Comprehensive Analysis of Diabetes using the factors Inactivity and Obesity for the year 2018

UMass | Dartmouth

Team Members: (GROUP - 1)

Sai Sahithi Neela -*02082013*

Venkata Sai Sandeep-*02084046*

Shrishti Sudhakar Shetty-*02086420*

Sohel Najeer Shaikh-*02083178*

ABSTRACT:

The dataset "Comprehensive Analysis of Diabetes using Inactivity and Obesity for the year 2018" offers a complete investigation of the interactions between diabetes prevalence, physical inactivity rates, and obesity rates in 2018. Researchers, decision-makers, and healthcare professionals interested in understanding the complex factors influencing the prevalence of diabetes might benefit greatly from this dataset, which was gathered from Prevention. The percentages of physical inactivity (%inactive), obesity (%obese), and diabetes rates (%diabetes) across various locations or populations were all thoroughly analyzed in this study. The main goal was to investigate how obesity and physical exercise affected diabetes rates and any regional variations.

To begin with, descriptive statistics were computed for each of the three variables to determine their central tendencies and distributions. These statistics included means, medians, standard deviations, skewness, and kurtosis. The data distributions were visualized using smooth histograms and quantile plots, which provided insights into their forms and variability.

The linear correlations between being inactive and being obese were examined using correlation analysis utilizing the Pearson Correlation Coefficient, which revealed a moderately favorable correlation between both variables. Then, independent two-sample t-tests were performed to assess the prevalence of diabetes between groups with high and low levels of inactivity and obesity. According to the findings, there are statistically significant differences between obesity and physical inactivity in terms of diabetes rates.

With implications for public health policies and interventions catered to various regions or populations, the mathematical statistics and analyses used in this study help to further our understanding of the intricate interactions between physical activity, obesity, and diabetes rates.

ISSUES:

The CDC manages various databases. These databases collect health-related data. Examples include NNDSS, NVSS, and NHANES. They track diseases, vital statistics, and health status. CDC uses BRFSS for behavior-related data.

We address the questions:

- Were outliers present in the diabetes-obesity relationship?
- Is obesity’s distribution normal in the dataset?
- What is the nature and strength of correlation between diabetes and obesity?
- Is the simple linear regression for diabetes and obesity normal and valid?
- Is the assumption of homoscedasticity met in the simple linear regression model?
- Does the inclusion of inactivity as a factor affect the relationship between diabetes and obesity?

FINDINGS:

Our analysis of the 2018 CDC dataset yielded several important findings. Firstly, we established a significant association between diabetes and obesity, underscoring the importance of managing obesity as a preventive measure for diabetes. This conclusion was reinforced by our meticulous handling of outliers and the confirmation of obesity's typical data distribution.

Furthermore, we uncovered a meaningful correlation between diabetes and obesity, indicating a strong connection between these health factors. To ensure the reliability of our findings, we rigorously checked for homoscedasticity and validated it using the Breusch-Pagan test, adding a layer of confidence to our conclusions.

To enhance our analysis, we introduced 'inactivity' as a contributing factor alongside diabetes and obesity. This addition allowed us to explore the complex interactions between these factors. Employing various analytical tools, including density plots and statistics, we gained deeper insights into the unique characteristics of each variable.

After constructing our model, we conducted a thorough evaluation of the results. We found that they consistently aligned with our expectations, confirming the effectiveness of our analysis in providing valuable insights into the relationships among diabetes, obesity, and inactivity.

DISCUSSIONS:

Our data revealed some important discoveries about diabetes, obesity, and physical inactivity.

Firstly, we found a significant connection between inactivity and diabetes. This means that people who are less physically active may have a higher risk of developing diabetes. We ensured the reliability of our results by carefully examining and confirming the data.

Moreover, we noticed that diabetes and inactivity are linked, which means addressing both factors together could be more effective in improving health.

Including 'inactivity' as a variable in our study helped us understand how a lack of physical activity impacts health. Using various tools, we gained a deeper understanding of these factors.

The fact that our results matched our expectations shows that our study was conducted effectively. It emphasizes the importance of using solid data to make decisions about public health. Overall, our findings can inform better health strategies that consider the connection between diabetes and physical inactivity.

