INSTITUTE OF
EMERGING
CAREERS

Data Analytics

Portfolio Project-2

Prepared By: Syed Ali Rehan

PHASES OF ANALYSIS

Data Pre processing and Cleaning

Exploratory Data Analysis (EDA)

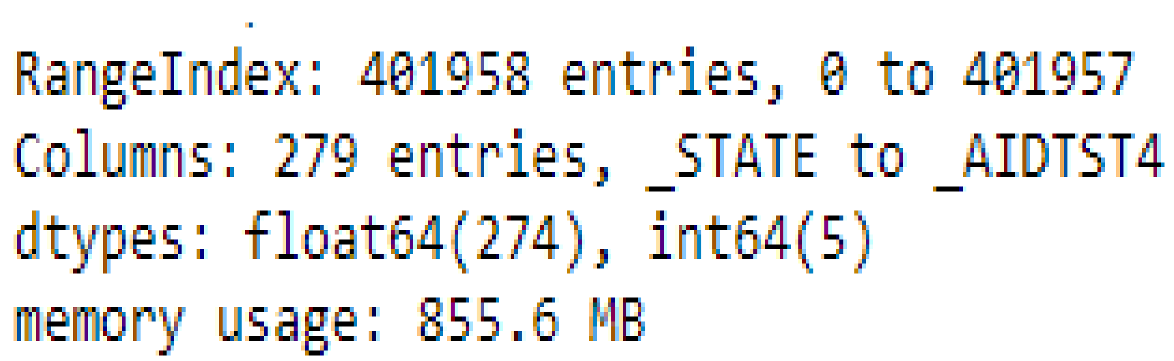
Hypothesis Testing

Regression Analysis

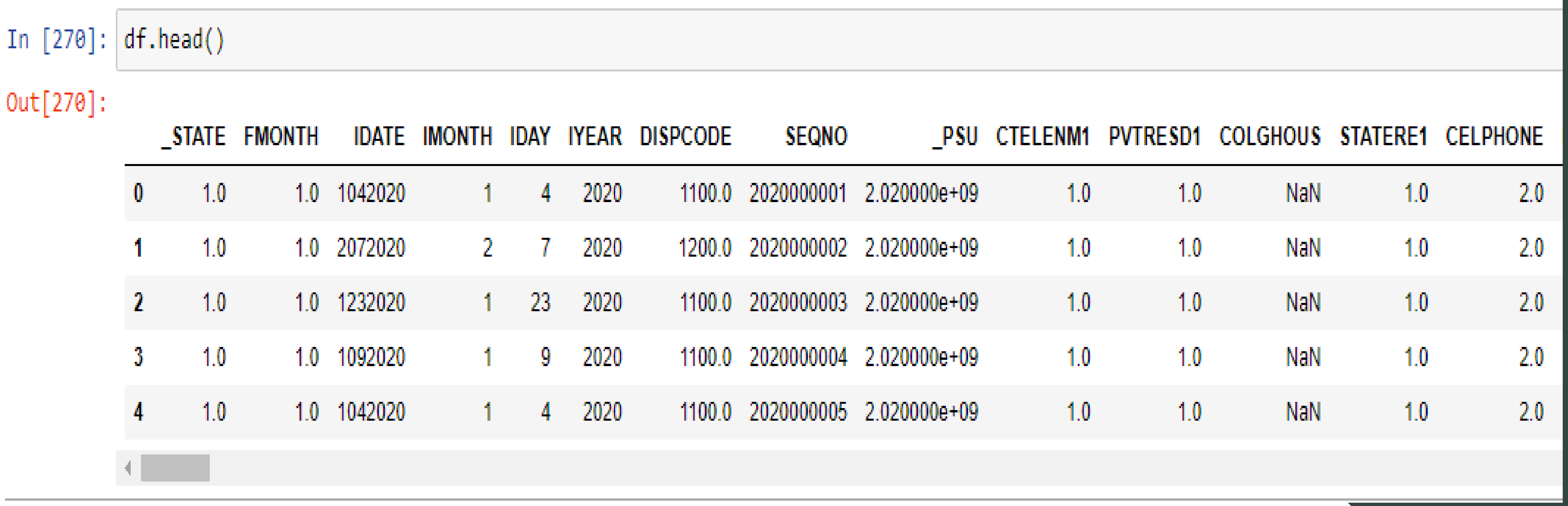
BEHAVIORAL RISK FACTOR SURVEILLANCE SYSTEM

- BRFSS is a nationwide health telephone survey system.
- It gathers data on health behaviors, chronic conditions, and preventive services usage.
- Established in 1984 with 15 states, it now covers all 50 states, D.C., and three territories.
- BRFSS conducts over 400,000 adult interviews annually, making it the world's largest continuous health survey system.

Over view of BRFSS Dataset?

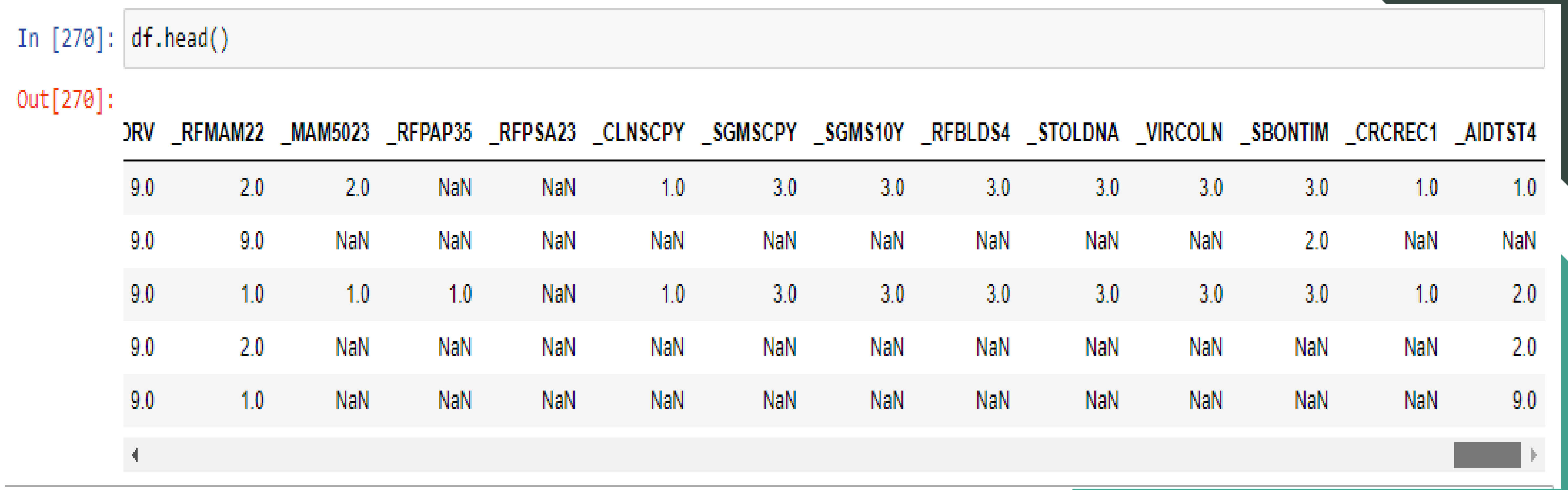


```
RangeIndex: 401958 entries, 0 to 401957
Columns: 279 entries, _STATE to _AIDTST4
dtypes: float64(274), int64(5)
memory usage: 855.6 MB
```



```
In [270]: df.head()
Out[270]:
```

	_STATE	FMONTH	IDATE	IMONTH	IDAY	IYEAR	DISPCODE	SEQNO	_PSU	CTELENM1	PVTRESD1	COLGHOUS	STATERE1	CELPHONE
0	1.0	1.0	1042020	1	4	2020	1100.0	2020000001	2.020000e+09	1.0	1.0	NaN	1.0	2.0
1	1.0	1.0	2072020	2	7	2020	1200.0	2020000002	2.020000e+09	1.0	1.0	NaN	1.0	2.0
2	1.0	1.0	1232020	1	23	2020	1100.0	2020000003	2.020000e+09	1.0	1.0	NaN	1.0	2.0
3	1.0	1.0	1092020	1	9	2020	1100.0	2020000004	2.020000e+09	1.0	1.0	NaN	1.0	2.0
4	1.0	1.0	1042020	1	4	2020	1100.0	2020000005	2.020000e+09	1.0	1.0	NaN	1.0	2.0



```
In [270]: df.head()
Out[270]:
```

DRV	_RFMAM22	_MAM5023	_RFPAP35	_RFPSA23	_CLNSCPY	_SGMSCPY	_SGMS10Y	_RFBLDS4	_STOLDNA	_VIRCOLN	_SBONTIM	_CRCREC1	_AIDTST4
9.0	2.0	2.0	NaN	NaN	1.0	3.0	3.0	3.0	3.0	3.0	3.0	1.0	1.0
9.0	9.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2.0	NaN	NaN
9.0	1.0	1.0	1.0	NaN	1.0	3.0	3.0	3.0	3.0	3.0	3.0	1.0	2.0
9.0	2.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2.0
9.0	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	9.0

Analyzing Stroke Risk Factors in the U.S. Using BRFSS Data: Implications for Public Health and Prevention

Public Health Significance: Stroke is a major public health concern in the U.S., making this analysis highly relevant.

Preventive Measures: Identifying stroke risk factors aids in developing targeted prevention strategies.

Healthcare Resource Allocation: Informs resource allocation for effective stroke prevention efforts.

At-Risk Populations: Identifies high-risk demographic groups for tailored interventions.

Healthcare Planning: Assists in healthcare planning for stroke patient needs.

Research Contribution: Adds to existing stroke risk factor research.

Improved Health Outcomes: Aims to reduce strokes, improving overall health and reducing healthcare costs.

STROKE

- A stroke, or brain attack, is an emergency where blood flow to the brain is blocked.
- The brain relies on constant oxygen and nutrient supply for proper function.
- Brief interruptions can cause brain cell death.
- Brain cell loss leads to functional impairment.
- Stroke can affect:
    1. Movement
    2. Speech
    3. Eating
    4. Thinking and memory
    5. Bodily functions
    6. Emotional control

RISK FACTORS

Non-modifiable Risk Factors:

- Age
- Gender
- Race

Modifiable Risk Factors:

- High Blood Pressure (Hypertension)
- Smoking:
- High Cholesterol:
- Diabetes:
- Obesity
- Physical In-activity
- Excessive Alcohol Consumption
- Heart Disease

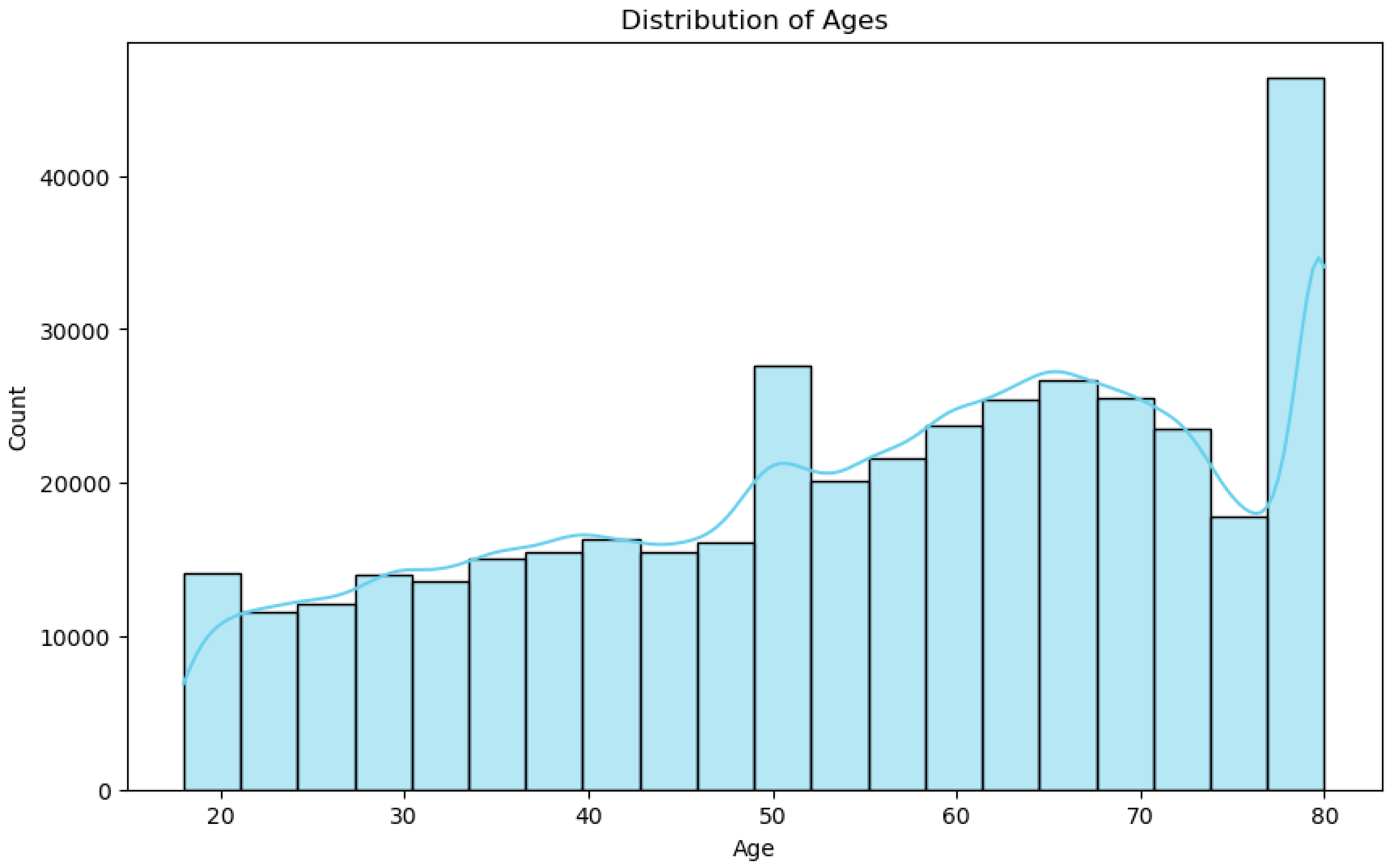
Risk factors columns available:

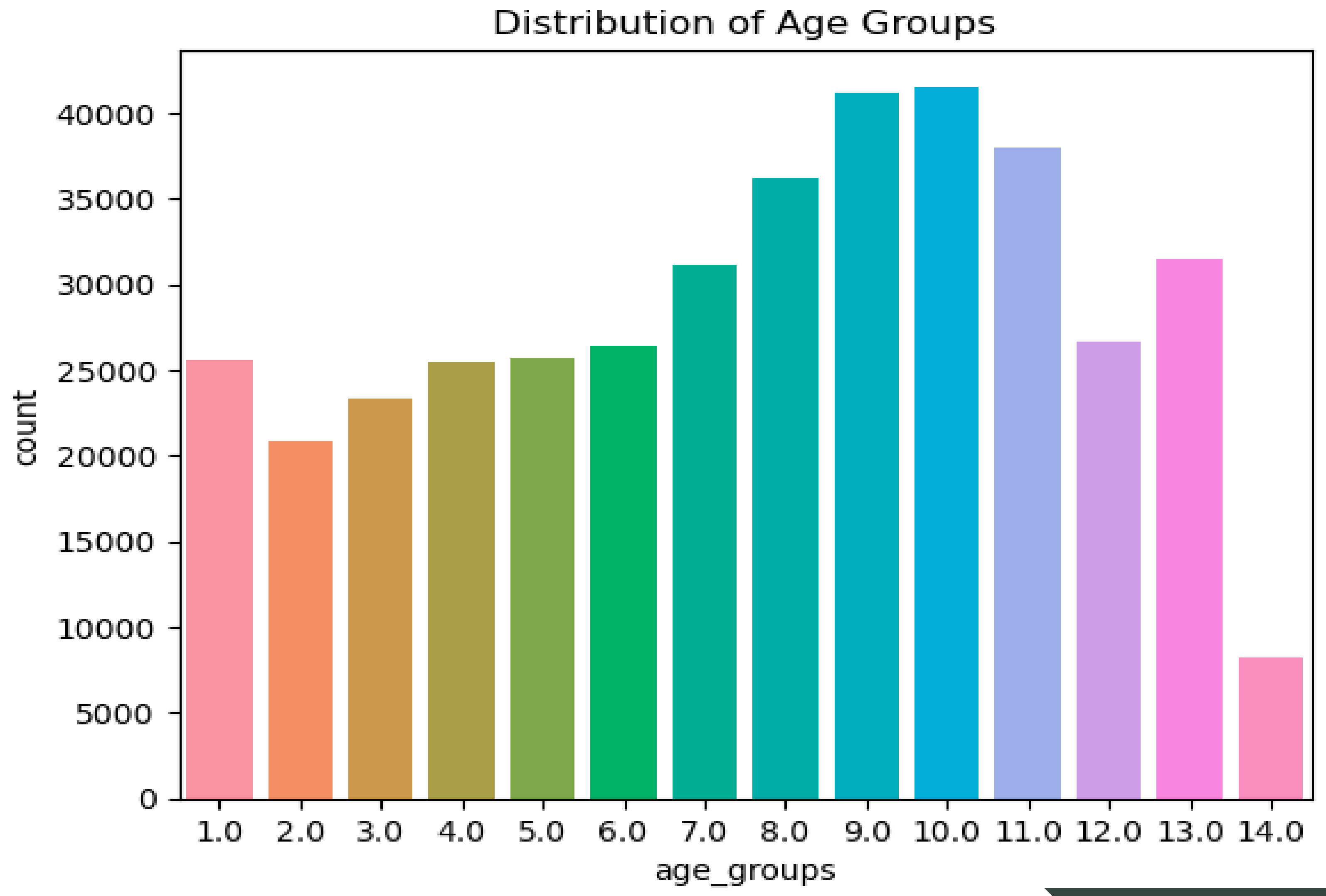
Non-modifiable Factors

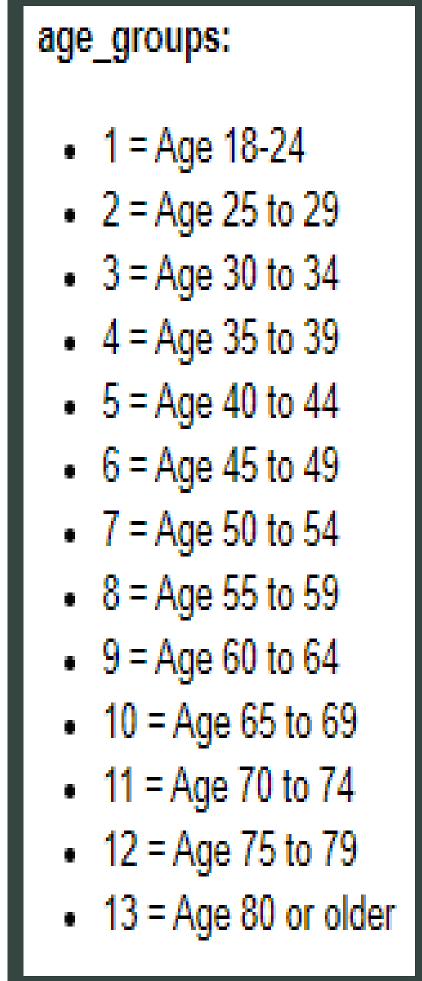
1) _AGE80 (ages in numbers)
2) _AGEG5YR (age groups)
3) _SEX (gender)
4) _IMPRACE (race)

