US Births (2016-2018)

Google Data Analytics Capstone

Gregory Van

2022-05-18

Introduction (Ask)

For the capstone project of Coursera's Google Data Analytics program, I decided to choose my own case study to showcase the complete steps of the data analysis process which are ask, prepare, process, analyze, share and act. While searching for a dataset for the project, I decided to choose something that was personal to me. My wife and I are expecting our first child later this year so I decided to choose a dataset relating to births in the United States. I also work with an obstetric ultrasound clinic to 3D print ultrasound models of babies so I want to discover trends and information around this field. Although I ultimately chose this project for my own personal interest, it can also support a business case.

Business Task

In this analysis, I will answer a few questions regarding births in the US and identify any trends among demographics or geographic location. These questions can potentially help organizations make business decisions such as who to target in a baby marketing ad campaign, where to open a obstetrics ultrasound clinic, or estimating the future market size of maternity wear by looking at trends in birthrates.

Questions that I will be looking to answer:

1) Which states have the most number of birth from 2016 to 2018? Which states have the highest birthrates?
2) What is the average age of mothers giving birth per state?
3) What is the average age of mothers giving birth by mother's ethnicity?
4) Over the three years, has birthrates decreased or increased?

To answer these questions, I used public data collected from the CDC and the United States Census Bureau.

Data Source (Prepare)

The primary data source I used was obtained from Google's BigQuery public datasets. The data is a subset of a larger data source from the Centers for Disease Control (CDC) which collects and compiles yearly data for live births in the United States. The data is collected from birth certificates which are required by law to be completed for all births. The dataset from BigQuery used in this analysis is called "county_natality_by_mother_race" and is part of "sdoh_cdc_wonder_natality". This dataset includes aggregated data about mothers and births by county in the United States from 2016 to 2018.

The dataset can be found here: Births Data Summary

The dataset has 12 columns and 8843 rows. To load the data into RStudio, I first downloaded the dataset from BigQuery using the export function. Then I uploaded the csv file into RStudio Cloud.

```
birth_data <- read.csv("natality_mother_race.csv")
```

Sample view of the data:

```
head(birth_data)
```

```
##         Year County_of_Residence County_of_Residence_FIPS Mothers_Single_Race
## 1 2018-01-01  Baldwin County, AL                     1003               Asian
## 2 2018-01-01  Calhoun County, AL                     1015               Asian
## 3 2018-01-01  Cochise County, AZ                     4003               Asian
## 4 2018-01-01     Yuma County, AZ                     4027               Asian
## 5 2018-01-01    Butte County, CA                     6007               Asian
## 6 2018-01-01   Madera County, CA                     6039               Asian
##   Mothers_Single_Race_Code Births Ave_Age_of_Mother Ave_OE_Gestational_Age_Wks
## 1                        A     36             30.11                       37.11
## 2                        A     15             28.33                       38.93
## 3                        A     29             31.86                       37.66
## 4                        A     29             32.31                       38.93
## 5                        A    200             29.13                       38.88
## 6                        A     58             31.95                       37.81
##   Ave_LMP_Gestational_Age_Wks Ave_Birth_Weight_gms Ave_Pre_pregnancy_BMI
## 1                       37.25              2991.81                 24.61
## 2                       39.07              3437.53                 25.59
## 3                       37.14              2934.03                 24.37
## 4                       39.28              3244.72                 24.62
## 5                       38.95              3259.41                 26.30
## 6                       37.83              3011.53                 25.93
##   Ave_Number_of_Prenatal_Wks
## 1                      10.42
## 2                       9.80
## 3                       7.96
## 4                      10.68
## 5                      10.70
## 6                      11.52
```

View attributes:

```
colnames(birth_data)
```

```
##  [1] "Year"                         "County_of_Residence"
##  [3] "County_of_Residence_FIPS"     "Mothers_Single_Race"
##  [5] "Mothers_Single_Race_Code"     "Births"
##  [7] "Ave_Age_of_Mother"            "Ave_OE_Gestational_Age_Wks"
##  [9] "Ave_LMP_Gestational_Age_Wks" "Ave_Birth_Weight_gms"
## [11] "Ave_Pre_pregnancy_BMI"        "Ave_Number_of_Prenatal_Wks"
```

The second data source I used was from the United States Census Bureau (data.census.gov). I was able to query the data I needed to a csv file directly from their database. I used this data source to get population data for each state from 2016 to 2018. This data will be used in the birthrate calculations.

```
population_data <- read.csv("state_populations.csv")
head(population_data)
```

```
##   STATE             NAME ABBREVIATION POPESTIMATE2016 POPESTIMATE2017
## 1     0    United States           US       322941311       324985539
## 2     0 Northeast Region           NR        56042330        56059240
## 3     0   Midwest Region           MR        67987540        68126781
## 4     0     South Region           SR       122351760       123542189
## 5     0      West Region           WR        76559681        77257329
```

```
## 6     1          Alabama           AL         4863525         4874486
##   POPESTIMATE2018
## 1       326687501
## 2        56046620
## 3        68236628
## 4       124569433
## 5        77834820
## 6         4887681
```

