1 Introduction

Miscarriage is a tragedy. Just as someone is preparing to be a parent, the prospect is taken from them. Miscarriage occurs surprisingly often; studies estimate that 10-20% of pregnancies end in miscarriage. In addition to its high rate of occurence, miscarriage carries many negative effects. Pregnancy loss may cause women to suffer from grief, anxiety, guilt, and self-blame, sometimes leading to post-traumatic stress disorder [3]. Despite miscarriage's high occurence rate, its cause is generally unknown. Rarely, a cause can be ascertained, but this is the exception rather than the rule. It is currently thought that chromosomal abnormalities are the primary cause of early miscarriage [4].

It is my opinion that more research can be done to determine the causes of miscarriage. This begins with understanding the risk factors involved – knowing the factors that are most correlated with miscarriage can give clues on what to study. Additionally, knowing more about the biggest risk factors for miscarriage could help doctors connect high-risk patients with mental health professionals.

In this project, I hope to use data to predict a person's risk of miscarriage. Some of the questions I hope to answer include:

- Who is at most risk for miscarriage?
- Are people of specific races, ethnicities, or socioeconomic statuses at higher risk for miscarriage?
- Do previous miscarriages increase your risk of subsequent miscarriage?
- Does smoking increase risk of miscarriage?
- Is use of contraception related to miscarriage rates?

2 Data Preparation

Because of some of the difficulties of obtaining medical data (such as HcG levels or other medical indicators), I chose to use survey data from the National Survey of Family Growth (NSFG). The NSFG was conducted by the National Center for Health Statistics (NCHS), which is affiliated with the Center for Disease Control and Prevention (CDC). This data set contains a plethora of data on family growth. Specifically, I'm examining the pregnancy data that they've obtained. The data I'm using come from 2015 - 2017 and are based on survey results obtained from 9,553 different people who were pregnant. The data contain information about race, miscarriage, smoking habits, religion, age, and around 240 other factors. After being parsed, I will list the factors that we choose to keep.

2.1 Data Scraping

The data is hosted by the U.S. government on an FTP (File Transfer Protocol) server. The following function scrapes the data files and saves them to the local filesystem.

```
def get_nsfg_data(files):
    """The NSFG data is provided on an FTP server that we pull from."""
    server_url = 'ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NSFG/'

    for file in files:
        with closing(request.urlopen(server_url+file)) as r:
            with open(file, 'wb') as f:
                shutil.copyfileobj(r, f)
```

2.2 Data Cleaning

The NSFG files are in a format that is difficult to parse. Specifically, the data is organized without any sort of column separation. They are stored as text files where each row corresponds to a single person's survey responses. Within that row, each survey response is recorded in a specific order, where an empty response is represented with whitespace. Stata dictionaries are provided that allow researchers to easily parse and load the data into Stata (software for statistical analysis). These dictionaries have entries that tell us how to interpret the data. For instance, one entry might say that the age of the respondent can be found on the 3rd and 4th entries of that row and should be interpreted as an integer. Below, we provide the function to convert the data files to .csv files for easy loading into the pandas software package. We utilize the Stata dictionaries to do so.

```
def create_csv(columns, column_range, data_file):
    """Create csv from NSFG data file.
    Parameters:
        columns (list): Contains the names of each data column.
        column_range (list): Contains the indices that separate the data in each
                             row
        data_file :
    """

    # Get the top line of the .csv file (i.e. column names)
    top_line = ','.join(columns)
    data_csv = [top_line+'\n']

    with open(data_file,'r') as data:
        data_raw = data.readlines()

    # Append each survey response to the list
    for line in data_raw:
        j = 0
        for idx in column_range[1:-1]:
            line = line[:idx+j] + ',' + line[idx+j:]
            j += 1 # Increment j since we just added another character
        data_csv.append(line)

    # Write to the file
    with open(data_file.replace('.dat','.csv'),'w') as data:
        data.writelines(data_csv)
```

To see all the code used in cleaning the data, please examine the appendix. Here, we will just summarize the process. Now that the data is in a readable format, we drop unneeded columns. The

data starts with 248 columns. I removed columns using the following process:

1. If most of the data in a column is NaN, we drop that column. The threshold we use is 60%. This brings the number of columns to only 122.
2. We drop several columns that have to do with measuring the year the data was recorded. This leaves us with 115 columns.
3. There are 42 columns that are imputation flags for another column, meaning that the NCHS recorded whether that variable was filled out in the questionnaire, or imputed using either 'Multiple Regression Imputation' or 'Logical Imputation'. We will drop these columns, under the assumption that the government properly imputed these data. Now, we have 73 columns.
4. After individually examining these columns, I chose to drop 46 more columns, bringing our total number of columns to 27. I felt justified dropping these columns because they could either be imputed using another column, or they were related to information about babies after they were born (e.g. the number of weeks that a mother breastfed her child).

