Project 2: An Investigation of Cardiovascular Risk Factors

Team Members: Tejasvi Kalakota, Mariam Elshenawy
Course: DS200, Section 8, Group 6
Date: 2023 Dec 12
Data Set: https://www.kaggle.com/datasets/captainozlem/framingham-chd-preprocessed-data/
GitHub Repository: https://github.com/UC-Berkeley-I-School/Project2_Elshenawy_Kalakota

Data Set:

The Framingham Heart Study was a cohort study started in 1948 by the National Heart, Lung, and Blood Institute, and it has been pivotal in transforming our understanding of cardiovascular health (Hong, 2023). It consists of decades of data, it's evolved into a multigenerational study examining patterns within cardiovascular diseases. By gathering genetic information and later expanding to include diverse populations, the FHS has unveiled a series of key risk features for heart disease. Findings from the FHS guide medical practices, and influence routine physicals to address factors like hypertension and cholesterol. Much of the data from the Framingham Study is publicly available but certain features need to be requested. For our project we use data from this study that has been extracted and joined into a single csv file that is available for public use.

Structure:

The dataset contains over 4,000 records with 16 attributes:

Demographics:
1) Sex: male or female (nominal).

2) Age: patient age (continuous).

Behavioral:
3) Education: 0=less than high school and high school degrees, 1=college degree and higher (nominal).
4) currentSmoker: 1=patient is a smoker or 0=non-smoker (nominal).
5) cigsPerDay: the number of cigarettes that the person smoked on average in one day (continuous).

Medical (history):
6) BPMeds: 1=on BP meds, 0=Not on BP meds (nominal).
7) prevalentStroke: 1=patient had previously had a stroke, 0=no history of stroke (nominal).
8) prevalentHyp: 1=patient was hypertensive, 0=not hypertensive (nominal).
9) diabetes: 1=patient had diabetes, 0=no diabetes (nominal).

Medical (current):
10) totChol: total cholesterol level (continuous)
11) sysBP: systolic blood pressure (continuous)
12) diaBP: diastolic blood pressure (continuous)
13) BMI: Body Mass Index (continuous)

14) Heart Rate: heart rate (continuous - in medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)
15) Glucose: glucose level (continuous)
Predict variable:
16) 10 year risk of coronary heart disease CHD (1=Yes, 0=No)

Introduction:

Heart Disease is one of the most debilitating and prevalent ailments plaguing the United States. The CDC reports that heart disease is the leading cause of death across all gender and most ethnic and racial groups in the United States (CDC, 2022). Behavioral Health factors like smoking and alcohol consumption are considered cardiovascular risk factors, but the exact causal relationship between cardiovascular disease between these factors is still not fully understood. A report published in 2010 outlines that smoking cigarettes stimulates the sympathetic nervous system and has an immediate hypertensive impact, which in turn may have perilous effects on blood pressure (Giannarelli, 2010). Similarly, a Meta-Analysis study conducted in 2019 found that dietary interventions resulting in weight reduction have contributed to improved glycemic control and lipid profiles (Liatis, 2019). Therefore, there is much value in investigating how core health behaviors associated with cardiovascular health like and smoking, are related to documenting cardiovascular health factors like cholesterol, blood pressure, and glucose control. Understanding the relationships between health behaviors and health factors like blood pressure is vital to developing novel therapeutics for treating cardiovascular disease as well as formulating community health interventions to reduce the risk of heart disease.

Primary Data Science Question
How do cardiovascular risk features (e.g., blood pressure, cholesterol) of patients who smoke vary from those who do not?

Sanity Check & Data Cleaning

Before starting our analysis, we wanted to perform a quick sanity check on the dataset. The Framingham data is known for its integrity, however, we wanted to complete our due diligence and check before proceeding. We will especially focus on attributes that are critical for the quality of our analysis.

```
Index(['male', 'age', 'education', 'currentSmoker', 'cigsPerDay', 'BPMeds',
       'prevalentStroke', 'prevalentHyp', 'diabetes', 'totChol', 'sysBP',
       'diaBP', 'BMI', 'heartRate', 'glucose', 'TenYearCHD'],
      dtype='object')
```

1. Missing Value Analysis
First, we will check the dataset for any missing values within each column, as this would negatively affect the quality of our analysis.

```
male               0
age                0
education          0
currentSmoker      0
cigsPerDay         0
BPMeds             0
prevalentStroke    0
prevalentHyp       0
```

```
diabetes            0
totChol             0
sysBP               0
diaBP               0
BMI                 0
heartRate           0
glucose             0
TenYearCHD          0
dtype: int64
```

It appears there are no missing values.

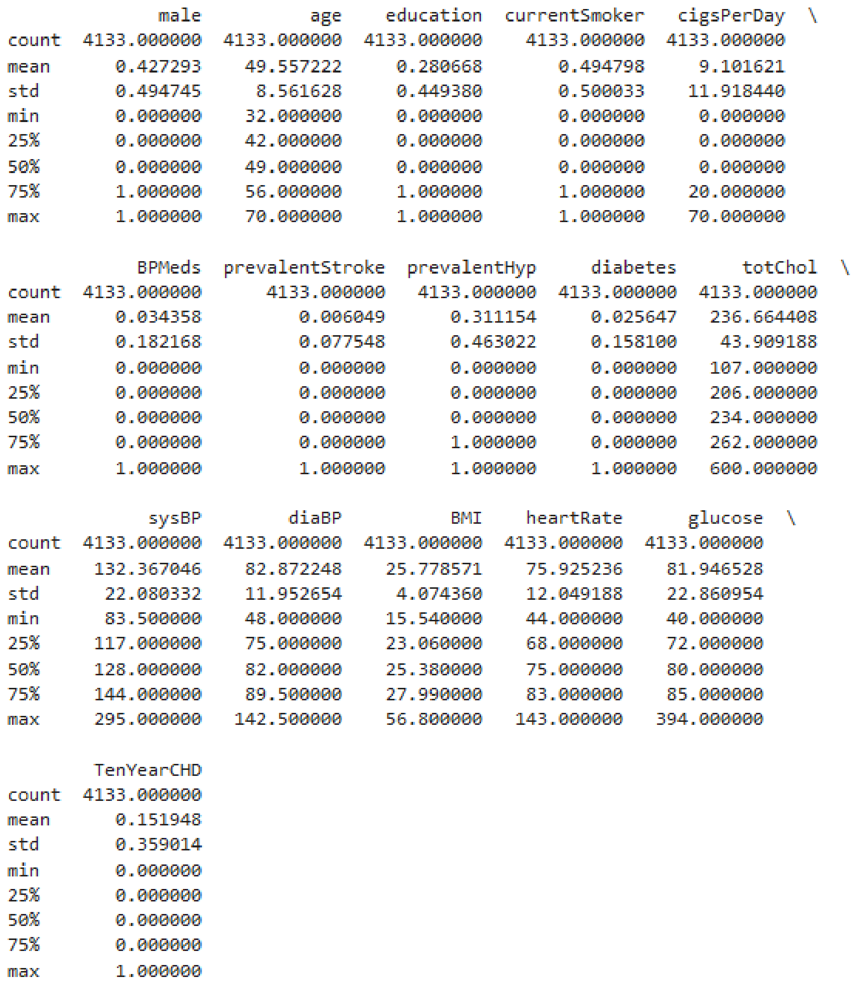
2. Data Type Validation

Next, we want to check the dtypes of each of our columns to make sure they are the correct type per attribute; e.g. integers, floats.

```
male                int64
age                 int64
education           int64
currentSmoker       int64
cigsPerDay          float64
BPMeds              float64
prevalentStroke     int64
prevalentHyp        int64
diabetes            int64
totChol             float64
sysBP               float64
diaBP               float64
BMI                 float64
heartRate           float64
glucose             float64
TenYearCHD          int64
dtype: object
```

3. Statistical Summary & Outliers

Next, we will generate a statistical summary for the dataset to visualize data distribution and identify outliers.



