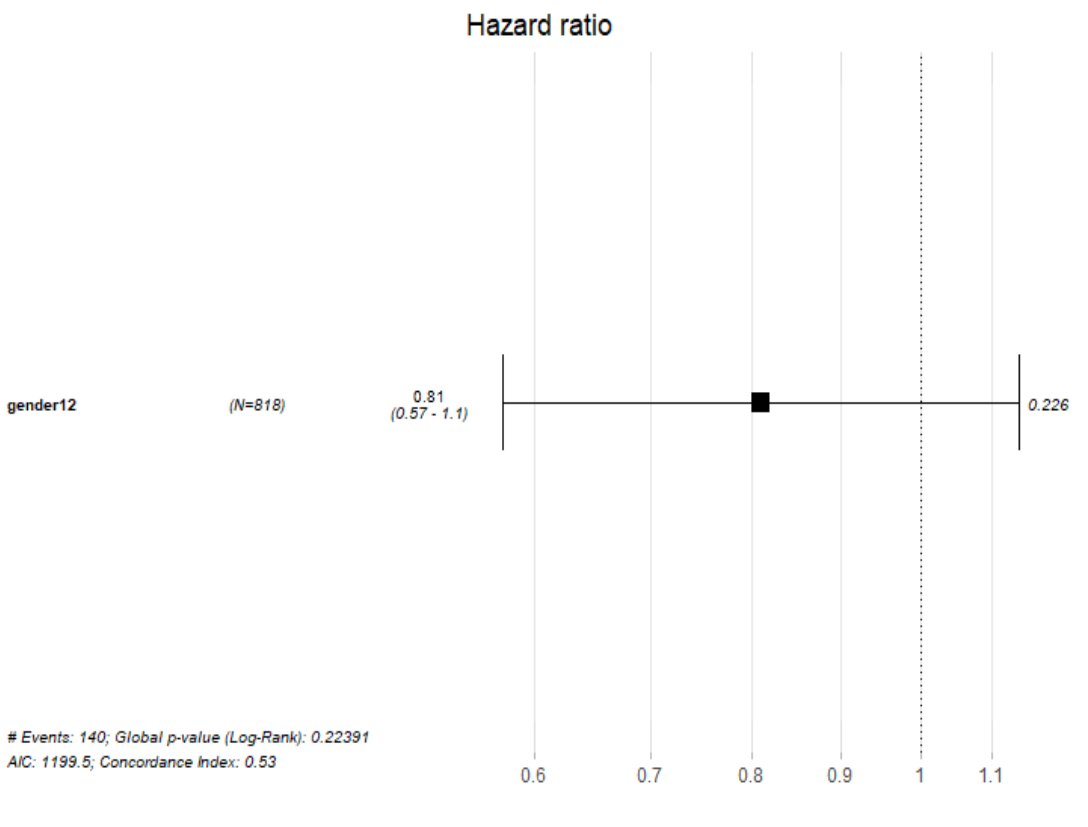

Event : Recovery

# Trial 1: Gender Grouping

- Considered censored data survival time/ recovery time as symptom onset date – reported date.
- Data set – Survival Status and Censored | Recovery Status and Censored




