CSE 578: DATA VISUALIZATION

Systems Documentation Report

1 INTRODUCTION

Data analysts: Abir Tawfeeq, Dima Alhaj, Tasneim Elkarakatly, Enes Bukte, Ruchi Rathod
Stakeholders: UVW College
Product owners: XYZ Corporation

We are a team of data analysts for the XYZ corporation. The objective of our current project is to develop an application for UVW college that helps to identify factors that affect an individual's salary. As a result, UVW college can use it to market the various programs provided by the college to reach out to individuals based on their income profile and bolster their enrollment.
Our team's goal is to identify patterns between various factors and salary to determine which factors of an individual affect their salary through data visualization techniques . We identify the important factors and present the top 4 most important features to the UVW executives of UVW college to help tailor their marketing strategies to bolster enrollment.

2 USER STORIES

2.1 Exploring the dataset

The United States Census Bureau has provided the adult dataset with the 14 features in total and our key variable is "income".
Firstly we explored the dataset to see what is the shape it was holding i.e how many columns and rows, the meaning of those attributes and what kind of data values are present in each column.
There are 15 columns as follows: age, workclass, fnlwgt, education, education-num, marital status, occupation, relationship, race, sex, capital gain, capital loss, hours per week, native country and income. Income is our class variable and the other 14 are our features or attributes. The types of data involved are numerical and categorical and description of these columns were given by the United States Census Bureau.
The key income value is $50,000, so individuals will be partitioned based on above(greater) or below(less) and equal to $50k salary. The value can be affected by other factors and our team evaluated them to find if there was any particular relation between income and other factors.

The original dataset contains a distribution of 24.90% entries labeled with >50k and 74.10% entries labeled with <=50k.

Income	Number	Percentage
<=50k	7508	24.10
>50k	22654	75.90

2.2 Data Preprocessing

Before we began our visualization on the data:

- The dataset had to be checked for any anomalous or null values before analysis. We loaded the data in Python and ran the info() function, this shows if there are null values, in which columns and how many along with the datatype of each attribute.

- We checked for noise and found a special character like “?” and replaced them with a null value.
- We cleaned the data by removing the missing or null data, columns/rows which contain only null entries, redundant columns/ rows, those columns which have only one unique value, etc.
- We had to convert some of the datatypes of the features from float to integer

2.3 Distributed Tasks among the Team Members

Initially we group data by key variable income to determine how many are earning more than and less than 50k. Our team selected 10 features to analyze , whether they determine an individual’s income. The 10 attributes are age, education, marital status, occupation, relationship, native country, sex, capital gain , hours per week and workclass.

1. Every team member had to understand and explore the dataset individually.
2. After the exploration task, we decided on analyzing the 10/14 attributes selected as a team that would affect the income class directly.
3. Each team member selected 2 or 3 attributes from 10 which were decided, to plot different graphs and come up with the results.
4. Each team member edited and added their inferences in parallel to the Executive Report and Systems Report.

2.4 Feature Analysis

We analyze all the attributes that the team selected and determine its impact on income. Before we begin our analysis we take into consideration some assumptions about the project.

Assumptions

1. Accurate data- We are assuming that the dataset provided to us is accurate which conveys the correct information. Dataset without accuracy can also lead to miscommunication.
2. Complete- Assuming that the dataset has the complete information. Incomplete data can also lead to inaccuracy of the data. It is also equally important to understand the full picture of the information that is tried to be conveyed.
3. Independent- each 14 features is considered independent and not affected by limiting factors.
4. Timeliness- We assume that the information provided to us is right at this moment of time. Data that is collected too early or late will lead to misinterpretation with inaccurate decision making.
5. Attributes Selection- Assuming that the attributes of factors which affect income most gives us more interpretable patterns. The feature selected and analysis done is considered unbiased.
6. Profitability- we assume that the analysis and findings will benefit UVW and bring them better outcomes in their marketing techniques. For example,we find that certain age groups earn more and hence UVW can tailor their marketing to reach out to more people in that age group and as a result get more enrollments.

1.Age- Represents the age of an individual. Around 90% of people in their twenties earn less than 50k a year, being the highest percentage of the people who earn less than 50k. In spite of that, around 60% of people in their forties earn more than 50k a year, being the highest percentage of the people who earn more than 50k a year. It shows a fairly strong relation with income.

2.**Education -** The percentage of the individuals with a professional school certificate is 0.6% in the <=50k income records.This percentage grows to 5.4% in the >50k income records and this is about 90% of the individuals who went to a professional.
The doctoral holders account for 0.4% of the people that earn less than 50k a year. Conversely, they account for 3.9% of the people who earn more than 50k a year . That means 80% of the people who pursued a Doctoral degree earn more than 50k a year. Moreover, 75% of the Master holder individuals earn more than 50k a year. Furthermore, 50% of the Bachelor holders earn more than 50k.Despite that, 90% of school students, 70% of higher education

education students, and around 65% of high school students, earn less than 50k a year. With increase in education level the income increases from <=50K to >50K, indicating a strong relation between them.

