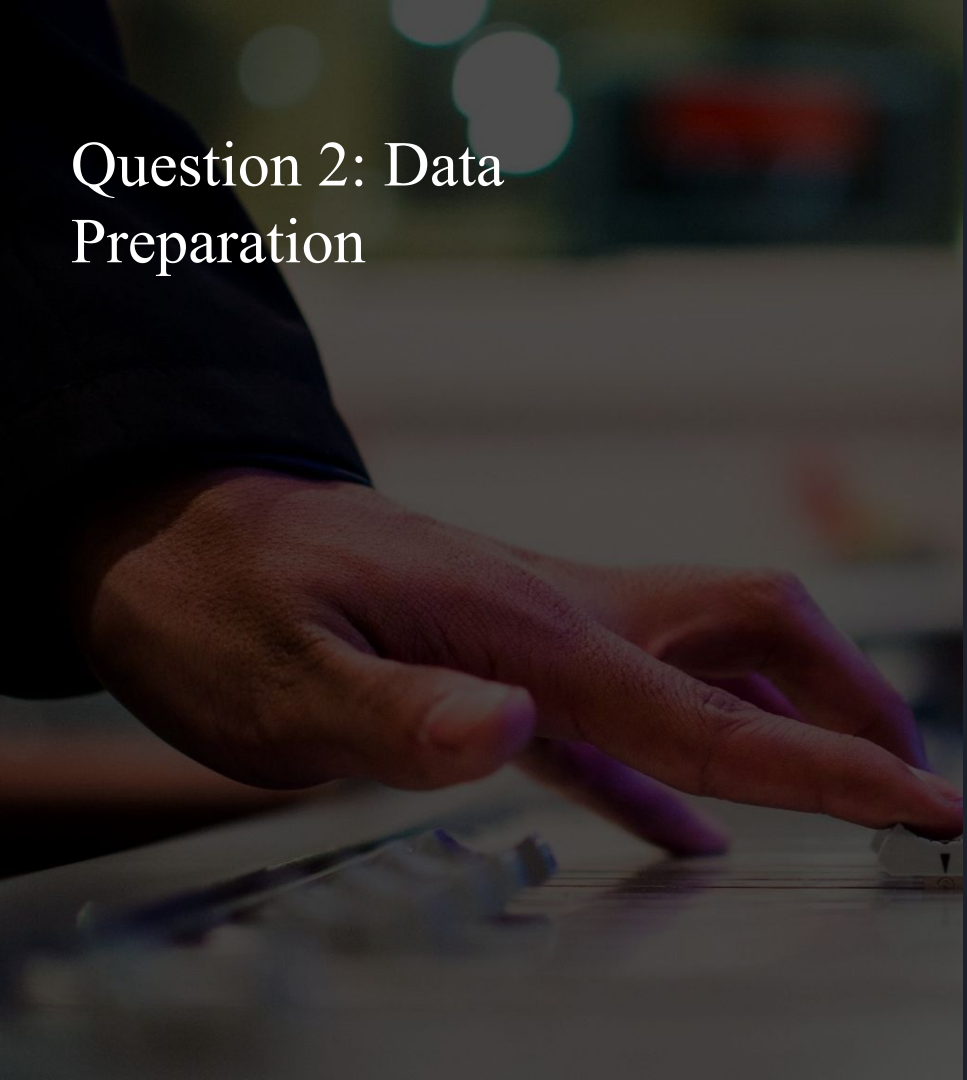
DS Practicum 3: Diabetes

Introduction:

Given a database consisting of profiles from patients from the CDC's research, we are to figure out patterns as to what engenders higher rates of diabetes based on correlation from attributes and fields taken from the data, such as financial status, education, general health as reported by the patient both physical and mental, with more to come. In order to do so, we must appropriately organize and sort the data, before doing exploratory analysis of its properties, visualizing the results to sift for patterns and trends, and present and defend any conclusions which may come of it.



