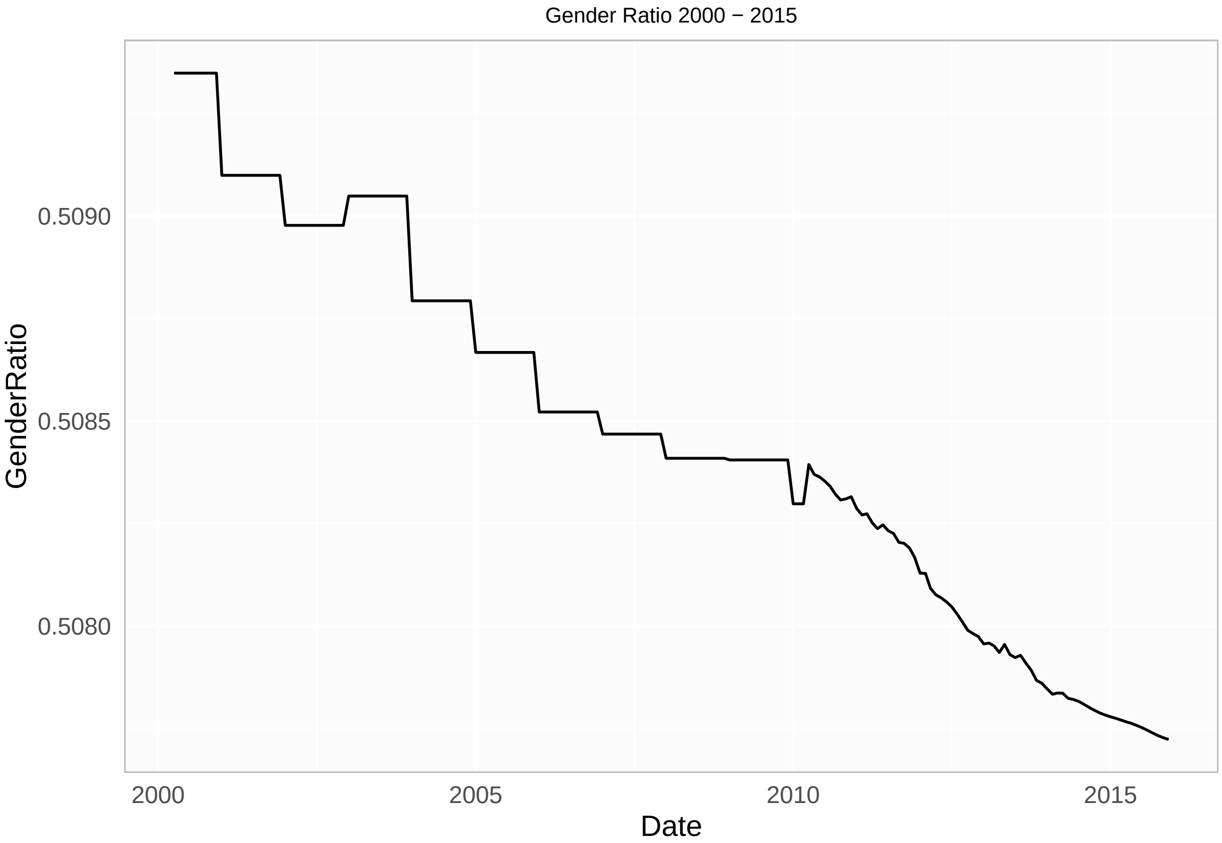
Natality Models Data Exploration

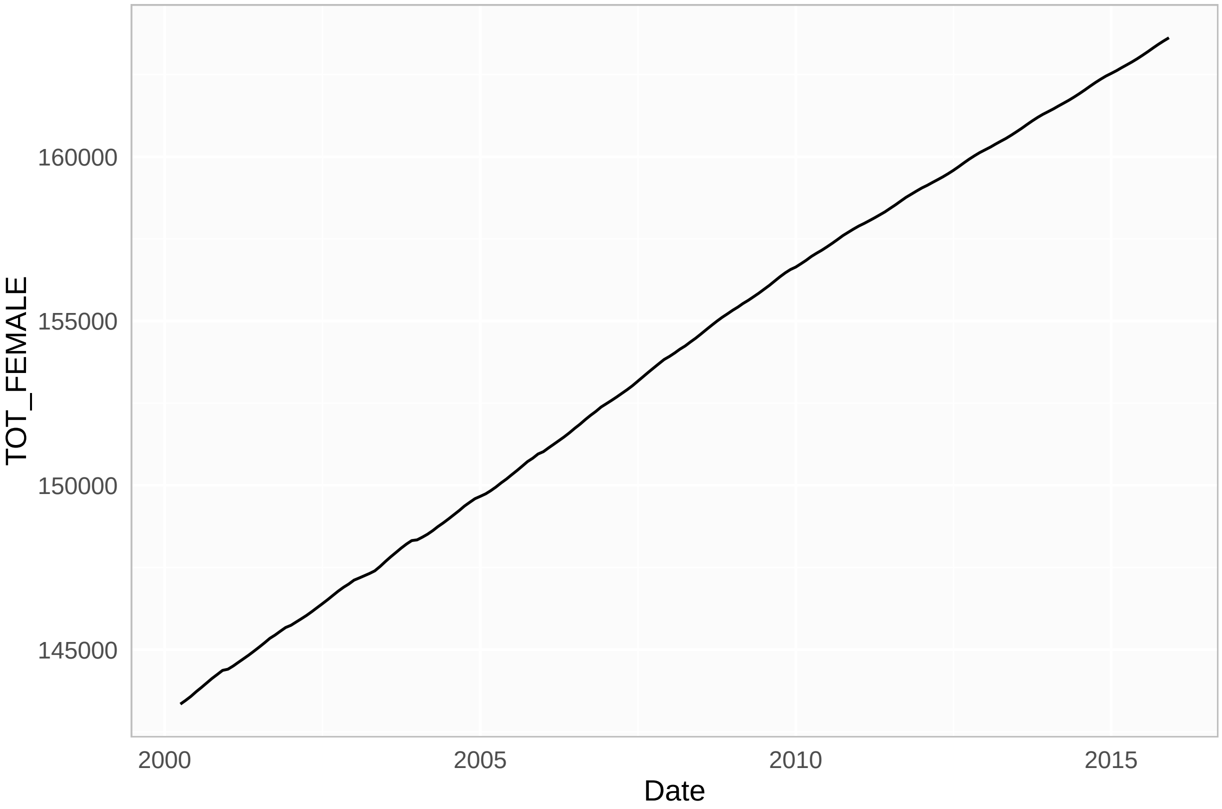
DATA 621: Business Analytics and Data Mining

