Data Analytics 2 Group Project

Words: 1374
Words: 598

Part I Obama-Clinton Case Study

Section 1 - The Problem

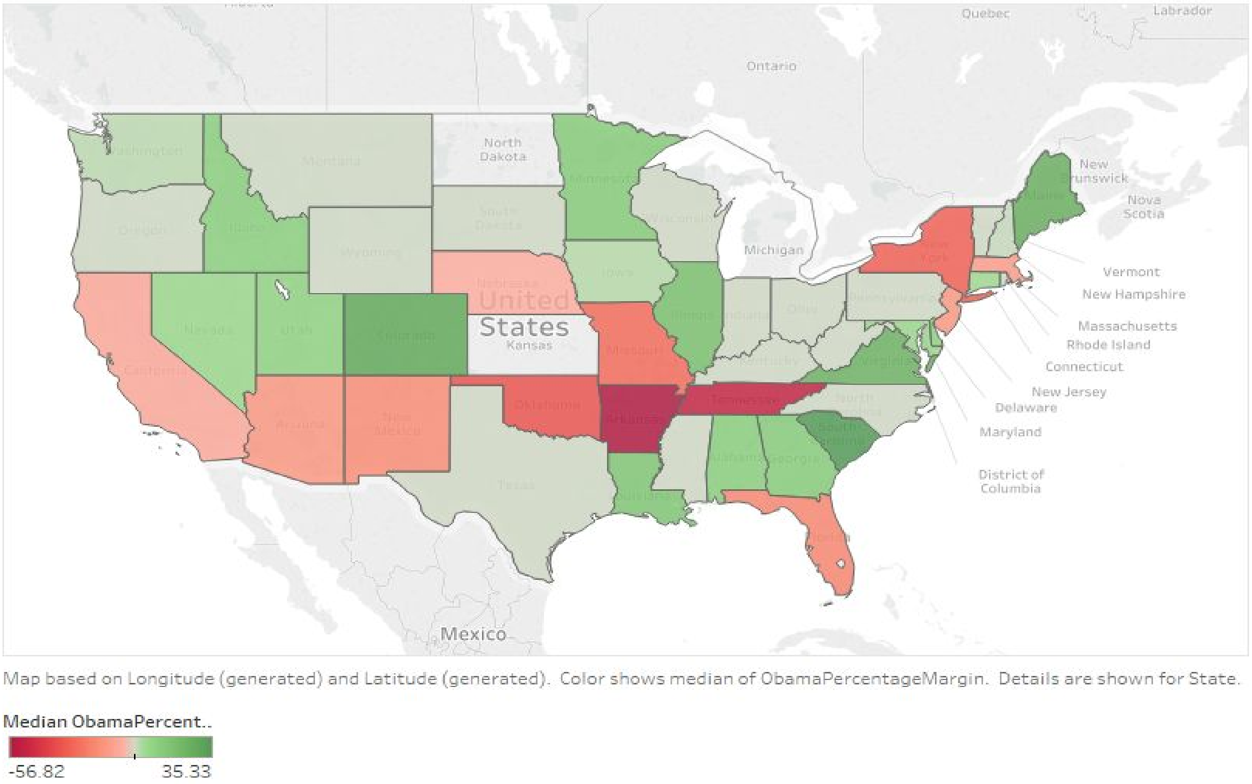
Our general problem is to identify and predict who Obama's team need to target to increase their chance of winning different counties. The target attribute we will focus on is the Obama margin rate. Swing states are the ones that we will be focusing on to see how they can ensure win. We want to specifically, explore how income and income-related variables impact the Obama margin and how this differs across parts of the US. Ultimately, we want to predict how the remaining counties will vote and use this to understand what variables are important in this prediction.

Target Attribute (1.1): (Obama Vote - Clinton Vote)*100/Total

Section 2 - Understanding the Data

The dataset has 41 variables with 2866 entries and it comes from the US Census Bureau. It contains explanatory data for each county, and who they voted for. It also contains data on those yet to vote.

