Exploring Diabetes Data by Analysing Obesity and Inactivity's Impact on Health Disparities

Addressed to:

Dr. Dylan George

Director for the Center for Forecasting and Outbreak Analytics (CFA)

Done by:

Meggna Karthikeyan

Issues:

1. **Data Understanding:**
    - What is the content and structure of the data provided by the Centers for Disease Control and Prevention (CDC)?
    - Can we explain the significance of the variables "Diabetes," "Obesity," and "Inactivity" in the context of public health?
    - Can we explain the current understanding of the causes of diabetes, especially in relation to factors like obesity and physical inactivity?

2. To analyse the given CDC Diabetes 2018 data to predict for people how the diabetes rate affects or has any relation with inactivity and obesity. What specific research questions or hypotheses can we address using this data?

3. How do these health metrics (Diabetes, Obesity, Inactivity) provide valuable insights into the health behaviors and outcomes of populations?

4. How does this comprehensive analysis contribute to our understanding of diabetes data, and what potential applications does it have in the field of public health?

5. How do the limitations of the dataset impact the accuracy and reliability of the linear and polynomial regression models?

6. What does the R-squared value of approximately 50% tell us about the dataset's ability to predict outcomes related to diabetes based on inactivity and obesity?

7. **Interpretation and Insights:**
    - What are the conclusions we drew from the analyses?
    - Are there any notable findings or patterns that stand out?

Findings:

We're still not quite sure what exactly causes diabetes, but we do know that being overweight and not moving around much can make it more likely to happen. Our data shows there's a little connection between being overweight, being inactive, and getting diabetes, but we need more data to say for sure.

The given CDC dataset has shown a correlation among obesity, physical inactivity, and diabetes. Nevertheless, to substantiate a robust correlation, additional data is needed.

To understand the relationship amongst them, we've used linear and polynomial regression models to forecast these variables against each other. However, our modelling outputs has constraints due to the inconsistent dataset.

For more clarity, we've employed data visualization techniques, particularly scatter plots that show regression lines generated via both linear and polynomial algorithms. These visualizations provide an intuitive grasp of the predictive capabilities of our models.

Despite having the models, the connection between these variables is not clear. A positive correlation coefficient between 0 and 1 suggests that as one variable increases, the other tends to increase as well, but the relationship is not very strong. A correlation coefficient of 0 would indicate no linear relationship between the variables. A correlation coefficient of 1 would indicate a perfect positive linear relationship, where one variable increases in direct proportion to the other.

In our case, a coefficient of 0.47 indicates a moderate, but not strong, positive correlation.

So, with a correlation coefficient of 0.47, we've concluded that there is a moderate positive correlation between the two variables, but it's not a very strong or highly predictive relationship. The correlation coefficient between obesity and physical inactivity hovers around 0.47, signifying a noteworthy but not all-encompassing interdependence. This indicates a positive correlation between two variables, but it is not a very strong correlation. To figure all this out, we've been using linear and polynomial regression models. But our dataset isn't huge, so these models aren't super accurate.

Finally, our findings underscore that, when considering the input values of %Inactivity and %Obesity, our dataset elucidates only approximately 50% of the outcomes accurately, as indicated by the R-squared value.

Discussions:

The findings from our analysis provide insights into the relationship between obesity, physical inactivity, and the risk of diabetes. The analysis has provided valuable insights into the relationship between obesity, physical inactivity, and the risk of diabetes. While the exact causes of diabetes continue to elude us, it is becoming increasingly evident that being overweight and leading a sedentary lifestyle can elevate the likelihood of developing this condition. These lifestyle factors, notably obesity and physical inactivity, are emerging as significant contributors to the diabetes puzzle.

Our study, based on the CDC dataset, has brought to light a correlation between obesity, physical inactivity, and the occurrence of diabetes. However, it is important to temper our expectations when considering this correlation. The correlation coefficient of 0.47, while indicating a moderate positive relationship, suggests that as obesity and physical inactivity increase, the risk of diabetes tends to follow suit. Yet, we must acknowledge that this relationship lacks the strength required to make highly predictive inferences about an individual's diabetes risk.

