A Deep Dive: Exploring the Impact of Maternal Characteristics on Preterm Birth through Multiple and Logistic Regression Analysis

D214: Data Analytics Graduate Capstone

Western Governs University

Table of Contents

Part I: Research Question

A1. Research Question

What maternal characteristics (e.g. age, ethnicity, education, socioeconomic status) significantly influence the likelihood of preterm birth?

A2. Justification

Data analysis through multiple regression and logistic regression modeling will allow for an objective assessment of the relationship between maternal characteristics and preterm birth. Utilizing population-level data, statical techniques can provide quantitative evidence to support or refute the association between variables with preterm birth. This analysis will investigate the relationship between material characteristics such as socioeconomic factors, medical conditions, and age on the likelihood of preterm birth. There are no definitive reasons for preterm births. However, there are risk factors that can increase the likelihood of women having a preterm birth. Some factors for delivering a preterm baby in the past include being pregnant with multiple fetuses, tobacco and/or substance abuse (*Premature Birth*, 2022). By identifying specific maternal factors related to preterm birth, healthcare providers can implement targeted interventions, monitoring, and care plans to mitigate the risk and improve patient outcomes.

A3. Context

Fuchs et al (2018) completed a similar study to examine maternal characteristics and preterm birth utilizing a multivariate logistic analysis (Fuchs et al., 2018). Fuchs et al found "chronic hypertension, assisted reproduction techniques, pre-gestational diabetes, invasive procedures in pregnancy, gestational diabetes and placenta praevia were linearly associated with increasing maternal age" (Fuchs et al., 2018). Fuchs et al concluded maternal age above 40 was associated with preterm birth and maternal age of 30-34 was associated with the lowest risk of preterm birth (Fuchs et al., 2018). Thus, logistic regression will be used to model the relationship between binary dependent variables and independent variables and predict binary outcomes in this analysis (Edgar & Manz, 2017b).

A4. Hypothesis

Given the project topic, exploring factors affecting preterm birth: a multiple and logistical regression analysis approach, this analysis will examine maternal characteristics and their relationship to the likelihood of preterm births. The hypotheses for this analysis are:

> ***Null hypothesis***- There is no significant association between individual maternal characteristics and the likelihood of preterm birth.
>
> ***Alternate Hypothesis***- There is a significant relationship between individual maternal characteristics and the likelihood of preterm birth.

The null hypothesis assumes there is no relationship between individual maternal characteristics and the likelihood of preterm birth. Thus, factors such as age, medications, medical conditions, body weight, education, and other variables associated with the mother are not associated with an increase or decrease in the risk of preterm birth. This hypothesis suggests any observed relationship is by chance and has no real underlying connection.

On the other hand, the alternative hypothesis proposes there is a significant relationship between individual maternal characteristics and the likelihood of preterm birth. Thus, certain maternal characteristics may be factors of preterm birth.

By conducting a multiple and logistic regression analysis the analysis aims to explore the relationship between various maternal characteristics and the likelihood of preterm birth.

This approach will allow for the control of confounding variables and examine the independent contribution of the different factors. This analysis will assess the statistical significance of these relationships and determine whether they support the alternative hypothesis or fail to reject the null hypothesis.

Part II: Data Collection

B1. Data Collected

Data containing maternal and paternal characteristics based on live births published by the Centers for Disease Control and Prevention's National Center for Health Statistics (NCHS) in 2020. The data is collected based on reported birth registry data from all 50 states and territories. "NCHS receives these files from the registration offices of all states, the two cities, and four territories through the Vital Statistics Cooperative Program" (*National Vital Statistics System (NVSS) - Health, United States*, n.d.). Births are requested to be reported promptly but laws vary from state to state, ranging from 24 hours to 10 days following birth (*National Vital Statistics System (NVSS) - Health, United States*, n.d.-b). Data can be retrieved from the website as a text (txt) file. However, the text files are coded based on a lengthy coding system published in the user guide on the CDC website. This analysis will utilize a downloadable CSV file from Kaggle for 2020 that has been encoded based on the NCHS user guidelines. The data set used consists of 180,992 observations and 39 columns. The variables in the dataset are continuous and categorical as shown below.

Variable Name	Description	Column Contents	Data Type
birth_year	year of birth	2020 Year of birth	categorical
birth_month	Month of birth	01 January 02 February 03 March 04 April 05 May 06 June 07 July 08 August 09 September 10 October 11 November 12 December	categorical
birth_time	time of birth	0000-2359 Time of Birth 9999 Not Stated	continuous
birth_place	Place of birth	1 Hospital 2 Freestanding Birth Center 3 Home (intended) 4 Home (not intended) 5 Home (unknown if intended) 6 Clinic / Doctor's Office 7 Other 9 Unknown	categorical
mother_age	Mother's age	12 10 – 12 years 13 - 49 years 50 years and over	continuous
marital_status	Martial status	1 Married 2 Unmarried	categorical
mother_education	Mother's education	1 8th grade or less 2 9th through 12th grade with no diploma 3 High school graduate or GED completed 4 Some college credit, but not a degree. 5 Associate degree (AA,AS) 6 Bachelor's degree (BA, AB, BS) 7 Master's degree (MA, MS, MEng, MEd, MSW, MBA) 8 Doctorate (PhD, EdD) or Professional Degree (MD, DDS, DVM, LLB, JD) 9 Unknown	categorical
father_age	Father's age	12 10 – 12 years 13-50 years 50 years and over	continuous

father_education	Father's education	1 8th grade or less 2 9th through 12th grade with no diploma 3 High school graduate or GED completed 4 Some college credit, but not a degree. 5 Associate degree (AA,AS) 6 Bachelor's degree (BA, AB, BS) 7 Master's degree (MA, MS, MEng, MEd, MSW, MBA) 8 Doctorate (PhD, EdD) or Professional Degree (MD, DDS, DVM, LLB, JD) 9 Unknown	categorical
interval_llb	Interval Since Last Live Birth	000-003 Plural delivery 004-300 Months since last live birth 888 Not applicable / no previous pregnancy 999 Unknown or not stated	continuous
cigarettes	Number of cigarettes before pregnancy	00-97 98 99 Number of cigarettes daily 98 or more cigarettes daily Unknown or not stated	continuous
mother_height	mother's height in inches	30-78 Height in inches 99 Unknown or not stated	continuous
mother_bmi	mother's pre-pregnancy body mass index	13.0-69.9 Body Mass Index 99.9 Unknown or not stated	continuous
pre_preg_weight	mother's pre-pregnancy weight in pounds	075-375 Weight in pounds	continuous
delivery_weight	mother's weight after delivery in pounds	100-400 Weight in pounds 999 Unknown or not stated	continuous
pre_preg_diabetes	diagnosis of diabetes prior to pregnancy	Y Yes N No U Unknown or not stated	categorical
gest_diabetes	pregnancy inducted diabetes	Y Yes N No U Unknown or not stated	categorical
pre_preg_hypertension	diagnosis of pregnancy prior to pregnancy	Y Yes N No U Unknown or not stated	categorical
gest_hypertension	pregnancy induced hypertension	Y Yes N No U Unknown or not stated	categorical
prev_preterm_birth	presence of previous preterm birth	Y Yes N No U Unknown or not stated	categorical
infertility_treatment	infertility treatments used	Y Yes N No U Unknown or not stated	categorical
prev_cesarian	previous cesarean delivery	Y Yes N No U Unknown or not stated	categorical
gonorrhea	maternal infection present and/or treated during pregnancy	Y Yes N No U Unknown or not stated	categorical
syphilis	maternal infection present and/or treated during pregnancy	Y Yes N No U Unknown or not stated	categorical
chlamydia	maternal infection present and/or treated during pregnancy	Y Yes N No U Unknown or not stated	categorical
hepatitis_b	maternal infection present and/or treated during pregnancy	Y Yes N No U Unknown or not stated	categorical
hepatitis_c	maternal infection present and/or treated during pregnancy	Y Yes N No U Unknown or not stated	categorical
labor_induction	induction of labor	Y Yes N No U Unknown or not stated	categorical
labor_augmentation	augmentation of labor	Y Yes N No U Unknown or not stated	categorical
steroids	received for fetal lung maturation received by the mother before delivery	Y Yes N No U Unknown or not stated	categorical
antibiotics	mother received during labor	Y Yes N No U Unknown or not stated	categorical

chorioamnionitis	clinical chorioamnionitis or maternal temperature >= 38 degrees Celsius (100.4 degrees Fahrenheit)	Y Yes N No U Unknown or not stated	categorical
anesthesia	epidural or spinal anesthesia during labor	Y Yes N No U Unknown or not stated	categorical
apgar5	Five Minute APGAR Score	00-10 A score of 0-10 99 Unknown or not stated	continuous
apgar10	Ten Minute APGAR Score	00-10 A score of 0-10 88 Not applicable 99 Unknown or not stated	continuous
plurality	if more than one infant shared the gestation and birth.	1 Single 2 Twin 3 Triplet 4 Quadruplet or higher	categorical
gender	infant gender	M Male F Female	categorical
infant_weight	infant weight at birth In grams	0227-8165 Number of grams	continuous

B2. Advantages and Disadvantages of Data-Gathering Methodology

This analysis used an existing database from NCHS. One advantage to using an existing database is the large same size. A disadvantage to using an existing database was limited variable options. The original data from NCHS includes a variety of variables. Many of these variables, such as race, were not used in the sourced data from Kaggle that was used in this analysis.

The primary limitation of this data set is it only contains data for one year. In a multiple regression analysis, multicollinearity can be a limitation. This can occur when independent variables are highly correlated with each other, making it challenging to determine the independent effect on the variable outcome (*Multiple Regression*, n.d.). This analysis focuses on finding significant relationships between maternal characteristics and preterm birth. This may show correlation, but causation is more difficult to determine due to confounding factors.

Some delimitations exist in completing this analysis. Data over several years could provide more statistically significant correlations over time. The data is also limited in maternal characteristics and paternal characteristics. More data regarding both parents prior to birth could provide more statistical significance in determining factors contributing to preterm birth. Socioeconomic factors such as receiving access to prenatal care, household income, and geographic location could provide more information contributing to preterm birth.

B3. Challenges in the data collection process

The data was sourced from Kaggle for public use. The data is published by the Centers for Disease Control and Prevention's National Center for Health Statistics (NCHS) and is a public use data published on their website. The original data is from 2020 natality micro-data and may be downloaded using their site tools. The dataset from Kaggle included continuous and categorical variables. Many categorical variables had qualitative characteristics/groupings that required encoding. One of the variables in the data set, previous cesarean delivery (prev_cesarian) was initially supposed to be a categorical variable with yes, no, and unknown responses. However, the Kaggle data source responses were 0 and 9. There was no reference for the interpretation of the 0 and 9 responses and appeared to be a data transformation error when coding from the original source. Thus, this variable was dropped and not used in for this analysis.

Part III: Data Extraction and Preparation

C1. Data Extraction and Preparation Process

The data was cleaned using Python by detecting duplicate data, missing values, outliers, and any other data quality issues in the churn data set. To start the data cleaning process the data types were determined. The data type of each variable is needed for understating because certain functions work only with specific functions. This includes the column names and the number of non-null values for each column.

```
df.info(file_path)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 180992 entries, 0 to 180991
Data columns (total 39 columns):
 #   Column                  Non-Null Count   Dtype  
---  ------                  --------------   -----  
 0   id                      180992 non-null  int64  
 1   birth_year              180992 non-null  int64  
 2   birth_month             180992 non-null  int64  
 3   birth_time              180969 non-null  float64
 4   birth_place             180977 non-null  float64
 5   mother_age              180992 non-null  int64  
 6   marital_status          180992 non-null  object 
 7   mother_education        178448 non-null  float64
 8   father_age              158408 non-null  float64
 9   father_education        154379 non-null  float64
 10  interval_llb            175142 non-null  float64
 11  cigarettes              180231 non-null  float64
 12  mother_height           180095 non-null  float64
 13  mother_bmi              177419 non-null  float64
 14  pre_preg_weight         177917 non-null  float64
 15  delivery_weight         178920 non-null  float64
 16  pre_preg_diabetes       180764 non-null  object 
 17  gest_diabetes           180764 non-null  object 
 18  pre_preg_hypertension   180764 non-null  object 
 19  gest_hypertension       180764 non-null  object 
 20  prev_preterm_birth      180764 non-null  object 
 21  infertility_treatment   180764 non-null  object 
 22  prev_cesarian           180992 non-null  int64  
 23  gonorrhea               180416 non-null  object 
 24  syphilis                180416 non-null  object 
 25  chlamydia               180416 non-null  object 
 26  hepatitis_b             180416 non-null  object 
 27  hepatitis_c             180416 non-null  object 
 28  labor_induction         180883 non-null  object 
 29  labor_augmentation      180883 non-null  object 
 30  steroids                180883 non-null  object 
 31  antibiotics             180883 non-null  object 
 32  chorioamnionitis        180883 non-null  object 
 33  anesthesia              180883 non-null  object 
 34  apgar5                  180219 non-null  float64
 35  apgar10                 180079 non-null  float64
 36  plurality               180992 non-null  int64  
 37  gender                  180992 non-null  object 
 38  infant_weight           180992 non-null  int64  
dtypes: float64(13), int64(7), object(19)
memory usage: 53.9+ MB
```

Once datatypes are known the data could then be cleaned. Cleaning and treating the data included detecting duplicates, and identifying missing values, and outliers. The data set sparsity is 1.06%.

