CHRONIC KIDNEY DISEASE SCREENING

IDS 506 | HEALTH INFO MANAGEMENT

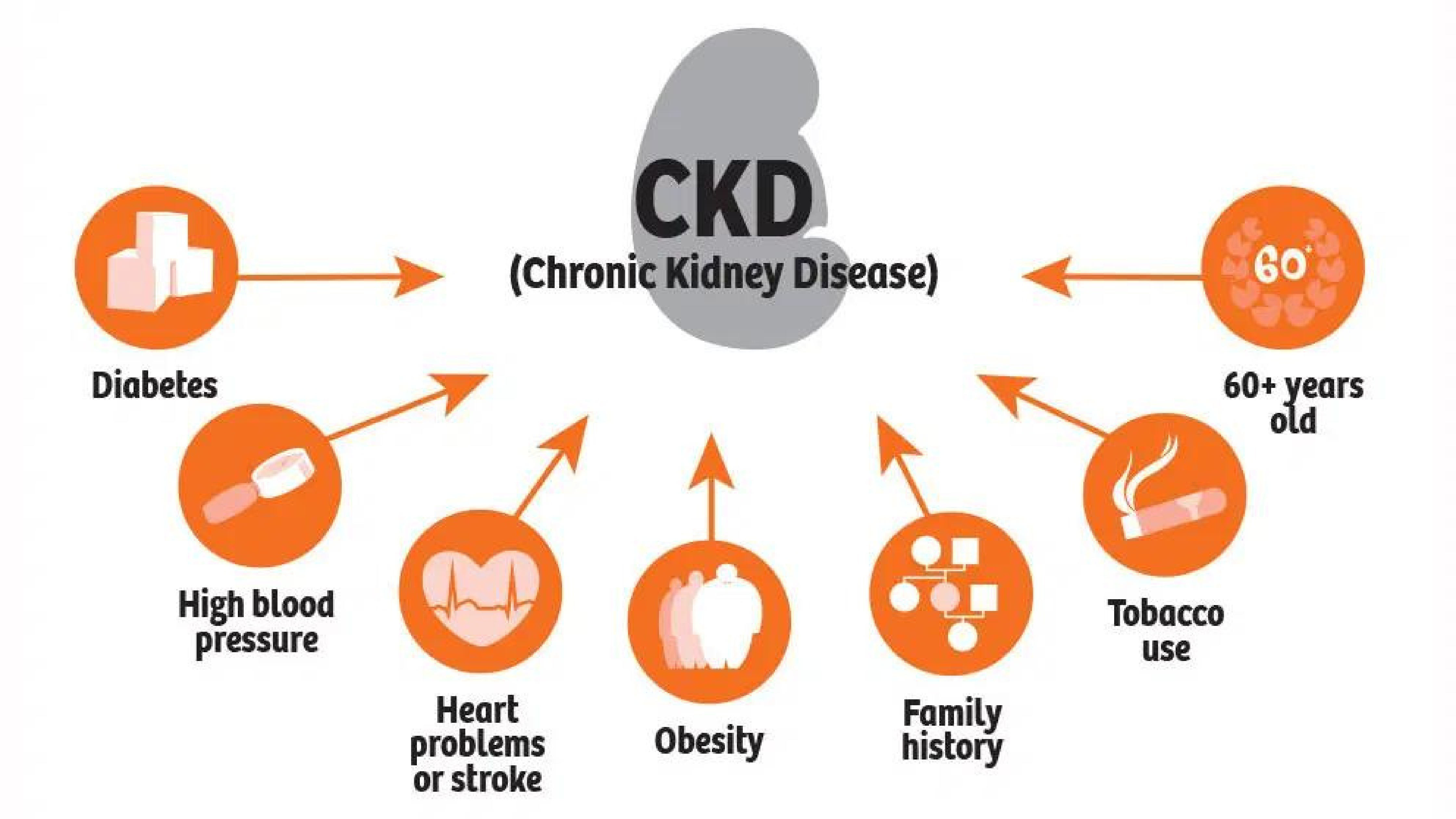
Presented By Anvesh Nadipelli

FEB 28, 2023

Feature Selection

We need to understand what are some of the different factors that causes CKD

Target Variable - CKD Binary



Feature Selection

To create a model that screens whether people will have a risk of Chronic Kidyney Diasease or not , we'll need to consider several relevant attributes. Here are a few relvant attributes from the dataset

Demographics

AGE - Ordinal Grouping
GENDER - Binary Variable
RACE - One Hot Encoding
Fam CVD- Binary
Fam Diabetes - Binary
Fam Hypertension - Binary

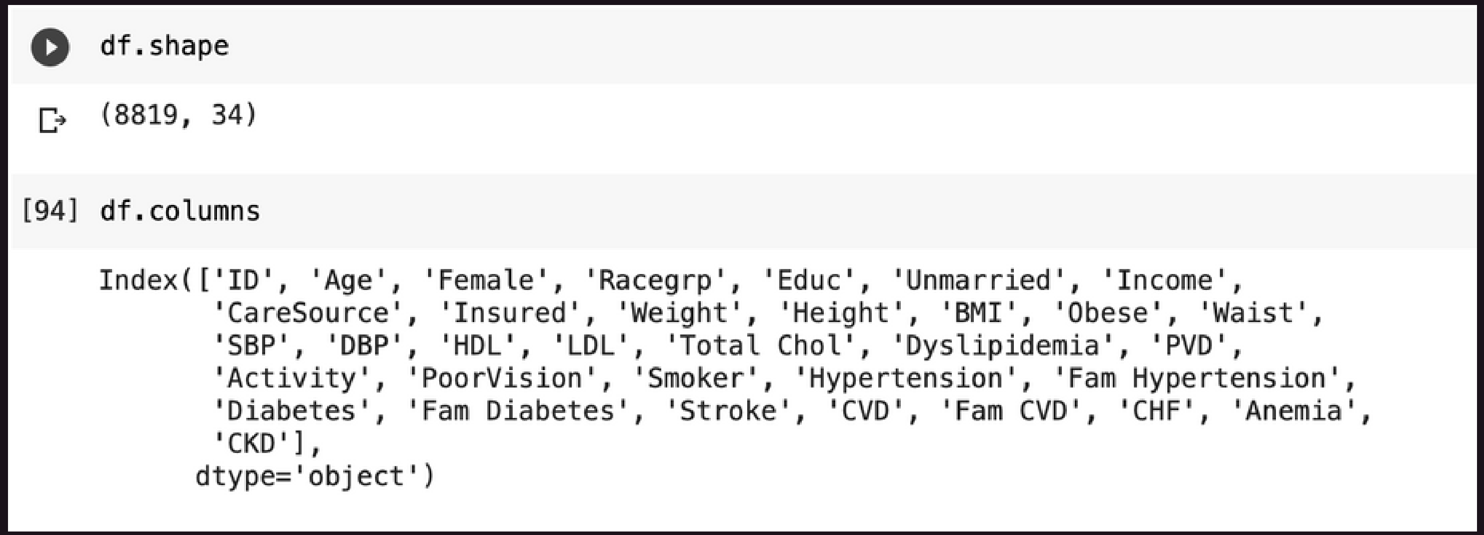
Lifestyle Factors

BMI- Ordinal Grouping
Activity - Ordinal Grouping
SMOKER -Binary Variable

Medical History

Total Cholesterol - Ordinal
Poor Vision
Hypertension
Stroke
CHF
Diabetes
PVD
Rest are Binary variables

Data Set QA



```
df.shape
```

```
(8819, 34)
```

```
[94] df.columns
```

```
Index(['ID', 'Age', 'Female', 'Racegrp', 'Educ', 'Unmarried', 'Income',
       'CareSource', 'Insured', 'Weight', 'Height', 'BMI', 'Obese', 'Waist',
       'SBP', 'DBP', 'HDL', 'LDL', 'Total Chol', 'Dyslipidemia', 'PVD',
       'Activity', 'PoorVision', 'Smoker', 'Hypertension', 'Fam Hypertension',
       'Diabetes', 'Fam Diabetes', 'Stroke', 'CVD', 'Fam CVD', 'CHF', 'Anemia',
       'CKD'],
      dtype='object')
```

