UCI Adult Income Dataset - Exploratory band Descriptive Analysis

In this notebook, we carry out an in-depth exploratory and descriptive analysis of the UCI Adult Income Dataset, a widely used dataset for income prediction tasks based on individual demographic and employment attributes.

This phase of analysis is essential for uncovering patterns, detecting potential biases, and gaining intuition about the dataset's structure before applying any modelling procedures. We examine the distribution of key numerical and categorical variables, investigate relationships between demographic features and income levels, and use visualizations to summarize insights. Particular focus is placed on income disparities across **age groups, geographical regions, races, and education-occupation combinations,** helping lay a solid foundation for downstream modeling and policy-relevant interpretation.

We begin our analysis by importing the core Python libraries required for **data handling, numerical computation, visualization,** and **directory management:**

- pandas: Enables efficient manipulation, filtering, and aggregation of structured tabular data, forming the backbone of our analysis pipeline.
- numpy: Provides support for fast numerical operations, array-based computation, and statistical routines.
- os: Facilitates interaction with the file system, allowing us to construct flexible and portable directory paths for data and output management.
- plotly.express: A high-level graphing library that enables the creation of interactive, publication-quality visualizations, which we use extensively to uncover patterns and present insights throughout the notebook.

```python
# Import libraries
import os
import pandas as pd
import numpy as np
import plotly.express as px
```

Define and Create Directory Paths

To ensure reproducibility and organized storage, we programmatically create directories if they don't already exist for:

- **raw data**
- **processed data**
- **results**
- **documentation**

These directories will store intermediate and final outputs for reproducibility.

```
# get woorking directory
Current_dir = os.getcwd()
# Go one directoty up to the root directory
project_root_dir = os.path.dirname(Current_dir)
# Define paths to the data files
data_dir = os.path.join(project_root_dir, 'data')
raw_dir = os.path.join (data_dir,  'raw')
processed_dir = os.path.join(data_dir,'processed')
# Define paths to results folder
results_dir = os.path.join(project_root_dir,'results')
# define paths to docs folder
docs_dir = os.path.join(project_root_dir, 'docs')

# create directories if they do not exist
os.makedirs(raw_dir, exist_ok = True)
os.makedirs(processed_dir, exist_ok = True)
os.makedirs(results_dir, exist_ok = True)
os.makedirs(docs_dir, exist_ok = True)
```

Loading the Cleaned Dataset

We load the cleaned version of the UCI Adult Income Dataset from the processed data directory into a Pandas DataFrame. The `head(10)` function shows the first ten records, giving a glimpse into the data columns such as `age`, `workclass`, `education_num`, etc.

```
adult_data_filename = os.path.join(processed_dir, "adult_cleaned.csv")
adult_df = adult_df = pd.read_csv(adult_data_filename)
adult_df.head(10)
```

	age	workclass	fnlwgt	education_num	marital_status	relationship	race	sex
0	39	state-gov	77516	13	single	single	white	male
1	50	self-employment	83311	13	married	male-spouse	white	male
2	38	private	215646	9	divorced or separated	single	white	male
3	53	private	234721	7	married	male-spouse	black	male
4	28	private	338409	13	married	female-spouse	black	female
5	37	private	284582	14	married	female-spouse	white	female
6	49	private	160187	5	divorced or separated	single	black	female
7	52	self-employment	209642	9	married	male-spouse	white	male
8	31	private	45781	14	single	single	white	female
9	42	private	159449	13	married	male-spouse	white	male

Check the shape of the dataset and datatype

Here, we examine the structure of the dataset:

- There are *32,513* entries and *16* variables.
- The dataset includes both **numerical** (e.g., `age`, `hours_per_week`) and **categorical** variables (e.g., `sex`, `education_level`).

