Towards An Analysis Of Factors Affecting Diabetes Across The U.S.

Addressed to: Dr Dylan George

Co-Authors:

Aman Thakur (02048327)
Sneh Pillai (02125179)
Tabeesh Fatiya (02119577)

Issues:

The Centre for Disease Control and Prevention (CDC) collects health data from state and local levels to group them together and analyze it, sharing the result back with the state and public to inform decisions that protect health.

This report used information from two sources: one from the CDC called the Behavioral Risk Factor Surveillance System (BRFSS), and another from the US Census Bureau's Population Estimates Program. They used this information to figure out how many people in different counties have diabetes, how many people were newly diagnosed with diabetes, how many people are obese, and how many people are not physically active in 2018.

We answer the following questions:

- Which factor (inactivity or obesity) has a stronger impact on the prevalence of diabetes?
- Can the data be used to inform public health policies targeting inactivity or obesity to reduce diabetes rates?
- Are there counties with unusually high or low diabetes rates given their inactivity and obesity levels?
- Do these outliers reveal any unique characteristics or interventions that might be contributing to their status?
- Can we use the inactivity and obesity percentage of counties to predict the diabetes percentage of a county?
- How accurate is the predictive model, and what are its limitations?
- Are there potential confounding variables that need to be considered?

Findings:

- We found that the level of inactivity has a stronger impact on diabetes compared to obesity. The data suggests that people being inactive is more closely related to the risk of diabetes. This conclusion is statistically supported through correlation coefficients and R-squared scores, indicating a more significant impact of inactivity on diabetes rates in our dataset.

- While our analysis provides valuable insights into the relationship between inactivity/obesity and diabetes, the intricacy of health as a factor demands a considerably broader and more nuanced understanding. Further investigations that consider genetic, socio-economic, and healthcare access factors are crucial towards developing targeted and effective public health policies. This comprehensive approach would work towards ensuring that interventions are not only evidence-based but also inclusive, addressing the broad variety of health determinants within our population.

- We identified approximately 12.4% of the total counties with higher-than-expected diabetes rates and 12.7% with lower-than-expected rates compared to predictions based on inactivity and obesity data. Further investigation, incorporating both quantitative and qualitative insights, is essential to identify the specific characteristics or interventions contributing to these deviations. This approach will inform targeted strategies for improving diabetes outcomes in these counties. Further investigation and collaboration with local stakeholders and health authorities is required to gain a more detailed understanding of the specific factors influencing diabetes in these outlier counties.

- Our analysis suggests that there is a relationship between inactivity and obesity percentages in counties and the diabetes percentage. The coefficients from the regression model indicate the direction and strength of this relationship. Inactivity and obesity percentages appear to be associated with deviations in the diabetes percentage.
  It seems that predicting diabetes percentages based on inactivity and obesity percentages is possible. Our analysis revealed a connection between these factors.
  How accurate is the predictive model?

- The model may be described as moderately accurate, explaining around 35% of the variation in diabetes rates in the counties using inactivity and obesity. The implication being that there are additional factors influencing diabetes rates that the model doesn't capture owing to .

- The model's accuracy is influenced by factors like the data not perfectly following a pattern (heteroskedasticity) and the residuals (the differences between predicted and actual values) not adhering perfectly to a normal distribution. These nuances in the data make predicting diabetes rates with absolute precision considerably challenging. Health outcomes are complex and influenced by numerous factors, and our model provides insights but acknowledges the inherent complexity of the issue.

Limitations:

- **Simplification**: The model oversimplifies the relationship between inactivity, obesity, and diabetes, potentially missing important nuances in real-world dynamics.
- **Generalization**: Its applicability to different populations or regions may be limited, as it's specifically trained on data from USA counties.
- **Limited Features**: The exclusion of factors like genetics, diet, and healthcare access may lead to an incomplete understanding of the major factors responsible for changes in diabetes rates.
- **Residual Distribution**: The assumption of normally distributed residuals is violated, casting doubt on the reliability of statistical inferences and predictions.

- **Causation**: While our model can reliably identify associations, it fails at establishing causation, limiting the ability to make conclusive statements about the impact of lifestyle factors on diabetes.

