Diabetes Exploration

STA 6247

Alejandro Suarez

Giancarlo Bambaren

**Background**

Diabetes is one of the most prevalent chronic illnesses affecting Americans today. Both Type I and Type II diabetes result in metabolic derangement in patients that include the inability to metabolize glucose (1).

It is a leading risk factor for a variety of illness including: heart disease, stroke, blindness, and kidney disease (2). According to the Centers for Disease Control (CDC) an estimated 30.4 million Americans are diabetic (9.4% of the US population) with an additional 84.1 million (33.9%) estimated to be suffering from pre-diabetes (3). A CDC report from 2017 also estimated that an estimated $245 billion dollars were spent on treating diagnosed patients (an average of $10,600 per patients) (4) while a University of Chicago report referenced in a CNN article estimated that those costs would rise to $336 billion by 2034 (1).

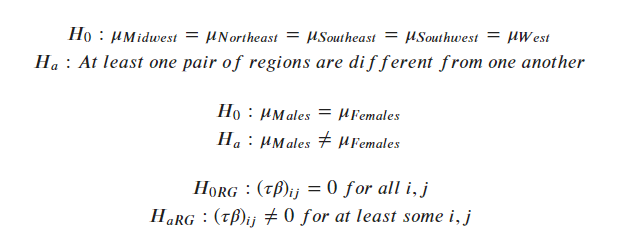
The purpose of this project is to highlight some of the disturbing trends in diabetes diagnoses across the United States over the past 20+ years and their manifestations across geographical regions and demographic groups.

We split the project into two parts. Part I consists of two two-factor ANOVA tests which test rates of diabetes between geographic and demographic categories while Part II consists of a time series analysis of those rates over the past 20+ years.

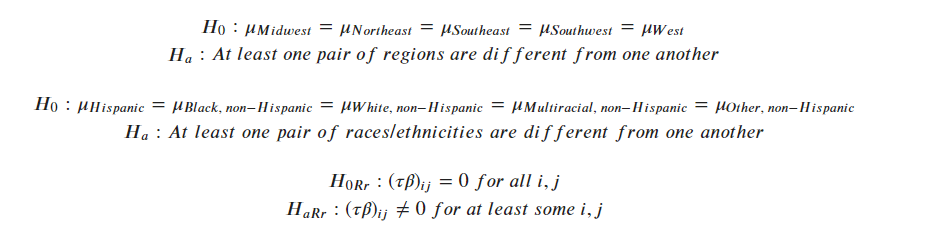
**Part I (ANOVA testing):**

We are testing whether there is a significant difference in the populations’ diabetes percentages between different regions in the United States based on gender. The second case being whether there is a significant difference in the US regions populations’ diabetes percentages based on race/ethnicity. We also sought to find whether there is a difference between each gender’s population diabetes percentage, and whether there is a difference between each race and ethnicity population diabetes percentage. Each of these two cases stated under the following hypothesis statements:

Case 1:



Case 2:



**Part II (Time series analysis):**

We extracted CDC data of the rates of diabetes for every state in the United States from 1994-2016 and performed a time series analysis of those rates for each geographic region. We also predicted the rates of those regions for 2017-2026 using Double Exponential Smoothing.

**Data Extraction and Pre-Processing**

**Part I:**

The data for this part originated from an exported CSV file that was queried using an application hosted on the CDC website (5). We could query very specific data values with names such as “Visits to dentist or dental clinic among adults aged >= 18 years with diagnosed diabetes” or “Adults with diagnosed diabetes aged >= 18 years who have taken a diabetes self-management course”, but seeing as this was a general, exploratory report, we focused on the rates of diabetes which is the “Prevalence of diagnosed diabetes among adults aged >= 18 years” statistic.

This statistic was available from 2011-2016 for every U.S. state and territory and was subdivided into gender (male/female) and ethnic (Black/Multiracial/Hispanic/White/Other) categories. Since the U.S. territories lacked adequate data for each of these categories for each of the years we decided to exclude these territories from our analysis and focus on the 51 U.S. states.

The final dataset loaded into R was (2177 x 11) values.

**Part II:**

The data for the time series portion of our analysis came from a CSV file from the CDC. Each row consisted of a state while each column consisted of data multiple values for each year from 1994-2014. In other words, each year consisted of: ”Number”, “Age-adjusted Percent”, “Age-adjusted Lower Confidence Limit”, “Age-adjusted Upper Confidence Limit”, “Age-adjusted Obesity Percent”. For our analysis we focused solely on the “Age-adjusted Percent”. In order to add the values for 2015 and 2016 we took the values from the dataset in Part I for those years and joined them accordingly for each state. There was no data cleaning involved, only deleting columns from the original CDC dataset. The final dataset that was loaded into R had (53 x 26) values.

**Final Datasets and Variables**

When doing the final pre-processing of our datasets we had to keep in mind the type of tests that were going to be made in order to test our hypothesis as well as the variables that were needed for each specific test.

In our first case hypothesis, the variables of interest were Regions (a factor), which includes the five regions in the United States (while excluding territories such as Guam and Puerto Rico since we are only focusing on the US mainland), as well as Gender (a second factor) for Male and Female. Lastly we also included DiabetesPercentage, which is the treatment variable of interest. In this case we ended up using Python to find the mean of each treatment across all the years (2011-2016) after grouping them by Region and Gender.