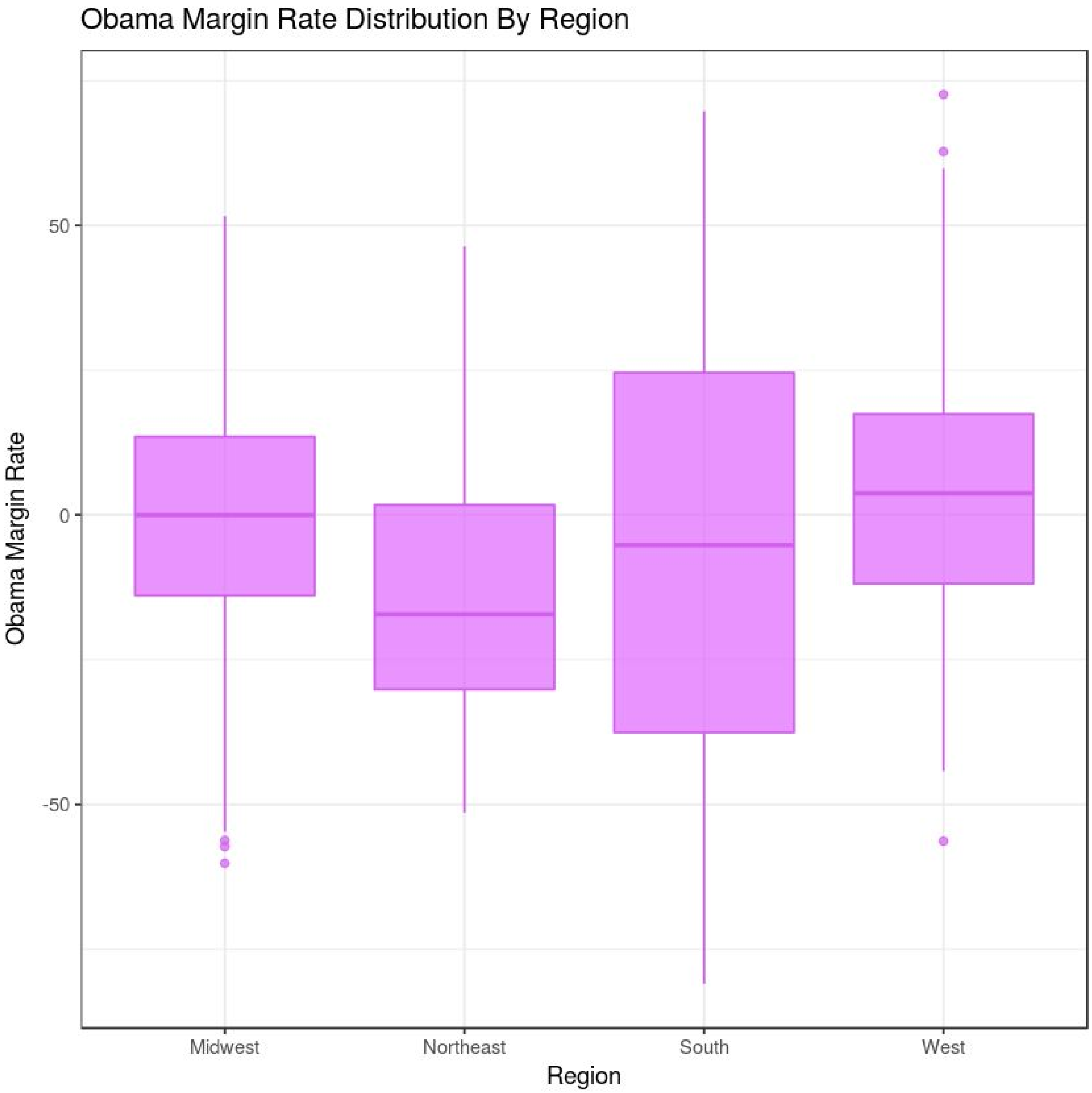
ObamaPercentageMargins by state



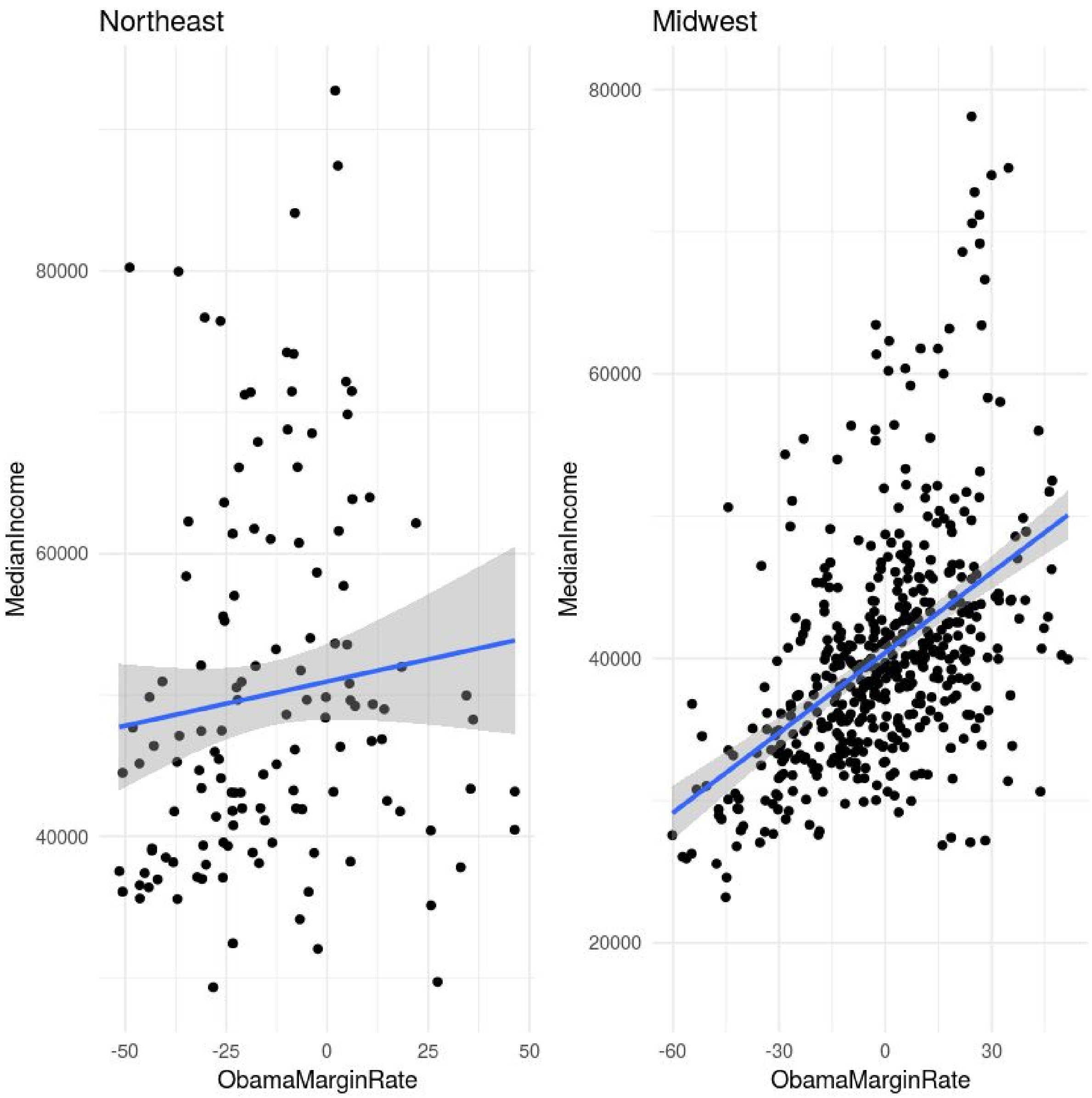
Map based on Longitude (generated) and Latitude (generated). Color shows median of ObamaPercentageMargin. Details are shown for State.

The map above shows the Obama margin between the states that have voted. As we can see a lot of the Northeast states are very tight with the potential of being swing states and thus of potential importance to Obama's administration..

We then looked at the distribution of the Obama margin rate by region to see which regions had the tightest margins and understand how different regions impacted the difference in the vote.



We see that the Midwest and the Northeast regions have the tightest distribution, and in fact, the margin rate of the Midwest region is actually 0. Whereas the Northeast and southern regions are actually ones where Obama is losing slightly on average, thus it would be interesting to explore variables that hold significant impact in these regions.

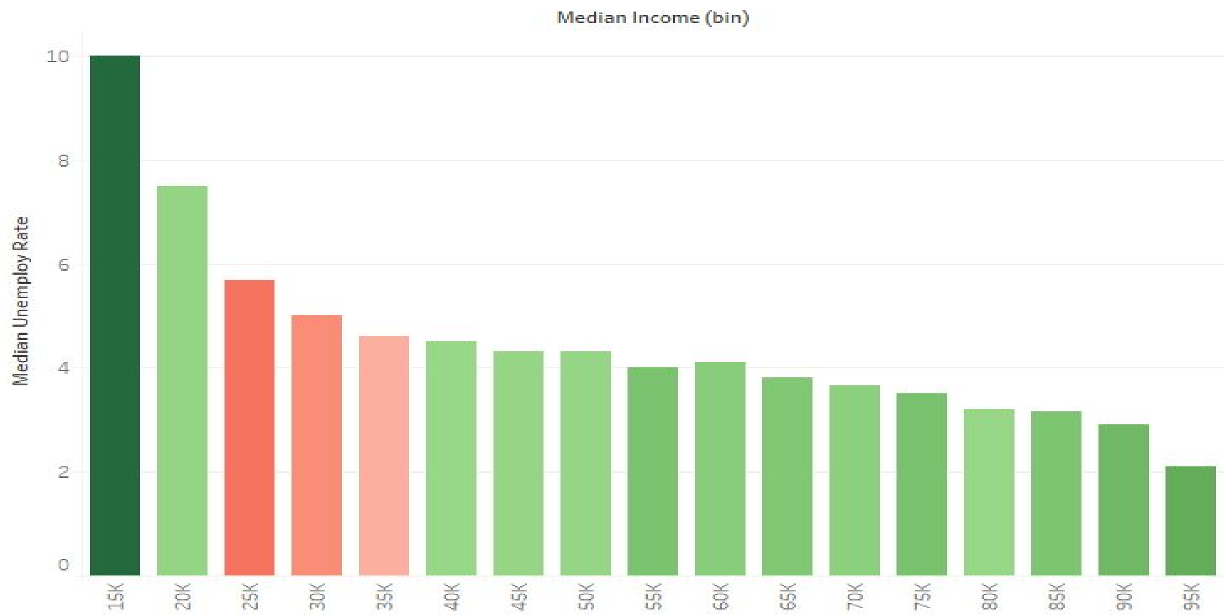


We then compared how the Obama margin is impacted by the income of individuals in different regions. As we can see income has a significantly stronger correlation in the Midwest region compared with the Northeast region where the correlation is a lot weaker. This suggest that income holds more impact in the Midwest regions.

How income variables impact Obama margin?

This graph shows the relationship between the unemployment rates, median income and the votes. As income increases the unemployment rate is lower, which is to be expected, we also see that it is the lowest income and highest income people that are voting for Obama, whilst the mid-income individuals tend to vote for Clinton.

Relation between Income, education and voting for Obama



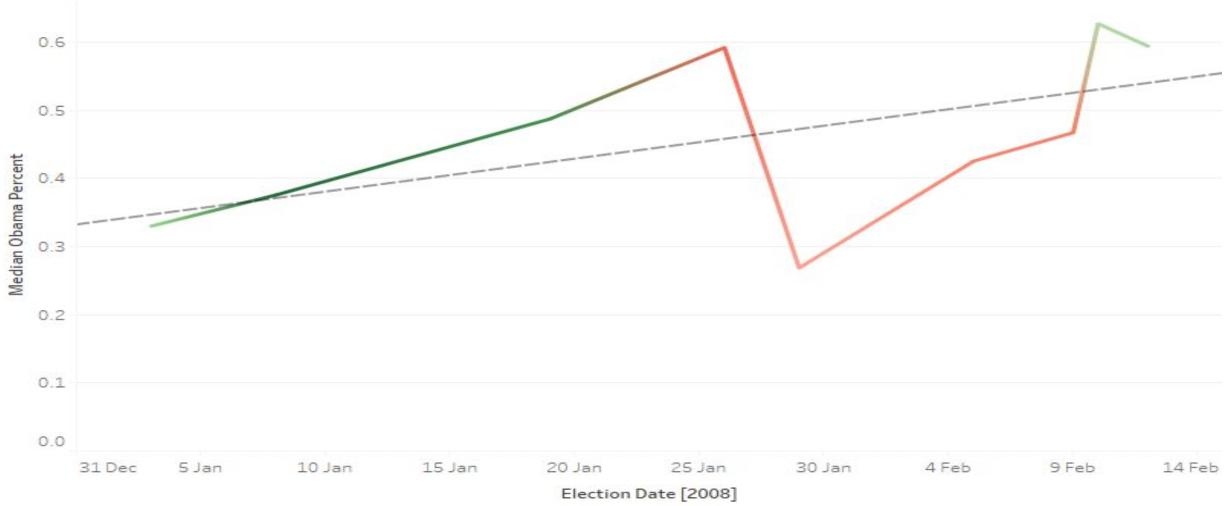
Median of Unemploy Rate for each Median Income (bin). Color shows median of ObamaPercentageMargin. The data is filtered on MalesPer100Females (bin), which excludes 191.1. The view is filtered on Median Income (bin), which excludes Null.

Median ObamaPercent..

-21.78 54.36

We then analyzed this relationship over time to see how it changed with each election date. We see that Obama was winning the first elections, which also happened to be in relatively higher income counties, whereas, he later began losing votes in Mid-income counties, showing that generally, higher income individuals were voting for Obama.

