Case Study 1: Maternal Smoking and Infant Mortality

Question

What is the difference in weight between babies born to mothers who smoked during pregnancy and those who did not? Is this difference important to the health of the baby?

Hypothesis

We hypothesize that, based on decades of medical research, smoking during pregnancy is detrimental to the health of the child and will tend to associate with lower birth weights.

Literature Review

We all know that smoking can be detrimental to one's health no matter if you're an avid chain-smoker or a person who experiences the consequences of passive smoking. Tobacco smoking is one of the most prevalent addictive habits, causing diseases of the respiratory tract such as lung cancer and cancers of the larynx and tongue. In addition to causing harm to the intended "active" smoker, new studies have shown that the effects of smoking can be much greater for expectant mothers. While the mechanisms by which by which smoking is related to preterm birth is unclear, there are more than 3,000 chemicals in cigarette smoke and the biological effects of most are unknown. However, both nicotine and carbon monoxide are powerful vasoconstrictors, and are associated with placental damage and decreased uteroplacental blood flow. Both pathways lead to fetal growth restriction and indicated preterm births (Goldenberg). Therefore, smoking done by mothers during pregnancy can play a role in their babies being born prematurely with a low birth weight. In order to validate this relationship, we will analyze the weights of babies born prematurely in conjunction with the smoking status of their mothers.

Maternal smoking, in addition to both alcohol consumption and dietary restrictions during the gestational period, and their effects on the health of the fetus have a number of opinions surrounding them regarding the benefits and disadvantages of consuming them in moderation. Therefore, through the use of a number of statistical methods our hope is to present our findings and draw conclusions in such a way so as to educate expectant mothers about how their smoking habits affect the health of their newborns.

Our statistical and numerical methods include finding the mean, minimum, maximum, standard deviation, skewness, kurtosis, and the first, second, and third quartiles of the baby weights for newborns whose mothers' smoked during the gestational period and for those who did not.

Furthermore, we chose to analyze our data graphically using histograms, box-plots, scatter plots, and kernel density estimations in order to spot any differences between the two aforementioned groups of infants. Finally, all of these methods are used in conjunction with one another to see if there is a difference in the weights of newborns who had mothers who smoked and those who did not.

Introduction of Data

The data that we are working with in this case study (babies23.txt) consists of all of the pregnancies that occurred between the years 1960 and 1967 among women in the Kaiser Health Plan in Oakland, California. The women that took part in this study are those that were enrolled in the Kaiser Health Plan, obtained prenatal health care in the San Francisco area, and delivered their child in any of the Kaiser hospitals in Northern California. Our set of data is made up of 1,236 male babies who all lived to be at least 28 days old. At their birth the length, weight, and head circumference of each baby was recorded. The weights are both numerical and discrete measured in ounces. In addition, the smoking status of the mother was recorded as a categorical variable as either a 0, 1, 2, 3 or 9. A 0 indicated that the mother had never smoked, a 1 indicated that the mother was an active smoker, a 2 indicated that the mother had smoked up until the point they became pregnant, a 3 indicated that they had smoked once before, and a 9 indicated that the smoking status of the mother was unknown.

Background

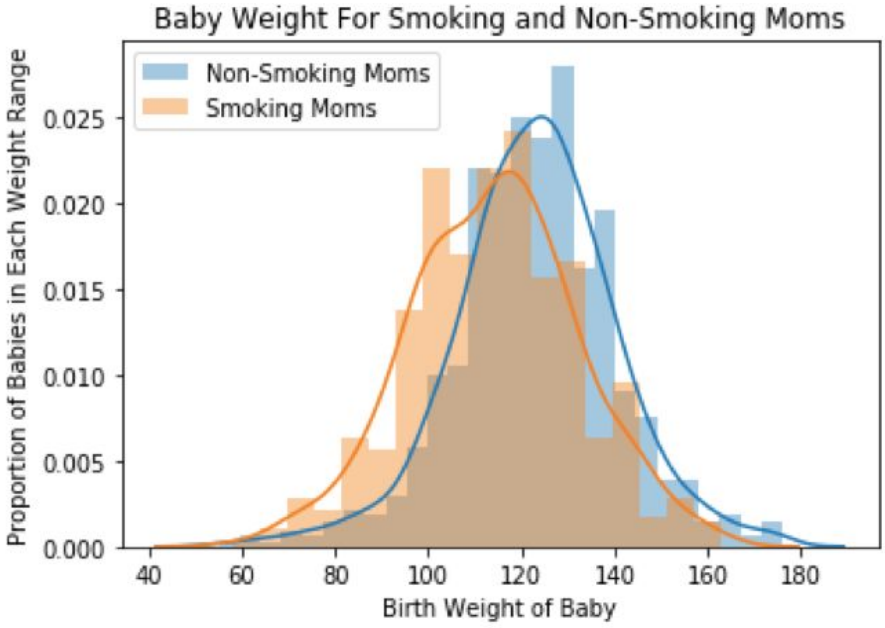
Fetal development can be influenced by any of the following: environmental changes, nutrition intake, genetic information, and uteroplacental cause. Following an exponential pattern, the growth of a fetus is measured by an infant's weight and gestational age. Born too soon, preterm infants are more vulnerable to organ injury, death, chronic illness, and neurodevelopmental disability than full term newborns (Butler). Preterm birth is a common, complex condition that results from multiple interactions between the maternal and fetal genomes and conditions in the intrauterine environment, the mother's body, and her external environment. Preterm birth is considered to be birth at less than 37 completed weeks of gestation.

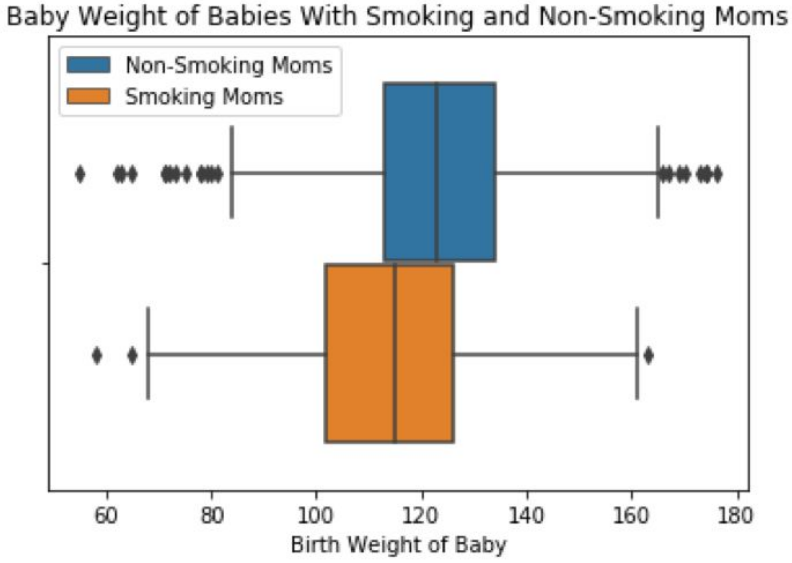
Studies show that babies with low weight at birth are at a higher risk of dying compared to those who are born with a normal weight. The main issue surrounding preterm birth is the lack of biological maturity that comes with extrauterine life. Infants born preterm have immature organ systems that need additional support to survive. Their immature organs are unable to perform complex physiological functions such as gas exchange and blood pressure control in conjunction with sustained breathing, crying, and digesting milk. While immaturity is one of the primary characteristics of preterm infants, the degree to which an infant is considered immature

varies. There is a biologic continuum, and similar gestational age and fetal size may not indicate similar levels of maturity (Butler). Different genotypes as well as different extrauterine environments lead to variation in both size and maturity of infants. Thus, we will further investigate the relationship between maternal smoking and the weights of newborns to see if these intrauterine experiences change the weights of those affected newborns.

