# Exploratory Data Analysis on Adult Income

## Contents

## Abstract

In this Project, we apply Exploratory Data Analysis (EDA) techniques to the adult_income.csv dataset. Here, we examine all the features, rename several columns, handle the missing data, duplicate data, visualize outliers, and handle them, we also visualize several features to extrapolate insights from our data

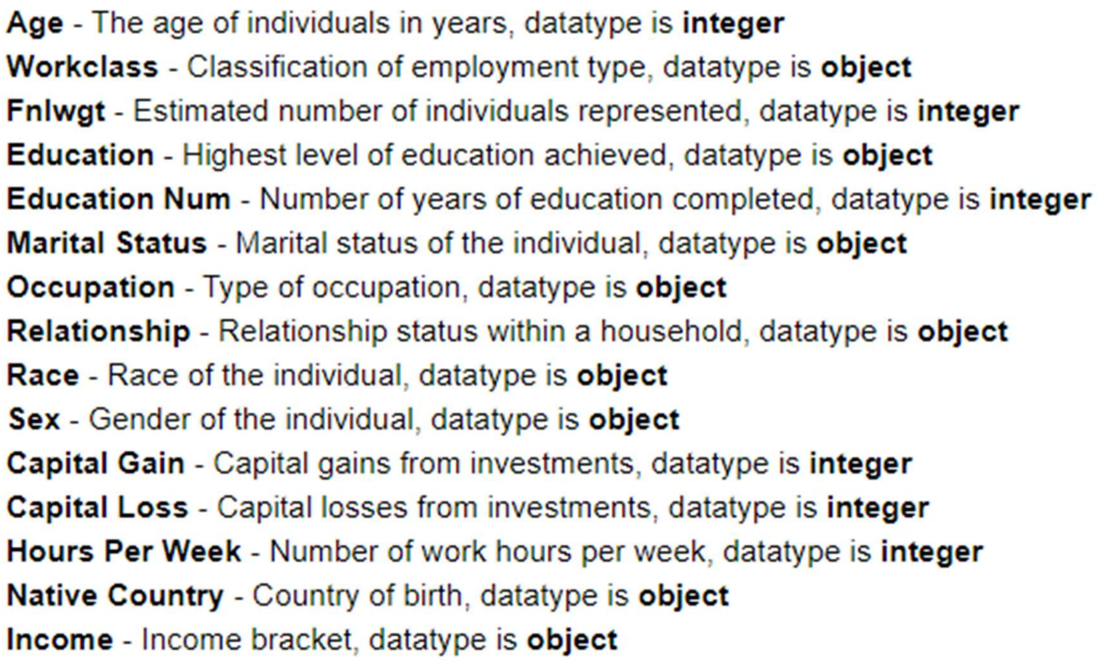
## Introduction

Performing Exploratory Data Analysis is very important. In this step of the Data Science Lifecycle, we Preprocess the data and find insights into all our Features. This step is highly necessary for the following steps of the Data Science Lifecycle. We perform the EDA of our adult_income.csv. We look at each feature and do a detailed analysis of it and the importance of the feature. Our focus will be on our income feature, which is categorical. We clean the dataset, handle missing and duplicate values, visualize outliers using Boxplots, and then handle them. We have also added several visualizations to gather insights from the data.

## Methods

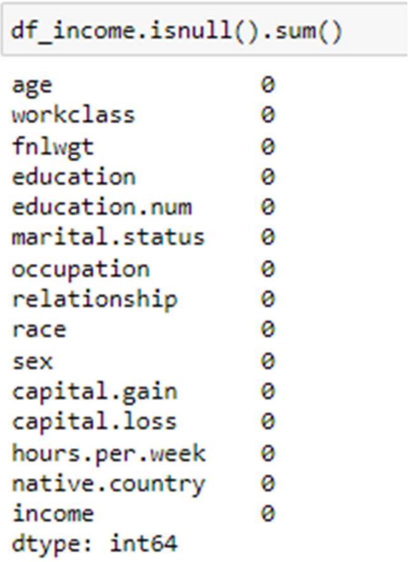
The fundamental goal of this report was to perform an Exploratory data analysis (EDA) in the adult_income.csv dataset. To do this, we use Jupyter Notebook and Python Programming. Let's walk through the entire process.

We import the required libraries, pandas, NumPy, matplotlib, and seaborn. To load the dataset, we use read_csv as our dataset is a CSV file. After this, we print our data frame, including its head and tail. Upon checking the shape of our data frame, we found 32561 rows and 15 columns. Digging deeper to check the datatypes, we see our Dataset Features and what they signify. Look at figure1 to get a better understanding of our features



**Age** - The age of individuals in years, datatype is **integer**
**Workclass** - Classification of employment type, datatype is **object**
**Fnlwgt** - Estimated number of individuals represented, datatype is **integer**
**Education** - Highest level of education achieved, datatype is **object**
**Education Num** - Number of years of education completed, datatype is **integer**
**Marital Status** - Marital status of the individual, datatype is **object**
**Occupation** - Type of occupation, datatype is **object**
**Relationship** - Relationship status within a household, datatype is **object**
**Race** - Race of the individual, datatype is **object**
**Sex** - Gender of the individual, datatype is **object**
**Capital Gain** - Capital gains from investments, datatype is **integer**
**Capital Loss** - Capital losses from investments, datatype is **integer**
**Hours Per Week** - Number of work hours per week, datatype is **integer**
**Native Country** - Country of birth, datatype is **object**
**Income** - Income bracket, datatype is **object**

Figure1: Dataset Description

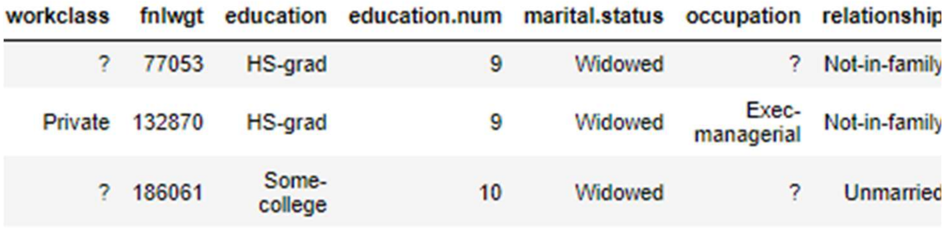


```
df_income.isnull().sum()
```

```
age               0
workclass         0
fnlwgt            0
education         0
education.num     0
marital.status    0
occupation        0
relationship      0
race              0
sex               0
capital.gain      0
capital.loss      0
hours.per.week    0
native.country    0
income            0
dtype: int64
```

Figure: Null Values Count

Our dataset does not appear to have any null values, but looking deeper, we find many '?'; this implies that our NA values have '?' instead. So, we can consider that these records are missing.



| workclass | fnlwgt | education | education.num | marital.status | occupation | relationship |
|---|---|---|---|---|---|---|
| ? | 77053 | HS-grad | 9 | Widowed | ? | Not-in-family |
| Private | 132870 | HS-grad | 9 | Widowed | Exec-managerial | Not-in-family |
| ? | 186061 | Some-college | 10 | Widowed | ? | Unmarried |

Figure: Presence of '?'

To find the Numerical and Categorical Features, we created two lists. Figure 2 shows the distribution of our categorical data, and Figure 3 shows the distribution of our Numerical data.

We observe from both these plots that we have some categorical features with many distinct categories. We observe class imbalance in many of the categorical features.

To find out more about our features and how to handle the missing data. We performed a detailed analysis of each feature, using visualizations.

