Adhithya Unni Narayanan
Tobias De Locht

KU LEUVEN

Project: Data visualization in data science

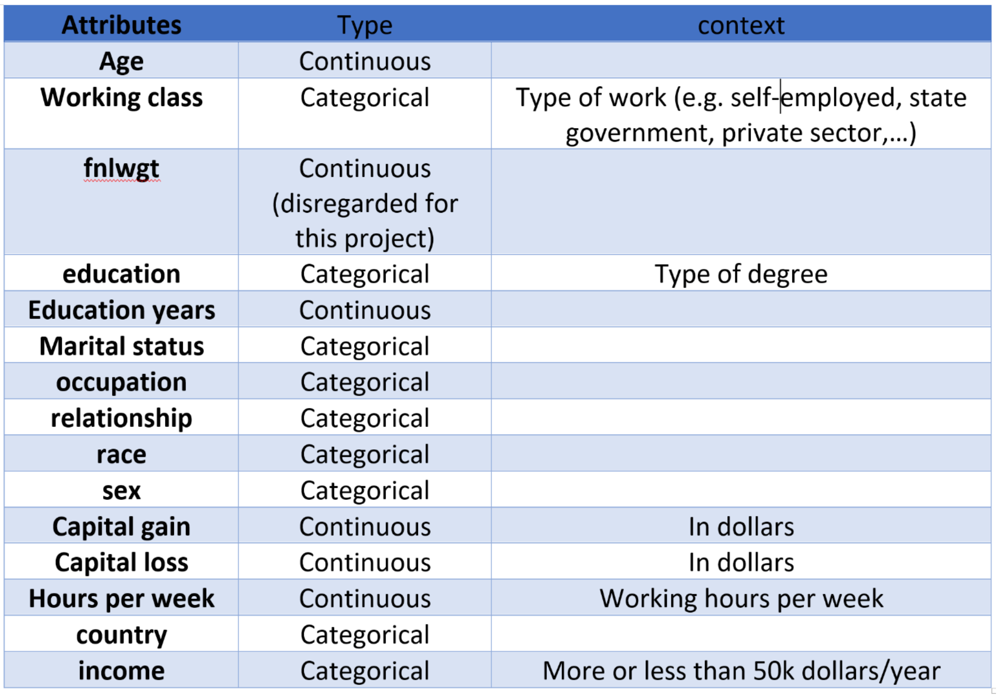
August 29
2021

I. Introduction

The amount of working hours has always been a point of discussion. It has been influenced by many events and ideologies throughout history and around the world. There has been the introduction of the 5-day work week or the fight for the women's right to work for example. But there are also some widespread thoughts about working like women work less than men or the paradox that on one hand immigrants steal jobs because they work more and cheaper but on the other hand that they don't work and live off the state. The aim of the project is to research which factors statistically have influence on working hours.

II. The data

The data for this project is obtained from UCI machine learning repository[1]. This data was extracted from the 1994 Census database with the intended use of predicting income (above or below 50k/year). We will however use the data to determine influential attributes on working hours. We will also be looking if the data reflects anything on the prominent question of disparity of working hours and income for different sex, races and education level. It's a multivariate dataset with both continuous and categorical attributes. Figure 1 is a summation of the data variables.



Attributes	Type	context
Age	Continuous	
Working class	Categorical	Type of work (e.g. self-employed, state government, private sector,…)
fnlwgt	Continuous (disregarded for this project)	
education	Categorical	Type of degree
Education years	Continuous	
Marital status	Categorical	
occupation	Categorical	
relationship	Categorical	
race	Categorical	
sex	Categorical	
Capital gain	Continuous	In dollars
Capital loss	Continuous	In dollars
Hours per week	Continuous	Working hours per week
country	Categorical	
income	Categorical	More or less than 50k dollars/year

Exploratory Analysis:

From the exploratory analysis conducted on the number of observations(data points) the most recurring country is the United States(about 29,000 out of 32,000 data points). The bar graph showing the number of data points per country seems to be skewed because of this reason. There is also an addition of an unknown field in the country variable '?' which is eliminated in the further analysis.

There is also interest in discovering possible correlations between the data attributes, figures 2 through 4 show some of the correlations made.

