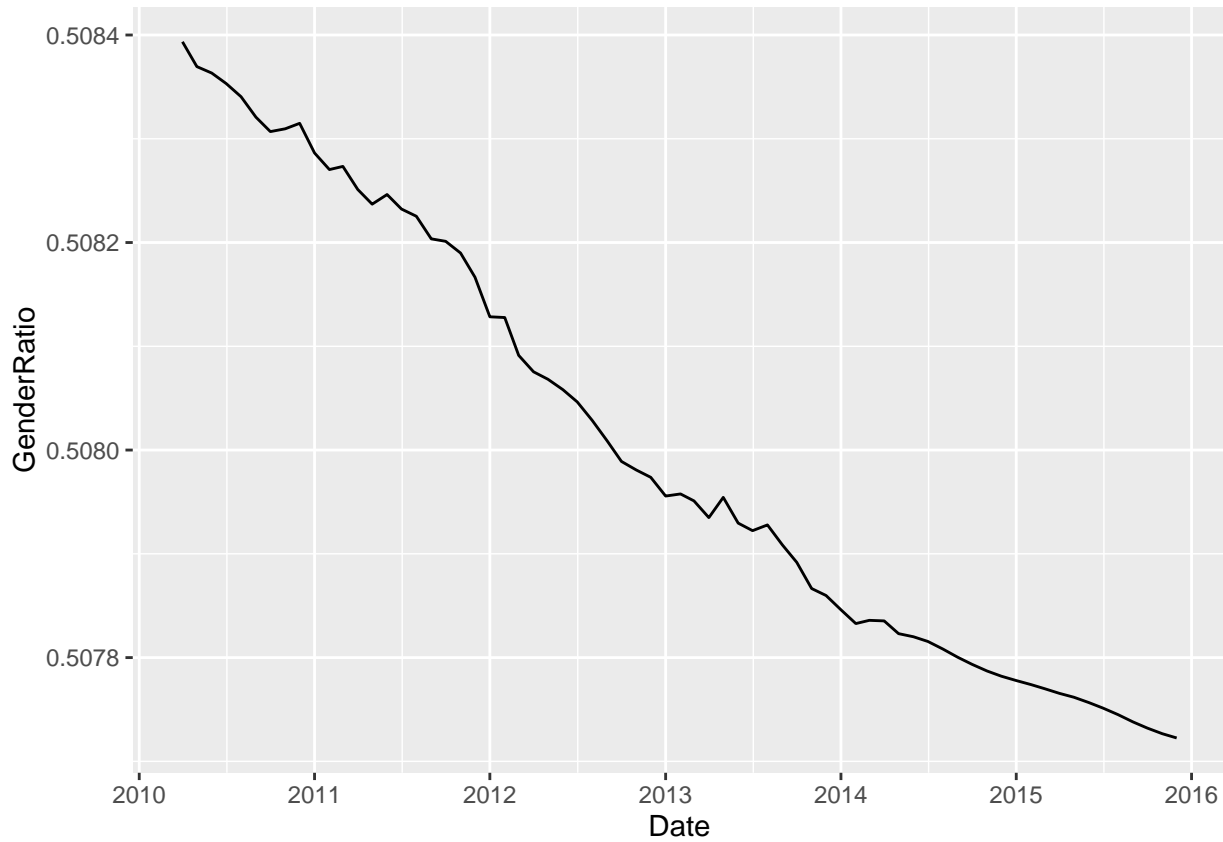


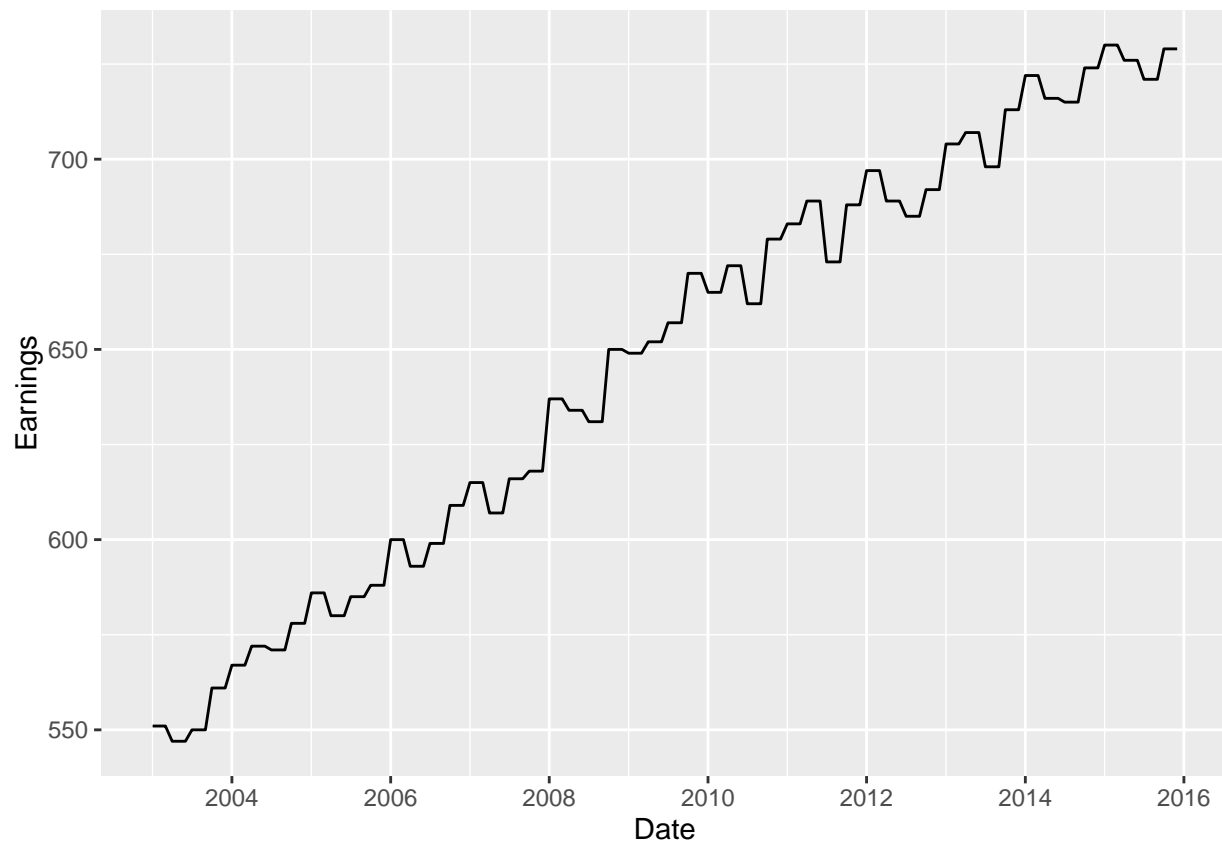
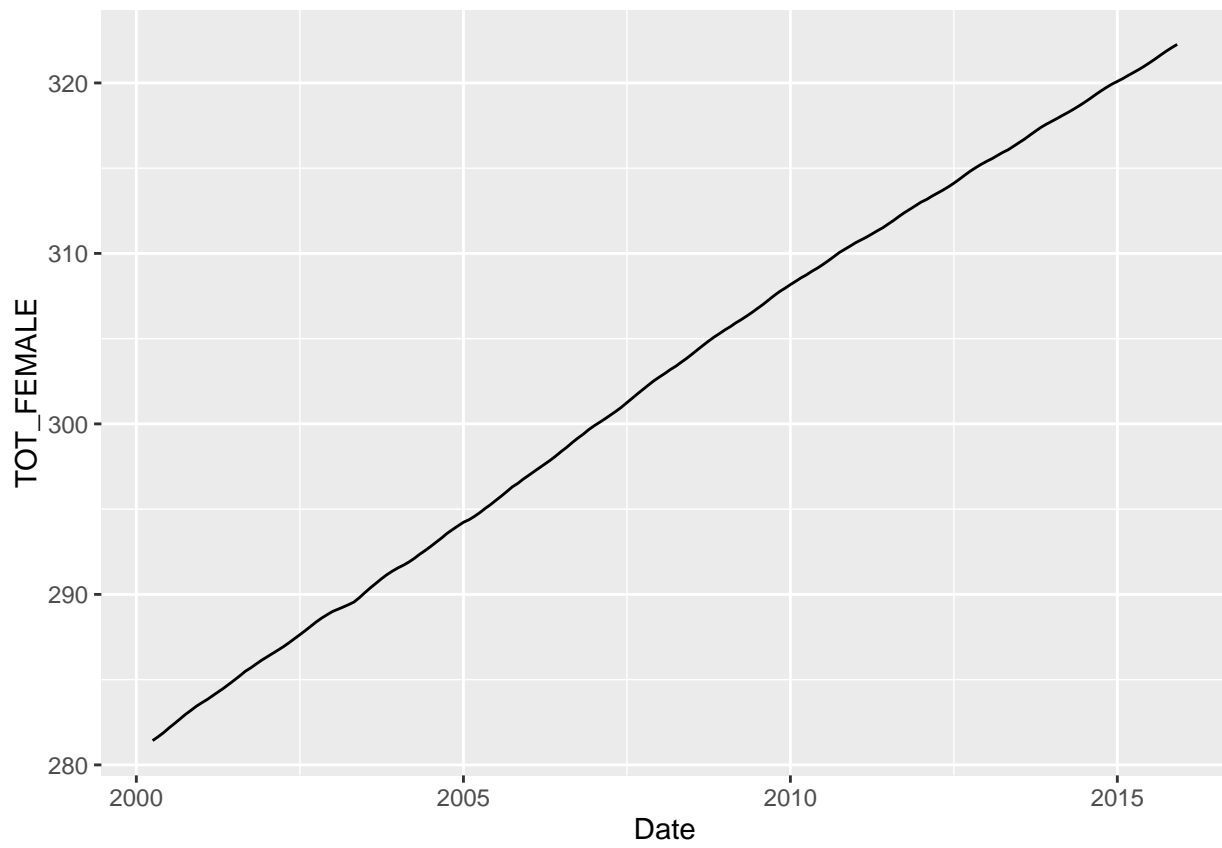
Natality Models Data Exploration

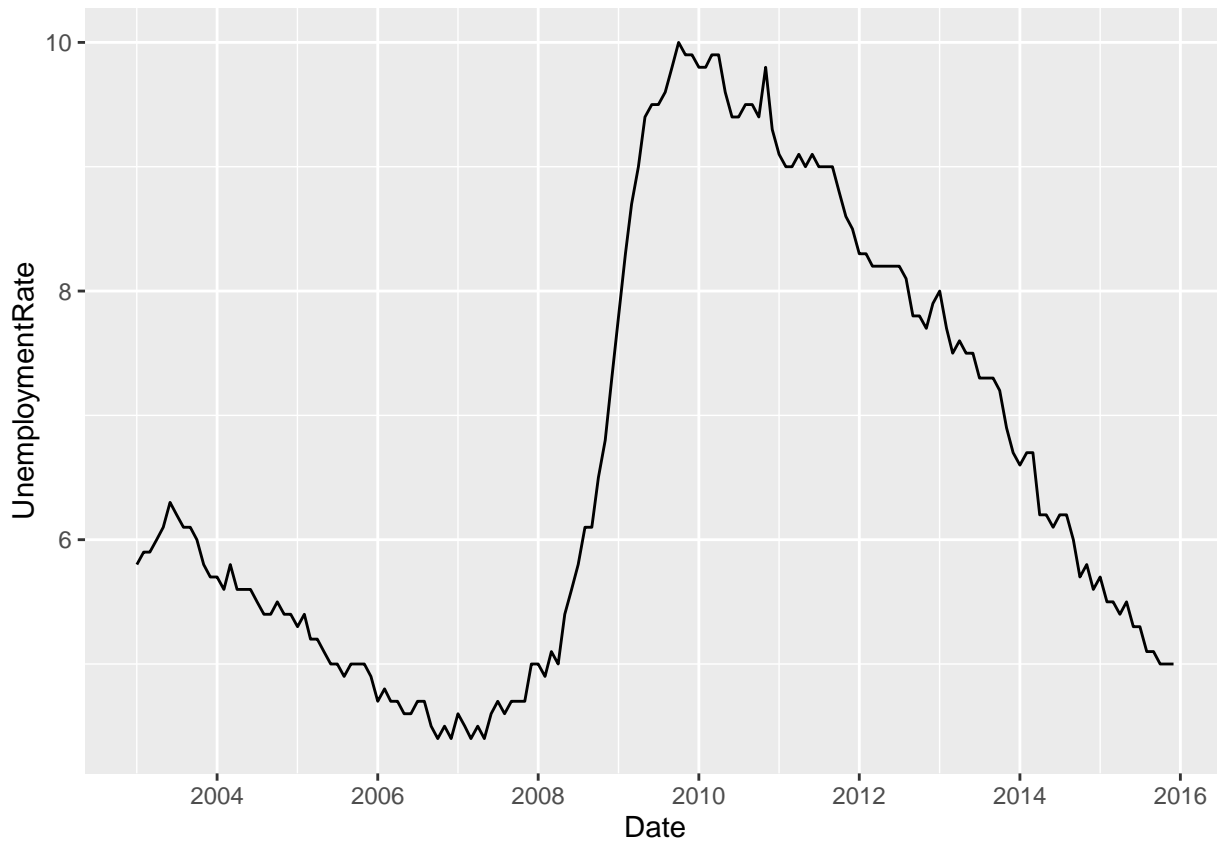
DATA 621: Business Analytics and Data Mining

Daniel Dittenhafer & Justin Hink

April 24, 2016







```