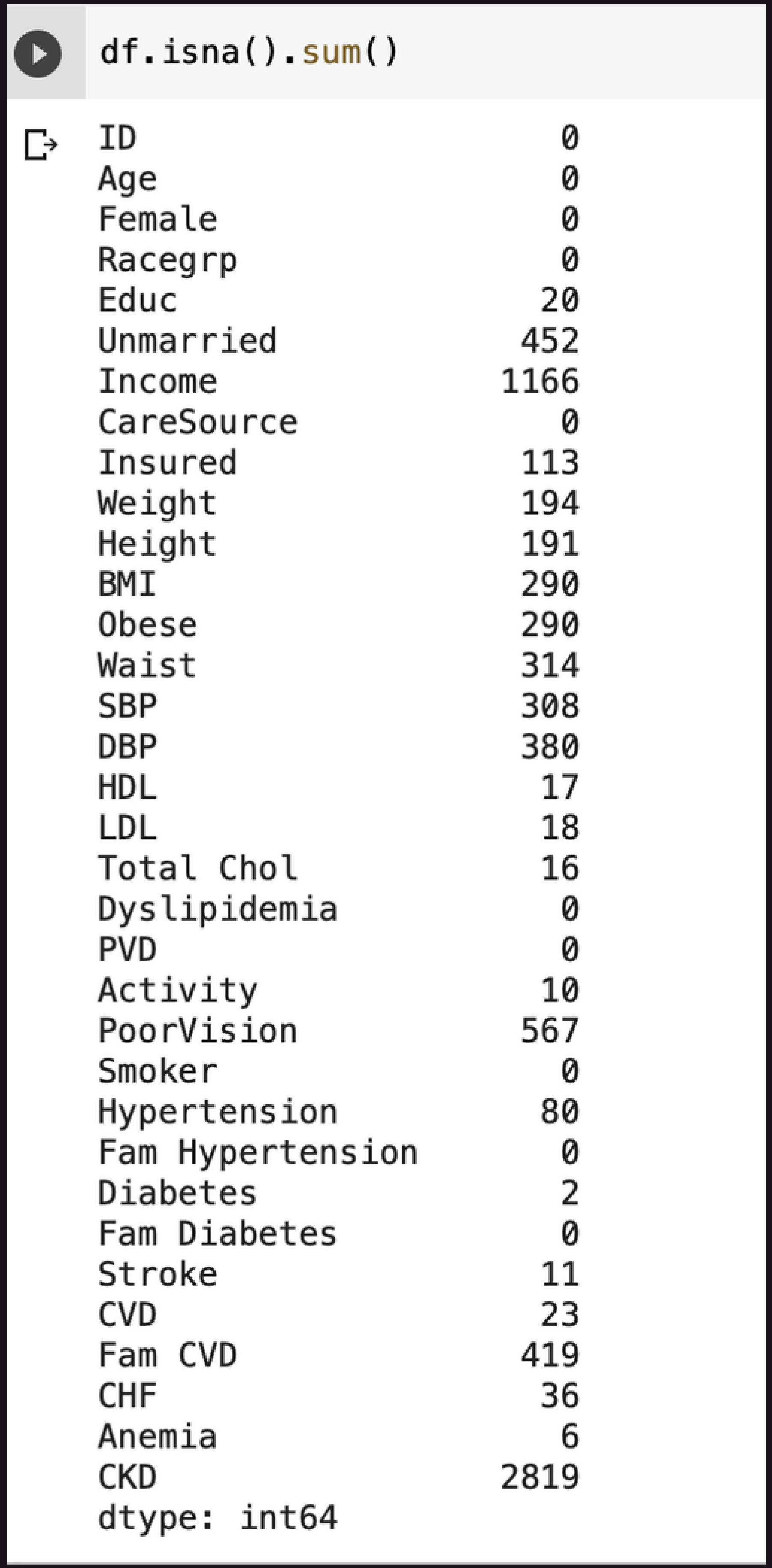
We removed all the irrelevant attributes. We have many null values or blanks in the dataset.

Once the null values are modified, Feature Engineering is done.

Age, BMI, and Total Cholesterol were continuous variables converted to ordinal variables.

- AGE: [18-34]: 1, [35-49]: 2, [50- 64]: 3, [65-74]: 4, [75+]: 5
- BMI [<24.9]: 0, [25-29.9]: 1, [>30]: 2
    - An underweight flag is created for BMI < 18.5
- Cholesterol [<200]: 0, [200-240]: 1, [>240: 2]

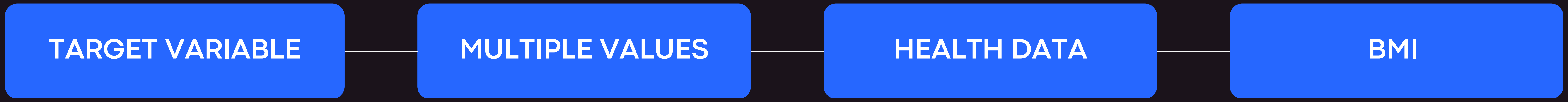
Racegrp - One Hot Encoded according to the different races.



```
df.isna().sum()
```

```
ID                    0
Age                   0
Female                0
Racegrp               0
Educ                 20
Unmarried           452
Income             1166
CareSource            0
Insured             113
Weight              194
Height              191
BMI                 290
Obese               290
Waist               314
SBP                 308
DBP                 380
HDL                  17
LDL                  18
Total Chol           16
Dyslipidemia          0
PVD                   0
Activity             10
PoorVision          567
Smoker                0
Hypertension         80
Fam Hypertension      0
Diabetes              2
Fam Diabetes          0
Stroke               11
CVD                  23
Fam CVD             419
CHF                  36
Anemia                6
CKD                2819
dtype: int64
```

Dealing with Null Values



TARGET VARIABLE

Removed all the rows where our target varibles were null.

