US County Obesity Rates and Possible Linkages

This project investigates US county obesity rates and possible economic and demographic linkages using 2016 data from the US Census Bureau, Bureau of Economic Analysis, and County Health Rankings & Roadmaps. All data are at the county level.

Technical Requirements

For this analysis, the following additional libraries were used within the Pandas and Jupyter Notebook:

- xlrd (pip install xlrd)
- seaborn (pip install seaborn)
- statsmodels.api (pip install -U statsmodels)

Variables

- US Census Bureau data:
    - SNAP Recipient (%) = (SNAP benefit recipients/population) * 100
- Bureau of Economic Analysis data:
    - Income Per Capita
- County Health Rankings & Roadmaps, Robert Wood Johnson Foundation:
    - Obesity (%)
    - Some College (%)
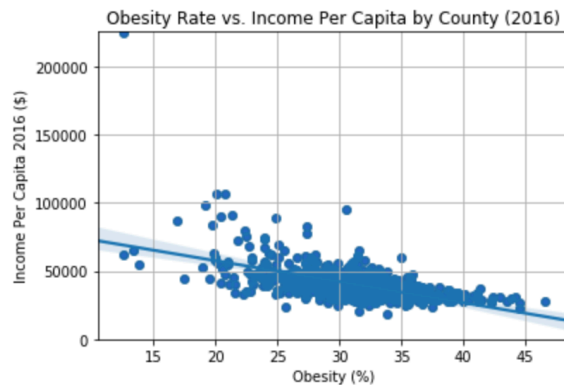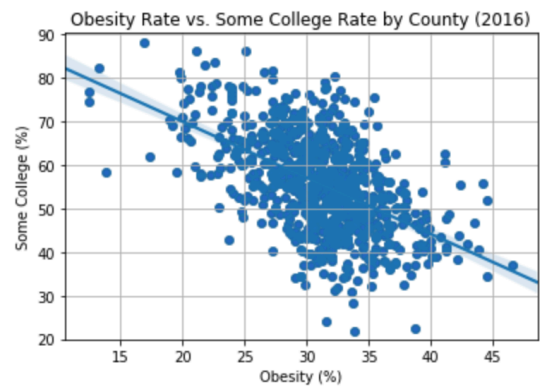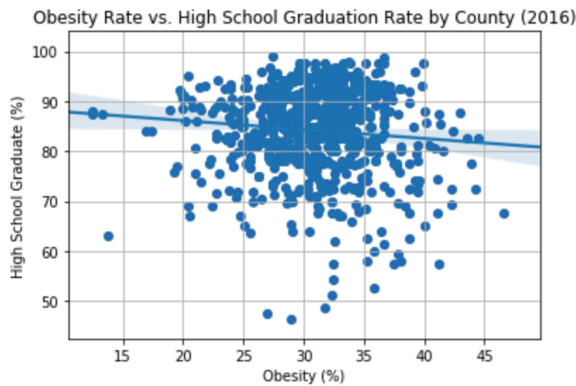   - High School Graduate (%)
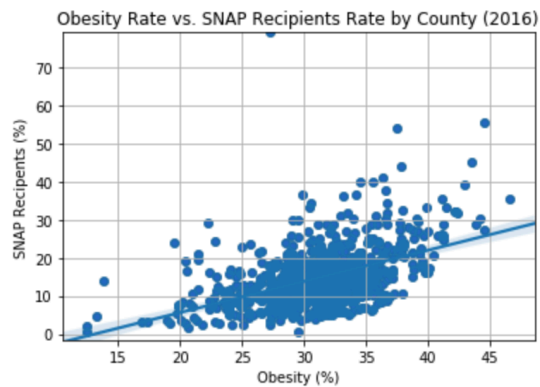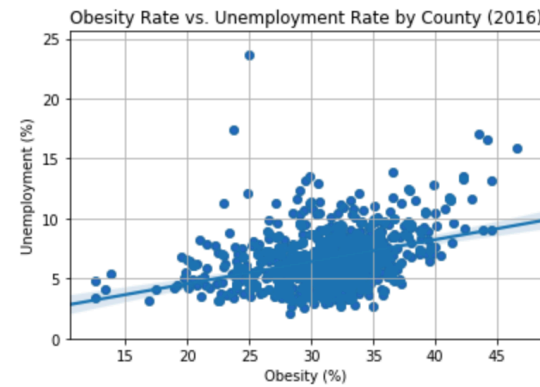   - Unemployment (%)

Hypothesis

Do obesity rates in US counties have strong correlations with demographical and economic characteristics such as SNAP benefits, education levels, income per capita, and unemployment rates?

Analysis

- **Scatter Plots and Trendlines**

The results of scatter plots and trendlines of the relationships of the obesity rates versus the other individual independent variables are as the following. The relationship of the obesity rates versus unemployment rates or SNAP recipient rates is a positive one. The relationship is negative between the obesity rates versus income per capita, some college or high school graduate percent. The high school graduate percent and obesity rate relationship is weaker compared to the other relationships. The relationships form around the range of 20% - 40% obesity rates.

