Predicting Heart Disease


(News-Medical.Net, 2022)

Sarah Casauria
Nicholas Chatjaval
Michael Dunne
Ramana Ganesula
Tan Wei Wen


(Beckerman, 2021)

Introduction

Our focus for this project

- Understanding personal health predictors of heart disease
- Descriptive analysis of heart disease dataset
- Comparison of heart disease risk between a variety of demographic parameters
- Building machine learning model to predict heart disease risk

Data wrangling

Data Source

- Found dataset on Kaggle "Personal Key Indicators of Heart Disease"
  - ~300,000 rows, 18 variables

- Original data source from Center for Disease Control and Prevention (CDC) contained over 200 variables

Machine Learning Process

Machine Learning Overview

- Build, fit, and test different machine-learning models
  - Logistic regression and Random Forest

- Utilised various python modules
  - Scikit-learn
  - Pandas
  - Pickle
  - Imbalanced-learn
  - Matplotlib and Seaborn

Data Preprocessing

- Lots of categorical data

- Reduced categorical options to avoid overfitting

```python
# Reduce the number of age categories to 3 and the number of diabetic categories to 2
age_cats = ["18 - 34", "35 - 64", "65 or older"]
df.replace({'AgeCategory' : {"18-24" : age_cats[0],
                             "25-29" : age_cats[0],
                             "30-34" : age_cats[0],
                             "35-39" : age_cats[1],
                             "40-44" : age_cats[1],
                             "45-49" : age_cats[1],
                             "50-54" : age_cats[1],
                             "55-59" : age_cats[1],
                             "60-64" : age_cats[1],
                             "65-69" : age_cats[2],
                             "70-74" : age_cats[2],
                             "75-79" : age_cats[2],
                             "80 or older" : age_cats[2]}}, inplace = True)

df.head()
```

	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Race	Diabetic	PhysicalActivit
0	No	16.60	Yes	No	No	3.0	30.0	No	Female	35 - 64	White	Yes	Ye
1	No	20.34	No	No	Yes	0.0	0.0	No	Female	65 or older	White	No	Ye
2	No	26.58	Yes	No	No	20.0	30.0	No	Male	65 or older	White	Yes	Ye
3	No	24.21	No	No	No	0.0	0.0	No	Female	65 or older	White	No	N
4	No	23.71	No	No	No	28.0	0.0	Yes	Female	35 - 64	White	No	Ye

Data Preprocessing

- Reduced dataset from 18 variables to 8 key variables of interest:
  - Age Category (3 options)
  - Sex (2 options)
  - General Health (5 options)
  - Smoking (2 options)
  - Diabetes (2 options)
  - Alcohol drinking (2 options)
  - Stroke (2 options)
  - Kidney disease (2 options)

Data Preprocessing

- Adapted modified LabelEncoder[1] method to convert all categorical variables to numerical values

Use LabelEncoder to encode the categorical variables

```
# Use LabelEncoder to encode the yes/no columns to 1/0
cat_columns = ["HeartDisease","Smoking", "AlcoholDrinking", "Stroke",
               "DiffWalking", "PhysicalActivity", "Asthma",
               "KidneyDisease","SkinCancer", "Diabetic", "AgeCategory","Race","GenHealth", "Sex"]

le=LabelEncoder()

encoded_df = df.copy()

# code snippet adapted from https://gsarantitis.wordpress.com/2019/07/16/how-to-persist-categorical-encoding-in-machine
dict_all = dict(zip([], []))

for col in cat_columns:
    temp_keys = encoded_df[col].values
#     print(temp_keys)
    temp_values = le.fit_transform(encoded_df[col])
#     print(temp_values)
    dict_temp = dict(zip(temp_keys, temp_values))
#     print(dict_temp)
    dict_all[col] = dict_temp
#     print(dict_all[col])

# print(dict_all['HeartDisease'])

for col in cat_columns:
    encoded_df.replace(dict_all[col], inplace=True)

# encoded_df.replace(dict_all[col], inplace=True)

encoded_df.head()
```

6]:

	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Race	Diabetic	PhysicalA
0	0	16.60	1	0	0	3.0	30.0	0	0	1	5	1	
1	0	20.34	0	0	1	0.0	0.0	0	0	2	5	0	
2	0	26.58	1	0	0	20.0	30.0	0	1	2	5	1	
3	0	24.21	0	0	0	0.0	0.0	0	0	2	5	0	
4	0	23.71	0	0	0	28.0	0.0	1	0	1	5	0	

[1] https://gsarantitis.wordpress.com/2019/07/16/how-to-persist-categorical-encoding-in-machine-learning-deployment-phase/

Data preprocessing

- StandardScaler to standardise the data

- Random undersampling to correct imbalanced dataset

Scale the data using StandardScaler

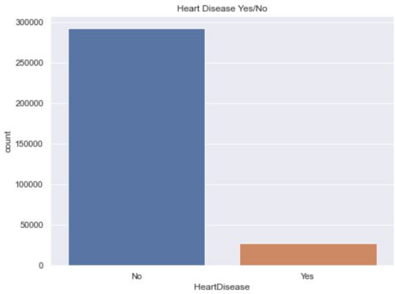
In [13]:
```
scaler = StandardScaler()
```

In [14]:
```
# Transform the training and testing data to the scaler
X_train = scaler.fit_transform(X_train)

X_test = scaler.fit_transform(X_test)
```

Use Random Undersampling to balance the data between heart disease sample and non heart disease sample

We do this because ~90% of the total data is classified as "No Heart Disease" and only ~10% is classified as "Yes Heart Disease". Since the original dataset is imbalanced, it is good practice to undersample the larger dataset to match the number of cases in the smaller dataset

In [15]:
```
# Plot the distribution of heart disease using sns countplot
sns.set(rc = {'figure.figsize':(8,6)})
sns.countplot(x="HeartDisease", data = df).set(title="Heart Disease Yes/No")
plt.show()
```



In [16]:
```
rus = RandomUnderSampler()
X_rus_train, y_rus_train = rus.fit_resample(X_train, y_train)

## We don't resample the testing data
print(X_rus_train.shape)
print(y_rus_train.shape)
```

```
(41060, 8)
(41060,)
```

