**Proposal For Data Analysis**

Project Title: Diabetes Prevalence In the United States

Team: The Miner League

Dataticians: Maxwell Carduner

Kalaivani Chandramohan

William Chirciu

Ramkumar Perumal

Charles Saporito

**Data Set Description**

We are using CDC county level diabetes related data to create a model to predict the diabetes prevalence within a county by using other county level data such as obesity prevalence, leisure-time physical inactivity rates, socio-economic indicators such as median age, income, Unemployment rate, education, and Median Household income.

**Variables for Analysis**

The Y-Variable for this analysis will be Diabetes Prevalence Percentage. Our proposed X-Variables (9, all numeric) for prediction of Diabetes Prevalence are as follows; Unemployment Rate, Household Median Income, Median Age, Obesity Prev %, Leisure Time Physical Inactivity Prev, Percent of adults with less than a high school diploma, Percent of adults with a high school diploma only, Percent of adults completing some college or associate's degree, Percent of adults with a bachelor's degree or higher.

**Number of Records in Proposed Sample**

For our proposed sample of Diabetes prevalence there are 3,139 observations for analysis.

**URLs**

Socio-economic county indicators from USDA: <https://www.ers.usda.gov/data-products/county-level-data-sets/>

Diabetes prevalence and lifestyle indicators from CDC:

<https://gis.cdc.gov/grasp/diabetes/DiabetesAtlas.html>

Median Age of Sample Data

<https://datausa.io/map/?level=county&key=age,age_moe,age_rank>

**Problem Statement**

30.3 million people had diabetes in the US according to a 2015 American Diabetes Association (ADA) study. This represents a national prevalence rate of 9.4%. Therefore, we have compiled county level prevalence rates with socio-economic and lifestyle indicators to see if they have a contributory effect to diabetes prevalence.

**Proposed Methodology**

1. Preprocess data (Complete)
   1. Merge datasets using FIPS Code (Zip Code)
   2. Remove observations where not all information is available
2. Examine distributions of all variables
   1. May need to transform

2. Examine scatterplots of all variables in the dataset to:

* 1. Check for linearity (target vs. all other predictors)
     1. If not linear, propose transformations
  2. Check for associations between predictors to see if anything is highly correlated
     1. If so, will need to check VIFs after building the regression model

4. Estimating model parameters:

1. Check the significance of predictors
   1. Remove any predictors with p-values > 0.05
2. Examine VIFs for multicollinearity
3. Examine studentized residuals and normal probability plot
   1. Check Linearity, Constant Variance, Independence, and Normality assumptions
   2. Identify Outliers and Influential Points
4. Re-run using model selection methods.
   1. Pick best model based on r squared, adjusted r squared, CP, F value, RMSE, etc. while keeping in mind as few variables as possible
5. Final Model on full data set

5. Partition data into training and test set

6. Evaluate goodness of fit on test set with final model(s)

**References**

Ezzati, Majid. 2016. Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4·4 million participants. <https://www.thelancet.com/journals/lancet/article/PIIS0140-6736%2816%2900618-8/fulltext> .Accesses April 6,2016.

Alexandra, Virginia. 2016.County-Level Data Reveal Three-Fold Difference in Diabetes Prevalence Across the U.S. <http://www.diabetes.org/newsroom/press-releases/2016/county-level-data-reveal.html> .Access August 23,2016.

Sanjay Basu,Paula Yoffe,Nancy Hills,Robert H.Lustig.2013. The Relationship of Sugar to Population-Level Diabetes Prevalence: An Econometric Analysis of Repeated CrossSectional Data. <https://www.medpagetoday.com/upload/2013/3/1/journal.pone.0057873.pdf> .Accessed Feburary 27,2013.

Statistics About Diabetes. 2018. <http://www.diabetes.org/diabetes-basics/statistics/> Accessed May 13, 2018.