ILLINOIS DATA SCIENCE CLUB

Heart Disease Classification

DATA DIVE

By Team CWMDSJ

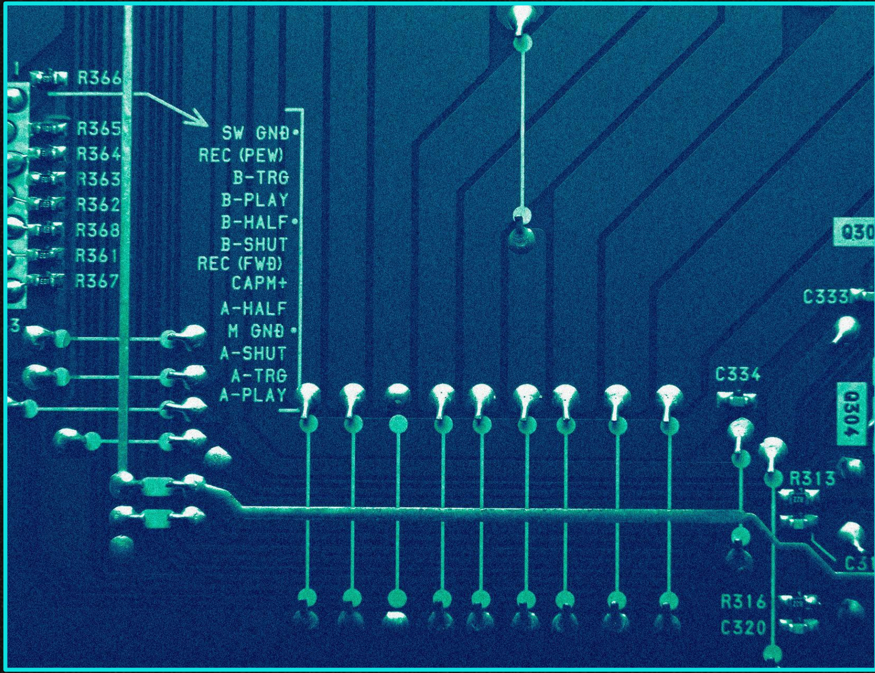


Table of contents

Problem Identification 01

Idk

Collecting Data 02

Idk

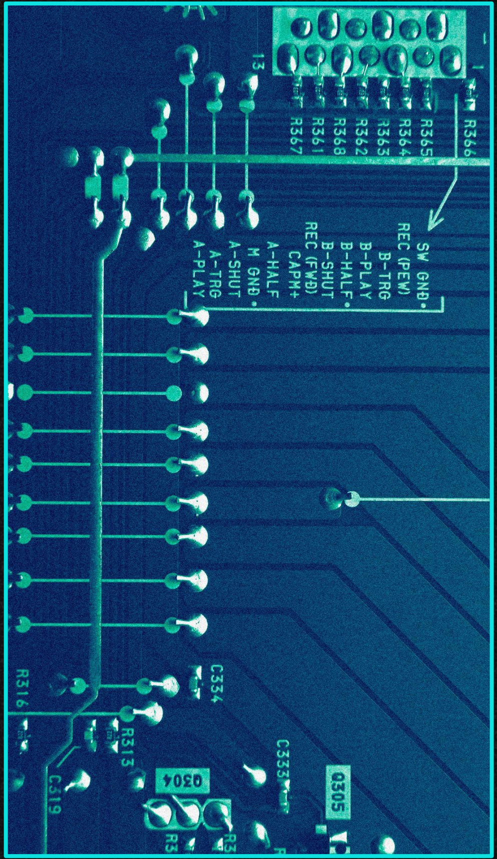
Processing Data 03

Idk

Modeling Data 04

idk

Data Analysis 05



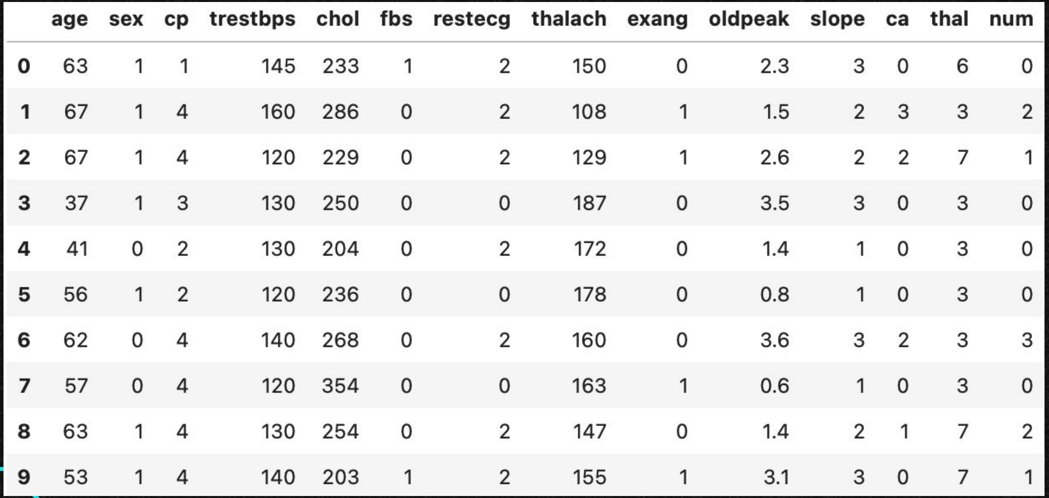
Problem Identification

- Why heart disease ?
    - The leading cause of death for both men and women in U.S.
    - Is heart disease preventable?

- To solve the questions? We should know:
    - What features are correlated with heart disease ?
    - How to classify the features that are correlated with heart disease?

Collecting Data

- Dataset from UC Irvine
  - Cleveland Database
- 14 Columns (more detail on next slide)
  - 13 possible explanatory variables
  - 1 response variable (num)



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
0	63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
1	67	1	4	160	286	0	2	108	1	1.5	2	3	3	2
2	67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
3	37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
4	41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
5	56	1	2	120	236	0	0	178	0	0.8	1	0	3	0