[Figure 1 in Appendix]

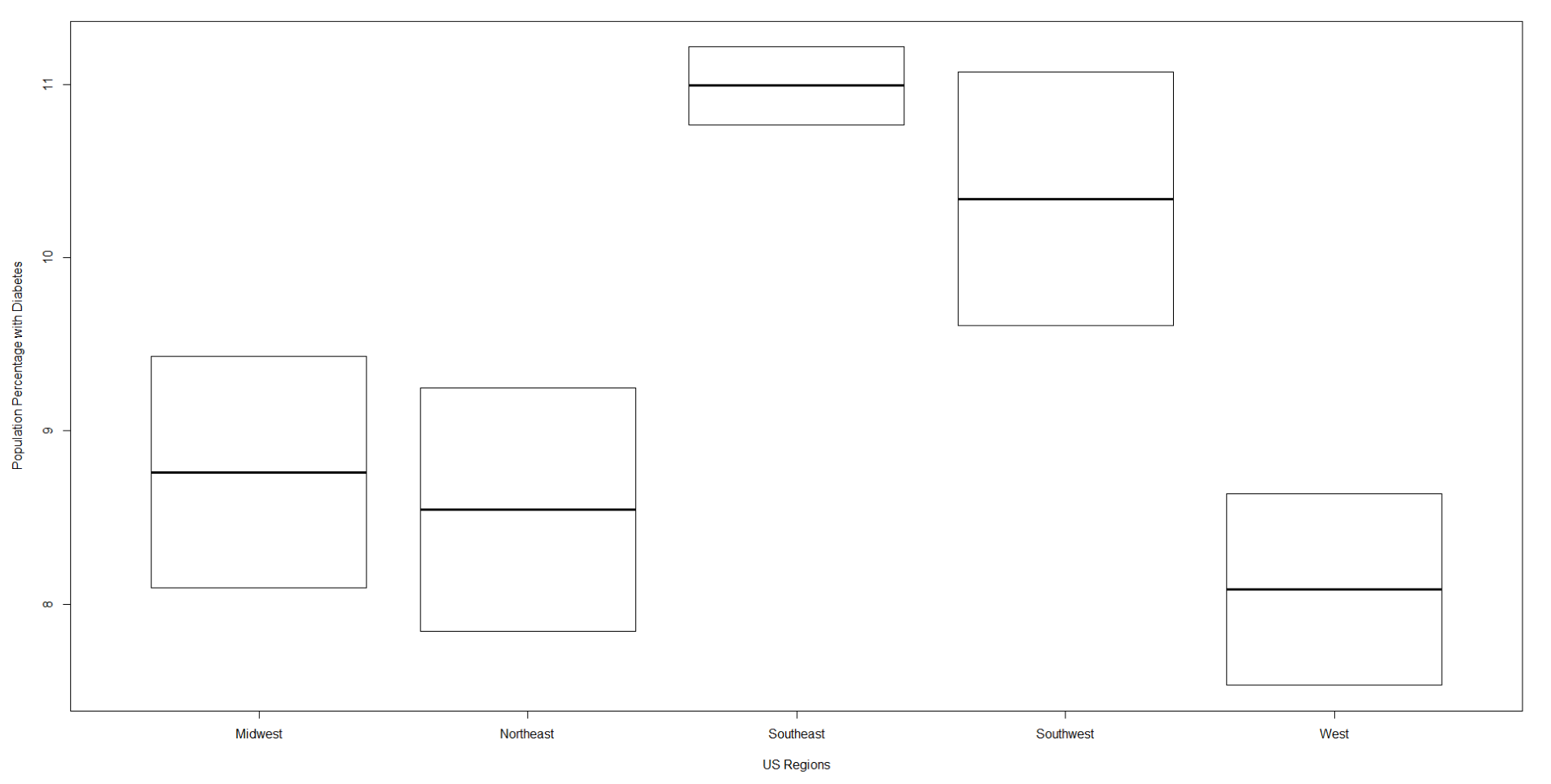
In our second case hypothesis, the variables of interest were Regions again, but this time Gender is replaced by RaceEthnicity (the second factor). RaceEthnicity is further broken down into Hispanic, Black non-Hispanic, Multiracial non-Hispanic, Other non-Hispanic, and White non-Hispanic. The treatment variable again is DiabetesPercentage. The same method in Python as in the first case hypothesis was used for the second case hypothesis.

[Figure 2 in Appendix]

