December 2022

Capstone Project Report

Big data analyzing using spark and pig

Prepared by: U Data team

TABLE OF CONTENT

Introduction

As the percentage of employment rate continue to grow and due to the noticeable expanding in organizations size and variety. Nowadays, governments and consultancy investigate in the factors that may affect the employees such as the age, education level, marital status and salaries.

This sort of information will contribute to help decision makers in many aspects such as adjusting requirements for jobs, gender equality, focusing on the sectors in demand and adaptation in the education system. Not to mention the poor outcome if this information was incorrect.

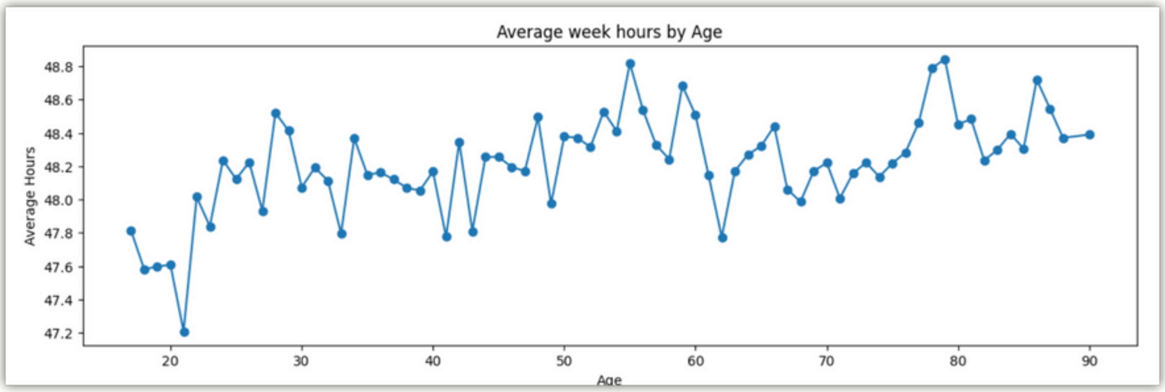
Our objective is to study a dataset that is not representative and highlight the affect of inaccurate information on the decisions and decision makers.

As one of the keys for the Vision which is increase the number of employees in Saudi Arabia, reduce the overall unemployment rate, and ensuring that the data is correct, representative, logical and meaningful, we used this dataset to visualize and understand how the false indication of data can affect the decision making.

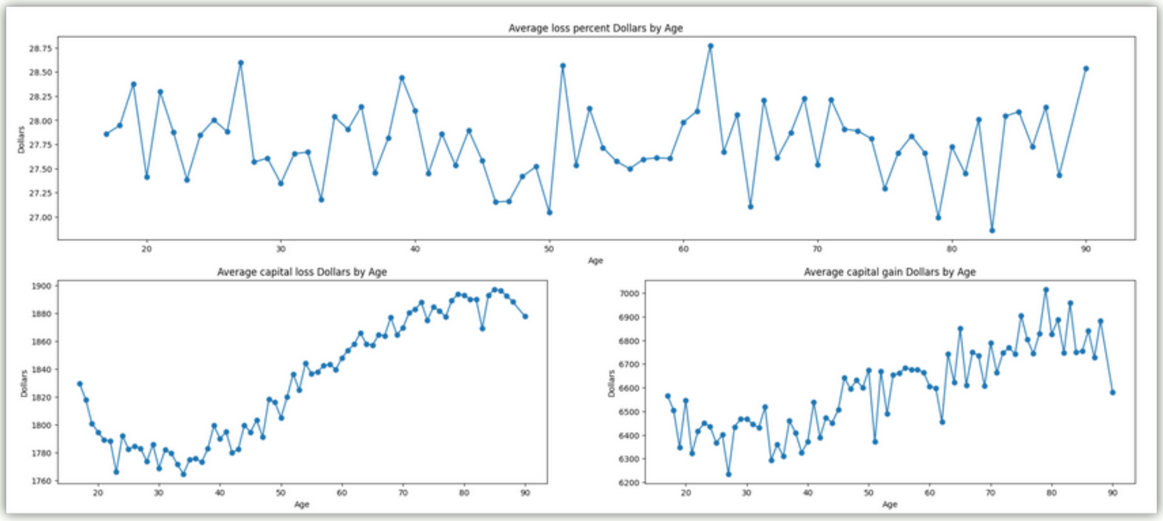
Data Review

The data we used is from UCI, the data set also known as "Census Income" dataset.
It contains 1M record and 15 column 'age','workclass','fnlwgt','education','education_num','marital_status','occupation','relationship','race','gender','capital_gain','capital_loss','hours_per_week','native_country','salary'

