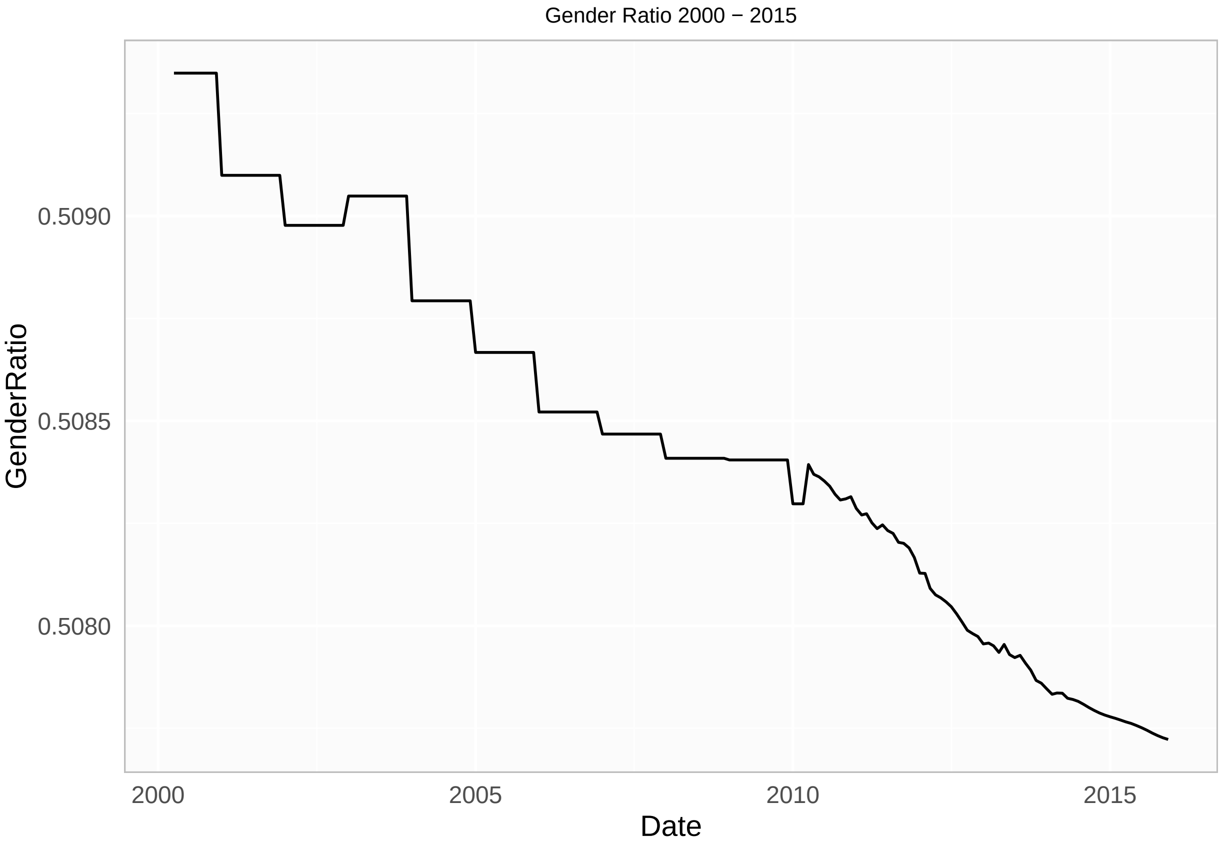
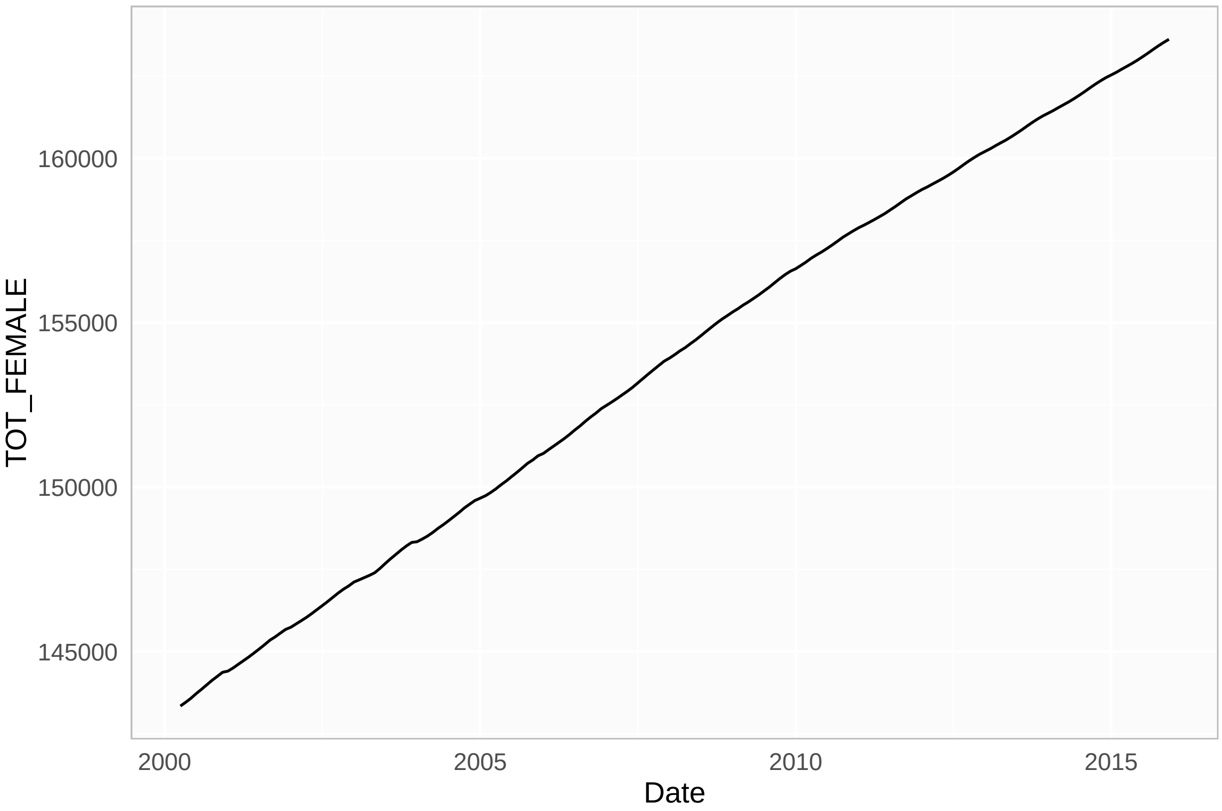
Natality Models Data Exploration

