**Analysis of Marriage Survival Rates based on Fixed Social Status Variables**

**Abstract**

Marriage is a vital institution in modern society - Supreme Court Justice Robert Kennedy called it a “keystone of our social order” in his decision to support gay marriage. Yet, despite the best intentions surrounding a marriage, many end in divorce. Considering the social fallout and public costs to a marriage ending, it would be valuable to be able to predict what factors lead to divorce. We used the Panel Study of Income Dynamics (PSID) data set to analyze how order of a marriage, age married, and decade married influence survival rates of marriage.1 We used Kaplan-Meier models to perform survival analysis on marriages and the log-rank test to evaluate the significance of our findings. We found that getting married at an older age, first marriages, and certain decades are correlated with better survival rates, though further research is needed to explore causal relationships.

**Background**

No one enters a marriage believing it will end in divorce, but despite our best intentions, nearly 50% of first marriages end this way. Marriage can be tricky and challenging at times, and the best we can do is playing the odds. Moreover, the social landscape surrounding marriages has changed dramatically during the past 50 years. The women’s liberation movement in the 1960s and the sexual revolution from the 1960s to 1980s have changed gender roles in marriage, as well as decreased the shame of being divorced and getting remarried. The promulgation of no-fault divorce laws in many states might also contribute to an increase in the divorce and remarriage rate. Better knowledge of the length of marriage is important to public policy considerations since divorce creates significant negative externalities related to children’s wellbeing. The major issue is that children have better physical, emotional, and academic outcomes when living in a healthy marriage.2 Based on prior research, the risk of divorce decreases as age increases.3,4 The literature also supports that first marriages last longer than subsequent marriages.5 Our hypotheses are that the older you are when you get married, the more likely your marriage is to survive; first marriages will be more likely to survive than later marriages; and that people who got married in earlier decades would last longer due to traditional values.

**Data Cleaning and Methods**

We conducted survival analysis on the length of marriage using data from the Panel Study of Income Dynamics (PSID) collected by the University of Michigan.1 The PSID is a longitudinal, survey-based study of socio-economic factors, originally intended to study the effects of income on U.S. families. The initial cohort is from 1968, and data has been gathered annually for the past 50 years. It followed the initial cohort and all members of their household. The survey data was gathered primarily through the phone. Due to the nature of longitudinal studies, there were numerous observations with missing data.

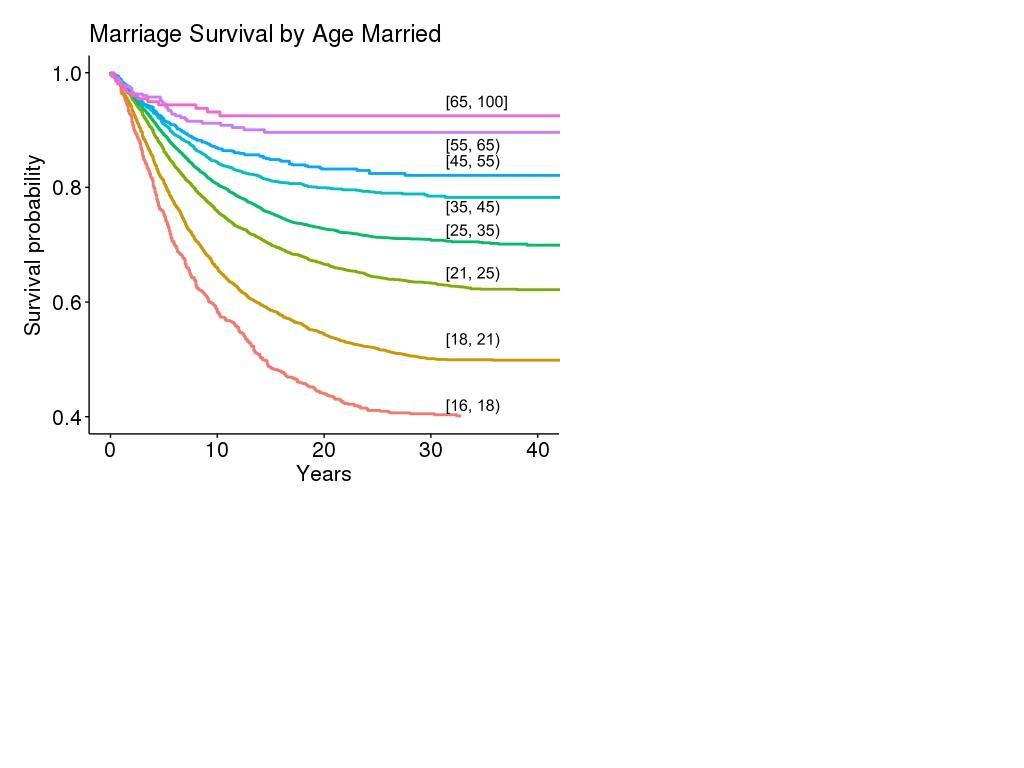
Cleaning our data involved relabeling columns to be more readable, filtering out irrelevant or unworkable observations, and making adjustments to imperfect but workable observations. First, we removed any data that fell outside our interest: namely unmarried individuals and marriages which took place outside the year range 1970 to 2009 (including people who did not know when they were married). After removing irrelevant data, we had 25,058 observations. Next, we removed observations that were missing crucial information, such as an unknown order of marriage or an unknown year of birth. This round of filtering brought our observation count down to 23,182. Filtering based on missing information resulted in us losing about 7.5% of our potential data. There is a risk of bias in removing these observations: it is possible that people who are less content in their marriage are less likely to remember or report relevant dates. However, the percentage removed is small enough that we think the study is still valid.

