BST 222 : Analysis of Factors Affecting Observed Breastfeeding Durations

Elsie Basa, Shirley Lin, Laura Wang

1 Introduction

1.1 Background

For newborns, current recommendations by the CDC are to primarily breastfeed a child for the first 6 months of their life and then move to gradually introduce solid foods into their diet. Previous studies have reported the mean and median breastfeeding times to be about 5.7 and 6 months respectively (Wasie Kasahun et al.). However, breastfeeding can continue for as long as the mother and child feel like they can continue for. In general, there are few drawbacks to breastfeeding for longer times, with some recommendations suggesting continuing breastfeeding for as long as 2 years ("Breastfeeding FAQs: How Much and How Often (for Parents) - Nemours KidsHealth"). Breastfeeding is considered beneficial for children, protecting them from certain diseases and past studies have shown that breastfed babies have lower risks of asthma, obesity, and sudden infant death syndrome (SIDS) ("Breastfeeding Benefits Both Baby and Mom | DNPAO | CDC"). While there are certainly multiple benefits to breastfeeding newborns, there's a lot of variability in the time until a mother chooses to wean her child. In some cases, mothers will only breastfeed their children for approximately a week before weaning them. The primary purpose of this analysis is to identify trends in the duration of the breastfeeding period as well as identify variables that appear to be associated with these trends.

1.2 Data Description

The data was sourced from the `KMsurv` package in R. The original data is from the National Longitudinal Survey of Youth (a survey which began in 1979 and ended in 1988). The package data is a cleaned and subsetted version of the original data.

Beginning in 1983, the women in the study population were asked about any pregnancies they had and various data about them were collected. For this dataset, only first-born children born after 1978 at a gestational age of 20 to 45 weeks were included. Then, the data was further narrowed down to only include responses from mothers who chose to breastfeed their children. After taking all of these factors into account, we end up with 927 total observations. For this analysis, the response (survival) is generated using the breastfeeding duration (weeks) and an indicator for whether or not the child was weaned.

The variables included in this dataset are shown in Table 1.

Table 1: Factor Levels for Categorical Variables

race	Race of mother (1=white, 2=black, 3=other)
poverty	Mother in poverty (1=yes, 0=no)
smoke	Mother smoked at birth of child (1=yes, 0=no)
alcohol	Mother used alcohol at birth of child (1=yes, 0=no)
agemth	Age of mother at birth of child
ybirth	Year of birth
yschool	Education level of mother (years of school)
pc3mth	Prenatal care after 3rd month (1=yes, 0=no)

It is important to note that, for this dataset in particular, formula feeding is not considered breastfeeding despite formula being a form of liquid nourishment for infants. The indicator for poverty only indicates whether or not the mother was in poverty at the time of birth of her child. Table 2 shows some very basic summary statistics for the outcome variables (`duration` and `delta`). When `delta` takes on a value of 1, it means that the child completed breastfeeding (was weaned) at the end of the reported period of time. Otherwise, the value of 0 meant that the information was either censored (unobserved) or there was loss to follow-up. All of the children who were reported to not have finished breastfeeding yet were born during or after 1984.

Table 2: Summary of Data Outcome

bfeed Outcome Summary	**N = 927**
duration (weeks)	
Mean	16
Median	10
Minimum-Maximum	1-192
delta (=1)	892 (96%)

[1] n (%)

Table 3: Summary of Data by Variable

bfeed Variable Summary	**0**, N = 35	**1**, N = 892
race		
white	28 (80%)	634 (71%)
black	4 (11%)	113 (13%)
other	3 (8.6%)	145 (16%)
yschool		
noHS	1 (2.9%)	219 (25%)
HSgrad	13 (37%)	425 (48%)
someCollege	21 (60%)	248 (28%)
poverty	3 (8.6%)	168 (19%)
alcohol	3 (8.6%)	76 (8.5%)
smoke	7 (20%)	263 (29%)
agemth		
Mean (SD)	25 (2)	21 (3)
pc3mth	8 (23%)	156 (17%)

[1] n (%)

General guidelines for pregnant mothers recommend seeking prenatal care after the third month of pregnancy. However, as seen in Table 3, only approximately 18% of the mothers chose to seek out

prenatal care after the third month of their pregnancies. Also, while a fairly low percentage of the mothers consumed alcohol around the time of the birth of their child, approximately 30% of them reported smoking at the time. However, the degree of their smoking, such as how much or how often they smoked, wasn't included in the dataset. Visual representations of these summary tables can be found in the exploratory data analysis portion of the paper.

2 Exploratory Data Analysis

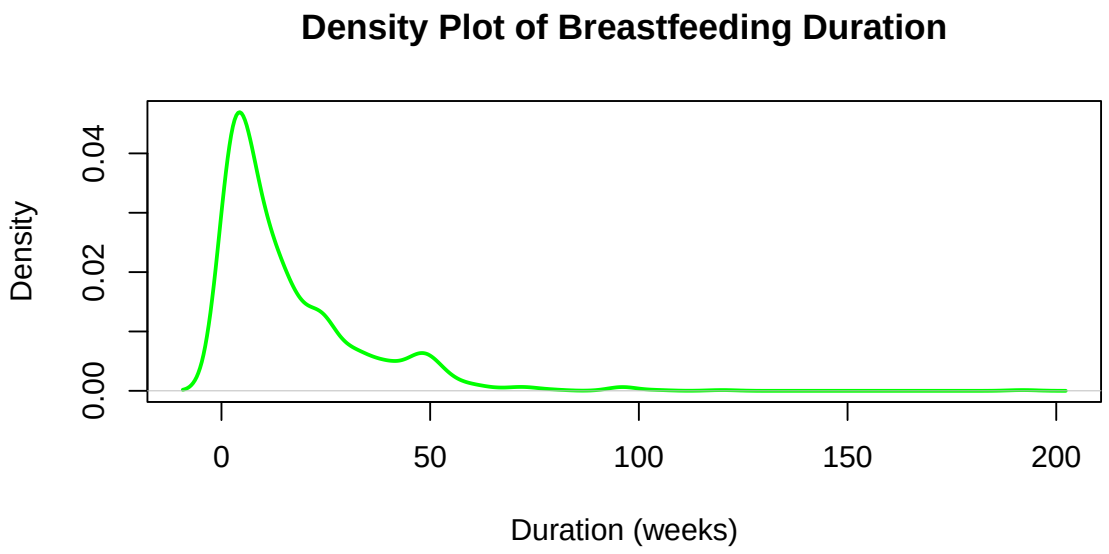


Figure 1: Density Plot of Breastfeeding Duration

Figure 1 is a density plot shows that almost all of the mothers in the study only breastfeed for no more than about 50 weeks, which is about a year. On the graph there is a spike at around 10-12 weeks indicating that the majority of mothers only breastfeed for about a few months.

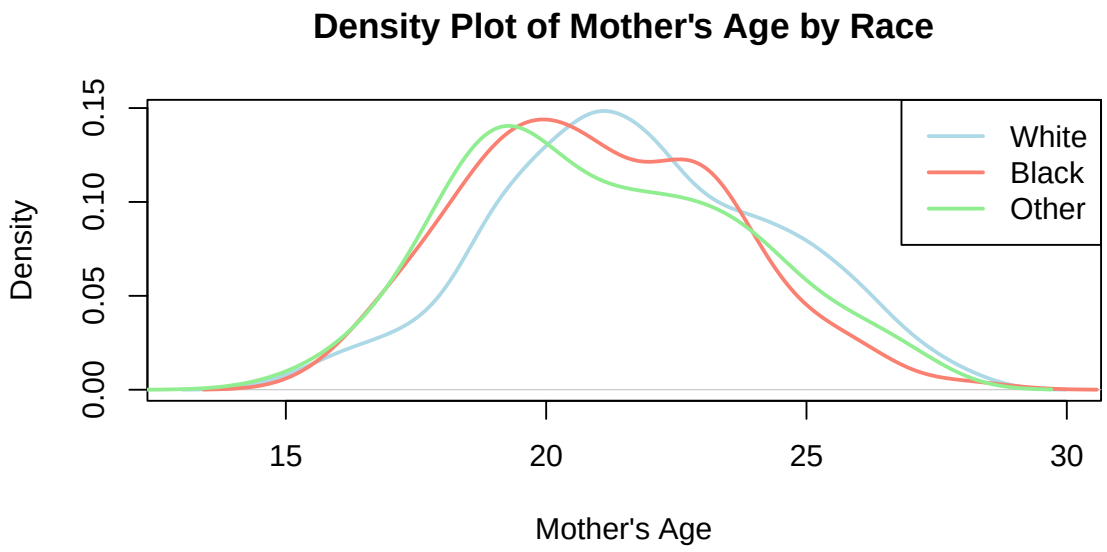


Figure 2: Density Plots Mother's Age by Race

Figure 2 shows the density of the mothers' age by race demonstrates that black mothers and mothers of other races tend to have their first kids younger than white mothers.

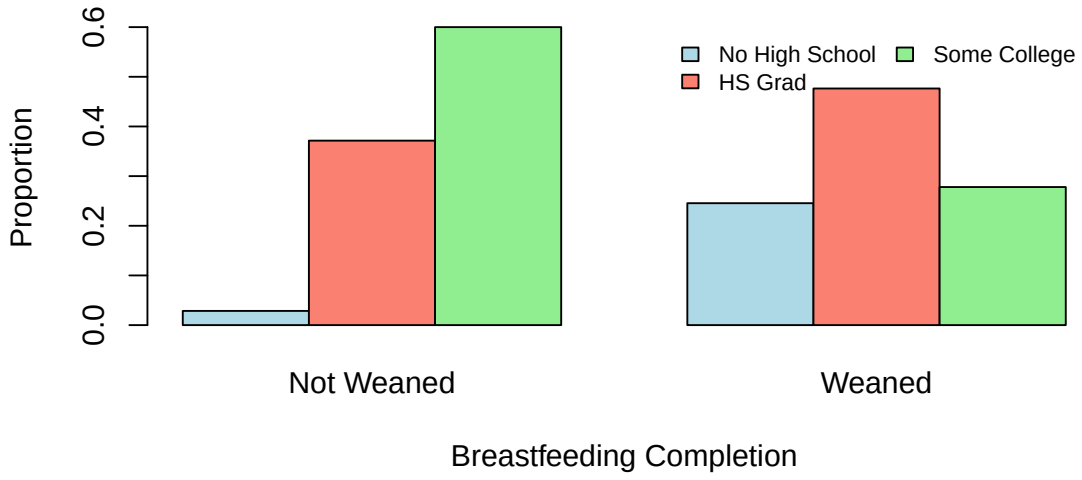


Figure 3: Bar Plot on Education proportion By Breastfeeding Status

Figure 3 compares the group of mothers who weaned their babies versus those who did not wean their babies. Those who didnn't wean their children had higher proportion of mothers who went some college in their education background. Mothers who weaned their babies had a higher proportion of those who did not finish high school.

