ANALYSIS ADULT DATA SET

EXPLORATORY DATA ANALYSIS

S15355 | K.K.D.S.N.GUNATHILAKA | ST3011- TAKE HOME ASSIGNMENT

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Welcome to the exploratory descriptive analysis of the Census Income dataset, also referred to as the Adult dataset, sourced from the UCI Machine Learning Repository. This dataset, extracted by Barry Becker from the 1994 Census database, serves as a valuable resource for predicting whether an individual's income exceeds $50,000 per year based on various census-related features.

The primary objective of this analysis is to gain insights into the socioeconomic factors that influence income levels, as well as to identify patterns and trends within the dataset. By leveraging statistical and visual exploration techniques, we aim to uncover meaningful relationships, highlight significant variables, and provide a comprehensive understanding of the data.

EXPLORATORY DATA ANALYSIS

☐ ATTRIBUTE INFORMATION: ADULT.CSV

Predictor variables considered for the analysis

Table 1: Predictor variables Descriptions

Qualitative/ Quantitative	Variable name	Description
Qualitative	Workclass	Work Class of the individual (Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked)
	Education	Education States of the individual (Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool)
	Education_num	Education Number of the individual
	Marital_status	Marital Status of the individual (Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse)
	Occupation	Occupation of the individual (Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces)
	Relationship	Relationship Status of the individual (Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried)
	Sex	Gender of the individual (Female / Male)
	Native_country	Native Country of the individual (United States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland,

		France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinidad &Tobago, Peru, Hong, Holand-Netherlands)
Quantitative	Age	The age of the individual in years
	Fnlwgt	Final weight of the individual. An estimate of the number of individuals in the population with the same demographics as this individual.
	Capital_gain	Capital gain in the previous years
	Capital_loss	Capital loss in the previous years
	Hours_per_week	Hours worked per week

Response variable/ Target variable of the analysis

Income: This is a Quantitative variable. This binary label salary encodes whether an individual earns more or less than $50,000. We refer to those who earn more than $50,000 as ">50K" and those who don't as "<=50K"

☐ DATA CLEANING

1) Handling Missing Values

In the Adult dataset, there are instances where unusual observations are represented by question marks (?). To address this, these question marks are replaced with NaN values. Upon inspecting missing values in the dataset, it is observed that the "Workclass," "Occupation," and "Native_country" variables, all of which are categorical, contain missing values.

To handle the missing data in these categorical variables, the approach chosen is to replace them with the mode value of each respective variable. The mode represents the most frequently occurring value in a categorical variable. Therefore, for "Workclass," "Occupation," and "Native_country," the mode values are imputed to ensure completeness in the dataset.

```
Age              False
Workclass         True
Fnlgwt           False
Education        False
Education_num    False
Marital_status   False
Occupation        True
Relationship     False
Race             False
Sex              False
Capital_gain     False
Capital_loss     False
Hours_per_week   False
Native_country    True
Income           False
dtype: bool
```

Figure 1: Checking Missing Values

2) Checking Duplicated Values

When we check the duplicated values of the data set, we can see 24 duplicated in here and we remove duplicated values from the data set.

```
df.duplicated().sum() # Number of duplicated rows
```

```
24
```

Figure 2:Checking Duplicated Rows

3) Checking Unique Values of the Variables

After handling missing values and the duplicated values when we check the unique values of the variables we can see there is nothing to change in it.

```
print(df.Workclass.unique())
```

```
[' State-gov' ' Self-emp-not-inc' ' Private' ' Federal-gov' ' Local-gov'
 ' Self-emp-inc' ' Without-pay' ' Never-worked']
```

```
print(df.Education.unique())
```

```
[' Bachelors' ' HS-grad' ' 11th' ' Masters' ' 9th' ' Some-college'
 ' Assoc-acdm' ' Assoc-voc' ' 7th-8th' ' Doctorate' ' Prof-school'
 ' 5th-6th' ' 10th' ' 1st-4th' ' Preschool' ' 12th']
```

```
print(df.Marital_status.unique())
```

```
[' Never-married' ' Married-civ-spouse' ' Divorced'
 ' Married-spouse-absent' ' Separated' ' Married-AF-spouse' ' Widowed']
```

```
print(df.Occupation.unique())
```

```
[' Adm-clerical' ' Exec-managerial' ' Handlers-cleaners' ' Prof-specialty'
 ' Other-service' ' Sales' ' Craft-repair' ' Transport-moving'
 ' Farming-fishing' ' Machine-op-inspct' ' Tech-support'
 ' Protective-serv' ' Armed-Forces' ' Priv-house-serv']
```

```
print(df.Relationship.unique())
```

```
[' Not-in-family' ' Husband' ' Wife' ' Own-child' ' Unmarried'
 ' Other-relative']
```

```
print(df.Native_country.unique())
```

```
[' United-States' ' Cuba' ' Jamaica' ' India' ' Mexico' ' South'
 ' Puerto-Rico' ' Honduras' ' England' ' Canada' ' Germany' ' Iran'
```

```
print(df.Income.unique())
```

```
[' <=50K' ' >50K']
```

Figure 3: Checking Unique values of the variables

☐ UNIVARIATE ANALYSIS

Histogram Plots for Continuous Variables

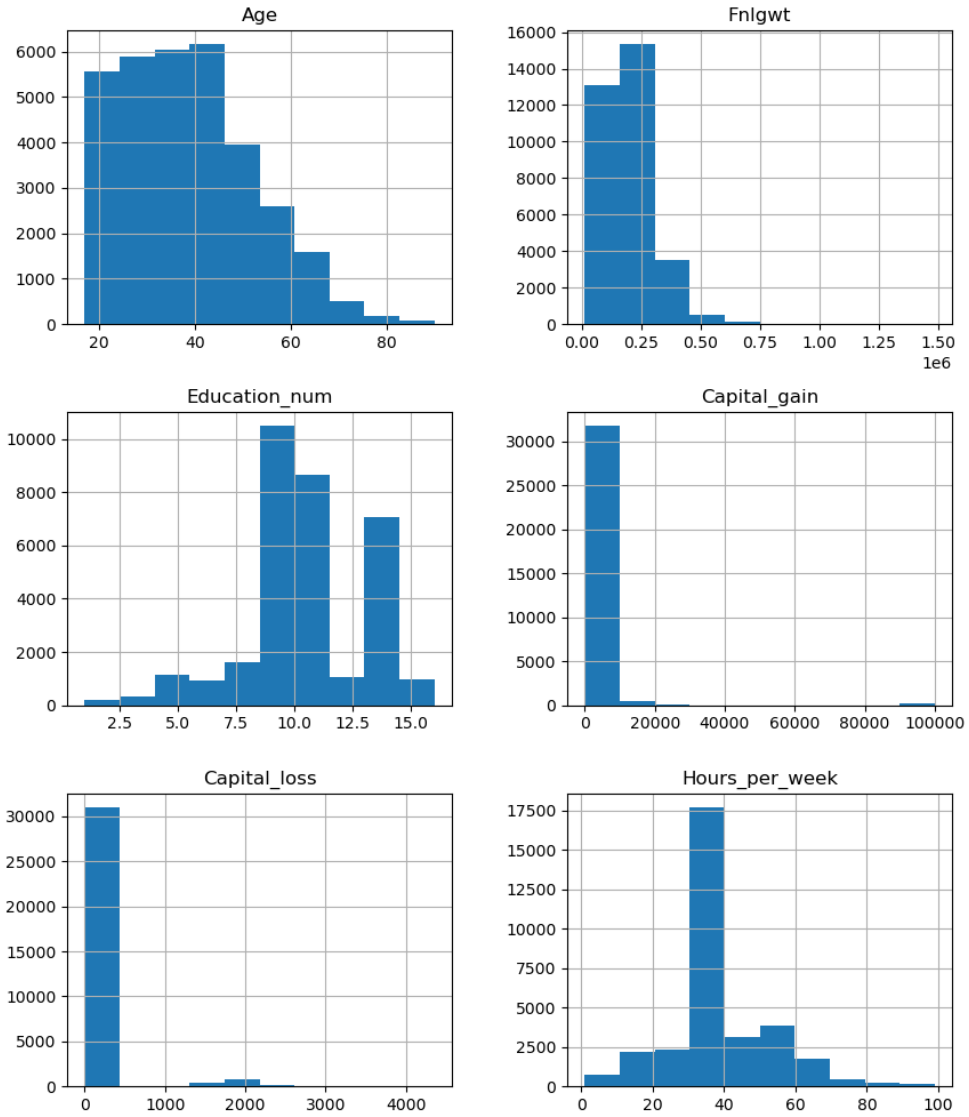


Figure 4: Histograms for variables

Upon examining the histograms for various variables in the dataset, distinctive patterns emerge, shedding light on the distributional characteristics of key attributes. The histogram for the "Age" variable indicates a positive skewness, suggesting that the majority of individuals tend to be younger, with a tail extending towards older ages. This positively skewed distribution implies that the dataset is concentrated towards the lower age range.

Conversely, the "Education Number" variable displays a negatively skewed distribution. This implies that a higher number of individuals are in less education number, while a smaller proportion have achieved higher levels of education, creating a left-skewed in the histogram.

The "Fnlwgt" variable also exhibits a positively skewed distribution. This skewness suggests that the weights assigned to observations vary, with a concentration towards lower values and a tail extending towards higher weights.

Examining the "Capital Gain" histogram reveals a bimodal distribution. A substantial number of individuals either report very small capital gains or have large gains, particularly at the extreme value of $10,000 or $99,000.

Similarly, the "Capital Loss" histogram unveils a concentration of values very small, with only a few instances of large losses, notably at $2,000. This pattern aligns with the characteristics observed in the capital gain variable.

Turning attention to the "Hours per Week" histogram, the data reveals a varied distribution within the range of 1 to 99 hours. Notably, a significant proportion of individuals (approximately 18,000) work within the standard 30-40 hours per week. However, the presence of outliers is apparent, with a few individuals working exceptionally long hours (80-100) highlighting potentially unusual work patterns within the dataset.

