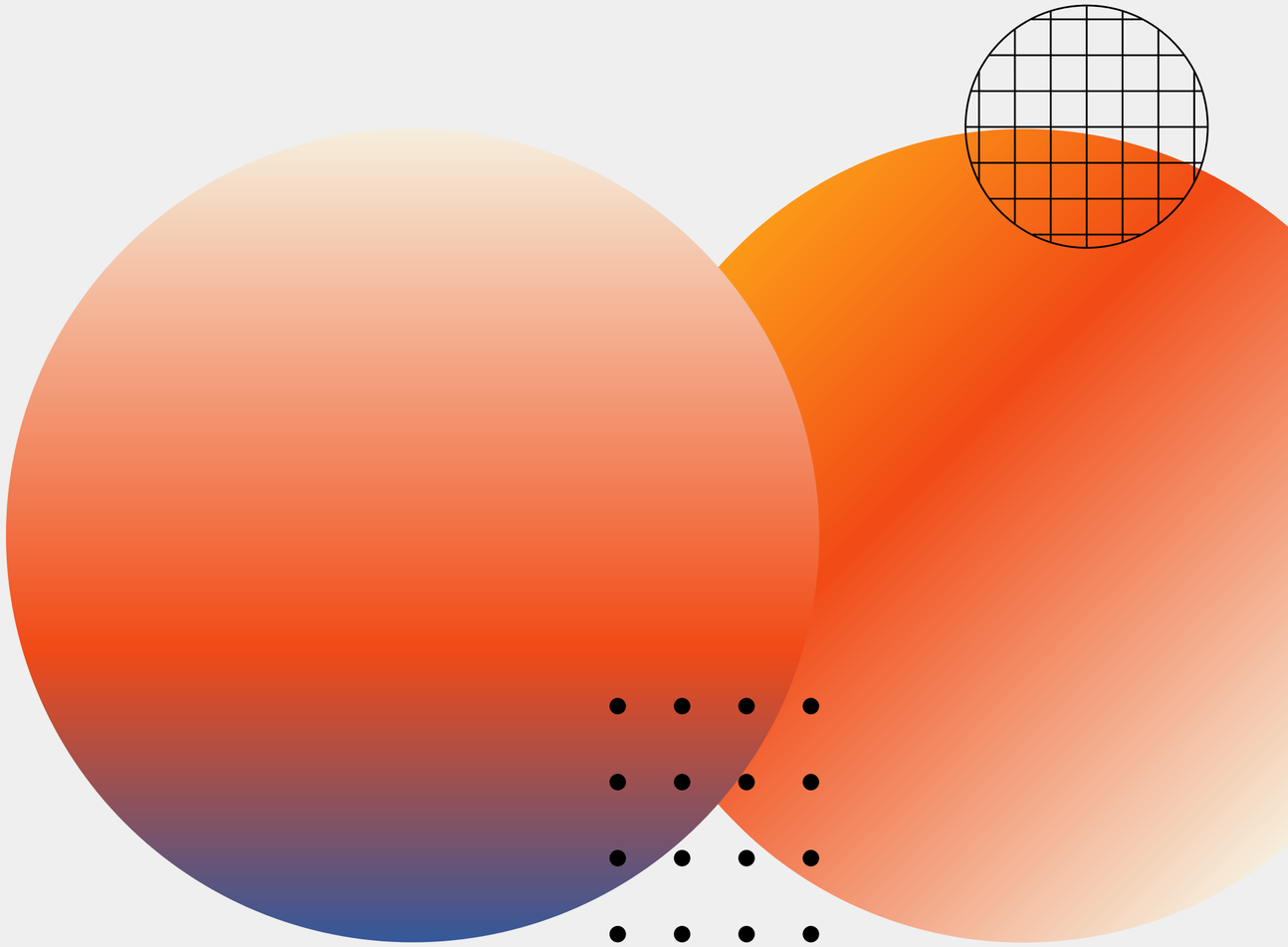
Cardiovascular Disease Risk Factors Analysis

by Jo Waters



Analysis Overview

Raw Look at the Data
Checking for Null Data
Formatting and Cleaning
Initial Patient Analysis
Guiding Questions
Conclusion

Fast Facts About Cardiovascular Diseases (CVDs)

1. Leading cause of death globally
2. Refers to a variety of conditions including stroke, heart attacks, and heart failure
3. In 2019, CVDs made up 32% of all global deaths

First Look

Our initial table has several data types including those recorded by medical personnel during intake, behaviors reported by patients, and CVD diagnosis status. Some of these are scale values, binary values, or just the recorded value. Looking at our available keys and table we have some units that aren't the most useful or readable, including age in days rather than years. We should clean some of this up to understand our data better.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 70000 entries, 0 to 69999
Data columns (total 13 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   id           70000 non-null  int64  
 1   age          70000 non-null  int64  
 2   gender       70000 non-null  int64  
 3   height       70000 non-null  int64  
 4   weight       70000 non-null  float64
 5   ap_hi        70000 non-null  int64  
 6   ap_lo        70000 non-null  int64  
 7   cholesterol  70000 non-null  int64  
 8   gluc         70000 non-null  int64  
 9   smoke        70000 non-null  int64  
 10  alco         70000 non-null  int64  
 11  active       70000 non-null  int64  
 12  cardio       70000 non-null  int64  
dtypes: float64(1), int64(12)
memory usage: 6.9 MB
```

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	3	17623	2	169	82.0	150	100	1	1	0	0	1	1
4	4	17474	1	156	56.0	100	60	1	1	0	0	0	0

```
id                0
age               0
gender            0
height            0
weight            0
ap_hi             0
ap_lo             0
cholesterol       0
gluc              0
smoke             0
alco              0
active            0
cardio            0
dtype: int64
```

Checking for Null

I ran a summation of null values to see where we are missing data. Thankfully we have all data points for all 70,000 patients. From here we need to begin to format our data for ease of use and reading.

Formatting Objectives

Convert to useful units

Make binary information more intuitive

Add columns for other helpful measures

Create separate dataframes for legibility

Formatting Continued

Age

Below is the UDF I made to convert age in days to years.

```
def day_to_yr(days):
    year = days // 365
    return year
```

Age Group

This function sorts our now age in years into age groups. The groups are based

```
def age_grouper(the_age):
    if the_age < 18:
        group = 'Adolescent'
    elif the_age < 24:
        group = 'Young Adult'
    elif the_age < 45:
        group = 'Adult'
    elif the_age < 65:
        group = 'Middle Adult'
    else:
        group = 'Older Adult'
    return group
```

BMI

While an imperfect measure, I added a column with calculated BMI to account for the relationship between height and weight

```
def hw_to_bmi(height, weight):
    bmi = weight / ((height/100)**2)
    return bmi
```

Binary

To help in the legibility of binary data and double checking my work, I formatted the binary behavioral values into strings.

```
def to_true_false(binary):
    t_or_f = 'True'
    if binary == 0:
        t_or_f = 'False'
    return t_or_f
```

```
def get_gender(gen_num):
    gender = 'Female'
    if gen_num == 2:
        gender = 'Male'
    return gender
```

Gender

Similar to the binary function this changes our non-base-zero binary values to the appropriate strings.

Refreshed Look

Now that we have formatted let's take another look at our numbers.