Sparsity

```
# Calculate missing value percentages for each variable
missing_percentages = df.isnull().sum() / len(df) * 100

# Sort variables based on missing value percentages
missing_percentages = missing_percentages.sort_values(ascending=Fals

# Print the results
print(missing_percentages)
```

```
father_education         14.703965
father_age               12.477900
interval_llb              3.232187
mother_bmi                1.974120
pre_preg_weight           1.698970
mother_education          1.405587
delivery_weight           1.144802
apgar10                   0.504442
mother_height             0.495602
apgar5                    0.427091
cigarettes                0.420461
hepatitis_c               0.318246
hepatitis_b               0.318246
chlamydia                 0.318246
syphilis                  0.318246
gonorrhea                 0.318246
infertility_treatment     0.125972
gest_hypertension         0.125972
prev_preterm_birth        0.125972
pre_preg_diabetes         0.125972
pre_preg_hypertension     0.125972
gest_diabetes             0.125972
anesthesia                0.060224
chorioamnionitis          0.060224
labor_induction           0.060224
labor_augmentation        0.060224
steroids                  0.060224
antibiotics               0.060224
birth_time                0.012708
birth_place               0.008288
plurality                 0.000000
gender                    0.000000
id                        0.000000
prev_cesarian             0.000000
birth_year                0.000000
marital_status            0.000000
mother_age                0.000000
birth_month               0.000000
infant_weight             0.000000
dtype: float64
```

```
# Calculate the percentage of missing values for the entire dataset
total_missing_percentage = df.isnull().sum().sum() / (df.shape[0] * df.shape[1]) * 100

# Print the result
print("Data Sparsity: {:.2f}%".format(total_missing_percentage))
```

```
Data Sparsity: 1.06%
```

Initially, the missing data was attempted to be treated by removing only rows with 25% or more missing data (Collins et al., 2001). However, when viewing the shape of the data set before and after this treatment, the shape remained the same and no missing data remained.

```
#Remove rows with >25% of missing data (Collins et al., 2001)

# Identify rows with more than 25% missing values
rows_to_delete = missing_percentages[missing_percentages > 25].index

# Delete the identified rows from the DataFrame
df = df.drop(rows_to_delete, axis=0)

# Optional: Reset the index if needed
df = df.reset_index(drop=True)

# Check the resulting DataFrame
df
```

The treatment of missing values then included deletion of missing values since there was a minimal amount in the dataset.

Treat Missing Values

```
# Detect missing values
missing_values = df.isnull().any(axis=1)

# Delete rows with missing values
df = df.dropna()
```

```
df.isnull().sum()
```

```
id                        0
birth_year                0
birth_month               0
birth_time                0
birth_place               0
mother_age                0
marital_status            0
mother_education          0
father_age                0
father_education          0
interval_llb              0
cigarettes                0
mother_height             0
mother_bmi                0
pre_preg_weight           0
delivery_weight           0
pre_preg_diabetes         0
gest_diabetes             0
pre_preg_hypertension     0
gest_hypertension         0
prev_preterm_birth        0
infertility_treatment     0
prev_cesarian             0
gonorrhea                 0
syphilis                  0
chlamydia                 0
hepatitis_b               0
hepatitis_c               0
labor_induction           0
labor_augmentation        0
steroids                  0
antibiotics               0
chorioamnionitis          0
anesthesia                0
apgar5                    0
apgar10                   0
plurality                 0
gender                    0
infant_weight             0
dtype: int64
```

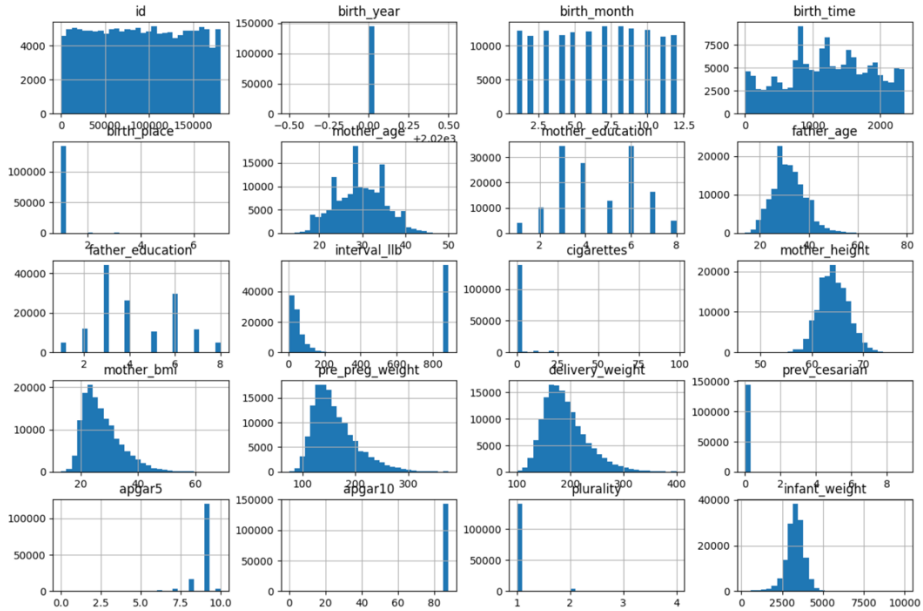
The descriptive statistics of the data were then viewed to further examine the distribution of the dataset by finding various statistical measures.

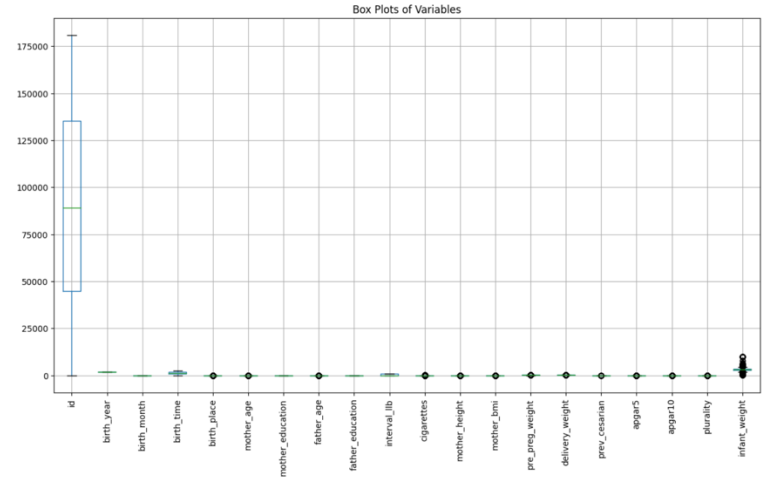
```
df.describe()
```

	id	birth_year	birth_month	birth_time	birth_place	mother_age	mother_education	father_age	father_education
count	144761.000000	144761.0	144761.000000	144761.000000	144761.000000	144761.000000	144761.000000	144761.000000	144761.000000
mean	89882.060203	2020.0	6.504121	1230.318256	1.035997	29.572806	4.574865	31.881819	4.289222
std	52068.390975	0.0	3.418868	628.934976	0.300862	5.623284	1.741075	6.734524	1.745398
min	1.000000	2020.0	1.000000	0.000000	1.000000	13.000000	1.000000	14.000000	1.000000
25%	44780.000000	2020.0	4.000000	801.000000	1.000000	26.000000	3.000000	27.000000	3.000000
50%	89254.000000	2020.0	7.000000	1237.000000	1.000000	30.000000	4.000000	32.000000	4.000000
75%	135251.000000	2020.0	9.000000	1731.000000	1.000000	34.000000	6.000000	36.000000	6.000000
max	180987.000000	2020.0	12.000000	2359.000000	7.000000	50.000000	8.000000	79.000000	8.000000

interval_llb	cigarettes	mother_height	mother_bmi	pre_preg_weight	delivery_weight	prev_cesarian	apgar5	apgar10	plurality
144761.000000	144761.000000	144761.000000	144761.000000	144761.000000	144761.000000	144761.000000	144761.000000	144761.000000	144761.000000
380.026692	0.715193	64.133600	27.400499	160.483162	189.872583	0.001554	8.789384	87.161273	1.027618
411.086556	3.927033	2.831993	6.726773	41.614455	41.044898	0.118263	0.755781	8.242547	0.168285
3.000000	0.000000	47.000000	13.100000	75.000000	100.000000	0.000000	0.000000	0.000000	1.000000
32.000000	0.000000	62.000000	22.500000	130.000000	160.000000	0.000000	9.000000	88.000000	1.000000
76.000000	0.000000	64.000000	25.800000	150.000000	183.000000	0.000000	9.000000	88.000000	1.000000
888.000000	0.000000	66.000000	31.000000	180.000000	212.000000	0.000000	9.000000	88.000000	1.000000
888.000000	98.000000	78.000000	68.700000	375.000000	400.000000	9.000000	10.000000	88.000000	4.000000

Outliers come from data entry errors, measurement errors, experimental errors, sampling errors, or novelties in the data (Lacrose & Lacrose, 2019). Outliers were visualized using histograms and boxplots. There were no significant outliers in the dataset.





The df.nunique() was then used to explore the uniqueness of the values within each column in the dataset. This also helps identify columns with high cardinality that might need special processing.

```
df.nunique()
```

```
id                        144761
birth_year                     1
birth_month                   12
birth_time                  1440
birth_place                    7
mother_age                    38
marital_status                 5
mother_education               8
father_age                    64
father_education               8
interval_llb                 288
cigarettes                    35
mother_height                 32
mother_bmi                   512
pre_preg_weight              300
delivery_weight              300
pre_preg_diabetes              2
gest_diabetes                  2
pre_preg_hypertension          2
gest_hypertension              2
prev_preterm_birth             2
infertility_treatment          2
prev_cesarian                  2
gonorrhea                      2
syphilis                       2
chlamydia                      2
hepatitis_b                    2
hepatitis_c                    2
labor_induction                2
labor_augmentation             2
steroids                       2
antibiotics                    2
chorioamnionitis               2
anesthesia                     2
apgar5                        11
apgar10                       12
plurality                      4
gender                         2
infant_weight               3394
dtype: int64
```

There are two types of data from this data set, quantitative and qualitative data. Qualitative or categorical data (e.g. yes/no) requires re-expression or encoding of numbers to perform statical modeling (Lacrose & Lacrose, 2019). There were categorical variables such as education that represented qualitative characteristics. These were transformed by assigning numeric codes.

birth_month

```
# Define the mapping dictionary for months
month_mapping = {
    'January': 1,
    'February': 2,
    'March': 3,
    'April': 4,
    'May': 5,
    'June': 6,
    'July': 7,
    'August': 8,
    'September': 9,
    'October': 10,
    'November': 11,
    'December': 12
}

# Convert the 'birth_month' column to a continuous variable
df['birth_month'] = df['birth_month'].replace(month_mapping)

df['birth_month']
```

birth_place

```
# Define the mapping dictionary for birth_place
birth_place_mapping = {
    'Hospital': 1,
    'Freestanding Birth Center': 2,
    'Home (intended)': 3,
    'Home (not intended)': 4,
    'Home (unknown if intended)': 5,
    'Clinic / Doctor's Office': 6,
    'Other': 7,
    'Unknown': 9
}

# Convert the 'birth_place' column to a continuous variable
df['birth_place'] = df['birth_place'].replace(birth_place_mapping).astype(int)

df['birth_place']
```

mother_education

```
# Create a dictionary to map the education categories to numerical values
mother_education_mapping = {
    '8th grade or less': 1,
    '9th through 12th grade with no diploma': 2,
    'High school graduate or GED completed': 3,
    'Some college credit, but not a degree': 4,
    'Associate degree (AA,AS)': 5,
    'Bachelor's degree (BA, AB, BS)': 6,
    'Master's degree (MA, MS, MEng, MEd, MSW, MBA)': 7,
    'Doctorate (PhD, EdD) or Professional Degree (MD, DDS, DVM, LLB, JD)': 8,
    'Unknown': 9
}

# Replace the values in the 'mother_education' column with the numerical encoding
df['mother_education'] = df['mother_education'].replace(mother_education_mapping)

# Convert the 'mother_education' column to integer type
df['mother_education'] = df['mother_education'].astype(int)

df['mother_education']
```

father_education

```
# Create a dictionary to map the education categories to numerical values
father_education_mapping = {
    '8th grade or less': 1,
    '9th through 12th grade with no diploma': 2,
    'High school graduate or GED completed': 3,
    'Some college credit, but not a degree': 4,
    'Associate degree (AA,AS)': 5,
    'Bachelor's degree (BA, AB, BS)': 6,
    'Master's degree (MA, MS, MEng, MEd, MSW, MBA)': 7,
    'Doctorate (PhD, EdD) or Professional Degree (MD, DDS, DVM, LLB, JD)': 8,
    'Unknown': 9
}

# Replace the values in the 'father_education' column with the numerical encoding
df['father_education'] = df['father_education'].replace(father_education_mapping)

# Convert the 'father_education' column to integer type
df['father_education'] = df['father_education'].astype(int)

df['father_education']
```

One hot encoding was then used to transform categorical data into nominal data to be used in the regression models. With one hot encoding, each categorical variable is represented as a binary vector where all elements are zero except the element corresponding to the category which is set to one utilizing dummy columns.