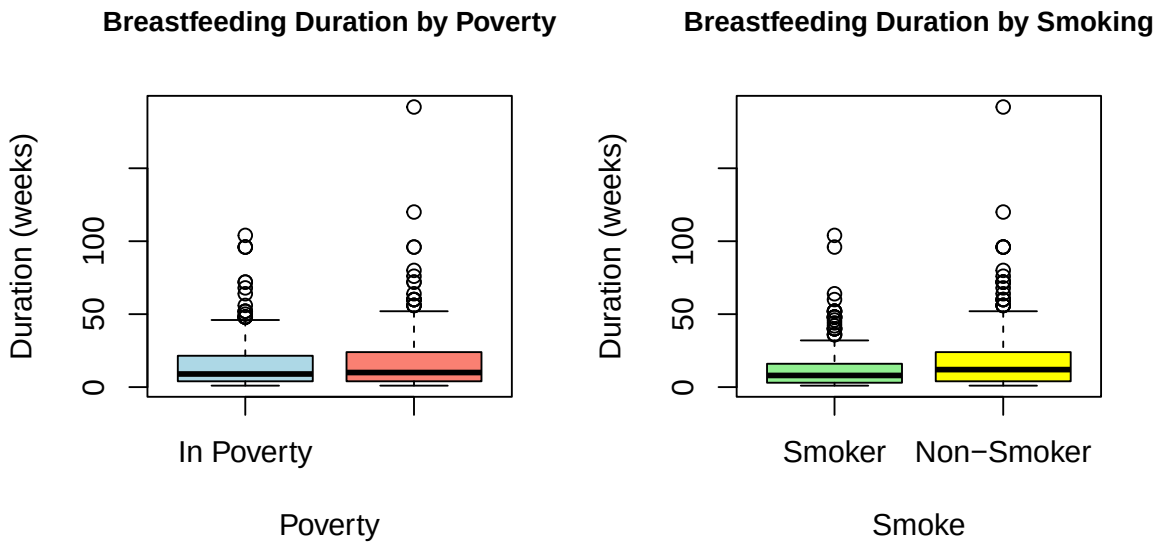


Figure 4: Box Plot for Categorical Variables

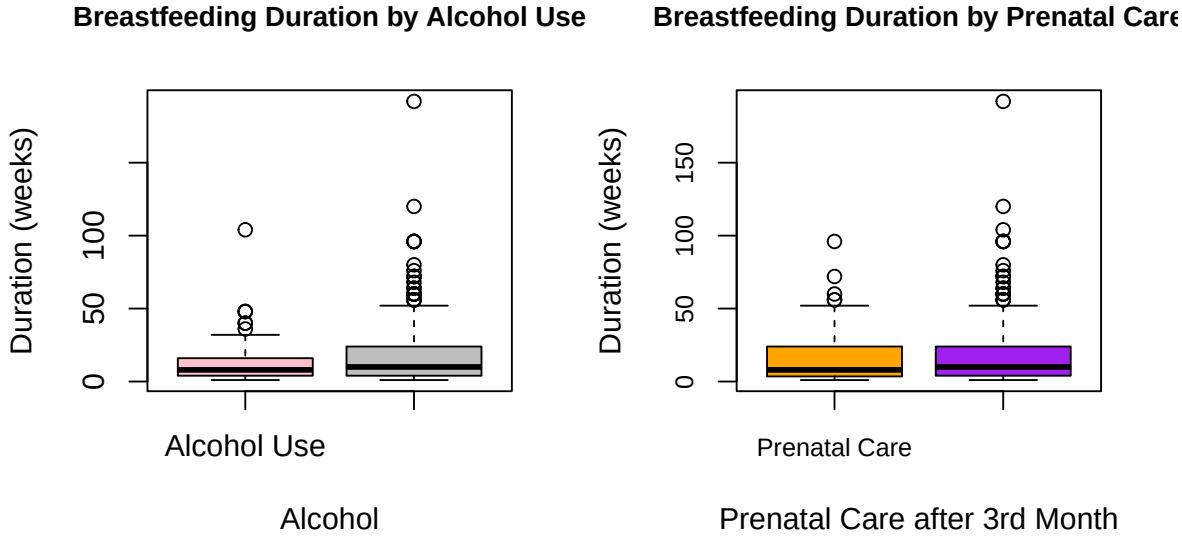


Figure 5: Box Plot for Categorical Variables

In figure Figure 4, the first box plot shows that the breastfeeding duration for those in poverty compared to those who are not in poverty are similar but those in poverty have a slightly lower and narrower breastfeeding duration range. The second box plot shows that mothers who smoked had a sligtly shorter breastfeeding duration and that the range for breastfeeding duration is also narrower compared to mothers who did not smoke.

In figure 5, the first box plot, demonstrates that mothers who use alcohol have a narrower breastfeeding distribution with most mothers breastfeeding betwween 0 and under 50 weeks. Mothers who did not consume alcohol tend to breastfeed between 0 and 52 weeks. Those mothers who consumed alcohol tend breastfeed for a shorter amount of time compared to those who did not. The second box plot

showed that the breastfeeding duration distribution of mothers who had prenatal care compared to those are very similar, with majority of both groups mainly brestfeeding for 0 weeks to 50 weeks.

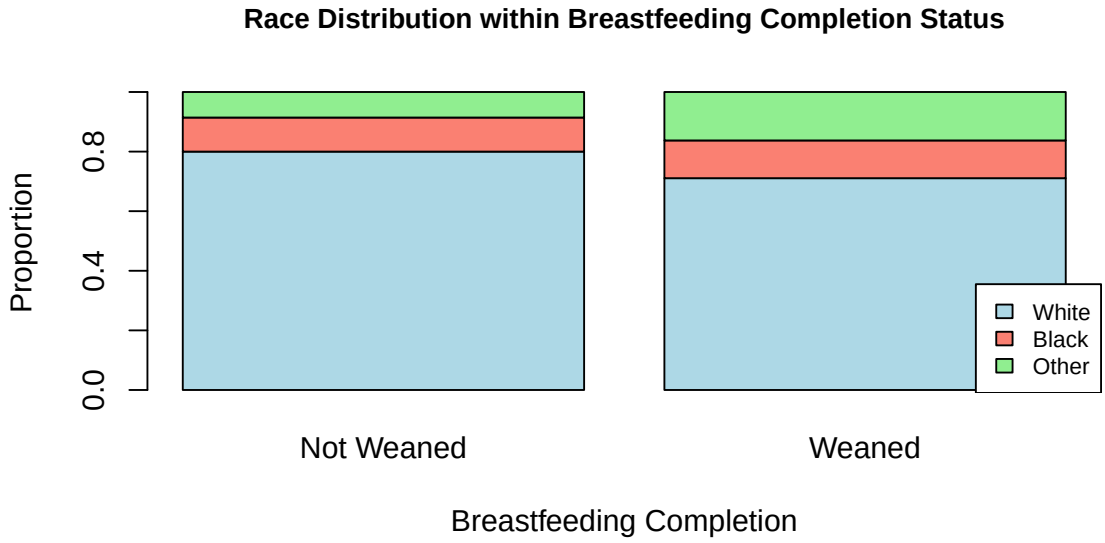


Figure 6: Box Plot of Race Distribution by Breastfeeding Status

Figure 6 compares the proportion of each race by breastfeeding status. The group of mothers who were able to wean have a high proportion of mothers of other race and black mothers compared to the group that did not wean.

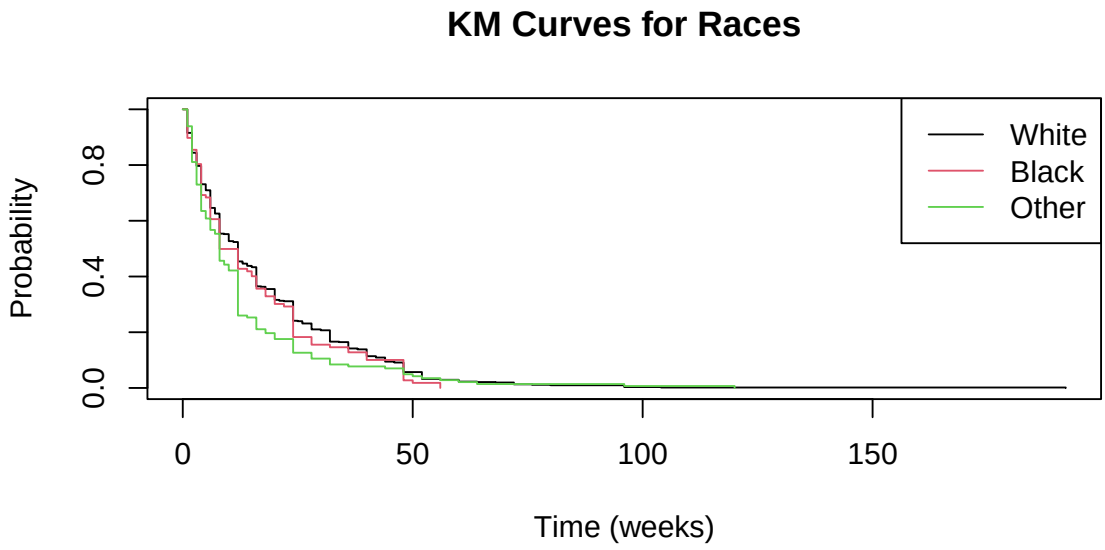


Figure 7: KM Curve - Race

Figure 7 compares the probability a mother doesn't wean for each of the different race categories. For each race, there is a steep decline around the 10-20 week mark indicating many mothers stop

breastfeeding around that time. Additionally, mothers of other race tend to wean sooner then black mothers and white mothers. Black mothers also weaned faster than white mothers.