Model Comparison

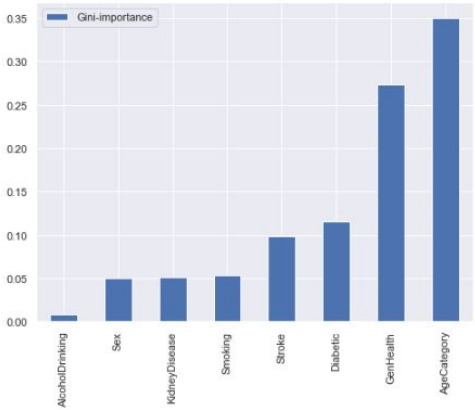
Random Forest

```python
# Score the training and testing data
print(f"Training Data Score: {rf.score(X_rus_train, y_rus_train)}")
print(f"Testing Data Score: {rf.score(X_test, y_test)}")
```

```
Training Data Score: 0.7589624939113493
Testing Data Score: 0.7145305132021664
```

```python
feats = {} # a dict to hold feature_name: feature_importance
for feature, importance in zip(X.columns, rf.feature_importances_):
    feats[feature] = importance #add the name/value pair

importances = pd.DataFrame.from_dict(feats, orient='index').rename(columns={0: 'Gini-importance'})
importances.sort_values(by='Gini-importance').plot(kind='bar', rot=90)
```

```
<AxesSubplot:>
```

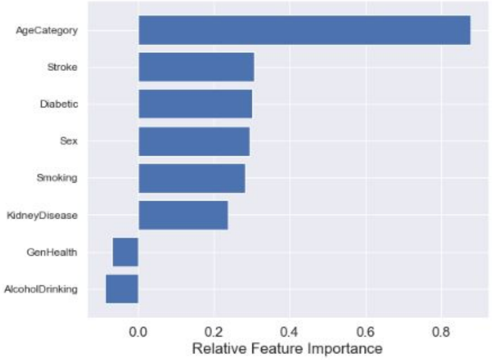


Logistic Regression

```python
# Print the r2 score for the test data
print(f"Training Data Score: {lg.score(X_train, y_train)}")
print(f"Testing Data Score: {lg.score(X_test, y_test)}")
```

```
Training Data Score: 0.7348298491532066
Testing Data Score: 0.7311285944789804
```

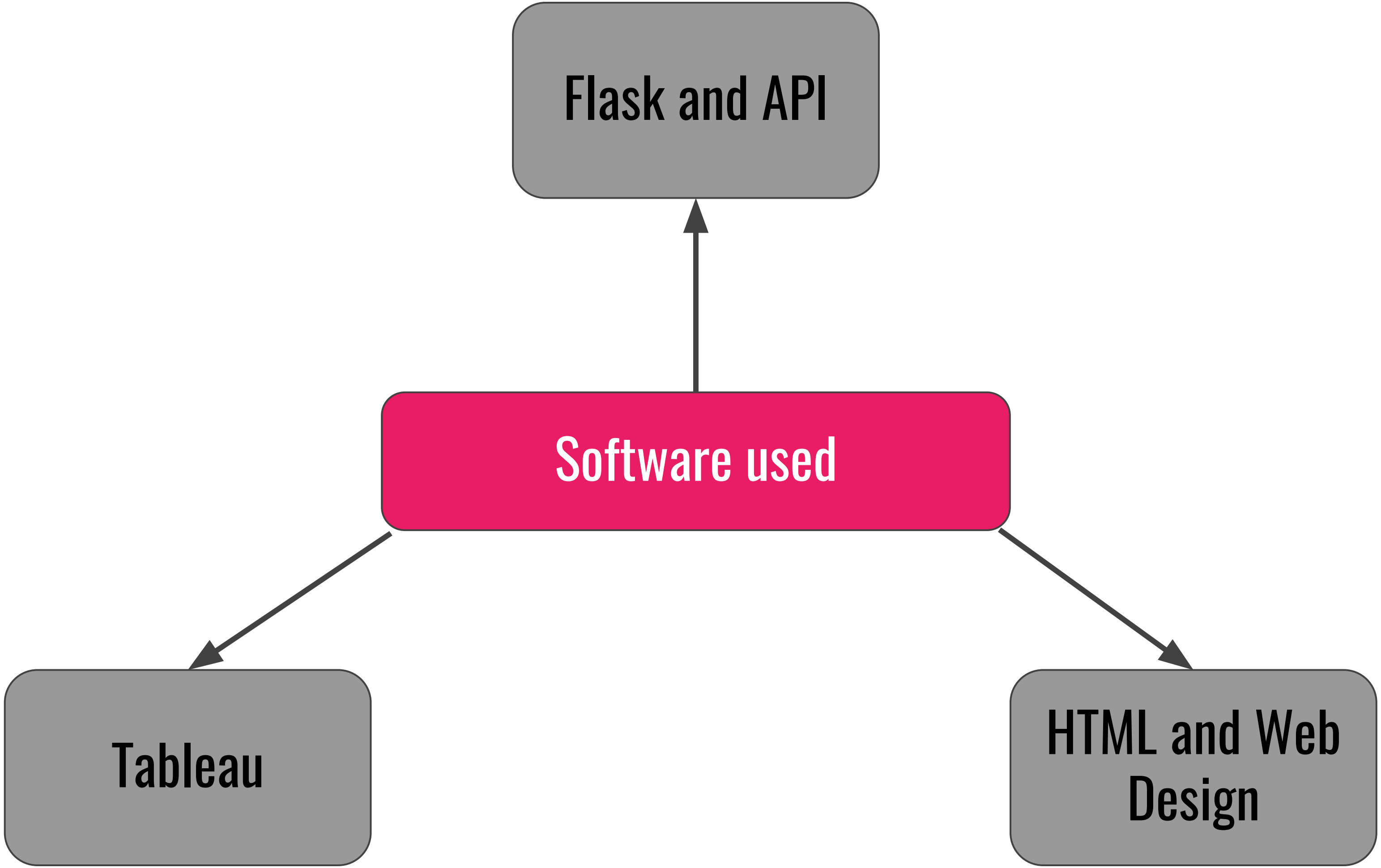
```python
feature_importance = (lg.coef_[0])
sorted_idx = np.argsort(feature_importance)
pos = np.arange(sorted_idx.shape[0]) + .5

featfig = plt.figure()
featax = featfig.add_subplot(1, 1, 1)
featax.barh(pos, feature_importance[sorted_idx], align='center')
featax.set_yticks(pos)
featax.set_yticklabels(np.array(X.columns)[sorted_idx], fontsize=12)
featax.set_xlabel('Relative Feature Importance')

plt.tight_layout()
plt.show()
```

