2/24/2018

An Analysis of Unemployment in the United States

With Evidence from the 2015 Census Bureau

Anum A Zaheer & Faisal Ahmad
DATA INTERPRETING SERVICES, LLC

Contents

Introduction 2

Description of Data 2

Variables of Interest 2

Potential Issues with Data & Handling 2

Summary Statistics 3

Distribution Review & Analysis 3

Box Plot Analysis by State 4

Correlation 4

Linear Model 4

Conclusion 4

Appendix 5

Introduction

The intention of this report is to explore the 2015 Census Data and discuss the most important 1-2 insights about the data set with the support of analytic tools: SAS and R. Additional insight and support are included within the appendix.

Description of Data

The dataset was found on Kaggle.com within their public domain. This dataset contains census data from across the country of the United States from year of 2015 and it was collected by the US Census Bureau. There are a total of 3,218 observations within the dataset.

Variables of Interest

There are a total of 37 variables included in the data set, however, we will only explore 13 variables of interest within this report. Each of the variables explored and their descriptions are listed below.

Name of Variable	**Description**
State	State, DC or Puerto Rico
County	County or county equivalent
TotalPop	Total population
Hispanic	% of population who identified as Hispanic
White	% of population who identified as White
Black	% of population who identified as Black
Native	% of population who identified as Native
Asian	% of population who identified as Asian
Pacific	% of population who identified as Pacific
IncomePerCap	Income per capita ($)
Poverty	% under poverty level
Employed	% employed (Age 16+)
Unemployment	Unemployment rate (%)

Potential Issues with Data & Handling

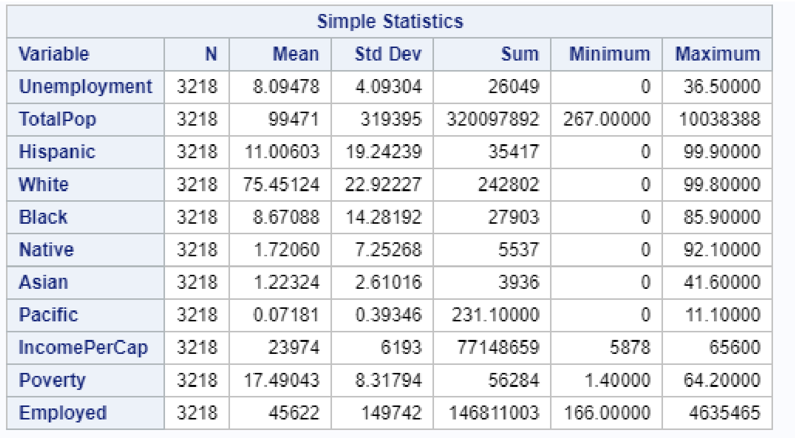
Some data or census tract lines were missing upon first glance of the entire dataset. For some values which may be unintuitive such as “IncomeErr” or “IncomePerCapErr,” we may need more information to understand the meaning of how the error was calculated and how both values can be interpreted in relation to the rest of the dataset.

To minimize problems with the data before starting the analysis, missing or null data was removed from the dataset. Any un-intuitive data which seemed unclear or vague in its description was not incorporated into the analysis.

For any potential problems of getting lost with the data due to variable repetition, this was resolved by running the code once again using distinct variables.

Summary Statistics

The overview of summary statistics for each variable of interest is listed below in Table 1. "N" is the number of observations of counties within the dataset.



Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Unemployment	3218	8.09478	4.09304	26049	0	36.50000
TotalPop	3218	99471	319395	320097892	267.00000	10038388
Hispanic	3218	11.00603	19.24239	35417	0	99.90000
White	3218	75.45124	22.92227	242802	0	99.80000
Black	3218	8.67088	14.28192	27903	0	85.90000
Native	3218	1.72060	7.25268	5537	0	92.10000
Asian	3218	1.22324	2.61016	3936	0	41.60000
Pacific	3218	0.07181	0.39346	231.10000	0	11.10000
IncomePerCap	3218	23974	6193	77148659	5878	65600
Poverty	3218	17.49043	8.31794	56284	1.40000	64.20000
Employed	3218	45622	149742	146811003	166.00000	4635465

Table 1: Means Procedure Output for Variables of Interest

By reviewing the summary statistics of the variables, some extreme values are: total population as low as 267 people in King County, Texas and as high as 10M in LA County, California. Also, the lowest rate of unemployment in the US is 0%, however the total population for the counties (Kenedy-TX, Slope-ND and Thomas-NB) with 0% unemployment are very low, less than 700 people. The highest rate of unemployment is Adjuntas County in Puerto Rico.