##      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.   :288999   Min.   :0.5078
## 1st Qu.:2005-12-24 06:00:00   1st Qu.:296931   1st Qu.:0.5082
## Median :2008-12-16 12:00:00   Median :305409   Median :0.5084
## Mean   :2008-12-15 17:00:00   Mean   :304885   Mean   :0.5084
## 3rd Qu.:2011-12-08 18:00:00   3rd Qu.:312854   3rd Qu.:0.5086
## Max.   :2014-12-01 00:00:00   Max.   :319925   Max.   :0.5090
##      TOT_FEMALE      TOT_MALE      Earnings      UnemploymentRate
## Min.   :289.0   Min.   :289.0   Min.   :547.0   Min.   : 4.400
## 1st Qu.:296.9   1st Qu.:296.9   1st Qu.:591.8   1st Qu.: 5.175
## Median :305.4   Median :305.4   Median :649.5   Median : 6.150
## Mean   :304.9   Mean   :304.9   Mean   :640.5   Mean   : 6.757
## 3rd Qu.:312.9   3rd Qu.:312.9   3rd Qu.:688.2   3rd Qu.: 8.300
## Max.   :319.9   Max.   :319.9   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

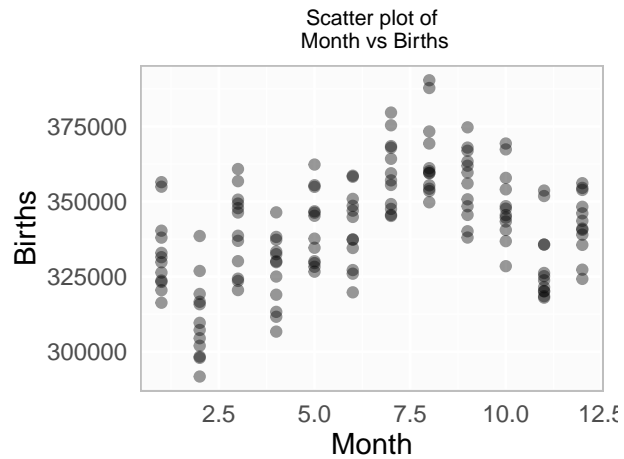
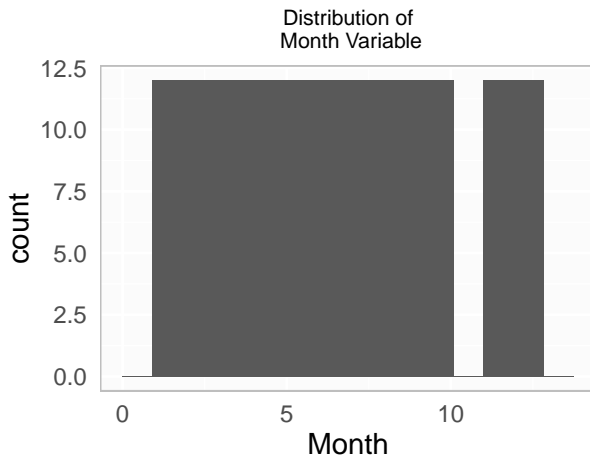
Births	1.0000000
Month	0.3646307
GenderRatio	0.2862173
TOT_POP	-0.3219328
TOT_FEMALE	-0.3219328
