Analyzing Demographic and Health Factors Influencing Diabetes Prevalence

PROJECT
REPORT

PREPARED FOR :

Statistics For Data Science

PREPARED BY GROUP 1

Ashraf Afana
Anu Girija
Amanjot Kandhola
Alan Mathews
April Uy

1. BUSINESS PROBLEM

Objective: Identify the key factors contributing to diabetes across different racial groups and genders.

Success Criteria: Understand the relationship between diabetes and potential risk factors such as BMI, HbA1c level, blood glucose level, smoking history, and hypertension.
Derive actionable insights that could inform public health interventions.

2. ANALYTICS PROBLEM

Reframe the problem as a classification and dependency analysis.

Use statistical models and hypothesis testing to evaluate: i.)Dependency of diabetes on race, ii.) Impact of variables like BMI, HbA1c, blood glucose, smoking, and hypertension on diabetes, iii.) Examine probabilities of diabetes within specific racial and BMI categories.

3. DATA

DATA SOURCE: The dataset is taken from https://www.kaggle.com/code/ailenenunez/diabetes-prediction-99-recall-rate/input
PREPARATION: The dataset was cleaned to: Consolidate race columns, Standardize gender labels, Exclude rows with missing or "No Info" smoking history data, Created BMI categories to enable analysis across weight groups.
EXPLORATION: Examined the relationship between diabetes, race, and other key predictors using pivot tables, contingency tables, and statistical models.

4. METHODOLOGY SELECTION

1- Descriptive Statistics: To summarize data distributions.

2- Pivot Tables: To calculate marginal probabilities of diabetes for specific groups.

3- Probability in each group between the success factors (predictors) AND gender.

4- Correlation matrix for being Diabetic /Male/Female across all races and predictors.

5- Logistic Regression: To model the likelihood of diabetes based on risk factors and calculate the significance of each variable. Logistic Regression model is created for each predictor as well.

6- Chi-Square Test: To evaluate the dependency of diabetes on race and gender.

7- Hypothesis test between different races.

5. MODEL BUILDING

Probability in each group between the success factors (predictors) AND gender :
Across all races, Obesity sticks out as a main factor for being diabetic, also having a record of (hbA1c_level, blood_glucose_level, heart_disease & smoking_history. Also, being Males had the highest probabilities with these factors for being diabetic over 20%

Correlation matrix for being Diabetic /Male/Female across all races and predictors.
The conclusion here, there is a strong correlation between being Male, Hispanic and has a record of hbA1c. for Female, being Asian and has a record of heart disease

5. MODEL BUILDING

Logistic Regression: To model the likelihood of diabetes based on risk factors and calculate the significance of each variable.

1. Hypertension

S-Curve Interpretation:

- The S-curve likely shows a steep increase in probability for higher values of hypertension (binary variable, 0 or 1).
- Individuals with hypertension (value = 1) have a significantly higher probability of being diabetic compared to those without hypertension.

Conclusion:

The logistic regression coefficient for hypertension (1.0946) translates to an odds ratio of approximately 2.99, indicating that individuals with hypertension are nearly 3 times more likely to have diabetes.

2. Heart Disease

S-Curve Interpretation:

- Similar to hypertension, the S-curve for heart disease (binary variable, 0 or 1) shows a jump in probability for those with heart disease.
- The curve indicates a significant increase in diabetes probability for individuals with heart disease.

Conclusion:

With a coefficient of 1.4649 and an odds ratio of approximately 4.33, individuals with heart disease are over 4 times more likely to have diabetes.

3. BMI (Body Mass Index)

S-Curve Interpretation:

- BMI has a continuous range, so the S-curve is more gradual.
- As BMI increases, the probability of being diabetic also increases, but the change is relatively modest compared to binary predictors.

Conclusion:

The coefficient for BMI is 0.0874, corresponding to an odds ratio of approximately 1.09. For every unit increase in BMI, the odds of being diabetic increase by about 9%.

5. MODEL BUILDING

4. HbA1c Level
S-Curve Interpretation:

- The S-curve for HbA1c levels shows a sharp increase in diabetes probability as HbA1c levels rise.
- This reflects the strong association between elevated HbA1c levels and diabetes.

