

Code Blue: Heart Attack: An Analysis on Heart Attack Data

Author: DeAndre Thomas

Professor: Dr. Debbie Landowski

Course: IST652

Date: June 12, 2022

**Introduction**

Heart disease, per the CDC, is the number one cause of death, followed by cancer, and COVID-19 (per recent data). Being that heart attack falls under the category of heart disease, it can be quite fatal. The intention of this work is to find the attributes of heart attack and where in the US are being most affected by this disease.

**Data**

Data for this project were found on Kaggle.com and CDC.org. The first data set, titled ‘heart.csv’ is an Excel Spreadsheets consisting of 304 rows and 15 columns (4,560 data points). This dataset contains data regarding the physical attributes/medical symptoms of a heart attack. Per Kaggle, the dataset contains 4 databases concerning heart diagnosis in the year 1988. Those databases are:

1. Cleveland Clinic Foundation (cleveland.data)
2. Hungarian Institute of Cardiology, Budapest (hungarian.data)
3. V.A. Medical Center, Long Beach, CA (long-beach-va.data)
4. University Hospital, Zurich, Switzerland (switzerland.data)

The additional data set from the CDC, titled ‘Heart\_Disease\_Mortality\_Data\_Among\_US\_Adults\_\_35\_\_\_by\_State\_Territory\_and\_County.csv’ is a compilatory one that includes demographic information of patients with heart disease from 2012-2014. This data is strictly United States data and consists of 59,076 rows and 19 columns.

**Questions**

1. What are the top five states effected by heart disease?
2. What gender is most prone to having heart attacks?
3. What are the top 3 factors in those that have had heart attacks?

**Data Cleaning**

The ‘heart.csv’ from Kaggle was fairly clean; there were no null values and no columns needed to be dropped. However, there column titles were changed to more understandable terms.

* ‘age': 'Age'
* sex': 'Gender'
* 'cp': 'Chest\_Pain'
* 'trtbps': 'Resting\_BP'
* 'chol': 'Cholesterol'
* 'fbs': 'Fasting\_BS'
* 'restecg': 'Resting\_ECG'
* 'thalachh': 'Max\_HR'
* 'exng': 'Exercise\_Ang'
* 'oldpeak': 'ST\_Depres'
* 'slp': 'Slope'
* 'caa': 'Num\_Vessels'
* 'thall': 'Hem\_Lev'
* 'output': 'HA\_Diag'

The CDC data set had several columns that needed to be renamed:

* 'LocationAbbr': 'State'
* 'LocationDesc': 'County'
* 'Stratification1': 'Gender'
* 'Stratification2': 'Race/Ethnicity'

Also, there were several columns that needed to be dropped due their irrelevancy to the analysis:

* 'GeographicLevel', 'DataSource', 'Class'
* 'Data\_Value\_Unit', 'Data\_Value\_Type'
* 'Data\_Value\_Footnote\_Symbol'
* 'Data\_Value\_Footnote'
* 'StratificationCategory1'
* 'StratificationCategory2', 'TopicID'
* 'LocationID', 'Location 1'

