Benjamin Kolber (bk2480)
Jonathan Bofman (jb4175)

April 18th, 2021

Can Money Buy Happiness?

Data Mining Final Project

Introduction

Covid has brought the world to a halt causing us to reevaluate how we, as a society, operate, work, and for the class of 21', graduate. Although online graduation is not ideal given four years of tedious work and expensive tuition bills, it's an eye opening experience for some. Rather than having the hubris and ego rise from a massive, over the top graduation commending students with 4.0 GPA's who painstakingly spent hours on end perfecting their exam taking skills, we will simply log into Zoom for a 45 minute tribute. In our opinion, this is a great and humbling experience, shedding light on what is truly important, and how the glory of grades can dissipate and become as unimportant as they should be. With this in mind, we set out to explore the relationship between what is truly important in life vs. our material desires. Since happiness and materialistic importance are hard to quantify, we decided to ask our parents what was the most critical decision they made that affected their happiness, which they promptly replied was marrying their significant other. whether this was a genuine answer or a forced one by our mothers, we took the initiative and set out to uncover the relationship between money, education and a successful marriage. For this purpose we decided to work with U.S census data, which is data generated by a yearly form sent to every American family with the purpose of producing data about the American people and economy. We hope that this paper will shed some light to current college students, and young grads, about the lack of need to think materialistically about what to study and how much to earn in order to live a happy life. On the other hand, we might prove the contrary and shed despair over this greedy and profane world.

Data

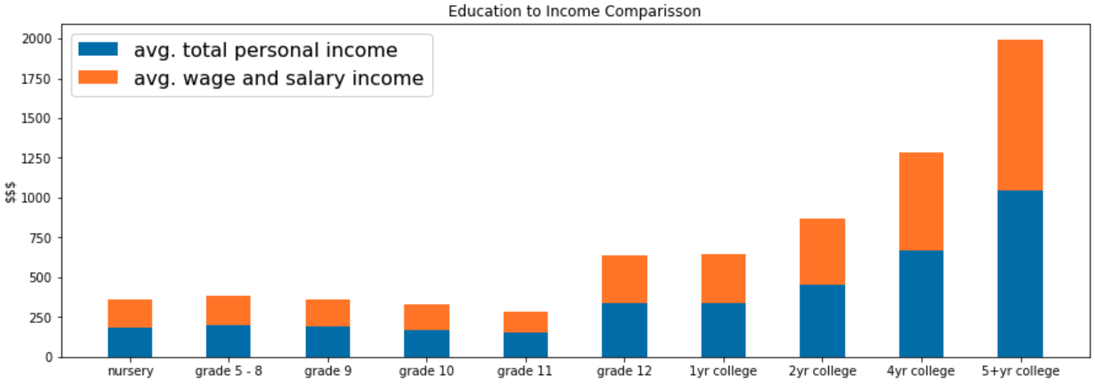
Census data can be found on the IPUMS website[1], which is essentially a massive database of yearly surveys dating back to the 1850's. The database is composed of answers to questions by American families ranging from race, education, and income to disability, times married and number of bathrooms in the household. The IPUMS database offers more than 1,000 variables to select from, which is why we decided to extract the features and years we were interested in and work with a subset from the main database. Due to the range of variables and years of recorded data, we determined this would suffice in terms of complexity over using two distinct datasets. For our initial extract, we decided to look at three main categories: Education, Income and marriage. Within each of these categories, we extracted the sub-features that, intuitively, we thought would make for an interesting exploration. More specifically, in the Income field we decided to look at personal wage, total personal income, welfare status and total family income. In the Education realm we extracted features such as level of education, degree type (if any), and public vs. private school attendance. These were the features we selected after studying the Census data, and staying hopeful of finding some relationship between the marriage features selected, which were number of times married, and current marriage status. These two 'data genres' came from two different demographic datasets, but help us define someone who is married, widowed, or divorced. We extracted the data ranging from the years 2000 to 2020, resulting in a dataset of 14 features and 32 million rows. Moreover, each row has a sub- feature called 'PERWT[2]', which is an indicator as to how many families each row pertains to, or in other words, the 58 million rows would actually represent approx. 5.8 Billion different logs based on the average PERWT across the entire data frame (~100). A summary of the dataset can be found in the 'Additional Materials' at the end of this report.

