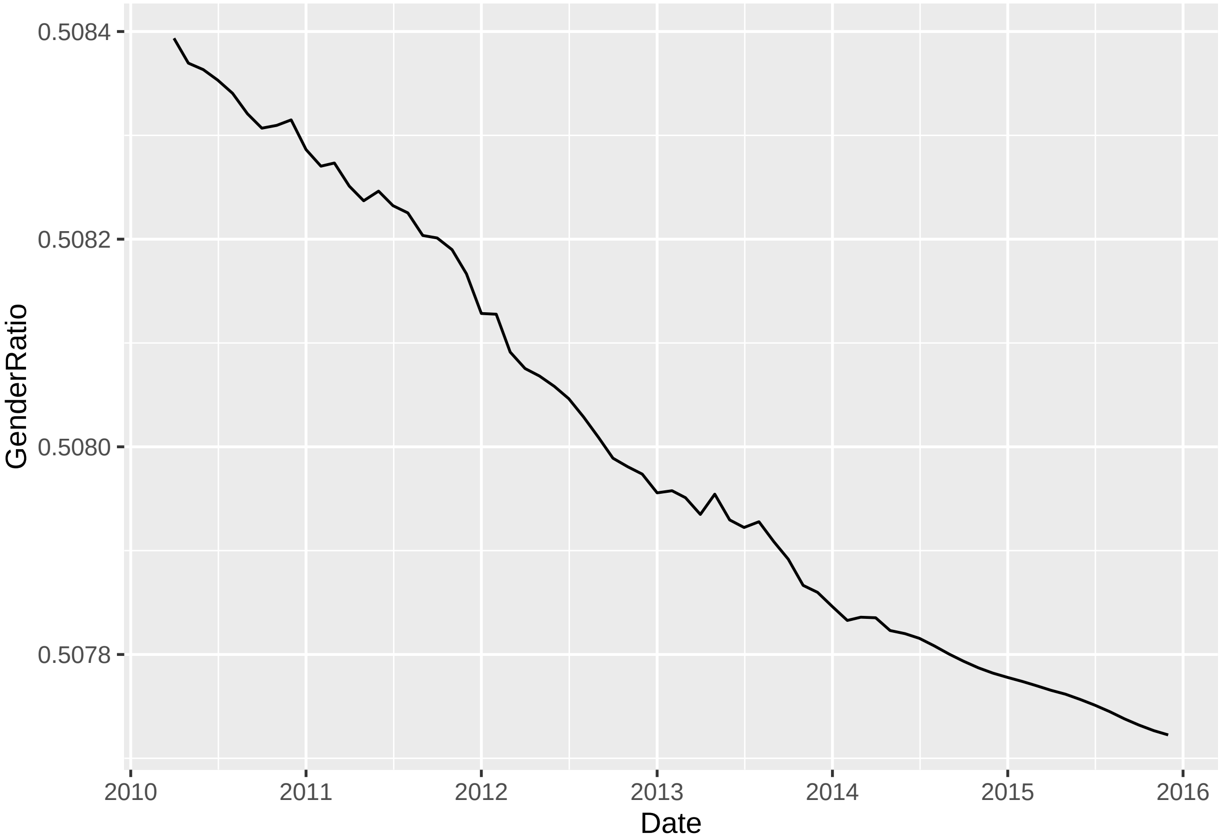
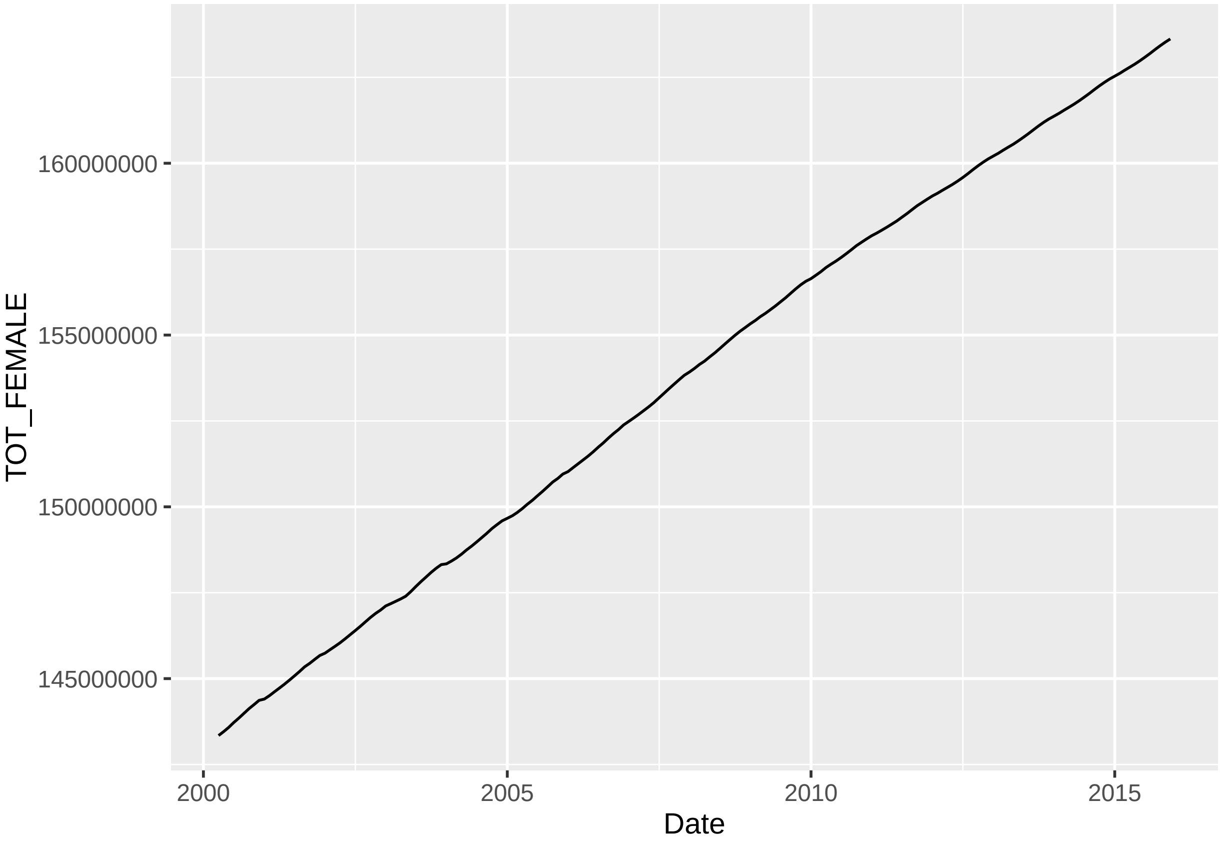
Natality Models Data Exploration