Modifiable Factors

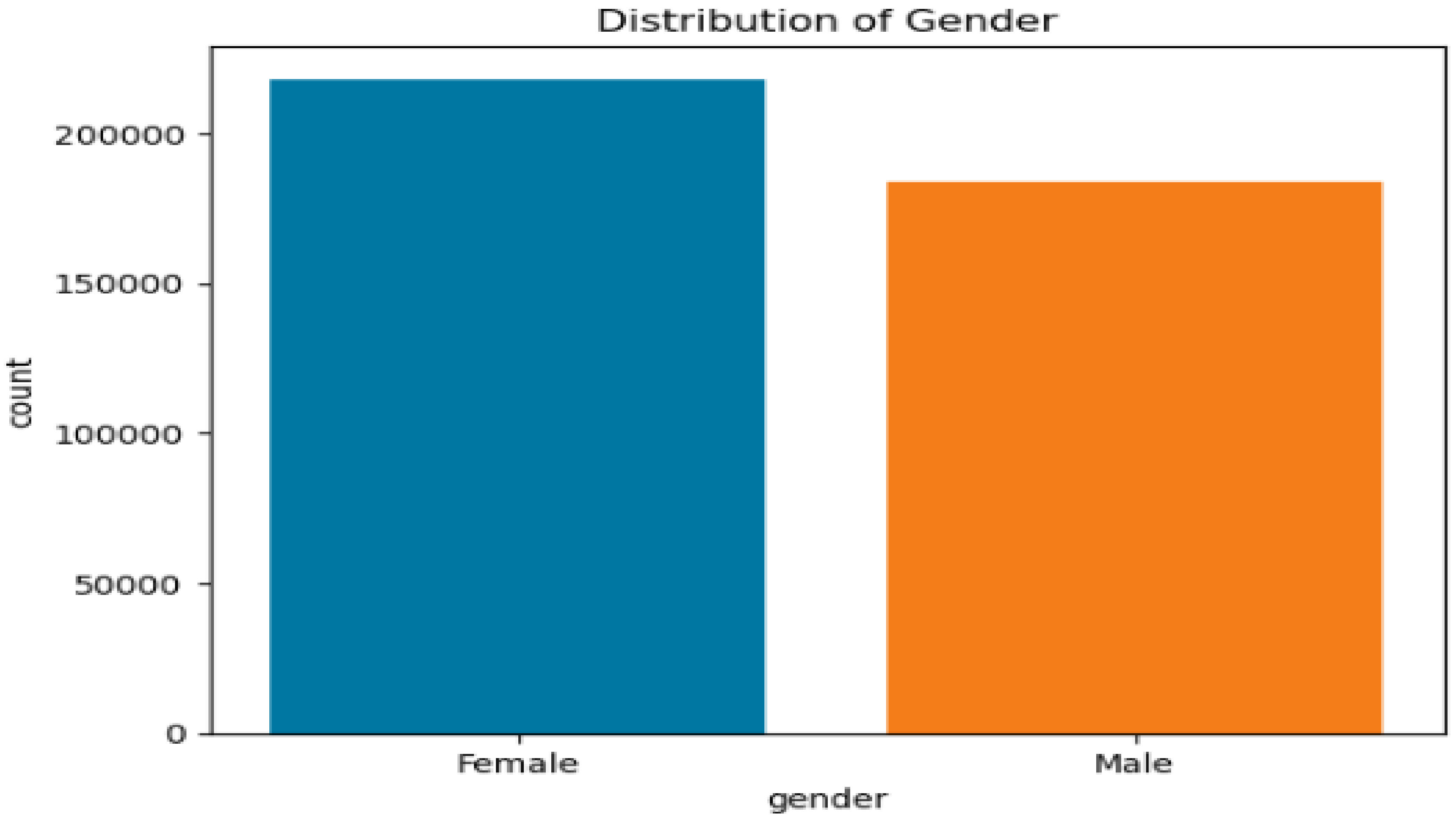
1) _SMOKER3 (smoking)
2) DIABETE4 (diabetes)
3) _RFBMI5 (obesity)
4) _TOTINDA (physical activity)
5) _RFDRHV7 (heavy drinker)
6) CVDINFR4 (heart attack)
7) CVDCRHD4 (angina or other CHD)

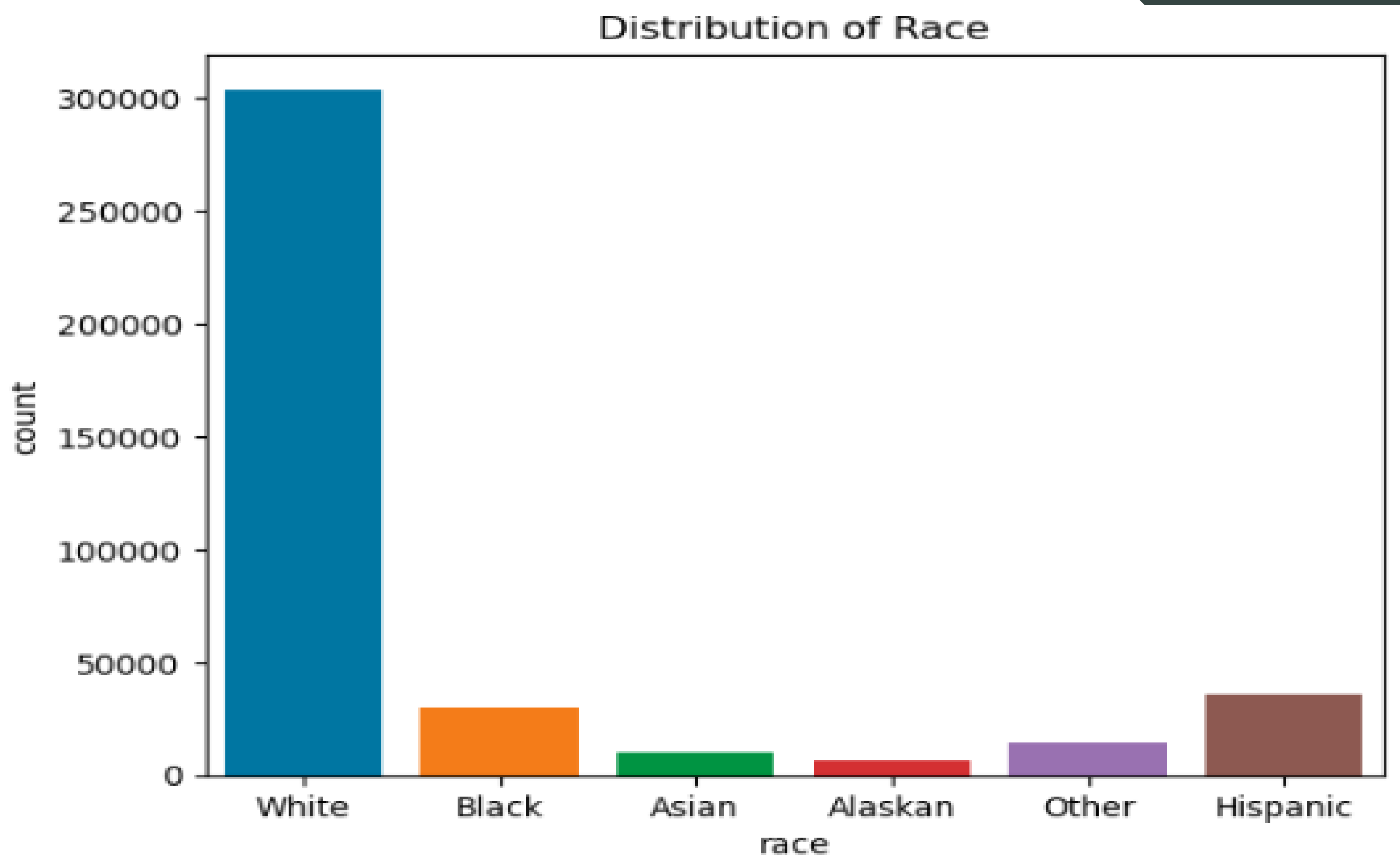


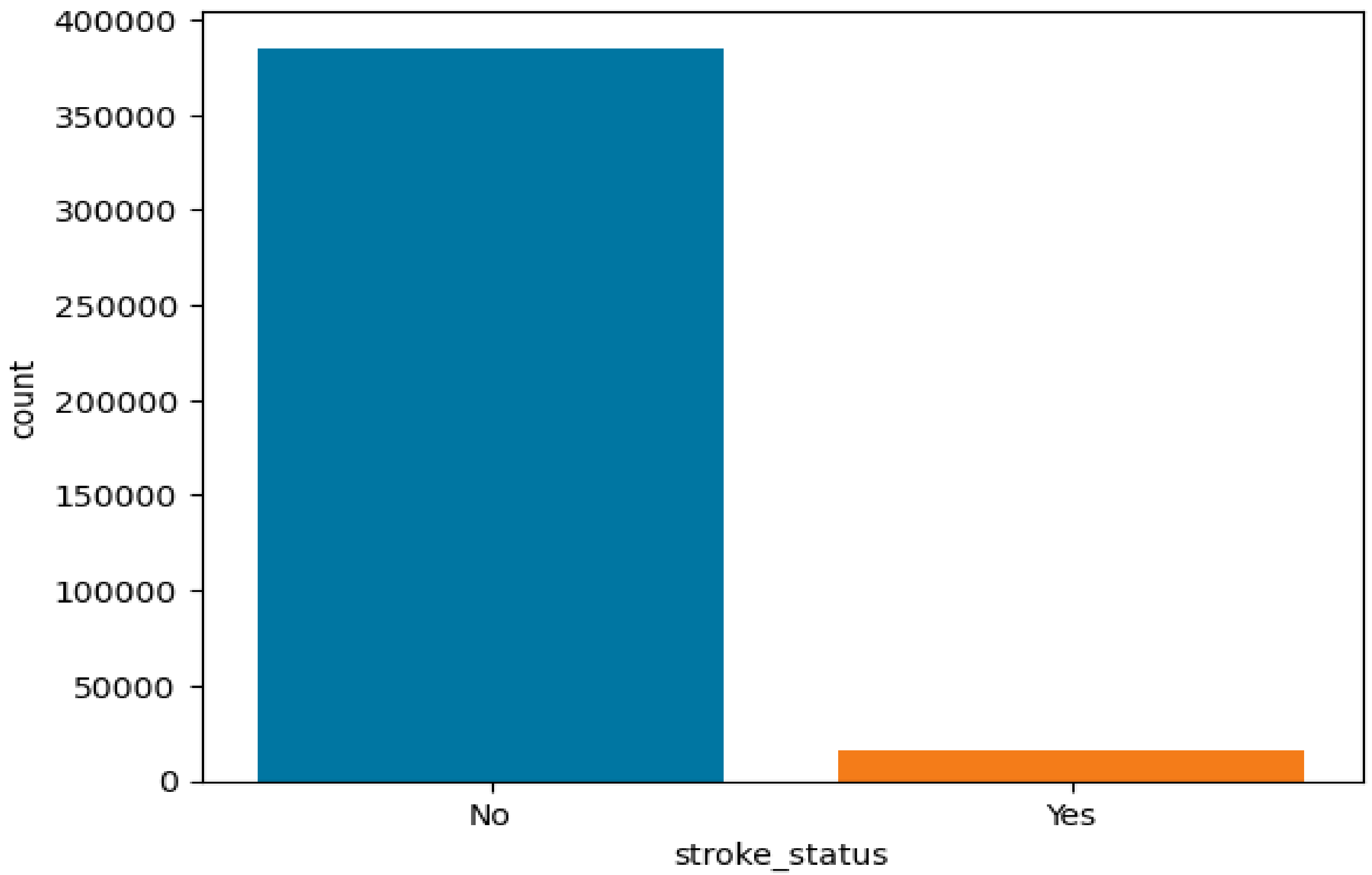



age_groups:

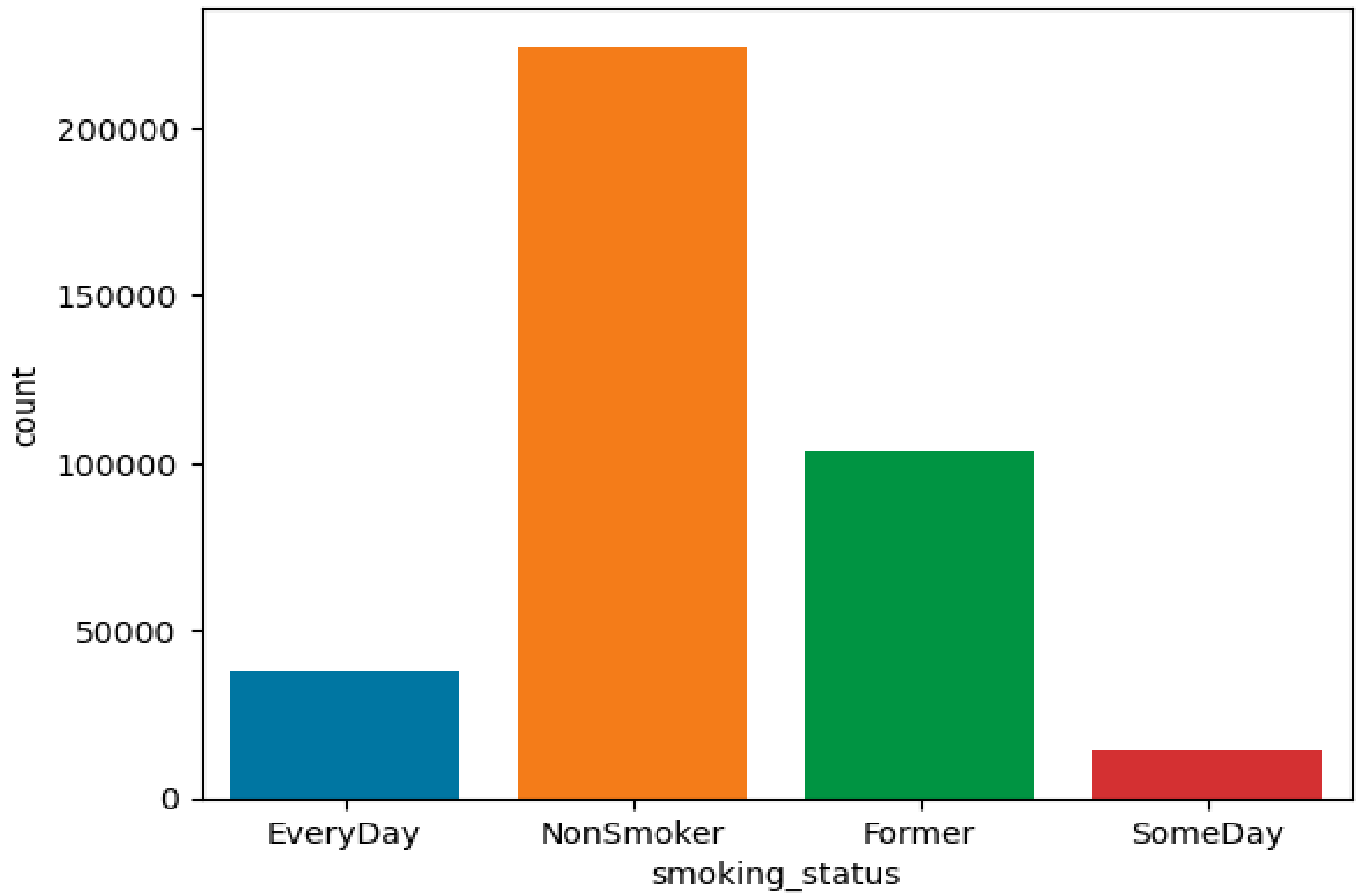
- 1 = Age 18-24
- 2 = Age 25 to 29
- 3 = Age 30 to 34
- 4 = Age 35 to 39
- 5 = Age 40 to 44
- 6 = Age 45 to 49
- 7 = Age 50 to 54
- 8 = Age 55 to 59
- 9 = Age 60 to 64
- 10 = Age 65 to 69
- 11 = Age 70 to 74
- 12 = Age 75 to 79
- 13 = Age 80 or older



