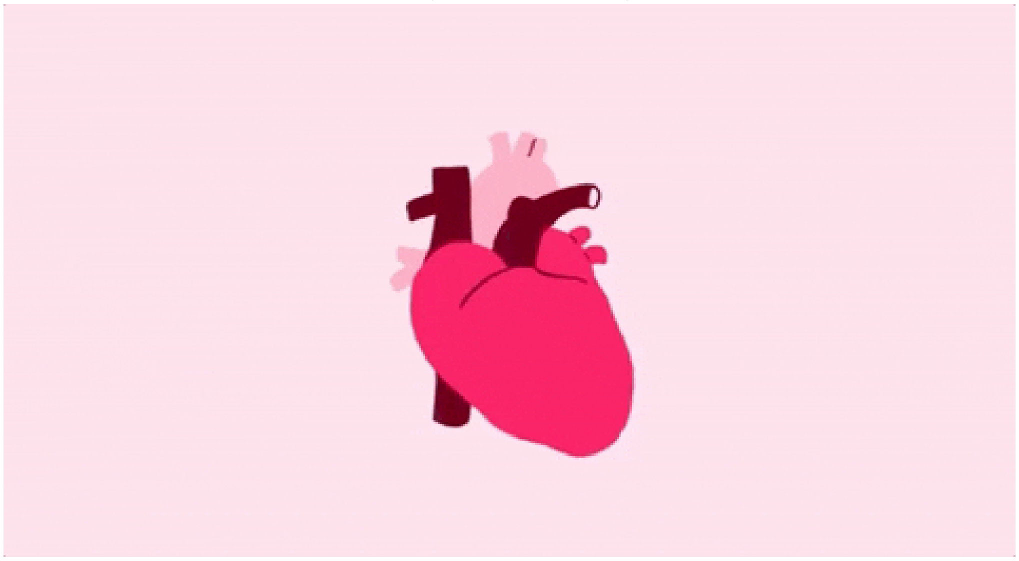
U of M Summer 2022
Data Analytics and Visualization Capstone Project
Group 2:
Lindsay Teeters
Kate Heise
Lalit Toshniwal
Nick Petcoff

In Hospital Mortality: Prediction of Heart Failure
Project Summary



Introduction/ Inspiration:

Heart failure is a leading cause for many deaths in America. The CDC reported that in 2018 close to 400,000 Americans died from some form of heart failure. During a study in 2012, it was determined that 6.2 million Americans had heart failure as a medical condition, leading to a cost of over $30 billion in health care costs (CDC, 2020).

Several of the team members work within healthcare, with interest in this field present, we wanted to take a look at a healthcare issue to see what drove the in-hospital mortality of heart failure patients. With heart failure becoming a greater issue for patients everyday, we wanted to take a look into the causes of heart failure.

Heart failure treatment is a major contributor to in-hospital mortality, and with that come great costs of care. The dataset that we used, had information on patients that either died in care or survived to discharge (Kaggle, 2021).

With this data we explored heart failure and how certain comorbidities and lab results can predict if a patient will survive while in hospital care to discharge, or if the patient died in care. This was completed by first doing an exploratory analysis to determine what factors correlate with the target variable of survival or in hospital death. This review was carried through visualizations utilizing Tableau and running machine learning methods to make predictions on the features with most correlation to the target variable. All of this is presented on a student built website.

Hypothesis/Expectations:

Our initial expectations for this data were that the greater the age the higher the mortality rate would be. The team also believed that the more comorbidities a patient had, the higher the likelihood that they would not survive to discharge.

With one of our team members being a nurse practitioner, it was believed that some laboratory tests would also be important in understanding and predicting the in-hospital mortality rates of heart failure patients.

Exploratory Data Analysis:

The data that we found on Kaggle.com was a csv that contained a little over 1170 rows and 51 columns. Upon our initial review, it seemed like a clean dataset. A majority of these columns were in numeric values. Nine of the columns were boolean values, in which they represented several comorbidities. If the boolean was listed as true then the patient had heart failure and the listed comorbidity. The remainder of the columns were listed as integers. There were no non-numeric

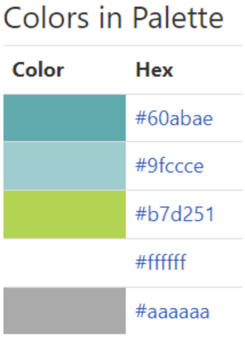
Once our group started looking into the data, we realized that many rows had missing values in the columns. We reviewed the data to determine how to move forward these NaN values. If we removed these NaN values, It would have drastically cut down the size of our dataset.

We determined that we would need to impute the missing NaN values so that we can review the data and have enough data points to make a prediction of mortality within hospital patients using machine learning models.

Initially we imputed the missing NaN values by taking the mean from each column, as they were all numeric, and imputing those averages into the missing columns fields. However, we also improved upon this by using linear regression for one of the columns (PCO2) that had a high degree of correlation with another column (Bicarbonate).

Within the dataset we did end up having to remove one entire row, as this row did not have a value added for the “outcome” column, which indicated if the patient survived to discharge. Since we could not reasonably predict the outcome of survival or discharge this was removed, which brought our row count of fully complete data to 1176. This new data set was used to complete the remainder of the EDA review, visualations, and machine learning.

Starting with the ERD review and into the other parts of the project, we used a color palette called Essential Wellness. We chose this color concept, as it was easy on the eyes and was colorful enough to be interesting and not be too overbearing.

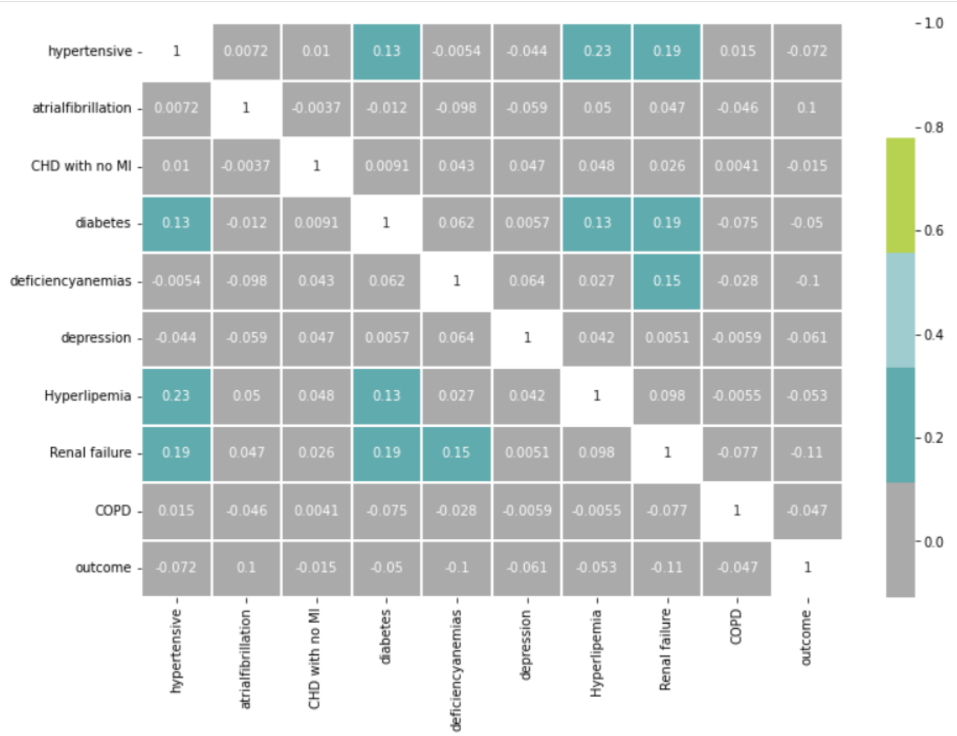


Colors in Palette

Color	Hex
■	#60abae
■	#9fccce
■	#b7d251
□	#ffffff
■	#aaaaaa

With the data having 51 rows, we first looked at the comorbidity section. By definition, a comorbidity is the presence of having more than one medical condition at one time. With the target variable being the “outcome, " we reviewed the correlation of these comorbidities using a heatmap.