These insights garnered from the histograms lay the foundation for a deeper understanding of the underlying dynamics within the Census Income dataset, guiding further exploration and analysis.

Count Plots for the Qualitative Variables

Analyzing the plots below provides valuable insights into the distribution of categorical variables in the dataset.

In the Workclass count plot, the private category stands out with the maximum count, while never worked and without pay categories are noticeably small in comparison. The bars representing these categories are almost negligible due to their minimal counts.

Examining the Education variable count plot, it is evident that HS-grad has the highest count, whereas pre-school has the minimum count. Additionally, the count of individuals with some college education surpasses those with a bachelor's degree, an interesting nuance within the dataset.

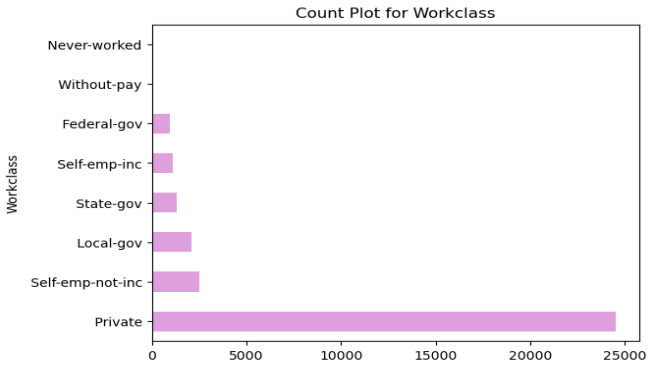
Moving to the Marital Status count plot, married civ spouses dominate with the maximum count, while married AF spouses exhibit the minimum count. Never married emerges as the second-largest category in the Adult dataset. Notably, the bar for married AF spouses is too small to be visible in the count plot.

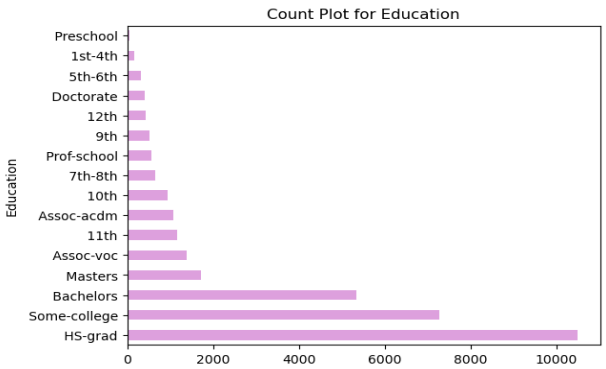
Within the Occupation category, Prof-specialty boasts the highest count, contrasting with the Armed Forces category, which has the minimum count.

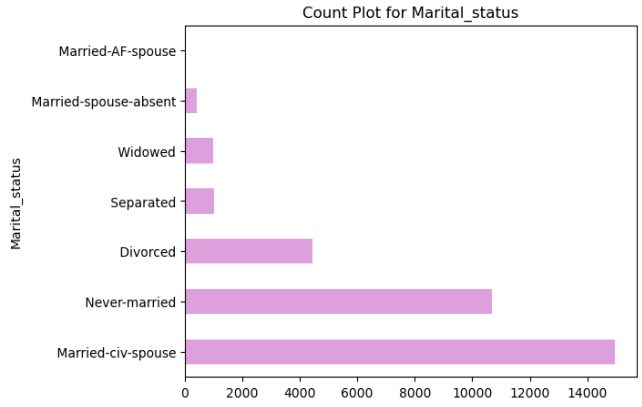
Analyzing the Relationship count plot reveals that the count of individuals classified as husbands is the highest, while other relative relationships have the minimum count.

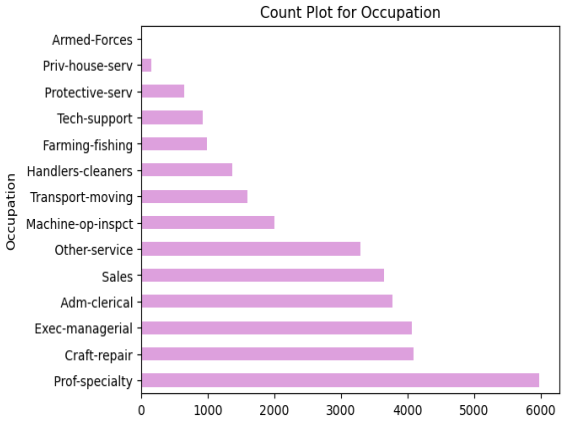
Shifting the focus to the Race count plot, the White category dominates with the maximum count, and the counts for other racial categories are minimal in comparison.

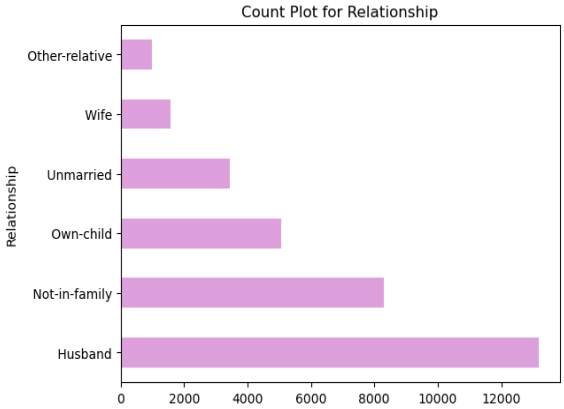
These observations, derived from the count plots, offer a comprehensive understanding of the distribution of categorical variables in the Census Income dataset, enabling nuanced insights into the demographic composition.

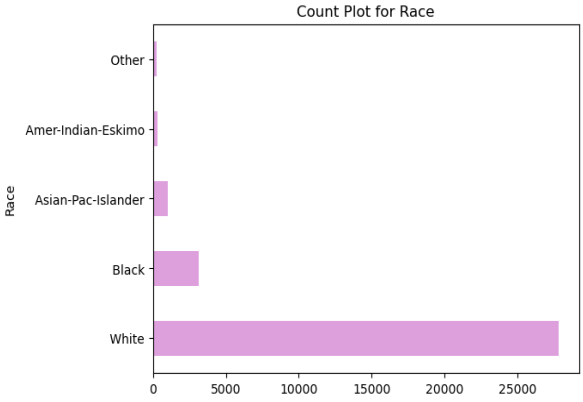












Cross tabs tables and Count Plots for the Variables

Income	<=50K	>50K	Total
Workclass			
Federal-gov	589	371	960
Local-gov	1476	617	2093
Never-worked	7	0	7
Private	19357	5152	24509
Self-emp-inc	494	622	1116
Self-emp-not-inc	1816	724	2540
State-gov	945	353	1298
Without-pay	14	0	14
Total	24698	7839	32537

Income	<=50K	>50K	Total
Education			
10th	871	62	933
11th	1115	60	1175
12th	400	33	433
1st-4th	160	6	166
5th-6th	316	16	332
7th-8th	605	40	645
9th	487	27	514
Assoc-acdm	802	265	1067
Assoc-voc	1021	361	1382
Bachelors	3132	2221	5353
Doctorate	107	306	413
HS-grad	8820	1674	10494
Masters	763	959	1722
Preschool	50	0	50
Prof-school	153	423	576
Some-college	5896	1386	7282
Total	24698	7839	32537

Income	<=50K	>50K	Total
Relationship			
Husband	7271	5916	13187
Not-in-family	7436	856	8292
Other-relative	944	37	981
Own-child	4997	67	5064
Unmarried	3227	218	3445
Wife	823	745	1568
Total	24698	7839	32537

Income	<=50K	>50K	Total
Race			
Amer-Indian-Eskimo	275	36	311
Asian-Pac-Islander	762	276	1038
Black	2735	387	3122
Other	246	25	271
White	20680	7115	27795
Total	24698	7839	32537

Income	<=50K	>50K	Total
Sex			
Female	9583	1179	10762
Male	15115	6660	21775
Total	24698	7839	32537

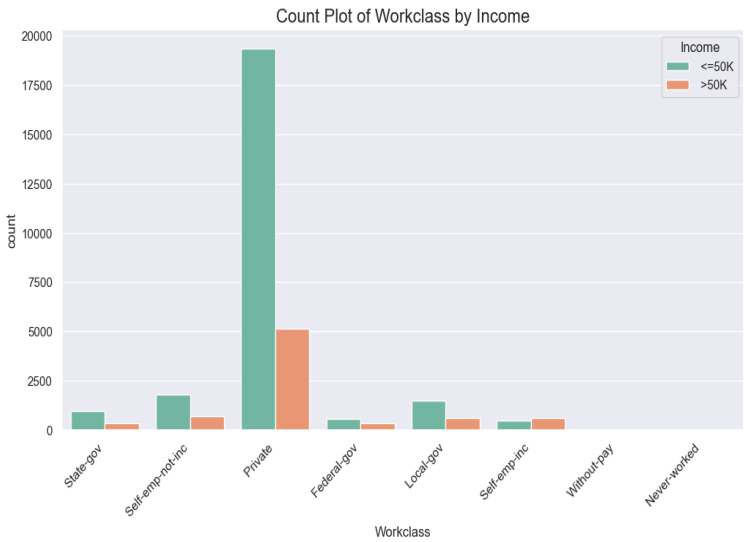
Income	<=50K	>50K	Total
Education_num			
1	50	0	50
2	160	6	166
3	316	16	332
4	605	40	645
5	487	27	514
6	871	62	933
7	1115	60	1175
8	400	33	433
9	8820	1674	10494
10	5896	1386	7282
11	1021	361	1382
12	802	265	1067
13	3132	2221	5353
14	763	959	1722
15	153	423	576
16	107	306	413
Total	24698	7839	32537