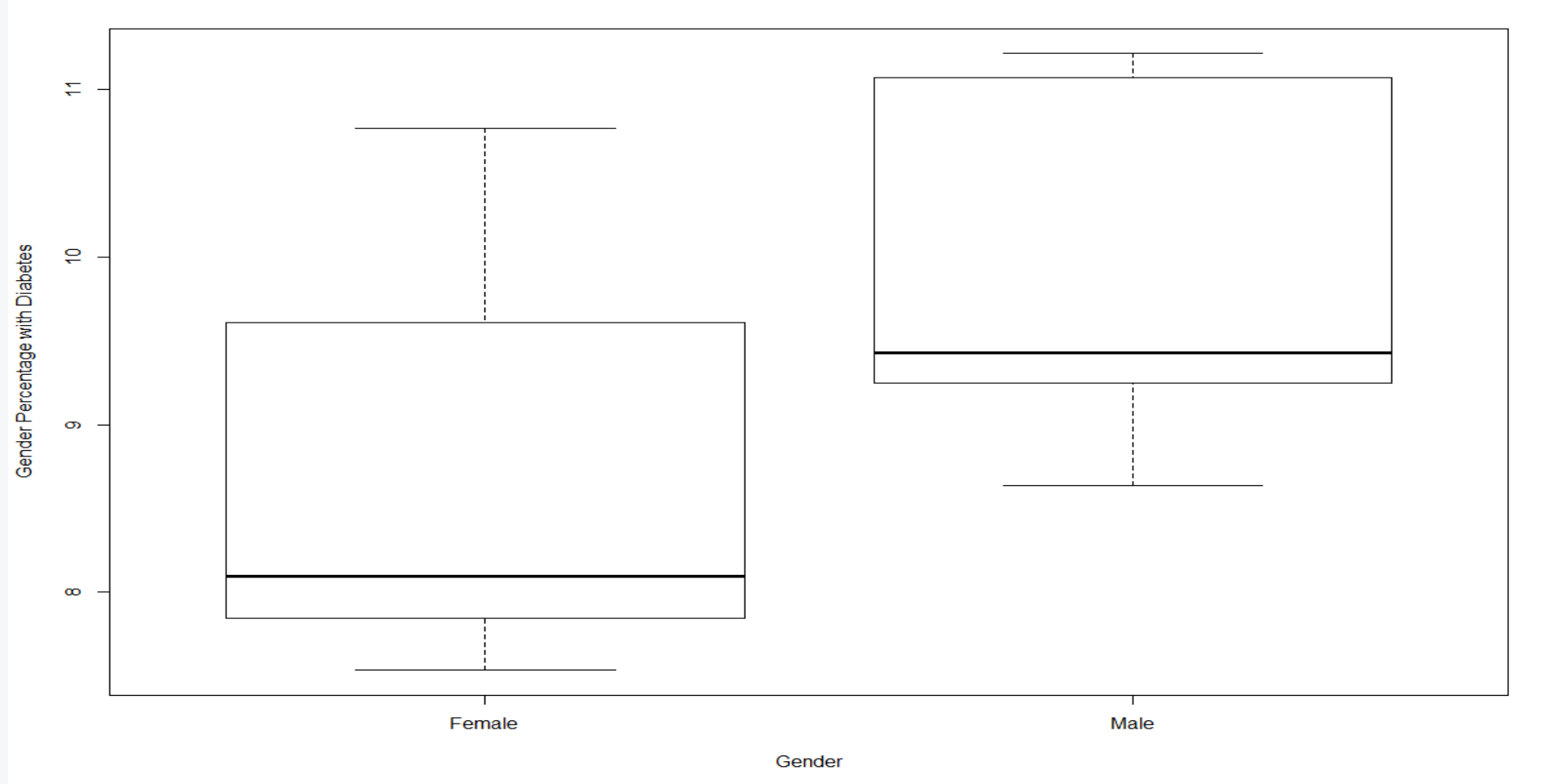
**Hypothesis Testing and Analysis**

**Case I**

To test the first and second hypothesis of our first case, we first plotted our factors by diabetes percentages using boxplots in order to see if we could identify a potential difference within groups.

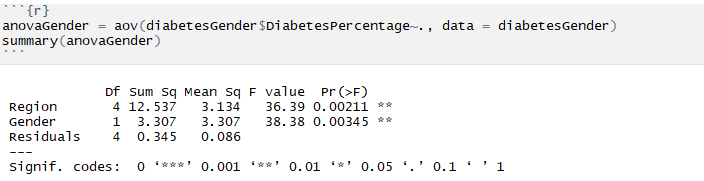


As seen by the boxplot of our Regions in the first case, it seemed as if there was a significant difference between some of the regions, with the southeast being the region of most difference, and southwest being a potential second region that has a significantly different percentage of population with diabetes.



When performing a boxplot of both genders, it didn’t seemed like there could be a significant difference between the percentage of population with diabetes of both genders. However we did observe that both genders were positively skewed.

After this initial observation, it was time to run the ANOVA test on both factors. The ANOVA test gave us the following p-values for both factors which seemed to reject the two first null hypothesis of our first case: Region 0.002, and Gender 0.003.



When plotting the ANOVA to check for our assumptions of variance and normality however, we could observe that although our residuals followed a normal distribution, the variance seemed like it could be potentially violated. A Levene Test on both factors would then be used to test for equal variance.

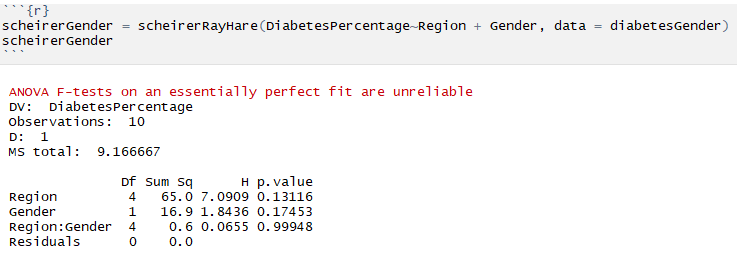
[Figure 3 in Appendix]

Although, the Levene test returned a p-value of 0.8511 for the Gender factor (meaning we could assume equal variance), it returned an extremely low p-value of 2.2e-16 for the Region factor, meaning it violated the equality of variance assumption.

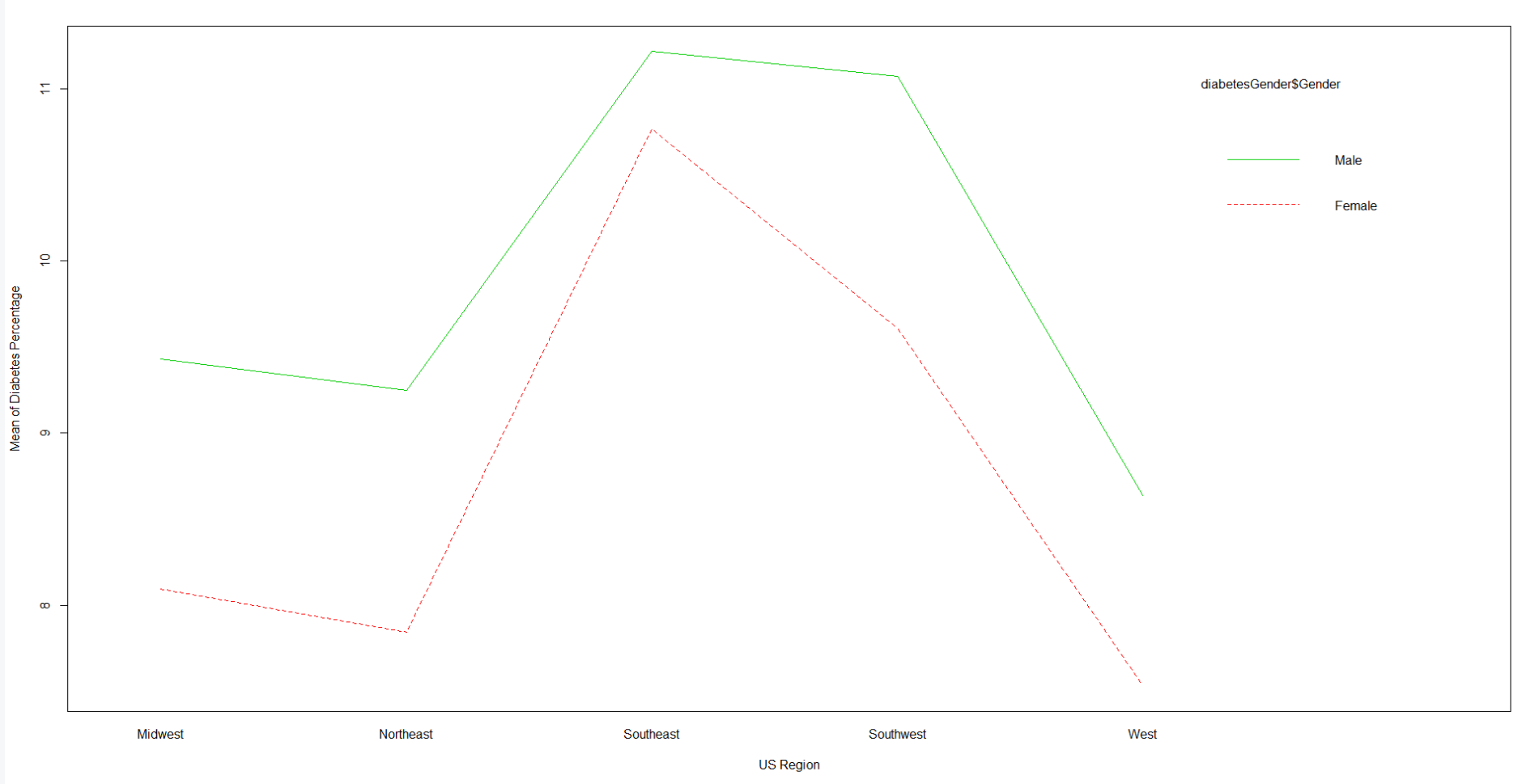
[Figure 4 in Appendix]