3.**Marital Status -** This represents an individual's marital-status. Around 90% of the never-married and divorced groups earn less than 50K a year. On the other hand, most of the high income people are married couples. The status affects the salary range but only one relation shows stronger correlation than others. Therefore, it is not very advantageous in determining income.

4.**Occupation -** This represents an individual's occupation type. Adm-clerical, Craft-repair and Other-service occupations have the highest proportion of low income. Prof-speciality and Exec-managerial are the most paid, around 50% earn more than 50K a year. Occupation acts as one of the top 4 attributes with a huge impact on the individual's income.

5.**Relationship- :** This represents an individual's relation to others. This factor has more effect on the income as seen in the graph. People that belong to a relationship "Husband", has approximately 75% count of the total people who are earning more than 50K. Also, people with a relationship of "Un-married" and "Own-Child" earn <=50K of the income.

6.**Native country-**There are 41 countries but the data distribution is in the US and Mexico mostly and majority of the people are earning less than or equal to 50K in these 2 countries. For remaining countries the number of people earning shown is very less. Therefore, it is not considered to be influential in determining income.

7.Sex- The 2 types of sex given are male and female. About 31% males are earning >50K whereas only 11% females are earning >50K. However, 89.1% females are earning <=50K and 69.4% males are earning <=50K. This relation between income and sex shows that sex influences income greatly.

8.**Hours per week-** These are the hours an individual person has reported to work per week. Looking at this factor, a good number of people were working more that 40 hours which leads into a prediction that they earned >50K.

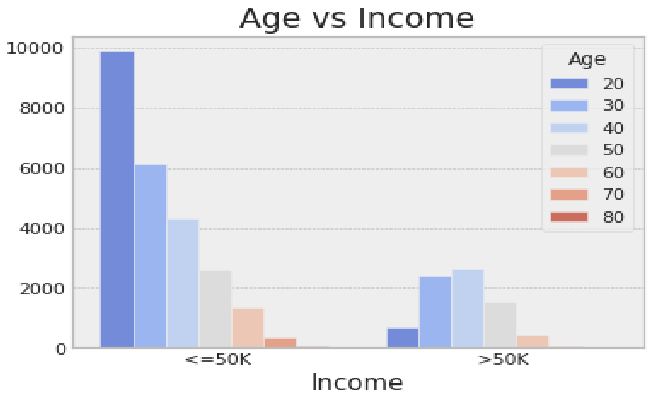
9.**Capital gain-** This is the capital gain for an individual. If the Capital is high, there is a probability that the person will earn >50K of income. Capital gain proves beneficial in determining income to a certain degree.

10. Workclass- A specific term used to describe a person's job situation. Private sector had the highest number of people for both >50K and <=50K and remaining work sectors have almost the similar percentage of people in both income groups. Only the private sector has the strongest influence on the income of individuals. Therefore it is considered to have little influence on the income.

2.5 Visualizations and Interpretations

Below are the graphs of 10 attributes that helped us conclude and predict the factors that affected income:

1.Age



The bar chart illustrates how many in each age group are earning either >50K or <=50K. Between ages 20-40 a decline is seen in the number of people earning <=50K but a rise is seen >50K. This shows a strong relation between age and income. From 50- 80 the number of people earning is very less in both groups. A bar chart helps group the ages and color them according to group to give a better picture of age group with income.