Conclusion:
The coefficient for HbA1c level is 1.9608, indicating that higher HbA1c levels are one of the strongest predictors of diabetes. The odds ratio is approximately 7.11, meaning a significant increase in diabetes probability for higher HbA1c levels.

5. Blood Glucose Level
S-Curve Interpretation:

- Blood glucose level also exhibits a positive relationship with diabetes probability, but the increase is more gradual compared to HbA1c.
- The S-curve shows that as blood glucose levels increase, the likelihood of diabetes also increases.

Conclusion:
The coefficient for blood glucose level is 0.0337, corresponding to an odds ratio of approximately 1.03. This indicates that for every unit increase in blood glucose, the odds of being diabetic increase by about 3%.

6. Smoking History
Categories: Current, Ever, Former, Never, Not Current
S-Curve Interpretation:

- Each smoking history category is represented by a dummy variable (binary, 0 or 1). The S-curves for these categories show how diabetes probability changes based on smoking history.
- Categories such as "Current" and "Ever" smoking history show higher probabilities of diabetes compared to others.

Conclusion:
Current Smoking: Coefficient = 0.8011; Odds Ratio ≈ 2.23

- Individuals who currently smoke are over twice as likely to have diabetes compared to non-smokers.

Ever Smoked: Coefficient = 1.0365; Odds Ratio ≈ 2.82

- Those who have smoked at some point in the past are nearly 3 times as likely to have diabetes.

5. MODEL BUILDING

Former Smoker: Coefficient = 1.0739; Odds Ratio ≈ 2.93

- Similar to "Ever Smoked," with slightly higher odds.

Never Smoked: Coefficient = 0.6955; Odds Ratio ≈ 2.00

- Individuals who never smoked still have higher odds compared to the baseline.

Not Current Smoker: Coefficient = 0.8838; Odds Ratio ≈ 2.42

- Not currently smoking still significantly increases diabetes probability compared to the baseline.

General Observations:

1. Predictors like HbA1c level, heart disease, and hypertension are the strongest drivers of diabetes probability, with steep S-curves and large coefficients.

2. Continuous predictors like BMI and blood glucose level show gradual increases in probability.

3. Binary predictors like smoking history categories, heart disease, and hypertension result in distinct jumps in the S-curve, reflecting their categorical nature.

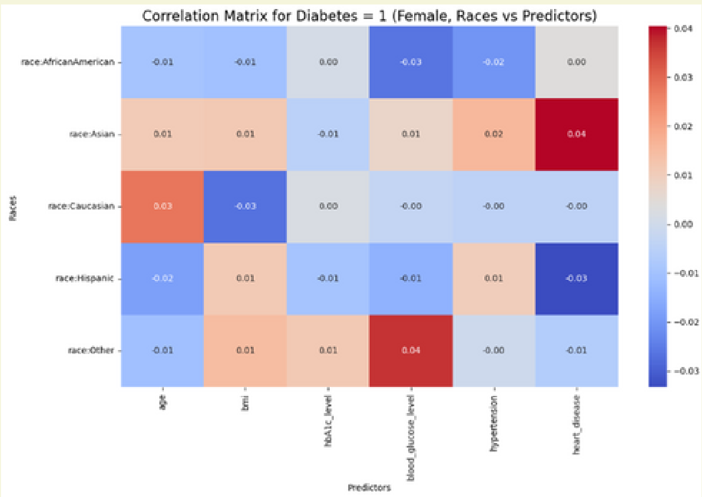
Chi-Square Test: To evaluate the dependency of diabetes on race and gender.

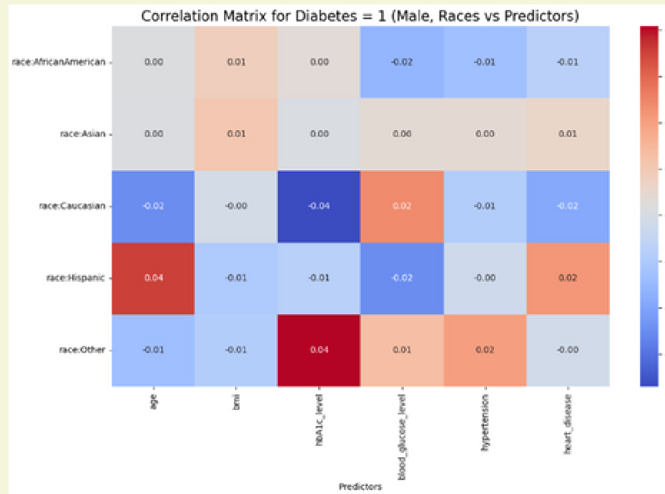
Fail to reject the null hypothesis indicating that being diabetic is independent from being Male/Female across the races

