MSIN0025 Data Analytics

Part I: Obama Clinton Case Study

Introduction

The case study aims to analyses the already casted votes within the democratic primaries, how this affected the areas which haven't voted yet and how the candidates' voter targeting strategy was adjusted according to these results as two critical states still need to vote and analysts say that whoever wins these two states will likely be the next POTUS. Therefore, adjusting one's strategy to appeal to the local communities can make or break the presidential race.

Focus Areas

1. As Obama has promised to improve healthcare, are counties with more people in need of healthcare more likely to vote for Obama?
2. Are there any similarities between the states that Bill Clinton won (1992 election) and the states Hillary Clinton won?
3. Is there any difference in Obama's support between the caucus and the primary system?
4. How do Obama's results correlate to bachelor degrees within the black population?

Target Attributes according to focus areas

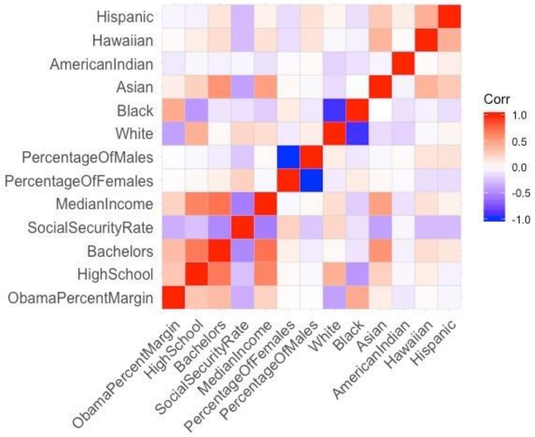
1. People in need of healthcare likely belong to either one of these groups: Age>65, disabled, no medical care, unemployed, no social security. The percentage against the populations of each variable were calculated per county to perform statistical analysis.
2. (also 3.) Using the regions, election type as well as introducing a conditional calculation determining the winner of each county.
3. The black percentage of the total population plays a role as well as the amount of bachelor degrees in the county

Understanding the Data

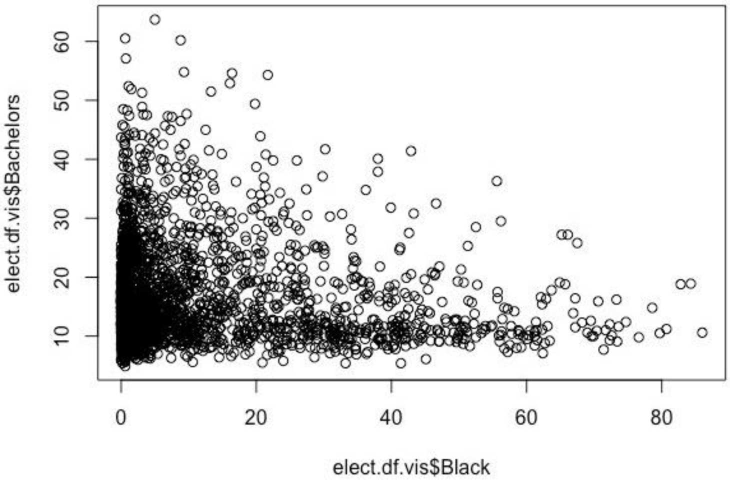
1. The Data was supplied by the U.S. Census Bureau which provides demographic information on 3141 counties. Out of these, 2868 had voted and therefore make voting data available, but 1131 had not voted when the dataset was created. The voting results of each county were matched up with the census data which also includes 41 attributes of which some are categorical and some numerical.

Data Analysis

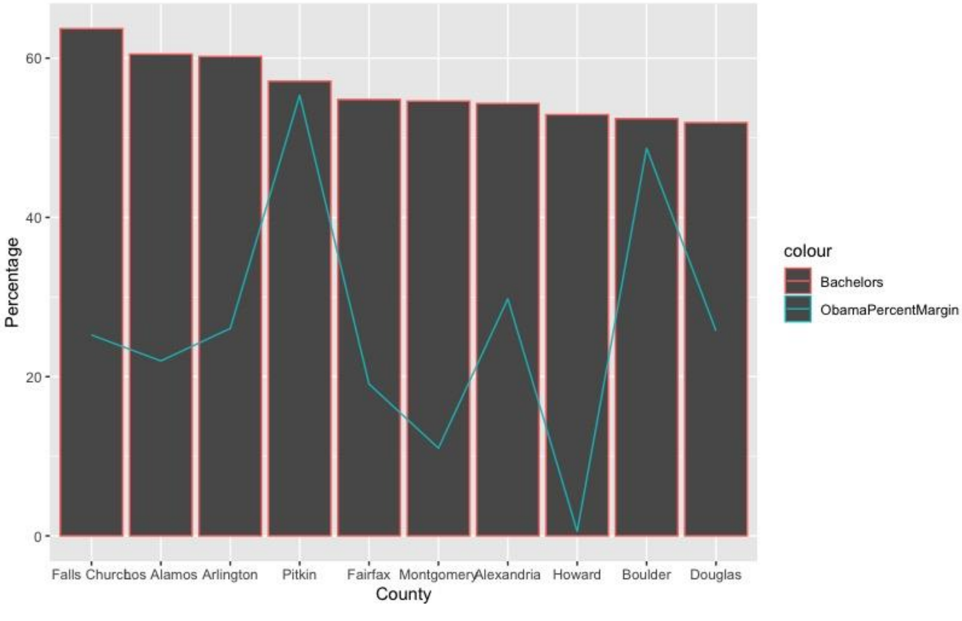
Education and Black



As we can see there is an especially negative correlation between the Black attribute and the attributes such as SocialSecurityRate, HighSchool and MedianIncome. This is likely due to black neighbourhoods or counties being poorer with less education. This can be backed up with a scatter plot showing that the higher the black population percentage the lower the average rate of bachelors is:

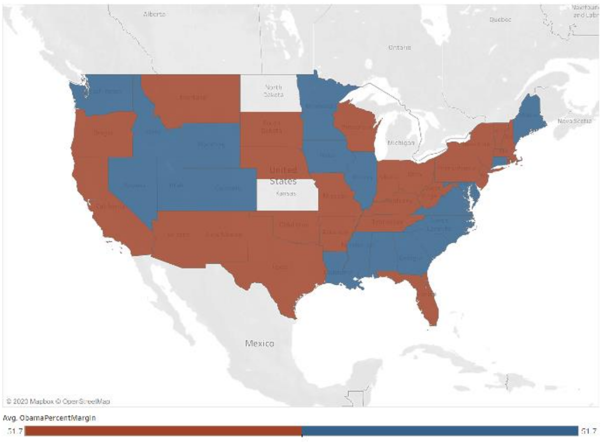


However, we can see with the graph below that Obama's victory does not only depend on the number of black people. It also depends on the level of education that is very positively correlated to Obama's votes. In fact, Obama won in all top ten counties with most people with a bachelor or more level of education. This represents a real strength for Obama who can spread his influence on more counties.