DATA 621: Business Analytics and Data Mining

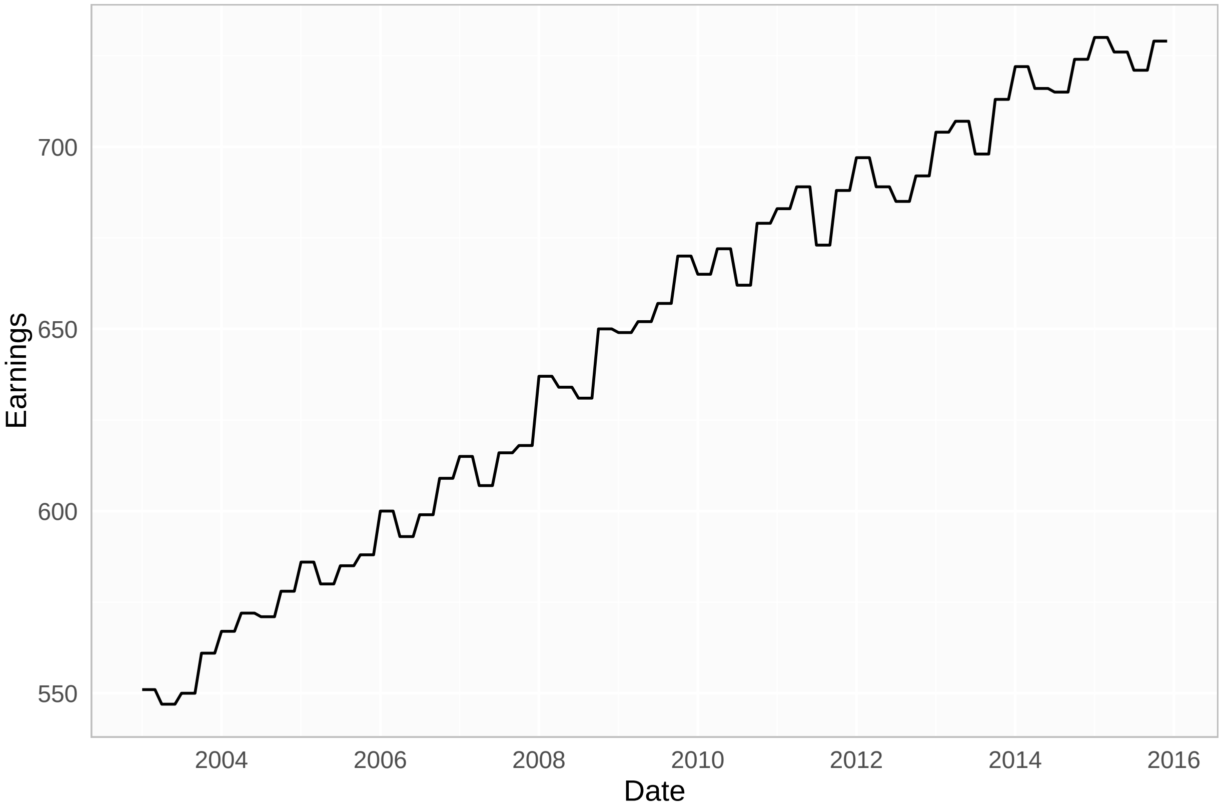
Daniel Dittenhafer & Justin Hink

April 24, 2016

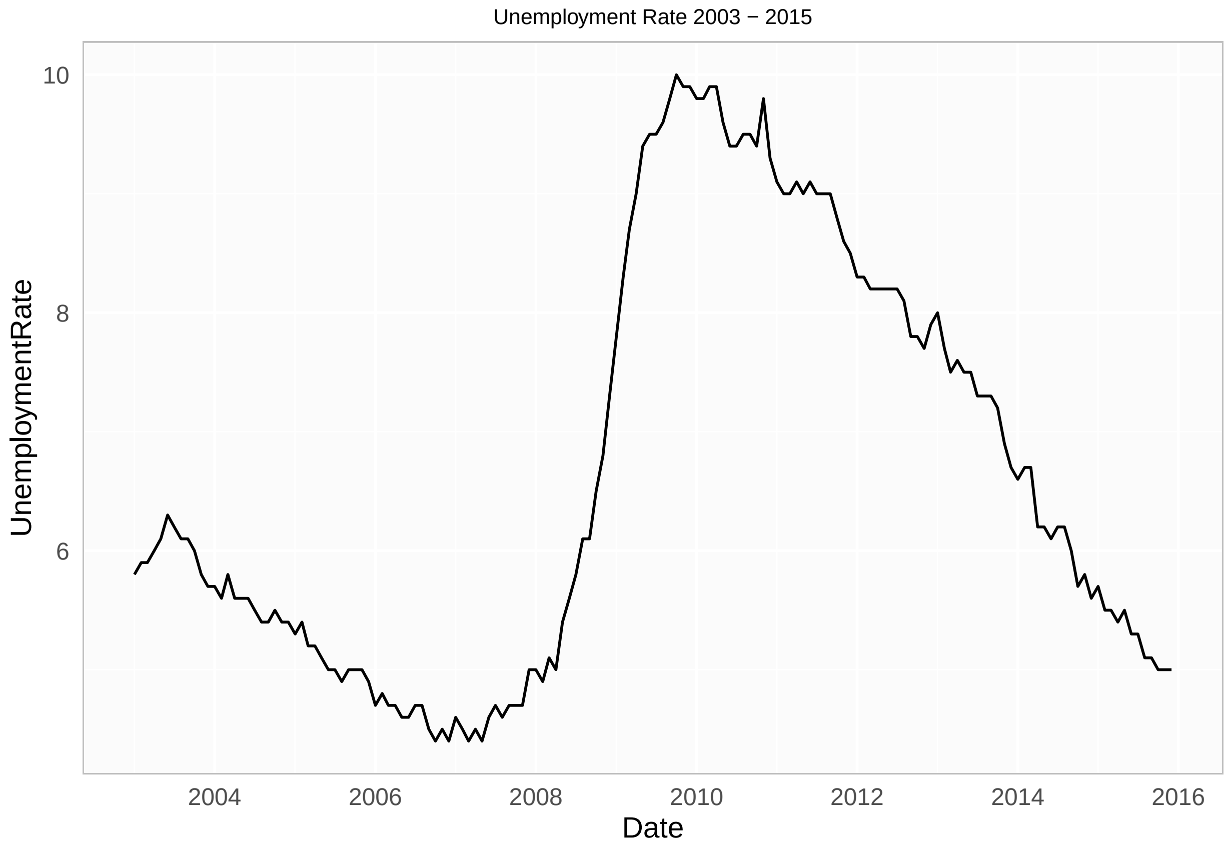




Female Population 2000 − 2015

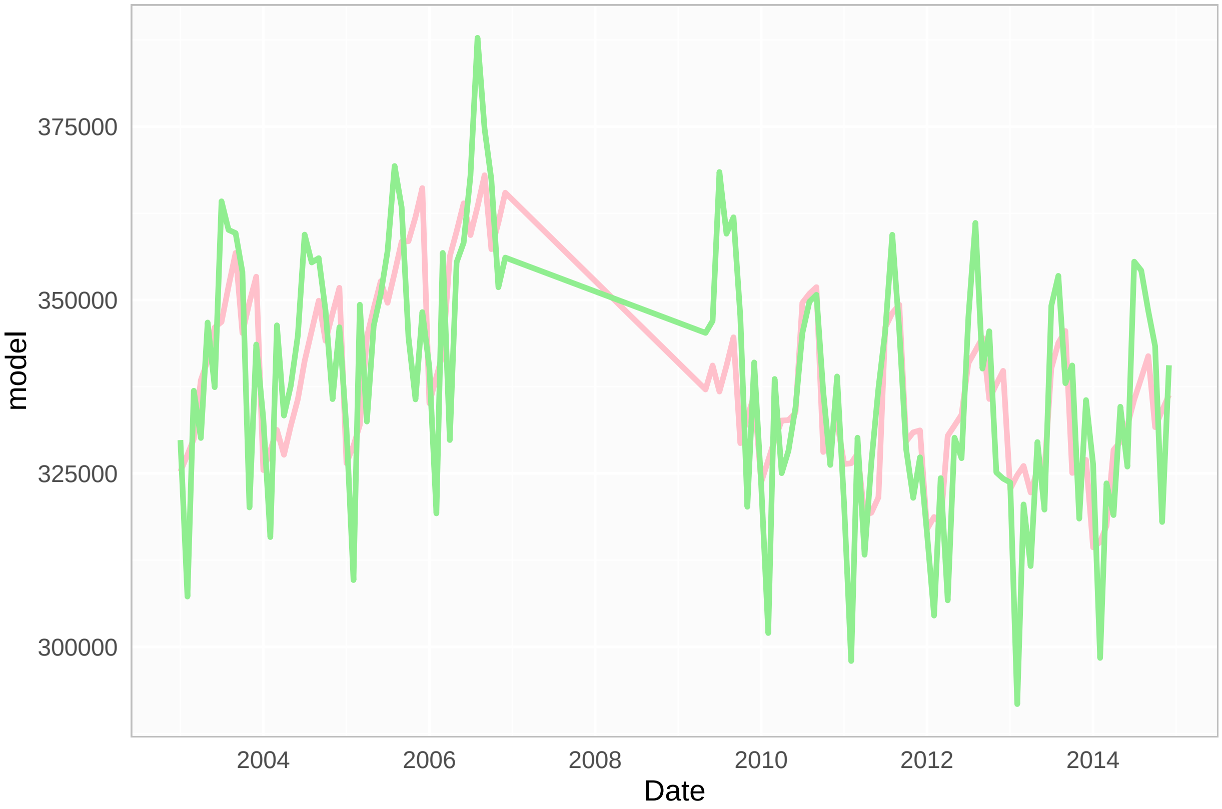


Women's Weekly Earnings 2003 − 2015



```
##       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
##  Min.   :20353   Min.   :547.0   Min.   : 4.400
##  1st Qu.:20398   1st Qu.:591.8   1st Qu.: 5.175
##  Median :21012   Median :649.5   Median : 6.150
##  Mean   :21120   Mean   :640.5   Mean   : 6.757
##  3rd Qu.:21787   3rd Qu.:688.2   3rd Qu.: 8.300
##  Max.   :22303   Max.   :724.0   Max.   :10.000
```

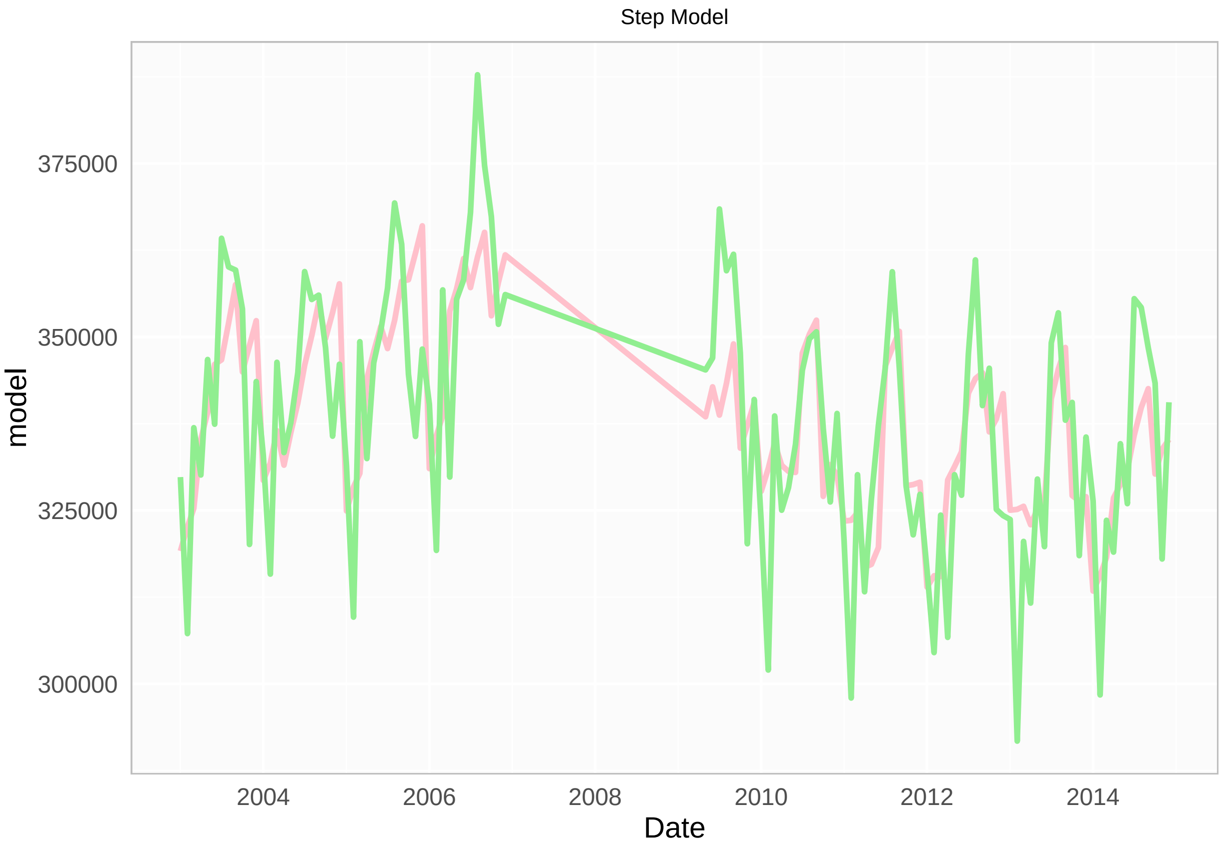
Signif Vars – High VIFs Model



```
##
## Call:
## lm(formula = Births ~ Month + GenderRatio + FEMALE_25_34 + FEMALE_35_44 +
##     Earnings, data = modelData)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -28812  -8656   1604   9136  33617
##
## Coefficients:
##                    Estimate    Std. Error t value      Pr(>|t|)
## (Intercept)   28468830.590   7403888.065   3.845      0.000202 ***
## Month             2692.415       409.664   6.572 0.00000000171 ***
## GenderRatio  -53512891.053  14212182.543  -3.765      0.000269 ***
## FEMALE_25_34       -12.120         7.322  -1.655      0.100707
## FEMALE_35_44       -17.276         9.708  -1.780      0.077918 .
## Earnings          -517.481       186.537  -2.774      0.006504 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14020 on 110 degrees of freedom
## Multiple R-squared:  0.4137, Adjusted R-squared:  0.387
## F-statistic: 15.52 on 5 and 110 DF,  p-value: 0.00000000001548
```

```
## Start:  AIC=2202.89
## Births ~ Month + (Year + Month + Date + TOT_POP + GenderRatio +
##     TOT_FEMALE + TOT_MALE + FEMALE_15_24 + FEMALE_25_34 + FEMALE_35_44 +
##     Earnings + UnemploymentRate) - Year - Date
```

```
##
##
## Step:  AIC=2202.89
## Births ~ Month + TOT_POP + GenderRatio + TOT_FEMALE + FEMALE_15_24 +
##     FEMALE_25_34 + FEMALE_35_44 + Earnings + UnemploymentRate
##
##                    Df  Sum of Sq         RSS    AIC
## - Month             1  140580486 17399005905 2201.8
## - FEMALE_25_34      1  234610171 17493035590 2202.4
## - TOT_POP           1  237725165 17496150583 2202.5
## - GenderRatio       1  242255960 17500681379 2202.5
## - TOT_FEMALE        1  242733933 17501159352 2202.5
## <none>                           17258425419 2202.9
## - FEMALE_15_24      1  422425278 17680850696 2203.7
## - UnemploymentRate  1  489250509 17747675928 2204.1
## - FEMALE_35_44      1 