Basic Analysis

To get an initial sense of how smoking during pregnancy affects their babies birth weight, the dataset found in 'babies.txt' was divided into a dataset containing data on babies whose mothers smoked during pregnancy and another containing data on babies whose mothers did not smoke during pregnancy. Histograms, with overlaid KDE plots, and box plots describing the frequency of birth weights among these two groups are shown below. Note that birth weight is in ounces.





Both distributions appear to be normally distributed. The mean birth weight of babies born to non-smoking moms appears to be higher than the mean birth weight of babies born to smoking moms and the standard deviations for these two groups seem to be similar.

To further investigate these two distributions, summary statistics were computed and appear below. Note mean, min, max and the quartiles are measured in ounces.

	Non-Smoking Moms	Smoking Moms
Mean	123.047	114.11
Min	55	58
1st Quartile	113	102
2nd Quartile	123	115
3rd Quartile	134	126
Max	176	163
Standard Deviation	17.3987	18.0989
Skew	-0.187363	-0.0336995
Kurtosis	1.05221	0.000407814
95% Confidence Interval	[88.25, 157.845]	[77.91, 150.31]

As seen in the above visualizations, the mean and quartiles of the non-smoking group were all slightly higher than those of the smoking group. In addition, both distributions have similar standard deviations and low skew. Low kurtosis values, in addition to a low skew, indicate that these distributions are approximately normal. As a result, the 95% confidence intervals shown in the table, calculated by taking two standard deviations from the mean, are reliable. These confidence intervals quantify the great overlap shown in the histogram above.

While having a lower mean birth weight, many of the babies born to smoking mothers are of normal weight. Many short and long-term health complications exist when a baby is classified as a “low birth weight baby.” This classification is given to babies who are born with a birth weight of less than 88 ounces. If there is a significant increase in the frequency of low weight births in the smoking group when compared to the non-smoking group, this points to a positive correlation between the frequency of low birth weight births and smoking.

To investigate these low weight birth differences, the frequency of low weight births for each group was calculated using the sample provided in babies.txt. The results are shown below

	Non-Smoking Moms	Smoking Moms
Actual Frequency of LW Births	0.02965	0.07438

In this sample, babies born to mothers who smoked during their pregnancy had roughly a two and a half times greater chance of being low weight. To see if this empirical result was an anomaly or the expected result, a Monte Carlo simulation was run. As previously observed, the distributions of each group are approximately normal. Knowing this, in conjunction with the previously calculated means and standard deviations for each group, 1000 sample birth weights were drawn from a normal distribution. The frequency of low weight births were then calculated from this sample. This procedure was run 100 times. The results of the first 10 trials are shown below:

	Non-Smoking Moms	Smoking Moms
S0 Underweight Freq	0.021	0.065
S1 Underweight Freq	0.032	0.082
S2 Underweight Freq	0.025	0.081
S3 Underweight Freq	0.021	0.087
S4 Underweight Freq	0.023	0.067
S5 Underweight Freq	0.018	0.082
S6 Underweight Freq	0.026	0.069
S7 Underweight Freq	0.019	0.075
S8 Underweight Freq	0.027	0.075
S9 Underweight Freq	0.016	0.068

In each of the 100 trials, the non-smoking group had less underweight births than the smoking group. The average frequency for the non-smoking group was 0.022 while the average frequency for the smoking group was 0.076.

I believe the estimated frequencies found in both the actual sample and Monte Carlo samples to be reliable. There was a clear difference in the frequency of low-weight births between the two groups. If the data were perturbed and several low-weight babies were moved from the smoking to non-smoking group, the frequencies would remain relatively unchanged due to the large sample size.

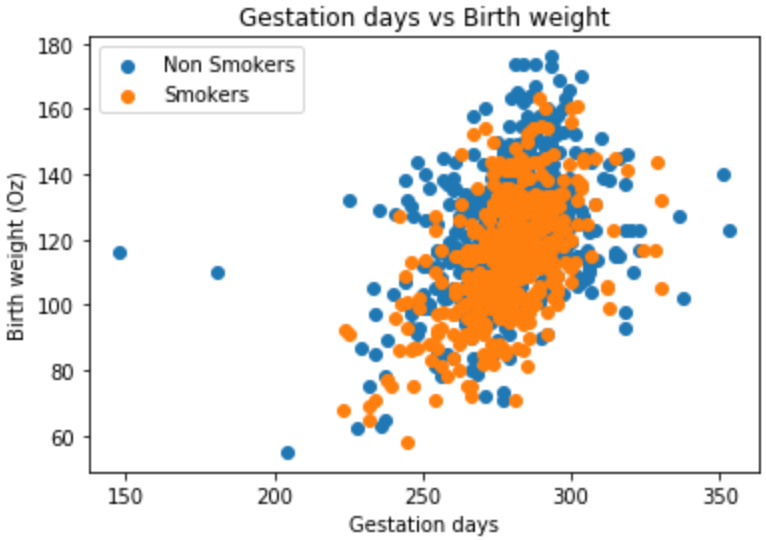
After analyzing the babies.txt dataset and the results of the Monte Carlo simulation, a notable correlation between whether or not a mom smoked and the birth weight of her baby was seen. In the dataset, babies born to mothers that smoked during pregnancy weighed approximately 9 ounces less. In addition, a much greater incidence in low birth weights was seen. The Monte Carlo simulation, in which the low weight birth frequencies for the smoking group were greater than the non-smoking group in all 100 trials, validated these results.

However, while both the actual sample means and Monte Carlo simulation sample means appear to indicate a strong correlation between the frequency of low weight births and whether or not the mother smoked, this sample cannot be generalized to the population as the data is from a very narrow subset of pregnant mothers (women who were part of the Oakland, California Kaiser Health Plan). To exhibit a reliable correlation between whether a mother smoked during pregnancy and the birth weight of their babies, a random sample of all pregnant mothers would have to be analyzed.

Advanced Analysis

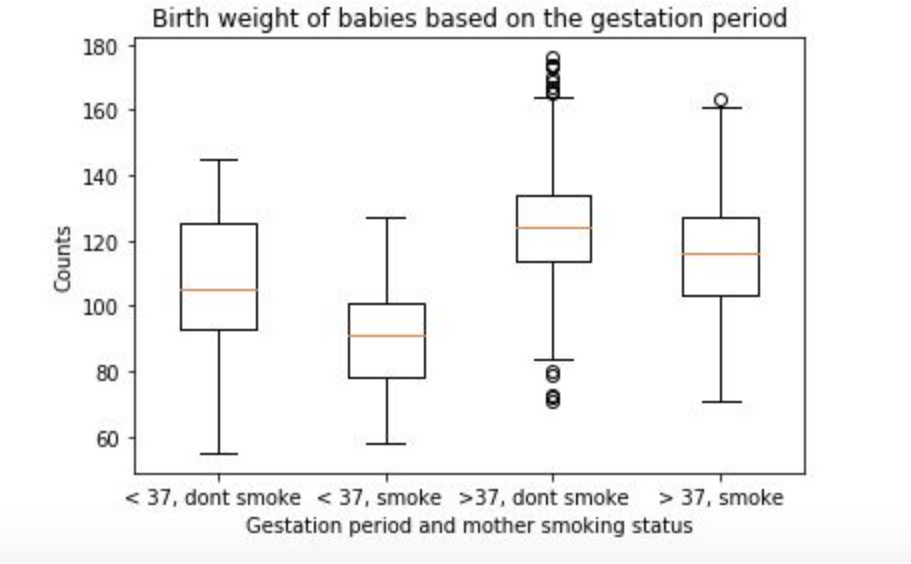
There are potentially many other factors that could affect the birth weight of the baby, for example premature babies are known to have much lower birth weights, age and diet of the mother and the number of births. However, in this dataset we will only be analysing mothers that had one child. So we will explore the other two factors.

