Stat426 HW1B

Data: Ecdat Wages Data: data(Griliches)
HTML: https://vincentarelbundock.github.io/Rdatasets/doc/Ecdat/Griliches.html

Single Numerical Variable:

Summary stats of select variables:

Summary statistics:

Column	n	Mean	Variance	Std. dev.	Std. err.	Median	Range	Min	Max	Q1	Q3
med	758	10.91029	7.5137381	2.7411199	0.099561957	12	18	0	18	9	12
iq	758	103.8562	185.46807	13.618666	0.49465223	104	91	54	145	95	114
kww	758	36.573879	53.322804	7.3022465	0.26522954	37	44	12	56	32	41
age	758	21.835092	8.8908673	2.9817557	0.10830225	22	14	16	30	20	24
school	758	13.405013	4.9810581	2.2318284	0.081063658	12	9	9	18	12	16
expr	758	1.7354288	4.4333092	2.1055425	0.076476747	0.96	11.444	0	11.444	0.277	2.442
tenure	758	1.8311346	2.8010373	1.67363	0.060788978	1	10	0	10	1	2
lw	758	5.6867388	0.18399755	0.42894935	0.015580142	5.684	2.446	4.605	7.051	5.38	5.991

Confidence Interval at 95% confidence level:

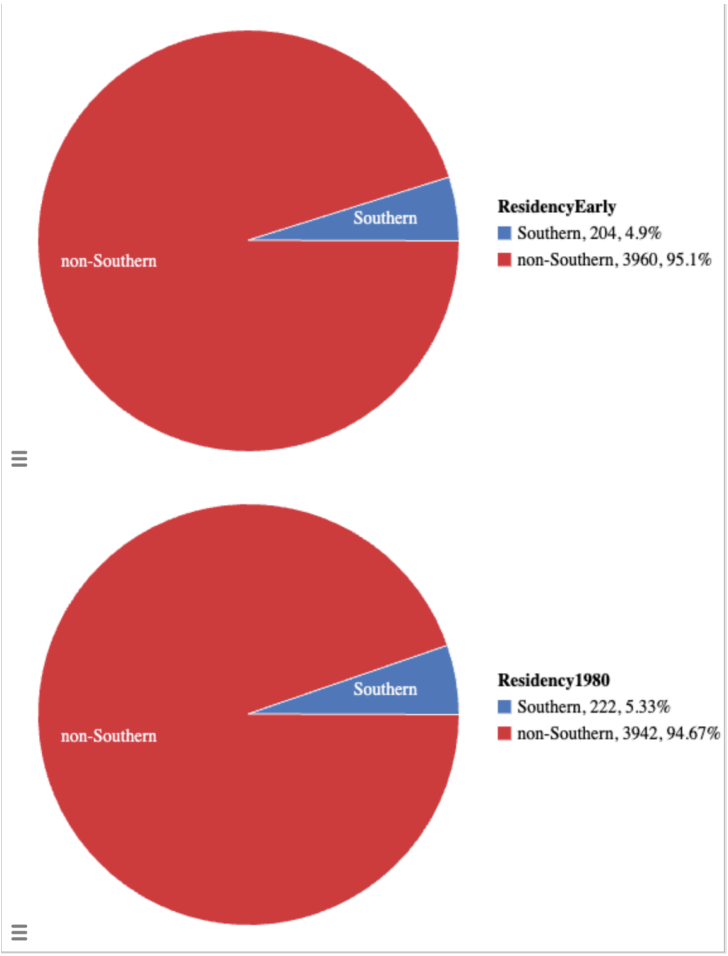
One sample T confidence interval:
μ : Mean of variable

95% confidence interval results:

Variable	Sample Mean	Std. Err.	DF	L. Limit	U. Limit
lw	5.6867388	0.015580142	757	5.6561534	5.7173242
lw80	6.8265554	0.014889211	757	6.7973264	6.8557845

I choose to select the means of logWage and logWage from 1980. I choose these variables because I wondered if these same variables may differ over time. The output shows that we can be 95% confident that the true population mean of logWage is between 5.66 and 5.72. We can also be 95% confident that the true population mean of logWage80 is between about 6.8 and 6.86. These findings are interesting because there seems to be a noticeable difference in means from the original sampling in 1976 to 1980. (I think the original data was sampled in 1976 and was then updated in 1980).

Pie Chart of Categorical Variable:
I re-coded some variables to be more descriptive: for residents living in the south (yes or no). For rns: residents living in the south we changed it to Southerner non-Southerner
I then did this for rns80 which was later sampled data in 1980. I changed the code the same way. Just to look at the data over time I made 2 pie charts to compare differences over time.



These charts show very small differences in residents over the sample difference period. To explore more I did a chi square to compare over time the people who are southerners and non-southerners.

Contingency table results:
Rows: ResidencyEarly
Columns: Residency1980

Cell format
Count (Expected count)