Income and how it impacts Obama rate over time



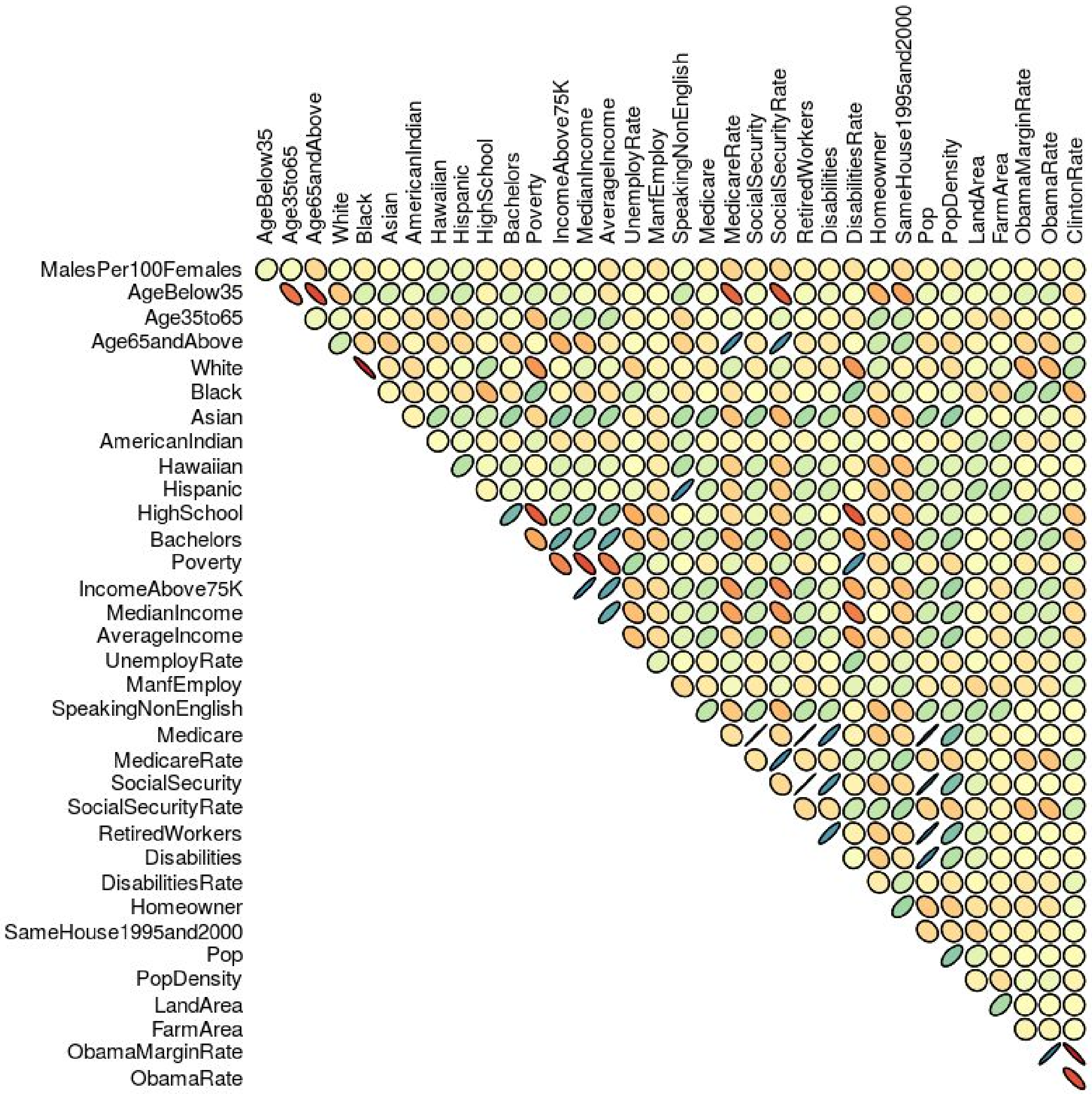
The trend of median of Obama Percent for Election Date. Color shows median of Median Income.

Median Median Income

29,581 50,335

Correlations analysis

We then wanted to understand how different variables in the dataset are related, to do this we used correlation plots and matrixes to understand these relationships (1.2).



Using these correlation matrices and this correlogram we identified some of the variables that we highly correlated to each other. For example, there was a correlation between similar variables, these are variables that described the same thing, but in different ways, for example, Disabilities and Disabilities Rate where the first is the absolute values of disabled people, and the second is the percentage. There was also a correlation between related variables, for example, being a bachelor was highly correlated to having a higher income. This shows us how some of these predictor variables are related which proved to be useful in choosing variables for our model.

Section 3 - Preparing the Data

First, we identified any missing values within the dataset. For example, the average income variable had 30 missing values. We chose to remove the average income variable as this was correlated with median income variable, as such we can use this instead. For missing values where variables had very few missing values, we replaced these with 0s as this tends to be a common reason for missing values. We also identified that the remaining NA values were in 2 records, thus, we removed those records (2.1).

Our chosen target attribute is the Obama rate margin, this was derived on both tableau and R using the the Obama, Clinton and Total votes variable found in the original data set. (Obama - Clinton) * 100/ Total Vote.

We then went on to split our data into train and test datasets. This is to allowed us to calculate the error of our models and compare how well they were doing. We first converted the election data variable into an actual date. Then we used this to separate the counties that had voted with ones that hadn't. Taking the counties that had voted we randomly split these into a train and test (75:25) (2.2).

Section 4 - Prediction Models

The first model chosen is the linear regression model. This model is relevant to the problem as it regresses the relation between our predictor variables (e.g. Income, Region etc.) and then it provides us with a relationship between these variables and our target attribute Obama margin rate.

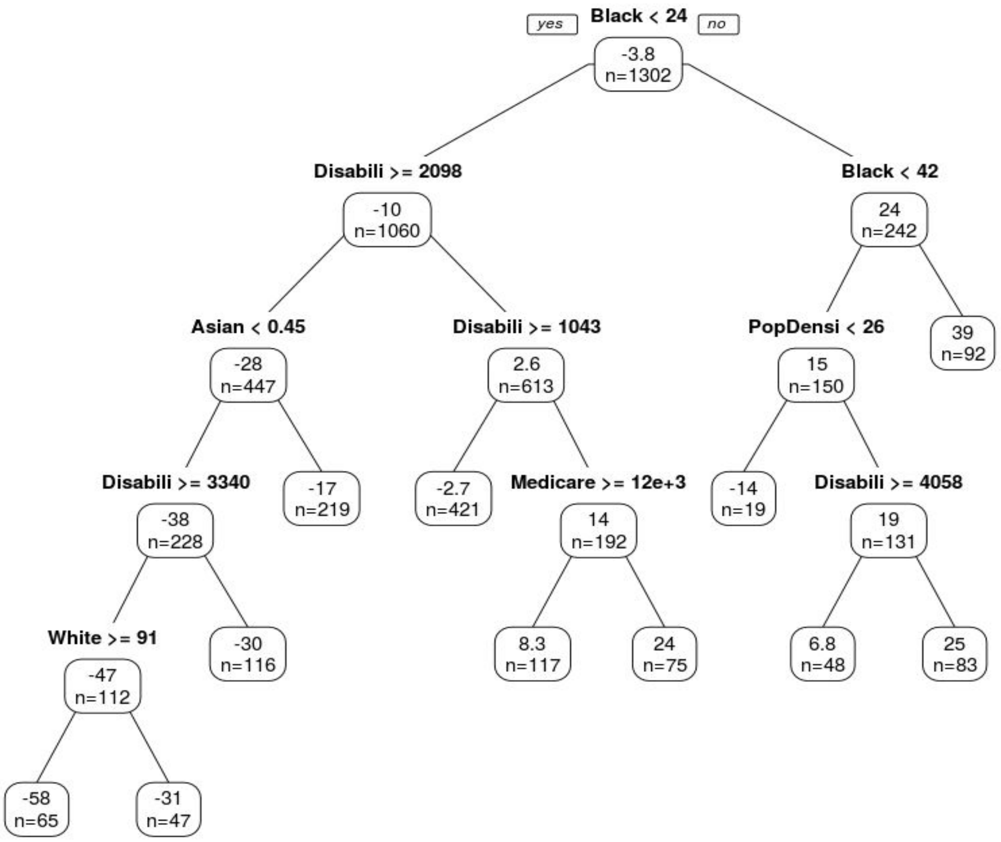
We first did a linear model with all the variables. However, we found that the output in terms of the relationships between the predictors and the target attributes was not consistent with the data description. To identify what attributes are relevant in predicting the Obama rate margin, we looked at the correlation between predictor variables, we then removed the variables that were strongly correlated as these impact the model. For example, disabilities and disability rate. Also, including multiple income variables made the output inconsistent thus we only included median income (3.1).

We then compared this model to a linear model formed backwards, which takes in these variables and removes variables that are not relevant in prediction. We also compare this to a forward linear regression model which starts with the minimum linear models and adds variables to reduce error.

Model	RSME	MAE
Linear Model (All Variables)	19.06	15.23
Linear Model (Chosen Variables, see 3.1)	21.43	16.87
Linear Model Backwards	19.2	15.35
Linear Model Forward	19.3	15.43

The linear model that predicted best was the one with all the variables, however, a key issue with this model is that the model correlations were not consistent with the data.

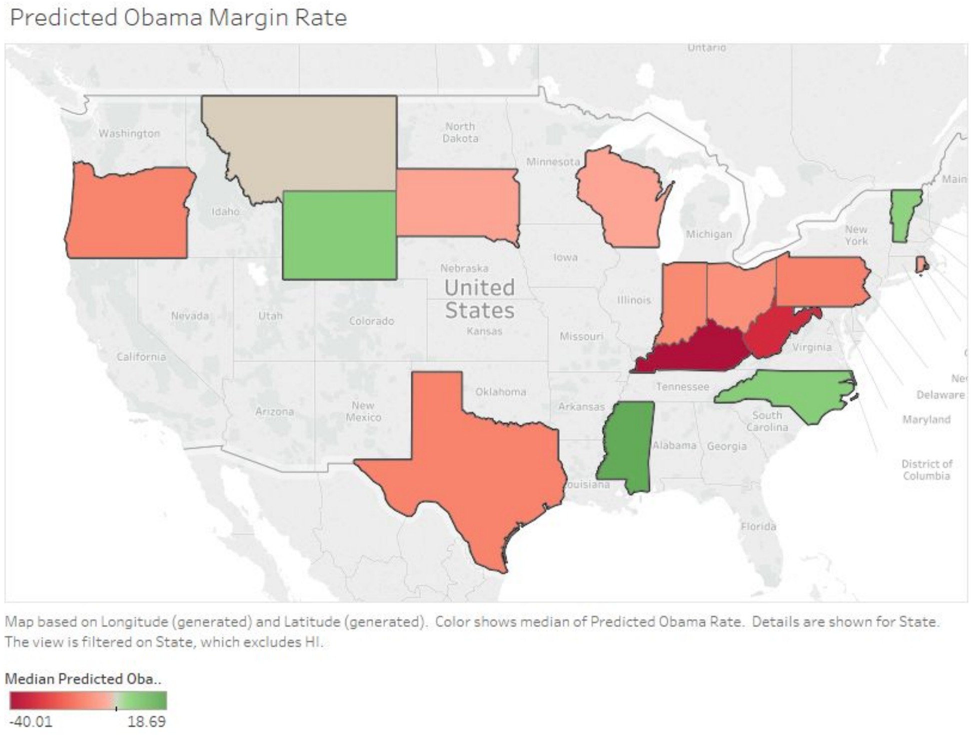
The second chosen model is a classification tree. These are easy to interpret algorithms identify the best ways to split data, and then allow you to predict new data. This model was constructed by looking at a tree with a large number of splits and then pruning this tree which means swapping decision nodes with leaf nodes to reduce the chance of overfitting the tree.