[1] https://usa.ipums.org/usa/

[2] *"PERWT indicates how many persons in the U.S. population are represented by a given person in an IPUMS sample."* - IPUMS.org

Exploration Part I – Education & Income

We wanted to start off simple, and as such our initial steps of the exploration was to see how much data was missing. We noticed that categories such as number of times married and degree type were missing between the years 2000 and 2008, and hence decided to drop them. In addition, we noticed some of the missing data in total income and total wage was labeled as 9999999 (=N/A), and hence also dropped all logs lacking this information. In total, we ended up working with data from the years 2009 to 2020, with about 34 million rows. Rather than jumping right into analyzing whether education and income directly impact the number of times a person gets married, we first wanted to solidify our trust in the data and our intuition, and took a look at income and education only.



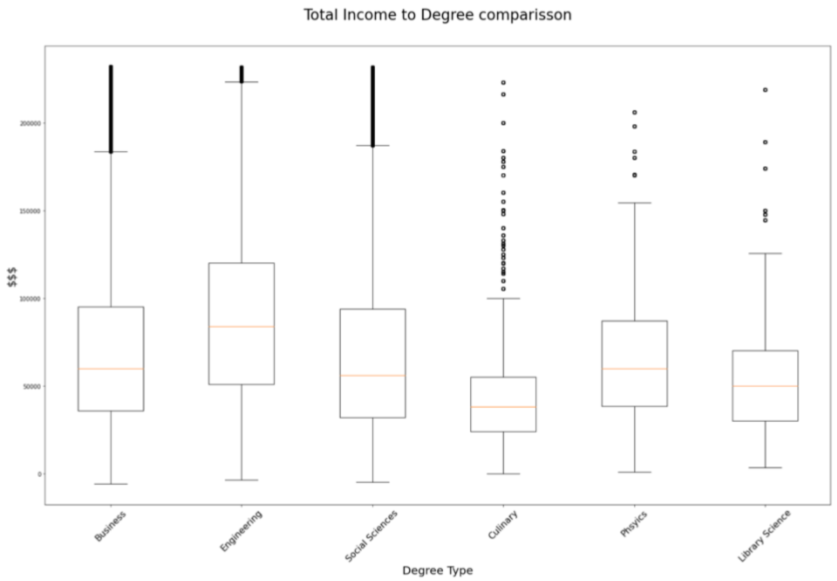
Per our intuition, more years spent in academia results in a significant increase in one's total personal income and salary. In addition, we noticed that the increase in wages for individuals with no undergrad (grade 12) and individuals with an undergrad (4yr college) are similar to the increase in wages of individuals with an undergrad compared to grad students (5+yr college). From this we can deduce that having a graduate degree can increase an undergrad's level of income as much as an undergraduate degree can increase a high school level income.

Comparing education levels and income is interesting, but a college student already knows going to college is a long term lucrative investment, and at this point is wondering how to maximize their long term return. To answer this question, our exploration took us into the meta- of education, comparing different degrees and income levels. As expected, individuals who studied engineering and business, on average, have higher incomes and wages than those who studied arts and education, but what was most interesting to us were the outliers. A good salary is important, but no multi- millionaire lives off of salaries. As such, we created two new metrics to evaluate the chances of becomes a 1- percenter given a degree, basing these metrics on the definition of an outlier as individuals making 200K a year or more per year (IPUMS max income is labeled as 250K+) . With this in mind, we created the following two metrics:

Outliers w.r.t degree = outliers in degree X / all people in degree X * 100

Outliers w.r.t all outliers = outliers in degree X / all outliers * 100

As such, we took note that the degrees with most outliers w.r.t degree were individuals who studied biology (16.4% of students who studied biology now make more than 200K a year), physical sciences (12.83%), social sciences (11.24%) and engineering (11.12%), while the degrees with lowest percent of outliers were Education (1.5%), Theology (1.66%), and Culinary (1.71%). More interestingly, most of the outliers aggregated in degrees such as business (25.84% of all outliers studied business), Engineering (13.25%), and Social Sciences (10.67%), and the least amount of outliers were present in Physics (0.01%), Library Science (0.01%), and Industrial Arts (0%).