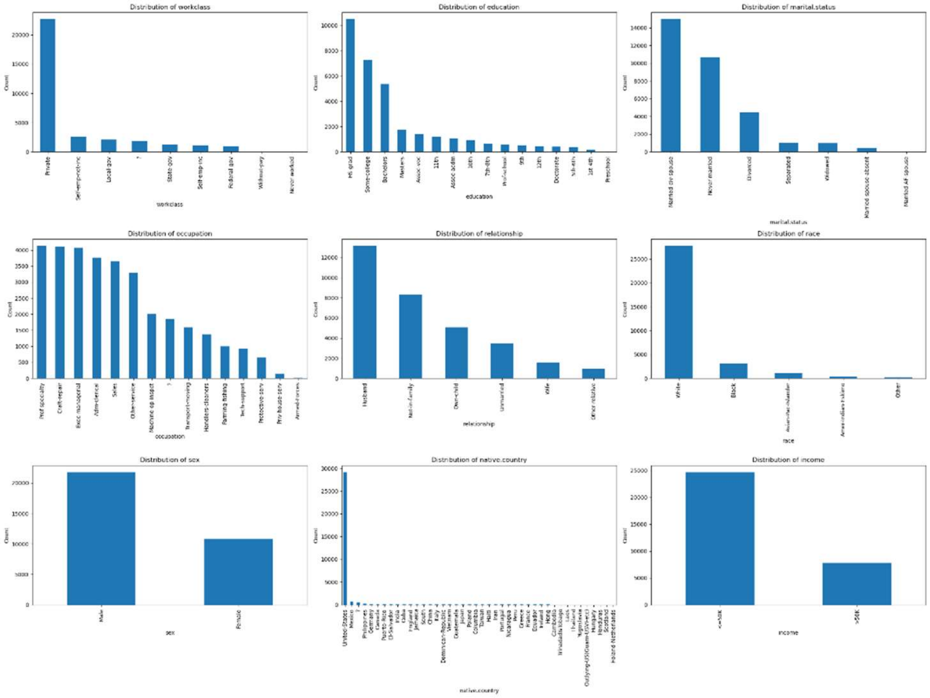


Figure2: Distribution of Categorical Features

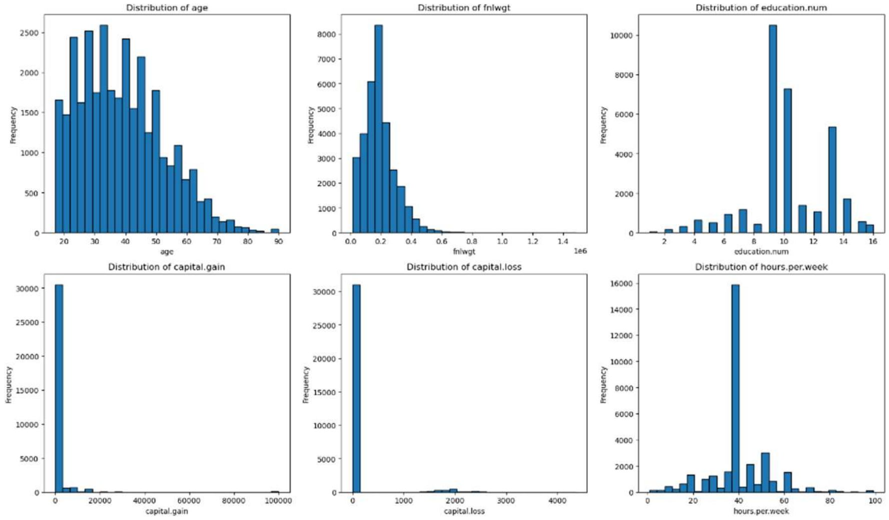


Figure3: Distribution of our Numerical Features

Now, we start data visualization of all our features.

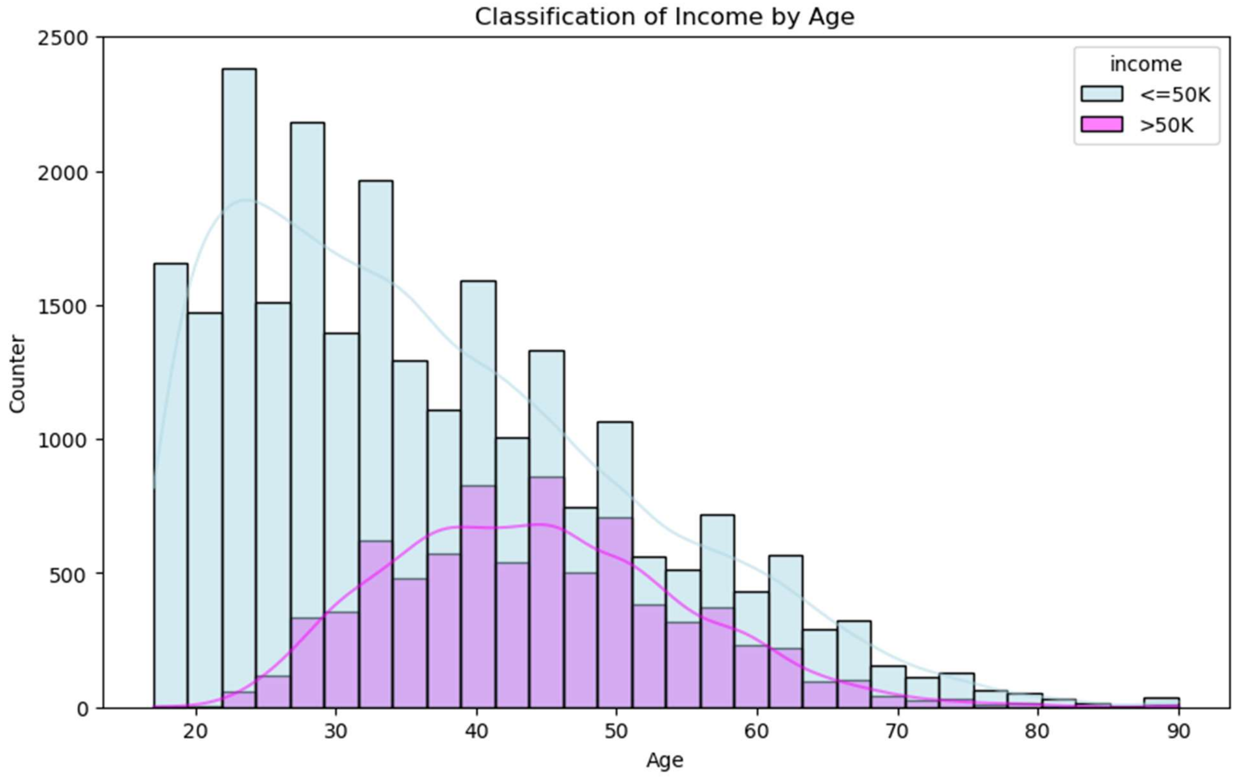


Figure 5: Distribution of Age with respect to Income

In Figure 5, we observe a higher count of those with income lower than 50k than those higher than 50k. Most people who earn less than 50k are in the age bracket 18- 3, and those who earn higher than 50k are in the range 35-55. As age increases, the distribution changes and the count for higher-income individuals increases.