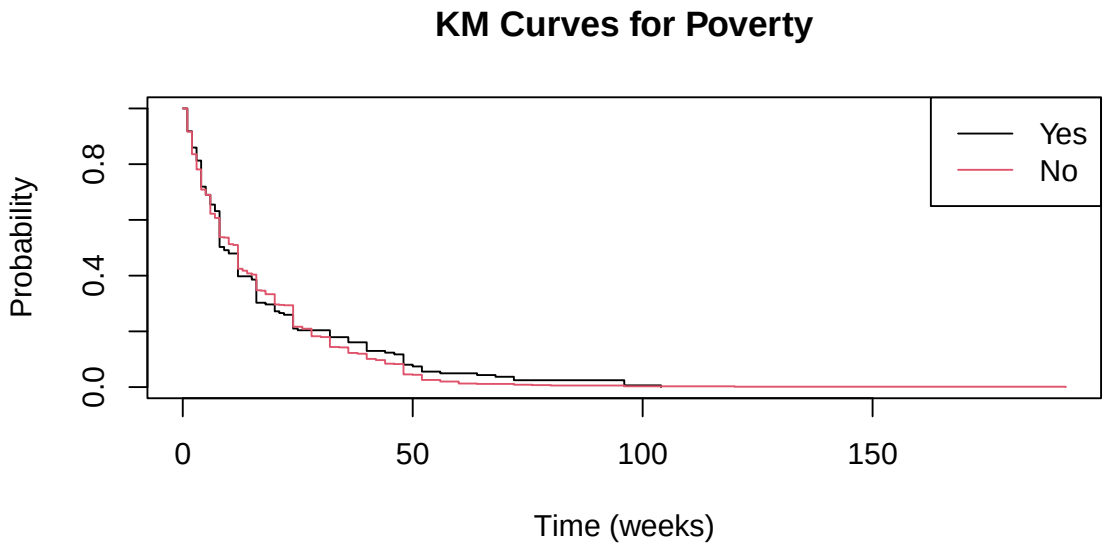


Figure 8: KM Curve - Poverty

Figure 8 compares the probability a mother doesn't wean depending if they are in poverty or not. For both groups, there is a steep decline around the 10-20 week mark indicating many mothers stop breastfeeding around that time. The Kaplen Meier curves are similar and do overlap til around week 30 where it can be seen that mother's who are not in poverty have a higher probability of weaning their child.

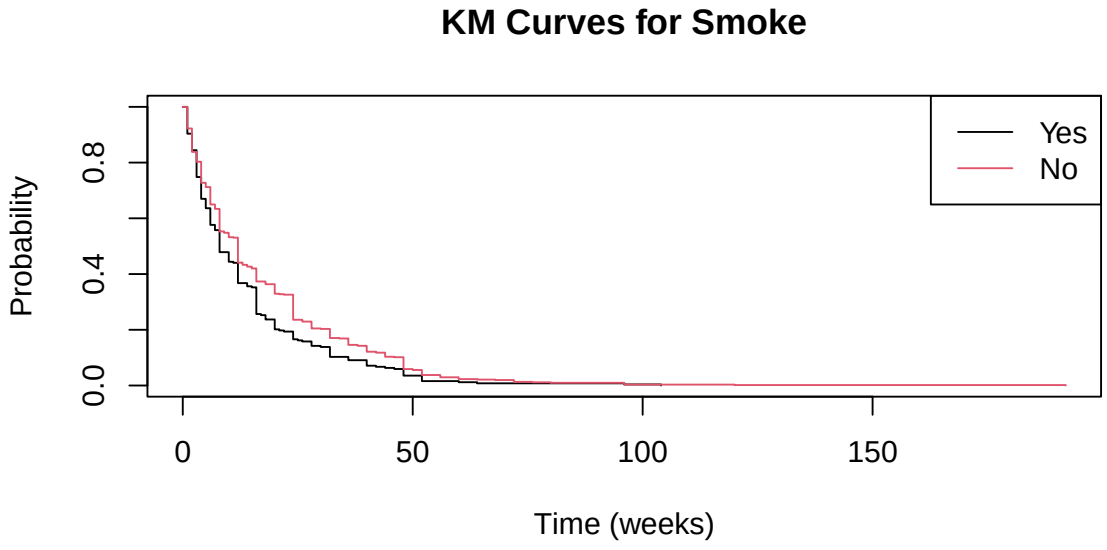


Figure 9: KM Curve - Smoke

Figure 9, compares the probability a mother does not wean for mothers who smoke and mothers who do not smoke. The curves have a steep decline around the 10-20 week mark indicating that many

mothers stop breastfeeding around that time period. The Kaplen Meier curves indicate that there is a higher probability that mothers who smoke will wean compare to those who do not smoke.

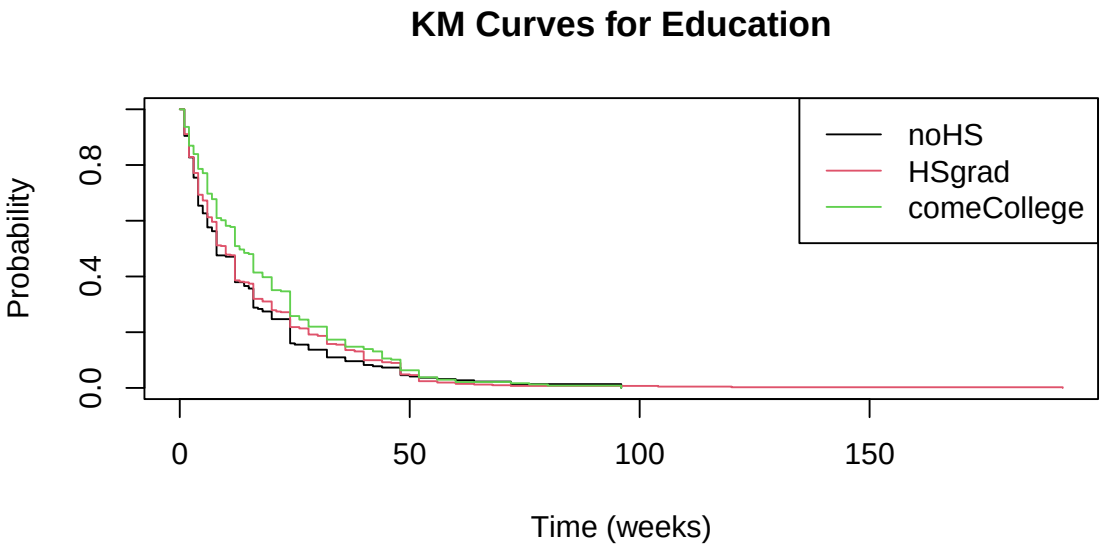


Figure 10: KM Curve - Education

Figure 10, compares the probability a mother does not wean for mothers in different education levels. The curves have a steep decline around the 10-20 week mark indicating that there is a low probability that mothers continue breastfeeding around that time period. Mothers with some college education are more likely to not wean compare to mothers with high school education and mothers with no high school education. For the KM curves for mothers with no high school education and those with only a high school degree have similar likelihoods of not weaning their child till around week 20. Around that time those without a high school education are more likely to wean than those without one.

3 Model Building

Cox proportional hazard regression models are used to analyze the weaning behavior because the coefficients can be used to quantify the effect and work for both quantitative and categorical variables simultaneously.

A survival object is created with the duration of breastfeeding in weeks and an indicator of completed breastfeeding as the event. This survival object is used throughout the model-building process. We built our model using backward elimination, which means we started with a full model that included all potential variables that could influence the likelihood of completing breastfeeding. We then removed the least significant variables based on AIC from drop1 and p values from the model's summary table each time. The ultimate goal was to keep only the most significant terms in our model. We evaluated the model using AIC, BIC, and global test results, including log-likelihood, Wald test, and Score test. For further improvement, we also explored potential interactions between significant terms.

Our full model with all variables has an AIC value of 10369.51 and a BIC value of 10412.65. From the full model, other race has a p-value of 0.00134, not in poverty has a p-value of 0.02986, not smoking has a p-value of 0.00191, and education level of getting some college has a p-value of 0.02702 according to our findings. Removing these variables will result in higher AIC values. However, the p-values of all the other variables are greater than the significance level of 0.05, and eliminating them will reduce the AIC. After removing insignificant terms repeatedly, the best model's remaining variables include race, poverty, and smoking. The AIC for this model is 10366.19, and the BIC is 10385.37. It is worth noting

that, while the year of school appears to be significant in the full model, it is no longer significant when combined with the three other significant variables. When the interactions are added, we did not observe either a drop in AIC value or an additional significant terms, so they do not appear to be essential in explaining mothers' likelihood of weaning.

However, the issue of multicollinearity arose when we examined the correlation among the three most significant variables in the model. We found a significant correlation between each pair of variables related to race, poverty, and smoke when we used Pearson's chi-squared test on a contingency table to assess the correlation between the categorical variables. The chi-squared test p-values for the correlation were 3.407e-05 for poverty and race, 1.009e-10 for race and smoke, and 0.006169 for poverty and smoke. Race and smoke had the lowest p-value, which indicates the strongest association between them.

As a result, even though this model has the lowest AIC and BIC values, multicollinearity severely restricts its applicability in real-world scenarios. This may result in problems like inflated standard errors and trouble determining the relative contributions of different variables. To address this problem and provide a more accurate analysis, we built three distinct models, each concentrating on one of the correlated variables, which are poverty, smoke, and race. In addition, we also checked if other variables are significant when modeled by itself.

