Data Analytics Capstone Topic Approval Form

Student Name: [redacted]

Student ID: [redacted]

Capstone Project Name: "A Deep Dive: Exploring the Impact of Maternal Characteristics on Preterm Birth through Multiple and Logical Regression Analysis"

Project Topic: Exploring factors affecting preterm birth: a multiple and logistical regression analysis approach. The analysis will examine maternal characteristics and their relationship on the likelihood of preterm births.

☒ **This project does not involve human subjects research and is exempt from WGU IRB review.**

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

Hypothesis:
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.

Context:
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. 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 and tobacco and/or substance abuse (*Premature Birth*, 2022). By identifying specific material factors related to preterm birth, healthcare providers can implement targeted interventions, monitoring, and care plans to mitigate the risk and improve outcomes. 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 procedure 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).

Data:
Data Context:

- The data is sourced from Kaggle for public use at: https://www.kaggle.com/datasets/shayta/usa-natality-2020
- 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 at http://www.cdc.gov/nchs/data_access/VitalStatsOnline.htm
- The data includes descriptive tabulations of data reported on birth certificates from 2020 in the United States.

Data Content:

- The data set has 180,993 rows/observations and 39 columns.
- This analysis will include variables relating to maternal characteristics (e.g. age, education, marital status); paternal characteristics (e.g. race, age, education); previous preterm birth (if the birth was

preterm or not full term); medical conditions (e.g. the presence of hypertension, diabetes, gonorrhea); or environmental factors (e.g. use of cigarettes, epidural or spinal anesthesia during labor).

- A data table of the dataset is 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	continous
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

The primary limitation of this data set is that it only includes data for one year. In 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 births. This may show correlations 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 births. Socioeconomic factors such as receiving access to prenatal care, household income, and geographic location could also provide more data contributing preterm birth.

Data Gathering:
Data containing maternal and paternal characteristics based on live births that was 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 its 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 TXT file. However, the txt files is coded based on a lengthy coding system published in the user guide. 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 for this analysis consists of slightly more than 180,000 observations and 39 columns. The data consist of continuous and categorical variables.

Cleaning and treating the data will first include detecting duplicates, and identifying missing values, and outliers. The data set sparsity is 1.06%. The treatment of missing values will include deletion of missing values since there are a minimal amount in this dataset. Outliers come from data entry errors, measurement errors, experimental errors, sampling errors, or novelties in the data (Lacrose & Lacrose, 2019). Outliers will be calculated with z-scores. 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). One hot encoding will thus be 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. Original columns will be dropped and dummy columns remain to be used in regression models.

Data Analytics Tools and Techniques:
Exploratory data analysis and a multiple regression analysis will be completed to analyze the relationship between individual maternal characteristics and the likelihood of preterm birth. Exploratory data analysis will help understand the data by identifying patterns and generating initial insights. Overall, EDA is beneficial because it allows for data exploration, visualization (e.g. scatterplot, box plot, heatmap, violin plot, interaction plot, partial dependence plot, forest plot), pattern recognition, feature engineering, missing data analysis, and primarily hypothesis generation. A 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 and independent variable rates (Turvey, 2013). 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.

The following tools will be used for this analysis:

- Jupyter Notebook- python coding
- Python (Anaconda environment) with the following packages:
    - pandas- Main package for data uploading and manipulation
    - numpy- Main package for working with arrays
    - matplotlib- visualization
    - seaborn- Advanced visualization
    - scikit-learn- building machine learning models and for model evaluation
    - statsmodel- statistical modeling and inference
    - SciPy- optimization and statical testing
- PowerPoint- presentation of findings
    - utilizes graphs made in Python as visualizations

Justification of Tools/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 to tailor. "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 birthing 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 test, ensuring the validity of the regression analysis results. Kolmogorov-Smirnov is a normality test 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.

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, the individual contributions can be assessed in identifying the significant predictors. 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.

Project Outcomes:
This analysis aims to generate findings that will provide insights into the significant influence of individual maternal characteristics associated with preterm birth. The analysis will determine which maternal characteristics are correlated to preterm birth. The deliverable will include regression coefficients that indicate the direction and strength of these relationships. The deliverable will also include p-values and confidence intervals along with other model performance evaluations that determine the significance of each predictor in relation to preterm birth. These findings will provide a significant understanding on which factors are associated with preterm birth and the business implications for healthcare organizations which can include cost management, resource allocation, quality improvement, risk management, value-based care services, and patient stratification and outcomes. Support for the alternative hypothesis is seen in Fuchs et al (2018) who found maternal characteristics such as maternal age (>40) and hypertension were confounders of preterm birth through the use of logistic regression.

Projected Project End Date: July 15, 2023

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

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

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

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

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

Course Instructor Signature/Date:

☒ The research is exempt from an IRB Review.

☐ An IRB approval is in place (provide proof in appendix B).

Course Instructor's Approval Status: Approved

Date: 6/22/2023

Reviewed by:

Comments: Click here to enter text.