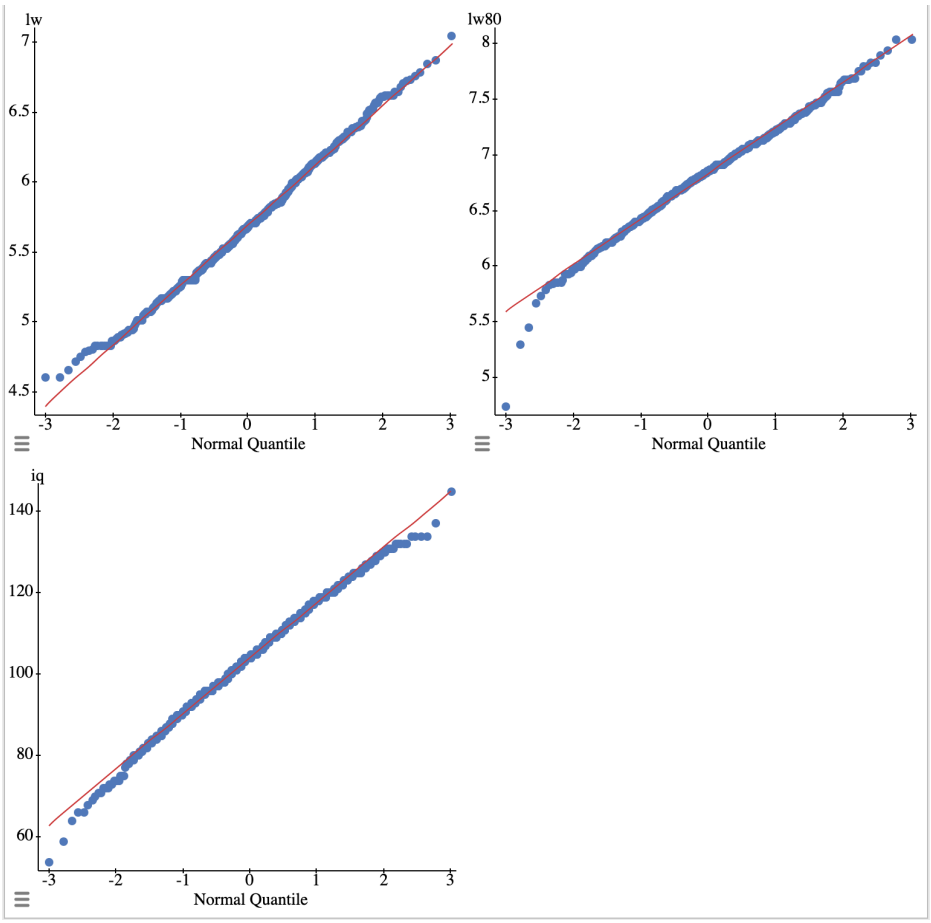
	Southern	non-Southern	Total
Southern	185 (10.88)	19 (193.12)	204
non-Southern	37 (211.12)	3923 (3748.88)	3960
Total	222	3942	4164

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	1	3096.3788	<0.0001

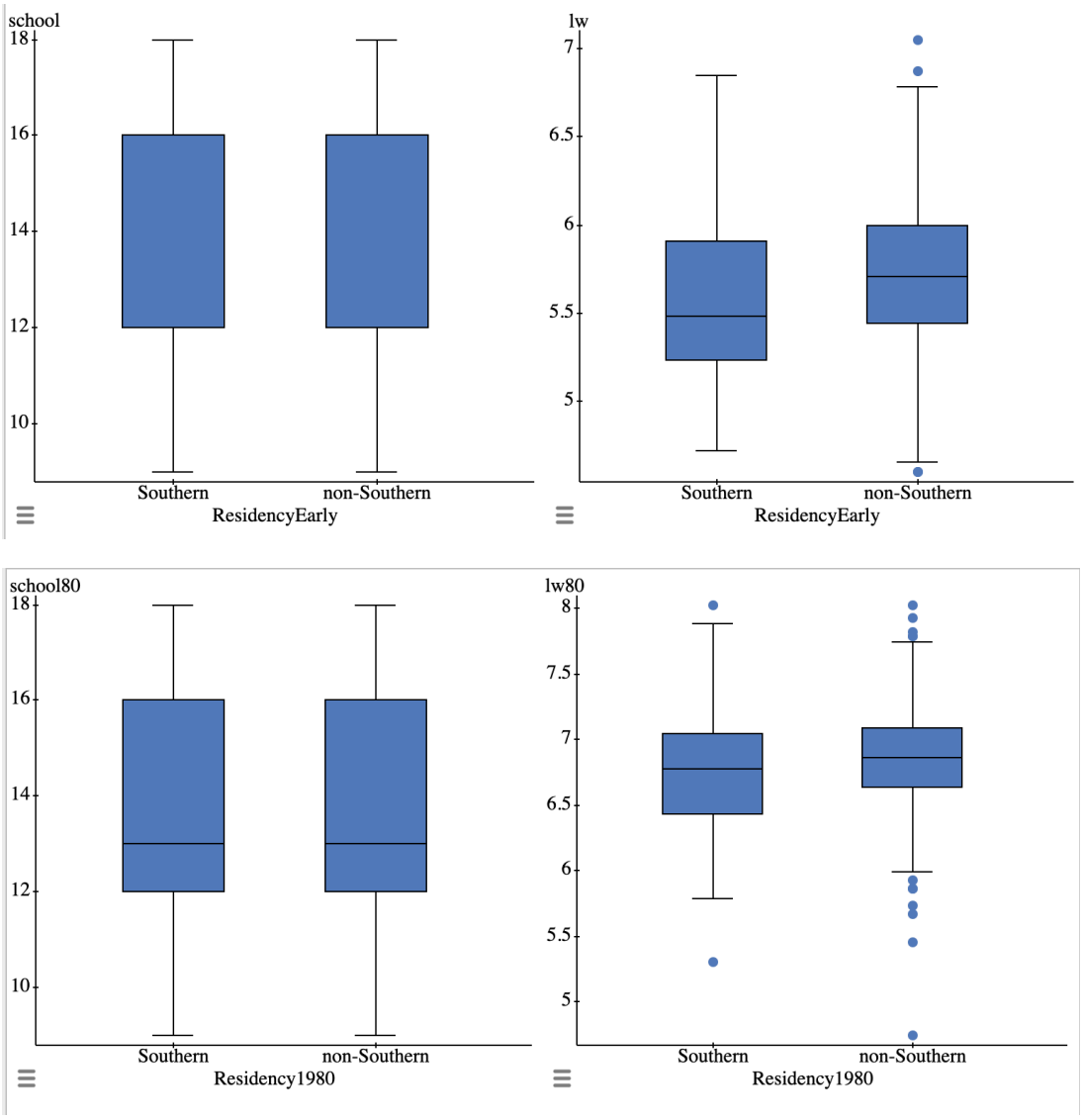
The chi square test demonstrated that 185 people are true southerners, 19 left the south, 37 moved to the south, and 3923 people were non southerners and never moved to the south. The results were significant with a very small p-value of <0.00001 meaning that within this sample, people tend to live in the same place over the duration of sampling–which I believe to be 4 years.