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio	age_group	bmi
0	0	50	Male	168	62.0	110	80	1	1	False	False	True	False	Middle Adult	21.967120
1	1	55	Female	156	85.0	140	90	3	1	False	False	True	True	Middle Adult	34.927679
2	2	51	Female	165	64.0	130	70	3	1	False	False	False	True	Middle Adult	23.507805
3	3	48	Male	169	82.0	150	100	1	1	False	False	True	True	Middle Adult	28.710479
4	4	47	Female	156	56.0	100	60	1	1	False	False	False	False	Middle Adult	23.011177

Patient Quick Hits

age

Unit : years
Mean: 52.84
IQR: 10

gender

of men: 24,470
# of women: 45,530

height

Unit : cm
Mean: 164.36
IQR: 11

weight

Unit : kg
Mean: 74.21
IQR: 17

BMI

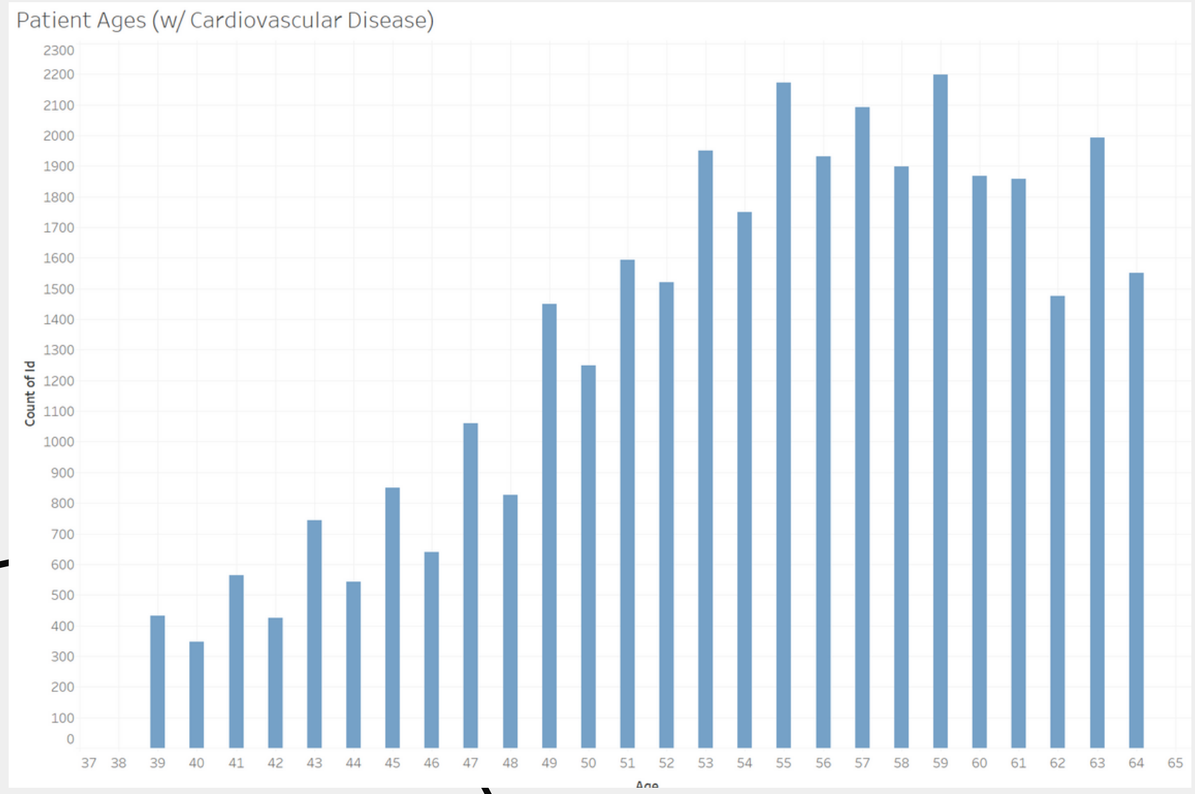
Unit : kg/m^2
Mean: 27.56
IQR: 6.35

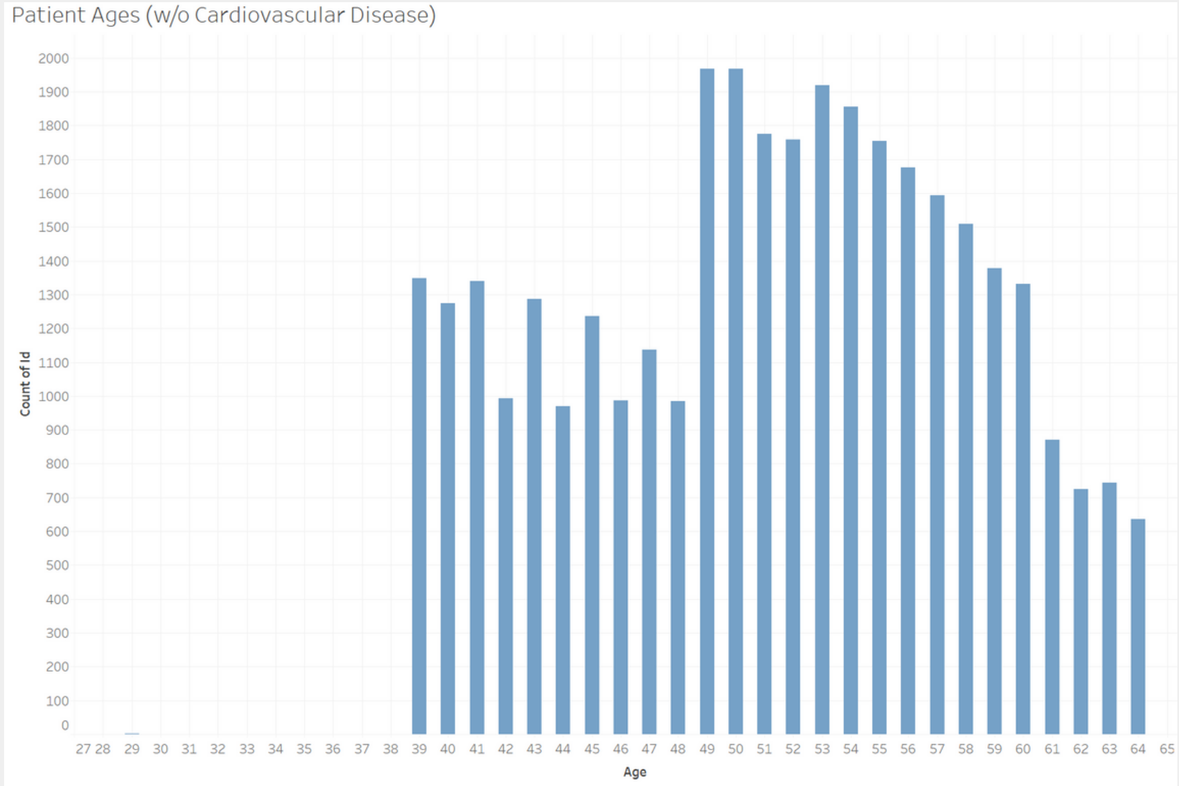
CVD Diagnosis

with CVD: 34,979
without CVD: 35,021

Age

Average age of those with CVD: 54.45
Average age of those without CVD: 51.23





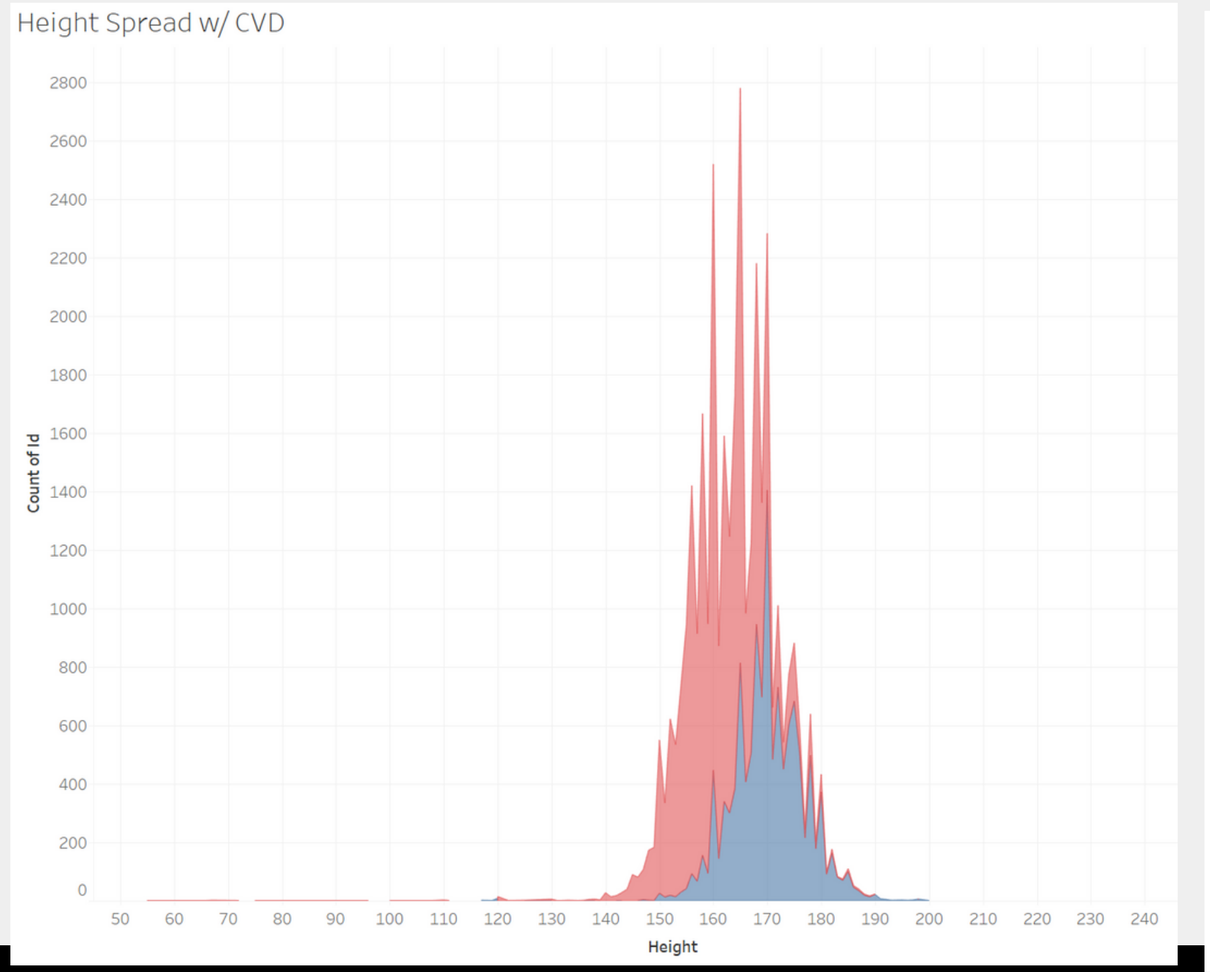
Gender

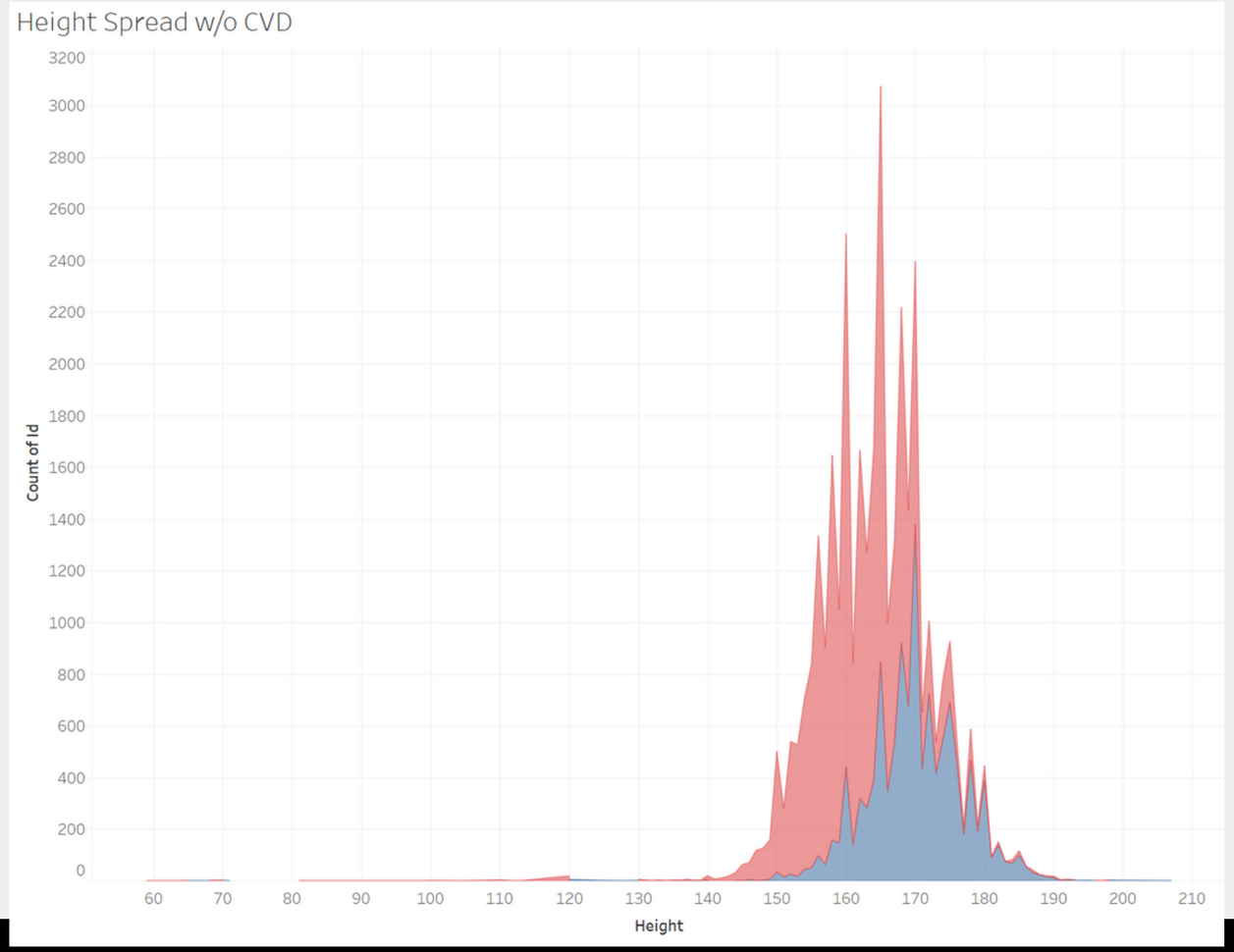
Rates of CVD are roughly even across our patient genders.

