Final Project: Predictive Analytics-MSA 8200

COVID-19 Analysis

Nimeelitha Akkiraju | Bennie Amani | Keerthi Bojja | Pranidhi Prabhat

April 29, 2020

Table of Contents

INTRODUCTION AND DATASET OVERVIEW

Within months, COVID-19 went from an epidemic to a pandemic. From the first identified case in December 2019 in China, the virus spread so fast and widely that it brought the entire world to a hold. Apart from its effect on the health of individuals, it has also scarred the economy to a great extent.

The datasource used in our analysis is accredited to : contact.sunky@gmail.com ; who has put together an excel sheet from various news sources citing COVID-19 cases in their country. He has made an extraordinary effort to put together data in different columns such as symptom onset date, exposure date, gender, recovered date or death date, etc. We considered this data set over other more cleaned data sets available online because this reflected the most heterogeneous data set and provided us an opportunity to deploy many time series learnings such as time series, random/fixed effect and survival analysis.

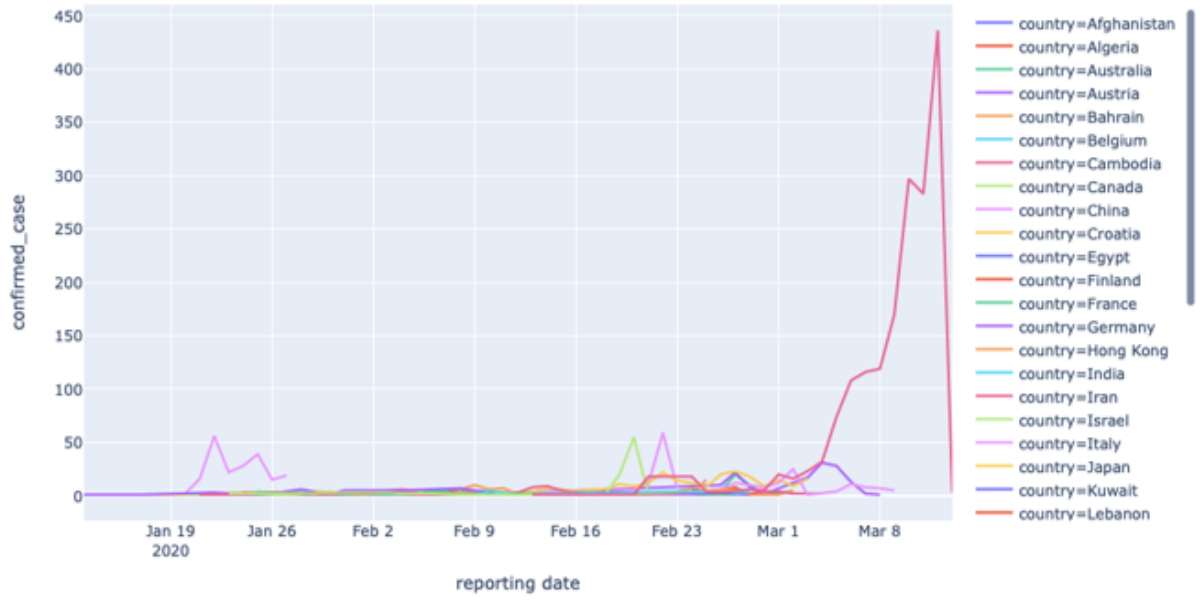
EXPLORATORY DATA ANALYSIS

We played around with various variables in our data and tried to find some insights from it.

We first saw that our confirmed cases data is spread across countries

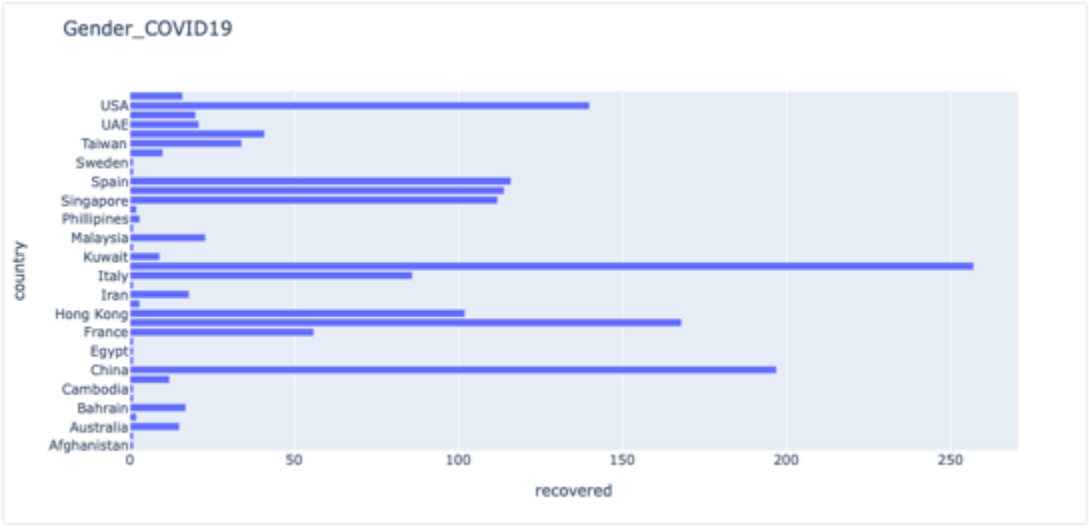
	country	confirmed_case
0	Afghanistan	1
1	Algeria	1
2	Australia	15
3	Austria	2
4	Bahrain	17
5	Belgium	1
6	Cambodia	1
7	Canada	12
8	China	197
9	Croatia	1
10	Egypt	1
11	Finland	1
12	France	56
13	Germany	168
14	Hong Kong	102
15	India	3
16	Iran	18
17	Israel	1
18	Italy	86
19	Japan	257
20	Kuwait	9
21	Lebanon	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 Analysis



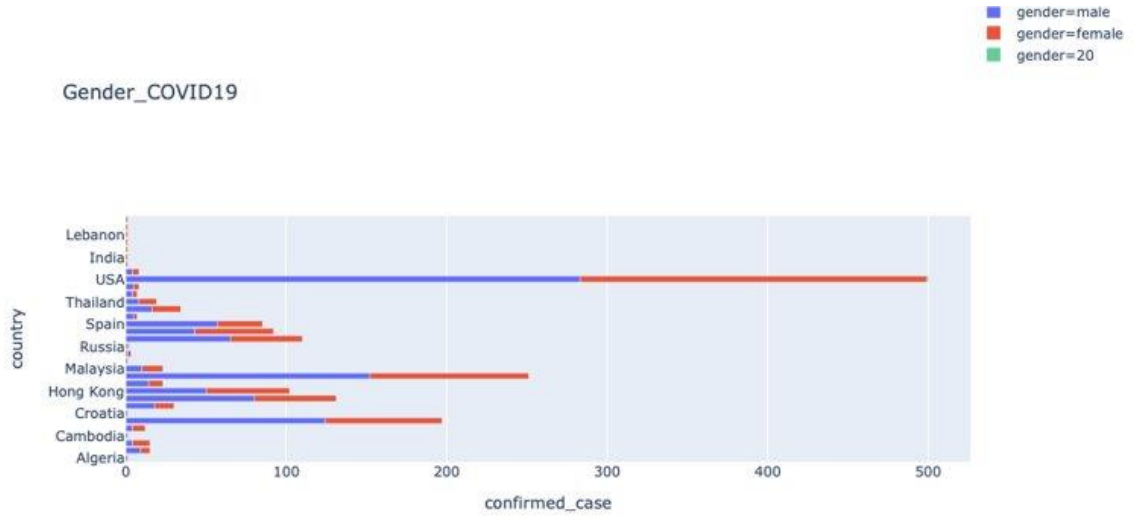
This graph shows the time series of the number of cases by country. With this kind of rich data country wise, we developed an idea of seeing the random and fixed effect of countries on the recovering of confirmed cases. We see a sharp decline in the number of cases towards the end because the data source that we have got was punching in the numbers manually and he stopped his research after that.

Country Wise Survival Numbers



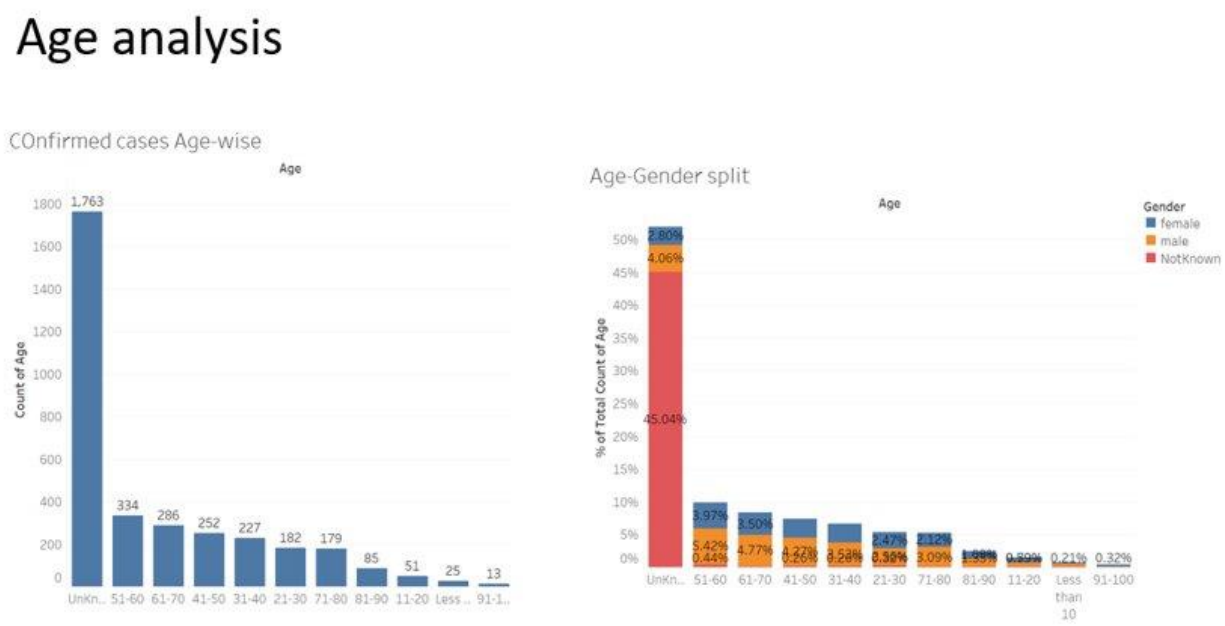
We then wanted to delve deep on the number of survival cases in each country. This graph gave us an idea of doing survival analysis on the data.

Gender Wise analysis



We then discovered the role of gender in the confirmed cases as well as the death and recovered cases.