Looking at the QQ plots:



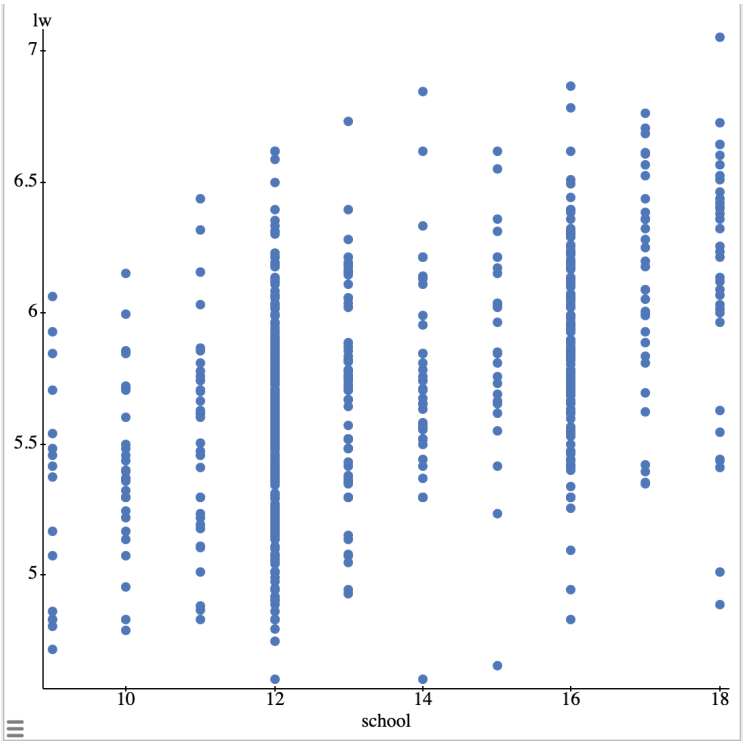
I choose, again, to look at the logWage and logWage80 and the trends along the red line are generally quite closely fit so we can assume the sample is normally distributed. I also wanted to look at IQ–because I find it to be an interesting variable–and this again seemed to be quite normal. Towards the tails of all the graphs, we can see some deviation in the trends, but I am assuming this data has been worked with because there are logs applied to some variables like the wage.

BoxPlots:
I wanted to look at the wage compared to education and then further group by location of living.

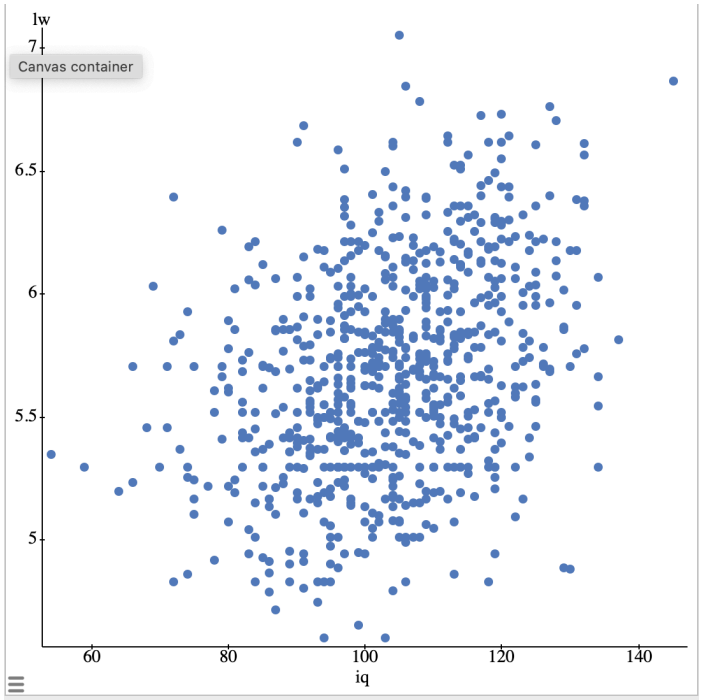


The top boxplots show the early sampling of data. We can see that both southerners and non-southerners have similar levels of education which appears to be a similar distribution as in the later sampling. Where the data seems to deviate over time is where we look at wages. In the early sampling (the top plots) we see that the mean wages for people who are non-southerners seems to be high than southerners. For the data sampled in 1980 we see that the mean wages are more similar but non-southerners have far more outliers than in the original sampling.

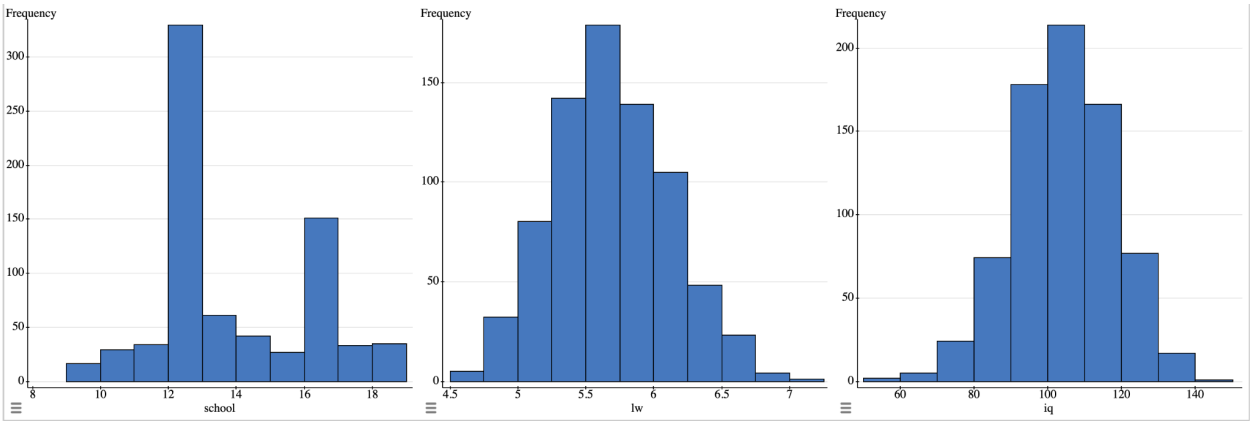
Schooling vs. LogWages:



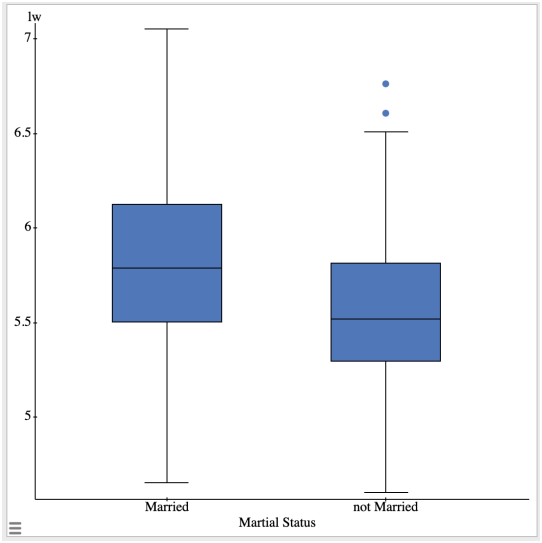
IQ vs. LogWages:



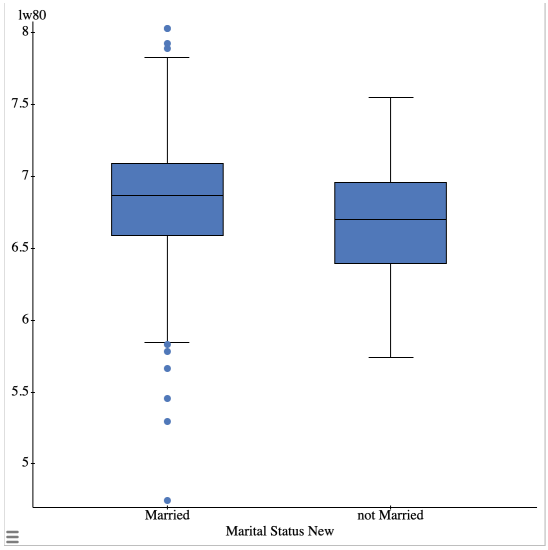
Histograms: School, LogWage, and IQ



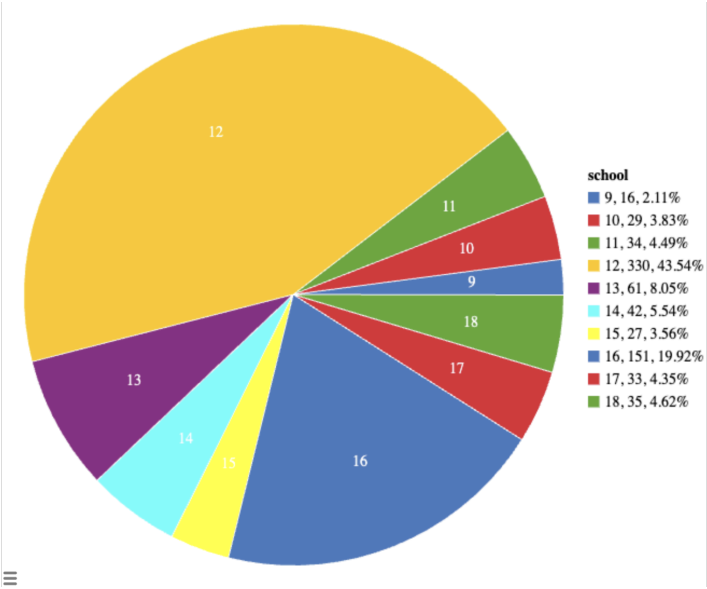
Box Plot for LogWages relative to marital status:

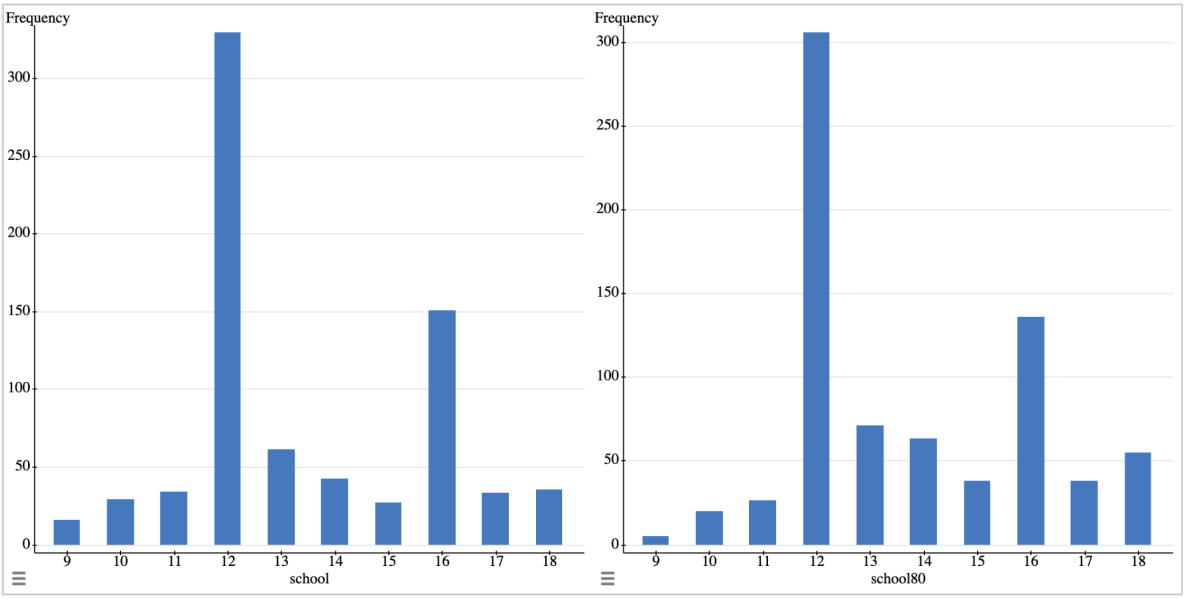


Boxplot for wages 1980 relative to marital status 1980:



Pie Chart by group & Bar Chart to follow up





95% CI by grouping:
Wages by location of living: Metro or non-Metro resident & Southerner or non-Southerner from the Old Data (1976)