Females without CVD: 22,914	Male without CVD: 12,107
Females with CVD: 22,616	Males with CVD: 12,363

Height

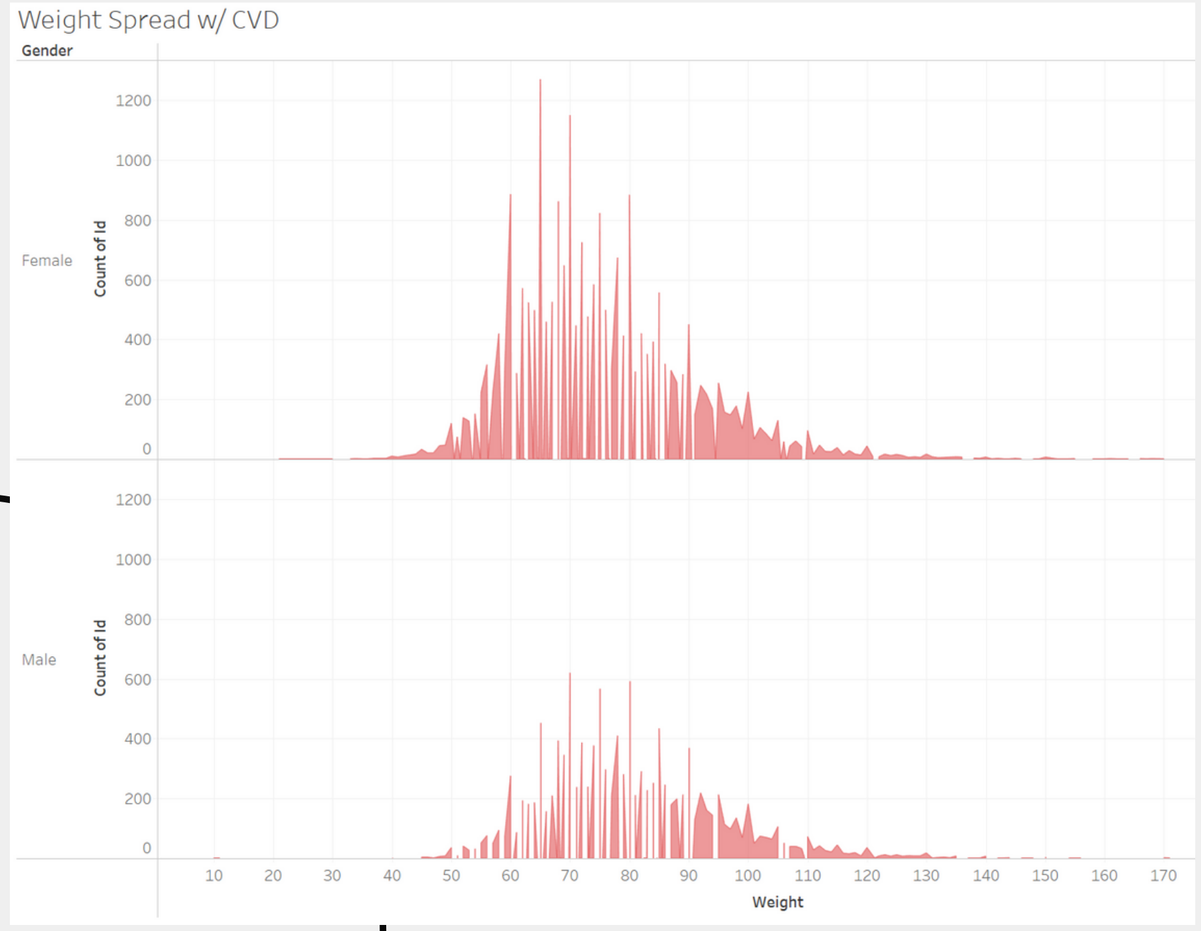
The mean height for both those with and without CVD is nearly the same

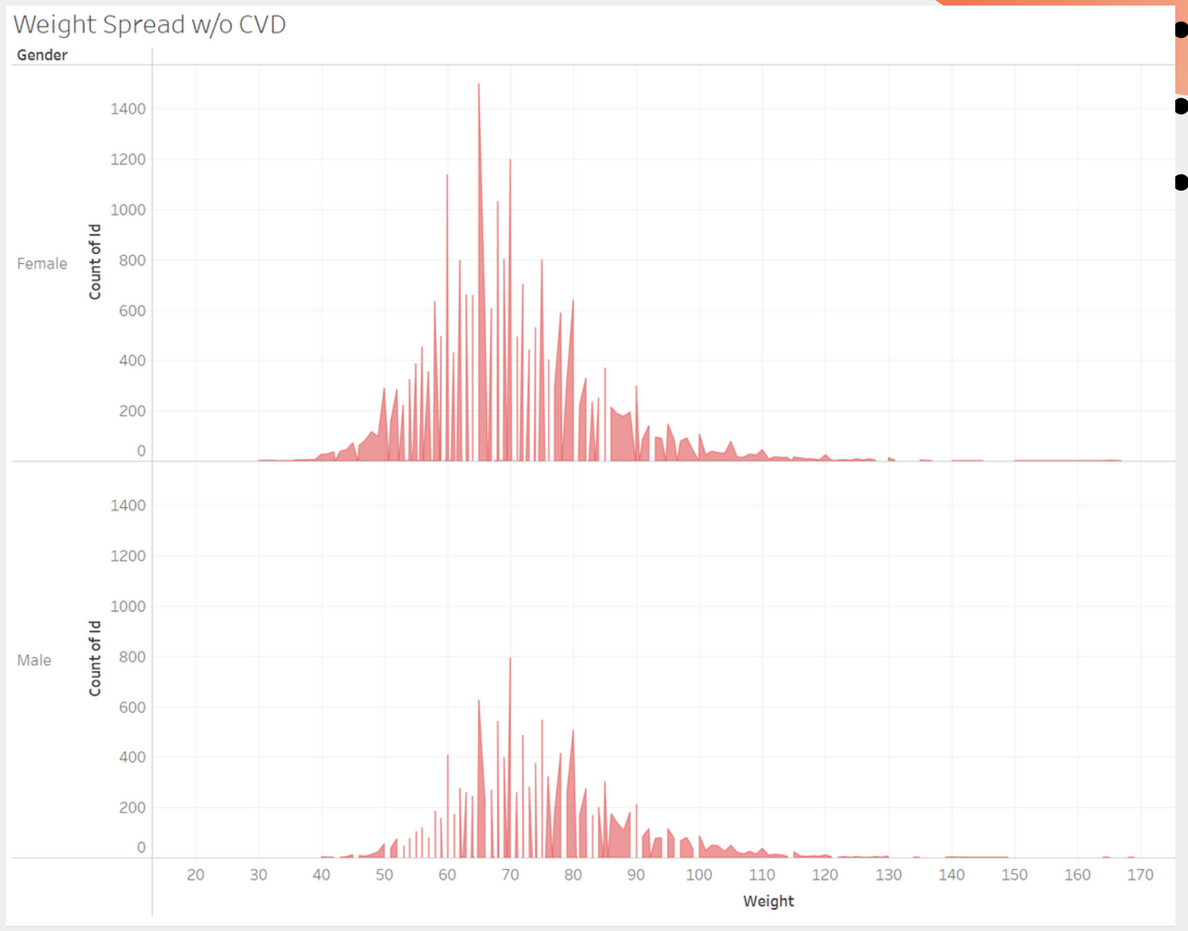




Weight

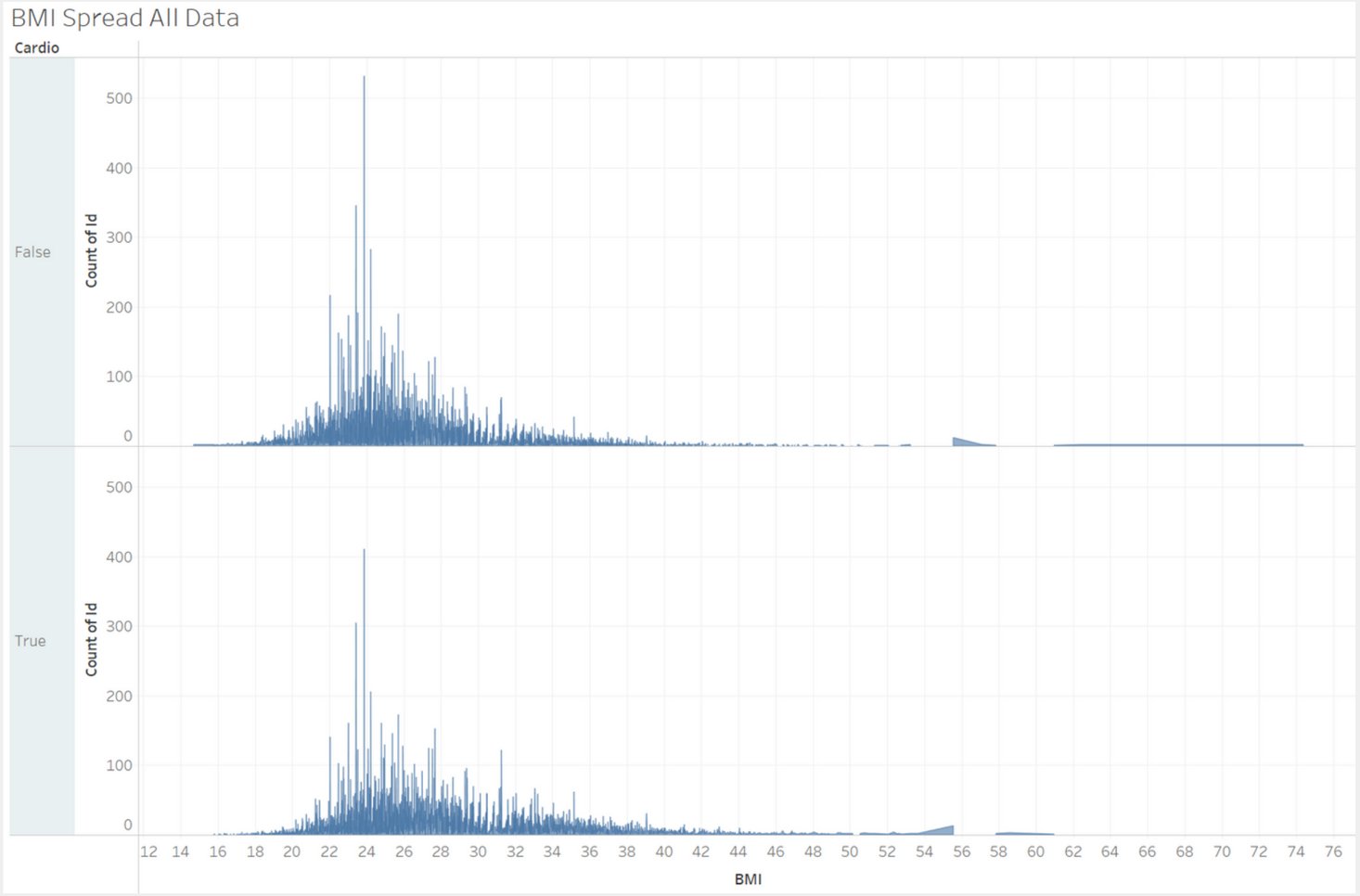
The average weight of those with CVD is 5kg higher than those without.

