CHRONIC KIDNEY DISEASE SCREENING

IDS 506 | HEALTH INFO MANAGEMENT

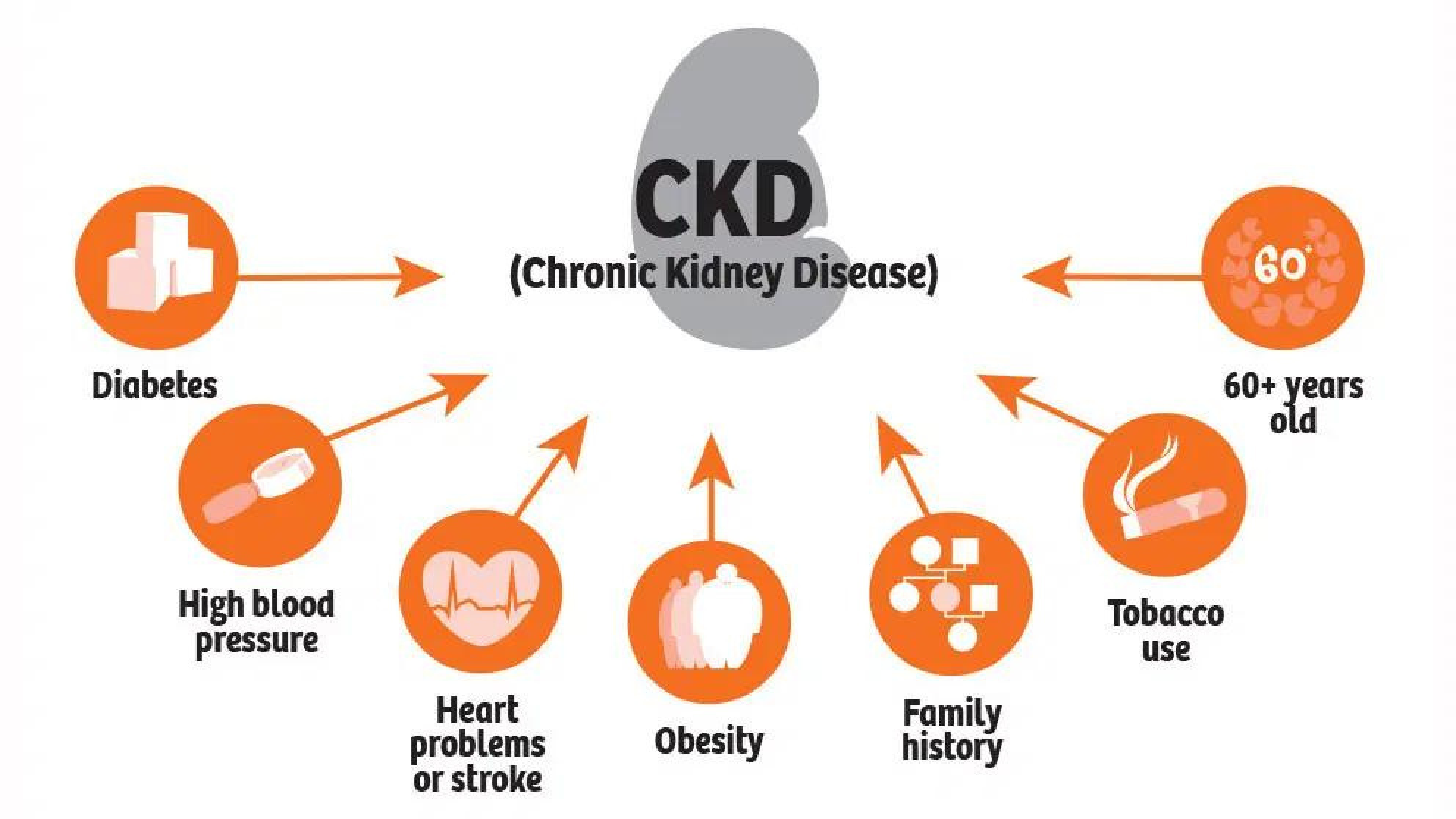
Presented By Anvesh Nadipelli

FEB 28, 2023

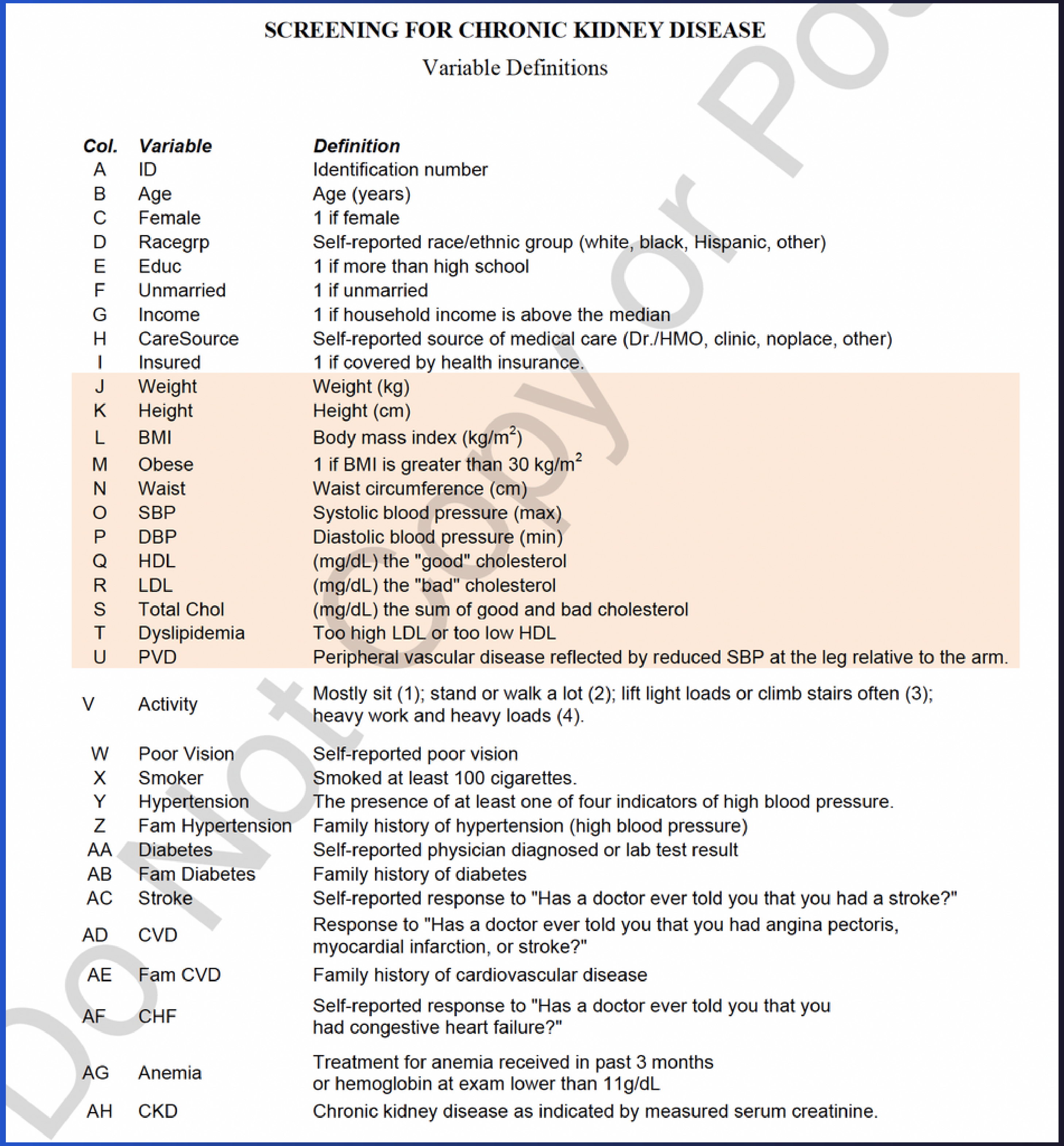
Feature Selection

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

Target Variable - CKD Binary



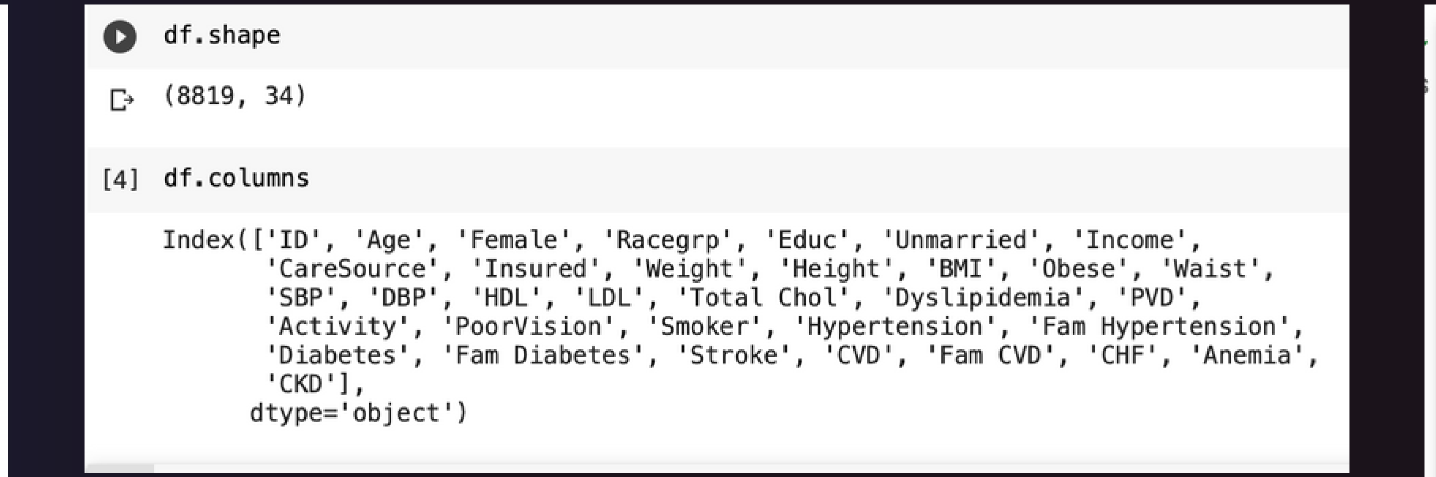
Initial QA



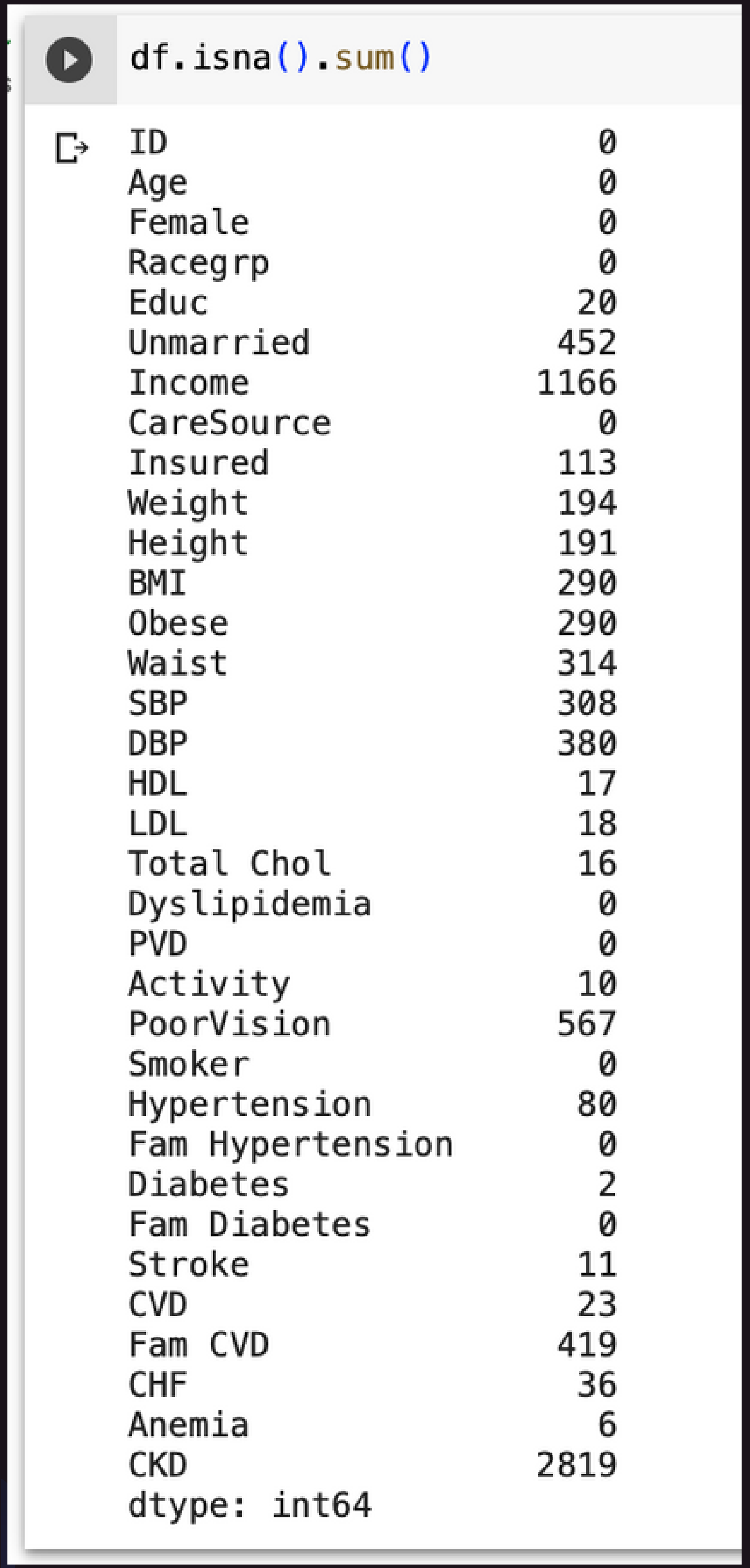
SCREENING FOR CHRONIC KIDNEY DISEASE

Variable Definitions

Col.	*Variable*	*Definition*
A	ID	Identification number
B	Age	Age (years)
C	Female	1 if female
D	Racegrp	Self-reported race/ethnic group (white, black, Hispanic, other)
E	Educ	1 if more than high school
F	Unmarried	1 if unmarried
G	Income	1 if household income is above the median
H	CareSource	Self-reported source of medical care (Dr./HMO, clinic, noplace, other)
I	Insured	1 if covered by health insurance.
J	Weight	Weight (kg)
K	Height	Height (cm)
L	BMI	Body mass index (kg/m^2)
M	Obese	1 if BMI is greater than 30 kg/m^2
N	Waist	Waist circumference (cm)
O	SBP	Systolic blood pressure (max)
P	DBP	Diastolic blood pressure (min)
Q	HDL	(mg/dL) the "good" cholesterol
R	LDL	(mg/dL) the "bad" cholesterol
S	Total Chol	(mg/dL) the sum of good and bad cholesterol
T	Dyslipidemia	Too high LDL or too low HDL
U	PVD	Peripheral vascular disease reflected by reduced SBP at the leg relative to the arm.
V	Activity	Mostly sit (1); stand or walk a lot (2); lift light loads or climb stairs often (3); heavy work and heavy loads (4).
W	Poor Vision	Self-reported poor vision
X	Smoker	Smoked at least 100 cigarettes.
Y	Hypertension	The presence of at least one of four indicators of high blood pressure.
Z	Fam Hypertension	Family history of hypertension (high blood pressure)
AA	Diabetes	Self-reported physician diagnosed or lab test result
AB	Fam Diabetes	Family history of diabetes
AC	Stroke	Self-reported response to "Has a doctor ever told you that you had a stroke?"
AD	CVD	Response to "Has a doctor ever told you that you had angina pectoris, myocardial infarction, or stroke?"
AE	Fam CVD	Family history of cardiovascular disease
AF	CHF	Self-reported response to "Has a doctor ever told you that you had congestive heart failure?"
AG	Anemia	Treatment for anemia received in past 3 months or hemoglobin at exam lower than 11g/dL