Question 2: Data Preparation

Field Alterations

Checking for Abnormal Values

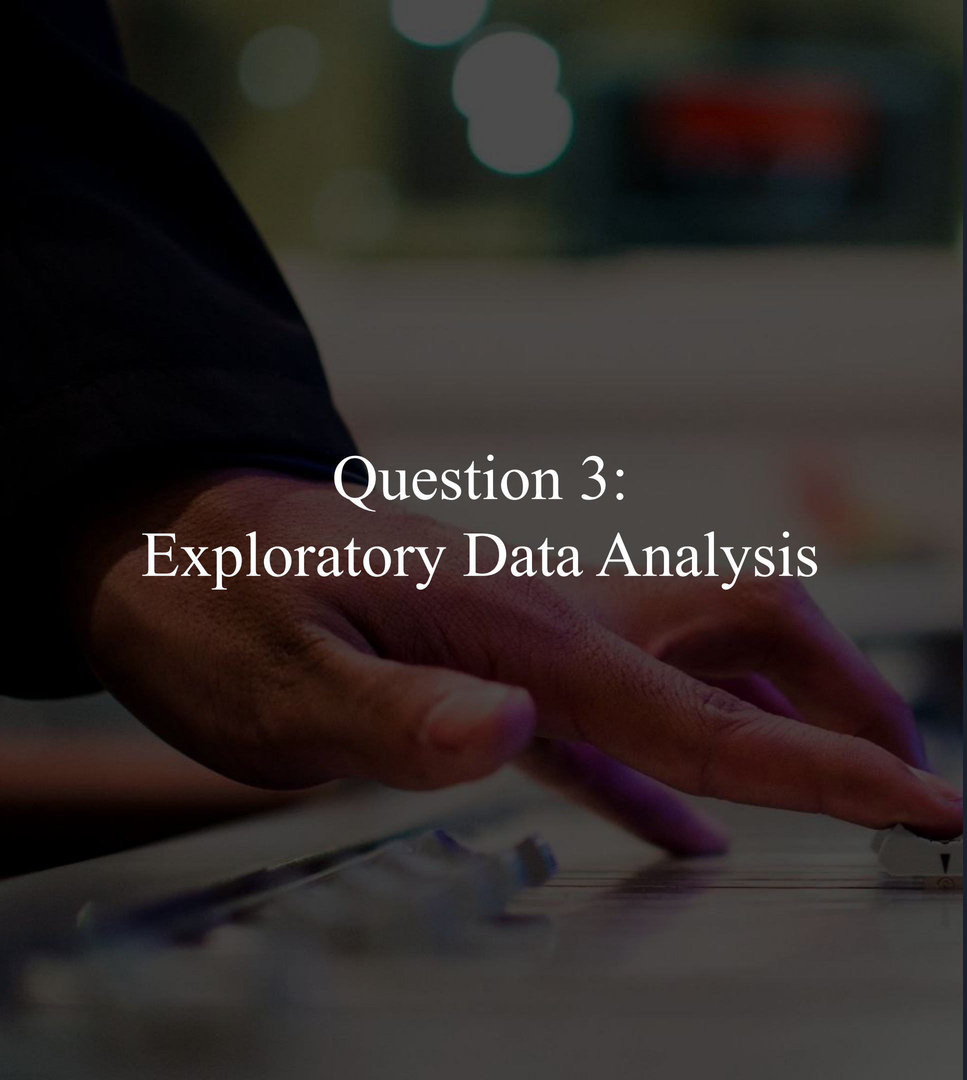
Field Alterations

```
Data columns (total 22 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   Diabetes_012          253680 non-null  float64
 1   HighBP                253680 non-null  float64
 2   HighChol              253680 non-null  float64
 3   CholCheck             253680 non-null  float64
 4   BMI                   253680 non-null  float64
 5   Smoker                253680 non-null  float64
 6   Stroke                253680 non-null  float64
 7   HeartDiseaseorAttack  253680 non-null  float64
 8   PhysActivity          253680 non-null  float64
 9   Fruits                253680 non-null  float64
 10  Veggies               253680 non-null  float64
 11  HvyAlcoholConsump     253680 non-null  float64
 12  AnyHealthcare         253680 non-null  float64
 13  NoDocbcCost           253680 non-null  float64
 14  GenHlth               253680 non-null  float64
 15  MentHlth              253680 non-null  float64
 16  PhysHlth              253680 non-null  float64
 17  DiffWalk              253680 non-null  float64
 18  Sex                   253680 non-null  float64
 19  Age                   253680 non-null  float64
 20  Education             253680 non-null  float64
 21  Income                253680 non-null  float64
dtypes: float64(22)
```

```
Data columns (total 22 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   Diabetes_012          253680 non-null  int32
 1   HighBP                253680 non-null  bool
 2   HighChol              253680 non-null  bool
 3   CholCheck             253680 non-null  bool
 4   BMI                   253680 non-null  int32
 5   Smoker                253680 non-null  bool
 6   Stroke                253680 non-null  bool
 7   HeartDiseaseorAttack  253680 non-null  bool
 8   PhysActivity          253680 non-null  bool
 9   Fruits                253680 non-null  bool
 10  Veggies               253680 non-null  bool
 11  HvyAlcoholConsump     253680 non-null  bool
 12  AnyHealthcare         253680 non-null  bool
 13  NoDocbcCost           253680 non-null  bool
 14  GenHlth               253680 non-null  int32
 15  MentHlth              253680 non-null  int32
 16  PhysHlth              253680 non-null  int32
 17  DiffWalk              253680 non-null  bool
 18  Sex                   253680 non-null  bool
 19  Age                   253680 non-null  int32
 20  Education             253680 non-null  int32
 21  Income                253680 non-null  int32
dtypes: bool(14), int32(8)
memory usage: 11.1 MB
```

