Exercise 6.1
Sanya Mohsini

The project goal is to examine trends and patterns in births across different academic groups, age groups and geographic locations within the United States.

Data Source

Data Source	The US Births 2016-2021 data set is sourced from Kraggle. This data set was originally sourced from the Centers for Disease Control and Prevention (CDC's) retrieval tool on the CDC Natality page. As the CDC is a government agency it is trustworthy and should not be biased. More information about how the source and data can be found here.
Data Collection	The data set has counts of live births occurring within the United States to U.S. residents among a variety of demographic characteristics such as state and country of residence, mother's age, mother's education, birth weight and birth gender. The data was derived from birth certificates issued in years 2016-2021. Note: It is unclear how the birth certificates are then collected, it might be manually which could led to some errors and time lag.
Contents	The US Births data set includes the following information: • State • State Abbreviation • Year of birth • Gender • Education Level of Mother • Education Level Code • Number of births • Average Age of Mother (years) • Average Birth Weight (g) It contains 5,496 rows and 9 columns.
Data Relevance	• This data set is very relevant to determining patterns/trends in birth rates across different states and education levels. We can also determine if average birth weight or average age of mother differs among states. The data set: • Provides geographic and annual birth statistics • Collected by a government entity, CDC

Limitations	For privacy concerns all statistics representing zero to nine births are suppressed
Ethics	• The CDC considers data privacy laws by suppressing births 0-9. PII is safeguarded and not shown in the data set. • The CDC considers data transparency by clearly stating how information was collected and stored. • Potentially the data set is not representative of everyone as it only looks at US residents there may be births in the US of non-residents.

Data Profile

Name of Variable	Description	Time Variable	Structure	Qualitative/ Quantitative	Data Type
state	Residential state of mother	Time-invariant	structured	qualitative	nominal
State Abbreviation	Residential state of mother abbreviated	Time-invariant	structured	qualitative	nominal
Year	Year of birth	Time-invariant	structured	qualitative	ordinal
Gender	Gender of baby	Time-invariant	structured	qualitative	binary
Education Level of Mother	Education level of mother	Time-invariant	structured	qualitative	ordinal
Education Level Code	Education level of mother coded	Time-invariant	structured	qualitative	ordinal
Number of births	Number of births born	Time-invariant	structured	quantitative	discrete
Average Age of Mother (years)	Average age of mother	Time-invariant	structured	quantitative	continuous
Average Birth Weight (g)	Average birth weight	Time-invariant	structured	quantitative	continuous

- *I would argue that all these variables are time invariant as they happened in the past and will not change. For instance, the average age of mother or the average birth weight in 2014 will not change.*

Data Cleaning/ Wrangling

- There are 5,496 rows and 9 columns.

- There were no missing values. Note that in education level there are values with unknown or not stated but this does not impact the analysis, so leaving as is. It appears that any suppressed data (representing zero to nine births) was potentially already removed from the data as there are no missing values.
- There were no duplicate values.
- Updated column names (shortened/clarified a few).
- Did not delete any columns, as I believe all are relevant.

Questions

- Which states have the highest number of births? Which states have the lowest number of births?
- How does level of education impact number of births?
- How do births vary year to year?
- Does age of mother impact number of births?
- Is there any correlation between average age of mother and birth weight?