Amrita School of Engneering

Dept of CSE

Foundations of Datascience

*https://github.com/flaviovdf/evcomp2018/raw/master/data/nc-est2015-agesex-res.csv*

*https://github.com/flaviovdf/evcomp2018/raw/master/data/educ\_inc.csv*

*https://github.com/flaviovdf/evcomp2018/raw/master/data/nba\_salaries.csv*

***Q1. Trends in the US Population.***

*(practice working with**large tables of data).*

Dataset: "Annual Estimates of the Resident Population by Single Year of Age and Sex for the United States."

The percent change jumps from about 3% for the overall population to almost 30% for the people in their late sixties and early seventies. This stunning change contributes to what is known as the graying of America. By far the greatest absolute change was among those in the 64-67 age group in 2014. What could explain this large increase? We can explore this question by examining the years in which the relevant groups were born. Those who were in the 64-67 age group in 2010 were born in the years 1943 to 1946.The attack on Pearl Harbor was in late 1941, and by 1942 U.S. forces were heavily engaged in a massive war that ended in 1945.Those who were 64 to 67 years old in 2014 were born in the years 1947 to 1950, at the height of the post-WWII baby boom in the United States. The post-war jump in births is the major reason for the large changes observed.

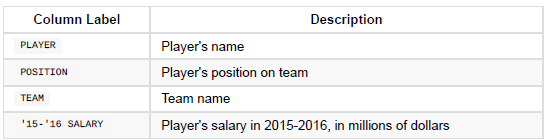
Requirement

1. Read the csv file as DataFrame.
2. Find size
3. Describe the data.
4. Re-label column years as “2014” and “2010”
5. Display first 10 rows.
6. Select 15th and 22nd rows.
7. Add columns change, percentage change.
8. Analyse the population changes (67-71) – show “graying of US” aspect
9. Analyse gender ratio of boys and girls among infants (age =0)
10. Find years where the ratio of females to males ranged from 2 to 4.
11. Plot the gender ratio versus age.

**Q2 Analyze National Basketball Association (NBA)**

"The NBA is the highest paying professional sports league in the world," reported CNN in

March 2016. The file nba\_salaries.csv contains the salaries of all National Basketball Association players in 2015-2016. Each row represents one player. The columns are:



The code for the positions is PG (Point Guard), SG (Shooting Guard), PF (Power Forward),

SF (Small Forward), and C (Center).

Data: [*https://www.statcrunch.com/app/index.php?dataid=1843341*](https://www.statcrunch.com/app/index.php?dataid=1843341) https://github.com/flaviovdf/evcomp2018/blob/master/data/nba\_salaries.csv

Requirement1: Simple Analysis

1. Read the csv file as DataFrame.
2. Find size
3. Sort on player
4. Sort on salary(increasing and decreasing)
5. Relabel column salary as “SALARY:
6. Display first 5 rows.
7. Select 3rd, 10th and 22nd rows.
8. Top 5 highest paid.
9. List those earned more than 10m USD.
10. Earnings of Stephen Curry
11. All Point Guards(PG) salary more than 15m USD.
12. Team “ Golden State Warriors”

Requirement2: Classifying by One Variable

1. How much money did each team pay for its players' salaries? (We have to group the rows by TEAM and then sum the salaries of the groups.)
2. How many NBA players were there in each of the five positions? (Classify by POSITION , and count – use grouping).
3. 15.What was the average salary of the players at each of the five positions?(Group by POSITION and take the mean of the salaries).

================================================================

**Q3.Education and Income of Californian Adults**

The State of California's Open Data Portal is a rich source of information about the lives of

Californians. A dataset on educational attainment and personal income among Californians over the years 2008 to 2014. The data are derived from the U.S. Census Current Population Survey. For each year, the table records the Population Count of Californians in many different combinations of age, gender, educational attainment, and personal income.

*Data: educ\_inc.csv*

*https://github.com/flaviovdf/evcomp2018/blob/master/data/educ\_inc.csv*

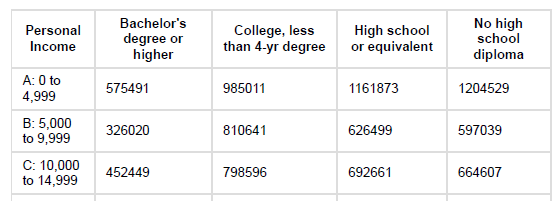
1. Analyse the data for the year 2014. (b) Consider adults (age >17)

* Each row of the table corresponds to a combination of age, gender, educational level, and income. There are 127 such combinations in all.
* As a first step focus on just one pair: educational attainment and personal income.

Requirement

1. Group the table by Educational Attainment and sum the Population Count in each category.
2. Analyse the percentage distribution of educational attainment among adult Californians.
3. Using pivot , get a contingency table (a table of counts) of adult Californians cross-classified by Educational Attainment and Personal Income.

*(see the power of pivot over other cross-classification methods. Each column of counts is a distribution of personal income at a specific level of educational attainment.Converting the counts to percents allows us to compare the four distributions)*.



*(Note: Over 35% of those with Bachelor's deg or higher had incomes of 75K and over, Vs fewer than 10% of the people in the other education categories had that level of income.)*

1. Draw a bar chart to compare the personal income distributions of adult Californians who have no high diploma with those who have completed a Bachelor's degree or higher. (The difference in the distributions would be striking. There is a clear positive association between educational attainment and personal income).

