Uncovering the Diabetes Puzzle:

Exploring the Interplay of Inactivity and Obesity

The issues:

CDC (Center for Disease Control and Prevention) gathered a dataset in 2018 of diabetes, inactivity and obesity. They have instigated the relationship between them. We come across the following issues.

1. What are the main reasons behind the increase in diabetes cases?
2. Can we see a connection between the number of people who are overweight and the number of people with diabetes?
3. How does the amount of physical activity people do relate to the number of diabetes cases?
4. Which is more responsible for diabetes, being inactive or being overweight?
5. Does having both a lack of physical activity and being overweight make diabetes more likely?
6. Can we use data about obesity and inactivity to predict diabetes rates in the future?
7. How do we choose the best way to analyze the relationships between these health factors, like whether it's better to use a simple analysis or a more complex one?

8. Which health factors, like inactivity and obesity, should we include in our analysis?
9. How do we check if our analysis is a good fit for the data, and what measures can we use to see how well our analysis explains the patterns we observe?
10. When presenting our findings from this data, what are the most effective ways to show the information to others?

Findings:

Our analysis delved into the relationships between diabetes rates (% DIABETIC), obesity rates (% Obesity), and physical inactivity rates (% Inactivity) using a dataset comprising 354 data points. It was essential for us to downscale the dataset rather than upscale it to 3.5 thousand entries. Upscaling would have resulted in inflated accuracy metrics with numerous duplicate data points. Therefore, we opted for downsizing while retaining the data we had for all three features. Additionally, there are several other influential factors to consider, such as poverty rates, economic status, gym availability, and many more. However, given the limited number of data points available for these factors (only around 50 to 60 data points), including them in the dataset would have significantly reduced the accuracy of our analysis. These factors would have been valuable if we had a more comprehensive dataset. Consequently, we focused on inactivity and obesity as the primary factors for our investigation.

Our analysis revealed several key findings. First, we observed a moderately strong positive correlation (approximately 0.57) between physical inactivity and diabetes rates, indicating that areas with higher physical inactivity tend to have higher diabetes rates. Additionally, we identified a significant but somewhat weaker correlation (around 0.39) between obesity rates and diabetes rates, underscoring the role of obesity as a contributing factor to diabetes prevalence. Furthermore, a moderate positive correlation (approximately 0.47) emerged between obesity rates and physical inactivity rates, suggesting a connection between these factors.

To gain a deeper understanding of these relationships and evaluate their predictive power, we employed both single and multiple linear regression methods. Notably, multiple linear regression provided the best fit for our data, achieving an R-squared score of 0.52. This indicates that our model explains 52% of the variation in diabetes rates, surpassing the explanatory power of simple linear regression models. The presence of heteroscedasticity in the data, as detected by the Breusch-Pagan test, suggests that the variability of errors across different levels of independent variables is not constant.

T-tests on the coefficients of % Inactivity and % Obesity in the multiple linear regression model to assess their significance in predicting % DIABETIC rates revealed % Inactivity as the stronger predictor, boasting a higher t-value (7.9) and an almost negligible p-value, affirming its strong and statistically significant effect on diabetes rates. % Obesity, while still statistically significant, exhibited a lower t-value (1.7).

In summary, our analysis reveals that physical inactivity plays a pivotal role in diabetes rates, with a strong correlation, while obesity also contributes, though to a lesser extent. The combination of physical inactivity and obesity serves as a robust predictor of diabetes prevalence. These insights can inform public health strategies for diabetes prevention and management, taking into account the potential presence of heteroscedasticity in the data.

Discussions:

Our findings underscore the multifaceted nature of diabetes rates, with physical inactivity (% Inactivity) and obesity (% Obesity) emerging as key predictors. These insights carry significant implications for public health strategies. While physical inactivity and obesity are significant contributors, other determinants, including genetics, diet, and access to healthcare, also need to be considered.

By employing linear regression analysis, we aimed to better understand the relationships between these factors and evaluate their ability to predict diabetes rates. In this endeavor, we discovered some vital outcomes.