Data Exploring

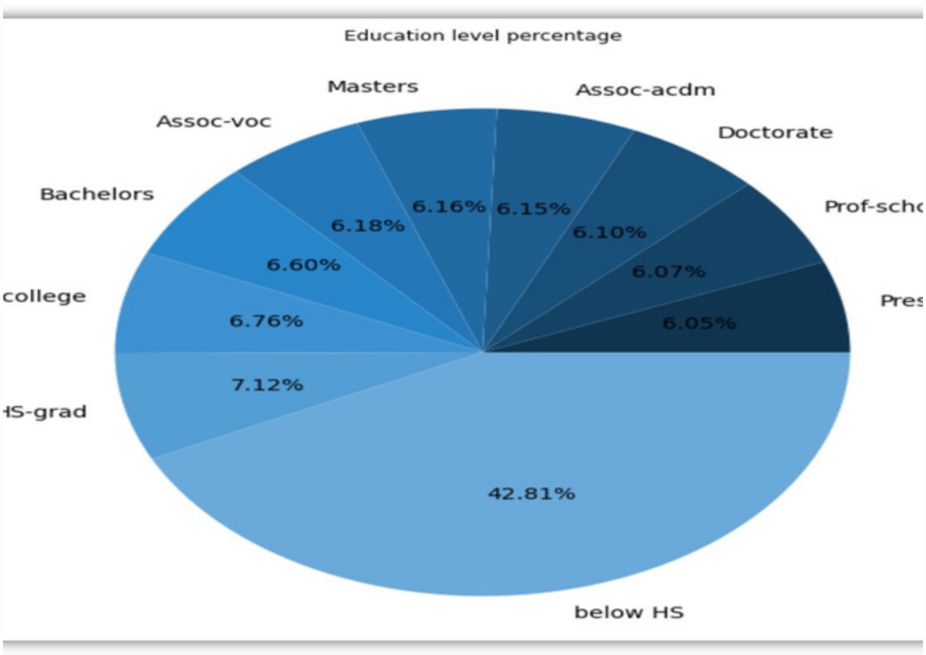


As we can see In this line chart, The Average weekly working hours remain almost the same. The average is between 47.2 and 48.8 hours per week