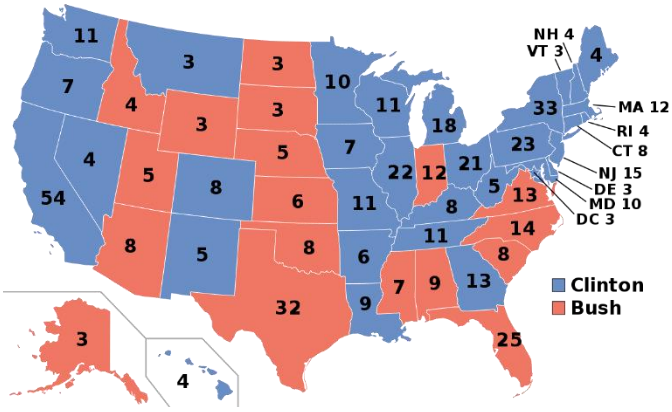


Hillary Clinton versus Bill Clinton

The casted votes that were combined, and predictions were generated in R with the Lasso to visualise the votes for Hillary Clinton (Red, 1st figure) versus the results her Husband achieved in 1992. There is a slight correlation, especially in the East and Western parts of the country. In the central states of the US, this correlation dissipates.



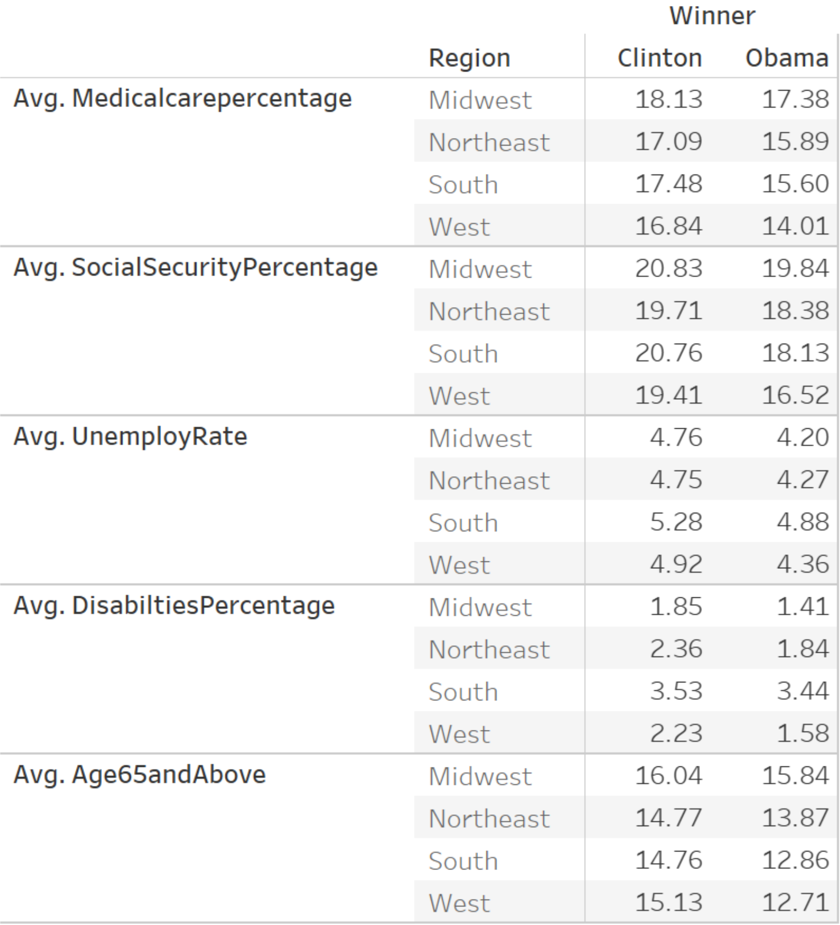
Hillary Clinton (Red) versus Obama (Blue)



Bill Clinton (Blue) versus Bush (Red); (Levy 2009)

Obama's Healthcare Promise

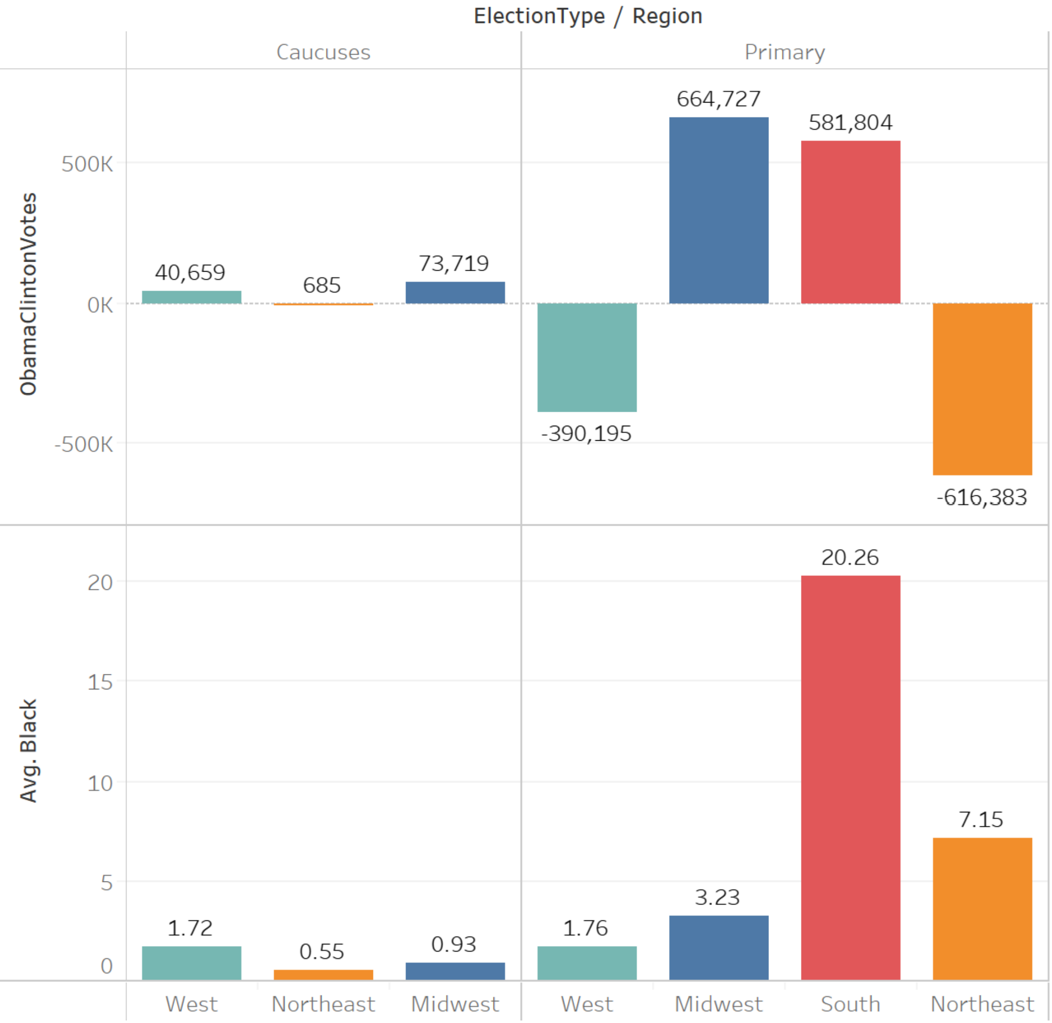
It was assumed that there had to be a correlation between people who would be rather in need of a national health care system and the votes casted for Obama as it was one of his promises throughout his election campaign. As we can see Obama was indeed the winner in the regions where proportionally less people have social security or medical care. Regions with higher unemployment, disabilities and elderly rate would rather vote for Clinton. This was unexpected.