Babies that are born before the 37th week are considered to be premature babies and tend to have lower birth weights and we will explore the relationship between gestation days and the birth weight for both smoking and non smoking mothers.



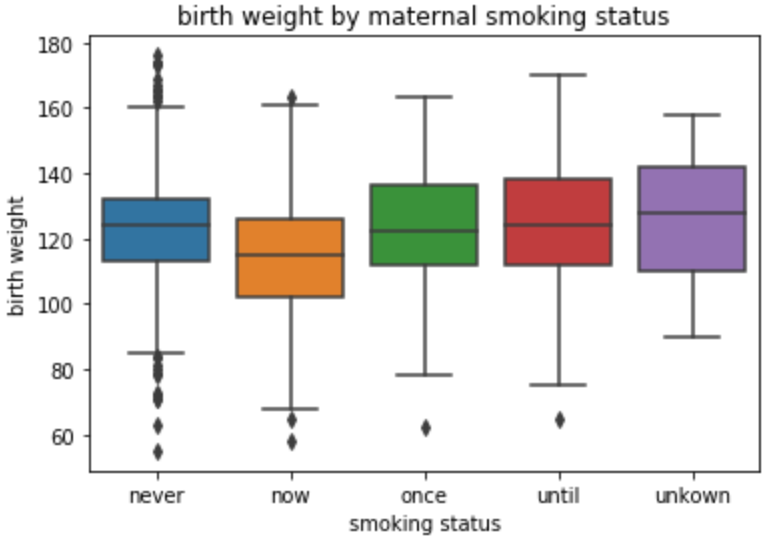
Pearson Correlation of gestation days and birth weight for non smoking moms = 0.34619714693191167
Pearson Correlation of gestation days and birth weight for smoking moms 0.49380755142347965

There does not seem to be any significant difference in the pearson coefficient for babies of smoking mothers and non smoking mothers. We can however see a reduction in birth weight for both type of babies that were born around and less than 260 gestation days (37 weeks). The average weight of a premature baby is known to be around 81 ounces.

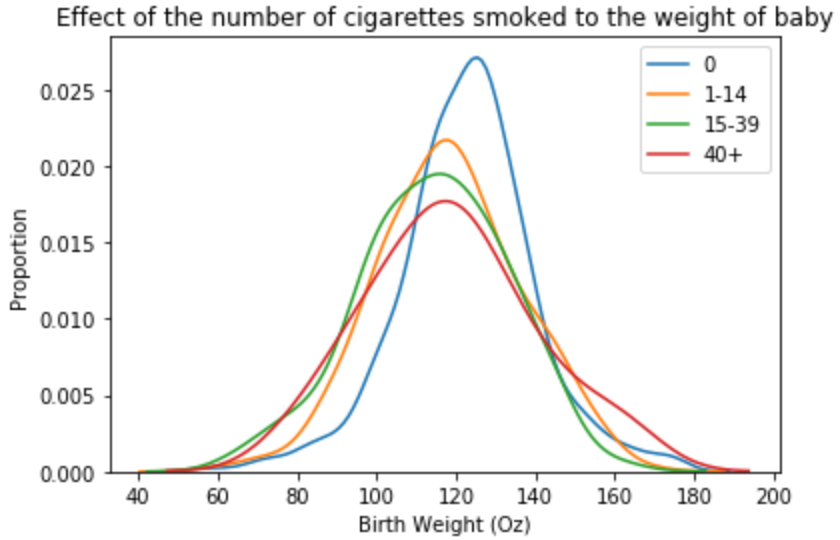


After plotting the boxplots for premature and non-premature babies for smoking and non-smoking mothers, we can see that our entire sample of premature babies have an average birth weight higher than the known average of 81.This is most likely due to homogeneity of the mothers picked for this study, as they were all from the San Francisco area and a part of the Kaiser Health Plan.

Our data from babies.txt does not give us enough information from the ‘smoke’ variable. Hence it would be more beneficial to use the babies23.txt, where the ‘smoke’ variable tells us if the mother smoked during pregnancy/ until pregnancy and so on.



After using the more extensive data from babies23 to plot the birth weight of babies based on their mother's smoking status, we can clearly see that mother's who smoked during their pregnancy had a marginal decrease in the birth weight of their babies. If the mother smoked once during her life, or smoked until the pregnancy, the babies weight seems to be unaffected. Further, on plotting the number of cigarettes the mother smoked in a day (graph below), we can also see that the number of cigarettes the mother smoked actually does not play any role in the birth weight. Some mothers, who smoked 40+ cigarettes a day had the same birth weight as those that smoked 1 a day. To show the birth weights in comparison to non smoking mothers, we did plot the weights of mothers who didn't smoke at all. The fact that the number of cigarettes smoked does not affect the birth weight is quite surprising, since it is known that smoking cigarettes causes problems with the placenta, which is the source of the babies nutrition and oxygen.



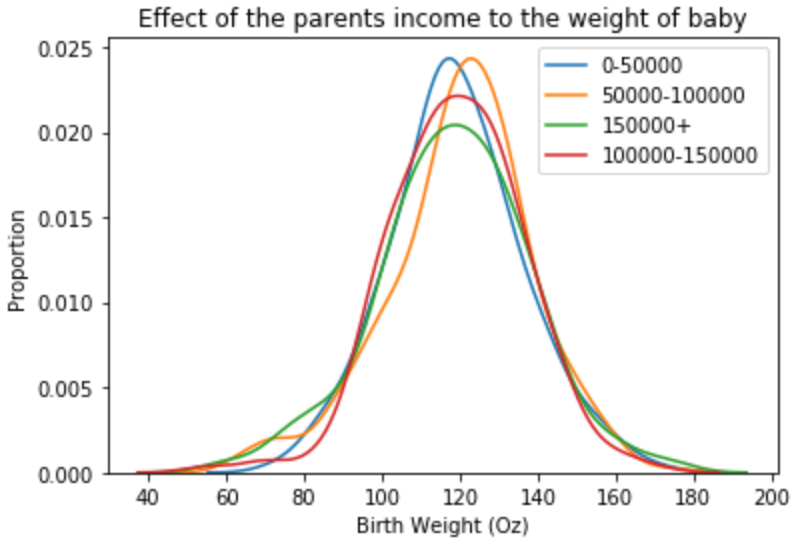
On checking the proportion of underweight and overweight babies for smoking and non-smoking mothers, we can truly see the hidden effect of smoking has on the birth weight of babies. 2/3rd of the healthy babies (in terms of birth weight) belong to non-smoking mothers, while 2/3rd of the unhealthy babies (in terms of birth weight) belong to smoking mothers.

