Examining the Triad: Diabetes, Sedentary Lifestyles, and Obesity - Insights from the CDC's 2018 Data

THE ISSUES:

In 2018, the CDC (Center for Disease Control and Prevention) compiled data highlighting the statistics related to Diabetes, Physical Inactivity, and Obesity. The overarching question this spark is whether there is an intricate relationship between the rising incidences of Diabetes and the prevailing levels of Obesity and Inactivity. This study embarks on a mission to address:

1. Given the model's significant predictive power, how might external factors not captured in the dataset influence diabetes rates, and to what extent can these predictive models be considered robust tools for health policymakers?
2. To what extent can more complex models, such as multiple linear regression, enhance our understanding by controlling for various confounding factors when analyzing the relationship between diabetes, obesity, and physical inactivity?
3. Given the observed relationship between regions of increased physical inactivity and higher rates of obesity and diabetes, how can we determine the extent to which sedentary behavior acts as a primary contributing factor?
4. Does data signify a tangible relationship between rates of diabetes, obesity, and lack of physical activity?
5. When observed county-wise, is there a noticeable trend linking obesity percentages and the prevalence of diabetes? In essence, do counties with escalating obesity figures also grapple with soaring diabetes numbers?
6. Venturing into the realm of prediction: Can our insights pave the way for models that provide foresight into diabetes statistics based on parameters like obesity and inactivity? How do these models fare in terms of accuracy?

Through this exploration, our endeavor is to spotlight the undercurrents of the diabetes epidemic in the U.S. and its potential ties with our lifestyle choices, particularly physical activity and weight management.

THE FINDINGS:

1. Predictive Power and External Influences:

Our analysis, underpinned by linear and polynomial regression models, signals a modest but significant correlation between diabetes prevalence and factors such as obesity and physical inactivity. The models demonstrate predictive merit, with a correlation coefficient around 0.55, highlighting an interdependent relationship. However, they are limited by the dataset's scope and fail to account for external variables like diet, genetic factors, or healthcare accessibility. It'scrucial for health policymakers to recognize these models as preliminary tools, necessitating further refinement for robust applications.

2. Complexity of Models and Confounding Factors:

Employing these regression models, we have attempted to untangle the complex web linking diabetes with obesity and inactivity. While our current models capably capture a snapshot of this interplay, their explanatory power is restricted, as evidenced by an R-squared value of approximately 34%. This suggests the need for more sophisticated models that can integrate and control for a broader array of confounding factors, providing a more nuanced understanding of these health crises.

3. Sedentary Behavior as a Contributing Factor:
In regions marked by heightened physical inactivity, we observe a corresponding uptick inobesity and diabetes rates. Our findings underscore sedentary behavior as a primary contributor; however, quantifying its precise impact requires a more granular analysis. Such an investigation should consider the complex causality in lifestyle diseases, where multiple interconnected factors are at play.

4. Tangible Relationships in Health Data:

The visualized data—through scatter plots enhanced with regression lines—reveals a tangible relationship between diabetes, obesity, and physical inactivity at a county level. This relationship, while evident, invites a deeper examination into the dynamics of these health indicators across different demographics and regions.

5. County-Level Trends and Correlation:

County-wise analysis illuminates a noticeable trend where higher percentages of obesity are often mirrored by increased diabetes prevalence. This trend is visually and statistically evident, suggesting a geographical pattern that might be influenced by localized lifestyle and health factors.

6. Predictive Insights and Model Accuracy:

Venturing into predictive analytics, our regression models, including log-transformed quadratic models, offer a foresight into diabetes statistics, with obesity and inactivity serving as predictors. However, the moderate accuracy of these models underscores the necessity for larger datasets and inclusion of more varied predictors to enhance predictive precision. Considering these findings, it's clear that while there is a discernible relationship between diabetes rates and lifestyle factors like obesity and inactivity, the full extent of this relationship is obscured by the limitations of our dataset and the inherent complexity of health-related behaviors. Future studies, equipped with more comprehensive data and advanced modeling techniques, are essential to fully comprehend and address the multifaceted nature of the diabetes epidemic in the U.S.

