Portfolio Project – My Study of the CDC’s “Chronic Disease Measurers for the 500 Largest American Cities” Project

Max P. Liphart

Course Number (MIS500-1) – Name of Course (Foundations of Data Analytics)

Colorado State University – Global Campus

Eric M. Straw, PhD

July 7, 2019

**Abstract**

Problem – for a subject of my own interest, first how do I collect data for chronic disease measures? I located data for the 500 largest American Cities (CDC 2016)? Next, what has been done – the Robert Wood Johnson Foundation and the Centers for Disease Control and Prevention collected data. Now, after reviewing the data I determined – for the 27 chronic disease measurers, as a percentage of the population, Birmingham measured a mean of 34.4, the United States measured a mean of 31.19 and Los Angeles measured a mean of 29.6 (percent of population who suffered from selected chronic disease measurers).

What does it mean – for the 27 chronic disease measurers, Los Angeles had the lowest mean.

Also, I determined the means for Los Angeles and Birmingham are not equal.

I utilized The Robert Wood Johnson Foundation and CDC Foundation [who] launched the 500 Cities Project in partnership with the Centers for Disease Control and Prevention (CDC 2015).

This project reports city and census tract-level data obtained using small area estimation methods, for 27 chronic disease measures for the 500 largest American cities. The project reports

the number of cities per state ranges from 1 to 121. The cities range in population from 42,417 in Burlington, Vermont to 8,175,133 in New York City, New York. Among these 500 cities, there are approximately 28,000 census tracts (subdivisions), for which data will be provided. The number of tracts per city ranges from 8 to 2,140.The project includes a total population(for the 500 cities) of 103,020,808, which represents 33.4% of the total 2010 United States population of 308,745,538. (CDC 2015)

**Detailed Information About the Project, the Site, and the Dataset Used**

**Project: 500 Cities: Local Data for Better Health, 2016 Release (CDC)**

The Centers for Disease Control and prevention (CDC) provides a complete dataset for the 500 Cities project 2016 release. This dataset includes 2013, 2014 model-based small area estimates for 27 measures of chronic disease related to unhealthy behaviors (5), health outcomes (13), and use of preventive services (9). Data were provided by the Centers for Disease Control and Prevention , Division of Population Health, Epidemiology and Surveillance Branch. (CDC)

The project was funded by the Robert Wood Johnson Foundation (RWJF) in conjunction with the CDC Foundation. It represents a first-of-its kind effort to release information on a large scale for cities and for small areas within those cities.

It includes estimates for the 500 largest US cities and approximately 28,000 census tracts within these cities. These estimates can be used to identify emerging health problems and to inform development and implementation of effective, targeted public health prevention activities. (CDC)

Data sources used to generate these measures include Behavioral Risk Factor Surveillance System (BRFSS) data (2013, 2014), Census Bureau 2010 census population data, and American Community Survey (ACS) 2009-2013, 2010-2014 estimates. (CDC 2016)

**The Site**

The site containsa 500 Cities Data Portal which contains a complete dataset for the 500 Cities project 2016 release. This dataset includes 2013, 2014 model-based small area estimates for 27 measures of chronic disease related to unhealthy behaviors. (CDC 2016)

**Dataset Used**

The dataset contains model-based small area estimates for 27 measures of chronic disease related to unhealthy behaviors retrieved from: (CDC 2016) https://chronicdata.cdc.gov/api/views/6vp6-wxuq/rows.csv?accessType=DOWNLOAD&bom=true&format=true&delimiter=%3B

**Provide the Code with Your Modification Comments**

**Main Dataset (SAS Code)**

\*Combine DATA and PROC to sort and summarize dataset;  
PROC SQL;  
CREATE TABLE WORK.query AS  
SELECT 'Year'n , StateAbbr , StateDesc , CityName , GeographicLevel , DataSource , Category , UniqueID , Measure , Data\_Value\_Unit , DataValueTypeID , Data\_Value\_Type , Data\_Value , Data\_Value\_Footnote\_Symbol , Data\_Value\_Footnote , Low\_Confidence\_Limit , High\_Confidence\_Limit , Population2010 , GeoLocation , CategoryID , MeasureId , CityFIPS , TractFIPS , Short\_Question\_Text , VAR25 FROM WORK.IMPORT;  
\*Execute the previous statements;  
RUN;  
\*Stop execution and exit;  
QUIT;  
  
  
  
  
\*Provides information about dataset, i.e. variables, attributes;  
PROC DATASETS NOLIST NODETAILS;  
\*Lists contents of dataset and print to directory;  
CONTENTS DATA=WORK.query OUT=WORK.details;  
\*Stop execution and exit;  
RUN;  
  
\*prints a listing of values of variables in dataset;  
PROC PRINT DATA=WORK.details;  
\*Execute the previous statements:  
RUN;

**Selected Dataset (Birmingham and Los Angeles)**

\* Generated Code (IMPORT) \*/  
/\* Source File: ExportCSV.csv \*/  
/\* Source Path: /folders/myfolders \*/  
/\* Code generated on: 7/6/19, 8:12 AM \*/  
  
%web\_drop\_table(WORK.IMPORT);  
  
  
/\*FIlename is where data can be found, the filepath\*/   
FILENAME REFFILE '/folders/myfolders/ExportCSV.csv';  
  
/\*import an external data file to a SAS data set\*/  
PROC IMPORT DATAFILE=REFFILE  
 /\*the type of data, a cvs file\*/  
 DBMS=CSV  
 /\*the output SAS data set and library\*/  
 OUT=WORK.IMPORT;  
 /\* yes means SAS variable names should be generated\*/  
 GETNAMES=YES;  
/\*execute the previous statements\*/  
RUN;  
  
/\* describes the structure of the data, dataset and variable level, where it is, get it and run it\*/  
PROC CONTENTS DATA=WORK.IMPORT; RUN;  
  
  
/\*open the table in work.import library\*/  
%web\_open\_table(WORK.IMPORT);

**Findings**

**Analysis Variables**

For this Portfolio Project, I did not copy the SAS code from the Milestone as I could not get the data from MEPS loaded into SAS. So, with the approval of the instructor I chose another data set and added comments to each portion of the code. I chose

the CDC’s Chronic Disease Measures for the 500 Largest American Cities Project: Birmingham, Los Angeles and United States data:





**Variable: Los\_Angeles**

| **N** | **Mean** | **Std Dev** | **Std Err** | **Minimum** | **Maximum** |
| --- | --- | --- | --- | --- | --- |
| 56 | 29.6000 | 24.7488 | 3.3072 | 2.8000 | 85.0000 |

| **Mean** | **95% CL Mean** | | **Std Dev** | **95% CL Std Dev** | |
| --- | --- | --- | --- | --- | --- |
| 29.6000 | 22.9722 | 36.2278 | 24.7488 | 20.8650 | 30.4225 |

| **DF** | **t Value** | **Pr > |t|** |
| --- | --- | --- |
| 55 | 8.95 | <.0001 |

**Variable: Birmingham**

| **N** | **Mean** | **Std Dev** | **Std Err** | **Minimum** | **Maximum** |
| --- | --- | --- | --- | --- | --- |
| 56 | 34.4161 | 25.3682 | 3.3900 | 3.8000 | 85.0000 |

| **Mean** | **95% CL Mean** | | **Std Dev** | **95% CL Std Dev** | |
| --- | --- | --- | --- | --- | --- |
| 34.4161 | 27.6224 | 41.2097 | 25.3682 | 21.3872 | 31.1839 |

| **DF** | **t Value** | **Pr > |t|** |
| --- | --- | --- |
| 55 | 10.15 | <.0001 |

**Difference: Birmingham - Los\_Angeles**

| **N** | **Mean** | **Std Dev** | **Std Err** | **Minimum** | **Maximum** |
| --- | --- | --- | --- | --- | --- |
| 56 | 4.8161 | 6.1462 | 0.8213 | -7.3000 | 17.1000 |

| **Mean** | **95% CL Mean** | | **Std Dev** | **95% CL Std Dev** | |
| --- | --- | --- | --- | --- | --- |
| 4.8161 | 3.1701 | 6.4620 | 6.1462 | 5.1817 | 7.5553 |

| **DF** | **t Value** | **Pr > |t|** |
| --- | --- | --- |
| 55 | 5.86 | <.0001 |

I am unable to run the Two-sample test as I ca not resolve the ERROR: The CLASS variable has more than two levels.

I chose to analyze Birmingham, Los Angeles and the United States for 27 measures of chronic disease related to unhealthy behaviors (Urban Institute May 2017) - the mean which is the percentage of the population that suffers from the 27 measures Birmingham has the highest score of 34.4 followed by the United States 31.19 and Los Angeles 29.6.

| **Tests for Normality** | | | | |
| --- | --- | --- | --- | --- |
| **Test** | **Statistic** | | **p Value** | |
| **Shapiro-Wilk** | **W** | 0.94717 | **Pr < W** | 0.0159 |
| **Kolmogorov-Smirnov** | **D** | 0.138347 | **Pr > D** | <0.0100 |
| **Cramer-von Mises** | **W-Sq** | 0.248542 | **Pr > W-Sq** | <0.0050 |
| **Anderson-Darling** | **A-Sq** | 1.265042 | **Pr > A-Sq** | <0.0050 |

| **N** | **Mean** | **Std Dev** | **Std Err** | **Minimum** | **Maximum** |
| --- | --- | --- | --- | --- | --- |
| 56 | 4.8161 | 6.1462 | 0.8213 | -7.3000 | 17.1000 |

| **Mean** | **95% CL Mean** | | **Std Dev** | **95% CL Std Dev** | |
| --- | --- | --- | --- | --- | --- |
| 4.8161 | 3.1701 | 6.4620 | 6.1462 | 5.1817 | 7.5553 |

| **DF** | **t Value** | **Pr > |t|** |
| --- | --- | --- |
| 55 | 5.86 | <.0001 |

All of the Tests for Normality are less than the alpha value of .05 so we can reject the null hypothesis.