	proportion over weight
smoke	
0	0.612403
1	0.378984
9	0.008613

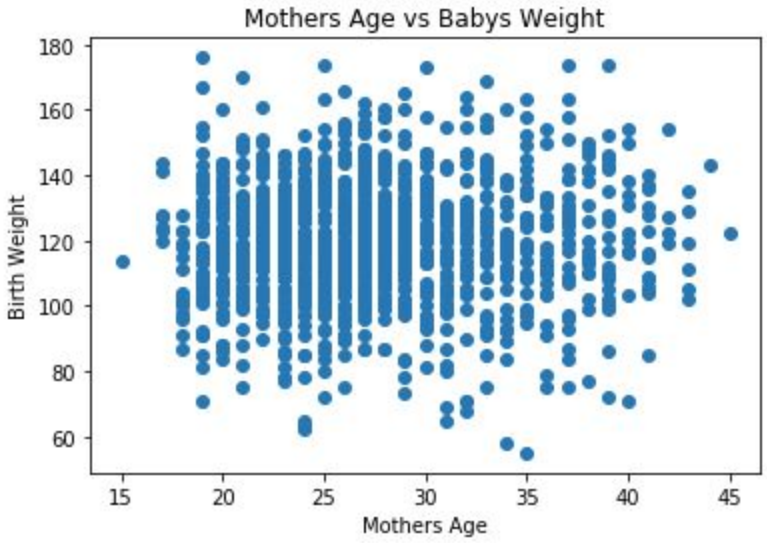
	proportion under weight
smoke	
0	0.354839
1	0.645161

We also explored the effect of family income on the birth weight of the babies. There are many reasons this factor could affect the babies weight. For starters, maybe the family didn't have

enough money to get the required medical care that is required for all mothers. Perhaps the mother had a virus that could have been cured, but didn't have the money for the cure. Further, the family could have potentially not had enough money to feed the mother 3 basic meals a day, which could take a toll on the birth weight of the baby. But after plotting out the data, we can see that the income of the family has absolutely no role in the final weight of the baby, as can be seen in the graph below.

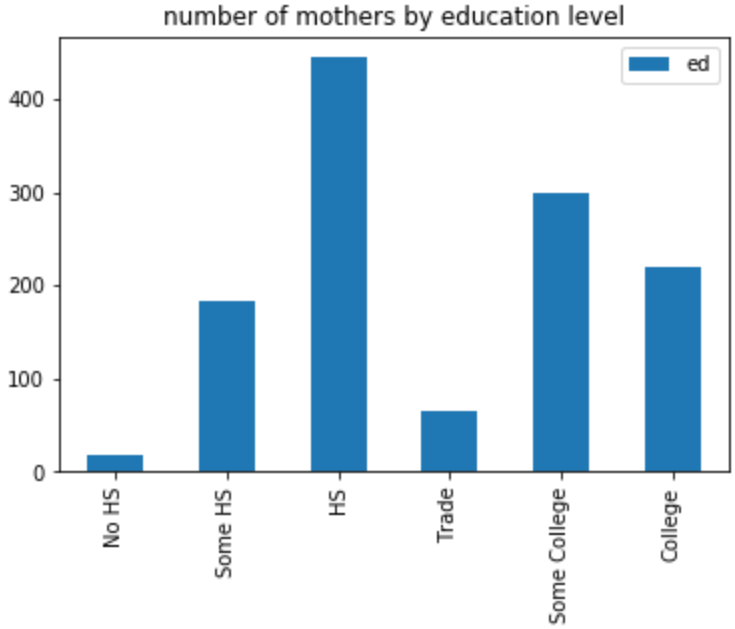


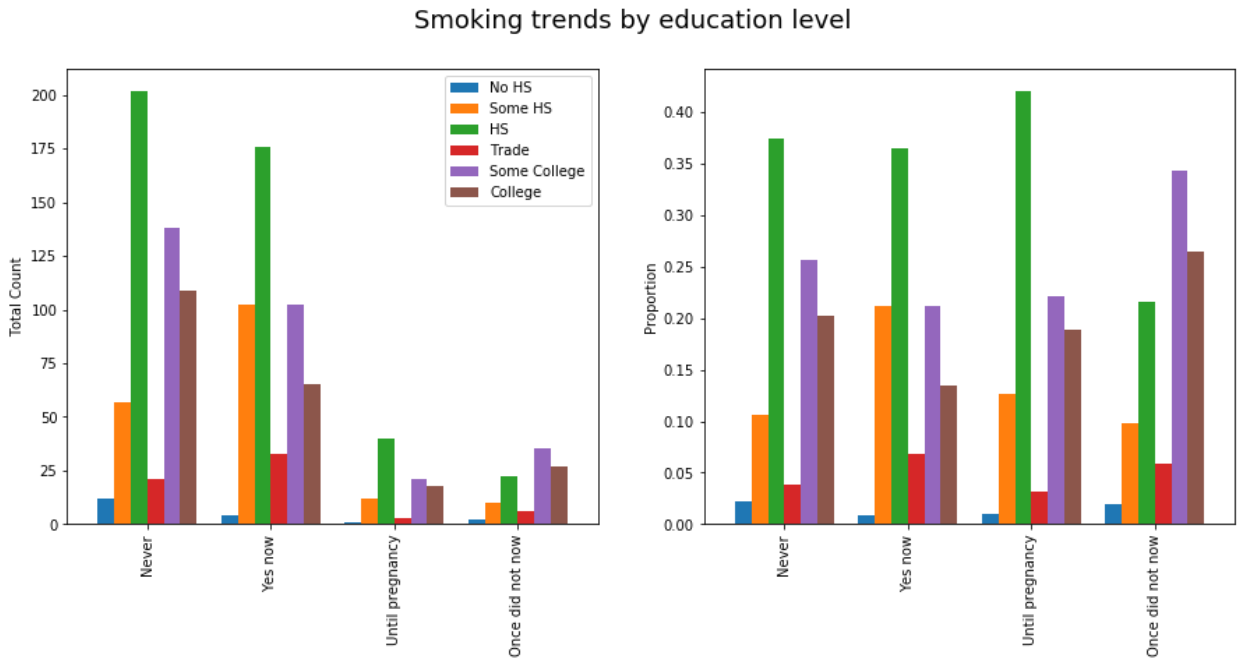
According to a lot of articles, teen mothers are more likely to have underweight babies. This is because teen mothers tend to gain less weight during their pregnancies. Teenage mothers are also more likely to have poor eating habits, in comparison to older women and are less likely to take recommended daily prenatal multivitamins to maintain adequate nutrition during pregnancy. Teens are also more likely to consume alcohol and other drugs, which can cause problems for the babies growth. To add to this, teenage mothers tend to receive prenatal care less than older women. During prenatal care, professionals inform mothers on what is best for them to ensure the baby is healthy.



On plotting the data, we see that there seems to be no correlation that younger mother's babies tend to have lower birth weights. The value of the pearson coefficient between the two variables is 0.029221376297396125 confirming that there is no correlation between the variables.