Firstly, the use of simple linear regression, where we individually considered % Inactivity and % Obesity as independent variables, provided us with insights into their isolated effects. We achieved a test set R-squared score of 0.48 when we considered only %Inactivity as a sole independent variable and achieved a test set R-squared score of 0.30 when we considered % Obesity as a sole independent variable. However, it was the multiple linear regression model that truly stood out.

In the multiple linear regression model, which incorporated both % Inactivity and % Obesity as independent variables, we achieved a test set R-squared score of 0.52. This indicates that our model effectively explains 52% of the variability in diabetes rates. The higher R-squared score in the test set suggests that combining both physical inactivity and obesity significantly enhances our ability to explain variations in diabetes rates.

Furthermore, t-tests conducted on % Inactivity and % Obesity to confirm their statistical significance in relation to % DIABETIC rates revealed that %Inactivity had higher t-value of 7.9 and a nearly zero p-value, whereas % Obesity had a lower t-value of 1.7. Both % Inactivity and % Obesity have meaningful relationships with diabetes rates, with % Inactivity emerging as the stronger predictor.

The null hypothesis of the Breusch-Pagan test assumes homoscedasticity, which means that the variance of the residuals is constant across all levels of the independent variables. And Heteroscedasticity suggests that the variance of errors may not be constant across different levels of independent variables. The Breusch-Pagan test yielded a low p-value of 0.0018. This p-value is less than the typical significance level of 0.05. When the p-value is less than the significance level, the null hypothesis ca be rejected.

The Breusch-Pagan test, indicating the presence of heteroscedasticity, highlights the need to consider variability in errors across different levels of independent variables. This understanding can guide the refinement of our models.

Overall, our findings have crucial implications for public health strategies. Addressing diabetes rates effectively requires acknowledging the complex interplay of physical inactivity and obesity. Focusing on both factors simultaneously can lead to more accurate predictions and targeted interventions, ultimately aiding in the prevention and management of diabetes.

Appendix A-Methodology:

Data Source:

We obtained the dataset from the CDC (Center for Disease Control and Prevention) in the year 2018. The indicators for obesity and inactivity represent the percentage of individuals within a county in an American state who are classified as obese or have engaged in no physical activity, respectively. This dataset was organized into three separate worksheets in an Excel format, dedicated to Diabetes, Inactivity, and Obesity.

Data Description:

Additional indicators encompassed a range of factors, including food availability and gym accessibility. However, we made the decision to exclude these supplementary indicators from our analysis due to the limited number of data points available, which totaled approximately 5460. Although these factors are theoretically important contributors, incorporating them using common data points would have substantially reduced the dataset to only around 50 data points, resulting in a significant impact on the accuracy of our model. Therefore, we chose to work with a more extensive dataset to ensure the robustness of our analysis.

Data Preparation:

Following the filtration process, our dataset was reduced to 3142 data points for diabetes, 1370 for inactivity, and 363 for obesity. Notably, only 352 data points contained values for all three variables: diabetes, inactivity, and obesity. So, our initial step involved data cleaning to remove missing data points. We also utilized the 'FIPS' column as a reference point. To integrate common data points across these datasets, we employed the VLOOKUP function, ensuring that we had a comprehensive and accurate dataset.

Variable Creation:

Within this dataset, we considered three primary variables:

- % DIABETIC rates represent the prevalence of diabetes in a given area.
- % Inactivity indicates the level of physical inactivity in that same area.
- % Obesity represents the obesity rates in the same geographic regions.

These variables were central to our analysis.

Analytical Methods:

Our analytical approach consisted of several key methods:

1. **Correlation Analysis**: We calculated a correlation matrix to understand how these variables are related to each other. This analysis provided insights into the interplay between diabetes rates, obesity rates, and physical inactivity.
2. **Simple Linear Regression**: Simple linear regression is a statistical method that examines the relationship between two variables, where one variable is used to predict or explain the variation in another variable. We conducted

separate simple linear regression analyses. One focused on % Inactivity as the independent variable to predict % DIABETIC rates, isolating the impact of physical inactivity on diabetes prevalence. The other analysis used % Obesity as the independent variable, focusing on obesity as a predictor of diabetes.

