Project Executive Report

Census Data Analysis For Salary Classification [UVW College]

Spring 2021

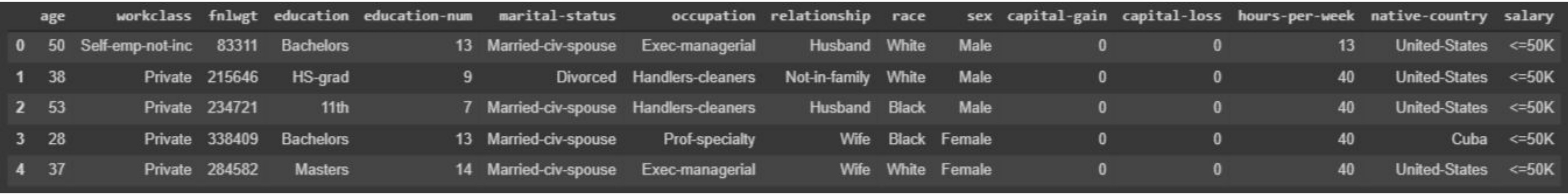
Group 12

Ajay Kannan
Aparokshith Rao
Raakesh Sureshkumar
Rithvik Chokkam
Sanjay Arivazhagan
Vivek Bellalacharvu Srinivasa Rao

Problem Definition

1. The UVW college requires XYZ company to analyze the Census dataset to get the best of influence of the various parameters either directly or indirectly over the prescribed boundary of $50,000
2. Process the provided United States Census data for further analysis.
3. Analyze the data using data aggregation and count iterators for getting an insight before proceeding with the visualizations
4. Generate insightful visualizations that provide information that is not evident in the tabular dataset. The visualizations must be of direct relation to the class of data based upon the two open salary ranges. The visualizations must be clear without occlusion or overlap with proper labelling and effective color selection

Dataset



	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	salary
0	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
1	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
2	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
3	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
4	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K

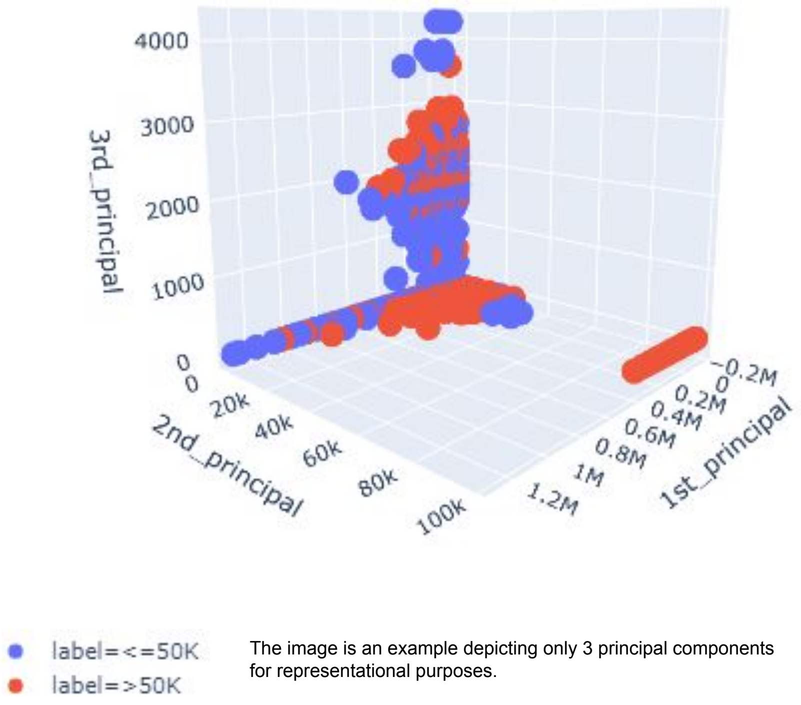
Description:

1. The main dataset contains 32561 records and a distribution of 12661 for the test dataset.
2. The grouping of the dataset is 34014 for the combined set of salaries less than $50K and 11208 for the combined set of salaries more than $50K.
3. The records contain 14 features in its entirety.
4. The used dataset contains random sampling of data constituting 50% each from classes of less and greater than $50K.