		Winner	
	Region	Clinton	Obama
Avg. Medicalcarepercentage	Midwest	18.13	17.38
	Northeast	17.09	15.89
	South	17.48	15.60
	West	16.84	14.01
Avg. SocialSecurityPercentage	Midwest	20.83	19.84
	Northeast	19.71	18.38
	South	20.76	18.13
	West	19.41	16.52
Avg. UnemployRate	Midwest	4.76	4.20
	Northeast	4.75	4.27
	South	5.28	4.88
	West	4.92	4.36
Avg. DisabiltiesPercentage	Midwest	1.85	1.41
	Northeast	2.36	1.84
	South	3.53	3.44
	West	2.23	1.58
Avg. Age65andAbove	Midwest	16.04	15.84
	Northeast	14.77	13.87
	South	14.76	12.86
	West	15.13	12.71

Caucus versus Primary

As for the question if the voting system makes a difference, in all states where caucuses are used Obama is winning but not many states deploy this method anymore therefore it is not a very impactful insight. Also, it is expected that Obama wins by a large majority in the south where the average percentage of black population is very high. Although the Midwest does not have a large black population it is Obama's largest supporting area. However, it does not have the largest population.



Data Preparation

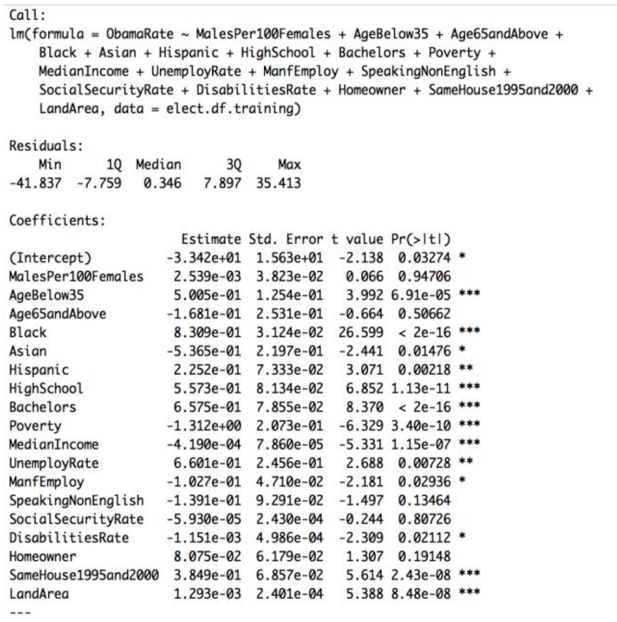
As four hypotheses were focused on, each time different attributes were accounted for. However, throughout the whole analysis, Obama and Clinton votes were utilised with the corresponding attributes to gain insight on the chosen problems which was achieved through the creation of tables, scatterplots, barplots and a correlation matrix.

As the voting data for Clinton and Obama were the core attributes, their individual rates were derived(number votes/number people), as well as margins for each county to conduct a more accurate analysis. In addition, to test if gender had an effect on the votes, we transformed the gender ratio into a percentage of males and females, which made the correlations -as they appear in the correlation matrix- possible. Moreover, to plot the parameters of Obama's victory, we had to predict the votes of counties that hadn't voted (NA values). We did so with a Lasso model and merged the results with the original dataset which only included the already casted votes. To do so, we first derived the ObamaRate, the ClintonRate and the ObamaPercentMargin from the election data. To deal with missing values we replaced the average income by median income, the missing values of black, Asian American-Indian, ManfEmploy, Disabilities DisabilitiesRate and farmarea were replaced by zero. For the attributes with few missing values (HighSchool and Poverty) we suppressed the line.

To use our data for linear regression and random forest we split it into two sub-datasets (test & train). For the regression tree, we have used the built-in k-folds function with 10 folds.

Generating and Testing Prediction Models

First, a linear regression model was constructed to predict the still missing ObamaRate in the counties which have not casted their votes yet. This enabled the identification of the most significant variables and their corresponding rate with which they effect Obama's vote rates. This model proves the selection of focus attributes as Education and Black are amongst the most significant variables for Obama's voter outturn.