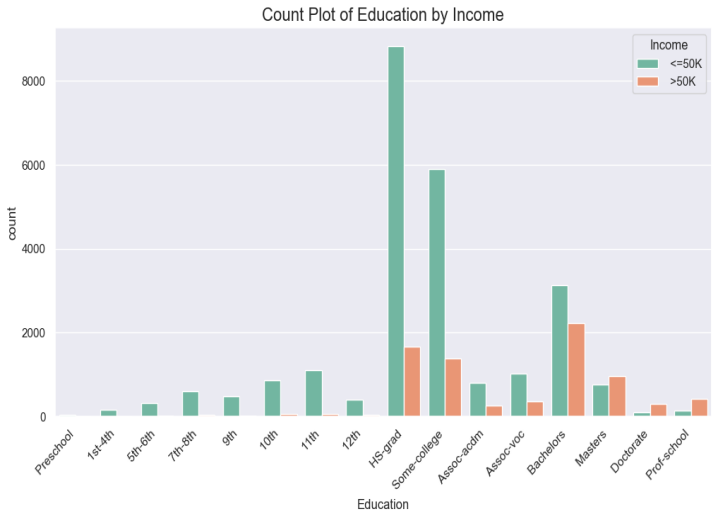
Income	<=50K	>50K	Total
Marital_status			
Divorced	3978	463	4441
Married-AF-spouse	13	10	23
Married-civ-spouse	8280	6690	14970
Married-spouse-absent	384	34	418
Never-married	10176	491	10667
Separated	959	66	1025
Widowed	908	85	993
Total	24698	7839	32537

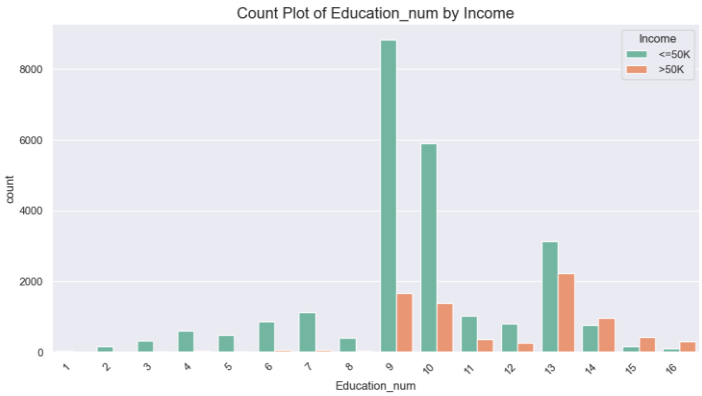
Income	<=50K	>50K	Total
Occupation			
Adm-clerical	3261	507	3768
Armed-Forces	8	1	9
Craft-repair	3165	929	4094
Exec-managerial	2097	1968	4065
Farming-fishing	877	115	992
Handlers-cleaners	1283	86	1369
Machine-op-inspct	1751	249	2000
Other-service	3154	137	3291
Priv-house-serv	146	1	147
Prof-specialty	3930	2049	5979
Protective-serv	438	211	649
Sales	2667	983	3650
Tech-support	644	283	927
Transport-moving	1277	320	1597
Total	24698	7839	32537

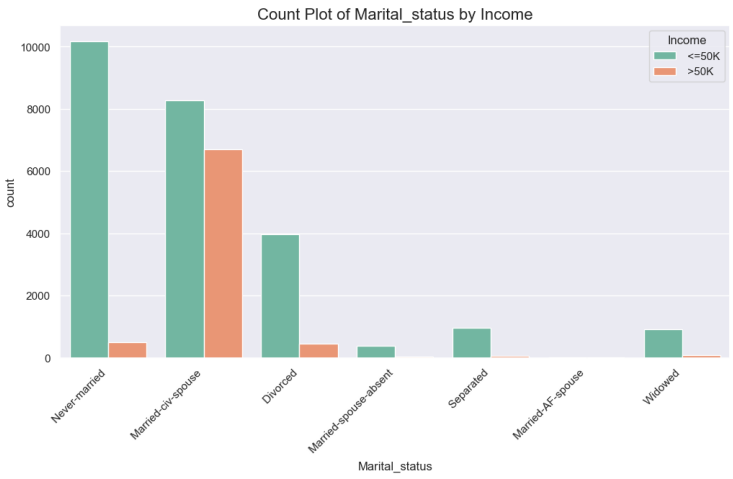
Income	<=50K	>50K	Total
Native_country			
Cambodia	12	7	19
Canada	82	39	121
China	55	20	75
Columbia	57	2	59
Cuba	70	25	95
Dominican-Republic	68	2	70
Ecuador	24	4	28
El-Salvador	97	9	106
England	60	30	90
France	17	12	29
Germany	93	44	137
Greece	21	8	29
Guatemala	59	3	62
Haiti	40	4	44
Holand-Netherlands	1	0	1
Honduras	12	1	13
Hong	14	6	20
Hungary	10	3	13
India	60	40	100
Iran	25	18	43
Ireland	19	5	24
Italy	48	25	73
Jamaica	71	10	81
Japan	38	24	62
Laos	16	2	18
Mexico	606	33	639
Nicaragua	32	2	34
Outlying-US(Guam-USVI-etc)	14	0	14
Peru	29	2	31
Philippines	137	61	198
Poland	48	12	60
Portugal	33	4	37
Puerto-Rico	102	12	114
Scotland	9	3	12
South	64	16	80
Taiwan	31	20	51
Thailand	15	3	18
Trinadad&Tobago	17	2	19
United-States	22420	7315	29735
Vietnam	62	5	67
Yugoslavia	10	6	16
Total	24698	7839	32537

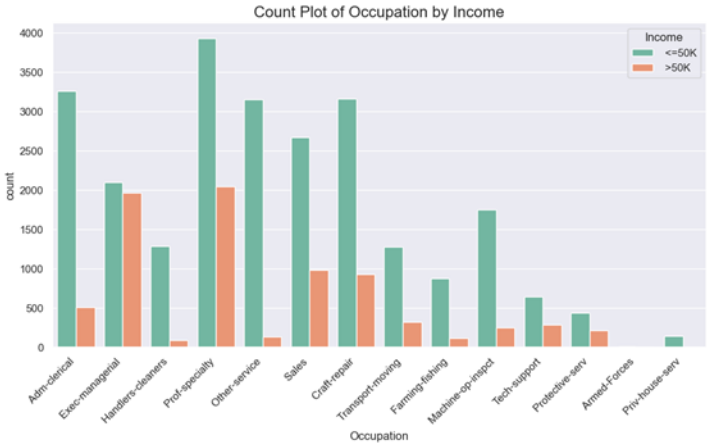
Figure 6: count tables for the variables

