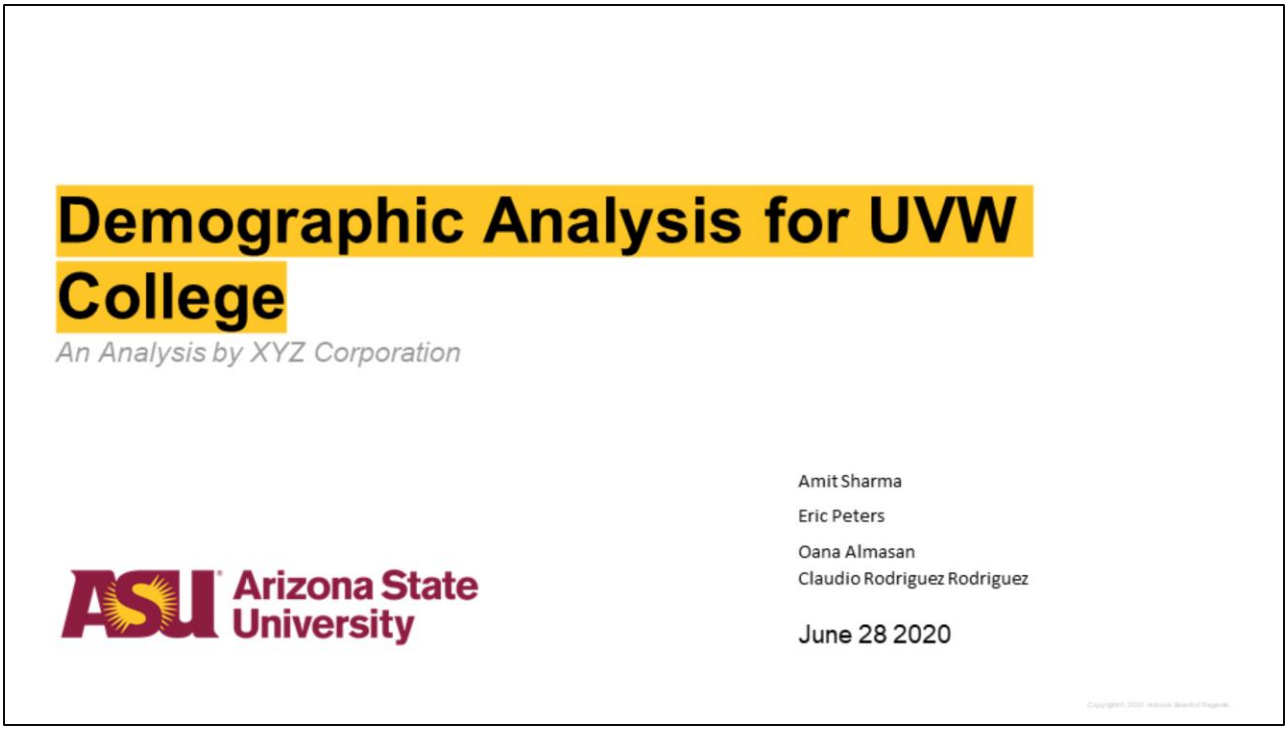


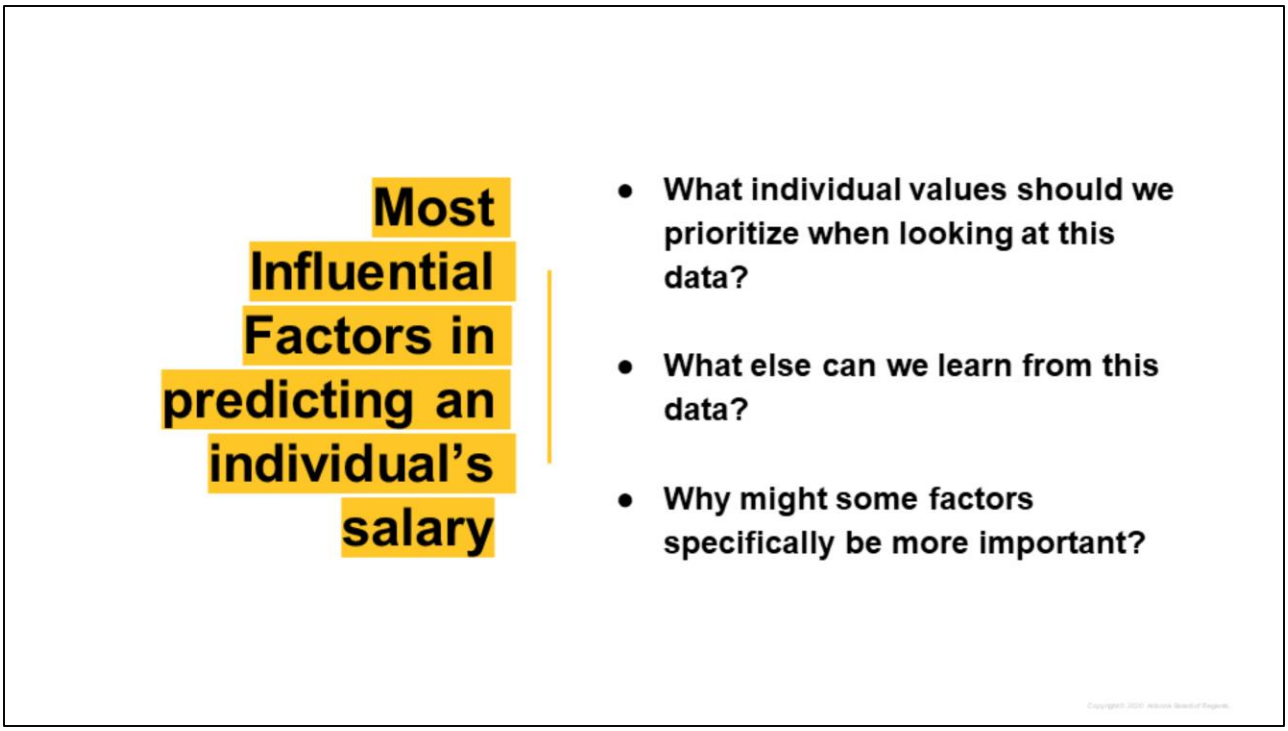
Using data from the 1994 census data, we have compiled a breakdown of the population based on a number of key factors. We found that the educational attainment of an individual had one of the highest impacts on salary, and have specifically broken out data around those points. We also examined job habits, such as job type and hours worked - and regional data, such as country of origin and ethnicity.

Summary of Findings

- **Interest in College Programs start at 18 years for all Races and Nationalities**
- **Higher Salaries are seen in White and Asian races**
- **Women have a higher degree of obtaining a Doctorate**
- **People from Asia have a higher chance of finishing their Bachelors**

Looking at the raw data, we set out to determine what values play the strongest role in determining whether or not a given individual will make over or under 50K/year. We used a machine learning algorithm to examine the data and determine which factors played the largest roles.

- People from Latin America and Caribbean have a lower education level
- High correlation with finished Bachelors and higher Salary
- Dropouts rates are high for all races

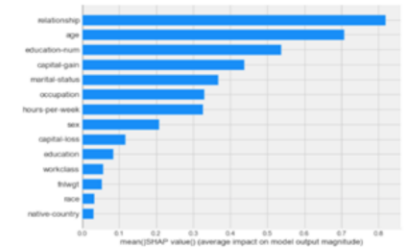


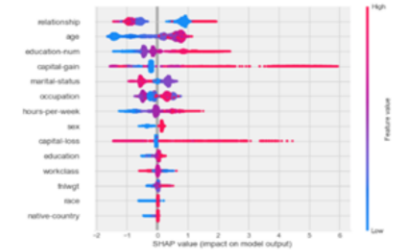
Looking at the raw data, we set out to determine what values play the strongest role in determining whether or not a given individual will make over or under 50K/year. We used a machine learning algorithm to examine the data and determine which factors played the largest roles.