Obesity Rate vs. Unemployment Rate by County (2016)

Obesity Rate vs. SNAP Recipients Rate by County (2016)

Obesity Rate vs. High School Graduation Rate by County (2016)

Obesity Rate vs. Some College Rate by County (2016)

Obesity Rate vs. Income Per Capita by County (2016)

- **Multiple Regression**

The results of the first multiple regression model show multicollinearity issues with positive sign on the coefficient of high school graduate rates. One or more variables are removed at times to resolve multicollinearity. The predictive model of obesity rate for this project has two predictors, SNAP recipient rate and some college percent. The significant F-test and P-value of the final model provides evidence for us to conclude that the obesity rate is significantly determined by SNAP recipient rate and some college percentage.

OLS Regression Results

Dep. Variable:	Obesity (%)	R-squared:	0.343
Model:	OLS	Adj. R-squared:	0.341
Method:	Least Squares	F-statistic:	193.4
Date:	Tue, 03 Sep 2019	Prob (F-statistic):	2.47e-68
Time:	17:47:21	Log-Likelihood:	-2048.8
No. Observations:	745	AIC:	4104.
Df Residuals:	742	BIC:	4117.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	37.4790	1.004	37.333	0.000	35.508	39.450
SNAP Recipents (%)	0.1674	0.021	8.113	0.000	0.127	0.208
Some College (%)	-0.1596	0.014	-11.079	0.000	-0.188	-0.131

Omnibus:	35.645	Durbin-Watson:	2.083
Prob(Omnibus):	0.000	Jarque-Bera (JB):	45.022
Skew:	-0.458	Prob(JB):	1.67e-10
Kurtosis:	3.782	Cond. No.	423.

- **Correlation Matrix**

The results of the correlation matrix support our aforementioned findings and conclude this project. The map reflects the findings using scatter plots and trendlines. It also shows correlations among independent variables which cause multicollinearity; e.g., positive strong correlations between unemployment rate and SNAP recipient rates and positive correlation between some college percent and income per capita.
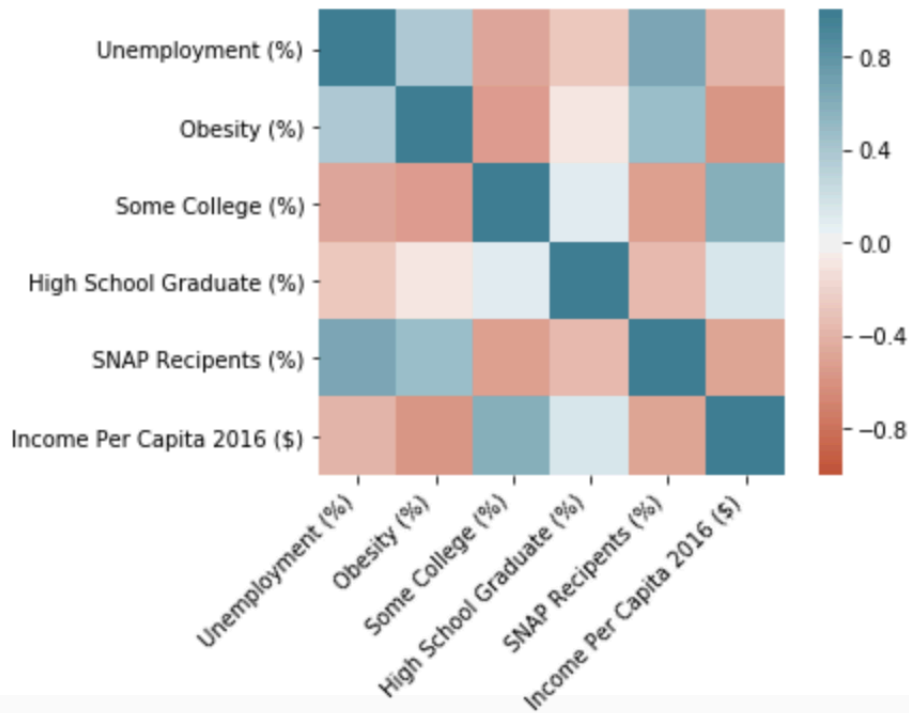
Conclusion

According to the above findings, there is evidence to conclude that US county SNAP benefits, education levels, income per capita, and unemployment rates are possible linkages to US county obesity rates. Some linkages are positive whereas others are negative and/or weak. The implication, however, suggests against using all variables as predictors of the US county obesity rates together in one model since some of these predictors also have strong correlations among themselves. A model which resolves multicollinearity is presented. However, further tests of the model residuals beyond this project are highly recommended prior to using the predictive model. Exploratory factor analysis and a systematic algorithmic method for finding a suitable model would also be helpful.