To deepen our understanding of this association, we employed linear and polynomial regression models. Unfortunately, our findings are hampered by the constraints imposed by the dataset. While scatter plots with regression lines offered some intuitive insights into the predictive capabilities of our models, the precise nature of the connection between these variables remains somewhat elusive.

The modest R-squared value, indicating that our dataset accurately explains only approximately 50% of the observed outcomes, underscores the limitations of our analysis. This suggests that there are additional factors beyond obesity and physical inactivity that play a role in the development of diabetes, and our dataset falls short of capturing the full complexity of these interactions.

In summary, our findings reveal a moderate positive correlation between obesity, physical inactivity, and the risk of diabetes. However, this correlation falls short of being a highly reliable predictor of diabetes risk. To gain a more comprehensive understanding of the myriad factors influencing diabetes, further data collection and research endeavours are essential.

Methods:

The data for this analysis was obtained from the Centres for Disease Control and Prevention (CDC), a reputable source for health-related data in the United States.

This dataset provides an extensive overview of health indicators across various counties, offering a rich foundation for statistical analysis and model building.

In terms of variable creation, each variable represents a specific health metric. "Diabetes" refers to the percentage of the population in a given county that has been diagnosed with diabetes, a chronic health condition that affects the body's ability to process blood glucose. "Obesity" indicates the percentage of individuals in the county who are classified as obese based on their Body Mass Index (BMI), a key measure of an individual's body fat based on height and weight. "Inactivity" represents the percentage of people who do not meet recommended physical activity levels, which can be a contributing factor to both obesity and diabetes. These variables are crucial as they offer measurable insights into the health behaviors and outcomes of populations, forming the basis for the statistical models employed in the analysis.

Analytic methods used in this study, such as linear regression and polynomial regression, allow us to examine and quantify the relationships between these variables. Linear regression models the relationship in a straight-line manner, while polynomial regression can model more complex, curvilinear relationships. By utilizing these methods, we can better understand how changes in obesity and inactivity are associated with changes in diabetes prevalence, which in turn can inform public health strategies and interventions.

To prepare the data for analysis, we merged and cleaned the necessary columns on CDC dataset, to gain insights into the relationships between various factors, ensuring data integrity and consistency. Our dataset included information on the percentage of obesity, inactivity, and diabetic individuals in different counties.

After that, we split the dataset into training and testing sets which was important in evaluating the performance of our model.

For modeling, we utilized the Linear Regression class from the sklearn.linear_model module, which allowed us to create a linear regression model. To assess the model's performance, we calculated R-squared values and mean absolute error, providing a quantitative measure of how well the model fitted the data.

After performing simple linear regression, we explored polynomial regression techniques to capture potential non-linear relationships between the input features and the target variable. This expanded our modeling capabilities and enabled us to better represent complex data patterns.

Our analysis also includes correlation values and basic statistics, that helps us the relationships between the variables and their distributions. Polynomial

regression, in particular, helped us to identify and model non-linear associations within the dataset.

Overall, this comprehensive analysis offers a deeper understanding of multiple linear regression and its practical application in studying diabetes data. The insights gained and methods employed provide a solid foundation for further research and informed decision-making in the critical field of public health.

Results:

Obtaining regression values from Excel for better analysis of data to gain its insights and further work on the actual code.

