The dataset shows the number of births in the United States between 1985 and 2015. Each row in the original dataset represents the number of births given in a county over a particular month of the year. The dataset was obtained from Kaggle.com.

Dictionary

Variable	Variable name	Measurement unit	Allowed values	Description
Index	Row index	Numeric - Integers	1-999999	The primary keys to identify each row
State	US State	Numeric - Integers	1-51	The numerical order of the state of residence of the mother
County	County	Numeric - Integers	>= 1001	The county of residence of the mother (FIPS County Code)
Month	Month of birth	Numeric - Integers	1-12	The month of the birth record
Year	Year of birth	Numeric - Integers	1985-2015	The year of the birth record
countyBirths	County births	Numeric - Integers	Integer	The aggregated number of births of all mothers living in a county for a particular month
stateBirths	State births	Numeric - Integers	Integer	The aggregated number of births for a state for a month-year combination

Data cleaning

For some reason, the data set includes records in which the states are 53, 54, 55, and 56. Those records account for 1.44% of the data set (4 620 out of 321 475 rows) and were removed.

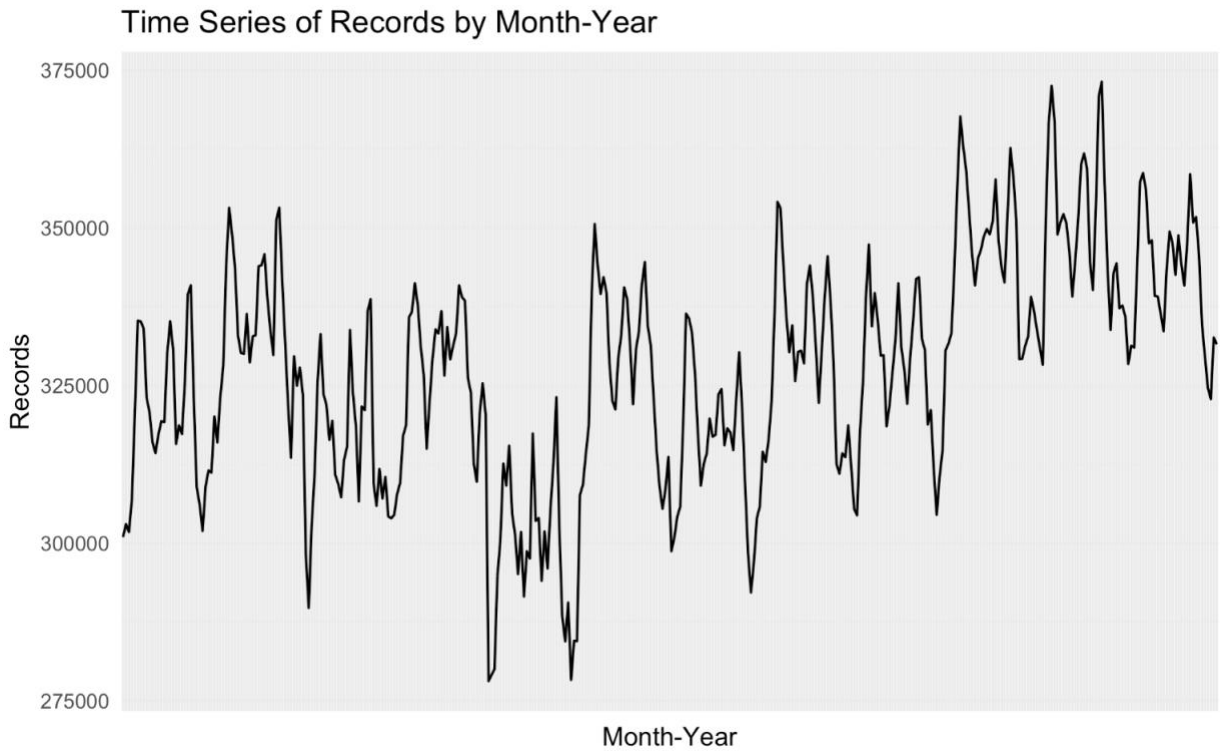
Data aggregation

The division of the number of births among different counties does not allow time series forecasts. Therefore, the aggregation of birth records among all counties for any specific month-year combination was executed, resulting in a new dataset of 372 rows, described in the following dictionary.

