Exploring the Interplay of Diabetes, Obesity, and Inactivity in US Counties

Kruthika Reddy Murthy - 02109003
Neeraj Reddy Neelapu – 02077978

1 The Issues

In our report, we delve into the health trends of 2018, with a special focus on diabetes, obesity, and physical inactivity across various counties in the USA. Our goal is to offer an easy-to-understand analysis of these health concerns, especially as they affect people with diabetes.
We've gathered our information from the Center for Disease Control and Prevention (CDC), a key source for up-to-date insights on diseases, especially their prevention. In our study, we aim to answer these important questions:

Link Between Health Issues: How are diabetes, obesity, and lack of physical activity related in U.S. counties?

Future Trends: Can we use the information about obesity and inactivity to predict future diabetes cases?

Data Reliability: How do we make sure our analysis is trustworthy, considering issues like unusual data points, missing information, and repeated data?

We've been careful to address these issues in our methods, making sure our conclusions are reliable and useful for understanding public health trends.

2 Findings

In our study, we explored the connection between diabetes rates and the rates of obesity and physical inactivity in different counties across the USA. Here's what we found:

We noticed a trend: places with higher obesity rates often have more cases of diabetes. Similarly, counties where people are less active also tend to have higher diabetes rates. This suggests that obesity and lack of exercise could be key factors in understanding diabetes in these areas.

By looking at the relationship between obesity and diabetes, and between inactivity and diabetes, we could predict where diabetes rates might increase. This is based on current levels of obesity and inactivity.

We made sure our analysis was solid by carefully dealing with any data issues like unusual data points, missing information, or repeated data. This helps us be confident in our findings.

3 Discussions

The discussion of our findings is presented in a straightforward manner, aiming to make it easily understandable. The positive correlations indicate that counties with higher obesity and inactivity rates tend to have higher diabetes rates. This observation emphasizes the need for a comprehensive approach to public health interventions, considering the interconnected nature of these factors.
In terms of implications, our results highlight the importance of addressing obesity and inactivity as preventive measures for diabetes. Public health initiatives targeting these lifestyle factors may contribute to reducing the prevalence of diabetes in different regions. The linear regression models used not only provide predictive insights but also emphasize the significance of obesity and inactivity in understanding diabetes rates.

4 Appendix A: Method

Data Collection:
The data was obtained from the Centers for Disease Control and Prevention (CDC). It provides county-level information on diabetes, obesity, and inactivity rates for the year 2018.

Variable Creation:
Diabetes Rate (% DIABETIC): This variable represents the percentage of the population in a given area diagnosed with diabetes, a condition characterized by high blood sugar levels.

Obesity Rate (% OBESE): This variable indicates the percentage of the population in an area classified as obese, defined as having a Body Mass Index (BMI) of 30 or higher.

Inactivity Rate (% INACTIVE): This variable shows the percentage of the population in a region reported as physically inactive, meaning they do not meet the recommended level of regular physical activity.

Analytic Methods
We used descriptive statistics to explore the data by using different methods as following:

- Data Preprocessing: Before analysis, the data undergoes preprocessing. This involves handling issues like duplicate records, missing values, and outliers. In your code, you handle outliers using the Interquartile Range (IQR) method and remove duplicates.

- Data Integration: The data from different sources (diabetes, obesity, and inactivity) are integrated by merging them based on common identifiers like "YEAR" and "FIPS" or "FIPDS," which represent the county codes.

- Exploratory Data Analysis (EDA): EDA is used to understand the data better. While not explicitly mentioned, this step may include visualizations, summary statistics, and checking for relationships between variables.

- Linear Regression: Linear regression is applied to analyze the relationship between two continuous variables. In this project, you perform linear regression for different combinations of variables, such as:
  % OBESE vs. % DIABETIC
  % INACTIVE vs. % DIABETIC
  % OBESE and % INACTIVE vs. % DIABETIC (Multiple Linear Regression)

- Train-Test Split: The data is split into training and testing sets to evaluate the performance of the regression models.

- Model Fitting: Linear regression models are created using the training data, and the coefficients (slopes and intercept) are estimated.