Model	RSME	MAE
Best Linear Regression Model	19.06	15.35
Regression Tree Model	22.32	17.67
LASSO Model	18.72	14.99

The final model is a Lasso model, this model regularises the coefficients of the standard linear model by reducing the impact of outliers using a scalar factor of lambda. This improved the accuracy of the original linear model and the Lasso model was actually our best model in terms of reducing error (3.3).

Section 5 - Conclusions and Recommendations



Map based on Longitude (generated) and Latitude (generated). Color shows median of Predicted Obama Rate. Details are shown for State. The view is filtered on State, which excludes HI.

Key conclusions:

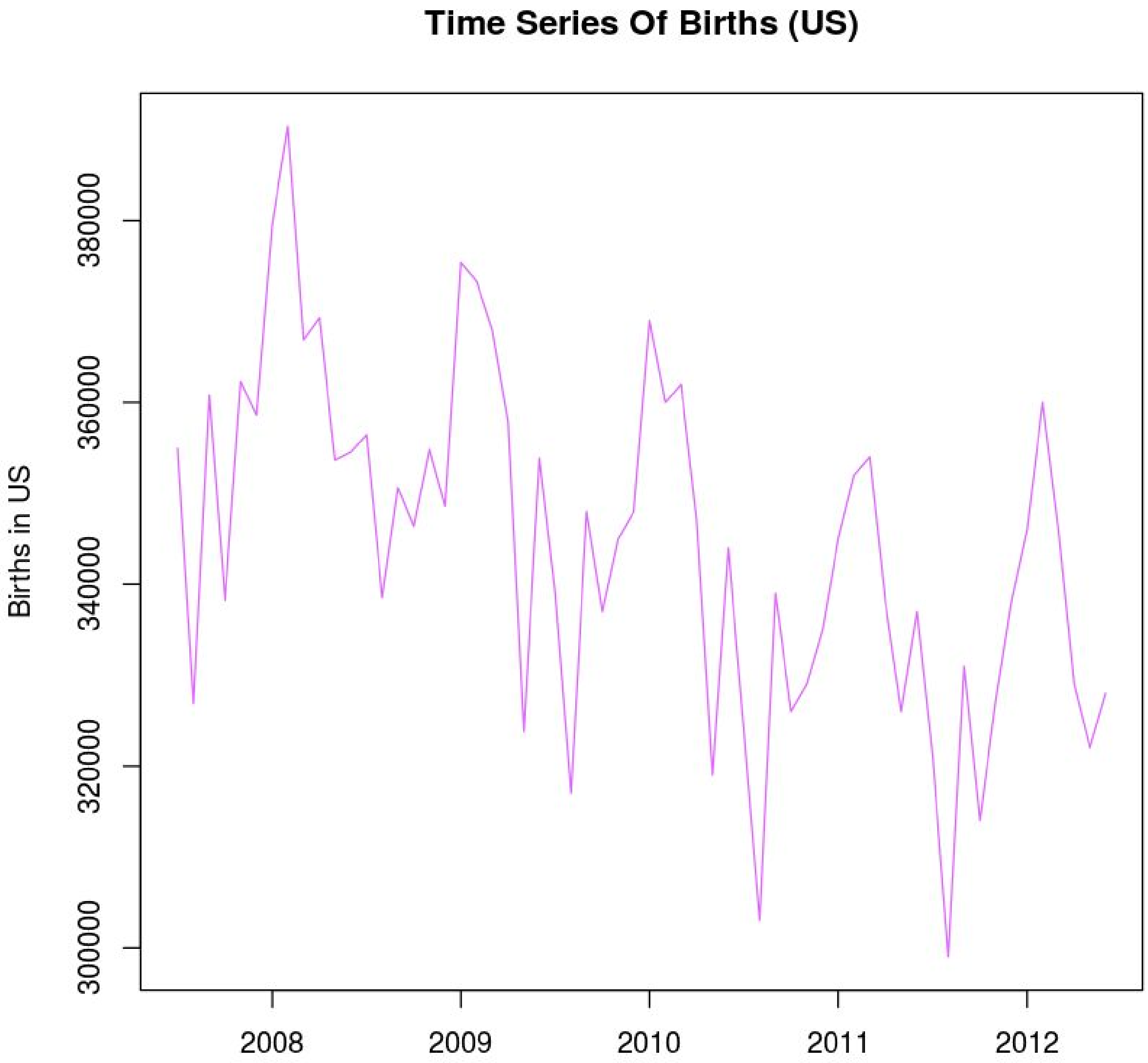
1. Key Regions: Northeast, South
2. Generally, very low income and high income individuals will vote for Obama, where as the mid-income individuals side more with clinton.
3. We see that income is a key variable that is positively correlated with the Obama margin rate. This is especially true for the Midwest region.
4. We also note that some of the other key variables that impact the Obama margin are: Black, Disability and Asian demonstrating that race oriented variables significantly influence Obama's margin.

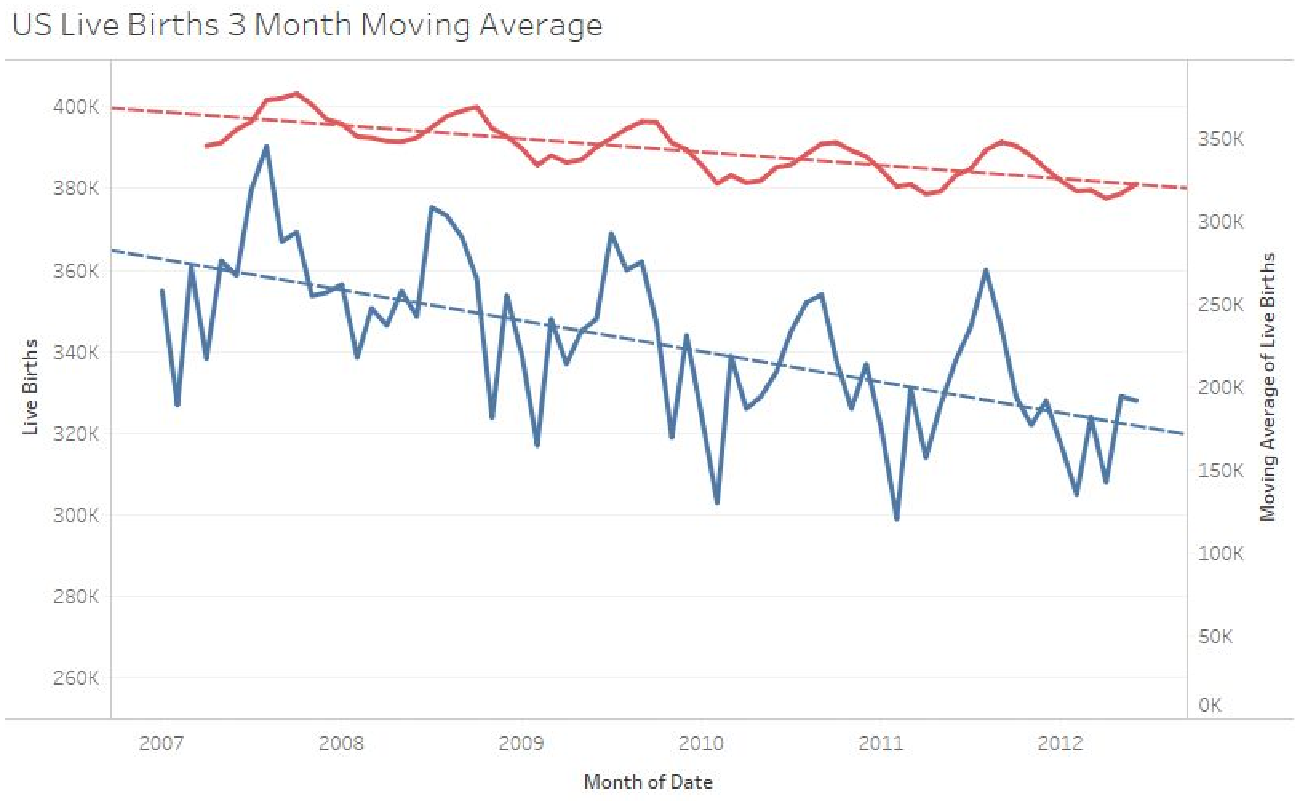
Recommendations:

For the Obama administration we recommend they focus on the key counties in the Midwest and Northeast region that have not voted. These are regions where many of the swing states occurred. To help influence the voting in Obama's favour we recommend focusing on income variables (e.g. Social security, median income, poverty etc.) and with in these we see the most value in focusing on mid-income individuals. This is where Obama has been losing most of his votes, and thus a key variable to target.

PART II - US Births Data
Data Exploration

US births is a time series dataset that contains the number of births in the US every month from 01/2007 to 06/2012.
First we used R to separate the year and month into two different columns (4.1) which allowed us to perform our time series analysis.





The trends of Live Births and Moving Average of Live Births from the previous 3 to the next 0 along Month of Date for Date Month. Color shows details about Live Births and Moving Average of Live Births from the previous 3 to the next 0 along Month of Date.