One Hot Encoding with Dummy Variables (continous variables)

```
# Create dummy variables in order to encode categorical, yes/no data points into 1/0 numerical values.
df['Dummy_marital_status'] = [1 if v == '1' else 0 for v in df['marital_status']]
df['Dummy_pre_preg_diabetes'] = [1 if v == 'Y' else 0 for v in df['pre_preg_diabetes']]
df['Dummy_gest_diabetes'] = [1 if v == 'Y' else 0 for v in df['gest_diabetes']]
df['Dummy_pre_preg_hypertension'] = [1 if v == 'Y' else 0 for v in df['pre_preg_hypertension']]
df['Dummy_gest_hypertension'] = [1 if v == 'Y' else 0 for v in df['gest_hypertension']]
df['Dummy_prev_preterm_birth'] = [1 if v == 'Y' else 0 for v in df['prev_preterm_birth']]
df['Dummy_infertility_treatment'] = [1 if v == 'Y' else 0 for v in df['infertility_treatment']]
df['Dummy_gonorrhea'] = [1 if v == 'Y' else 0 for v in df['gonorrhea']]
df['Dummy_syphilis'] = [1 if v == 'Y' else 0 for v in df['syphilis']]
df['Dummy_chlamydia'] = [1 if v == 'Y' else 0 for v in df['chlamydia']]
df['Dummy_hepatitis_b'] = [1 if v == 'Y' else 0 for v in df['hepatitis_b']]
df['Dummy_hepatitis_c'] = [1 if v == 'Y' else 0 for v in df['hepatitis_c']]
df['Dummy_labor_induction'] = [1 if v == 'Y' else 0 for v in df['labor_induction']]
df['Dummy_labor_augmentation'] = [1 if v == 'Y' else 0 for v in df['labor_augmentation']]
df['Dummy_steroids'] = [1 if v == 'Y' else 0 for v in df['steroids']]
df['Dummy_antibiotics'] = [1 if v == 'Y' else 0 for v in df['antibiotics']]
df['Dummy_chorioamnionitis'] = [1 if v == 'Y' else 0 for v in df['chorioamnionitis']]
df['Dummy_anesthesia'] = [1 if v == 'Y' else 0 for v in df['anesthesia']]
df['Dummy_gender'] = [1 if v == 'M' else 0 for v in df['gender']]
df['Dummy_steroids'] = [1 if v == 'Y' else 0 for v in df['steroids']]
```

Original columns were then dropped and dummy columns remained to be used in the regression models.

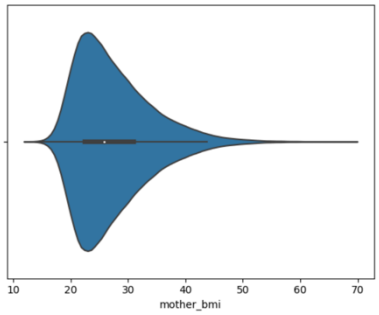
```
# Drop original categorical features from dataframe and unnecessary columns
df = df.drop(columns=['id','prev_cesarian', 'marital_status', 'pre_preg_diabetes', 'gest_diabetes','pre_preg_hypertensi
                      'infertility_treatment','prev_preterm_birth', 'gonorrhea', 'syphilis', 'chlamydia', 'hepatitis_b'
                     'hepatitis_c','labor_induction', 'labor_augmentation','antibiotics', 'chorioamnionitis', 'steroids
                     'anesthesia', 'gender'])
```

C2. Tools and Techniques

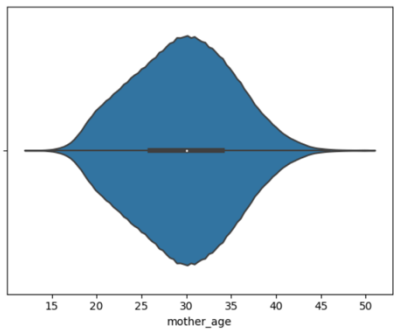
Exploratory data analysis (EDA) and multiple regression analysis were used to analyze the relationship between individual maternal characteristics and the likelihood of preterm birth. Exploratory data analysis provides an understanding of the data by identifying patterns and generating initial insights. Overall, EDA is beneficial because it allows for data exploration, visualization, pattern recognition, feature engineering, missing data analysis, and primarily hypothesis generation.

Univariate statistics is the statical analysis of a single variable at one time (Bruce et al., 2020). Below is the distribution of all independent variables in this analysis. Histograms and boxplots were used to visualize the distribution of each variable.

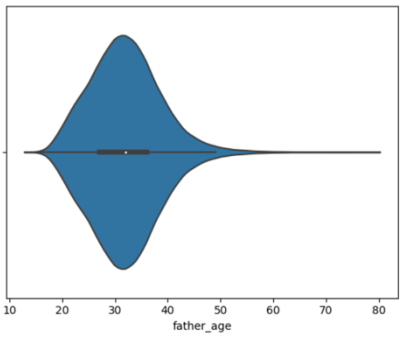
```
sns.violinplot(x=df['mother_bmi'])
<AxesSubplot: xlabel='mother_bmi'>
```



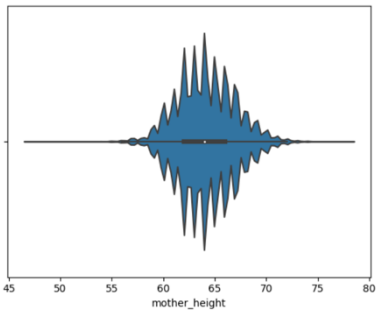

```
sns.violinplot(x=df['mother_age'])
<AxesSubplot: xlabel='mother_age'>
```



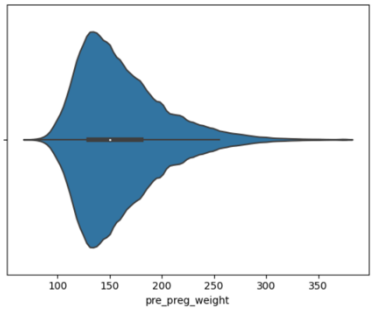

```
sns.violinplot(x=df['father_age'])
<AxesSubplot: xlabel='father_age'>
```



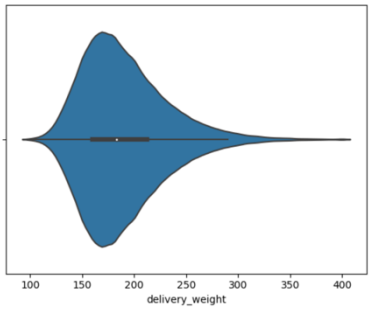

```
sns.violinplot(x=df['mother_height'])
<AxesSubplot: xlabel='mother_height'>
```



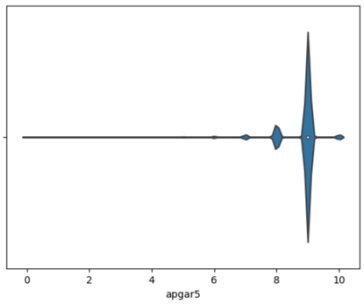

```
sns.violinplot(x=df['pre_preg_weight'])
<AxesSubplot: xlabel='pre_preg_weight'>
```



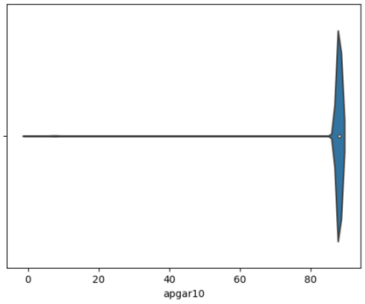

```
sns.violinplot(x=df['delivery_weight'])
<AxesSubplot: xlabel='delivery_weight'>
```



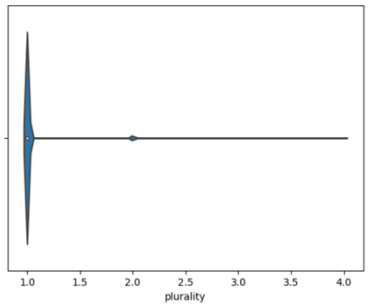

```
sns.violinplot(x=df['apgar5'])
<AxesSubplot: xlabel='apgar5'>
```



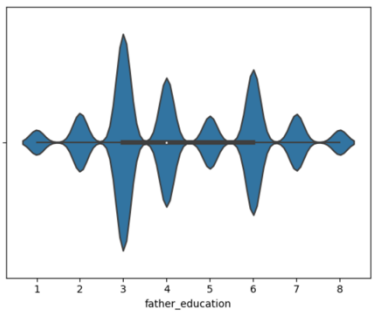

```
sns.violinplot(x=df['apgar10'])
<AxesSubplot: xlabel='apgar10'>
```



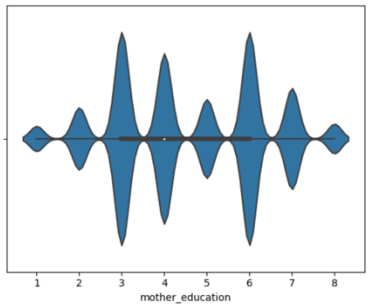

```
sns.violinplot(x=df['plurality'])
<AxesSubplot: xlabel='plurality'>
```



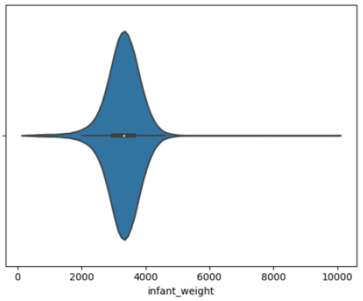

```
sns.violinplot(x=df['father_education'])
<AxesSubplot: xlabel='father_education'>
```



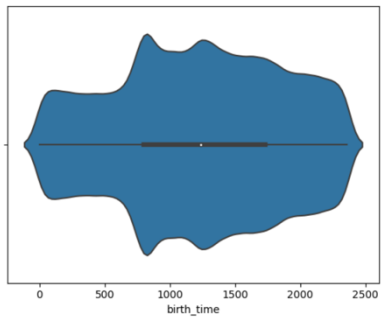

```
sns.violinplot(x=df['mother_education'])
<AxesSubplot: xlabel='mother_education'>
```




```
sns.violinplot(x=df['infant_weight'])
<AxesSubplot: xlabel='infant_weight'>
```