Discussion:

- The positive coefficients for inactivity and obesity suggest a positive association with diabetes rates. This successfully aligns with existing literature indicating a connection between sedentary behavior, obesity, and the prevalence of diabetes.
- The r-squared values indicate that your model explains a moderate portion of the variability in diabetes rates. This suggests that other factors beyond inactivity and obesity contribute to diabetes rates. Consider exploring and incorporating additional variables to improve the model's explanatory power.
- The presence of heteroskedasticity in the initial model suggests that the variability of the errors is not constant across all levels of the independent variables. The log transformation of diabetes and sqrt transformation of obesity seems to have addressed this issue to a certain extent.
- The positive associations between inactivity, obesity, and diabetes rates suggest that specific interventions targeting these lifestyle factors could potentially contribute to reducing diabetes rates. Public health policies promoting physical activity and healthy weight management could prove to be beneficial.
- The findings indicate room for further investigation. Consider exploring interactions between variables, incorporating more features, and examining the impact of potential outliers or influential data points. Like collaboration with local stakeholders and health authorities to gain a more detailed understanding of the specific factors influencing diabetes rates in these outlier counties.
- The accuracy of the predictive model can be described using a metric called (R-squared). It gives us a sense of how well the model explains the variability in diabetes rates based on inactivity and obesity. In our analysis, before transformation The R-squared value was around 0.35 for the training set and 0.27 for the test set. This suggests that the model explains about 35% of the variability in diabetes rates in the training set and 27% in the test set.
- The counties identified with higher-than-expected diabetes rates compared to predictions based on inactivity and obesity levels would require a comprehensive investigation to uncover unique characteristics or interventions contributing to their status. This analysis would involve reviewing existing literature and public health reports for each state, examining demographic and socioeconomic factors, evaluating healthcare infrastructure, considering cultural and lifestyle influences, and investigating public health interventions and policies targeting physical activity and diabetes prevention.

Appendix A: Method

Data Collection: Data on diagnosed diabetes, new cases, obesity, and inactivity at the county level were collected through a survey of adults aged 18 or older during the year 2018. Self-reports determined diabetes status, new cases, obesity (BMI ≥30), and physical inactivity.

Variable collection: The three variables in the data are “Inactivity Percentage”, “Obesity Percentage” and “Diabetes Percentage”. The dataset included county-level data for these variables, along with FIPS codes as identifiers. The process involved merging three separate tables based on FIPS codes to create a consolidated dataset with common data points.
Analytic methods:

Data Merging: Was performed by merging three tables (inactivity percentage, obesity percentage, and diabetes percentage) using a common identifier (FIPS).
This step provided access to a consolidated dataset with relevant information from all three tables.

Descriptive Statistics: Involved the calculation of a 5-point summary (minimum, 25th percentile, median, 75th percentile, and maximum) for inactivity, obesity, and diabetes percentages.
Created three separate histogram plots to visualize the distribution of each variable.
Identified that inactivity is a little left-skewed, obesity is more left-skewed, and diabetes is right-skewed.

Data Splitting: You split the data into training and testing sets (80% training, 20% testing) to evaluate your model's performance on unseen data.
Used random_state to ensure reproducibility.

Linear Regression Modeling:
Applied multiple linear regression with inactivity and obesity (X) as predictors and diabetes percentage (y) as the target variable. Obtained coefficients for the intercept, inactivity, and obesity. Fitted the model on the training set and evaluated its performance on both training and testing sets using R-squared values.

Residual Analysis: Calculated residuals for the training set. Conducted the Breusch-Pagan test to check for heteroskedasticity, which indicated evidence of heteroskedasticity with a p-value.

Cross-Validation: Applied 5-fold cross-validation to assess the model's generalizability.

Obtained an array of R2 values for each fold. Calculated the range, standard deviation, and interquartile range of the cross-validated R2 values.

Transformation to Address Heteroskedasticity: Applied log transformation to the diabetes data and square root transformation to the obesity data to address heteroskedasticity. Obtained new coefficients for the transformed model. Fitted the transformed model on the training set and evaluated its performance on the testing set.

Residual Analysis Post-Transformation: Calculated residuals for the training set post-transformation. Conducted the Breusch-Pagan test again, finding no evidence of heteroskedasticity (p-value: 0.23).