[Figure 5 in Appendix]

With the variance assumption violated, we could not assume the ANOVA results to be reliable. As a result we opted to use the Scheirer-Ray-Hare test (a multifactor extension of the Kruskal-Wallis) as the non-parametric function in order to re-test our hypothesis under a more reliable setting.

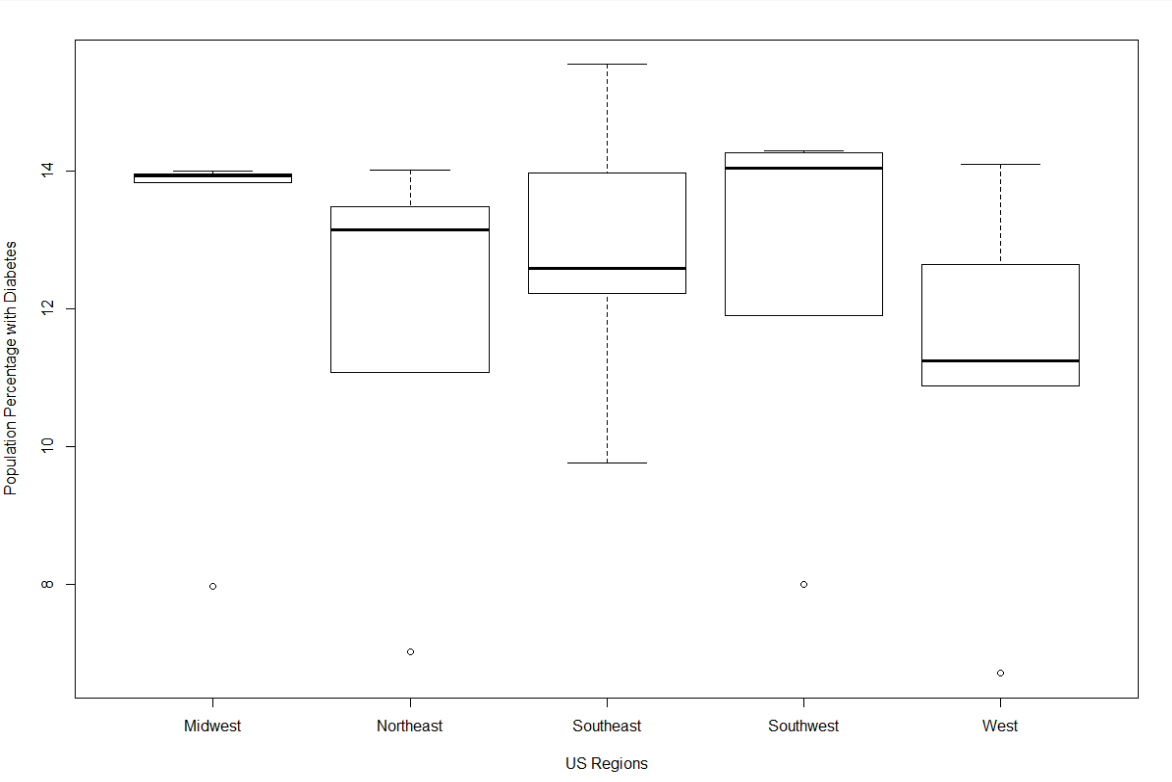


Based on the p-values obtained under the Scheirer-Ray-Hare test of 0.1312 for Region and 0.1745 for Gender, it was safe to conclude there is no significant difference in the means of diabetes percentages between regions and between genders. The Gender p-value backing our initial assumption from the boxplots. Due to the p-values, no post-hoc analysis was performed. The Scheirer-Ray-Hare test also provided us with an interaction p-value of 0.9995, rejecting our third null-hypothesis in our first case. This is further proven by plotting the interaction between the Region and Gender variables.

