1. **DATA SOURCE**

The data source for our group assignment is from (<http://www.cs.toronto.edu/~delve/data/adult/desc.html>)

The dataset is a description of adult income along with an individual’s personal information. An individual’s annual income results from various factors ranging from an individual’s education level, age, gender, occupation etc.

This data was interesting because it provided the data for us to group people into categories for analysis and comparison.

1. **ANALYSIS**

Our dataset had to be merged in the first place, as it was already split into two separate parts (train and test) on the UCI website. After merging, the data consisted of a total of 48842 entries, with 15 columns. Out of these, 3620 entries had null values for a few columns and therefore, we dropped these entries. Our analysis first started with performing univariate analysis to explore each factor individually. Afterwards, bivariate analysis was performed to find out which combinations of factors led to an individual’s increased chances of earning a better income.

Once this data was obtained, we used a KNN model to predict whether an individual would earn more or less than $50k a year.

A challenge we faced while performing this analysis was deciding which factors we should focus on as many of them showed signs of correlation despite being potentially redundant such as marital status vs relationship. This issue was solved through discussion with team members to choose the factors that we thought were the most meaningful.

1. **CONCLUSION**

The conclusion we reached from our analysis was that while there were a number of factors that appeared to impact an individual’s income, the factors that were the most prevalent to us were Age, Sex, Occupation, Work Class, Years of Education, Hours worked per week, and Marital Status. These factors were chosen to train our KNN Model.

When it came to answering our previous sub-questions, interesting observations were made on which factors saw the highest incomes.

Looking at the different working classes we see that the *self-emp-not-inc* shows the highest representation out of classes that earned >50k at 55.41% followed by *Federal-gov*. This is despite both being overshadowed by the *Private* working class when looking at the entire dataset population.

Marital Status showed an interesting story where individuals that were *Married-civ-spouse* and *Married-AF-spouse* were both nearly 4 times as prevalent as other marital statuses at 45.42% and 43.75% respectively when it came to individuals that earned over $50k a year.

Males were significantly overrepresented over Females out of individuals that earned over $50k a year at nearly a 3:1 ratio with 31.24% of the male population earning more than 50k while only 11.35% of the female population managed to do the same.

Age was also another interesting factor to look at as it showed a parabola starting from the age of 20 all the way to 80 with the peak being located from the 40-60 range.

Finally, there was a strong trend where having more years of education as well as higher levels of education increased an individual’s odds of earning over $50k a year with a noticeable increase after 9 years of study.