US_STATS_MANN

2024-02-11

Part 1

Introduction

The U.S. Census Bureau is responsible for keeping track of data related to population, the economy, social status, housing, education, and much more. And luckily, since this is government data, it is almost all open-sourced. This makes it a prime target for performing some tests to gain insight into the status of U.S. society. In this project, I will be specifically focusing on education, finance and industry data in order to determine where the country currently stands, where the trends lie, and what kinds of correlations exist in the data.

Imagine a family that might want to move to a certain state or a certain district. They might be interested in the average educational level of the area, average salary ranges, and perhaps the types of jobs that are available there. Or, imagine a city councilperson that is trying to determine what sectors are currently growing, and what sectors are diminishing. This project would help people like them to make informed decisions that impact those around them.

Research Questions

Here are some potential research questions that will be addressed:

- What are the overall average educational levels in the US?
- What areas of the country are more educated? Less?
- What percentage of the country is middle-, upper- and lower-class?
- Where is the size of the middle class the lowest in the country?
- What is the correlation between education level and financial class?
- What percentage of the country works in each type of industry?
- What are the most common industries for men? For women?
- Which parts of the country have the most IT jobs?
- What is the correlation between the percentage of manufacturing jobs and education level?

Approach

The overall approach used in this project will be to import data tables into R, clean the data, and then join the data tables together in order to answer the questions posed above. Averages can be determined through basic R functions, such as `mean()` and `median()`. Maps for states can be drawn using the `maps` and `mapdata` libraries, which offer a way of graphically representing data by a specific region. Correlations can be found through built-in correlation tests, such as `cor.test()`, which contains different methods such as Pearson, Kendall, and Spearman. If we want to make projections for future data, we could also use regression to model the data by working with the `lm()` or `glm()` functions from the `Metrics` library. Each of these approaches will be used for the various research questions posed above. For example, to demonstrate which areas of the country are more and less educated, we will use a heat map overlaid on the map of the US. And to determine the overall average education level, all we need to do is generate the mean of the entire set of data.

How the Approach Addresses the Problem

This approach should be able to quickly give us insights into all of the research questions posed above, as well as give us the proper statistics to answer the overall problem related to how the country currently stands. Once we have answered each of the research questions, we can use the charts, tables, and maps generated in order to explain the bigger picture. We could then use the same approach in order to extrapolate trends that might occur in the future. Because the data set is so diverse, it can be used for many different purposes to address our problem. A politician might need to know about education levels and industries in their district in order to appeal to the highest amount of voters. A business owner might need to know the same information in order to construct a targeted advertising campaign. Since there are many contrasting goals involved with the research problem, maintaining a broad approach is the best method of adapting to the widest possible audience.

Datasets

This project will be using data derived from the U.S. Census, which was then further cleaned and provided to Kaggle here (https://www.kaggle.com/datasets/mittvin/u-s-census-dataset-education-finance-industry). This data was collected by internet, mail, phone, and in-person interviews as a part of the American Community Survey throughout every district in the U.S. The purpose of this survey is to collect and produce information on social, economic, housing, and demographic characteristics about our nation's population every year

There are three separate csv files of data: one for education, one for finance, and one for industry.

Each file contains columns for the year (2019-2021) and the congressional district that the census survey was taken. Congressional district codes are created by attaching the district number to the state code and then separating by an underscore. There is a row entry for each of 437 congressional districts (this includes Washington D.C. and Puerto Rico, even though they are non-voting districts).

The education file contains three additional columns: - Bachelors_degree_or_higher : has a bachelor's degree or higher - high_school_or_some_degree : has a high school diploma or some college - Less_than_high_school_graduate : has not graduated high school

Unfortunately, the education data set is the only one to not contain any information on the year 2019, only 2020-2021.

The finance file contains 11 additional columns that represent the counts of yearly salary for various categories ranging from <$5,000 to >$150,000.

The industry file contains 41 additional columns that are separated into three categories: total, male, and female. Each column gives either the total number of jobs in a specific industry, the total number of men in that industry, or the total number of women in that industry. There are also two columns just labeled 'male' and 'female' that give the total number of men and women overall.

Missing data has already been cleaned, and apart from the lack of 2019 education data, there is no data that is missing from any of the tables.

Required Packages

In this project, we will need to use the following packages:

```
library(dplyr)
library(tidyr)
library(ggplot2)
library(maps)
library(mapdata)
library(Metrics)
library(corrplot)
library(data.table)
```

`dplyr` and `tidyr` will help to group, aggregate, melt, and join data in a more efficient manner. `ggplot2` will be used to plot data. `maps` and `mapdata` will be used to create heat maps of the united states. `Metrics` might be necessary if we need to perform regression analysis. `corrplot` will be used for plotting correlations in aesthetically pleasing charts.

Plots and Table Needs

The main plots that will be necessary in this project will be:

- pie charts for percentages data
- bar charts for categorical data
- state heat maps
- correlation heat maps

Other values can be computed on an individual basis or presented in a simple table.

Future Steps

The main things that I will need to learn/review in order to complete this project are how to make heat maps, how to use state map data, and review how to make similar heat maps for correlation tables. I will also need to review how to join data tables together so that I can perform certain analyses on the data as a whole.

To make a heat map, I will need to import the `ggplot2`, `maps` and `mapdata` libraries, merge the map data with my own data sets, and then plot the data using `geom_polygon()` and `scale_fill_gradient()`. In order to make a heat map of correlation, all I will need to do is use the `corrplot` package to plot the correlation. As for joining data tables together, the `left_join()` function via the `dplyr` package should do the trick.

Part 2

Importing and Cleaning Data

First, we need to import all of the three csv files and turn them into data frames:

```
# import the three csv files and convert them to data frames
edu <- read.csv("education.csv")
fin <- read.csv("finance.csv")
ind <- read.csv("industry.csv")
```

Next, let's convert all the column names to lower case:

```
# convert column names to all lower case
colnames(edu) <- tolower(colnames(edu))
colnames(fin) <- tolower(colnames(fin))
colnames(ind) <- tolower(colnames(ind))
```

In the financial data, the column names initially contained dollar signs, then after conversion via the `read.csv()` function, the dollar signs were converted to `.` and `x`. Here, we use the `gsub()` function to get rid of these unwanted characters:

```
# take out unnecessary characters from the finance data frame
colnames(fin) <- gsub("\\.|x", "", colnames(fin))
```

Next, we need to deal with the district/state codes, which are kept in the 'cd' column. Here, all we need to do is separate the column into the district number and the state abbreviation, with the separator being an underscore:

```
# separate the cd column into two columns for the district number and state code
edu <- separate(edu, cd, into = c("district", "state"), sep = "_", remove = TRUE)
fin <- separate(fin, cd, into = c("district", "state"), sep = "_", remove = TRUE)
ind <- separate(ind, cd, into = c("district", "state"), sep = "_", remove = TRUE)
```

One more thing to keep in mind is that all of the data for 2019 is missing for the education data. Before we attempt to merge all of our data together, we will first fill up all of the education data for 2019 with NA values:

```
# fill in empty data (NA values) into the education df for the year 2019
empty_2019 <- edu %>% distinct(state, district) %>% mutate(year = 2019, bachelors_degree_or_higher = NA, high_sch
ool_or_some_degree = NA, less_than_high_school_graduate = NA)
edu <- rbind(empty_2019, edu)
```

Finally, all we need to do is merge each of these data frames together based on the unique values of year, state, and district:

```
# merge all three data frames together based on year, state and district
efi <- merge(merge(edu, fin, by = c("year", "state", "district")),
                   ind, by = c("year", "state", "district"))
```

What does the final data set look like?

