

**Predicting Annual Wages of Individual based on Public Use Microdata Samples**

**Project Proposal**

**March,14 2018**

**MIS 749 Business Analytics**

*Team Mean Squares*

**Amit Gupte**

**Monish Thakore**

**Mayuri Kudale**

**Kanchan Pathak**

**Team Information**

**Team Name:**

Team Mean Squares

**Team Member Names:**

* Amit Gupte
* Mayuri Kudale
* Kanchan Pathak
* Monish Thakore

**Data and Problem Explained**

**Problem Statement:**

To predict the annual wages of individual based on various independent variables available in Public Use Microdata Samples provided by Census Bureau. The Model built to do this will help to identify the predictors that drive Wages of Individuals in California.

In future this model could be used for Predicting salaries of individuals based education levels, Field of Degree and other factors on Hiring Websites like Glassdoor and LinkedIn.

**Problem Type:** Regression Problem

**Data Sources:**

* American Community Survey Data Published by Census Board
* Link: https://www.census.gov/programs-surveys/acs/data/pums.html

**Data Preparation:**

* **Response Variable**: Annual Wages (In Dataset as WAGP)
* **Number of Independent Variables/Columns** :**200** Independent Variables in Dataset but a lot of them will be filtered out based on p-value analysis and correlation values in later stages. We would need to select a set of relevant independent variable using Predictor: Subset Selection Techniques and trimming the Irrelevant variables out.
* **Number of Rows in Data:** 376,036 (Dataset for California Population)
* **Types of Predictors:** The Dataset is a mix of quantitative and categorical variables. Please find a few examples below

**Note:** Detailed Data Dictionary for Dataset Attached in email.

* + CIT: Citizenship status (Categorical)
  + CITWP: Year of naturalization (Quantitative)
    - bbbb. Not eligible - not naturalized
    - 1928 .1928 or earlier (Bottom-coded)
    - 1929 .1929 - 1933
    - 1934 .1934 - 1939
    - 1940 .1940 - 1942
    - 1943 .1943 - 1944
    - 1945 .1945
    - 1946 .1946 – 1947
  + SCHL: Educational attainment (Categorical)
    - bb. N/A (less than 3 years old)
    - 01. No schooling completed
    - 02. Nursery school, preschool
    - 03. Kindergarten
    - 04. Grade 1
    - 05. Grade 2
    - 06. Grade 3
    - 07. Grade 4
    - 08. Grade 5
    - 15 .12th grade - no diploma
    - 16. Regular high school diploma
    - 17. GED or alternative credential
    - 18. Some college, but less than 1 year
    - 19 .1 or more years of college credit, no degree
    - 20. Associate degree
    - 21. Bachelor’s degree
    - 22. Master’s degree
    - 23. Professional degree beyond a bachelor's degree
    - 24. Doctorate degree etc
  + WKHP: Usual hours worked per week (Quantitative)
  + FOD2P: Recoded field of degree (Categorical)
    - bbbb. N/A (less than bachelor's degree)
    - 1100. GENERAL AGRICULTURE
    - 1101. AGRICULTURE PRODUCTION AND MANAGEMENT
    - 1102. AGRICULTURAL ECONOMICS
    - 1103. ANIMAL SCIENCES and SO ON