Distribution Review & Analysis

A histogram was created to graphically represent the distribution of unemployment throughout the US. Figure 1 below shows the unemployment distribution is somewhat symmetric or looks similar to a normal distribution shape, but the distribution is more toward the left direction (with the tail of the output at the right), therefore, the distribution is positively skewed.

By visually looking at Figure 1, the center of the distribution is at a range of 6-8%, and this can be supported by the Summary Statistics in Table 1 where it states that unemployment has a mean of 8.09% within the dataset. There are some instances where the unemployment rate is beyond 20%, but the number of occurrences as shown within the histogram are not significant.

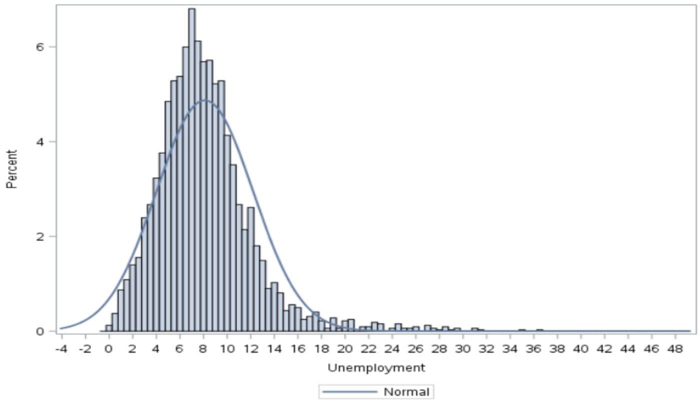


Figure 1: Histogram of Unemployment in US (2015)

Other histogram comparisons are included within the appendix of this report to explore the relationship between race and unemployment and poverty and unemployment.

Box Plot Analysis by State

Figure 2 is a graphical representation of unemployment by each state using box plots. Upon review of the box plot results, South Dakota, North Dakota, Minnesota are the least unemployed, whereas Puerto Rico and Alaska are highly unemployed.

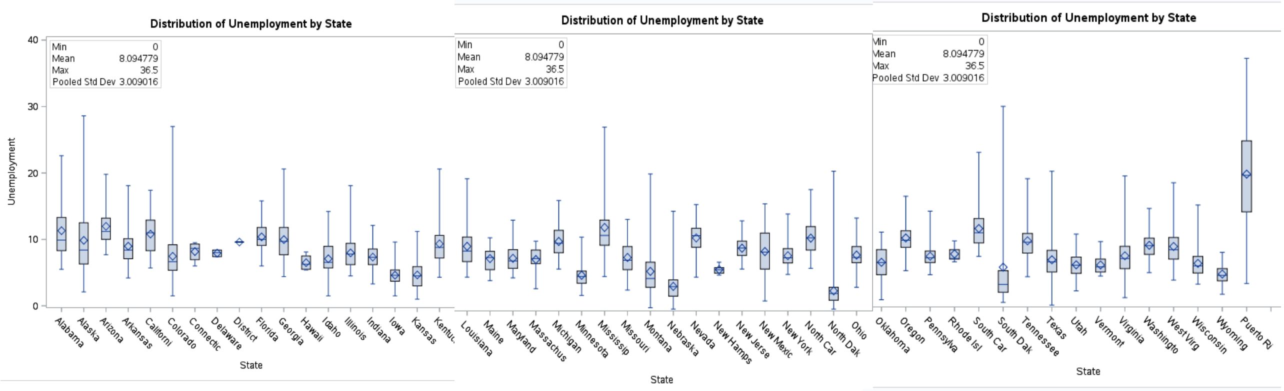
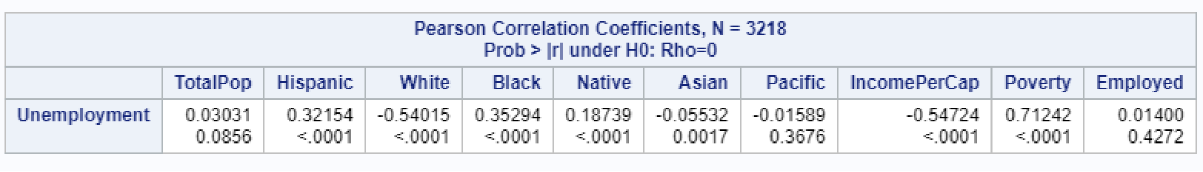


Figure 2: Box Plot Analysis Results for Unemployment by State

Correlation

By running a correlation coefficients analysis as shown in Table 2 below, we can see a positive correlation in Hispanics and blacks for unemployment. Although we may not have enough background on the possible relationship between the two variables for each race group, based on the dataset provided, the analysis is showing that there seems to be relationship between unemployment and those who identify as Hispanic or black. It may be of interest to use a variety of more measurables in order to explore the background of this observation for further support.



