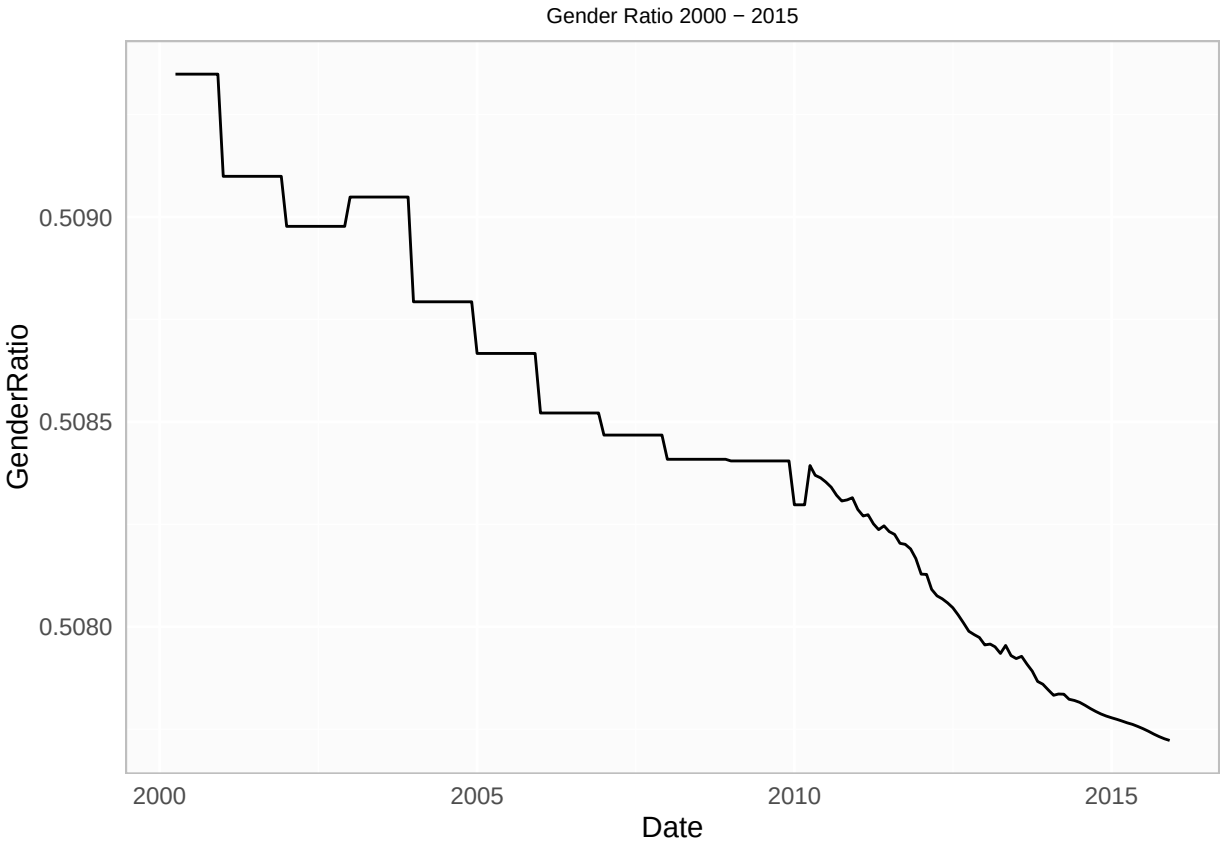
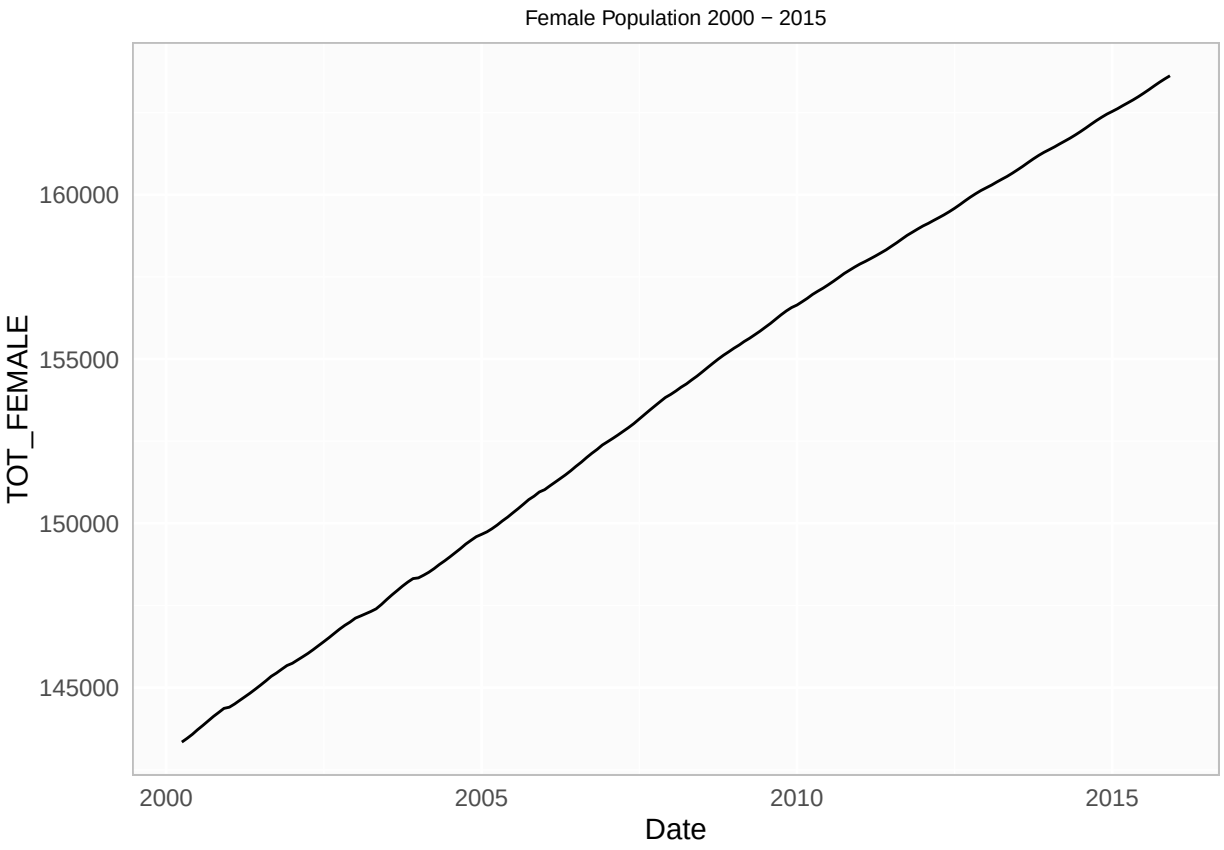
Natality Models Data Exploration

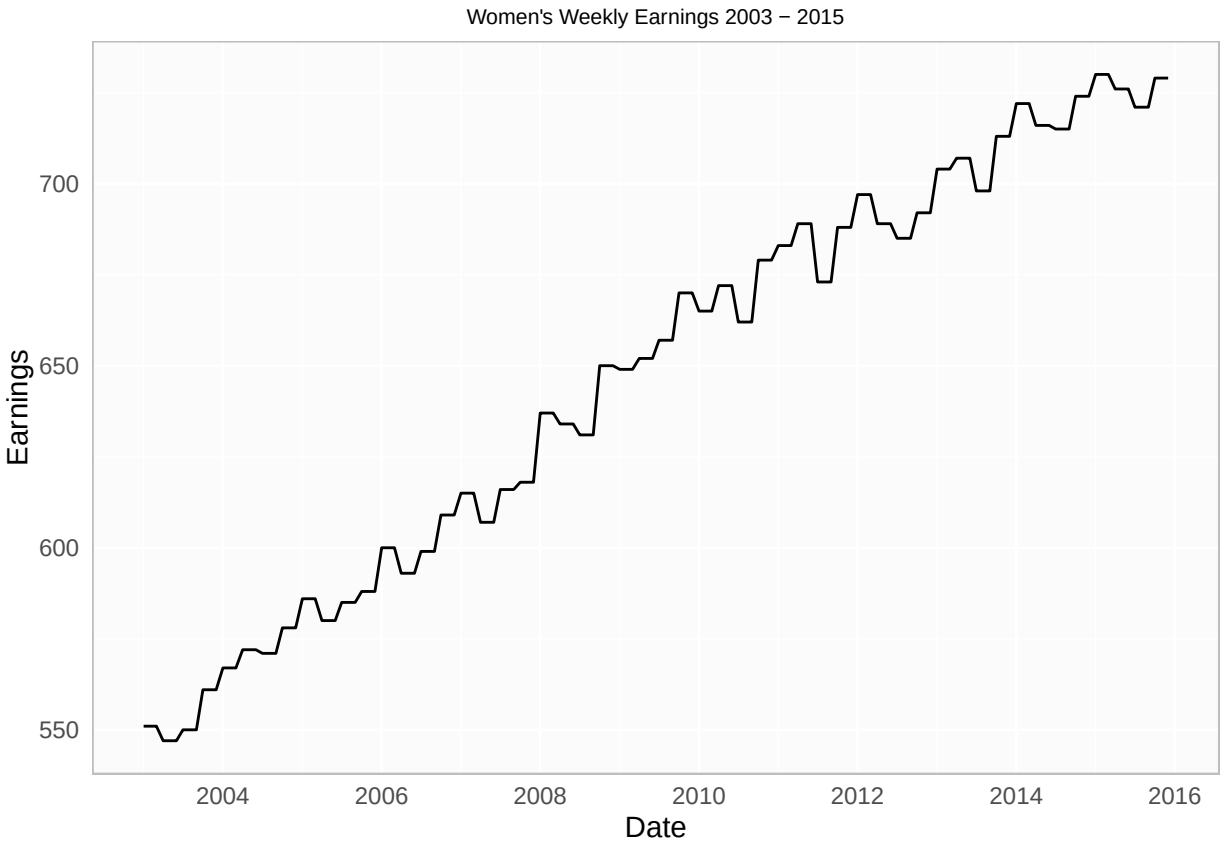
DATA 621: Business Analytics and Data Mining

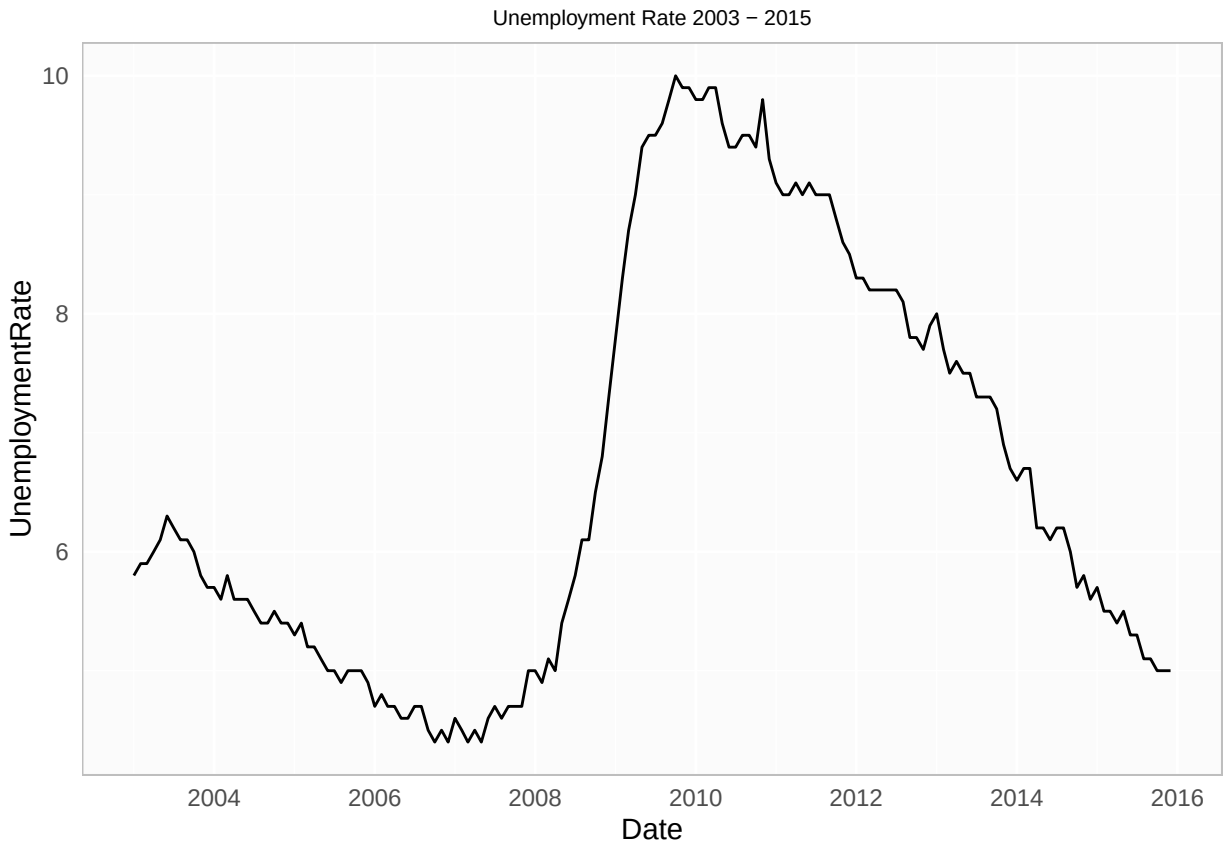
Daniel Dittenhafer & Justin Hink

April 24, 2016




Female Population 2000 − 2015
TOT_FEMALE
160000
155000
150000
145000
2000
2005
2010
2015
Date


Women's Weekly Earnings 2003 − 2015
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.   :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
```

```
## Start:  AIC=2170.85
## Births ~ Month + (Year + Month + Date + TOT_POP + GenderRatio +
##     TOT_FEMALE + TOT_MALE + FEMALE_15_24 + FEMALE_25_34 + FEMALE_35_44 +
##     Earnings + UnemploymentRate + Month9Ago) - Year - Date
##
##
## Step:  AIC=2170.85
## Births ~ Month + TOT_POP + GenderRatio + TOT_FEMALE + FEMALE_15_24 +
##     FEMALE_25_34 + FEMALE_35_44 + Earnings + UnemploymentRate +
##     Month9Ago
##
##                    Df  Sum of Sq         RSS    AIC
## - TOT_POP           1          3 12870423478 2168.8
## - TOT_FEMALE        1       8815 12870432291 2168.8
## - GenderRatio       1      34250 12870457725 2168.8
## - UnemploymentRate  1   26081865 12896505341 2169.1
## - Earnings          1   53062010 12923485485 2169.3
## - FEMALE_35_44      1  126190466 12996613942 2170.0
## - FEMALE_15_24      1  217415291 13087838767 2170.8
## <none>                           12870423476 2170.8
## - FEMALE_25_34      1  621713147 13492136622 2174.3
## - Month             1 1888434497 14758857972 2184.7