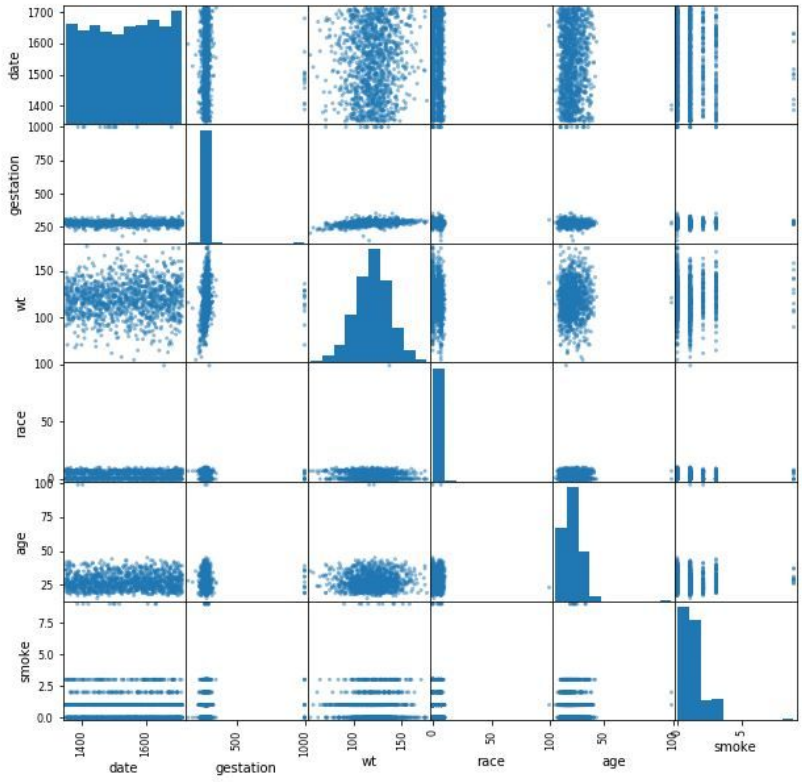
On plotting the counts of the education levels of the mothers, we see that a lot of the mothers in our dataset had just a high school education. Perhaps this may result in them being less aware of the ramifications of smoking and may skew the overall results of our data analysis.





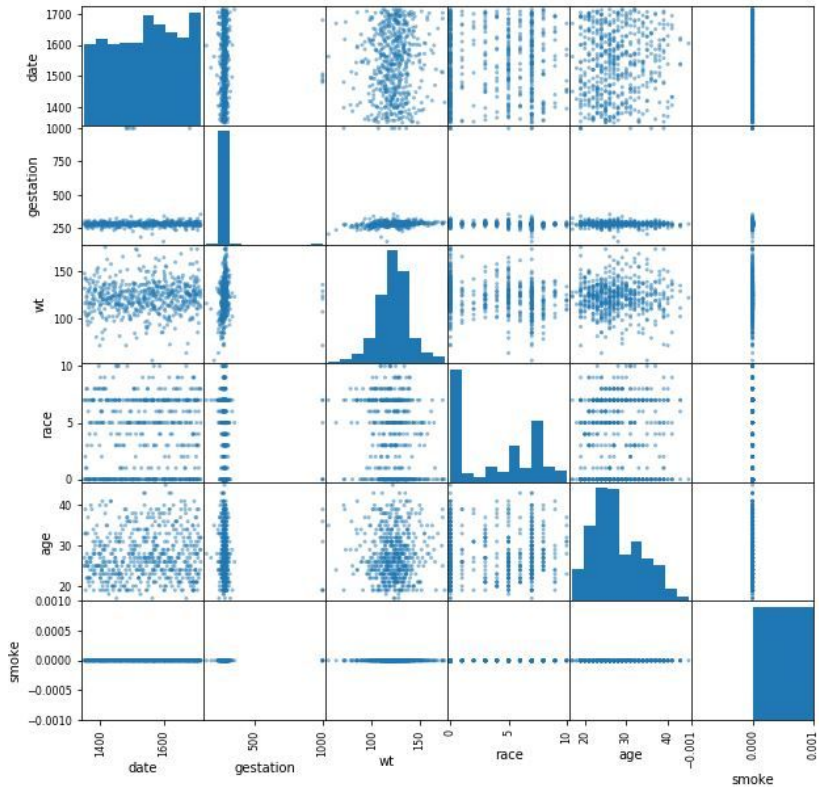
The general trend in this is that mothers who stopped their education at high school tend to be smoking more before and during their pregnancies than in comparison to the other categories. This must be due to the lack of education about how harmful cigarettes are for not only themselves, but their babies. A lot of college educated mothers smoked a while ago, but don't anymore. This may have been due to the stress of their college courses. But clearly, they know the harmful effects of cigarettes and stopped smoking. One way the government could counter this is by having informative sessions about the harmful effects of smoking cigarettes. Further, they could set up free prenatal care for mothers, where professionals can help mothers with what they should and shouldn't do during their pregnancies.

Included is some more analysis on the babies23 dataset, particularly with scatter matrices. Here is the scatter matrix for the entire babies23 dataset.



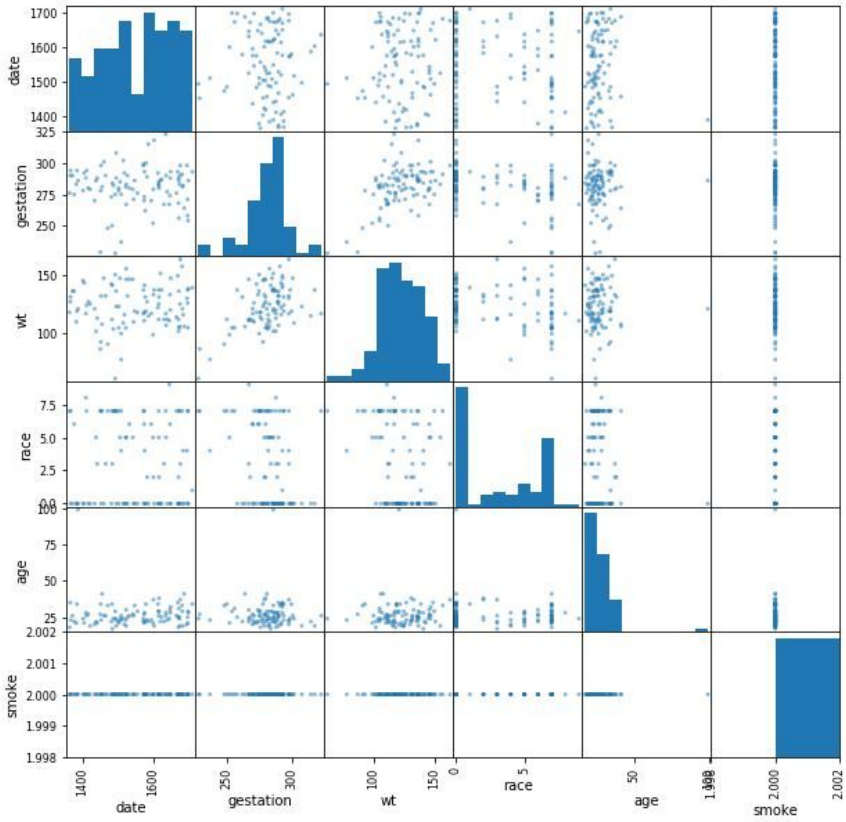
Looking at the scatter matrix of the date, gestation, wt, race, age, and smoke columns of the entire babies23 dataset, we can see that the wt histogram follows a normal distribution. This normal distribution can be verified by the scatterplot of date versus wt. In this scatterplot, we see a higher concentration of data points towards the center of the plot, with some extreme values on both the left and right sides of the plot. So, if each data point dropped to the x-axis to make a histogram of the wt values, that histogram would follow a normal distribution, which is visually verified in the wt histogram shown above. The reason that any plot that plots smoke values has striations in the plot, meaning that there are clearly defined lines in the plot, is because smoke values are discrete values (0=never, 1=yes now, 2=until pregnancy, 3=once did not now, 9=unknown), so there would be large concentrations of each discrete value in the smoke plots.