Measure Names
🟦 Live Births
🟥 Moving Average of Live Births from the previous 3 to the next 0 along Month of Date

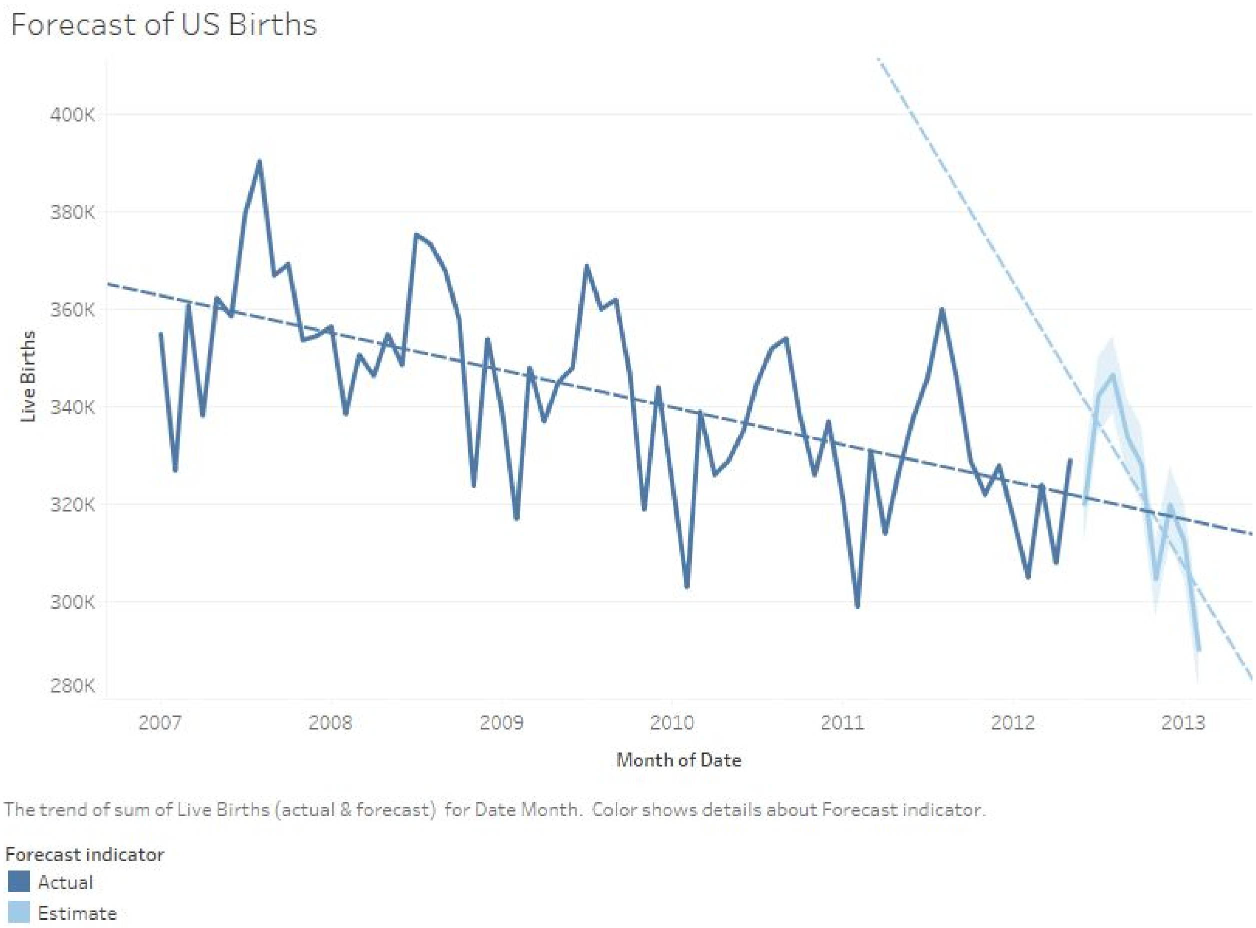
From these plots we observe a downtrend in US births. This is mainly due to the fact people are in times of economic uncertainty which leads to reduced conception (Rettner, 2019).

We also see a recurring pattern which could represent seasonality. If we look at each year we see that the highs occur toward the mid-end of the year, whilst the lows occur at the start. We also calculated the average monthly births across this time period and found that the highest was August. After looking at external studies it turns out that there are some key biological factors that lead to this trend. In winter males produce higher quality sperm, and women's ovum is in a better condition for eggs (Nelson, 2017). These factors together increase the chances of conception during winter, thus, increasing the number of births in late summer (Tita, 2001).

Models

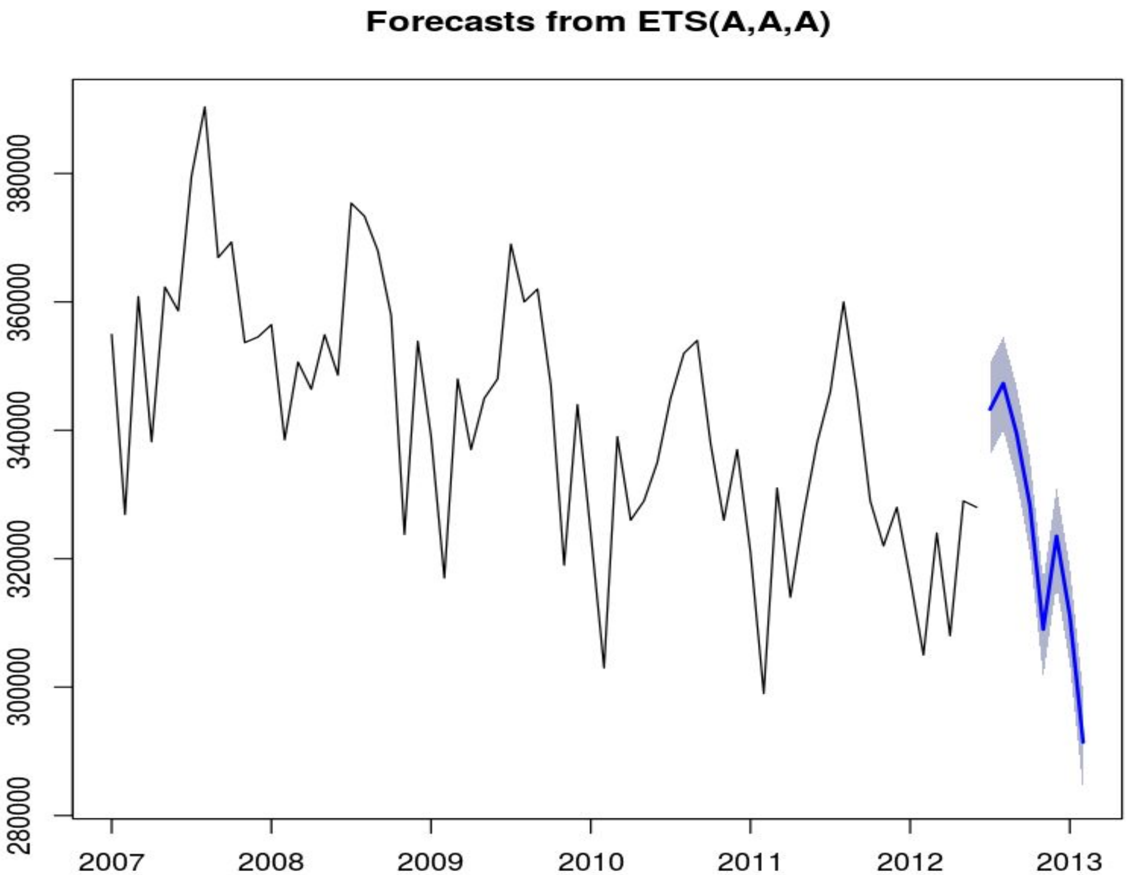
Model	RSME
AAN Model	15584
AAA Model	5622

We ran two exponential smoothing models which assign smaller weights to older data, as generally more recent data has stronger predicting power. The AAA model accounts for both seasonality and trends in data, whilst AAN only accounts for trends. As discussed before, the data has both a downtrend and seasonality, thus the AAA model predicted better.
Using R and Tableau we forecasted US births to 02/2013. We see that the forecast holds similar seasonality and also maintains the downtrend although it is steeper in the forecast. Forecast mean: 291394.