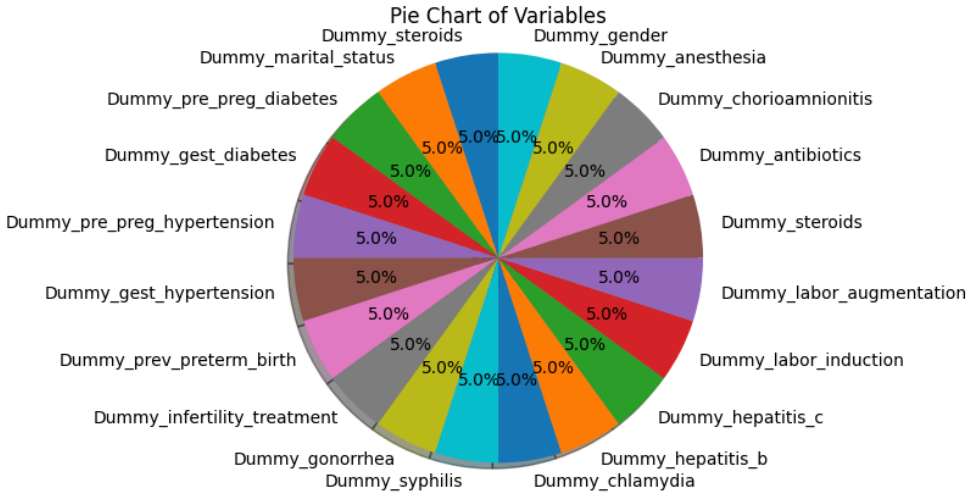


```
sns.violinplot(x=df['birth_time'])
<AxesSubplot: xlabel='birth_time'>
```

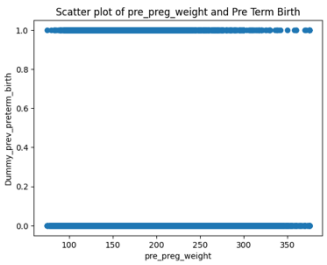


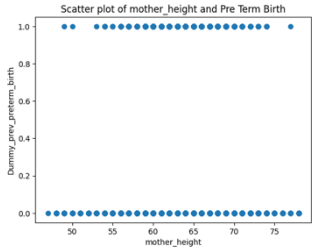
Pie Chart

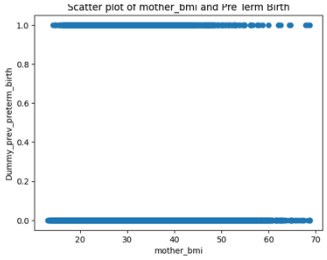
```
3]: frequencies = [df[variable].value_counts().sum() for variable in variables]
    plt.pie(frequencies, labels=variables, autopct='%1.1f%%', shadow=True, startangle=90)
    plt.axis('equal')
    plt.title('Pie Chart of Variables')
    plt.show()
```

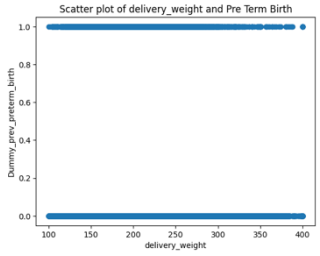


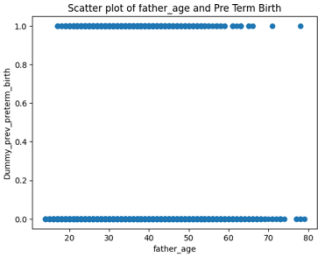
Bivariate statistical analysis refers to the statistical analysis of two variables at once (Bruce et al., 2020). A scatterplot was created for each independent variable to examine the relationship with the dependent variable.

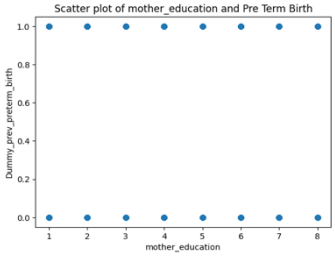




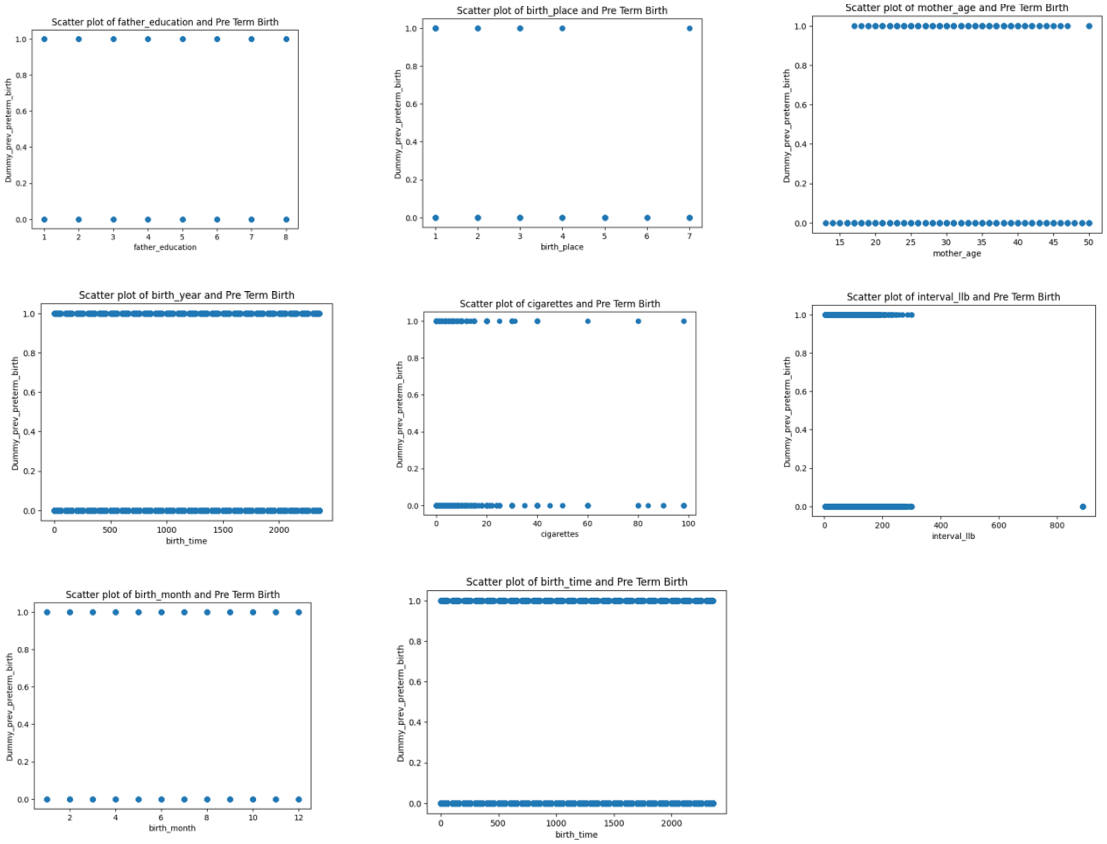




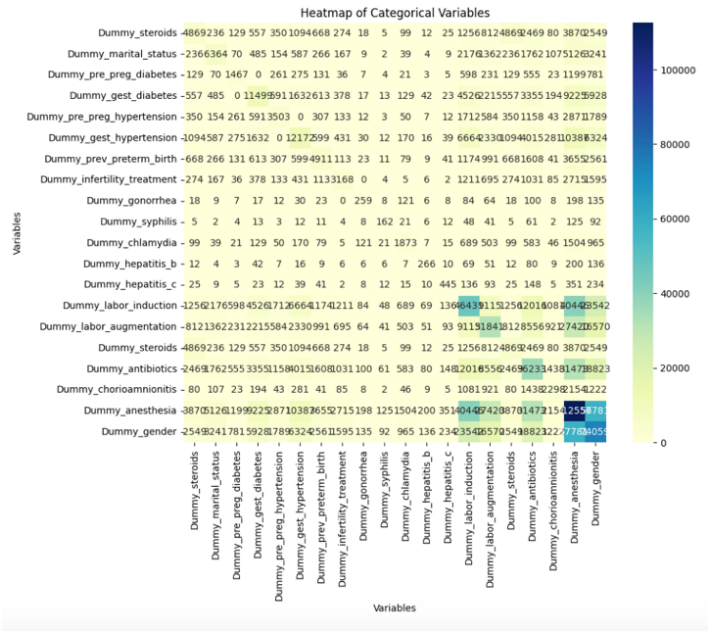


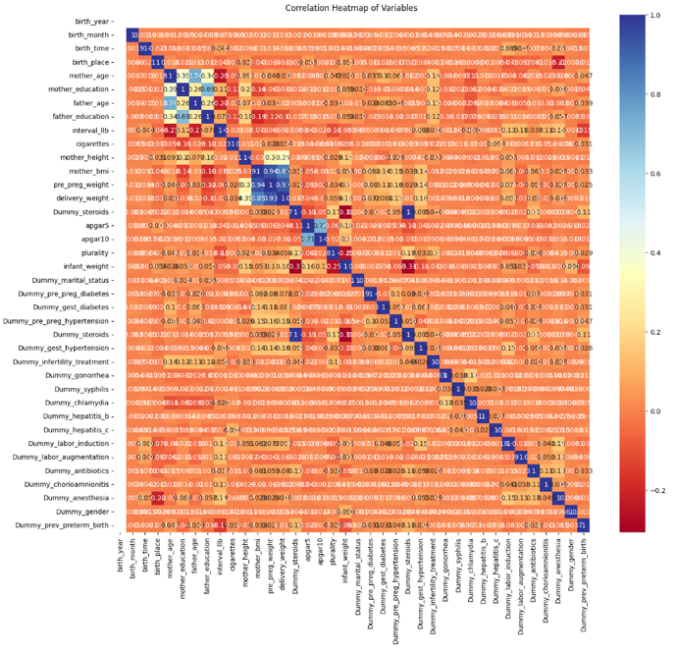


Maternal Characteristics and the Likelihood of Preterm Birth



Heatmaps:





Kolmogorov-Smirnov Test is a normality test used to assess whether a sample of data follows a specific distribution or if it differs significantly from that distribution.

Kolmogorov-Smirnov Test

```
from scipy.stats import kstest, norm

np.random.seed(123)
sample = np.random.normal(loc=0, scale=1, size=100)
ks_stat, p_value = kstest(sample, norm.cdf)
print("Kolmogorov-Smirnov test:")
print("KS statistic:", ks_stat)
print("p-value:", p_value)
```

```
Kolmogorov-Smirnov test:
KS statistic: 0.10665330380822602
p-value: 0.19116286378085373
```

Shapiro-Wilk Test is also a normality test used to assess the distribution of the data. This test was also used due to the smaller data set.

Shapiro-Wilk Test

```
from scipy.stats import shapiro

# Perform the Shapiro-Wilk test
statistic, p_value = shapiro(df['Dummy_prev_preterm_birth'])

# Print the results
print("Shapiro-Wilk test statistic:", statistic)
print("P-value:", p_value)
```

```
Shapiro-Wilk test statistic: 0.17307919263839722
P-value: 0.0
```

C3. Justification for Tools and Techniques

Python is an open-sourced programing language used for analysis and development. Python has a consistent syntax that makes coding and debugging user-friendly for beginners. Python has a simple syntax and readability. Python is flexible and has the ability to import packages and tailor them to user needs. "Python being a general-purpose tool encourages participation from users outside the Data Science community which enhances package availability" (Brittain et al, 2018). Although SAS has many preferred advantages, Python is the preferred choice for this dataset due to its smaller data size and beginner-friendly useability. Python is a general-purpose programming language while R is a statistical programming language. This makes Python more versatile and used for a wide range of tasks such as machine learning (Luna, 2022).

EDA will provide a comprehensive understanding of the dataset, including distributions, patterns, and potential outliers. EDA will also add in selecting relevant variables for the multiple regression analysis. Multiple regression analysis relies on assumptions of linearity, independence, and normality of residuals. EDA will for these assumptions to be checked through the examination of scatter plots, residual plots, and normality tests, ensuring the validity of the regression analysis results. Kolmogorov-Smirnov and Shapiro-Wilk tests are normality tests that will be used as part of the EDA process to assess the goodness of fit between an observed data sample and a specified probability distribution. Regression analysis allows for the quantitative assessment of the relationship between the dependent and independent variables.

Part IV: Analysis

D1. Description of Technique Used

In determining the relationship between maternal characteristics and preterm birth this analysis will utilize a multiple regression analysis and a logistic regression analysis. Multiple regression analysis can help identify significant predictors and quantify their impact on preterm birth. Multiple regression analysis is appropriate for analyzing the relationship between the dependent (preterm birth) and independent variable (maternal characteristics) rates (Turvey, 2013). Logistic regression is a statistical method used to analyze the relationship between a binary dependent variable and the independent variable(s). In this analysis, the dependent variable is a categorical variable (preterm birth) and has binary values of 0 and 1 (coded from yes/no responses). The logistic regression model uses the sigmoid function to estimate the probability of the dependent variable being in a particular category based on the values of the independent variable (Lacrose & Lacrose, 2019).