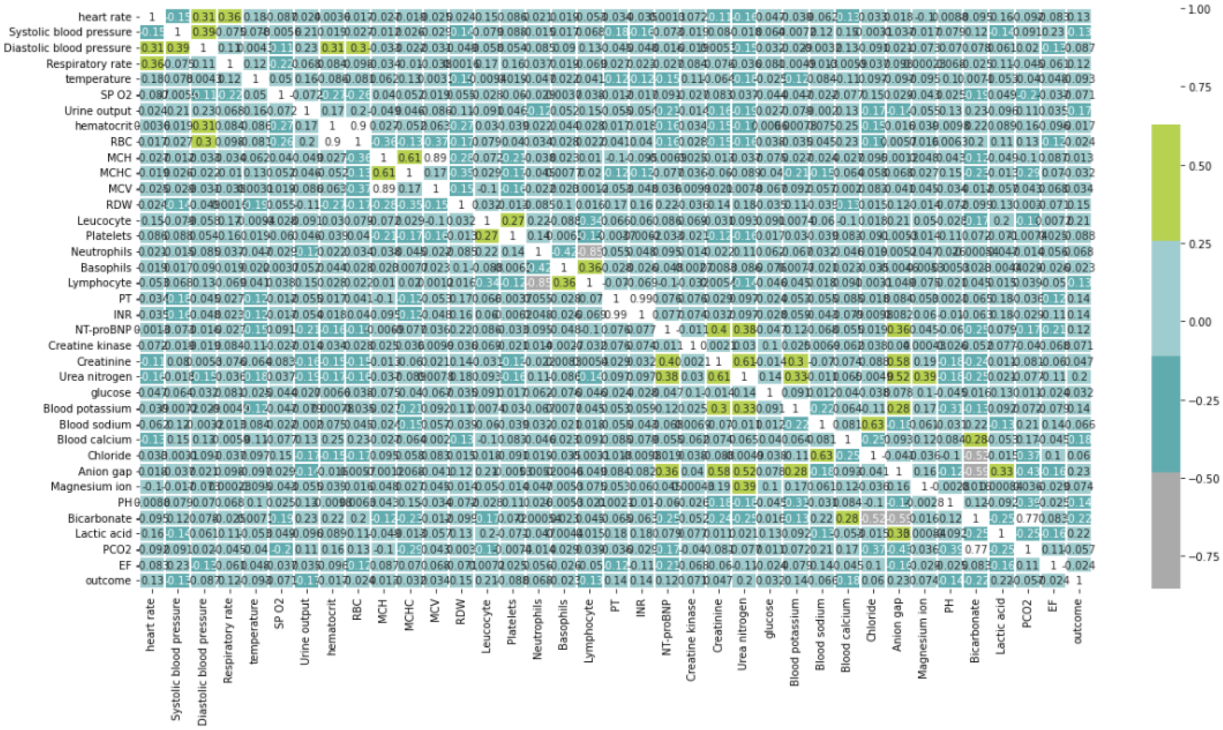
Comorbidity Heatmap:



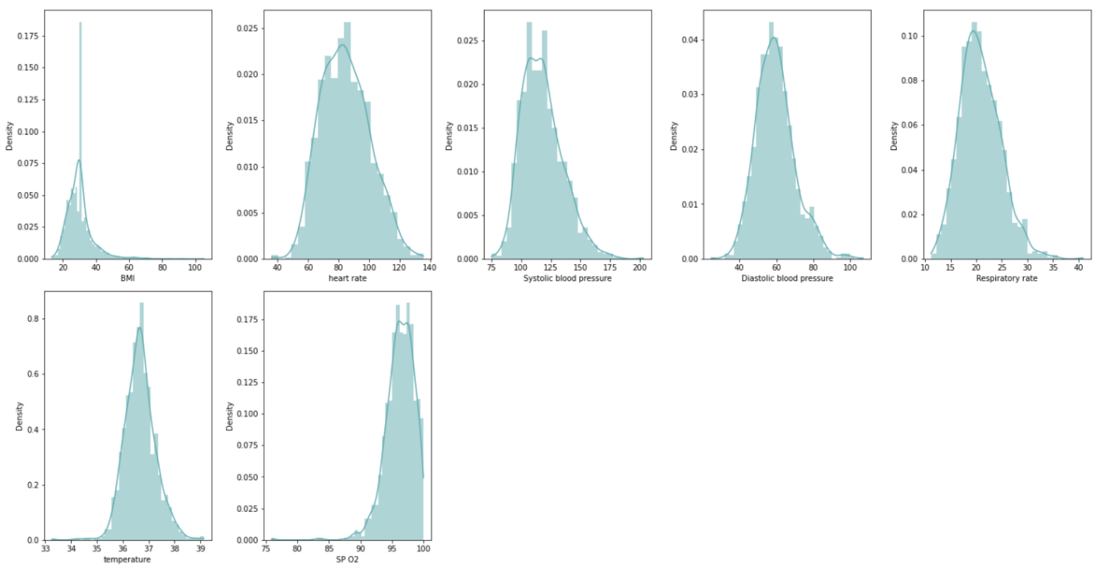
	hypertensive	atrialfibrillation	CHD with no MI	diabetes	deficiencyanemias	depression	Hyperlipemia	Renal failure	COPD	outcome
hypertensive	1	0.0072	0.01	0.13	-0.0054	-0.044	0.23	0.19	0.015	-0.072
atrialfibrillation	0.0072	1	-0.0037	-0.012	-0.098	-0.059	0.05	0.047	-0.046	0.1
CHD with no MI	0.01	-0.0037	1	0.0091	0.043	0.047	0.048	0.026	0.0041	-0.015
diabetes	0.13	-0.012	0.0091	1	0.062	0.0057	0.13	0.19	-0.075	-0.05
deficiencyanemias	-0.0054	-0.098	0.043	0.062	1	0.064	0.027	0.15	-0.028	-0.1
depression	-0.044	-0.059	0.047	0.0057	0.064	1	0.042	0.0051	-0.0059	-0.061
Hyperlipemia	0.23	0.05	0.048	0.13	0.027	0.042	1	0.098	-0.0055	-0.053
Renal failure	0.19	0.047	0.026	0.19	0.15	0.0051	0.098	1	-0.077	-0.11
COPD	0.015	-0.046	0.0041	-0.075	-0.028	-0.0059	-0.0055	-0.077	1	-0.047
outcome	-0.072	0.1	-0.015	-0.05	-0.1	-0.061	-0.053	-0.11	-0.047	1

To further our review, we looked at several laboratory tests as well. With there being 40 plus lab tests available for review the original heatmap was crowded and understanding that not all of these values would be needed in a machine learning or visualization they were narrowed down to the ones that had the greatest correlation to the outcome.

