FINAL PROJECT REPORT

By - Ayush Bhatia | Aakash Ahuja

INTRODUCTION:

Productivity is essentially the efficiency in which a company or economy can transform resources into goods, potentially creating more from less. Productivity is defined as the per capita income per hour. Essentially, higher productivity will boost the GDP of a country and eventually contributes towards the economic growth of the country. There could be many monetary factors that could influence the productivity of a nation and there might be a direct relationship between productivity and these factors. Although, we believe that the employees' state of mind and environmental factors might also have an impact on the productivity of the nation as a whole. Additionally, in the current COVID-19 scenario, there might be a significant change in employee's contribution towards productivity. So, we wished to analyze these trends by considering some of these measures from broad categories like Environment, Health, Society, etc.

In this project, we wish to answer the following questions:

1. Is there any direct relationship between the productivity of different states and different factors like Air Pollution, Sleep, Drinking Alcohol, Smoking, Sleep deprivation etc.
2. Given that there is a relationship between productivity and other factors, how can we predict the productivity of a given state in the United States ?

DATA DESCRIPTION:

The data was gathered from different online sources as the values of the measures/factors we considered were not readily available in a single website. Here's a website where we got the data for different measures [https://www.americashealthrankings.org/health-topics]. We collected the productivity values of all the states in the United States from the years 2007 to 2017 and the corresponding measure values were also extracted from different sources. The following are the list of measures and the description of what their values signify:

Air Pollution Value: Average exposure of the general public to particulate matter of 2.5 microns or less (PM2.5) measured in micrograms per cubic meter (3-year estimate)

Binge Drinking Value: Percentage of adults who reported either binge drinking (having four or more [women] or five or more [men] drinks on one occasion in the past 30 days)

Diabetes Value: Percentage of adults who reported being told by a health professional that they have diabetes (excluding prediabetes and gestational diabetes)

High Blood Pressure Value: Percentage of adults who reported being told by a health professional that they have High Blood Pressure

High Cholestrol Value: Percentage of adults who reported being told by a health professional that they have High Cholestrol

Obesity Value: Percentage of adults with a body mass index of 30.0 or higher based on reported height and weight

Physical Inactivity Value: Percentage of adults who reported doing no physical activity or exercise other than their regular job in the past 30 days

Smoking Value: Percentage of adults who are smokers (reported smoking at least 100 cigarettes in their lifetime and currently smoke daily or some days)

Violent Crime Value: Number of murders, rapes, robberies and aggravated assaults per 100,000 population

Frequent Mental Distress Value: Percentage of adults who reported their mental health was not good 14 or more days in the past 30 days

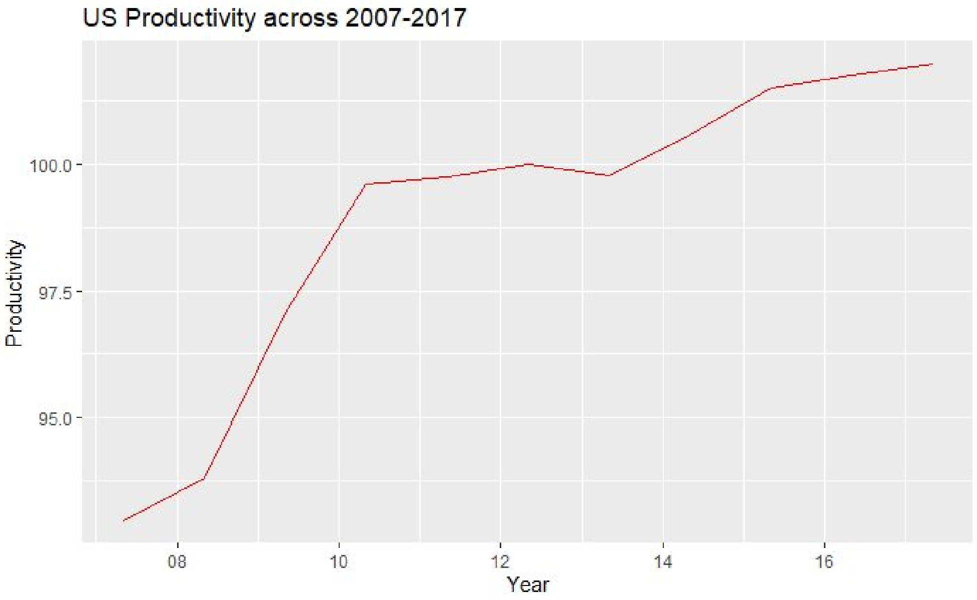
Insufficient Sleep Value: Percentage of adults who reported sleeping less than seven hours in a 24-hour period on average

Water Fluoridation Value: Percentage of population served by community water systems that receive fluoridated water

MODELING AND ANALYSIS:

Initially, we plotted the general trend of productivity for the entire United States across all the years:

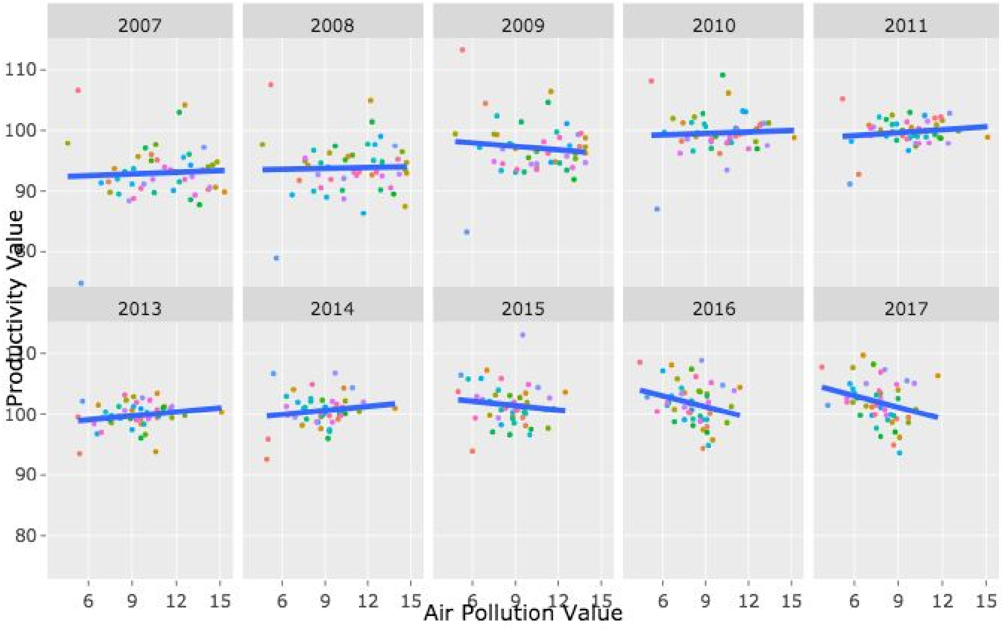
PLOT 1:



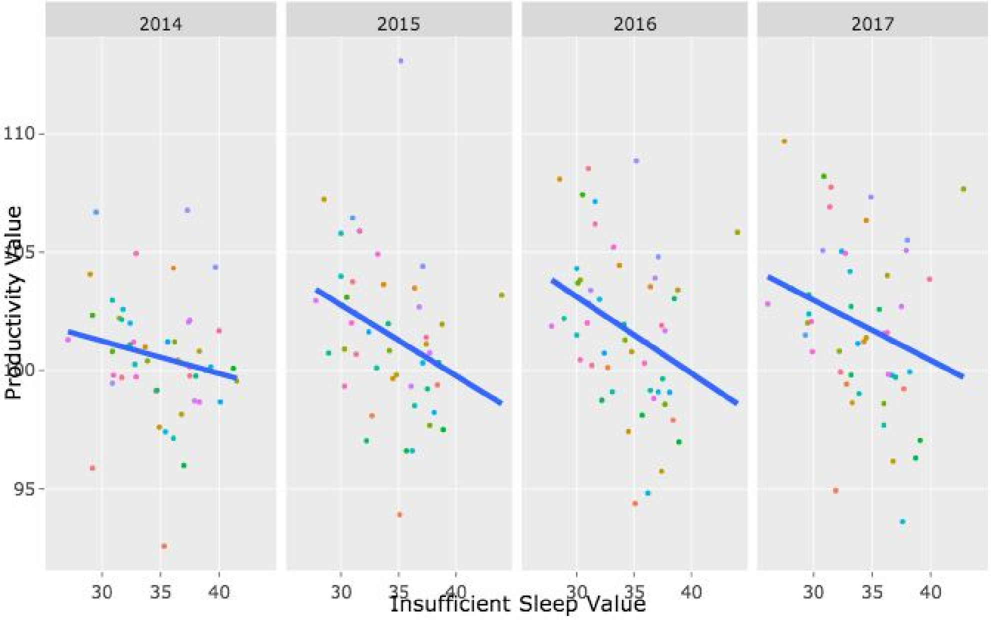
The given plot shows a general trend of increase in productivity in the US across the years.

Now, we need to determine the relationship between productivity and the measures for every year. To perform this, we plot a graph between the productivity and the different measure values faceted by each year. The plots for some of the measures looks as follows:

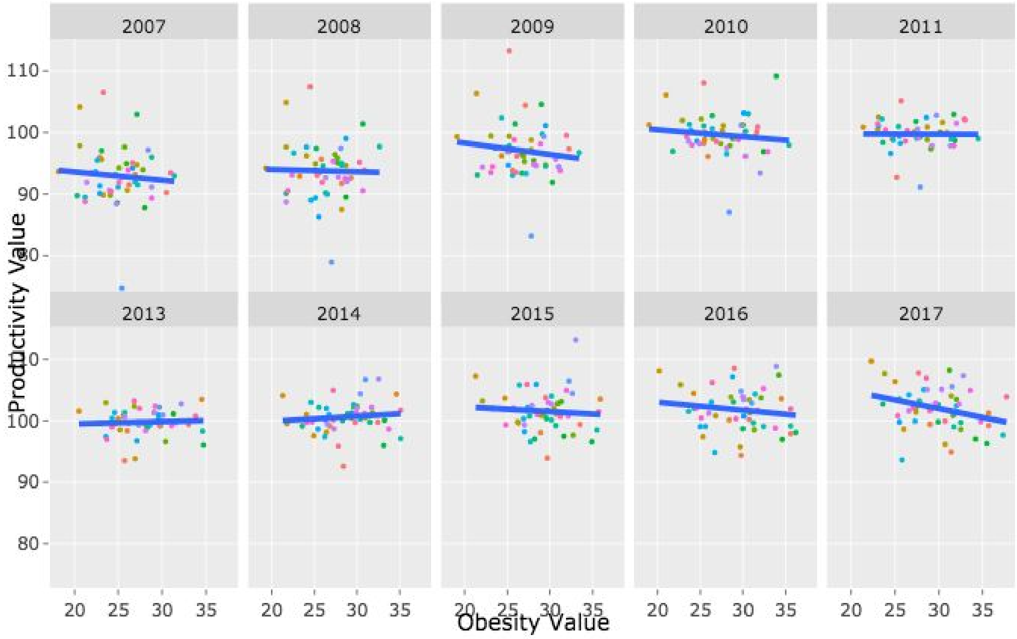
PLOT 2.1



PLOT 2.2



PLOT 2.3



The other plots have been embedded in the Appendix(HTML & RMD).

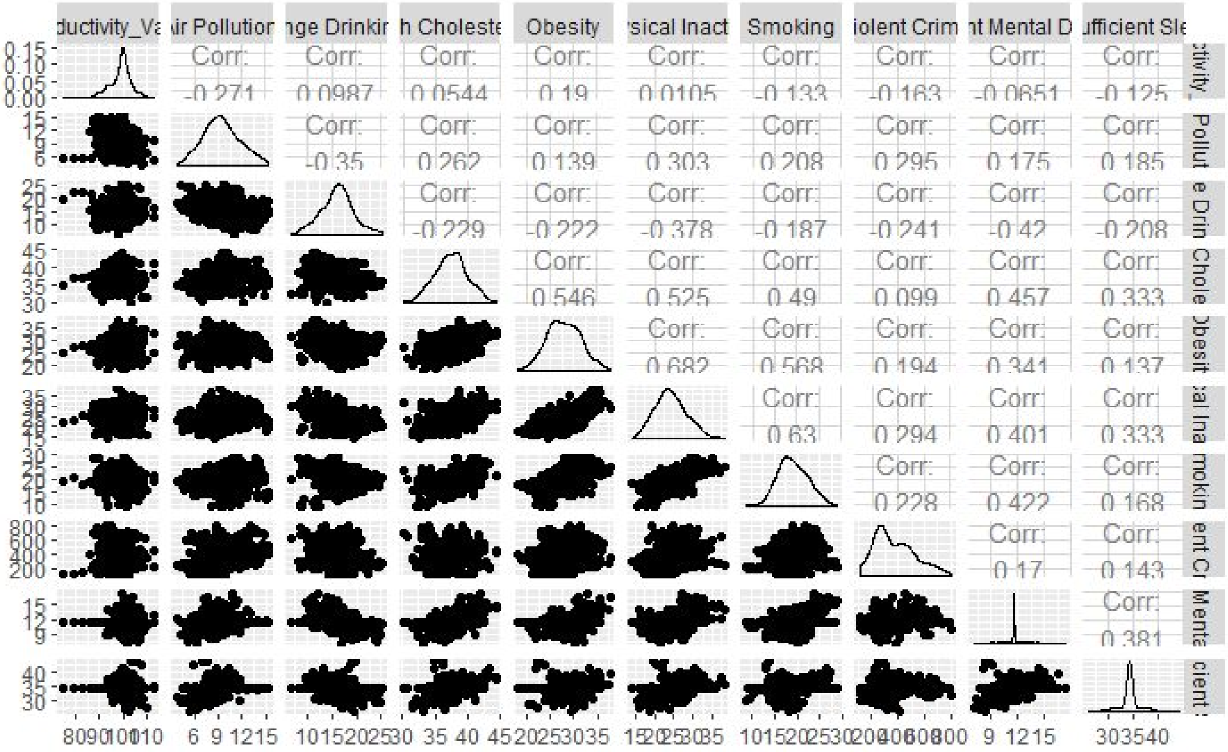
After plotting the relationship graphs for all the 12 measures, we can truncate the number of measures by analyzing the relationship trends. Among the measures, Diabetes and High Blood Pressure don’t show any clear trend and hence, no direct/indirect impacts on Productivity. So, we can remove these two measures from our analysis. Also, there seems to be insufficient data for Water fluoridation values for all the years to conclude any relationship and hence, this variable is also eradicated from our model. It does show an interesting trend though & further in the future we would be interested in considering this variable

Few Insights:

- There can be observed that there are some mixed trends for the measure “Cholesterol Value”. But, we will retain this for our analysis and figure out if the model shows a similar picture.
- The color of the plot represents the 50 states in the US. Some states like Wyoming, North Dakota and Alaska show some weird trends for all the measures.