Age-Wise analysis

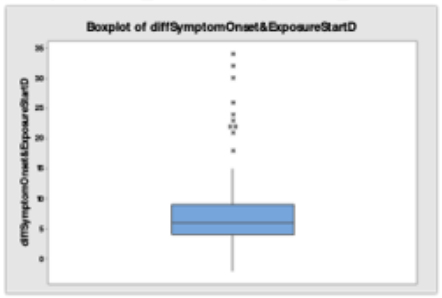


To run this analysis, we binned the age into age groups of 11-20, 21-30 etc as it made more sense to see what age groups are getting affected rather than individual age numbers. Firstly, the first bar (Unknown gender) in the graphs shows evidence of lack of sufficient data and presence of large null values. We can see from first graph that most affected age group is of 51-60 years followed by 61-70 years. The second graph shows distribution of gender in the given age groups.

Date Column Analysis

There is a general rule about quarantining oneself for 14 days if they are exposed to the virus. We wanted to check if this is true from the data available. Below box plot confirms this rule as the range of time taken to show symptoms is exactly 14 days and the average time is 8-9 days

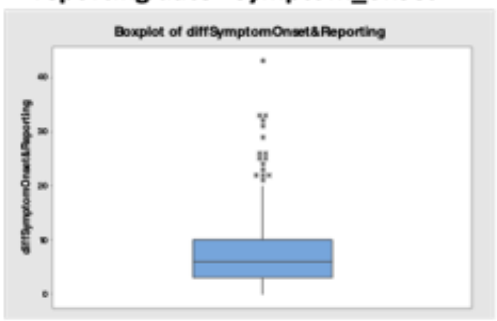
symptom_onset - exposure_start



Avg Time taken to show symptoms: 8-9 days

Also, we wanted to know the time taken for reporting the illness after the symptoms have started to see if early reporting would help stop spreading the virus. The average time taken is 6-7 days, so people should report the symptoms as soon as possible and quarantine themselves for at least 14 days.

reporting date - symptom_onset



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

Calculated the difference between Exposure Start Date & End Date to identify if there is any relation between the exposure period and death of the patient. We see that there is a correlation between death & exposure period

exposure_end - exposure_start

Avg Exposure Period: 4-5 days

```
Correlation between death and
Exposure Period = -0.039
```

DATA CLEANING & PRE-PROCESSING

Though this data was spread across different columns, there were issues with missing data and hence we had to run different data processing for different analysis.

TIME SERIES ANALYSIS

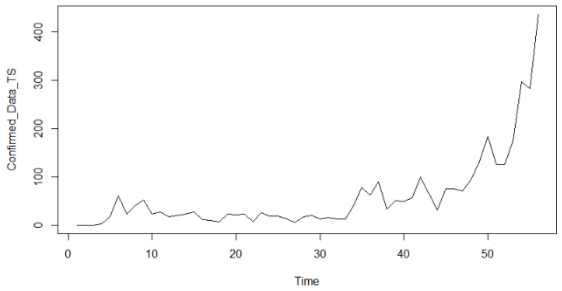
With the nature of data that we have, we initially wanted to do a time series forecasting to predict the following cases:

1. Number of confirmed cases
2. Number of deaths
3. Number of recovered cases

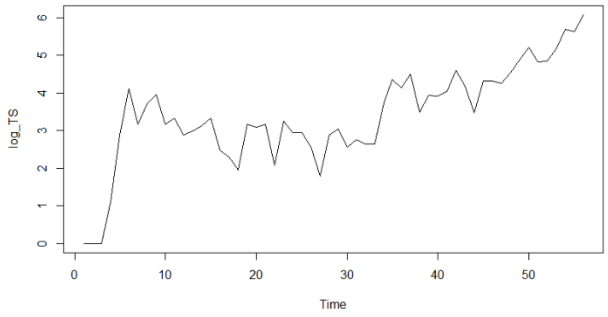
So, we transformed the data into time series object for each of the above points and plotted them to analyze the data. Further details of time series modeling are explained below

Predicting Number of Confirmed Cases:

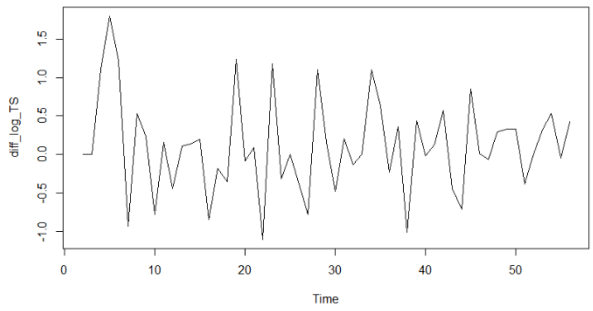
We plotted the initial time series and we can see from below plot that the number of cases have been increasing exponentially after 30 days of the start of this study



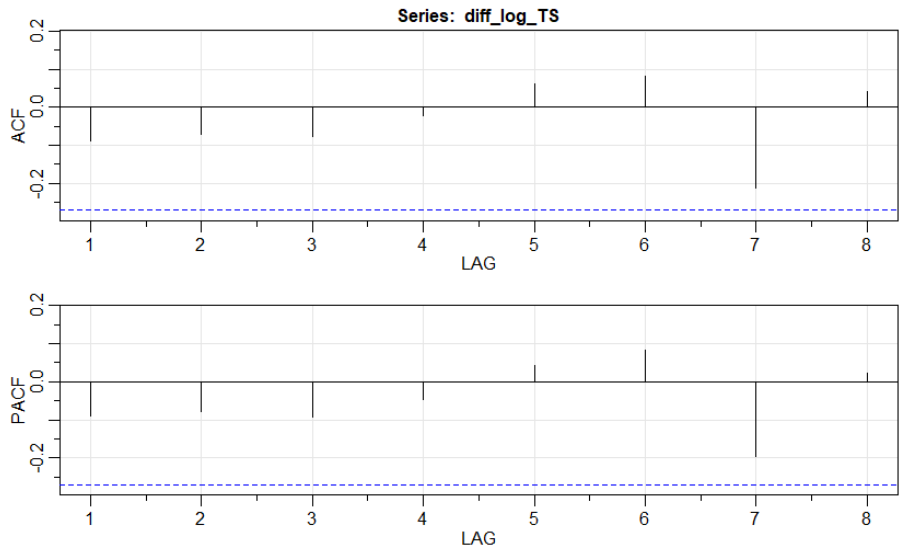
We can also see variance present in the data as from above plot and applied log to remove the variance. Below is the series after applying log.



We can clearly see an upward trend in the above plot and hence applied differencing to remove trend from the data

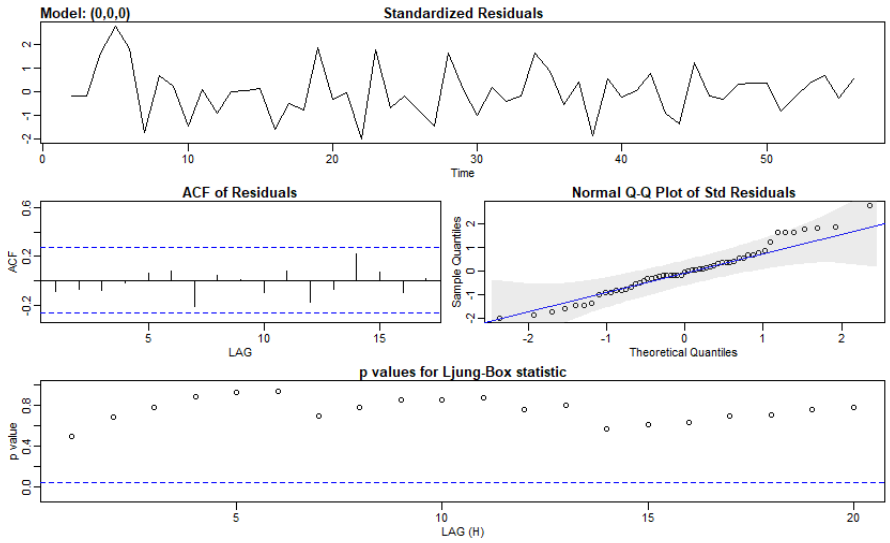


Next step was to plot the ACF/PACF Plots to identify if there is any correlation between the time lags and visually check the type of model to apply for the prediction



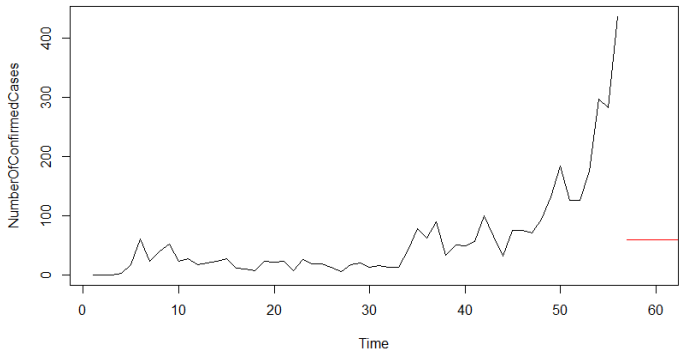
From above result, ARIMA(0,0,0) looks like the best fit since it has the lowest AIC value = 2.24

We fit an ARIMA(0,0,0) model and below are the results of it. We can see from residual plot that no additional information is retained and from Ljung-Box statistic, we can also see that all the p-values are above the blue line and residuals follow normal distribution.



Prediction for the next 1 week: After finalizing on (0,0,0) model, we ran model on the original time series data.

