STAT 8000
Death Master File

Hunter Perlman
Elliot Outland
Minjiao Yang
Mohammad Toutiaee

Submitted to Dr. Jaxk Reeves on May 8th, 2019

Contents

List of Figures

List of Tables

Summary

The purpose of this study is to provide a better understanding of the exact distribution of birthdays in the Social Security Administrations Death Master File. The data used in this analysis is divided into two subsets of approximately equal size, with the first containing the birthdays of people who died prior to 1916 and the second containing the birthdays of people who died after January 1st, 1916. Using this data, we fit a curve to the seasonal effect for each data-set which yields distinct patterns in the frequency of births over the year. In the pre-1916 data-set, the seasonal variation is substantial, with steep peaks and valleys. In the post-1916 data-set, the seasonal variation is much smaller. In both data-sets, the global maximum for birth frequency is in early-to-mid September, which is approximately nine months after the holiday season and matches closely with our expectations. Further, we have reason to believe that both data-sets contain a number of erroneously-reported birthdays, with the pre-1916 data having far more than the post-1916 data.

1 Introduction

Many people maintain the naive belief that humans have equal chance to be born on any given date within a year. This would give each births a probability of 1/365 of falling on a particular date, or 1/366 in a leap year. This analysis will address the naive hypothesis that birthdays should be uniformly distributed across the year by looking at the data-set collected in the Death Master File by the Social Security Administration. The data-set used in this project contains the recorded birthdates of the 89.7 million people in the death master file. This is presented as a frequency count for each of the 366 possible birthdays in a year. After adjusting for missing data (people with unknown birthdays), leap days (February 29th), and non-existent dates (e.g. November 31st) in the data-set, we find that the distribution of birth-dates over the remaining 365 days is not, in fact, uniform.

This difference from the uniform distribution can be explained in terms of seasonal and non-seasonal factors. In this report, the term seasonal factors is used to describe how the birth distribution varies naturally throughout the year, and roughly corresponds to a sinusoidal curve, likely caused by the holiday season and variations in temperature throughout the year. By contrast, the non-seasonal factors described in this report include all statistically significant deviations from the seasonal pattern. For example, some non-seasonal factors include incorrectly-reported birthdays and data entry errors. We seek, in this study, to use statistical reasoning to detect and explain the seasonal pattern as well as nonseasonal departures from the seasonal patterns. In this pursuit, we hope to answer the following questions:

(a) Is the distribution of birth dates before and after 1916 different, and in what ways ?

(b) Is there a seasonal effect are more babies born in certain seasons of the year ?

(c) Is there any evidence of digit transposition in dates (such as 12 instead of 21) ?

(d) Is there any evidence of fictitious birth dates being entered ?

(e) After correcting the data in any way that is appropriate, what is your estimate of the least likely day of the year on which to be born (excluding Feb. 29th)?

2 Data Summary and Exploratory Data Analysis

The data-set for this problem is a frequency table of all the recorded birth dates for the 89.7 million people listed in the Social Security Administrations Death Master File as of June 2011. Each persons reported birthdate is recorded as a combination of month and day. The data-set also contains a number of entries from people who report either an unknown birthday or an invalid birthday, such February 31. The overall data-set of 89.7 million birthdates is subdivided into two data-sets, one for people who were born before 1916, and another for those who were born on and after January 1, 1916. These subsets of the data are of approximately equal sample sizes, at around 45 million people for both sub-data-sets.

Table 1: Death Master Data Summary.

	Pre-1916	Post-1916
Total	45,607,021	44,181,194
Missing Data	148,732	15,122
Feb. 29th Increase	83,520	100,659
Adjusted Total	45,541,809	44,266,731
Adjusted Mean	124,431	120,947
Adjusted SD	12,152	5,180

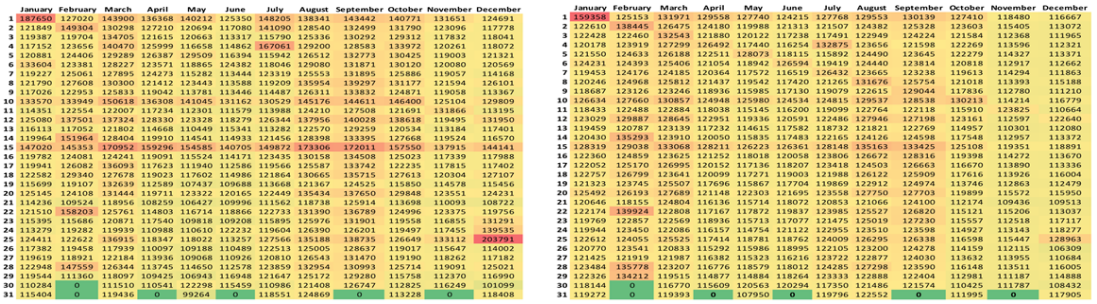

Figure 1: Heatmaps of Pre- and Post-1916 Data.

In the cleaning phase of the analysis, entries corresponding to an unknown birthday were removed. This caused a small loss of potential information, as individuals reported as knowing just their birth month or day of the month were excluded from the final analysis. These partially unknown birthdates corresponded to an extremely small fraction of the data, however, so their removal should be minimally impactful. Similarly, birth dates corresponding to six non-existent days of the year (Feb. 30, Feb. 31, Apr. 31, Jun. 31, Sep. 31, and Nov. 31) were removed. As with the partially unknown birth dates, these entries could have been placed within the self-reported month, for example treating February 30th as February 28th, but again, this problem was only observed in a small number of cases, so it was not necessary to attempt to salvage any theoretical information. Overall, the missing and invalid data removed from the data-set corresponds to only 0.33% of the pre-1916 data and 0.03% in the post 1916 data. It is not likely that this small amount of missing data affects the analysis of either data-set in any way.

Leap days present an unusual problem in this data-set, as they only occur about every four years. While leap days could have been simply removed from the data-set because they are not exactly comparable to the other days of the year, it was ultimately decided that the frequency value for February 29th is best analyzed by multiplying the observed frequency by four. This choice makes the leap day roughly comparable to the other days of the year.

After conducting these procedures, the two data-sets contain approximately the same number of birthday entries, at around 45 million. Perhaps the biggest difference between the two data-sets is the standard deviation of the birthdate frequency, as shown in Table 1. The pre-1916 data has approximately twice the standard deviation of the post-1916 data. We believe that this difference in standard deviation is related to the other major difference between the data-sets: the number of missing birthdays. The most plausible explanation of the higher standard deviation in the earlier data is that a large number of people report a birthday that is not the day on which they were actually born. It seems apparent that there is a high degree of bias in the incorrect reporting of birthdays, with certain days of the month and holidays appearing more often than would be expected.

The final version of the data used in the analysis is shown in the heatmaps in Figure 1. Even a cursory glance at the heatmap reveals certain patterns in the data, which we formally analyze later in this report. Further, as anticipated by the standard deviations, it is easy to observe that there are more obvious patterns in the pre-1916 data-set than in the post-1916 data-set. For example, there appears to be a much stronger bias towards holidays, such as January 1st and December 25th in the pre-1916 data-set. Similarly, the row-effects, such as the over-abundance of birthdays on the 15th of each month, appear much stronger in the pre-1916 data. We believe

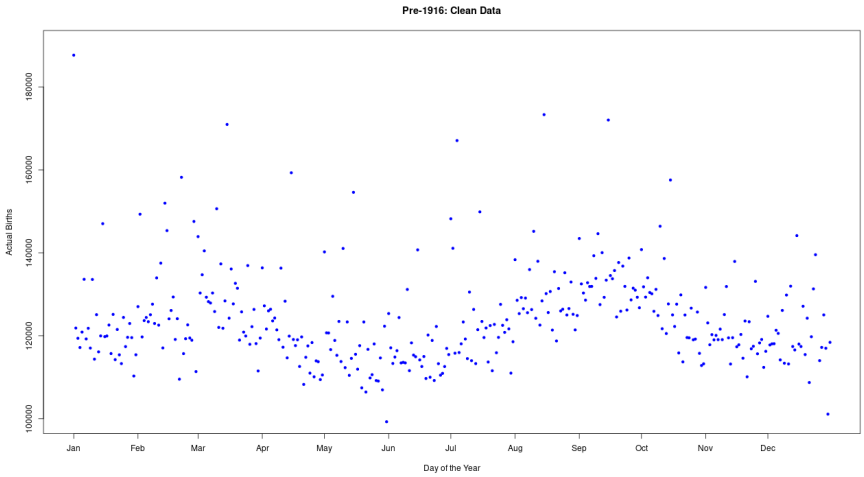

Figure 2: Clean Data for Pre-1916 Births.