From the dates provided, we created the variables of interest: length of marriage, age group when married,[[1]](#footnote-1) order of marriage, and decade married. When an observation was missing a month (for birth, death, marriage, or divorce), we defaulted to the first month of the year.[[2]](#footnote-2) Truncating missing months helped preserve some of the data we would have otherwise removed. Unfortunately, removing missing months occasionally created some issues: When an individual got married and divorced in the same year, but was missing divorce month, the truncation would result in a negative marriage length. In this case, we set the marriage length to zero. The full data cleaning process can be found in Appendix C.

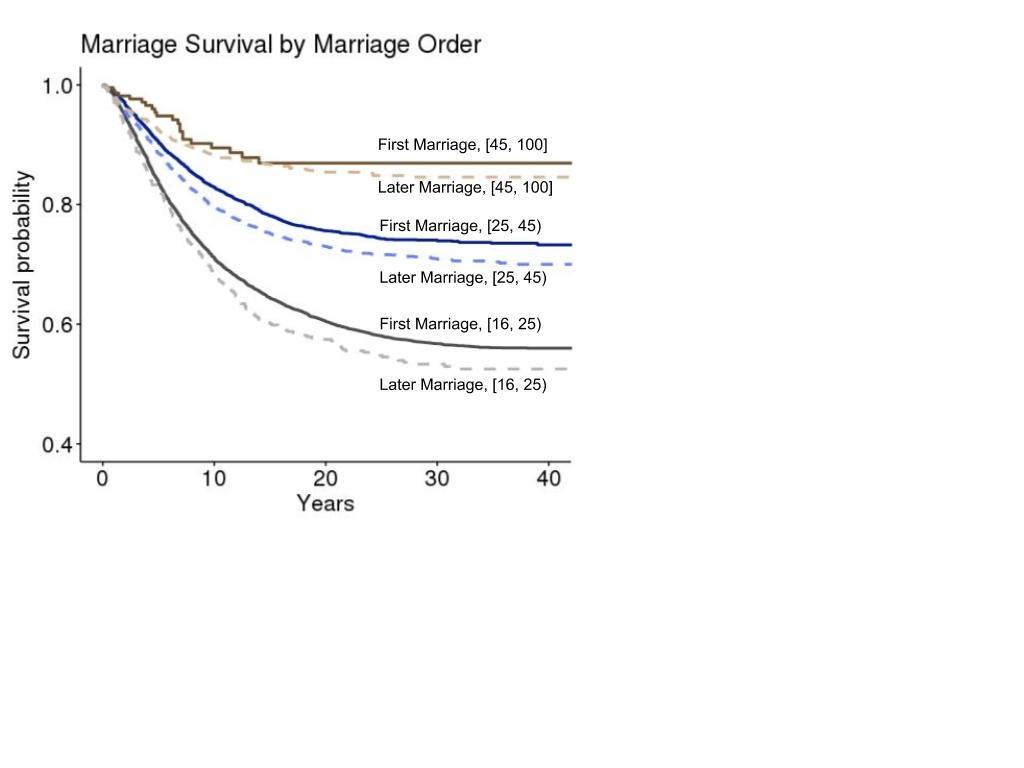
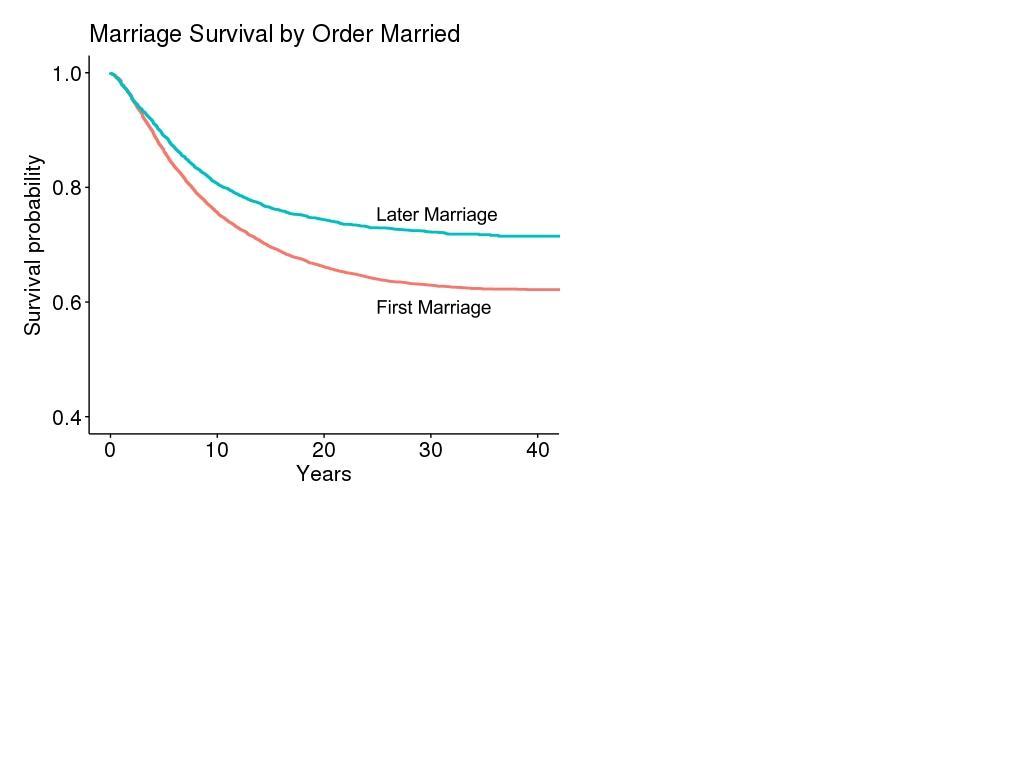
**Results and Discussion**