Daniel Dittenhafer & Justin Hink

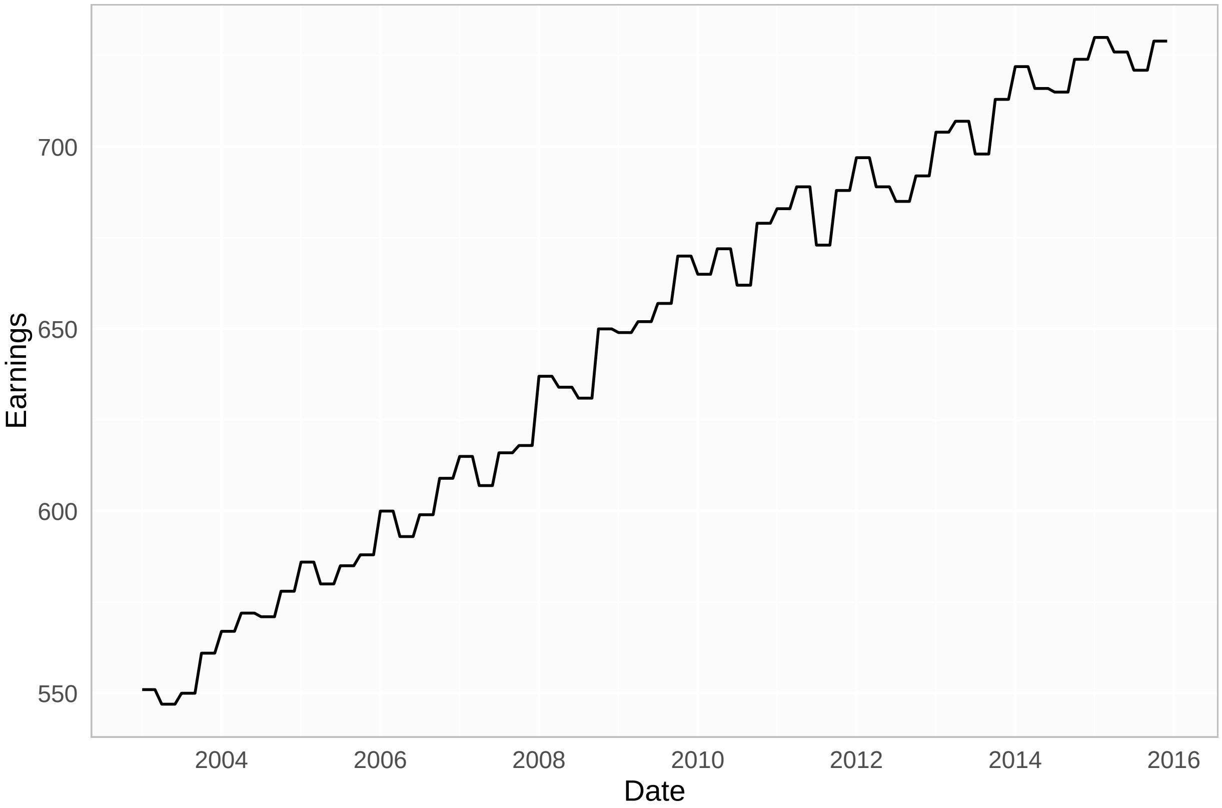
April 24, 2016

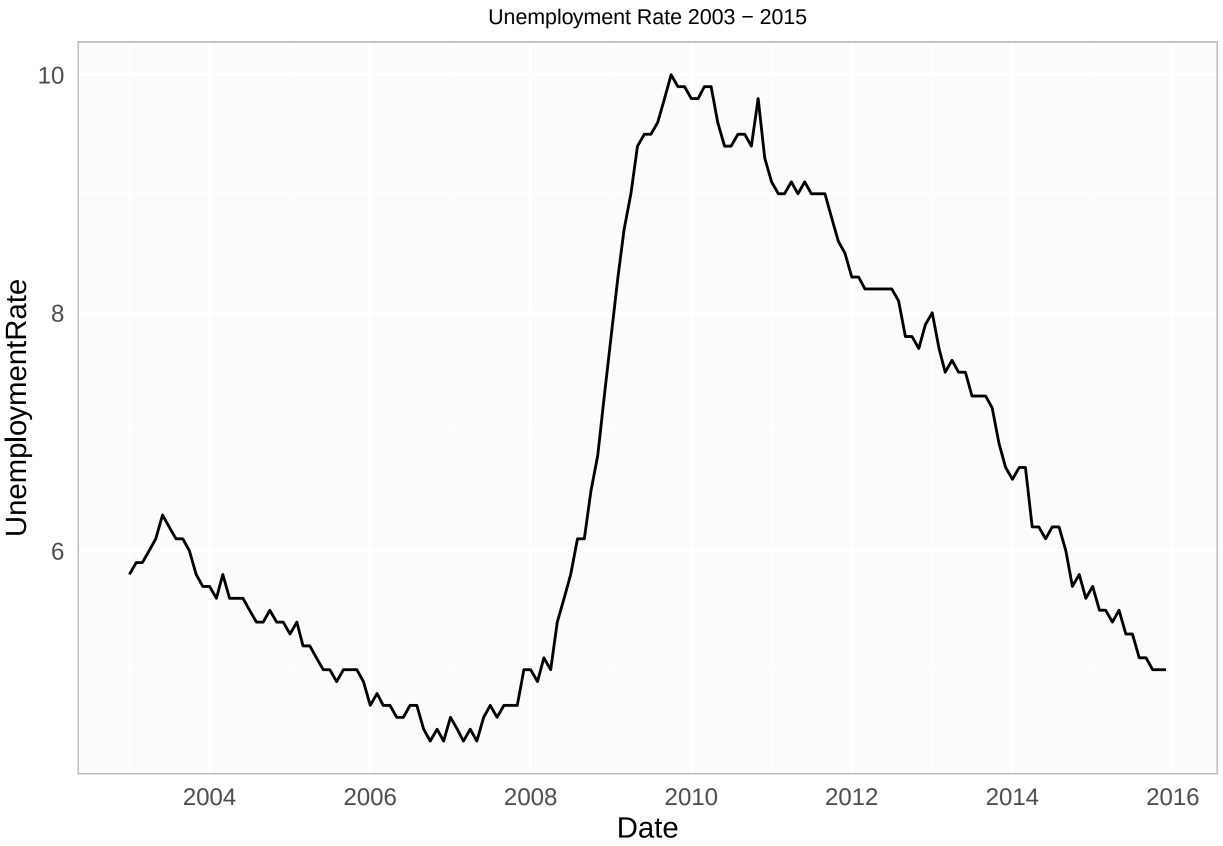


Female Population 2000 − 2015

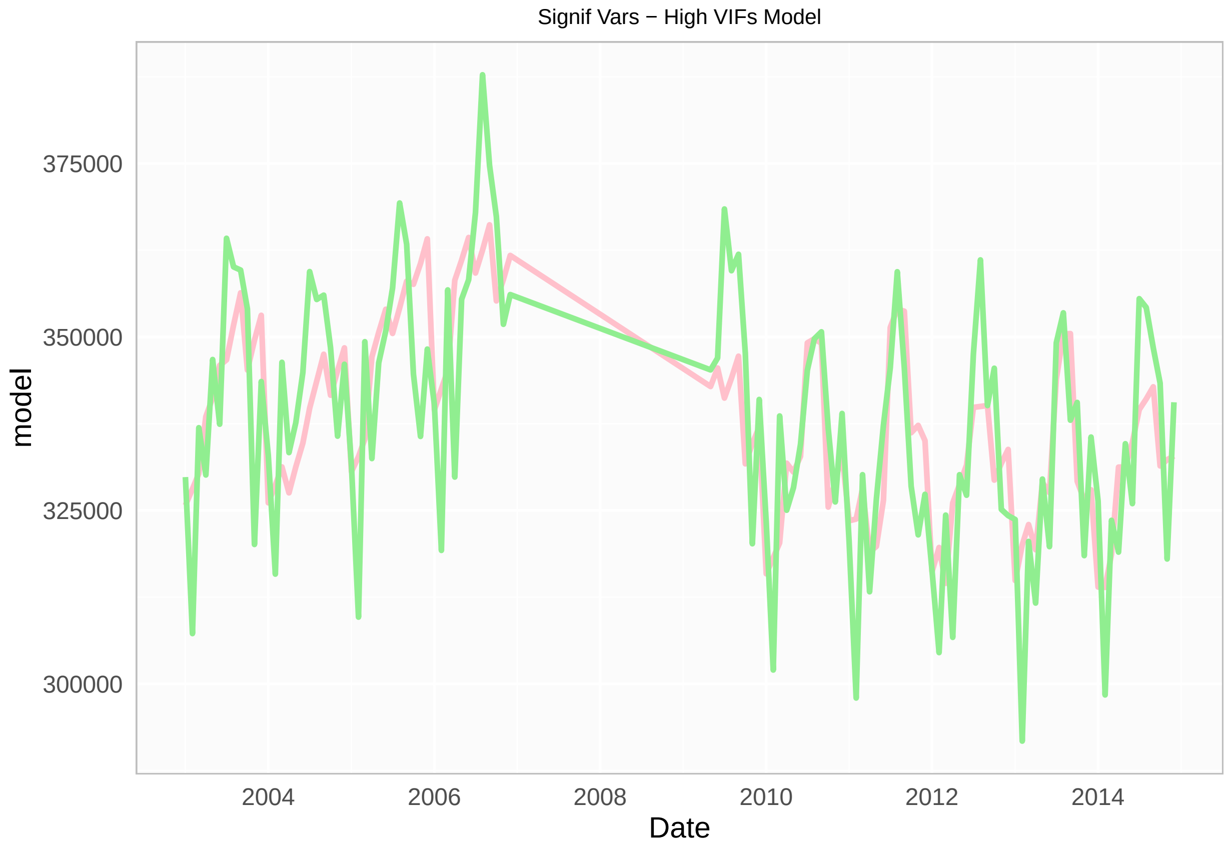


Women's Weekly Earnings 2003 − 2015





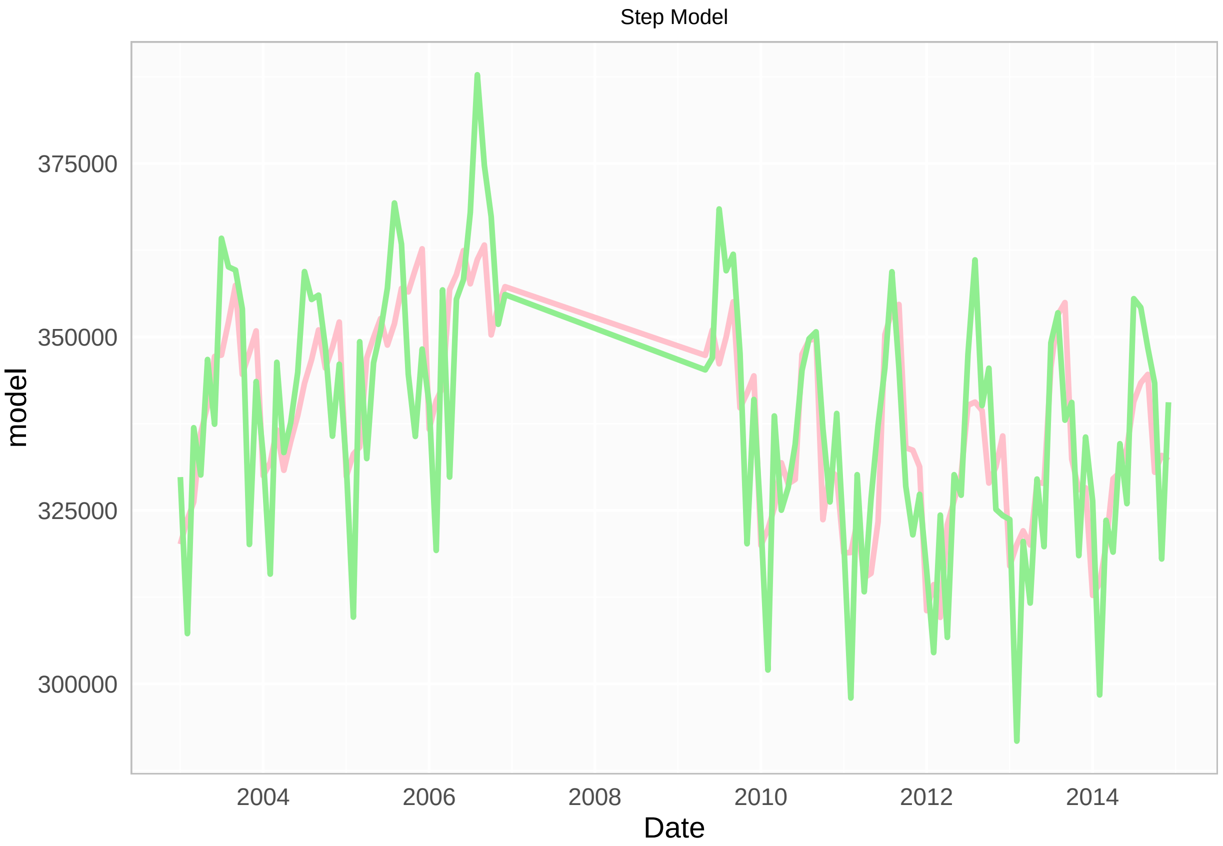
```
##       Year          Month           Births      
##  Min.   :2003   Min.   : 1.00   Min.   :291748  
##  1st Qu.:2006   1st Qu.: 3.75   1st Qu.:327115  
##  Median :2008   Median : 6.50   Median :342176  
##  Mean   :2008   Mean   : 6.50   Mean   :341157  
##  3rd Qu.:2011   3rd Qu.: 9.25   3rd 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.   :20180   Min.   :19501  
##  1st Qu.:151007   1st Qu.:145925   1st Qu.:20791   1st Qu.:19607  
##  Median :155272   Median :150137   Median :21204   Median :20143  
##  Mean   :154997   Mean   :149888   Mean   :21051   Mean   :20279  
##  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 :21019   Median :649.5   Median : 6.150  
##  Mean   :21125   Mean   :640.5   Mean   : 6.757  
##  3rd Qu.:21762   3rd Qu.:688.2   3rd Qu.: 8.300  
##  Max.   :22207   Max.   :724.0   Max.   :10.000  
```



```
##
## Call:
## lm(formula = Births ~ Month + GenderRatio + FEMALE_25_34 + FEMALE_35_44 +
##     Earnings, data = modelData)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -28450  -8431   1826   9569  32900
##
## Coefficients:
##                   Estimate    Std. Error t value      Pr(>|t|)
## (Intercept)   31747268.966   7732672.010   4.106 0.00007773342 ***
## Month             2479.362       382.665   6.479 0.00000000268 ***
## GenderRatio  -59536698.713  14776694.547  -4.029      0.000103 ***
## FEMALE_25_34       -12.774         7.131  -1.791      0.075999 .
## FEMALE_35_44       -23.507        10.722  -2.192      0.030450 *
## Earnings          -626.129       200.041  -3.130      0.002239 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13910 on 110 degrees of freedom
## Multiple R-squared:  0.4228, Adjusted R-squared:  0.3966
## F-statistic: 16.11 on 5 and 110 DF,  p-value: 0.000000000006754
```

```
