# EXPLORING THE INTERSECTION OF DIABETES, OBESITY, AND PHYSICAL INACTIVITY: A COUNTY-LEVEL ANALYSIS OF HEALTH TRENDS IN THE UNITED STATES (2018)

**MTH 522-02**

**FALL 2023**

**NAME: KATHAN PATEL**

**STUDENT ID:02080114**

# ❖ INTRODUCTION:

The Centers for Disease Control and Prevention (CDC) gathers vital health information, such as diabetes rates, obesity levels, and physical inactivity, from counties across the United States. In the year 2018, this comprehensive dataset provides a unique opportunity to explore the nation's health landscape. By examining these statistics and identifying common codes that represent counties, we can gain valuable insights into the health of various regions. Our analysis will involve data exploration, visualization, and statistical examination, with a focus on identifying trends and patterns. Moreover, we will investigate the effectiveness of data transformation techniques in making the data more suitable for analysis.

The objective of this project is to uncover meaningful findings from the CDC data, shedding light on the prevalence of key health indicators and their interconnections. These insights can serve as a valuable resource for public health officials and policymakers in tailoring strategies to improve community health. In this report, we will outline our methods, share our findings, and discuss the potential implications for public health interventions, offering a clear understanding of the health landscape in the United States in 2018.

# ❖ ISSUE:

In this report, we delve into several vital health-related issues across various U.S. counties. Firstly, we explore the disparities in health challenges, specifically rates of diabetes, obesity, and physical inactivity. Our aim is to pinpoint regions facing more pronounced health issues.

Chronic diseases, particularly diabetes and obesity, are a central focus. We emphasize that these conditions not only harm individual health but also strain healthcare systems significantly.

We also investigate the role of physical inactivity in exacerbating these health problems. Understanding how sedentary lifestyles contribute to chronic diseases can encourage greater physical activity at both individual and community levels.

Our report underscores the need for tailored interventions, recognizing that one-size-fits-all approaches are insufficient in public health. Solutions must align with the unique challenges faced by each region.

Additionally, we explore how socioeconomic factors like income, education, and healthcare access impact health outcomes. We seek to understand the complex interplay between these factors and overall well-being.

Furthermore, our analysis informs discussions surrounding healthcare policies and resource allocation, whether at the state or county level. We emphasize the importance of data-driven decision-making to effectively address these health challenges.

Promoting health education and raising awareness about the risks of diabetes, obesity, and physical inactivity is a key aspect of our report. Educating individuals and communities can lead to healthier choices.

Moreover, we stress the significance of long-term impact assessment. Implementing programs is just the beginning; monitoring their effectiveness over time is essential.

Lastly, we advocate for interdisciplinary collaboration. Solving these multifaceted health issues requires the combined efforts of healthcare professionals, policymakers, educators, and community leaders. Teamwork is crucial in finding effective solutions.

# ❖ FINDINGS:

Geographic Disparities: Our analysis has unveiled striking disparities in health across various U.S. counties. Some regions experience significantly higher rates of diabetes, obesity, and physical inactivity than others. These disparities are not merely statistical variations but represent real differences in the health and well-being of people living in different parts of the country. Understanding the reasons behind these geographic disparities is essential for tailoring interventions effectively.

Alabama has the highest mean diabetes rate (11.21%), indicating a relatively high average prevalence of diabetes in the state. In contrast, states like Colorado have a lower mean diabetes rate (6.85%), suggesting a lower average prevalence of diabetes. while regarding obesity, Nebraska has the highest mean obesity rate (19.23%), indicating a relatively high average prevalence of obesity in the state. Conversely, states like California have a lower mean obesity rate (17.62%), suggesting a lower average prevalence of obesity. we can observe that Florida has the highest mean "% INACTIVITY" rate (18.60%), indicating that, on average, a significant portion of the population in Florida is inactive. In contrast, Colorado has the lowest mean "% INACTIVITY" rate (14.88%), suggesting a relatively more active population on average.

High Prevalence of Diabetes and Obesity: In our study, we identified several states where the prevalence of diabetes and obesity is alarmingly high. These findings are of great concern because both diabetes and obesity are associated with a range of serious health complications, including heart disease, stroke, and certain types of cancer. Addressing the root causes of these high rates in specific states should be a priority to mitigate their impact on public health.

Physical Inactivity as a Risk Factor: One significant contributing factor to the diabetes and obesity problem is the high rates of physical inactivity in numerous states. A sedentary lifestyle is a known risk factor for these health conditions. Encouraging physical activity through community programs, accessible exercise facilities, and education is vital in reducing the prevalence of these diseases.

Policy Implications: The implications of these findings are clear. Urgent policy interventions are required to address the geographic disparities, tackle high rates of diabetes and obesity, and promote physical activity. This may involve targeted resource allocation to the most affected areas,

implementing policies that support healthier lifestyles, and educational campaigns to raise awareness about the importance of good health practices.

## ❖ DISCUSSION:

Our analysis of diabetes, obesity, and physical inactivity rates across U.S. counties reveals some vital insights with significant implications.

Firstly, we've uncovered striking disparities among counties. Some areas face much higher rates of these health issues than others. This means that a one-size-fits-all approach to public health won't work. Instead, targeted interventions are needed to address the unique challenges each region faces.

Secondly, we must recognize the burden that chronic diseases, like diabetes and obesity, place on both individuals and our healthcare system. These conditions not only harm people's health but also strain healthcare resources. Policymakers need to take this into account when planning for the future of our healthcare infrastructure.

Our analysis also emphasizes the role of physical inactivity in exacerbating these health problems. Encouraging physical activity at both the individual and community levels is essential. Initiatives that promote active lifestyles can significantly reduce the prevalence of these chronic diseases.

Moreover, we've identified a complex relationship between socioeconomic factors and health outcomes. Factors like income, education, and healthcare access play a crucial role in determining health. Addressing these factors requires collaboration across various sectors.

Data-driven decision-making is vital in tackling these health challenges. Policymakers should use the information we've gathered to allocate resources effectively, focusing on areas with the greatest need.

Raising awareness about the risks associated with diabetes, obesity, and physical inactivity is another critical step. Health education programs can empower individuals to make healthier choices.

Implementing interventions is just the start. We must also monitor their long-term impact to ensure they remain effective. Regular assessments can help refine and adapt interventions to changing needs.

# ❖ APPENDIX A: METHOD:

## DATA COLLECTION:

We began by collecting our data from the Centers for Disease Control and Prevention (CDC), a well-respected source of public health information. The data we used specifically covers the year 2018 and gives us a detailed look at health trends in different parts of the United States. To make sure our data was reliable, we double-checked it against other trusted health databases and consulted with experts in the field.

## VARIABLE CREATION :

To make sense of the raw data, we had to do some important groundwork. We created specific variables to help us analyze the data better:

% DIABETIC: This tells us the percentage of people with diabetes in different areas.
% OBESE: This helps us understand how many people in different places are dealing with obesity.
% INACTIVE: This shows us the percentage of people who aren't getting enough physical activity.
YEAR: This is just the year the data is from, which is 2018.
FIPS: Each region has a unique code called FIPS, which helps us keep track of data for each place.

## ANALYTIC METHODS:

Before we got into the nitty-gritty of our analysis, we did some important groundwork:

Data Completeness: We made sure there were no missing pieces of data by using techniques to fill in any gaps.
Data Transformation: Some of our data was a bit tricky to work with, so we used a technique called Box-Cox transformation to make it easier to handle.
Variable Augmentation: We introduced two new variables, COUNTY_F and STATE_F, to help us look at data on a county and state level.

The main part of our analysis involved using something called linear regression. It's a fancy statistical method that helps us understand how different things, like obesity and physical inactivity, affect the rate of diabetes. We made sure to follow all the rules of statistics to make sure our findings were accurate and made sense.

## LINEAR REGRESSION:

Linear regression is a tool that helps us figure out how one thing is connected to another. In our case, we used it to answer questions like, "How does obesity affect the rate of diabetes?" and "Does being physically inactive make diabetes more common?"

The main idea behind linear regression is to draw the best straight line that shows how two or more things are related. For us, it was about understanding how the percentage of people with diabetes (% DIABETIC) is linked to factors like obesity (% OBESE), physical inactivity (% INACTIVE), and other stuff.

Linear regression gives us numbers that are easy to understand. For instance, it can tell us how much the diabetes rate goes up if the obesity rate increases by one percent. It helps us see clear patterns and connections in the data.

By using linear regression, we aimed to find meaningful insights in the data. It helped us identify which factors are most connected to diabetes rates and provided a basis for giving advice on public health.

## GEOSPATIAL ANALYSIS:

In addition to linear regression, we also looked at the data from a geographical perspective. We used special tools to see how health trends vary across different parts of the country. This helped us create maps that show which areas have higher or lower health indicators.

Our detailed method, which included linear regression and looking at the data on maps, makes sure our findings are reliable. It helps us understand the world of public health using careful data handling, advanced statistics, and exploring data on a map.

# ❖ APPENDIX B: RESULTS

- **DIABETIC CORRELATION HEAT MAP:**

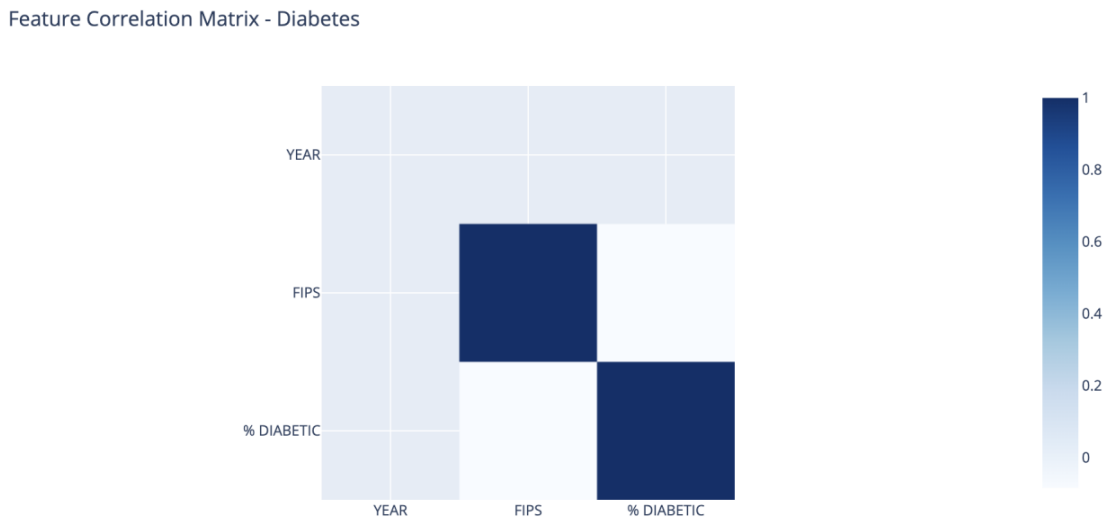



**FIG : 1.1**

The correlation heatmap provides insights into how health indicators are related. Darker colors indicate stronger correlations, while lighter colors suggest weaker or no correlations. In this heatmap, we can see that diabetic is only correlated with itself and it is the same with FIPS. We can see that based on this correlation heat map we did not find any correlation within a different variable. There is some correlation between Diabetic and Year along the side FIPS and Year but it seems very small which cannot be used in any future analysis.

- **OBESITY CORRELATION HEAT MAP:**

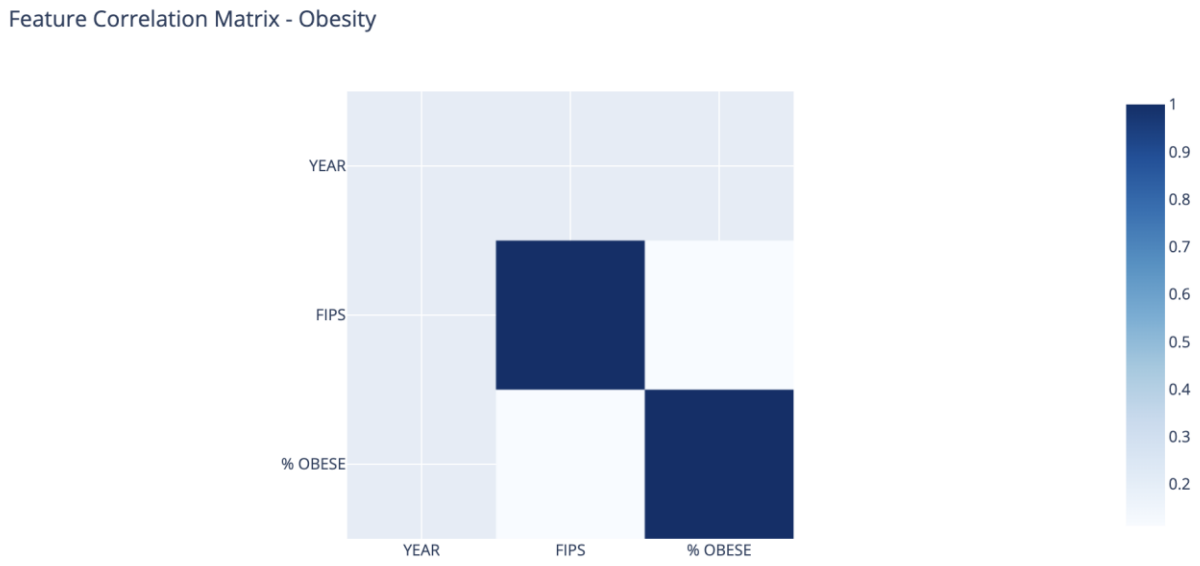



**FIG : 1.2**

In this heatmap also we can see that Obesity is only correlated with itself and it is the same with FIPS. We can see that based on this correlation heat map we did not find any correlation within a different variable. There is some correlation between Diabetic and Year along the side FIPS and Year but it seems very small which cannot be used in any future analysis.

- **INACTIVE CORRELATION HEAT MAP :**

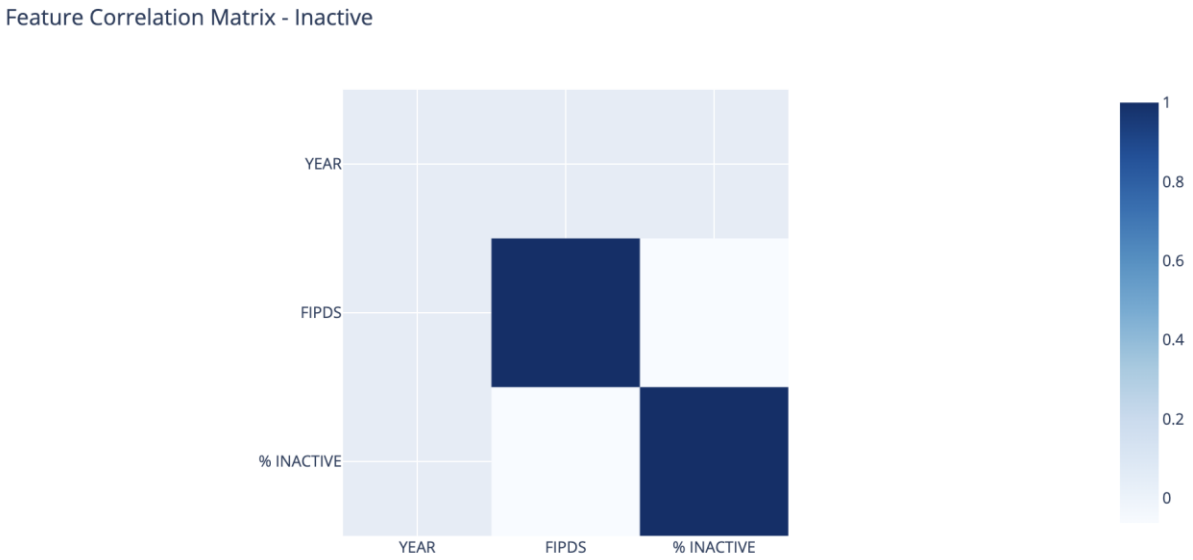



**FIG : 1.3**

In simpler terms, this correlation heatmap indicates that Inactive doesn't have strong relationships with other variables, and while there are some weak correlations between diabetes, FIPS codes, and the year, these correlations are not substantial enough to be meaningful for future analysis. It suggests that there may not be significant connections between these variables in your dataset.

- **PEAR PLOT OD DIABETIC FEATURE:**

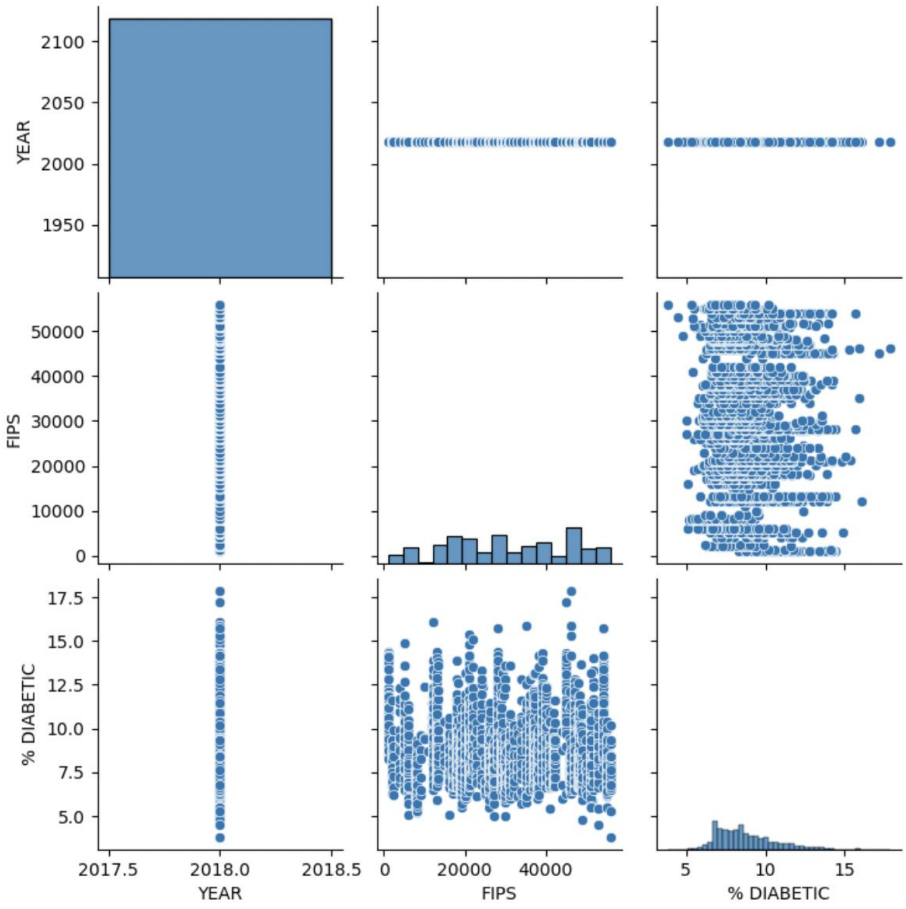



**FIG : 2.1**

I generate a pair heatmap for the "Diabetes" dataset (df1). The diagonal of the heatmap displays histograms for each numerical variable in the dataset, giving you an idea of their individual distributions. The lower triangle of the heatmap shows scatterplots for pairwise combinations of numerical variables. Each scatterplot represents the relationship between two variables, with points scattered according to their joint values. The upper triangle of the heatmap is a mirrored version of the lower triangle, which helps avoid redundancy in the visualization.

- **PEAR PLOT OD OBESITY FEATURE:**

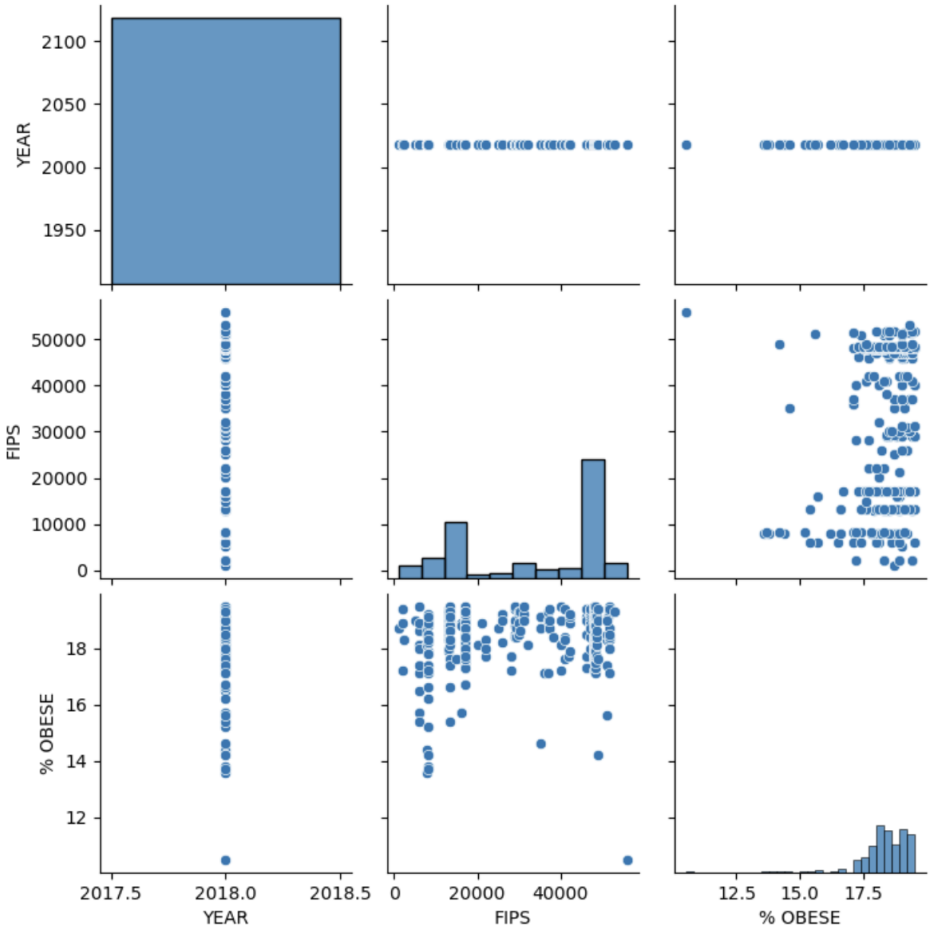



**FIG: 2.2**

I generate a pair heatmap for the "obese" dataset (df1). In this heatmap the diagonal of the heatmap displays histograms for each numerical variable in the dataset, giving you an idea of their individual distributions. The lower triangle of the heatmap shows scatterplots for pairwise combinations of numerical variables. Each scatterplot represents the relationship between two variables, with points scattered according to their joint values. The upper triangle of the heatmap is a mirrored version of the lower triangle, which helps avoid redundancy in the visualization.

