1. Assume that you are hired to investigate the causal effect between being raised in high-poverty neighborhoods in the US and future outcomes during adulthood (such as health, well-being, social networks and economic self-sufficiency). Employ the sources in the Datasets file in Blackboard and succinctly answer the following questions:

(a) Mention a suitable dataset that can help you answer the question above. Provide its name and the website where it can be downloaded.

It's called County-level Employment Rates by Gender and Parent Income Quintile and other County-level Covariates

Found here:
https://opportunityinsights.org/data/?geographic_level=0&topic=108&paper_id=0#resource-listing

https://opportunityinsights.org/wp-content/uploads/2018/04/gender_cty_readme.pdf

(b) What is the sample size in this dataset? Is this a reasonable number for your research?
The sample size = 3,139

This is more than enough because a sample size of 30 is more than enough to make a sound conclusion about a population.

(c) Briefly describe the data you found in part (a). Using the codebook discuss which variables are crucial to answer the research question posed in the statement above (no more than 10 lines).

The dataset I found reports on the employment rates of age 30 by the county of childhood for children whose parents are permanent residents sorted by child's gender and parent's income quintile

The most important variables are cty and [outcome]_q[quintile]_[gender][fam] because it allows us to build a regression model to investigate the causal effect between being raised in high-poverty neighborhoods in the US and future outcomes during adulthood.

2. Suppose you are a researcher interested in studying the relationship between household characteristics and future educational outcomes of children. You have been advised that one dataset which satisfies your requirements is the Early Childhood Longitudinal Study, Birth Cohort. Try to find the data through the sources talked about in the Stata lecture. In order to answer the following questions, additionally you will need to locate the codebooks of this database. (Note: You do not need the data, the codebooks and webpage pdfs contain all information you require)

(a) Briefly describe the objectives of this study and the different rounds of the survey. Mention the methods employed for data collection. At what ages are the interviews conducted? (Your answer should not exceed 10 lines).

ECLS-B was designed to provide policy makers, researchers, child care providers, teachers, and parents with detailed information about children's early life experiences. The data collected for ECLS-B focuses on children's health, development, care, and education during the formative years from birth to kindergarten. Information about these children was collected when they were about 9 months old(2001-02), 2 years old(2003-04), and 4 years(2005-06).

In every round of data collection, children participated in assessment activities and parent respondents were asked about themselves, their families and their children. In addition, when the child is 2 years old, their child care/early education providers were asked to provide information about their own experience and training and their setting's learning environment. When the ECLS-B children were in kindergarten, their teachers were asked to provide information about children's early learning experiences and the school and classroom environments.

(b) Describe which are the restrictions for the use of this database.
Due to NCES's confidentiality legislation, ECLS-B case level data are available only to qualified researchers who are granted a restricted-use data license. When presenting analyses, preparing manuscripts, publishing ECLS-B results, or sending email correspondence (including to NCES staff), analysts must comply with ECLS-B rounding rules. Unweighted sample sizes must be rounded to the nearest 50. Further, all presentations and manuscripts prepared using ECLS-B restricted-use data must be sent to the NCES Data Security Office (IESData.Security@ed.gov) for disclosure review prior to publication or presentation, as is required by the terms of the NCES restricted-use data license.

(c) How many children are classified as low birth weight in the first round of the survey?

Birth weight			
Normal birth weight	7,850	3,696,100	93
Moderately low birth weight	1,650	247,900	6
Very low birth weight	1,150	51,400	1

(d) Describe the groups of variables available in the first round. Classify them in child characteristics, mother characteristics and household characteristics.

child characteristics: Child's sex, Child's race/ethnicity, Child's assessment age, Birth Weight, Born premature

mother characteristics: Plurality, Maternal age at child's birth, Mother's education

household characteristics: Poverty status, Primary language spoken in home

(e) Choose two variables you could employ as baseline characteristics of the household. Describe how these variables would be relevant for studying future outcomes of children.

I would employ poverty status and primary language spoken in home as baseline as the characteristic of household because these two factors could affect the future of the child's success. When a family is in poverty, it can affect the future child's abilities to perform well in school or in a career. When a child is in a multi language household or a primary language household, it can affect the future learning abilities of the child or it might hinder the child's abilities.

(f) Calculate the nonresponse rate between the initial number of individuals interviewed and the two following rounds of the survey.