Here are the first few rows over the first 10 columns:

```
tail(efi[1:10])
```

```
##      year state district bachelors_degree_or_higher high_school_or_some_degree
## 1306 2021    WI        6                     181235                      427977
## 1307 2021    WI        7                      79792                      235135
## 1308 2021    WI        8                      98908                      245395
## 1309 2021    WV        1                     101244                      367015
## 1310 2021    WV        2                     124320                      370968
## 1311 2021    WY        0                      81404                      211910
##      less_than_high_school_graduate less_than_5000 5000_to_9999 10000_to_14999
## 1306                          42508           9086         5322          13696
## 1307                          26886           5422         4311          10355
## 1308                          26275           5819         3460           8782
## 1309                          64940          16633        16335          23485
## 1310                          47006          11109        11781          18037
## 1311                          22337           5334         5102           7506
##      15000_to_19999
## 1306          13794
## 1307           9733
## 1308           8836
## 1309          23969
## 1310          17800
## 1311           8046
```

The total dimensions of the data frame are 1311 rows by 58 columns.

```
dim(efi)
```

```
## [1] 1311   58
```

Apart from the year, state and district, all of the values for each column in the data frame are numerical counts of total people.

Here is a list of all of the column names:

```
colnames(efi)
```

```
##  [1] "year"
##  [2] "state"
##  [3] "district"
##  [4] "bachelors_degree_or_higher"
##  [5] "high_school_or_some_degree"
##  [6] "less_than_high_school_graduate"
##  [7] "less_than_5000"
##  [8] "5000_to_9999"
##  [9] "10000_to_14999"
## [10] "15000_to_19999"
## [11] "20000_to_24999"
## [12] "25000_to_34999"
## [13] "35000_to_49999"
## [14] "50000_to_74999"
## [15] "75000_to_99999"
## [16] "100000_to_149999"
## [17] "150000_or_more"
## [18] "total_agriculture_forestry_fishing_hunting_mining"
## [19] "total_construction"
## [20] "total_manufacturing"
## [21] "total_wholesale_trade"
## [22] "total_retail_trade"
## [23] "total_transportation_warehousing_utilities"
## [24] "total_information"
## [25] "total_finance_insurance_realestate_rental_leasing"
## [26] "total_professional_scientific_management_administrative_waste_management_services"
## [27] "total_educationalservices_healthcare_socialassistance"
## [28] "total_arts_entertainment_recreation_accommodation_foodservices"
## [29] "total_otherservices_except_public_administration"
## [30] "total_public_administration"
## [31] "male"
## [32] "male_agriculture_forestry_fishing_hunting_mining"
## [33] "male_construction"
## [34] "male_manufacturing"
## [35] "male_wholesale_trade"
## [36] "male_retail_trade"
## [37] "male_transportation_warehousing_utilities"
## [38] "male_information"
## [39] "male_finance_insurance_realestate_rental_leasing"
## [40] "male_professional_scientific_management_administrative_waste_management_services"
## [41] "male_educationalservices_healthcare_socialassistance"
## [42] "male_arts_entertainment_recreation_accommodation_foodservices"
## [43] "male_otherservices_except_public_administration"
## [44] "male_public_administration"
## [45] "female"
## [46] "female_agriculture_forestry_fishing_hunting_mining"
## [47] "female_construction"
## [48] "female_manufacturing"
## [49] "female_wholesale_trade"
## [50] "female_retail_trade"
## [51] "female_transportation_warehousing_utilities"
## [52] "female_information"
## [53] "female_finance_insurance_realestate_rental_leasing"
## [54] "female_professional_scientific_management_administrative_waste_management_services"
## [55] "female_educationalservices_healthcare_socialassistance"
## [56] "female_arts_entertainment_recreation_accommodation_foodservices"
## [57] "female_otherservices_except_public_administration"
## [58] "female_public_administration"
```

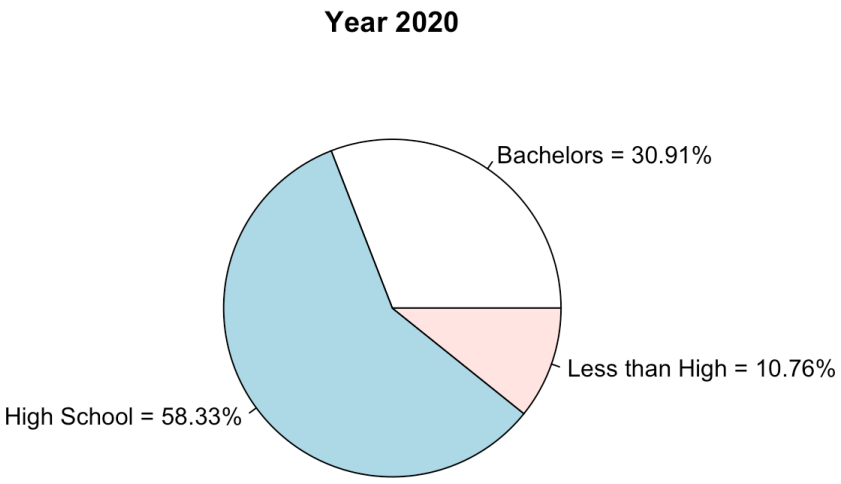
What information is not self-evident?

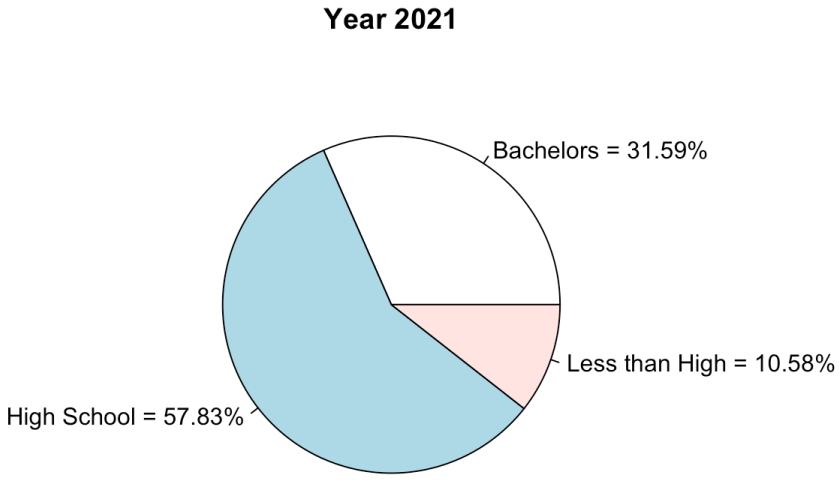
Before we talk about what is not self-evident, let's first address what is self-evident. We can immediately find information on particular districts in each state, such as the total number of people with bachelor's degrees, people with an income over $150,000, and the number of women in public administration. Anything else will require further calculations and groupings.