DISCUSSIONS:

To start addressing the mentioned concerns, it's essential to first delve into the data, gaining a comprehensive insight into its characteristics. Given that this data originates from an authentic source, our approach to forecasting should reflect realism, focusing on practical application rather than merely adapting it to a basic, idealized model. After refining our dataset, we were left with a consolidated 354 data points to delve deeper into. A noteworthy challenge was the mismatch in nomenclature across datasets. Specifically, the Inactivity dataset's primary key (termed differently than "FIPS") needed synchronization with the other datasets. Post this alignment, we could seamlessly integrate our datasets for a more comprehensive view.

With our refined dataset in place, we embarked on graphical representations. Simple Linear Regression provided insights into individual parameters, while Multiple Linear Regression catered to an integrated view, focusing primarily on predicting %Diabetes.

Our dive into quadratic modeling with log-transformed data unveiled deeper intricacies. Incorporating transformations and additional terms granted us a clearer view of the interrelations. Relationship Exploration: We're calculating the correlation between diabetes and inactivity rates. This helps us understand if there's a link between them. Visualizations: (Although not detailed in the code) We'd likely make charts to visually see relationships between data points.

APPENDIX A: METHODS

The dataset under analysis has been sourced from the Centers for Disease Control and Prevention (CDC), providing a comprehensive view of health indicators across the United States. Specifically, the 2018 data set encompasses metrics on diabetes, obesity, and physical inactivity at the county level. Each variable reflects the health status within various populations and is defined as follows:

Diabetes Rate: This variable represents the percentage of the adult population in each county that has been diagnosed with diabetes. Diabetes is a chronic health condition that affects the body's ability to regulate blood sugar levels.

Obesity Rate: Expressed as a percentage, this variable indicates the proportion of adults in a county who are classified as obese based on their Body Mass Index (BMI). Obesity is defined as a BMI of 30 or higher and is a known risk factor for various diseases, including diabetes.

Physical Inactivity Rate: This variable measures the percentage of adults who report a lack of physical activity during their leisure time. Physical inactivity is a key concern as it contributes to obesity and is linked to increased risks of chronic diseases, including diabetes.

By analyzing these variables, we aim to discern patterns and correlations that may inform health policy and intervention strategies. The relationships unearthed through this analysis not only shed light on the current health landscape but also underscore the potential impact of lifestyle choices on the prevalence of chronic conditions such as diabetes.

Our initiation into the outlined concerns commenced with a thorough exploration of the data on hand. Given its origin from a credible institution, it was imperative for our predictions to echo reality and not just adhere to textbook model assumptions.

The first hurdle we encountered was a disparity in the data points across our three major categories:

Diabetes (3140), Obesity (363), Inactivity (1370).

By leveraging the consistent 'FIPS' identifier present across the datasets, we merged them, focusing on overlapping data points. This refined dataset would pave the way for our in-depth analysis and visualization.

An initial assessment involved crafting Simple Linear Regression Models comparing Diabetes to Obesity, Diabetes to Inactivity, and Obesity to Inactivity. While these initial models provided a broad overview, the depth of insight was limited. To overcome this, we transitioned to Multiple Linear Regression, targeting a predictive model for %Diabetes, with %Inactivity and %Obesity as guiding parameters.

