SUPPORT 2 Dataset Overview

Coursework 2

Objectives

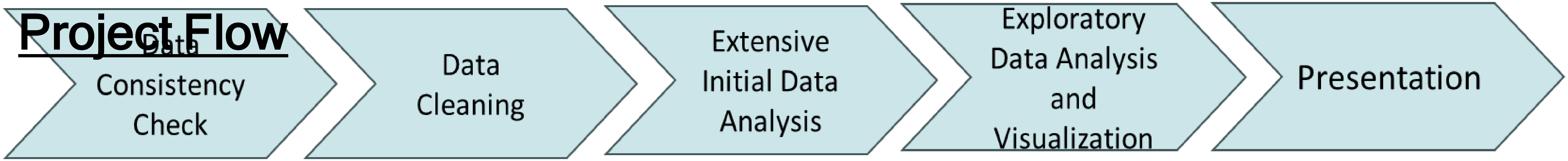
Perform an initial data analysis and exploratory data analysis using Python to derive meaningful insights from the health dataset.

Purpose of the Dataset

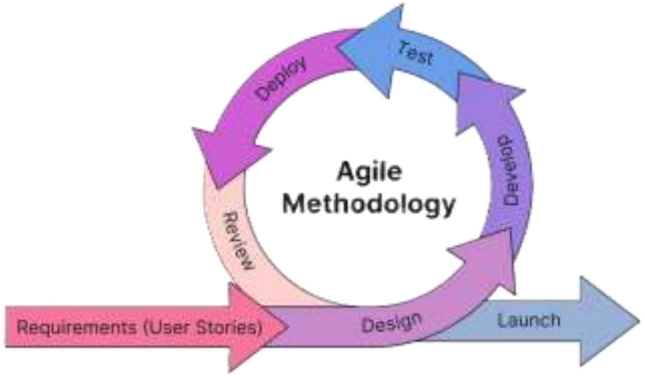
To create and test a model that predicts 180-day survival for seriously ill hospitalized adults (Phase I of SUPPORT) and compare its accuracy with an existing system and doctors' own predictions (Phase II of SUPPORT).

Dataset Link: [UCI Machine Learning Repository]

Project Flow



Methodology and Technology



pandas

matplotlib

NumPy

Microsoft Teams

WhatsApp

seaborn

Google colab

jupyter

plotly

Exploring and Cleaning the Dataset

Checking and Formatting Data Types

- Change the data type for 'age' column to integer number.
- Round the columns with decimal numbers to two decimal places
- Change the format for 'surv2m, surv6m, prg2m, and prg6m' columns. Since they represent the percentile, rename the column titles and present the percentage of them for better understanding of the values.

```
ID            int64
age         float64
death         int64
sex          object
hospdead      int64
slos          int64
d.time        int64
dzgroup      object
dzclass      object
num.co        int64
edu         float64
income       object
scoma       float64
charges     float64
totcst      float64
totmcst     float64
avtisst     float64
race         object
sps         float64
aps         float64
surv2m      float64
surv6m      float64
hday          int64
diabetes      int64
dementia      int64
ca           object
prg2m       float64
prg6m       float64
dnr          object
```

```
df['age'] = df['age'].astype(int)
```

```
#roundup the charges, totcst, totmcst columns to 2 decimal points
cost = ['charges', 'totcst', 'totmcst']

for i in cost:
    df[i] = df[i].round(2)
```

```
'''Changing format for the columns below. First rename the columns, then multiplying
them by 100 to show suitable percentages.'''

percentage = ['surv2m', 'surv6m', 'prg2m', 'prg6m']

df.rename(columns={col: f'percentage_{col}' for col in percentage}, inplace=True)

for col in percentage:
    df[f'percentage_{col}'] = (df[f'percentage_{col}'] * 100).round(1)
```

```
# Checking duplicates
df.duplicated().sum()
```

```
0
```

Checking Missing Values and Outliers

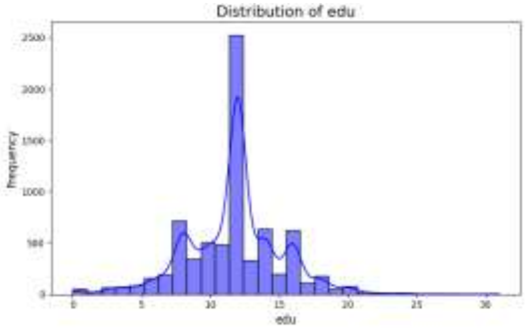
```
# Count of missing values
missing_count = (df.isnull().sum() + (df == '').sum())

# Percentage of missing values
missing_percentage = ((missing_count / len(df)) * 100).round(2)

# Checking for potential outliers for numerical columns only
numerical_df = df.select_dtypes(include=['number'])  # Filter for numerical columns
Q1 = numerical_df.quantile(0.25)
Q3 = numerical_df.quantile(0.75)
IQR = Q3 - Q1
outlier_counts = ((numerical_df < (Q1 - 1.5 * IQR)) | (numerical_df > (Q3 + 1.5 * IQR))).sum()

# Creating a DataFrame for the results
summary_df = pd.DataFrame({
"Missing": missing_count,
"% Missing": missing_percentage,
"Outliers": outlier_counts})

# Percentage of outliers
summary_df['% Outliers'] = summary_df['Outliers'] / len(df) * 100
summary_df['% Outliers'] = summary_df['% Outliers'].round(2)

summary_df = summary_df.sort_values(by="Missing", ascending=False)

# Print the combined table
print(summary_df)
```

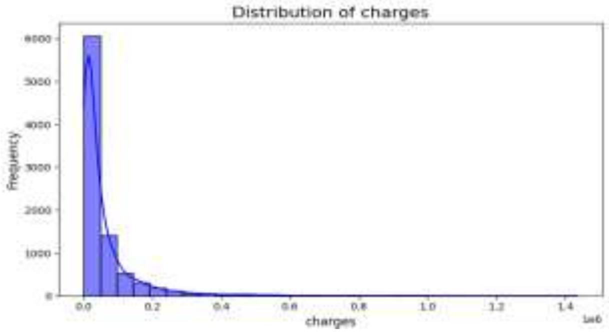
	Missing	% Missing	Outliers	% Outliers
adlp	5641	61.95	149.0	1.64
urine	4862	53.40	92.0	1.01
glucose	4500	49.42	272.0	2.99
bun	4352	47.80	267.0	2.93
totmcst	3475	38.17	495.0	5.44
alb	3372	37.03	15.0	0.16
income	2982	32.75	NaN	NaN
adls	2867	31.49	0.0	0.00
bili	2601	28.57	926.0	10.17
pafi	2325	25.54	90.0	0.99
ph	2284	25.09	260.0	2.86
percentage_prg2m	1649	18.11	0.0	0.00
edu	1634	17.95	199.0	2.19
percentage_prg6m	1633	17.94	0.0	0.00
sfdm2	1400	15.38	NaN	NaN
totcst	888	9.75	749.0	8.23
wblc	212	2.33	399.0	4.38
charges	172	1.89	912.0	10.02
avtisst	82	0.90	43.0	0.47
crea	67	0.74	987.0	10.84
race	42	0.46	NaN	NaN
dnrday	30	0.33	799.0	8.78
dnr	30	0.33	NaN	NaN
sod	1	0.01	256.0	2.81
sps	1	0.01	283.0	3.11
scoma	1	0.01	1955.0	21.47
temp	1	0.01	14.0	0.15
hrt	1	0.01	40.0	0.44
meanbp	1	0.01	6.0	0.07
resp	1	0.01	313.0	3.44
aps	1	0.01	178.0	1.95
percentage_surv2m	1	0.01	307.0	3.37
percentage_surv6m	1	0.01	0.0	0.00

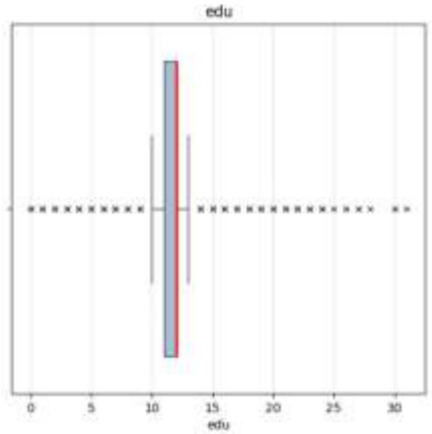
Handling Missing Values and Outliers

```python
# Imputing missing values based on the recommended normal fill in values.
df['alb']   = df['alb'].fillna(value=3.5)
df['pafi']  = df['pafi'].fillna(value=333.3)
df['bili']  = df['bili'].fillna(value=1.01)
df['crea']  = df['crea'].fillna(value=1.01)
df['bun']   = df['bun'].fillna(value=6.5)
df['wblc']  = df['wblc'].fillna(value=9)
df['urine'] = df['urine'].fillna(value=2502)
```

```python
# 'charges' column has a few negative values,
# it would be better to remove them and replace them with the median.
df.loc[df['totmcst'] < 0, 'totmcst'] = np.nan
df['totmcst'] = df['totmcst'].fillna(df['totmcst'].median())
```








```python
# Fill missing values for 'edu' with its mean
df['edu'] = df['edu'].fillna(df['edu'].mean())

# Fill missing values for the other columns with their medians
columns_to_fill_with_median = ['charges', 'totcst', 'totmcst']
for col in columns_to_fill_with_median:
    df[col] = df[col].fillna(df[col].median())
```

```python
# 'edu' column has 25+ years for the education level.
# Based on the researches and the outliers of the 'edu' column,
# imputing 25 year would be better for the upper level.
df[['edu']].boxplot()
df.loc[df['edu'] > 25, 'edu'] = 25
```