SUMMARY OUTPUT

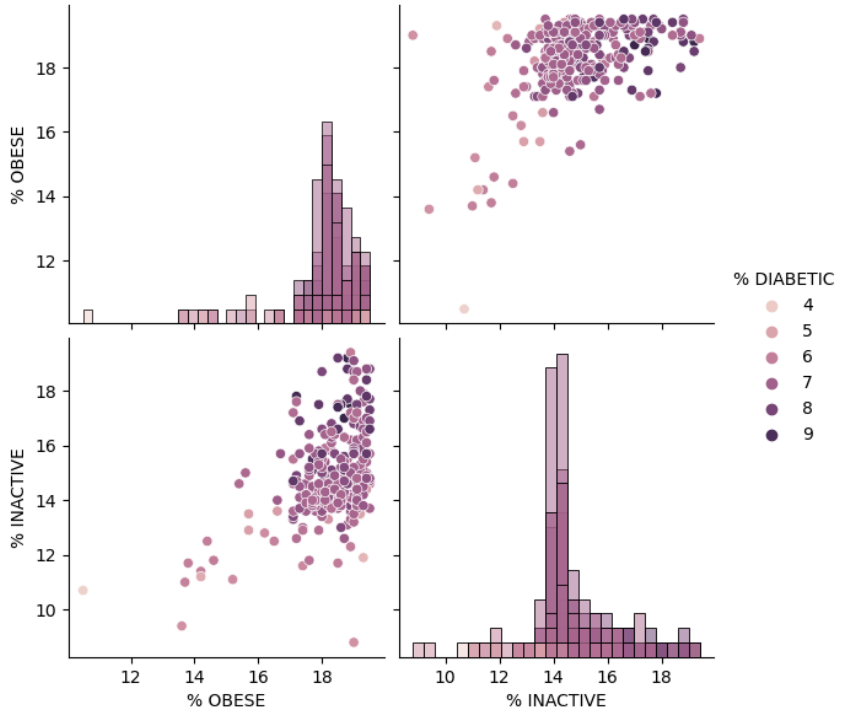
Regression Statistics	
Multiple R	0.583729108
R Square	0.340739671
Adjusted R Square	0.336983202
Standard Error	0.593139754
Observations	354

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	63.82442908	31.91221454	90.70743327	1.75587E-32
Residual	351	123.4869834	0.351814767		
Total	353	187.3114124			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	1.653599152	0.562219841	2.94119672	0.003486461	0.547855775	2.759342528	0.547855775	2.759342528
% OBESE	0.111062966	0.034797846	3.191662086	0.001542263	0.042624457	0.179501475	0.042624457	0.179501475
% INACTIVE	0.232469919	0.023193817	10.02292618	5.8475E-21	0.186853582	0.278086257	0.186853582	0.278086257

Output for plotting Data points for the given Dataset



Output for Basic Statistics

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

Output for Standard Deviation, Skewness and Kurtosis

```
Standard Deviation of Residuals: 0.727412045335248
Skewness of Residuals: -0.04881196867705582
Kurtosis of Residuals: 2.7884224412882883
```

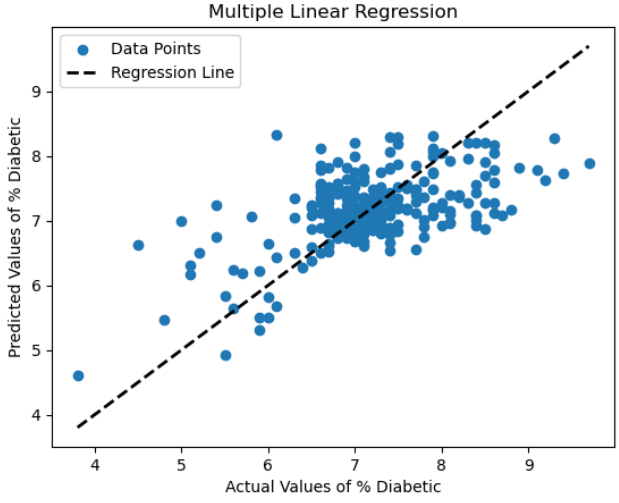
Output for Correlation matrix

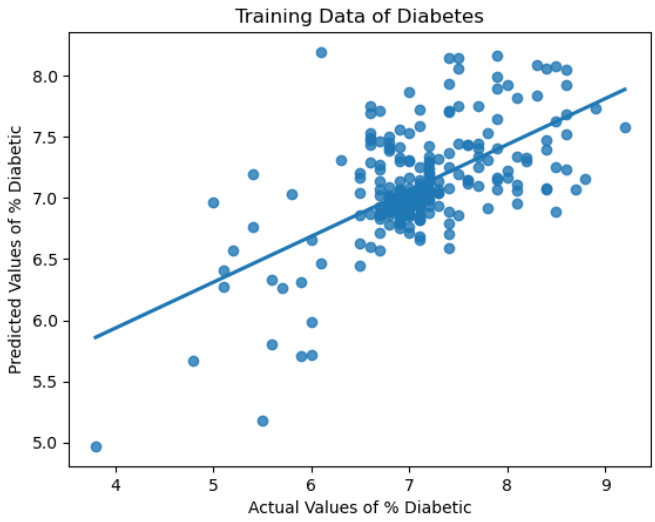
	% DIABETIC	% OBESE	% INACTIVE
% DIABETIC	1.000000	0.389941	0.567104
% OBESE	0.389941	1.000000	0.472656
% INACTIVE	0.567104	0.472656	1.000000