4 Results

4.1 Results Summary

Table 4: Coefficients for Race Model

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
raceblack	0.1105978	1.116946	0.1023577	1.080503	0.2799182
raceother	0.2543888	1.289673	0.0924692	2.751066	0.0059402

Table 5: Coefficients for Smoke Model

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
smokeno	-0.2270217	0.7969035	0.073742	-3.078593	0.0020798

Table 6: Coefficients for Year of School Model

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
yschoolHSgrad	-0.0755235	0.9272579	0.0834309	-0.9052227	0.3653474
yschoolsomeCollege	-0.2331438	0.7920397	0.0928262	-2.5116177	0.0120179

4.2 Results Interpretation

The p-value for other races is 0.00594, which is lower than our significance level of 0.05. The coefficient for `raceother` is 0.25439, which implies that when holding other variables constant, mothers of other races are approximately exp(0.25439) =1.28967 times more likely to wean than our reference group of white mothers. For the same reason, black mothers are approximately exp(0.11060) =1.11695 times more likely to wean than white mothers. However, the difference between black and white mothers is not as significant as that between mothers of other races and white mothers.

After that, not smoking has a coefficient of -0.22702. This indicates that when holding other variables constant, mothers who did not smoke during the born of their first child have a hazard approximately exp(-0.22702) = 0.79690 times that of mothers who reported smoking with a p-value of 0.00208.

We found that when only having poverty in the model, the p-value is 0.379, which is larger than the significance level. Although a 1.1681 times likelihood to wean is observed in mothers not in poverty in the reduced model with race, smoke, and poverty, the effect of poverty disappears in the new model. Therefore, mothers' financial situation at the time of their first child's birth is not an important indicator of weaning by itself and is removed from the significant models for this study.

In addition to the variables identified as significant in the best reduced model, we also discovered that the degree of education of mothers has an impact on the probability of breastfeeding completion. The summary table reveals that mothers with some college education are exp(-0.0755235)=0.7920397 times as likely to wean compared to the reference level of mothers who did not complete high school education, with a p-value of 0.0120179. Similarly, mothers who completed high school but did not attend college exhibit a likelihood of exp(-0.2331438)=0.9272579 to wean compared to those who did not complete high school education. However, this difference is not statistically significant in comparison to the preceding comparison.

5 Model Checking

5.1 Assumption

The Schoenfeld residual plots and the cox.zph function are used to test the proportional hazards assumption for the Cox regression model. The models with race, smoke, and year of school have p-values of 0.29, 0.83, and 0.055 respectively, all of which are greater than 0.05.

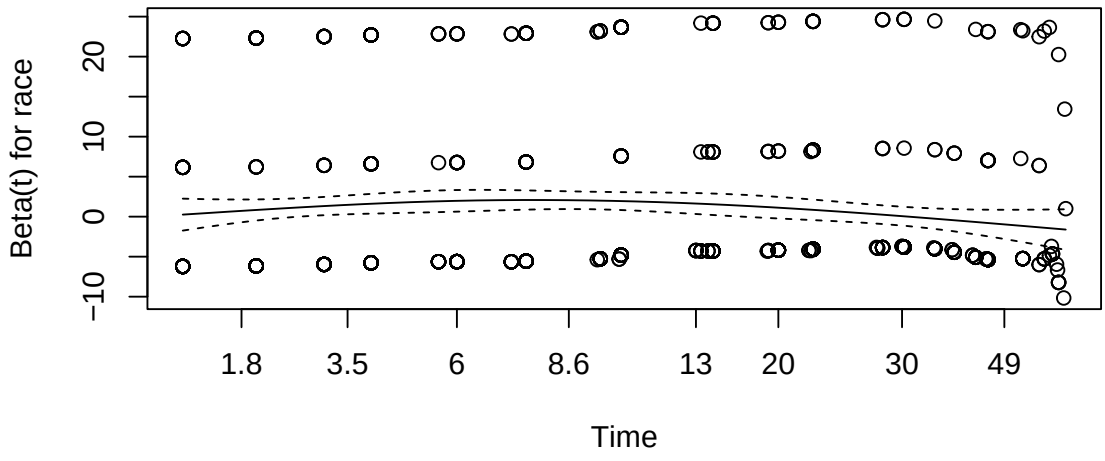


Figure 11: Schoenfeld Residuals for Race

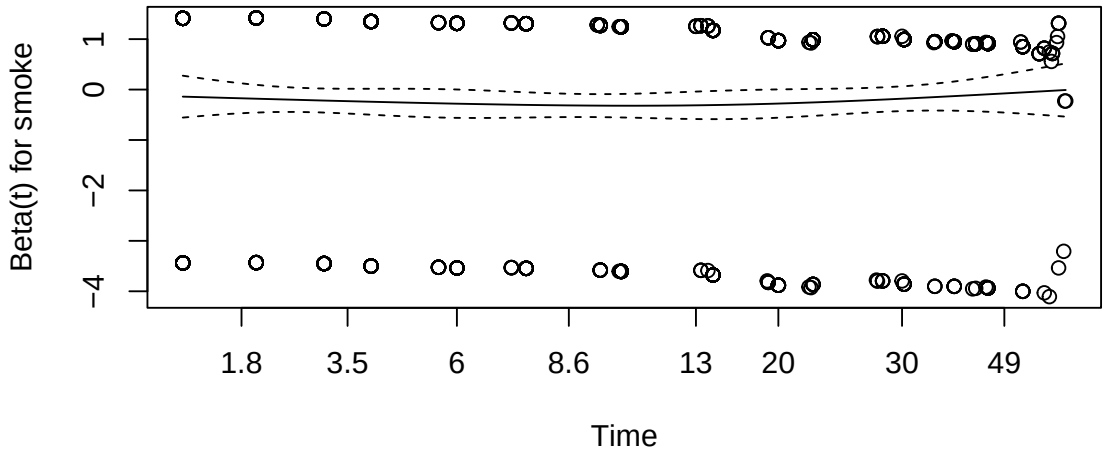


Figure 12: Schoenfeld Residuals for Smoke Status

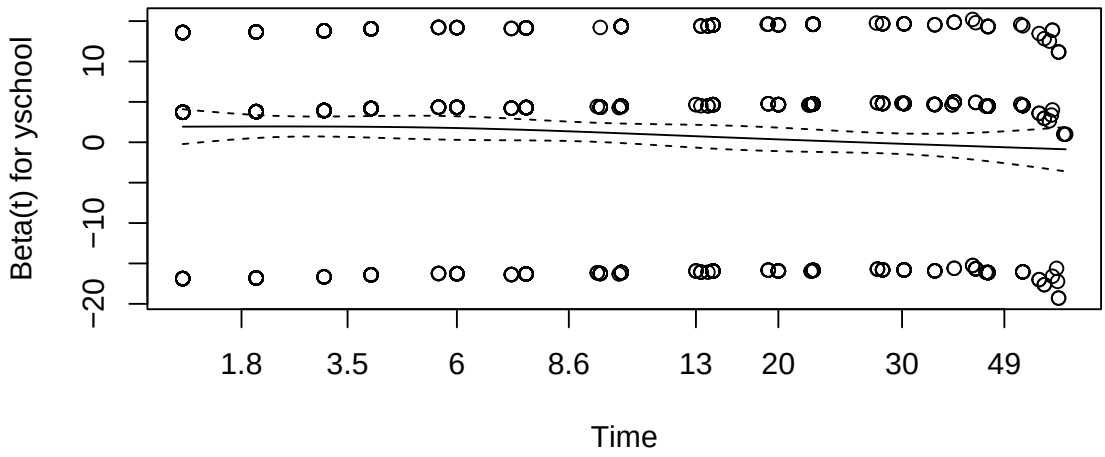
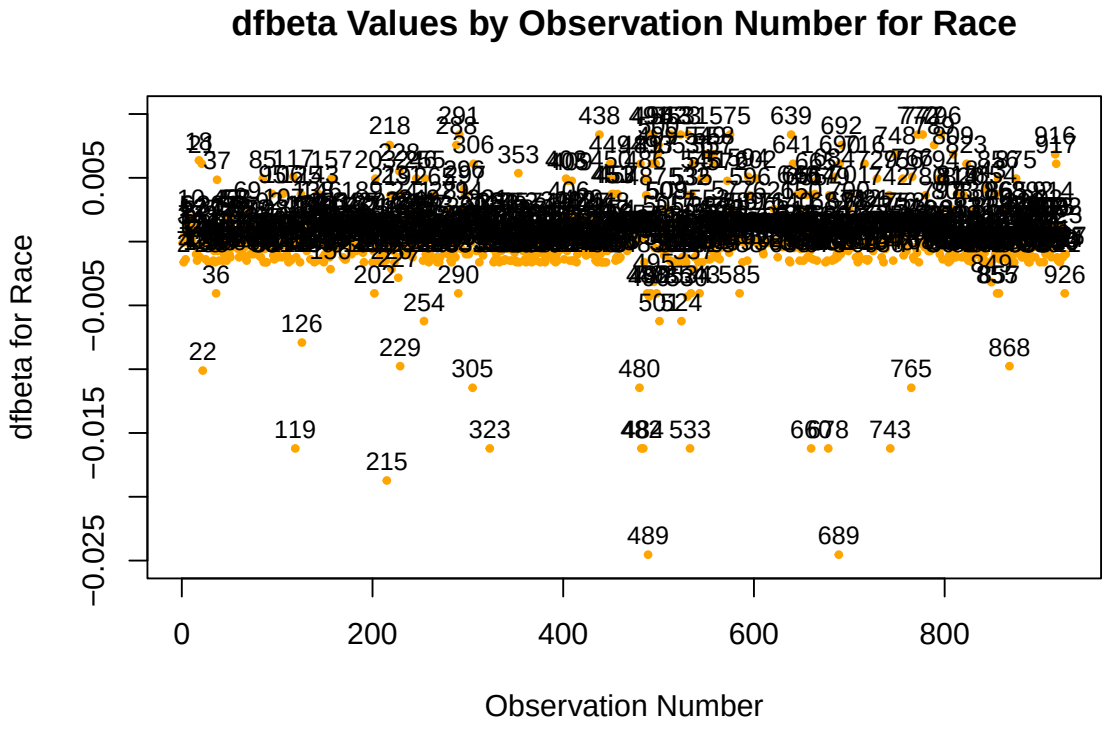
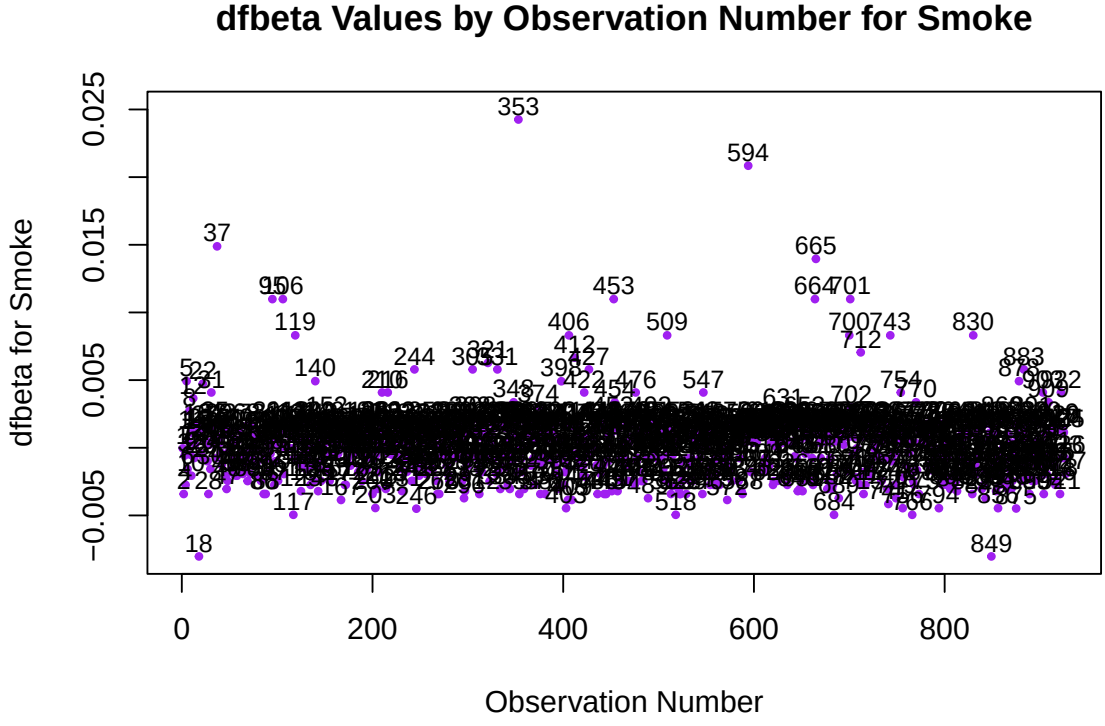


