Subject: Querying, data presentation, data visualization

FINAL PROJECT

Topic: Factors affecting the mortality rate of COVID-19 patients

Lecturer: Dr. Jarosław Olejniczak

Student: Ky Anh Le

Student ID: 140044

I. Topic

1. Selection of the topic justification

The COVID-19 pandemic is a global event that has profoundly impacted all aspects of life, from public health to the economy and society. Analyzing data on COVID-19 not only helps us better understand how the pandemic spreads, the factors affecting infection and mortality rates, but also provides a foundation for devising effective prevention and management strategies in the future.

Additionally, this topic allows for the exploration and application of advanced data analysis tools while raising awareness of the importance of data-driven decision-making. Therefore, the topic is not only urgent but also has high practical value.

2. Project objective

The objective of this project is to analyze the factors influencing the mortality rate among COVID-19 patients. By identifying key determinants such as demographic characteristics, pre-existing health conditions, medical interventions, and other relevant factors, this analysis aims to provide insights into the underlying causes of fatal outcomes. The results can be used to support healthcare strategies, improve patient management, and minimize mortality rates in future pandemic responses.

II. Dataset

1. Source

The dataset was provided by the Mexican government: https://datos.gob.mx/busca/dataset/informacion-referente-a-casos-covid-19-en-mexico and downloaded from the Kaggle platform.

2. Data summary

The dataset contains 1048575 rows, 21 total columns:

- sex: 1 for female and 2 for male.
- age: of the patient.
- classification: covid test findings. Values 1-3 mean that the patient was diagnosed with covid in different degrees. 4 or higher means that the patient is not a carrier of covid or that the test is inconclusive.
- patient type: type of care the patient received in the unit. 1 for returned home and 2 for hospitalization.
- pneumonia: whether the patient already have air sacs inflammation or not.
- pregnancy: whether the patient is pregnant or not.
- diabetes: whether the patient has diabetes or not.
- copd: indicates whether the patient has Chronic obstructive pulmonary disease or not.
- asthma: whether the patient has asthma or not.
- inmsupr: whether the patient is immunosuppressed or not.
- hypertension: whether the patient has hypertension or not.
- cardiovascular: whether the patient has heart or blood vessels related disease.
- renal chronic: whether the patient has chronic renal disease or not.
- other disease: whether the patient has other disease or not.
- obesity: whether the patient is obese or not.
- tobacco: whether the patient is a tobacco user.
- usmer: indicates whether the patient treated medical units of the first, second or third level.
- medical unit: type of institution of the National Health System that provided the care.
- intubed: whether the patient was connected to the ventilator.
- icu: indicates whether the patient had been admitted to an Intensive Care Unit.
- date died: If the patient died indicate the date of death, and 9999-99-99 otherwise.

Note that in the Boolean features, 1 means "yes" and 2 means "no", values as 97 and 99 are missing data.

```
     USMER        MEDICAL_UNIT          SEX          PATIENT_TYPE    DATE_DIED
 Min.   :1.000   Min.   : 1.000   Min.   :1.000   Min.   :1.000   Length:1048575
 1st Qu.:1.000   1st Qu.: 4.000   1st Qu.:1.000   1st Qu.:1.000   Class :character
 Median :2.000   Median :12.000   Median :1.000   Median :1.000   Mode  :character
 Mean   :1.632   Mean   : 8.981   Mean   :1.499   Mean   :1.191
 3rd Qu.:2.000   3rd Qu.:12.000   3rd Qu.:2.000   3rd Qu.:1.000
 Max.   :2.000   Max.   :13.000   Max.   :2.000   Max.   :2.000
    INTUBED        PNEUMONIA           AGE            PREGNANT        DIABETES
 Min.   : 1.00   Min.   : 1.000   Min.   :  0.00   Min.   : 1.00   Min.   : 1.000
 1st Qu.:97.00   1st Qu.: 2.000   1st Qu.: 30.00   1st Qu.: 2.00   1st Qu.: 2.000
 Median :97.00   Median : 2.000   Median : 40.00   Median :97.00   Median : 2.000
 Mean   :79.52   Mean   : 3.347   Mean   : 41.79   Mean   :49.77   Mean   : 2.186
 3rd Qu.:97.00   3rd Qu.: 2.000   3rd Qu.: 53.00   3rd Qu.:97.00   3rd Qu.: 2.000
 Max.   :99.00   Max.   :99.000   Max.   :121.00   Max.   :98.00   Max.   :98.000
      COPD            ASTHMA          INMSUPR        HIPERTENSION    OTHER_DISEASE
 Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
 1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.: 2.000
 Median : 2.000   Median : 2.000   Median : 2.000   Median : 2.000   Median : 2.000
 Mean   : 2.261   Mean   : 2.243   Mean   : 2.298   Mean   : 2.129   Mean   : 2.435
 3rd Qu.: 2.000   3rd Qu.: 2.000   3rd Qu.: 2.000   3rd Qu.: 2.000   3rd Qu.: 2.000
 Max.   :98.000   Max.   :98.000   Max.   :98.000   Max.   :98.000   Max.   :98.000
 CARDIOVASCULAR      OBESITY       RENAL_CHRONIC       TOBACCO       CLASIFFICATION_FINAL
 Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   :1.000
 1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.:3.000
 Median : 2.000   Median : 2.000   Median : 2.000   Median : 2.000   Median :6.000
 Mean   : 2.262   Mean   : 2.125   Mean   : 2.257   Mean   : 2.214   Mean   :5.306
 3rd Qu.: 2.000   3rd Qu.: 2.000   3rd Qu.: 2.000   3rd Qu.: 2.000   3rd Qu.:7.000
 Max.   :98.000   Max.   :98.000   Max.   :98.000   Max.   :98.000   Max.   :7.000
      ICU
 Min.   : 1.00
 1st Qu.:97.00
 Median :97.00
 Mean   :79.55
 3rd Qu.:97.00
 Max.   :99.00
```