Data Manipulation (Process)

```
install.packages("tidyverse")
```

```
install.packages("usmap")
```

```
install.packages("gridExtra")
```

```
library(tidyverse)
library(stringr)
library(usmap)
library(ggplot2)
library(gridExtra)
```

Create Subset of Data

I will only select columns of interest that will help answer the business questions.

```
birth_data_subset <- subset(birth_data, select=c("Year", "County_of_Residence", "Mothers_Single_Race",
head(birth_data_subset)
```

```
##         Year County_of_Residence Mothers_Single_Race Ave_Age_of_Mother Births
## 1 2018-01-01  Baldwin County, AL               Asian             30.11     36
## 2 2018-01-01  Calhoun County, AL               Asian             28.33     15
## 3 2018-01-01  Cochise County, AZ               Asian             31.86     29
## 4 2018-01-01     Yuma County, AZ               Asian             32.31     29
## 5 2018-01-01    Butte County, CA               Asian             29.13    200
## 6 2018-01-01   Madera County, CA               Asian             31.95     58
```

Formatting

All the dates show up as the 1st of January but the column header indicates it should only be year. Therefore, I will adjust the dates to display the year only.

```
birth_data_subset$Year <- format(as.Date(birth_data_subset$Year, format = "%Y-%m-%d"), "%Y")
head(birth_data_subset)
```

```
##   Year County_of_Residence Mothers_Single_Race Ave_Age_of_Mother Births
## 1 2018  Baldwin County, AL               Asian             30.11     36
## 2 2018  Calhoun County, AL               Asian             28.33     15
## 3 2018  Cochise County, AZ               Asian             31.86     29
## 4 2018     Yuma County, AZ               Asian             32.31     29
## 5 2018    Butte County, CA               Asian             29.13    200
## 6 2018   Madera County, CA               Asian             31.95     58
```

For this analysis, my main interest is the state of residence thus, I will remove the county from the residence attribute.

```
birth_data_subset$County_of_Residence = str_sub(birth_data_subset$County_of_Residence,-2)
colnames(birth_data_subset)[2] <- "State"
head(birth_data_subset)
```

```
##   Year State Mothers_Single_Race Ave_Age_of_Mother Births
## 1 2018    AL               Asian             30.11     36
## 2 2018    AL               Asian             28.33     15
## 3 2018    AZ               Asian             31.86     29
## 4 2018    AZ               Asian             32.31     29
## 5 2018    CA               Asian             29.13    200
## 6 2018    CA               Asian             31.95     58
```

Data Validation

Now I will look through the data to examine it and clean up the data if necessary. First, I will check if there are any null values.

```
sum(is.na(birth_data_subset))
```

```
## [1] 0
```

The dataset does not have null values. Next, I will check the dataset for errors or inconsistencies. I will check the number of unique states in our dataset.

```
length(unique(birth_data_subset$State))
```

```
## [1] 51
```

I was expecting 50 values since there are 50 states. I will look into the values further to see why there are 51 unique values.

```
sort(unique(birth_data_subset$State))
```

```
##  [1] "AK" "AL" "AR" "AZ" "CA" "CO" "CT" "DC" "DE" "FL" "GA" "HI" "IA" "ID" "IL"
## [16] "IN" "KS" "KY" "LA" "MA" "MD" "ME" "MI" "MN" "MO" "MS" "MT" "NC" "ND" "NE"
## [31] "NH" "NJ" "NM" "NV" "NY" "OH" "OK" "OR" "PA" "RI" "SC" "SD" "TN" "TX" "UT"
## [46] "VA" "VT" "WA" "WI" "WV" "WY"
```

It appears "DC" is in included in our dataset which is why there are 51 values. Next, I will look into mother's race for inconsistencies.

```
unique(birth_data_subset$Mothers_Single_Race)
```

```
## [1] "Asian"
## [2] "White"
## [3] "More than one race"
## [4] "Black or African American"
## [5] "American Indian or Alaska Native"
## [6] "Native Hawaiian or Other Pacific Islander"
```

The categories shown above appear acceptable. Now I will look at the quantitative values to check for errors by seeing if there are any extreme outliers.

```
summary(birth_data_subset$Ave_Age_of_Mother)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.00   26.92   28.31   28.49   30.11   35.60
```

```
summary(birth_data_subset$Births)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    10.0    42.5   167.0  1309.8  1120.0 89459.0
```

The dataset does show any obvious errors from the summary.

For the population_data dataset, I will remove the rows that are neither states or DC. This will help with joining the dataset later.

```
population_data = subset(population_data, STATE != 0)
head(population_data)
```

```
##    STATE       NAME ABBREVIATION POPESTIMATE2016 POPESTIMATE2017
## 6      1    Alabama           AL         4863525         4874486
## 7      2     Alaska           AK          741456          739700
## 8      4    Arizona           AZ         6941072         7044008
## 9      5   Arkansas           AR         2989918         3001345
## 10     6 California           CA        39167117        39358497
## 11     8   Colorado           CO         5539215         5611885
##    POPESTIMATE2018
## 6          4887681
## 7           735139
## 8          7158024
## 9          3009733
## 10        39461588
## 11         5691287
```

I will now move on to the analysis section and look for answers to the questions outlined in the business task.

