Sushmitha Mohana
EM-622
CRISP DM Phases

Pre-requisite steps: Install all required packages and invoke them

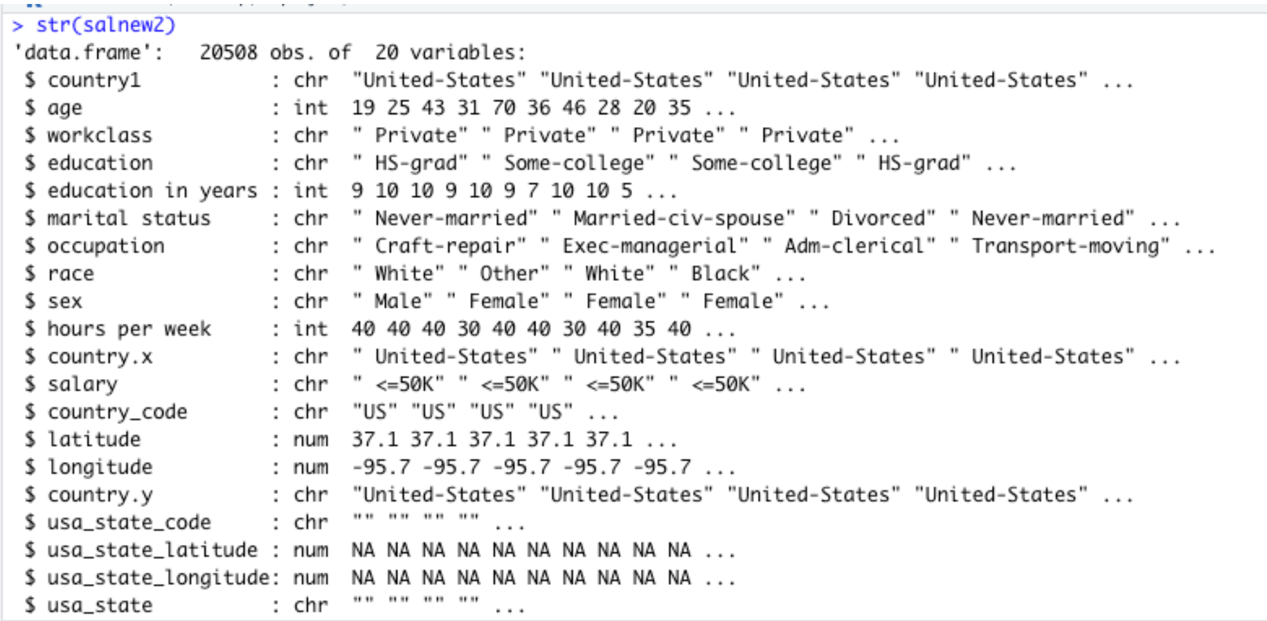
Phase 1: Business Understanding
The objective of this project is to determine the demographic and characteristics of a population in poverty or at risk of poverty. The inferences from the project can be used by government agencies to formulate policies to improve the lives of persons experiencing poverty or who are at risk of poverty.

Phase 2: Data Understanding

I have used two datasets for the purpose of this project:

Data set 1: ‘adult.csv’ containing 48,842 rows of census data, including information about the per annum salary of the individual, which can be >$50k or <=$50k

