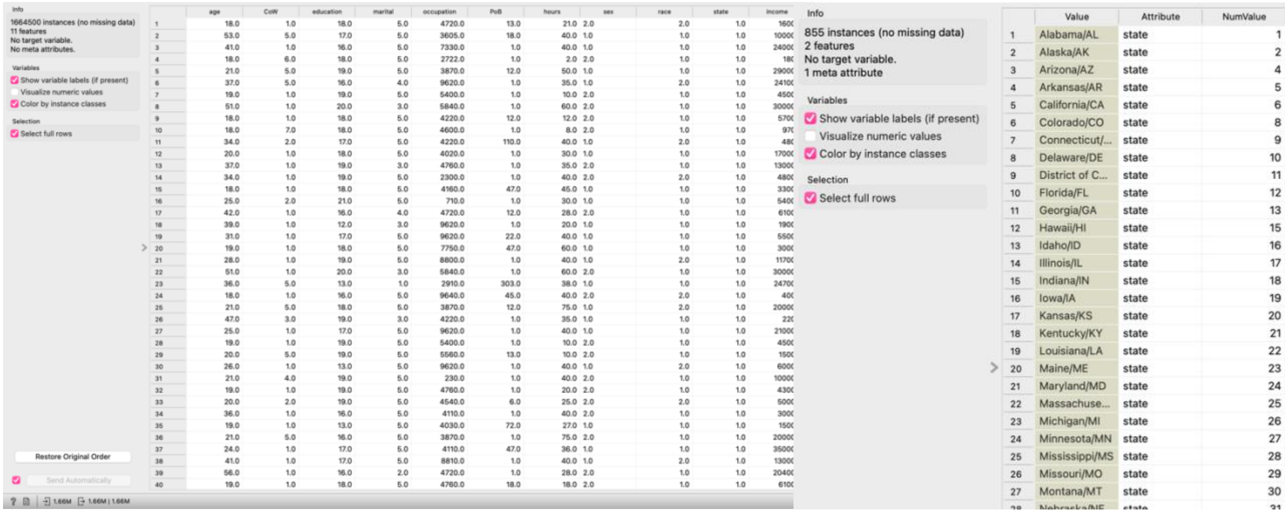
Analysis of Income.

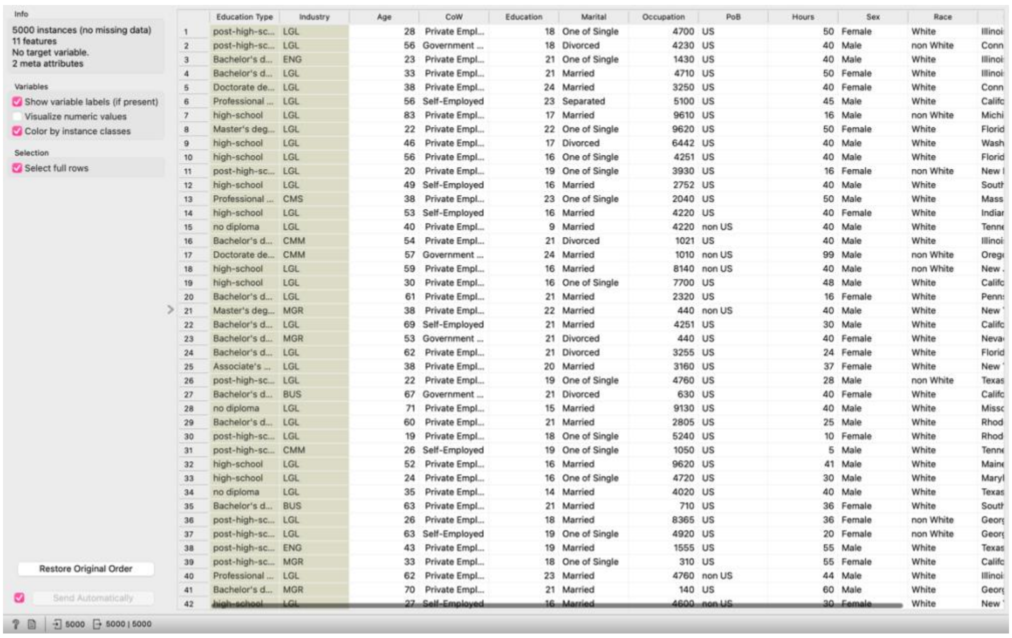
This report aims to analyse census data from the United States. This research, using a Public Use Microdata Sample (PUMS), will evaluate various feature statistics to ascertain whether there are any significant data trends pertaining to income. Orange 3, a data mining software, will be used throughout to preprocess the data, analyse the fairness in income distribution, predict income and gain a wider understanding of demographics of US elections.

Part 1. Preprocessing.

The file **Census_data.csv** is loaded in orange, the original data has 1664500 instances. This is reduced to 5000 instances with the data sampler widget to make processing easier.



The file **Attribute_values.csv** is also loaded in orange, where select rows and select columns are used to isolate then merge the features with Census_data.csv.



Following this, a visual analysis was carried out to determine any outliers and gain a deeper understanding of the data.

Feature Statistics

Name	Distribution	Mean	Mode	Median	Dispersion	Min.	Max.	Missing
Age		43.2758	54	44	0.353751	17	94	0 (0 %)
Education		18.5936	21	19	0.175012	1	24	0 (0 %)
Occupation		4176.35	2310	4200	0.637687	10	9830	0 (0 %)
Hours		38.4268	40	40	0.346461	1	99	0 (0 %)
Income		56108.1	50000	39000	1.26853	150	830000	0 (0 %)

Above shows the summary of the numerical statistics from the data set. The category Age shows the mean as 43.2758, mode as 54, median as 44. This shows that the average age of the data set is 43, the most frequent age is 54. With the mean of 43.2758 and the median of 44, this indicates that the data is symmetrically distributed, showing no extreme skewness. Education has a mean of 18.5936, a mode of 21 and a median of 19. This indicates that 18.5936 is the average amount of years spent in education, with the most being 21 years. With the median being close value to mean this also suggest an evenly distributed data, which will be less affected by the outliers. The dispersion being 0.175012 demonstrates this. The closer to 0 the dispersion values are the more likely the data points have lower variability. Occupation has a dispersion value of 0.637687, showing that the values are more spread out, having moderate variability, which is shown in the distribution column. Hours has a dispersion value of 0.346461, showing that the data is grouped around the mean also indicating low variability, like Age and Education. Income has a dispersion value of 1.26853. As this is greater than one, it shows high variability which means the income ranges widely.

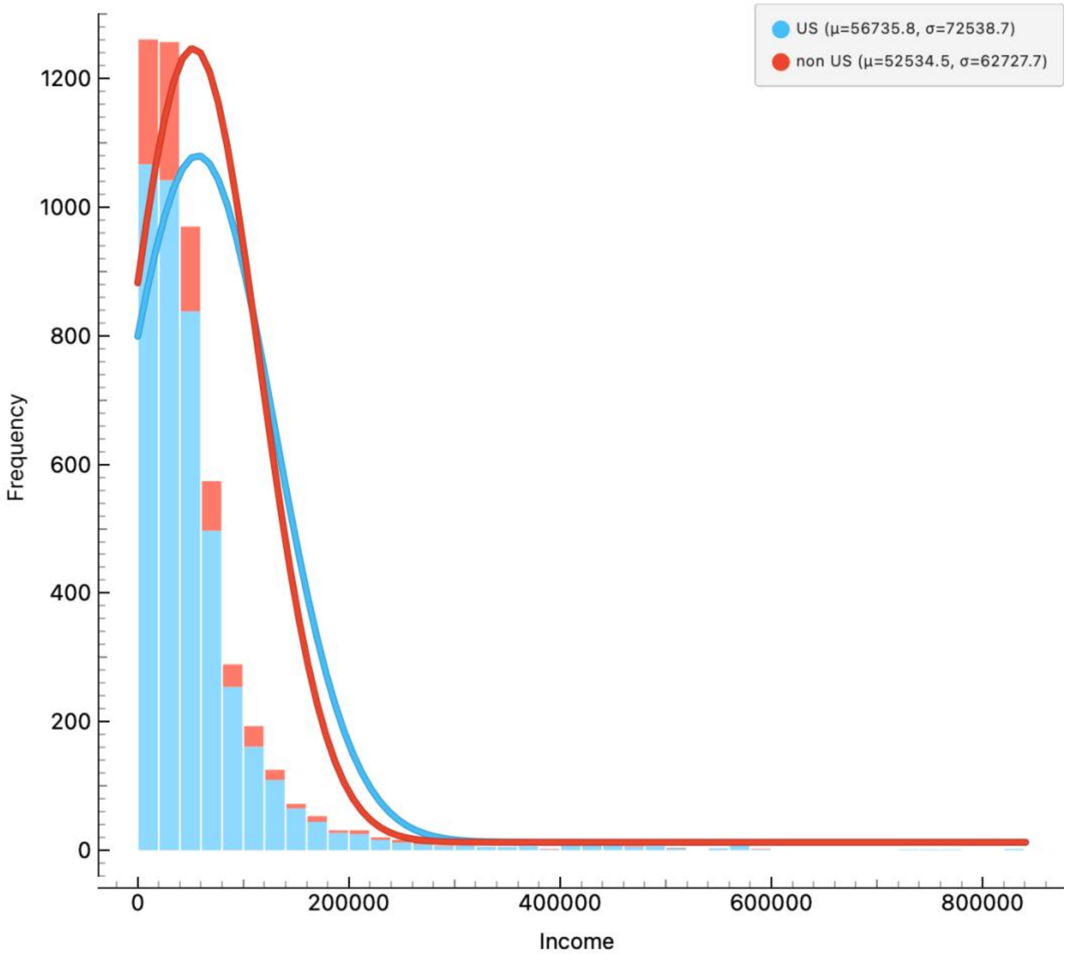
Feature Statistics

Name	Distribution	Mean	Mode	Median	Dispersion	Min.	Max.	Missing
CoW			Private Employee		0.778			0 (0 %)
Marital			Married		1.09			0 (0 %)
PoB			US		0.422			0 (0 %)
Sex			Male		0.692			0 (0 %)
Race			White		0.523			0 (0 %)
State			California		3.5			0 (0 %)
Education Type			high-school		1.82			0 (0 %)
Industry			LGL		0.936			0 (0 %)

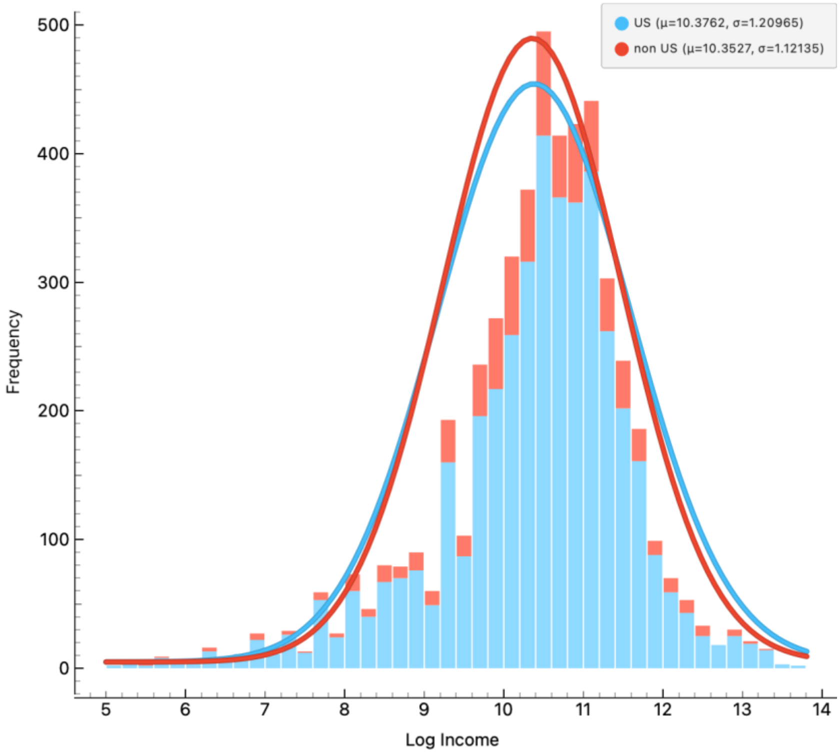
From the categorical data we can gather, that there is high variability between the data. With PoB having the lowest dispersion value of 0.422 and State having the highest value of 3.5

Part 2. Fairness in income distribution.

