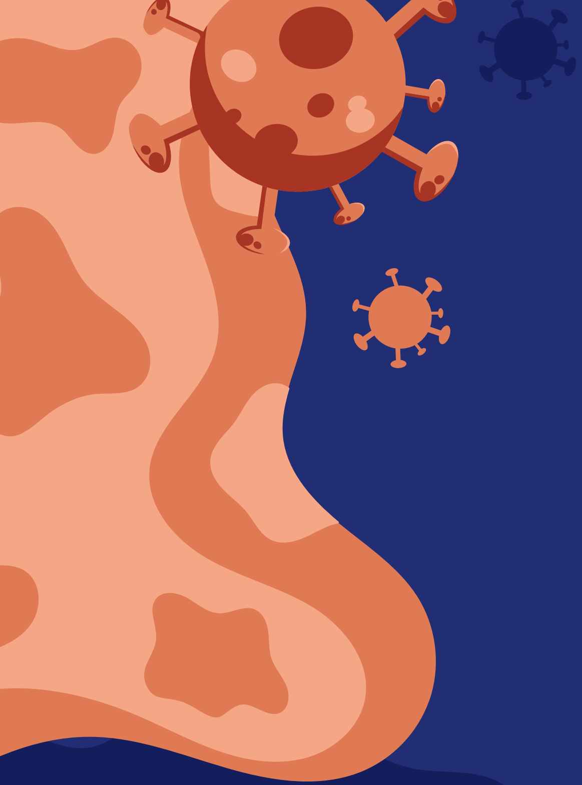
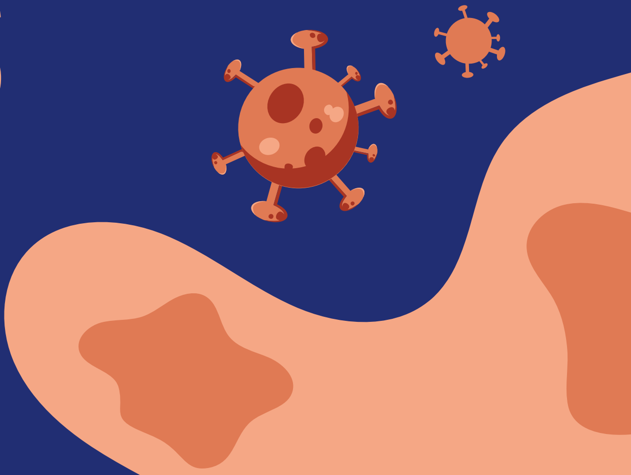


STATISTICAL ANALYSIS: CORONAVIRUS



Disclaimer
correct data and information may be present (obtained from kaggle.com via KCDC). This presentation is not meant to be used
r research purposes. The project was conducted mainly to explore new statistical tools and derive insights from the data

TABLE OF CONTENTS

01

Introduction

-Background about the virus
-Dataset and variables used

02

EDA

-Overview of data from KCDC

03

Statistical Analysis

-T-test
- Hazard Rate Model

04

Network Analysis

Who infected whom?

05

Conclusion

-Key Insights
-Limitations

01

INTRODUCTION

COVID-19 Quick Facts

Origin: Wuhan, Hubei Province, China

Virus is thought to spread mainly from person-to-person

Incubation period ranges from 1-14 days (average ~5 days)

Coronaviruses are zoonotic; transmitted between animals and people

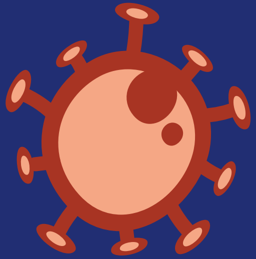
Common signs: Respiratory symptoms, fever, cough

Can cause pneumonia, severe acute respiratory syndrome, kidney failure and even death

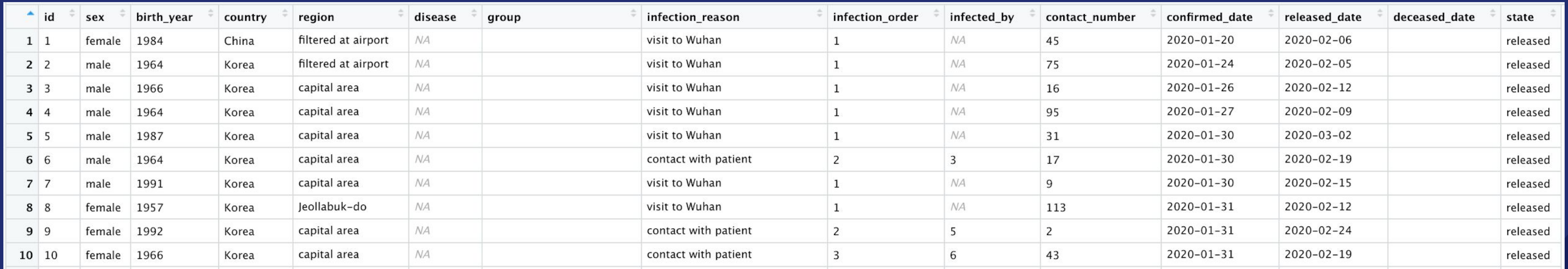
DATASET

COVID-19 in South Korea (KCDC)

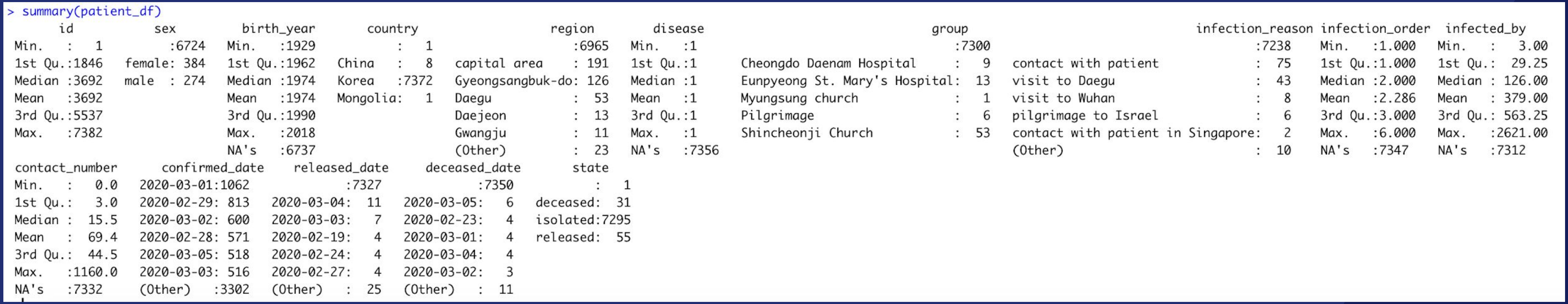
- Patient.csv (15 variables)
- Route.csv (7 variables)