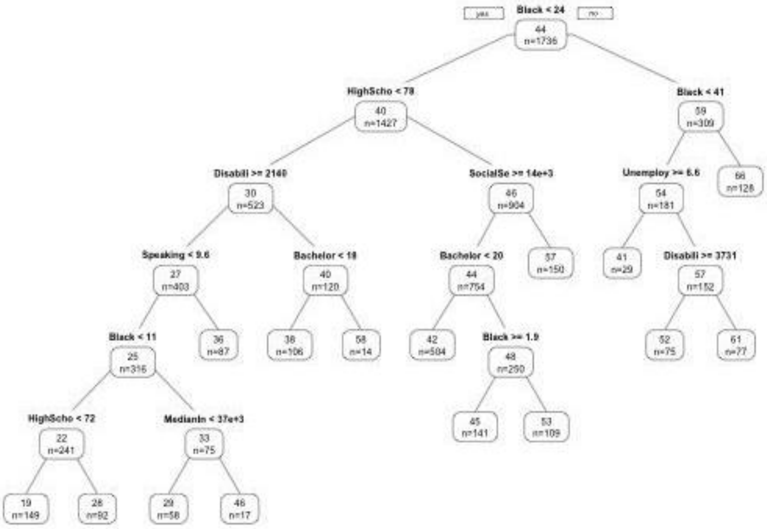


```
Call:
lm(formula = ObamaRate ~ MalesPer100Females + AgeBelow35 + Age65andAbove +
    Black + Asian + Hispanic + HighSchool + Bachelors + Poverty +
    MedianIncome + UnemployRate + ManfEmploy + SpeakingNonEnglish +
    SocialSecurityRate + DisabilitiesRate + Homeowner + SameHouse1995and2000 +
    LandArea, data = elect.df.training)

Residuals:
    Min      1Q  Median      3Q     Max
-41.837  -7.759   0.346   7.897  35.413

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          -3.342e+01  1.563e+01  -2.138  0.03274 *
MalesPer100Females    2.539e-03  3.823e-02   0.066  0.94706
AgeBelow35            5.005e-01  1.254e-01   3.992 6.91e-05 ***
Age65andAbove        -1.681e-01  2.531e-01  -0.664  0.50662
Black                 8.309e-01  3.124e-02  26.599  < 2e-16 ***
Asian                -5.365e-01  2.197e-01  -2.441  0.01476 *
Hispanic              2.252e-01  7.333e-02   3.071  0.00218 **
HighSchool            5.573e-01  8.134e-02   6.852 1.13e-11 ***
Bachelors             6.575e-01  7.855e-02   8.370  < 2e-16 ***
Poverty              -1.312e+00  2.073e-01  -6.329 3.40e-10 ***
MedianIncome         -4.190e-04  7.860e-05  -5.331 1.15e-07 ***
UnemployRate          6.601e-01  2.456e-01   2.688  0.00728 **
ManfEmploy           -1.027e-01  4.710e-02  -2.181  0.02936 *
SpeakingNonEnglish   -1.391e-01  9.291e-02  -1.497  0.13464
SocialSecurityRate   -5.930e-05  2.430e-04  -0.244  0.80726
DisabilitiesRate     -1.151e-03  4.986e-04  -2.309  0.02112 *
Homeowner             8.075e-02  6.179e-02   1.307  0.19148
SameHouse1995and2000  3.849e-01  6.857e-02   5.614 2.43e-08 ***
LandArea              1.293e-03  2.401e-04   5.388 8.48e-08 ***
---
```

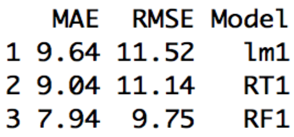
The second model is a regression tree which allows to rank the importance of the predictor variables based on their position within the tree but also allows for a fast prediction as no calculation is needed. The regression tree backs up the assumption that Black is the most important attribute for votes as it is the root node.



The final prediction model is a random forest. It was sued for its reputation of being very precise with predictions which is why it was used to predict the votes for Obama and Clinton for the counties which have not voted.

To configure my regression tree, the prune method has been used to find the cp and therefore have an optimal number of splits to avoid overfitting, this also allowed the improvement of the MAE and RMSE error rates. The ten-folds method was also applied to avoid overfitting. When it comes to the random forest, 500 trees were constructed to improve the error rates and increase the prediction precision.

The comparison of these three error rates supports the assumptions that the random forest has the highest precision followed by the linear regression (lm1) and the regression tree (rt1).



```
   MAE  RMSE Model
1 9.64 11.52   lm1
2 9.04 11.14   RT1
3 7.94  9.75   RF1
```

Conclusion

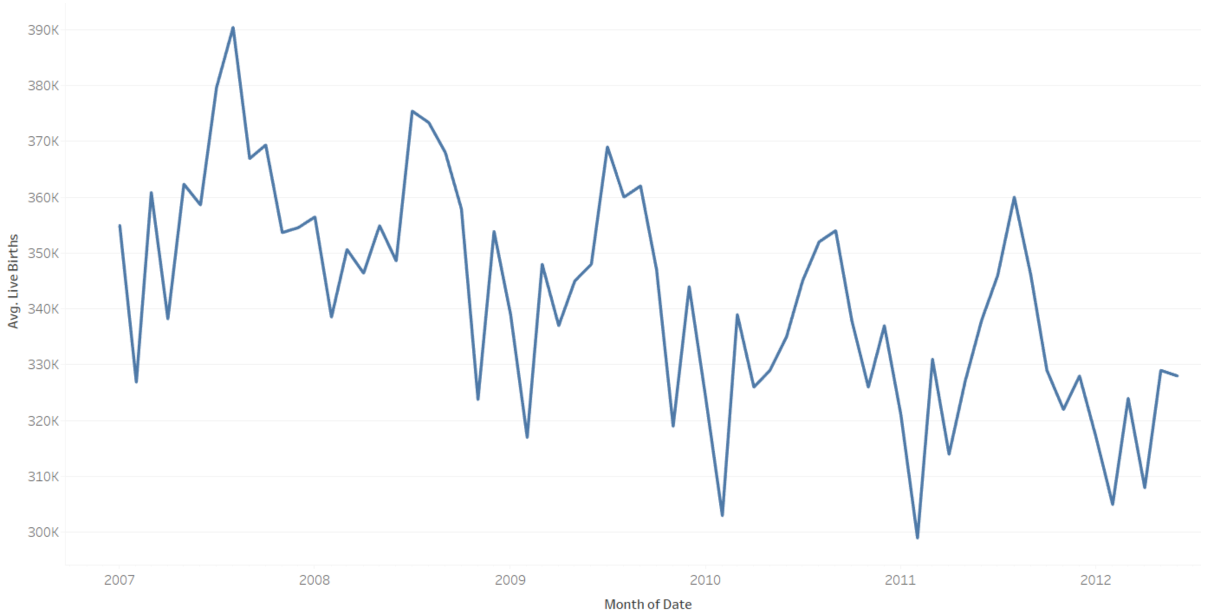
As identified and expected Obama has a very strong following amongst the black communities as well as in counties with a high level of education. Obama is not doing well with the older generation. This may be since he is black. This indicates that for elderly people the different race is a more negative factor than the gender of Hillary Clinton. However, as we can see in the linear regression the age group above 65 has a low significance for Obama and therefore, he should not suddenly focus on this group. But something that is apparent that the young age group below 35 has a high significance for his results and also has a high rate. But in the regression tree it shows not a very high importance. Therefore, Obama should focus more on targeting the young voters.