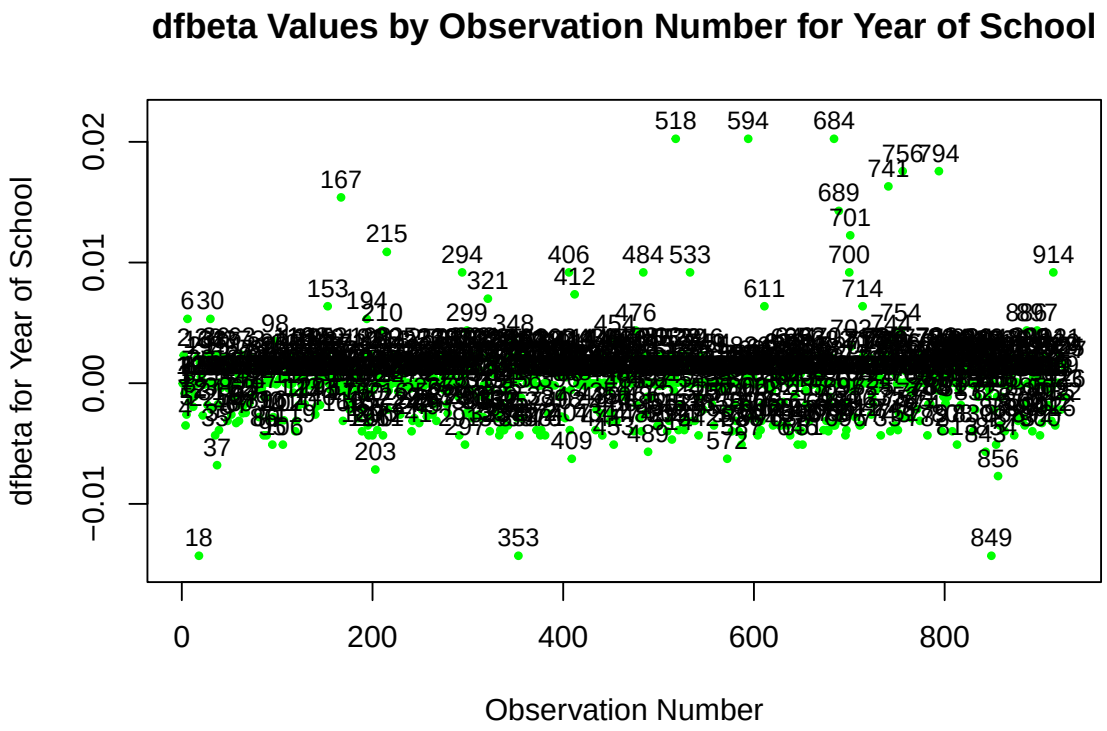
Figure 13: Schoenfeld Residuals for Year of School

The Schoenfeld residual plots shown in Figures 11, 12, and 13 do not exhibit any clear signs of non-proportionality because the residuals are randomly distributed around the y=0 lines in both plots. These results suggest that the evidence is insufficient to conclude that the hazards are non-proportional for the two models.

5.2 Outliers


dfbeta Values by Observation Number for Race
dfbeta for Race
0.005
−0.005
−0.015
−0.025
0
200
400
600
800
Observation Number


dfbeta Values by Observation Number for Smoke
dfbeta for Smoke
0.025
0.015
0.005
−0.005
0
200
400
600
800
Observation Number



We used DFBETA to assess how the exclusion of a specific data point impacts the regression coefficients. The larger the absolute value of DFBETA is, the more influential the point becomes and the greater the probability that it is an outlier.

Observations 18, 353, 489, 518, 594, 684, 689, and 849 in our analysis deviate from the majority of the data points, indicating that these observations are likely to be outliers and need further investigation. Table 7 displays the detailed information of these unusual observations and will be discussed later.

Table 7: Unusual Observations Detail

	duration	delta	race	poverty	smoke	alcohol	agemth	ybirth	yschool	pc3mth
18	192	1	white	no	no	no	21	78	HSgrad	no
353	104	1	white	yes	yes	yes	20	83	HSgrad	no
489	56	1	black	no	no	no	23	84	HSgrad	yes
518	96	1	other	yes	no	no	19	79	noHS	no
594	96	1	white	yes	yes	no	18	78	noHS	yes
684	96	1	white	yes	no	no	19	80	noHS	no
689	56	1	black	yes	no	no	17	80	noHS	no
849	120	1	other	no	no	no	22	80	HSgrad	no

6 Conclusion

As shown by the multiple single-variate models, the mother's race, smoking status, and years of education played a role in determining whether or not they weaned their child. While performing the analysis, the correlation between poverty status and the race of the mother had to be considered. Because the data is from the 70s, it makes sense, intuitively, that race is associated with poverty status. Therefore, the poverty status appears to be significant when considered in a model with race but rather insignificant otherwise. Additionally, smoking status was also slightly correlated with race

but not to the same extent. To take these relationships into account, the conclusions regarding the associations between smoking status and race with breastfeeding duration were drawn from hazard ratios from separate models. This was so we could have the most accurate final results. `yschool` was rather significant in the "full" model but it was also important to consider it within the context of the mother's age. The age of the mother and her years of education have a non-randomly correlated relationship. A person cannot have more years of education than she is old. However, in the context of this analysis, the age of the mother didn't appear to be a significant influence on breastfeeding duration on its own or in tandem with the years of education. Generally, it can be concluded that smoking mothers tend to wean their children faster than non-smoking mothers and that non-white mothers had a tendency to wean sooner than white mothers.

6.1 Discussion

There are a few possible explanations for the results. First of all, in regards to the trends with the smokers versus non-smokers, previous research has shown that smoking mothers' breast milk tends to dry up sooner than smoking mothers. This can can essentially force them too wean their child after a certain point as they can no longer provide any breast milk ("Tobacco and E-Cigarettes | Breastfeeding | CDC"). As our findings follow similar trends, with smoking mothers weaning sooner, this is a viable explanation for the findings. In the case of the mother's race being significant, perhaps it is a combination of things such as socioeconomic status, occupation, and other lifestyle factors that make a difference. As mentioned before, in the 1970s, socioeconomic status was still tied to one's race. However, generally, poverty status was associated with weaning later than those not in poverty. This could be explained by the fact that solid foods are more expensive than breastfeeding, which is free so long as the mother can produce breast milk. One additional hypothesis is that mothers with certain occupations may be more likely to wean early due to the nature of their occupations. For example, it may be more difficult for women who work in offices to bring their child to work and breastfeed them. However, we do not have any data on the occupation of these women so none of these hypotheses can be proven. Thus, while there are general trends being observed, there isn't enough data to make strong conclusions about the causes of the trends. Cox models can only show associations in these cases, so more powerful statistical methods may be required.

On the topic of the potential outliers identified during the sensitivity analysis, there are a few reasons for why they have been pointed out as such. First of all, inputs such as the duration of 195 weeks for observation 18 could be input errors. If they aren't errors, then these observations could simply be the result of natural variations arising from unobserved or immeasurable differences between all of the mothers in the study. It is completely possible, though arguably a bit unusual, for women to breastfeed for prolonged periods of time. There have been instances of women breastfeeding for more than 5 years, so this might just be an observation of this somewhat rare instance.

Additional research on this topic is not a waste of effort either. Because breastfeeding is largely beneficial for infants and their development, further research into the the main limitations mothers face when deciding how long to breastfeed their children could be beneficial. Stronger causal associations could help lawmakers and such adjust or create legislation that helps give mothers the accommodations they need to be able to raise a healthy generation of children.