Overview of KCDC Dataset



	id	sex	birth_year	country	region	disease	group	infection_reason	infection_order	infected_by	contact_number	confirmed_date	released_date	deceased_date	state
1	1	female	1984	China	filtered at airport	NA		visit to Wuhan	1	NA	45	2020-01-20	2020-02-06		released
2	2	male	1964	Korea	filtered at airport	NA		visit to Wuhan	1	NA	75	2020-01-24	2020-02-05		released
3	3	male	1966	Korea	capital area	NA		visit to Wuhan	1	NA	16	2020-01-26	2020-02-12		released
4	4	male	1964	Korea	capital area	NA		visit to Wuhan	1	NA	95	2020-01-27	2020-02-09		released
5	5	male	1987	Korea	capital area	NA		visit to Wuhan	1	NA	31	2020-01-30	2020-03-02		released
6	6	male	1964	Korea	capital area	NA		contact with patient	2	3	17	2020-01-30	2020-02-19		released
7	7	male	1991	Korea	capital area	NA		visit to Wuhan	1	NA	9	2020-01-30	2020-02-15		released
8	8	female	1957	Korea	Jeollabuk-do	NA		visit to Wuhan	1	NA	113	2020-01-31	2020-02-12		released
9	9	female	1992	Korea	capital area	NA		contact with patient	2	5	2	2020-01-31	2020-02-24		released
10	10	female	1966	Korea	capital area	NA		contact with patient	3	6	43	2020-01-31	2020-02-19		released



```
> summary(patient_df)
       id            sex        birth_year       country                   region        disease                                  group                                  infection_reason infection_order  infected_by
 Min.   :   1         :6724   Min.   :1929           :   1                   :6965   Min.   :1                                    :7300                                    :7238   Min.   :1.000   Min.   :   3.00
 1st Qu.:1846   female: 384   1st Qu.:1962   China   :   8   capital area    : 191   1st Qu.:1      Cheongdo Daenam Hospital     :   9   contact with patient             :  75   1st Qu.:1.000   1st Qu.:  29.25
 Median :3692   male  : 274   Median :1974   Korea   :7372   Gyeongsangbuk-do: 126   Median :1      Eunpyeong St. Mary's Hospital:  13   visit to Daegu                   :  43   Median :2.000   Median : 126.00
 Mean   :3692                 Mean   :1974   Mongolia:   1   Daegu           :  53   Mean   :1      Myungsung church             :   1   visit to Wuhan                   :   8   Mean   :2.286   Mean   : 379.00
 3rd Qu.:5537                 3rd Qu.:1990                   Daejeon         :  13   3rd Qu.:1      Pilgrimage                   :   6   pilgrimage to Israel             :   6   3rd Qu.:3.000   3rd Qu.: 563.25
 Max.   :7382                 Max.   :2018                   Gwangju         :  11   Max.   :1      Shincheonji Church           :  53   contact with patient in Singapore:   2   Max.   :6.000   Max.   :2621.00
                              NA's   :6737                   (Other)         :  23   NA's   :7356                                        (Other)                          :  10   NA's   :7347    NA's   :7312
 contact_number      confirmed_date    released_date     deceased_date      state
 Min.   :   0.0   2020-03-01:1062              :7327              :7350           :   1
 1st Qu.:   3.0   2020-02-29: 813   2020-03-04:  11   2020-03-05:   6   deceased:  31
 Median :  15.5   2020-03-02: 600   2020-03-03:   7   2020-02-23:   4   isolated:7295
 Mean   :  69.4   2020-02-28: 571   2020-02-19:   4   2020-03-01:   4   released:  55
 3rd Qu.:  44.5   2020-03-05: 518   2020-02-24:   4   2020-03-04:   4
 Max.   :1160.0   2020-03-03: 516   2020-02-27:   4   2020-03-02:   3
 NA's   :7332     (Other)   :3302   (Other)   :  25   (Other)   :  11
```

02

EXPLORATORY DATA ANALYSIS

196,640

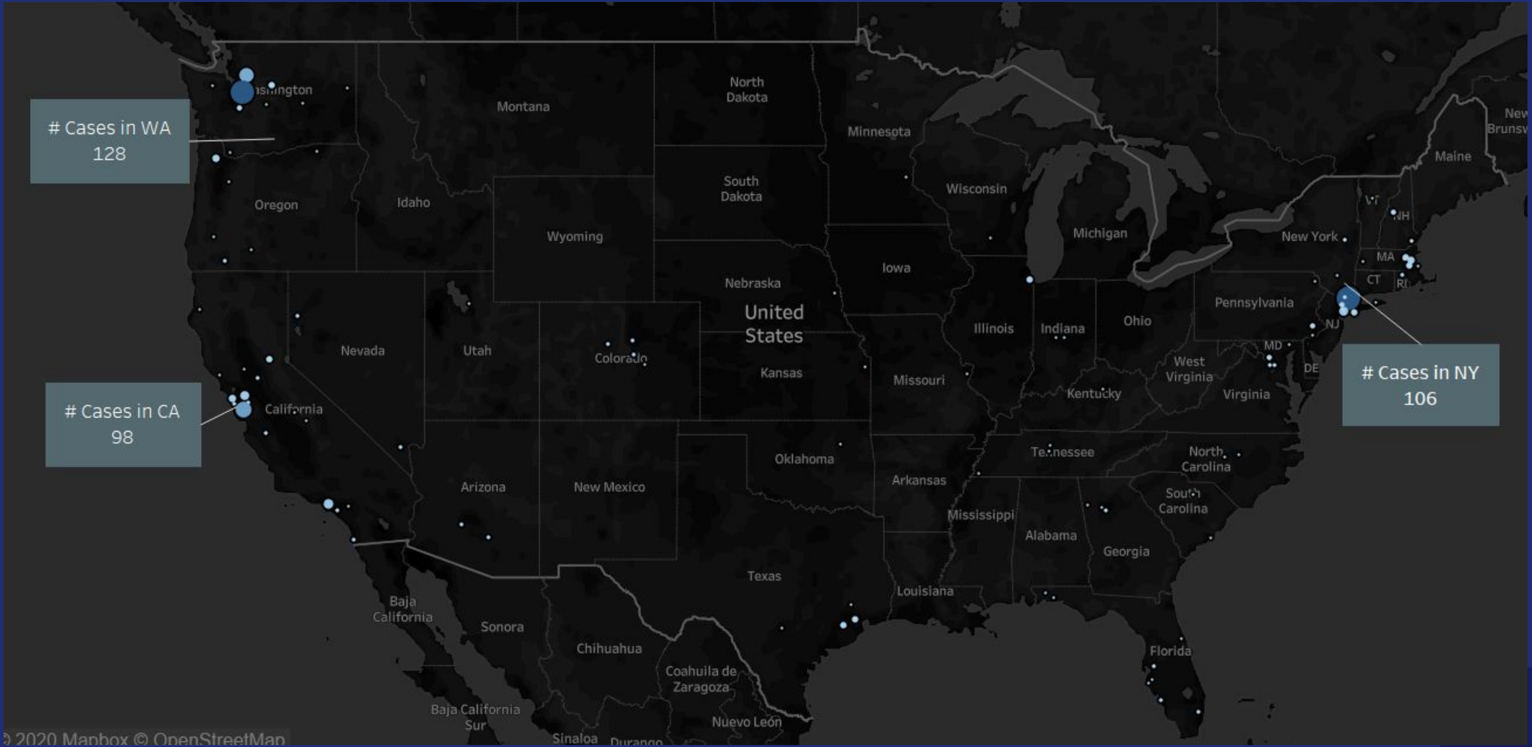
Global Cases

(as of 3/17/2020)