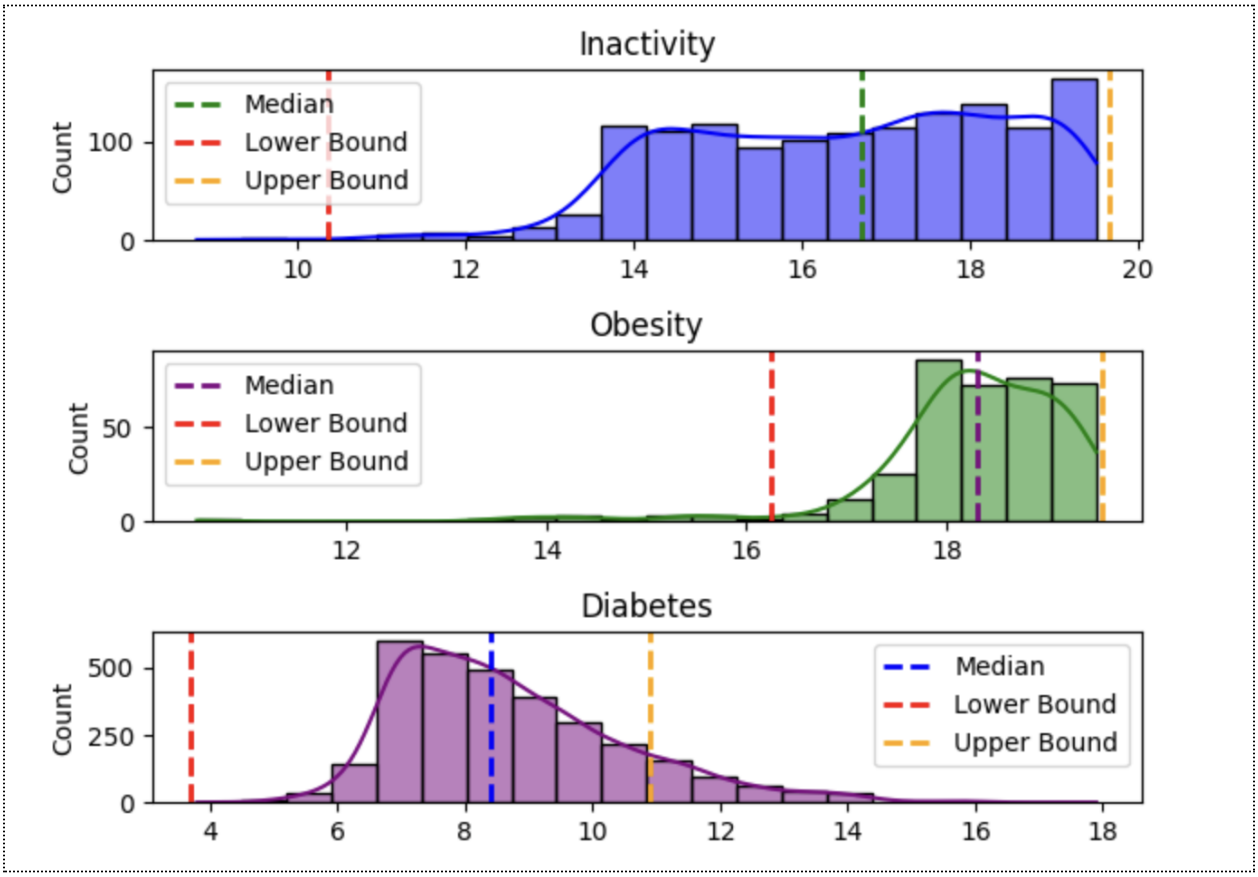
Cross-Validation Post-Transformation: Applied 5-fold cross-validation on the transformed model and assessed the range, standard deviation, and interquartile range of the cross-validated R2 values.

Appendix B: Results

The descriptive statistics reveal key characteristics of the data. Inactivity percentages displayed a left-skewed distribution (skewness: -0.342). Obesity percentages exhibited substantial skewness and kurtosis (skewness: -2.685, kurtosis: 12.322), while diabetes percentages were right-skewed (skewness: 0.974) and moderately kurtotic (kurtosis: 1.032).

```
First Quartile for inac  : 15.0
Median for inac  : 16.7
Third Quartile for inac  : 18.1
IQR for inac     : 3.1000000000000014
Lowerbound for inac      : 10.349999999999998
Upper bound for inac     : 19.650000000000002


First Quartile for obes  : 17.9
Median for obes  : 18.3
Third Quartile for obes  : 19.0
IQR for obes     : 1.1000000000000014
Lowerbound for obes      : 16.249999999999996
Upper bound for obes     : 19.55


First Quartile for diab  : 7.3
Median for diab  : 8.4
Third Quartile for diab  : 9.7
IQR for diab     : 2.3999999999999995
Lowerbound for diab      : 3.7000000000000006
Upper bound for diab     : 10.899999999999999
```

Figures such as histograms were used to visually convey these distributions.



Moving to the linear regression model, the initial model provided coefficients for inactivity (B1: 0.238) and obesity (B2: 0.102), indicating their impact on diabetes percentages.

```
Intercept: 1.7026414354405093
Coefficient for Inactivity (B1): 0.23778887846349625
Coefficient for Obesity (B2): 0.1024931241432882
```

The R-squared values were 0.3465 for the training set and 0.2657 for the testing set.
The Breusch-Pagan test suggested heteroskedasticity (P-value = 0.0118).

```
Breusch-Pagan Test Results:
P-value: 0.011803102323772647
Heteroskedasticity detected (reject null hypothesis)
```