III. Data Processing

- Convert data type of variables
- Convert “sex” variable from 1, 2 to Woman, Man
- Convert “patient_type” variable from 1, 2 to “returned home”, “hospitalization”
- Convert “classification_final” variable from number to “Positive”, “Negative”
- Convert boolean variables from 1, 2 to Yes, No
- Convert missing data from 97, 99 to NA
- Impute missing data
- Drop “intubed”, “icu” variables as they contains more than 85% missing data
- Create “survive” column from date_died column: if the date-died is 9999-99-99 then “Yes”, else “No”

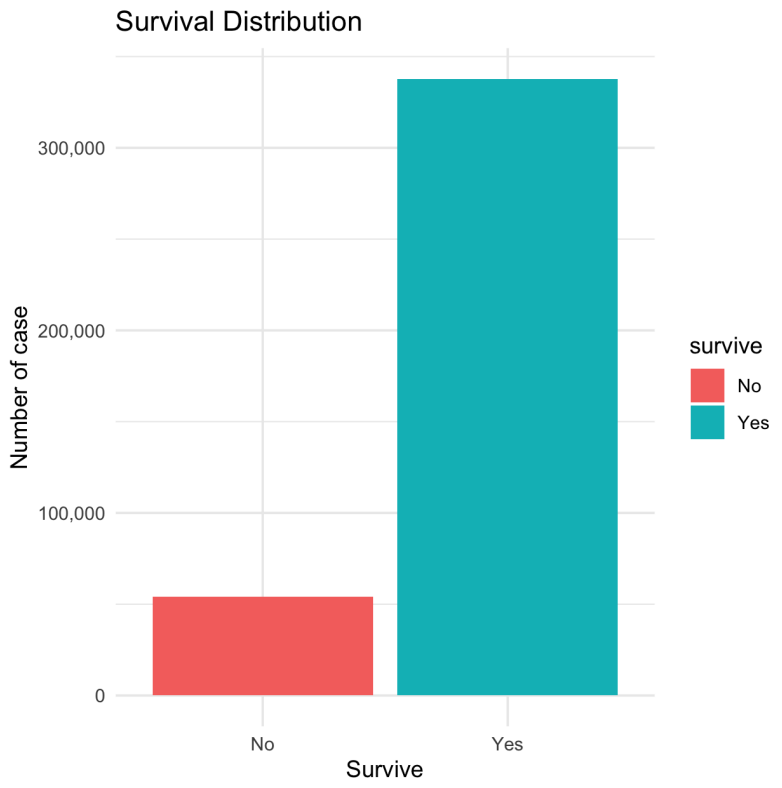
- Create “comorbidities” column from disease columns
- Create “num_comorbidities” column
- Create “high-risk” column: if the patient is older than 60 or has more than 2 comorbidities then “Yes”, else “No”
- Drop unnecessary columns: Some columns such as “usmer”, “medical_unit”, “date_died” are not related to the project’s objective
- Filter dataset with classification_final is “Positive”

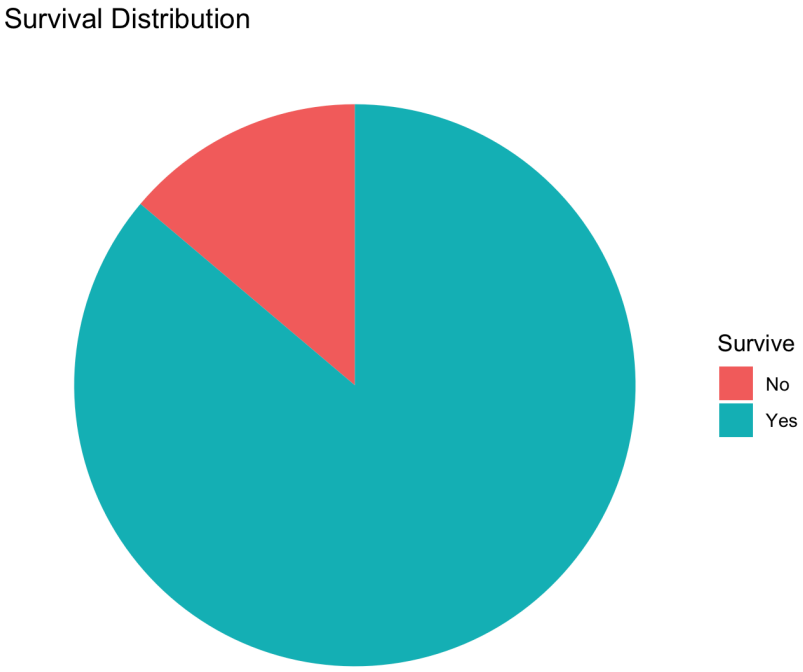
After these data processing steps, we have the dataset contains 391979 and 20 variables

```
> colnames(df)
 [1] "sex"                 "patient_type"        "pneumonia"
 [4] "age"                 "pregnant"            "diabetes"
 [7] "copd"                "asthma"              "inmsupr"
[10] "hipertension"        "other_disease"       "cardiovascular"
[13] "obesity"             "renal_chronic"       "tobacco"
[16] "clasiffication_final" "survive"            "comorbidities"
[19] "num_comorbidities"   "high_risk"
```