6	62	0	4	140	268	0	2	160	0	3.6	3	2	3	3
7	57	0	4	120	354	0	0	163	1	0.6	1	0	3	0
8	63	1	4	130	254	0	2	147	0	1.4	2	1	7	2
9	53	1	4	140	203	1	2	155	1	3.1	3	0	7	1

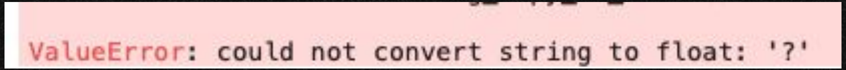
Column Descriptions

- Age - age in years
- Sex - (1 = male; 0 = female)
- CP - chest pain type
- Trestbps - resting blood pressure (in mm Hg on admission to the hospital)
- Chol - serum cholesterol in mg/dl
- FBS - (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- Restecg - resting electrocardiographic results
- Thalach - maximum heart rate achieved
- Exang - exercise induced angina (1 = yes; 0 = no)
- Oldpeak - ST depression induced by exercise relative to rest
- Slope - the slope of the peak exercise ST segment
- Ca - number of major vessels (0-3) colored by fluoroscopy
- Thal - 1 = normal; 2 = fixed defect; 3 = reversible defect
- Num - artery diameter (0-4)

Data Cleaning

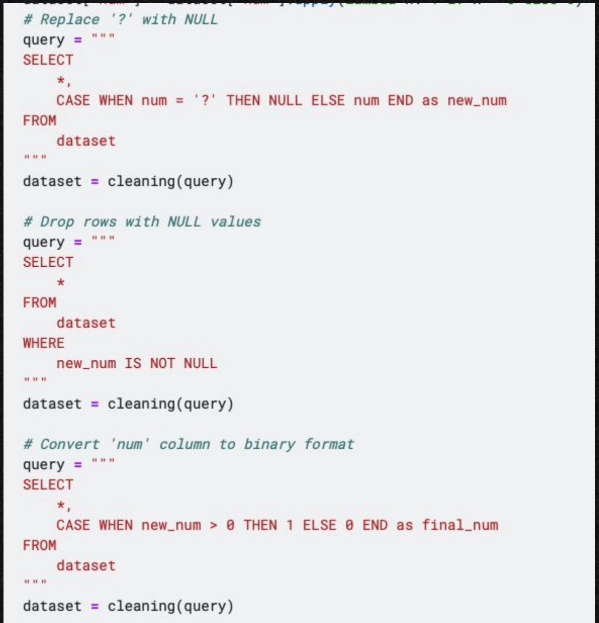
Our data was uncleaned

- Our data type was in the wrong format for the model
- We needed to get rid of categories showing little correlation
- We wanted a clear target - yes or no (1 or 0)



```
ValueError: could not convert string to float: '?'
```