Data set 2: ‘A1.csv’ containing the latitude and longitude data of countries. This dataset has been used for geographical plots only.
Examined the dataset ‘adult.csv’ to understand data attributes. I noticed inaccurately named columns with leading and trailing spaces. Noticed cell value ‘?’ For some columns.



```
> str(salnew2)
'data.frame':	20508 obs. of  20 variables:
 $ country1           : chr  "United-States" "United-States" "United-States" "United-States" ...
 $ age                : int  19 25 43 31 70 36 46 28 20 35 ...
 $ workclass          : chr  " Private" " Private" " Private" " Private" ...
 $ education          : chr  " HS-grad" " Some-college" " Some-college" " HS-grad" ...
 $ education in years : int  9 10 10 9 10 9 7 10 10 5 ...
 $ marital status     : chr  " Never-married" " Married-civ-spouse" " Divorced" " Never-married" ...
 $ occupation         : chr  " Craft-repair" " Exec-managerial" " Adm-clerical" " Transport-moving" ...
 $ race               : chr  " White" " Other" " White" " Black" ...
 $ sex                : chr  " Male" " Female" " Female" " Female" ...
 $ hours per week     : int  40 40 40 30 40 40 30 40 35 40 ...
 $ country.x          : chr  " United-States" " United-States" " United-States" " United-States" ...
 $ salary             : chr  " <=50K" " <=50K" " <=50K" " <=50K" ...
 $ country_code       : chr  "US" "US" "US" "US" ...
 $ latitude           : num  37.1 37.1 37.1 37.1 37.1 ...
 $ longitude          : num  -95.7 -95.7 -95.7 -95.7 -95.7 ...
 $ country.y          : chr  "United-States" "United-States" "United-States" "United-States" ...
 $ usa_state_code     : chr  "" "" "" "" ...
 $ usa_state_latitude : num  NA NA NA NA NA NA NA NA NA NA ...
 $ usa_state_longitude: num  NA NA NA NA NA NA NA NA NA NA ...
 $ usa_state          : chr  "" "" "" "" ...
```

Fig: column names with respective data types of census data in ‘adult.csv’

Phase 3: Data Preparation

- Checked and removed rows with Null and NA values. Accurately named the columns. Removed columns I found unnecessary for the analysis.
- Joined data frames when the analysis required it

Phase 4: Modeling

- I have tried to drill down on the characteristics of the population earning<= 50K/year based on birth country, race, sex, and marital status. Using each visualization as a stepping stone for the next filtration.

■ Plot 1:
The plot answers a simple question- do people not born in the U.S. suffer an economic disadvantage compared to persons born in the U.S.?

For the visualization purpose, I framed the question in an easy-to-understand manner:
What is the country of origin of individuals earning less than $50,000 per annum?

Contrary to my expectation, it was found that the majority(20,000+) of persons in this category (<=50K salaries) were born in the U.S.
I found that using an interactive world map would be the best way to identify places with the highest concentration of persons. Zooming in to a geographical zone(America, Asia, Europe) gives a country-wise breakdown of the population belonging to the category