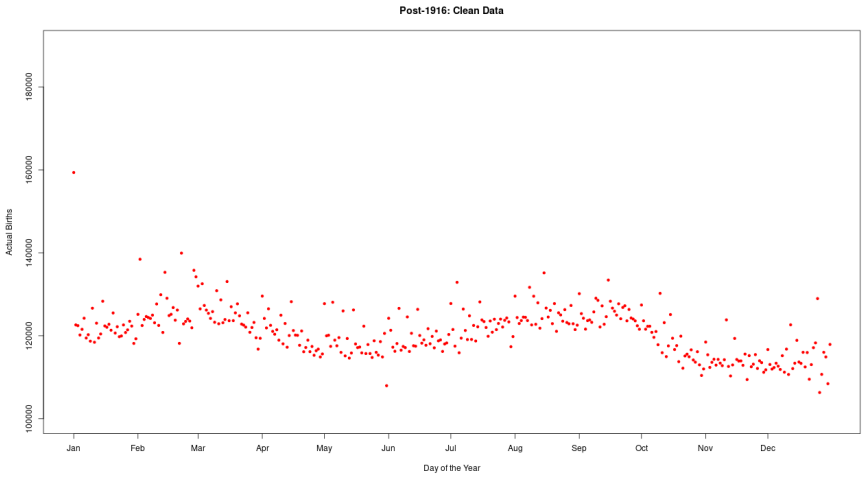

Figure 3: Clean Data for Post-1916 Births.

that these effects are caused by the preponderance of erroneously-reported birthdays in the pre-1916 data. This study makes the case that reporting accuracy is a significant problem in the Death Master File.

In addition to these non-seasonal patterns, even an untrained eye can pick up on a relationship between season and birthday frequency in the data. The plot of the data with respect to day of the year (1-366) is shown in Figure 2 and Figure 3. These plots also seem to demonstrate the differences in standard deviations in the data-sets, as the apparent seasonal trend is much clearer in the post-1916 data.

3 Analysis

As we seen have in the exploratory data analysis, birth trends appear to be subject to two classes of effects: seasonal and non-seasonal. It is well-known that birth patterns naturally vary across the year. This natural variation is described as the seasonal effect, and should be most strongly influenced by the environmental conditions of nine months prior. In this analysis, seasonal effects are analyzed by looking at the relationship between day of the year (1-366) and birth frequency.

We describe factors that cannot be accounted for by day of the year as non-seasonal effects. These non-seasonal effects include factors such as whether the birthday is a holiday such as

Christmas, or corresponds to an unlucky number such as the 13th of each month. We believe that these non-seasonal effects influence the observed frequency of births by affecting the probability that a person falsely reports a given day as their birthday. Further, for more recent births, we believe that non-seasonal effects may be influenced by the application of certain elective procedures, which for example may not be available on holidays.

In addition to the day of the month, we consider two types of days that warrant special attention: holidays and days with repeated digits (e.g., January 1st or February 2nd). In the following analysis, we treat holidays as a factor variable with levels for New Years Day, Valentines Day, Presidents Day, St. Patricks Day, Independence Day, Christmas, and New Years Eve. We also create an indicator variable for repeated digits, with a value of 1 if digits are repeated and 0 otherwise. Since the patterns in the data are sufficiently distinct between the pre-1916 data and post-1916, we conduct the succeeding analyses separately for each data-set.

3.1 Analysis of Seasonal Variation

In our statistical analysis, we begin by isolating the seasonal variation in the number of births. This step is crucial for later analysis, as it gives us the baseline for comparisons to actual observations. The statistical model for seasonal variation is created by first controlling for the outsized share of births on certain days of the month and holidays using an ANOVA model. Essentially, the model gives a perfect fit for holidays, such as Christmas, and then predicts each day of the year as the average number of births observed for that day of the month, for example the 15th. The exact equation for this model is shown in Equation (1).

$$\begin{gathered}\text{Births}_i = \beta_0 + \beta_1 \times I(\text{Day}_j) + \beta_2 \times I(\text{Holiday}_k) + \beta_3 \times I(\text{Repeat}) + \epsilon_i, \qquad (1)\\ \text{where} \quad i = 1, 2, \ldots, 366,\\ j = 1, 2, \ldots, 31,\\ k = \text{New Year's Day, Christmas Day, Independence Day, Valentine's Day,}\\ \text{New Year's Eve, Washington's Birthday and St. Patrick's Day.}\end{gathered}$$

Baseline for the proposed model is Jan. 1st , and we have 3 predictors in the model, Day_j with 31 levels, Holiday_k with 8 levels and Repeat is the indicator of repeating digits for the month and day $(01/01, 02/02, 03/03, \ldots,$ etc) with 2 levels.

Table 2: ANOVA Table for Pre-1916

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
factor(day)	30	25099923102.75	836664103.42	16.57	0.0000
factor(holiday)	7	11668077996.88	1666868285.27	33.02	0.0000
factor(rpt)	1	623184512.39	623184512.39	12.35	0.0005
Residuals	327	16506844900.14	50479648.01		

Table 3: ANOVA Table for Post-1916

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
factor(day)	30	2712469938.08	90415664.60	4.12	0.0000
factor(holiday)	7	1842406061.34	263200865.91	12.01	0.0000
factor(rpt)	1	597404444.03	597404444.03	27.25	0.0000
Residuals	327	7168752596.08	21922790.81		

where j = New Year's Day, Christmas Day, Independence Day, Valentine's Day, New Year's Eve, Washington's Birthday and St. Patrick's Day.
We next use the model to control for non-seasonal variations. Even without the running a regression, the residuals (with the mean added back to it) look much closer to a seasonal pattern than the original data, as it appears that much of the initial scattering was caused by the preponderance of erroneously-reported birthdays assigned to certain days of the month and

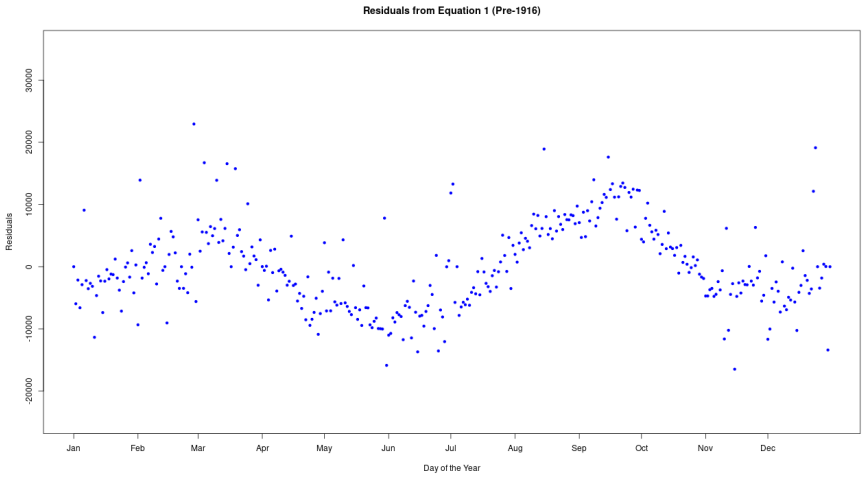

Figure 4: Residuals for Pre-1916 Births.