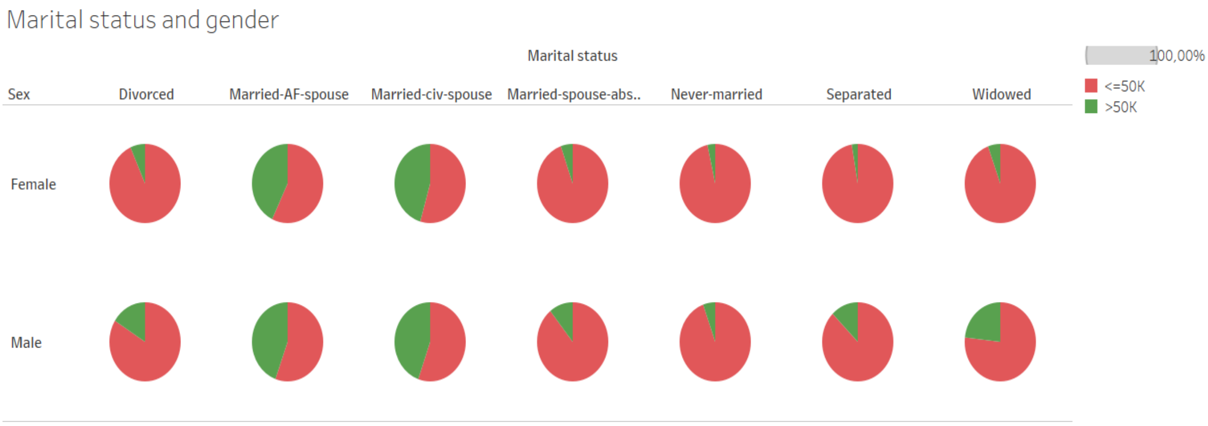


Figure 2: Marital Status and gender

Figure 2 shows the difference in earning capacity for males and females with respect to their marital status. It is pretty evident that for most of the marital status, males have more earning capability than females.

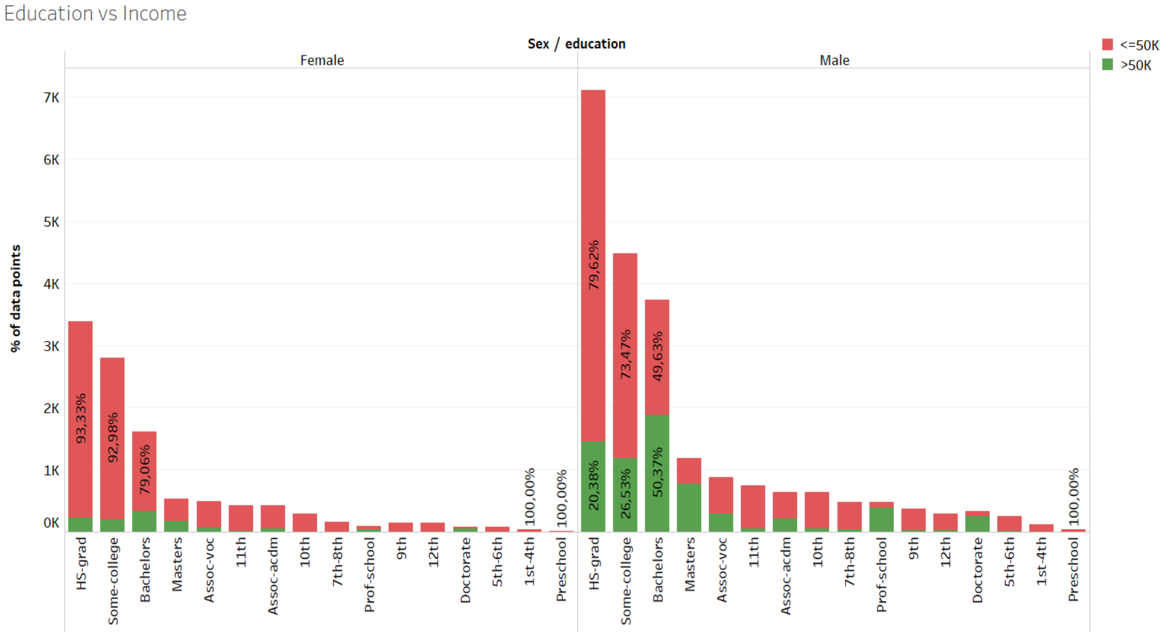


Figure 3: Education v/s Income for males and females

Figure 3 adds to the findings regarding the difference in the earning capacity for females when compared to males even when they are equally qualified. The Figure shows that 20% of HS-graduate males are likely to earn >50k while only 6% of females are able to earn >50k with the same level of education.