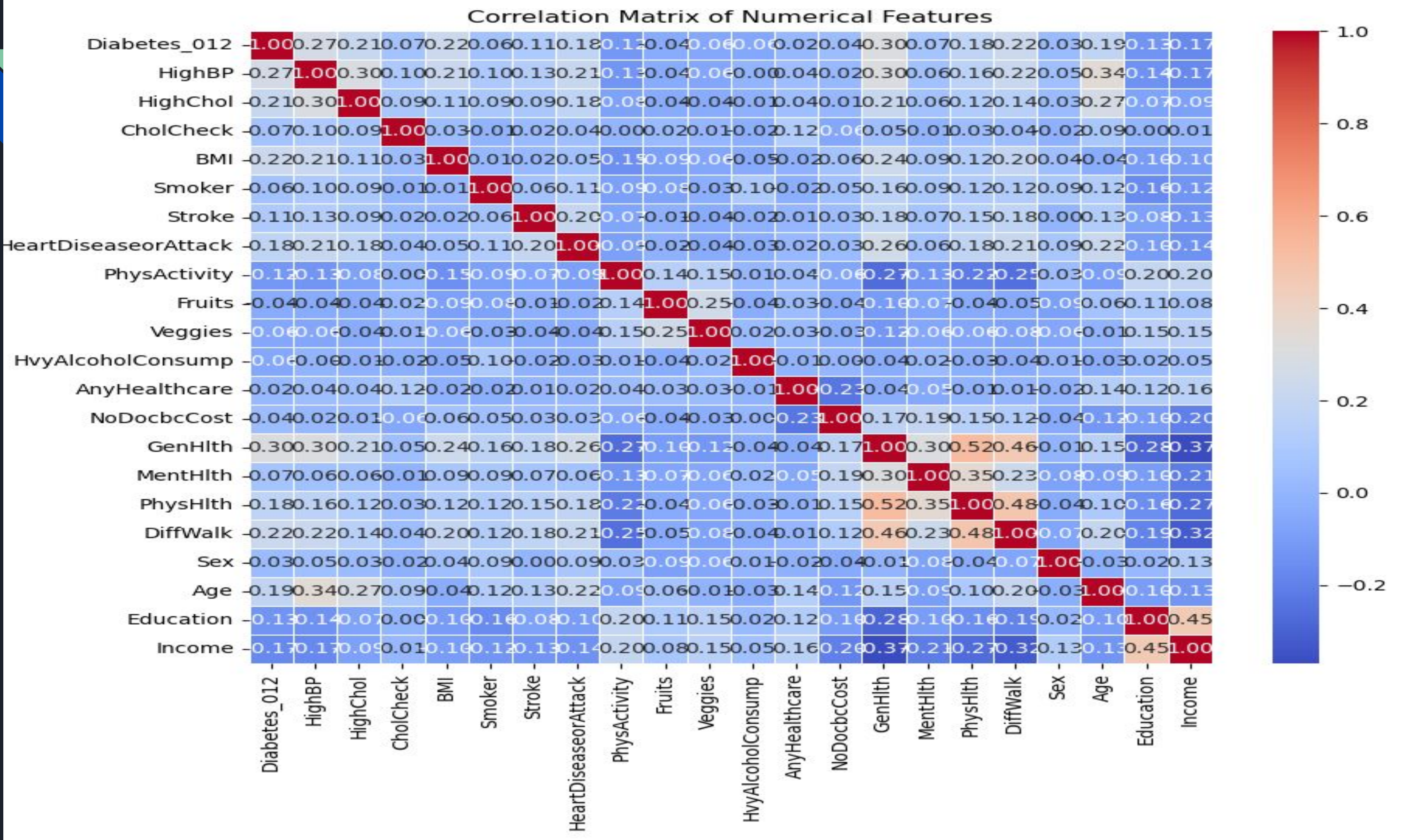
Checking for Abnormal Values

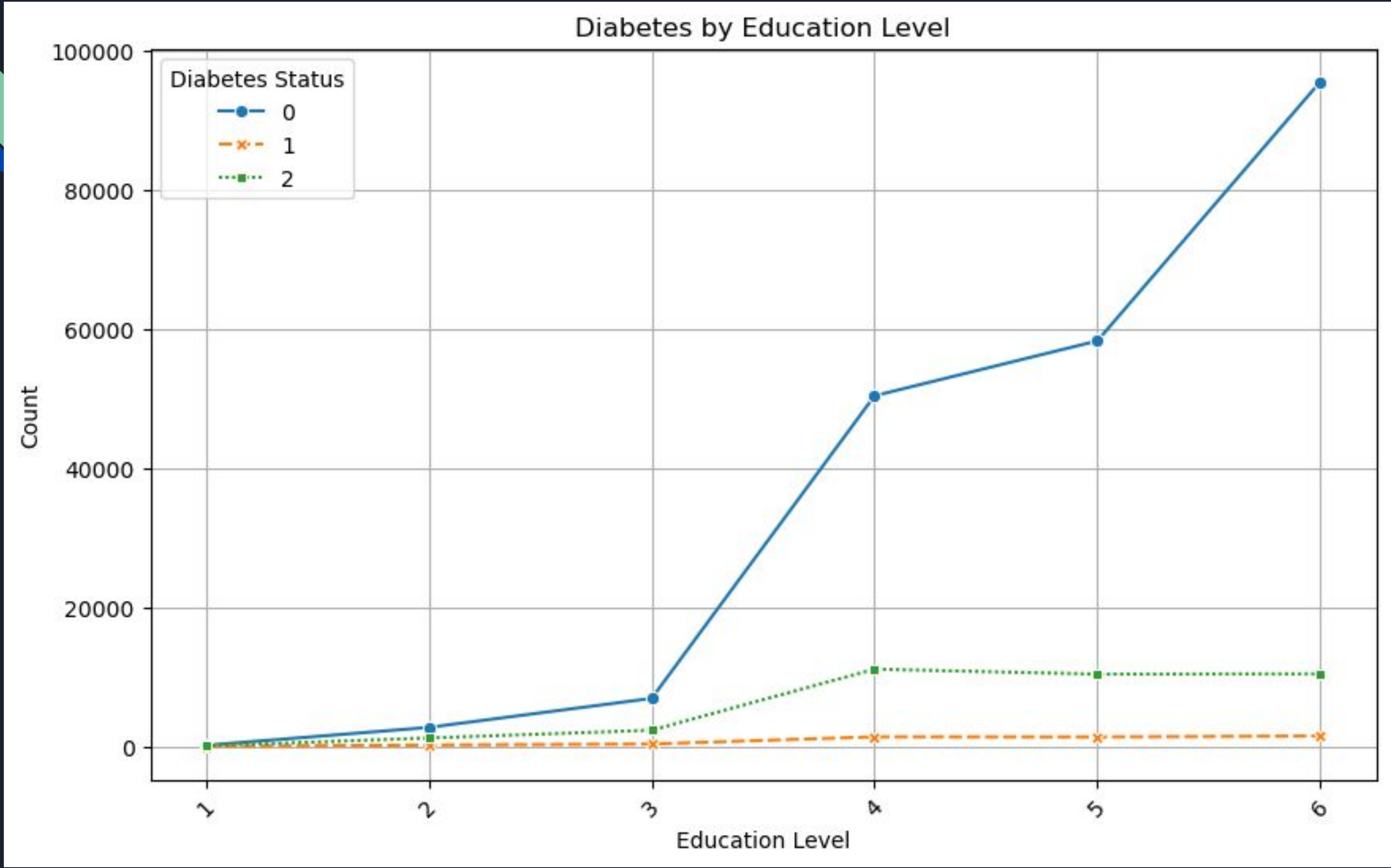
```
Max and Min of BMI:
Max: 98
Min: 12

Max and Min of MentHlth:
Max: 30
Min: 0

Max and Min of PhysHlth:
Max: 30
Min: 0

Max and Min of Age:
Max: 13
Min: 1

Max and Min of Income:
Max: 8
Min: 1
```