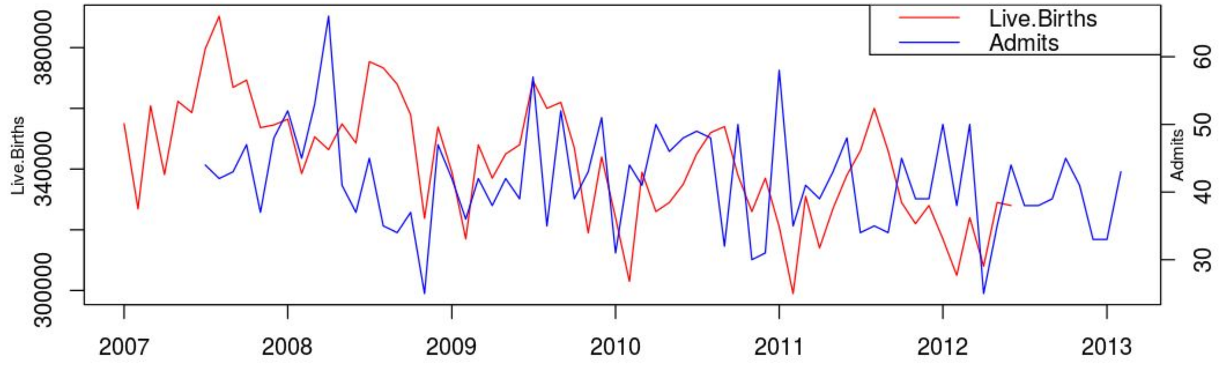


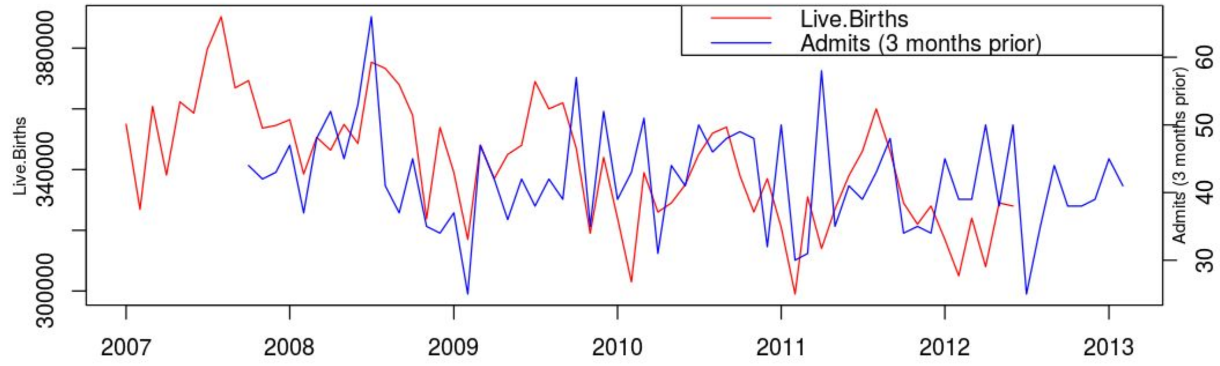
Forecast of US Births

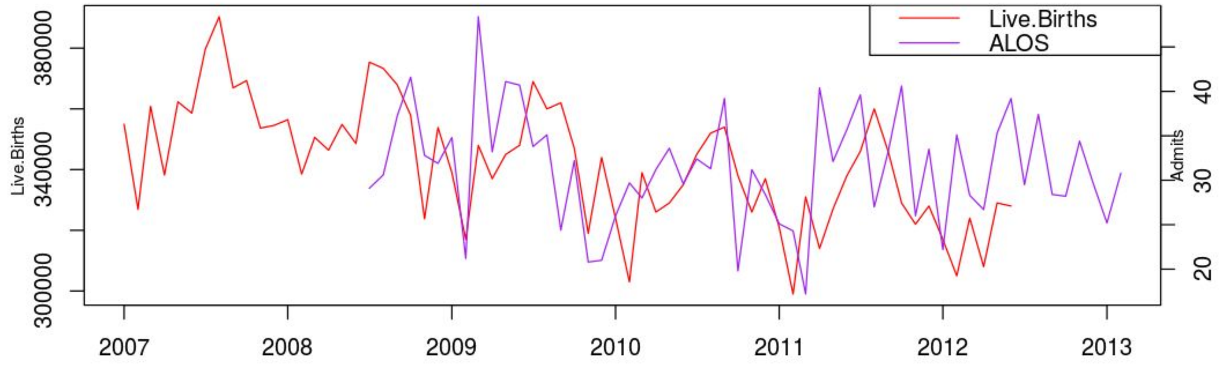
The trend of sum of Live Births (actual & forecast) for Date Month. Color shows details about Forecast indicator.

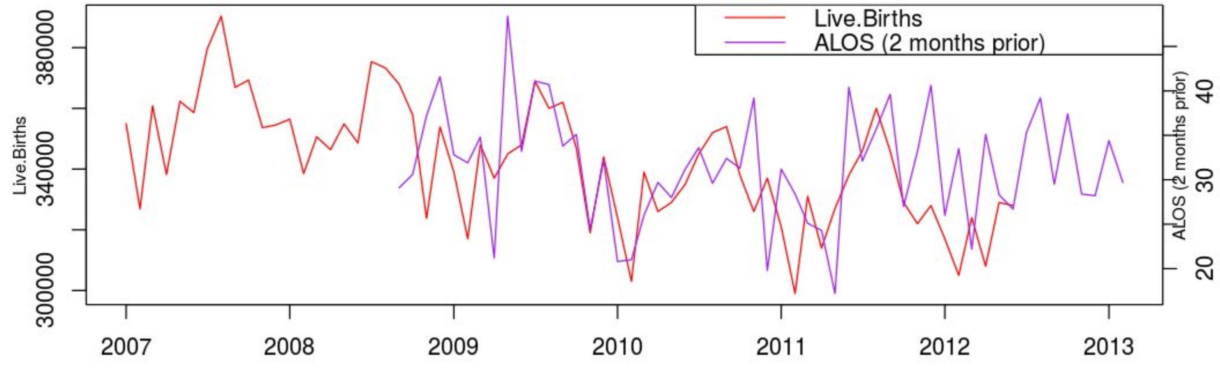

Forecasts from ETS(A,A,A)
380000
360000
340000
320000
300000
280000
2007
2008
2009
2010
2011
2012
2013

Comparing US Births with NICU data

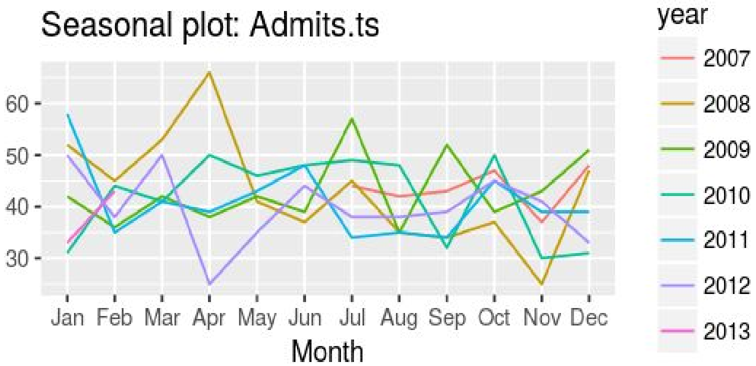


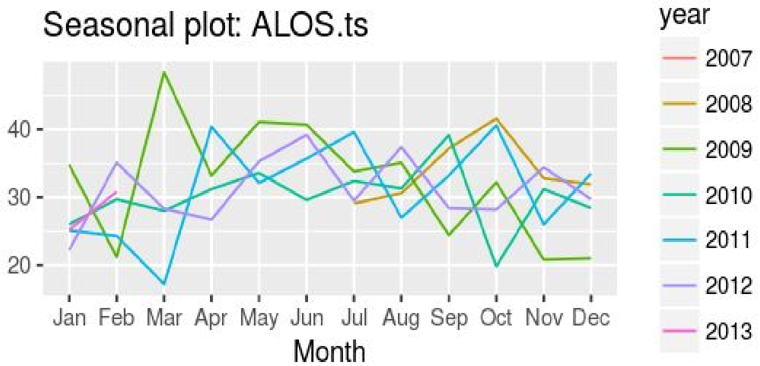


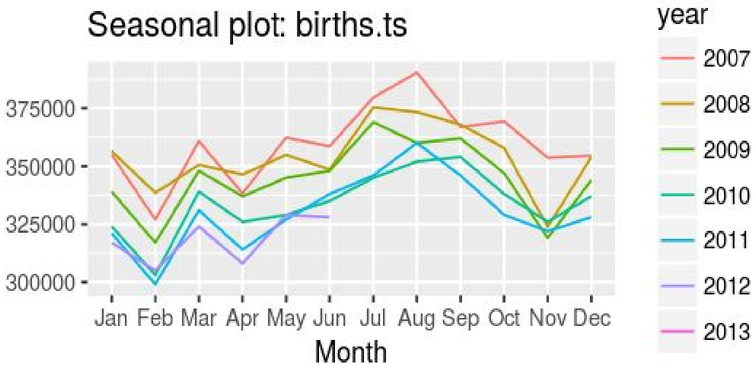




We compared the time series of plots of live births against both ALOS and admissions. We observed that the peaks and lows were happening 2-3 months apart. Thus, we identified the optimal offset by comparing the correlations (4.4). We then compared the monthly averages of each variable to identify the months where the maximum and minimum occurred (see table). We see that both of these time series follow similar seasonal trends (below). We observe that when births are low so to are ALOS and admissions with a 2- 3 month lag. However, there is a key difference. If we look at 2009 we see that admissions were slumping while birth rates were peaking. However, after some research we see that this was just after the financial crisis, meaning people were less likely to go to the NICU as they were worried about spending in uncertain times (Illinois, 2010). Also, hospitals would have been struggling with financing which would mean they would take less patients (Illinois, 2010).

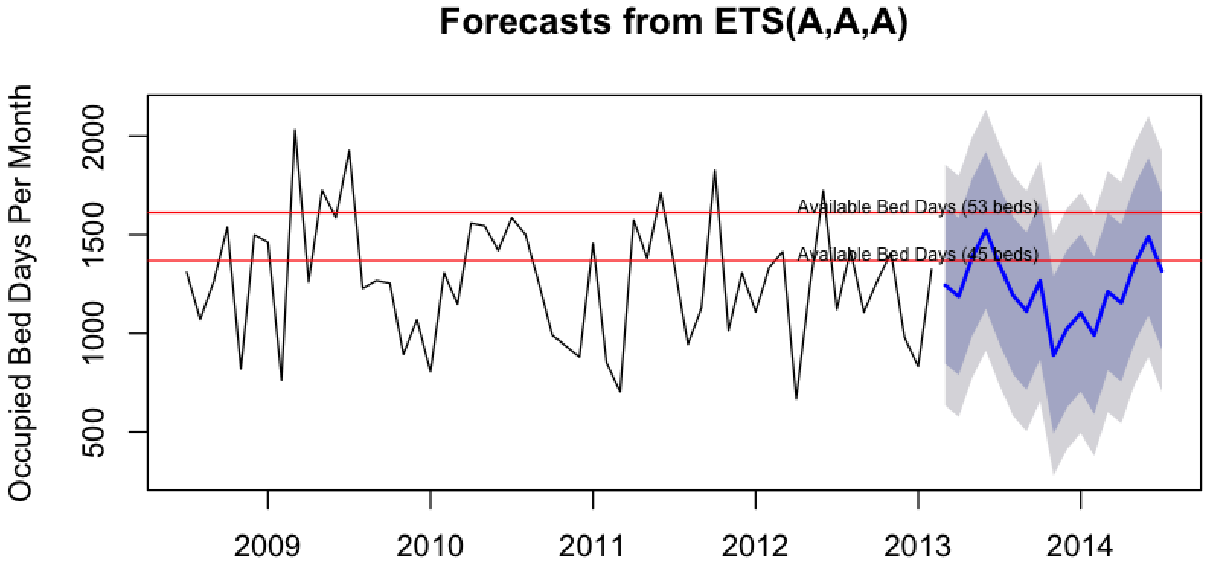






Data	Maximum	Minimum
US Births	August	February
ALOS	June	January
Admissions	March	November

Conclusions And Recommendations



Statistic	Beds Required
Mean	44
conf=80%	53
conf=95%	64

The above indicates that the NICU should go ahead with this expansion as it will likely be needed to meet demand. However, if we look at the US births data, we see a longer term downtrend in the number of births. This indicates that the NICU forecast may not be accounting for this longer term fall in US births (which is correlated with admissions and ALOS) thus, it may not be a good idea to continue with the expansion. We recommend that they pause the expansion to analyse the new data coming in and create new forecasts and conclusions.