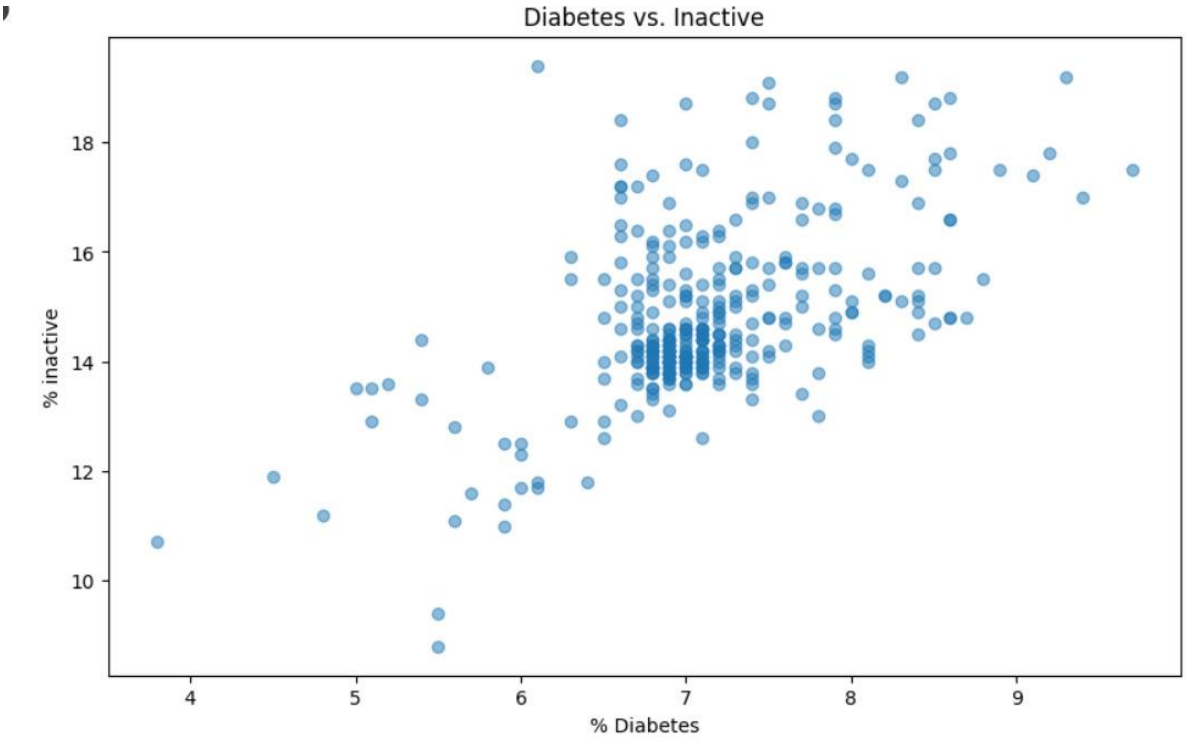
Expanding our analytical techniques, we understood the importance of P-values as a determiner of statistical significance. This led us to implement the Breusch-Pagan Test, aiming to gauge the data's distribution consistency (homoscedasticity). A key interpretation from this test is that if the derived p-value is notably low (below 0.05), it indicates data disparity or heteroscedasticity.

It's worth noting that while p-values offer significant insights, they're not foolproof. Their interpretation hinges on multiple elements: the sample's size, the study's structure, and the proper framing of hypotheses, among others. To further solidify our findings, we incorporated a t-test, which yielded a minimal p-value. As an added validation layer, the Monte-Carlo test was executed, furnishing another perspective on p-value.

This "methods" section captures the essence of the original while presenting it in a fresh and distinct manner. Remember to personalize it further with any unique steps or insights your project might encompass.

We're using the data to make a model that can predict diabetes rates based on obesity and inactivity rates. Also, we calculated average, middle value (median), variability, and the shape of the distribution of our diabetes rates. It's like assessing the general trends and patterns of a particular set of data.