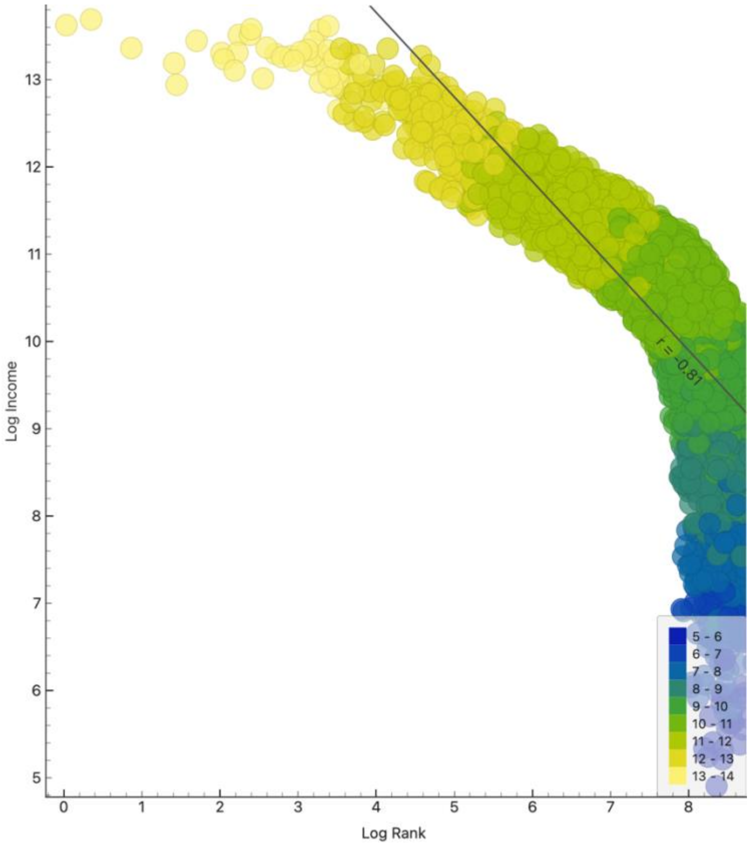
The Histogram below shows the distribution of income in the US and Non-US. Focusing on income distribution in the US, the histogram is skewed significantly to the right. This highlights that only a small proportion of people in the US are high income earners, the majority having lower incomes. The blue line represents the fitted normal distribution with a mean of $56,735.80 and a SD of $72,538.70. With the SD being higher than the mean, it shows high variability confirming that income distribution in US is right-skewed.



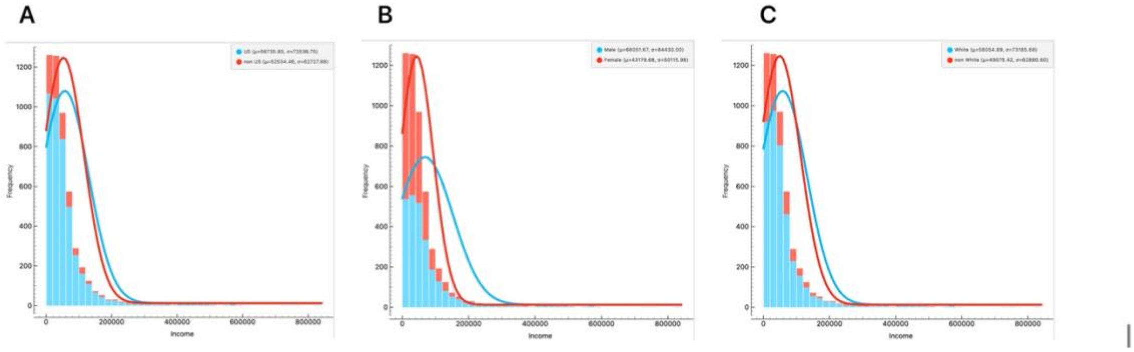
The Histogram below shows the log of income. This transformation follows a log normal distribution, indicating that large absolute changes are less common than small percentage changes. The US log mean is 10.3762 and the SD is 1.20965. Using $e^{10.3762}$ we get ≈ $32,000. Compared to the original income data, log-income is more representative of income distribution in the US as it highlights how income is varied.



The Zipf plot examines how income follows Zipf's Law. The x axis represents the log rank while the y axis represents the log income. The colours further highlight the log income, for easier readability. The downward curve suggests the inequality is prevalent between income. While highest earners follow Zipf's Law, the lower earners have a distribution that is simpler. With a slope of -0.81 it is more likely that the distribution follows the power-law rather than Zipf's law.

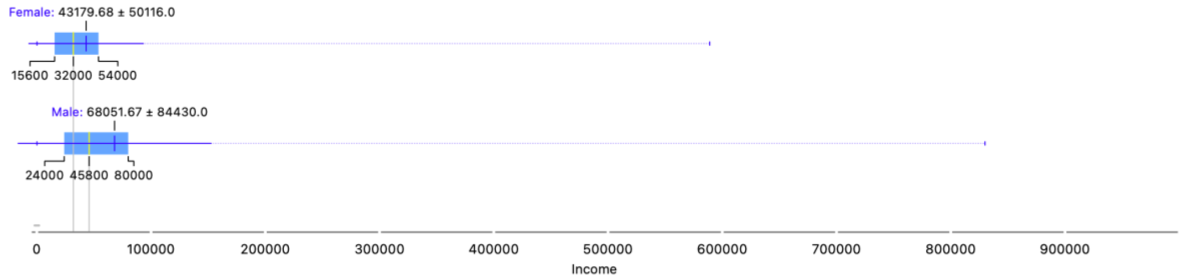


<u>Does Sex, PoB (place of birth) and Race affect an individual's income?</u>

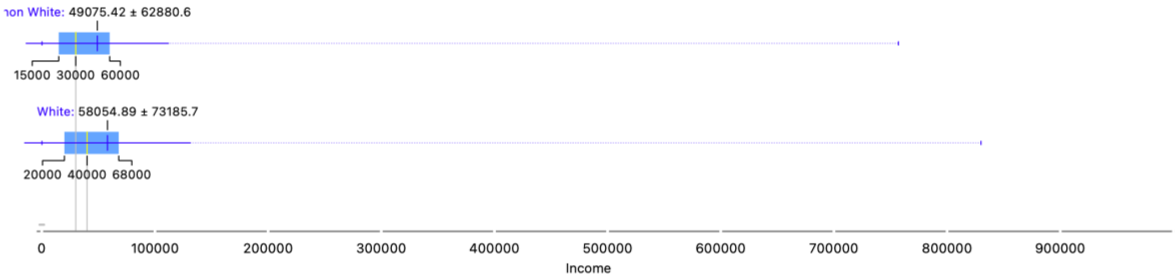


A: Right-skewed distribution (PoB: Income) **B**: Right-skewed distribution (Sex: Income) **C**: Right-skewed distribution (Race: Income).

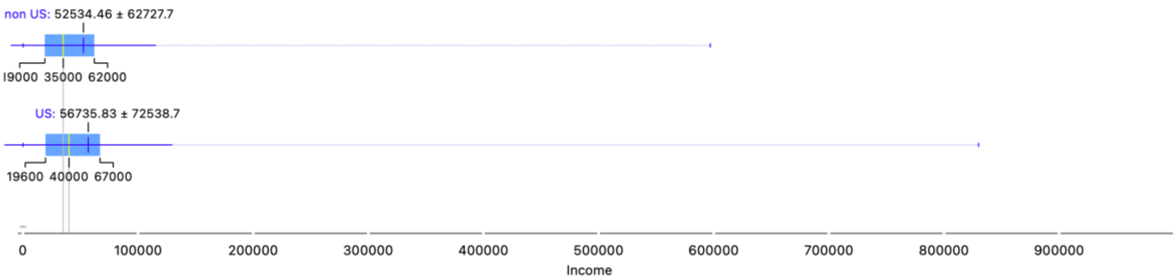
The Null hypothesis (Sex): There is no disparity of the mean income between male and female.
Female: 43179.68 ± 50116.0
Male: 68051.67± 84430.0
Student's T-test: t = 12.778 (p=0.000, N=5000)
Interpretation: We can reject the null hypothesis as there is disparity between male and female. With a p value < 0. 001 the disparity is statistically significant. This is further confirmed when looking at the mean income between male and female, $43,179.68 and $68,051.67 respectively. This shows that men on average are higher earners (≈ 57.6% more). Both SD suggest a high variability of the mean incomes between male and females, with males' income being more varied, suggesting some a larger disparity between higher and lower earners.



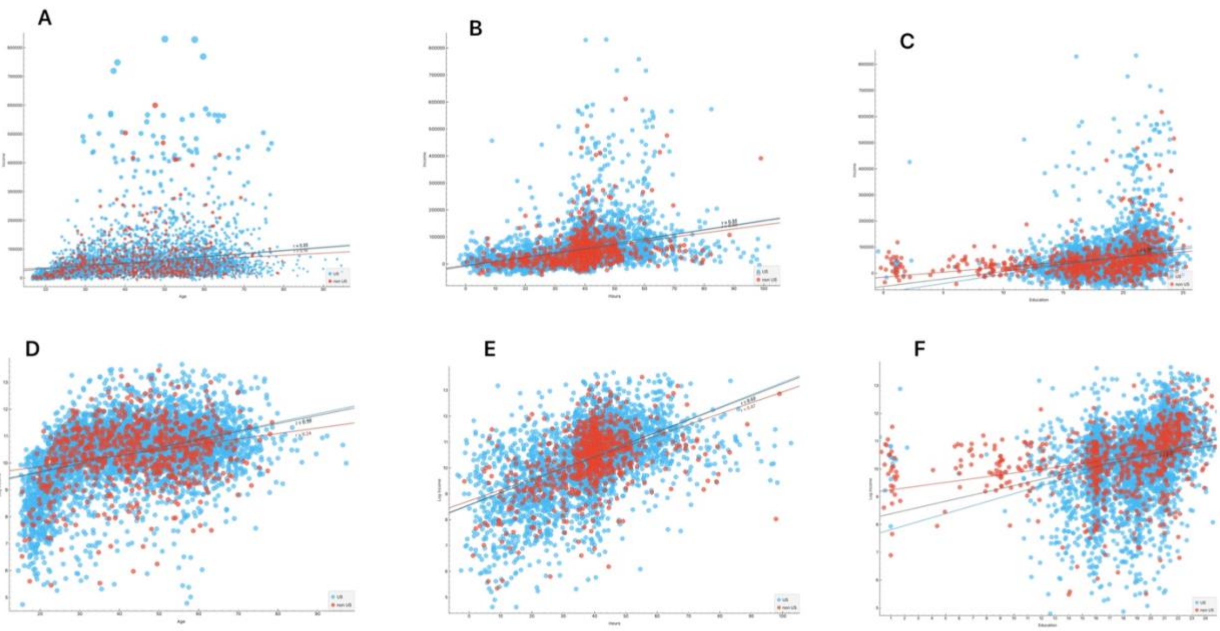
The Null hypothesis (Race): There is no significant difference in mean income between different races.
Non-white: 49075.42 ± 62880.6
White: 58054.89 ± 73185.7
Student's T-test: t = 4.010(p=0.000, N=5000)
Interpretation: Null hypothesis is rejected. Again, a p value < 0. 001 indicates that the difference is statistically significant. Looking at mean income white individuals are higher earners than non-white earning $58,054.89. With a difference of $8,979.47, white individuals earn ≈ 18.3% more on average. High SD from both groups confirms income is right-skewed.