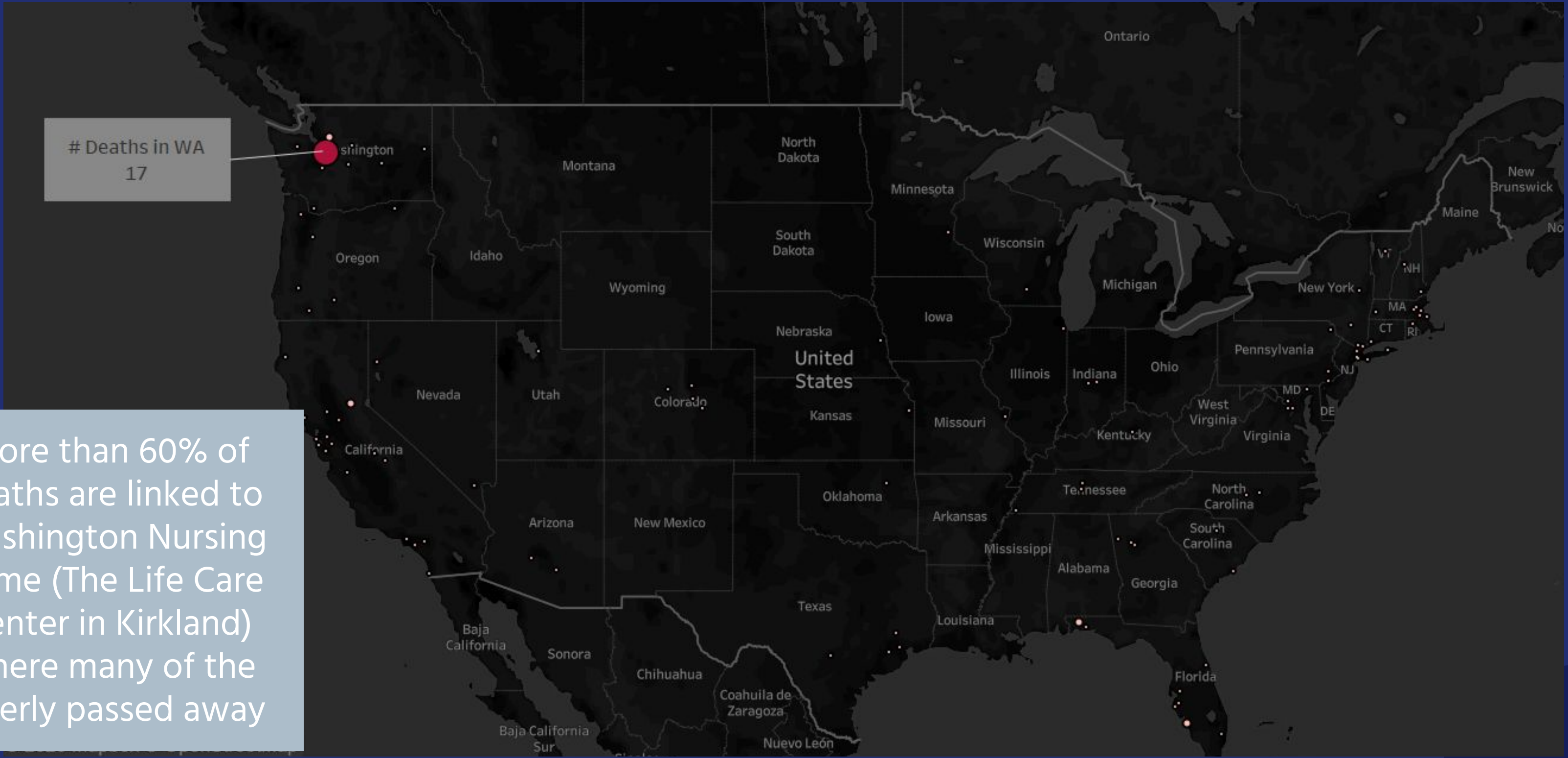
153

Countries affected

Confirmed Cases in US

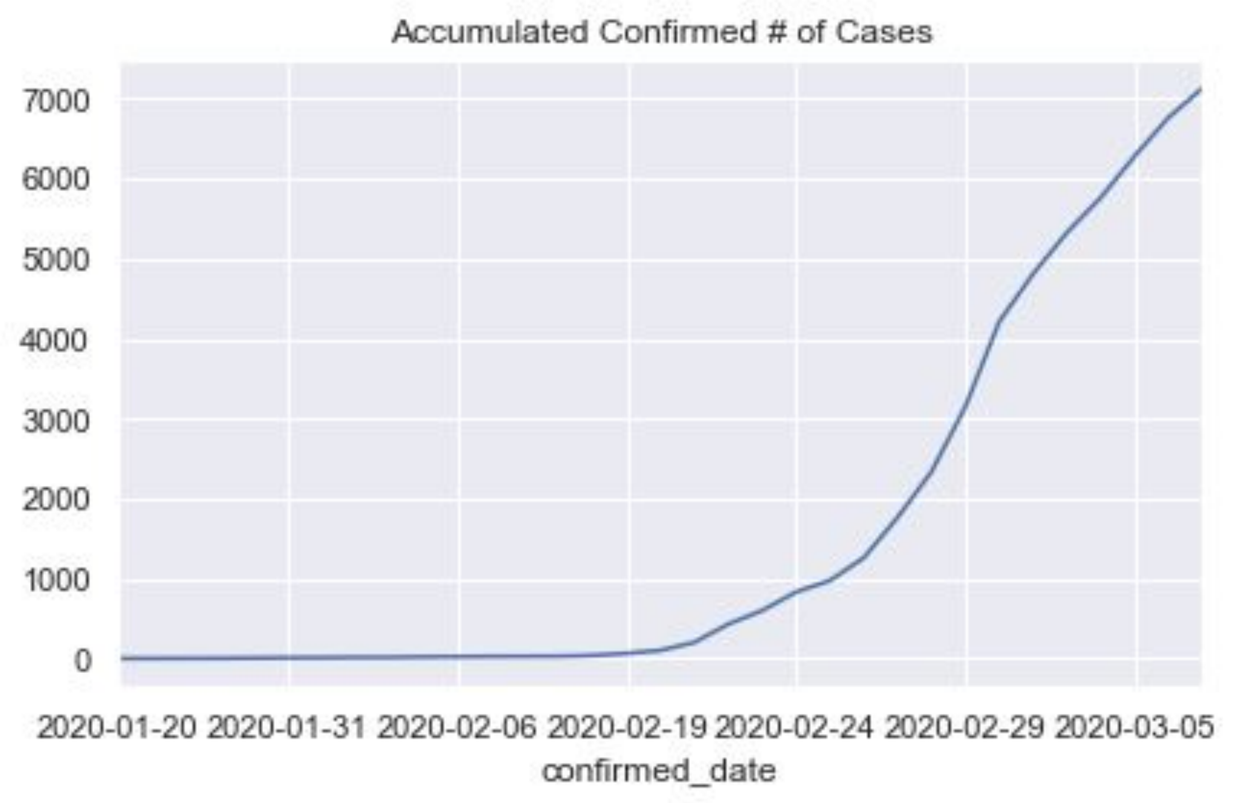


Death by State in US



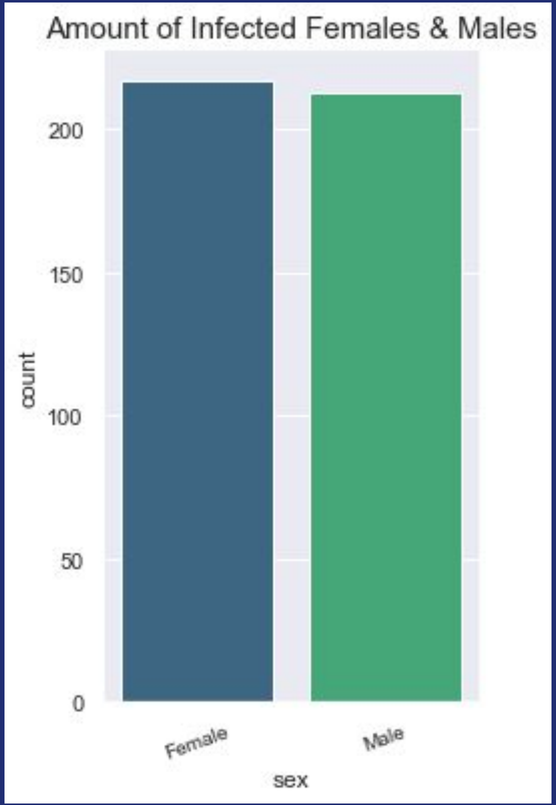
More than 60% of deaths are linked to Washington Nursing home (The Life Care Center in Kirkland) where many of the elderly passed away

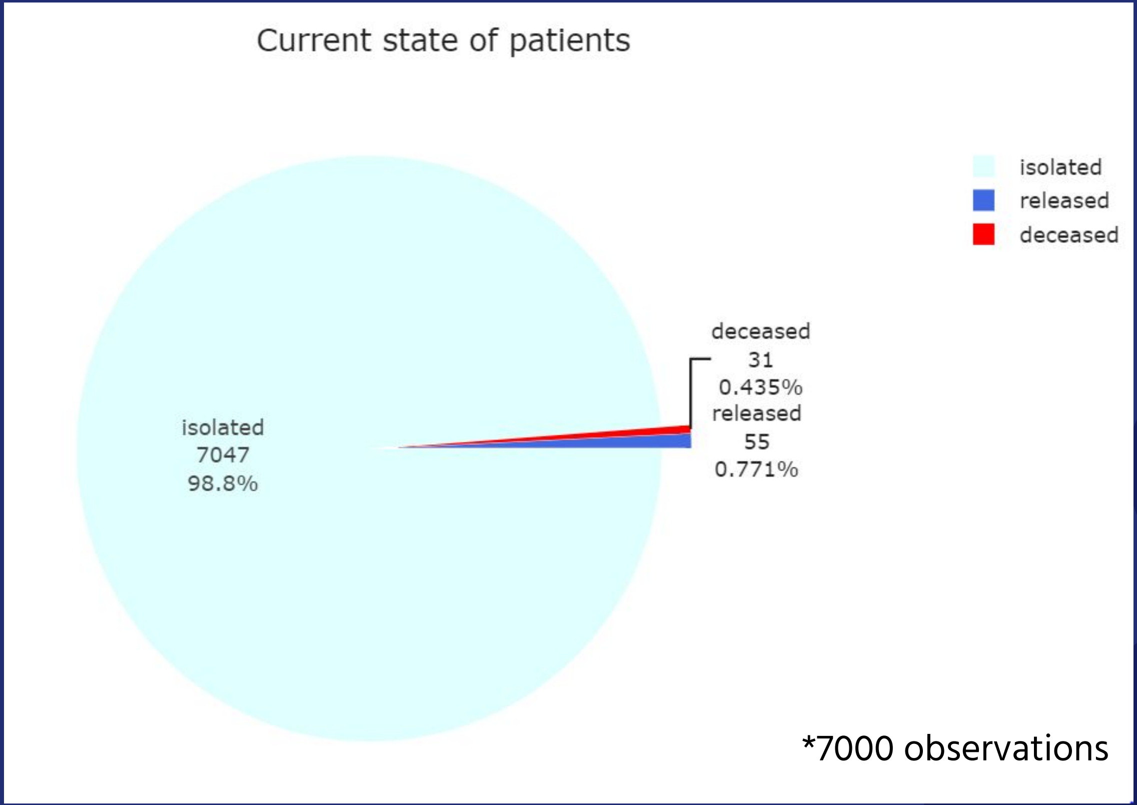
Cases over time in South Korea



After Feb 24th there is an exponential increase in the number of confirmed cases reaching just over 7000 infected individuals