6. RESULTS

Correlation matrix for being Diabetic /Male/Female across all races and predictors.

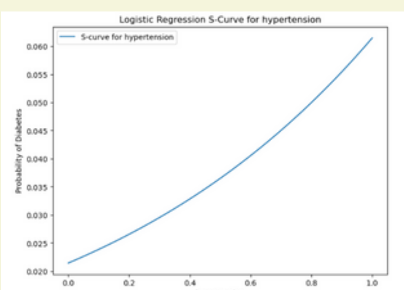
The conclusion here, there is a strong correlation between being Male, Hispanic and has a record of hbA1c. For Female, being Asian and has a record of heart disease.

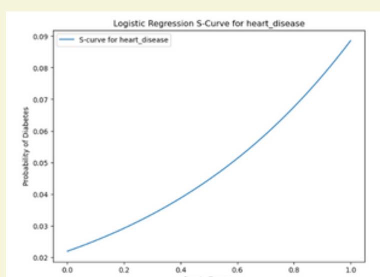


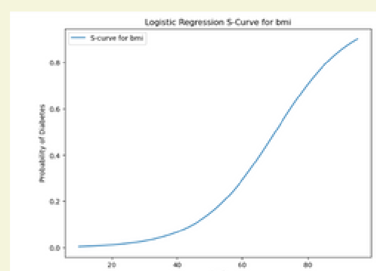


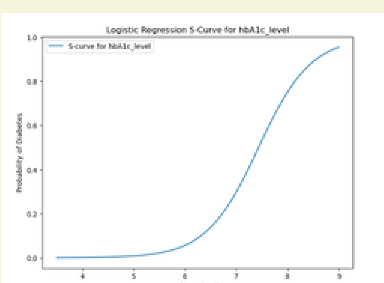
Logistic Regression: To model the likelihood of diabetes based on risk factors and calculate the significance of each variable.

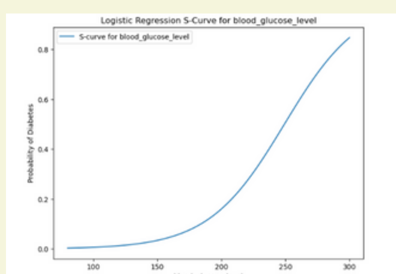
Logistic Regression model is created for each predictor as well.