Now that we discovered a clear correlation between income and education, and can also interpret what degrees within education can maximize one's long term income, we moved on and tried to find some hidden relationships between a good marriage, income and education.

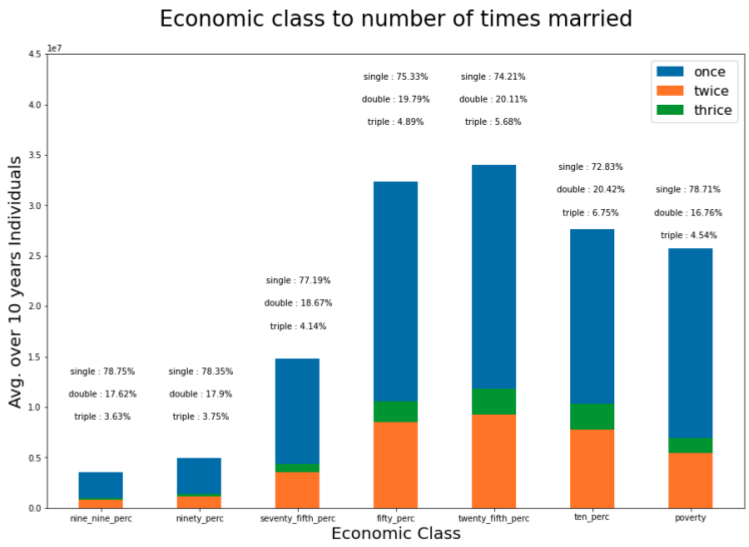
Exploration Part II – Marriage and Income and features

Since we know that education is highly correlated to income, we set out to explore whether income and a good marriage are correlated. From there, we hoped to imply that choice of degree can, in fact, lead to a better marriage. For this purpose we set out to clean out the data further and create new features to better define economic standings and a good marriage. In terms of economic standing, we created a new categorical feature based on the U.S individual income percentiles for 2020[3], described as such:

99th percentile: total income **above 200K**
90th percentile: total income **above 125K**
75th percentile: total income **above 75k**
50th percentile: total income **above 40k**
25th percentile: total income **above 20k**
10th percentile: total income **above 8k**
Rest: total income **less than 8K**

After generating this new feature, we decided to eliminate all individuals under the age of 18 in order to not confuse children, with an income of zero, as individuals in poverty. Using these features, we can now visualize in a vague matter how well each economic class is doing in terms of number of times individuals get married:

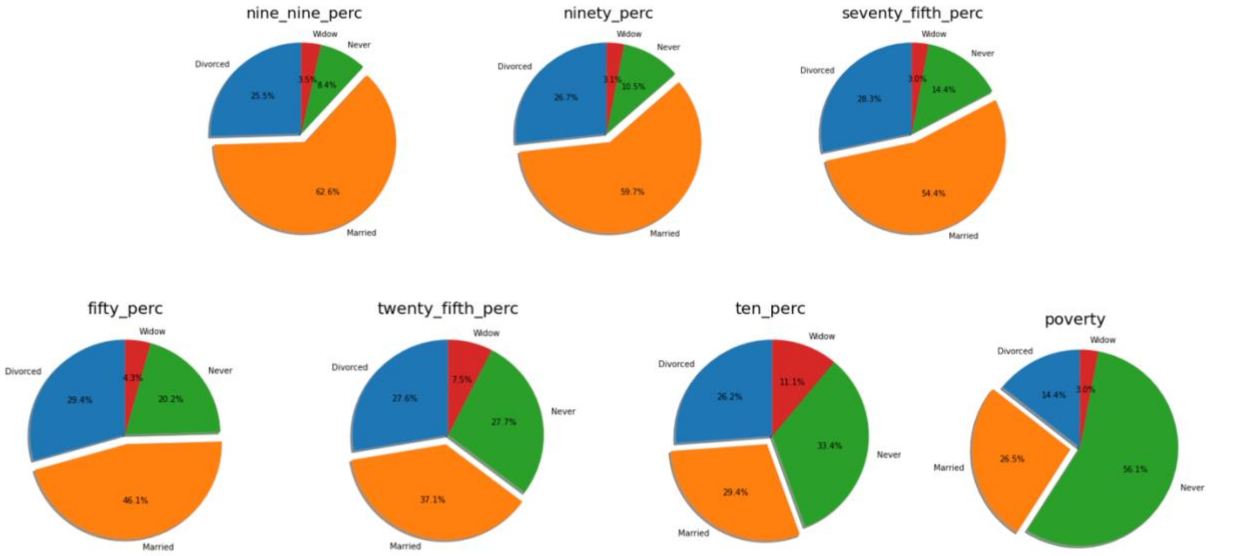
[3] https://dqydj.com/average-median-top-individual-income-percentiles/



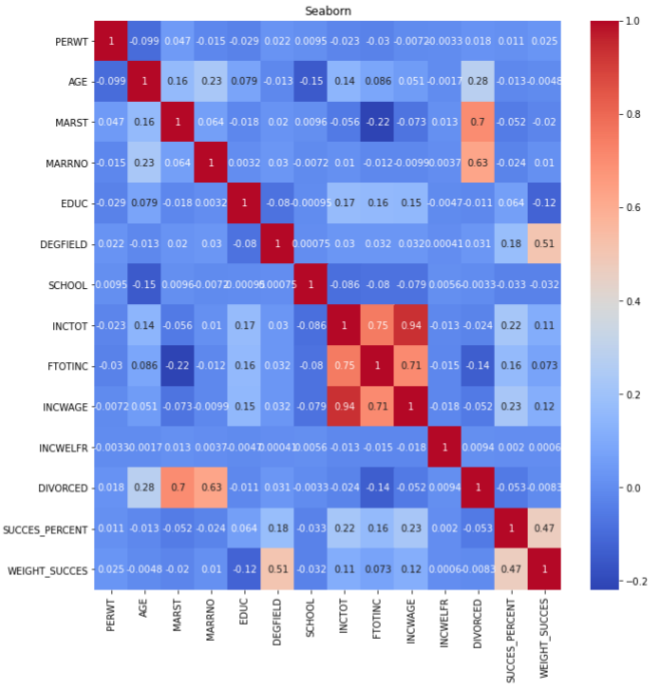
This graph was not very insightful, only as much as showing that there is a similar distribution for number of times marrying (once between 73%-79%, twice between 16%-20% and thrice between 3%-6%) between different economic classes, indicating that number marriages is independent from economic class. We also investigated the inverse to this plot, meaning the distribution of economic classes w.r.t all people married once, twice etc. which also result in a similar, balanced distribution. This made us shift our focus from number of times married to creating a new feature focusing on quality of marriage. Quality of marriage will indicates whether an individual has gotten divorced sometime in their life at least once. For this, we created a new feature by applying logic to the number of times married (**MARRNO**) and marital status (**MARST**), as follows:

```
If MARST == 1 (Married, spouse present) or MARST == 2 (Married, spouse absent):
    Check MARRNO:
        If MARRNO > 1 then we categorize as DIVORCED
        If MARRNO==1 then we categorize as MARRIED
If MARST == 3 (Separated) MARST  == 4 (Divorced), then we categorize it as DIVORCED
If MARST == 6 (Never married), then we categorize it as NEVER.
If MARST == 5 (Widowed), then we categorize it as WIDOW.
```

Below is the distribution of the created features amongst the predefined economic class levels based on percentiles.