We used Kaplan-Meier curves and the log-rank test to explore our hypothesis. For the analysis, we treated intact marriages and death of spouse as censored events. The log-rank test yielded a significant p-value (<0.05) for our three explanatory variables. Thus, we can reject the null hypothesis that age when married, order of the marriage, and decade married have no effect on the survival chances of marriage. However, it is worth noting that a large dataset (24,000+ observations), we are more likely to get significant results.

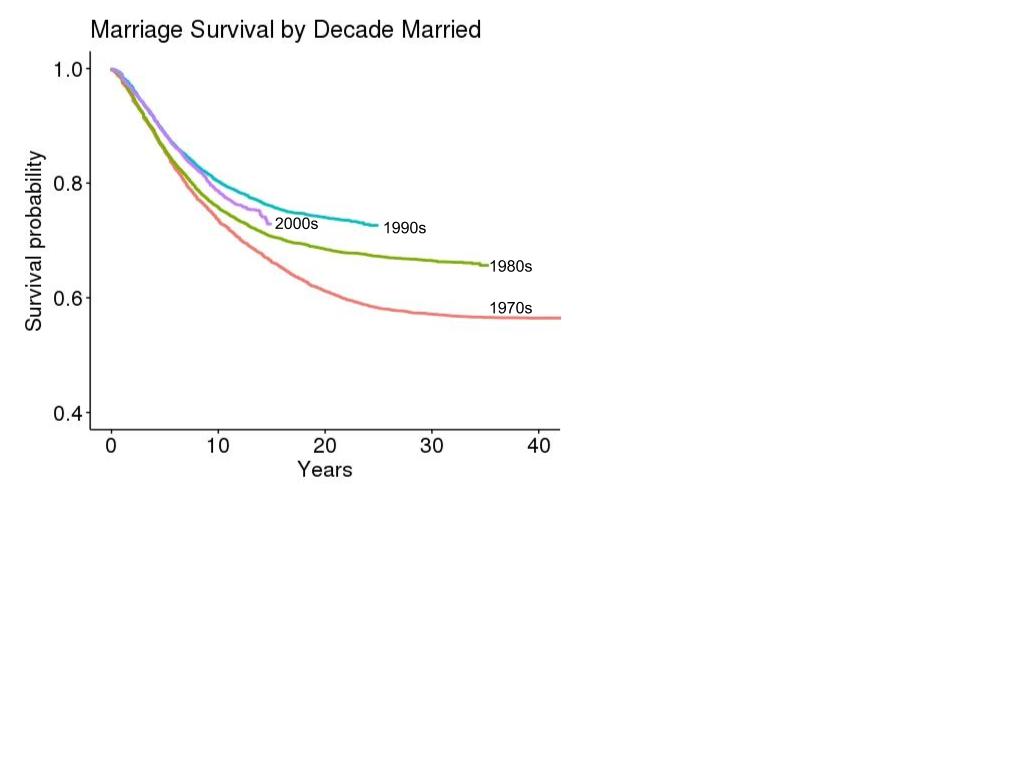
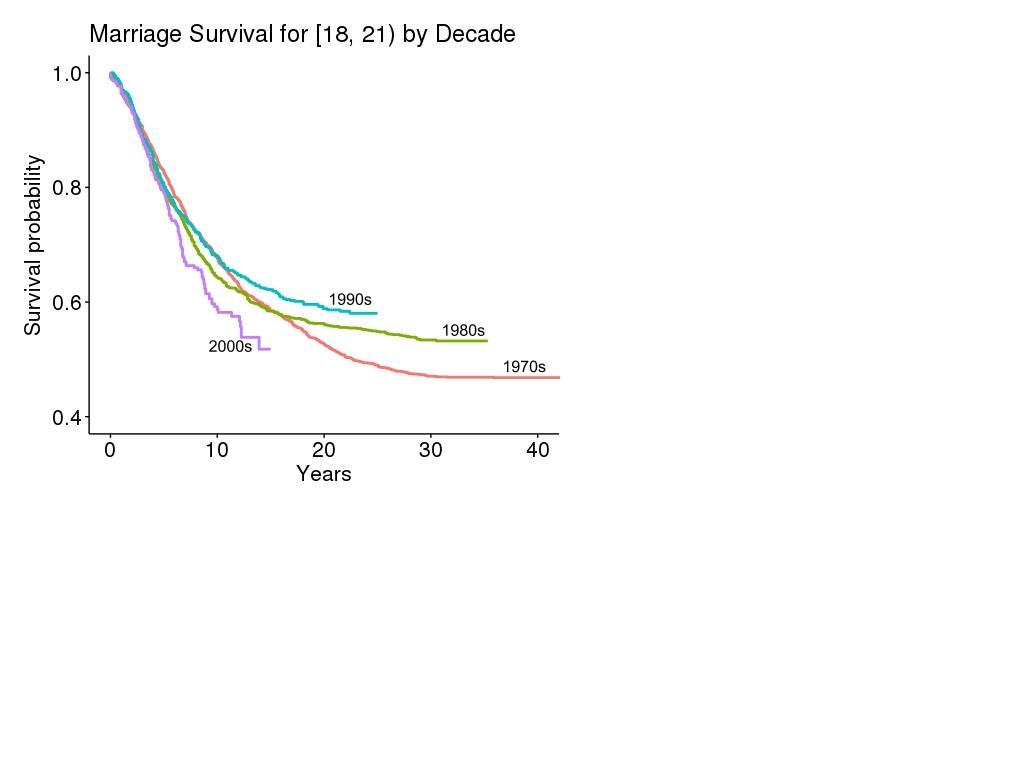
We sought to corroborate our findings and gain more insight into the differences by looking at Kaplan-Meier curves and probability of survival at five-year intervals throughout the marriage (see Appendix B for survival probabilities). We decided to use five-year intervals for a few reasons: first, we wanted an outlier resistant measure due to the large number of right censored observations. Second, the median was an inappropriate measure due to the high survival rate of marriages, the median is an inappropriate statistic for exploring the data.



