1. First of all, I choose Bachelor\_Degree\_Majors from <https://www.kaggle.com/datasets>

“This dataset contains information on the number of students that majored in different topics of study at universities in the United States in 2019 and is broken down by age group, sex, and state.”



I chose this dataset because I am very interested in the data inside this data set.

I wanted to know the relationship between gender and bachelor’s degree, the relationship between gender and major choice, the relationship between age and bachelor’s degree, and the relationship between age and major choice. Even inside this data set, there is information about states, and I can explore what majors are preferred by people in each state. This dataset has too much information for me to analyze.

**Preparing the data**

I downloaded the data set on my computer and opened it in excel. The purpose is to check if there are any gaps or missing data in it.

So, I saved this data set on my computer and opened it with RStudio.

Then, I use

#Validation

any(is.na(degree))

Make sure again that there is no n/a inside.

Graphical user interface

Description automatically generated

Analysis on categorical variable

1. Percentage of male and female in all state over the age of 25 with a bachelor's degree in 2019

Chart, pie chart

Description automatically generated

**male and female in each major choice**

**Chart, bubble chart

Description automatically generated**

1. Mean，mini, max of the number of people above 25 with a bachelor's degree, within all states.

Chart, box and whisker chart

Description automatically generated

1. In mosaicplot compare New York State and California, choose Science and Engineering major and Business major as age above 25.

Chart, treemap chart

Description automatically generated

**random samples**

Chart, histogram

Description automatically generated

Compare their means and sd,

the means of the population and the means of sample mean distributions are close.

As the number of sample size increases, the standard deviation of the

sample mean distribution decreases

Graphical user interface

Description automatically generated with low confidence

**sampling methods**

Text

Description automatically generated

The first one is Srswr, the second one is systematic sampling and the third one is stratified.

I did three sample screens using three methods and I found that because of the method of systematic and Stratified, the proportion of the selected sexes is very close to the original data, therefore, the conclusion should be very similar to the conclusion with all the data

**Implementation of additional feature**

**Chart, bar chart

Description automatically generated**

**Graphical user interface, text

Description automatically generated**