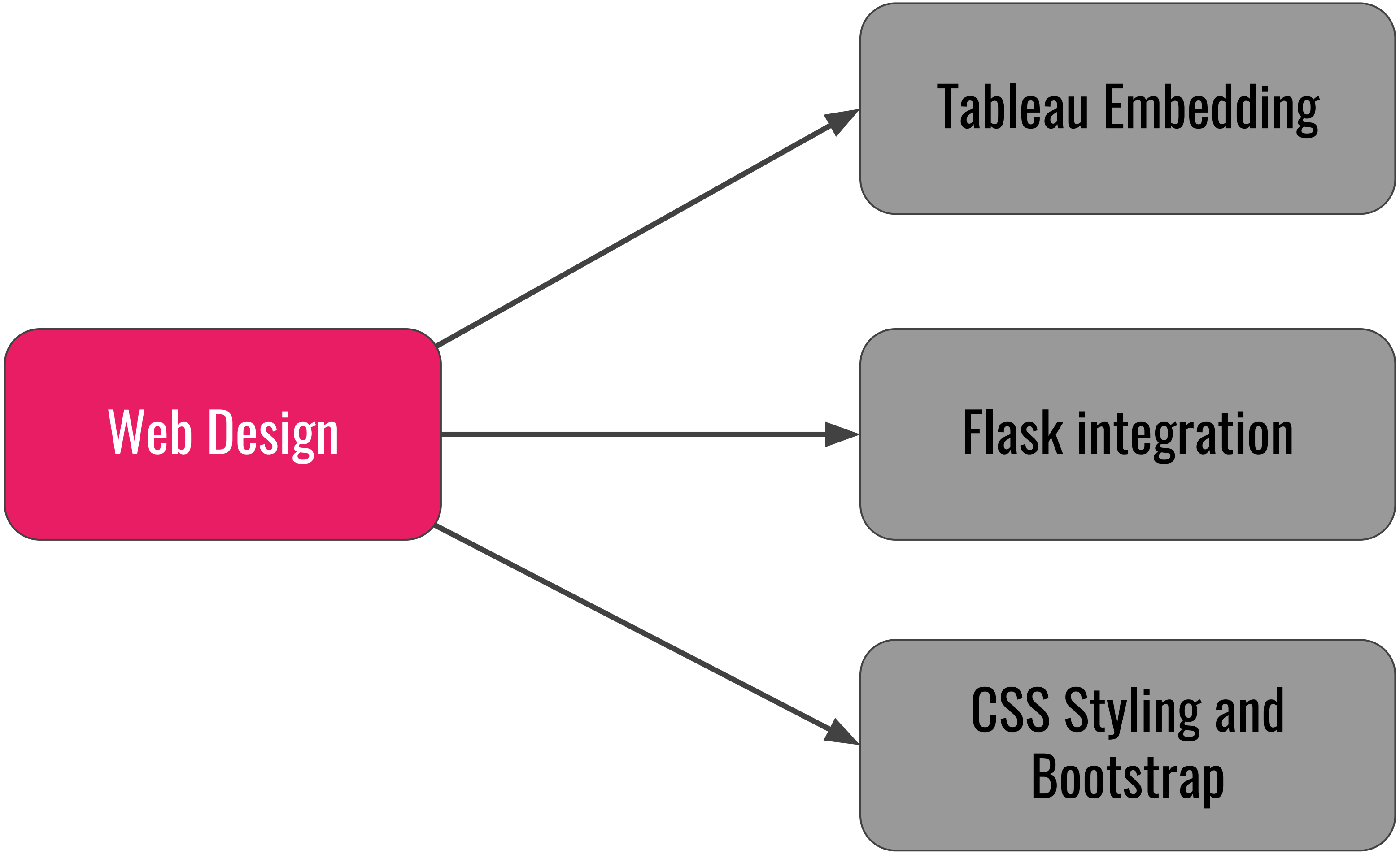


Machine Learning Conclusion

- Logistic regression model more accurate than random forest
- Saved logistic regression model, label encoder dictionary, and standard scaler for flask integration

Software and HTML Approach


Flask and API
Software used
Tableau
HTML and Web Design

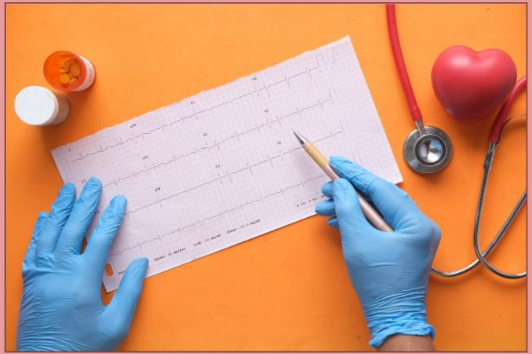

Web Design
Tableau Embedding
Flask integration
CSS Styling and Bootstrap

Web Design

Predictors of Heart Disease Home Predictor Tool Visualisations ▾ GitHub Repository

Predictors of Heart Disease

An Analysis on the impacts of heart disease on others, including a variety of different factors.