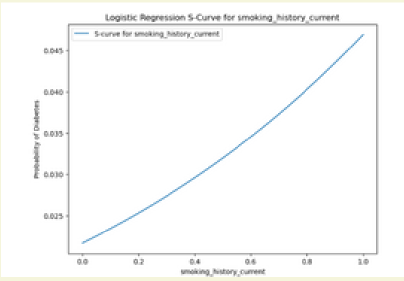


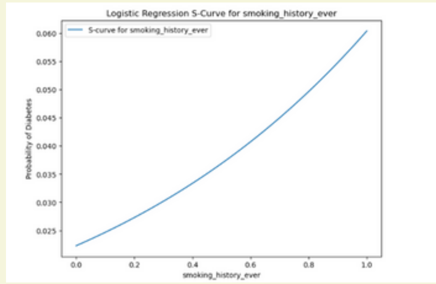


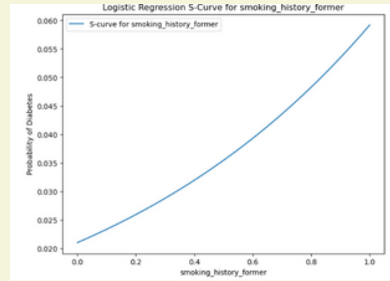


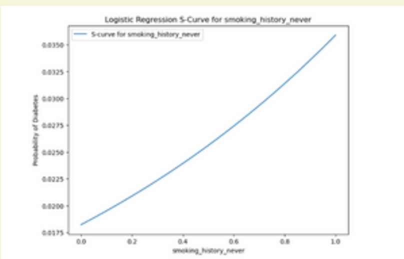


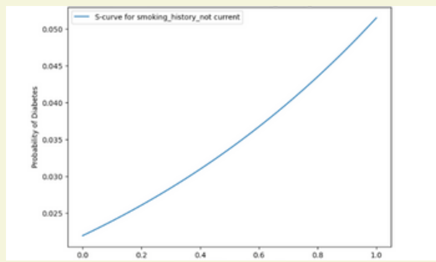
6. RESULTS











odd- Ratio's for race : African American

Optimization terminated successfully.
Current function value: 0.134693
Iterations 9

Logit Regression Results

Dep. Variable:	diabetes	No. Observations:	20223
Model:	Logit	Df Residuals:	20212
Method:	MLE	Df Model:	10
Date:	Sun, 08 Dec 2024	Pseudo R-squ.:	0.5458
Time:	19:53:55	Log-Likelihood:	-2723.9
converged:	True	LL-Null:	-5996.9
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-22.2877	0.479	-46.537	0.000	-23.226	-21.349
hypertension	1.0946	0.100	10.969	0.000	0.899	1.290
heart_disease	1.4649	0.130	11.279	0.000	1.210	1.719
bmi	0.0874	0.005	17.371	0.000	0.078	0.097
hbA1c_level	1.9608	0.061	31.938	0.000	1.840	2.081
blood_glucose_level	0.0337	0.001	33.332	0.000	0.032	0.036
smoking_history_current	0.8011	0.135	5.918	0.000	0.536	1.066
smoking_history_ever	1.0365	0.173	5.992	0.000	0.697	1.376
smoking_history_former	1.0739	0.125	8.623	0.000	0.830	1.318
smoking_history_never	0.6955	0.101	6.887	0.000	0.498	0.893
smoking_history_not current	0.8838	0.155	5.715	0.000	0.581	1.187

True Negatives (TN): 18321
False Positives (FP): 134
False Negatives (FN): 735
True Positives (TP): 1033

6. RESULTS

odd- Ratio’s for race : Asian

```
Optimization terminated successfully.
         Current function value: 0.130810
         Iterations 9
                           Logit Regression Results
==============================================================================
Dep. Variable:               diabetes   No. Observations:                20015
Model:                          Logit   Df Residuals:                    20004
Method:                           MLE   Df Model:                           10
Date:                Sun, 08 Dec 2024   Pseudo R-squ.:                  0.5577
Time:                        19:53:55   Log-Likelihood:                -2618.2
converged:                       True   LL-Null:                       -5919.2
Covariance Type:            nonrobust   LLR p-value:                     0.000
===============================================================================================
                                coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
const                       -22.3723      0.494    -45.266      0.000     -23.341     -21.404
hypertension                  1.1856      0.102     11.621      0.000       0.986       1.386
heart_disease                 1.5530      0.128     12.113      0.000       1.302       1.804
bmi                           0.0846      0.005     16.301      0.000       0.074       0.095
hbA1c_level                   1.9856      0.064     30.942      0.000       1.860       2.111
blood_glucose_level           0.0344      0.001     34.198      0.000       0.032       0.036
smoking_history_current       0.4120      0.145      2.846      0.004       0.128       0.696
smoking_history_ever          0.7517      0.180      4.177      0.000       0.399       1.104
smoking_history_former        0.9525      0.125      7.599      0.000       0.707       1.198
smoking_history_never         0.5431      0.098      5.549      0.000       0.351       0.735
smoking_history_not current   0.6160      0.154      3.988      0.000       0.313       0.919
===============================================================================================
True Negatives (TN): 18131
False Positives (FP): 141
False Negatives (FN): 708
True Positives (TP): 1035
```

