Rhode Island 5 Year Population Prediction Report

Abhishek Khatti

Western Governors University

# Contents

[Contents 2](#_Toc36654797)

[Synopsis 3](#_Toc36654798)

[The United States Population Raw Data Set 3](#_Toc36654799)

[Project Assumptions & Structure 3](#_Toc36654800)

[Data Munging 4](#_Toc36654801)

[Re-Structure Raw Data Set 4](#_Toc36654802)

[Building Home State Dataset 6](#_Toc36654803)

[Importing Required Libraries 6](#_Toc36654804)

[Feature Selection 7](#_Toc36654805)

[Check Distribution for Rhode Island 2010 Data 7](#_Toc36654806)

[Build a Linear Regression Model 9](#_Toc36654807)

[Load Training Data & Prepare for Model 9](#_Toc36654808)

[Checking Linear Regression Relationship 10](#_Toc36654809)

[Check for Linear Relationship 10](#_Toc36654810)

[Train-Test Split of Training Data Set 11](#_Toc36654811)

[Create Linear Regression Model & Statistical Descriptions 11](#_Toc36654812)

[Regression Diagnostics 12](#_Toc36654813)

[Model Performance 14](#_Toc36654814)

[Predicting Population for Rhode Island State 16](#_Toc36654815)

[Loading Feature Values for Predictions 16](#_Toc36654816)

[Saving the Final Output in a File 16](#_Toc36654817)

[Final Outputs & Conclusions 17](#_Toc36654818)

[Console Output for R Script 17](#_Toc36654819)

[Final Output File with Predicted Populations 20](#_Toc36654820)

[Conclusion 21](#_Toc36654821)

[Acknowledgment 21](#_Toc36654822)

[References 23](#_Toc36654823)

Rhode Island 5 Year Population Prediction Report

# Synopsis

The United States analyzes demographic information from the U.S. population. The U.S. Census Bureau presents estimates of the population size of each U.S. state and region. A significant portion of crucial decisions are made using these estimated population dynamics, including the investments in new infrastructures, such as schools and hospitals; establishing new job training centers; opening or closing schools and senior centers; and adjusting the emergency services to the size and characteristics of the demographics of metropolitan and other areas, states, or the country as a whole. A linear regression model is to be developed and used to predict the population for a U.S. state for the next 5 years ("Performance Assessment - BOM1").

# The United States Population Raw Data Set

The United States Population raw data set is available in a Comma Separated Value format (.CSV) file. The file consists of a total of 67 rows and 10 columns. Out of 67 rows, there are 57 observations for different Geographical Areas in United State and 1 for Headers. Out of 10 columns, there 9 features that provide population estimate details by year.

In conclusion, the data set file consists of 2 features, the name of the Geographical Area and Population Estimates by Years. The structure of the raw data file is not optimal to be directly loaded into any statistical software for analysis and would need to be rectified to be in a structured format.

# Project Assumptions & Structure

The project structure follows the standard data science template, which consists of folder structures specific to a phase of the project. For example, the “data” folder will house all data set related files. The “data” folder will have sub-folders depicting the type of data file it would consist of. For instance, “raw” folder will contain all raw data set files, “processed” folder would consist all files which are produced as an outcome of processing on raw data set files, “external” folder would consist final output files that can be released as an outcome of the analysis and so forth.

Some of the important folders that are used in this analysis are:

1. data/raw: Contains raw population data by Geographic Area
2. data/processed: Contains processed data set files as a result of Data Munging activities
3. data/external: Contains final output files as a result of the analysis and Linear Regression Model
4. src/r\_code: Contains R Script file used to perform analysis and prediction. The name of the R script file is “*wgu\_population\_predictions\_ri.R*”

Some important assumptions considered for this analysis & prediction are:

1. The analysis & predictions will be performed in R.
2. A linear regression methodology is to be used for predicting population
3. The Home State of the author is Rhode Island and all analyses and predictions will be specific to only this state.

# Data Munging

## Re-Structure Raw Data Set

The raw data set has a variety of extra information, like “Table 1” header, footer and citation information and a few merged cells. The raw data file is placed under the “raw” sub-folder of the “data” folder, as “*population\_raw.csv*”.

