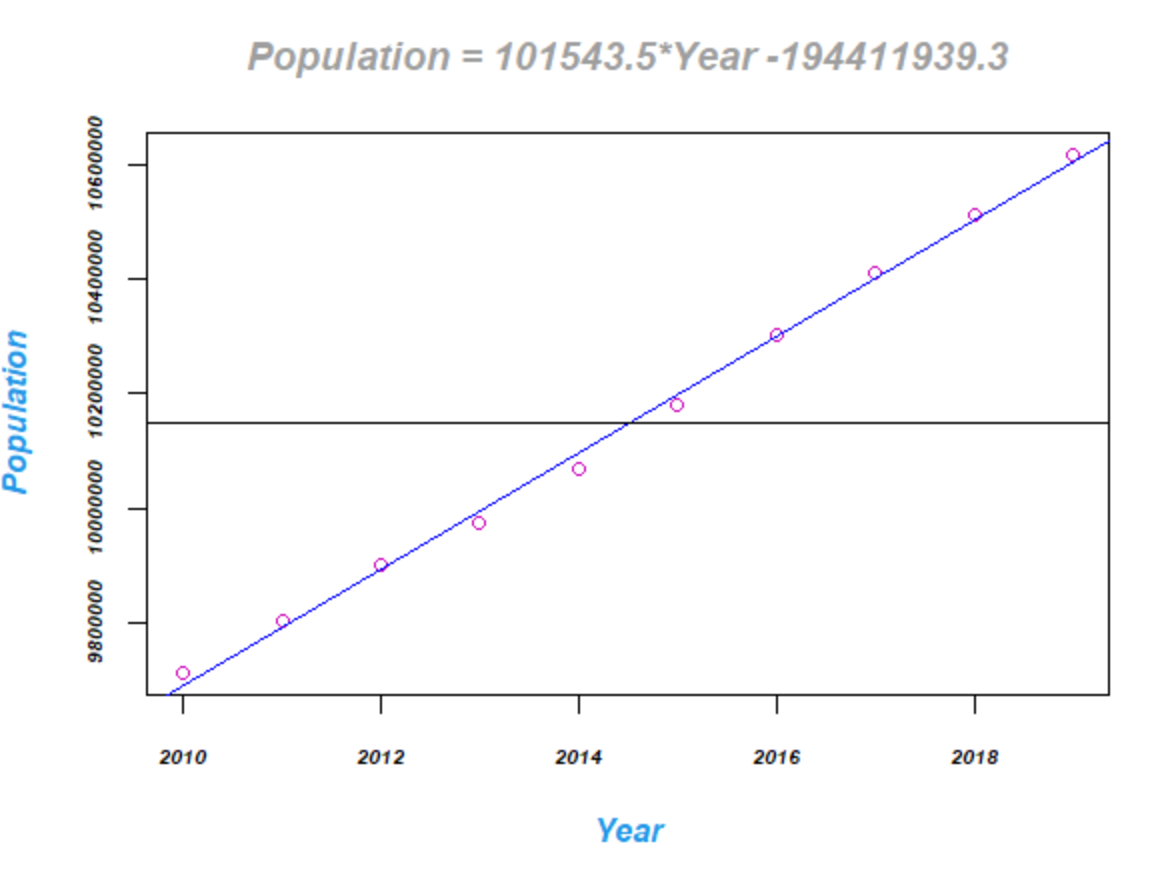
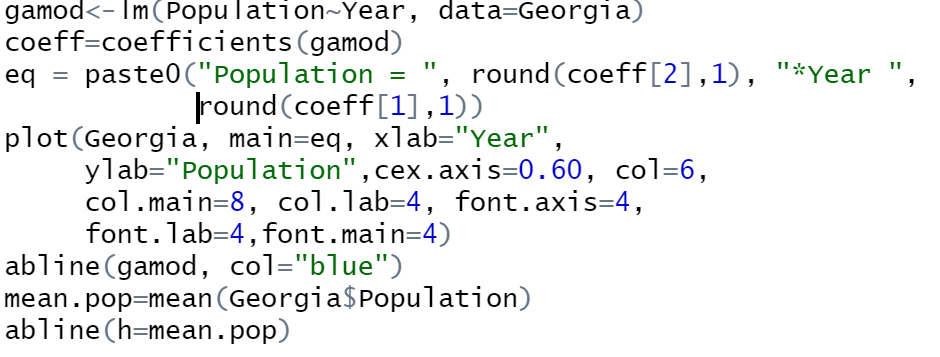
**Linear Regression Model to predict the Population of Georgia in 2025**