The difference between the two means (Los Angeles and Birmingham) is 4.8161 , the confidence level at 95% confidence is 3.1701 through 6.4620 ( 95% Confidence Level) . The minimum difference is -7.3000 and the maximum difference is 17.1000. The degrees of freedom are observations (56) – 1 or 55.

Using data structure presentations from UCLA Institute for Digital Research & Education I noted: (UCLA Institute for Digital Research, 2019) :

The t Value of 5.89- it is the ratio of the sample mean and the standard error of the mean (4.8161/ 0.8213). The p-value <.0001 is less than .05 the chosen alpha value ( based on 95% of the area of a critical normal distribution – within 1.96 standard deviations of the mean) so I reject the null hypothesis , the means are not equal, so I accept the alternative hypothesis.

As Pr>|t| - p-value 2 tailed probability computed using a t distribution – or the probability of observing a greater absolute value of t under NULL HYPOTHESIS, so .0001 is less than .05 (alpha value) so I reject the null hypothesis – the means are not the same and they are not equal. Therefor I accept the alternate hypothesis.







**Code to Generate SAS Summary Statistics**

SAS Code

/\*  
 \*  
 \* Task code generated by SAS Studio 3.8   
 \*  
 \* Generated on '7/7/19, 5:56 AM'   
 \* Generated by 'sasdemo'   
 \* Generated on server 'LOCALHOST'   
 \* Generated on SAS platform 'Linux LIN X64 2.6.32-754.6.3.el6.x86\_64'   
 \* Generated on SAS version '9.04.01M6P11072018'   
 \* Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/75.0.3770.100 Safari/537.36'   
 \* Generated on web client 'http://localhost:10080/SASStudio/38/main?locale=en\_US&zone=GMT-07%253A00'   
 \*  
 \*/  
  
/\*To suppress the writing of the title of the procedure\*/

ods noproctitle;  
/\* Create statistical graphics\*/

ods graphics / imagemap=on;  
/\*Create output from work.import, create values of the data set\*/  
proc means data=WORK.IMPORT chartype mean std min max n vardef=df;  
 var Birmingham;  
 class Los\_Angeles;  
/\*Execute the program\*/

run;

**Code to Generate Hypotheses**

**Null Hypothesis**

Null Hypothesis, Towards Data Science (2019) states,”… a null hypothesis , proposes that no significant difference exists in a set of given observations.”

We need a so-called critical value – Minitab Express Support (2019) indicates it is a point compared to the test statistic so as to determine whether or not to reject the null hypothesis. I choose 1.96 “...which is based on 95% of the area of a critical normal distribution is within 1.96 standard deviations of the mean.”

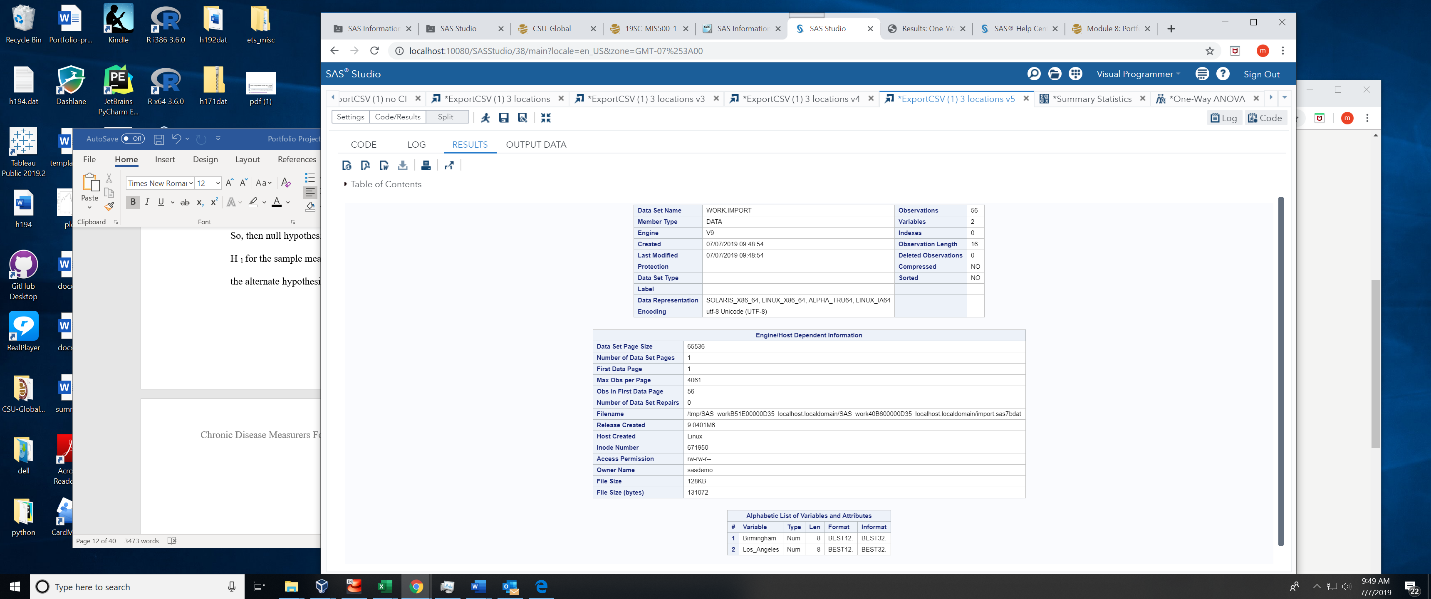
**The Null Hypothesis Statement**

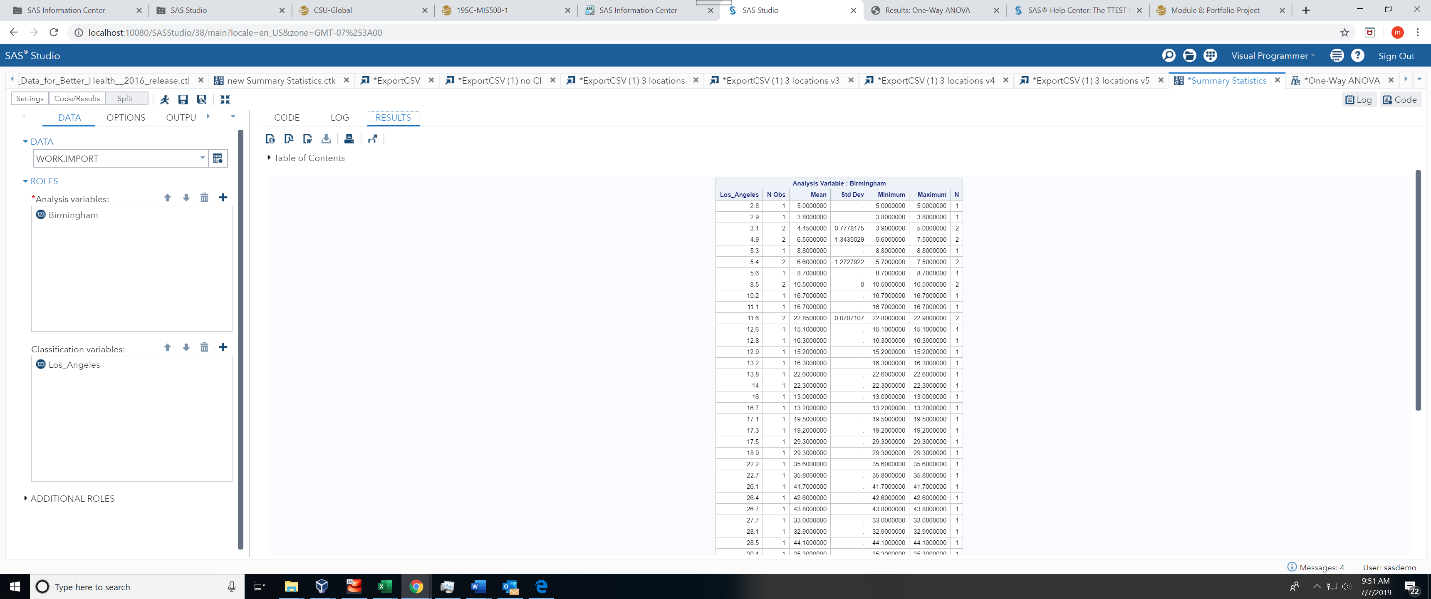
H 0 for the sample means of Birmingham and Los Angeles

So, then null hypothesis H 0 - is the sample means are equal and

H 1 for the sample means of Birmingham and Los Angeles are not the same, the alternate hypothesis: is the sample means are not equal.

**Screenshots from the Execution**





**Code to Generate SAS Summary Statistics**

**Variables**

The numeric variables are: year, data\_value, Low and High\_confidence \_Limit and Population (2010)

**SAS Code**

\*  
 \*  
 \* Task code generated by SAS Studio 3.8   
 \*  
 \* Generated on '7/7/19, 9:34 AM'   
 \* Generated by 'sasdemo'   
 \* Generated on server 'LOCALHOST'   
 \* Generated on SAS platform 'Linux LIN X64 2.6.32-754.6.3.el6.x86\_64'   
 \* Generated on SAS version '9.04.01M6P11072018'   
 \* Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/75.0.3770.100 Safari/537.36'   
 \* Generated on web client 'http://localhost:10080/SASStudio/38/main?locale=en\_US&zone=GMT-07%253A00'   
 \*  
 \*/  
  
ods noproctitle;  
ods graphics / imagemap=on;  
  
proc means data=WORK.IMPORT chartype mean std min max n vardef=df;  
 var Birmingham;  
 class Los\_Angeles;  
run;

**Code to Generate Hypotheses for a Two-Sample t Test – I used a OneWay Anova**

**One-Way Anova**

**SAS Code**

/\*  
 \*  
 \* Task code generated by SAS Studio 3.8   
 \*  
 \* Generated on '7/7/19, 6:36 AM'   
 \* Generated by 'sasdemo'   
 \* Generated on server 'LOCALHOST'   
 \* Generated on SAS platform 'Linux LIN X64 2.6.32-754.6.3.el6.x86\_64'   
 \* Generated on SAS version '9.04.01M6P11072018'   
 \* Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/75.0.3770.100 Safari/537.36'   
 \* Generated on web client 'http://localhost:10080/SASStudio/38/main?locale=en\_US&zone=GMT-07%253A00'   
 \*  
 \*/  
  
Title;  
ods noproctitle;  
ods graphics / imagemap=on;  
  
proc glm data=WORK.IMPORT;  
 class Los\_Angeles;  
 model Birmingham=Los\_Angeles;  
 means Los\_Angeles / hovtest=levene welch plots=none;  
 lsmeans Los\_Angeles / adjust=tukey pdiff alpha=.05;  
 run;  
quit;

**Summary of Data**

|  |  |  |  |
| --- | --- | --- | --- |
| Mean number: | 1 2 | | 2 |
|  | Birmingham Los Angeles | | Los Angeles |
|  |  | |  |
| Sample Size | 56 56 | | 56 |
| Sample Mean | 34.4 29.6 | | 29.6 |
| Sample Standard Deviation | 25.4 24.7 | | 24.7487667 |
|  |  | |  |
| HO mean1 = mean2 |  | |  |
| H1 the means are not equal | |  | |

The means are not equal, so I reject the null hypothesis in favor of the alternate hypothesis.