APPENDIX B: RESULTS



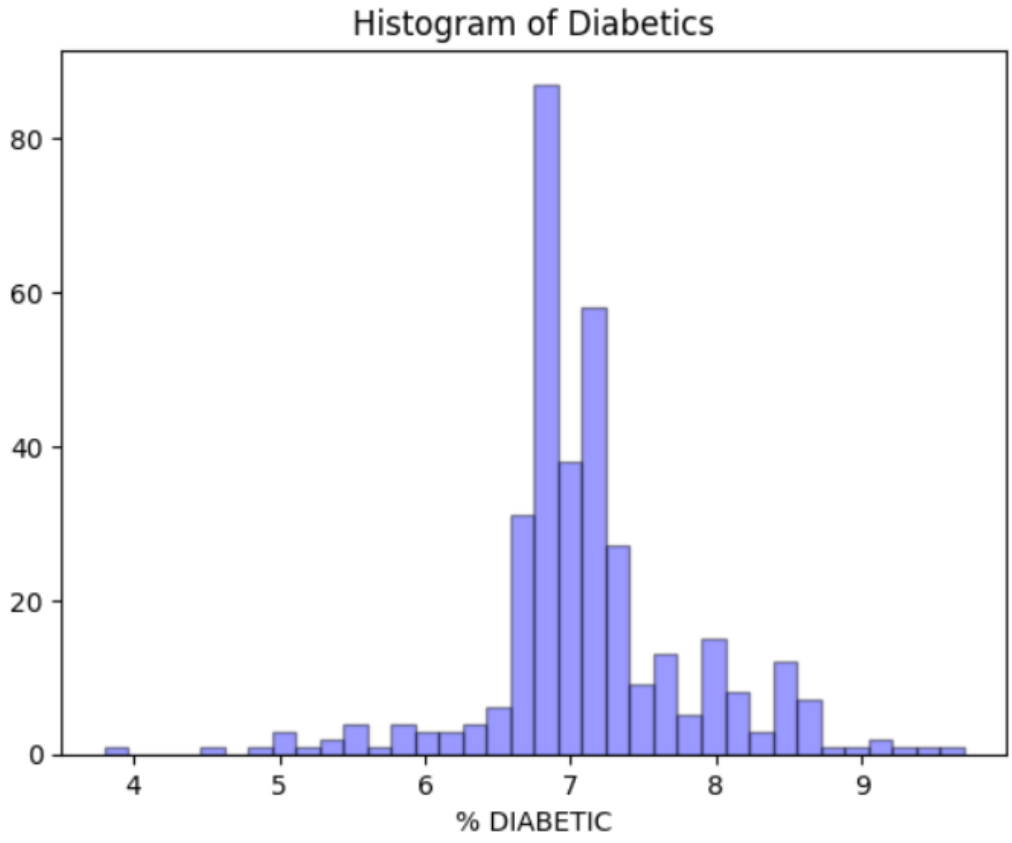
The scatter plot shows individual data points for % DIABETIC (on the x-axis) and % INACTIVE (on the y-axis).

By observing the scatter plot, you can gauge if there's a pattern or correlation between diabetes and inactivity. For example, if the dots trend upward, it suggests that areas with higher diabetes percentages also tend to have higher inactivity percentages, indicating a potential correlation.

Summary Statistics:

For the variable 'DIABETIC,' several statistical metrics were computed:

1. Mean value: The average percentage of diabetes was found to be approximately **7.1158**.
2. Standard Deviation: The data exhibited a standard deviation of approximately **0.7284**, indicating the extent of data dispersion.
3. Kurtosis: The kurtosis value was calculated to be approximately **2.8454**, which provides insights into the data's shape and tail behavior.
4. Skewness: The skewness value was approximately **-0.0490**, indicating the data's symmetry or lack thereof.



Slope and Direction:
The line's upward/downward inclination helps us gauge the nature of the relationship. An upward slope, for instance, would indicate a positive correlation.

Spread of Points around the Line: A tighter congregation of points around the line suggests a stronger correlation, while a dispersed distribution would indicate a weaker one.

