# CSE 578 : Data Visualization

# MARKETING PROFILES GENERATION BASED ON INCOME

SYSTEM DOCUMENTATION REPORT


**Group 2 (The Serial Visualizers)**

Arushi Gaur (1219396022)
Kartik Paigwar (1219739716)
Madhu Shreepathihalli Shivakumar (1219396334)
Natesh Tyagi (1219775492)
Subrahmanya Sai Krishna Madduri (1223013168)


## Roles and Responsibilities

| Responsibility | Team Member | Date |
|---|---|---|
| Installation and Docker Setup | Kartik | March 20, 2022 |
| Understanding the *adult* dataset | All members | March 23, 2022 |
| Cleaning the dataset | Natesh and Kartik | March 25, 2022 |
| Discussing and finalizing the user stories | All members | March 30, 2022 |
| Visualization of Income with Sex | Natesh | April 1, 2022 |
| Visualization of Income with Education | Natesh | April 1, 2022 |
| Visualization of Income with Hours Per Week | Kartik | April 1, 2022 |
| Visualization of Income with Marital Status | Kartik | April 2, 2022 |
| Visualization Income with Age and Sex | Arushi | April 2, 2022 |
| Visualization of Income with Sex and Race | Arushi | April 3, 2022 |
| Visualization of income with age, education and hours worked | Subrahmanya | Apr 4, 2022 |
| Visualization of income with age, education and capital gain | Madhu | Apr 5, 2022 |
| Multivariate Analysis | Madhu | Apr 5, 2022 |
| Executive and System Report | All members | Apr 8 - Apr 15, 2022 |

## Team Goals and Business Objective

In today's world, with the increase in the importance of marketing, targeted marketing is the need of the hour. XYZ Corporation uses data to develop marketing profiles of people which are used for marketing purposes. A project from UVW College about using salary as a key factor to promote its degree programs requires the generation of marketing profiles from the data provided by the United States Census Bureau. The salary of interest is $50,000, we need to figure out what factors must be considered to determine whether the salary will be above or below $50,000.

The business objective of the college is to determine the target audience to market its degree programs based on the given information like gender, workclass, race. To achieve this objective, we have compared various factors like education, gender, age with income how they impact them. To visualize the comparison and for effective impact analysis, we have plotted multiple different graphs for different parameters based on the type of relationship. For example we plotted Mosaic plots to visualize the relationship between sex and income.

Understanding this relationship will help the UVW College in marketing its degree programs. For example, if we know that white females under 30 with high school diplomas are more likely to earn less than $50,000, UVW College can advertise the appropriate program for females in that demographic location.

## Assumptions

We are using the United States Census Bureau dataset for year 1994 and have assumed the following points:

1. **Correctness**: We are assuming that the data is correct and precise.
2. **Distribution**: We are assuming that the data distribution is not skewed to avoid any misleading information from the visualizations.
3. **Independent Features**: We are assuming that all the 14 key parameters are always independent of each other, that is there is no correlation.
4. **Working Professionals**: We are assuming that the dataset includes only working professionals, that is people with less than $50k income does not include non working candidates.