IV. Data Analysis

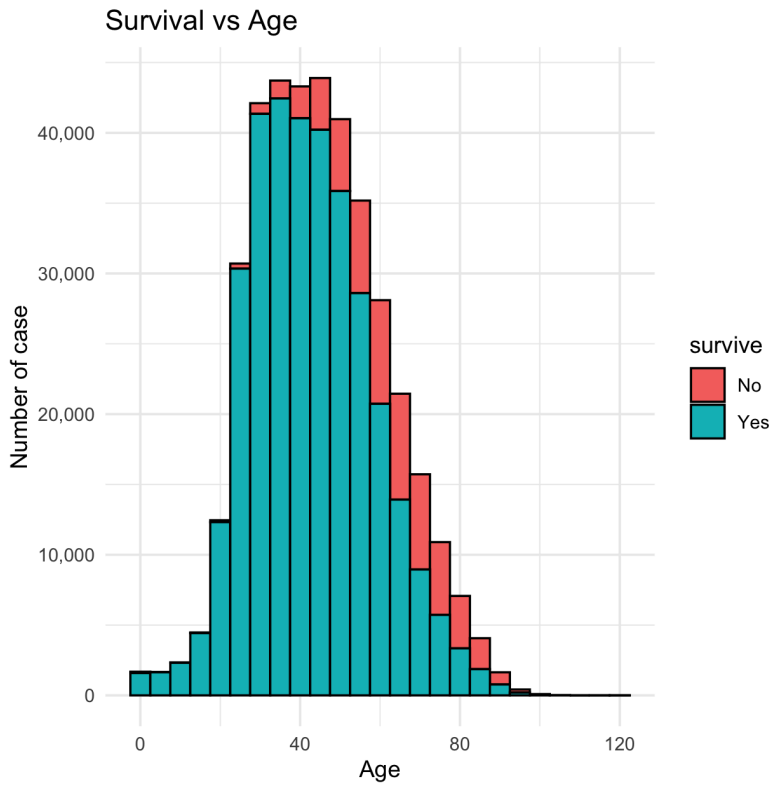
1. Death rate





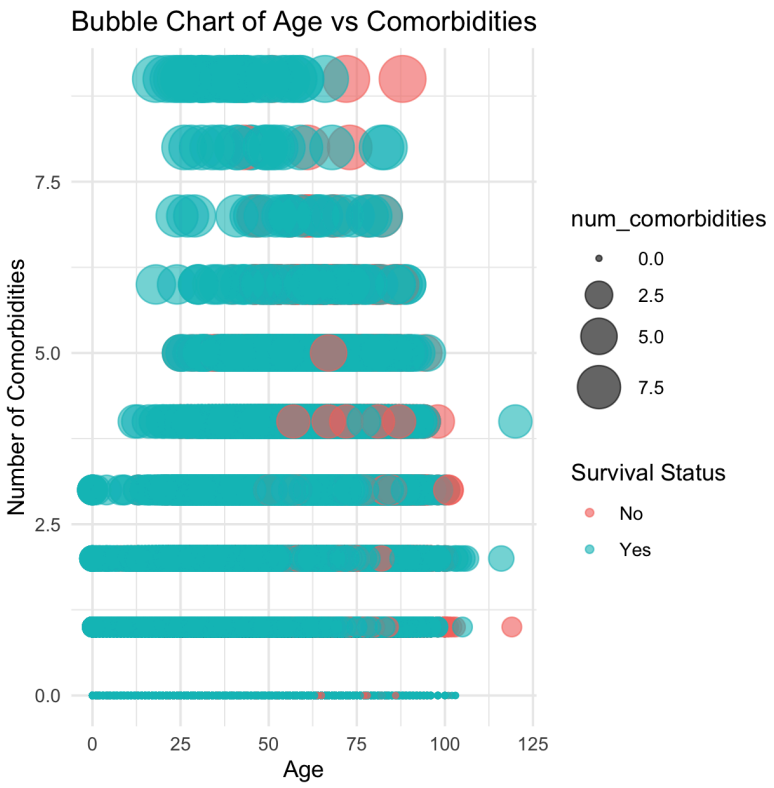
The death rate for COVID-19 positive cases is about 15%