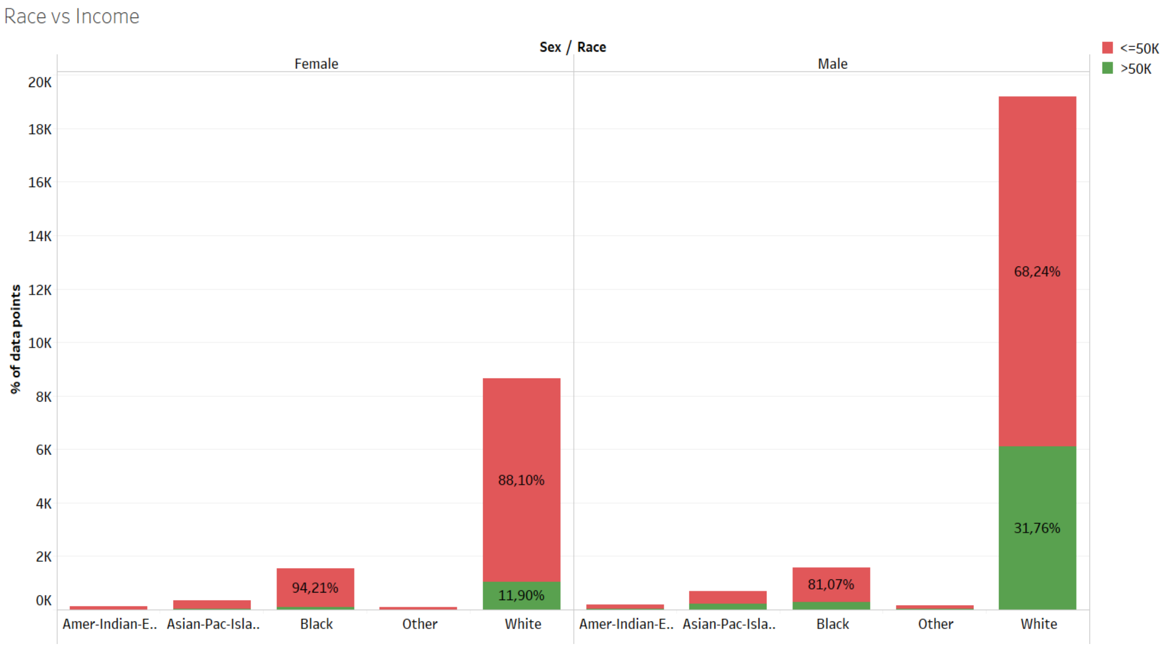


Figure 4: Race v/s Income for males and females

Apart from gender, race is one important factor that insinuates inequalities. Figure 4 shows Race v/s income for both the genders. It can be seen that gender when combined with race is a huge factor that contributes to differences. Here, it can be seen that when 12% white females are likely to earn >50k when 32% white males are. But only 5% black women are likely to earn >50k while 19% of black men are capable of doing the same. Similar trend prevails for all the races.
Note: This summary can also be biased because most of the data points in the dataset are from the United States and the race distribution is not equal throughout the dataset.

III. Research Questions:

The aim of this project is to answer the question which features are determinant for the amount of working hours. We start from 3 ideas on working, the idea that men work more than women, People in western civilization work less than people in other parts of the world and finally that a higher level education leads to working less.

The idea that men work more than women is historically based on the fact that women have been responsible for childcare and household. Still in modern day society this is often still seen as the role of the woman.
The second idea comes from the strong democratic and social organization of western civilization. A consequence is the strong presence of union representation. This in regard has led to many rights of the working force
The third idea is that people with a higher education are in high demand on the market so they have a strong negotiation position and higher salary. This leads to the possible conclusion that therefore they can work less than people with a lower education.

The questions asked in this project are if these three ideas can be supported by statistical evidence or if there are other factors in play.

IV. Exploring the design space

The exploration of the design space has been done on the three ideas on working hours. Each of these statements will be analysed and from there other potential influencing features will be researched.

Delving into the first statement regarding the notion of men working more than women, the data indicates that men work for 42.43 hrs per week on average and that of females is 36.41. The distribution of the working hour disparity is given in the map below. From the map (Figure 5), it can be said that the data supports the disparity in the working hours with respect to gender but it should be noted that the number of female observations in the dataset is less compared to males.