Correlation Analysis

The dataset had values of different data types

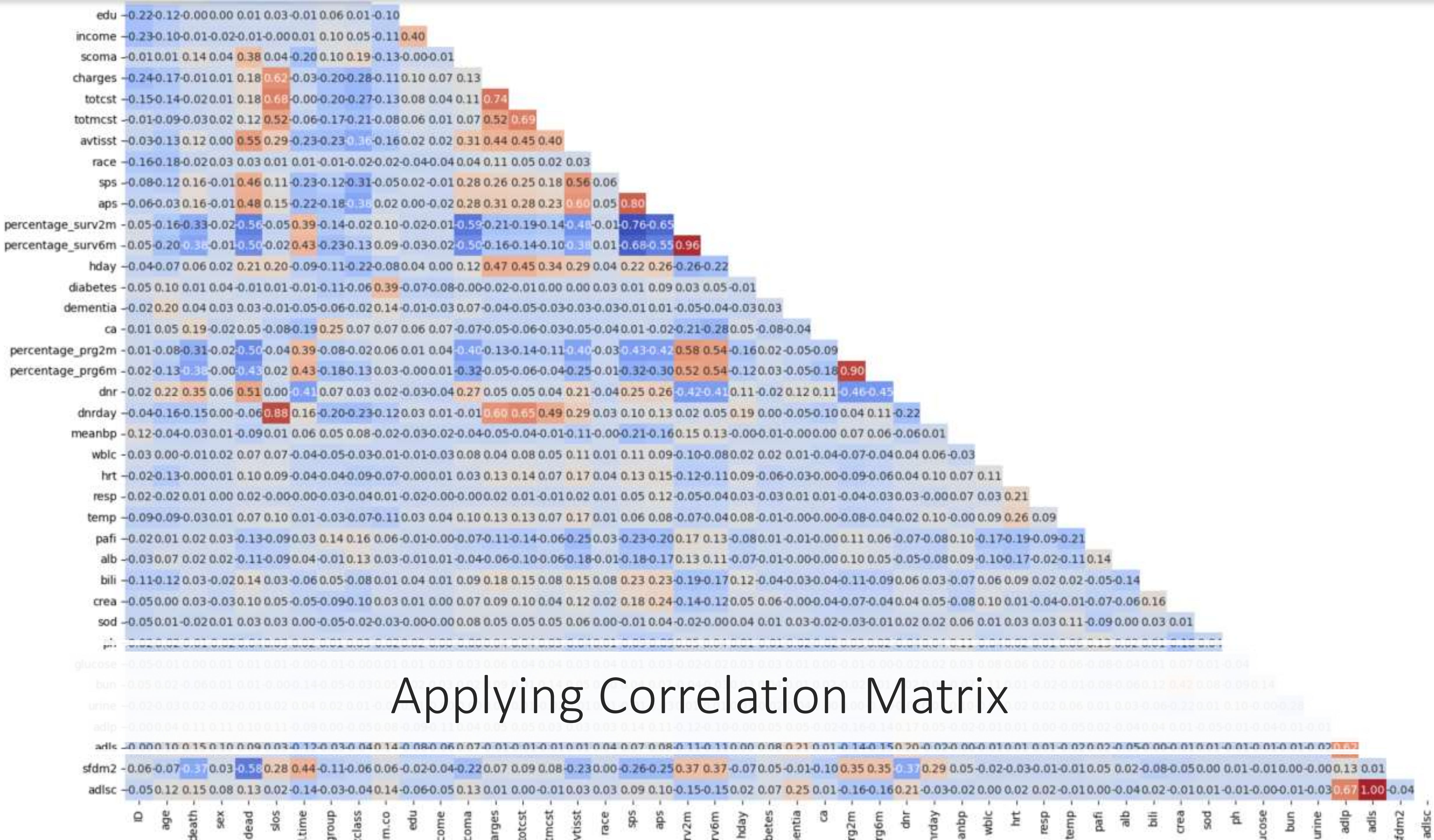
```
df.dtypes
```

ID	int64
age	float64
death	int64
sex	object
hospdead	int64
slos	int64
d.time	int64
dzgroup	object
dzclass	object
num.co	int64
edu	float64
income	object
scoma	float64
charges	float64

All categorical values were changed to numerical

```
def replace_values_dz_class(df, column):
    def replace_value(x):
        if x == "Coma":
            return 4
        elif x == "Cancer":
            return 3
        elif x == "COPD/CHF/Cirrhosis":
            return 2
        elif x == "ARF/MOSF":
            return 1
        else:
            return np.nan

    df[column] = df[column].apply(replace_value)
    return df

df_corr = replace_values_dz_class(df_corr, "dzclass")
df_corr["dzclass"].value_counts()
```

Applying Correlation Matrix



Strongest Correlations

1 Imputed Activities of Daily Living Calibrated to Surrogate - Activities of Daily Living filled out by surrogate

0.96 SUPPORT model 6-month survival estimate at day 3 - SUPPORT model 2-month survival estimate at day 3

0.95 Total micro cost - Total ratio of costs to charges

0.9 Physician's 6-month survival estimate - Physician's 2-month survival estimate

?

0.88 Day of DNR order - Days from Study Entry to Discharge

0.87 Total ratio of costs to charges - Hospital charges

0.81 Total micro cost - Hospital charges

0.8 APACHE III day 3 physiology score - SUPPORT physiology score on day 3

0.77 Total ratio of costs to charges - Days from Study Entry to Discharge

0.77 Total micro cost - Days from Study Entry to Discharge

0.73 Day of DNR order - Total ratio of costs to charges

0.72 Day of DNR order - Total micro cost

0.68 Blood urea nitrogen levels measured at day 3 - serum creatinine levels measured at day 3

0.67 Imputed Activities of Daily Living Calibrated to Surrogate - Index of Activities of Daily Living filled out by the patient

0.65 The patient's disease category - The patient's disease sub category

0.62 Day of DNR order - Hospital charges

0.62 Activities of Daily Living filled out by surrogate - Index of Activities of Daily Living filled out by the patient51515

Visualization

- Patient Demographics
- Diseases
- Mortality and Survival Factors
- Medical Expenditure

Research for Patient Demographics

- In this section, we investigate the relationship between age, gender, and socioeconomic factors with survival rates.

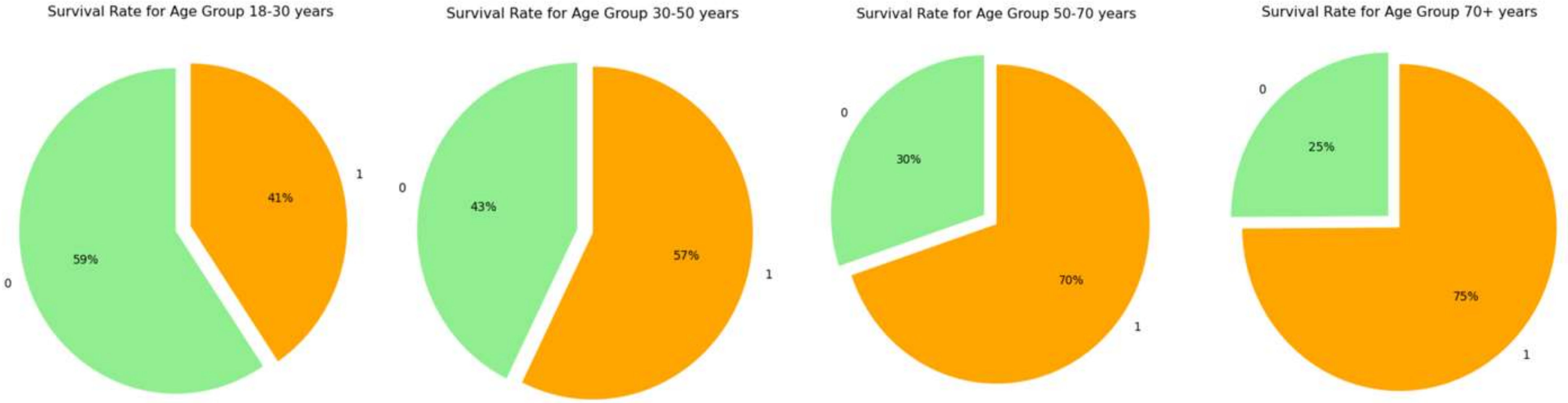
Age Distribution Among All Patients

- Individual age values were grouped into 4 categories by using 'if' function. The youngest group was indicated as 18-30 years, followed by 30-50 years, 50-70 years and 70+ years.

- Among all age groups of patients, those aged 50-70 years have the highest proportion. This group is followed by patients aged 70+, who represent the second-largest proportion. Patients between 30-50 years come next, showing a moderate contribution to the overall distribution. Finally, the 18-30 age group has the smallest proportion.

- This pattern shows that middle-aged and older adults are the most affected or most frequently observed in the study population.

Survival Rate Based on Patients' Age

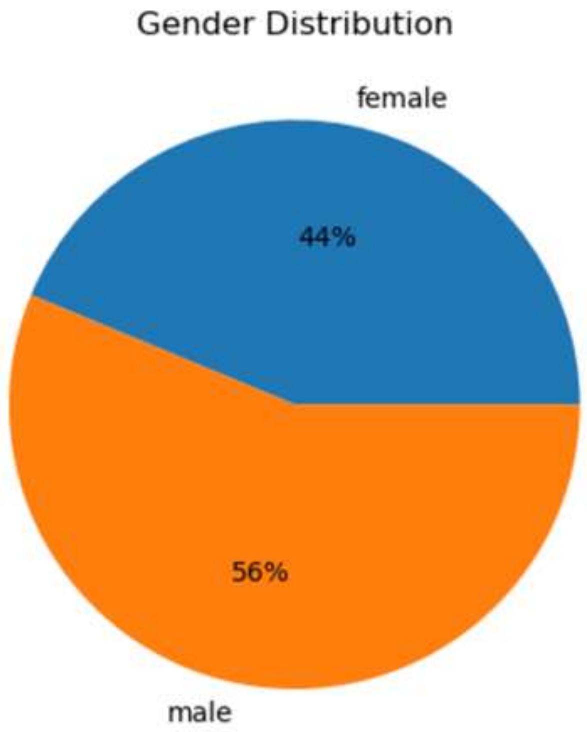


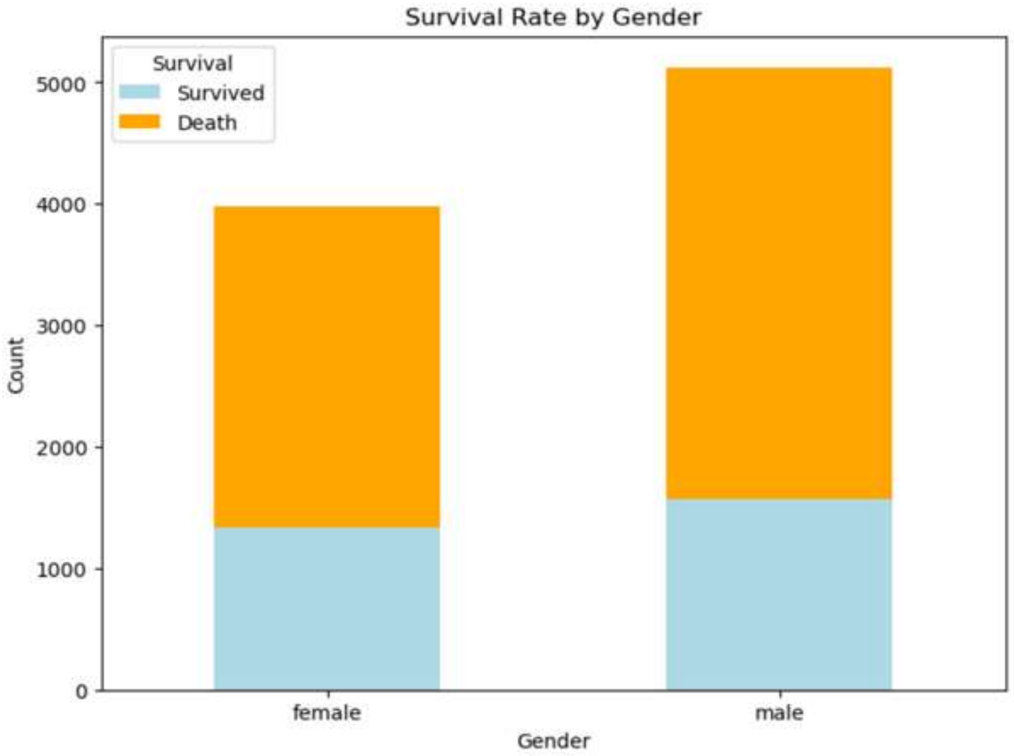
0=Survived, 1=Died

- The investigation of the death rate across each age group suggests that the death rate among individuals aged 70+ is comparatively higher in the study. This is followed by the age groups 50-70 years, 30-50 years, and 18-30 years.