Pearson Correlation Coefficients, N = 3218
Prob > |r| under H0: Rho=0

	TotalPop	Hispanic	White	Black	Native	Asian	Pacific	IncomePerCap	Poverty	Employed
Unemployment	0.03031 0.0856	0.32154 <.0001	-0.54015 <.0001	0.35294 <.0001	0.18739 <.0001	-0.05532 0.0017	-0.01589 0.3676	-0.54724 <.0001	0.71242 <.0001	0.01400 0.4272

Table 2: Correlation Analysis in Race and Unemployment

As shown in the correlation coefficients table, poverty and unemployment also have a high correlation which seems to make sense logically since the two variables are likely to share a connection, interdependence or a relationship.

A similar observation for blacks, Hispanics and whites as discussed earlier can be made in the scatter plots with 95% prediction ellipses in Figure 3 (appendix). Majority of whites seem to show a lower chance of unemployment in comparison to those who identify as black or Hispanic.

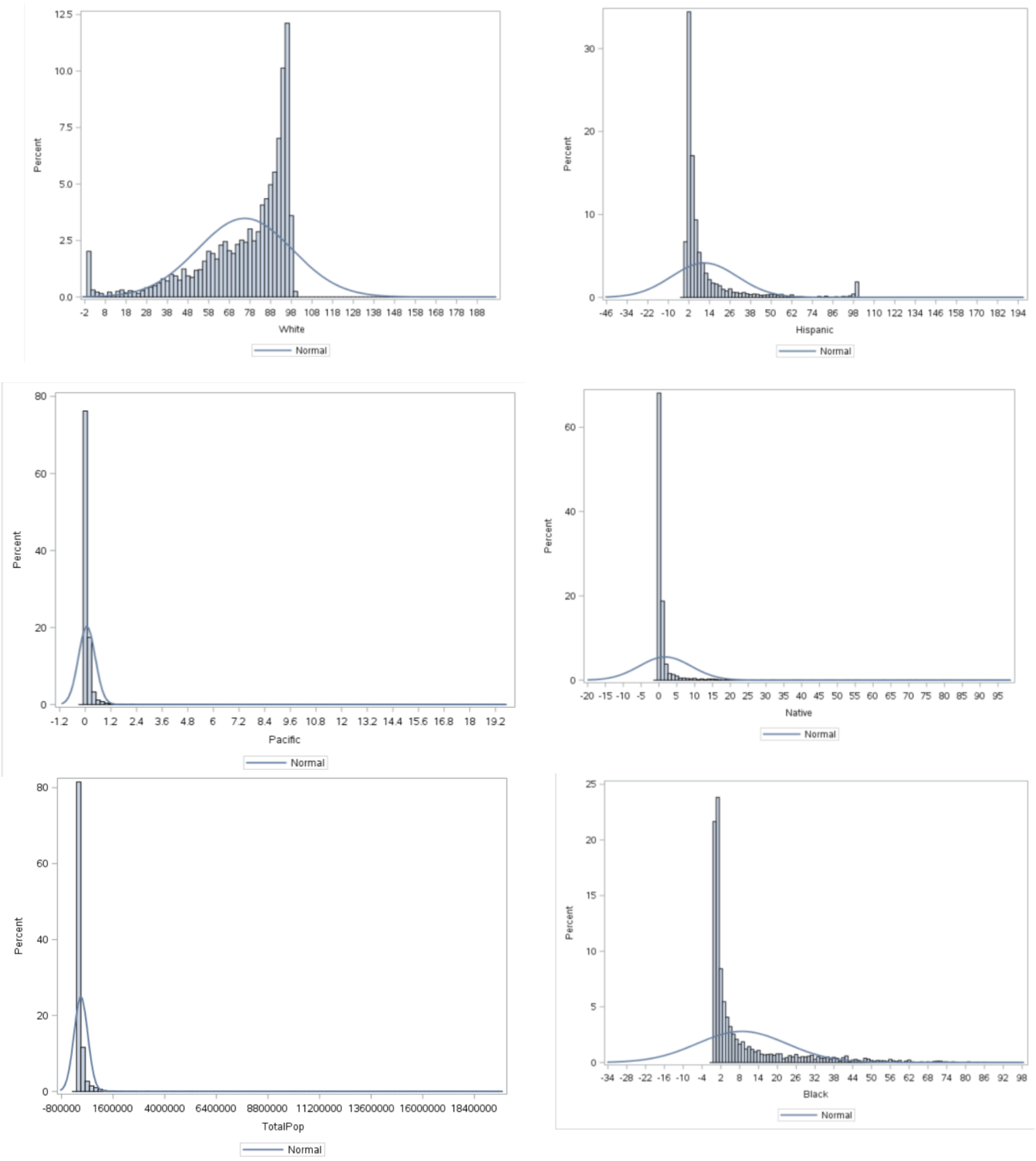
Linear Model

Based on the observations from the correlation matrix against the unemployment variable, we tried to model a multivalued linear model based on TotalPop, Hispanic, White, Poverty and Black covariates (Figure 4 in appendix): The resulting P values are very significant for all the covariates except black variables which is not that significant. The results are an adjusted R-square value of 54.16% which means that our fitted model predicts 51% of the true model.

Conclusion

From the above analysis, we found that in 58% of counties in USA have a unhealthy unemployment rate of over 6%. Our observations showed that counties with high unemployment are high in Hispanic and Black population. Puerto Rico with almost 99% Hispanic population is contributing to the high number of Hispanics being unemployed, whereas counties with less than 1% population have a 90% average white population.