Fixed Using SQLDF



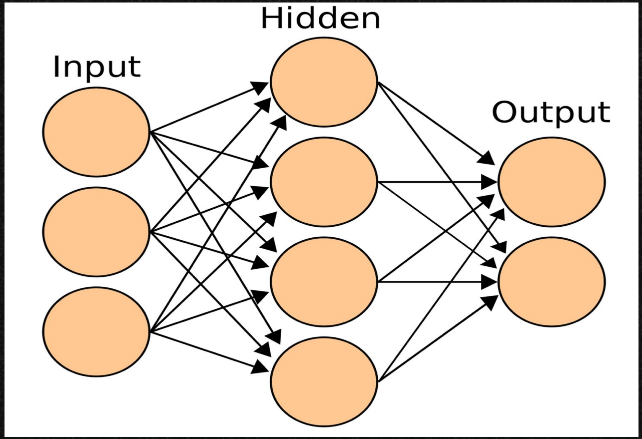
```
# Replace '?' with NULL
query = """
SELECT
    *,
    CASE WHEN num = '?' THEN NULL ELSE num END as new_num
FROM
    dataset
"""
dataset = cleaning(query)

# Drop rows with NULL values
query = """
SELECT
    *
FROM
    dataset
WHERE
    new_num IS NOT NULL
"""
dataset = cleaning(query)

# Convert 'num' column to binary format
query = """
SELECT
    *,
    CASE WHEN new_num > 0 THEN 1 ELSE 0 END as final_num
FROM
    dataset
"""
dataset = cleaning(query)
```

Machine Learning Model

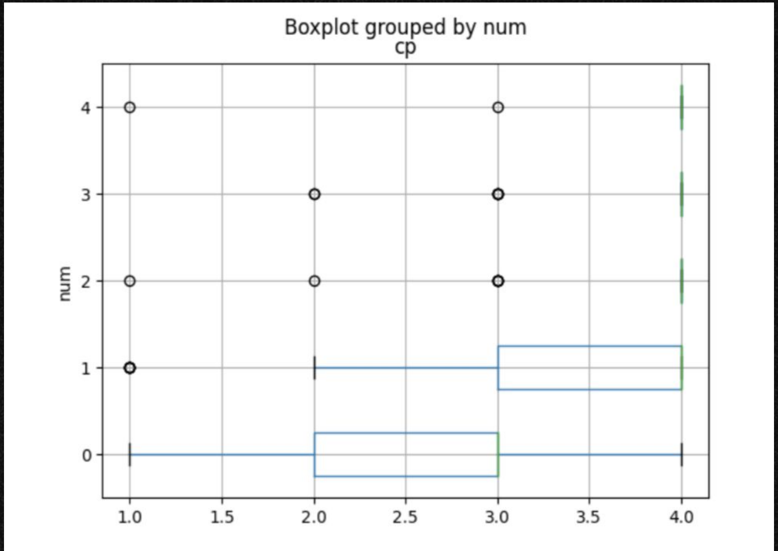
- The MLP we use has two hidden layers, with 32 neurons in the first layer and 16 neurons in the second layer.
- The activation function for the neurons is ReLU (Rectified Linear Unit), and the optimizer used for training the network is Adam.
- The model determines the weights of each layer based on previous iterations of epoch training.

