CSE 578: EXECUTIVE REPORT
MARKETING PROFILES

Shivani Priya (1219492950)
Atin Singhal (1217358454)
Varoon Parthasarathy (1216905976)
Avinash Khatwani (1216635004)
Uttam Bhat(1217142043)
Anil Kumar (1219401495)

PROBLEM STATEMENT

– – –

- To develop marketing profiles of individuals where salary is chosen as a key factor to determine criteria for marketing its degree programs.
- $50,000 should be the key number for salary.
- Data for analysis is provided by United States Census Bureau.
- Determine the features that are relevant for people making income above or below $50,000.

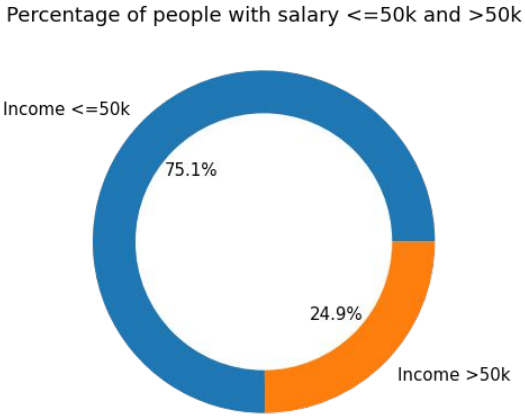
INITIAL ANALYSIS

Data Cleaning & Preprocessing:

- Removed irrelevant symbols & characters from the data.
- Checked for NULL values in the dataset.
- Trimmed the white spaces from the values of the features.
- Divided data into two major categories- ***income > 50K*** and ***income <= 50K***.

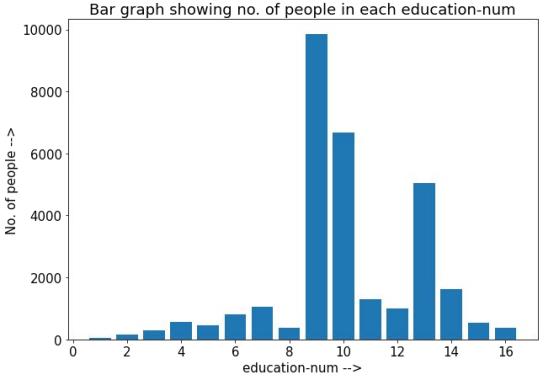
Initial analysis was done after plotting ***heatmap*** and ***scatter plot*** for the given dataset. Top 8 features based on highest correlation factor are as follows:

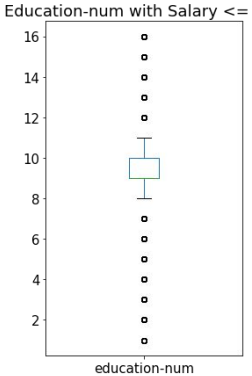
- education-num (0.335286)
- relationship (-0.251003)
- age (0.241998)
- hours-per-week (0.229480)
- capital-gain (0.221196)
- sex (0.216699)
- marital-status (-0.193518)
- education (0.078987) -> capital-loss had higher correlation than education but we were interested in comparing the results from education with education-num.

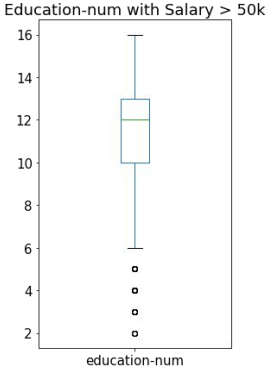


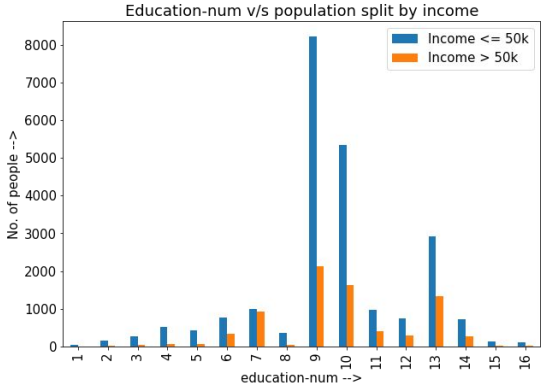
Initial thoughts: If we have to use this data for predicting the income, we'll have to figure out a way to make sure that the model is not biased as the dataset is highly skewed. We have a lot more people who earn <= 50k (75.1%) compared to people who make >50k (24.9%).