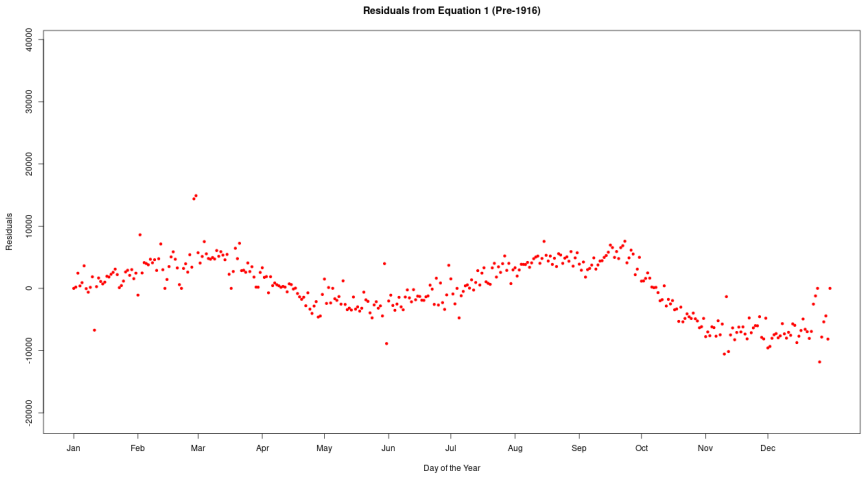

Figure 5: Residuals for Post-1916 Births.

holidays. This effect is much more pronounced in the pre-1916 data-set than the post-1916 data, as it appears that this data has many more inaccurate birthdays. The effect of this approach is shown in the plots in Figure 4 and Figure 5.

After adding the mean back to the residuals from the ANOVA model for scaling purposes, we next fit a smoothing spline to the residuals. Under the smoothing spline model, we fit a series of polynomial equations over segments of the data, with the constraints that the fitted curve and its first two derivatives be continuous throughout the data. Essentially, we require that the curve be smooth, without kinks or discontinuities. Additionally, in order to ensure that our curve accurately models seasonal effects, we apply the constraint that the curve between the first and last day of the year be continuous. We must also specify the number of knots (i.e., where the coefficients describing the curve can change); we find that the smoothing spline with 8 knots (at days 1, 53, 105, 157, 209, 313, and 366) yields the best fit for both data-sets, and have plotted the splines in Figure 6.

From the plot in Figure 6, we can see that the distribution of births over the year is bimodal in both data-sets, but that the peaks and troughs are more extreme in the pre-1916 data-set. This aligns with our expectations; the post-1916 data-set is more reliable from a data-collection perspective, and so should fit more closely to the seasonal trend. We also note that there are fewer overall births counted in the post-1916 data-set, and so it is sensible that the post-1916 curve is at most points below the pre-1916 curve. The peaks and valleys of both curves are described in Table 4 and the most extreme outliers are described in Table 6.

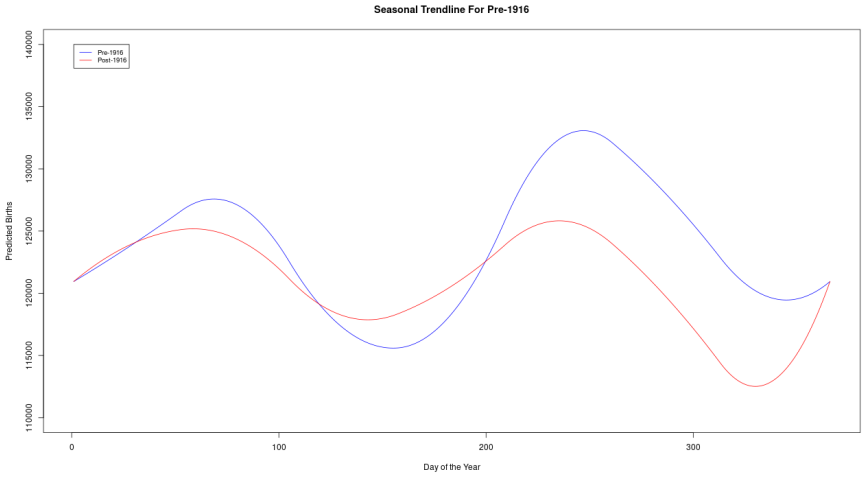

Figure 6: Spline Fits for Seasonal Trends in Pre- and Post-1916 Births.

Table 4: Peaks and Valleys in Seasonal Trends.

	Pre-1916	**Post-1916**
1st Peak	Mar. 1	Mar. 3
2nd Peak	Sep. 11	Sep. 3
1st Valley	May 27	May 16
2nd Valley	Nov. 23	Nov 26

Table 5: The Estimate of the Most and Least Likely Day of the Year on which to be Born.

	Pre-1916	**Post-1916**
Most Likely Day	Sep. 11	Sep. 3
Least Likely Day	May 27	Nov. 26

Table 6: Overall Outliers in the Full Model.

	Pre-1916	**Post-1916**
	Positive	
(1)	Dec. 24	Feb. 29
(2)	Feb. 28	Feb. 28
(3)	Jul. 2	May 30
	Negative	
(1)	Dec. 31	Jan. 11
(2)	Nov. 15	Dec. 26
(3)	Feb. 15	May 31

It is worth noting that the set of outliers in the data-sets shown in Table 2 and Table 3 are somewhat similar. One observation of possible interest is that consecutive days in May appear among both the most positively as well as the most negatively extreme residual values. This could be caused by a number of people reporting their birthday as May 30th instead of May 31st, on the mistaken belief that there is only 30 days in May. This would produce the observed result by pulling some births from May 31st to May 30th. We also see that February 28th appears as one of the most positive outliers in both the pre-1916 and post-1916 data-sets. We believe that the preponderance of birthdays on February 28th may be due to the leap day, as many people may be avoiding having the leap day as their actual birthday. This could be accomplished through elective procedures or simply misreporting one's birthday as being on the 28th instead of the 29th. Lastly, there exist a large number of outliers around the holiday season. We believe that this effect is caused at least partially by the nature of the model, as it essentially treats the days that correspond to fixed holidays (i.e., Christmas) as having an average number of births. During the holiday season, there is a preponderance of omitted holidays, and it is therefore likely that these days are outliers because the model has insufficient data to make a quality projection during the holiday season. It is also possible, especially in the pre-1916 data, that a number of people simply rounded their birthday to a holiday inadvertently, with for example December 26th being recorded as December 25th. The presence of July 2nd as an outlier in the pre-1916 data is likely due to the historical association with the independence of the United States, with it being an early contender for the celebration of independence day.

3.2 Analysis of Non-Seasonal Variation

The next step in our analysis was to use the seasonal model created above to better understand the non-seasonal trends. This was done by manually extracting the residuals of the model, taking them as the difference between the seasonal-effect curve and the actual number of births. This number, again with the mean adjusted to fit the overall mean of the corresponding data-set, was used in an ANOVA analysis similar to the one employed in equation (1) in the previous subsection. This procedure yields the following results.