Cleaning and Preprocessing:

1. The noise with the ‘?’ was found and removed.
2. New data frames were created for the capital-gain and capital-loss by removing the 0s from the columns. There were 2712 records for the capital-gain and 1519 belonging to the capital-loss.
3. Due the sparse nature of the 2 columns, these columns are neglected for further analysis.
4. The final set is converted to a pandas dataframe for further analysis using Python.
5. Redundant columns were identified “Education” & “Education-num” which are categorical with 16 values.

Major Influencing Factors

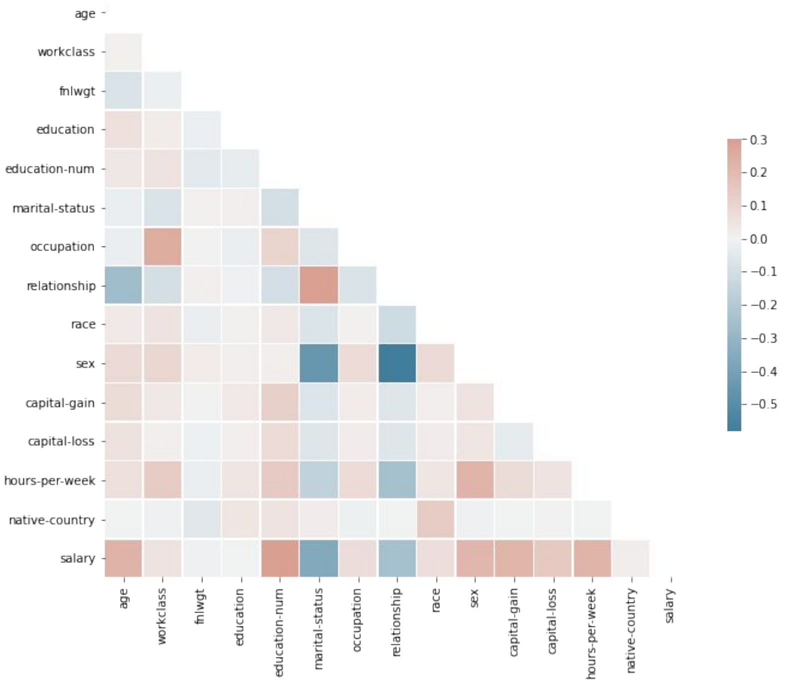


The image is an example depicting only 3 principal components for representational purposes.

A Principal Component Analysis was done with the 14 features mentioned above to get the most 9 significant factors under consideration:

- **Age (Nominal)** [16, 99]
- **Education (Categorical)** [' Bachelors', ' HS-grad', ' 11th', ' Masters', ' 9th', ' Some-college', ' Assoc-acdm', ' Assoc-voc', ' 7th-8th', ' Doctorate', ' Prof-school', ' 5th-6th', ' 10th', ' 1st-4th', ' Preschool', ' 12th']
- **Relationship (Categorical)** [' Married-civ-spouse' ' Divorced' , ‘Married-spouse-absent', ' Never-married' ' Separated' ' Married-AF-spouse' ' Widowed']
- **Occupation (Categorical)** [15 variants]
- **Race (Categorical)** [' White', ' Black', ' Asian-Pac-Islander', ' Amer-Indian-Eskimo', ' Other']
- **Sex (Categorical)** [Male, Female]
- **Hours per week (Nominal)**
- **Native country (Categorical)** [42 variants]
- **Final weight (Nominal)**

Multivariate Analysis



1. All the parameters were considered for this analysis to produce a heatmap matrix

2. The red boxes have higher correlation compared to the blue ones

3. The highest correlations with the salary (required trait) can be found in age, education-num, sex, capital-gain and hours-per-week