- Model Evaluation: The performance of the models is assessed using metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and hypothesis tests for heteroscedasticity. You also check the coefficients for significance.

- Kurtosis and Skewness: Kurtosis and skewness of variables are calculated. These statistics describe the shape and symmetry of the data distribution.

- Residual Analysis: The skewness, kurtosis, median, and standard deviation of residuals are examined to understand the model's predictive performance.

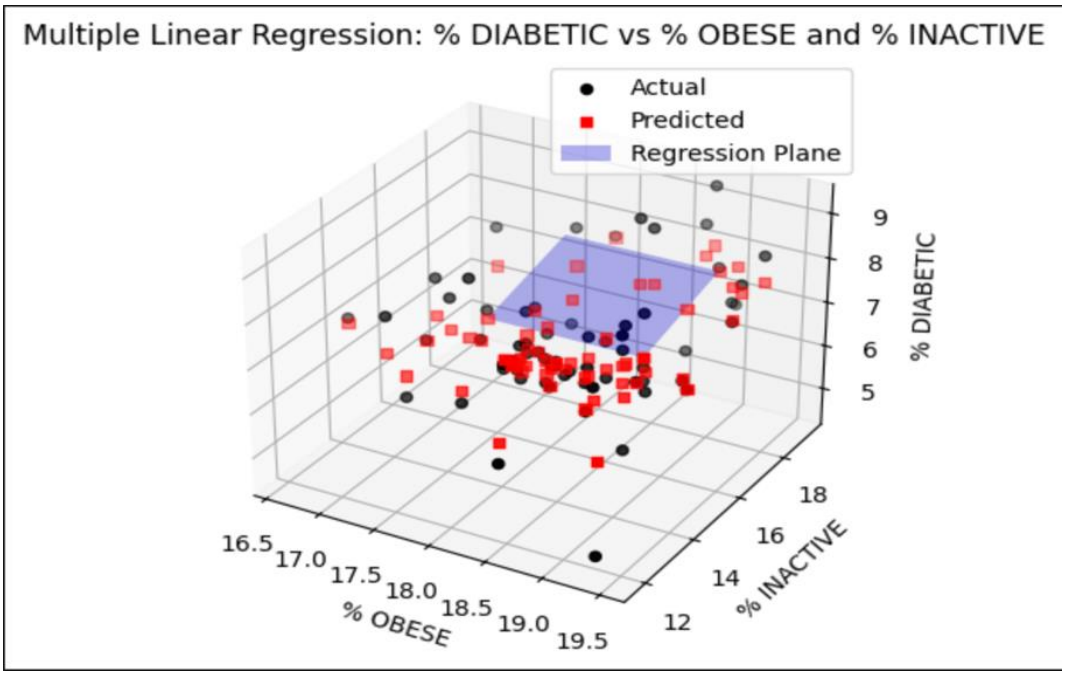
- Visualization: You create scatterplots of actual vs. predicted values and regression lines. For multiple linear regression, you visualize the regression plane in 3D space.

- Statistical Tests: You conduct statistical tests such as the Breusch-Pagan test for heteroscedasticity.

5 Appendix B: Results

Our dataset initially consisted of 3,142 data points. After merging relevant datasets, the number of data points was reduced to 354. These data points included information about the percentage of diabetic individuals (%DIABETIC), the percentage of obese individuals (%OBESE), and the percentage of inactive individuals (%INACTIVE) for all U.S. counties in the year 2018.

Prior to performing data transformation and preprocessing steps, we had 354 data points. Following these steps, the number of data points further decreased to 339.



The 3D scatter plot displays actual versus predicted diabetes rates against obesity and inactivity rates. The red dots represent actual observed values, and the blue plane shows the predicted values from the multiple linear regression model. The closer the dots to the plane, the more accurate the predictions.