One sample T confidence interval:
Group by: City Residency New
μ : Mean of lw

95% confidence interval results:

City Residency New	Sample Mean	Std. Err.	DF	L. Limit	U. Limit
Metro	5.7465918	0.018331223	533	5.7105815	5.7826021
non-Metro	5.5440536	0.027278507	223	5.4902969	5.5978102

One sample T confidence interval:
Group by: ResidencyEarly
μ : Mean of lw

95% confidence interval results:

ResidencyEarly	Sample Mean	Std. Err.	DF	L. Limit	U. Limit
Southern	5.5810833	0.031801682	203	5.5183794	5.6437873
non-Southern	5.7256444	0.017543317	553	5.6911847	5.7601041

Wages 1980 by location of living: Metro or non-Metro resident & Southerner or non-Southerner from the New Data (1980)

One sample T confidence interval:
Group by: City Residency 1980
μ : Mean of lw80

95% confidence interval results:

City Residency 1980	Sample Mean	Std. Err.	DF	L. Limit	U. Limit
Metro	6.889537	0.017160023	539	6.8558283	6.9232458
non-Metro	6.6705459	0.026829125	217	6.6176668	6.7234249

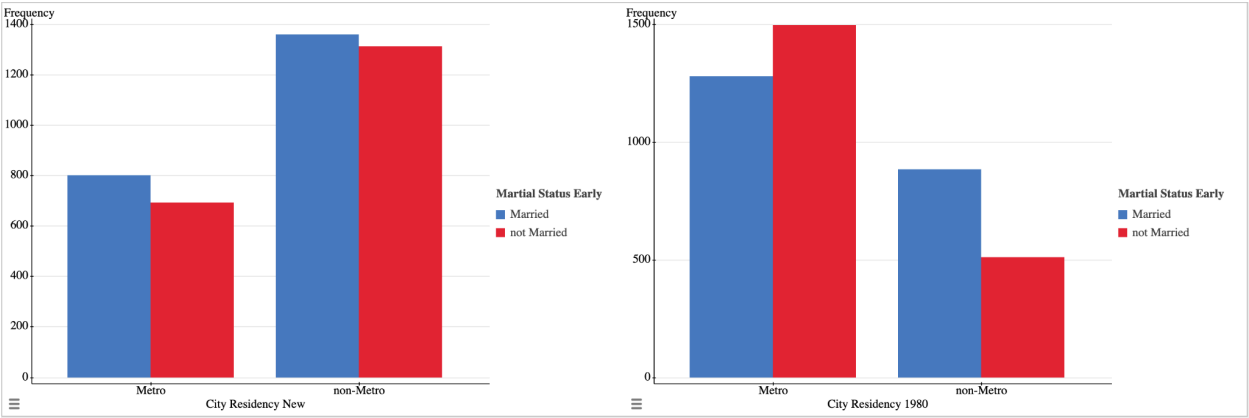
One sample T confidence interval:
Group by: Residency1980
μ : Mean of lw80

95% confidence interval results:

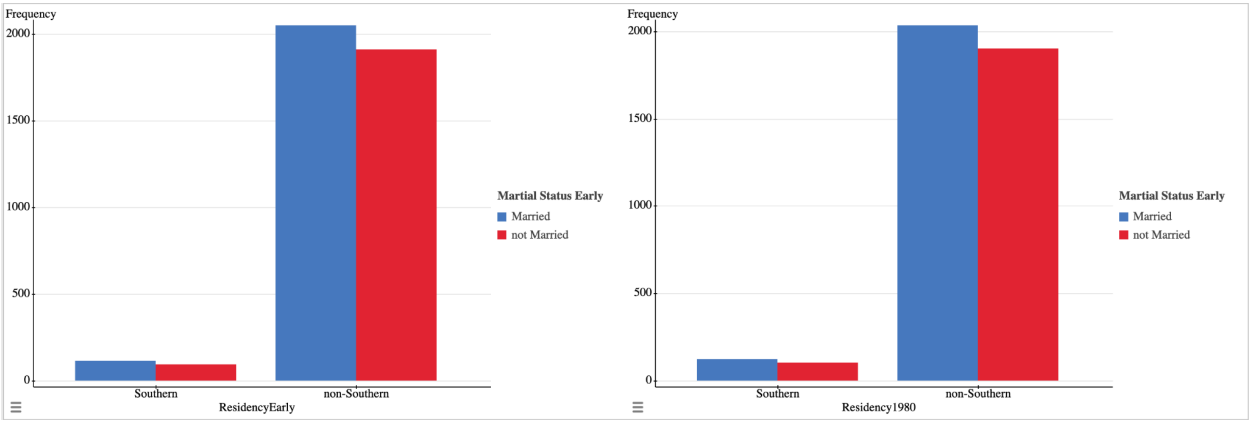
Residency1980	Sample Mean	Std. Err.	DF	L. Limit	U. Limit
Southern	6.7506486	0.030242285	221	6.6910485	6.8102488
non-Southern	6.8579944	0.016757423	535	6.825076	6.8909128

I wanted to compare these wages over time from the older data from 1976 to the newer data from 1980. I wanted to further compare the intervals based on the location of residency from the applicant. Based on the location grouping in the older data we see that there is a numerical difference in wage intervals (may or may not be significant). We also see a numerical difference in wages from the older 1976 data to the 4 year older 1980 data.

