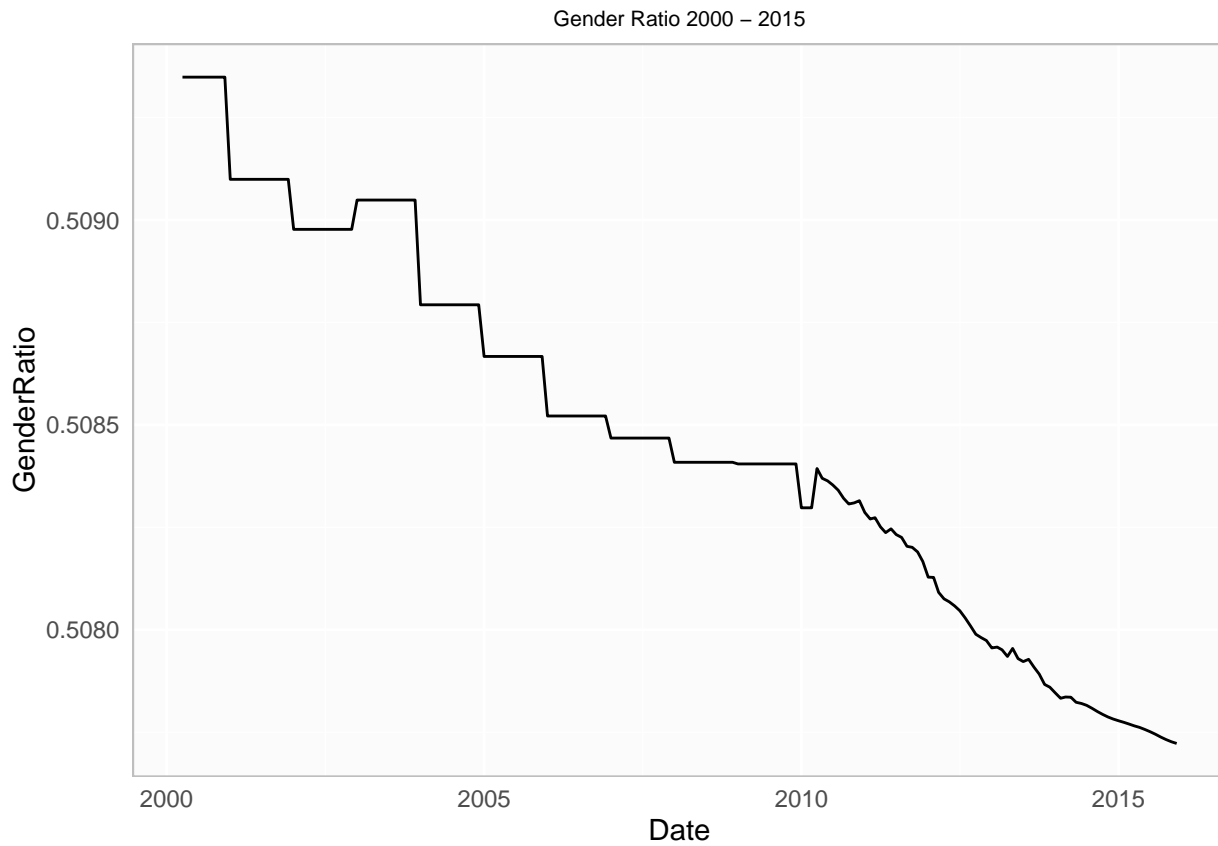


# Natality Models Data Exploration

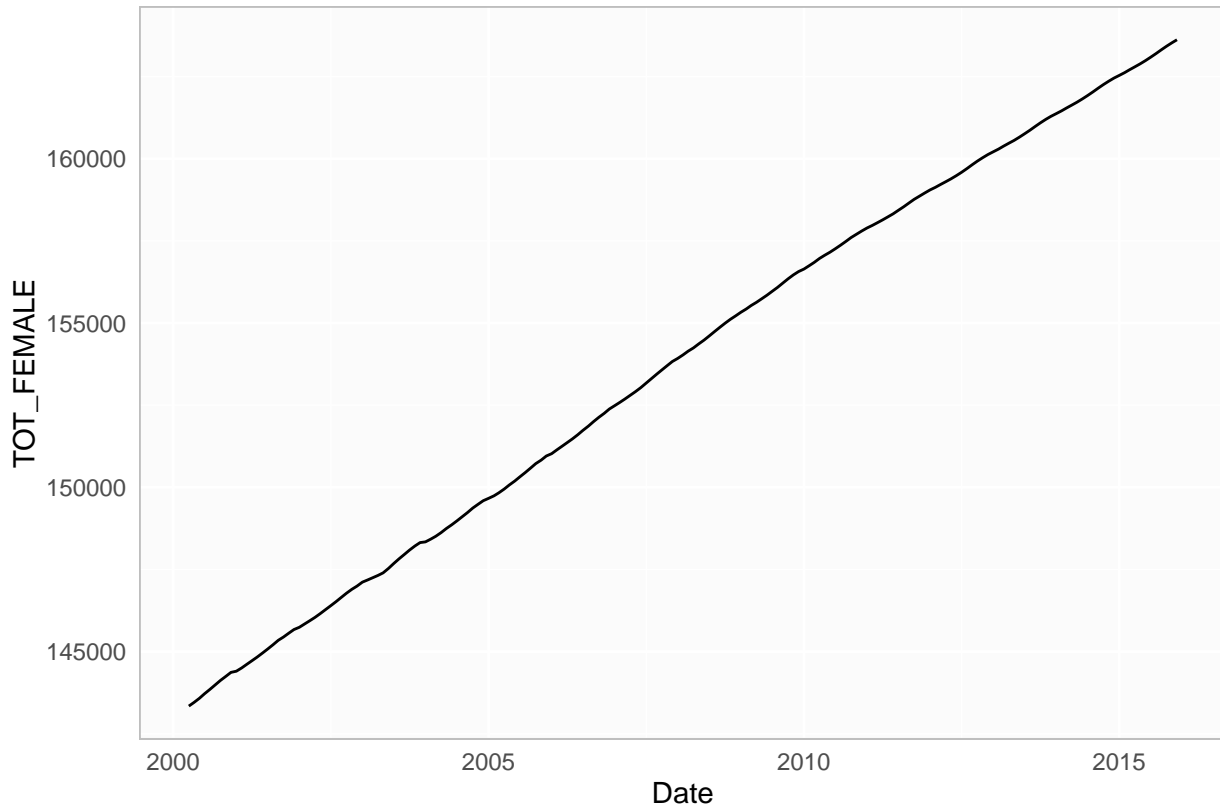
DATA 621: Business Analytics and Data Mining