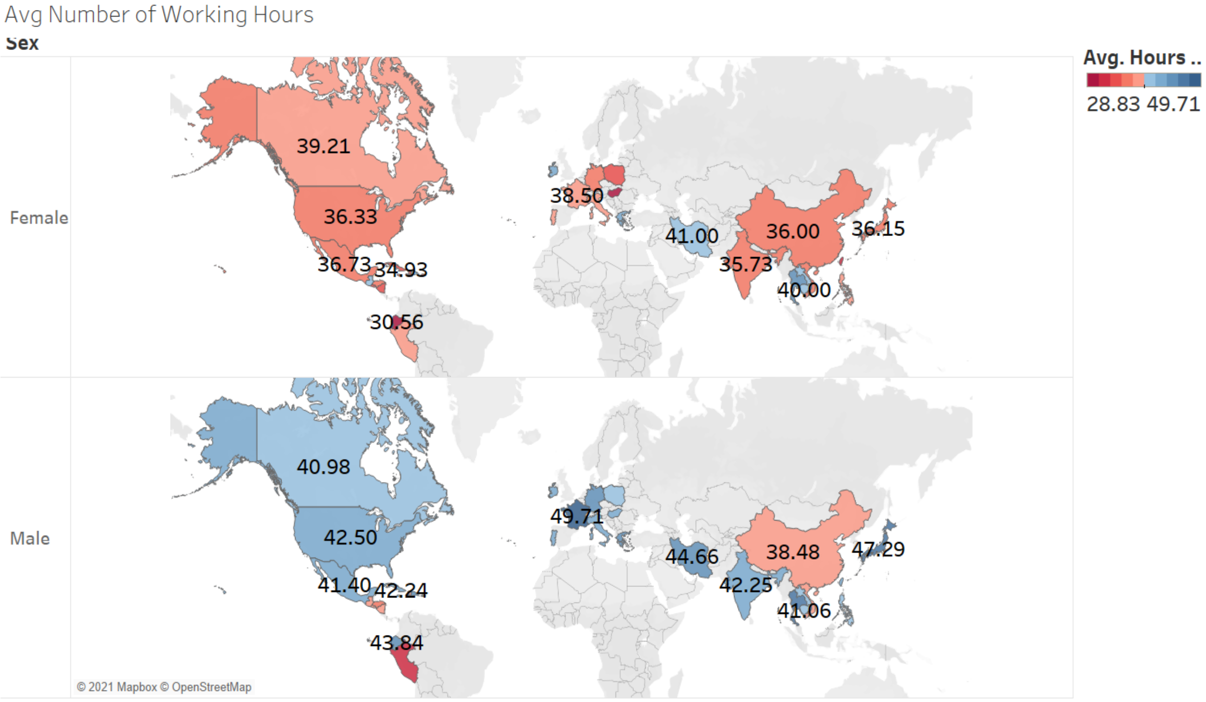


Figure 5: Avg number of working hours by gender

Moving into the notion of less working hours for Western countries, it is shown that countries like France and Greece have the highest average. Seconding them are the countries like Japan and Iran(Figure 6).

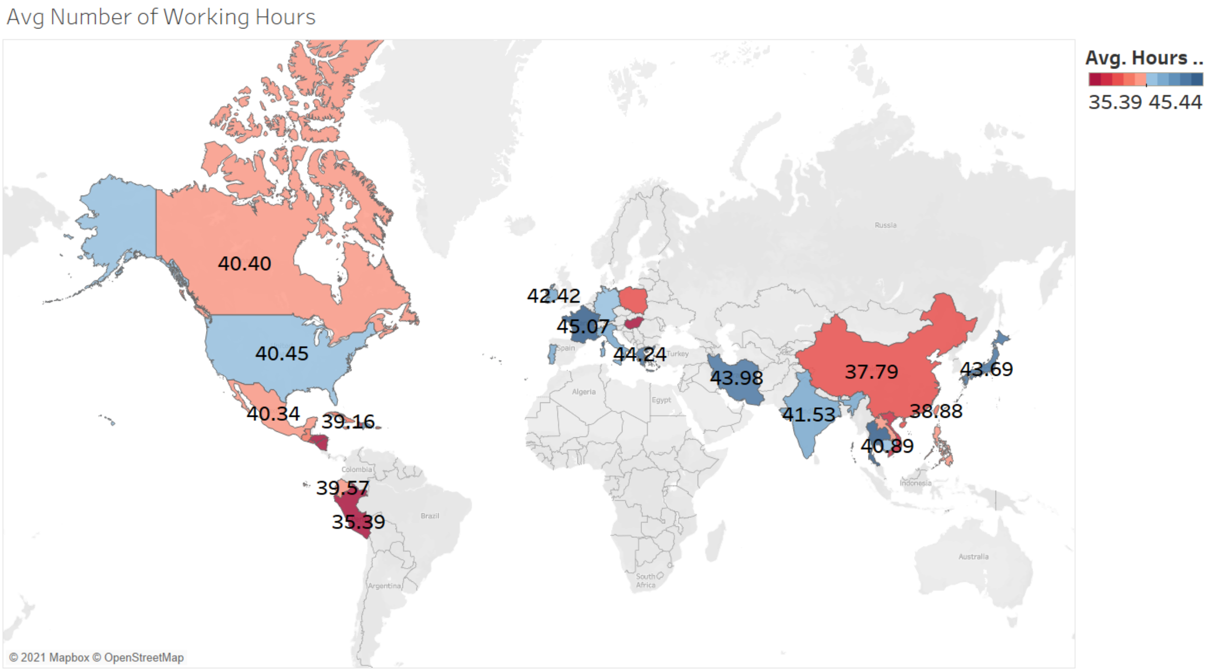


Figure 6: Average number of working hours

But it is to be noted that these basic statistics for average number of working hours cannot be taken as a reliable reference since it is rather evident that the data is skewed.

For exploring the final statement of educational level having an influence on the working hours, a simple linear model is devised keeping the variable average working hours as response variable and avg education in years as explanatory variable.