Event : Death

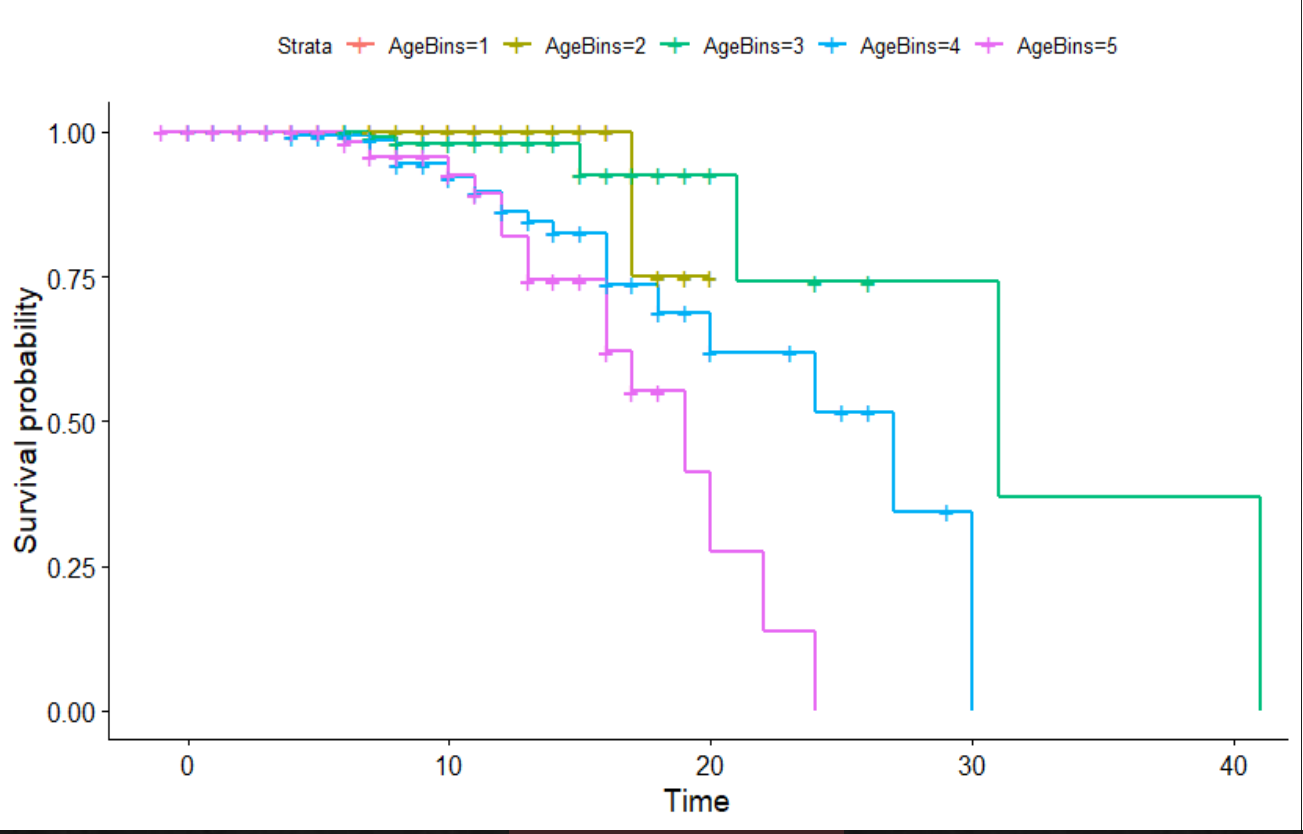



Event : Recovery

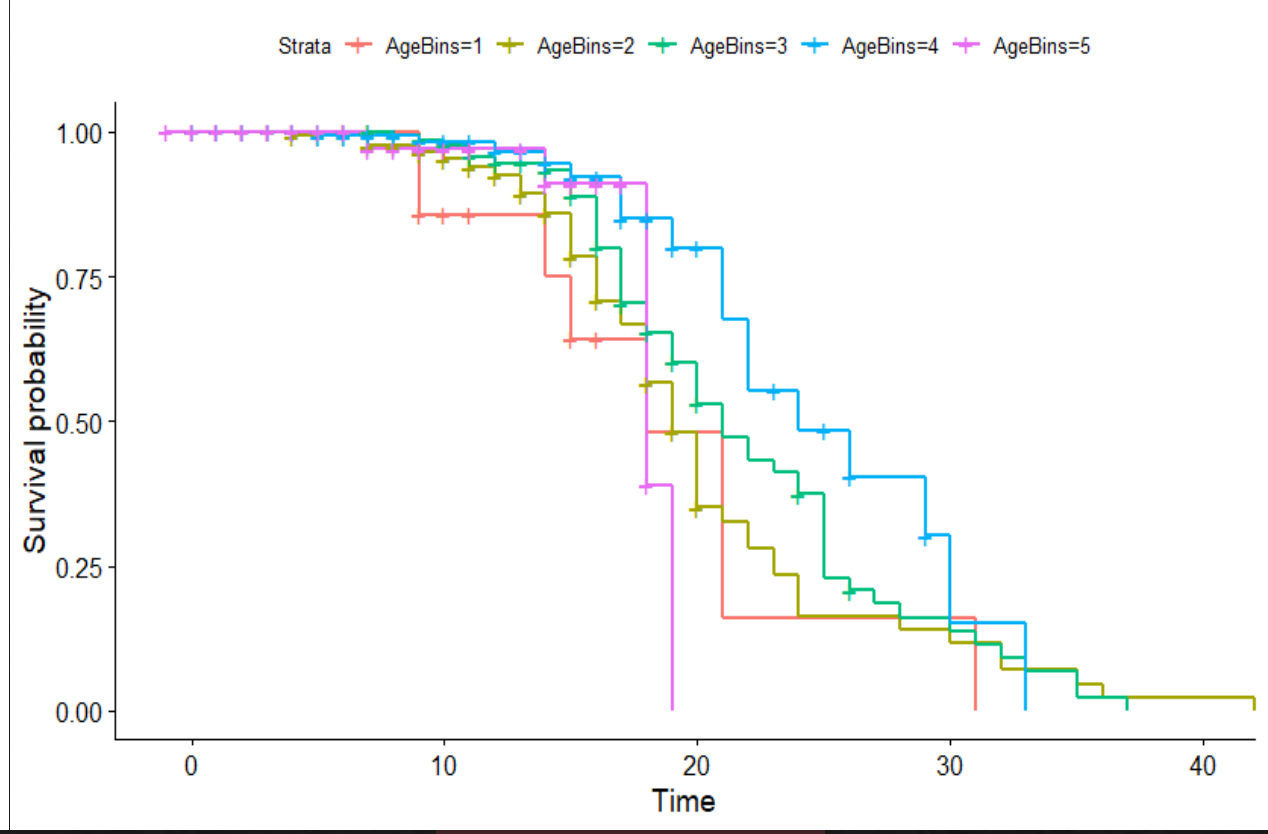
# Trial 1: Gender Grouping – Hazard Ratio