1073238233 18331663652 2207.9
## - Earnings          1 5423161788 22681587207 2232.6
##
## Step:  AIC=2201.83
## Births ~ TOT_POP + GenderRatio + TOT_FEMALE + FEMALE_15_24 +
##     FEMALE_25_34 + FEMALE_35_44 + Earnings + UnemploymentRate
##
##                    Df  Sum of Sq         RSS    AIC
## - FEMALE_25_34      1  157515257 17556521162 2200.9
## <none>                           17399005905 2201.8
## - GenderRatio       1  510346910 17909352815 2203.2
## - TOT_POP           1  513484168 17912490073 2203.2
## - TOT_FEMALE        1  522483397 17921489302 2203.3
## - FEMALE_15_24      1  531486162 17930492067 2203.3
## - UnemploymentRate  1  675041804 18074047709 2204.2
## - FEMALE_35_44      1 2924431465 20323437370 2217.8
## - Earnings          1 8000839474 25399845379 2243.7
##
## Step:  AIC=2200.87
## Births ~ TOT_POP + GenderRatio + TOT_FEMALE + FEMALE_15_24 +
##     FEMALE_35_44 + Earnings + UnemploymentRate
##
##                    Df  Sum of Sq         RSS    AIC
## <none>                           17556521162 2200.9
## - FEMALE_15_24      1  417077960 17973599122 2201.6
## - UnemploymentRate  1  657746838 18214268000 2203.1
## - GenderRatio       1 1154633167 18711154329 2206.3
## - TOT_POP           1 1155512172 18712033334 2206.3
## - TOT_FEMALE        1 1162646725 18719167887 2206.3
## - FEMALE_35_44      1 3195731081 20752252243 2218.3
## - Earnings          1 7913534201 25470055363 2242.0
```



```
##
## Call:
## lm(formula = Births ~ TOT_POP + GenderRatio + TOT_FEMALE + FEMALE_15_24 +
##     FEMALE_35_44 + Earnings + UnemploymentRate, data = modelData)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -33400  -6663    734   8974  29709
##
## Coefficients:
##                       Estimate     Std. Error t value       Pr(>|t|)
## (Intercept)        896454593.57   337856822.49   2.653        0.00917 **
## TOT_POP               -2928.23        1098.31  -2.666        0.00885 **
## GenderRatio      -1770699468.75   664400877.15  -2.665        0.00888 **
## TOT_FEMALE             5794.49        2166.70   2.674        0.00865 **
## FEMALE_15_24            -95.65          59.72  -1.602        0.11213
## FEMALE_35_44             81.00          18.27   4.434 0.000022341275 ***
## Earnings              -1578.71         226.27  -6.977 0.000000000253 ***
## UnemploymentRate       6735.09        3348.28   2.012        0.04676 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12750 on 108 degrees of freedom
## Multiple R-squared:  0.5242, Adjusted R-squared:  0.4933
## F-statistic:    17 on 7 and 108 DF,  p-value: 0.000000000000005533
```