The problem now is that much of our data consists of answers to survey questions that are recorded as numbers. We will need to go through each of these questions and convert the numbers to categories (strings). The information on how the survey questions were encoded can be found at the following website:
https://www.icpsr.umich.edu/icpsradmin/nsfg/variableGroupParent/14242?studyNumber=10001.

2.3 Description of Data

Originally, I wanted to look at medical data to answer some of the questions in my proposal. However, finding good medical data is very difficult. This dataset is a really good one, and I think some great information can be gleaned from it.

The following quotation from the CDC website (https://www.cdc.gov/nchs/nsfg/) explains what this dataset is for:

> The National Survey of Family Growth (NSFG) gathers information on family life, marriage and divorce, pregnancy, infertility, use of contraception, and men's and women's health. The survey results are used by the U.S. Department of Health and Human Services and others to plan health services and health education programs, and to do statistical studies of families, fertility, and health.

The 27 columns that remain after my data cleaning are described in the following table:

CASEID	Case Identification Number
PREGORDR	Pregnancy Order (number, e.g. 4 = 4th pregnancy)
BORNALIV	Number of babies born alive from this pregnancy
BIRTHWGT_LB1	Birthweight in Pounds - 1st baby from this pregnancy
BIRTHWGT_OZ1	Birthweight in Ounces - 1st baby from this pregnancy
EVUSEINT	Use any birth control method in pregnancy interval?
STOPDUSE	Before you became preg, stop using all b.c. methods?
WANTBOLD	Before pregnancy, mother want to have baby in future?
HPWNOLD	Before pregnancy, father want to have baby in future?
COHPBEG	Was R living w/father at beginning of pregnancy
COHPEND	Was R living w/father when preg ended/baby was born
PRGLNGTH	Duration of completed pregnancy in weeks
OUTCOME	Pregnancy outcome
AGECON	Age at Time of Conception
WANTPART	Wantedness of Pregnancy – Respondent's Partner
NEWWANTR	Wantedness of Pregnancy - Respondent
RMARITAL	Marital Status
EDUCAT	Education (Completed Years of Schooling)
RACE	Race
HISPANIC	Hispanic Origin (Yes or No)
PREGNUM	CAPI-based total number of pregnancies
PUBASSIS	Whether R received public assistance in prior calendar year
POVERTY	Poverty level income
LABORFOR	Labor force status
RELIGION	Current religious affiliation
METRO	Place of residence (Metropolitan / Nonmetropolitan)
WGT2015_2017	Final weight for the 2015-2017 NSFG (explained later)

2.4 Suitability of Data

This dataset does have some problems if we want to extrapolate its findings to the general public. The surveyors deliberately over-sampled certain populations (specifically minorities such as blacks and teenagers). This is because they were worried that they wouldn't have enough predictive power for these minorities if they surveyed them "to scale". In order to increase their precision for these subgroups of the population, they over-sampled these groups. This could cause problems in our statistical analysis if we hope to extrapolate the findings to the general U.S. population. In order to account for this oversampling of minorities (and hence undersampling of non-minorities), the dataset includes weights that can be used to generalize the data to the total U.S. population. The column 'WGT2015_2017' contains the weights that should be used for each sample. The NSFG explains the weights in the following way:

> For purposes of description, it may be useful to observe that the final weight can be interpreted as the number of persons in the population that an individual NSFG respondent represents. A final weight for a teenage Hispanic female of 2,000 means that this sample respondent represents herself and 1,999 other similar women in the population. The NSFG 2015-2017 final weights are values greater than 1, and when

> summed across a subgroup or the total sample are expected to provide an estimate of the total number of persons in that subgroup in the U.S. household population.

For more information on weights and how they were calculated, one can review the survey's documentation at https://www.cdc.gov/nchs/data/nsfg/PUF3-NSFG-2015-2017-Weighting-Design_02Oct2019.pdf.