Bar Plot using Categorical Data:
Married or not married based on location of residence: Metro or non metro 1976 & 1980



Married or not married based on location of residence: South or non-South 1976 & 1980



Two Sample T-test between 1976 logWage and 1980 logWage:

Two sample T hypothesis test:

μ_1 : Mean of lw

μ_2 : Mean of lw80

$\mu_1 - \mu_2$: Difference between two means

$H_0 : \mu_1 - \mu_2 = 0$

$H_A : \mu_1 - \mu_2 \neq 0$

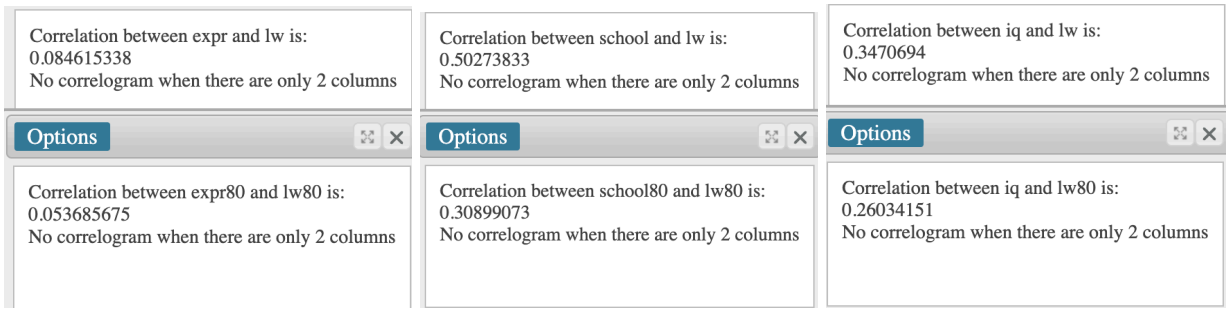
(without pooled variances)

Hypothesis test results:

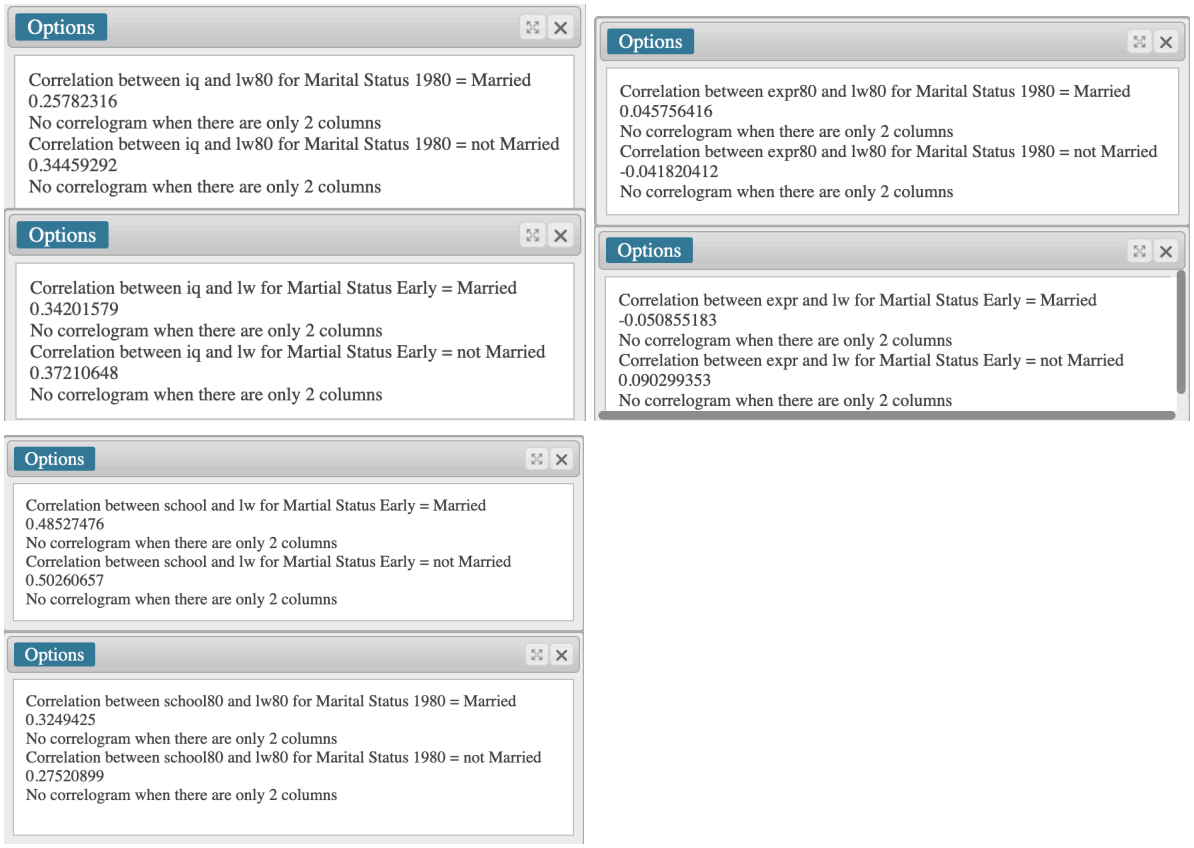
Difference	Sample Diff.	Std. Err.	DF	T-Stat	P-value
$\mu_1 - \mu_2$	-1.1398166	0.021550625	1510.8955	-52.89019	<0.0001

Correlation between numerical variables:

I attempted numerous tests of correlation, but none of these results are necessarily strong.



Correlation between numerical variables by group (Marital Status):



Once again, I ran the same few tests as above, and these were the outputs. These outputs do not seem to demonstrate any eye-catching relationship of correlation between the few numerical variables even when further divided by group of marital status.

Two Paired Numerical variables:

Paired T hypothesis test:

$\mu_D = \mu_1 - \mu_2$: Mean of the difference between lw and lw80

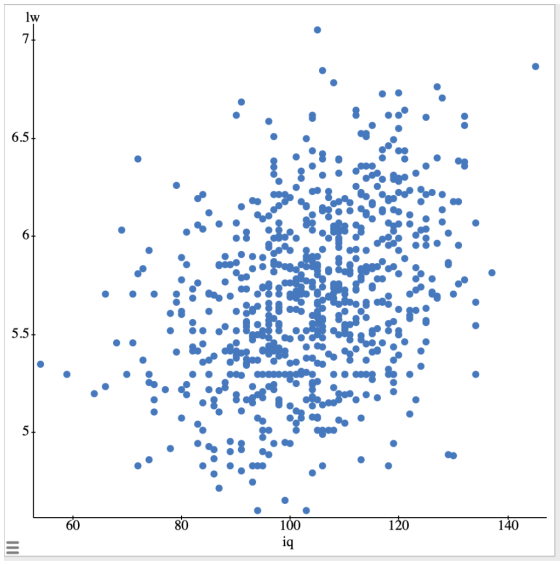
$H_0 : \mu_D = 0$

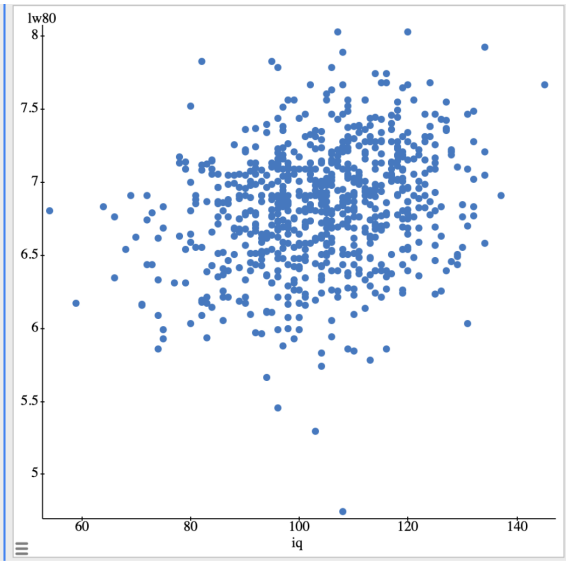
$H_A : \mu_D \neq 0$

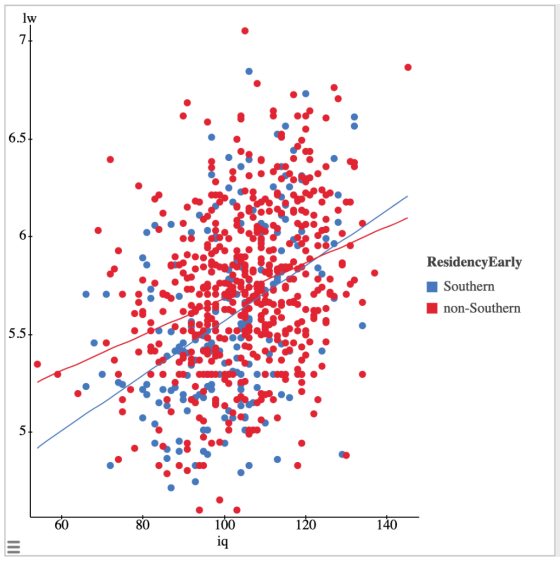
Hypothesis test results:

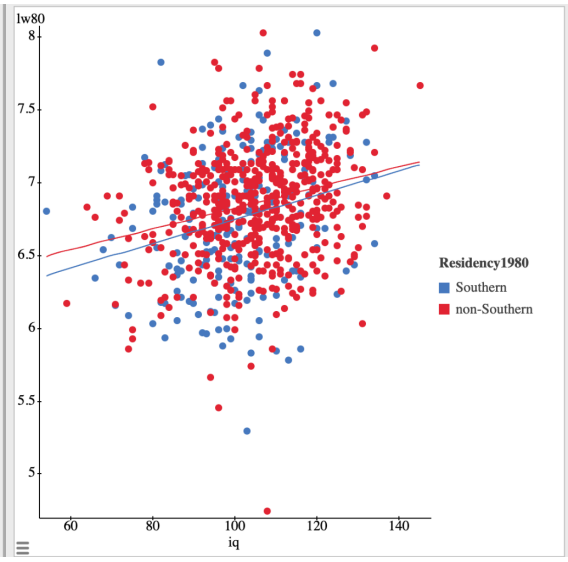
Difference	Mean	Std. Err.	DF	T-Stat	P-value
lw - lw80	-1.1398166	0.016172238	757	-70.47983	<0.0001

Scatterplots for numerical variables: basic and advanced plots









Z-test with 2 samples grouped:
I did logWage by marital status

Two sample Z hypothesis test:
μ_1 : Mean of lw where "Martial Status Early" = "Married" (Std. dev. not specified)
μ_2 : Mean of lw where "Martial Status Early"="not Married" (Std. dev. not specified)
$\mu_1 - \mu_2$: Difference between two means
$H_0 : \mu_1 - \mu_2 = 0$
$H_A : \mu_1 - \mu_2 \neq 0$

Hypothesis test results:

Difference	n_1	n_2	Sample mean	Std. err.	Z-stat	P-value
$\mu_1 - \mu_2$	390	368	0.25562892	0.029674592	8.6144037	<0.0001

ANOVA:

Analysis of Variance results:
Responses: lw
Factors: Martial Status Early

Response statistics by factor

Martial Status Early	n	Mean	Std. Dev.	Std. Error
Married	390	5.8108436	0.43244307	0.021897606
not Married	368	5.5552147	0.38418233	0.020026889

ANOVA table

Source	DF	SS	MS	F-Stat	P-value
Martial Status Early	1	12.372663	12.372663	73.701651	<0.0001
Error	756	126.91348	0.16787498		
Total	757	139.28614			

Tukey HSD results (95% level)
Married subtracted from

	Difference	Lower	Upper	P-value
not Married	-0.25562892	-0.31408306	-0.19717477	<0.0001

In this ANOVA test, I looked to see if marital status had any significant effect on logWages. The output shows that people who are married have a higher logWage than people who are not married. We can be sure that these results are statistically significant due to the very small p-value of <0.0001. On average people who are married have an increased logWage of .256 over people who are not married.

Analysis of Variance results:
Responses: lw80
Factors: Marital Status 1980

Response statistics by factor

Marital Status 1980	n	Mean	Std. Dev.	Std. Error
Married	681	6.8437841	0.40643505	0.015574625
not Married	77	6.6741818	0.4117919	0.046928043

ANOVA table

Source	DF	SS	MS	F-Stat	P-value
Marital Status 1980	1	1.9899044	1.9899044	12.014149	0.0006
Error	756	125.21634	0.16563008		
Total	757	127.20624			

Tukey HSD results (95% level)
Married subtracted from

	Difference	Lower	Upper	P-value
not Married	-0.16960232	-0.26565937	-0.073545279	0.0006

When I ran the ANOVA again with the later data of logWage80 and marital status 1980, we still see the same trend in logWages between people who are married and people who are not married. The difference in logWages is actually larger in the later years of 1980. If a person is married in 1980 then their logWage is .267 higher than a person who is not married. Once again this finding is statistically significant due to the very small p-value of 0.0006.