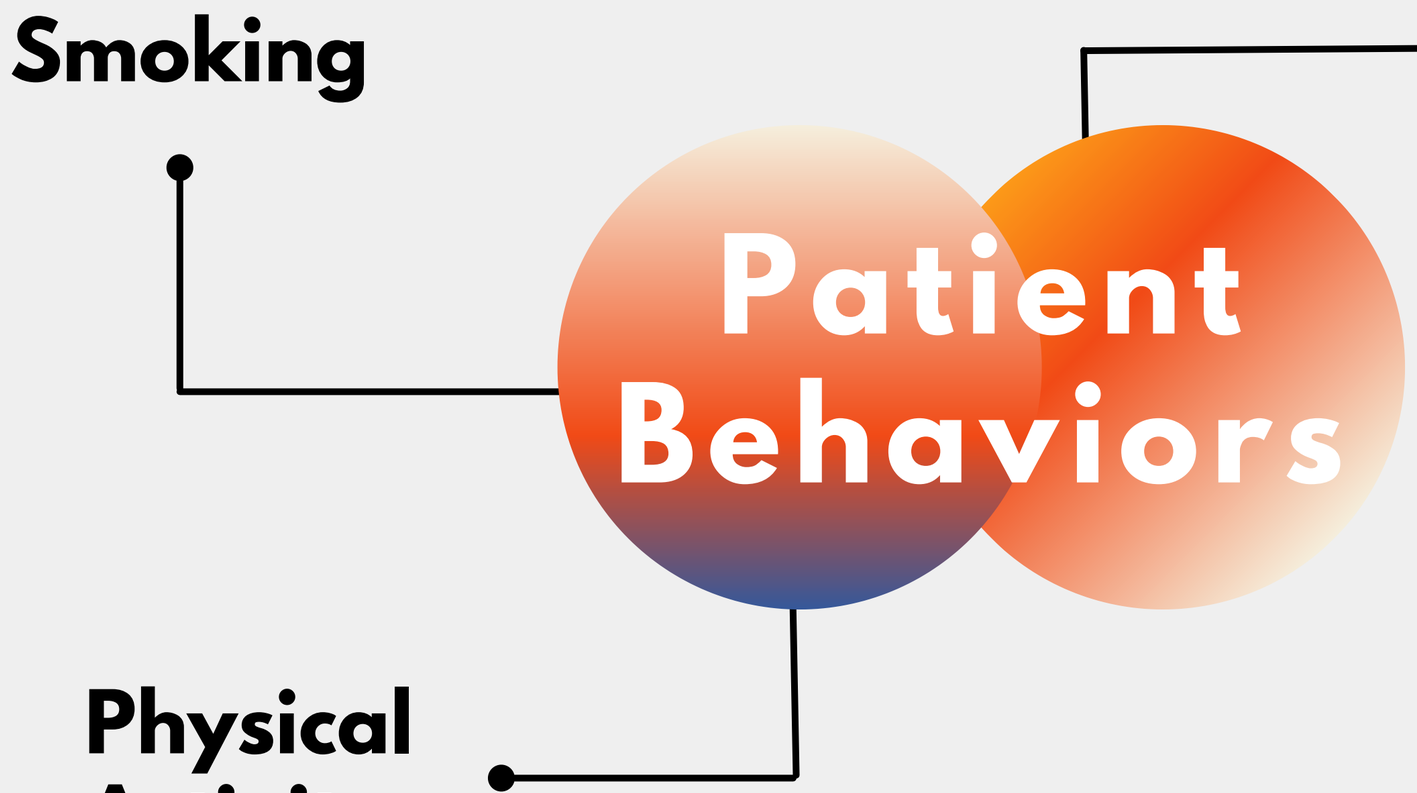




BMI

The average BMI for patients with cardiovascular disease is 2 kg/m2 higher than their counterparts.

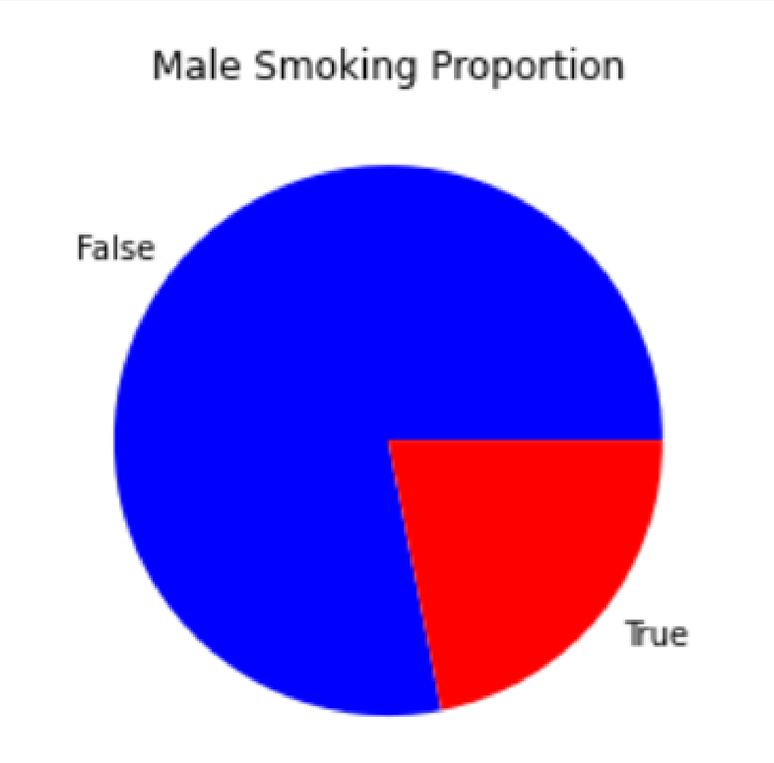


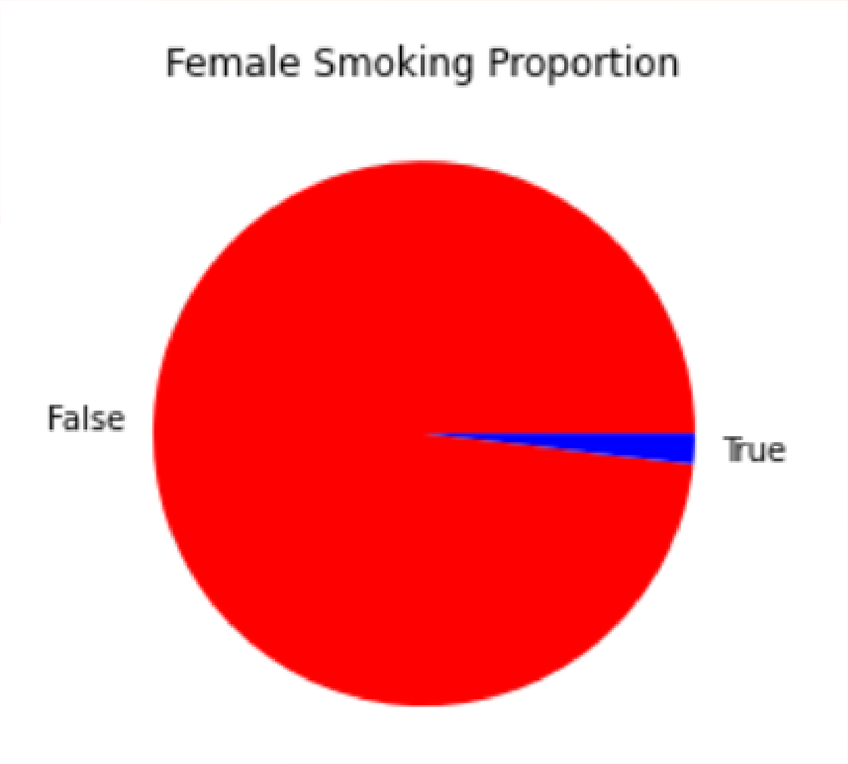


All of these patient behaviors are self reported and are represented with binary values.