The Kaplan-Meier curves for age married confirmed the result that we expected based on previous literature: the older the individual was when they began a marriage, the more likely their marriage is to survive. Nor is the improvement to chances minor: one comparison shows that getting married between the ages of 21 and 25 instead of 16 and 18 results in about a 17.5% increase in chances of being together at 10 years and an increase of 22.5% at 20 years. Improvements to marriage survival rates do decline as age increases. There is also the confounding variable of death rates. It is possible that older age groups have better survival rates partially due to the large amount of right-censored data in these groups. However, the trends in our data show more of a difference in marriage survival rates by age group than could be accounted for by increased mortality rates.



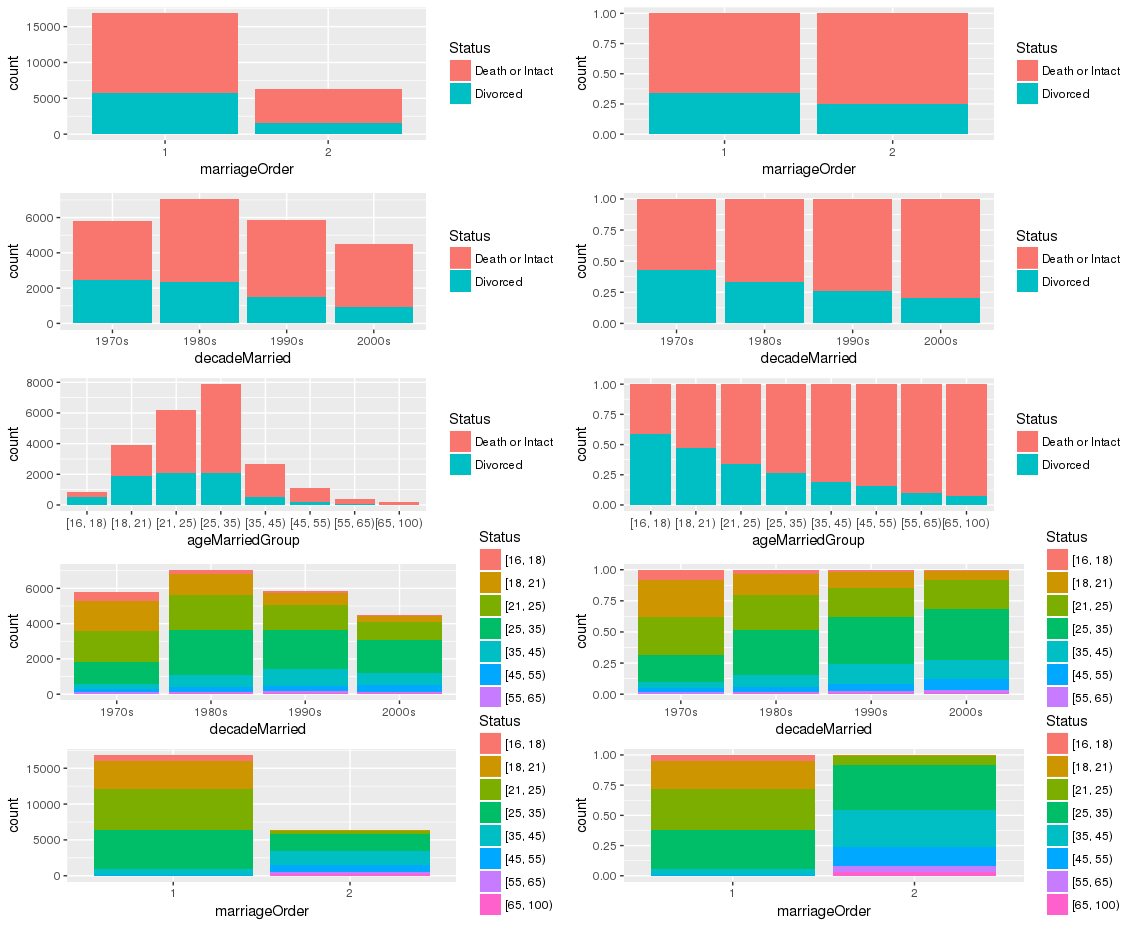
The Kaplan-Meier curve and survival probabilities for marriage order surprised us: it appears that second (and subsequent) marriages have a much higher chance of survival than first marriages. This result contradicted all of the established literature we had read. Upon further reflection, we realized that there was likely a confounding variable at work: second marriages have to take place after first marriages (i.e. when the individual is older), and we had already found that the individual’s age at marriage has a significant impact on the chance that the marriage will survive. To confirm this, we made a histogram looking at the age distribution of first and second marriages (Appendix A). After confirmation, we adjusted our study by creating Kaplan-Meier curves and survival probability tables (Appendix B) with both age married and order of the marriage. Once we did this, it became clear that within most age categories, first marriages had a much better chance of surviving than second marriages; our initial result was simply a case of Simpson’s Paradox.