Gender Related Research

- Based on the gender research, the male group has 56% of all patients while the female group has 44%. When investigating whether there is a relationship between gender and mortality rate, no significant difference was observed between genders.

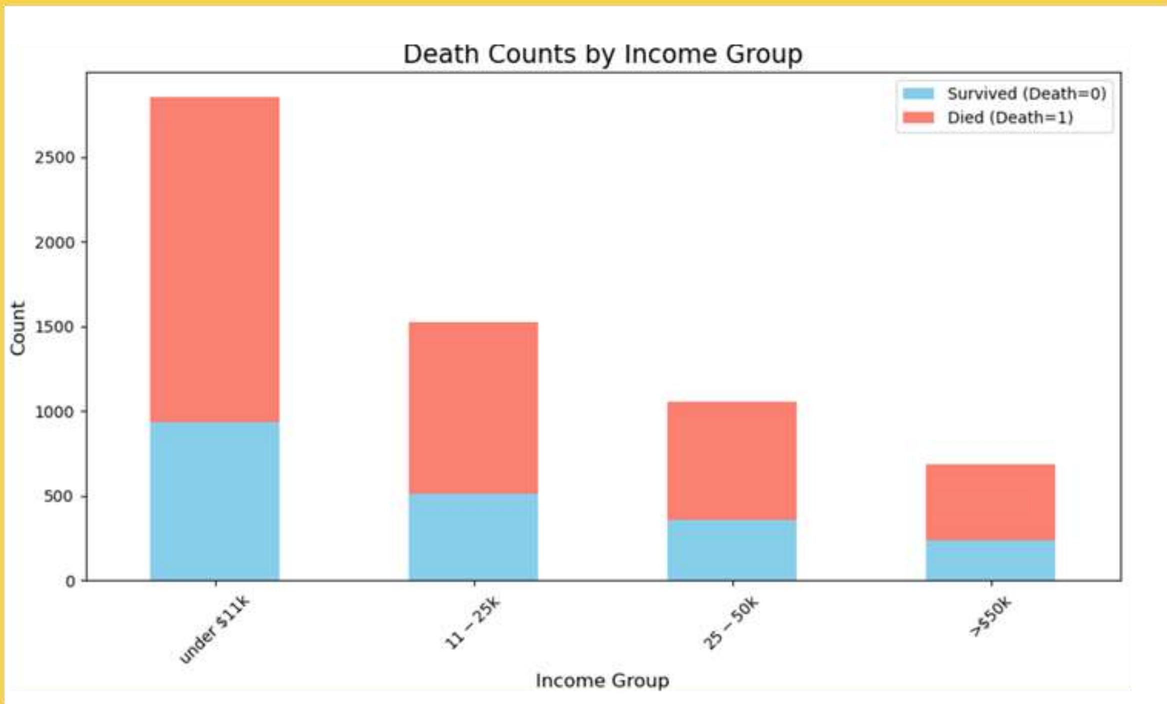


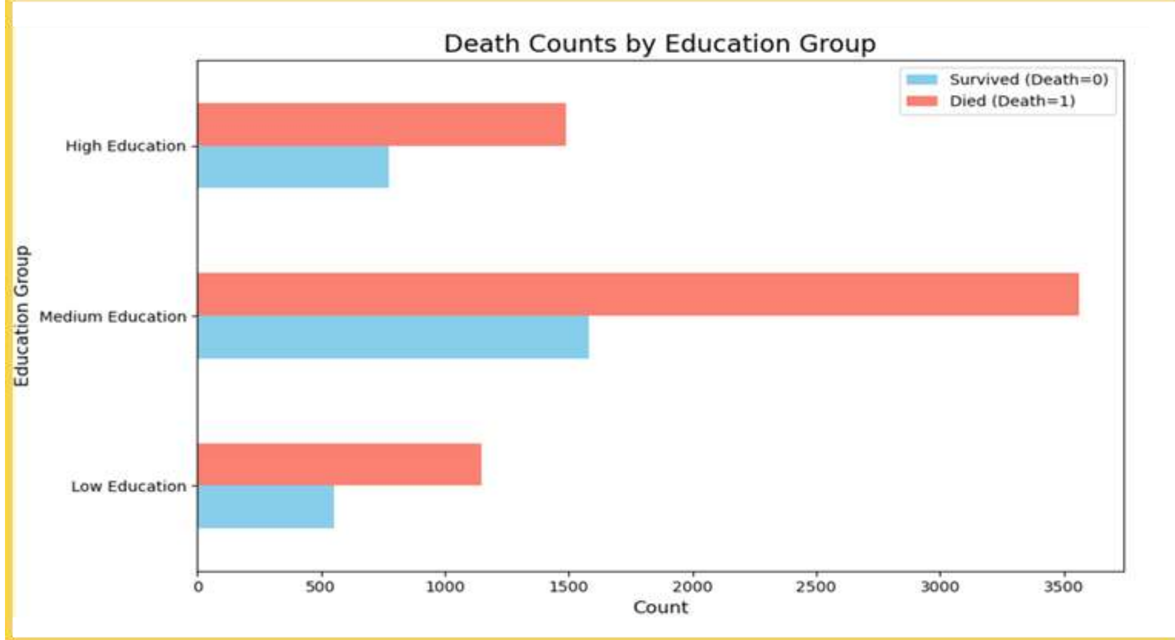


Socioeconomic Research

•When examining income among patients, it is observed that the proportion of patients with low income is higher, and this proportion decreases inversely with income levels.

•On the other hand, based on the research on education level, a medium level of education is the most common among the patients.





Research for Diseases

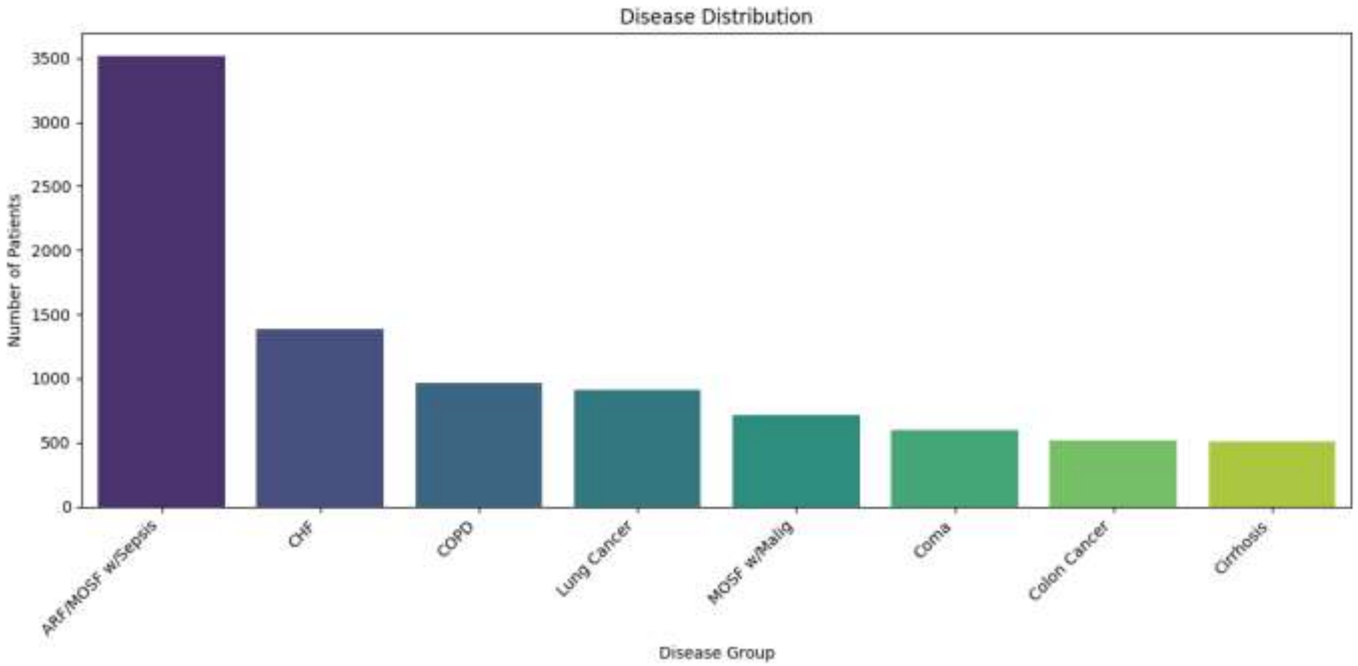
- In this section, we investigate the relationship between Comorbidities, Physiological Factors, Hospital Mortality Rates and Length of Stay with Diseases.

Disease Prevalence

- ARF/MOSF with Sepsis is the most prevalent disease with almost 3,500 patients.

- Least prevalent diseases include Cirrhosis and Colon Cancer, each with approximately 500 patients.