Laboratory Test Heatmap



Upon this review of data, the tests were narrowed down to several that had better correlating to the outcome some histograms were developed.

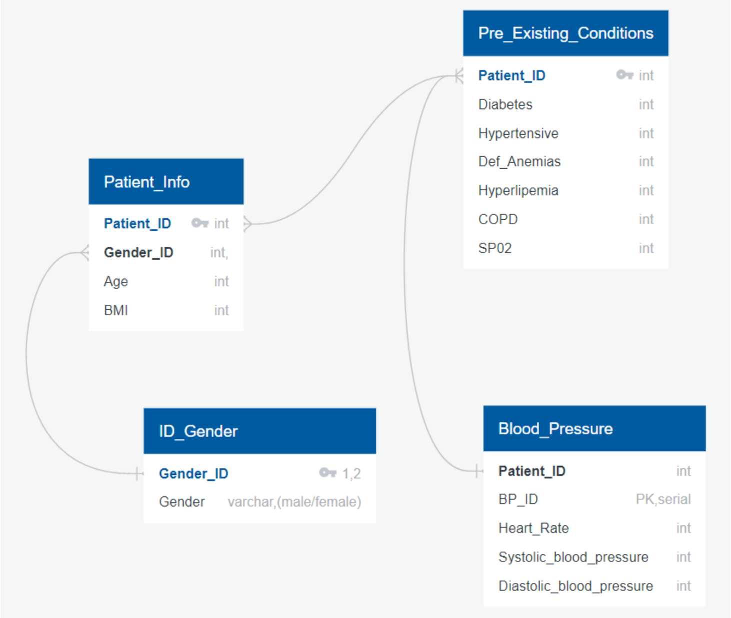


Through this review it could be seen that not all of the columns would be used within the machine learning module training.

DATABASE SECTION:

Fortunately for our project, our one dataset contained all the items we needed. We did not have to worry about joining multiple datasets together. Within our data set, we were able to generate a couple of different relation tables. Our first one contained the Patient Information. Our foreign keys were Patient ID and Gender ID.

The patient ID allowed three of the tables to be linked (Patient Info, Pre-Existing Conditions, and Blood Pressure. Blood Pressure was able to be a primary key between the tables. The Gender ID was able to link our first table (patient info) together with the ID Gender table. The Gender ID table transformed the numeric gender value into the proper Male/Female genders.



Patient_Info	
Patient_ID	int
Gender_ID	int,
Age	int
BMI	int

Pre_Existing_Conditions	
Patient_ID	int
Diabetes	int
Hypertensive	int
Def_Anemias	int
Hyperlipemia	int
COPD	int
SP02	int

ID_Gender	
Gender_ID	1,2
Gender	varchar,(male/female)

Blood_Pressure	
Patient_ID	int
BP_ID	PK,serial
Heart_Rate	int
Systolic_blood_pressure	int
Diastolic_blood_pressure	int

Machine Learning:

We tested following supervised machine learning classification models using all the provided demographic, vital and lab results variables - a total of 50 features:

- Logistic Regression
- Decision Tree Classifier,
- Random Forest Classifier,
- AdaBoost Classifier,
- Extra Trees Classifier,
- Gradient Boosting Classifier,
- Light GBM Classifier,
- XGBoost Classifier,
- Support Vector Machines,
- K- Nearest Neighbors Classifier,
- Deep Neural Network

Since we had class imbalance (most patients admitted to the ICU survived) in the existing dataset, we tried random under-sampling, random over-sampling and combination sampling with SMOTEENN.

To evaluate each combination of sampling method and machine learning model we reviewed the following metrics:

- Confusion matrix
- Classification report
- Accuracy score
- ROC curve
- Root mean squared error (RMSE)
- R-squared (R2)
- Mean Absolute Error (MAE)

Due to the class imbalance problem, we selected the SMOTEENN sampling method with Extra Trees Classifier model as this combination provided a good balance between the average precision and recall value for the minority class. Since our model is intended to be used by clinicians, we also wanted to balance model prediction accuracy and model interpretation. So, we decided against using Deep Neural Networks even though we were able to get more accurate predictions using the model.

Once we finalized the sampling and machine learning model combination, we used the correlation matrix to reduce the number of features. We dropped the features that were highly correlated with each other while selecting the features that had the strongest relationship with the target variable. We also used feature importance coefficients to further improve the model.

Our final model is 75% accurate, has a recall value of 0.68 for minority class, and takes 10 features to make predictions.

Visualizations with Tableau:

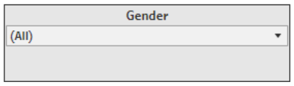
With using the powerful tool that is Tableau, we could take our data visualizations in many directions. We chose to do two dashboards. One focused on survival and comorbidities and the other focused on heart failure and patient statistics. Within the visualizations we also used the Essential Wellness Color Palette that was present throughout the project.

Dashboard 1- Hospital Mortality: Comorbidities

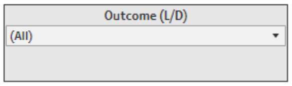
The dataset we used had nine comorbidities. What this dashboard presents is a look at the outcome of the patients care, gender, age groups, and the comorbidities. Why we chose to do this type of review was to show the correlations between having heart failure and other medical conditions and showing that relationship.

The struggle with this dashboard was presenting the comorbidity data in a meaningful and interesting way. With the amount of information within the dataset, it was important to show a set of visualizations that a user could use and interpret and pull relevant details out of. The comorbidity information was represented as a Boolean data type, so representing that visual became a small challenge to develop engaging visuals.

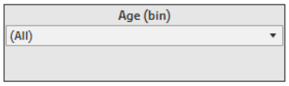
This dashboard will allow the user to filter each visual by gender, age group, and outcome of in hospital care (lived or died).