TOT_MALE	-0.3219328
Year	-0.3593053
Earnings	-0.3697992
UnemploymentRate	-0.3862666

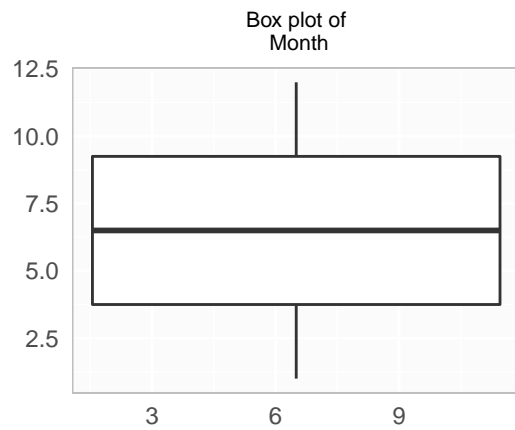
1.3 Variable Month

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



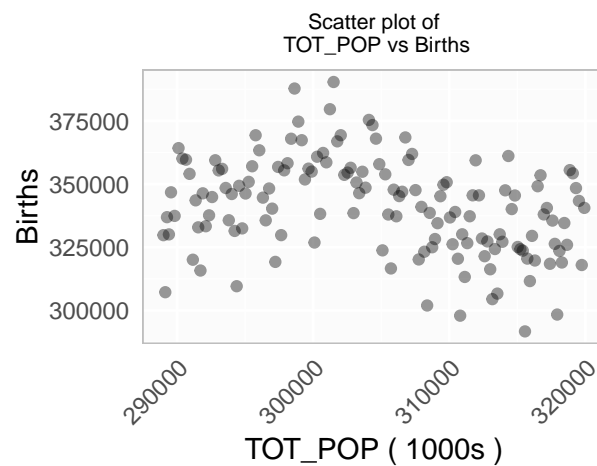
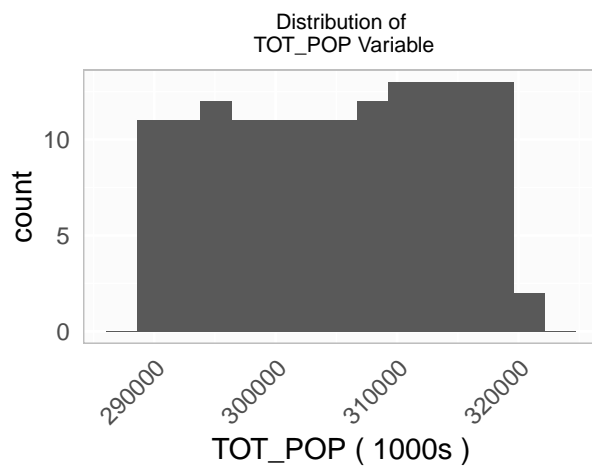


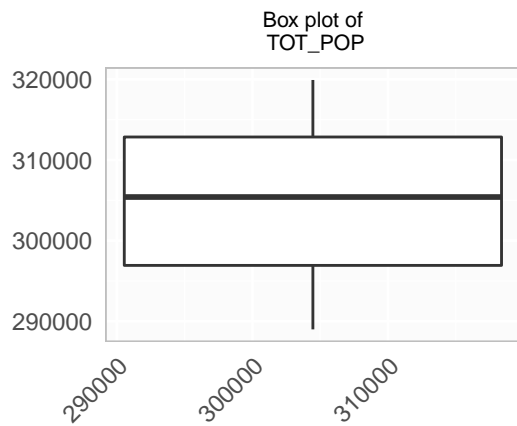
1.4 Variable TOT_POP