## - Month9Ago         1 4388001943 17258425419 2202.9
##
## Step:  AIC=2168.85
## Births ~ Month + GenderRatio + TOT_FEMALE + FEMALE_15_24 + FEMALE_25_34 +
##     FEMALE_35_44 + Earnings + UnemploymentRate + Month9Ago
##
##                    Df  Sum of Sq         RSS    AIC
## - GenderRatio       1   21504953 12891928431 2167.1
## - UnemploymentRate  1   30580604 12901004082 2167.1
## - Earnings          1   62059970 12932483448 2167.4
## - FEMALE_35_44      1  215698484 13086121962 2168.8
## <none>                           12870423478 2168.8
## - FEMALE_15_24      1  391382936 13261806414 2170.3
## - TOT_FEMALE        1  659892002 13530315480 2172.7
## - FEMALE_25_34      1  933441638 13803865116 2175.0
## - Month             1 2435164771 15305588250 2187.0
## - Month9Ago         1 4625727105 17496150583 2202.5
##
## Step:  AIC=2167.05
## Births ~ Month + TOT_FEMALE + FEMALE_15_24 + FEMALE_25_34 + FEMALE_35_44 +
##     Earnings + UnemploymentRate + Month9Ago
##
##                    Df  Sum of Sq         RSS    AIC
## - UnemploymentRate  1    9345872 12901274303 2165.1
## - Earnings          1   61267152 12953195583 2165.6
## - FEMALE_35_44      1  194193618 13086122049 2166.8
## <none>                           12891928431 2167.1
## - FEMALE_15_24      1  373926406 13265854837 2168.4
## - TOT_FEMALE        1  723243317 13615171748 2171.4
## - FEMALE_25_34      1  949077775 13841006206 2173.3
## - Month             1 2453251656 15345180087 2185.3
## - Month9Ago         1 4619101833 17511030264 2200.6
##
## Step:  AIC=2165.13
## Births ~ Month + TOT_FEMALE + FEMALE_15_24 + FEMALE_25_34 + FEMALE_35_44 +
```