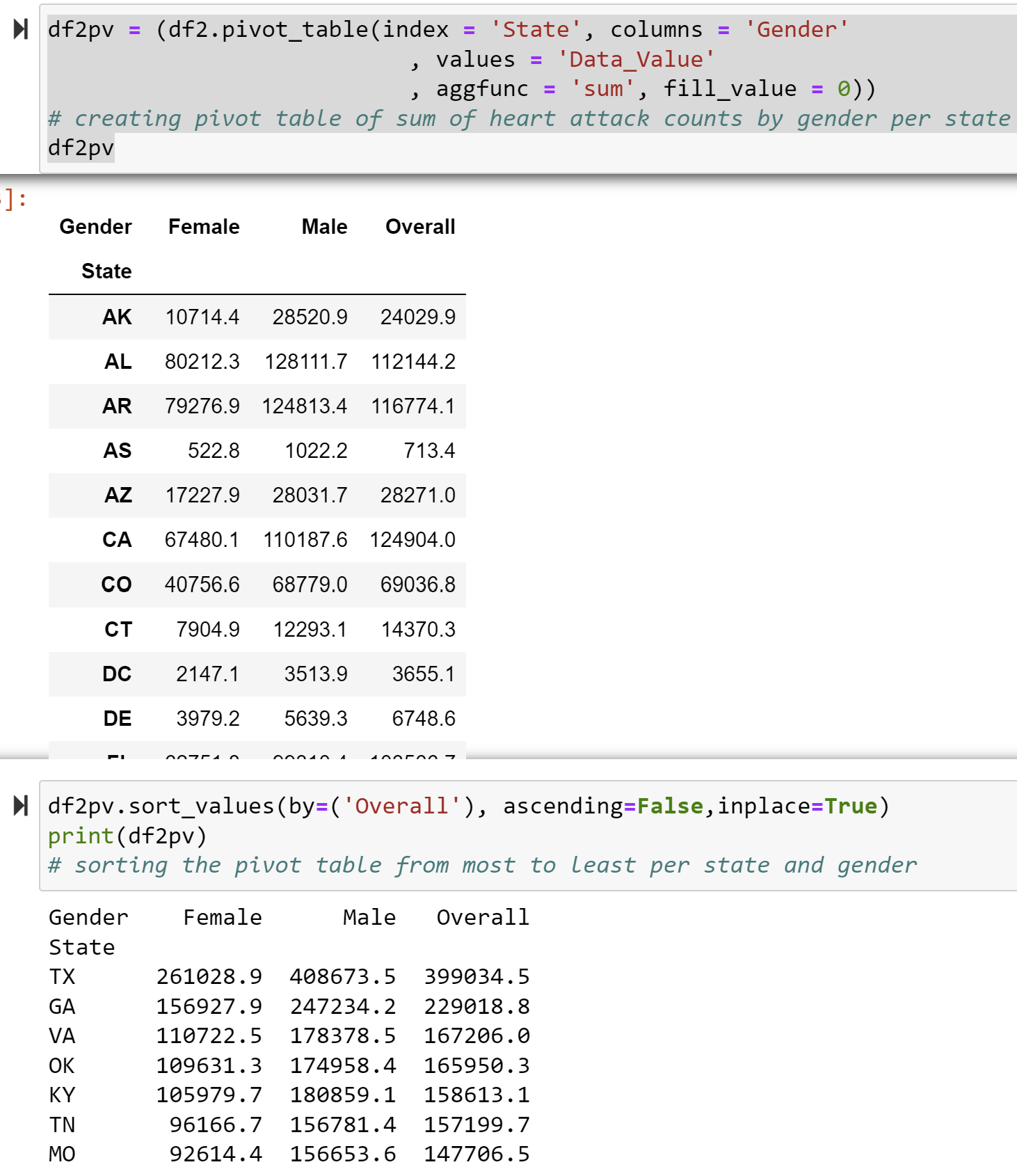
Last, there were rows in the CDC data set within the ‘Data\_Value’ column that were removed. It was decided to remove these values and not replace them due to not wanting to skew the data in analysis.

**Program Description**

Once proper cleaning was complete, a pivot table was created to answer the first two questions:

1. What areas (states) are most effected by heart disease?
2. What gender is most prone to having heart attacks?

After doing so, a list was created of sum amounts of people per gender and overall (both male and female accounted), and the list was printed by states in alphabetical order. Following, that list was sorted in ascending order.



Following, in effort to create a heat map to display the distribution of those diagnosed with having a heart attack in the United States, the pivot table previously created was converted into a pandas data frame. This data frame was then used to create a heat map of the United States that demonstrated states with the most diagnoses in dark purple and states with the least diagnoses in light yellow/light green. This program is also interactive where the user can hover over each state to see the overall sum amounts of diagnoses.

Graphical user interface

Description automatically generated with medium confidence

This heat map of the United States demonstrates that the most people affected by heart disease are in the state of Texas. This visual also demonstrates the program’s interactive feature of displaying the overall amount of people affected by hovering over the state.

Next, a stacked bar charts were created to demonstrate the amount sum amounts of diagnoses per gender in each state.

Chart

Description automatically generated

This plot demonstrates that Texas had the most people (male, female, and overall) affected by heart disease.

To answer Question 3 (What are the top 3 factors in those that have had heart attacks?), a pair plot was created to find visual correlations and collinearity between the variables of the Kaggle data set (‘heart.csv’.