Fig. 1

2.Education

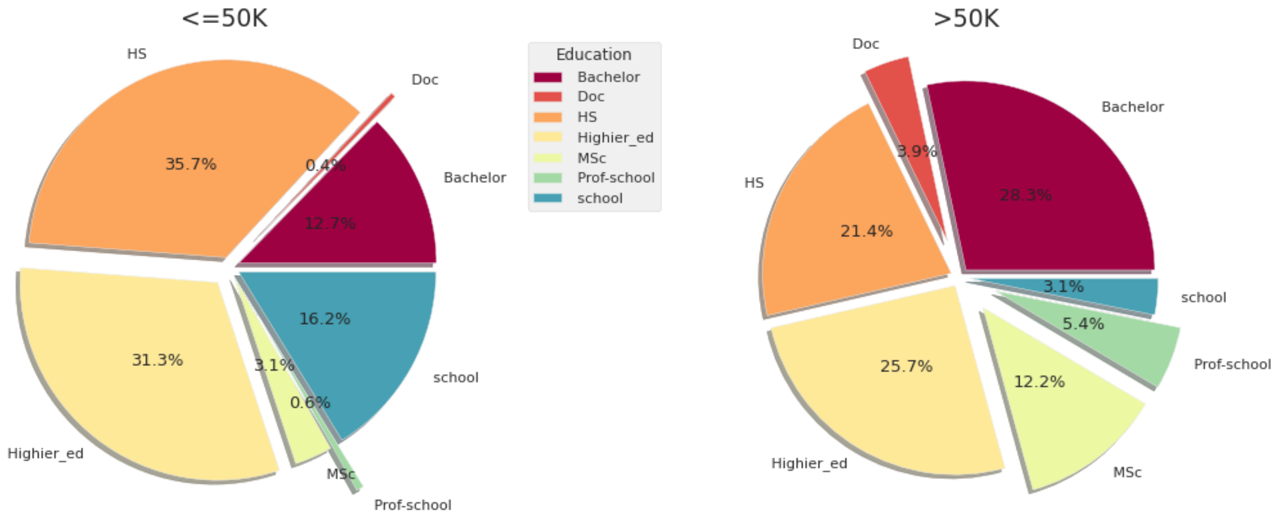


Fig 2

The pie charts above show the contribution of education to an individual's income. Exploded pie chart gives an overall understanding of education level with income by looking at the pie slice sizes, the gaps help compare the very small groups also and clear distinction can be made between increasing education level and income. The graphs show that 90% of the professional schools students, 80% of the people who pursued a Doctoral degree, 75% of the Master certified people and 50% of the Bachelor holders earn more than 50k a year. In contrast, 90% of school students, 70% of higher education education students, and around 65% of high school students earn less than 50k a year.

3.Marital Status

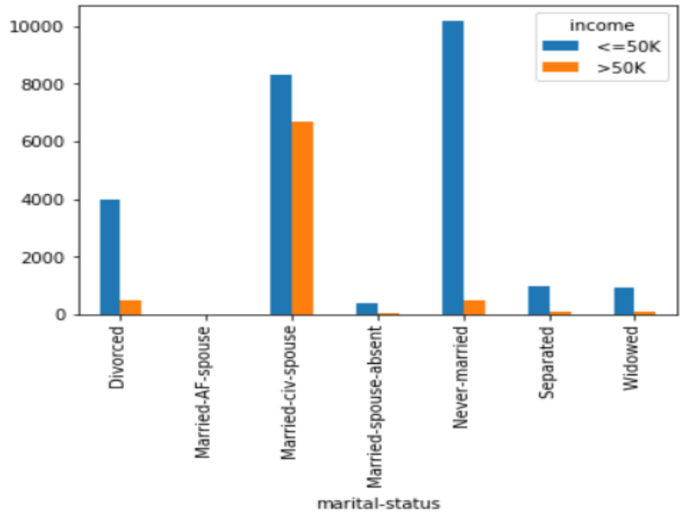


Fig. 3

As seen in **Fig 3**, we can say that around 90% of the never-married and divorced groups are earning less than 50K a year. On the other hand, most of the highest income earning (>50K) individuals are married couples. Marital status don’t have a great impact on the income compared to other attributes.A bar chart with color code for 2 income groups helps analyse categorical data with income.

4.Occupation

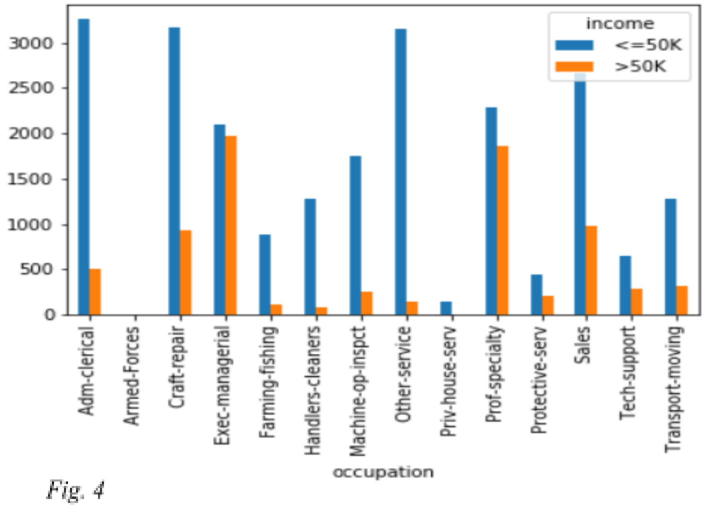
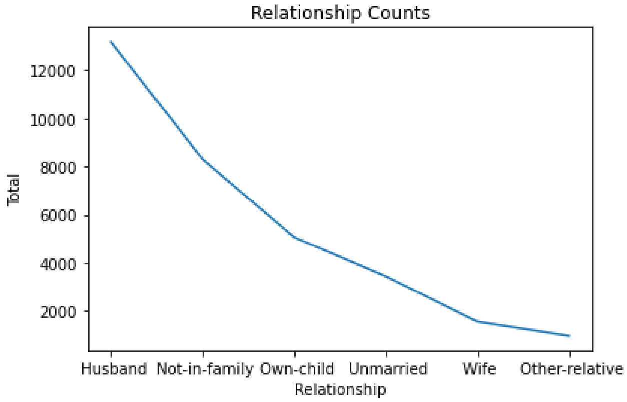


Fig. 4

As seen in **Fig 4**, ”Adm-clerical, “Craft-repair” and “Other-service” occupations have the highest proportion of low income. Prof-speciality and Exec-managerial are the most paid, around 50% earn more than 50K a year. Occupation acts as one of the top 4 attributes with a huge impact on the individual's income. With 14 categories in occupation it is too many for a pie chart therefore a grouped bar chart is more reasonable and the data range is also clear to show the difference between the salary groups.