2. Death rate by Age group

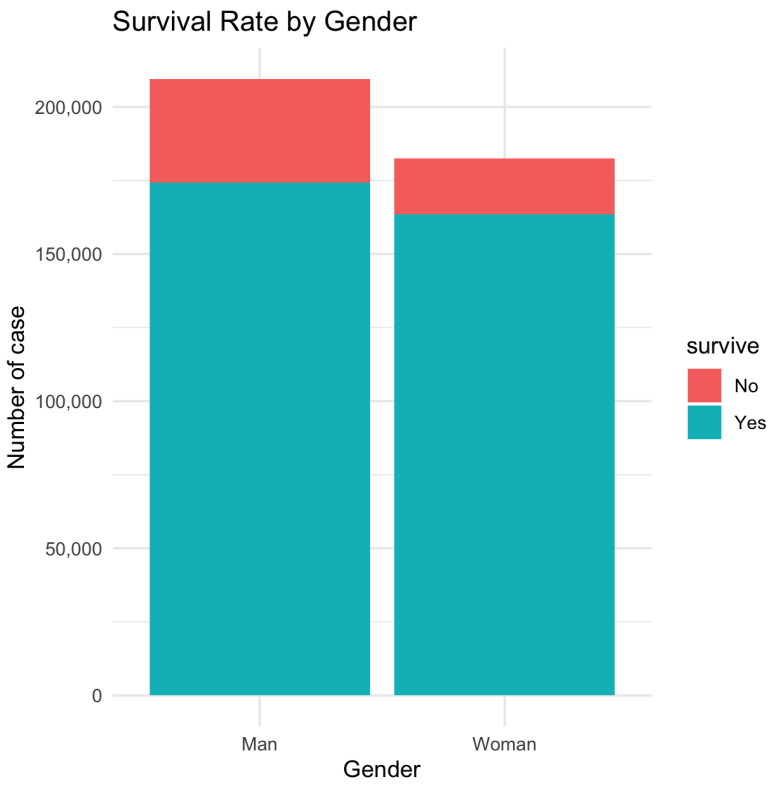


Death rate gradually increases with age and is especially high in the elderly group (> 60 years old).

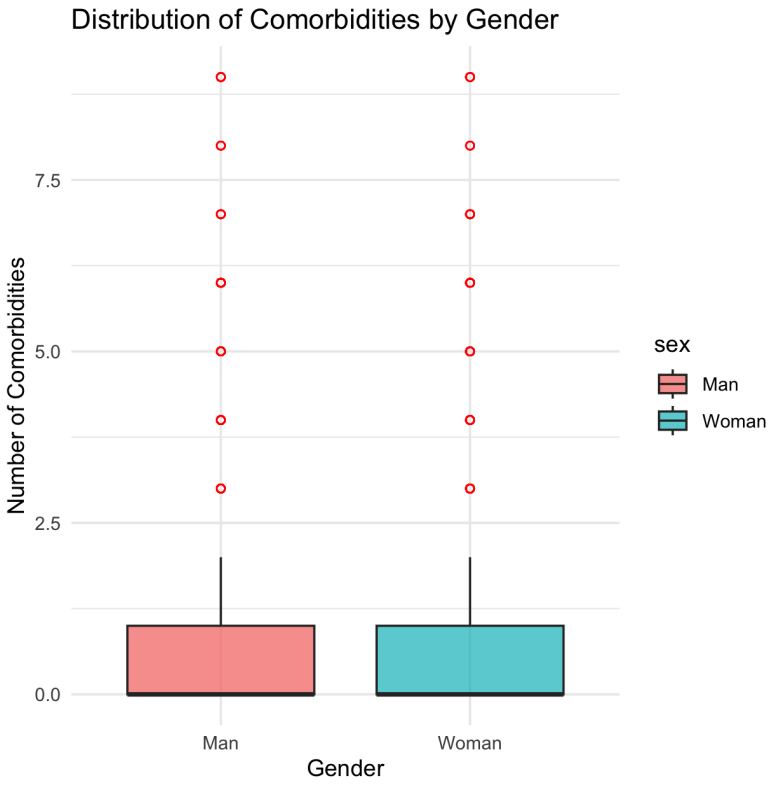
This is understandable because at this age, people's resistance also gradually decreases

	Groups with particularly high mortality rates are often found in groups with multiple underlying health conditions and the elderly