My main purpose is to test if two groups of data differ from each other significantly for one or more characteristics. (Biau, Jones and Porcher 2009)

The null hypothesis (H0) is there is no difference between the groups and means are the same. The alternative hypothesis (H1) is that there is a difference between the means.

| **Class Level Information** | | |
| --- | --- | --- |
| **Class** | **Levels** | **Values** |
| **Los\_Angeles** | 50 | 14 16 53 65 85 2.8 2.9 3.1 4.9 5.3 5.4 5.6 8.5 10.2 11.1 11.6 12.6 12.8 12.9 13.2 13.8 16.7 17.1 17.3 17.5 18.9 22.2 22.7 26.1 26.4 26.7 27.7 28.1 28.5 30.4 32.6 33.3 33.9 36.6 36.8 58.8 59.2 61.5 62.1 66.2 67.5 72.1 74.1 74.4 79.5 |

Using the UCLA framework (UCLA Institute for Digital Research & Education), the p-Value of <.0001 is less than the alpha value of .05. The Sum of Squares is computed by summing the squared differences between each group (city). DF is degrees of freedom: observations 56 – model 49 error 6 for a corrected total of 55.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Number of Observations Read** | | | 56 | |
| **Number of Observations Used** | | | 56 | |
| **Source** | **DF** | **Sum of Squares** | | **Mean Square** | | **F Value** | **Pr > F** |
| **Model** | 49 | 35377.34054 | | 721.98654 | | 246.76 | <.0001 |
| **Error** | 6 | 17.55500 | | 2.92583 | |  |  |
| **Corrected Total** | 55 | 35394.89554 | |  | |  |  |

| **R-Square** | **Coeff Var** | **Root MSE** | **Birmingham Mean** |
| --- | --- | --- | --- |
| 0.999504 | 4.970081 | 1.710507 | 34.41607 |

| **Source** | **DF** | **Type I SS** | **Mean Square** | **F Value** | **Pr > F** | |
| --- | --- | --- | --- | --- | --- | --- |
| **Los\_Angeles** | 49 | 35377.34054 | 721.98654 | 246.76 | | <.0001 | |

| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Los\_Angeles** | 49 | 35377.34054 | 721.98654 | 246.76 | <.0001 |

The goal of tests for differences between the means and to attempt to quantify the differences. (SAS Support)

Here the F value 246.76 and p<.0001 seems to state the model accounts for a significant portion of the variability in the so-called dependent variable.



| **Levene's Test for Homogeneity of Birmingham Variance ANOVA of Squared Deviations from Group Means** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Los\_Angeles** | 0 | 0 | . | . | . |
| **Error** | 0 | 0 | . |  |  |

The p-Value of <.0035 is less than the alpha value of .05

| **Welch's ANOVA for Birmingham** | | | |
| --- | --- | --- | --- |
| **Source** | **DF** | **F Value** | **Pr > F** |
| **Los\_Angeles** | 4.0000 | 278.45 | 0.0035 |
| **Error** | 2.0142 |  |  |

| **Level of Los\_Angeles** | **N** | **Birmingham** | |
| --- | --- | --- | --- |
| **Mean** | **Std Dev** |
| **14** | 1 | 22.3000000 | . |
| **16** | 1 | 13.0000000 | . |
| **53** | 1 | 70.1000000 | . |
| **65** | 1 | 76.9000000 | . |
| **85** | 1 | 85.0000000 | . |
| **2.8** | 1 | 5.0000000 | . |
| **2.9** | 1 | 3.8000000 | . |
| **3.1** | 2 | 4.4500000 | 0.77781746 |
| **4.9** | 2 | 6.5500000 | 1.34350288 |
| **5.3** | 1 | 8.8000000 | . |
| **5.4** | 2 | 6.6000000 | 1.27279221 |
| **5.6** | 1 | 8.7000000 | . |
| **8.5** | 2 | 10.5000000 | 0.00000000 |
| **10.2** | 1 | 16.7000000 | . |
| **11.1** | 1 | 16.7000000 | . |
| **11.6** | 2 | 22.8500000 | 0.07071068 |
| **12.6** | 1 | 15.1000000 | . |
| **12.8** | 1 | 16.3000000 | . |
| **12.9** | 1 | 15.2000000 | . |
| **13.2** | 1 | 16.3000000 | . |
| **13.8** | 1 | 22.6000000 | . |
| **16.7** | 1 | 13.2000000 | . |
| **17.1** | 1 | 19.5000000 | . |
| **17.3** | 1 | 19.2000000 | . |
| **17.5** | 1 | 29.3000000 | . |
| **18.9** | 1 | 29.3000000 | . |
| **22.2** | 1 | 35.6000000 | . |
| **22.7** | 1 | 35.8000000 | . |
| **26.1** | 1 | 41.7000000 | . |
| **26.4** | 1 | 42.6000000 | . |
| **26.7** | 1 | 43.8000000 | . |
| **27.7** | 1 | 33.0000000 | . |
| **28.1** | 1 | 32.9000000 | . |
| **28.5** | 1 | 44.1000000 | . |
| **30.4** | 1 | 35.3000000 | . |
| **32.6** | 1 | 39.2000000 | . |
| **33.3** | 1 | 26.0000000 | . |
| **33.9** | 1 | 26.7000000 | . |
| **36.6** | 1 | 41.8000000 | . |
| **36.8** | 1 | 41.5000000 | . |
| **58.8** | 1 | 52.6000000 | . |
| **59.2** | 1 | 52.5000000 | . |
| **61.5** | 1 | 64.1000000 | . |
| **62.1** | 1 | 64.7000000 | . |
| **66.2** | 1 | 76.9000000 | . |
| **67.5** | 1 | 81.3000000 | . |
| **72.1** | 1 | 74.8000000 | . |
| **74.1** | 1 | 75.8000000 | . |
| **74.4** | 1 | 77.7000000 | . |
| **79.5** | 2 | 81.0000000 | 3.67695526 |

| **Los\_Angeles** | **Birmingham LSMEAN** | **LSMEAN Number** |
| --- | --- | --- |
| **14** | 22.3000000 | 1 |
| **16** | 13.0000000 | 2 |
| **53** | 70.1000000 | 3 |
| **65** | 76.9000000 | 4 |
| **85** | 85.0000000 | 5 |
| **2.8** | 5.0000000 | 6 |
| **2.9** | 3.8000000 | 7 |
| **3.1** | 4.4500000 | 8 |
| **4.9** | 6.5500000 | 9 |
| **5.3** | 8.8000000 | 10 |
| **5.4** | 6.6000000 | 11 |
| **5.6** | 8.7000000 | 12 |
| **8.5** | 10.5000000 | 13 |
| **10.2** | 16.7000000 | 14 |
| **11.1** | 16.7000000 | 15 |
| **11.6** | 22.8500000 | 16 |
| **12.6** | 15.1000000 | 17 |
| **12.8** | 16.3000000 | 18 |
| **12.9** | 15.2000000 | 19 |
| **13.2** | 16.3000000 | 20 |
| **13.8** | 22.6000000 | 21 |
| **16.7** | 13.2000000 | 22 |
| **17.1** | 19.5000000 | 23 |
| **17.3** | 19.2000000 | 24 |
| **17.5** | 29.3000000 | 25 |
| **18.9** | 29.3000000 | 26 |
| **22.2** | 35.6000000 | 27 |
| **22.7** | 35.8000000 | 28 |
| **26.1** | 41.7000000 | 29 |
| **26.4** | 42.6000000 | 30 |
| **26.7** | 43.8000000 | 31 |
| **27.7** | 33.0000000 | 32 |
| **28.1** | 32.9000000 | 33 |
| **28.5** | 44.1000000 | 34 |
| **30.4** | 35.3000000 | 35 |
| **32.6** | 39.2000000 | 36 |
| **33.3** | 26.0000000 | 37 |
| **33.9** | 26.7000000 | 38 |
| **36.6** | 41.8000000 | 39 |
| **36.8** | 41.5000000 | 40 |
| **58.8** | 52.6000000 | 41 |
| **59.2** | 52.5000000 | 42 |
| **61.5** | 64.1000000 | 43 |
| **62.1** | 64.7000000 | 44 |
| **66.2** | 76.9000000 | 45 |
| **67.5** | 81.3000000 | 46 |
| **72.1** | 74.8000000 | 47 |
| **74.1** | 75.8000000 | 48 |
| **74.4** | 77.7000000 | 49 |
| **79.5** | 81.0000000 | 50 |