```
##     Earnings + Month9Ago
##
##                 Df  Sum of Sq         RSS    AIC
## - Earnings       1   52274314 12953548617 2163.6
## <none>                        12901274303 2165.1
## - FEMALE_15_24   1  375307004 13276581307 2166.5
## - FEMALE_35_44   1  429544716 13330819019 2166.9
## - FEMALE_25_34   1 1015402971 13916677274 2171.9
## - TOT_FEMALE     1 1056868571 13958142874 2172.3
## - Month          1 4045243569 16946517872 2194.8
## - Month9Ago      1 5204873539 18106147842 2202.4
##
## Step:  AIC=2163.6
## Births ~ Month + TOT_FEMALE + FEMALE_15_24 + FEMALE_25_34 + FEMALE_35_44 +
##     Month9Ago
##
##                 Df  Sum of Sq         RSS    AIC
## <none>                        12953548617 2163.6
## - FEMALE_15_24   1  449263111 13402811728 2165.6
## - FEMALE_35_44   1  470163367 13423711984 2165.7
## - TOT_FEMALE     1 1005006137 13958554753 2170.3
## - FEMALE_25_34   1 1080850158 14034398775 2170.9
## - Month          1 5450576199 18404124816 2202.3
## - Month9Ago      1 9808376654 22761925270 2227.0
```


```
##
## Call:
## lm(formula = Births ~ Month + Month9Ago + FEMALE_25_34 + UnemploymentRate,
```

```
##     data = modelData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25786.0  -9235.8    552.6   8038.5  25521.9
##
## Coefficients:
##                     Estimate Std. Error t value             Pr(>|t|)
## (Intercept)      466234.450  31177.372  14.954 < 0.0000000000000002 ***
## Month              2737.050    315.725   8.669   0.0000000000000405 ***
## Month9Ago          