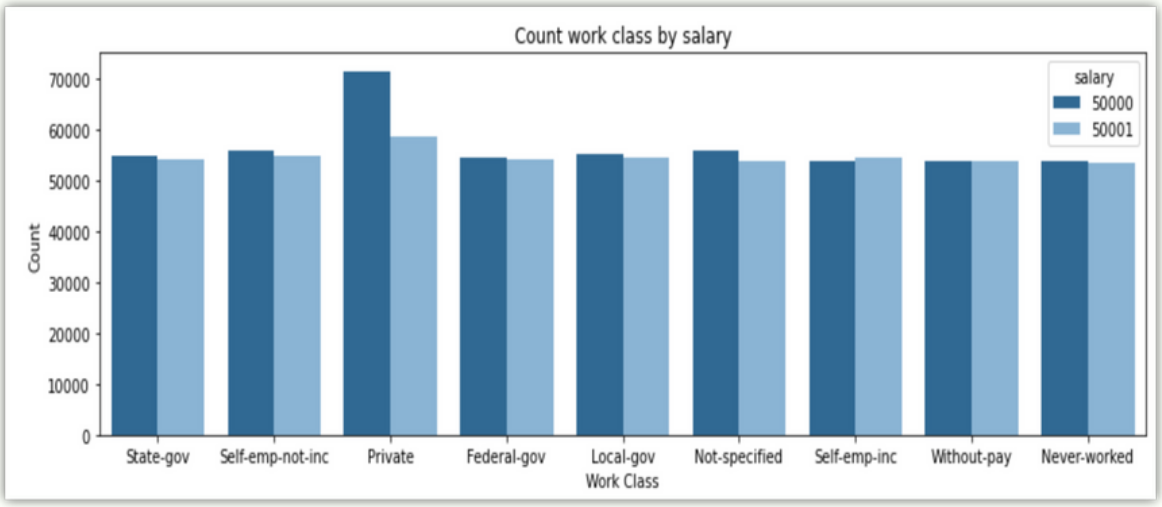


these charts show the Average capital loss percent, capital gain, and capital loss by age. We didn't notice a significant difference in the percentages, it ranged between 26.86% and 28.77% . Worth to notice that people who are 62,27, and 51 years have the highest percentages in average loss respectively. Both average capital gain and loss are increasing with age.

Data Exploring

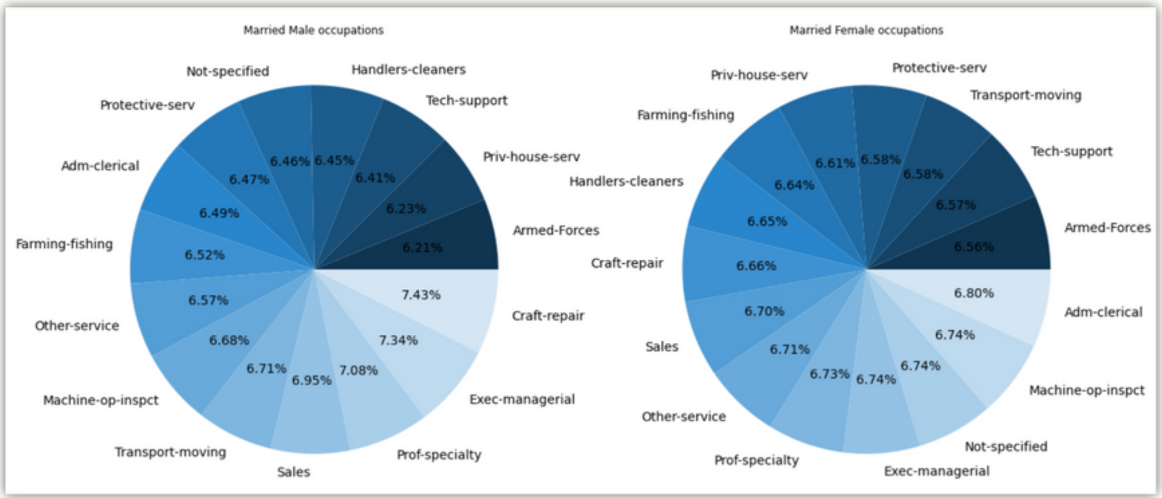


This pie chart shows education percentage for each education level. The education values are equal around 6% except for a slight increase in the High School, Some-College,and Bachelors are around 7%.