AH	CKD	Chronic kidney disease as indicated by measured serum creatinine.

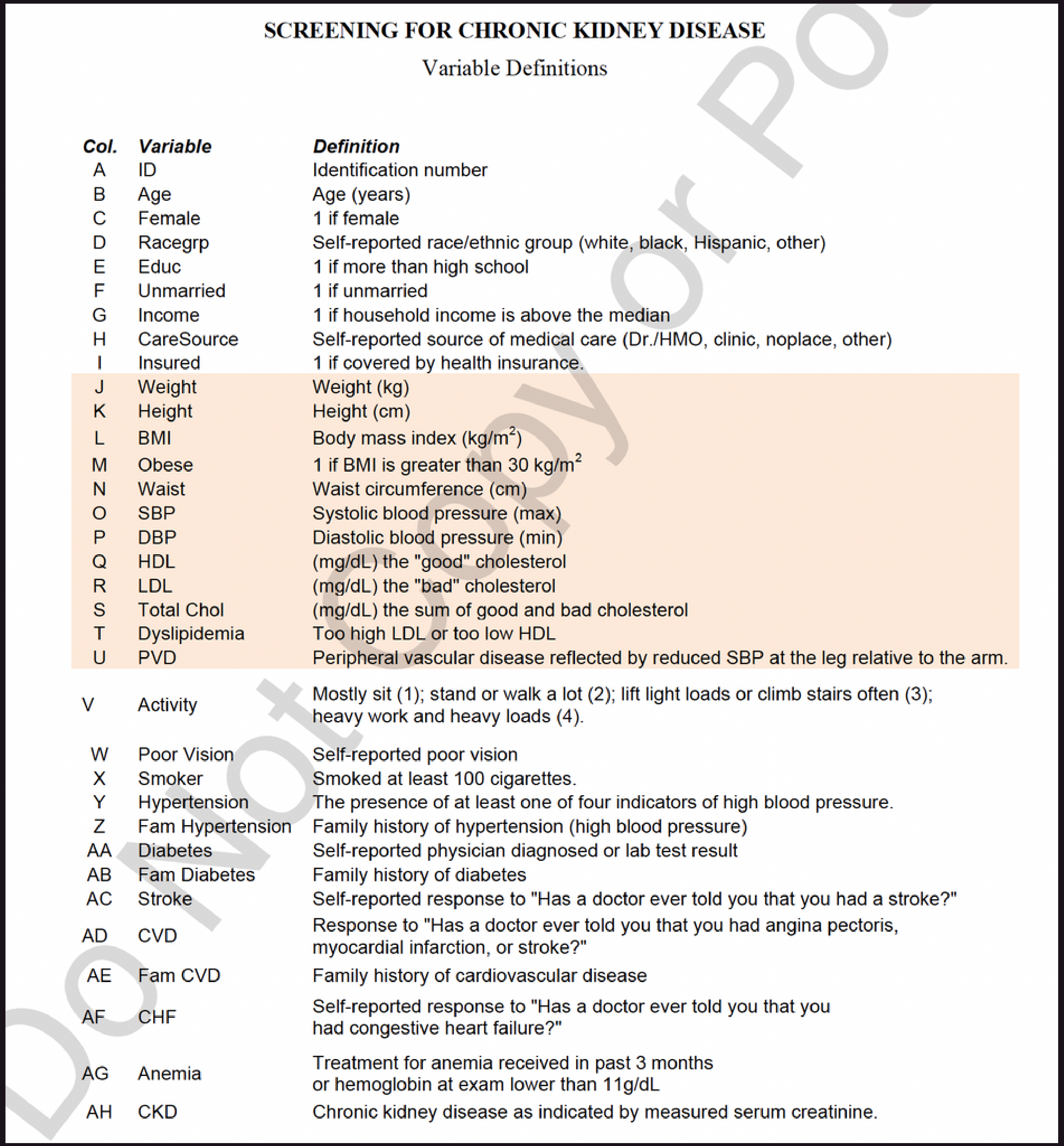


```
df.shape

(8819, 34)

[4] 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')
```



```
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
```

Considerations for Final Data Set



SCREENING FOR CHRONIC KIDNEY DISEASE

Variable Definitions

Col.	*Variable*	*Definition*
A	ID	Identification number
B	Age	Age (years)
C	Female	1 if female
D	Racegrp	Self-reported race/ethnic group (white, black, Hispanic, other)
E	Educ	1 if more than high school
F	Unmarried	1 if unmarried
G	Income	1 if household income is above the median
H	CareSource	Self-reported source of medical care (Dr./HMO, clinic, noplace, other)
I	Insured	1 if covered by health insurance.
J	Weight	Weight (kg)
K	Height	Height (cm)
L	BMI	Body mass index (kg/m^2)
M	Obese	1 if BMI is greater than 30 kg/m^2
N	Waist	Waist circumference (cm)
O	SBP	Systolic blood pressure (max)
P	DBP	Diastolic blood pressure (min)
Q	HDL	(mg/dL) the "good" cholesterol
R	LDL	(mg/dL) the "bad" cholesterol
S	Total Chol	(mg/dL) the sum of good and bad cholesterol
T	Dyslipidemia	Too high LDL or too low HDL
U	PVD	Peripheral vascular disease reflected by reduced SBP at the leg relative to the arm.
V	Activity	Mostly sit (1); stand or walk a lot (2); lift light loads or climb stairs often (3); heavy work and heavy loads (4).
W	Poor Vision	Self-reported poor vision
X	Smoker	Smoked at least 100 cigarettes.
Y	Hypertension	The presence of at least one of four indicators of high blood pressure.
Z	Fam Hypertension	Family history of hypertension (high blood pressure)
AA	Diabetes	Self-reported physician diagnosed or lab test result
AB	Fam Diabetes	Family history of diabetes
AC	Stroke	Self-reported response to "Has a doctor ever told you that you had a stroke?"
AD	CVD	Response to "Has a doctor ever told you that you had angina pectoris, myocardial infarction, or stroke?"
AE	Fam CVD	Family history of cardiovascular disease
AF	CHF	Self-reported response to "Has a doctor ever told you that you had congestive heart failure?"
AG	Anemia	Treatment for anemia received in past 3 months or hemoglobin at exam lower than 11g/dL
AH	CKD	Chronic kidney disease as indicated by measured serum creatinine.