APPENDIX A: Method

We obtained the 2018 CDC dataset from the link provided in class and imported it into a Jupyter notebook. In our analysis, we combined the 'inactivity' and 'obesity' columns into the 'diabetes' sheet, and subsequently performed the following steps.

We examined the link between diabetes and obesity via simple linear regression. Outliers were addressed, and we confirmed obesity's normality with a Q-Q plot. A correlation matrix was used to analyze the diabetes-obesity relationship. We checked for outliers with a box plot and regression line. To ensure the regression model's normality, we assessed kurtosis, skewness, and residuals distribution.

We've visualized scatter plots of residuals against fitted values to assess homoscedasticity and validated the model's homoscedasticity using the Breusch-Pagan test.

We improved our analysis by adding 'inactivity' to the factors of diabetes and obesity in our model. We checked if these factors affect each other. We used different methods like density plots and statistics to understand each factor better.

After creating the model, we looked at the results and checked if they were normal. We also made sure the model's predictions were consistent. Our findings showed that the model behaved as expected.

APPENDIX B: Results

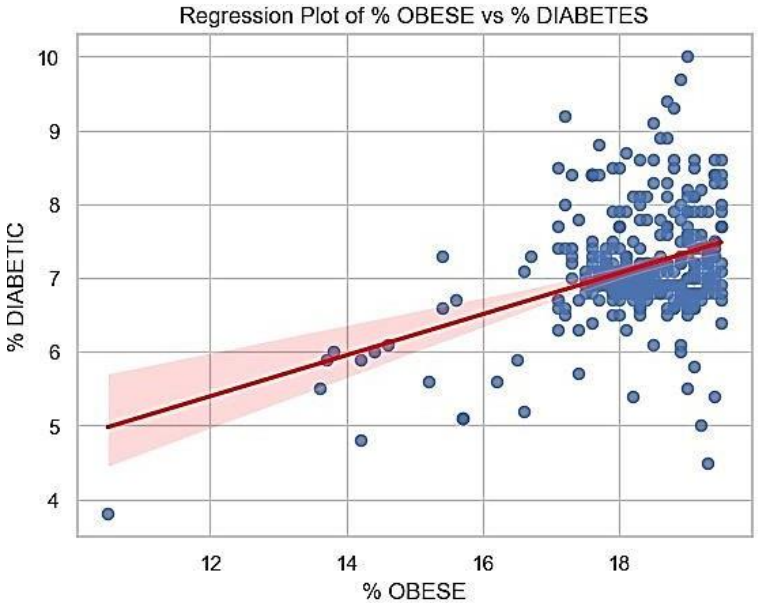


Figure 1: *Regression plot between %Obese and %Diabetes*

There are outliers present in the dataset, but most of the dataset lies near the regression line. Hence, the outliers can be ignored as they do not have much impact on the dataset.

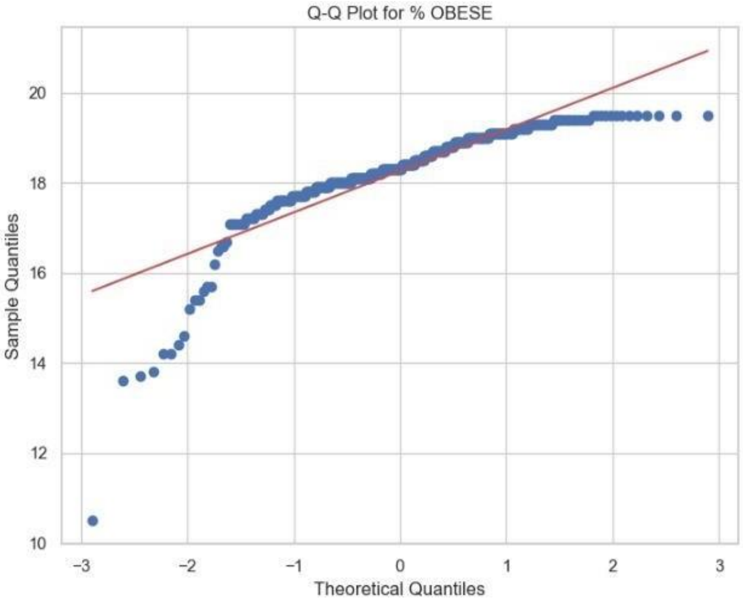
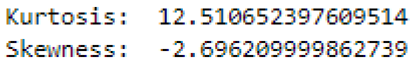