Understanding data types and null entries is essential before proceeding with analysis.

```
adult_df.shape
```

```
(32514, 16)
```

```
adult_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32514 entries, 0 to 32513
Data columns (total 16 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   age              32514 non-null  int64
 1   workclass        32514 non-null  object
 2   fnlwgt           32514 non-null  int64
 3   education_num    32514 non-null  int64
 4   marital_status   32514 non-null  object
 5   relationship     32514 non-null  object
 6   race             32514 non-null  object
```

```
 7   sex               32514 non-null  object
 8   capital_num       32514 non-null  int64
 9   capital_loss      32514 non-null  int64
 10  hour_per_week     32514 non-null  int64
 11  income            32514 non-null  object
 12  education_level   32514 non-null  object
 13  occupation_group  32514 non-null  object
 14  native_region     32514 non-null  object
 15  age_group         32514 non-null  object
dtypes: int64(6), object(10)
memory usage: 4.0+ MB
```

Summary Statistics: Numerical Variables

This summary provides a snapshot of key distribution characteristics. We see that:

- Age ranges from 17 to 90, with a mean of 38.6 years. It is slightly right-skewed (positively skewed). While the average age is approximately 38.6 years, an examination of the percentiles reveals that the majority of individuals are clustered in the younger to middle-age range, with fewer observations in the older age brackets. This skewed age distribution might suggest labor force participation is concentrated in specific age groups, which could reflect broader demographic or economic realities.
- Capital gains/losses are highly skewed, with most values at 0 (the 75th percentile is 0). This indicates that a small number of individuals report very large gains or losses, especially evident in the capital gain variable which reaches up to $99,999. These variables act as proxies for wealth-related income that goes beyond regular wages or salaries. Individuals with non-zero values for capital gains or losses often represent a distinct socioeconomic subset of the population — typically more financially literate, or with access to investment assets. The stark inequality in their distributions mirrors real-world disparities in asset ownership and investment returns.
- The dataset has individuals working anywhere from 1 to 99 hours per week, with a median of 40. This aligns with the standard full-time work week in many countries (8 hours per day for 5 working days). The mean is slightly above that at 40.4 hours, suggesting a mild right skew, with a small subset of individuals working significantly longer hours. The mode is also 40, further reinforcing the prevalence of full-time work. A non-trivial number of individuals report working very few hours, possibly due to part-time work, unemployment, or semi-retirement. On the other extreme, some report working more than 45 hours per week, which may indicate multiple jobs, weekend-work, self-employment, or informal labor, and could reflect socio economicecessity.

```
adult_df.describe()
```

	age	fnlwgt	education_num	capital_num	capital_loss	hour_per_week
count	32514.000000	3.251400e+04	32514.000000	32514.000000	32514.000000	32514.000000
mean	38.589746	1.897964e+05	10.081626	1079.206619	87.430030	40.440949
std	13.639033	1.055780e+05	2.571975	7390.514416	403.237687	12.349994
min	17.000000	1.228500e+04	1.000000	0.000000	0.000000	1.000000
25%	28.000000	1.178330e+05	9.000000	0.000000	0.000000	40.000000
50%	37.000000	1.783630e+05	10.000000	0.000000	0.000000	40.000000
75%	48.000000	2.370615e+05	12.000000	0.000000	0.000000	45.000000
max	90.000000	1.484705e+06	16.000000	99999.000000	4356.000000	99.000000

Categorical Variables

```python
adult_df.describe(include="object")
```

	workclass	marital_status	relationship	race	sex	income	education_level	occ
count	32514	32514	32514	32514	32514	32514	32514	325
unique	8	4	5	5	2	2	8	5
top	private	married	male-spouse	white	male	<=50k	secondary-school graduate	wh
freq	22650	14984	13178	27772	21758	24678	10484	165

```python
adult_df['workclass'].value_counts()
```

```
workclass
private            22650
self-employment     3656
local-gov           2093
unknown             1836
state-gov           1298
government           960
voluntary             14
unemployment           7
Name: count, dtype: int64
```

```python
adult_df['workclass'].value_counts(normalize=True)
```

```
workclass
```

```
private           0.696623
self-employment   0.112444
local-gov         0.064372
unknown           0.056468
state-gov         0.039921
government        0.029526
voluntary         0.000431
unemployment      0.000215
Name: proportion, dtype: float64
```

```python
adult_df['marital_status'].value_counts(normalize=True)
```

```
marital_status
married                 0.460848
single                  0.327705
divorced or separated   0.180907
widowed                 0.030541
Name: proportion, dtype: float64
```

```python
adult_df['relationship'].value_counts(normalize=True)
```

```
relationship
male-spouse         0.405302
single              0.360706
own-child           0.155595
female-spouse       0.048225
extended-relative   0.030172
Name: proportion, dtype: float64
```

```python
adult_df['marital_status'].value_counts(normalize=True)
```

```
marital_status
married                 0.460848
single                  0.327705
divorced or separated   0.180907
widowed                 0.030541
Name: proportion, dtype: float64
```

```python
adult_df['race'].value_counts(normalize=True)
```

```
race
white                         0.854155
black                         0.096020
asian or pacific islander     0.031925
american indian or eskimo     0.009565
other                         0.008335
Name: proportion, dtype: float64
```