Word count : 1347

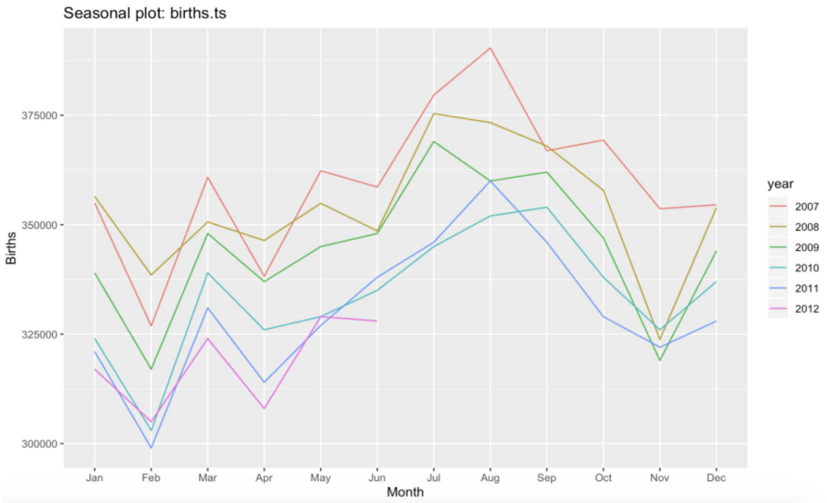
Part II: NICU Case Study

Investing US Births from January 2007 to June 2012

US Births data set is comprised of 66 observations of two variables which are number of births and date (given as numerical: YYYYMM). In order to perform statistical analysis in R, a separate date column attribute was created from the Yr_Mo column that will be used later in the analysis for merging datasets. A new time-series object was also created from January 2007 to June 2012 to plot the following time-series graph to have a general understanding of the trends and seasonality patterns.



An overall decline in birth rates can be seen from 2007 to 2012 however with an oscillating seasonal pattern. To further investigate these fluctuations a seasonal plot was created where data points are plotted against the seasons for separate years.



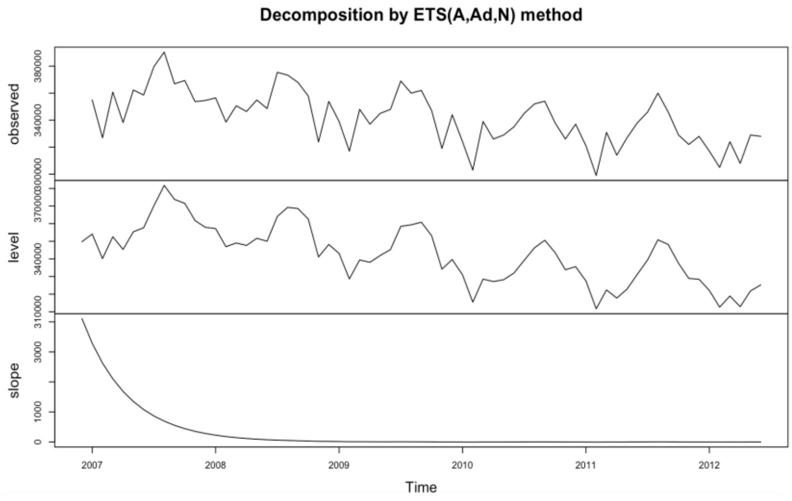
The reason for the decline in births were the effects of the Great Recession in 2008. This made people focus on their careers rather than personal relationships. It also makes many couples hesitant to get a baby which introduces additional costs. (Stack 2019)

The maximum births happen in the 7th-10th months. This is due to the seasons as people spend more time indoors during winter which indirectly increases their sexual activities. High summer temperatures also decrease the female fertility. (Anand 2000)

Forecasting US Births up to February 2013

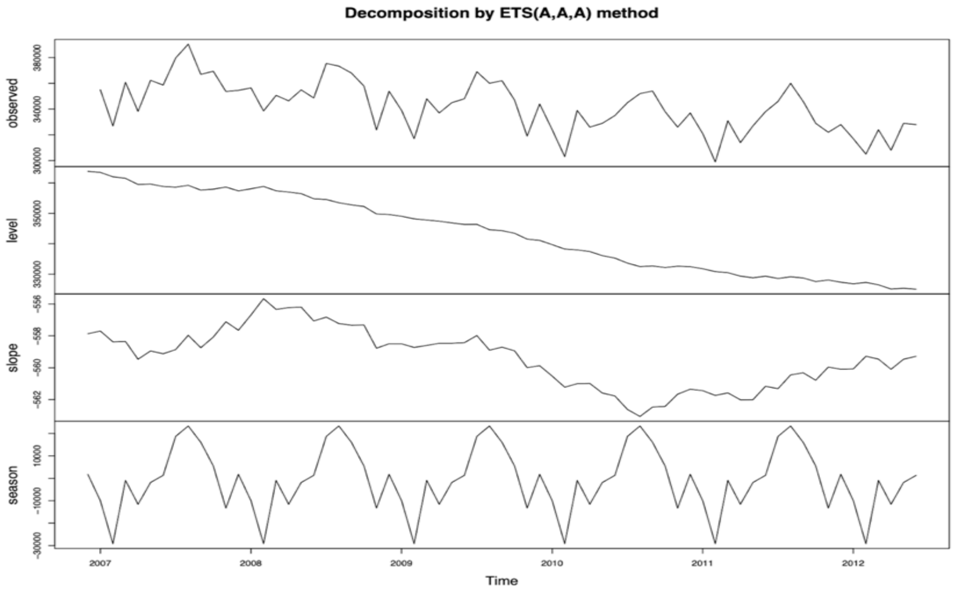
Both a trend (AAN) and a seasonality (AAA) model were created, along with their RMSE error rates, to see which model would be the best fit for US Births data set to predict births up to February 2013.