- Death Hazard Ratio = 0.43
- Hazard for Female < Hazard for Male

- Recovery Hazard Ratio = 0.81
- Recovery Rate for female and male is very close

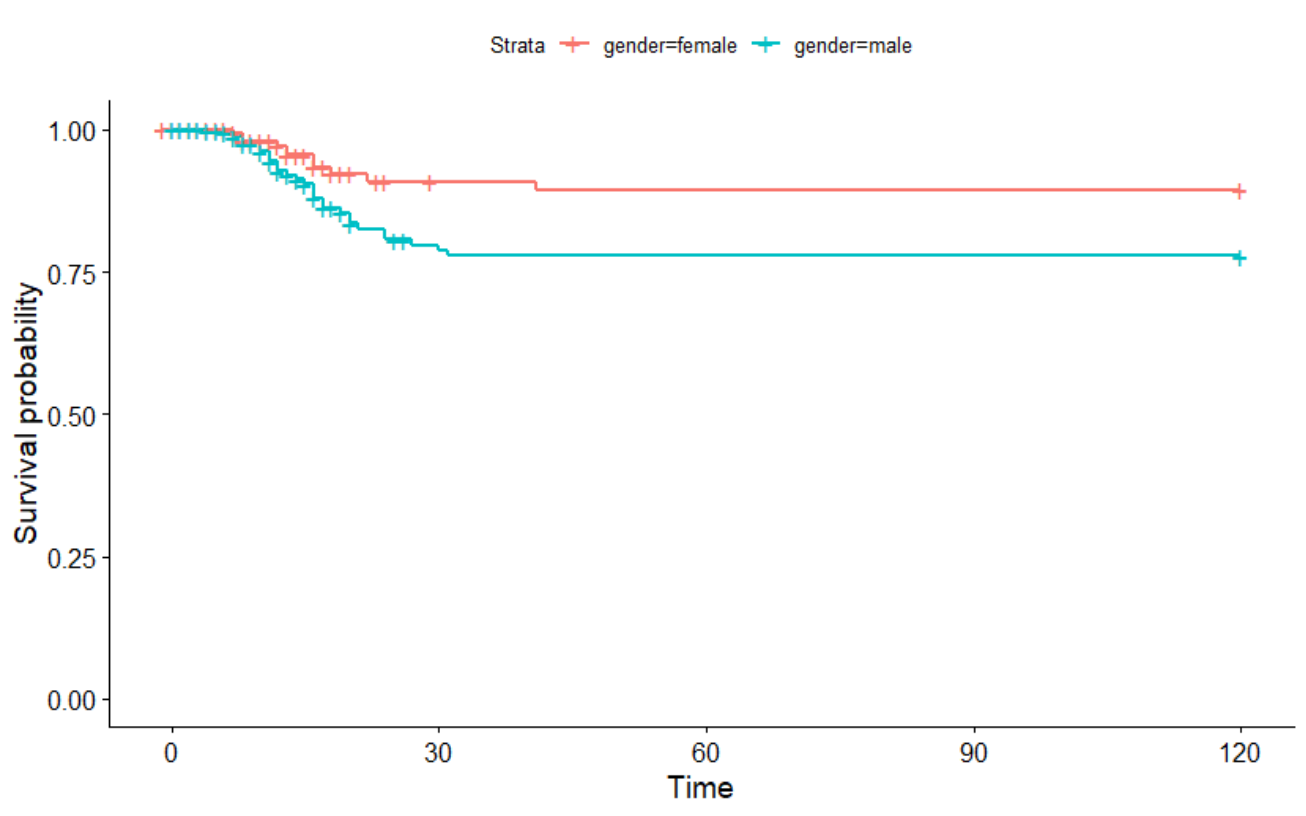



Event : Death

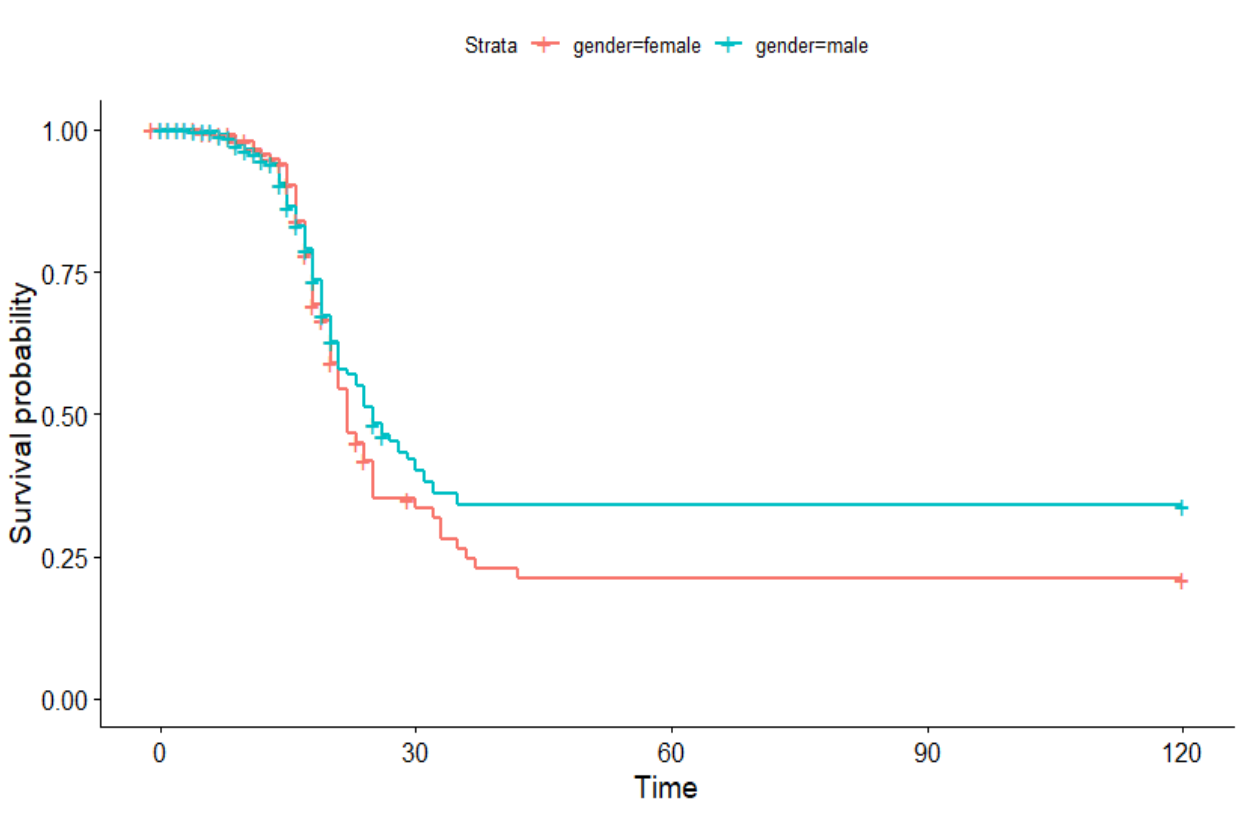



Event : Recovery

# Trial 1: Age-Bins Grouping



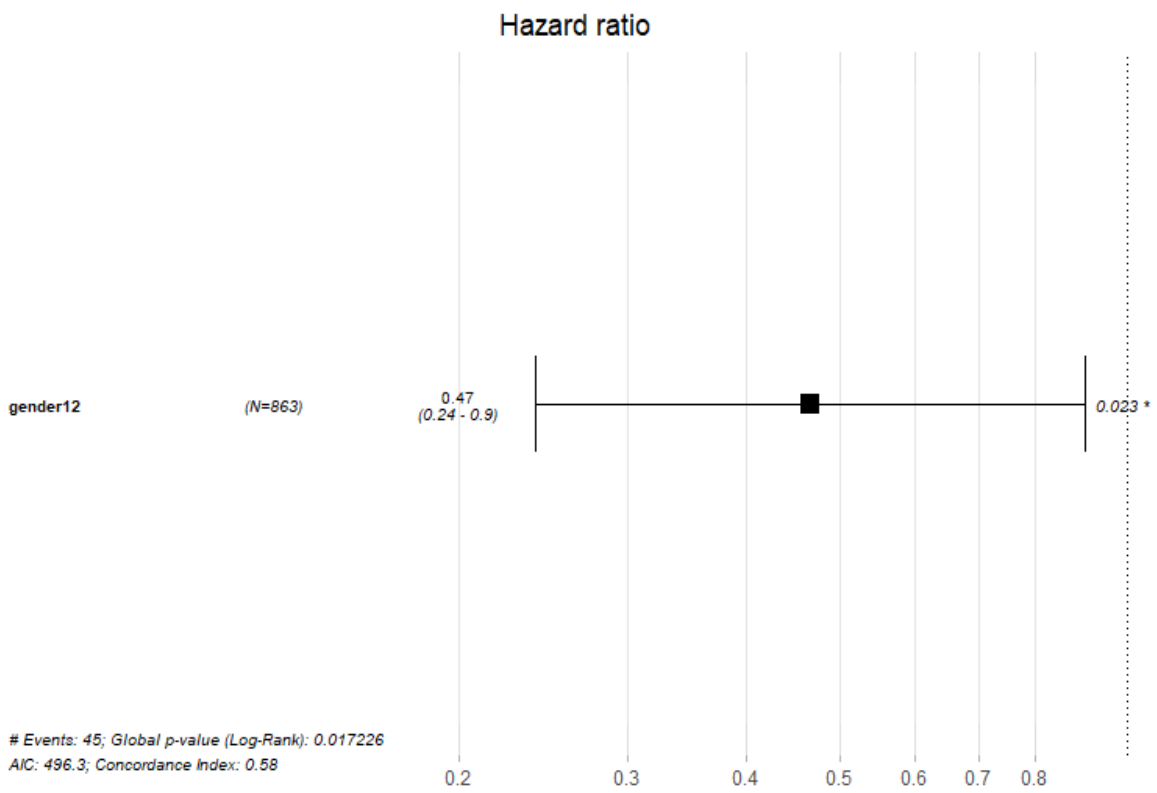