Most impactful predictors of salary

We can see that relationship (the role an individual plays in their family) is most impactful when predicting salary.

Age and Educational Attainment closely follow.

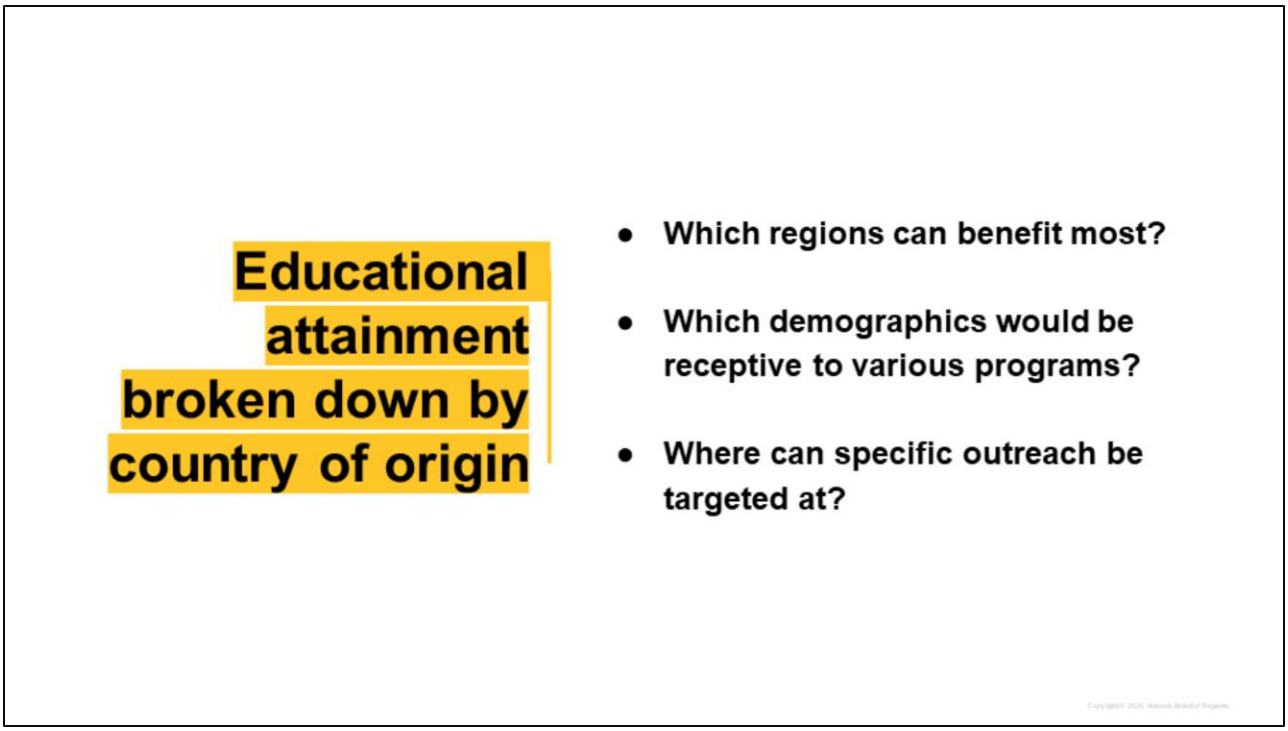




	column_name	shap_importance
7	relationship	0.746656
0	age	0.626939
4	education-num	0.515205
10	capital-gain	0.433327
5	marital-status	0.369107
6	occupation	0.340632
12	hours-per-week	0.339498
9	sex	0.200037
11	capital-loss	0.126448
3	education	0.0864223
2	fnlwgt	0.0674727
1	workclass	0.0562161
8	race	0.0274393
13	native-country	0.0158496

The breakdown of values here is very clear. We can say, with some certainty, that there is some data relative to the relationship (or family role) that plays a significant role in determining if an individual will be making more or less than 50k - as covered in the system documentation, this may be a factor of misreporting; it appears that husbands/wives are *significantly* more likely to earn >50k, which may indicate that those numbers reflect household income, and not individual income. Beyond that, we see that age, educational attainment, and career-specific factors come into play.