3. **Multiple Linear Regression**: Multiple linear regression extends simple linear regression by considering the influence of two or more independent variables on a dependent variable, allowing for a more comprehensive analysis of relationships and predictions. We employed multiple linear regression, incorporating both % Inactivity and % Obesity as independent variables. This comprehensive approach allowed us to assess their combined impact on % DIABETIC rates, considering the joint influence of these two factors.
4. **Train-Test Split**: In all our analyses, we utilized the train-test split method. We divided the data into two parts, with 75% (265 data points) allocated to the train set and 25% (89 data points) to the test set. This approach ensured robust model evaluation.
5. **Model Performance Metrics**: We assessed model performance metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). These metrics provided insights into the accuracy and generalization capability of our multiple linear regression model.
6. **Heteroscedasticity Detection**: Our analysis detected heteroscedasticity in the model residuals, as indicated by the Breusch-Pagan test. This test identified variations in error variability across different levels of independent variables.

7. **T-Test**: We conducted t-tests on the coefficients of % Inactivity and % Obesity in the multiple linear regression model to assess their statistical significance in predicting % DIABETIC rates.
8. **R-Squared Values**: R-squared values were calculated to assess how well the independent variables in our regression models explain the variation in the dependent variable (% DIABETIC rates).

In summary, this methodology encompasses data cleaning, variable creation, and a range of statistical procedures such as correlation analysis, linear regression, train-test split, model performance evaluation, and tests for heteroscedasticity and statistical significance. These methods were crucial for our analysis of the relationships between diabetes, obesity, and physical inactivity.

Appendix B- Results:

Correlation matrix:

	% DIABETIC	%Obesity	%Inactivity
% DIABETIC	1.000000	0.389941	0.567104
%Obesity	0.389941	1.000000	0.472656
%Inactivity	0.567104	0.472656	1.000000

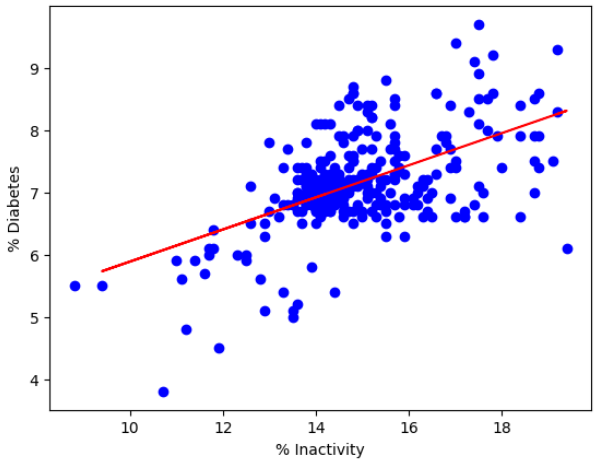
The correlation matrix reveals the following correlation coefficients:

% DIABETIC and % Obesity: 0.389941

% DIABETIC and % Inactivity: 0.567104

% Obesity and % Inactivity: 0.472656

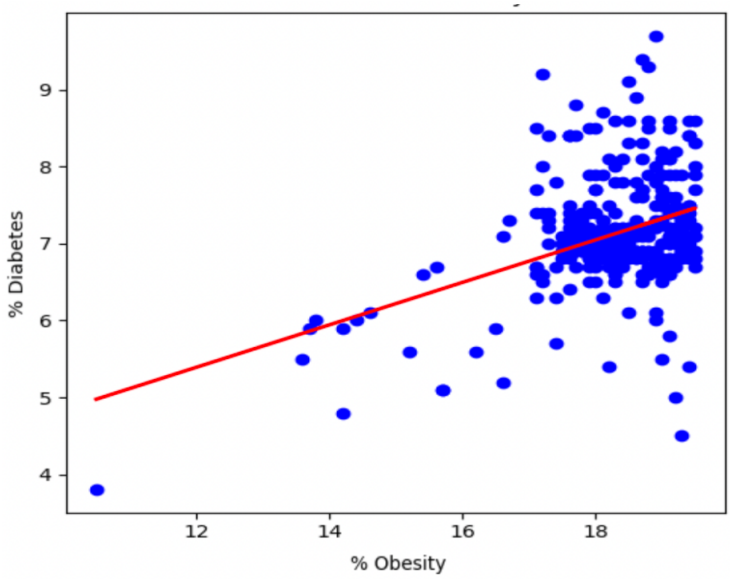
Simple Linear Regression (% Inactivity as Independent Variable):