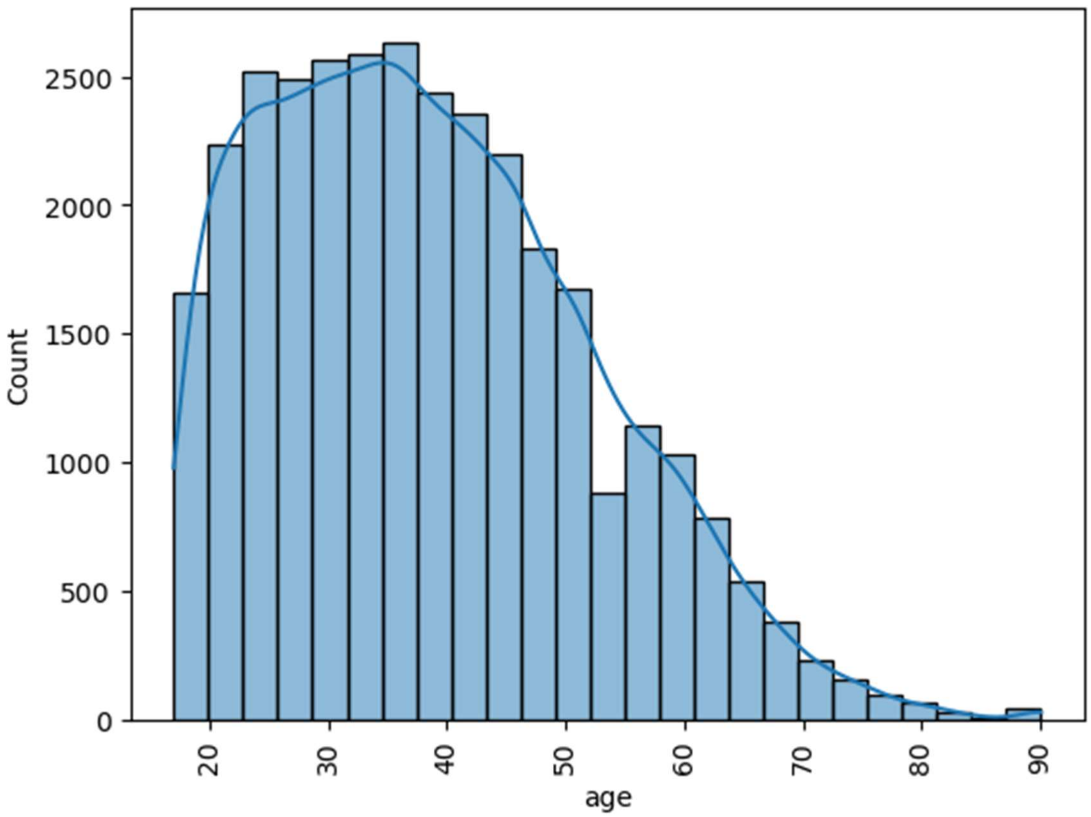


Figure 6: Countplot of Age (Using Seaborn)

We also found some unique observations about age. There are 73 unique ages in our dataset in the range of 17-90. Most of the records are in the range 25-45. Our data's top 5 most common ages are 36, 31, 24, 23, and 35. Observe Figure 6.

Now we look at our Sex Feature, which tells us whether our record is of a Male or Female. We find a higher count of Male records in our dataset. We create a pie chart to compare it with the income and find out the percentage of people who earn higher than 50k vs lower than 50k. Check out Figure 7 and Figure 8.

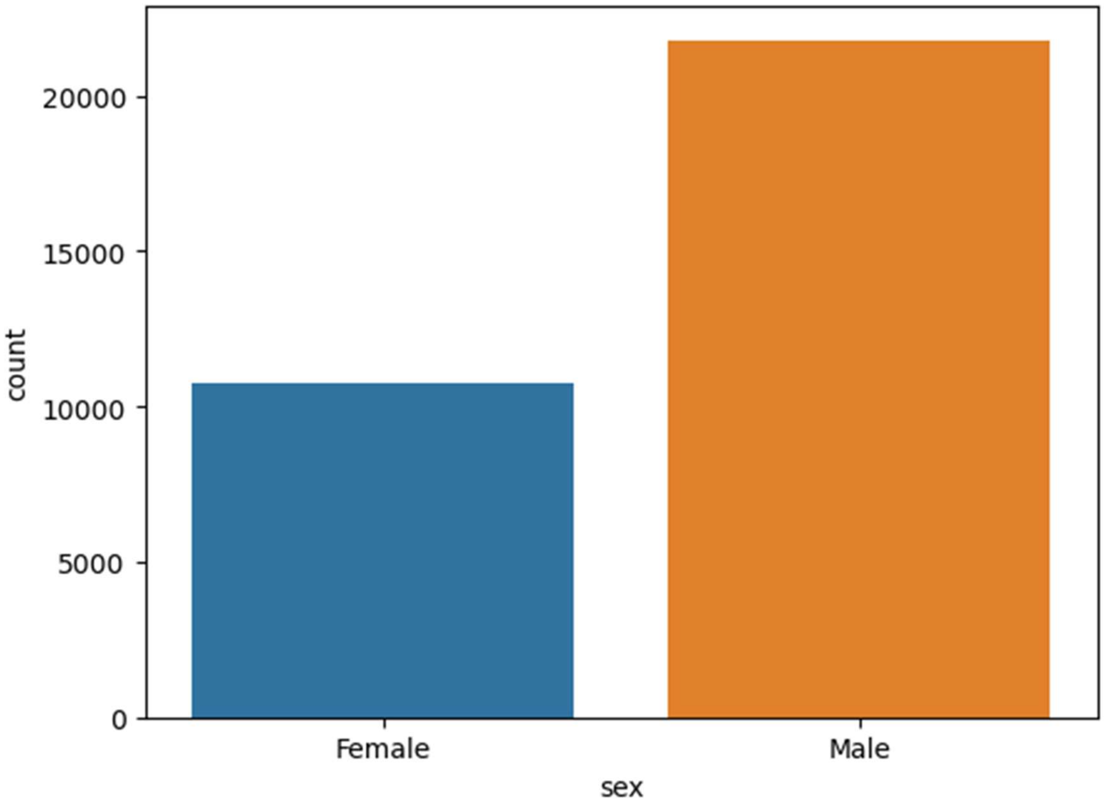


Figure 6: Countplot of Female and Male(Using Seaborn)

The number of males is more than females.

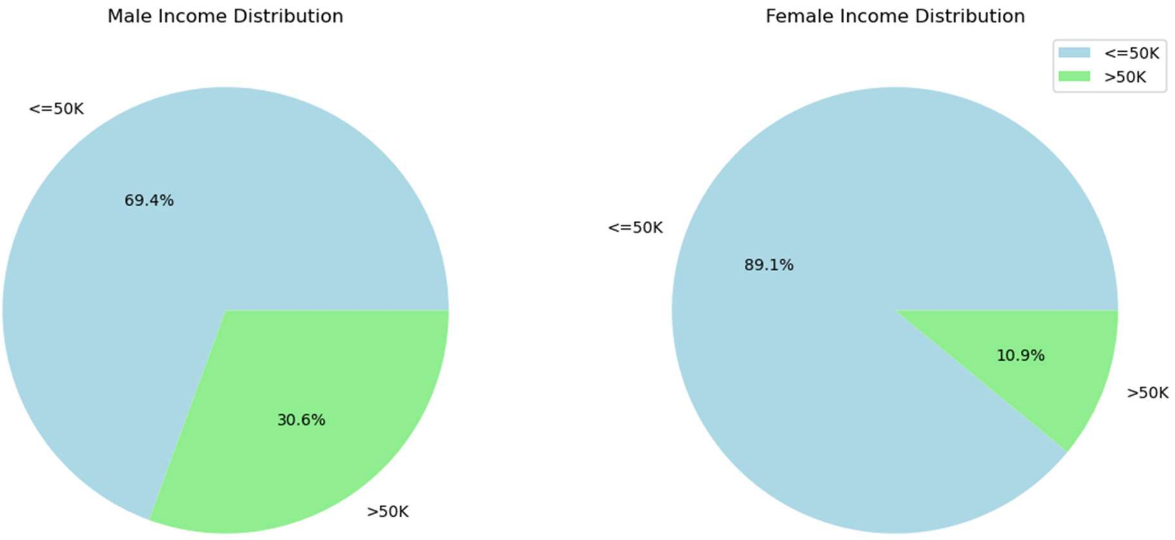


Figure 7: Pie chart for Male Female Distribution with respect to Income

The percentage of those earning less than 50k is greater for males and females, but there is a higher percentage of records for males who earn more than 50k.

