Priyam Dalwadi(1001994810)
Kelsey Nguyen (1001836837)
DASC 5300-001

Project I Report

Overall Status

Looking at the data. The shape of the data was 32561 x 15 originally. This data included missing values that sum up to 2399x15 . The data types of each column in the data is shown with the figure: Here , memory consumption of the data is around 3.7+ mb. The target or dependent variable is salary, the rest are independent variables.

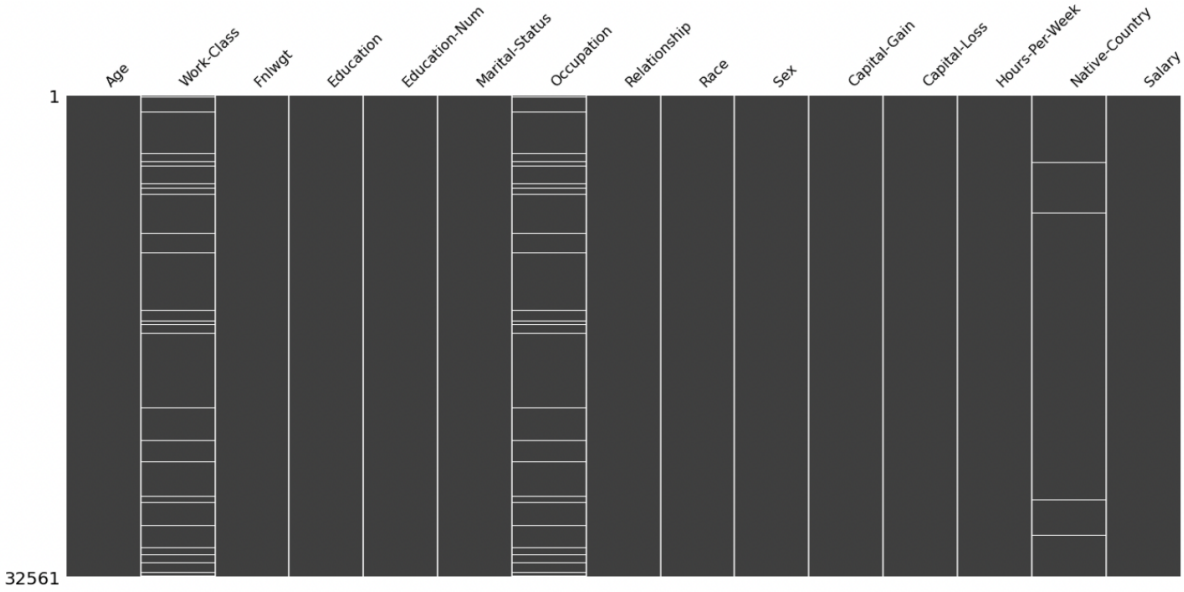
```
data.info()
data.shape
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32561 entries, 0 to 32560
Data columns (total 15 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Age             32561 non-null  int64
 1   Work-Class      30725 non-null  object
 2   Fnlwgt          32561 non-null  int64
 3   Education       32561 non-null  object
 4   Education-Num   32561 non-null  int64
 5   Marital-Status  32561 non-null  object
 6   Occupation      30718 non-null  object
 7   Relationship    32561 non-null  object
 8   Race            32561 non-null  object
 9   Sex             32561 non-null  object
 10  Capital-Gain    32561 non-null  int64
 11  Capital-Loss    32561 non-null  int64
 12  Hours-Per-Week  32561 non-null  int64
 13  Native-Country  31978 non-null  object
 14  Salary          32561 non-null  object
dtypes: int64(6), object(9)
memory usage: 3.7+ MB

(32561, 15)
```