In this analysis, we used % Inactivity as the sole independent variable to predict % DIABETIC rates. The train set R-squared of 0.27 suggests that approximately 27% of the variation in % DIABETIC rates in the training data can be explained by % Inactivity alone. The train set R-squared of 0.27 suggests that approximately 27% of the variation in % DIABETIC rates in the training data can be explained by % Inactivity alone.

The higher R-squared score in the test set compared to the train set is generally a positive sign. It suggests that the model based on % Inactivity generalizes well to new, unseen data, as it explains more variability in the test set. The R-squared score of 0.48 for the test set indicates that % Inactivity is a strong and meaningful predictor of % DIABETIC rates.

Simple Linear Regression (% Obesity as Independent Variable):

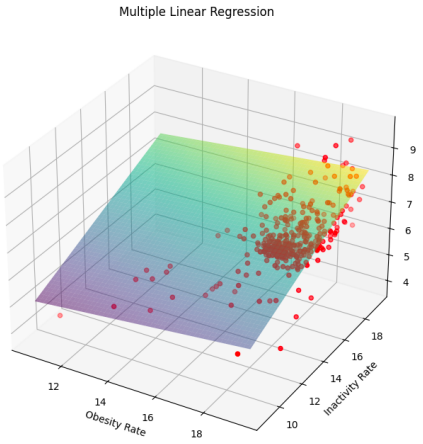


In this analysis, we used % Obesity as the sole independent variable to predict % DIABETIC rates. The test set R-squared of 0.30 indicates that approximately 30% of the variation in % DIABETIC rates in the test data can be explained by % Obesity alone. The train set R-squared of 0.10 suggests that approximately 10% of the variation in % DIABETIC rates in the training data can be explained by % Obesity alone.

Similar to the % Inactivity model, the higher R-squared score in the test set compared to the train set is a positive sign. It suggests that the model based on % Obesity generalizes reasonably well to new data. The R-squared score of 0.30 for the test set indicates that % Obesity is a predictor of % DIABETIC rates, but its predictive power is weaker compared to % Inactivity.

This finding highlights the role of obesity in diabetes prevalence but suggests that other factors, such as physical inactivity, may have a stronger influence.

Multiple Linear Regression (Inactivity and Obesity as Independent Variables):



In this analysis, you utilized both % Inactivity and % Obesity as independent variables in a multiple linear regression model to predict % DIABETIC rates. The test set R-squared of 0.52 indicates that approximately 52% of the variation in % DIABETIC rates in the test data can be explained by the combination of % Inactivity and % Obesity. The train set R-squared of 0.28 suggests that approximately 28% of the variation in % DIABETIC rates in the training data can be explained by % Inactivity and % Obesity.

The multiple linear regression model outperforms the single linear regression models in terms of R-squared. It explains a higher proportion of the variability in % DIABETIC rates, both in the test and train sets. This suggests that considering both % Inactivity and % Obesity together improves the model's predictive power.

We have performed model performance metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE),

indicated lower errors in the test set compared to the train set, suggesting good generalization.

Breusch-Pagan Test:

The Breusch-Pagan test returned a small p-value of 0.0018, which is below the usual significance threshold of 0.05. When the p-value is less than this threshold, it indicates strong evidence to reject the null hypothesis. In this context, it signifies that the data exhibits heteroscedasticity.