AAN Model



RMSE = 15551.1

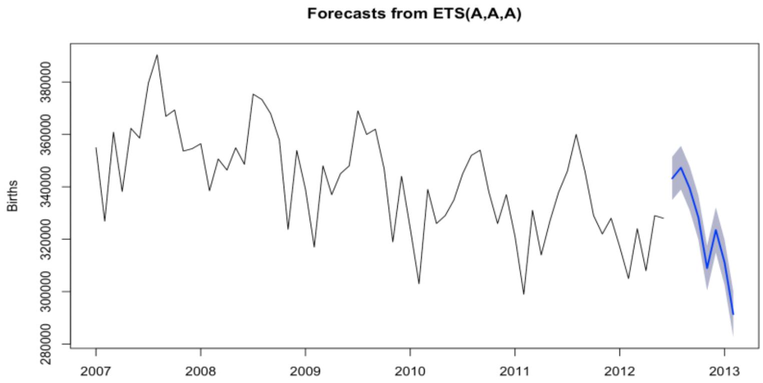
AAA Model



RMSE = 5602.903

The seasonality model is better model for forecasting than the trend model. It is our best model because it predicts based on seasonal variation, while forecasting in trend models tend to follow a linear trend based on tangible data points from the past.

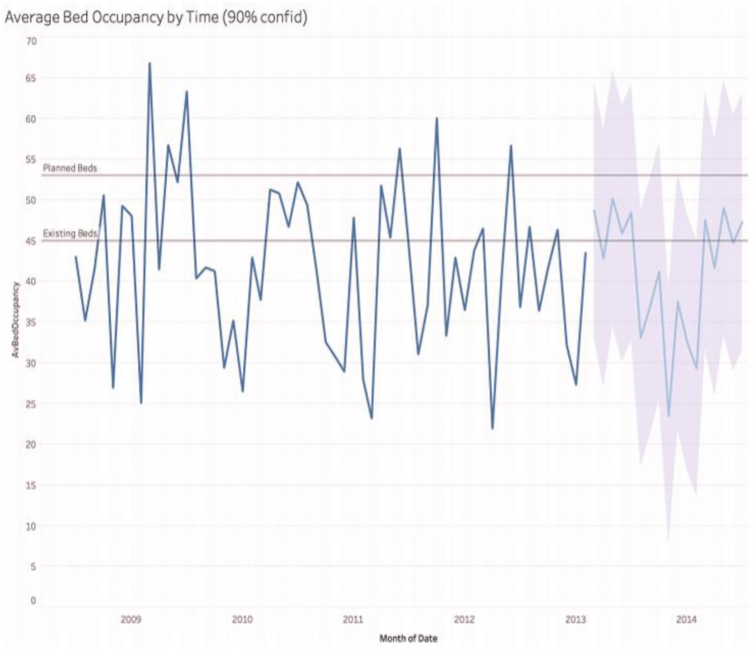
Forecasting in R



Above: forecasting plot of US Births from June 2012 to February 2013 with 80% confidence level done with the AAA model. Below are the predictions from June 2012 to February 2013:

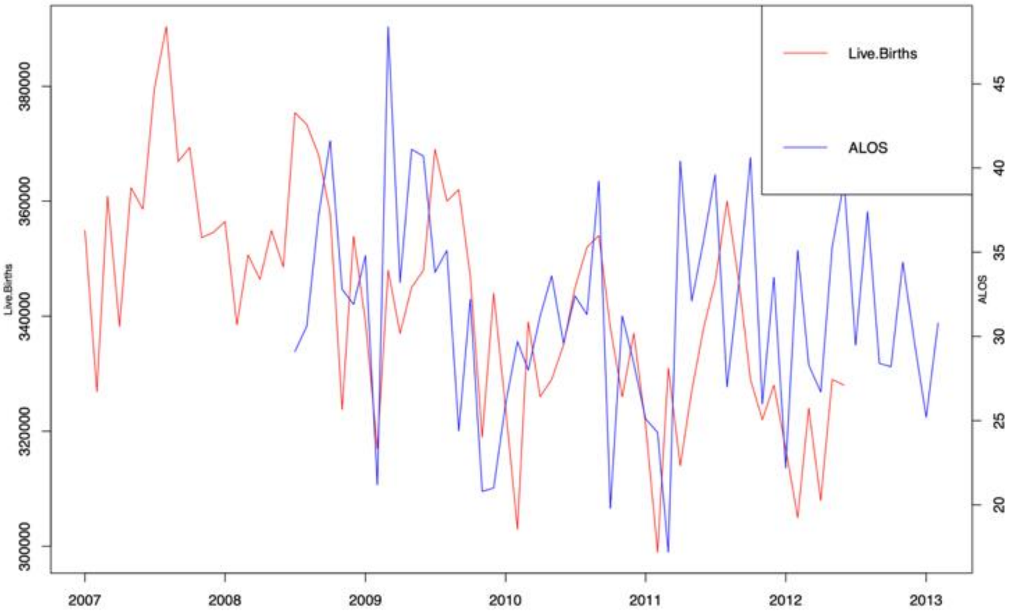
Jun 2012	343213.5
Aug 2012	347295.2
Sep 2012	339456.8
Oct 2012	328322.3
Nov 2012	308942.4
Dec 2012	323508.4
Jan 2013	311140.1
Feb 2013	291363.3

Forecasting in Tableau

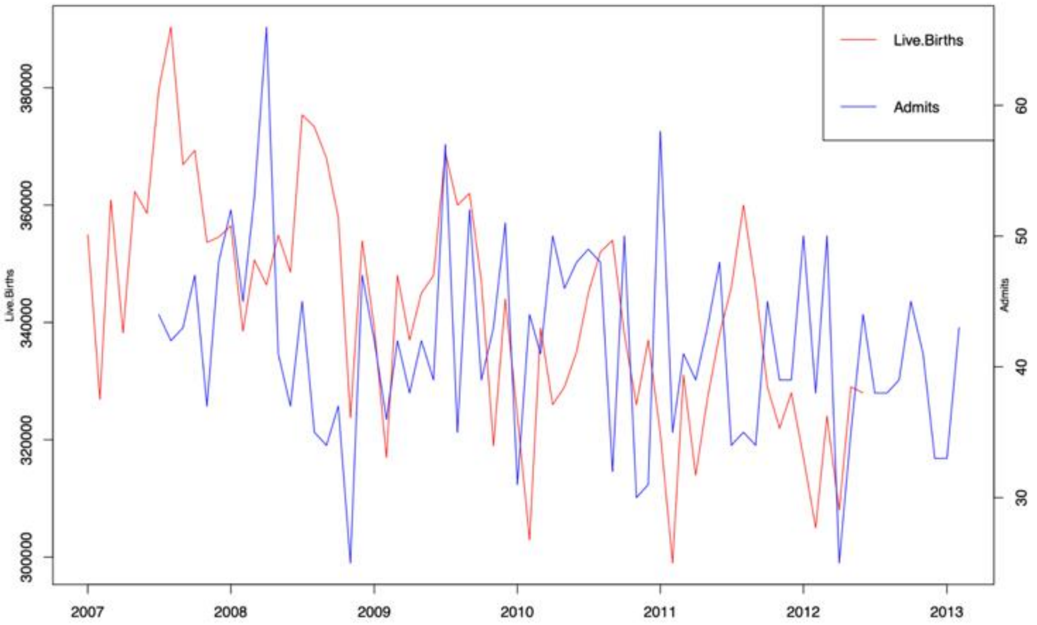


Comparing Seasonality Patterns in US Births, NICU Admissions and NICU ALOS

Live Births versus ALOS



Live Births vs Admits



	US Births	Admits	ALOS
Maximum	7th-10th	3rd-5th	4th-10th
Minimum	2nd-3rd	9th-11th	2nd-3rd