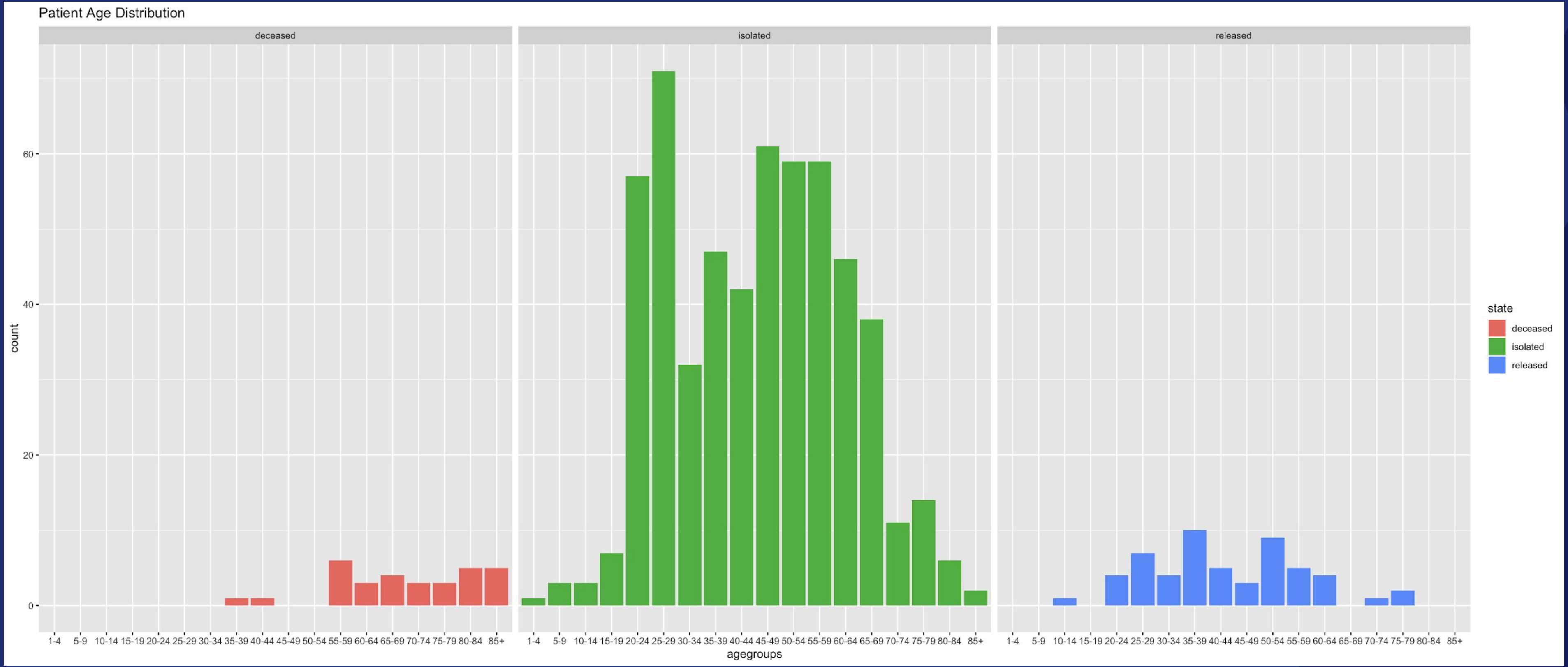
Overview of Infected Patients



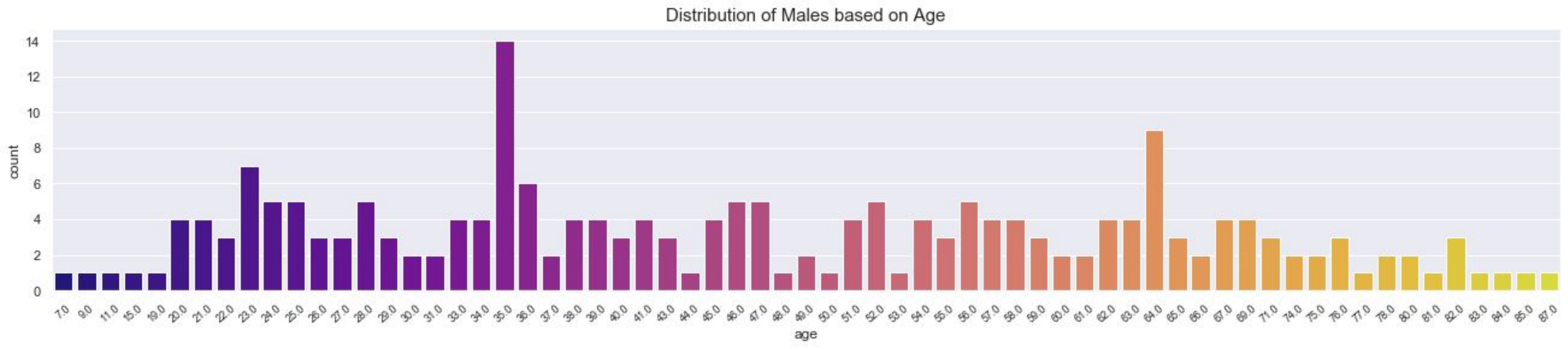


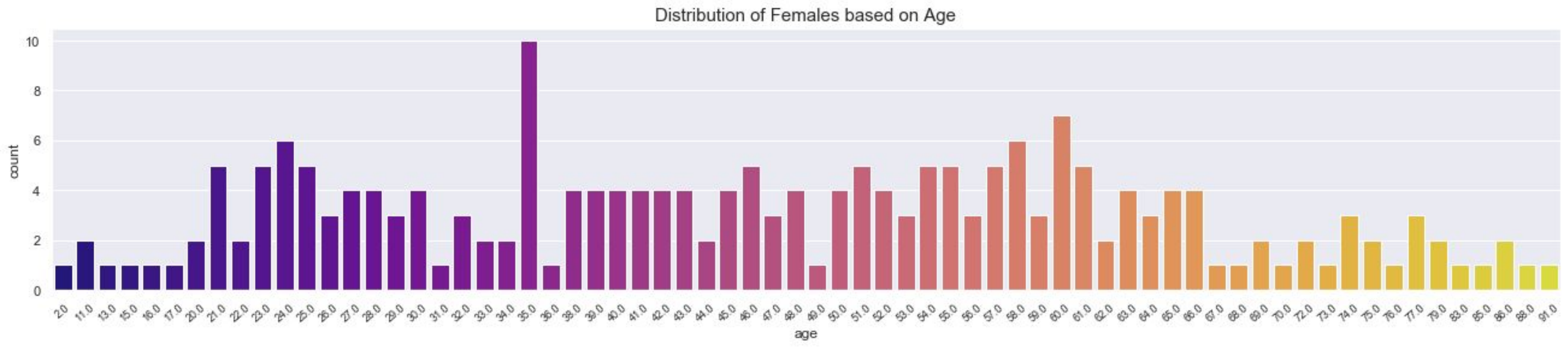
*7000 observations

Current State of Patients by Age

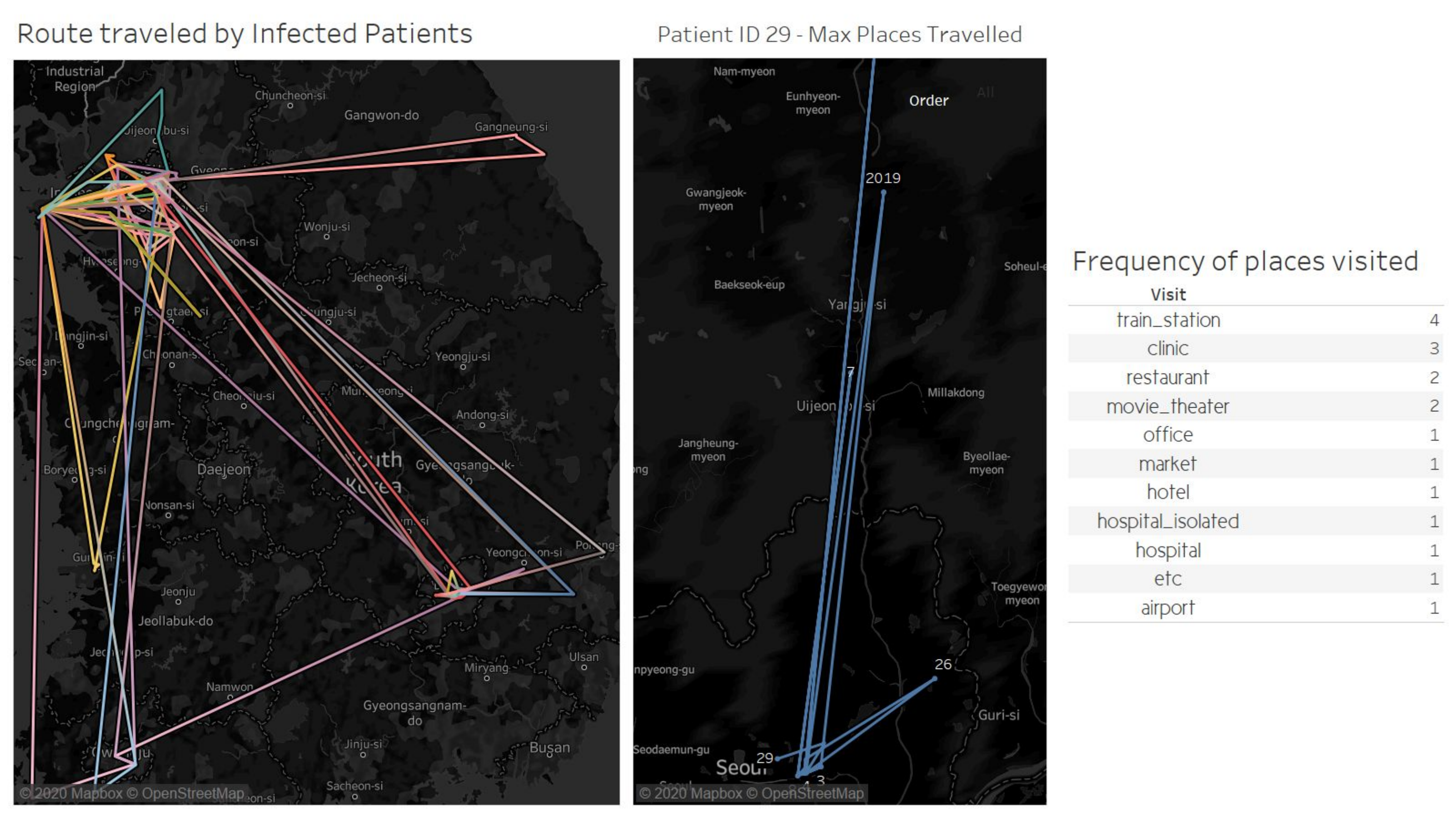


Overview of Infected Patients





Route Travelled by 56 Infected Patients



Route traveled by Infected Patients

Patient ID 29 - Max Places Travelled

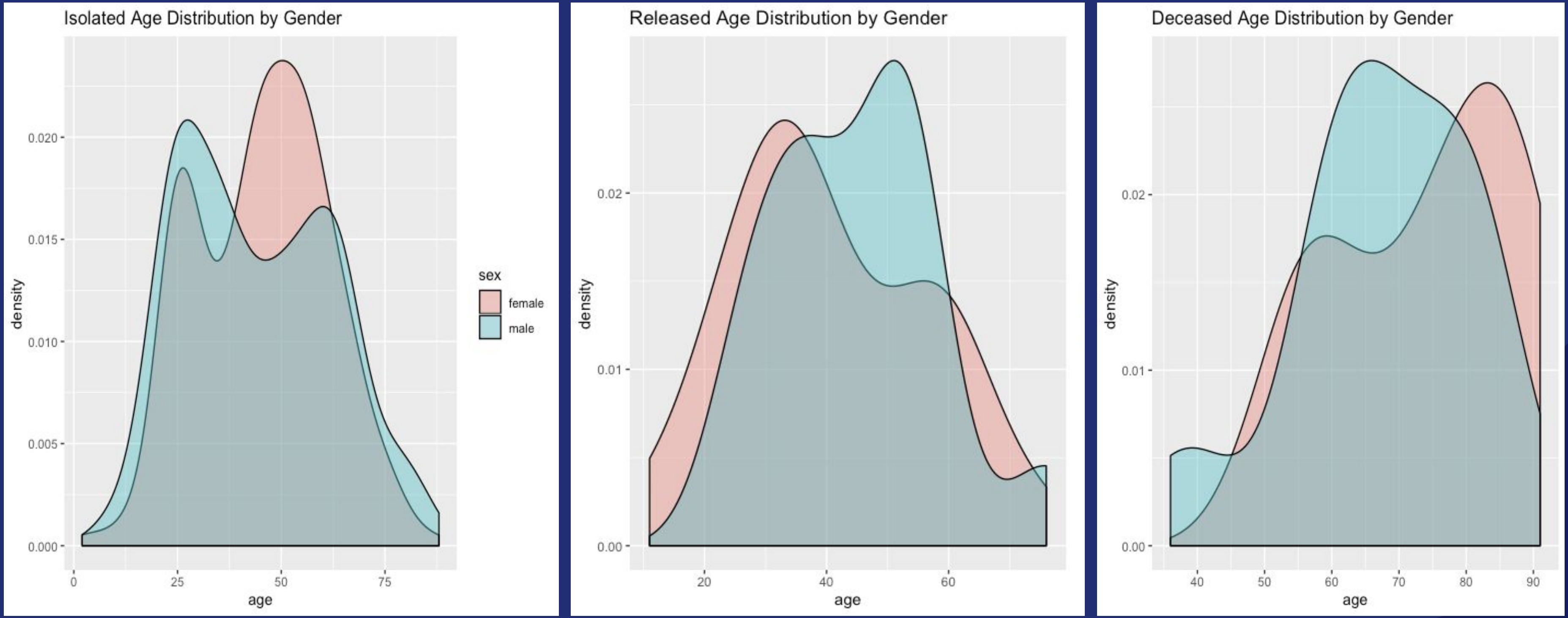
Frequency of places visited

Visit	
train_station	4
clinic	3
restaurant	2
movie_theater	2
office	1
market	1
hotel	1
hospital_isolated	1
hospital	1
etc	1
airport	1

03

STATISTICAL ANALYSIS

Density Graph: Age Densities by Gender and State of Patients



- Different age ranges have fluctuations in densities for female and male patients
- Relatively normal distribution of those who died for males with a mean of ~67 years old and a st. dev of 13.28

T-Test

```
> t.test(age ~ sex, data =
isolated_df)

        Welch Two Sample t-test

data:  age by sex
t = 0.8862, df = 418.4, p-value =
0.376
alternative hypothesis: true
difference in means is not equal to
0
95 percent confidence interval:
 -1.613900  4.263837
sample estimates:
mean in group female   mean in
group male
          45.35693           44.03196
```

