Analysis of the Relationship Between the Number of Fast-Food Restaurants And Average Obesity Rate by County

Data Preparation

For this project, the analysis of the data was performed using R Studio, but the cleaning of the data was done using a combination of Excel and R Studio. Many of the larger data preparation steps, such as removing unwanted columns or renaming columns, was done in Excel as this was more efficient for these datasets. Once the data had been prepared and formatted with Excel, it was imported into R Studio to be further cleaned, sorted, manipulated, and analyzed.

This project utilized multiple datasets for the analysis. One dataset provided location information for fast food restaurants such as city, state, and zip code, and was imported into R Studio as *df_ff*. Another dataset included obesity information by state and county; this was imported as *df_obesity*. Since these two datasets did not share county, city, or zip code information, a third dataset was required that included geographical information for states such as state, city, zip code, and county, imported as *dfUS*.

For these datasets to be used for an analysis, preprocessing needed to occur which included cleaning, sorting, and establishing consistency among data types and formatting for all city, state, and county names. Due to some datasets having state names listed out and others listed as abbreviations, state names that were listed out were converted to abbreviations for simplicity. Additionally, one dataset listed county names with "County" after the name of the county (i.e., "Johnson County"). This had to be removed to provide consistency between datasets. Zip code information was also inconsistent among datasets and was converted to character data types to provide consistency. The character data type was chosen due to the potential for zip codes to be hyphenated, which would cause issues if zip codes were set as integers or other numeric data type.

Another aspect of the data preparation was calculating the average obesity rate using the given male and female averages. This was average was performed in R Studio and the *avg_obesity* column was added to *df_obesity*. All counties without obesity or restaurant information were provided with NA values to fill the blanks and removed.

Once each dataset was cleaned and prepared, the datasets could be combined to one dataset that would associate appropriate county information with the number of fast-food restaurants and obesity rate for each county and state. The first step in doing this involved combining the *df_ff* and *dfUS* datasets so that the fast-food restaurant locations could be associated with a particular county. This operation was done using R Studio by performing an inner join of the *df_ff* and *dfUS* datasets which matched the fast-food restaurants to the correct county using correlated state, city, and zip code information. In doing this, a new dataset called *df* was created and would become the main dataset for the analysis.

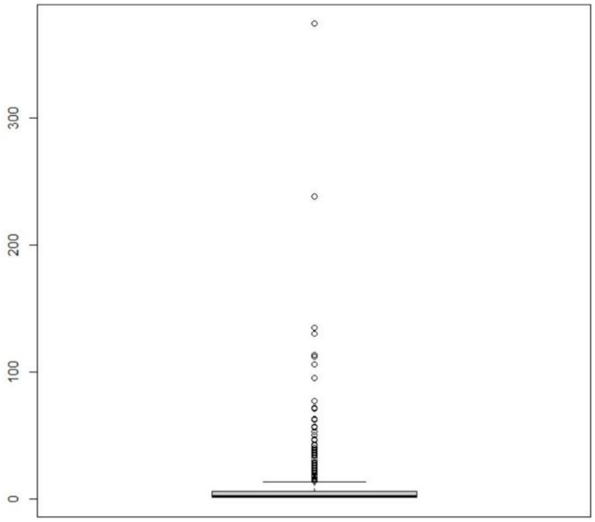
Before performing the analysis, it was preferable to merge all the data into one combined dataset to ensure all data was properly matched with the correct state and county. This meant that the average obesity rate needed to be included into *df*. However, the average obesity first needed to be calculated using the values for both female and male obesity rates provided by the *df_obesity* dataset. Once calculated, these values were merged into the *df* dataset using corresponding state and county information.

Analysis

With all data properly cleaned, ordered, and combined into a single dataset, statistical analysis could be more easily performed. However, prior to analysis, an alternative and null hypothesis must be declared as well as the significance value. For this analysis, it was determined that the significance value would be 0.05 while the alternative hypothesis will state that there is a correlation between the number of fast-food restaurants and the average rate of obesity for a given county. The null hypothesis for the analysis will state that there is no correlation between the two variables.