Checking the correlation of our selected 9 measures(attributes) with respect to the productivity(Target value):

PLOT 3



The correlation graph shows that Air pollution has the highest correlation with Productivity value, with a negative value of -0.271. Obesity shows a high positive correlation with a value of 0.19. Also, Physical Inactivity doesn't seem to have any clear impact on the productivity value. (Weird? We would like to get more data in the future to validate this fact)

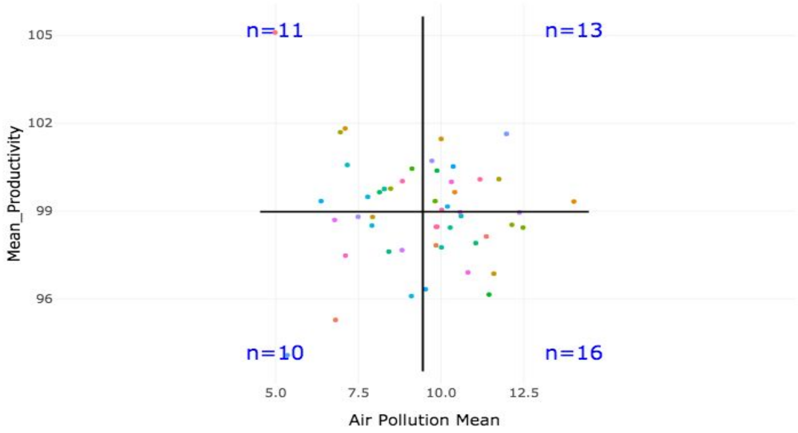
To delve further into our analysis, we check the change in the average measure values for all the states faceted by different Measures. Here, we selected the 9 measures that seem to have some visual relationship with productivity.

Plotting the mean measure values vs the mean productivity values for all the states over all the years.
For the graphs below, we divided a given plot into four quadrants by using the mean values of productivity and the corresponding mean of measure value.
NOTE: Lower Measure Value is better. As the Measure Value increases that means that that measure is getting worse for a particular state

Following is the interpretation of the four quadrants:

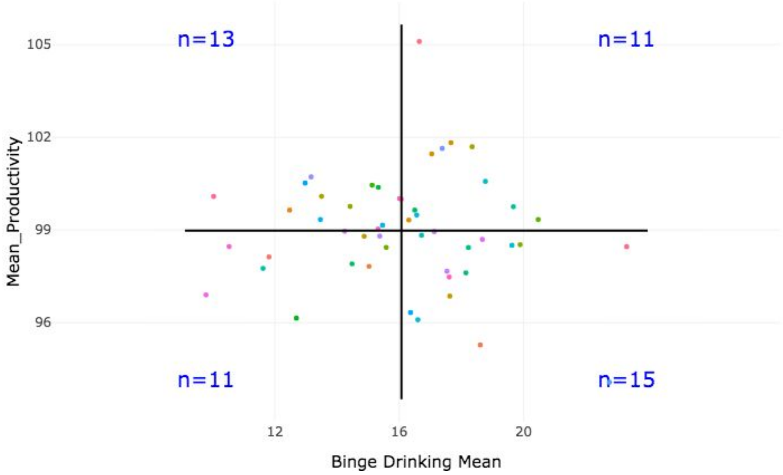
- **Quadrant 1:** High productivity and high measure value. This indicates that productivity is more for a greater measure value.
- **Quadrant 2:** High productivity and low measure value. This indicates that productivity is more for a lower measure value. **DESIRABLE TO SHOW TREND**
- **Quadrant 3:** Low productivity and low measure value. This indicates that productivity is lower for a lower measure value.
- **Quadrant 4:** Low productivity and high measure value. This indicates that productivity is lower for greater measure value. **DESIRABLE TO SHOW TREND**

This interpretation of the quadrant helps us visualize the relationship to a greater extent.



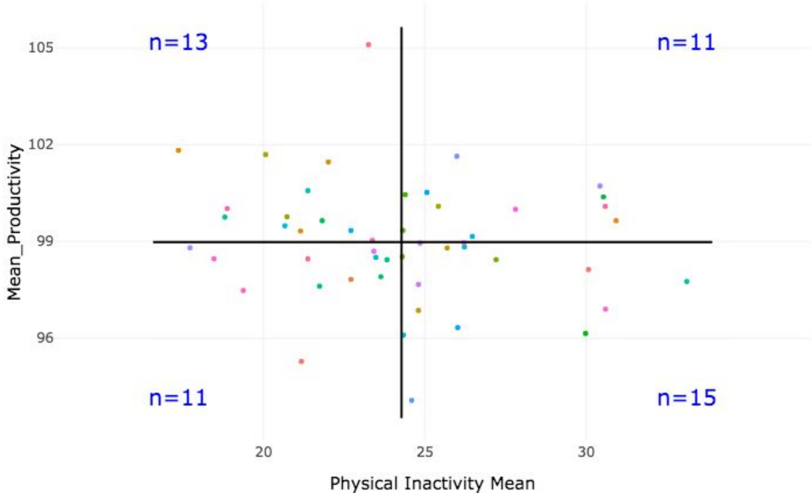
PLOT 4.1

There are more points in the 2nd and 4th quadrant combined. This shows a negative relationship between Air Pollution & Productivity