high temperatures and humidity in the 6th and 7th month make bacterial infections more likely for pregnant women leading to higher ALOS. Heatwaves also lead to more early births. These premature born babies need to stay for a long time in incubators to get strong.

Live.Births and Admits share a decreasing trend but they can diverge. Starting from 2011, the most admits occur on the beginning of the year. Before 2011, allergies are the main infant hospitalization reason during the 3rd to 5th months. However, since 2011 the flu season, starting in the beginning of the year became the most common reason of infant hospitalization.

Recommendation to the COO of Garfield Children's Hospital

Forecasts:

Mean Admissions ANN	Mean Admissions AAN	Mean ALOS ANN Forecast	Mean ANN Forecast Occupied Bed Days 06.2014
41.7	37.7	28.4	42.1

They should stop adding extra beds. As identified the downward trend of US birth rates indicate that the bed requirements will not increase. If possible, the hospital could change the amount of beds based on seasonality. Ergo increasing the capacity during the highest birth rate months between June or August. This is however unlikely as the beds will already be in the possession and in the hospital thus resulting in more costs and work to move them around.

Word count: 595; total = 1942

Code Appendix

PART I: Clinton Obama Case study

```
# Load the data file directly from the internet
library(curl)
library(rattle)
x <-
curl("https://raw.githubusercontent.com/awhiter/TeachingRepos/master/datasets/Obama.csv")
elect.df  <- read.csv(x)

# Create new derived target attributes

elect.df$ObamaRate <- 100 * elect.df$Obama / elect.df$TotalVote

elect.df$ClintonRate <- 100 * elect.df$Clinton / elect.df$TotalVote

elect.df$ObamaPercentMargin <- 100 * (elect.df$Obama - elect.df$Clinton) /
  elect.df$TotalVote
```

1. IMPUTING MISSING VALUES:

```
# Missing values for AverageIncome are replaced by the
# MedianIncome for that same record

elect.df$AverageIncome <- ifelse(is.na(elect.df$AverageIncome),
                                 elect.df$MedianIncome,
                                 elect.df$AverageIncome)

# Missing values for the following list of attributes
# are replaced by 0.

for (attr in c("Black","Asian","AmericanIndian","ManfEmploy",
               "Disabilities","DisabilitiesRate","FarmArea"))
{elect.df[[attr]] <- ifelse(is.na(elect.df[[attr]]),
                            0,
                            elect.df[[attr]])}

# There still remain several attributes with 1 or 2 missing values.
# It turns out that all these final missing values are in 2 records.
# The following codes removes these records entirely.
```

```
elect.df <- elect.df[is.na(elect.df$HighSchool)==FALSE,]
elect.df <- elect.df[is.na(elect.df$Poverty)==FALSE,]
```

2. TEST OF LINEAR REGRESSION

```
lm1 <- lm(ObamaRate ~ MalesPer100Females+AgeBelow35+Age65andAbove+
          Black+Asian+Hispanic+HighSchool+Bachelors+Poverty+
          MedianIncome+UnemployRate+ManfEmploy+SpeakingNonEnglish+
          SocialSecurityRate+DisabilitiesRate+
          Homeowner+SameHouse1995and2000+LandArea,
        data = elect.df.training)

summary(lm1)

# first set library path to include the following directory
# if running on Azure LinuxDataScience VM ...

.libPaths('/home/vmuser/R/x86_64-pc-linux-gnu-library/3.2')

# The Metrics package includes the mae and rmse functions.
# Install Metrics if needed..
# install.packages("Metrics")

library(Metrics)

genError <- function(prediction, actual)
  cat('MAE =', signif(mae(actual,prediction),4),
      ' RMSE =', signif(rmse(actual,prediction),4), "\n")

lm1.pred <- predict(lm1, elect.df.test)

genError(lm1.pred, elect.df.test$ObamaRate)

model.results <- data.frame(MAE=9.64, RMSE=11.52, Model="lm1")
```

3. REGRESSION TREE

```
rt1 <- rpart(ObamaRate ~ MalesPer100Females+AgeBelow35+Age65andAbove+
            Black+Asian+Hispanic+HighSchool+Bachelors+Poverty+
            MedianIncome+UnemployRate+ManfEmploy+SpeakingNonEnglish+
            SocialSecurityRate+DisabilitiesRate+
            Homeowner+SameHouse1995and2000+LandArea,
          data = elect.df.known, method = "anova", maxdepth = 18,minsplit = 15,cp = 0.001,
xval = 10)
```

```
rt1.pred <- predict(rt1, elect.df.test)
genError(rt1.pred, elect.df.test$ObamaRate)
model.results <- data.frame(MAE = 8.852, RMSE = 10.91, Model = "rt1")

optimalCP <- function(rt.model){
  df<-as.data.frame(rt.model$cptable)
  minerr <- min(df[,"xerror"])
  minerr.xstd <- df[df$xerror==minerr,"xstd"]
  df[df$xerror<minerr+minerr.xstd,][1,"CP"]} # We use this function to compute or optimal cp

optimalCP(rt1)


 rt1.opt <- prune(rt1, cp=optimalCP(rt1))
prp(rt1.opt, type = 1, extra = 1)

rt1.opt.pred <- predict(rt1.opt, elect.df.test)

genError(rt1.opt.pred, elect.df.test$ObamaRate)


plotcp(rt1,upper = "splits")
model.results <- rbind(model.results, data.frame(MAE=9.04,RMSE=11.14, Model="RT1"))
```