Smoking

A far greater proportion of men smoke than women, and within each gender the rates of CVD are roughly the same across smoking and non-smoking groups.





Alcohol Use

Similar to smoking, a larger proportion of men consume alcohol than women, and these rates are roughly the same across those with and without CVD.

Active

The rates of patients reporting physical activity are the same across genders. Rates of CVD are also the same across those who are physically active and those who are not.

CVD v. Active Lifestyle

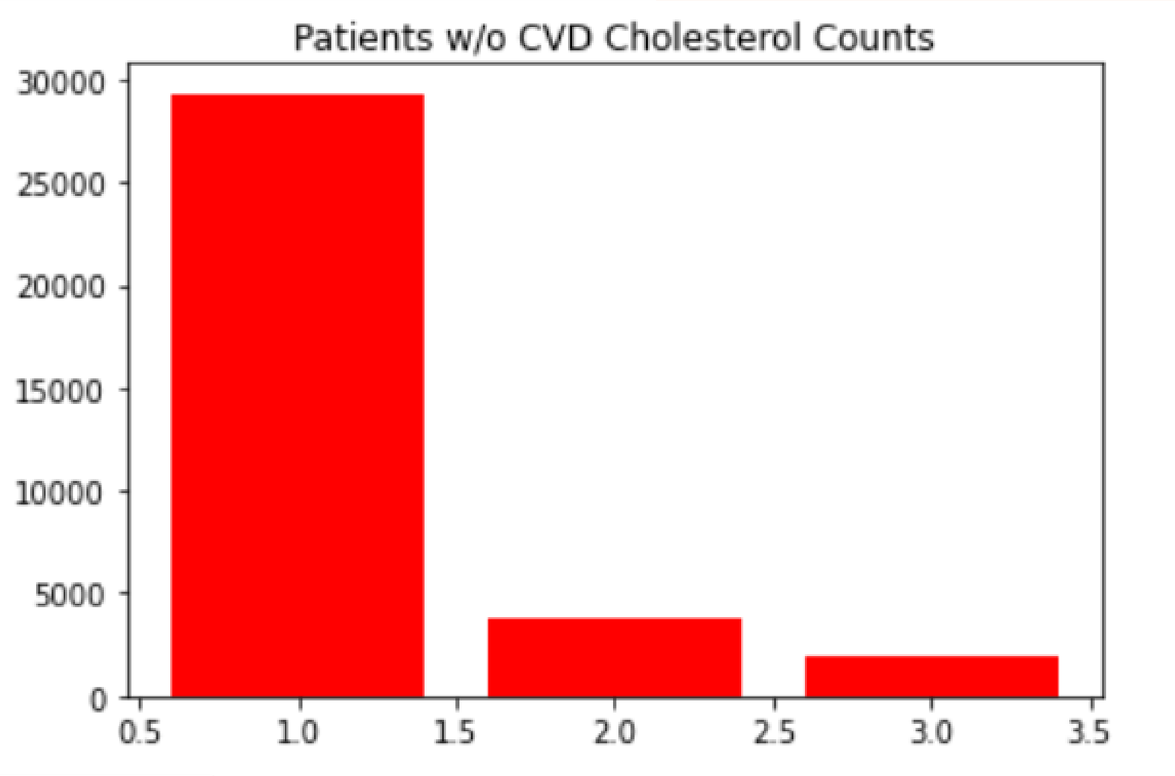
	Cardio	
Active	False	True
False	6,378	7,361
True	28,643	27,618

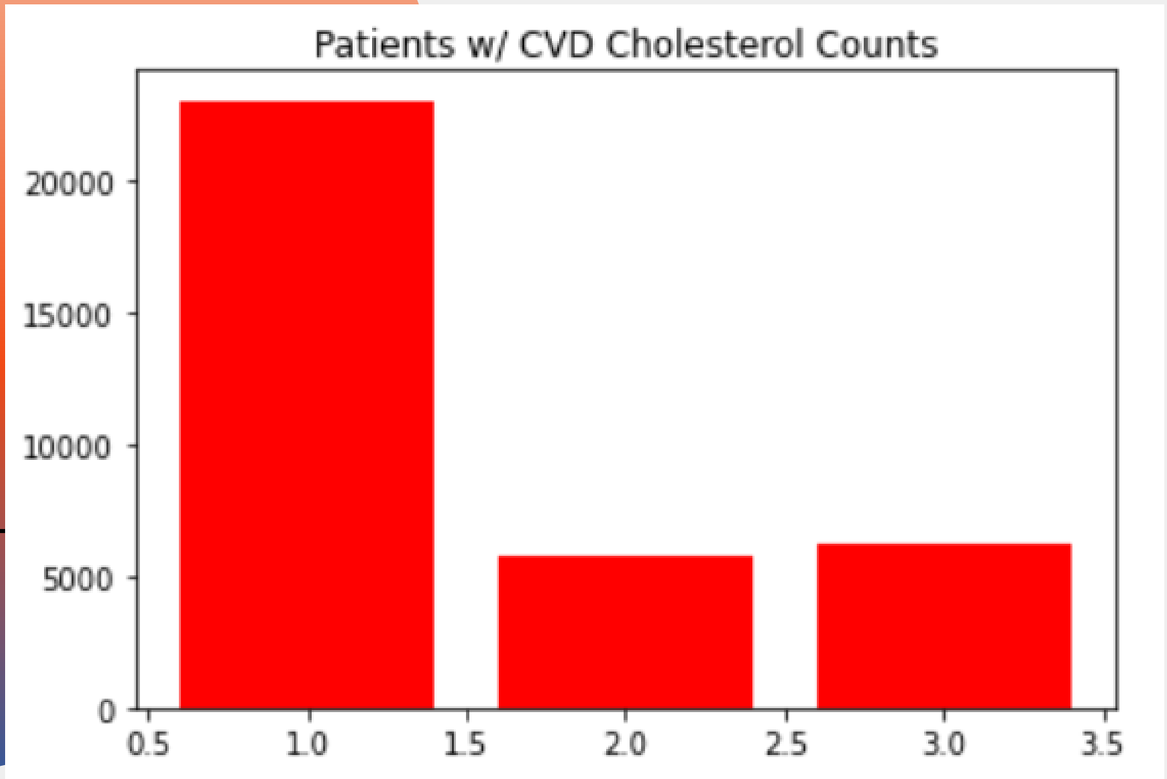
Cardio Observations

1. Cholesterol Levels-
   1: normal, 2: above normal, 3: well above normal
2. Glucose Levels-
   1: normal, 2: above normal, 3: well above normal
3. Systolic (ap_hi) and Diastolic (ap_lo) Blood Pressure

Cholesterol Levels

The proportion of patients with above normal and well above normal cholesterol levels is greater in the patients with CVD.