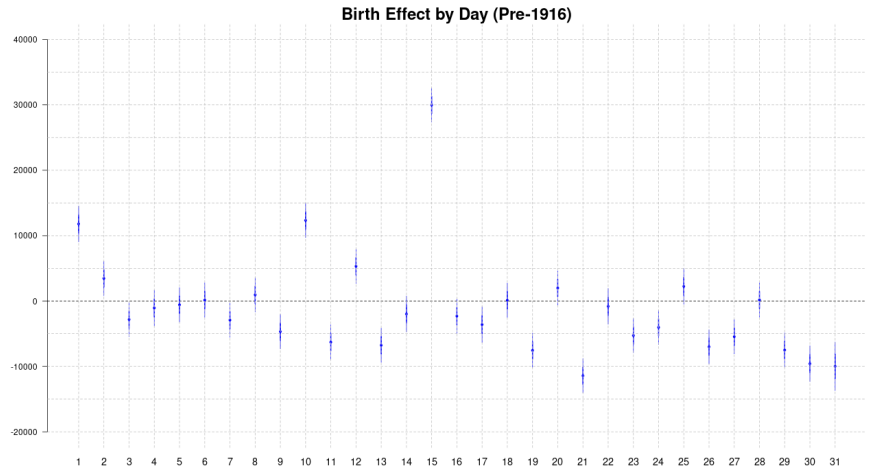

Figure 7: Coefficients for Day Effect in Pre-1916 Births.

Plotting the coefficients in Figures 7-10, we see obvious trends. The 1st, 10th, and 15th appear to have far more births than other days, and the 13th, 21st, and 31st appear to have noticeably fewer births than other days. Holidays and repeating days are also positively associated with birthdays. We also note that the magnitudes of the effects in the post-1916 data-set are much smaller than those in the pre-1916 data-set. These estimates for the effect of certain days of the month and holidays are found by running a regression of these factors on the residuals from the seasonal model (which is calculated as the actual frequency of birthdays minus the fitted value of the seasonal model) and comparing the results to the baseline which is the overall

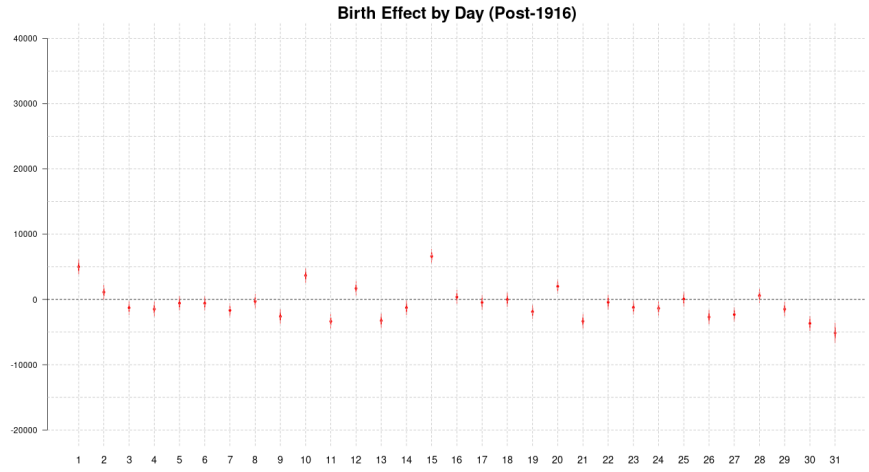

Figure 8: Coefficients for Day Effect in Post-1916 Births.

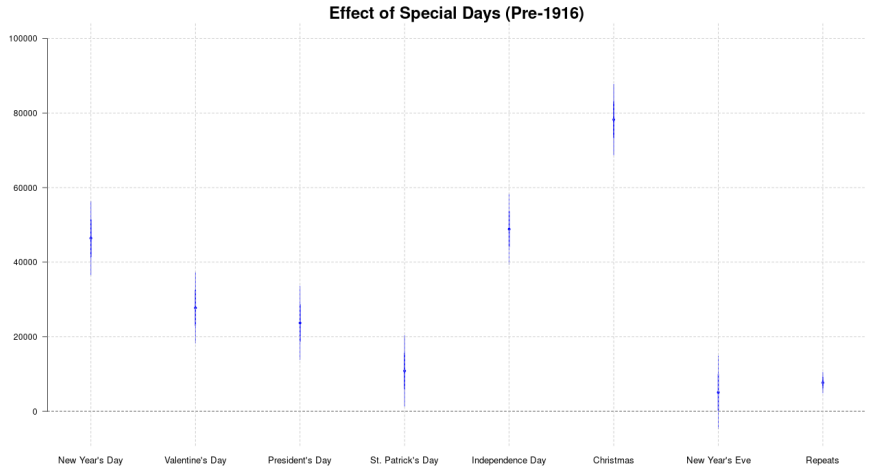

Figure 9: Coefficients for Effect of Special Days in Pre-1916 Births.

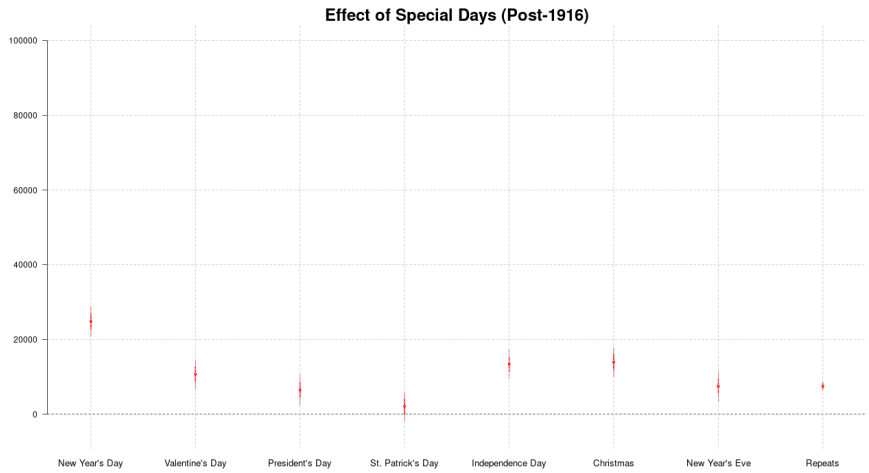

Figure 10: Coefficients for Effect of Special Days in Post-1916 Births.

mean of the residuals of the seasonal model. The error bound was determined by comparing

these results to a t-distribution ($\alpha = 0.05$). The results suggest that most of these days are statistically distinct from the overall mean.

Despite the limited data available, we attempted to investigate the possibility that some birthday entries in the data-set were the result of transposed digits. This would mean that two or more digits or sets of digits were swapped from the actual birthday when entered into the database. Some examples of what this could result in include writing March 2nd instead of February 3rd or writing the 12th instead of the 21st for the day of the month. We tested this hypothesis in two ways. First we looked at whether the first twelve days of the month are more common than the other days, as this could be the result of swapping month and day in the data-set. Second, we tested this possibility by looking at whether days of the month with ascending digits (12, 13, etc) are more common than days with descending digits (21, 31, etc). This test was based on the tendency for people to prefer ascending digits when writing numbers, and we would therefore expect that for example, the 12th has a number of additional birthdays added to it from people who were actually born on the 21st but accidentally transposed the digits of their birthday. Both of these hypotheses seem to be true ($\alpha = 0.05$). These results were tested by the model in equation (1). , and can be verified using the coefficients table in the output appendix.

After careful analysis, it appears that some of the difference in standard deviation between the data-sets seems to be caused by a larger frequency of fictitious birthdays in the pre-1916 data-set. While we cannot directly estimate the number of fictitious birthdays in the data, we were able to get a rough estimate by assuming that our estimated seasonal pattern was the actual trend and then taking the sum of the excess births above the trend-line. This decision to sum over only the entries with births above the trend-line was based on the theory that birthdays are being moved from the real birthday to a fictitious birthday. This means that this approach should give a conservative estimate of how much the observed birthday distribution differs from the seasonal trend-line. This estimate notably excludes fictitious birthdays that are placed in less common dates.