Income Distribution

Given that `income` is the target variable, most of the analysis hereafter will be based on it. We first of all examine the income distribution in the dataset.

This pie chart visualizes the overall income split: 76% of individuals earn 50K, while 24% earn >50K. This means that nearly 3 out of 4 individuals fall into the lower income bracket (<=50K). This shows that there is a significant imbalance.

```python
adult_df_Income = adult_df.groupby('income').size().reset_index(name='total')
adult_df_Income
```

	income	total
0	<=50k	24678
1	>50k	7836

```python
fig = px.pie(adult_df_Income, names='income',values='total', title='Overall Income Distribut
fig.show()
```

```
Unable to display output for mime type(s): application/vnd.plotly.v1+json, text/html
```

Income by Age Group

The bar chart visualizes the income distribution across age groups, using percentages within each group. There is an evident pattern in terms of income progression over the years with a gradual increase in terms of the number of people earning >50K starting from 0 amongst those aged 18 and below, peaking between 36 and 60 years, then declining after 60 years but not to zero.

All individuals under 18 earn <=50K, likely due to being students, minors, or ineligible for full-time employment. Extremely few young adults (2.1%) exceed 50K, as most are early in their careers, pursuing education, or in entry-level jobs. For the 26-35 age group, there's a noticeable improvement — roughly 1 in 5 individuals in this group earn >50K, reflecting early career progression and accumulation of qualifications/experience. A substantial income increase is seen in the 36-45 age group: over a third now earn >50K. This is typically considered prime earning age where individuals settle into stable, higher-paying positions. Highest proportion of >50K earners is seen amongst individuals aged between 46 and 60— nearly 4 in 10. This reflects career maturity, peak seniority levels, and accumulated experience. There's a drop-off in high incomes as many transition to retirement, part-time, or less demanding roles in the age group 61-75. Yet about 1 in 4 still earn >50K. Most in 76+ age group earn <=50K, likely due to retirement, pensions, or fixed incomes — but a small minority still earn higher incomes, possibly through continued work or investments.

```
adult_df_Income_age = adult_df.groupby(['age_group','income']).size().reset_index(name='total
adult_df_Income_age
```

	age_group	income	total_by_age
0	18-25	<=50k	5334
1	18-25	>50k	114
2	26-35	<=50k	6910
3	26-35	>50k	1591
4	36-45	<=50k	5230
5	36-45	>50k	2771
6	46-60	<=50k	4479
7	46-60	>50k	2809
8	61-75	<=50k	1580
9	61-75	>50k	511
10	76+	<=50k	200
11	76+	>50k	40
12	<18	<=50k	945

```
total_per_group = adult_df_Income_age.groupby('age_group')['total_by_age'].transform('sum')
adult_df_Income_age['percentage'] = (adult_df_Income_age['total_by_age']/total_per_group)*10
adult_df_Income_age
```