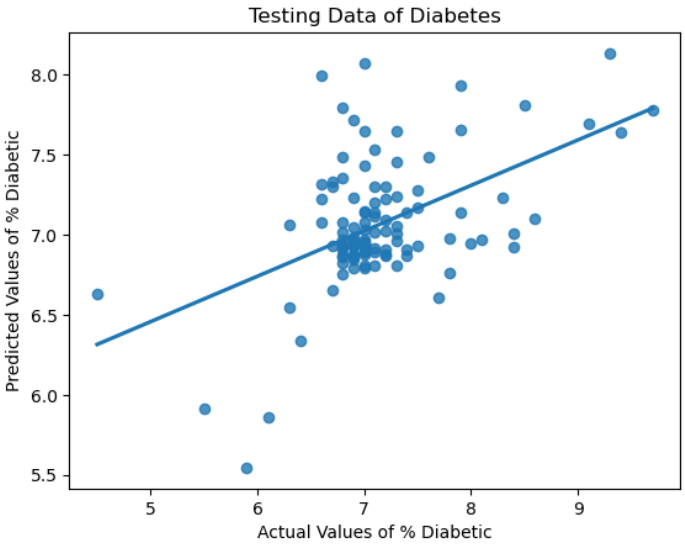
Output for KFold Cross-validation mean R-squared value

```
5-fold cross-validation mean R-squared value: 0.30131914105371216
```

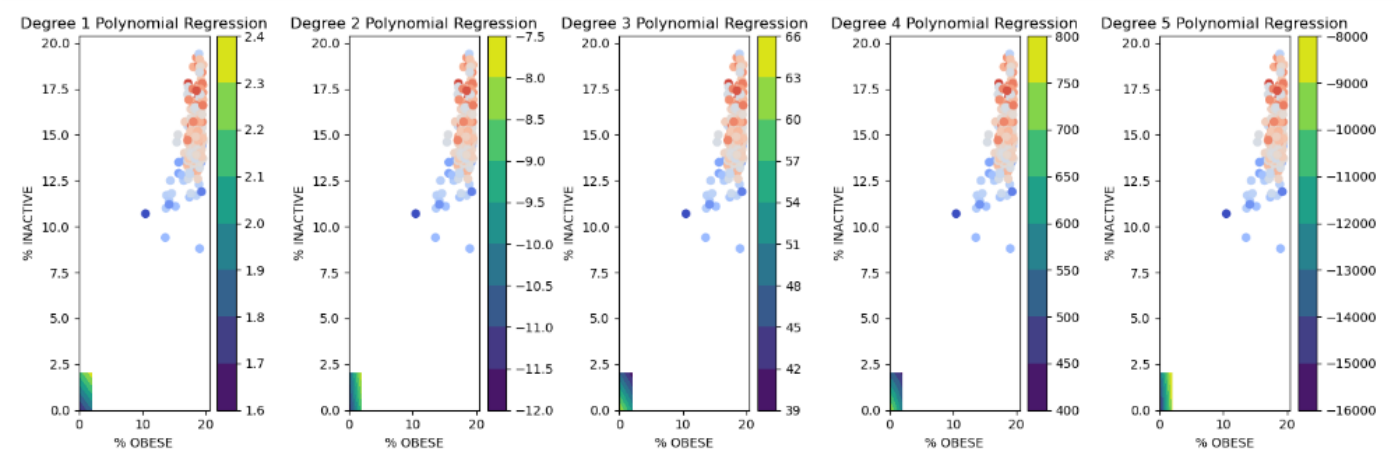
Outputs for Multiple Linear Regression







Output for Polynomial Regression



Output for R2 values of Polynomial Regression V/S Degrees

```
R2 Value for Polynomial Regression (Degree 1): 0.3672741054718953
R2 Value for Polynomial Regression (Degree 2): 0.4040979297531323
R2 Value for Polynomial Regression (Degree 3): -1.1701480988319473
R2 Value for Polynomial Regression (Degree 4): -51.26825515624212
R2 Value for Polynomial Regression (Degree 5): -57.14783147971228
```