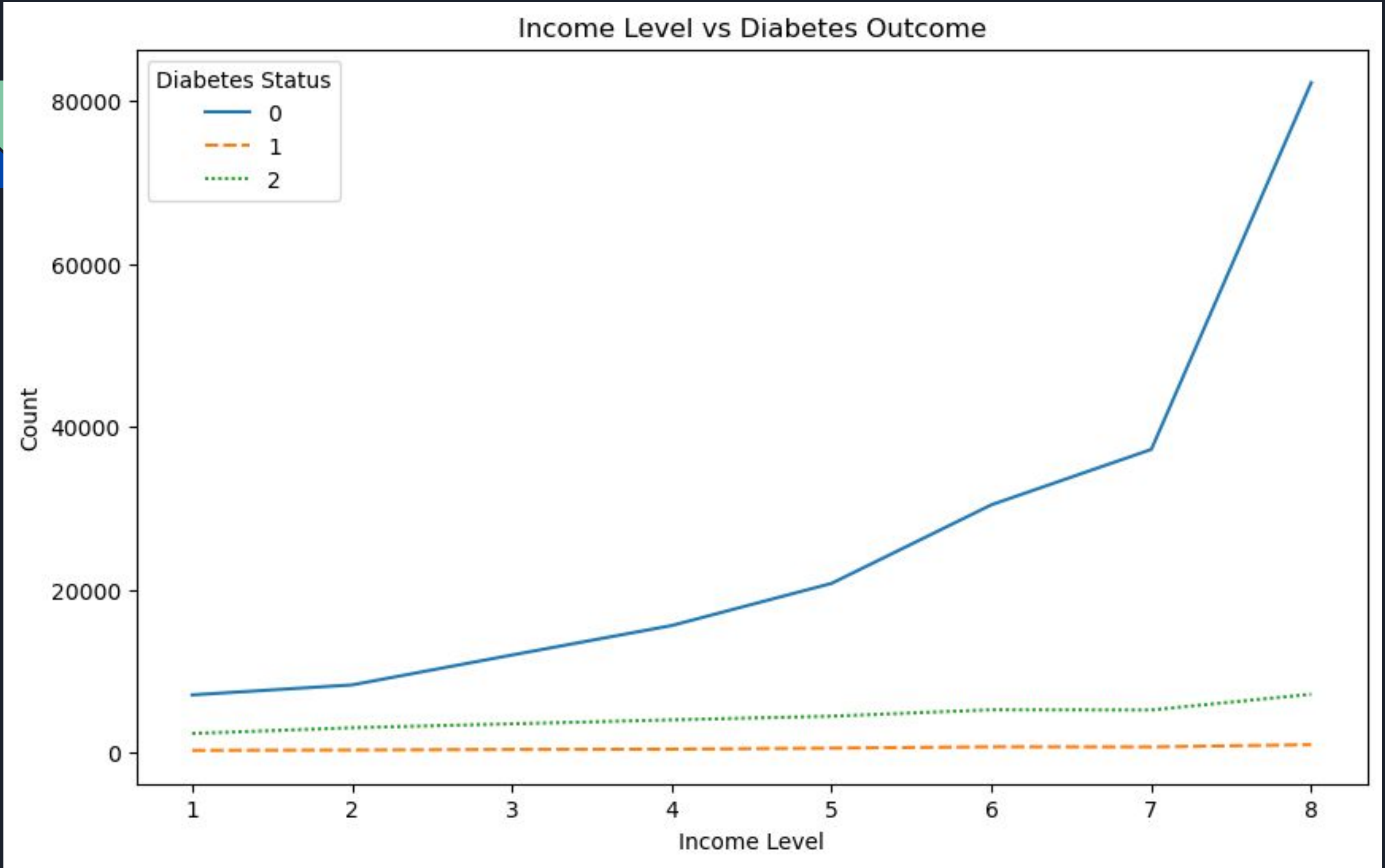
Question 3:
Exploratory Data Analysis

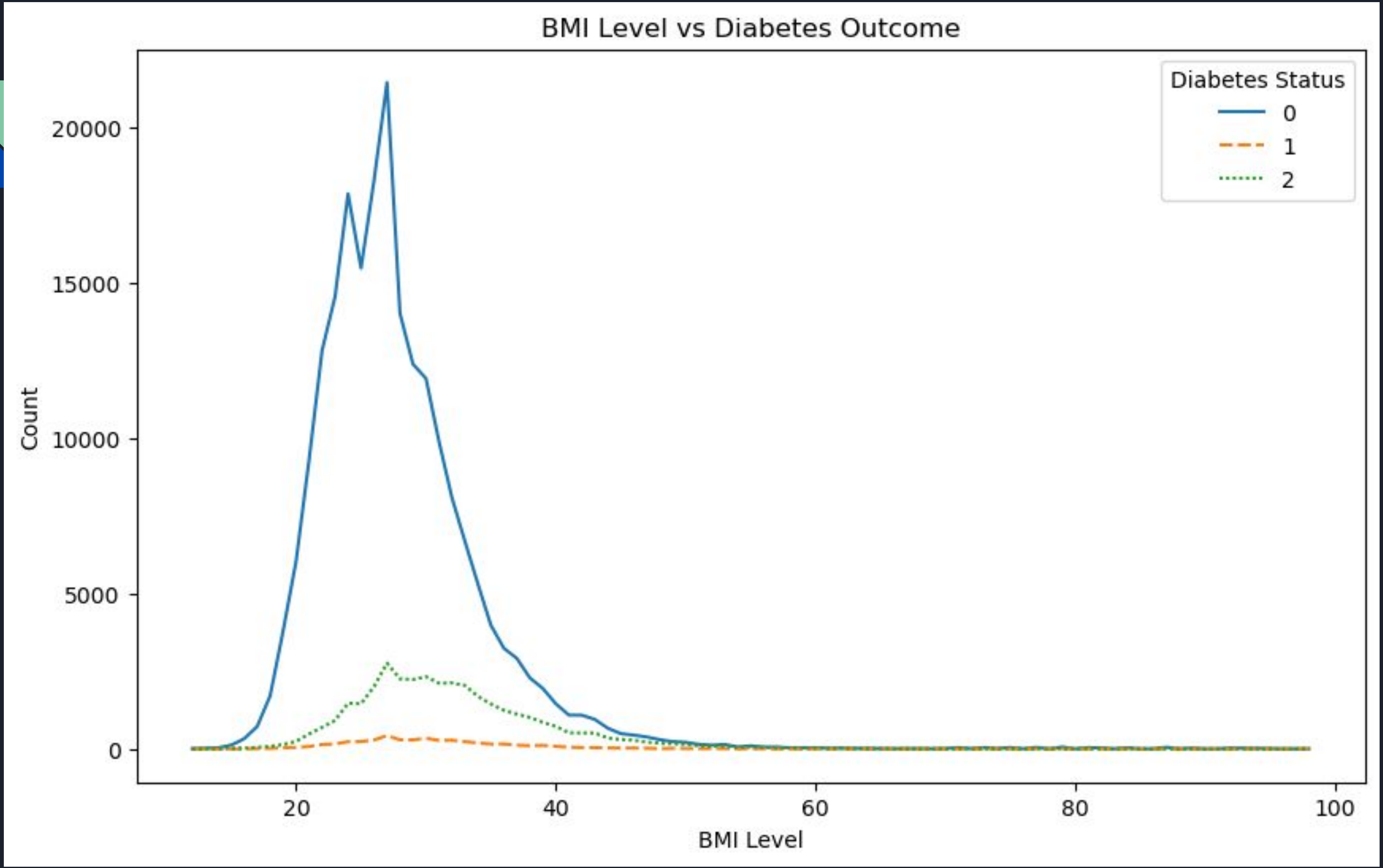
1. Correlation Matrix
2. Diabetes by Income
3. Diabetes by Education
4. Diabetes by BMI
5. Diabetes by GenHlth