| **Least Squares Means for effect Los\_Angeles Pr > |t| for H0: LSMean(i)=LSMean(j)  Dependent Variable: Birmingham (*the additional data was redacted)*** | | | | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **i/j** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** |
| **1** |  | 0.3457 | <.0001 | <.0001 | <.0001 | 0.0272 | 0.0195 | 0.0112 | 0.0212 | 0.0865 | 0.0216 | 0.0838 | 0.0831 | 0.8840 |
| **2** | 0.3457 |  | <.0001 | <.0001 | <.0001 | 0.5163 | 0.3570 | 0.2868 | 0.6021 | 0.9877 | 0.6114 | 0.9846 | 0.9999 | 0.9970 |
| **3** | <.0001 | <.0001 |  | 0.7074 | 0.0556 | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 |
| **4** | <.0001 | <.0001 | 0.7074 |  | 0.5014 | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 |
| **5** | <.0001 | <.0001 | 0.0556 | 0.5014 |  | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 |
| **6** | 0.0272 | 0.5163 | <.0001 | <.0001 | <.0001 |  | 1.0000 | 1.0000 | 1.0000 | 0.9959 | 1.0000 | 0.9970 | 0.7790 | 0.1561 |
| **7** | 0.0195 | 0.3570 | <.0001 | <.0001 | <.0001 | 1.0000 |  | 1.0000 | 0.9996 | 0.9449 | 0.9995 | 0.9526 | 0.5565 | 0.1051 |
| **8** | 0.0112 | 0.2868 | <.0001 | <.0001 | <.0001 | 1.0000 | 1.0000 |  | 0.9999 | 0.9430 | 0.9998 | 0.9521 | 0.4372 | 0.0704 |
| **9** | 0.0212 | 0.6021 | <.0001 | <.0001 | <.0001 | 1.0000 | 0.9996 | 0.9999 |  | 1.0000 | 1.0000 | 1.0000 | 0.8857 | 0.1551 |
| **10** | 0.0865 | 0.9877 | <.0001 | <.0001 | <.0001 | 0.9959 | 0.9449 | 0.9430 | 1.0000 |  | 1.0000 | 1.0000 | 1.0000 | 0.5313 |
| **11** | 0.0216 | 0.6114 | <.0001 | <.0001 | <.0001 | 1.0000 | 0.9995 | 0.9998 | 1.0000 | 1.0000 |  | 1.0000 | 0.8940 | 0.1581 |
| **12** | 0.0838 | 0.9846 | <.0001 | <.0001 | <.0001 | 0.9970 | 0.9526 | 0.9521 | 1.0000 | 1.0000 | 1.0000 |  | 1.0000 | 0.5163 |
| **13** | 0.0831 | 0.9999 | <.0001 | <.0001 | <.0001 | 0.7790 | 0.5565 | 0.4372 | 0.8857 | 1.0000 | 0.8940 | 1.0000 |  | 0.6488 |
| **14** | 0.8840 | 0.9970 | <.0001 | <.0001 | <.0001 | 0.1561 | 0.1051 | 0.0704 | 0.1551 | 0.5313 | 0.1581 | 0.5163 | 0.6488 |  |
| **15** | 0.8840 | 0.9970 | <.0001 | <.0001 | <.0001 | 0.1561 | 0.1051 | 0.0704 | 0.1551 | 0.5313 | 0.1581 | 0.5163 | 0.6488 | 1.0000 |
| **16** | 1.0000 | 0.1740 | <.0001 | <.0001 | <.0001 | 0.0112 | 0.0079 | 0.0032 | 0.0062 | 0.0373 | 0.0063 | 0.0360 | 0.0260 | 0.6583 |
| **17** | 0.6422 | 1.0000 | <.0001 | <.0001 | <.0001 | 0.2661 | 0.1784 | 0.1280 | 0.2868 | 0.7868 | 0.2923 | 0.7713 | 0.9163 | 1.0000 |
| **18** | 0.8311 | 0.9993 | <.0001 | <.0001 | <.0001 | 0.1784 | 0.1198 | 0.0816 | 0.1809 | 0.5938 | 0.1844 | 0.5779 | 0.7241 | 1.0000 |
| **19** | 0.6585 | 1.0000 | <.0001 | <.0001 | <.0001 | 0.2574 | 0.1725 | 0.1233 | 0.2761 | 0.7713 | 0.2814 | 0.7556 | 0.9040 | 1.0000 |
| **20** | 0.8311 | 0.9993 | <.0001 | <.0001 | <.0001 | 0.1784 | 0.1198 | 0.0816 | 0.1809 | 0.5938 | 0.1844 | 0.5779 | 0.7241 | 1.0000 |
| **21** | 1.0000 | 0.3137 | 0.0001 | <.0001 | <.0001 | 0.0250 | 0.0180 | 0.0102 | 0.0193 | 0.0786 | 0.0196 | 0.0761 | 0.0744 | 0.8451 |
| **22** | 0.3687 | 1.0000 | <.0001 | <.0001 | <.0001 | 0.4868 | 0.3347 | 0.2658 | 0.5655 | 0.9808 | 0.5746 | 0.9765 | 0.9997 | 0.9985 |
| **23** | 1.0000 | 0.7556 | <.0001 | <.0001 | <.0001 | 0.0630 | 0.0436 | 0.0266 | 0.0547 | 0.2180 | 0.0557 | 0.2108 | 0.2415 | 1.0000 |
| **24** | 0.9997 | 0.8019 | <.0001 | <.0001 | <.0001 | 0.0692 | 0.0477 | 0.0294 | 0.0609 | 0.2409 | 0.0620 | 0.2330 | 0.2709 | 1.0000 |



**Following the GitHub Instructions for uploading source files, upload your source files to your GitHub Account**

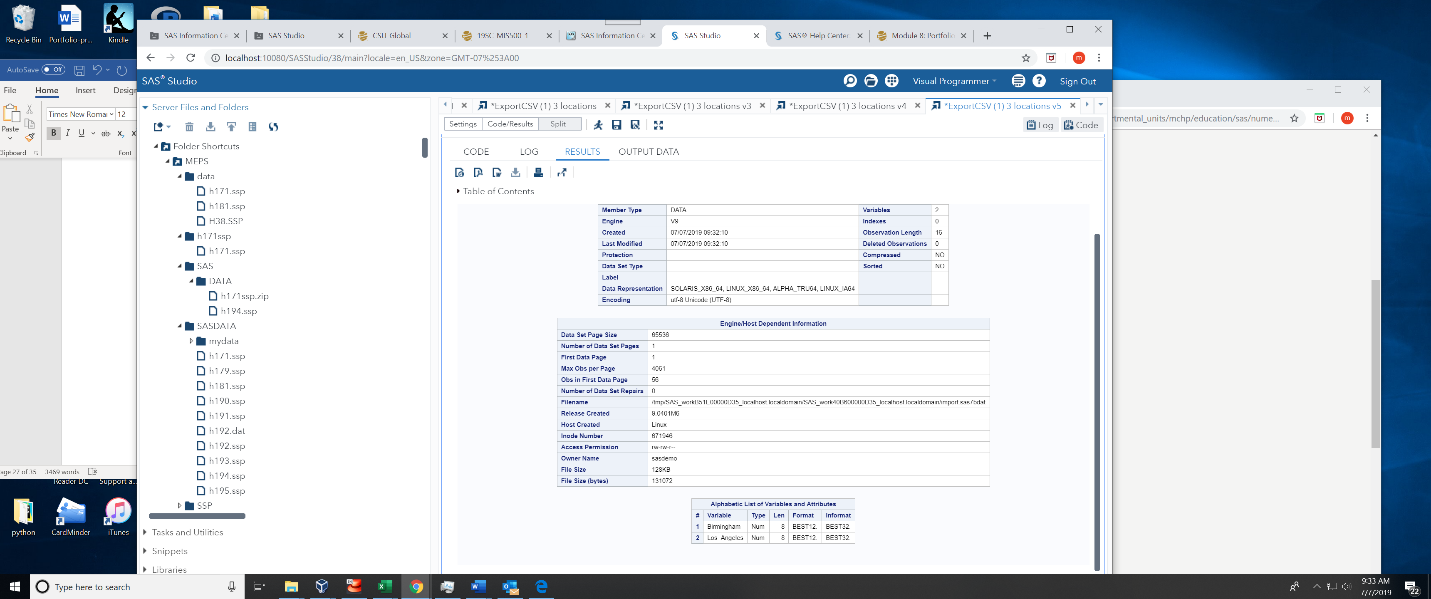
[@mpaulip](https://github.com/mpaulip)

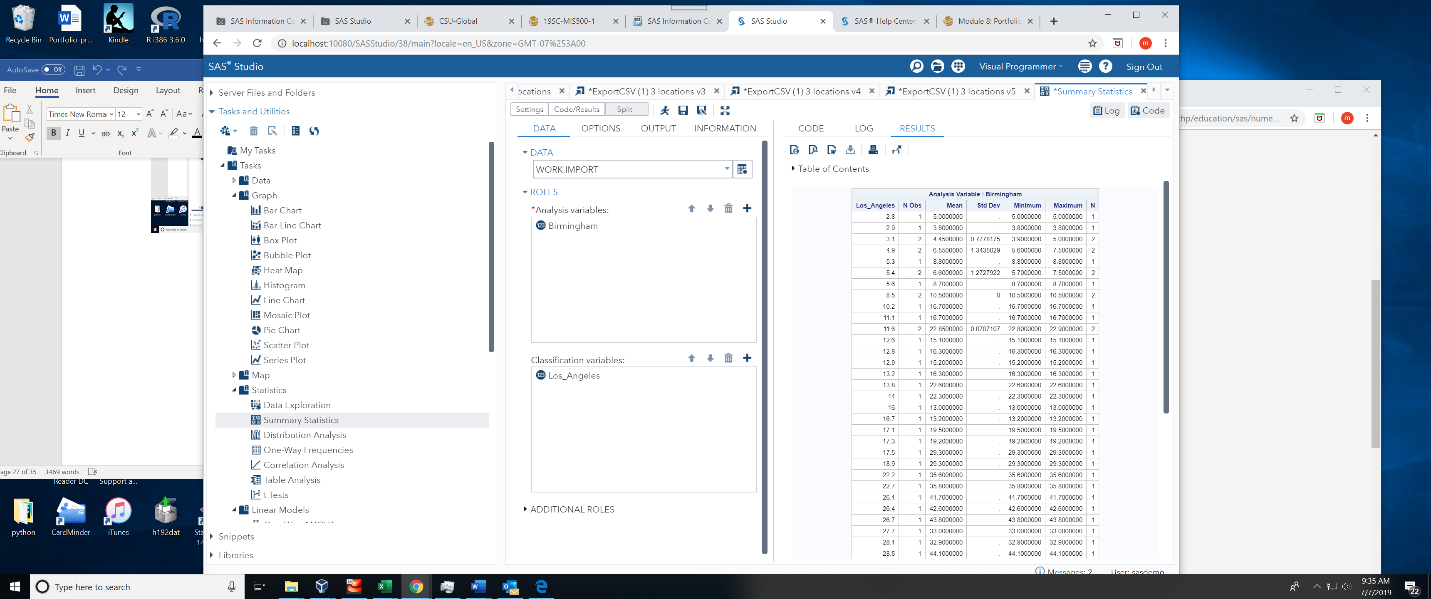
[**mpaulip**](https://github.com/mpaulip/Portfolio-Project/commits?author=mpaulip) [Add files via upload](https://github.com/mpaulip/Portfolio-Project/commit/5c3ae649e2e51f7a94a709d8e2625b5c625785c1)