We'll take a look at these factors more closely.



Looking at educational attainment (or years of education) by country of origin, we can get a sense of which regions may specifically be more or less likely to have individuals making >50k
We can also use this data to generate more specific marketing based on nationality, or deliver location-based marketing

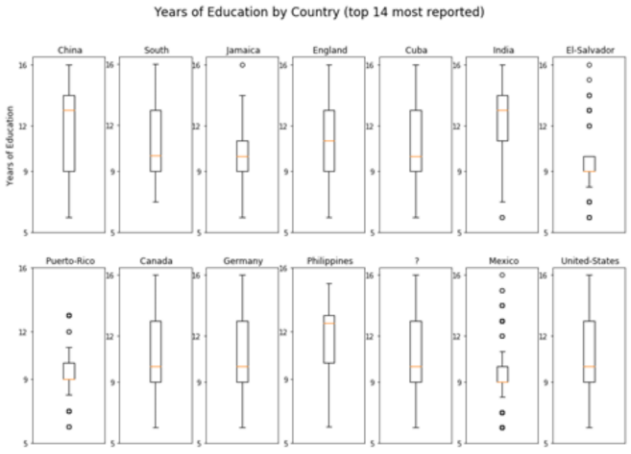
Average level of educational attainment by region (country of origin)

This graph indicates the average level of education, by region. This also lets us see which regions are being reported by our data. A majority of the regions reflect some HS or HS-grad, which is in line with the data that we have.
Let's take a look at the top 14 locations represented

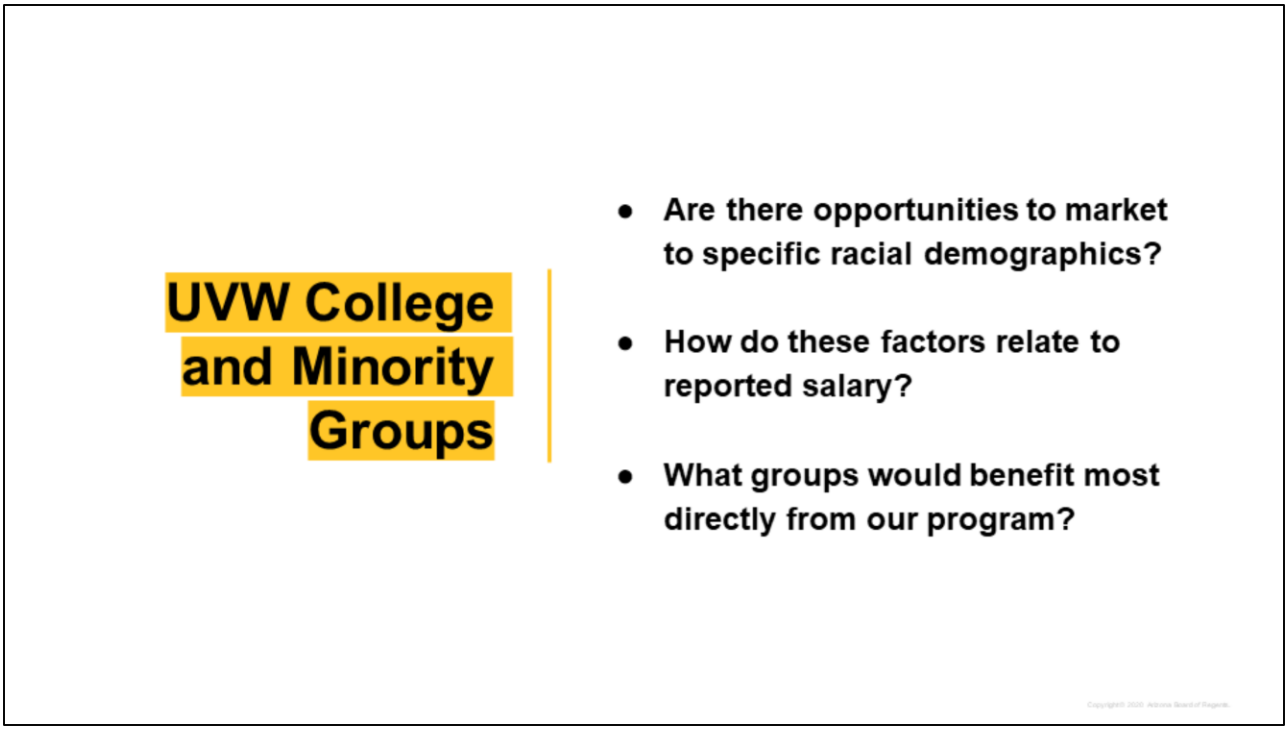
Quartile breakdown of the most common Countries of origin