- **PEAR PLOT OD INACTIVE FEATURE:**

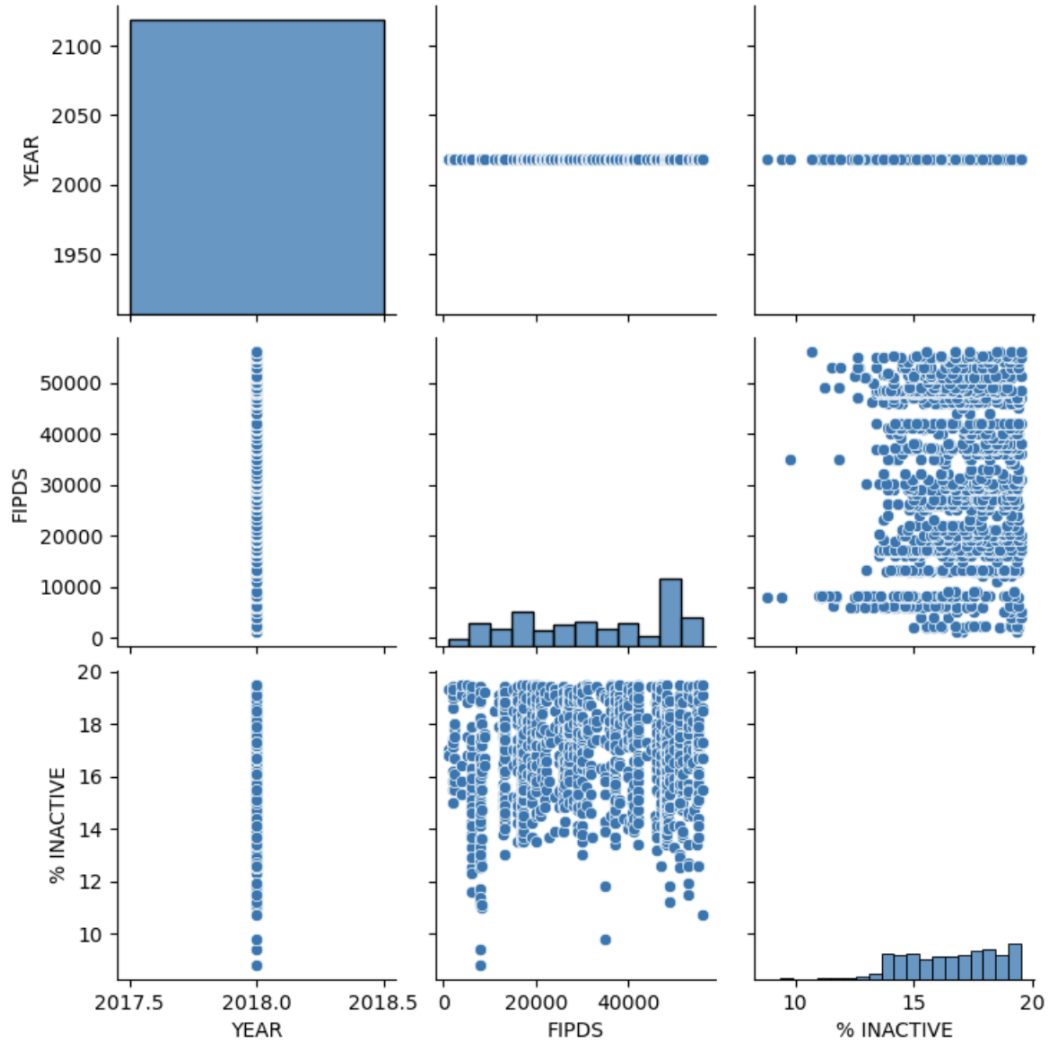



**FIG: 2.3**

I generate a pair heatmap for the "Inactive" dataset (df1). In this heatmap the diagonal of the heatmap displays histograms for each numerical variable in the dataset, giving you an idea of their individual distributions. The lower triangle of the heatmap shows scatterplots for pairwise combinations of numerical variables. Each scatterplot represents the relationship between two variables, with points scattered according to their joint values. The upper triangle of the heatmap is a mirrored version of the lower triangle, which helps avoid redundancy in the visualization

- **BOX – PLOT FOR DIABETIC** :

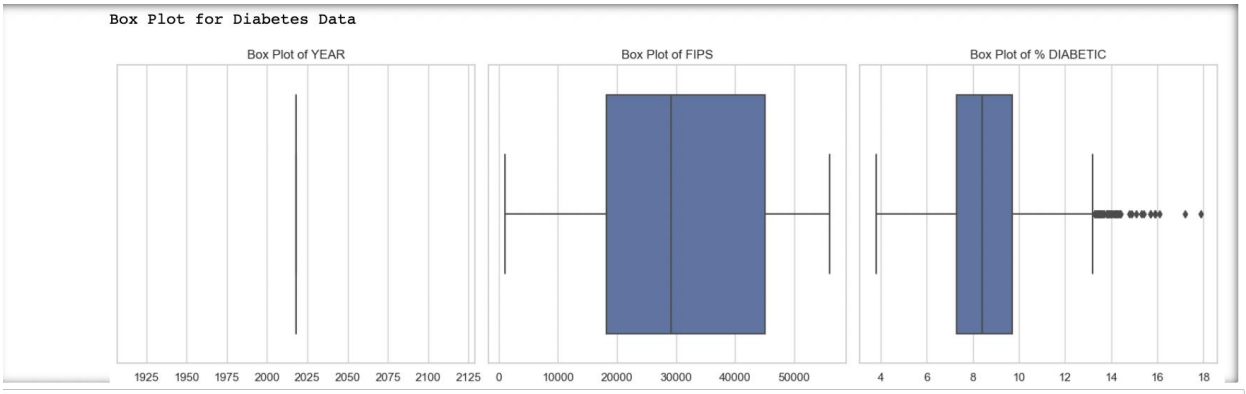



**FIG: 3.1**

% DIABETIC: This boxplot visualizes diabetes rate distribution, showcasing medians, interquartile ranges, and outliers. It helps pinpoint areas with notable diabetes concerns.

- **BOX – PLOT FOR OBESITY**:

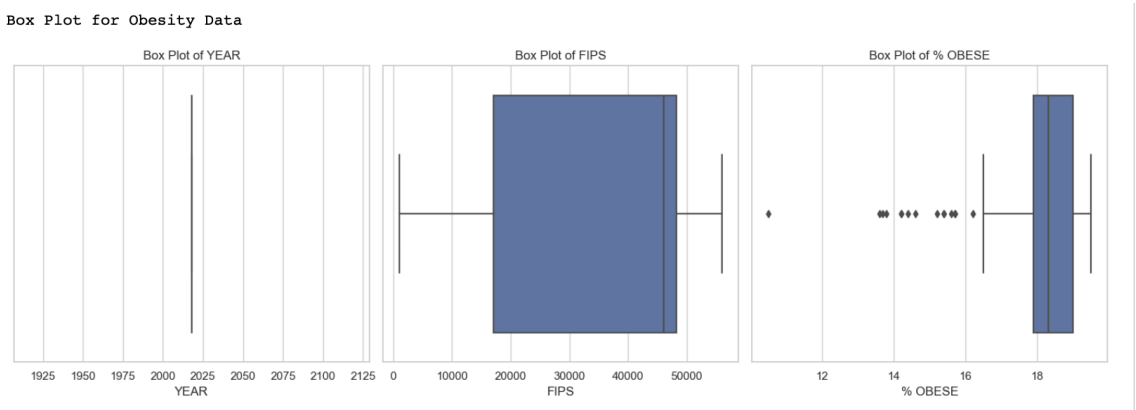



**FIG: 3.2**

% OBESITY: The boxplot displays the spread of obesity rates, highlighting median rates, interquartile ranges, and outliers. It aids in identifying regions with concerning or lower obesity prevalence.

- **BOX – PLOT FOR INACTIVE** :

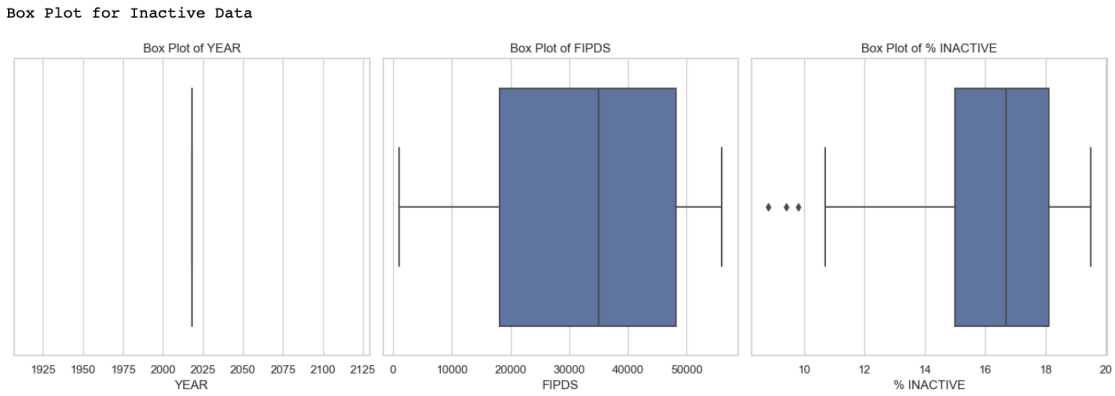



**FIG: 3.3**

% INACTIVE: The boxplot reveals inactivity rate variation, featuring medians, interquartile ranges, and outliers. It identifies regions with significant inactivity levels and successes in reducing them.

- **HISTOGRAM FOR DIABETIC :**

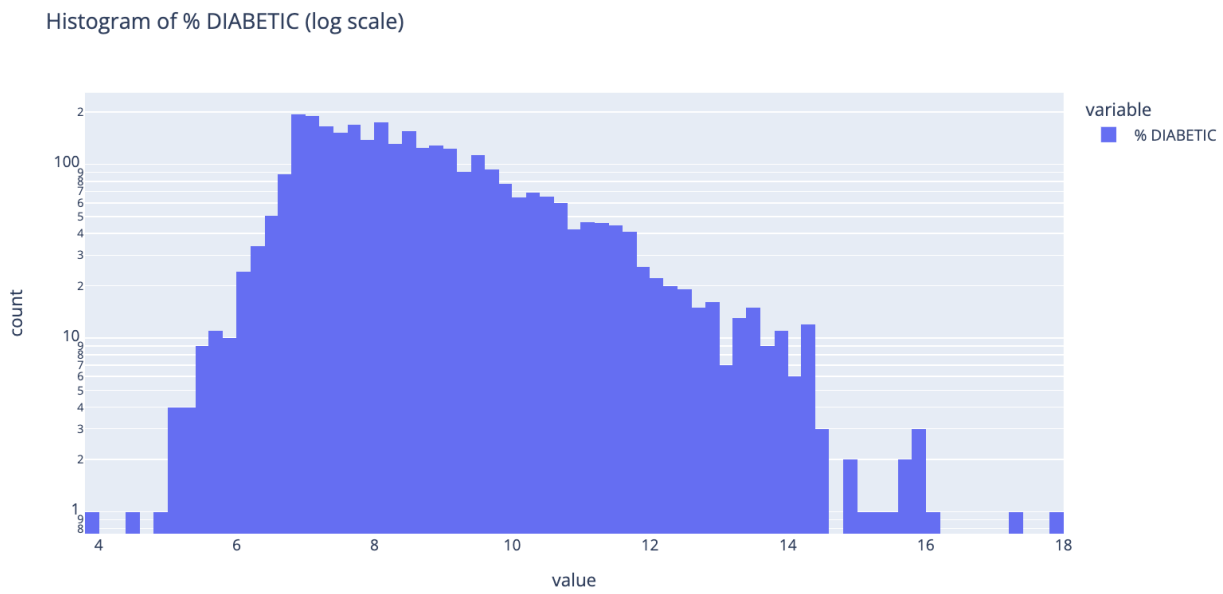



**FIG: 4.1**

Histogram of % DIABETIC (log scale): This histogram effectively visualizes the distribution of diabetes rates across counties and states. The logarithmic scale on the y-axis condenses the data, enabling us to discern the varying prevalence of diabetes across regions. Higher bars in the histogram signify a greater concentration of counties within specific diabetes rate ranges.

- **HISTOGRAM FOR OBESITY :**

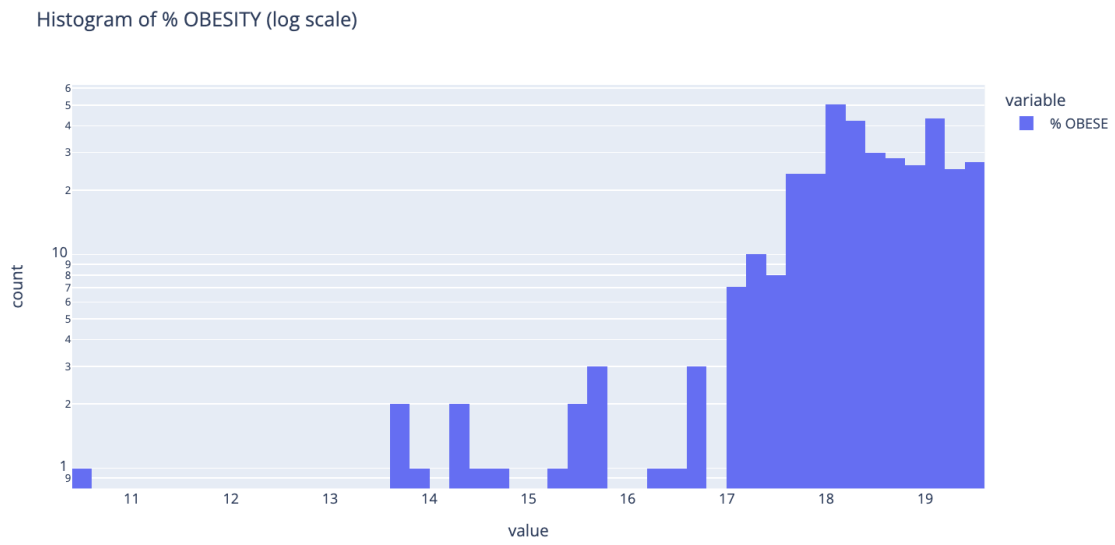



**FIG: 4.2**

Histogram of % OBESE (log scale): Similarly, this histogram illuminates the distribution of obesity rates among counties. The logarithmic scale on the y-axis aids in visualizing a wide range of values, offering insights into the prevalence of obesity across different regions. Tall bars indicate a significant number of counties within specific obesity rate ranges.

- **HISTOGRAM FOR INACTIVE:**

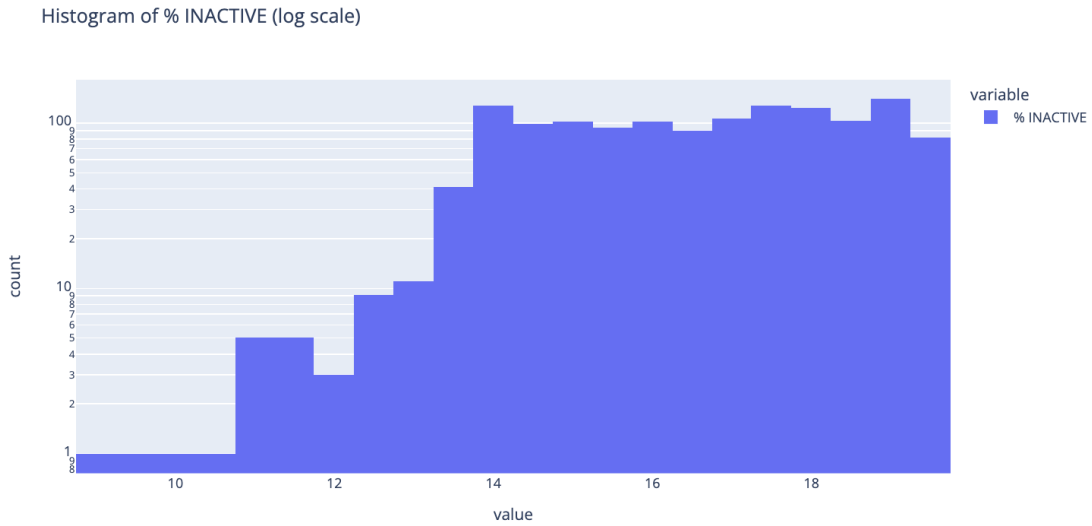



**FIG: 4.3**

Histogram of % INACTIVE (log scale): In this histogram, we explore the distribution of inactivity rates across counties and states. The logarithmic scale on the y-axis enhances visualization, allowing us to assess the prevalence of physical inactivity. Taller bars highlight a higher concentration of counties with specific inactivity rates.

- # EXAMINE THE VARIATIONS IN THE PREVALENCE OF DIABETES BY STATE:

| STATEW | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| Illinois | 102.0 | 7.552941 | 0.736153 | 6.3 | 7.000 | 7.35 | 8.000 | 9.6 |
| Massachusetts | 14.0 | 7.378571 | 0.789526 | 6.2 | 6.750 | 7.45 | 8.000 | 8.3 |
| Vermont | 14.0 | 7.378571 | 0.880715 | 6.2 | 6.700 | 7.40 | 7.850 | 9.3 |
| Minnesota | 87.0 | 7.811494 | 0.908029 | 5.0 | 7.150 | 7.80 | 8.400 | 9.9 |
| Wisconsin | 72.0 | 7.480556 | 0.969484 | 5.5 | 6.800 | 7.30 | 8.025 | 10.4 |

**FIG: 5.1**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Maryland | 24.0 | 10.012500 | 1.972708 | 6.8 | 8.675 | 9.90 | 11.550 | 13.3 |
| Kentucky | 120.0 | 9.450833 | 1.986128 | 6.4 | 7.900 | 9.00 | 10.800 | 15.4 |
| New Mexico | 33.0 | 8.712121 | 2.162718 | 5.8 | 7.300 | 8.00 | 9.600 | 15.9 |
| South Dakota | 66.0 | 8.381818 | 2.433036 | 6.3 | 7.025 | 7.70 | 8.400 | 17.9 |
| District of Columbia | 1.0 | 8.700000 | NaN | 8.7 | 8.700 | 8.70 | 8.700 | 8.7 |

**FIG: 5.2**

The mean "% DIABETIC" rate represents the average diabetes prevalence for each state. For instance, Alabama has the highest mean diabetes rate (11.21%), indicating a relatively high average prevalence of diabetes in the state. In contrast, states like Colorado have a lower mean diabetes rate (6.85%), suggesting a lower average prevalence of diabetes.

The standard deviation (std) measures the variability or spread of diabetes rates within each state. States with higher standard deviations exhibit greater variability in diabetes rates, indicating that the rates may vary significantly between regions or years within the state. Conversely, states with lower standard deviations have more consistent diabetes rates.

For example, while Alabama has a high mean diabetes rate, it also has a relatively high standard deviation (1.57), indicating variability in diabetes rates across different regions or years. In contrast, states like Vermont have a similar mean diabetes rate but a lower standard deviation (0.88), suggesting more consistency in diabetes rates.

These statistics provide valuable information for public health officials and researchers, helping them understand the prevalence and variability of diabetes across different states and aiding in the development of targeted interventions and healthcare policies to address this health concern.

**Now, we will display a line plot that focuses on the diabetes rates in the TOP TWO states with the HIGHEST prevalence. This visual representation will provide a clear overview of the diabetes trends in these states.**

**STATE : ILLINOIS**

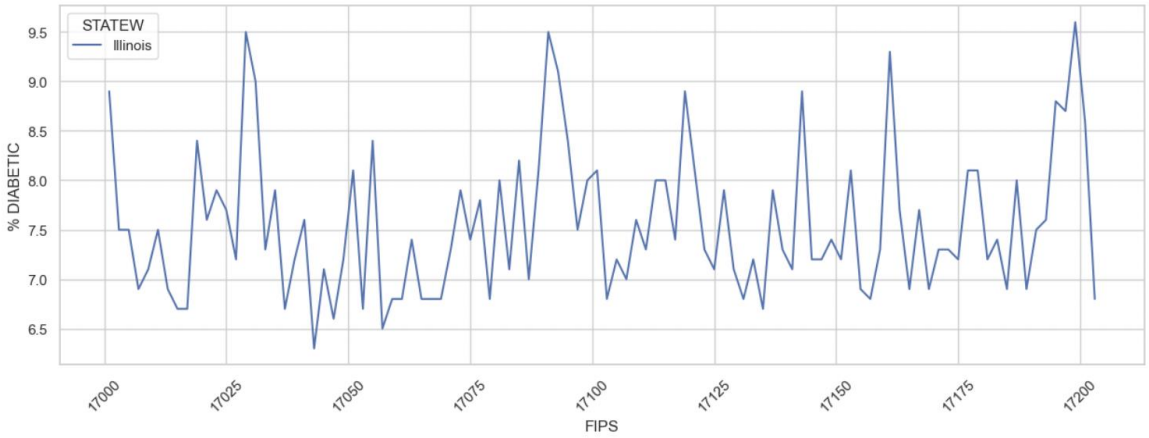



**FIG: 5.3**

The line graph depicts a conspicuous pattern of frequent fluctuations in %diabetic rates across various FIPS regions, resembling a series of peaks and valleys. This observed trend implies that diabetes prevalence undergoes regular oscillations, which could be attributed to cyclic elements or localized factors that exert influence on different counties.

**STATE : MASSACHUSETTS**

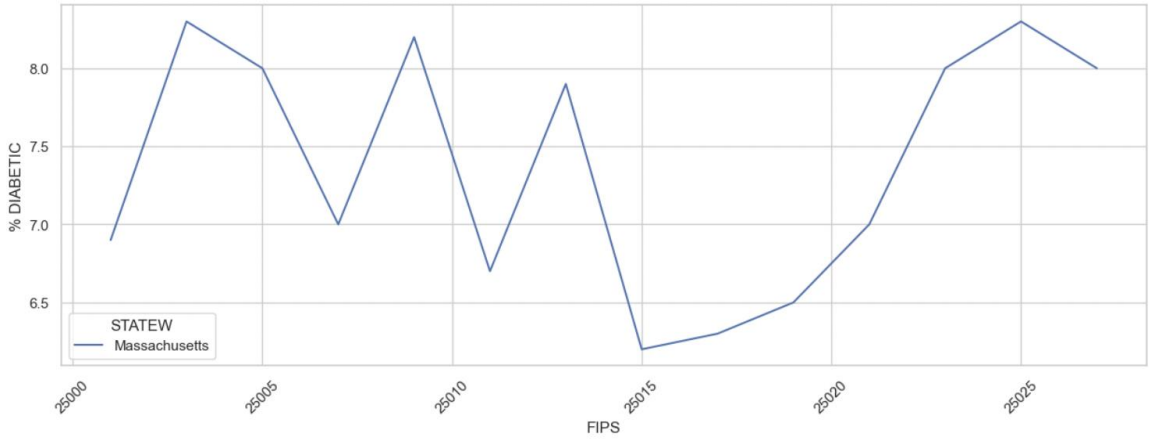



**FIG: 5.4**

The line chart illustrates a fluctuating trend in diabetes prevalence, with some counties showing a notably low rate initially, while others exhibit alarmingly high levels of diabetes cases. This observed pattern hints at the possibility that diabetes prevalence undergoes periodic fluctuations, which could be attributed to cyclic factors or region-specific factors impacting various counties.