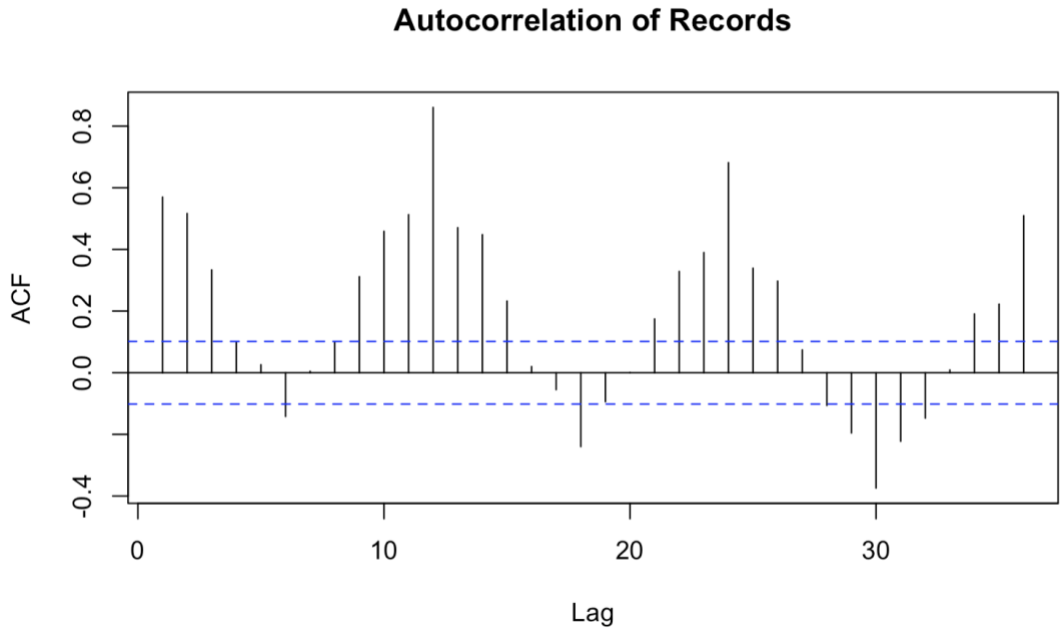
Variable	Variable name	Measurement unit	Allowed values	Description
MonthYear	Month and year of birth	mm-yyyy	mm: 1-12 yyyy: 1985-9999	The month of the year in which the birth record was collected
Records	Birth records	Numeric - Integers	Integer	The total of birth records for a month-year combination of the US

A time series was plotted for the dataset. The ACF plot indicates the birth data has a periodic characteristic.

Time series



ACF plot



Data collection methodology

The dataset was obtained from "Kaggle.com" and was collected by the National Bureau for Economic Research and the National Center for Health Statistics. The organization collected the data to understand birth patterns in the US in order to make appropriate decisions regarding individual health. The website from which the data was downloaded did not indicate the data collection methodology of this dataset. However, regarding birth records, the United States government gathers birth data through two main sources: (1) the National Center for Health Statistics (NCHS) and state vital records offices. The values suggest the data was gathered monthly for each county in the US.

Purpose for choosing the dataset

I want to gain insights into the population growth pattern of the US, specifically the monthly birth growth. It is interesting to explore how the number of births per month in the US changed over the examined period and to figure out which months or periods of the year have the higher numbers of newborn babies. Also, the data is until 2015, which means forecasts can be executed for 2016-2024. The actual data until 2024 will help assess the forecasting accuracy besides the residuals.