**Project Proposal:**

GROUP 3: KIRAN RANGWANI, LOGAN GUNNING, VASU SHARMA

**1. A succinct summary of idea in the form of a short descriptive title. (e.g., How does gender proportion impact average wage in a county?)**

“The Impact of Varying Levels of English Language Proficiency on an Individual’s Median Income”

This study compares the impact of different levels of English proficiency on an individual’s median income. Specifically, the study investigates whether there is a link between English proficiency and higher median income. The study's findings may have policy implications that direct attention to the value of language skill and its potential influence on earnings outcomes.

**2. Data: Describe the respondent- or county-level Census data you plan to use. Explain where it can be found (e.g., URL). Also explain. why these data will work for your project.**

The Census data that we are using is respondent level, and contains information on various demographic, social, and economic variables including income, age, sex, race, educational attainment, English proficiency, marital status, and occupation. The data can be accessed from the data.usa.gov website, which provides an interface to access and download data from a variety of federal agencies. The data is for all 50 states of the U.S., and it is for the year 2021. The data are useful for our project because they are recent, accessible, and include, according to the existing literature, relevant variables to address our research question.

URL:<https://data.census.gov/mdat/#/search?ds=ACSPUMS1Y2021&vv=PINCP,%2aAGEP&cv=RACWHT,RACBLK,MAR,SEX&rv=SCHL,POWSP&nv=ENG,RACPI,RACNH,RACAIAN,RACSOR,RACASN,SOCP&wt=PWGTP>

**3. Population: Who do you wish to study? (e.g., If you plan to analyze a particular subset of U.S. counties.)**

The project focuses on the characteristics of individuals residing in the United States with particular attention paid to English proficiency among respondents. More specifically, we are interested in studying the characteristics of individuals with varying degrees of English fluency, and the influence of these variables on an individual’s income. Our population data contains information for all 50 U.S. States.

We have created a dataset for the entire population where we performed data cleaning and added dummy columns wherever necessary. This dataset is created as per our current understanding and can be modified as our analysis moves forward.

<https://drive.google.com/file/d/1f1Xg6v2rRFXv83rns9RsYT6TgDNC92JK/view?usp=share_link>

**4. Sample: If you are analyzing a subset of respondents or U.S. counties, how do these data relate to the U.S. population? Likely the sample will be different in some important ways from our target. population. Be sure to address any ways they might be different.**

**Note: if you analyze counties, you must choose at least 250 counties for your sample. For example, you can choose to examine the lower Midwest, which includes Illinois with 102 counties, Indiana with 92 counties, and Ohio with 88 counties and totals 282 counties.**

We are planning to analyze all states on the East Coast of the United States as our sample. This constitutes roughly 36% of the population. Our sample is likely to differ demographically, economically, politically, and/or socially in comparison to the population, although the East Coast is relatively diverse. This is likely to influence to some extent the frequency of certain characteristics relative to the population.

The sample will consist of individual respondents from the following states:

1. Maine

2. New Hampshire

3. Massachusetts

4. Rhode Island

5. Connecticut

6. New York

7. New Jersey

8. Delaware

9. Maryland

10. Virginia

11. North Carolina

12. South Carolina

13. Georgia

14. Florida

**5. Response variable: What is the response variable you will be evaluating? Exactly how the variable is measured.**

The response variable or the variable that is dependent and the one that is influenced by one or more independent variables is Income. In our dataset, the income isthe *“*Total person's income (signed, use ADJINC to adjust to constant dollars)*”.* Our population data will have the income of people across the 50 states of the United States of America and our sample data will have the income of people in the east coast of the United States.

As for our response variable, we have a total amount in earnings for all people aged 16 and over and then the income ranges from $1 to $4209995.

We have dropped all the negative values because they denote loss for individuals and income of people with less than 15 years of age. We chose to do this because in the case of our project negative values don’t hold a lot of meaning as we are analyzing the impact level of English proficiency on income **earned.** We Haveimplemented log transformation in the income and created a new column namely transformed\_income. Log transformation of income is important because it helps to normalize the income and helps to reduce the impact of extreme values and makes the distribution of the variable more symmetric. Overall, log transformation of income can help to improve the accuracy and interpretability of statistical analyses involving income data.

To perform this transformation, we have added 1 to every row in this column to avoid because log of 0 is undefined.

Chart, waterfall chart

Description automatically generated

Chart, histogram

Description automatically generated

 **6. Explanatory variables: What variables will be you attempting to use to explain the response variable? Exactly how the variable is measured.**

Explanatory variables, also known as independent variables, are used to predict the outcome of response variables. These variables can be either continuous or categorical and may have a positive or negative impact on the response variable. Essentially, these variables are the factors that we want to test.