The Kaplan-Meier curves for decade married showed an interesting trend: marriage survival rate improved by decade until it peaked in the 1990s, and started declining in the 2000s. To verify that we weren’t dealing with another instance of Simpson’s paradox, we graphed survival rates by decade within each age category and found that the trends held in most age categories. Even so, we would not attribute a causal relationship between decade married and survival rates of the marriage - there are too many confounding variables for decade, including economic climate and changes in social norms. The differences in survival rates corresponding to decades might merit speculation about what societal level factors lead to different rates of marital success and serve as a foundation for further investigation.**References**

1. Panel Study of Income Dynamics - Data Center, simba.isr.umich.edu/Zips/ZipMain.aspx.
2. Anderson, Jane. "The impact of family structure on the health of children: effects of divorce." *The Linacre Quarterly* 81, no. 4 (2014): 378-387.
3. Chan, Tak Wing, and Brendan Halpin. “Union Dissolution in the United Kingdom.” *International Journal of Sociology*, vol. 32, no. 4, 2002, pp. 76–93., doi:10.1080/15579336.2002.11770260.
4. Lehrer, Evelyn. “Age at Marriage and Marital Instability: Revisiting the Becker–Landes–Michael Hypothesis.” *Journal of Population Economics*, Springer, 1 Jan. 1970, econpapers.repec.org/RePEc:spr:jopoec:v:21:y:2008:i:2:p:463-484.
5. Poortman, Anne-Rigt, and Torkild Hovde Lyngstad. “Dissolution Risks in First and Higher Order Marital and Cohabiting Unions.” *Social Science Research*, vol. 36, no. 4, 2007, pp. 1431–1446., doi:10.1016/j.ssresearch.2007.02.005.
6. Killewald, Alexandra. "Money, work, and marital stability: Assessing change in the gendered determinants of divorce." *American Sociological Review* 81, no. 4 (2016): 696-719.