The model is as follows:

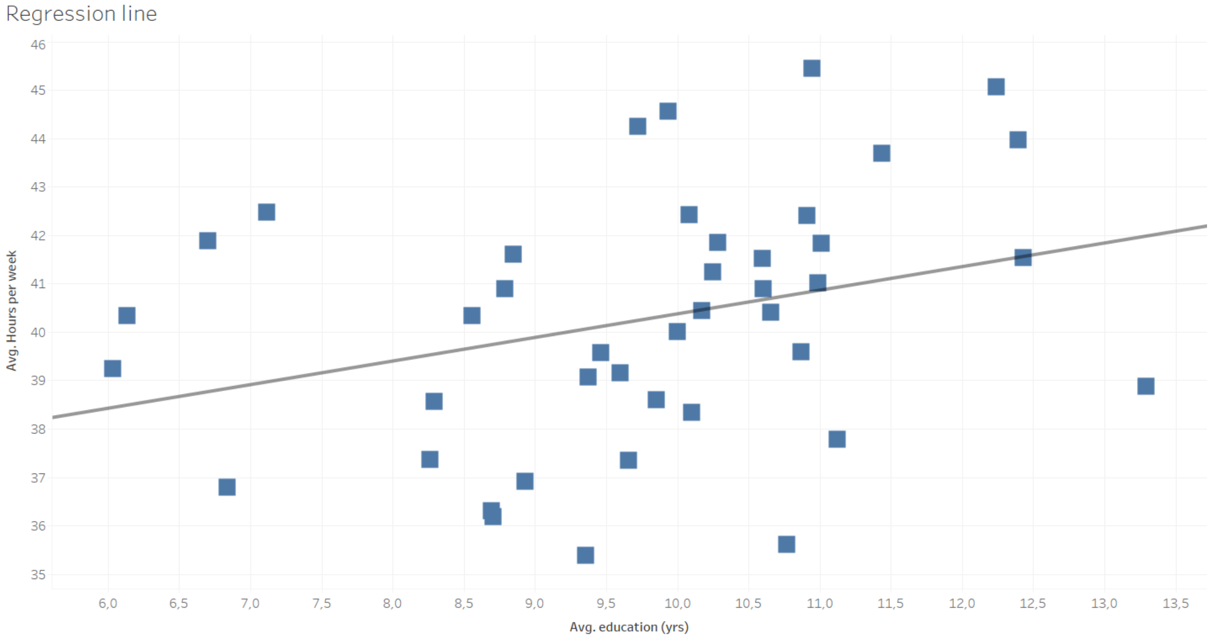
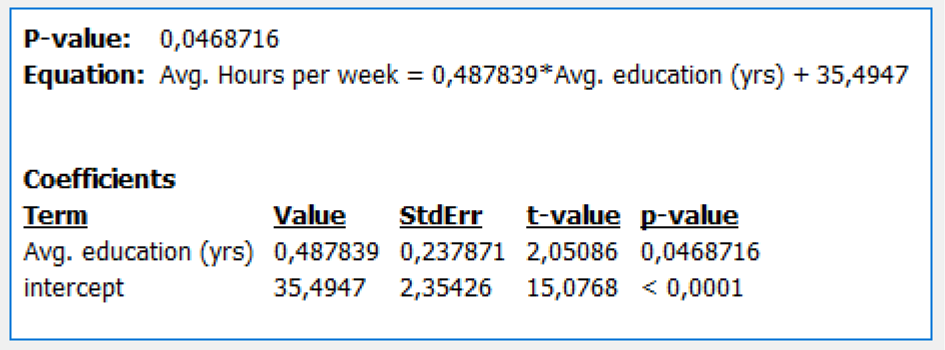


Figure 7: Regression line

The regression values is as shown:



P-value: 0,0468716

Equation: Avg. Hours per week = 0,487839*Avg. education (yrs) + 35,4947

Coefficients

Term	Value	StdErr	t-value	p-value
Avg. education (yrs)	0,487839	0,237871	2,05086	0,0468716
intercept	35,4947	2,35426	15,0768	< 0,0001

Figure 8: Summary statistic of simple linear regression

The fit summary indicates that the Average education in years significantly influences the Average hours per week but the coefficient of determination is 0.09 which establishes the fact the correlation between the 2 variables is very very small. Since these results are a bit contradictory we will further investigate the effect of education but in addition with other attributes like sex, occupation and income. We looked at the income variable to see if there exists any difference in the income earned by the 2 genders. It is striking that a huge difference in the income can be noticed for the same occupational sector between males and females. For example, when 60% of exec-managerial male earn >50k USD per year, only 25% of females from the same occupation earns the same amount. This trend prevails in almost all the occupations and it is pretty evident from figure 9.
While on the gender disparity topic, there are many paradigms such as age, marital status to explore and see if there are any interesting relationships that are prevailing. So the design space exploration has led us to see if the latter exists. This has also inspired us to explore more variables which are specified in the final design section.