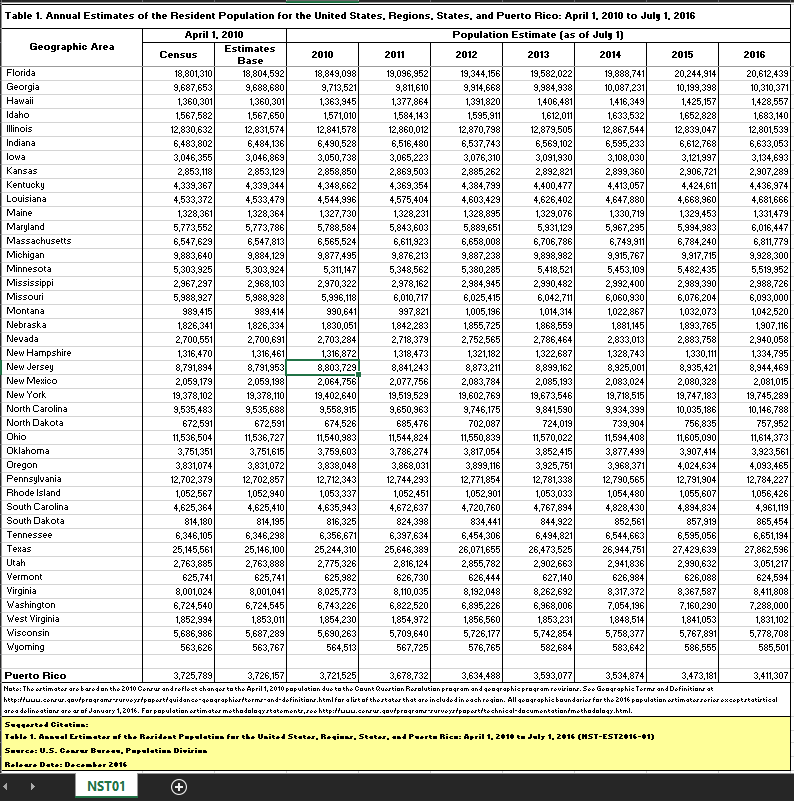


Figure 1. Raw Data Set for Population by Geographical Area

These unwanted details will have to be removed from the file before any other data munging activities could be performed. Besides, the population estimates contain commas separating the thousands. This format would need to be fixed as well since the final training data set would be saved in Comma Separated Values (CSV) format and retaining the former format for population estimates would hinder the smooth functioning of the script.

All the above edits to the raw data file are performed manually in Microsoft Excel. The final processed file is then saved as a separate file under the “processed” sub-folder within the “data” folder as “*population.csv*”.

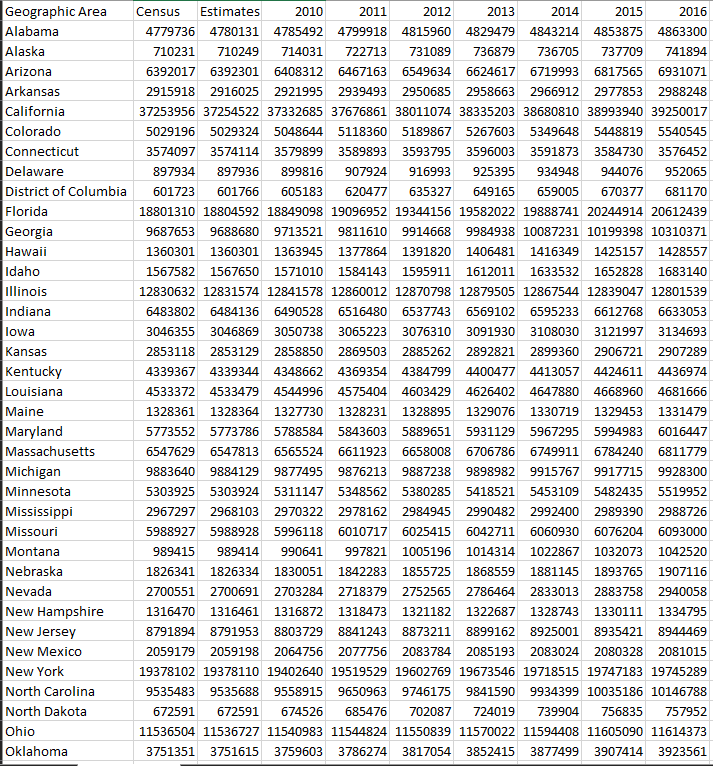


Figure 2. Intermediate Processed Data Set for Population by Geographical Area

## Building Home State Dataset

The processed data file generated in the previous steps can be used to filter out only Rhode Island specific observations. This would have features for Census, Estimates Baseline and yearly population estimates from 2010 to 2016. The final processed file is then saved as a separate file under “processed” sub-folder within “data” folder as “*population\_ri.csv*”



Figure 3. RI Home State Training Data Set

## Importing Required Libraries

For completing analysis smoothly and easily, R provides a huge number of libraries to import additional functionalities in addition to standard ones. The libraries that are being used in this analysis are:

1. dplyr
2. base
3. datasets
4. stats
5. utils

## Feature Selection

The Census and Estimates Base features apply to April 2010 and there is an additional population estimate feature for July 2010. A decision must be made whether to include Census and Estimates Base in creating the learning model.

One option would be to average the three feature values specific to the year 2010. Another option is to check if the July 2010 estimate is a fair representation of all 3 population estimates. If this follows a normal distribution and July 2010 falls under 1 or 2 Standard Deviations of the distribution, we can drop Census and Estimates Base features from the model and just use July 2010 Population Estimates.