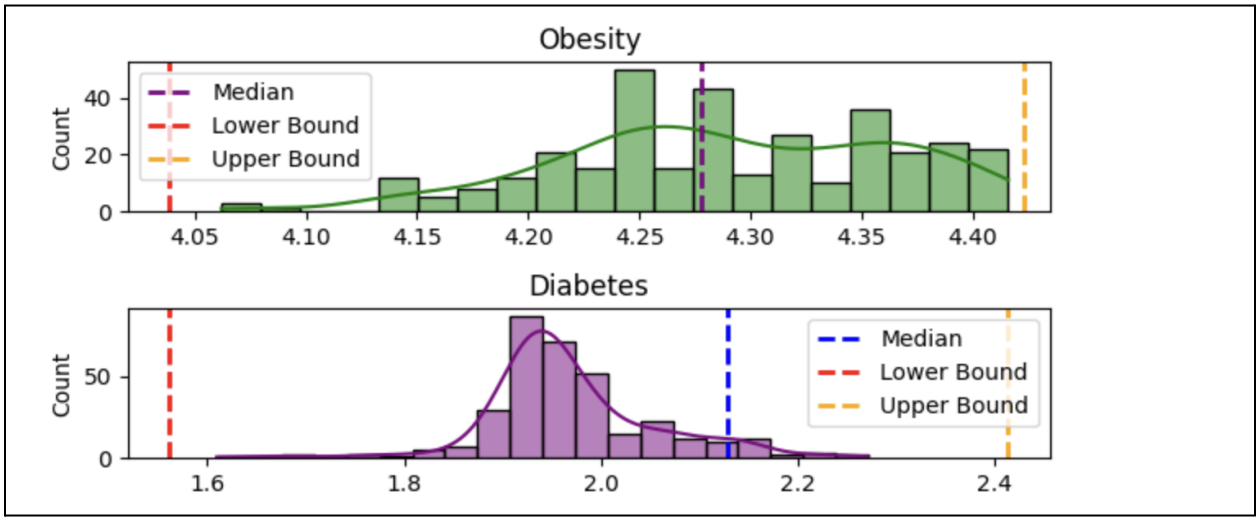
Cross-validation with a 5-fold approach yielded an R-squared range of 0.2924, standard deviation of 0.0940, and interquartile range of 0.0528.

```
Range of R-squared:  0.2924070839879689
Standard Deviation of R-squared:  0.0940412889655816
Interquartile Range of R-squared:  0.05280956747528154
```

To address heteroskedasticity, log-transformations on diabetes and square root-transformations on obesity were applied. The transformed model exhibited improved R-squared values (training: 0.3665, testing: 0.262) and showed no evidence of heteroskedasticity (Breusch-Pagan P-value = 0.2300).

```
Breusch-Pagan Test Results:
P-value: 0.23002700235252826
No evidence of heteroskedasticity
```

Histogram plot after the transformations are given below:



Cross-validation results mirrored those of the initial model.

Appendix C: Data and code

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
```

```
data_inactivity = pd.read_csv('inactivity.csv', usecols=['FIPS', '% INACTIVE']) #FIPDS
data_obesity = pd.read_csv('obesity.csv', usecols=['FIPS', '% OBESE'])
data_diabetes = pd.read_csv('diabetes.csv', usecols=['FIPS', '% DIABETIC',])

inac_ob = pd.merge(data_inactivity, data_obesity, on='FIPS', how='inner')

dataset = pd.merge(inac_ob, data_diabetes, on='FIPS', how='inner')
dataset
```

	FIPS	% INACTIVE	% OBESE	% DIABETIC
0	1011	17.0	18.7	9.4
1	2068	16.2	18.9	6.8
2	2105	15.0	19.4	7.3
3	2195	17.8	17.2	9.2
4	2230	15.8	18.3	6.6
...	...	...	...	...
349	51820	16.6	19.5	8.6
350	51830	15.7	18.0	8.5
351	51840	16.1	19.4	6.9
352	53055	11.9	19.3	4.5

```
obes_data = data_obesity.iloc[:, 1].values
diab_data = data_diabetes.iloc[:, 1].values
inac_data = data_inactivity.iloc[:, 1].values
```

```
First Quartile for inac  : 15.0
Median for inac  : 16.7
Third Quartile for inac  : 18.1
IQR for inac     : 3.1000000000000014
Lowerbound for inac      : 10.349999999999998
Upper bound for inac     : 19.650000000000002


First Quartile for obes  : 17.9
Median for obes  : 18.3
Third Quartile for obes  : 19.0
IQR for obes     : 1.1000000000000014
Lowerbound for obes      : 16.249999999999996
Upper bound for obes     : 19.55


First Quartile for diab  : 7.3
Median for diab  : 8.4
Third Quartile for diab  : 9.7
IQR for diab     : 2.3999999999999995
Lowerbound for diab      : 3.7000000000000006
Upper bound for diab     : 10.899999999999999
```

```python
import numpy as np
from scipy.stats import skew, kurtosis

data = np.random.randn(100)

skewness = skew(inac_data) #inactivity data
kurt = kurtosis(inac_data)

for i in ["inac","obes", "diab"]:
    skewness = skew(locals()[f"{i}_data"]) #inactivity data
    kurt = kurtosis(locals()[f"{i}_data"])
    print("Skewness "+ i , skewness)
    print("Kurtosis "+ i , kurt)
    print("\n")
```