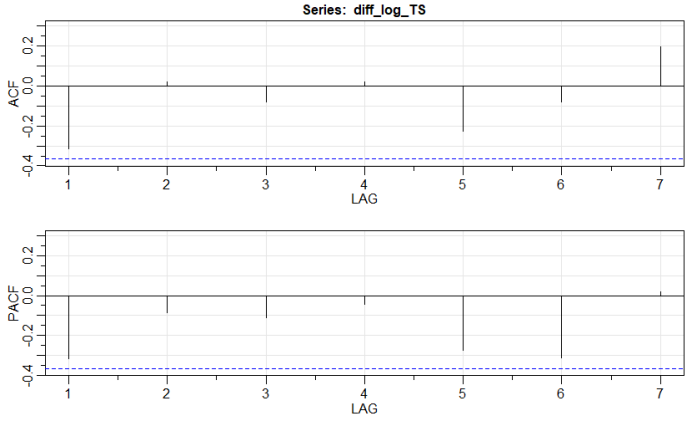
predict <- predict(arima(Confirmed_Data_TS, order = c(0,0,0)), n.ahead = 7)



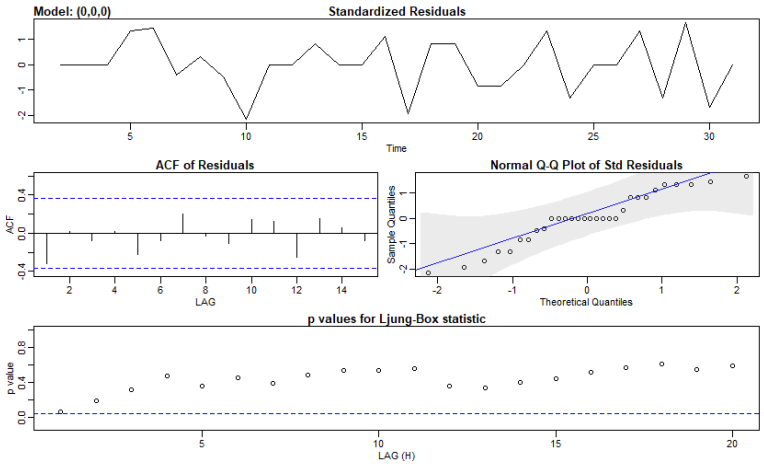
In the above plot, we can see that the red line which is a prediction of next one week from 57-63 days is tending to average too early because past data is available only for less than 60 days

Predicting Number of Deaths

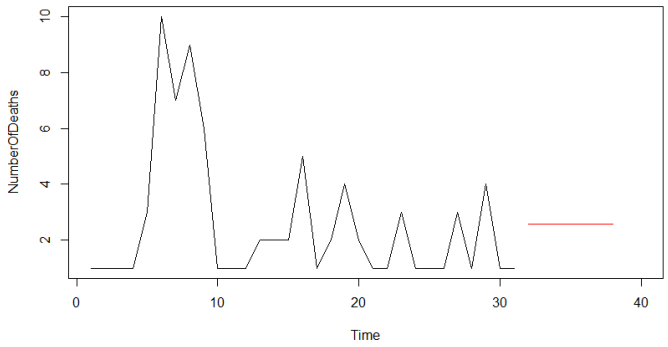
To predict number of deaths, we took time series data of death reported date. Like in case of confirmed cases, we plotted the time series, sew that there is variance and trend in the data, so we applied log & difference and made the data stationary and plotted ACF & PACF plots to identify the model



The finalized model from above ACF-PACF plots is ARIMA(0,0,0). Below are results of the model.

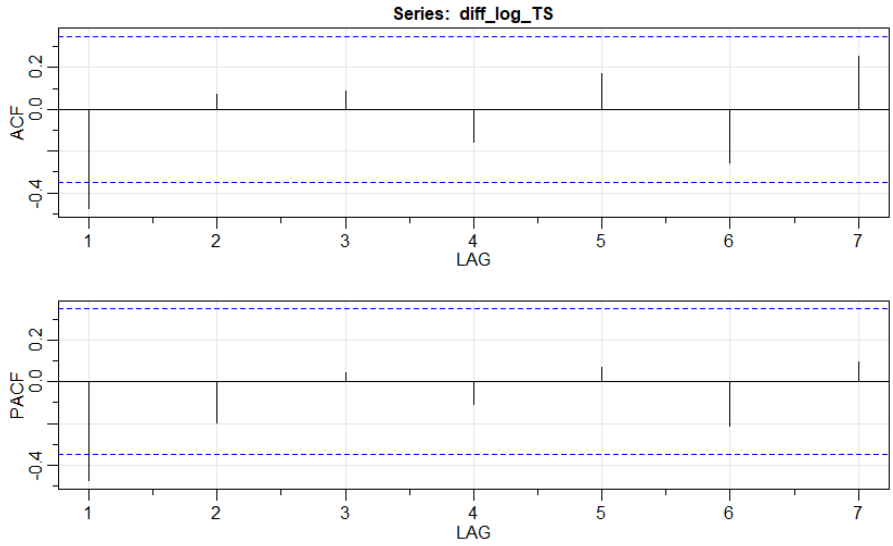


Prediction of number of deaths for next 1 week on original time series data:

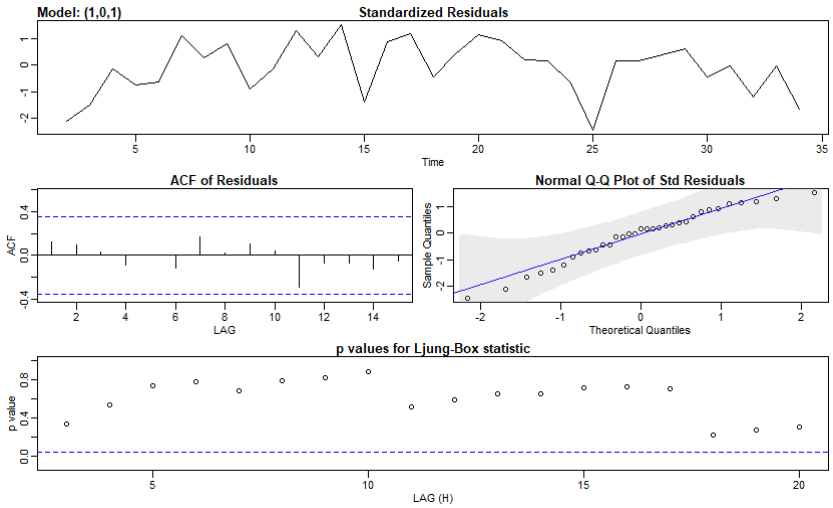


Predicting Number of Recovered Cases:

To predict number of recovered cases, we took time series data of recovered reported date. Like in case of confirmed cases, we plotted the time series, sew that there is variance and trend in the data, so we applied log & difference and made the data stationary and plotted ACF & PACF plots to identify the model

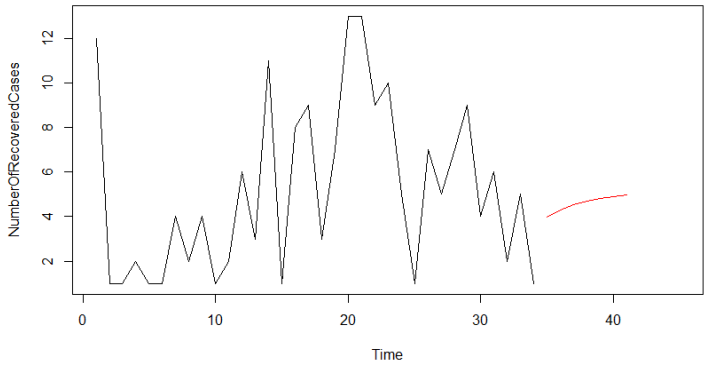


As there is 1 spike sticking out of the blue line in both ACF and PACF plots, we finalized model to be ARIMA(1,0,1)



Here as well, we see no information retained in residual plots and all the p-values are above the blue line and the residuals are following normal distribution.

Prediction for recovered cases for next 1 week on original time series data:



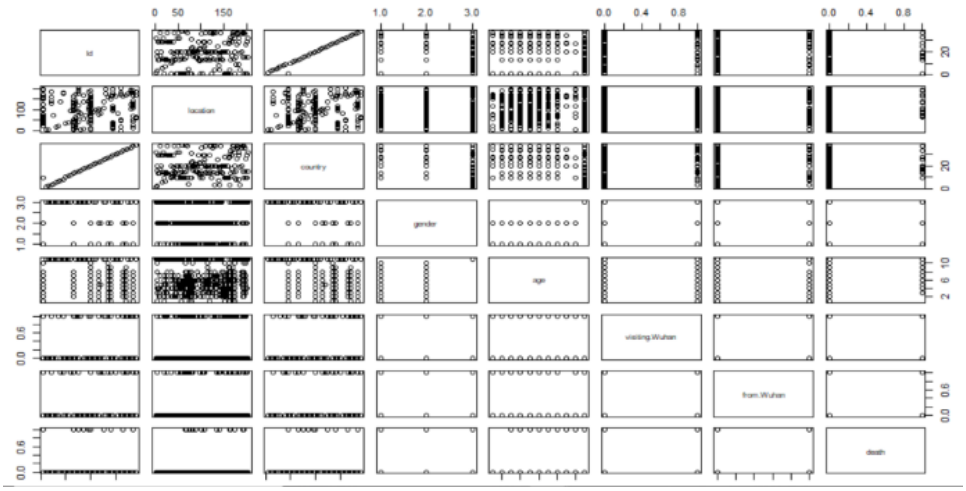
FIXED EFFECT & RANDOM EFFECT

This section explains the process of measuring the random and fixed effect of some x variables on the y variable, which is death.

Data

Initial task was to select the data to run our fixed/random effect models. We wanted to see if there are any subjects that can have fixed/random effects on y variable, death. Looking at the data, we realized we have few columns like country and age that can be factored and see if they have any such effects. Hence, we decided to look at country level effect and age level effect on death. But majority of data have null values and therefore we filtered subset of data that has death information.

The x variables chosen for this analysis were selected based on the fact that they are not always the same within a cluster and are assumed to affect the y variable in some way. They are age, gender, visiting.wuhan (whether the person went to Wuhan), from.wuhan (whether the person is coming from Wuhan) and country. Below is the set of data that we used.



Country level effect:

We first ran a linear regression model by factoring the country and looked at significance of results

```
> 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
```

As highlighted above, few countries like China, Singapore, Japan show significance in terms of p-value. This means that even after controlling for all other factors, for example in China, we can say probability of death is 0.13 times more. Therefore, we can confirm country level effect is present in predicting death and we must run a fixed effect model to include these effects in prediction.

To further confirm, we decided to run random and fixed effects model to see their results. Our initial approach was to construct the Omega matrix and find out the effects in manual method. We were stuck here for quite some time because we were in impression that we should have equal samples from each country in order to have panel data. But after further research, we found lmer(Linear mixed effects regression) R package that made things little easy. The lmer model from the lmerTest package is a mixed effects model, which means that it includes both fixed effects and random effects.

Below are results from running lmer model:

```
> 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 ***
```

From random effects result, we see the variance of country is not exactly equal to 0. When we run rand(model), we can also see results confirming the significance of fixed effects

```
> 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
```

Null hypothesis: Variance of country is 0

We can see since p<alpha, we reject the null hypothesis and confirm the fixed effects of country on death.

Age and Country level effect:

We also wanted to check if age has fixed or random effects on death. So, we followed same method and ran a linear model by factoring age:

```
> 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
```

We can see that for age as well, the variance is not equal to 0. Let us look rand() results of this model

```
> 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
```

Null hypothesis: Variance of age is 0

We can see since p<alpha, we reject the null hypothesis and confirm the fixed effects of age on death.

From running these two models, we can see that there is country and age level effects on death.