Here is the scatter matrix for babies with non-smoking mothers.



Looking at the scatter matrix of the date, gestation, wt, race, age, and smoke columns of the babies23_never_smoked dataset, we can see that the wt histogram follows a normal distribution. This normal distribution can be verified by the scatterplot of date versus wt. In this scatterplot, we see a higher concentration of data points towards the center of the plot, with some extreme values on both the left and right sides of the plot. So, if each data point dropped to the x-axis to make a histogram of the wt values, that histogram would follow a normal distribution, which is visually verified in the wt histogram shown above. The reason that any plot that plots smoke values has a striation in the plot, meaning that there is a clearly defined line in the plot, is because smoke values are discrete values (in this case, since we are working with non-smoking mothers, we only work with the smoke value of 0=never), so there would be a large concentration of the smoke value as 0 in the smoke plots. There was also an outlier for a gestation value around 1000 for babies with non-smoking mothers. In terms of race, there is a large concentration of values for white as well as black. In terms of age, there is a large concentration of low values for age (around 20 to 30), so a lot of non-smoking mothers have babies when they are younger (around 20 to 30). In addition, there are no outliers for age.

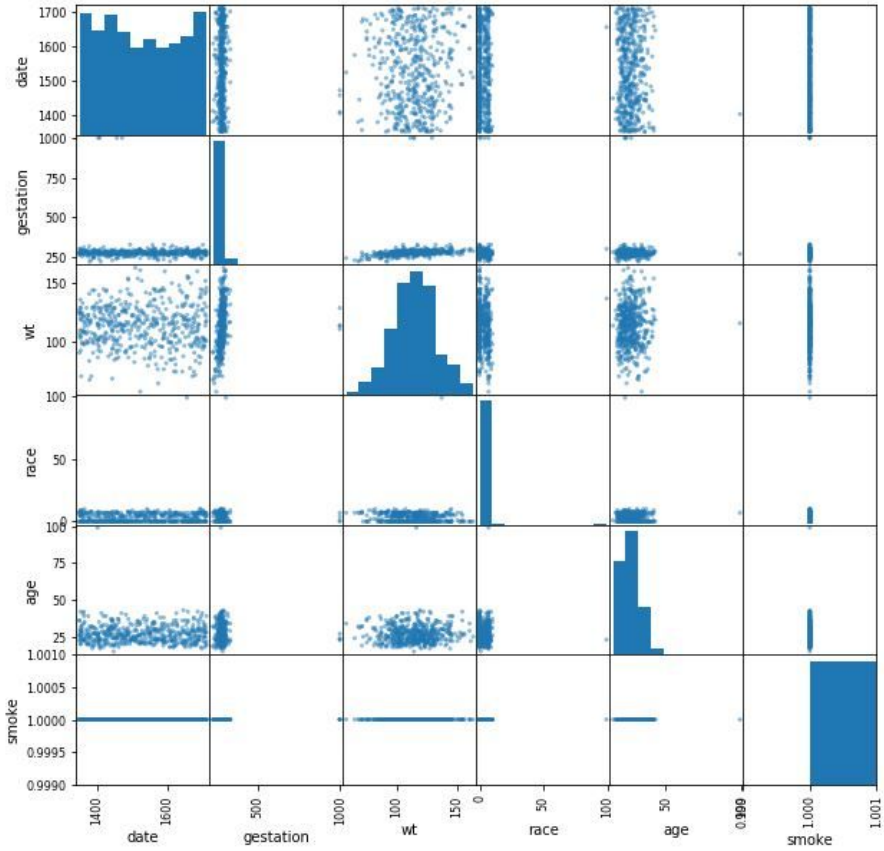
Here is the scatter matrix for babies with mothers who smoked until pregnancy.



Looking at the scatter matrix of the date, gestation, wt, race, age, and smoke columns of the babies23_until_pregnancy dataset, we can see that the wt histogram does not follow a normal distribution. This distribution can be verified by the scatterplot of date versus wt. In this scatterplot, we see a higher concentration of data points towards the right side of the plot, meaning that there were more babies with mothers who smoked until pregnancy that had higher than average birth weights than babies with smaller than average birth weights. So, if each data point dropped to the x-axis to make a histogram of the wt values, that histogram would not follow a normal distribution, which is visually verified in the wt histogram shown above. The reason that any plot that plots smoke values has a striation in the plot, meaning that there is a clearly defined line in the plot, is because smoke values are discrete values (in this case, since we are working with mothers who smoked until pregnancy, we only work with the smoke value of 2=until pregnancy), so there would be a large concentration of the smoke value as 0 in the smoke plots. There was also no outliers for gestation values for babies with mothers who smoked until pregnancy, compared to gestation values for both the entire babies23 dataset and for babies with non-smoking mothers. In those two cases, there was an outlier gestation value of about 1000. In terms of race, there is a large concentration of values for white as well as black.

In terms of age, there is a large concentration of low values for age (less than 50), so a lot of mothers who smoke until pregnancy have babies when they are younger (less than 50). In addition, there is an outlier of an age value around 100.

Here is the scatter matrix for babies with mothers who smoke now.

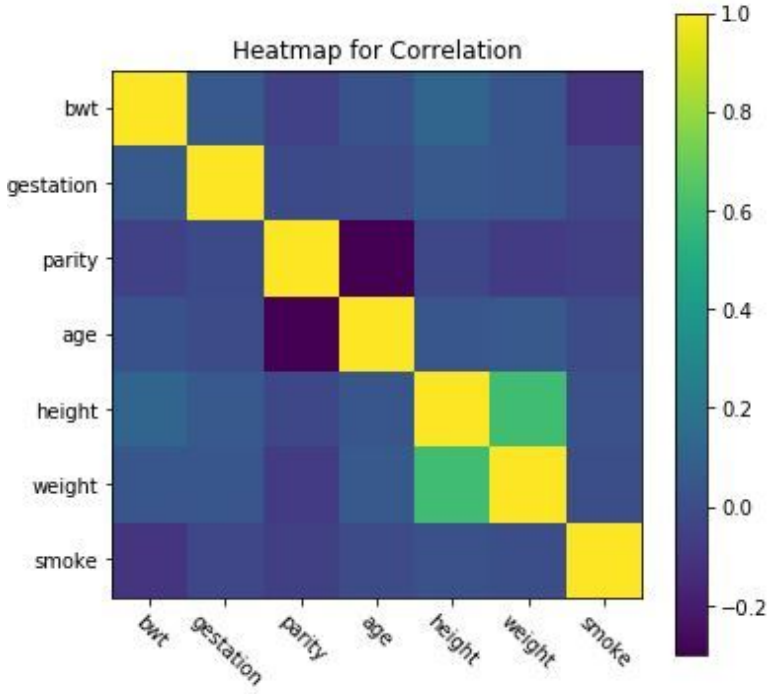


Looking at the scatter matrix of the date, gestation, wt, race, age, and smoke columns of the babies23_smoke_now dataset, we can see that the wt histogram follows a normal distribution. This distribution can be verified by the scatterplot of date versus wt. In this scatterplot, we see a higher concentration of data points towards the center of the plot, with some extreme values on both the left and right sides of the plot. So, if each data point dropped to the x-axis to make a histogram of the wt values, that histogram would follow a normal distribution, which is visually verified in the wt histogram shown above. The reason that any plot that plots smoke values has a striation in the plot, meaning that there is a clearly defined line in the plot, is because smoke values are discrete values (in this case, since we are working with mothers who smoke now, we only work with the smoke value of 1=yes now), so there would be a large concentration of the smoke value as 1 in the smoke plots. There was also no outliers for gestation values for babies with mothers who smoke now, compared to gestation values for both the entire babies23 dataset and for babies with non-smoking mothers. In those two cases, there was an outlier gestation