Ultimately, we estimate that approximately 2.7% of the birthdates in the pre-1916 are fictitious, although the real number might be much higher. Using the same methodology for the post-1916 data-set is gives an estimate of 1.1%, which is in line with our expectation. However, interpreting this number is not as straightforward as it was for the pre-1916 data, as much of the deviation from the seasonal trend-line could be caused by elective procedures, and would therefore reflect actual birthdays.

4 Conclusion

After analyzing the data, our conclusion is that the two data-sets have very dissimilar distributions of births. In the aggregate, the pre-1916 data-set has a much larger standard deviation than the post-1916 data-set. After running all the relevant analysis, our conclusion is that this higher standard deviation is at least partially caused by the steeper hills and valleys in the pre-1916 seasonal trend-line.

We believe that the seasonal peaks and valleys are shallower in the post-1916 data-set due to the invention of air conditioning and the decline of agriculture as a common profession. It is our understanding that these factors would at least partially reduce the significance of the significance of seasons in determining activity. It would seem that this trend pushes the natural distribution of birthdays towards a more uniform distribution, thereby lowering the standard deviation of births.

On the other hand, much of the difference in standard deviation seems to be caused by the larger number of apparently fictitious birthdays in the pre-1916 data-set, as it seems that some birthdates, such as the 15th of every month, acquire outsize proportions of the total number of births. Our final estimate was that the pre-1916 data deviates from the seasonal model by about 2.7% while the post-1916 data deviates by as little as 1.1%. Again, it should be emphasized that the apparent deviation from the seasonal trend-line in the post-1916 data-set could be at least partially accounted for by elective procedures, which became fairly common during the 20th century and could be used to offset a birthday by a number of days. It seems likely that these procedures would reduce the number of births on holidays for birthdays in the late 20th

century, due to the inability to get non-emergency hospital procedures on a holiday.

We calculated the number of potentially-erroneous birthdays as the sum of the difference between the predicted values of the seasonal model, minus the overall mean of the data-set, where this difference is positive. The proportion of fictitious birthdays was calculated by dividing this number by the overall number of birthdays recorded within the data-set. While we do not believe that this is the actual number of fictitious birthdays, we believe that it is a decent estimate, as it accounts for pull towards certain dates that cannot be explained by the seasonal trends alone.

After comparing these two datasets, we would expect that more modern data would more closely follow the seasonal pattern. This would reflect the fact that actual birthdays are much more well-document in modern times, due to strict documentation of births, as it would be more difficult to get in a situation where one simply would not know their own birthday and would be unable to find it out from a some hospital or government document.

5 Appendix

Detailed View of Heatmaps

	January	February	March	April	May	June	July	August	September	October	November	December
1	187650	127020	143900	136368	140212	125350	148205	138341	143442	140771	131651	124691
2	121849	149304	130298	127210	120694	117080	141090	128540	132499	131790	123096	117778
3	119387	119704	134705	121615	120663	113317	115790	125336	130292	129312	117832	118041
4	117152	123656	140470	125999	116658	114862	167061	129200	128583	133972	120261	118072
5	120881	124406	129289	126387	129509	116394	115942	126512	132773	130425	119003	121321
6	133604	123381	128227	123571	118865	124382	118046	129080	131871	130120	120080	120569
7	119227	125061	127895	124273	115282	113444	123319	125553	131895	125886	119057	114168
8	121790	127608	130300	121412	123443	113588	119209	135954	139297	131177	121594	126101
9	117026	122953	125833	119042	113781	113446	114487	126311	133832	124871	119058	113367
10	133570	133949	150618	136308	141045	131162	130529	145176	144611	146400	125104	129809
11	114351	122554	122007	117234	112301	111579	113988	124210	127508	121691	131866	113195
12	125080	137501	137324	128330	123328	118279	126344	137956	140028	138618	119495	131950
13	116113	117052	121802	114668	110449	115341	113282	122570	129259	120534	113184	117401
14	119964	151964	128404	119910	114541	114933	121456	128398	133395	127668	119524	116570
15	147020	145353	170952	159296	154585	140705	149872	173306	172011	157550	137915	144141
16	119782	124081	124241	119091	115524	114171	123435	130158	134508	125023	117339	117988
17	119941	126082	136093	117623	111940	112586	119566	125587	133742	122235	117815	117402
18	122582	129340	127678	119023	117602	114986	121864	130665	135715	127613	120304	127107
19	115699	119107	132639	112589	107437	109688	113668	121367	124525	115850	114578	115456
20	125145	124108	131444	119711	123322	120165	122449	135434	137650	129848	123551	124231
21	114236	109524	118956	108259	106427	109996	111562	118738	125914	113698	110093	108722
22	121510	158203	125761	114803	116714	118866	122733	131390	136789	124996	123375	119756
23	115395	115686	120871	117540	109818	109208	115895	125976	131901	119558	116855	131291
24	113279	119282	119939	110988	110610	122232	119604	126390	126201	119497	117455	139535
25	124411	122622	136915	118347	118022	113257	127566	135188	138735	126649	133112	203791
26	117382	119458	117939	110097	109188	110489	122513	125005	128637	119017	115647	114002
27	119619	118921	122184	113936	109068	110926	120810	126543	131470	119190	118262	117182
28	122948	147559	126344	113745	114650	112578	123859	132954	130993	125714	119091	125021
29	119544	111360	118097	109425	106943	116948	121647	125172	129280	115758	112370	116990
30	110284	0	111510	110541	122298	115459	110986	121408	126747	112825	116249	101099
31	115404	0	119436	0	99264	0	118551	124869	0	113228	0	118408

Figure 11: Detailed Heatmap for Pre-1916 Births.

	January	February	March	April	May	June	July	August	September	October	November	December
1	159358	125153	131971	129558	127740	124215	127768	129553	130139	127410	118480	116667
2	122610	138445	126475	124180	119988	121313	121507	124382	125328	123603	115405	113072
3	122428	122460	132543	121880	120122	117238	117491	122949	124224	121584	112368	111965
4	120178	123919	127299	126492	117440	116254	132875	123656	121598	122269	113596	112321
5	121550	124633	126188	122511	128073	118115	115892	124490	123645	122279	114327	113371
6	124231	124393	125406	121054	118942	126594	119419	124440	123814	120818	112917	112662
7	119453	124176	124185	120364	117572	116519	126432	123665	123238	119613	114294	111863
8	120246	124968	125812	121437	119542	117420	121265	131676	125754	121018	113393	115188
9	118687	123126	123246	118936	115985	117130	119079	122615	129044	117836	112780	111210
10	126634	127660	130857	124948	125980	124534	124815	129537	128538	130213	114214	116779
11	118433	122488	122884	118038	115145	116200	119099	122764	122118	115910	123825	110664
12	123029	129887	128645	122951	119336	120591	122486	127946	127198	123161	112597	122640
13	119459	120787	123139	117232	114615	117582	118732	121821	122769	114957	110301	112080
14	120430	135293	123910	120050	115835	117483	122165	124126	124598	117548	112957	113372
15	128319	129038	133068	128211	126223	126361	128148	135163	133425	125108	119351	118891
16	122360	124859	123625	121252	118018	120058	123806	126672	128316	119398	114272	113670
17	122052	125170	126995	120152	117136	118207	123418	124503	126663	116670	113890	113336
18	122757	126799	123641	120099	117271	119003	121988	126122	125909	117616	113926	116004
19	121323	123745	125507	117696	115867	117704	119869	122912	124974	113746	112863	112479
20	125492	126193	127689	121148	122303	121695	123558	127750	127703	119899	115572	115950
21	120646	118155	124804	116136	115714	118072	120853	121066	124100	112174	109436	109513
22	122174	139924	122808	117167	117872	119837	123985	125527	126820	115121	115206	113037
23	119769	122857	122569	118936	115713	117077	121475	125019	127230	115557	112518	117117
24	119944	123450	122086	116157	114754	121122	122955	123510	123598	114927	113143	118277
25	122612	124055	125525	117414	118781	118762	124009	126295	126338	116598	115447	128963
26	120770	123541	120833	115292	115986	118995	122105	123200	124278	114159	112115	106309
27	121425	121919	121987	116382	115323	116216	123722	122877	124030	113632	113955	110684
28	123484	135778	123207	116776	118579	118012	124285	127298	123590	116148	113511	116005
29	122326	134212	119515	114877	114884	118264	123333	122888	122404	112981	111187	114888
30	118144	0	116770	115609	120563	120294	117350	121486	121574	110425	111787	108432
31	119272	0	119393	0	107950	0	119796	122552	0	111995	0	117905

Figure 12: Detailed Heatmap for Post-1916 Births.