Event : Death



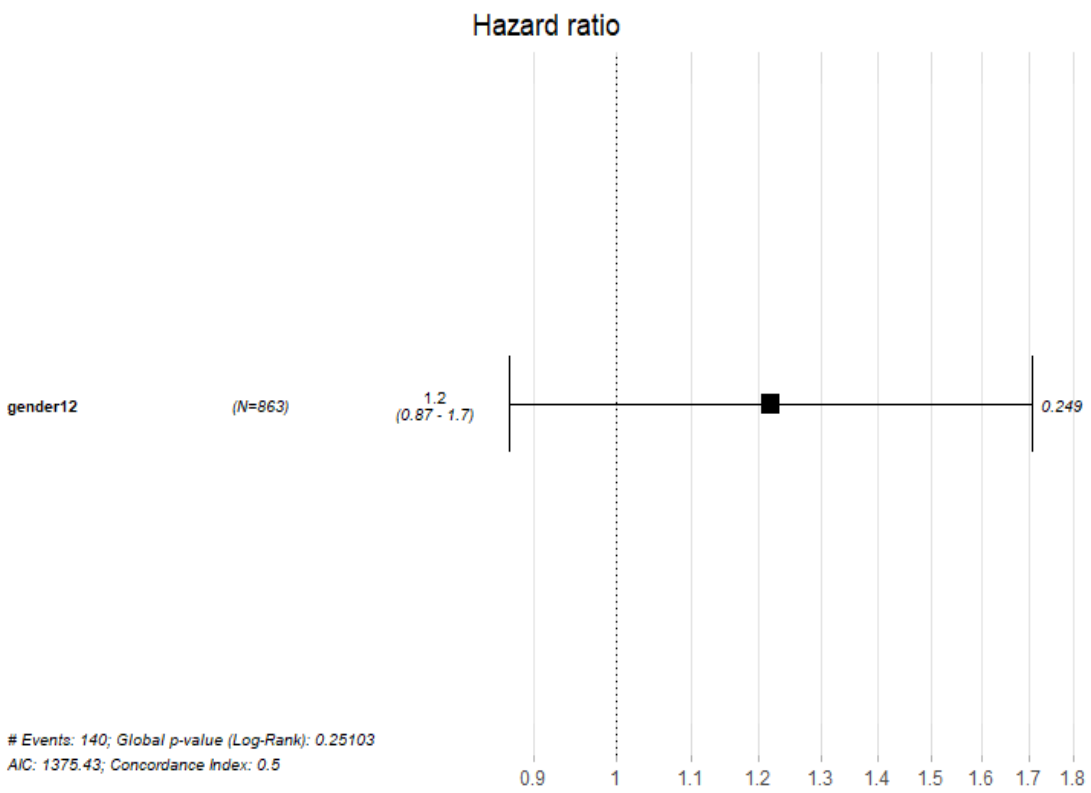

Event : Recovery

# Trial 2: Gender Grouping

- Considered censored data survival time/ recovery time as symptom onset date – reported date
- Data set – Survival ,Censored, Recovery | Non – event treated as infinity