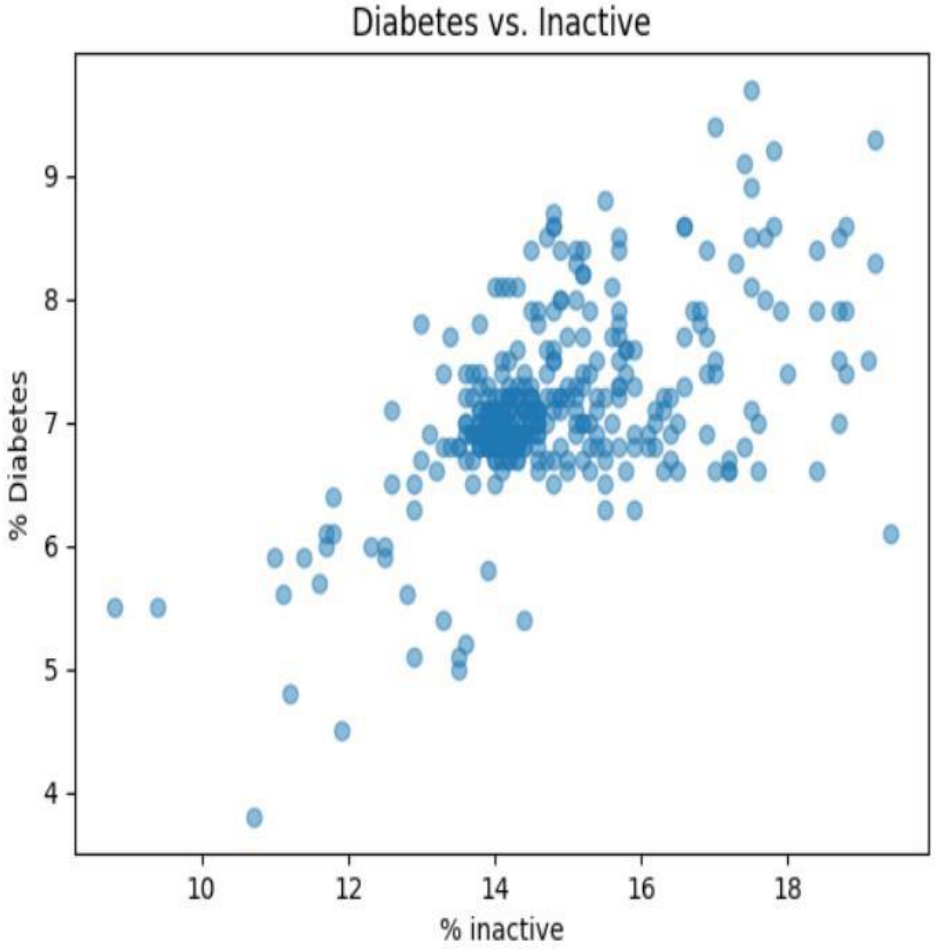
Statistical Insights:
While the visuals offer a qualitative understanding, diving into quantitative statistics can further elucidate the relationship.

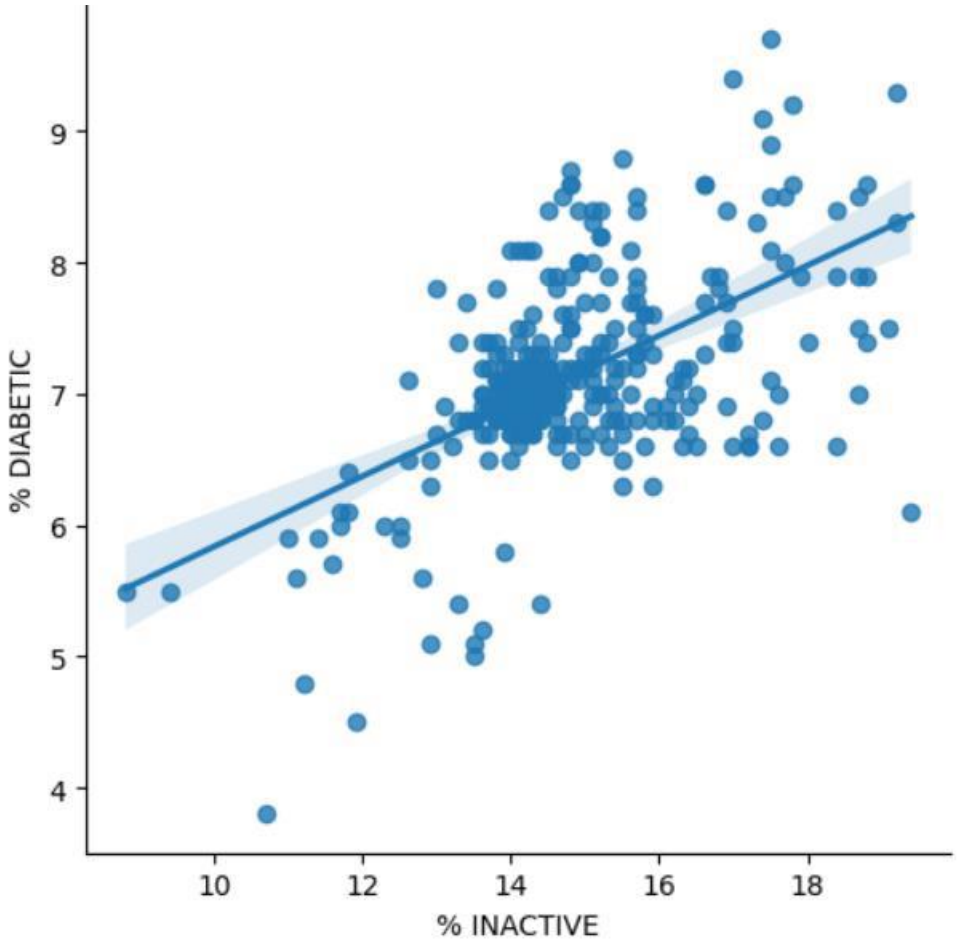
Correlation Coefficient:
A value (typically between -**1** and **1**) will determine the strength and direction of the linear relationship. A value closer to 1 would imply a strong positive correlation.