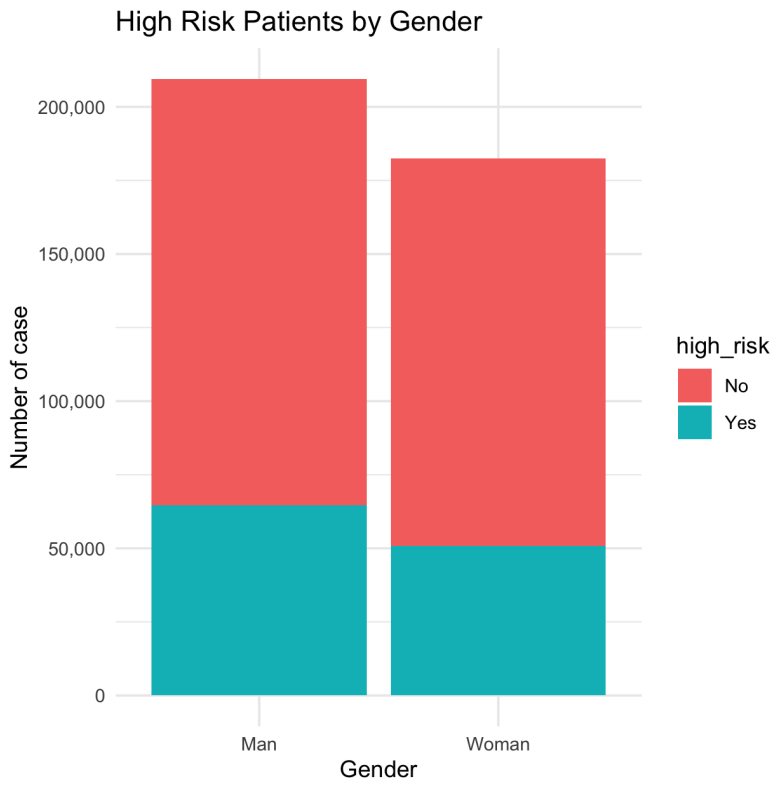
3. Death rate by Gender



The number of positive COVID-19 cases as well as the death rate in men are higher than in women

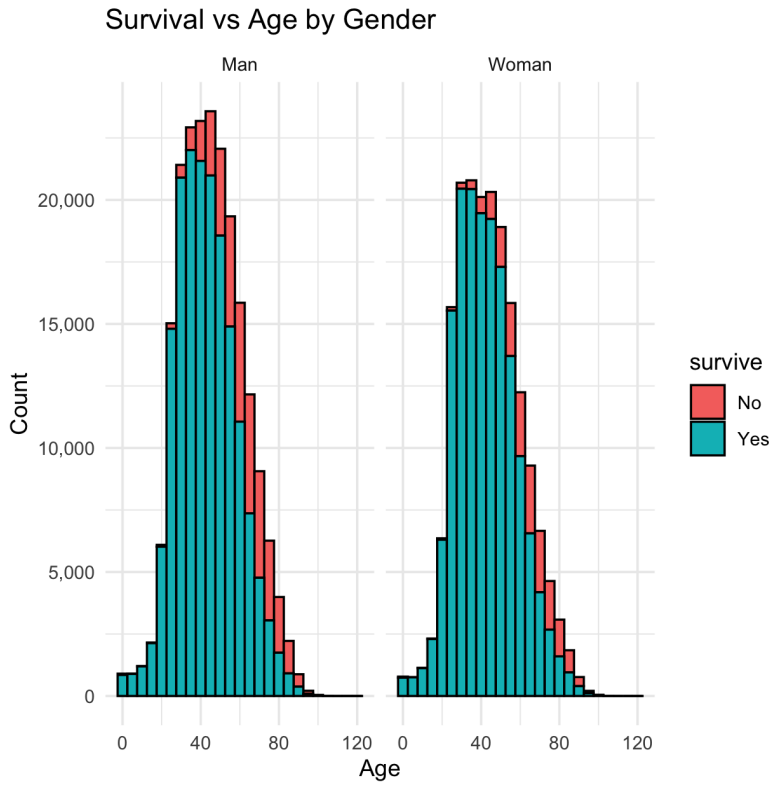


There is no difference between the number of underlying diseases in men compared to women



The number of dangerous cases is higher for men than for women

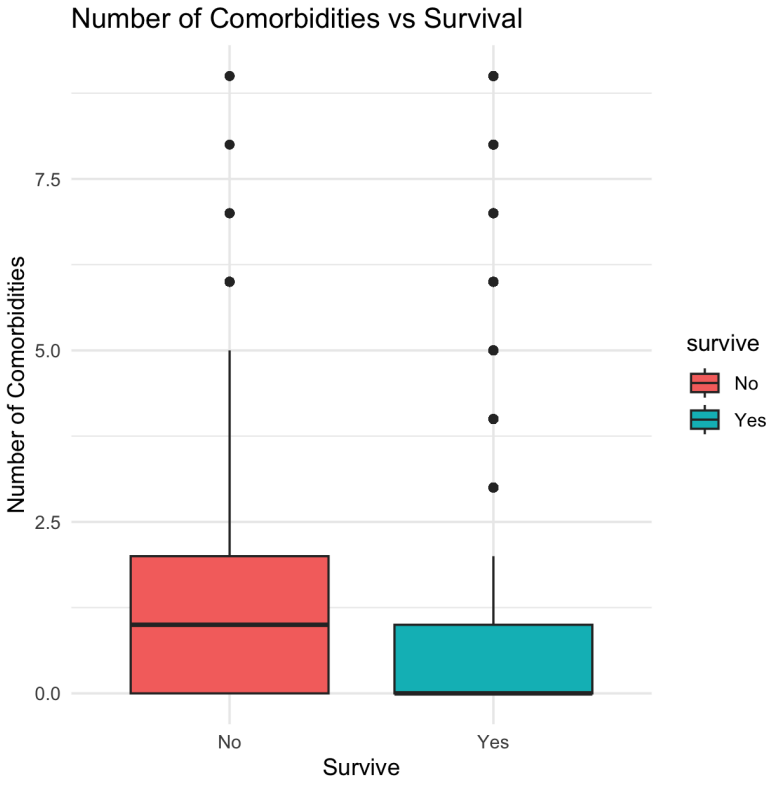
This partly explains why the death rate for COVID-19 patients is higher in men