The Null hypothesis (PoB): The mean income of US residents and non-US residents are the same.
Non-US: 52534.46 ± 62727.7
US: 56735.83 ± 72538.7
Student's T-test: t = 1.647 (p=0.100, N=5000)
Interpretation: Null hypothesis is not rejected as there is not a significant difference between US residents and non-US, with the p value > 0.05 and a t value of 1.647. US residents only earn ≈ 8% more on average. Rather than underlying income disparity, the mean difference could be the result of random variance.



<u>**Correlations between income and (1) age, (2) hours worked and (3) education.**</u>



A: Correlation between Age and Income. There is a weak positive correlation (US: r = 0.23, Non-US: r= 0.16) between Age and Income. Income rises with age but there is a wide range of variation. Residents of US (blue) are higher earners than non-US (red).
B: Correlation between Hours worked and Income. There is a moderate positive correlation (US: r= 0.32, Non-US: r= 0.30). Higher income is linked to more hours worked.
C: Correlation between Education and Income. There is a weak positive correlation (US: r= 0.26, Non-US: r= 0.26). The financial benefits of higher education seem to be greater for Us residents.
D: Scatter plot of Age and Log Income showing a weak positive correlation with US residents having a greater correlation.
E: Scatter plot of Hours worked and Log Income showing a stronger positive correlation with US residents having a higher correlation.
F: Correlation of Education and Log Income showing a moderate positive corelation.

	Correlation	Independent Variable	Dependent Variable	Uncorrected p	FDR
1	0.313	Hours	Income	5.66434e-114	9.91259e-114
2	0.262	Education	Income	2.10235e-79	2.94329e-79
3	0.225	Age	Income	2.12379e-58	2.12379e-58

Age and Income: with r = 0.225, there is a weak positive correlation.
Hours and Income: with r = 0.313, there is a moderate positive correlation.
Education and Income: with r = 0.262, there is a weak positive correlation

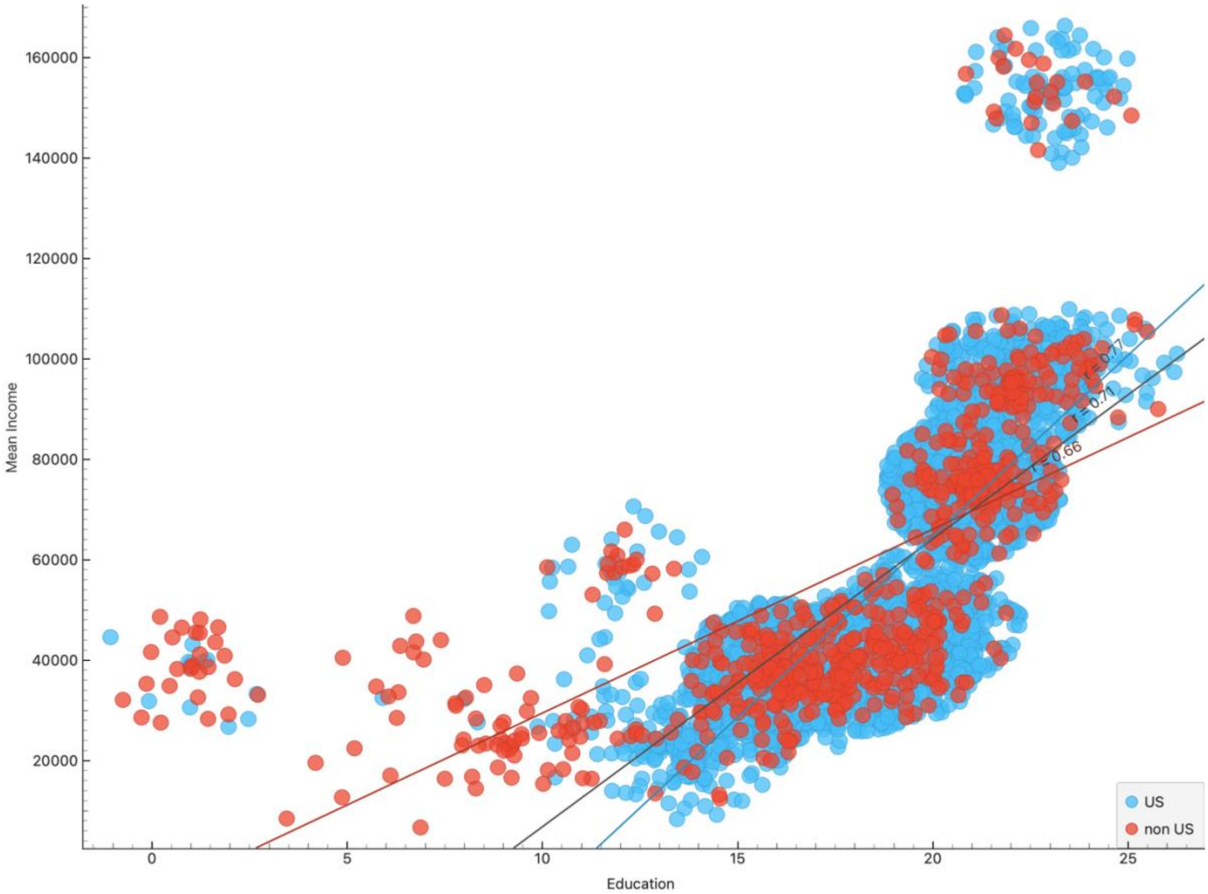
	Correlation	Independent Variable	Dependent Variable	FDR	Uncorrected p
1	0.694	Income	Log Income	0	0
2	0.52	Hours	Log Income	0	0
3	0.383	Age	Log Income	8.46765e-174	6.04832e-174
4	0.307	Education	Log Income	1.54756e-109	1.32648e-109

Age and Log Income: with r = 0.383, there is a moderate positive correlation.
Hours and Log Income: with r = 0.52, there is a moderately strong correlation.
Education and Log Income: with r = 0.307, there is a moderate positive correlation.

The log income results are more representable of real income values. The log income also reduces the higher values and disperses the lower values. Overall, the log income transformed the independent variables towards a more linear relationship. Hours worked are a strong indicator of earnings. With age and education being a moderate indicator.

Part 3. Predicting income.

The plot below shows the relationship between education and mean income, the data is separated as US residents and non-US.



There is a positive correlation between the two groups, indicating the length of education will have an influence on income. The affects are more prominent in the US, indicated by the slope. Those who spent longer time in education are seen to have a significant increase in earnings. This plot shows outliers, where some who have little/more education are earning significantly higher/lower than predicted. There is a stronger correlation between education and income amongst US residents.

The graph below shows the relationship between education and log mean income, grouped by US and non-US residents. There are fewer outliers between education and log mean income than mean income. Following a similar pattern, this plot shows that there is a positive correlation, further highlighting how education can impact income overall. As mentioned before, the log income seems to be better fitted.

This table shows what the mean income is as well as log income for each education level. No diploma has the lowest earning potential with the mean income at $23,266.90. Those who go on to do a professional degree have the potential to become top earners with a mean income of $153,082.00. In contrast, the log mean income shows that having a doctorates degree will make you have the potential to be the highest earners, with the mean income at $129,365.89. The lowest earners are those with no diploma with mean income at $20,556.24.

Education Type	Mean Income ($)	Raw Data Log Mean Income	Log Mean Income ($)
Associate's degree	48,425.10	10.4659	35,098.02
Bachelor's degree	74,457.80	10.7839	48,237.89
Doctorate degree	98,824.60	11.7704	129,365.89
Master's degree	95,212.30	11.1156	67,211.52
Professional degree	153,082.00	11.4188	91,016.86
High school	36,128.00	9.95188	20,991.65
No diploma	23,266.90	9.93092	20,556.24
Post-high-school	39,548.20	10.0137	22,330.30

A

	Mean Income - First value	Education
1	38728.9	1
2	27900	2
3	23000	4
4	10833.3	5
5	31000	6
6	40357.1	7
7	32166.7	8
8	23266.9	9
9	29000	10
10	27481.9	11
11	57166.7	12
12	22619.5	13
13	22019.7	14
14	32590.6	15
15	37895.3	16
16	36128	17
17	39548.2	18
18	41367.2	19
19	48425.1	20
20	74457.8	21
21	95212.3	22
22	153082	23
23	98824.6	24

B

Mean Log Income - First value	Education
9.98043	1
9.9202	2
10.0432	4
8.84828	5
10.2159	6
10.5708	7
10.1729	8
9.93092	9
9.86488	10
9.78623	11
10.2593	12
9.28429	13
9.13309	14
9.84664	15
10.0802	16
9.95188	17
10.0137	18
10.0623	19
10.4659	20
10.7839	21
11.1156	22
11.4188	23
11.1704	24

A: Mean Income for each year of Education.
B: Log Mean Income for each year of Education.

To estimate the monetary value for education by income and years per education multiple linear regressions were carried out. Below are tables showing, the coefficients of log income for education type, followed by the coefficient for each year overall of log income.

	name	coef
1	intercept	10.6064
2	Education Type=Associate's degree	-0.140474
3	Education Type=Bachelor's degree	0.177519
4	Education Type=Doctorate degree	0.564047
5	Education Type=Master's degree	0.509207
6	Education Type=Professional degree	0.812402
7	Education Type=high-school	-0.654502
8	Education Type=no diploma	-0.675468
9	Education Type=post-high-school	-0.59273

	name	coef
1	intercept	9.50086
2	Education	0.0484855

Below are tables showing the monetary value for normal income.

	name	coef
1	intercept	71118.1
2	Education Type=Associate's degree	-22692.9
3	Education Type=Bachelor's degree	3339.69
4	Education Type=Doctorate degree	27706.5
5	Education Type=Master's degree	24094.2
6	Education Type=Professional degree	81963.8
7	Education Type=high-school	-34990.1
8	Education Type=no diploma	-47851.2
9	Education Type=post-high-school	-31569.9

This table shows that for each year of education you have the potential of earning an additional $3,138.77 overall.

	name	coef
1	intercept	4819.89
2	Education	3138.77

Although estimating a monetary value for income is beneficial there are still things to consider. By assuming that there is a linear relationship between income and education, there is a risk of overestimating or underestimating the effect of education. Other features like marital, sex, race etc. could be influencing income however without being able to control these features, we are unsure if education alone is affecting income.