5.Relationship



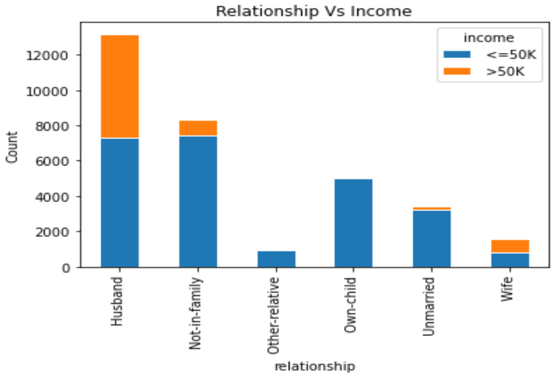


Fig 5

Fig 6

From **Fig 5,** we can see that we have the majority of “Husband” relationships with a total count of approximately 13500 and those in a relationship with “wife” and “other-relatives” we have less count with approximately 2000.Individuals with “Own-Child” are earning <=50K. As seen in **Fig 6**, people in a relationship with “Husband” earn half of <=50K and the remaining half earn >50K.The line chart helps understand the distribution of data. The stacked bar chart makes it simpler to compare between two groups of income for the relationship attribute and shows the stark contrast between different relationships and how much they earn.

6.Native country

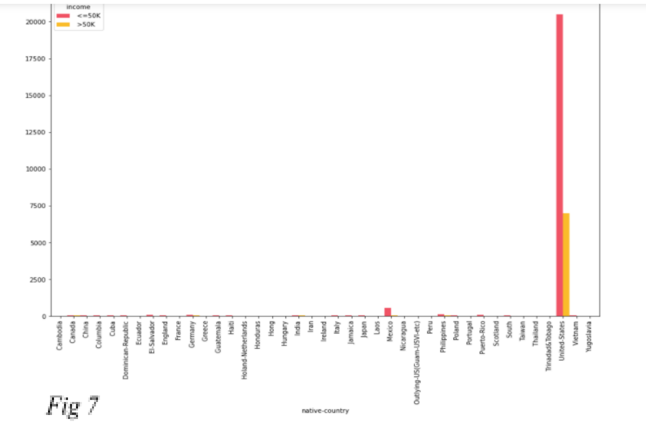


Fig 7

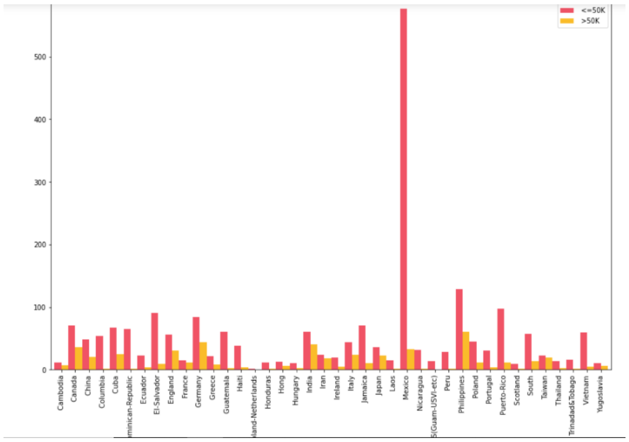
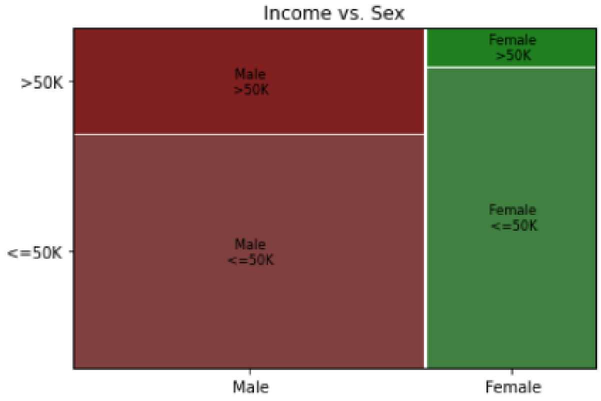


Fig 8

As seen in **Fig 7 and Fig 8,** the United States and Mexico have the majority of individuals earning. Comparing the earning rates for both of the countries, in **Fig 7**, the US has approximately 40% of the highest income compared to those earning <=50K. On the other hand, in **Fig 8**, in the Mexico country we have very less around almost 5% of the people that are earning >50K than <=50K. Since the country attribute has about 41 names,which is a large category it is more interpretable using a grouped bar chart. Moreover, it will help compare every country with the 2 income groups.

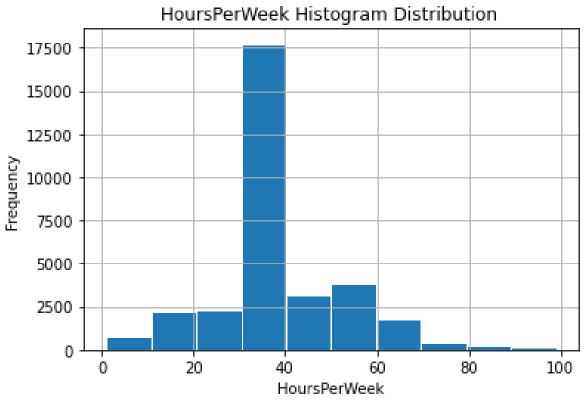
7.Sex



In this **Fig,** we can see that overall, both males and females are earning <=50K mostly but far more males are earning >50K than females. Moreover, more females are earning <=50K than male. A mosaic plot helps recognize the connection between sex and income. Since sex has only 2 categories and so does income it is more easy to plot it on mosiac plot and no cluttered data. The plot is very self explanatory and easy to interpret hence a mosaic plot is selected.