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
288998.8	304885.4	9171.506	305409.3	319925.2



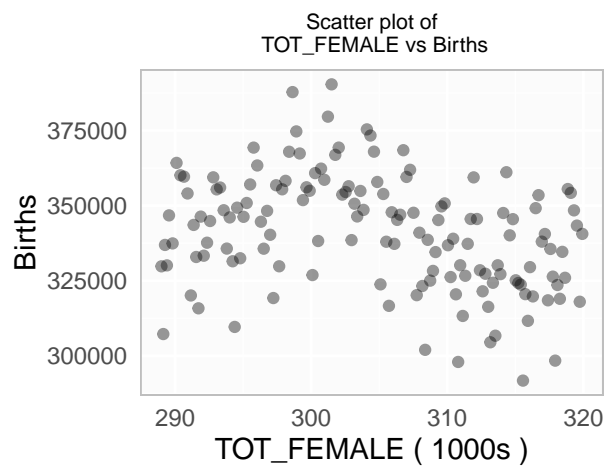
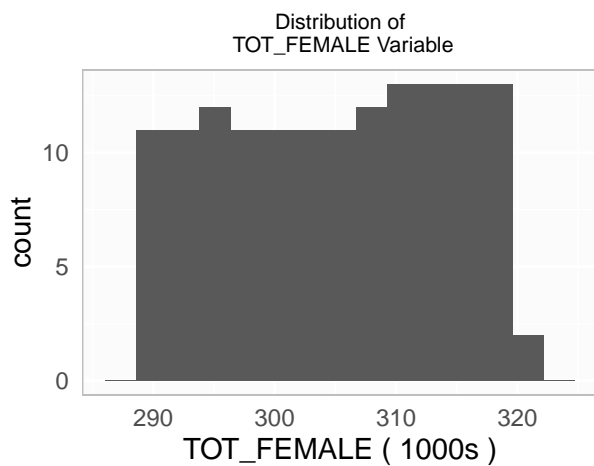


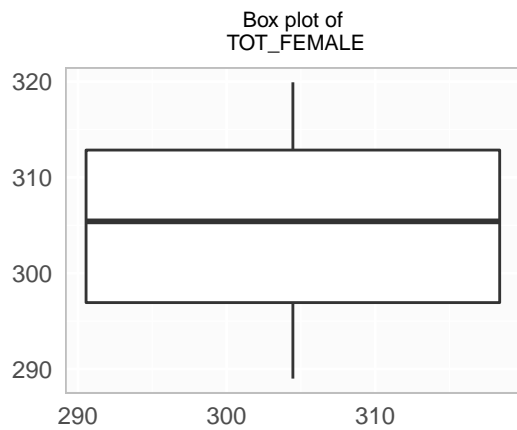
1.5 Variable TOT_FEMALE

The *TOT_FEMALE* variable is the total population of females per month as estimated by the Census Bureau.