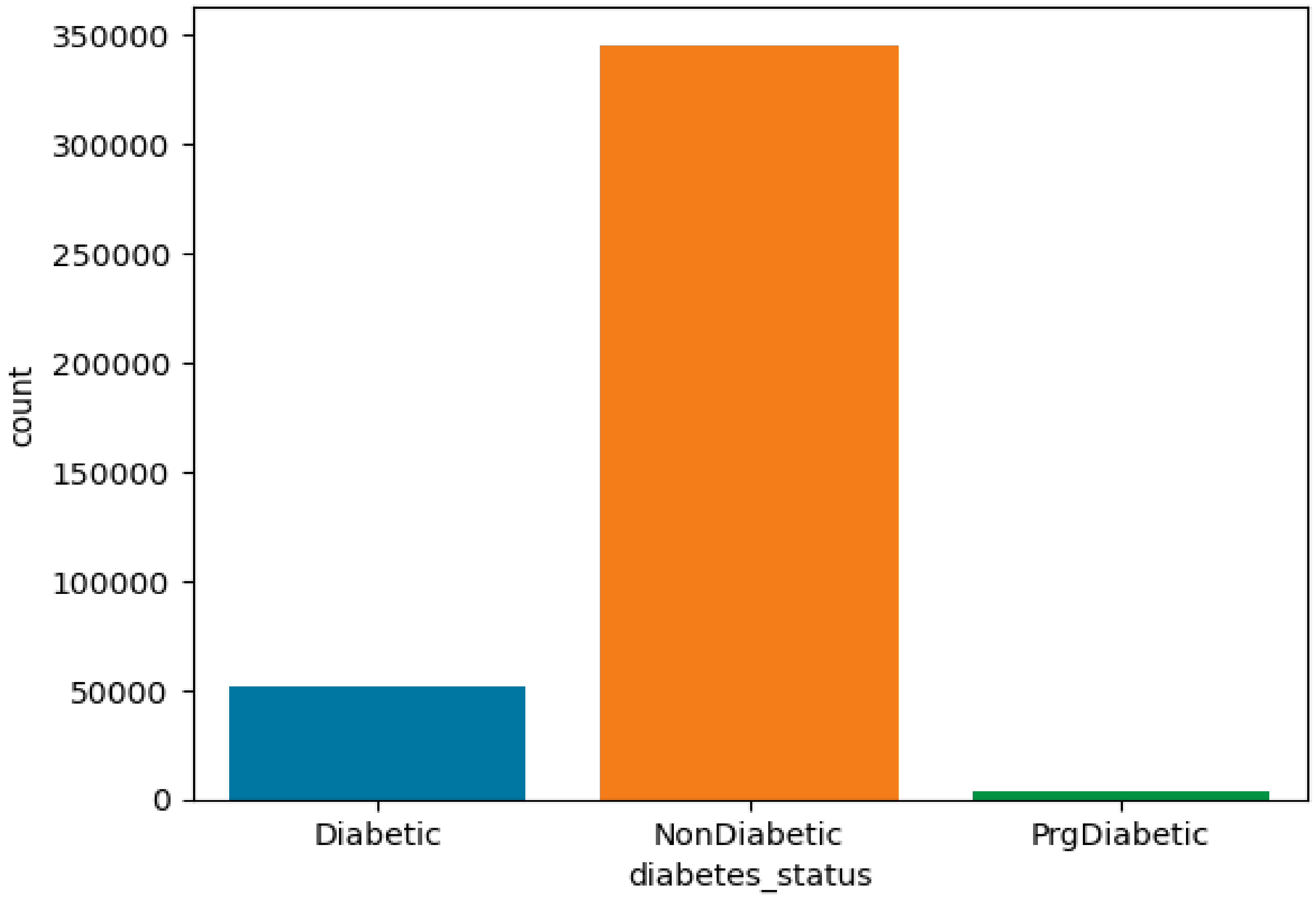




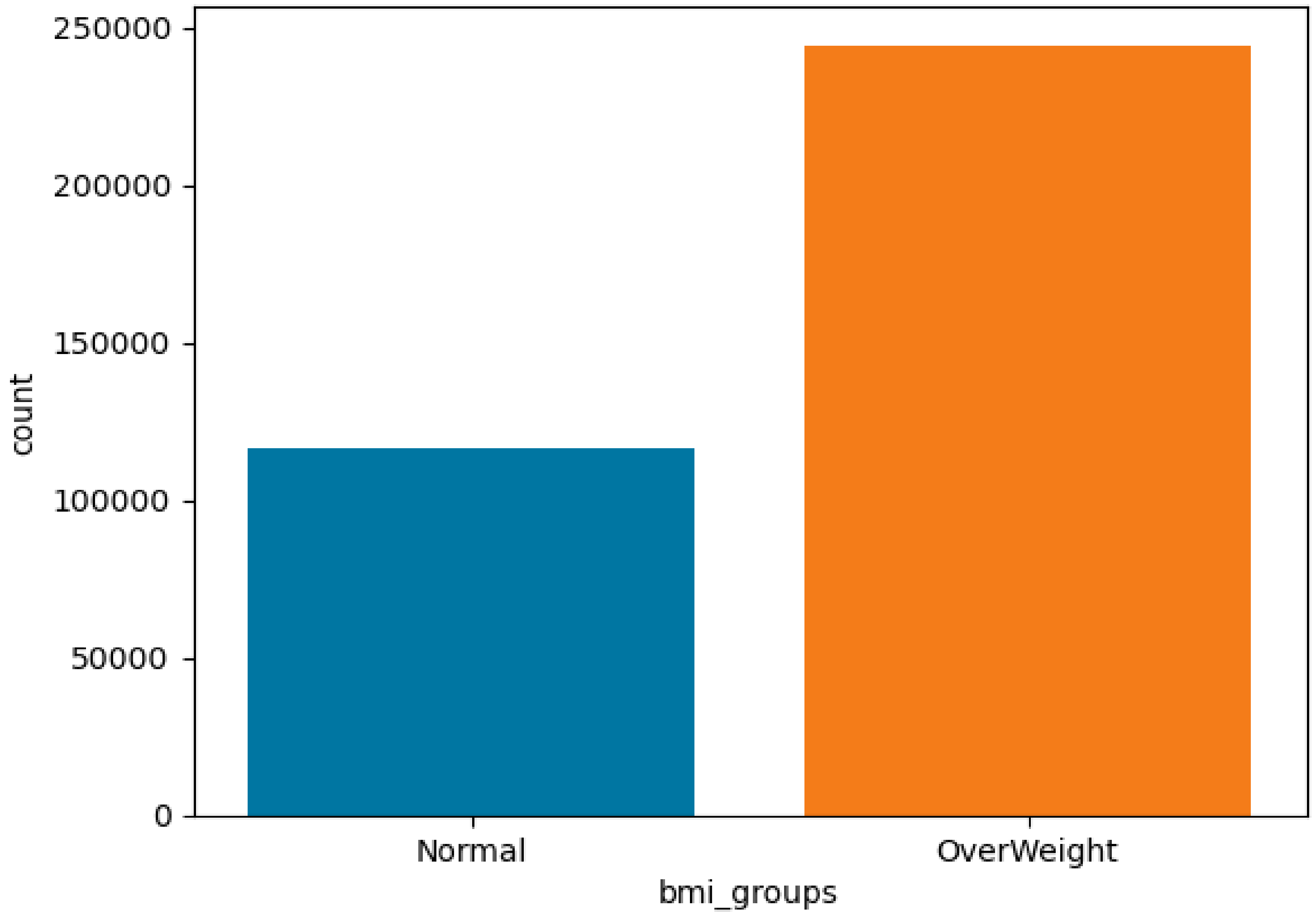
Distribution of Stroke Status



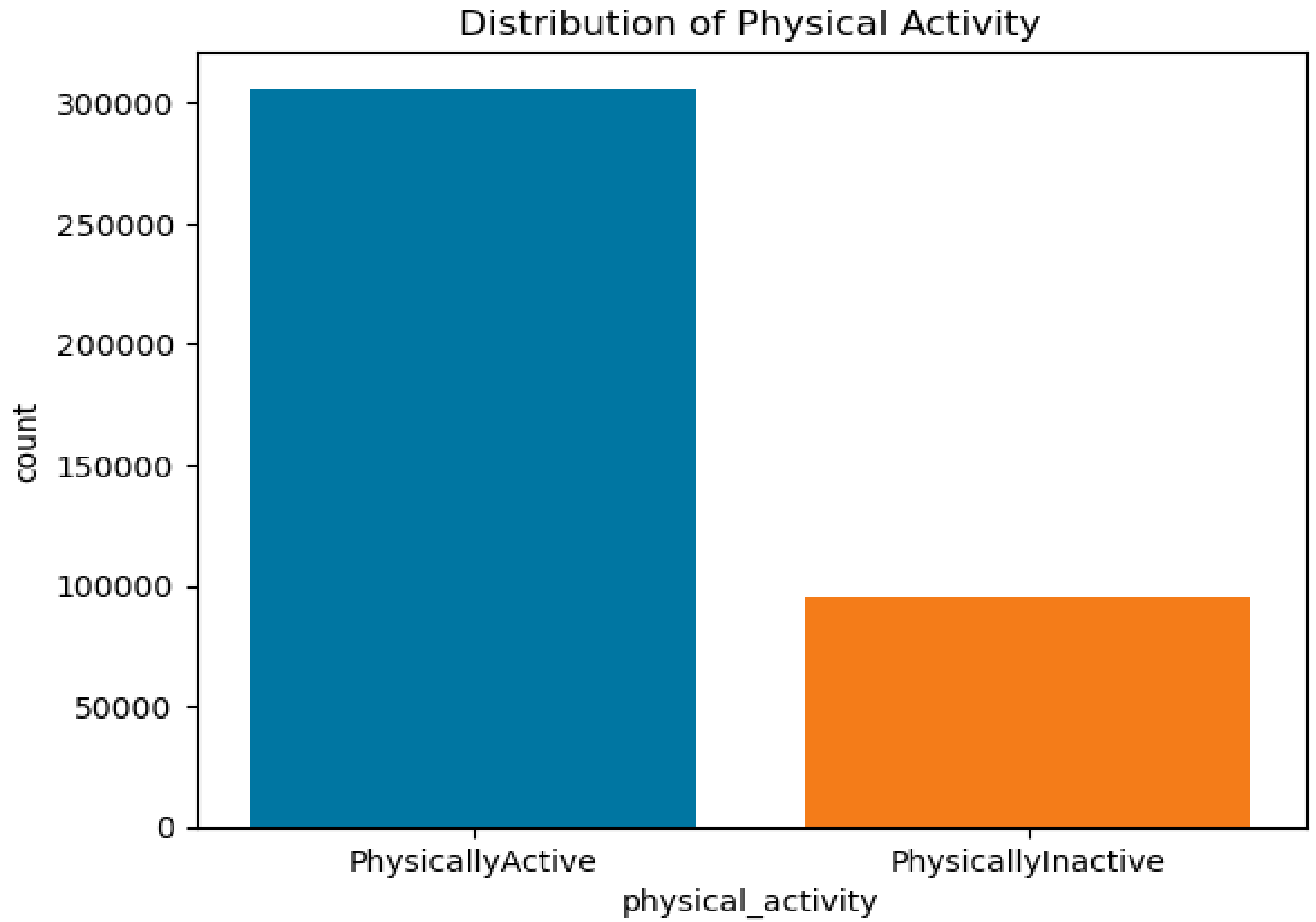
Distribution of Smoking Status

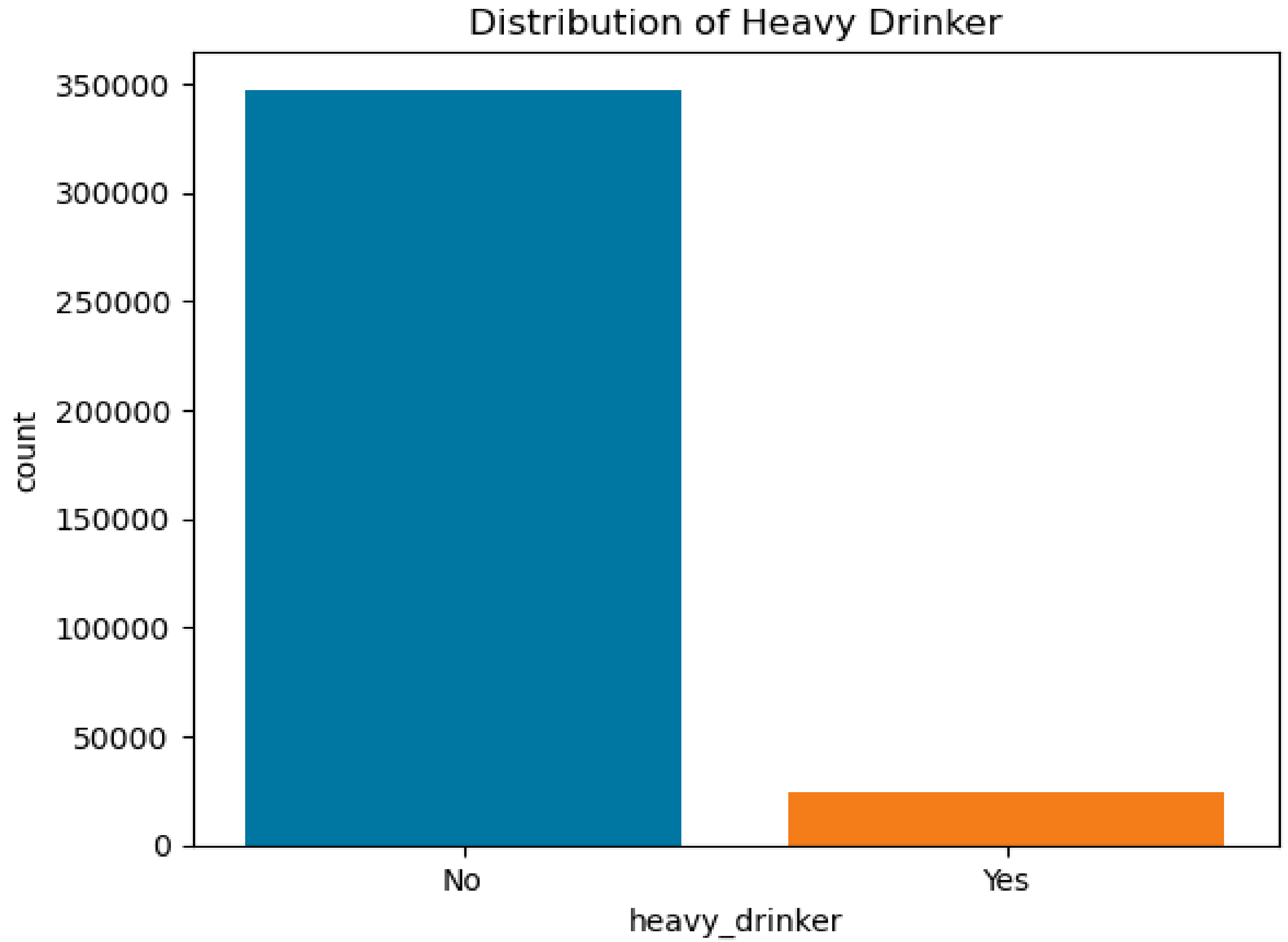


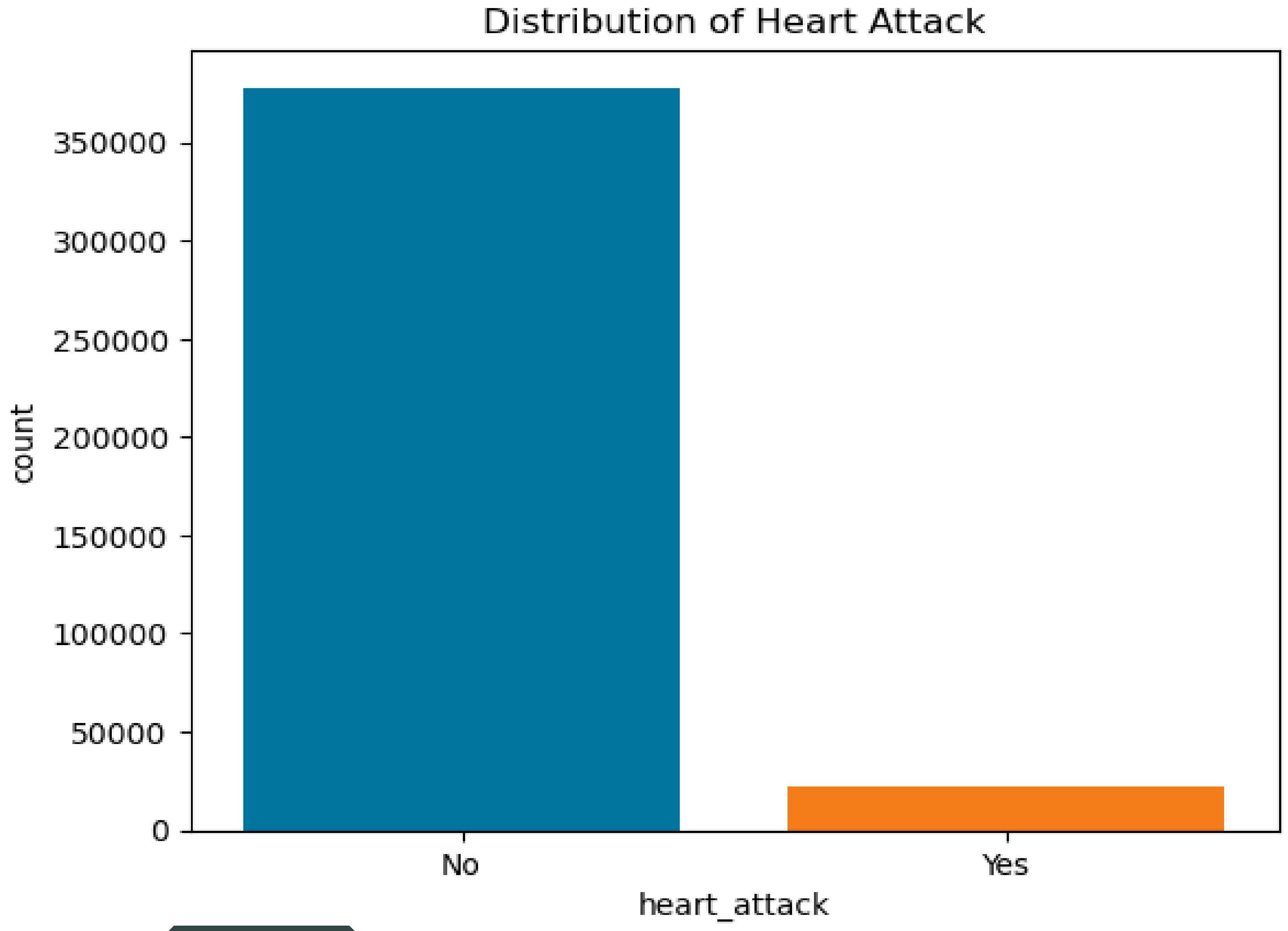
Distribution of Diabetes Status

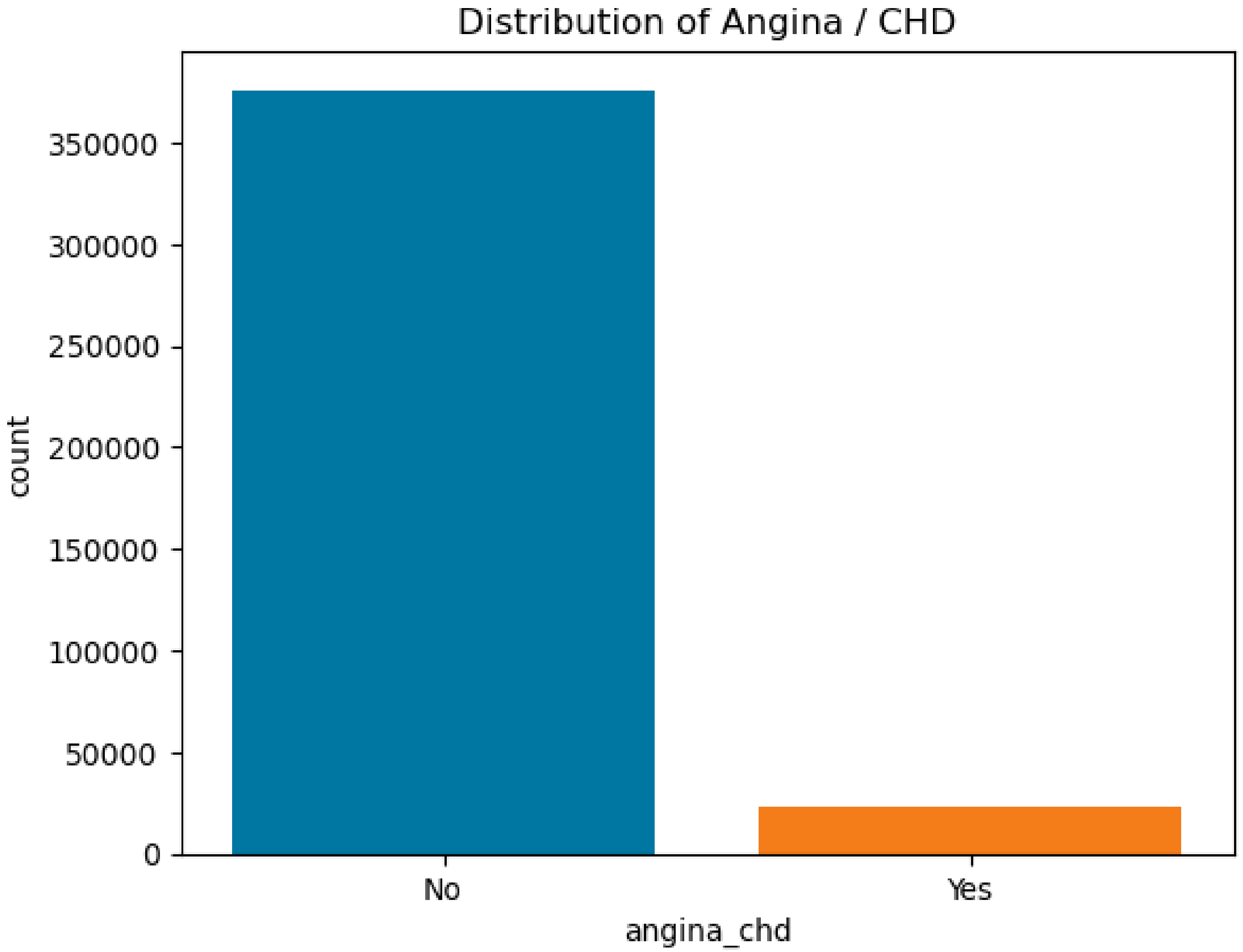


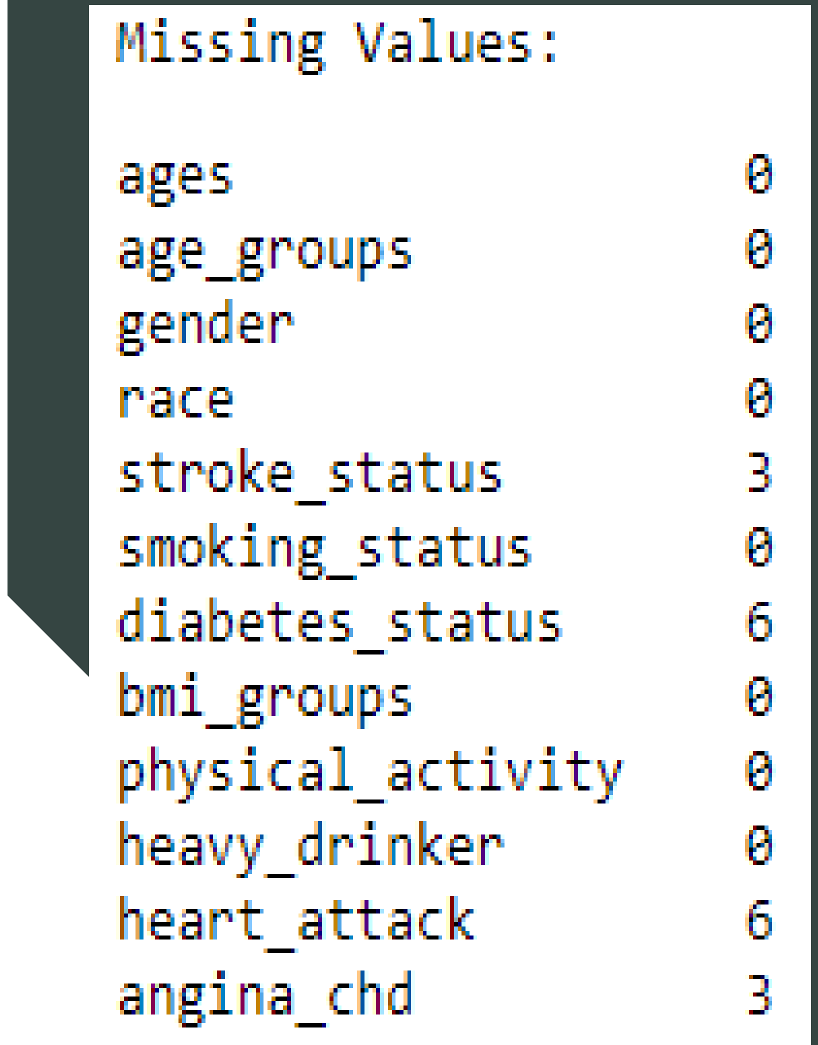
Distribution of BMI Groups







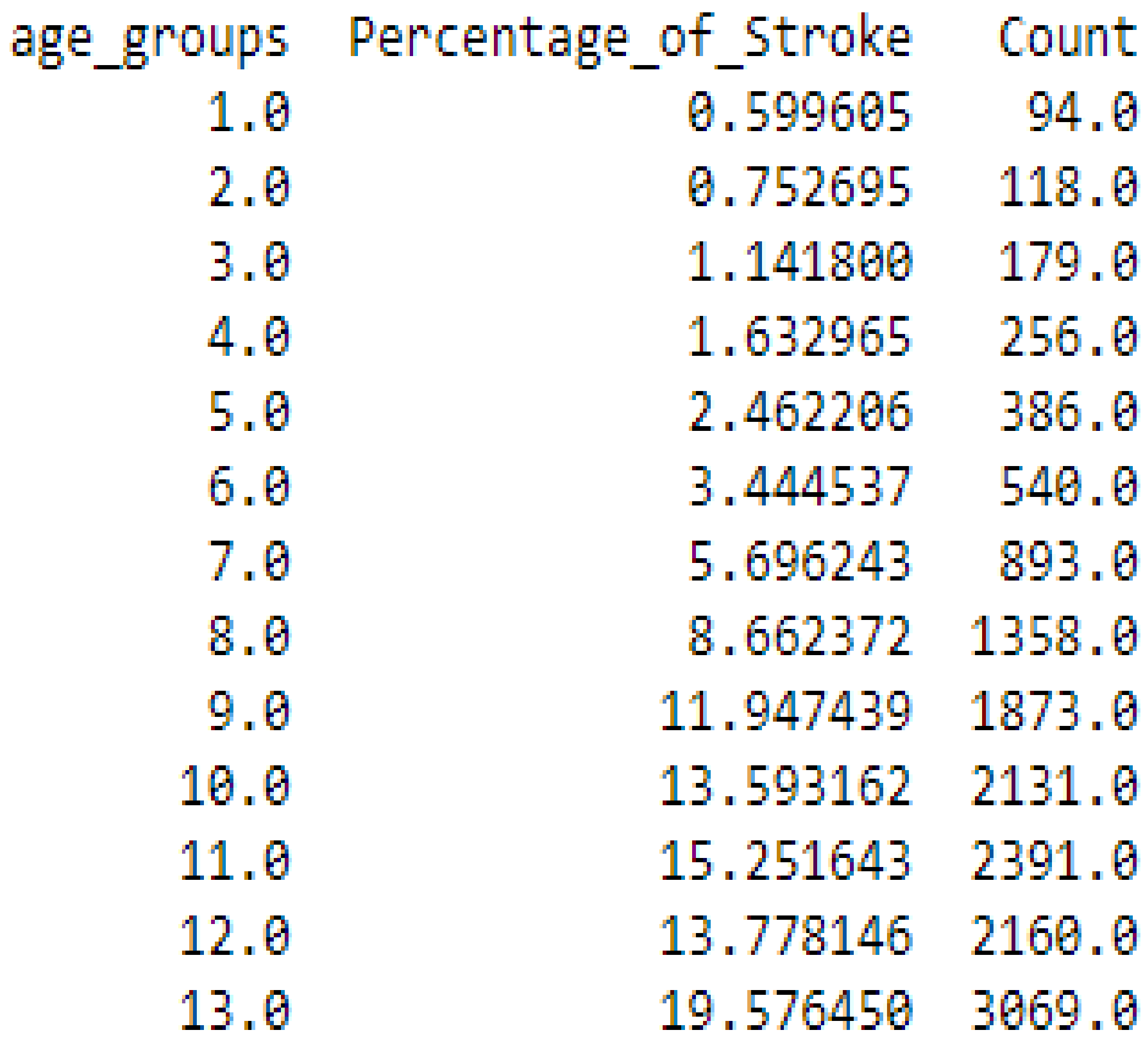




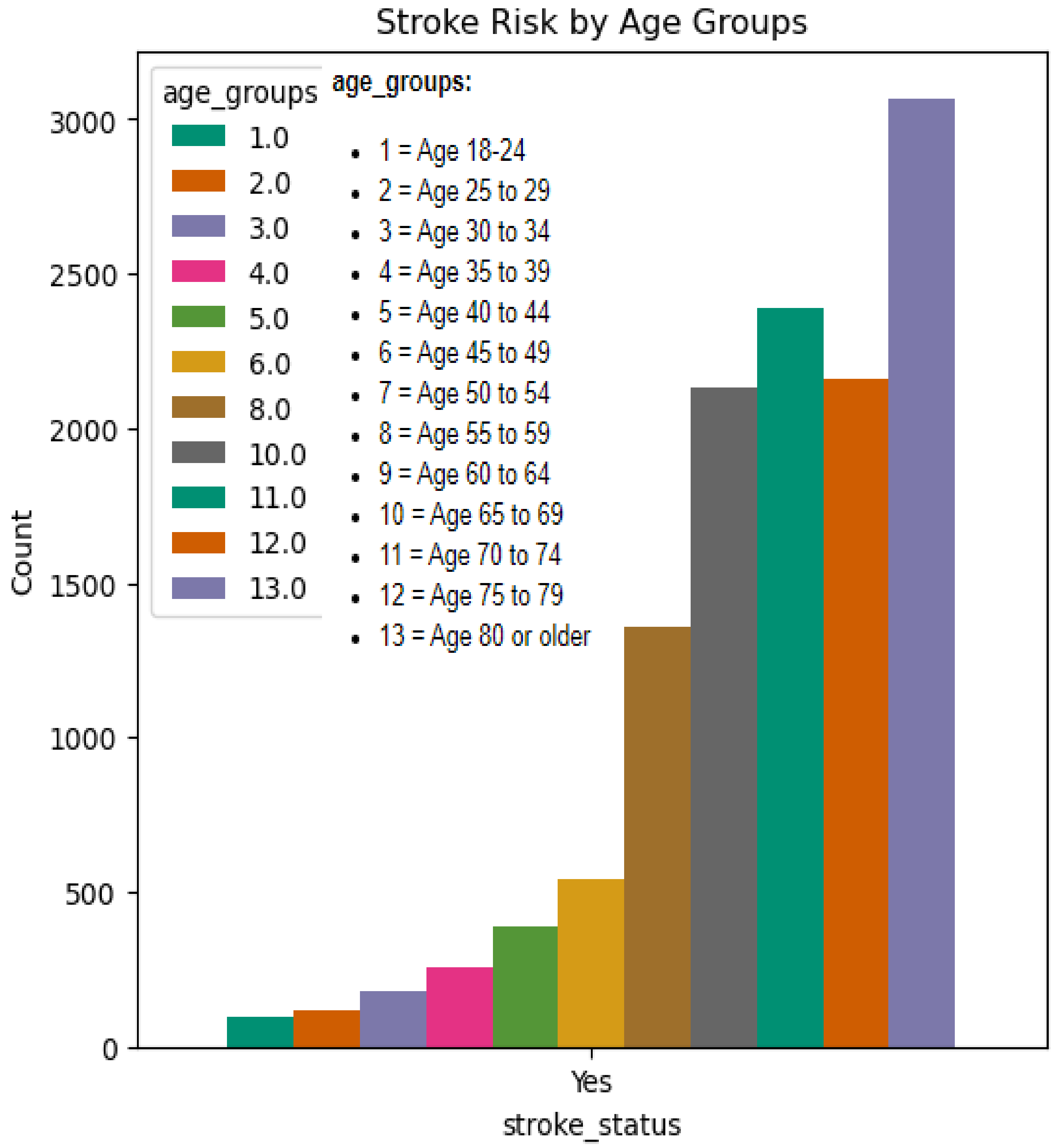
```
Missing Values:

ages                 0
age_groups           0
gender               0
race                 0
stroke_status        3
smoking_status       0
diabetes_status      6
bmi_groups           0
physical_activity    0
heavy_drinker        0
heart_attack         6
angina_chd           3
```

EFFECT OF NON-MODIFIABLE RISK FACTORS ON INDIVIDUALS WITH A HISTORY OF STROKE

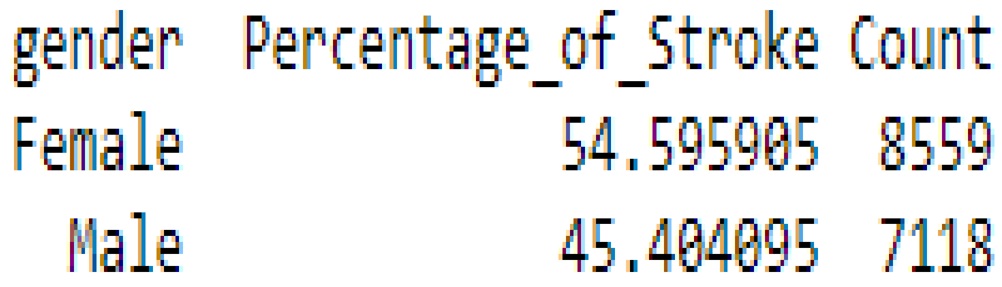


