C997 9/20/2020

James Shea

Task A:

To complete this task, I used the book R for Data Science to help me learn R. To complete the linear regression analysis, I first had to download the data from the Census Bureau website. Then, I used R to clean the data. Finally, I was able to create my model. To create the linear regression model, I used the code screenshot below. The results of that model are also screenshot below.

```
#-----------------------------------------
# Create Regression Model
#-----------------------------------------
regression_model=lm(population~years)
#-----------------------------------------
```

```
> regression_model

Call:
lm(formula = population ~ years)

Coefficients:
(Intercept)        years
  -25154377        18249
```

Task B:

Data was originally downloaded from the census bureau website as provided in the description of the task. Data was then imported into R using the read.csv command (screenshot below). Data was cleaned using R (screenshot below). Cleaning the data resulting in removing most of the data as most of it was unnecessary for this analysis. Many columns and most rows were deleted. This was done in R by simply making a new data frame from the original raw data that only included the necessary data. A screenshot of the results is also below.

```
#---------------------------
# Read Data File (Part B)
#---------------------------
raw <- read.csv("raw.csv")
#---------------------------
```

```
#-------------------------------------------------------------------------------------------------------------------
# Clean Data (Part B)
#-------------------------------------------------------------------------------------------------------------------
clean <- as.data.frame(raw[c(41), c(8:17)])
clean
#-------------------------------------------------------------------------------------------------------------------
# Save Data
#-------------------------------------------------------------------------------------------------------------------
population <- c(11539336, 11544663, 11548923, 11576684, 11602700, 11617527, 11634370, 11659650, 11676341, 11689100)
years <- c(2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019)
#-------------------------------------------------------------------------------------------------------------------
```

```
> clean
   POPESTIMATE2010 POPESTIMATE2011 POPESTIMATE2012 POPESTIMATE2013 POPESTIMATE2014 POPESTIMATE2015 POPESTIMATE2016 POPESTIMATE2017 POPESTIMATE2018 POPESTIMATE2019
41        11539336        11544663        11548923        11576684        11602700        11617527        11634370        11659650        11676341        11689100
> #-------------------------------------------------------------------------------------------------------------------
> # Save Data
> #-------------------------------------------------------------------------------------------------------------------
> population <- c(11539336, 11544663, 11548923, 11576684, 11602700, 11617527, 11634370, 11659650, 11676341, 11689100)
> years <- c(2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019)
> population
 [1] 11539336 11544663 11548923 11576684 11602700 11617527 11634370 11659650 11676341 11689100
> years
 [1] 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019
>
```

Task C:

The summary command (screenshot below) was used to tabulate a statistical description of the model, screenshot below.

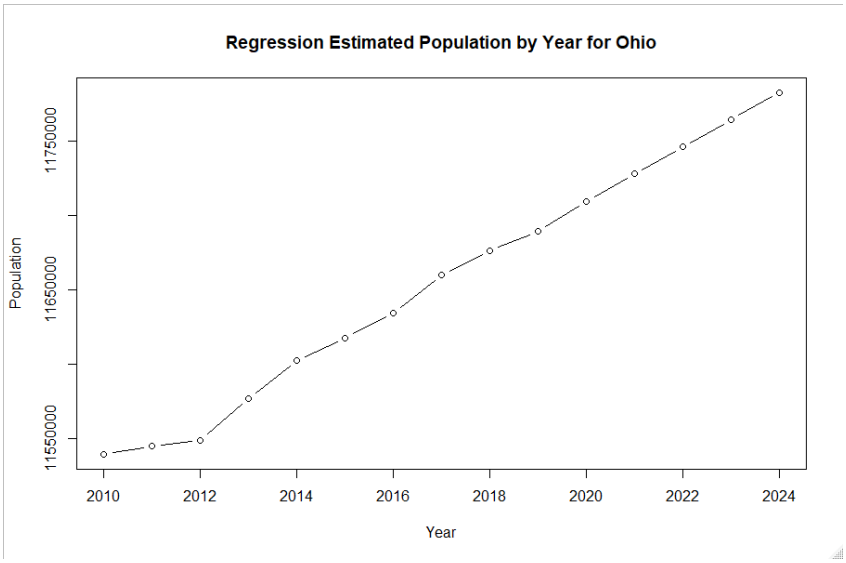
```
#---------------------------------------
# Summary of Regression Model (Part C)
#---------------------------------------
summary(regression_model)
#---------------------------------------
```

```
> summary(regression_model)

Call:
lm(formula = population ~ years)

Residuals:
     Min       1Q   Median       3Q      Max
-14383.0  -1946.9   -460.4   3378.0  12528.7

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.515e+07  1.649e+06  -15.25 3.39e-07 ***
years        1.825e+04  8.188e+02   22.29 1.74e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7437 on 8 degrees of freedom
Multiple R-squared:  0.9842,    Adjusted R-squared:  0.9822
F-statistic: 496.8 on 1 and 8 DF,  p-value: 1.736e-08
```

Task D:

Future population size was predicted using the linear regression from part A. A screenshot of just the population values for the next 5 years as well as a chart is provided below using screenshots as the task did not specify how to provide these results.

```
> #---------------------------------------
> estimated
       1        2        3        4        5
11709301 11727550 11745799 11764049 11782298
> #---------------------------------------
```



Task E:

References:

Wickham, H., & Grolemund, G. (2017). *R for Data Science*. Beijing: O'Reilly.

U.S. Census Bureau (2019). 2019 National and State Population Estimates. Retrieved from https://www.census.gov/newsroom/press-kits/2019/national-state-estimates.html.