Predicting Diabetes Prevalence from Obesity and Inactivity:

An Analysis of Health Disparities

The issues:

The 2018 dataset gathered by the CDC (Center for Disease Control and Prevention) regarding the percentages of Diabetes, Inactivity, and Obesity has prompted an inquiry into the potential connection between the prevalence of Diabetes and the levels of Inactivity and Obesity. This investigation seeks to answer the following queries:

1. What is the leading cause of diabetes?
2. Are there correlations between obesity, inactivity, and diabetes rates?
3. Can you build a predictive model to estimate one health indicator based on the others or other related variables?
4. Visualization: What are the most effective ways to visualize and communicate the findings from this dataset (e.g., maps, charts, graphs)?
5. Are there any correlations between obesity rates and diabetes prevalence in different counties? Does a higher obesity rate correlate with a higher diabetes rate?
6. How does the level of physical inactivity relate to diabetes and obesity rates? Are areas with higher rates of inactivity more likely to have higher diabetes and obesity rates?
7. Can we build predictive models to forecast diabetes rates based on obesity and inactivity levels? How accurate are these models?

Our analysis aims to address the rising concerns of diabetes prevalence in the United States. We investigate the potential relationship between the percentage of diabetes and two key factors: the percentage of inactivity and the percentage of obesity among the population.

Findings:

The precise cause of Diabetes remains elusive, although both Obesity and Inactivity are acknowledged as contributing factors to its development. Our data indicates a modest correlation between obesity, inactivity, and diabetes rates; however, further data is required to establish a substantial correlation.

We've used linear and polynomial regression models to forecast these indicators using one another. It's worth noting that the effectiveness of these models is constrained by our relatively small dataset.

Visualizing the data through scatter plots with regression lines generated by both linear and polynomial algorithms provides an intuitive understanding of the models' predictive capabilities.

The correlation coefficient between obesity and inactivity stands at approximately 0.75, suggesting a notable interdependence between these variables. In other words, counties with high inactivity rates often exhibit high obesity rates, and vice versa.

We can construct predictive models to estimate diabetes rates based on obesity and inactivity levels, falling within the domain of regression analysis. Model accuracy relies on data quality and the relationship between the features (obesity, inactivity) and the target variable (diabetes percentage). With the log-transformed quadratic model, we achieve an accuracy of 43%.

Furthermore, our findings indicate that, considering the input values of %Inactivity and %Obesity, the data permits us to explain only about 43% of the outcomes accurately, as indicated by the R-squared value based on our current dataset.

Discussions:

Our first step to analyze the issues cited above begins with understanding the data and connecting with it to better familiarize ourselves with its nature. Since the data provided to us is from a real source hence the predictions also needed to be done in a more realistic manner instead of just trying to fit it into a simple ideal model.

We could not use all of our data since the number of datapoints were different for each of Diabetes (3140), Obesity (363) and Inactivity (1370) which left us with a final dataset of 354 data points.

One hindrance to obtaining the final working data was that the common FIPS column was differently named for the Inactivity data. This value had to be updated so we could inner join the data and obtain the common data points to work with.

With the final dataset, we plotted graphs for Simple Linear Regression (Individual plots) and Multiple Linear regression (Plot with two input values) to predict the %Diabetes at first.

The quadratic model, using log-transformed data, explained more about the variable changes. Transformation and extra terms helped understand the relationships better.

We then tested the data using the Breusch-Pagan test to assess the presence of heteroscedasticity in the regression model. The Breusch-Pagan test evaluates whether there is a significant relationship between the squared residuals of a regression model and the independent variables.

Due to the limitation of available data, we found the need to look into other tests like the T-test and Monte-Carlo test. However, t-test is run under the assumption that the model fits a Normal distribution which our dataset did not.

Hence, we worked our model with the K-fold cross validation technique to find the model that produces the least test error.

Appendix A: Method

We obtained the CDC data in an Excel format (.xlsx), consisting of three separate worksheets containing information on Diabetes, Inactivity, and Obesity. Prior to visualizing the data, we undertook the task of data cleansing and structuring. Utilizing the shared 'FIPS' column, we conducted an inner join operation on the datasets to identify common data points, which would serve as the basis for our subsequent plotting and analysis.