Blood Pressures

Normal Values

- American Heart Association

Systolic	Diastolic
<= 120	<= 80

Patients without CVD

Mean Values

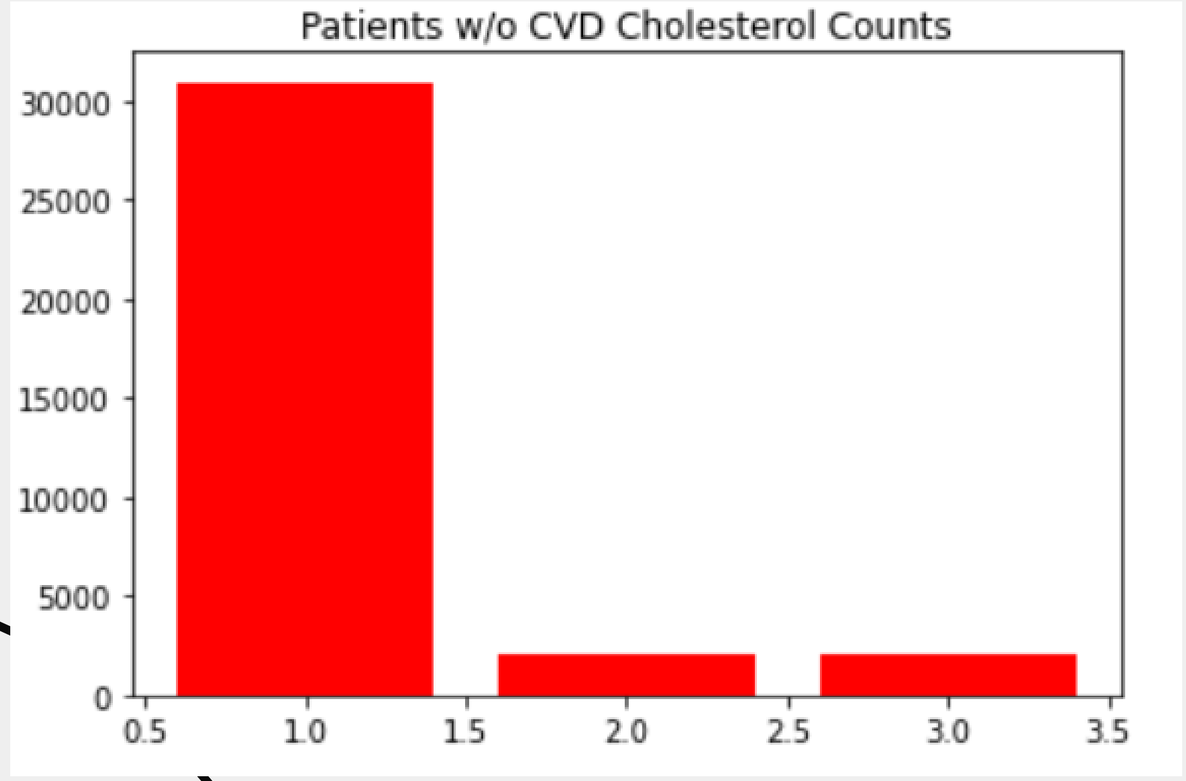
Systolic	Diastolic
120.43	84.25

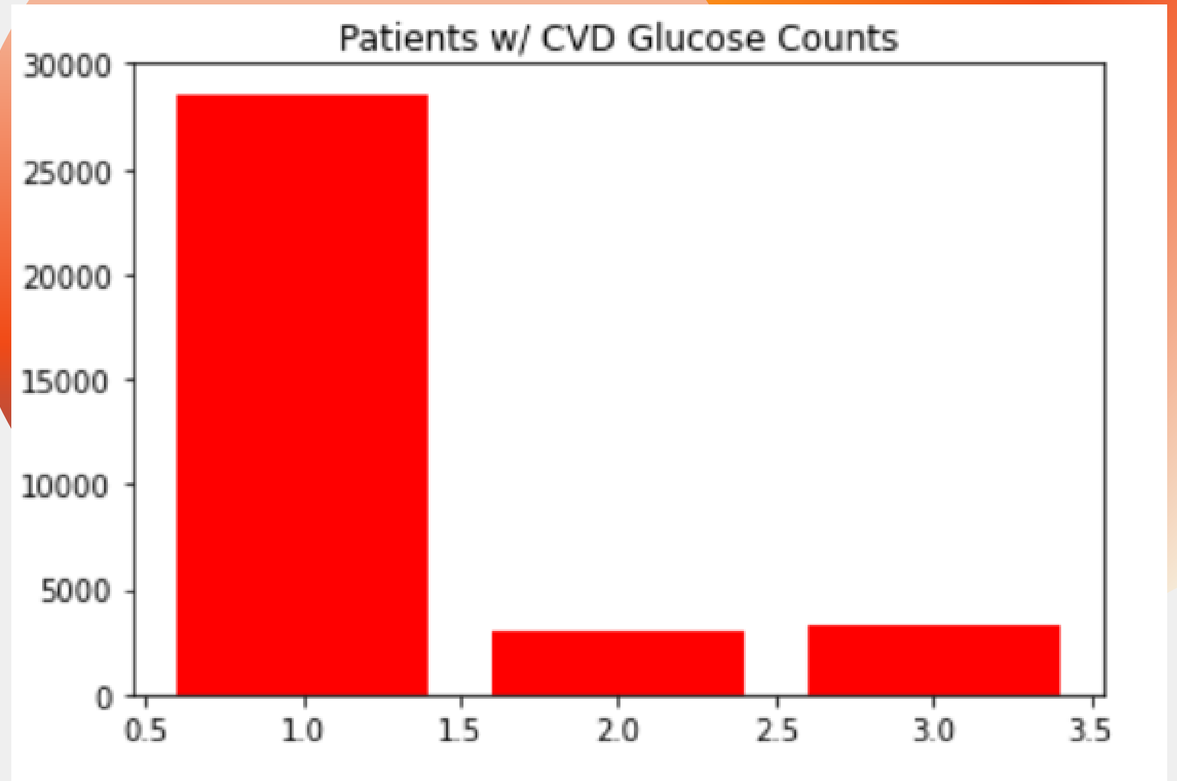
Patients with CVD

Mean Values

Systolic	Diastolic
137.21	102.02

Glucose Levels





Patients with CVD Quick Hits

age

Unit : years
Mean: 54.45
IQR: 9

gender

of men: 12,363
# of women: 22,616

height

Unit : cm
Mean: 164.27
IQR: 19

weight

Unit : kg
Mean: 76.82
IQR: 19

BMI

Unit : kg/m^2
Mean: 28.56
IQR: 7

Question 1

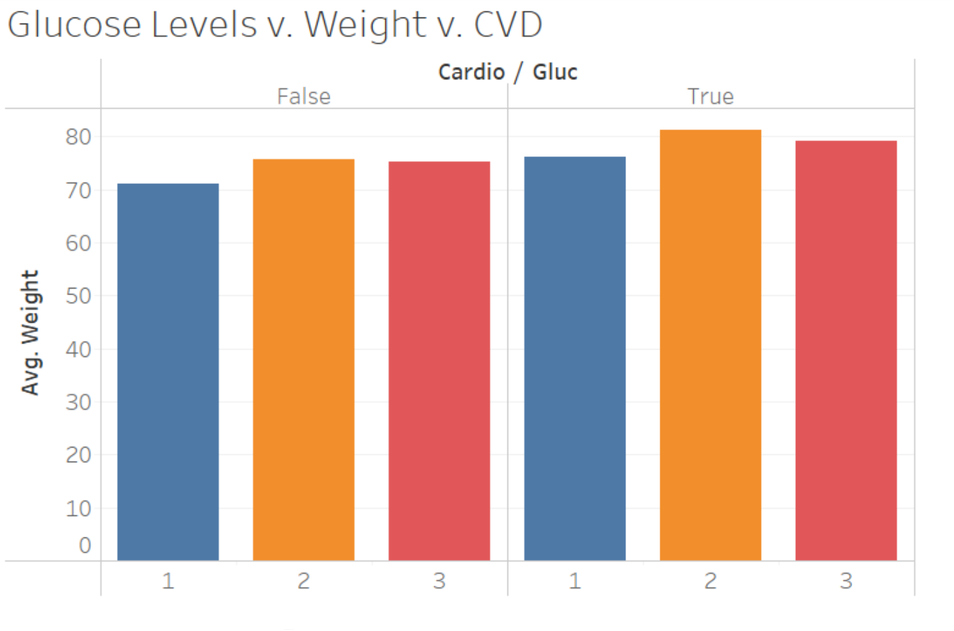
Relationship between Blood Pressure, Active Lifestyle, and CVD rates.