```
Skewness inac -0.34204159975018034
Kurtosis inac -0.5490325254959423


Skewness obes -2.6850558229853996
Kurtosis obes 12.322509149363517


Skewness diab 0.9744494449218979
Kurtosis diab 1.0317351879435321
```

```python
fig, axes = plt.subplots(3, 1)

sns.histplot(inac_data, bins=20, kde=True, color='blue', ax=axes[0])
axes[0].axvline(median_inac, color='green', linestyle='dashed', linewidth=2, label='Median')
axes[0].axvline(lower_bound_inac, color='red', linestyle='dashed', linewidth=2, label='Lower Bound')
axes[0].axvline(upper_bound_inac, color='orange', linestyle='dashed', linewidth=2, label='Upper Bound')
axes[0].set_title('Inactivity')
axes[0].legend()

sns.histplot(obes_data, bins=20, kde=True, color='green', ax=axes[1])
axes[1].axvline(median_obes, color='purple', linestyle='dashed', linewidth=2, label='Median')
axes[1].axvline(lower_bound_obes, color='red', linestyle='dashed', linewidth=2, label='Lower Bound')
axes[1].axvline(upper_bound_obes, color='orange', linestyle='dashed', linewidth=2, label='Upper Bound')
axes[1].set_title('Obesity')
axes[1].legend()

sns.histplot(diab_data, bins=20, kde=True, color='purple', ax=axes[2])
axes[2].axvline(median_diab, color='blue', linestyle='dashed', linewidth=2, label='Median')
axes[2].axvline(lower_bound_diab, color='red', linestyle='dashed', linewidth=2, label='Lower Bound')
axes[2].axvline(upper_bound_diab, color='orange', linestyle='dashed', linewidth=2, label='Upper Bound')
axes[2].set_title('Diabetes')
axes[2].legend()

plt.tight_layout()
plt.show()
```

```
X = dataset.iloc[:, 1:-1].values
y = dataset.iloc[:, -1].values
```

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)
```

```
from sklearn.linear_model import LinearRegression
regressor = LinearRegression()
regressor.fit(X_train, y_train)
```

▾ LinearRegression

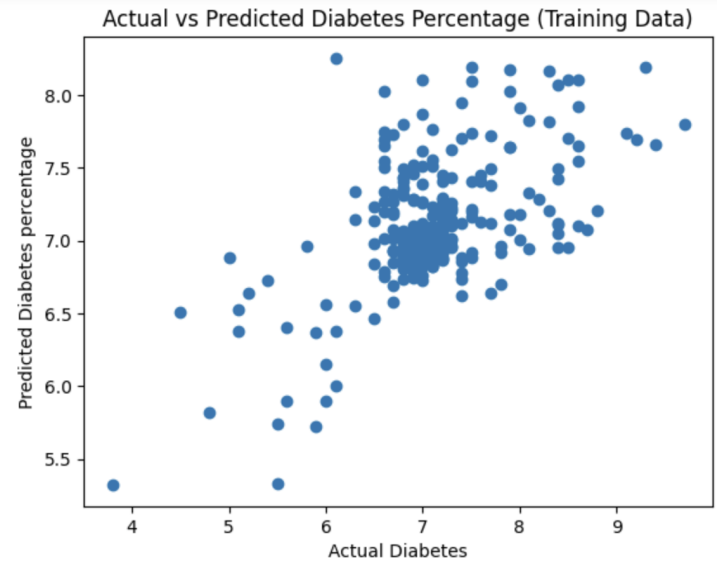
LinearRegression()

```
print("Intercept:", regressor.intercept_)
print("Coefficient for Inactivity (B1):", regressor.coef_[0])
print("Coefficient for Obesity (B2):", regressor.coef_[1])
```

```
Intercept: 1.7026414354405093
Coefficient for Inactivity (B1): 0.23778887846349625
Coefficient for Obesity (B2): 0.1024931241432882
```

```
y_pred_train = regressor.predict(X_train) # prediction for training set
y_pred_test = regressor.predict(X_test)
np.set_printoptions(precision=2)
# print(np.concatenate((y_pred.reshape(len(y_pred),1), y_test.reshape(len(y_test),1)),1))
```

```
import matplotlib.pyplot as plt

plt.scatter(y_train, y_pred_train)
plt.xlabel("Actual Diabetes")
plt.ylabel("Predicted Diabetes percentage")
plt.title("Actual vs Predicted Diabetes Percentage (Training Data)")
```