There are seven independent variables that we will be incorporating in our project which will also have our variable of interest, i.e. Ability to speak English. We will also have a variable for state as part of independent variables.

We define independent variables as:

1. **AGE**: The variable in the dataset is “AGEP”. In this column we have ages starting from 0 years to 96 years old. We will be considering individuals of the age of 18 years or above. This is because people in this age bracket could come into the labor force i.e., individuals who are 18 years or above are supposed to be earning.
2. **SEX**: The variable is “SEX”. The gender is in two categories, i.e., Male and Female. The raw data downloaded from the US census website has recoded values for age i.e., 1 for Male and 2 for Female, hence in our observations we will have values of 1 and 2 only.
3. **RACE: In** the raw dataset that has been extracted, we have 8 types of races namely American Indian and Alaska Native recode, Black or African American, Some other race recode, Other Pacific Islander recode, White recode, Native Hawaiian recode, Hawaiian and Pacific Islander, Native Hawaiian. All these columns have been recoded to 0 and 1 where 0 means that a person is not from that race and 1 means that the person is from that race. These are already present as separate variables in the raw data.
4. **MARITAL STATUS**: The variable is “MAR”. The raw data contains information on 5 different types of marital status, with the following code indicating the following

1 represents individuals who are married,

2 represents individuals who are widowed,

3 represents individuals who are divorced,

4 represents individuals who are separated,

5 represents individuals who are never married or under 15 years old.

To simplify the data, we will create a new variable called a "dummy variable" for marital status. This means that we will use 1 to indicate that a person belongs to a certain marital status category and 0 to indicate that they do not belong to that category.

1. **ABILITY TO SPEAK ENGLISH**: This is our variable of interest. We want to understand the effect of this on our response variable i.e., income. The focus of our hypothesis will be based on this variable.

The raw data is represented as:

The variable "ABILITY TO SPEAK ENGLISH" is measured on a scale of 0 to 4, where each value represents a level of proficiency in the English language.

0 represents individuals who are less than 5 years old or speak only English.

1 represents those who speak English very well.

2 represents those who speak English well.

3 represents those who do not speak English well.

4 represents those who do not speak English at all.

We will create four separate variables, one for each response category, to recode these values into dummy variables (very well, well, not well, not at all). We will assign a value of 1 to the corresponding variable for each respondent if their response falls into that category, and a value of 0 otherwise. This will generate **five** new binary variables for statistical analysis.

To recode the "very well" category, for instance, we will create a new variable called "ENGLISH VERY WELL" and assign a value of 1 to it for all respondents who answered "1" (very well) on the original variable, and a value of 0 to all other respondents. This process will be repeated for the remaining three categories, yielding a total of four dummy variables.

1. **EDUCATIONAL ATTAINMENT:** The variable is SCHL. The raw data we extracted includes 25 different levels of education, ranging from no schooling to a Doctorate Degree. However, we will only be considering individuals who have completed 12th grade or higher. This is because we believe that at least this level of education is necessary to earn. Additionally, we plan to group certain educational levels together and create dummy variables for these groups. Specifically, we will group together individuals who have no diploma or a regular high school diploma as well as those who have some college but less than one year, or one or more years of college credit without a degree. We will also group together individuals who have a master’s degree or a professional degree beyond a bachelor’s degree. We believe these groups are similar in terms of their educational attainment levels. This will allow us to create fewer dummy variables while still capturing the important variations in educational attainment.

1. **OCCUPATION**: The variable is Standard Occupational Classification (SOC) codes for 2018 and later based on 2018 SOC codes, “SOCP”. This is a great variable for our research. To think intuitively, the level of English proficiency required to earn well is indirectly dependent on income. The amount of English required to carry out your work duties is different based on occupation. The SOCP system assigns each occupation a unique digit code based on a set of criteria such as the type of work performed, and the skills required. These codes are then used to classify occupations into broader categories or "major groups" with similar characteristics.

The data is represented in numerical digit codes of different occupations. The raw data on the website also has categories of alphabetical codes for ease of understanding. Therefore, we group all similar unique codes and create dummy variables based on their alphabetical codes.