## User Stories

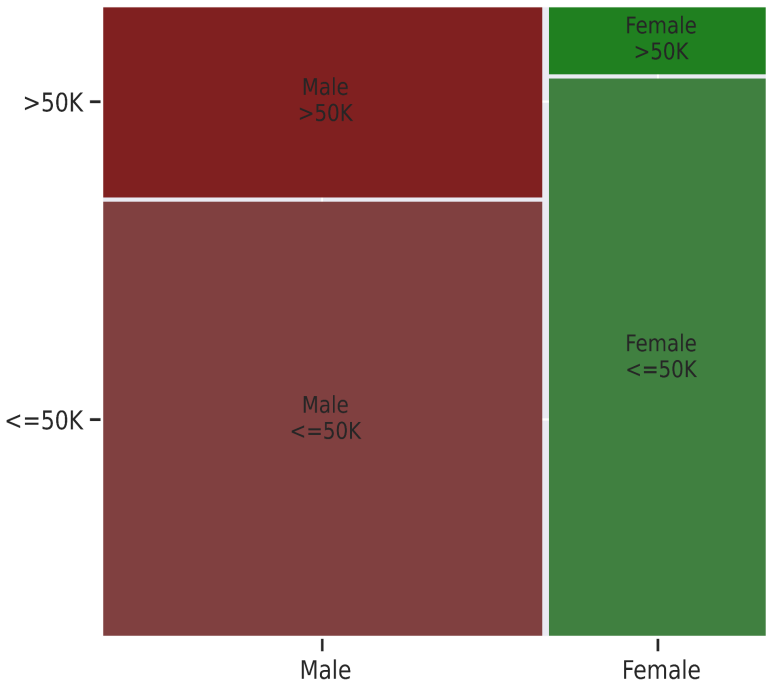
To achieve the business objective and address the concern of UVW College of performing targeted marketing, we are using the following user stories:

1. As a part of the UVW Marketing team, I want to understand the disparity in income due to **sex**.
2. As a part of the UVW Marketing team, I want to understand how income is affected by the highest level of **Education** of a person.
3. As a part of the UVW Marketing team, I want to see whether more **hours per week** always means more income.
4. As a part of the UVW Marketing team, I want to understand how the **marital status** of a person impacts the amount of money they are making.
5. As a part of the UVW Marketing team, I want to understand how **Age and Sex** together impact the income of an individual.
6. As a part of the UVW Marketing team, I want to understand how **Sex and Relationship** impact the

income of an individual.

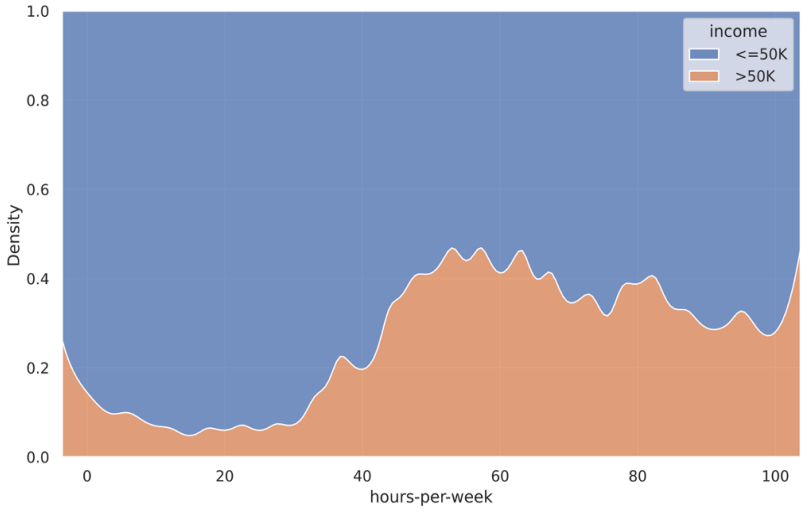
7. As a part of the UVW Marketing team, I want to understand how **Education, hours worked per week, capital gains and sex** impact the income of the indicidual.
8. As a part of the UVW Marketing team, I want to understand how **Age, education and hours per week** affect the income of an individual.
9. As a part of the UVW Marketing team, I want to understand how **fnlwgt** affects the income of an individual.