1 Data Exploration

The unified data set for this project contains 144 rows of data with 1 response variable and 12 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

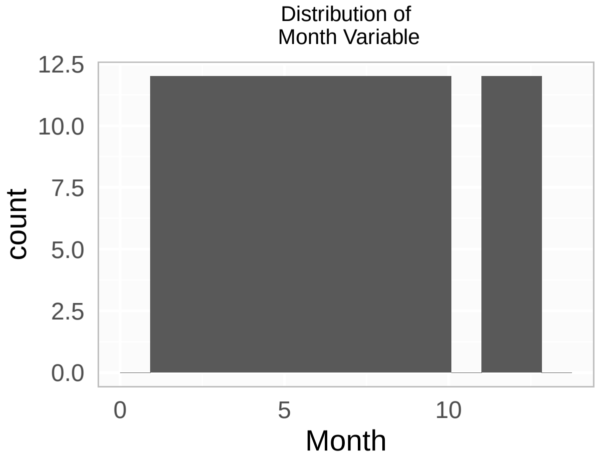
Births	1.0000000
FEMALE_35_44	0.3880661
Month	0.3646307
GenderRatio	0.2862173
FEMALE_15_24	-0.2307949
TOT_MALE	-0.3214851
TOT_POP	-0.3219328
TOT_FEMALE	-0.3223760
Year	-0.3593053
Earnings	-0.3697992
UnemploymentRate	-0.3862666
FEMALE_25_34	-0.3879287

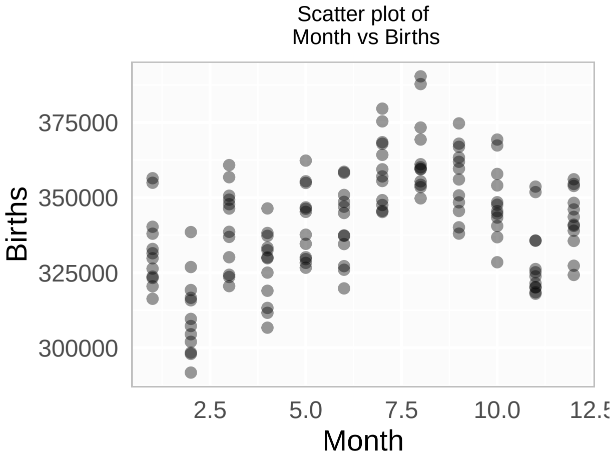
1.3 Variable Month

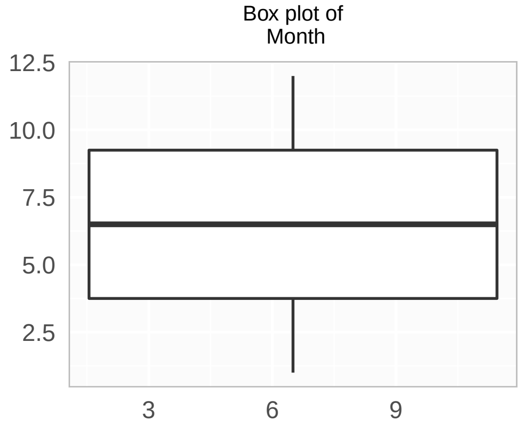
The *Month* variable is the month of birth. As one should expect, the distribution is uniform, but we can see some seasonality to the relationship between *Births* and *Month* with July and August being high frequency birth months.