For example:

1. All digits starting with 11 are named under “Managers”

2. All digits starting with 131 are named under “Business”

3. All digits starting with 132 are named under “Financial”

4. All digits starting with 15 are named under “IT”

5. All digits starting with 17 are named under “Engineering”

6. All digits starting with 19 are named under “Science”

7. All digits starting with 21 are named under “Coucelling”

8. All digits starting with 23 are named under “Legal”

9. All digits starting with 25 are named under “Education”

10. All digits starting with 27 are named under “Entertainment”

11. All digits starting with 29 are named under “Medical”

12. All digits starting with 31 are named under “HealthCare”

13. All digits starting with 33 are named under “PRT”

14. All digits starting with 35 are named under “EATERY”

15. All digits starting with 37 are named under “CLN”

16. All digits starting with 39 are named under “PRS”

17. All digits starting with 41 are named under “Sales”

18. All digits starting with 43 are named under “Official”

19. All digits starting with 45 are named under “FFF”

20. All digits starting with 474 are named under “CON”

21. All digits starting with 475 are named under “EXT”

22. All digits starting with 49 are named under “RPR”

23. All digits starting with 51 are named under “Production”

24. All digits starting with 53 are named under “TRN”

25. All digits starting with 55 are named under “MIL”

Since these alphabetical codes are easier to understand, we decided to group it under those respective names itself

Ex: for FIN, we can create “finance” as the variable and assign a value of 1 to it for all respondents who are under FIN on the original variable “SOCP”, and a value of 0 to all other respondents. This process will be repeated for the remaining 23 categories.

As per initial analysis, the category “MGR” is too broad for managers. For example: the “Lodging Manager” and “Chief executive legislators” are clubbed under the same category, it appears that different managerial roles are clubbed together and not professions. This might dilute our analysis hence we might drop it after we perform our experiment (we want to make sure we have strong reason to drop it out)

1. **STATE:** The State variable is used in our analysis so as to represent the Population and Sample that we choose.

The state variable is Place of work - State or foreign country recode (POWSP).

Our variable has unique numeric values for each state of the United States. The values range from 001 to 0056. The values 003,007 are missing hence we have a value more than 050. Apart from these, we have values for Puerto Rico, Mexico, Europe, East Asia etc. represented in unique numerical values. We are dropping this value and will only retain 50 states of the United States.

The 50 states will form our population. As for our sample, we will create a subset of the East coast of the United States from the 50 states. (Defined in 4.)

We will create a dummy variable for each of the 12 states in the east coast.

Example, for New York, we can create a variable New\_york, and assign a value of 1 to it for all respondents who answered New York as their place of work on the original variable, and a value of 0 to all other respondents. This process will be repeated for the remaining 13 categories, yielding a total of 14 dummy variables.

**6. Hypothesis: What do you expect the relationship between the explanatory and response variables to be? Why do you expect this? This is where you should cite any existing literature on the topic.**

**You will likely benefit by first making a long tentative list of topics you are interested in, and then determining if these topics are doable by searching for Census data (scratching off those that are not). Once you have a “short-list,” you should come see me and we can figure out together what might be the best project for your proposal. After you submit your proposal, I will return it with comments. You must address my comments and resubmit your proposal.**

**Literature Review:**

There is a robust set of literature that establishes the relationship between English language proficiency and economic assimilation, defined as earnings. Chiswick and Miller’s study, “Language in the Labor Market: The Immigrant Experience in Canada and the United States” (1990), found that “fluency in the dominant language has a large positive effect on earnings, independent of other personal characteristics and country of origin” (4). Cadena et al. (2015) came to this same conclusion as well and found age of immigration to be a determinant of English proficiency which minimizes earning deficits with native-born counterparts. Gill and Ahmad (2019) quantified the relationship, finding that “the results for English-language proficiency show that individuals with no proficiency earn about 26% less than individuals with native-level proficiency.” Generally, the literature indicates a positive correlation of English language proficiency and earnings, suggesting economic assimilation to be facilitated linguistically.

**Hypothesis:**

Based on this existing literature, we hypothesize that median income will be greater for those with higher levels of English proficiency.

**References:**

Cadena, B. C., Duncan, B., & Trejo, S. J. (2015). Chapter 22 - The Labor Market Integration and Impacts of US Immigrants. In B. R. Chiswick & P. W. Miller (Eds.), Handbook of the Economics of International Migration 1197–1259. [https://doi:10.1016/B978-0-444-53768-3.00022-9](about:blank)

Chiswick B. R., & Miller P. W. (1990). Language in the Labor Market: The Immigrant Experience in Canada and the United States, Working Paper 784, Economics Department, Queen's University.

Gill, F., & Ahmad, W. (2019). The Earnings Disadvantage of 21st Century Immigrants in the United States. The American Economist, 64(1), 31–44. <https://www.jstor.org/stable/26725792>