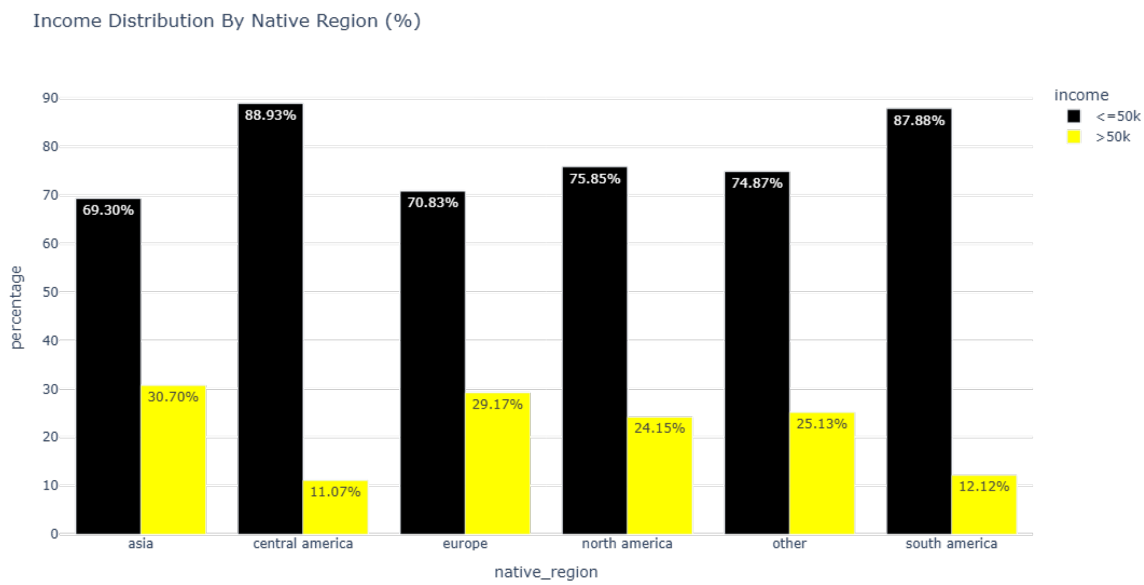
	age_group	income	total_by_age	percentage
0	18-25	<=50k	5334	97.907489
1	18-25	>50k	114	2.092511

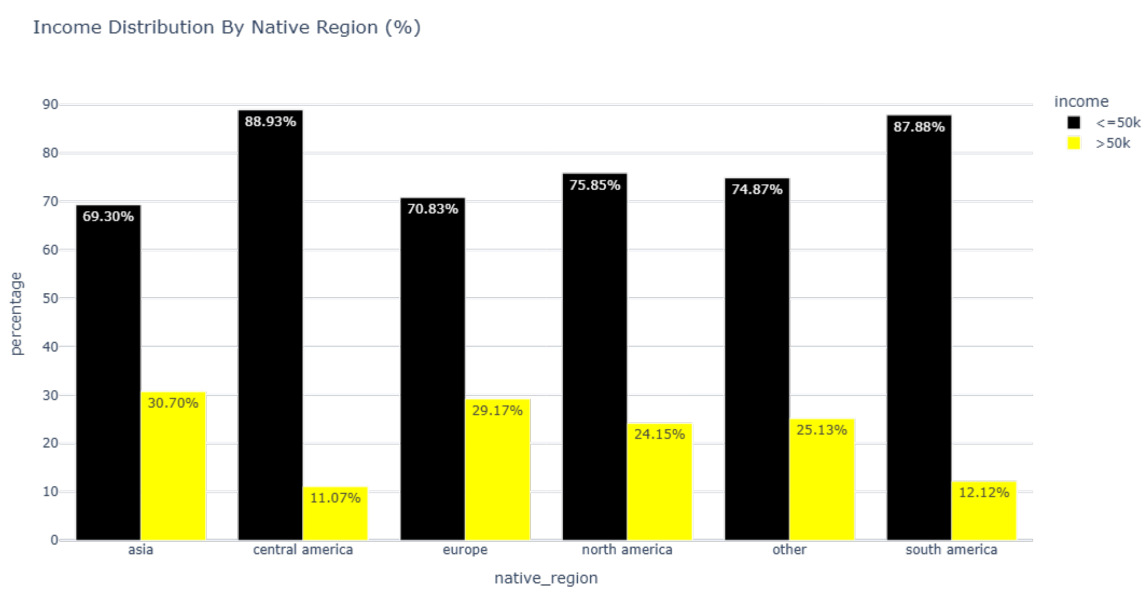
	age_group	income	total_by_age	percentage
2	26-35	<=50k	6910	81.284555
3	26-35	>50k	1591	18.715445
4	36-45	<=50k	5230	65.366829
5	36-45	>50k	2771	34.633171
6	46-60	<=50k	4479	61.457190
7	46-60	>50k	2809	38.542810
8	61-75	<=50k	1580	75.561932
9	61-75	>50k	511	24.438068
10	76+	<=50k	200	83.333333
11	76+	>50k	40	16.666667
12	<18	<=50k	945	100.000000