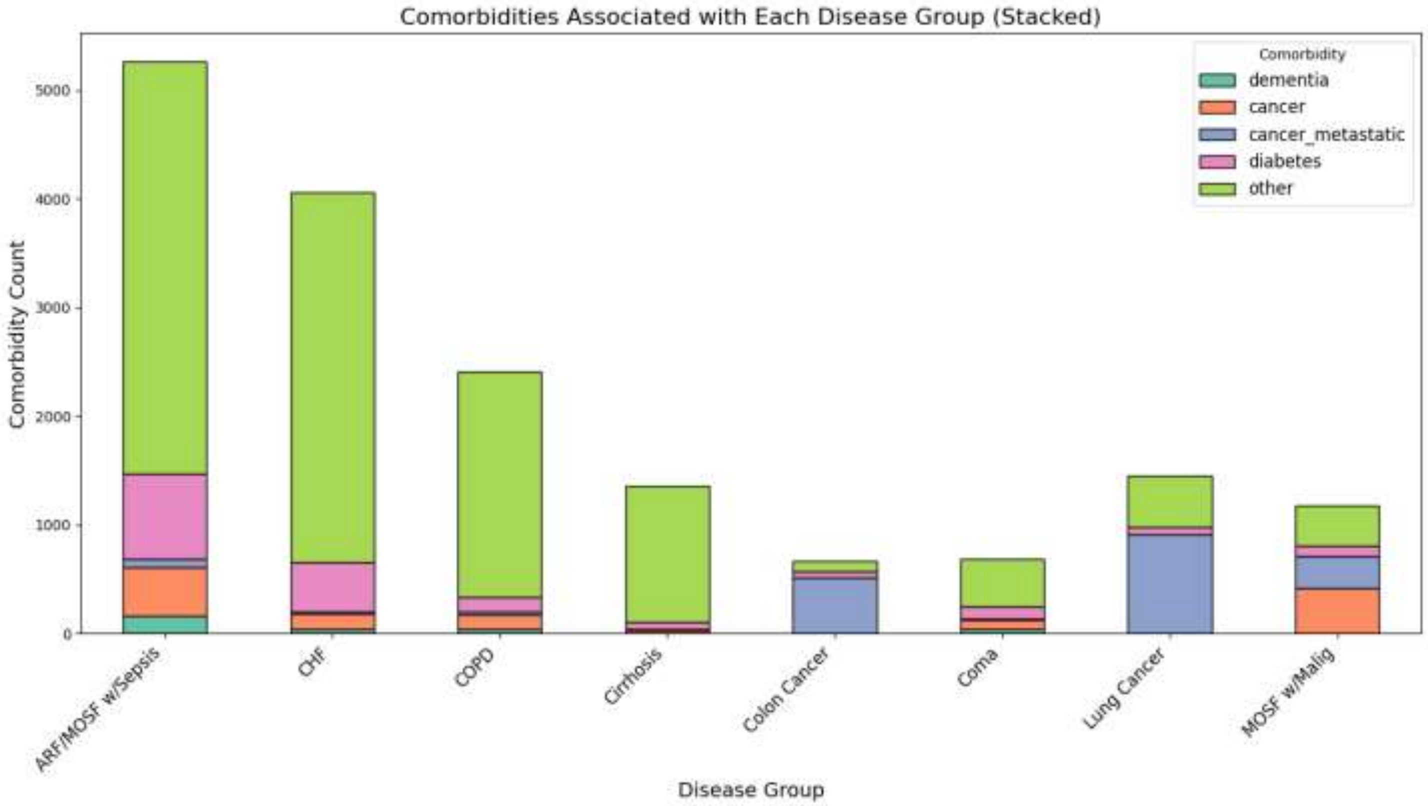
- These results could be due to,
  - **critical nature [**frequent hospitalizations or prolonged treatment durations]
  - **rarer conditions** or better-managed outpatient cases.



Association of Comorbidities to Diseases

The highest Comorbidity count is associated with ARF/ MOSF w/Sepsis.
Highest Total Burden of Patients with,

- ➢ **Diabetes as Comorbidity** – ARF/MOSF w/Sepsis & CHF [cardiovascular and metabolic disorders]
- ➢ **Dementia as Comorbidity** - ARF/MOSF w/Sepsis [chronic and critical conditions]
- ➢ **Cancer as Comorbidity** – Lung Cancer, MOSF w/Malig, Colon Cancer [chronic diseases and malignancy-driven systemic effects]



Correlation of Physiological Factors and Diseases

- No strong correlations observed between diseases and physiological variables.

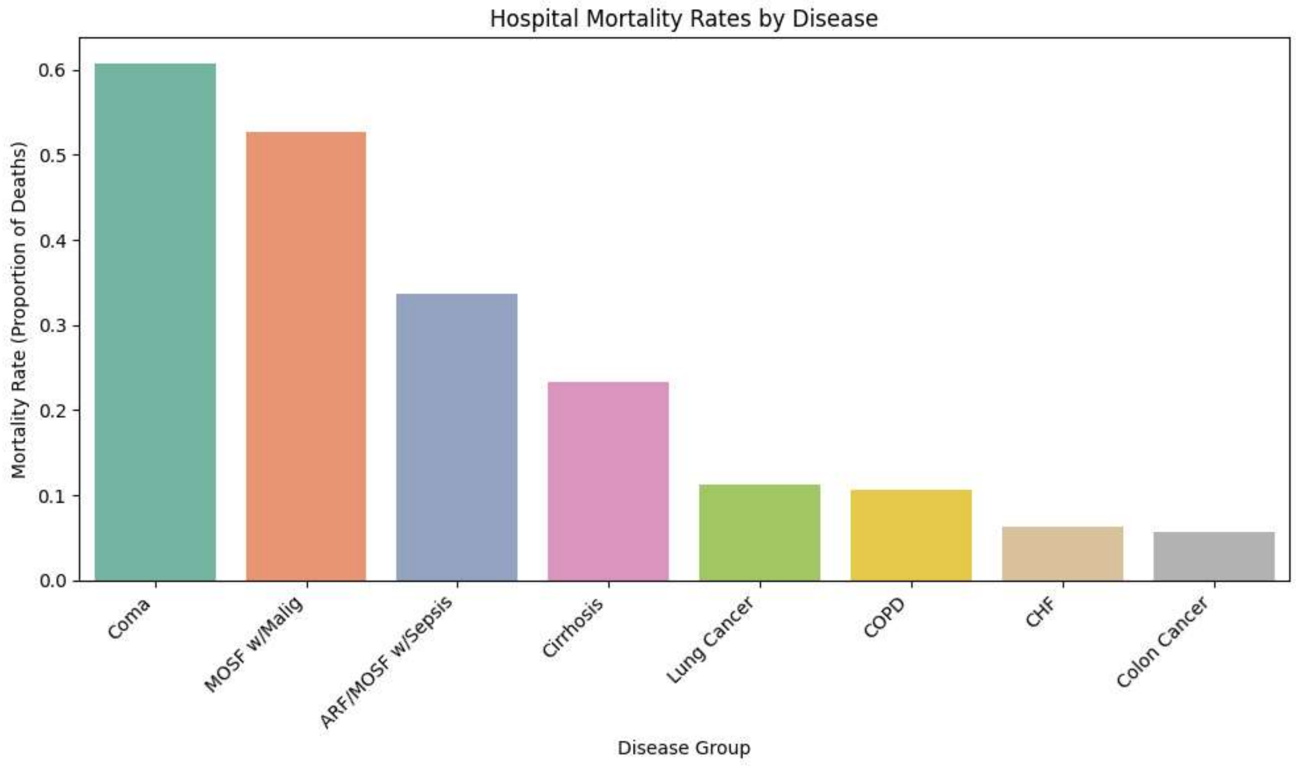
- The results could be due to,
  - Aspects of diseases not solely defined by individual physiological variables.
  - Imputation of Missing values introduce Bias and reduce Variability.

Diseases by Hospital Mortality Rates

- Highest Death rate in the hospital was recorded among patients who were in Coma.

- Lowest Death Rate in hospital was for Colon Cancer.

- These results could be due to,
  - Limited recovery potential/severe physiological derangements
  - Treatment complexity
  - Surgical and medical management effectivity

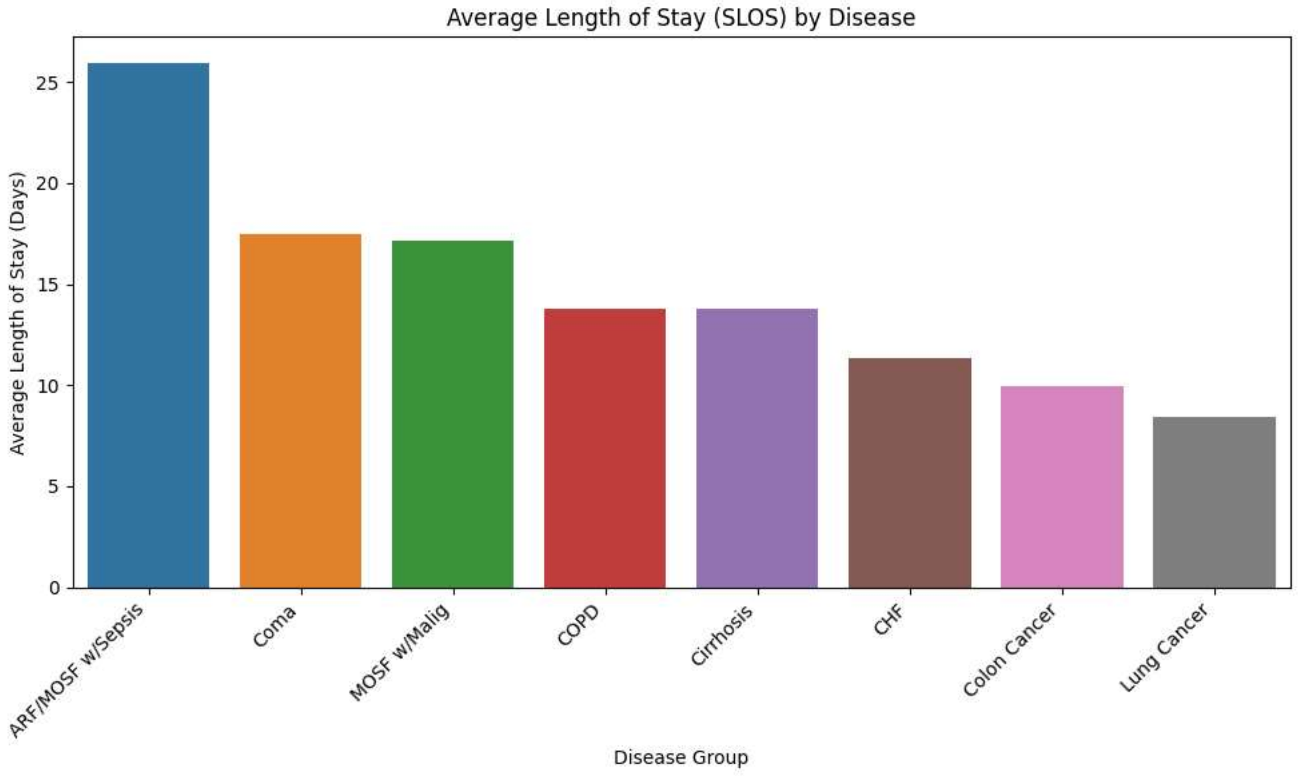


Diseases by Hospital Length of Stay

- Patients who stayed in the Hospital the most were patients with ARF/MOSF w/Sepsis