Overall, trends in mortality rates for both sexes by age group are relatively similar

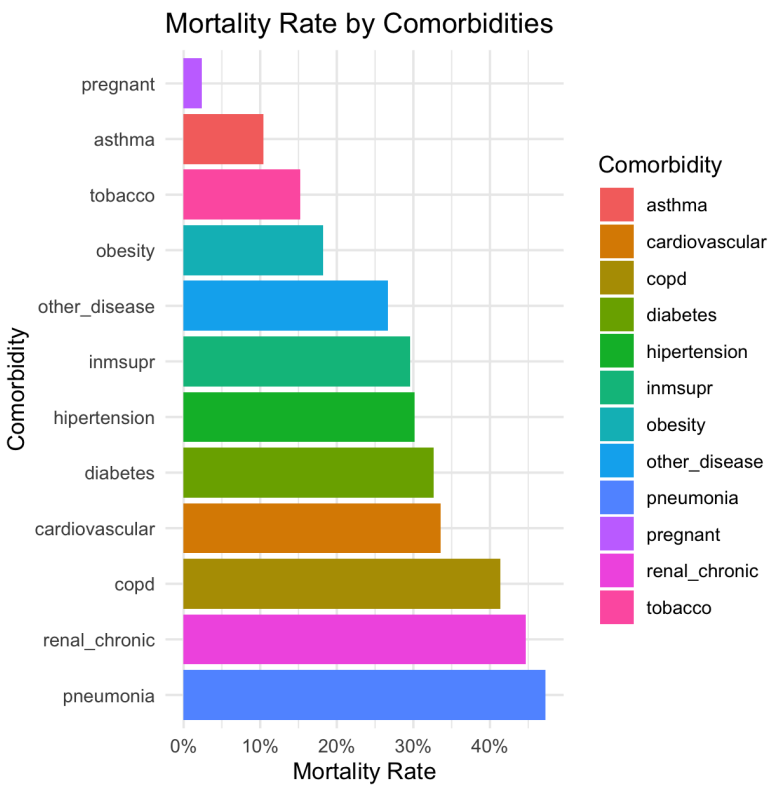
However, it can be seen that the mortality rate of the elderly male group is relatively high compared to this age group of women.

4. Death rate by comorbidity



Patients who died often had a higher number of underlying diseases.

There are also cases with no underlying disease but still died when positive for COVID-19 and vice versa, there are also cases with many underlying diseases but still cured of this disease.

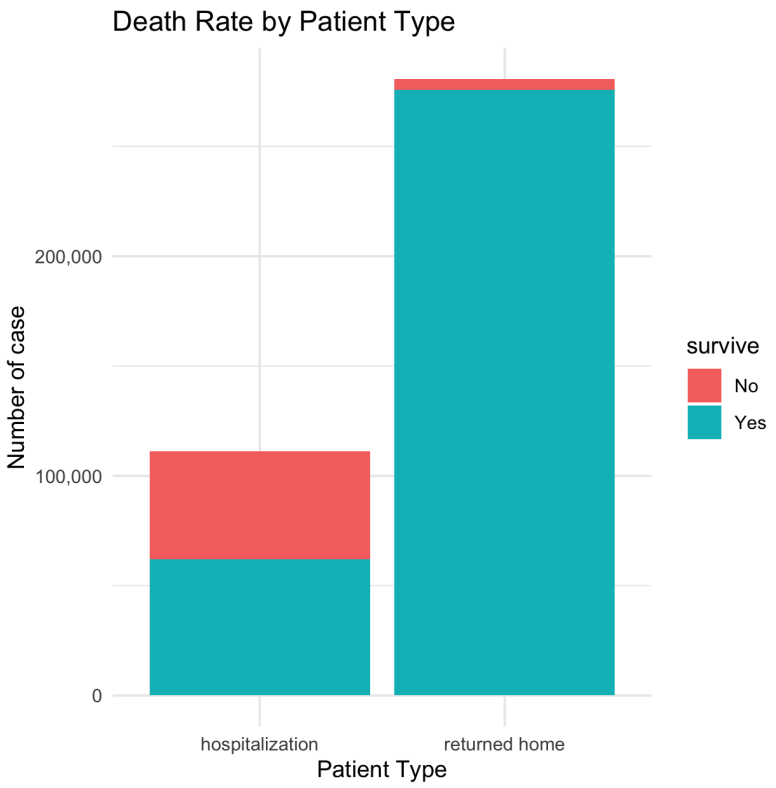


The underlying diseases with the highest mortality rate when patients get covid-19 are:

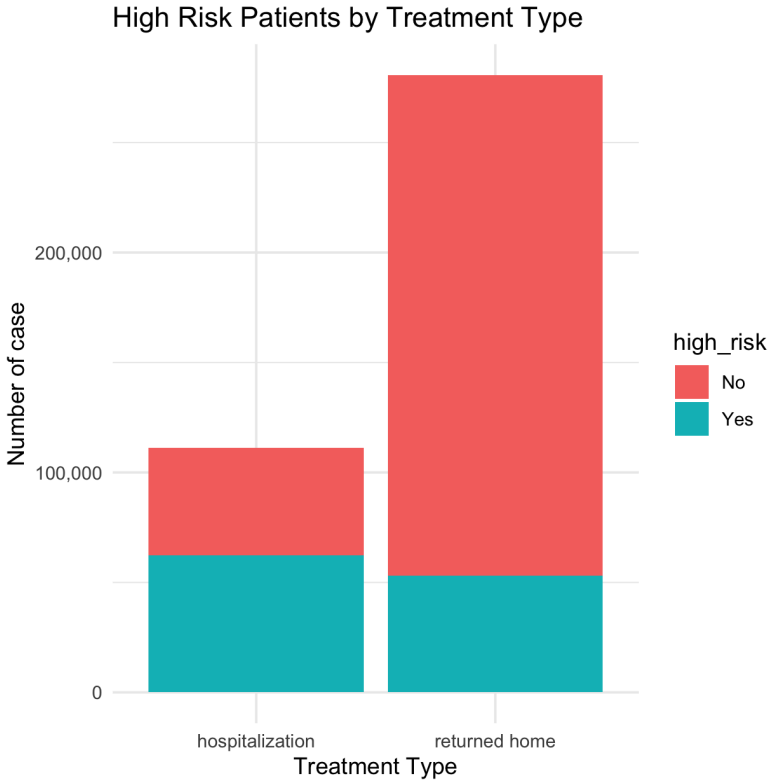
Pneumonia: ~47%
Renal chronic: ~ 45%
COPD: ~ 42%

There are many other underlying diseases with mortality rates ~ 30% and under

5. Death rate between hospitalized and non-hospitalized patients

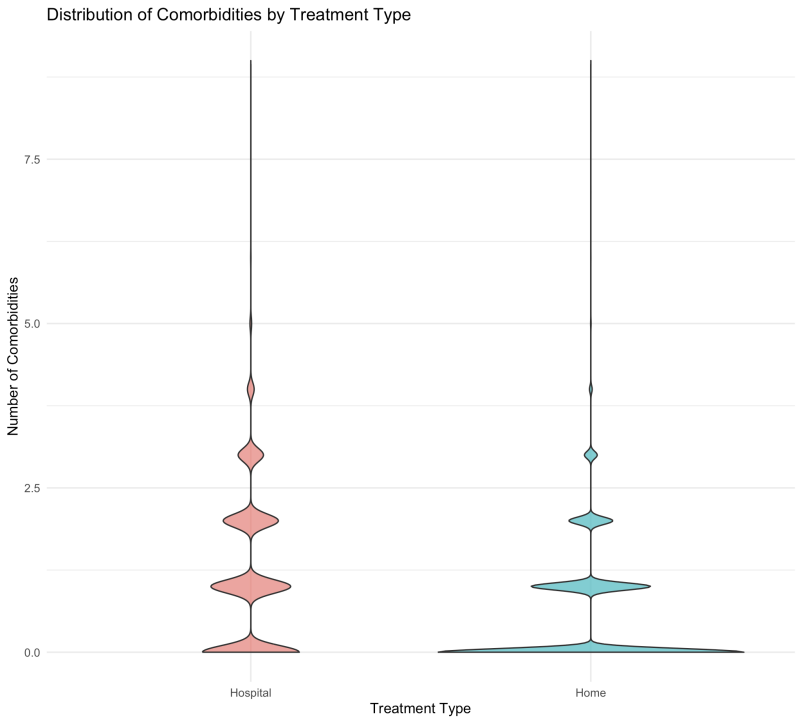


The number of patients treated at the hospital is only 1/3 of the number of patients treated at home. The death rate for patients with COVID-19 when treated in the hospital is also very high, at 40%.