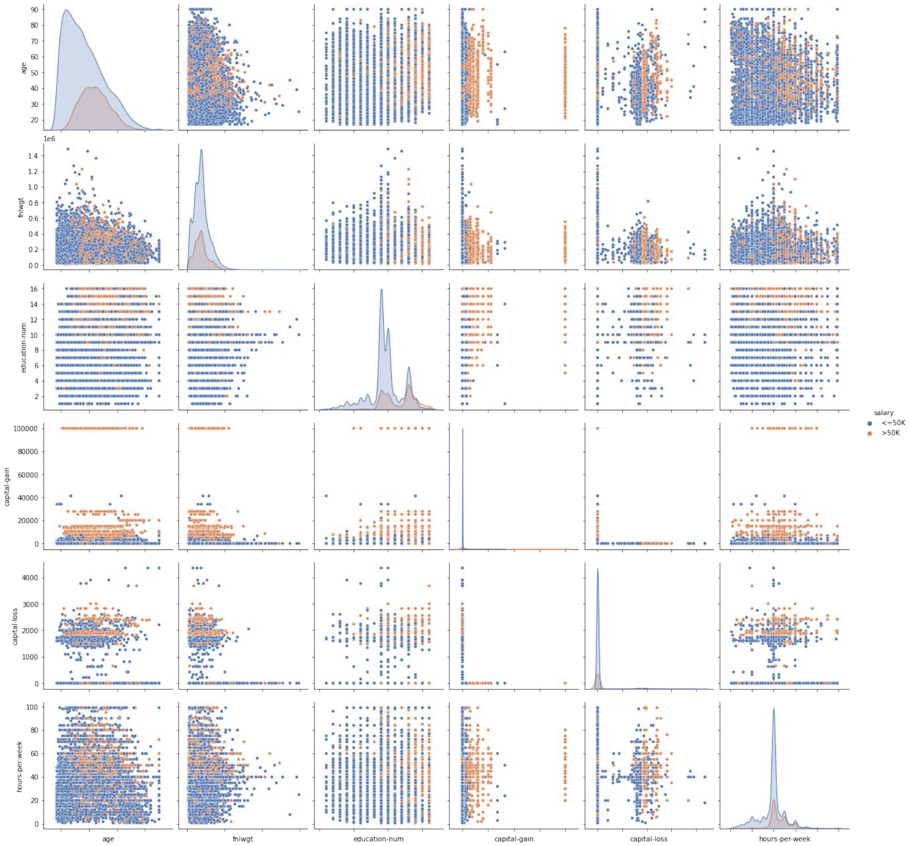
4. Apart from these indirectly, the variables relationship and marital status are also found to be correlated

Scatter matrix plot

1. This scatter plot was created to provide an eye-opener on new domains to explore based on the aggregation of the variables age, education-num, capital-gain and hours-per-week

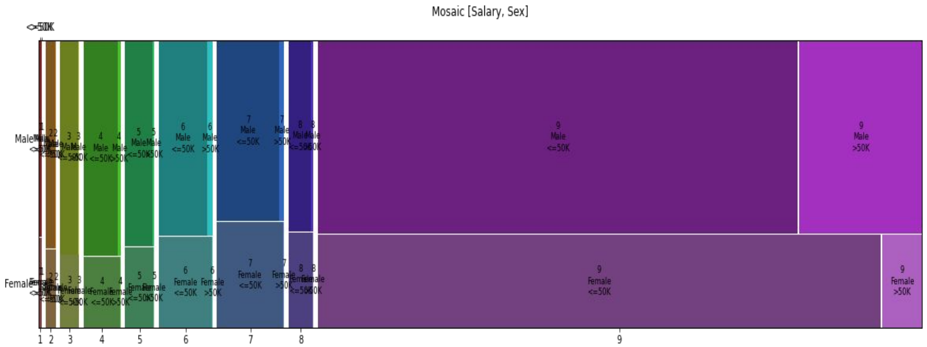
2. For histogram, the segregation based on the salary range gives us clear patterns and effective troughs can be found in the variables education-num and hours-per-week (histogram).