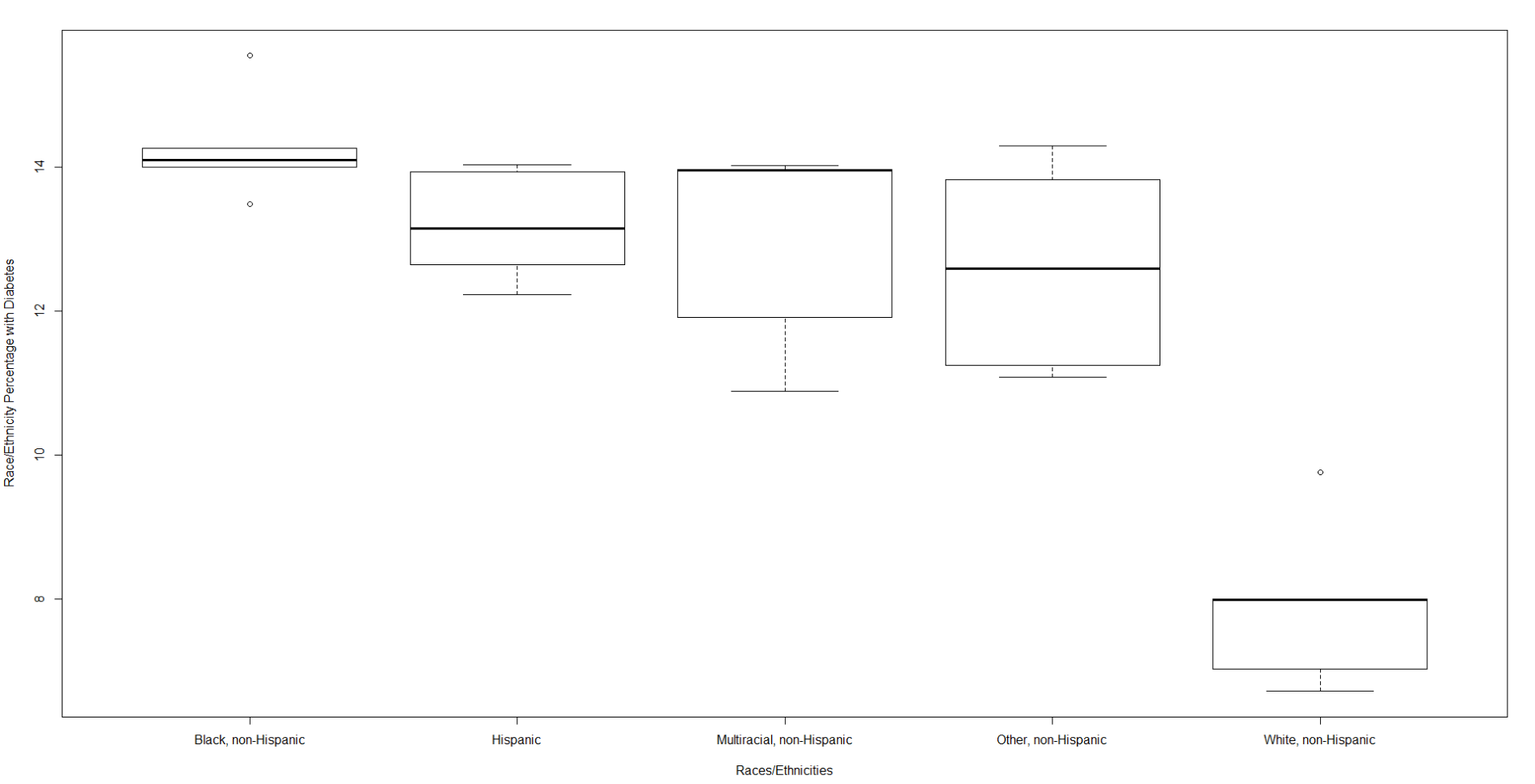


**Case II**

Just as in the first case, we started this second case analysis by plotting boxplots of our factors by diabetes percentages in order to identify any potential differences within these groups.

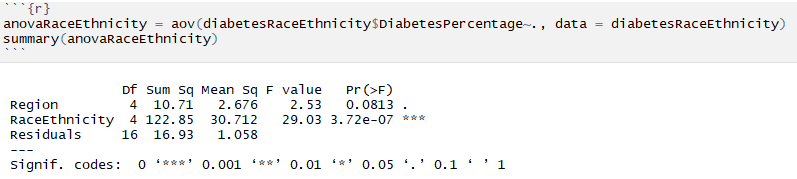


As seen in the boxplot of Regions in this case, we can observe no significant difference between the regions in terms of population percentage with diabetes. Given this, we expect a p-value greater than 0.05 from our ANOVA test.



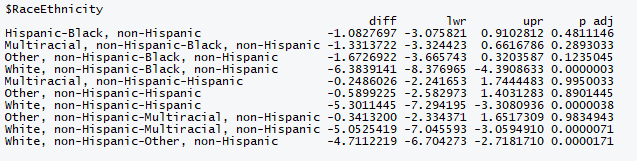
The contrary applies to the boxplot of Race/Ethnicity. Based on this boxplot, the initial expectation is that white, non-Hispanics, have a far lower diabetes percentage from all the other Races/Ethnicities. A significant p-value is expected to be obtained from our ANOVA test, and we expect this to be the case as well during our post-hoc analysis.

When running the ANOVA analysis for this case, as expected, we observed a p-value for Region of 0.0813 which is greater than our alpha level of 0.05. If all ANOVA assumptions were to be held, we would fail to reject our first null hypothesis. Our significant p-value of 3.72e-07 for Race/Ethnicity also agreed with our initial assumptions from the boxplot that there is a significant difference between at least one pair of races/ethnicities. If this holds, then we would reject our second null hypothesis.