Event : Death

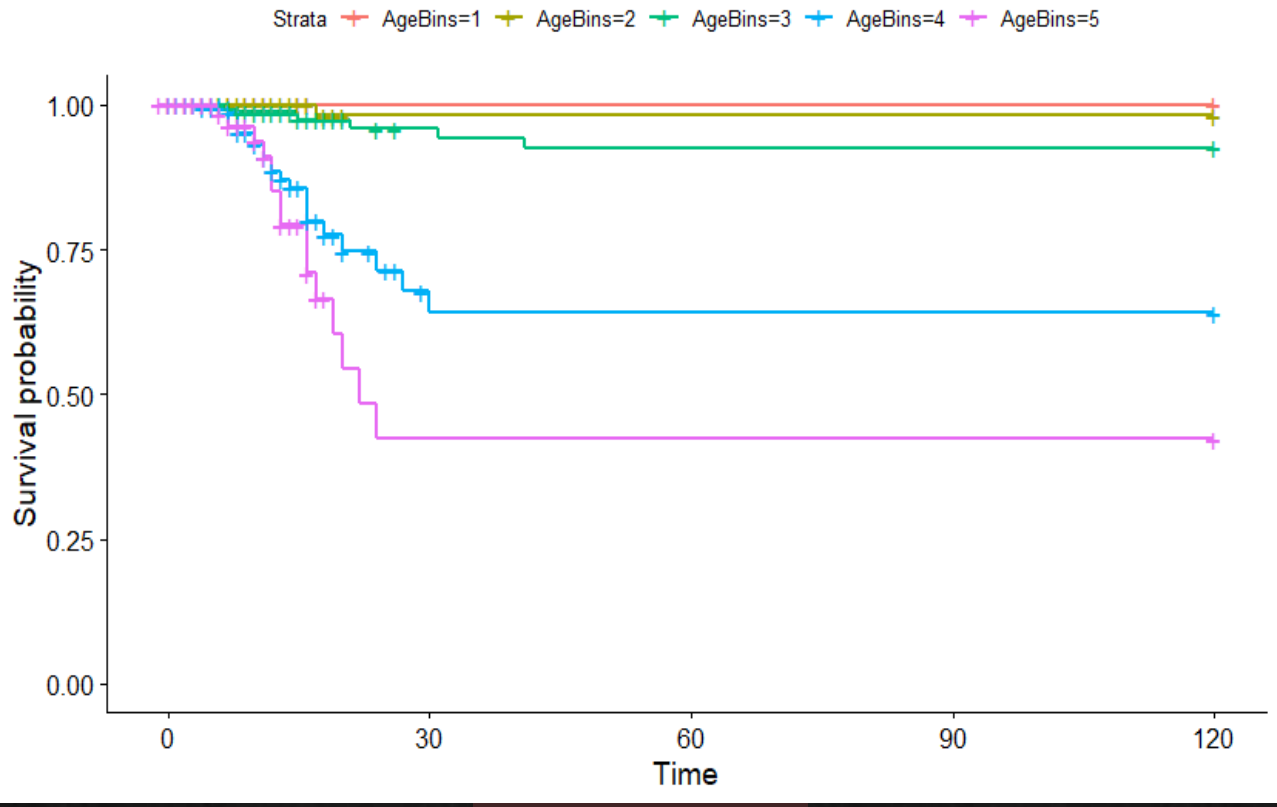



Event : Recovery

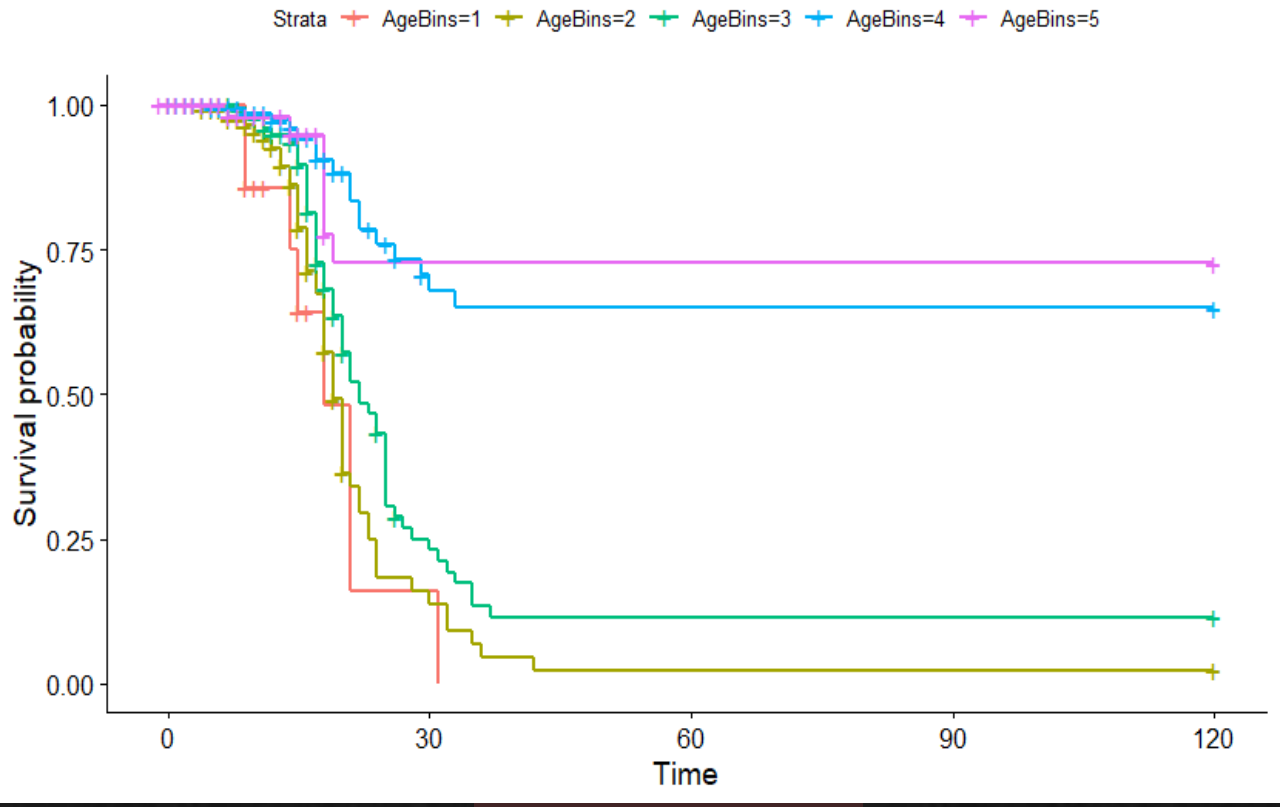
# Trial 2: Gender Grouping – Hazard Ratio