Analysis (Analyze)

Q1: Birth Totals and Birthrates

Total births in the United States from 2016 to 2018.

```
total_births <- birth_data_subset %>% group_by(State)  %>%
                    summarise(Births= sum(Births))
total_births[order(total_births$Births, decreasing = TRUE),]
```

```
## # A tibble: 51 x 2
##    State  Births
##    <chr>   <int>
##  1 CA    1415191
##  2 TX    1158130
##  3 NY     689884
##  4 FL     669663
##  5 IL     448330
##  6 PA     412253
##  7 OH     409714
##  8 GA     385107
##  9 NC     359383
## 10 MI     334425
## # ... with 41 more rows
```

As expected, California had the highest number of births in the country since they are the most populous state. However, what would be more useful is seeing the birthrates per capita. To do this, first I will get the birthrates for each state for each year. Then I will average the yearly birthrates for each state.

```r
state_population <- subset(population_data, select=c("ABBREVIATION", "POPESTIMATE2016", "POPESTIMATE201
colnames(state_population) <- c("State", "Pop_2016", "Pop_2017", "Pop_2018")
state_population
```

```
##    State Pop_2016 Pop_2017 Pop_2018
## 6     AL  4863525  4874486  4887681
## 7     AK   741456   739700   735139
## 8     AZ  6941072  7044008  7158024
## 9     AR  2989918  3001345  3009733
## 10    CA 39167117 39358497 39461588
## 11    CO  5539215  5611885  5691287
## 12    CT  3578141  3573297  3571520
## 13    DE   948921   956823   965479
## 14    DC   685815   694906   701547
## 15    FL 20613477 20963613 21244317
## 16    GA 10301890 10410330 10511131
## 17    HI  1427559  1424393  1420593
## 18    ID  1682380  1717715  1750536
## 19    IL 12820527 12778828 12723071
## 20    IN  6634304  6658078  6695497
## 21    IA  3131371  3141550  3148618
## 22    KS  2910844  2908718  2911359
## 23    KY  4438182  4452268  4461153
## 24    LA  4678135  4670560  4659690
## 25    ME  1331317  1334612  1339057
## 26    MD  6003323  6023868  6035802
## 27    MA  6823608  6859789  6882635
## 28    MI  9950571  9973114  9984072
## 29    MN  5522744  5566230  5606249
## 30    MS  2987938  2988510  2981020
## 31    MO  6087135  6106670  6121623
## 32    MT  1040859  1052482  1060665
## 33    NE  1905616  1915947  1925614
## 34    NV  2917563  2969905  3027341
## 35    NH  1342307  1348787  1353465
## 36    NJ  8870827  8885525  8886025
## 37    NM  2091630  2091784  2092741
## 38    NY 19633428 19589572 19530351
## 39    NC 10154788 10268233 10381615
## 40    ND   754434   754942   758080
## 41    OH 11634370 11659650 11676341
## 42    OK  3926331  3931316  3940235
## 43    OR  4089976  4143625  4181886
## 44    PA 12782275 12787641 12800922
## 45    RI  1056770  1055673  1058287
## 46    SC  4957968  5021268  5084156
## 47    SD   862996   872868   878698
## 48    TN  6646010  6708799  6771631
## 49    TX 27914410 28295273 28628666
## 50    UT  3041868  3101042  3153550
## 51    VT   623657   624344   624358
## 52    VA  8410106  8463587  8501286
## 53    WA  7294771  7423362  7523869
## 54    WV  1831023  1817004  1804291
```

```
## 55    WI  5772628  5790186  5807406
## 56    WY   584215   578931   577601
```

```
state_births <- birth_data_subset %>% group_by(Year, State)  %>%
                  summarise(Births = sum(Births))
# Need to make the data from long to wide
state_births = spread(state_births, Year, Births)
colnames(state_births) <- c("State", "Birth_2016", "Birth_2017", "Birth_2018")
state_births
```

```
## # A tibble: 51 x 4
##    State Birth_2016 Birth_2017 Birth_2018
##    <chr>      <int>      <int>      <int>
##  1 AK         11209      10445      10086
##  2 AL         56813      58883      57666
##  3 AR         38250      37507      37007
##  4 AZ         84495      81828      80684
##  5 CA        488751     471587     454853
##  6 CO         66576      64357      62823
##  7 CT         35987      35184      34697
##  8 DC          9858       9560       9212
##  9 DE         10961      10841      10595
## 10 FL        224815     223468     221380
## # ... with 41 more rows
```

```
birthrate = merge(state_births, state_population, by="State")
birthrate$Ave_Birthrateper1000 = round((1000 * (birthrate$Birth_2016 / birthrate$Pop_2016 + birthrate$Bi
birthrate[order(birthrate$Ave_Birthrateper1000, decreasing = TRUE),]
```

```
##    State Birth_2016 Birth_2017 Birth_2018 Pop_2016 Pop_2017 Pop_2018
## 45    UT      50464      48585      47209  3041868  3101042  3153550
## 29    ND      11379      10734      10630   754434   754942   758080
## 1     AK      11209      10445      10086   741456   739700   735139
## 42    SD      12262      12122      11890   862996   872868   878698
## 8     DC       9858       9560       9212   685815   694906   701547
## 44    TX     397841     381862     378427 27914410 28295273 28628666
## 30    NE      26577      25804      25467  1905616  1915947  1925614
## 19    LA      63098      60940      59527  4678135  4670560  4659690
## 37    OK      52586      50205      49794  3926331  3931316  3940235
## 14    ID      22450      22161      21388  1682380  1717715  1750536
## 17    KS      38034      36503      36228  2910844  2908718  2911359
## 3     AR      38250      37507      37007  2989918  3001345  3009733
## 26    MS      37887      37317      36956  2987938  2988510  2981020
## 16    IN      83007      82079      81569  