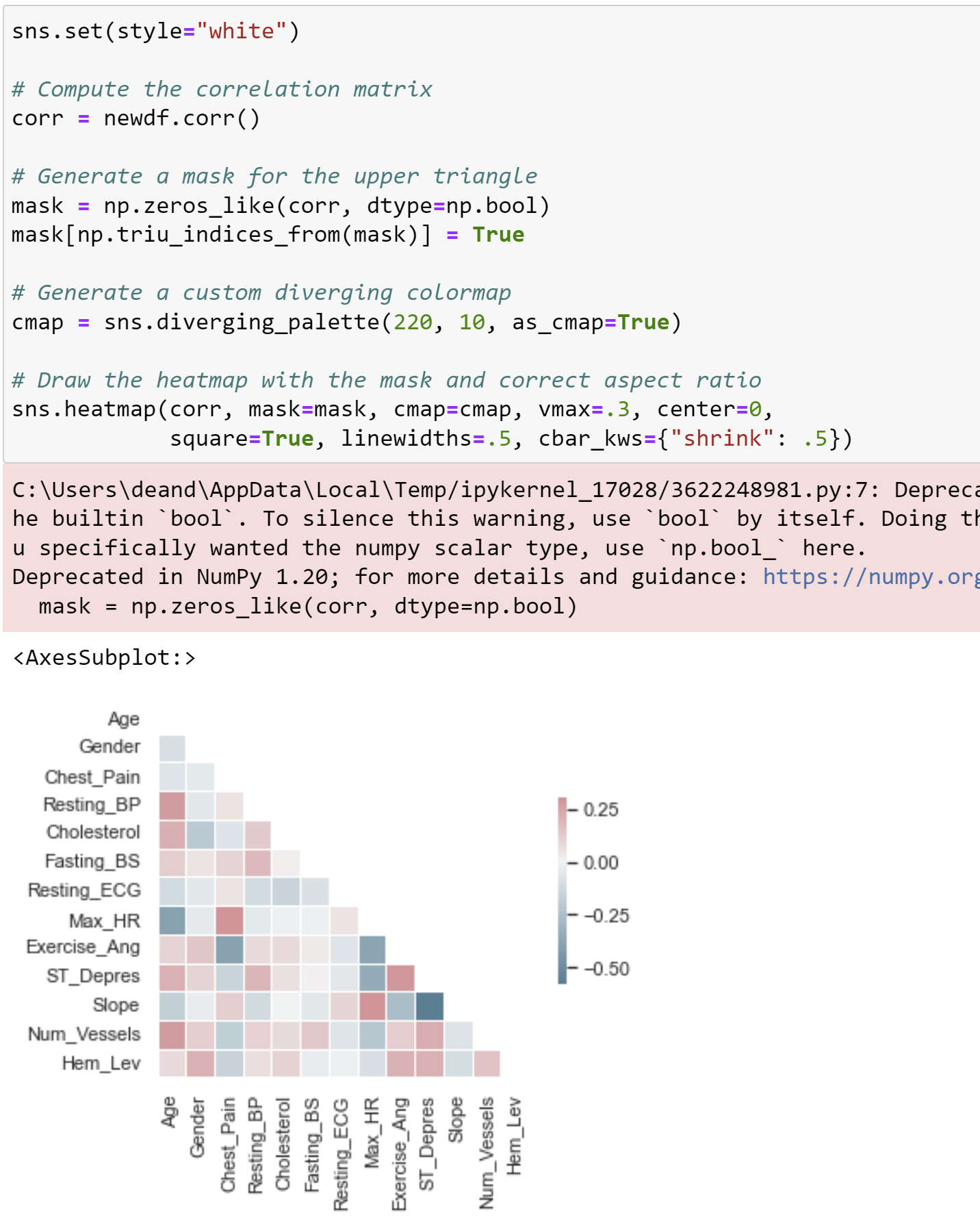
Diagram

Description automatically generated

The results of this visual were inconclusive. Being so, a correlation heat maps were created to observe the correlations per gradients in color and numerical values.

Chart, treemap chart

Description automatically generated



The heat maps demonstrates that the variables do not show high positive correlation. However, there are several strong negative correlations, including

* Exercise induced angina (chest pain): Heart Attack Diagnosis
* Age: Maximum Heart Rate
* ST Depression (reading on EKG/heart monitoring device): Heart Attack Diagnosis

Nonetheless, the variables in this model do not demonstrate high multicollinearity.

Finally, a linear regression model was created using the XGBoost Regressor. This model provided the top attributes in someone being diagnosed with a heart attacked.

Graphical user interface, text, application

Description automatically generated

**Results**

1. What are the top 5 states effected by heart disease?

* Texas, Georgia, Virginia, Oklahoma, Kentucky

1. What gender is most prone to having heart attacks?

* Male

1. What are the top 3 factors in those that have had heart attacks?

* Hemoglobin Levels, Chest Pain, and Thalassemia (a blood disorder that causes your body to have less hemoglobin (cells that carry oxygen) than normal).

**Conclusion & Recommendations**

Being that modern medicine is constantly evolving with medical treatment, dietary recommendations, nutrition, lifestyle recommendations, genetic evolution, working with a more up-to-date data set would be beneficial in analysis. Also, working with an up-to-date global dataset, such as the one obtained from Kaggle (‘heart.csv’), would allow effective sampling to derive at a statistics-based conclusion on population.

**References**

1. *Heart Disease Dataset*. (2019, June 6). Kaggle. https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset
2. *Heart Disease Mortality Data Among US Adults (35+) by State/Territory and County | Chronic Disease and Health Promotion Data & Indicators*. (2016, September 20). Https://Chronicdata.Cdc.Gov/Heart-Disease-Stroke-Prevention/Heart-Disease-Mortality-Data-Among-US-Adults-35-by/R35g-Znws. https://chronicdata.cdc.gov/Heart-Disease-Stroke-Prevention/Heart-Disease-Mortality-Data-Among-US-Adults-35-by/r35g-znws
3. *WONDER Message*. (2020). Https://Wonder.Cdc.Gov/Controller/Datarequest/D76;jsessionid=5135CC1057F85421F7D31B1322B4. https://wonder.cdc.gov/controller/datarequest/D76;jsessionid=5135CC1057F85421F7D31B1322B4
4. *Simplest Way to Create a Choropleth Map by U.S. States in Python*. (2022). Https://Towardsdatascience.Com/Simplest-Way-of-Creating-a-Choropleth-Map-by-u-s-States-in-Python-F359ada7735e. https://towardsdatascience.com/simplest-way-of-creating-a-choropleth-map-by-u-s-states-in-python-f359ada7735e