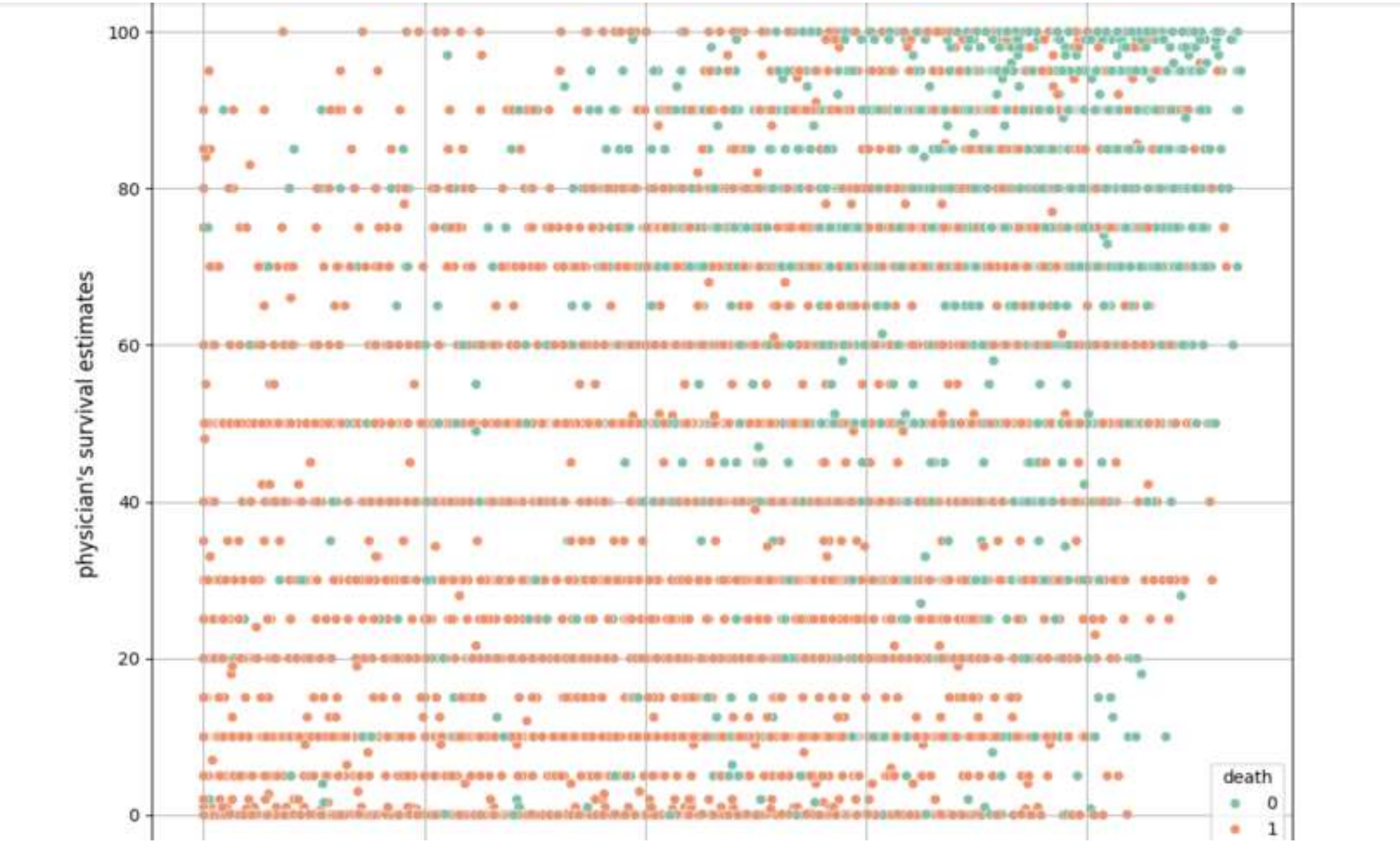
- Patients who stayed in the hospital the least were patients with Lung Cancer.

- These results could be due to,
    - Extended critical care and recovery time
    - Focused interventions or advanced disease progression



Mortality/Survival Factors

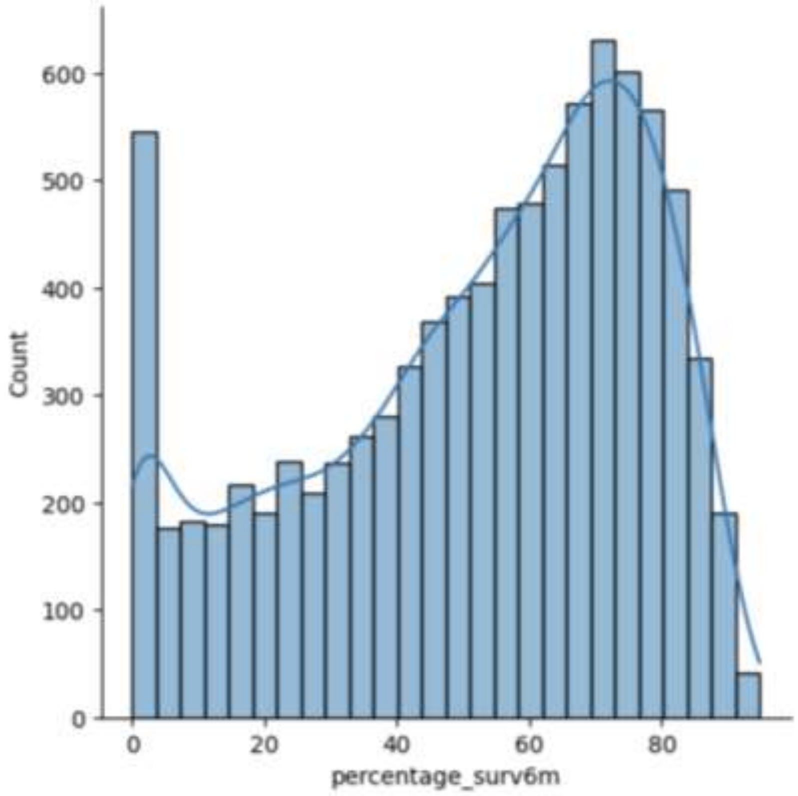
Comparison of physician's survival estimates and predictions made by SUPPORT model in terms of actual deaths



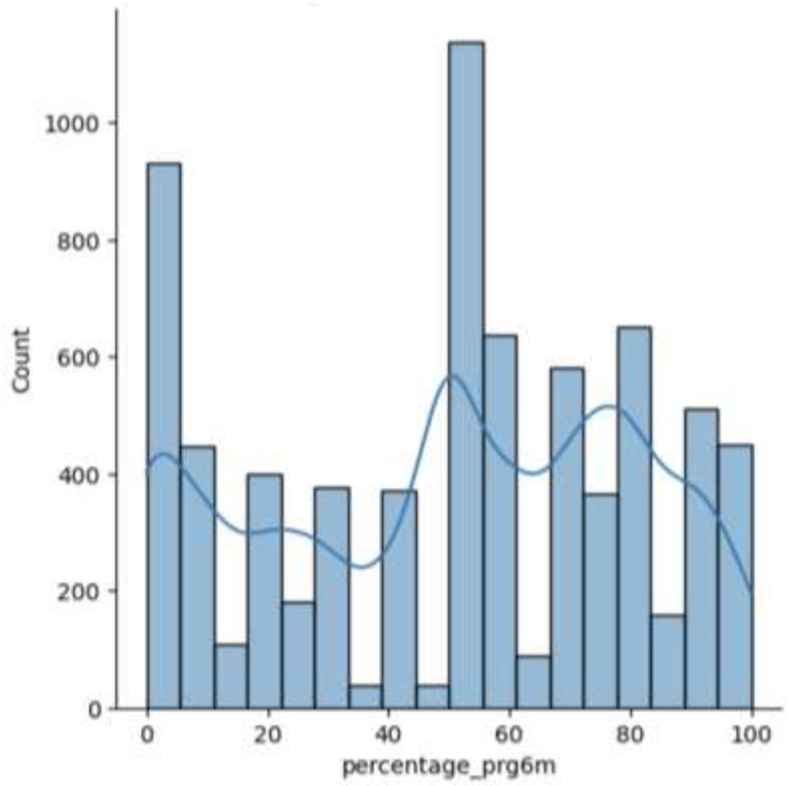
Physician's 6 months survival estimates are slightly less precise.