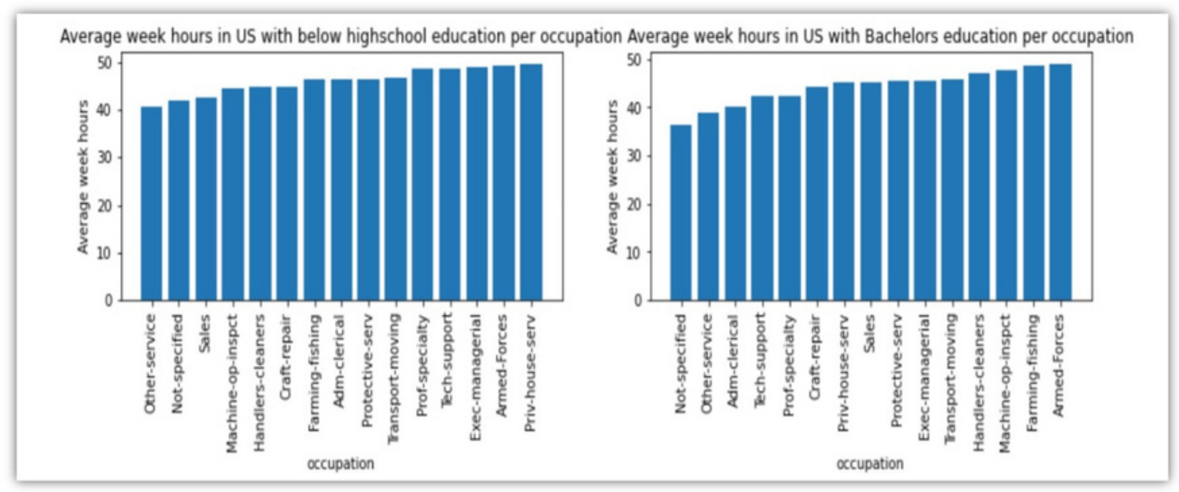


This bar chart shows the work class by salary. There's no significant difference in the working sectors except that in the Private sector there were around 70k people who have have salary less than or equal 50K.

Data Exploring

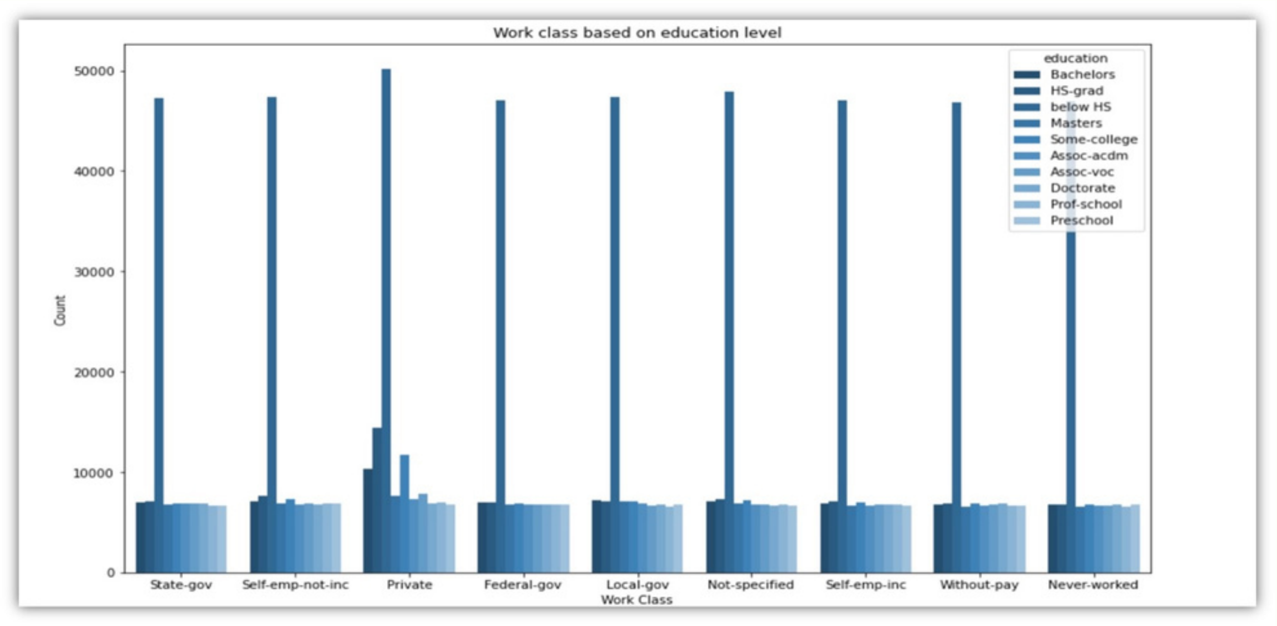


These two pie charts show the working sectors occupied by married male and female. There is no significant difference in the occupation count in each gender. The occupation equally diverse in each gender approximately.



These two bar charts show the Average working hours by occupation in US between those who have below HS and Bachelors educations. As we can see highest average working hours was in armed force. In addition, Bachelors degree holders have average weekly working hours in Handlers-cleaners higher than who don't have high school degree.