R Output

Seasonal Model

```
##############################

#### Pre-1916 Seasonal Model ####


##############################


## must run all previous code for this to work
## R input and output for the natural variation

## attempts to explain the residuals of the previous model

## overall mean added back to it for scaling



## sets up a new model with the constraint that the fitted values for 1/1 = 12/31
## sets up the constraint matrix
## forces 1/1 and 12/31 to have expected value as average
con <- matrix((c(0,1,120947.4,
                 0,366,120947.4
              )), nrow=2, ncol=3, byrow=TRUE)

## uses 8 knots
> use_knots <- c(1,53,105,157,209,261,313,366)

## sets up a new model with the constraint that the fitted values for 1/1 = 12/31
## sets up the constraint matrix
## forces 1/1 and 12/31 to have expected value as average
> mdl2 <- cobs(birthdata$totday,birthdata$unex, pointwise = con, knots=use_knots)

> summary(mdl2)
COBS regression spline (degree = 2) from call:
cobs(x = birthdata$totday, y = birthdata$unex, knots = use_knots,     pointwise = con)
{tau=0.5}-quantile;  dimensionality of fit: 8 from {8}
x$knots[1:7]:   0.999635,  53.000000, 105.000000, ... , 366.000365
with 2 pointwise constraints
coef[1:8]: 120947.4, 123171.1, 130329.6, 107026.7, 135242.6, ... , 120947.5
R^2 = 54.81% ;  empirical tau (over all): 178/366 = 0.4863388 (target tau= 0.5)

## manually calculates actual RMSE
> sqrt((sum((birthdata$births - mdl$fitted.values)^2))/366)
[1] 11172.64

##############################

#### Post-1916 Seasonal Model ####

##############################


## must run all previous code for this to work
## R input and output for the natural variation

## attempts to explain the residuals of the previous model

## overall mean added back to it for scaling
```

```
## this is the spline model for seasonal variation
## attempt to predict the residuals of the model in equation (1)
## with only the day of the year, overall mean added for scaling
> mdl2 <- cobs(birthdata$totday,birthdata$unex, pointwise = con, knots=use_knots)

## renames the model fitted and resid as mdl
## for legacy code compatibility
> mdl$fitted.values <- fitted(mdl2)
> mdl$residuals <- resid(mdl2)

## gets summary of the model
> summary(mdl2)
COBS regression spline (degree = 2) from call:
cobs(x = birthdata$totday, y = birthdata$unex, knots = use_knots,     pointwise = con)
{tau=0.5}-quantile;  dimensionality of fit: 9 from {9}
x$knots[1:8]:   0.999635,  53.000000, 105.000000, ... , 366.000365
with 2 pointwise constraints
coef[1:9]: 120947.3, 124573.7, 125669.5, 116652.1, 119986.3, ... , 120947.6
R^2 = 77.94% ;  empirical tau (over all): 191/366 = 0.5218579 (target tau= 0.5)

## gets true RMSE for seasonal model only
> sqrt((sum((birthdata$births - mdl$fitted.values)^2))/366)
[1] 4189.988
```