### Check Distribution for Rhode Island 2010 Data

The density curve can be created for the 2010 population estimates for Rhode Island state (Prabhakaran, "Linear Regression - A Complete Introduction in R with Examples", 2019). It is created by the “plot” function and is displayed in the following commands. It then generates a density curve for the 2010 population estimates. Besides, the script also shows the standard deviation interval for density distribution, along with the difference between July 2010 population estimate and the right standard deviation interval.

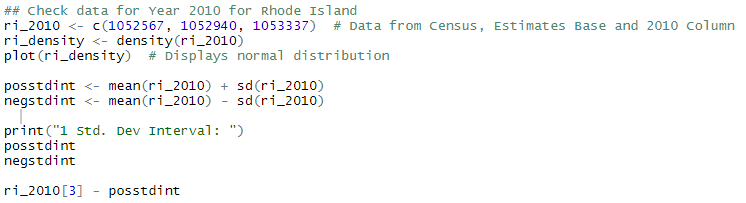


Figure 4. Commands for Density Curve & Std. Dev. Intervals

It can be observed that the distribution looks normal and all normal distribution rules can be applied to the year 2010 population data.

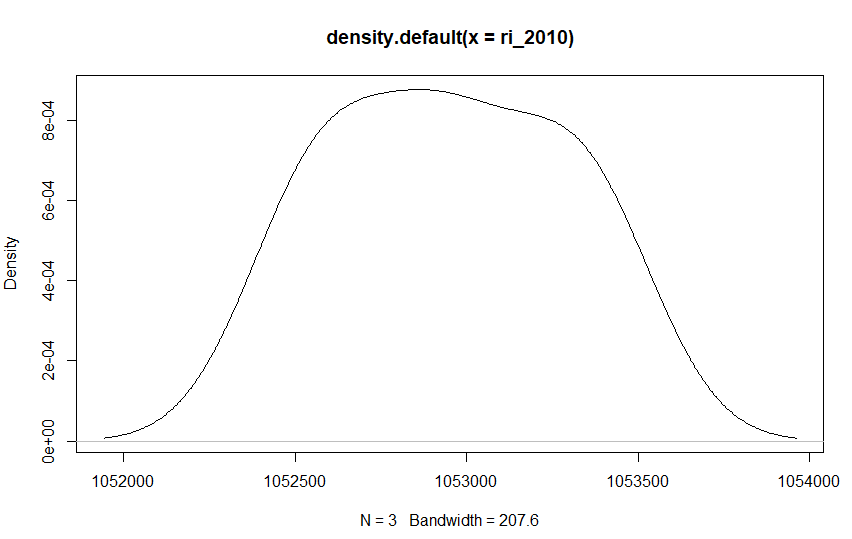


Figure 5a. Density Curve for 2010 Population Estimates

The values at one standard deviation (σ) intervals for the above distribution are (1052563, 1053333). Thus, it can be inferred that the population estimate as of July 2010 is approximately within 1 standard deviation range and there is a 68% chance that the estimate is correct. Thus, the Census and Estimates Base features can be dropped and the 2010 feature as of July 2010 can be retained as is. The Geographic Area feature can also be dropped since the analysis is only for one state. The data can be restructured and transposed into 2 features, Year and Population.

The final processed file is placed under the “processed” sub-folder under “data” as “*population\_ri\_train.csv*”.

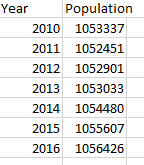


Figure 5b. Final Population Training Data for RI

# Build a Linear Regression Model

## Load Training Data & Prepare for Model

The final processed file is loaded into the R statistical software, using the “read.csv” function into the “trainingdata” vector ("Reading and Writing CSV Files").



Figure 6. Load Training Data into R Script

To be able to pass the required predictor and response variables to build the Linear Regression model, the training data loaded into R needs to be split into their respective vectors. The predictor variable in this care is “Year” and the response variable is “Population”. The below command splits the training data in their respective vectors.

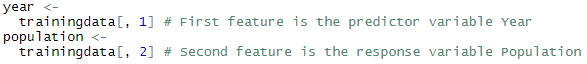


Figure 7. Split Features from Training Data into their resp. Vectors

## Checking Linear Regression Relationship

Before proceeding with building a linear regression model for predictions, the relationship between the predictor and the response variable will have to be checked. Also, the linear regression assumptions will have to be checked if the data does not pose any unwanted biases.

### Check for Linear Relationship

The relationship can be checked easily by using “pairs” and “cbind” functions and passing the predictor and response predictors. The command and its respective plot are shown below. It can be inferred that there is some form of relationship between the two variables.