Classification models were applied to predict incomes. The table below shows income grouped by "high-income" or "low-income".

Info

5000 instances (no missing data)
12 features
Target with 2 values
No meta attributes.

Variables

- [x] Show variable labels (if present)
- [ ] Visualize numeric values
- [x] Color by instance classes

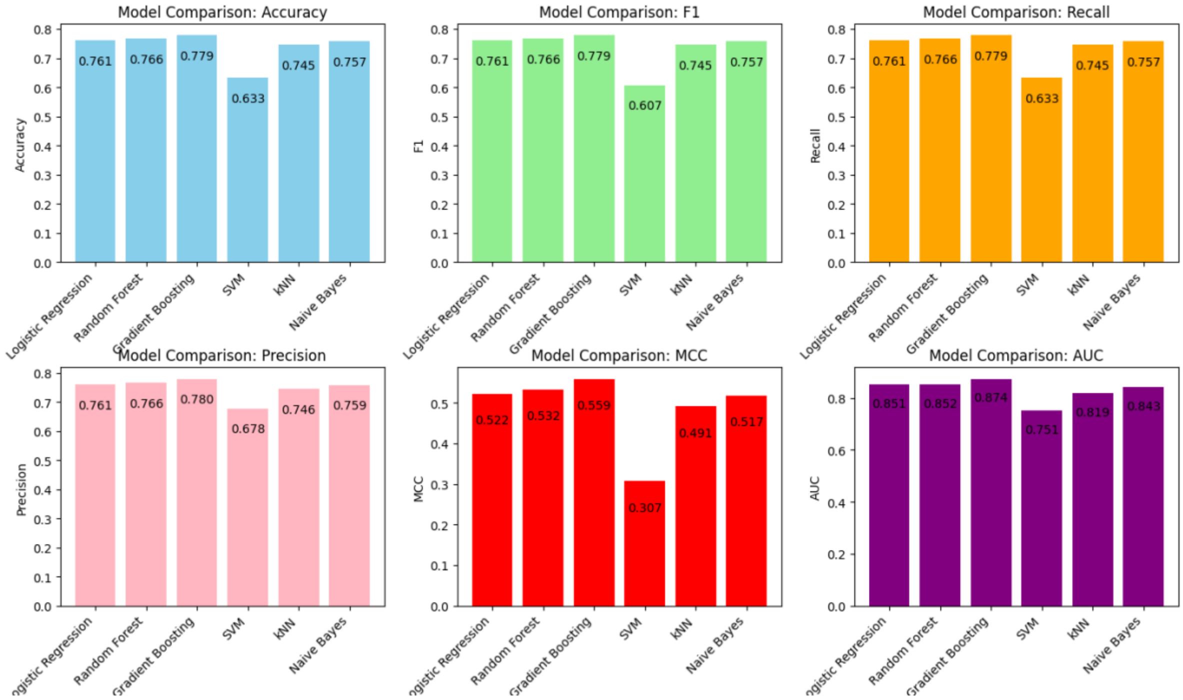
Selection

- [ ] Select full rows

	Income Classification	Education Type	Race	CoW	State	Industry	Occupation	Marital	Sex	Education
1	High Income	post-high-sc...	White	Private Empl...	Illinois	LGL	4700	One of Single	Female	18
2	High Income	post-high-sc...	non White	Government ...	Connecticut	LGL	4230	Divorced	Male	18
3	High Income	Bachelor's d...	White	Private Empl...	Illinois	ENG	1430	One of Single	Male	21
4	High Income	Bachelor's d...	White	Private Empl...	Illinois	LGL	4710	Married	Female	21
5	High Income	Doctorate de...	White	Private Empl...	Connecticut	LGL	3250	Married	Female	24
6	Low Income	Professional ...	White	Self-Employed	California	LGL	5100	Separated	Male	23
7	High Income	high-school	non White	Private Empl...	Michigan	LGL	9610	Married	Male	17
8	High Income	Master's deg...	White	Private Empl...	Florida	LGL	9620	One of Single	Female	22
9	High Income	high-school	White	Private Empl...	Washington	LGL	6442	Divorced	Male	17
10	Low Income	high-school	White	Private Empl...	Florida	LGL	4251	One of Single	Male	16
11	Low Income	post-high-sc...	non White	Private Empl...	New Hamps...	LGL	3930	One of Single	Female	19
12	Low Income	high-school	White	Self-Employed	South Carolina	LGL	2752	Married	Male	16
13	High Income	Professional ...	White	Private Empl...	Massachuse...	CMS	2040	One of Single	Male	23

Six model classifiers were used to predict whether an individual was a high earner or a low earner. Those were: Logistic Regression, Random Forest, Gradient Boosting, SVM, kNN and Naïve Bayes.

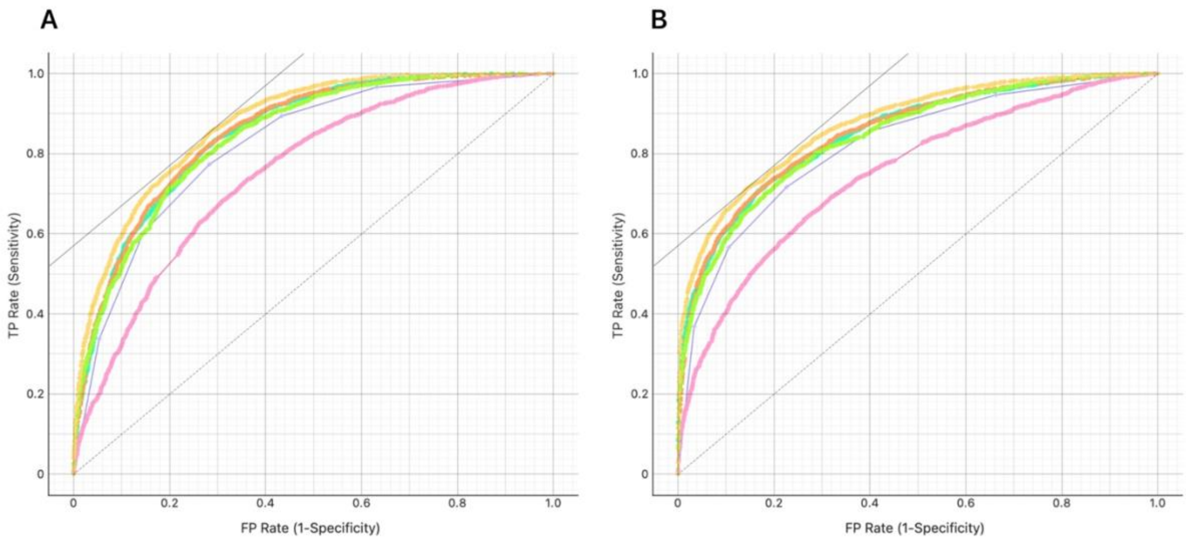
The barplot below shows a model comparison. When choosing the best fit model these factors were considered: Accuracy, F1, Recall, Precision, MCC and AUC. Overall, the model Gradient boosting was the best fit as it had higher values across all these factors. The values can be seen in the plot below.



Below shows the results of an ROC analysis to determine the performance of each classification model.

Key for ROC: **Yellow** – Gradient Boosting, **Green** – Naïve Bayes, **Pink** – SVM, **Purple** – kNN, **Orange** – Random Forest, **Blue** – Logistic regression.

As Gradient Boosting had the most area under curve, this was seen as the best fit model.



A: Represents “High-income”. **B**: Represents: “Low-income”.

A Confusion matrix was also carried out.
A: Number of instances.
B: Proportion of actual.
Interpretation: For number of instances, the model has correctly classified 2001 individuals of high income and 1894 of low income. For the proportion of actual, the model has correctly classified 80.1% individuals of high income and 75.7% individuals of low income.

A

		Predicted		
		High Income	Low Income	Σ
Actual	High Income	2001	497	2498
	Low Income	608	1894	2502
	Σ	2609	2391	5000

B

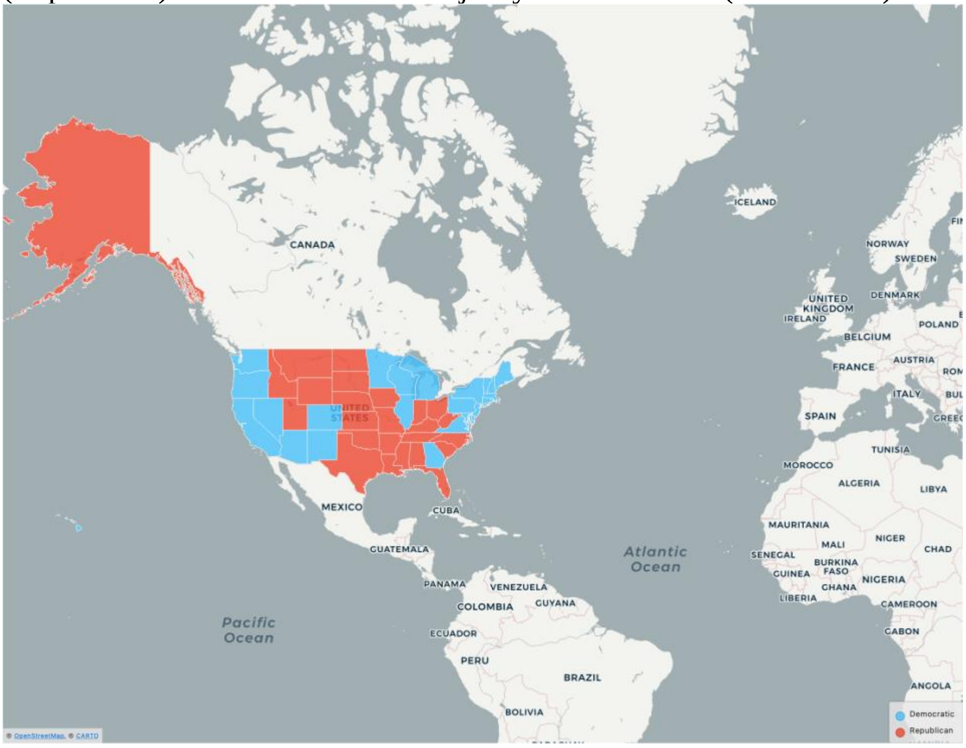
		Predicted		
		High Income	Low Income	Σ
Actual	High Income	80.1 %	19.9 %	2498
	Low Income	24.3 %	75.7 %	2502
	Σ	2609	2391	5000

As education is not the only feature to have an impact on income, the features were ranked on the importance to income. With hours worked being the most feature to impact income and education being last.