D2. Calculations and Output

Multiple Regression Analysis:

The initial multiple regression model contained 35 independent variables to be analyzed against the dependent variable, preterm birth. The model r-squared is 0.046 indicating approximately 4.6% of the variability in the outcome variable can be explained by the independent variable. The F-statistic demonstrates the overall significance of the regression model. The model has a F-statistic of 197. A low p-value associated with the F-statistic suggest that the model is statistically significant. The initial model provides the estimated coefficients who represent the average change in the dependent variable associated with a one-unit increase in the corresponding independent variable. The model also provides p-values for each independent variable. P-values less than 0.05 indicate statistical significance for the individual independent variables. The AIC and BIC measure model fitness and can be used for model comparison. The initial model AIC was -.9.069e+04 and the BIC was -9.033e+04. Lower AIC and BIC values indicate better model fit. The reduced model had minimal change with an R-squared of 0.045, AIC of -9.063e+04, and BIC -9.042e+04.

OLS Regression Results

Dep. Variable:	Dummy_prev_preterm_birth	R-squared:	0.046
Model:	OLS	Adj. R-squared:	0.045
Method:	Least Squares	F-statistic:	197.8
Date:	Thu, 29 Jun 2023	Prob (F-statistic):	0.00
Time:	18:03:01	Log-Likelihood:	45381.
No. Observations:	144761	AIC:	-9.069e+04
Df Residuals:	144725	BIC:	-9.033e+04
Df Model:	35		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
birth_year	4.782e-05	2.23e-05	2.146	0.032	4.14e-06	9.15e-05
birth_month	-0.0002	0.000	-1.767	0.077	-0.001	2.62e-05
birth_time	-5.905e-07	7.44e-07	-0.794	0.427	-2.05e-06	8.68e-07
birth_place	-0.0056	0.002	-3.505	0.000	-0.009	-0.002
mother_age	0.0007	0.000	4.885	0.000	0.000	0.001
mother_education	-0.0021	0.000	-5.377	0.000	-0.003	-0.001
father_age	-9.368e-05	0.000	-0.902	0.367	-0.000	0.000
father_education	-0.0009	0.000	-2.310	0.021	-0.002	-0.000
interval_llb	-6.779e-05	1.25e-06	-54.019	0.000	-7.02e-05	-6.53e-05
cigarettes	0.0008	0.000	6.459	0.000	0.001	0.001
mother_height	0.0012	0.001	1.713	0.087	-0.000	0.003
mother_bmi	0.0014	0.001	1.725	0.085	-0.000	0.003
pre_preg_weight	-0.0003	0.000	-2.540	0.011	-0.001	-7.92e-05
delivery_weight	0.0002	3.27e-05	4.835	0.000	9.4e-05	0.000
apgar5	-0.0030	0.001	-3.370	0.001	-0.005	-0.001
apgar10	6.873e-05	8.09e-05	0.850	0.395	-8.98e-05	0.000
plurality	-0.0279	0.003	-9.430	0.000	-0.034	-0.022
infant_weight	-2.488e-05	9.31e-07	-26.711	0.000	-2.67e-05	-2.31e-05
Dummy_marital_status	0.0075	0.002	3.301	0.001	0.003	0.012
Dummy_pre_preg_diabetes	0.0388	0.005	8.269	0.000	0.030	0.048
Dummy_gest_diabetes	0.0143	0.002	8.142	0.000	0.011	0.018
Dummy_pre_preg_hypertension	0.0365	0.003	11.711	0.000	0.030	0.043
Dummy_steroids	0.0762	0.003	27.658	0.000	0.071	0.082
Dummy_gest_hypertension	0.0120	0.002	6.865	0.000	0.009	0.015
Dummy_infertility_treatment	0.0069	0.003	2.130	0.033	0.001	0.013
Dummy_gonorrhea	0.0380	0.011	3.399	0.001	0.016	0.060
Dummy_syphilis	0.0181	0.014	1.299	0.194	-0.009	0.045
Dummy_chlamydia	0.0055	0.004	1.307	0.191	-0.003	0.014
Dummy_hepatitis_b	-0.0079	0.011	-0.722	0.470	-0.029	0.013
Dummy_hepatitis_c	0.0352	0.008	4.168	0.000	0.019	0.052
Dummy_labor_induction	-0.0059	0.001	-5.635	0.000	-0.008	-0.004
Dummy_labor_augmentation	0.0062	0.001	5.386	0.000	0.004	0.008
Dummy_antibiotics	0.0086	0.001	7.837	0.000	0.006	0.011
Dummy_chorioamnionitis	0.0049	0.004	1.294	0.196	-0.003	0.012
Dummy_anesthesia	0.0006	0.001	0.501	0.616	-0.002	0.003
Dummy_gender	0.0034	0.001	3.669	0.000	0.002	0.005

Omnibus:	135023.409	Durbin-Watson:	1.986
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3575424.287
Skew:	4.800	Prob(JB):	0.00
Kurtosis:	25.374	Cond. No.	1.23e+05

To check for multicollinearity among the independent variables, the variance inflation factor for each variable can be calculated (Massaron, 2016). The VIF measures how much the variance of the estimated regression coefficient is increased due to collinearity. A VIF greater than 5 or 10 indicates that the variable is highly collinear with other variables and may need to be removed from the model (Massaron, 2016). The initial model was reduced using the variance inflation factor of each independent variable greater than 5 (birth_year, mother_height, mother_bmi, pre_preg_weight, delivery_weight).

	Variable	VIF
0	birth_year	9376.091684
1	birth_month	1.001329
2	birth_time	1.013059
3	birth_place	1.066944
4	mother_age	2.618896
5	mother_education	2.137324
6	father_age	2.263369
7	father_education	2.020449
8	interval_llb	1.231404
9	cigarettes	1.033994
10	mother_height	17.902290
11	mother_bmi	129.978056
12	pre_preg_weight	149.661950
13	delivery_weight	8.342230
14	apgar5	2.114315
15	apgar10	2.056415
16	plurality	1.143712
17	infant_weight	1.336345
18	Dummy_marital_status	1.003670
19	Dummy_pre_preg_diabetes	1.022148
20	Dummy_gest_diabetes	1.045376
21	Dummy_pre_preg_hypertension	1.062457
22	Dummy_steroids	1.142068
23	Dummy_gest_hypertension	1.091472
24	Dummy_infertility_treatment	1.046770
25	Dummy_gonorrhea	1.032560
26	Dummy_syphilis	1.005379
27	Dummy_chlamydia	1.045415
28	Dummy_hepatitis_b	1.003209
29	Dummy_hepatitis_c	1.013530
30	Dummy_labor_induction	1.091011
31	Dummy_labor_augmentation	1.039322
32	Dummy_antibiotics	1.050256
33	Dummy_chorioamnionitis	1.030432
34	Dummy_anesthesia	1.118465
35	Dummy_gender	1.013411

Another variable selection procedure is the backward elimination method. This involves removing independent variables from the model one at a time based on their statical

significance until only statically significant variables remain (Massaron, 2016). This approach is justified since the goal is to identify the most important independent variables while minimizing the number of irrelevant variables. Also, those variables that had a p-values of greater than 0.05 were removed to get a final reduced multiple regression model. In the reduced model the R-squared value was 0.045. The F-statistic increased to 341.5 and the log-likelihood slightly increased to 45335. The AIC is -9.063e+04 and the BIC is -9.042e+04. The reduced model had 21 independent variables. The variables father_education, Dummy_martial_status, Dummy_infertility_treatment had p-value of less than 0.05. All the other independent variables in the model had p-values of 0.00. Therefore, all independent variables in the final multiple regression model were statically significant. These included: birthplace, mother age, mother education, father education, time since last pregnancy, cigarette use, steroids, plurality, infant weight, marital status, diabetes prior to pregnancy, gestation diabetes, hypertension prior to pregnancy, gestational hypertension, received infertility treatments, gonorrhea, hepatitis c, labor augmentation, received antibiotics, and infant gender.

OLS Regression Results

Dep. Variable:	Dummy_prev_preterm_birth	R-squared:	0.045
Model:	OLS	Adj. R-squared:	0.045
Method:	Least Squares	F-statistic:	341.5
Date:	Thu, 29 Jun 2023	Prob (F-statistic):	0.00
Time:	16:34:10	Log-Likelihood:	45335.
No. Observations:	144761	AIC:	-9.063e+04
Df Residuals:	144740	BIC:	-9.042e+04
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1540	0.006	27.402	0.000	0.143	0.165
birth_place	-0.0056	0.002	-3.600	0.000	-0.009	-0.003
mother_age	0.0006	9.71e-05	5.960	0.000	0.000	0.001
mother_education	-0.0020	0.000	-5.288	0.000	-0.003	-0.001
father_education	-0.0010	0.000	-2.656	0.008	-0.002	-0.000
interval_11b	-6.764e-05	1.23e-06	-55.036	0.000	-7.01e-05	-6.52e-05
cigarettes	0.0008	0.000	6.603	0.000	0.001	0.001
Dummy_steroids	0.0388	0.001	28.258	0.000	0.036	0.042
plurality	-0.0246	0.003	-8.423	0.000	-0.030	-0.019
infant_weight	-2.422e-05	8.84e-07	-27.404	0.000	-2.6e-05	-2.25e-05
Dummy_marital_status	0.0076	0.002	3.354	0.001	0.003	0.012
Dummy_pre_preg_diabetes	0.0392	0.005	8.374	0.000	0.030	0.048
Dummy_gest_diabetes	0.0136	0.002	7.825	0.000	0.010	0.017
Dummy_pre_preg_hypertension	0.0372	0.003	12.127	0.000	0.031	0.043
Dummy_gest_hypertension	0.0122	0.002	7.138	0.000	0.009	0.016
Dummy_infertility_treatment	0.0067	0.003	2.059	0.039	0.000	0.013
Dummy_steroids	0.0388	0.001	28.258	0.000	0.036	0.042
Dummy_gonorrhea	0.0410	0.011	3.718	0.000	0.019	0.063
Dummy_hepatitis_c	0.0357	0.008	4.225	0.000	0.019	0.052
Dummy_labor_augmentation	0.0064	0.001	5.651	0.000	0.004	0.009
Dummy_antibiotics	0.0091	0.001	8.417	0.000	0.007	0.011
Dummy_gender	0.0036	0.001	3.816	0.000	0.002	0.005

Omnibus:	135112.921	Durbin-Watson:	1.986
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3583306.760
Skew:	4.804	Prob(JB):	0.00
Kurtosis:	25.400	Cond. No.	5.63e+18

A metric to analyze the reduced regression model is the mean squared error (MSE). The MSE is a measure of the average squared difference between the predicted and actual values (Massaron, 2016). A lower MSE indicates a better model performance. The MSE of the multiple regression model is 0.0313. This indicates that on average the squared difference between the predicted and actual value is small. The lower MSE suggest the model is capable of capturing the patterns and trends in the data leading to more accurate predictions of preterm birth in relation to maternal characteristics. The model's root mean squared error (RMSE) is 0.177, representing on average the difference between the actual values and the predicted values. A lower RMSE indicates a small prediction error.

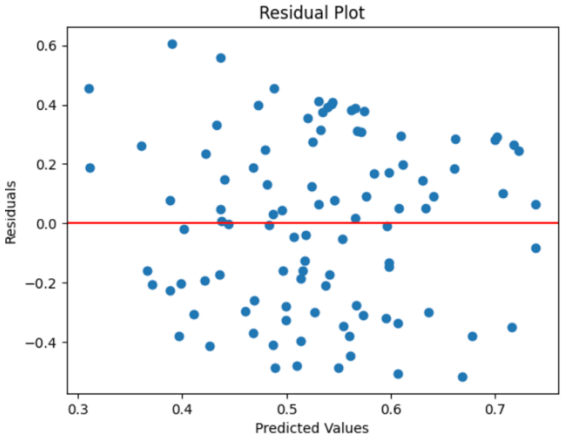
Standardized coefficients were calculated due to the measured variables being on different scales. Getting the standardized coefficient allows for comparison of the relative importance of variables in predicting the outcome variable, regardless of the unit of measurement. The standardized coefficients indicate variables with a high impact on the dependent variable, preterm birth. The variable, gonorrhea, had a standard coefficient of 0.969 indicating a

strong positive relationship with an increase in the likelihood of preterm birth. The variable, diabetes prior to pregnancy, had a standard coefficient of 0.392 indicating a moderately positive relationship associated with an increase in the likelihood of preterm birth. The variable hypertension prior to birth, had a standard coefficient of 0.242, indicating a positive relationship with an increased likelihood of preterm birth. The variable, steroids, had a standard coefficient of 0.215 which had a positive association with an increased likelihood of preterm birth. The variable hepatitis C had a coefficient of 0.645 suggesting a positive association with an increased in likelihood of preterm birth.

birth_place	-1.861906e-02
mother_age	1.029380e-04
mother_education	-1.176788e-03
father_education	-5.684292e-04
interval_llb	-1.645462e-07
cigarettes	2.020219e-04
Dummy_steroids	2.153224e-01
plurality	-1.459100e-01
infant_weight	-4.197016e-08
Dummy_marital_status	3.714188e-02
Dummy_pre_preg_diabetes	3.916578e-01
Dummy_gest_diabetes	5.035981e-02
Dummy_pre_preg_hypertension	2.422968e-01
Dummy_gest_hypertension	4.382765e-02
Dummy_infertility_treatment	4.573416e-02
Dummy_steroids	2.153224e-01
Dummy_gonorrhea	9.692908e-01
Dummy_hepatitis_c	6.445325e-01
Dummy_labor_augmentation	1.548160e-02
Dummy_antibiotics	2.109594e-02
Dummy_gender	7.142076e-03

Coefficients represent the estimated effects of the independent variables on the dependent variables. The coefficients for the multiple regression analysis were 0.14648006, 0.21760186, 0.19794293, -0.09386233, and 0.07351614. The intercept is the value of the dependent variable when all the independent variables are zero. This model has an intercept of 0.275.