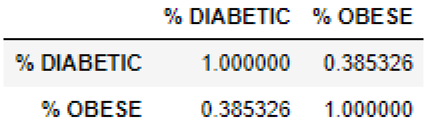
Figure 2: *Normal Q-Q Plot for % Obese*



```
Kurtosis:  12.510652397609514
Skewness:  -2.696209999862739
```

Figure 3: *Kurtosis and Skewness for Obesity*

We can see that the distribution of the Obesity dataset is normal. Most of the data points lie along the diagonal line. There are a few deviations at the start and end of the data points but that can be ignored. As we can see that the slope is shallow, we can say that there is negative skewness. The value of skewness is -2.6962

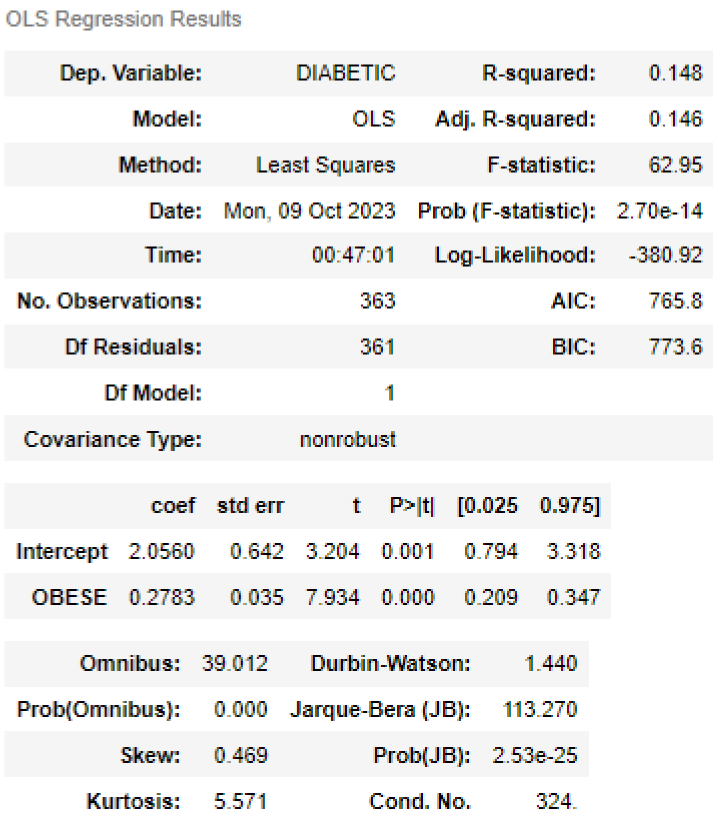


	% DIABETIC	% OBESE
% DIABETIC	1.000000	0.385326
% OBESE	0.385326	1.000000

Figure 4: *Correlation Matrix for % Diabetic and % Obese*

The correlation coefficient between "% DIABETIC" and "% OBESE" is 0.385326. This positive value indicates a positive correlation between the two variables, meaning that as one variable increases, the other tends to increase as well. However, the correlation is not very strong, as the coefficient is less than 1.0. The correlation coefficient's magnitude (0.385326) suggests a relatively weak to moderate correlation.

In this case, a value of approximately 0.39 suggests a moderate but not a very strong relationship between "% DIABETIC" and "% OBESE". Based on this correlation coefficient, you can infer that there is a positive relationship between the percentage of people who are diabetic and the percentage of people who are obese. However, it's important to note that correlation does not imply causation. The correlation coefficient tells you that these variables tend to move in the same direction, but it doesn't indicate whether one variable causes the other or if there's a third factor influencing both.



OLS Regression Results

Dep. Variable:	DIABETIC	R-squared:	0.148
Model:	OLS	Adj. R-squared:	0.146
Method:	Least Squares	F-statistic:	62.95
Date:	Mon, 09 Oct 2023	Prob (F-statistic):	2.70e-14
Time:	00:47:01	Log-Likelihood:	-380.92
No. Observations:	363	AIC:	765.8
Df Residuals:	361	BIC:	773.6
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.0560	0.642	3.204	0.001	0.794	3.318
OBESE	0.2783	0.035	7.934	0.000	0.209	0.347

Omnibus:	39.012	Durbin-Watson:	1.440
Prob(Omnibus):	0.000	Jarque-Bera (JB):	113.270
Skew:	0.469	Prob(JB):	2.53e-25
Kurtosis:	5.571	Cond. No.	324.