Table 4: TOT_FEMALE Variable Statistics

min	mean	stdev	median	max
288.9988	304.8854	9.171506	305.4093	319.9252



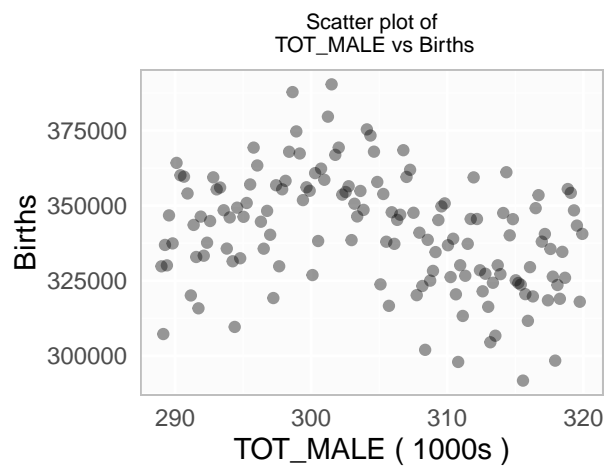
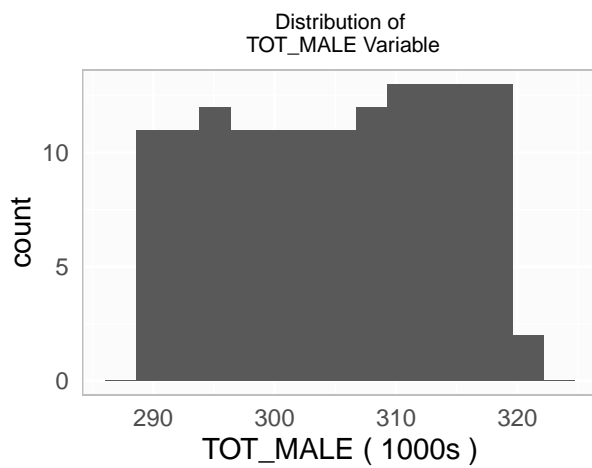


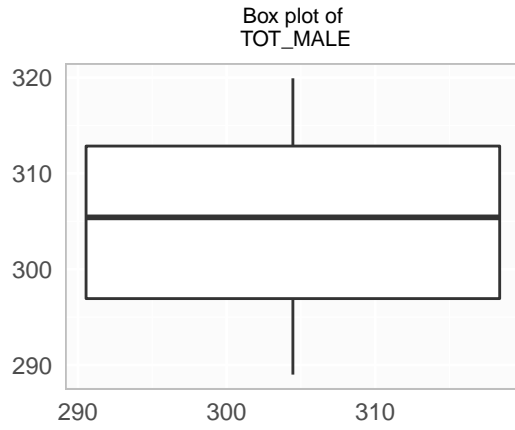
1.6 Variable TOT_MALE

The *TOT_MALE* variable is the total population of females per month as esimated by the Census Bureau.

Table 5: TOT_MALE Variable Statistics

min	mean	stdev	median	max
288.9988	304.8854	9.171506	305.4093	319.9252



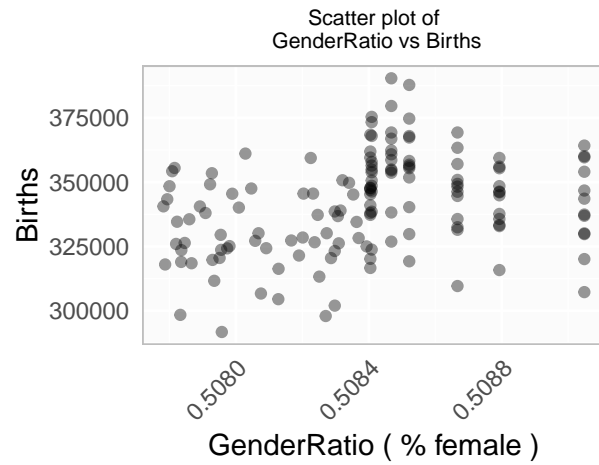
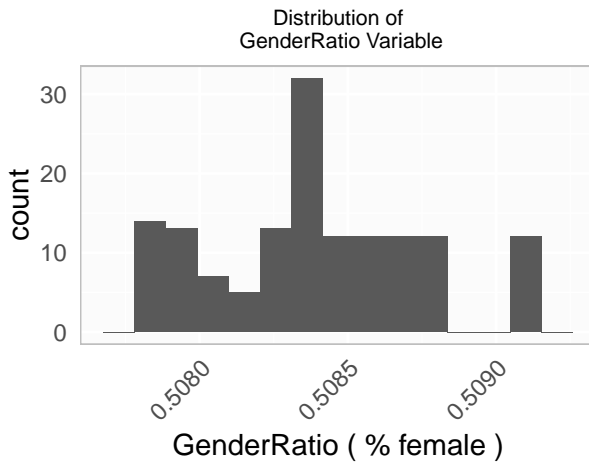


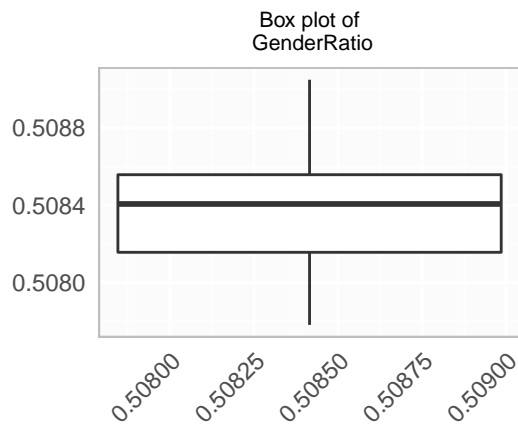
1.7 Variable GenderRatio

The *GenderRatio* variable is the percentage of the total population which are females per month derived from data from the Census Bureau. In cases where month data was not available, the annual gender ratio was computed and applied to the monthly total population.