OLS(ORDINARY LEAST SQUARES) REGRESSION RESULTS FOR DIABETIC VS OBESE AND INACTIVE MODEL:

OLS Regression Results

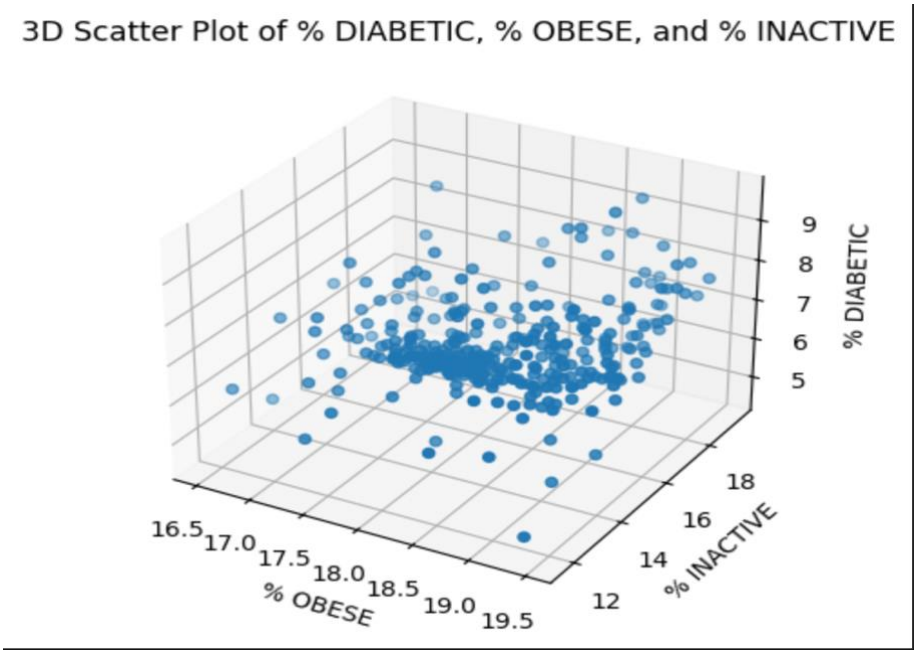
Dep. Variable:	% DIABETIC	R-squared:	0.164
Model:	OLS	Adj. R-squared:	0.157
Method:	Least Squares	F-statistic:	26.21
Date:	Mon, 09 Oct 2023	Prob (F-statistic):	4.00e-11
Time:	12:20:17	Log-Likelihood:	-231.57
No. Observations:	271	AIC:	469.1
Df Residuals:	268	BIC:	480.0
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.2446	0.990	4.286	0.000	2.295	6.195
% OBESE	0.0063	0.058	0.110	0.913	-0.107	0.120
% INACTIVE	0.1881	0.028	6.693	0.000	0.133	0.243

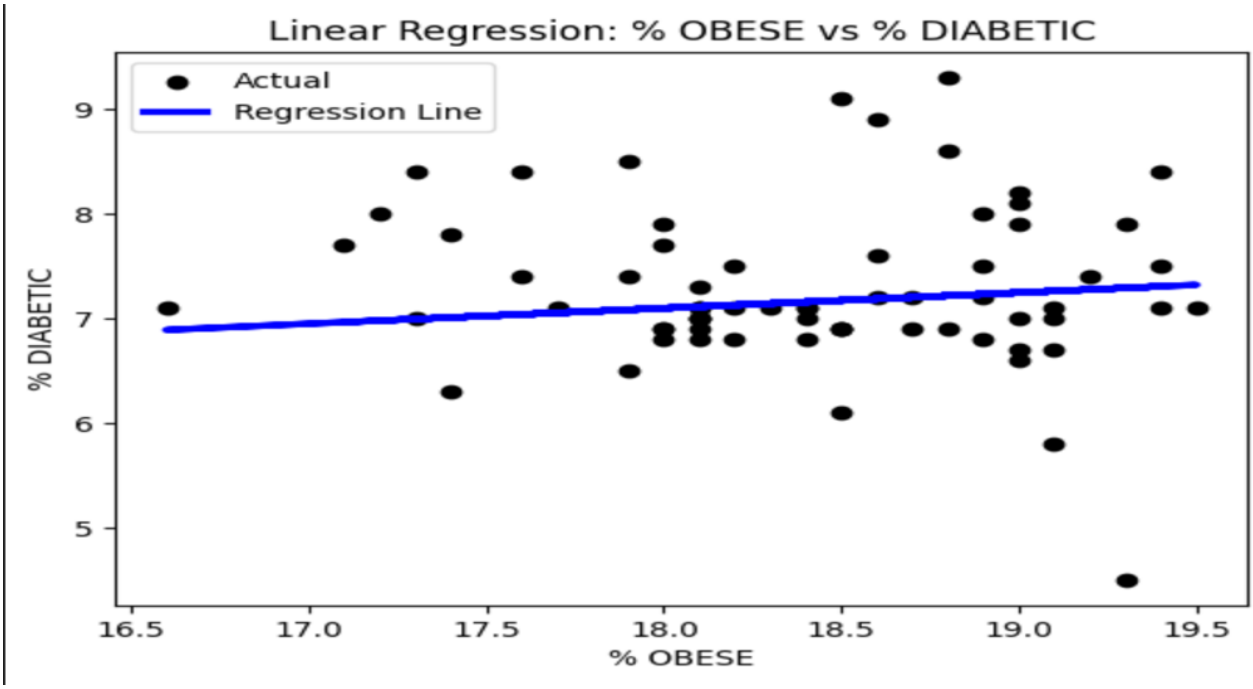
...