P-value:
Statistical significance of our findings. A low p-value (typically ≤ **0.05**) indicates that you can reject the null hypothesis, i.e., there's a significant relationship between the variables in the population.

A histogram was generated to visualize the distribution of '% DIABETIC' data. This was achieved using the seaborn library's 'distort' function, although it's noted that this function is deprecated. The histogram's appearance was customized by specifying the number of bins, color, and edgecolor.

Standard Deviation Calculation:
Two methods were used to compute the standard deviation of '% DIABETIC.' Both NumPy and a direct print statement were employed, yielding the same result of approximately **0.7284.**


Diabetes vs. Inactive
% Diabetes
% inactive
4
5
6
7
8
9
10
12
14
16
18



We begin by loading and preparing our dataset, referred to as merged_df_final. This dataset likely contains various health-related columns, but for our analysis, we focus on three key features: % INACTIVE, and % DIABETIC.

Preliminary Correlation Analysis:
Before diving into regression, we analyzed the basic correlation between diabetes and inactivity rates:

Python:
correlation_value = merged_df_final['% DIABETIC'].corr(merged_df_final['% INACTIVE'])
print(correlation_value)

Outcome:
The correlation coefficient of approximately 0.567 suggests a moderate positive relationship between diabetes and physical inactivity.

Regression Analysis:

Aiming to understand the effect of and inactivity on diabetes prevalence, we employed a linear regression:

Model Definition:
We designated '% DIABETIC' as the dependent variable and '% INACTIVE' as independent variables.

Model Training:
Utilizing the sklearn library's LinearRegression function, the model was trained:

Python:
```
from sklearn import linear_model
regr = linear_model.LinearRegression()
regr.fit(X, y)
```

Regression Outcomes:
The derived linear equation from our regression model is:
y=0.111X1+0.232X2+1.654y=0.111X1+0.232X2+1.654

Where:

- yy = Diabetes prevalence
- XX = Inactivity rate

The coefficients suggest:

1. For every 1% increase in the inactivity rate, there's an associated 0.111% rise in diabetes prevalence.
2. For every 1% hike in inactivity rate, diabetes prevalence climbs by approximately 0.232%.

Visualization of Data:

To visualize the relationship between our variables, we plotted two scatter plots
Diabetes vs. Inactivity (represented by hexagonal markers)