## Univariate Visualizations



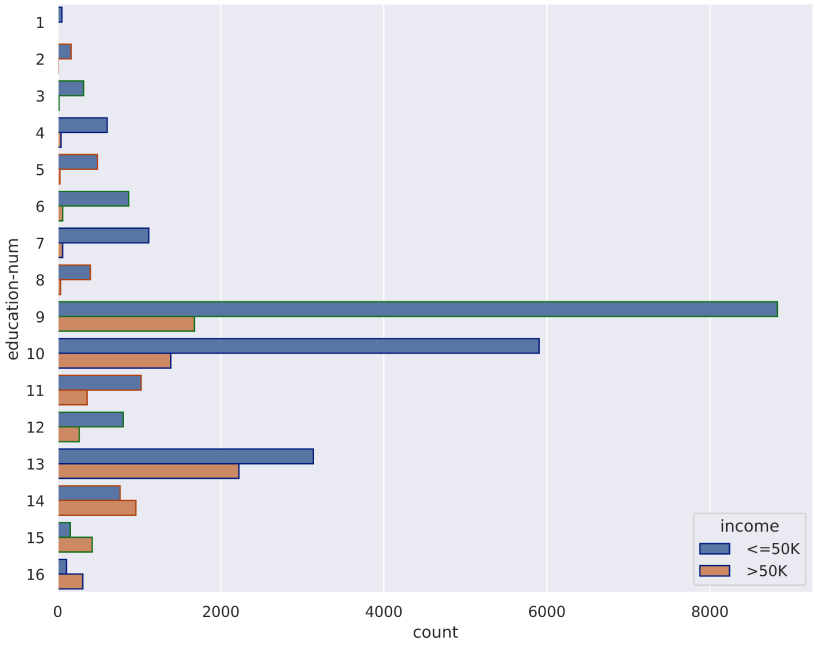

**SEX**

- Mosaic Plot
  - Features covered: Sex
- Inferences:
  - Most of the people >50% regardless of their gender, earn less than $50K.
  - Males have a higher proportion of earning more than $50K in comparison to their counterparts.
  - Females have a very less proportion of their population earning more than $50K, so, UVW college can focus on creating awareness amongst the female mass about the benefits of a college level education.




**HOURS-PER-WEEK**

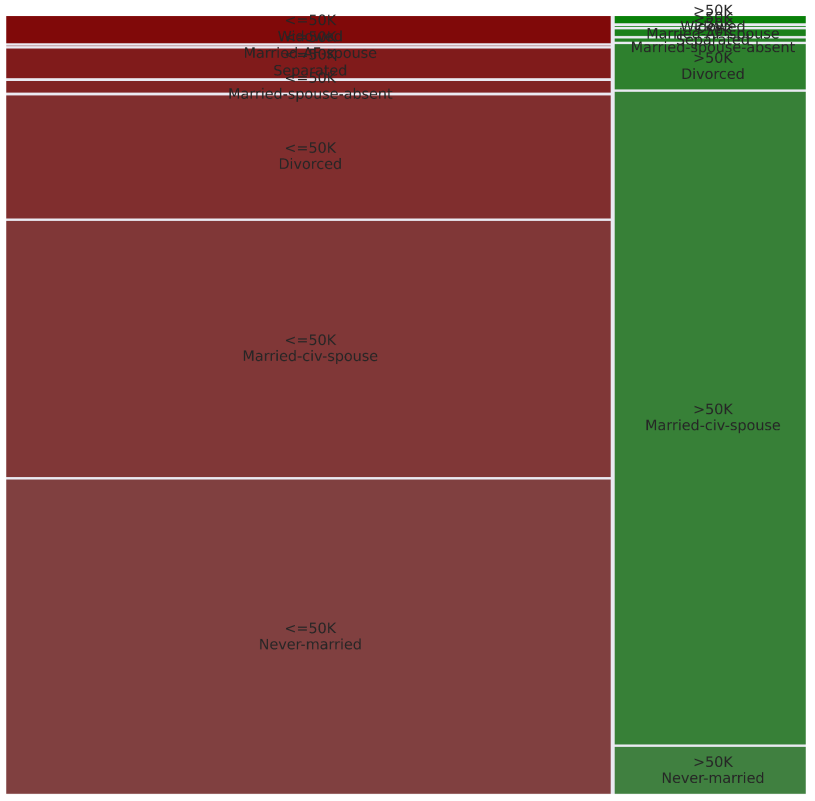
- Density Plot
  - Features covered: hours-per-week
- Inferences
  - The people who work less than ~30 hours/week generally earn less than $50k.
  - Almost 40% of the people who work more than ~50 hours/week earn more than $50k.