value of about 1000. In terms of race, there is an outlier value of 99=unknown. In terms of age, there is a large concentration of low values for age (less than 50), so a lot of mothers who smoke until pregnancy have babies when they are younger (less than 50). In addition, there is an outlier of an age value around 100.

Here's the correlation matrix and corresponding heatmap of each variable in the babies dataset.

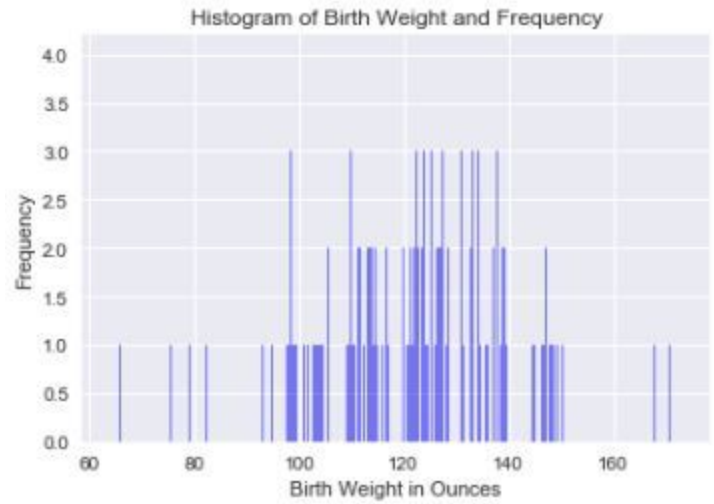
	bwt	gestation	parity	age	height	weight	smoke
bwt	1.000000	0.062504	-0.046107	0.029041	0.125541	0.046764	-0.097609
gestation	0.062504	1.000000	-0.008475	-0.003265	0.065981	0.049459	-0.021770
parity	-0.046107	-0.008475	1.000000	-0.298859	-0.015540	-0.069002	-0.051491
age	0.029041	-0.003265	-0.298859	1.000000	0.048518	0.062732	0.000940
height	0.125541	0.065981	-0.015540	0.048518	1.000000	0.601003	0.023704
weight	0.046764	0.049459	-0.069002	0.062732	0.601003	1.000000	0.012604
smoke	-0.097609	-0.021770	-0.051491	0.000940	0.023704	0.012604	1.000000

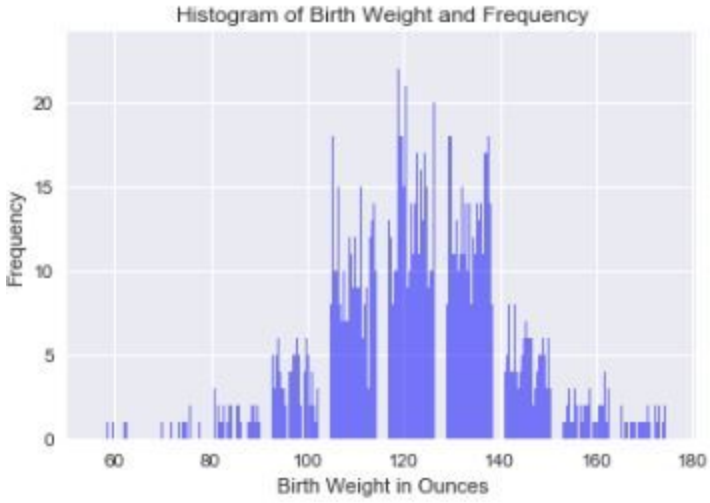


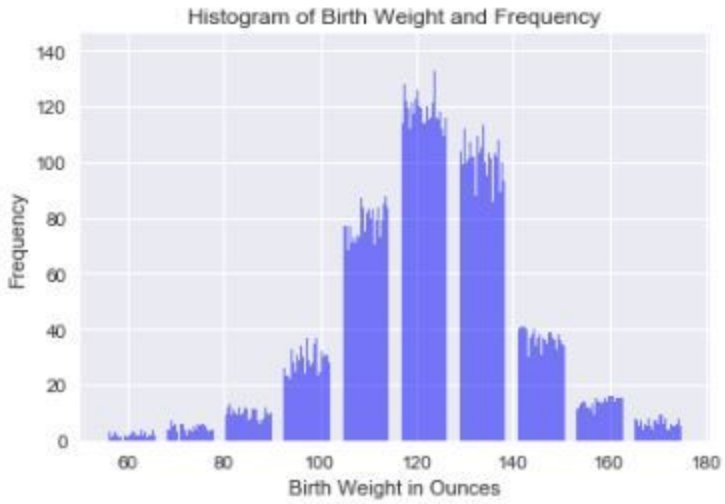
As we can see from both the correlation matrix and heatmap of each variable of the babies dataset, there is a slight positive correlation between height and weight, which makes sense for

babies since as height increases, weight generally increases as well. In addition, there is a negative correlation between age and parity (-0.298859), which is somewhat larger than other negative correlations shown in the correlation matrix.

Here is the Central Limit Theorem and normal approximations used with data for babies with non-smoking mothers. The following histograms show frequency versus birth weight in ounces using sample sizes of 5, 50, and 400, respectively. For our implementation, we ran through 100 iterations of sampling, and for each iteration, we got a sample with size of 5, 50, and 400, respectively.