Of the 42 reported countries of origin, these 14 had the highest response rates

The trends predicted here are more likely to be trends of the whole population



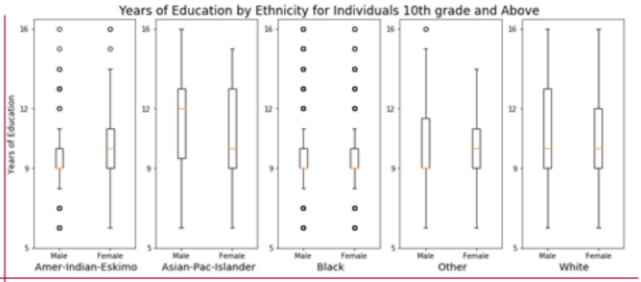
Here we see the locations which had the highest number of reported values. We can see some strong trends in this data, with China reflecting an overall higher level of education, and locations like Puerto Rico reporting a majority of individuals with a high school degree or lower. Due to the numbers reported for these locations, we can more readily expect these to be more representative of their populations

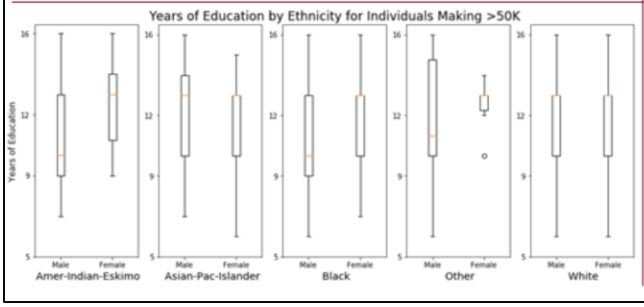


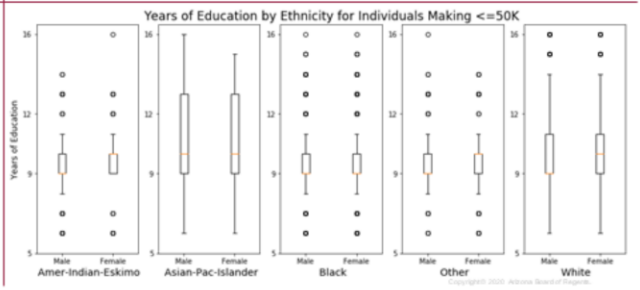
Hand-in-hand with country of origin - we decided to analyze the data on the metric of ethnicity and gender, relative to income and educational attainment
This helped us to identify any areas where there were clear trends based on ethnicity or sex

Quartile breakdown by Ethnicity and Salary

Individuals with more years or education are regularly on the high end of Salary, with some outliers showing in the <=50K range.