Why Poor Vision is considered:

- Sudden Visual Deterioration is 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, HDL, and LDL are not considered.
- Anemia is irrelevant for screening. People with CKD have a higher chance of getting anemia but not vice versa. [2*]

Final Features

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

Binary variables:

- Poor Vision
- Hypertension
- Stroke
- CHF [Chronic Heart Faliure]
- Diabetes
- Dy
- PVD [Peripheral vascular disease]

Dealing with Null Values

TARGET VARIABLE

Removed all the rows where our target variables were null.
Stored them in different files for prediction.

MULTIPLE VALUES

Then removed rows where multiple columns have missing values.

HEALTH DATA

Removed all the rows where health data was null.
I imputed where I could.

Hypertension values are imputed based on Systolic and Diastolic Blood Pressure.
BMI values are imputed using Height and Weight where ever possible.

Feature Engineering

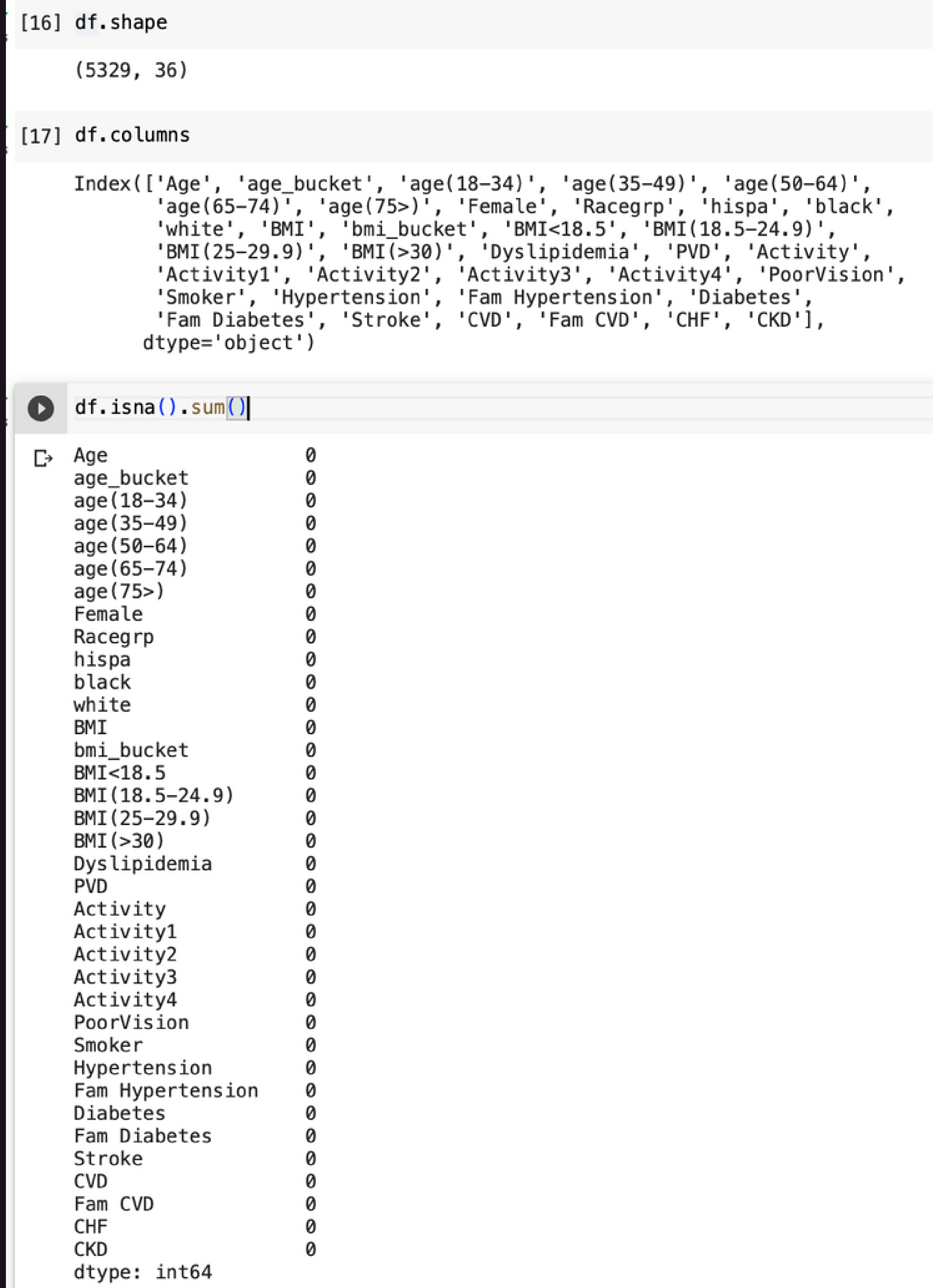
Once the null values are handled, Feature Engineering is done.

Age and BMI were continuous variables converted to ordinal variables.

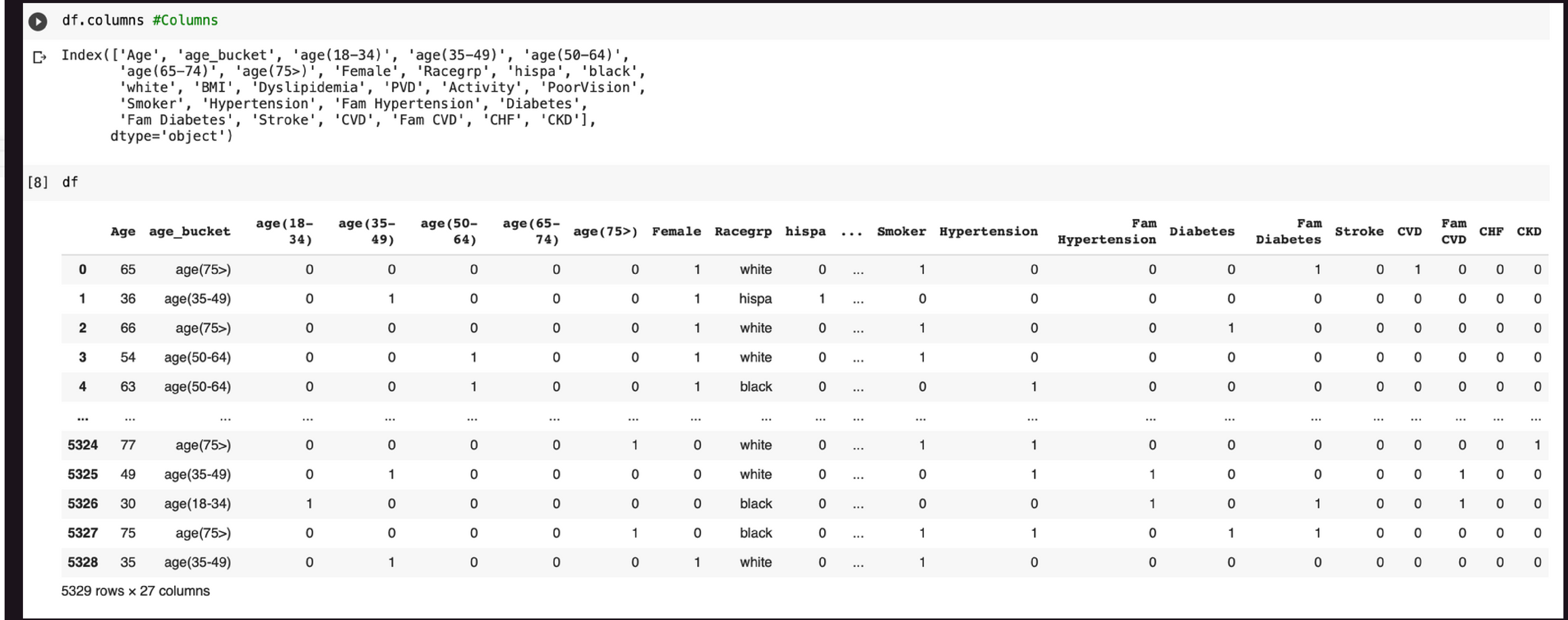
- AGE: [18-34], [35-49], [50- 64], [65-74], [75+]: 5
- BMI [<18.5], [18.5- 24.9], [25-29.9], [>30]

Racegrp - One Hot Encoded according to the different races.