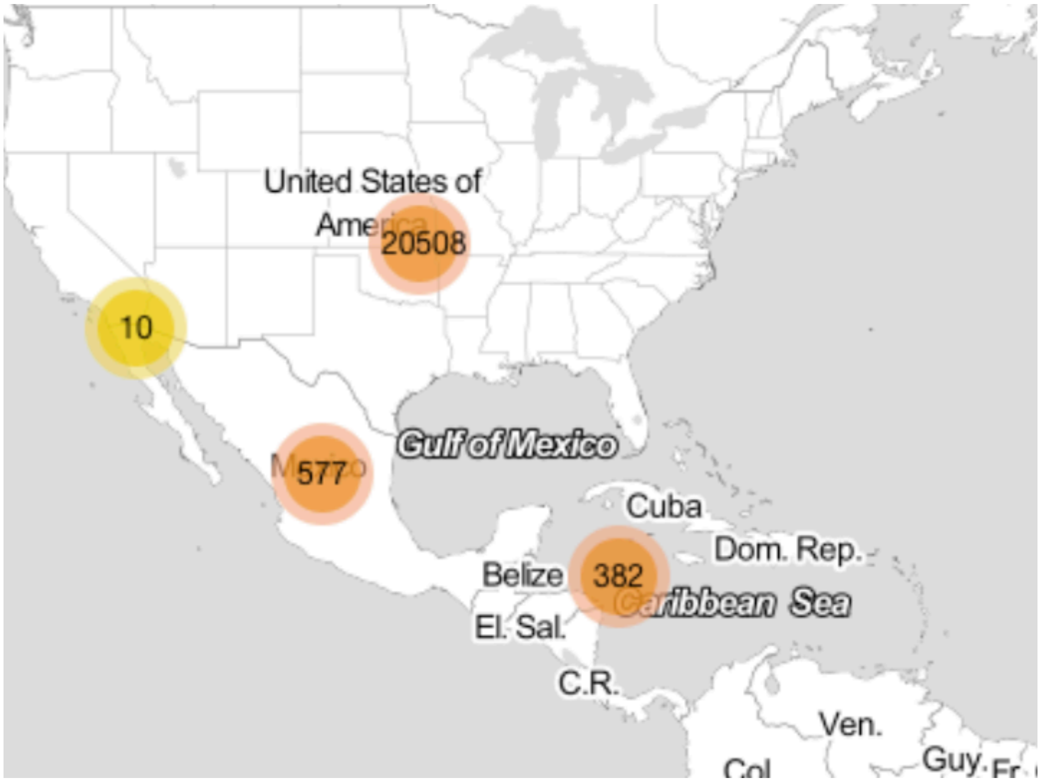
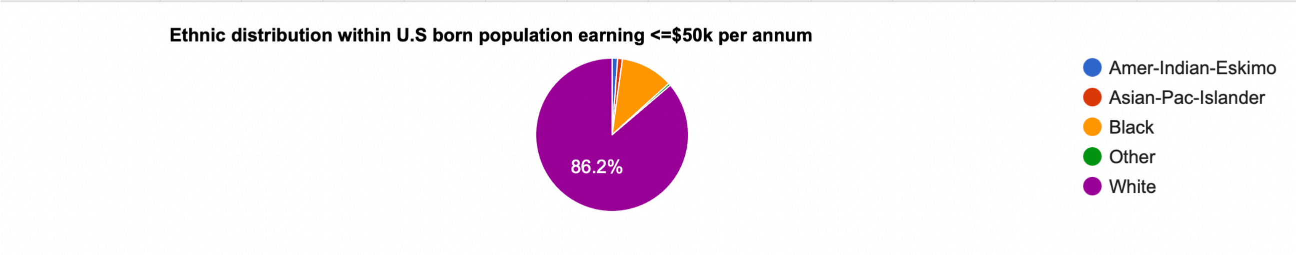


Fig1 : plot1

■ Plot 2

I tried to drill down further and know the ethnic group/race of the person born in the U.S belonging to this category. 5 ethnic groups were identified, and 86.2% of the persons were found to be white.I chose a pie chart to represent the data because I wanted to show a 'percentage of the whole population.'



Data: race • Chart ID: PieChartID98573ef9eaa • googleVis-0.7.0
R version 4.2.1 (2022-06-23) • Google Terms of Use • Documentation and Data Policy

Fig2: Interactive Plot-2

- Plot-3 :

For the identified category in the previous plot, how does poverty /poverty risk vary with gender and age?

I observed from the above graph that the count of persons in poverty or at risk of poverty decreases with age, irrespective of gender, and more men are prone to poverty than Women of the same age.

I have used a scatterplot as it helps in identifying the relationship between two or more variables and resurfacing potential co-relations between them.