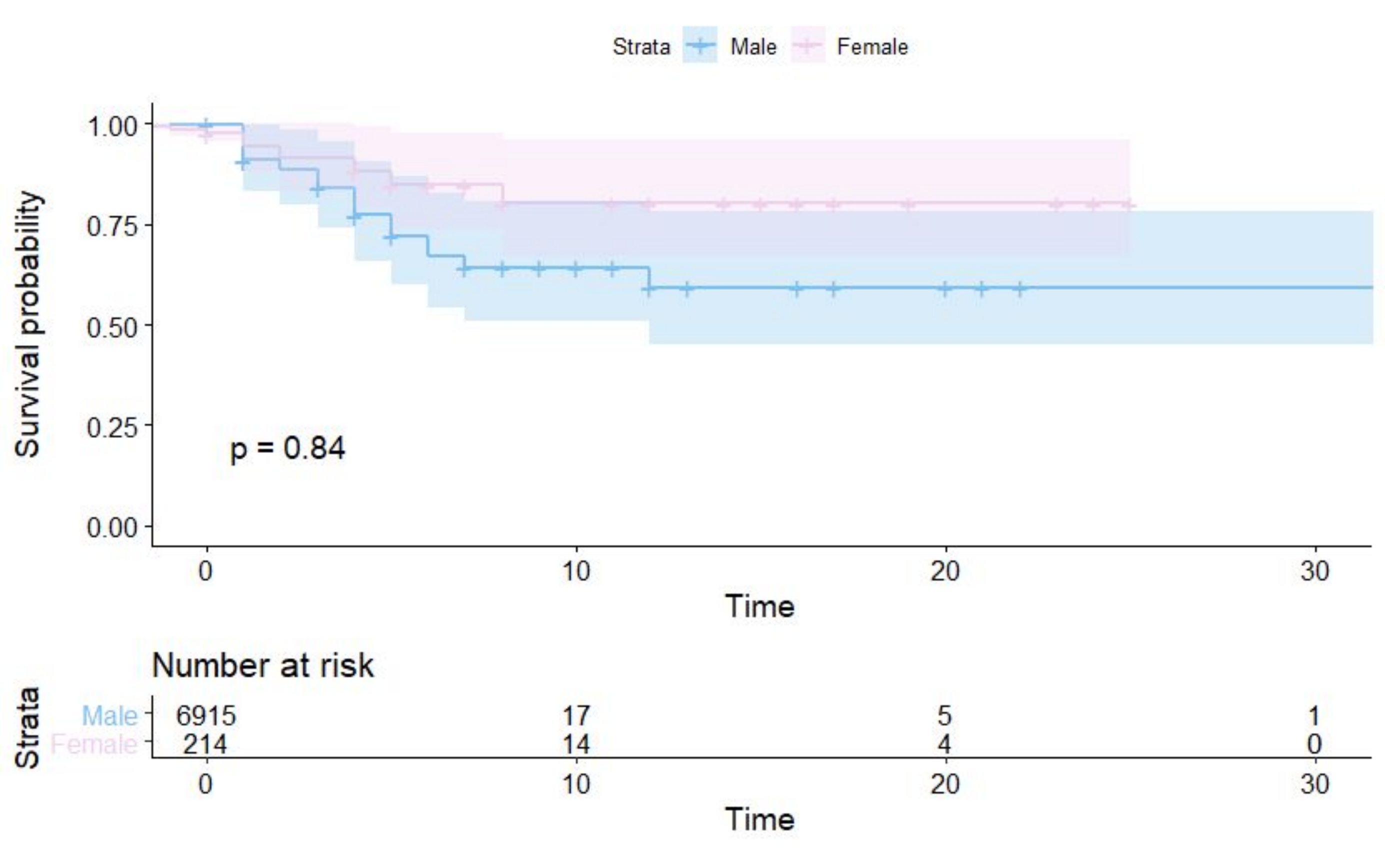
```
> t.test(age ~ sex, data =
released_df)

        Welch Two Sample t-test

data:  age by sex
t = -1.0726, df = 49.307, p-value =
0.2887
alternative hypothesis: true
difference in means is not equal to 0
95 percent confidence interval:
 -12.289713   3.735336
sample estimates:
mean in group female   mean in
group male
          40.65385           44.93103
```

```
> t.test(age ~ sex, data =
deceased_df)

        Welch Two Sample t-test

data:  age by sex
t = 1.2634, df = 17.502, p-value =
0.223
alternative hypothesis: true
difference in means is not equal to 0
95 percent confidence interval:
 -4.349813 17.406955
sample estimates:
mean in group female   mean in
group male
          74.10000           67.57143
```

For each of these t-tests, using a significance level of 0.05, we fail to reject the null and conclude that the true difference in means is equal to 0

Cox Hazard Rate Model : Survival Curve for Gender

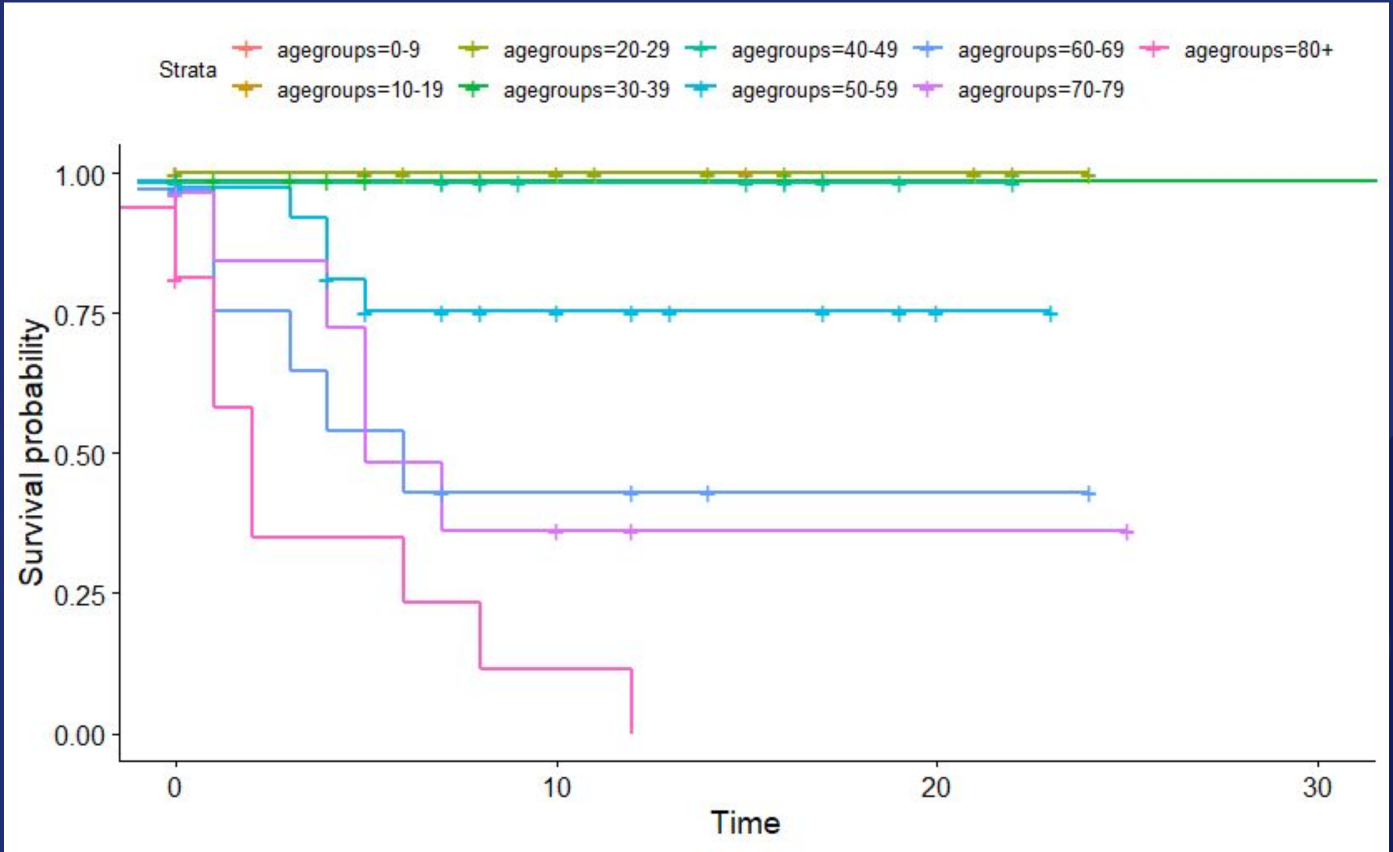
Sex is not statistically significant - no major difference in terms of survival rate for gender