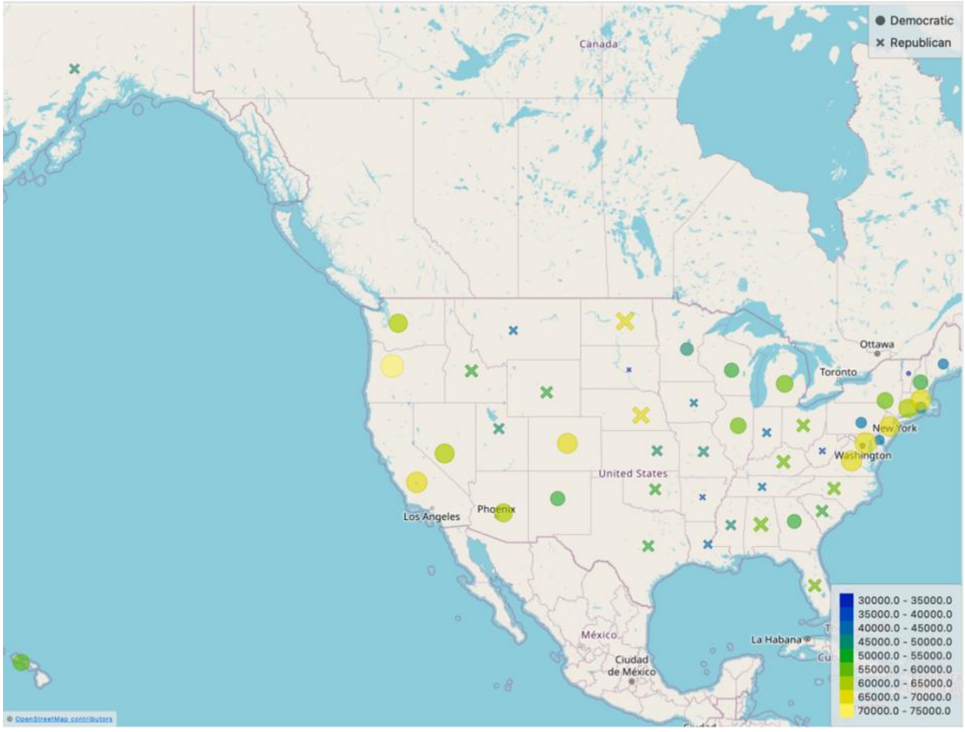
		#	Gra...ing
1	N Hours		0.361
2	N Occupation		0.203
3	N Age		0.169
4	N Education		0.142

Part 4. Demographics of US elections.

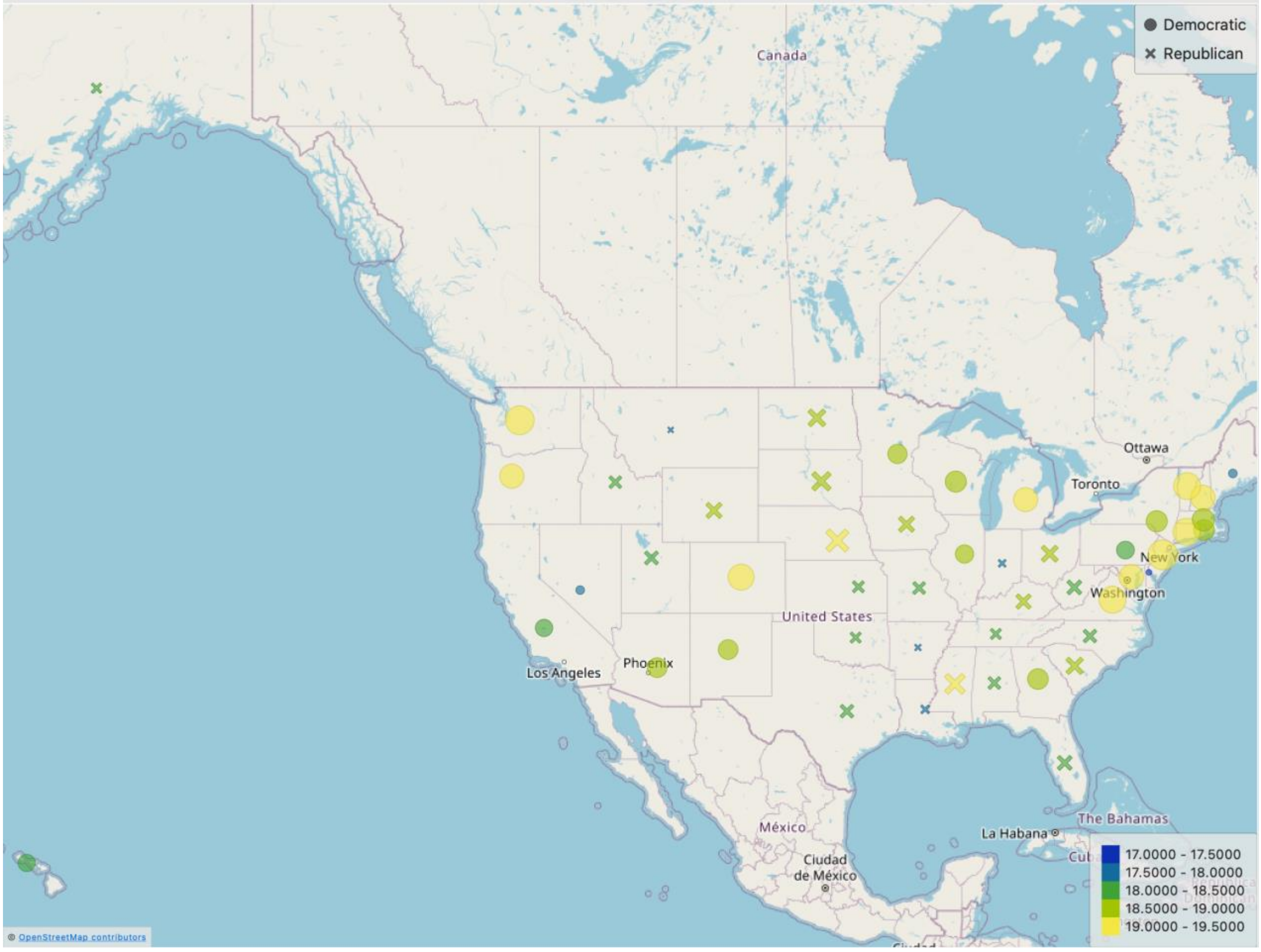
A map to show the 2020 US election result by state. Red states are majority win for Trump (Republican) and blue states are majority win for Biden (Democratic).



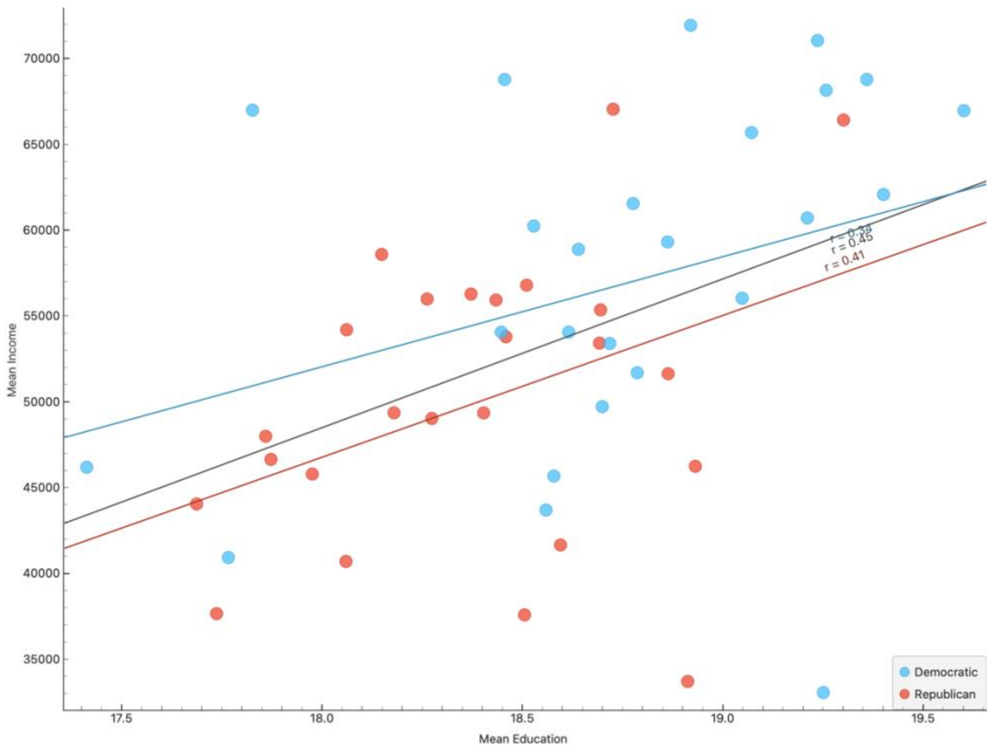
A map to show the mean income by state. The shapes represent the political party that won (Democratic: Circle, Republican: Cross). The size of shape and colours both represent the income. Lower income is associated with states that voted for Trump while higher earning states was associated with Biden.

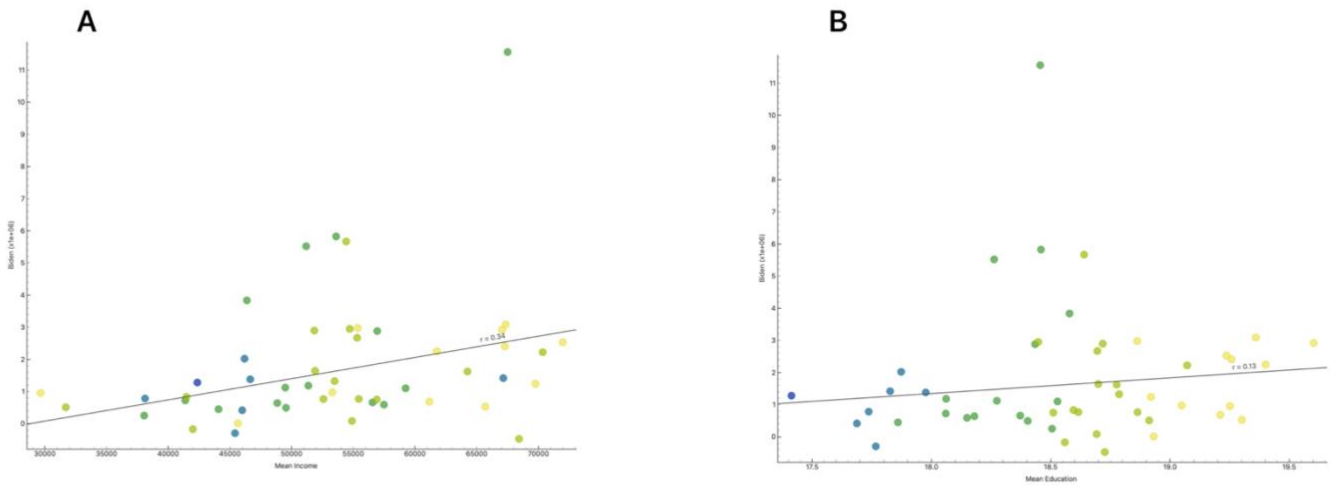


A map to show the mean education level by state. The yellow circles indicate that higher education was associated with states that vote for Biden (Democratic) and lower was associated with voting for Trump (Republican).