*Daniel Dittenhafer & Justin Hink*

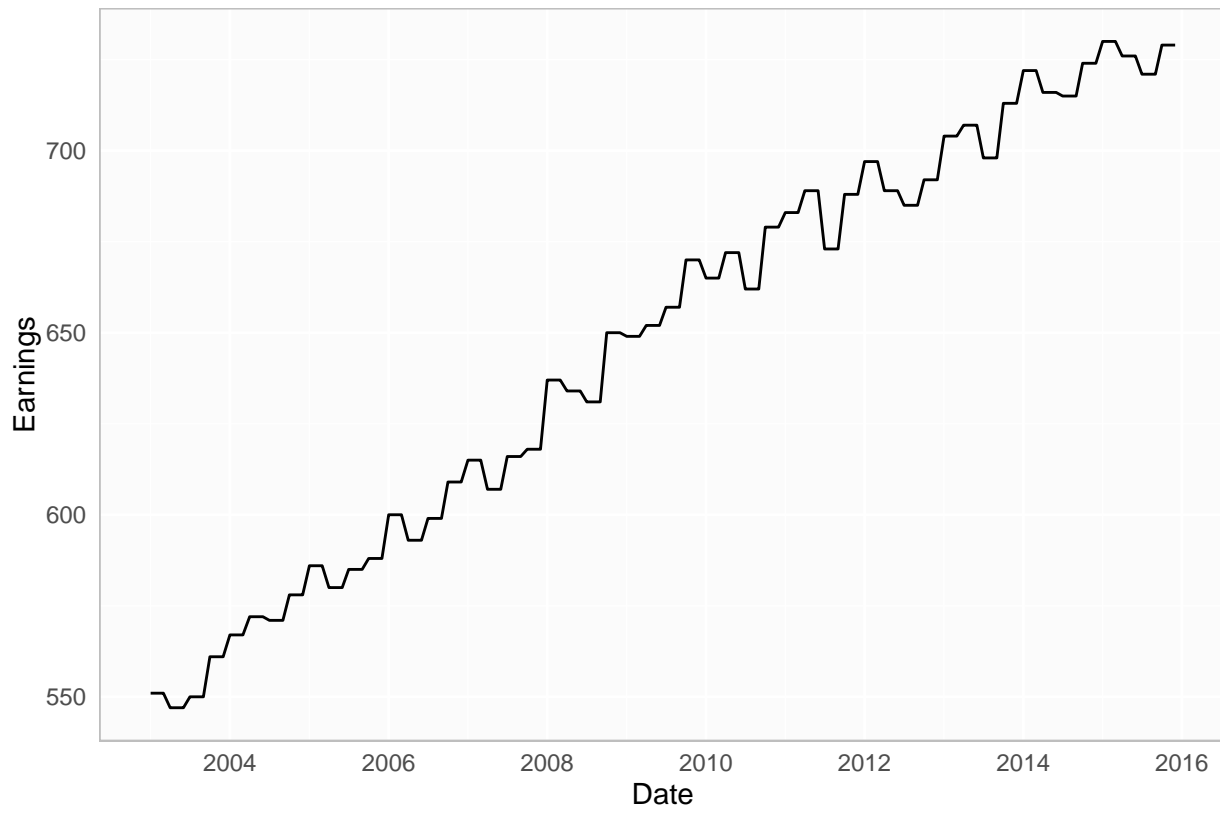
*April 24, 2016*

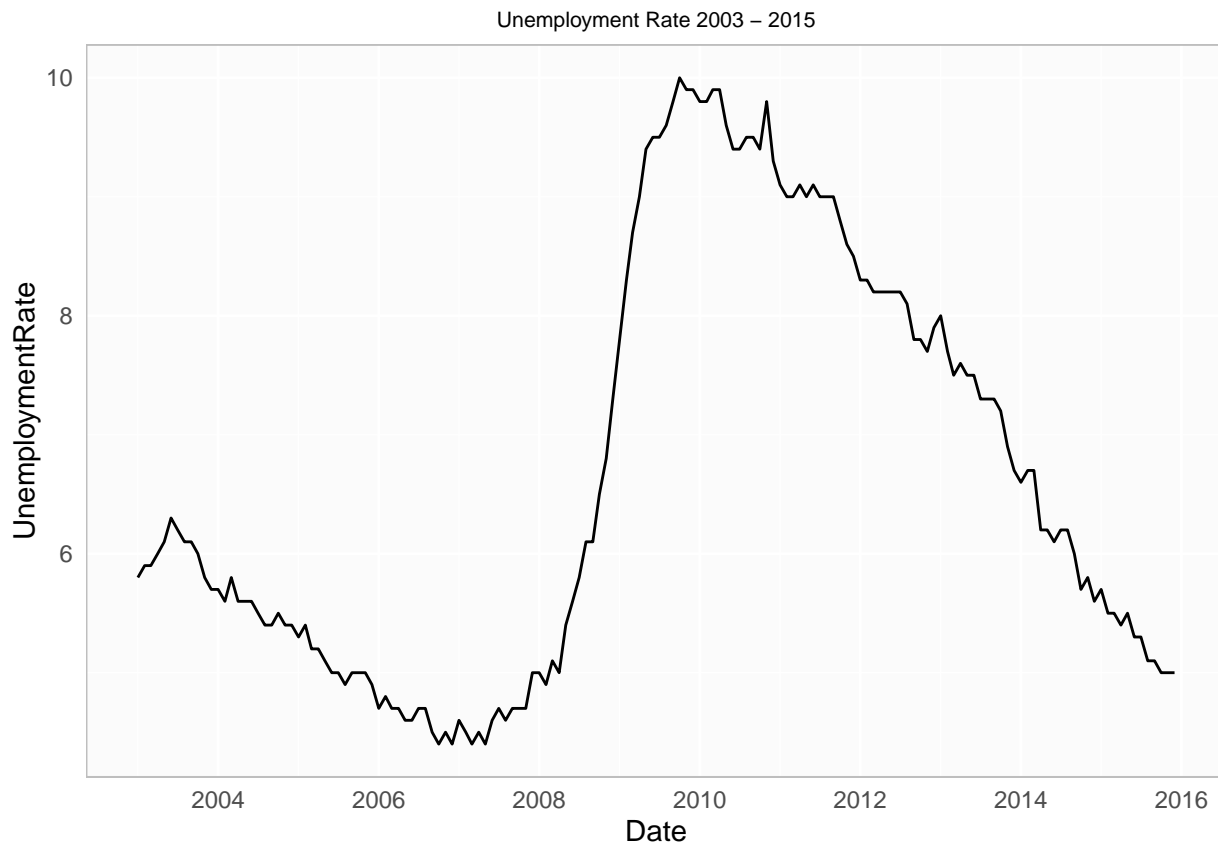


Female Population 2000 – 2015



Women's Weekly Earnings 2003 – 2015





```
##      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      FEMALE_15_24      FEMALE_25_34
## Min.   :147114   Min.    :141884   Min.    :20103   Min.    :19426
## 1st Qu.:151007   1st Qu.:145925   1st Qu.:20743   1st Qu.:19591
## Median :155272   Median :150137   Median :21201   Median :20142
## Mean   :154997   Mean    :149888   Mean    :21047   Mean    :20274
## 3rd Qu.:158979   3rd Qu.:153875   3rd Qu.:21414   3rd Qu.:20892
## Max.   :162452   Max.    :157473   Max.    :21489   Max.    :21646
##      FEMALE_35_44      Earnings      UnemploymentRate      Month9Ago
## Min.   :20353   Min.    :547.0   Min.    : 4.400   Min.    : 1.00
## 1st Qu.:20398   1st Qu.:591.8   1st Qu.: 5.175   1st Qu.: 3.75
## Median :21012   Median :649.5   Median : 6.150   Median : 6.50
## Mean   :21120   Mean    :640.5   Mean    : 6.757   Mean    : 6.50
## 3rd Qu.:21787   3rd Qu.:688.2   3rd Qu.: 8.300   3rd Qu.: 9.25
## Max.   :22303   Max.    :724.0   Max.    :10.000   Max.    :12.00
```

# 1 Data Exploration

The unified data set for this project contains 144 rows of data with 1 response variable and 13 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