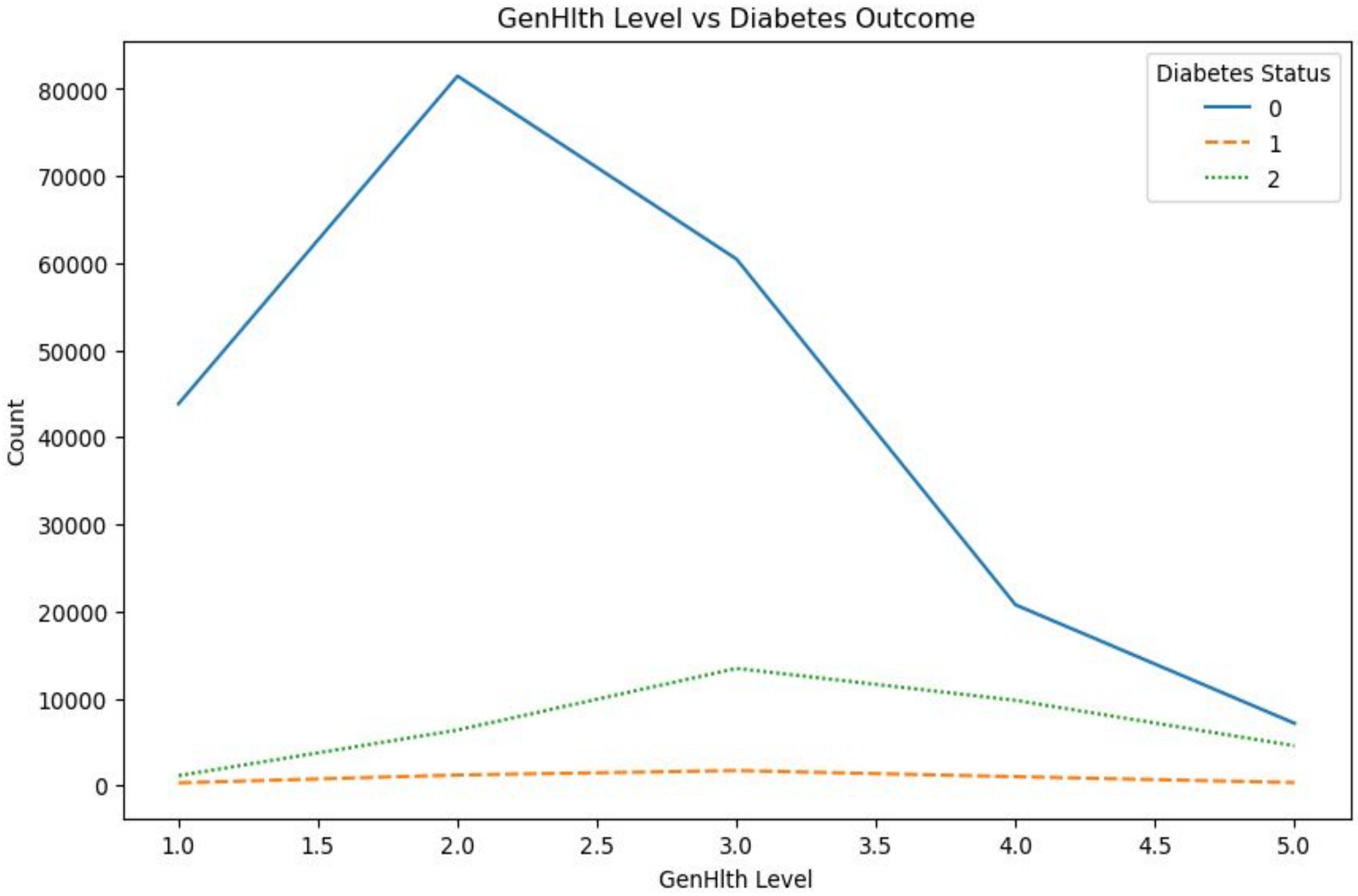

Correlation Matrix of Numerical Features
Diabetes_012
HighBP
HighChol
CholCheck
BMI
Smoker
Stroke
HeartDiseaseorAttack
PhysActivity
Fruits
Veggies
HvyAlcoholConsump
AnyHealthcare
NoDocbcCost
GenHlth
MentHlth
PhysHlth
DiffWalk
Sex
Age
Education
Income
1.0
0.8
0.6
0.4
0.2
0.0
−0.2



Diabetes by Education Level
Diabetes Status
0
1
2
100000
80000
60000
40000
20000
0
Count
1
2
3
4
5
6
Education Level


Income Level vs Diabetes Outcome
Diabetes Status
0
1
2
80000
60000
40000
20000
0
Count
1
2
3
4
5
6
7
8
Income Level

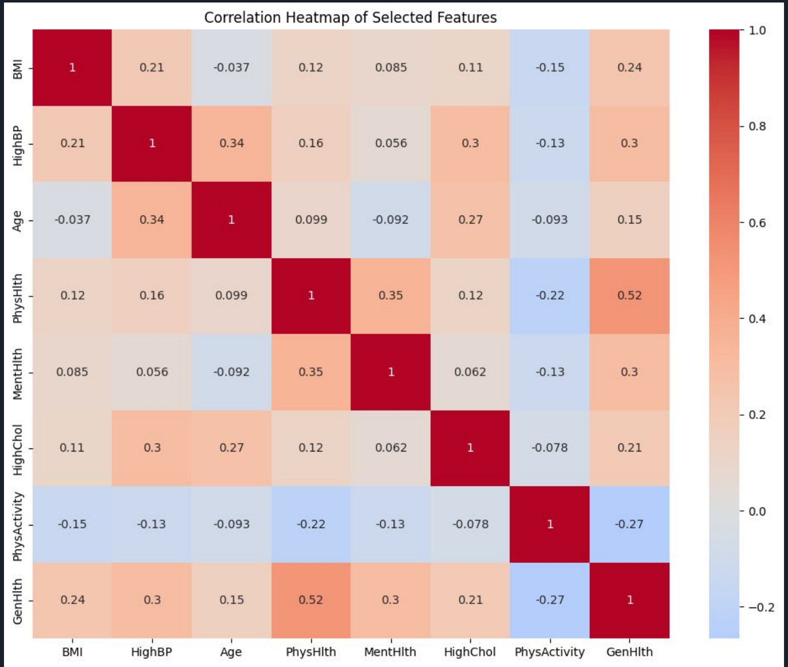

BMI Level vs Diabetes Outcome
Diabetes Status
0
1
2
Count
20000
15000
10000
5000
0
20
40
60
80
100
BMI Level


GenHlth Level vs Diabetes Outcome
Diabetes Status
0
1
2
Count
80000
70000
60000
50000
40000
30000
20000
10000
0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
GenHlth Level

Question 4:

Perform significance tests to determine if the patterns that are detected above are statistically significant.

Three Significant Features Selected:



BMI and Diabetes

High Blood Pressure and Diabetes

Physical and Mental Health Correlation

Question 4 Bonus

Machine learning models can benefit greatly from feature engineering. Create a new feature that can be included in the model and perform significance testing to determine if it's statistically significant. Explain the results and justify if the feature will be included in the ML model. If you decide that you will not include the new feature in the ML model, explain the reasons.