**Tehmina A Dangra**

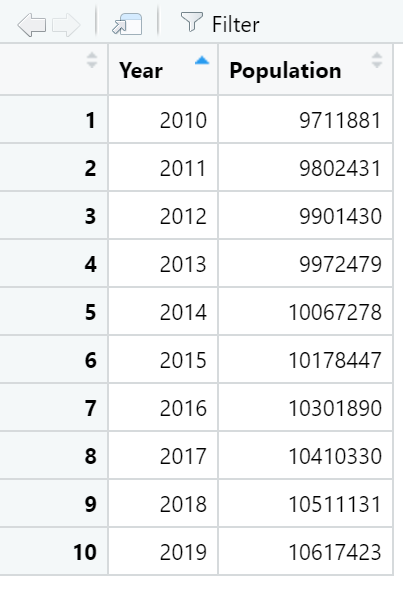
**Task 1 C997**

**Linear Regression Model for the State of Georgia to Predict the Population in 2025**

This paper will explain the use of R to make predictions about the population for the state of Georgia in five years (2025). A simple linear regression model was used to predict the population for the state of Georgia.

The data was cleaned and imported in R, to predict the population of Georgia in five years. The lm() function of R was used to create the linear regression of the given data set. The plot() function was applied to plot the linear regression and the abline() function was used to show the line of regression and the baseline mean population of Georgia from Year 2010-20.

The table shown below was imported in Rstudio. The Current Estimates Data dataset (nst-est2019-01.xlsx) was extracted from the U.S. Census Bureau and saved as ‘Georgia.997.xlsx’. The Excel file ‘Georgia.997.xlsx’ has the population estimates for every state from the year 2010-2019 on the sheet ‘NSTO1’. It is evident from the given data that the last census was in April 2010. It is also apparent that the data stored is in the wide format. The data extracted pertaining to the state of Georgia is on a new sheet named ‘Georgia’ to alleviate the noise. Since the data from the Census Bureau was in wide format by default, to make it compatible with R, the structure of the data was converted to the long format using the pivot table in Excel, which is stored, on the sheet ‘Georgia\_Long’. The table from the sheet ‘Georgia\_Long’ was saved in a new Excel file named ‘Georgia.xlsx’ so it could be imported in R.

The saved manipulated data file in R folder on the computer. The file was imported into Rstudio using the option File->Import Data set -> From Excel. The table shown below was imported in Rstudio.