We approached the project by first importing the data and reading the csv file using **data.read_csv()** function. In this we passed a list of column names for our convenience. Next, we obtained a random sample which was implemented by using **data.sample()** function. The seed value we passed is 261197 to get a random sample and we copied this dataset into **random.csv file** and then knew we had to clean the data by finding the “?” values. While trying to find “ ?” values we passed the arguments in syntax **pandas.read_csv()** that **na_values = “ ?”**. The argument which we passed in the function converted “ ?” values into NaN values. Using the **.isnull().sum()** function made it easier to isolate the null values to be able to extract and place them into a different file. After finding the null values we saved them into the file named **“missingvalues.csv”** and then removed them from the data set. The total number of rows of missing values that was eliminated was 2399. From further analysis we found that 1836 missing values came from the column work-class, 1843 missing values came from the occupation column and 583 missing value rows came from the native-country column. The total number of missing values contributes to 7 percent of the data. Another way to deal with the missing values is to reassign the values on each of the columns so that there is a greater sample size to evaluate with. After removing the missing values, the total dimensions of the data are now 30,162 rows with 15 columns. Missing values can be visualized in the picture below by using **missingno** library. In the picture below horizontal white lines shows us missing values.

Priyam Dalwadi(1001994810)
Kelsey Nguyen (1001836837)
DASC 5300-001



Then, to be able to plot the data into histograms we had to convert the data set values of age into an integer and group them by the specific category. After that, we used the **data.replace()** function to reassign the integers with the age group string values (“<20”, “21-40”, “41-60”, “>60”). To create the histogram of age groups, we used the **seaborn** and **matplotlib** libraries to create the visualizations. For counting the number of people under each age group we used the **countplot** function in seaborn. We then grouped the count of people in the work class and age groups to be able to create a **pie chart.**

File Descriptions

- EDA-final.ipynb

This file contains all the commands and graphs and shows how we approached analyzing the census data. This file includes

- Missingvalues.csv

This file contains the rows of missing values of columns Work-class, Occupation and native country that we have extracted from the census data.

- RandomSample

We assigned 100 values to the seed variable and random state to our birth date that is 11261997 and we saved a sample data set in a csv file.

Division of Labor

We divided the analysis of the project between the both of us and developed the algorithms in 5 days. We then spent 2 weeks trying to develop the code for creating the visualizations. We then spent 1 week refining the visualizations to be able to analyze and interpret.

Priyam Dalwadi(1001994810)
Kelsey Nguyen (1001836837)
DASC 5300-001

Problems encountered and how you handled them

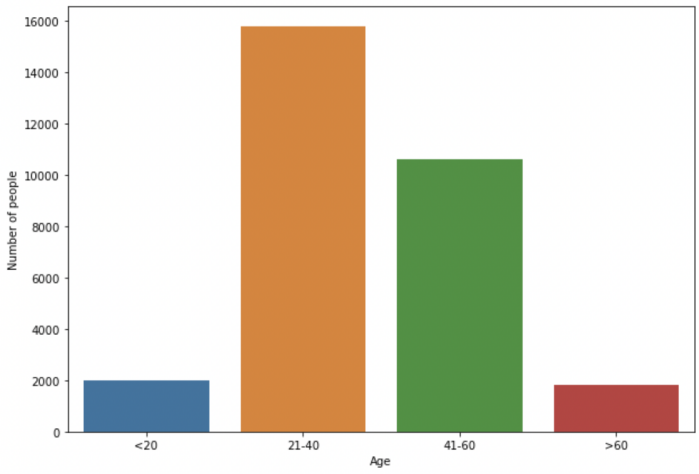
We directly read the data and started doing data exploratory analysis by finding NaN missing values. So we checked individually for the values in each column and the missing values were " ?" instead of Nan. So we converted those values by reading documentation.

Another problem we faced was the number of females and males. When we group the number of males and females it shows that there are half as many women in the data compared to men in education, profession, workclass.

We also had difficulty dividing the age of people into 4 groups. We ran a for loop to convert the range of ages into categorical variables. While we ran a for loop it throws an error of **'list' object has no attribute 'loc'**. To fix this error we used **data.astype()** function which has capability to convert any existing column to a categorical type.