SURVIVAL ANALYSIS

With the given data set, there are two key events – 1) Death 2) Recovery. We decided to incorporate Survival Analysis technique to study the amount of time it takes before a death/recovery occurs for subjects with COVID-19 disease. This analysis has a unique advantage of including the cases that have experienced the event and the cases that have not. It also provides the base to add strata of different features on the event and see if one groups performs better than the other. For the case in point, we decided to incorporate survival rate and the recovery rate of male vs female and age bins

Here is our key analysis overview in brief:

1) Who has a better survival rate – male/female (Event: Death)

2) Who has a better recovery rate – male female (Event: Recovery)

3) Who has a better survival rate –age brackets 0-20 /20-40 /40-60/ 60-80/ 80+ (Event: Death)

4) Who has a better recovery rate – age brackets 0-20 /20-40 /40-60/ 60-80/ 80+ (Event: Recovery)

Trial 1 : We first ran the analysis on segregated recovery and death data along with censored 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

Following was the data pre-processing for Survival Analysis :

1) Extracted Age details from Summary text Column and filled missing data in Age column
2) Fixed the date of death, Null or blank imputed with 0 or with date mentioned in the summary
3) Fixed the date of recovery column, Null or blank imputed with 0 or with date mentioned in the summary
4) Put formula – Recovery_ Time & Survival_Time if 0 – as of date – symptom onset else recovered date – symptom onset
5) Put Recovery_Status & Survival_Status value based on event happen (2) or censor(1)
6) Drop records that say 2/30/1899 as recovery dates – 12 rows
7) For recovery as 1 – found the date of recovery by reading the summary – Last 4 data rows do not have the date in summary – dropped the two rows
8) Imputed the blank symptom onset date with hospital visited date
9) For remaining symptom onset date with NA/0 equated to Exposure end date for 228 rows
10) Deleted the remaining 2405 rows with blank symptom onset date
11) For Death 0 but recovered 1 manually change the survival time to 120
12) For Death 1 but recovered 0 manually change the survival time to 120
13) Dropped records of Unknown Gender
14) Add column name gender12 to store male as 1 & female as 2
15) Create a column agebin that stores 1 for age <20; 2 for age (20 –40), 3 for age (40-60), 4 for age (60-80), 5(80+)

Trial 1 :

Death Recovery Censored – independently - without infinity

Survival Analysis (Event:Death)

```
> Survival_Data[160:170]
 [1]  1+  6+  1+  5+  4+ 10+ 12+ 12+  2+  3+ 22
```

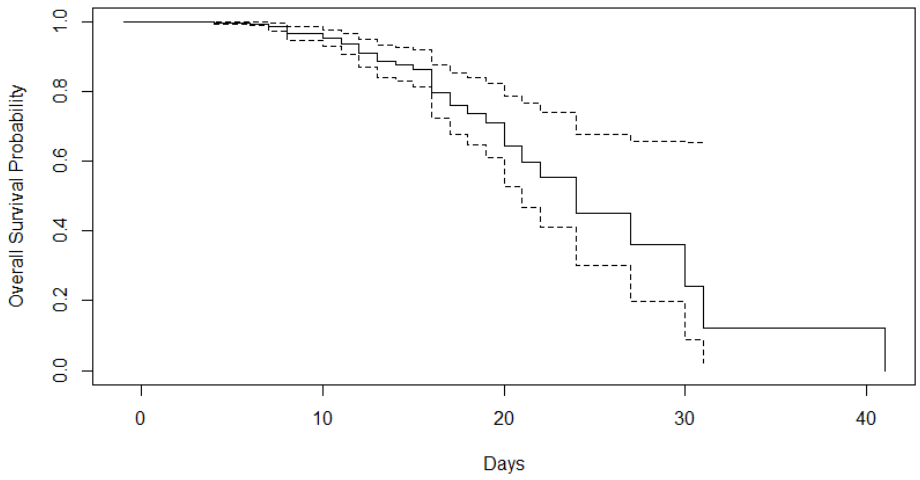
In the output above:

22 : indicates that the subject has died in 22 days after he/she got the disease

3+ : indicates that the subject is censored i.e the survival time for this person is at least 3 days

Kaplan Mier Survival Plot for 1 group: Shows the chance of survival of all the subjects after a certain time period :

plot(survfit(Survival_Data ~ 1),xlab = "Days",ylab = "Overall Survival Probability")



Observations in the plot:

1. All the observations are alive for the first 5 days from the beginning of symptom onset date, so here the survival probability is 100%
2. We can also see that the chances of survival after 40 days is close to zero.
3. Dashed lines are the confident interval which is symmetric around the Kaplan Mier estimated line

Data from the output parameters of Kaplan Mier Estimator:

	time	n.event	n.censor	n.risk
1	-1	0	1	723
2	0	0	18	722
3	1	0	52	704
4	2	0	73	652
5	3	0	65	579
6	4	1	49	514
7	5	0	62	464
8	6	1	56	402
9	7	3	36	345
10	8	6	40	306

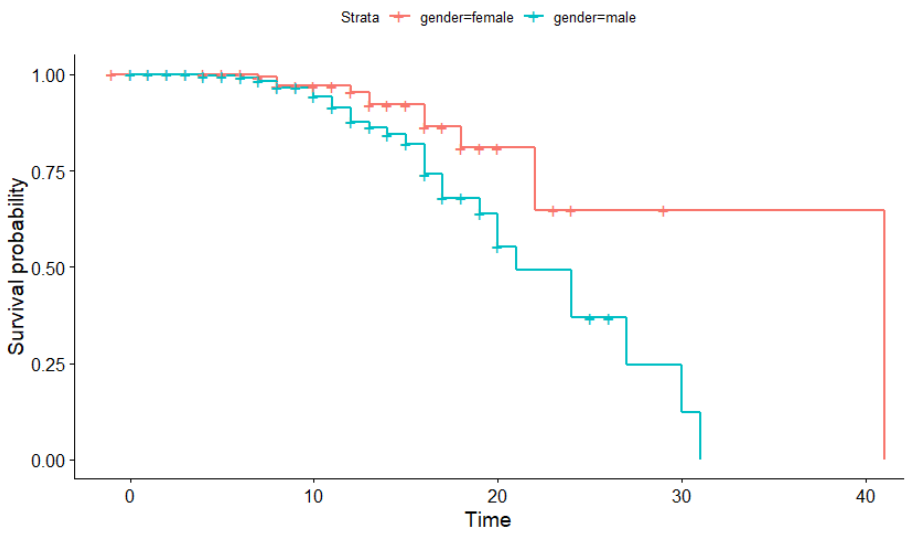
Interpretation:

Row 1: There are 514 subjects in the study as of day 6 and 1 event(death) happened on that day

Row 16: There are 306 subjects in the study as of 10 days, 6 event(death) & 40 censored (out of study) cases happened on that day

Kaplan Mier Survival Plot for 2 groups: Shows the chance of survival of 2 groups (Male & Female) of subjects after a certain time period. This is a good visual check to see the significance between the two groups

ggsurvplot(survfit(Survival_Data ~ df$gender),data = df)



No matter what the cutoff for the time is, Survival Probability for female is always greater than the survival probability for male. I.e. Death Rate for male is higher. Also, there is slight overlap in the beginning so we can say that both group have equal Survival Probability in the early days of the disease being attacked.

We can also do a log rank test to do a statistical check to see if there is any systematic pattern btw gender & survival probability:

```
> survdiff(Survival_Data ~ df$gender)
Call:
survdiff(formula = Survival_Data ~ df$gender)

                    N Observed Expected (O-E)^2/E (O-E)^2/V
df$gender=female  297       12       20      3.19      6.13
df$gender=male    426       33       25      2.54      6.13

 Chisq= 6.1  on 1 degrees of freedom, p= 0.01
```

Null hypothesis for this test is that there is no difference in the survival curve for male vs. female

Since p-value is 0.01, null hypothesis is rejected, and we can conclude that there is some difference of Survival Probability for male & female as we saw in the plot

Prediction:

30 days Survival Rate

We are interested in what happened to the subject after 1 month of their attack with this virus.

The solution is to look at the Survival Curve Plot from Kaplan Mier Estimator and see the Survival Rate or we can also extract the survival probability from the output and check in the data set

```
> summary(survfit(Survival_Data ~ 1), times = 30)
Call: survfit(formula = Survival_Data ~ 1)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   30      3      43    0.241   0.123        0.089        0.655
```

We can see that the survival rate at 30 days is 24.1% without considering recovered cases

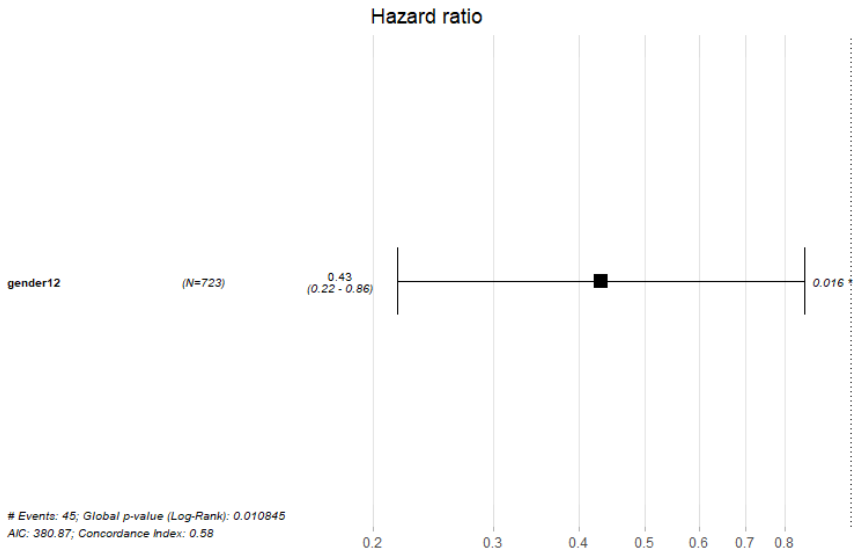
Hazard Ratio

Cox regression model and hazard ratio helps us to evaluate how a variable like Gender will influence the survival rate

```
> out2 = coxph(Surv(Survival_Time, SurvivalStatus) ~ gender12, data = df)
> out2
Call:
coxph(formula = Surv(Survival_Time, SurvivalStatus) ~ gender12,
    data = df)

            coef exp(coef) se(coef)      z      p
gender12 -0.8417    0.4310   0.3501 -2.404 0.0162

Likelihood ratio test=6.49  on 1 df, p=0.01085
n= 723, number of events= 45
```

Firstly, p is low, so gender is very useful in terms of explaining the survival time which is the same conclusion we get by running the log rank test

Second, we can see that the Hazard Ratio = 0.4310. Higher Hazard Ratio means faster the death/failure rate. Since 0.4310 < 1, it means that the loss/hazard for female is less than the loss/hazard for male.



