C997 task submission

Jesse Ashby

##Section B - Data preparation and import into R

I downloaded the April 1, 2010 to July 1, 2018 census data from the site specified in the task requirements (United States Census Bureau, 2018). The file was setup with one row per geographic summary level and had columns for population estimates for each year between 2010 and 2018. The components of each year's estimate also each had their own columns.

My goal in preparing the data for import into R was to extract only the relevant data and then transform it to meet the rules of tidy data (Grolemund & Wickham, 2017). I started by opening the file in Microsoft Excel and removing all of the rows that were not for the state I live in (Texas). I then removed all columns except for the population estimate for each year. This left me with 1 row and 9 columns. I pivoted this row to 9 rows in 1 column. This setup my response variable of population estimate, but I lacked the needed explanatory variable of time. I added a new column and manually entered the year of the population estimate.

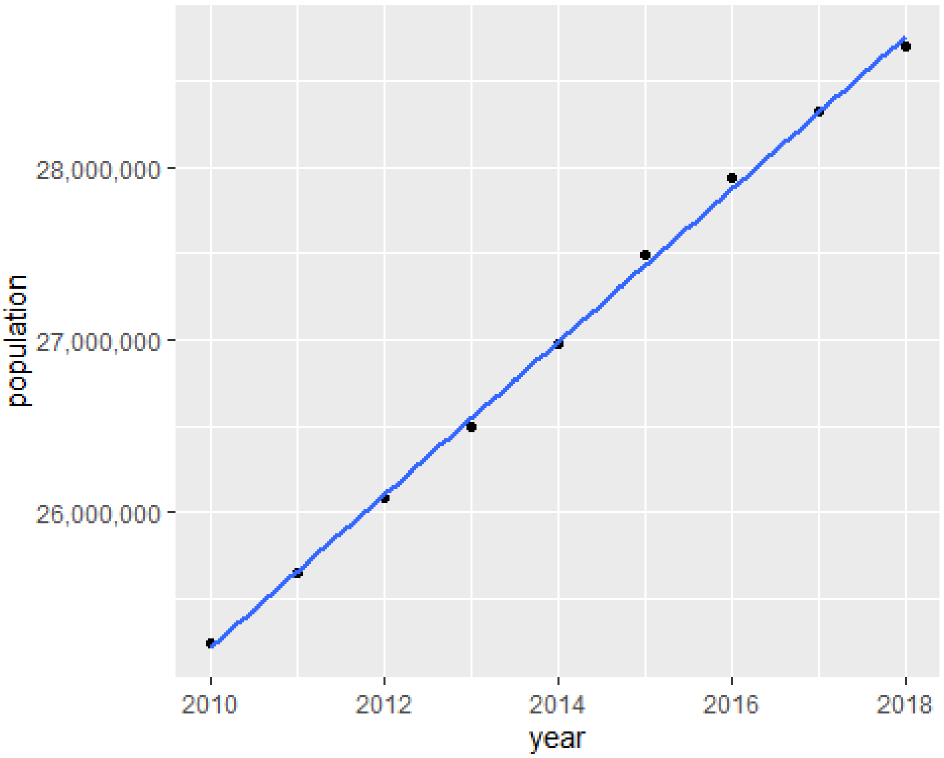
To import the data into R, I used the read_csv function to get the data into a tibble data structure, useful for working with data in the tidy format.

```
library(tidyverse)

pops <- read_csv("texasPopEstimates.csv")
pops

## # A tibble: 9 x 2
##    year popestimate
##   <dbl>       <dbl>
## 1  2010    25242679
## 2  2011    25646227
## 3  2012    26089620
## 4  2013    26489464
## 5  2014    26977142
## 6  2015    27486814
## 7  2016    27937492
## 8  2017    28322717
## 9  2018    28701845
```

##Section A - Linear regression analysis

```
library(modelr)
require(scales)

# (Grolemund & Wickham, 2017)
```

```
popsModel <- lm(popestimate ~ year, data = pops)
popsModel

##
## Call:
## lm(formula = popestimate ~ year, data = pops)
##
## Coefficients:
## (Intercept)         year
##  -864516531       442654

vis <- ggplot(pops, aes(year, popestimate)) +
  geom_point() +
  scale_y_continuous (labels = comma) +
  ylab('population') +
  geom_smooth(method = "lm", se = FALSE)

vis
```