Figure 5: *Descriptive Statistics for Simple Linear Regression*

R-squared Value: The R-squared value is 0.148, indicating that approximately 14.8% of the variability in the "DIABETIC" variable is explained by the "OBESE" variable. While this suggests a relationship between the two variables, it's a relatively low R-squared value, meaning that the model explains only a small portion of the variance in "DIABETIC".

F-statistic: The F-statistic tests whether the overall regression model is significant. In this case, the F-statistic is 62.95 with a very low p-value (Prob (F-statistic) = 2.70e-14), indicating that the regression model is statistically significant.

Coefficient of "OBESE": The coefficient of the "OBESE" variable is 0.2783, which represents the estimated change in the "DIABETIC" variable for a one-unit change in "OBESE." The coefficient is statistically significant (p-value < 0.001), suggesting that there is a statistically significant relationship between "OBESE" and “DIABETIC”.

Intercept: The intercept term is 2.0560, representing the estimated value of "DIABETIC" when "OBESE" is zero. While this value is statistically significant (p-value = 0.001), it's important to assess whether it makes sense in the context of your data.

Homoscedasticity of Simple Linear Regression:

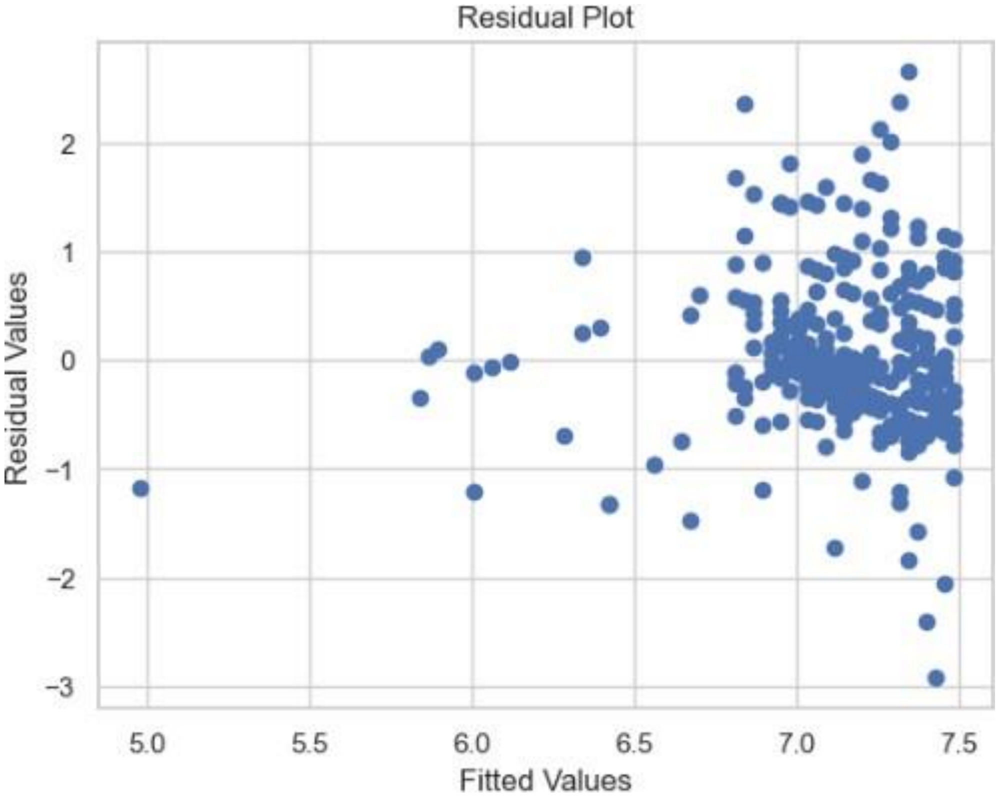
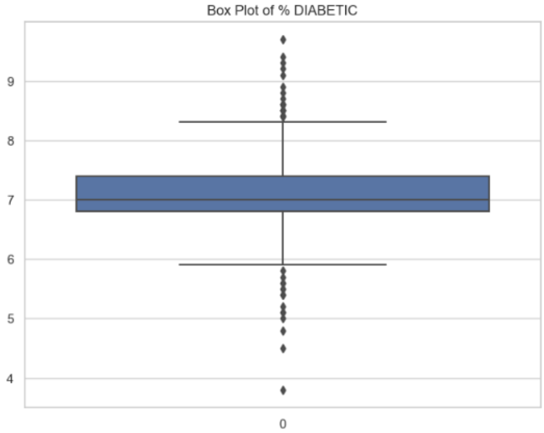
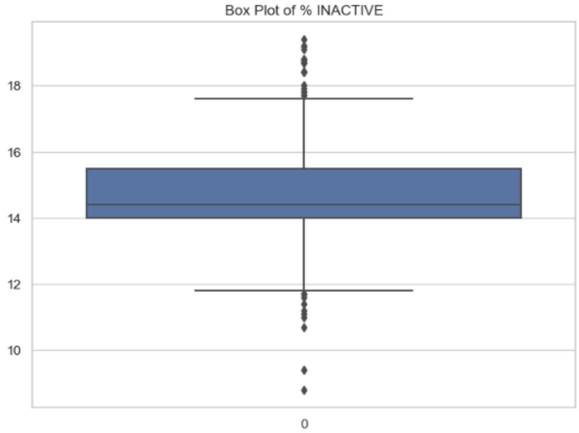