Gender
(All)



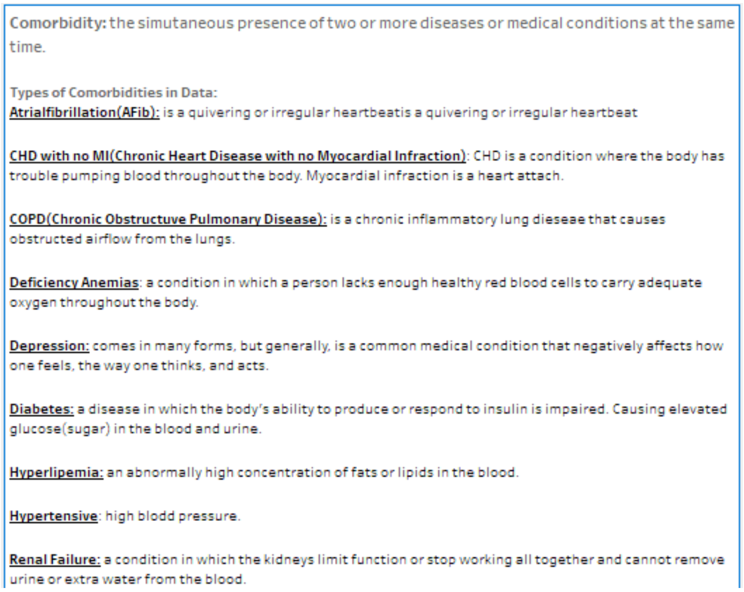
Outcome (L/D)
(All)



Age (bin)
(All)

We chose to present three separate visualizations on this relationship. The first being the "Comorbidities per Age Group." This data is represented in a bar chart.

To start we chose to bring in definitions of what a comorbidity is, and the definition of each one present in the dataset. Originally, we had one-third of the dashboard filled with these definitions, however this led to a wordy and crowded dashboard. We ended up changing these definitions into a pop-up. So, when the user of the dashboard wants to learn more then can click on a table on the dashboard.



Comorbidity: the simutaneous presence of two or more diseases or medical conditions at the same time.

Types of Comorbidities in Data:
Atrialfibrillation(AFib): is a quivering or irregular heartbeatis a quivering or irregular heartbeat

CHD with no MI(Chronic Heart Disease with no Myocardial Infraction): CHD is a condition where the body has trouble pumping blood throughout the body. Myocardial infraction is a heart attach.

COPD(Chronic Obstructuve Pulmonary Disease): is a chronic inflammatory lung dieseae that causes obstructed airflow from the lungs.

Deficiency Anemias: a condition in which a person lacks enough healthy red blood cells to carry adequate oxygen throughout the body.

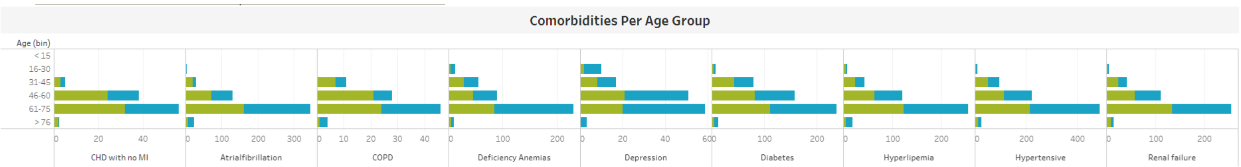
Depression: comes in many forms, but generally, is a common medical condition that negatively affects how one feels, the way one thinks, and acts.

Diabetes: a disease in which the body's ability to produce or respond to insulin is impaired. Causing elevated glucose(sugar) in the blood and urine.

Hyperlipemia: an abnormally high concentration of fats or lipids in the blood.

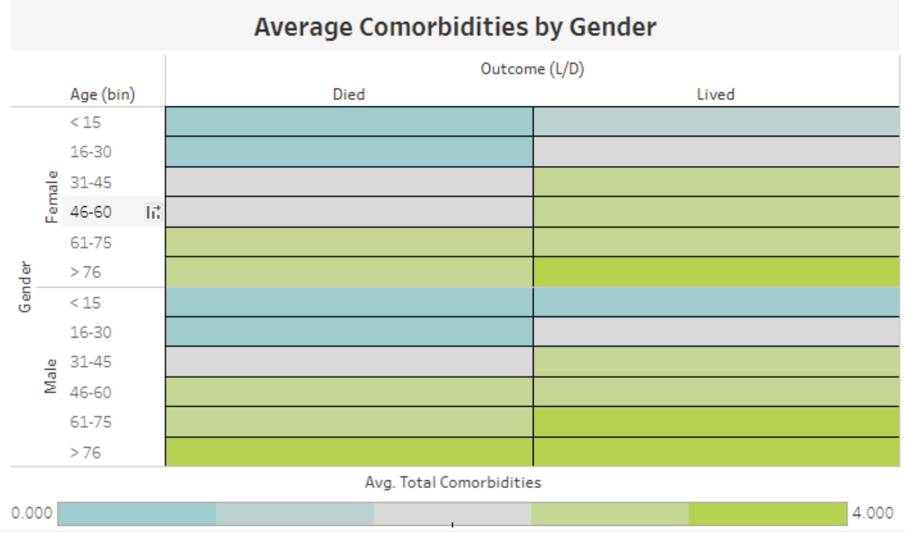
Hypertensive: high blodd pressure.

Renal Failure: a condition in which the kidneys limit function or stop working all together and cannot remove urine or extra water from the blood.



As shown in the visual, the older a patient was the more likely they would not survive to hospital discharge. The 61–75 year age range has the most deaths per comorbidity. The colors in the visual also represent the genders. Green represents male and blue represents females. The data shows that males did not survive while in care most of the time in the 61-75 age range per comorbidity being present.

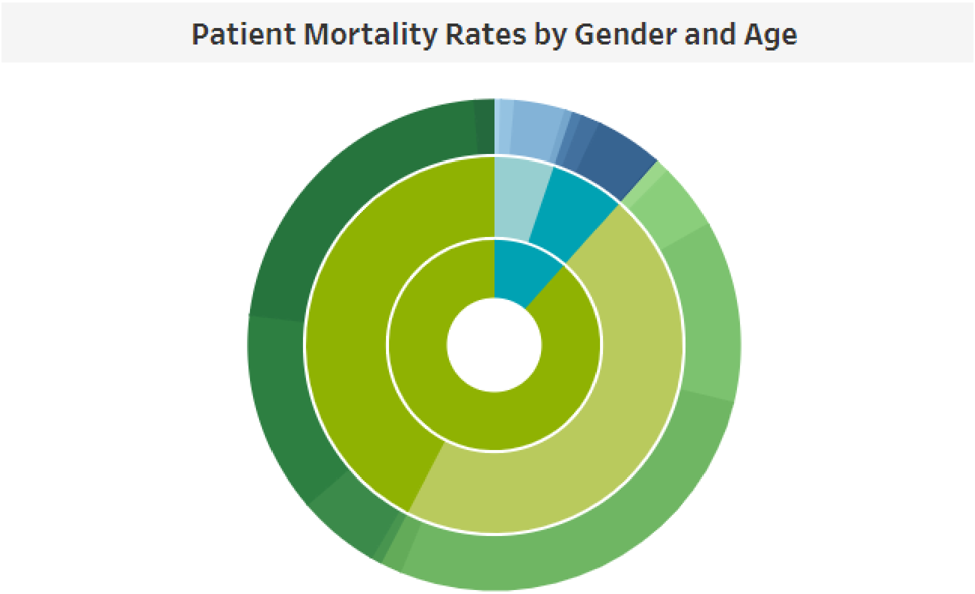
The second visual we chose to present was “Average Comorbidity by Gender.” This data is represented as a heatmap.