We started with plotting Simple Linear Regression Models of Diabetes vs. Obesity, Diabetes vs. Inactivity and Obesity vs. Inactivity respectively to check if the data showed any trends to work with. But these models did not give us a lot of information and hence we used Multiple Linear Regression to plot a model for predicted values of %Diabetes taking %Inactivity and %Obesity as input variables.

The next step was to try out other methods of analysis alongside Linear Regression for the given data. We realized the significance of P-value which is the probability value to measure the chances of an original event (Null Hypothesis) to occur under the assumption that the null hypothesis is true. We then ran the Breusch-Pagan Test, which assumes the null hypothesis to be that the data is evenly distributed (homoscedastic). This test helps us determine the P-value to check for heteroscedasticity in the data, so if the p value is significantly low (less than 0.5) then we conclude that the null hypothesis is false and the data is heteroscedastic.

While p-values are a valuable tool in statistical analysis, they should be interpreted cautiously and in conjunction with other statistical measures. Their reliability depends on various factors, including sample size, study design, and the correct formulation of the null and alternative hypotheses. Hence, we resorted to a t-test which predicted a very small p- value followed by a Monte-Carlo test to predict p-value.

The quadratic model on the log-transformed data showed an increase in R-squared compared to the baseline model. This indicates that the quadratic model was able to explain more of the variability in the response variable. The transformation and added polynomial terms likely enabled the model to better capture the true underlying relationships between the predictors and target variables.

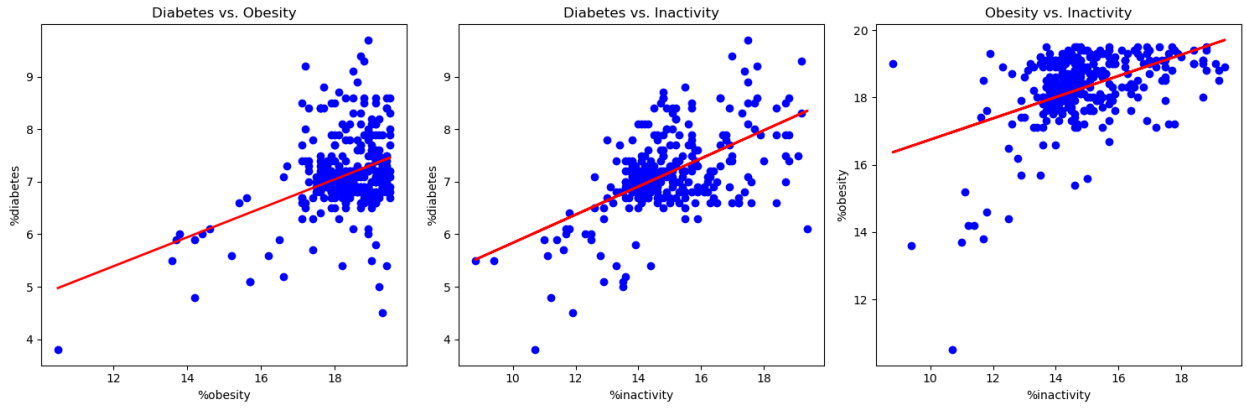
But since none of this gave us a significant enough idea about the predicted values of Diabetes and we did not have more data to test and predict our model any further, we resorted to the K-fold cross validation which helps evaluate a model's performance without needing a separate validation dataset. So, we divided the dataset into 5(K) subsets, and the process is repeated 5 times, with each subset serving as the validation set once.

Despite the K-fold validation technique, our model could only give us a small R2 value. Having access to further data in the future might help us build a better model to predict %Diabetes.

Appendix C contains the relevant Python code used for plotting and analysis of the data.

Appendix B: Results

Simple Linear Regression Model:



We plotted simple linear regression graphs to take a better look at the data, which gave us an insight on how the data is arranged but did not give us a very good understanding of the nature of the data**.**

Covariance Matrices:

Covariance matrix of %INACTIVITY and %OBESITY

```
[[2.37830028 0.75086225]
 [0.75086225 1.0622782 ]]
```