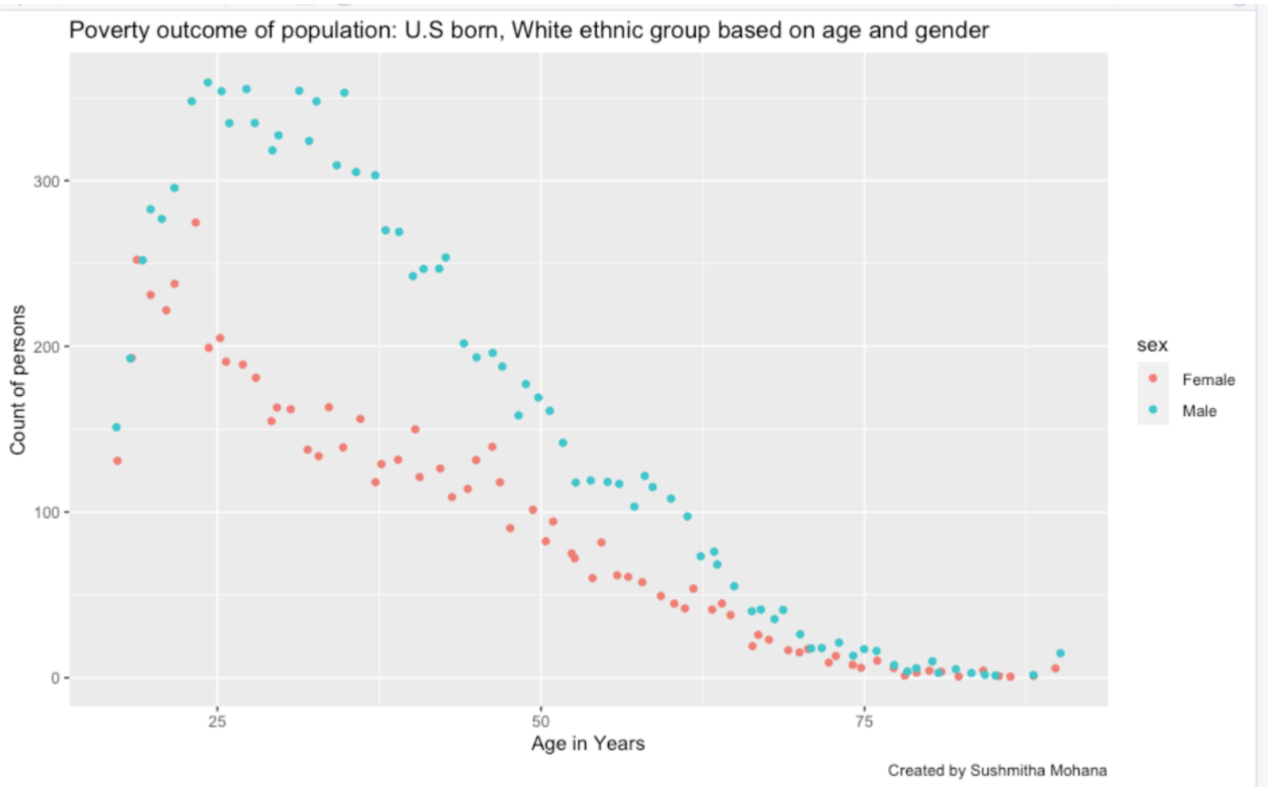


Fig3: Plot 3

Plot 4: Now that I understood the gender of the category, I wanted to understand if the profession plays a role in poverty/poverty risk

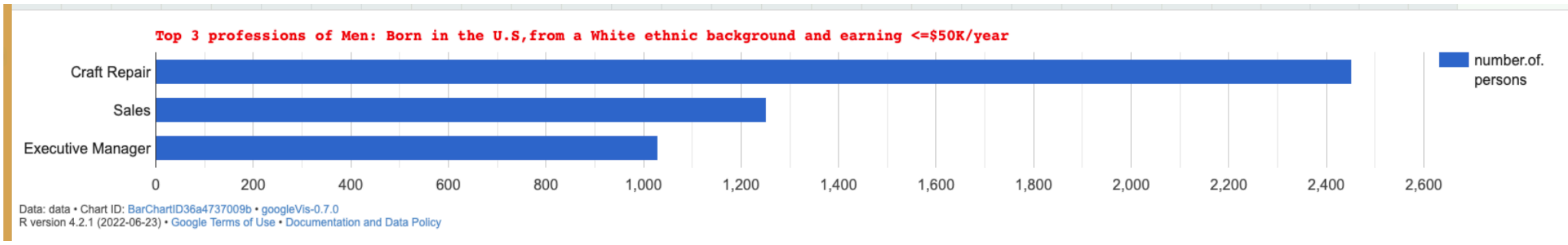


Fig 4: Plot 4

It was found that the count of Men in Craft Repair, Sales, and Managerial executive occupations was the highest in this population.

It is uncanny that Executive managers are in the top 3 of this list because the salary range for Higher Executive Manager positions is ~77K/year and upwards. Lower-level Executive managers make <50K/year. We do not have enough information from the data to decide to which Level these Executive Managers belong. The same applies to Sales professionals as well.

Therefore I can exclude ‘occupation’ while understanding the features of the population segment.