## EDUCATION-NUM

- Horizontal bar chart
  - Features covered: education-num

- Inferences
  - As the education-num increases, the proportion of the number of people with income over 50k increases.
  - This inference is intuitive as the level and quality of education of a person increase, their income increases proportionally.
  - As the education-num gets significantly higher, the number of people having income over 50k become more as compared to their counterparts.

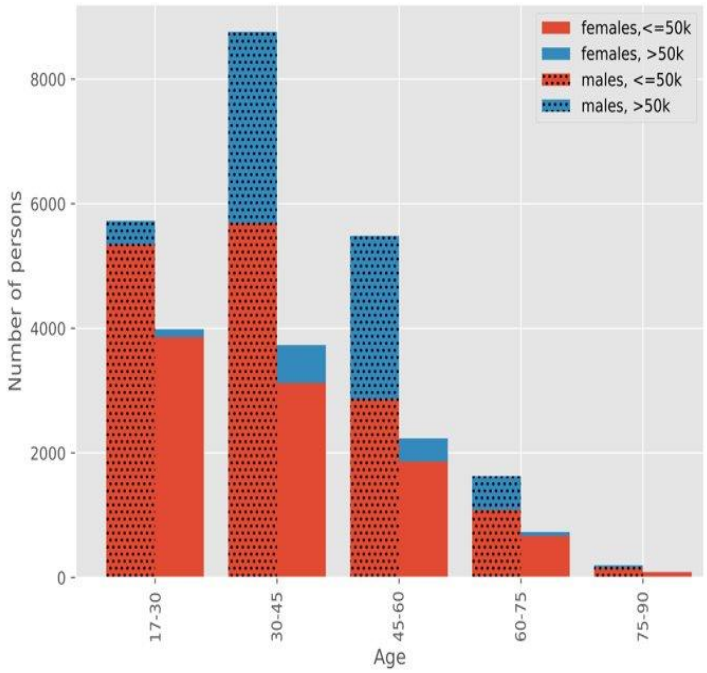



## MARITAL STATUS

- Mosaic Plot
  - Features covered: Marital Status

- Inferences:
  - If a person earns more than >50k they are more likely married and have a civilian spouse. UVW can target children of these people for admission in expensive degree programs.
  - If a person earns less than <=50k they are more likely never married or divorced.