odd- Ratio’s for race : Caucasian

```
Optimization terminated successfully.
         Current function value: 0.133431
         Iterations 9
                           Logit Regression Results
==============================================================================
Dep. Variable:               diabetes   No. Observations:                19876
Model:                          Logit   Df Residuals:                    19865
Method:                           MLE   Df Model:                           10
Date:                Sun, 08 Dec 2024   Pseudo R-squ.:                  0.5375
Time:                        19:53:55   Log-Likelihood:                -2652.1
converged:                       True   LL-Null:                       -5733.9
Covariance Type:            nonrobust   LLR p-value:                     0.000
===============================================================================================
                                coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
const                       -21.8708      0.490    -44.609      0.000     -22.832     -20.910
hypertension                  1.0854      0.102     10.620      0.000       0.885       1.286
heart_disease                 1.2365      0.133      9.286      0.000       0.976       1.498
bmi                           0.0700      0.005     13.831      0.000       0.060       0.080
hbA1c_level                   1.9712      0.065     30.303      0.000       1.844       2.099
blood_glucose_level           0.0342      0.001     33.971      0.000       0.032       0.036
smoking_history_current       0.5789      0.138      4.187      0.000       0.308       0.850
smoking_history_ever          0.8356      0.176      4.745      0.000       0.490       1.181
smoking_history_former        0.9075      0.128      7.115      0.000       0.658       1.158
smoking_history_never         0.6027      0.100      6.015      0.000       0.406       0.799
smoking_history_not current   0.6843      0.151      4.534      0.000       0.388       0.980
===============================================================================================
True Negatives (TN): 18082
False Positives (FP): 124
False Negatives (FN): 716
True Positives (TP): 954
```

6. RESULTS

odd- Ratio's for race : Hispanic

```
Optimization terminated successfully.
         Current function value: 0.129833
         Iterations 9
                           Logit Regression Results
==============================================================================
Dep. Variable:               diabetes   No. Observations:                19888
Model:                          Logit   Df Residuals:                    19877
Method:                           MLE   Df Model:                           10
Date:                Sun, 08 Dec 2024   Pseudo R-squ.:                  0.5509
Time:                        19:53:55   Log-Likelihood:                -2582.1
converged:                       True   LL-Null:                       -5749.2
Covariance Type:            nonrobust   LLR p-value:                     0.000
===============================================================================================
                                  coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
const                         -22.5589      0.500    -45.113      0.000     -23.539     -21.579
hypertension                    1.3124      0.099     13.213      0.000       1.118       1.507
heart_disease                   1.5484      0.128     12.066      0.000       1.297       1.800
bmi                             0.0892      0.005     17.797      0.000       0.079       0.099
hbA1c_level                     1.9896      0.065     30.409      0.000       1.861       2.118
blood_glucose_level             0.0342      0.001     33.526      0.000       0.032       0.036
smoking_history_current         0.5225      0.138      3.790      0.000       0.252       0.793
smoking_history_ever            0.5630      0.191      2.950      0.003       0.189       0.937
smoking_history_former          1.0094      0.126      8.027      0.000       0.763       1.256
smoking_history_never           0.5297      0.101      5.223      0.000       0.331       0.728
smoking_history_not current     0.7142      0.153      4.679      0.000       0.415       1.013
===============================================================================================
True Negatives (TN): 18053
False Positives (FP): 159
False Negatives (FN): 697
True Positives (TP): 979
```

Chi-Square Test: To evaluate the dependency of diabetes on race and gender.

```
Chi-Square Test for Gender: Male
Observed Frequencies Table:
                    Diabetes = 1  Diabetes = 0
race:AfricanAmerican         838          7576
race:Asian                   862          7435
race:Caucasian               783          7361
race:Hispanic                793          7491
race:Other                   763          7528

Expected Frequencies Table:
                    Diabetes = 1  Diabetes = 0
race:AfricanAmerican  820.278687   7593.721313
race:Asian            808.872387   7488.127613
race:Caucasian        793.956457   7350.043543
race:Hispanic         807.605021   7476.394979
race:Other            808.287449   7482.712551

Chi-Square Statistic: 7.5623
P-value: 0.1090
Degrees of Freedom: 4
Conclusion: Fail to reject the null hypothesis - Diabetes status is independent of race for Male.

Chi-Square Test for Gender: Female
Observed Frequencies Table:
                    Diabetes = 1  Diabetes = 0
race:AfricanAmerican         930         10877
race:Asian                   881         10835
race:Caucasian               887         10836
race:Hispanic                883         10718
race:Other                   880         10825

Expected Frequencies Table:
                    Diabetes = 1  Diabetes = 0
race:AfricanAmerican  899.559827  10907.440173
race:Asian            892.626657  10823.373343
race:Caucasian        893.159977  10829.840023
race:Hispanic         883.864958  10717.135042
race:Other            891.788581  10813.211419

Chi-Square Statistic: 1.4945
P-value: 0.8276
Degrees of Freedom: 4
Conclusion: Fail to reject the null hypothesis - Diabetes status is independent of race for Female.
```