- Death Hazard Ratio = 0.47
- Hazard for Female < Hazard for Male

- Recovery Hazard Ratio = 1.2
- Recovery Rate for female and male is very close




Event : Death




Event : Recovery

# Trial 2: Age-Bins Grouping

# Conclusion

- Time Series Analysis – Forecasted the confirmed, death and survival numbers
- Countries & Age have a significant effect on the number of deaths
- Female have a higher Survival Probability than Male after contracting the disease
- There is no difference in the Recovery Rate for Male & Female
- Age group >60 are more prone to death after getting the disease
- Age group <40 recover the fastest after getting the disease

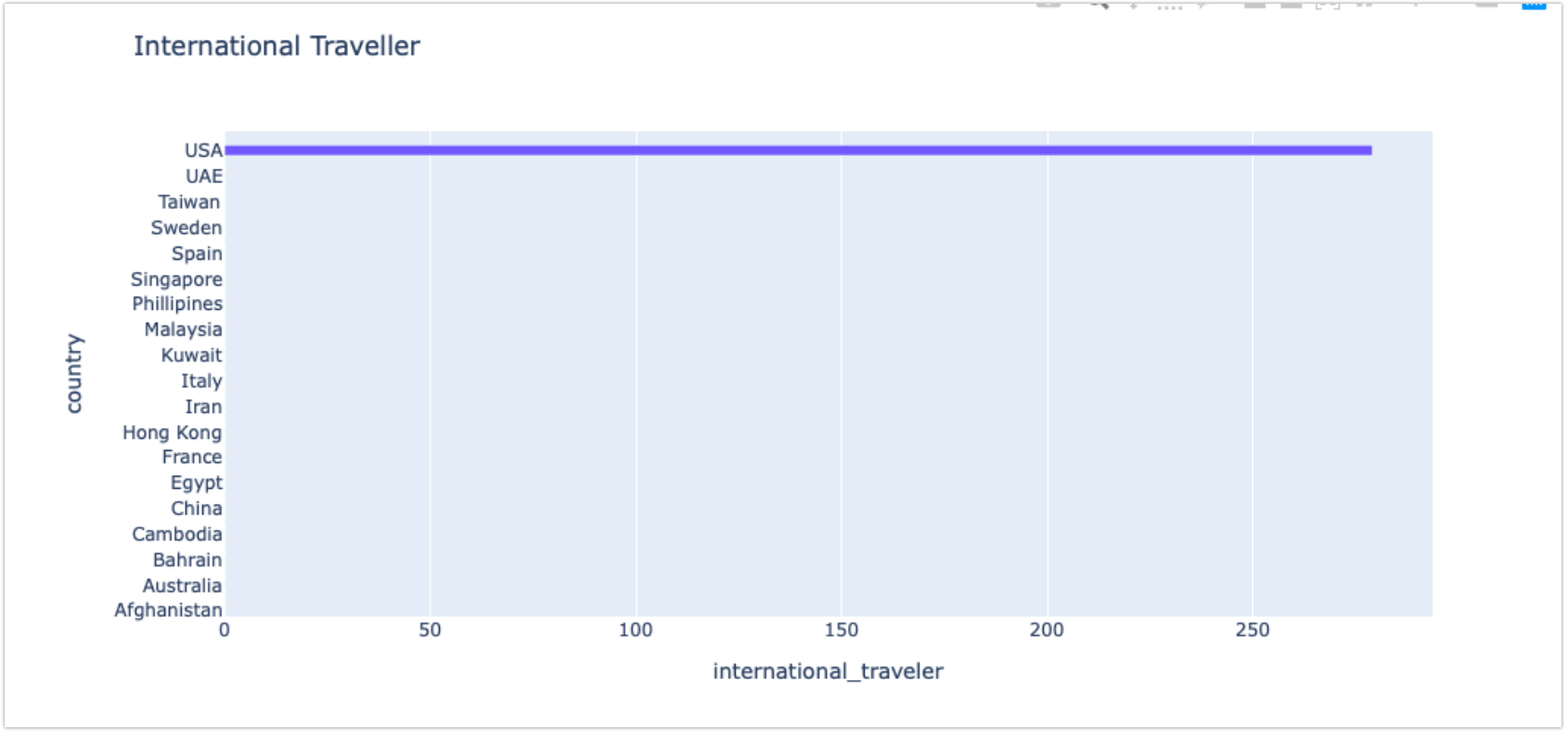
# Future Work

- Improving the time series predictive model
- Try to collect more data in the international traveller column to see the effect of international traveller on the number of confirmed cases
- Conduct random/fixed effect tests in granular level looking for location level effects