Another potential problem with the suitability of this data stems from the fact that it is a survey. Sometimes people misrepresent the truth on surveys. This might be to appear better or it might be because the person's recollection isn't true. Often, the survey questions pertained to a pregnancy that was a few years in the past, which could have caused the respondent to forget details or give incorrect details. Additionally, the survey asked if a pregnancy ended in abortion. Since abortion is of questionable legality depending on the state and can be stigmatized, it's possible (and in my opinion, likely) that the number of abortions is underrepresented. Also, some instances of reported miscarriage may have actually been abortions.

Despite some of the problems with the dataset, this dataset is a very good one. The NSFG specifically traveled throughout the entire U.S., sampling individuals from every state. They further subdivided states into county-based areas, each of which was visited and sampled. This means that the survey should have good predictive power for the entire U.S. (although extrapolating internationally could be unwise).

3 Feature Engineering

The first step is one-hot encoding all data fields that are not numeric. These data fields include marital status, race, religion, whether the child was wanted, the outcome of the pregnancy, and others. These fields should be one-hot encoded so that regression can be performed on them.

```
one_hot_fields = ['EVUSEINT', 'STOPDUSE', 'WANTBOLD', 'HPWNOLD',
                  'COHPBEG', 'COHPEND', 'OUTCOME', 'WANTPART',
                  'NEWWANTR', 'RMARITAL', 'RACE', 'LABORFOR',
                  'RELIGION']

for field in one_hot_fields:
    one_hot = pd.get_dummies(df[field],prefix=field)
    df.drop(columns=[field],inplace=True)
    df = df.join(one_hot)
```

Add birthweights – originally, birthweights are recorded in two separate columns, one for pounds and one for ounces. We will convert both to kilograms and add them together in a new column. This is a good decision because it doesn't make sense to look at either of these columns separately. Just the number of ounces (less than a pound) that a baby weighed probably won't have any predictive power.

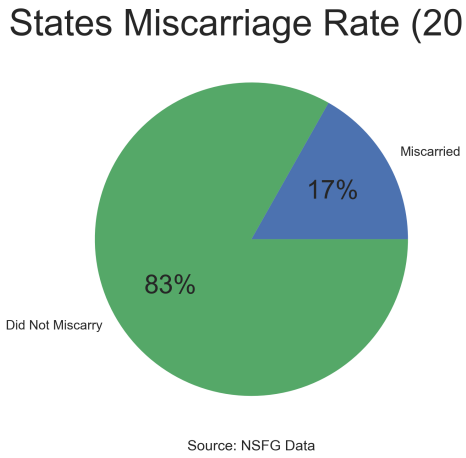
```
cols['BIRTHWGT_KG'] = 'Birthweight in Kilograms'
df['BIRTHWGT_KG'] = (
    pd.to_numeric(df['BIRTHWGT_LB1'])/2.205
    + pd.to_numeric(df['BIRTHWGT_OZ1'])/35.74)
df.drop(columns=['BIRTHWGT_LB1', 'BIRTHWGT_OZ1'],inplace=True)
```

Add a column that shows whether a specific person (denoted by CASEID) has ever had a failed pregnancy before the current pregnancy. This feature could be very useful in predicting future miscarriages.

```
df['HAS_HAD_MISC'] = 0
for case_id in set(df['CASEID']):
    has_had_miscarriage = 0
    bornaliv_col = df[df['CASEID'] == case_id]['BORNALIV']
    for idx in bornaliv_col.index:
        df.loc[idx, 'HAS_HAD_MISC'] = has_had_miscarriage
        if df.loc[idx, 'BORNALIV'] == 0:
            has_had_miscarriage = 1
```