The results from the regression analysis indicate that both obesity and inactivity have a statistically significant relationship with diabetes rates. The constant coefficient suggests a baseline diabetes rate, with obesity and inactivity contributing to variations in rates across counties. Inactivity seems to have a stronger relationship with diabetes rates compared to obesity, as indicated by its larger coefficient and significant p-value. However, the R-squared value suggests that only a portion of the variation in diabetes rates can be explained by these factors alone.

CORRELATION ANALYSIS:



The 3D scatter plot in Figure 1.2 suggests a relationship between diabetes, obesity, and inactivity rates across counties. The blue dots, which represent county data, are spread out, indicating variability in how these health issues correlate. The plot illustrates that as obesity and inactivity rates increase, diabetes rates tend to rise as well. The distribution of data points suggests variability in how these factors interact across different regions, highlighting the complexity of their relationship.



The scatter plot shows the actual diabetes rates in relation to obesity rates for various counties. The blue line represents the regression line — a line of best fit calculated by the linear regression model. The plot suggests a positive relationship between obesity rates and diabetes rates, as indicated by the slope of the regression line. However, the data points are quite spread out around the regression line, which implies that the predictive power of obesity rates on diabetes rates, while existent, is not very strong.

OLS REGRESSION RESULTS FOR OBESE VS DIABETIC MODEL

OLS Regression Results

Dep. Variable:	% DIABETIC	R-squared:	0.024
Model:	OLS	Adj. R-squared:	0.020
Method:	Least Squares	F-statistic:	6.561
Date:	Mon, 09 Oct 2023	Prob (F-statistic):	0.0110
Time:	12:20:23	Log-Likelihood:	-252.52
No. Observations:	271	AIC:	509.0
Df Residuals:	269	BIC:	516.2
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.4232	1.068	4.143	0.000	2.321	6.525
% OBESE	0.1485	0.058	2.561	0.011	0.034	0.263

Omnibus:	45.872	Durbin-Watson:	2.149
Prob(Omnibus):	0.000	Jarque-Bera (JB):	118.701
Skew:	0.771	Prob(JB):	1.68e-26
Kurtosis:	5.852	Cond. No.	526.

The regression output indicates a positive association between obesity rates (% OBESE) and diabetes rates (% DIABETIC), with a coefficient of 0.1485. This means that for each percentage point increase in the obesity rate, the diabetes rate increases by approximately 0.1485 percentage points. However, the R-squared value is very low (0.024), meaning that obesity rates alone do not explain much of the variation in diabetes rates. The model's predictive power is limited, suggesting that other factors beyond obesity also play a significant role in diabetes rates.