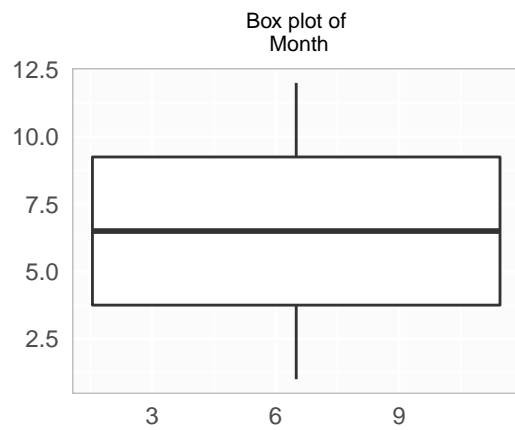
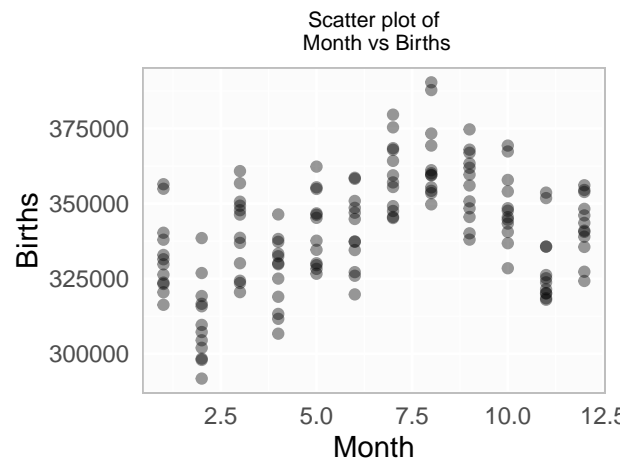
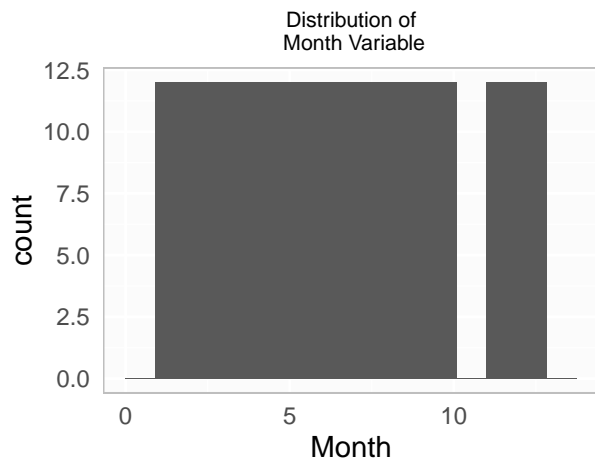
|                  |            |
|------------------|------------|
| Births           | 1.0000000  |
| FEMALE_35_44     | 0.3880661  |
| Month            | 0.3646307  |
| GenderRatio      | 0.2862173  |
| FEMALE_15_24     | -0.2307949 |
| TOT_MALE         | -0.3214851 |
| TOT_POP          | -0.3219328 |
| TOT_FEMALE       | -0.3223760 |
| Year             | -0.3593053 |
| Earnings         | -0.3697992 |
| UnemploymentRate | -0.3862666 |
| FEMALE_25_34     | -0.3879287 |

## 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  |

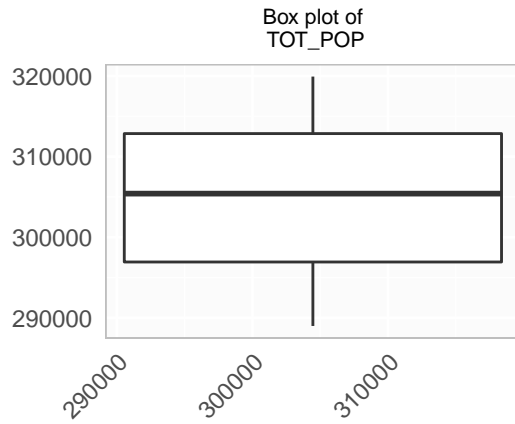
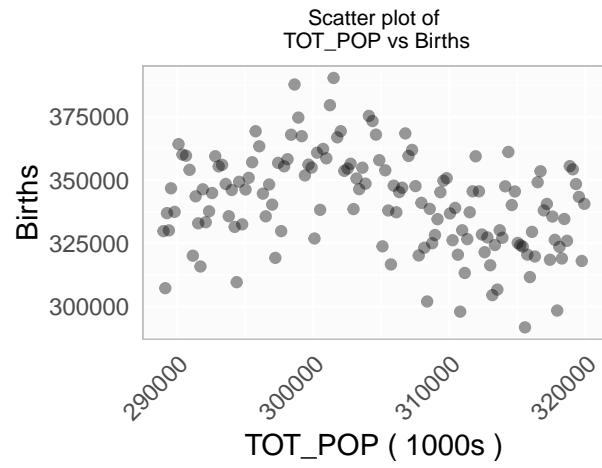
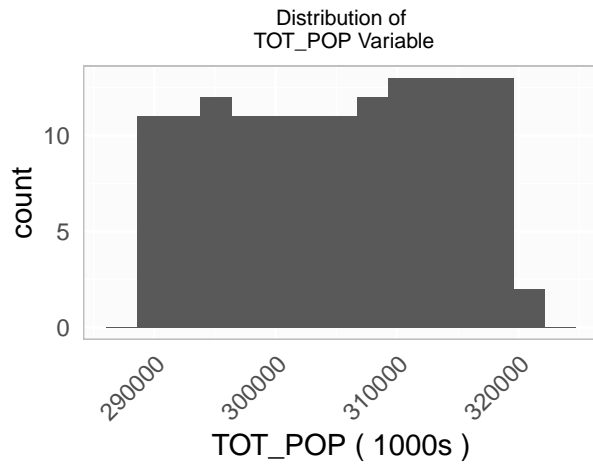


## 1.4 Variable TOT\_POP

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 |

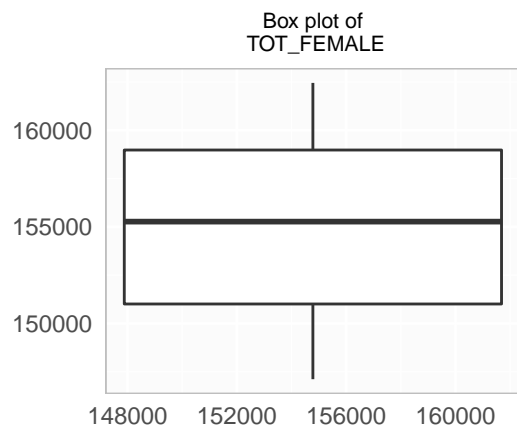
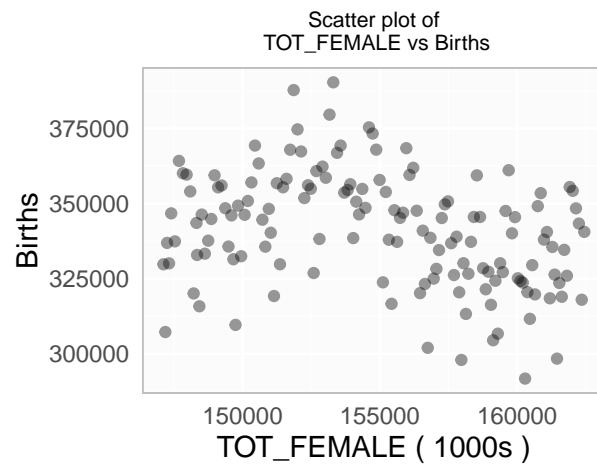
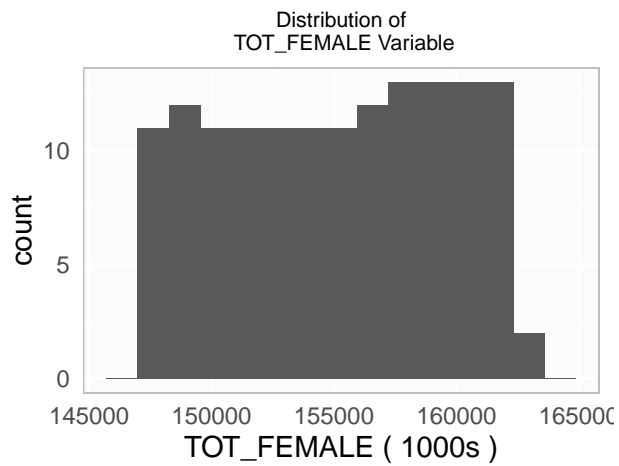