Let's analyze each of the features:

Most people work in the Private sector

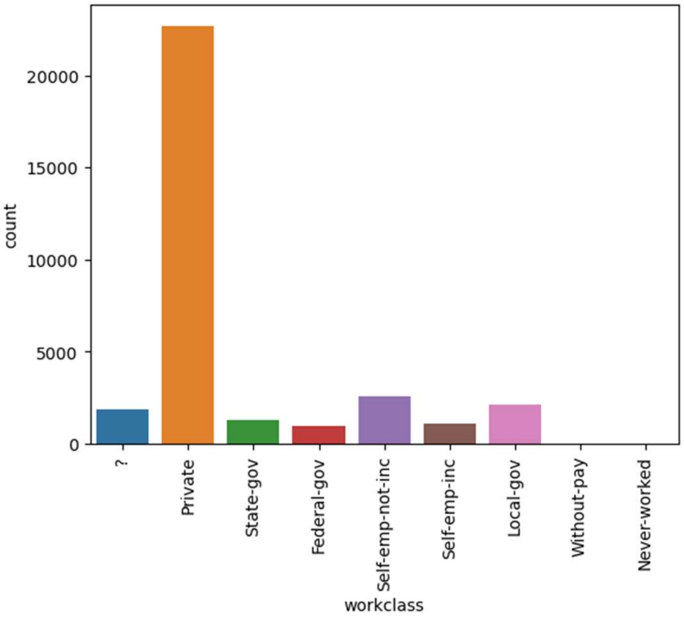


Figure 8: Countplot of Workclass

There is a higher count of HS-Grad – High School Graduates, Some college and Bachelors

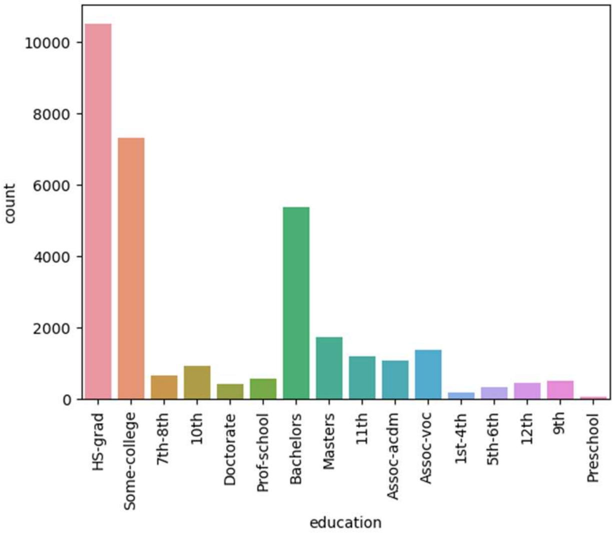


Figure 9: Countplot of Education

A higher count of Married, Never-married, and Divorced people

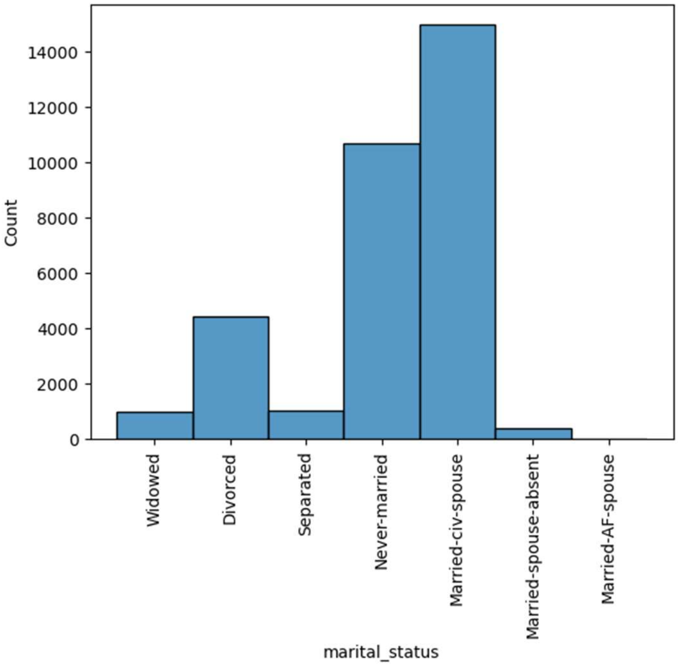


Figure 10: Countplot of Marital Status

Almost All Records are from the USA.

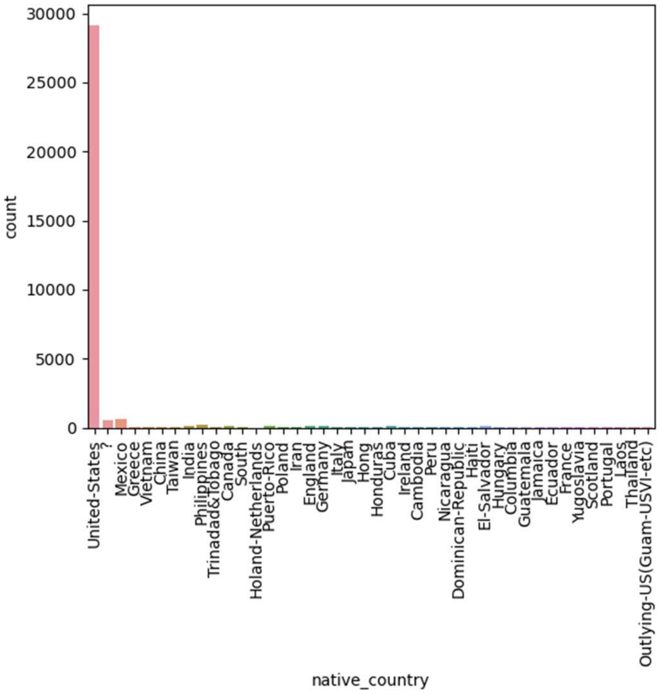
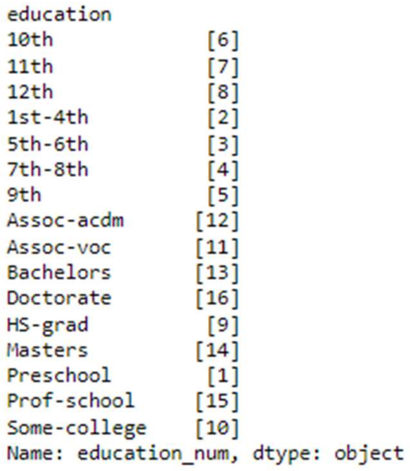