CORRELATION ANALYSIS:

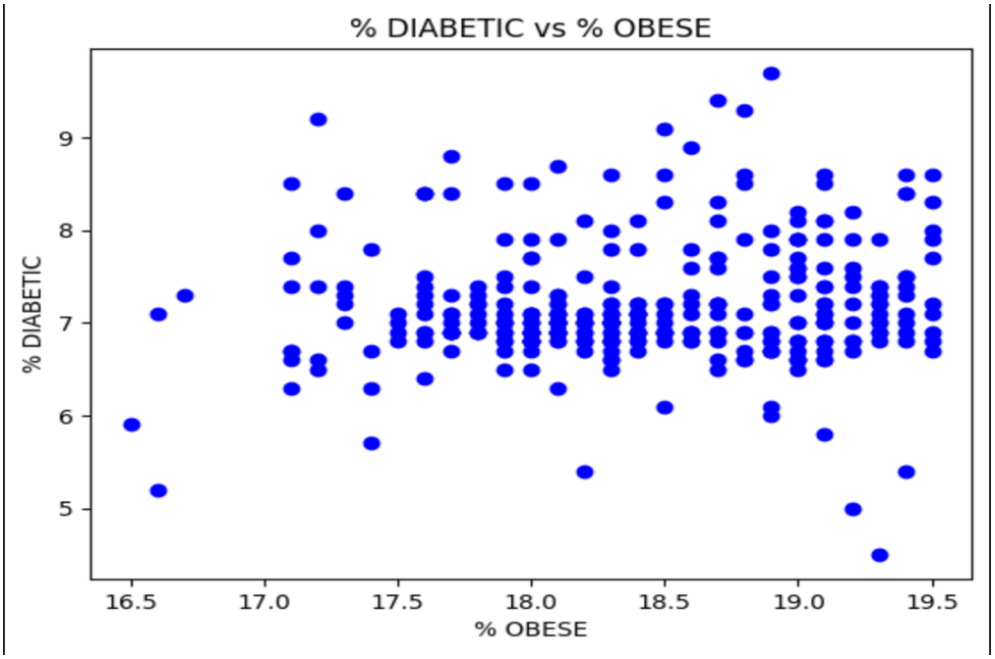
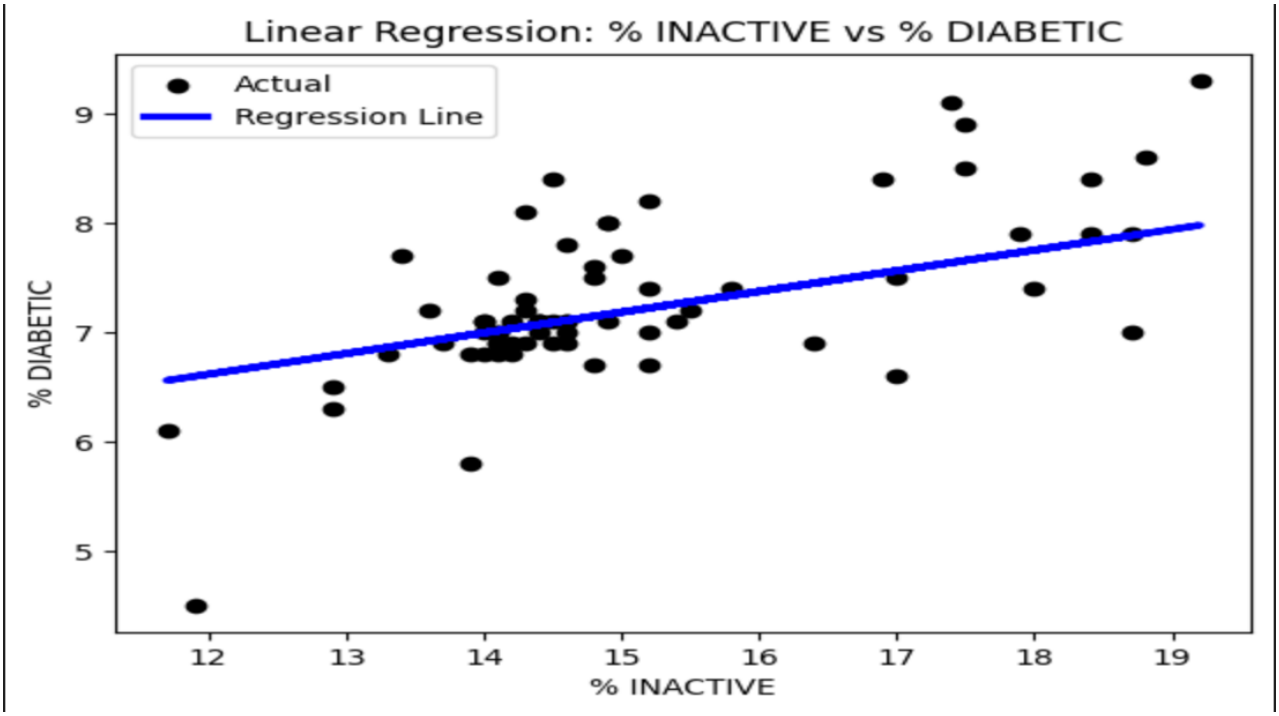