# Questions

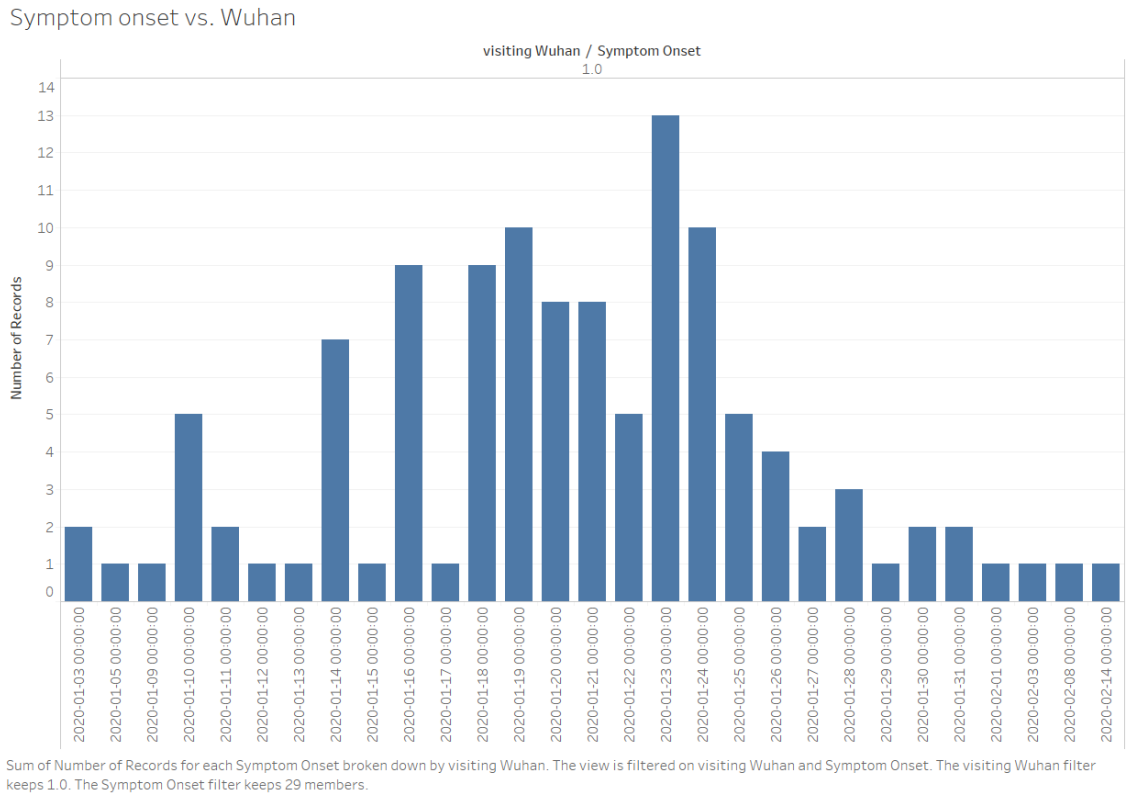
# Appendix

# International Traveller



# Insights Column: "visiting Wuhan"

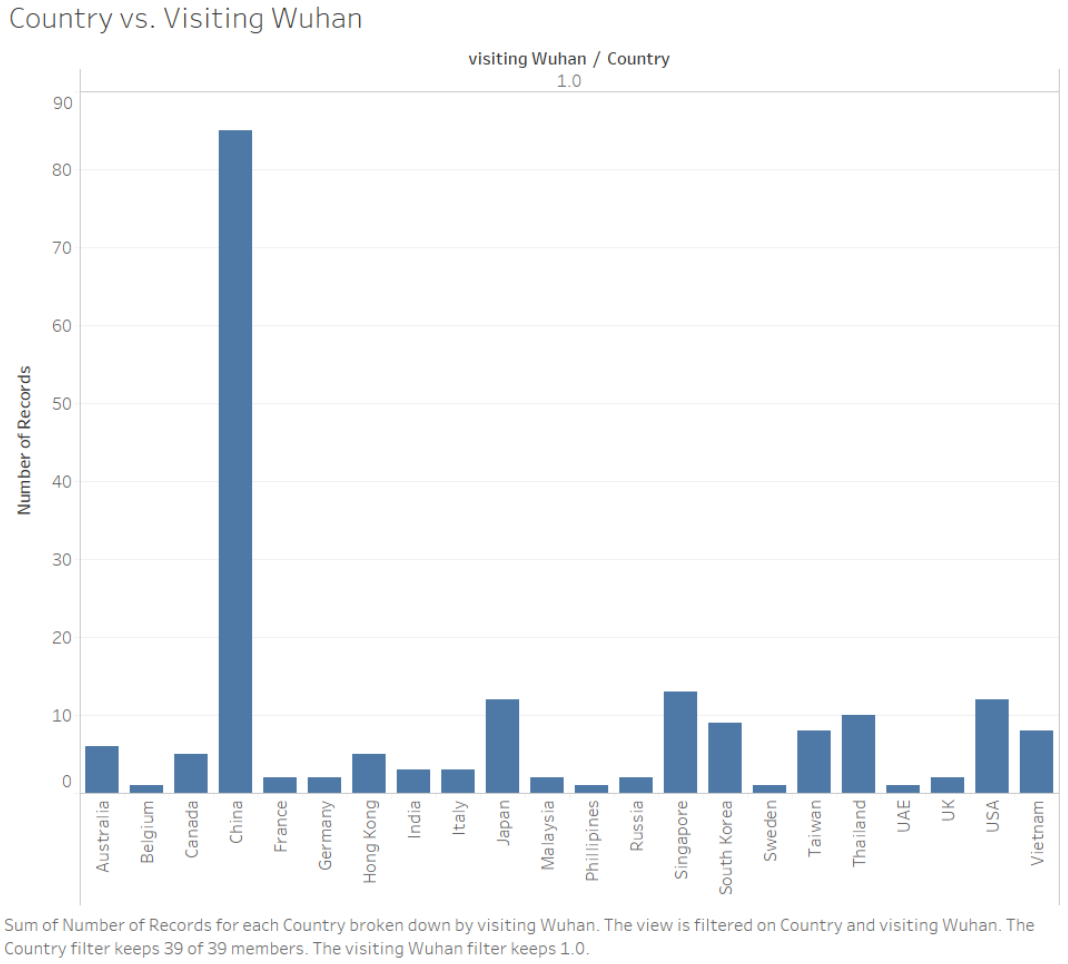
- When did the people that visited Wuhan start showing symptoms?