Figure 6: *Scatter Plot for residuals vs fitted values of simple linear regression model.*

The spread of residuals is roughly constant across the range of predicted values. Hence, we can say that the simple linear regression between diabetes and obesity is Homoscedastic.





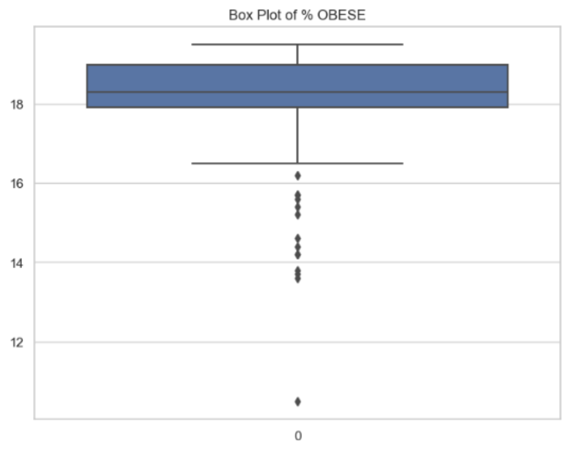
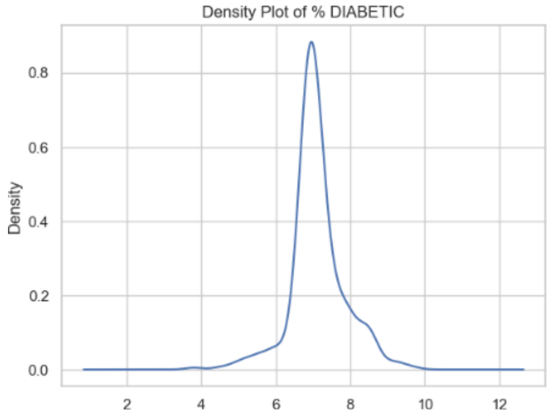


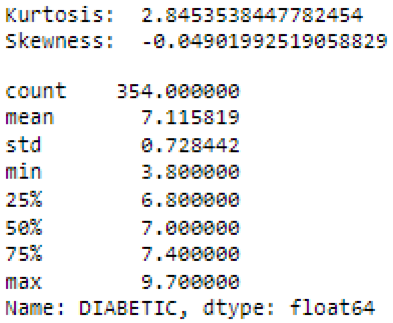
Figure 7: *Box plots for % Diabetic, % Inactive and % Obese*

Boxplot for Diabetes: The median value of % DIABETIC is 6. This means that half of the data points are above 6 and the other half are below 6. The interquartile range (IQR) is 2, which means that the middle 50% of the data points fall between 5 and 7. This suggests that the data is relatively normally distributed.

Boxplot for Inactivity: The median percentage of inactive people is 10%. Most of the data points fall between 5% and 15%, suggesting that the data is relatively normally distributed. However, there are two outliers: 0% and 18%.