## Start:  AIC=2191.88
## 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=2191.88
## 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   51038653 15747246625 2190.3
## <none>                           15696207973 2191.9
## - FEMALE_15_24      1  442328301 16138536274 2193.1
## - UnemploymentRate  1  903214824 16599422796 2196.4
## - FEMALE_25_34      1 1207762257 16903970230 2198.5
## - GenderRatio       1 1371939085 17068147058 2199.6
## - TOT_POP           1 1421315009 17117522982 2199.9
## - TOT_FEMALE        1 1426898973 17123106946 2200.0
## - FEMALE_35_44      1 2692118045 18388326017 2208.2
## - Earnings          1 5449575587 21145783559 2224.4
##
## Step:  AIC=2190.26
## Births ~ TOT_POP + GenderRatio + TOT_FEMALE + FEMALE_15_24 +
##     FEMALE_25_34 + FEMALE_35_44 + Earnings + UnemploymentRate
##
##                    Df  Sum of Sq         RSS    AIC
## <none>                           15747246625 2190.3
## - FEMALE_15_24      1  391851471 16139098096 2191.1
## - UnemploymentRate  1 1124642314 16871888939 2196.3
## - FEMALE_25_34      1 1279051719 17026298344 2197.3
## - GenderRatio       1 1849895239 17597141864 2201.1
## - TOT_POP           1 1910392854 17657639479 2201.5
## - TOT_FEMALE        1 1920753522 17668000148 2201.6
## - FEMALE_35_44      1 3226913215 18974159840 2209.9
## - Earnings          1 7853387188 23600633813 2235.2
```



```
##
## Call:
## lm(formula = Births ~ TOT_POP + GenderRatio + TOT_FEMALE + FEMALE_15_24 +
##     FEMALE_25_34 + FEMALE_35_44 + Earnings + UnemploymentRate,
##     data = modelData)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -28330  -7477   1600   8264  26637
##
## Coefficients:
##                         Estimate     Std. Error t value        Pr(>|t|)
## (Intercept)       1686532426.77   477772862.33   3.530        0.000614 ***
## TOT_POP                -5786.65        1606.11  -3.603        0.000479 ***
## GenderRatio      -3344339133.96   943294211.09  -3.545        0.000583 ***
## TOT_FEMALE              11406.78        3157.46   3.613        0.000463 ***
## FEMALE_15_24               97.83          59.95   1.632        0.105675
## FEMALE_25_34              187.08          63.46   2.948        0.003927 **
## FEMALE_35_44              253.74          54.19   4.683 0.0000083561470 ***
## Earnings                -1556.51         213.08  -7.305 0.0000000000521 ***
## UnemploymentRate         8928.11        3229.70   2.764        0.006717 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12130 on 107 degrees of freedom
## Multiple R-squared:  0.5732, Adjusted R-squared:  0.5413
## F-statistic: 17.96 on 8 and 107 DF,  p-value: < 0.00000000000000022
```