**Appendix A**



**Appendix B**

Table 1: Marriage Survival Rates by Age

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Age Group** | **5  Years** | **10  Years** | **15  Years** | **20  Years** | **25  Years** | **30  Years** |
| [16, 18) | 0.752 | 0.581 | 0.484 | 0.440 | 0.411 | 0.405 |
| [18, 21) | 0.809 | 0.660 | 0.587 | 0.545 | 0.520 | 0.502 |
| [21, 25) | 0.862 | 0.758 | 0.700 | 0.666 | 0.643 | 0.633 |
| [25, 35) | 0.891 | 0.805 | 0.754 | 0.727 | 0.712 | 0.707 |
| [35, 45) | 0.911 | 0.845 | 0.813 | 0.800 | 0.792 | 0.786 |
| [45, 55) | 0.915 | 0.867 | 0.847 | 0.831 | 0.823 | 0.820 |
| [55, 65) | 0.943 | 0.909 | 0.893 | 0.893 | 0.893 | 0.893 |
| [65, 100) | 0.941 | 0.929 | 0.922 | 0.922 | 0.922 | 0.922 |

Table 2a: Marriage Survival Rates by Marriage Order

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Marriage Order** | **5  Years** | **10  Years** | **15  Years** | **20  Years** | **25  Years** | **30  Years** |
| First | 0.861 | 0.754 | 0.695 | 0.660 | 0.639 | 0.628 |
| Later | 0.890 | 0.806 | 0.764 | 0.744 | 0.730 | 0.722 |

Table 2b: Marriage Survival Rates by Marriage Order and Age

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Marriage Order** | **Age Group** | **5  Years** | **10  Years** | **15  Years** | **20  Years** | **25  Years** | **30  Years** |
| First | [16, 18) | 0.750 | 0.580 | 0.482 | 0.438 | 0.409 | 0.403 |
| Later | [16, 18) | 1.000 | 0.714 | 0.714 | 0.714 | 0.571 | 0.571 |
| First | [18, 21) | 0.811 | 0.664 | 0.592 | 0.549 | 0.524 | 0.507 |
| Later | [18, 21) | 0.676 | 0.432 | 0.332 | 0.318 | 0.302 | 0.286 |
| First | [21, 25) | 0.863 | 0.760 | 0.704 | 0.670 | 0.647 | 0.637 |
| Later | [21, 25) | 0.853 | 0.729 | 0.647 | 0.618 | 0.592 | 0.578 |
| First | [25, 35) | 0.900 | 0.822 | 0.772 | 0.746 | 0.733 | 0.729 |
| Later | [25, 35) | 0.869 | 0.764 | 0.713 | 0.685 | 0.666 | 0.658 |
| First | [35, 45) | 0.921 | 0.870 | 0.845 | 0.835 | 0.817 | 0.812 |
| Later | [35, 45) | 0.907 | 0.836 | 0.801 | 0.788 | 0.782 | 0.776 |
| First | [45, 55) | 0.937 | 0.880 | 0.848 | 0.848 | 0.848 | 0.848 |
| Later | [45, 55) | 0.912 | 0.865 | 0.847 | 0.828 | 0.820 | 0.816 |
| First | [55, 65) | 1.000 | 0.944 | 0.944 | 0.944 | 0.944 | 0.944 |
| Later | [55, 65) | 0.939 | 0.906 | 0.889 | 0.889 | 0.889 | 0.889 |
| First | [65, 100) | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| Later | [65, 100) | 0.939 | 0.926 | 0.919 | 0.919 | 0.919 | 0.919 |

Table 3a: Marriage Survival Rates by Decade

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Decade** | **5 Years** | **10  Years** | **15  Years** | **20  Years** | **25  Years** | **30  Years** |
| 1970s | 0.854 | 0.734 | 0.662 | 0.611 | 0.582 | 0.571 |
| 1980s | 0.856 | 0.757 | 0.707 | 0.685 | 0.672 | 0.665 |
| 1990s | 0.885 | 0.803 | 0.759 | 0.740 | 0.726 | NA |
| 2000s | 0.886 | 0.785 | 0.729 | NA | NA | NA |