Figure 2.2 Correlation analysis of %OBESE VS %DIABETIC

The plot shows a positive correlation between the two variables, meaning that countries with a higher percentage of obese people also tend to have a higher percentage of diabetic people. This is likely due to a number of factors, including the fact that obesity is a major risk factor for type 2 diabetes.

The plot also shows a large variation in the data, with some countries having a much higher percentage of obese and diabetic people than others. This variation is likely due to a number of factors, including differences in diet, lifestyle, and access to healthcare.



The scatter plot shows the relationship between the percentage of inactive population (% INACTIVE) and the percentage of population with diabetes (% DIABETIC). The blue line represents the linear regression model's prediction of how inactivity rates might predict diabetes rates. The positive slope of the regression line suggests that as inactivity rates increase, diabetes rates also tend to increase. The data points are quite close to the regression line, which indicates a stronger relationship between inactivity and diabetes compared to obesity and diabetes, as seen in previous analyses.

OLS REGRESSION RESULTS FOR INACTIVITY VS DIABETIC MODEL

OLS Regression Results

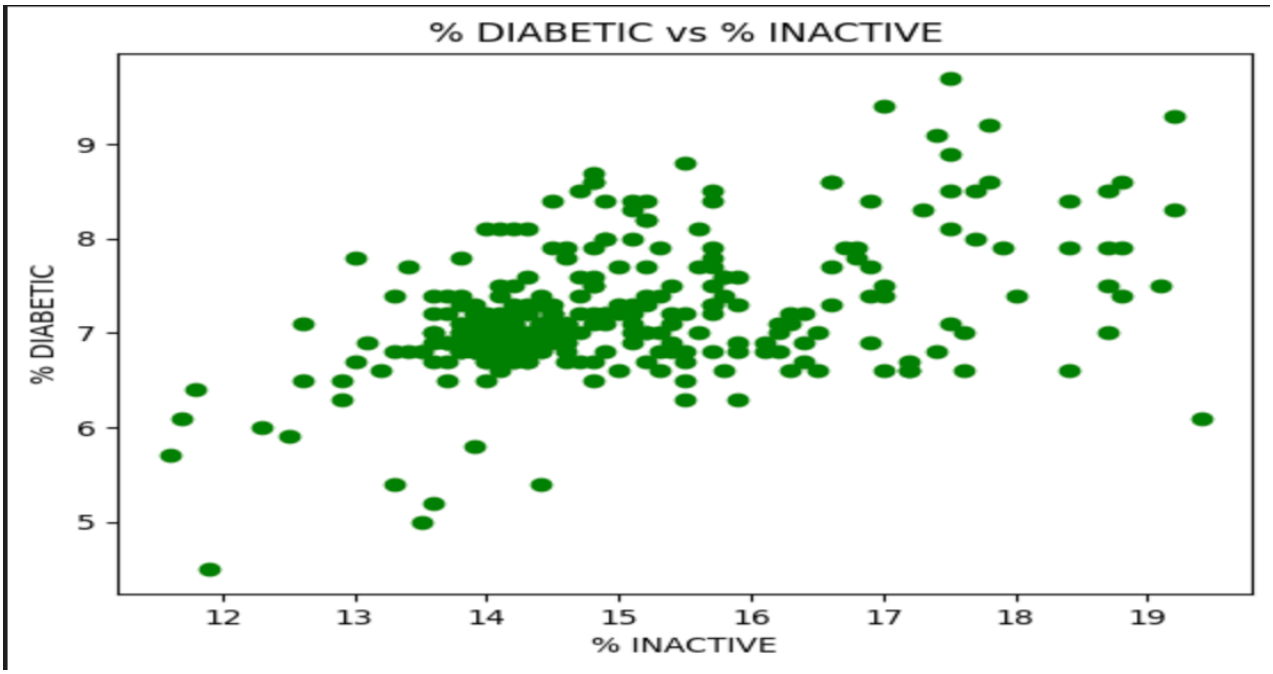
Dep. Variable:	% DIABETIC	R-squared:	0.164
Model:	OLS	Adj. R-squared:	0.160
Method:	Least Squares	F-statistic:	52.61
Date:	Mon, 09 Oct 2023	Prob (F-statistic):	4.34e-12
Time:	12:20:32	Log-Likelihood:	-231.58
No. Observations:	271	AIC:	467.2
Df Residuals:	269	BIC:	474.4
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.3445	0.389	11.164	0.000	3.578	5.111
% INACTIVE	0.1892	0.026	7.253	0.000	0.138	0.241