Let's begin to tackle some of our research questions. This will help demonstrate some of the information that is not already obvious.

One question is: What are the overall average educational levels in the US?

```
# select the educational data from 2020 and 2021
ed_dat <- efi %>% select(year, bachelors_degree_or_higher, high_school_or_some_degree, less_than_high_school_grad
uate) %>% filter(year > 2019)

# calculate the total number of people in the US that fall into each category
tot_vals <- ed_dat %>% group_by(year) %>% summarize(b = sum(bachelors_degree_or_higher), h = sum(high_school_or_s
ome_degree), l = sum(less_than_high_school_graduate))

# iterate by year, generating a pie chart of the percentage of people
# with each education level in the US overall
for (i in 1:nrow(tot_vals)) {
  year_data <- tot_vals[i, ]
  year <- year_data$year
  year_data <- year_data[-1]
  year_data <- as.numeric(as.character(year_data))

  pie_labs <- paste0(c("Bachelors","High School","Less than High"), " = ", round(100 * year_data/sum(year_data),
2), "%")
  pie(year_data, main = paste("Year", year), labels = pie_labs)
}
```

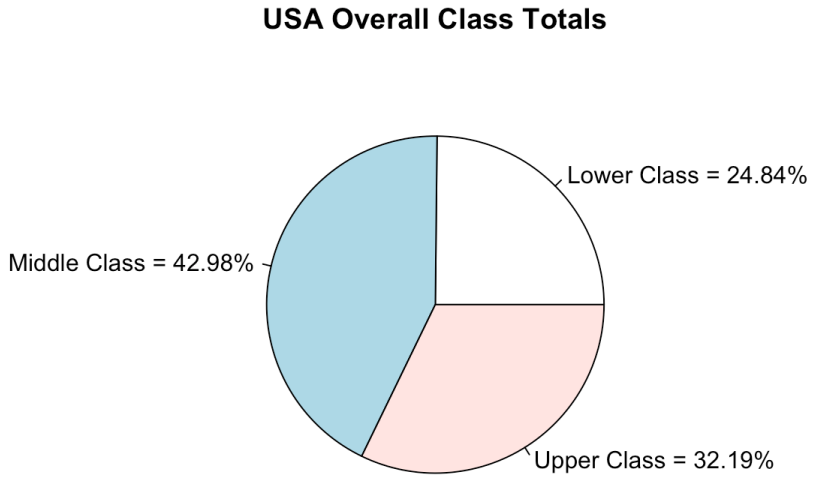




Another question is: What percentage of the country is middle-, upper- and lower-class?

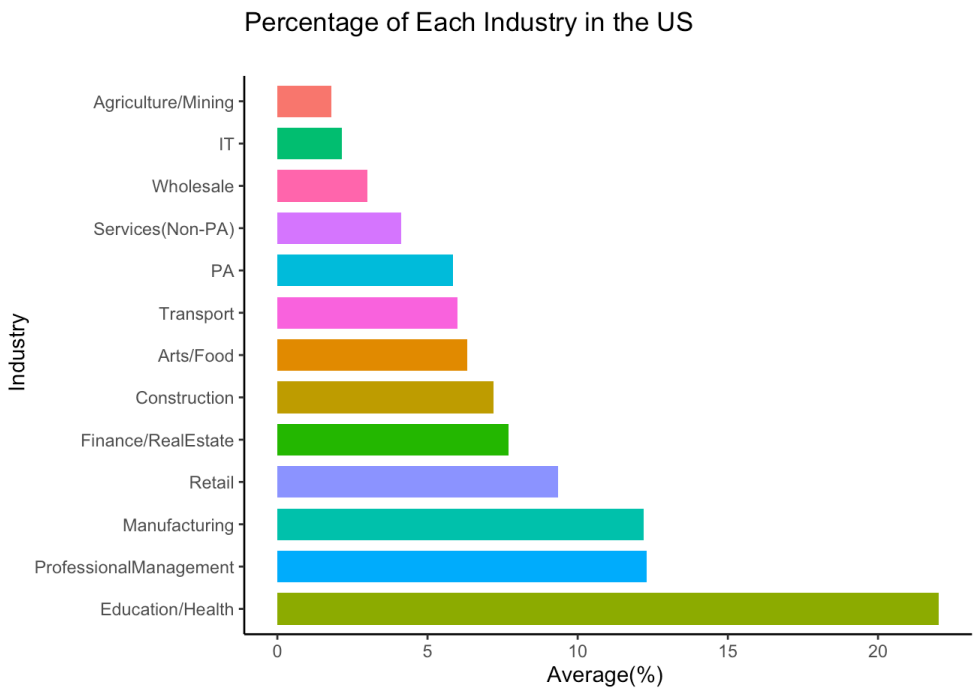
```
# define lower class = <$35,000; middle class = >$35,000 and <$100,000;
# upper class = >$100,000
efi <- efi %>% mutate(lower_class = `5000_to_9999` + `10000_to_14999` + `15000_to_19999` + `20000_to_24999` + `25
000_to_34999`, middle_class = `35000_to_49999` + `50000_to_74999` + `75000_to_99999`, upper_class = `100000_to_14
9999` + `150000_or_more`)
```

```
# get the total sum of each class category
class_sum <- efi %>% summarize(low = sum(lower_class), mid = sum(middle_class), up = sum(upper_class))

# gather the class data
class_dat <- class_sum[1, ]
class_dat <- as.numeric(as.character(class_dat))

# generate the labels for our pie chart
pie_labs <- paste0(c("Lower Class","Middle Class","Upper Class"), " = ", round(100 * class_dat/sum(class_dat),
2), "%")

# generate the pie chart
pie(class_dat, main = paste("USA Overall Class Totals"), labels = pie_labs)
```



Another: What percentage of the country works in each type of industry?