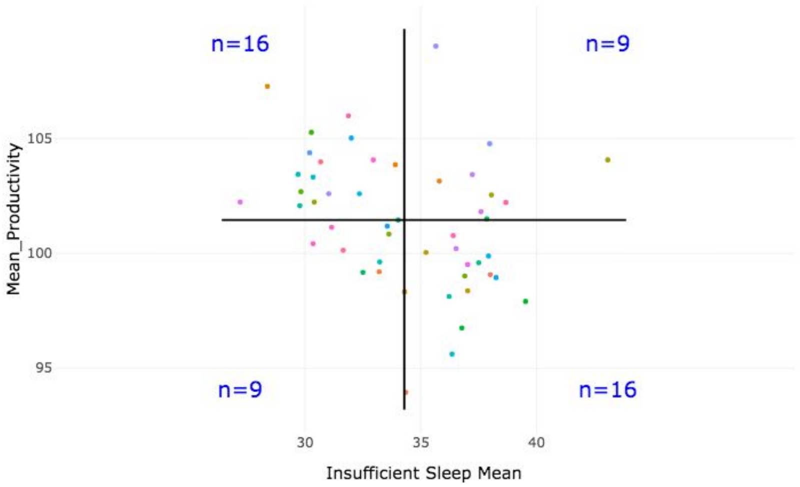
PLOT 4.2

There are more points in the 2nd and 4th quadrant combined. This shows a negative relationship between Binge Drinking & Productivity



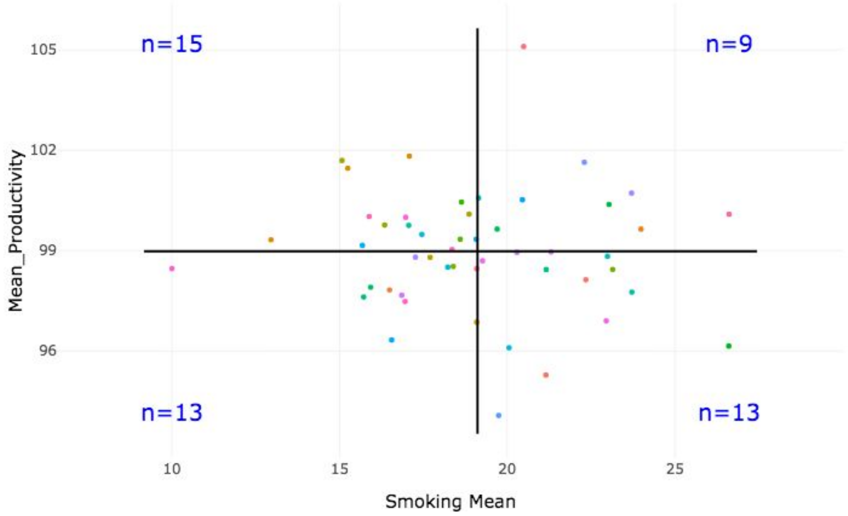
PLOT 4.3

There are more points in the 2nd and 4th quadrant combined. This shows a negative relationship between Physical Inactivity & Productivity



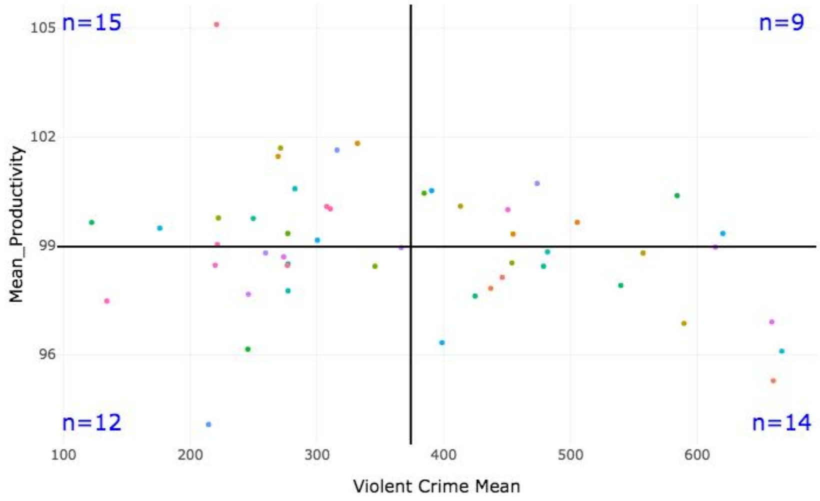
PLOT 4.4

There are more points in the 2nd and 4th quadrant combined. This shows a negative relationship between Insufficient Sleep & Productivity



