US Adult Income

```
train_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32561 entries, 0 to 32560
Data columns (total 15 columns):
Age               32561 non-null int64
Workclass         32561 non-null object
fnlwgt            32561 non-null int64
Education         32561 non-null object
Education Num     32561 non-null int64
Marital Status    32561 non-null object
Occupation        32561 non-null object
Relationship      32561 non-null object
Race              32561 non-null object
Gender            32561 non-null object
Capital Gain      32561 non-null int64
Capital Loss      32561 non-null int64
Hours Per Week    32561 non-null int64
Native Country    32561 non-null object
Income Bracket    32561 non-null object
dtypes: int64(6), object(9)
memory usage: 3.7+ MB
```

The following dataset has a total of 15 columns and 32561 rows. There are no missing values but there are '?' values present within the dataset. I will be replacing the '?' values to Nan and dropping them from the dataset.

```
In [6]: # There were no "nan" values found in this data other than "?", this will be converted to "nan" and then dropped from t
        train = train_df.replace(' ?', np.nan).dropna()

In [7]: train.drop('fnlwgt', axis=1, inplace=True)
```

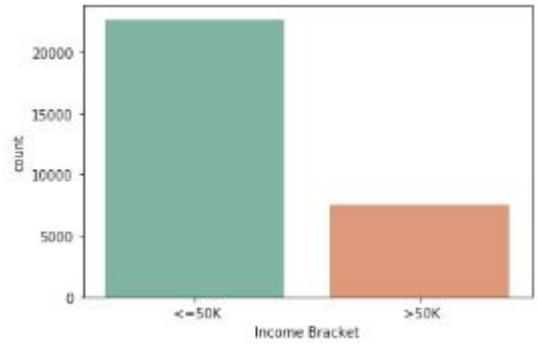
'Fnlwgt' is dropped from the dataset as it has no relationship towards towards the income bracket.

```
In [8]: train.describe()
```

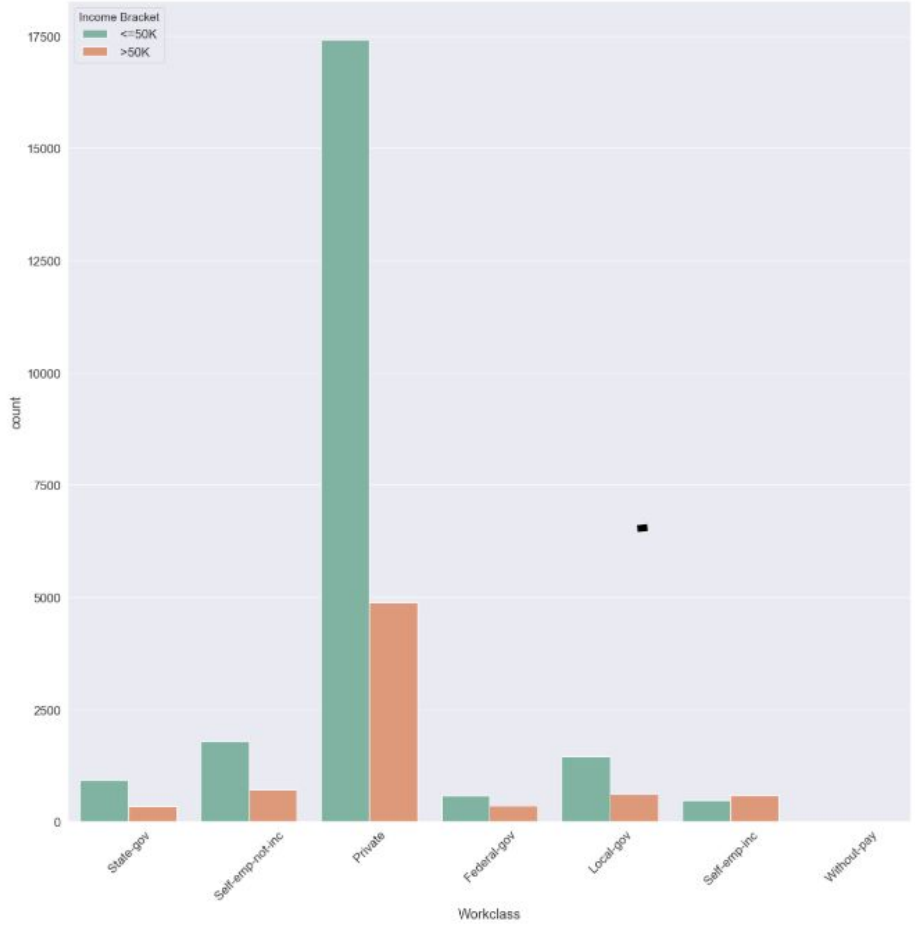
Out[8]:

	Age	Education Num	Capital Gain	Capital Loss	Hours Per Week
count	30162.000000	30162.000000	30162.000000	30162.000000	30162.000000
mean	38.437902	10.121312	1092.007858	88.372489	40.931238
std	13.134665	2.549995	7406.346497	404.298370	11.979984
min	17.000000	1.000000	0.000000	0.000000	1.000000
25%	28.000000	9.000000	0.000000	0.000000	40.000000
50%	37.000000	10.000000	0.000000	0.000000	40.000000
75%	47.000000	13.000000	0.000000	0.000000	45.000000
max	90.000000	16.000000	99999.000000	4356.000000	99.000000

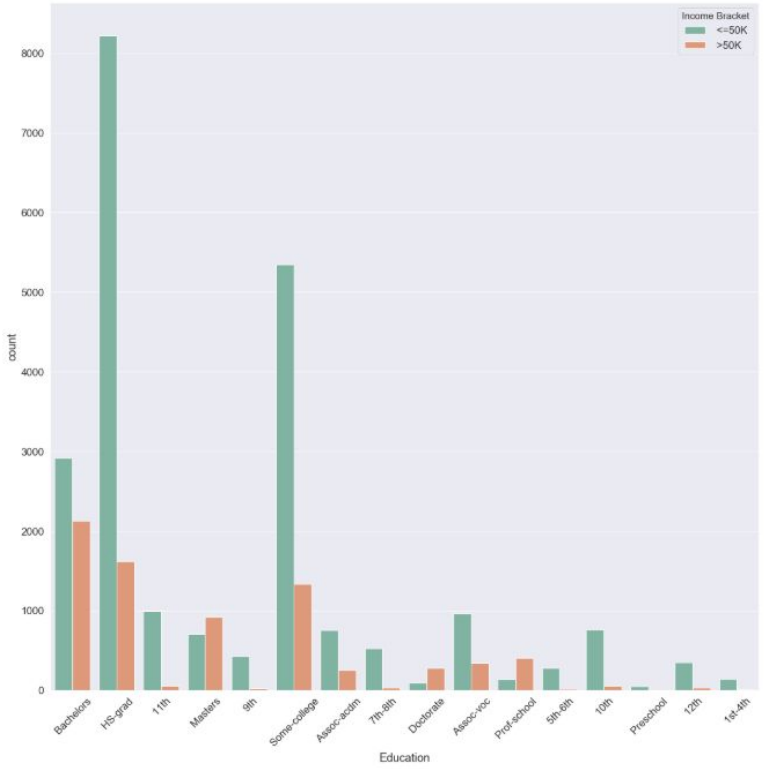
The figure above shows some statistics from the dataset.