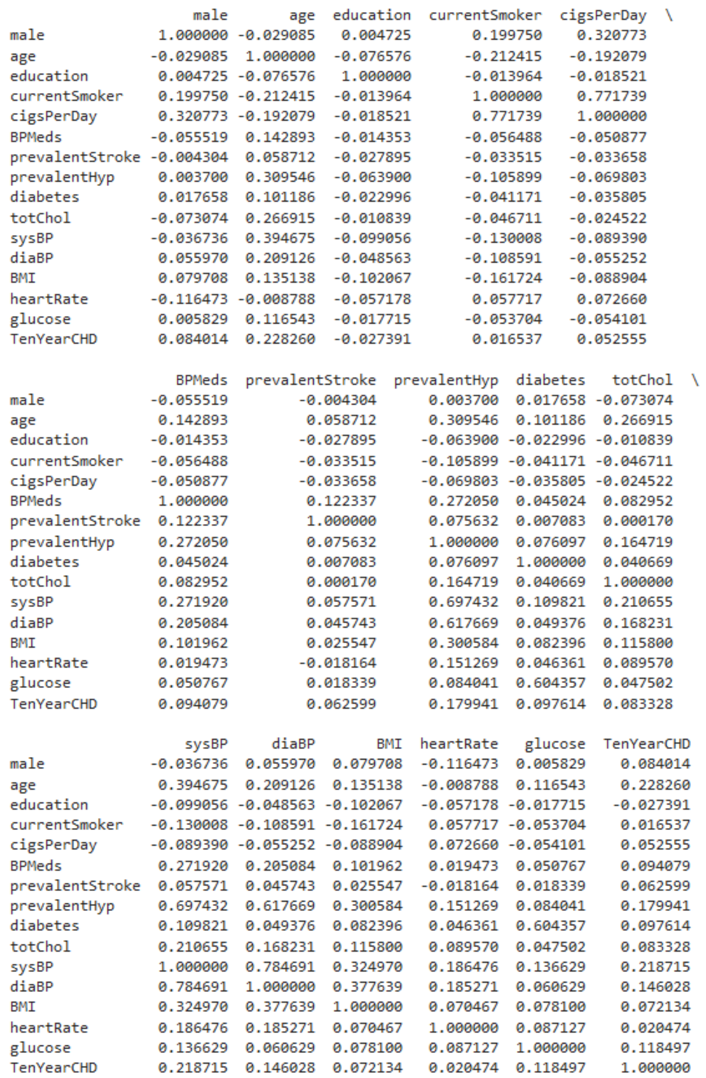
```
              male          age    education  currentSmoker   cigsPerDay  \
count  4133.000000  4133.000000  4133.000000    4133.000000  4133.000000
mean      0.427293    49.557222     0.280668       0.494798     9.101621
std       0.494745     8.561628     0.449380       0.500033    11.918440
min       0.000000    32.000000     0.000000       0.000000     0.000000
25%       0.000000    42.000000     0.000000       0.000000     0.000000
50%       0.000000    49.000000     0.000000       0.000000     0.000000
75%       1.000000    56.000000     1.000000       1.000000    20.000000
max       1.000000    70.000000     1.000000       1.000000    70.000000

            BPMeds  prevalentStroke  prevalentHyp     diabetes      totChol  \
count  4133.000000      4133.000000   4133.000000  4133.000000  4133.000000
mean      0.034358         0.006049      0.311154     0.025647   236.664408
std       0.182168         0.077548      0.463022     0.158100    43.909188
min       0.000000         0.000000      0.000000     0.000000   107.000000
25%       0.000000         0.000000      0.000000     0.000000   206.000000
50%       0.000000         0.000000      0.000000     0.000000   234.000000
75%       0.000000         0.000000      1.000000     0.000000   262.000000
max       1.000000         1.000000      1.000000     1.000000   600.000000

             sysBP        diaBP          BMI    heartRate      glucose  \
count  4133.000000  4133.000000  4133.000000  4133.000000  4133.000000
mean    132.367046    82.872248    25.778571    75.925236    81.946528
std      22.080332    11.952654     4.074360    12.049188    22.860954
min      83.500000    48.000000    15.540000    44.000000    40.000000
25%     117.000000    75.000000    23.060000    68.000000    72.000000
50%     128.000000    82.000000    25.380000    75.000000    80.000000
75%     144.000000    89.500000    27.990000    83.000000    85.000000
max     295.000000   142.500000    56.800000   143.000000   394.000000

        TenYearCHD
count  4133.000000
mean      0.151948
std       0.359014
min       0.000000
25%       0.000000
50%       0.000000
75%       0.000000
max       1.000000
```

Key Takeaways:

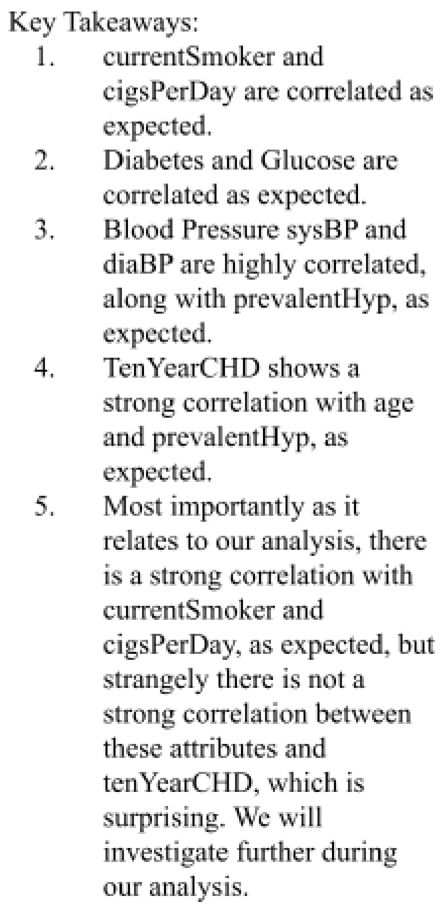
1. Age range is 32 to 70 years old - within the normal range.
2. CigsPerDay has a max of 70, seems high, but on investigation, that is about 3.5 packs per day, which is plausible.
3. BPMeds, PrevalentStroke, PrevalentHyp, Diabetes, TenYearCHD all are binary values and summary appears normal.
4. Total Cholesterol max value is 600 which is very high, may be an outlier, but on investigation, it is plausible.
5. sysBP and diaBP - max sysBP is 295 and max diaBP is 142.5. Both are extremely high and would indicate severe hypertension or possible outliers, but upon investigation, they are plausible.
6. BMI range is 15 to 56 - both are plausible in a population, 56 is severe morbid obesity, 15 is underweight.
7. Heart Rate - ranges from 44 to 143, all are normal range, 143 is likely tachycardia.
8. Glucose - max was 394, which is high, but plausible in uncontrolled diabetes.

4. Correlation Analysis

Next, we'll generate a correlation matrix to identify relationships between variables.



```
                    male       age  education  currentSmoker  cigsPerDay  \
male            1.000000 -0.029085   0.004725       0.199750    0.320773   
age            -0.029085  1.000000  -0.076576      -0.212415   -0.192079   
education       0.004725 -0.076576   1.000000      -0.013964   -0.018521   
currentSmoker   0.199750 -0.212415  -0.013964       1.000000    0.771739   
cigsPerDay      0.320773 -0.192079  -0.018521       0.771739    1.000000   
BPMeds         -0.055519  0.142893  -0.014353      -0.056488   -0.050877   
prevalentStroke -0.004304  0.058712  -0.027895      -0.033515   -0.033658   
prevalentHyp    0.003700  0.309546  -0.063900      -0.105899   -0.069803   
diabetes        0.017658  0.101186  -0.022996      -0.041171   -0.035805   
totChol        -0.073074  0.266915  -0.010839      -0.046711   -0.024522   
sysBP          -0.036736  0.394675  -0.099056      -0.130008   -0.089390   
diaBP           0.055970  0.209126  -0.048563      -0.108591   -0.055252   
BMI             0.079708  0.135138  -0.102067      -0.161724   -0.088904   
heartRate      -0.116473 -0.008788  -0.057178       0.057717    0.072660   
glucose         0.005829  0.116543  -0.017715      -0.053704   -0.054101   
TenYearCHD      0.084014  0.228260  -0.027391       0.016537    0.052555   

                  BPMeds  prevalentStroke  prevalentHyp  diabetes   totChol  \
male           -0.055519        -0.004304      0.003700  0.017658 -0.073074   
age             0.142893         0.058712      0.309546  0.101186  0.266915   
education      -0.014353        -0.027895     -0.063900 -0.022996 -0.010839   
currentSmoker  -0.056488        -0.033515     -0.105899 -0.041171 -0.046711   
cigsPerDay     -0.050877        -0.033658     -0.069803 -0.035805 -0.024522   
BPMeds          1.000000         0.122337      0.272050  0.045024  0.082952   
prevalentStroke  0.122337         1.000000      0.075632  0.007083  0.000170   
prevalentHyp    0.272050         0.075632      1.000000  0.076097  0.164719   
diabetes        0.045024         0.007083      0.076097  1.000000  0.040669   
totChol         0.082952         0.000170      0.164719  0.040669  1.000000   
sysBP           0.271920         0.057571      0.697432  0.109821  0.210655   
diaBP           0.205084         0.045743      0.617669  0.049376  0.168231   
BMI             0.101962         0.025547      0.300584  0.082396  0.115800   
heartRate       0.019473        -0.018164      0.151269  0.046361  0.089570   
glucose         0.050767         0.018339      0.084041  0.604357  0.047502   
TenYearCHD      0.094079         0.062599      0.179941  0.097614  0.083328   

                   sysBP     diaBP       BMI  heartRate   glucose  TenYearCHD  
male           -0.036736  0.055970  0.079708  -0.116473  0.005829    0.084014  
age             0.394675  0.209126  0.135138  -0.008788  0.116543    0.228260  
education      -0.099056 -0.048563 -0.102067  -0.057178 -0.017715   -0.027391  
currentSmoker  -0.130008 -0.108591 -0.161724   0.057717 -0.053704    0.016537  
cigsPerDay     -0.089390 -0.055252 -0.088904   0.072660 -0.054101    0.052555  
BPMeds          0.271920  0.205084  0.101962   0.019473  0.050767    0.094079  
prevalentStroke  0.057571  0.045743  0.025547  -0.018164  0.018339    0.062599  
prevalentHyp    0.697432  0.617669  0.300584   0.151269  0.084041    0.179941  
diabetes        0.109821  0.049376  0.082396   0.046361  0.604357    0.097614  
totChol         0.210655  0.168231  0.115800   0.089570  0.047502    0.083328  
sysBP           1.000000  0.784691  0.324970   0.186476  0.136629    0.218715  
diaBP           0.784691  1.000000  0.377639   0.185271  0.060629    0.146028  
BMI             0.324970  0.377639  1.000000   0.070467  0.078100    0.072134  
heartRate       0.186476  0.185271  0.070467   1.000000  0.087127    0.020474  
glucose         0.136629  0.060629  0.078100   0.087127  1.000000    0.118497  
TenYearCHD      0.218715  0.146028  0.072134   0.020474  0.118497    1.000000  
```