7 Bibliography

"Breastfeeding Benefits Both Baby and Mom | DNPAO | CDC." Centers for Disease Control and Prevention, 7 Sept. 2023, https://www.cdc.gov/nccdphp/dnpao/features/breastfeeding-benefits/index.html.

"Breastfeeding FAQs: How Much and How Often (for Parents) - Nemours KidsHealth." Nemours KidsHealth - the Web's Most Visited Site about Children's Health, https://kidshealth.org/en/parents/breastfeed-often.html. Accessed 7 Dec. 2023.

"Frequently Asked Questions (FAQs) | Breastfeeding | CDC." Centers for Disease Control and Prevention, 18 Apr. 2023, https://www.cdc.gov/breastfeeding/faq/index.htm.

"Kaplan Meier Curve • Simply Explained - DATAtab." Online Statistics Calculator: Hypothesis Testing, t-Test, Chi-Square, Regression, Correlation, Analysis of Variance, Cluster Analysis, https://datatab.net/tutorial/kaplan-meier-curve. Accessed 4 Dec. 2023.

"Tobacco and E-Cigarettes | Breastfeeding | CDC." Centers for Disease Control and Prevention, 16 June 2023, https://www.cdc.gov/breastfeeding/breastfeeding-special-circumstances/vaccinations-medications-drugs/tobacco-and-e-cigarettes.html.

Wasie Kasahun, Abebaw, et al. "Predictors of Exclusive Breastfeeding Duration among 6–12 Month Aged Children in Gurage Zone, South Ethiopia: A Survival Analysis | International Breastfeeding Journal | Full Text." BioMed Central, 21 Apr. 2017, https://internationalbreastfeedingjournal.biomedcentral.com/articles/10.1186/s13006-017-0107-z.