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

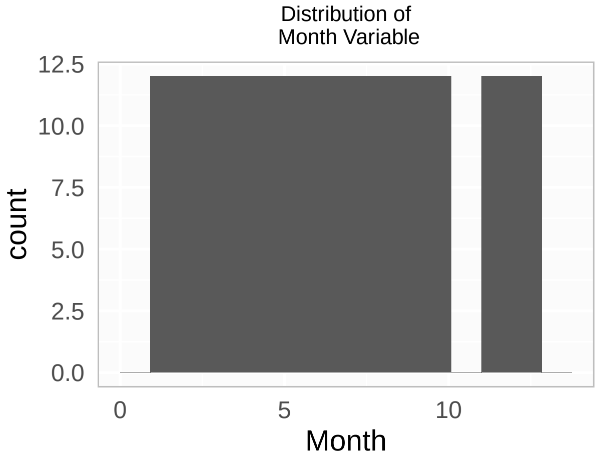
Births	1.0000000
FEMALE_35_44	0.3724631
Month	0.3646307
GenderRatio	0.2862173
FEMALE_15_24	-0.2572348
TOT_MALE	-0.3214851
TOT_POP	-0.3219328
TOT_FEMALE	-0.3223760
Year	-0.3593053
Earnings	-0.3697992
UnemploymentRate	-0.3862666
FEMALE_25_34	-0.4037382

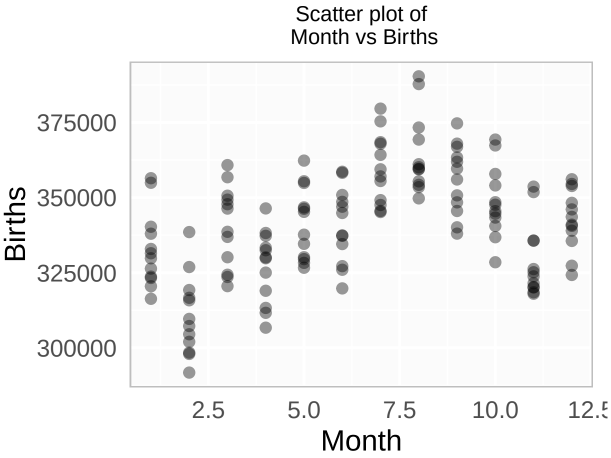
1.3 Variable Month

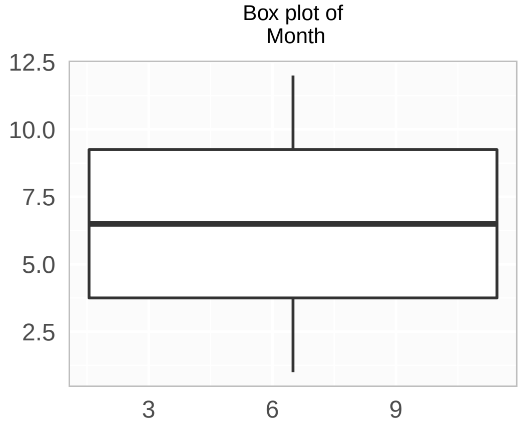
The *Month* variable is the month of birth. As one should expect, the distribution is uniform, but we can see some seasonality to the relationship between *Births* and *Month* with July and August being high frequency birth months.

