Logo

Description automatically generated

**INTERMEDIATE ANALYTICS**

**ALY 6015.80428.202235**

**Instructor: Roy Wada**

**FINAL PROJECT: PROPOSAL/DATASET SELECTION**

**Date: 04/24/2022**

**BY**

**GROUP 9**

**Kiran Kasu|Abhinav Jain|Kelvina Patheni|Tanusha Erpula**

**Introduction**

**Dataset:** National Longitudinal Survey of Youth 1997

**Description:** The NLSY97 Cohort is a longitudinal project that follows the lives of a sample of American youth born between 1980-1984; 8,984 respondents were ages 12-17 when first interviewed in 1997. This ongoing cohort has been surveyed 19 times and is now interviewed biennially. Data are now available from Round 1 (1997-98) to Round 19 (2019-2020), along with a COVID-19 supplement fielded in 2021.

**Sample Introduction:   
  
Official Title:** National Longitudinal Survey of Youth 1997 (NLSY97), part of the National Longitudinal Surveys (NLS) program.

**Age of cohort:** Born between 1980 and 1984. At the time of first interview, respondents' ages ranged from 12 to 18. The respondents were 34 to 40 at the time of their round 19 interviews (Round 19 is the most recent data release, fielded in 2019-2020).

**Number of respondents in survey:**8,984 individuals initially interviewed in round 1. More than 77 percent (6947) of the round 1 sample were interviewed in round 19. For an overview of sample sizes by round, see table "NLSY97 Sample Sizes by Subsample, Race/Ethnicity & Gender."

**Gender:**4,599 (51%) males and 4,385 (49%) females in initial survey.

**Race/ Ethnicity (in initial survey):**

* Non-black/non-Hispanic: 4,665 (51.9%)
* Black non-Hispanic: 2,335 (26%)
* Hispanic or Latino: 1,901 (21.2%)
* Mixed: 83 (0.9%)

**Sample Sizes by subgroups:**

Table

Description automatically generated

**Figure 1: NLSY97 Samples Sizes by Subsamples, Race\Ethnicity & Gender**

**Other Documentation:**   
  
The NLSY97 provides significant insight into employment and unemployment trends that other surveys do not track, such as initiation into the job market and skill sets to prepare for transition into other jobs, periods of unemployment–particularly repeated periods of unemployment–and academic or military experiences. In addition, because longitudinal surveys, such as the NLSY, track the same individuals over time, the NLSY gives a more in-depth and complete picture of the labor market and provides unique, important insight into the experiences of adolescents as they enter the job market for the first time.

The NLS bibliography database houses nearly 6,700 summaries of ongoing and completed research projects, including published journal articles, government reports, policy analyses, monographs, working papers, conference presentations, doctoral dissertations, and master’s theses. Team members of the NLS Program routinely attend professional meetings of economists, demographers, social policy analysts, sociologists, and developmental psychologists, promoting the use of these data through presentations and poster sessions. In the past four years, the NLSY management team members have successfully offered four variations on courses at the University of Chicago’s Harris School of Public Policy Studies that principally feature NLS data.

Because of the success of the data collection - high response rates and solid data quality, combined with the many related topic areas well covered within the questionnaire, key staff members have been consultants to other birth cohorts and presented survey information internationally, in particular in England, Australia, Korea, China, and Germany.

**ANALYSIS**

**What are the possible questions/ problems that we can solve through this dataset?**

1. No of ‘white’ people and ‘black or African American’ and ‘Something else’ in years 1997-2012.
2. No. of ‘African American or Black’ people born in year 2000?
3. Descriptive Statistics of both Numerical and Categorical values in the dataset.
4. No of children in household whose race is white.
5. Total wages of people who are highest in the number of ‘race’ in year 1999.

(First find which race has highest number and then evaluate for the total wages of that race in year 1999)

1. No of Hispanic whose marital status is ‘Never married, cohabiting’ from year 1997-2012.
2. Educational degree of black whose marital status is ‘divorced, cohabiting.
3. No of total group of people whose wages are between 50k to 70k in the age group between 15-25 born in year 1998?
4. Age of white people when having a first child whose employment status is ‘Employed’.
5. Total number of arrests happened in ‘Rural’ in year 2012 who are ‘Not in Labor force’

Task 2: identify important sub-groups (for example, by gender or by year) for sub-group analysis

* Groupby KEY\_SEX and YEAR
* Groupby KEY\_Race and Wage\_
* Groupby CV\_Sample\_Type and Age
* Groupby CV\_MARSTAT\_COLLAPSED and Year
* Groupby CV\_ESR\_COLLAPSED\_ and AGE

**Descriptive Statistics**

The term “descriptive statistics” refers to the analysis, summary, and presentation of findings related to a data set derived from a sample or entire population. Descriptive statistics comprises three main categories – Frequency Distribution, Measures of Central Tendency, and Measures of Variability.

Descriptive statistics allow for the ease of data visualization. It allows for data to be presented in a meaningful and understandable way, which allows for a simplified interpretation of the data set in question. However, raw data would be challenging to analyze, and trend and pattern determination may be challenging to perform. In addition, raw data makes it challenging to visualize what the data is showing.