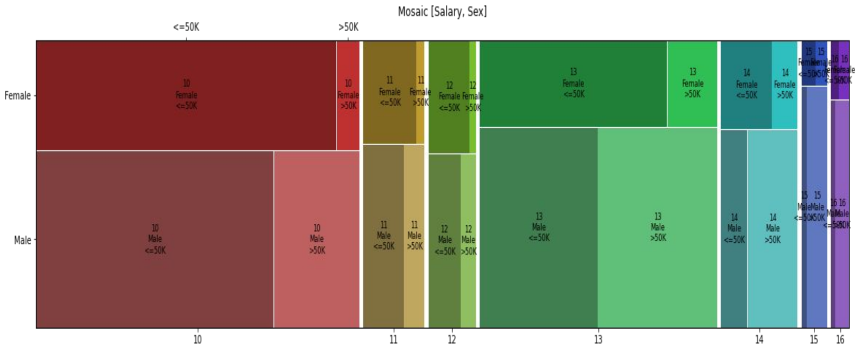
3. According to the scatter plot, the capital-gain seems to be of significance too



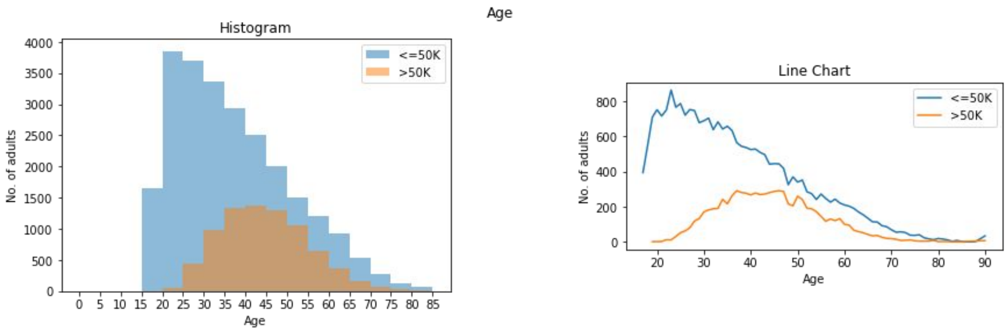
Sex

1. The mosaic plot was plotted to analyze the variance of the salary range count based on the fact if the adult is a Male / Female and education number
2. The plot shows the ratio of number of people earning more than 50k to less than 50k is only more in cases of masters and doctorate degree holders. For all other education levels, the ratio is lesser i.e more people are earning lesser than 50k compared to more. So, from the data, having a degree of master's or above increases the chances of earning over $50k.





Age



1. The insights made here goes beyond the dataset by implicitly categorizing it into (17-25), (25, 60), and 60 & beyond as retired.

2. It is observed that people in the age group of 17-25 are most likely to receive a salary less than or equal to 50 thousand. It is also observed that there is a relatively higher chance that individuals in the age group 25 to 65 are more likely to receive a salary greater than 50 thousand.

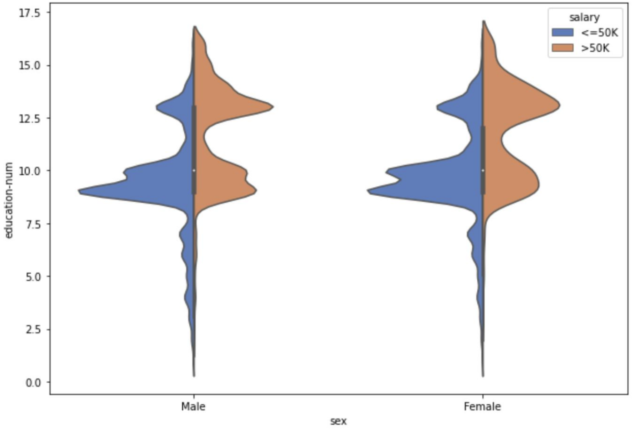
3. There are fewer people above the age of 65 with a salary. This may be due to the fact that many people in skilled occupations retire around 65 years