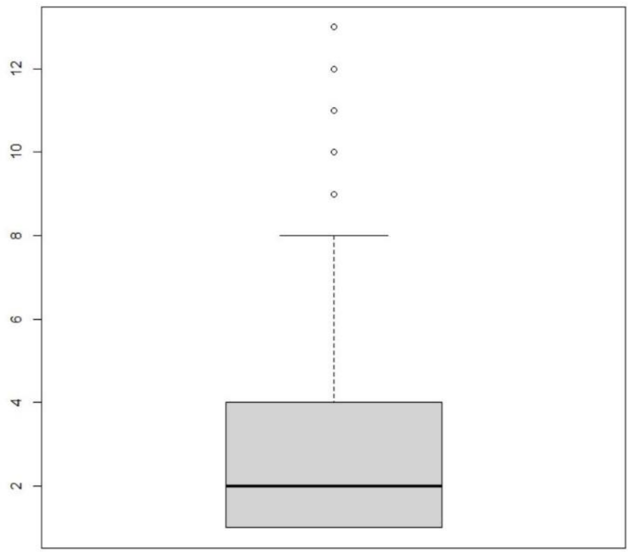
At the beginning of the analysis, a box plot was created for the number of fast-food restaurants to determine what outliers may exist within the data. From this plot, it was discovered that several outliers existed in the dataset that could potentially impact the analysis. As a result, these outliers were removed from the dataset. It should be noted that the removal of these outliers is not believed to have negatively impacted the analysis. This can be seen below in *Figure 1* and *Figure 2*.

Figure 1.



Note: *Figure 1* shows the boxplot representing the number of restaurants by county.

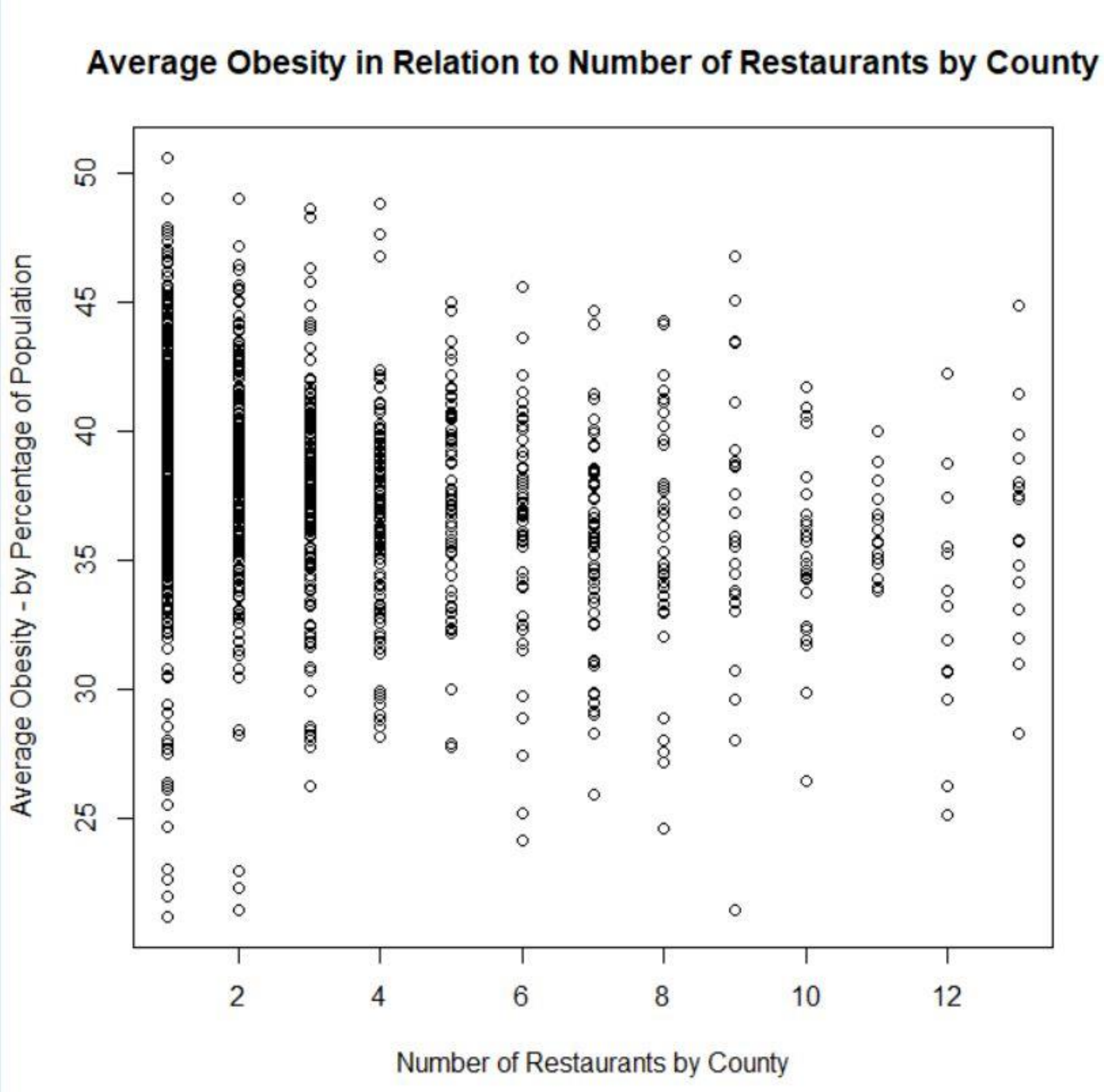
Figure 2.