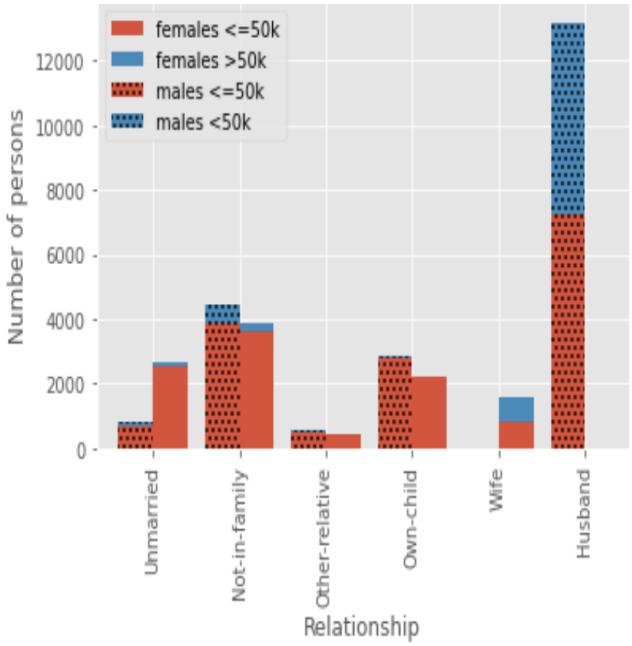
# Multivariate Visualization




**SEX & AGE**

- Stacked Bar Chart
  - Features covered: Age, Sex

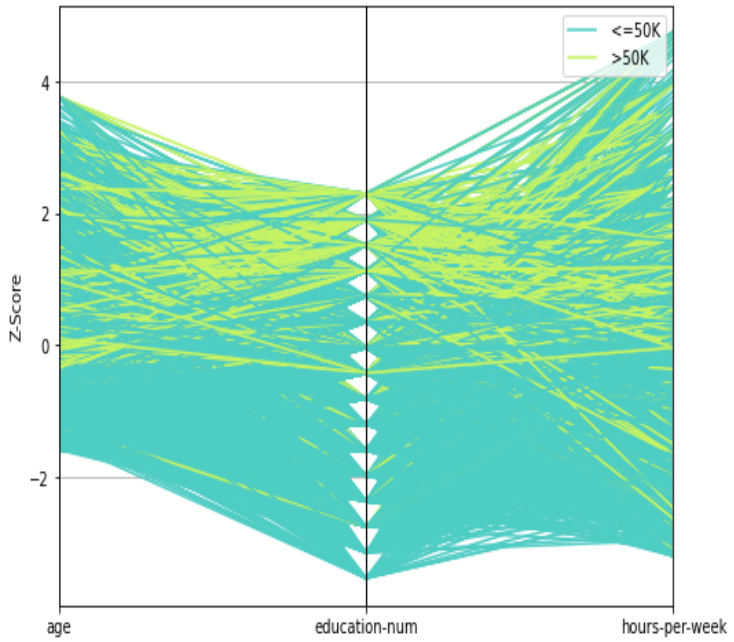
- Inferences:
  - Across all the age groups, males have a higher proportion of income greater than $50K in comparison to females.
  - We can see the above effect more pronounced in the two age groups: 30-45 & 45-60. UVW college can then marketeer their advertisements focused on these two groups.
  - People in the age group of 17-30 regardless of their gender generally earn less than $50k owing to their younger age.




**Relationship & SEX**

- Stacked Bar Chart
  - Features covered: Sex, Relationship

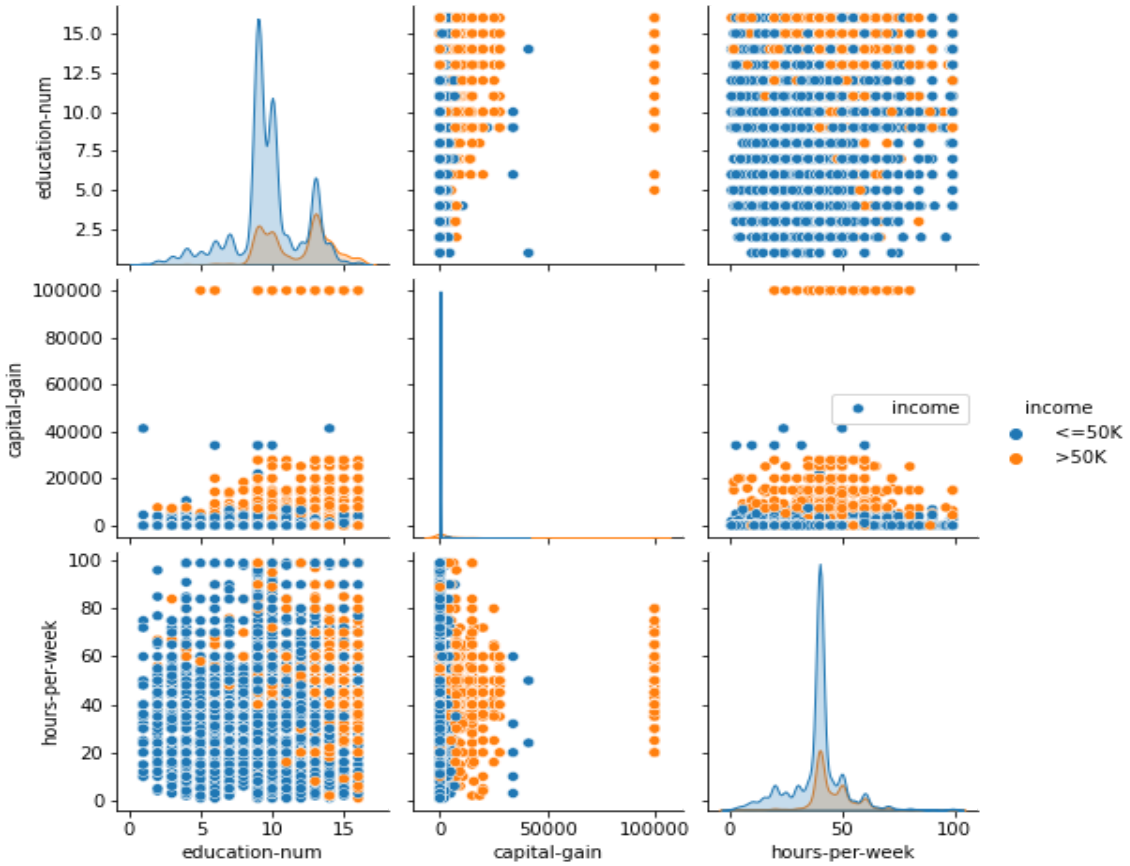
- Inferences:
  - Married people are more likely to earn more than $ 50k.
  - Own-Child is likely to earn less than $50k and hence can benefit from the UVW College's degree program.
  - Very High proportion of unmarried females are earning less than $50k and these can be the target audience for UVW College.