Fig 9

8.Hours per week



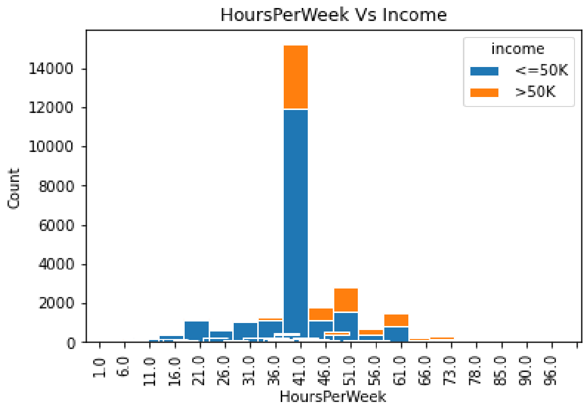


Fig 10

Fig 11

From the histogram distribution , the majority of people are working or reporting to work 40 hours per week. Regardless of these numbers we can see from the bar chart in HoursPerWeek Vs Income that most of the percentage count between 36-40 hours earn less income that is <=50K. **In Fig 10** histogram shows the data distribution and since the data is mostly skewed to the left it was interpretable as a stacked bar chart. Other charts have been tried but due to the distribution it was not easy to interpret. Secondly, the hours can be made into bins and compared with income.

9.Capital gain

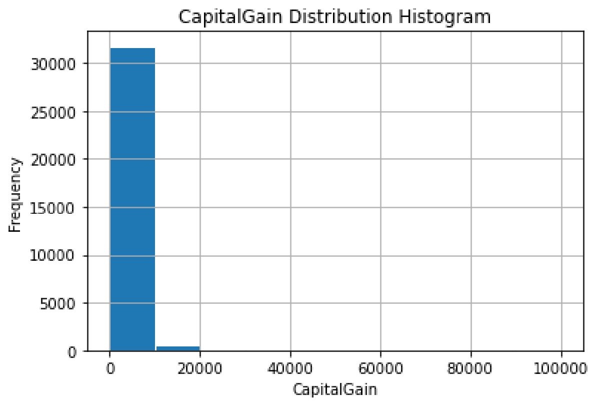


Fig 12

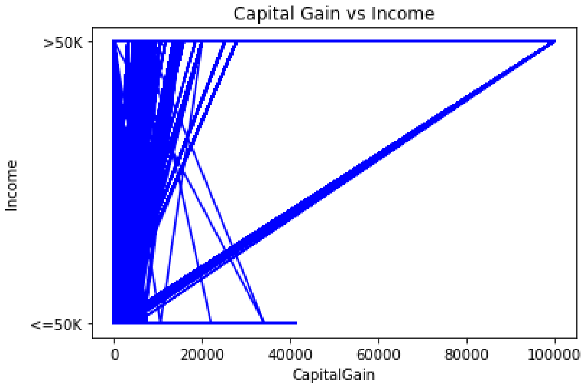


Fig 13

From **Fig 12** we can see that Capital Gain doesn't have a normal distribution , it is positively skewed, people do not really have a capital Gain so this doesn't affect the Income of the individual. Secondly, in the line chart **Fig 13** we can see that people who have some Capital Gain are mostly earning >50K. The variance in lower capital gain for those earning less than 50K and more capital gain indicating salary >50k is very understandable from the line chart

10. Workclass:

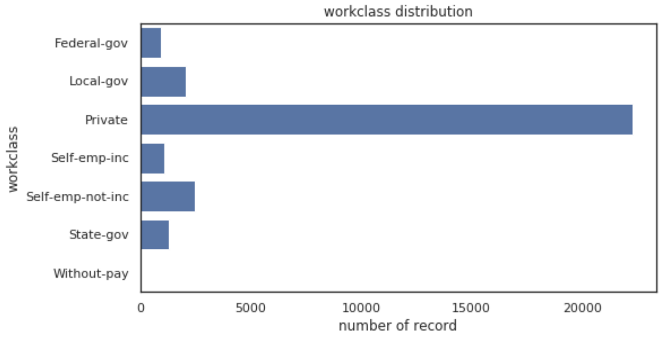


Fig 14

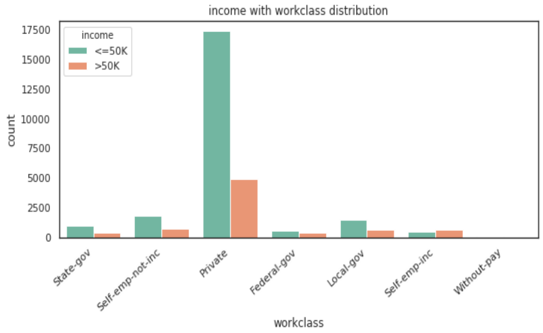


Fig 15

There are a greater number of people working in the private sector, as seen in **Fig 14**.