This review is represented in our color palette, with the darkest blue being low average, which is 0, and the darkest green being the largest average, which was 4. This visual was also split further by gender, outcome, and age group. The data presents in a way showing that the younger a patient is the lower the comorbidities present. The older the patient the more comorbidities diagnosed.

Furthermore, it shows us that the patients that did not survive, interestingly, had on average less comorbidities present.

The final visualization on this dashboard was “Patient Mortality Rates by Gender and Age.” This data was represented as a donut chart with three levels.


Patient Mortality Rates by Gender and Age

The color palette on this visual shows those patients that lived in green hues and those who did not survive to hospital discharge in blue hues. The first level, or the innermost ring, shows the outcome of survival to discharge by lived or died. Again, those who died are in blue. The second level, or middle ring, shows the outcome of survival by gender. The final level, or outermost ring, shows the outcome of survival to discharge by age group.

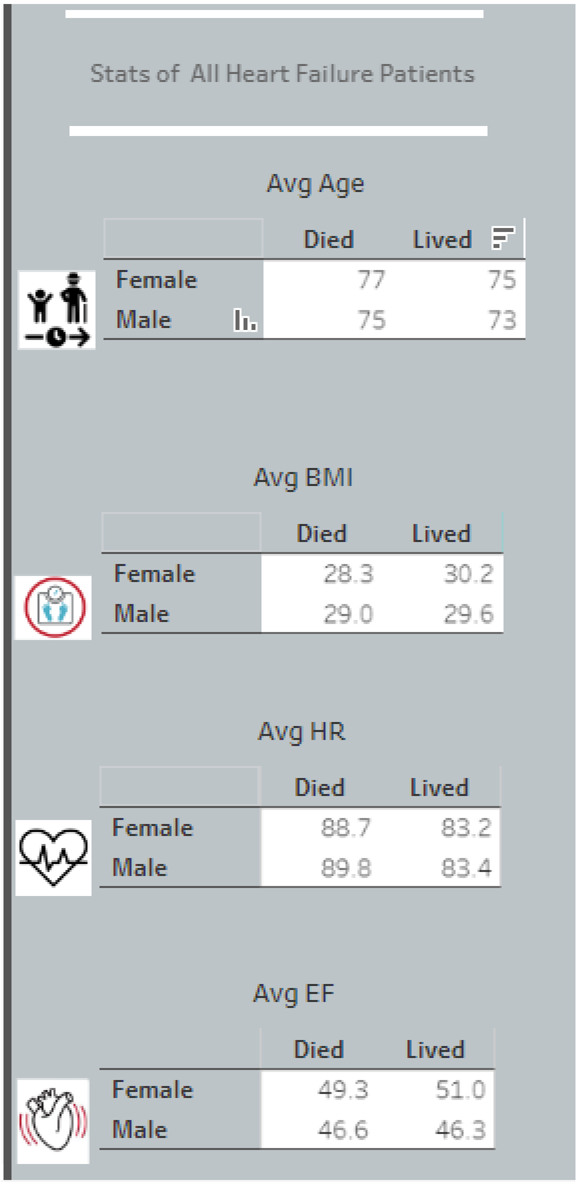
The data shows us that a majority of the patients survived to hospital discharge. Of those who did not survive, females perished more so than male. And finally, those age groups that are older represent the greater number of non-survival patients.

<u>Dashboard 2: Hospital Mortality in Heart Failure Patients</u>

The dataset had 40 plus columns of lab data, with that many datasets the ERD and machine learning part of the project were very important to determine what we could support our visualizations with.

The struggle that occurred while creating this dashboard was selecting just a couple of the key factors. Blood pressure was another factor that would have been good to display on this dashboard. The structure with the blood pressure graph was not configured correctly, so it was removed.

The first visual present called “Stats of Heart Failure Patients,” shows us data where you can see the total number of patients in the dataset and their average age, BMI (body mass index), heart rate, and EF (ejection fraction). The data was further displayed by those patients who lived or died within care.



Stats of All Heart Failure Patients

Avg Age

	Died	Lived
Female	77	75
Male	75	73

Avg BMI

	Died	Lived
Female	28.3	30.2
Male	29.0	29.6

Avg HR

	Died	Lived
Female	88.7	83.2
Male	89.8	83.4

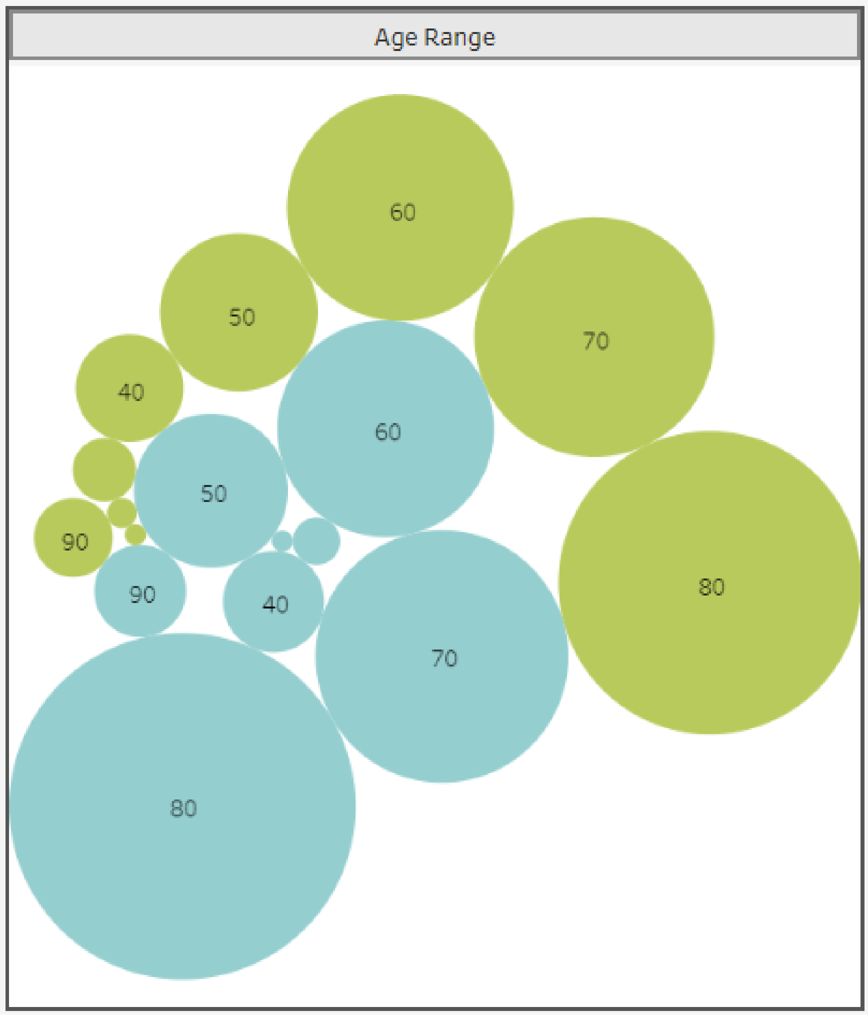
Avg EF

	Died	Lived
Female	49.3	51.0
Male	46.6	46.3

These are a couple of key factors that may play a part in the heart failure process. As can be seen with this visual, the average age per gender was fairly similar with males having a lower average age. BMI and heart rate averages for both genders were also similar as well. The only major difference was the EF, where males presented with lower amounts.