Table 2: Month Variable Statistics

min	mean	stdev	median	max
1	6.5	3.464102	6.5	12





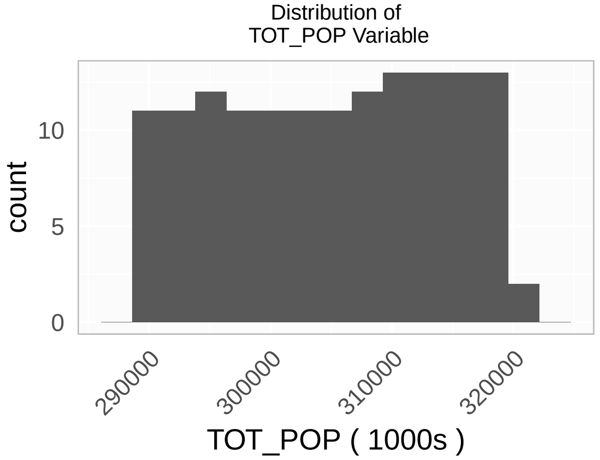


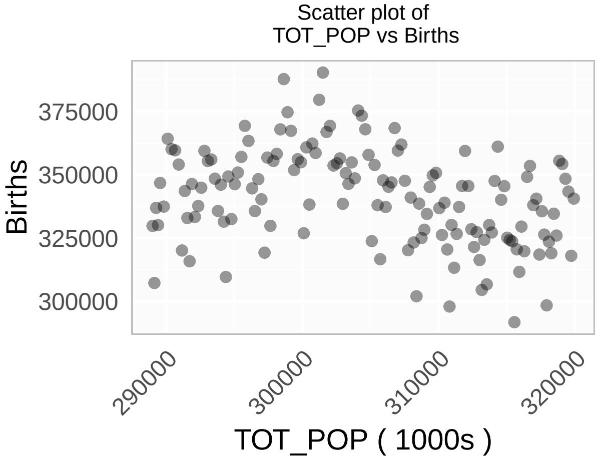
1.4 Variable TOT_POP

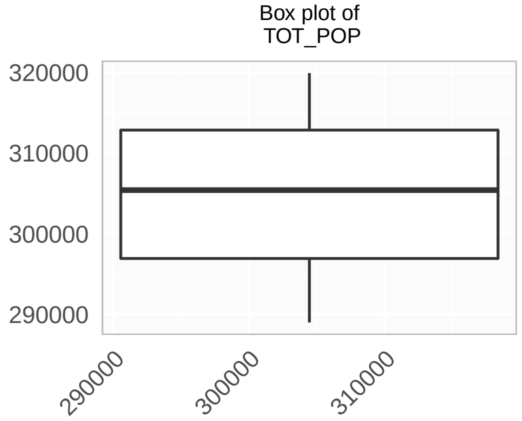
The *TOT_POP* variable is the total population per month as esimated by the Census Bureau.

Table 3: TOT_POP Variable Statistics

min	mean	stdev	median	max
288998.8	304885.4	9171.506	305409.3	319925.2





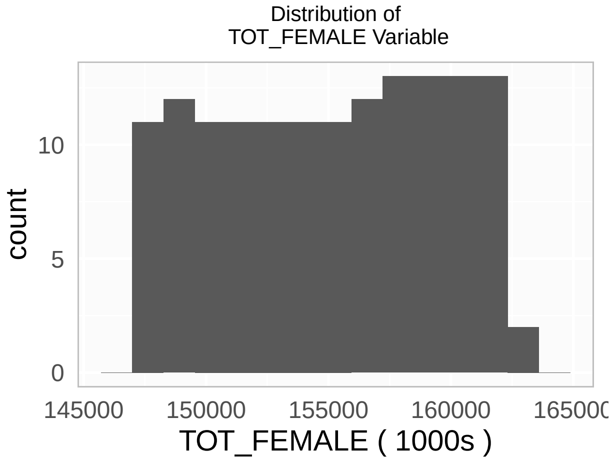


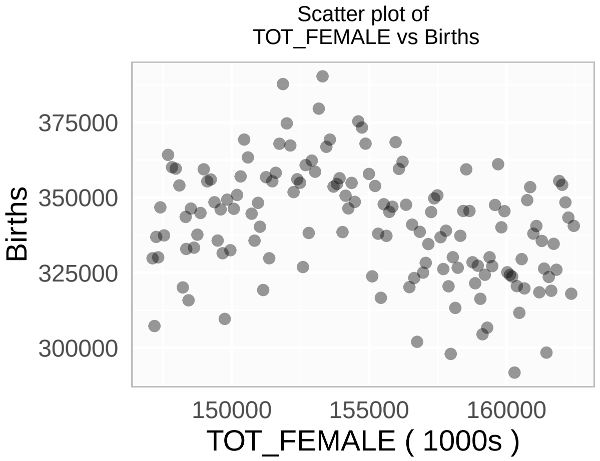
1.5 Variable TOT_FEMALE

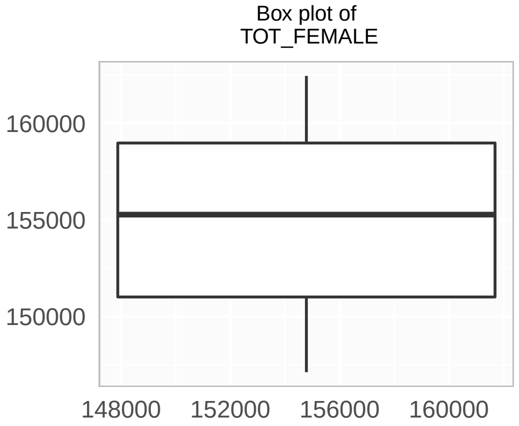
The *TOT_FEMALE* variable is the total population of females per month as estimated by the Census Bureau.

Table 4: TOT_FEMALE Variable Statistics

min	mean	stdev	median	max
147114.4	154997.1	4561.405	155272.1	162452.2





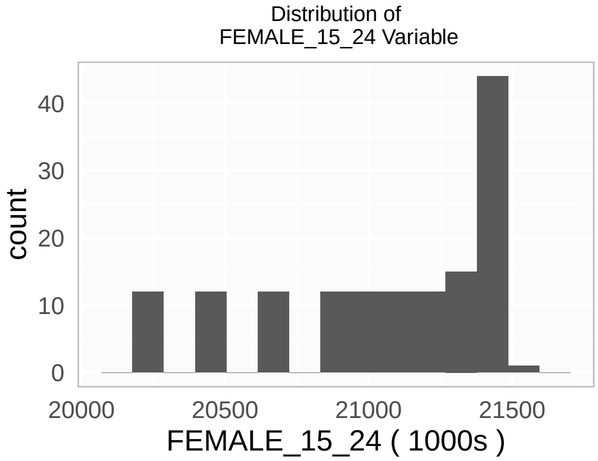


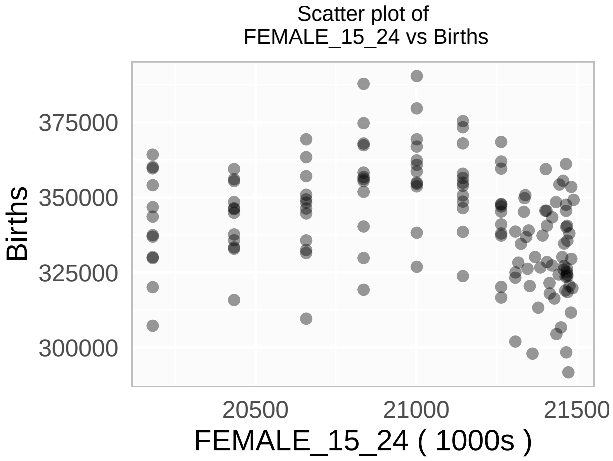
1.6 Variable FEMALE_15_24

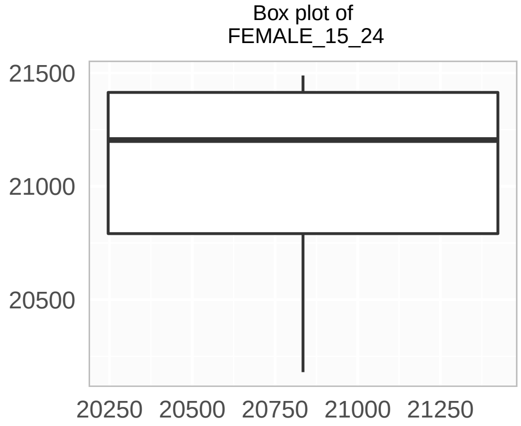
The *FEMALE_15_24* variable is the total population of females ages 15-24 per month as estimated by the Census Bureau.