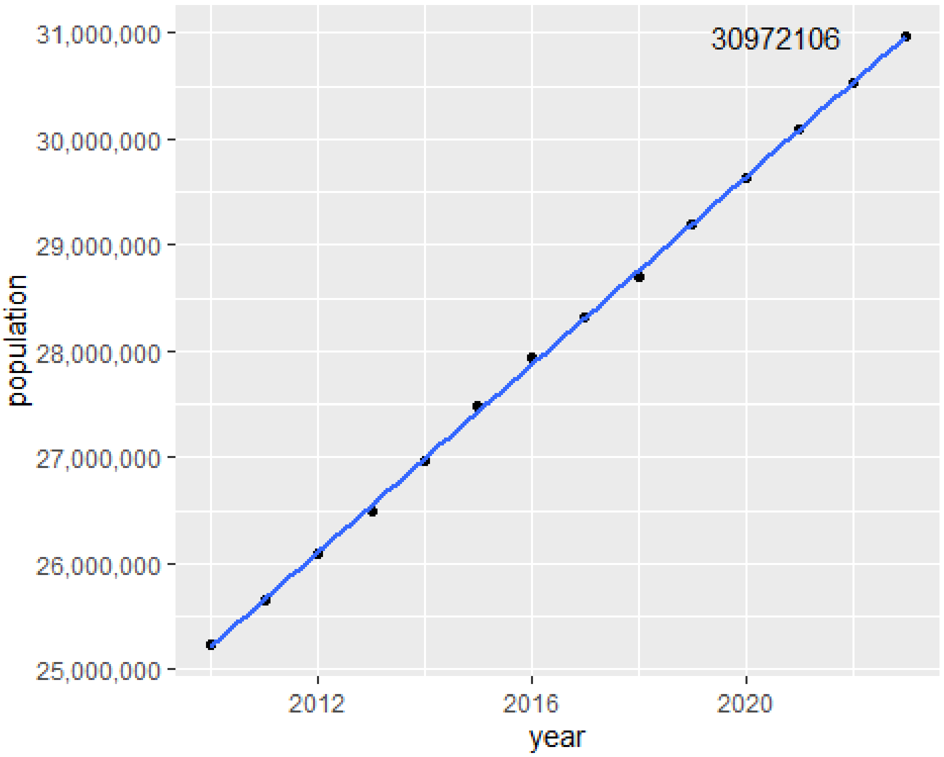
##Section C - Statistical description using summary()

```
summary(popsModel)

##
## Call:
## lm(formula = popestimate ~ year, data = pops)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max 
## -56992 -14034 -11080  25072  63962 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -864516531   11948914  -72.35 2.53e-11 ***
## year            442654       5933   74.61 2.04e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 45960 on 7 degrees of freedom
## Multiple R-squared:  0.9987, Adjusted R-squared:  0.9986 
## F-statistic:  5567 on 1 and 7 DF,  p-value: 2.044e-11
```

##Section D - 5 year prediction for population of Texas

```
next5years <- data_frame('year' = c((max(pops$year) + 1): (max(pops$year) +
5)))

## Warning: `data_frame()` is deprecated as of tibble 1.1.0.
## Please use `tibble()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.

next5pops <- data_frame('popestimate' = predict(popsModel, next5years,
                                                type = "response"))
next5estimates <- data.frame('years' = next5years, 'popestimate' = next5pops)
pops <- rbind(pops, next5estimates)

vis <- ggplot(pops, aes(year, popestimate)) +
  geom_point() +
  geom_text(aes(label=ifelse(year == max(pops$year),
                             as.character(as.integer(pops$popestimate)),'')),
            hjust=1.5) +
  scale_y_continuous (labels = comma) +
  ylab('population') +
  geom_smooth(method = "lm", se = FALSE)

vis

## `geom_smooth()` using formula 'y ~ x'
```



##References:

Grolemund, G, and Wickham, H. "R For Data Science." R For Data Science, O'Reilly, 2017, r4ds.had.co.nz/index.html.

United States Census Bureau. (2018). National, State, and Puerto Rico Commonwealth Totals Datasets: Population, population change, and estimated components of population change: April 1, 2010 to July 1, 2018 [CSV file]. Retrieved from https://www2.census.gov/programs-surveys/popest/datasets/2010-2018/national/totals/nst-est2018-alldata.csv