Appendix

References

Nelson, B. (2017). *This is why so many babies are born in September*. [online] Business Insider. Available at: https://www.businessinsider.com/why-are-there-so-many-babies-born-in-september-2017-12?r=US&IR=T [Accessed 31 Jan. 2019].

TITA, A., HOLLIER, L. and WALLER, D. (2001). Seasonality in Conception of Births and Influence on Late Initiation of Prenatal Care. *Obstetrics & Gynecology*, 97(6), pp.976-981.

Melina, R. (2010). *In Which Month Are the Most Babies Born?*. [online] Live Science. Available at: https://www.livescience.com/32728-baby-month-is-almost-here-.html [Accessed 31 Jan. 2019].

Rettner, R. (2018). *US Birth Rate Hits All-Time Low: What's Behind the Decline?*. [online] Live Science. Available at: https://www.livescience.com/62592-birth-rate-declines-2017.html [Accessed 31 Jan. 2019].

Illinois, D. (2010). *Hospitals in Distress: How the Economy has Affected Financing of Health Care – Illinois Business Law Journal*. [online] Publish.illinois.edu. Available at: https://publish.illinois.edu/illinoisblj/2010/03/16/hospitals-in-distress-how-the-economy-has-affected-financing-of-health-care/ [Accessed 31 Jan. 2019].

Code Appendix

PART I - Obama - Clinton Case

1.1 Loading and Prepearing data:

```
elect.df  <- read.csv('Obama.csv')
summary(elect.df)

elect.df$ObamaMarginRate <-  (elect.df$Obama - elect.df$Clinton) *100 / elect.df$TotalVote
elect.df$ObamaRate <- 100 * elect.df$Obama / elect.df$TotalVote
elect.df$ClintonRate <- 100 * elect.df$Clinton / elect.df$TotalVote

dim(elect.df)

summary(elect.df$ObamaMarginRate)
```

1.2 Data Descriptions

Correlogram Code:

```
# Libraries
library(ellipse)            # first run install.packages('ellipse') if needed
library(RColorBrewer)       # first run install.packages('RColorBrewer') if needed

data=cor(elect.df[,c(10:44)],use="complete.obs")

# Build a Pannel of 100 colors with Rcolor Brewer
my_colors <- brewer.pal(5, "Spectral")
my_colors=colorRampPalette(my_colors)(100)


plotcorr(data , col=my_colors[data*50+50], mar = c(0,0,0,0),cex.lab=0.75 , type = "upper" ,
diag=FALSE)
```

Correlation Matrix Code:

```
CorMtrix <-cor(dplyr::select_if(elect.df, is.numeric))
round(CorMtrix, 2)

cor1 <-cor(elect.df[,c(10:30)])
round(cor1, 2)
```

Correlation Plots Code:

```
pairs(~ Poverty+
        MedianIncome + MalesPer100Females+ ManfEmploy + DisabilitiesRate +
UnemployRate+ MedicareRate+ RetiredWorkers,
    data = elect.df,
    pch=".")
```

1.3 Data Description: Region and Income analysis

```
library(datasets)
library(ggplot2)

fill <- "mediumorchid1"
line <- "mediumorchid2"

Plot1 <- ggplot(elect.df, aes(x = Region, y = ObamaMarginRate)) +
      geom_boxplot(fill = fill, colour = line, alpha = 0.7)
Plot1 <- Plot1 + scale_x_discrete(name = "Region") +
        scale_y_continuous(name = "Obama Margin Rate")
Plot1 <- Plot1 + ggtitle("Obama Margin Rate Distribution By Region")

Plot1 <- Plot1 + theme_bw()
Plot1


library(gridExtra)
elect.dfNE <- elect.df[elect.df$Region == "Northeast",]
elect.dfMW <- elect.df[elect.df$Region == "Midwest",]

Plot2 <- ggplot(elect.dfNE, aes(x = ObamaMarginRate, y = MedianIncome)) +
        geom_point() + geom_smooth(method=lm) + theme_minimal() + scale_color_gradient(low
= "#0091ff", high = "#f0650e")
Plot2 <- Plot2 + ggtitle("Northeast")

Plot3 <- ggplot(elect.dfMW, aes(x = ObamaMarginRate, y = MedianIncome)) +
        geom_point() + geom_smooth(method=lm) + theme_minimal() + scale_color_gradient(low
= "#0091ff", high = "#f0650e")
Plot3 <- Plot3 + ggtitle("Midwest")

grid.arrange(Plot2, Plot3, nrow = 1)
```

2.1 Data Prepearations and Cleaning

```
# Allows us to identify the NA values, of course there will be missing vote data, as some
counties are yet to vote.

countNAs <- function (v) sum(ifelse(is.na(v),1,0))

elect.countNAs <- sapply(elect.df, countNAs)

elect.countNAs[elect.countNAs != 0]

# Remove average income variable
elect.df$AverageIncome <- NULL

# 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 remove these records entirely.

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

2.2 Creating Train and Test Data

```
# Coverts date variable into an actual date format
elect.df$ElectionDate <- as.Date(elect.df$ElectionDate,
                        format="%m/%d/%Y")

#Creates Known and unknown dates for votes
elect.df.known <- elect.df[elect.df$ElectionDate <
                    as.Date("2/19/2008", format = "%m/%d/%Y"), ]

elect.df.unknown <- elect.df[elect.df$ElectionDate >=
                     as.Date("2/19/2008", format = "%m/%d/%Y"), ]



# Find the number of rows in the known dataset
nKnown <- nrow(elect.df.known)

# Set the seed for a random sample
set.seed(201)

# Randomly sample 75% of the row indices in the known dataset
rowIndicesTrain <- sample(1:nKnown,
                   size = round(nKnown*0.75),
                   replace = FALSE)

# Split the training set into the training set and the test set using these indices.

elect.df.training <- elect.df.known[rowIndicesTrain, ]

elect.df.test <- elect.df.known[-rowIndicesTrain, ]
```

Models

3.1 Linear Model

```
lmall <- lm(ObamaMarginRate ~ MalesPer100Females+AgeBelow35+Age35to65+Age65andAbove+

White+Black+Asian+AmericanIndian+Hawaiian+Hispanic+HighSchool+Bachelors+Poverty+IncomeAbove75K+

MedianIncome+AverageIncome+UnemployRate+ManfEmploy+SpeakingNonEnglish+Medicare+MedicareRate+

SocialSecurity+SocialSecurityRate+RetiredWorkers+Disabilities+DisabilitiesRate+Homeowner+
        SameHouse1995and2000+Pop+PopDensity+LandArea+FarmArea, data = elect.df.training)
summary(lmall)
lmall.pred <- predict(lmall, elect.df.test)

genError(lmall.pred, elect.df.test$ObamaMarginRate)



lm1 <- lm(ObamaMarginRate ~ MalesPer100Females+AgeBelow35+Age35to65+
          White+Black+Asian+AmericanIndian+
           Hawaiian+Hispanic+Poverty+
          MedianIncome+
           UnemployRate+
          MedicareRate+
           RetiredWorkers+
           DisabilitiesRate+Homeowner+
           Pop+PopDensity+LandArea,
        data = elect.df.training)

summary(lm1)

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$ObamaMarginRate)


lm.step.backward <- step(lmall, direction = "backward")
summary(lm.step.backward)
lm.step.backward.pred <- predict(lm.step.backward, elect.df.test)

genError(lm.step.backward.pred, elect.df.test$ObamaMarginRate)

lm.min <- lm(ObamaMarginRate ~ 1,
        data = elect.df.training)


lm.step.forward <- step(lm.min,
                    direction='forward',
                    scope=ObamaMarginRate ~
MalesPer100Females+AgeBelow35+Age35to65+Age65andAbove+

White+Black+Asian+AmericanIndian+Hawaiian+Hispanic+HighSchool+Bachelors+Poverty+Inco
meAbove75K+

MedianIncome+UnemployRate+ManfEmploy+SpeakingNonEnglish+Medicare+MedicareRate+

SocialSecurity+SocialSecurityRate+RetiredWorkers+Disabilities+DisabilitiesRate+Homeowner+
        SameHouse1995and2000+Pop+PopDensity+LandArea+FarmArea)

summary(lm.step.forward)

lm.step.forward.pred <- predict(lm.step.forward, elect.df.test)

genError(lm.step.forward.pred, elect.df.test$ObamaMarginRate)
```

3.2 Regression Trees

```
install.packages("rpart.plot")
library(rpart)
library(rpart.plot)

rt <- rpart(ObamaMarginRate ~ MalesPer100Females+AgeBelow35+Age35to65+
         White+Black+Asian+AmericanIndian+ Hawaiian+Hispanic+Poverty+ MedianIncome+
UnemployRate+MedicareRate+RetiredWorkers+
        DisabilitiesRate+Homeowner+ Pop+PopDensity+LandArea,
        data = elect.df.training, cp = 0.001)  # Fits a regression tree.