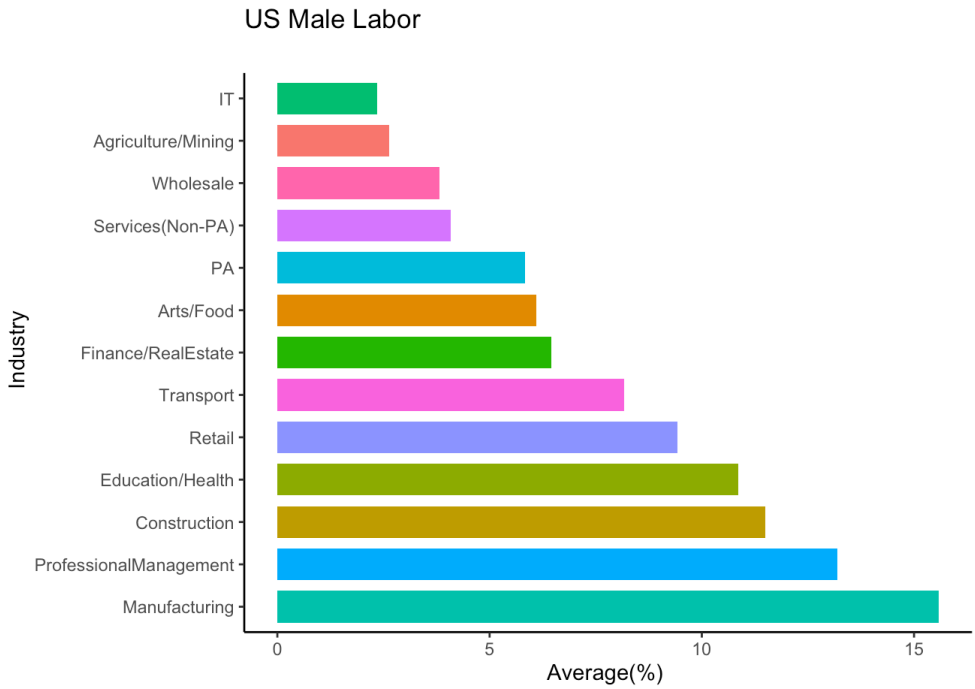
```
# calculate the total sums of the people in each industry
industries <- efi %>% summarize(sum(total_agriculture_forestry_fishing_hunting_mining), sum(total_construction),
sum(total_manufacturing), sum(total_wholesale_trade), sum(total_retail_trade), sum(total_transportation_warehousi
ng_utilities), sum(total_information), sum(total_finance_insurance_realestate_rental_leasing), sum(total_professi
onal_scientific_management_administrative_waste_management_services), sum(total_educationalservices_healthcare_so
cialassistance), sum(total_arts_entertainment_recreation_accommodation_foodservices), sum(total_otherservices_exc
ept_public_administration), sum(total_public_administration))

# abbreviate the industry names for readability
ind_names <- c("Agriculture/Mining", "Construction", "Manufacturing", "Wholesale", "Retail", "Transport", "IT",
"Finance/RealEstate", "ProfessionalManagement", "Education/Health", "Arts/Food", "Services(Non-PA)", "PA")

colnames(industries) <- ind_names
vals <- industries[1,]

# flip the data frame so that we have two columns: industry and avg
industries <- data.frame(Industry = ind_names, Avg = as.numeric(vals))
industries$Avg = 100 * industries$Avg / sum(industries$Avg)

# plot the bar plot in descending order
ggplot(industries, aes(x=reorder(Industry,-Avg), y=Avg, fill=Industry)) +
  geom_bar(stat = "identity", width = 0.75, show.legend = FALSE) +
  coord_flip() +
  labs(x = "\n Industry", y = "Average(%) \n", title = "Percentage of Each Industry in the US \n") +
  theme_classic()
```



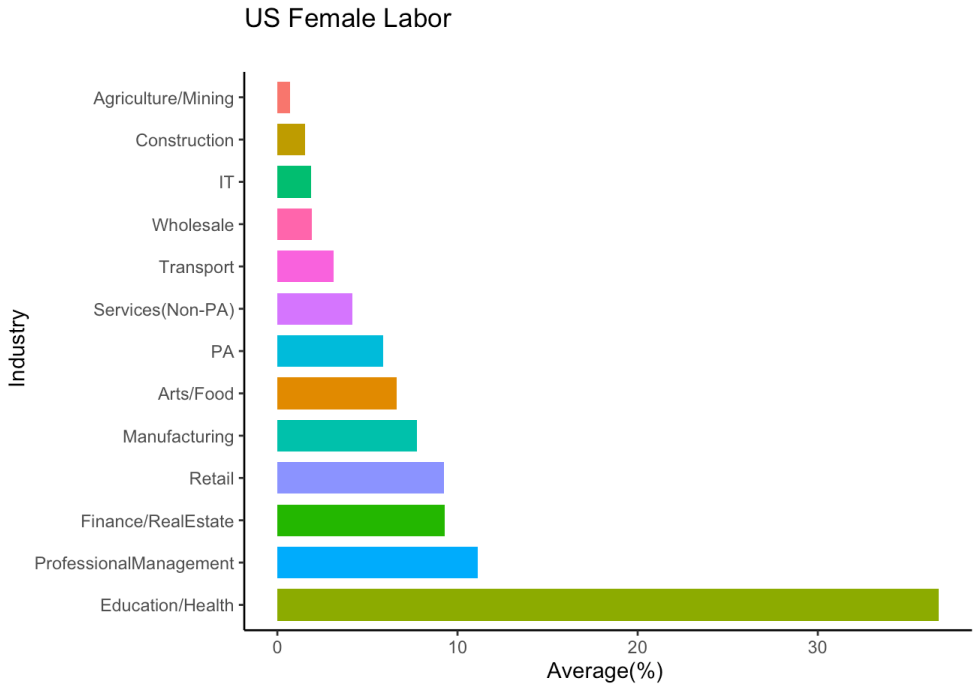
And another: What are the most common industries for men? For women?

```
# here all we need to do is just replace 'total' with 'male' and 'female' using
# gsub, then run the same exact code we did for the overall population
fs <- "summarize(sum(total_agriculture_forestry_fishing_hunting_mining), sum(total_construction), sum(total_manuf
acturing), sum(total_wholesale_trade), sum(total_retail_trade), sum(total_transportation_warehousing_utilities),
sum(total_information), sum(total_finance_insurance_realestate_rental_leasing), sum(total_professional_scientific
_management_administrative_waste_management_services), sum(total_educationalservices_healthcare_socialassistanc
e), sum(total_arts_entertainment_recreation_accommodation_foodservices), sum(total_otherservices_except_public_ad
ministration), sum(total_public_administration))"

men <- gsub("total", "male", fs)
women <- gsub("total", "female", fs)

male_ind <- efi %>% summarize(sum(male_agriculture_forestry_fishing_hunting_mining), sum(male_construction), sum
(male_manufacturing), sum(male_wholesale_trade), sum(male_retail_trade), sum(male_transportation_warehousing_util
ities), sum(male_information), sum(male_finance_insurance_realestate_rental_leasing), sum(male_professional_scien
tific_management_administrative_waste_management_services), sum(male_educationalservices_healthcare_socialassista
nce), sum(male_arts_entertainment_recreation_accommodation_foodservices), sum(male_otherservices_except_public_ad
ministration), sum(male_public_administration))

female_ind <- efi %>% summarize(sum(female_agriculture_forestry_fishing_hunting_mining), sum(female_constructio
n), sum(female_manufacturing), sum(female_wholesale_trade), sum(female_retail_trade), sum(female_transportation_w
arehousing_utilities), sum(female_information), sum(female_finance_insurance_realestate_rental_leasing), sum(fema
le_professional_scientific_management_administrative_waste_management_services), sum(female_educationalservices_h
ealthcare_socialassistance), sum(female_arts_entertainment_recreation_accommodation_foodservices), sum(female_oth
erservices_except_public_administration), sum(female_public_administration))
```

```
# run the same code, just with male and female data
colnames(male_ind) <- ind_names
colnames(female_ind) <- ind_names
male_vals <- male_ind[1,]
fem_vals <- female_ind[1,]

male_ind <- data.frame(Industry = ind_names, Avg = as.numeric(male_vals))
male_ind$Avg = 100 * male_ind$Avg / sum(male_ind$Avg)

female_ind <- data.frame(Industry = ind_names, Avg = as.numeric(fem_vals))
female_ind$Avg = 100 * female_ind$Avg / sum(female_ind$Avg)

ggplot(male_ind, aes(x=reorder(Industry,-Avg), y=Avg, fill=Industry)) +
  geom_bar(stat = "identity", width = 0.75, show.legend = FALSE) +
  coord_flip() +
  labs(x = "\n Industry", y = "Average(%) \n", title = "US Male Labor \n") +
  theme_classic()
```




```
ggplot(female_ind, aes(x=reorder(Industry,-Avg), y=Avg, fill=Industry)) +
  geom_bar(stat = "identity", width = 0.75, show.legend = FALSE) +
  coord_flip() +
  labs(x = "\n Industry", y = "Average(%) \n", title = "US Female Labor \n") +
  theme_classic()
```