Background

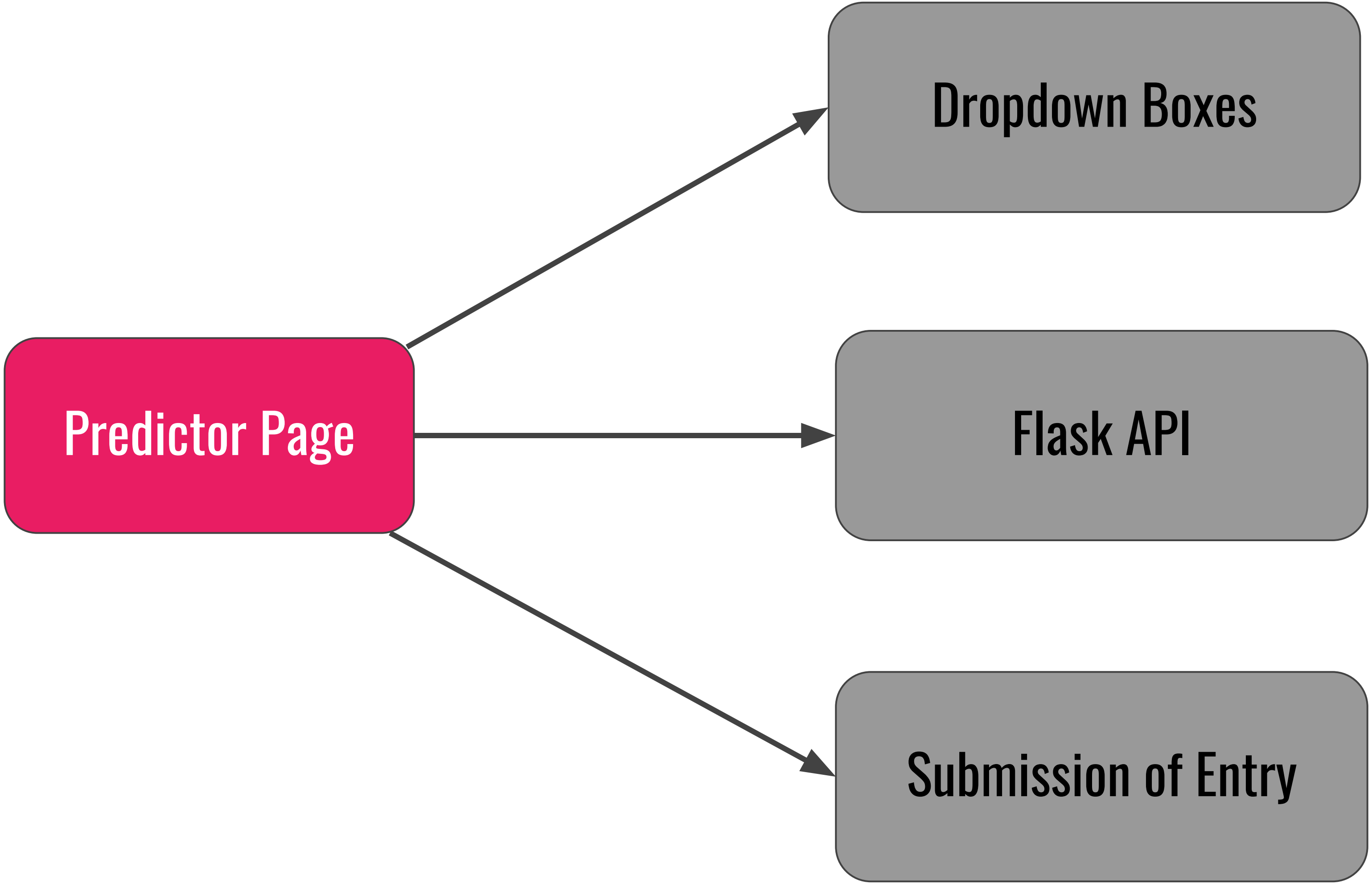
Heart disease is one of the common diseases that caused many deaths around the world, in which are varied between a variety of demographics. This project discusses how heart disease has been predicted based on a variety of different activities that people have engaged in, as well as a variety of different health factors.

Project Purpose

To build a machine-learning model that can predict heart disease risk using different health parameters.

Methods

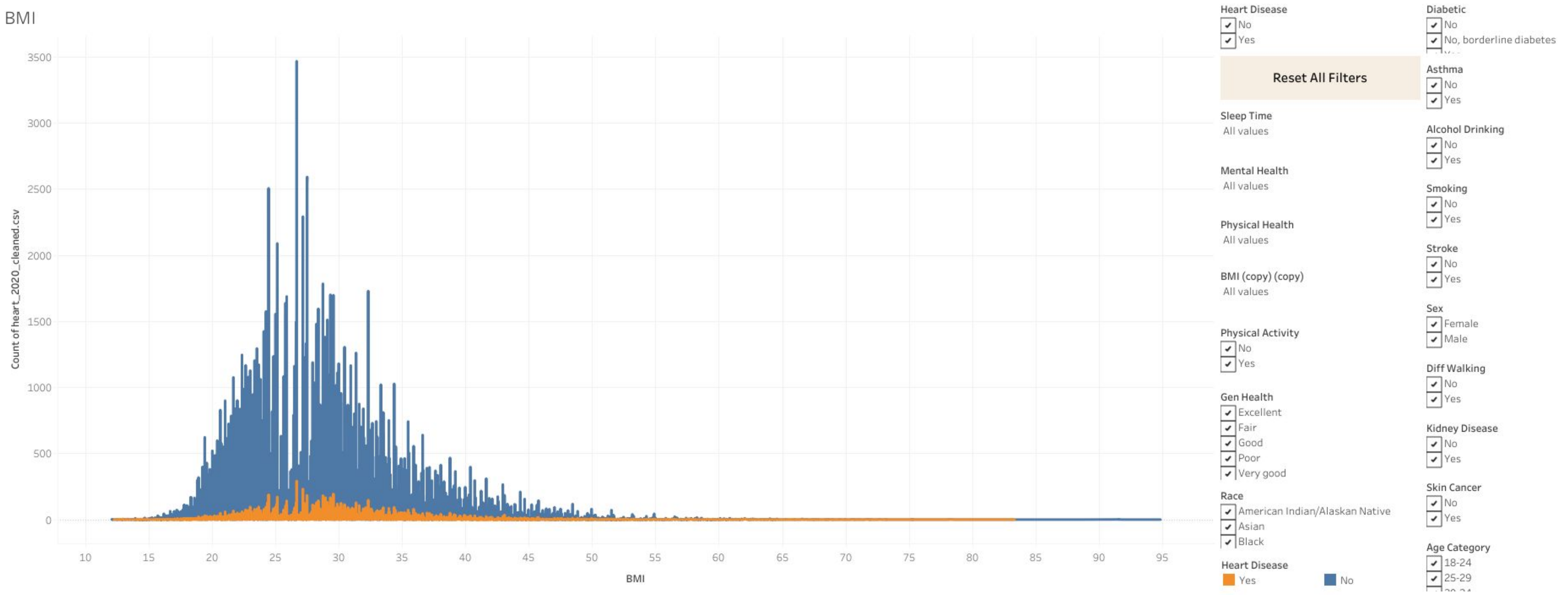
For the machine-learning model, we used a dataset containing survey responses from over 300,000 people in the United States. Respondents were surveyed on various health parameters, including BMI, Age, Physical activity, General health, and various medical conditions.
This data can be freely accessed on Kaggle. The Kaggle data is based on an original, larger dataset available on the Centers for Disease Control and Prevention (CDC) website.
We reduced the dataset down to the following 8 parameters that we wished to use for our predictor tool:


Predictor Page
Dropdown Boxes
Flask API
Submission of Entry

Predictor Page

Visual Analyses

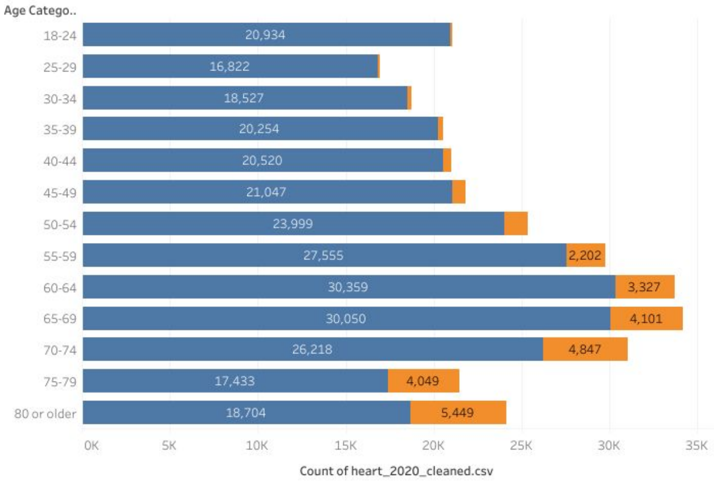
Analysis 1



BMI
Count of heart_2020_cleaned.csv
3500
3000
2500
2000
1500
1000
500
0
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
BMI
Heart Disease
No
Yes
Reset All Filters
Sleep Time
All values
Mental Health
All values
Physical Health
All values
BMI (copy) (copy)
All values
Physical Activity
No
Yes
Gen Health
Excellent
Fair
Good
Poor
Very good
Race
American Indian/Alaskan Native
Asian
Black
Heart Disease
Yes
No
Diabetic
No
No, borderline diabetes
Asthma
No
Yes
Alcohol Drinking
No
Yes
Smoking
No
Yes
Stroke
No
Yes
Sex
Female
Male
Diff Walking
No
Yes
Kidney Disease
No
Yes
Skin Cancer
No
Yes
Age Category
18-24
25-29

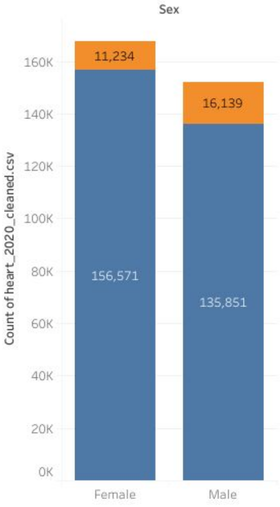
Analysis 2

Age



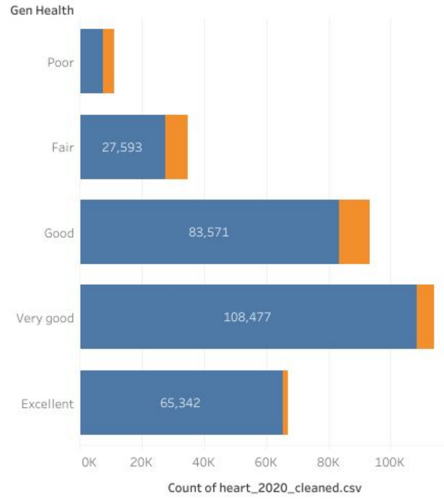
	Age Category										
Heart Disease	18-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74
Yes	130	133	226	296	486	744	1,383	2,202	3,327	4,101	4,84
No	20,934	16,822	18,527	20,254	20,520	21,047	23,999	27,555	30,359	30,050	26,21

Gender



	Sex	
Heart Disea..	Female	Male
Yes	11,234	16,139
No	156,571	135,851

General Health



	Gen Health				
Heart Disea..	Poor	Fair	Good	Very good	Excellent
Yes	3,850	7,084	9,558	5,381	1,500
No	7,439	27,593	83,571	108,477	65,342

Heart Disease
- No
- Yes

Reset All Filters

Sleep Time
All values

Mental Health
All values

Physical Health
All values

BMI (copy) (copy)
All values

Physical Activity
- No
- Yes

Gen Health
- Excellent
- Fair
- Good
- Poor

Race
- American Indian/Alaskan Native
- Asian
- Black
- Hispanic

Heart Disease
- No
- Yes

Diabetic
- No
- No, borderline diabetes

Asthma
- No
- Yes

Alcohol Drinking
- No
- Yes

Smoking
- No
- Yes

Stroke
- No
- Yes

Sex
- Female
- Male

Diff Walking
- No
- Yes

Kidney Disease
- No
- Yes

Skin Cancer
- No
- Yes

Age Category
- 18-24
- 25-29

Analysis 3

Difficulty Walking

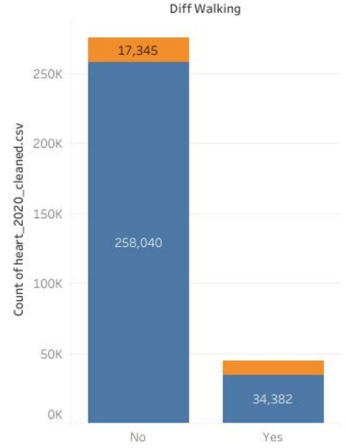
Heart Disea..	Diff Walking: Yes	Diff Walking: No
Yes	10,028	17,345
No	34,382	258,040