```
age_groups  Percentage_of_Stroke   Count
       1.0              0.599605    94.0
       2.0              0.752695   118.0
       3.0              1.141800   179.0
       4.0              1.632965   256.0
       5.0              2.462206   386.0
       6.0              3.444537   540.0
       7.0              5.696243   893.0
       8.0              8.662372  1358.0
       9.0             11.947439  1873.0
      10.0             13.593162  2131.0
      11.0             15.251643  2391.0
      12.0             13.778146  2160.0
      13.0             19.576450  3069.0
```

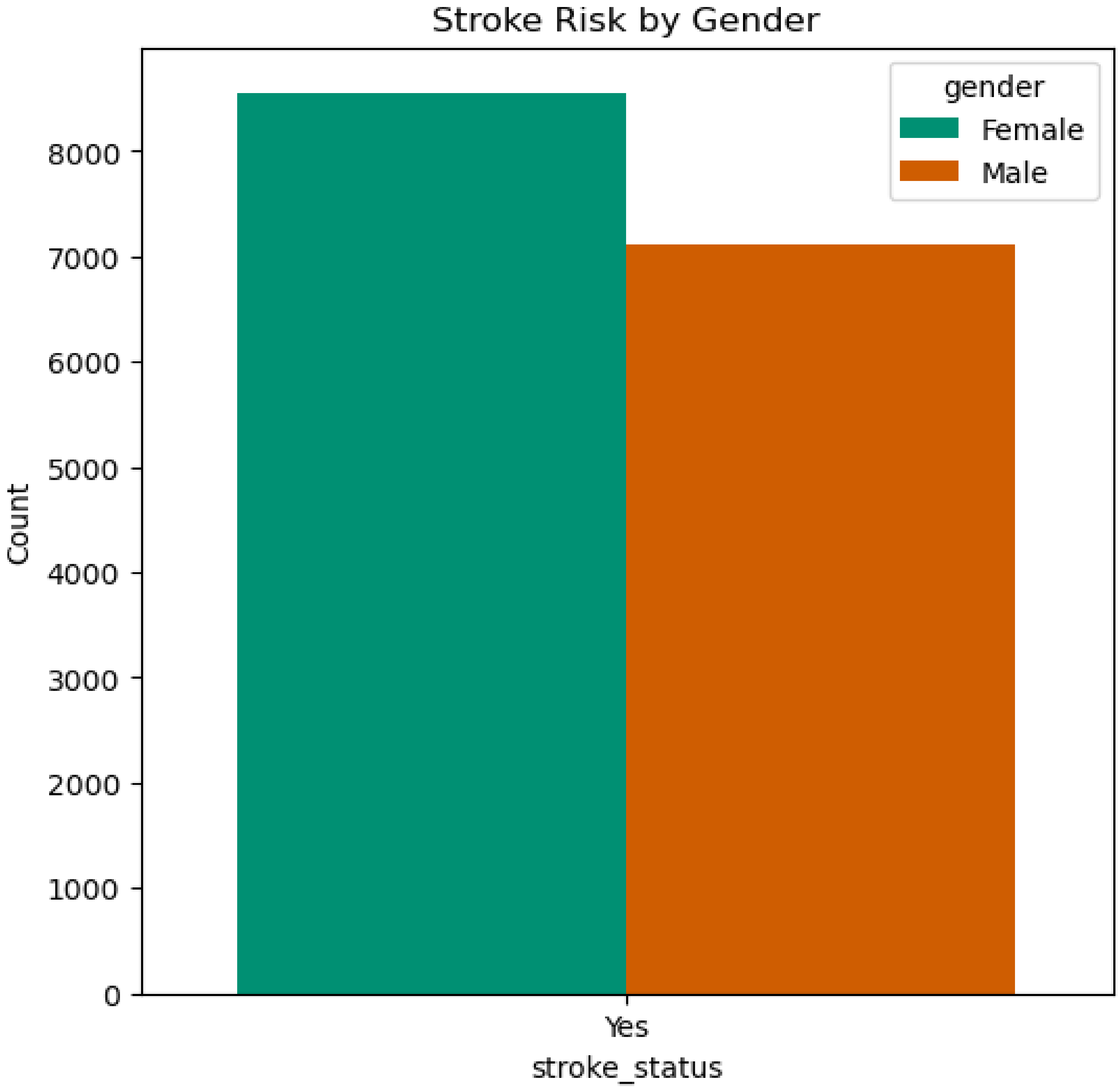


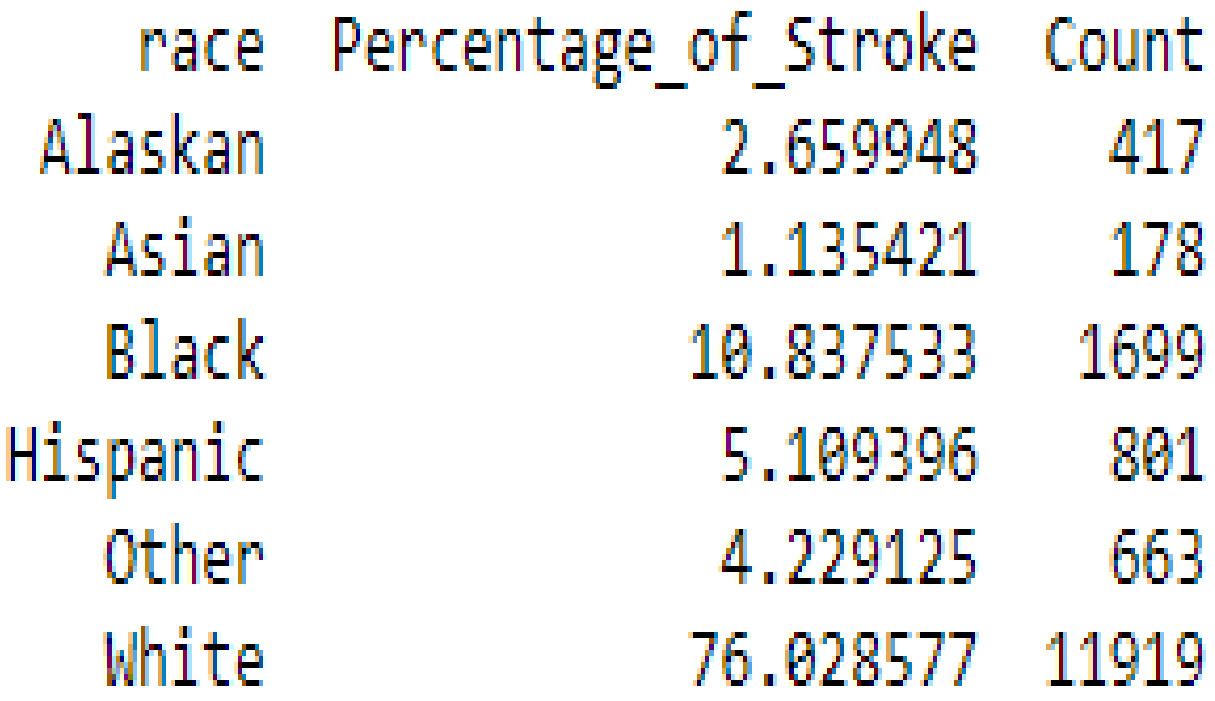
age_groups:

- 1 = Age 18-24
- 2 = Age 25 to 29
- 3 = Age 30 to 34
- 4 = Age 35 to 39
- 5 = Age 40 to 44
- 6 = Age 45 to 49
- 7 = Age 50 to 54
- 8 = Age 55 to 59
- 9 = Age 60 to 64
- 10 = Age 65 to 69
- 11 = Age 70 to 74
- 12 = Age 75 to 79
- 13 = Age 80 or older