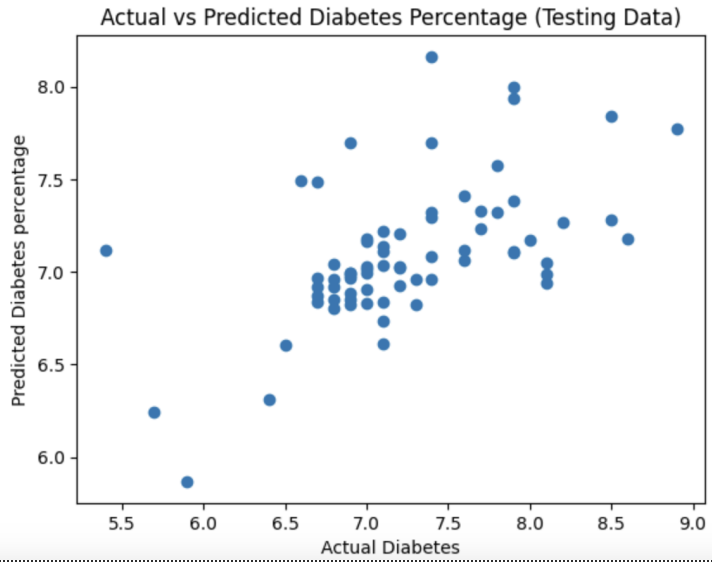



```
from sklearn.metrics import r2_score
r2_score(y_train, y_pred_train)
```

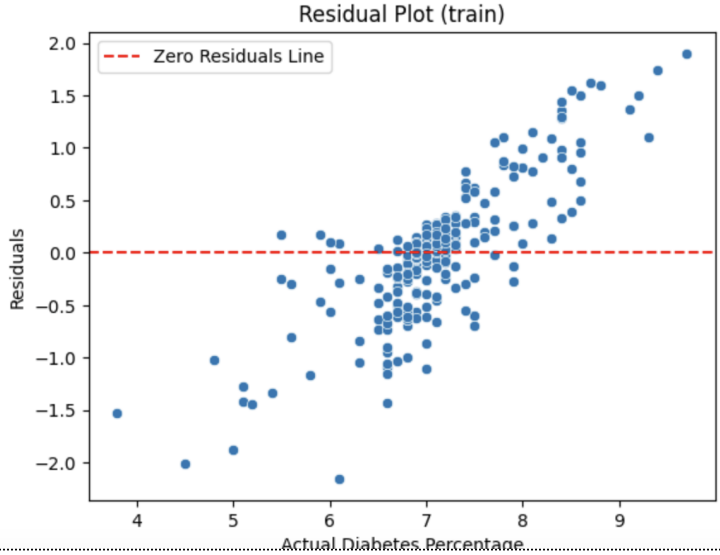
```
0.3465457795941539
```

```
plt.scatter(y_test, y_pred_test) #Testing scatter for testing set
plt.xlabel("Actual Diabetes")
plt.ylabel("Predicted Diabetes percentage")
plt.title("Actual vs Predicted Diabetes Percentage (Testing Data)")
```

```
Text(0.5, 1.0, 'Actual vs Predicted Diabetes Percentage (Testing Data)')
```




```
#Residual plot
residuals_train = y_train - y_pred_train
sns.scatterplot(x=y_train, y=residuals_train)
plt.axhline(y=0, color='r', linestyle='--', label='Zero Residuals Line')
plt.title('Residual Plot (train)')
plt.xlabel('Actual Diabetes Percentage')
plt.ylabel('Residuals')
plt.legend()
plt.show()
```