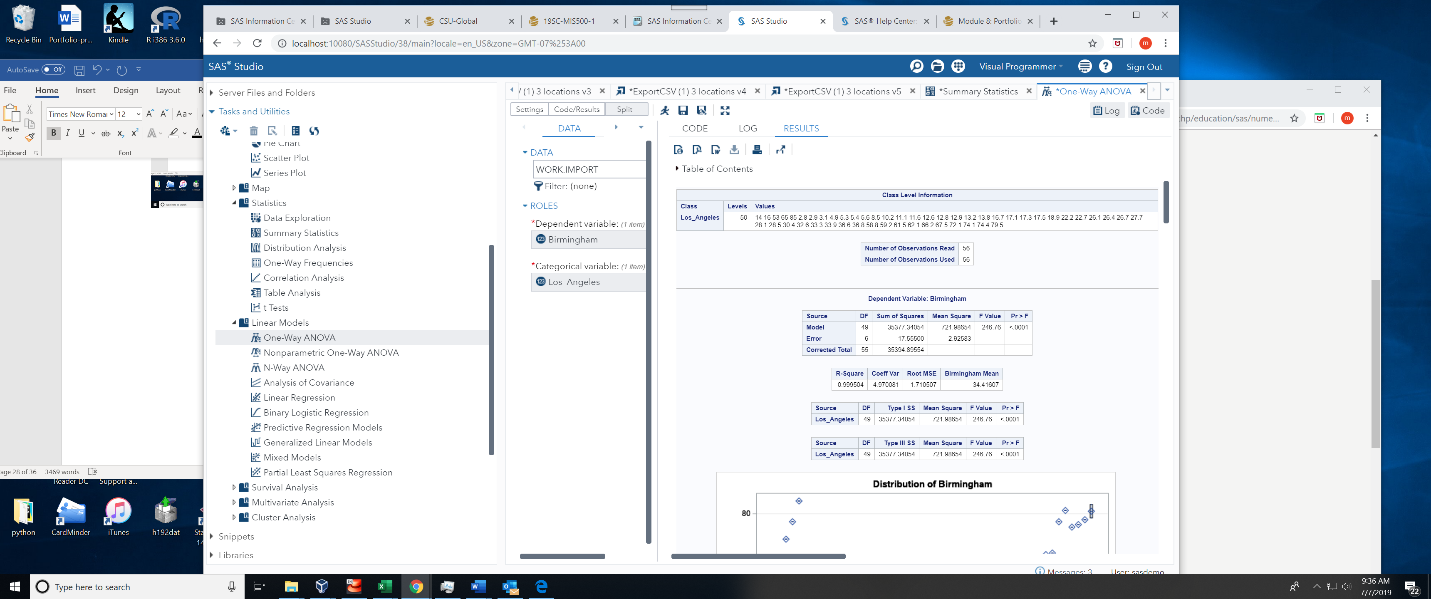
Latest commit [5c3ae64](https://github.com/mpaulip/Portfolio-Project/commit/5c3ae649e2e51f7a94a709d8e2625b5c625785c1) now

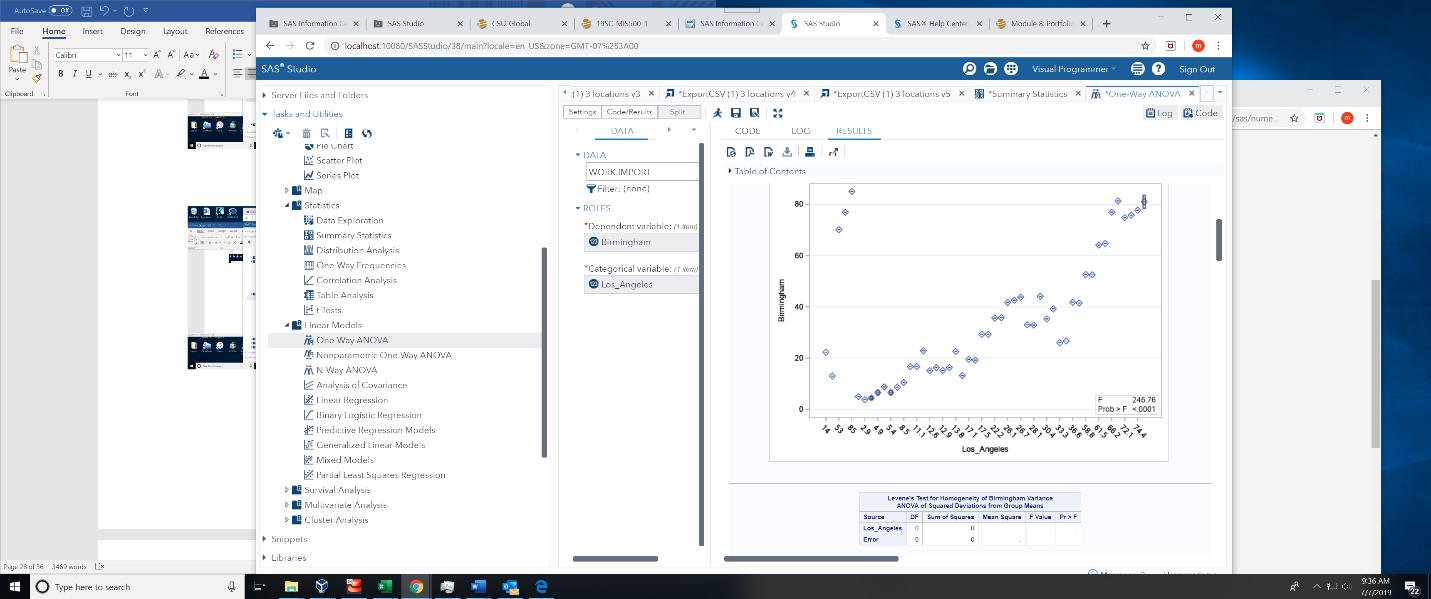
| **Type** | **Name** | **Latest commit message** | **Commit time** |
| --- | --- | --- | --- |
|  | [Portfolio Project v3.docx](https://github.com/mpaulip/Portfolio-Project/blob/master/Portfolio%20Project%20v3.docx) |  |  |