Table 2: Month Variable Statistics

min	mean	stdev	median	max
1	6.5	3.464102	6.5	12





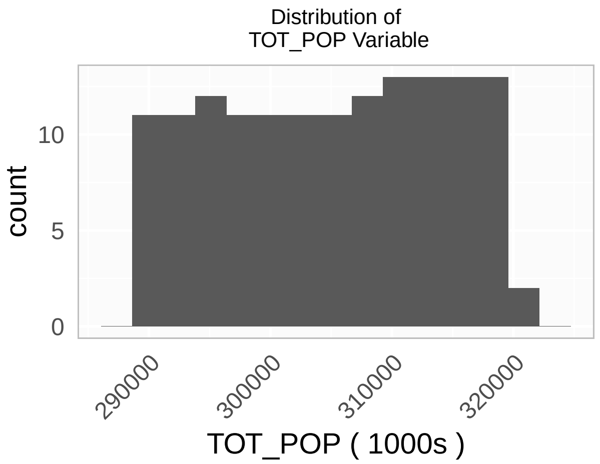


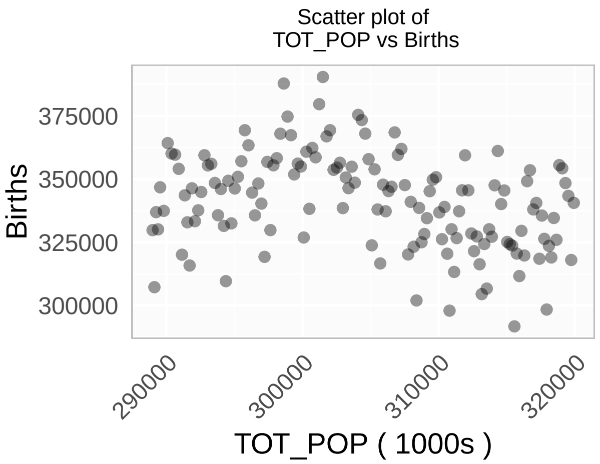
1.4 Variable TOT_POP

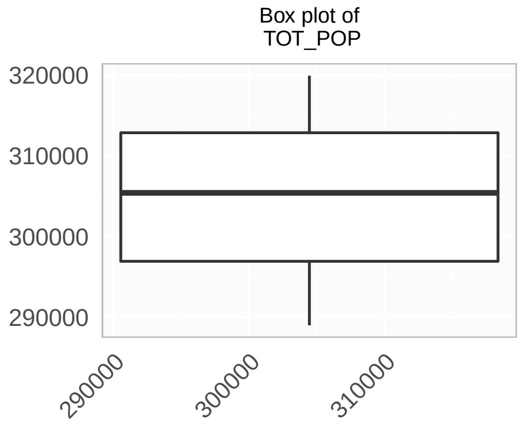
The *TOT_POP* variable is the total population per month as esimated by the Census Bureau.

Table 3: TOT_POP Variable Statistics

min	mean	stdev	median	max
288998.8	304885.4	9171.506	305409.3	319925.2





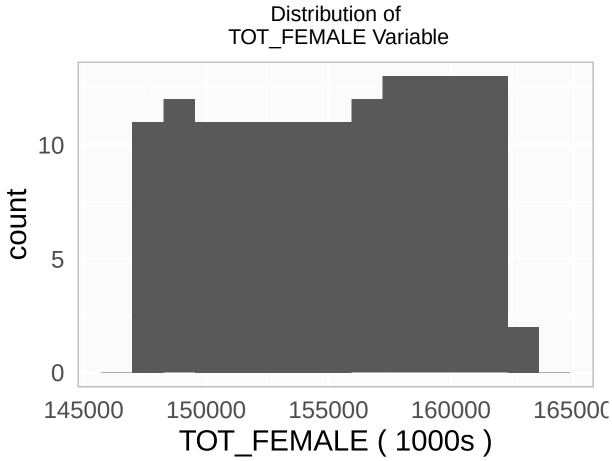


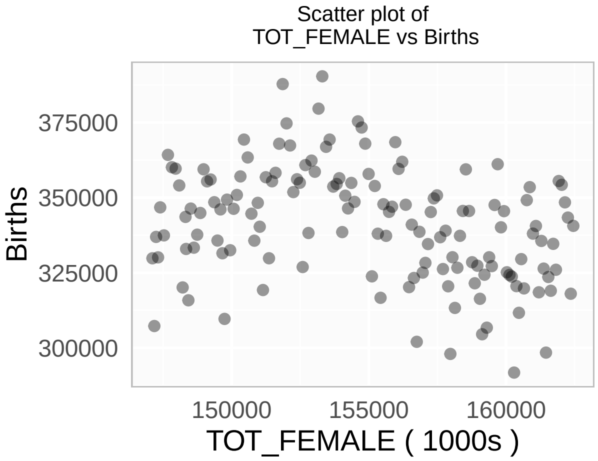
1.5 Variable TOT_FEMALE

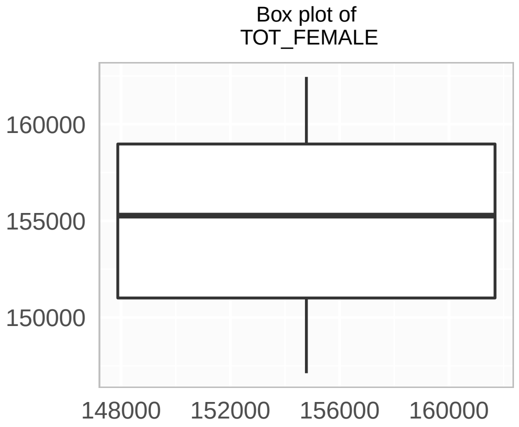
The *TOT_FEMALE* variable is the total population of females per month as estimated by the Census Bureau.

Table 4: TOT_FEMALE Variable Statistics

min	mean	stdev	median	max
147114.4	154997.1	4561.405	155272.1	162452.2





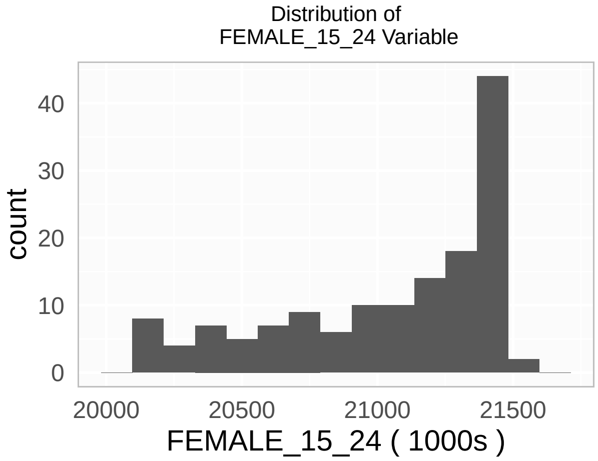


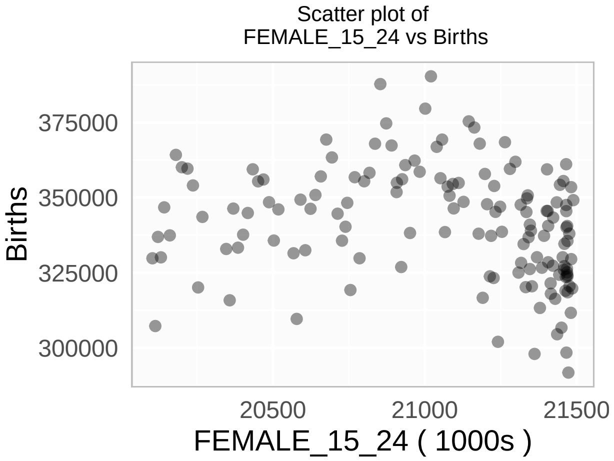
1.6 Variable FEMALE_15_24

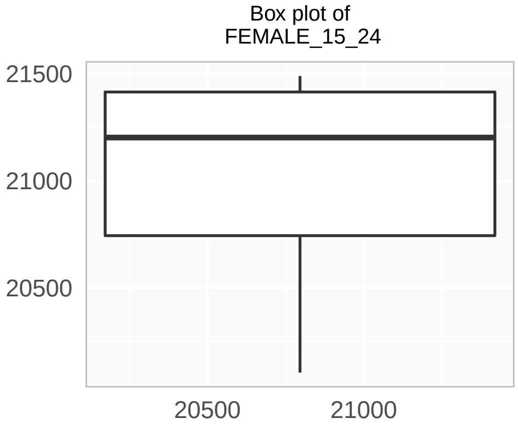
The *FEMALE_15_24* variable is the total population of females ages 15-24 per month as estimated by the Census Bureau.