6634304  6658078  6695497
## 11    GA     129925     129116     126066 10301890 10410330 10511131
## 24    MN      69712      68573      67311  5522744  5566230  5606249
## 18    KY      55408      54716      53899  4438182  4452268  4461153
## 12    HI      18034      17497      16943  1427559  1424393  1420593
## 13    IA      39369      38393      37768  3131371  3141550  3148618
## 34    NV      36260      35748      35674  2917563  2969905  3027341
## 25    MO      74642      72993      73212  6087135  6106670  6121623
## 43    TN      80768      80967      80715  6646010  6708799  6771631
## 5     CA     488751     471587     454853 39167117 39358497 39461588
## 51    WY       7379       6894       6562   584215   578931   577601
## 21    MD      73048      71575      71010  6003323  6023868  6035802
```

```
## 46    VA     102374     100305      99760  8410106  8463587  8501286
## 48    WA      90489      87557      86072  7294771  7423362  7523869
## 2     AL      56813      58883      57666  4863525  4874486  4887681
## 35    NY     234148     229592     226144 19633428 19589572 19530351
## 36    OH     137993     136707     135014 11634370 11659650 11676341
## 15    IL     154337     149269     144724 12820527 12778828 12723071
## 4     AZ      84495      81828      80684  6941072  7044008  7158024
## 28    NC     120610     119984     118789 10154788 10268233 10381615
## 6     CO      66576      64357      62823  5539215  5611885  5691287
## 32    NJ     102550     101131     101103  8870827  8885525  8886025
## 33    NM      24669      23737      23012  2091630  2091784  2092741
## 41    SC      57258      56935      56550  4957968  5021268  5084156
## 9     DE      10961      10841      10595   948921   956823   965479
## 27    MT      12271      11791      11504  1040859  1052482  1060665
## 49    WI      66572      64905      63997  5772628  5790186  5807406
## 23    MI     113202     111304     109919  9950571  9973114  9984072
## 39    PA     139211     137553     135489 12782275 12787641 12800922
## 10    FL     224815     223468     221380 20613477 20963613 21244317
## 38    OR      45504      43599      42168  4089976  4143625  4181886
## 50    WV      19065      18669      18240  1831023  1817004  1804291
## 20    MA      71258      70632      69055  6823608  6859789  6882635
## 40    RI      10780      10620      10486  1056770  1055673  1058287
## 7     CT      35987      35184      34697  3578141  3573297  3571520
## 22    ME      12672      12266      12271  1331317  1334612  1339057
## 47    VT       5752       5651       5430   623657   624344   624358
## 31    NH      12251      12089      11972  1342307  1348787  1353465
##    Ave_Birthrateper1000
## 45                15.742
## 29                14.441
## 1                 14.319
## 42                13.876
## 8                 13.754
## 44                13.655
## 30                13.547
## 19                13.103
## 37                12.934
## 14                12.821
## 17                12.687
## 3                 12.528
## 26                12.521
## 16                12.341
## 11                12.336
## 24                12.316
## 18                12.285
## 12                12.281
## 13                12.263
## 34                12.083
## 25                12.058
## 43                12.047
## 5                 11.996
## 51                11.967
## 21                11.938
## 46                11.920
## 48                11.880
```

```
## 2                11.853
## 35               11.742
## 36               11.716
## 15               11.698
## 4                11.687
## 28               11.668
## 6                11.508
## 32               11.440
## 33               11.379
## 41               11.337
## 9                11.285
## 27               11.279
## 49               11.254
## 23               11.182
## 39               10.744
## 10               10.662
## 38               10.577
## 50               10.265
## 20               10.258
## 40               10.056
## 7                 9.873
## 22                9.291
## 47                8.990
## 31                8.978
```

Q2: Mothers Average Age by State

Now I will find the average age of mothers giving birth per state. To do this, I cannot simply group by state and then take the mean of the mother's age because the value of the mother's age is an average of the county of residence and each county has a different number of total births. It is more accurate to take a weighted average instead by taking the sum of births times the mother's average age and then divide by the total number of births. For instance, if county A has 30 births and the average age is 30 and county B has 15 births and the average age is 24, the average age for these two counties should show as 28 and not 27.

```
ave_age <- birth_data_subset %>% group_by(State)  %>%
                    summarise(Average_Age = round(sum(Births*Ave_Age_of_Mother)/sum(Births), digits = 2)
ave_age[order(ave_age$Average_Age),]
```

```
## # A tibble: 51 x 2
##    State Average_Age
##    <chr>       <dbl>
##  1 MS           26.8
##  2 AR           26.9
##  3 WV           27.0
##  4 OK           27.3
##  5 AL           27.3
##  6 KY           27.3
##  7 LA           27.4
##  8 NM           27.5
##  9 TN           27.7
## 10 IN           27.8
## # ... with 41 more rows
```