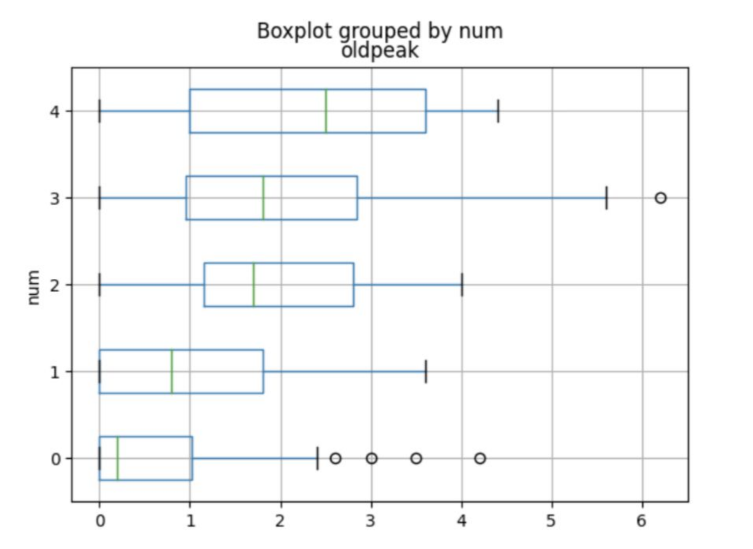


Exploratory Data Analysis

Question: Which categories show the highest correlation with the classification.

Approach: Visualize each category via a box plot.

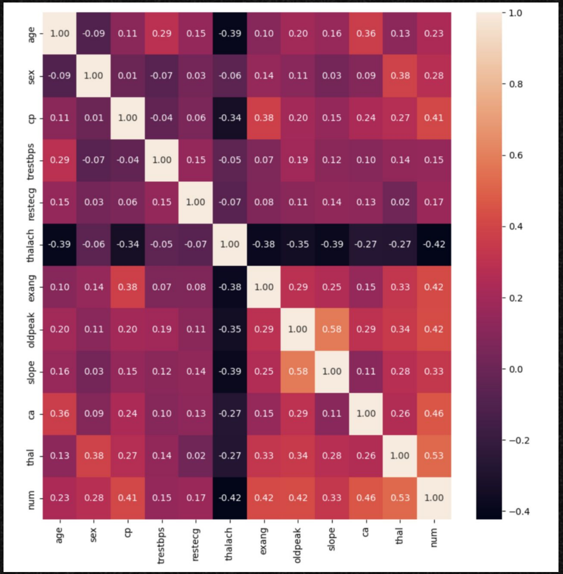




**num on the y axis is the target variable, showing severity of heart disease (0 being none, 4 being fatal)

Feature Analysis & Statistical Analysis

- Another tool we used to determine correlation between target and features was a correlation matrix
- Using this we could tell columns that are strongly correlated based on the heat map
- By removing features with little correlation, we were able to increase our accuracy by ~10%



	age	sex	cp	trestbps	restecg	thalach	exang	oldpeak	slope	ca	thal	num
age	1.00	-0.09	0.11	0.29	0.15	-0.39	0.10	0.20	0.16	0.36	0.13	0.23
sex	-0.09	1.00	0.01	-0.07	0.03	-0.06	0.14	0.11	0.03	0.09	0.38	0.28
cp	0.11	0.01	1.00	-0.04	0.06	-0.34	0.38	0.20	0.15	0.24	0.27	0.41
trestbps	0.29	-0.07	-0.04	1.00	0.15	-0.05	0.07	0.19	0.12	0.10	0.14	0.15
restecg	0.15	0.03	0.06	0.15	1.00	-0.07	0.08	0.11	0.14	0.13	0.02	0.17
thalach	-0.39	-0.06	-0.34	-0.05	-0.07	1.00	-0.38	-0.35	-0.39	-0.27	-0.27	-0.42
exang	0.10	0.14	0.38	0.07	0.08	-0.38	1.00	0.29	0.25	0.15	0.33	0.42
oldpeak	0.20	0.11	0.20	0.19	0.11	-0.35	0.29	1.00	0.58	0.29	0.34	0.42
slope	0.16	0.03	0.15	0.12	0.14	-0.39	0.25	0.58	1.00	0.11	0.28	0.33
ca	0.36	0.09	0.24	0.10	0.13	-0.27	0.15	0.29	0.11	1.00	0.26	0.46
thal	0.13	0.38	0.27	0.14	0.02	-0.27	0.33	0.34	0.28	0.26	1.00	0.53
num	0.23	0.28	0.41	0.15	0.17	-0.42	0.42	0.42	0.33	0.46	0.53	1.00

Project Takeaway

- Utilized Python libraries such as Pandas, NumPy, Matplotlib, and Seaborn for data preprocessing, cleaning, and visualization in a heart disease classification project built off a Kaggle available dataset
- Conducted data manipulation with SQLDF, handling missing values and transforming categorical variables into binary
- Leveraged Scikit-learn to implement a Neural Network model, achieving a 95% accuracy rate in heart disease prediction