In **Fig 15** the percentage of earning >50K is almost the same in all job groups except for the private sector. “Federal government” is regarded as the most powerful in the public sector, which reasons for the higher chance of making >50K. The term "self-employed incorporated" refers to a person who owns their own business, which is a category with an almost unrestricted earnings potential. As seen in Fig 15 The work class vs income graph all the workers are in private and the income of those > 50K and <= 50K is very high so the private sector is the strongest sector. The grouped bar chart compares and contrasts the amount of people earning either >50K or <=50K very clearly.

2.6 Executive and Systems Report

After completion of the visualizations all team members worked on the system documentation report explaining our steps in achieving our objective. Meanwhile, the team worked on the executive report to create a summary of top 4 important features for presenting our findings and give insights on the data to UVW executives.

3 QUESTIONS

1. What factor or attributes to use and to which attributes we can give more priority?
   There are 14 features in the dataset where 6 of them are continuous and the rest 7 are categorical. We can say that not all the 14 factors are important for predicting the income. We as a team, explored and analyzed each of the attributes individually and ranked them based on the priority that would affect the income more.

2. What kind of Visualizations to use for analyzing the data properly?
   There are different data types that need to be first looked into and decide accordingly which graphs will be best suited for them.
   For categorical data, visualizations like the pie chart and mosaic plots are better to use as these plots represent the categorical data well.
   For continuous data, visualizations like the histogram, box plots, scatterplots and line chats are used.

3. How to use this data to infer and analyze the visualization?
   The data set is highly skewed and not normally distributed. For example the "Capital-Gain" attribute is highly skewed because most of the values have 0 capital gain and it has approximately 29825 of total values from a total of 32538 entries.

4 CONCLUSION

In conclusion, we have studied the data, understood the client requirements and analysed the data to retrieve useful information about income and features that influence income. The analysis and given visual representation of data has helped gain insights about which factors and how strongly these factors affect income.
Overall , looking at this dataset we can conclude and predict that for "Class label" Income', the highest affecting factors are education,sex, age and occupation. Talking about the factors which do not impact the income much is native country. Other factors that do not have a very strong direct relation with income are marital-status, hours per week, capital gain and workclass.
With these discoveries about the data it will help UVW executives to make better informed decisions for their marketing strategies and help promote UVW college so they will reach out to the right individuals' and boost their enrollment.

5 FUTURE WORK

- As the future work we are planning to build prediction models based on this data to predict an indian's income precisely for more detailed marketing. We are planning to create machine learning models for predicting using different classification and regression algorithms, assuming that this particular dataset requires supervised machine learning algorithms.
- In the future we also plan to work on the prediction model with multivariate data, for example we predict income of males between 50-60 years working in the private sector.

APPENDIX

A. Exploring the dataset:

-
  ```
  #load important libraries
  import pandas as pd
  import numpy as np
  from collections import Counter
  from matplotlib import pyplot
  ```

-
  ```
  #load data
  df = pd.read_csv('adult.data',header=None, sep="\t", na_values=' ?')
  ```

-
  ```
  # summarize the shape of the dataset
  print(df.shape)
  ```

-
  ```
  #add heading to columns
  df.columns = ["age", "workclass", "fnlwgt", "education", "education-num", "marital-status",
  "occupation", "relationship", "race", "sex", "capital-gain", "capital-loss", "hours-per-week",
  "native-country", "income"]
  df.head()
  print(df)
  print(df.dtypes)
  ```

-
  ```
  #Checking for null values and datatypes for the columns
  df.isin(['?']).sum(axis=0)
  print(df.shape)
  print(df.columns)
  print(df.info())
  ```

-
  ```
  # Replacing null values  for numeric columns  using median
  median = df1['CapitalGain'].median()
  df1['CapitalGain'].fillna(median, inplace=True)
  median2 = df1['HoursPerWeek'].median()
  df1['HoursPerWeek'].fillna(median, inplace=True)
  ```

-
  ```
  # summarize the class distribution of income
  target = df.values[:,-1]
  counter = Counter(target)
  for k,v in counter.items():
          per = v / len(target) * 100
          print('Class=%s, Count=%d, Percentage=%.3f%%' % (k, v, per))
  ```

-
  ```
   #filtering noisy data
  df = df[df["workclass"] != ' ?']
  df = df[df["education"] != ' ?']
  df = df[df["marital-status"] != ' ?']
  df = df[df["occupation"] != ' ?']
  df = df[df["relationship"] != ' ?']
  df = df[df["race"] != ' ?']
  df = df[df["sex"] != ' ?']
  df = df[df["native-country"] != ' ?']
  ```

```
#replacing ‘? ‘  with  null values
df['native-country'] = df['native-country'].replace(' ?',np.nan)
df['workclass'] = df['workclass'].replace(' ?',np.nan)
df['occupation'] = df['occupation'].replace(' ?',np.nan)

#dropping rows with null values
df.dropna(how='any',inplace=True)
```