Table 3b: Marriage Survival Rates by Decade and Age

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Decade** | **Age Group** | **5 Years** | **10  Years** | **15  Years** | **20  Years** | **25  Years** | **30  Years** |
| 1970s | [16, 18) | 0.756 | 0.582 | 0.485 | 0.427 | 0.383 | 0.377 |
| 1980s | [16, 18) | 0.793 | 0.638 | 0.536 | 0.498 | 0.486 | 0.481 |
| 1990s | [16, 18) | 0.697 | 0.550 | 0.450 | 0.450 | NA | NA |
| 2000s | [16, 18) | 0.548 | 0.242 | NA | NA | NA | NA |
| 1970s | [18, 21) | 0.825 | 0.676 | 0.584 | 0.526 | 0.490 | 0.471 |
| 1980s | [18, 21) | 0.795 | 0.643 | 0.584 | 0.561 | 0.549 | 0.534 |
| 1990s | [18, 21) | 0.802 | 0.680 | 0.622 | 0.588 | 0.580 | NA |
| 2000s | [18, 21) | 0.789 | 0.587 | 0.518 | NA | NA | NA |
| 1970s | [21, 25) | 0.873 | 0.759 | 0.698 | 0.648 | 0.619 | 0.608 |
| 1980s | [21, 25) | 0.850 | 0.750 | 0.688 | 0.662 | 0.644 | 0.634 |
| 1990s | [21, 25) | 0.857 | 0.765 | 0.720 | 0.698 | 0.683 | NA |
| 2000s | [21, 25) | 0.873 | 0.758 | 0.677 | NA | NA | NA |
| 1970s | [25, 35) | 0.882 | 0.792 | 0.721 | 0.673 | 0.652 | 0.647 |
| 1980s | [25, 35) | 0.876 | 0.788 | 0.745 | 0.722 | 0.711 | 0.706 |
| 1990s | [25, 35) | 0.907 | 0.831 | 0.784 | 0.765 | 0.748 | NA |
| 2000s | [25, 35) | 0.897 | 0.801 | 0.748 | NA | NA | NA |
| 1970s | [35, 45) | 0.898 | 0.814 | 0.765 | 0.742 | 0.726 | 0.718 |
| 1980s | [35, 45) | 0.903 | 0.838 | 0.806 | 0.799 | 0.795 | 0.791 |
| 1990s | [35, 45) | 0.933 | 0.861 | 0.836 | 0.822 | 0.810 | NA |
| 2000s | [35, 45) | 0.895 | 0.842 | 0.799 | NA | NA | NA |
| 1970s | [45, 55) | 0.900 | 0.867 | 0.844 | 0.820 | 0.820 | 0.813 |
| 1980s | [45, 55) | 0.883 | 0.838 | 0.829 | 0.815 | 0.806 | 0.806 |
| 1990s | [45, 55) | 0.930 | 0.901 | 0.878 | 0.866 | 0.841 | NA |
| 2000s | [45, 55) | 0.931 | 0.847 | 0.827 | NA | NA | NA |
| 1970s | [55, 65) | 0.923 | 0.889 | 0.868 | 0.868 | 0.868 | 0.868 |
| 1980s | [55, 65) | 0.916 | 0.882 | 0.871 | 0.871 | 0.871 | 0.871 |
| 1990s | [55, 65) | 0.943 | 0.920 | 0.908 | 0.908 | 0.908 | NA |
| 2000s | [55, 65) | 0.981 | 0.938 | NA | NA | NA | NA |
| 1970s | [65, 100) | 0.871 | 0.836 | 0.801 | 0.801 | 0.801 | 0.801 |
| 1980s | [65, 100) | 0.934 | 0.911 | 0.911 | 0.911 | 0.911 | 0.911 |
| 1990s | [65, 100) | 0.953 | 0.953 | 0.953 | 0.953 | 0.953 | NA |
| 2000s | [65, 100) | 0.978 | 0.978 | NA | NA | NA | NA |

**Appendix C**

marriageData <- **read\_csv**("~/RawData.csv")

*#Relabel the columns of marriageData*  
marriageLabels <-**c**("Observation\_Number","Release\_Number","Interview\_Number\_of\_Individual",

"Person\_Number\_of\_Individual","Individual\_Sex","Month\_Born","Year\_Born",

"Interview\_Number\_of\_Spouse","Person\_Number\_of\_Spouse","Order\_of\_This\_Marriage",

"Month\_Married","Year\_Married","Status\_of\_This\_Marriage","Month\_Widowed\_or\_Divorced",

"Year\_Widowed\_or\_Divorced","Month\_Separated","Year\_Separated",

"Year\_Most\_Recently\_Reported\_Marriage","Number\_of\_Marriages\_of\_this\_Individual",

"Last\_Known\_Marital\_Status","Number\_of\_Marriage\_Records")  
**colnames**(marriageData)<-marriageLabels

*#Removes all observations with unknown order data and people who were never married.*   
*#Removes all observations of marriages that occurred outside of the range 1970-2009*

*#(This also removes all observations where the year married is unknown)*   
*#Removes all observations where the birth year is unknown*  
*#Removes all observations where the status of a marriage is "Don’t know", multiple concurrent marriages, never married, or separated.*

*#Removes all observations with unknown year of divorce or widowing.*   
marriageData <- marriageData %>%   
  **filter**(Order\_of\_This\_Marriage < 14) %>%  
  **filter**(Year\_Married >= 1970 & Year\_Married < 2010) %>%  
  **filter**(Year\_Born < 3000)  %>%  
  **filter**(Status\_of\_This\_Marriage < 5) %>%  
  **filter**(Year\_Widowed\_or\_Divorced != 9998)

*# If the month of birth is unknown, set Month\_Born to 1.  Otherwise, leave it unchanged*  
marriageData$Month\_Born <-

**ifelse**(marriageData$Month\_Born == 98, 1, marriageData$Month\_Born)  
*#If the month of marriage is unknown, set Month\_Married to 1.  Adjust for the seasons: Winter = 1, Spring = 4, Summer = 7, Fall = 10. Otherwise, leave it unchanged*  
marriageData$Month\_Married <-