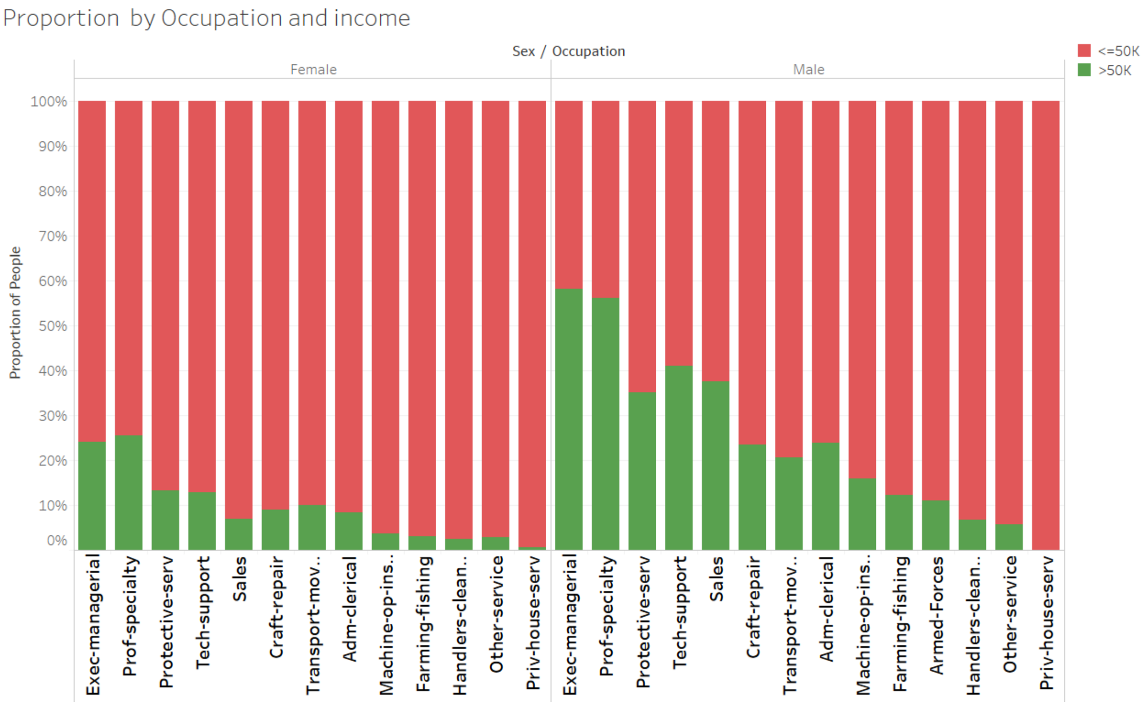


Figure 9: proportion by occupation and income

The result of the first analysis shows that the first idea is indeed backed up by statistical evidence. The second idea on the other hand has no merit after analysis. The third and final idea has given the most ambiguous results. This will be further explored by introducing other variables into the equation. The final result of this analysis will be explained in the next section.

Finally our results were compared with similar statistics produced by ourworldindata[2] and eurostats[3]. This led to some insights, especially since these statistics are based on more up to date datasets. One aspect worth noting is that there is a significant correlation between hours of work and GDP per capita. This attribute isn't present in our dataset but is clearly an influential feature on working hours. A second discovery is a clear evolution of less working hours in europe. This can explain why the older dataset for this project didn't confirm with the notion that people in western civilization work less. This asks for follow-up analysis.

V. Final designs

On discussing the gender aspects and disparities, an interesting association came on the exploration of the dataset with the variables Average hours per week and average age for both males and females separately. The trend line is as shown:

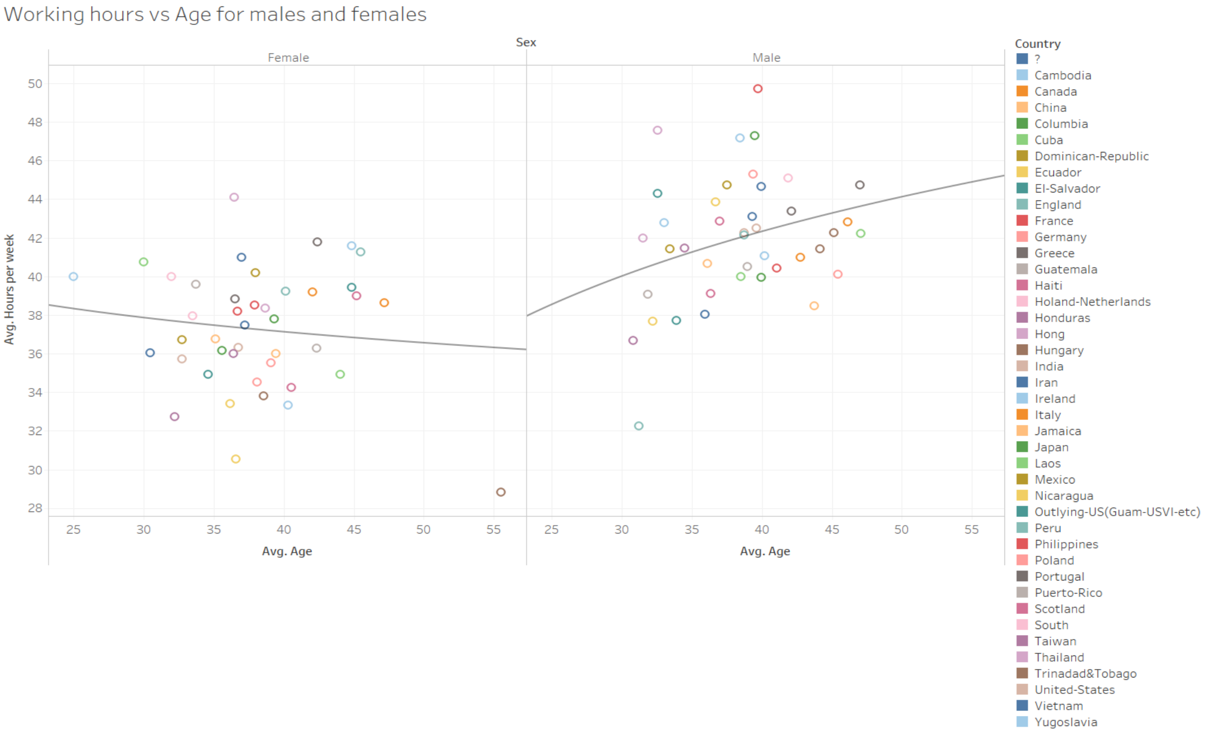


Figure 10: Average working hours v/s average age for males and females

It is found that for males, there is an increasing trend for average working hours with age while for females, there is a decreasing trend.

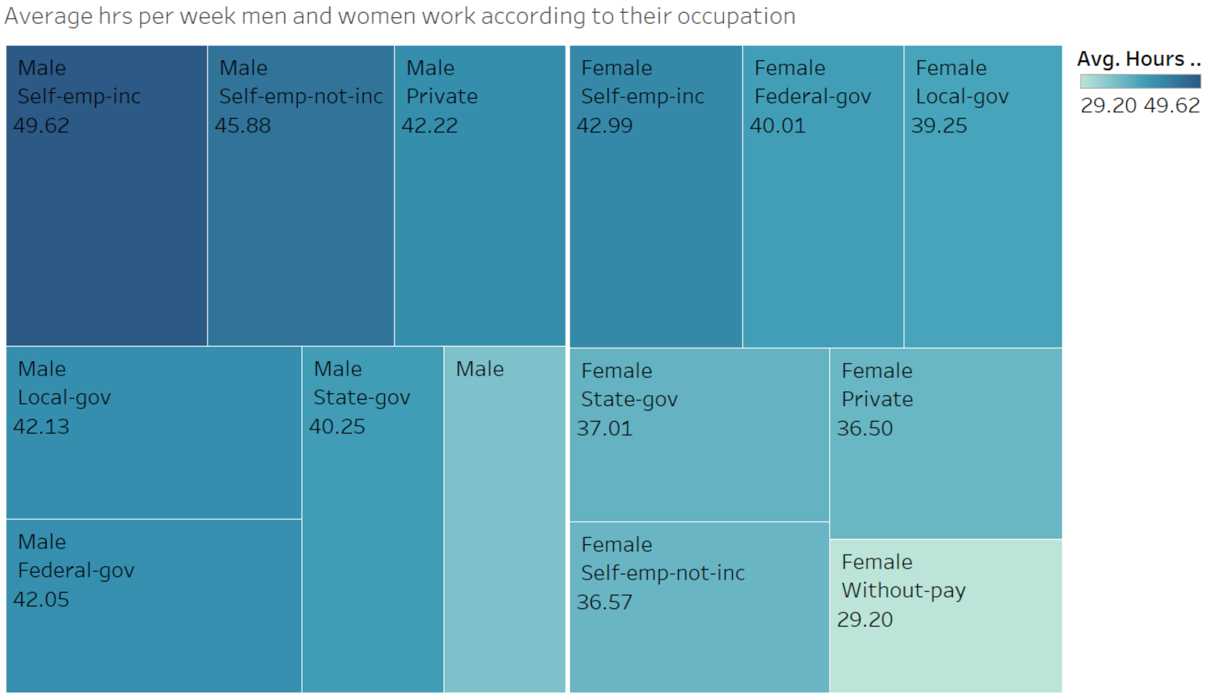


Figure 11: Average hours per week for men and women according to the occupation

The treemap in Figure 11 indicates the average number of hours men and women work according to their occupation and it's evident that for similar jobs, women work less.