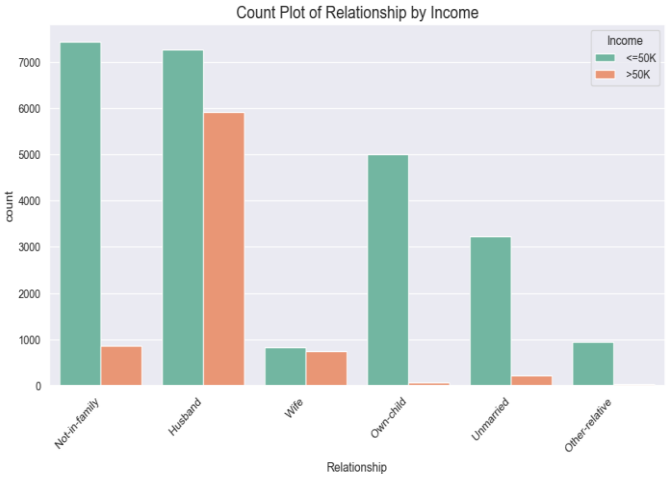


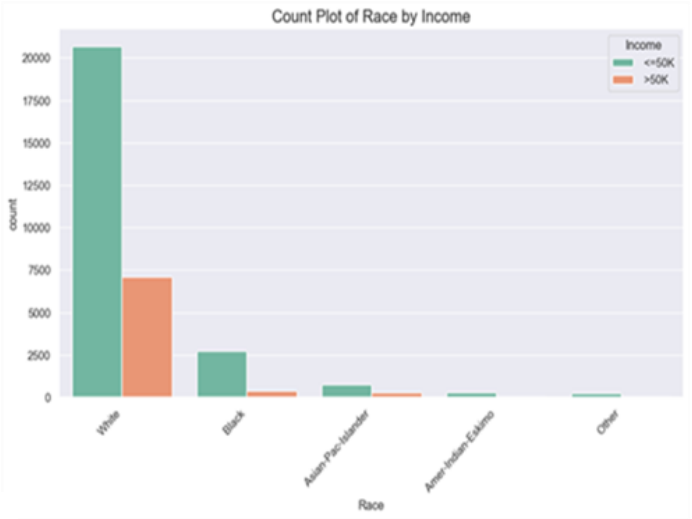












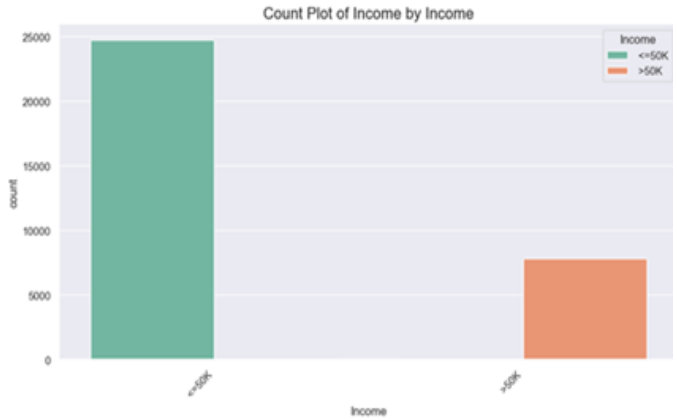


Figure 7: Count plots for variables base on Income

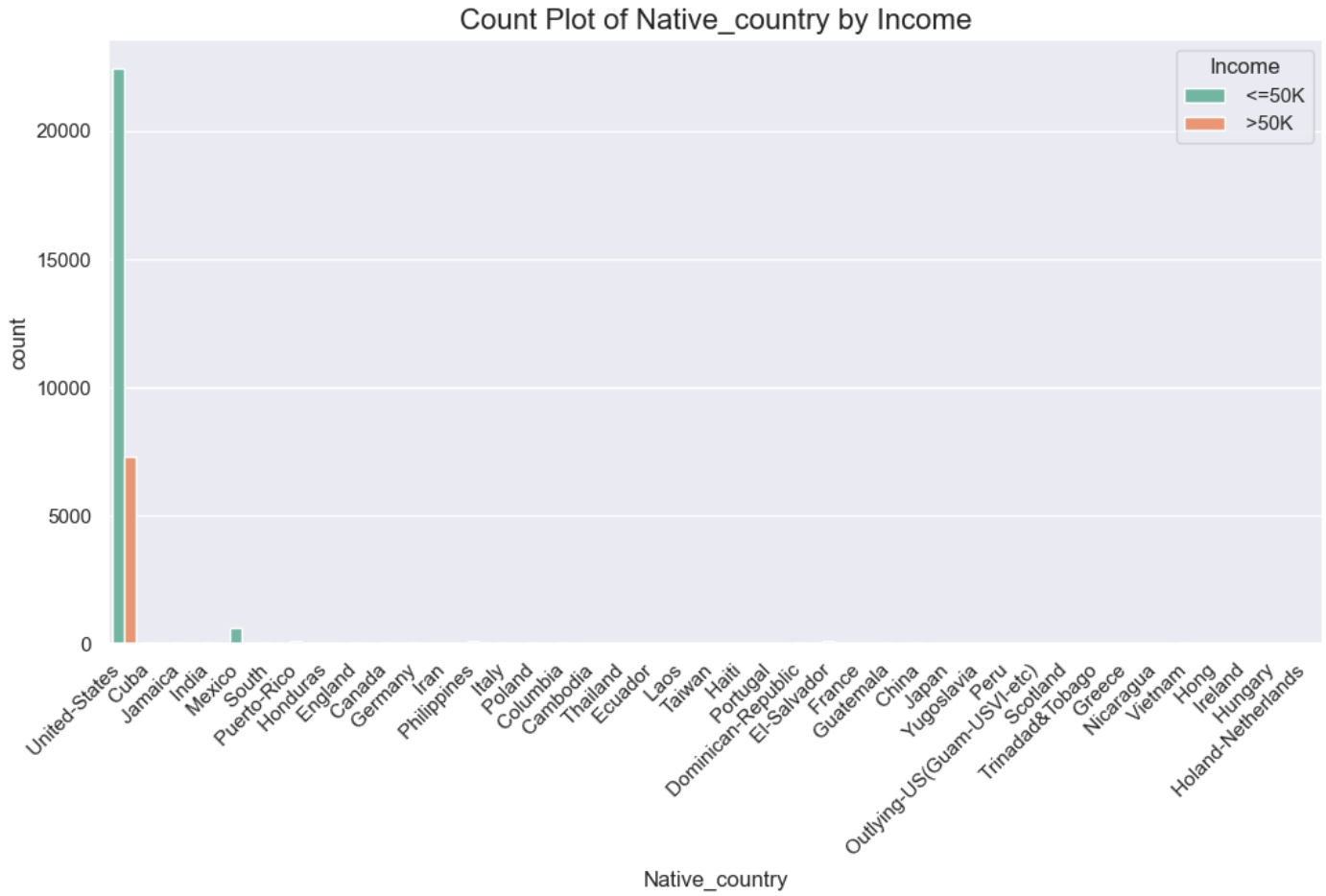


Figure 8: Count plot for Native country by income

Examining the cross-tabulation tables and the Count Plots, notable patterns emerge, shedding light on the relationship between various categorical variables and income brackets.

In the Work class table and the work class plot, the count of individuals with (income >50K) group count is higher than those with (income <=50K) group count, except for the "Never worked" and "Without pay" groups where there are no count instances of income >50K. Additionally, within the "Private" group, the count of individuals with (income <=50K) group exceeds those with (income >50K) group.

Turning to the Education variable, individuals with a Master's degree predominantly have income >50K, while those with a bachelor's degree are more likely to have income <=50K. Interestingly, those with a professional school education exhibit a higher count of (income >50K) group. Similar trends are observed in the education num variable.

In the Marital Status variable, individuals who are divorced, separated, or never married are more likely to have a count of (income <=50K) group compared to those with a count of (income >50K). Moreover, for all categories in marital status, the count of income <=50K exceeds that of the count of income >50K.

Analyzing the Occupation variable, income <=50K dominates across all categories, with "Prof-specialty" having the highest count of income >50K compared to other categories.

In the Relationship variable, the "Husband" category has the highest count of income >50K, while the "Not in family" category is prevalent in the income <=50K group. The "Own child" group represents a higher count of income (<=50K) group compared to income (>50K) group of it.

Considering the Race variable, individuals of White ethnicity have the highest count of income >50K compared to other racial categories. For Black individuals, the count of income <=50K is higher than that of income >50K, unlike in the White category.

In the Sex variable, there are more details for males than females. However, the count of males with income >50K is higher than females.

The count plot provides insights into the distribution of income in the dataset, categorizing individuals into <=50K and >50K income groups. The majority of the dataset represents individuals with income <=50K, with a notably larger count compared to the >50K group.

Examining the Native Country variable, it's notable that the "Holand-Netherlands" and "Outlying-US" categories do not have any instances of income >50K. The United States category has the highest count of details, and most of the individuals with income >50K are also represented within the United States group.

In the Native Country plot, only the columns for the US, Cuba, and Mexico are represented. However, when examining the count table for the native country representation, we observe that values other than the US exist in the "Native country" column, albeit in smaller amounts. To focus on the count values for the other countries in the plot, we exclude the US country, allowing for a more detailed examination of the distribution among the remaining countries. The following plot represents these;

Upon examining the Income-based native country distribution plot, it becomes apparent that after the US, Mexico has the highest count. However, within the Mexico category, the count of individuals in the >50K income group is comparatively small in the <=50K income group. Notably, the lowest count is attributed to Holland Netherlands, as also depicted in this plot. In summary, for all countries, the count of individuals in the <=50K income group is smaller than that in the >50K income group.

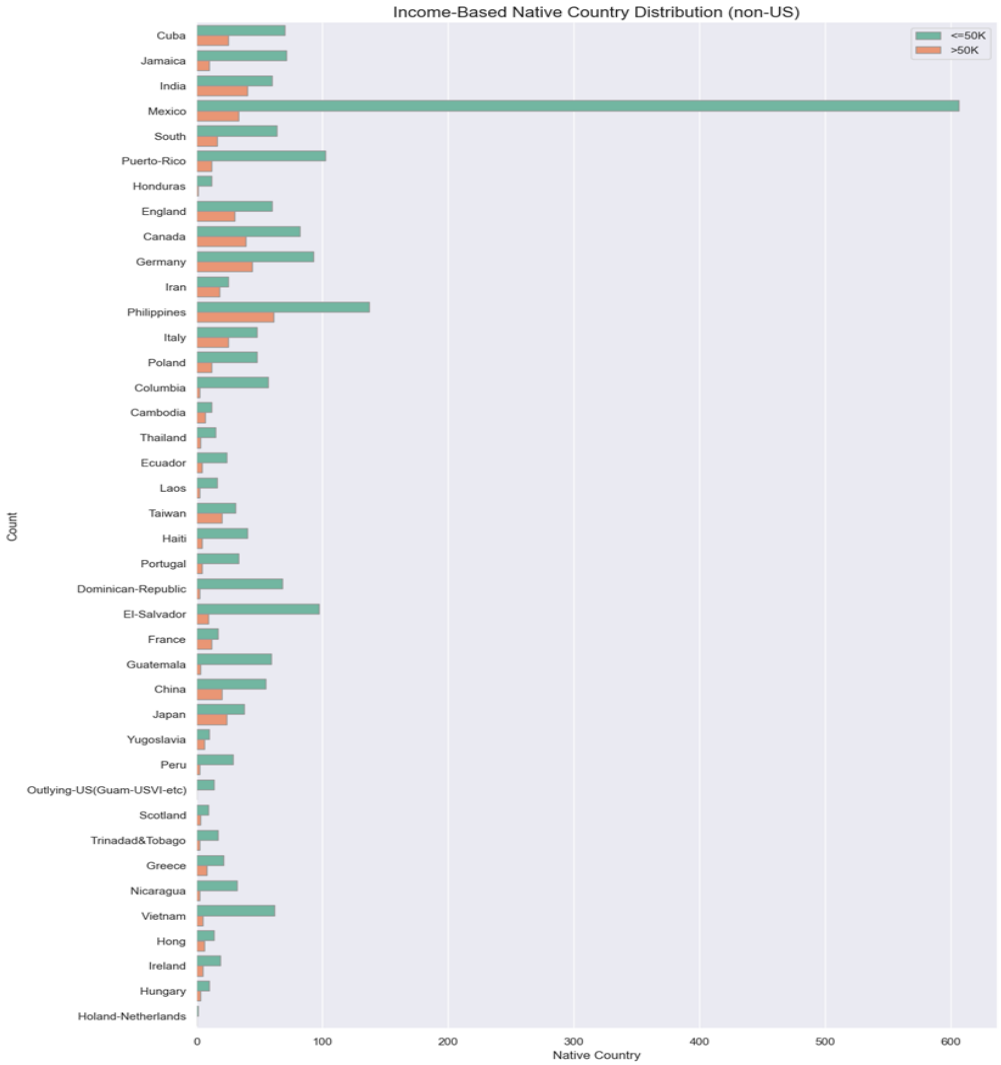


Figure 9:Count plot for Native Country by Income without US

Distribution of Sex Variables in the Dataset

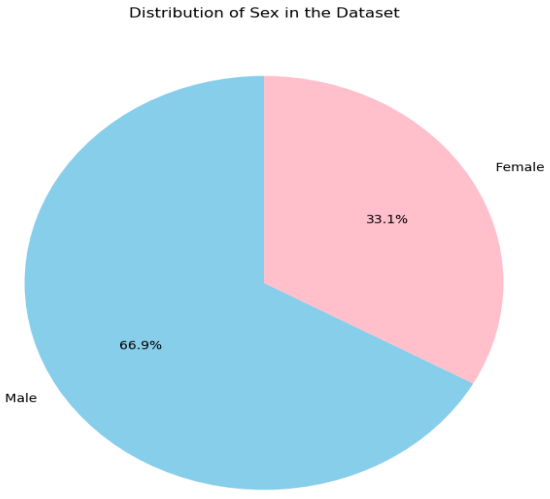


Figure 10: Pie Chart for Gender

The pie chart illustrates the distribution of the "Sex" variable in the Adult dataset. Notably, the majority of details in the dataset are associated with male individuals, constituting 66.9%, while females account for 33.1%.