2625.253    307.381   8.541   0.0000000000000792 ***
## FEMALE_25_34         -7.509      1.638  -4.586   0.0000119515458828 ***
## UnemploymentRate  -1507.504    693.527  -2.174               0.0319 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11440 on 111 degrees of freedom
## Multiple R-squared:  0.606,  Adjusted R-squared:  0.5918
## F-statistic: 42.68 on 4 and 111 DF,  p-value: < 0.00000000000000022

##            Month        Month9Ago     FEMALE_25_34 UnemploymentRate
##         1.034251         1.025088         1.346553         1.339489
```

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

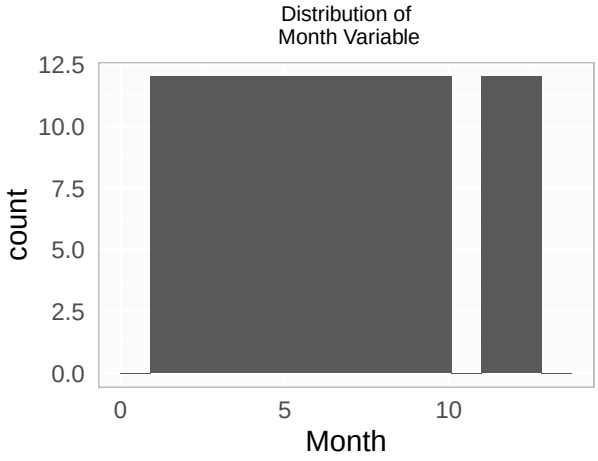
Births	1.0000000
FEMALE_35_44	0.3880661
Month	0.3646307
GenderRatio	0.2862173
FEMALE_15_24	-0.2307949
TOT_MALE	-0.3214851
TOT_POP	-0.3219328
TOT_FEMALE	-0.3223760
Year	-0.3593053
Earnings	-0.3697992
UnemploymentRate	-0.3862666
FEMALE_25_34	-0.3879287

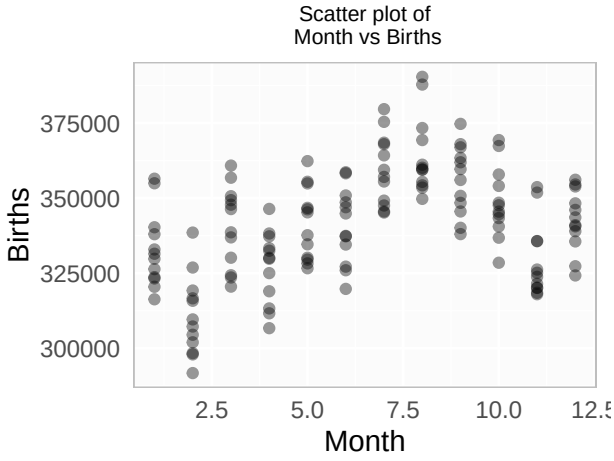
1.3 Variable Month

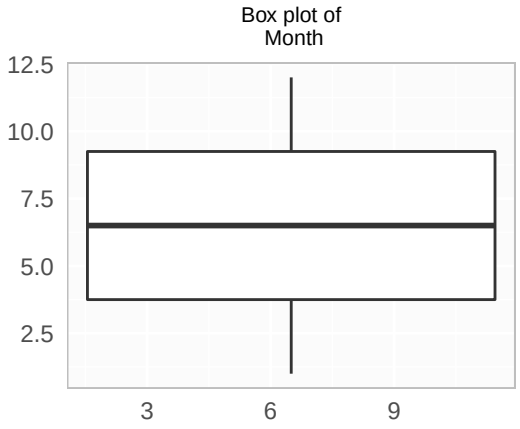
The *Month* variable is the month of birth. As one should expect, the distribution is uniform, but we can see some seasonality to the relationship between *Births* and *Month* with July and August being high frequency birth months.

Table 2: Month Variable Statistics

min	mean	stdev	median	max
1	6.5	3.464102	6.5	12





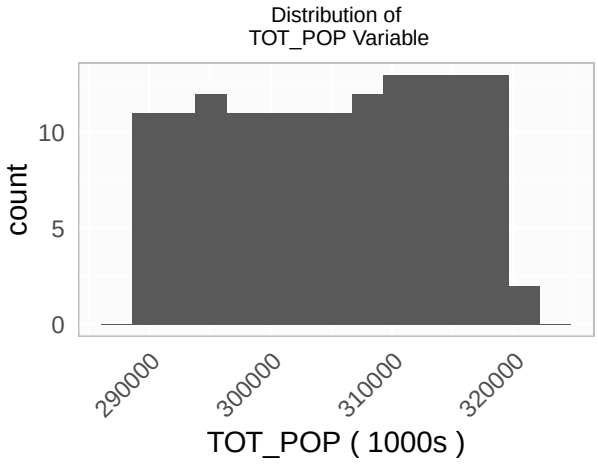


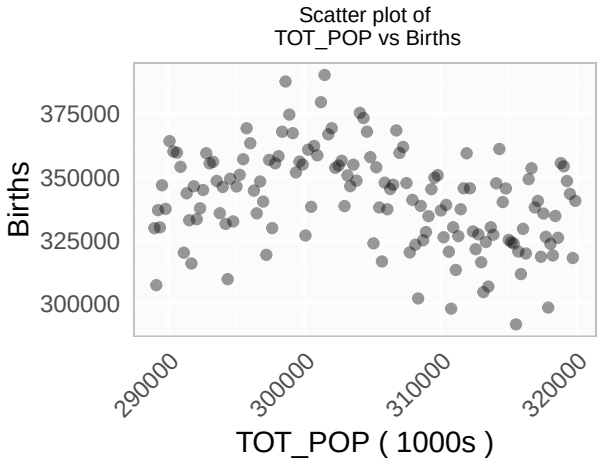
1.4 Variable TOT_POP

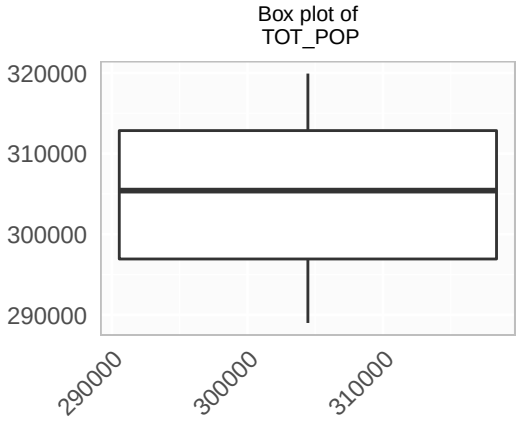
The *TOT_POP* variable is the total population per month as esimated by the Census Bureau.

