Iliana Vasslides

EDA Lab Writing

2. The variables I chose to look at are `wrkstat`, `marital`, `educ`, `polviews`, `happy`, `satfin`, and `realinc.` These represent labor force status, marital status, highest level of school completed, political views in terms of how liberal or conservative, general happiness, satisfaction with financial situation, and family income, respectively. I think that all of these variables could be interconnected with each other, and show some correlation that would be interesting to analyze and graph. General happiness could be affected by marital status and labor force status. It could possibly be affected by political views, if a person leans strongly one way and is strongly affected by election results. Additionally, general happiness could be assumed to be almost directly correlated with satisfaction with financial situation. These two variables measure the feelings of satisfaction in general and then more specifically, with financial situations. I think there might also be a correlation or relationship between the highest level of education, labor force status, satisfaction with financial situation, and family income. It can be assumed that someone with a high level of education is working full time in the labor force and mostly satisfied with their financial situation. On the other hand, it can also be assumed that someone with a low level of education is either working full-time or not working at all. This is one of the reasons why I chose this combination of variables. While there are many relationships between happiness, marital status, education level, political views, work status, and financial situation that can be correctly assumed, it is not unlikely for there to be outliers. These outliers could be people who are having some other problems in life such as family, friends, or partner issues, that then affect their general happiness, even if they are satisfied with their financial status. These types of outliers are what I think would make for interesting analysis.

4. The first numeric summary is a cross table between `happy` and `satfin`, or general happiness compared to satisfaction of financial situation. We can see that “Pretty happy” had the highest frequency for every satisfaction rating, except for the “Inapplicable” category. “Pretty happy” holds a big lead in every satisfaction rating category except for “Pretty well satisfied”, where the “Very happy” category has a very close frequency.

satfin happy	Inapplicable	More or less satisfied	No answer	Not satisfied at all	Pretty well satisfied
Inapplicable	4383	0	0	0	0
No answer	0	171	40	113	95
Not too happy	0	3166	27	4937	1260
Pretty happy	0	18097	132	10128	9456
Very happy	0	8553	86	2831	8915

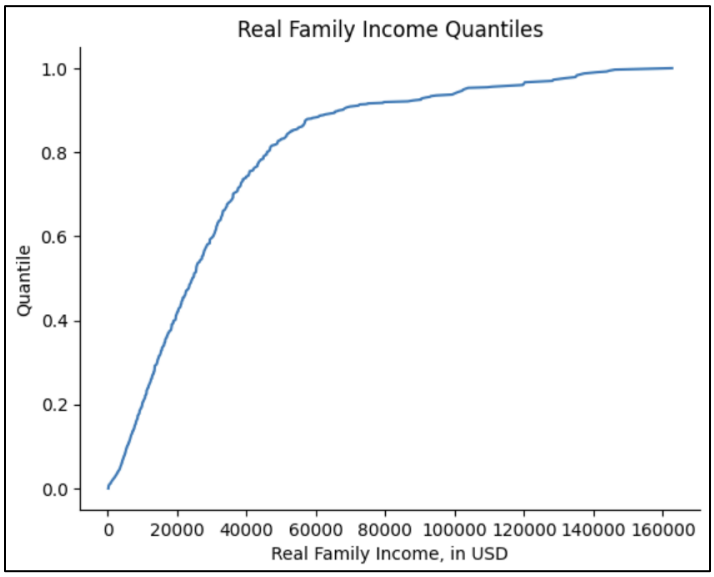
The second numeric summary is another cross table, this one comparing highest education level and satisfaction of financial situation. The highest education level of “High School” holds the highest frequency for each satisfaction rating. Looking beyond that, we can see that “4 years of college” has the next highest frequency for the “More or less satisfied” as well as the “Pretty well satisfied” rankings. “Some high school” has the next highest frequency for the “Not satisfied at all” ranking.

satfin educ	Inapplicable	More or less satisfied	No answer	Not satisfied at all	Pretty well satisfied
1 year of college	401	2477	21	1652	1354
2 years of college	598	3529	29	2134	1918
3 years of college	227	1358	17	861	844
4 years of college	667	4051	30	1768	3478
5 years of college	152	1046	6	411	777
6 years of college	193	1187	9	434	1122
7 years of college	69	454	3	167	419
8+ years of college	126	668	2	264	743
Elementary school	13	150	1	132	114
High school	1239	9083	86	5712	5281
Intermediate school	87	1138	15	668	878
No answer	15	97	20	65	66
No formal schooling	10	67	6	52	42
Some elementary school	34	347	5	224	191
Some high school	476	3614	28	2981	2025
Some intermediate school	76	721	7	484	474