**Now, we will display a line plot that focuses on the diabetes rates in the BOTTOM TWO states with the LOWEST prevalence. This visual representation will provide a clear overview of the diabetes trends in these states.**

**STATE : FLORIDA**

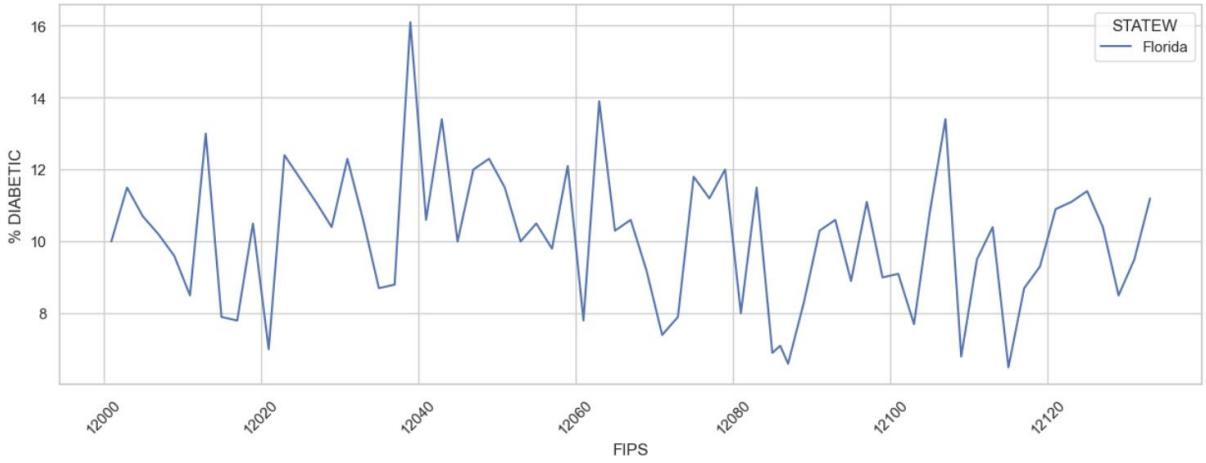



**FIG: 5.5**

The line chart of Florida State illustrates a prominent and recurring pattern of %diabetic rate fluctuations across numerous FIPS regions, forming a sequence of crests and troughs. This apparent trend suggests that diabetes prevalence experiences periodic oscillations, possibly influenced by cyclic components or localized factors that impact various counties.

**STATE: MARYLAND**

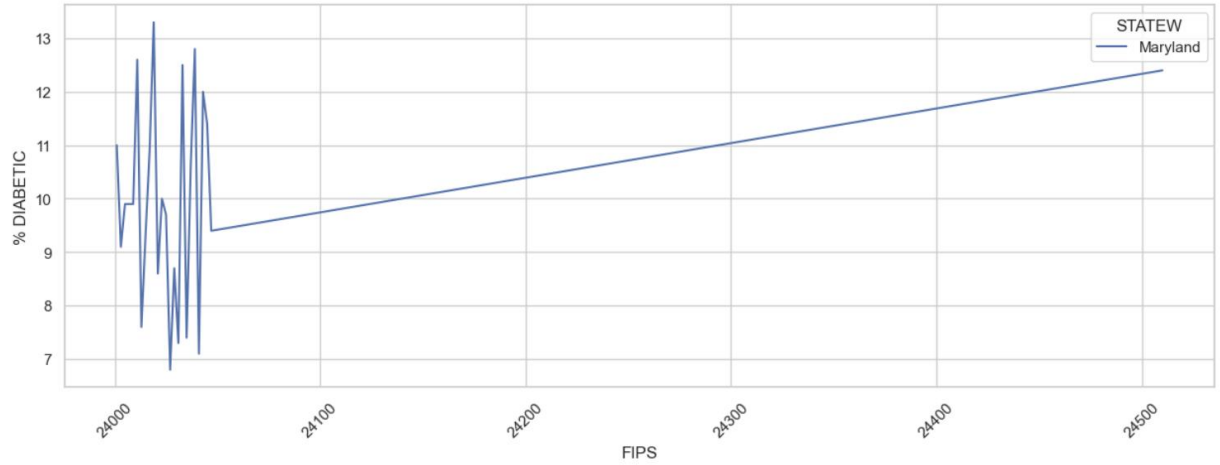



**FIG: 5.6**

The line plot of Maryland State displays a sudden sharp and steep rise in %diabetic rates across FIPS regions from the beginning, suggesting a rapid and significant increase in diabetes prevalence as we move across counties and then showed a gradual increment.

- # EXAMINE THE VARIATIONS IN THE PREVALENCE OF OBESITY BY STATE:

| STATE | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **Nebraska** | 3.0 | 19.233333 | 0.251661 | 19.0 | 19.100 | 19.20 | 19.350 | 19.5 |
| **Louisiana** | 3.0 | 18.000000 | 0.300000 | 17.7 | 17.850 | 18.00 | 18.150 | 18.3 |
| **Mississippi** | 2.0 | 17.450000 | 0.353553 | 17.2 | 17.325 | 17.45 | 17.575 | 17.7 |
| **Montana** | 6.0 | 18.966667 | 0.355903 | 18.5 | 18.675 | 19.05 | 19.275 | 19.3 |
| **Missouri** | 10.0 | 18.870000 | 0.452278 | 18.4 | 18.500 | 18.75 | 19.300 | 19.5 |

**FIG: 5.7**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Nevada** | 1.0 | 18.100000 | NaN | 18.1 | 18.100 | 18.10 | 18.100 | 18.1 |
| **New York** | 1.0 | 17.100000 | NaN | 17.1 | 17.100 | 17.10 | 17.100 | 17.1 |
| **North Dakota** | 1.0 | 18.400000 | NaN | 18.4 | 18.400 | 18.40 | 18.400 | 18.4 |
| **Washington** | 1.0 | 19.300000 | NaN | 19.3 | 19.300 | 19.30 | 19.300 | 19.3 |
| **Wyoming** | 1.0 | 10.500000 | NaN | 10.5 | 10.500 | 10.50 | 10.500 | 10.5 |

FIG: 5.8

The mean "% OBESITY" rate represents the average obesity prevalence for each state. For instance, Nebraska has the highest mean obesity rate (19.23%), indicating a relatively high average prevalence of obesity in the state. Conversely, states like California have a lower mean obesity rate (17.62%), suggesting a lower average prevalence of obesity.

The standard deviation (std) measures the variability or spread of obesity rates within each state. States with higher standard deviations exhibit greater variability in obesity rates, indicating that the rates may vary significantly between regions or years within the state. Conversely, states with lower standard deviations have more consistent obesity rates.

For example, while Nebraska has a high mean obesity rate, it also has a relatively low standard deviation (0.25), indicating consistency in obesity rates across different regions or years. In contrast, states like Colorado have a similar mean obesity rate but a higher standard deviation (1.82), suggesting more variability in obesity rates.

These statistics provide valuable information for public health officials and researchers, helping them understand the prevalence and variability of obesity across different states and aiding in the development of targeted interventions and healthcare policies to address this health concern.

**We will now showcase a line plot concentrating on Obesity rates within the TWO STATES that exhibit the HIGHEST prevalence. This graphical depiction aims to offer a concise overview of the Obesity trends occurring within these states during a specified time period.**

**STATE:NEBRASKA**

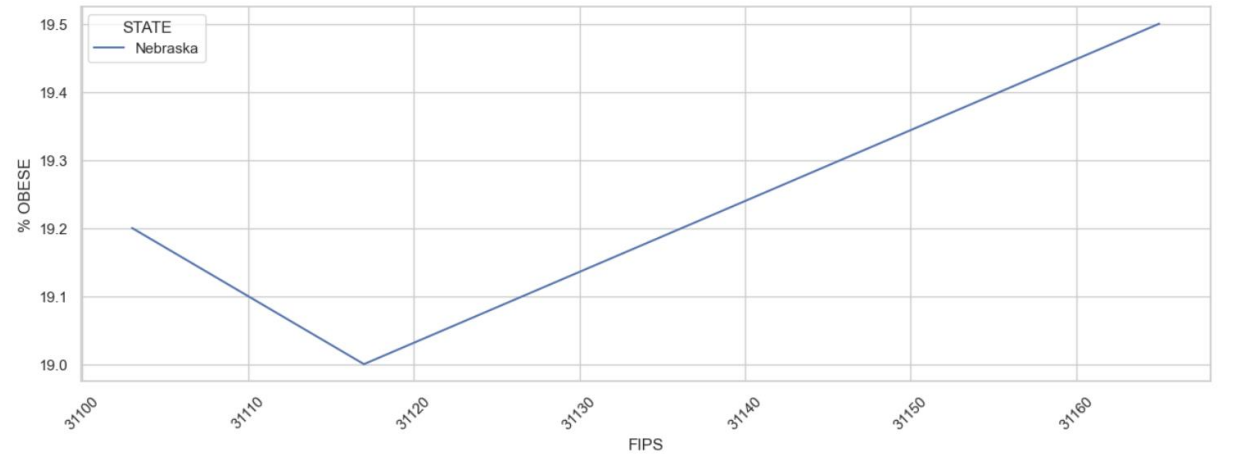



FIG: 5.9

The line chart exhibits an initial steep downward slope for the first few data points, followed by a subsequent steep upward slope. This pattern implies an abrupt decrease in a certain metric or variable, succeeded by a rapid and significant increase. Such a trend may signify a notable shift or change in the observed phenomenon over time or across data points.

**STATE: LOUISIANA**

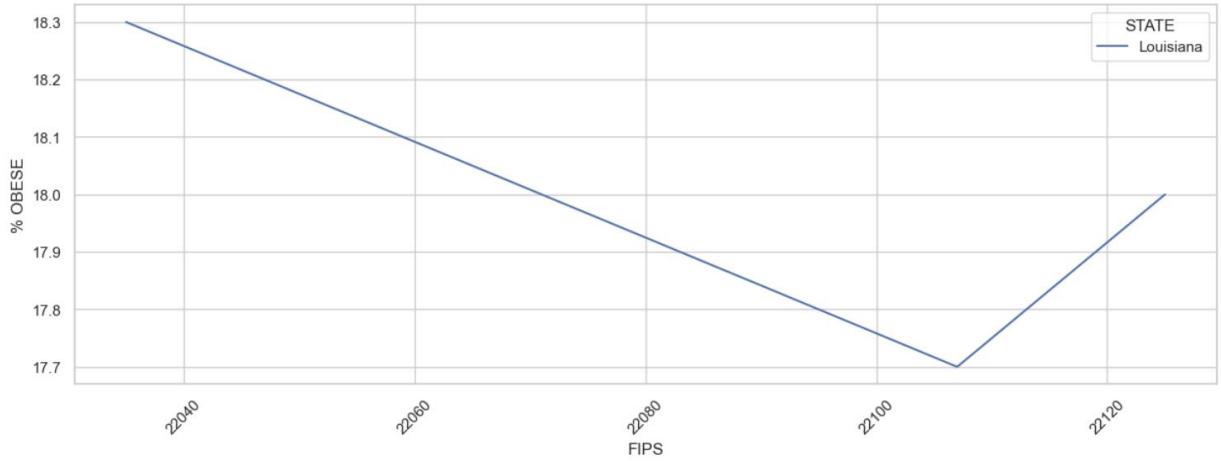



FIG: 5.10

In this plot, the initial downward slope followed by an upward slope signifies a declining trend in obesity rates in certain counties, succeeded by a subsequent increase in obesity rates in different counties. This pattern suggests variations in obesity prevalence among counties within the state, with some experiencing a decrease while others observe an increase over time.

**We will now showcase a line plot concentrating on Obesity rates within the TWO STATES that exhibit the LOWEST prevalence. This graphical depiction aims to offer a concise overview of the Obesity trends occurring within these states during a specified time period.**

We are unable to present line plots for bottom states in this analysis due to a standard deviation (std) of zero in their obesity rates. These zero standard deviations indicates that the obesity rates for these states remain unchanged and exhibit no variability throughout the specified time frame. Consequently, there is no meaningful variation to visualize in the form of line plots for these states. However, it is noteworthy that these states consistently maintain the lowest obesity rates throughout the analyzed period, underscoring the stability and consistency of their low prevalence. While these states may not reveal trends in obesity rates, their consistently favorable health outcomes are an important aspect of this study.

- # EXAMINE THE VARIATIONS IN THE PREVALENCE OF INACTIVE BY STATE:

| STATE | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| Florida | 3.0 | 18.600000 | 0.624500 | 17.9 | 18.350 | 18.80 | 18.950 | 19.1 |
| Rhode Island | 3.0 | 17.433333 | 0.709460 | 16.8 | 17.050 | 17.30 | 17.750 | 18.2 |
| New Hampshire | 5.0 | 18.240000 | 0.743640 | 17.4 | 17.900 | 18.10 | 18.400 | 19.4 |
| Connecticut | 5.0 | 17.480000 | 1.089495 | 16.4 | 16.700 | 17.50 | 17.600 | 19.2 |
| Indiana | 11.0 | 17.627273 | 1.094615 | 16.1 | 16.950 | 17.80 | 18.100 | 19.5 |

FIG: 5.11

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Washington | 31.0 | 16.300000 | 2.180978 | 11.5 | 15.350 | 16.90 | 17.750 | 19.4 |
| Colorado | 60.0 | 14.876667 | 2.466478 | 8.8 | 13.100 | 14.95 | 16.625 | 19.4 |
| Wyoming | 9.0 | 16.955556 | 2.684730 | 10.7 | 16.700 | 17.30 | 18.600 | 19.5 |
| New Mexico | 15.0 | 15.826667 | 2.710948 | 9.8 | 14.250 | 16.00 | 17.800 | 19.1 |
| District of Columbia | 1.0 | 18.500000 | NaN | 18.5 | 18.500 | 18.50 | 18.500 | 18.5 |

**FIG: 5.12**

From the table, we can observe that Florida has the highest mean "% INACTIVITY" rate (18.60%), indicating that, on average, a significant portion of the population in Florida is inactive. In contrast, Colorado has the lowest mean "% INACTIVITY" rate (14.88%), suggesting a relatively more active population on average.

However, it's essential to consider the standard deviation alongside the mean. For example, even though Florida has a high mean inactivity rate, its relatively low standard deviation (0.62) suggests that this rate is relatively consistent across different regions or years within the state. On the other

hand, states like Colorado, with a low mean but higher standard deviation (2.47), may have more significant variability in inactivity rates.

The table provides valuable insights into the distribution and central tendency of inactivity rates across different states, helping policymakers and researchers understand the variations in physical inactivity levels.

**We will present a line chart highlighting obesity rates in the top TWO STATES with the HIGHEST prevalence. This visual representation aims to provide a succinct summary of obesity trends in these states over a specified timeframe.**

**STATE: FLORIDA**

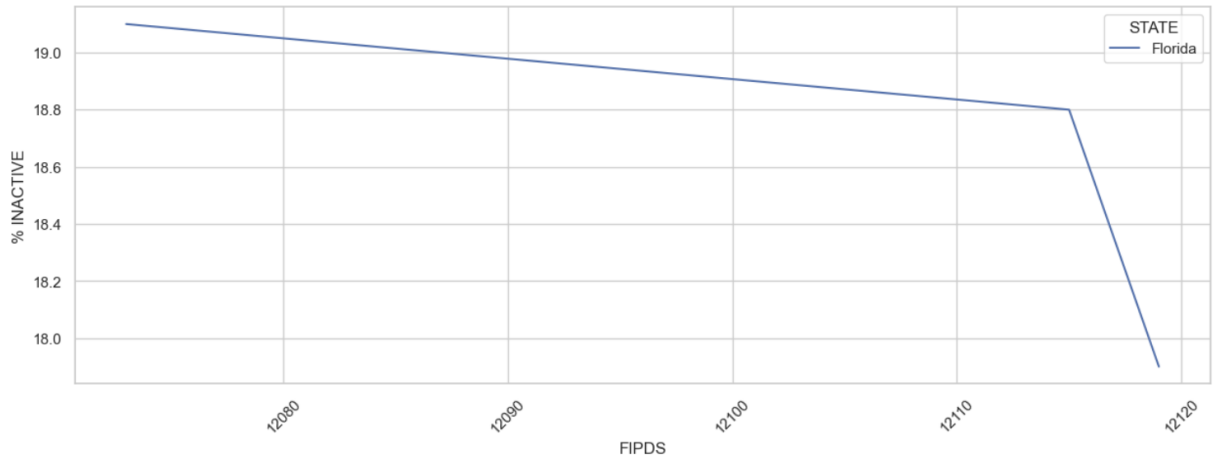



**FIG: 5.13**

The line plot for Florida's "% INACTIVE" rates exhibits an interesting trend. As we move along the x-axis, which represents different counties in Florida (denoted by FIPS codes), there is a gradual decrease in inactivity rates. This initial gradual decrease suggests a general trend of decreasing physical inactivity across most counties in Florida. However, at a certain point along the x-axis, there is a sudden and steep downward slope. This indicates that some counties in Florida experience a significant drop in inactivity rates, potentially showcasing successful initiatives or programs promoting physical activity.

**STATE: RHODE ISLAND**

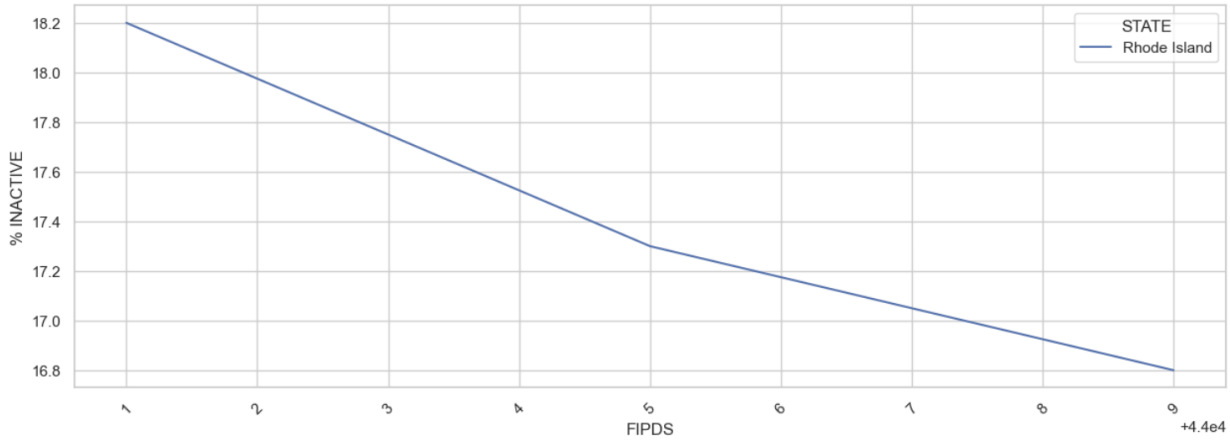



**FIG: 5.14**

The line plot for Rhode Island's "% INACTIVE" rates shows a continuous downward slope across the x-axis. The x-axis represents different counties or regions within Rhode Island. This consistent downward trend indicates that, as we move from one county to another, there is a continuous reduction in physical inactivity rates. It suggests that Rhode Island as a whole experiences a decline in physical inactivity across its various regions.

- **CREATED VISUALIZATION TO COMPARE THE DISTRIBUTION OF OUR FEATURE BOTH BEFORE AND AFTER APPLYING THE BOX-COX TRANSFORMATION:**

## 1. DIABETES:

**BEFORE:**

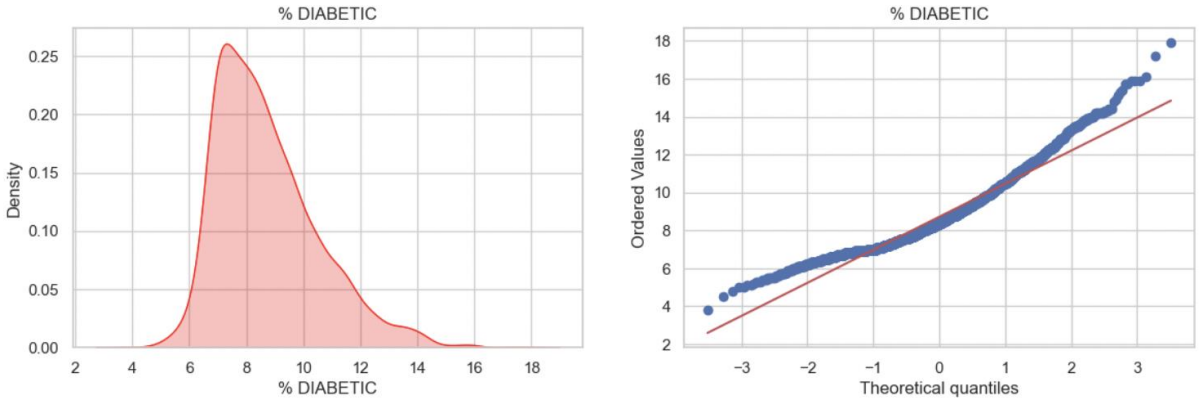



**FIG: 6.1**

The KDE (Kernel Density Estimation) plot for diabetic data exhibits right skewness. This indicates that the distribution of diabetic rates is skewed towards the higher end of the scale. In a right-skewed distribution, most data points are concentrated on the left side of the peak, with a long tail extending to the right. This suggests that there are fewer instances of lower diabetic rates, while a larger portion of the data falls into the higher range. The right skewness in the diabetic data implies that, on average, diabetic rates tend to be higher, but there are still outliers or extreme values with lower diabetic rates.

The Q-Q plot for diabetes data reveals a distinct pattern where data points initially deviate from a straight line at the plot's edges but align closely with the line in the middle. This indicates that the central portion of the data has been transformed to approximate a normal distribution. However, deviations at the plot's ends suggest that extreme values or outliers in the dataset's tails remain, despite the transformation's effectiveness in normalizing the central data range.

**AFTER :**

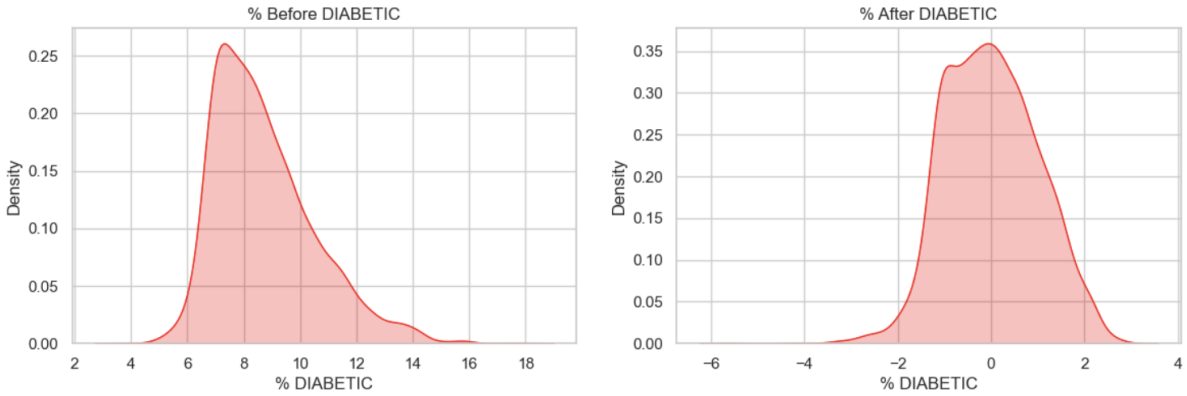



**FIG: 6.2**

After applying the Box-Cox transformation to the diabetic data, the KDE plot shows a more symmetrical and bell-shaped distribution. This transformation helps to normalize the data, making it closer to a normal distribution. The skewness is reduced, and the data points are more evenly distributed around the peak, indicating a better fit to a normal distribution. This transformation is often used to stabilize variances and improve the normality of data, which can be beneficial for statistical analysis and modeling.