## 1.5 Variable TOT\_FEMALE

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      |
|----------|----------|----------|----------|----------|
| 147114.4 | 154997.1 | 4561.405 | 155272.1 | 162452.2 |

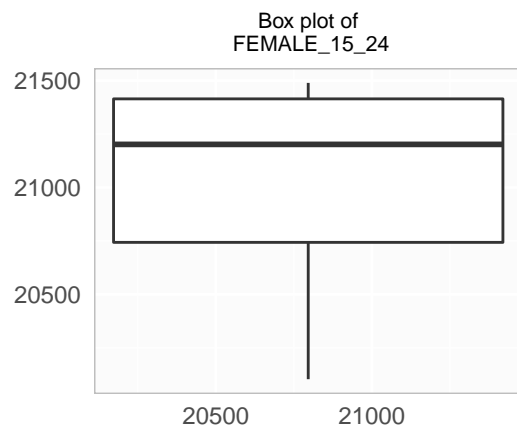
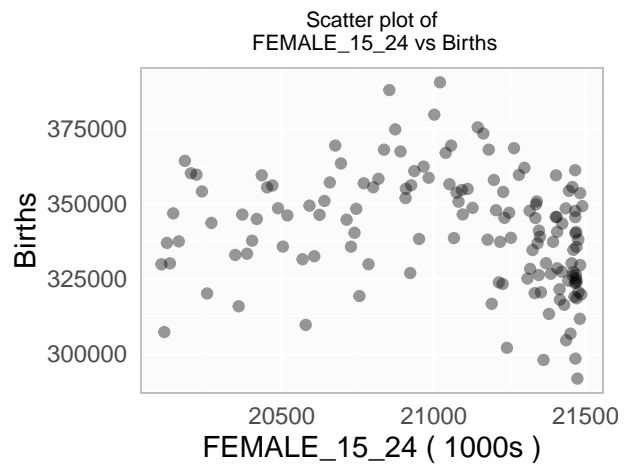
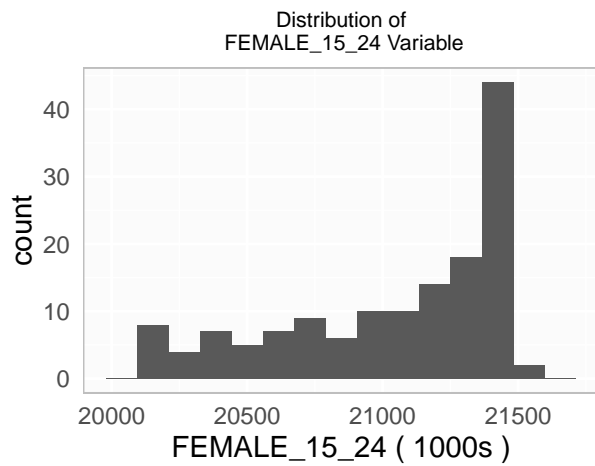


## 1.6 Variable FEMALE\_15\_24

The *FEMALE\_15\_24* variable is the total population of females ages 15-24 per month as estimated by the Census Bureau.

Table 5: FEMALE\_15\_24 Variable Statistics

| min      | mean    | stdev    | median   | max     |
|----------|---------|----------|----------|---------|
| 20103.14 | 21046.7 | 422.1778 | 21201.43 | 21489.1 |



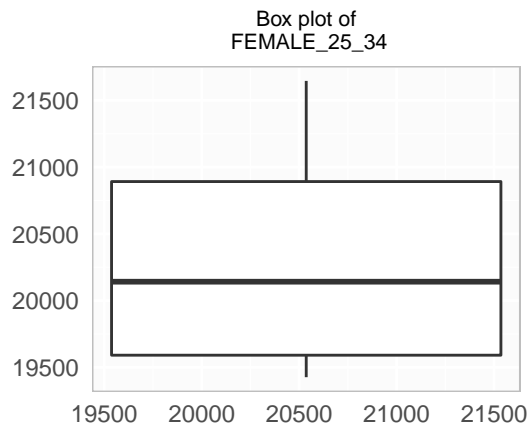
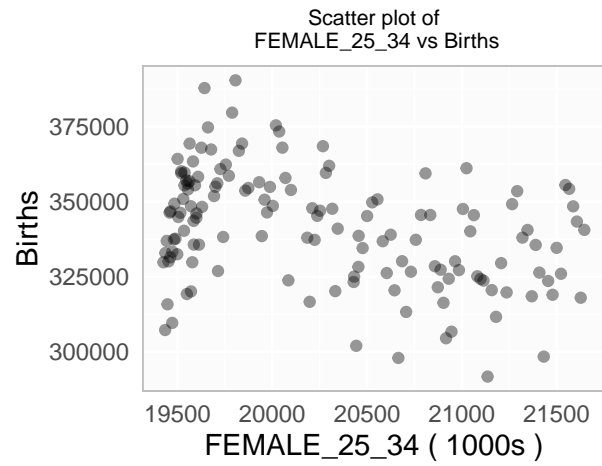
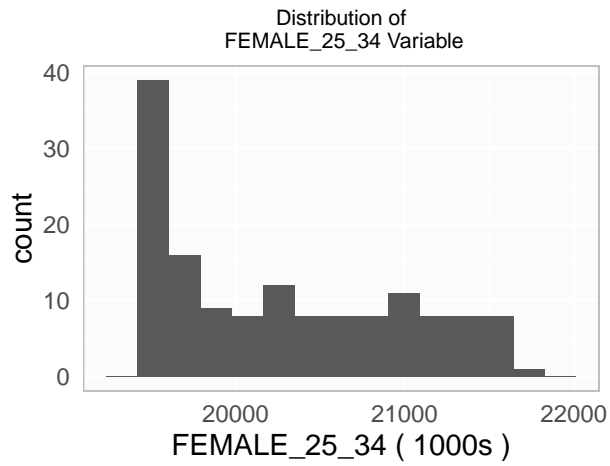
## 1.7 Variable FEMALE\_25\_34

The *FEMALE\_25\_34* variable is the total population of females ages 25-34 per month as estimated by the Census Bureau.