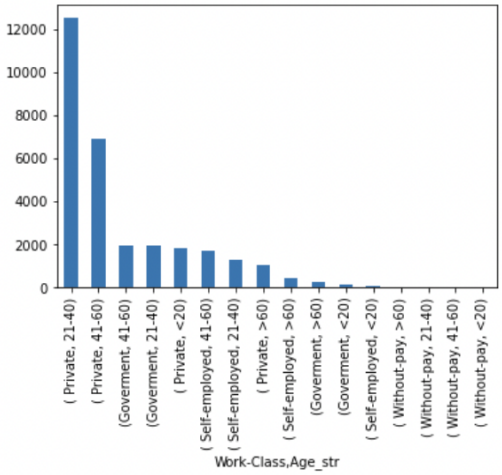
Analysis of results

Part 4a. Histogram of Age groups



From this histogram we can see that the age group with the highest count of workers are in the group with people aged 21-40. The group with the second highest count of workers is in the group with people aged 41-60. The age group with the third highest count of workers are in the group with people <21 and the group with the lowest count of workers are in the group with people who are >60. The results we have obtained from the histogram are intuitive. We should expect the groups with the highest count of workers to be with the people aged 21-40 and 41-60 because that is in range with the working age population. The people in the age group <21 could still be in grade school and may not be focused on working yet but are able to if they want. The people in the age group >60 may have retired from the workforce or just stopped working due to old age.

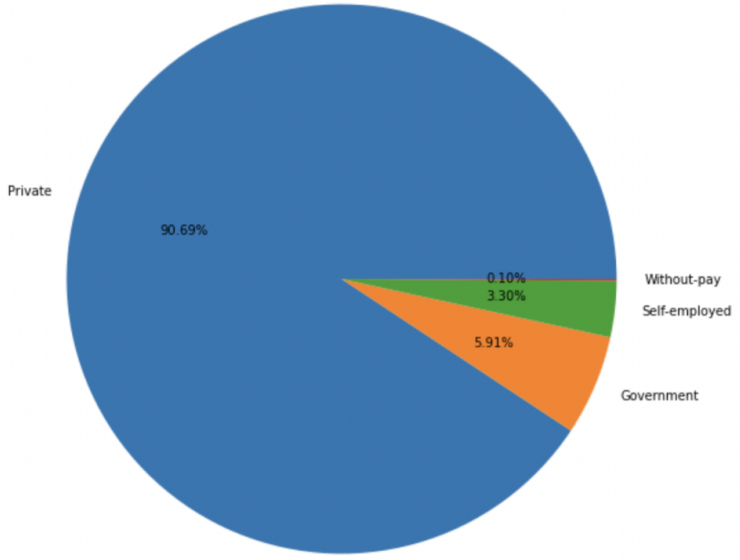
Part 4b. Number of people in professions split by age groups



Priyam Dalwadi(1001994810)
Kelsey Nguyen (1001836837)
DASC 5300-001

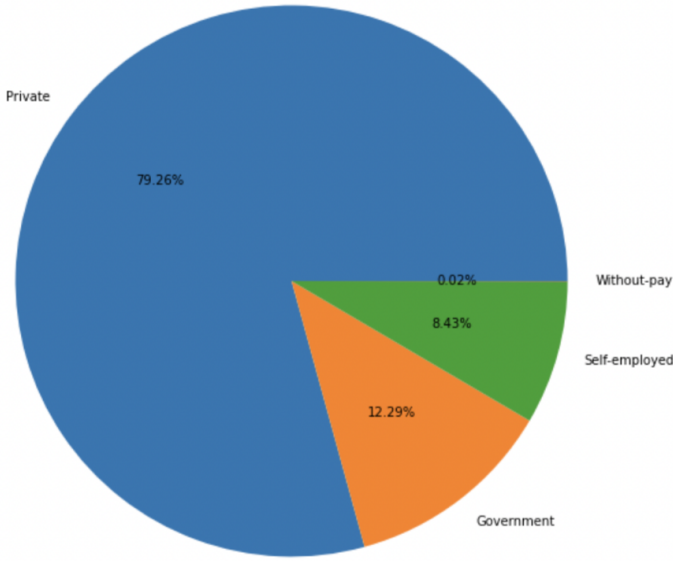
To first approach creating the pie charts, we found the count of the people comparing work-class and age shown with the image on the right. Each work-class was further separated by the 4 age groups (<20, 21-40, 41-60, >60) to find the specific count value. From the chart shown on the right, we were able to obtain the counts from each workclass and create an individual pie chart for each age group by plotting the counts using matplotlib. The image showing a bar graph between work class and age shows the distribution of count values between work class and age. The image on the far right shows the exact count values and the values are visualized with the bar graph next to it and will be further analyzed with the proceeding bar graphs for each age group and profession.