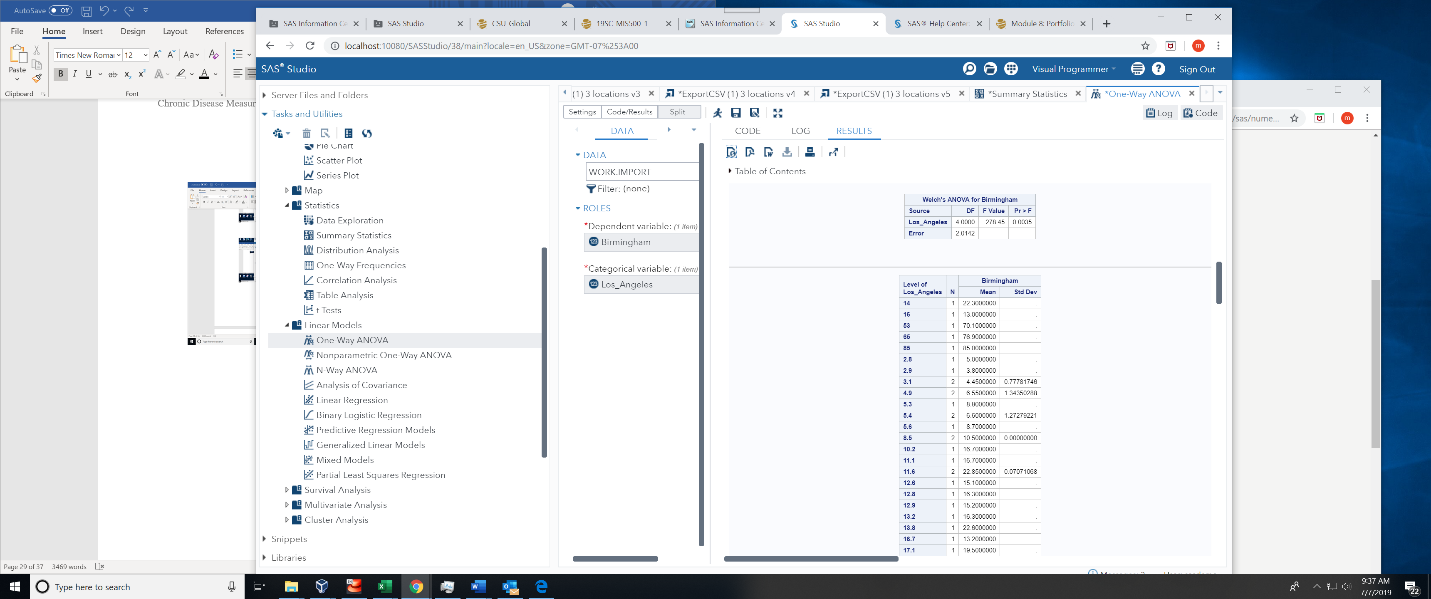
**Screenshots from The Output Result**

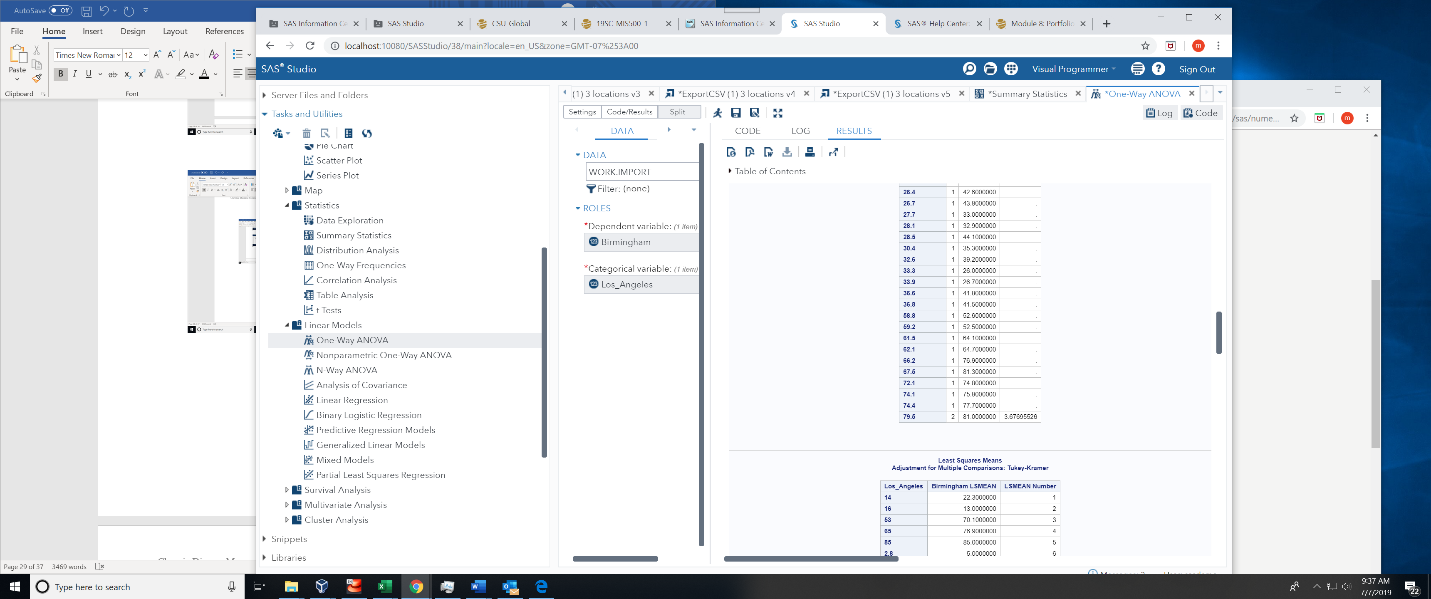


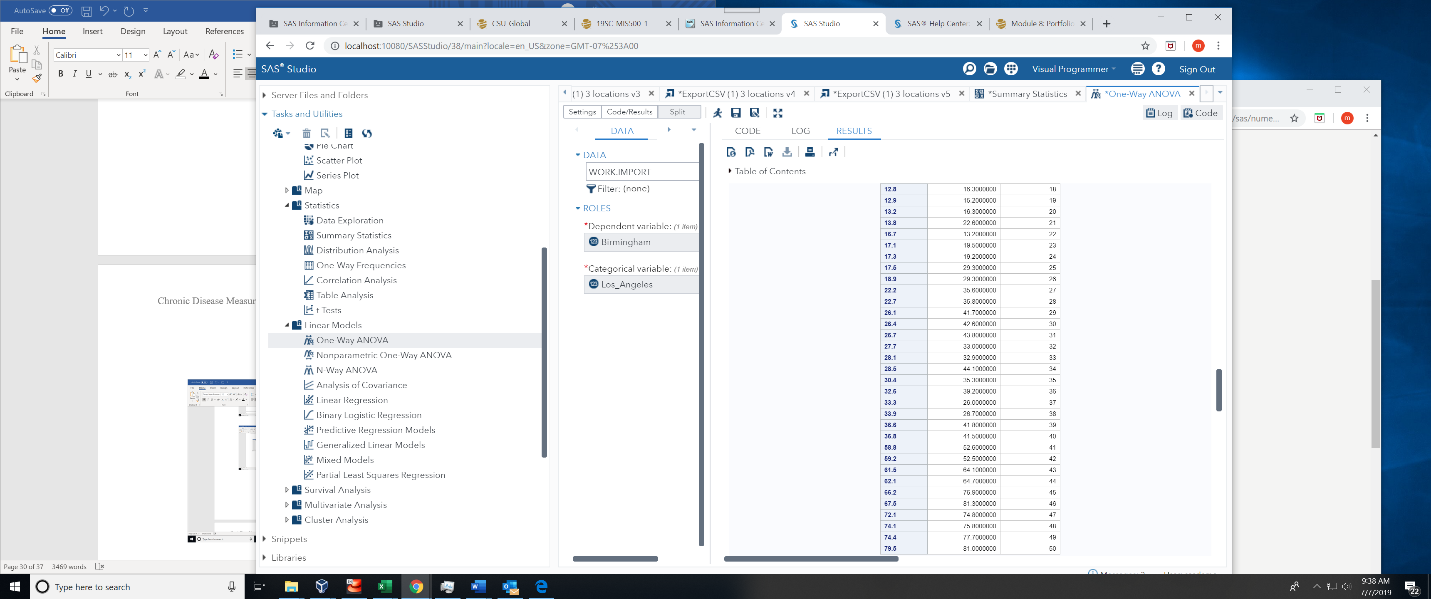


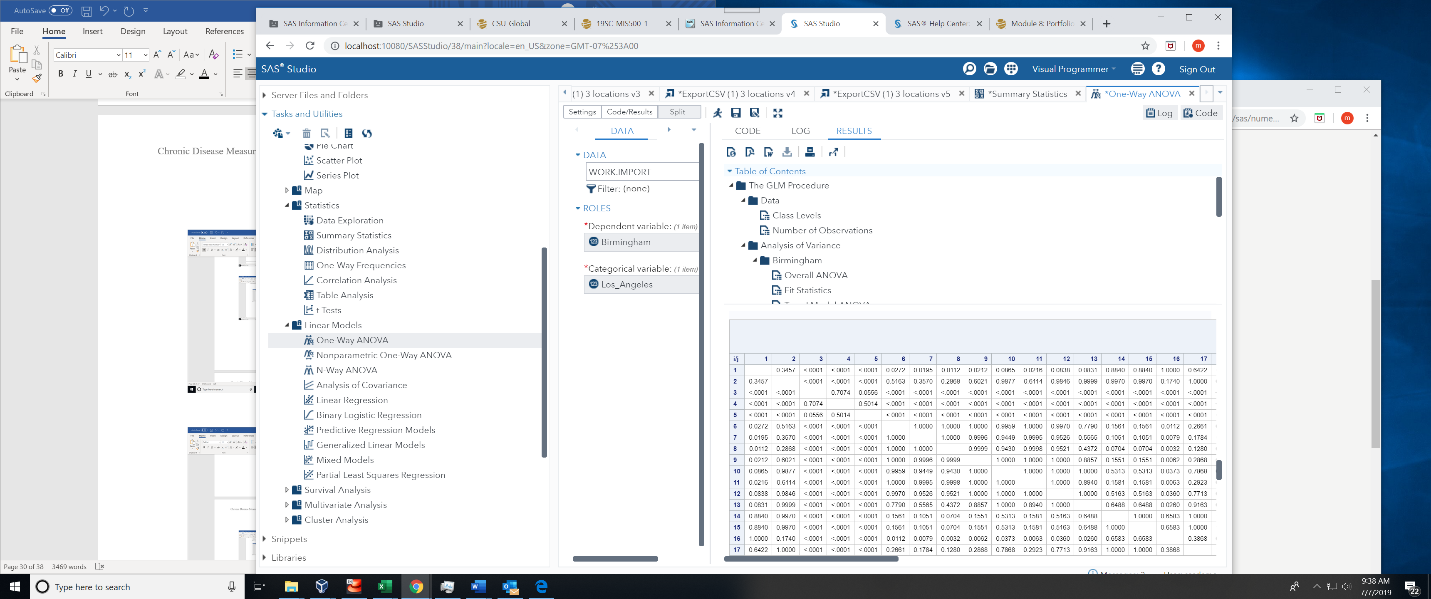


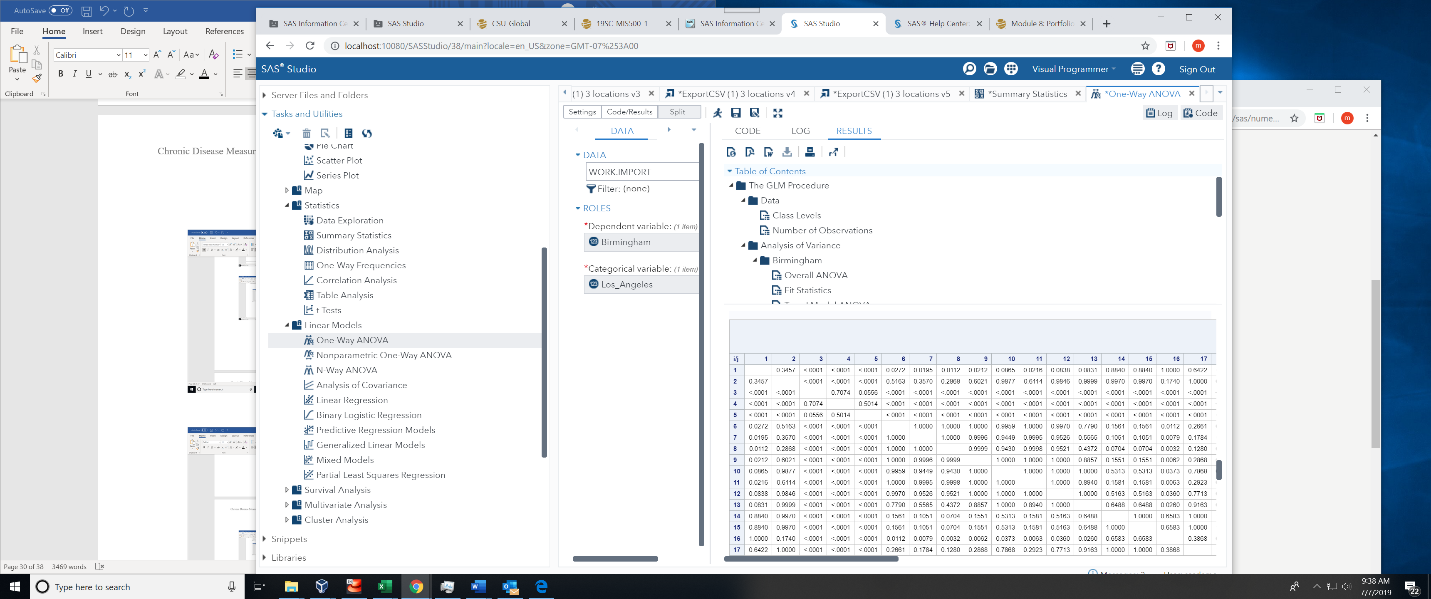