Sum of Number of Records for each Symptom Onset broken down by visiting Wuhan. The view is filtered on visiting Wuhan and Symptom Onset. The visiting Wuhan filter keeps 1.0. The Symptom Onset filter keeps 29 members.

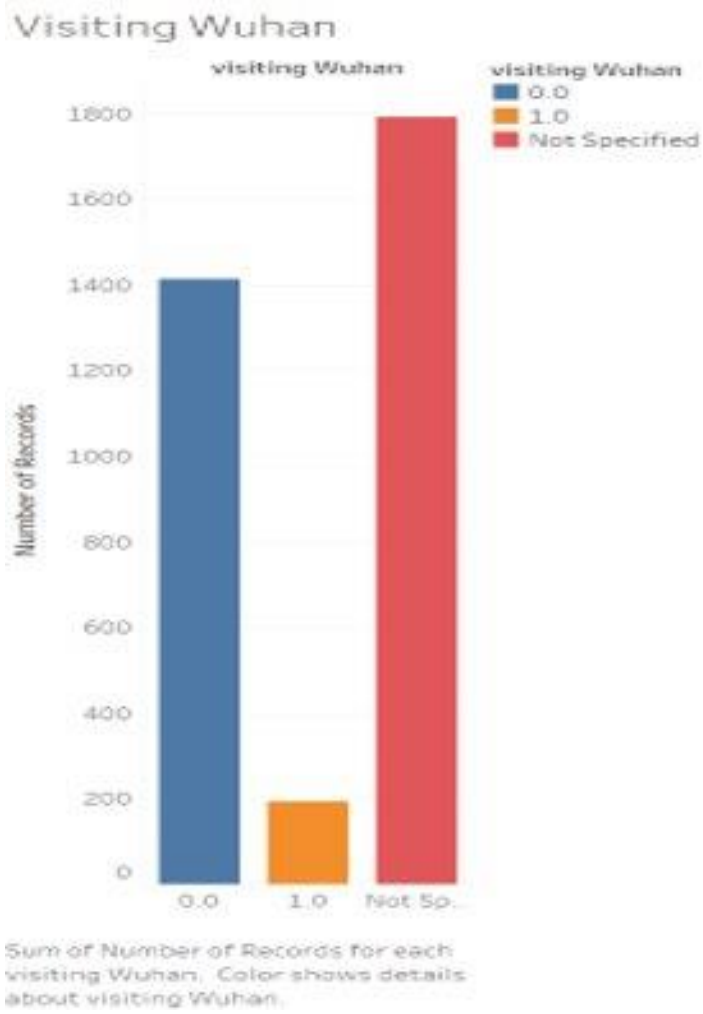
# Insights Column: "visiting Wuhan"

- How many confirmed cases in each country are  related to Wuhan



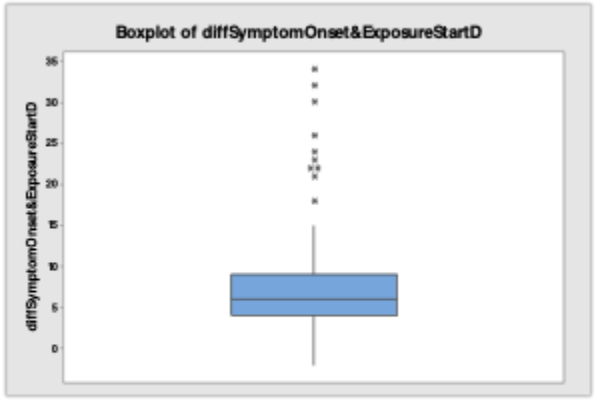
# Insights Column: "visiting Wuhan"

- How many people with confirmed cases visited Wuhan



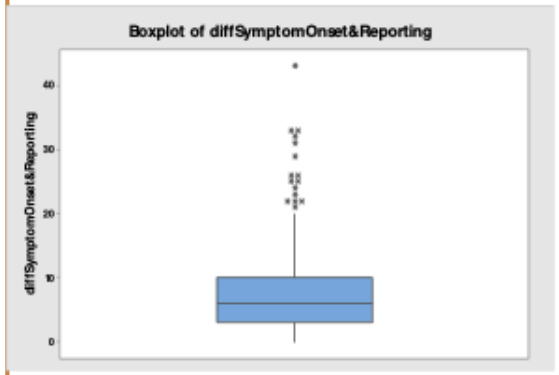
# Date Columns Analysis

**symptom_onset - exposure_start**



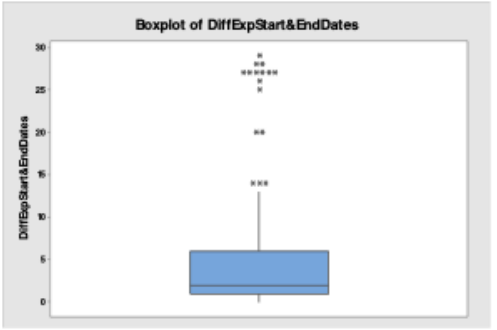

**Avg Time taken to show symptoms: 8-9 days**

**reporting date - symptom_onset**




**Time taken to report after symptoms started: 7-8 days**

**exposure_end - exposure_start**




**Avg Exposure Period: 4-5 days**

Correlation between death and Exposure Period = -0.039