Response rate is given but nonresponse rate is not given: nonresponse rate = 100 - Response rate

The weighted unit response rate for the 9-month parent interview = 74.1%
The weighted unit nonresponse rate for the 9-month parent interview = 100 - 74.1 = 25.9%

The weighted unit response rate for the 2-year parent interview = 93.1%
The weighted unit nonresponse rate for the 2-year parent interview = 100 - 93.1 = 6.9%

The weighted unit response rate for the preschool parent interview = 91.3%
The weighted unit nonresponse rate for the preschool parent interview = 100 - 91.3 = 8.7%

(g) Suppose you are interested in studying how socio-emotional skills are developed before the age of two. Describe which assessments included in this study could be employed for this purpose. Does the study have similar assessments for higher ages?

I would look at assessments for 9 months, 2 years, and preschool and look at the parent interview questions/answers(section HI015 because it tells us household income).
Yes the study also has similar assessments for higher ages.

(h) Describe which measurements can be used to analyze the cognitive skills of children in kindergarten.

Care and Education setting where the child spends most hours

3. This problem asks you to work directly with Stata. Suppose you are a researcher interested in studying the labor market outcomes of recent college graduates. One public-use, suitable dataset for this purpose is the National Survey of College Graduates (NSCG). In order to answer the following questions, you will need to use the attached documentation to identify the variables of interest.

(a) Explore the survey using the interview questionnaire. Based on this, write down one scientific question (related to the topic mentioned above) which could be answered using the NSCG.

Does obtaining a degree affect employment opportunities?

(b) Use the interview questionnaire provided with the database to identify the variables related to hours worked per week, weeks worked per year and year earnings. Notice that information about weeks worked can be derived using two variables. Also note that the NSCG15 value of 98 for hrs worked per week = logical skip.

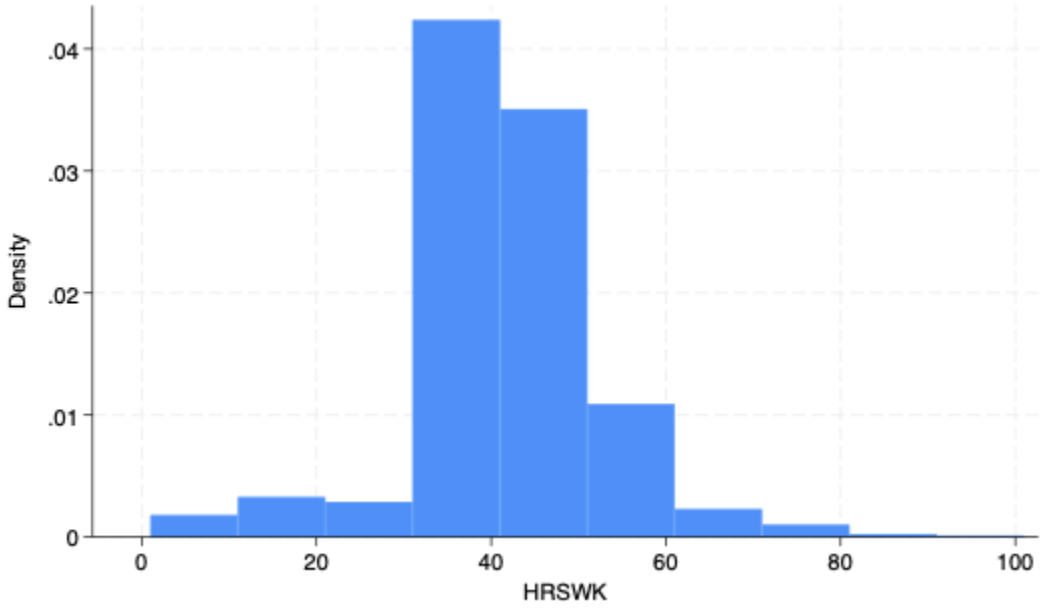
hrswk is hours worked per week
wkslyr is weeks worked per year
earn is year earnings

(c) After handling invalid values properly, create a table showing the mean and standard deviation of the three variables described in part (b) for men and women separately.