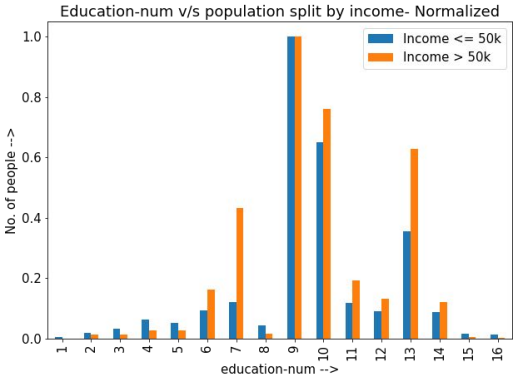
EDUCATION-NUM











Education-num has the **highest correlation** with Income at 0.33 so we decided to explore and see if this will be relevant for our purpose of prediction.

Visualizations: Bar graph, Box plot

Inferences: People earning <=50k will be in the range of 8-11 years of education (with the median being 9), whereas people earning >50k will be in the range of 6-16 years of education (with the median being 12).

Also, people who have >10 years of education are more likely to earn >50k.

Taking the above results into consideration, education-num should be a good feature to include while training the model.

RELATIONSHIP

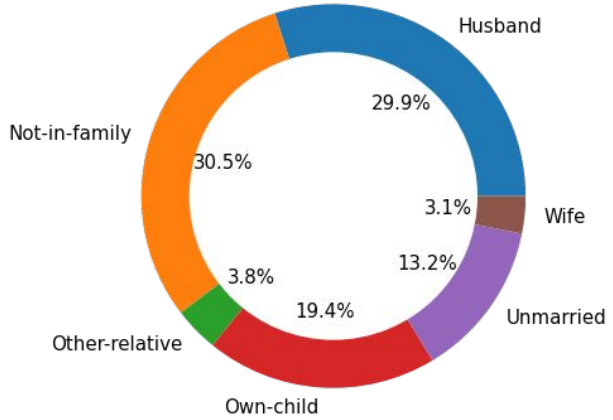
Relationship has the **second highest correlation** with Income at -0.25. Even though the value of correlation is negative, it should not impact the end goal- selecting features to train model for prediction.

Visualizations: Donut plots

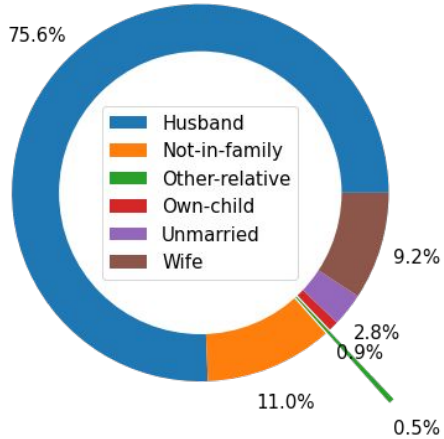
Inferences:

- Most of the people earning <= 50k are either husbands or not in a family.
- More than 3/4th of the people making > 50k are husbands.
- ‘Husband’ and ‘Wife’ are the only categories which have a higher probability amongst the people making >50k. The other categories (not-in-family, other-relative, unmarried and own-child) have a higher probability of being amongst the people making <50k.

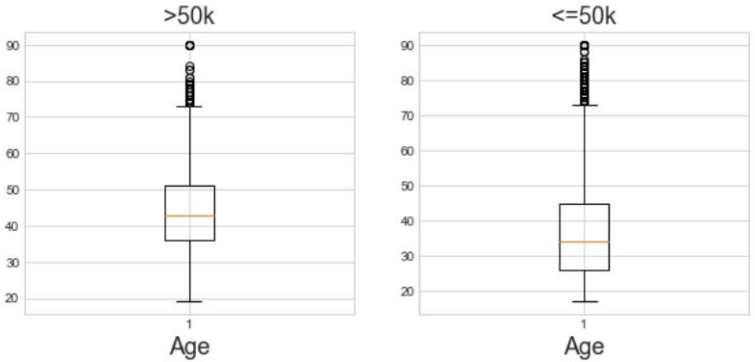
Relationship Status for People with Income <=50k