Appendix



Extension of Figure 1: Exploring the data using histograms

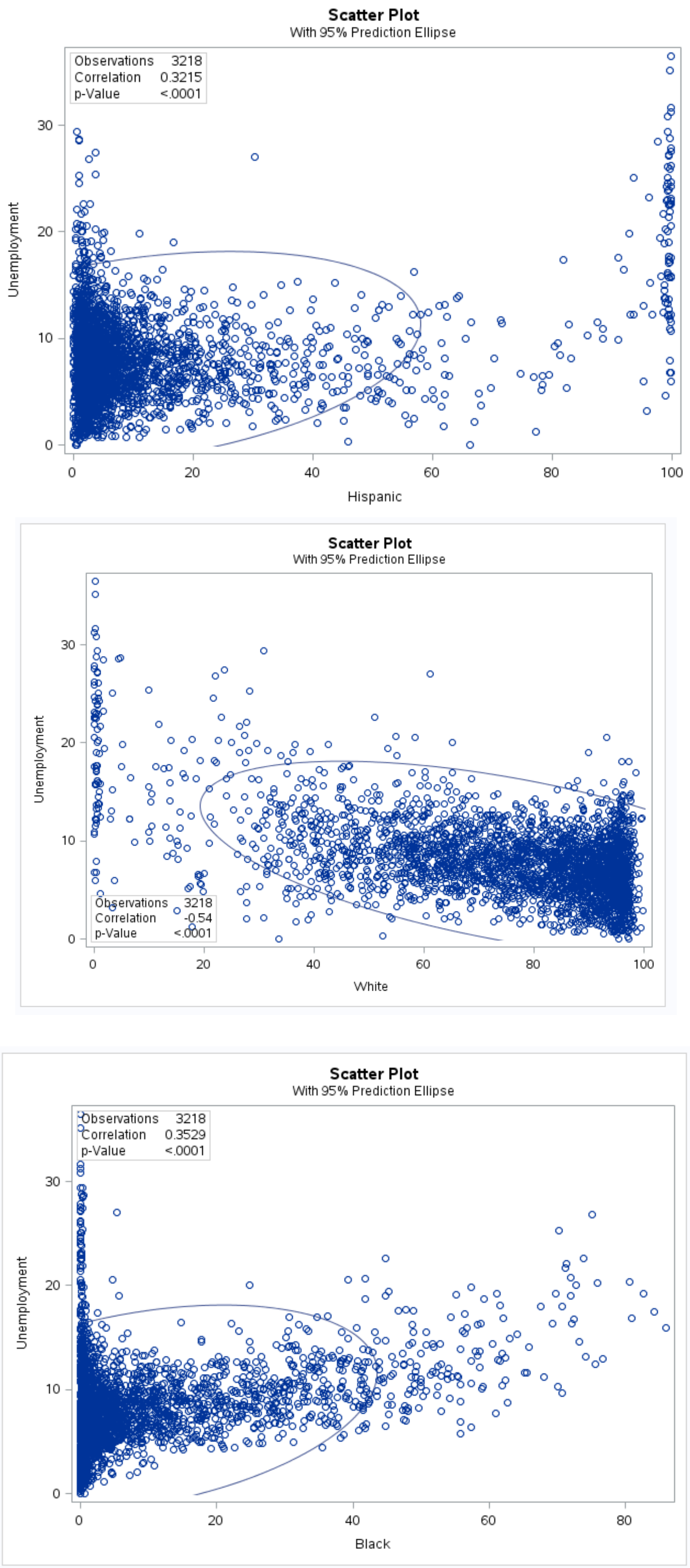
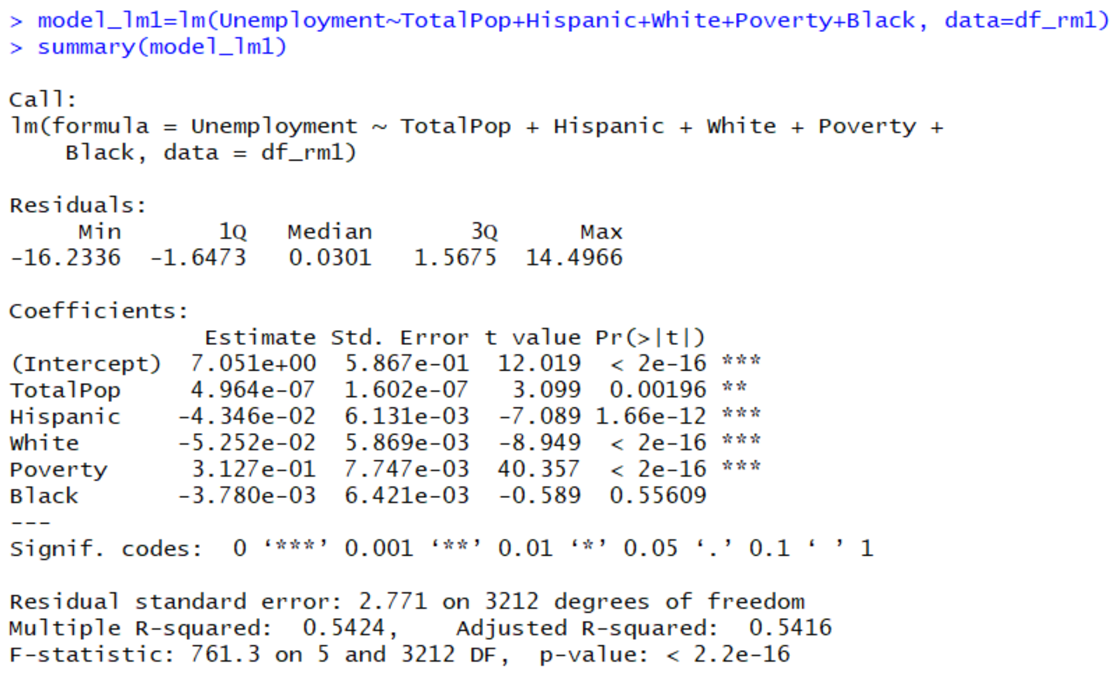


Figure 3: Scatter Plots with 95% Prediction Ellipses



```
> model_lm1=lm(Unemployment~TotalPop+Hispanic+White+Poverty+Black, data=df_rm1)
> summary(model_lm1)

Call:
lm(formula = Unemployment ~ TotalPop + Hispanic + White + Poverty + 
    Black, data = df_rm1)

Residuals:
     Min       1Q   Median       3Q      Max 
-16.2336  -1.6473   0.0301   1.5675  14.4966 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.051e+00  5.867e-01  12.019  < 2e-16 ***
TotalPop     4.964e-07  1.602e-07   3.099  0.00196 ** 
Hispanic    -4.346e-02  6.131e-03  -7.089 1.66e-12 ***
White       -5.252e-02  5.869e-03  -8.949  < 2e-16 ***
Poverty      3.127e-01  7.747e-03  40.357  < 2e-16 ***
Black       -3.780e-03  6.421e-03  -0.589  0.55609    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.771 on 3212 degrees of freedom
Multiple R-squared:  0.5424,	Adjusted R-squared:  0.5416 
F-statistic: 761.3 on 5 and 3212 DF,  p-value: < 2.2e-16
```

Figure 4: Multivalued Linear Model Results

Note:
Summary Statistics and Correlation were produced using SAS Studio
While as the rest of the results were obtained using R Studio