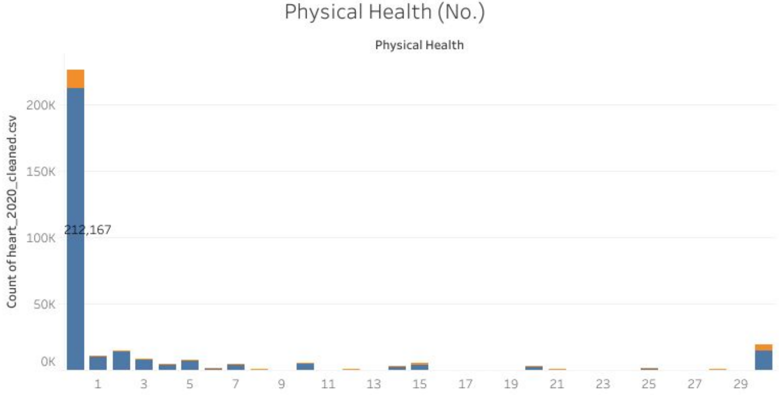
Physical Health

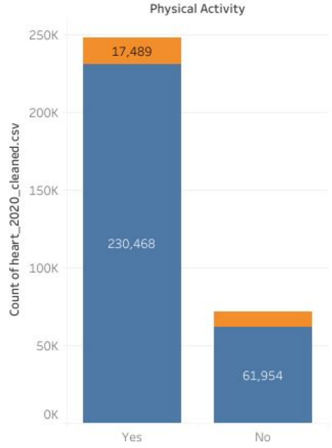
&

Pyshical Activity (Yes/No)

Heart Disea..	Physical Activity: Yes	Physical Activity: No
Yes	17,489	9,884
No	230,468	61,954







Physical Health (No.)

Physical Health

Hear..	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Yes	14,422	605	1,169	843	494	896	173	465	120	37	838	9	104	10	312	930	21	21	33	9	641	99	15	
No	212,167	9,884	13,711	7,774	3,974	6,710	1,097	4,164	804	143	4,615	76	501	81	2,581	4,082	114	89	134	26	2,575	527	74	

Heart Disease
- [x] No
- [x] Yes

Reset All Filters

Sleep Time
All values

Mental Health
All values

Physical Health
All values

BMI (copy) (copy)
All values

Physical Activity
- [x] No
- [x] Yes

Gen Health
- [x] Excellent
- [x] Fair
- [x] Good
- [x] Poor

Race
- [x] American Indian/Alaskan Native
- [x] Asian
- [x] Black
- [x] Hispanic

Heart Disease
Yes No

Diabetic
- [x] No
- [x] No, borderline diabetes
- [x] Yes

Asthma
- [x] No
- [x] Yes

Alcohol Drinking
- [x] No
- [x] Yes

Smoking
- [x] No
- [x] Yes

Stroke
- [x] No
- [x] Yes

Sex
- [x] Female
- [x] Male

Diff Walking
- [x] No
- [x] Yes

Kidney Disease
- [x] No
- [x] Yes

Skin Cancer
- [x] No
- [x] Yes

Age Category
- [x] 18-24
- [x] 25-29
- [x] 30-34

Total Percent Analysis

High risk variables

- Brain Stroke - 36.4%
- General Health (Poor) - 34.1%
- Kidney Disease - 29.3%
- Diabetes - 22%
- Difficulty Walking - 22.58%

Low risk variables

- Alcohol - 5.24%
- Mental Health (30) - 13.16%

	Stroke	
Heart Disea..	Yes	No
Yes	4,389	27,373
No	7,680	292,422

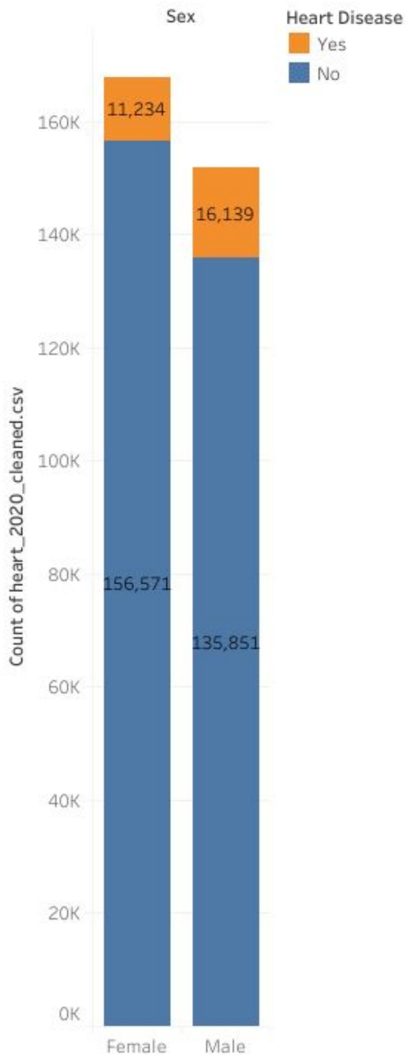
	Gen Health				
Heart Disea..	Poor	Fair	Good	Very good	Excellent
Yes	3,850	7,084	9,558	5,381	1,500
No	7,439	27,593	83,571	108,477	65,342