Table 5: FEMALE_15_24 Variable Statistics

min	mean	stdev	median	max
20103.14	21046.7	422.1778	21201.43	21489.1





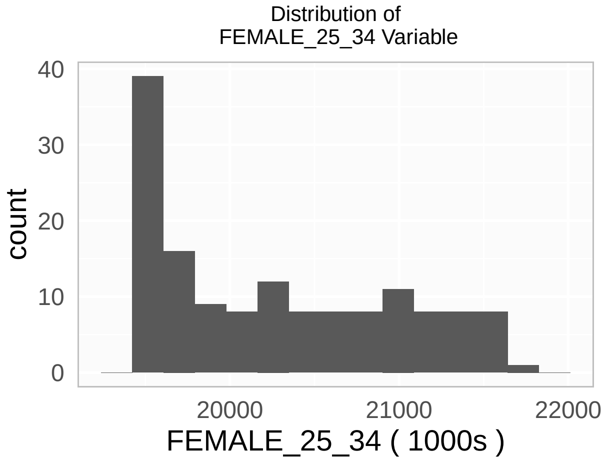


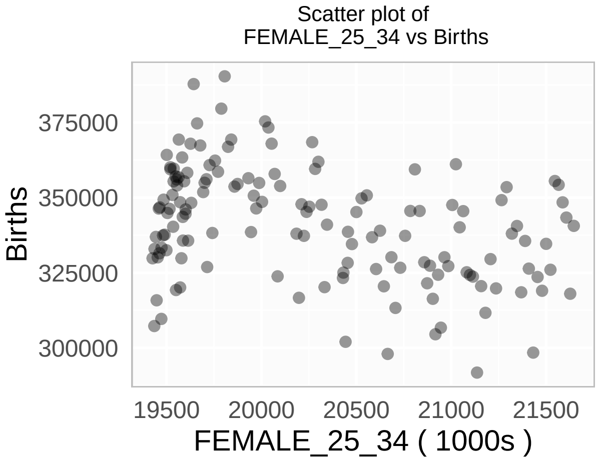
1.7 Variable FEMALE_25_34

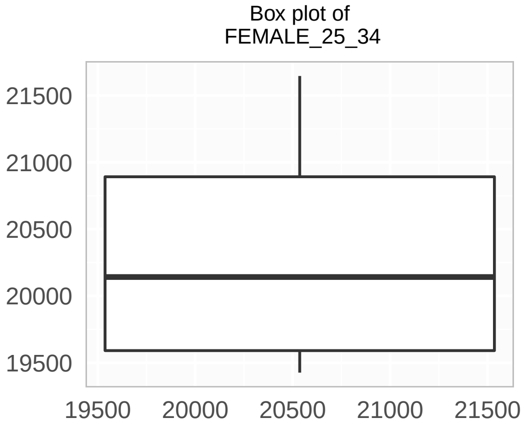
The *FEMALE_25_34* variable is the total population of females ages 25-34 per month as estimated by the Census Bureau.

Table 6: FEMALE_25_34 Variable Statistics

min	mean	stdev	median	max
19426.37	20274.31	701.1676	20141.73	21646.13





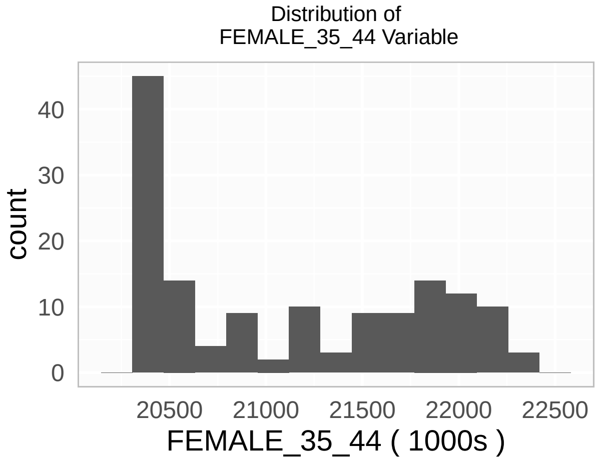


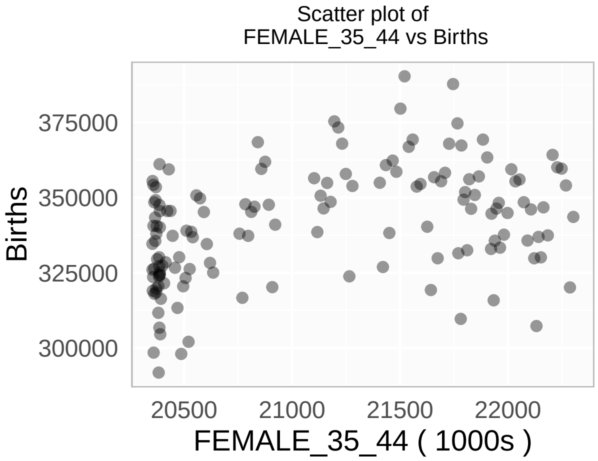
1.8 Variable FEMALE_35_44

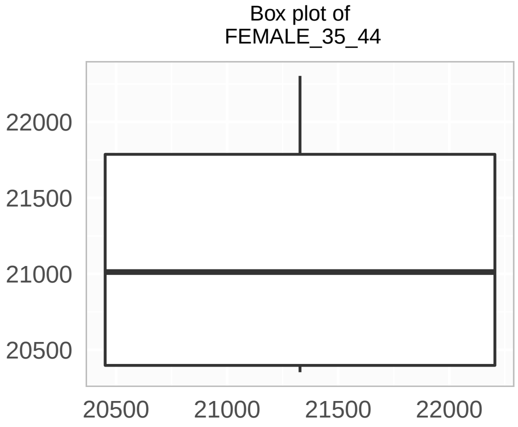
The *FEMALE_35_44* variable is the total population of females ages 35-44 per month as estimated by the Census Bureau.