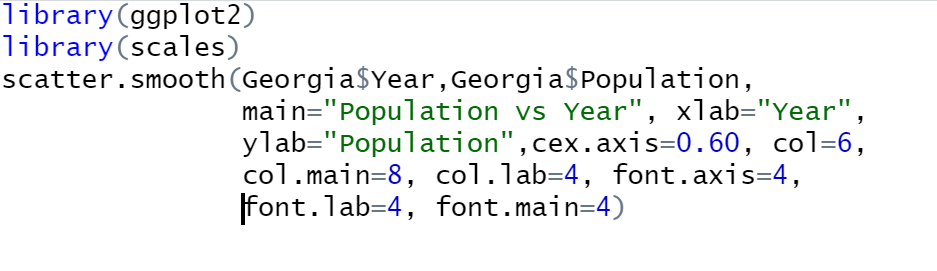
**Linear Regression:**

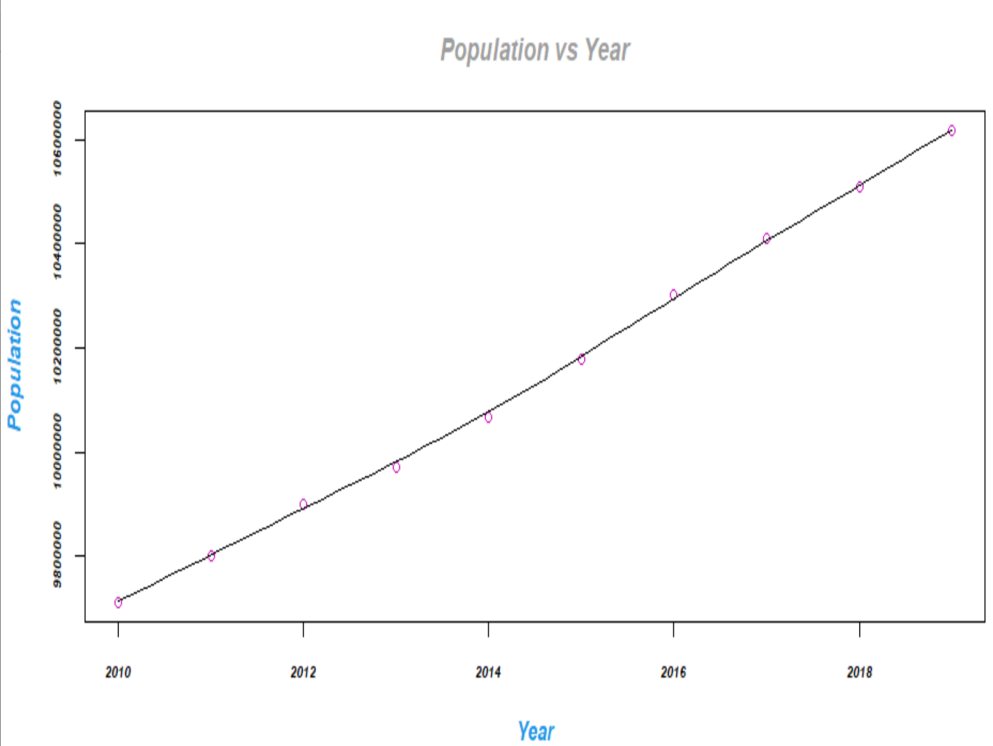
Linear Regression can be an adequate statistical model candidate to predict the population for the next five years. However, to ensure that linear regression is an acceptable choice to model the given data, it must meet the four assumptions of Linear regression, “Building a linear regression model is only half of the work. To be usable in practice, the model should conform to the assumptions of linear regression” (Prabhakaran, 2016).

1. Linearity: The relationship between the population in GA and the year must be linear. The linearity of the data can be verified using scatter plots.
2. Constant Variance: The variance for residuals is the same for any ‘x’ value. Homoscedasticity can be verified from scatter plots and residual plots.
3. Independence: The observations in the data must be independent of each other. The residuals plots can be used to check for independence.
4. Normal distributions: The error terms in the data must be distributed normally. The normality can be verified using histograms or density plots.

**To verify linearity:**

A scatter plot was created to check for linearity in the data given using the code below in R.





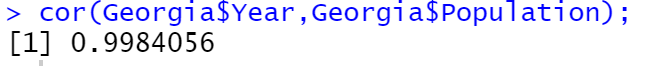
It is evident from the graph plotted above that the data satisfies the linear assumption for linear regression. The scatter plot also helps identify two important aspects of data.

1. It gives an insight that the population has a positive correlation with Year.
2. It is evident from the graph that there is no presence of outliers in the data.

However, these two statistical aspects of the data can be further investigated before applying the linear regression.