Here below we have performed descriptive statistics to the chosen dataset.   
  
 **Created two tables of descriptive statistics**

#Created all sample

**Table 1: All Sample table**

Table

Description automatically generated

#One by Group Tables:

**Groupby KEY\_SEX and YEAR**

A picture containing circle

Description automatically generated

**Groupby KEY\_Race and Wage\_**

Table

Description automatically generated

**Groupby CV\_Sample\_Type and Age**

A picture containing table

Description automatically generated

**Groupby CV\_MARSTAT\_COLLAPSED and Year**

Text, table

Description automatically generated

**Groupby CV\_ESR\_COLLAPSED\_ and AGE**

Table

Description automatically generated

**We have performed few analysis and created charts comparing the variables in the dataset**

**Chart 1:** Bar Chart for Comparision of Race and Sample Type



**Figure 2: Comparision of Race and Sample type**

**Chart2:** for the comparison of Gender and Year



**Figure 3: Comparision of Gender and Year**

**Chart3:** Comparison of RACE and RACE\_ETHNICITY  
  


**Figure 4: Comparision of Race & Race\_Ethinicity**

**Chart4:** Comparison of AGE and Marital Status



**Figure 5: Comparision of Age & Marital Status**

**Predictor Variables:**   
  
Columns that contain measurements made on each case are referred to as variables. When we’re trying to predict the value of one variable based on its relationships with the others, we use terms to distinguish between the variable we want to predict and the ones we are using to make the predictions. Statisticians call the variable we’re trying to predict the dependent variable, while the variables we used to make these predictions are the independent variables.

In data science, you might be more likely to hear the term outcome or response variable for the dependent variable and predictor variables or features for the independent variables.

**Dummy Variables:** **Dummy variable** in R programming is a type of variable that represents a characteristic of an experiment. A dummy variable is either 1 or 0 and 1 can be represented as either True or False and 0 can be represented as False or True depending upon the user. This variable is used to categorize the characteristic of an observation.

We are taking KEY\_RACE, KEY\_SEX,KEY\_ETHINICTY,KEY\_RACE\_ETHINICITY as a predictor variables and converting into a dummy variables which would be useful for us to implement the regression model to find the best accuracy.

**Analytical Methods**

The language is built specifically for statistical analysis and data mining. R analytics is not just used to analyze data, but also to create software and applications that can reliably perform statistical analysis. In addition to the standard statistical tools, R includes a graphical interface. As such, it can be used in a wide range of analytical modeling including classical statistical tests, lineal/non-lineal[modeling](https://www.sisense.com/glossary/data-modeling/), data clustering, time-series analysis, and more.

**Regression:** Regression analysis is a group of statistical processes used in R programming and statistics to determine the relationship between dataset variables. Regression analysis is used to determine the relationship between the dependent and independent variables of the dataset. Regression analysis helps understand how dependent variables change when one of the independent variables changes and other independent variables are kept constant. This helps build a regression model and, further, helps forecast the values regarding a change in one of the independent variables.

**Correlation:** Correlation is a bivariate analysis that measures the strength of association between two variables and the direction of the relationship.  In terms of the strength of relationship, the value of the correlation coefficient varies between +1 and -1.  A value of ± 1 indicates a perfect degree of association between the two variables.  As the correlation coefficient value goes towards 0, the relationship between the two variables will be weaker.  The direction of the relationship is indicated by the sign of the coefficient; a + sign indicates a positive relationship and a – sign indicates a negative relationship. Usually, in statistics, we measure four types of correlations: Pearson Correlation, Kendall rank correlation, Spearman correlation, and the Point-Biserial correlation.

**Conclusion**

In this assignment we have chosen a dataset from a list of datasets and have worked together as a group and created a report describing the data, understanding the dataset, and analyzing the possible questions that we can answer through the data. We have performed the descriptive statistics for the dataset and identified any predictor variables identified from the original data. And finally, we discussed the analytical methods or plans that are mentioned in the course to work on data. We would in future work on the dataset and analyze it more and perform different analysis with the help of the analytical methods mentioned in the report.

**References**

* National longitudinal survey of youth - 1997 | NORC.org*. (n.d.). NORC at the University of Chicago | Research You Can Trust | NORC.org.*<https://www.norc.org/Research/Projects/Pages/national-longitudinal-survey-of-youth-1997.aspx>
* National longitudinal survey of youth | 1997*. (n.d.). National Longitudinal Surveys | A Program of the U.S. Bureau of Labor Statistics*.  
  <https://www.nlsinfo.org/content/cohorts/nlsy97>
* LiveBook · Manning. (n.d.). liveBook · Manning.   
  <https://livebook.manning.com/concept/r/predictor-variable>
* Dummy variables in R programming. (2020, August 5). GeeksforGeeks.   
  <https://www.geeksforgeeks.org/dummy-variables-in-r-programming/?ref=gcse>
* Group by one or more variables — group\_by. (n.d.). A Grammar of Data Manipulation dplyr.   
  <https://dplyr.tidyverse.org/reference/group_by.html>
* Subset rows using column values — filter. (n.d.). A Grammar of Data Manipulation dplyr.   
  <https://dplyr.tidyverse.org/reference/filter.html>