The Hazard Ratio graph shows that the dying rate for female is 0.43 times as that of male at any given time period

Kaplan Mier Survival Plot for Age groups: Shows the chance of survival of age groups (<20 as 1, 20-40 as 2, 40-60 as 3, 60-80 as 4, >80 as 5) of subjects after a certain time period. This is a good visual check to see the significance between the different groups

ggsurvplot(survfit(Survival_Data ~ df$AgeBins),data = df)

Please refer to Figure 1 in Section "Screenshots" at the end of the document

In the first 15 days cutoff, we see that the survival probability for age bin 20- 40 is leading, but then 40-60 age group shows higher survival probability . I.e. Death Rate is higher for 40 – 60 age group as compared to greater than 60 age group. Also, there is slight crossing in the beginning so we can't say which group has higher Survival Probability in the early days of the disease being attacked.

Log Rank Test

We can also do a log rank test to do a statistical check to see if there is any systematic pattern beteenw age & survival probability:

```
> survdiff(Survival_Data ~ df$AgeBins)
Call:
survdiff(formula = Survival_Data ~ df$AgeBins)

                N Observed Expected (O-E)^2/E (O-E)^2/V
df$AgeBins=1   19        0    0.576     0.576     0.598
df$AgeBins=2  164        1    4.926     3.129     3.725
df$AgeBins=3  248        6   15.080     5.468     9.483
df$AgeBins=4  202       23   18.108     1.322     2.369
df$AgeBins=5   90       15    6.310    11.968    14.605

 Chisq= 24.3  on 4 degrees of freedom, p= 7e-05
```

Null hypothesis for this test is that there is no difference in the survival curve for different age groups.

Since p-value is very small, null hypothesis is rejected, and we can conclude that there is some difference of Survival Probability for different age bins as we saw in the plot.

Analysis of Recovery Rate

To transform our data into Recovery object, we need two 'y' variables - Recovery Time & Recovery Status

Recovery Time - Can be recovery time or censor time depending on the status of subject whether they have recovered or still recovering. Recovery time is the time taken for recovery & censor time is the last time we saw a subject to be recovering.

'Surv' function in R helps us to transform these 2 'y' variables into a survival object and makes it easy to read the data

```
> Recovery_Data[135:145]
 [1]  9+ 17+  6+  4+ 10+  4+  5+  6+  5+ 16  11
```

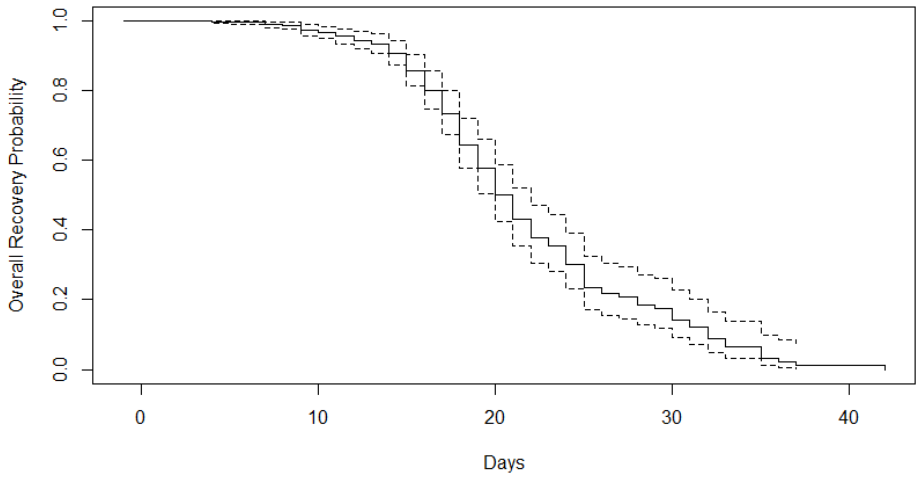
In the output above:

11 : indicates that the subject has recovered in 11 days after he/she got the disease

9+ : indicates that the subject is censored I.e the recovery time for this person is at least 9 days and we don't know if the person is alive or recovered after 9 days

Kaplan Mier Survival Plot for 1 group: Shows the chance of survival of all the subjects after a certain time period :

plot(survfit(Recovery_Data ~ 1),xlab = "Days",ylab = "Overall Recovery Probability")



Observations in the plot:

· All the observations are recovering for the first 5 days from the beginning of symptom onset date, so here the probability is 100%

· We can also see that the chances of recovery after 40 days diminishes to 0

· Dashed lines is the confidence interval which is symmetric around the Kaplan Mier estimated line

Data from the output parameters of Kaplan Mier Estimator:

	time	n.event	n.censor	n.risk
1	-1	0	1	818
2	0	0	18	817
3	1	0	52	799
4	2	0	73	747
5	3	0	65	674
6	4	1	49	609
7	5	1	62	559
8	6	0	56	496
9	7	3	36	440
10	8	1	40	401

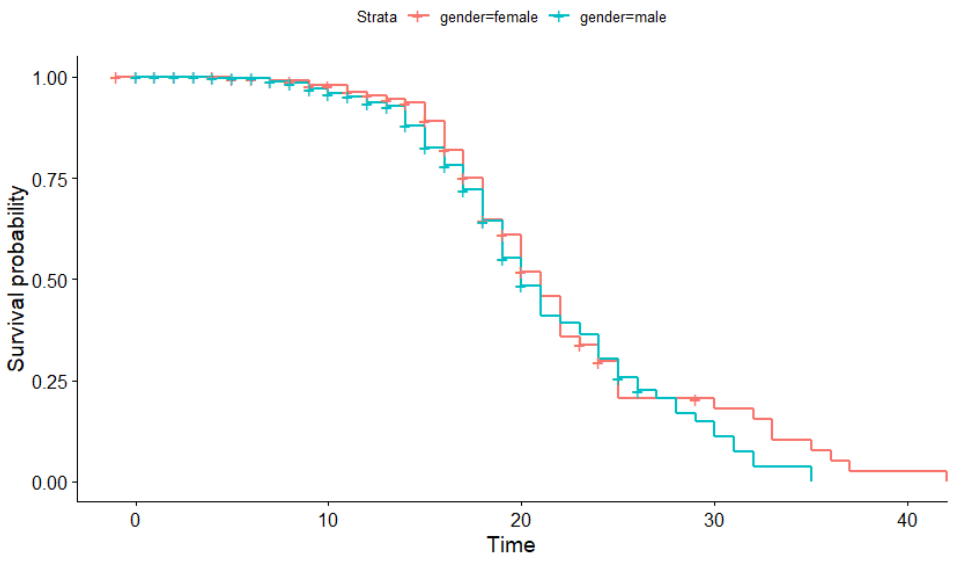
Interpretation:

Row 1: There are 609 subjects in the study as of day 6 and 1 event(recovered) happened on that day

Row 10: There are 401 subjects in the study as of 10 days, 1 recovery happened on that day and 40 left the study.

Kaplan Mier Survival Plot for 2 groups: Shows the chance of recovery of 2 groups (Male & Female) of subjects after a certain time period. This is a good visual check to see the significance between the two groups

ggsurvplot(survfit(Recovery_Data ~ df$gender),data = df)



Recovery probability for gender 1(male) crosses is more or less the same than the recovery probability of the females by cutoff time of 27 days. After that female recovers better.

Log Rank Test

We can also do a log rank test to do a statistical check to see if there is any systematic pattern btw gender & recovery probability:

```
> survdiff(Recovery_Data ~ df$gender)
Call:
survdiff(formula = Recovery_Data ~ df$gender)

                    N Observed Expected (O-E)^2/E (O-E)^2/V
df$gender=female  346       61     67.8     0.676       1.5
df$gender=male    472       79     72.2     0.634       1.5

 Chisq= 1.5  on 1 degrees of freedom, p= 0.2
```

Null hypothesis for this test is that there is no difference in the recovery curve for male vs. female

Since p-value is 0.2, null hypothesis cannot be rejected, and we can conclude that recovery rate for both men and women is almost the same.

Prediction:

30 days Recovery Rate

We are interested in what happened to the subject after 1 month of their attack with this virus.

The solution is to look at the Recovery Curve Plot from Kaplan Mier Estimator and see the Recovery Rate or we can also extract the recovery probability from the output and check in the data set

```
> summary(survfit(Recovery_Data ~ 1), times = 30)
Call: survfit(formula = Recovery_Data ~ 1)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   30     16     127    0.143  0.0337       0.0902        0.227
```

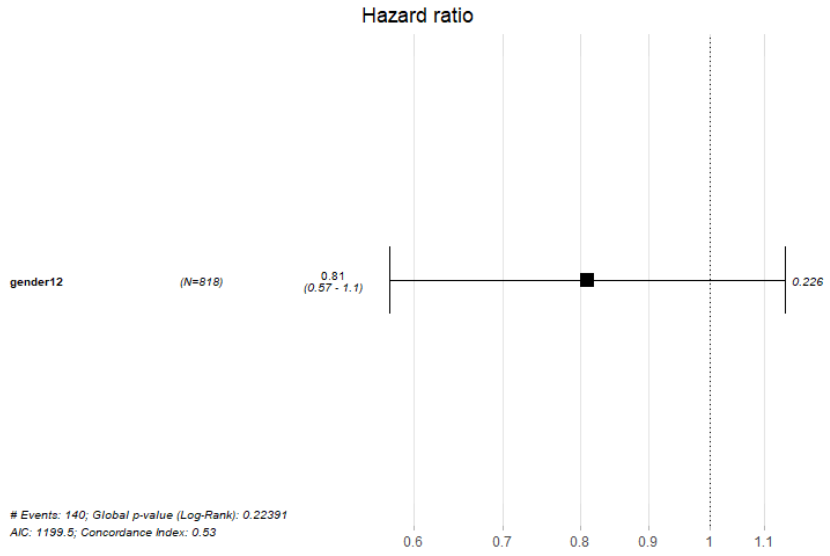
We can see that the recovery rate at 30 days is 14.3% without considering death cases

Hazard Ratio using Cox regression model:

```
> out2 = coxph(Surv(Recovery_Time, RecoveryStatus) ~ gender12, data = df)
> out2
Call:
coxph(formula = Surv(Recovery_Time, RecoveryStatus) ~ gender12,
    data = df)

           coef exp(coef) se(coef)      z     p
gender12 -0.2117    0.8092   0.1749 -1.211 0.226

Likelihood ratio test=1.48  on 1 df, p=0.2239
n= 818, number of events= 140
```