Curves of Physician's Survival Estimates and Predictions by SUPPORT Model

SUPPORT Model Predictions



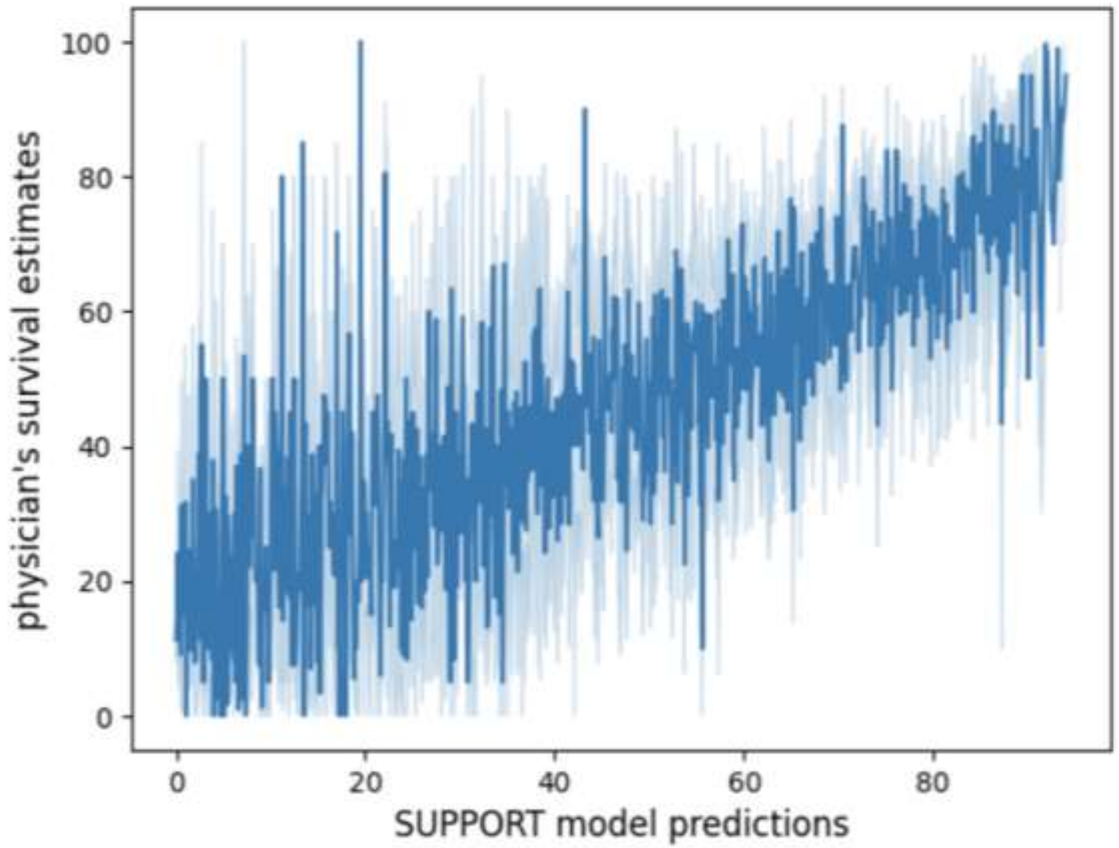
Physician's Survival Estimates



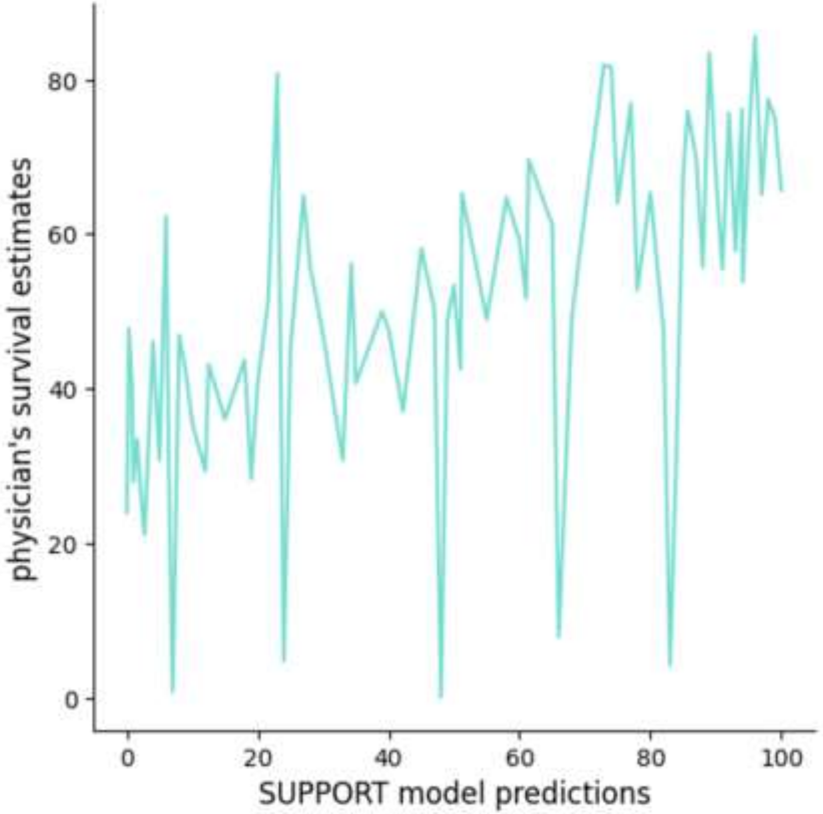
SUPPORT models 6 months survival predictions are more consistent. Predictions made by SUPPORT model give higher survival estimates than those made by physicians. Highest counts of estimates given by SUPPORT model are 70% and by physicians are 50%. Second highest is 0% for the both

Compare Physician's Survival Estimates and Predictions by SUPPORT Model

SUPPORT Model Predictions

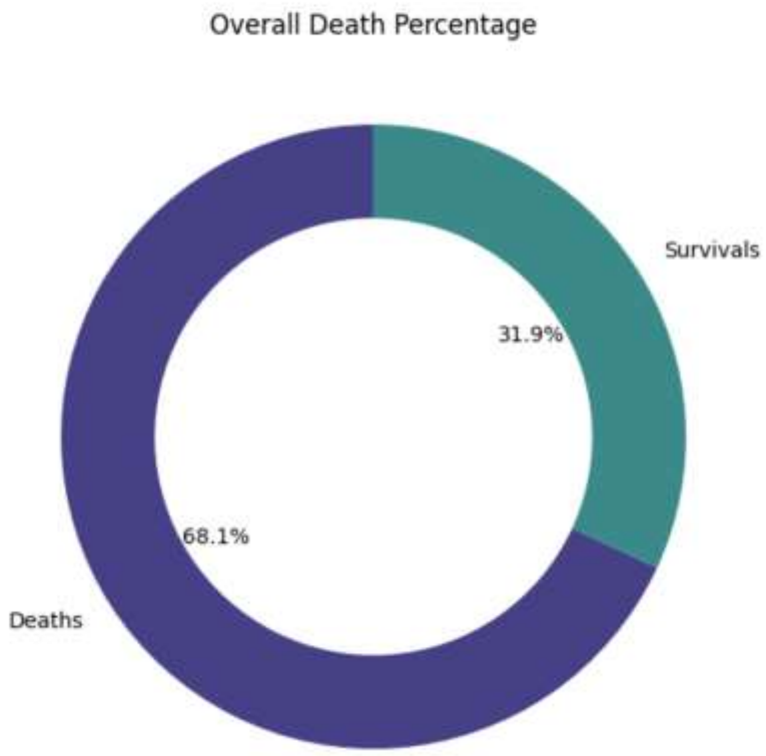


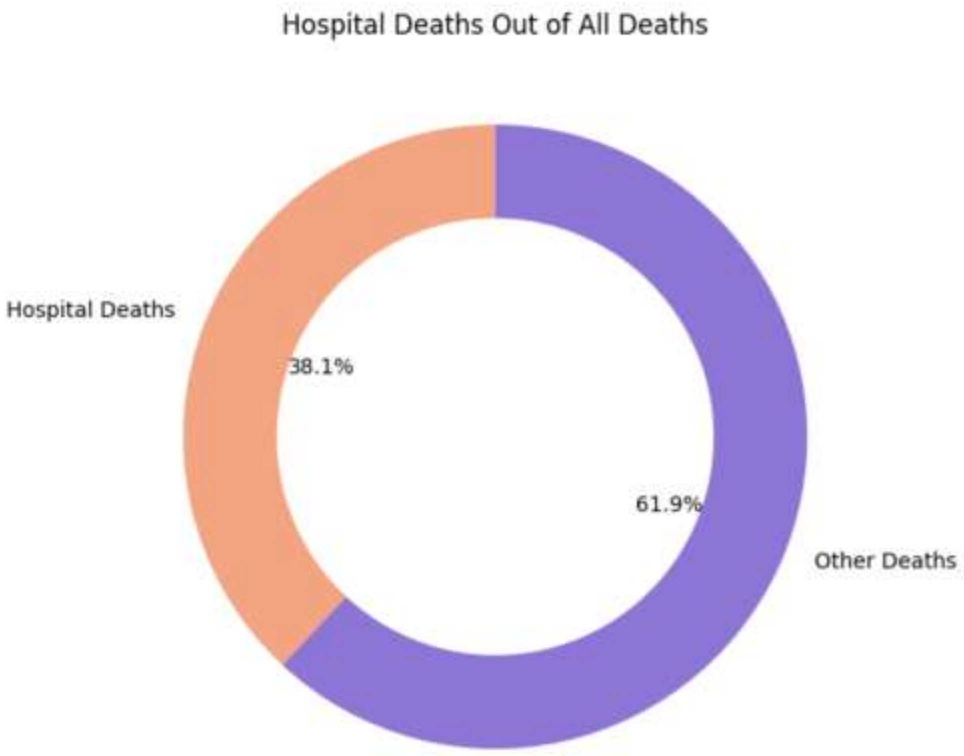
Physician's Survival Estimates



The main differences are observed at 20% vs 80% and 50% and 60%, which could result from physicians' rounding percentage. Further information on how the estimates and predictions are given is necessary.

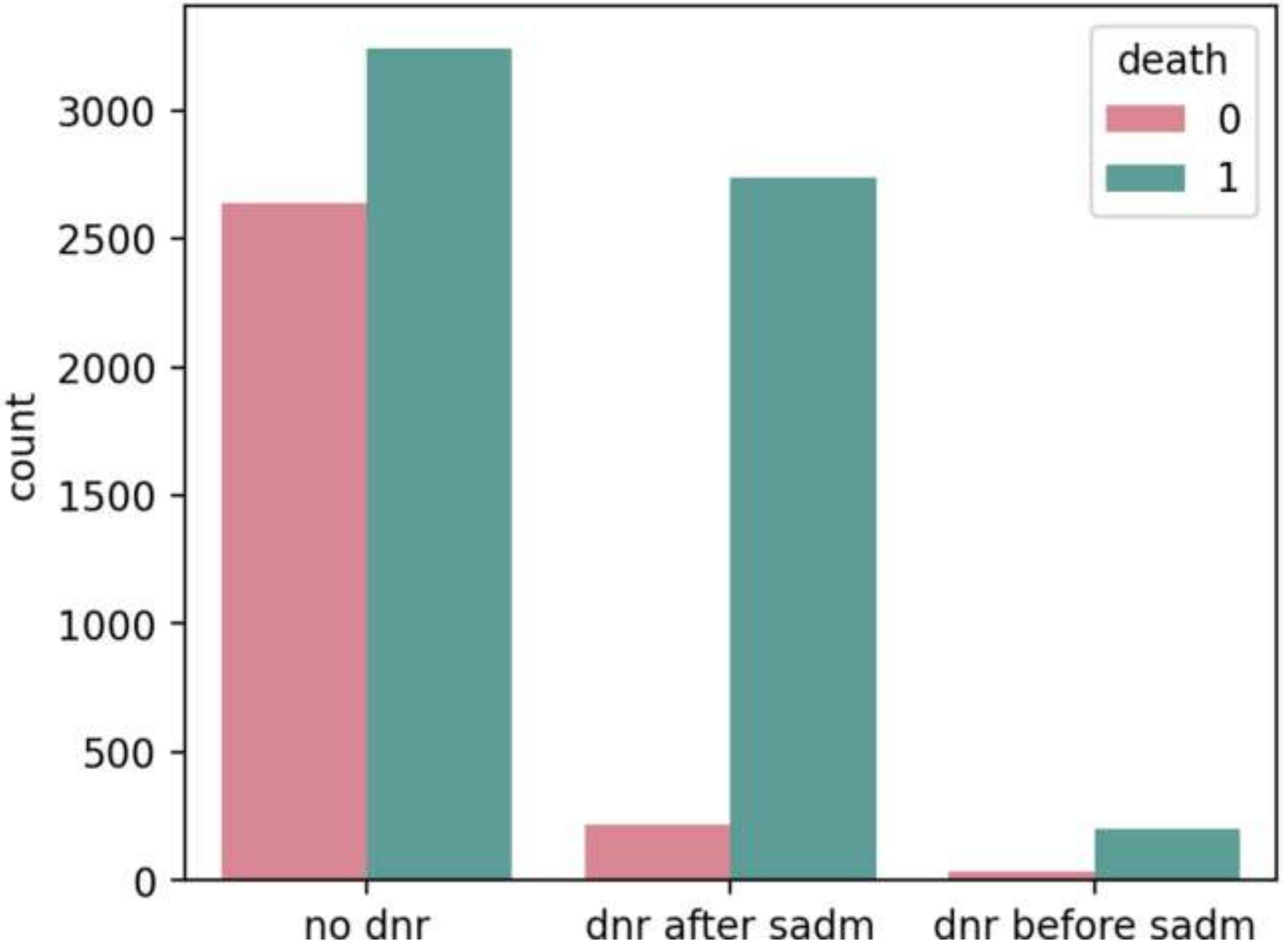
What Percentage of Deaths Were Deaths in Hospital?





Deaths in hospital are relatively low, which can potentially mean higher chances of survival if a patient is in hospital.