☐ BIVARIATE ANALYSIS

Paired Plot for the Variables

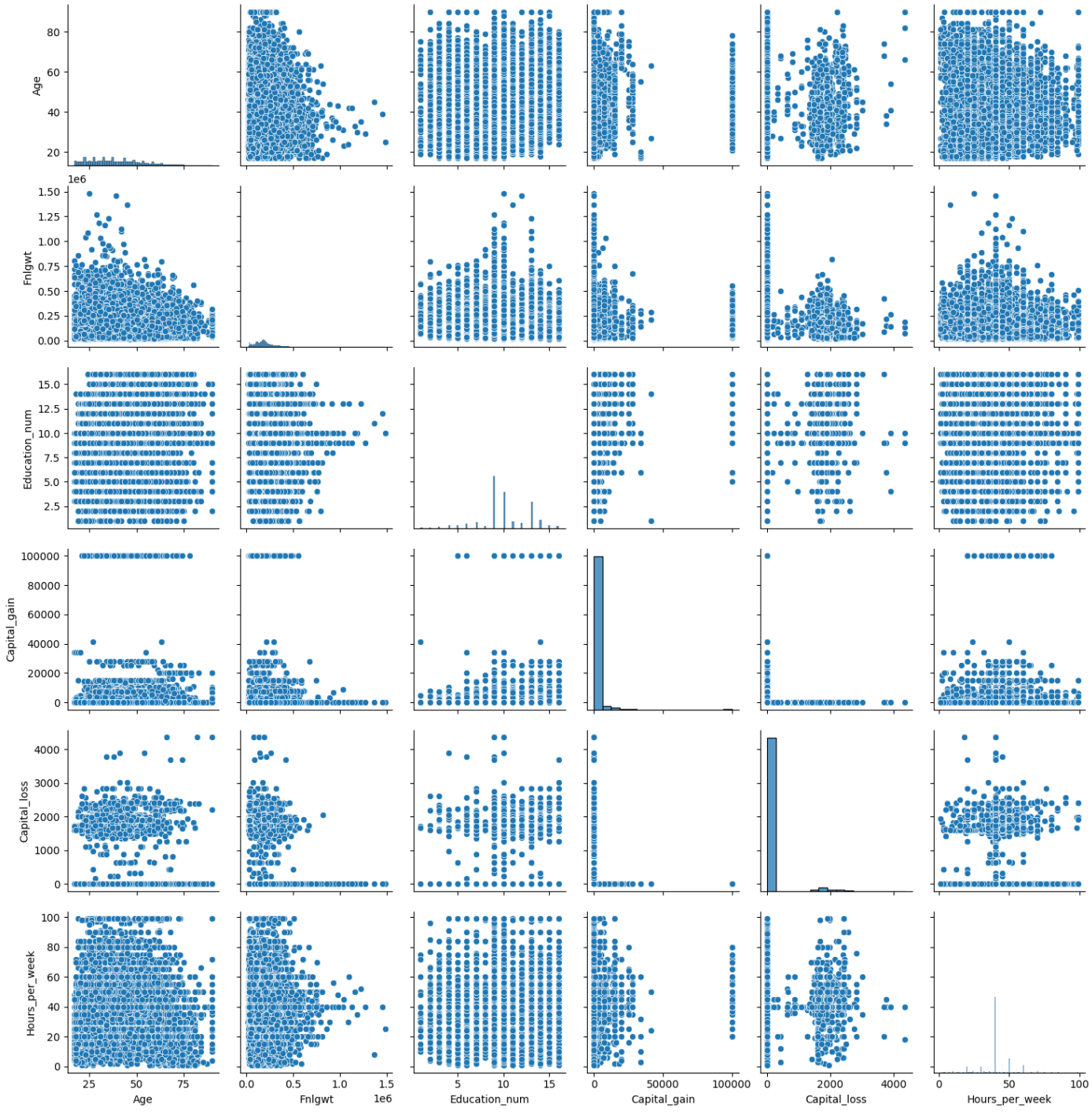


Figure 11: Pair Plot for the Variables

The pair plot shows the relationships between six numerical variables from the Adult Census Income dataset: age, capital gain, capital loss, education num, hours per week, and income. Each variable is plotted against each other variable, creating a grid of 36 scatter plots. But from the scatter plot, can't see any clear relationship between the variables.

Correlation Map for the Variables

Figure 12: Correlation Map for the Continuous Variables

In the above pair plot, we cannot specialty identify the relationship between two variables. Then when we draw a correlation Map for the variable, we can we correlation value the between these variables. Here it's that most of the variables have a very weak positive correlation between two variables. But in between the (Fnlgwt and Age) variable, (Captial_losss and Capital_gain) variable, (Fnlwgt and Education_num) variable represent in very weak negative relationships.

Density plots for the Income based on Hours_per_week and the Age Variables

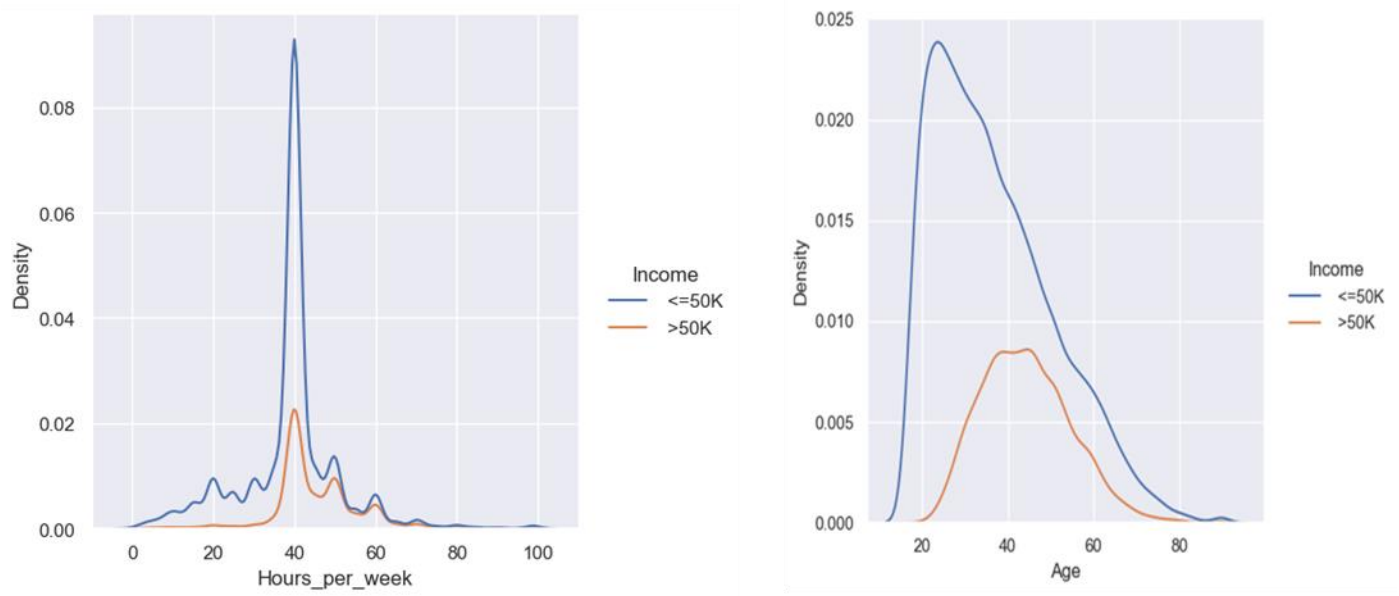


Figure 13: Density plot for Income based on Hours_per_wekk and the Age variables

The two density plots show the distribution of hours worked per week for two groups of people: those with an income of less than or equal to $50,000 and those with an income of more than $50,000. The density plots are a smoothed version of histograms, and they show the probability density of a variable at each point in its range.

The density plot for the lower-income group shows that most people in this group work between 20 and 60 hours per week. The density plot for the higher-income group shows that people in this group are more likely to work long hours. The peak of the density plot is at around 40 hours per week, and there is a long tail that extends out to 100 hours per week.

It is important to note that these density plots are only based on a small sample of data, and they may not be representative of the entire population. However, suggests that there is a relationship between income and hours worked per week. People with higher incomes are more likely to work long hours like 40 hours per week than people with lower incomes. As we can see in a (>50K) income density curve small amount of the counts represent less than approximately 35 hours.

The Age-based Income Density plot reveals that the curve for this age group peaks at around 40 hours per week. The count of individuals in the <=50K income group surpasses that in the >50K group. Notably, for individuals beyond the age of 70, the count of those in the >50K income group is minimal, suggesting a decline in employment at this age. This may imply that individuals in this age range have accumulated capital and are not actively employed. Conversely, the count of individuals in the <=50K income group remains significant, indicating that they have to continue working because potentially lack substantial capital.