Omnibus:	21.526	Durbin-Watson:	2.058
Prob(Omnibus):	0.000	Jarque-Bera (JB):	54.653
Skew:	0.317	Prob(JB):	1.36e-12
Kurtosis:	5.106	Cond. No.	168.

The regression output presents an analysis of the relationship between the inactivity rate (% INACTIVE) and the diabetes rate (% DIABETIC). The coefficient for % INACTIVE is 0.1892, which suggests that as the inactivity rate increases by one percentage point, the diabetes rate is predicted to increase by roughly 0.1892 percentage points. The R-squared value is 0.164, indicating that approximately 16.4% of the variability in diabetes rates can be explained by the inactivity rate alone. The p-value is very small (almost 0), which shows a statistically significant relationship between these two variables. The model also suggests some presence of heavier tails in the data distribution, as indicated by the kurtosis value (5.106).

CORRELATION ANALYSIS:



This scatter plot illustrates the relationship between the percentage of the population that is inactive (% INACTIVE) and the percentage with diabetes (% DIABETIC). The green dots, which represent individual data points, show a general upward trend, also shows positive correlation between the percentage of people with diabetes and the percentage of people who are inactive, indicating that higher inactivity rates might be associated with higher diabetes rates. However, the spread of the dots suggests that while there is a trend, it is not a perfectly linear relationship and other factors may also play a role in determining diabetes prevalence.

6 Appendix C: Data and Code

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn import metrics
from sklearn.metrics import mean_squared_error
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import kurtosis, skew
from statsmodels.stats.diagnostic import het_breuschpagan
import statsmodels.api as sm
from sklearn.preprocessing import PolynomialFeatures
from sklearn.linear_model import LinearRegression
import scipy.stats as stats
# Specify the path to your Excel file
excel_file_path = r'C:\\Users\\Kruthika reddy\\Desktop\\mth\\cdc-diabetes-2018 (1).xlsx'
# Read data from Excel sheets
df_obesity = pd.read_excel(excel_file_path, sheet_name='Obesity')
df_diabetic = pd.read_excel(excel_file_path, sheet_name='Diabetes')
df_inactive = pd.read_excel(excel_file_path, sheet_name='Inactivity')

# Before  preprocessing and transformation of data
# Merge the datasets on FIPS code and YEAR
print('df_diabetic:', df_diabetic.shape)
print('df_obesity:', df_obesity.shape)
df_merged = pd.merge(df_diabetic, df_obesity, on=['YEAR', 'FIPS'])
print('After First Merge:', df_merged.shape)
df_merged = pd.merge(df_merged, df_inactive, left_on=['YEAR', 'FIPS'], right_on=['YEAR',
'FIPDS'])
print('After Second Merge:', df_merged.shape)
print('Number of Rows and Columns in df_merged:', df_merged.shape)

# After preprocessing and transformation of data
# Function to handle outliers using IQR
def handle_outliers(df, column_name):
    Q1 = df[column_name].quantile(0.25)
    Q3 = df[column_name].quantile(0.75)
    IQR = Q3 - Q1