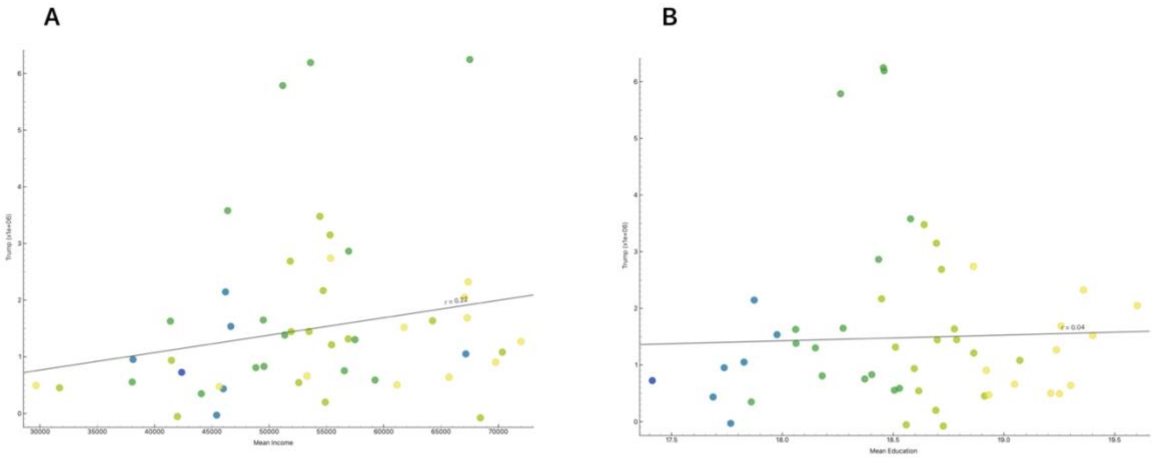


The scatter graph below the relationship between education and income. It is grouped by voting party. There is a positive correlation, showing the effect of education on income. The trend line for the democratic party indicates higher income and education than of the republic party.





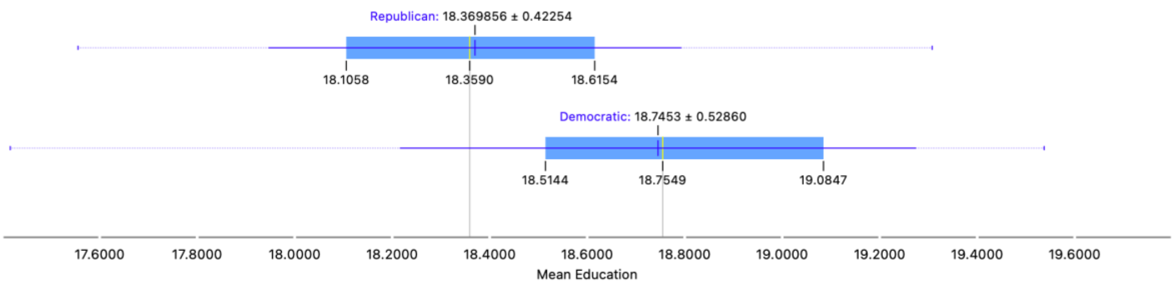
A: Scatter graph showing the relationship between Biden voters and Income. Positive correlation (0.34). **B:** Scatter graph showing the relationship between Biden voters and Education. Positive correlation (0.13). A lower correlation in education suggests that income has more of an impact on voting decisions.

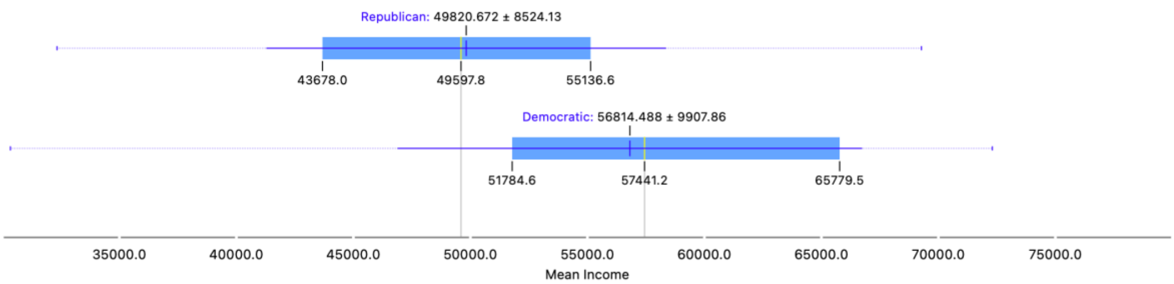


A: Scatter graph showing the relationship between Trump voters and Income. Positive correlation (0.22). **B:** Scatter graph showing the relationship between Biden voters and Education. Weak Positive correlation (0.03).

The box plots below show the mean education and mean income between the two-party voters. In both categories Democratic voters have a higher median than republican voters.

Education T-test: 2.774 (p=0.008, N =50)
Interpretation: p < 0.05 means education has a strong correlation between Biden voters.
Income t-test: 2.676 (p=0.10, N=10)
Interpretation: p > 0.05 means there is insufficient evidence to infer that income significantly correlates.





From the statistics we can see that low income states voted for Trump and states with higher education voted for Biden in the 2020 election

Part 5. Own data mining.

How does CoW affect income distribution across regions/states?

Null Hypothesis 1: The CoW has no significant impact on income between regions/states.

Null Hypothesis 2: The highest earning individual comes from a region that has the highest overall income between regions.

States were grouped and assigned a new feature. The new feature was "Region", where sates were grouped as either "North", "South", "West", "Midwest", "Northeast" or "Other". This made identifying income distribution across the US easier. Below is the raw data table.

	Income	CoW	State	Education Type	Industry	Marital	Race	Sex	Region
1	40000	Private Empl...	Illinois	post-high-sc...	LGL	One of Single	White	Female	Midwest
2	45200	Government ...	Connecticut	post-high-sc...	LGL	Divorced	non White	Male	Northeast
3	58010	Private Empl...	Illinois	Bachelor's d...	ENG	One of Single	White	Male	Midwest
4	68000	Private Empl...	Illinois	Bachelor's d...	LGL	Married	White	Female	Midwest
5	75000	Private Empl...	Connecticut	Doctorate de...	LGL	Married	White	Female	Northeast
6	21500	Self-Employed	California	Professional ...	LGL	Separated	White	Male	West
7	45400	Private Empl...	Michigan	high-school	LGL	Married	non White	Male	Midwest
8	50000	Private Empl...	Florida	Master's deg...	LGL	One of Single	White	Female	South
9	40000	Private Empl...	Washington	high-school	LGL	Divorced	White	Male	West
10	25000	Private Empl...	Florida	high-school	LGL	One of Single	White	Male	South
11	5400	Private Empl...	New Hamps...	post-high-sc...	LGL	One of Single	non White	Female	Northeast
12	25000	Self-Employed	South Carolina	high-school	LGL	Married	White	Male	South
13	40000	Private Empl...	Massachuse...	Professional ...	CMS	One of Single	White	Male	Northeast
14	18000	Self-Employed	Indiana	high-school	LGL	Married	White	Female	Midwest
15	28000	Private Empl...	Tennesse	no diploma	LGL	Married	White	Male	Other
16	151000	Private Empl...	Illinois	Bachelor's d...	CMM	Divorced	White	Male	Midwest
17	392000	Government ...	Oregon	Doctorate de...	CMM	Married	non White	Male	West
18	25000	Private Empl...	New Jersey	high-school	LGL	Married	non White	Male	Northeast
19	72000	Private Empl...	California	high-school	LGL	One of Single	White	Male	West
20	20000	Private Empl...	Pennsylvania	Bachelor's d...	LGL	Married	White	Female	Northeast
21	100000	Private Empl...	New York	Master's deg...	MGR	Married	White	Male	Northeast

This pivot table shows the mean income of different CoW categories by each region.

		CoW				
Region	Mean	Private Employee	Government Employee	Self-Employed	No Pay	Total
	Midwest	51737.732	46368.451	66805.804	52000.0	**52550.219**
	Northeast	55542.548	59361.507	56023.297	27700.0	**56167.928**
	Other	40524.688	38268.519	23243.846	29800.0	**38361.971**
	South	54211.791	53813.790	68575.038	573.333	**55567.610**
	West	58283.654	57713.074	80109.396	39636.667	**60598.732**
	Total	**54762.843**	**54237.459**	**68894.407**	**34663.889**	**56108.143**

For Midwest, the highest earners are self-employed, and the lowest earner have no paid jobs, although the average income suggests they could be getting income from other means like inheritance or government pay. Northeasters who work in the government are likely to be paid more than any other region ($ 59,36.50). The highest earners overall appear to be self-employed individuals who reside in the west earning an average of $80,109.396.

Income ranked highest to lowest by CoW:
1. Self-employed **($68, 894.41**)
2. Private Employee **($54,762.84**)
3. Government Employee **($54,237.46**)
4. No Pay **($34,663.90**)

Income highest to lowest by region:
1. West **($60,598.73**)
2. Northeast **($56,167.93**)
3. South **($55,567.61**)
4. Midwest **($52,550.22**)
5. Other **($38,361.97**)

Below is another pivot table showing the overall breakdown of income between states and CoW.

	CoW				
Mean	**Private Employee**	**Government Employee**	**Self-Employed**	**No Pay**	**Total**
Alabama	47071.818	50855.556	196960.0	840.0	**59567.797**
Alaska	37212.5	16865.0	244000.0	?	**46858.462**
Arizona	66516.714	49800.769	45400.0	?	**61733.830**
Arkansas	36335.357	37312.5	35236.667	?	**36289.333**
California	62361.634	72018.765	88382.237	50583.333	**67054.695**
Colorado	65488.889	42910.0	101853.0	?	**65845.429**
Connecticut	65917.353	58788.889	52700.0	?	**62213.019**
Delaware	45175.0	36975.0	?	?	**42441.667**
Florida	52676.009	56609.722	67526.098	?	**55136.557**
Georgia	51321.058	38833.913	72141.765	?	**51784.583**
Hawaii	47952.381	87680.0	21000.0	?	**57441.250**
Idaho	41615.455	42715.0	411000.0	38530.0	**54854.643**
Illinois	55252.545	55955.909	62274.286	?	**56035.865**
Indiana	40686.7	37129.167	81450.0	?	**43678.033**
Iowa	41658.182	41529.286	36500.0	?	**41374.918**
Kansas	44965.385	36447.5	101474.0	?	**49088.462**
Kentucky	52861.538	41660.833	113300.0	?	**56865.439**
Louisiana	46584.211	33124.444	31080.0	?	**43786.197**
Maine	39617.778	45000.0	57880.0	?	**43700.8**
Maryland	61455.769	92065.333	58300.0	?	**68970.424**
Massachusetts	67269.681	71368.421	39322.857	?	**66288.417**
Michigan	57057.619	40507.333	76745.789	154000.0	**58648.786**
Minnesota	49302.879	44155.556	62287.5	?	**49996.265**