Data Exploring

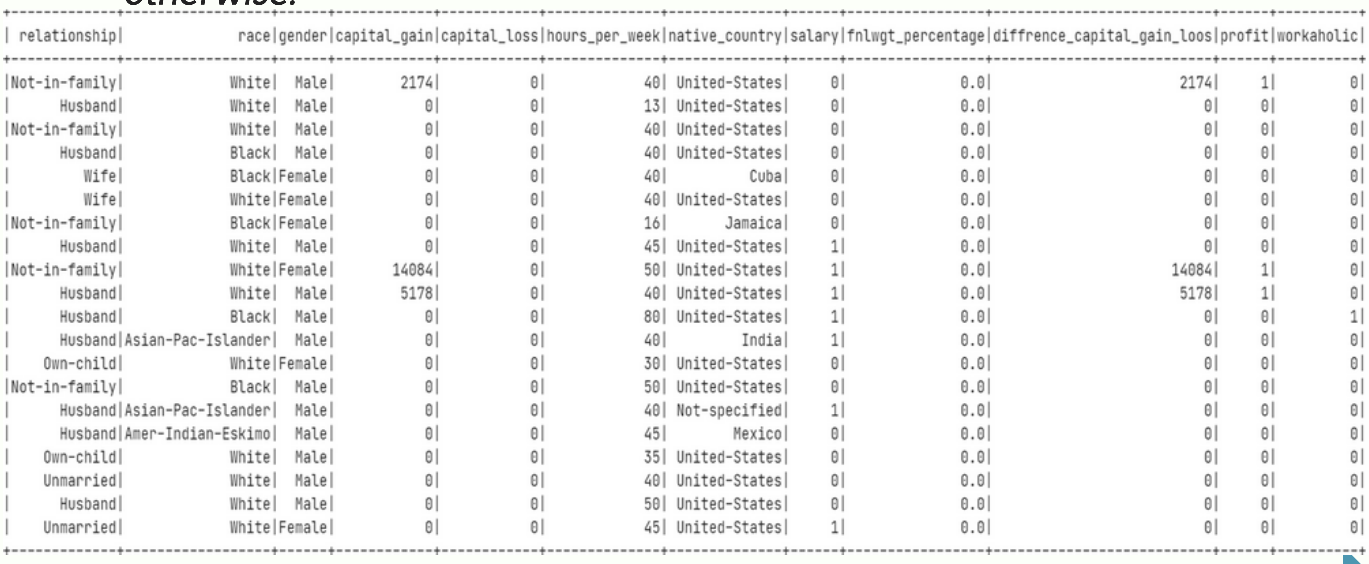


Employing all education levels is almost constant in all sectors. However, There is signigicant demand in the private sector for Bachelor, master and high-school graduates

Data pre-processing

The dataset is almost clean with no null values. However, there was inconsistent input .