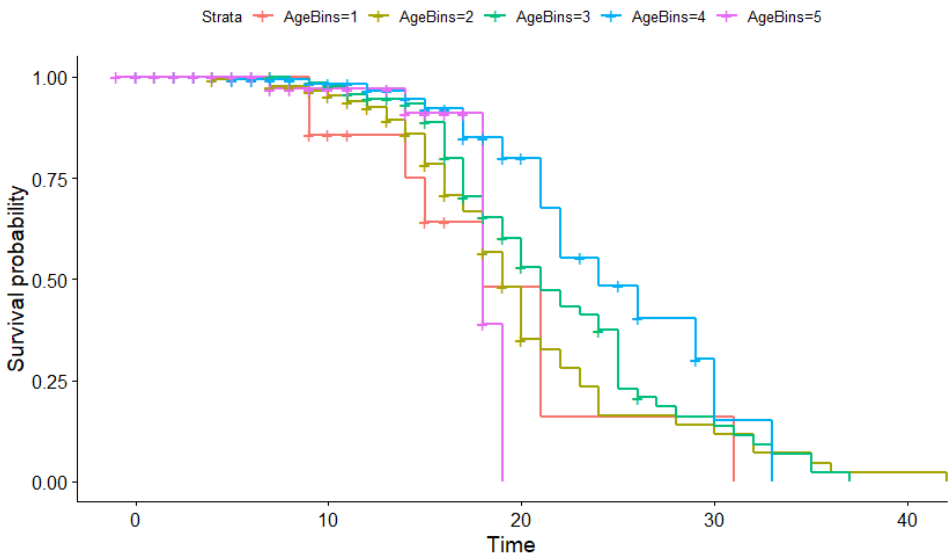
As we see p-value is 0.223, gender doesn't play much role in explaining the recovery time which is similar to the result obtained by the log rank test. Hazard ratio = 0.81 again means that the recovery rate for female and male is same. Below plot gives a snapshot of the same conclusion.

ggforest(out2, data = df)



Kaplan Mier Survival Plot for Age groups: Shows the chance of recovery of age groups (<20 as 1, 20-40 as 2, 40-60 as 3, 60-80 as 4, >80 as 5) of subjects after a certain time period. This is a good visual check to see the significance between the different groups

ggsurvplot(survfit(Recovery_Data ~ df$AgeBins),data = df)



In the first 32 days cutoff, we see that the recovery probability for age bin 40- 60 is mostly leading, except around 17-18 cutoff where 80 + is leading. Less then 40 age group shows higher recovery probability after 18 days. I.e. Recovery Rate is higher for 20 – 60 age group as compared to greater than 60 age group. Also, there is a lot of crossing in the beginning so we can't say which group has higher Recovery Probability in the early days of the disease being attacked.

Log Rank Test

We can also do a log rank test to do a statistical check to see if there is any systematic pattern btw age & survival probability:

```
> survdiff(Recovery_Data ~ df$AgeBins)
Call:
survdiff(formula = Recovery_Data ~ df$AgeBins)

                N Observed Expected (O-E)^2/E (O-E)^2/V
df$AgeBins=1   27        8     4.82     2.095     2.368
df$AgeBins=2  214       51    42.10     1.881     3.067
df$AgeBins=3  299       57    59.59     0.113     0.220
df$AgeBins=4  196       17    28.31     4.517     6.267
df$AgeBins=5   82        7     5.18     0.642     0.729

 Chisq= 10.2  on 4 degrees of freedom, p= 0.04
```

Null hypothesis for this test is that there is no difference in the recovery curve for different age groups

Since p-value is very small, null hypothesis is rejected, and we can conclude that there is some difference of Recovery Probability for different age bins as we saw in the plot

Trial 2 :

Death Recovery Censored – independently - with the other event treated as infinity(120)

Survival Analysis (Event:Death)

```
> Survival_Data[130:180]
 [1]   5+   3+   6+   7+   9+   9+  17+   6+   4+  10+   4+   5+   6+   5+ 120+ 120+ 120+   2+ 120+ 120+
20    0+
[23]   1+ 120+  10+   4+   3+   6+  14+   2+   9+   1+   5+   6+   7+  11+   1+   6+   1+   5+   4+  10+
12+  12+
[45]   2+   3+  22   15+  13+  11+  27
```

In the output above:

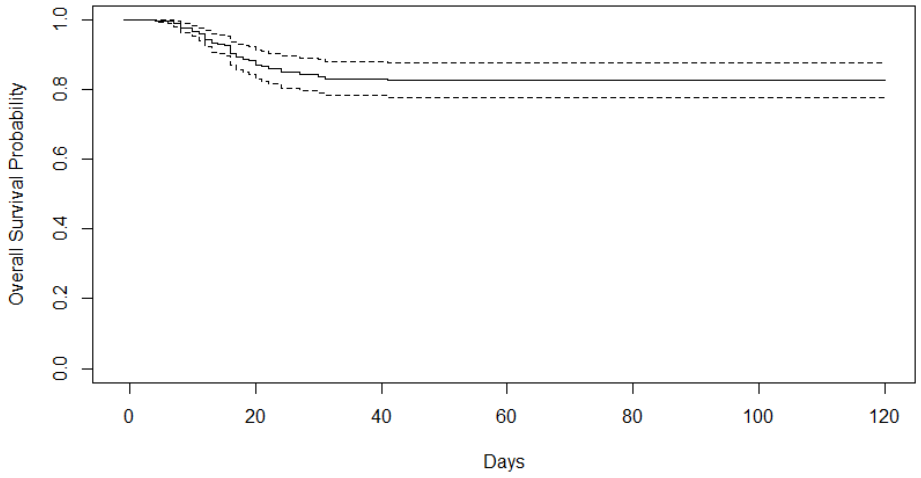
22 : indicates that the subject has died in 22 days after he/she got the disease

3+ : indicates that the subject is censored I.e the survival time for this person is at least 3 days

120+ : Indicates recovered cases

Kaplan Mier Survival Plot for 1 group: Shows the chance of survival of all the subjects after a certain time period :

plot(survfit(Survival_Data ~ 1),xlab = "Days",ylab = "Overall Survival Probability")



Observations in the plot:

1. All the observations are alive for the first 2 days from the beginning of symptom onset date, so here the survival probability is 100%
2. We can also see that the chances of survival after 40 days drops to 80% and stays the same till the end .
3. Dashed lines are the confident interval which is symmetric around the Kaplan Mier estimated line

Data from the output parameters of Kaplan Mier Estimator:

Please refer to Figure 2 in Section “Screenshots” at the end of the document

Interpretation:

Row 1: There are 863 subjects in the study as of day 6 and 1 event(death) happened on that day

Row 16: There are 446 subjects in the study as of 10 days, 6 event(death) & 40 censored (out of study) cases happened on that day

Kaplan Mier Survival Plot for 2 groups: Shows the chance of survival of 2 groups (Male & Female) of subjects after a certain time period.

ggsurvplot(survfit(Survival_Data ~ df$gender),data = df)

Please refer to Figure 3 in Section “Screenshots” at the end of the document

No matter what the cutoff for the time is, Survival Probability for female is always greater than the survival probability for male. I.e. Death Rate for male is higher. Also, there is slight crossing in the beginning so we can't say which group has higher Survival Probability in the early days of the disease being attacked.

Log Rank Test

We can also do a log rank test to do a statistical check to see if there is any systematic pattern btw gender & survival probability:

Please refer to Figure 4 in Section "Screenshots" at the end of the document

Null hypothesis for this test is that there is no difference in the survival curve for male vs. female

Since p-value is 0.02, null hypothesis is rejected, and we can conclude that there is some difference of Survival Probability for male & female as we saw in the plot

Prediction:

30 days Survival Rate

We are interested in what happened to the subject after 1 month of their attack with this virus.

The solution is to look at the Survival Curve Plot from Kaplan Mier Estimator and see the Survival Rate or we can also extract the survival probability from the output and check in the data set

```
> summary(survfit(Survival_Data ~ 1), times = 30)
Call: survfit(formula = Survival_Data ~ 1)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   30    143      43    0.838  0.0244        0.791        0.887
```

We can see that the survival rate at 30 days is 83.8% without considering recovered cases

Hazard Ratio

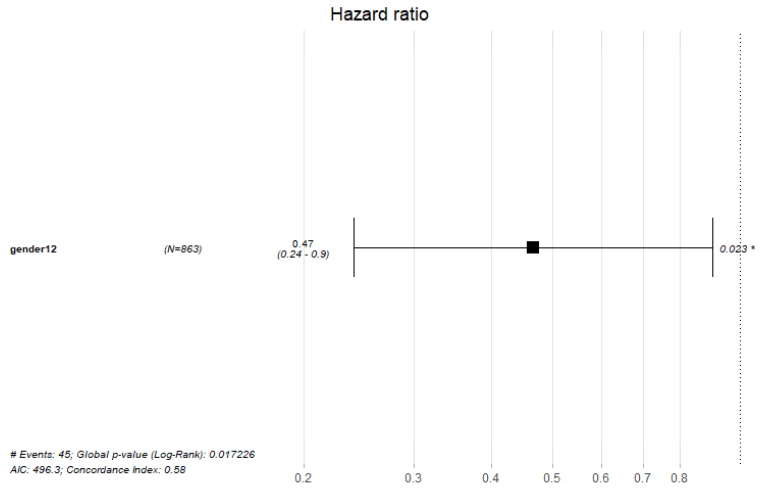
Cox regression model and hazard ratio helps us to evaluate how a variable like Gender will influence the survival rate

```
> out2 = coxph(Surv(Survival_Time, SurvivalStatus) ~ gender12, data = df)
> out2
Call:
coxph(formula = Surv(Survival_Time, SurvivalStatus) ~ gender12,
    data = df)

            coef exp(coef) se(coef)      z      p
gender12 -0.7639    0.4658   0.3372 -2.265 0.0235

Likelihood ratio test=5.67  on 1 df, p=0.01723
n= 863, number of events= 45
```

Firstly, p is low, so gender is useful in terms of explaining the survival time which is the same conclusion we get by running the log rank test

Second, we can see that the Hazard Ratio = 0.47. Higher Hazard Ratio means faster the death/failure rate. Since 0.47 < 1, it means that the loss/hazard for female is less than the loss/hazard for male.

