title: “Final Project” author: Courtney Taylor output: pdf_document

description: National Longitudinal Survey of Youth (1979 – 2012) is a longitudinal project that follows a sample of American youth born between 1957-64 on various life aspects from 1979 to 2012. The data set provided below is a subset of this database, focusing on variables of 4 main topics: socioeconomic status, employment, education, and marriage. Some recommended statistical analysis techniques to be applied are multiple regression, time series analysis, logistic regression, and ANOVA.

```
library(readr)
library(ggplot2)
library(tidyverse)
```

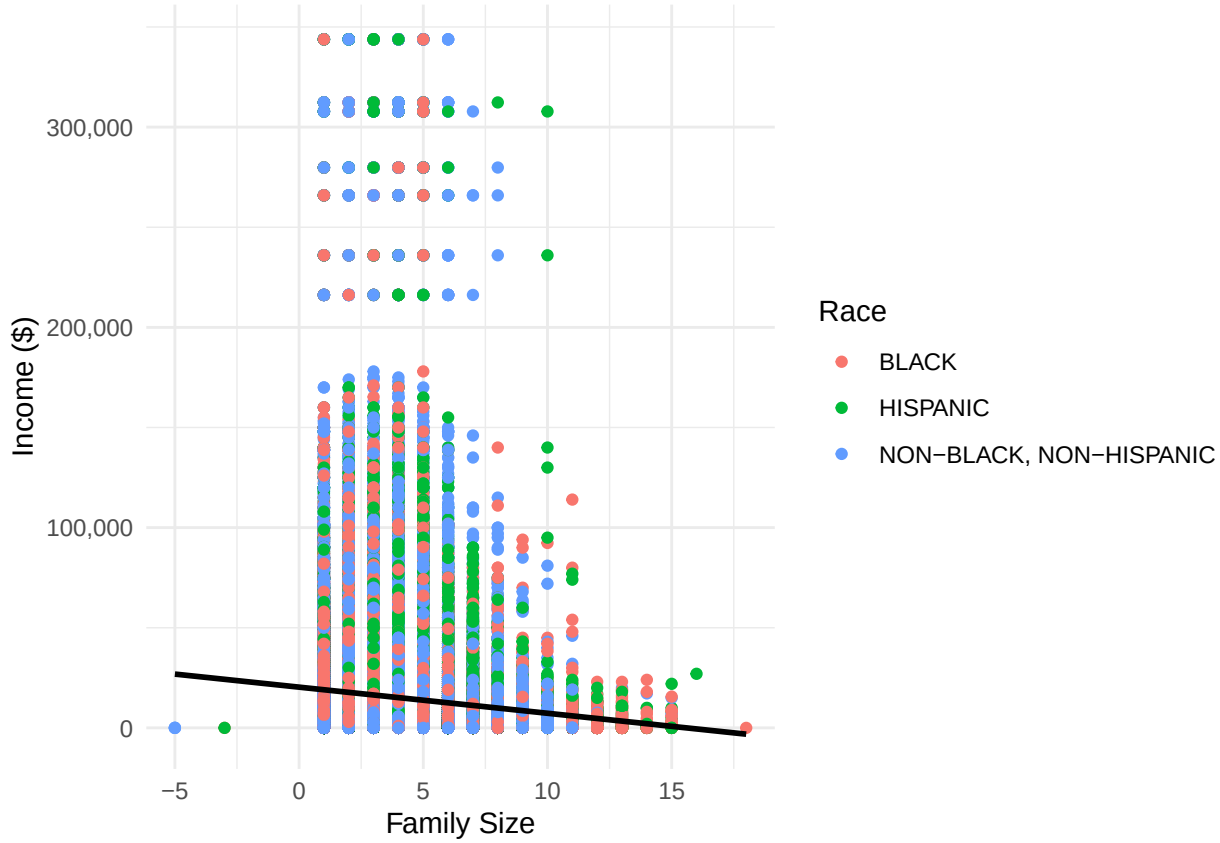
```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v stringr   1.5.0
## v forcats   1.0.0     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dplyr)
youth_dataset_long_form <- read_csv("~/Desktop/Final Project/youth-dataset-long-form.csv")
```

```
## Rows: 243071 Columns: 36
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (18): COUNTRY_OF_BIRTH, SAMPLE_RACE, SAMPLE_SEX, C1DOB_Y, HAVING_HEALTHP...
## dbl (18): ID, YEAR, YEAR_OF_BIRTH, FAMSIZE_, TNFI_, WKSUEMP_PCY_, JOBSNUM_, ...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
library(ggplot2)

ggplot(youth_dataset_long_form, aes(x = FAMSIZE_, y = INCOME_, color = SAMPLE_RACE)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE, color = "black") +
  labs(
    x = "Family Size",
    y = "Income ($)",
    color = "Race"
  ) +
  theme_minimal() +
  scale_y_continuous(labels = scales::comma)
```




```
ggplot
```

```
## function (data = NULL, mapping = aes(), ..., environment = parent.frame())
## {
##     UseMethod("ggplot")
## }
## <bytecode: 0x12cc00c50>
## <environment: namespace:ggplot2>
```

Explanation for Visual: The visual above represents a scatterplot representing the relationship between income and family size, and the points are color coded to the race that each individual is. The regression line is here to show the relationship between the two variables to further predicte the value of the dependent variable for a given value of the independent variable. The results from the scatterplot show that first, most of the individuals within this dataset are below the $200,000 income level. It also shows that while every race is spread throughout the amount of children each family has, black people and hispanic people tend to be the only ones who have more than 11 children within their family. It is also seen that there is a decline in income level in relation to the amount of children a family has. For every child born into a family, the income level will decrease.