Table 5: FEMALE_15_24 Variable Statistics

min	mean	stdev	median	max
20180.29	21051.25	418.9959	21204.35	21489.1





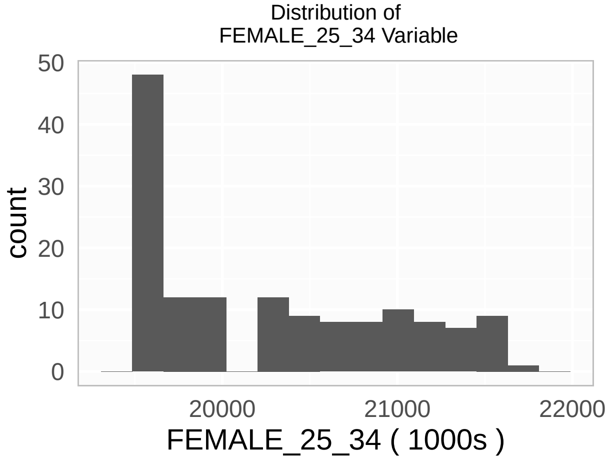


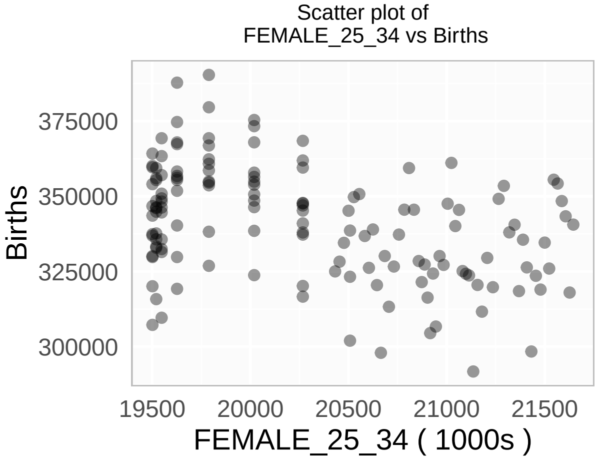
1.7 Variable FEMALE_25_34

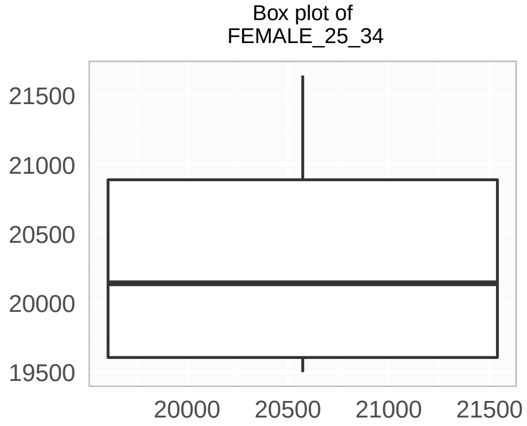
The *FEMALE_25_34* variable is the total population of females ages 25-34 per month as estimated by the Census Bureau.

Table 6: FEMALE_25_34 Variable Statistics

min	mean	stdev	median	max
19500.92	20278.64	698.1041	20143.42	21646.13





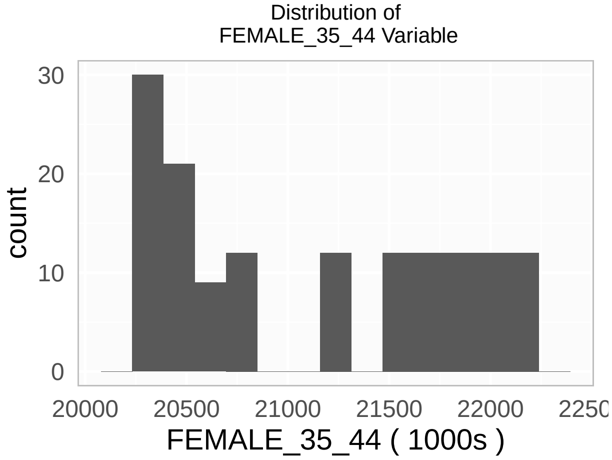


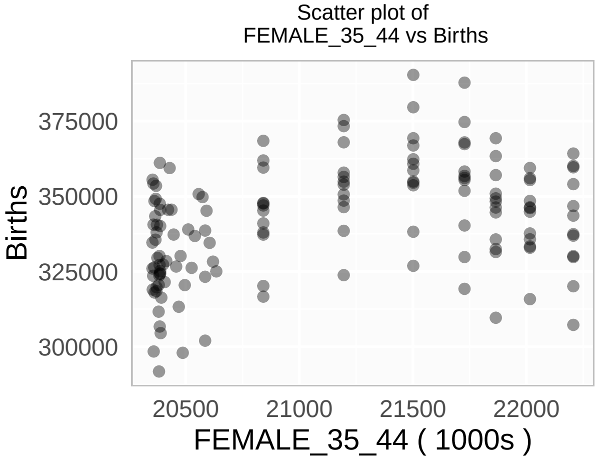
1.8 Variable FEMALE_35_44

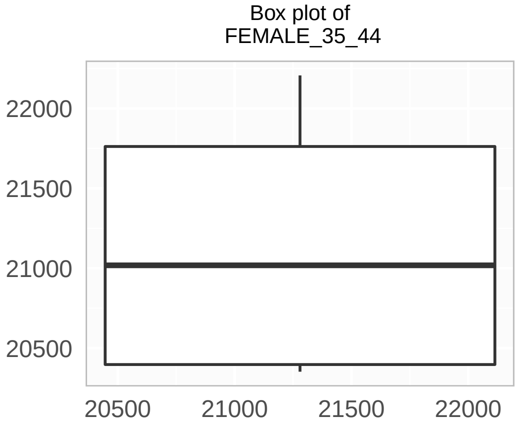
The *FEMALE_35_44* variable is the total population of females ages 35-44 per month as estimated by the Census Bureau.