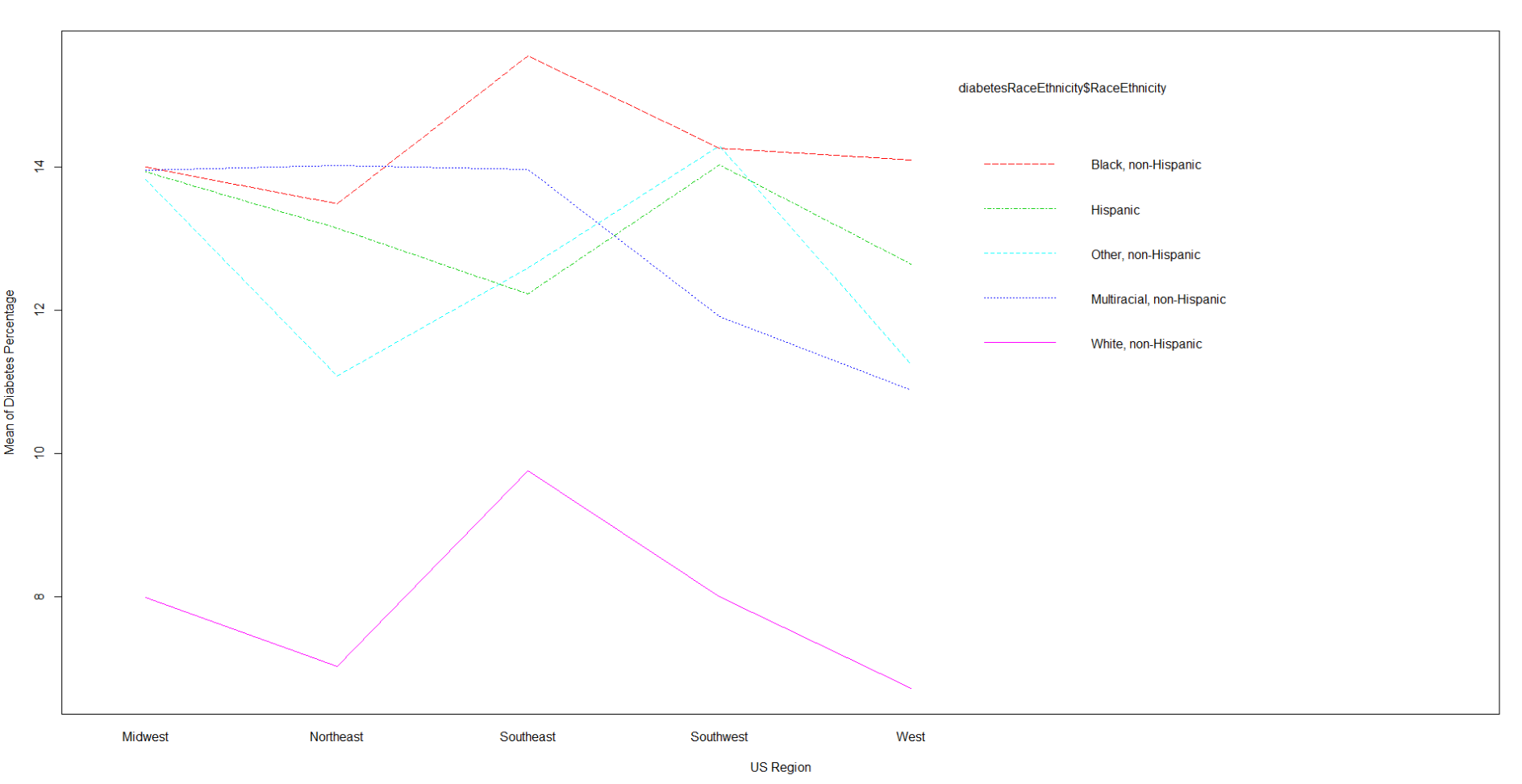
After plotting our ANOVA, we observed a potential equal variance assumption just as in the first case. However, when running the Levene Test for both Region and Race/Ethnicity, we obtained high p-values of 0.99 and 0.6909 respectively. This meant our ANOVA assumptions were correct, failed to reject our first null hypothesis, rejected our second null hypothesis, and were ready to start a post-hoc analysis using TukeyHSD.

[Figure 6 in Appendix]



Our post-hoc analysis backed up our boxplot assumption that the White, non-Hispanic group was significantly different from the rest of the races/ethnicities. This is because all of the difference in pairs for the White, non-Hispanic group have significant p-values far lower than our alpha level of 0.05. The specific pair can be seen in the above table.

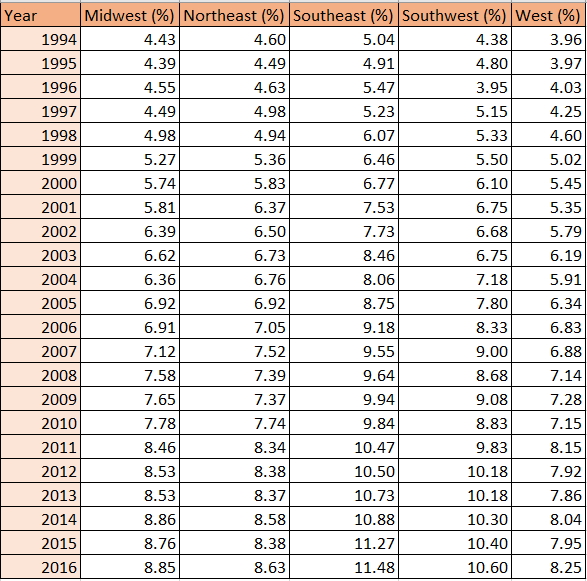
As for our third hypothesis test, we concluded there was an interaction between Region and Race/Ethnicity. This is apparent when looking at the interaction plot below. Therefore we were able to reject our third null-hypothesis. Unfortunately due to the high degrees of freedom of 16, and the small dataset size, a p-value could not be obtained from our ANOVA function for the interaction effect.



**Time Series Plot**

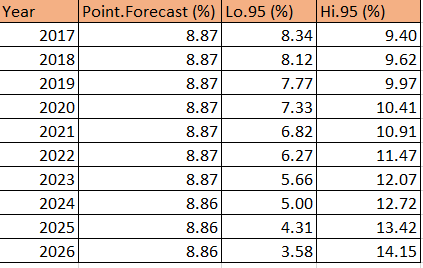
For each U.S. region (Northeast, Southeast, Southwest, Midwest, West) we plotted the rates of diabetes across time (1994-2016) and used Double Exponential Smoothing to forecast the rates from 2017-2026 using the “forecast” package. In addition to point estimates for the next 10 years of diabetes rates, we generated 95% confidence intervals for these years.

Below we have the average diabetes rates per region.