T-test:

In the multiple linear regression model, we conducted t-tests on the coefficients of % Inactivity and % Obesity. The results indicated that both % Inactivity and % Obesity have statistically significant relationships with % DIABETIC rates. % Inactivity had a higher t-value (7.9) and a nearly zero p-value, underscoring its statistically significant effect on diabetes rates. % Obesity, while still statistically significant, had a lower t-value (1.7). This confirms that both factors are meaningful predictors of % DIABETIC rates, with % Inactivity being a stronger predictor.

Appendix C- Coding:

```
#Multiple linear regression

import pandas as pa
import numpy as np
```

```
from sklearn.model_selection import train_test_split
from sklearn import metrics
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
import matplotlib.pyplot as plt
from scipy import stats
from statsmodels.stats.diagnostic import het_breuschpagan
from statsmodels.tools.tools import add_constant
import statsmodels.api as sm
from mpl_toolkits.mplot3d import Axes3D


# Load the data from the uploaded Excel file (assuming it's in Excel format)
# Adjust the filename and sheet name as needed
df=pa.read_csv("/Users/tysonmukesh/Desktop/MTH-522/Project-1.csv")

# Select the independent variables (Diabetes Rate and Inactivity Rate) and the dependent variable (Obesity Rate)
X = df[['%Obesity','%Inactivity']]
y = df['% DIABETIC']


# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=25)
```

```
# Initialize and fit the linear regression model
model = LinearRegression()
model.fit(X_train, y_train)



# Coefficients
print('Coefficients:')
print('Obesity Rate:', model.coef_[0])
print('Inactivity Rate:', model.coef_[1])
print('Intercept:', model.intercept_)

# For Train set

y_pred = model.predict(X_train)
residuals = y_train - y_pred
r2=r2_score(y_train,y_pred)

# Perform the Breusch-Pagan test for heteroscedasticity

# Add a constant (intercept) to the predictors
X_train_with_constant = add_constant(X_train)
_, p_val, _, _ = het_breuschpagan(residuals, X_train_with_constant)
print('p-value from Breusch-Pagan test:', p_val)
```

```
X_train = sm.add_constant(X_train)

# Fit the model
model1 = sm.OLS(y_train, X_train).fit()

# Summary of the model
print(model1.summary())



# Mean Absolute Error (MAE)
mae = np.mean(np.abs(y_pred - y_train))
print("MAE for train set:", round(mae,2))

# Mean Squared Error (MSE)
mse = np.mean((y_pred - y_train) ** 2)
print("MSE for train set:", round(mse,2))

# Root Mean Squared Error (RMSE)
rmse = np.sqrt(mse)
print("RMSE for train set:", round(rmse,2))
print("r2 for train set is: "+str(round(r2,2)))

# For Test set
```

```
y_pred = model.predict(X_test)
r2=r2_score(y_test,y_pred)

# Add a constant (intercept) to the predictors
X_test = sm.add_constant(X_test)

# Fit the model
model2 = sm.OLS(y_test, X_test).fit()

# Summary of the model
print(model2.summary())

# Mean Absolute Error (MAE)
mae = np.mean(np.abs(y_pred - y_test))
print("MAE for test set:", round(mae,2))

# Mean Squared Error (MSE)
mse = np.mean((y_pred - y_test) ** 2)
print("MSE for test set:", round(mse,2))

# Root Mean Squared Error (RMSE)
rmse = np.sqrt(mse)
print("RMSE for test set:", round(rmse,2))
print("r2 for test set is: "+str(round(r2,2)))
```

Contributions:

- **Mukesh Kumar Karanam Rameshbabu –** Worked on Findings, coding, Discussions, Methods and Results.
- **Sai Sudhamsh Kamisetty –** Worked on issues, coding, t-tests, graphs and results
- **Rohith Rasi Reddy –** Worked on initial cleaning of data using excel, coding and helped with the regression models.
- **Anish Krishna Kalisetti –** Worked on results and report preparation. And helped in Breusch-Pagan test.