## 2. OBESITY:

**BEFORE:**

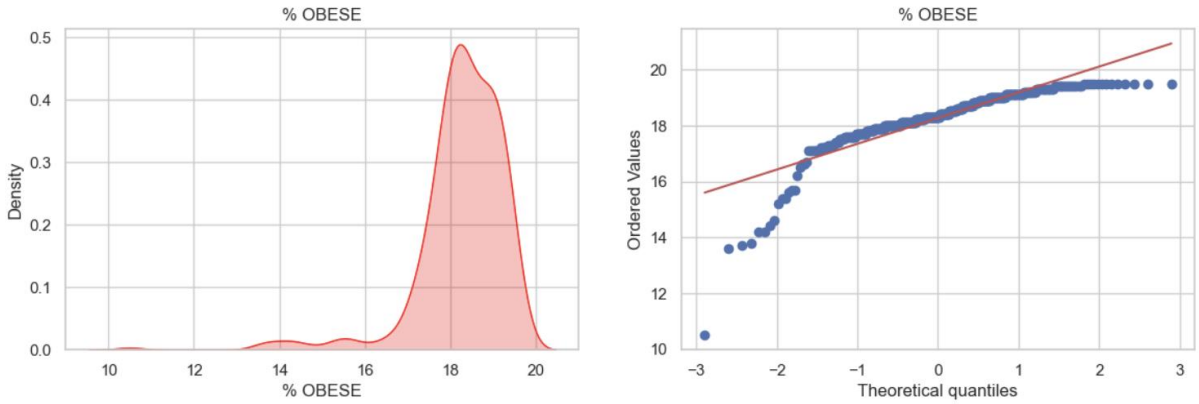



**FIG: 6.3**

The KDE (Kernel Density Estimation) plot for obesity data exhibits left skewness. This means that the distribution of obesity rates is skewed towards the lower end of the scale. In a left-skewed distribution, the majority of data points are concentrated on the right side of the peak, with a long tail extending to the left. This indicates that there are fewer instances of higher obesity rates, while a larger portion of the data falls into the lower range. This skewness suggests that, on average, obesity rates tend to be lower, but there are still outliers or extreme values with higher obesity rates.

In the case of the obesity data, the Q-Q plot displays a noticeable trend where, at the outer edges of the plot, data points diverge from a straight line, while in the middle, they closely follow this line. This observation implies that a transformation has been applied to make the central part of the data resemble a normal distribution. Nevertheless, discrepancies at the plot's extremes indicate that there are still outliers or extremely high/low values in the tails of the dataset, even though the transformation has successfully normalized the central range of data.

**AFTER:**

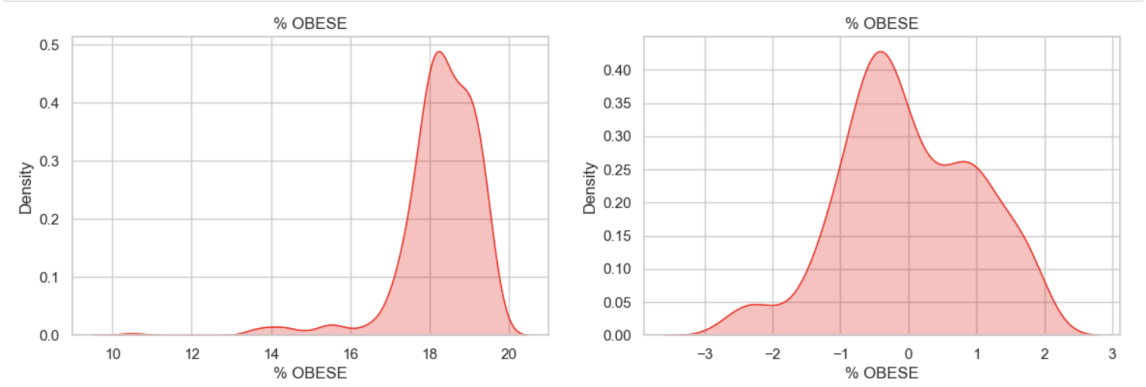



**FIG: 6.4**

After applying the Box-Cox transformation, the shape of the distribution changed significantly. It evolved into a more balanced and symmetric bell-shaped curve. This transformation helped make the data distribution closer to a normal distribution, where the majority of data points are concentrated around the center, and fewer extreme values exist on either side. In simpler terms, it made the obesity rate data behave in a way that's easier to analyze statistically. This kind of transformation is commonly used to make data more suitable for various statistical techniques and modeling.

## 3. INACTIVITY:

**BEFORE:**

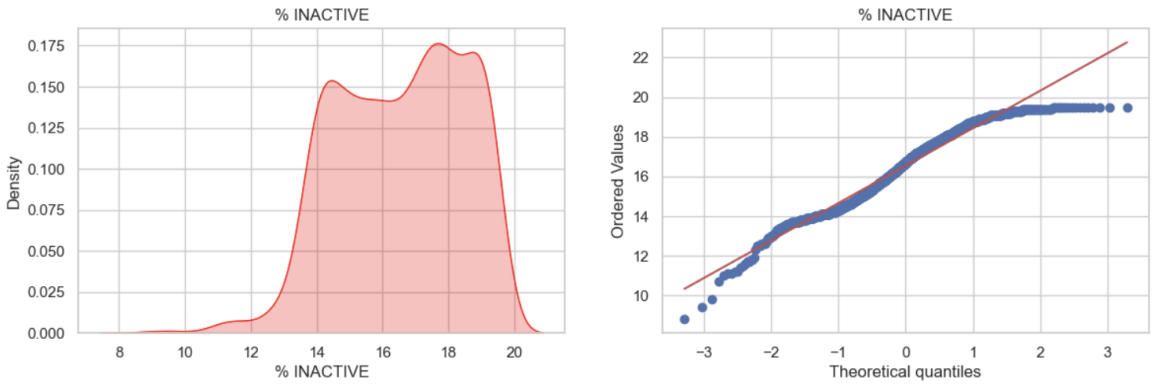



**FIG: 6.5**

The KDE (Kernel Density Estimation) plot for inactive data exhibits left skewness. In a left-skewed distribution, the majority of data points are concentrated on the right side of the peak, with a long tail extending to the left. This indicates that there are fewer instances of higher inactivity rates, while a larger portion of the data falls into the lower range. The left skewness suggests that, on average, inactivity rates tend to be lower, but there are still outliers or extreme values with higher inactivity rates.

## AFTER:

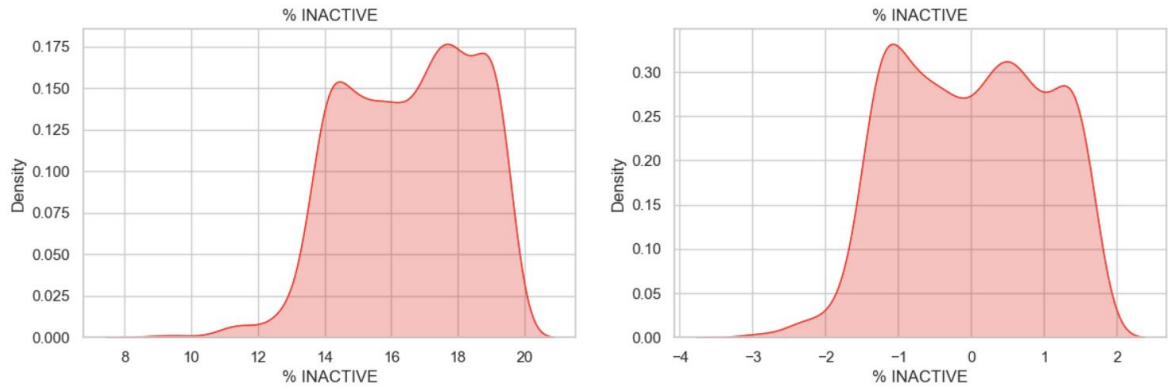



**FIG: 6.6**

Following the Box-Cox transformation, the distribution of the inactivity rates changed notably. It transformed into a more symmetric, bell-shaped curve, resembling a normal distribution. This transformation was effective in reducing the left-skewness and making the data distribution more suitable for statistical analysis. In simpler terms, it made the inactivity rate data behave in a manner that is easier to work with statistically and allowed for more reliable modeling and analysis. The Box-Cox transformation is a valuable tool in data preprocessing to improve the normality of distributions and enhance the effectiveness of statistical methods.

- **GRAPHICAL REPRESENTATION OF DIABETES INCIDENCE ACROSS THE UNITED STATES THROUGH MAP:**

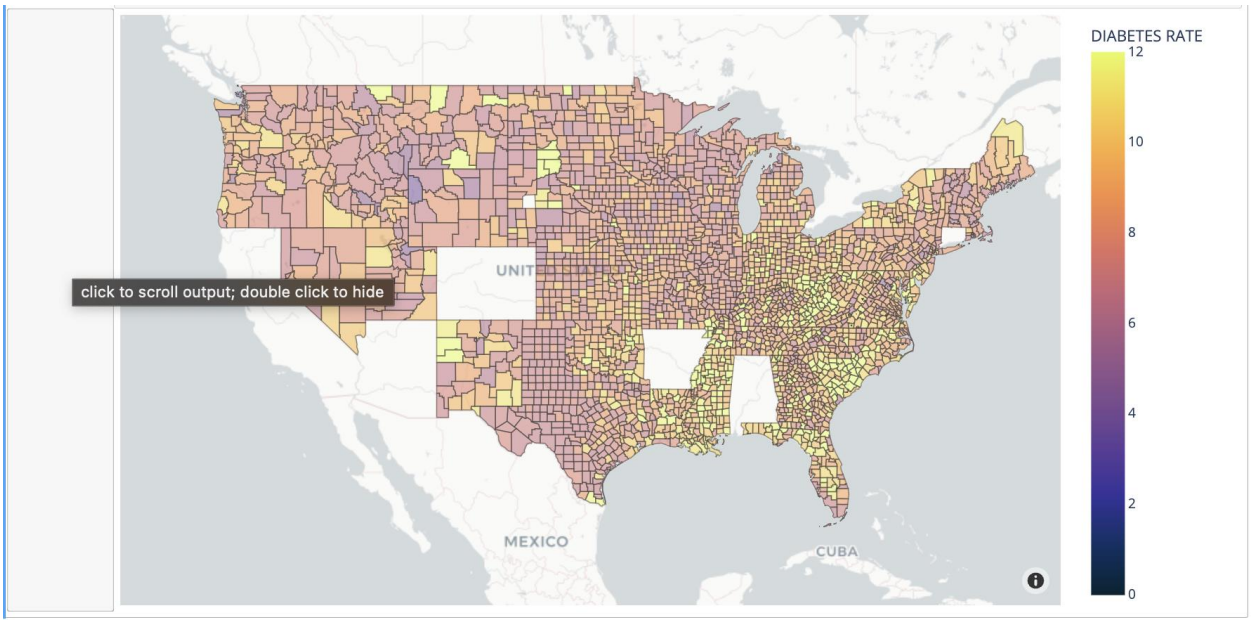



**FIG: 7.1**

In the geospatial heatmap, darker shades represent lower diabetes rates, while lighter shades indicate higher rates. This map visually illustrates the significant disparities in diabetes prevalence across different regions of the United States. Darker regions typically have lower diabetes rates and may be associated with factors such as healthier lifestyles, better healthcare access, and lower poverty levels. In contrast, lighter regions indicate higher diabetes rates, often linked to factors like obesity, limited healthcare access, and higher poverty rates. This heatmap provides valuable insights for public health officials, policymakers, and researchers to target interventions and resources effectively in areas with higher diabetes prevalence.

These variations can be attributed to a complex interplay of factors, including socio-economic conditions, access to healthcare, and lifestyle choices. For instance, states with darker shades may have higher rates of obesity, limited access to healthcare facilities, or higher poverty levels, all of which contribute to increased diabetes prevalence.

Conversely, regions with lighter shades may have better healthcare infrastructure, healthier lifestyle choices, and lower poverty rates, resulting in reduced diabetes prevalence. This heatmap serves as a valuable tool for public health officials, policymakers, and researchers to identify areas in need of targeted interventions and resources to address the diabetes epidemic effectively.

The geospatial heatmap created using Plotly for diabetes prevalence in the United States reveals notable regional disparities. States like Louisiana, Alabama, and Mississippi have significantly higher diabetes rates, as indicated by their lighter colors on the map. Conversely, states such as Colorado, Vermont, and Massachusetts exhibit lower prevalence with darker shades.

- **GRAPHICAL REPRESENTATION OF OBESITY INCIDENCE ACROSS THE UNITED STATES THROUGH MAP:**

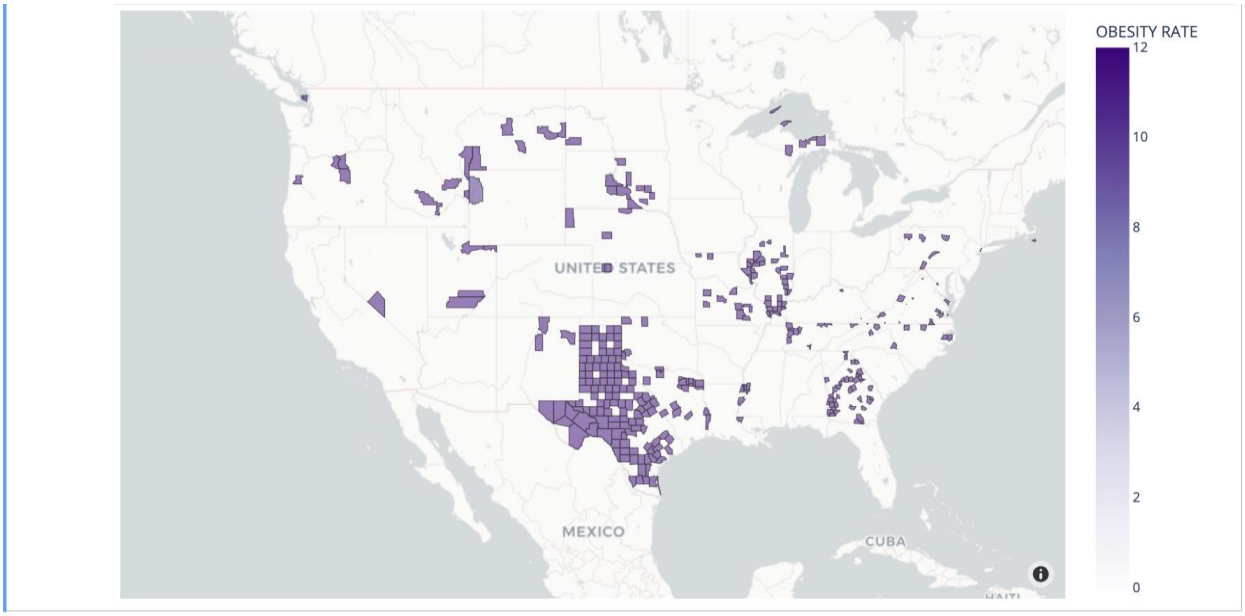



**FIG: 7.2**

The geospatial heatmap visually portrays the prevalence of obesity across various regions of the United States. In this representation, darker shades on the map signify higher obesity rates, while lighter shades indicate lower rates. The map reveals substantial variations in obesity prevalence across the nation, with darker regions signifying areas with elevated obesity rates and lighter regions representing areas with relatively lower rates.

These disparities can be attributed to a multitude of factors, including socio-economic conditions, healthcare accessibility, and lifestyle choices. For example, regions with darker shades may experience higher obesity rates due to factors such as limited access to healthy food options, sedentary lifestyles, and economic challenges. Conversely, areas with lighter shades may have lower obesity rates, often linked to better access to nutritious food, active lifestyles, and improved socio-economic conditions.

This geospatial heatmap serves as a valuable resource for public health professionals, policymakers, and researchers. It aids in identifying regions that require targeted interventions and resources to combat the obesity epidemic effectively, promoting healthier lives and reducing the burden on healthcare systems.

The regions with darker shades, such as Louisiana and Nebraska, may experience higher obesity rates due to factors such as limited access to healthy food options, sedentary lifestyles, and economic challenges. Conversely, areas with lighter shades, such as Hawaii and New York, may have lower obesity rates, often linked to better access to nutritious food, active lifestyles, and improved socio-economic conditions.

- **GRAPHICAL REPRESENTATION OF INACTIVITY INCIDENCE ACROSS THE UNITED STATES THROUGH MAP:**

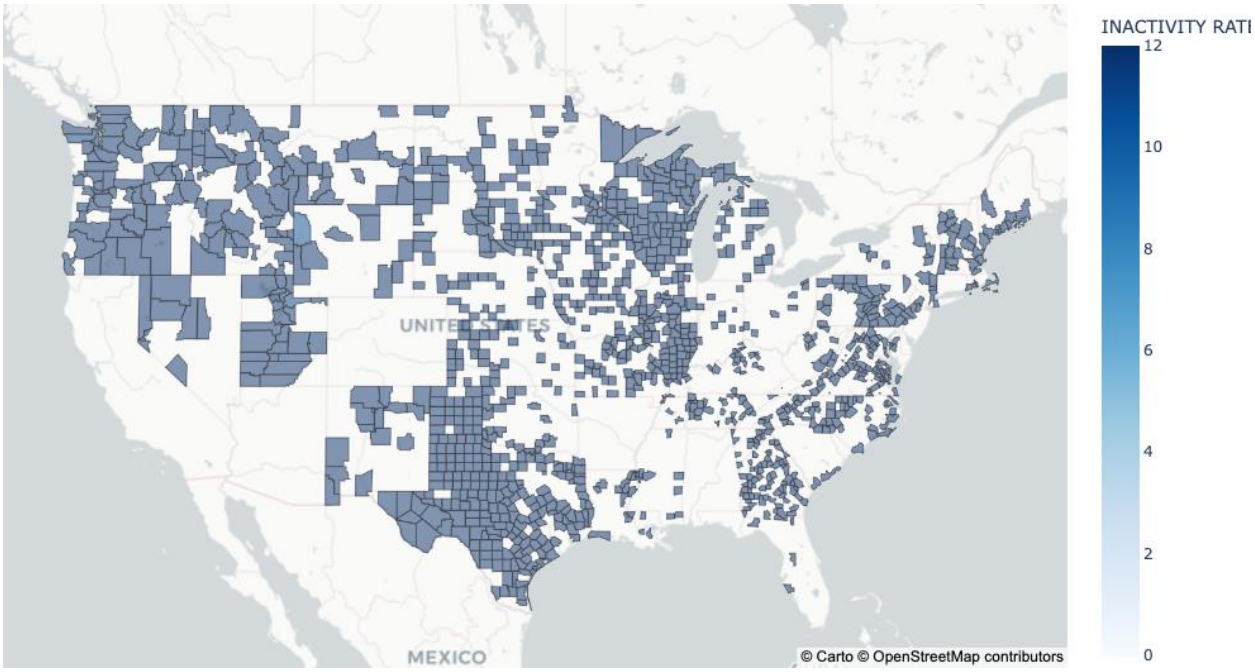



**FIG: 7.3**

The geospatial heatmap provides a clear and informative visualization of physical inactivity prevalence across various regions of the United States. Darker shades on the map signify areas with higher rates of physical inactivity, while lighter shades indicate regions with lower levels of inactivity. This visualization underscores the significant disparities that exist across the country, shedding light on the urgent need for targeted interventions to address this public health concern.

The disparities in physical inactivity rates can be attributed to a multitude of factors. Regions with darker shades, such as Arkansas, Louisiana, and Mississippi, may face challenges such as limited access to recreational facilities, socioeconomic inequalities, and environments that discourage physical activity. In contrast, areas with lighter shades, including Colorado, New Mexico, and Maryland, may benefit from better infrastructure that promotes active lifestyles, greater awareness of the importance of physical activity, and improved access to fitness resources.

This geospatial heatmap serves as a crucial tool for public health officials, policymakers, and researchers. By pinpointing regions with high levels of physical inactivity, such as the southern states, it allows for the development of tailored interventions and resources to encourage physical activity and combat the negative health effects associated with sedentary behavior. Ultimately, this visualization supports efforts to improve the overall health and well-being of communities throughout the United States.

- # DATA MODELING

  - ## DATA PREPARATION AND FEATURE EXTRACTION:

The analysis involves three datasets, each representing different health-related factors: diabetes prevalence (% Diabetic), obesity rates (% Obese), and physical inactivity rates (% Inactive) across various regions in the United States. We began by extracting unique FIPS codes from each dataset, serving as unique identifiers for different regions or counties.

  - ## MERGING DATASETS:

After obtaining FIPS codes, all three datasets were merged using these codes to create a comprehensive dataset called 'features. Data cleaning steps were implemented, including dropping unnecessary columns and ensuring column name consistency.

  - ## HANDLING MISSING DATA:

We identified missing values in the 'features' dataset and adopted a random imputation strategy. This involved generating random values from normal distributions with means and standard deviations matching the respective column to fill in missing data.

  - ## LINEAR REGRESSION MODEL:

We selected features (independent variables) for the linear regression analysis, including 'YEAR,' 'FIPS,' '% OBESE,' '% INACTIVE,' 'COUNTY_F,' 'STATE_F,' 'STATE_CODE,' and 'COUNTY_CODE.' The target variable (dependent variable) was '% DIABETIC,' representing diabetes prevalence. The dataset was split into a training set (70%) and a test set (30%) to facilitate model training and evaluation.

  - ## TRAINING THE LINEAR REGRESSION MODEL:

A linear regression model was created using Scikit-learn's Linear Regression class. The model was trained on the training data to learn the relationship between the independent variables (features) and the target variable ('% DIABETIC').

  - ## MODEL COEFFICIENTS AND INTERCEPTS:

After model training, we examined the coefficients associated with each feature. These coefficients indicate the impact of a one-unit change in the corresponding feature on the target variable, assuming other features are constant. Additionally, we determined the model's intercept, representing the predicted value of the target variable when all features are set to zero.

  - ## MODEL EVALUATION:

To assess model performance, we used it to make predictions on the test set. A scatterplot was created to visualize the relationship between the true values of '% DIABETIC' and the model's predicted values. The black regression line in the plot illustrates this linear relationship.

- **METRICS:**

Two key metrics were computed to gauge model accuracy. Mean Absolute Error (MAE): This metric measures the average absolute difference between the true and predicted values, quantifying prediction accuracy.