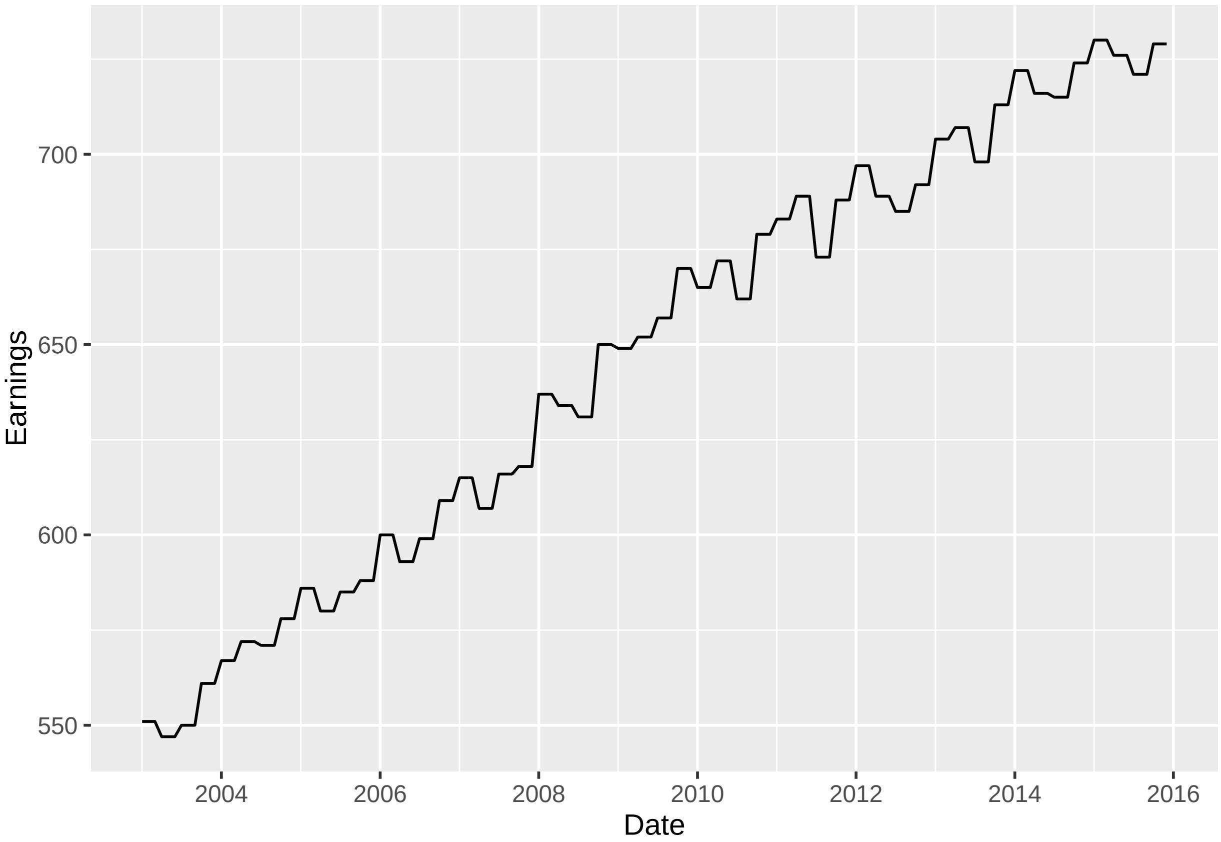
DATA 621: Business Analytics and Data Mining