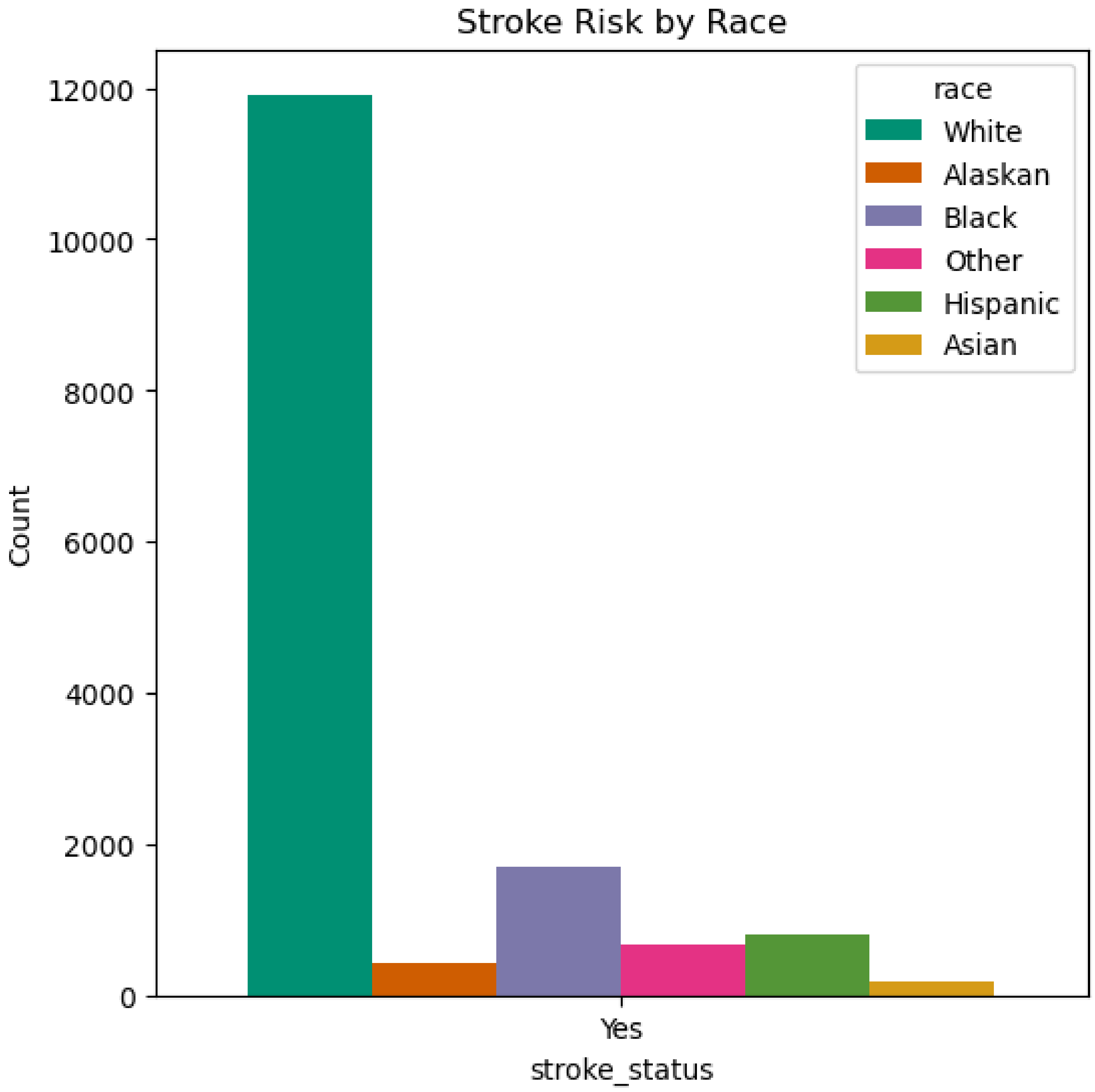


```
gender  Percentage_of_Stroke Count
Female             54.595905  8559
  Male             45.404095  7118
```

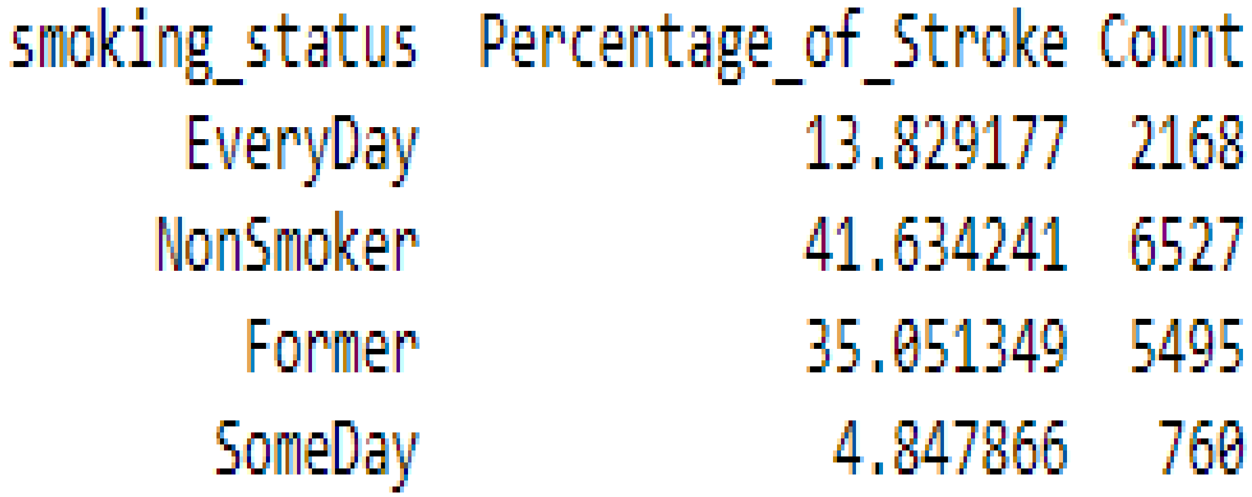




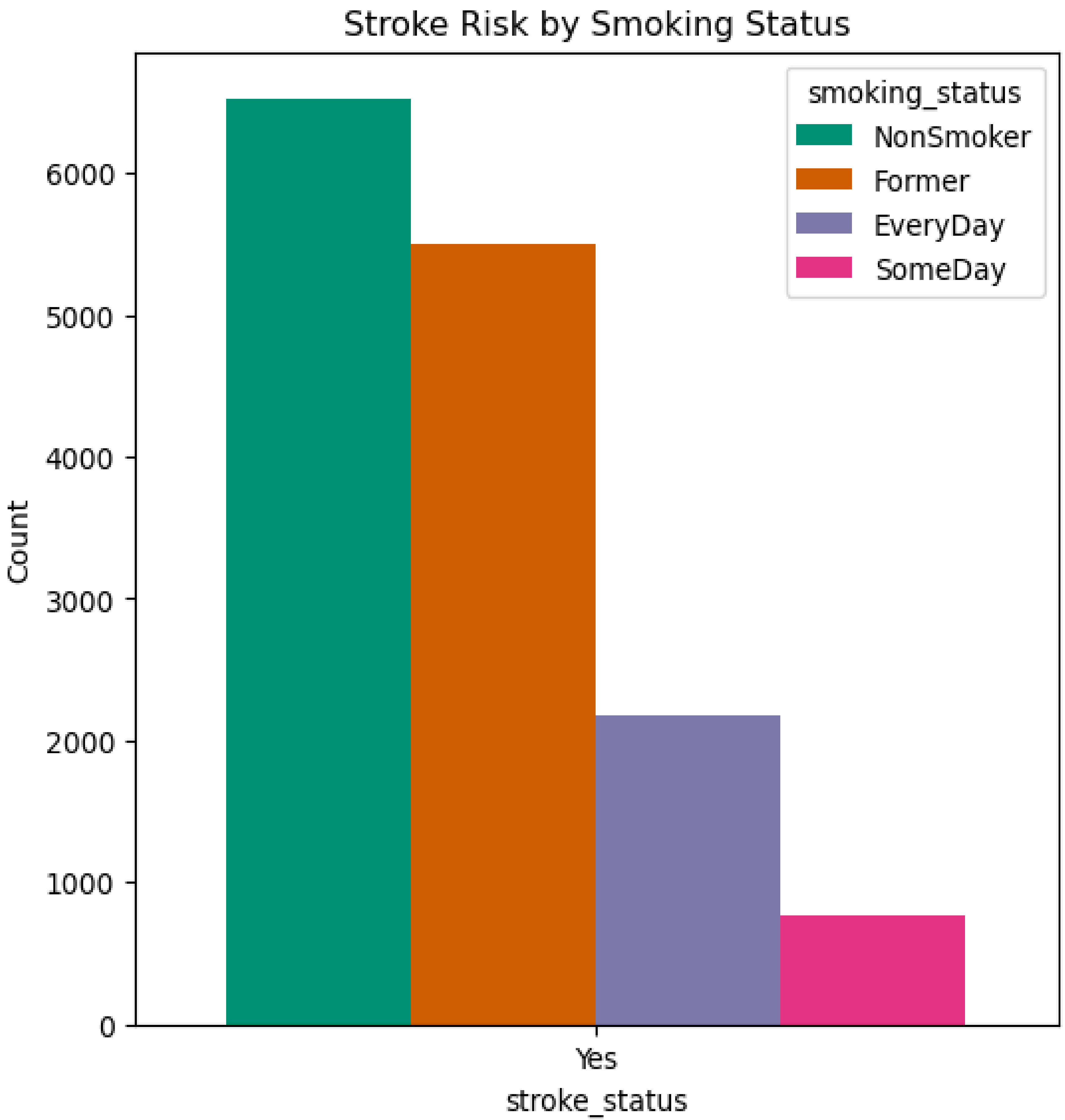
```
     race  Percentage_of_Stroke  Count
  Alaskan              2.659948    417
    Asian              1.135421    178
    Black             10.837533   1699
 Hispanic              5.109396    801
    Other              4.229125    663
    White             76.028577  11919
```

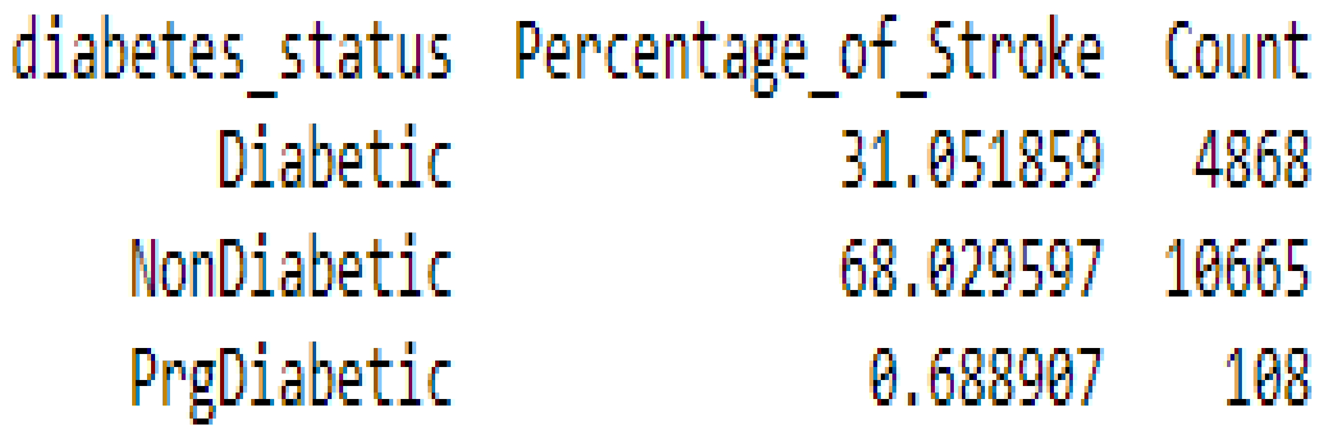


EFFECT OF MODIFIABLE RISK FACTORS ON INDIVIDUALS WITH A HISTORY OF STROKE

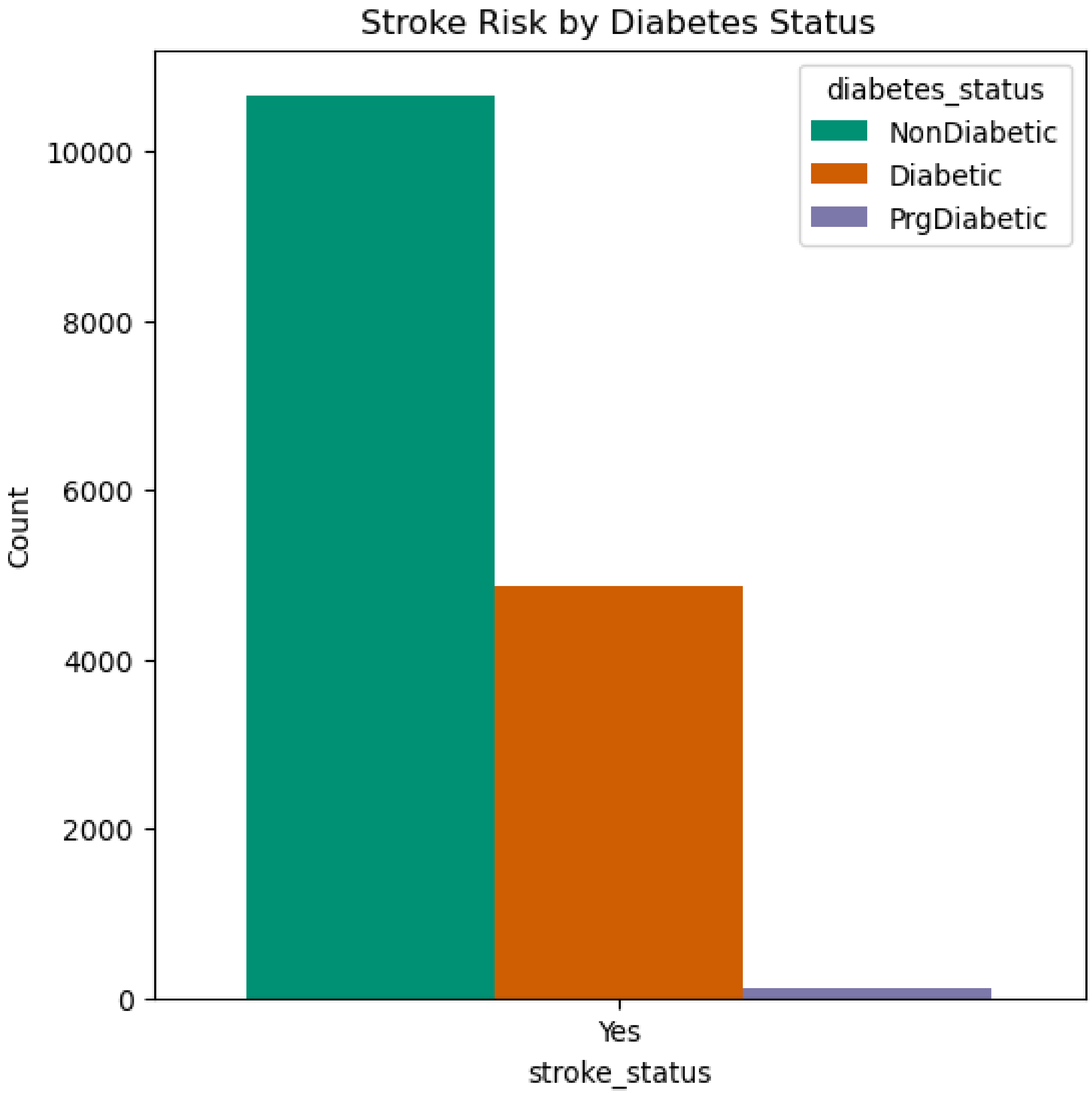


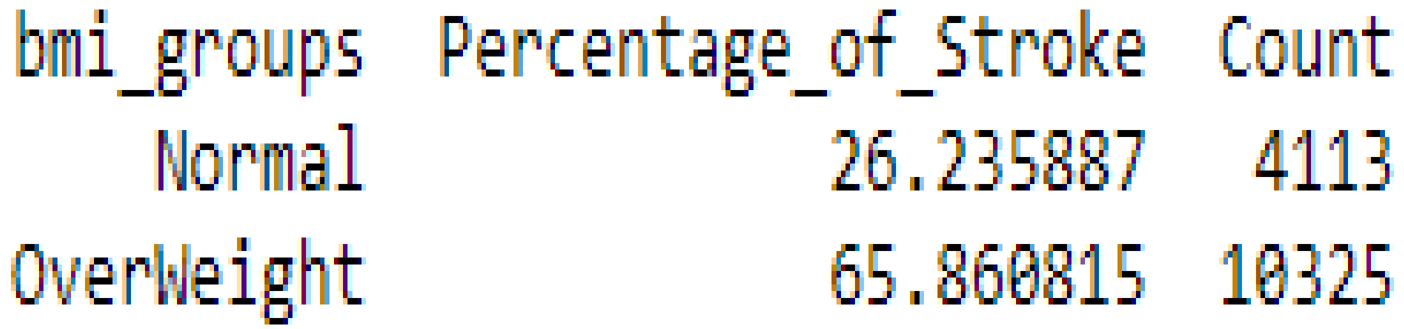
```
smoking_status  Percentage_of_Stroke Count
      EveryDay             13.829177  2168
     NonSmoker             41.634241  6527
        Former             35.051349  5495
       SomeDay              4.847866   760
```



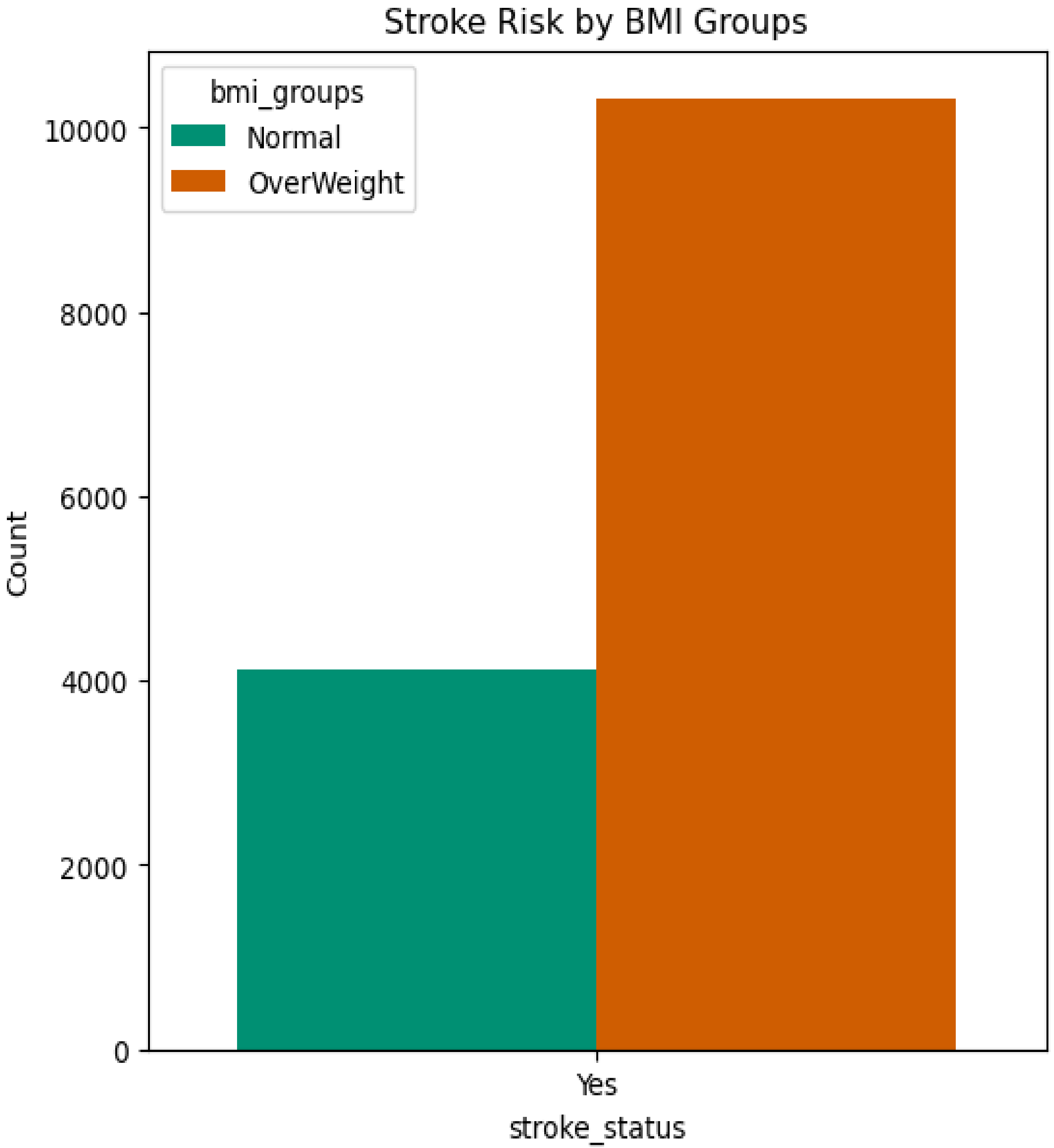


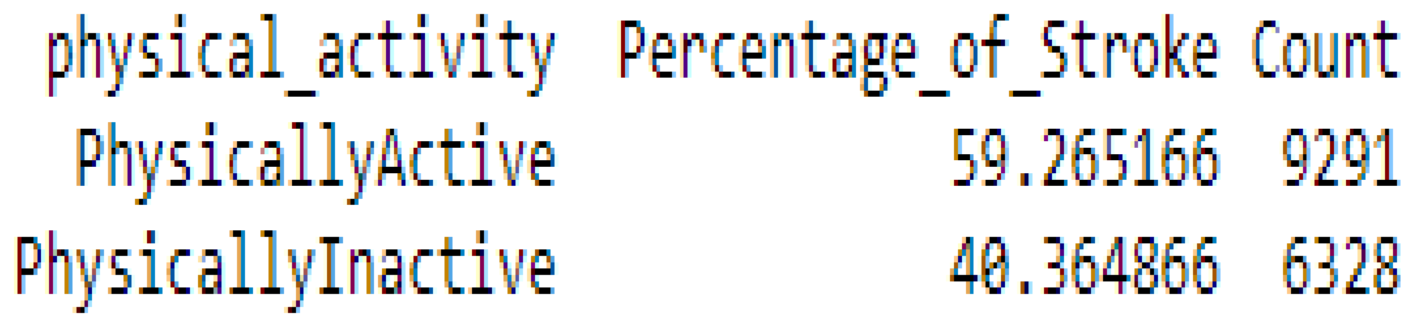
```
diabetes_status  Percentage_of_Stroke  Count
       Diabetic             31.051859   4868
    NonDiabetic             68.029597  10665
    PrgDiabetic              0.688907    108
```