Table 7: FEMALE_35_44 Variable Statistics

min	mean	stdev	median	max
20353.37	21124.66	683.2824	21018.67	22206.7





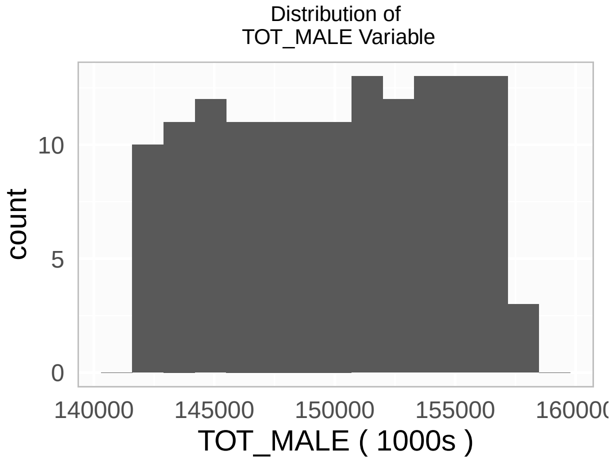


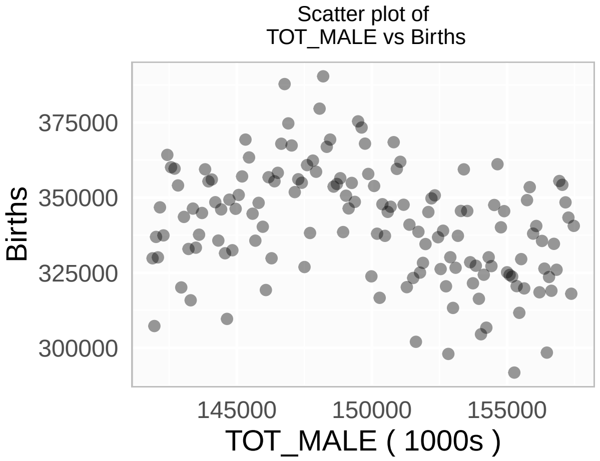
1.9 Variable TOT_MALE

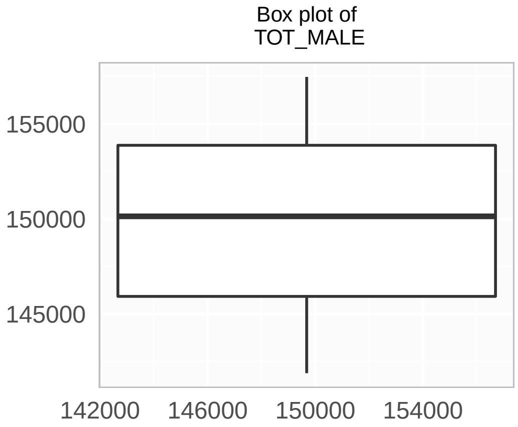
The *TOT_MALE* variable is the total population of females per month as esimated by the Census Bureau.

Table 8: TOT_MALE Variable Statistics

min	mean	stdev	median	max
141884.4	149888.3	4610.232	150137.2	157472.9





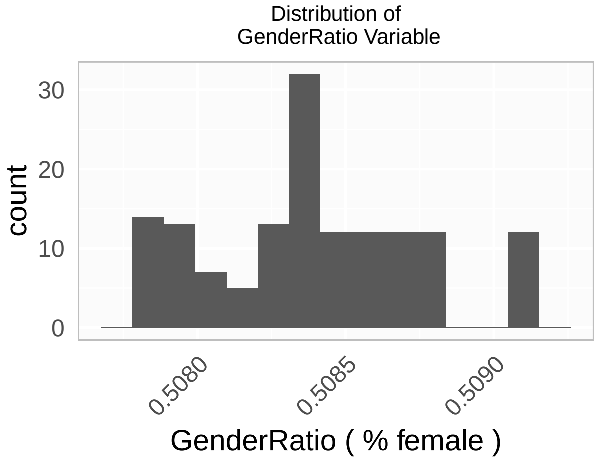


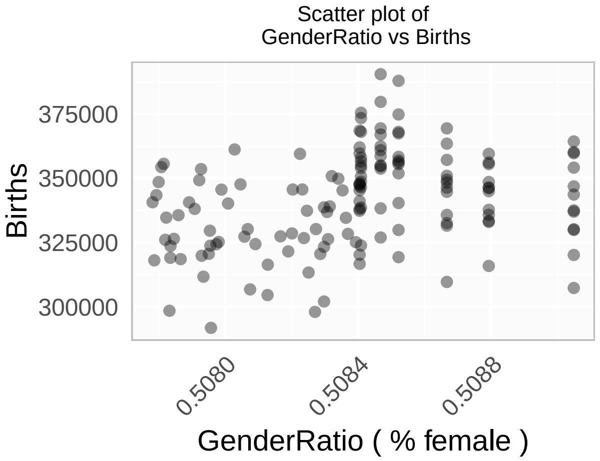
1.10 Variable GenderRatio

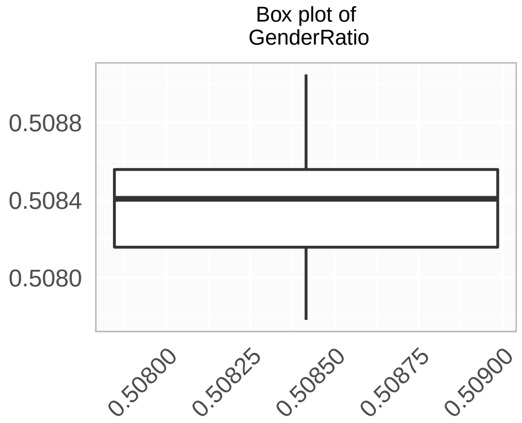
The *GenderRatio* variable is the percentage of the total population which are females per month derived from data from the Census Bureau. In cases where month data was not available, the annual gender ratio was computed and applied to the monthly total population.

Table 9: GenderRatio Variable Statistics

min	mean	stdev	median	max
0.507782	0.5083882	0.0003426	0.5084067	0.5090486





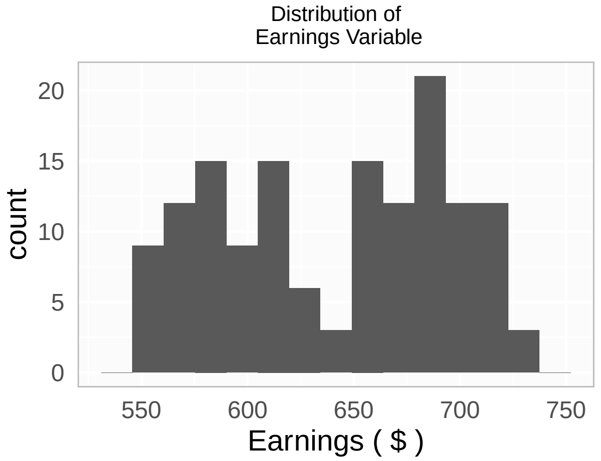


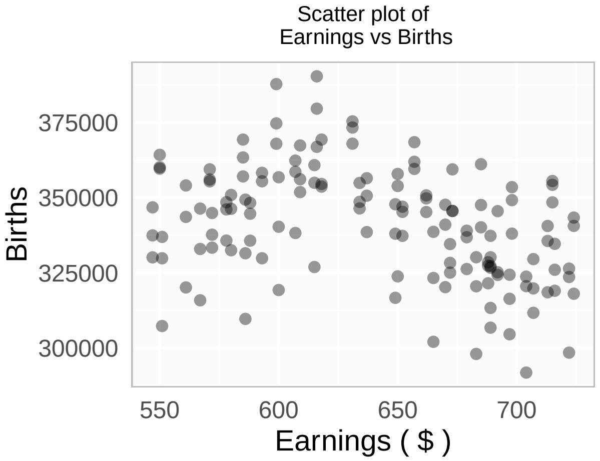
1.11 Variable Earnings

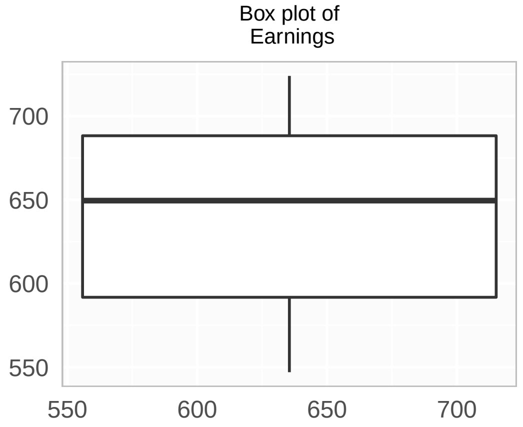
The *Earnings* variable is womoen's weekly earnings in current dollars based on data from the Bureau of Labor Statistics. The original values were provided quarterly and were expanded to a monthly format for data analysis purposes.