Figure 11: Countplot of Native Country

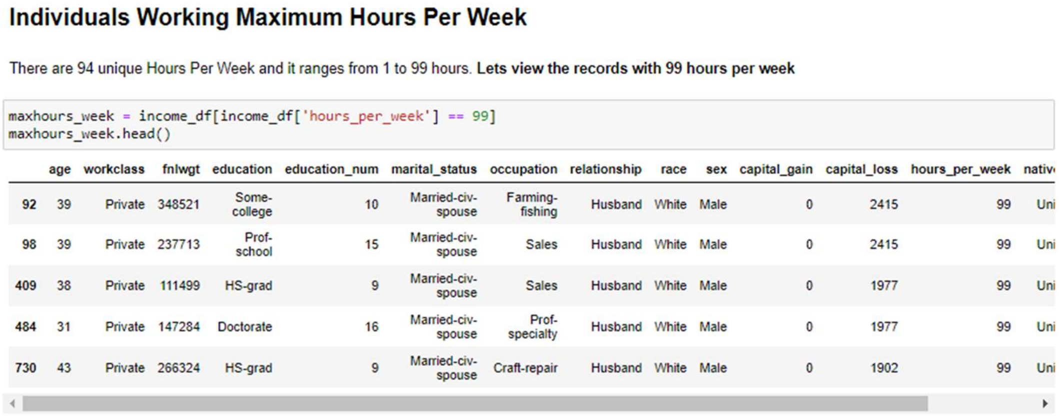
Education and Education Num Mapping



```
education
10th              [6]
11th              [7]
12th              [8]
1st-4th           [2]
5th-6th           [3]
7th-8th           [4]
9th               [5]
Assoc-acdm       [12]
Assoc-voc        [11]
Bachelors        [13]
Doctorate        [16]
HS-grad           [9]
Masters          [14]
Preschool         [1]
Prof-school      [15]
Some-college     [10]
Name: education_num, dtype: object
```

Figure 12: Education and Education Num Mapping

The highest number of hours per week in our data is 99 Hours, so we create a new data frame of these records to dig deeper. 85 people work 99 hours; 66 are males, and 19 are females. Only 25 of the total number of individuals earn more than 50k.



## Individuals Working Maximum Hours Per Week

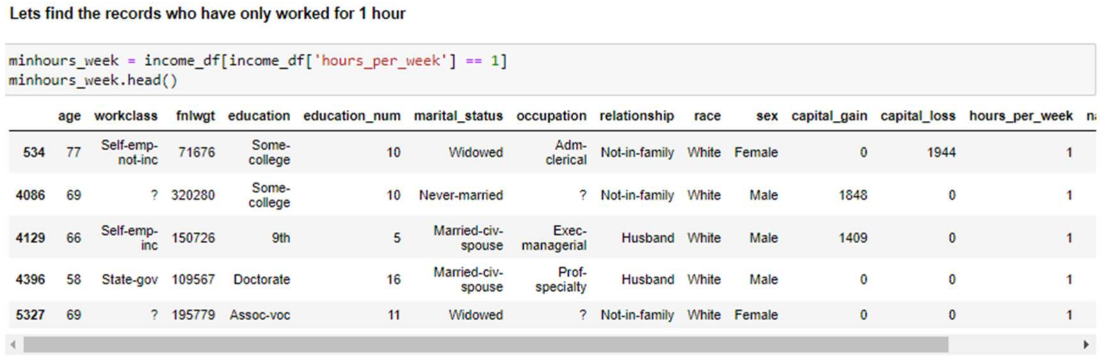
There are 94 unique Hours Per Week and it ranges from 1 to 99 hours. **Lets view the records with 99 hours per week**

```
maxhours_week = income_df[income_df['hours_per_week'] == 99]
maxhours_week.head()
```

| | age | workclass | fnlwgt | education | education_num | marital_status | occupation | relationship | race | sex | capital_gain | capital_loss | hours_per_week | nativ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **92** | 39 | Private | 348521 | Some-college | 10 | Married-civ-spouse | Farming-fishing | Husband | White | Male | 0 | 2415 | 99 | Uni |
| **98** | 39 | Private | 237713 | Prof-school | 15 | Married-civ-spouse | Sales | Husband | White | Male | 0 | 2415 | 99 | Uni |
| **409** | 38 | Private | 111499 | HS-grad | 9 | Married-civ-spouse | Sales | Husband | White | Male | 0 | 1977 | 99 | Uni |
| **484** | 31 | Private | 147284 | Doctorate | 16 | Married-civ-spouse | Prof-specialty | Husband | White | Male | 0 | 1977 | 99 | Uni |
| **730** | 43 | Private | 266324 | HS-grad | 9 | Married-civ-spouse | Craft-repair | Husband | White | Male | 0 | 1902 | 99 | Uni |

Figure 13: Maximum Hours per week dataframe

The people who work for only 1 hour a week. There are only 20 records where the people work for only 1 hour per week. Among these, 18 earn less than 50k, and only 2 earn more than 50k. By digging deeper, we find that both these people are just 19 years old; one is a male, and the other a female.

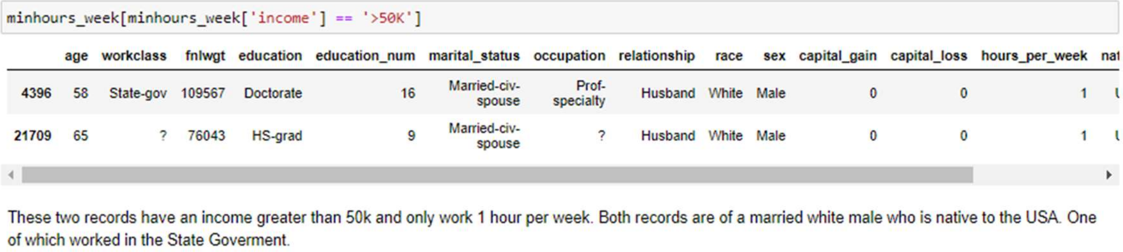


**Lets find the records who have only worked for 1 hour**

```
minhours_week = income_df[income_df['hours_per_week'] == 1]
minhours_week.head()
```

| | age | workclass | fnlwgt | education | education_num | marital_status | occupation | relationship | race | sex | capital_gain | capital_loss | hours_per_week | n |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **534** | 77 | Self-emp-not-inc | 71676 | Some-college | 10 | Widowed | Adm-clerical | Not-in-family | White | Female | 0 | 1944 | 1 | |
| **4086** | 69 | ? | 320280 | Some-college | 10 | Never-married | ? | Not-in-family | White | Male | 1848 | 0 | 1 | |
| **4129** | 66 | Self-emp-inc | 150726 | 9th | 5 | Married-civ-spouse | Exec-managerial | Husband | White | Male | 1409 | 0 | 1 | |
| **4396** | 58 | State-gov | 109567 | Doctorate | 16 | Married-civ-spouse | Prof-specialty | Husband | White | Male | 0 | 0 | 1 | |
| **5327** | 69 | ? | 195779 | Assoc-voc | 11 | Widowed | ? | Not-in-family | White | Female | 0 | 0 | 1 | |

Figure 14: Minimum Hours per week



```
minhours_week[minhours_week['income'] == '>50K']
```

| | age | workclass | fnlwgt | education | education_num | marital_status | occupation | relationship | race | sex | capital_gain | capital_loss | hours_per_week | nat |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **4396** | 58 | State-gov | 109567 | Doctorate | 16 | Married-civ-spouse | Prof-specialty | Husband | White | Male | 0 | 0 | 1 | U |
| **21709** | 65 | ? | 76043 | HS-grad | 9 | Married-civ-spouse | ? | Husband | White | Male | 0 | 0 | 1 | U |

These two records have an income greater than 50k and only work 1 hour per week. Both records are of a married white male who is native to the USA. One of which worked in the State Goverment.