Note: *Figure 2* shows the boxplot representing the number of restaurants by county with outliers removed. In part, the outliers were removed to allow for a more focused analysis and graphical representations of the data and the relationships between variables.

Once outliers were removed from the dataset, a proper correlation analysis could be performed. The first step in performing this analysis was to generate a scatter plot for the average obesity rate in relation to the number of fast-food restaurants. The scatter plot can be seen below in *Figure 3.*

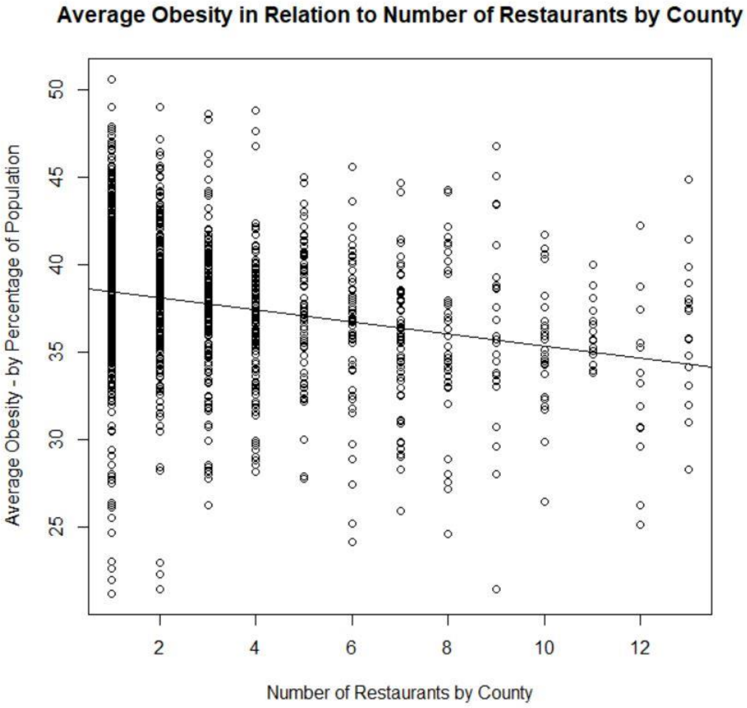
Figure 3.



Note: *Figure 3* shows the scatterplot representing the average obesity rate as a percentage of population in relation to the number of restaurants by county.

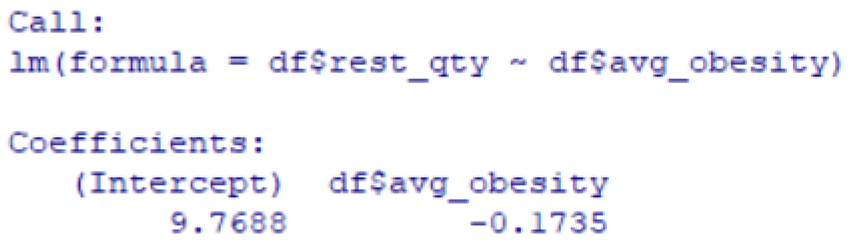
From the above scatter plot, it can be estimated that there is a strong, slightly negative linear relationship between the average rate of obesity and the number of restaurants for a given county. To further confirm this, a regression analysis is performed, and a regression line is created for the dataset and added to the plot. This regression line can be used to visualize the trend of the data as well as predict values for the dataset. This can be seen below in *Figure 4* & *Figure 5*.