Table 6: FEMALE\_25\_34 Variable Statistics

| min      | mean     | stdev    | median   | max      |
|----------|----------|----------|----------|----------|
| 19426.37 | 20274.31 | 701.1676 | 20141.73 | 21646.13 |



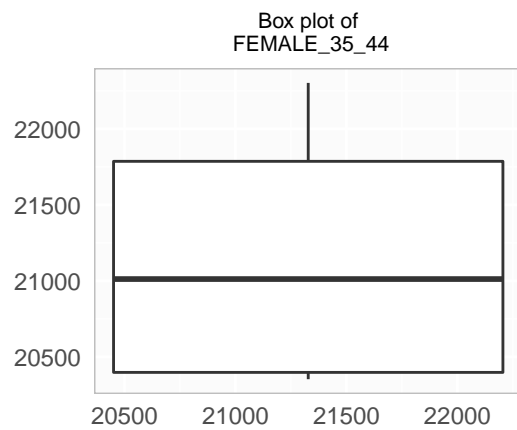
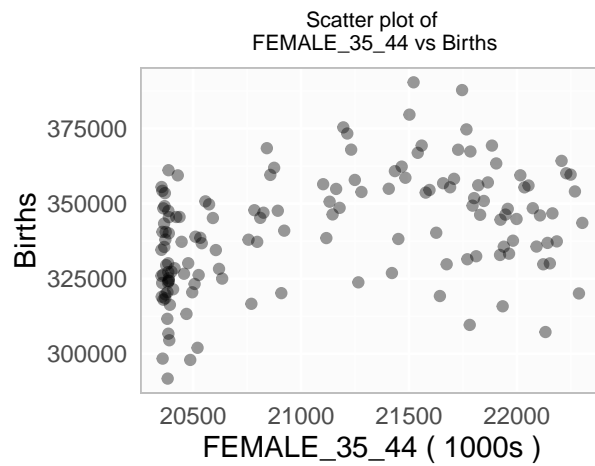
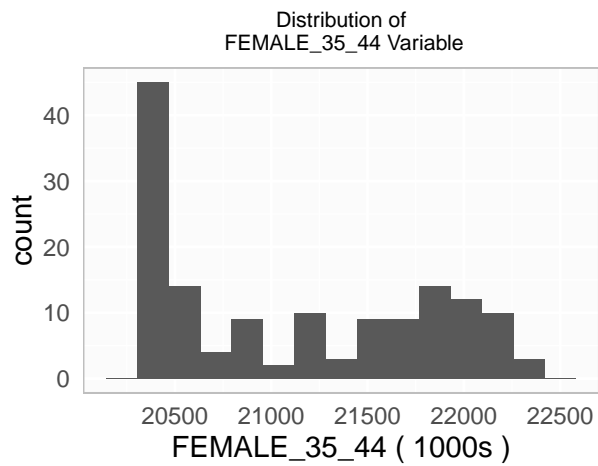


## 1.8 Variable FEMALE\_35\_44

The *FEMALE\_35\_44* variable is the total population of females ages 35-44 per month as estimated by the Census Bureau.

Table 7: FEMALE\_35\_44 Variable Statistics

| min      | mean     | stdev    | median   | max      |
|----------|----------|----------|----------|----------|
| 20353.37 | 21120.04 | 683.5963 | 21012.17 | 22302.87 |

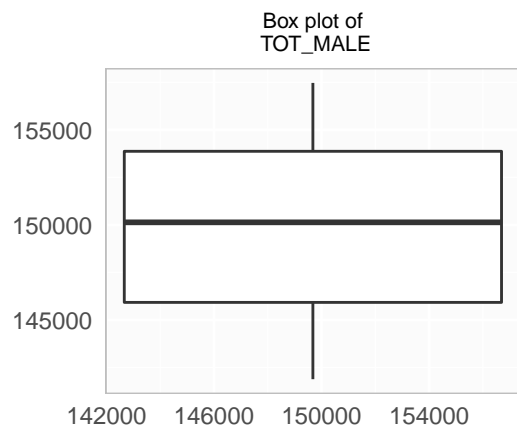
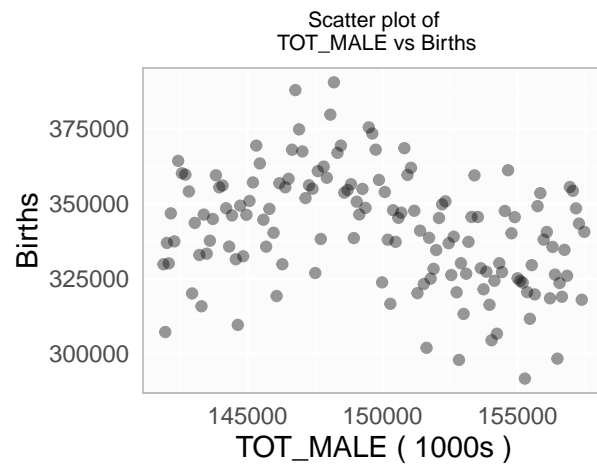
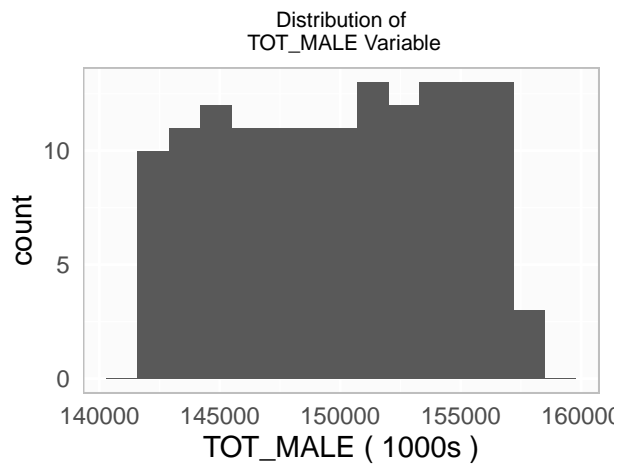


## 1.9 Variable TOT\_MALE

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

Table 8: TOT\_MALE Variable Statistics

| min      | mean     | stdev    | median   | max      |
|----------|----------|----------|----------|----------|
| 141884.4 | 149888.3 | 4610.232 | 150137.2 | 157472.9 |