6. RESULTS

Prevalence of Diabetes by Race

The analysis shows the percentage of individuals with diabetes within each racial group in the dataset. The prevalence rates are as follows:
African American: 8.74%
Asian: 8.71%
Caucasian:8.40%
Hispanic:8.43%
Other:8.22%
The diabetes prevalence rates across all racial groups are relatively similar, ranging between 8.2% and 8.7%. African Americans have the highest prevalence(8.74%), followed closely by Asians(8.71%).

Comparison of Diabetes Prevalence Between Genders within Each Race

Across all racial groups, males generally have a higher prevalence of diabetes compared to females. For example: Among African Americans, diabetes prevalence is 9.96% for males compared to 7.88% for females. Similarly, among Asians, the prevalence for males is 10.39%, while for females, it is 7.52%. This pattern of higher diabetes prevalence in males is consistent across all racial groups. Within each gender, the prevalence of diabetes varies slightly by race:
Among males, Asians have the highest prevalence (10.39%) followed by African Americans(9.96%) and Hispanics(9.57%). Among females, African Americans have the highest prevalence(7.88%), followed closely by Hispanics(7.61%) and Caucasians(7.57%). The results suggest that gender plays a role in diabetes prevalence, with males being more affected across all racial groups. However, the differences in prevalence between races are relatively small within each gender.

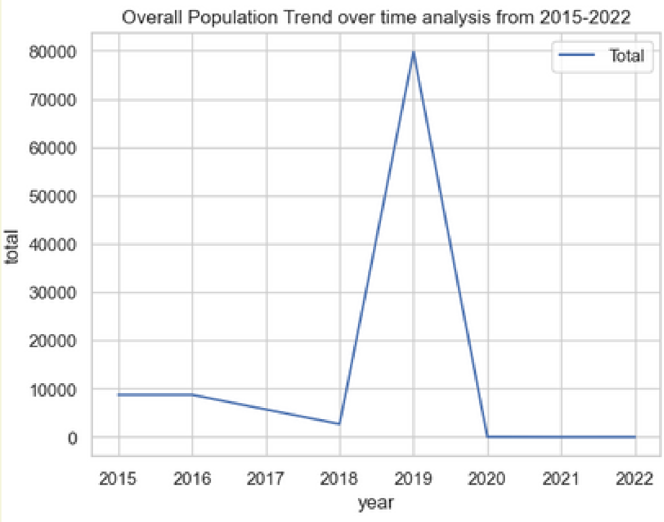
6. RESULTS

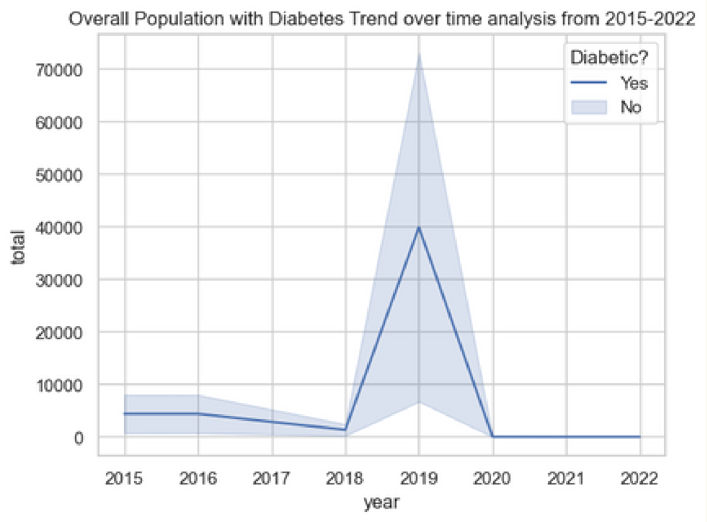
Bootstrapping: Confidence Intervals for Diabetes Proportion