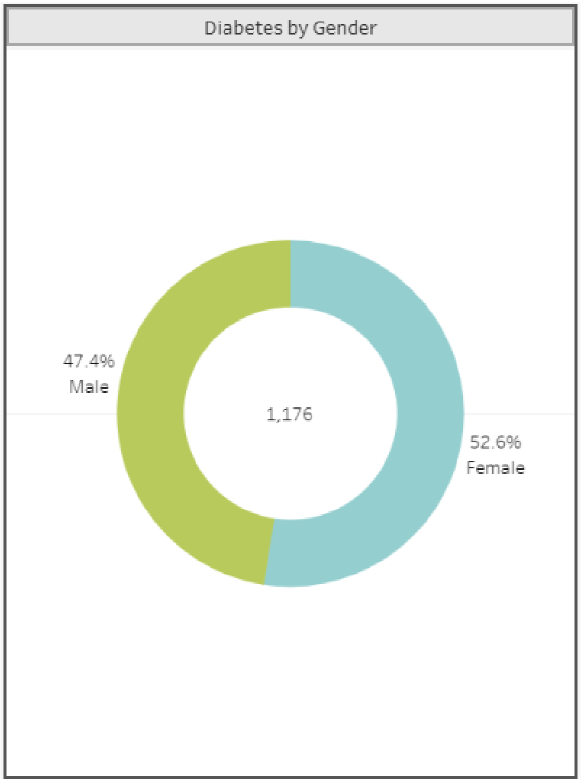
The remainder of the dashboard is filterable via age group, which are presented in increments of 10, as well as those patients that lived or died. This allows the dashboard to have a more detailed look at age groups instead of a broader view.

The second visual present is an “Age Group '' and is presented as a bubble graph.



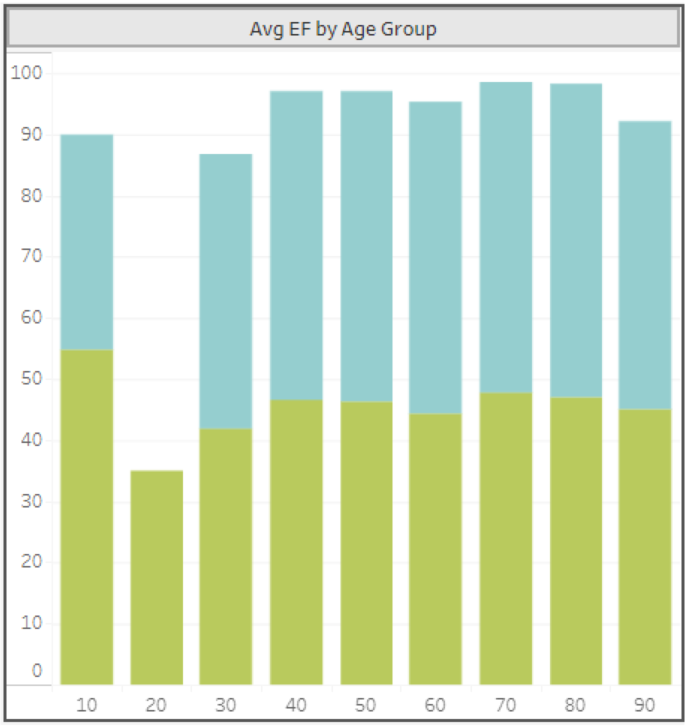
Using the color palette of essential wellness, blue represents female patients, and the green represents males. Each bubble presents a count of patients in each age range. This visualization allows us to see varying sizes of bubbles, with the larger the bubble the greater that age is present within the data. Here, it can be seen that both male and females with the age of 80 - 89 are represented to a greater extent within the dataset.

The third visual is a doughnut chart, “Diabetes by Gender,” that breaks down the gender of patients with diabetes.

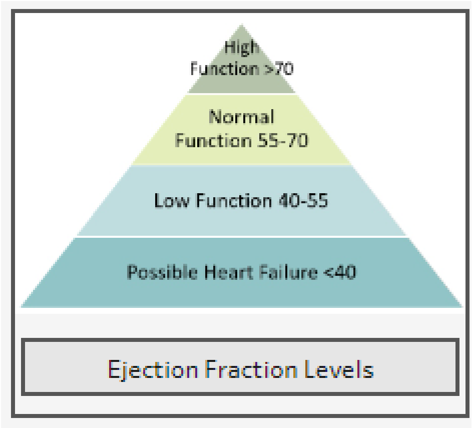


Within the dataset of 1176 patients, females make up a majority of diabetic patients with heart failure.

The third visual is a stacked bar graph called "Average EF by Age."

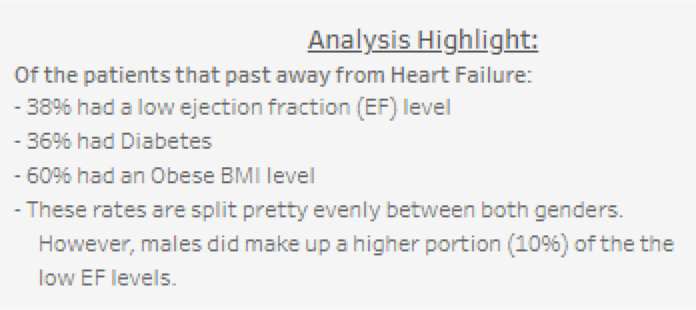


This captures the ejection fraction for both male/females within the age groups. The color is consistent with the other visuals. To help further aid the user an additional image was provided to contain a breakout of the EF levels. This image can help direct the user in interpreting the EF data of the patients.



With ejection fraction it is important to understand that the lower the number the worse off a patient's probability of survival is.

Lastly, an analysis highlight was added onto the dashboard to pull out the highlights of the data shown.



Analysis Highlight:

Of the patients that past away from Heart Failure:

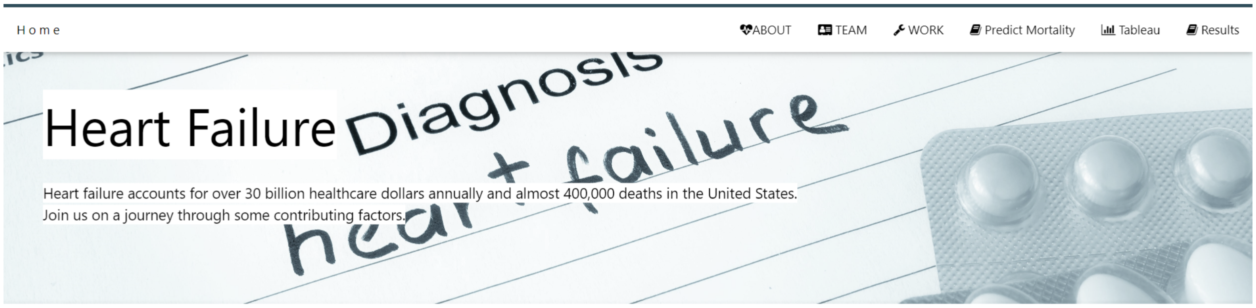
- 38% had a low ejection fraction (EF) level
- 36% had Diabetes
- 60% had an Obese BMI level
- These rates are split pretty evenly between both genders. However, males did make up a higher portion (10%) of the the low EF levels.