## AGE & EDUCATION-NUM & HOURS/WEEK

- Parallel Coordinate Plot
  - Features Covered: Age, hours/week, education-num
- Inferences
  - There are a significant number of people with less education-num who are working more hours, regardless of their age and yet earning less than 50K.
  - If a person has an average age, high education-num and works for more hours than average, generally earns more than $50K.

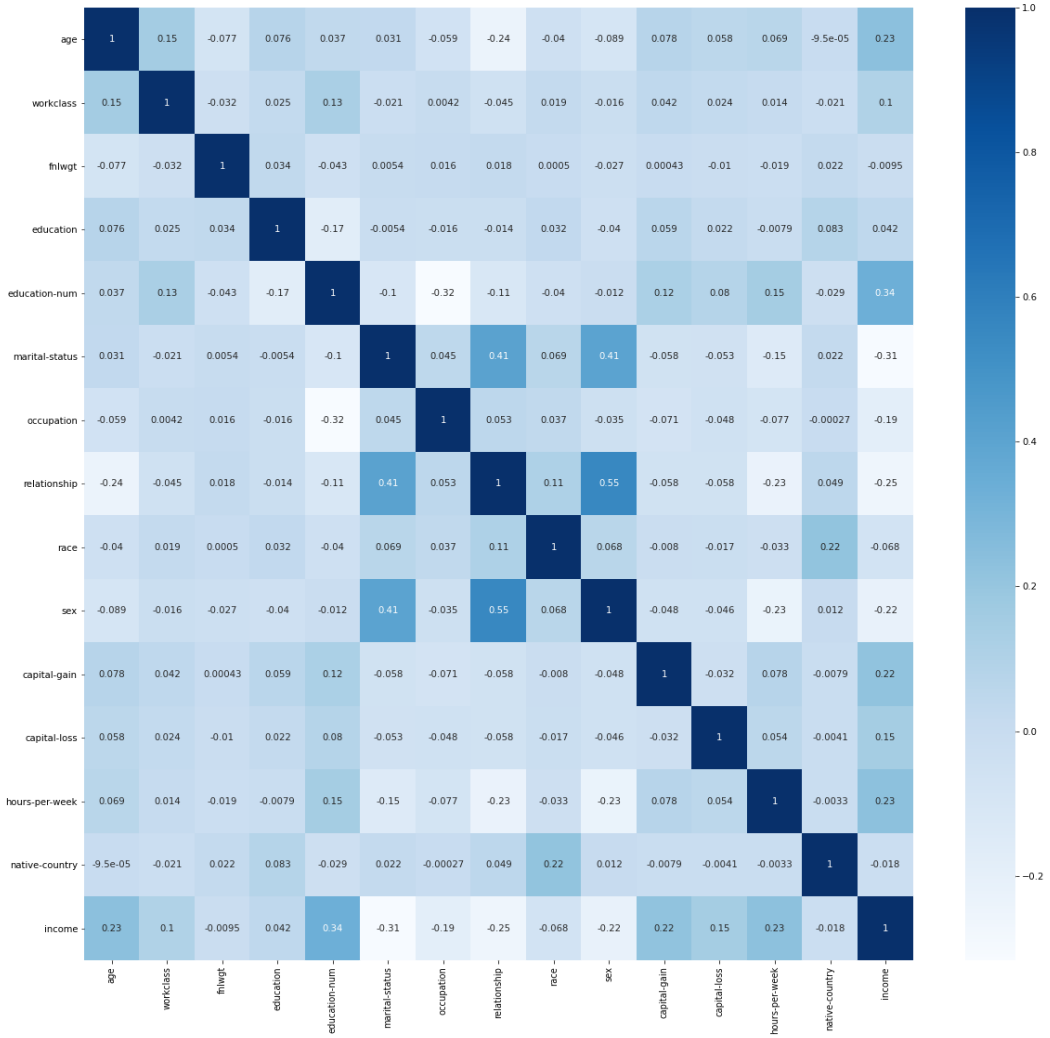
## CAPITAL-GAIN & EDUCATION-NUM & HOURS/WEEK