Summary table with Summary Stats on Numericals BY GROUPING variable:

Summary statistics for iq:
Group by: Martial Status Early

Martial Status Early	Mean	Std. dev.	Median	IQR	Mode
Married	104.23077	13.317097	105	18	104
not Married	103.45924	13.938302	103	18.5	98

Summary statistics for school:
Group by: Martial Status Early

Martial Status Early	Mean	Std. dev.	Median	IQR	Mode
Married	13.669231	2.3923243	12	4	12
not Married	13.125	2.013747	12	3	12

Summary statistics for expr:
Group by: Martial Status Early

Martial Status Early	Mean	Std. dev.	Median	IQR	Mode
Married	2.3492462	2.4560797	1.462	3.447	0
not Married	1.0849158	1.3877365	0.462	1.451	0

Summary statistics for tenure:
Group by: Martial Status Early

Martial Status Early	Mean	Std. dev.	Median	IQR	Mode
Married	2.2282051	1.9150934	2	2	1
not Married	1.4103261	1.2430085	1	0	1

Summary statistics for lw:
Group by: Martial Status Early

Martial Status Early	Mean	Std. dev.	Median	IQR	Mode
Married	5.8108436	0.43244307	5.7885	0.62	5.298
not Married	5.5552147	0.38418233	5.521	0.519	5.298

Summary statistics for iq:
Group by: Marital Status 1980

Marital Status 1980	Mean	Std. dev.	Median	IQR	Mode
Married	103.67107	13.358095	104	17	104
not Married	105.49351	15.741015	105	25	Multiple modes

Summary statistics for school80:
Group by: Marital Status 1980

Marital Status 1980	Mean	Std. dev.	Median	IQR	Mode
Married	13.660793	2.2131436	13	4	12
not Married	14.116883	2.2003231	14	4	12

Summary statistics for expr80:
Group by: Marital Status 1980

Marital Status 1980	Mean	Std. dev.	Median	IQR	Mode
Married	11.576643	4.2456527	11.292	6.473	1
not Married	9.7812468	3.5197122	9.055	3.806	No mode

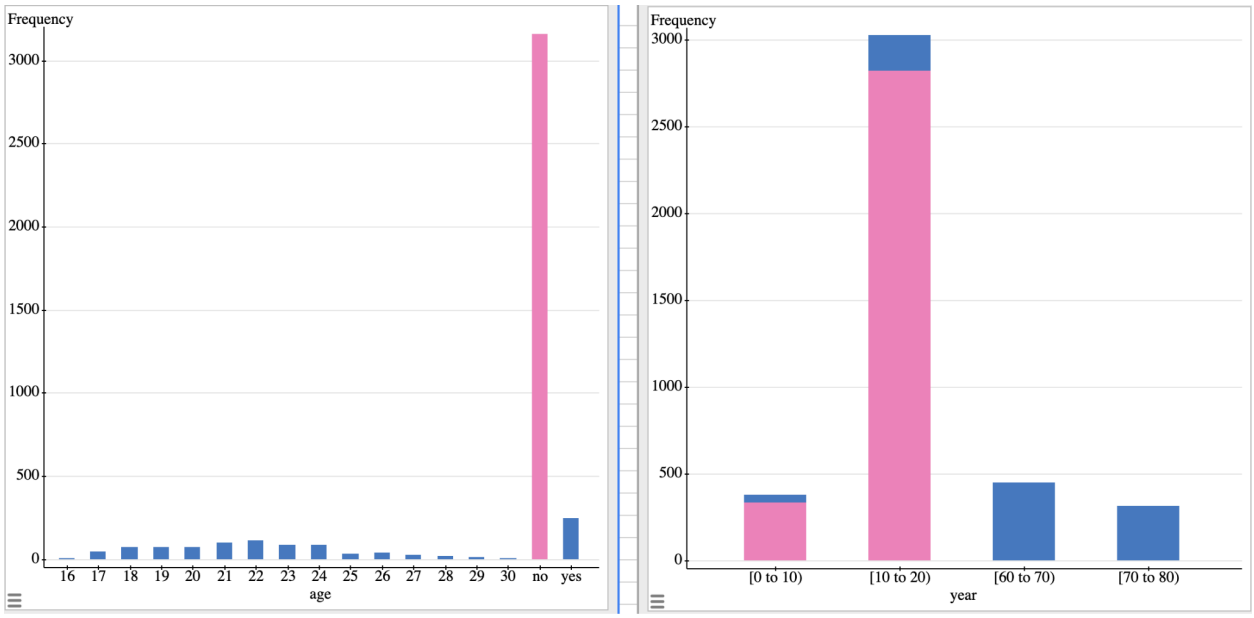
Summary statistics for tenure80:
Group by: Marital Status 1980

Marital Status 1980	Mean	Std. dev.	Median	IQR	Mode
Married	7.4552129	5.0849958	7	8	2
not Married	6.5454545	4.683525	5	7	1

Summary statistics for lw80:
Group by: Marital Status 1980

Marital Status 1980	Mean	Std. dev.	Median	IQR	Mode
Married	6.8437841	0.40643505	6.869	0.5	6.908
not Married	6.6741818	0.4117919	6.7	0.56	Multiple modes

Remaining Questions:

1. 

I found this in my data. Looking at the year dates that the survey data was obtained on the right, and the people who entered their age when the survey was taken. It seems that a lot of the sample may have said no to entering their age at the time the survey was taken, but that a lot of the data was actually collected in the 1910s to 1920s. I believe that this is what the bar graphs are showing. Very confounding. I am not sure this entirely has an impact on the data, but it can explain some of the behavior of the surveyors.

2. I also did not look into some of my other variables like tenure, experience, knowledge of the world, schooling, mothers education. These variables may be necessary later on in deeper discoveries and explaining some of the things I found in this writeup.
3. I wish I had a gender variable. I think that gender can easily be a confounding variable in this dataset.