5.0.1 Non-seasonal Model

```
######################
## Nonseasonal Model  ##
######  Pre-1916 #######
######################
> ## runs a model predicting the number of births by
> ## day of the month and holiday and repeating digits
> ## gives a perfect fit for holidays
> daymodel <- lm(births~as.factor(day) + holiday + rpt,data=birthdata)
> summary.aov(daymodel)
                Df    Sum Sq   Mean Sq F value   Pr(>F)
as.factor(day)  30 2.510e+10 8.367e+08   16.57  < 2e-16 ***
holiday          7 1.167e+10 1.667e+09   33.02  < 2e-16 ***
rpt              1 6.232e+08 6.232e+08   12.35 0.000504 ***
Residuals      327 1.651e+10 5.048e+07
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
> summary.lm(daymodel)

Call:
lm(formula = births ~ as.factor(day) + holiday + rpt, data = birthdata)

Residuals:
     Min       1Q   Median       3Q      Max
-16477.2  -4690.7   -754.8   4655.0  22937.7

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      136359.2     2142.2  63.654  < 2e-16 ***
as.factor(day)2   -8555.8     2971.2  -2.880 0.004245 **
as.factor(day)3  -14825.3     2971.2  -4.990 9.85e-07 ***
as.factor(day)4  -12604.8     3035.9  -4.152 4.21e-05 ***
as.factor(day)5  -12587.9     2971.2  -4.237 2.95e-05 ***
as.factor(day)6  -11841.8     2971.2  -3.986 8.31e-05 ***
as.factor(day)7  -14903.1     2971.2  -5.016 8.68e-07 ***
as.factor(day)8  -11035.4     2971.2  -3.714 0.000240 ***
```

```
as.factor(day)9  -16657.5     2971.2  -5.606 4.40e-08 ***
as.factor(day)10    365.3     2971.2   0.123 0.902224
as.factor(day)11 -18250.0     2987.5  -6.109 2.85e-09 ***
as.factor(day)12  -6638.7     2971.2  -2.234 0.026136 *
as.factor(day)13 -18721.3     2965.8  -6.312 8.94e-10 ***
as.factor(day)14 -14108.0     3029.5  -4.657 4.68e-06 ***
as.factor(day)15  18033.0     2965.8   6.080 3.34e-09 ***
as.factor(day)16 -14247.4     2965.8  -4.804 2.37e-06 ***
as.factor(day)17 -15948.4     3029.5  -5.264 2.55e-07 ***
as.factor(day)18 -11819.3     2965.8  -3.985 8.32e-05 ***
as.factor(day)19 -19475.6     2965.8  -6.567 2.02e-10 ***
as.factor(day)20  -9937.7     2965.8  -3.351 0.000900 ***
as.factor(day)21 -23348.8     2965.8  -7.873 5.15e-14 ***
as.factor(day)22 -13023.5     3029.5  -4.299 2.27e-05 ***
as.factor(day)23 -17193.0     2965.8  -5.797 1.59e-08 ***
as.factor(day)24 -15941.5     2965.8  -5.375 1.46e-07 ***
as.factor(day)25  -9557.0     3029.5  -3.155 0.001756 **
as.factor(day)26 -18911.3     2965.8  -6.377 6.17e-10 ***
as.factor(day)27 -17349.9     2965.8  -5.850 1.19e-08 ***
as.factor(day)28 -11737.8     2965.8  -3.958 9.28e-05 ***
as.factor(day)29 -19398.0     2965.8  -6.541 2.36e-10 ***
as.factor(day)30 -21867.7     3029.5  -7.218 3.71e-12 ***
as.factor(day)31 -21233.8     3605.9  -5.889 9.66e-09 ***
holiday1          43703.7     7728.6   5.655 3.41e-08 ***
holiday45         29712.8     7420.8   4.004 7.71e-05 ***
holiday53         27280.1     7728.6   3.530 0.000476 ***
holiday77         15682.2     7420.8   2.113 0.035335 *
holiday186        43306.7     7423.4   5.834 1.30e-08 ***
holiday360        76988.8     7420.8  10.375  < 2e-16 ***
holiday366         3282.7     7674.2   0.428 0.669111
rpt                7587.2     2159.4   3.514 0.000504 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Residual standard error: 7105 on 327 degrees of freedom
Multiple R-squared:  0.6937, Adjusted R-squared:  0.6581
F-statistic: 19.49 on 38 and 327 DF,  p-value: < 2.2e-16


######################
## Nonseasonal Model  ##
#####  Post-1916 #######
######################

> ## runs a model predicting the number of births by
> ## day of the month and holiday and repeating digits
> ## gives a perfect fit for holidays
> daymodel <- lm(births~as.factor(day) + holiday + rpt,data=birthdata)
> summary.aov(daymodel)
               Df    Sum Sq   Mean Sq F value   Pr(>F)
as.factor(day)  30 2.712e+09  90415665   4.124 6.04e-11 ***
holiday          7 1.842e+09 263200866  12.006 1.21e-13 ***
rpt              1 5.974e+08 597404444  27.250 3.18e-07 ***
Residuals      327 7.169e+09  21922791
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
> summary.lm(daymodel)

Call:
lm(formula = births ~ as.factor(day) + holiday + rpt, data = birthdata)

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-11822.9  -2975.5    402.1   3507.7  14898.8

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         126241       1412  89.423  < 2e-16 ***
as.factor(day)2      -3835       1958  -1.958 0.051032 .
as.factor(day)3      -6256       1958  -3.195 0.001535 **
as.factor(day)4      -6460       2001  -3.229 0.001369 **
as.factor(day)5      -5604       1958  -2.862 0.004479 **
as.factor(day)6      -5636       1958  -2.878 0.004259 **
as.factor(day)7      -6746       1958  -3.445 0.000645 ***
as.factor(day)8      -5384       1958  -2.750 0.006300 **
as.factor(day)9      -7721       1958  -3.943 9.84e-05 ***
as.factor(day)10     -1468       1958  -0.750 0.453988
as.factor(day)11     -8515       1969  -4.325 2.03e-05 ***
as.factor(day)12     -3488       1958  -1.781 0.075775 .
as.factor(day)13     -8452       1954  -4.324 2.03e-05 ***
as.factor(day)14     -6926       1996  -3.469 0.000593 ***
as.factor(day)15      1368       1954   0.700 0.484602
as.factor(day)16     -4882       1954  -2.498 0.012976 *
as.factor(day)17     -6132       1996  -3.072 0.002308 **
as.factor(day)18     -5313       1954  -2.719 0.006906 **
as.factor(day)19     -7184       1954  -3.676 0.000277 ***
as.factor(day)20     -3329       1954  -1.703 0.089501 .
as.factor(day)21     -8686       1954  -4.444 1.21e-05 ***
as.factor(day)22     -6282       1996  -3.146 0.001805 **
as.factor(day)23     -6588       1954  -3.371 0.000839 ***
as.factor(day)24     -6748       1954  -3.452 0.000628 ***
as.factor(day)25     -4802       1996  -2.405 0.016726 *
as.factor(day)26     -8109       1954  -4.149 4.26e-05 ***
as.factor(day)27     -7729       1954  -3.954 9.41e-05 ***
as.factor(day)28     -4852       1954  -2.482 0.013549 *
as.factor(day)29     -6928       1954  -3.545 0.000450 ***
as.factor(day)30     -9656       1996  -4.837 2.04e-06 ***
as.factor(day)31     -9415       2376  -3.962 9.13e-05 ***
holiday1             25688       5093   5.044 7.59e-07 ***
holiday45            15977       4890   3.267 0.001202 **
holiday53            12536       5093   2.461 0.014359 *
holiday77             6886       4890   1.408 0.160049
holiday186           13094       4892   2.677 0.007814 **
holiday360            7523       4890   1.538 0.124917
holiday366            1079       5057   0.213 0.831235
rpt                   7429       1423   5.220 3.18e-07 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Residual standard error: 4682 on 327 degrees of freedom
Multiple R-squared:  0.4182, Adjusted R-squared:  0.3506
F-statistic: 6.185 on 38 and 327 DF,  p-value: < 2.2e-16
```