- **MEAN SQUARED ERROR(MSE) AND ROOT MEAN SQUARED ERROR(RMSE):** These metrics provide insight into the overall model accuracy by quantifying the average squared error and its square root.

- **INTERPRETING THE RESULTS:**

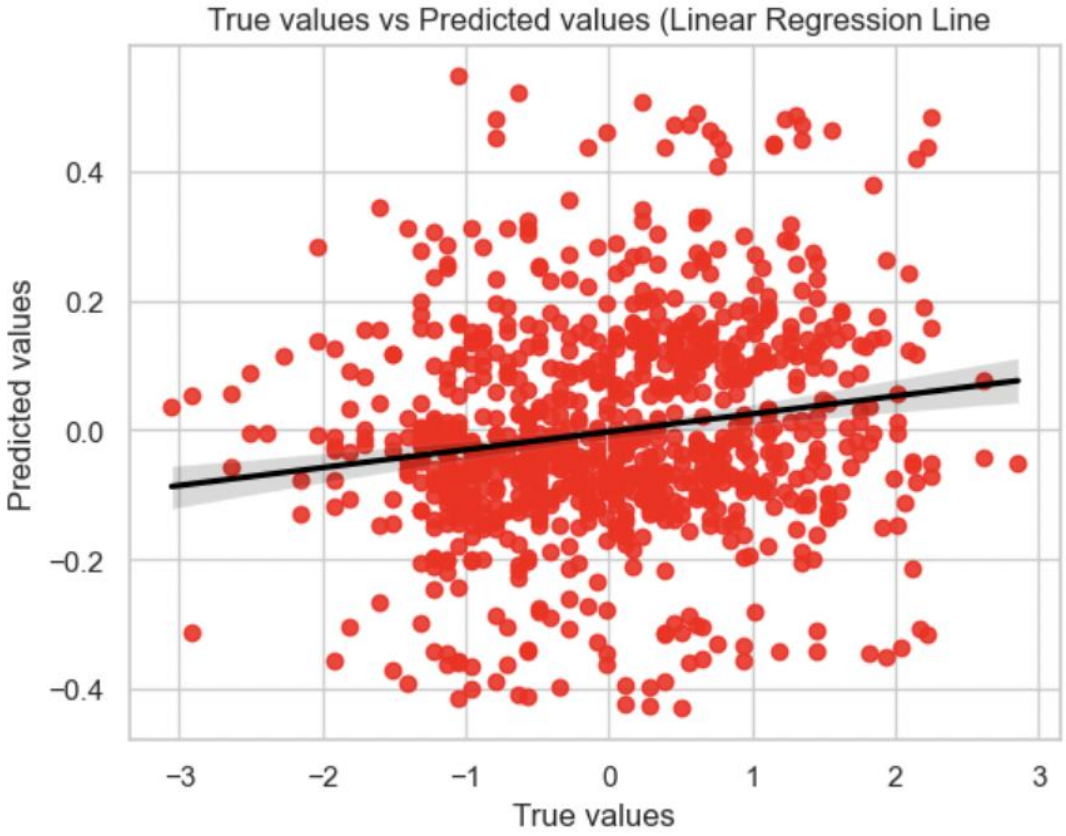



**FIG: 8.1**

Certainly, it's important to acknowledge that the linear regression model used in this analysis does not provide a particularly strong predictive power for diabetes prevalence based on the selected features. Despite examining various variables such as obesity rates ('% OBESE') and physical inactivity rates ('% INACTIVE'), the model struggles to capture meaningful correlations with diabetes prevalence ('% DIABETIC'). This lack of strong correlation between the variables suggests that additional factors beyond those considered in this model significantly influence diabetes rates. Consequently, the model's limited ability to predict diabetes prevalence emphasizes the complexity of this health issue, underlining the importance of considering a broader range of contributing factors in future analyses and interventions. The points seem like scatter hence it is difficult to build a model.

- **FINAL REMARKS:**

In conclusion, it's important to acknowledge the limitations of our analysis due to the constraints of the dataset. The dataset encompassed 3,142 unique FIPS codes for diabetes prevalence (% Diabetic), 363 FIPS codes for obesity rates (% Obesity), and 1,370 FIPS codes for physical inactivity (% Inactive). These relatively limited numbers of unique identifiers suggest that our analysis covered only a subset of regions within the United States. This constrained dataset size may have impacted the robustness and generalizability of our linear regression model.

A more comprehensive dataset, encompassing a wider range of geographic regions and factors, would have provided a more accurate representation of the complex interplay between various variables and diabetes prevalence. With such expanded data, we could have potentially developed a more reliable predictive model that accounts for a broader spectrum of influences on diabetes rates. Therefore, while our analysis provides valuable insights into the relationships between certain variables and diabetes prevalence, it is essential to recognize that the model's performance may have been hindered by data limitations.

# ❖ APPENDIX C: CODE

## Import Libraries

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split

import plotly.express as px
import plotly.graph_objects as go
from plotly.subplots import make_subplots

from sklearn.preprocessing import PowerTransformer
import scipy.stats as stats

import warnings
warnings.filterwarnings("ignore")

from sklearn.metrics import mean_squared_error as mse
from sklearn.metrics import mean_absolute_error as mae

from urllib.request import urlopen
import json
```

## Read Files from the data given

```
xls = pd.ExcelFile('cdc_diabetes_2018.xlsx')
df1 = pd.read_excel(xls, 'Diabetes')
df2 = pd.read_excel(xls, 'Obesity')
df3 = pd.read_excel(xls, 'Inactivity')
```

# Data Overview

```
In [3]: df1.head()
```

Out[3]:

| | YEAR | FIPS | COUNTY | STATEW | % DIABETIC |
|---|---|---|---|---|---|
| 0 | 2018 | 1001 | Autauga County | Alabama | 9.5 |
| 1 | 2018 | 1003 | Baldwin County | Alabama | 8.4 |
| 2 | 2018 | 1005 | Barbour County | Alabama | 13.5 |
| 3 | 2018 | 1007 | Bibb County | Alabama | 10.2 |
| 4 | 2018 | 1009 | Blount County | Alabama | 10.5 |

```
In [4]: df2.head()
```

Out[4]:

| | YEAR | FIPS | COUNTY | STATE | % OBESE |
|---|---|---|---|---|---|
| 0 | 2018 | 1011 | Bullock County | Alabama | 18.7 |
| 1 | 2018 | 2068 | Denali Borough | Alaska | 18.9 |
| 2 | 2018 | 2105 | Hoonah-Angoon Census Area | Alaska | 19.4 |
| 3 | 2018 | 2195 | Petersburg Census Area | Alaska | 17.2 |
| 4 | 2018 | 2230 | Skagway Municipality | Alaska | 18.3 |

```
In [5]: df3.head()
```

Out[5]:

| | YEAR | FIPDS | COUNTY | STATE | % INACTIVE |
|---|---|---|---|---|---|
| 0 | 2018 | 1011 | Bullock County | Alabama | 17.0 |
| 1 | 2018 | 1029 | Cleburne County | Alabama | 19.3 |
| 2 | 2018 | 1037 | Coosa County | Alabama | 16.8 |
| 3 | 2018 | 1063 | Greene County | Alabama | 16.8 |
| 4 | 2018 | 2013 | Aleutians East Borough | Alaska | 19.2 |

# Check for missing values

```
print("\nDiabetes")
print(df1.isnull().sum())
print("\nObesity")
print(df2.isnull().sum())
print("\nInactive")
print(df3.isnull().sum())
```

```
Diabetes
YEAR          0
FIPS          0
COUNTY        0
STATEW        0
% DIABETIC    0
dtype: int64

Obesity
YEAR       0
FIPS       0
click to scroll output; double click to hide
STATE      0
% OBESE    0
dtype: int64

Inactive
YEAR          0
FIPDS         0
COUNTY        0
STATE         0
% INACTIVE    0
dtype: int64
```

# Data Statistics

In [7]:
```
print("\nDiabetes")
df1.describe()
```

```
Diabetes
```

Out[7]:

| | YEAR | FIPS | % DIABETIC |
|---|---|---|---|
| count | 3142.0 | 3142.000000 | 3142.000000 |
| mean | 2018.0 | 30383.649268 | 8.719796 |
| std | 0.0 | 15162.508374 | 1.794854 |
| min | 2018.0 | 1001.000000 | 3.800000 |
| 25% | 2018.0 | 18177.500000 | 7.300000 |
| 50% | 2018.0 | 29176.000000 | 8.400000 |
| 75% | 2018.0 | 45080.500000 | 9.700000 |
| max | 2018.0 | 56045.000000 | 17.900000 |

In [8]:
```
print("\nObesity")
df2.describe()
```

```
Obesity
```

Out[8]:

| | YEAR | FIPS | % OBESE |
|---|---|---|---|
| count | 363.0 | 363.000000 | 363.000000 |
| mean | 2018.0 | 33817.906336 | 18.264738 |
| std | 0.0 | 16810.094056 | 1.038311 |
| min | 2018.0 | 1011.000000 | 10.500000 |
| 25% | 2018.0 | 17020.000000 | 17.900000 |
| 50% | 2018.0 | 46061.000000 | 18.300000 |
| 75% | 2018.0 | 48262.000000 | 19.000000 |
| max | 2018.0 | 56039.000000 | 19.500000 |

In [9]:
```
print("\nInactive")
df3.describe()
```

```
Inactive
```

Out[9]:

| | YEAR | FIPDS | % INACTIVE |
|---|---|---|---|
| count | 1370.0 | 1370.000000 | 1370.000000 |
| mean | 2018.0 | 32715.840876 | 16.543358 |
| std | 0.0 | 15891.407141 | 1.926010 |
| min | 2018.0 | 1011.000000 | 8.800000 |
| 25% | 2018.0 | 18076.000000 | 15.000000 |
| 50% | 2018.0 | 35012.000000 | 16.700000 |
| 75% | 2018.0 | 48196.500000 | 18.100000 |
| max | 2018.0 | 56039.000000 | 19.500000 |

# Features dtypes

```
print("\nDiabetes")
df1.info(z)
```

```
Diabetes
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3142 entries, 0 to 3141
Data columns (total 5 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   YEAR         3142 non-null   int64
 1   FIPS         3142 non-null   int64
 2   COUNTY       3142 non-null   object
 3   STATEW       3142 non-null   object
 4   % DIABETIC   3142 non-null   float64
dtypes: float64(1), int64(2), object(2)
memory usage: 122.9+ KB
```

```
print("Obesity")
df2.info()
```

```
Obesity
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 363 entries, 0 to 362
Data columns (total 5 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   YEAR      363 non-null    int64
 1   FIPS      363 non-null    int64
 2   COUNTY    363 non-null    object
 3   STATE     363 non-null    object
 4   % OBESE   363 non-null    float64
dtypes: float64(1), int64(2), object(2)
memory usage: 14.3+ KB
```

```
print("\nInactive")
df3.info()
```

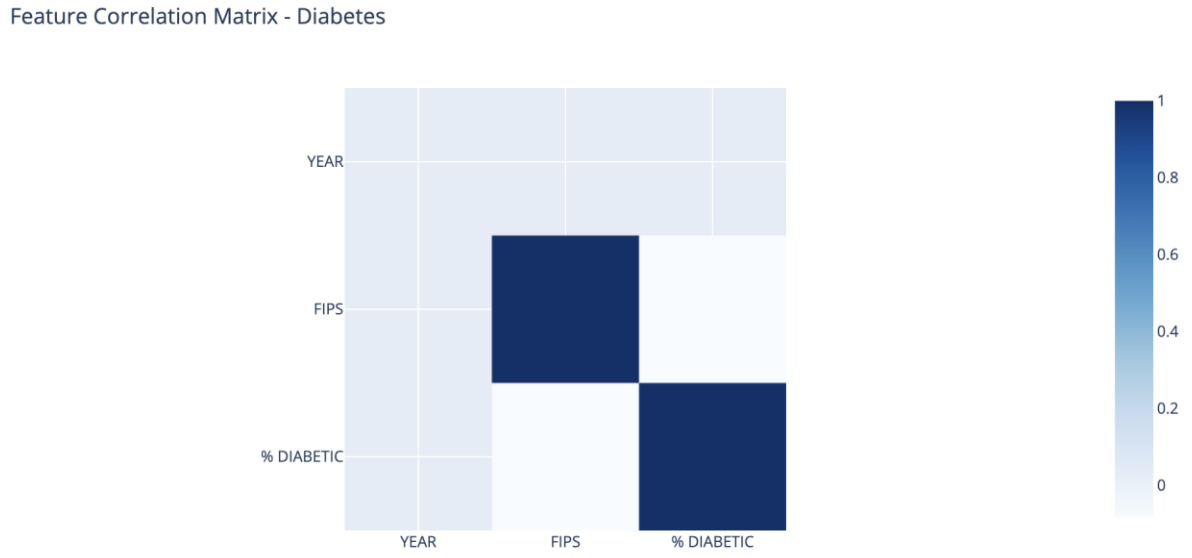
```
Inactive
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1370 entries, 0 to 1369
Data columns (total 5 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   YEAR         1370 non-null   int64
 1   FIPDS        1370 non-null   int64
 2   COUNTY       1370 non-null   object
 3   STATE        1370 non-null   object
 4   % INACTIVE   1370 non-null   float64
dtypes: float64(1), int64(2), object(2)
memory usage: 53.6+ KB
```

# Shape of Data

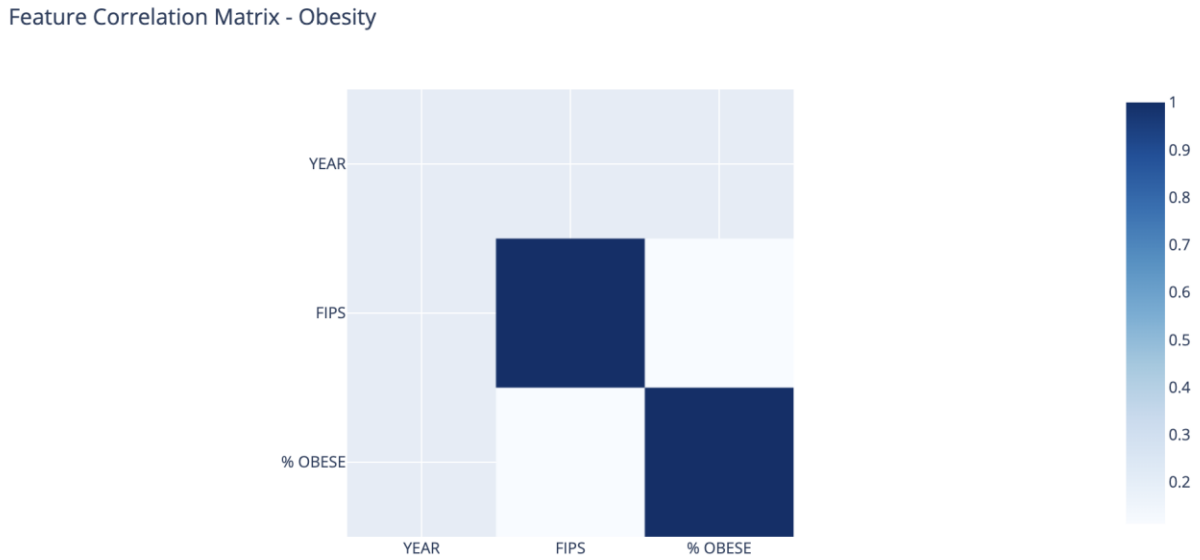
```
print('The total shape of the Diabetes data is: ',df1.shape)
print('The total shape of the Obesity data is: ',df2.shape)
print('The total shape of the Inactive starter data is: ',df3.shape)
```

# Correlation values

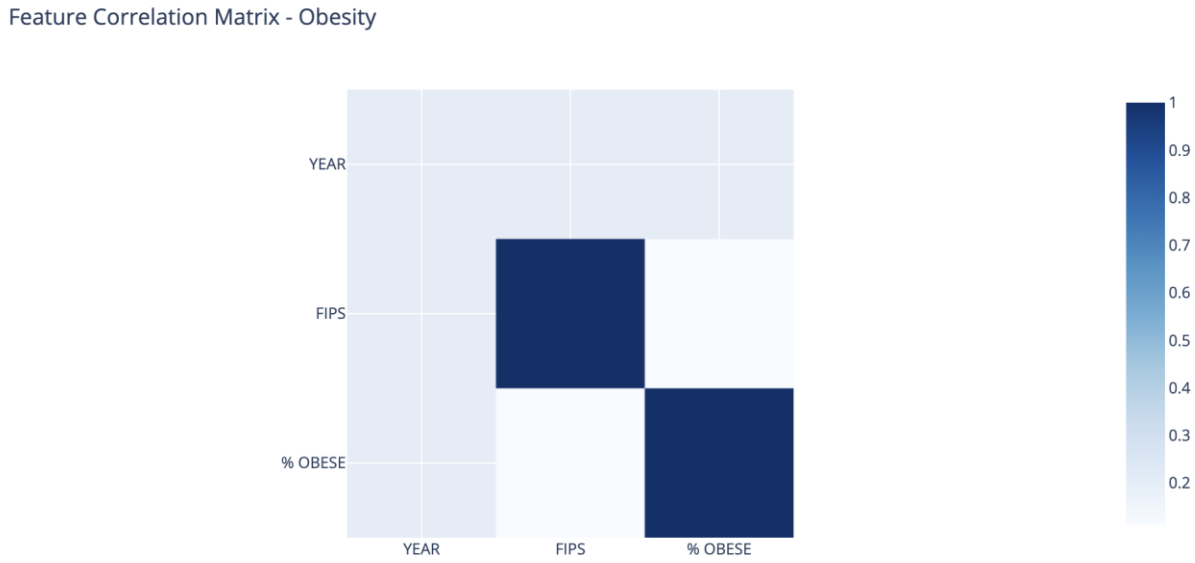
```
corr_matrix = df1.corr()
fig = px.imshow(corr_matrix, color_continuous_scale='blues', title="Feature Correlation Matrix - Diabetes")
fig.show()
```




```
corr_matrix = df2.corr()
fig  =  px.imshow(corr_matrix,  color_continuous_scale='blues',
title="Feature Correlation Matrix - Obesity")
fig.show()
```

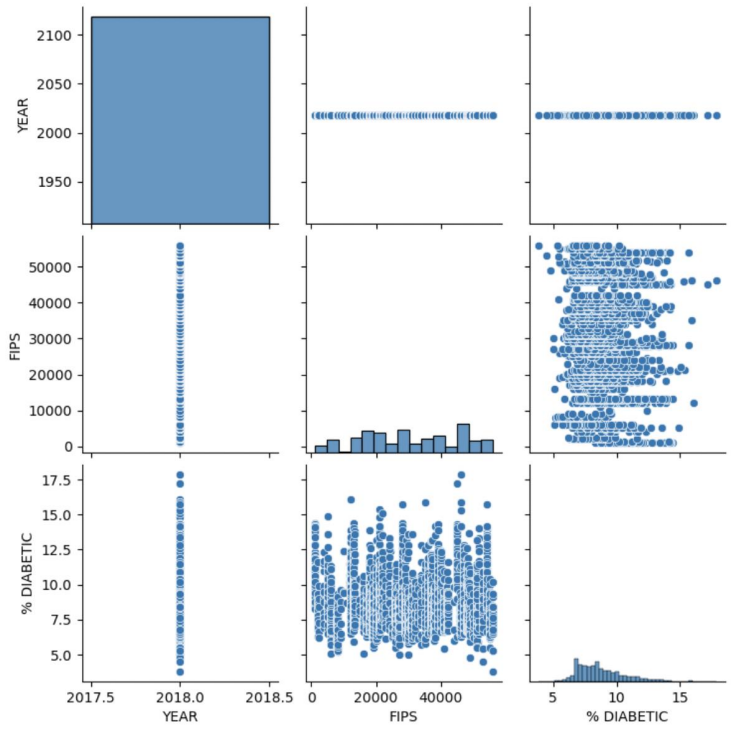



```
corr_matrix = df3.corr()
fig  =  px.imshow(corr_matrix,  color_continuous_scale='blues',
title="Feature Correlation Matrix - Inactive")
fig.show()
```

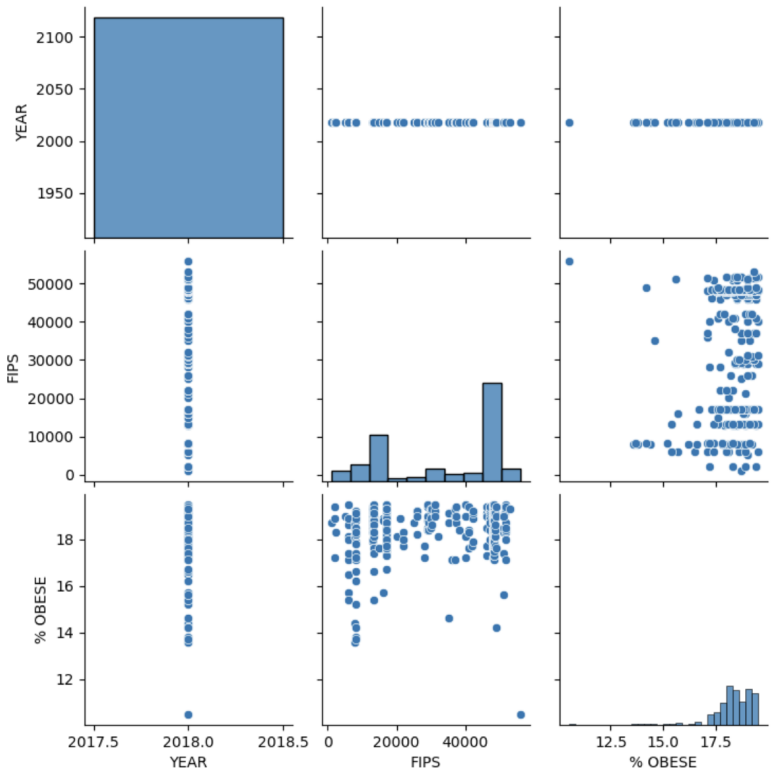



# Exploratory Data Analysis

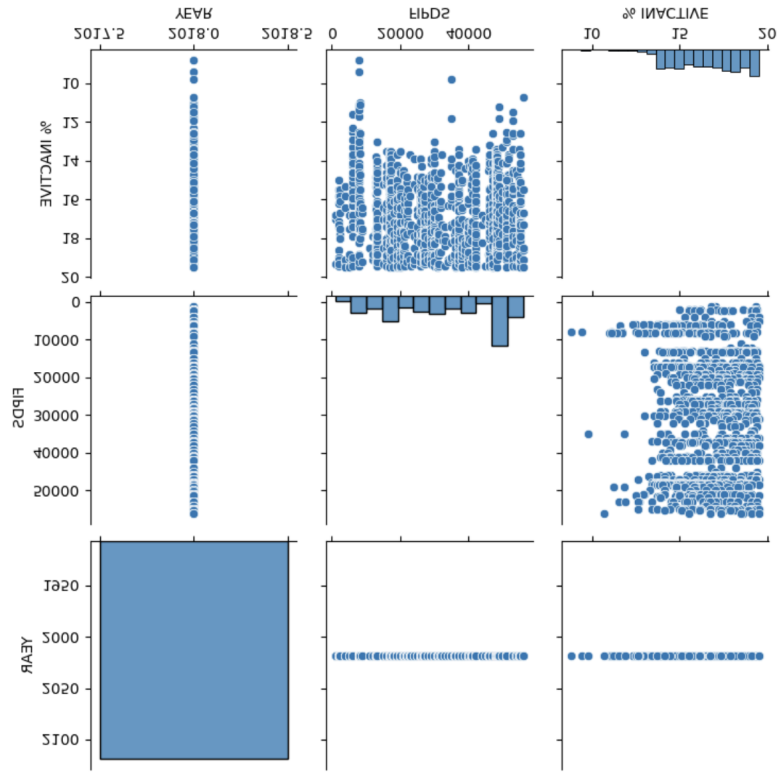
```
ns.pairplot(df1)
```