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

genError(rt.pred, elect.df.test$ObamaMarginRate)

plotcp(rt,upper = "splits")

rt.opt <- prune(rt, cp=0.00855)
prp(rt.opt, type = 1, extra = 1)

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

genError(rt.opt.pred, elect.df.test$ObamaMarginRate)
```

3.3 Lasso Model

```
startCol <- which(names(elect.df)=="MalesPer100Females")
endCol <- which(names(elect.df)=="FarmArea")
xknown <- as.matrix(elect.df.known[, startCol:endCol])
yknown <- elect.df.known$ObamaRateMargin
library(glmnet)
lm.lasso <- glmnet(xknown, yknown, family = "gaussian")

plot(lm.lasso, xvar = "lambda", label = TRUE)
coef(lm.lasso, s = exp(0))

set.seed(101)

lm.lasso.cv <- cv.glmnet(xknown, yknown, nfolds = 5, family = "gaussian")

lm.lasso.cv$lambda.min
(minLogLambda <- log(lm.lasso.cv$lambda.min))

coef(lm.lasso.cv, s = "lambda.min")
# Coefficients of the regularized linear regression with an optimal lambda.

plot(lm.lasso, xvar = "lambda", label = TRUE)
abline(v = log(lm.lasso.cv$lambda.min))

xtest <- as.matrix(elect.df.test[, startCol:endCol])
```

```
lm.lasso.cv.pred <- predict(lm.lasso.cv, newx = xtest, s = "lambda.min")

genError(lm.lasso.cv.pred, elect.df.test$ObamaRateMargin)
```

Part II - US Births Case

4.1 Loading and preparing data set

```
USBirths.df <- read.csv("US_Births.csv")
head(USBirths.df)

summary(USBirths.df)

library(tidyr)
USBirths.df <- extract(USBirths.df, Yr_Mo, into = c("Year", "Month"), "(.{4})(.{2})")
head(USBirths.df)
```

4.2 Data Exploration (Time Series)

```
# Making the births variable a time series
LiveBirths.ts <- ts(USBirths.df$Live.Births,
          start = c(2007, 01),
          end = c(2012, 06),
          freq = 12)

plot(LiveBirths.ts, ylab = "Births in US", col = "mediumorchid1", main ="Time Series Of Births
(US)")
```

4.3 Modeling

AAA Model

```
.libPaths("/usr/local/lib/R/site-library")
library("forecast")

(USBirths.ets.AAA <- ets(LiveBirths.ts, model = "AAA"))
rmse.ets <- function (etsmodel) cat("RMSE = ", sqrt(etsmodel$mse))

rmse.ets(USBirths.ets.AAA)

plot(USBirths.ets.AAA)
```

AAN Model

```
(USBirths.ets.AAN <- ets(LiveBirths.ts, model = "AAN"))
rmse.ets(USBirths.ets.AAN)
plot(USBirths.ets.AAN)
```

Predictions Using AAA Model

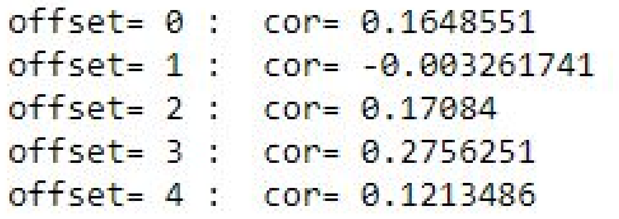
```
USBirths.ets.AAA.pred <- forecast(USBirths.ets.AAA, h = 9. Level = 80)
plot(USBirths.ets.AAA.pred)

USBirths.ets.AAA.pred$mean[8]
```

4.4 Comparing US Births to NICU data

```
us.df <- read.csv("US_Births.csv")

nicuA.df <- read.csv("NICU_A.csv")

us.df$Date <- as.Date(paste(as.character(us.df$Yr_Mo),"1",sep=""),
                         format="%Y%m%d")

nicuA.df$Date <- as.Date(paste(as.character(nicuA.df$Year), nicuA.df$Month,"1",sep=""),
                          format="%Y%b%d")
us.nicuA.df <- merge(us.df, nicuA.df, by="Date", all=TRUE)

library(repr)
options(repr.plot.width=9, repr.plot.height=4)  # change plot size to 7 x 5

## Admits vs US Births
plot(us.nicuA.df$Date, us.nicuA.df$Live.Births, type="l", col="red", xlab=NA, ylab=NA)
par(new = T)
plot(us.nicuA.df$Date, us.nicuA.df$Admits, type="l", col="blue", axes=F, xlab=NA, ylab=NA)
axis(side = 4)
mtext(side = 4, line = 0, 'Admits', cex=0.75)
mtext(side = 2, line = 2, 'Live.Births', cex=0.75)
legend("topright",
    legend=c("Live.Births", "","Admits"),
    col=c("red", "white","blue"),lty=c(1,1,1))

##Shifted plot
plot(us.nicuA.df$Date, us.nicuA.df$Live.Births, type="l", col="red", xlab=NA, ylab=NA)
par(new = T)
```

```
offset=3   # used to set the month offset for Admits - +3 means use the value from 3 months ago
plot(us.nicuA.df$Date, c(rep(NA,offset),head(us.nicuA.df$Admits, nrow(us.nicuA.df)-offset)),
type="l", col="blue", axes=F, xlab=NA, ylab=NA)
axis(side = 4)
mtext(side = 4, line = 0, 'Admits (3 months prior)', cex=0.75)
mtext(side = 2, line = 2, 'Live.Births', cex=0.75)
legend("topright",
      legend=c("Live.Births", "","Admits (3 months prior)"),
      col=c("red", "white","blue"),lty=c(1,1,1))
```

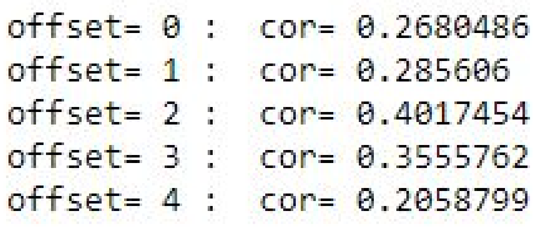
```
for (offset in 0:4)
   cat("offset=",offset, ":  cor=", cor(us.nicuA.df$Live.Births,
c(rep(NA,offset),head(us.nicuA.df$Admits, nrow(us.nicuA.df)-offset)), use="complete.obs"), "\n")
```



```
offset= 0 :  cor= 0.1648551
offset= 1 :  cor= -0.003261741
offset= 2 :  cor= 0.17084
offset= 3 :  cor= 0.2756251
offset= 4 :  cor= 0.1213486
```

```
## ALOS vs US Births
plot(us.nicuA.df$Date, us.nicuA.df$Live.Births, type="l", col="red", xlab=NA, ylab=NA)
par(new = T)
plot(us.nicuA.df$Date, us.nicuA.df$ALOS, type="l", col="purple", axes=F, xlab=NA, ylab=NA)
axis(side = 4)
mtext(side = 4, line = 0, 'Admits', cex=0.75)
mtext(side = 2, line = 2, 'Live.Births', cex=0.75)
legend("topright",
      legend=c("Live.Births", "","ALOS"),
      col=c("red", "white","purple"),lty=c(1,1,1))

plot(us.nicuA.df$Date, us.nicuA.df$Live.Births, type="l", col="red", xlab=NA, ylab=NA)
par(new = T)
offset=2 # used to set the month offset for Admits - +3 means use the value from 3 months ago
plot(us.nicuA.df$Date, c(rep(NA,offset),head(us.nicuA.df$ALOS, nrow(us.nicuA.df)-offset)),
type="l", col="purple", axes=F, xlab=NA, ylab=NA)
axis(side = 4)
mtext(side = 4, line = 0, 'ALOS (2 months prior)', cex=0.75)
mtext(side = 2, line = 2, 'Live.Births', cex=0.75)
```

```
legend("topright",
       legend=c("Live.Births", "","ALOS (2 months prior)"),
       col=c("red", "white","purple"),lty=c(1,1,1))

for (offset in 0:4)
    cat("offset=",offset, ":  cor=", cor(us.nicuA.df$Live.Births,
c(rep(NA,offset),head(us.nicuA.df$ALOS, nrow(us.nicuA.df)-offset)), use="complete.obs"), "\n")
```



```
offset= 0 :  cor= 0.2680486
offset= 1 :  cor= 0.285606
offset= 2 :  cor= 0.4017454
offset= 3 :  cor= 0.3555762
offset= 4 :  cor= 0.2058799
```

```
## Seasonality Plots
par(mfrow=c(3,1))

Admits.ts <- ts(us.nicuA.df$Admits, start = c(2007, 1), freq = 12)

p1 <- ggseasonplot(Admits.ts)

births.ts <- ts(us.nicuA.df$Live.Births, start = c(2007, 1), freq = 12)

p2 <- ggseasonplot(births.ts)

ALOS.ts <- ts(us.nicuA.df$ALOS, start = c(2007, 1), freq = 12)

p3 <- ggseasonplot(ALOS.ts)

multiplot(p1, p2, p3, cols=2)

###Uses multiplot function used from R cookbook:
```

http://www.cookbook-r.com/Graphs/Multiple_graphs_on_one_page_(ggplot2)/