Pie chart for age below 21 and Professions



From this pie chart we can conclude that most of the people in this age group work in a private profession. The next largest profession is government and self-employed. This is a reasonable observance because private profession jobs include the services and goods industries.Compared to the other age groups, there is little variation among work professions and that can be attributed to the fact that people at that age may be focused on school.

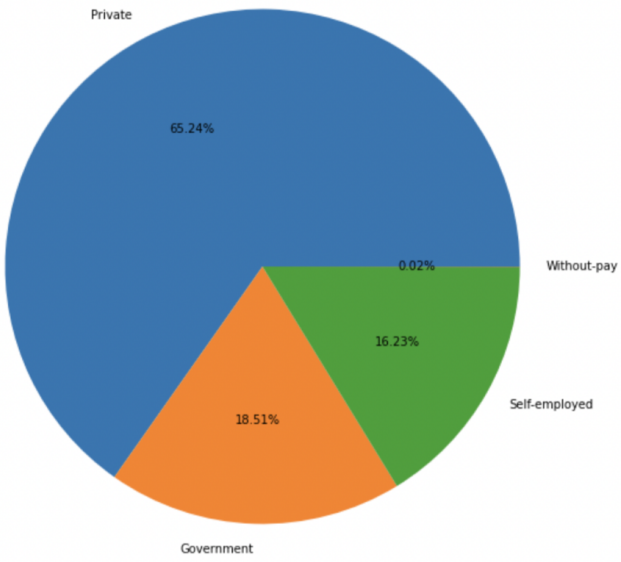
Pie chart for age 21-40 and Professions



For this pie chart, there is still a majority of the people that work in the private profession for this age group. There seems to be more spread between the other work classes compared to the pie chart for people under 20.For example , government jobs require certain experience in an age group of <20 which they lack. The section of the chart that has very little

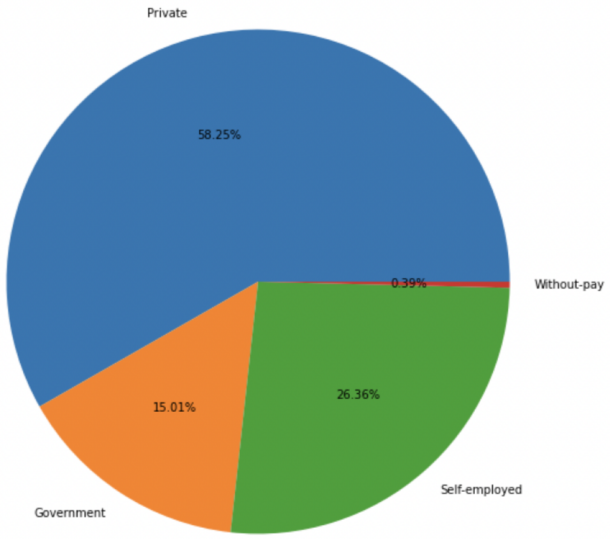
values is the without pay group, indicating that there is a large tendency for people to work between these ages. The increase in variation between the work classes can be attributed to workers gathering more experience.

Pie chart for age 41-60 and Professions



For this pie chart, the largest section is the private profession section. The next largest section is the government and then the self-employed, and the smallest section is the profession without pay. For this age group, the amount of self-employed workers just about doubles from the percentage of self-employed workers in the age group 21-40. This shows an influx of workers transitioning into self-employment at this age which could be attributed to people changing professions with more experience

Pie chart for age above 60 and Professions



For this pie chart, the largest section is the private profession. The next largest section is the self employed, and then government and then those in a profession that is without pay. Out of all of the age groups, the age group above 60 has the largest percentage of people that are in the profession without pay. This is a reasonable observance because around this age people will be eligible for