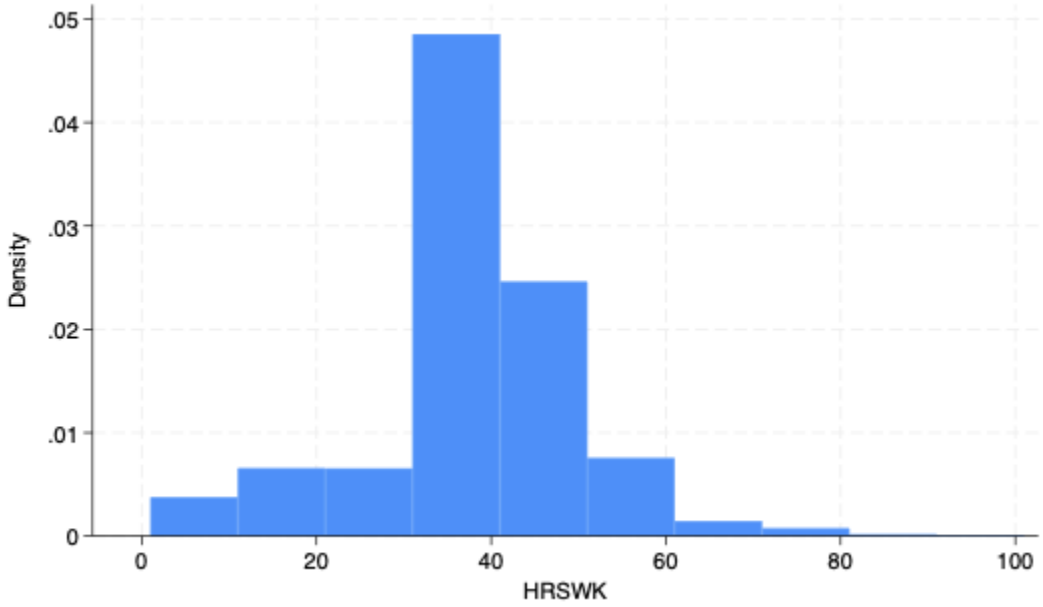
summarize hrswk wkslyr earn

Variable	Obs	Mean	Std. dev.	Min	Max
hrswk	76,814	42.02291	11.90968	1	96
wkslyr	91,000	90.94916	21.0914	1	98
earn	91,000	1134031	3047246	0	9999998

(d) In order to see the distribution of hours worked per week, crate a histogram of this variable for men and women. Plot the density in the y-axis and use a bin width of 10 for the x-axis.
For males:



For females:



(e) Create a new variable lnhourwage defined as the (natural) logarithm of year earnings divided by total hours worked during the year. Produce a table showing the mean, standard deviation

and percentiles 10th and 90th of this variable for men and women separately. Drop observations which yield a negative value of this variable.

lnhourwage

	Percentiles	Smallest		
1%	7.965546	2.062634		
5%	9.952278	2.079442		
10%	10.69195	2.120264	Obs	76,426
25%	11.4204	2.525729	Sum of wgt.	76,426
50%	11.97278		Mean	11.79433
	Largest		Std. dev.	1.042562
75%	12.36819	17.70733		
90%	12.72487	17.99501	Variance	1.086935
95%	12.9968	18.40048	Skewness	-1.545767
99%	13.89062	18.75715	Kurtosis	9.597071

(f) Use the interview questionnaire to identify the variable which indicates whether a respondent changed employer and/or job between 2013 and 2015, as well as the variables describing the reason of change in case the employer is different between these two years. What is the proportion of respondents who stayed with the same employer and job during this period?

(g) As a researcher, you are also interested in studying how the gender wage gap varies across major fields. Using the variable related to the first bachelor degree (nbamemg) and your variable lnhourwage create a table showing the mean hourly wage for women and men across different majors. Which is the one that presents the higher wage gap?