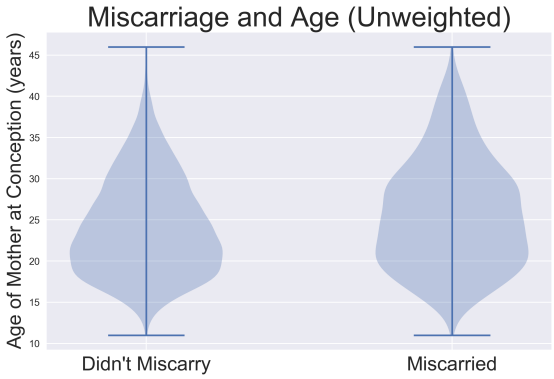
4 Analysis

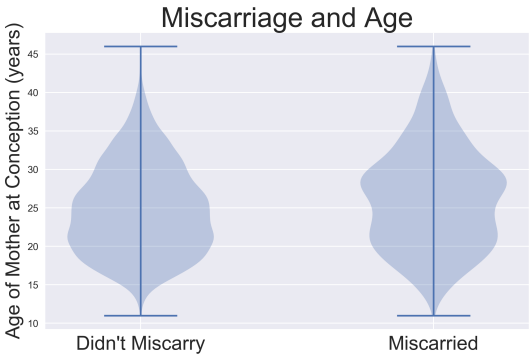
First, we estimate the percent of all pregnancies ending in miscarriage in the U.S. in the years 2015-2017. In order to obtain the correct percentages, we must account for the weight column. Specifically, we multiply each instance of miscarriage by the appropriate weight, and divide by the sum of all the weights, giving the estimated percentage for the entire U.S.



According to the data, a little under 20% of all pregnancies end in miscarriage. In my opinion, this is a much higher rate than most people expect (it's certainly higher than I expected before studying this topic). However, this rate aligns with the literature [2] and indicates trustworthy data.

Now, we want to find the columns that are most highly correlated with miscarriage. However, we cannot use the direct implementation of correlation from the pandas software. This is because pandas doesn't have an option to weight the correlation calculations. Using the unweighted data to compute correlation is problematic because the correlation values could be larger or smaller after weighting the data. For instance, look at the difference between the following two violin plots. These plots display the distributions of healthy pregnancies with mothers' ages at conception vs. miscarriages with mothers' ages.





We note that with the unweighted data, it seems that miscarriage is only slightly more likely as a woman ages. However, with the weighted data we see that miscarriage becomes much more common as women age. The reason for the discrepancy is most likely because the researchers over-sampled younger women, which is why the unweighted plot is skewed toward younger ages.

To correct for this effect, we compute the correlation coefficients using a weighted Pearson correlation,

$$\rho_{XY} = \frac{s_{XY}}{\sqrt{s_X s_Y}},$$

where

$$s_{XY} = \frac{\sum_i w_i (x_i - m_X)(y_i - m_Y)}{\sum_i w_i}$$

is the weighted covariance, using weighted means given by

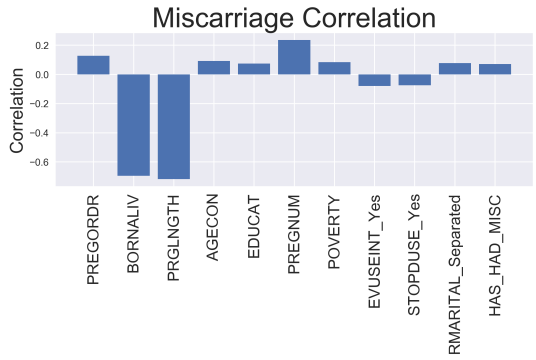
$$m_X = \frac{\sum_i w_i x_i}{\sum_i w_i}, \quad m_Y = \frac{\sum_i w_i y_i}{\sum_i w_i}.$$

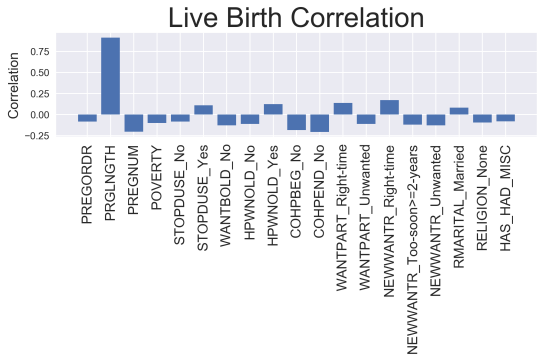
Our implementation is as follows:

```
def w_mean(x,w):
    """Computes the weighted mean of data x with weights w."""
    return np.sum(x*w)/np.sum(w)

def w_cov(x,y,w):
    """Computes the weighted covariance of data x,y with weights w."""
    m_X = w_mean(x,w)
    m_Y = w_mean(y,w)
    return np.sum(w*(x-m_X)*(y-m_Y))/np.sum(w)

def w_corr(x,y,w):
    """Computes the weighted Pearson correlation between data x,y
    with weights w.
    """
    s_XY = w_cov(x,y,w)
    s_X = w_cov(x,x,w)
    s_Y = w_cov(y,y,w)
    return s_XY/sqrt(s_X*s_Y)
```