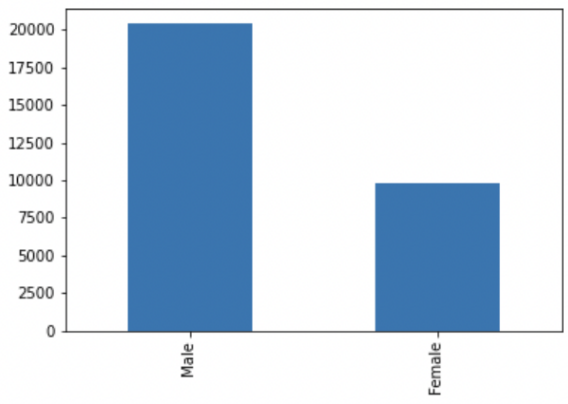
retirement or may be a part of a non- profit organization.

Among all of the pie charts, it looks like as the age increases then the amount of variation between jobs also increases. When comparing all of the pie charts, there is a majority of people in every age group that works in the private profession, but as the age increases the amount of people in the government and self-employed professions increases and the amount of people that work in the private profession decreases. This is a reasonable observance because jobs in the private profession include the service industry which typically has a larger percentage of younger workers. Another attribute that supports the amount of private profession workers decreasing with age is that as the workers gain more experience/knowledge they will be able to seek other professions like government or self-employment as these types of professions require the right experience.

Part 4c. Comparison of male/female education (in 4 groups)

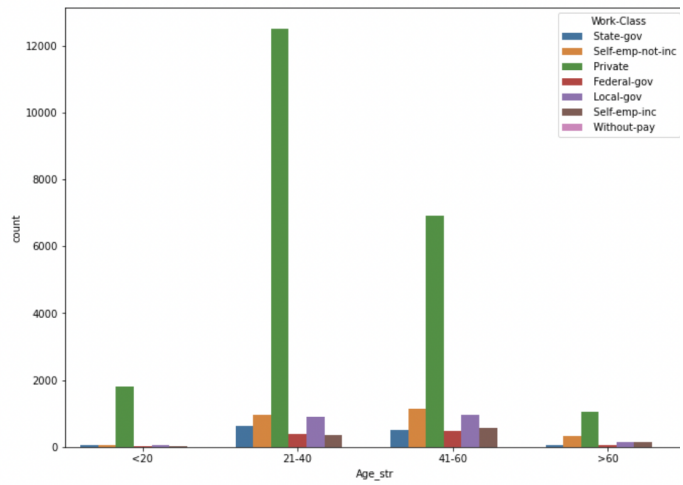
When we add all of the female and male counts that are in the data set we can see that there are far more males than females. Between some education levels there is a large count gap between males and females. As shown with the image on the right, the male education count for highschool graduates is 6734 and the count for female highschool graduates is 3106. The count for males with bachelor's degrees is 3522 and the count for females with bachelors is 1522. The count for males with masters is 1118 and the count for females with masters is 509. The count for males in professional school is 455 and the count of females in professional school is 87. The reason why we chose those education levels is because those education levels have the greatest variation. There is a clear distinction that there are twice as many males in every level of education compared to females. One way I can attribute the gap to is that they sampled twice as many males compared to females which can be shown with the image on the bottom right. The image shows the total count values between males and females in the census data. Also if we look at the data of bachelor category we can see that the number of females who were pursuing bachelors

```
: Sex      Education
  Male      HS-grad          6734
            Some-college     4171
            Bachelors        3522
  Female    HS-grad          3106
            Some-college     2507
            Bachelors        1522
  Male      Masters          1118
            Assoc-voc         852
            11th              677
            Assoc-acdm        613
            10th              570
  Female    Masters           509
  Male      Prof-school       455
  Female    Assoc-voc         455
  Male      7th-8th           425
  Female    Assoc-acdm        395
            11th              371
  Male      9th               336
            Doctorate         294
            12th              255
  Female    10th              250
  Male      5th-6th           219
  Female    7th-8th           132
            12th              122
            9th               119
  Male      1st-4th           108
  Female    Prof-school        87
            Doctorate          81
            5th-6th            69
            1st-4th            43
  Male      Preschool          31
  Female    Preschool          14
dtype: int64
```