Figure 15: Youngest people to earn more than 50k

We create a violinplot using seaborn to see the capital gain for both income classes. We find that most records have a very low capital gain. But few records in the >50k income class have a high capital gain. We observe this in the violin plot. Our scatterplot also shows us a similar result.

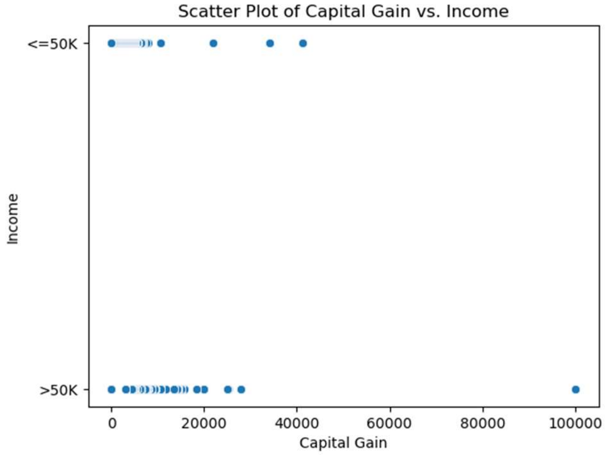


Figure: Scatterplot of Capital Gain vs Income

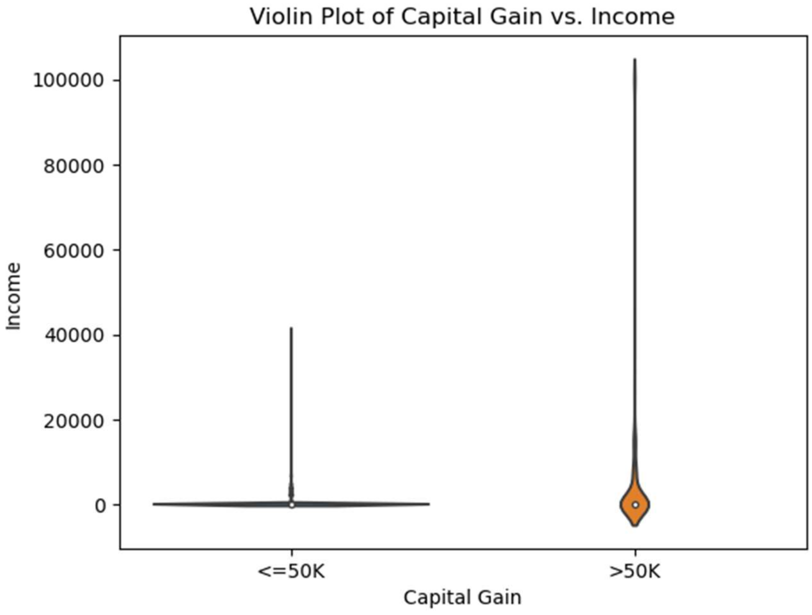


Figure: Violinplot of Capital Gain for both income Groups

In a violinplot of Capital Loss vs Income we observe that the capital loss for the income class <=50k is higher and has a bigger count also.

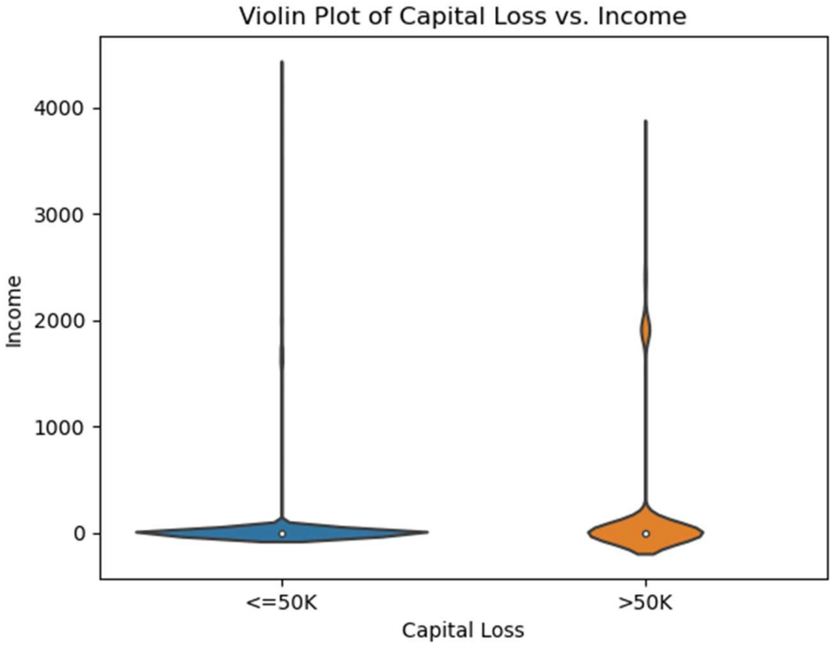
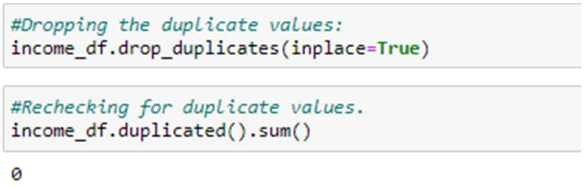


Figure: Violinplot of Capital Gain vs Income

Now that we have seen each feature, we find the number of duplicates in our dataset. We have 24 duplicate values. We remove them by dropping them.



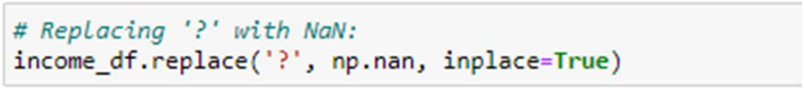
```
#Dropping the duplicate values:
income_df.drop_duplicates(inplace=True)
```

```
#Rechecking for duplicate values.
income_df.duplicated().sum()
```

```
0
```

Figure 16: Dropping Duplicates

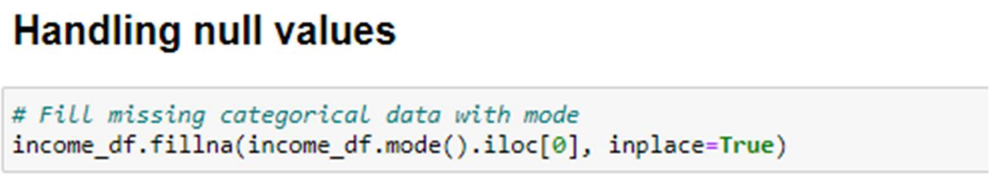
Now, we start analyzing the null values in our dataset. At first, it appears as if there are no null values. Upon closer inspection, we find that we have '?' instead of null values. We replace them with NA



```
# Replacing '?' with NaN:
income_df.replace('?', np.nan, inplace=True)
```

Figure: Replacing '?' with NA

We can notice that the missing values are more for the lower-income (<=50k) class. There does not seem to be any reason why this data can be missing. Records for both classes are missing; the lower income class has more missing values, but also, the distribution of our data is such that we have more records for <=50k compared to greater than 50k. As it is missing completely at random, we handle them by imputation with mode