Figure 4.



Note: *Figure 4* shows the scatter plot with a regression line for the average obesity rate in relation to the number of restaurants for a given county.

Figure 5.

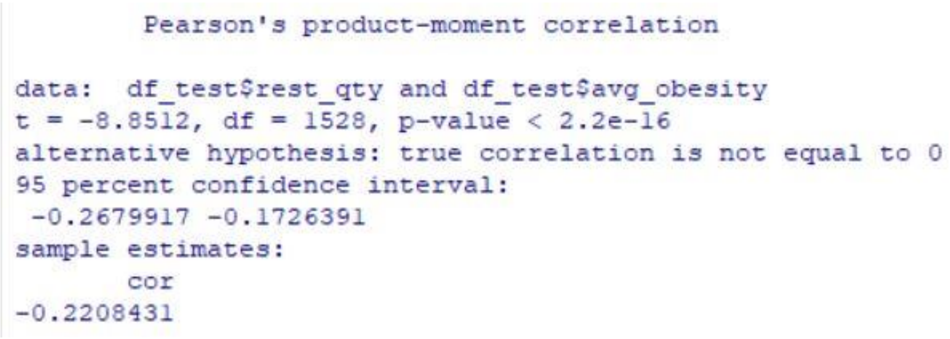


```
Call:
lm(formula = df$rest_qty ~ df$avg_obesity)

Coefficients:
   (Intercept)  df$avg_obesity  
        9.7688         -0.1735  
```

Note: *Figure 5* shows the intercept and slope of the regression line shown in the plot above in *Figure 4.*

As predicted, the regression line has a slightly negative slope indicating a very linear negative relationship between the average obesity rate and number of fast-food restaurants by county. To further determine the strength of this relationship it is necessary to run a correlation analysis to determine the t-value, p-value, and confidence intervals of the dataset. This analysis can be seen below in *Figure 6*.

Figure 6.



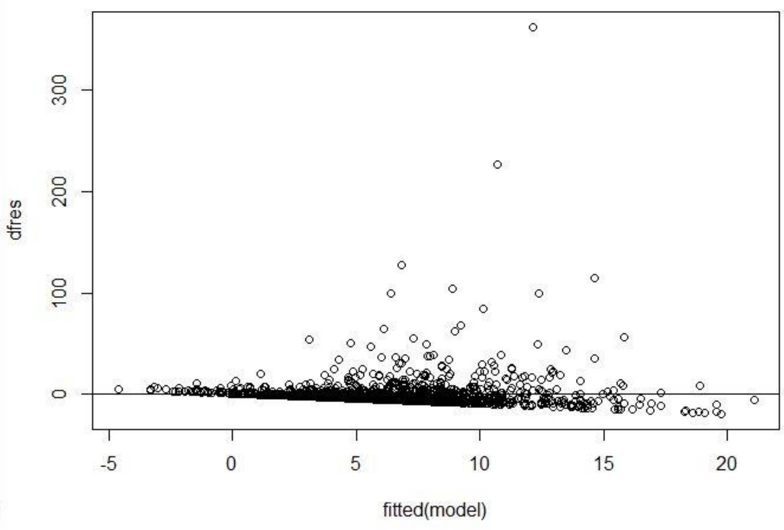
```
	Pearson's product-moment correlation

data:  df_test$rest_qty and df_test$avg_obesity
t = -8.8512, df = 1528, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2679917 -0.1726391
sample estimates:
       cor 
-0.2208431 
```

Note: *Figure 6* shows the correlation analysis between the average obesity rate and the number of fast-food restaurants by county.

As can be seen in *Figure 6*, the p-value for the dataset is 2.2^-16, which is very close to zero and below the initially declared significance level of 0.05. As a result, it can be concluded that there is a statistically significant relationship between the average obesity rate and number of fast-food restaurants. Additionally, the absolute t-value of 8.85 further supports the statistical significance of the relationship between the two variables. When compared to the value of 1.96 provided by the t-value distribution table, the t-value provided by the dataset is greater providing sufficient evidence to reject the null hypothesis.