Violon Plot of Age by Income Variable

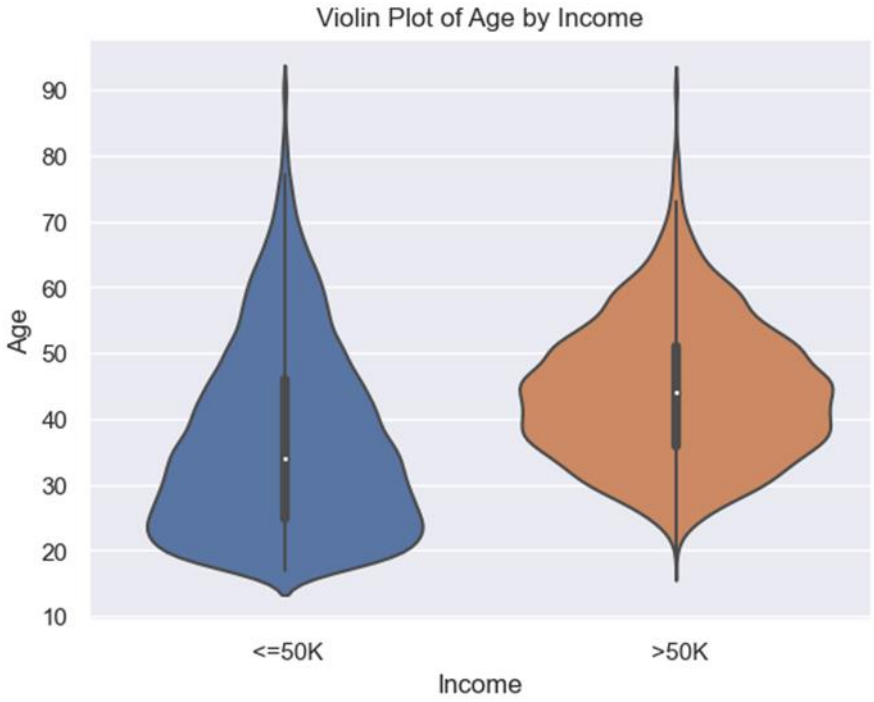


Figure 14: Violin Plot of Age by Income

It shows the distribution of age for people with different income levels. The violin plot itself is made up of two parts: the kernel density plot and the box plot.

People with higher incomes tend to be younger. The median age for people with incomes above $50,000 is approximately 45, while the median age for people with incomes below $50,000 is about 35.

The distribution of ages is wider for people with lower incomes. The box and whiskers for the lower-income group are wider than the box and whiskers for the higher-income group. This means that there is a greater range of ages in the lower-income group.

There are outliers in both income groups. The whiskers extend beyond the boxes in both groups, which means that there are a few people in each group who have ages that are much lower or much higher than the median.

Overall, this plot suggests that there is a relationship between age and income. People with higher incomes tend to be younger, and the distribution of ages is wider for people with lower incomes.

Bar plots for Income vs. Categorical variables

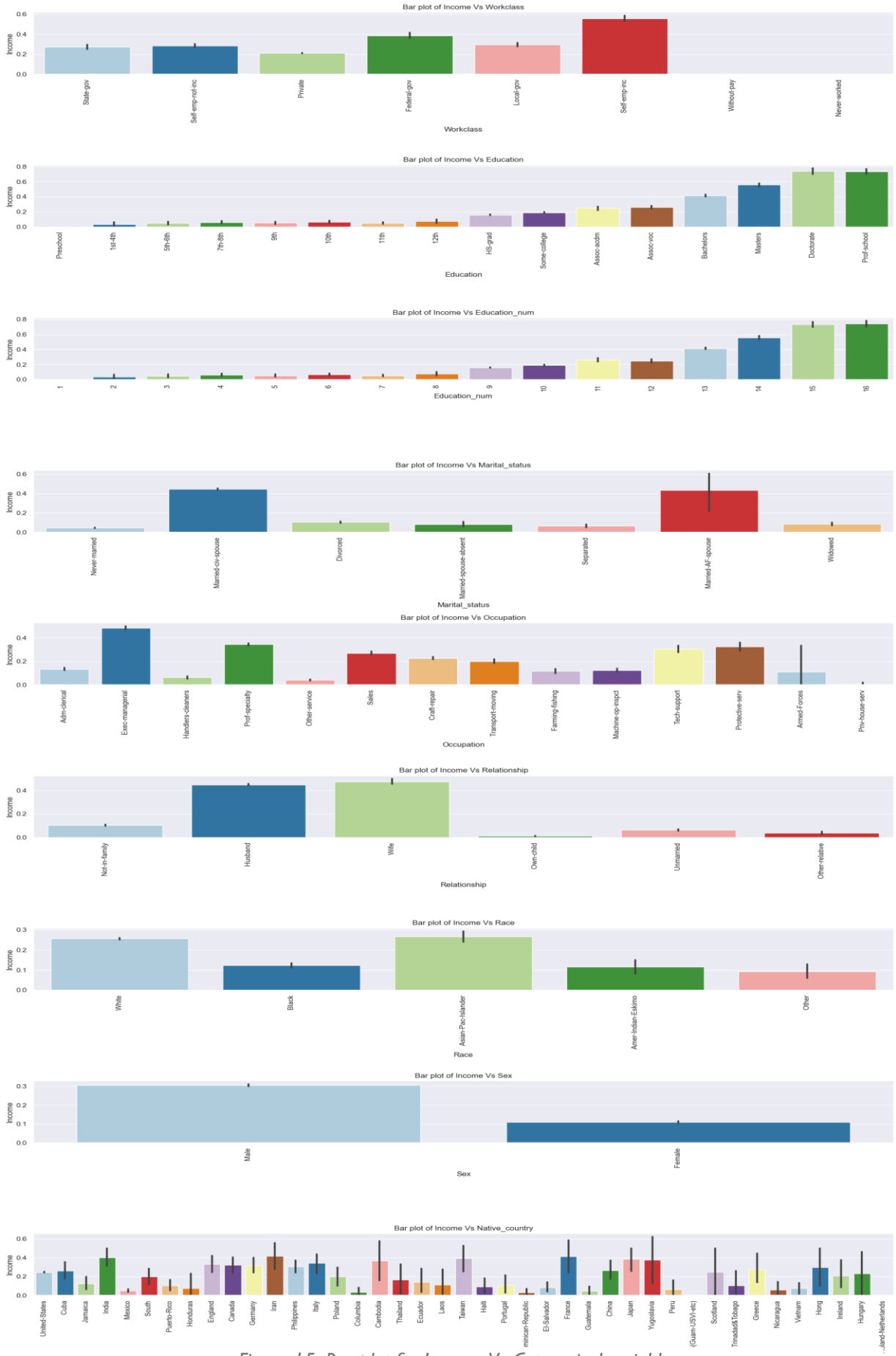


Figure 15: Bar plot for Income Vs Categorical variables

In the Income vs. Work Class plot, the highest income is observed in the Self-Emp-Inc category, while the Never worked and without pay categories exhibit the lowest income levels. Notably, government categories tend to have higher incomes compared to the private category.

The Bar plot depicting Income vs. Education reveals a clear trend: as education levels progress from lower to higher, including categories like 'Preschool,' '1st-4th,' '5th-6th,' and so forth, income also increases. This positive correlation between education and income is a noteworthy observation.

Similarly, the Education Number vs. Income plots reinforce the notion that as the education number increases, so does income. This positive relationship is visually evident in the plots.

In the Married vs. Income plot, individuals categorized as never married tend to have the lowest income, while Married-Civ-Spouses exhibit the highest income. Additionally, separated individuals tend to earn more than those who are divorced.

The Occupation vs. Income plot highlights that individuals in the Exec-managerial occupation earn the most, while those in the Priv-House-Serv earn the lowest income. Prof-Specialty follows as the second-highest earning occupation.

According to the Income vs. Relationship plot, individuals classified as wives tend to have the highest income, followed by husbands, while the own-child category exhibits the lowest income.

In the Income vs. Race plot, individuals belonging to the Asian-Pac-Islander race category have the highest income. Notably, the White race category has a lower income than Asian-Pac-Islander but a higher income than the Black race category.

Examining the Income vs. Sex plot, it's evident that males have higher incomes than females, indicating a gender-based income disparity.

Considering the Income vs. Native Country plot, individuals in Iran have the highest income, while the Dominican Republic represents the lowest income among native countries. China and India rank as the second and third highest, respectively.

☐ MULTIVARIATE ANALYSIS

Income>50K counts in Education categories vs. Work class categories Table and Heat Map

Table 2: Income>50K count in Education categories vs Work class categories Table

Income						
Workclass	Federal-gov	Local-gov	Private	Self-emp-inc	Self-emp-not-inc	State-gov
Education						
10th	NaN	1.0	49.0	3.0	7.0	2.0
11th	1.0	2.0	45.0	4.0	7.0	1.0
12th	NaN	2.0	25.0	1.0	3.0	2.0
1st-4th	NaN	NaN	5.0	NaN	1.0	NaN
5th-6th	NaN	1.0	9.0	2.0	4.0	NaN
7th-8th	NaN	1.0	20.0	5.0	14.0	NaN
9th	1.0	3.0	19.0	NaN	4.0	NaN
Assoc-acdm	19.0	28.0	176.0	18.0	18.0	6.0
Assoc-voc	15.0	25.0	269.0	19.0	21.0	12.0
Bachelors	95.0	162.0	1540.0	171.0	163.0	90.0
Doctorate	15.0	17.0	143.0	29.0	31.0	71.0
HS-grad	73.0	90.0	1164.0	119.0	179.0	49.0
Masters	47.0	173.0	552.0	57.0	59.0	71.0
Prof-school	23.0	19.0	179.0	78.0	106.0	18.0
Some-college	82.0	93.0	957.0	116.0	107.0	31.0