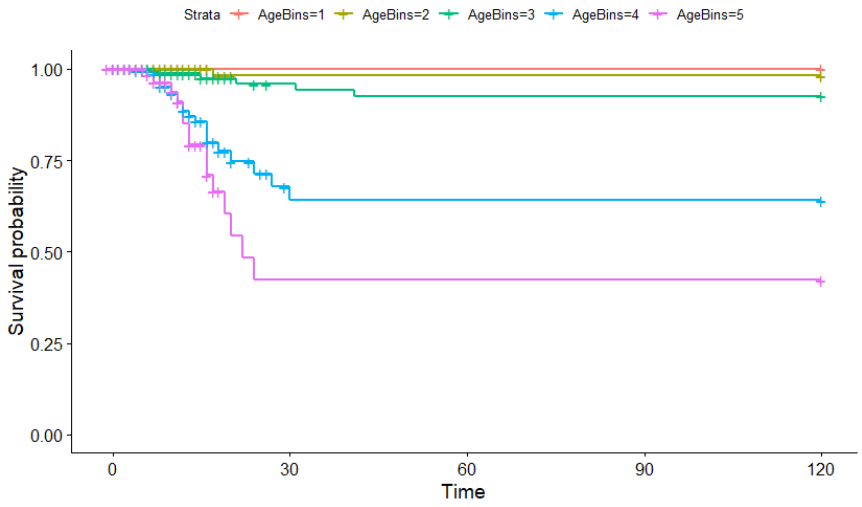
ggforest(out2, data = df)



The Hazard Ratio graph shows that the dying rate for female is 0.47 times as that of male at any given time period

Kaplan Mier Survival Plot for Age groups: Shows the chance of survival of age groups (<20 as 1, 20-40 as 2, 40-60 as 3, 60-80 as 4, >80 as 5) of subjects after a certain time period. This is a good visual check to see the significance between the different groups

ggsurvplot(survfit(Survival_Data ~ df$AgeBins),data = df)



We see that the survival probability for age bin less than 20 is leading while as the age increases the survival probability reduces with a minimum of 0.45 . I.e. Death Rate is highest for higher age group of more than 60 years. Also, there is slight crossing in the beginning so we can't say which group has higher Survival Probability in the early days of the disease being attacked.

Log Rank Test

We can also do a log rank test to do a statistical check to see if there is any systematic pattern btw age & survival probability:

```
> survdiff(Survival_Data ~ df$AgeBins)
Call:
survdiff(formula = Survival_Data ~ df$AgeBins)

                N Observed Expected (O-E)^2/E (O-E)^2/V
df$AgeBins=1   27        0     1.84      1.84      1.94
df$AgeBins=2  215        1    12.35     10.44     14.62
df$AgeBins=3  305        6    16.20      6.42     10.14
df$AgeBins=4  219       23    10.54     14.73     19.75
df$AgeBins=5   97       15     4.06     29.46     32.81

 Chisq= 65  on 4 degrees of freedom, p= 3e-13
```

Null hypothesis for this test is that there is no difference in the survival curve for different age groups.

Since p-value is very small, null hypothesis is rejected, and we can conclude that there is some difference of Survival Probability for different age bins as we saw in the plot.

Analysis of Recovery Rate

To transform our data into Recovery object, we need two ‘y’ variables - Recovery Time & Recovery Status

Recovery Time - Can be recovery time or censor time depending on the status of subject whether they have recovered or still recovering. Recovery time is the time taken for recovery & censor time is the last time we saw a subject to be recovering. For cases where we know that another event ‘death’ has happened, we change the survival time as infinity which is 120 in this

case.

‘Surv’ function in R helps us to transform these 2 ‘y’ variables into a survival object and makes it easy to read the data

```
> Recovery_Data[140:150]
 [1]   4+   5+   6+   5+  16   11    4    2+  24   20  120+
```

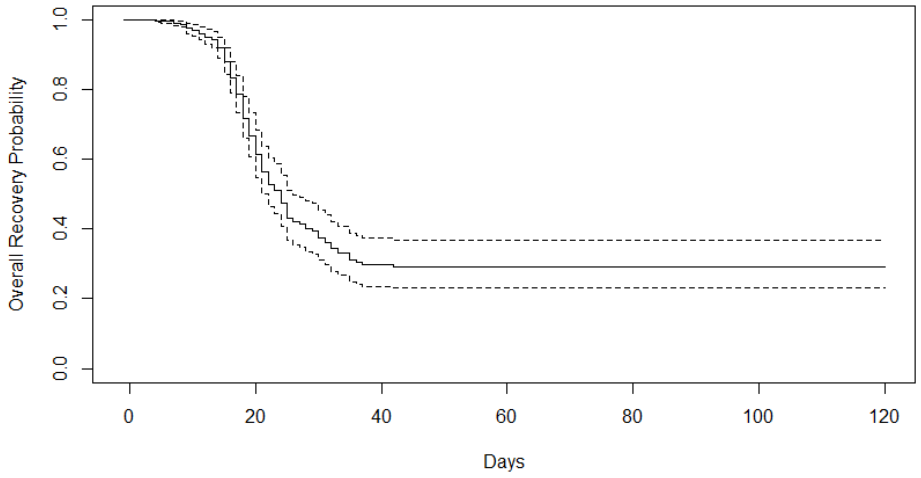
In the output above:

11 : indicates that the subject has recovered in 11 days after he/she got the disease

4+ : indicates that the subject is censored I.e the recovery time for this person is at least 4 days and we don't know if the person is alive or dead after4 days

Kaplan Mier Survival Plot for 1 group: Shows the chance of survival of all the subjects after a certain time period :

plot(survfit(Recovery_Data ~ 1),xlab = "Days",ylab = "Overall Recovery Probability")



Observations in the plot:

· All the observations are recovering for the first couple of days from the beginning of symptom onset date, so here the probability is 100%

· We can also see that the chances of recovery after 40 days diminishes to30 % and remains the same

· Dashed lines is the confidence interval which is symmetric around the Kaplan Mier estimated line

Data from the output parameters of Kaplan Mier Estimator:

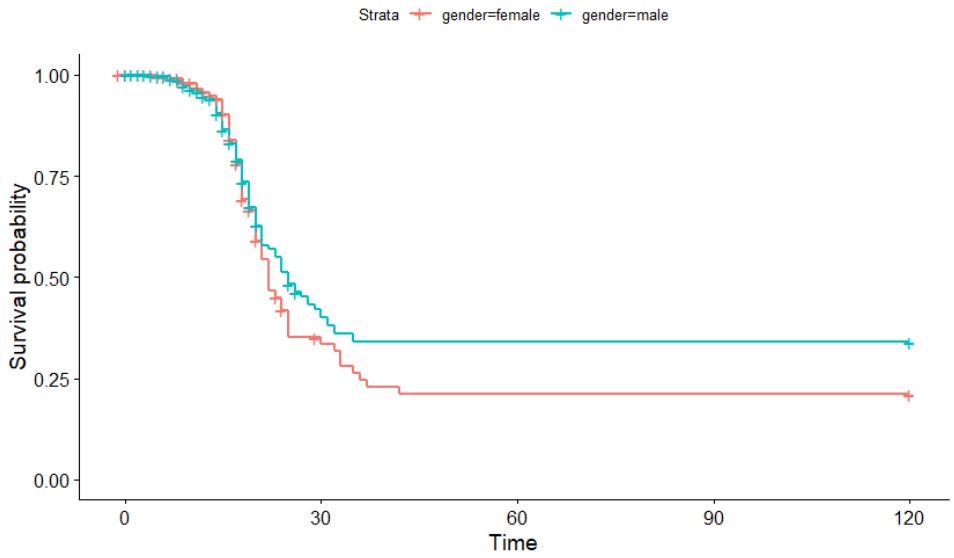
	time	n.event	n.censor	n.risk
1	-1	0	1	863
2	0	0	18	862
3	1	0	52	844
4	2	0	73	792
5	3	0	65	719
6	4	1	49	654
7	5	1	62	604
8	6	0	56	541
9	7	3	36	485
10	8	1	40	446

Interpretation:

Row 1: There are 654 subjects in the study as of day 6 and 1 event(recovered) happened on that day

Row 10: There are 446 subjects in the study as of 10 days, 1 recovery happened on that day

Kaplan Mier Survival Plot for 2 groups: Shows the chance of recovery of 2 groups (Male & Female) of subjects after a certain time period. This is a good visual check to see the significance between the two groups



Recovery probability for gender 1(male) is more or less the same than the recovery probability of the females by cutoff time of 37 days. After that female recovers better.

Log Rank Test

We can also do a log rank test to do a statistical check to see if there is any systematic pattern between gender & recovery probability:

```
> survdiff(Recovery_Data ~ df$gender)
Call:
survdiff(formula = Recovery_Data ~ df$gender)

                  N Observed Expected (O-E)^2/E (O-E)^2/V
df$gender=female 358       61     54.6     0.742      1.29
df$gender=male   505       79     85.4     0.475      1.29

 Chisq= 1.3  on 1 degrees of freedom, p= 0.3
```

Null hypothesis for this test is that there is no difference in the recovery curve for male vs. female

Since p-value is 0.3, null hypothesis cannot be rejected, and we can conclude that recovery rate for both men and women is almost the same.

Prediction:

30 days Recovery Rate

We are interested in what happened to the subject after 1 month of their attack with this virus.

The solution is to look at the Recovery Curve Plot from Kaplan Mier Estimator and see the Recovery Rate or we can also extract the recovery probability from the output and check in the data set

```
> summary(survfit(Recovery_Data ~ 1), times = 30)
Call: survfit(formula = Recovery_Data ~ 1)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   30     61     127    0.376  0.0365        0.311        0.454
```

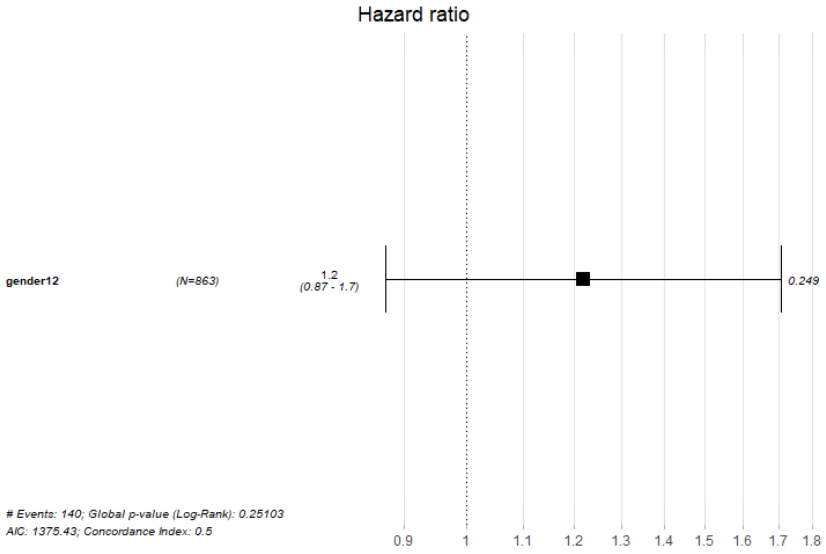
We can see that the survival rate at 30 days is 37.6% without considering death cases

Hazard Ratio using Cox regression model:

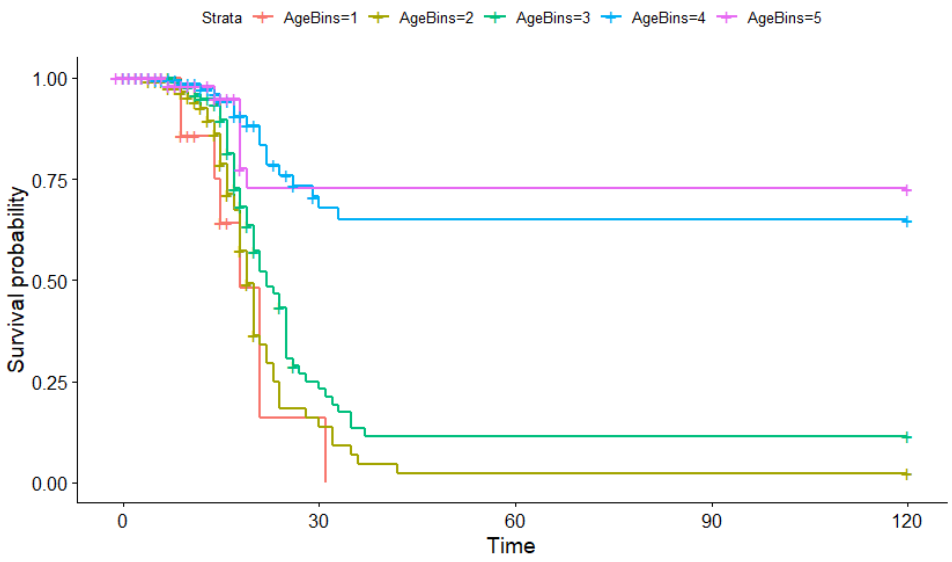
```
> out2 = coxph(Surv(Recovery_Time, RecoveryStatus) ~ gender12, data = df)
> out2
Call:
coxph(formula = Surv(Recovery_Time, RecoveryStatus) ~ gender12,
    data = df)

           coef exp(coef) se(coef)     z     p
gender12 0.1977    1.2186   0.1714 1.153 0.249

Likelihood ratio test=1.32  on 1 df, p=0.251
n= 863, number of events= 140
```

As we see p-value is 0.251, gender doesn't play much role in explaining the recovery time which is similar to the result obtained by the log rank test. Hazard ratio = 1.2 again means that the recovery rate for female and male is same. Below plot gives a snapshot of the same conclusion.

ggforest(out2, data = df)



Kaplan Mier Survival Plot for Age groups: Shows the chance of recovery of age groups (<20 as 1, 20-40 as 2, 40-60 as 3, 60-80 as 4, >80 as 5) of subjects after a certain time period. This is a good visual check to see the significance between the different groups



The youngest age group is recovering in 30 days. But, the higher age groups are either taking time or not recovering at all in the long run. Also, there is a lot of crossing in the beginning so we can't say which group has higher Recovery Probability in the early days of the disease being attacked.

Log Rank Test

We can also do a log rank test to do a statistical check to see if there is any systematic pattern between age & recovery probability:

```
> survdiff(Recovery_Data ~ df$AgeBins)
Call:
survdiff(formula = Recovery_Data ~ df$AgeBins)

                N Observed Expected (O-E)^2/E (O-E)^2/V
df$AgeBins=1   27        8     3.11      7.70      8.27
df$AgeBins=2  215       51    25.46     25.62     33.39
df$AgeBins=3  305       57    42.99      4.57      6.99
df$AgeBins=4  219       17    46.38     18.61     29.53
df$AgeBins=5   97        7    22.06     10.28     13.04

 Chisq= 73.3  on 4 degrees of freedom, p= 5e-15
```

Null hypothesis for this test is that there is no difference in the recovery curve for different age groups

Since p-value is very small, null hypothesis is rejected, and we can conclude that there is some difference of Recovery Probability for different age bins as we see in the plot

So, this analysis in sync with the assumption that is currently in the world that we should actually keep the 60's age group people safe from this virus because their healing/recovery power is very less than the other age groups

CONCLUSION

We had a clear set of questions as our objective when we started working on this Covid-19 data set and like you have seen, by deploying different kind of models we got great insights few of which are in-line with the assumptions in the society regarding the virus.

For Time Series Analysis, we forecasted the number of confirmed cases, deaths & survival numbers for the future

Then we wanted to understand if there is a country level & age level effect on the death rate. Based on the fixed & random effect analysis we found that there is a significant effect of countries & ages on the number of deaths

And lastly by performing Survival Analysis, we concluded that:

- 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

NEXT STEPS

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

TEAM CONTRIBUTION:

The team sat together to brainstorm on the exploratory data analysis and how the data could be dissection to derive most insights. We then divided the project into the following sections and picked up each part.

a. Time Series Analysis
b. Fixed & Random Effect
c. Survival Analysis

We shared the results and the problem in each section and tried to reach to a solution. The whole team contributed equally to the success of this project.

Screenshots:

We included this section because there was some problem with pasting some images at their respected places

Figure 1:

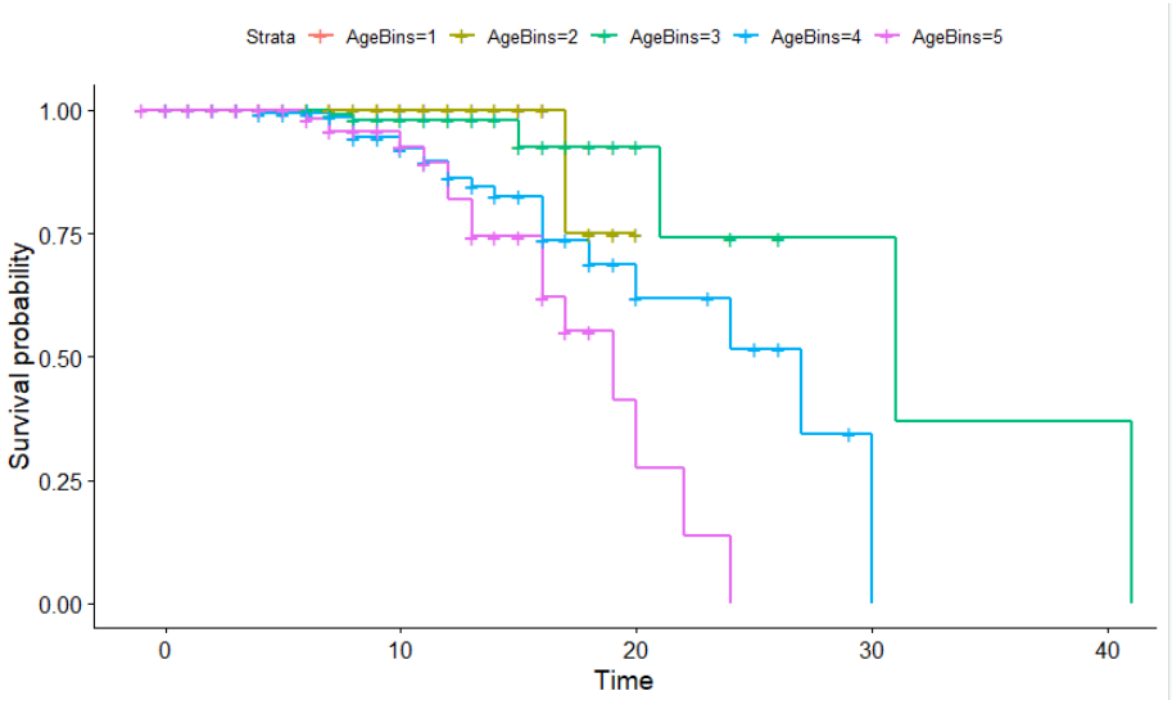


Figure 2:

	time	n.event	n.censor	n.risk
1	-1	0	1	863
2	0	0	18	862
3	1	0	52	844
4	2	0	73	792
5	3	0	65	719
6	4	1	49	654
7	5	0	62	604
8	6	1	56	542
9	7	3	36	485
10	8	6	40	446

Figure 3:

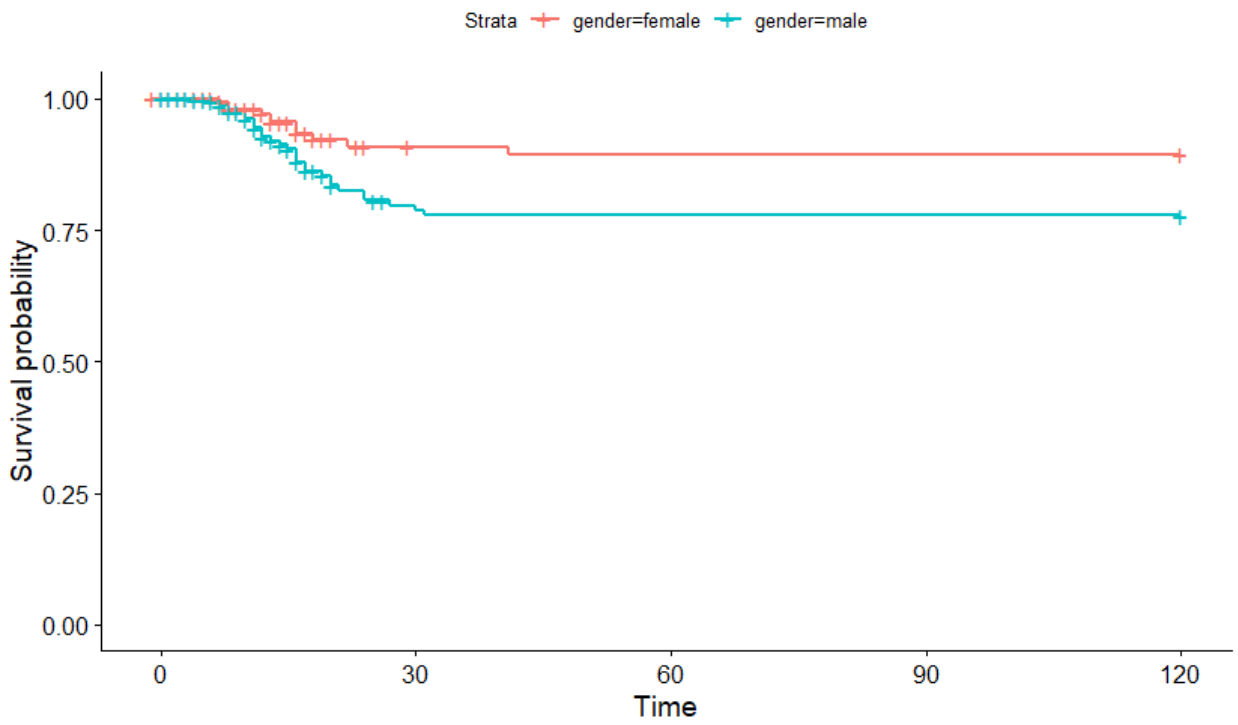


Figure 4:

```
> survdiff(Survival_Data ~ df$gender)
Call:
survdiff(formula = Survival_Data ~ df$gender)

                    N Observed Expected (O-E)^2/E (O-E)^2/V
df$gender=female  358       12     19.7      3.01       5.4
df$gender=male    505       33     25.3      2.35       5.4

 Chisq= 5.4  on 1 degrees of freedom, p= 0.02
```