Part 2: R

I. The state I live in is Colorado. Below is a linear regression analysis with R to predict the size of the population for just Colorado for 2020, using the existing population data.

```
> # Create a linear model
> coData2.lm <- lm(popCol ~ years)
>
> # Create a barplot
> barplot(popCol, names.arg = years, main = "Colorado Population 2010-2018", xlab = "Year", ylab =
 "Number of People", ylim = c(4000000,6000000), xpd = FALSE, col = c("lightblue"))
>
> # Create a linear regression plot
> plot(popCol ~ years, main = "Colorado Population 2010-2018", xlab = "Year", ylab = "Number of Pe
ople")
> abline(coData2.lm, col = "blue")
>
> # Predict the population size for 2020
> newData <- data.frame(years = 2020)
> predict(coData2.lm, newData, interval = "prediction")
      fit     lwr     upr
1 5860307 5828981 5891633
```

Figure 1 - Code from R script that creates the linear model and predicts population for 2020

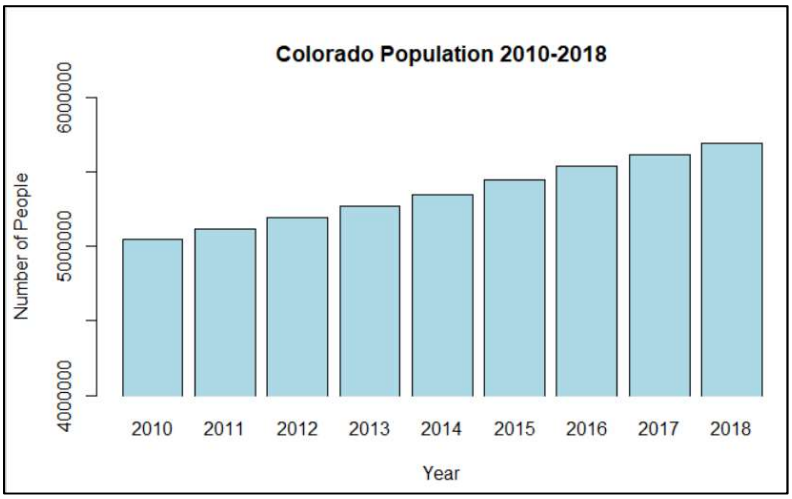

Figure 2 - Bar plot of the population data

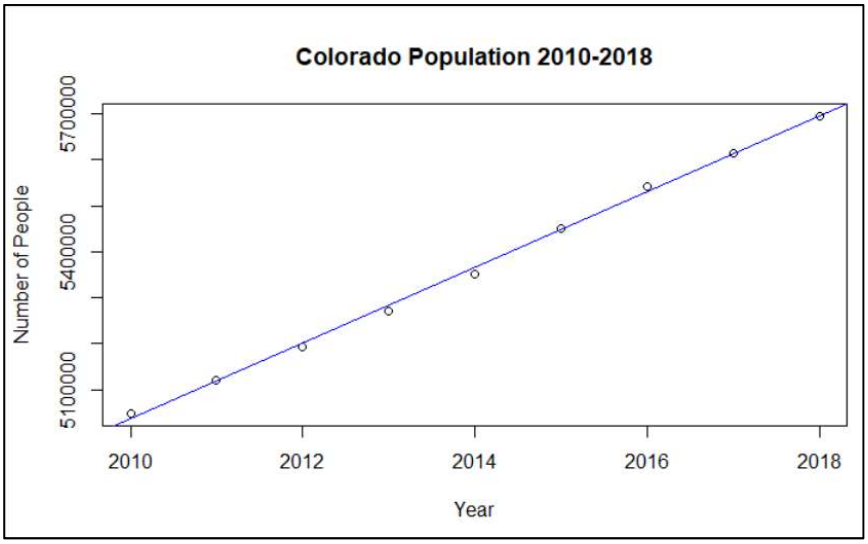

Figure 3 - Linear model of the population data

J. The csv file (Datasets) was imported into R with the code shown below. The header was skipped as I had decided to create the years with a separate vector. The 'popCol' vector is created by only looking at the row for Colorado and selecting just the population totals for each year that were available in the dataset.

```
> # Import the population data - skipping the header as it is not needed
> popData <- read.csv("nst-est2018-popchg2010_2018.csv", header = TRUE, skip = 1)
>
> # Pulls the population estimates for just Colorado
> popCol <- as.numeric(popData[10, c(7:15)])
>
> # Create a vector of the years
> years <- c(2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018)
```

Figure 4 - Code from the R script that imports the data and creates vectors

Data	
coData2.lm	List of 12
newData	1 obs. of 1 variable
newData2	1 obs. of 1 variable
popData	56 obs. of 61 variables
Values	
popCol	num [1:9] 5048281 5121771 5193721 5270482 5351218 ...
years	num [1:9] 2010 2011 2012 2013 2014 ...

Figure 5 - Data and values created by the code in the R script

K. The 'summary()' function is included in the R script that tabulates a statistical description.

```
> # Output the linear model information
> summary(coData2.lm)

Call:
lm(formula = popCol ~ years)

Residuals:
   Min     1Q Median     3Q    Max
-14334  -6913   2973   4096  12566

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -160707210    2633229  -61.03 8.33e-11 ***
years            82459       1308   63.07 6.62e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10130 on 7 degrees of freedom
Multiple R-squared:  0.9982,    Adjusted R-squared:  0.998
F-statistic:  3978 on 1 and 7 DF,  p-value: 6.618e-11
```

Figure 6 - Output of the summary function

L. The predicted population size for the state of Colorado in five years is roughly 6.2 million, as shown below.

```
> # Predict the population size for 2024 (5 years from now)
> newData2 <- data.frame(years = 2024)
> predict(coData2.lm, newData2, interval = "prediction")
      fit     lwr     upr
1 6190144 6150230 6230057
```

Figure 7 - Code from R script that predicts population size in five years

References

Datasets. (December 7th, 2018). *Population change and rankings: April 1, 2010 to July 1, 2018 (NST-EST2018-popchg2010-2018)* [Data file] Retrieved from: https://www2.census.gov/programs-surveys/popest/datasets/2010-2018/national/totals/nst-est2018-popchg2010_2018.csv