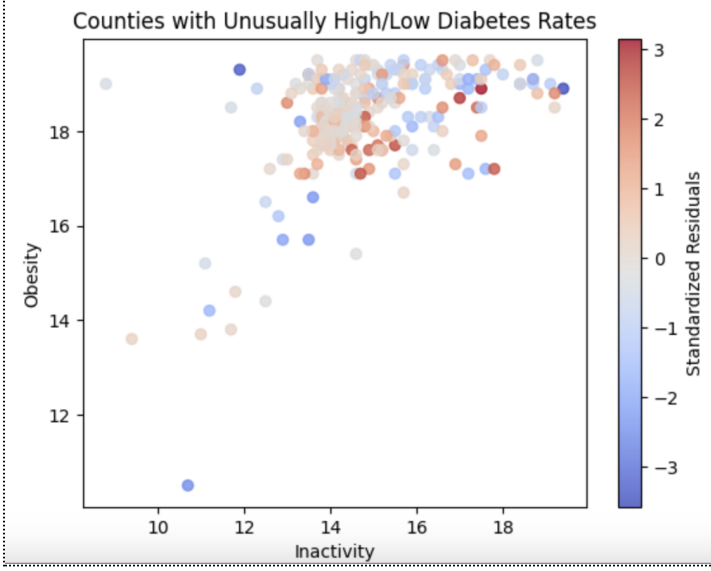



```
from sklearn.metrics import mean_squared_error
residuals_train_std = residuals_train / np.sqrt(mean_squared_error(y_train, y_pred_train))
```

```python
from scipy.stats import zscore

z_scores = zscore(residuals_train)
outliers = (np.abs(z_scores) > 1)

# Count of counties with higher than expected diabetes percentage (given inactivity and obesity)
count_higher = np.sum(residuals_train_std > 1)

count_lower = np.sum(residuals_train_std < -1)

total_count = len(residuals_train_std)

# Proportion of count_higher
proportion_higher = count_higher / total_count

# Proportion of count_lower
proportion_lower = count_lower / total_count

print("Count of counties higher than expected diabetes percentage (given inactivity and obesity):", count_higher)
print("Count of counties lower than expected diabetes percentage (given inactivity and obesity):", count_lower)
print("Total count of counties:", total_count)
print("Proportion of counties higher than expected:", proportion_higher)
print("Proportion of counties lower than expected:", proportion_lower)
```

```
Count of counties higher than expected diabetes percentage (given inactivity and obesity): 35
Count of counties lower than expected diabetes percentage (given inactivity and obesity): 36
Total count of counties: 283
Proportion of counties higher than expected: 0.12367491166077739
Proportion of counties lower than expected: 0.127208480565371
```

```python
plt.figure()
plt.scatter(X_train[:, 0], X_train[:, 1], c=residuals_train_std, cmap='coolwarm', alpha=0.8)
plt.colorbar(label='Standardized Residuals')
plt.xlabel('Inactivity')
plt.ylabel('Obesity')
plt.title('Counties with Unusually High/Low Diabetes Rates')
plt.show()
```



```
squared_residuals = residuals_train**2
```

```
from statsmodels.stats.diagnostic import het_breuschpagan
import statsmodels.api as sm
# Add a constant term to X_train for the intercept
X_train_with_constant = sm.add_constant(X_train)

# Perform Breusch-Pagan test
_, p_value, _, _ = sm.stats.diagnostic.het_breuschpagan(squared_residuals, X_train_with_constant)

print("Breusch-Pagan Test Results:")
print(f"P-value: {p_value}")

#Significance value
alpha = 0.05
if p_value < alpha:
    print("Heteroskedasticity detected (reject null hypothesis)")
else:
    print("No evidence of heteroskedasticity")
```

```
Breusch-Pagan Test Results:
P-value: 0.011803102323772647
Heteroskedasticity detected (reject null hypothesis)
```

```
from sklearn.model_selection import cross_val_score

# 5 fold cross validation
cross_val_r2 = cross_val_score(regressor, X_train, y_train, cv=5, scoring='r2')
cross_val_r2
```

```
array([0.46, 0.17, 0.28, 0.33, 0.31])
```

```
cv_range = np.max(cross_val_r2) - np.min(cross_val_r2)
print("Range of R-squared: ", cv_range)
cv_std = np.std(cross_val_r2)
print("Standard Deviation of R-squared: ", cv_std)
q75, q25 = np.percentile(cross_val_r2, [75 ,25])
cv_iqr = q75 - q25
print("Interquartile Range of R-squared: ", cv_iqr)
```

```
Range of R-squared:  0.2924070839879689
Standard Deviation of R-squared:  0.0940412889655816
Interquartile Range of R-squared:  0.05280956747528154
```

Transformation:

```
X = dataset.iloc[:, 1:-1].values
X[:, 1] = np.sqrt(X[:, 1])
y = np.log(dataset.iloc[:, -1].values)
```

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)
```

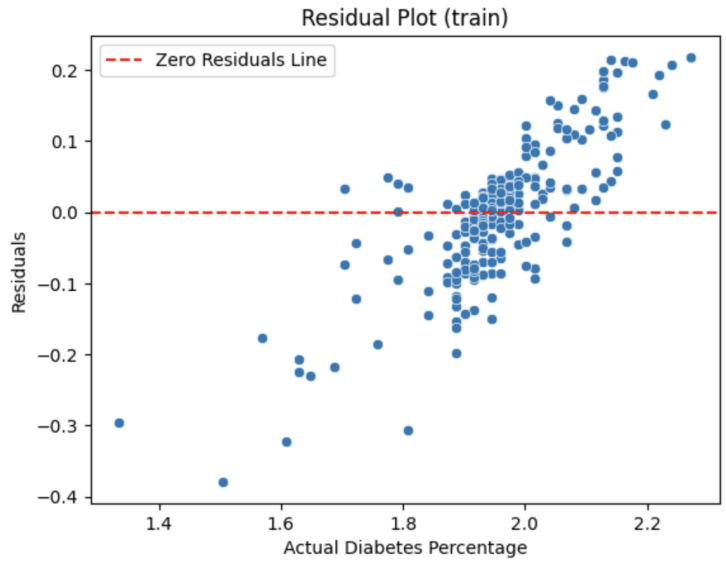
```
from sklearn.linear_model import LinearRegression
regressor = LinearRegression()
regressor.fit(X_train, y_train)
```