**Correlation:**

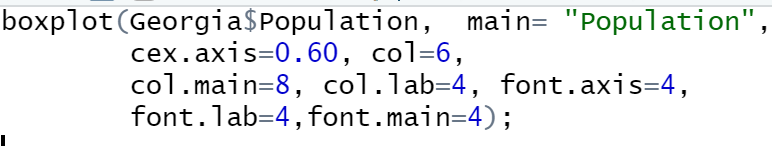
The degree of correlation between the year and the population will help us understand the trend of data. The cor() function was used to identify the strength of the relationship between variables, and the correlation coefficient can be observed below:

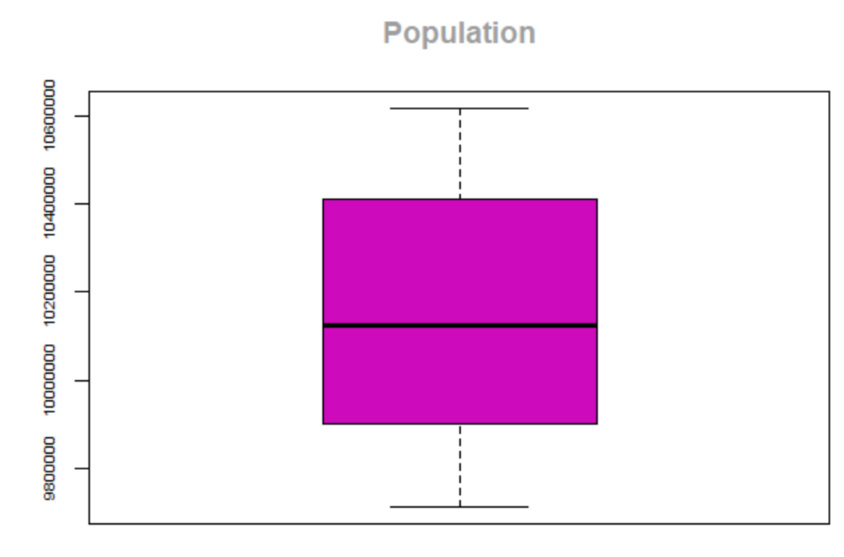


The correlation coefficient suggests a strong positive relationship as it is tremendously close to +1; this means that as the number of years increases, the population in the state of Georgia also increases.

**Presence of outliers:**

The presence of outliers can significantly affect the regression model; therefore, it is necessary to check for the presence of outliers. Though the graph shows no presence of an outlier, it can be further verified with the box plot. The boxplot() function was used to create the boxplot shown below:

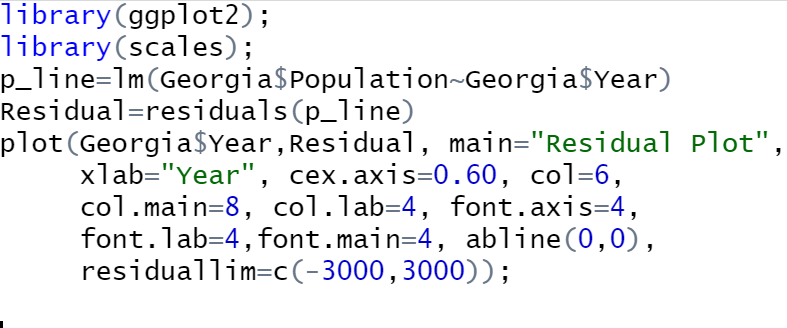




The box plot above confirms that outliers are not present in the data.

**To verify Constant Variance and independence:**

Residual plots were created using the residual function in R as shown in the code below:

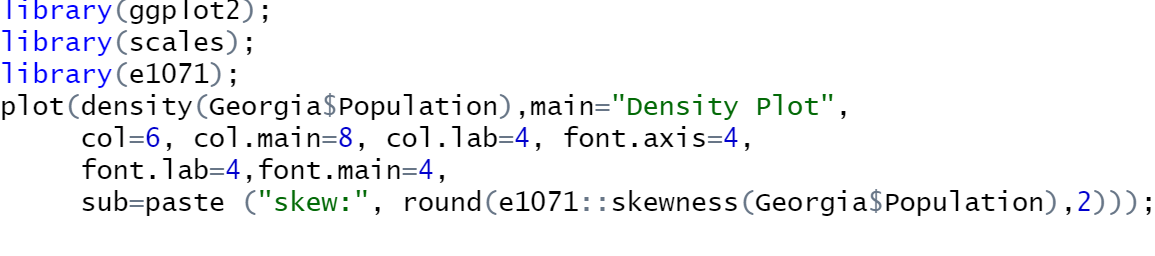
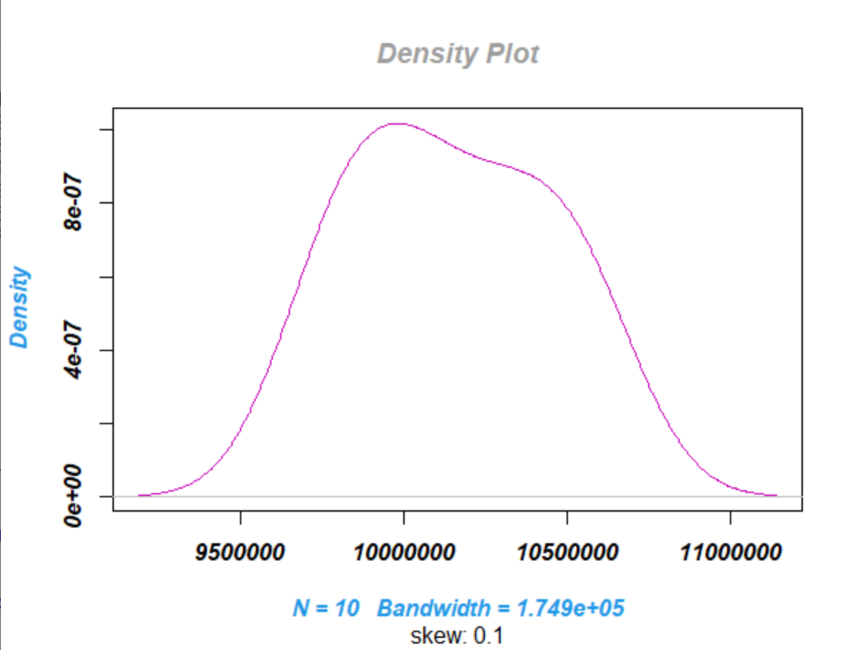




The above residual plots verify that the observations are independent and have a variance of zero, therefore, linear regression will be a good model fit for the given data.

**To check for normality:**

A density plot was created to check for normal distribution of the data. The code used in R representing the graph is below:

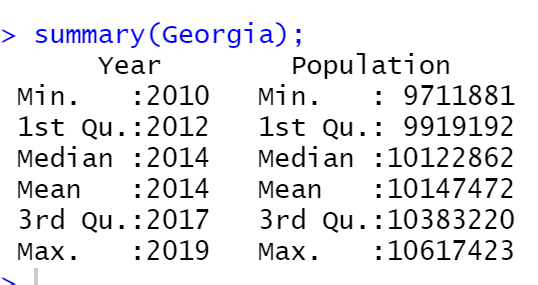


It is evident from the density plot that despite being slightly skewed, it can be concluded, that the data follows a normal distribution and is a fit for linear regression model.

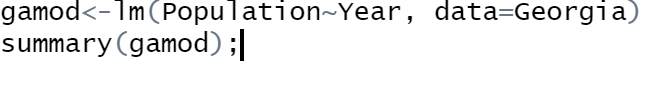
The different graphical representations to investigate data suggest that linear regression will be a powerful model to predict the population of Georgia for the Year 2025. As stated in *Assumptions of Linear Regression*, “A Linear Regression model’s performance characteristics are well understood and backed by decades of rigorous research. The model’s predictions are easy to understand, easy to explain and easy to defend” (Date, 2020).