So, as is seen above, we have quickly generated a whole lot of new data that we can analyze when we get to later sections of the project.

What are different ways you could look at this data?

There are numerous ways we can go about looking at our data and analyzing it. We can use summary statistics like mean, median and mode to offer a snapshot of the data's central tendency and dispersion. We can also use correlation tests like Pearson's R to find out the strength of relationships between different variables. Heat maps will give us a nice and clear visual representation of our data that makes it easier to identify trends. In addition, regression models allow us to predict future outcomes and assess which variables have the greatest predictive impact.

Each of these techniques will be used to look at the data in this project.

How do you plan to slice and dice the data?

As we have already seen earlier, I merged each of the three data sets related to education, finance and industry all into a single data frame. This will make it easier to compare information between the three initial data frames all at once. I have also already started creating new variables such as classes - lower-, middle- and upper-class, as well as separate the district codes into district number and state.

I plan to group data based on education levels, financial classes, industry and other factors, group it by year, and group it by state. Grouping data by state will make it easier to generate heat maps that demonstrate the differences between different regions of the country.

How could you summarize your data to answer key questions?

Let's look at another one of the questions: What is the correlation between education level and financial class?

```
# create a column for the total number of people for all education levels
efi <- efi %>% mutate(edu_total = bachelors_degree_or_higher + high_school_or_some_degree + less_than_high_school
_graduate)

# create the same type of column for class levels
efi <- efi %>% mutate(class_total = lower_class + upper_class + middle_class)
```

```
# now create columns for the proportion of people with bachelors degrees, high
# school only, and less than high school
efi <- efi %>% mutate(bach_rate = bachelors_degree_or_higher/edu_total,  high_rate = high_school_or_some_degree/e
du_total, less_rate = less_than_high_school_graduate/edu_total)

# do the same thing with income class
efi <- efi %>% mutate(low_rate = lower_class/class_total, up_rate = upper_class/class_total, mid_rate = middle_cl
ass/class_total)
```

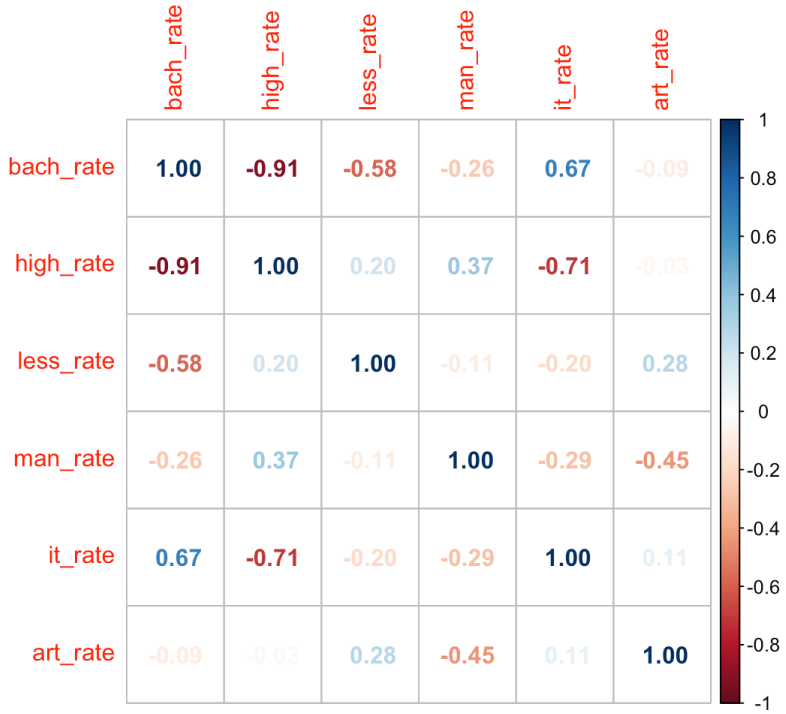
```
# use these columns
using_cols <- c("bach_rate", "high_rate", "less_rate", "low_rate", "mid_rate", "up_rate")

# generate a correlation plot
efi_after_19 <- efi %>% filter(year > 2019)
corr_mat <- cor(efi_after_19[, using_cols])
corrplot(corr_mat, method = "number")
```

I could summarize the data in this instance by noticing a substantial positive correlation between the proportion of upper class people with the proportion of people with bachelor's degrees. Likewise, a larger proportion of middle class and lower class people also tends to have a higher proportion of people with only a high school or less than high school education.

And what about the other correlation question: - What is the correlation between the percentage of manufacturing jobs and education level?

```
# here we are doing the same thing with the rates of three different industries:
# manufacturing, information (IT), and arts/entertainment/food
dummy_df <- efi %>% mutate(man_rate = (male_manufacturing + female_manufacturing) / (male + female), it_rate = (m
ale_information + female_information)/(male + female), art_rate = (male_arts_entertainment_recreation_accommodati
on_foodservices + female_arts_entertainment_recreation_accommodation_foodservices)/(male + female))
```

```
using_cols <- c("bach_rate", "high_rate", "less_rate", "man_rate", "it_rate", "art_rate")

efi_after_19 <- dummy_df %>% filter(year > 2019)
corr_mat <- cor(efi_after_19[, using_cols])
corrplot(corr_mat, method = "number")
```



Here we see a strong correlation between the proportion of people working in IT to the proportion with a college degree. What is also interesting is that there is an even stronger negative relationship between the proportion of IT workers and the proportion of middle-class income earners. There also seems to be a moderate correlation between the proportion of middle class earners to the proportion with only a high school degree.