The bootstrapping analysis provided a 95% confidence interval for the proportion of individuals with diabetes in the dataset:
Confidence Interval: [0.0832, 0.0867]

The analysis suggests that the proportion of individuals with diabetes in the dataset is likely between 8.32% and 8.67%, with 95% confidence. The result provides a reliable estimate of diabetes prevalence in the population represented by the dataset. These insights can be used to compare prevalence rates across subgroups or inform public health interventions. The bootstrapping results reinforce the datasets's reliability in estimating diabetes prevalence, offering a clear and precise range for further analysis and reporting.

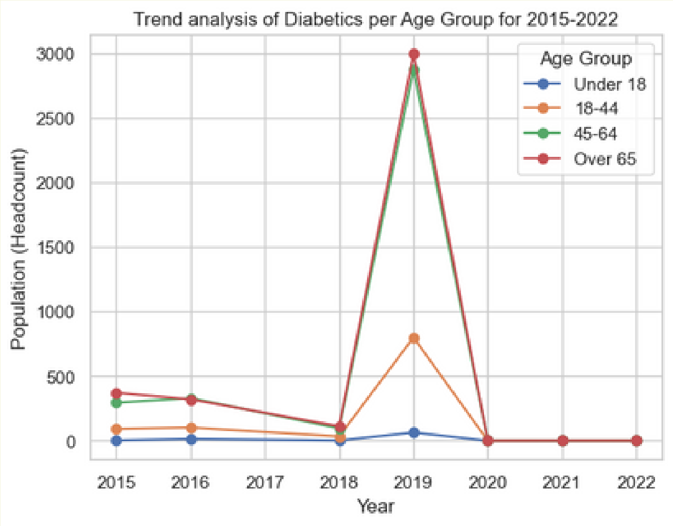
Time Trend Analysis

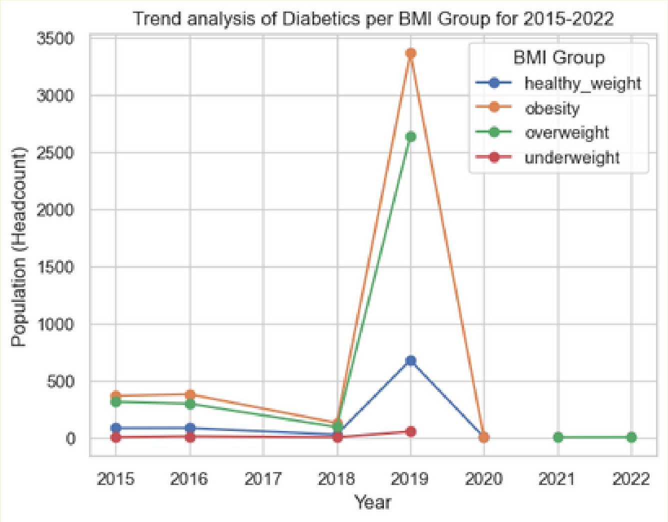


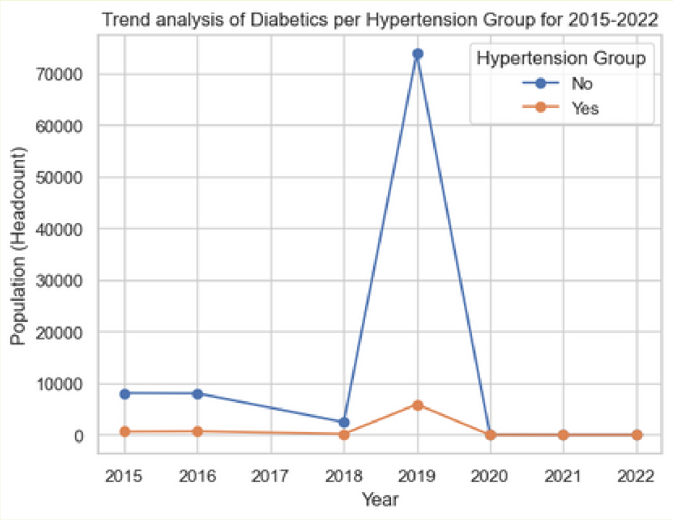


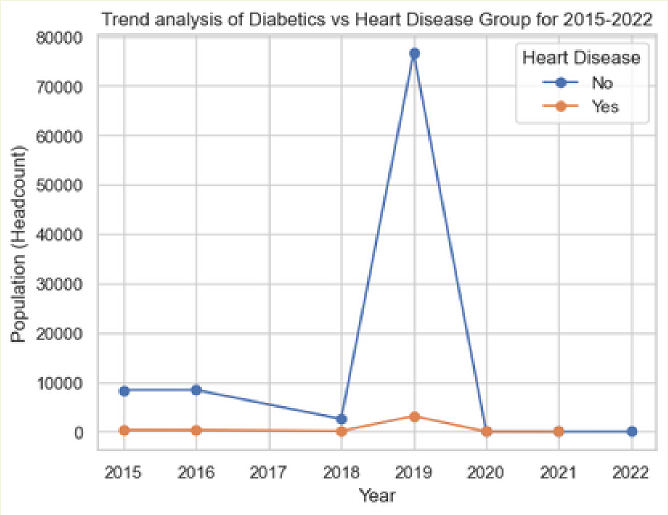
6. RESULTS

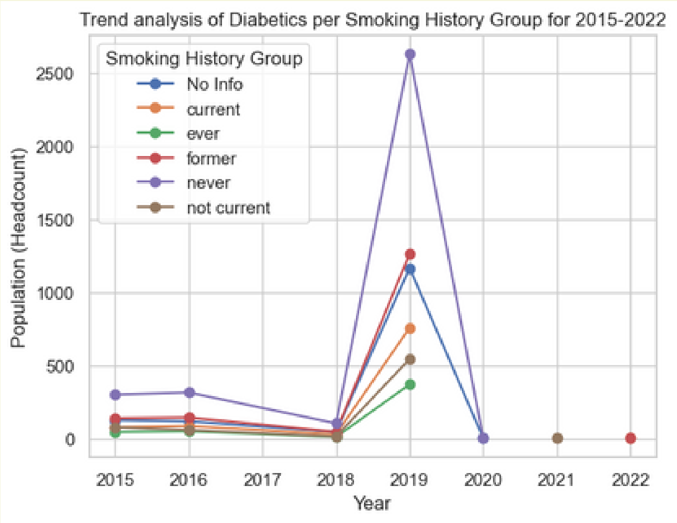
Time Trend Analysis

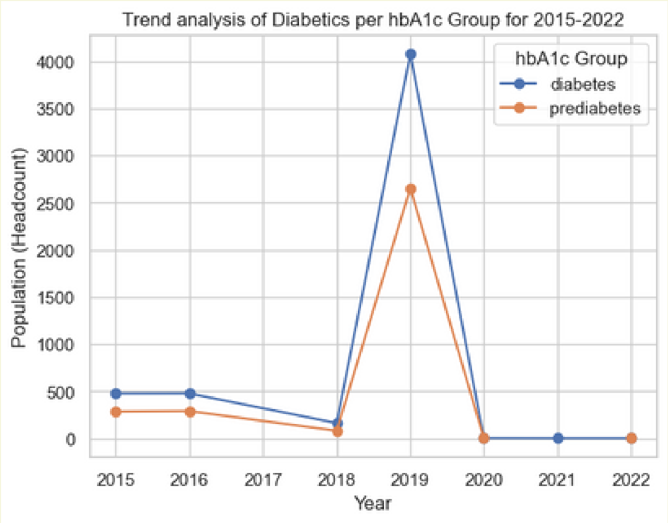












6. RESULTS

Conclusions:

High-Risk Groups :

- Age: Older populations ("45-64" and "Over 65")
- BMI: "Obesity" and "Overweight" groups.
- Health Conditions: Hypertension and heart disease play a secondary role.

Preventive Focus :

- Target populations with high BMI, older age, and comorbidities to prevent and manage diabetes effectively.