- *Replacing (?) mark with not-specified in: (workclass-occupation-native country).*
- *Replacing range education level with one level.*
- *Replace the values for salary to be 0 if it was less that 50k, and 1 if it was more than 50k.*
- *Create a new column (Percentage of fnlwgt).*
- *Create (profit) column which have 3 values: (0) if the difference between gain and lose is greater than 0,(1)if the difference between lose and gain if equal 0, and (-1) if the difference less than 0 .*
- *Workholic column which is a boolean column that have (1) if the total working hours is equal or more than 70 and (0) otherwise.*



```
+-------------+------------------+------+------------+------------+--------------+--------------+------+-----------------+---------------------------+------+----------+
| relationship|              race|gender|capital_gain|capital_loss|hours_per_week|native_country|salary|fnlwgt_percentage|diffrence_capital_gain_loos|profit|workaholic|
+-------------+------------------+------+------------+------------+--------------+--------------+------+-----------------+---------------------------+------+----------+
|Not-in-family|             White|  Male|        2174|           0|            40| United-States|     0|              0.0|                       2174|     1|         0|
|      Husband|             White|  Male|           0|           0|            13| United-States|     0|              0.0|                          0|     0|         0|
|Not-in-family|             White|  Male|           0|           0|            40| United-States|     0|              0.0|                          0|     0|         0|
|      Husband|             Black|  Male|           0|           0|            40| United-States|     0|              0.0|                          0|     0|         0|
|         Wife|             Black|Female|           0|           0|            40|          Cuba|     0|              0.0|                          0|     0|         0|
|         Wife|             White|Female|           0|           0|            40| United-States|     0|              0.0|                          0|     0|         0|
|Not-in-family|             Black|Female|           0|           0|            16|       Jamaica|     0|              0.0|                          0|     0|         0|
|      Husband|             White|  Male|           0|           0|            45| United-States|     1|              0.0|                          0|     0|         0|
|Not-in-family|             White|Female|       14084|           0|            50| United-States|     1|              0.0|                      14084|     1|         0|
|      Husband|             White|  Male|        5178|           0|            40| United-States|     1|              0.0|                       5178|     1|         0|
|      Husband|             Black|  Male|           0|           0|            80| United-States|     1|              0.0|                          0|     0|         1|
|      Husband|Asian-Pac-Islander|  Male|           0|           0|            40|         India|     1|              0.0|                          0|     0|         0|
|    Own-child|             White|Female|           0|           0|            30| United-States|     0|              0.0|                          0|     0|         0|
|Not-in-family|             Black|  Male|           0|           0|            50| United-States|     0|              0.0|                          0|     0|         0|
|      Husband|Asian-Pac-Islander|  Male|           0|           0|            40| Not-specified|     1|              0.0|                          0|     0|         0|
|      Husband|Amer-Indian-Eskimo|  Male|           0|           0|            45|        Mexico|     0|              0.0|                          0|     0|         0|
|    Own-child|             White|  Male|           0|           0|            35| United-States|     0|              0.0|                          0|     0|         0|
|    Unmarried|             White|  Male|           0|           0|            40| United-States|     0|              0.0|                          0|     0|         0|
|      Husband|             White|  Male|           0|           0|            50| United-States|     0|              0.0|                          0|     0|         0|
|    Unmarried|             White|Female|           0|           0|            45| United-States|     1|              0.0|                          0|     0|         0|
+-------------+------------------+------+------------+------------+--------------+--------------+------+-----------------+---------------------------+------+----------+
```

Building model

We build 4 models of each classification and regression models. We used ***salary*** *as our target and the results were: 51% for Gradient Boosted Trees classifier and 50% for decision tree classifier. In the regression model we used the* ***capital gain*** *as our target and the results were 79% in Gradient Boosted Trees regressor and 78% in Decision Tree Regressor.*

Pig Questions

We have picked 5 top questions that we believe it can be applicable in case of real- data values :

- ***Count of each gender?***
- ***The numbers of instance in each country?***
- **How many races are there?**
- ***Count workaholic who works more than or equal 70 hours a week and they are between 20 - 40 Years old ?***
- **Count work-class and gender?**

Insights we have come to:

- *There are 1M individuals. The number of males are 505923 representing 50.69%. While the number of females 494077 which represent 49.4%. The dataset shows almost equal distribution of males and females .*
- *There are 42 countries. The most occuring country is the U.S 51942 5.19%. The least occurring is Laos 22788. The rest of the countries occured around 23,000 around 2.3%.*
- *The count of white race is 221239, representing 22.21% which is the highest.The count of Other races is 193367, representing 19.33%, which is the lowest.The count of Amer-Indian-Eskimo is 194177, which form 19.41% of the total number . The count of Asian-Pac-Islander is 194364, which represent 19.43%. The count of Blacks are 196853, representing 19.68%. The dataset shows almost equal distribution between races.*
- *There is only 7.7% workaholics who works more than 70 hours a week .*
- *The private sector, local government and state government were the top three sectors that both Men and Women worked at. The minority of both genders work as self-employed.*

Conclusion

False indication have a big impact on the decision-making process, if the data was biased towards specific sample, and this sample is not representative, it will lead to bad decision.

1. Clone the method we approached on Saudi data that is approved from The Ministry Of Human Resources And Social Development
2. Add more detailed columns for a better investigating and insights- driven