Table 3: TOT_POP Variable Statistics

min	mean	stdev	median	max
288998.8	304885.4	9171.506	305409.3	319925.2





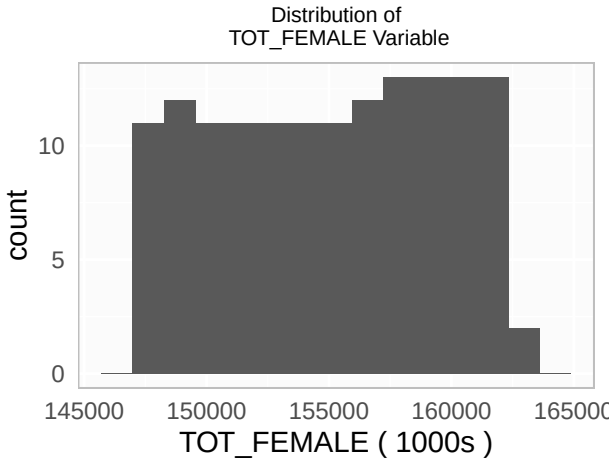


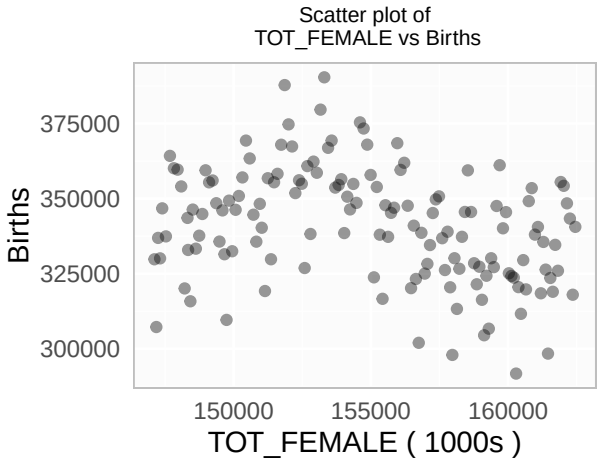
1.5 Variable TOT_FEMALE

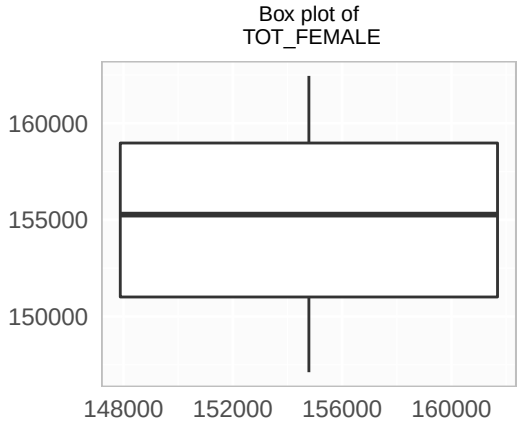
The *TOT_FEMALE* variable is the total population of females per month as estimated by the Census Bureau.

Table 4: TOT_FEMALE Variable Statistics

min	mean	stdev	median	max
147114.4	154997.1	4561.405	155272.1	162452.2





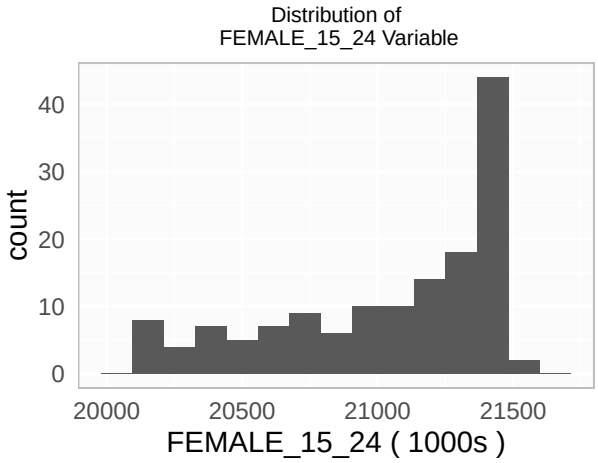


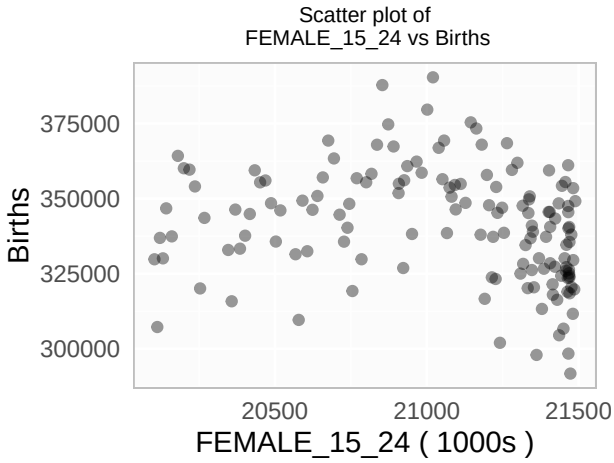
1.6 Variable FEMALE_15_24

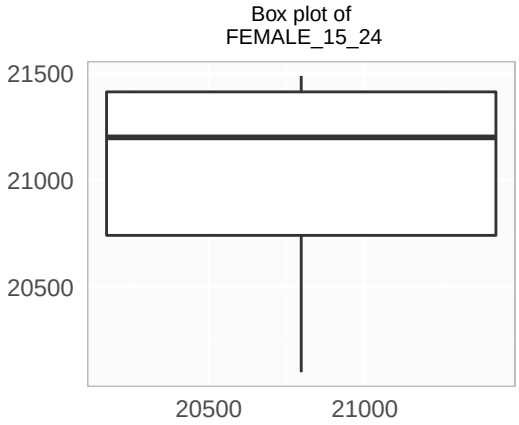
The *FEMALE_15_24* variable is the total population of females ages 15-24 per month as estimated by the Census Bureau.

Table 5: FEMALE_15_24 Variable Statistics

min	mean	stdev	median	max
20103.14	21046.7	422.1778	21201.43	21489.1





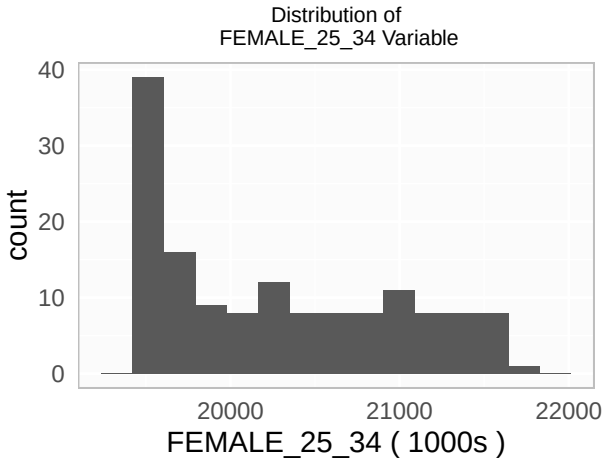


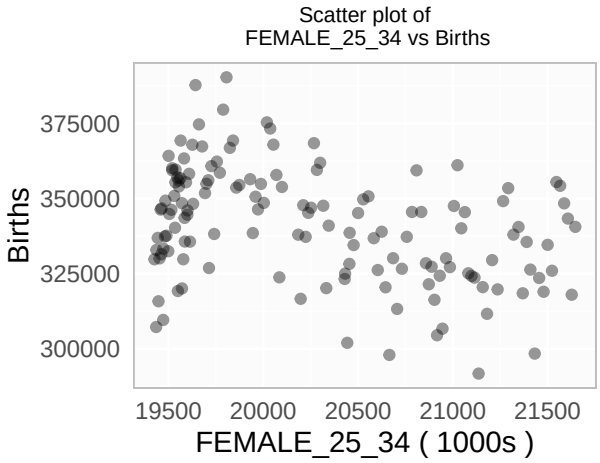
1.7 Variable FEMALE_25_34

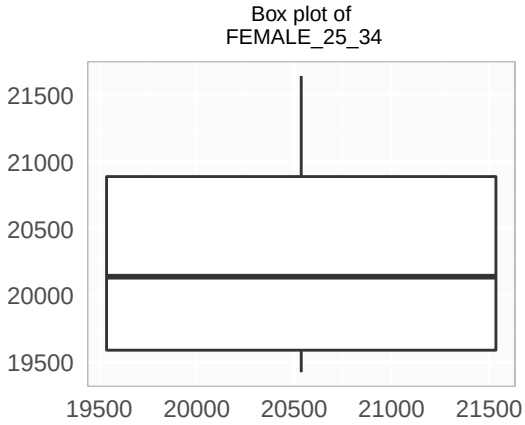
The *FEMALE_25_34* variable is the total population of females ages 25-34 per month as estimated by the Census Bureau.