Let's also do a correlation between class and some industries:

```
using_cols <- c("low_rate", "mid_rate", "up_rate", "man_rate", "it_rate", "art_rate")

corr_mat <- cor(dummy_df[, using_cols])
corrplot(corr_mat, method = "number")
```

Again, we see similar trends. The proportion of people working in IT has a positive relationship with the percentage of upper class earners, and it has a negative relationship with the percentage of middle and lower class income earners. We also see a moderately negative relationship with the proportion of art/food industry workers and the proportion that work in manufacturing.

In the final step to this project, I will go further in assessing the reason's behind these correlations, the trends that have been found, and how to

interpret the charts and plots that have been generated.

What types of plots and tables will help you to illustrate the findings to your questions?

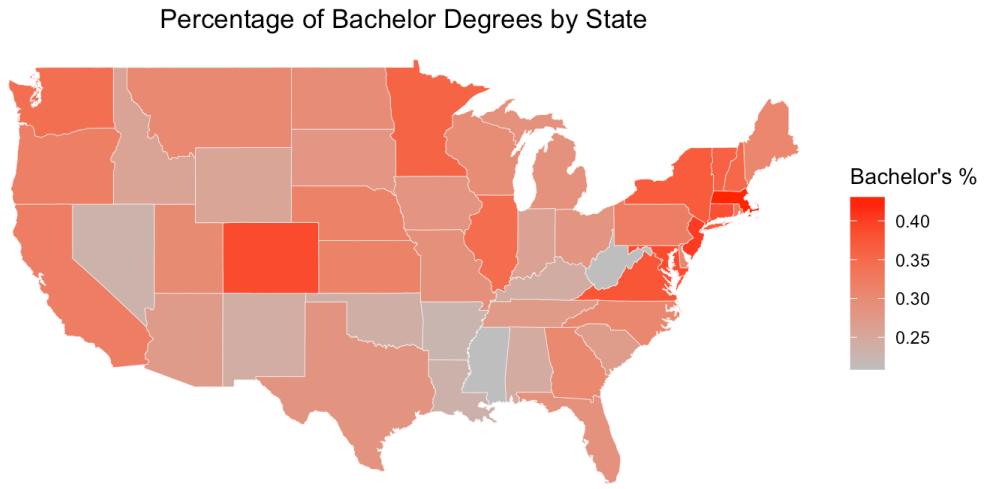
As can be seen above, I have already produced some pie charts, horizontal bar charts, and correlation plots to help illustrate the findings to my research questions. I will also include some heat maps of the USA and some accompanying tables. See below:

```
# this code generates a df with state and the percentage of bachelors degrees
# in that state
state_ed <- efi %>% group_by(state) %>% filter(year>2019) %>% summarize(bach_pct = sum(bachelors_degree_or_highe
r) / sum(edu_total))
state_ed <- state_ed %>% select(state, bach_pct)
state_ed %>% arrange(desc(bach_pct)) %>% head(10)
```

```
## # A tibble: 10 × 2
##    state bach_pct
##    <chr>    <dbl>
##  1 DC       0.575
##  2 MA       0.430
##  3 NJ       0.404
##  4 MD       0.388
##  5 CO       0.387
##  6 CT       0.383
##  7 VA       0.376
##  8 NY       0.369
##  9 VT       0.361
## 10 MN       0.358
```

```
# this data frame has a list of state names and state codes, so we can
# convert state codes to state names for the purpose of merging our data
state_df <- read.csv("states.csv")

# merge the two data frames by state name and bachelors percentage
state_ed <- merge(state_ed, state_df, by=c("state")) %>% select(state_name, bach_pct)

# make the names lower case
state_ed$state_name <- tolower(state_ed$state_name)
```

```
# generate the map data for the USA, and merge by region
map_data <- map_data("state")
merged_data <- merge(map_data, state_ed, by.x = "region", by.y = "state_name", all.x = TRUE, sort = FALSE)
merged_data <- merged_data[order(merged_data$order),]

# generate a heat map using ggplot
heat_map <- ggplot(data = merged_data) +
  geom_polygon(aes(x = long, y = lat, fill = bach_pct, group = group),
           color = "white", size = 0.1) +
  scale_fill_gradient(low = "gray", high = "red", name = "Bachelor's %") +
  coord_map() +
  labs(title = "Percentage of Bachelor Degrees by State") +
  theme_void() +
  theme(plot.title = element_text(hjust = 0.5))

heat_map
```

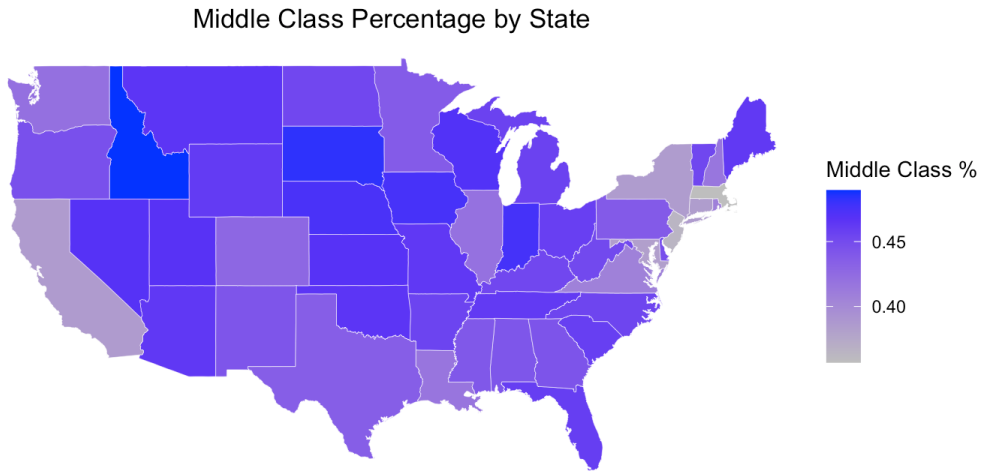



```
# now do the same process with the proportion of middle class earners
state_mid <- efi %>% group_by(state) %>% summarize(mid_pct = sum(middle_class) / sum(class_total))
state_mid <- state_mid %>% select(state, mid_pct)
state_mid %>% arrange(mid_pct) %>% head(10)
```

```
## # A tibble: 10 × 2
##    state mid_pct
##    <chr>   <dbl>
##  1 PR      0.312
##  2 DC      0.320
##  3 MA      0.357
##  4 NJ      0.365
##  5 MD      0.382
##  6 CT      0.384
##  7 NY      0.385
##  8 CA      0.386
##  9 HI      0.402
## 10 VA      0.408
```

```
# again, transform the state ID to state name
state_mid <- merge(state_mid, state_df, by=c("state")) %>% select(state_name, mid_pct)

state_mid$state_name <- tolower(state_mid$state_name)
```

```
# perform the same operations to generate anothe heat map
merged_data <- merge(map_data, state_mid, by.x = "region", by.y = "state_name", all.x = TRUE, sort = FALSE)
merged_data <- merged_data[order(merged_data$order),]

heat_map <- ggplot(data = merged_data) +
  geom_polygon(aes(x = long, y = lat, fill = mid_pct, group = group),
           color = "white", size = 0.1) +
  scale_fill_gradient(low = "gray", high = "blue", name = "Middle Class %") +
  coord_map() +
  labs(title = "Middle Class Percentage by State") +
  theme_void() +
  theme(plot.title = element_text(hjust = 0.5))

heat_map
```




```
# finally, do the same thing with the percentage of IT workers
state_it <- efi %>% group_by(state) %>% summarize(it_amt = sum(male_information + female_information)/sum(male +
female))
state_it <- state_it %>% select(state, it_amt)
state_it %>% arrange(desc(it_amt)) %>% head(10)
```