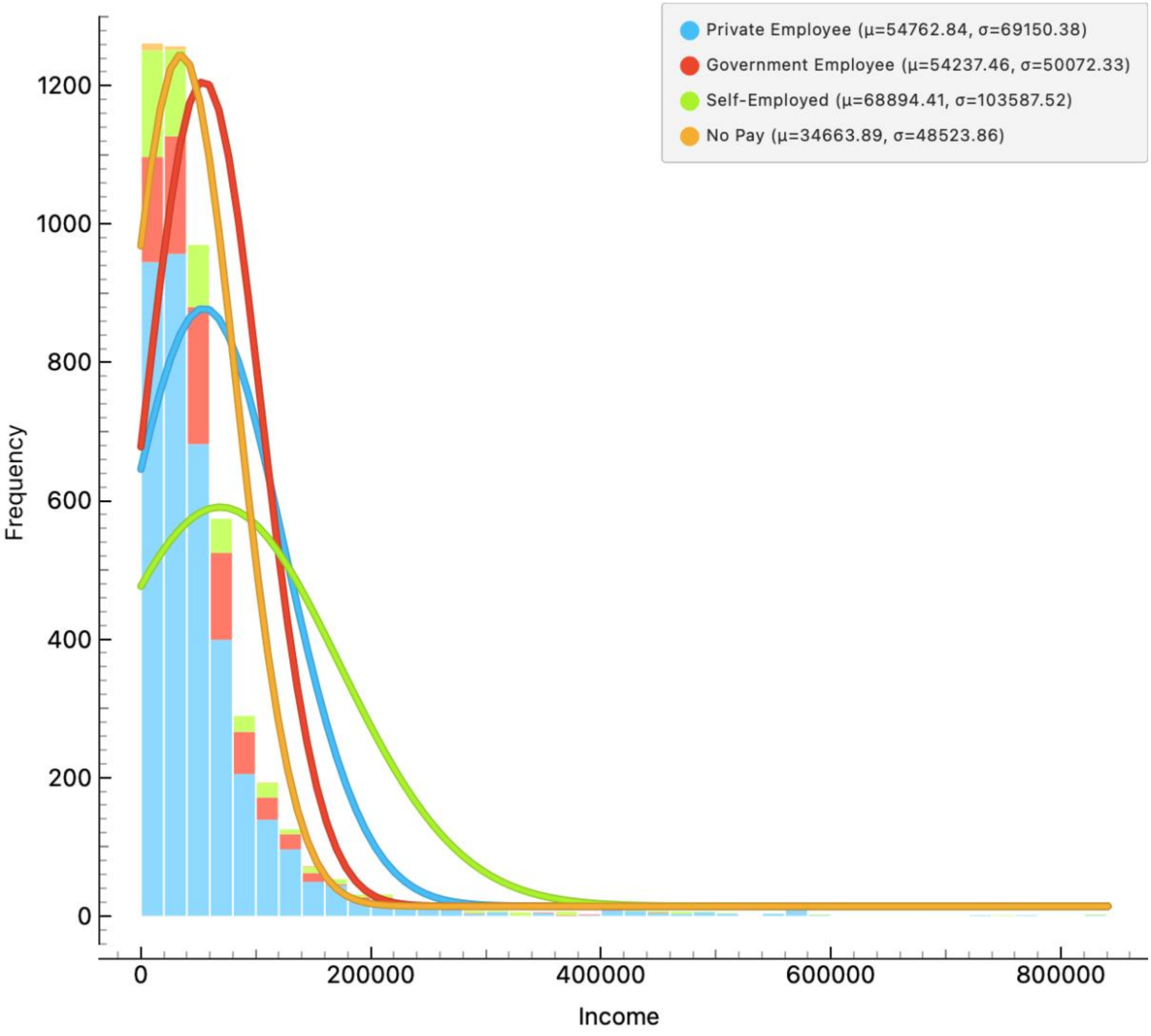
State					
Mississippi	44667.619	27120.0	78366.667	?	**48244.375**
Missouri	49130.227	41645.556	67120.0	?	**49519.655**
Montana	45547.143	27800.0	60000.0	?	**43392.222**
Nebraska	72881.923	56416.667	47665.714	?	**65822.821**
Nevada	66367.105	58800.0	48666.667	?	**64271.277**
New Hampshire	52552.353	65650.0	48067.5	?	**53930.4**
New Jersey	62881.146	88840.909	51460.0	?	**65779.478**
New Mexico	51866.190	62031.818	34500.0	?	**53042.778**
New York	55983.684	52387.910	79828.571	27700.0	**57213.488**
North Carolina	51678.211	52986.0	75544.211	?	**55175.683**
North Dakota	86446.250	47200.0	20200.0	?	**69274.615**
Ohio	55221.301	49078.182	67426.471	13000.0	**55383.226**
Oklahoma	51137.5	32110.0	105100.0	?	**51611.111**
Oregon	75278.222	66840.0	56600.0	?	**72303.871**
Pennsylvania	45760.4	39680.476	35982.381	?	**44225.760**
Rhode Island	43141.176	65650.0	?	?	**45510.526**
South Carolina	59111.176	40529.412	37285.714	?	**52862.267**
South Dakota	35643.333	?	7175.0	33000.0	**32333.333**
Tennesse	45059.306	40545.238	27237.0	29800.0	**42287.404**
Texas	49978.7	50081.091	66487.5	4300.0	**51628.447**
Utah	57372.727	42703.333	17177.5	?	**49597.812**
Vermount	32566.667	?	17000.0	?	**30342.857**
Virginia	71667.670	60139.118	29853.750	700.0	**66205.685**
Washington	58077.701	53007.917	91853.333	?	**62529.318**
West Virginia	44444.0	28000.0	?	180.0	**37514.783**
Wisconsin	51359.625	56091.667	66776.667	8000.0	**52851.569**
Wyoming	66066.667	29600.0	?	10400.0	**53369.231**
Puerto Rico	25038.889	30300.0	6400.0	?	**24819.231**
Total	**54762.843**	**54237.459**	**68894.407**	**34663.889**	**56108.143**

Top 5 Income by state:

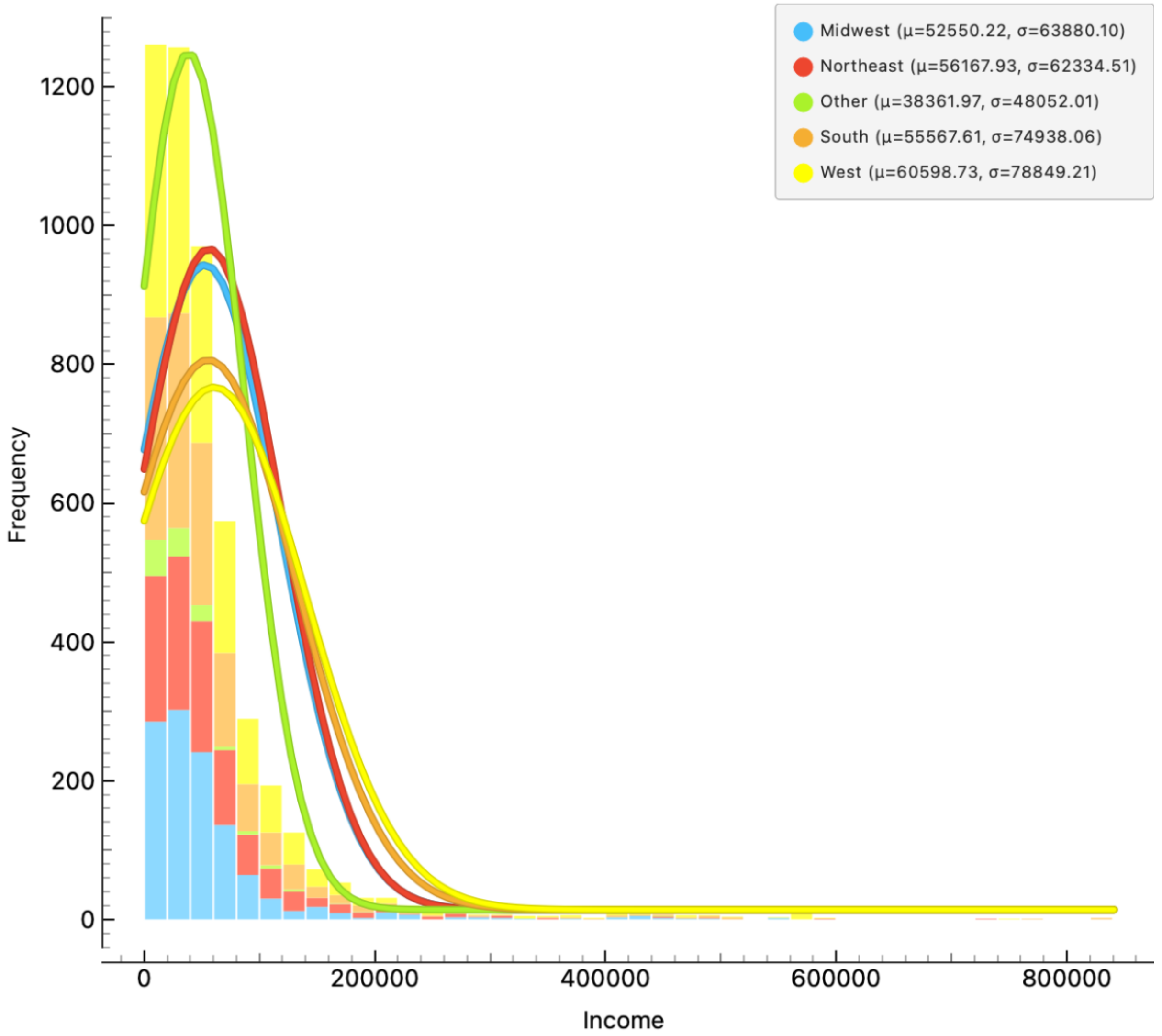
1. Oregon **($72,303.87**)
2. North Dakota **($69,274.62**)
3. Maryland **($68,970.47**)
4. California **($34,663.90**)
5. Massachusetts ($66,288.41)

The lowest earning US state is Vermont with an average income of $30,342.86. Puerto Rico is classed as a US territory not a state otherwise has the lowest income of $24,819.23.

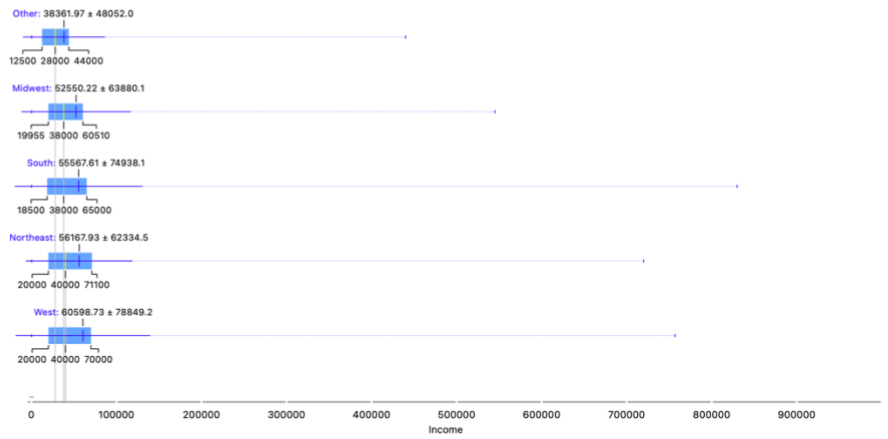
Below is a histogram showing a right-skewed distribution between CoW and income. With self-employed earning the highest $68,894.41 and self-employed having the most significant variation with a SD of $103,587.52, this indicates earners in this category have an extreme income imbalance.