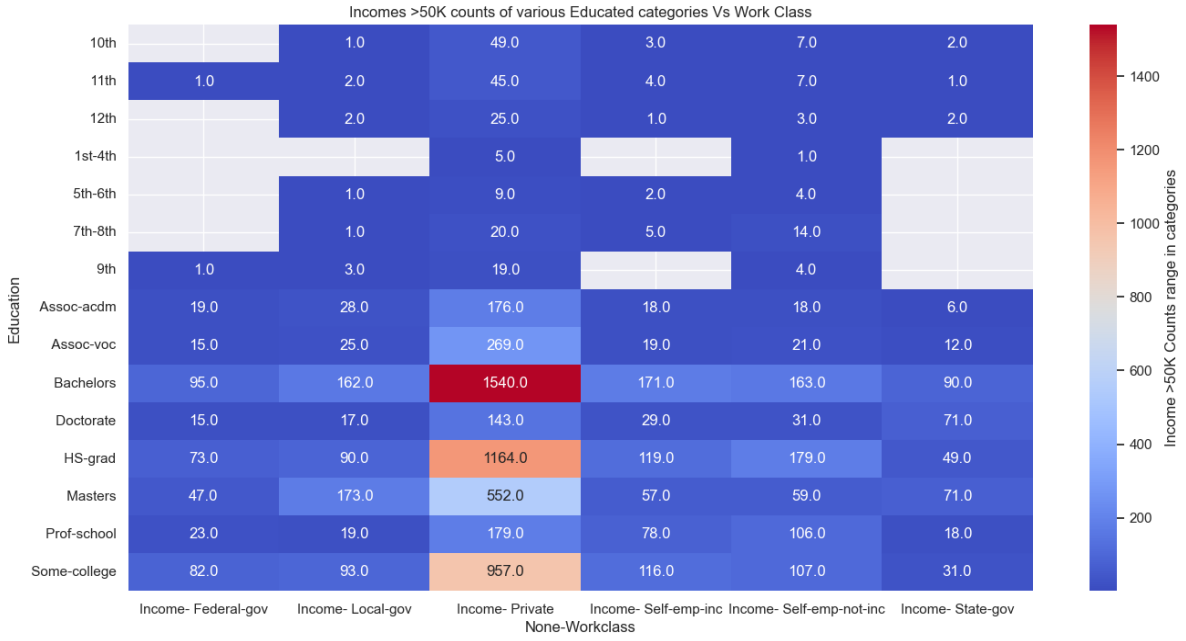


Figure 16 Income>50K counts in Education categories vs Work class categories Heat Map:

According to this table, we observe the counts of individuals with Income >50K categorized by Education levels and Work class. Notably, the Private sector stands out with the highest count, where 1540 individuals have an income >50K. Following closely is the HS-grad category in the Private sector, representing a substantial count of 1164. This relationship is also visually represented in the heatmap. Examining the heatmap, we find that in the Private sector, the counts for Some college and Masters are 957 and 552, respectively, ranking as the third and fourth largest counts. In this heat map not showing any value or color represent NaN values of the above table.

Income>50K counts in Occupation categories vs. Work class categories Table and Heat Map

Table 3: Income>50K counts in Occupation categories vs. Work class categories Table

	Income					
Workclass	Federal-gov	Local-gov	Private	Self-emp-inc	Self-emp-not-inc	State-gov
Occupation						
Adm-clerical	101.0	33.0	321.0	9.0	16.0	27.0
Armed-Forces	1.0	NaN	NaN	NaN	NaN	NaN
Craft-repair	21.0	40.0	721.0	38.0	95.0	14.0
Exec-managerial	92.0	102.0	1295.0	254.0	144.0	81.0
Farming-fishing	2.0	2.0	30.0	15.0	64.0	2.0
Handlers-cleaners	2.0	7.0	73.0	NaN	3.0	1.0
Machine-op-inspct	2.0	2.0	224.0	5.0	11.0	5.0
Other-service	3.0	12.0	100.0	6.0	12.0	4.0
Priv-house-serv	NaN	NaN	1.0	NaN	NaN	NaN
Prof-specialty	95.0	254.0	1198.0	121.0	210.0	171.0
Protective-serv	14.0	135.0	30.0	2.0	1.0	29.0
Sales	5.0	3.0	684.0	160.0	128.0	3.0
Tech-support	25.0	15.0	221.0	2.0	11.0	9.0
Transport-moving	8.0	12.0	254.0	10.0	29.0	7.0

Figure 17: Income>50K counts of various Occupation categories Vs Work Class Heat Map

In this table, we explore the counts of individuals with an income greater than $50K, categorized by Occupation levels and Work class. Notably, within the Private sector, the highest count is observed in the Executive-managerial category, where 1295 individuals earn over $50K. This trend is visually reinforced in the heatmap, where the corresponding cell is highlighted in red. Following closely, the second-largest count is associated with the Prof-specialty occupation in the Private sector, totaling 1198. Additionally, the Craft repair occupation in the Private sector represents the third-largest count for individuals with an income exceeding $50K.

Income>50K counts in Occupation Race vs. Education Table and Heat Map

Table 4: Income>50 K counts in Occupation Race vs. Education Table

	Income														
Education	**10th**	**11th**	**12th**	**1st-4th**	**5th-6th**	**7th-8th**	**9th**	**Assoc-acdm**	**Assoc-voc**	**Bachelors**	**Doctorate**	**HS-grad**	**Masters**	**Prof-school**	**Some-college**
Race															
Amer-Indian-Eskimo	NaN	2.0	NaN	NaN	NaN	NaN	NaN	1.0	1.0	8.0	2.0	11.0	3.0	2.0	6.0
Asian-Pac-Islander	1.0	1.0	1.0	NaN	3.0	NaN	1.0	8.0	9.0	97.0	18.0	34.0	43.0	27.0	33.0
Black	6.0	7.0	5.0	1.0	NaN	2.0	4.0	19.0	18.0	96.0	9.0	86.0	40.0	8.0	86.0
Other	1.0	NaN	NaN	NaN	1.0	NaN	NaN	2.0	NaN	5.0	1.0	2.0	2.0	4.0	7.0
White	54.0	50.0	27.0	5.0	12.0	38.0	22.0	235.0	333.0	2015.0	276.0	1541.0	871.0	382.0	1254.0

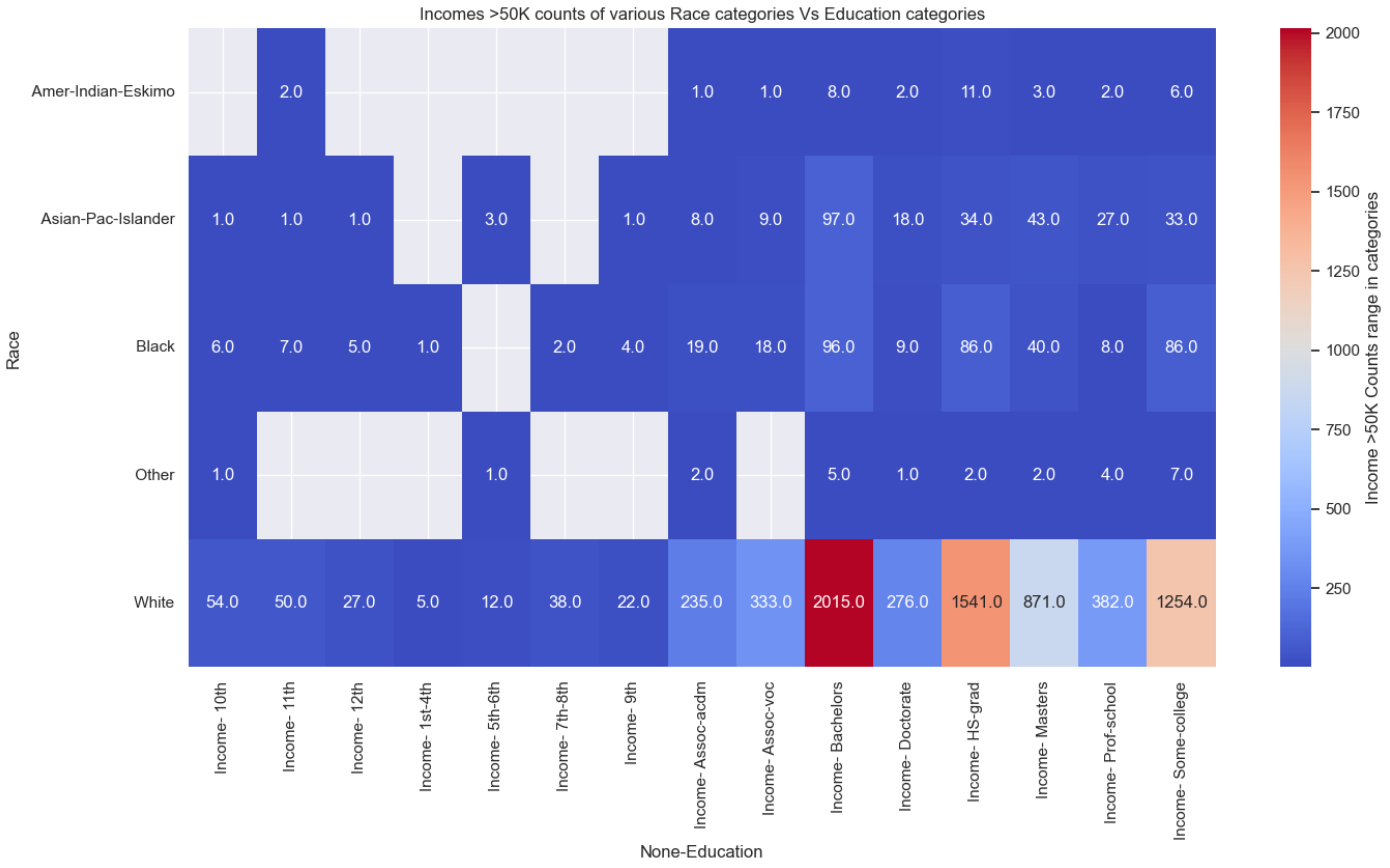


Figure 18: Income>50K counts in Occupation Race vs. Education Heat Map

This table and heatmap illustrate the counts of individuals with an income greater than $50K, categorized by Race and Education levels. Notably, the highest count is observed in the "White" race category with a "Bachelors" education level, totaling 2015. The second-highest count corresponds to the "HS-grad" education level within the "White" race category, accounting for 1541. Following closely, the third-largest count is associated with the "Some college" education level in the "White" race category, representing 1254 individuals. These observations highlight that the highest counts are predominantly concentrated within the "White" race category and various education levels.