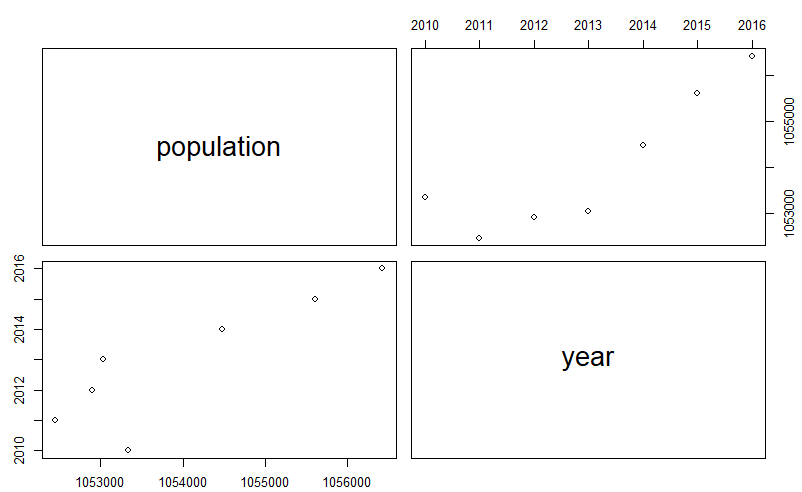


Figure 8a & 8b. Command for Scatterplot & Scatterplot graphs

### Train-Test Split of Training Data Set

Before building the linear regression model, the training data set is split into train and test sub-sets (Prabhakaran, "Linear Regression - A Complete Introduction in R with Examples", 2019). This is optional, but it is considered a best practice to perform this split. The simple reason is to check the performance of the model and build the best model for prediction. This split is performed by the below commands in the R script. This command set will perform the 80-20 percent split on the original training data.

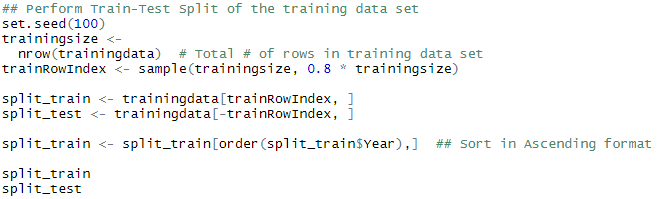


Figure 9. Train-Test Split of original Training Data

### Create Linear Regression Model & Statistical Descriptions

Creating a linear regression model is straightforward and simple using R commands. The “lm” function is used to create the linear regression model (Prabhakaran, "Linear Regression - A Complete Introduction in R with Examples", 2019). Besides, to get the statistical descriptions of the model, the “summary” function is very handy to get all the details in one go (Prabhakaran, "Linear Regression - A Complete Introduction in R with Examples", 2019). Below are the commands that are used to create the model and get its corresponding statistical descriptions.



Figure 10. Create Linear Regression Model

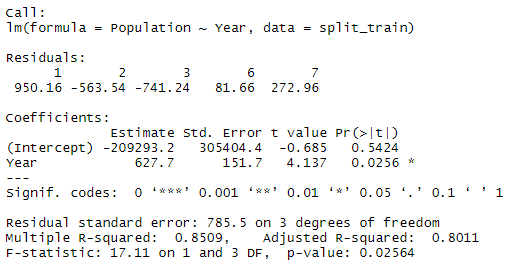


Figure 11. Summary Statistics for Linear Regression Model

The summary statistics show us the Intercept parameter as -209293.2 and the Year parameter co-efficient as 627.7, with its p-value as 0.0256. The p-value is less than the commonly used significance value of 0.5, meaning there is a significant relationship between two variables. The Residual Standard Error is 785.5 on 3 degrees of freedom. As a rule of thumb, smaller Residual Standard Error means the predictions are better. The F-statistic is 17.11 on 1 and 3 DF with its p-value as 0.02564 which is less than the significance value of 0.5. It means the model is relevant and better than the intercept only model. The Multiple R-Squared value of the model for the training data is 85.09%, meaning the model will give predictions with approximately 85% accuracy ("Assessing the Fit of Regression Models", 2020).

### Regression Diagnostics

Before proceeding with using this model to use for validations and finally predicting populations for the Rhode Island state, the model needs to be confirmed that linear regression assumptions hold up. Although the reading the plots are subjective to the reader of the residual plots, it is a good practice to check the residual plots to find any glaring violations of assumptions. The assumptions checked for this model are:

1. Mean of Zero
2. Heteroscedasticity
3. Independence
4. Normal Probability Residual

The commands used to generate the plots are below

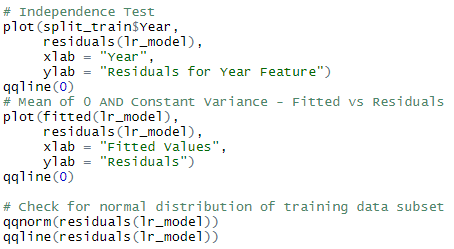


Figure 12. Check Linear Regression Assumptions

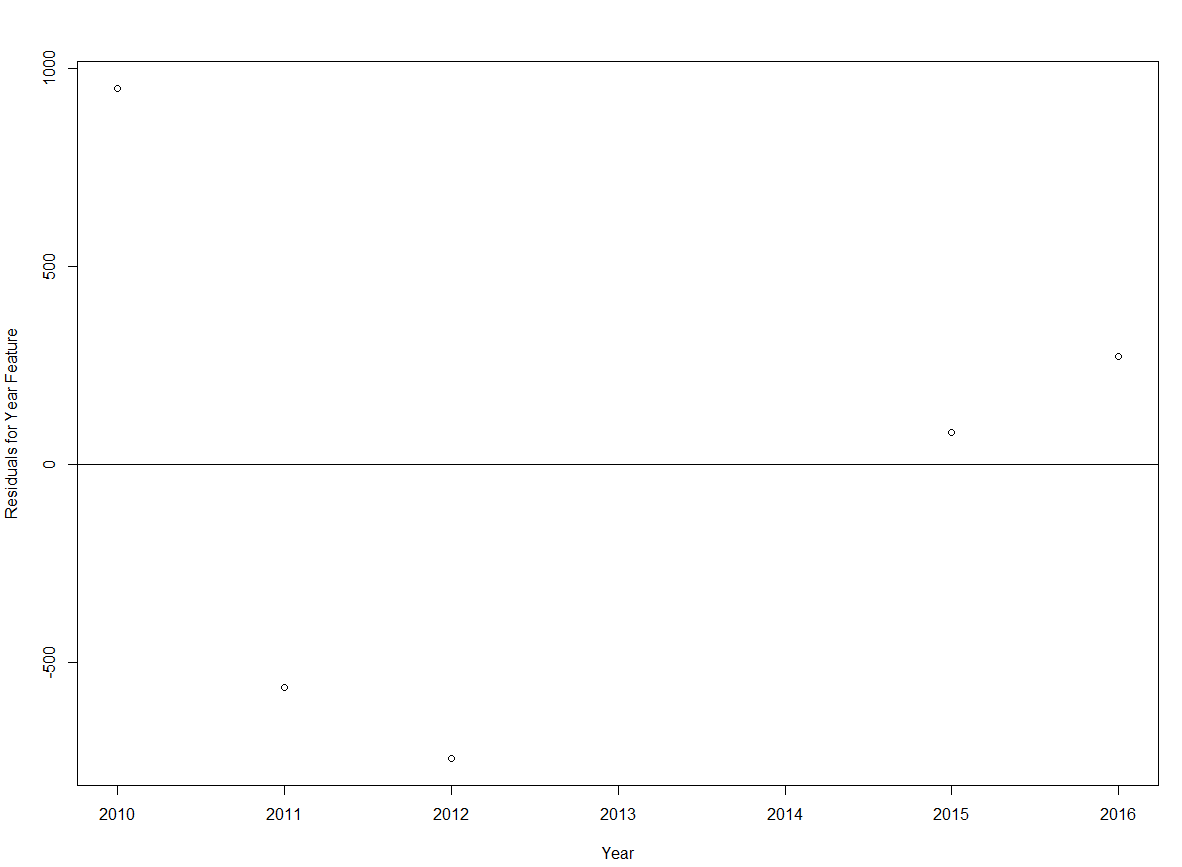


Figure 13a. Independence Test

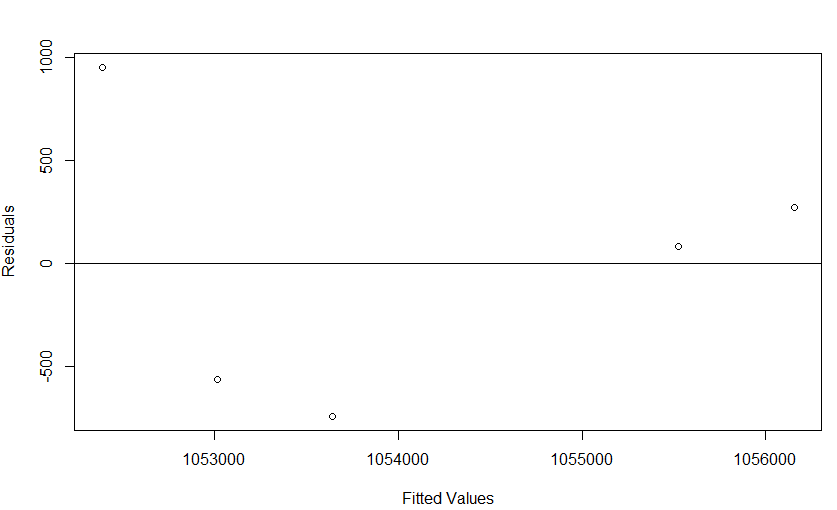


Figure 13b. Mean of Zero & Heteroscedasticity Test

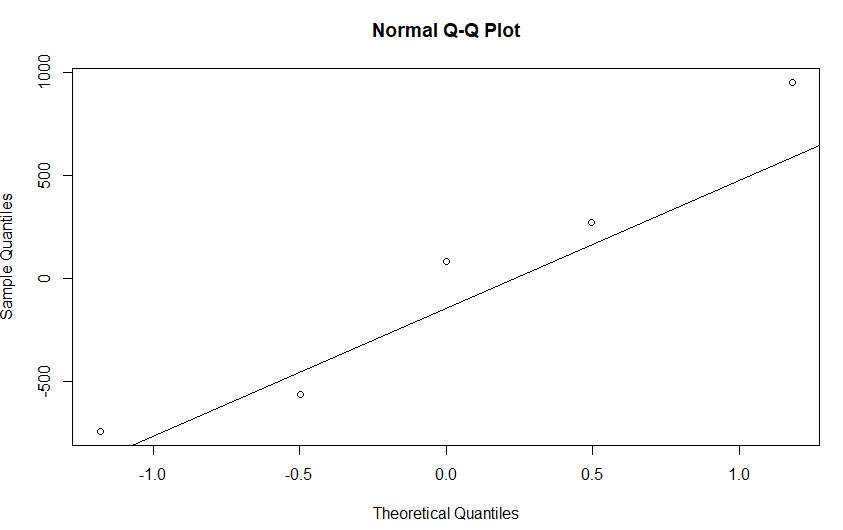


Figure 13c. Normal Probability Test

Since the total number of data points is less than 30, the regression assumptions are loosely hold up and the model can be used for predictions.

### Model Performance

The model performance can be assessed by checking regression errors and min-max accuracy scores. Smaller Regression errors and higher min-max accuracy numbers dictate that the model is performing well. The functions “regr.eval” from DMWR package can be used to get MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error) and MAPE (Mean Absolute Percentage Error). The “apply” and “mean” functions can be used to calculate the min-max accuracy score. If the output score is closer to 1, the model is better. The commands to calculate these scores and their actual outputs are below (Prabhakaran, "Linear Regression - A Complete Introduction in R with Examples", 2019).

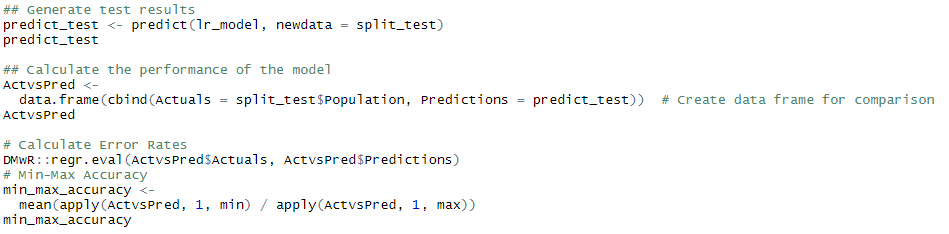


Figure 14a. Check Model Performance

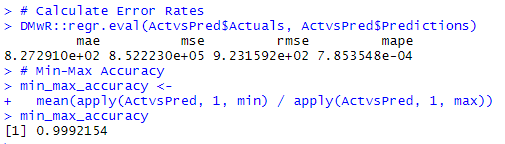


Figure 14b. Model Performance Output

The above code creates a vector “predict\_test” with the predictions for the validation data sub-set from the train-test split. It also creates a data frame with two columns. The first column Actual Population values from the validation data sub-set, and the second column is for Predictions for the validation data sub-set. This data frame is then used to calculate the regression errors and min-max accuracy scores. The regression errors are comparatively smaller to the population numbers and the min-max accuracy score is very close to 1.

Now that the model performance is evaluated and considered a good model, it can be used for performing final predictions on the population for Rhode Island state.

# Predicting Population for Rhode Island State

## Loading Feature Values for Predictions

The feature values will have to be loaded into the statistical software to be passed to the Linear Regression model for predicting the values. The feature values are stored in “population\_test.csv” under the “processed” subfolder of “data”. These feature values are loaded into the “testfeatures” vector. The code for this is below.



Figure 15. Load Feature Values for Prediction

These feature values are passed to the “predict” function for the lr\_model. The output population predictions for Rhode Island state are saved to the “fiveyearpredictions” vector. The output is shown below.



Figure 16a. Generate Predicted Values for 5 Years



Figure 16b. Predicted Values for 5 Years

## Saving the Final Output in a File

The final predictions can be saved in an output file. This output can be either used as a final output of the analysis or further used for advanced analytics or visualizations. The final output is first written in a proper structure, where each predicted population is displayed against its respective year. This can be done by concatenating the “testfeatures” vector with the “fiveyearpredictions” vector into a data frame. This data frame can be written to the file.

The “write.csv” function can be used to finally data frame to a CSV format file ("Reading and Writing CSV Files"). The commands and their respective outputs are shown below.