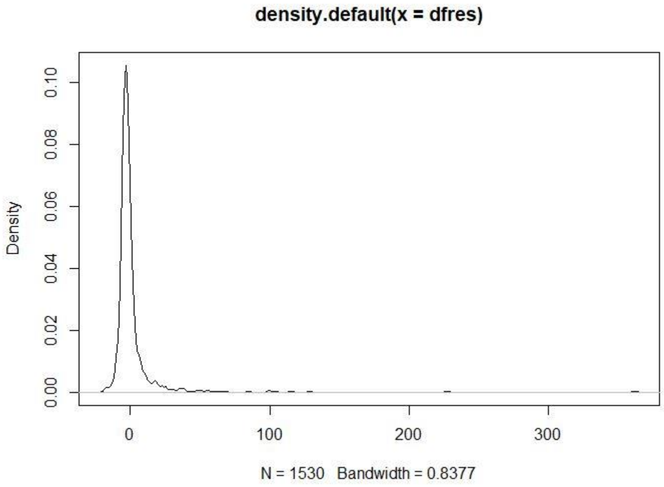
However, while this relationship is statistically significant, and there is sufficient evidence to reject the null hypothesis, evidence provided by the -0.22 correlation coefficient, or R value, suggests that the relationship between the two variables is not very strong, being that the correlation coefficient is closer to zero than -1 or 1. To further test the strength of the correlation between the variables and the regression line, the residuals should be analyzed as well as the R and R^2 values for the dataset. *Figure 7* and *Figure 8* below show the residual plots for the data.

Figure 7.



Note: *Figure 7* shows the residual plot showing the difference between the observed values of the average obesity data and the regression line values.

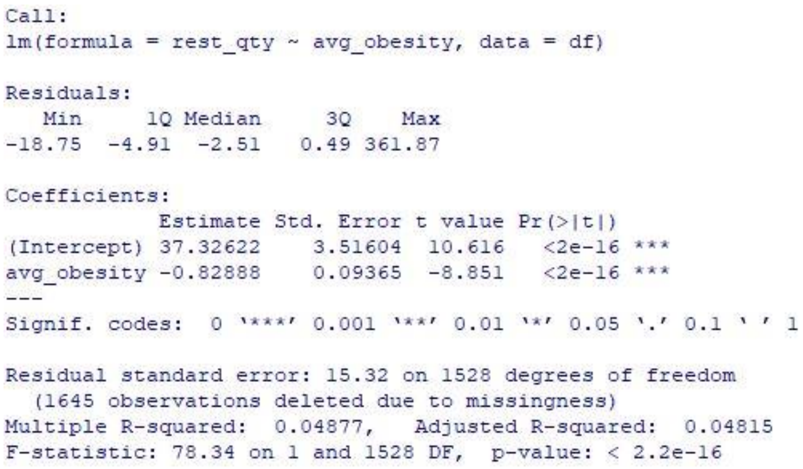
Figure 8.



Note: *Figure 8* shows the density of the residual values.

As can be seen above in *Figure 8*, with the exception of the outlier values, the residual values appear to be fairly normally distributed, indicating that the regression line is appropriate for the model. Additionally, the R and R^2 values for the dataset should be considered as well. In *Figure 9* below, it can be observed that the R^2 value for the dataset is 0.48. This means that 48% of the variance in obesity can be explained by the number of fast-food restaurants within a given county, further supporting the evidence to reject the null hypothesis.

Figure 9.



```
Call:
lm(formula = rest_qty ~ avg_obesity, data = df)

Residuals:
   Min     1Q Median     3Q    Max 
-18.75  -4.91  -2.51   0.49 361.87 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 37.32622    3.51604  10.616   <2e-16 ***
avg_obesity -0.82888    0.09365  -8.851   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.32 on 1528 degrees of freedom
  (1645 observations deleted due to missingness)
Multiple R-squared:  0.04877,	Adjusted R-squared:  0.04815 
F-statistic: 78.34 on 1 and 1528 DF,  p-value: < 2.2e-16
```

Note: *Figure 9* shows the summary statistics for the dataset including the R^2 values.

Based on the above analysis, as well as the summary statistics presented in *Figure 9*, there is sufficient evidence to reject the null hypothesis and conclude that there is a relationship between the average rate of obesity and the number of fast-food restaurants for a given county. However, the relationship appears to be a negative linear relationship, which was contrary to what was expected by the author. Considerations for this analysis include variations in people moving from county to county and state to state, as well as inconsistencies or inaccuracies in restaurant data. Additional considerations include difference in rural and metropolitan areas as well as levels of average physical activity and common transportation methods for various counties, cities, and states.