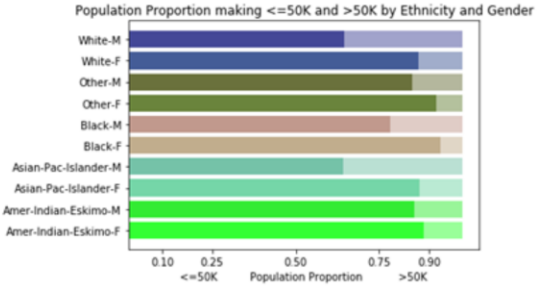


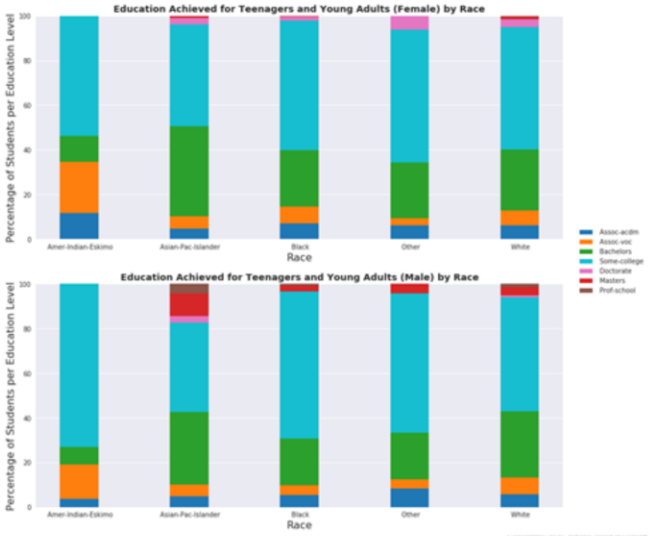
All data here is representative of individuals with 10th grade or higher education.

To tackle business concerns with ecological fallacies, we broke down the dataset with more specific targeting of gender groups.

Salaries and educational breakdown by Race and Sex

Minority Groups seem to have a higher degree of Dropout rates.



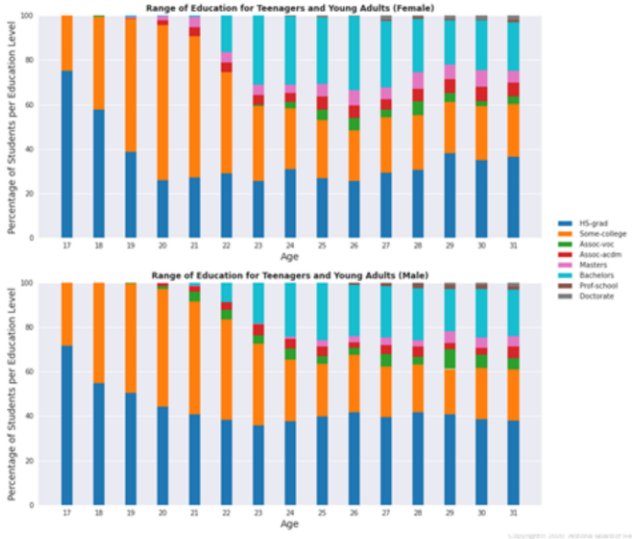


All data here is representative of individuals with a high school degree or higher.

We can observe that Amer-Indian-Eskimo Group starts College but they have a hard time finishing it.

Critical Ages to target College Programs

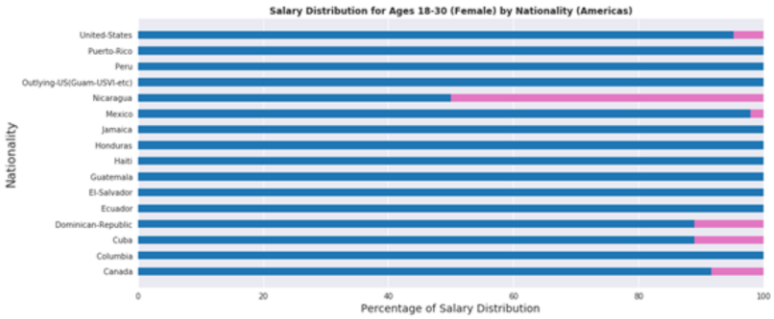
- Biggest increase in Bachelor programs starting around age 17 to 18
- High School demand is stable across all age groups but dropout rate is high.
- Women show demand in Masters Programs.