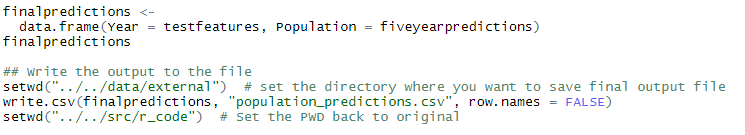


Figure 17a. Write Final Predicted Population to File

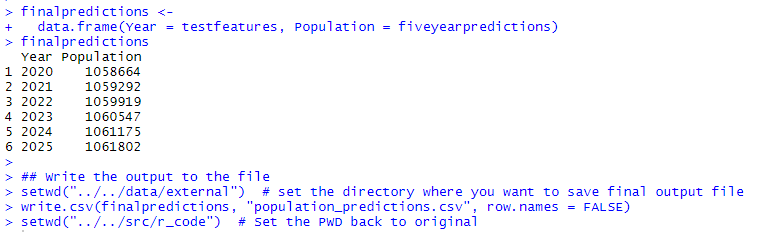
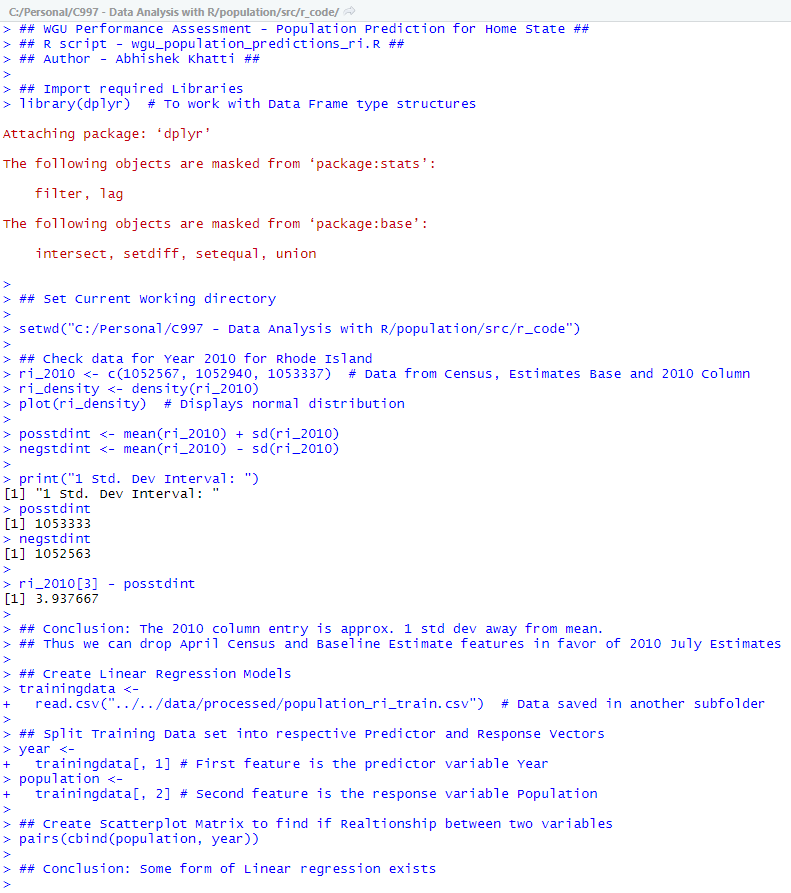


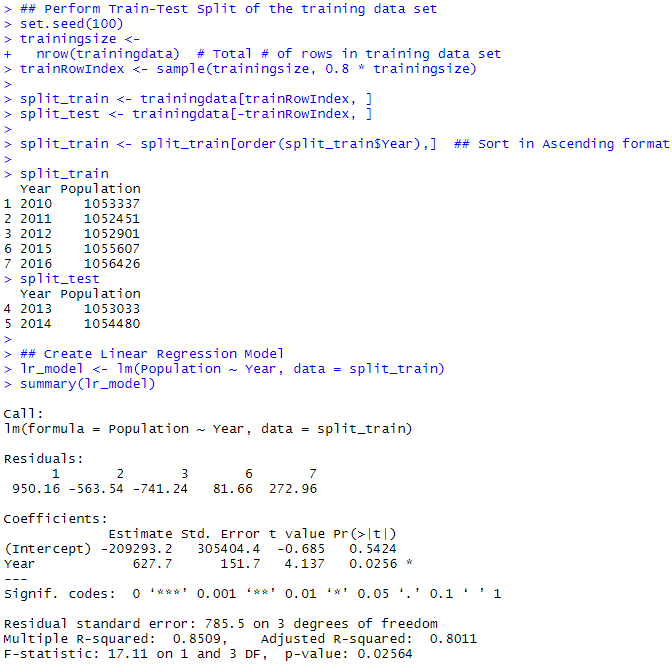
Figure 17b. Final Predicted 5 Year Population