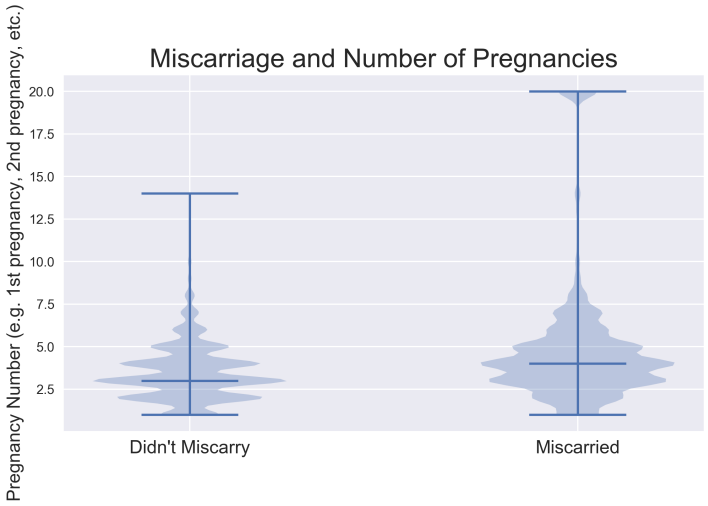


After calculating the weighted correlations, we plot the most highly correlated columns above.

Analyzing the above graphs, we see that some of the most highly correlated variables with miscarriage are obvious: being born alive is negatively correlated, and pregnancy length is negatively correlated. Some more interesting correlations include past miscarriages, age at conception, education, which pregnancy this is (e.g. 1st, 2nd...), and poverty. Perhaps counterintuitively, the pregnancy number seems to be more highly correlated with miscarriage than the mother's age.

A baby being born alive (i.e. not abortion, ectopic pregnancy, miscarriage, or other) has most of the same most highly correlated values, but is also negatively correlated with use of birth control and being nonreligous, and positively correlated with the baby being wanted. These are some very intriguing variables. It makes sense that the baby being wanted would be positively correlated with it being born alive since an unwanted baby is more likely to be aborted and recorded as either an abortion or some other type of pregnancy loss.

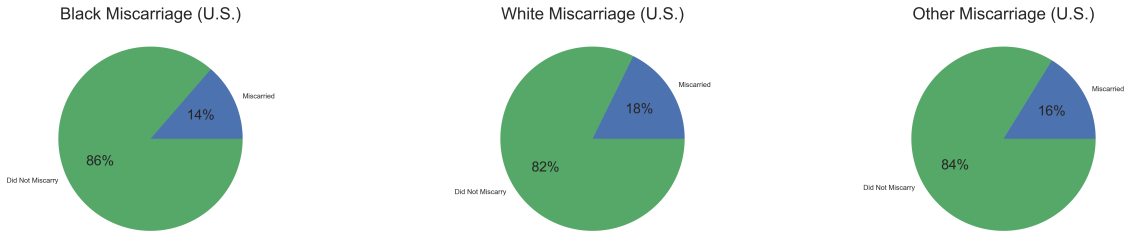
The following plot examines miscarriage based on pregnancy order. We might ask, is someone more likely to have a miscarriage if this is a later pregnancy (i.e. a 6th pregnancy rather than a 1st pregnancy)?



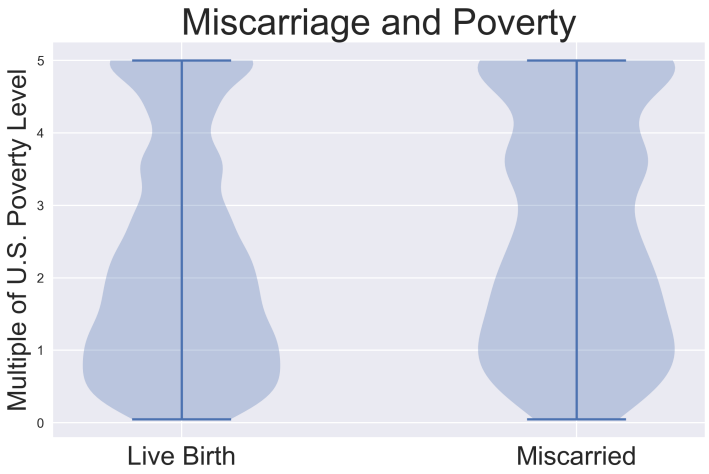
The first aspect of this plot that we notice is that the violin plots are wavy. This is caused by the fact that the y-axis, pregnancy number, can only be integer-valued. We also note that the median