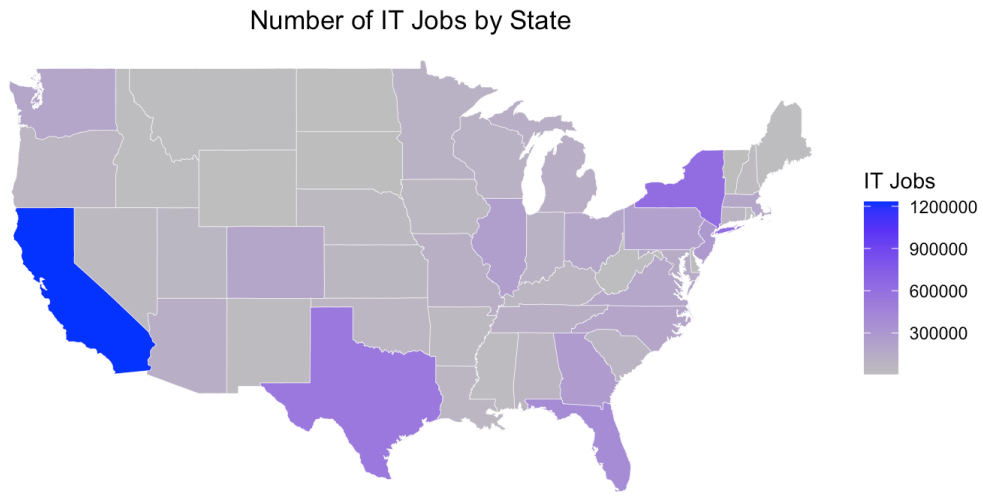
```
## # A tibble: 10 × 2
##    state it_amt
##    <chr>  <dbl>
##  1 DC    0.0396
##  2 CA    0.0321
##  3 CO    0.0310
##  4 NY    0.0309
##  5 NJ    0.0299
##  6 GA    0.0255
##  7 MA    0.0252
##  8 WA    0.0249
##  9 CT    0.0227
## 10 AK    0.0224
```

```
# just to change things up, we will look at the total number of IT jobs
# per state instead
state_it <- efi %>% group_by(state) %>% summarize(it_amt = sum(male_information + female_information))
state_it <- state_it %>% select(state, it_amt)
state_it %>% arrange(desc(it_amt)) %>% head(10)
```

```
## # A tibble: 10 × 2
##    state  it_amt
##    <chr>   <int>
##  1 CA    1232685
##  2 NY     623647
##  3 TX     537297
##  4 FL     393079
##  5 NJ     286992
##  6 GA     274752
##  7 IL     250630
##  8 PA     232230
##  9 NC     194657
## 10 CO     191605
```

```
# change state codes to state names
state_it <- merge(state_it, state_df, by=c("state")) %>% select(state_name, it_amt)

state_it$state_name <- tolower(state_it$state_name)
```

```
# merge the data and generate the heat map
merged_data <- merge(map_data, state_it, by.x = "region", by.y = "state_name", all.x = TRUE, sort = FALSE)
merged_data <- merged_data[order(merged_data$order),]

heat_map <- ggplot(data = merged_data) +
  geom_polygon(aes(x = long, y = lat, fill = it_amt, group = group),
           color = "white", size = 0.1) +
  scale_fill_gradient(low = "gray", high = "blue", name = "IT Jobs") +
  coord_map() +
  labs(title = "Number of IT Jobs by State") +
  theme_void() +
  theme(plot.title = element_text(hjust = 0.5))

heat_map
```



Do you plan on incorporating any machine learning techniques to answer your research questions?

Although I don't plan on using almost any machine learning techniques to answer my research questions, I can generate a linear regression model like the one below in order to further assess the relationship between some of the variables:

```
ed_mod <- lm(data = efi, bach_rate ~ up_rate + total_information)
summary(ed_mod)
```

```
##
## Call:
## lm(formula = bach_rate ~ up_rate + total_information, data = efi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.17770 -0.03292 -0.00378  0.03187  0.37095
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.039e-02  6.605e-03   6.115 1.46e-09 ***
## up_rate           7.379e-01  2.109e-02  34.989  < 2e-16 ***
## total_information 4.582e-06  5.116e-07   8.957  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06105 on 871 degrees of freedom
##   (437 observations deleted due to missingness)
## Multiple R-squared:  0.6974, Adjusted R-squared:  0.6967
## F-statistic:  1003 on 2 and 871 DF,  p-value: < 2.2e-16
```

The above model predicts the proportion of bachelor's degree holders from the proportion of upper class earners and the total number of IT workers. It is statistically significant (p-value < 0.001), and the two predictor variables account for nearly 70% of the variance (Adjusted R-squared ~ 70%).

The kind of assessment uses some basic machine learning techniques like regression modeling in order to predict an outcome based on a limited pool of predictor variables. If we gather new data in the future from subsequent ACS surveys, we better assess the strength of the model that has been created.

What questions do you have now, that will lead to further analysis or additional steps?

After reviewing the charts, tables and maps that have been produced, I have come up with some questions that will help lead to further analysis in the final step of the project:

- How does education level compare to income class? How can we use the pie charts and the correlation charts to extrapolate further?
- What are the most popular industries overall? How and why are they different for men and women as compared with the total population?
- How do the proportion of workers in certain industries correlate with the education level of the area, as well as the income classes of the area?
- What do the heat maps of the USA tell us about certain regions of the country?

These should all lead to a better overall assessment of the status of the country in the three key areas of education, finance, and industry.

Part 3

Introduction

In this research paper, we analyzed some of the vast array of data provided by the U.S. Census Bureau during their annual American Community Survey. Focusing specifically on indicators within the realms of education, finance, and industry, we aimed to offer valuable insights into the current state of U.S. society and economy.