Key Takeaways:

1. currentSmoker and cigsPerDay are correlated as expected.
2. Diabetes and Glucose are correlated as expected.
3. Blood Pressure sysBP and diaBP are highly correlated, along with prevalentHyp, as expected.
4. TenYearCHD shows a strong correlation with age and prevalentHyp, as expected.
5. Most importantly as it relates to our analysis, there is a strong correlation with currentSmoker and cigsPerDay, as expected, but strangely there is not a strong correlation between these attributes and tenYearCHD, which is surprising. We will investigate further during our analysis.

5. Data Integrity Check

Here, we will do a quick check to see, especially for our binary attributes for our analysis, if there are any issues with unique values outside of the expected 0, 1.

Unique values in education: [1 0]
Unique values in currentSmoker: [0 1]
Unique values in BPMeds: [0. 1.]
Unique values in prevalentStroke: [0 1]
Unique values in prevalentHyp: [0 1]
Unique values in diabetes: [0 1]
Unique values in TenYearCHD: [0 1]

The above all check out across all of these columns. The only oddity is the float for BPMeds, but we don't anticipate it will cause any issues with our analysis.

6. Duplicate Record Check

Finally, we will assess if there are any duplicate records present.

Summary of Sanity Check & Data Cleaning Needs

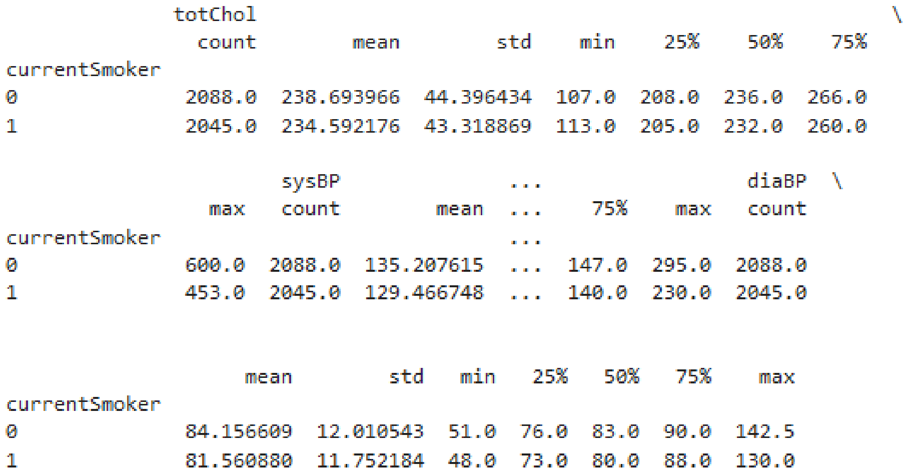
Cleaning Actions Required: None.

1. No missing values.
2. Data types are valid.
3. Ranges and Stats are valid and plausible. There were some high values, but none that were outside of the realm of possibility.
4. Correlations are valid.
5. No issues with unique values in key binary columns.
6. No duplicate records.

Now that we have completed our sanity check, and there are no glaring data cleaning needs, we will move onto our analysis, first starting with an initial exploration.

Initial Exploration

Let us first begin with a general exploration of basic statistics for coronary heart disease risk factors against currentSmoker and see what we get.



```
              totChol                                                     \
                count        mean        std    min    25%    50%    75%
currentSmoker
0              2088.0  238.693966  44.396434  107.0  208.0  236.0  266.0
1              2045.0  234.592176  43.318869  113.0  205.0  232.0  260.0

                       sysBP              ...                diaBP  \
                 max   count        mean  ...    75%    max  count
currentSmoker                             ...
0              600.0  2088.0  135.207615  ...  147.0  295.0  2088.0
1              453.0  2045.0  129.466748  ...  140.0  230.0  2045.0


                    mean        std   min   25%   50%   75%    max
currentSmoker
0              84.156609  12.010543  51.0  76.0  83.0  90.0  142.5
1              81.560880  11.752184  48.0  73.0  80.0  88.0  130.0
```

On first look, it seems counterintuitive to what we thought we would see, which is that non-smokers seem to have a marginally higher cholesterol, sysBP, and diaBP than smokers. The difference appears negligible at this point though, and we need more context.

We then stratify each of the respective features by gender, smoking status, and risk of Coronary Heart Disease

TenYear CHDRisk	Sex	Smoking Status	age	sysBP	diaBP	heartRate	BMI	cigsPerDay
No Risk	**Female**	**NonSmoker**	50.71	133.86	83.15	76.62	26.10	0.00
No Risk	**Female**	**Smoker**	46.70	127.19	79.50	77.64	24.28	13.95
No Risk	**Male**	**NonSmoker**	50.03	131.25	84.14	72.03	26.88	0.00
No Risk	**Male**	**Smoker**	47.24	127.91	82.12	75.21	25.63	21.89
CHDRisk	**Female**	**NonSmoker**	57.51	152.10	88.74	76.32	27.24	0.00
CHDRisk	**Female**	**Smoker**	50.83	137.37	83.83	79.59	25.73	15.76
CHDRisk	**Male**	**NonSmoker**	55.42	141.42	87.59	73.30	27.15	0.00
CHDRisk	**Male**	**Smoker**	52.20	140.55	86.69	76.94	26.00	22.97

Key Points from table above:

- Cigarettes Per Day
    - Females smoked on average fewer Cigarettes per Day compared to their male counterparts among those with and without a risk of CHD(coronary heart disease)
    - Both sexes within CHD Risk group on average smoked more Cigarettes per Day compared with the No Risk group counterparts
- Heart Rate
    - Females have elevated heart rate compared to male counterparts in all subgroups
    - With both CHD risk groups: Males and Females who smoke have a slightly elevated heart rate compared to males and females who do not
- Cholesterol
    - In all subgroups Non-smokers appear to have a slightly higher Cholesterol values compared to smokers, and this is particularly apparent with Females who have a risk of CHD

For the remainder of this report we will focus on three categories of cardiovascular features: Hypertension, Cholesterol, and Resting Heart Rate. We will generate box plots, histograms, and scatterplots to better visualize and contextualize these features against smoking status