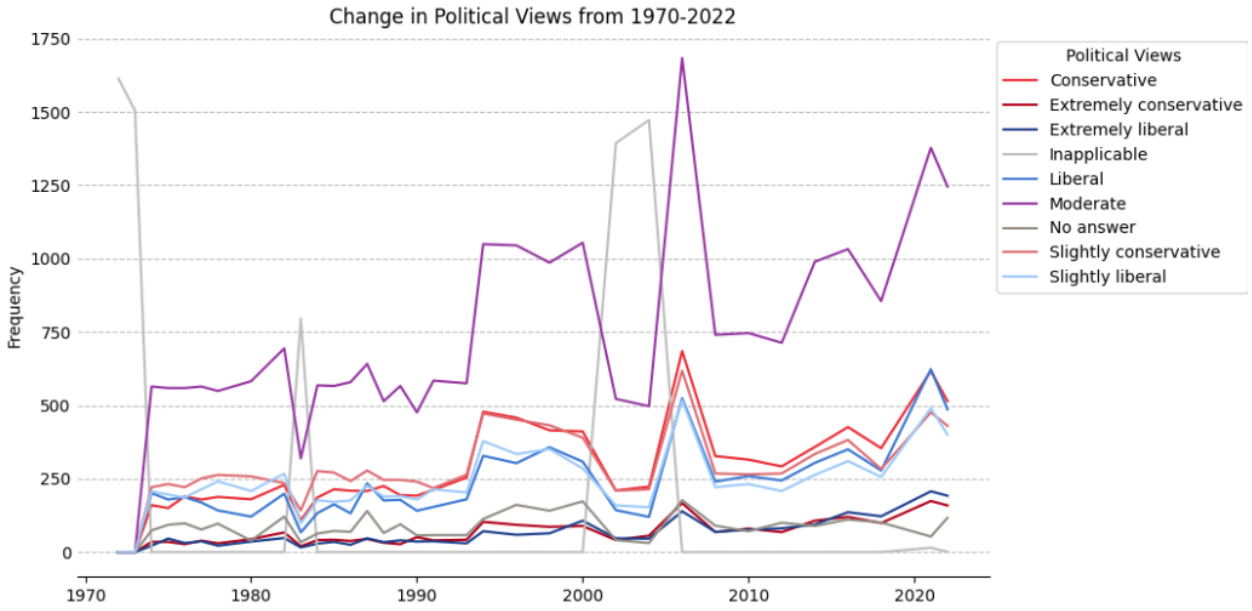
The last numeric summary is a cross table between marital status and labor force status. The "married" category holds the highest frequency for "Has job, but not actively working", "Keeping house," "Retired," "Working full time", and "Working part time." For "Keeping house" and "Working full time," the "married" category leads by a very large gap. The "Never married" category holds the highest frequency for "Student" and "Unemployed."

wrkstat	Has job, but not actively working	Keeping house	No answer	Other	Retired	Student	Unemployed	Working full time	Working part time
marital									
Divorced	225	765	3	392	1440	136	444	5379	858
Married	841	6762	11	547	5117	408	796	19461	3653
Never married	352	1126	12	393	808	1579	1155	8292	2187
No answer	1	9	9	2	10	2	1	14	3
Separated	54	403	1	104	238	47	155	1194	245
Widowed	83	1699	0	205	3273	15	70	927	484

The first visual is a empirical distribution function plot that displays the quantiles for the real family income variable. We can see that there is a steep increase below approximately $50,000. After this point, the line begins to level off at the 80th percentile. Of course, the line does increase to show the 100th percentile but it is a slow process to $160,000.



The second is a visual line graph, which shows the change in political views from 1970 to 2022. The frequency is plotted on the y-axis, with each different political view represented by a different colored line. Disregarding the “Inapplicable” category, the “Moderate” view consistently has a higher frequency than any of the other political views.



5. These variables, general happiness, labor force status, marital status, highest education level, political views, satisfaction in financial situation, and real family income, can all be considered correlated in some way or another. Analyzing this data can prove or disprove some stereotypical assumptions that people make about certain socioeconomic classes or outlook on life. According to this data, approximately 52% of people from 1972-2022 would consider themselves pretty happy. In my opinion, this percentage is around what I would expect the number to be. A little less than 30% of people said they were “very happy”, which I think provides a positive outlook on the population. Along with general happiness expectations, the data also confirmed my assumptions that approximately 50% of the population works full time. There is a large gap between working full time, at 48.72% to the next highest percentage, which is retired at 15.04%. This makes sense, because the largest generation, the baby boomers, are either in or reaching retirement age, which explains why this category has a higher percentage than students or unemployed.

Moving on to analyzing the results of cross tables that were presented in the numeric summary section, the general happiness and satisfaction in financial situation cross table is an example where the results were expected, but also some surprised arose. The “pretty happy” category for happiness held the highest frequency for every satisfaction ranking. This was expected for the “more or less satisfied” and “pretty well satisfied” categories, but not the “not satisfied at all” category. It is a bit unexpected that people would be pretty happy in general but not satisfied at all with their financial situation. The rest of the data suggests that these two variables would be positively correlated, however this specific intersection proves otherwise.

When analyzing the real family income variable, it became evident that there might be some outliers affecting the statistical description and central tendencies of the data. From looking at

the labor force status variable, we know that there is a portion of the population that is retired, unemployed, or a student. These groups of people most likely have no income because they aren't working. In this case, they could be considered outliers because technically they could enter their real income as $0. Even though these data points would be included because it represents part of the population, there is a chance that they could be removed from this scenario because they are a special case of students or retired. Regardless of these possible outliers, this data shows the mean real family income as $32,480. This is much lower than the actual average family income has been in recent years, suggesting that either the incomes from the mid 20th century are having a large impact on the mean, or these possible outliers are having a large impact on the mean.