Q3: Mother's Average Age by Race

The dataset has women identifing as one of 6 possible ethnicity groups.

```
ave_age_race <- birth_data_subset %>% group_by(Mothers_Single_Race)  %>%
                    summarise(Average_Age = round(sum(Births*Ave_Age_of_Mother)/sum(Births), digits = 2)
ave_age_race[order(ave_age_race$Average_Age),]
```

```
## # A tibble: 6 x 2
##   Mothers_Single_Race                       Average_Age
##   <chr>                                           <dbl>
## 1 American Indian or Alaska Native                 26.9
## 2 More than one race                               27.4
## 3 Black or African American                        27.7
## 4 Native Hawaiian or Other Pacific Islander        27.8
## 5 White                                            28.9
## 6 Asian                                            31.5
```

Q4: Trajectory of Birthrates from 2016 to 2018

To understand the birthrate trajectories, I have to find the birthrates for the entire US for each year of 2016, 2017 and 2018. Then I can see if it increases or decreases.

```
yearly_birthrates <- birth_data_subset %>% group_by(Year)  %>%
                    summarise(Births= sum(Births))
yearly_birthrates$Population <- c(sum(population_data$POPESTIMATE2016), sum(population_data$POPESTIMATE
yearly_birthrates$Birthrate <- 1000*yearly_birthrates$Births/yearly_birthrates$Population
yearly_birthrates
```

```
## # A tibble: 3 x 4
##   Year   Births Population Birthrate
##   <chr>   <int>      <int>     <dbl>
## 1 2016  3940811  322941311      12.2
## 2 2017  3852740  324985539      11.9
## 3 2018  3788947  326687501      11.6
```

Results (Share)

Q1: Map of Birth Totals and Birthrates

From the results of the data, California had the highest births between 2016 and 2018. This was followed by Texas. The following heat map visually shows that states with higher populations had the higher number of births which is not of surprise.

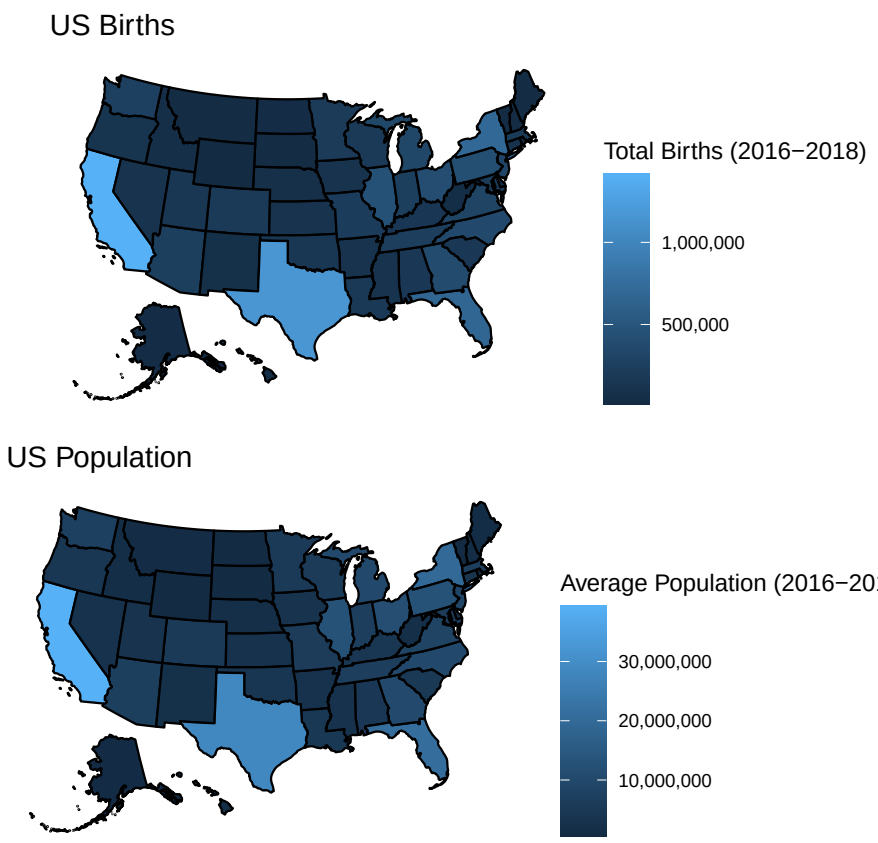
```
colnames(total_births)[1] <- "state"
colnames(birthrate)[1] <- "state"

birthmap <- plot_usmap(data = total_births, values = "Births", color = "black") +
  scale_fill_continuous(name = "Total Births (2016-2018)", label = scales::comma) +
  theme(legend.position = "right") + ggtitle("US Births")

birthrate$Ave_Population = round(((birthrate$Pop_2016 + birthrate$Pop_2017 + birthrate$Pop_2018) / 3),

popmap <- plot_usmap(data = birthrate, values = "Ave_Population", color = "black") +
  scale_fill_continuous(name = "Average Population (2016-2018)", label = scales::comma) +
  theme(legend.position = "right") + ggtitle("US Population")
```