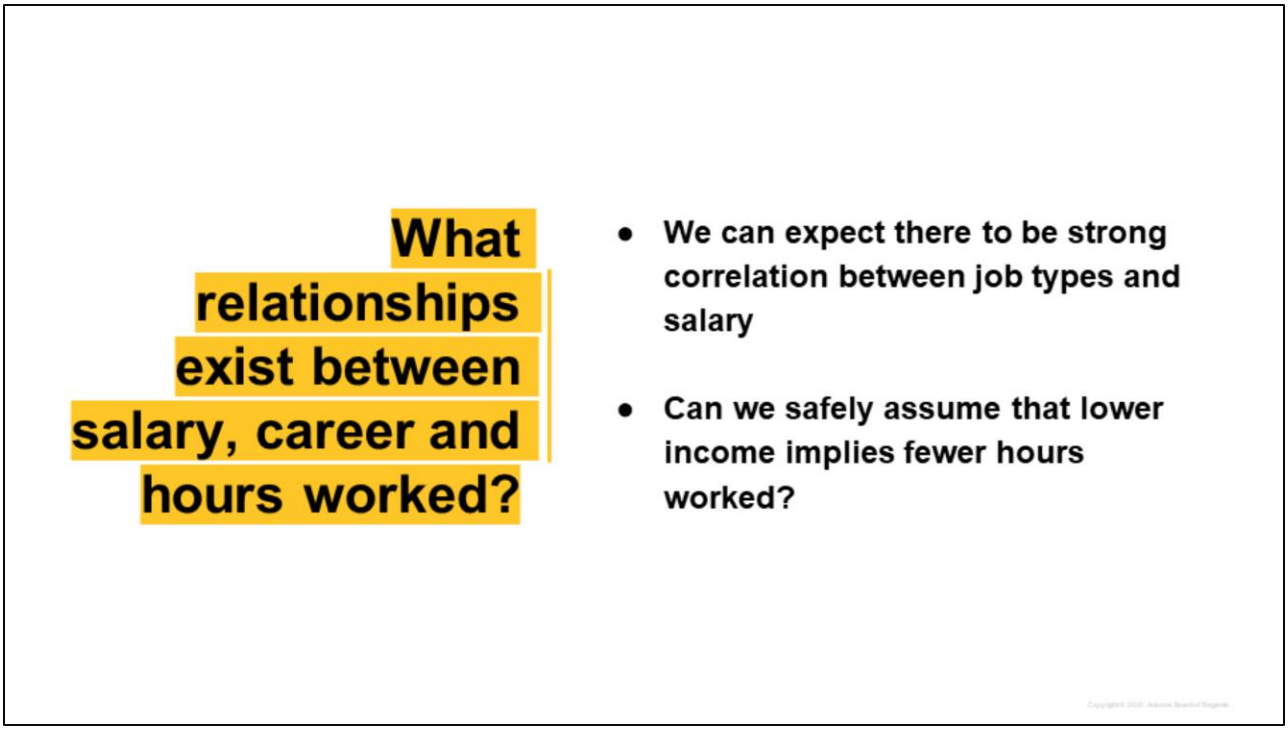
We observe the biggest increase in Bachelor programs starting around age 17 to 18, which is consistent with the common age High School ends, but we see High School maintaining popularity over the years. Finally we observe an increase in seriousness of Masters Programs for Women in their early twenties.

Minority Groups of Interest for UVW College

- American Countries contain the lowest salary rating for both Male and Female Populations.
- We also have Eastern European Countries as minority groups but our dataset didn't show African Minority Groups.



A minority group would be a group of people that has a disadvantage compared to other groups. They would be in the same age groups as other groups looking for College Education (18-30 years old).