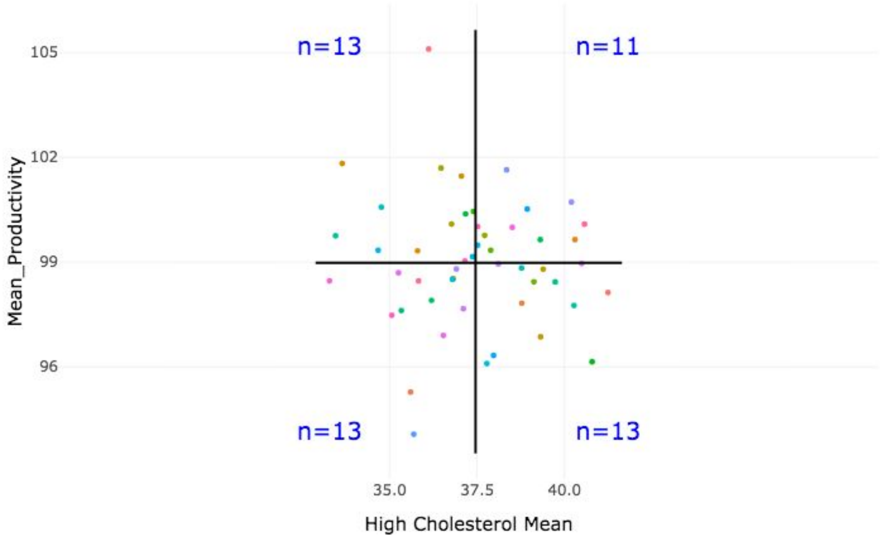
PLOT 4.5

There are more points in the 2nd and 4th quadrant combined. This shows a negative relationship between Smoking & Productivity



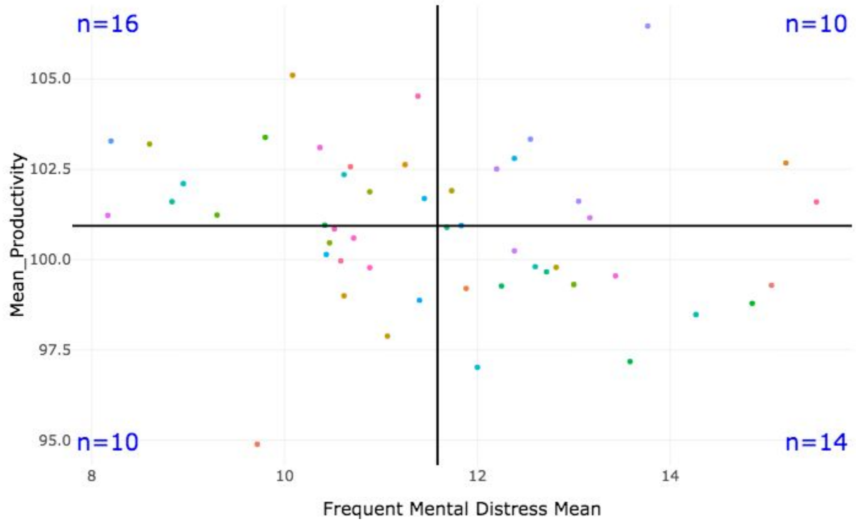
PLOT 4.6

There are more points in the 2nd and 4th quadrant combined. But most of the points are closer to the x and y axis and hence, don’t depict a concrete relationship.



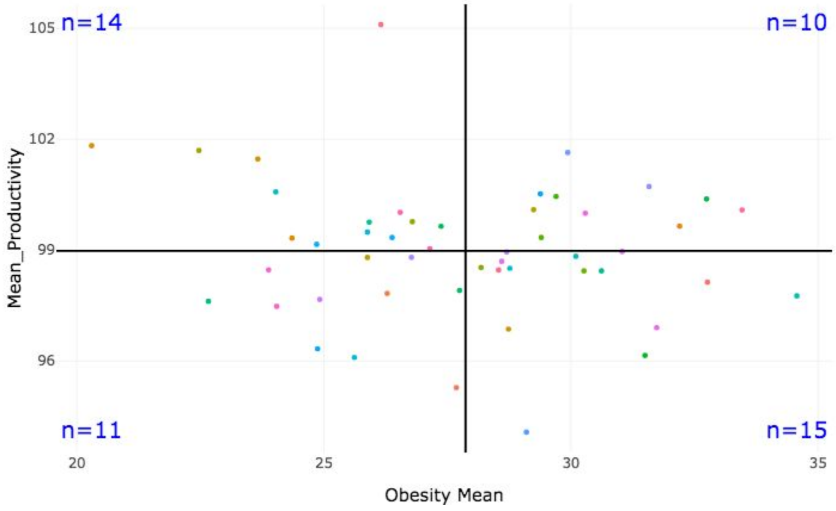
PLOT 4.7

There are almost equal points in all the quadrants and there cannot be seen any relationship between Cholesterol & Productivity



PLOT 4.8

There are more points in the 2nd and 4th quadrant combined. This shows a negative relationship between Frequent Mental Distress & Productivity



PLOT 4.9

There are more points in the 2nd and 4th quadrant combined. But most of the points are closer to the x and y axis and hence, don’t depict a concrete relationship between Obesity and Productivity.