Potential Correlation Between the Do Not Resuscitate Order and Mortality

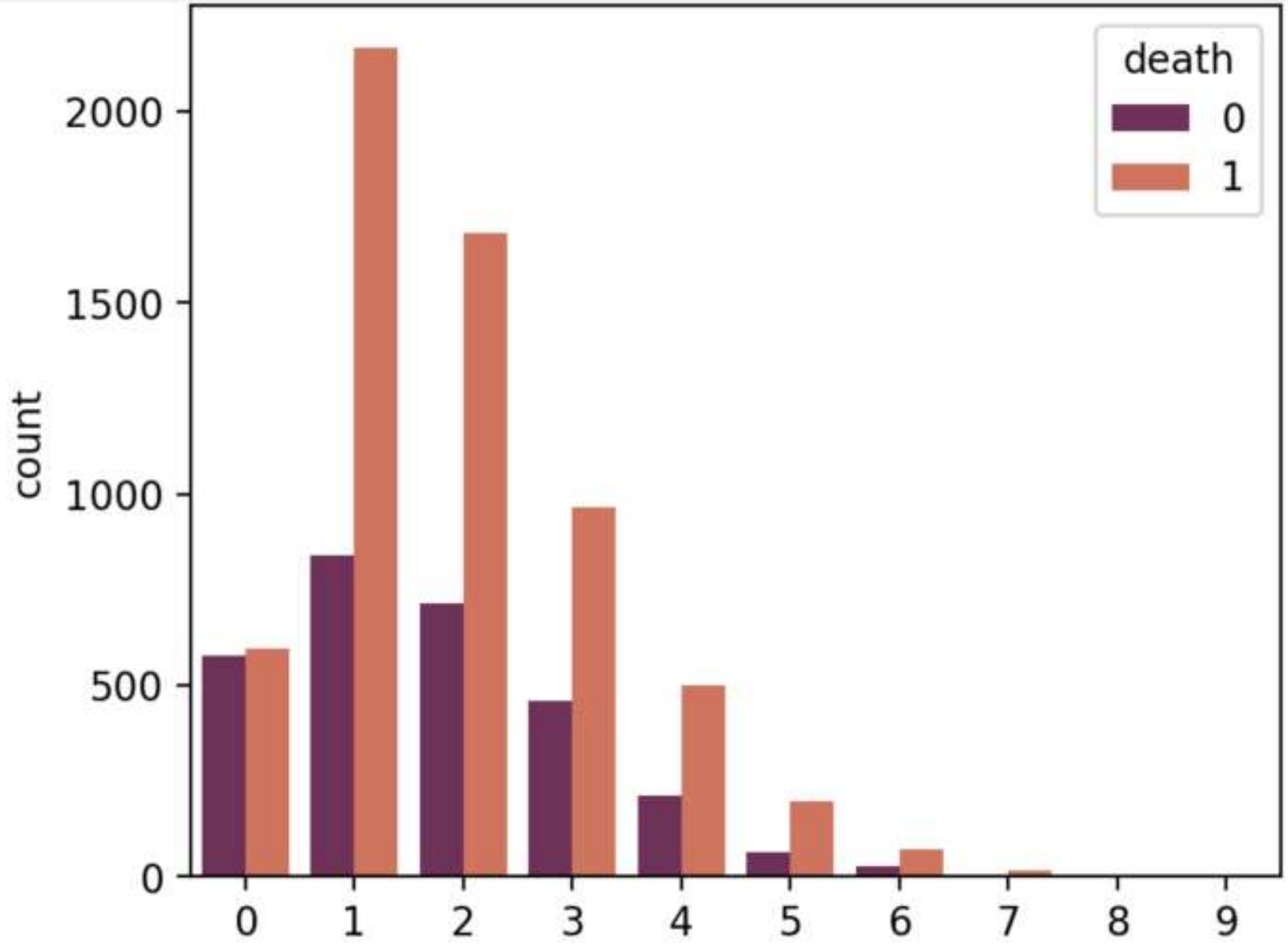


Mortality is significantly higher for patients who gave the Do Not Resuscitate Order after first admission compared to those who had the Do Not Resuscitate Order before first admission or did not have it at all.

Comparison of Total Days in Hospital and the Day of the Do Not Resuscitate Order in Terms of Deaths

There a connection between the day of Do Not Resuscitate order and length of stay in hospital but no correlation between them in the number of deaths.

Potential Correlation Between Number of Comorbidities and Mortality



The number of comorbidities does not seem to be linked to mortality, with the highest mortality with 1 and 2 comorbidities and almost 50% chance of survival with zero comorbidities.

Are the hospital charges different for people from different races?

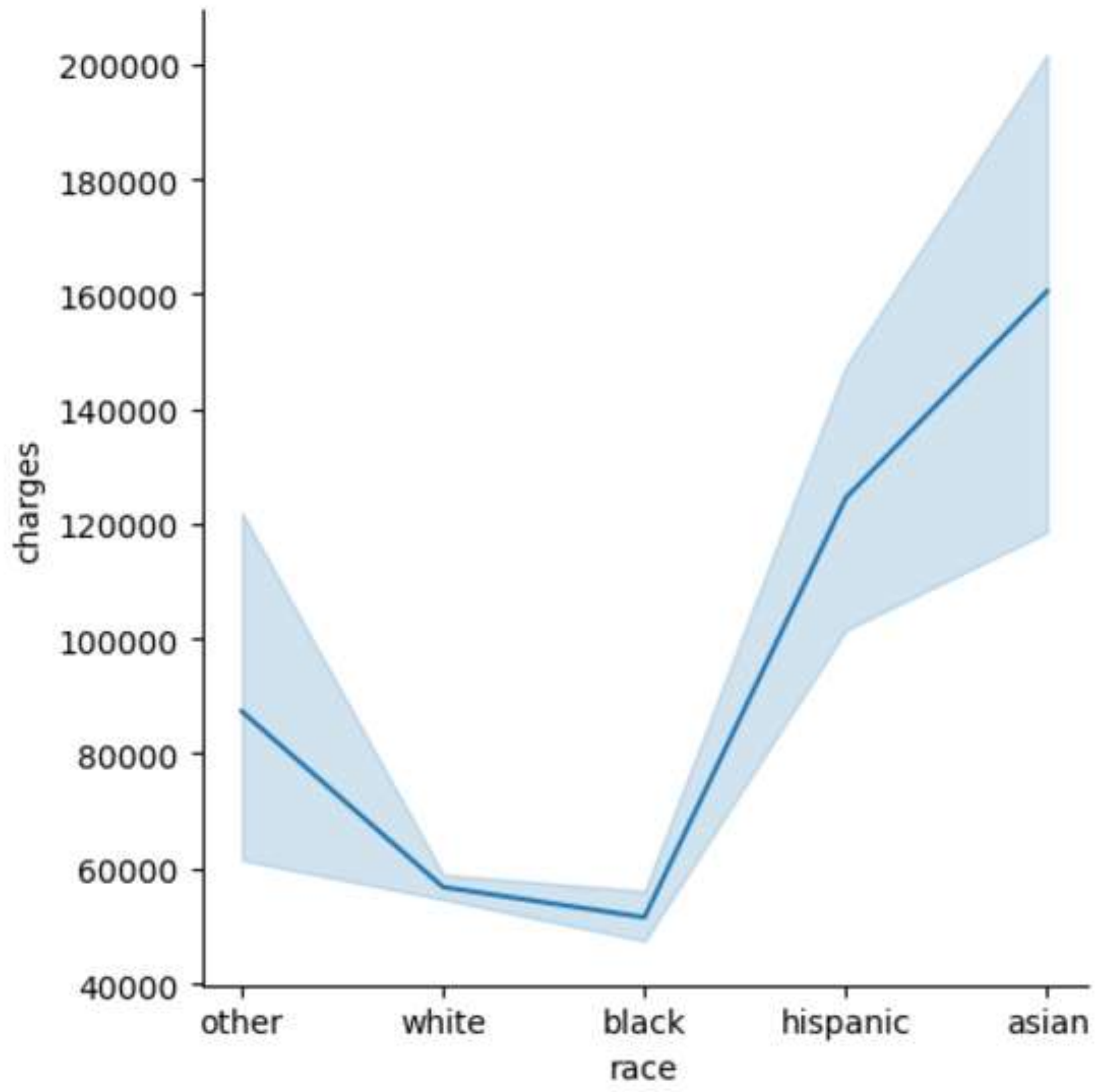
Purpose:

To find out whether people from different races were charged differently

If ‘Yes’ then how the difference in the charges are related to mainstream diseases

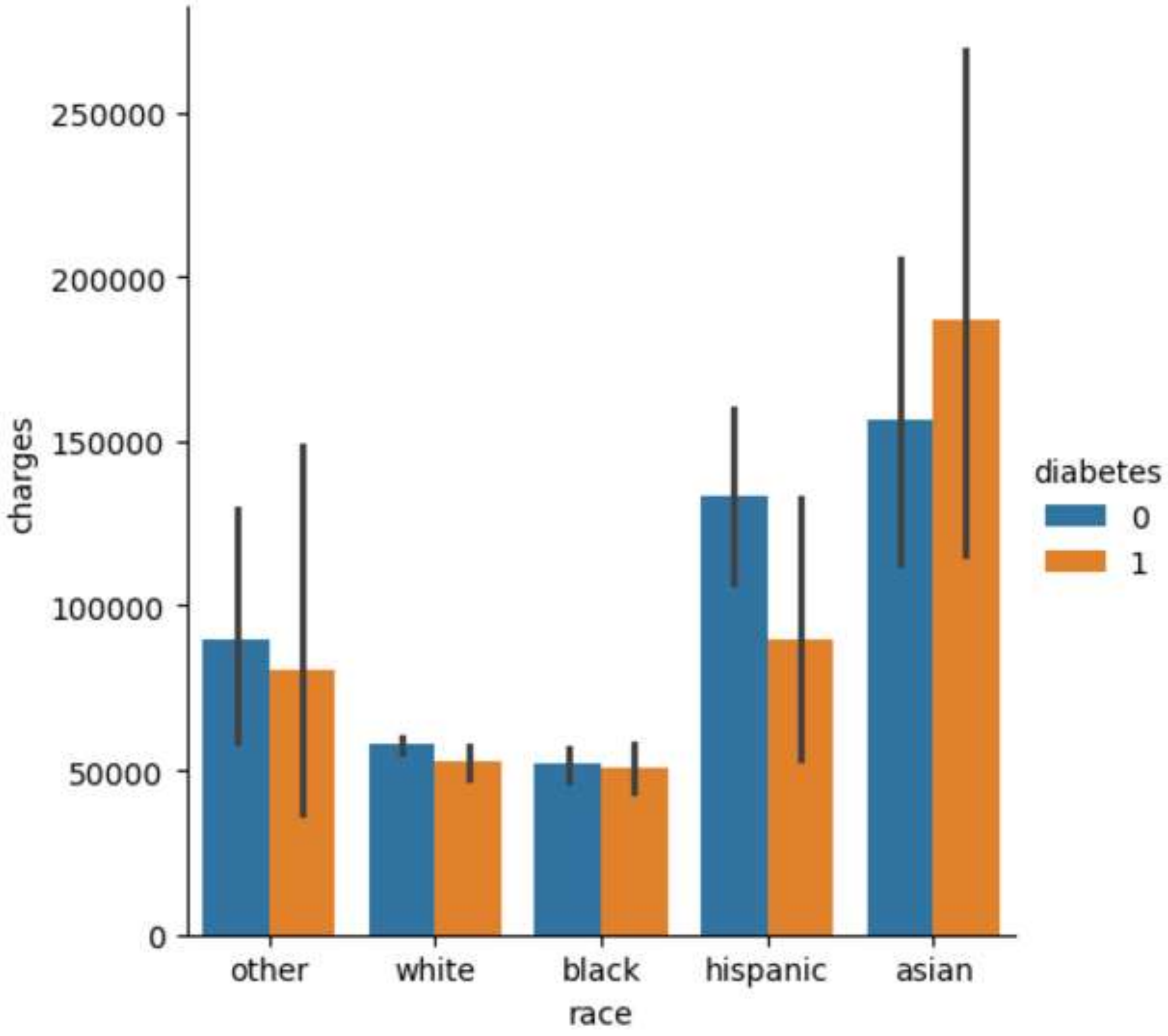
Hospital Charges charged to different races