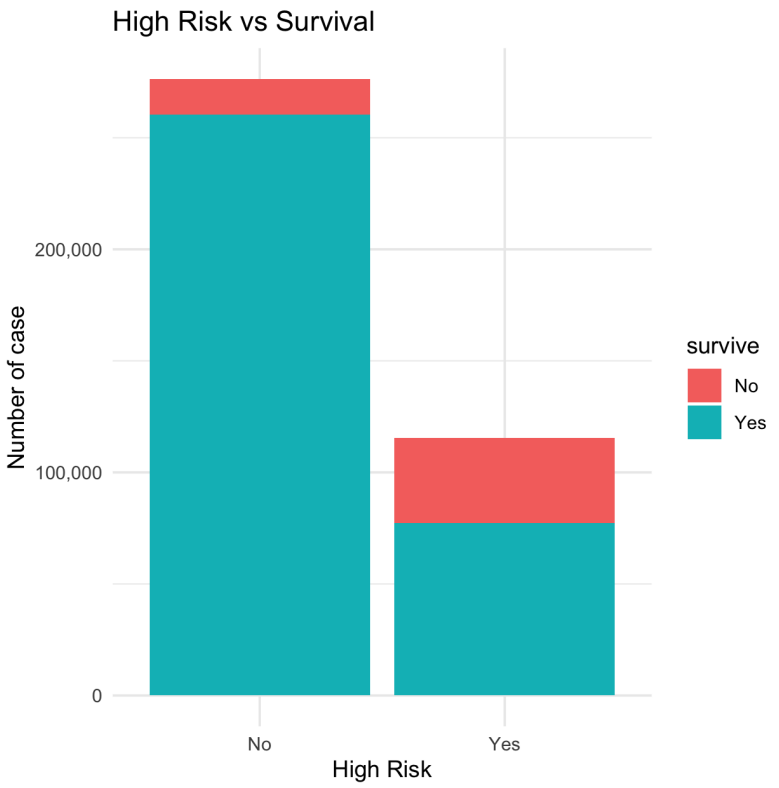


The number of high risk cases in both types of patients (hospital treatment and home treatment) is almost the same.

This shows that the context in hospitals during the epidemic is extremely overloaded when they cannot serve

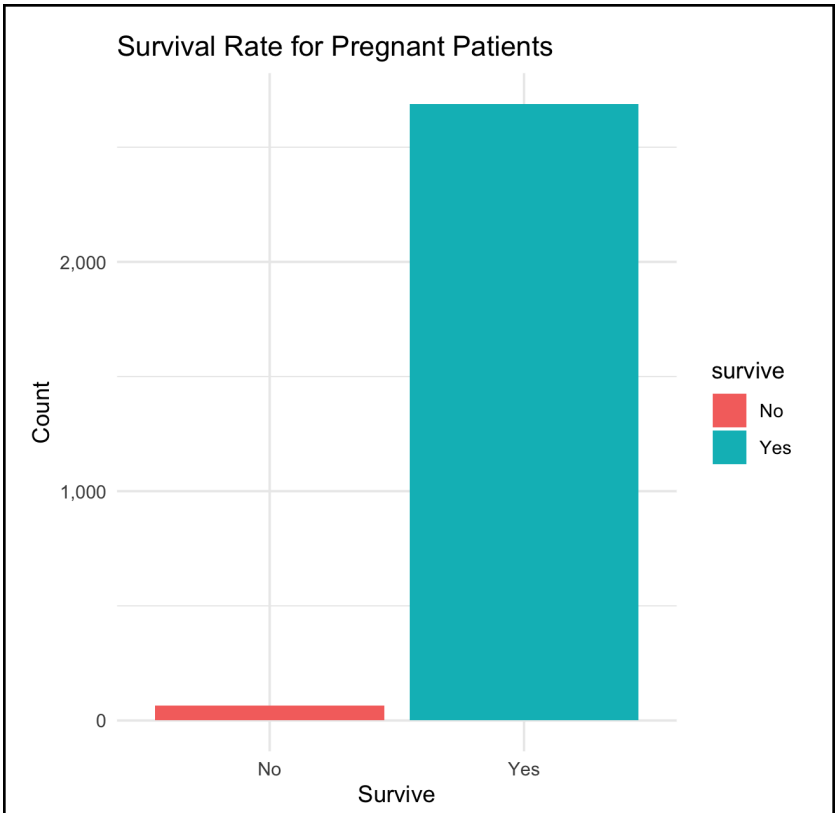
	all COVID-19 cases in general as well as high-risk cases in specific.
	Patients with a high number of underlying diseases are often in the dangerous group and are often given priority for treatment at the hospital

6. Death rate by risk



The mortality rate in high risk cases is relatively high, at ~ 27%.

7. Death rate for pregnant patient



Females who are positive with COVID-19 during pregnancy are not in a very dangerous situation as the death rate during pregnancy is not really high.

V. Conclusion

The project has analyzed the factors influencing the mortality rate among COVID-19 patients, providing key insights to support management and reduce mortality during pandemics. Key findings include a gradual increase in death rates with age, particularly among the elderly; higher severity and mortality rates in men compared to women; and underlying conditions such as pneumonia, chronic kidney disease, and COPD contributing to the highest mortality rates when combined with COVID-19.

The analysis also highlights the healthcare system's overload during the pandemic, with a high death rate (~40%) among patients treated in hospitals and a significant mortality rate (~27%) in high-risk cases. Patients with multiple underlying conditions were often prioritized for hospital treatment, emphasizing the need for more effective healthcare resource management.

From these findings, healthcare strategies should focus on supporting elderly patients, individuals with multiple underlying conditions, and improving care conditions both at home and in hospitals. These measures will help mitigate the impacts of similar pandemics in the future.