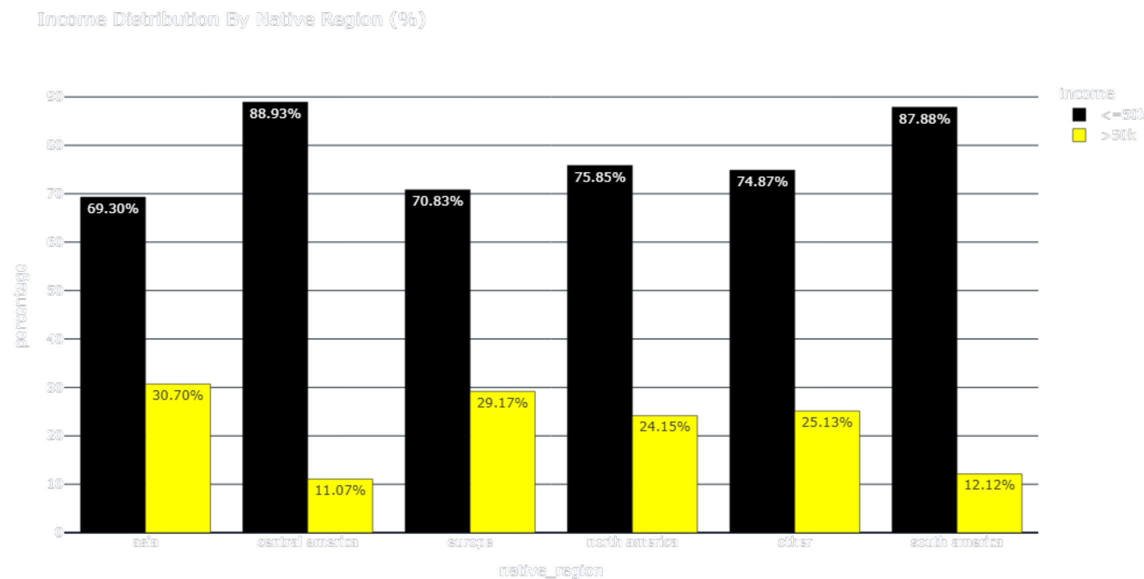
```
fig = px.bar(
    adult_df_Income_age,
    x = 'age_group',
    y = 'percentage',
    color = 'income',
    title='Income Distribution by Age Group(%)',
    barmode='group',
    color_discrete_sequence=px.colors.sequential.RdBu,
    text='percentage'
)
fig.update_traces(texttemplate = '%{text:.2f}%')
fig.show()
```