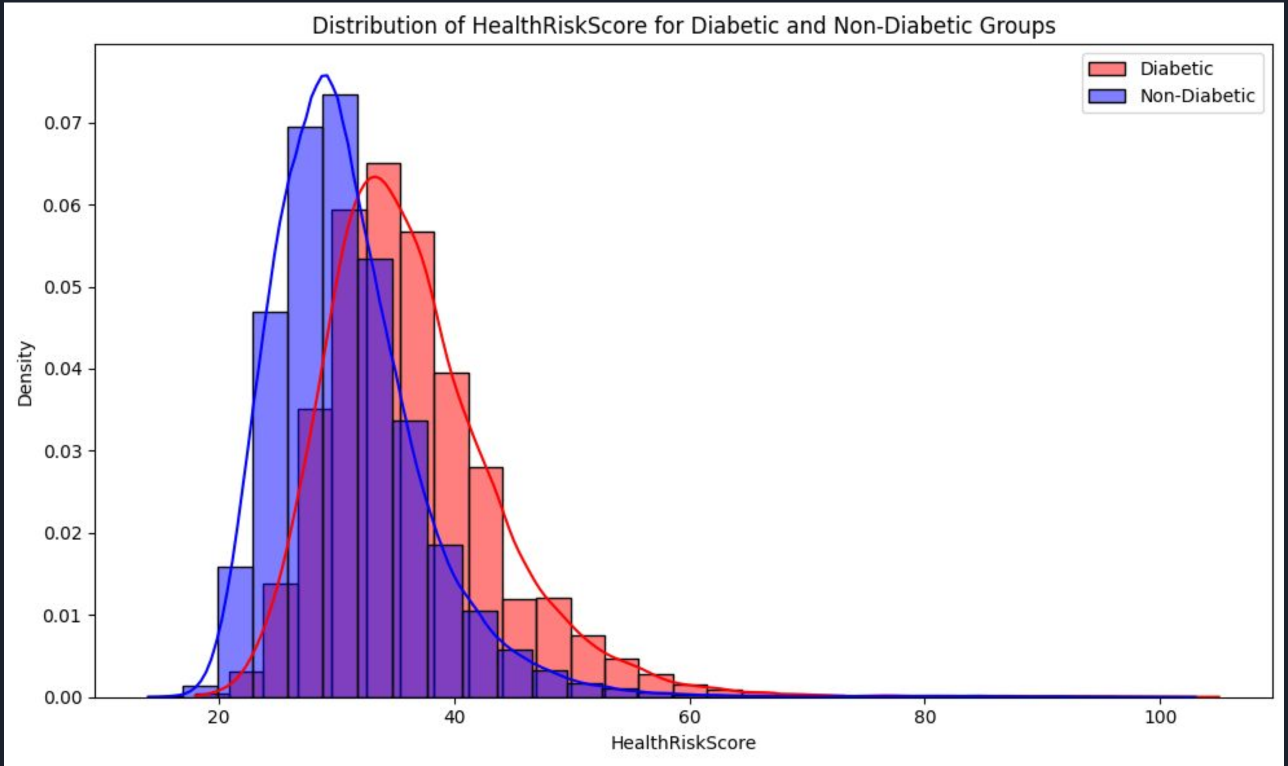
Question 6

Building the Models

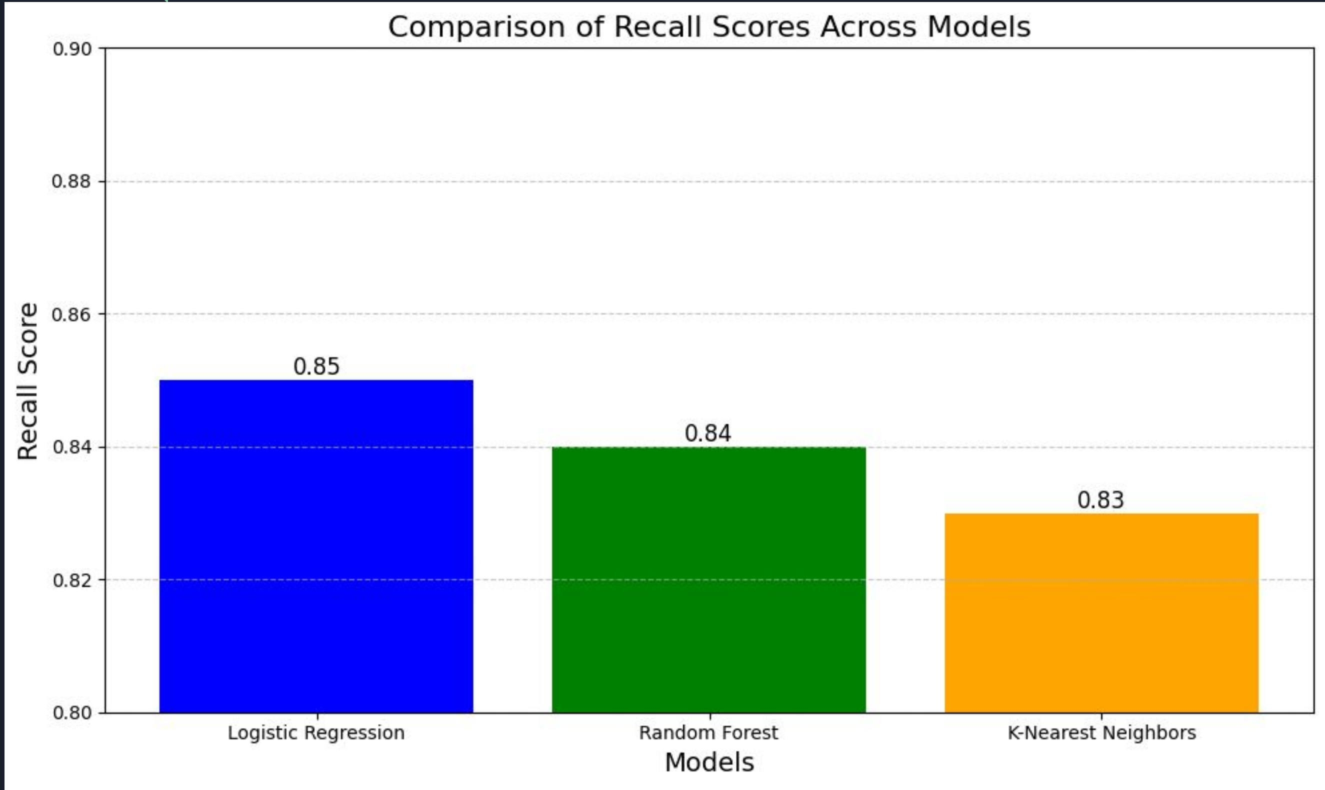
- Logistic Regression Model
- K-Nearest Neighbors (KNN)
- Random Forest Classifier