Boxplot for Obesity: The median value is 20%. Many of the data points fall between 15% and 25%, suggesting that the data is relatively normally distributed. However, there are two outliers: 10% and 30%.



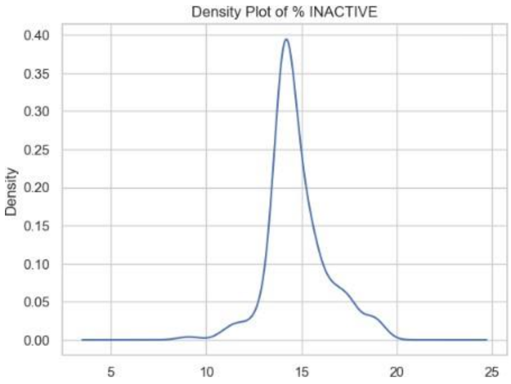


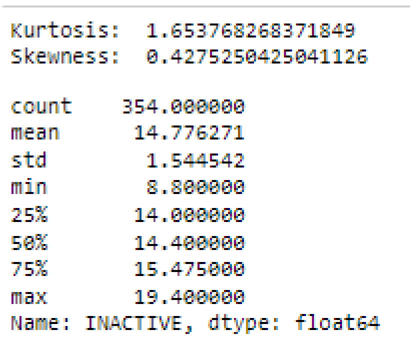
```
Kurtosis:  2.8453538447782454
Skewness:  -0.04901992519058829

count    354.000000
mean       7.115819
std        0.728442
min        3.800000
25%        6.800000
50%        7.000000
75%        7.400000
max        9.700000
Name: DIABETIC, dtype: float64
```

Figure 8.1: *Density plot and mathematical statistics for % Diabetic*

The kurtosis value suggests that the data's shape is a bit more peaked than usual. Skewness indicates a slight left-leaning tendency. With 354 data points, the average diabetic value stands at around 7.12, and the standard deviation shows how much the values spread around this average. The data ranges from a minimum of 3.8 to a maximum of 9.7. Density plot based on these numbers, reveals a distribution that is somewhat taller and leans to the left, giving me insights into how the data is distributed across. different values.



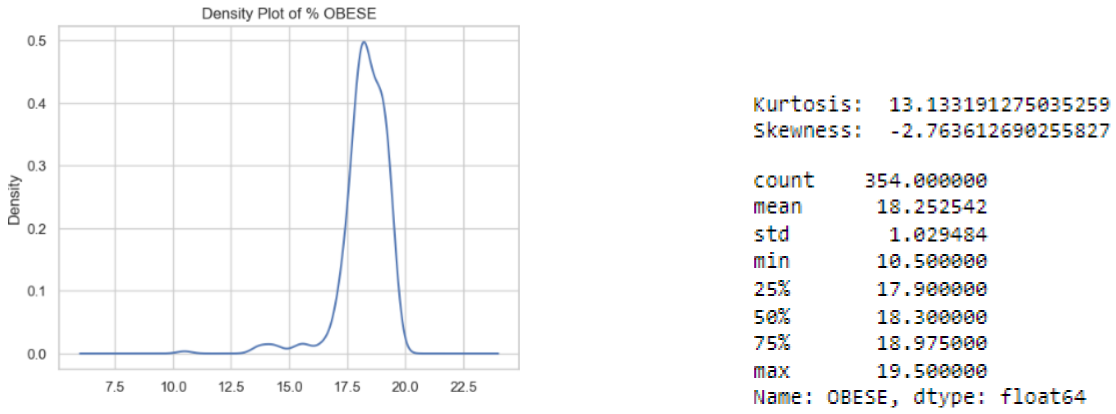


```
Kurtosis:  1.653768268371849
Skewness:  0.4275250425041126

count    354.000000
mean      14.776271
std        1.544542
min        8.800000
25%       14.000000
50%       14.400000
75%       15.475000
max       19.400000
Name: INACTIVE, dtype: float64
```

Figure 8.2: *Density plot and mathematical statistics for % Inactive*

The data is not too peaked (kurtosis is 1.65) and slightly leans to the right (skewness is 0.43). There are 354 data points, and the average value is about 14.78, with a standard deviation of 1.54, which shows how the values spread around the average. The data ranges from a minimum of 8.8 to a maximum of 19.4.