A residual plot is a graphical representation of the errors. A well-fitting model will have residuals that are randomly distributed around zero, meaning the model is able to explain the variation in the dependent variable based on the independent variables (Massaron, 2016). Based on the residual plot below there is a random and evenly spread distribution of residual indicating the model is capturing underlying relationships between the independent variables and the dependent variable adequately.



Cross-validation was used to validate the performance and generalization ability of the model. The cross-validation scores for the multiple regression model are -0.01900525, -0.25505279, 0.2439426, 0.05838873, 0.09044186. Each score represents the performance

of the model on a specific fold of the cross-validation. The mean score of the model is 0.0237 which is the average of the cross-validation scores. A lower mean score suggests poor performance while a higher mean score, closer to 1, suggests better performance and the ability to generalize unseen data. Therefore, this model suggests a low performance in accurately predicting maternal characteristics associated with preterm birth.

Cook's distance is a measure used to assess the influence of individual data points on the fitted regression model by measuring the difference between the estimated regression coefficients with and without the inclusion of each observation (Massaron, 2016). A large cook's distance means the corresponding data points have a large influence on the regression coefficients and may be an outlier driving the data. A cook's distance of greater than 1 may indicate a significant impact on the model's results. Some values obtained from completing Cook's distance on the multiple regression suggest that some observations have a larger impact on the regression model than others and may need to be investigated further to determine whether they are outliers or have some other issue that needs to be addressed.

```
Cook's distance: [2.97695452e-02 3.99320708e-03 7.93508528e-03 2.84644945e-02
 8.60072725e-03 2.96866120e-03 4.78134184e-02 3.42630148e-04
 2.22439881e-05 8.84949733e-04 4.34236969e-03 2.39283637e-03
 1.50957890e-02 6.71871509e-03 1.37669132e-02 7.14973517e-03
 2.63094513e-03 2.94405695e-03 1.15083961e-02 3.51765905e-02
 1.62854931e-02 6.92108473e-02 2.71041950e-02 2.18158228e-03
 5.08997771e-03 2.99762952e-03 2.11416492e-02 3.42432964e-03
 8.84718389e-04 2.34154159e-03 1.87639113e-02 2.34851268e-02
 8.01133395e-03 4.29153415e-03 3.43950461e-06 1.72935703e-02
 2.44381168e-03 1.31155743e-02 1.03027209e-02 8.11618853e-03
 9.17412262e-03 4.66043213e-04 2.25045926e-02 1.74460951e-02
 1.57765385e-02 1.30005295e-02 4.95109754e-03 1.29085800e-03
 4.27729580e-04 8.76955154e-04 3.70157457e-03 1.36235901e-04
 1.76595330e-02 1.12768535e-03 3.45972082e-03 1.21986973e-02
 6.24713634e-05 3.06769881e-03 1.93479211e-04 2.94231987e-02
 1.55549739e-03 1.27723360e-02 2.81960716e-03 1.61728744e-02
 1.47545791e-02 1.61112301e-02 1.05931708e-02 2.18205874e-03
 8.48823501e-06 1.03047424e-02 3.00267849e-04 1.28835873e-02
 2.43041681e-02 4.20502336e-03 2.20500229e-02 1.35061555e-02
 1.25681748e-02 5.49127350e-03 1.72951828e-02 4.83257623e-02
 3.60022771e-03 1.21862655e-03 1.12044385e-02 3.15152116e-03
 1.99172270e-03 1.04385566e-02 2.95670723e-02 8.22541843e-03
 6.88282148e-03 7.73050041e-03 1.73852192e-02 3.49878557e-03
 7.07089080e-03 1.88237180e-02 3.87249263e-03 3.87668341e-04
 2.05108024e-02 7.18610234e-04 3.09905385e-03 2.45438035e-03]
```

An analysis of variance (ANOVA) test can be performed before running a multiple regression model to obtain information about the relationship between the dependent and independent variables when the relationship is not linear. This can help identify which variables are most likely to be useful in predicting the response variable. An ANOVA test can be used to test whether the assumptions of the regression model are met (Bruce, 2020). Variables such as mother_age, mother_education, interval_llb, cigarettes, Dummy_steroids, infant_weight, Dummy_pre_preg_diabetes, Dummy_gest_diabetes, Dummy_pre_preg_hypertension, Dummy_gest_hypertension, Dummy_antibiotics, and Dummy_gender had p-values (associated with the F-statistic) close to 0, indicating they have a significant effect on preterm births. Variables such as birth_place, father_education, Dummy_marital_status, Dummy_infertility_treatment, Dummy_gonorrhea, Dummy_hepatitis_c, Dummy_labor_induction, and Dummy_labor_augmentation had p-values less than 0.05, indicating their significance in the model. The residual has a large degree of freedom (144739.0) and a small mean square value (0.0313) indicating there is a large portion of the variation in the dependent variable is unexplained.

	df	sum_sq	mean_sq	F	PR(>F)
birth_place	1.0	0.588074	0.588074	18.791007	1.459515e-05
mother_age	1.0	10.674423	10.674423	341.084942	4.531415e-76
mother_education	1.0	22.250158	22.250158	710.969919	2.946131e-156
father_education	1.0	1.457012	1.457012	46.556615	8.935732e-12
interval_llb	1.0	82.822204	82.822204	2646.457406	0.000000e+00
cigarettes	1.0	2.579937	2.579937	82.437964	1.103575e-19
Dummy_steroids	1.0	51.100769	51.100769	1632.847243	0.000000e+00
plurality	1.0	0.112827	0.112827	3.605224	5.760031e-02
infant_weight	1.0	25.271914	25.271914	807.525518	3.840992e-177
Dummy_marital_status	1.0	0.381878	0.381878	12.202337	4.774393e-04
Dummy_pre_preg_diabetes	1.0	2.969741	2.969741	94.893537	2.041191e-22
Dummy_gest_diabetes	1.0	2.502110	2.502110	79.951114	3.881545e-19
Dummy_pre_preg_hypertension	1.0	4.349997	4.349997	138.997535	4.561695e-32
Dummy_gest_hypertension	1.0	1.726736	1.726736	55.175209	1.108526e-13
Dummy_infertility_treatment	1.0	0.145912	0.145912	4.662398	3.083175e-02
Dummy_gonorrhea	1.0	0.482277	0.482277	15.410413	8.655027e-05
Dummy_hepatitis_c	1.0	0.582525	0.582525	18.613692	1.601727e-05
Dummy_labor_induction	1.0	1.078706	1.078706	34.468400	4.341688e-09
Dummy_labor_augmentation	1.0	0.958328	0.958328	30.621899	3.140757e-08
Dummy_antibiotics	1.0	2.242742	2.242742	71.663391	2.575607e-17
Dummy_gender	1.0	0.437454	0.437454	13.978171	1.850166e-04
Residual	144739.0	4529.679177	0.031295	NaN	NaN

Logistic Regression Analysis:

A logistic regression is a statistical method used to analyze the relationship between a binary dependent variable and the independent variable(s). The initial logistic regression model contained 35 independent variables to be analyzed against the dependent variable, preterm birth. The number of observations used in the analysis was 144,761. The pseudo R-squared value was 0.1757 which indicates the goodness of fit of the model. The log-likelihood value was -17,676 which is a measure of how well the model fit the current data. Higher values indicate a better fit meaning the model is more likely to generate the observed data. The model indicates convergence. The LL-Null value of -21,444 represents the log-likelihood of a null model with no predictors. A lower value indicates a better fit. The difference between the log-likelihood value and the LL-Null value can be used to assess the improvement of the model compared to the null model. The model has a LLR p-value of 0.000 suggesting the model has a significant impact on the dependent variable. Many variables had p-values greater than 0.05 and were removed from the initial model. The variables father_education, Dummy_martial_status, Dummy_infertility_treatment, dummy_honorrhea, Dummy_hepatitis, and Dummy_gender had a p-value of less than 0.05. All the other independent variables have p-values of exactly 0.00. All these variables have very small p-values likely indicating their significance.

Akaike Information Criterion (AIC) is a model evaluation metric used to measure the relative quality of a statistical model while taking into account both the goodness of fit and the complexity of the model (Massaron, 2016).

A lower AIC value indicates a better model fit with a larger difference in AIC values indicating a greater difference in model quality. This can be used to compare models to select the best model. The AIC for the initial model was 35424.834 and for the reduced model the AIC was 35426.638. There was a minimal change (difference of 1.804) in the AIC from the initial model to the reduced model. Therefore, it is possible that both models provide a similar fit for the data.

Dep. Variable:	Dummy_prev_preterm_birth	**No. Observations:**	144761
Model:	Logit	**Df Residuals:**	144725
Method:	MLE	**Df Model:**	35
Date:	Thu, 29 Jun 2023	**Pseudo R-squ.:**	0.1757
Time:	18:03:07	**Log-Likelihood:**	-17676.
converged:	True	**LL-Null:**	-21444.
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
birth_year	-0.0001	0.001	-0.179	0.858	-0.001	0.001
birth_month	-0.0077	0.004	-1.738	0.082	-0.016	0.001
birth_time	-4.13e-05	2.49e-05	-1.658	0.097	-9.01e-05	7.51e-06
birth_place	-0.2703	0.074	-3.631	0.000	-0.416	-0.124
mother_age	0.0276	0.004	6.797	0.000	0.020	0.036
mother_education	-0.0709	0.012	-5.692	0.000	-0.095	-0.046
father_age	-0.0007	0.003	-0.212	0.832	-0.007	0.006
father_education	-0.0356	0.012	-2.867	0.004	-0.060	-0.011
interval_llb	-0.0067	0.000	-25.071	0.000	-0.007	-0.006
cigarettes	0.0157	0.003	5.677	0.000	0.010	0.021
mother_height	0.0051	0.022	0.235	0.814	-0.037	0.047
mother_bmi	0.0132	0.024	0.562	0.574	-0.033	0.059
pre_preg_weight	-0.0061	0.004	-1.472	0.141	-0.014	0.002
delivery_weight	0.0054	0.001	5.235	0.000	0.003	0.007
Dummy_steroids	0.9337	0.055	16.861	0.000	0.825	1.042
apgar5	-0.0330	0.027	-1.233	0.218	-0.085	0.019
apgar10	0.0033	0.002	1.402	0.161	-0.001	0.008
plurality	-0.9401	0.073	-12.811	0.000	-1.084	-0.796
infant_weight	-0.0007	2.83e-05	-26.020	0.000	-0.001	-0.001
Dummy_marital_status	0.2178	0.067	3.248	0.001	0.086	0.349
Dummy_pre_preg_diabetes	0.7234	0.104	6.977	0.000	0.520	0.927
Dummy_gest_diabetes	0.3698	0.048	7.726	0.000	0.276	0.464
Dummy_pre_preg_hypertension	0.6459	0.071	9.146	0.000	0.508	0.784
Dummy_gest_hypertension	0.3881	0.050	7.757	0.000	0.290	0.486
Dummy_infertility_treatment	0.3561	0.106	3.344	0.001	0.147	0.565
Dummy_gonorrhea	0.6012	0.252	2.387	0.017	0.108	1.095
Dummy_syphilis	0.2860	0.339	0.843	0.399	-0.379	0.951
Dummy_chlamydia	0.1855	0.129	1.440	0.150	-0.067	0.438
Dummy_hepatitis_b	-0.2717	0.355	-0.766	0.444	-0.967	0.423
Dummy_hepatitis_c	0.5439	0.175	3.099	0.002	0.200	0.888
Dummy_labor_induction	-0.1677	0.036	-4.651	0.000	-0.238	-0.097
Dummy_labor_augmentation	0.2306	0.038	6.091	0.000	0.156	0.305
Dummy_antibiotics	0.2568	0.033	7.675	0.000	0.191	0.322
Dummy_chorioamnionitis	0.1435	0.175	0.818	0.413	-0.200	0.487
Dummy_anesthesia	0.0028	0.036	0.078	0.938	-0.067	0.073
Dummy_gender	0.0976	0.030	3.223	0.001	0.038	0.157