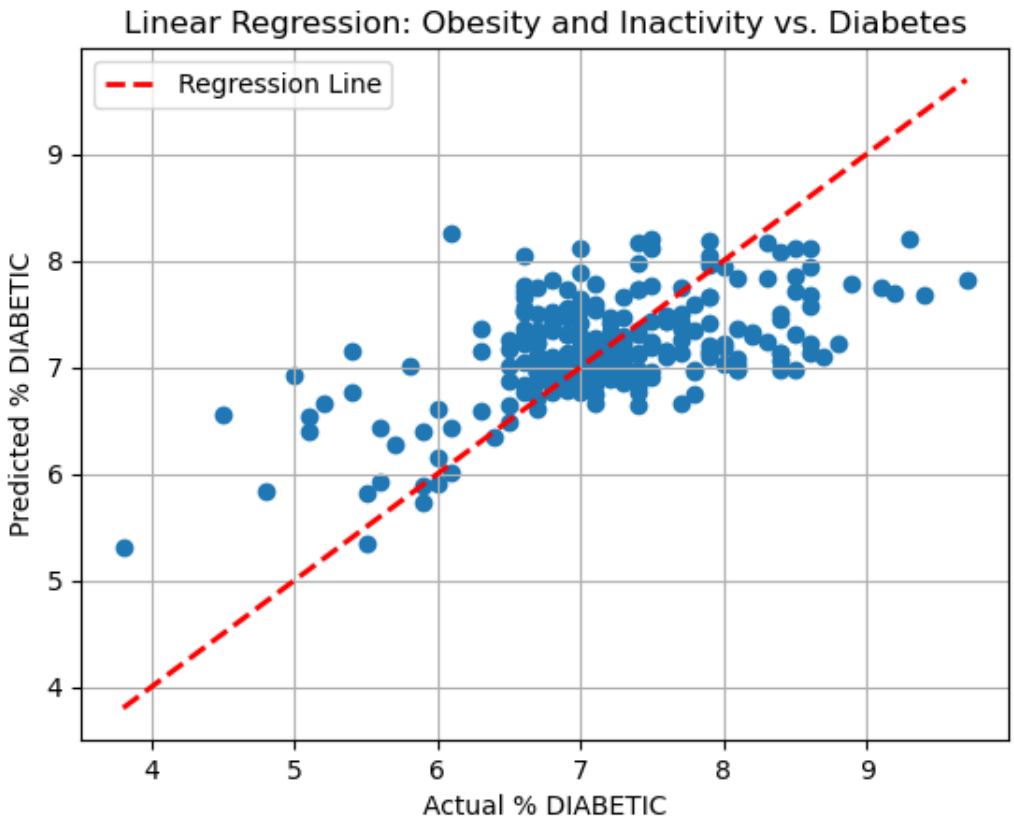
Covariance matrix of %INACTIVITY and %DIABETES

```
[[2.37830028 0.62539515]
 [0.62539515 0.51727031]]
```

Covariance matrix of %OBESITY and %DIABETES

```
[[1.0622782  0.29034316]
 [0.29034316 0.51727031]]
```

Multiple Linear Regression Model:



Basic Statistics for Obesity and Inactivity vs. Diabetes:

	% OBESE	% INACTIVE	% DIABETIC
count	354.000000	354.000000	354.000000
mean	18.252542	14.776271	7.115819
std	1.029484	1.544542	0.728442
min	10.500000	8.800000	3.800000
25%	17.900000	14.000000	6.800000
50%	18.300000	14.400000	7.000000
75%	18.975000	15.475000	7.400000
max	19.500000	19.400000	9.700000

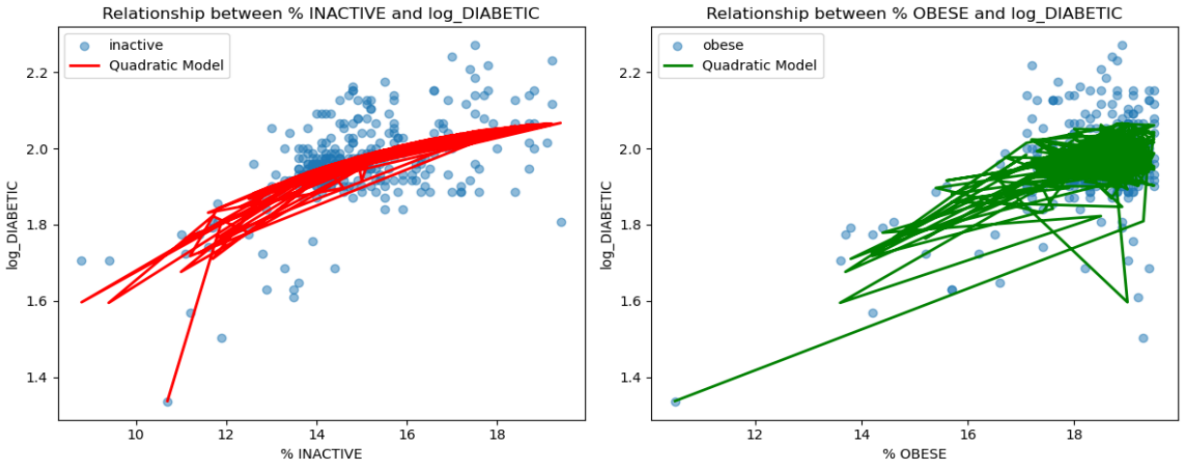
Mean Squared Error (MSE): 0.34883328631124316