R Code

```
####################
## Import Library ##
####################
library(tidyverse)
library(readxl)
library(lubridate)
library(splines)
library(boot)
library(pls)
```

```
library(ggplot2)
library(cobs)

####################
## Importing Data ##
####################

## import the pre-1916 data as: ssdm_pret1916
## import the post-1916 data as: ssdm_post1916

## use this function for running the code below
## the dataset analyzed should be called birthdata
## it should be coded with the number of births
## and the day of the year and which month
birthdata <- ssdm_post_1916
birthdata <- ssdm_pre_1916
View(birthdata)

#####################
## Variable Adding ##
#####################

## adds the day of the year
for(i in 1:nrow(birthdata))
{
  birthdata$totday[i] = i
}

## adds day of the month
birthdata$day <- 0
day=1
birthdata$births[birthdata$month==2]
count=1
for(i in 1:12)
{
  day=1
  for(g in 1:length(birthdata$births[birthdata$month==i]))
  {
    birthdata$day[count] = g
    day = day + 1
    count=count+1
  }
}

#####################
### Day Variation ###
#####################

## initialize the repeating digits
## is 0 by default
## is 1 for 1/1,1/11,2/2, etc
birthdata$rpt <- 0
for(i in 1:nrow(birthdata))
{
  if(birthdata$month[i]==birthdata$day[i])
  {
    birthdata$rpt[i] <- 1
  }
  if((birthdata$month[i]==1)&(birthdata$day[i])==11)
  {
    birthdata$rpt[i] <- 1
  }
```

```
  if((birthdata$month[i]==2)&(birthdata$day[i])==22)
  {
    birthdata$rpt[i] <- 1
  }
  if((birthdata$month[i]==11)&(birthdata$day[i])==11)
  {
    birthdata$rpt[i] <- 1
  }
}

## initialize the holidays
## first as a factor variable with the levels
## being day of the year corresponding to holidays
## then also as a boolean with 0 def and 1 if it is a holiday
birthdata$holiday <- 0
birthdata$is_holiday <- 0
for(i in 1:nrow(birthdata))
{
  if((i==1)|
     (i==45)|
     (i==53)|
     (i==77)|
     (i==186)|
     (i==360)|
     (i==366))
  {
    birthdata$holiday[i] <- i
    birthdata$is_holiday[i] <- 1
  }
}
birthdata$holiday <- as.factor(birthdata$holiday)


######################
##### Modeling ######
######################

## runs a model predicting the number of births by
## day of the month and holiday and repeating digits
## gives a perfect fit for holidays
daymodel <- lm(births~as.factor(day) + holiday + rpt,data=birthdata)
summary.aov(daymodel)
summary.lm(daymodel)

## plots the residuals
plot(birthdata$totday,(daymodel$residuals+mean(birthdata$births)), ylim=c(100000,160000))

## sets the unex var as the residuals of the daymodel regression
## adds the mean, for plot scaling, used for only the seasonal trendlines
birthdata$unex <- (daymodel$residuals+mean(birthdata$births))

## runs the regression on the residuals of the daymodel regression
## fits a spline model with 12 df
## this is the legacy model that does not connect the 1/1 with 12/31
mdl <- glm(unex~bs(totday, df = 12),data=birthdata) #fitting spline
summary(mdl)

## sets up a new model with the constraint that 1/1 = 12/31
## sets up the constraint matrix
## forces 1/1 and 12/31 to have expected value as average
con <- matrix((c(0,1,120947.4,
                 0,366,120947.4
```

```
              )), nrow=2, ncol=3, byrow=TRUE)

con <- matrix((c(0,1,120947.4,
                 0,366,120947.4
              )), nrow=2, ncol=3, byrow=TRUE)


## uses 8 knots for pre-1916 and post-1916 using method 2
use_knots <- c(1,53,105,157,209,261,313,366)

## this is the main model in the analysis
## it connects 1/1 and 12/31
## mdl2 <- cobs(birthdata$totday,birthdata$unex, pointwise = con, nknots=8)
mdl2 <- cobs(birthdata$totday,birthdata$unex, pointwise = con, knots=use_knots)

## renames the model fitted and resid as mdl
## for legacy code compatibility
mdl$fitted.values <- fitted(mdl2)
mdl$residuals <- resid(mdl2)
mdl2$coef
summary(mdl2)

## gets true RMSE for seasonal model only
sqrt((sum((birthdata$births - mdl$fitted.values)^2))/366)


######################
##### Graphing ######
######################

## graph the results of the second model 'mdl'
## use the one appropriate for the dataset
plot(birthdata$totday,mdl$fitted.values, ylim=c(110000,140000),pch=20,
    xlab = 'Day␣of␣the␣Year', ylab = 'Predicted␣Births',
    main='Seasonal␣Trendline␣For␣Post-1916')
plot(birthdata$totday,mdl$fitted.values, ylim=c(110000,140000),pch=20,
    xlab = 'Day␣of␣the␣Year', ylab ='Predicted␣Births',
    main='Seasonal␣Trendline␣For␣Pre-1916')

## plot the original data for post 1916
plot(birthdata$totday,ssdm_post_1916$births, ylim=c(100000,190000),pch=20, xaxt='n',
    xlab = 'Day␣of␣the␣Year', ylab = 'Actual␣Births', main='Post-1916:␣Original␣Data',
    col='red')
axis(1, at=c(1,32, 61,92,122,153,183,214,245,275,306,336),
    labels=c('Jan', 'Feb', 'Mar','Apr','May','Jun','Jul','Aug','Sep','Oct','Nov','Dec'))

## plot the original data for pre 1916
plot(birthdata$totday,ssdm_pre_1916$births, ylim=c(100000,190000),pch=20, xaxt='n',
    xlab = 'Day␣of␣the␣Year', ylab = 'Actual␣Births', main='Pre-1916:␣Original␣Data',
    col='blue')
axis(1, at=c(1,32, 61,92,122,153,183,214,245,275,306,336),
    labels=c('Jan', 'Feb', 'Mar','Apr','May','Jun','Jul','Aug','Sep','Oct','Nov','Dec'))

## graph the results of the smoothed seasonal model 'mdl'
## adds loess trendlines to make it clear
plot(birthdata$totday,mdl$fitted.values, ylim=c(110000,140000),pch=20, xaxt='n',
    xlab = 'Day␣of␣the␣Year', ylab = 'Predicted␣Births',
    main='Seasonal␣Trendlines', type='n')
lo1 <- loess(premodel$fitted.values~birthdata$totday)
lo2 <- loess(postmodel$fitted.values~birthdata$totday)
lines(lo1,col='blue',lwd=4)
lines(lo2,col='red',lwd=4)
```

```
legend("topleft",c("before 1916","after 1916"),
    pch=c(-1,-1),col = c("red","blue"),cex=1,lty=c(1,1))
axis(1, at=c(1,32, 61,92,122,153,183,214,245,275,306,336),
    labels=c('Jan', 'Feb', 'Mar','Apr','May','Jun','Jul','Aug','Sep','Oct','Nov','Dec'))

## check the residuals vs fitted
plot(birthdata$totday,(birthdata$births-mdl$fitted.values), ylim=c(-8000,8000))

## checks for dist of residuals, should be somewhat normal
## not entirely normally distributed, but not too skewed
hist((birthdata$births-mdl$fitted.values), breaks=60)

## checks for bias in the model
mean(birthdata$births-mdl$fitted.values)

## gets the most common birthdate, seasonal trends only
which.max(mdl$fitted.values)

## plots the residuals of just the trendline
plot(birthdata$totday,(birthdata$births - mdl$fitted.values), ylim=c(-20000,20000))
plot(birthdata$totday,mdl$fitted.values, ylim=c(100000,160000))

## checks for digit transposition
## looks at whether the first
mdl <- glm(births~(day<=12),data=birthdata) #fitting spline
summary(mdl)

##############
### RMSE ####
##############

## gets rmse
## about 4x as much for pre-1916 as post
birthdata$pred <- daymodel$fitted.values -mean(daymodel$fitted.values) +
    mdl$fitted.values - mean(mdl$fitted.values)
    + mean(birthdata$births)
sqrt(sum((birthdata$births - birthdata$pred)^2)/366)

## sets up manual residuals
birthdata$resid <- birthdata$births - birthdata$pred
mean(birthdata$resid)
plot(birthdata$resid,pch=20)

## pre 1916
# local max: March 11
# minimum: May 27
# maximum: Sep 11
# local min: Nov 23

## post 1916
# local max: March 3
# local min: May 16
# maximum: Sep 3
# minimum: Nov 26

## extra plots with loess lines
plot(birthdata$totday,birthdata$births, ylim=c(100000,200000),pch=20, xaxt='n',
    xlab = 'Day of the Year', ylab = 'Births', main='Post-1916 Data With Trendlines',
        col='red')
plot(birthdata$totday,birthdata$births, ylim=c(100000,200000),pch=20, xaxt='n',
    xlab = 'Day of the Year', ylab = 'Births', main='Pre-1916 Data With Trendlines',
        col='blue')
```

```
lo2 <- loess(mdl$fitted.values~birthdata$totday)
lines(lo2,col='black',lwd=4)
axis(1, at=c(1,32, 61,92,122,153,183,214,245,275,306,336),
    labels=c('Jan', 'Feb', 'Mar','Apr','May','Jun','Jul','Aug','Sep','Oct','Nov','Dec'))

## Summary Stats pre-1916
sd(ssdm_pre_1916$births)
mean(ssdm_pre_1916$births)
sum(ssdm_pre_1916$births)

## summary stats post-1916
sd(ssdm_post_1916$births)
mean(ssdm_post_1916$births)
sum(ssdm_post_1916$births)

## gets the outliers orders by magnititude
residdat <- birthdata[order(birthdata$resid),]

## pre1916
# most neg resid
# 365: -12690 -> 12/30
# 320: -12053 -> 11/15
#  46: -10656 -> 02/15
# most pos resid
# 359:  19957 -> 12/24
#  59:  18683 -> 02/28
# 184:  16876 -> 07/02

## post1916
# most neg resid
#  11: -7378 -> 01/11
# 361: -6578 -> 12/26
# 152: -6525 -> 05/31
# most pos resid
#  60:  9727 -> 02/29
#  59:  9237 -> 02/28
# 151:  6368 -> 05/30

## method 1
## takes the sum of excess births caused by the day and holiday factors
(sum(daymodel$fitted.values[daymodel$fitted.values>=mean(birthdata$births)]-
    mean(birthdata$births)))/sum(birthdata$births)

## method 2
## takes the sum of the difference between actual births and the seasonal trend
## using only positive differences
sum((birthdata$births - mdl$fitted.values)[(birthdata$births - mdl$fitted.values) >= 0])/
    sum(birthdata$births)

which.max(mdl$fitted.values)
birthdata[248,]
```