- The plot evidently shows the difference between the charges incurred to people from different races
- People belonging to the white and black race have been charged less as compared to other races
- People from the Asian race have been charged the most as determined from the available data
- The results can be due to different reasons:
- People from different races coming to the hospital for different needs
- People from different races are more prone to the diseases than others



Hospital charges to different races with the comparison of the diabetes

- The plot shows that the Asians with diabetes are the people who have incurred the most charges from the hospital
- Other races, as compared, are charged less when differentiated based on diabetes
- This suggests that the hospital has give more attention to the Asian people with diabetes
- Meanwhile people from other races do not have much difference when the data is segregated in terms of diabetes
- People belonging to the Hispanic race shows a different trend. This can probably mean that their illness is less affected due to diabetes



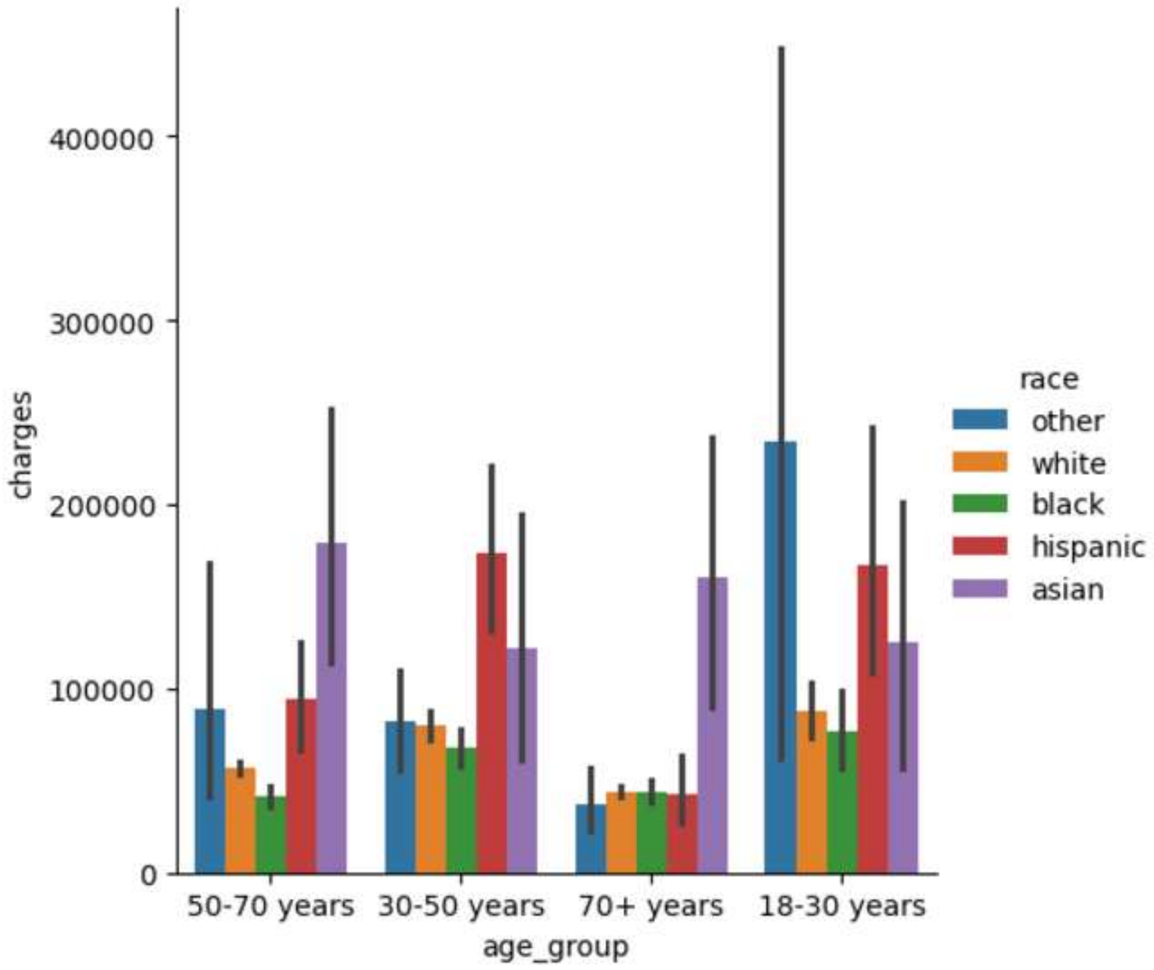
Hospital charges incurred to different age groups from different races

- We wanted to see the hospital charges differing with different age group in different races.
- The plot evidently shows that the age group 18-30 years and 30-50 years old has been charged by the hospital more
- The age group 70+ years has been charged the least as collective

This plot can suggest that the difference in charges with age and race can be due to different reasons:

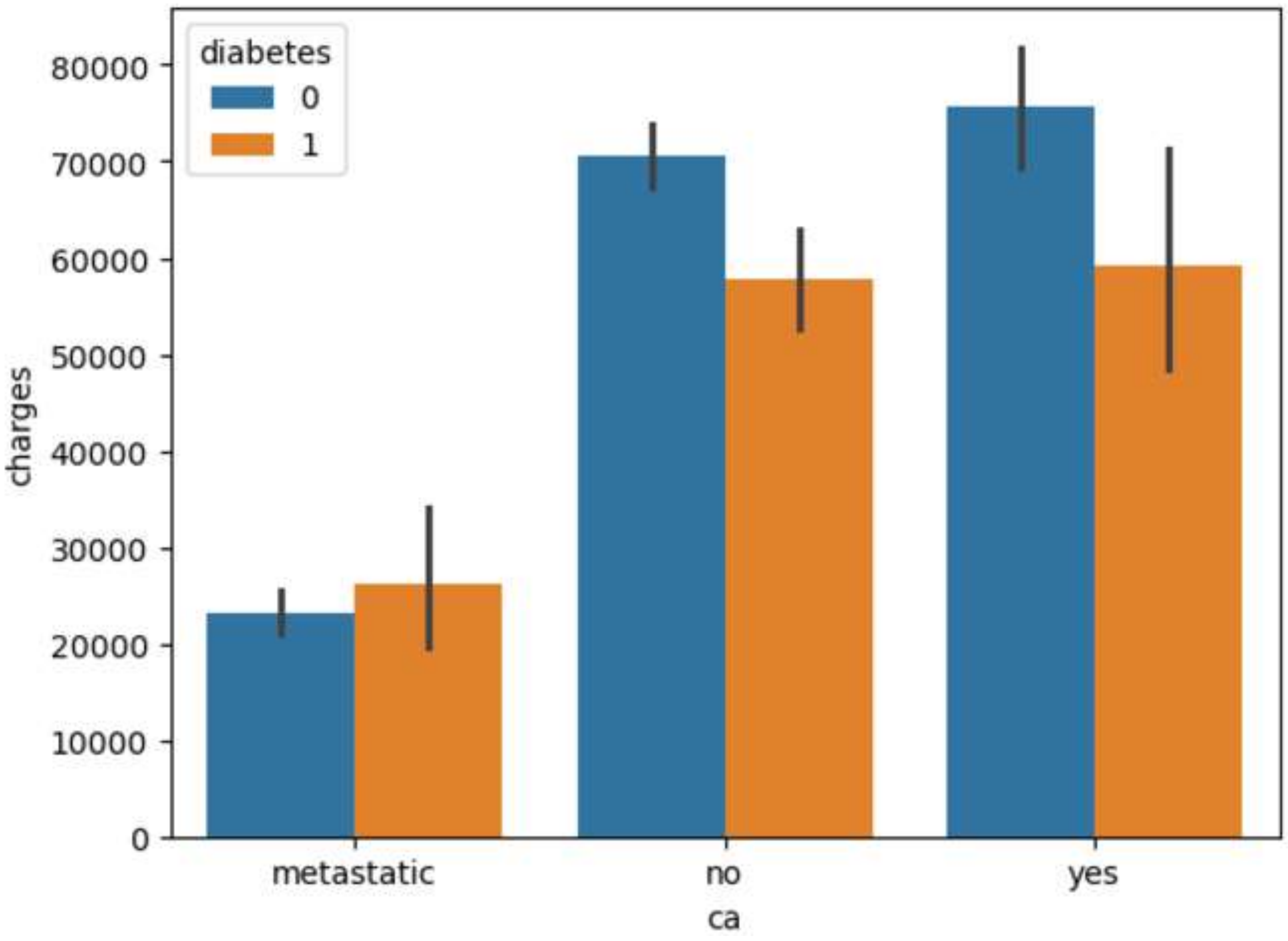
- The living habits of people from different races and ages
- The genes and their body development

This plot reflects how the hospitals must be careful with people from different ages and races



Hospital charges relation to cancer segregated with diabetes

- Assumption: People with diabetes and cancer would have incurred the most hospital charges
- The plot shows that the patients who had metastatic cancer were charged less than those who either had cancer or no cancer at all
- People with cancer who did not have diabetes were charged more than those who had diabetes
- Non-diabetic patients who did not have cancer were charged less than non-diabetic patients with cancer



Recommendations

- It is significant that mortality rate is related to age. We can conclude that the elderly group should be more cautious, and additional measures should be taken for their well-being.
- Targeted preventions can be identified by observing specific comorbidities that may dominate certain disease groups.
- 6 months survival estimates made by SUPPORT model are more precise and even. Further information on how the estimates and predictions are given is necessary.
- Staying in hospital can potentially mean higher chances of survival.
- Further research is necessary on why mortality is significantly higher for patients who gave the Do Not Resuscitate Order after first admission.

Challenges

- Due to the lack of domain knowledge, some of the missing values couldn't be imputed.
- Decision for keeping, transforming, or removing outliers is highly dependent on domain knowledge. Some of the outliers transformed or removed based on the insight from the data set. However, most of the outliers were left unchanged to prevent the loss of any critical information.
- Distribution of uneven amount of patient subgroup data(Ex: Different Disease Groups, Patient Demographics) makes it difficult and limits the conclusions we can draw out of them.

Q&A