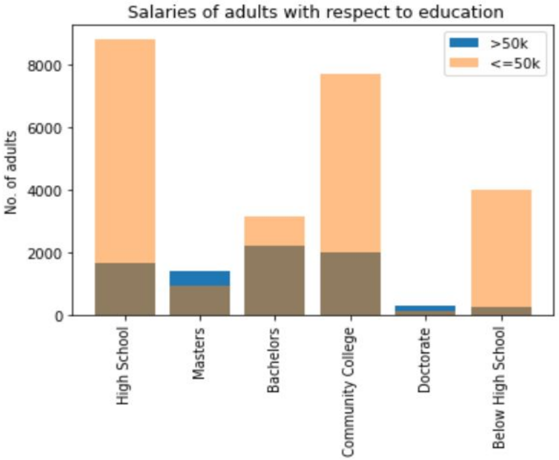
Education

1. The violin chart and an occluded bar chart has been included for analyzing the education.

2. The major inferences made were regarding the violin chart displaying bulges at the education numbers of [10, 12] and [14, 15]. This shows that majority of the people included in the dataset with a salary < $50k can be found doing some form of college education. The other bulge can confirm the increased distribution of adults doing a Masters or PhD earning greater than $50k. However, the pattern seems to remain the same for both Men and Women

3. For all other education levels, the ratio is lesser i.e more people are earning lesser than 50k compared to more. So, from the data, having a degree of master's or above increases the chances of earning over $50k