APPENDIX

Python Code for Analyzing the adult data set

```
# <ins><font color="purple"><center>EDA Adult Data Set</center></font></ins>


## <ins> Python Librarys </ins>

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

%matplotlib inline


## <ins> Add column names to adult csv data set </ins>


column_names = ['Age','Workclass','Fnlgwt','Education','Education_num','Marital_status',

        'Occupation','Relationship','Race','Sex','Capital_gain','Capital_loss',

        'Hours_per_week','Native_country','Income']

df = pd.read_csv('adult.csv', names=column_names)

df.head()


## <ins> Handling missing values </ins>


df[df == ' ?'] = np.nan

df.isna().any()


df.info()


### <font color="green">Recoding ? as a Mode value of Categorial Variables </font>


for col in ['Workclass', 'Occupation', 'Native_country']:

    df[col].fillna(df[col].mode()[0], inplace=True)


df.isna().any()
```

<ins> Checking the duplicates </ins>

df.duplicated() # True for duplicated in rows

df.duplicated().sum() # Number of duplicated rows

len(df)

df.drop_duplicates(inplace=True) # Dropping duplicated rows

len(df)

df.duplicated().sum()

<ins> Search unique values in Variables </ins>

print(df.Workclass.unique())

print(df.Education.unique())

print(df.Marital_status.unique())

print(df.Occupation.unique())

print(df.Relationship.unique())

print(df.Native_country.unique())

print(df.Income.unique())

<ins> Exploratory Data Analysis </ins>

 Univariate Analysis

 Histograms for Continuous Data

```
Histogram_graphs=df.select_dtypes(include=['int'])
Histogram_graphs.hist(figsize=(10,12))
plt.show()
```

 Draw a pair plot for dataset

```
sns.pairplot(df)
plt.show()
```

<ins> Counts plots for variables <ins>

 Workclass, Education, Marital_status, Occupation, Relationship, Race

```
categorical_columns = ['Workclass', 'Education', 'Marital_status', 'Occupation', 'Relationship', 'Race']

for colname in categorical_columns:
    plt.title('Count Plot for ' + colname)

    (df[colname].value_counts().head(20).plot(kind='barh', color='plum'))
    plt.show()

categorical_columns = ['Workclass', 'Education', 'Marital_status', 'Occupation', 'Relationship', 'Race']

# Set up the matplotlib figure with subplots
fig, axes = plt.subplots(nrows=3, ncols=2, figsize=(10, 20))
```

```
# Flatten the axes for easier iteration
axes = axes.flatten()

# Loop through the categorical columns and create count plots
for i, column in enumerate(categorical_columns):
    sns.countplot(y=column, data=df, ax=axes[i])
    axes[i].set_title(f'Count Plot for {column}')

# Adjust layout to prevent overlapping
plt.tight_layout()

# Show the plot
plt.show()

### <font color='green'> Income </font>

plt.figure(figsize=(4,4))
sns.countplot(x="Income", data=df)

##  <ins> Pie Chart for Gender </ins>

sex_counts = df['Sex'].value_counts()

# Plotting the pie chart
plt.figure(figsize=(8, 8))
plt.pie(sex_counts, labels=sex_counts.index, autopct='%1.1f%%', startangle=90, colors=['skyblue', 'pink'])
plt.title('Distribution of Sex in the Dataset')
plt.show()
```

<ins> Count Plots Based on Income </ins>

```
# Categorical columns in the dataset
categorical_columns = ['Workclass', 'Education', 'Education_num', 'Marital_status', 'Occupation', 'Relationship', 'Race', 'Sex', 'Native_country', 'Income']

# Set the style of seaborn for better visualization
sns.set(style="darkgrid")

# Plot count plots for each categorical variable
for column in categorical_columns:
    plt.figure(figsize=(12, 6))

    # Set the order for 'Education' variable
    if column == 'Education':
        order = sorted(df['Education'].unique(), key=lambda x: [' Preschool', ' 1st-4th', ' 5th-6th', ' 7th-8th', ' 9th', ' 10th', ' 11th', ' 12th',  ' HS-grad', ' Some-college', ' Assoc-acdm', ' Assoc-voc', ' Bachelors', ' Masters', ' Doctorate', ' Prof-school'].index(x))
        sns.countplot(x=column, hue='Income', data=df, palette='Set2', order=order)
    else:
        sns.countplot(x=column, hue='Income', data=df, palette='Set2')

    plt.title(f'Count Plot of {column} by Income', fontsize=16)
    plt.xticks(rotation=45, ha='right')
    plt.show()

plot = sns.catplot(data=df.query('Native_country != " United-States"'), y='Native_country', hue="Income", kind="count",
                   palette="Set2", edgecolor=".6", legend=False,
                   height=16, aspect=.8, orient='v');
plot.set_xlabels('Native Country');
plot.set_ylabels('Count');
plt.legend(loc='upper right', labels=['<=50K', '>50K']);
plt.title('Income-Based Native Country Distribution (non-US)', fontsize=16);
```

<ins> Cross Table Based on Sex </ins>

column_name=["Age", "Workclass", "Education", "Education_num", "Marital_status", "Occupation", "Relationship", "Race", "Sex", "Native_country"]

for column in column_name:

    if column != 'Income':

        print(pd.crosstab(df[column], df['Income'], margins=True, margins_name='Total'))

        print("\n")

<font color='orange'> Bivariate Analysis </color>

<ins> Density Plots Hours_per_week and Age Based on Income </ins>

sns.displot(x='Age', hue='Income', data=df, kind='kde')

sns.displot(x='Hours_per_week', hue='Income', data=df, kind='kde')

<ins>Violine Plot for Age Based on Income</ins>

sns.violinplot(x = 'Income', y = 'Age', data = df, size = 6)

plt.title('Violin Plot of Age by Income')

<ins> Bar Plot For Income Vs Categoriacal Variables </ins>

df1 = df.copy()

Convert 'Income' column to binary values (1 if '>50K', 0 otherwise)

df1['Income'] = df['Income'].apply(lambda x: 1 if x == ' >50K' else 0)

education_order = [' Preschool', ' 1st-4th', ' 5th-6th', ' 7th-8th', ' 9th', ' 10th', ' 11th', ' 12th', ' HS-grad', ' Some-college', ' Assoc-acdm', ' Assoc-voc', ' Bachelors', ' Masters', ' Doctorate', ' Prof-school']

List of categorical variables to include in the plot

categorical_vars = ['Workclass', 'Education', 'Education_num', 'Marital_status', 'Occupation', 'Relationship', 'Race', 'Sex', 'Native_country']

```
# Determine the number of rows and columns dynamically

num_plots = len(categorical_vars)

num_cols = min(1, num_plots)

num_rows = (num_plots - 1) // num_cols + 1


# Create a grouped bar plot for each categorical variable

plt.figure(figsize=(18, 4 * num_rows))

for i, var in enumerate(categorical_vars, 1):

    plt.subplot(num_rows, num_cols, i)

    sns.barplot(x=var, y="Income", data=df1, order=education_order if var == 'Education' else None,
palette="Paired")

    plt.xticks(rotation=90)

    plt.title(f"Bar plot of Income Vs {var}")

plt.tight_layout()

plt.show()


# Categorical columns in the dataset

categorical_columns = ['Workclass', 'Education', 'Education_num', 'Marital_status', 'Occupation', 'Relationship', 'Race',
'Sex', 'Native_country']


# Set the style of seaborn for better visualization

sns.set(style="darkgrid")


# Plot count plots for each categorical variable

for column in categorical_columns:

    plt.figure(figsize=(12, 6))


    # Set the order for 'Education' variable

    if column == 'Education':

        order = sorted(df['Education'].unique(), key=lambda x: [' Preschool', ' 1st-4th', ' 5th-6th', ' 7th-8th', ' 9th', '
10th', ' 11th', ' 12th',  ' HS-grad', ' Some-college', ' Assoc-acdm', ' Assoc-voc', ' Bachelors', ' Masters', ' Doctorate', '
Prof-school'].index(x))

        sns.barplot(x=column, y='Income', data=df1, palette='Set2', order=order)
```

```
    else:
        sns.barplot(x=column, y='Income', data=df1, palette='Set2')

    plt.title(f'Count Plot of {column} by Income', fontsize=16)
    plt.xticks(rotation=45, ha='right')
    plt.show()
```

<ins> Heat Map </ins>

```
numerical_columns = df.select_dtypes(include=['int64', 'float64'])

# Create a correlation matrix
correlation_matrix = numerical_columns.corr()

# Set up the matplotlib figure
plt.figure(figsize=(12, 10))

# Create a heatmap using seaborn to visualize the correlation matrix
sns.heatmap(correlation_matrix, annot=True, cmap="YlGn", fmt=".2f", linewidths=.5)

# Show the plot
plt.title("Correlation Plot of Numerical Features")
plt.show()
```

 Multivariate Analysis

```
mult_df = df.where(df.Income == " >50K").pivot_table(values=['Income'],
                                                     index='Education',
                                                     columns='Workclass',
                                                     aggfunc='count')

mult_df.sort_index()
```

```
plt.figure(figsize=(16, 8))

sns.heatmap(mult_df.sort_index(), annot=True, fmt='.1f', cbar_kws= {'label':'Income >50K Counts range in categories'}, cmap='coolwarm')

plt.title('Incomes >50K counts of various Educated categories Vs Work Class')



mult_df1 = df.where(df.Income == ' >50K').pivot_table(values=['Income'],
                                                      index='Occupation',
                                                      columns='Workclass',
                                                      aggfunc='count')


mult_df1.sort_index()


plt.figure(figsize=(16, 8))

sns.heatmap(mult_df1.sort_index(), annot=True, fmt='.1f', cbar_kws= {'label':'Income >50K Counts range in categories'}, cmap='coolwarm')

plt.title('Incomes >50K counts of various Occupation categories Vs Work Class')


mult_df2 = df.where(df.Income == ' >50K').pivot_table(values=['Income'],
                                                      index='Race',
                                                      columns='Education',
                                                      aggfunc='count')


mult_df2.sort_index()


plt.figure(figsize=(16, 8))

sns.heatmap(mult_df2.sort_index(), annot=True, fmt='.1f', cbar_kws= {'label':'Income >50K Counts range in categories'}, cmap='coolwarm')

plt.title('Incomes >50K counts of various Race categories Vs Education categories')
```

REFERENCES

- Lecture notes
- https://archive.ics.uci.edu/dataset/2/adult
- Matthes, E. (2023). Python crash course: A hands-on, project-based introduction to programming.