<u>Visuals</u>

Blood Pressure

Figure 1

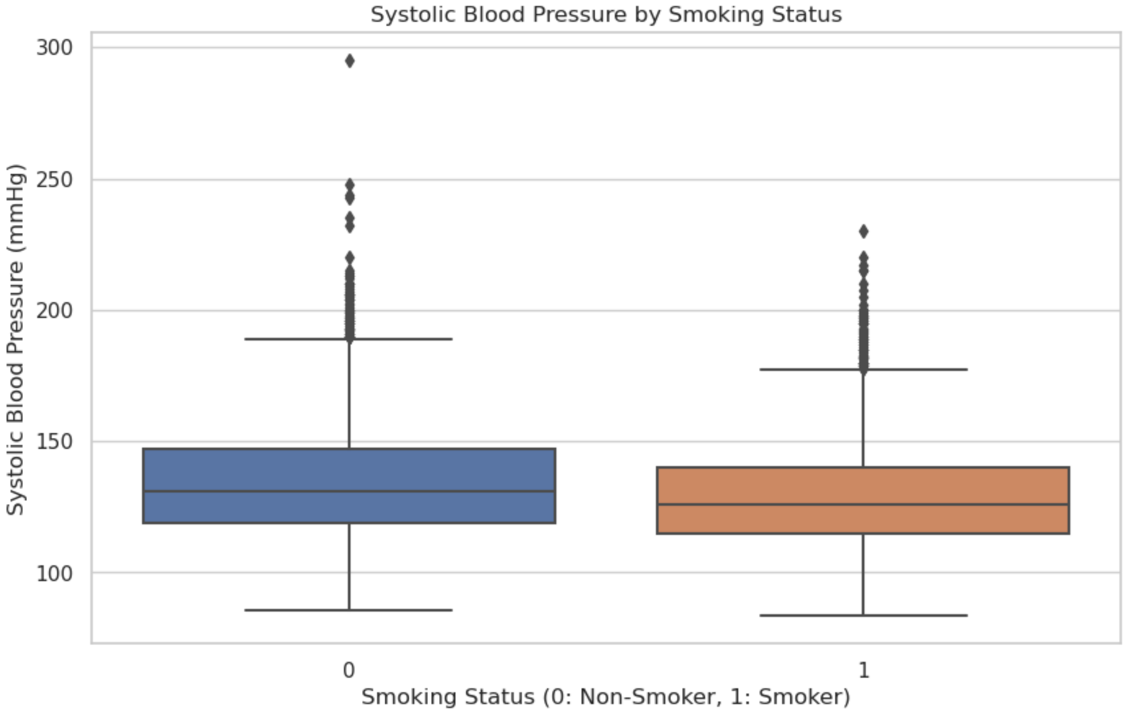


Figure 2

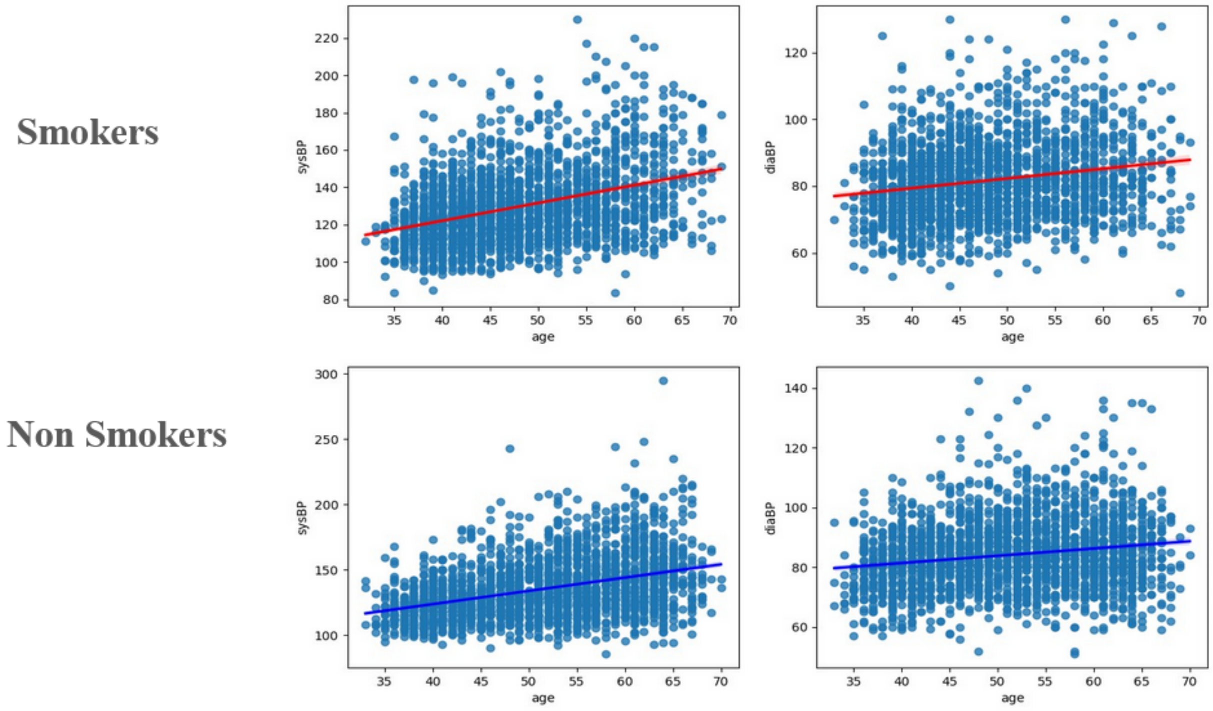


Figure 1 and 2 Description: Both Smokers and Non-Smokers have a slight upward trend of BP as age increases. Smokers have a slightly steeper systolic BP slope as age progresses, but overall with this dataset, it appears the Non-Smokers have a slightly higher BP than the Smokers. Most notably, the scatterplot spread of the Non-Smokers seems greater than with the Smokers, exhibiting greater variability among this group.

Figure 3

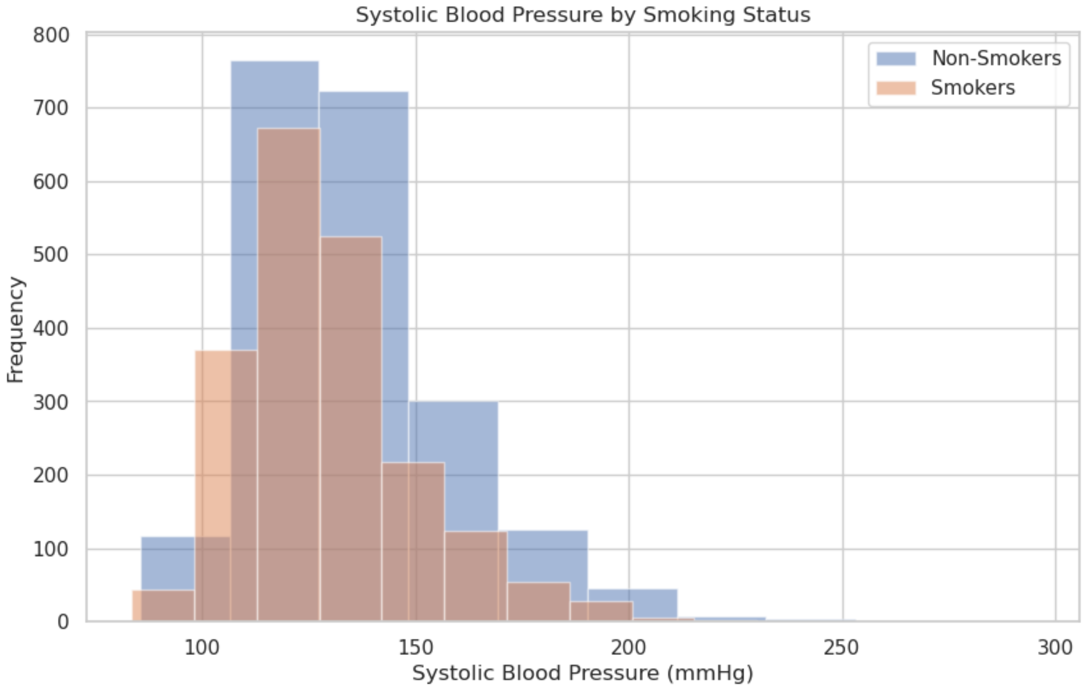


Figure 3 Description: This Systolic BP histogram shows how the Non-Smoker group seems to, again, have a slightly wider distribution and variability than the Smokers. Let us move onto Cholesterol and see if we discover the same trends.

Cholesterol

Figure 4

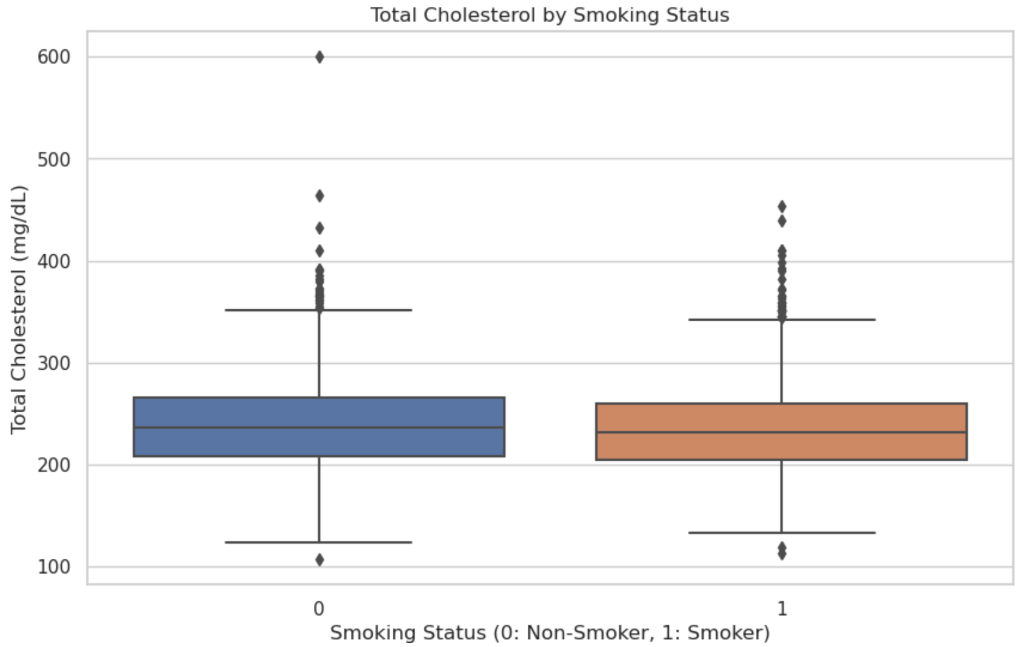


Figure 4 Description: Total Cholesterol by Smoking Status also exhibited the same marked variability among the non-smoker group, where there seems to be the high 600 anomaly we noticed in our sanity check. This hardly tells the full story, but it is a helpful note to make that the non-smoker group seems to have a wider interquartile range on the box plots than the smoker group, and that trend carries with the histograms.