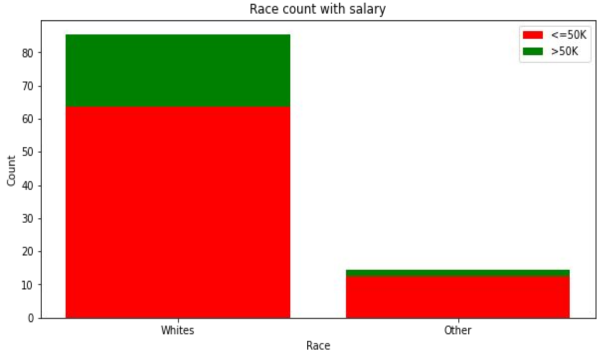


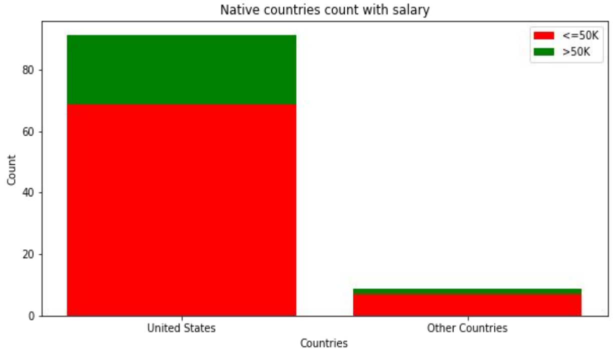


Race and Native Country

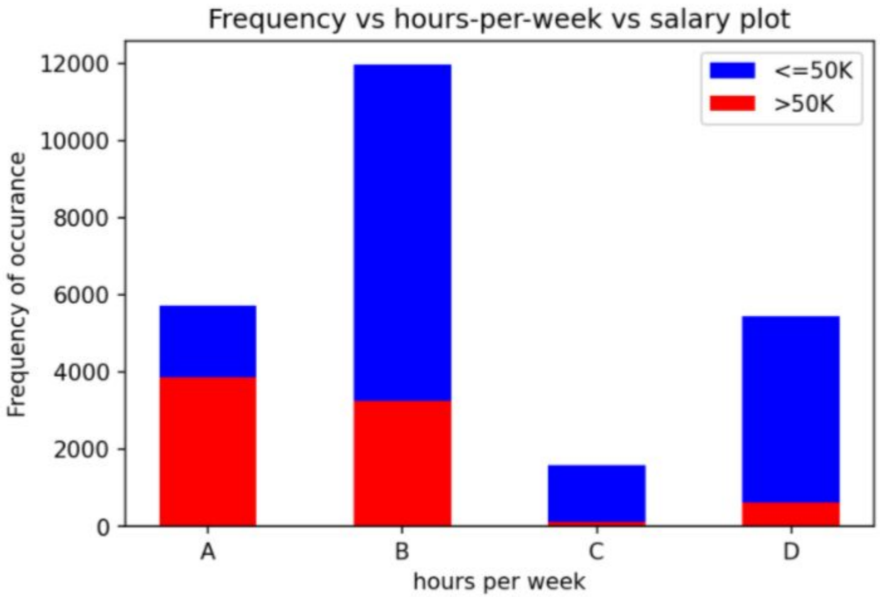
1. There is an aggregation of the other categories within the race column and the native-country column to analyze this particular data to establish the unordinary skew in the data.

2. Bar plots have been an efficient analysis to display the skewness with the United States dataset for salary distribution based on race and country of origin. There seems to be a high skew with the ratio maintained for each category of [Whites / Other] & [United States / Other countries]





Hours per week



A = Greater than 40 hours (Over-time)

B = Equal to 40 hours (Full-time)

C = Between 20 and 40 hours (Part-time)

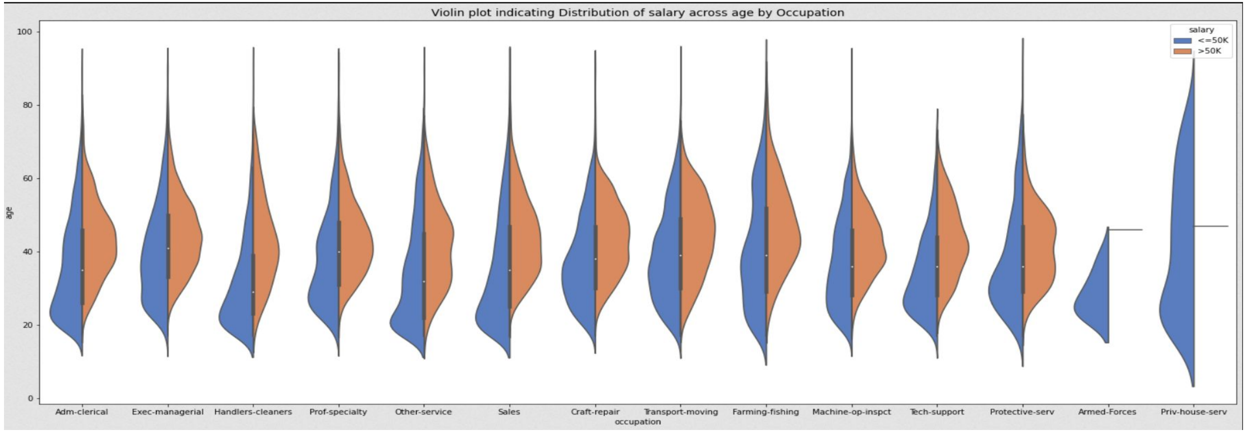
D = Less than 20 hours (Contract)

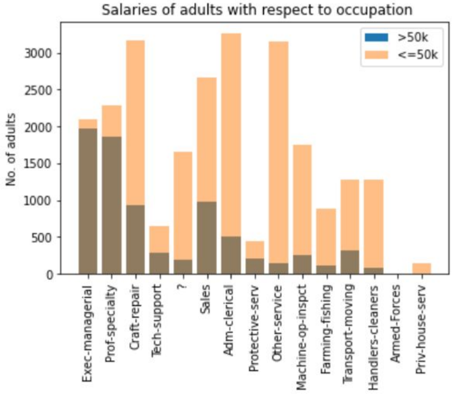
1. In one of the analyses, the data is segregated into 4 categories as A (Over-time), B (Full-time), C (Part-time), D (Contract).