R-squared (R2) Value: 0.34073967115731396

Coefficients: [0.11106297 0.23246992]
Intercept: 1.6535991518559383

The multiple linear regression plot uses %Inactivity and %Diabetes as input variables to forecast %Diabetes, providing us with a way to compare the actual and predicted %Diabetes. Nevertheless, this model yields an R-squared value of 0.34, which is still below our desired level of satisfaction.

Quadratic model on log-transformed data:



R-squared (quadratic model on log-transformed data): 0.43

The model we created by applying a logarithmic transformation to the data for %Diabetes, alongside Inactivity and Obesity, yielded an R-squared value of 0.43, the highest we've seen thus far. Unless any other method produces a higher value, we will consider this model as the most suitable for predicting %Diabetes using our dataset..

Breusch-Pagan Test Output:

Breusch-Pagan Test p-value: 3.555846910402186e-05
Heteroscedasticity is detected (reject the null hypothesis).

The Breusch-Pagan Test used to measure the p-value and hence the pattern of residuals gives us a very small p-value which indicates that the data is heteroscedastic in nature.

T-Test output:

Reject the null hypothesis. There is a significant difference between the datasets.
T-statistic: -8.586734600367794
P-value: 1.960253729590773e-17

On running the t-test, we also get a very small p-value for the data points which also adds to the fact that the data is heteroscedastic in nature. However, It is important to note that the T-test result is not highly reliable as the data points are not normally distributed.

K-fold validation test Output:

Fold 1: R-squared = 0.3947
Fold 2: R-squared = 0.4278
Fold 3: R-squared = 0.1305
Fold 4: R-squared = 0.2200
Fold 5: R-squared = 0.1851
Mean R-squared: 0.2716
Standard Deviation of R-squared: 0.1180

The K-fold cross validation method was used with an iteration of 5-folds to find the R-squared value in an attempt to have better predictions of %Diabetes. The R squared value received (0.27)

was not the highest we have achieved so far (0.43 with quadratic model), hence this method was rejected.

Appendix C: code

```
import numpy as np
import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import PolynomialFeatures
from sklearn.metrics import r2_score
import matplotlib.pyplot as plt
merged_df['log_DIABETIC'] = np.log(merged_df['% DIABETIC'])
X = merged_df[['% INACTIVE', '% OBESE']]
y = merged_df['log_DIABETIC']
poly = PolynomialFeatures(degree=2)
X_poly = poly.fit_transform(X)
model = LinearRegression()
model.fit(X_poly, y)
y_pred = model.predict(X_poly)
r2 = r2_score(y, y_pred)
print(f'R-squared (quadratic model on log-transformed data): {r2:.2f}')
```

Output: R-squared (quadratic model on log-transformed data): 0.43

Contribution:

Tiyasa Saha: Worked on the Issues, Discussion, Methods and Results sections. Also self-plotted the graphs to analyze the data using the various methods discussed in the report.

Kanishka Patre : Worked on the Issues, Findings, Methods, Code and Result sections. Plotted graphs and used various regression analysis models and tests to analyze the data.

Srikanth Koncherry : Worked on identifying issues, writing code for and the analysis models and producing the graphs for them

Gautam Marathe : Worked on initial analyses, cleaning data, analyzing and looking for different models to fit the data on, testing various fits for their errors, testing non linear models on the data to describe trends between predictors.