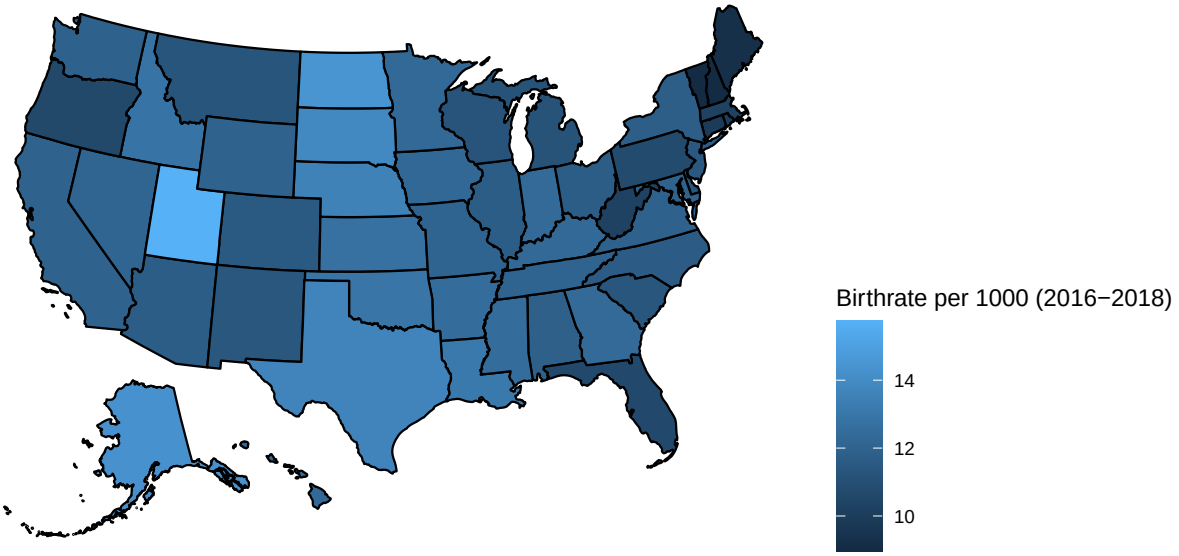
```
grid.arrange(
  birthmap,
  popmap,
  nrow = 2
  )
```



Plotting the map with birthrate data shows which states have the highest births per capita. Utah, North Dakota, Alaska, and South Dakota top the list here. Many of these states are located in the central part of the US. New Hampshire, Vermont, Maine, Connecticut and Rhode Island are at the bottom of the list. The northeast in general tends to have lower birthrates.

```
plot_usmap(data = birthrate, values = "Ave_Birthrateper1000", color = "black") +
  scale_fill_continuous(name = "Birthrate per 1000 (2016-2018)", label = scales::comma) +
  theme(legend.position = "right") + ggtitle("US Population")
```