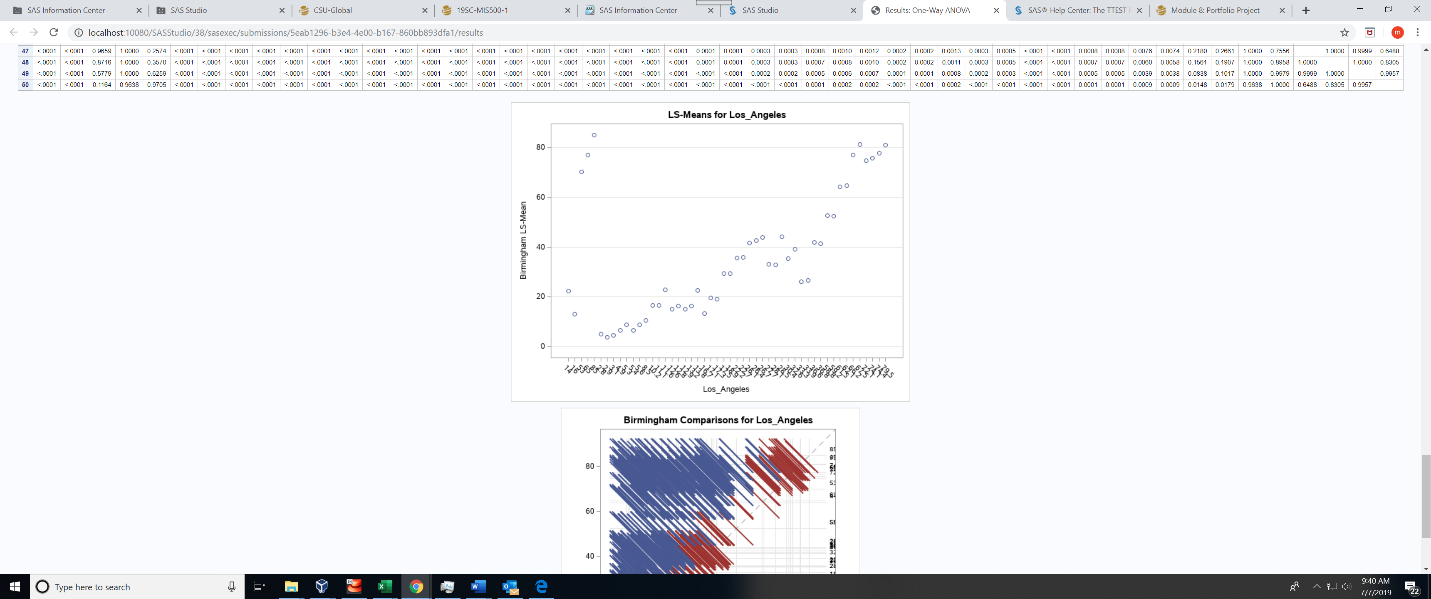


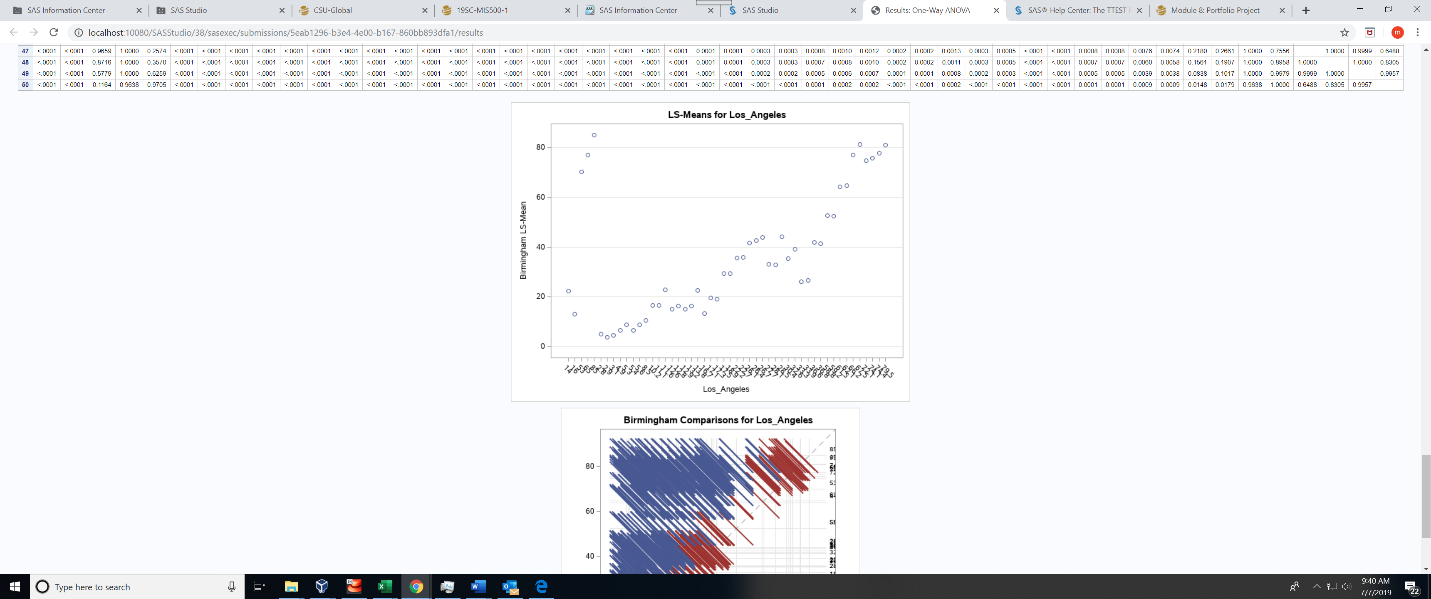








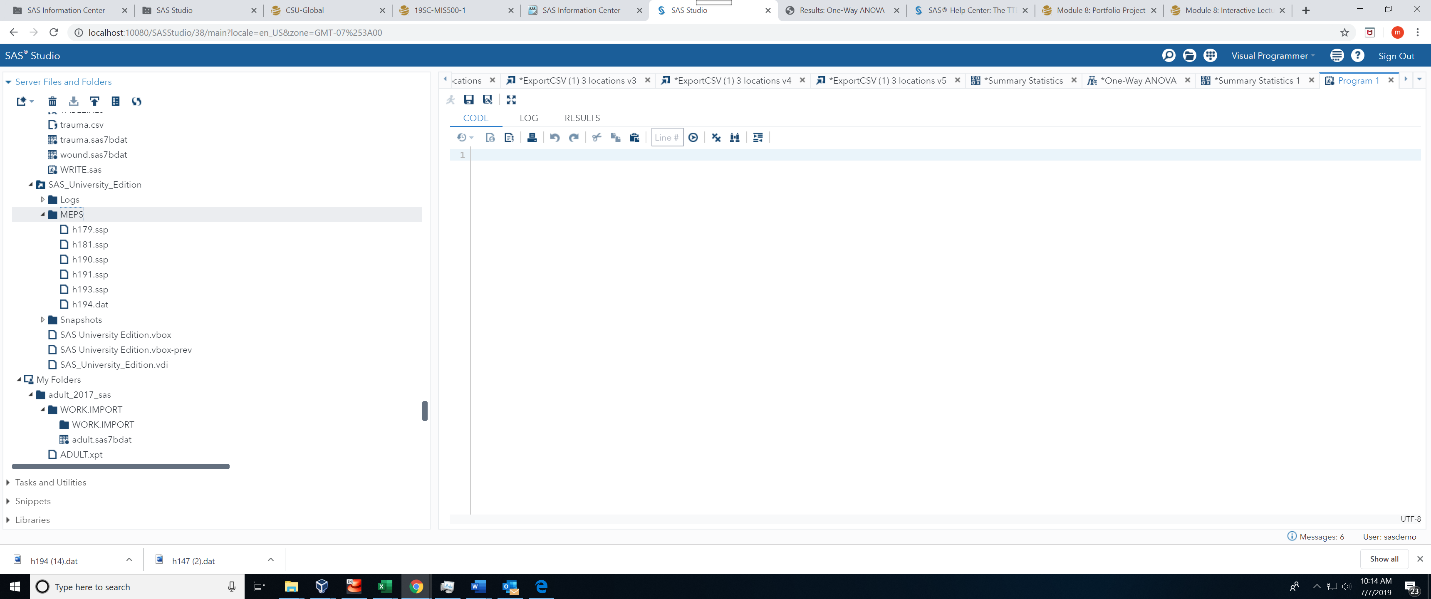




**If for any reason you could not get the code to run, please include the screenshots of your work and explanations of the issue(s).**