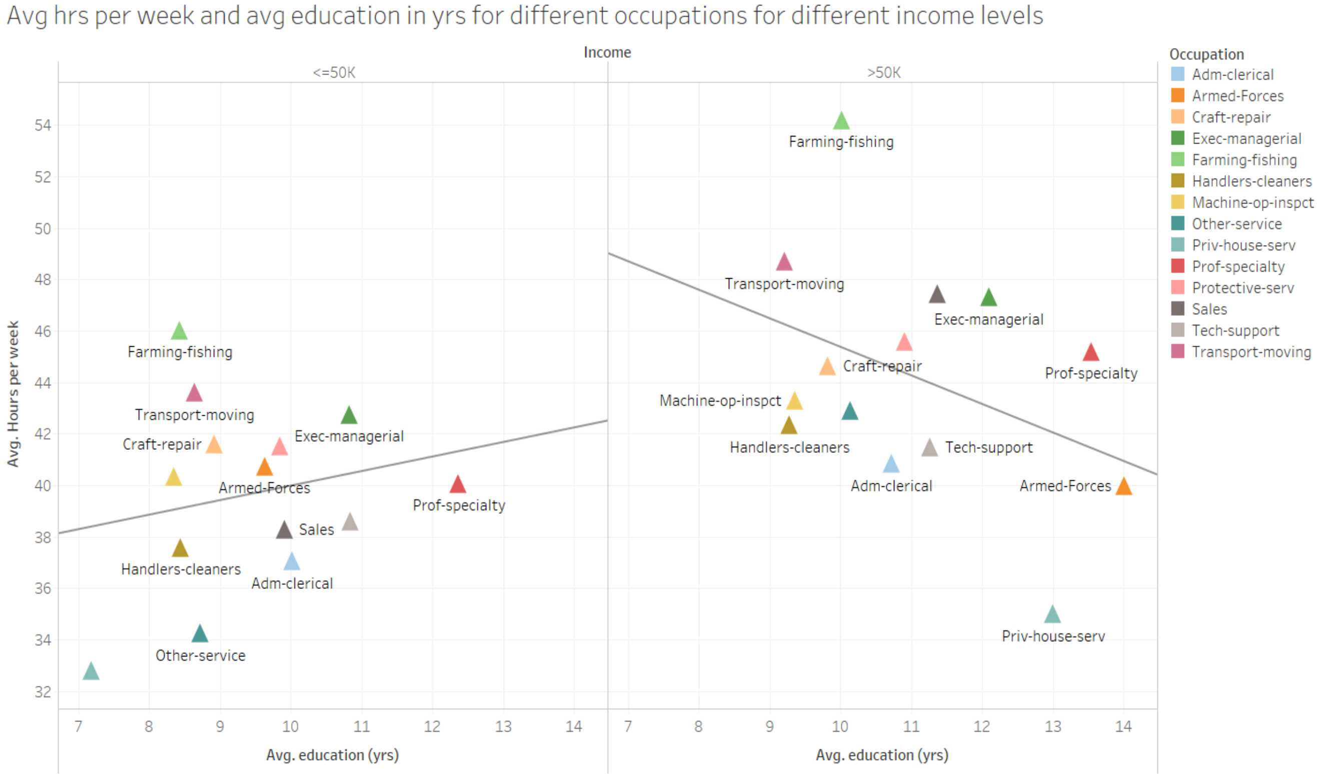


Figure 12:Average hours per week and average education in years for different occupations for different income levels

Figure 12 indicates that for people earning <=50k, as their average years of education increases, for occupations like prof-speciality which requires educational qualifications thus more years of education, there is no significant increase in the average hours per week while for jobs which require less to no educational qualification, there is a significant increase in the average hours per week. This can also be interpreted that less educational qualification may bring more working hours. But for people earning more than 50k, the average hour per week difference between occupation that requires more education in years and those which doesn't increases. That is, with more earnings comes more working for jobs that ask for less educational qualifications and not so much for occupations that require more educational qualifications.

VI. Conclusion

The conclusion to the first question, "is there a difference in hours of work between men and women?", was answered in the exploration phase as well as in comparison with other work. There is a significant disparity between working hours between men and women with more hours for men.

The statement that there are less working hours in western civilization was rejected in the exploration phase. However there was research that indicated that the used dataset wasn't representative in the present time. This question should be investigated further for correct conclusions.

The influence of education on working hours gave ambiguous results that needed further investigations into different attributes. The following results came to the surface. The first result showed that women work less with increasing age while men work longer with increasing age. A second influential variable is the occupation type.

While the results are different in relation to gender, there are significant differences in working hours depending on the occupation type.

The final influence detected was the correlation between working hours and education years in regard with income. There is a (partial) linear increasing correlation for low incomes while there is a linear decreasing correlation for high incomes.

Appendix A: contributions

While we both worked together on the exploration of the design space and the report, Adhithya was responsible for almost every aspect of the graph implementations. She had more insight and skills in creating the needed statistical visualizations. Tobias was cooperative and provided ideas and insights for the hypothesis which helped the visualisations and the conclusions. Arailyn was not a part of this project.

Appendix B: link to video

No video has been made because of the lack of interactive visualizations

Appendix C: references

[1]https://archive.ics.uci.edu/ml/datasets/Census+Income?fbclid=IwAR31LITznOdQ00eMjpzaYLI3pZawqe09ImwUAjtriLTqxWdMz21ok81olwQ
[2]https://ourworldindata.org/working-hours
[3]https://ec.europa.eu/eurostat/statistics-explained