US Population



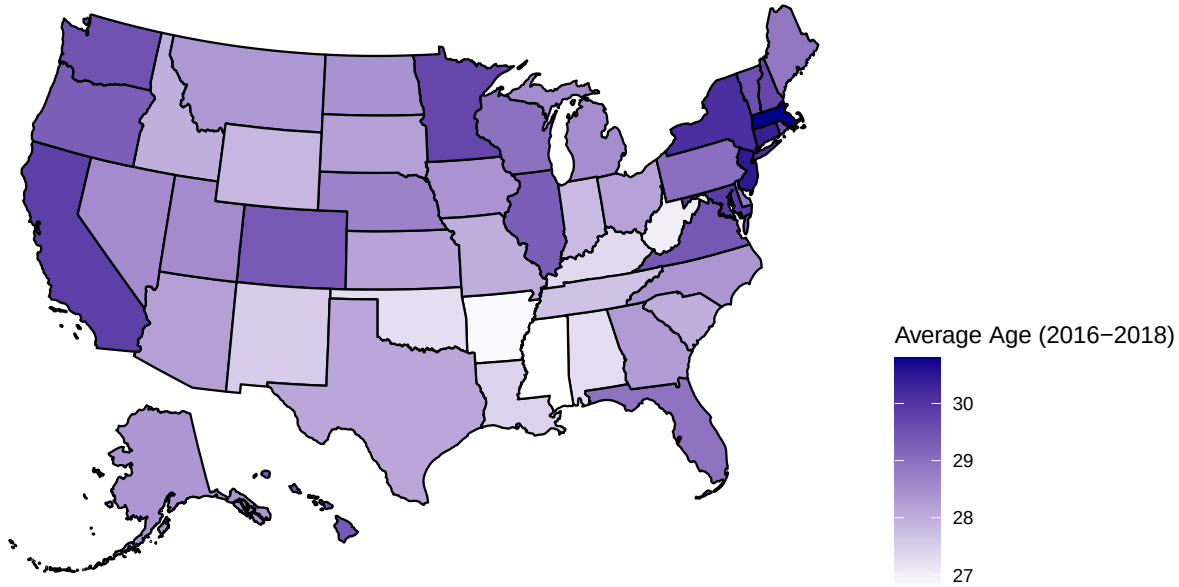
Q2: Map of Mothers Average Age by State

Now we will look at the average age of mothers at the time of birth by states. The plot shows that in states in the northeast, women tend to be a bit older when giving birth. Massachusetts, DC, New Jersey, Connecticut, New York and Maryland have the oldest average age. The states with the youngest average age are Mississippi, Arkansas, West Virginia and Oklahoma.

```
colnames(ave_age)[1] <- "state"

plot_usmap(data = ave_age, values = "Average_Age", color = "black") +
  scale_fill_continuous(low = "white", high = "darkblue", name = "Average Age (2016-2018)", label = scal
  theme(legend.position = "right") + ggtitle("Average US Mother's Age Giving Birth")
```

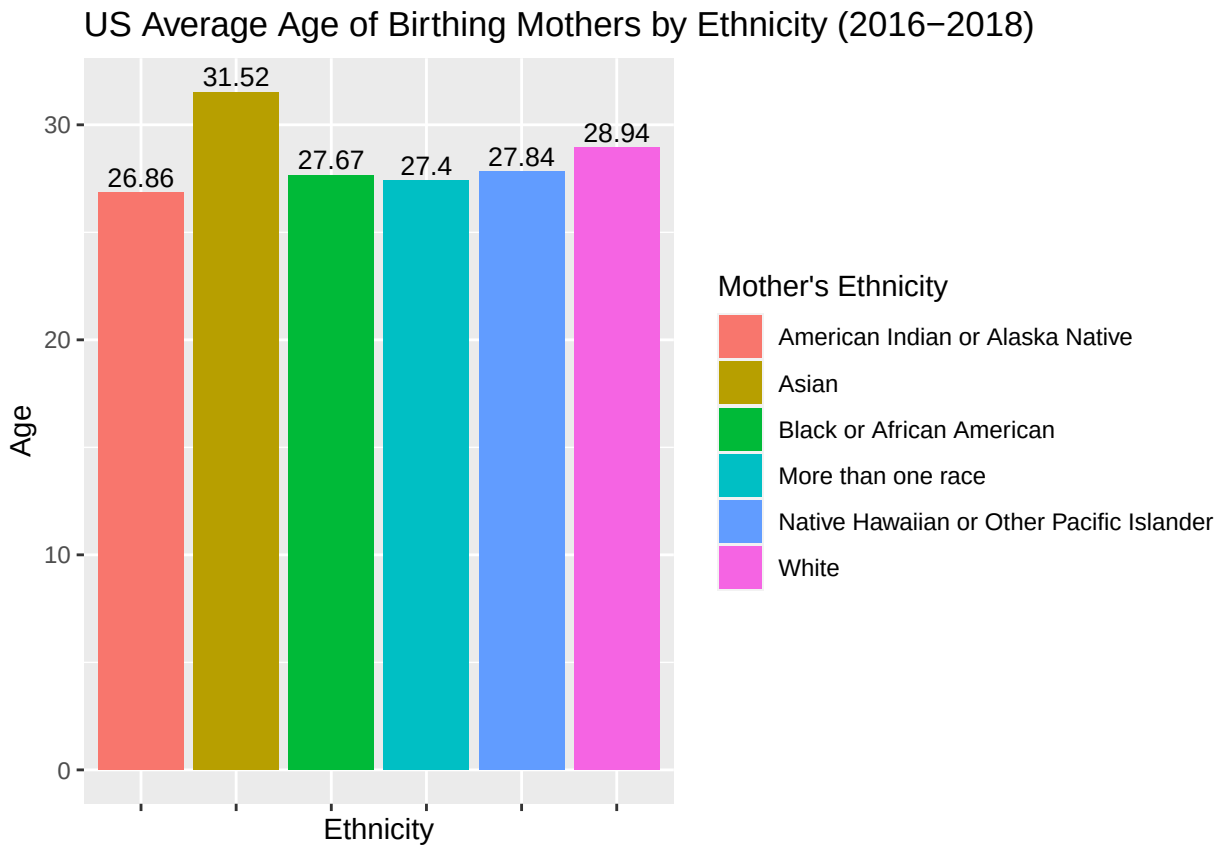
Average US Mother's Age Giving Birth



Q3: Chart of Mother's Average Age by Race

To compare the mother's age by ethnicity, the bar chart below gives a visual depiction. Asian mothers tend to have babies when they are older. American Indian/Alaska Native mothers tend to have babies at a younger age.