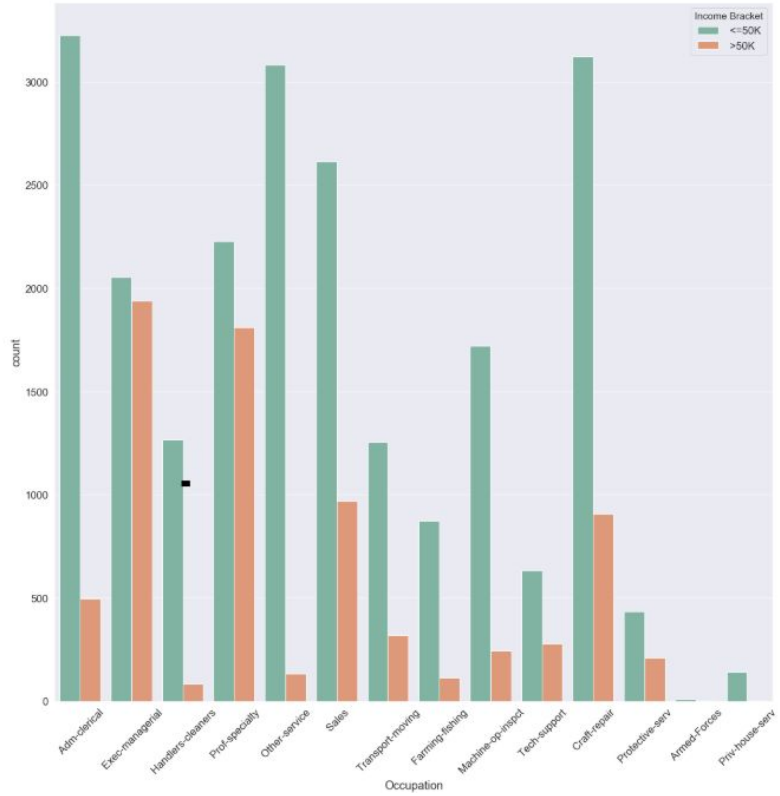
The figure above shows that there are far more individuals making <=50K than in comparison to >50K. Less than half make >50K.



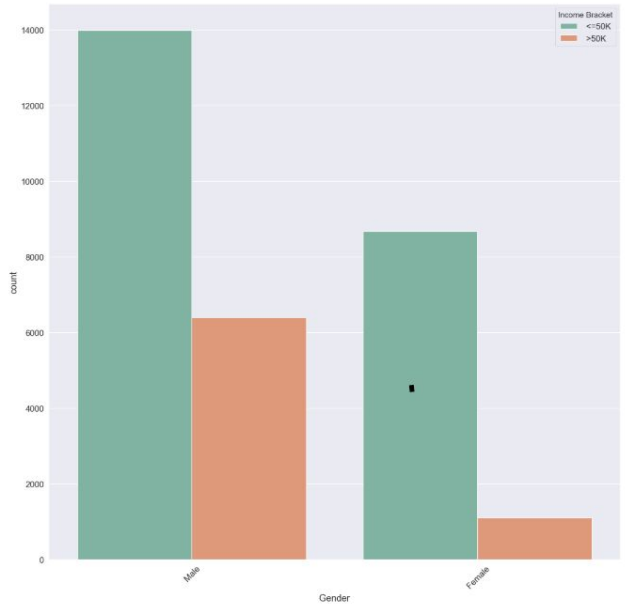
The figure above shows that most people are working in the private sector. Self employed workers have more individuals making over 50K. In all the other working classes, there is a huge gap between <=50K and >50K, most people in these sectors make <=50K.