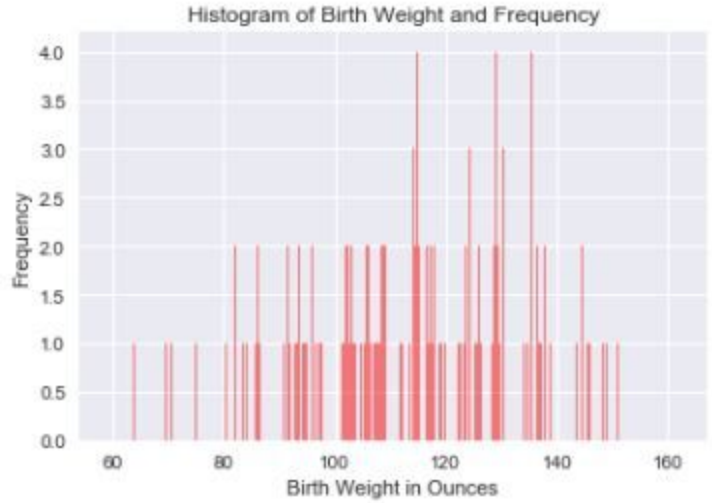


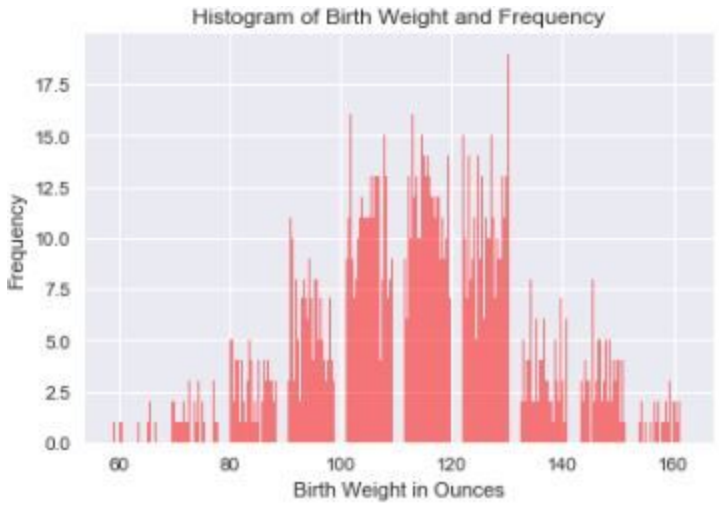


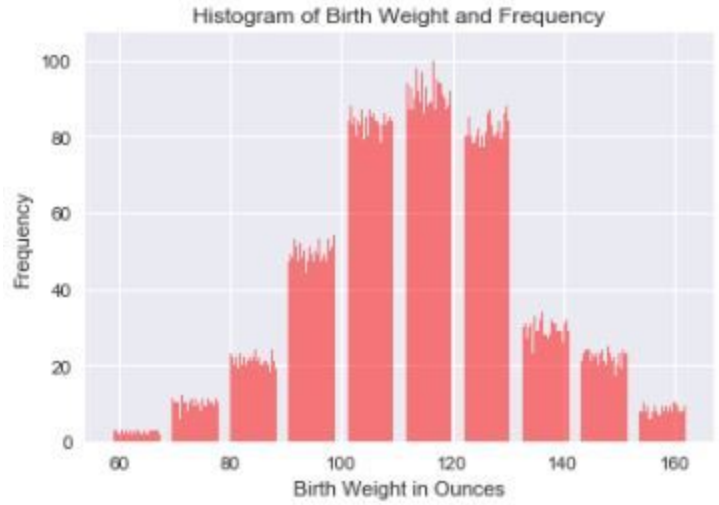


As we can see from the above histograms, the data for babies with non-smoking mothers follows the Central Limit Theorem since as the sample size increases, the corresponding histogram more accurately reflects a normal distribution.

Here is the Central Limit Theorem and normal approximations used with data for babies with smoking mothers. The following histograms show frequency versus birth weight in ounces using sample sizes of 5, 50, and 400, respectively. For our implementation, we ran through 100 iterations of sampling, and for each iteration, we got a sample with size of 5, 50, and 400, respectively.

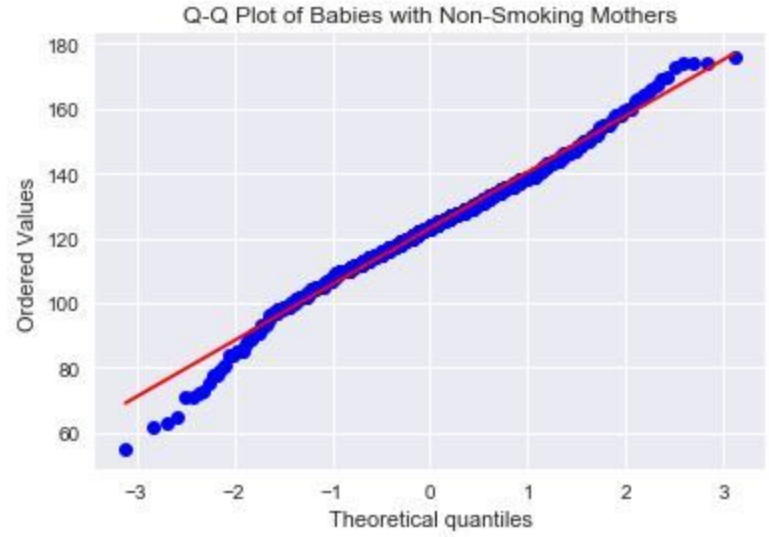


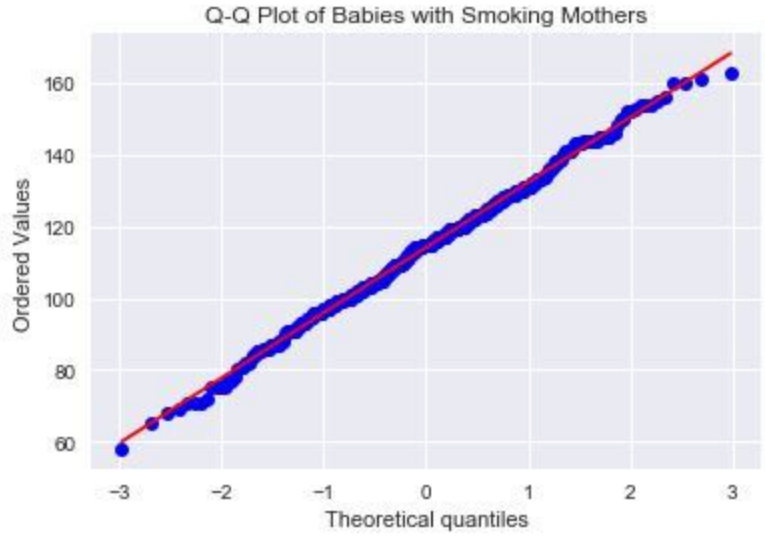




As we can see from the above histograms, the data for babies with smoking mothers follows the Central Limit Theorem since as the sample size increases, the corresponding histogram more accurately reflects a normal distribution.

Here are the Q-Q plots for babies with non-smoking mothers and babies with smoking mothers, respectively:





Based on the Q-Q plots above, we see that the histograms for both babies with non-smoking mothers and babies with smoking mothers generally follow a normal histogram, which is shown by the red lines. This pattern is shown since the blue data points generally follow the red lines.

Conclusion

The analysis performed on the given babies datasets strongly indicates that there is a relationship between smoking during pregnancy and lower birth weight of the baby. This would suggest a rejection of our hypothesis that no correlation exists. Our basis analysis offers numerous, robust indications that there is a distinct difference in birth weights between smoking and non-smoking mothers. Further, in our advanced analysis, we assessed the impact of other variables including mother's age, mother's family income, and mother's education and retained our indication that smoking during pregnancy has a relationship with lower birth weight. Controlling for these variables suggests a causal relationship, though we cannot be sure without performing a randomized experiment.

These results confirm historical analysis that generally shows that smoking during pregnancy leads to lower birth weights. As lower birth weights can be detrimental to the health of the baby, most medical recommendations advocate against smoking during pregnancy. We would second these recommendations. In practice, it is important to continue to update and confirm even well-established hypotheses, especially when it comes to issues of health and safety. We believe our analysis is honest and rigorous and, if reproduced, would yield similar results and conclusions.

Works Cited

https://www.ncbi.nlm.nih.gov/pubmed/20669423
http://www.nber.org/data/vital-statistics-natality-data.html