Heart Rate

Figure 5A: What is the mean heart rate of men and women who do and do not smoke?

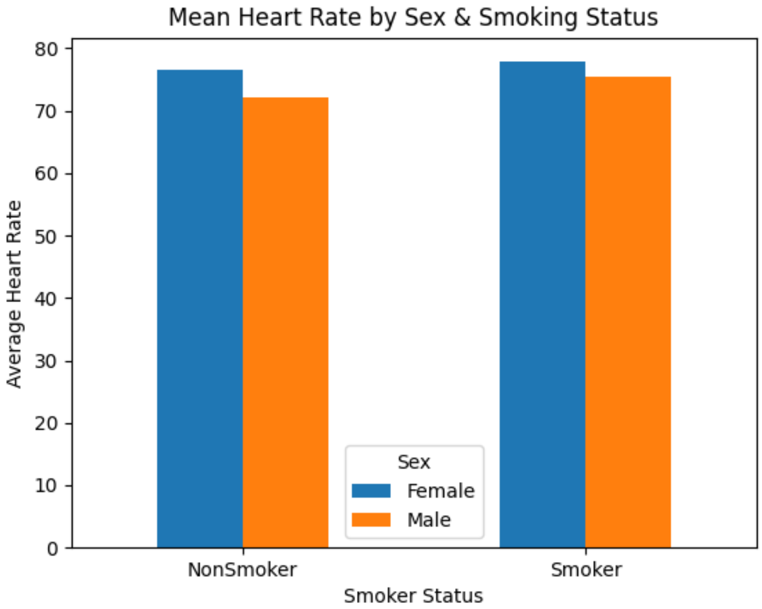


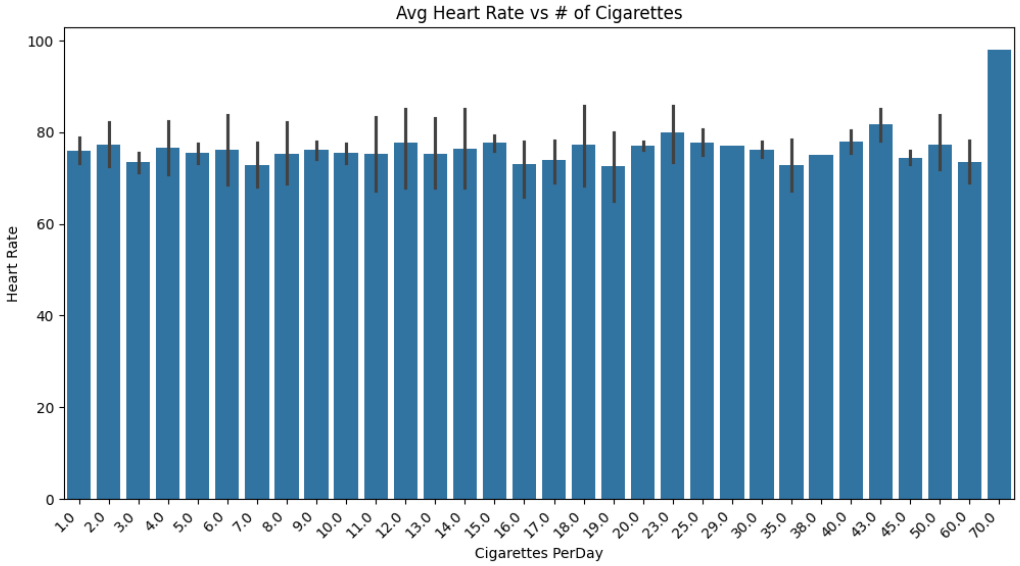
Figure 5B: What is the average heart rate per smoking status, per gender, and 10-year risk of future (CHD) coronary heart disease?

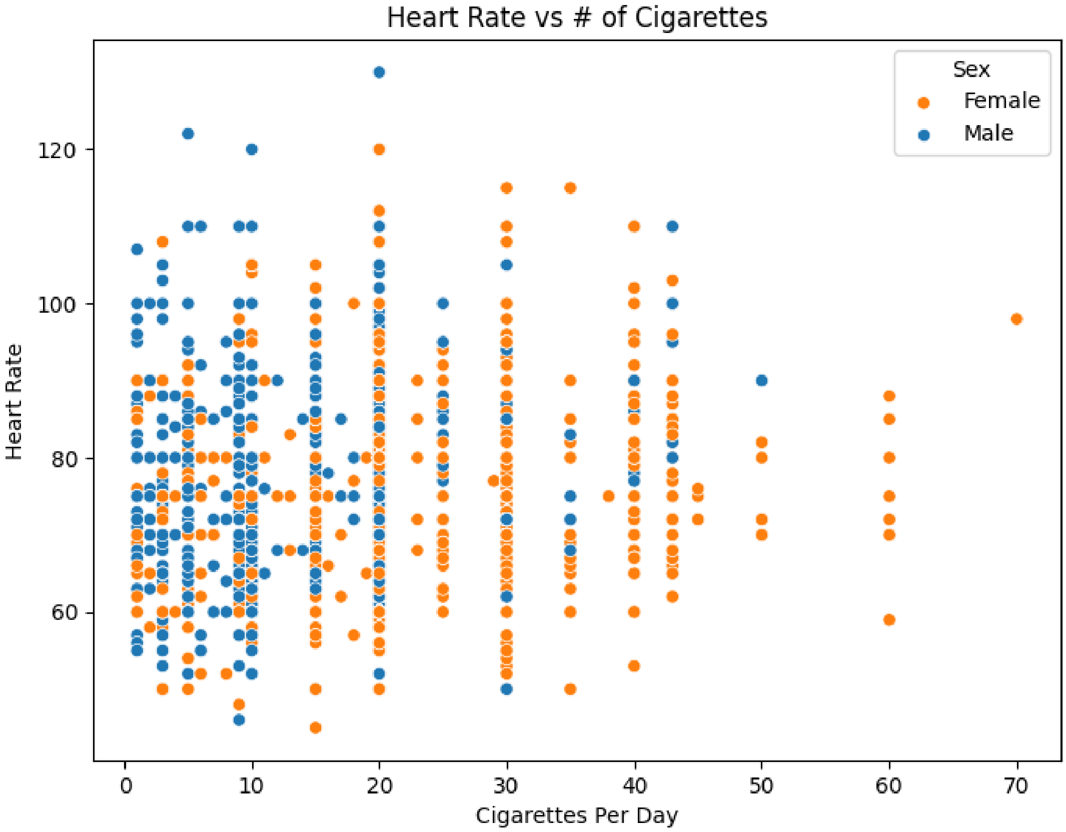
Smoking Status	Sex	No_CHD	CHD_Risk	All
NonSmoker	**Female**	76.61581921	76.32291667	76.57651992
	Male	72.03193277	73.30252101	72.24369748
Smoker	**Female**	77.6375	79.58715596	77.85237614
	Male	75.20975057	76.93721973	75.55837104
All		75.7625139	76.53032659	75.87898089

Figure 5A and 5B Description:

- Females have an elevated heart rate compared to their male counterparts in all subgroups
- Within both CHD risk groups: Males and Females who smoke have a slightly elevated heart rate compared to males and females who do not
- The results are consistent with existing literature. **Linneberg** et al suggest that part of the cardiovascular risk of smoking may operate through increasing resting heart rate

Figure 6A and 6B: How does heart rate differ based on the number of cigarettes a patient consumes?