B. Code for checking distribution of data:

-
  ```
  # select columns with numerical data types
  num_ix = df.select_dtypes(include=['int64', 'float64']).columns
  # select a subset of the dataframe with the chosen columns
  subset = df[num_ix]
  # create a histogram plot of each numeric variable
  subset.hist()
  pyplot.show()
  ```

-
  ```
   # Creating scatter plots for all pairs of variables.
  g = sns.PairGrid(data=df1, hue='income')
  g.map(plt.scatter)
  ```

C. Code for income vs. age

```
#Sorting the values of age
  df_sort =df.sort_values('Age', ascending=True)
#Bining the data for every 10 years
bins = [16.9, 30.0, 40.0, 50.0, 60.0, 70.0, 80.0, 90.0]
labels = [20,30,40,50,60,70, 80]
df_sort['Age']= pd.cut(df_sort['Age'], bins, labels= labels)
#Plotting the Age vs Income
sns.countplot(df['Salary'], palette='coolwarm', hue='Age', data=df_sort).set(title='Age vs Income',
xlabel='Income', ylabel='')
```

D. Code for income vs. workclass

```
df = pd.read_csv("adult.data", header=None, sep=", ")
df.columns = ["age", "workclass", "fnlwgt", "education", "education-num", "marital-status", "occupation",
"relationship", "race", "sex", "capital-gain", "capital-loss", "hours-per-week", "native-country", "income"]
df = df[df["workclass"] != '?']
df = df.drop_duplicates()
df['workclass'] = df['workclass'].replace(' ?',np.nan)
df['income2'] = np.where(df['income'] == '>50K', 0, 1)
df_workclass=df.groupby(["workclass"])["workclass"].count().reset_index(name="number of record")
sns.barplot(x="number of record", y="workclass", data=df_workclass, label="Total", color="b")
```

```
plt.title('workclass distribution')
plt.show()
plt.title('income vs workclass')
sns.set(rc={'figure.figsize':(10,4)})
sns.set_style("white")
g = sns.countplot(x="workclass",hue="income", data=df, palette="Set2",orient="h")
#sns.despine()
g = g.set_xticklabels(g.get_xticklabels(), rotation=40, horizontalalignment='right')
plt.show()
```

E. Code for income vs. education

```
#Grouping the education attributes
df.education = df.education.str.replace('Preschool', 'school')
df.education = df.education.str.replace('1st-4th', 'school')
df.education = df.education.str.replace('5th-6th', 'school')
df.education = df.education.str.replace('7th-8th', 'school')
df.education = df.education.str.replace('9th', 'school')
df.education = df.education.str.replace('10th', 'school')
df.education = df.education.str.replace('11th', 'school')
df.education = df.education.str.replace('12th', 'school')
df.education = df.education.str.replace('HS-grad', 'HS')
df.education = df.education.str.replace('Assoc-voc', 'Highier_ed')
df.education = df.education.str.replace('Assoc-acdm', 'Highier_ed')
df.education = df.education.str.replace('Prof-school', 'Prof-school')
df.education = df.education.str.replace('Some-college', 'Highier_ed')
df.education = df.education.str.replace('Bachelors', 'Bachelor')
df.education = df.education.str.replace('Masters', 'MSc')
df.education = df.education.str.replace('Doctorate', 'Doc')

#Dividing data frame among less or more than 50k
s='<=50K'
poor_df=df.loc[df['Salary'].str.lower().str.contains(s, case=False)]
poor_df['Salary']=0
r='>50K'
rich_df=df.loc[df['Salary'].str.lower().str.contains(r, case=False)]
rich_df['Salary']=1
#Grouping the values by income
poor_gr = poor_df.groupby('education').agg('count')
poor_gr.reset_index()
poor_gr = poor_gr.reset_index()
poor_gr.columns = ['education','Age', 'Workclass', 'fnlwgt'
,'educatin-num','Maritual_status','occupation','relation','ace','Sex','Capital-gain','capital-loss','hours-per-week
','native-country','Salary']# change column names
rich_gr = rich_df.groupby('education').agg('count')
rich_gr.reset_index()
rich_gr = rich_gr.reset_index()
rich_gr.columns = ['education','Age', 'Workclass', 'fnlwgt'
,'educatin-num','Maritual_status','occupation','relation','ace','Sex','Capital-gain','capital-loss','hours-per-week
','native-country','Salary']# change column names
import matplotlib.ticker as ticker
import matplotlib.cm as cm
import matplotlib as mpl
from matplotlib.gridspec import GridSpec

import matplotlib.pyplot as plt
%matplotlib inline

plt.rc('axes', labelsize=15)
plt.rcParams.update({'font.size':13})
poor_labels = poor_gr.education


# Make square figures and axes
plt.figure(1, figsize=(15,15))
the_grid = GridSpec(2, 2)


cmap = plt.get_cmap('Spectral')
colors = [cmap(i) for i in np.linspace(0, 1, 8)]

explode=[.05,.2,.05,.1,.05,.3,.05]
plt.subplot(the_grid[0, 1], aspect=5, title='<=50K ')


poor_pie = plt.pie(poor_gr['Salary'], labels=poor_labels,explode=explode, autopct='%1.1f%%',
pctdistance=0.5,shadow=True, colors=colors, labeldistance=1.1)


plt.axis('equal')

plt.legend( title='Education',loc='upper right', bbox_to_anchor=(1,0,0.4,1))
plt.suptitle('Income vs Education', fontsize=20)


plt.show()


plt.rc('axes', labelsize=15)
plt.rcParams.update({'font.size':13})
rich_labels = rich_gr.education
# Make square figures and axes
plt.figure(1, figsize=(15,15))
the_grid = GridSpec(2, 2)


cmap = plt.get_cmap('Spectral')
colors = [cmap(i) for i in np.linspace(0, 1, 8)]

explode=[.05,.2,.05,.05,.2,.2,.05]
plt.subplot(the_grid[0, 1], aspect=5, title='>50K ')
```

```
poor_pie = plt.pie(rich_gr['Salary'], labels=rich_labels,explode=explode, autopct='%1.1f%%',
pctdistance=0.5,shadow=True, colors=colors, labeldistance=1.1)


plt.axis('equal')

#plt.legend( title='Education',loc='upper right', bbox_to_anchor=(1,0,0.3,1))
#plt.suptitle('Education vs Income', fontsize=15)


plt.show()
```