pregnancy number for miscarried pregnancies is slightly higher than the median number for non-miscarried pregnancies. We also note that there are some outliers we're dealing with. Specifically, there is a single woman who had 20 pregnancies. Her 19th and 20th were both miscarriages. This outlier skews the "Miscarried" side of the plot and might be the reason why pregnancy number is more highly correlated with miscarriage than age.

Are miscarriage rates different between races?

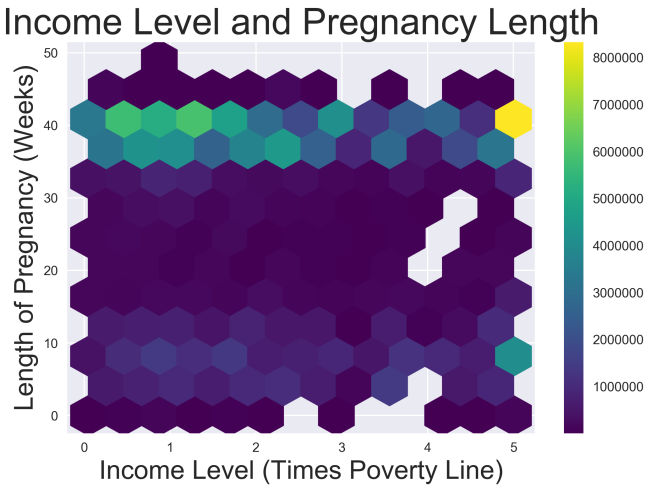


It looks like miscarriage rates are not significantly higher from one race to another. Granted, the survey only partitioned by blacks, whites, or others. It's possible that further partitioning may reveal differences.

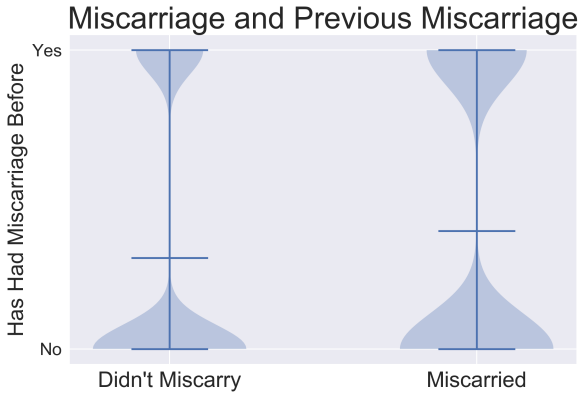


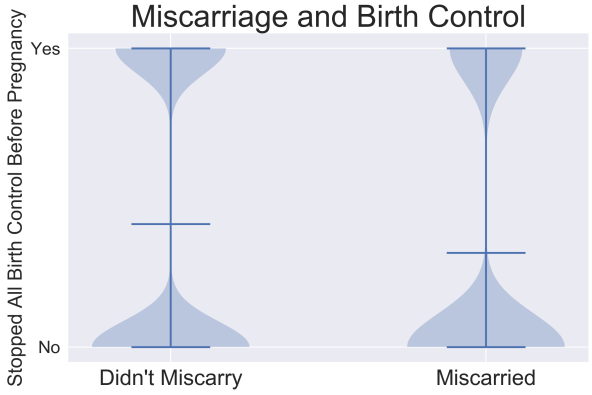
Another interesting factor that could influence miscarriage is poverty. Strangely, the correlation coefficient between miscarriage and income is positive. This indicates that being richer correlates with a higher miscarriage rate. The density plot above seems to indicate such a phenomenon. Pregnancies that result in live births are skewed toward lower income levels, meaning that more poor people tend to have such pregnancies. However, pregnancies that end in miscarriage are more evenly distributed among all income levels. This is definitely the opposite of what I would expect to see, and further analysis is warranted. Also, the above plot shows a large density concentration for both pregnancy outcomes right at 5 times the poverty level. At first, I wondered why this was. So, I read through the survey's documentation and found that if a person's income was greater than 5 times the poverty level, it was only recorded as a 5. Since the poverty level in the U.S. is around $25,000 for a family of four [1], there are many Americans making more than five times the poverty level income. This explains the high density in that area. Perhaps with data that contains higher income levels, we could better analyze the correlation between poverty and miscarriage. The

following graph visualizes the same phenomenon a different way, plotting the densities of income level and pregnancy length.