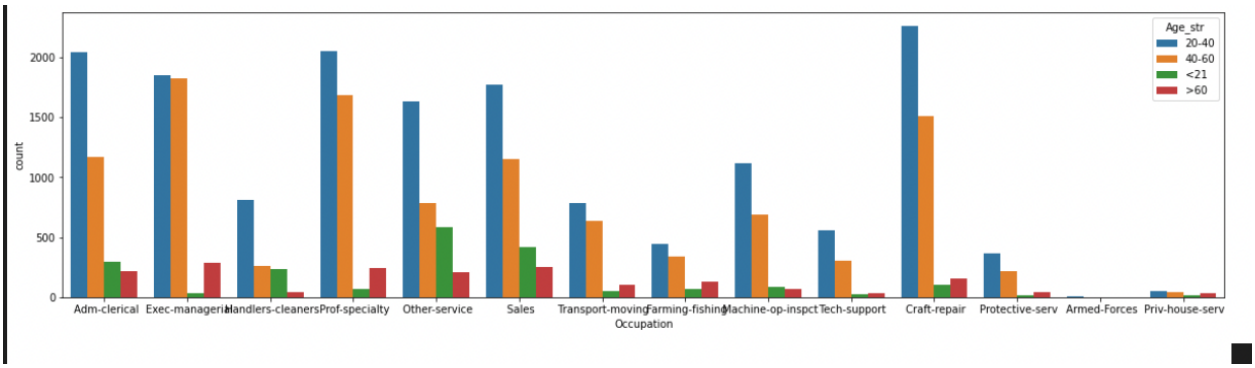


Priyam Dalwadi(1001994810)
Kelsey Nguyen (1001836837)
DASC 5300-001
significantly decreased in masters.Another thing that could be attributed to the gap disparity could be the timing of the census. Education opportunities for males and females may be drastically different which could have caused the education gap. The count values of females are slowly decreasing as the education level increases.

Part 4d. Top 5 occupations and professions for the population



With this histogram, we evaluated the top professions among all of the age groups. It shows that from all of the work classes that most people work in the private, self-emp-not-inc, local-gov, state-gov, and self-emp-inc. For the histogram below, we are able to visualize the most common



top 5 profession fields =
1. Private
2.Self-emp-not-inc
3.local-gov
4.state-gov
5.self-emp-inc

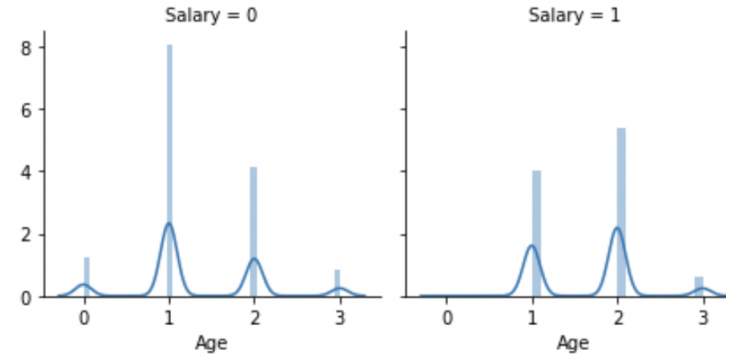
top 5 occupations =
1.Craft-repair
2.Prof-specialty
3.Adm-clerical
4.Exec-managerial
5.sales

We used matplotlib to plot a histogram that compares data occupation and data age to be able to visualize the top occupations for the population. After analyzing the