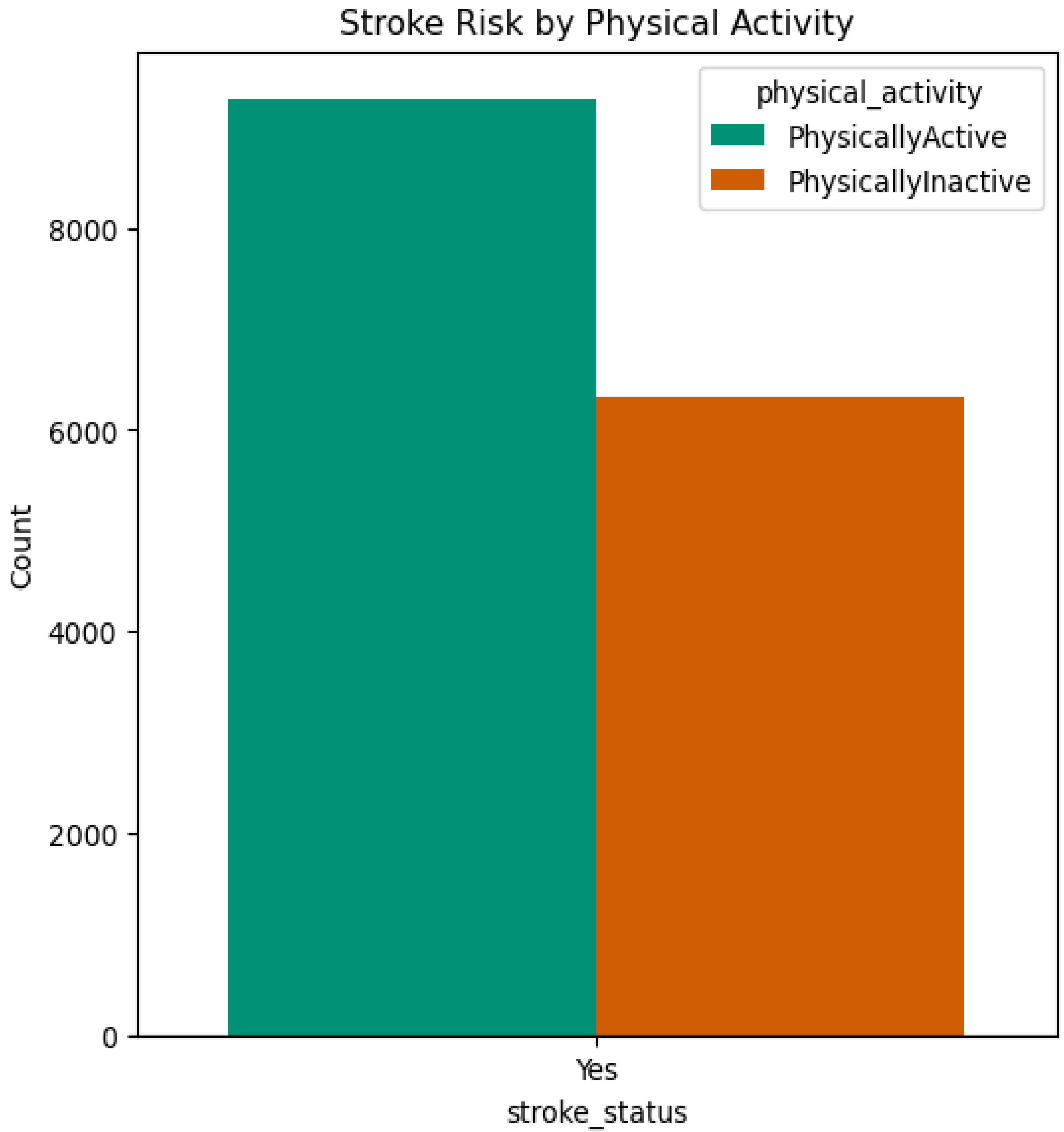


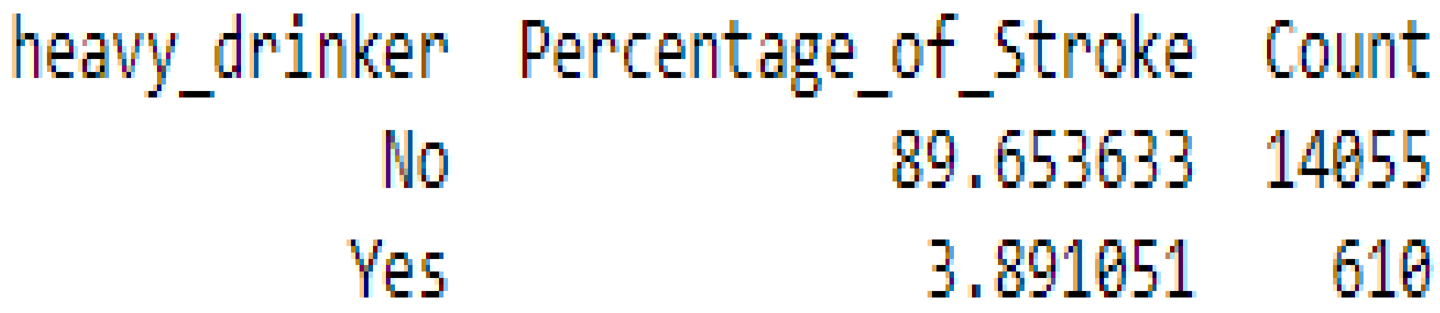
```
bmi_groups  Percentage_of_Stroke  Count
    Normal             26.235887   4113
OverWeight             65.860815  10325
```



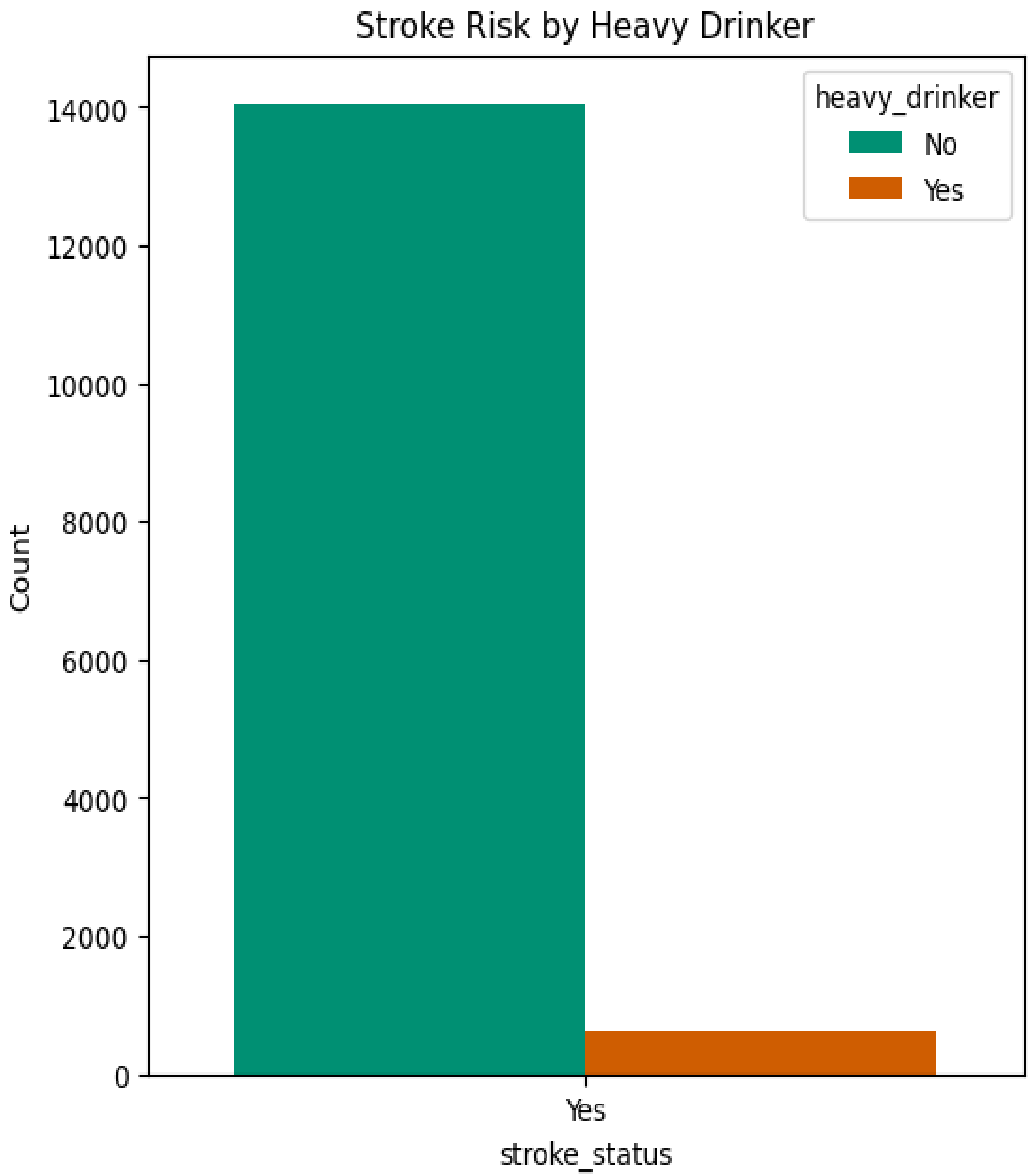


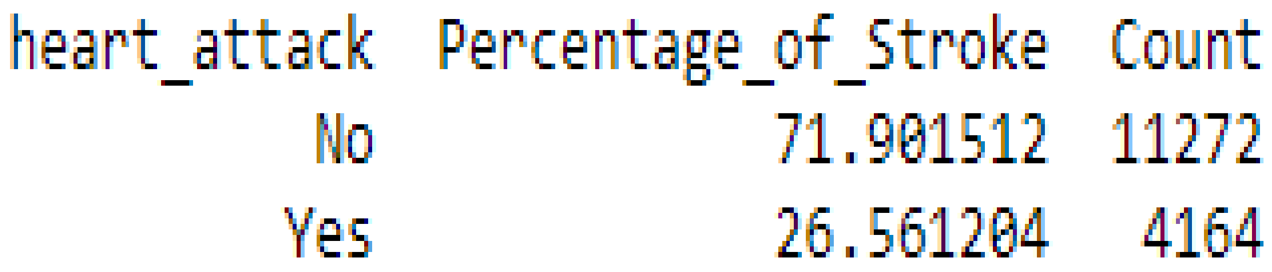
```
  physical_activity  Percentage_of_Stroke Count
    PhysicallyActive             59.265166  9291
  PhysicallyInactive             40.364866  6328
```



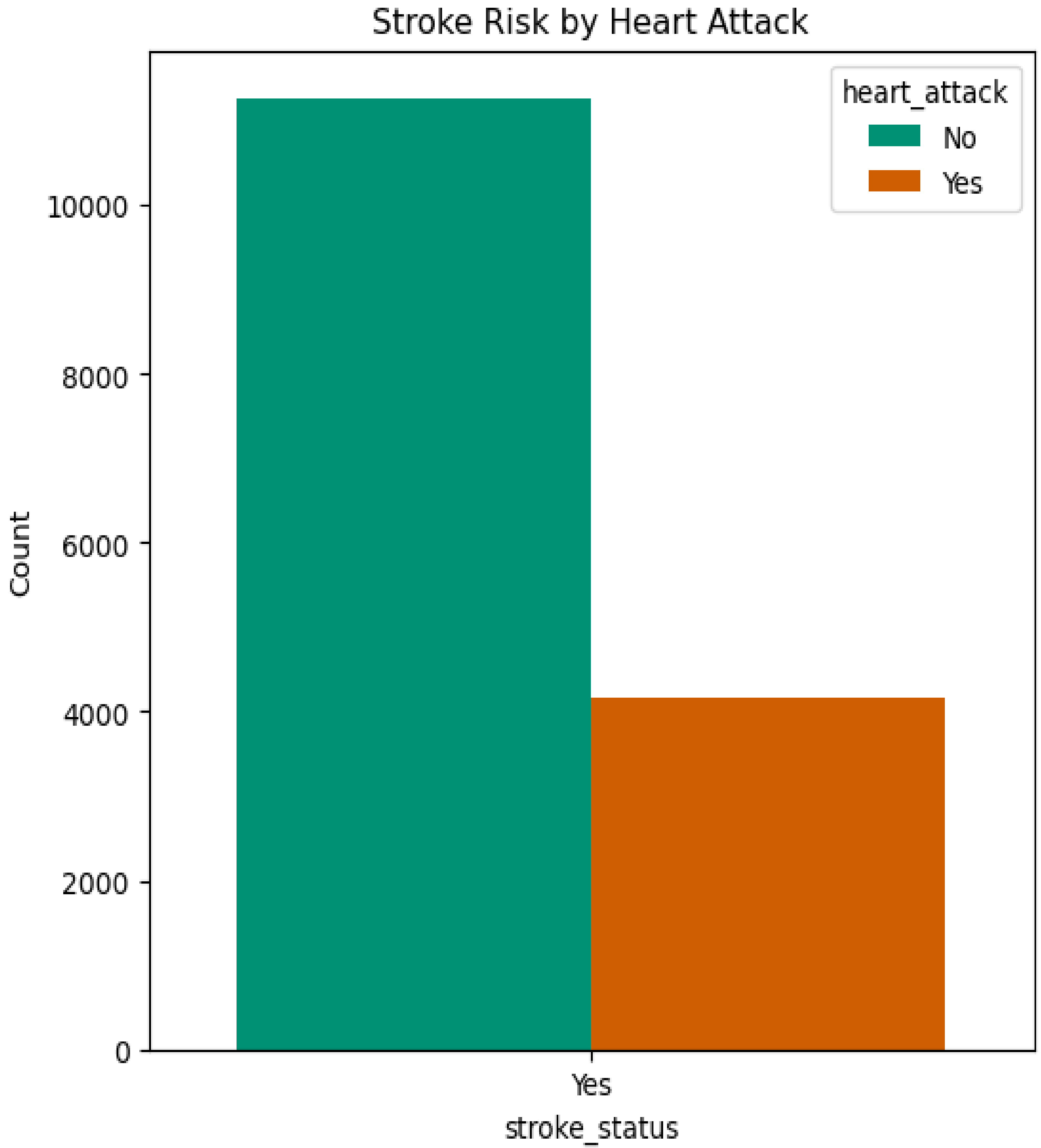


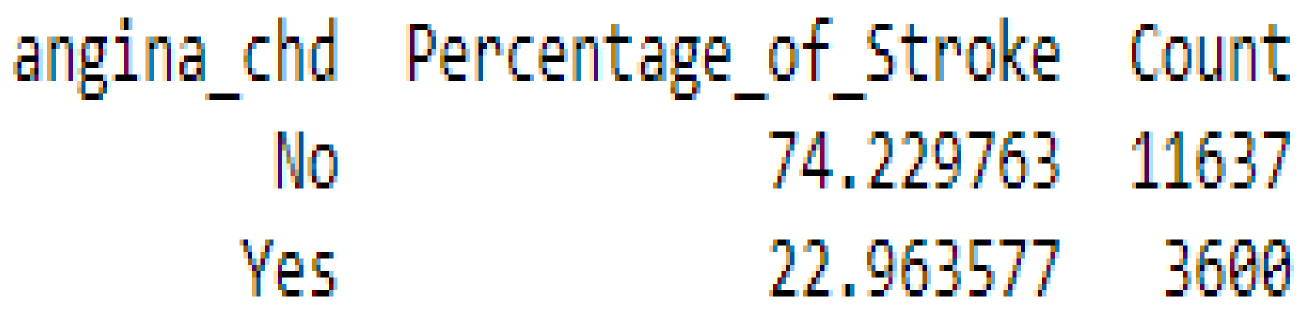
```
heavy_drinker  Percentage_of_Stroke  Count
           No             89.653633  14055
          Yes              3.891051    610
```



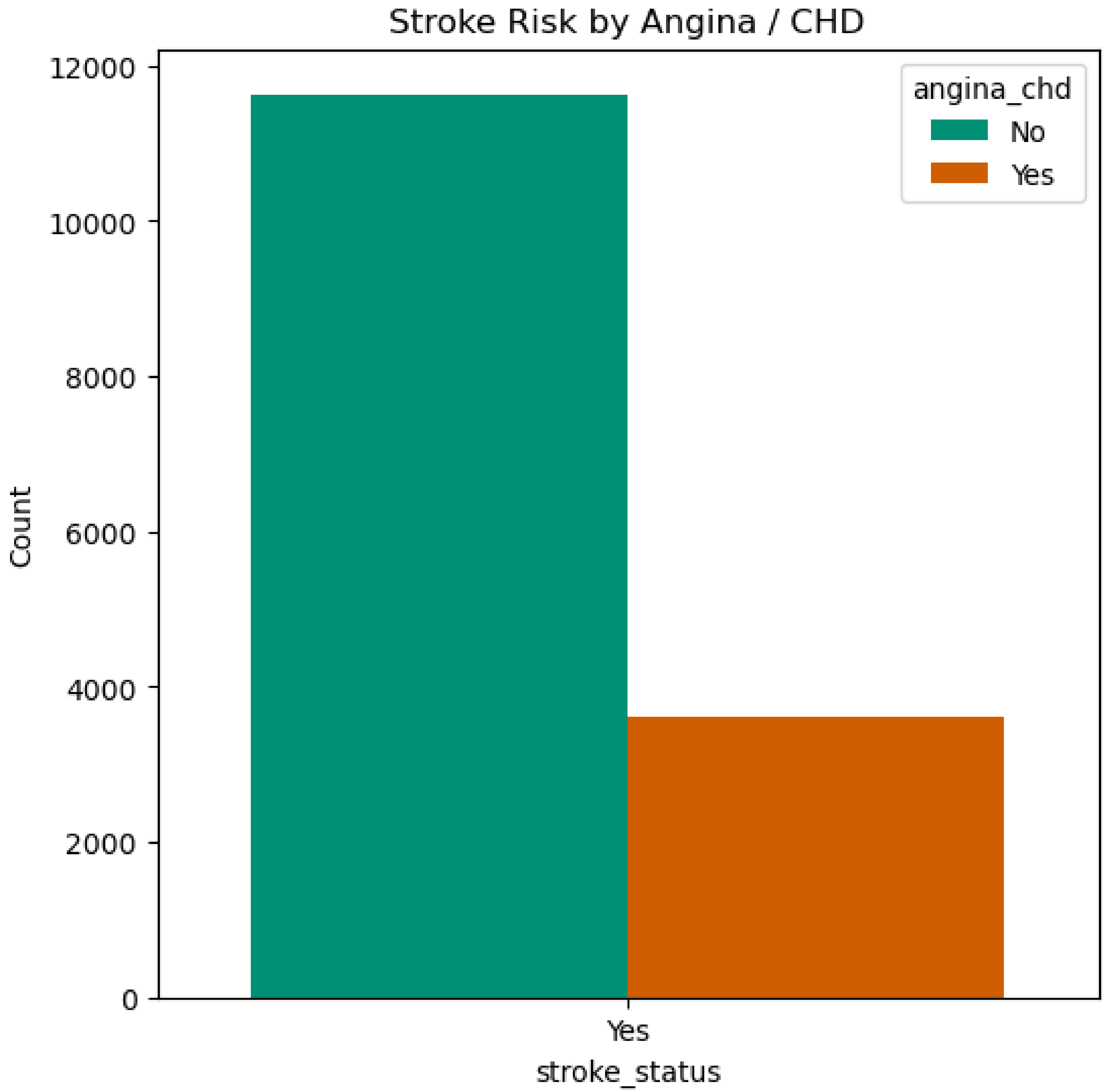


```
heart_attack  Percentage_of_Stroke  Count
          No             71.901512  11272
         Yes             26.561204   4164
```





```
angina_chd  Percentage_of_Stroke  Count
        No             74.229763  11637
       Yes             22.963577   3600
```



HYPOTHESIS TESTING

METHOD

REASON

CHI SQUARE

A statistical test used to determine if there is a significant association or relationship between two categorical variables in a dataset.