## Handling null values

```
# Fill missing categorical data with mode
income_df.fillna(income_df.mode().iloc[0], inplace=True)
```

Figure 17: Handling Null Values by Mode Imputation

We then find the correlation heatmap; in this, we notice no highly correlated features. Here is what we found:

- capital_gain and education_num have a weak positive correlation of 0.12
- similary education_num and hours_per_week have a weak positive correlation of 0.15
- There's a small positive correlation (0.078), indicating that older individuals might have slightly higher capital gains

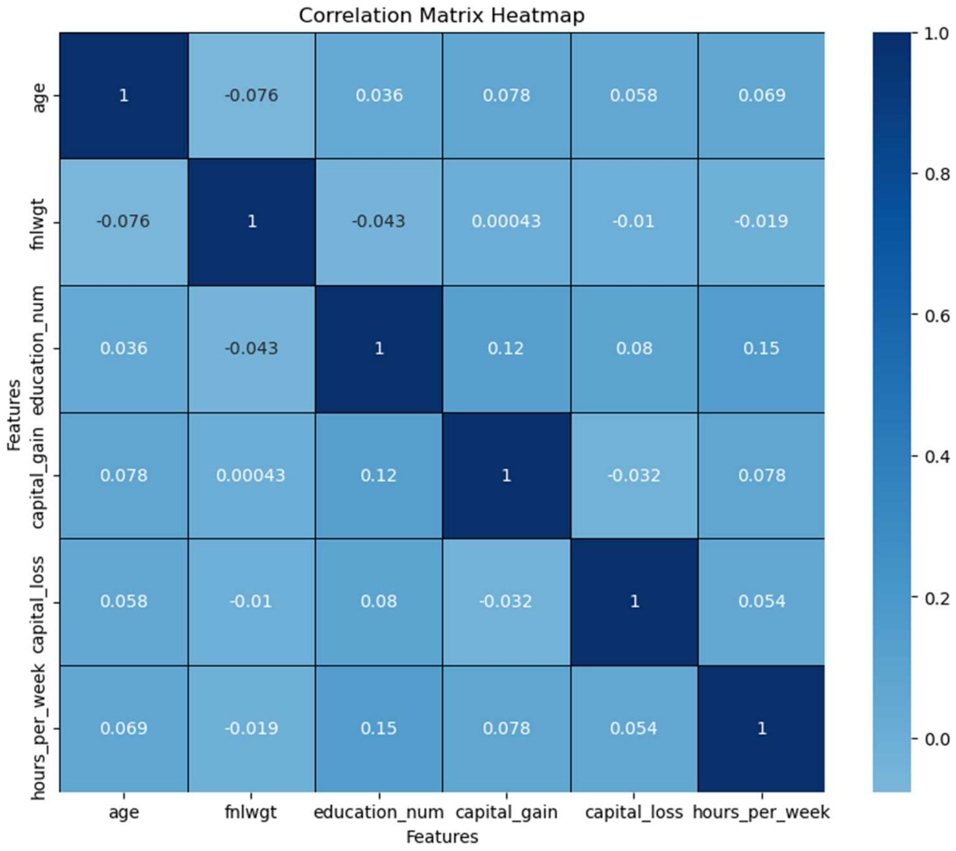


Figure 18: Correlation Heatmap

A Boxplot for income with respect to education number; from this, we can understand that most people with an income greater than 50k have an education of at least 10th. There are some outliers where an individual with a lower education has a high income.

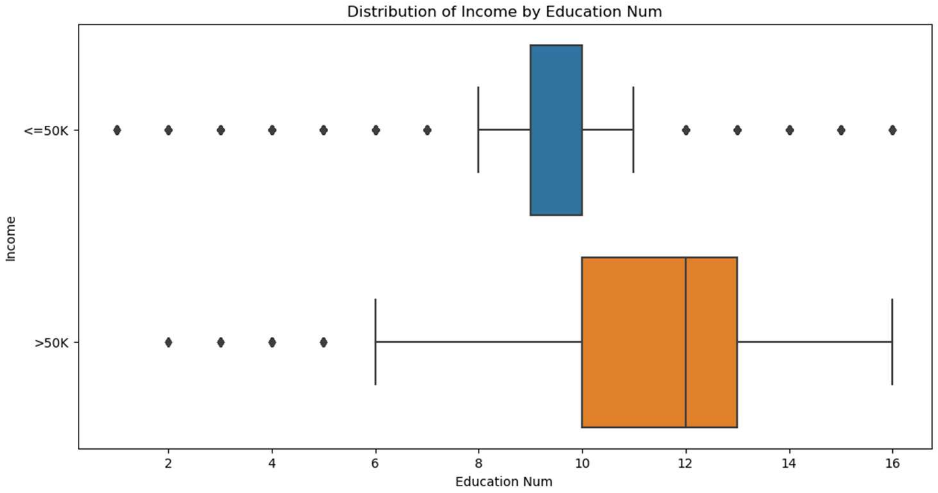


Figure 19: Distribution of Income with Education Num

From this, we can observe

- For individuals earning <=50K, the education_num values are spread across lower educational levels (ranging from Preschool [1] to some college [10]).
- For individuals earning >50K, the education_num values are concentrated around higher educational levels (ranging from some college [10] to Doctorate [16]).

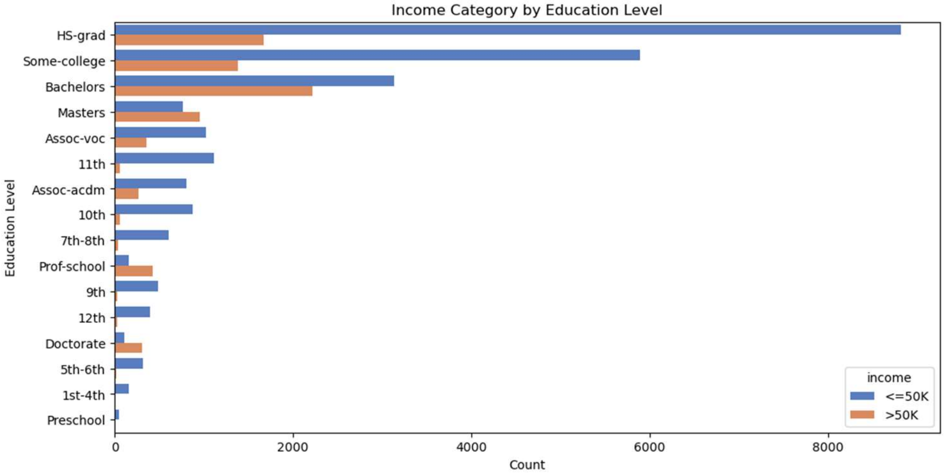


Figure 20: Income category with Education Level

From Figure 20, we observe that our data has most records from HS-grad, but we also observe that with higher education like Masters, Prof School, and Doctorate, the count for income >50k is more. Similarly, those with lower education, like Preschool, 5th-6th, 1st-4th, 7th-8th, 10th, 9th, 11th, Assoc-acdm, and 12th, have higher counts of income <=50k.