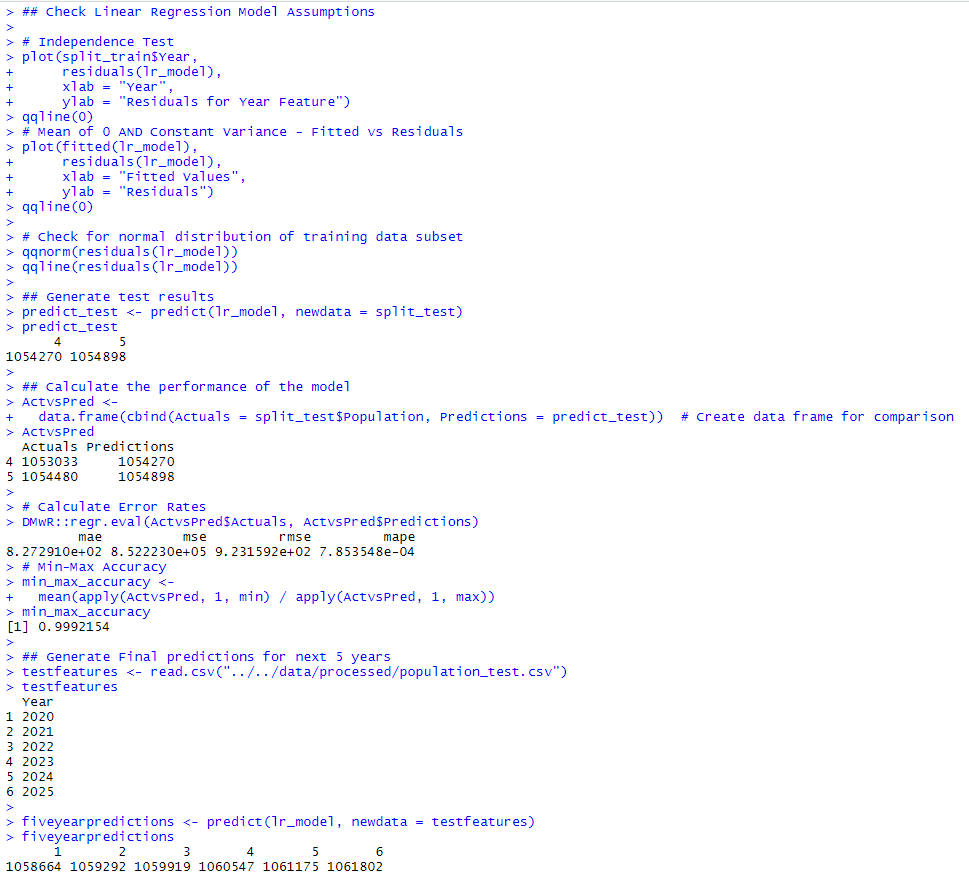
# Final Outputs & Conclusions

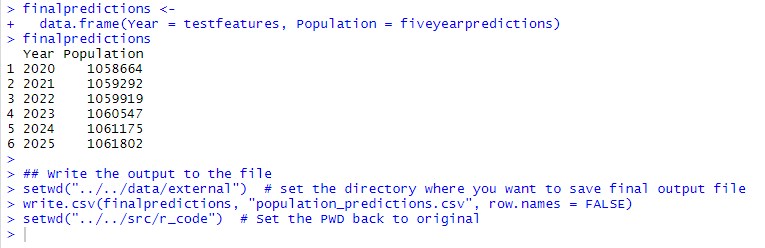
## Console Output for R Script

Below is the final console output for the R script that predicts Rhode Island State population for the next 5 years.









## Final Output File with Predicted Populations

The final output file with 5-year population predictions for Rhode Island state is shown below. The output file is placed under the “external” sub-folder under the “data” folder of the project as “*population\_predictions.csv*”.

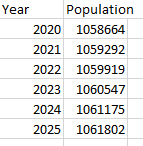


Figure 17b. Final Predicted 5 Year Population

## Conclusion

Based on a small training data set of 7 years starting from 2010, a linear regression model was built and used for predicting population for Rhode Island State for the next 5 years, until 2025. The predicted population numbers are then saved in a Comma Separated Value (CSV) format file, which shows the years for which the population is to be predicted, and its respective population predictions. This output file is saved under the “external” sub-folder under the “data” folder.

# Acknowledgment

The raw data set was obtained by Dr. William Sewell through the Course Announcements link.

The document template for this report was provided by the Writing Center team from Student Success Centers, with additional details referred to APA style formatting from WGU Knowledge Center (WGU Knowledge Center, solene, "APA Document Formatting", 2018).

All the sources listed under the “References” section of this report were used to refer details and instructions to complete this analysis and report.

Finally, sincere gratitude to “Dr. William Sewell” for his exceptional guidance, support, and encouragement throughout this course.

# References

PERFORMANCE ASSESSMENT - BOM1. (n.d.). Retrieved from

https://tasks.wgu.edu/student/001101075/course/14380006/task/1148/overview

Prabhakaran, S. (2019, October 22). Linear Regression - A Complete Introduction in R with Examples. Retrieved from https://www.machinelearningplus.com/machine-learning/complete-introduction-linear-regression-r/

Reading and Writing CSV Files. (n.d.). Retrieved from https://swcarpentry.github.io/r-novice-inflammation/11-supp-read-write-csv/

Assessing the Fit of Regression Models. (2020, January 16). Retrieved from https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/

Kumar, A. (2014, October 18). R Programming Fundamentals. Retrieved from https://app.pluralsight.com/library/courses/r-programming-fundamentals/table-of-contents

Renze, M. (2016, October 25). Data Science with R. Retrieved from https://app.pluralsight.com/library/courses/r-data-science/table-of-contents