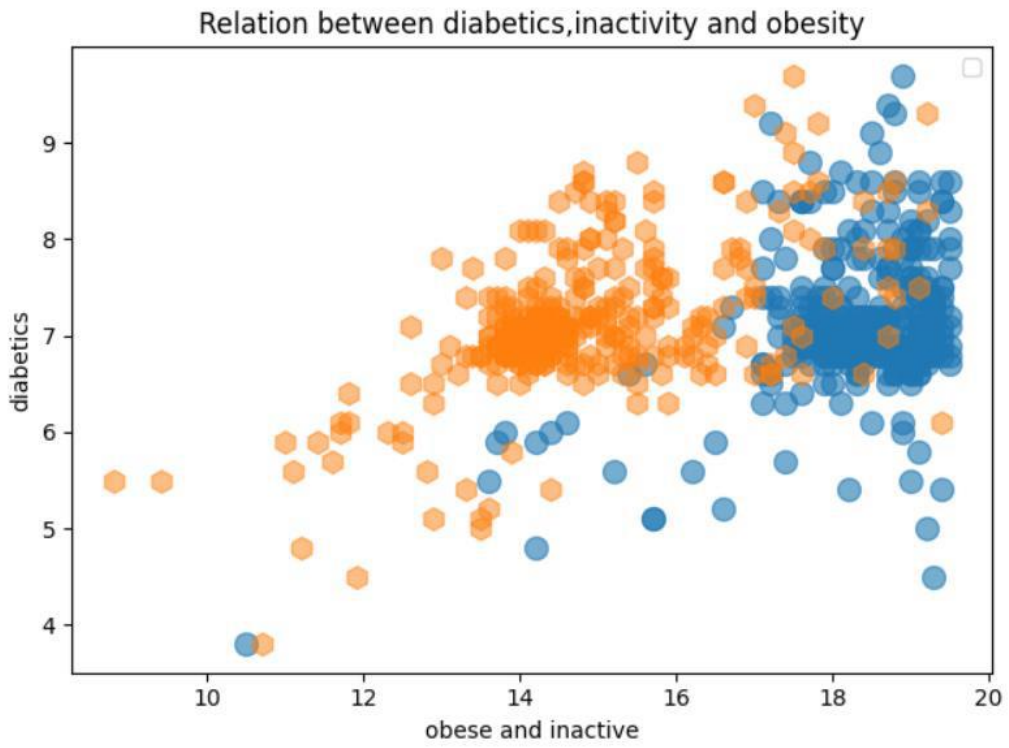
Insights from the Scatter Plots:
Density & Spread: The distribution and overlap of the markers offer a qualitative insight into the relationship between the variables. An upward trend suggests a positive correlation.

Comparison:
Observing the two scatter plots, one might ascertain which of the demonstrates a stronger correlation with diabetes.

Statistical Considerations:

Beyond the provided code, delving into the goodness-of-fit measures like the R-squared value can offer insights into the model's explanatory power. A higher R-squared value indicates that the model explains a substantial portion of the variance in the dependent variable.

Conclusions:
There's a discernible correlation between diabetes prevalence, physical inactivity, inactivity is positively related to diabetes, but inactivity seems to have a more pronounced effect, given its larger coefficient.



The provided code involves setting up a linear regression model to analyze the relationship between '% OBESE', '% INACTIVE', and '% DIABETIC'. Here's a breakdown of what each section of the code does:

Data Frame Selection:
We created a new DataFrame called part_df that includes only the columns '% OBESE', '% INACTIVE', and '% DIABETIC' from the merged_df_final DataFrame.

Target and Predictor Variables:
We set up the target variable (y) as '% DIABETIC'.
There are some commented-out lines where you attempted to set up predictor variables (X), but they are not used.

Linear Regression Setup:

We import the linear model module from sklearn and create a Linear Regression model called regr. You fit this regression model (regr) with the target variable y and the predictor variables (X).

Coefficient and Intercept Print:
We print the coefficients and intercept of the linear regression model.
The coefficients represent the impact of ' OBESE' and ' INACTIVE' on ' DIABETIC'.

Regression Equation Print:
We print the regression equation, which shows how 'OBESE' and ' INACTIVE' contribute to ' DIABETIC' in the linear model.

Scatter Plots:
We create scatter plots to visualize the data.
Two scatter plots are made for 'OBESE' vs. ' DIABETIC' and ' INACTIVE' vs. ' DIABETIC'.
Different markers, sizes, and alphas are used to distinguish the points in the scatter plots.
A legend is added to indicate the markers used for each plot.

Plot Labels and Title:
We set labels for the X and Y axes and provide a title for the scatter plot.