Thus, we can observe the importance of education with respect to income

A boxplot of all our Numerical Features to find the outliers. We observe that only fnlgwt and capital gain have outliers. We ignore the outliers in Capital Gain because most records have a low capital gain, which is why it appears to have outliers.
To handle the outliers in fnlgwt we find our Q1, Q3 and then find our Interquartile range(IQR)

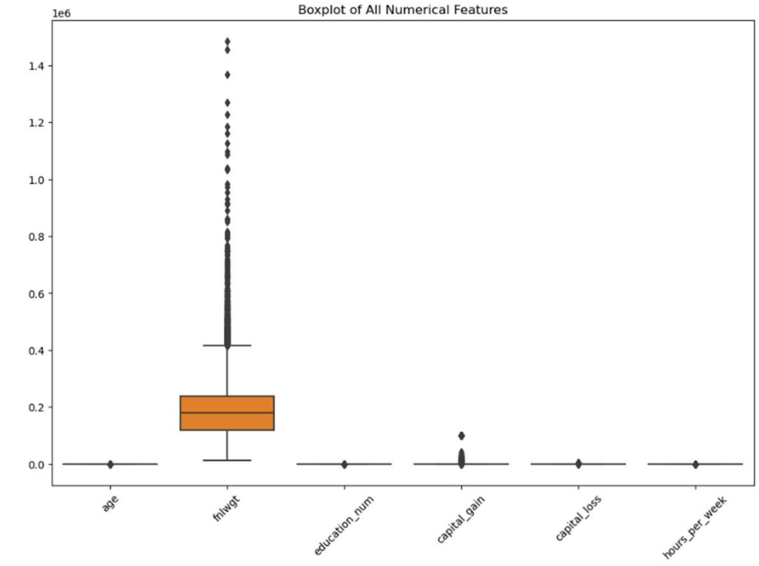


Figure 21: Boxplot of all Numerical Features

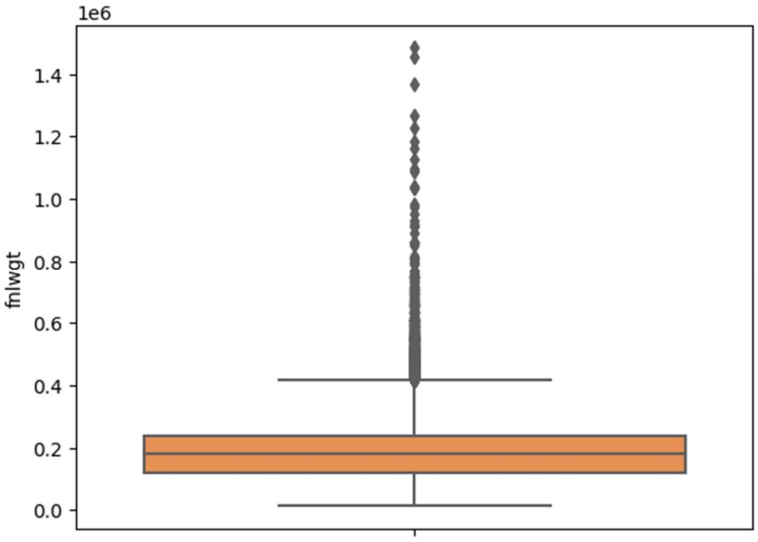
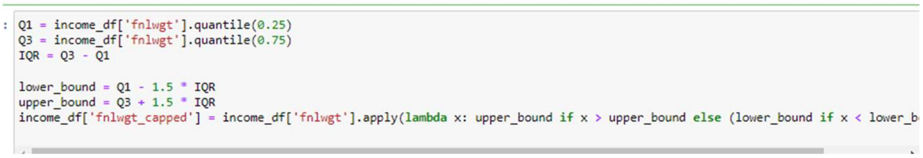


Figure 22: Boxplot of fnlgwt



```
Q1 = income_df['fnlwgt'].quantile(0.25)
Q3 = income_df['fnlwgt'].quantile(0.75)
IQR = Q3 - Q1

lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
income_df['fnlwgt_capped'] = income_df['fnlwgt'].apply(lambda x: upper_bound if x > upper_bound else (lower_bound if x < lower_b
```

Figure 23: Handling the outliers

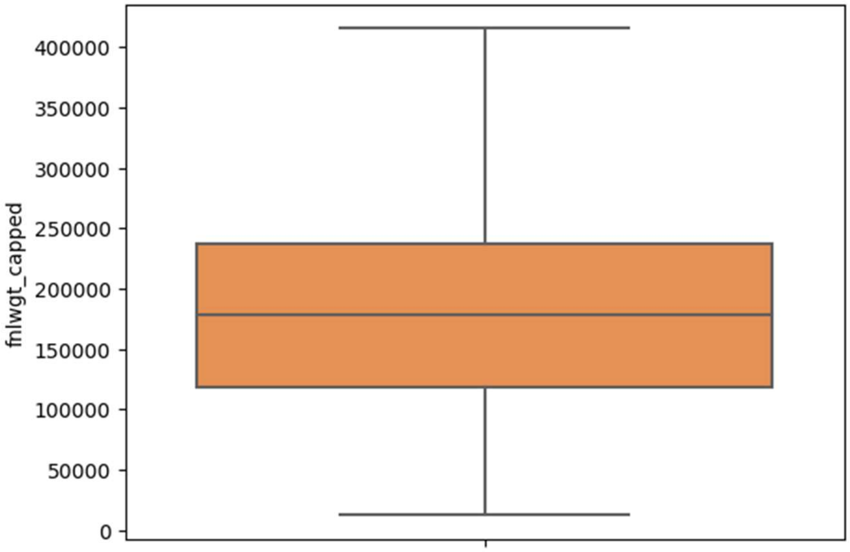


Figure 24: After handling outliers

After handling the outliers and creating the feature ‘fnlgwt’ capped, we drop ‘fnlgwt’ as it is no longer necessary.

## Results

We discovered several insights while performing EDA on our data. Here is a summary of our observations

- There are several categorical and numerical features.
- There are many features where a single category predominates – like Native Country, Race, Sex, Marital Status, and Relationship
- There is a major class imbalance for Income; this will affect any models we create in the future.
- In Education Feature, we see a higher count of high school graduates; further analysis shows that the count of >50k increases for higher education levels.
- We find outliers in our Hours per Week feature. There are only 20 records of people who work only 1 hour per week; among them, we see only two records who earn more than 50k and their age is just 19. We also found out the number of people who work more than 99 hours per week.
- No features have any strong positive or Negative Correlation
- Capital gain and loss for most records are very low. Hence, they appear to have outliers. But the capital gain is more in ‘>50k’ records, and the capital loss is more in the ‘<=50k’ records.

## Conclusions and Future Work

In this activity, we completed a detailed EDA for our adult_income.csv. We start by observing each feature and fixing some of its column names. We gained clear insights after a detailed analysis of each feature using sufficient visualizations. Furthermore, we imputed the missing data with mode imputation and dropped the duplicate data. Ultimately, we find our outliers and cap them between the lower and upper bound. During this activity, we made some very interesting insights, which I've summarized above.

To continue this project, we can create a classification model using a Random Search Classifier, the KNN Classifier, and other classification algorithms. With the help of our EDA, we have found the important features. Moreover, our EDA gives us a clear picture of how to encode our categorical features. If our model does not predict accurately, we can try some Hyper Parameter Tuning Techniques like Grid Search CV

## References

Pmarcelino. (2022, April 30). Comprehensive data exploration with python. Kaggle. https://www.kaggle.com/pmarcelino/comprehensive-data-exploration-with-python

Adult. UCI Machine Learning Repository. (n.d.-b). https://archive.ics.uci.edu/ml/datasets/adult

Becker,Barry and Kohavi,Ronny. (1996). Adult. UCI Machine Learning Repository. https://doi.org/10.24432/C5XW20.