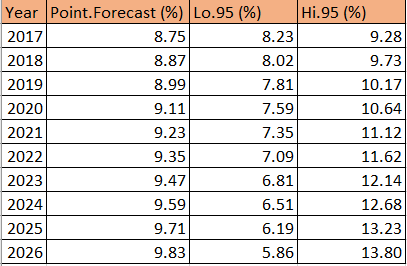


Below are the point estimates of diabetes rates for each region over the next 10 years with (95%) confidence intervals.

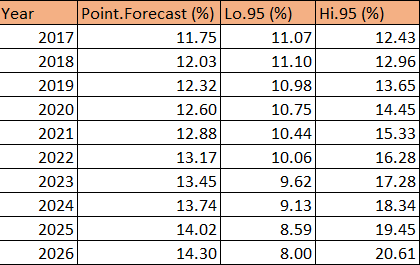
**Midwest**



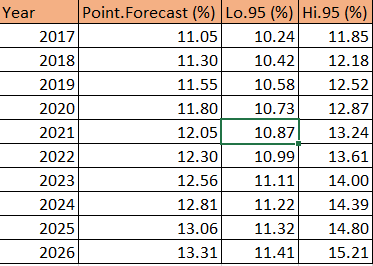
**Northeast**



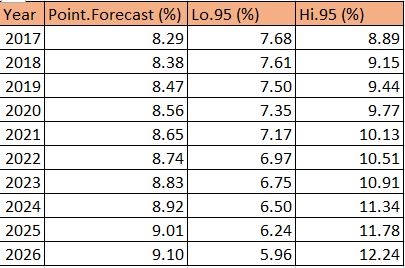
**Southeast**

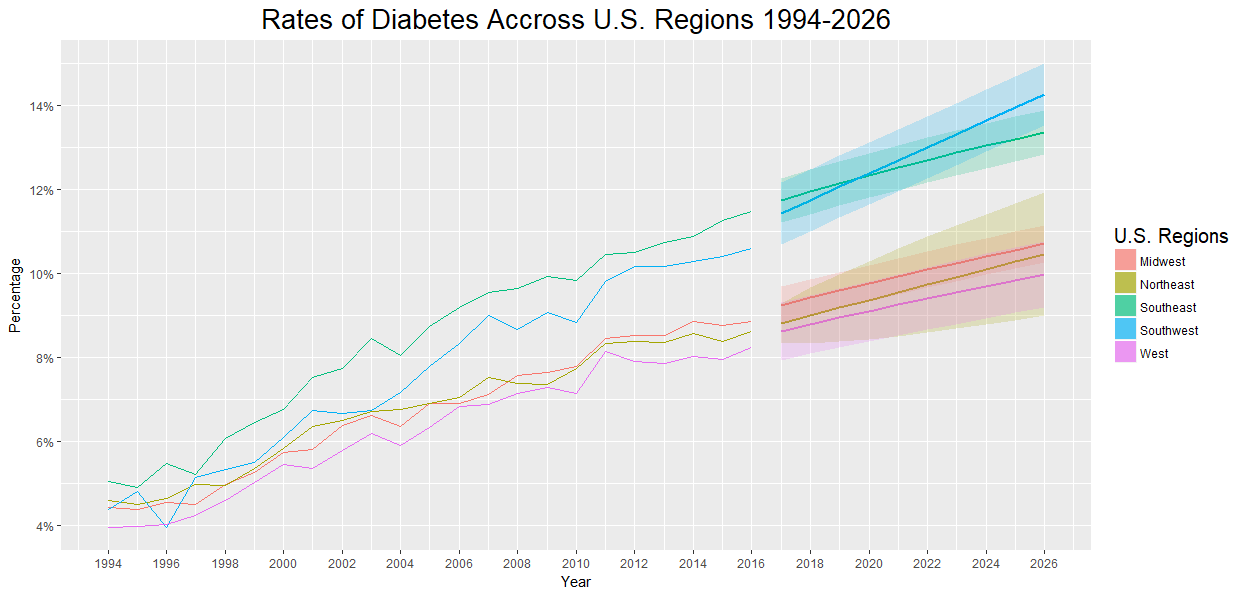


**Southwest**



**West**





There is a very clear, disturbing upward trend in the rates of diabetes across all regions in the U.S. The estimates confidence intervals over the next 10 years indicate that the Southwestern states will surpass the Southeastern states in diabetes rates. The Northeastern, Western and Midwestern states appeared to subside in their growth rates over the last 6 years but are still anticipated to grow again.

**Conclusion**

From our ANOVA tests on our two cases we can see that there is no significant difference in diabetic rates among the different US regions. Both cases concluded this in the failure to reject the first null hypothesis. This shows that the diabetes epidemic is affecting the US mainly as a whole, and is not concentrated in specific regions. There is no indication that gender has a relationship with a person’s propensity to be diabetic, as there is no significant difference between males and females based on our second hypothesis conclusion in the first case.

However, there does seem to be a big difference between races/ethnicities with White, non-Hispanics, having the lowest rate of diabetes. The present interaction effect between race/ethnicity and region, indicates that a region’s race’s concentration has a high impact in determining what that region’s population diabetic percentage is. It would be interesting to further explore why there is such a big difference between White, non-Hispanics and every other race/ethnicity. Taking into consideration each race/ethnicities traditional diet composition could provide us with an important window into this significant difference.

Although this was an initial, exploratory analysis this analysis could be enriched with more detailed data at the demographic level. We only had access to Gender/Region or Ethic Group/Region data, having diabetes rates of each of these categories broken down into subcategories would have given us an opportunity to perform multi factor analysis across more than two categories. Nonetheless, the rates of diabetes are alarming and their projected rates into the future show no sign of slowing down.