Logit Regression Results

Dep. Variable:	Dummy_prev_preterm_birth	**No. Observations:**	144761
Model:	Logit	**Df Residuals:**	144739
Method:	MLE	**Df Model:**	21
Date:	Thu, 29 Jun 2023	**Pseudo R-squ.:**	0.1750
Time:	18:03:07	**Log-Likelihood:**	-17690.
converged:	True	**LL-Null:**	-21444.
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
birth_place	-0.2734	0.066	-4.113	0.000	-0.404	-0.143
mother_age	0.0266	0.003	9.709	0.000	0.021	0.032
mother_education	-0.0709	0.012	-5.736	0.000	-0.095	-0.047
father_education	-0.0346	0.012	-2.815	0.005	-0.059	-0.011
interval_llb	-0.0066	0.000	-25.163	0.000	-0.007	-0.006
cigarettes	0.0160	0.003	5.814	0.000	0.011	0.021
delivery_weight	0.0016	0.000	4.426	0.000	0.001	0.002
Dummy_steroids	0.9382	0.055	17.146	0.000	0.831	1.045
plurality	-0.8971	0.062	-14.381	0.000	-1.019	-0.775
infant_weight	-0.0007	2.33e-05	-30.974	0.000	-0.001	-0.001
Dummy_marital_status	0.2167	0.067	3.233	0.001	0.085	0.348
Dummy_pre_preg_diabetes	0.7104	0.103	6.881	0.000	0.508	0.913
Dummy_gest_diabetes	0.3546	0.047	7.491	0.000	0.262	0.447
Dummy_pre_preg_hypertension	0.6433	0.070	9.133	0.000	0.505	0.781
Dummy_gest_hypertension	0.3961	0.050	7.948	0.000	0.298	0.494
Dummy_infertility_treatment	0.3576	0.105	3.391	0.001	0.151	0.564
Dummy_gonorrhea	0.6915	0.243	2.843	0.004	0.215	1.168
Dummy_hepatitis_c	0.5584	0.175	3.194	0.001	0.216	0.901
Dummy_labor_induction	-0.1735	0.036	-4.867	0.000	-0.243	-0.104
Dummy_labor_augmentation	0.2260	0.038	6.016	0.000	0.152	0.300
Dummy_antibiotics	0.2605	0.033	7.881	0.000	0.196	0.325
Dummy_gender	0.0994	0.030	3.295	0.001	0.040	0.159

The statistical significance can be determined from the logistic regression coefficients and p-values which are expressed in terms of odds ratios. Each of the variable p-values were less than 0.05 suggesting statical significance. The coefficients represent the change in the log-odds of the dependent variable per unit increase of each of the independent variables while holding all other variables constant. The coefficients can be used to make predictions on new data using the same model. The intercept coefficient is 0.0233 and is the log-odds when all categorical variables are at their references levels and all continuous variables are at zero. The larger the absolute value of a coefficient, the grater its impact on the dependent variable. A positive coefficient indicate there is an increase in the independent variable and is associated with an increase in the log-odd of the dependent variable while a negative coefficient indicates the opposite. Variables with relatively large coefficients that had the highest impact based on their coefficient were:

- Dummy_steroids: coefficient= 0.9382, a one unit increase in steroid use is associated with a significant increase int he log-odds of preterm birth.
- Dummy_pre_preg_diabets: coefficient= 0.7104, a one unit increase in diabetes prior to pregnancy is associated with an increase in the log-odds of preterm. birth.
- Dummy_gonorrhea: coefficient= 0.6915, a one unit increase in gonorrhea is associated with an increase in the log-odds of preterm birth.

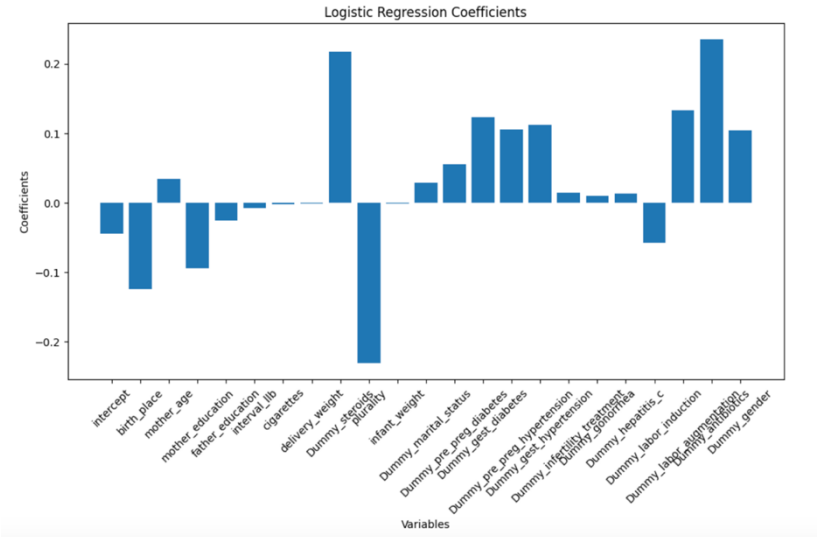
The cross-validation scores of the logistic model were -0.80656088, -0.08675117, 0.00616065, -0.06012151, 0.0689787. Larger positive values (closer to 1) indicate better performance, while negative values suggest poor performance. The mean score for the model was -0.176 which suggest the model is demonstrates poor performance.

A confusion matrix is a table that is used to evaluate the performance of a classification model. It provides a summary of the predictions made by the model compared to the actual class labels of instances. The confusion matrix provides the following information: the model correctly predicted 27,946 instances as true positive, the model correctly predicted 1 instance as true negative, the model incorrectly predicted 1,007 instances as false positive, and the model incorrectly predicted 0 instances as a false negative.

```
Confusion Matrix:
[[27945     1]
 [ 1007     0]]
```

The intercept of the logistic regression model was -3.471 which means that when all the independent variables are zero, the log-odds of the positive class (1) compared to the negative class (0) is -3.471. The coefficients are: -0.0440814,-0.12427593, 0.03419068, -0.09472079, -0.02564572, -0.00798864, -0.00263918, -0.00114053, 0.21780572, -0.23094571, -0.00077642, 0.02831237, 0.05589069, 0.12312164, 0.10590176, 0.11250995, 0.01430617, 0.00994238, 0.01307448, -0.05779718, 0.1335582, 0.23508944, 0.10402321. These values represent the change in the log-odds of the positive class compared to the negative class for each independent variable while holding the others constant. As the independent variable increases, the log-odds of the positive class increase, indicating a higher probability of belonging to the positive class. With negative coefficients, as the independent variable increases, the log-odds of the positive class decrease, indicating a lower probability of belonging to the positive class.

The above coefficients can be plotted on a graph to identify the variables that have the most significant impact on the prediction of preterm birth. Positive coefficients suggest there is an increase in that variable's value in association with a higher likelihood of preterm birth. Positive variables included: mother_age, 'Dummy_steroids', 'Dummy_marital_status', 'Dummy_pre_preg_diabetes', 'Dummy_gest_diabetes', 'Dummy_pre_preg_hypertension', 'Dummy_gest_hypertension', 'Dummy_infertility_treatment','Dummy_gonorrhea', 'Dummy_hepatitis_c', 'Dummy_labor_augmentation', 'Dummy_antibiotics', 'Dummy_gender'. The height of each bar represents the magnitude of the coefficient for the corresponding variable. Thus, the larger the bar the stronger the effect of the variable. The variables with the highest values are Dummy_antibiotics (0.2351and Dummy_steroids (0.2178). All other variables (infant_weight, delievery_weight, cigarettes, interval_llb, father_education, dummy_labor_induction, mother_education, birth_place, plurality) had negative coefficients. If a coefficient is negative, then as the variable increases in value there is an association with a lower likelihood of preterm birth.




```
intercept: -0.0440814
birth_place: -0.12427593
mother_age: 0.03419068
mother_education: -0.09472079
father_education: -0.02564572
interval_llb: -0.00798864
cigarettes: -0.00263918
delivery_weight: -0.00114053
Dummy_steroids: 0.21780572
plurality: -0.23094571
infant_weight: -0.00077642
Dummy_marital_status: 0.02831237
Dummy_pre_preg_diabetes: 0.05589069
Dummy_gest_diabetes: 0.12312164
Dummy_pre_preg_hypertension: 0.10590176
Dummy_gest_hypertension: 0.11250995
Dummy_infertility_treatment: 0.01430617
Dummy_gonorrhea: 0.00994238
Dummy_hepatitis_c: 0.01307448
Dummy_labor_induction: -0.05779718
Dummy_labor_augmentation: 0.1335582
Dummy_antibiotics: 0.23508944
Dummy_gender: 0.10402321
```

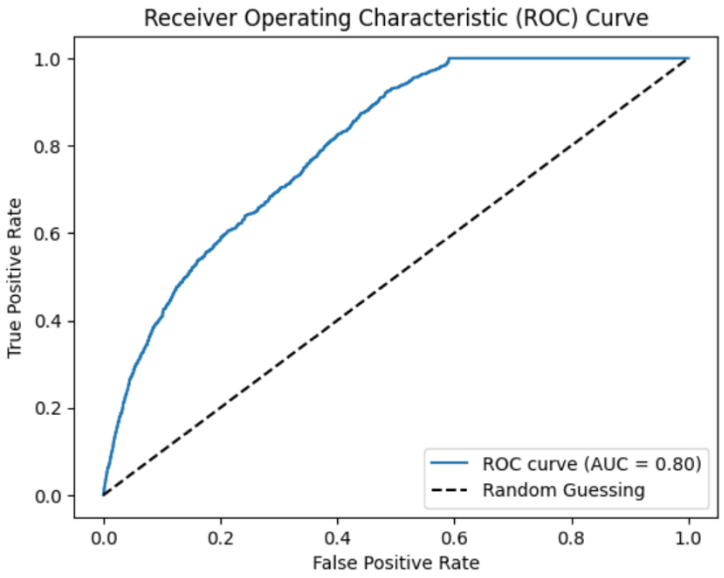
Variable importance was calculated using the premutation importance method which provides a measure of the importance of each feature in predicting the dependent variable, preterm birth. Higher values indicate greater importance. A weight of zero indicates the independent variable has no importance in predicting the dependent variable. Based on the values the variables with non-zero weights (indicating some importance in predicting the likelihood of preterm birth) were: delivery weight, infant gender, labor augmentation, mother's age, gestational diabetes, use of antibiotics, gestational hypertension, and hypertension prior to pregnancy. These independent variables had the highest impact on predicting the likelihood of preterm birth, while the other variables had zero importance in the prediction.

Weight	Feature
0.0000 ± 0.0000	delivery_weight
0.0000 ± 0.0000	Dummy_gender
0.0000 ± 0.0000	Dummy_labor_augmentation
0.0000 ± 0.0001	mother_age
0.0000 ± 0.0000	Dummy_gest_diabetes
0.0000 ± 0.0000	Dummy_antibiotics
0.0000 ± 0.0000	Dummy_gest_hypertension
0.0000 ± 0.0000	Dummy_pre_preg_hypertension
0 ± 0.0000	Dummy_steroids
0 ± 0.0000	Dummy_pre_preg_diabetes
0 ± 0.0000	intercept
0 ± 0.0000	Dummy_labor_induction
0 ± 0.0000	Dummy_marital_status
0 ± 0.0000	Dummy_gonorrhea
0 ± 0.0000	Dummy_hepatitis_c
0 ± 0.0000	Dummy_infertility_treatment
-0.0000 ± 0.0000	cigarettes
-0.0000 ± 0.0000	birth_place
-0.0000 ± 0.0000	plurality
-0.0000 ± 0.0000	interval_llb
	... 3 more ...