Priyam Dalwadi(1001994810)
Kelsey Nguyen (1001836837)
DASC 5300-001

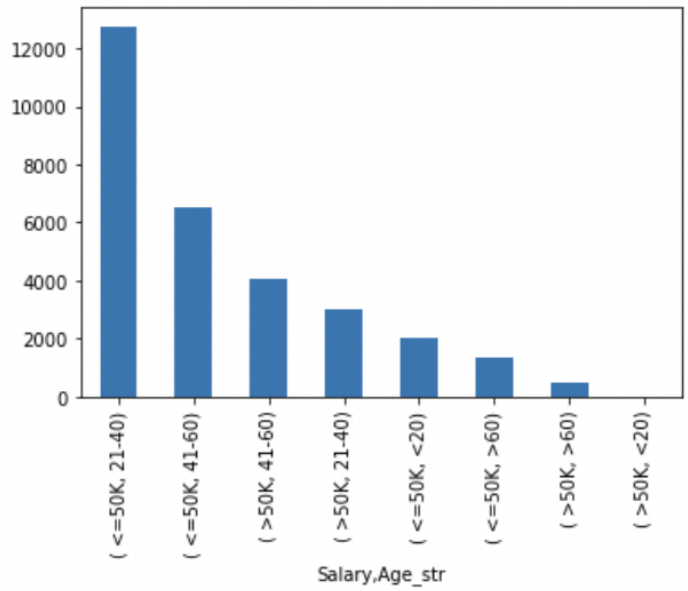
histogram, we can conclude that the top occupations are Craft-repair, Prof-specialty, Adm-clerical, Exec-managerial, and Sales. These occupations have the highest frequencies of people between all of the age groups.

Further analysis



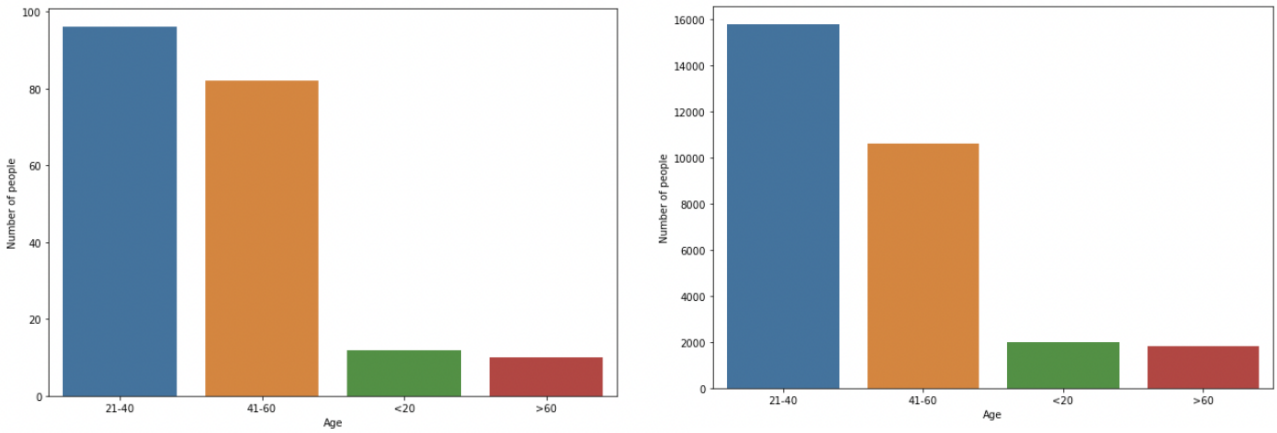

```
Salary = 0 (<=50k) | X-axis: 0 = <21
Salary = 1 (>50k)  | Age     1 = 21-40
                   |         2 = 40-60
                   |         3 = >60
```

In this data the target variable is Salary and others are independent variables. So the Salary variable is dependent on all other variables. We plot a graph of 4 ages of categories <21, 20-40,40-60 and >60. The graph on the left side targets the age group of having a salary less than 50k. It can be observed that the age group 21-40 has the maximum number of people whose salary falls under 50k. I think this is because people of this age group are just starting their professional career where salary is not initially high.The lowest on both the graphs are age group >60 where, i think people of this age group might retire, join charity, or non profit organisation. Also there are very few people in the age group of >60 who are having less than and greater than 50k salary. Furthermore, if we look at the age group of 40-60, they are less in number in the left graph and number increases significantly on the right graph. So we can assume that as people get more experience in their profession they are more likely to get more salary. Also as years pass, if we look at the positive side, we can assume that there is growth in their business because of that age group 40-60 are more likely to earn more than 50k.



Priyam Dalwadi(1001994810)
Kelsey Nguyen (1001836837)
DASC 5300-001

Random sample vs Whole data



Here we plot 2 graphs of random sample vs whole data.
We can clearly conclude the result that both graphs are similar to each other. The only significant difference between the sample vs the whole data bar graphs is that the values between 41-60 for the whole data is less than the random sample.