Education is a good indicator of societal progress and development. By assessing the highest level of education attained by the population, we hoped to gain a better understanding of the overall educational landscape. Finance plays a critical role in shaping the dynamics of a society and its economy. By looking at the income levels of the population, we then correlated that with education data to generate further insights. Industry data further supplemented our view of how the economy of a region operates, and what kinds of job opportunities are available in a region. All of these sectors of society helped to produce a better picture of the current state of the country.

By aggregating and assessing data from these three key domains, we provided valuable insights for various stakeholders, including policymakers, businesses, educators, and individuals. Whether it's a family considering a relocation decision or a local government official seeking to understand economic trends, our research aims to give everyone the information they need to make important decisions. Through this analysis, our hope is that the contribution to informed decision-making will improve the capabilities of individuals, communities, and our overall society.

The Problem

The problem statement we addressed was to determine trends between the sectors of education, finance and industry, correlations between them, and to use that information to determine the overall state of the country and of individual states. In order to concretely address this rather abstract problem statement, we also put forward nine research questions that helped solve this problem:

- What are the overall average educational levels in the US?
- What areas of the country are more educated? Less?
- What percentage of the country is middle-, upper- and lower-class?
- Where is the size of the middle class the lowest in the country?
- What is the correlation between education level and financial class?
- What percentage of the country works in each type of industry?
- What are the most common industries for men? For women?
- Which parts of the country have the most IT jobs?
- What is the correlation between the percentage of manufacturing jobs and education level?

How the Problem was Addressed

In order to address the problem, we looked at three different sets of data between the years 2019-2021 in the domains of education, finance and industry. Each set of data listed the total count of people that fell into various categories among each of the 437 district across the USA (Puerto Rico and Washington D.C. are included in the data, in addition to the usual 435 congressional districts). The education data included counts of the number of people who have achieved certain levels of education. The finance data contained counts of people who fell into certain income ranges. The industry data included counts of people who worked in various industrial sectors, as well as the same information separated by gender.

The methodology used to assess this data was to read each of the three csv files into data frames, then merge all three of those data frames into a single data frame. Because of missing 2019 data for education, that data frame had to be filled with NA values for that year. Once the data was aggregated, it was then cleaned and edited to make for easier manipulation later on. Pie charts were generated for education level and class distribution. Bar charts were made for industry allocation overall, as well as by gender. Correlation charts were produced for one-to-one interactions among important variables in each of the data sets. Lastly, heat maps of the USA were formed for visualizing the distributions of various categories across the country.

A simple linear regression model was also created to help predict the proportion of bachelor's degree holders in an area by using the proportion of upper class individuals as well as the number of people working in IT. This could be used to assess the education levels of an area without actually needing to know what they are. We would simply need to look at the income distribution of the area.

Analysis and Findings

From the year 2020 to 2021, the percentage of the population with a bachelor's rose from 30.9% to 31.6%, while the percentage with only a high school degree fell from 58.3% to 57.8%. The percentage of people with less than a high school diploma stayed relatively the same at around 10.6%. This shows that although more people are getting higher education, there are still many who are not even getting a high school education.

For class distributions, those making <$35k (lower class) accounted for 24.8% of the population, >$35k and <$100k (middle class) for 43%, and >$100k (upper class) for 32.2%. This would seem to indicate that income distribution is relatively balanced. However, it is important to keep in mind that this does not account for the average cost of living and relative buying power.

In terms of industry, education and health made up the bulk of the overall distribution of labor, accounting for 23%, higher than all other industries by more than 10%. Wholesale, IT and agriculture were the smallest sectors, each accounting for less than 5% of total labor. What is more interesting, though, is the difference between male and female labor. The most popular industries for men were manufacturing, management and construction, which were all around 10-15% of the total labor distribution. On the other hand, the most popular industry for women was health & education, accounting for over 38% of the labor share. This was 25% more than the next highest-represented industry. This massive number of women working in the education and health sector helps demonstrate more clearly why it is also the most popular sector overall.

Correlation testing also gave us some new insights into the relationships among the data. There was a substantial positive correlation between the proportion of upper class people with the proportion of people with bachelor's degrees. Likewise, a larger proportion of middle class and lower class people also tends to have a higher proportion of people with only a high school or less than high school education. In addition, there was a strong correlation between the proportion of people working in IT to the proportion with a college degree. There was an even stronger negative relationship between the proportion of IT workers and the proportion of middle-class income earners. Finally, we saw that the proportion of people working in IT has a positive relationship with the percentage of upper class earners, and it has a negative relationship with the percentage of middle and lower class income earners.

In the USA heat maps, we also derived some useful information that sheds some light on how different parts of the country are related. The percentage of bachelor's degrees was the highest along the east coast, particularly in New England, with Colorado, Minnesota and Illinois also producing a higher number of college graduates. The lowest percentage of college degrees were found in the south. Conversely, the part of the country with the highest proportion of middle-class earners was the the midwest, with places like South Dakota, Indiana and Idaho leading the pack. The east coast and California had the smallest proportion of middle-class earners. When looking at total IT jobs, California had the most by far, even when adjusting for the ratio of IT jobs to the overall population. Colorado had the 3rd highest proportion of IT jobs, despite having only the 10th highest number of overall IT jobs.

Implications

More Americans are graduating with college degrees, while there are still many that are not graduating high school. Add this to the fact that there is a strong positive correlation between higher education and higher income for each education and income level. This would implicate that there may be a growing disparity between the upper and lower classes in the future. As the more highly educated continue to get richer, the less educated will be left with lower wages and bleaker future prospects.

The fact that the largest industry in America is now Healthcare and Education has a lot to do with the growing impact of women in the workplace. As fewer and fewer women opt to become stay-at-home mothers and instead pursue careers, these industries will only continue to grow. One other surprising thing about the industrial data, though, is that the proportion of IT jobs is still quite small. We would expect that as we become ever more dependent on technology, this number will continue to rise. But with the advent of ever more impressive AI technology, who knows what the trajectory of these industries will truly be?

Limitations

There were many limitations that had to be taken into account while conducting this research project. One is that all of the 2019 data was missing for the education csv file. Another is that all the information was given in the form of blanket counts, without taking into account other demographic features like race, age and gender. The industrial data did account for gender, but that was the only data set to do so. Having more statistics for each district or state such as average income or average demographics would have helped with adjusting our analysis to adjust for variations among different regions. Additionally, the maps package for rendering
the US map did not account for places like Alaska, Hawaii and Puerto Rico, which were represented in the data frame but not graphically.

Concluding Remarks

Overall, the direction of the country seems to be doing all right. This data was collected during the peak of the COVID-19 pandemic, and yet the proportion of bachelor's degree earners continued to rise, the income levels of the country continued to stay relatively stable, and there appeared to be strong participation in important sectors such as healthcare, education and manufacturing. Income has a strong positive relationship with education level, as well as with the proportion of IT jobs. This is a particularly useful bit of information to keep in mind for people looking to transition into the tech industry or for employers to recruit bright young talent. Through this research paper, industry leaders and policy makers will be better informed on to make critical decisions about the future direction of the country.