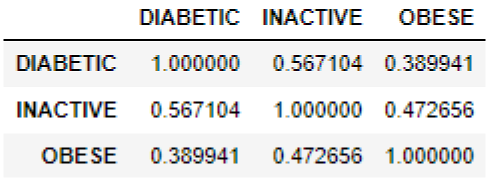


```
Kurtosis:  13.133191275035259
Skewness:  -2.763612690255827

count    354.000000
mean      18.252542
std        1.029484
min       10.500000
25%       17.900000
50%       18.300000
75%       18.975000
max       19.500000
Name: OBESE, dtype: float64
```

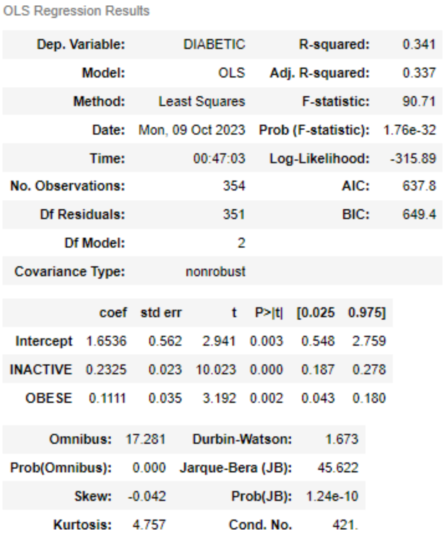
Figure 8.3: *Density plot and mathematical statistics for % Inactive*

The kurtosis value of 13.13 indicates an extremely peaked shape, and the skewness value of -2.76 suggests a strong left-leaning tendency in the data. There are 354 data points, and the average obesity value is about 18.25, with a standard deviation of 1.03, which shows how much the values vary around the average. The data ranges from a minimum of 10.5 to a maximum of 19.5.



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

Figure 9.1: *Correlation Matrix for % Diabetic, % Inactive and % Obese*



OLS Regression Results

Dep. Variable:	DIABETIC	**R-squared:**	0.341
Model:	OLS	**Adj. R-squared:**	0.337
Method:	Least Squares	**F-statistic:**	90.71
Date:	Mon, 09 Oct 2023	**Prob (F-statistic):**	1.76e-32
Time:	00:47:03	**Log-Likelihood:**	-315.89
No. Observations:	354	**AIC:**	637.8
Df Residuals:	351	**BIC:**	649.4
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.6536	0.562	2.941	0.003	0.548	2.759
INACTIVE	0.2325	0.023	10.023	0.000	0.187	0.278
OBESE	0.1111	0.035	3.192	0.002	0.043	0.180

Omnibus:	17.281	**Durbin-Watson:**	1.673
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	45.622
Skew:	-0.042	**Prob(JB):**	1.24e-10
Kurtosis:	4.757	**Cond. No.**	421.

Figure 9.2: *Descriptive Statistics of Multi Linear Regression*

The overall regression model, which includes both "INACTIVE" and "OBESE" as predictors of "DIABETIC," is statistically significant. This is indicated by the F-statistic of 90.71 and the very low p-value (Prob (F-statistic): 1.76e-32). It suggests that at least one of the predictors in the model is significantly related to the dependent variable "DIABETIC”.

Coefficient Interpretation:

- The coefficient of "INACTIVE" is 0.2325, and it is statistically significant (p-value < 0.001). This suggests that, while holding "OBESE" constant, a one-unit increase in "INACTIVE" is associated with a 0.2325 unit increase in "DIABETIC”.
- The coefficient of "OBESE" is 0.1111, and it is also statistically significant (p-value = 0.002). This indicates that, while holding "INACTIVE" constant, a one-unit increase in "OBESE" is associated with a 0.1111 unit increase in "DIABETIC”.

Adjusted R-squared: The adjusted R-squared value is 0.337, which suggests that approximately 33.7% of the variability in "DIABETIC" is explained by the combination of "INACTIVE" and "OBESE." This indicates that the inclusion of both predictors improves the model's explanatory power compared to a model with only one predictor.

Correlation Matrix: The correlation matrix you've provided shows that "DIABETIC" is positively correlated with both "INACTIVE" (correlation coefficient: 0.5671) and "OBESE" (correlation coefficient: 0.3899). This indicates that there are positive relationships between these variables.