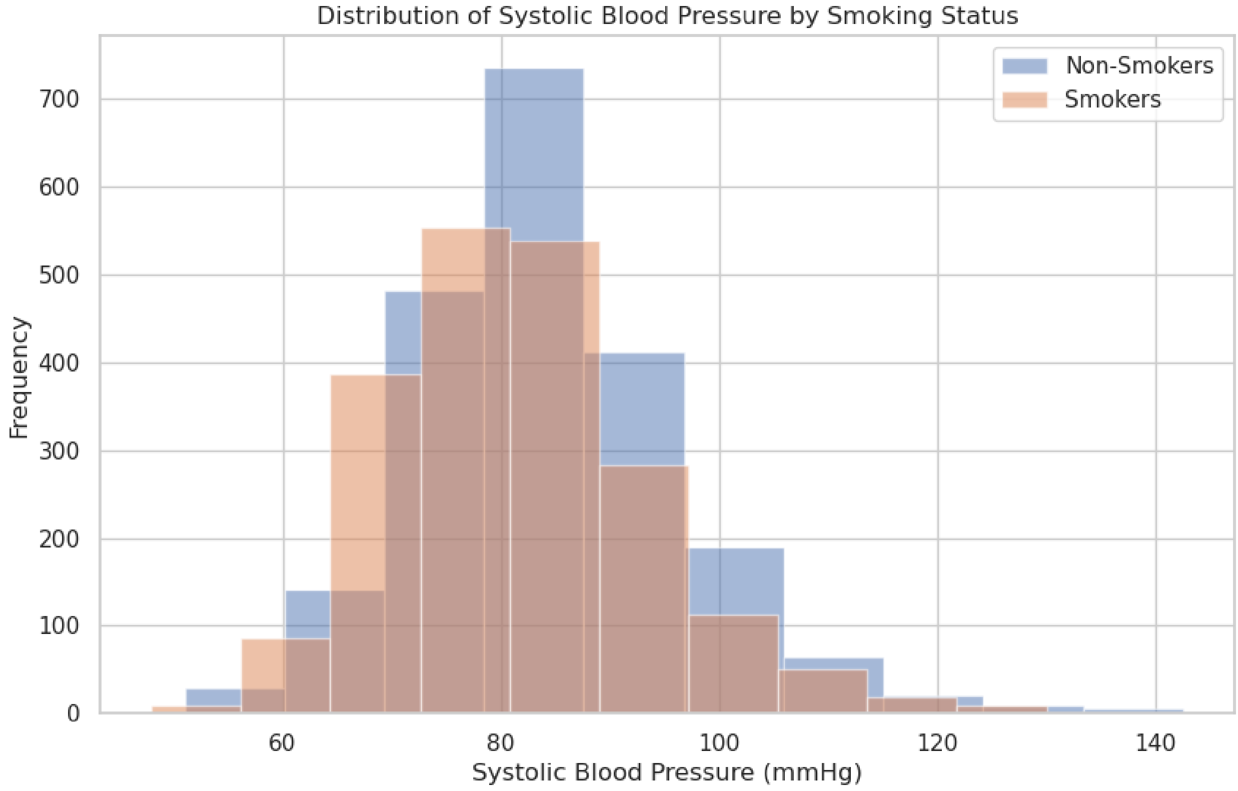
6A and 6B Description: At first glance it may appear that patients who smoke 70 cigarettes have a noticeably higher heart. However, the second scatter plot showcases how there is only one patient who smokes 70 cigarettes a day and this data point alone may not be a good measure of the average heart rate of an individual who smokes 70 cigarettes. Regardless, both visuals combined showcase that there is not a clear pattern or correlation between the number of cigarettes one smokes a day and their resting heart rate.

Conclusion:

In summary, we identified some interesting patterns with this dataset as we explored smoking behaviors and cardiovascular health. Our results of our exploratory analysis are consistent with former meta-analysis studies which indicate that smoking is causally related to higher levels of resting heart rate, but not to alterations in blood pressure and risk of hypertension (Linneberg 2015). A more surprising component of our analysis is that there was no apparent correlation between the number of cigarettes one smokes and the propensity for a higher or even average resting heart rate. Ultimately, the data here suggests that the relationship between smoking and cardiovascular health is likely more complex than initially expected, and we would need additional information about lifestyle choices and other healthcare habits to better forumate true correlations. For next steps, we would want to refine our risk models and expand our understanding through statistical analysis.

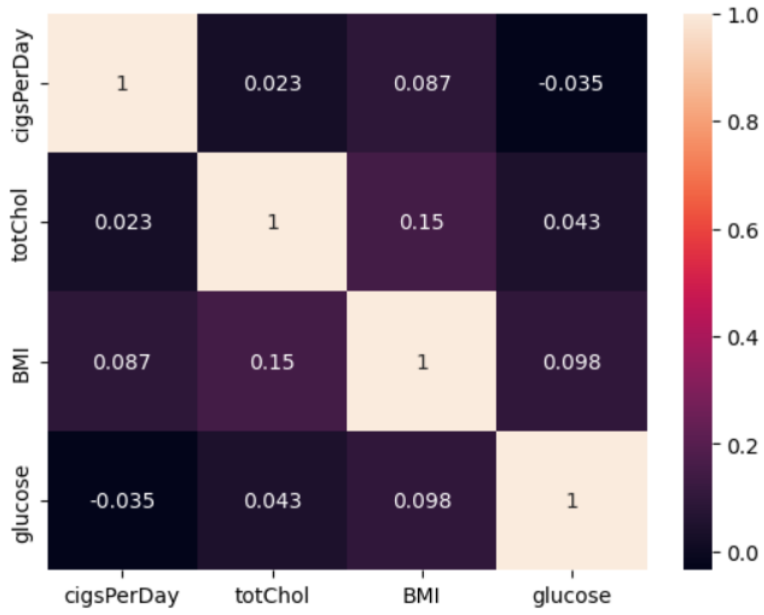
Appendix

A. Figure 1



Appendix Figure 1 Description: This histogram was created to assess the distribution of Diastolic BP according to smoking status. The same trends identified during the Initial Exploration was discovered, where the Non-Smoker group seemed to have a wider variability than the Smoker group.

A.Figure 2



Appendix Figure 2 Description: There is not a strong correlation between any of the cardiovascular features outlined in the heatmap above

---

Bibliography

Linneberg, Allan et al. "Effect of Smoking on Blood Pressure and Resting Heart Rate: A Mendelian Randomization Meta-Analysis in the CARTA Consortium." *Circulation. Cardiovascular genetics* 8.6 (2015): 832–841. Web.

Virdis, A., Giannarelli, C., Neves, M. F., Taddei, S., & Ghiadoni, L. (2010). Cigarette smoking and hypertension. Current pharmaceutical design, 16(23), 2518–2525. https://doi.org/10.2174/138161210792062920

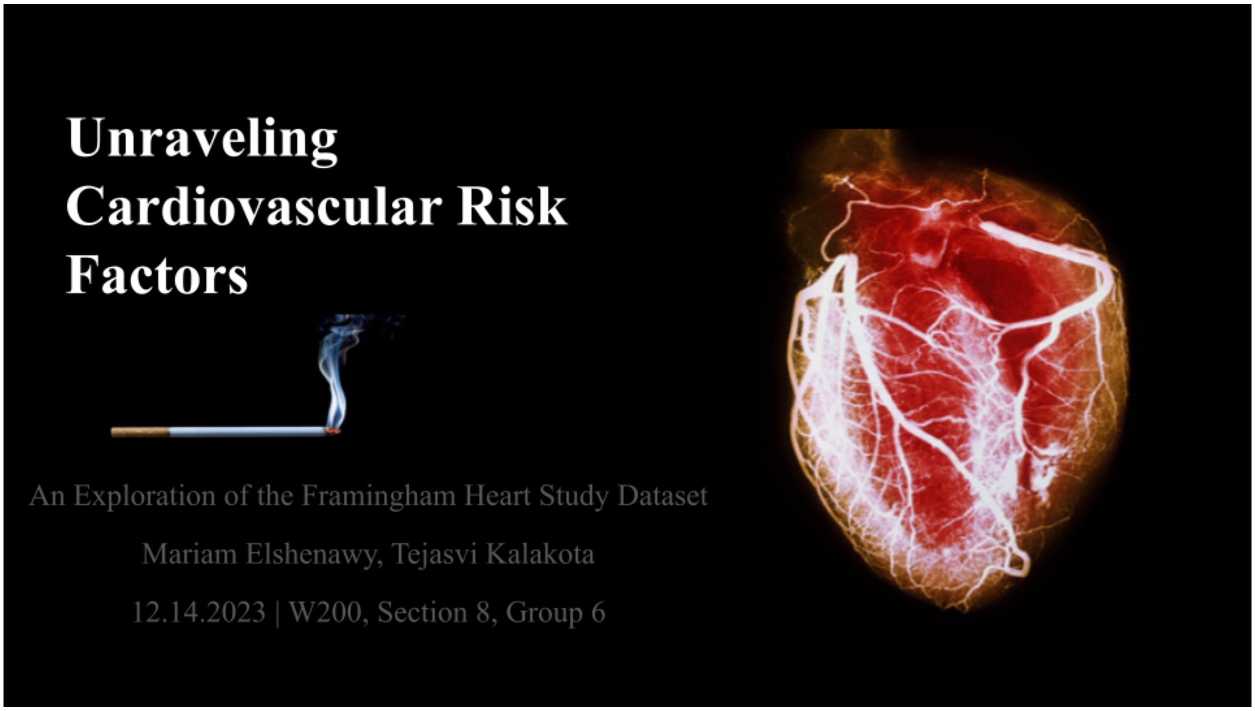
Zhou, C., Wang, M., Liang, J., He, G., & Chen, N. (2022). Ketogenic Diet Benefits to Weight Loss, Glycemic Control, and Lipid Profiles in Overweight Patients with Type 2 Diabetes Mellitus: A Meta-Analysis of Randomized Controlled Trails. International journal of environmental research and public health, 19(16), 10429. https://doi.org/10.3390/ijerph191610429

Moser M. (1999). Hypertension treatment and the prevention of coronary heart disease in the elderly. *American family physician*, *59*(5), 1248–1256.