#We now will code our third model a random forest which will normally give us a better accuracy than
# our regression tree as:

```
rf1 <- randomForest(ObamaRate ~ MalesPer100Females+AgeBelow35+Age65andAbove+
                    Black+Asian+Hispanic+HighSchool+Bachelors+Poverty+
                    MedianIncome+UnemployRate+ManfEmploy+SpeakingNonEnglish+
                    SocialSecurityRate+DisabilitiesRate+
                    Homeowner+SameHouse1995and2000+LandArea,
                   data = elect.df.training , ntree = 500)

rf1.pred <- predict(rf1, elect.df.test)
genError(rf1.pred, elect.df.test$ObamaRate)

model.results <-rbind(model.results, data.frame(MAE = 7.94, RMSE = 9.75, Model = "RF1"))
```

#Visualisation of our Regression tree:

```
library(rattle)
fancyRpartPlot(rt1, main = "ObamaRate prediction regression tree", palettes = c("Blues"))
```

PART II: NICU case study

```
# First, import the US Births (births.df) and NICU (baby.df) datasets to R
```

```
# Install the following packages
library(dplyr)
library(ISLR)
library(ggplot2)
library(zoo)
library(forecast)
library(repr)
```

US BIRTHS

1. CLEANING UP THE DATA

```
# Format the date variable to create a separate “date” attribute
births.df$Date <- as.Date(paste(as.character(births.df$Yr_Mo),"1",sep=""),
                      format="%Y%m%d")
```

```
# Delete the old Yr_Mo column after date attribute has been created
births.df = select(births.df,-c(1))
```

2. ANALYZING TRENDS AND PATTERNS IN THE DATA SET

```
# Create a time-series object from the Live Births variable
births.ts <- ts(births.df$Live.Births, start = c(2007, 1), end = c(2012, 6), freq = 12)
```

```
# Plot a time-series graph to visualize patterns
autoplot(births.ts, ylab = "Births")
```

```
# Create a seasonal Plot
ggseasonplot(births.ts, ylab = "Births")
```

3. FORECASTING

```
# Create RMSE function by taking the square root of the mean square error and assigning it to a new function
rmse.ets <- function (etsmodel) cat("RMSE = ", sqrt(etsmodel$mse))
```

```
# Create an ANN model and plot it
(births.ets.ANN <- ets(births.ts, model = "ANN"))
plot(births.ets.ANN)
```

```
# Create an ANN model and plot it
(births.ets.AAN <- ets(births.ts, model = "AAN"))
plot(births.ets.AAN)
```

```
# Create an AAA model and plot it
births.ets.AAA <- ets(births.ts, model = "AAA")
plot(births.ets.AAA)
```

```
# Calculate the RMSE of both models and compare
rmse.ets(births.ets.AAA)
rmse.ets(births.ets.AAN)
```

```
# Forecasting done with both models
births.forecast.AAA <- forecast(births.ets.AAA, h = 8, level = c(80))
births.forecast.AAN <- forecast(births.ets.AAN, h = 8, level = c(80))
```

```
# Plot side by side the prediction models for both AAA and ANN
par(mfrow = c(1, 2))
plot(births.forecast.AAA,ylab="Births")
plot(births.forecast.AAN,ylab="Births")
```

```
# Print the predictions of both models for the following 8 months
births.forecast.AAA$mean
births.forecast.AAN$mean
```

NICU

4. CLEANING UP THE DATA

```
# Format the date variable to create a separate “date” attribute
baby.df$Date <- as.Date(paste(as.character(baby.df$Year), baby.df$Month,"1",sep=""),
                    format="%Y%b%d")
```

```
# Create a time-series object from the variable Admits
admits.ts <- ts(baby.df$Admits, start = c(2007, 7), end = c(2013, 2), freq = 12)
```

```
# Plot a time-series graph of Admits
autoplot(admits.ts, ylab = "Admits")
```

```
# Merge the baby.df and births.df datasets
merged.df <- merge(births.df, baby.df, by="Date", all=TRUE)
```

```
#Plot the  merged datasets: Live Births vs Admits
options(repr.plot.width=9, repr.plot.height=4)
plot(merged.df$Date, merged.df$Live.Births, type="l", col="red", xlab=NA, ylab=NA)
par(new = T)
plot(merged.df$Date, merged.df$Admits, type="l", col="blue", axes=F, xlab=NA, ylab=NA)
plot(merged.df$Date, merged.df$ALOS, type="l", col="green", xlab=NA, ylab=NA)
axis(side = 4)
mtext(side = 4, line = 0, 'Admits', cex=0.75)
mtext(side = 4, line = 0, 'ALOS', cex=0.75)
mtext(side = 2, line = 2, 'Live.Births', cex=0.75)


legend("topright",
     legend=c("Live.Births", "","Admits","","ALOS"),
     col=c("red", "white","blue","green"),lty=c(1,1,1,1))
```

```
# Plot merged datasets: Live Births vs ALOS
plot(merged.df$Date, merged.df$Live.Births, type="l", col="red", xlab=NA, ylab=NA)
par(new = T)
plot(merged.df$Date, merged.df$ALOS, type="l", col="blue", axes=F, xlab=NA, ylab=NA)
axis(side = 4)
mtext(side = 4, line = 0, 'ALOS', cex=0.75)
mtext(side = 2, line = 2, 'Live.Births', cex=0.75)
```

```
legend("topright",
       legend=c("Live.Births", "","ALOS"),
       col=c("red", "white","blue"),lty=c(1,1,1))


# Create a scatterplot of Admits vs Live Births Scatterplot
plot(merged.df$Live.Births, merged.df$Admits, col="blue", xlab="Live Births", ylab="Admits")


# Create a scatterplot of ALOS vs Live Births Scatterplot
plot(merged.df$Live.Births, merged.df$ALOS, col="blue", xlab="Live Births", ylab="ALOS")


#Create a scatterplot of ALOS vs Admits Scatterplot
plot(merged.df$Admits, merged.df$ALOS, col="blue", xlab="Admits", ylab="ALOS")
```

Bibliography

Anand, K. 2000. *Indian Pediatrics.* 03. Accessed 17 02, 2020. https://indianpediatrics.net/march2000/march-306-312.htm.

Holland, Kimberley. 2019. *Healthline.* 22 01. Accessed 02 09, 2020. https://www.healthline.com/health-news/why-does-the-u-s-have-such-a-low-birth-rate.

Levy, Michael. 2009. *Encyclopedia Britannica.* 09 11. Accessed 02 16, 2020. https://www.britannica.com/event/United-States-presidential-election-of-1992.

Stack, Liam. 2019. *New York Times.* 17 05. Accessed 02 09, 2020. https://www.nytimes.com/2019/05/17/us/us-birthrate-decrease.html.