A model evaluation metric used to evaluate the performance of binary classifiers is precision, recall, and F1-scores which are found in the classification report. These metrics can be used to assess the model's ability to correctly classify instances of each class. The precision of term birth (class 0) is 0.97 which means that out of all instances predicted as term births, 97% were classified correctly. The precision of preterm birth (class 1) is 0.14 meaning only 14% of the instances predicted as preterm births were predicted correctly. Thus, the model had a high precision in predicting term births but not preterm births. This is consistent with a recall of 1.00 for class 0 and 0.00 for class 1. The F1-score is the harmonic mean of the precision and recall. The F1-score of class 0 was 0.98 which indicates a good balance between the precision and recall for term births. However, the F1-score for class 1 is 0.00 suggesting poor performance in predicting instances of preterm birth. The report also indicates there were 27,946 instances of term births and 1,007 instances of preterm births. The accuracy of the model is 0.97 which demonstrates the model correctly predicts the class for 97% of the instances.

	precision	recall	f1-score	support
0	0.97	1.00	0.98	27946
1	0.14	0.00	0.00	1007
accuracy			0.97	28953
macro avg	0.55	0.50	0.49	28953
weighted avg	0.94	0.97	0.95	28953

The receiver operating characteristic (ROC) curve is a graphical representation of the performance of a binary classifier as the discrimination threshold varies (Narkhede, 2022). The closer the ROC curve is to the upper left corner of the plot, the better the classifier's performance. The area under the ROC curve (AUC-ROC) is used to evaluate the overall performance of the classifier. The AUC-ROC for the model is 0.80. The closer the score is to 1.0 the higher the classifier performance. The AUC-ROC of 0.08 suggests the model has relatively good discrimination ability in ditgi9nushing between the two classes.



A model comparison between the logistic model and a random forest model was completed to assess the predictive accuracy of the logistic model. As previously discussed, the logistic model correctly predicted 97% of the instances but was only able to correctly identify one instance of preterm birth. The random forest model's accuracy was also high at 0.9652, indicating it correctly predicted 97% of the cases. The random forest model had a precision of 0.375, which correctly classified preterm births 38% of the time. This is an improvement compared to the logistic regression model. The recall of the random forest model is 0.00298, meaning the model was able to identify only 0.298% of the actual preterm birth cases. The F1-score was also very low at 0.00591 representing a balance between the precision and recall. The random forest model performed better than the logistic regression model in terms of precision for the positive class. However, the recall and F1-scores were similar for both models suggesting both models may not be effectively capturing the cases of preterm birth.

Regulation techniques are used to prevent overfitting and improve generalization performance. Overfitting occurs when the model learns to fit the training too closely. Regulation techniques introduce a penalty term to discourage them from taking large values and help control the complexity of the model and reduce the influence of noise or irrelevant features (Brownlee, 2019a). The accuracy score which measures the overall correctness of the model's predictions, is all the same; indicating the models performed similarly in terms of overall prediction correctness and were effective in preventing overfitting and improving the generalizing of the models.

```
L2 Regularization Accuracy: 0.9644251027527372
L1 Regularization Accuracy: 0.9644251027527372
Elastic Net Regularization Accuracy: 0.9644251027527372
```

D3. Justification of Analysis Technique

Multiple regression analysis with ANOVA allows for hypothesis testing in determining the statical significance of individual maternal characteristics as predictors of preterm birth. By examining the p-values associated with each predictor variable, the individual contributions can be assessed in identifying the significant predictors. An advantage of multiple regression analysis is it provides a quantitative assessment of the relationship between variables, allowing for the estimation of the strength and direction of the relationships. A disadvantage

of multiple regression analysis is overfitting when there are too many independent variables in the model without sufficient justification.

An ANOVA is used in conjunction with multiple regression analysis to examine the relationship between continuous variables to assess if there are significant differences in mean values between different characteristics (*What Is Analysis of Variance (ANOVA)?*, n.d.). A logistic regression model will be used to produce coefficients for each independent variable to indicate the direction and magnitude of their influence on the probability of the dependent variable.

A logistic regression analysis is appropriate for this analysis because the dependent variable is binary/categorical and will provide the probability of preterm birth associated with the dependent variable. Logistic regression analysis allows for the assessment of the likelihood of a particular outcome based on the independent variable and provides a quantitative measure of the association. A disadvantage of a logistic regression assumes a linear relationship between the log-odds of the dependent and the independent variables.

Random forest modeling is known for its high prediction accuracy and robustness. A random forest model was used as a comparison to the logistic regression model to examine the model's performance. A random forest can also handle complex relationships and interactions between variables effectively. Both models' performance can be assessed and provide insights into the logistics regression model's ability to capture relationships in the data.

Part V: Data Summary and Implications

E1. Results

This analysis aimed to investigate maternal characteristics associated with preterm birth using multiple regression and logistic regression modeling. The analysis indicates several variables were found to be significant in the prediction of preterm birth (indicated by p-values <0.05). These maternal characteristics included: birthplace, mother age, mother education, father education, the interval between the last live birth, cigarettes use, delivery weight, use of steroids, plurality (multiple births), infant weight, material status, pre-pregnancy diabetes, gestational diabetes, pre-pregnancy hypertension, gestational hypertension, infertility treatment, gonorrhea, hepatitis C, labor induction, labor augmentation, antibiotics usage, and gender. These findings provide insights into the factors that may contribute to preterm birth.

The null hypothesis stated that there is no significant association between individual maternal characteristics and the likelihood of preterm birth. The alternative hypotheses proposed there is a significant relationship between individual maternal characteristics and the likelihood of preterm birth. Based on the p-values and coefficients found from the multiple regression model and logistic regression model there was statistical significance (p-values less than 0.05) in maternal characteristics studied and the likelihood of preterm birth. Therefore, the null hypothesis can be rejected.

The logistic regression coefficients indicate the direction and magnitude of the association between each independent variable and the log-odds of preterm birth. The positive coefficients were: steroid use, marital status, diabetes prior to pregnancy, hypertension prior to pregnancy, gestational hypertension, infertility treatments, gonorrhea, hepatitis C, labor augmentation, antibiotic use, and gender. The negative coefficients were birthplace, mother's education, father's education, the interval between the last live birth, cigarette use, delivery weight, plurality, infant weight, and labor induction. Positive coefficients

indicate there is an increase in the respective independent variable associated with an increase in the likelihood of preterm birth, while negative coefficients suggest a decrease in the likelihood of preterm birth.

Variable importance suggested the independent variables in this analysis had very small weights and their variation had a negligible impact on predicting preterm birth. This means the independent variables used in this analysis had little to no impact on predicting preterm birth.

Multiple regression analysis indicated an overall good model as a prediction of preterm birth. The multiple regression model provided a MSE value of 0.313. Suggesting the predicted values deviated from the actual values by a relatively small amount. The residual squared error value of 0.177 suggests there is some remaining amount of variation in the data that is not explained by the multiple regression model.

The logistic regression model struggled to correctly identify instances of preterm birth (positive class), indicated by a low recall, precision, and F1-score. However, the model performed well in predicting the term birth (negative class). The confusion matrix also provided two classes: preterm birth (positive class) and term birth (negative class). Thus, the logistic model does not identify any instances of preterm birth, only term births. The estimated probability of preterm birth for this given set of predictor values is 0.56%. This could mean the selected maternal characteristics might not be strong predictors of preterm birth. This could also mean the model needs to be improved to better capture the relationship between maternal characteristics and preterm birth.

These findings provide valuable insights into the association between maternal characteristics and preterm birth, which can guide future interventions and efforts to reduce preterm birthing incidents. Overall, this study emphasized the complex nature of predicting preterm birth and the need for further research and model improvement to better understand and capture the maternal characteristics contributing to premature births.

E2. Limitations

The regression models used in the analysis relied on assumptions such as linearity, independence, and absence of multicollinearity. If these are not present it may affect the validity of the results. A larger data set could provide more data and a better-fit model. The limited number of independent variables used in the analysis may not fully capture all the factors that influence preterm birth. It is possible that other maternal factors not included in the analysis may be responsible for preterm birth. There were many variables showing significance in preterm birth. However, variable selection did not fully capture the complexity of factors that contribute to preterm birth. This analysis is based on retrospective data which may not establish a clear temporal relationship between the independent variables and preterm birth. This study focused solely on preterm birth as an independent variable which may limit the generalizability of the findings. Preterm birth can have various causes and risk factors. Not all women who have had a previous preterm birth will have the same underlying factors contributing to subsequent preterm births.

E3. Recommended Course of Action

The models did well at predicting term births but not preterm births in relation to maternal characteristics. It is recommended that model be expanded with more independent variables. This analysis is limited in the number and variety of maternal characteristics used in the regression models. It is recommended there be external validation. Validating the findings and model on another dataset that is similar should be done to assess the generalizability of

the results. Other possible factors influencing preterm birth that could include race, substance abuse, inadequate prenatal care, healthcare resource availability, or hormonal imbalances.

E4. Approach for Future Study

Further investigation should be conducted to determine maternal characteristics influencing preterm birth to understand specific factors and characteristics that contribute to preterm birth to provide valuable insights on prevention and intervention strategies. One future approach for further studying would be to complete a longitudinal analysis. A longitudinal analysis can provide insight into patterns and changes in risk factors associated with preterm birth and offer multiple time points during pregnancy and during postpartum care for a dynamic assessment and investigation into maternal characteristics. Another approach for future study would be to look further into causal inferences. Using casual inference methods can establish a causal relationship between specific risk factors and preterm birth.

Part VI: Sources

Brittain, Jim; Cendon, Mariana; Nizzi, Jennifer; and Pleis, John (2018) "Data Scientist's Analysis Toolbox: Comparison of Python, R, and SAS Performance," *SMU Data Science Review*: Vol. 1: No. 2, Article 7. https://scholar.smu.edu/datasciencereview/vol1/iss2/7

Brownlee, J. (2019a). Use Weight Regularization to Reduce Overfitting of Deep Learning Models. *MachineLearningMastery.com*. https://machinelearningmastery.com/weight-regularization-to-reduce-overfitting-of-deep-learning-models/

Bruce, P., Bruce, A., & Gedeck, P. (2020a). Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python (2nd ed.). O'Reilly Media.

Collins, L. M., Schafer, J. A., & Kam, C. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330–351. https://doi.org/10.1037/1082-989x.6.4.330

Data Access - Vital Statistics Online. (n.d.). https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm#Tools

Daniel T. Larose, & Chantal D. Larose. (2019). Data Science Using Python and R. Wiley.

Edgar, T. F., & Manz, D. O. (2017). Exploratory Study. In Elsevier eBooks (pp. 95–130). https://doi.org/10.1016/b978-0-12-805349-2.00004-2

Fuchs, F., Monet, B., Ducruet, T., Chaillet, N., & Audibert, F. (2018). Effect of maternal age on the risk of preterm birth: A large cohort study. PLOS ONE, 13(1), e0191002. https://doi.org/10.1371/journal.pone.0191002

GeeksforGeeks. (2022). Violin plot using Seaborn in Python. *GeeksforGeeks*. https://www.geeksforgeeks.org/violinplot-using-seaborn-in-python/

Luna, J. C. (2022, December 28). *Python vs R for Data Science: Which Should You Learn?* https://www.datacamp.com/blog/python-vs-r-for-data-science-whats-the-difference

Massaron, L. (2016). *Regression analysis with python: Learn the art of regression analysis with python*. Packt Publishing.

Multiple Regression. (n.d.). https://home.csulb.edu/~msaintg/ppa696/696regmx.htm

Narkhede, S. (2022, March 5). Understanding AUC - ROC Curve - Towards Data Science. *Medium*. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

National Vital Statistics System (NVSS) - Health, United States. (n.d.). https://www.cdc.gov/nchs/hus/sources-definitions/nvss.htm

Premature Birth. (2022, November 1). Centers for Disease Control and Prevention. https://www.cdc.gov/reproductivehealth/features/premature-birth/index.html#:~:text=Some%20risk%20factors%20for%20preterm,has%20to%20be%20delivered%20early

Turvey, B. E. (2013). Multivariate Analysis of Forensic Fraud, 2000–2010. In Elsevier eBooks (pp. 157–182). https://doi.org/10.1016/b978-0-12-408073-7.00009-4

USA Natality 2020. (2022, April 20). Kaggle. https://www.kaggle.com/datasets/shayta/usa-natality-2020

What is Analysis of Variance (ANOVA)? (n.d.). TIBCO Software. https://www.tibco.com/reference-center/what-is-analysis-of-variance-anova#:~:text=Sign%20In-,What%20is%20Analysis%20of%20Variance%20(ANOVA)%3F,the%20means%20of%20different%20groups