8 Appendix

8.1 Plots

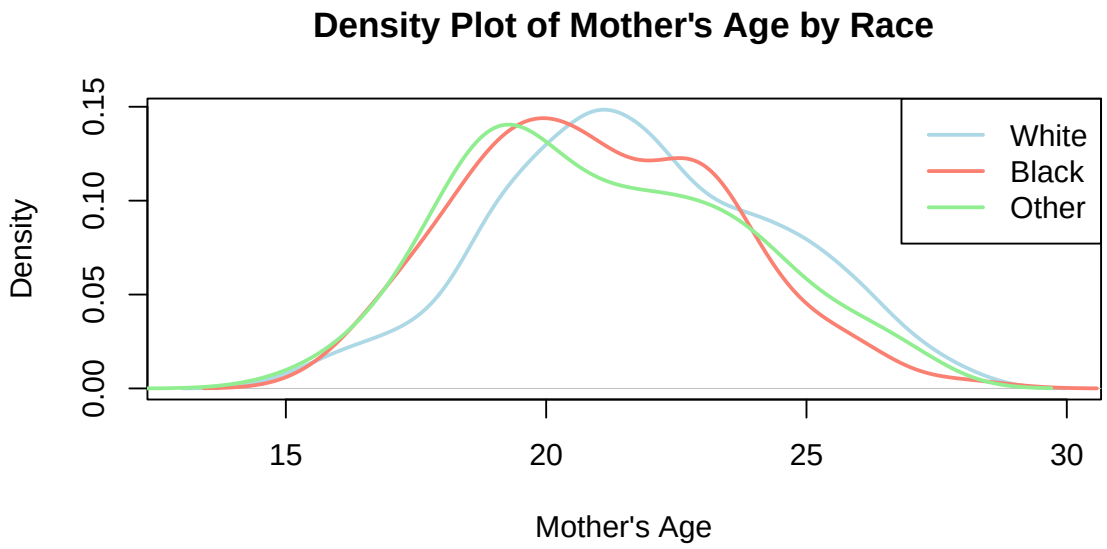


Figure S1: Density Plots Mother's Age by Race

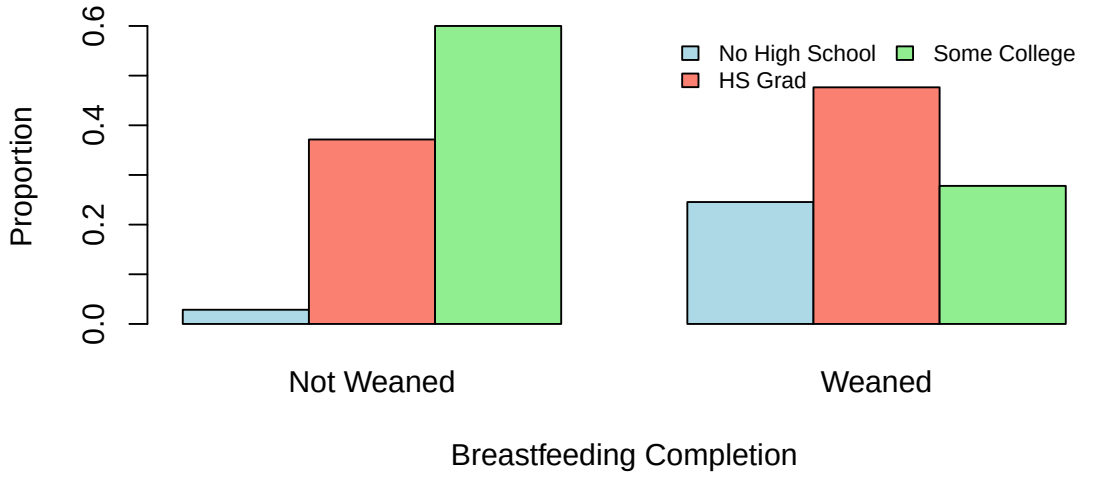


Figure S2: Bar Plot on Education proportion By Breastfeeding Status

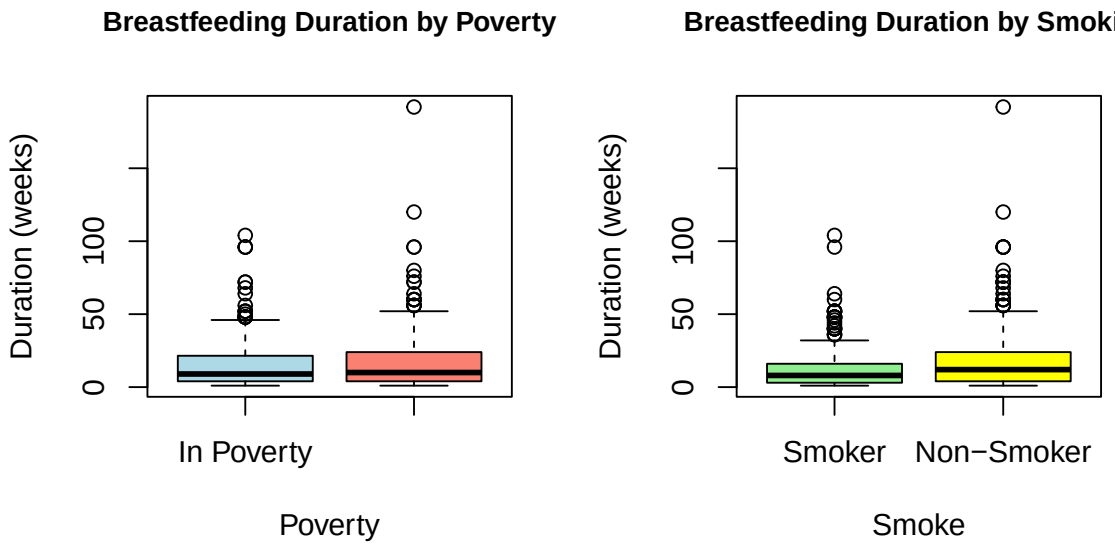


Figure S3: Box Plot for Categorical Variables

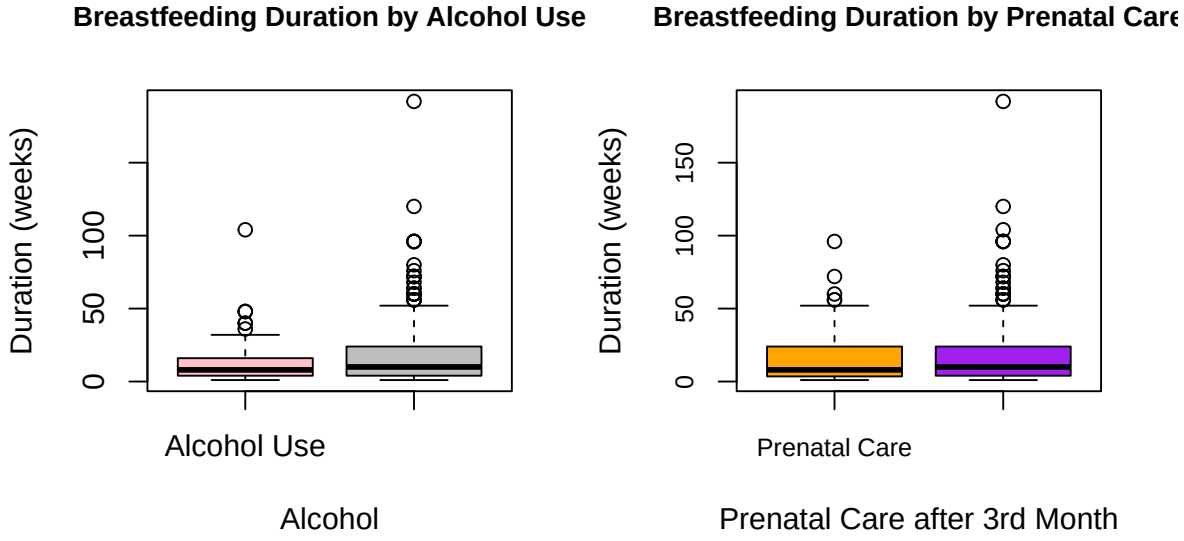


Figure S4: Box Plot for Categorical Variables

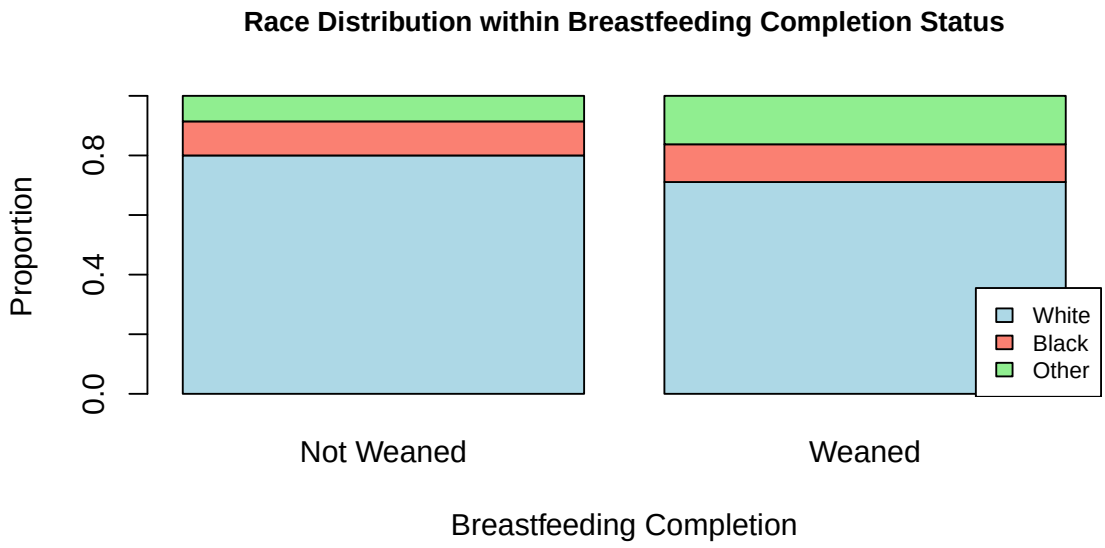


Figure S5: Box Plot of Race Distribution by Breastfeeding Status

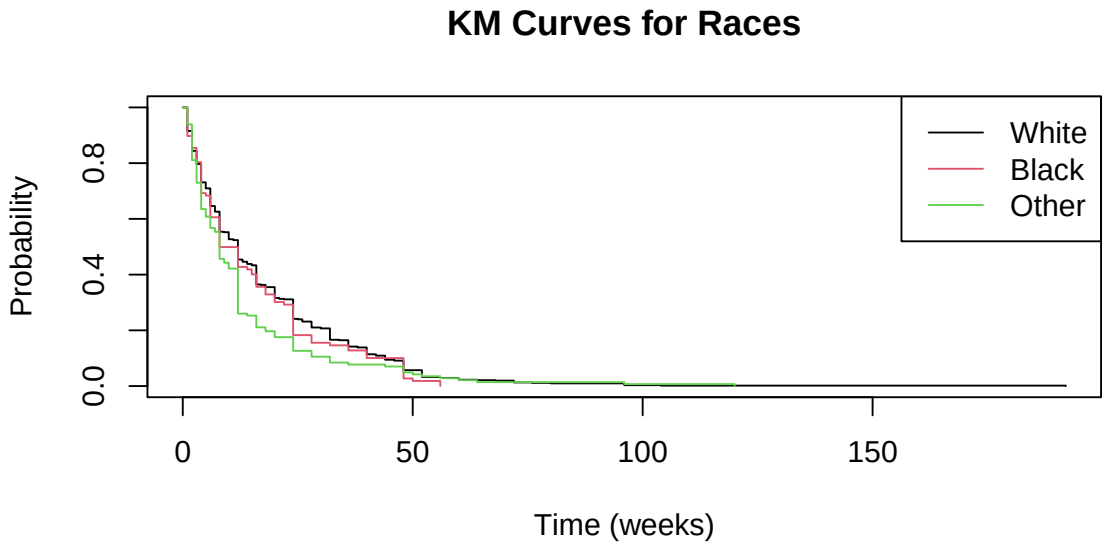


Figure S6: KM Curve - Race

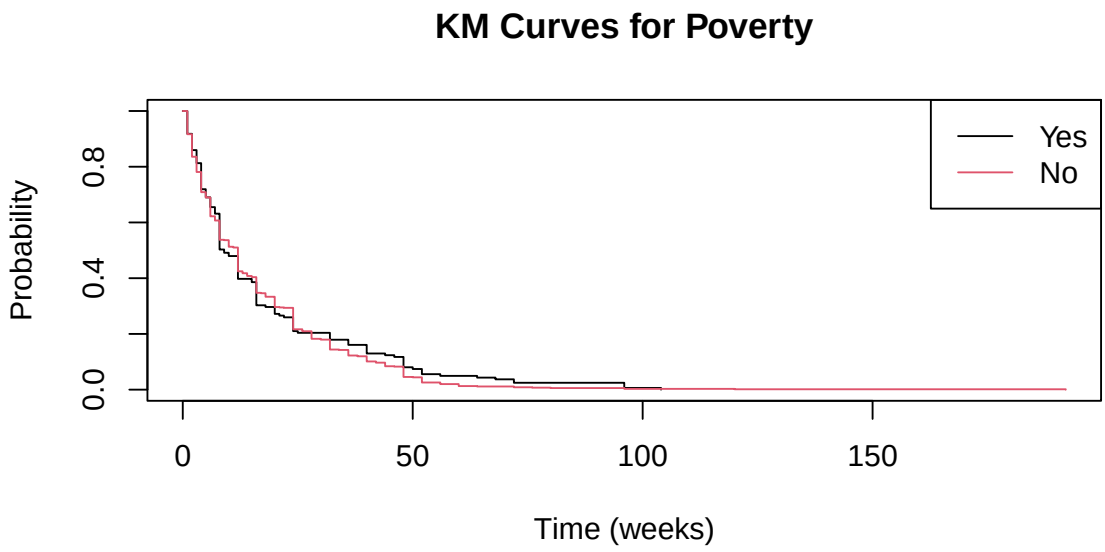