MULTIPLE VALUES

Then removed rows where multiple columns have missing values

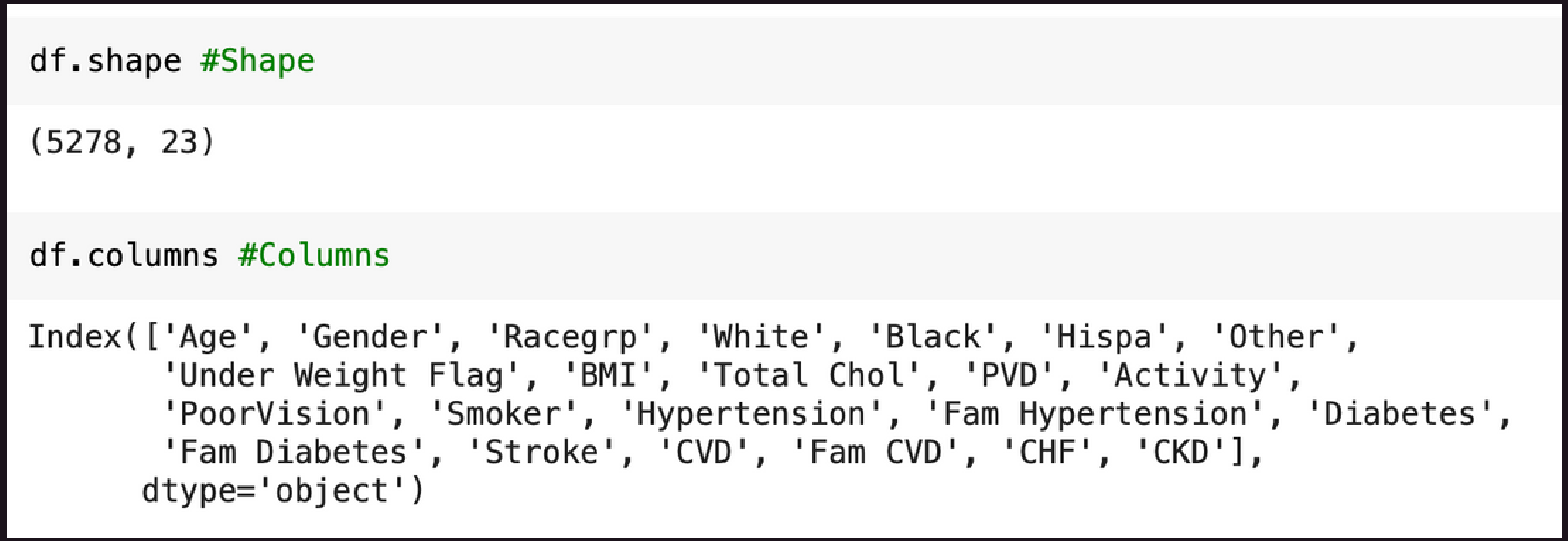
HEALTH DATA

Removed all the rows where health data was null

BMI

Imputed Average BMI based on gender for CKD label 1 Data

Final Data Set | After QA



```
df.shape #Shape
```

```
(5278, 23)
```

```
df.columns #Columns
```

```
Index(['Age', 'Gender', 'Racegrp', 'White', 'Black', 'Hispa', 'Other',
       'Under Weight Flag', 'BMI', 'Total Chol', 'PVD', 'Activity',
       'PoorVision', 'Smoker', 'Hypertension', 'Fam Hypertension', 'Diabetes',
       'Fam Diabetes', 'Stroke', 'CVD', 'Fam CVD', 'CHF', 'CKD'],
      dtype='object')
```

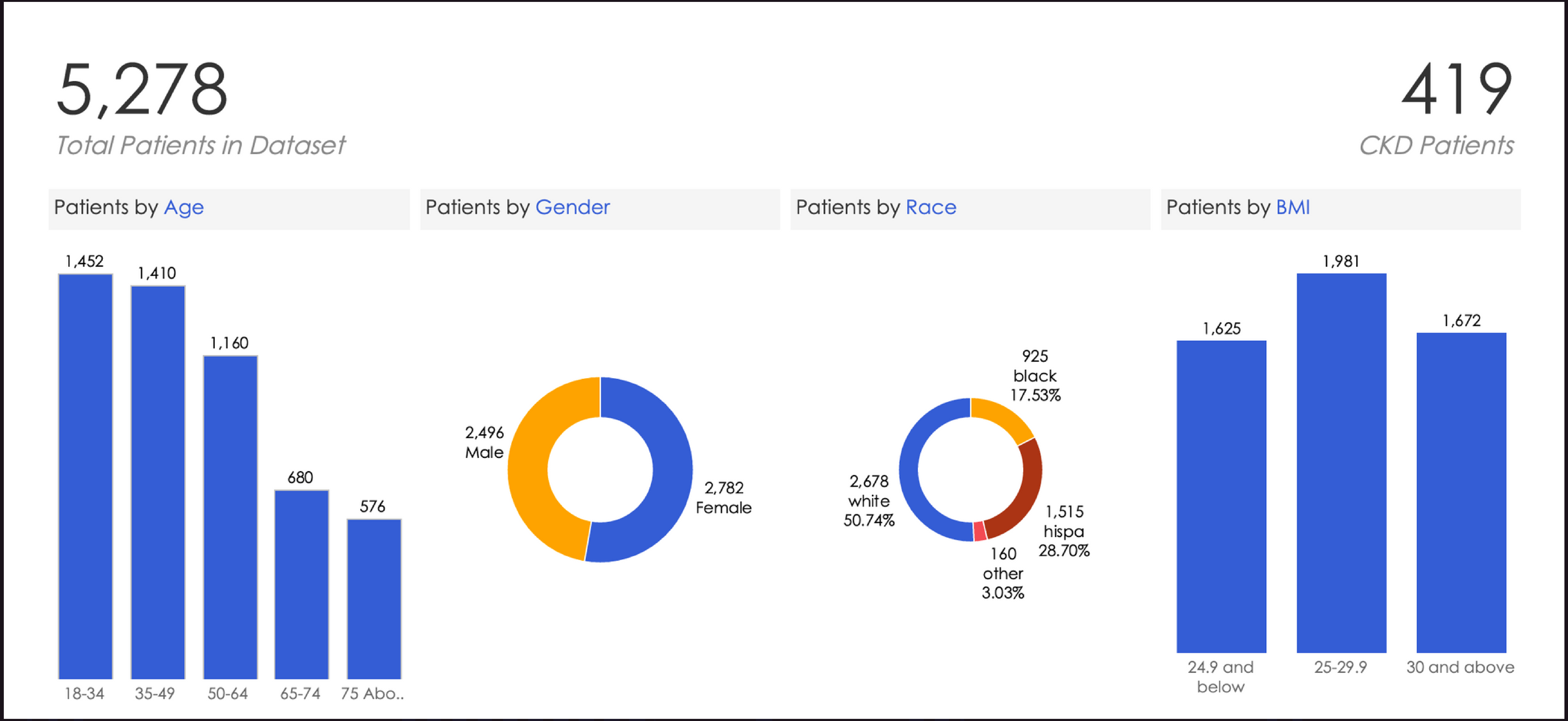
Why Poor Vision is considered:
Sudden Visual Deterioration as the First Symptom of Chronic Kidney Failure [1*]

Attributes that are not considered from the dataset

- Educ, Unmarried, Income, CareSource, and Insured are irrelevant whether to predict CKD or not.
- Attributes [Weight, Height, and waist] are correlated with BMI, which is already considered.
- The obese attribute was a flag for BMI greater than 30. BMI ordinal considers all the BMI groups instead of just looking for people with a BMI greater than 30.
- SBP and DBP are irrelevant. The hypertension parameter is already present.
- Total Chol is the sum of HDL and LDL. HDL and LDL are not considered. Dyslipidemia is a flag for Total Chol > 240.
- Anemia is irrelevant for screening. People with CKD have higher chances of getting anemia but not other way around. [2*]

Uni Variate Analysis

Demographics Data



5,278

Total Patients in Dataset

419

CKD Patients

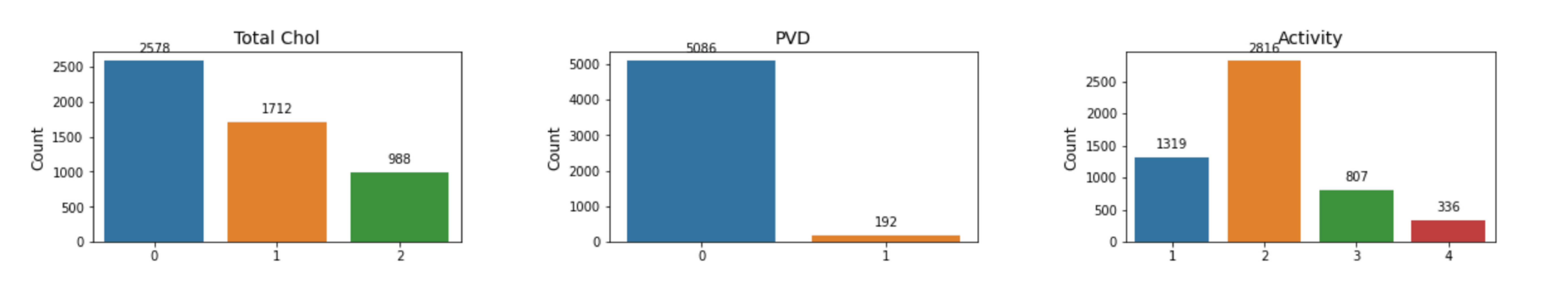
Patients by Age

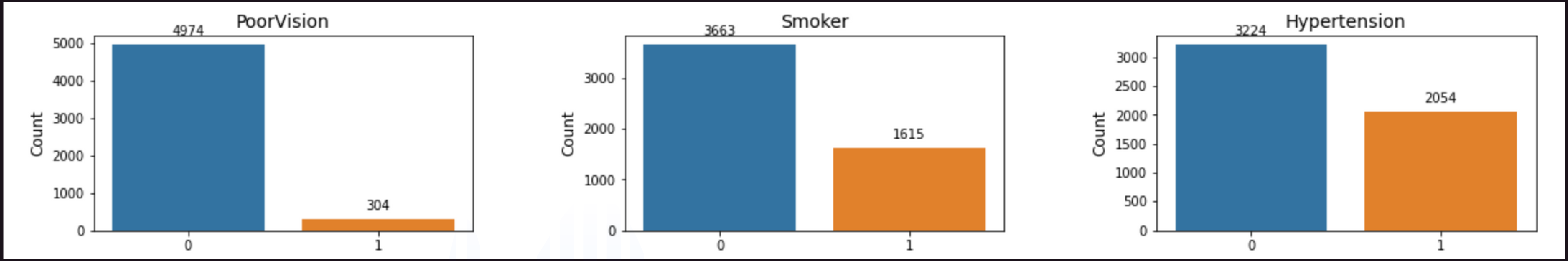
Patients by Gender

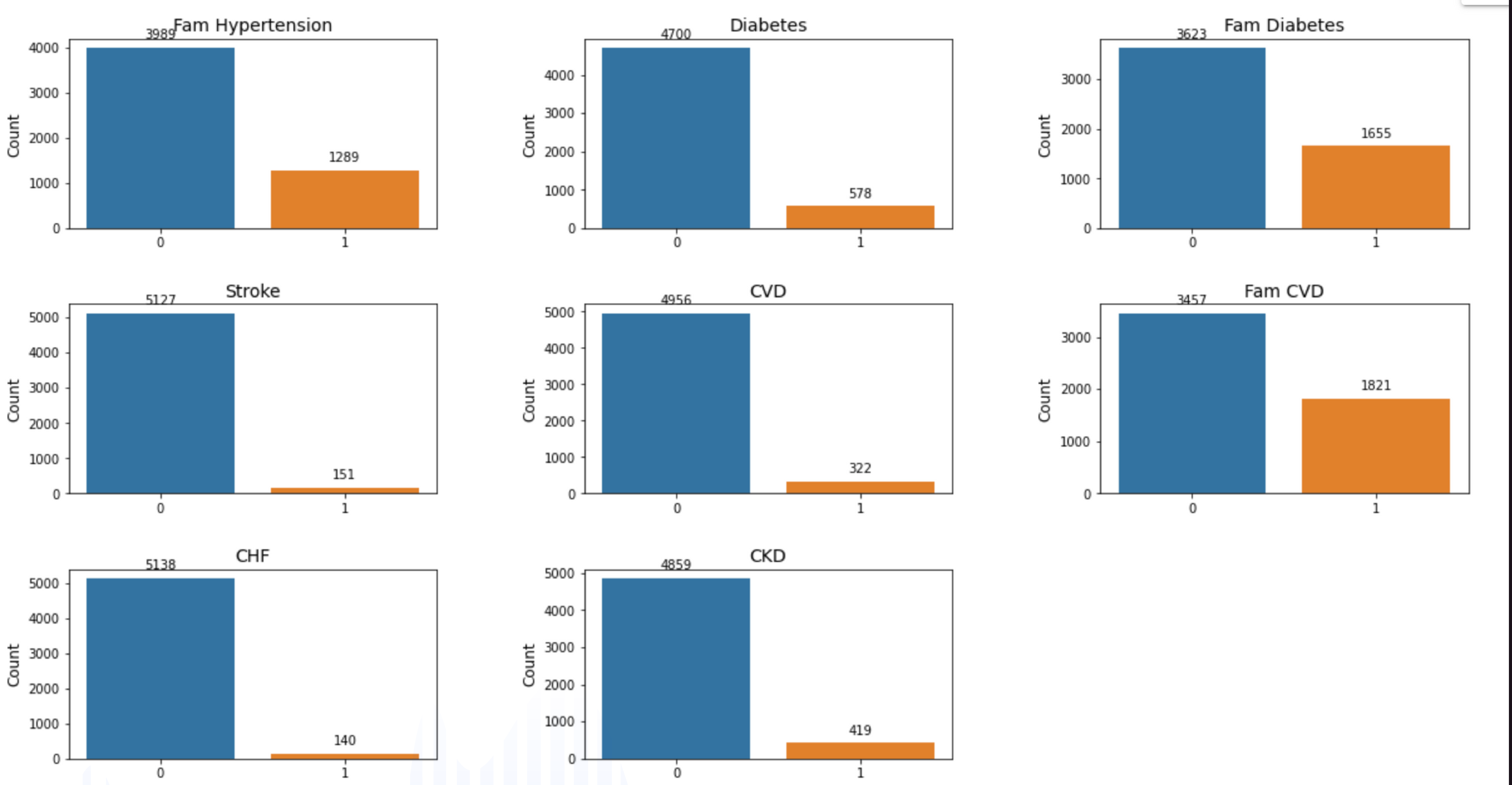
Patients by Race

Patients by BMI

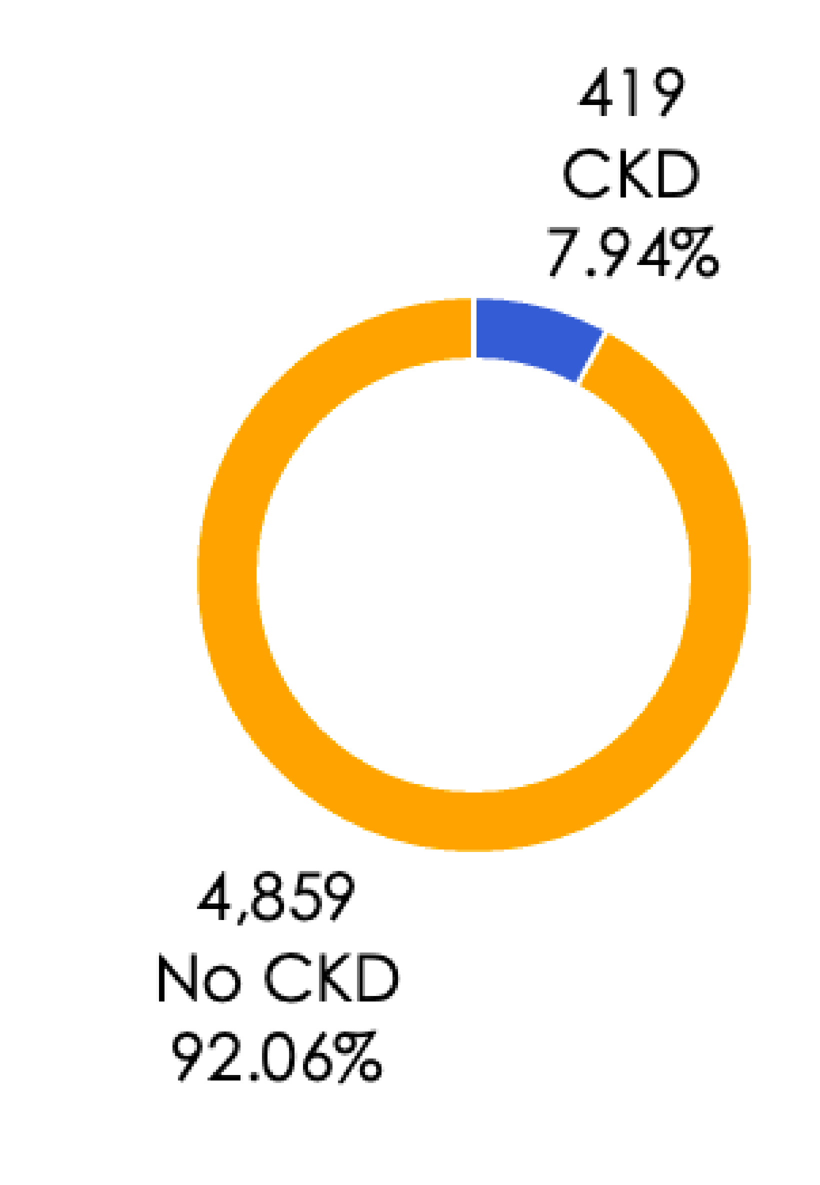
Count Plots -Health



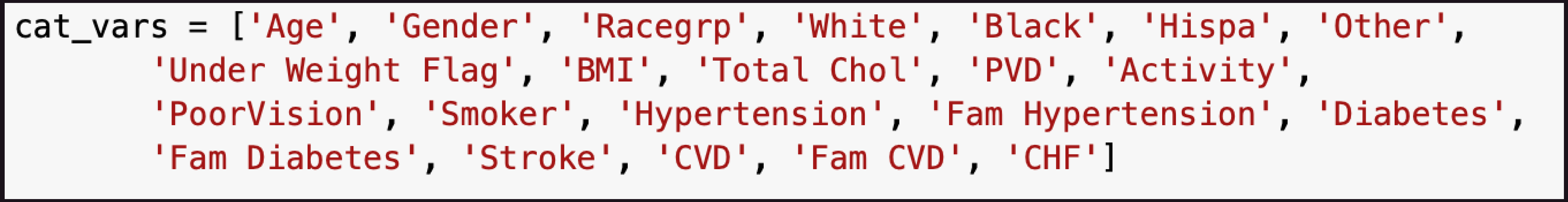



Fam Hypertension
3989
1289
Diabetes
4700
578
Fam Diabetes
3623
1655
Stroke
5127
151
CVD
4956
322
Fam CVD
3457
1821
CHF
5138
140
CKD
4859
419
Count
0
1
0
1000
2000
3000
4000
5000

TargetVariable - Label Imbalance

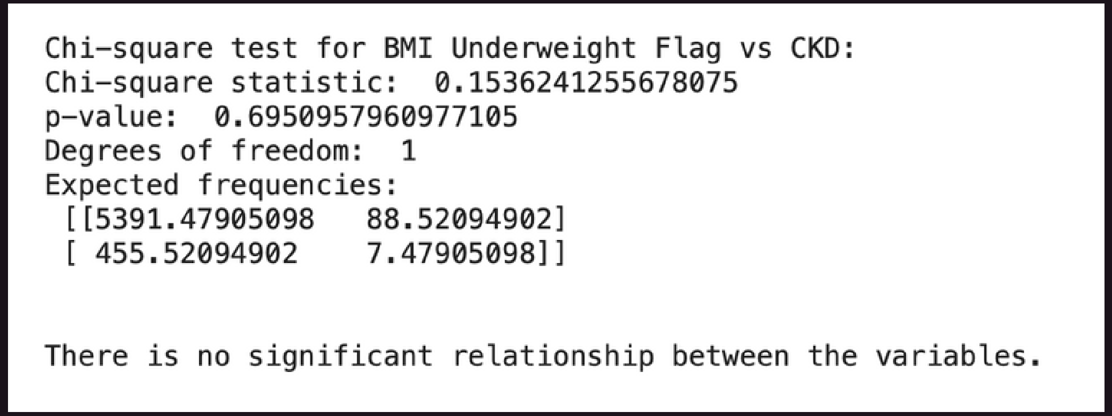


Bi Variate Analysis

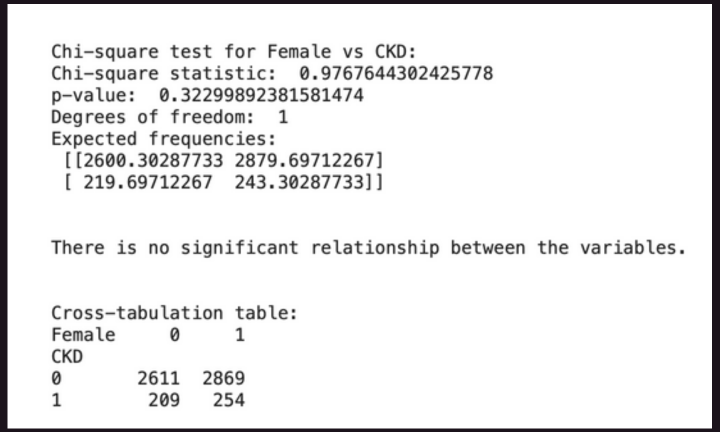


```
cat_vars = ['Age', 'Gender', 'Racegrp', 'White', 'Black', 'Hispa', 'Other',
        'Under Weight Flag', 'BMI', 'Total Chol', 'PVD', 'Activity',
        'PoorVision', 'Smoker', 'Hypertension', 'Fam Hypertension', 'Diabetes',
        'Fam Diabetes', 'Stroke', 'CVD', 'Fam CVD', 'CHF']
```

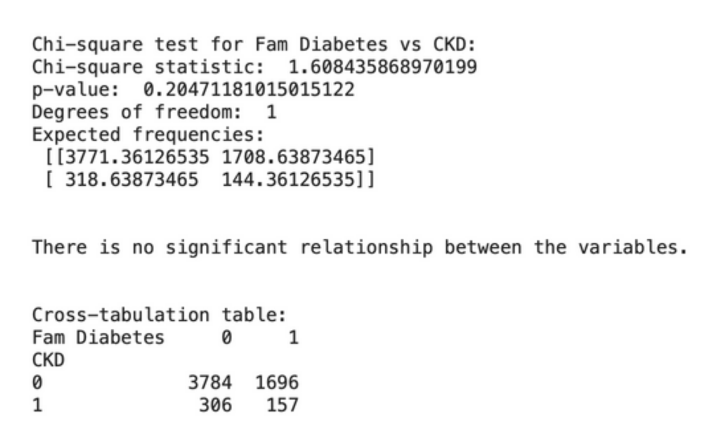
Variables for BI Variate Analysis - All of them are categorical. Performed CHI Square tests to determine any significant relationship with the target variable.



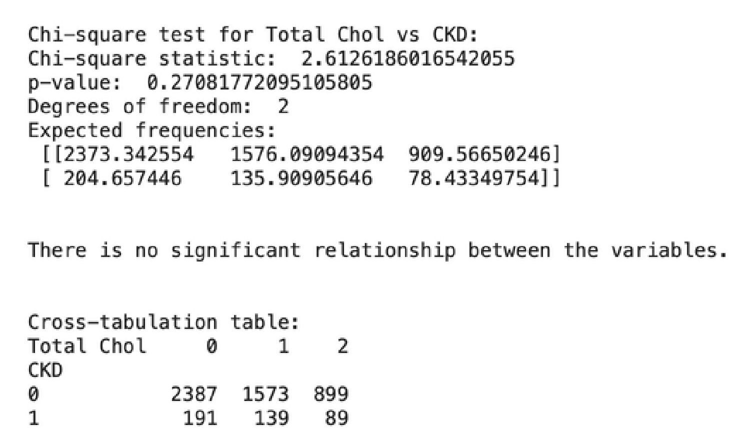
```
Chi-square test for BMI Underweight Flag vs CKD:
Chi-square statistic:  0.1536241255678075
p-value:  0.6950957960977105
Degrees of freedom:  1
Expected frequencies:
 [[5391.47905098   88.52094902]
 [ 455.52094902    7.47905098]]


There is no significant relationship between the variables.
```



```
Chi-square test for Female vs CKD:
Chi-square statistic:  0.9767644302425778
p-value:  0.32299892381581474
Degrees of freedom:  1
Expected frequencies:
 [[2600.30287733 2879.69712267]
 [ 219.69712267  243.30287733]]


There is no significant relationship between the variables.


Cross-tabulation table:
Female     0     1
CKD
0       2611  2869
1        209   254
```



```
Chi-square test for Fam Diabetes vs CKD:
Chi-square statistic:  1.608435868970199
p-value:  0.20471181015015122
Degrees of freedom:  1
Expected frequencies:
 [[3771.36126535 1708.63873465]
 [ 318.63873465  144.36126535]]


There is no significant relationship between the variables.


Cross-tabulation table:
Fam Diabetes     0     1
CKD
0             3784  1696
1              306   157
```



```
Chi-square test for Total Chol vs CKD:
Chi-square statistic:  2.6126186016542055
p-value:  0.27081772095105805
Degrees of freedom:  2
Expected frequencies:
 [[2373.342554   1576.09094354  909.56650246]
 [ 204.657446    135.90905646   78.43349754]]


There is no significant relationship between the variables.


Cross-tabulation table:
Total Chol     0     1    2
CKD
0           2387  1573  899
1            191   139   89
```

These are four variable which had no significant relationship with the our target variable. Gender, Underweight Flag, Family Diabetic Histroy and Total Chol. Removed these variables from our final model.

2,054
Hypertension Patients

151
Stroke Patients

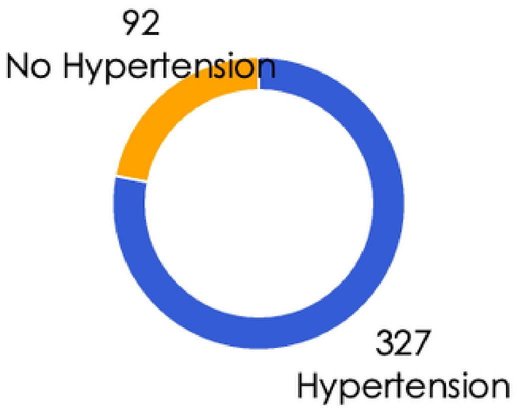
322
CVD Patients

140
CHF Patients

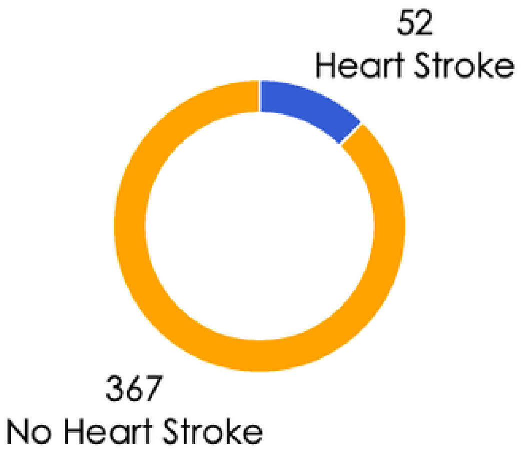
192
PVD Patients

578
Diabetes Patients

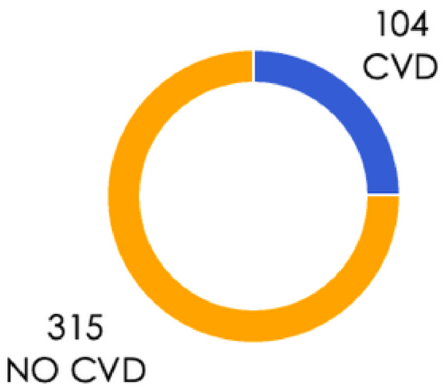
CKD and Hypertension



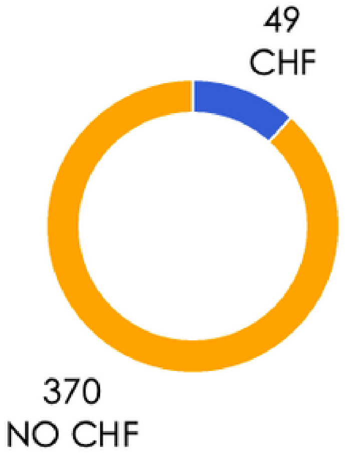
CKD and Stroke



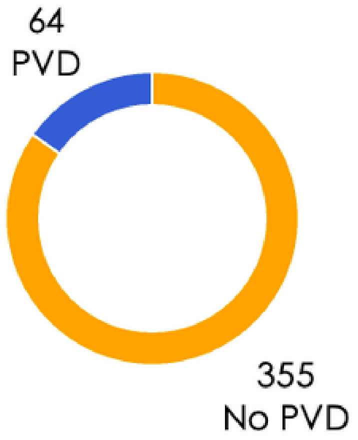
CKD and CVD



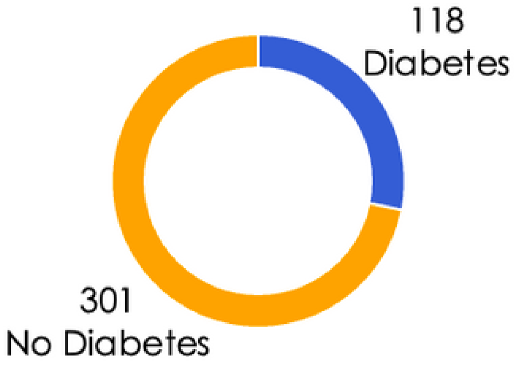
CKD and CHF



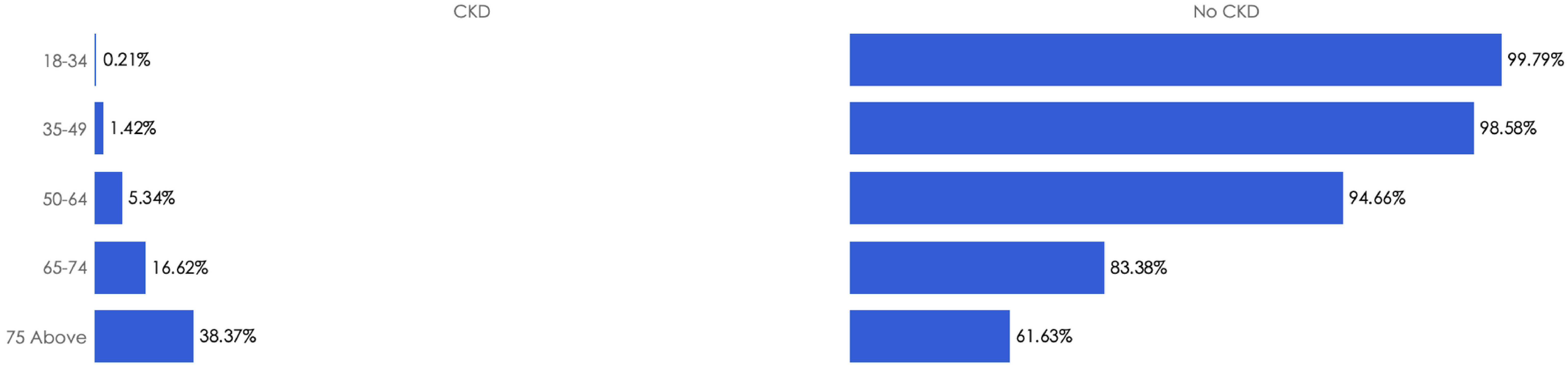
CKD and PVD



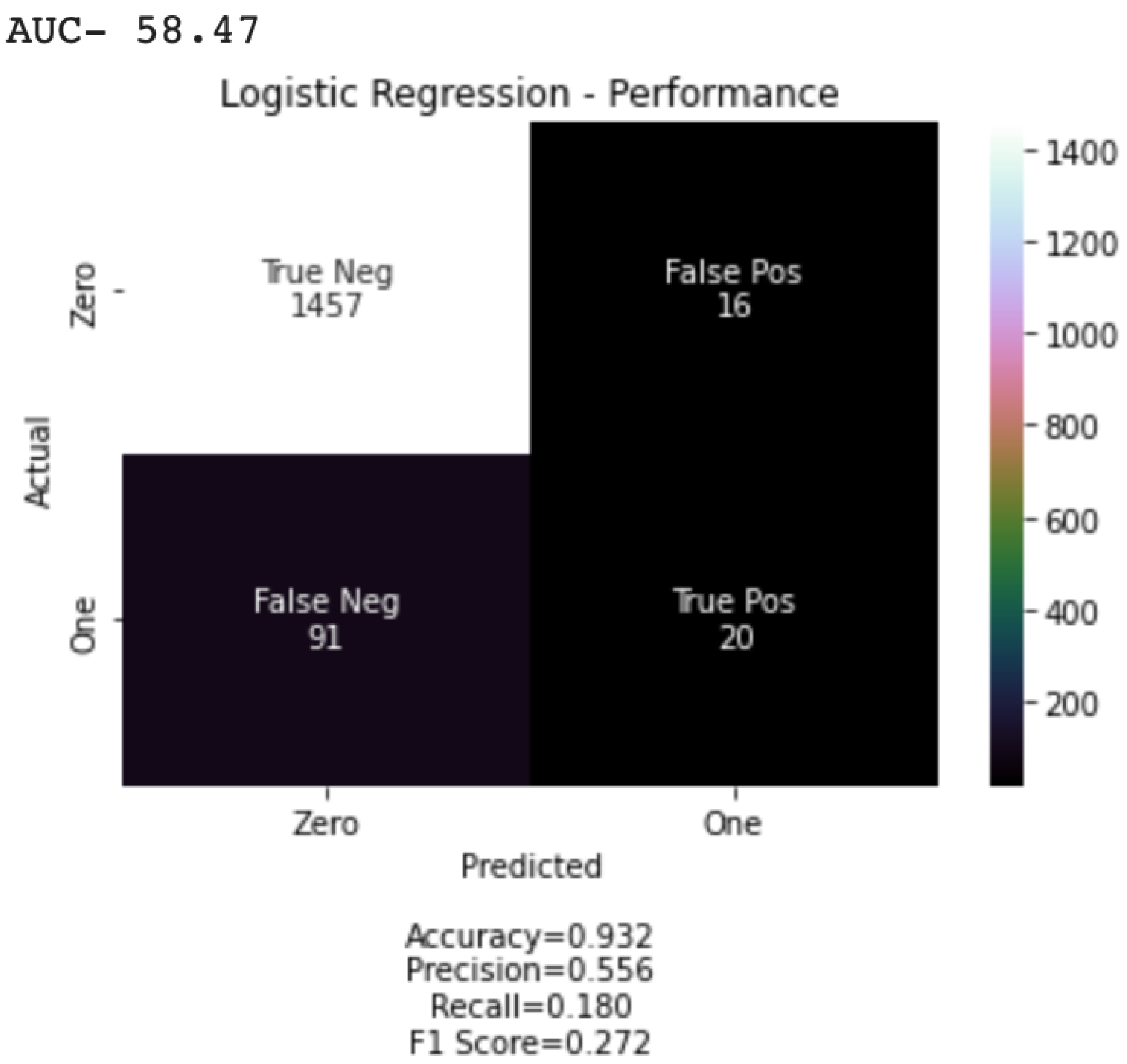
CKD and Diabetes



CKD and Age

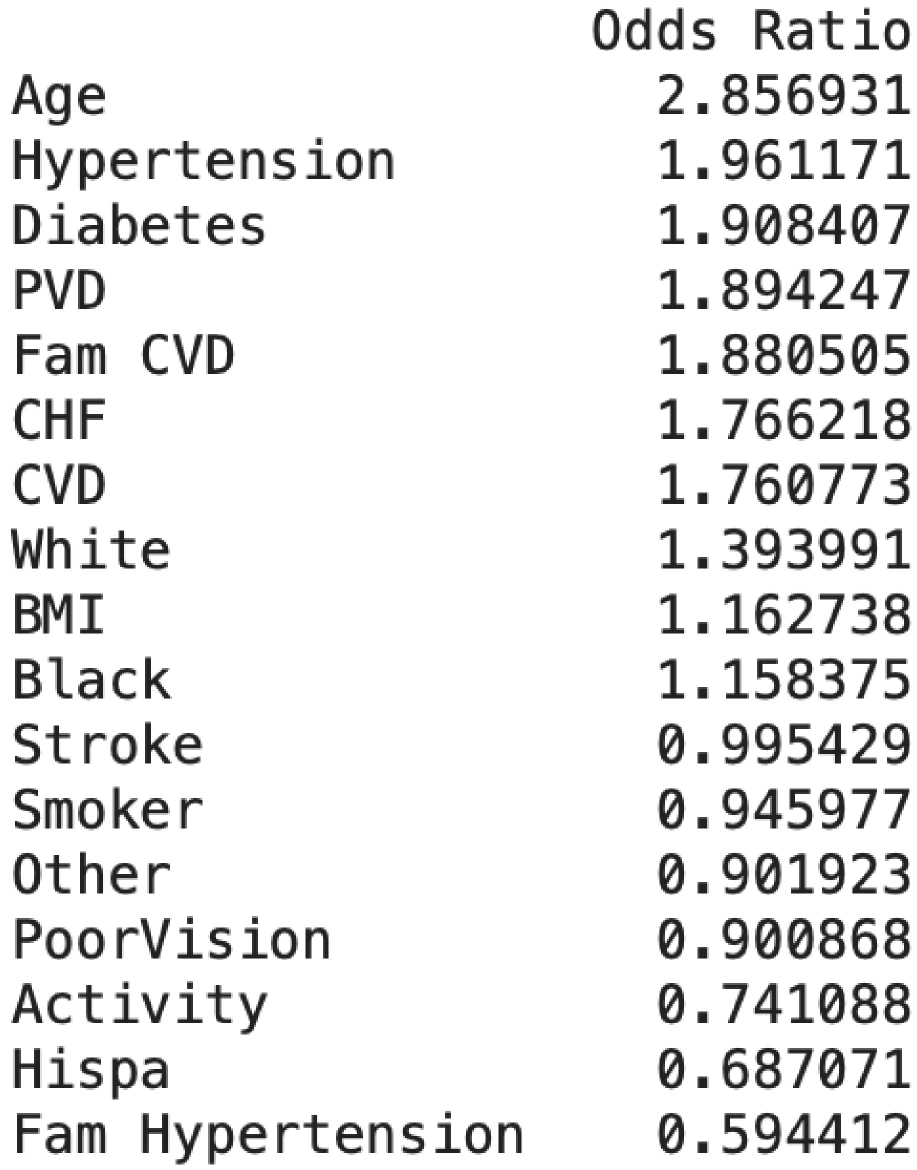


Model Built Logistic Regression



```
AUC- 58.47
```

Accuracy=0.932
Precision=0.556
Recall=0.180
F1 Score=0.272

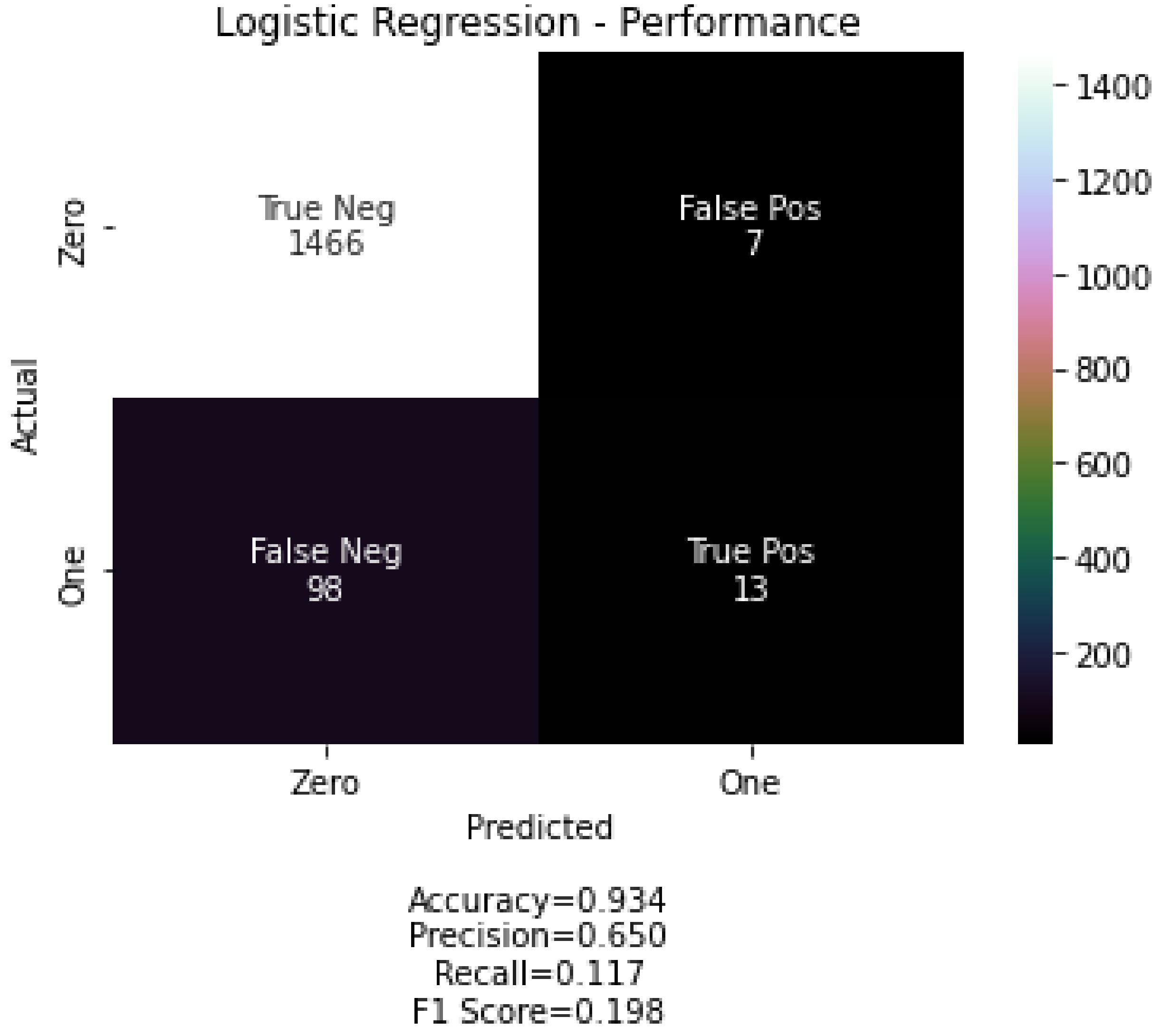


```
                  Odds Ratio
Age                 2.856931
Hypertension        1.961171
Diabetes            1.908407
PVD                 1.894247
Fam CVD             1.880505
CHF                 1.766218
CVD                 1.760773
White               1.393991
BMI                 1.162738
Black               1.158375
Stroke              0.995429
Smoker              0.945977
Other               0.901923
PoorVision          0.900868
Activity            0.741088
Hispa               0.687071
Fam Hypertension    0.594412
```

- Low AUC for ROC was expected as Data was Highly Imbalanced.
- Logistic Regression gave importance to the attributes found to be significant with CKD during EDA.
- The negative odds ratio for attributes Smoker, Stroke, PoorVision, and Fam Hypertension - Weird Result.

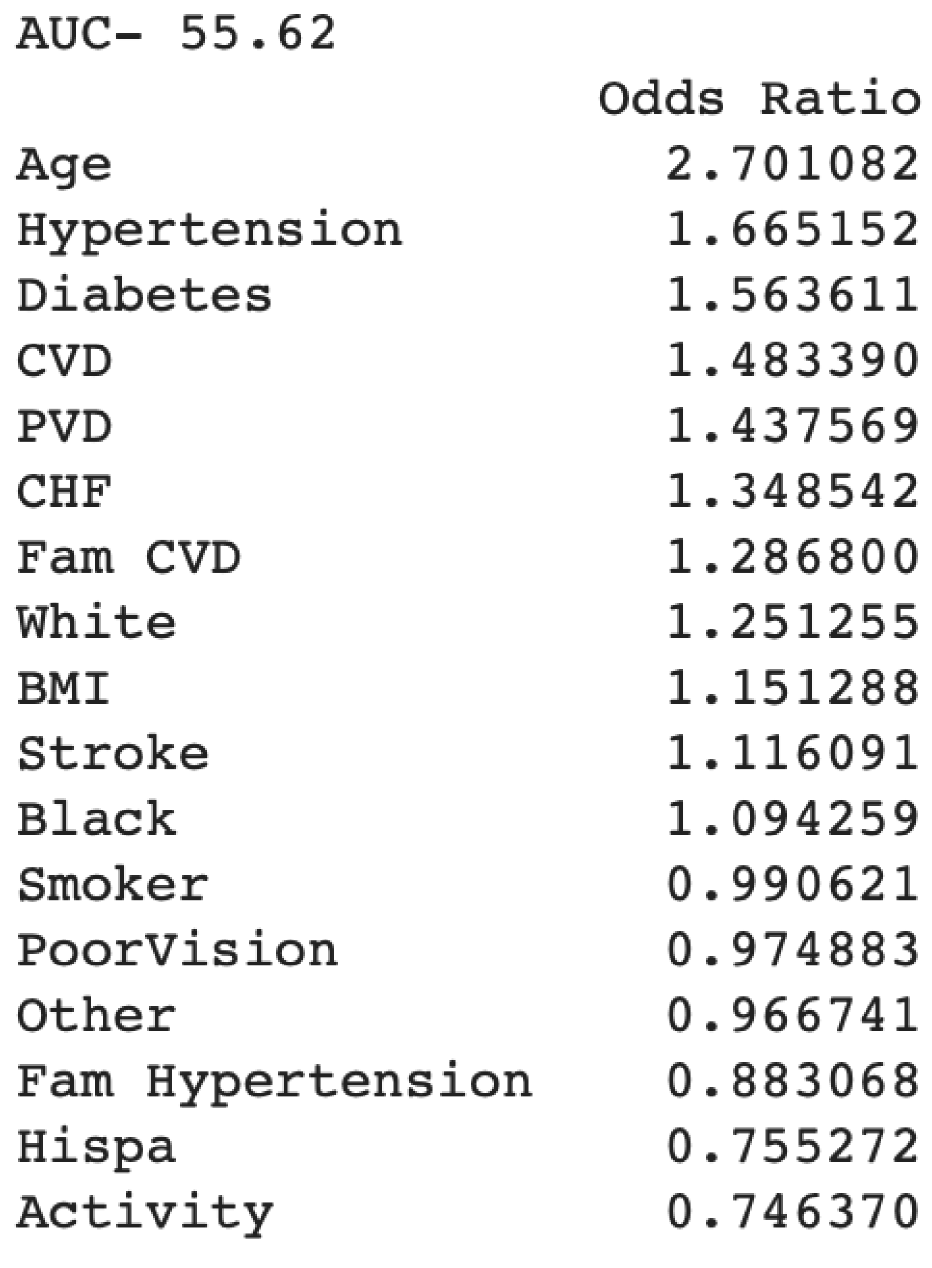
Logistic Regression | 10 Cross Validation

AUC- 55.2



Logistic Regression - Performance

Accuracy=0.934
Precision=0.650
Recall=0.117
F1 Score=0.198



```
AUC- 55.62
                      Odds Ratio
Age                     2.701082
Hypertension            1.665152
Diabetes                1.563611
CVD                     1.483390
PVD                     1.437569
CHF                     1.348542
Fam CVD                 1.286800
White                   1.251255
BMI                     1.151288
Stroke                  1.116091
Black                   1.094259
Smoker                  0.990621
PoorVision              0.974883
Other                   0.966741
Fam Hypertension        0.883068
Hispa                   0.755272
Activity                0.746370
```

- **Similar Odds ratio to our original model**
- **Decrease in AUC under ROC with 10 Fold Cross Validation**

References

1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4777964/
2. https://www.niddk.nih.gov/health-information/kidney-disease/anemia