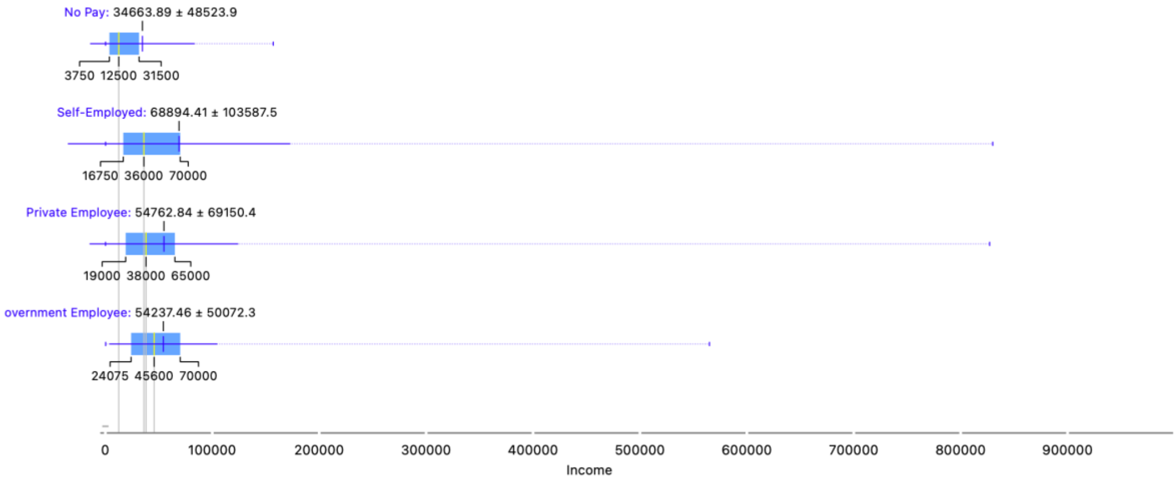


Below is a histogram showing a right-skewed distribution between regions and income. The south region has the most significant variation with the SD being $74,938.06 and average income $55,567.61, suggesting an unequal distribution of income.



A box plot of income distribution between regions and income. Although the south region has the highest earner over $800,000, this is not representative. The histogram showed the SD being $74,938.06, with the average only being $55,567.61, we can confirm the highest earner is not representative of income in the South. Interestingly the Northeast is on the lower scale of earnings ($56,167.93) but has the least variance between the region, with a SD of $62,334.51. Therefore, we can reject the null hypothesis the highest earning individual comes from a region that has the highest overall income between regions.





A box plot of income distribution between Cow and regions.

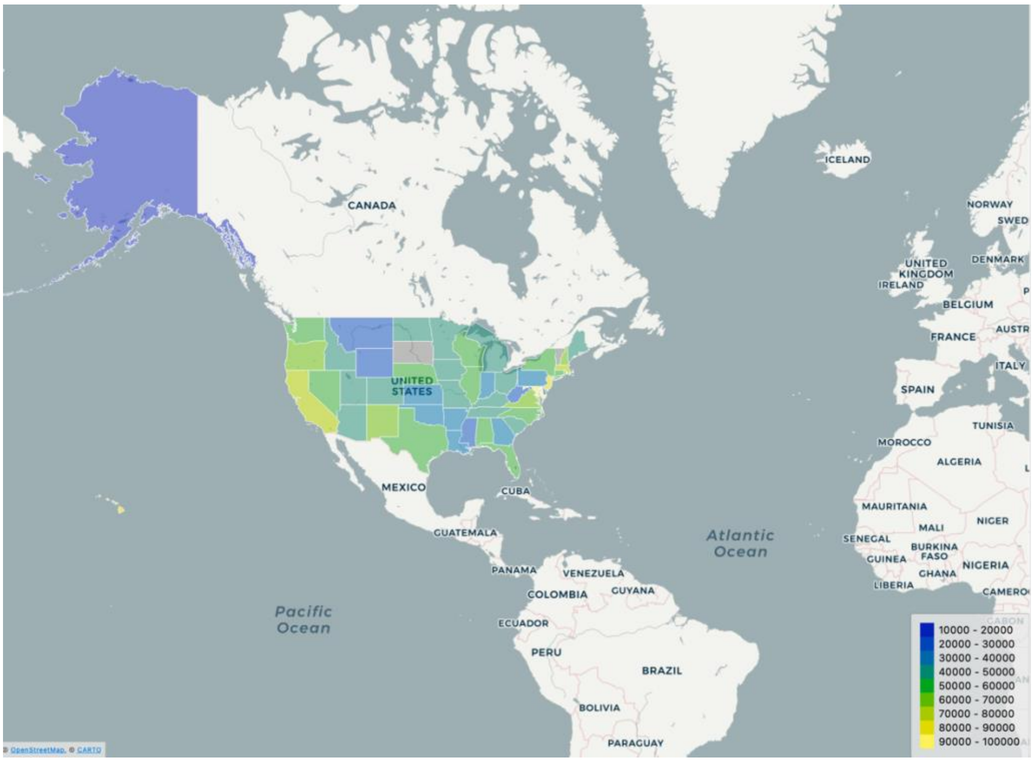
CoW ANOVA: 6.898(p=0.000, N=5000)
Region ANOVA: 4. 443 (p=0.001, N=5000)

Null Hypothesis 1 is rejected as there is evidence that CoW has significant impact on income between regions/states. With the f-value being 6.898, it shows a high variance between CoW and the regions. As $p < 0.001$, the differences seen are unlikely to be from random variation.

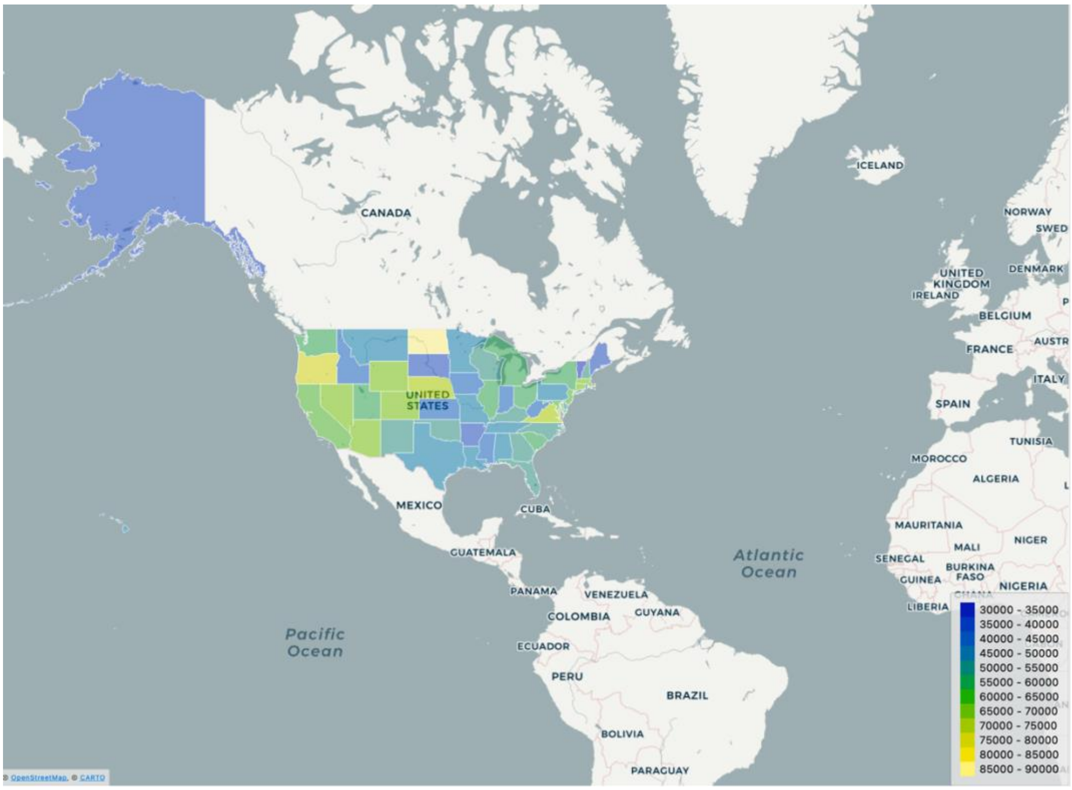
Null Hypothesis 2 is rejected. With the f-value being 4.443, it shows a high variance. Also unlikely to be from random variation $p = 0.001$.

Below are maps showing the highest earners by Cow for each state.

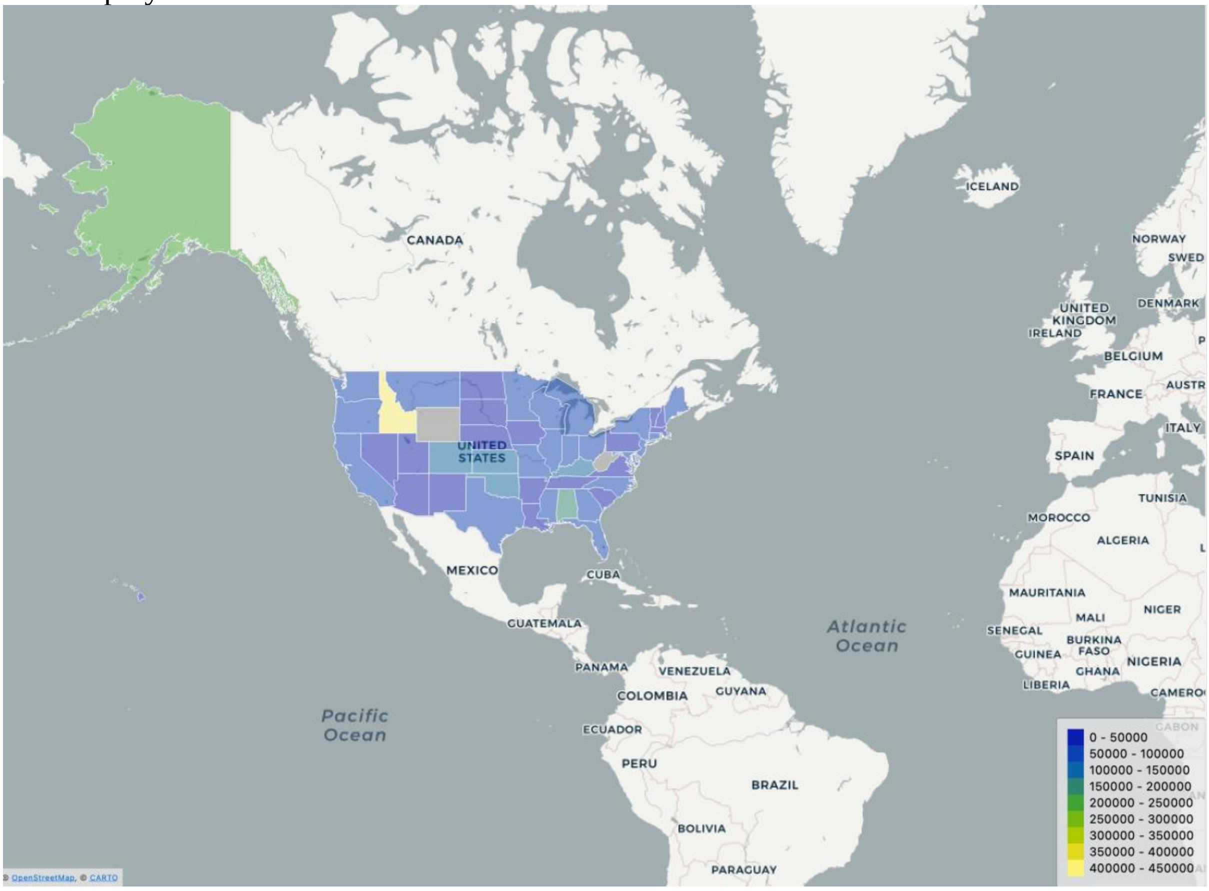
Government:



Private:



Self-employed:



There were a lot of missing values for “No Pay”, therefore this map was excluded.