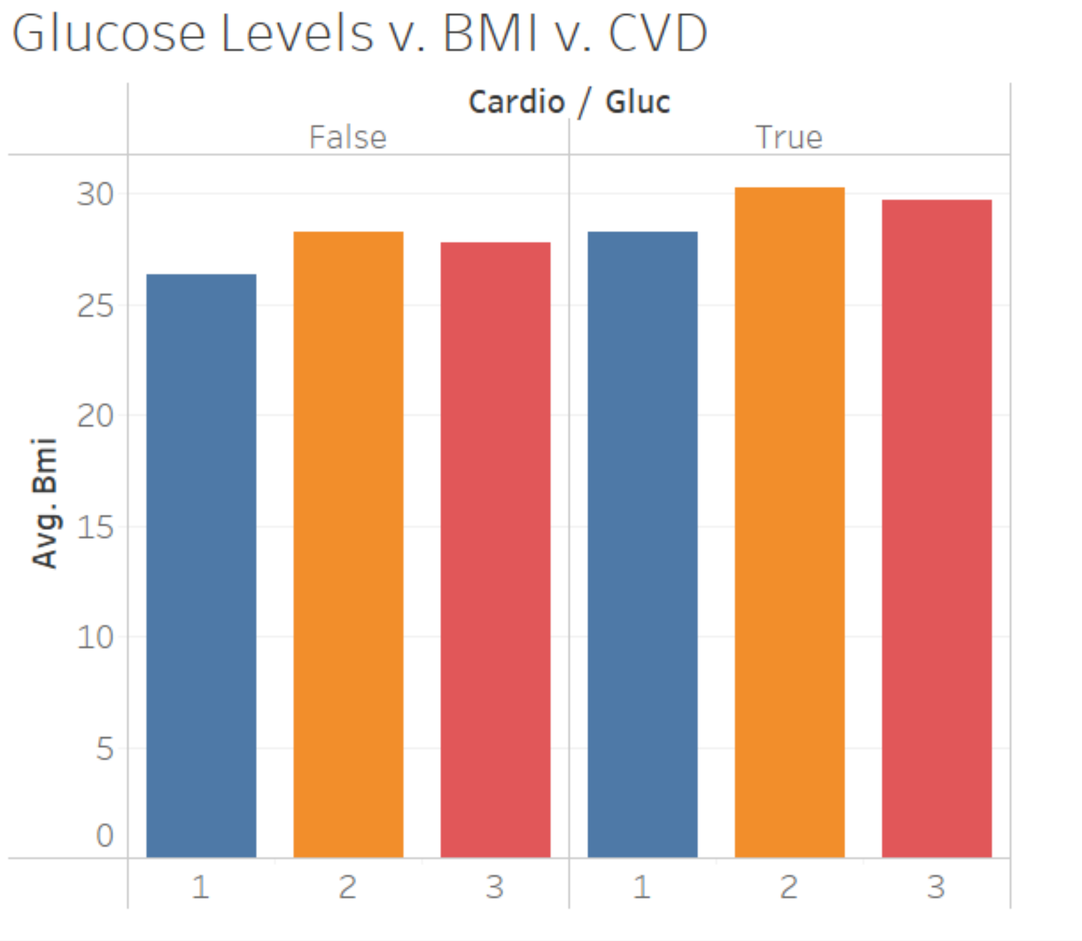
Patients without CVD	Active	Not Active	Overall
Systolic	120.18	121.55	120.43
Diastolic	87.05	83.63	84.25

Patients with CVD	Active	Not Active	Overall
Systolic	137.77	135.13	137.21
Diastolic	111.02	101.53	109.02

Question 2

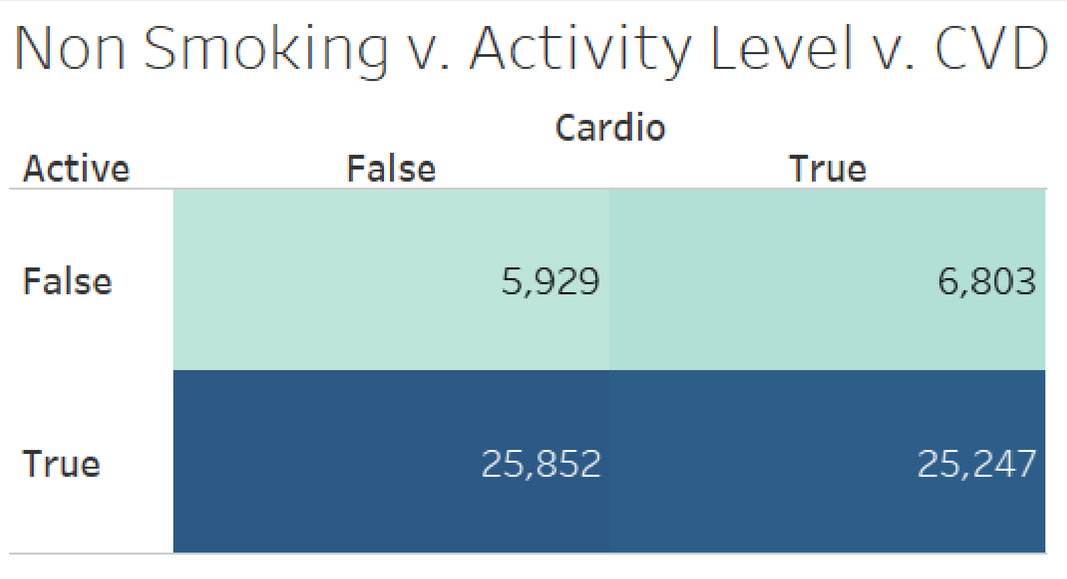
Glucose Levels v. Weight
Measures v. CVD

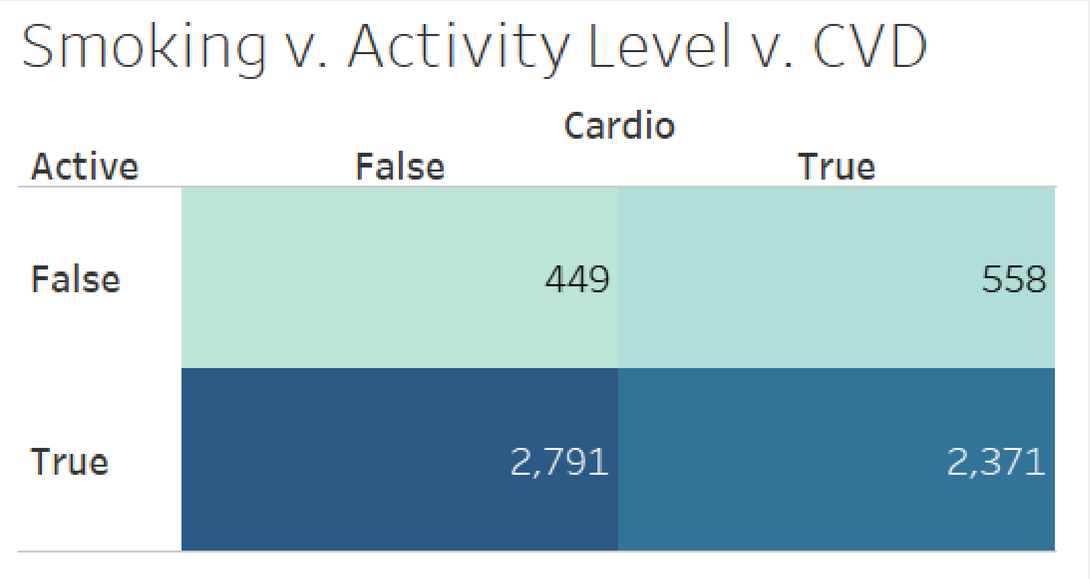




Question 3

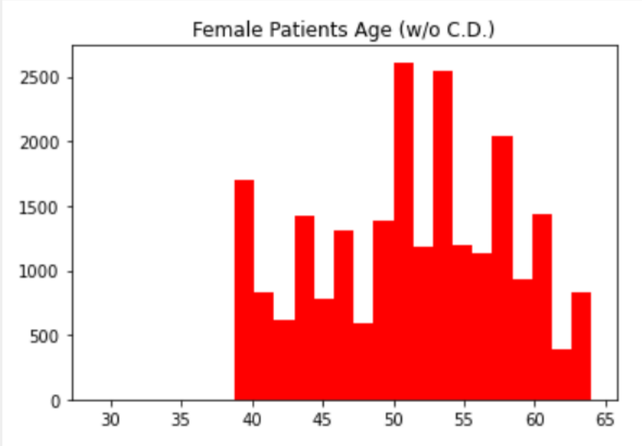
Smoking v. Physical Activity v. CVD

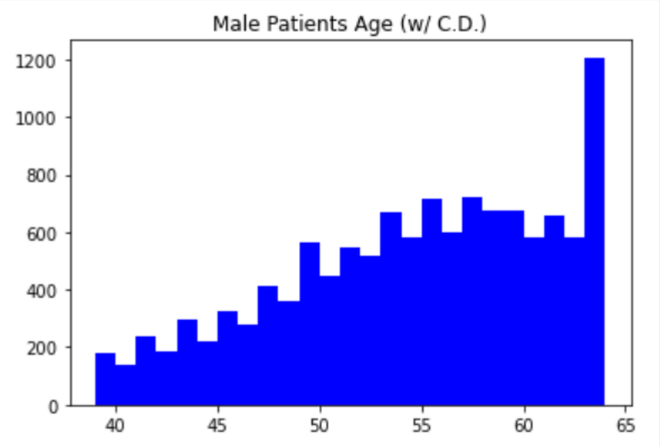




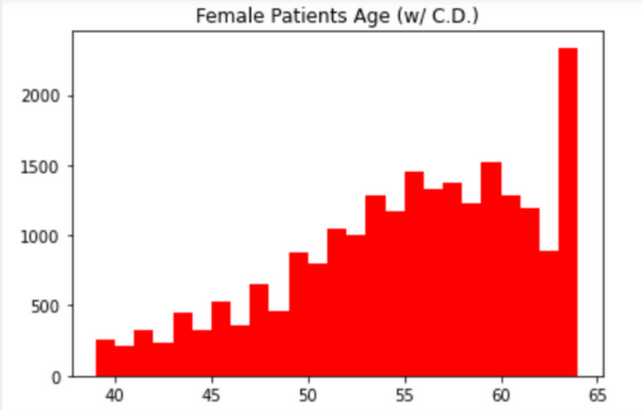
Question 4

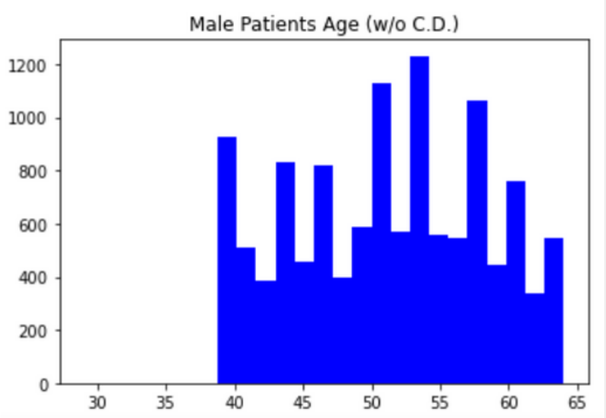
Gender v. Age v. CVD





Female Average Ages
With CVD: 54.66
Without CVD: 51.27