```
sns.pairplot(df2)
```




```
sns.pairplot(df3)
```



# Diabetes - Correlation Matrix

```
df1.corr()
```

Out[20]:

| | YEAR | FIPS | % DIABETIC |
|---|---|---|---|
| **YEAR** | NaN | NaN | NaN |
| **FIPS** | NaN | 1.000000 | -0.083521 |
| **% DIABETIC** | NaN | -0.083521 | 1.000000 |

# Obesity - Correlation Matrix

```
df2.corr()
```

Out[21]:

| | YEAR | FIPS | % OBESE |
|---|---|---|---|
| **YEAR** | NaN | NaN | NaN |
| **FIPS** | NaN | 1.000000 | 0.112354 |
| **% OBESE** | NaN | 0.112354 | 1.000000 |

# Inactive - Correlation Matrix

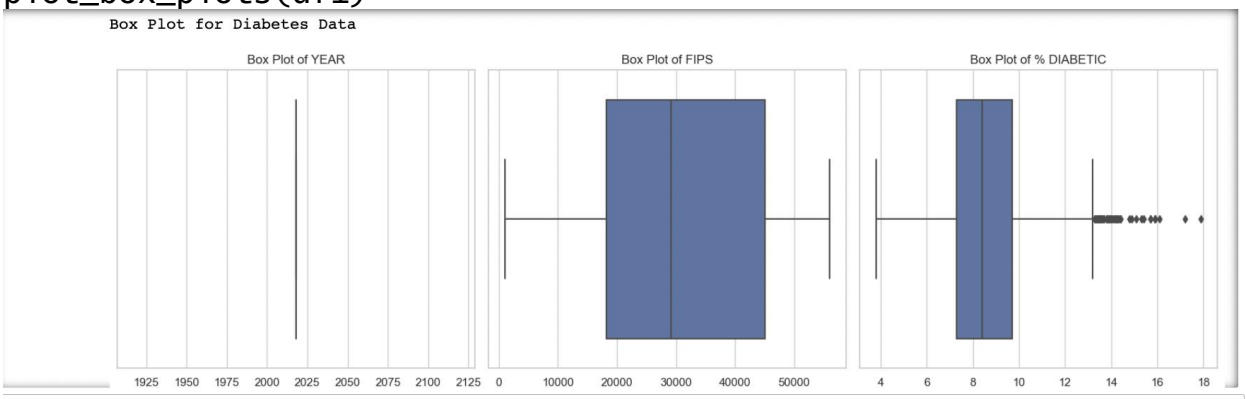
```
df3.corr()
```

| | YEAR | FIPDS | % INACTIVE |
|---|---|---|---|
| **YEAR** | NaN | NaN | NaN |
| **FIPDS** | NaN | 1.00000 | -0.06312 |
| **% INACTIVE** | NaN | -0.06312 | 1.00000 |

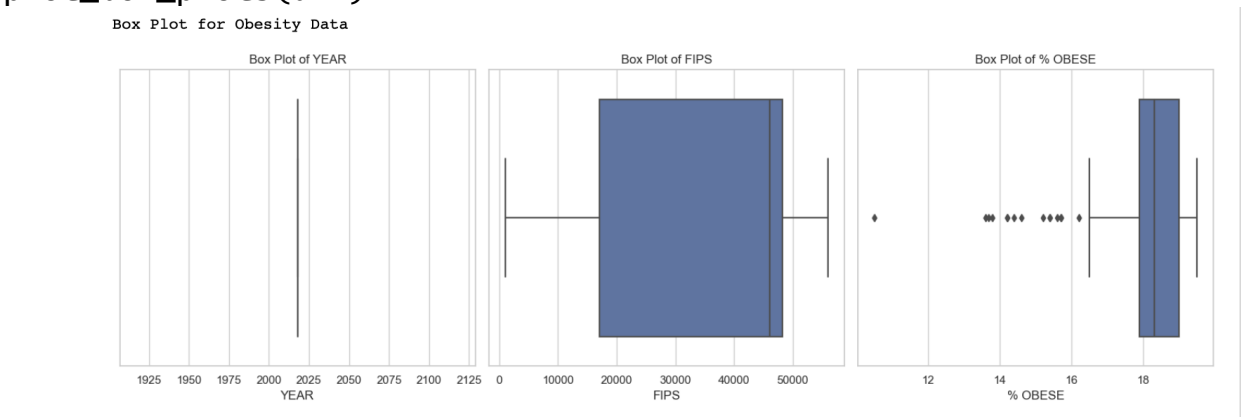
# Box Plot

```
def plot_box_plots(df):
    numeric_columns = df.select_dtypes(include=[pd.np.number])
    sns.set(style="whitegrid")
    num_features = numeric_columns.shape[1]
    num_rows = (num_features + 2) // 3
    num_cols = min(num_features, 3)

    # Create a figure with subplots
    fig, axes = plt.subplots(num_rows, num_cols, figsize=(15, 5 *
num_rows))

    # Flatten the axes if there's only one row
    if num_rows == 1:
        axes = axes.reshape(1, -1)

    # Loop through the columns and plot box plots
    for i, col in enumerate(numeric_columns.columns):
        row_index = i // num_cols
        col_index = i % num_cols

        ax = axes[row_index, col_index]

        sns.boxplot(x=df[col], ax=ax)
        ax.set_title(f'Box Plot of {col}')
```

```
        ax.set_xlabel(col)

    # Remove any empty subplots
    for i in range(num_features, num_rows * num_cols):
        fig.delaxes(axes.flatten()[i])

    # Adjust the layout and display the plots
    plt.tight_layout()
    plt.show()

print("Box Plot for Diabetes Data")
plot_box_plots(df1)
```

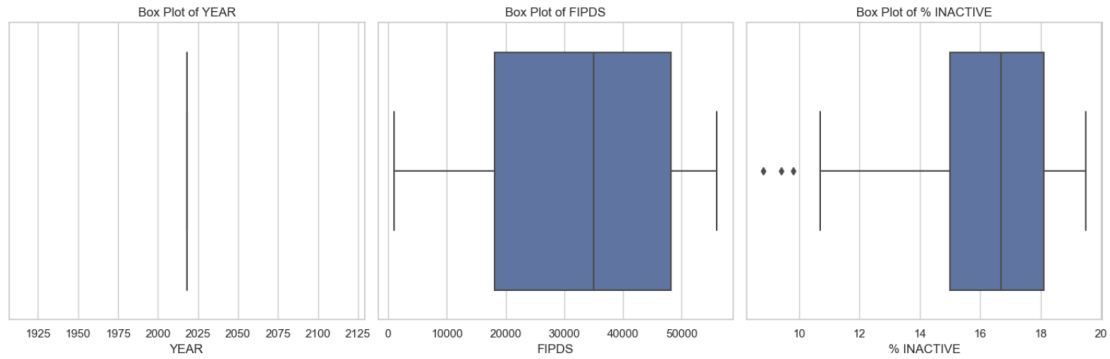



```
print("Box Plot for Obesity Data")
plot_box_plots(df2)
```



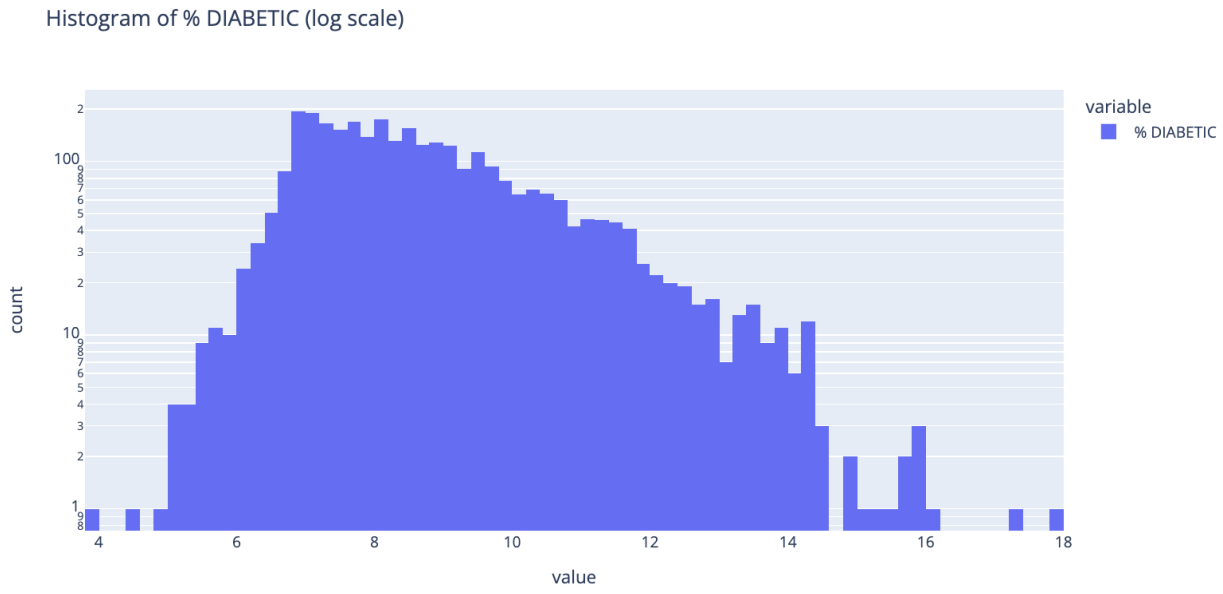

```
print("Box Plot for Inactive Data")
plot_box_plots(df3)
```

Box Plot for Inactive Data

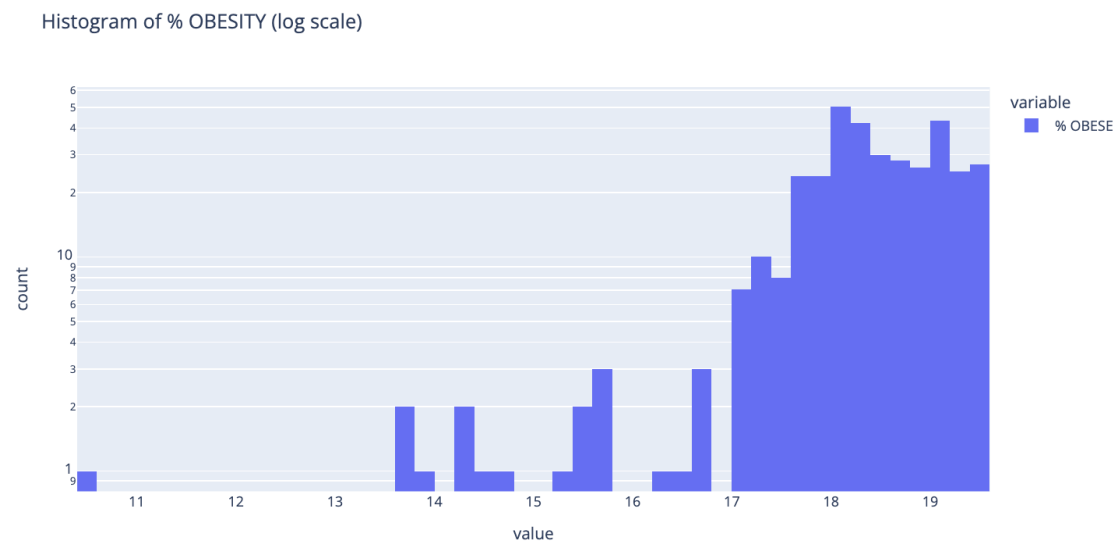



# Histogram

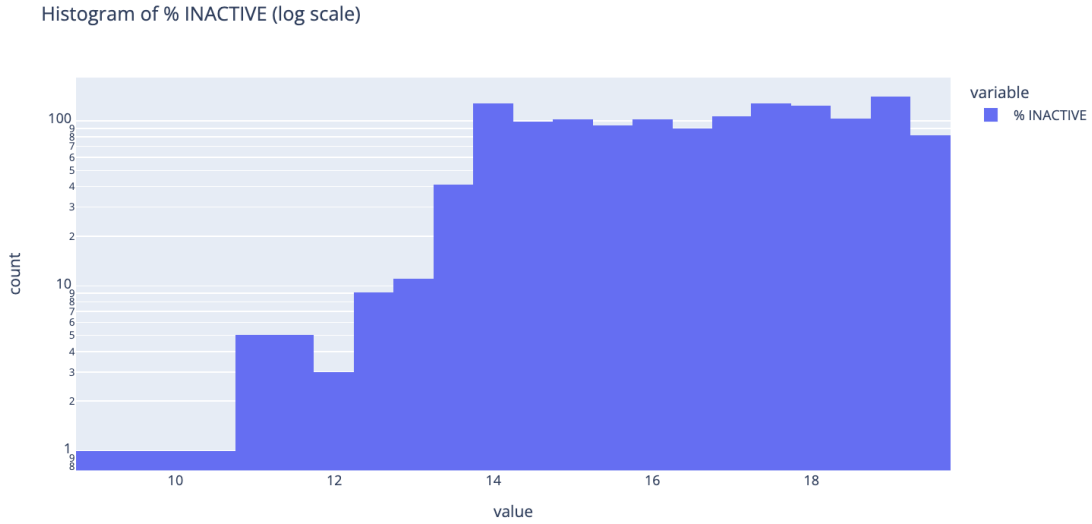
```
fig = px.histogram(df1["% DIABETIC"], log_y=True, title='Histogram of % DIABETIC (log scale)')
fig.update_layout(yaxis_type='log')
fig.show()
```




```
fig = px.histogram(df2["% OBESE"], log_y=True, title='Histogram of % OBESITY (log scale)')
fig.update_layout(yaxis_type='log')
fig.show()
```




```
fig = px.histogram(df3["% INACTIVE"], log_y=True, title='Histogram of % INACTIVE (log scale)')
fig.update_layout(yaxis_type='log')
fig.show()
```




# Unique State in diabetic

```
df1['STATEW'].nunique() , df1['STATEW'].unique()
```

```
Out[30]: (51,
          array(['Alabama', 'Alaska', 'Arizona', 'Arkansas', 'California',
                 'Colorado', 'Connecticut', 'Delaware', 'District of Columbia',
                 'Florida', 'Georgia', 'Hawaii', 'Idaho', 'Illinois', 'Indiana',
                 'Iowa', 'Kansas', 'Kentucky', 'Louisiana', 'Maine', 'Maryland',
                 'Massachusetts', 'Michigan', 'Minnesota', 'Mississippi',
                 'Missouri', 'Montana', 'Nebraska', 'Nevada', 'New Hampshire',
                 'New Jersey', 'New Mexico', 'New York', 'North Carolina',
                 'North Dakota', 'Ohio', 'Oklahoma', 'Oregon', 'Pennsylvania',
                 'Rhode Island', 'South Carolina', 'South Dakota', 'Tennessee',
                 'Texas', 'Utah', 'Vermont', 'Virginia', 'Washington',
                 'West Virginia', 'Wisconsin', 'Wyoming'], dtype=object))
```

# Unique State in obesity

df2['STATE'].nunique() , df2['STATE'].unique()

```
Out[31]: (33,
          array(['Alabama', 'Alaska', 'Arkansas', 'California', 'Colorado',
                 'Georgia', 'Hawaii', 'Idaho', 'Illinois', 'Kansas', 'Kentucky',
                 'Louisiana', 'Massachusetts', 'Michigan', 'Mississippi',
                 'Missouri', 'Montana', 'Nebraska', 'Nevada', 'New Mexico',
                 'New York', 'North Carolina', 'North Dakota', 'Oklahoma', 'Oregon',
                 'Pennsylvania', 'South Dakota', 'Tennessee', 'Texas', 'Utah',
                 'Virginia', 'Washington', 'Wyoming'], dtype=object))
```

# Unique State in Inactivity

df3['STATE'].nunique() , df3['STATE'].unique()

```
Out[32]: (50,
          array(['Alabama', 'Alaska', 'Arizona', 'Arkansas', 'California',
                 'Colorado', 'Connecticut', 'District of Columbia', 'Florida',
                 'Georgia', 'Hawaii', 'Idaho', 'Illinois', 'Indiana', 'Iowa',
                 'Kansas', 'Kentucky', 'Louisiana', 'Maine', 'Maryland',
                 'Massachusetts', 'Michigan', 'Minnesota', 'Mississippi',
                 'Missouri', 'Montana', 'Nebraska', 'Nevada', 'New Hampshire',
                 'New Jersey', 'New Mexico', 'New York', 'North Carolina',
                 'North Dakota', 'Ohio', 'Oklahoma', 'Oregon', 'Pennsylvania',
                 'Rhode Island', 'South Carolina', 'South Dakota', 'Tennessee',
                 'Texas', 'Utah', 'Vermont', 'Virginia', 'Washington',
                 'West Virginia', 'Wisconsin', 'Wyoming'], dtype=object))
```

# Let's Check differences by State for % DIABETIC

```
sort_std = df1.groupby(['STATEW']).describe()['% DIABETIC'].sort_values('std').index
```

```
each_state = df1.groupby(['STATEW']).describe()['% DIABETIC'].sort_values('std')

each_state
```

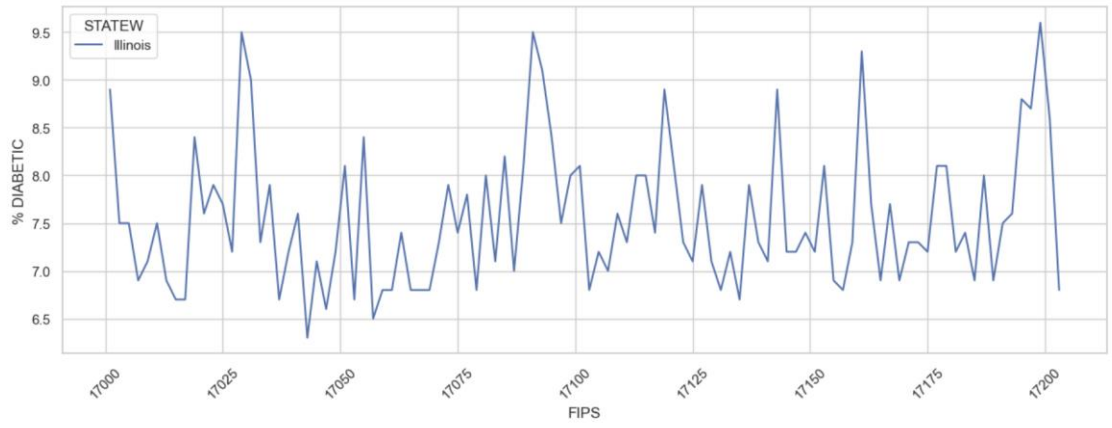
Out[33]:

| STATEW | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| Illinois | 102.0 | 7.552941 | 0.736153 | 6.3 | 7.000 | 7.35 | 8.000 | 9.6 |
| Massachusetts | 14.0 | 7.378571 | 0.789526 | 6.2 | 6.750 | 7.45 | 8.000 | 8.3 |
| Vermont | 14.0 | 7.378571 | 0.880715 | 6.2 | 6.700 | 7.40 | 7.850 | 9.3 |
| Minnesota | 87.0 | 7.811494 | 0.908029 | 5.0 | 7.150 | 7.80 | 8.400 | 9.9 |
| Wisconsin | 72.0 | 7.480556 | 0.969484 | 5.5 | 6.800 | 7.30 | 8.025 | 10.4 |
| Colorado | 64.0 | 6.845313 | 0.984794 | 5.2 | 6.000 | 6.80 | 7.500 | 9.3 |
| Hawaii | 5.0 | 8.740000 | 1.001499 | 7.9 | 8.100 | 8.40 | 8.900 | 10.4 |
| Pennsylvania | 67.0 | 8.350746 | 1.041760 | 6.7 | 7.500 | 8.40 | 9.100 | 11.6 |
| Iowa | 99.0 | 7.906061 | 1.106458 | 5.5 | 7.050 | 7.80 | 8.600 | 11.2 |
| Idaho | 44.0 | 7.943182 | 1.115722 | 5.1 | 7.100 | 7.70 | 8.700 | 10.6 |

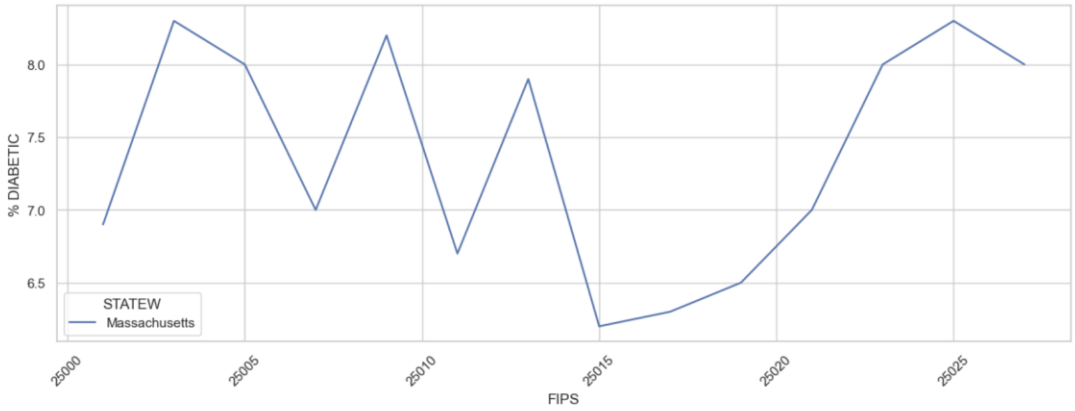
# Top 5

## For Illinois