```
print("Intercept:", regressor.intercept_)
print("Coefficient for Inactivity (B1):", regressor.coef_[0])
print("Coefficient for Obesity (B2):", regressor.coef_[1])
```

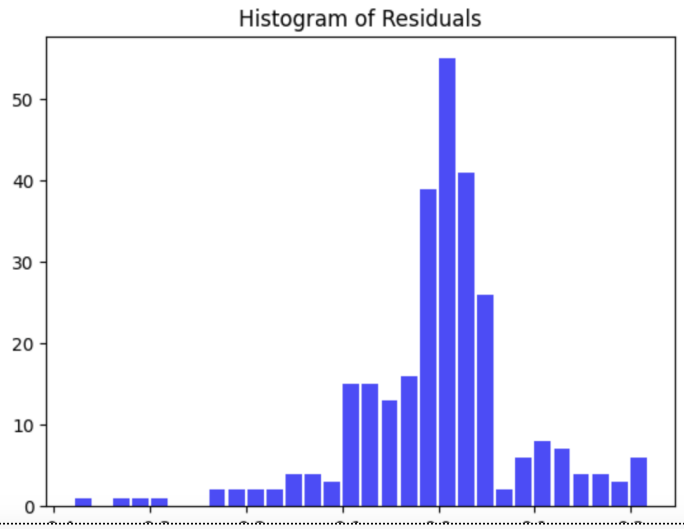
```
Intercept: 0.686564462589341
Coefficient for Inactivity (B1): 0.03203370607290505
Coefficient for Obesity (B2): 0.18562871685278498
```

```
from sklearn.metrics import r2_score
r2_score(y_train, y_pred_train)
```

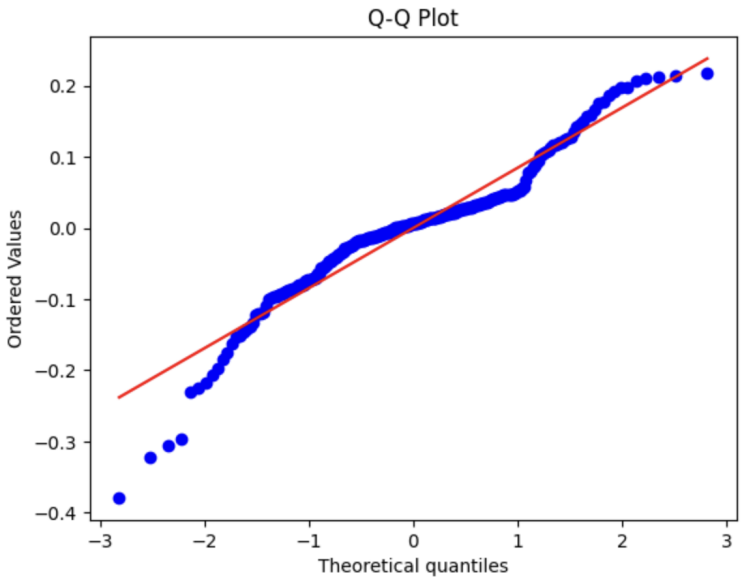
```
0.36493625545062025
```




```
import scipy.stats as stats
plt.hist(residuals, bins='auto', color='blue', alpha=0.7, rwidth=0.85)
plt.title('Histogram of Residuals')
plt.show()

# Q-Q plot
stats.probplot(residuals, dist="norm", plot=plt)
plt.title('Q-Q Plot')
plt.show()
```




```
residuals = y_train - y_pred_train
sns.scatterplot(x=y_train, y=residuals)
plt.axhline(y=0, color='r', linestyle='--', label='Zero Residuals Line')
plt.title('Residual Plot (train)')
plt.xlabel('Actual Diabetes Percentage')
plt.ylabel('Residuals')
plt.legend()
plt.show()
```




```python
#Check for Normality
from scipy.stats import shapiro

_, p_value = shapiro(residuals)
print("Shapiro-Wilk p-value:", p_value)
if p_value < 0.05:
    print("Not normal")
else:
    print("Normal")
```

```
Shapiro-Wilk p-value: 5.73020131344748e-10
Not normal
```

```python
from statsmodels.stats.diagnostic import het_breuschpagan
import statsmodels.api as sm
# Add a constant term to X_train for the intercept
X_train_with_constant = sm.add_constant(X_train)

# Perform Breusch-Pagan test
_, p_value, _, _ = sm.stats.diagnostic.het_breuschpagan(squared_residuals, X_train_with_constant)

print("Breusch-Pagan Test Results:")
print(f"P-value: {p_value}")

alpha = 0.05
if p_value < alpha:
    print("Heteroskedasticity detected (reject null hypothesis)")
else:
    print("No evidence of heteroskedasticity")
```

```
Breusch-Pagan Test Results:
P-value: 0.23002700235252826
No evidence of heteroskedasticity
```

Contributions:

All co-authors played an equal part towards the creation of the project.