Udacity, Nanodegree: Data Scientist

Capstone Project:

Maternal Health - Maternal Sepsis by Select Risk Factors

By Laura Hagg, 07.06.2023

Table of contents

1. The Dataset

The dataset contains information on the occurrence of maternal sepsis among live births during the pregnancy, delivery, and postpartum periods from 2016 to 2018. It provides counts, rates, and measures of association related to specific risk factors and the incidence of maternal sepsis, identified through administrative means.

The dataset is available on the HealthData.gov website. You can access the dataset using the following link: Link to the dataset.

Additional information about this dataset can be found here. The provided link includes details about the columns and their data types.

The dataset consists of 16 columns and 585 rows.

1.1 Columns in the Dataset

In this part I will discuss the important columns of the dataset. Additionally, a data dictionary with column descriptions is available in the repository under the filename "HDNY_MaternalSepsis_RiskFactors_DataDictionary_v2.pdf".

Maternal Window:

Here are the 3 stages of maternity:

- Pregnancy
- Delivery
- Postpartum (= 42 days after the delivery)

For every maternal window, data on the risk factor had been collected.

Risk Factor, Risk Factor Type:

The dataframe contains 5 different risk factor types: 'Bateman Comorbidities', 'Elixhauser Comorbidities', 'Other Comorbidities', 'Demographics', and 'Obstetric'.

For my exploration, I did not select risk factors based on these types individually, but rather grouped them together:

For the first set of questions (3.1 – 3.4), I grouped all comorbidities together and investigated the columns where Risk Factor Strata = 'yes'. This means focusing on women who had these types of comorbidities.

For the last question, I took the risk factor from the 'Demographics' type and combined it with the risk factor 'Trimester Beginning Prenatal Care'. Since this risk factor is not divided into 'yes' and 'no', I examined it in question 3.5.

Any Sepsis/Severe Sepsis p-value:

This column shows the p-value for each risk factor, comparing it to the reference group of women who do not have the risk factor (indicated by risk factor strata = 'no'). It's important to note that the presence of missing values in this column is not indicative of data cleaning issues, as the NaN values serve as references for the comparison.

In the upcoming questions, I will investigate whether there are statistically significant risk factors associated with the occurrence of any or severe sepsis. I have chosen a significance threshold of p-value = 0.05 for determining significance.

Live Births and Incidence:

These columns represent the number or percentage of live births for women with or without the specific risk factor. The incidence of sepsis is presented in two separate columns: the number of incidences (for any or severe sepsis) and the incidence per 100,000 live births. This separation will be particularly relevant in question 3.4.

2. Data Cleaning

The dataset is generally clean, but there were a few tasks I performed before working with the data.

2.1 Changing the datatype of the p-value to a float

The issue arises from the columns 'Any_Sepsis_p-value' and 'Severe_Sepsis_p-value', where non-floating point numbers such as '<0.0001' are present. To facilitate better analysis, I converted these values to '0.0001'.

2.2 Removing duplicated information regarding Tobacco and Alcohol use

Due to data collection from different sources ('SPARCS', 'Birth certificate', 'SPARCS & Birth certificate'), redundant information regarding tobacco and alcohol use exists. To ensure data integrity in our analysis, I decided to exclude the risk factors collected solely from the birth certificate, thus eliminating duplicated information.

In the dataframe, there are two risk factors that appear twice:

1. Tobacco use (& Tobacco Use)
2. Alcohol use (& Alcohol abuse)

The reason for this duplication is that the data is collected from two different sources: SPARCS and Birth Certificate.

To address this issue, I have three options:

1. Keep the data as it is.
2. Combine the data together.
3. Remove one of the risk factors.

After careful consideration, I have decided to remove one of these risk factors to avoid duplication in the data analysis. Specifically, I will remove the risk factors obtained from the 'Birth Certificate' data

source. This decision is based on the fact that the 'Tobacco use' and 'Alcohol use' risk factors from the 'SPARCS' source have a simpler structure, with two options (yes and no), while the risk factors from the 'Birth Certificate' source have three options (yes, no, and unknown). By removing the risk factors from the 'Birth Certificate' source, the analysis will be easier and more consistent when exploring tobacco and alcohol use.

3. Questions

In this part of my discisson I will anser the folling questions:

- What are the ten most risk factors to get a sepsis during the maternal window?
- From this top ten, which are statistically significant (p-value < 0.05)?
- What are the top ten risk factors with the highest incidence for a statistically significant sepsis (any or severe)?
- Is there a relationship between the Number of Live Births and the Sepsis Incidence?
- Do other Risk Factors have a significant risk of getting a sepsis?

To address these questions, I opted to create visualizations in the form of bar plots and scatter plots. These visuals aim to highlight the most prevalent risk factors, identify the frequently occurring significant risk factors, and examine any potential relationships between the risk factors and the incidence of sepsis.

I created a uniform color schema for the visuals:

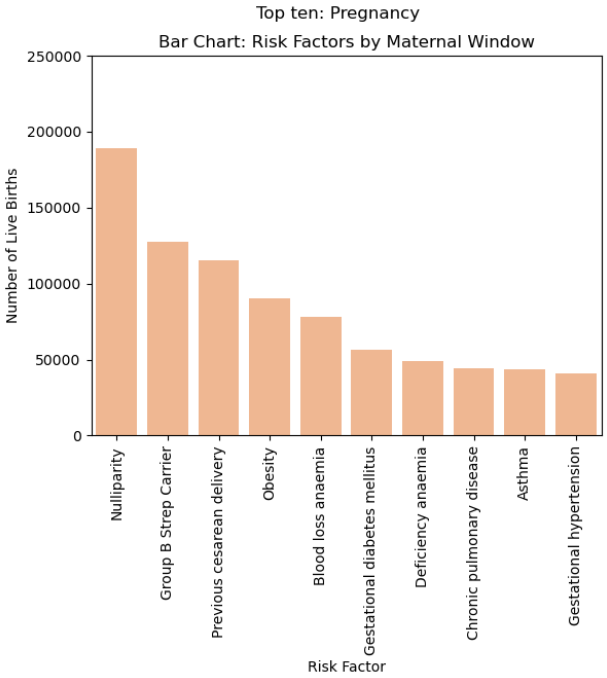
Maternal Window	*Color*
Combination of all	blue
Pregnancy	orange
Delivery	green
Postpartum	violet

3.1 What are the ten most risk factors to get a sepsis during the maternal window?

With this question I want to show the ten most often risk factors during the three maternal windows (pregnancy, delivery and postpartum). Top ten means the highest number of livebirths with women who have these risk factors. For the first question I did not take into account if the sepsis is ‘any sepsis or ‘severe sepsis’.

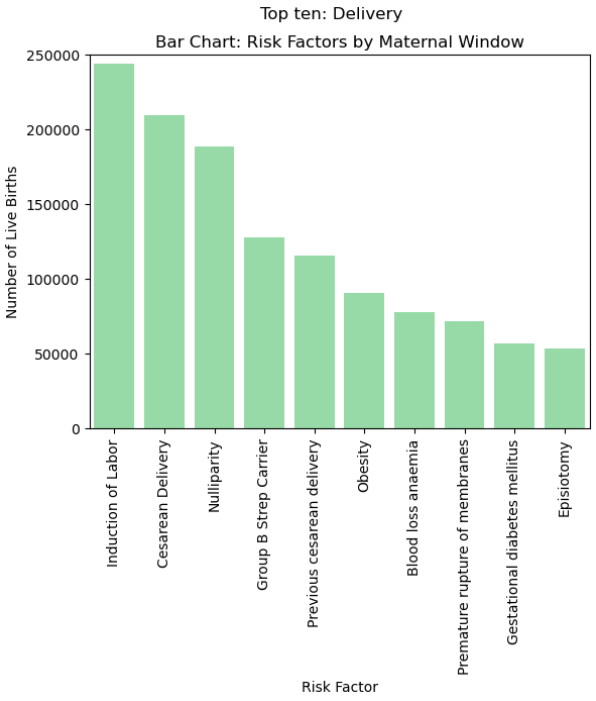
<u>During pregnancy:</u>

- 'Nulliparity',
- 'Group B Strep Carrier',
- 'Previous cesarean delivery',
- 'Obesity',
- 'Blood loss anemia',
- 'Gestational diabetes mellitus',
- 'Deficiency anemia',
- 'Chronic pulmonary disease',
- 'Asthma',
- 'Gestational hypertension'



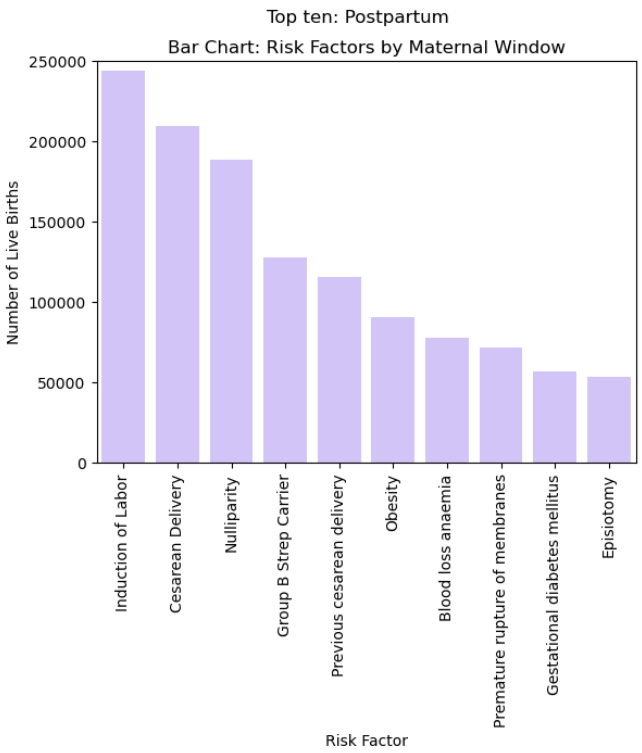
<u>During delivery:</u>

- 'Induction of Labor',
- 'Cesarean Delivery',
- 'Nulliparity',
- 'Group B Strep Carrier',
- 'Previous cesarean delivery',
- 'Obesity',
- 'Blood loss anemia',
- 'Premature rupture of membranes',
- 'Gestational diabetes mellitus',
- 'Episiotomy'



Postpartum:

- 'Induction of Labor',
- 'Cesarean Delivery',
- 'Nulliparity',
- 'Group B Strep Carrier',
- 'Previous cesarean delivery',
- 'Obesity',
- 'Blood loss anaemia',
- 'Premature rupture of membranes',
- 'Gestational diabetes mellitus',
- 'Episiotomy'



As you can see during delivery and postpartum there are the same top ten risk factors.

In the next step, I examined the top ten risk factors for each maternity window. The most frequently observed risk factors across all three stages of maternity are:

- Previous cesarean delivery
- Nulliparity (having no previous pregnancies)
- Group B Strep Carrier
- Blood loss anemia
- Gestational diabetes mellitus
- Obesity

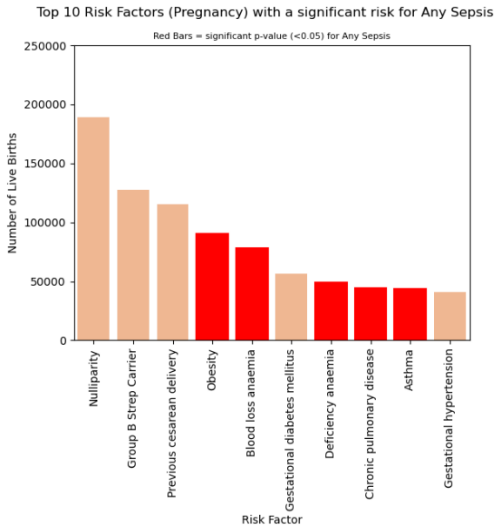
These six risk factors were found to be the most commonly occurring factors during pregnancy, delivery, and postpartum, based on the findings of this study.

3.2 From this top ten, which are statistically significant (p-value < 0.05)?

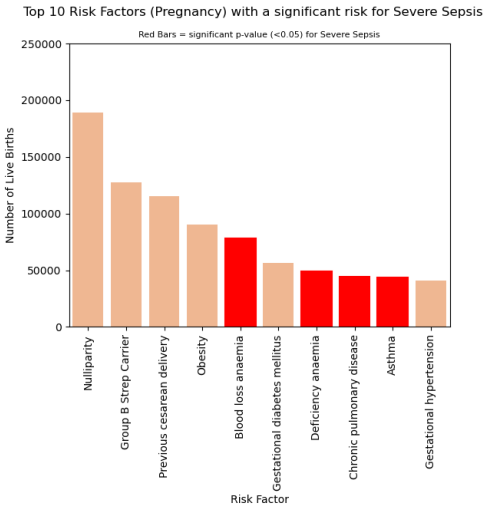
The first question just answered how many women had each risk factor, but it did not determine if there is a statistically significant risk of developing sepsis. The next visuals will show which of these risk factors have a p-value < 0.05 for both any sepsis and severe sepsis.

3.2.1 During Pregnancy

Any Sepsis



Severe Sepsis



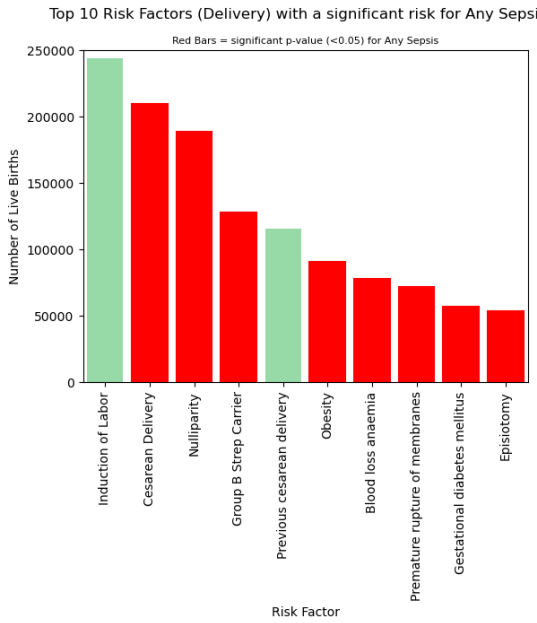
During pregnancy, there is a significant risk for any sepsis and severe sepsis if women have risk factors like:

- 'Asthma',
- 'Blood loss anaemia',
- 'Chronic pulmonary disease',
- 'Deficiency anaemia'

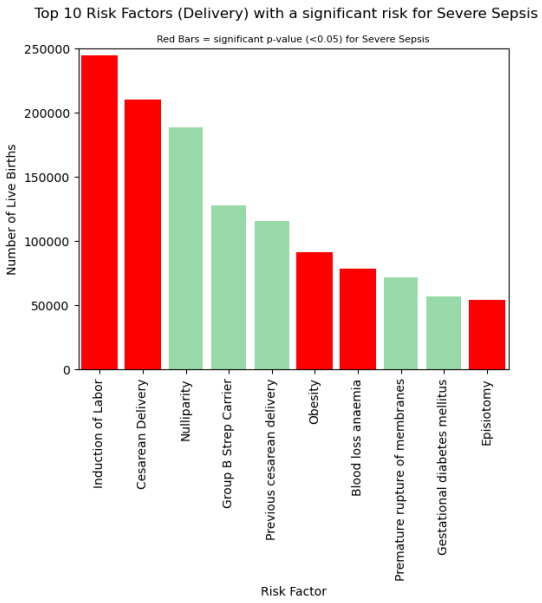
'Obesity' is only a risk factor for any sepsis, but not for severe.

3.2.2 During Delivery

Any Sepsis



Severe Sepsis



During delivery there is a significant risk for both, any and severe, sepsis, if women have the following risk factors:

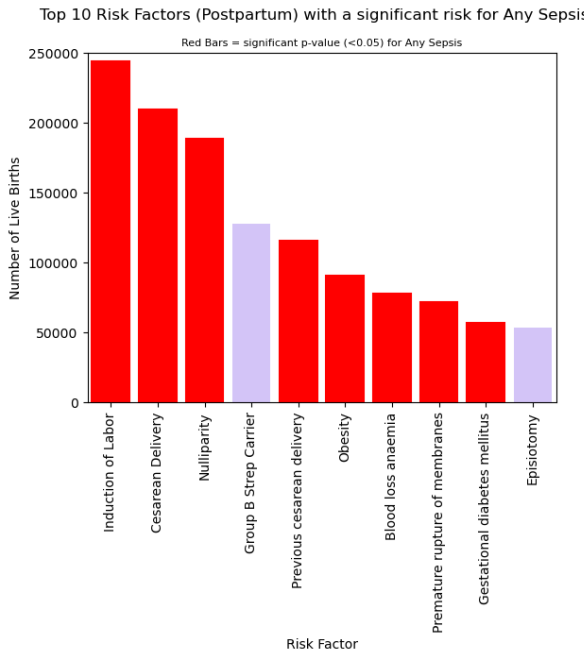
- 'Cesarean Delivery',
- 'Group B Strep Carrier',
- 'Obesity',
- 'Blood loss anaemia'

'Obesity', 'Episiotomy', 'Nulliparity' and 'Premature rupture of membranes' are the risk factors, that have a statistically signicant risk for any sepsis, but not for a severe sepsis.

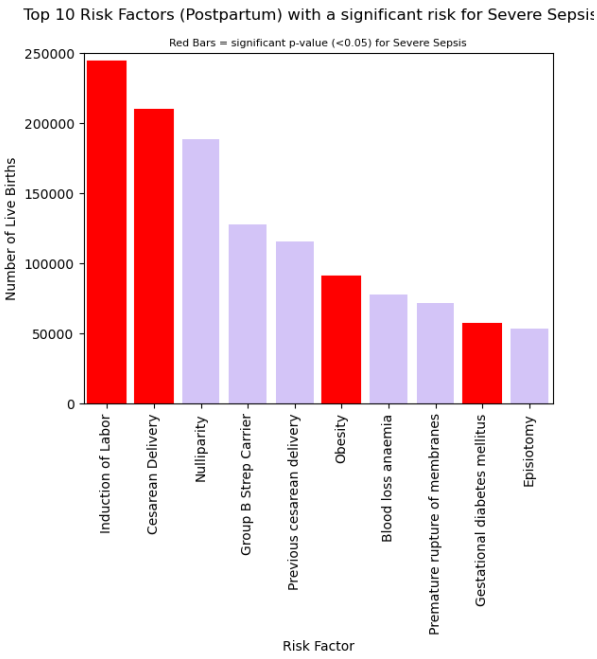
‘Induction of labor’ leads statistically significant to a severe sepsis, but not to any sepsis.

3.3.3 Postpartum

Any Sepsis



Severe Sepsis



Postpartum there is a significant risk for both, any and severe, sepsis, if women have the following risk factors:

- 'Gestational diabetes mellitus',
- 'Obesity',
- 'Cesarean Delivery',
- 'Induction of Labor'

'Previous cesarean delivery','Blood loss anaemia','Nulliparity' and 'Premature rupture of membranes' only lead to any sepsis, but not severe.

3.3 What are the top ten risk factors with the highest incidence for a statistically significant sepsis (any or severe)?

In the following sections (3.3.1 to 3.3.3), you will find bar charts depicting the top ten significant risk factors for each maternal window, categorized by any sepsis and severe sepsis. The specific risk factors are not listed here as they are visually presented in the charts. Additionally, I will provide combined lists of the risk factors discussed in this question.

Here are the risk factors that are statistically significant for any sepsis across all maternal windows:

- 'Pulmonary edema / Acute heart failure',
- 'Paralysis',
- 'Shock',
- 'Temporary tracheostomy',
- 'Chronic renal disease',
- 'Adult respiratory distress syndrome',
- 'Chronic congestive heart failure',
- 'History of Sepsis (w/in 1yr prior to start of pregnancy)',
- 'Puerperal cerebrovascular disorders',
- 'Peripheral vascular disorders',
- 'Pulmonary hypertension'
- 'Hysterectomy',
- 'Acute renal failure',
- 'Amniotic fluid embolism',
- 'Metastatic cancer',
- 'Conversion of cardiac rhythm',
- 'Acute myocardial infarction',
- 'Cardiac arrest/ventricular fibrillation',
- 'Air and thrombotic embolism',
- 'Weight loss',
- 'Fluid and electrolyte disorders',
- 'Pulmonary circulation disorders',
- 'Ventilation',

And this are the risk factors that are statistically significant for a severe sepsis across all maternal windows:

- 'Pulmonary edema / Acute heart failure',
- 'Paralysis',
- 'Shock',
- 'Organ Transplant',
- 'Temporary tracheostomy',
- 'Adult respiratory distress syndrome',
- 'Chronic congestive heart failure',
- 'History of Sepsis (w/in 1yr prior to start of pregnancy)',
- 'Puerperal cerebrovascular disorders',
- 'Peripheral vascular disorders',
- 'Acute renal failure',
- 'Amniotic fluid embolism',
- 'Metastatic cancer',
- 'Conversion of cardiac rhythm',
- 'Acute myocardial infarction',
- 'Cardiac arrest/ventricular fibrillation',
- 'Air and thrombotic embolism',
- 'Weight loss',
- 'Fluid and electrolyte disorders',
- 'Pulmonary circulation disorders',
- 'Cardiac arrhythmias',
- 'Chronic ischemic heart disease',
- 'Ventilation',
- 'Pulmonary hypertension'

And this risk factors are both – statistically significant for any sepsis and a severe sepsis:

- 'Pulmonary edema / Acute heart failure',
- 'Paralysis',
- 'Shock',
- 'Temporary tracheostomy',
- 'Chronic congestive heart failure',
- 'Adult respiratory distress syndrome',
- 'History of Sepsis (w/in 1yr prior to start of pregnancy)',
- 'Puerperal cerebrovascular disorders',
- 'Peripheral vascular disorders',
- 'Acute renal failure',
- 'Amniotic fluid embolism',
- 'Metastatic cancer',
- 'Conversion of cardiac rhythm',

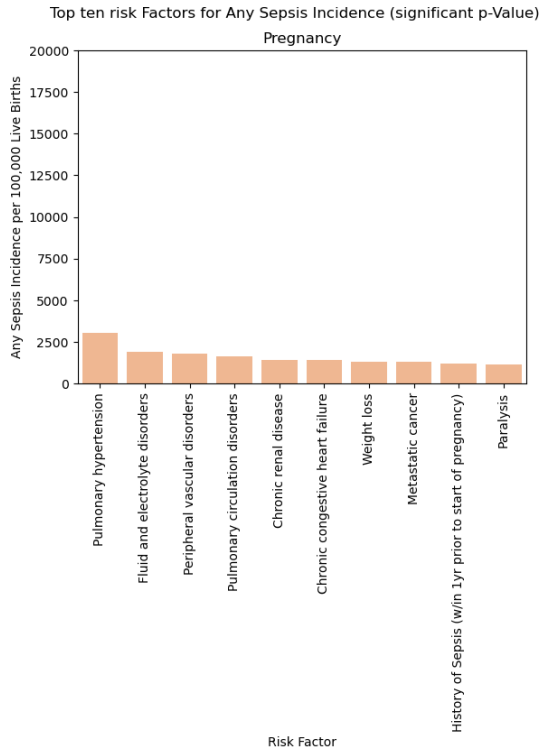
- 'Acute myocardial infarction',
- 'Cardiac arrest/ventricular fibrillation',
- 'Air and thrombotic embolism',
- 'Weight loss',
- 'Fluid and electrolyte disorders',
- 'Pulmonary circulation disorders',
- 'Ventilation',
- 'Pulmonary hypertension'

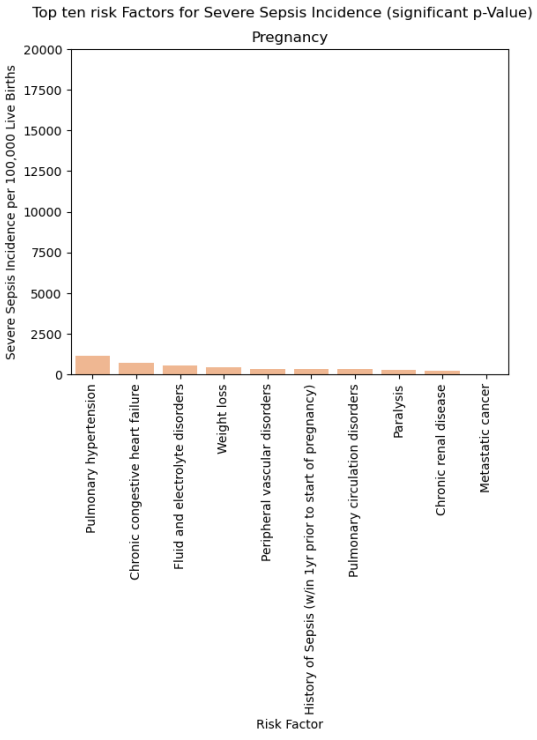
These are several lists containing numerous risk factors that are statistically significant for sepsis (any or severe). However, none of these risk factors appear in the top ten risk factor lists. Therefore, these risk factors listed here are not very common occurrences.

However, none of the risk factors exhibit a significant p-value for any sepsis across all maternal window stages. Similarly, there are no risk factors with a significant p-value for severe sepsis across all maternal window stages.

3.3.1 Pregnancy

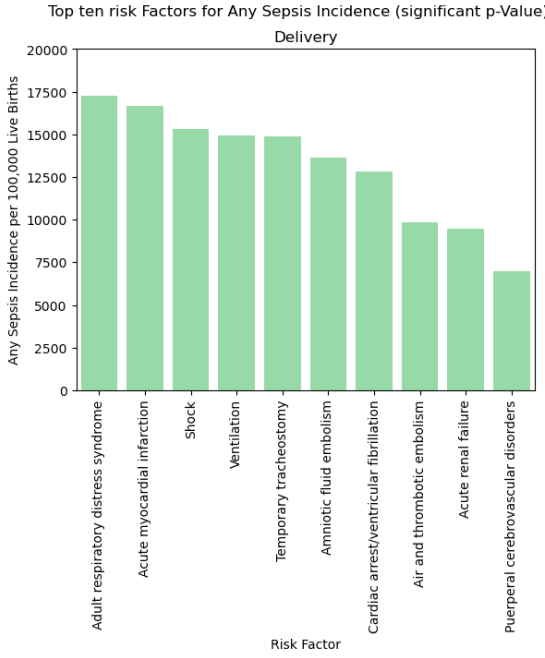
This two barplots show the top ten significant risk factors during pregnancy for any sepsis (left) and a severe sepsis (right).

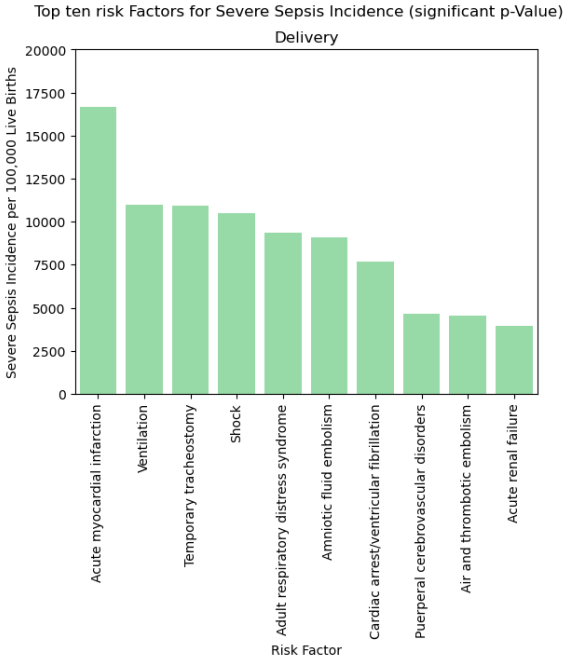




3.3.2 Delivery

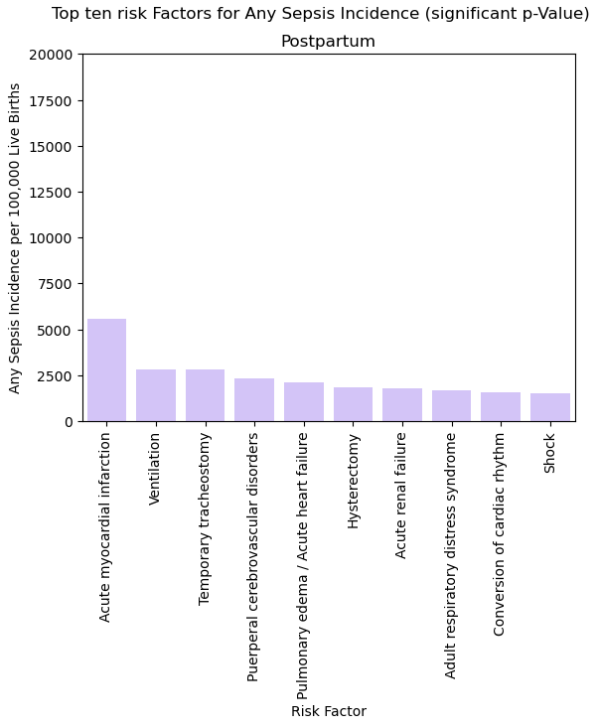
This two barplots show the top ten significant risk factors during delivery for any sepsis (left) and a severe sepsis (right).

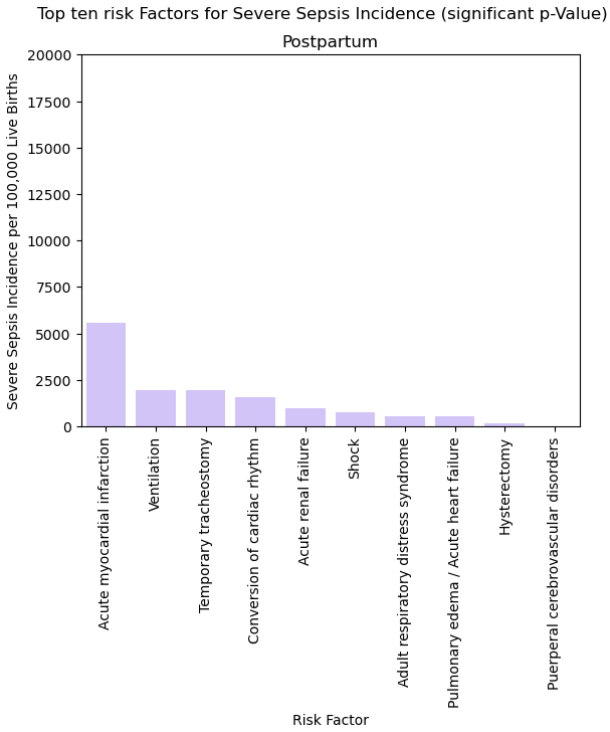




3.3.3 Postpartum

This two barplots show the top ten significant risk factors postpartum for any sepsis (left) and a severe sepsis (right).



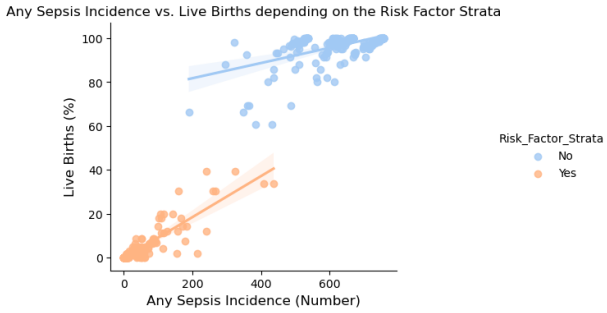


The key observation from studying these visuals is that the incidence of sepsis is significantly higher during the delivery stage compared to the pregnancy or postpartum stages.

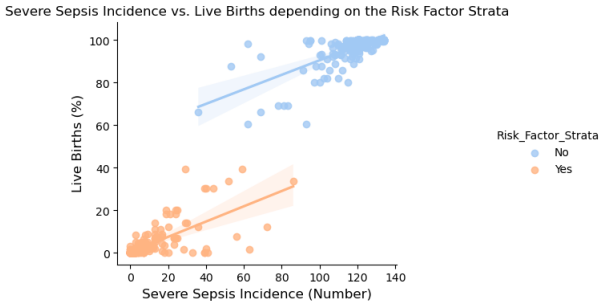
3.4 Is there a relationship between the Number of Live Births and the Sepsis Incidence?

The question asks whether there is a relationship between the number of live births and the incidence of any or severe sepsis. The scatterplots will show the relationship between the percentage of live births and the incidence of sepsis, both in terms of percentage and per 100,000 live births, for women who had the risk factor ('yes') and those who did not ('no').

Any Sepsis



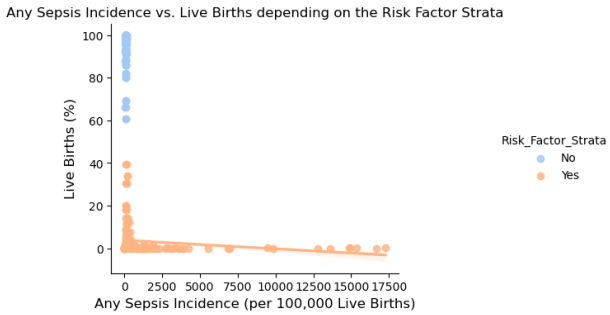
Severe Sepsis



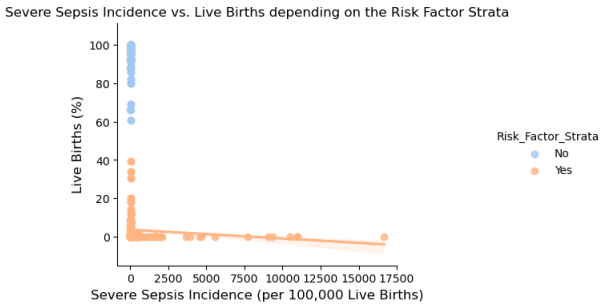
Upon initial observation, it appears that there is a linear relationship between the percentage of live births and the incidence of sepsis for each risk factor group. Specifically, a higher percentage of live births corresponds to a higher incidence of sepsis. Additionally, it is worth noting that women without the risk factor (represented by blue points) tend to have a higher incidence of sepsis.

To further investigate and confirm this relationship, I conducted a more detailed analysis. Instead of considering the total number of sepsis incidences, I focused on the incidence of sepsis per 100,000 live births. This approach allows for a more accurate and meaningful comparison across different risk factor groups.

Any Sepsis



Severe Sepsis



Upon further analysis, an additional relationship becomes evident:

- Women without any risk factor have a sepsis incidence rate of approximately 0.
- Only women with the specific risk factors exhibit higher sepsis incidences.
- Furthermore, there is an inverse relationship between the sepsis incidence rate and the live birth rate. In other words, as the sepsis incidence rate increases, the rate of live births decreases.

3.5 Do other Risk Factors have a significant risk of getting a sepsis?

This question will answer, if there are risk factors getting a sepsis beside the comorbidities we took a look in the first questions.

The Risk Factors I will analyze in this question are:

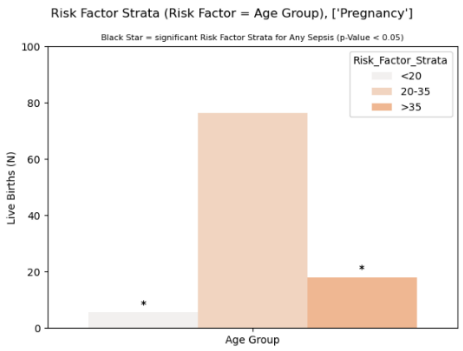
- 'Age Group'
- 'Education'
- 'Race/Ethnicity'
- 'Region of Residence'
- 'Trimester Beginning Prenatal Care'

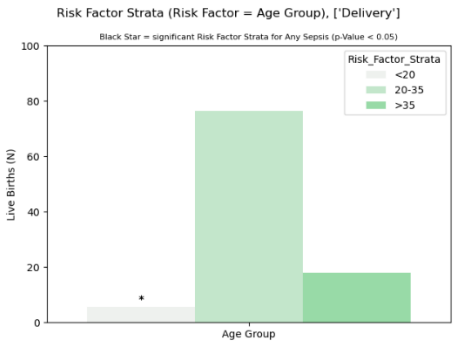
The star (*) on the bars shows, if there the p-value for any sepsis or severe sepsis is statistically significant.

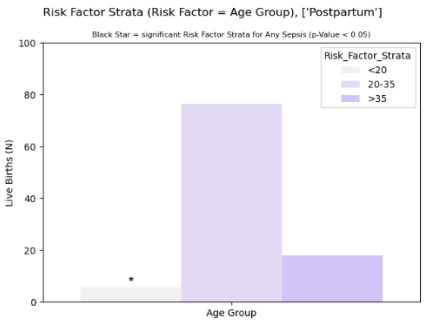
In the following chapters, there will be a total of 6 risk factors discussed. Each risk factor will be analyzed for all 3 maternity windows and both any and severe sepsis, resulting in a total of 6 pictures for each risk factor. Due to the large number of combinations, I will focus on highlighting the main aspects related to these risk factors.

3.5.1 Age Group

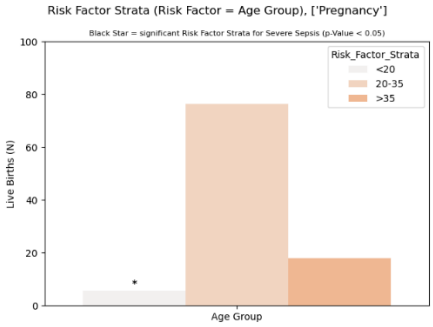
<u>Any Sepsis</u>

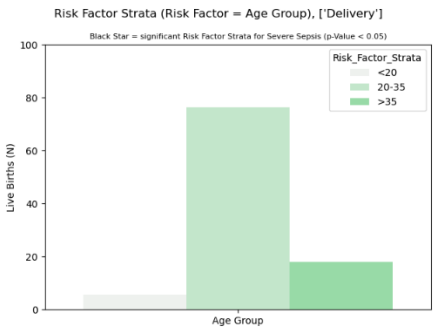


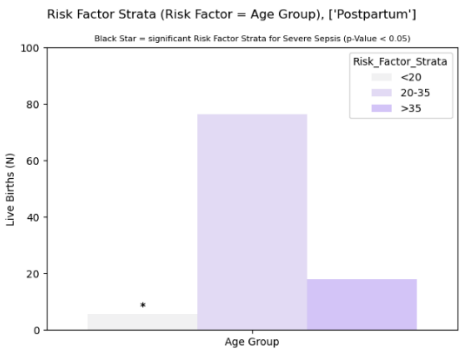




<u>Severe Sepsis</u>





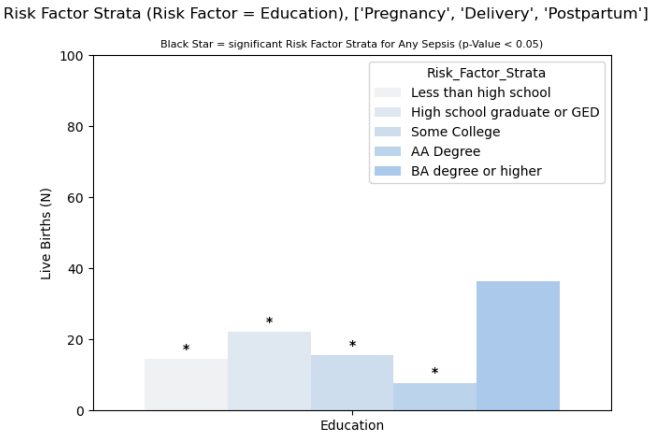


The reference for calculating the p-value for this groups if risk factors is the risk factor strata ’20-35’.

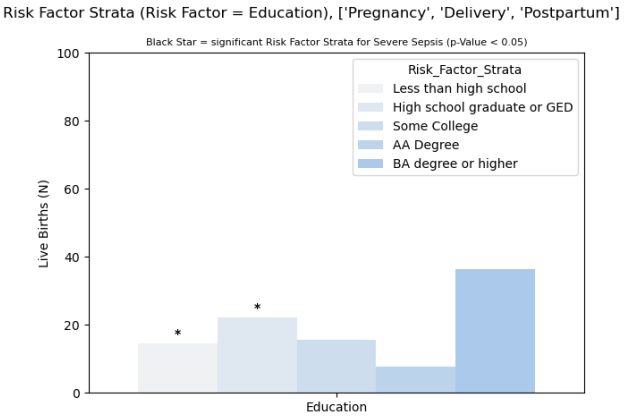
For women who are younger than 20 years and in various stages of pregnancy and during delivery or postpartum, there appears to be a significant risk of developing any or severe sepsis across nearly all combinations.

3.5.2 Education

Any Sepsis



Severe Sepsis



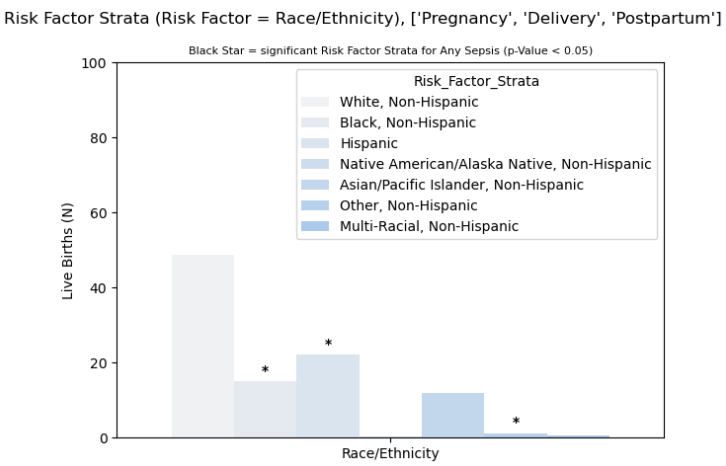
Due to the large number of combinations for the Risk Factor 'Education', I have created a consolidated diagram that includes all three maternity windows in a single bar plot.

The reference point for calculating the p-value is the Risk factor stratum 'BA degree or higher'.

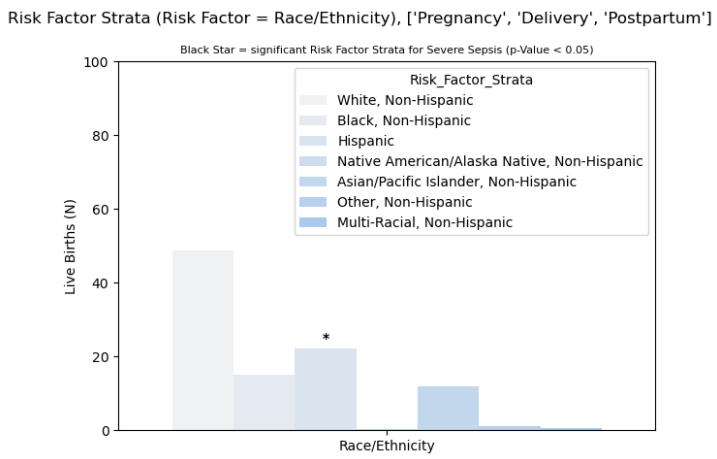
In summary, there seems to be a potential effect where lower education levels are associated with a higher likelihood of developing any or severe sepsis. However, to validate this initial observation, further calculations such as regressions or statistical tests are required.

3.5.3 Race/Ethnicity

Any Sepsis



Severe Sepsis



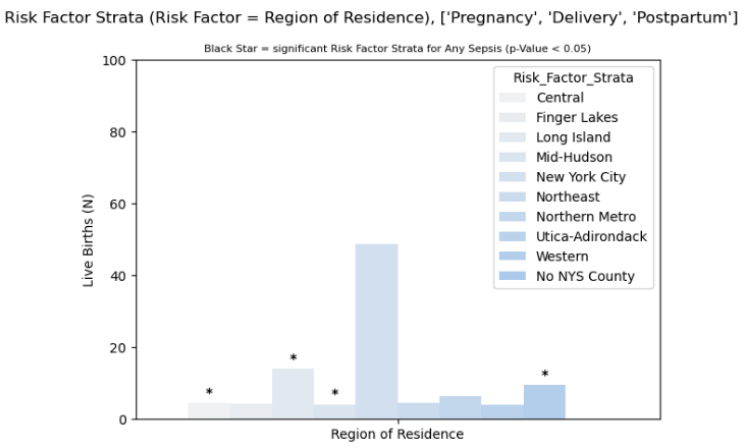
Similar to the education risk factor, the barplots for the risk factor 'Race/Ethnicity' are presented as combined visuals. This approach allows for a comprehensive comparison across all three maternity windows. By using this merged visualization, we can explore the relationship between race/ethnicity and the risk of developing any or severe sepsis. However, it is important to note that additional analyses, such as regression or statistical tests, are necessary to further investigate and confirm these findings.

In this analysis, the reference group for calculating the p-values is 'White, Non-hispanic'.

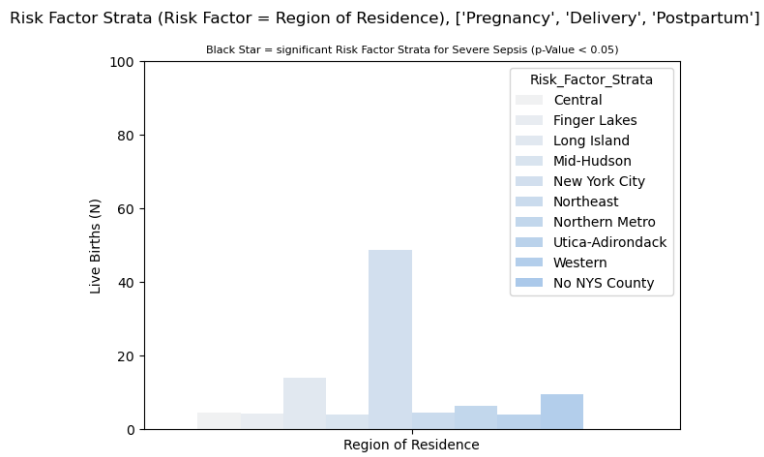
Interestingly, there appears to be a significant risk for Hispanic women in developing any or severe sepsis. However, it is important to note that this project does not aim to delve into the reasons behind the differences in p-values among different race/ethnicity groups. The focus of this project lies elsewhere, and further investigation into the underlying factors contributing to these disparities is beyond its scope.

3.5.4 Region of Residence

Any Sepsis



Severe Sepsis



Similar to the previous chapter, the risk factor 'Region of Residence' is visualized in a combined bar plot for all three maternity stages. The reference group in this analysis is the region with the highest bar, 'New York City'.

The left visualization, which focuses on the significance of getting any sepsis, reveals that some regions (Central, Long Island, Mid Hudson, and non-NYS Country) have significant p-values, indicating a potential risk.

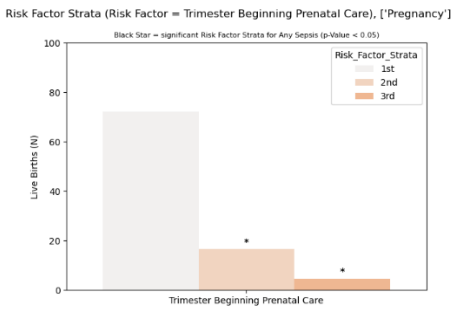
On a positive note, there is no variation in the likelihood of developing severe sepsis based on the region where women reside.

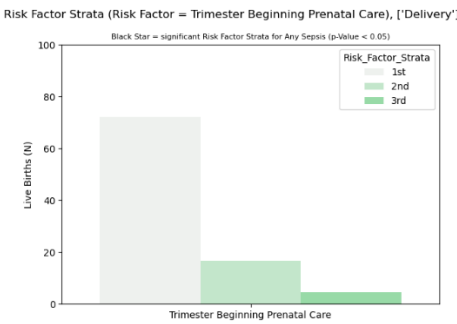
3.5.5 Trimester Beginning Prenatal Care

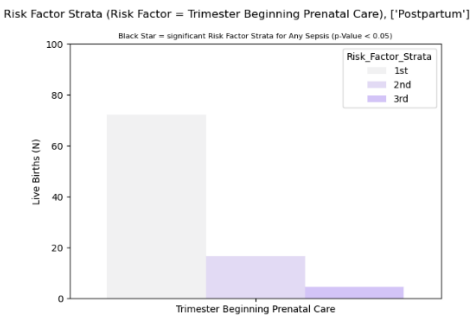
The visuals for the risk factor 'Trimester Beginning Prenatal Care' are divided into the three maternity stages once again.

The reference group in this analysis is the Trimester with the highest bar (1st trimester).

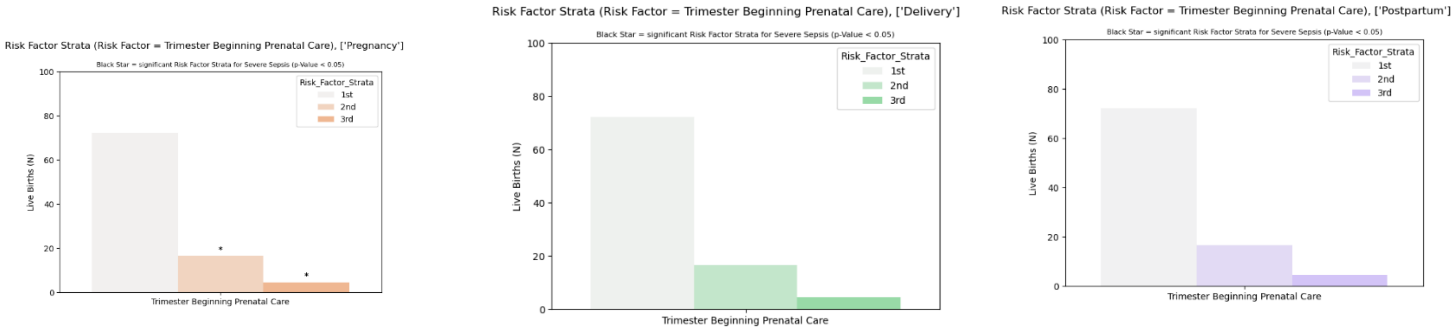
Any Sepsis







Severe Sepsis



These bar plots demonstrate that there is a significantly higher risk of experiencing any or severe sepsis during pregnancy if women initiate their prenatal care later in the 2nd or 3rd trimester. However, there is no difference in the significance of sepsis risk based on the timing of prenatal care initiation during delivery or postpartum.

Fortunately, the majority of women initiate their prenatal care early in pregnancy, which is good news. However, conducting a thorough analysis of recent literature would be beneficial in order to definitively conclude that early initiation of prenatal care leads to lower risks of complications during maternity.

3.5.6 Conclusion of the other risk factors

However, these plots provide additional valuable information:

1. Most of the women (around 80%) are between 20 and 35 years old when they are pregnant or giving birth. This suggests that the majority of women in the population fall within this age range.

2. Approximately 70% of women start their prenatal care in the first trimester. This indicates that a significant portion of women initiate their prenatal care early in their pregnancy.

4. Conclusion

In total, 94 risk factors were examined in this dataset to determine if there is a statistically significant risk for any or severe sepsis. Here are the key insights from my exploration:

1. None of the risk factors in the statistically significant top ten list are included in the list of most common risk factors. This indicates that the risk factors that are most commonly significant (represented by red bars) are not the same as the most frequently occurring risk factors.

2. Demographic factors exhibit varying influences on the significance of developing sepsis. Younger women and those who initiate prenatal care after the first trimester appear to be associated with significant risks of sepsis. Additionally, factors such as education, race/ethnicity,

and region of residence have distinct impacts on sepsis significance. Further investigations are required to determine if any patterns exist within these factors.

3. Risk factors with a very high incidence are rare, indicating a low number of live births associated with them.