Table 6: GenderRatio Variable Statistics

min	mean	stdev	median	max
0.507782	0.5083882	0.0003426	0.5084067	0.5090486



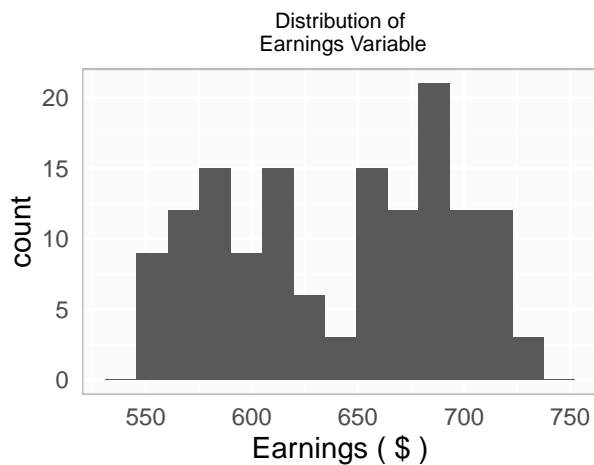


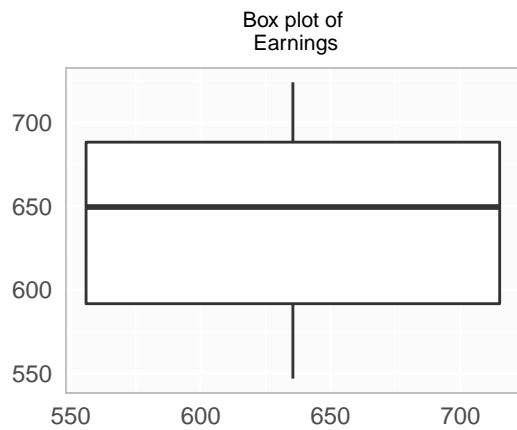
1.8 Variable Earnings

The *Earnings* variable is women's weekly earnings in current dollars based on data from the Bureau of Labor Statistics. The original values were provided quarterly and were expanded to a monthly format for data analysis purposes.

Table 7: Earnings Variable Statistics

min	mean	stdev	median	max
547	640.5417	53.55213	649.5	724





1.9 Variable UnemploymentRate

The *UnemploymentRate* variable is the unemployment rate per month (U3) based on data from the Bureau of Labor Statistics.

Table 8: UnemploymentRate Variable Statistics

min	mean	stdev	median	max
4.4	6.756944	1.789466	6.15	10