    # Filtering values between Q1-1.5*IQR and Q3+1.5*IQR
    df_no_outliers = df[(df[column_name] >= Q1 - 1.5 * IQR) & (df[column_name] <= Q3 + 1.5
* IQR)]

    return df_no_outliers

# Function to preprocess and transform the data
```

```
def preprocess_and_transform(df):
    # Check for and handle duplicates
    df = df.drop_duplicates()

    # Check for missing values and handle them (if needed)
    df = df.dropna()  # Uncomment this line if you want to remove rows with missing values

    # Handle outliers
    df = handle_outliers(df, df.columns[4])  # Assuming the numerical column is at index 4

    # Add any additional preprocessing or transformation steps here

    return df

# Preprocess and transform each dataset
df_obesity = preprocess_and_transform(df_obesity)
df_diabetic = preprocess_and_transform(df_diabetic)
df_inactive = preprocess_and_transform(df_inactive)


# Plot the actual data points
ax.scatter(X_test_diabetic_vs_obese_inactive['% OBESE'],
           X_test_diabetic_vs_obese_inactive['% INACTIVE'],
           y_test_diabetic_vs_obese_inactive, c='black', marker='o', label='Actual')

# Plot the predicted values
ax.scatter(X_test_diabetic_vs_obese_inactive['% OBESE'],
           X_test_diabetic_vs_obese_inactive['% INACTIVE'],
           y_pred_diabetic_vs_obese_inactive, c='red', marker='s', label='Predicted')

# Create a meshgrid for the plane
grid_x, grid_y = np.meshgrid(X_test_diabetic_vs_obese_inactive['% OBESE'],
                             X_test_diabetic_vs_obese_inactive['% INACTIVE'])
grid_z = (model_diabetic_vs_obese_inactive.coef_[0] * grid_x +
          model_diabetic_vs_obese_inactive.coef_[1] * grid_y +
          model_diabetic_vs_obese_inactive.intercept_)

# Plot the regression plane
ax.plot_surface(grid_x, grid_y, grid_z, alpha=0.3, rstride=100, cstride=100, color='blue',
label='Regression Plane')
ax.set_xlabel('% OBESE')
ax.set_ylabel('% INACTIVE')
ax.set_zlabel('% DIABETIC')
ax.set_title('Multiple Linear Regression: % DIABETIC vs % OBESE and % INACTIVE')

plt.legend()
```

```
plt.show()

# Create an OLS (Ordinary Least Squares) model
ols_model_multiple_regression = sm.OLS(y_train_diabetic_vs_obese_inactive,
X_with_constant_multiple_regression)

# Fit the OLS model
ols_results_multiple_regression = ols_model_multiple_regression.fit()

# Print the summary which contains p-values
print(ols_results_multiple_regression.summary())


import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D

# Assuming df_merged contains the DataFrame with columns '% DIABETIC', '% OBESE', '%
INACTIVE'
fig = plt.figure()
ax = fig.add_subplot(111, projection='3d')
# Scatter plot
ax.scatter(df_merged['% OBESE'], df_merged['% INACTIVE'], df_merged['% DIABETIC'])
# Set labels
ax.set_xlabel('% OBESE')
ax.set_ylabel('% INACTIVE')
ax.set_zlabel('% DIABETIC')
# Set title
ax.set_title('3D Scatter Plot of % DIABETIC, % OBESE, and % INACTIVE')
plt.show()
```

FULL CODE: colab link-
https://colab.research.google.com/drive/1Fn1ZMAqXBDbQ1OoJLhMQDhKS_yTLWBnh?usp=sharing

7 Author Contributions

In this project, both authors made significant contributions. Being responsible for data collection, preprocessing, and conducting the linear regression analysis. Contributing to data analysis, interpretation of results, and visualizations, including the creation of graphs and charts. The two authors collaborated closely in merging datasets, performing statistical tests, and discussing the implications of the findings. Their combined efforts ensured a comprehensive and insightful analysis of the relationships between diabetes, obesity, and physical inactivity across U.S. counties in 2018.

8 Reference:

1. Reference: “An Introduction to Statistical Learning with Applications in Python”, Chapter 4.
2. Applied Logistic Regression, Hosmer & Lemeshow.