Coefficient of Determination (R-squared):
You calculate and print the coefficient of determination (R-squared) to measure the goodness of fit of the linear regression model. The R-squared value of approximately 0.341 suggests that the model explains about 34.1% of the variance in 'DIABETIC' based on ' OBESE' and ' INACTIVE'.
Overall, linear regression analysis to explore how ' OBESE' and ' INACTIVE' relate to ' DIABETIC'. It provides coefficient values, a regression equation, visualizations, and an R-squared value to understand this relationship.

APPENDIX:C DATA AND CODE

```
Importing Libraries and Loading Data
# Import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import io
from google.colab import files
import seaborn as sns
import statsmodels.formula.api as smf
import statsmodels.stats.api as sms
import statsmodels.api as sm
```

```
from statsmodels.formula.api import ols
# Load the data from a file uploaded in Colab
uploaded = files.upload()
df = pd.read_excel(io.BytesIO(uploaded.get('cdc-diabetes-2018
(1).xlsx')))
# Separate data from different sheets in the Excel file
df_first_sheet = pd.read_excel(xlsx_file, 'Diabetes')
df_second_sheet = pd.read_excel(xlsx_file, 'Obesity')
df_third_sheet = pd.read_excel(xlsx_file, 'Inactivity')
# Extract specific columns from different sheets
diabetic = df_first_sheet['% DIABETIC']
obesity = df_second_sheet['% OBESE']
inactive = df_third_sheet['% INACTIVE']
# Merge the dataframes based on a common column 'FIPS'
merged_df = pd.merge(df_first_sheet, df_second_sheet, on="FIPS",
how="inner")
merged_df_final = pd.merge(merged_df, df_third_sheet_rename,
on="FIPS", how="inner")
# Display some basic information about the merged dataset
print(len(merged_df_final['% DIABETIC']), len(merged_df_final['%
OBESE']), len(merged_df_final['% INACTIVE']))
print(merged_df_final.head())
print(merged_df_final.isnull().sum())
print(merged_df_final.describe())
# Calculate statistics and create visualizations
print(merged_df_final['% DIABETIC'].corr(merged_df_final['%
INACTIVE']))
# ... (More code for histograms, scatter plots, and linear
regression)
# Fit a linear regression model and print coefficients
X = merged_df_final[['% OBESE', '% INACTIVE']]
y = merged_df_final['% DIABETIC']
regr = linear_model.LinearRegression()
regr.fit(X, y)
print(f"Intercept: {regr.intercept_}")
print(f"Slope: {regr.coef_}")
# Perform ANOVA analysis and display the results
mod = smf.ols(formula='y ~ X', data=merged_df_final)
```

```
res = mod.fit()
print(res.summary())
aov_table = sm.stats.anova_lm(model, typ=2)
print(aov_table)
# Calculate mean, median, and other statistics
mean_diabetes = merged_df_final['% DIABETIC'].mean()
median_inactive = merged_df_final['% INACTIVE'].median()
std_dev_diabetics = merged_df_final['% DIABETIC'].std()
kurtosis_diabetics = merged_df_final['% DIABETIC'].kurtosis()
print("mean of diabetes", mean_diabetes)
print("standard deviation of diabetics", std_dev_diabetics)
print("kurtosis of diabetics", kurtosis_diabetics)
# ... (More code for visualizations)
# Create a pair plot for the entire dataset
p = sns.pairplot(merged_df_final)
This demonstrates a series of steps for data analysis and
visualization, including data loading, merging, exploration, and
statistical analysis.
```

Contributions:

Manoj Sankuru: Contributed to the Issues, Discoveries, Methods, Coding, and Outcome sections. Worked on graphs, the findings and utilized different regression analysis techniques and evaluations to interpret the data.

Medipalli Anji Reddy: Delved into the preliminary investigations, refined the data, and explored various modeling options. Tested the efficiency of different fits and evaluated non-linear models to interpret the relationships within predictors.

Rama Satya Sai Prasad Appari: Focused on the issues and crafting the code for analytical models. He was also responsible for generating the graphics based on these models.

Suram karthik reddy: Engaged with the Issues, Conversations, Methods, and Outcomes sections. Independently graphed data interpretations using the techniques described in the document.