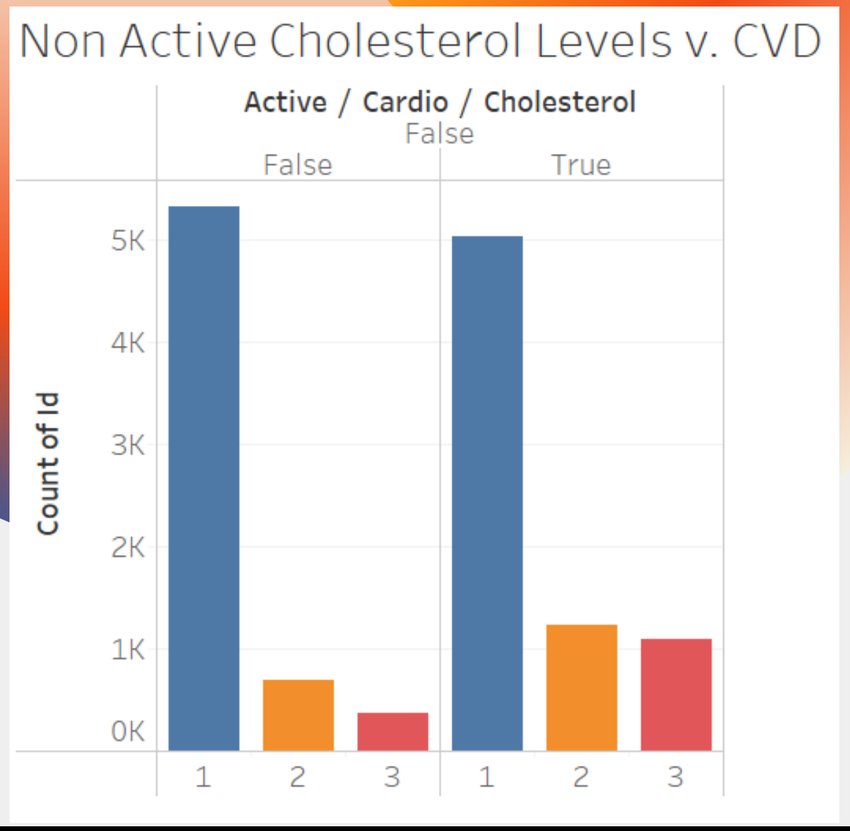


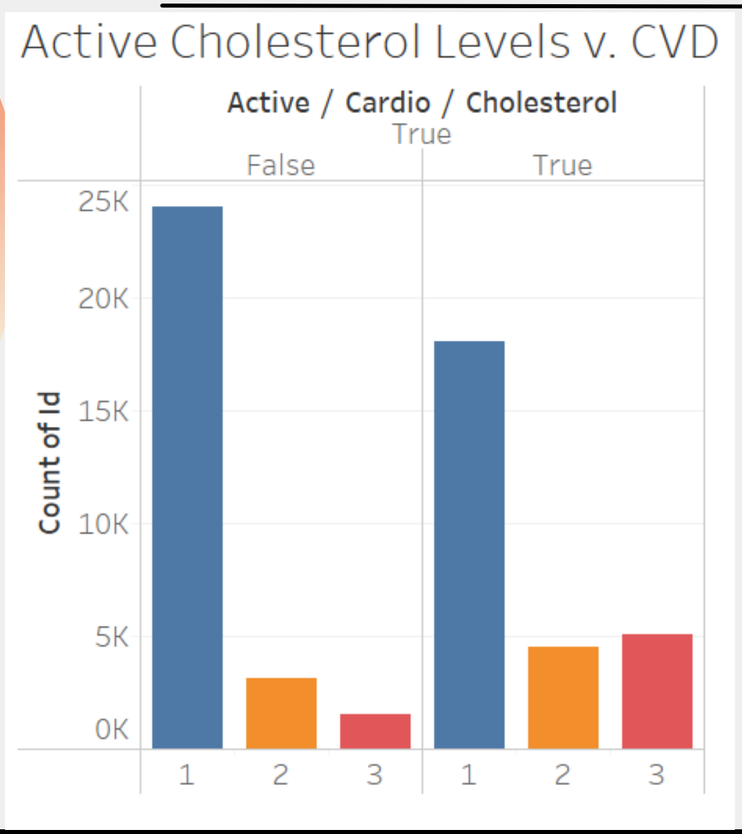


Male Average Ages
With CVD: 54.07
Without CVD: 51.16

Question 5

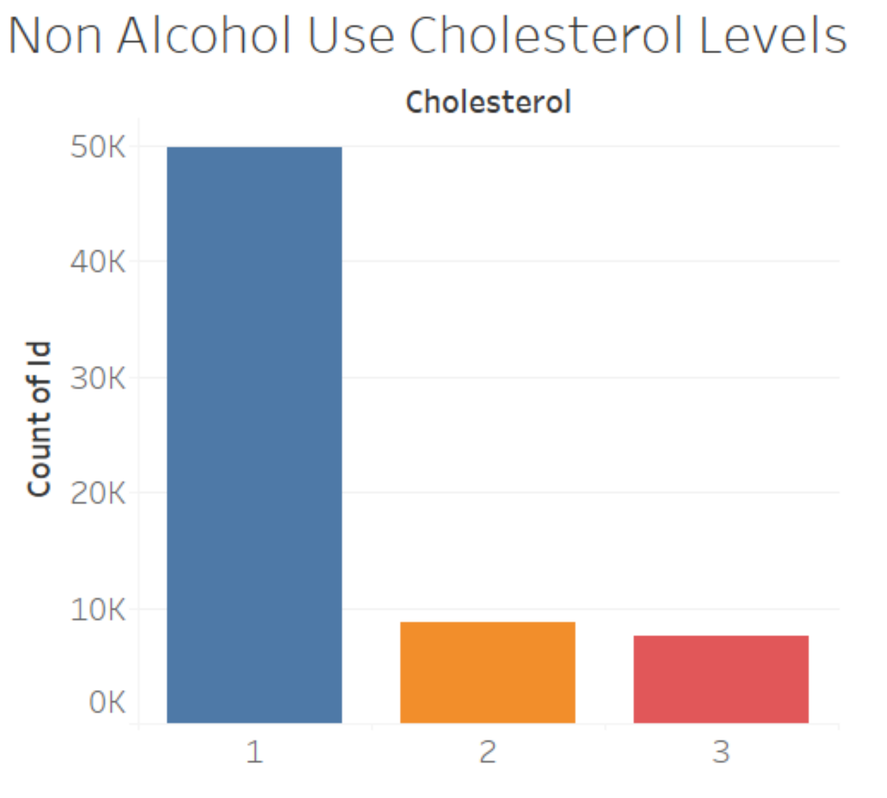
Physical Activity v. Cholesterol v. CVD

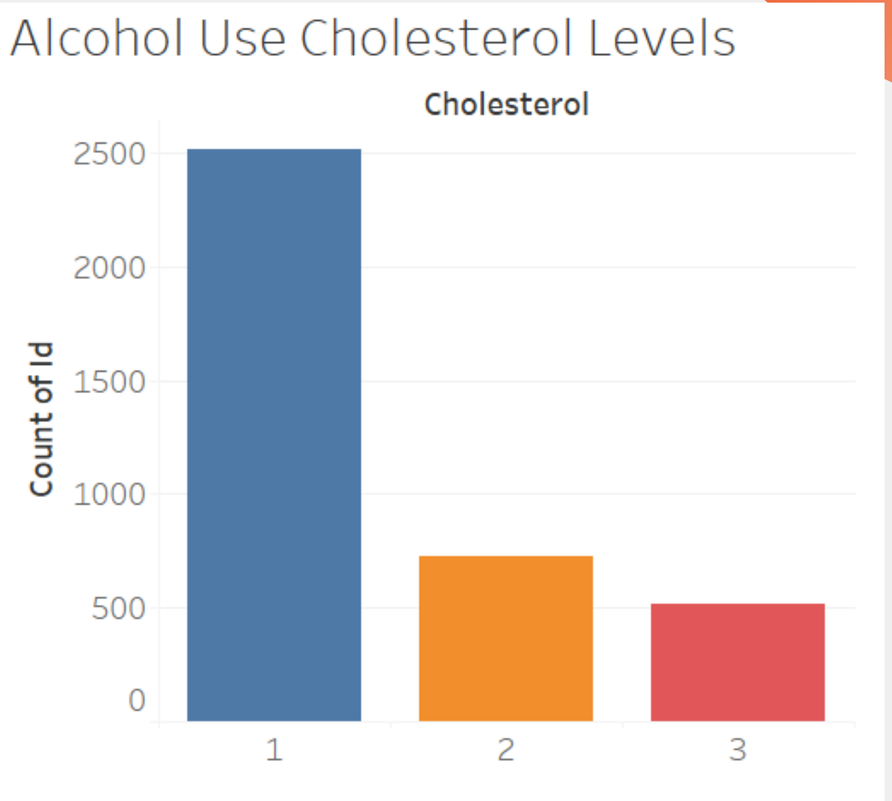




Question 6

Alcohol v. Cholesterol Levels



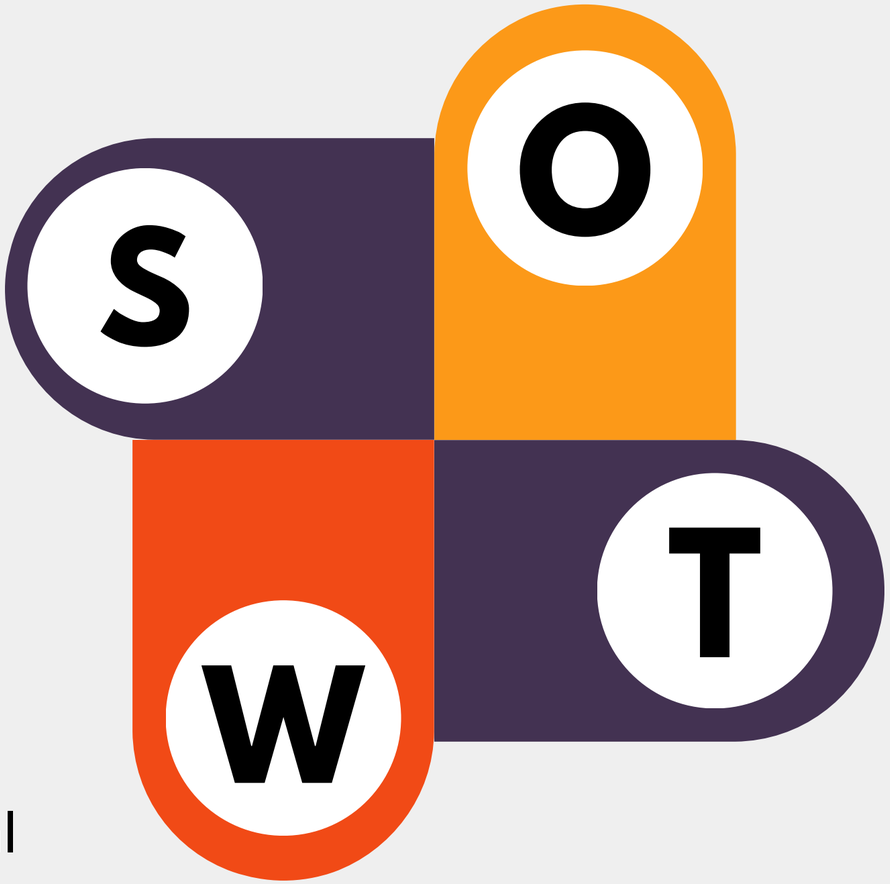


Strengths

Thorough patient data with multiple sources and types of data

Weaknesses

Limited view on behavioral reports due to binary options



Opportunities

Good guidance/reference for further research questions

Threats

Confusing causation and correlation when looking at data we have pre-existing assumptions about

Thank you

Do you have any questions?