Figure S7: KM Curve - Poverty

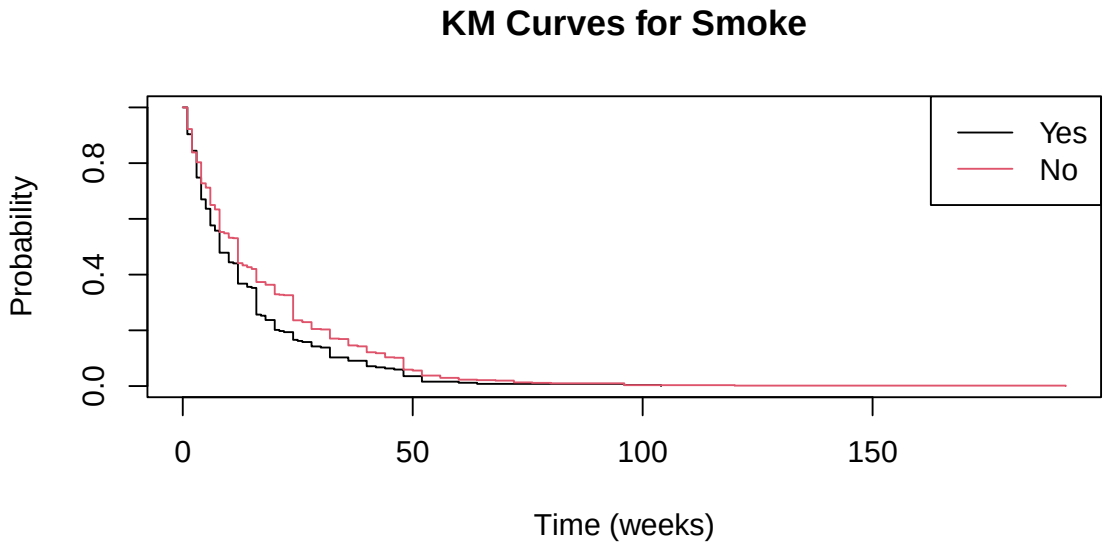


Figure S8: KM Curve - Smoke

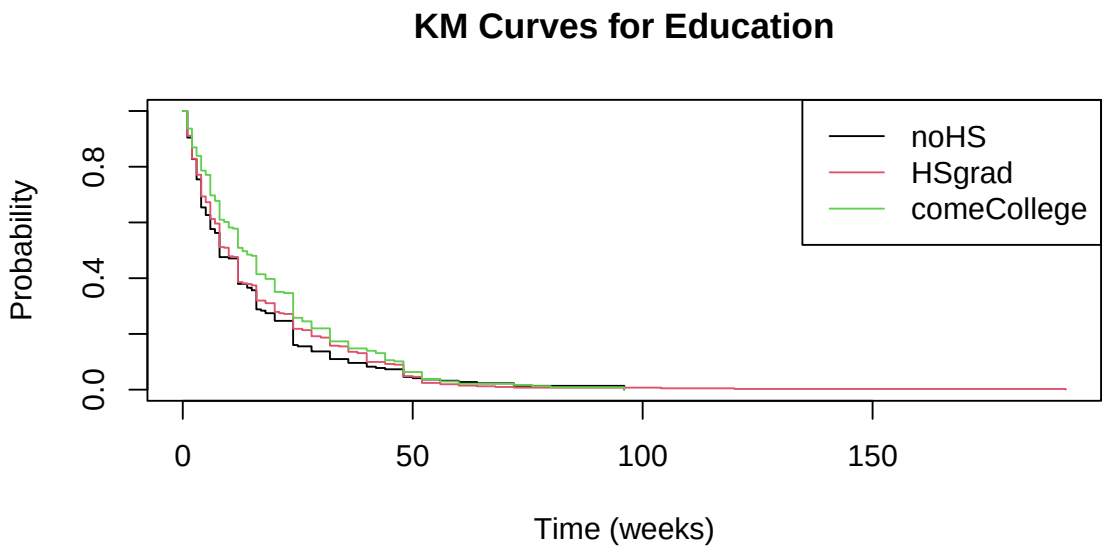


Figure S9: KM Curve - Education

8.1.1 Additonal Plots

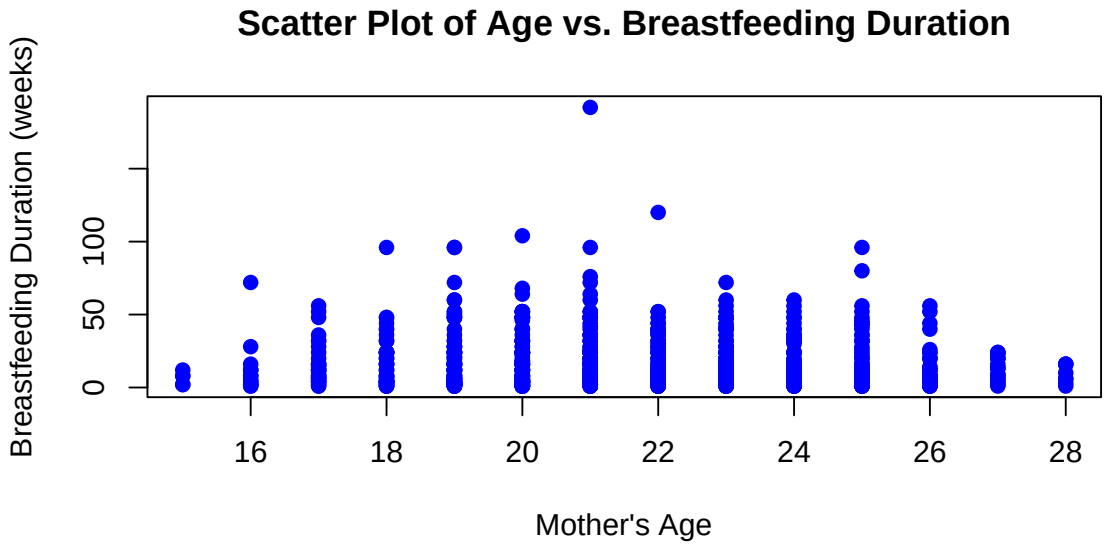


Figure S10: Scatter Plot of Age Vs Duration

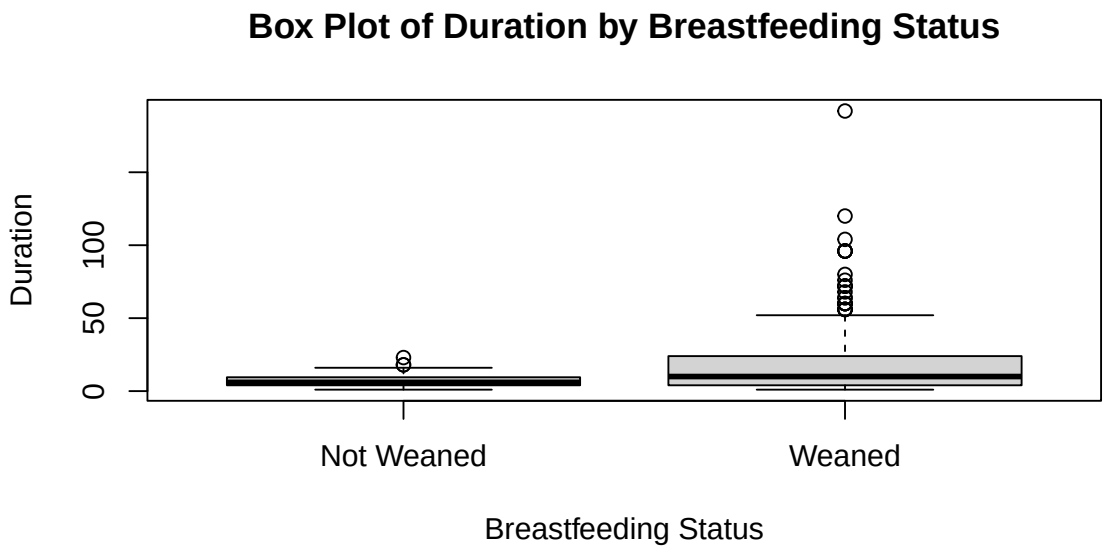


Figure S11: Box Plot of Duration by Breastfeeding Status

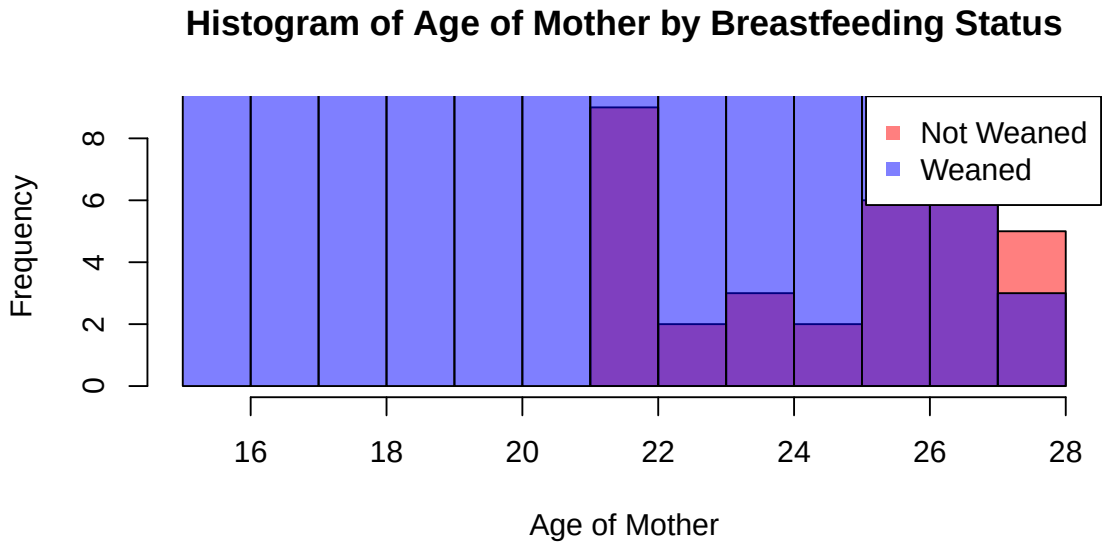


Figure S12: Histogram of Age by Breastfeeding Status

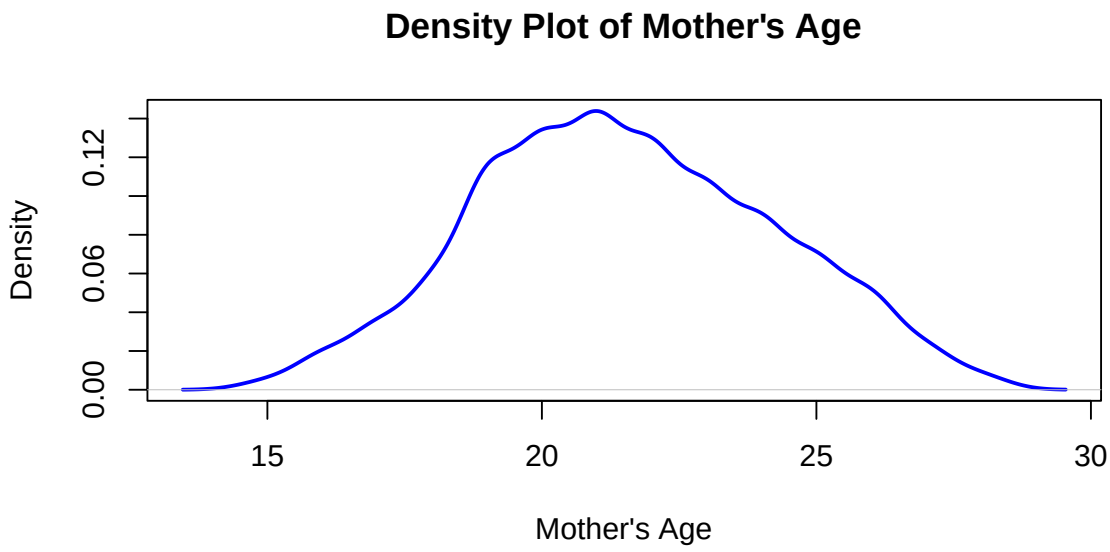


Figure S13: Density Plots for Mothers Age

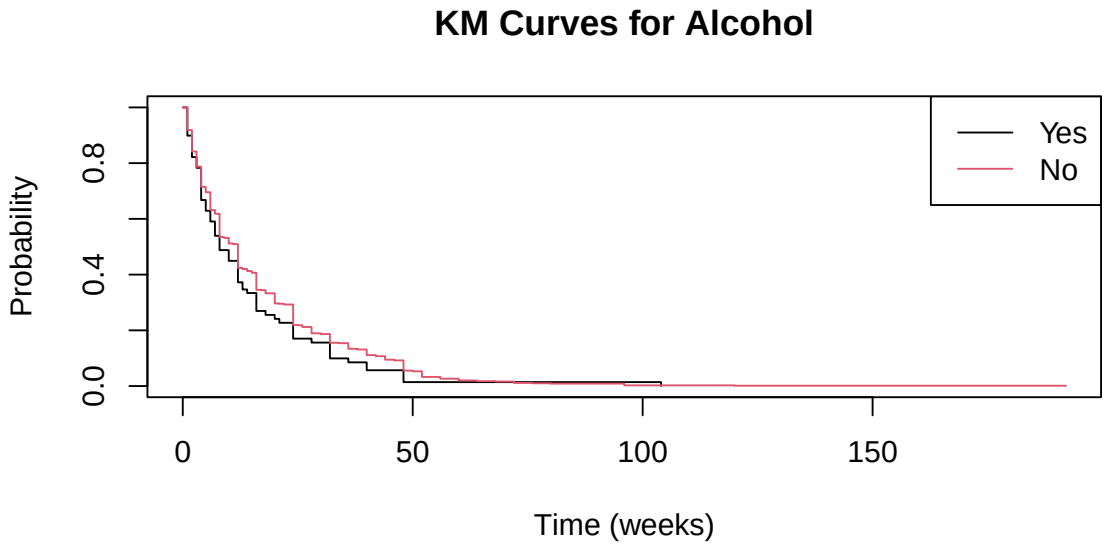


Figure S14: KM Curve - Alcohol