I have chosen to make a horizontal bar chart because I found it to be the best while making a comparison. In my case, it helped compare the count of persons belonging to a particular profession.

■ Plot 5: I wanted to know the majority number of years of education within this population segment and the majority marital status within the education segment.

From the plot below, it is evident that most persons from this category had nine years of education. It is to be noted that, according to the data, these 9 years exclude preschool to grade 3. Therefore 9 years of education means education up to High school only. Furthermore most persons belonging to this category were married to a civilian spouse.

I have chosen to make an interactive stacked bar chart to answer the question because I wanted to find the majority population count within a particular subcategory. For example: through the stacked chart we can say that there are 2,389 persons with 9 years of education, married to a civilian spouse

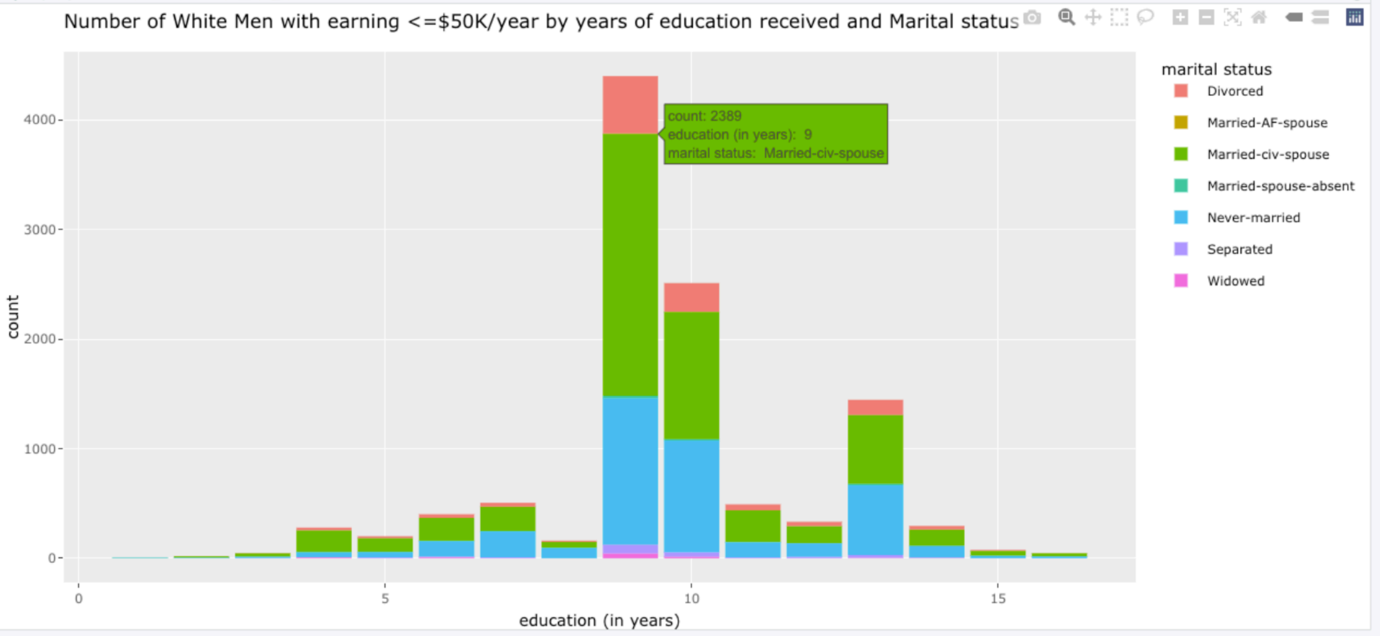


Fig: Plot 5

Conclusion:

This project concludes that U.S born White men, whose maximum education level is a High School diploma (9 years of education, excluding preschool to grade 3) and who is married to a Civilian Spouse, are at the highest risk of being in /already in poverty.

Drawing lessons from the inferences of this project. The government of the United States should focus on the following:

- Educational Initiatives to boost the White Male population to pursue a college education

- Design Schemes to benefit Men with civilian spouses, especially if the male partner is White

Please note that :

- Though I have added the code to change the legend of the scatter plot, the visualization does not reflect the change.