F. Code for income vs. marital status

```
df = df.groupby(['marital-status','income']).size()
df = df.unstack()
df.plot(kind='bar')
```

G. Code for income vs. occupation

```
df=df.groupby(['occupation','income']).size()
df=df.unstack()
df.plot(kind='bar')
```

H. Code for income vs. relationship

```
pt = df1[['relationship','income']].pivot_table(index='relationship', columns='income',
                                   aggfunc=len, fill_value=0)
pt.plot.bar(stacked=True, edgecolor='white')
plt.ylabel('Count')
plt.title('Relationship Vs Income')
```

I. Code for income vs. sex

```
import pandas as pd
import numpy as np
import statsmodels
import matplotlib.pyplot as plt
from statsmodels.graphics.mosaicplot import mosaic
df = pd.read_csv('adult.data',header=None, sep="\t")
df.columns = ["age", "workclass", "fnlwgt", "education", "education-num", "marital-status", "occupation",
"relationship", "race", "sex", "capital-gain", "capital-loss", "hours-per-week", "native-country", "income"]
df.head()
df_gender=df[['sex', 'income']].groupby("sex").count()
print(df_gender)
print(df_gender.dtypes)
```

```
mosaic(df, ['sex', 'income'],title="Income vs. Sex")
plt.show()
```

J. Code for income vs. capital gain

```
#Line Chart for Capital Gain vs income
Capital = df1['CapitalGain']
income = df1['income']
plt.plot(Capital, income, color='Blue')
plt.xlabel('CapitalGain')
plt.ylabel('Income')
plt.title('Capital Gain vs Income')
plt.show()
```

K. Code for income vs. hours-per-week

```
pt = df1[['HoursPerWeek','income']].pivot_table(index='HoursPerWeek', columns='income',
                                aggfunc=len, fill_value=0)
pt.plot.bar(stacked=True, width=5.00, edgecolor='white')
plt.xticks(np.arange(0, 96, 5))
plt.ylabel('Count')
plt.title('HoursPerWeek Vs Income')
```

L. Code for income vs. native country

```
#load important libraries
import pandas as pd
import numpy as np
import statsmodels
import matplotlib.pyplot as plt

#load data
df = pd.read_csv('adult.data',header=None, sep="\t")
#print(df)

#add heading to columns
df.columns = ["age", "workclass", "fnlwgt", "education", "education-num", "marital-status", "occupation",
"relationship", "race", "sex", "capital-gain", "capital-loss", "hours-per-week", "native-country", "income"]
df.head()
df = df[df["native-country"] != ' ?']
df['native-country'] = df['native-country'].replace(' ?',np.nan)
df.dropna(how='any',inplace=True)
print(df['native-country'].groupby([df['native-country']]).count())
grouped_single = df.groupby(['native-country', 'income'], as_index=False).agg({'age': ['count']})
grouped_single.columns = list(map(".join, grouped_single.columns.values))
grouped_single.columns = grouped_single.columns.to_flat_index()

grouped_single = grouped_single.pivot(index='native-country', columns='income', values='agecount')
#plot bar graph
grouped_single.plot.bar(color=[(240/255,83/255,101/255),
(250/255,188/255,42/255)],figsize=(15,10),width=1)
```

```
df = df[df["native-country"] != ' United-States']
grouped_single = df.groupby(['native-country', 'income'], as_index=False).agg({'age': ['count']})
grouped_single.columns = list(map(".join, grouped_single.columns.values))
grouped_single.columns = grouped_single.columns.to_flat_index()
grouped_single = grouped_single.pivot(index='native-country', columns='income', values='agecount')
#plot bar graph
grouped_single.plot.bar(color=[(240/255,83/255,101/255),
(250/255,188/255,42/255)],figsize=(15,10),width=1)
```