Web App Section:

The work that was put into the review, visualization, and prediction modules, needed a home to display the work and results that came from the In Hospital Heart Failure Prediction dataset.

We set out to develop a website that would be able to display all facets of the project. The website was originally developed through using a localized flask app, utilizing html, JavaScript, and CSS. With the ultimate goal of launching Heroku, which hosts this website for users to review the prediction and visualizations.

There are several sections within the report with a navigation bar at the top to aid the user in reviewing the full website and its different sections. Also within this section is a quick blurb about what is going to be presented. Also presented is the team itself.



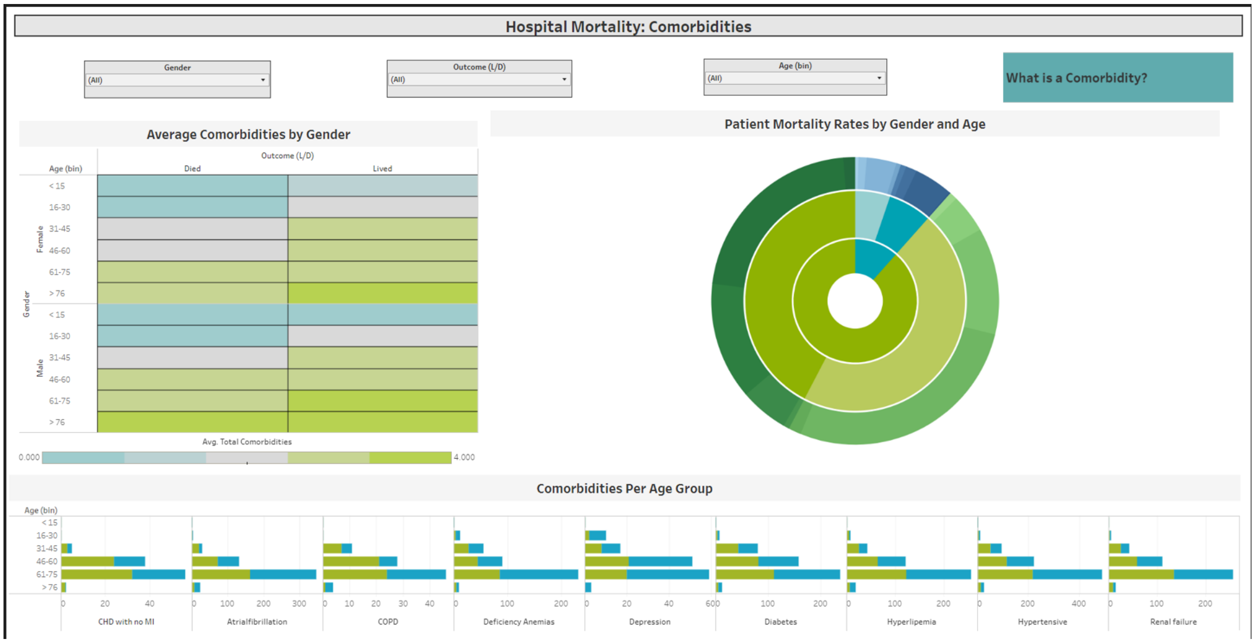
Home ABOUT TEAM WORK Predict Mortality Tableau Results

Heart Failure

Heart failure accounts for over 30 billion healthcare dollars annually and almost 400,000 deaths in the United States.
Join us on a journey through some contributing factors.

The user can scroll down through the presented information or use the navigation links on the top of the screen, both will provide the user with the same results.

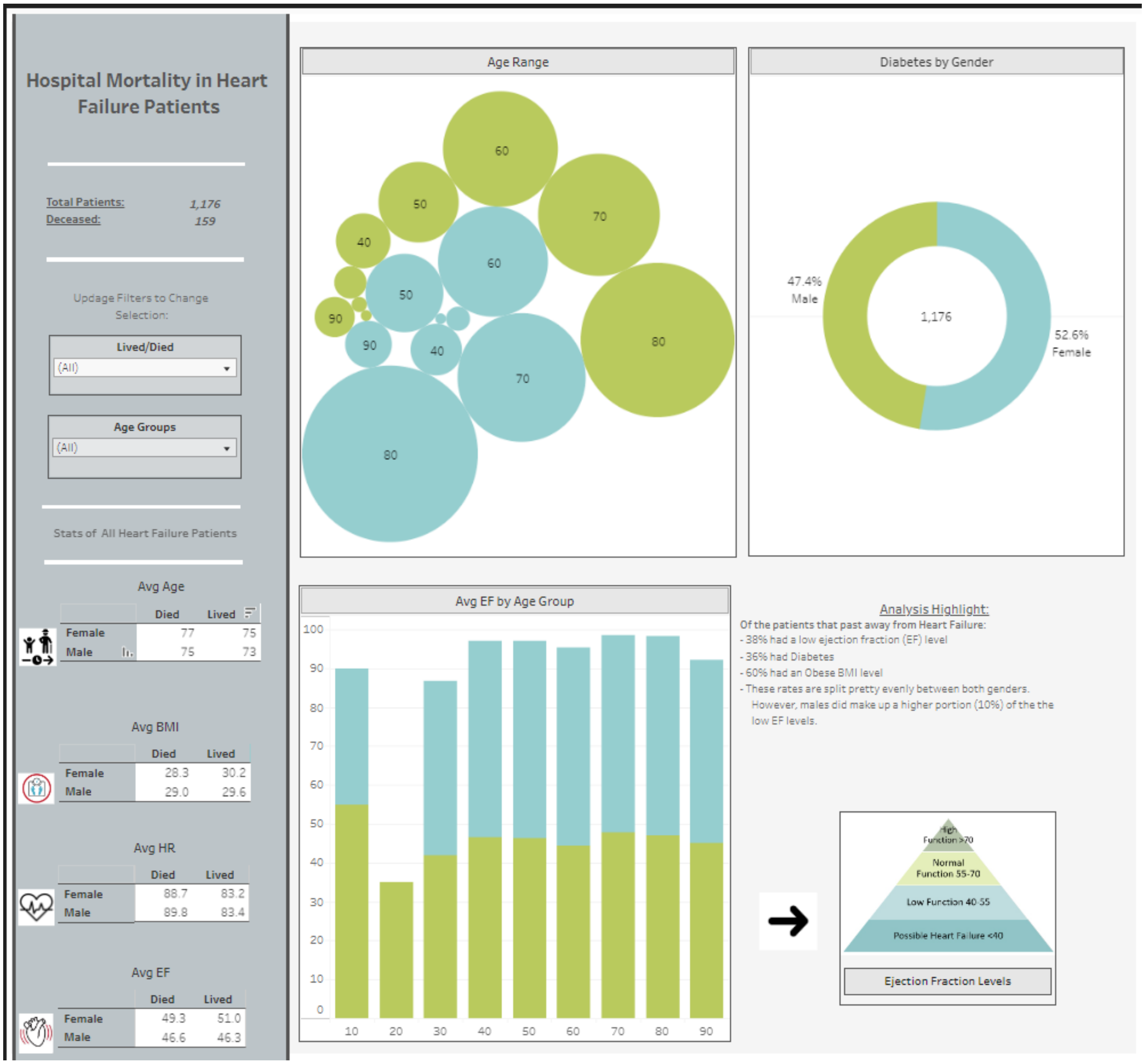
The Tableau section of the report allows the user to review several visualizations that were created to represent the dataset. The first being a dashboard showing the relationship between heart failures and comorbidities. This is a user-friendly report that is filterable on three aspects of the data, age, gender, and if the patient died within care or died.