Insights:

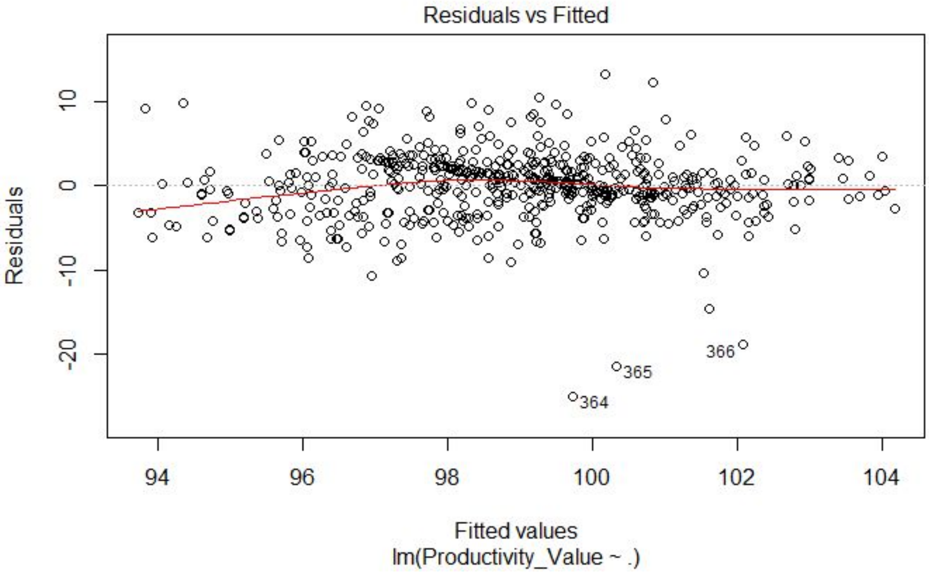
- As we know that a lower measure value(like low Air pollution values) should indicate a positive impact on productivity, it is easy to identify those measures with a larger number of points in the 2nd and 4th quadrant.
- This is due to the fact that more points in the 2nd and 4th quadrant denotes a negative relationship between measure and productivity value. Whereas, more points in the 1st and 3rd quadrant indicates a positive relationship between measure and productivity value.
- It is also noteworthy that more number of points concentrated around the origin indicates no direct/indirect relationship for that particular measure.

Modeling:

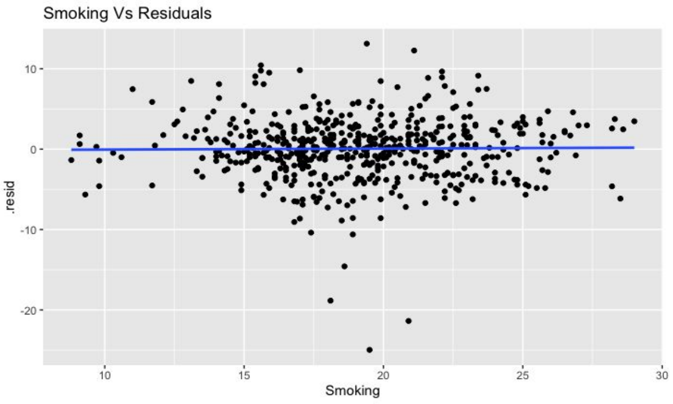
After observing the linear relationship between different measures and the corresponding productivities, it can be represented as a linear model with all the 9 selected measures as attributes and the “Productivity Value” as the target variable or the value to be predicted.

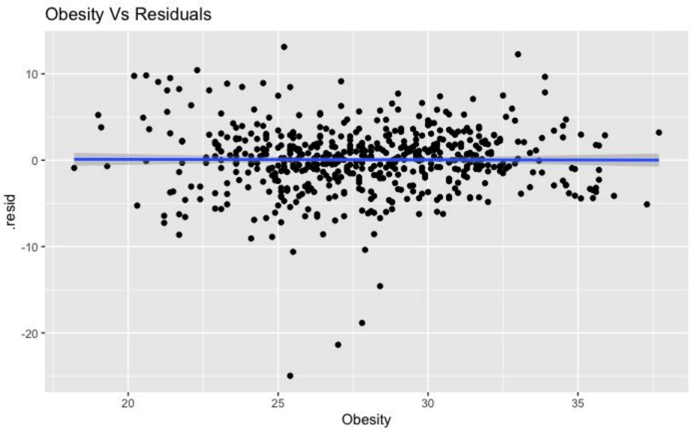
We tried modeling using different models like Linear Model(LM), General Linear Model(GLM) and Robust Linear Model(RLM)
Although, the best results were obtained using RLM.