Koliaki, C., Liatis, S., & Kokkinos, A. (2019). Obesity and cardiovascular disease: Revisiting an old relationship. Metabolism, 92, 98-107. doi: 10.1016/j.metabol.2018.10.008

Centers for Disease Control and Prevention. (2021, February 17). About heart disease. Retrieved from https://www.cdc.gov/heartdisease/about.htm

Presentation:



Unraveling Cardiovascular Risk Factors

An Exploration of the Framingham Heart Study Dataset

Mariam Elshenawy, Tejasvi Kalakota

12.14.2023 | W200, Section 8, Group 6

Introduction

Intended Audience: Healthcare professionals and Public Health Policy Officials

Background:

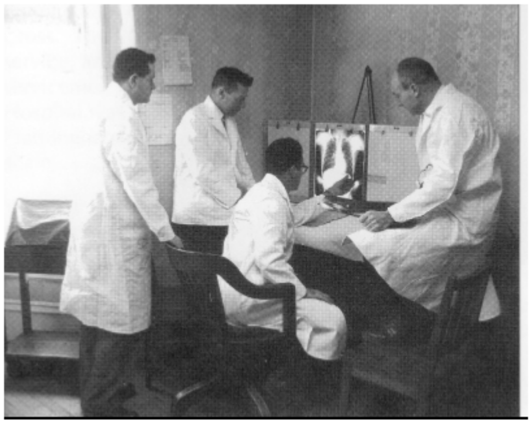
- Community Health Outreach (NJMS APSEA)
- Prevalence of cardiovascular ailments
- Causation between Cardiovascular ailments and risk factors is not fully understood
- **Importance:** Heart Disease Detection, Prevention, and Treatment

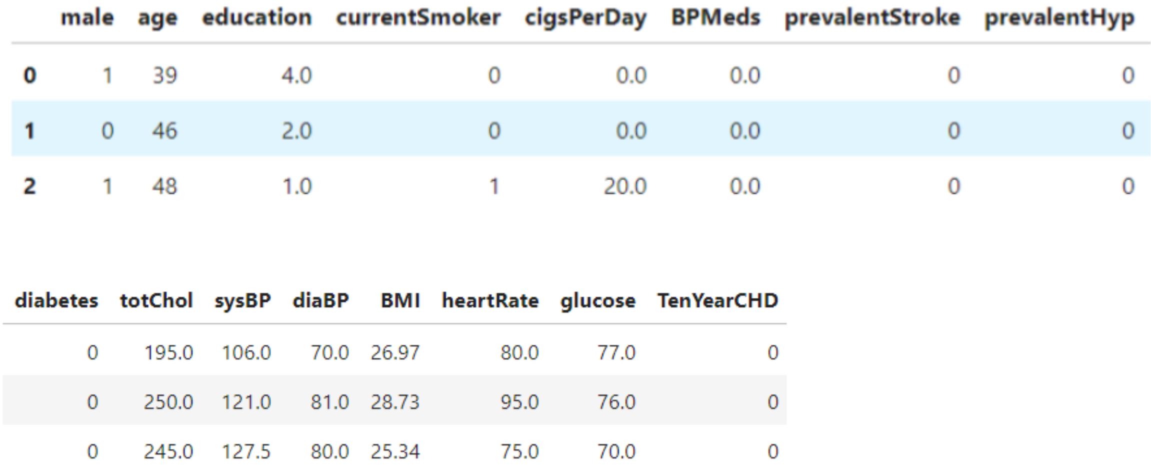
Primary Question: How do cardiovascular assessment features (e.g., blood pressure, cholesterol, heart rate) of patients who smoke vary from those who do not?

Dataset

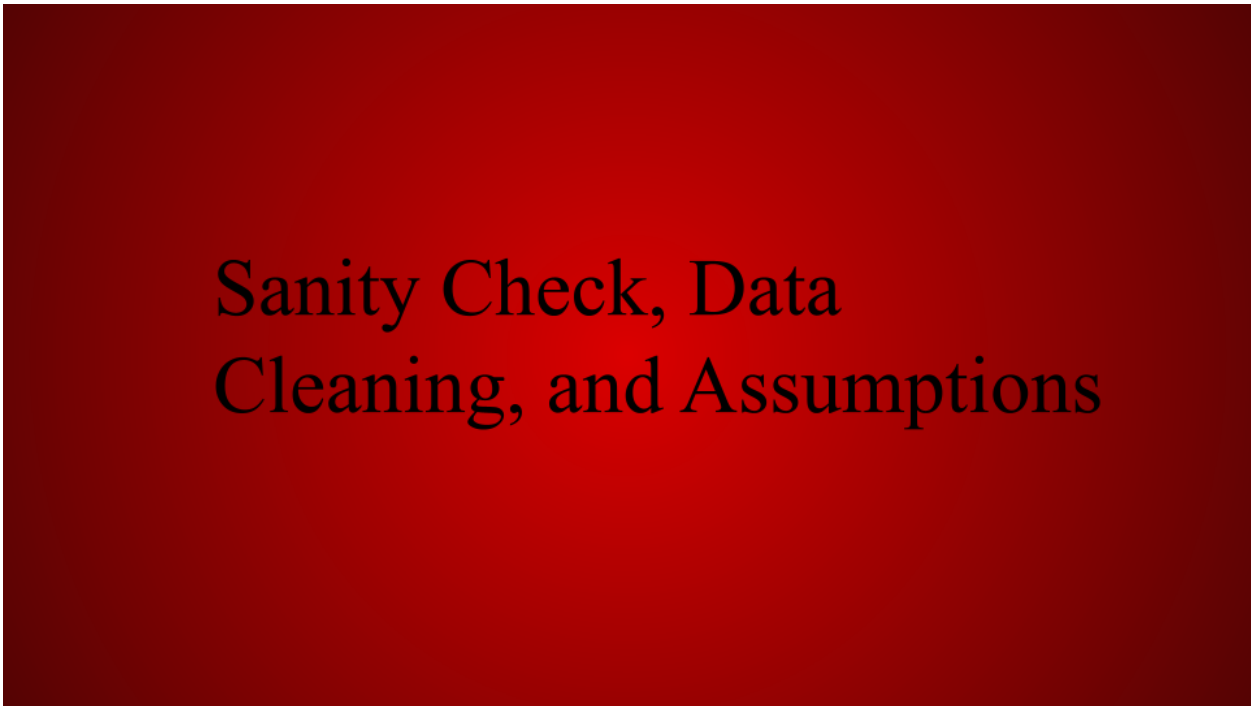
Framingham Heart Study

- Longitudinal Design
- Multigenerational
- 4,000 + records and 15 features
- Cardiovascular Events
- Cardiovascular Assessment Features
- Shaped clinical research surrounding Cardiovascular Health
- Kaggle Data Set extracts and compiles data from the cohort study, and organizes it into a CSV file





	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
0	1	39	4.0	0	0.0	0.0	0	0	0	195.0	106.0	70.0	26.97	80.0	77.0	0
1	0	46	2.0	0	0.0	0.0	0	0	0	250.0	121.0	81.0	28.73	95.0	76.0	0
2	1	48	1.0	1	20.0	0.0	0	0	0	245.0	127.5	80.0	25.34	75.0	70.0	0

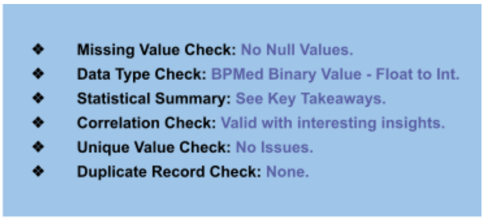


Sanity Check, Data Cleaning, and Assumptions

Assumptions

- Accurate and consistent self-reporting
- Dataset completeness
- Duplicate records were data entry error
- Typical distribution of cardiovascular risk factors
- Other healthcare and lifestyle factors unaccounted for, but may influence cardiovascular health
- Outliers that were physiologically feasible were likely anomalies

Sanity Check & Data Cleaning

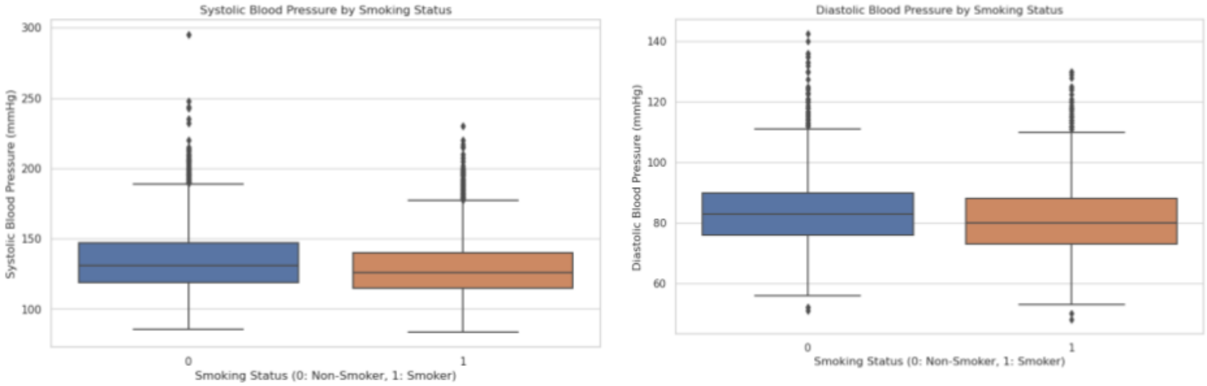


- **Missing Value Check:** No Null Values.
- **Data Type Check:** BPMed Binary Value - Float to Int.
- **Statistical Summary:** See Key Takeaways.
- **Correlation Check:** Valid with interesting insights.
- **Unique Value Check:** No Issues.
- **Duplicate Record Check:** None.