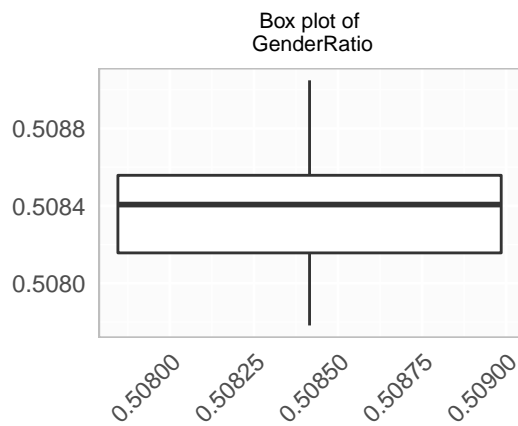
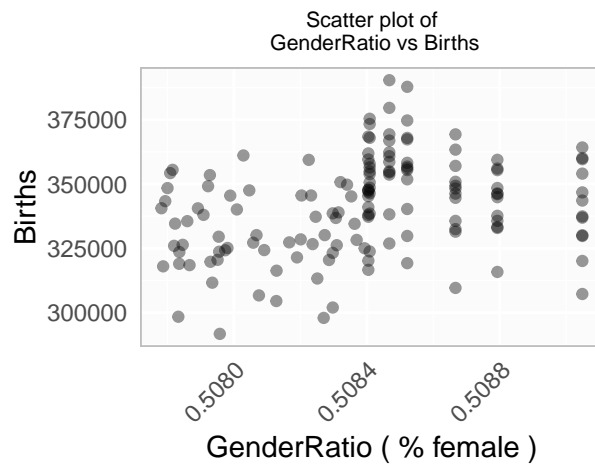
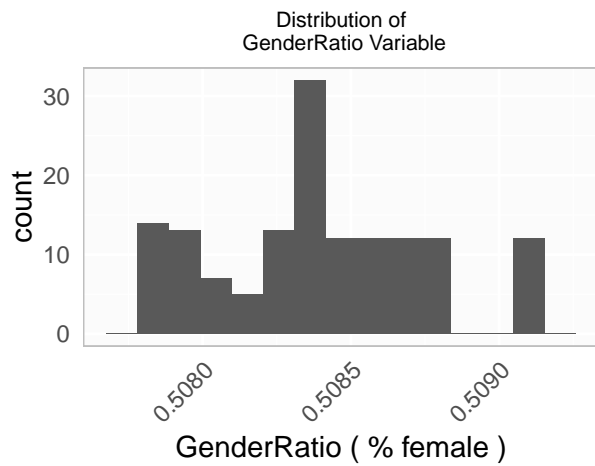


## 1.10 Variable GenderRatio

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 9: GenderRatio Variable Statistics

| min      | mean      | stdev     | median    | max       |
|----------|-----------|-----------|-----------|-----------|
| 0.507782 | 0.5083882 | 0.0003426 | 0.5084067 | 0.5090486 |

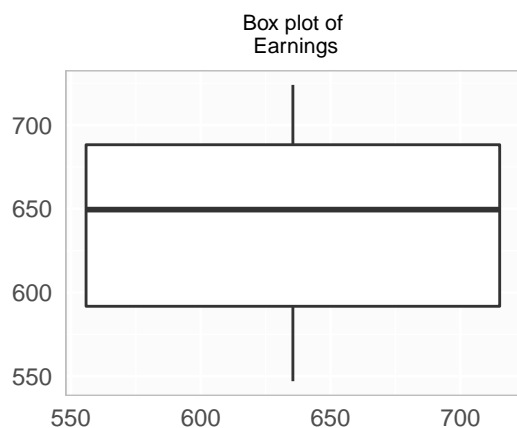
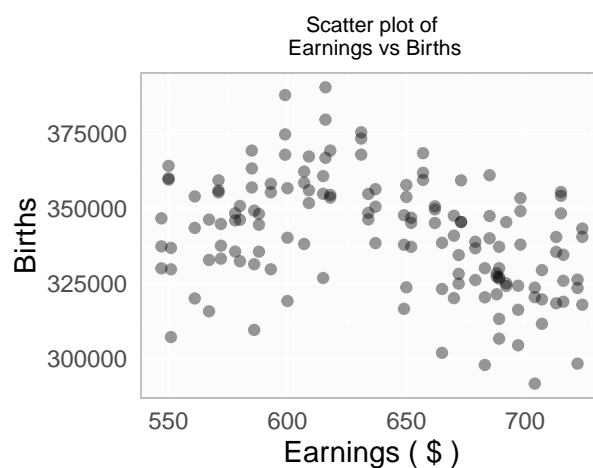
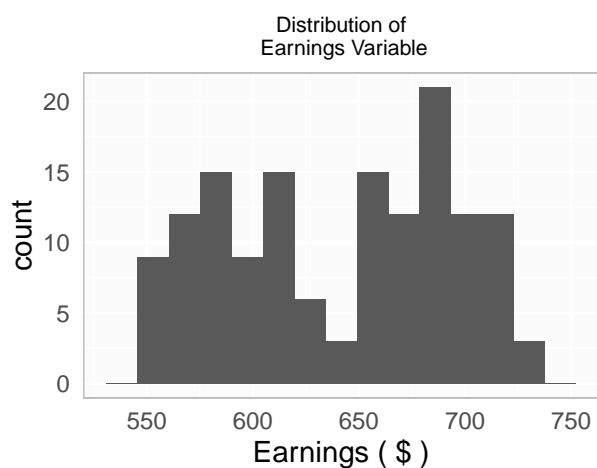


### 1.11 Variable Earnings

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 10: Earnings Variable Statistics

| min | mean     | stdev    | median | max |
|-----|----------|----------|--------|-----|
| 547 | 640.5417 | 53.55213 | 649.5  | 724 |

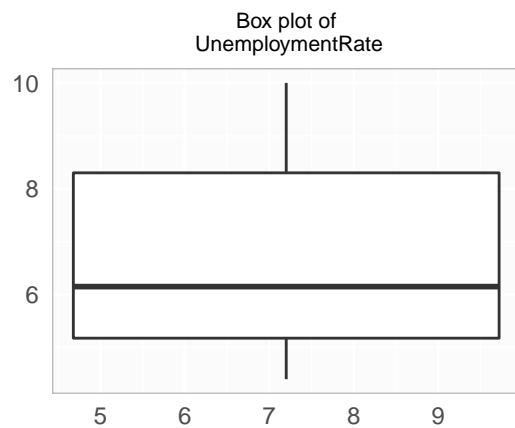
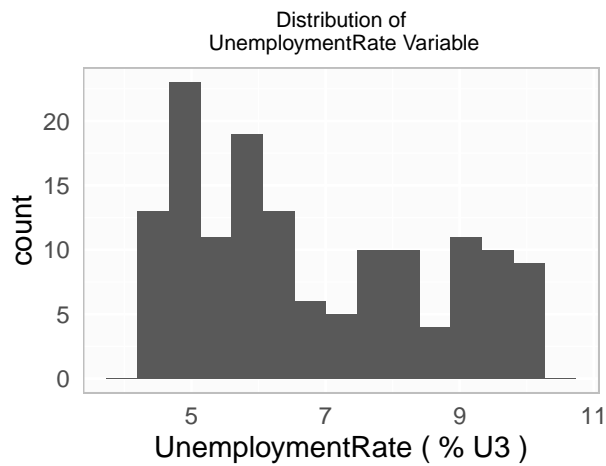


## 1.12 Variable UnemploymentRate

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

Table 11: UnemploymentRate Variable Statistics

| min | mean     | stdev    | median | max |
|-----|----------|----------|--------|-----|
| 4.4 | 6.756944 | 1.789466 | 6.15   | 10  |



## 2 Build Models

### 2.1 All Variables Linear Model

The first multiple linear regression model uses all 10 predictor variables. The adjusted  $R^2$  value for this model is 0.61129.