Table 10: Earnings Variable Statistics

min	mean	stdev	median	max
547	640.5417	53.55213	649.5	724





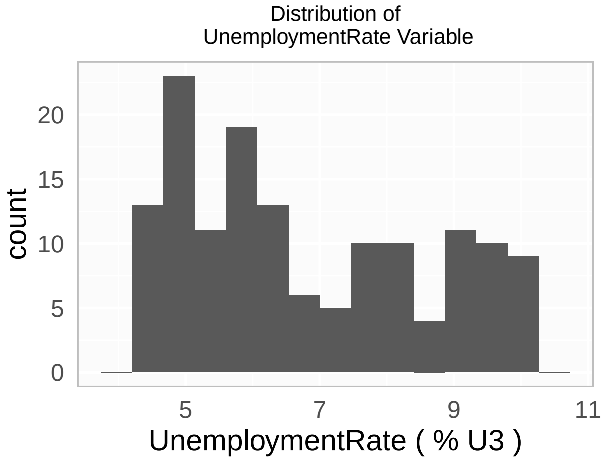


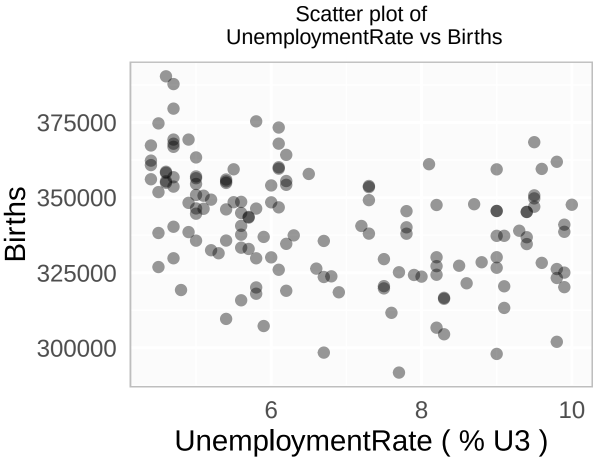
1.12 Variable UnemploymentRate

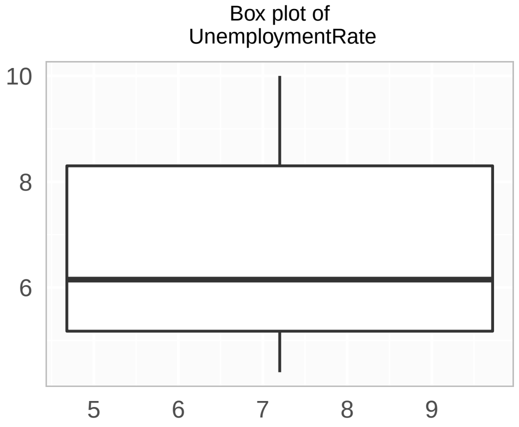
The *UnemploymentRate* variable is the unemployment rate per month (U3) based on data from the Bureau of Labor Statistics.

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 0.53847.

Table 12: All Variables Linear Model Coefficient Estimates

	Estimate	Pr(>|t|)
Intercept *	1570551806.7736	0.0030744
Month	332.2193	0.5583912
TOT_POP *	-5397.4043	0.0024933
GenderRatio *	-3115292895.6142	0.0029458
TOT_FEMALE *	10638.2549	0.0024468
FEMALE_15_24	109.1647	0.0868401
FEMALE_25_34 *	182.9208	0.0051644
FEMALE_35_44 *	243.4659	0.0000437
Earnings *	-1486.2078	0.0000000
UnemploymentRate *	8355.6395	0.0151169

Table 13: All Variables Linear Model VIFs

Month	2.9383515
TOT_POP	242043388.5994964
GenderRatio	118028.3612887
TOT_FEMALE	231599223.5213172
TOT_MALE	80414.6712485
FEMALE_15_24	686.3108235
FEMALE_25_34	1437.7536154
FEMALE_35_44	26150.2256700
Earnings	30883.8583972
UnemploymentRate	0.7886929

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 0.51309.

Table 14: Signficant Variables Linear Model Coefficient Estimates

	Estimate	Pr(>|t|)
Intercept *	1205109027.2476	0.0098101
TOT_POP *	-4291.5282	0.0069034
GenderRatio *	-2393295347.7861	0.0093938
TOT_FEMALE *	8465.0929	0.0067176
FEMALE_15_24	117.2511	0.0586028
FEMALE_25_34 *	141.8762	0.0267537
FEMALE_35_44 *	183.8783	0.0003143
Earnings *	-1427.7852	0.0000000

Table 15: Signficant Variables Linear Model VIFs

TOT_POP	183515878.4284
GenderRatio	87459.4194
TOT_FEMALE	175339848.1945
FEMALE_15_24	596.7110
FEMALE_25_34	1668.1899
FEMALE_35_44	1007.2717
Earnings	117.3696

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 0.49745.

Table 16: High Correlation Variables Linear Model Coefficient Estimates

	Estimate	Pr(>|t|)
Intercept *	-3603113.53644	0.0013897
FEMALE_25_34 *	-43.12873	0.0000021
UnemploymentRate	4423.89607	0.0553434

	Estimate	Pr(>|t|)
FEMALE_35_44 *	57.64305	0.0182079
Earnings *	-1382.82922	0.0000001
Month	626.73765	0.1887064
TOT_FEMALE *	28.71991	0.0000002

Table 17: High Correlation Variables Linear Model VIFs

FEMALE_25_34	29.937341
UnemploymentRate	11.799762
FEMALE_35_44	231.305667
Earnings	143.381588
Month	1.892072
TOT_FEMALE	484.853806

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 0.5413.

Table 18: Step Linear Model Coefficient Estimates

	Estimate	Pr(>|t|)
Intercept *	1686532426.76931	0.0006141
TOT_POP *	-5786.64719	0.0004791
GenderRatio *	-3344339133.95835	0.0005829
TOT_FEMALE *	11406.77777	0.0004633
FEMALE_15_24	97.82689	0.1056746
FEMALE_25_34 *	187.08443	0.0039269
FEMALE_35_44 *	253.73618	0.0000084
Earnings *	-1556.50623	0.0000000
UnemploymentRate *	8928.11161	0.0067175

Table 19: Step Linear Model VIFs

TOT_POP	206987914.65025
GenderRatio	100878.10698
TOT_FEMALE	197808003.35084
FEMALE_15_24	605.02199
FEMALE_25_34	1786.85517
FEMALE_35_44	1287.23116
Earnings	123.25596
UnemploymentRate	25.85011

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	206647888	0.5130879	7	TBD
All Variables	293016215	0.5384704	10	TBD
Step	303897019	0.5412971	8	TBD
High Cor	309970787	0.4974482	6	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.