1. Age range 32 to 70 years old.
2. CigsPerDay max of 70, about 3.5 packs per day.
3. BPMeds, PrevalentStroke, PrevalentHyp, Diabetes, TenYearCHD all are binary values.
4. totCholesterol max value 600.
5. sysBP and diaBP - max sysBP is 295 and max diaBP is 142.5.
6. BMI range is 15 to 56.
7. heartRate - ranges from 44 to 143.
8. Glucose - max was 394.

Blood Pressure

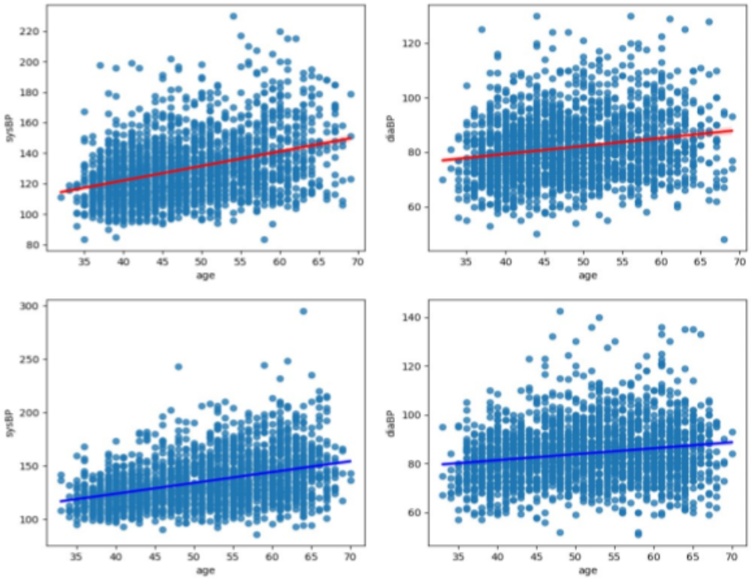
Initial Exploration: Blood Pressure vs. Smoking Status

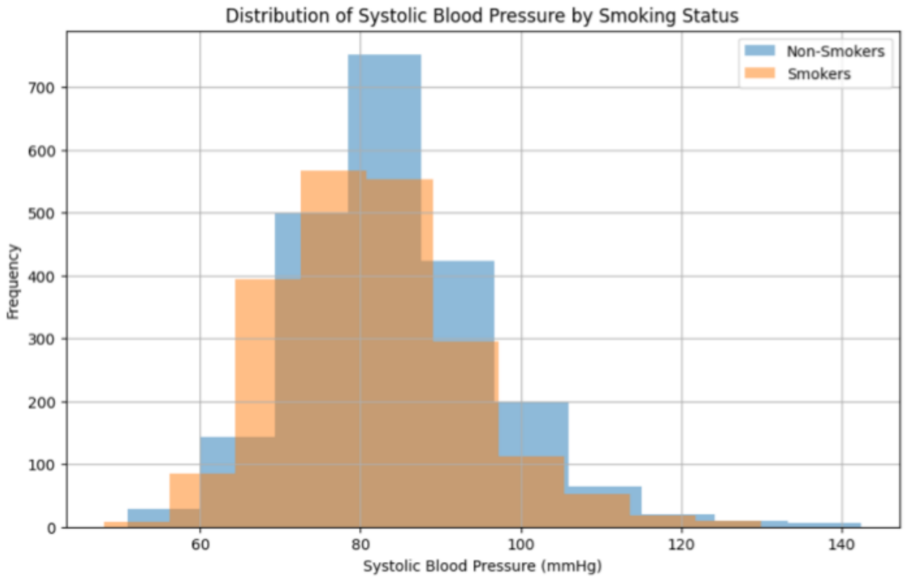


How does Blood Pressure change with age between smokers and non smokers?

Smokers

Non Smokers



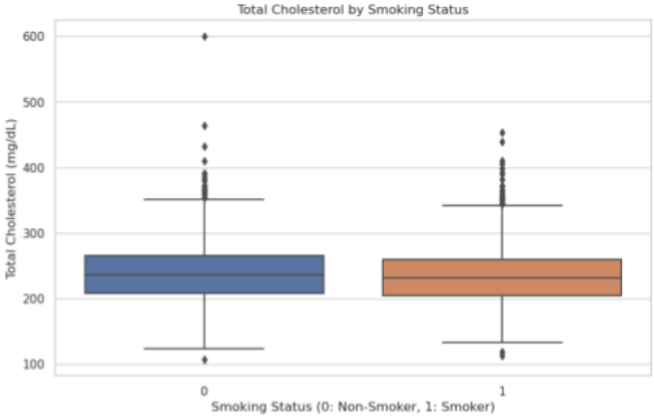


Cholesterol

Initial Exploration: Cholesterol vs. Smoking Status

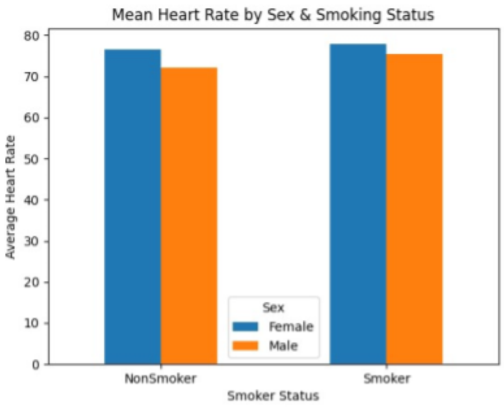
- Marked variability among the non-smoker group; wider interquartile range.

- A few of the high variables we noticed in our sanity check are present in non-smoker group.



Heart Rate

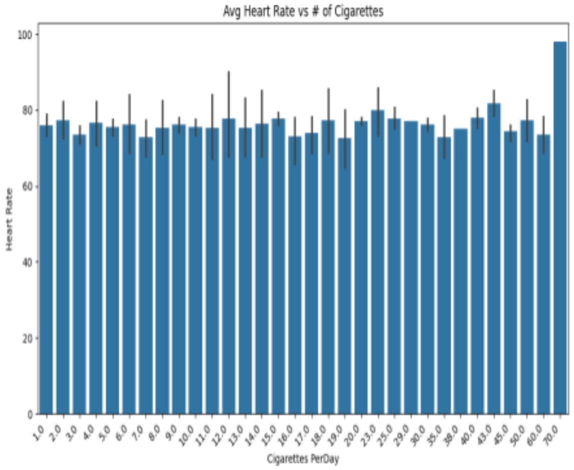
1. What is the mean heart rate of men and women who do and do not smoke?
2. What is the average heart rate per smoking status, per gender, and 10-year risk of future (CHD) coronary heart disease?

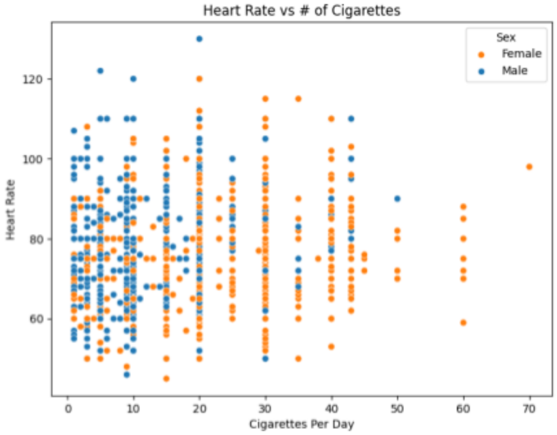


Avg Heart Rate

Smoking Status	Sex	No_CHD	CHD_Risk	All
NonSmoker	Female	76.61581921	76.32291667	76.57651992
	Male	72.03193277	73.30252101	72.24369748
Smoker	Female	77.6375	79.58715596	77.85237614
	Male	75.20975057	76.93721973	75.55837104
All		75.7625139	76.53032659	75.87898089

How does heart rate differ based on the number of cigarettes a patient consumes?





Next Steps

- Our data provides a framework for further exploring the relationship between adverse risky health behaviors and cardiovascular features
- Developing risk models
- Statistical Significance Testing
- Education and Awareness Programs