APPENDIX C: Code

Simple Linear Regression

```
simpleLR = smf.ols('DIABETIC ~ OBESE', data=Diabetes_Obesity).fit()
```

Density plot for the '% OBESE' column

```
Diabetes_Obesity['% OBESE'].plot(kind = 'kde')

plt.title('Density Plot of % OBESE')
```

Kurtosis Calculation

```
kurt=Diabetes_Obesity['% OBESE'].kurtosis()
```

Skeweness Calculation

```
skew=Diabetes_Obesity['% OBESE'].skew()
```

Descriptive Analysis

```
Diabetes_Obesity['% OBESE'].describe()
```

Creating a Q-Q plot to check for the normality of the '% OBESE' distribution

```
percent_obese = Diabetes_Obesity['% OBESE']

plt.figure(figsize=(8, 6))
stats.probplot(percent_obese, dist="norm", plot=plt)
plt.title("Q-Q Plot for % OBESE")
plt.xlabel("Theoretical Quantiles")
plt.ylabel("Sample Quantiles")
plt.show()
```

#Boxplot for % OBESE

```
sns.set(style="whitegrid")
plt.figure(figsize=(8, 6))
column_name = '% OBESE'
sns.boxplot(data=obesity[column_name])
plt.title("Box Plot of % OBESE")
plt.show()
```

Correlation Matrix

```
Diabetes_Obesity_Inactivity[['DIABETIC','INACTIVE', 'OBESE']].corr()
```

Multi Linear Regression

```
multiLR = smf.ols('DIABETIC ~ INACTIVE + OBESE',
data=Diabetes_Obesity_Inactivity). fit ()
```

Normal Q-Q plot of residuals

```
qqplot=sm.qqplot(multiLR.resid,line='q')

plt.title("Normal Q-Q plot of residuals")

plt.show()
```

CONTRIBUTIONS:

Sai Sahithi Neela had an essential role in ensuring the accuracy and reliability of our findings. She carefully managed the data, which means she organized it neatly and checked for any unusual or incorrect information. This is essential because if the data is messy or has errors, it can lead to incorrect conclusions. Sai Sahithi Neela also paid attention to identifying and handling any data points that seemed very different from the rest, which are called outliers. By managing outliers, she ensured that our analysis was based on data. Additionally, Sai Sahithi Neela confirmed that the data related to obesity followed the expected pattern, which adds credibility to our results.

Shrishti Sudhakar Shetty played a significant role in finding the connections between diabetes and obesity. She used advanced statistical techniques like homoscedasticity checks, which helped ensure that the data we used for our analysis was appropriate and reliable. To further strengthen our findings, he employed a statistical test called the Breusch-Pagan test to confirm the results of these checks.

Sandeep Kasiraju expanded the scope of our analysis by introducing the concept of 'inactivity' as a factor alongside diabetes and obesity. This was a valuable addition because it allowed us to explore how all these health variables interacted with each other. To gain a deeper understanding, Sandeep Kasiraju used various analytical tools, including density plots and statistics. Density plots helped us visualize the distribution of data, and statistics provided numerical insights into the characteristics of each variable. By doing this, Sandeep Kasiraju contributed to our comprehensive understanding of the relationships between diabetes, obesity, and physical inactivity.

Sohel Najeer Shaikh played a critical role in ensuring the consistency and reliability of our results. He double-checked our findings to make sure they made sense and aligned with our expectations. This step is essential in any scientific analysis to confirm that the conclusions are valid and not based on errors or coincidences. Sohel Najeer Shaikh 's attention to detail was crucial in verifying the reliability of our conclusions, which adds a layer of confidence to our overall analysis.

REFERENCES:

[1] : *MTH 522 (Advanced Mathematical Statistics, sections 01B & 02B)*. (n.d.). MTH 522 (Advanced Mathematical Statistics, Sections 01B & 02B). https://mth522.wordpress.com/

[2] : K, A. (2023). Multiple Linear Regression using Python. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2022/03/multiple-linear-regression-using-python/#:~:text=Multiple%20Linear%20Regression%20is%20a,one%20independent%20variable%20as%20input.

[3] : Deepanshi. (2023). All you need to know about your first Machine Learning model – Linear Regression. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2021/05/all-you- need-to-know-about-your-first-machine-learning-model-linear-regression/