Table 7: FEMALE_35_44 Variable Statistics

min	mean	stdev	median	max
20353.37	21120.04	683.5963	21012.17	22302.87


Distribution of
FEMALE_35_44 Variable


Scatter plot of
FEMALE_35_44 vs Births

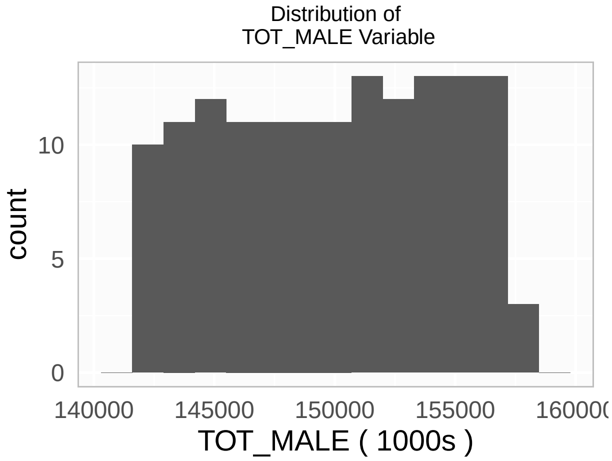

Box plot of
FEMALE_35_44

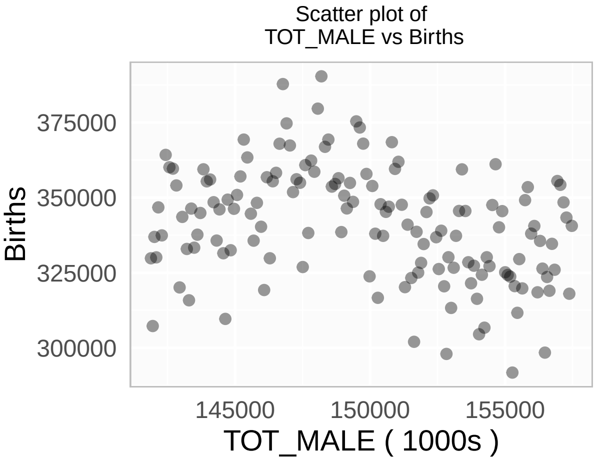
1.9 Variable TOT_MALE

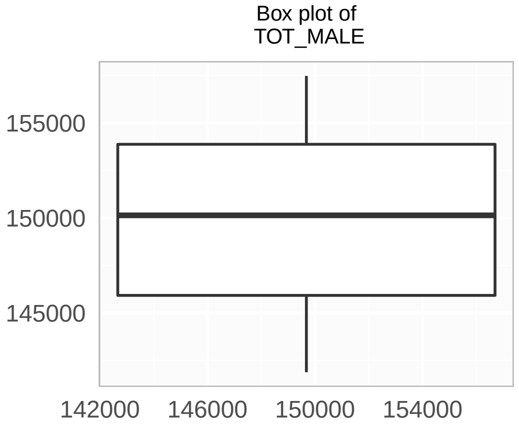
The *TOT_MALE* variable is the total population of females per month as esimated by the Census Bureau.

Table 8: TOT_MALE Variable Statistics

min	mean	stdev	median	max
141884.4	149888.3	4610.232	150137.2	157472.9





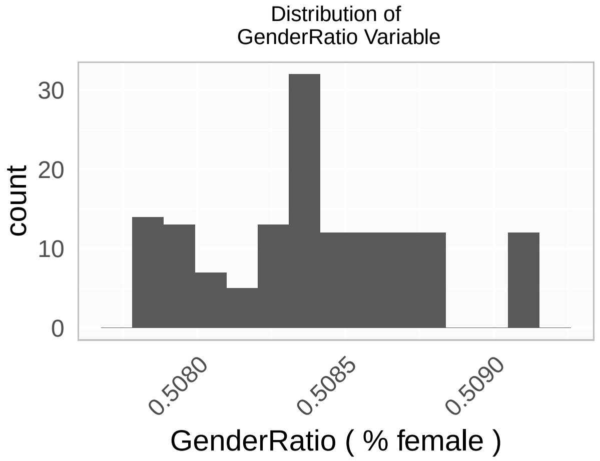


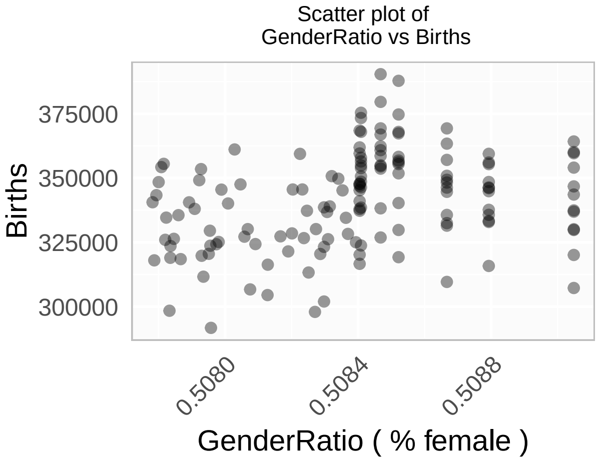
1.10 Variable GenderRatio

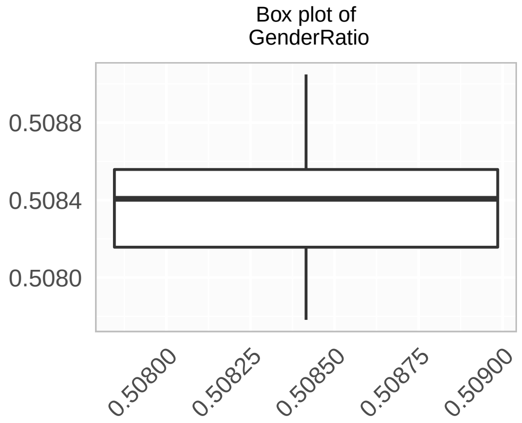
The *GenderRatio* variable is the percentage of the total population which are females per month derived from data from the Census Bureau. In cases where month data was not available, the annual gender ratio was computed and applied to the monthly total population.

Table 9: GenderRatio Variable Statistics

min	mean	stdev	median	max
0.507782	0.5083882	0.0003426	0.5084067	0.5090486


Distribution of GenderRatio Variable


Scatter plot of GenderRatio vs Births

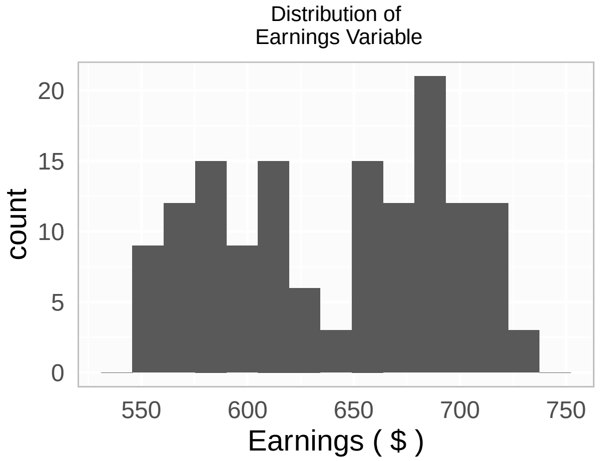

Box plot of GenderRatio

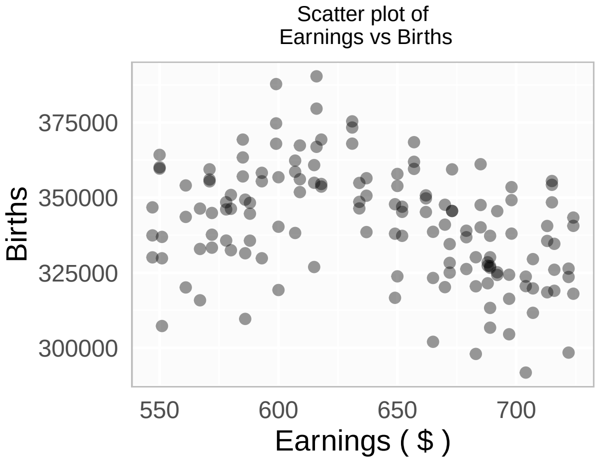
1.11 Variable Earnings

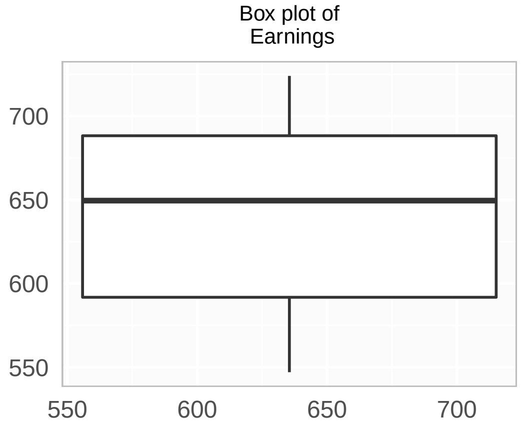
The *Earnings* variable is womoen's weekly earnings in current dollars based on data from the Bureau of Labor Statistics. The original values were provided quarterly and were expanded to a monthly format for data analysis purposes.