2. The inference includes a huge distribution of data concentrated towards the Full-time jobs. Also, within the full-time job holders, the adults earning less than $50k seems to be higher in number

3. Among the distribution, the people doing part-time seem to have the least amount demographic distribution

Occupation



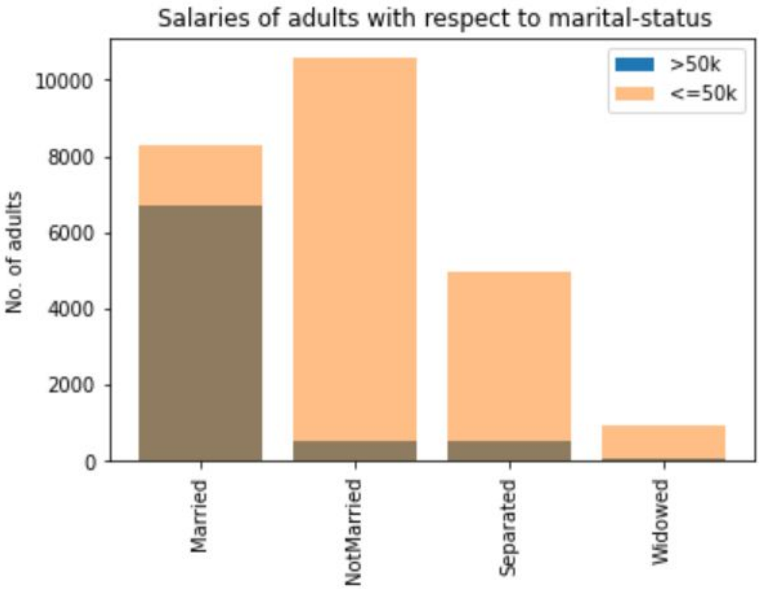


1. In the violin plot, the armed forces group has the least age distribution. As expected, armed forces have lots of employees in the age group 21-30. None of the armed forces personnel earn more than 50 thousand. This might be a job characteristic. While the age distribution in private house services is significantly large, experience does not translate to higher salaries.

2. Apart from armed forces and private house services, other occupations seem to value experience and it can be observed with the increase in people receiving salaries greater than $50k

3. The bar chart shows people with occupation “Exec-managerial” have the highest ratio of earning greater than $50k to earning less than $50k. We can conclude that holding a position in the “Exec-managerial” category increases the chances of earning greater than $50k compared to any other type of occupation

Relation



It is shown that in the data, people who are married have the highest ratio of people earning greater than $50k compared to those earning lesser than $50k. So, we can infer from this that being married increases the chances of earning more than $50k

Summary and Key Insights

1. Adults with full time disrupt the demography of people with about 75% people earning less than $50k representing about 25% of the entire population.

2. Adults who have done a higher education are found to posses jobs with salaries greater than $50k though the overall sample count is less

3. The male adults have a positive trend towards the salary range being greater than $50k. However, the trend in women starts the same as in men (with a higher ratio earning less than $50k) and remains constant as the age increases with diminishing count.

4. A major skew can be found in the data for race and country of origin with the inclination towards the whites and the adults from the United States.

Future Scope

1. A machine learning model can be built to analyze the dataset for prediction of the salary above or below the $50,000 mark. Pipelines can be built using the pre-existing models like the Naive Bayes, Decision Tree, Ensemble, Logistic Regression and the kind. This can be done by choosing the 5 best parameters influencing the trend by performing a Principal Component Analysis. The dataset has already been divided into train and test which is in the ratio 4:1.

2. Additional tools and technology like Tableau, Power BI and d3.js can be used for following advantages:
   a. Creation of interactive dashboards for better visualizations and getting useful inferences
   b. Using aggregation and merge functions to analyze the data in depth and produce better insights.
   c. Collect multiple tables of data that has different perspectives on the salary deviations