- Scatter Plot Matrix
  - Features Covered: capital-gain, hours/week, education-num
- Inferences
  - We can clearly see that as the education-num increases, the capital-gain is also increasing which in-turn has its effect on the income of a person which is generally above $50K.
  - If the education-num is low and the person works for less number of hours, generally the person earns less than $50K.
  - If your education-num is low, even if you work for more than 50 hours per week, the person's income is generally going to be less than $50K.

## Correlation Analysis (Heatmap)

Correlation between one feature to another can be calculated using a covariance matrix. We plotted the covariance matrix in the form of heatmap. We can see from the bottom row that income has the highest correlation with education num then age, hours-per-week and capital-gain subsequently. We can also see that fnlwgt has approximately zero covariance which makes it a redundant feature.



## Questions

1. What kind of visualization to be used to represent the data?
   We performed in depth study of multiple visualizations and analyzed the data type for the visualizations. For categorical features, we used plots like mosaic plots and for numerical data, we used bar charts, density plots and scatter plots.

2. How to fill the missing data entries in the dataset?
   We filled the missing categorical data with *mode* of the feature column while for numerical data we used the *mean* value of the feature.

3. How to convert the categorical data against numerical data?
   We convert the categorical data into numerical value by assigning a unique number to the category in range of total no. of categories.

4. Which parameters to analyze from the given set of 14 key variables?
   Each parameter affects the income in its own way. We tried to choose the variables that are either impacting the income directly or play an important role in understanding the income distribution so that UVW College can address the targeted audience. We plotted the heatmap for covariance matrix and identified the important features based on their correlation values.

## Not Doing

There was a requirement to predict the income of an individual based on the given set of inputs which would have required training a machine learning model and dividing the dataset into training and testing dataset. We did not implement this machine learning model for multiple reasons. Firstly, the fourteen key variables of the problem statement are independent of each other, they might affect income together in a certain pattern but they themselves are independent of each other. Since it is not possible to determine the exact relationship, it will not be possible to understand the impact of the variables in the final result. For example, gender and age affect income, females in the age group 30-60 are more likely to earn less than $50k than males in the same age group. But we can't determine the exact contribution of gender and age group in the final result.

Secondly, we have assumed certain things about the dataset like data not being skewed, data including only working professionals. Even if these assumptions are not true, inferences from visualizations will still be helpful but the predictions of the model will be wrong resulting in wrong deductions.

## References


1. https://seaborn.pydata.org
2. https://www.kaggle.com/learn/data-cleaning
3. https://www.kaggle.com/learn/pandas


## Appendix

- **GitHub Repository** : https://github.com/kartikpaigwar/CSE578Project