**Sources**

(1) <http://www.cnn.com/2009/HEALTH/11/26/diabetes.projections/>

(2) <https://www.healthline.com/health/type-2-diabetes/statistics#6>

(3)<https://www.cdc.gov/diabetes/data/statistics/statistics-report.html>

(4) <https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf>

(5) <https://chronicdata.cdc.gov/Chronic-Disease-Indicators/U-S-Chronic-Disease-Indicators-Diabetes/f8ti-h92k>

**Appendix**

Figure 1 First Case Final Dataset:

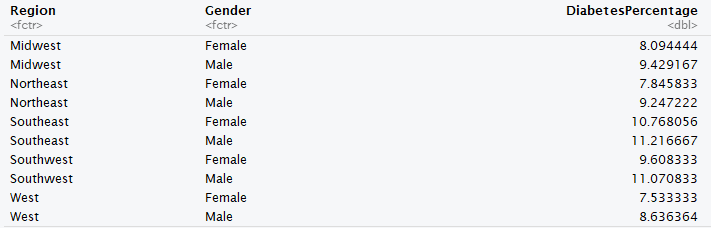


Figure 2 Second Case Final Dataset:

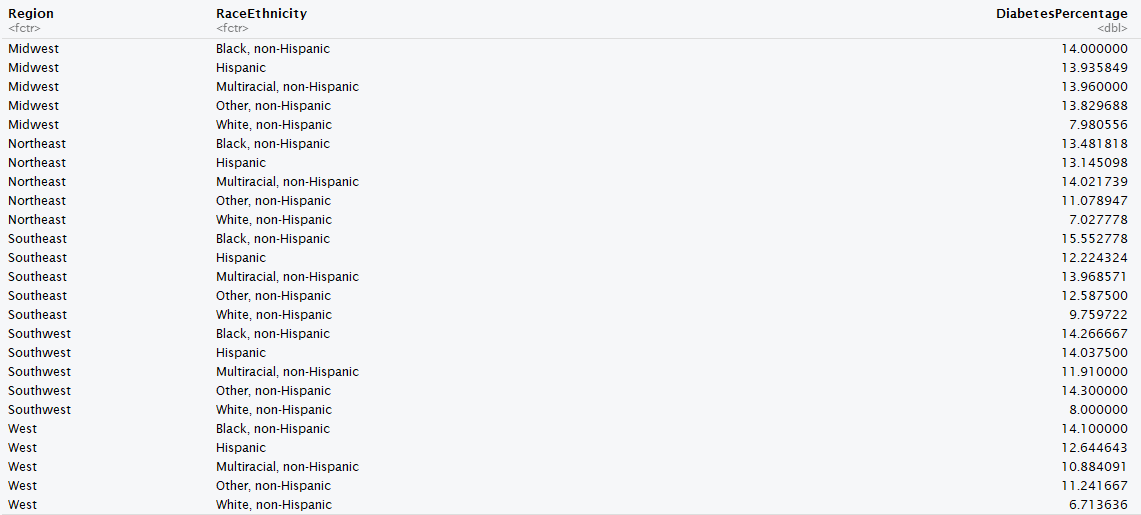


Figure 3 First Case ANOVA Residual Plots:

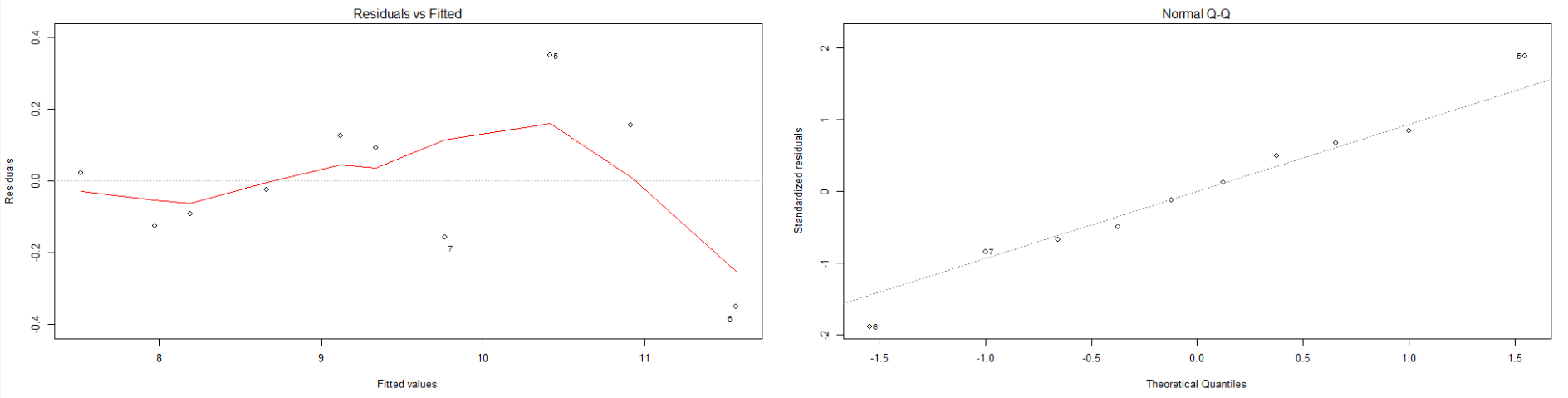


Figure 4 Levene Test for Region in First Case:

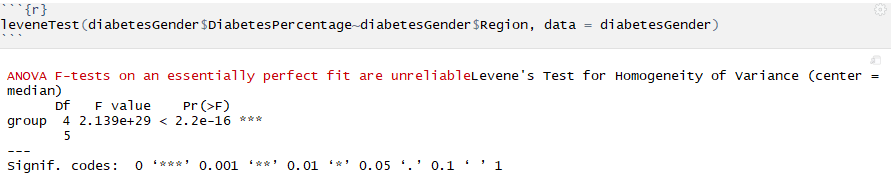


Figure 5 Levene Test for Gender in First Case:

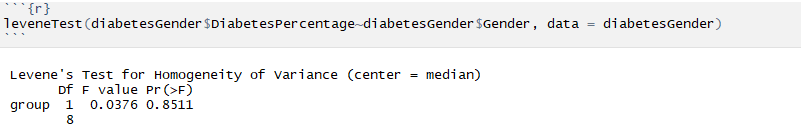


Figure 6 Second Case ANOVA Residual Plots:

