COMP5310 Project Stage 1

Mental health in the pregnancy during the COVID-19

Contents

1 Data Cleaning and Exploration

1.1 Topic and research question

The COVID-19 pandemic has profoundly affected global mental health, disproportionately impacting pregnant individuals. Amidst this crisis, our research investigates the psychological effects on expectant mothers, focusing on how socioeconomic status, pandemic-related health perceptions, and neonatal outcome concerns drive their anxiety and depression. Our pivotal question is: **How do socioeconomic factors, perceived pandemic risks, and neonatal outcome concerns influence the mental health of pregnant women during the COVID-19 era?**

Addressing this question promises to unravel the complex mental health ramifications of the pandemic on expectant mothers, benefiting key stakeholders:

1. **Academic Researchers**: Our insights enable researchers to explore the pandemic's psychological dimensions further, enriching academic discourse and sparking targeted investigations into maternal stressors during health crises.
2. **Clinical Practitioners**: Therapists and healthcare professionals gain crucial data to refine intervention strategies, offering personalized support to mitigate pregnant women's mental health challenges.
3. **Policymakers**: With a deeper comprehension of the psychological impact on pregnant women, policymakers can develop targeted supports, crafting policies that address the unique needs of expectant mothers in times of crisis.

Our findings aim to facilitate a multidisciplinary approach to understanding and addressing the psychological impacts of the pandemic, contributing to a foundation for future resilience among pregnant women and their families.

1.2 Data description

1.2.1 Provenance of the data

Our primary dataset was sourced from the Kaggle platform [1] in January 2024, discovered by our team two months later. Its genesis traces back to a comprehensive survey conducted between April 2020 and April 2021, targeting pregnant individuals across Canada [1]. This dataset, initially compiled and shared via the Open Science Framework in April 2023 [2], was the result of collaborative efforts to reach participants through diverse channels. The survey aimed at those 17 years and older, pregnant, residing in Canada, within 35 weeks of gestation, and proficient in English or French.

1.2.2 Data license

The dataset is governed by the **Creative Commons Attribution 4.0 International** (CC BY 4.0) license. This license permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. It enables both commercial and non-commercial reapplication and adaptation, demanding only that users credit the creators, link to the license terms, and denote any modifications. This openness facilitates widespread academic and practical application, promoting the advancement of knowledge and interventions based on our findings.

1.2.3 Data structure and metadata

Our dataset comprises a collection of 10,772 entries across 16 meticulously selected variables, aimed at painting a comprehensive picture of the psychological well-being of pregnant women amidst the COVID-19 pandemic. This dataset is particularly rich, featuring a balanced mix of quantitative and categorical data types, including scales for depression and anxiety (Edinburgh Postnatal Depression Scale [2], PROMIS Anxiety [3]), socioeconomic factors (Household Income, Maternal Education), perceptions of health risks (Threaten Life, Threaten Baby Danger, Threaten Baby Harm), and newborn health metrics (Birth Length, Birth Weight, NICU Stay). A

[1] Dataset on the Kaggle platform: Mental health in the pregnancy during the COVID-19.
[2] Dataset on the Open Science Framework platform: Pregnancy During the COVID-19 Pandemic Study.

detailed data dictionary A.1 provided in Appendix A serves as a guide, offering clear definitions, data types, and relevance of each attribute, facilitating an in-depth understanding of the dataset's structure and potential analytical value.

1.3 Data quality and cleaning

1.3.1 Data quality assurance and cleaning

Initial Assessment Upon initial assessment, the dataset presented a significant challenge with up to 50% missing data in some variables, necessitating a nuanced approach to cleaning. Our strategy involved a preliminary elimination of surveys lacking data in more than 10 of the 15 non-identifier attributes, prioritizing completeness in variables critical to our analysis. Additionally, we standardized variable formats for consistency, including the conversion of the *Delivery Date* attribute to a uniform naming convention and adjusting the scale of birth weights from grams to kilograms for analytical convenience.

Normalization and Encoding The normalization phase included transforming string-dated attributes to datetime objects to facilitate temporal analyses. Recognizing the importance of categorical variables in our study, we employed one-hot encoding for the *Language* variable to retain informational value without resorting to binary reduction. This approach ensures that machine learning models can interpret the data without loss of nuance in linguistic diversity.

Advanced Cleaning Techniques To tackle the complex issue of data integrity, particularly concerning the significant proportion of missing values, our strategy was developed to ensure the highest standards of data cleanliness without compromising the dataset's integrity or distributional characteristics.

For categorical variables, we designated a new category, *Unknown*, for missing entries, thus retaining the data's structural integrity. This approach was particularly applied to variables such as *Household Income* and *Maternal Education*. Additionally, for variables like *Delivery Mode* and *NICU Stay*, which are inherently binary but plagued with missing entries, we refrained from arbitrary imputation. Instead, we employed one-hot encoding, a decision that preserved the original data's nuance and allowed for a more inclusive analysis that accounts for the unknowns without bias. Subsequently, the total number of columns in our dataset increased to 21, ensuring a coherent augmentation of our categorical data handling framework.

The challenge with continuous data was its susceptibility to distortion through naive imputation techniques. Given the dataset's considerable size and the significant percentage of missing values, simple fill-ins like mean or median could drastically shift the overall data distribution, leading to misleading analyses. Our strategy involved a more sophisticated approach using predictive modeling techniques that consider the underlying patterns and relationships within the data. Kernel Density Estimation (KDE) was particularly useful for independent variables such as *Maternal Age*, *Delivery Date*, and *Gestational Age At Birth*, allowing us to impute missing values based on the probability density function of each variable, thus maintaining their original distribution.

A notable concern was the relational dynamics between certain variables, especially *Birth Length* and *Birth Weight*, which exhibit a natural correlation. Preliminary analysis revealed a Pearson correlation coefficient of approximately 0.47, indicating a moderate positive relationship. To preserve this relationship, we first employed linear regression for cases where one variable was missing and the other was present, within a truncated range that maximized their correlation. For missing values beyond this range, the k-nearest neighbors (kNN) algorithm offered a solution by imputing based on the proximity to the nearest valid entries, taking into account the distance and thus ensuring a more accurate fill-in. Finally, to address any remaining gaps, an iterative imputer with a random forest regressor was deployed, leveraging the interconnectedness of the dataset's variables to predict missing values with high accuracy. This multi-pronged approach ensured that the imputation process did not erode the dataset's internal consistency or its foundational relationships.

1.3.2 Data ingestion

Leveraging the versatile Python Pandas library, we efficiently ingested the dataset from its original .csv format, transitioning seamlessly to a cleaned and analytically ready state. This process was underpinned by a clear

schema that facilitated not only the initial data cleaning efforts but also subsequent exploratory data analysis and predictive modeling, ensuring a coherent and reliable foundation for our research.

1.4 Exploratory data analysis (EDA)

1.4.1 Description

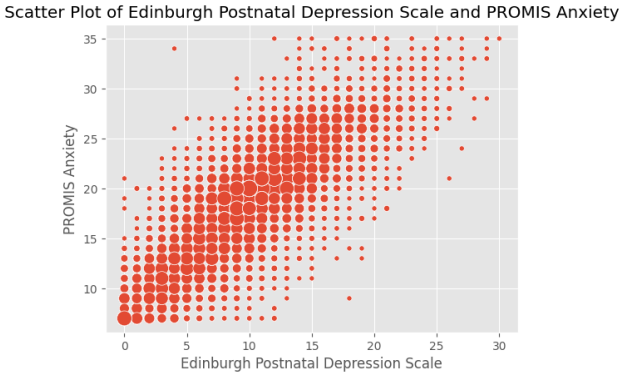
Our EDA commenced with the calculation of a correlation matrix, a pivotal step that allowed us to discern both positive and negative relationships between variables. This matrix was instrumental in identifying the interplay of attributes within our dataset, setting the stage for deeper analysis directly tied to our research question.

To address our research question, we honed in on the *Edinburgh Postnatal Depression Scale* and *PROMIS Anxiety* as our focal points, noting their significant correlation, the highest among all attributes examined. Given their integer nature, visualizing a direct scatter plot posed challenges due to potential point overlap. To circumvent this, we adjusted point sizes based on frequency, unveiling a linear relationship between these attributes, particularly within their principal components, as depicted in Figure 1a.

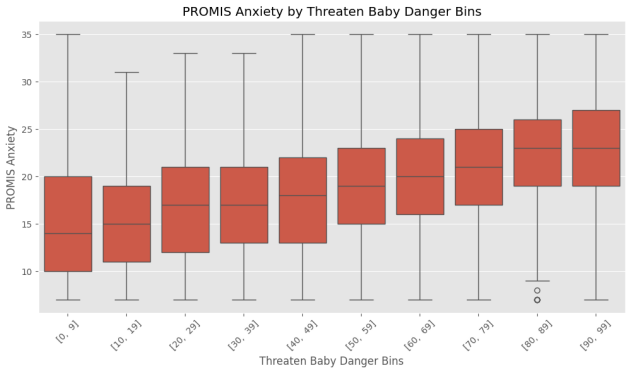
Further analysis revealed *Threaten Baby Danger* as a key factor related to both depression and anxiety levels. Figure 1b illustrates a clear pattern: higher levels of perceived danger to the baby correlate with increased anxiety and depression levels among pregnant individuals. This trend is also observed with the attribute Threaten Baby Harm, underscoring the impact of external threats on mental health.