Since all are variables of analysis are categorical therefore we have used chi square test to test the significance relationship between two variables

IS THERE A RELATIONSHIP BETWEEN SMOKING STATUS AND THE OCCURRENCE OF HEART ATTACKS?

HYPOTHESIS 1

NULL - HYPOTHESIS (Ho)

Null Hypothesis (H0): There is no relationship between Smoking and Occurance of Heart Attack.

ALTERNATE - HYPOTHESIS (Ha)

Alternate Hypothesis (Ha): There is a significant relationship between Smoking and Occurance of Heart Attack.

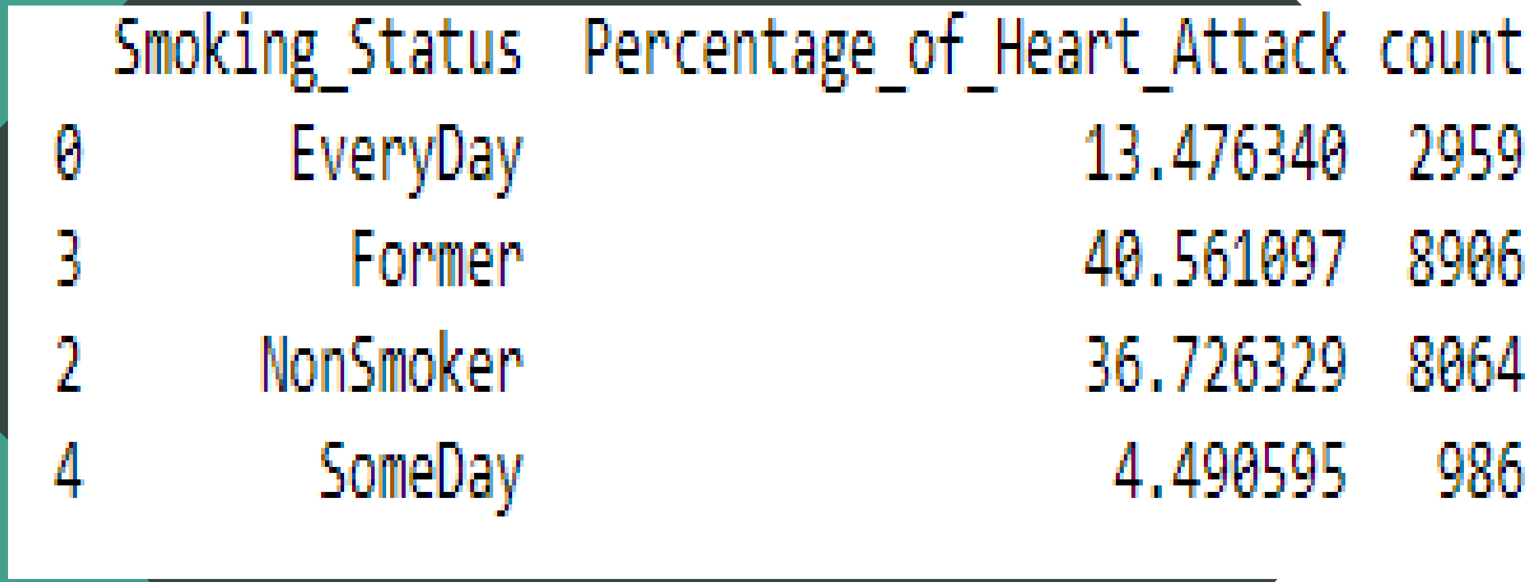
INTERPRETATION

P-VALUE

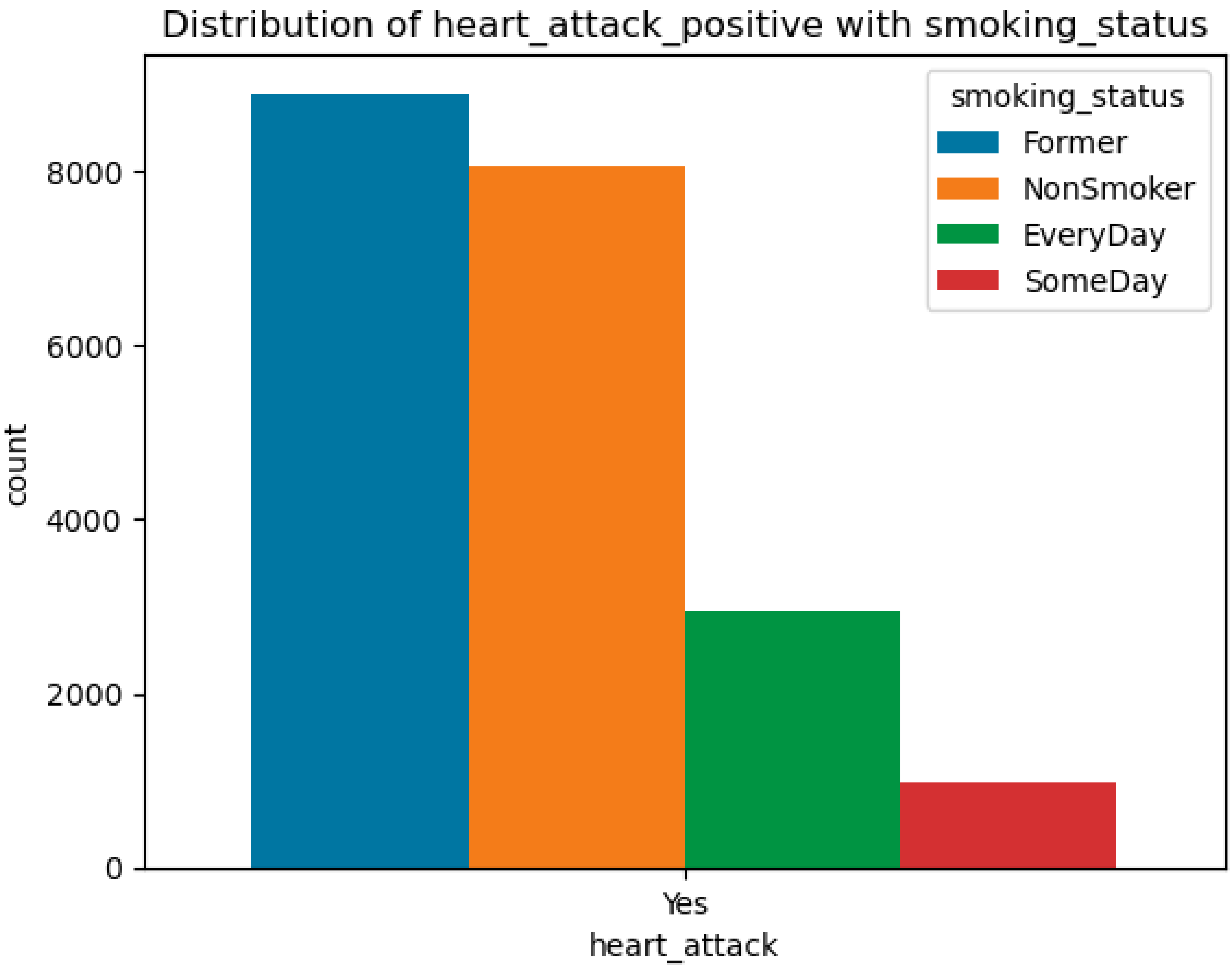
0.0

The very low p-value (p = 0.0) indicates that the association between smoking status and the occurrence of heart attacks is statistically significant.

GRAPHICAL REPRESENTATION



```
  Smoking_Status  Percentage_of_Heart_Attack  count
0       EveryDay                   13.476340   2959
3         Former                   40.561097   8906
2      NonSmoker                   36.726329   8064
4        SomeDay                    4.490595    986
```



Distribution of heart_attack_positive with smoking_status

DOES HIGHER BMI GROUP (OBESE / OVERWEIGHT) SHOW AN INCREASED PREVALENCE OF DIABETES?

HYPOTHESIS 2

NULL - HYPOTHESIS (Ho)

Null Hypothesis (H0): There is no association between Higer BMI group and prevelence of Diabetes.

ALTERNATE - HYPOTHESIS (Ha)

Alternate Hypothesis (Ha): There is a significant association between Higer BMI group and prevelence of Diabetes.

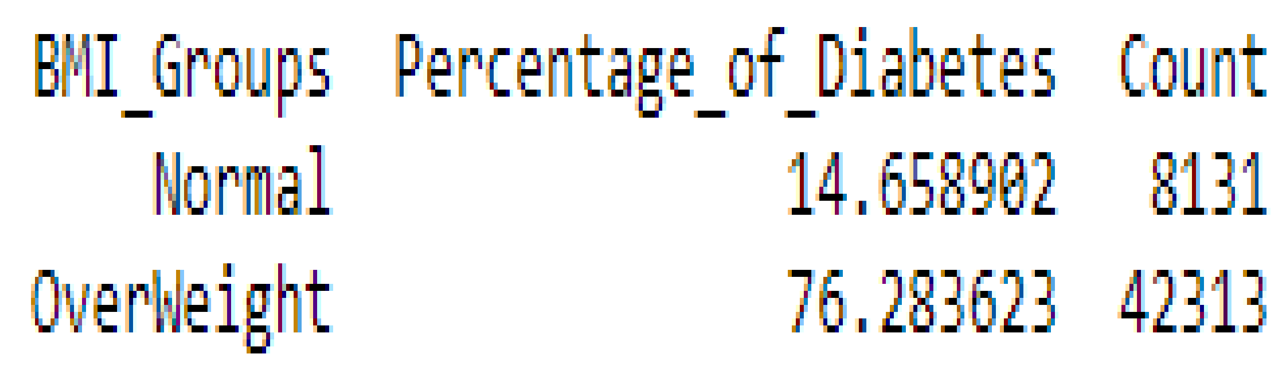
INTERPRETATION

P-VALUE

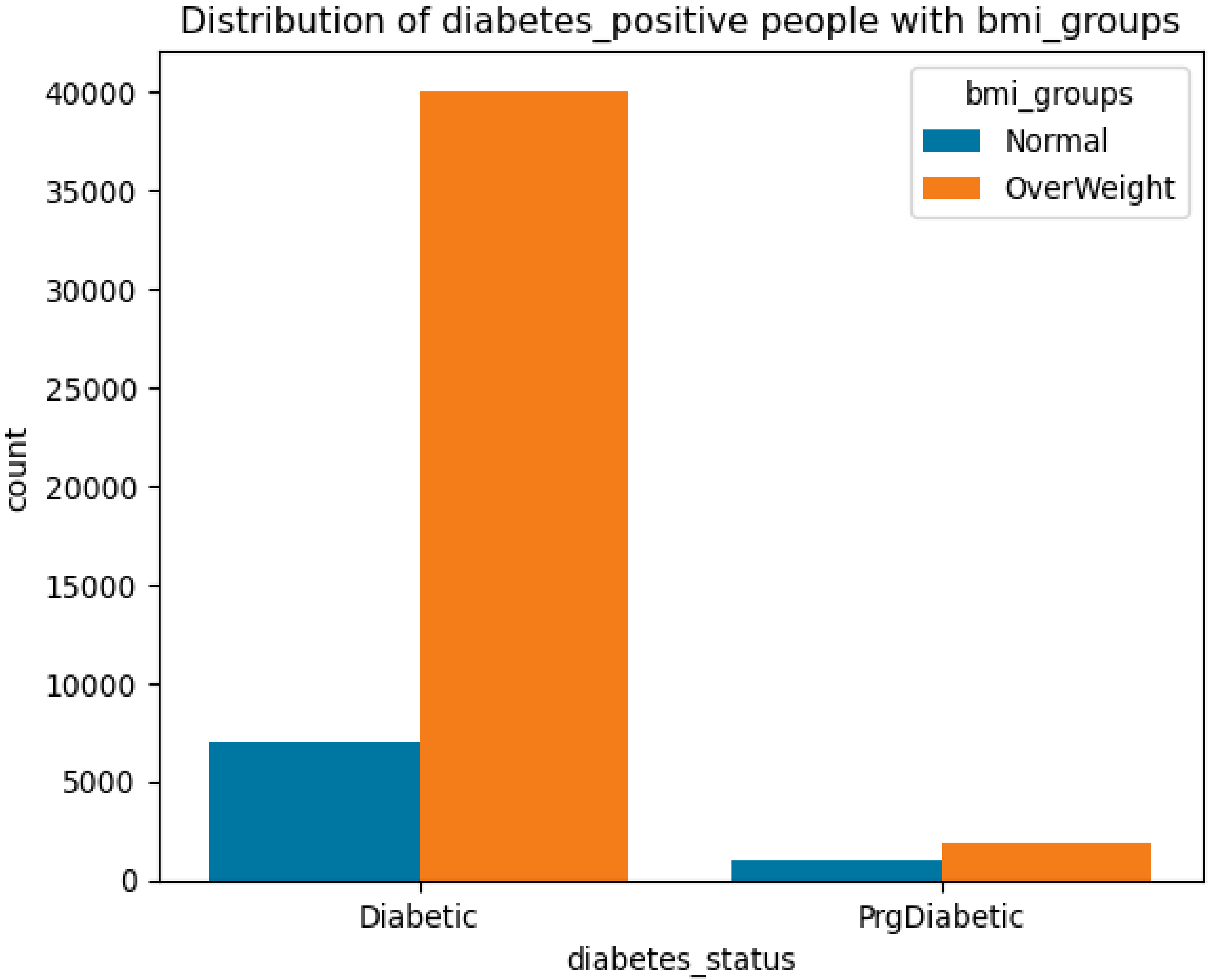
0.0

The very low p-value (p = 0.0) indicates that the association between BMI group (Obese/Overweight) and the prevalence of diabetes is statistically significant.

GRAPHICAL REPRESENTATION



```
BMI_Groups  Percentage_of_Diabetes  Count
    Normal               14.658902   8131
OverWeight               76.283623  42313
```



IS THERE ANY ASSOCIATION BETWEEN PHYSICAL ACTIVITY AND THE LIKELIHOOD OF ANGINA OR CORONARY HEART DISEASE?

HYPOTHESIS 3

NULL - HYPOTHESIS (Ho)

There is no association between physical activity and the likelihood of angina or coronary heart disease?

ALTERNATE - HYPOTHESIS (Ha)

There is some association between physical activity and the likelihood of angina or coronary heart disease?

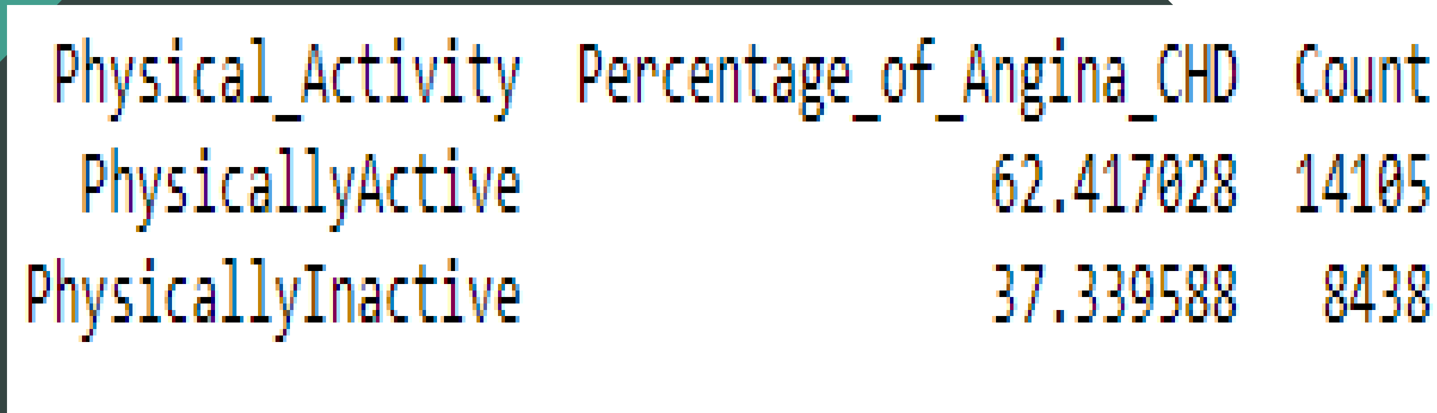
INTERPRETATION

P-VALUE

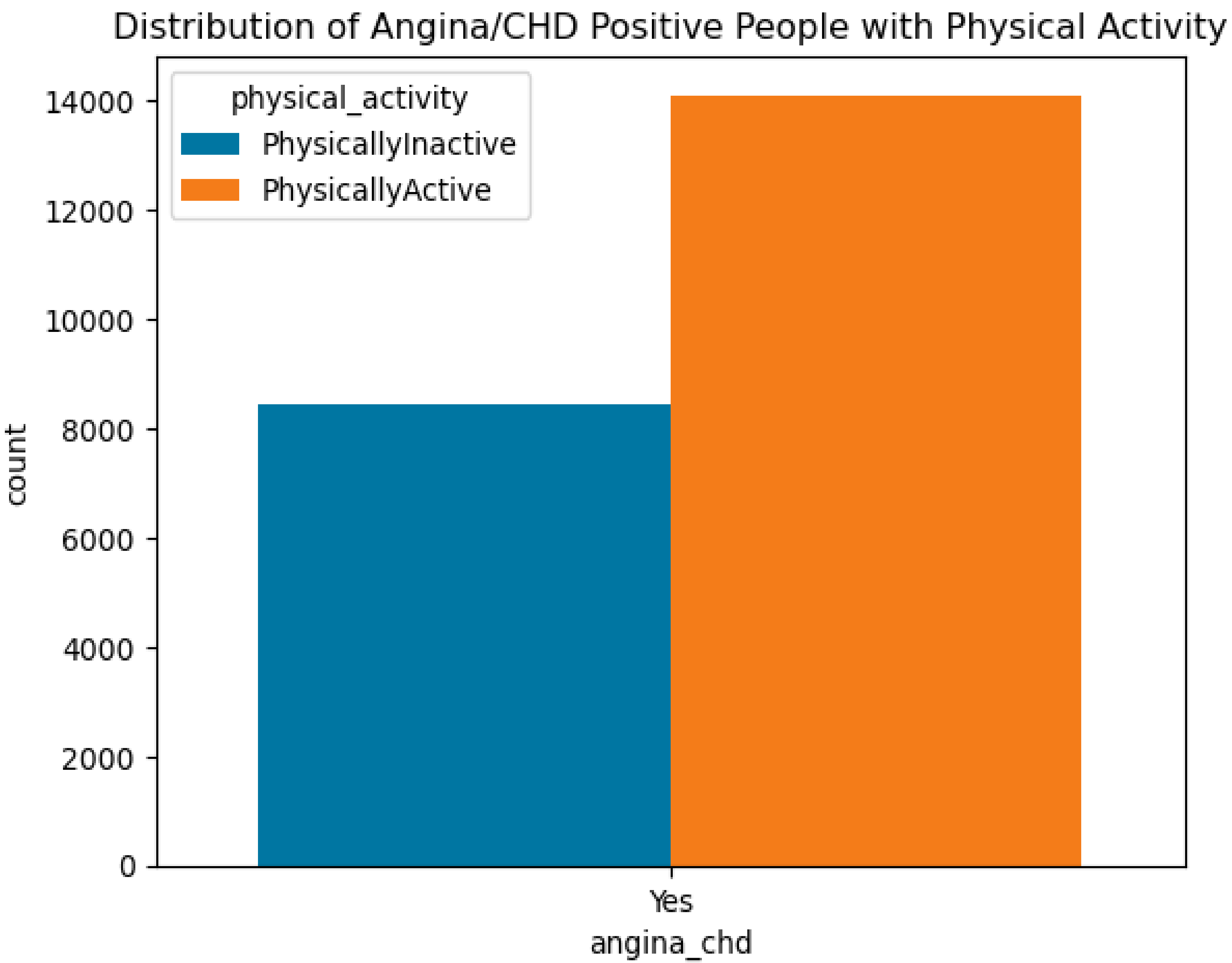
0.0

The very low p-value (p = 0.0) indicates that the association between physical activity and the likelihood of angina or coronary heart disease is statistically significant.