```
plt.figure(figsize=(15,5))
sns.lineplot(data=df1[df1['STATEW'].isin(['Illinois'])], x='FIPS',y='% DIABETIC', hue='STATEW')
locs,labels = plt.xticks()
plt.setp(labels, rotation=45)
plt.show()
```



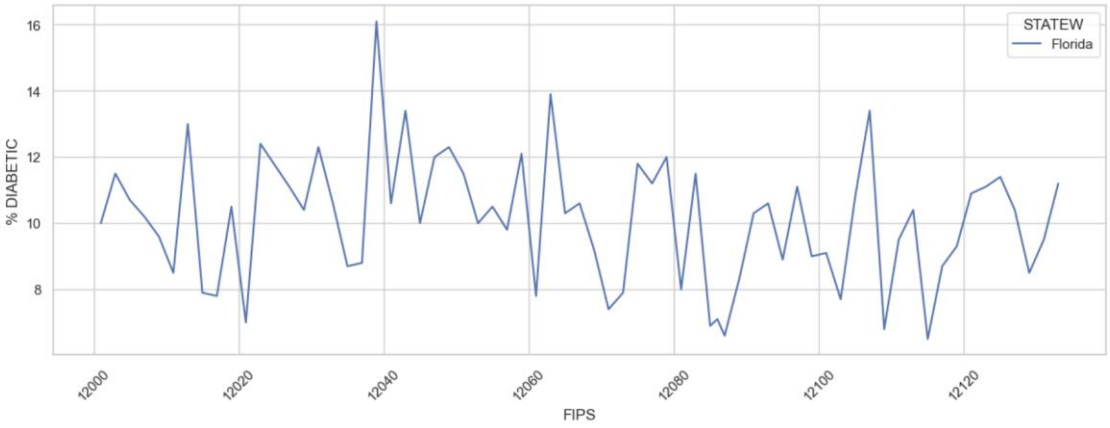
## For Massachusetts




# Bottom 5

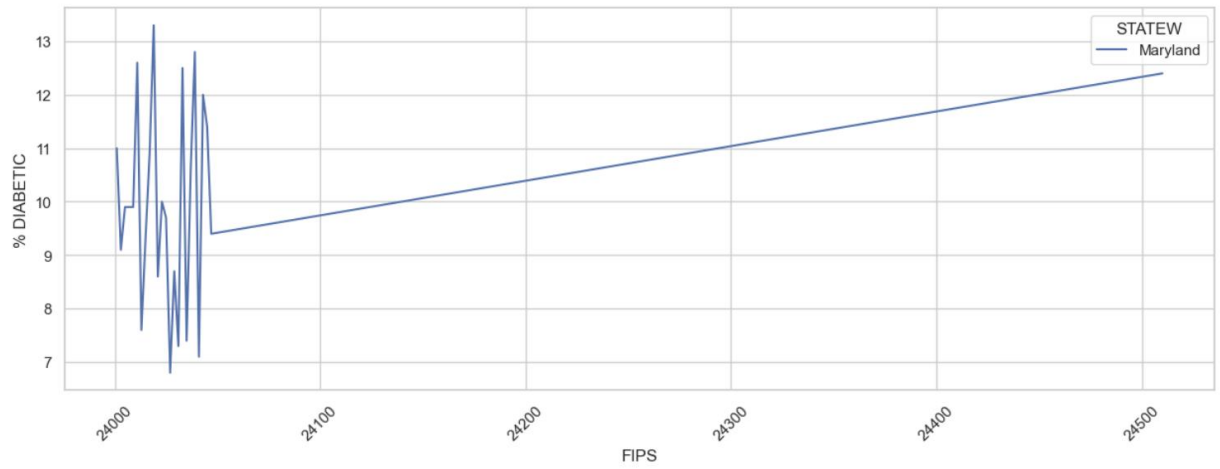
## For Florida

```
plt.figure(figsize=(15,5))
sns.lineplot(data=df1[df1['STATEW'].isin(['Florida'])],
x='FIPS',y='% DIABETIC', hue='STATEW')
locs,labels = plt.xticks()
plt.setp(labels, rotation=45)
plt.show()
```




## For Maryland

```
plt.figure(figsize=(15,5))
sns.lineplot(data=df1[df1['STATEW'].isin(['Maryland'])],
x='FIPS',y='% DIABETIC', hue='STATEW')
locs,labels = plt.xticks()
plt.setp(labels, rotation=45)
plt.show()
```




# Let's Check differences by State for % OBESE

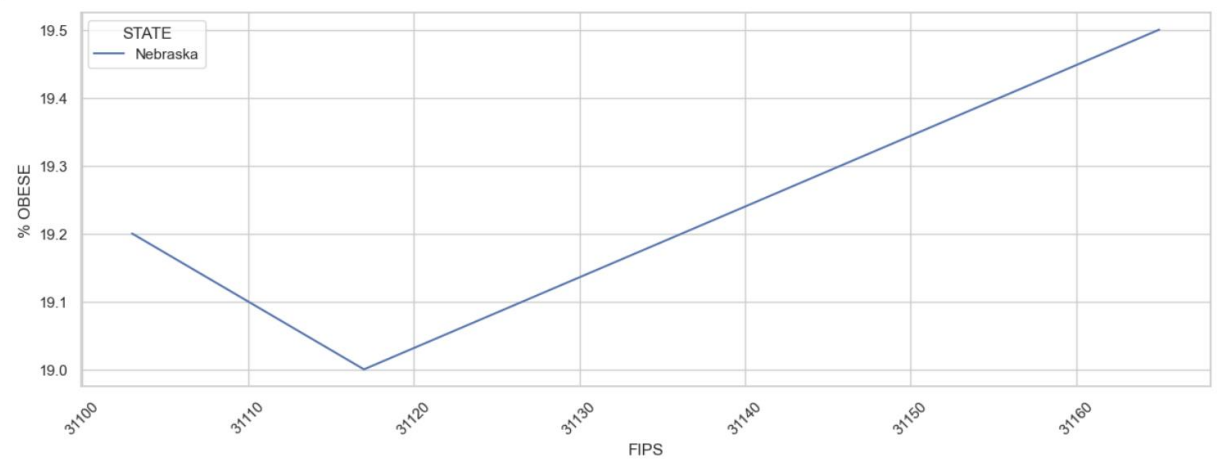
# % OBESE

```
sort_std = df2.groupby(['STATE']).describe()['%
OBESE'].sort_values('std').index

each_state = df2.groupby(['STATE']).describe()['%
OBESE'].sort_values('std')

each_state
```

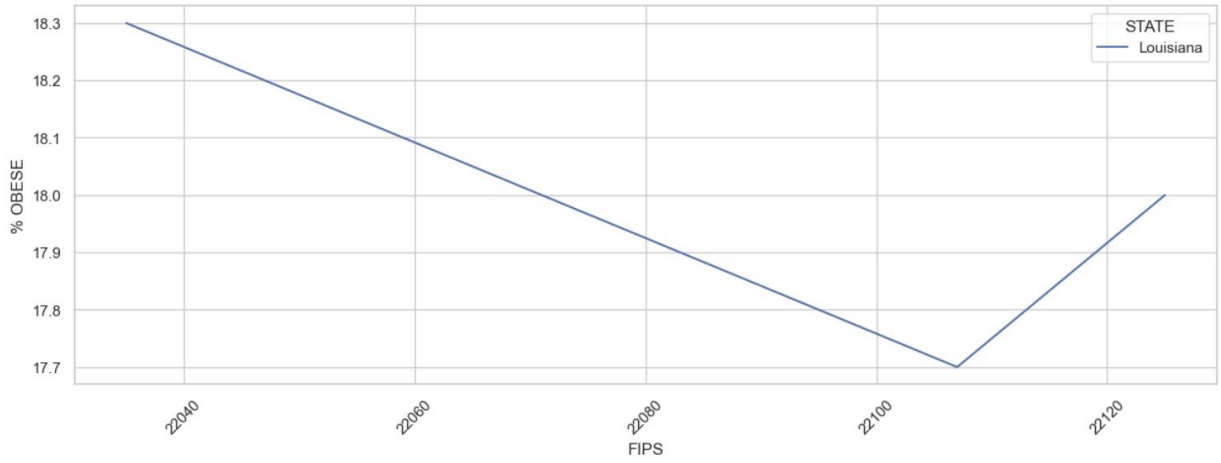
| STATE | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **Nebraska** | 3.0 | 19.233333 | 0.251661 | 19.0 | 19.100 | 19.20 | 19.350 | 19.5 |
| **Louisiana** | 3.0 | 18.000000 | 0.300000 | 17.7 | 17.850 | 18.00 | 18.150 | 18.3 |
| **Mississippi** | 2.0 | 17.450000 | 0.353553 | 17.2 | 17.325 | 17.45 | 17.575 | 17.7 |
| **Montana** | 6.0 | 18.966667 | 0.355903 | 18.5 | 18.675 | 19.05 | 19.275 | 19.3 |
| **Missouri** | 10.0 | 18.870000 | 0.452278 | 18.4 | 18.500 | 18.75 | 19.300 | 19.5 |

# TOP 5

```
plt.figure(figsize=(15,5))
sns.lineplot(data=df2[df2['STATE'].isin(['Nebraska'])], x='FIPS',y='%
OBESE', hue='STATE')
locs,labels = plt.xticks()
plt.setp(labels, rotation=45)
plt.show()
```




```
plt.figure(figsize=(15,5))
sns.lineplot(data=df2[df2['STATE'].isin(['Louisiana'])], x='FIPS',y='%
OBESE', hue='STATE')
locs,labels = plt.xticks()
plt.setp(labels, rotation=45)
plt.show()
```




# Bottom 5

Standard Deviation is NAN - So No Plots

# Let's Check differences by State for % INACTIVE

# % INACTIVE

```
sort_std = df3.groupby(['STATE']).describe()['%
INACTIVE'].sort_values('std').index

each_state = df3.groupby(['STATE']).describe()['%
INACTIVE'].sort_values('std')

each_state
```

| STATE | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| Florida | 3.0 | 18.600000 | 0.624500 | 17.9 | 18.350 | 18.80 | 18.950 | 19.1 |
| Rhode Island | 3.0 | 17.433333 | 0.709460 | 16.8 | 17.050 | 17.30 | 17.750 | 18.2 |
| New Hampshire | 5.0 | 18.240000 | 0.743640 | 17.4 | 17.900 | 18.10 | 18.400 | 19.4 |
| Connecticut | 5.0 | 17.480000 | 1.089495 | 16.4 | 16.700 | 17.50 | 17.600 | 19.2 |
| Indiana | 11.0 | 17.627273 | 1.094615 | 16.1 | 16.950 | 17.80 | 18.100 | 19.5 |

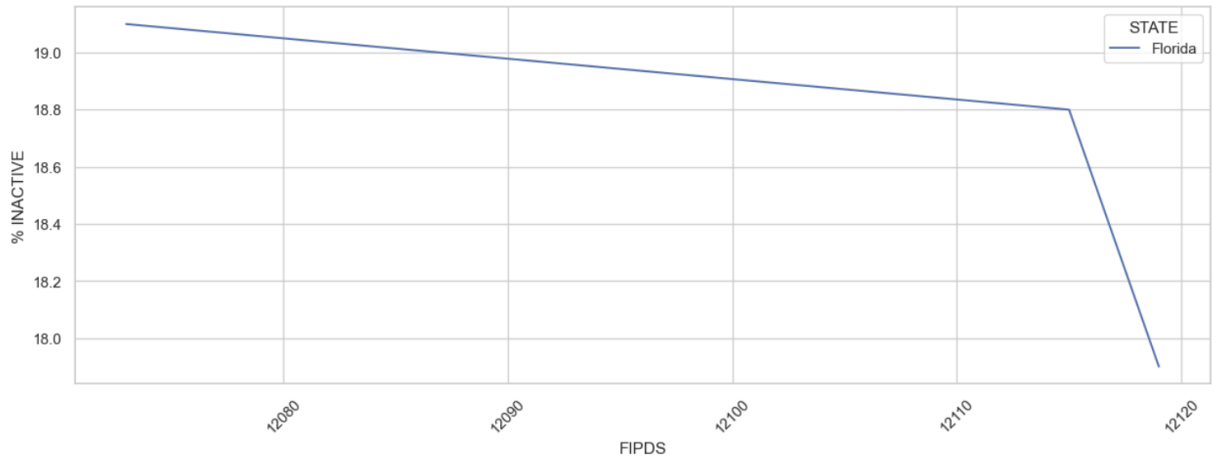
```
df3.head()
```

Out[51]:

| | YEAR | FIPDS | COUNTY | STATE | % INACTIVE |
|---|---|---|---|---|---|
| 0 | 2018 | 1011 | Bullock County | Alabama | 17.0 |
| 1 | 2018 | 1029 | Cleburne County | Alabama | 19.3 |
| 2 | 2018 | 1037 | Coosa County | Alabama | 16.8 |
| 3 | 2018 | 1063 | Greene County | Alabama | 16.8 |
| 4 | 2018 | 2013 | Aleutians East Borough | Alaska | 19.2 |

# For Florida

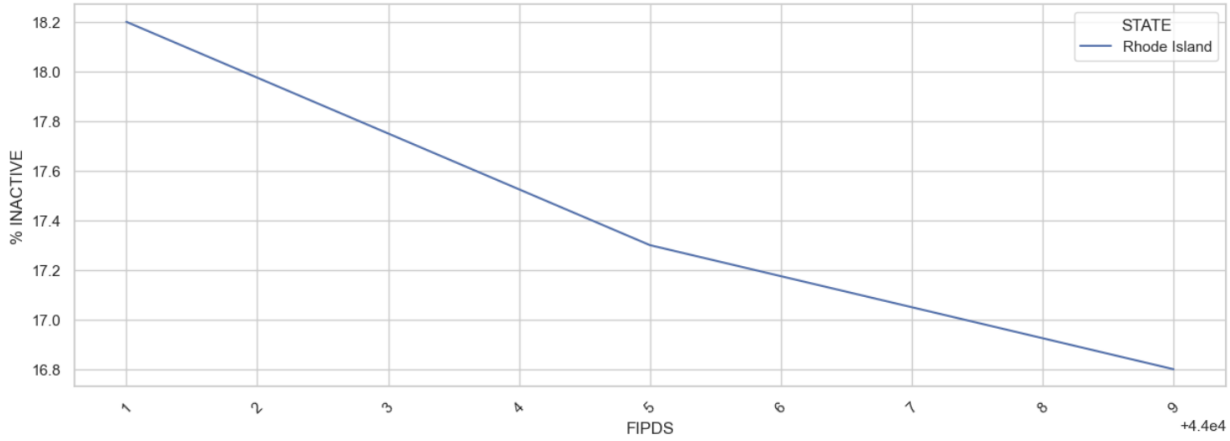
```
plt.figure(figsize=(15,5))
sns.lineplot(data=df3[df3['STATE'].isin(['Florida'])], x='FIPDS',y='%
INACTIVE', hue='STATE')
locs,labels = plt.xticks()
plt.setp(labels, rotation=45)
plt.show()
```




# For Rhode Islan

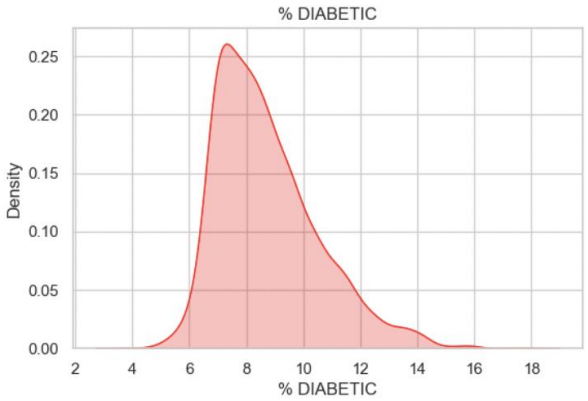
```
plt.figure(figsize=(15,5))
```

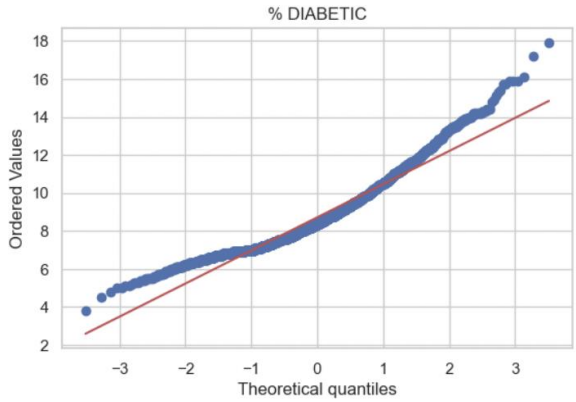
```
sns.lineplot(data=df3[df3['STATE'].isin(['Rhode Island'])],
x='FIPDS',y='% INACTIVE', hue='STATE')
locs,labels = plt.xticks()
plt.setp(labels, rotation=45)
plt.show()
```



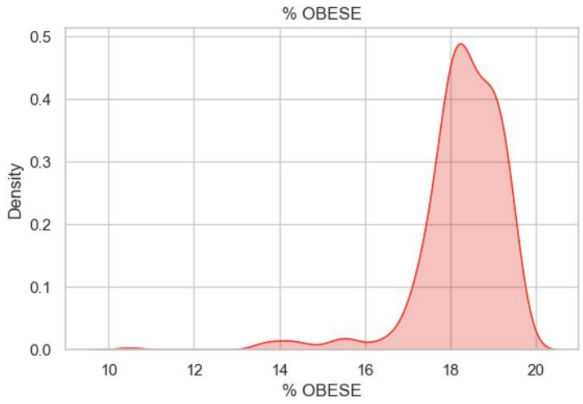

# Let's visualized distribution of our feature

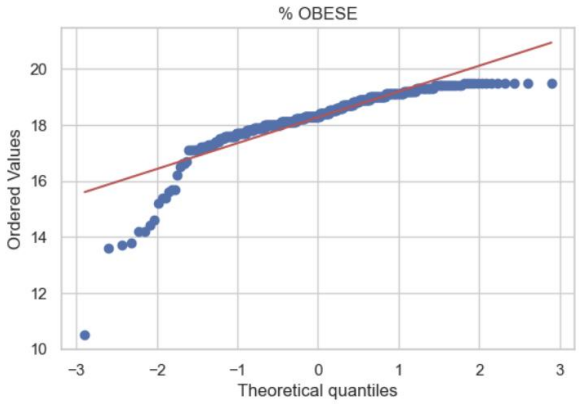
```
import scipy.stats as stats
plt.figure(figsize=(14,4))
plt.subplot(121)
sns.kdeplot(df1['% DIABETIC'], shade=True, color='red')
plt.title('% DIABETIC')

plt.subplot(122)
stats.probplot(df1['% DIABETIC'],dist="norm", plot=plt)
plt.title('% DIABETIC')
plt.show();
```



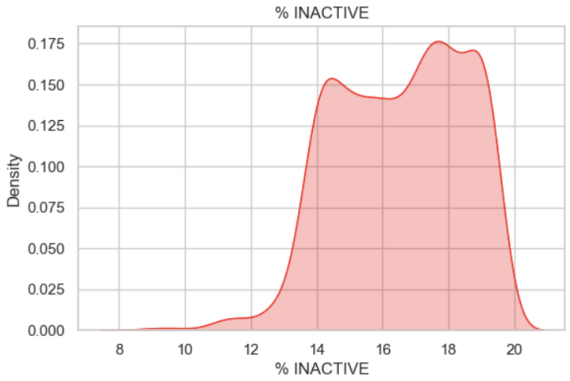



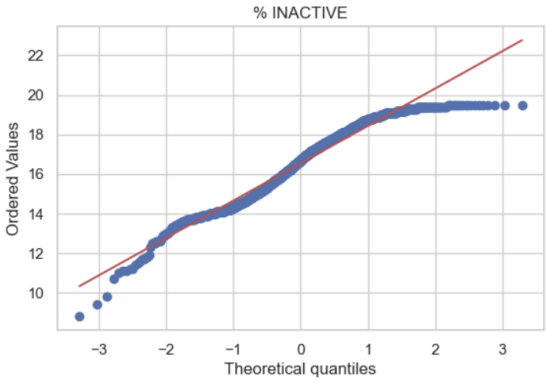

```
plt.figure(figsize=(14,4))
plt.subplot(121)
sns.kdeplot(df2['% OBESE'], shade=True, color='red')
plt.title("% OBESE")

plt.subplot(122)
stats.probplot(df2['% OBESE'],dist="norm", plot=plt)
plt.title("% OBESE")
plt.show();
```




```
plt.figure(figsize=(14,4))
plt.subplot(121)
sns.kdeplot(df3['% INACTIVE'], shade=True, color='red')
plt.title("% INACTIVE")

plt.subplot(122)
stats.probplot(df3['% INACTIVE'],dist="norm", plot=plt)
plt.title("% INACTIVE")
plt.show();
```




# Box - Cox Transformation

## FOR DIABETES

```
#Applying Box-Cox Tranform
pt = PowerTransformer(method='box-cox')
df1_transformed = pt.fit_transform(1+df1["%
DIABETIC"].to_numpy().reshape(-1, 1))
df1_transformed = pd.DataFrame(df1_transformed, columns=['% DIABETIC'])
pd.DataFrame({'cols':"% DIABETIC", 'Box_Cox_lamdas':pt.lambdas_})
```

Out[62]:

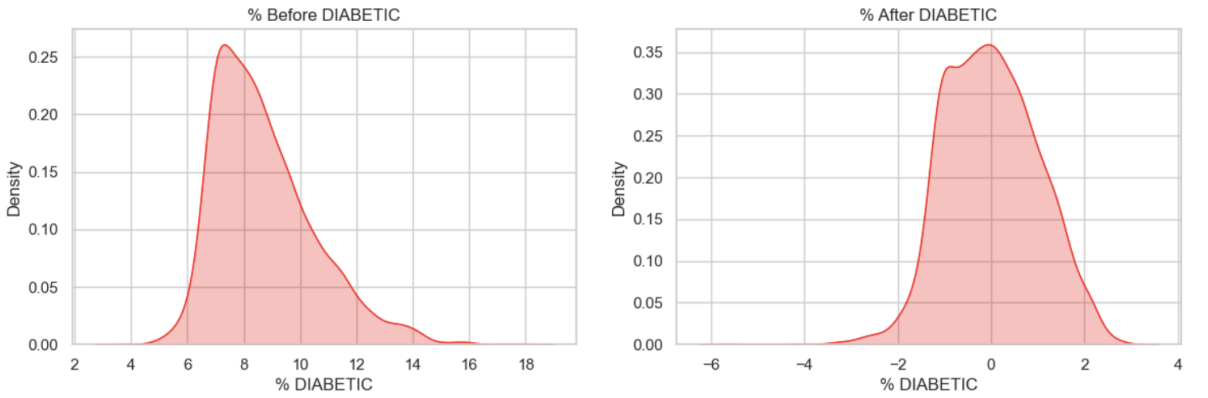
| | cols | Box_Cox_lamdas |
|---|---|---|
| 0 | % DIABETIC | -0.982737 |

```
df1_transformed.head()
```

Out[63]:

| | % DIABETIC |
|---|---|
| 0 | 0.603006 |
| 1 | -0.016331 |
| 2 | 2.068290 |
| 3 | 0.934292 |
| 4 | 1.064029 |