Cox Hazard Rate Model : Survival Curve for Age Groups

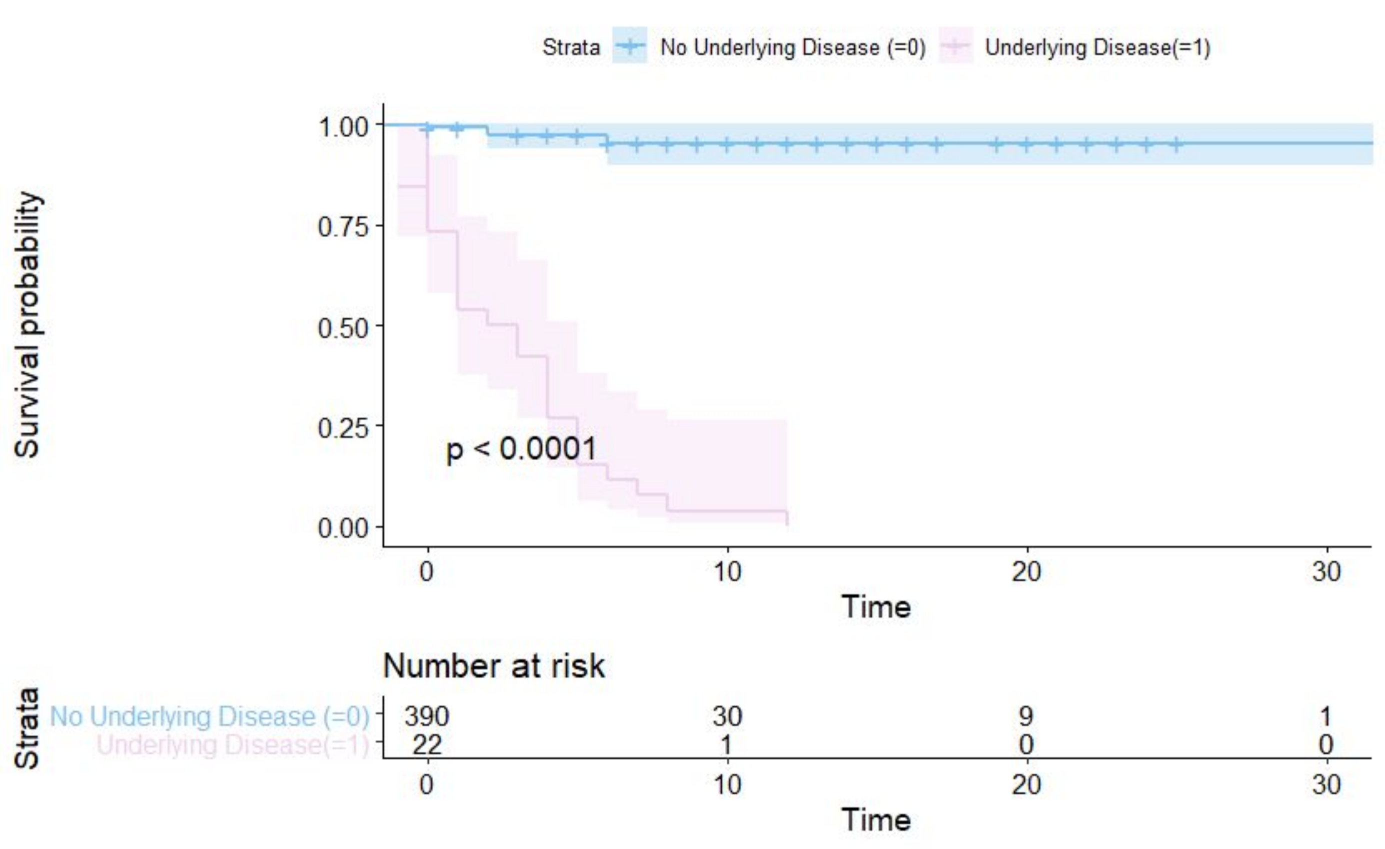
Age wasn't statistically significant

But, survival analysis graph shows older age survival rate is low compared to populations in younger age groups



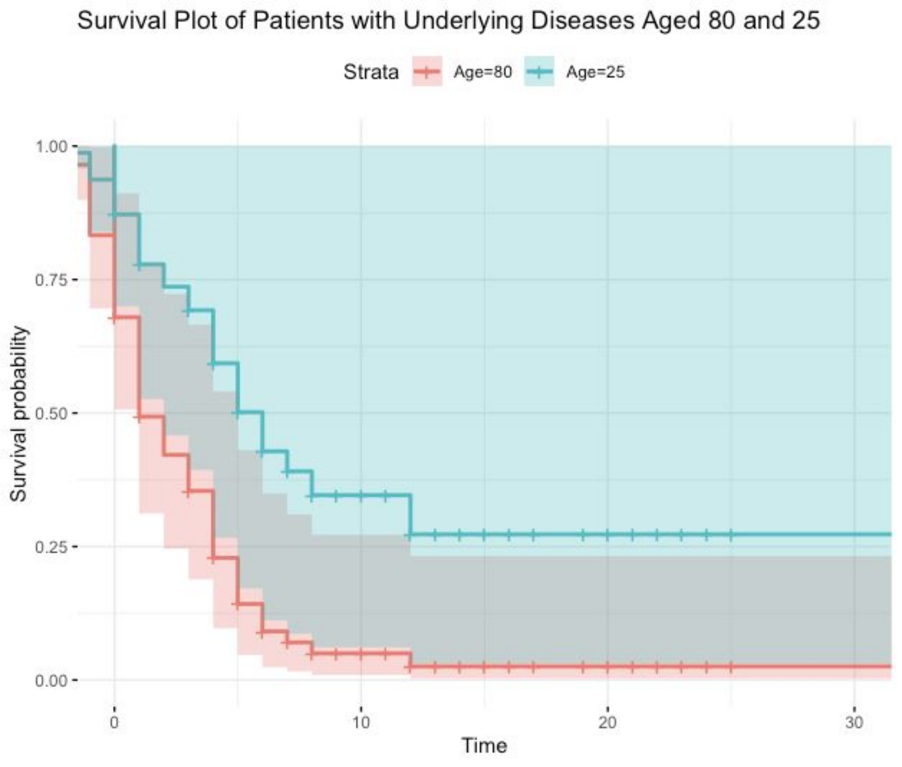
Cox Hazard Rate Model : Survival Curve for Underlying Diseases

Having an underlying disease is statistically significant
This indicates a **strong** relationship between patients' underlying disease and increased risk of death.

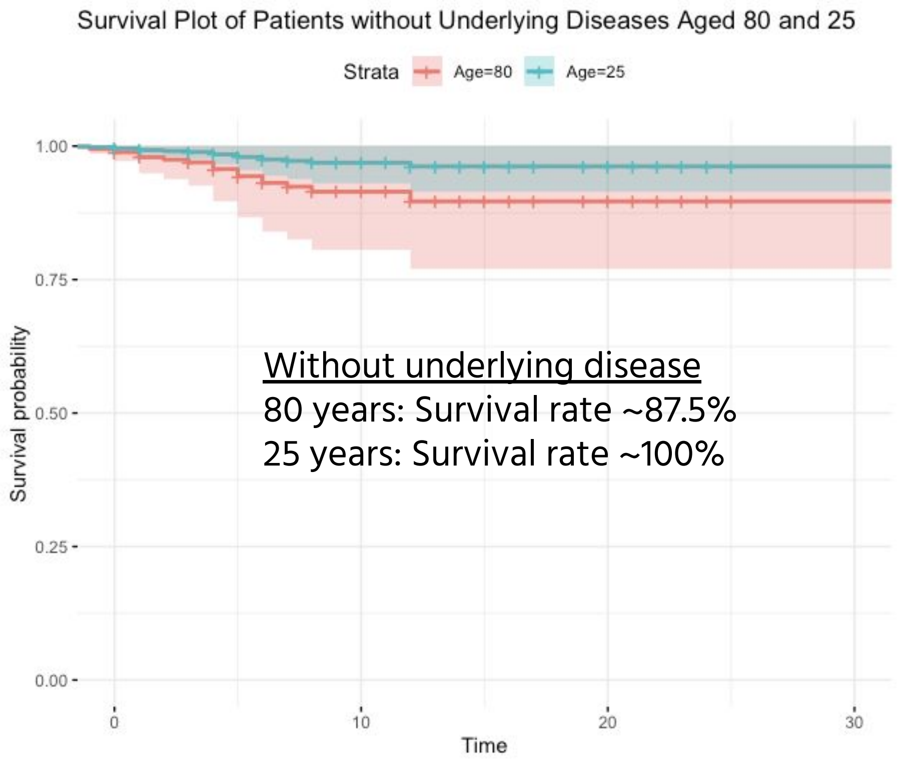


Survival Curves by Age

Those aged 80 with an underlying disease have a very low survival rate of almost 0 after 12 days