Model Results

- Even with the optimal parameters (n_neighbors=3, weights='distance'), the KNN Classifier struggled to balance precision and recall, as evidenced by its recall (39.54%) and F1-score (40.22%), despite its 81.22% accuracy.
- Despite having a similar recall (38.64%) and F1-score (39.65%), the Logistic Regression scored somewhat better in accuracy (84.75%), indicating that it might not adequately capture the underlying relationships between attributes and the multiclass target.
- The Random Forest Classifier was also a balanced model, with the best accuracy (85.26%) and demonstrating enhanced recall (42.15%) and F1-score (43.78%) with optimal settings (max_depth=20, n_estimators=50).

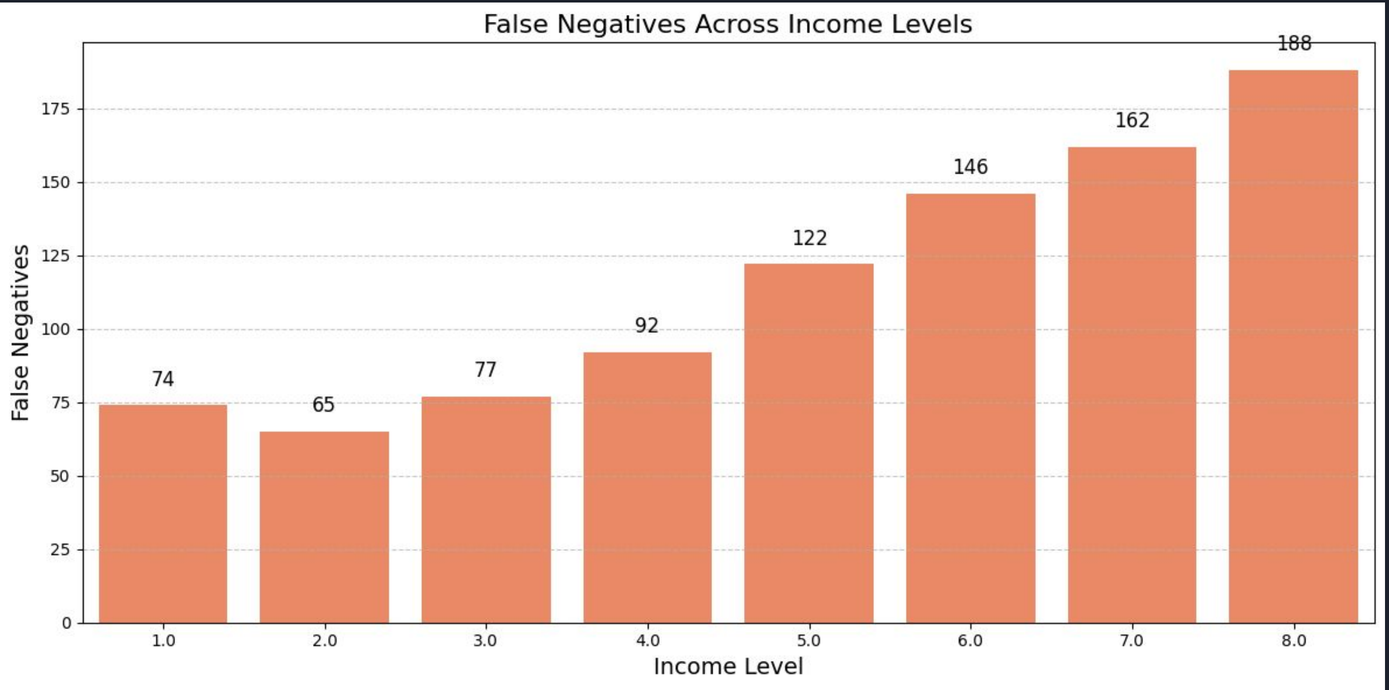

Distribution of HealthRiskScore for Diabetic and Non-Diabetic Groups
Diabetic
Non-Diabetic
Density
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
20
40
60
80
100
HealthRiskScore

Question 7



- Random Forest Classifier - 0.84
- K- Nearest Neighbors Classifier - 0.83
- ***<u>Logistic Regression -0.85</u>***

Question 8



- Income levels with fewer samples may have lower recall due to imbalanced representation.

- False negatives disproportionately affect lower-income groups, highlighting inequity in model predictions.