**ifelse**(marriageData$Month\_Married == 98, 1,

**ifelse**(marriageData$Month\_Married == 21, 1,

**ifelse**(marriageData$Month\_Married ==22, 4,

**ifelse**(marriageData$Month\_Married == 23, 7,

**ifelse**(marriageData$Month\_Married == 24, 10,

marriageData$Month\_Married)))))  
  
*# If the month of divorce or widowing is unknown, set Widowed\_or\_Divorced to 1. Adjust for the seasons: Winter = 1, Spring = 4, Summer = 7, Fall = 10.*

*Otherwise, leave it unchanged*  
marriageData$Month\_Widowed\_or\_Divorced <-

**ifelse**(marriageData$Month\_Widowed\_or\_Divorced == 98, 1,

**ifelse**(marriageData$Month\_Widowed\_or\_Divorced == 21, 1,  
**ifelse**(marriageData$Month\_Widowed\_or\_Divorced ==22, 4,

**ifelse**(marriageData$Month\_Widowed\_or\_Divorced == 23, 7,

**ifelse**(marriageData$Month\_Widowed\_or\_Divorced == 24, 10,

      marriageData$Month\_Widowed\_or\_Divorced)))))

*#Alters Order\_Of\_This\_Marriage so that it is 1 if it is a first marriage and 2 if it is*

*subsequent marriage*  
marriageData$marriageOrder <-

**as.factor**(**ifelse**(marriageData$Order\_of\_This\_Marriage > 1, 2, 1))

*#Creates a variable for the year and month an individual was born*  
marriageData$birthYrM <-

marriageData$Year\_Born+((marriageData$Month\_Born-1)  /12)  
*#Creates a variable for the year and month an individual got married*  
marriageData$marriageYrM <-

marriageData$Year\_Married + ((marriageData$Month\_Married-1) /12)  
  
*#Creates a variable for the age an individual was when they got married*  
marriageData$ageMarried <- marriageData$marriageYrM - marriageData$birthYrM   
*#Creates a variable for the year and month an individual got divorced or widowed.*  
marriageData$divorcedYrM <- marriageData$Year\_Widowed\_or\_Divorced +

((marriageData$Month\_Widowed\_or\_Divorced-1) /12)  
*#Removes all observations of individuals who were married at age 16 or younger*  
marriageData <- **filter**(marriageData, ageMarried >= 16)  
*#Creates a variable that assigns individuals to groups as described in the codebook*  
marriageData$ageMarriedGroup<-  
   **as.factor**(**ifelse**(marriageData$ageMarried < 18,  
         "[16, 18)",  
         **ifelse**(marriageData$ageMarried < 21,  
         "[18, 21)",  
         **ifelse**(marriageData$ageMarried < 25,  
         "[21, 25)",  
         **ifelse**(marriageData$ageMarried < 35,  
         "[25, 35)",  
         **ifelse**(marriageData$ageMarried < 45,  
         "[35, 45)",  
         **ifelse**(marriageData$ageMarried < 55,  
         "[45, 55)",  
         **ifelse**(marriageData$ageMarried < 65,  
         "[55, 65)",  
         "[65, 100)"))))))))  
  
*#Creates a variable that groups individuals based on their sex*  
marriageData$sex <-

**as.factor**(**ifelse**(marriageData$Individual\_Sex ==1, "Male","Female"))  
  
*#Creates a variable that groups individuals by what decade they were married*   
marriageData$decadeMarried <-

**as.factor**(**ifelse**(marriageData$Year\_Married < 1980,  
         "1970s",  
         **ifelse**(marriageData$Year\_Married < 1990,  
         "1980s",  
         **ifelse**(marriageData$Year\_Married < 2000,  
         "1990s", "2000s"))))  
  
*#Creates a boolean variable that is true if an individual was widowed or*

*their marriage is still intact and false if they have gotten divorced*  
marriageData$isCensored <- marriageData$Status\_of\_This\_Marriage < 4

*#Creates a variable that records the length of an individual's marriage*  
marriageData$marriageLength <-  
  **ifelse**(marriageData$isCensored & marriageData$Status\_of\_This\_Marriage == 1,  
         2015 - marriageData$marriageYrM,

        marriageData$divorcedYrM - marriageData$marriageYrM)  
  
*#Setting all negative marriage lengths to 0 to account for any individuals*

*who were married and divorced in the same year, but didn't report their month of marriage*  
marriageData$marriageLength <-

**ifelse**(marriageData$marriageLength < 0, 0, marriageData$marriageLength)

*# Selects final variables of interest and intermediate variables*  
FinalData <-

**select**(marriageData, marriageLength, isCensored, sex, decadeMarried,

ageMarriedGroup, marriageOrder, birthYrM, marriageYrM, divorcedYrM,

ageMarried)  
**write.csv**(FinalData, "FinalData.csv")

1. \* Age group when married refers to the age of the person who answered the survey, and does not take into account their spouse’s age. The same is true for order of marriage. [↑](#footnote-ref-1)
2. We also defaulted to the first month of the season if season was provided instead of month. [↑](#footnote-ref-2)