Table 6: FEMALE_25_34 Variable Statistics

min	mean	stdev	median	max
19426.37	20274.31	701.1676	20141.73	21646.13





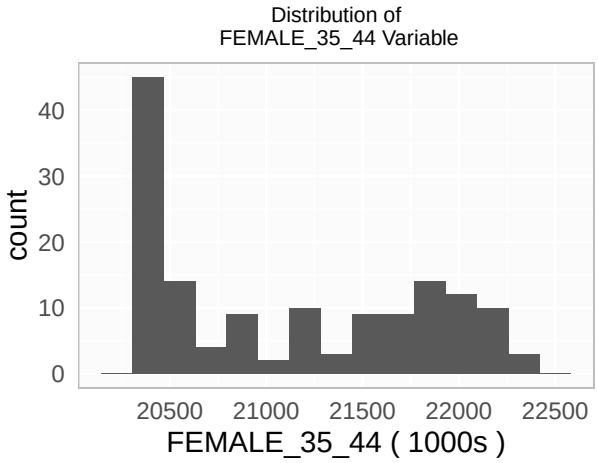


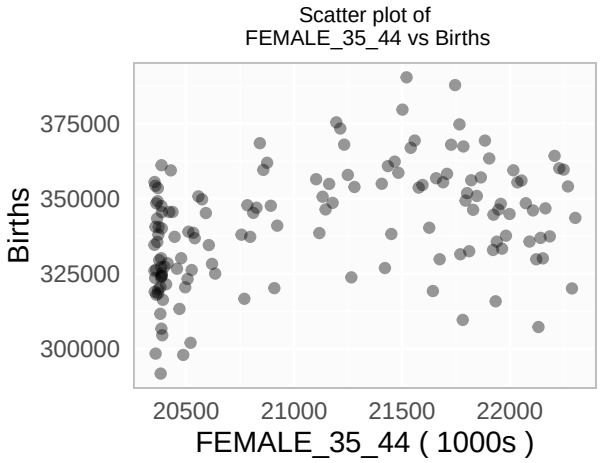
1.8 Variable FEMALE_35_44

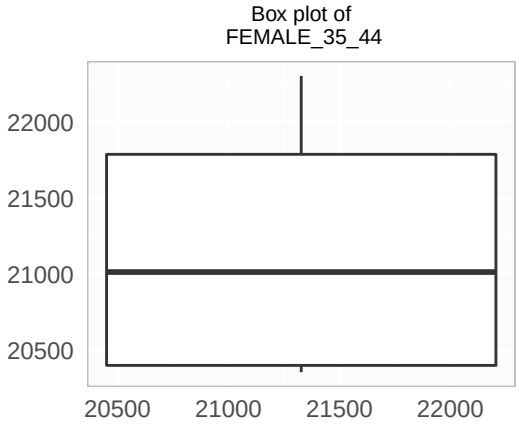
The *FEMALE_35_44* variable is the total population of females ages 35-44 per month as estimated by the Census Bureau.

Table 7: FEMALE_35_44 Variable Statistics

min	mean	stdev	median	max
20353.37	21120.04	683.5963	21012.17	22302.87





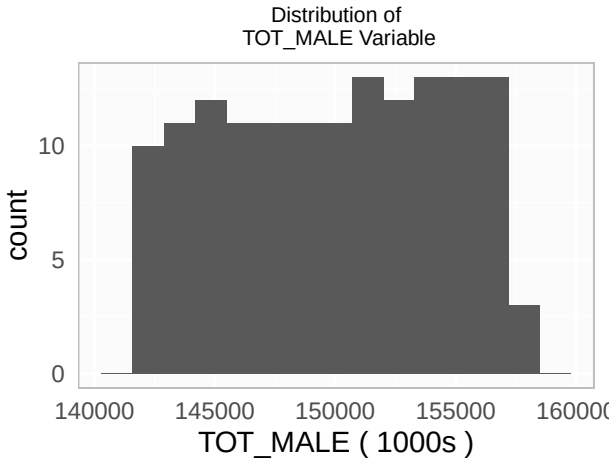


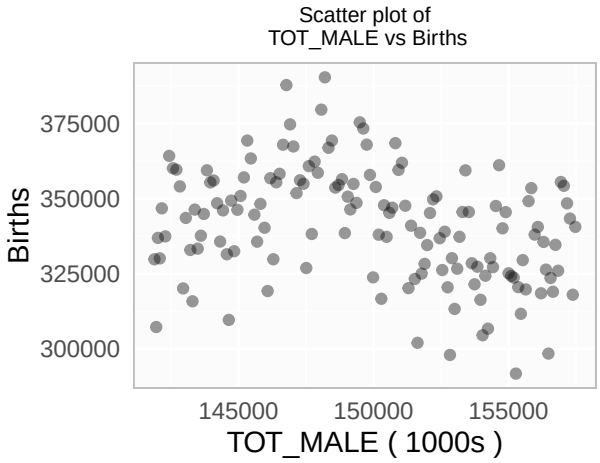
1.9 Variable TOT_MALE

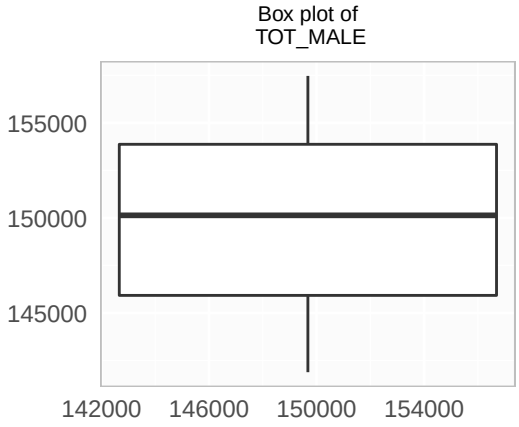
The *TOT_MALE* variable is the total population of females per month as esimated by the Census Bureau.