	Alcohol Drinking	
Heart Disea..	Yes	No
Yes	1,141	26,232
No	20,636	271,786

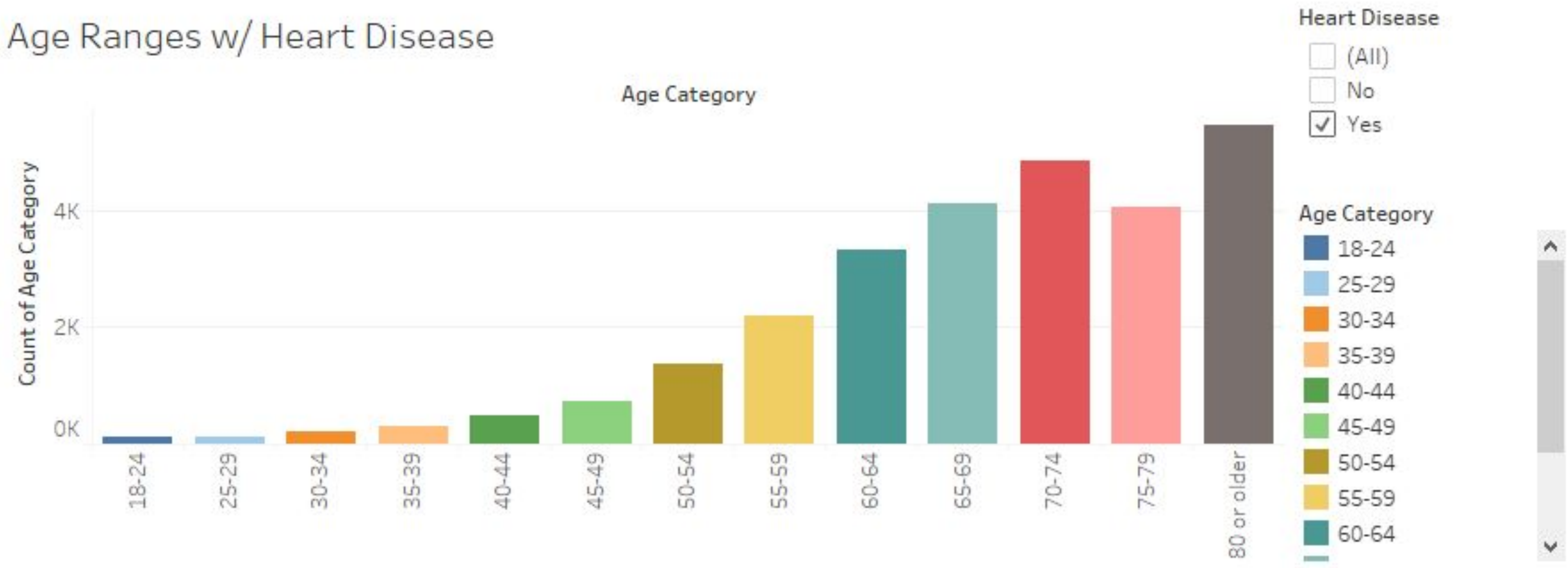
Gender Analysis

	Sex	
Heart Disea..	Female	Male
Yes	11,234	16,139
No	156,571	135,851

- Male Heart disease Risk at 10.6%, while Female heart disease risk was at 6.7%.
- Males have high cases of kidney disease, diabetes and high BMI average leading to heart disease risk.
- More data and research is needed to find why Male have high health issues than females.

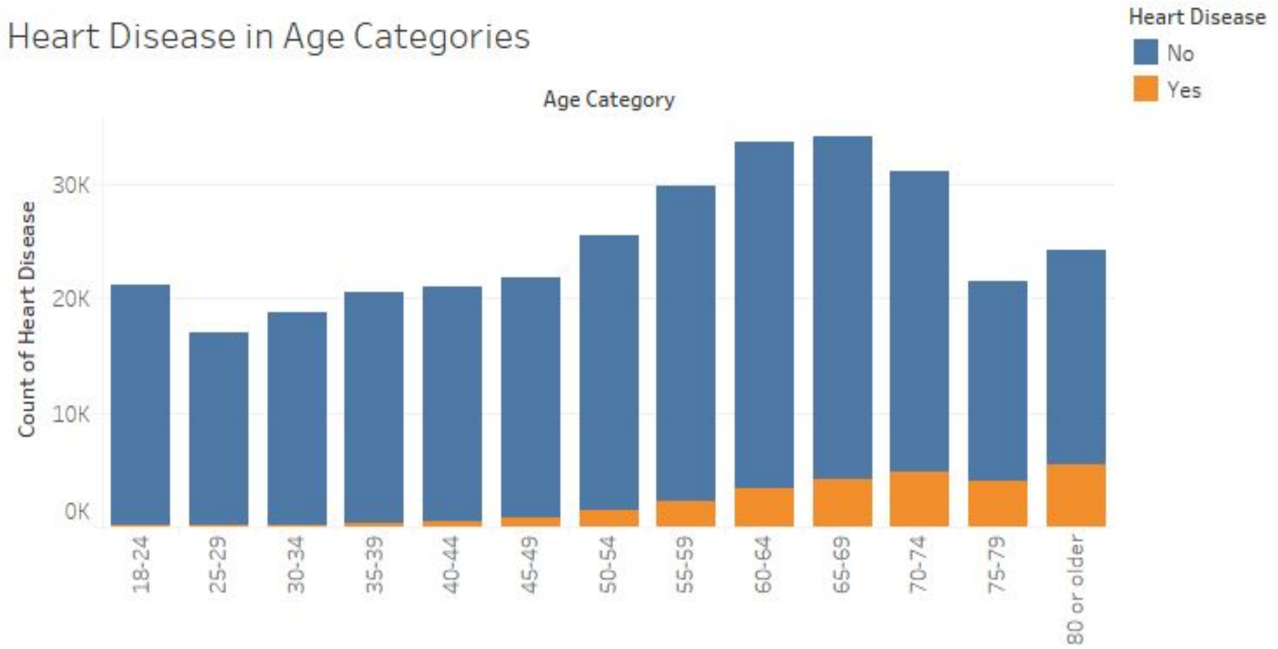


Analysis 4


Age Ranges w/ Heart Disease
Age Category
Count of Age Category
4K
2K
0K
18-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
80 or older
Heart Disease
(All)
No
Yes
Age Category
18-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64

Analysis 5

	Heart Disease	
Age Category	No	Yes
18-24	99.38%	0.62%
25-29	99.22%	0.78%
30-34	98.79%	1.21%
35-39	98.56%	1.44%
40-44	97.69%	2.31%
45-49	96.59%	3.41%
50-54	94.55%	5.45%
55-59	92.60%	7.40%
60-64	90.12%	9.88%
65-69	87.99%	12.01%
70-74	84.40%	15.60%
75-79	81.15%	18.85%
80 or older	77.44%	22.56%