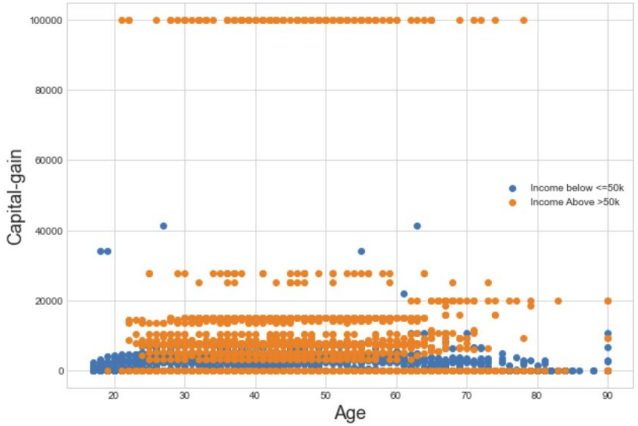


Relationship Status for People with Income >50k



AGE





After the initial analysis, it was determined that age has a **high correlation** with Income at 0.24. On further analysis, we formed the following insights:

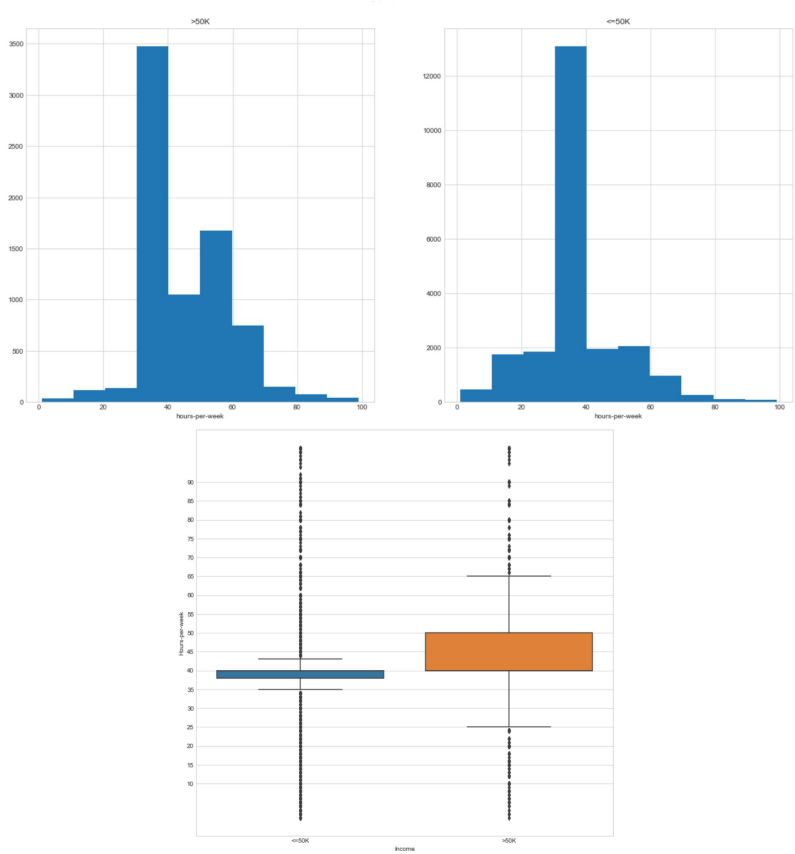
Box Plot: Visualization of the **Age** Feature.

- The Box Plot provides insights over the distribution of age for the two classes of income.
- Individual with a higher age are more likely to have an income above 50k when compared to the other class.

Scatter Plot: This plot covers the Age and Capital Gain Feature with the **Age** on X axis and **Capital Gain** on Y axis.

- The two classes of data seem to be well separated.
- People with a higher capital gain seem to be the ones earning more than 50k.

HOURS PER WEEK



The initial analysis revealed that hours per week was highly correlated with Income with a value of 0.23. So we decided to explore further.

Visualizations: Box plot and Bar Graph.

Inferences:

- The majority of people who earn less than 50k, do their work close to 40 hours a week on average.
- People who earn more than 50k, work more than 40 hours (less than 50 hours) on average per week.
- The majority of people work for 40 hours per week on average from both the categories(less than 50K and greater than 50K).
- From the box plot, we can see that those who earn more than 50K tend to work 7 hours more on average compared to those who earn less than 50K.


THANK YOU