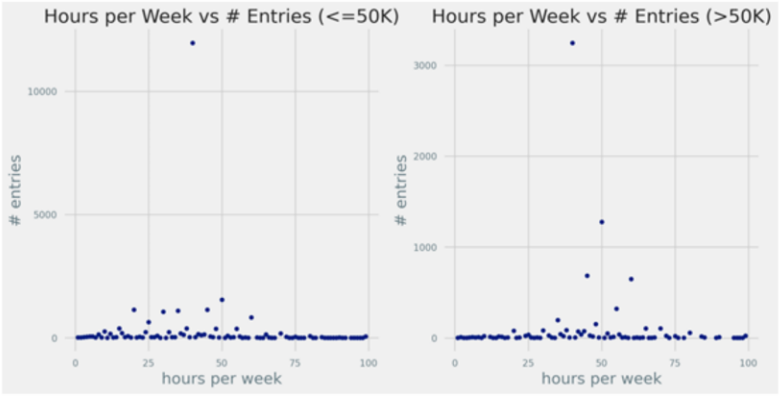
We have examined the data based on job type and hours worked in a week, to find areas where specific groups may benefit from college-level programs.
Typically, we expect lower-wage positions to require less work, but can we conclude that individuals making less than 50k have more time available during the week?

Hours worked per week, by salary

Both graphs indicate a large number of individuals working 40 hours per week

Individuals making less than 50K a year: can expect to be working 20-60 hours per week

Individuals making more than 50K a year: can expect to be working 40-60 hours per week



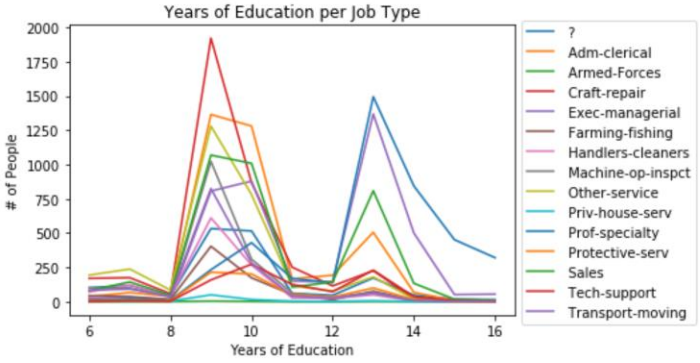
Looking at the two graphs, we can see that both groups have a very large number of individuals working 40 hours per week.

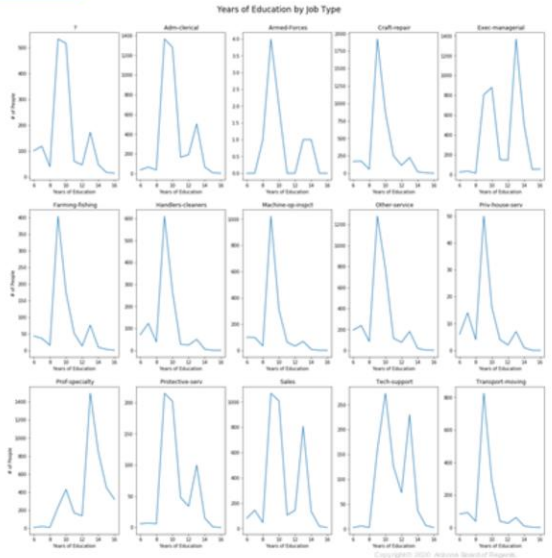
Looking at the distribution of points other than that value, however, yields the following:

People making less than 50k are fairly evenly spread between 20 and 60 hours. This could be the result of one or more part-time jobs, or long-hour low paying jobs

People making more than 50k primarily fall between 30 and 60 hours, with a strong majority between 40 and 60. These numbers are likely indicative of salaried positions.

Education and type of career





Looking at the peak at the 9 year mark, we can see that craft-repair is the most common job for high school graduates, strongly represented for that level of education and weakly represented for all higher levels

Transport, machine op, and handlers/cleaners follow a very similar trend. Individuals in these roles are significantly less likely to have any higher education and have the lowest proportional numbers of individuals making >50K

Questions?

Questions about the data or the methodology behind the analysis