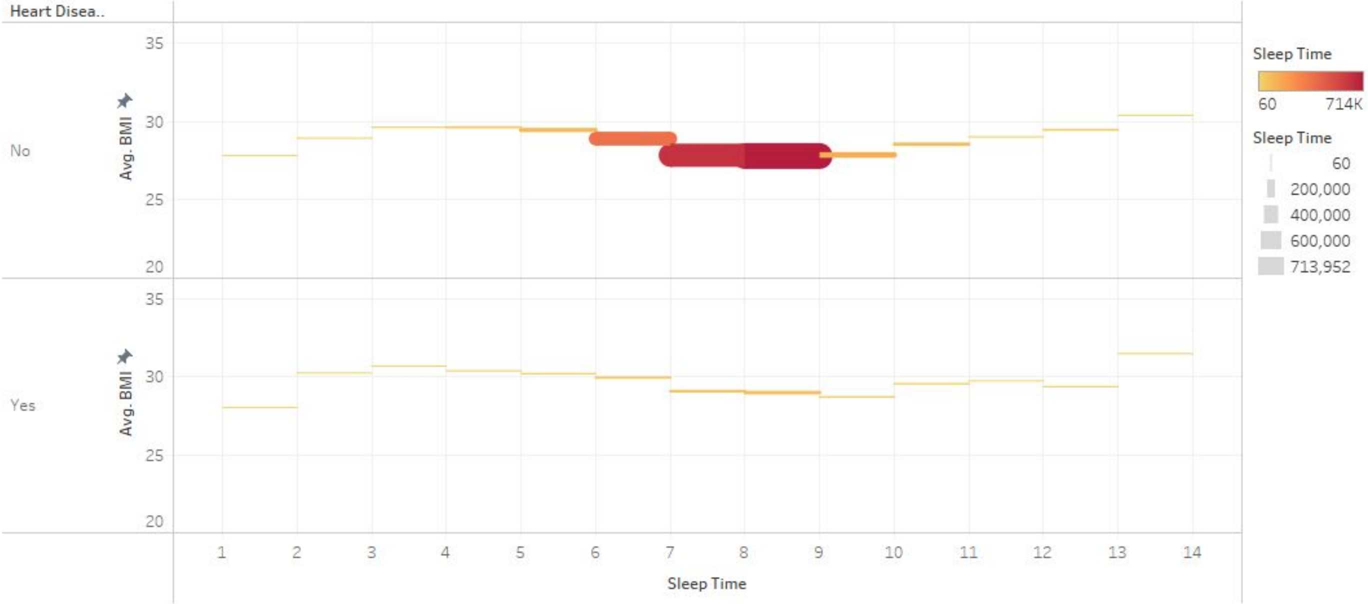


Analysis 6

Average Sleep Time

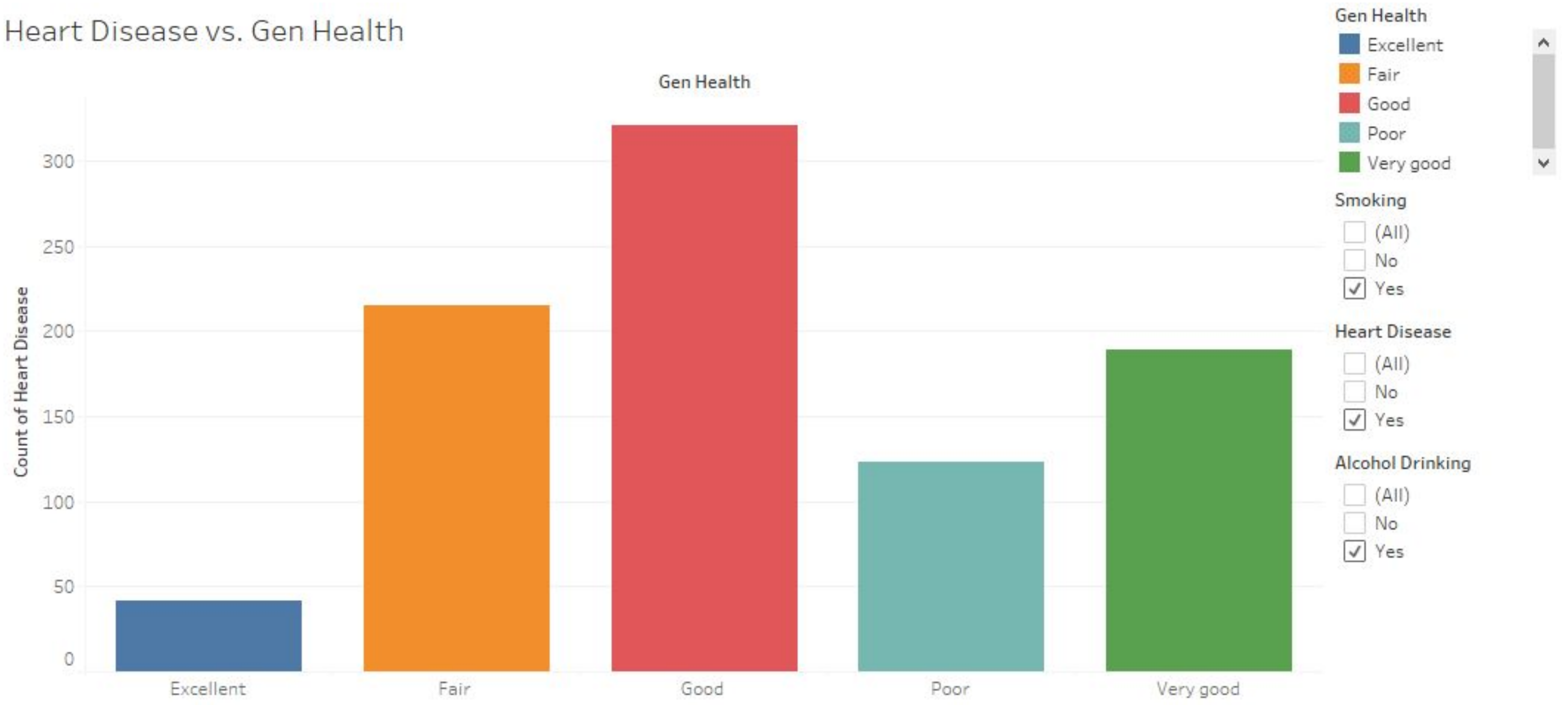
Heart Disease	
Yes	7.13616
No	7.09342

Sleep Time Comparison between Heart Disease/No Heart Disease



Analysis 7

Heart Disea..	Alcohol Drinking	Smoking	
Yes	No	No	40.50%
		Yes	55.34%
	Yes	No	0.92%
		Yes	3.25%



App Demonstration

The End!

Q & A Time

References

Beckerman, J. (2021, July 29). *How Heart Disease Affects Your Body*. Retrieved from WebMD: https://www.webmd.com/heart-disease/ss/slideshow-heart-disease-affects-body

News-Medical.Net. (2022). *Heart Disease*. Retrieved from News Medical Life Sciences: https://www.news-medical.net/condition/Heart-Disease

https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

https://www.cdc.gov/brfss/annual_data/annual_2020.html