Survival Plot of Patients with Underlying Diseases Aged 80 and 25



Survival Plot of Patients without Underlying Diseases Aged 80 and 25

Without underlying disease

80 years: Survival rate ~87.5%

25 years: Survival rate ~100%

04

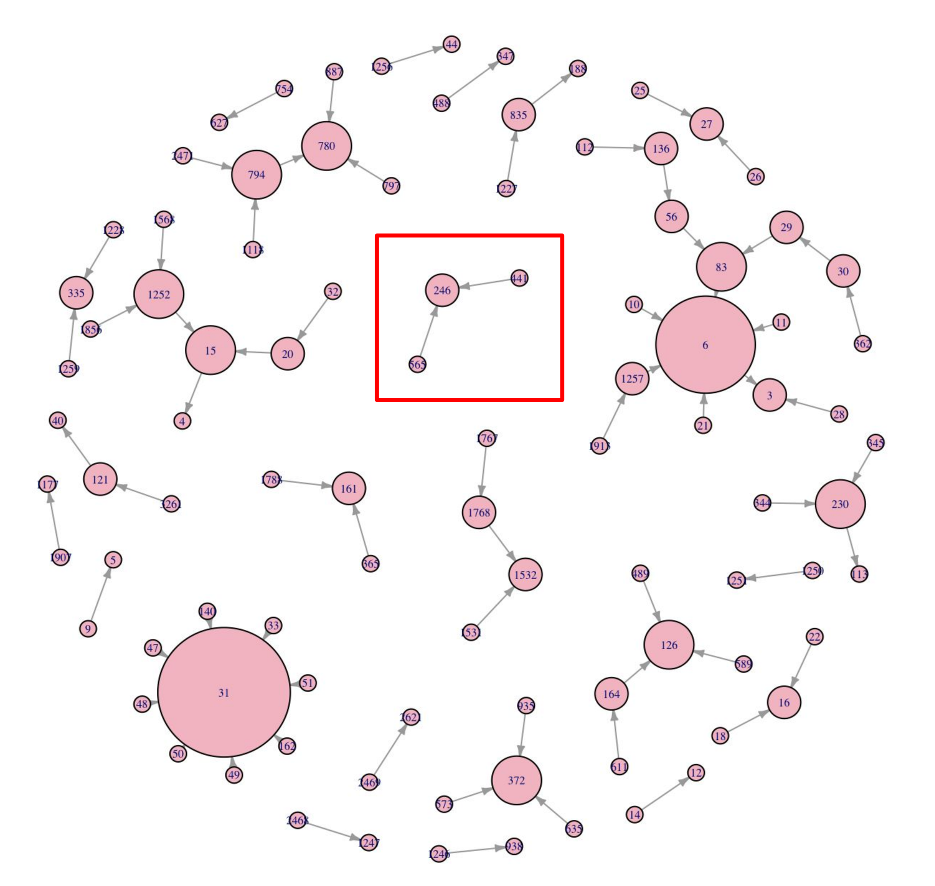
NETWORK ANALYSIS

Network Analysis Graph

node = patient id

- Arrowhead points to node of who the patient was infected by

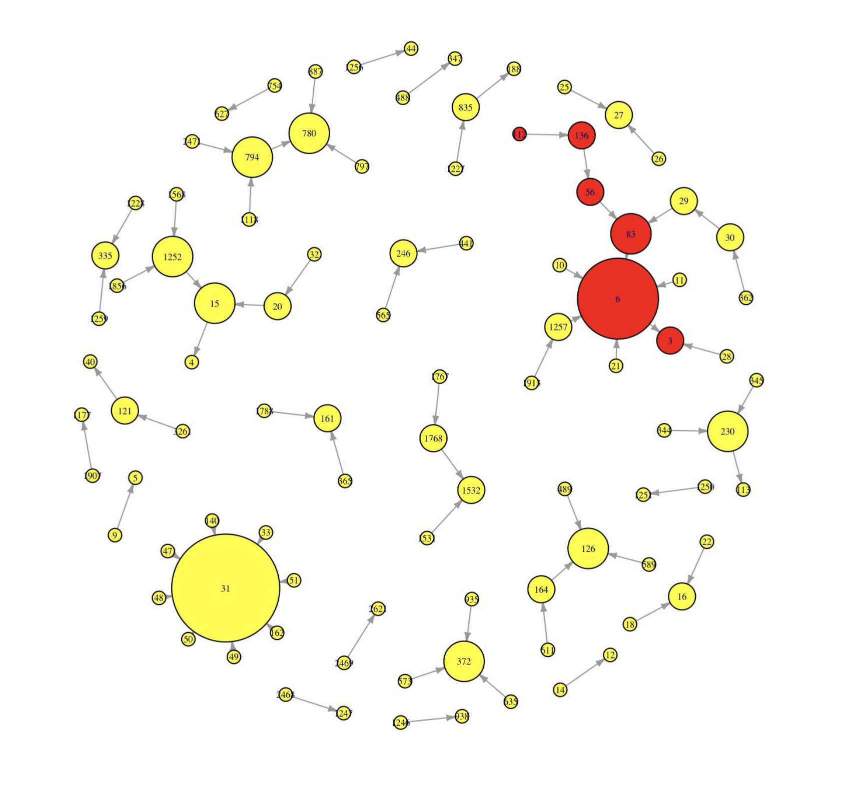
- For example, in the red box, patient 565 and 141 were both infected by patient 246.



Diameter

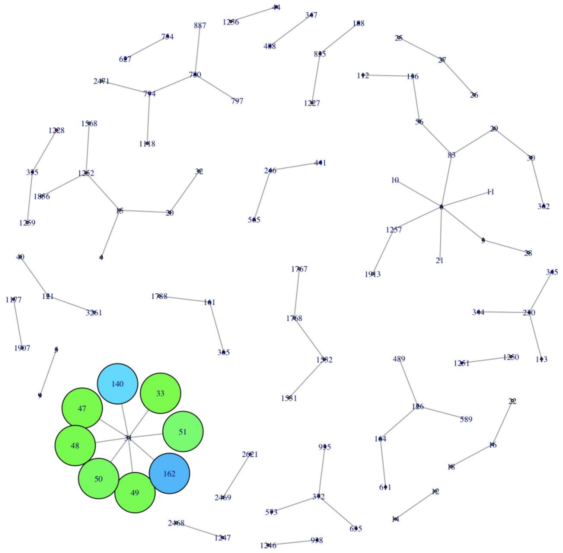
Diameter is the longest distance between two vertices

- diameter = 5
- infection could reach 5 steps

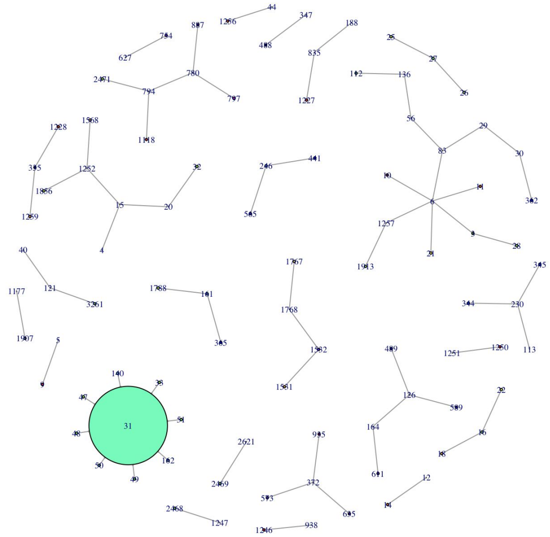


Hubs and Authority

Hubs



Authority

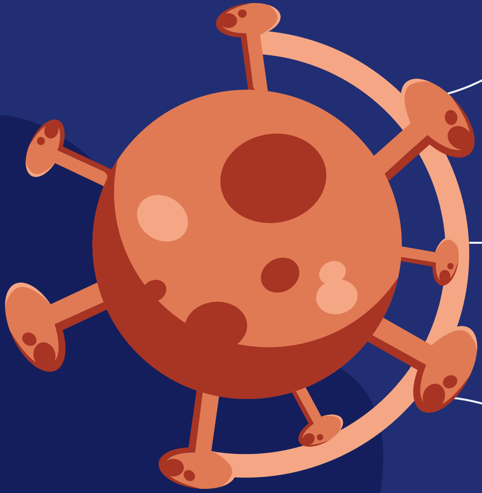


Patient 31 → infected most people

05

CONCLUSION

Key Insights



01

Age

People of all age ranges can contract the Coronavirus, but the elderly have a lower chance of surviving

02

Underlying Diseases

Patients with underlying diseases have a lower chance of surviving the Coronavirus

03

Gender

There are no statistically significant differences in survival rate between males and females who contracted the Coronavirus

Limitations

01

Missing Values

Affected network graph and Hazard model

02

Insights are Largely Based Off of South Korean Dataset

Other countries might yield different results

03

Data Might Not Reflect True Value

Different countries have different methods of testing for COVID-19, which might affect reporting