Table 10: Earnings Variable Statistics

min	mean	stdev	median	max
547	640.5417	53.55213	649.5	724





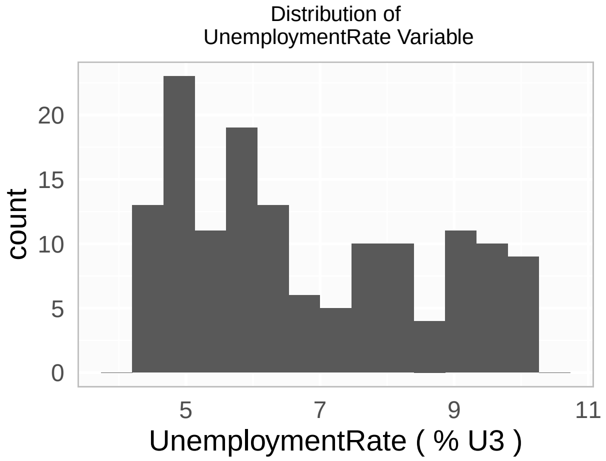


1.12 Variable UnemploymentRate

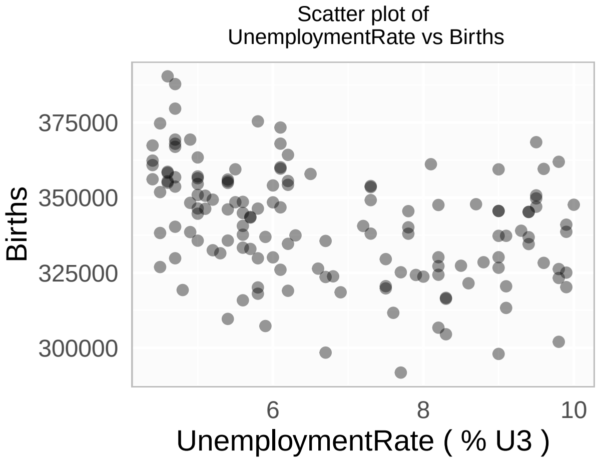
The *UnemploymentRate* variable is the unemployment rate per month (U3) based on data from the Bureau of Labor Statistics.

Table 11: UnemploymentRate Variable Statistics

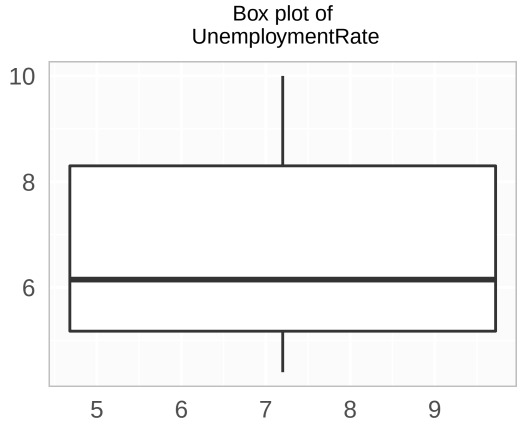
min	mean	stdev	median	max
4.4	6.756944	1.789466	6.15	10



Distribution of UnemploymentRate Variable



Scatter plot of UnemploymentRate vs Births



Box plot of UnemploymentRate

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 0.49254.

Table 12: All Variables Linear Model Coefficient Estimates

	Estimate	Pr(>|t|)
Intercept	527377744.96006	0.2275515
Month	560.49655	0.3548904
TOT_POP	-1714.42312	0.2296045
GenderRatio	-1043221418.58050	0.2252467
TOT_FEMALE	3411.13517	0.2247929
FEMALE_15_24	-101.62980	0.1102088
FEMALE_25_34	-48.54466	0.2326601
FEMALE_35_44 *	63.21939	0.0116395
Earnings *	-1478.32035	0.0000001
UnemploymentRate	6007.83544	0.0859187

Table 13: All Variables Linear Model VIFs

Month	3.036506
TOT_POP	146007289.016866
GenderRatio	74955.232693
TOT_FEMALE	139977071.322651
TOT_MALE	72980.616443
FEMALE_15_24	252.850240
FEMALE_25_34	244.885968
FEMALE_35_44	25942.166841
Earnings	29475.962010
UnemploymentRate	0.815039

2.2 Signficant Variables Linear Model

The second multiple linear regression model uses predictor variables indicated as significant from the All Variables model. The adjusted R^2 value for this model is 0.47839.

Table 14: Signficant Variables Linear Model Coefficient Estimates

	Estimate	Pr(>|t|)
Intercept	447550392.10991	0.2426656
TOT_POP	-1566.11592	0.2133924
GenderRatio	-887242452.20354	0.2389139
TOT_FEMALE	3117.56224	0.2081650
FEMALE_15_24	-57.38770	0.3138821
FEMALE_25_34	-36.08217	0.3620445
FEMALE_35_44 *	47.93124	0.0000322
Earnings *	-1477.49061	0.0000000

Table 15: Signficant Variables Linear Model VIFs

TOT_POP	110470539.51333
GenderRatio	55961.53742
TOT_FEMALE	105774732.35076
FEMALE_15_24	483.86089
FEMALE_25_34	610.56768
FEMALE_35_44	46.92232
Earnings	118.73239

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 0.49415.

Table 16: High Correlation Variables Linear Model Coefficient Estimates

	Estimate	Pr(>|t|)
Intercept *	-2929795.91132	0.0011256
FEMALE_25_34 *	-42.89940	0.0000031
UnemploymentRate	3023.19686	0.1026913

	Estimate	Pr(>|t|)
FEMALE_35_44 *	42.01009	0.0230580
Earnings *	-1363.94318	0.0000001
Month	760.40107	0.1496743
TOT_FEMALE *	26.45528	0.0000001

Table 17: High Correlation Variables Linear Model VIFs

FEMALE_25_34	30.783323
UnemploymentRate	7.583020
FEMALE_35_44	131.753678
Earnings	144.176105
Month	2.299555
TOT_FEMALE	391.468595

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 0.49333.

Table 18: Step Linear Model Coefficient Estimates

	Estimate	Pr(>|t|)
Intercept *	896454593.57492	0.0091721
TOT_POP *	-2928.22791	0.0088520
GenderRatio *	-1770699468.74884	0.0088771
TOT_FEMALE *	5794.48744	0.0086514
FEMALE_15_24	-95.65110	0.1121266
FEMALE_35_44 *	81.00347	0.0000223
Earnings *	-1578.71082	0.0000000
UnemploymentRate *	6735.08673	0.0467621

Table 19: Step Linear Model VIFs

TOT_POP	87629209.36379
GenderRatio	45307.39811
TOT_FEMALE	84328003.05117
FEMALE_15_24	552.19195
FEMALE_35_44	132.17785
Earnings	125.83397
UnemploymentRate	25.15273

3 Select Models

A validation data set (VS) was created from a subset of the full dataset for use in the mulitple 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 20: Linear Model Validation Error Results

Model	VS Error	Adj R^2	Variables	VIF
Significant	227183351	0.4783943	7	TBD
High Cor	278969733	0.4941529	6	TBD
All Variables	370642958	0.4925352	10	TBD
Step	436852858	0.4933298	7	TBD

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