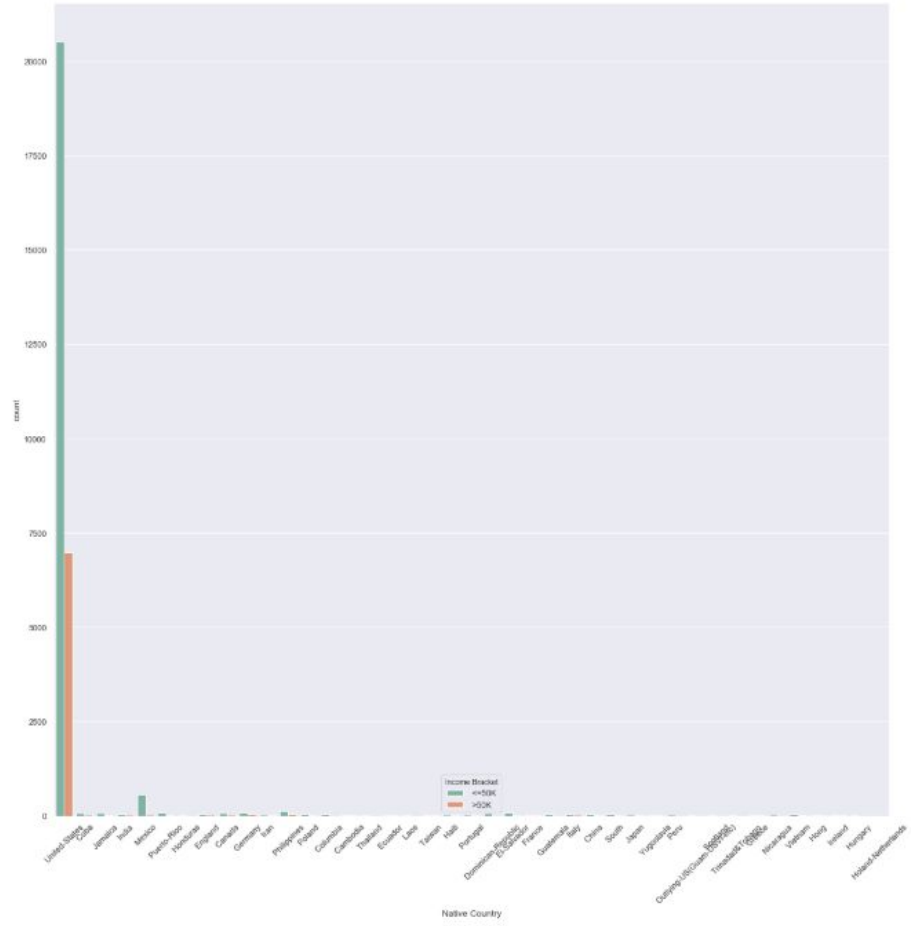
From the figure above, most individuals from this dataset have an education of highschool or more.



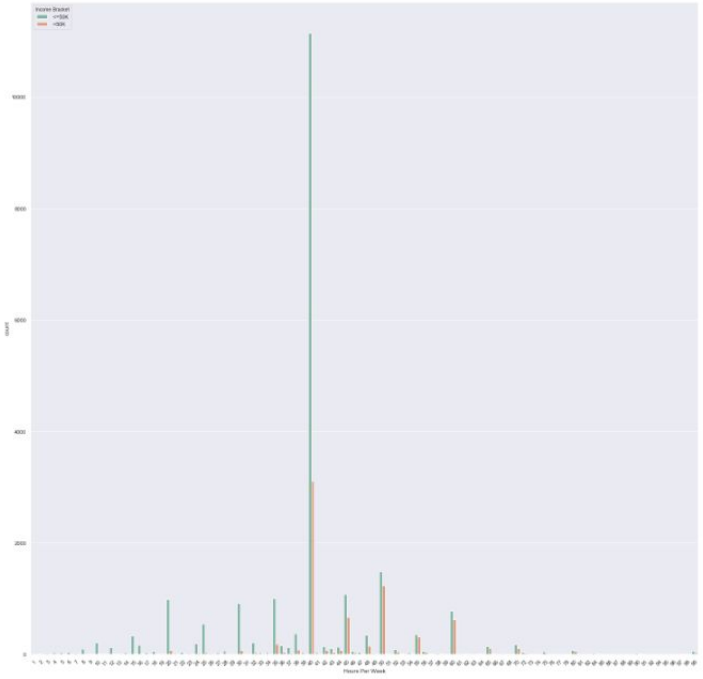
The figure above shows the histogram of the different occupations in the dataset for incomes <=50K and >50K. Handlers-cleaners have the greatest difference of income. Exec-managerial has the least difference as there are far more people in this occupation making >50K.



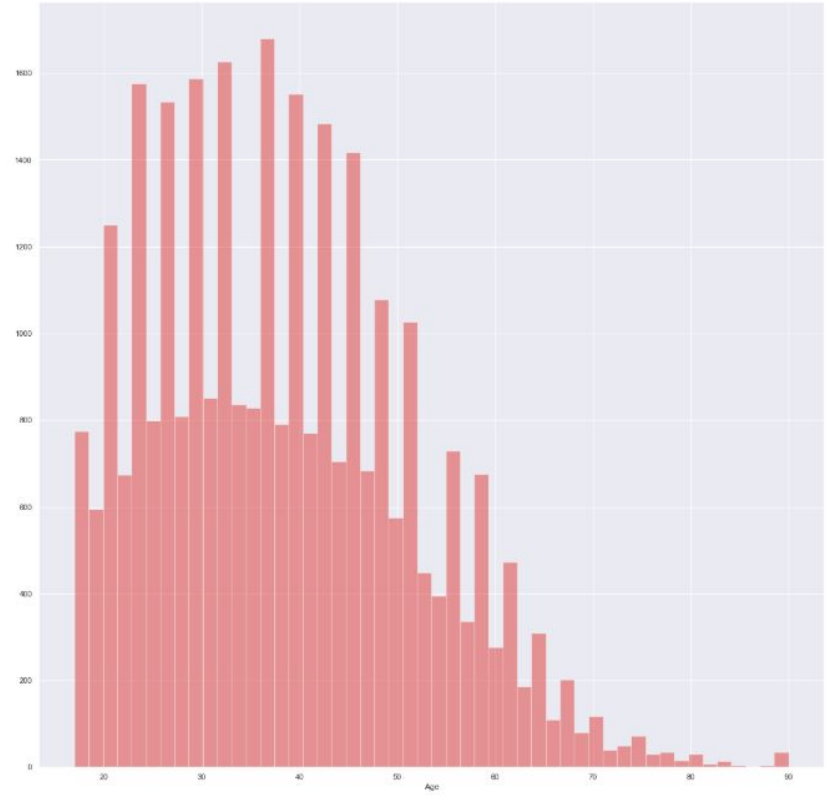
The figure above shows the histogram for males and females in the dataset for incomes <=50K and >50K. There are far less females in ratio making >50K in comparison to males.



The figure above shows the histogram for individuals and their native countires for incomes <=50K and >50K. Most individuals native country is United States and second highest is Mexico.



The figure above shows the histogram of the "Hours Per Week" in the dataset for incomes <=50K and >50K. Most individuals are working 40 hours per week. A large amount of individuals who make >50K seem to work 40 hours or more per week.



The figure above shows the histogram of the different ages in the dataset. There is a wide age gap in this dataset, from 17 years old to 90 years old.

```
# The columns "Income Bracket" needs to be turned into binary (0 and 1) inorder to use it for the correlation matrix ar
y_train = train['Income Bracket']
train['Income Bracket'] = y_train.replace([' <=50K',' >50K' ] , [0,1] )
```

```
corr_matrix = train.corr()
corr_matrix
```

	Age	Education Num	Capital Gain	Capital Loss	Hours Per Week	Income Bracket
Age	1.000000	0.043526	0.080154	0.060165	0.101599	0.241998
Education Num	0.043526	1.000000	0.124416	0.079646	0.152522	0.335286
Capital Gain	0.080154	0.124416	1.000000	-0.032229	0.080432	0.221196
Capital Loss	0.060165	0.079646	-0.032229	1.000000	0.052417	0.150053
Hours Per Week	0.101599	0.152522	0.080432	0.052417	1.000000	0.229480
Income Bracket	0.241998	0.335286	0.221196	0.150053	0.229480	1.000000

I tuned the Income bracket into binary (0 and 1) in order to use it for the correlation matrix