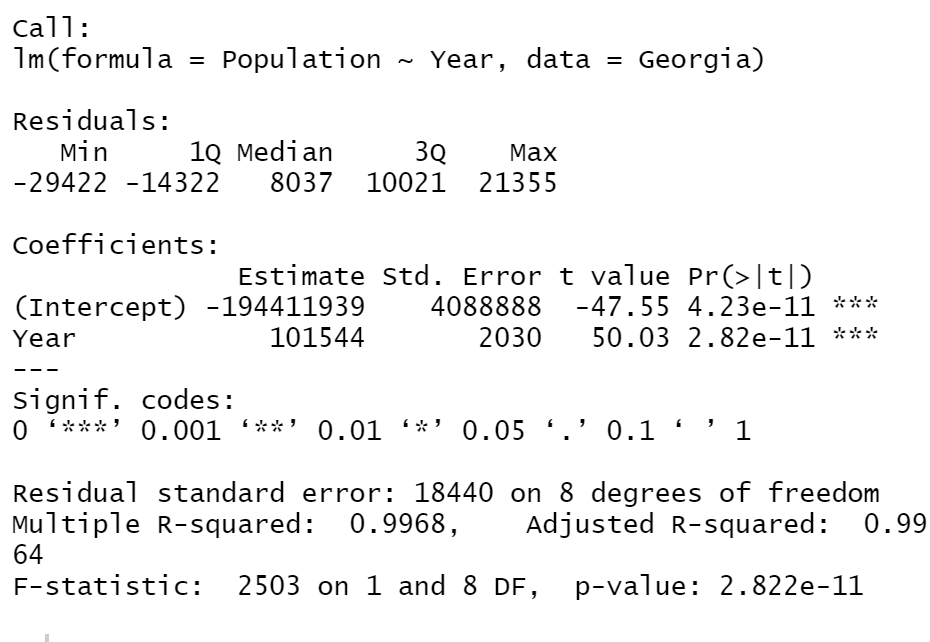
The function lm() was used to plot the linear regression, as mentioned above in the report. The summary function in R was called to analyze the statistical measures on the data set Georgia by using the code below:





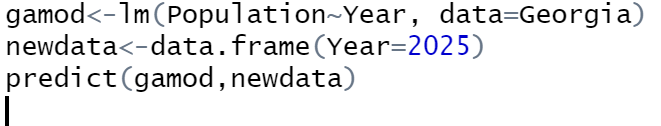
The summary function was also called on the Linear model using the code below:

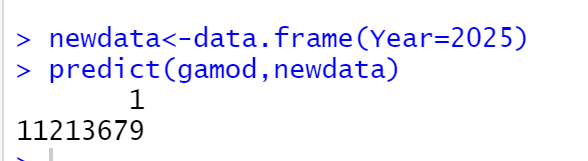




The summary of the linear model indicates that the Linear model created is a good fit, as the adjusted R2 is very close to 1.

Once the linear regression model was created, the prediction() function in R was used to predict the population of Georgia in the Year of 2025. To successfully use this function, along with the Linear Mode, a data frame with Year value of 2025 was passed to estimate the population in five years. The code used for prediction is below:





Based on the prediction, the population of Georgia will be **1,112,13679** in the Year 2025.

**References:**

Date, S. (2020, August 07). Assumptions of Linear Regression. Retrieved September 07, 2020, from <https://towardsdatascience.com/assumptions-of-linear-regression-5d87c347140>

Prabhakaran, S. (2016). Assumptions of Linear Regression. Retrieved September 07, 2020, from http://r-statistics.co/Assumptions-of-Linear-Regression.html

Great. We offer free e-learning courses to students. You do have to create a SAS Profile using your .edu email address. I did not see one in our system. Then you would register for the SAS Academic Hub. The SAS Academic Hub is available for all university students interested in learning SAS. To access free e-learning and online resources, please [click here](https://www.sas.com/en_us/learn/academic-programs/resources/free-sas-e-learning.html#for-students). To request access, scroll down and click the "Register for the SAS Academic Hub" box located under the For Students tab. To validate your access, your SAS Profile must use an email address associated with your university domain. You will need to accept the license agreement. Then you will be taken to the Virtual Learning Environment (VLE) where you can access the Hub.

To access the courses you will log into the SAS Profile associated with the .edu email address and go to that Profiles My Training.