Final Dataset



```
[16] df.shape

     (5329, 36)

[17] df.columns

     Index(['Age', 'age_bucket', 'age(18-34)', 'age(35-49)', 'age(50-64)',
            'age(65-74)', 'age(75>)', 'Female', 'Racegrp', 'hispa', 'black',
            'white', 'BMI', 'bmi_bucket', 'BMI<18.5', 'BMI(18.5-24.9)',
            'BMI(25-29.9)', 'BMI(>30)', 'Dyslipidemia', 'PVD', 'Activity',
            'Activity1', 'Activity2', 'Activity3', 'Activity4', 'PoorVision',
            'Smoker', 'Hypertension', 'Fam Hypertension', 'Diabetes',
            'Fam Diabetes', 'Stroke', 'CVD', 'Fam CVD', 'CHF', 'CKD'],
           dtype='object')

     df.isna().sum()

     Age                  0
     age_bucket           0
     age(18-34)           0
     age(35-49)           0
     age(50-64)           0
     age(65-74)           0
     age(75>)             0
     Female               0
     Racegrp              0
     hispa                0
     black                0
     white                0
     BMI                  0
     bmi_bucket           0
     BMI<18.5             0
     BMI(18.5-24.9)       0
     BMI(25-29.9)         0
     BMI(>30)             0
     Dyslipidemia         0
     PVD                  0
     Activity             0
     Activity1            0
     Activity2            0
     Activity3            0
     Activity4            0
     PoorVision           0
     Smoker               0
     Hypertension         0
     Fam Hypertension     0
     Diabetes             0
     Fam Diabetes         0
     Stroke               0
     CVD                  0
     Fam CVD              0
     CHF                  0
     CKD                  0
     dtype: int64
```

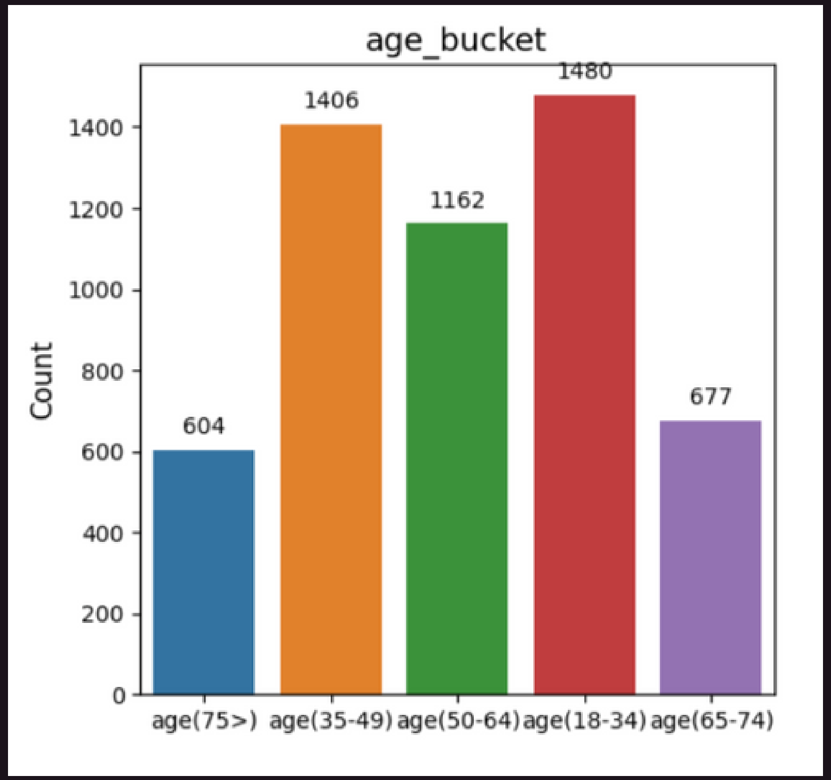


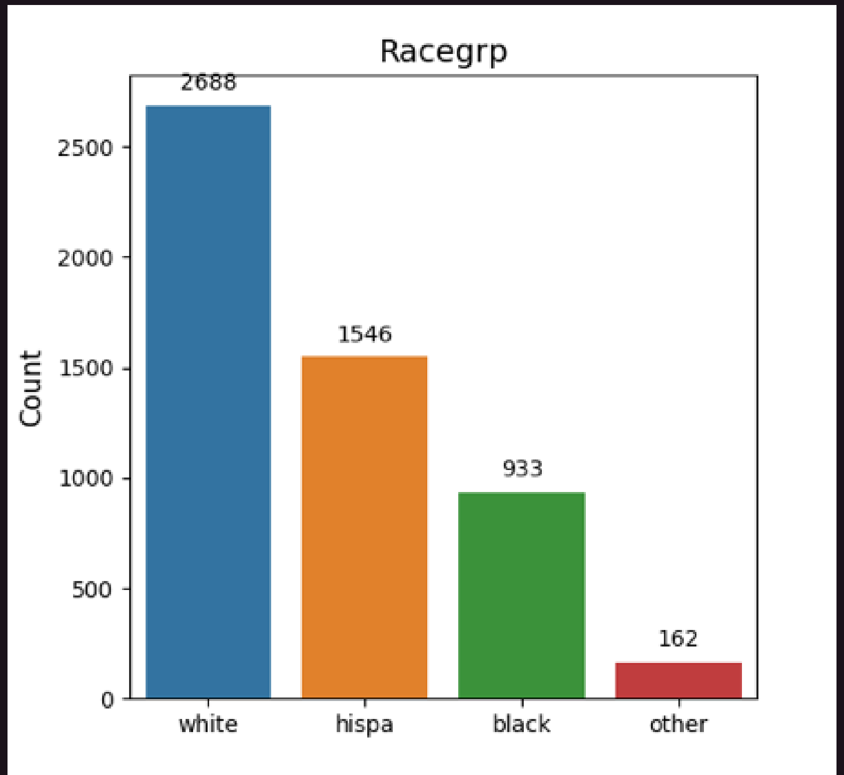
```
df.columns #Columns

Index(['Age', 'age_bucket', 'age(18-34)', 'age(35-49)', 'age(50-64)',
       'age(65-74)', 'age(75>)', 'Female', 'Racegrp', 'hispa', 'black',
       'white', 'BMI', 'Dyslipidemia', 'PVD', 'Activity', 'PoorVision',
       'Smoker', 'Hypertension', 'Fam Hypertension', 'Diabetes',
       'Fam Diabetes', 'Stroke', 'CVD', 'Fam CVD', 'CHF', 'CKD'],
      dtype='object')

[8] df
```

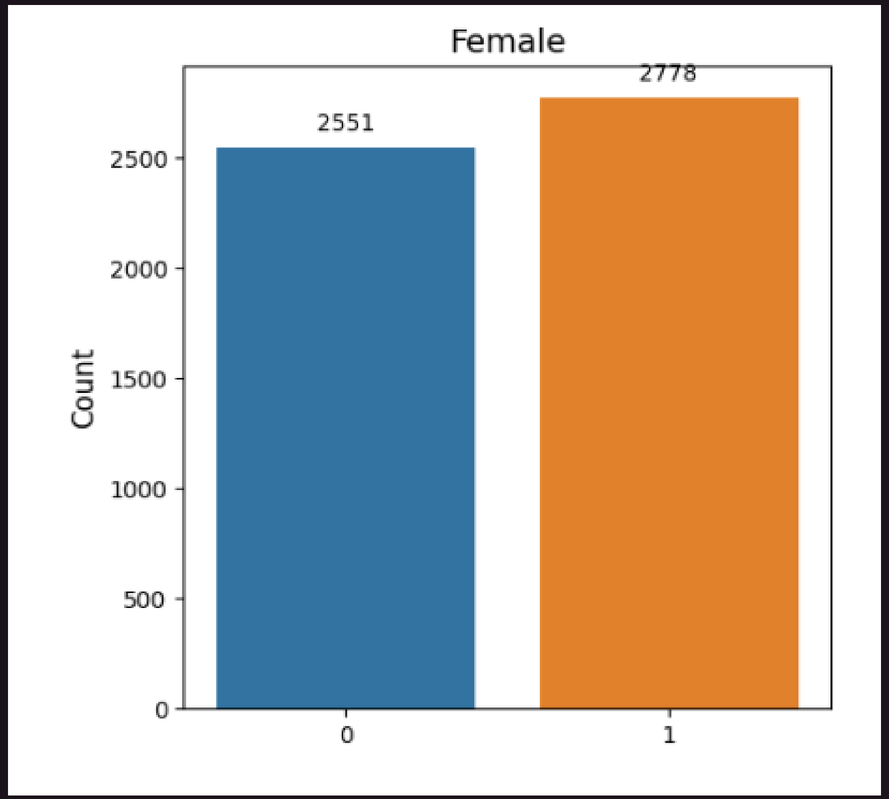
	Age	age_bucket	age(18-34)	age(35-49)	age(50-64)	age(65-74)	age(75>)	Female	Racegrp	hispa	...	Smoker	Hypertension	Fam Hypertension	Diabetes	Fam Diabetes	Stroke	CVD	Fam CVD	CHF	CKD
0	65	age(75>)	0	0	0	0	0	1	white	0	...	1	0	0	0	1	0	1	0	0	0
1	36	age(35-49)	0	1	0	0	0	1	hispa	1	...	0	0	0	0	0	0	0	0	0	0
2	66	age(75>)	0	0	0	0	0	1	white	0	...	1	0	0	1	0	0	0	0	0	0
3	54	age(50-64)	0	0	1	0	0	1	white	0	...	1	0	0	0	0	0	0	0	0	0
4	63	age(50-64)	0	0	1	0	0	1	black	0	...	0	1	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5324	77	age(75>)	0	0	0	0	1	0	white	0	...	1	1	0	0	0	0	0	0	0	1
5325	49	age(35-49)	0	1	0	0	0	0	white	0	...	0	1	1	0	0	0	0	1	0	0
5326	30	age(18-34)	1	0	0	0	0	0	black	0	...	0	0	1	0	1	0	0	1	0	0
5327	75	age(75>)	0	0	0	0	1	0	black	0	...	1	1	0	1	1	0	0	0	0	0
5328	35	age(35-49)	0	1	0	0	0	1	white	0	...	1	0	0	0	0	0	0	0	0	0

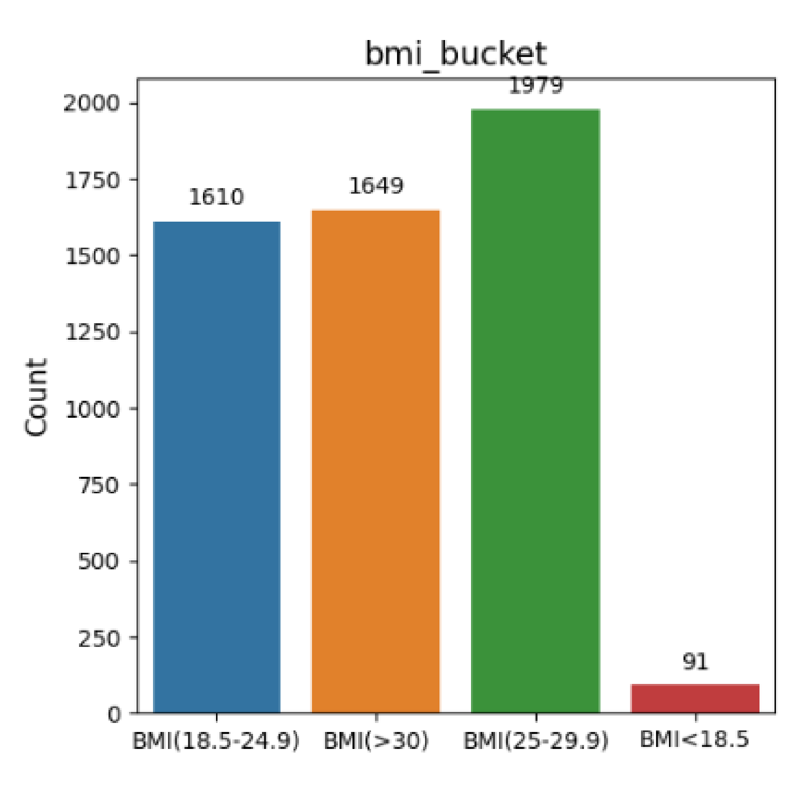
5329 rows × 27 columns

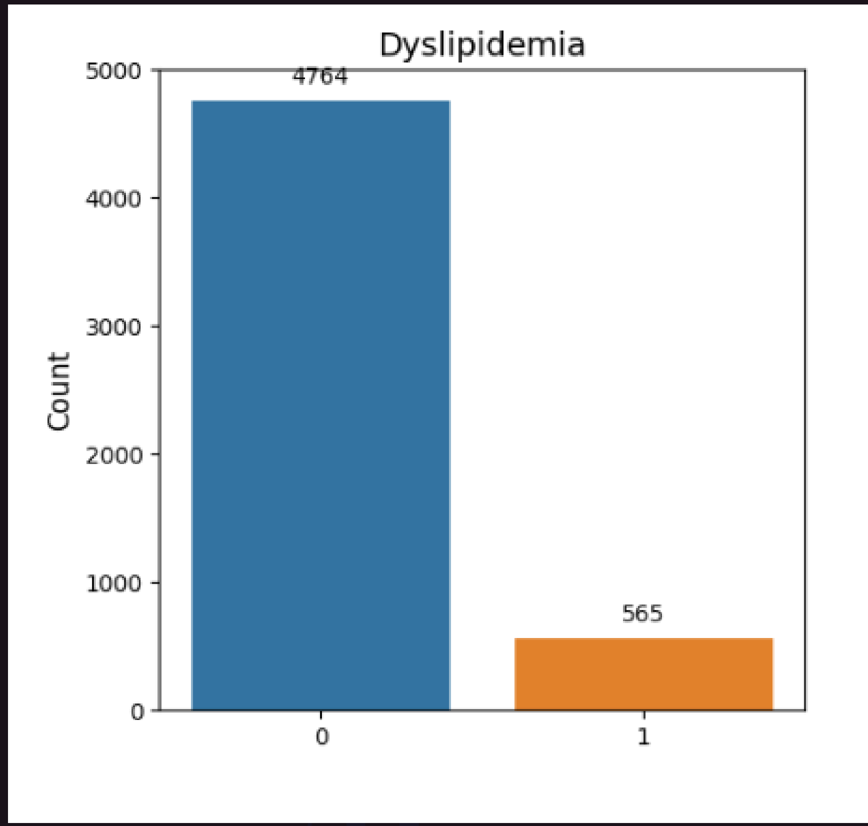
Uni Variate Analysis

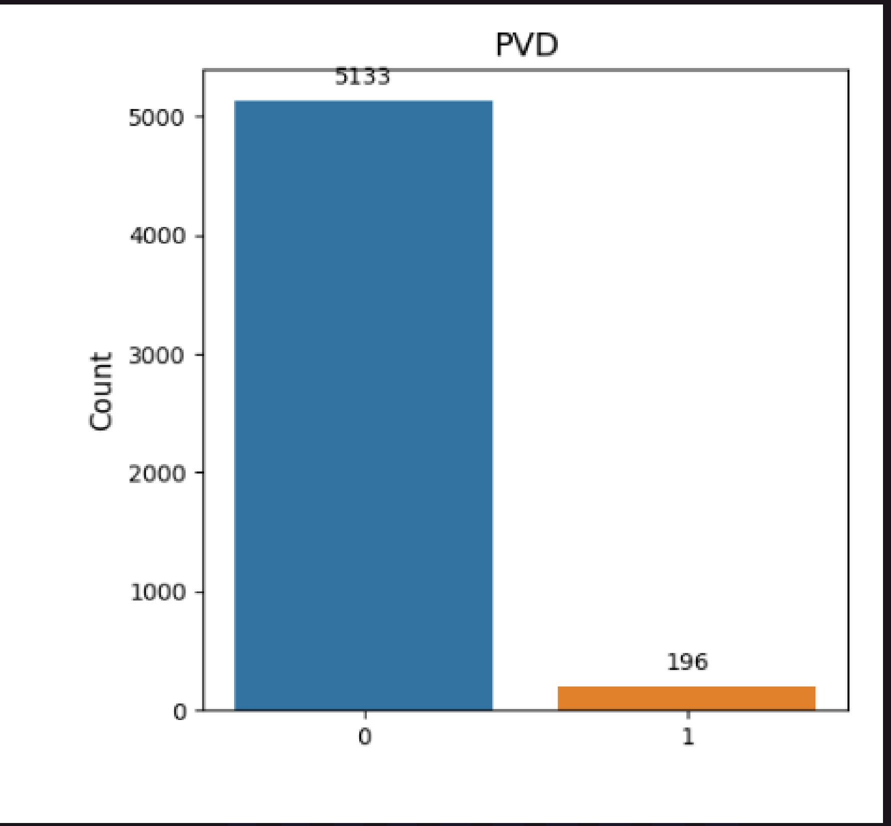

age_bucket

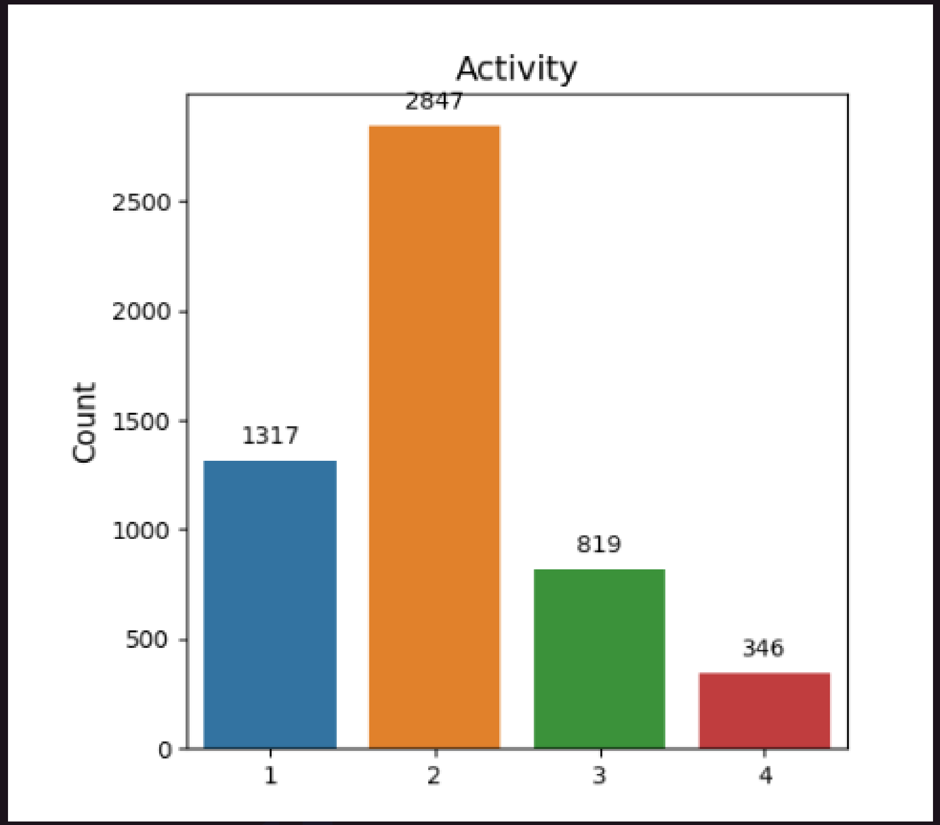

Racegrp

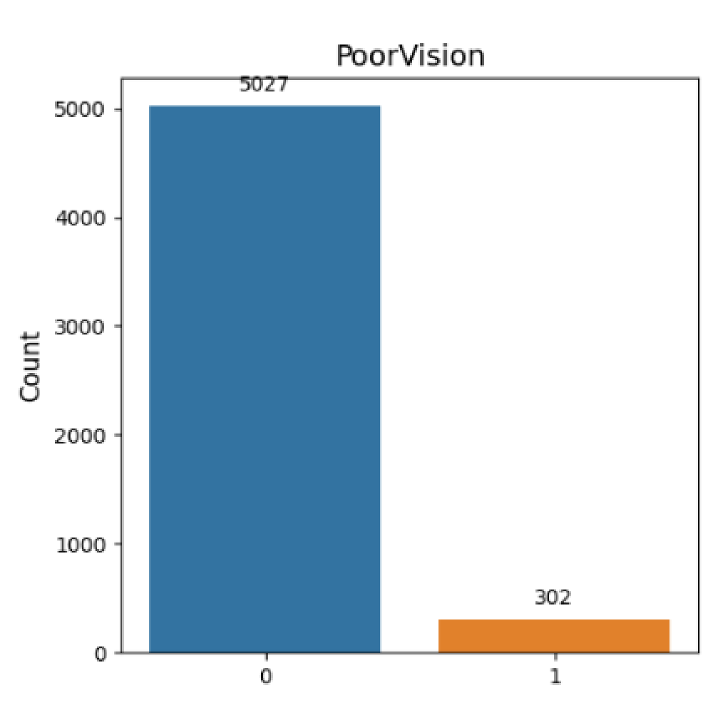

Female


bmi_bucket

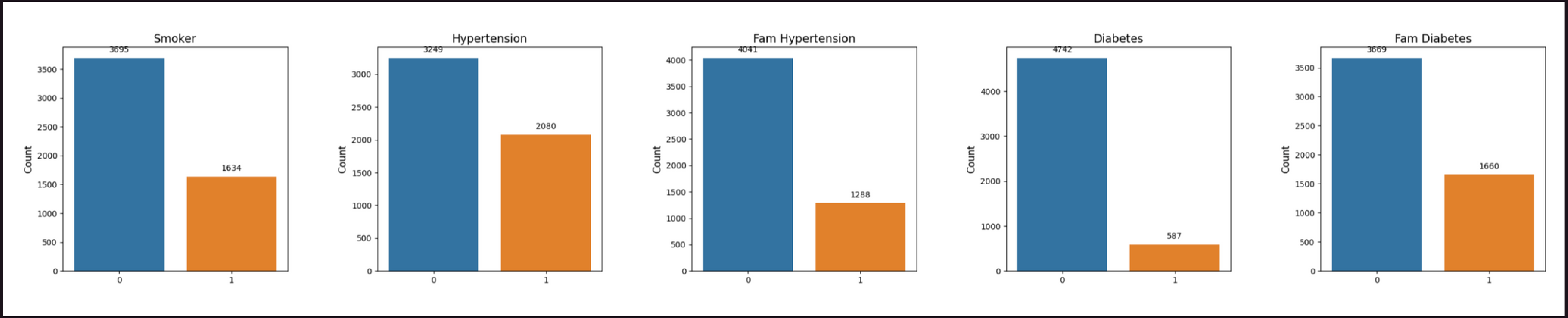

Dyslipidemia

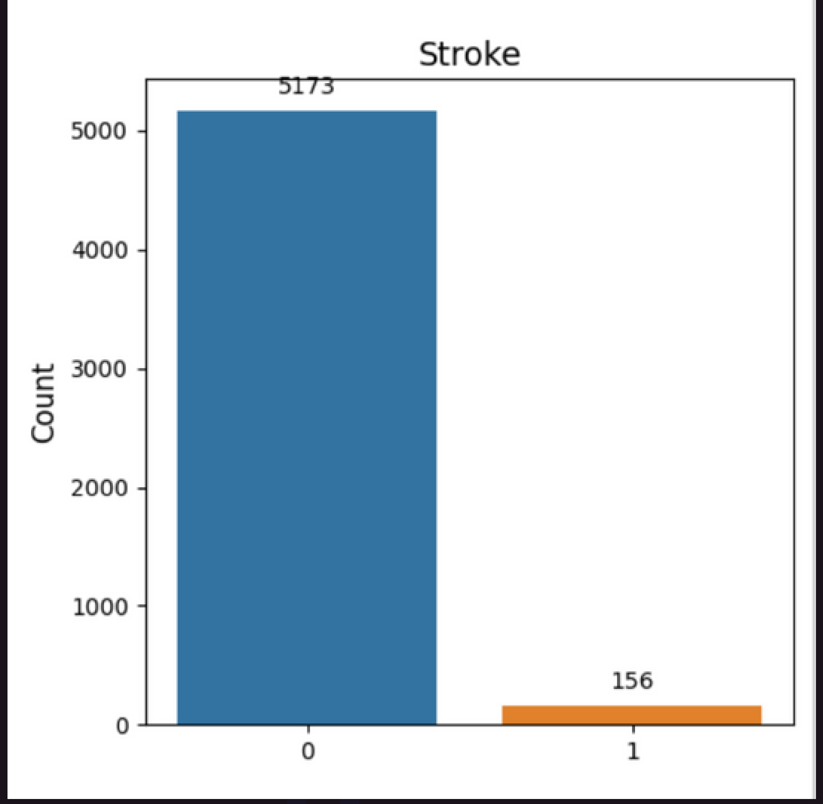

PVD

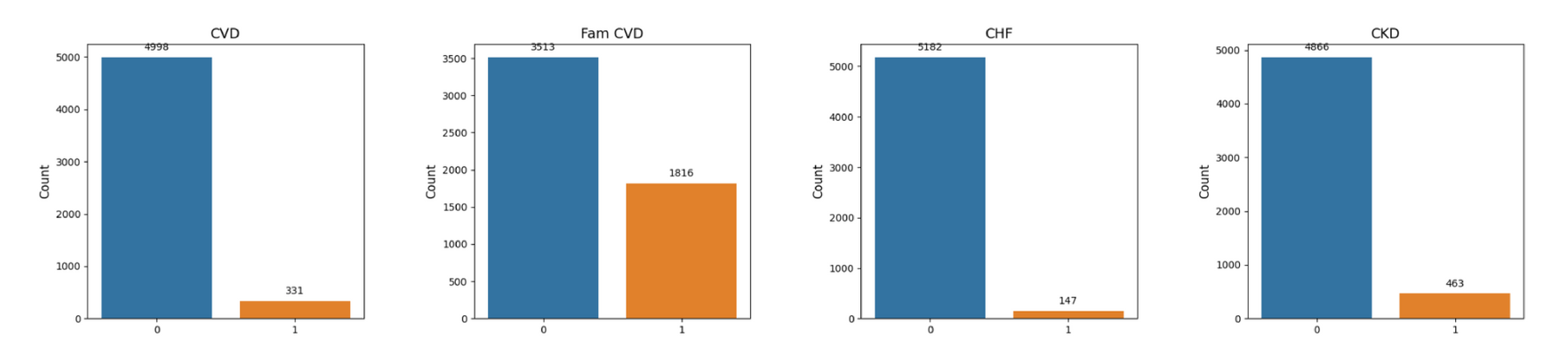

Activity


PoorVision

Univariate Analysis

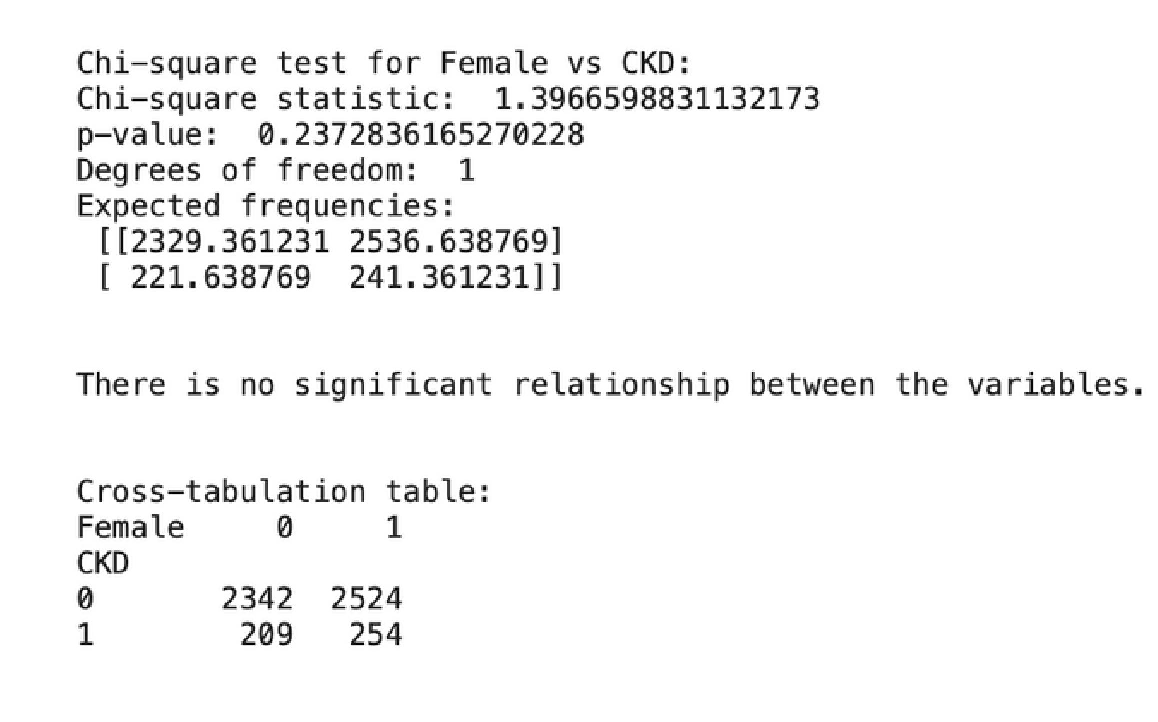




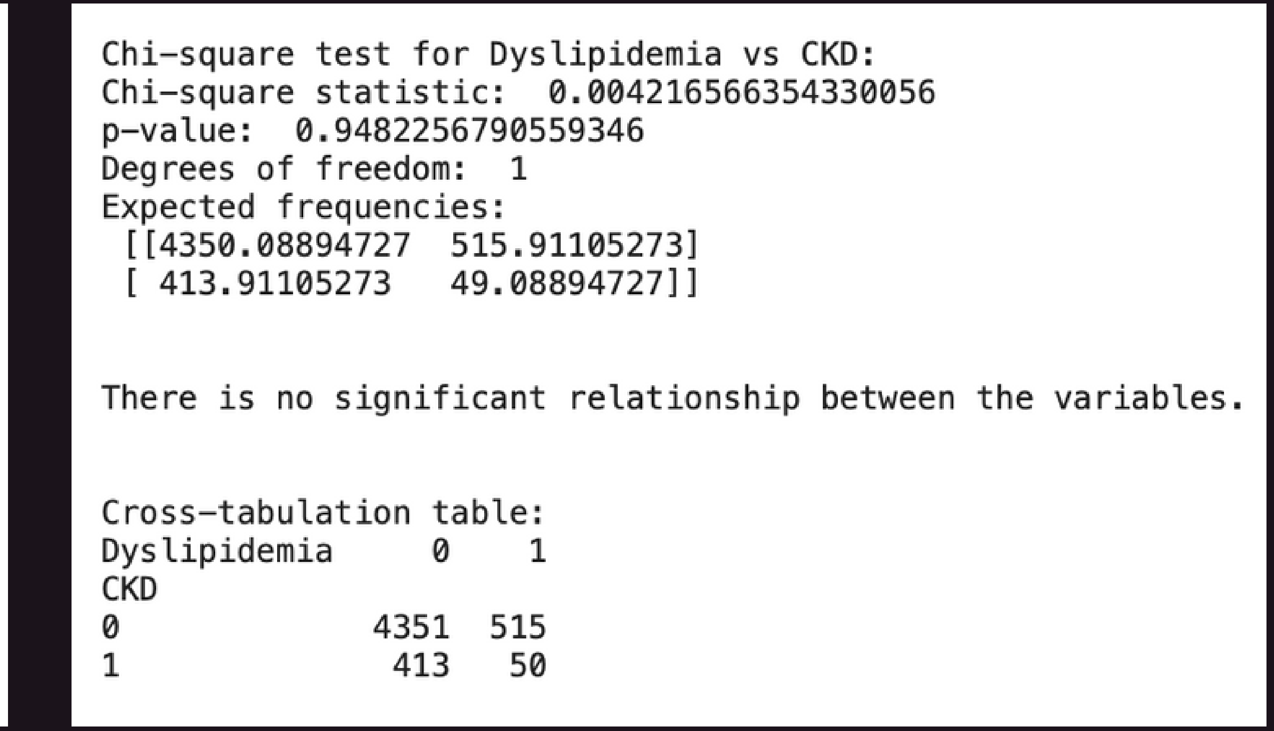


Bi Variate Analysis

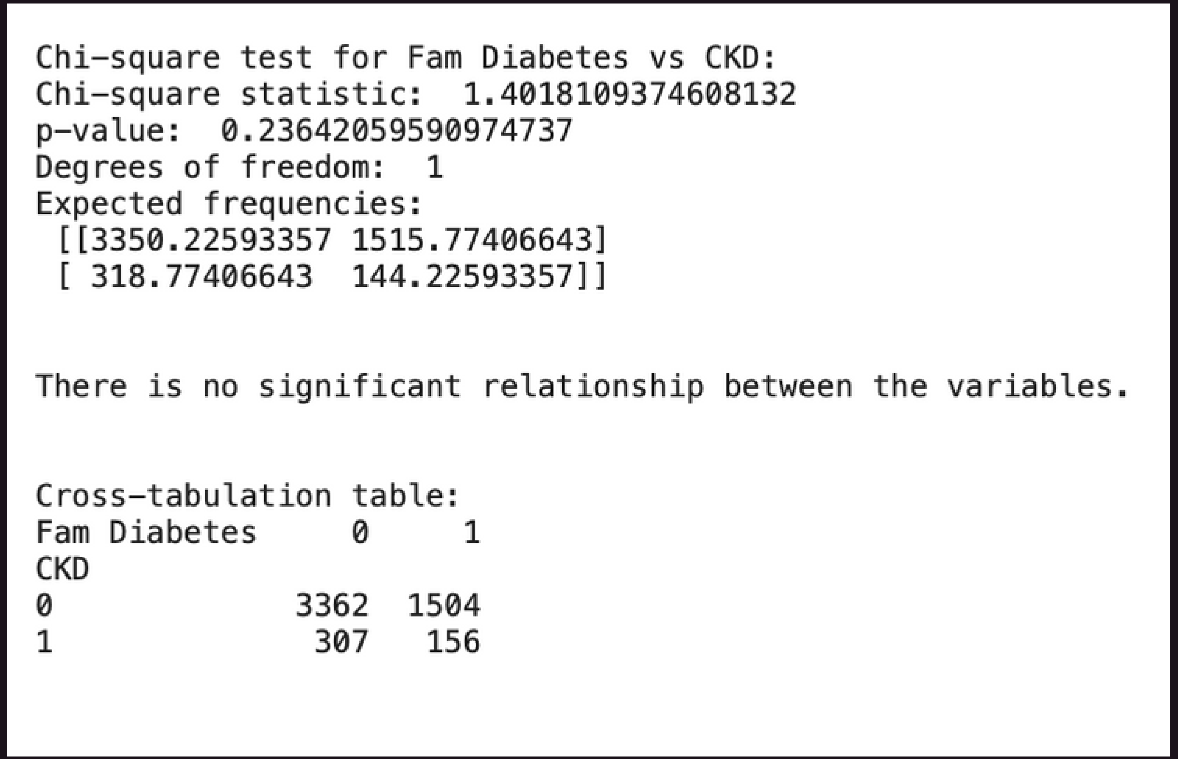
Variables for BI Variate Analysis - All of them are categorical. Performed CHI Square tests to determine any significant relationship with the target variable.



```
Chi-square test for Female vs CKD:
Chi-square statistic:  1.3966598831132173
p-value:  0.2372836165270228
Degrees of freedom:  1
Expected frequencies:
 [[2329.361231 2536.638769]
 [ 221.638769  241.361231]]


There is no significant relationship between the variables.


Cross-tabulation table:
Female     0     1
CKD
0       2342  2524
1        209   254
```



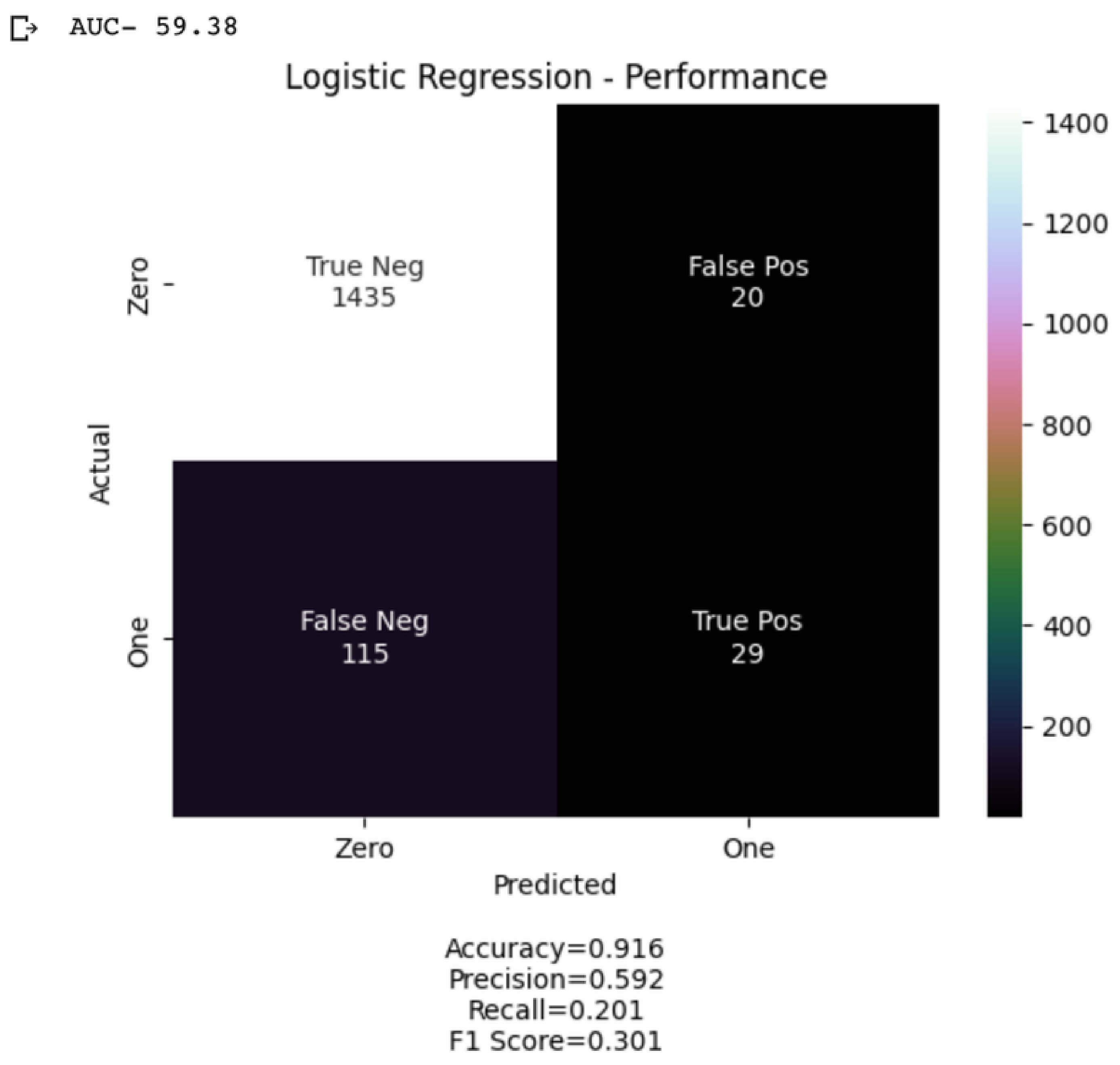
```
Chi-square test for Dyslipidemia vs CKD:
Chi-square statistic:  0.004216566354330056
p-value:  0.9482256790559346
Degrees of freedom:  1
Expected frequencies:
 [[4350.08894727  515.91105273]
 [ 413.91105273   49.08894727]]


There is no significant relationship between the variables.


Cross-tabulation table:
Dyslipidemia     0    1
CKD
0             4351  515
1              413   50
```



```
Chi-square test for Fam Diabetes vs CKD:
Chi-square statistic:  1.4018109374608132
p-value:  0.23642059590974737
Degrees of freedom:  1
Expected frequencies:
 [[3350.22593357 1515.77406643]
 [ 318.77406643  144.22593357]]


There is no significant relationship between the variables.


Cross-tabulation table:
Fam Diabetes     0     1
CKD
0             3362  1504
1              307   156
```

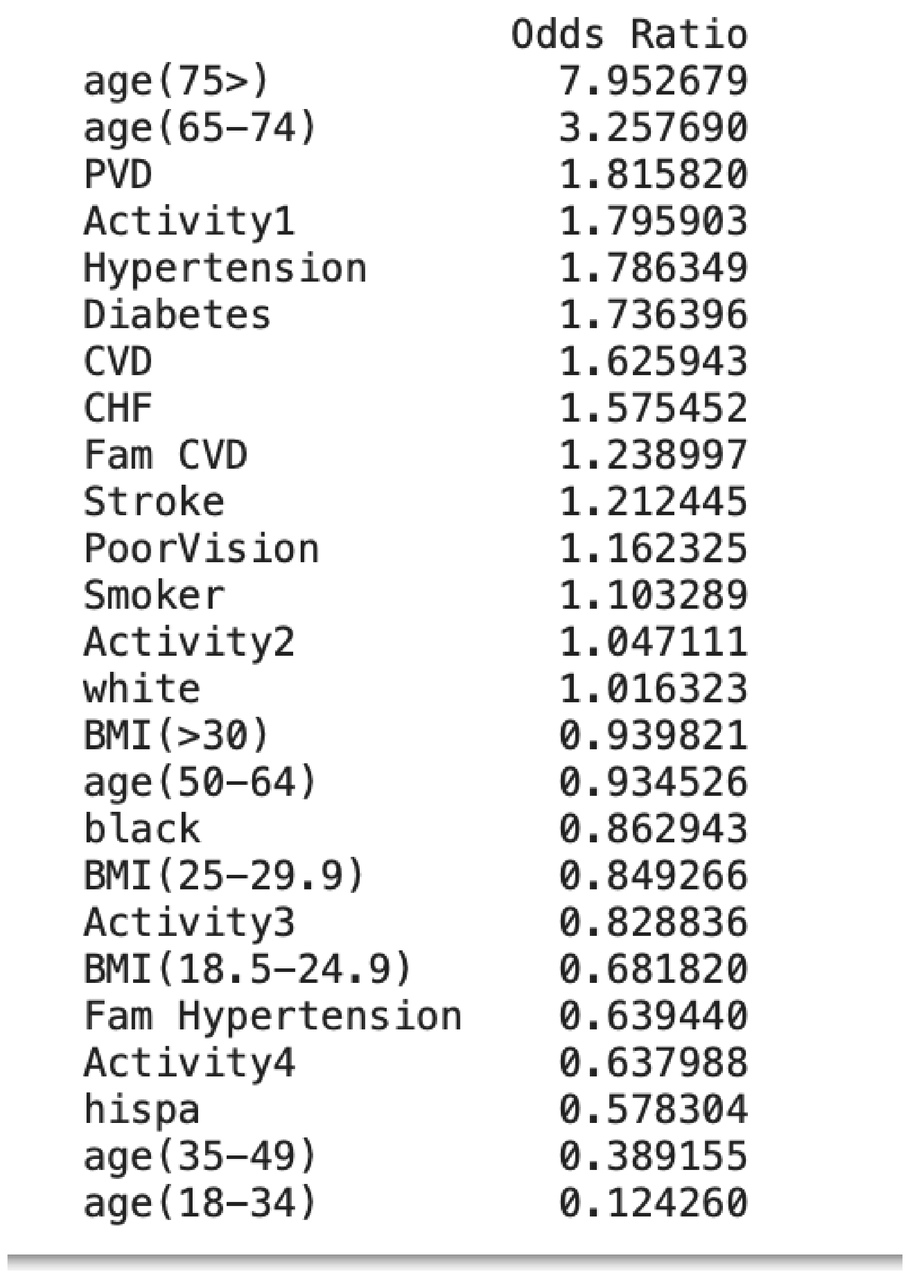
These are three variables that had no significant relationship with our target variable.
Gender, Family Diabetic History, and Dyslipidemia. Removed these variables from our final model.

Model Built Logistic Regression



```
AUC- 59.38
```

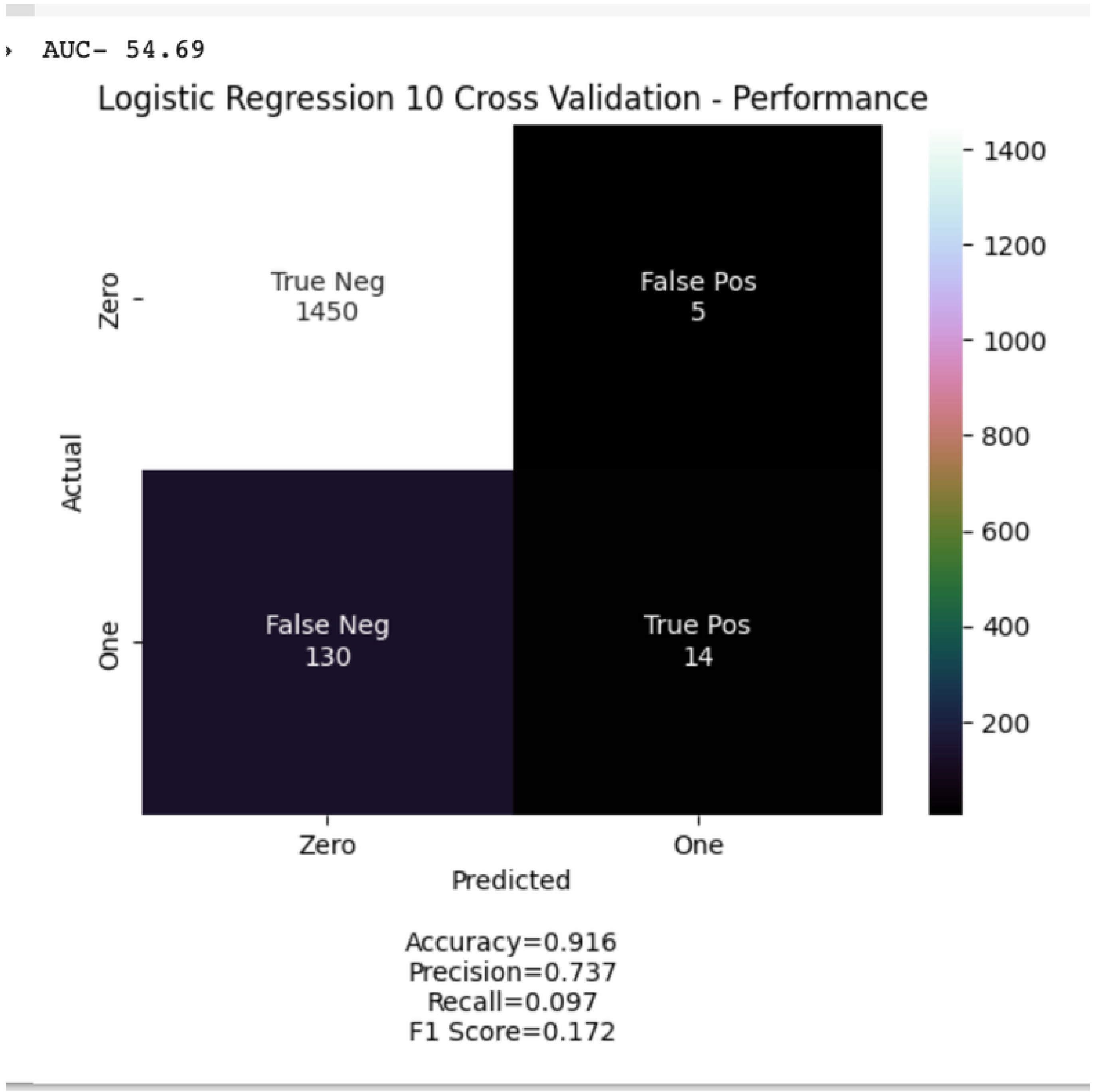
Accuracy=0.916
Precision=0.592
Recall=0.201
F1 Score=0.301



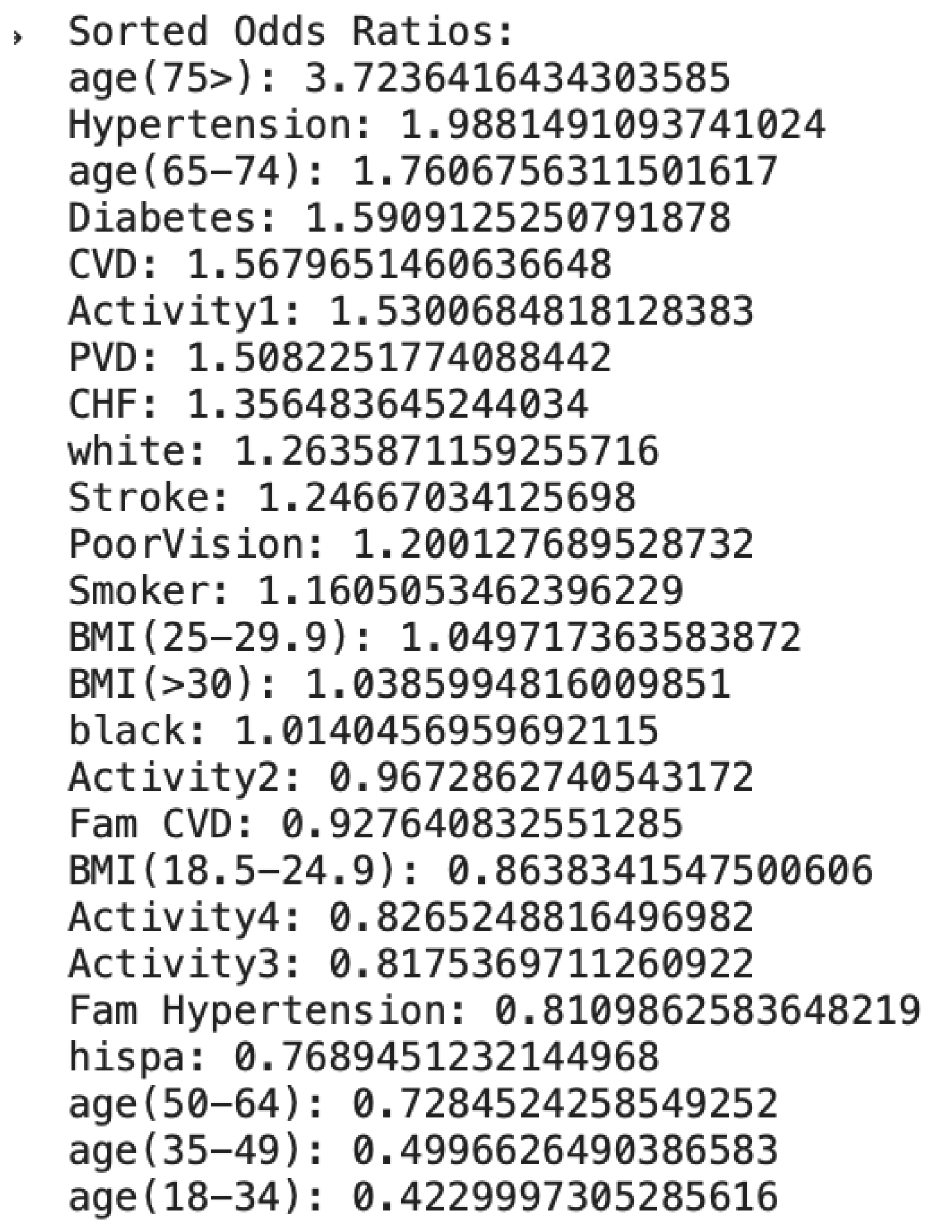
```
                    Odds Ratio
age(75>)              7.952679
age(65-74)            3.257690
PVD                   1.815820
Activity1             1.795903
Hypertension          1.786349
Diabetes              1.736396
CVD                   1.625943
CHF                   1.575452
Fam CVD               1.238997
Stroke                1.212445
PoorVision            1.162325
Smoker                1.103289
Activity2             1.047111
white                 1.016323
BMI(>30)              0.939821
age(50-64)            0.934526
black                 0.862943
BMI(25-29.9)          0.849266
Activity3             0.828836
BMI(18.5-24.9)        0.681820
Fam Hypertension      0.639440
Activity4             0.637988
hispa                 0.578304
age(35-49)            0.389155
age(18-34)            0.124260
```

- 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 Fam Hypertension and BMI>30 have weird odds ratio.

Logistic Regression | 10 Cross Validation



```
AUC- 54.69
```



```
Sorted Odds Ratios:
age(75>): 3.7236416434303585
Hypertension: 1.9881491093741024
age(65-74): 1.7606756311501617
Diabetes: 1.5909125250791878
CVD: 1.567965146063664
Activity1: 1.5300684818128383
PVD: 1.5082251774088442
CHF: 1.356483645244034
white: 1.2635871159255716
Stroke: 1.24667034125698
PoorVision: 1.200127689528732
Smoker: 1.1605053462396229
BMI(25-29.9): 1.049717363583872
BMI(>30): 1.0385994816009851
black: 1.014045695969211
Activity2: 0.9672862740543172
Fam CVD: 0.927640832551285
BMI(18.5-24.9): 0.8638341547500606
Activity4: 0.8265248816496982
Activity3: 0.8175369711260922
Fam Hypertension: 0.8109862583648219
hispa: 0.7689451232144968
age(50-64): 0.7284524258549252
age(35-49): 0.4996626490386583
age(18-34): 0.4229997305285616
```

References

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