```
plt.figure(figsize=(14,4))
plt.subplot(121)
sns.kdeplot(df1["% DIABETIC"],shade=True, color='red')
plt.title("% Before DIABETIC")

plt.subplot(122)
sns.kdeplot(df1_transformed['% DIABETIC'],shade=True, color='red')
plt.title("% After DIABETIC")

plt.show()
```




# FOR OBESITY

```
#Applying Box-Cox Tranform
pt = PowerTransformer(method='box-cox')
df2_transformed = pt.fit_transform(1+df2["% OBESE"].to_numpy().reshape(-1, 1))
df2_transformed = pd.DataFrame(df2_transformed, columns=['% OBESE'])
pd.DataFrame({'cols':"% OBESE", 'Box_Cox_lamdas':pt.lambdas_})
```

Out[65]:

| | cols | Box_Cox_lamdas |
|---|---|---|
| 0 | % OBESE | 10.130292 |

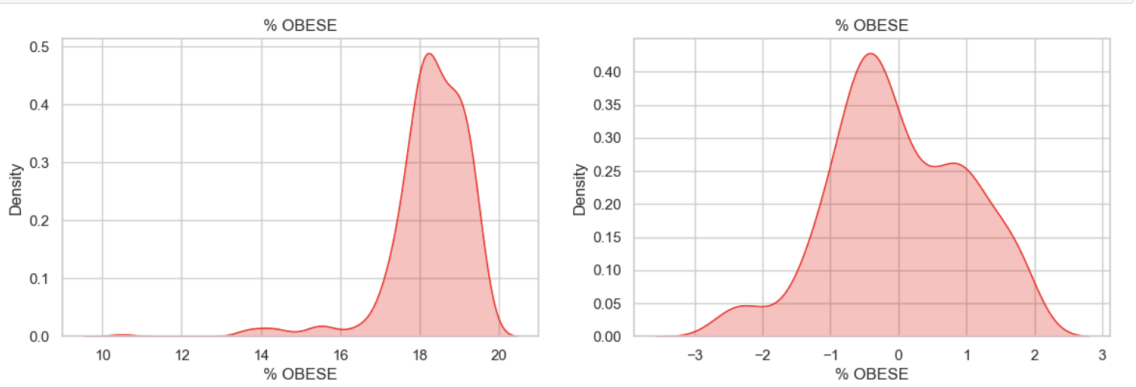
```
df2_transformed.head()
```

Out[66]:

| | % OBESE |
|---|---|
| 0 | 0.367667 |
| 1 | 0.694183 |
| 2 | 1.653615 |
| 3 | -1.304219 |
| 4 | -0.200955 |

```
plt.figure(figsize=(14,4))
plt.subplot(121)
sns.kdeplot(df2["% OBESE"],shade=True, color='red')
plt.title("% OBESE")
```

```
plt.subplot(122)
sns.kdeplot(df2_transformed['% OBESE'],shade=True, color='red')
plt.title("% OBESE")

plt.show()
```




# INACTIVE

```
#Applying Box-Cox Tranform
pt = PowerTransformer(method='box-cox')
df3_transformed = pt.fit_transform(1+df3["%
INACTIVE"].to_numpy().reshape(-1, 1))
df3_transformed = pd.DataFrame(df3_transformed, columns=['% INACTIVE'])
pd.DataFrame({'cols':"% INACTIVE", 'Box_Cox_lamdas':pt.lambdas_})
```

Out[68]:

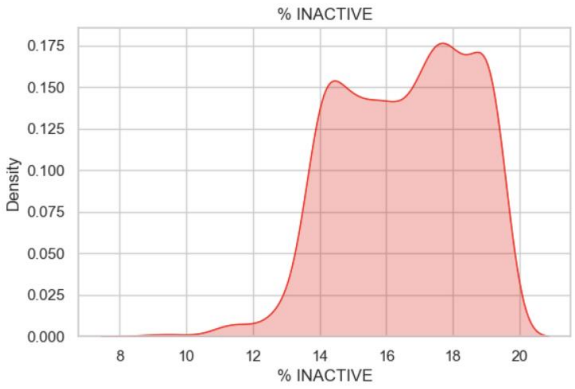
| | cols | Box_Cox_lamdas |
|---|---|---|
| 0 | % INACTIVE | 2.288737 |

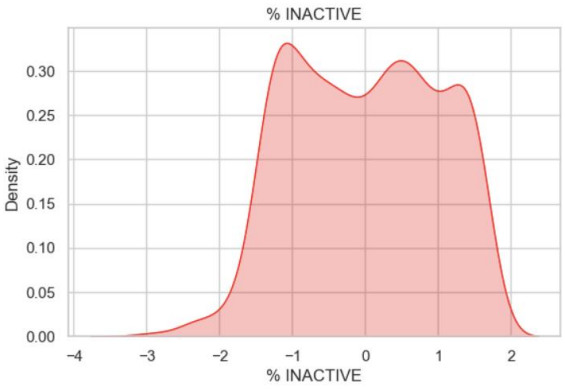
```
df3_transformed.head()
```

Out[69]:

| | % INACTIVE |
|---|---|
| 0 | 0.174055 |
| 1 | 1.537728 |
| 2 | 0.065378 |
| 3 | 0.065378 |
| 4 | 1.474027 |

```
plt.figure(figsize=(14,4))
plt.subplot(121)
sns.kdeplot(df3["% INACTIVE"],shade=True, color='red')
plt.title("% INACTIVE")

plt.subplot(122)
sns.kdeplot(df3_transformed['% INACTIVE'],shade=True, color='red')
plt.title("% INACTIVE")

plt.show()
```

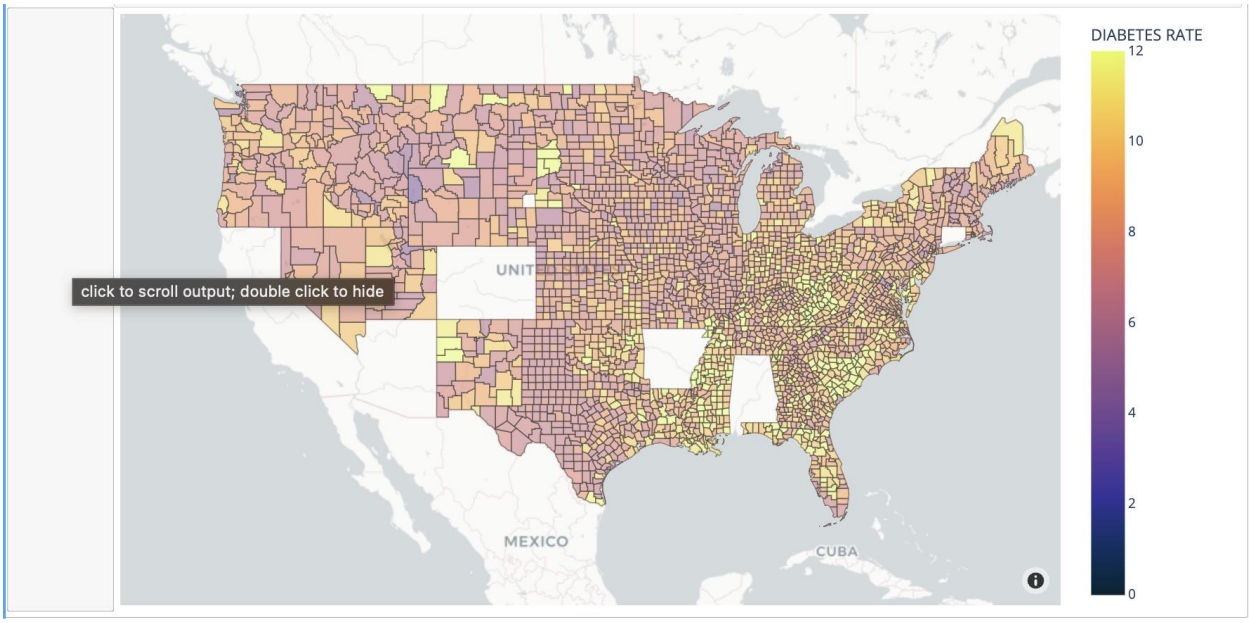



```
from urllib.request import urlopen
import json
with
urlopen('https://raw.githubusercontent.com/plotly/datasets/master/geojson
-counties-fips.json') as response:
    counties = json.load(response)
```

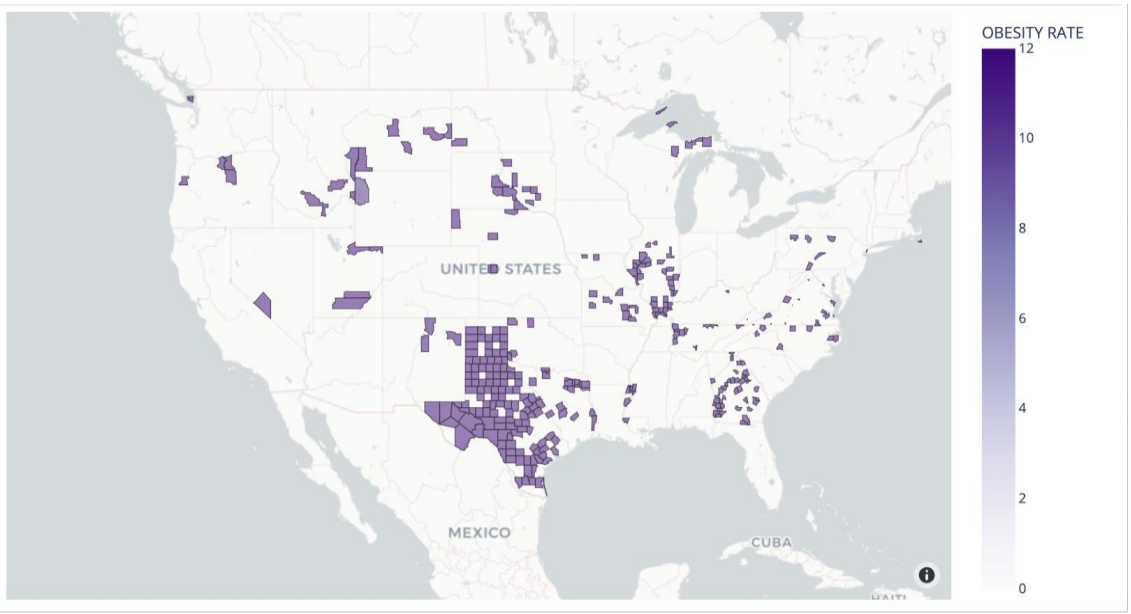
## VISUALIZING DIABETIC RATE ON UNITED STATES MAP

```
with
urlopen('https://raw.githubusercontent.com/plotly/datasets/master/geojson
-counties-fips.json') as response:
    counties = json.load(response)

fig = px.choropleth_mapbox(df1, geojson=counties, locations='FIPS',
color='% DIABETIC',
                           color_continuous_scale="thermal",
                           range_color=(0, 12),
                           mapbox_style="carto-positron",
                           zoom=3, center = {"lat": 37.0902, "lon": -
95.7129},
                           opacity=0.5,
                           labels={'% DIABETIC':'DIABETES RATE',
'COUNTY':'COUNTY'}
                          )
fig.update_layout(margin={"r":0,"t":0,"l":0,"b":0})
fig.show()
```



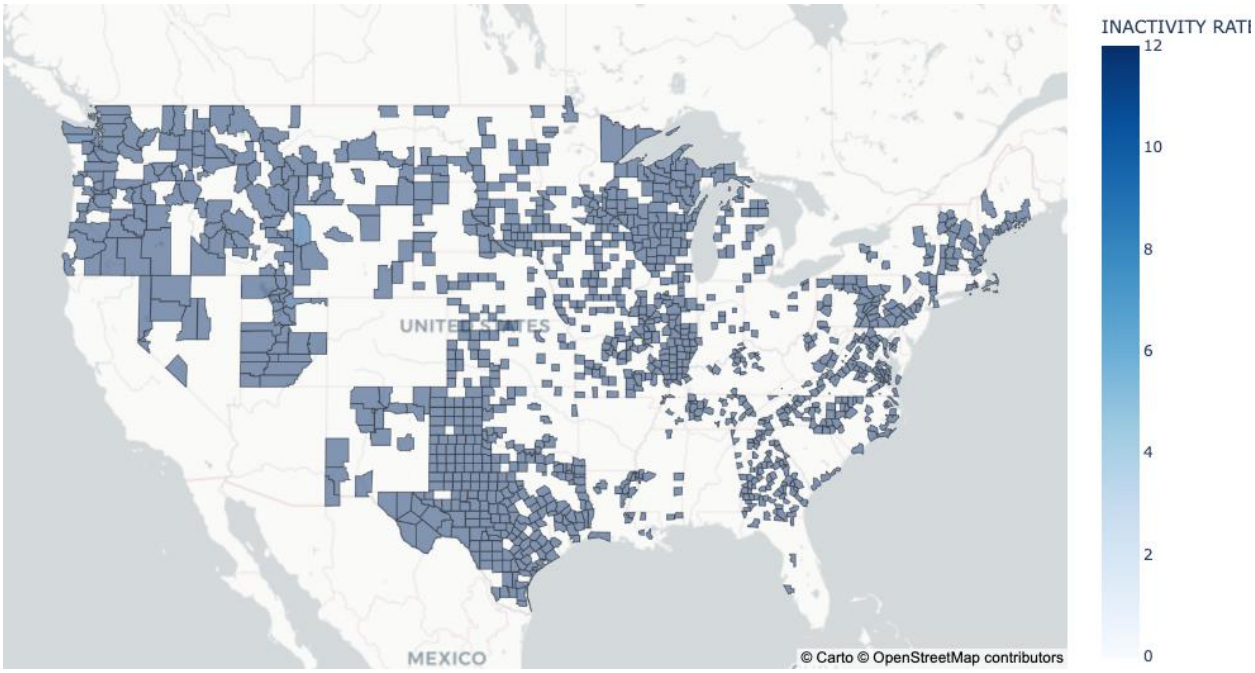

# VISUALIZING OBSEITY RATE ON UNITED STATES MAP

```
with
urlopen('https://raw.githubusercontent.com/plotly/datasets/master/geojson
-counties-fips.json') as response:
    counties = json.load(response)

fig = px.choropleth_mapbox(df2, geojson=counties, locations='FIPS',
color='% OBESE',
                           color_continuous_scale="purples",
                           range_color=(0, 12),
                           mapbox_style="carto-positron",
                           zoom=3, center = {"lat": 37.0902, "lon": -
95.7129},
                           opacity=0.5,
                           labels={'% OBESE':'OBESITY RATE'}
                          )
fig.update_layout(margin={"r":0,"t":0,"l":0,"b":0})
fig.show()
```




## VISUALIZING INACTIVITY RATE ON UNITED STATES MAP

```
with
urlopen('https://raw.githubusercontent.com/plotly/datasets/maste
r/geojson-counties-fips.json') as response:
    counties = json.load(response)

fig = px.choropleth_mapbox(df3, geojson=counties,
locations='FIPDS', color='% INACTIVE',
                           color_continuous_scale="blues",
                           range_color=(0, 12),
                           mapbox_style="carto-positron",
                           zoom=3, center = {"lat": 37.0902,
"lon": -95.7129},
                           opacity=0.5,
                           labels={'% INACTIVE':'INACTIVITY
RATE'}
                          )
fig.update_layout(margin={"r":0,"t":0,"l":0,"b":0})
fig.show()
```




```
df1.columns
```

```
Out[75]: Index(['YEAR', 'FIPS', 'COUNTY', 'STATEW', '% DIABETIC'], dtype='object')
```

# FEATURE EXTRACTION

## Unique FIPS Code present in each dataframe

```
% Diabetic
df1['FIPS'].nunique()
```

```
Out[76]: 3142
```

```
% Obesity
df2['FIPS'].nunique()
```

```
Out[77]: 363
```

```
% Inactive
df3['FIPDS'].nunique()
```

```
Out[78]: 1370
```

```
df3.rename(columns={'FIPDS':'FIPS'}, inplace=True)
```

# Let's merge all this dataset

```
features = pd.merge(df1,df3, how='outer', on='FIPS')
features = pd.merge(features, df2, how='outer', on='FIPS')
```

```
features = features.drop(['YEAR_y','COUNTY_y', 'STATE_x',
'YEAR', 'COUNTY', 'STATE_y',], axis=1)
features.rename(columns={'FIPDS':'FIPS', 'COUNTY_x': 'COUNTY',
'STATEW':'STATE', 'YEAR_x':'YEAR'}, inplace=True)
features['COUNTY_F'] = (features['COUNTY'] +
features['STATE']).factorize()[0]
features['STATE_F'] = features['STATE'].factorize()[0]
features['% DIABETIC'] = df1_transformed['% DIABETIC']
features['% OBESE'] = df2_transformed['% OBESE']
features['% INACTIVE'] = df3_transformed['% INACTIVE']
features['STATE_CODE'] = features['FIPS'] // 1000
features['COUNTY_CODE'] = features['FIPS'] % 1000
features.isnull().sum()
```

Out[86]:
```
YEAR             0
FIPS             0
COUNTY           0
STATE            0
% DIABETIC       0
% INACTIVE    1772
% OBESE       2779
COUNTY_F         0
STATE_F          0
STATE_CODE       0
COUNTY_CODE      0
dtype: int64
```

# Function to impute missing values

```
def imputing_missing_column(dataframe):
    for column in dataframe.columns:
        if dataframe[column].isnull().sum() > 0:
            mean = dataframe[column].mean()
            std = dataframe[column].std()
            random_values = np.random.normal(loc=mean,
scale=std, size=dataframe[column].isnull().sum())
            dataframe[column] =
dataframe[column].fillna(pd.Series(random_values,index=dataframe
[column][dataframe[column].isnull()].index))
imputing_missing_column(features)
```

```
features['% INACTIVE'] = features['% INACTIVE'].fillna(features['% INACTIVE'].mean()) features['% OBESE'] = features['% OBESE'].fillna(features['%
OBESE'].mean())
```

```
features.isnull().sum()
```

Out[89]:
```
YEAR           0
FIPS           0
COUNTY         0
STATE          0
% DIABETIC     0
% INACTIVE     0
% OBESE        0
COUNTY_F       0
STATE_F        0
STATE_CODE     0
COUNTY_CODE    0
dtype: int64
```

```
features.columns
```

Out[90]:
```
Index(['YEAR', 'FIPS', 'COUNTY', 'STATE', '% DIABETIC', '% INACTIVE',
       '% OBESE', 'COUNTY_F', 'STATE_F', 'STATE_CODE', 'COUNTY_CODE'],
      dtype='object')
```

```
#Features we are using for Linear regression Analysis
```

```
X = features[['YEAR', 'FIPS', '% OBESE','% INACTIVE'
,'COUNTY_F', 'STATE_F', 'STATE_CODE', 'COUNTY_CODE']]
y = features['% DIABETIC']

#Split into Training set and Test set
x_train, x_test, y_train, y_test = train_test_split(X, y,
test_size = 0.3)

#linear regression model
model = LinearRegression()

#Fit the model
model.fit(x_train, y_train)
```

Out[91]:
```
LinearRegression()
```

```
print(model.coef_)
```

```
[ 0.00000000e+00 -2.47476643e-04  5.37822263e-03 -7.05533452e-03
 -6.51676761e-04  2.98003490e-01 -2.43291501e-07 -4.18514235e-06]
```

```
print(model.intercept_)
```

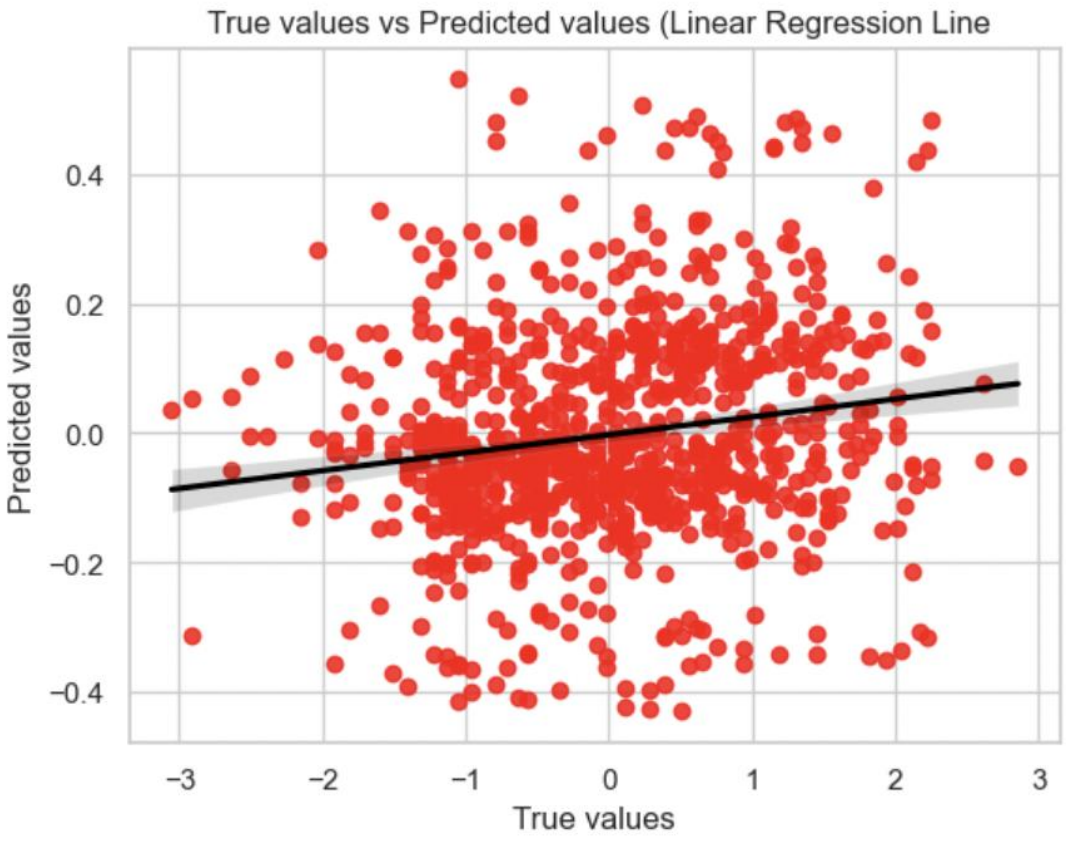
```
0.7230919418351535
```

```
pd.DataFrame(model.coef_, X.columns, columns=['Coeff'])
```

Out[94]:

| | Coeff |
|---|---|
| YEAR | 0.000000e+00 |
| FIPS | -2.474766e-04 |
| % OBESE | 5.378223e-03 |
| % INACTIVE | -7.055335e-03 |
| COUNTY_F | -6.516768e-04 |
| STATE_F | 2.980035e-01 |
| STATE_CODE | -2.432915e-07 |
| COUNTY_CODE | -4.185142e-06 |

```
predictions = model.predict(x_test)
sns.regplot(y_test, predictions, scatter_kws={"color": "red"},
line_kws={"color": "black"})
plt.title(" True values vs Predicted values (Linear Regression
Line")
plt.xlabel('True values')
plt.ylabel('Predicted values')
plt.show()
```




```
from sklearn import metrics
metrics.mean_absolute_error(y_test, predictions)

#MSE
print("The mean squared error value for our model:
",np.round(metrics.mean_squared_error(y_test, predictions),4))

#RMSE
print("The root mean squared error value for our model:
",np.round(np.sqrt(metrics.mean_squared_error(y_test,
predictions)),4))
```

```
The mean squared error value for our model:  0.9385
The root mean squared error value for our model:  0.9688
```