Hospital Mortality: Comorbidities

Gender: (All)
Outcome (L/D): (All)
Age (bin): (All)
What is a Comorbidity?

The second Tableau dashboard allows the user to review the dataset through the review of statistics of the patient. For example, BMI, heart rate, ejection fraction, are present on the dashboard. To aid the user in analyzing the data the user can filter the entire dashboard by age and if the patient survived within hospital care or died.



Hospital Mortality in Heart Failure Patients

Total Patients: 1,176
Deceased: 159

Updage Filters to Change Selection:

Lived/Died: (All)

Age Groups: (All)

Stats of All Heart Failure Patients

Avg Age

	Died	Lived
Female	77	75
Male	75	73

Avg BMI

	Died	Lived
Female	28.3	30.2
Male	29.0	29.6

Avg HR

	Died	Lived
Female	88.7	83.2
Male	89.8	83.4

Avg EF

	Died	Lived
Female	49.3	51.0
Male	46.6	46.3

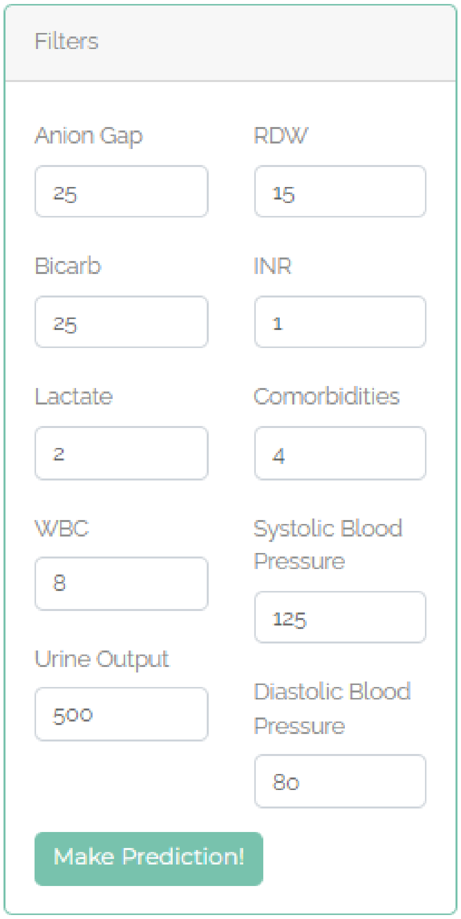
Analysis Highlight:

Of the patients that past away from Heart Failure:
- 38% had a low ejection fraction (EF) level
- 36% had Diabetes
- 60% had an Obese BMI level
- These rates are split pretty evenly between both genders. However, males did make up a higher portion (10%) of the the low EF levels.

The final major part of the website reviews the predictability of in hospital care of heart failure patients. The results of this part of the project is driven by the machine learning module developed early in the project development.

With this part of the project, we assume that the user would be a clinical individual, such as a doctor, nurse, resident, nurse practitioner, etc. The prediction tool uses patient statistics, such as comorbidities, blood pressure (systolic and diastolic), urine output, and other items that can be seen below.

The amounts listed in the filters already would be the average amount of a normally healthy patient. Once the user put in all the inputs the results will run and the prediction of survivability will be populated for the user to proceed with patient care.



Filters

Anion Gap: 25

RDW: 15

Bicarb: 25

INR: 1

Lactate: 2

Comorbidities: 4

WBC: 8

Systolic Blood Pressure: 125

Urine Output: 500

Diastolic Blood Pressure: 80

Conclusions/call to action:
Through the initial analysis of the dataset, visualizations through Tableau, and using machine learning techniques, we were able to provide the user of the website an in-depth review of in hospital care for heart failure and to predict if that patient would survive to discharge.

Within that review we could see that the older the patient is the more likely they would be to die within care. Adding onto that older patients typically have one or more comorbidities as well as heart failure. Women were more present in the data then men and may have more patients that did not survive to discharge, but this is a small data set when compared to what information could be added to this type of review. Men died in greater numbers at younger ages.

Through the machine learning module of SMOTEEN with Extra Trees Classifier, we were able to put together a prediction module that had the features within the dataset that correlated to our target variable of outcome.

Being able to predict mortality rates within hospital care would save the time of the doctors, nurses, and other hospital care staff. We are by no means saying that this prediction module would take the place of the much-needed care these patients need. However, as stated by the CDC, heart failure is a large issue within the medical community. With the 2010's having $6 million plus heart failure patients, with 400,000 of those patients passing in 2018, and the average yearly cost of this being greater than $30 million, being able to help alleviate the time, funding, and potential grief of those affected will be a greater problem as the population in America ages.

Limitations:

Data was incomplete for many patients. The missing data was present in several columns and not located solely in one column. If we removed all rows with NaN values, we would have lost more than 50% of the data. Which led the team to take the average of each column and impute in the NaN values with those average values. Linear regression was also employed on a few columns to fill in the NaN values after the exploratory database was first started. Using the imputation process allowed our team to use a majority of the data.

Another area in which we were hampered was determining what information out of the 51 columns of comorbidity and patient stats would we use in the project. The dataset was run through several regressions during the ERD process and the machine learning process to determine what stats or comorbidities had the greatest correlation to that of our target variable, the outcome.

Future work:

With unlimited time and resources, this work could continue to be a model used for frontline practitioners to focus on those patients with the highest risk of mortality in the hospital.

The dataset we used was limited to a little over 100 patients. Using a much larger database of patients this model could further the accuracy that we were able to develop within our models and could create a great baseline for developing a prediction tool of in hospital survivability of heart failure patients.

Other information that could be pulled in could be regional data, additional comorbidities to see if they correlate to survival rates, as well as natural language capability with patient records to determine if other non-lab work information is present, but not represented by standard testing.

References

CDC. (2020, September 8). *Heart Failure*. Retrieved from CDC.gov: https://www.cdc.gov/heartdisease/heart_failure.htm

Kaggle. (2021). In Hospital Mortality Prediction. Retrieved from Kaggle.com: https://www.kaggle.com/datasets/saurabhshahane/in-hospital-mortality-prediction