Table 12: All Variables Linear Model Coefficient Estimates

|                  | Estimate          | Pr(> t )  |
|------------------|-------------------|-----------|
| Intercept        | -107271160.660574 | 0.7809893 |
| Month *          | 2217.241706       | 0.0006441 |
| TOT_POP          | 311.447873        | 0.8070423 |
| GenderRatio      | 210610661.121019  | 0.7815692 |
| TOT_FEMALE       | -595.803571       | 0.8124195 |
| FEMALE_15_24     | -30.890272        | 0.5719171 |
| FEMALE_25_34     | -74.057899        | 0.0513369 |
| FEMALE_35_44     | 5.958233          | 0.7613287 |
| Earnings         | -84.575677        | 0.7558396 |
| UnemploymentRate | -2839.165366      | 0.1719684 |
| Month9Ago *      | 2600.816380       | 0.0000000 |

Table 13: All Variables Linear Model VIFs

|                  |                   |
|------------------|-------------------|
| Month            | 4.4144849         |
| TOT_POP          | 120650287.7957136 |
| GenderRatio      | 57975.4107139     |
| TOT_FEMALE       | 115783897.4451604 |
| TOT_MALE         | 55884.8603223     |
| FEMALE_15_24     | 217.5425337       |
| FEMALE_25_34     | 170.6400888       |
| FEMALE_35_44     | 30258.7816115     |
| Earnings         | 10734.8169565     |
| UnemploymentRate | 0.5494916         |
| Month9Ago        | 4.3240384         |

## 2.2 Significant Variables Linear Model

The second multiple linear regression model uses predictor variables indicated as significant from the All Variables model. The variables selected here were based on the All Variables model prior to the inclusion of the Month9Ago generated variable which appears to have affected the significant variables. The adjusted  $R^2$  value for this model is 0.47714.

Table 14: Significant Variables Linear Model Coefficient Estimates

|                | Estimate          | Pr(> t )  |
|----------------|-------------------|-----------|
| Intercept      | 514568208.68924   | 0.2046065 |
| TOT_POP        | -1765.33807       | 0.1855940 |
| GenderRatio    | -1018540784.31186 | 0.2017759 |
| TOT_FEMALE     | 3501.25182        | 0.1822808 |
| FEMALE_15_24   | -34.77073         | 0.5600065 |
| FEMALE_25_34   | -15.46851         | 0.6855427 |
| FEMALE_35_44 * | 47.00020          | 0.0000874 |
| Earnings *     | -1277.29819       | 0.0000000 |

Table 15: Significant Variables Linear Model VIFs

|              |                |
|--------------|----------------|
| TOT_POP      | 97350400.65649 |
| GenderRatio  | 47225.03676    |
| TOT_FEMALE   | 93360042.80857 |
| FEMALE_15_24 | 405.25307      |
| FEMALE_25_34 | 481.02800      |
| FEMALE_35_44 | 40.62177       |
| Earnings     | 88.00442       |

## 2.3 High Correlation Variables Linear Model

The third multiple linear regression model uses the six predictor variables with the highest correlation. The adjusted  $R^2$  value for this model is 0.47889.

|  | Estimate | Pr(> t ) |
|--|----------|----------|
|--|----------|----------|

Table 16: High Correlation Variables Linear Model Coefficient Estimates

|                  | Estimate       | Pr(> t )  |
|------------------|----------------|-----------|
| Intercept *      | -1962175.32283 | 0.0203610 |
| FEMALE_25_34 *   | -47.74490      | 0.0000000 |
| UnemploymentRate | 713.91408      | 0.6637059 |
| FEMALE_35_44     | 22.29315       | 0.2121244 |
| Earnings *       | -1182.65357    | 0.0000014 |
| Month            | 988.80549      | 0.0908859 |
| TOT_FEMALE *     | 22.88310       | 0.0000006 |

Table 17: High Correlation Variables Linear Model VIFs

|                  |            |
|------------------|------------|
| FEMALE_25_34     | 20.572017  |
| UnemploymentRate | 6.361599   |
| FEMALE_35_44     | 96.756161  |
| Earnings         | 100.430108 |
| Month            | 2.785238   |
| TOT_FEMALE       | 257.428151 |

## 2.4 Step Linear Model

The *step* function was used to produce the next multiple linear regression model. The adjusted  $R^2$  value for this model is 0.62001.

Table 18: Step Linear Model Coefficient Estimates

|                    | Estimate      | Pr(> t )  |
|--------------------|---------------|-----------|
| Intercept *        | 320506.481281 | 0.0000000 |
| Month *            | 2531.412763   | 0.0000000 |
| TOT_POP *          | 2.630363      | 0.0000005 |
| FEMALE_25_34 *     | -39.214446    | 0.0000000 |
| UnemploymentRate * | -2955.157296  | 0.0000197 |
| Month9Ago *        | 2645.246189   | 0.0000000 |

Table 19: Step Linear Model VIFs

|                  |           |
|------------------|-----------|
| Month            | 1.063277  |
| TOT_POP          | 18.523696 |
| FEMALE_25_34     | 17.820910 |
| UnemploymentRate | 1.427085  |
| Month9Ago        | 1.028006  |

## 2.5 Significant Variables Minus Linear Model

The next model was aimed at removing variables with multicollinearity evidenced by the high VIFs we'd seen on earlier models. The adjusted  $R^2$  value for this model is 0.36007.



Table 20: Significant Variables Minus Linear Model Coefficient Estimates

|                | Estimate        | Pr(> t )  |
|----------------|-----------------|-----------|
| Intercept *    | 36913929.49606  | 0.0000048 |
| Month *        | 2863.98730      | 0.0000000 |
| GenderRatio *  | -69493643.00948 | 0.0000069 |
| FEMALE_25_34 * | -25.38478       | 0.0006337 |
| FEMALE_35_44   | -20.81506       | 0.0506143 |
| Earnings *     | -479.99715      | 0.0138361 |

Table 21: Significant Variables Minus Linear Model VIFs

|              |           |
|--------------|-----------|
| Month        | 1.415159  |
| GenderRatio  | 13.281871 |
| FEMALE_25_34 | 14.099426 |
| FEMALE_35_44 | 27.701955 |
| Earnings     | 56.329990 |

## 2.6 Significant Variables Limited Linear Model

A manual review of features and the introduction of a 9 month lag variable brought us to the next model. The adjusted  $R^2$  value for this model is 0.52589.

Table 22: Significant Variables Limited Linear Model Coefficient Estimates

|                    | Estimate      | Pr(> t )  |
|--------------------|---------------|-----------|
| Intercept *        | 468350.706460 | 0.0000000 |
| Month *            | 2490.804603   | 0.0000000 |
| Month9Ago *        | 2649.171132   | 0.0000000 |
| FEMALE_25_34 *     | -7.181586     | 0.0003757 |
| UnemploymentRate * | -2190.844273  | 0.0030120 |

Table 23: Significant Variables Limited Linear Model VIFs

|                  |          |
|------------------|----------|
| Month            | 1.062619 |
| Month9Ago        | 1.028000 |
| FEMALE_25_34     | 1.400205 |
| UnemploymentRate | 1.360376 |

## 3 Select Models

A validation data set (VS) was created from a subset of the full dataset for use in the multiple linear regression. This VS data set was used to perform a level of independent validation of the previously described models. The validation metric for the multiple linear regression models is the mean squared error from the validation set.

The results of the multiple linear regression model validation are shown below.

Table 24: Linear Model Validation Error Results

| Model               | VS Error  | Adj R <sup>2</sup> | Variables | VIF |
|---------------------|-----------|--------------------|-----------|-----|
| All Variables       | 161072343 | 0.6112935          | 11        | BAD |
| Step                | 172024296 | 0.6200088          | 5         | BAD |
| Significant Limited | 176416016 | 0.5258888          | 4         | OK  |
| High Cor            | 212269346 | 0.4788873          | 6         | BAD |
| Significant         | 227028994 | 0.4771385          | 7         | BAD |
| Significant Minus   | 231634851 | 0.3600735          | 5         | BAD |

Based on the criteria of least complex model with lowest validation error, highest  $R^2$  and no multicollinearity issues, the Significant Limited model is favored for further investigation.