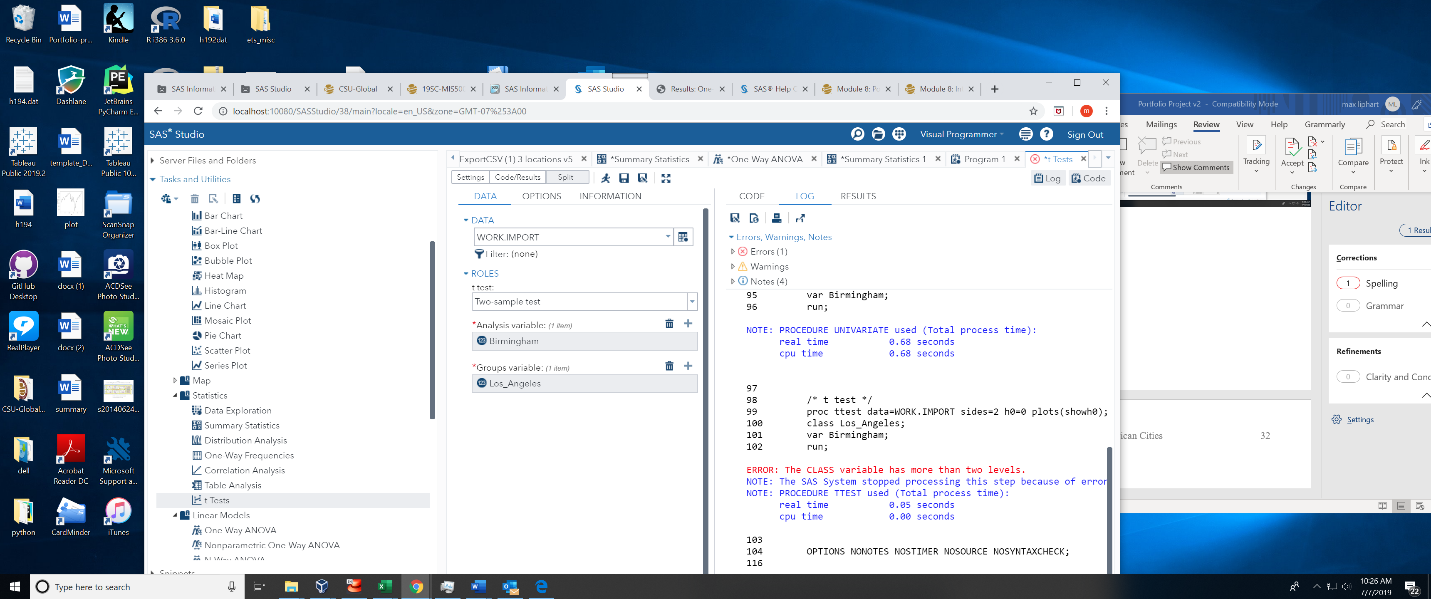
**MEPS**

My original project to analyze the MEPS data was unsuccessful as I was not able to load .ssp or .dat data into SAS to perform the work. I researched the issue online, including SAS help web sites, and had multiple tutoring sessions, I spent approximately 20 hours with the process and was unable to solve the problem.



**Two-Sample test**

I was unable to perform the two-sample Test for my two cities and after much research was unable to solve the problem “Class variable has more than two levels”:



**What I did to get this project completed**

I spent significant time reviewing the 500 City Data (CDC) site trying to understand what the goal of the project was, and the datasets collected. I loaded the datasets into SAS via excel and .csv data and analyzed the provided data. I wanted to compare means of different datasets and further analyzed data from Birmingham and Los Angeles to determine which state has the lowest percent of the measurers of chronic diseases and found Los Angeles was lower than both Birmingham and the United States. I would like to plan to further analyze the data in the future as it is very interesting and helpful.

**References**

Biau, D.J., Jolles, B.M. & Porcher, R. Clin Orthop Relat Res (2010) 468: 885. (accessed July 2019) from <https://doi.org/10.1007/s11999-009-1164-4>

Corianne, Payton Scally, Kathryn L.S. Pettit Olivia Arena May 2017, 500 Cities: Local data for better health, Urban Institute, (accessed July 2019) from: <http://www.urban.org/sites/default/files/publication/90376/500-cities-project_1.pdf>

Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Division of Population Health, National Center for Chronic Disease Prevention and Health Promotion. 500 Cities Project Data [online]. 2018 [accessed July 2019]. URL: <https://www.cdc.gov/500cities>.

Centers for Disease Control and Prevention, CDC 24/7: Saving Lives. Protecting People. 500 Cities: Local Data for Better Health, 2018 release [online]. 2018 [accessed July 2019] URL: <https://chronicdata.cdc.gov/500-Cities/500-Cities-Local-Data-for-Better-Health-2018-relea/6vp6-wxuq>

Centers for Disease Control and Prevention, Recognized Lifestyle Change Program (accessed July 2019) from: <https://nccd.cdc.gov/DDT_DPRP/Programs.aspx> and BRFSS <https://www.cdc.gov/brfss/index.html>

Conduent Healthy Communities Institute, Data and Dashboards, Data Sources and Demographics, (accessed on line July 2019), URL: <https://help.healthycities.org/hc/en-us/categories/202738768-Data-and-Dashboards>

Minitab Express for Mac and PC retrieved July 2019 from[**https://www.minitab.com/enus/products/express/**](https://www.minitab.com/enus/products/express/)

Robert Wood Johnson Foundation, CDC Foundation and RWJF Launch 500 Cities Project, (accessed July 2019) from: <https://www.rwjf.org/en/library/articles-and-news/2017/03/cdc-foundation-and-robert-wood-johnson-foundation-to-launch-500-.html>

SAS University Edition: Help Center (accessed June 2019) retrieved from: <https://support.sas.com/software/products/university-edition/faq/main.htm>

SAS University Edition Program (accessed June 2019) retrieved from: <http://localhost:10080/SASStudio/38/main?locale=en_US&zone=GMT-07%253A00>

Towards Data Science, Sharing concepts, ideas, and codes (accessed July 2019) from: <https://towardsdatascience.com/>

United States Census Bureau, American fact Finder (accessed June 2019) from: <https://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t> and <https://factfinder.census.gov/faces/nav/jsf/pages/guided_search.xhtml> and ACS <https://www.census.gov/programs-surveys/acs>

UCLA Institute for Digital Research & Education, SAS Learning Modules, Fundamentals of Using SAS (Part I and II) (accessed July 2019) from: <https://stats.idre.ucla.edu/sas/modules/>

UCLA Institute for Digital Research & Education, Choosing the Correct Statistical Test in SAS,STATA, SPSS and R (accessed July 2019) from: <https://stats.idre.ucla.edu/other/mult-pkg/whatstat/>

**Appendix 1**

**500 Cities: Local Data for Better Health (CDC 2019)**

**More About the Project**

[](https://www.cdc.gov/500Cities/index.htm?s_cid=dph-500cities-003)

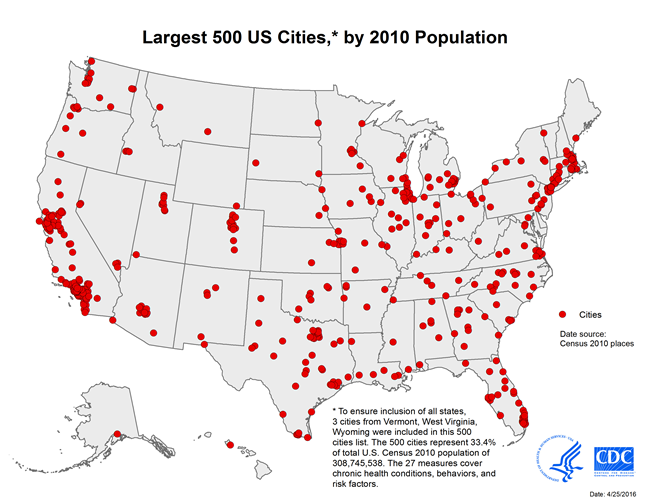
In 2015, The Robert Wood Johnson Foundation and CDC Foundation launched the 500 Cities Project in partnership with the Centers for Disease Control and Prevention (CDC). In 2018, this partnership was extended through 2020.

Project Purpose:

* This project reports city and census tract-level data obtained using small area estimation methods, for 27 chronic disease measures for the 500 largest American cities.
* The data are published through a public, interactive “500 Cities” website that allows users to view, explore, and download city- and tract-level data.
* Although limited data are available at the county and metropolitan levels, this project represents a first-of-its kind data analysis to release information on a large scale for cities and for small areas within cities. This system complements existing surveillance data necessary to more fully understand the health issues affecting the residents of that city or census tract.
* These high-quality, small-area epidemiologic data can be used both by individual cities and groups of cities as well as other stakeholders to help develop and implement effective and targeted prevention activities; identify emerging health problems; and establish and monitor key health objectives. For example, city planners and elected officials may want to use this data to target neighborhoods with high rates of smoking or other health risk behaviors for effective interventions.

Cities:

* The project will deliver data for the 497 largest American cities and will include data from the largest cities in Vermont (Burlington – population: 42,417), West Virginia (Charleston – population: 51,400) and Wyoming (Cheyenne – population: 59,466) to ensure inclusion of cities from all the states; bringing the total to 500 cities.
* The number of cities per state ranges from 1 to 121.
* The cities range in population from 42,417 in Burlington, Vermont to 8,175,133 in New York City, New York.
* Among these 500 cities, there are approximately 28,000 census tracts, for which data will be provided. The tracts range in population from less than 50 to 28,960, and in size from less than 1 square mile to more than 642 square miles. The number of tracts per city ranges from 8 to 2,140.
* The project includes a total population of 103,020,808, which represents 33.4% of the total United States population of 308,745,538.
* [List of 500 Cities pdf icon[PDF-314KB]](https://www.cdc.gov/500cities/pdf/500-Cities-Listed-by-State.pdf)



**Unique Value of the 500 Cities Project**

* The 500 Cities Project reflects innovations in generating valid small area estimates for population health.
* It provides data for cities, many of which cover multiple counties or do not follow county boundaries, and for census tracts for the first time. These data will be filterable (by city and/or tracts; as well as by measure) and downloadable for use in separate analyses by the end-users.
* The project will enable retrieval, visualization, and exploration of a uniformly- defined selected city and tract-level data for the largest 500 US cities for conditions, behaviors, and risk factors that have a substantial effect on population health.