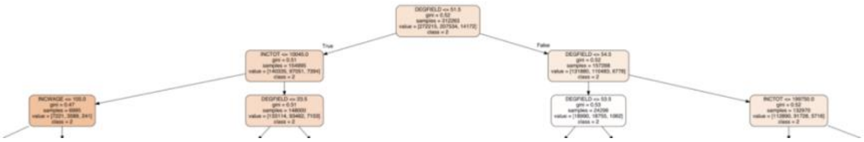
The first thing to notice after plotting the economic classes w.r.t to the new married feature is that as the economic level goes down, the percentage of married people goes down. This is understandable, as people from higher economic standings can afford to get married. Another interesting insight is that divorce rates, as proven above, are also not skewed when it comes to economic classes. A different angle to that statement would be that in the bottom 25th and 10th percentiles have a much larger percent of widows, which is also logical given poorer living and medical conditions. It seems that income and a good marriage might not be related after all, better depicted by the following correlation map:



From the correlation heat map we can notice that, as expected, that the total income and total wage are highly correlated, as is weighted success (percent of outliers w.r.t to all outliers) with degree (since it was generated from the degree distributions). What is also noticeable is the relatively low, but still correlated features of outliers with income, wage and degree. This is also expected as all three factors were used to generate our new features. All in all, the correlation plot did not offer too much insight, beyond what we already knew. This led us to try and train a model and find significant features in a different manner.

Training a Model

For the final part of our investigation, we decided to train some models to see if our new generated feature can be predicted from total income, level of education, degree, income from welfare, and yearly wage. We hoped that these models would attain some good accuracy scores, which then would lead us to interpreting the significant features. For this problem, both Ridge and Lasso performed terribly, indicating that these features might not be as predictive as we thought, until we trained a random forest classifier. We went with classifier over regressor since we were predicting class labels in the form of our new feature. After running the random forest classifier in a fivefold cross validation, we received an avg. accuracy of 50%, with some cases going up to 72%. When investigating the inner working of the tree, we noticed that the most important features, those existing in the top branches, were income wage (0.4 importance) and degree studied (0.33%), as seen in the first layers of our decision tree:



From here we can conclude that in a rough sense, income and degree can in some manner help determine a person’s marriage, but not really. The importance of these features can relate to an imbalanced dataset, or just pure chance since we are dealing with 50 -50 odds across 4 classes.

Conclusion

Starting off with basic and intuitive reasoning, we established that degree type and income are highly correlated. Not only does the degree determine the average income and wage, it can also indicate the chance of a person becoming an outlier, meaning an individual who makes so much they are way above any standard averages. Despite increasing ones chance of becoming rich, economic standing will not influence at all someone's chances of getting a divorce as we have seen. Albeit divorce is not influenced, getting married or even losing a loved one does take effect in lower economic classes, and hence degree type can be a significant factor in determining someone's chances of getting married and living a long and healthy life. All in all, degree will without a doubt affect your income, and income will affect the comfort of your life, which will determine how you will raise a family and the health concerns that will follow. But regardless of material success, I hope we proved that a good marriage cannot be bought, and therefore conclude that money cannot buy happiness, as a good marriage is up to you and you alone.

Abstract

Our initial motivation for this project was combining data science and life science together. Given our canceled semester, quarantine, and lot of time to self- reflect, we wanted to quantify that grades and a good job do not necessarily relate to a well lived life. As described in the 'Data' section, we worked on this problem using US Census data from the IPUMS website, which would be the data mining aspect of this project, tweaking continuously the features we are extracting and testing our hypothesis on. Discovery would be the segmentation's we performed into degree types, outliers of different fields and the new feature we created (DIVORCED) and how it pertains to different economic classes. Once explored, our first approach was correlation, which led to regression, which led to ensemble models in order to better understand the significant features in our dataset. At first, we believed that degree could be correlated to income, which can be correlated to a lower divorce rate. Once we realized that divorce rate is consistent across all economic classes, we realized that the only

actual harm of being in a low economic class is the probability of getting married, and not losing a loved one.