Additionally, our exploration into the distribution of depression and anxiety scores across different *Delivery Mode* and *NICU Stay* conditions revealed largely overlapping distributions, suggesting minimal impact from these factors. However, a notable difference emerged when analyzing anxiety scores in relation to *Language*.

Our analysis also extended to categorical attribute pairs through a heatmap approach, effectively showcasing the frequency of each attribute combination. This revealed a trend between *Maternal Education* and *Household Income*, with higher education levels often correlating with higher income. A strong association was also noted between *NICU Stay* (No) and *Delivery Mode* (Vaginally), indicating a pattern worth further exploration.



(a) Scatter Plot



(b) Box Plot

Figure 1: Supporting Figures

1.4.2 Discussion

The insights gleaned from our EDA are foundational for the subsequent modeling phase. The strong correlation between our primary targets suggests the potential for multitask learning approaches in modeling, promising a more nuanced understanding of the underlying patterns. The pronounced effect of external threats, particularly those concerning infant safety, on maternal mental health is a critical finding, highlighting areas for targeted interventions. The relationships unearthed between various attributes underscore the necessity for thoughtful feature engineering prior to model development. This preparation is key to enhancing model accuracy and relevance, ensuring that our predictive analysis is both robust and deeply informative. Our EDA has not only illuminated direct answers to our research question but also paved the way for a sophisticated modeling approach that will further our understanding of the intricate dynamics at play in maternal mental health during the COVID-19 pandemic.

References

[1] C. Lebel, L. Tomfohr-Madsen, G. Giesbrecht, *et al.*, "Prenatal mental health data and birth outcomes in the pregnancy during the covid-19 pandemic dataset," *Data in Brief*, vol. 49, p. 109 366, 2023, ISSN: 2352-3409. DOI: https://doi.org/10.1016/j.dib.2023.109366. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352340923004857.

[2] J. L. Cox, J. M. Holden, and R. Sagovsky, "Detection of postnatal depression: Development of the 10-item edinburgh postnatal depression scale," *The British journal of psychiatry*, vol. 150, no. 6, pp. 782–786, 1987.

[3] D. Cella, S. W. Choi, D. M. Condon, *et al.*, "Promis® adult health profiles: Efficient short-form measures of seven health domains," *Value in health*, vol. 22, no. 5, pp. 537–544, 2019.

A Appendix

Table A.1: Data Dictionary

Attribute name	Data type	Unit	Description	Range
OSF ID (Primary Key)	Integer	N/A	Open Science Framework ID number	N/A
Maternal Age	Floating point	Year	Maternal age at intake	N/A
Household Income	String	Canadian dollar	What is the total household income, before taxes and deductions, of all the household members from all sources in 2019	Less than $20, 000, $20,000-$39,999, $40,000-$69,999, $70,000-$99,999, $100,000-$124,999, $125,000-$149,999, $150,000-$174,999, $175,000-$199,999, $200,000+
Maternal Education	String	N/A	N/A	Less than high school diploma, High school diploma, College/trade school, Undergraduate degree, Masters degree, Doctoral Degree
Edinburgh Postnatal Depression Scale	Integer	N/A	N/A	N/A
PROMIS Anxiety	Integer	N/A	Higher scores indicate greater severity of anxiety	[7, 35]
Gestational Age At Birth	Float pointing	Week	Gestational age at birth	N/A
Delivery Date	Datetime	N/A	N/A	N/A
Birth Length	Floating point	Centimeter	N/A	N/A
Birth Weight	Floating point	Gram	N/A	N/A
Delivery Mode	String	N/A	N/A	Vaginally, Caesarean-section (c-section)
NICU Stay	Boolean	N/A	Was your infant admitted to the NICU?	Yes, No
Language	String	N/A	Survey language	English, French
Threaten Life	Integer	N/A	How much do (did) you think your life is (was) in danger during the COVID-19 pandemic?	[0, 100]
Threaten Baby Danger	Integer	N/A	How much do (did) you think your unborn baby's life is (was) in danger at any time during the COVID-19 pandemic?	[0, 100]
Threaten Baby Harm	Integer	N/A	How much are you worried that exposure to the COVID-19 virus will harm your unborn baby?	[0, 100]