Table 8: TOT_MALE Variable Statistics

min	mean	stdev	median	max
141884.4	149888.3	4610.232	150137.2	157472.9





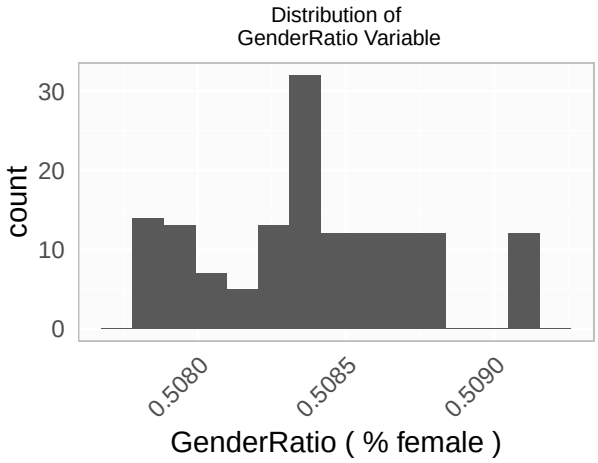


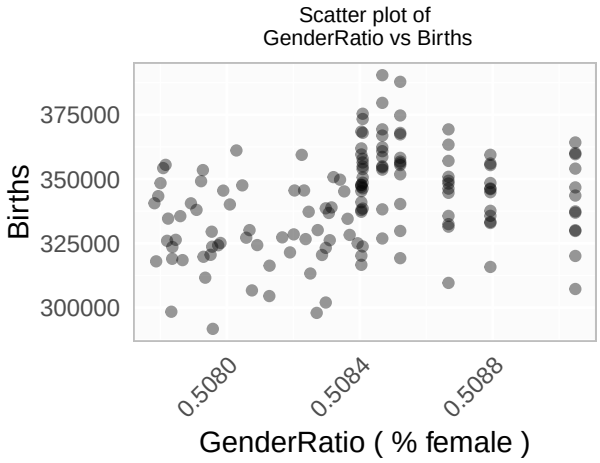
1.10 Variable GenderRatio

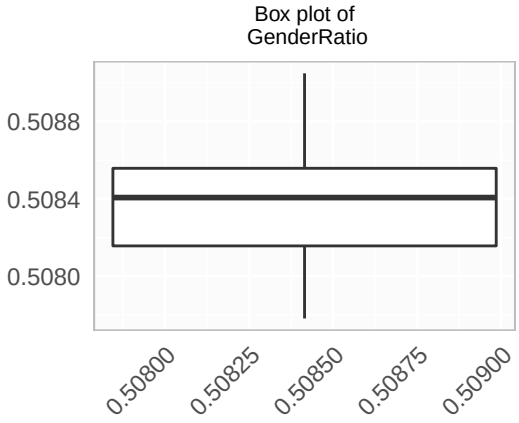
The *GenderRatio* variable is the percentage of the total population which are females per month derived from data from the Census Bureau. In cases where month data was not available, the annual gender ratio was computed and applied to the monthly total population.

Table 9: GenderRatio Variable Statistics

min	mean	stdev	median	max
0.507782	0.5083882	0.0003426	0.5084067	0.5090486





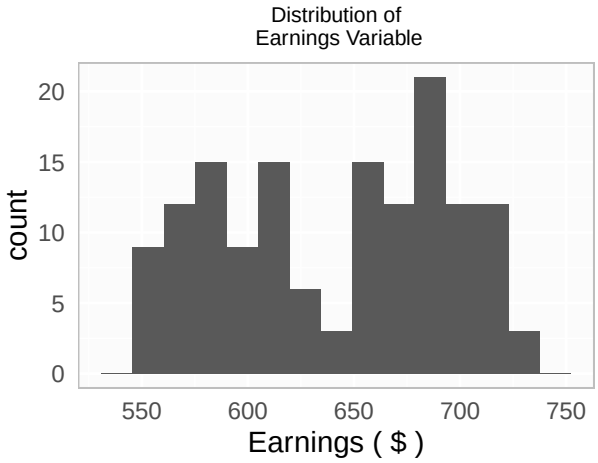


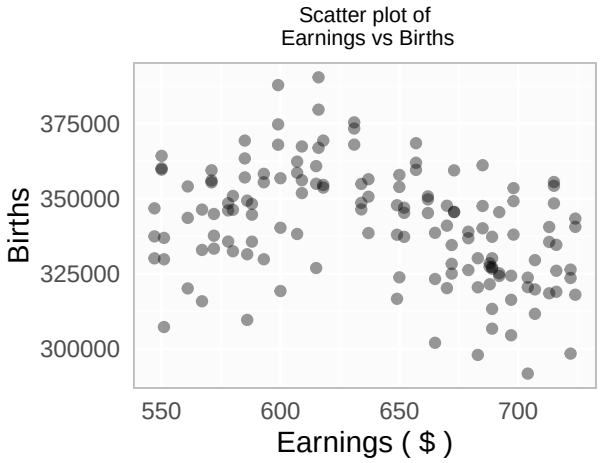
1.11 Variable Earnings

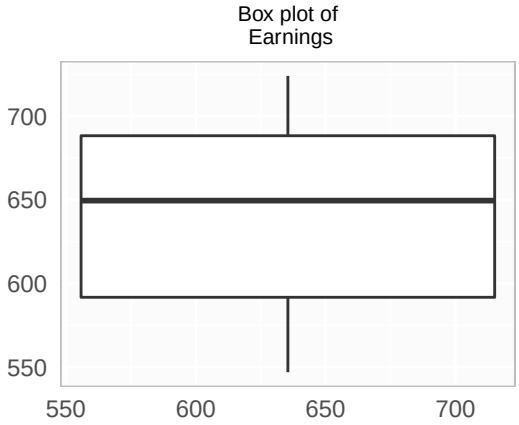
The *Earnings* variable is womoen's weekly earnings in current dollars based on data from the Bureau of Labor Statistics. The original values were provided quarterly and were expanded to a monthly format for data analysis purposes.