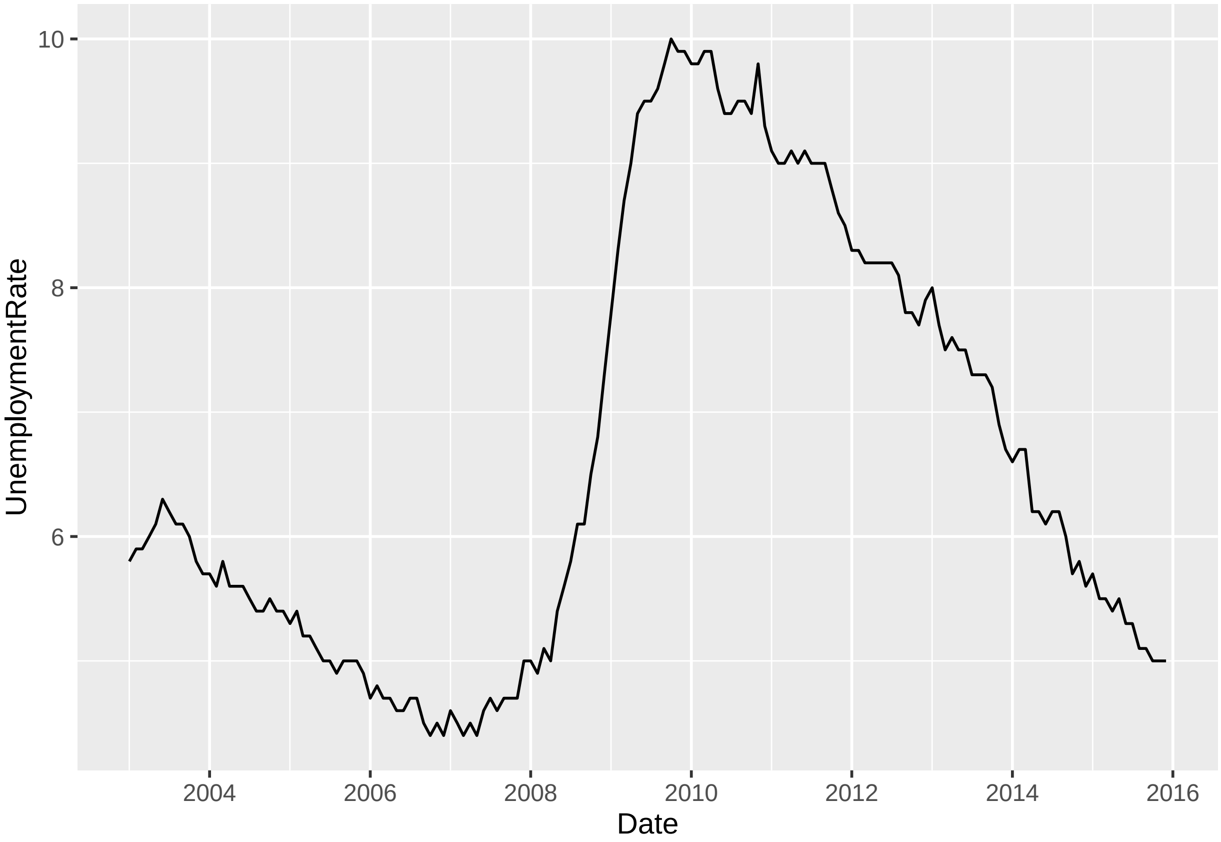
Daniel Dittenhafer & Justin Hink

April 24, 2016




TOT_FEMALE
160000000
155000000
150000000
145000000
2000
2005
2010
2015
Date


Earnings
700
650
600
550
2004
2006
2008
2010
2012
2014
2016
Date



```
##       Year          Month           Births
##  Min.   :2003   Min.   : 1.00   Min.   :291748
##  1st Qu.:2006   1st Qu.: 3.75   1st Qu.:327115
##  Median :2008   Median : 6.50   Median :342176
##  Mean   :2008   Mean   : 6.50   Mean   :341157
##  3rd Qu.:2011   3rd Qu.: 9.25   3rd Qu.:354900
##  Max.   :2014   Max.   :12.00   Max.   :390378
##       Date                        TOT_POP             GenderRatio
##  Min.   :2003-01-01 00:00:00   Min.   :288998781   Min.   :0.5078
##  1st Qu.:2005-12-24 06:00:00   1st Qu.:296931251   1st Qu.:0.5082
##  Median :2008-12-16 12:00:00   Median :305409338   Median :0.5084
##  Mean   :2008-12-15 17:00:00   Mean   :304885450   Mean   :0.5084
##  3rd Qu.:2011-12-08 18:00:00   3rd Qu.:312853653   3rd Qu.:0.5086
##  Max.   :2014-12-01 00:00:00   Max.   :319925152   Max.   :0.5090
##    TOT_FEMALE           TOT_MALE            Earnings     UnemploymentRate
##  Min.   :147114424   Min.   :141884357   Min.   :547.0   Min.   : 4.400
##  1st Qu.:151006749   1st Qu.:145924501   1st Qu.:591.8   1st Qu.: 5.175
##  Median :155272146   Median :150137192   Median :649.5   Median : 6.150
##  Mean   :154997130   Mean   :149888319   Mean   :640.5   Mean   : 6.757
##  3rd Qu.:158978789   3rd Qu.:153874864   3rd Qu.:688.2   3rd Qu.: 8.300
##  Max.   :162452248   Max.   :157472904   Max.   :724.0   Max.   :10.000
```

1 Data Exploration

The unified data set for this project contains 144 rows of data with 1 response variable and 9 predictor variables. An exploration of this data follows.

1.1 Missing Values

An analysis of missing values in the data set revealed 0 variables with incomplete data.

1.2 Correlations

The following table shows Pearson's *r* correlation coefficients between the numeric independent variables and the response variable *Births*.

Table 1: Pearson's r Correlation Coefficients

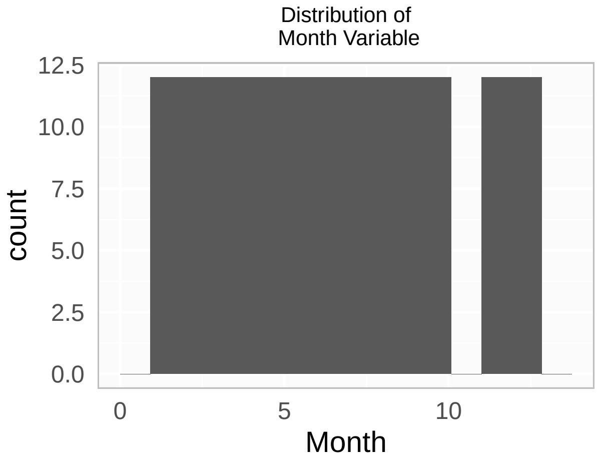
Births	1.0000000
Month	0.3646307
GenderRatio	0.2862173
TOT_MALE	-0.3214851
TOT_POP	-0.3219328
TOT_FEMALE	-0.3223760
Year	-0.3593053
Earnings	-0.3697992
UnemploymentRate	-0.3862666