### 3.1 Evaluation: Significant Limited Linear Model

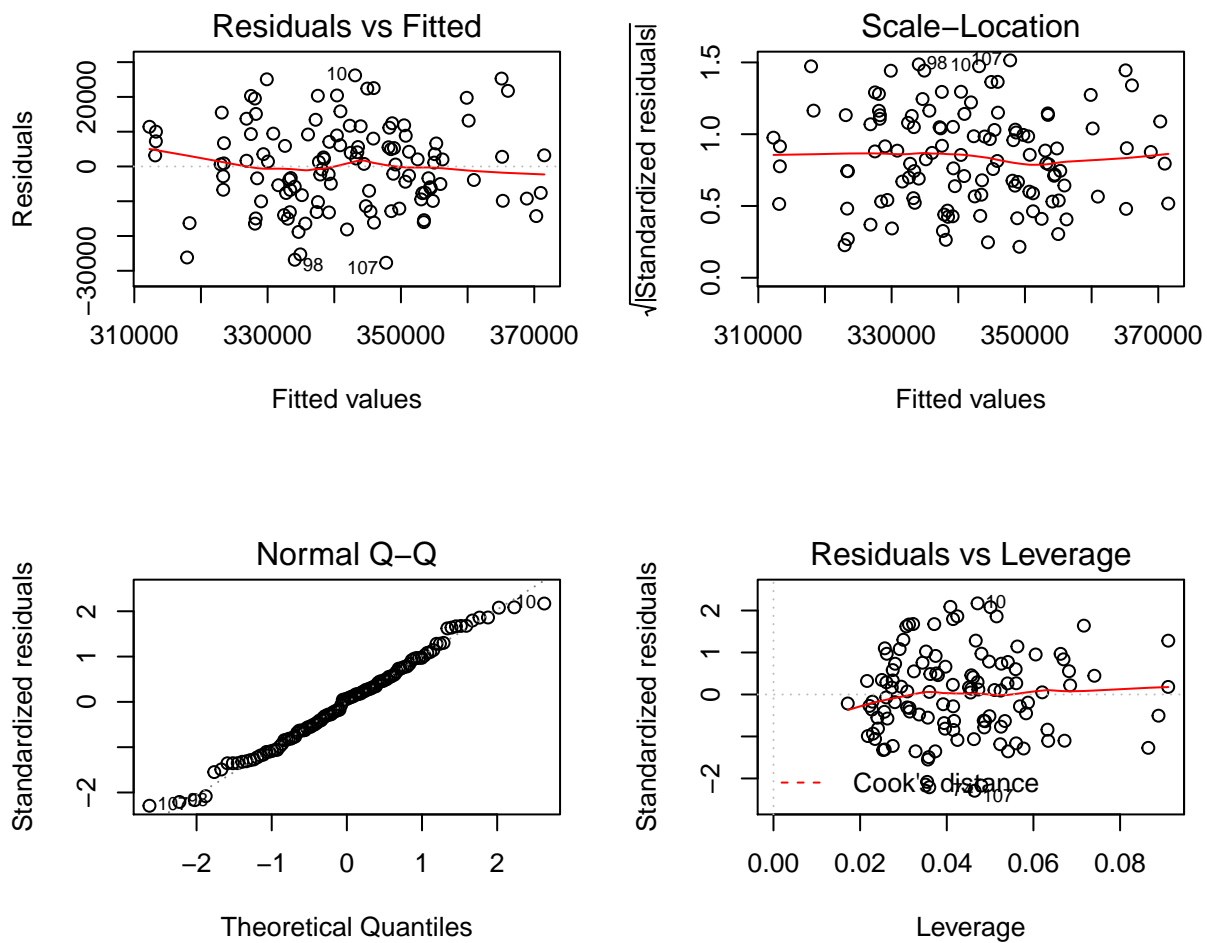
The Significant Limited model has an F-statistic of 32.89 and a mean squared error (MSE) of 146358133.79.

$$\begin{aligned}
 y_{births} = & 468350.7064601 + 2490.8046026x_{Month} \\
 & + 2649.1711324x_{Month9Ago} \\
 & - 7.181586x_{FEMALE\_25\_34} \\
 & - 2190.8442727x_{UnemploymentRate}
 \end{aligned}$$

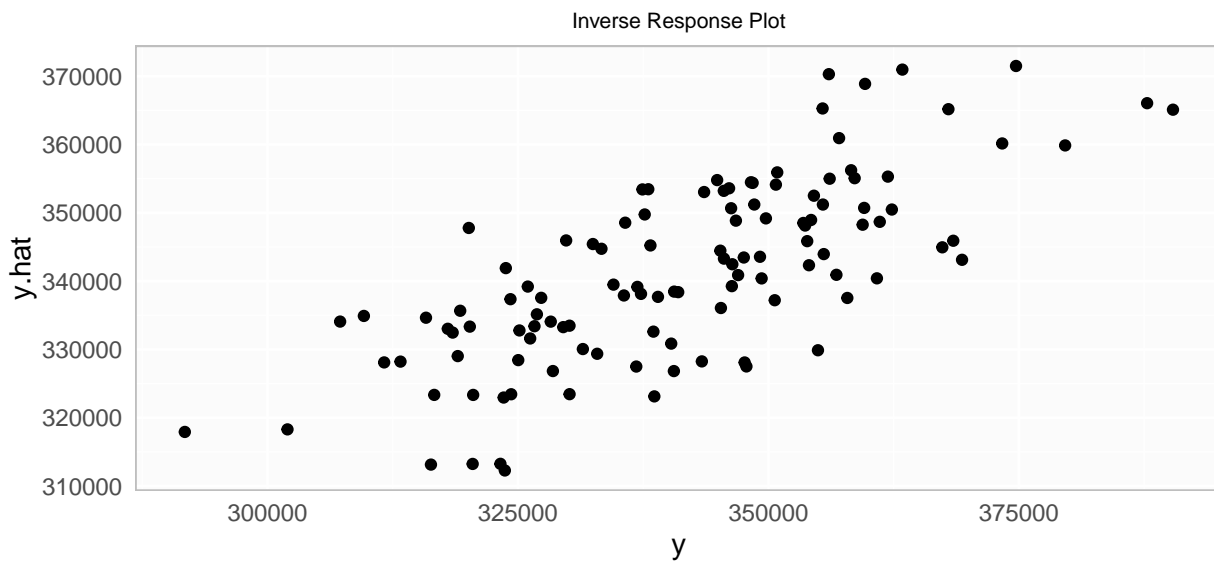
We can interpret the coefficients in the following manner. Holding all other predictors constant, for variable:

- *Month*, as the month of the year increased, a 2490.8 increase in births would occur.
- *Month9Ago*, as the 9 month lagged month of the year increased, a 2649.17 increase in births would occur.
- *FEMALE\_25\_34*, a unit increase in the population of females age 25-34 would yield a 7.18 decrease in births.
- *UnemploymentRate*, a unit increase in the *UnemploymentRate* related to a 2190.84 decrease in births.

Linear regression diagnostic plots are shown below. Residuals appear to be normally distributed and variance seems to be fairly constant. There are definitely some outliers which the model is not capturing.



Looking at the inverse response plot, there does appear to be a good linear pattern to the predicted response versus actual.



Running a more targeted auto-correlation analysis with R's *acf* function shows a possible pattern of negative and positive residual oscillation and inparticular at the lag of 12 .

## Significant Limited Linear Model Auto-correlation Plot