Means, Standard Deviations and Frequencies of lnhourwage

NBAMEMG	GENDER F	GENDER M	Total
1	11.695639	12.065727	11.940078
	1.0342117	.90467935	.9665803
	2180	4241	6421
2	11.567104	11.786427	11.670634
	1.0300273	1.0624478	1.0511126

```
          |      4419       3951 |      8370
----------+----------------------+----------
        3 | 11.605673  11.959047 | 11.835687
          | .97422959  .97876026 | .99149132
          |      1595       2974 |      4569
----------+----------------------+----------
        4 | 11.463384  11.809445 |  11.60557
          | 1.1133166  1.1092468 | 1.1245752
          |      9310       6493 |     15803
----------+----------------------+----------
        5 | 11.974905  12.209226 | 12.161777
          | .81728479  .74152551 | .76328478
          |      3436      13532 |     16968
----------+----------------------+----------
        6 | 11.677304  11.911296 | 11.757823
          | 1.0365532  .97771073 | 1.0226872
          |      5133       2693 |      7826
----------+----------------------+----------
        7 | 11.377098  11.788487 | 11.566149
          | 1.1617183  1.0571154 | 1.1335266
          |      7815       6645 |     14460
----------+----------------------+----------
        8 | 11.636037  12.112991 | 11.915704
          | 1.2676961  1.0663895 | 1.1773021
          |       831       1178 |      2009
----------+----------------------+----------
    Total | 11.564665  11.985514 |  11.79433
          | 1.0842618  .96574855 | 1.0425615
          |     34719      41707 |     76426
```

.
.
Highest wage gap seems to be when nbamemg is at 8.

(h) Run a regression of hourly wages on education separately for men and women. How does the parameter of education differ across gender?

For men:

Source	SS	df	MS	Number of obs	=	41,707
				F(1, 41705)	=	92.35
Model	85.9410169	1	85.9410169	Prob > F	=	0.0000
Residual	38812.0045	41,705	.930631928	R-squared	=	0.0022
				Adj R-squared	=	0.0022
Total	38897.9456	41,706	.932670253	Root MSE	=	.96469

lnhourwage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
educ	.0196062	.0020402	9.61	0.000	.0156073	.0236051
_cons	11.70736	.0293278	399.19	0.000	11.64988	11.76484

For female:

Source	SS	df	MS	Number of obs	=	34,719
				F(1, 34717)	=	134.13
Model	157.082014	1	157.082014	Prob > F	=	0.0000
Residual	40658.2165	34,717	1.17113277	R-squared	=	0.0038
				Adj R-squared	=	0.0038
Total	40815.2985	34,718	1.17562355	Root MSE	=	1.0822

lnhourwage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
educ	.0300697	.0025964	11.58	0.000	.0249807	.0351587
_cons	11.13045	.0379395	293.37	0.000	11.05609	11.20482

The coefficient, std error and t values of education varies by gender.

(i) Create a variable of potential experience ptlexper, defined as age-education-6. Run a regression of hourly wages on education, potential experience and potential experience squared separately for men and women. Interpret your results.

For male:

Source	SS	df	MS	Number of obs	=	41,707
				F(3, 41703)	=	723.94
Model	1925.47176	3	641.823919	Prob > F	=	0.0000
Residual	36972.4738	41,703	.886566285	R-squared	=	0.0495
				Adj R-squared	=	0.0494
Total	38897.9456	41,706	.932670253	Root MSE	=	.94158

lnhourwage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
educ	.0241988	.0019976	12.11	0.000	.0202834	.0281142
ptlexper	.0836396	.0019356	43.21	0.000	.0798459	.0874334
ptlexper_sq	-.0011898	.0000298	-39.87	0.000	-.0012483	-.0011313
_cons	10.41749	.0402862	258.59	0.000	10.33853	10.49646

For female:

Source	SS	df	MS	Number of obs	=	34,719
				F(3, 34715)	=	262.00
Model	903.656834	3	301.218945	Prob > F	=	0.0000
Residual	39911.6416	34,715	1.14969442	R-squared	=	0.0221
				Adj R-squared	=	0.0221
Total	40815.2985	34,718	1.17562355	Root MSE	=	1.0722

lnhourwage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
educ	.0323963	.0025886	12.52	0.000	.0273226	.03747
ptlexper	.0647701	.002552	25.38	0.000	.0597681	.0697721
ptlexper_sq	-.0010491	.0000414	-25.37	0.000	-.0011302	-.0009681
_cons	10.27797	.0520461	197.48	0.000	10.17596	10.37998

R-squared = 0.0495 for men and R-squared = 0.0221
This means that 4.95% of the variation in hourly wages is explained by the model for men and 2.21% of the variation of hourly wages is explained by the model.

For men since the coefficient for education and ptlexper is positive. It means that for additional year of education, the hourly wage increases by approximately 2.42% in hourly wages and for additional year of potential experience is associated with about a 8.36% hourly wages.

For female since the coefficient for education and ptlexper is positive. It means that for additional year of education, the hourly wage increases by approximately 3.24% in hourly wages and for additional year of potential experience is associated with about a 6.47% hourly wages.