GRAPHICAL REPRESENTATION



```
 Physical_Activity  Percentage_of_Angina_CHD  Count
  PhysicallyActive                 62.417028  14105
PhysicallyInactive                 37.339588   8438
```



CONCLUSION

"The implications of these findings underscore the importance of identifying and addressing these risk factors, as they are crucial in the development of effective stroke prevention strategies. Addressing modifiable risk factors becomes a key focus for improving public health outcomes related to stroke."

REGRESSION ANALYSIS

METHOD

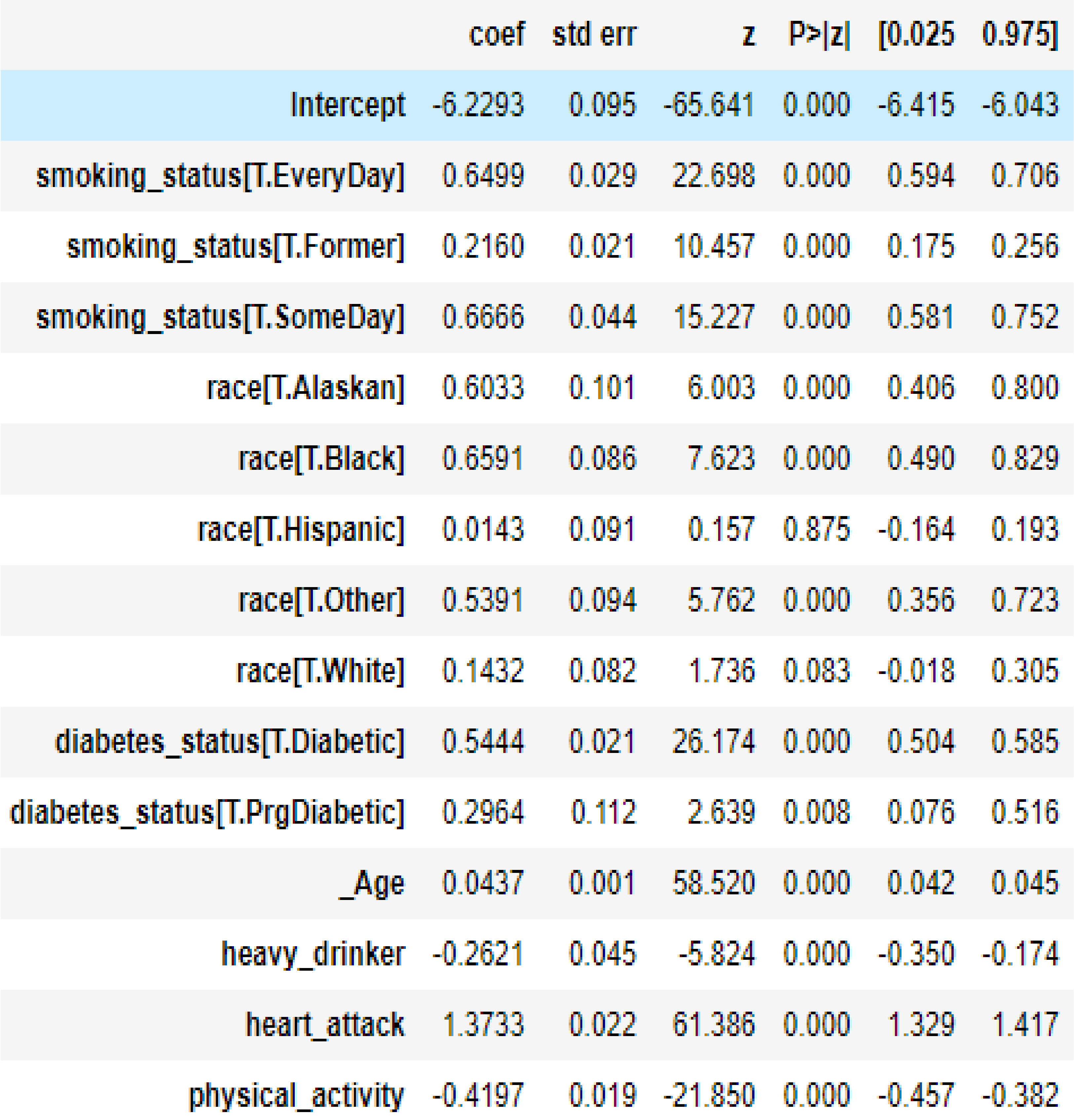
LOGISTICS REGRESSION

- Statistical method for modeling binary outcomes (yes/no or 0/1).
- Utilizes predictor variables to estimate the probability of a specific event.
- Determines the relationship between independent variables and event probability.

REASON

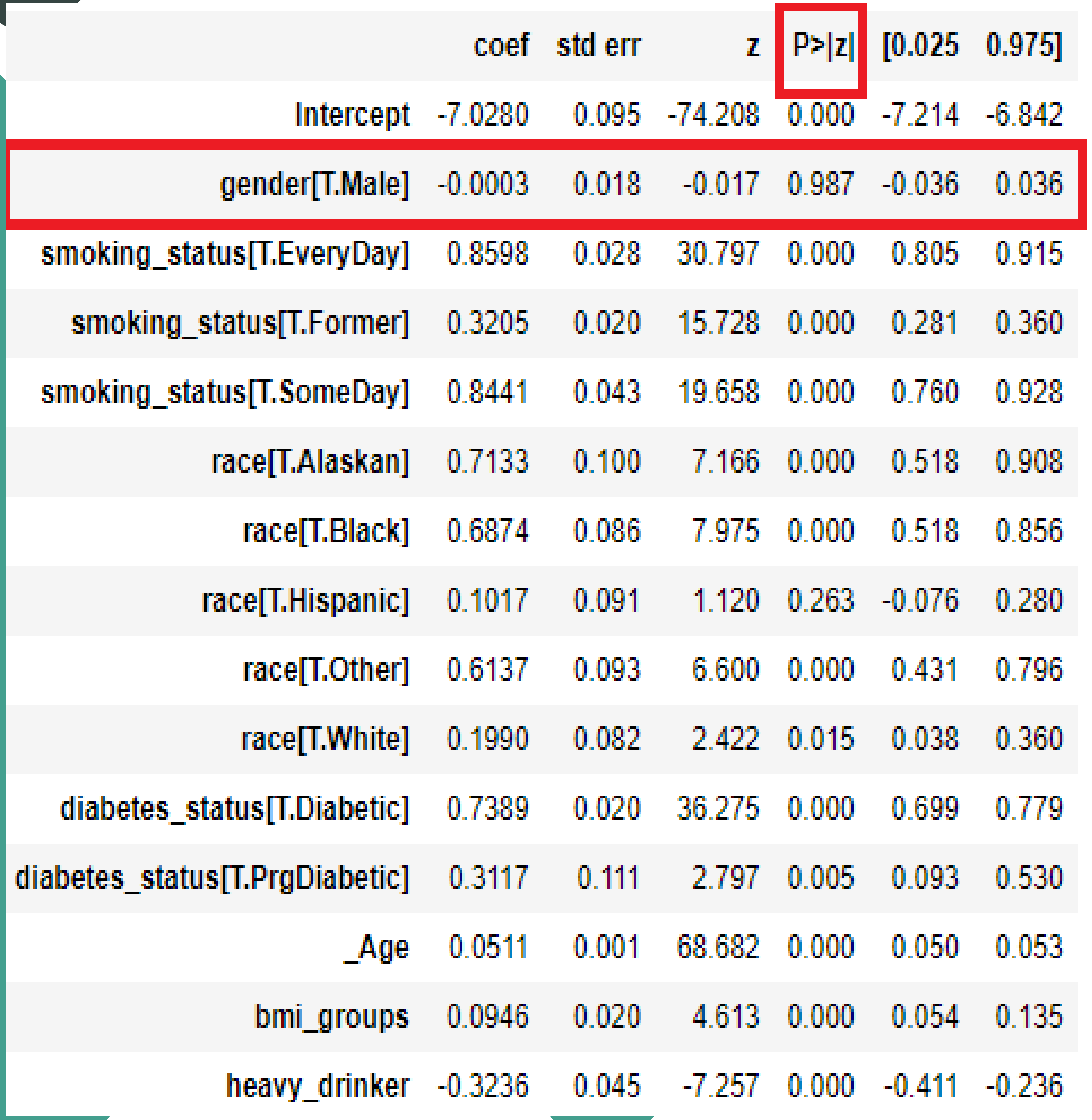
- Logistic regression was selected due to the binary nature of "stroke_status" (0 for "no" and 1 for "yes").
- It is well-suited for modeling binary outcomes, making it appropriate for studying the relationship with predictor variables.

LOGISTIC REGRESSION RESULTS

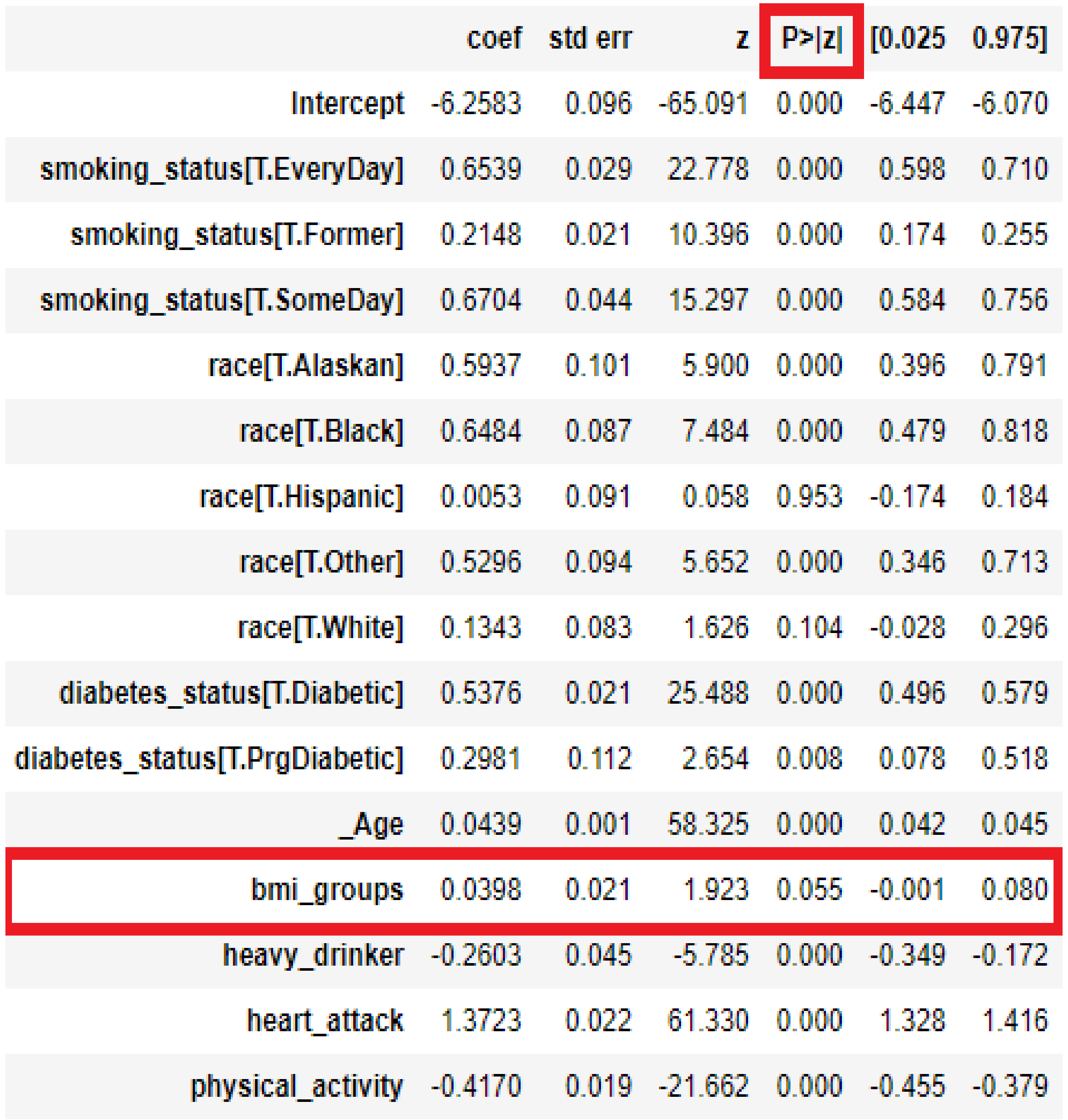


	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-6.2293	0.095	-65.641	0.000	-6.415	-6.043
smoking_status[T.EveryDay]	0.6499	0.029	22.698	0.000	0.594	0.706
smoking_status[T.Former]	0.2160	0.021	10.457	0.000	0.175	0.256
smoking_status[T.SomeDay]	0.6666	0.044	15.227	0.000	0.581	0.752
race[T.Alaskan]	0.6033	0.101	6.003	0.000	0.406	0.800
race[T.Black]	0.6591	0.086	7.623	0.000	0.490	0.829
race[T.Hispanic]	0.0143	0.091	0.157	0.875	-0.164	0.193
race[T.Other]	0.5391	0.094	5.762	0.000	0.356	0.723
race[T.White]	0.1432	0.082	1.736	0.083	-0.018	0.305
diabetes_status[T.Diabetic]	0.5444	0.021	26.174	0.000	0.504	0.585
diabetes_status[T.PrgDiabetic]	0.2964	0.112	2.639	0.008	0.076	0.516
_Age	0.0437	0.001	58.520	0.000	0.042	0.045
heavy_drinker	-0.2621	0.045	-5.824	0.000	-0.350	-0.174
heart_attack	1.3733	0.022	61.386	0.000	1.329	1.417
physical_activity	-0.4197	0.019	-21.850	0.000	-0.457	-0.382

SOME ITERATIONS



	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-7.0280	0.095	-74.208	0.000	-7.214	-6.842
gender[T.Male]	-0.0003	0.018	-0.017	0.987	-0.036	0.036
smoking_status[T.EveryDay]	0.8598	0.028	30.797	0.000	0.805	0.915
smoking_status[T.Former]	0.3205	0.020	15.728	0.000	0.281	0.360
smoking_status[T.SomeDay]	0.8441	0.043	19.658	0.000	0.760	0.928
race[T.Alaskan]	0.7133	0.100	7.166	0.000	0.518	0.908
race[T.Black]	0.6874	0.086	7.975	0.000	0.518	0.856
race[T.Hispanic]	0.1017	0.091	1.120	0.263	-0.076	0.280
race[T.Other]	0.6137	0.093	6.600	0.000	0.431	0.796
race[T.White]	0.1990	0.082	2.422	0.015	0.038	0.360
diabetes_status[T.Diabetic]	0.7389	0.020	36.275	0.000	0.699	0.779
diabetes_status[T.PrgDiabetic]	0.3117	0.111	2.797	0.005	0.093	0.530
_Age	0.0511	0.001	68.682	0.000	0.050	0.053
bmi_groups	0.0946	0.020	4.613	0.000	0.054	0.135
heavy_drinker	-0.3236	0.045	-7.257	0.000	-0.411	-0.236



	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-6.2583	0.096	-65.091	0.000	-6.447	-6.070
smoking_status[T.EveryDay]	0.6539	0.029	22.778	0.000	0.598	0.710
smoking_status[T.Former]	0.2148	0.021	10.396	0.000	0.174	0.255
smoking_status[T.SomeDay]	0.6704	0.044	15.297	0.000	0.584	0.756
race[T.Alaskan]	0.5937	0.101	5.900	0.000	0.396	0.791
race[T.Black]	0.6484	0.087	7.484	0.000	0.479	0.818
race[T.Hispanic]	0.0053	0.091	0.058	0.953	-0.174	0.184
race[T.Other]	0.5296	0.094	5.652	0.000	0.346	0.713
race[T.White]	0.1343	0.083	1.626	0.104	-0.028	0.296
diabetes_status[T.Diabetic]	0.5376	0.021	25.488	0.000	0.496	0.579
diabetes_status[T.PrgDiabetic]	0.2981	0.112	2.654	0.008	0.078	0.518
_Age	0.0439	0.001	58.325	0.000	0.042	0.045
bmi_groups	0.0398	0.021	1.923	0.055	-0.001	0.080
heavy_drinker	-0.2603	0.045	-5.785	0.000	-0.349	-0.172
heart_attack	1.3723	0.022	61.330	0.000	1.328	1.416
physical_activity	-0.4170	0.019	-21.662	0.000	-0.455	-0.379

Covariates	Log Odds	Odds Ratio	Interpretation
Intercept	-6.258	0.002	The baseline odds of having a stroke when all predictors are at their reference levels is about 0.002
smoking_status[T.EveryDay]	0.6539	1.92	Every day smokers have about 1.92 higher odds of having a stroke compared to the non smokers
smoking_status[T.Former]	0.2148	1.24	Former smokers have about 1.24 higher odds of having a stroke compared to the non smokers
smoking_status[T.SomeDay]	0.6704	1.96	Some day smokers have about 1.96 higher odds of having a stroke compared to the non smokers
race[T.Alaskan]	0.5937	1.81	Alaskans have about 1.81 higher odds of having a stroke compared to Asians
race[T.Black]	0.6484	1.91	Black people have about 1.91 higher odds of having a stroke compared to Asians
race[T.Hispanic]	0.0053	1.01	Hispanic people have about 1.01 higher odds of having a stroke compared to Asians
race[T.White]	0.5296	1.7	White people have about 1.7 higher odds of having a stroke compared to Asians
race[T.Other]	0.1343	1.14	Races other than mentioned have about 1.343 higher odds of having a stroke compared to Asians
diabetes_status[T.Diabetic]	0.5376	1.7	Diabetic individuals have about 1.7 higher odds of having a stroke compared to Non-diabetic individuals
diabetes_status[T.PrgDiabetic]	0.2981	1.35	Diabetic Pregnant Women have about 1.35 higher odds of having a stroke compared to Non-diabetic individuals
_Age	0.0439	1.04	For every one year age increase the odds of having a stroke increase by 1.04
heavy_drinker	-0.2603	0.77	Heavy drinkers have about 1.04 lower odds of having a stroke compared to individuals that are not heavy drinkers
heart_attack	1.3723	3.94	Individuals that had Heart attack have about 3.94 higher odds of having a stroke compared to individuals that didn't have Heart attack
physical_activity	-0.417	0.66	Physically active Individuals have about 0.66 lower odds of having a stroke compared to physically inactive individuals

THANK YOU

FOR YOUR ATTENTION