Finally, we plot the densities of miscarriage with previous miscarriage and birth control.





5 Conclusion

Our analysis agrees with the current literature in that age is positively correlated with miscarriage. However, our analysis showed that pregnancy number is more highly correlated (however, this may have been due to outliers). Also, we found a similar number of overall miscarriages compared to the literature.

Some interesting correlations that we found were education, poverty, use of birth control, and whether the baby was wanted. These could be correlated with miscarriage because they are correlated with abortion, and some abortions may have been recorded as miscarriages.

Having analyzed the data, we answer some of the questions asked in the introduction:

- Who is at most risk for miscarriage? Every woman is at risk for miscarriage. However, older women and those who've had multiple previous pregnancies are more at risk. Those with a history of miscarriage are also slightly more at risk.
- Are people of specific races, ethnicities, or socioeconomic statuses at higher risk for miscarriage? It seems that miscarriage is not correlated with being black, white or hispanic. It may be useful to examine other races/ethnicities. Also, poverty may be negatively correlated with miscarriage.
- Do previous miscarriages increase your risk of subsequent miscarriage? Yes, it seems that they do. Having a previous miscarriage seems to be positively correlated with having a subsequent miscarriage.
- Does smoking increase risk of miscarriage? Although the NSFG contained data on smoking habits, the majority of respondents didn't respond to this question. So, we weren't able to analyze possible correlations between smoking and miscarriage. This would be a good area in which to do more research.
- Is use of contraception related to miscarriage rates? Stopping the use of birth control before becoming pregnant is (barely) negatively correlated with miscarriage.

Now, even though there were correlations, most of the data were not highly correlated with miscarriage. This indicates that miscarriage may be mostly random. As the current scientific consensus holds that miscarriage is mainly caused by genetic/chromosomal abnormalities, our low correlations seem to support that consensus (or at least not contradict it). Since genetic abnormalities are thought to occur mostly randomly, this would support the randomness that we see in the data.

In conclusion, this data set may not be the most useful for predicting miscarriage. Our correlations are relatively low, and miscarriage randomly affects people. For future analysis, it could be very useful to examine health data such as HcG levels or ultrasound images. I believe that such analysis could yield more useful, predictive results. Until then, further study on the effects of age and repeated pregnancies on pregnancy outcomes could prove very fruitful in understanding and combating miscarriage.

References

[1] U.S. Department of Health & Human Services. "2019 Poverty Guidelines". In: (). URL: https://aspe.hhs.gov/2019-poverty-guidelines.

[2] Elina Hemminki and Erja Forssas. "Epidemiology of miscarriage and its relation to other reproductive events in Finland". In: *American Journal of Obstetrics and Gynecology* 181.2 (1999), pp. 396–401. ISSN: 0002-9378. DOI: https://doi.org/10.1016/S0002-9378(99)70568-5. URL: http://www.sciencedirect.com/science/article/pii/S0002937899705685.

[3] Gail Erlick Robinson. "Pregnancy loss". In: *Best Practice & Research Clinical Obstetrics & Gynaecology* 28.1 (2014). Perinatal Mental Health: Guidance for the Obstetrician-Gynaecologist, pp. 169–178. ISSN: 1521-6934. DOI: https://doi.org/10.1016/j.bpobgyn.2013.08.012. URL: http://www.sciencedirect.com/science/article/pii/S1521693413001247.

[4] *The John Hopkins Manual of Gynecology and Obstetrics*. 2. Lippincott Williams & Wilkins, 2012, pp. 438–439. ISBN: 9781451148015.

Appendix

The appendix consists of four jupyter notebooks that contain code that will collect, clean, feature engineer, and display all the data used in this project. Feel free to check it out and perform your own analysis. It is hosted on github at https://github.com/paskett/predicting_miscarriage.