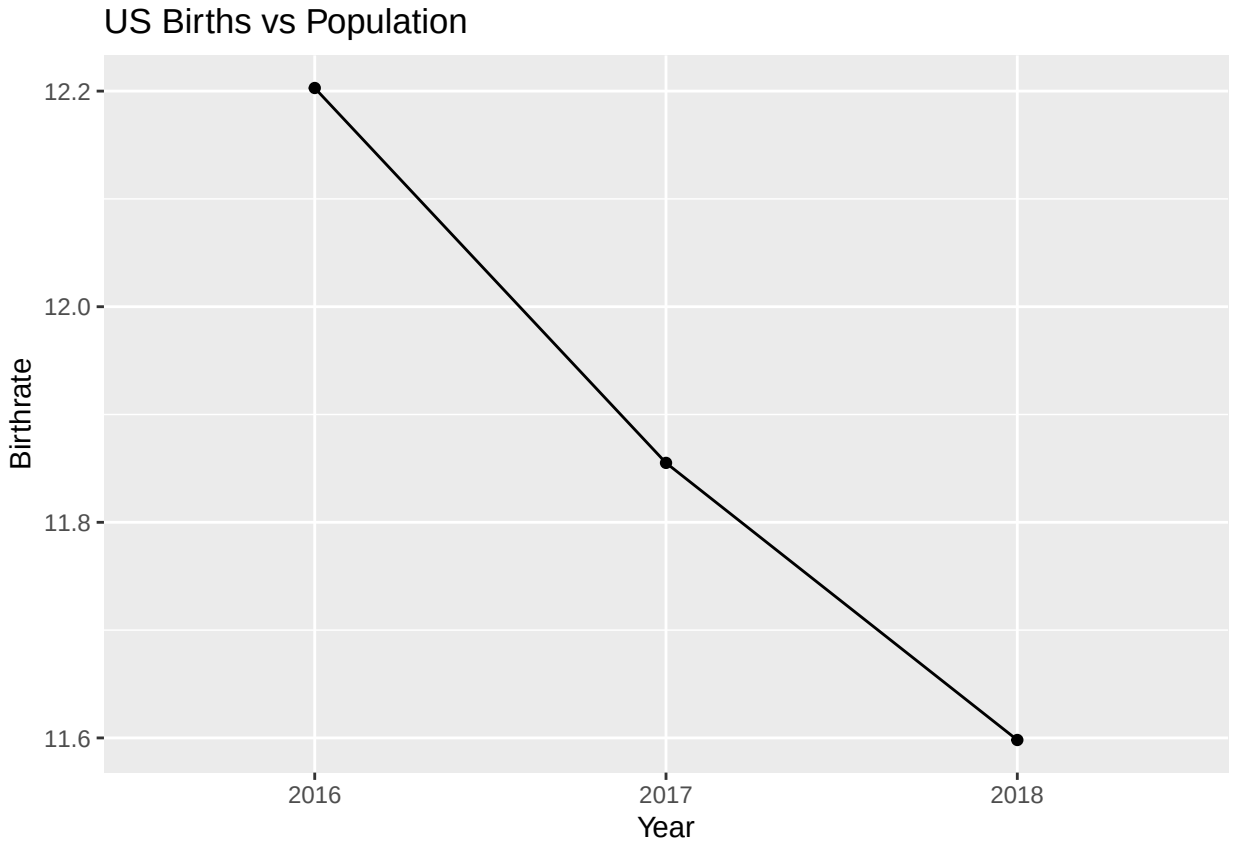
```
ggplot(data=ave_age_race, aes(x=Mothers_Single_Race, y=Average_Age, fill=Mothers_Single_Race)) +
  geom_bar(stat="identity") + geom_text(aes(label=Average_Age), vjust=-0.3, size=3.5) + theme(axis.text
  ggtitle("US Average Age of Birthing Mothers by Ethnicity (2016-2018)") +
  xlab("Ethnicity") + ylab("Age") + labs(fill = "Mother's Ethnicity")
```



Q4: Line Chart of Birthrates from 2016 to 2018

Finally to look at the trends of birth as a whole for the United States, I will plot the yearly birthrates from 2016 to 2018. The chart shows that the birthrates are decreasing as a whole over this time. In fact, the number of births decreased each year while the population grew.

```
ggplot(yearly_birthrates, aes(x=Year, y=Birthrate, group=1)) +
  geom_line() +
  geom_point() +
  ggtitle("US Births vs Population")
```



Conclusion (Act)

From the analysis, for the years 2016-2018, it was found that states with the highest populations had the highest number of births. However, Utah had the highest birthrate per capita and in general, the central region of the United States typically had higher birthrates. The analysis also showed that women in Mississippi and Arkansas tend to have babies at a younger age and states in the northeast like Massachusetts, DC and New Jersey tend to have women of older age giving birth. In addition, American Indian/Alaska Native mothers have babies when they are younger and Asian mothers tend to have babies when they are older. Finally, the data showed that both birthrates and total births have been decreasing from 2016 to 2018 in the United States.

This information can be used to perform marketing analysis. It gives demographics about women who are having babies which in turn can be used for advertisement or to help drive other business decisions. This can also be a good starting point to look further into root causes of why Asian women tend to have babies at a later age than the rest of the population or why birthrates are decreasing.

To improve the analysis, there are a few things that can be looked at but would require additional datasets. For birthrate, instead of including everyone in the population, it may be better to calculate it based on population of just females in a certain age range like from 18-45 years old. Additionally, expanding the dataset to include more years would allow us to make predictions such as estimating populations in the future.