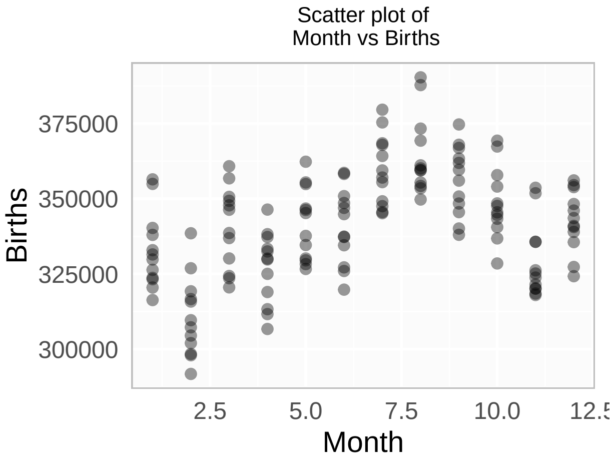
1.3 Variable Month

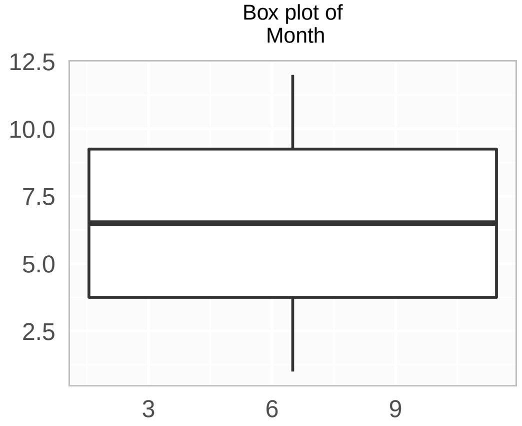
The *Month* variable is the month of birth. As one should expect, the distribution is uniform, but we can see some seasonality to the relationship between *Births* and *Month* with July and August being high frequency birth months.

Table 2: Month Variable Statistics

min	mean	stdev	median	max
1	6.5	3.464102	6.5	12





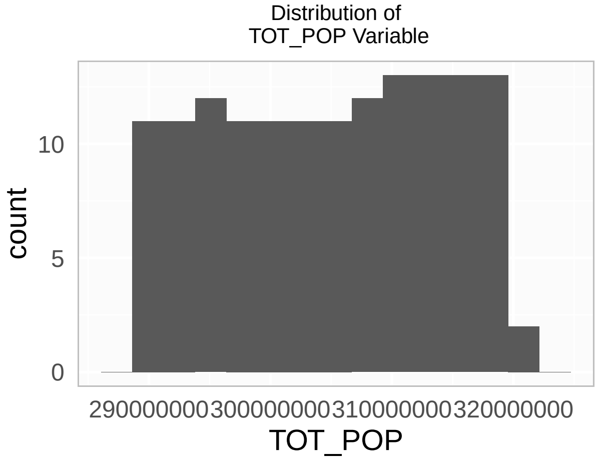


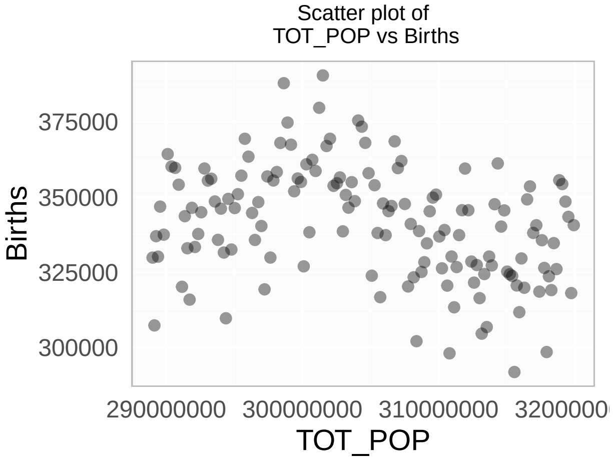
1.4 Variable TOT_POP

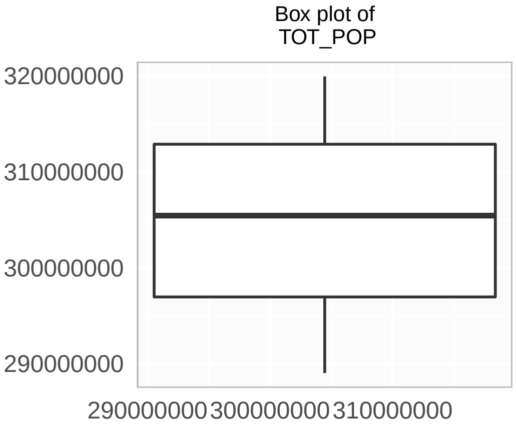
The *TOT_POP* variable is the total population per month as esimated by the Census Bureau.

Table 3: TOT_POP Variable Statistics

min	mean	stdev	median	max
288998781	304885450	9171506	305409338	319925152






Box plot of
TOT_POP
320000000
310000000
300000000
290000000
290000000
300000000
310000000