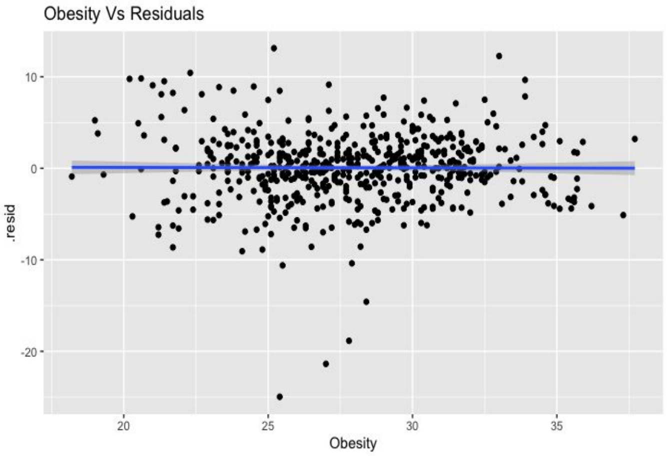
The following is the residual plot for the fitted Robust Linear Model (for all the measures):

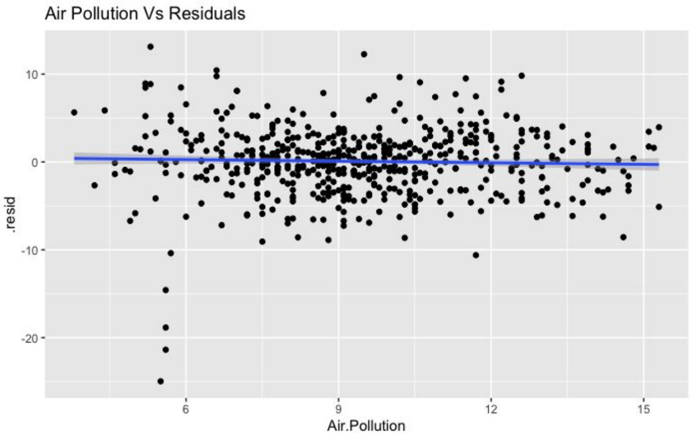


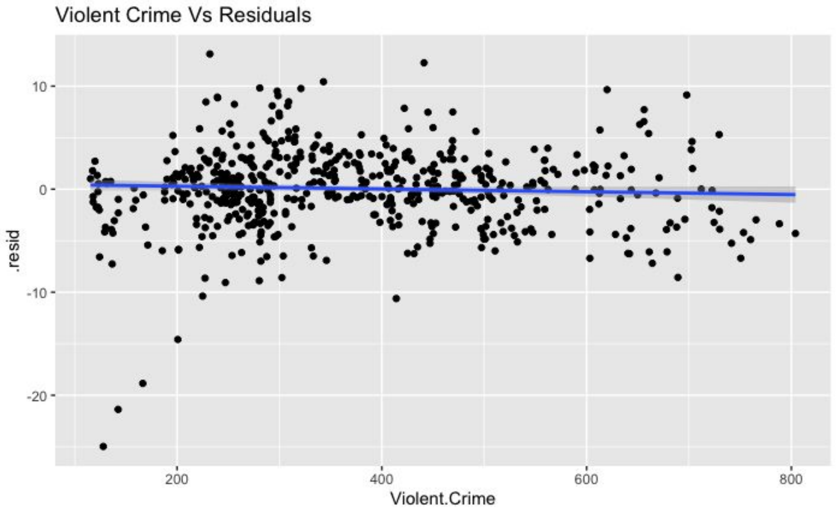
The following are the residual plots for the measures that seem correlated with the productivity:


Smoking Vs Residuals


Obesity Vs Residuals


Obesity Vs Residuals


Air Pollution Vs Residuals


Violent Crime Vs Residuals

CONCLUSION:

The effect of multiple predictors(like Air Pollution, Smoking, Obesity, etc.) were explored. Although, there were a few abnormalities in some of the predictors like Drinking as it was observed that excessive drinking has a positive impact on the productivity of a particular nation. Also, some outliers were observed while plotting the residuals plot as the outliers were not handled while fitting the model.
Some additional attributes or parameters like Screen usage time per day/year could have also brought an interesting insight to our analysis, but it was difficult to extract this data from the online sources.

Future Scopes:

- The visualizations can be enhanced by using the United States map in which we can integrate a dropdown for the “Measures” and a scrollbar to select the year from the given 10 years. This will provide the user to select his/her choice of parameters and will trigger the color of the state based on that particular measure value.
- We can include other attributes like size of family, number of work hours or some other personal factors that might directly/indirectly impact the productivity of a particular region.
- We can perform randomized control experiments on a group of people by adjusting a particular measure and then checking the productivity changes based on that.
- There was a measure “Water Fluoridation levels” for which we had the data for just two years, but there was a relationship portrayed even with that insufficient data. So, we wish to study its impacts on Productivity if we get the data.
- We can extract data from some other sources to validate our findings or maybe improve our analysis.