Table 10: Earnings Variable Statistics

min	mean	stdev	median	max
547	640.5417	53.55213	649.5	724





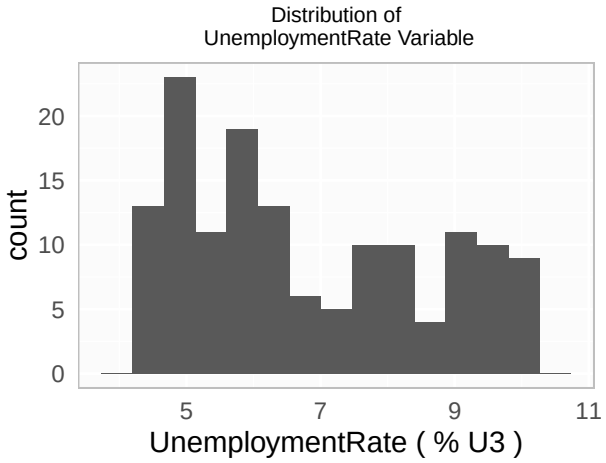


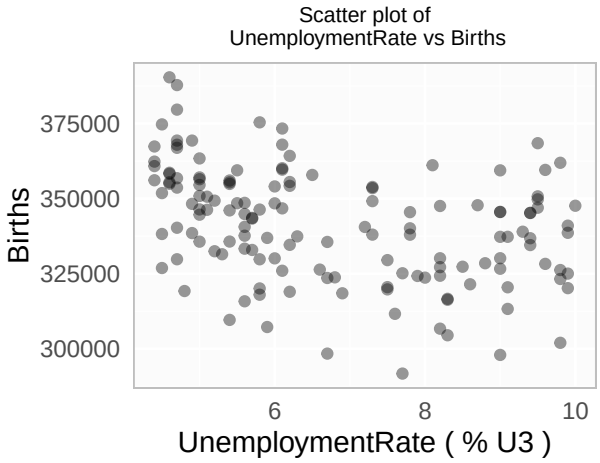
1.12 Variable UnemploymentRate

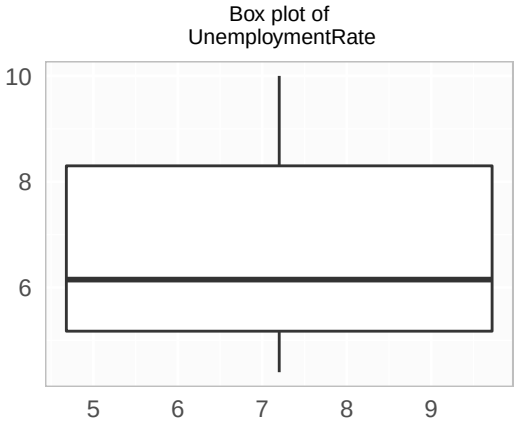
The *UnemploymentRate* variable is the unemployment rate per month (U3) based on data from the Bureau of Labor Statistics.

Table 11: UnemploymentRate Variable Statistics

min	mean	stdev	median	max
4.4	6.756944	1.789466	6.15	10







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 0.61796.

Table 12: All Variables Linear Model Coefficient Estimates

	Estimate	Pr(>|t|)
Intercept	6253281.6293451	0.9871368
Month *	2373.8746261	0.0001552
TOT_POP	0.1850169	0.9998835
GenderRatio	-12733902.1443100	0.9866951
TOT_FEMALE	21.1107054	0.9932498
FEMALE_15_24	-73.1850866	0.1858054
FEMALE_25_34 *	-79.9013043	0.0263951
FEMALE_35_44	22.7387984	0.3126103
Earnings	-202.6374817	0.5120129
UnemploymentRate	1431.3148764	0.6455496
Month9Ago *	2483.2833009	0.0000000

Table 13: All Variables Linear Model VIFs

Month	4.0547769
TOT_POP	153919112.9554873
GenderRatio	78993.0075166
TOT_FEMALE	147624728.1306689
TOT_MALE	73530.9824819
FEMALE_15_24	258.4907689
FEMALE_25_34	269.4435927
FEMALE_35_44	49816.9993515
Earnings	31383.0346822
UnemploymentRate	0.5125583
Month9Ago	4.2399970

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

	Estimate	Pr(>|t|)
UnemploymentRate	3023.19686	0.1026913
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 0.6296.

Table 18: Step Linear Model Coefficient Estimates

	Estimate	Pr(>|t|)
Intercept	60785.36420	0.9067120
Month *	2527.26252	0.0000000
TOT_FEMALE *	19.42614	0.0044077
FEMALE_15_24	-72.24008	0.0544331
FEMALE_25_34 *	-79.54210	0.0031896
FEMALE_35_44 *	17.38382	0.0492024
Month9Ago *	2673.92121	0.0000000

Table 19: Step Linear Model VIFs

Month	1.592251
TOT_FEMALE	1096.462376
FEMALE_15_24	292.403835
FEMALE_25_34	384.935528
FEMALE_35_44	41.377411
Month9Ago	1.035743

2.5 Significant Variables Minus Linear Model

The next model was aimed at removing variables with multicolinearity evidenced by the high VIFs we'd seen on earlier models. The adjusted R^2 value for this model is 0.38705.

Table 20: Significant Variables Minus Linear Model Coefficient Estimates

	Estimate	Pr(>|t|)
Intercept *	28468830.58995	0.0002023
Month *	2692.41457	0.0000000
GenderRatio *	-53512891.05265	0.0002689
FEMALE_25_34	-12.11996	0.1007071
FEMALE_35_44	-17.27610	0.0779181
Earnings *	-517.48100	0.0065042

Table 21: Significant Variables Minus Linear Model VIFs

Month	1.159557
GenderRatio	17.136699
FEMALE_25_34	17.925565
FEMALE_35_44	30.852861
Earnings	70.693120

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 0.59182.

Table 22: Significant Variables Limited Linear Model Coefficient Estimates

	Estimate	Pr(>|t|)
Intercept *	466234.449673	0.0000000
Month *	2737.050149	0.0000000
Month9Ago *	2625.253130	0.0000000
FEMALE_25_34 *	-7.509284	0.0000120
UnemploymentRate *	-1507.503984	0.0318537

Table 23: Significant Variables Limited Linear Model VIFs

Month	1.034251
Month9Ago	1.025088
FEMALE_25_34	1.346553
UnemploymentRate	1.339489

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 24: Linear Model Validation Error Results

Model	VS Error	Adj R^2	Variables	VIF
Step	216248955	0.6295984	6	BAD
Significant	227183351	0.4783943	7	BAD
All Variables	240910313	0.6179553	11	BAD
High Cor	278969733	0.4941529	6	BAD
Significant Limited	313706164	0.5918160	4	OK
Significant Minus	318353505	0.3870452	5	BAD

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.