Additional Materials - IPUMS data extract summary

<table>
<tr><th>Features</th><th>Description</th></tr>
<tr><td>PERNUM</td><td>When combined with SAMPLE and SERIAL, PERNUM uniquely identifies each person within the IPUMS. PERNUM is a 4-digit numeric variable which numbers all persons within each household consecutively in the order in which they appear on the original census or survey form.</td></tr>
<tr><td>PERWT</td><td>PERWT indicates how many persons in the U.S. population are represented by a given person in an IPUMS sample.</td></tr>
<tr><td>AGE</td><td>Person age</td></tr>
<tr><td>MARST</td><td>Marital status:
<table>
<tr><th>Code</th><th>Label</th></tr>
<tr><td>1</td><td>Married, spouse present</td></tr>
<tr><td>2</td><td>Married, spouse absent</td></tr>
<tr><td>3</td><td>Separated</td></tr>
<tr><td>4</td><td>Divorced</td></tr>
<tr><td>5</td><td>Widowed</td></tr>
<tr><td>6</td><td>Never married/single</td></tr>
</table>
</td></tr>
<tr><td>MARRNO</td><td>Times married:
<table>
<tr><th>Code</th><th>Label</th></tr>
<tr><td>0</td><td>N/A</td></tr>
<tr><td>1</td><td>Married once</td></tr>
<tr><td>2</td><td>Married twice (or more)</td></tr>
</table>
</td></tr>
</table>

3	Married thrice (or more)

SCHOOL

School attendance:

Code	Label
0	N/A
1	No, not in school
2	Yes, in school
3	Missing

EDUC

Educational attainment:

Code	Label
0	N/A or no schooling
1	Nursery to Grade 4
2	Grade 5-8
3	Grade 9
4	Grade 10
5	Grade 11
6	Grade 12
7	1 year of college
8	2 years of college
9	3 years of college
10	4 years of college
11	5 years of college

EDUCD

Detailed Educational attainment:

Code	Label
061	12 Grade, no diploma
063	Regular High School diploma
064	GED or alternative credential
065	Some college, but less than 1 year
071	1 or more years of college credit, no degree
081	Associate's degree
101	Bachelor's degree
114	Master's degree

115	Professional degree beyond a bachelor's degree
116	Doctoral degree

SCHLTYPE

Public or private school:

Code	Label
0	N/A
1	Not enrolled
2	Public school
3	Private School

DEGFIELD

Field of degree:

Code	Label
0	N/A
11	Agriculture
13	Environment and Natural Resources
14	Architecture
15	Area, Ethnic, and Civilization Studies
19	Communications
20	Communication Technologies
21	Computer and Information Sciences
22	Cosmetology Services and Culinary Arts
23	Education Administration and Teaching
24	Engineering
25	Engineering Technologies
26	Linguistics and Foreign Languages
29	Family and Consumer Sciences
32	Law
33	English Language, Literature, and Composition
34	Liberal Arts and Humanities
35	Library Science
36	Biology and Life Sciences
37	Mathematics and Statistics
38	Military Technologies
40	Interdisciplinary and Multi-Disciplinary Studies (General)

	41	Physical Fitness, Parks, Recreation, and Leisure
	48	Philosophy and Religious Studies
	49	Theology and Religious Vocations
	50	Physical Sciences
	51	Nuclear, Industrial Radiology, and Biological Technologies
	52	Psychology
	53	Criminal Justice and Fire Protection
	54	Public Affairs, Policy, and Social Work
	55	Social Sciences
	56	Construction Services
	57	Electrical and Mechanic Repairs and Technologies
	58	Precision Production and Industrial Arts
	59	Transportation Sciences and Technologies
	60	Fine Arts
	61	Medical and Health Sciences and Services
	62	Business
	64	History
INCTOT	**Total personal income**	
FTOTINC	**Total family income**	
INCWAGE	**Wage and salary income**	
INCWELFR	**Welfare (public assistance) income**	