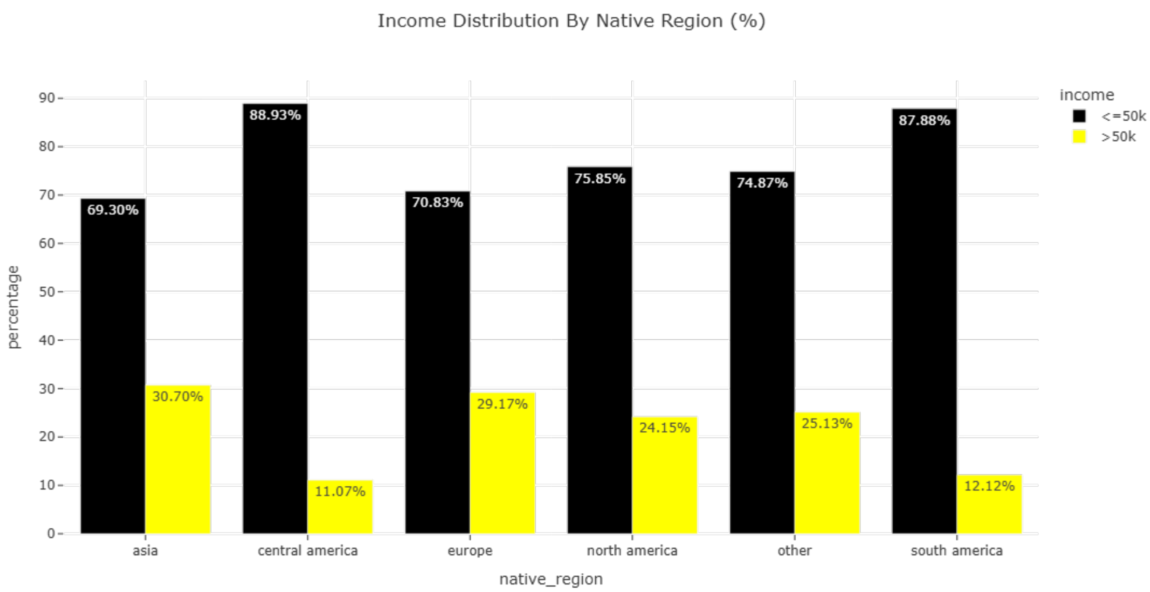
```
Unable to display output for mime type(s): application/vnd.plotly.v1+json, text/html
```

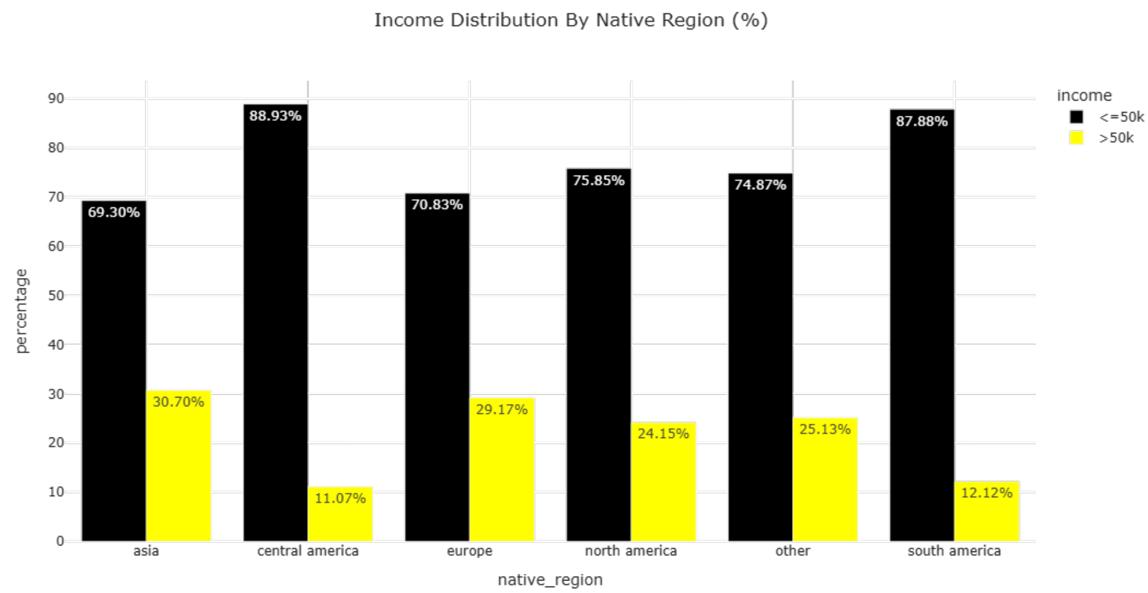
```
themes = ["plotly", "plotly_white", "plotly_dark", "ggplot2", "seaborn", "simple_white", "pre

for theme in themes:
    fig.update_layout(template=theme)

    fig.show()
```

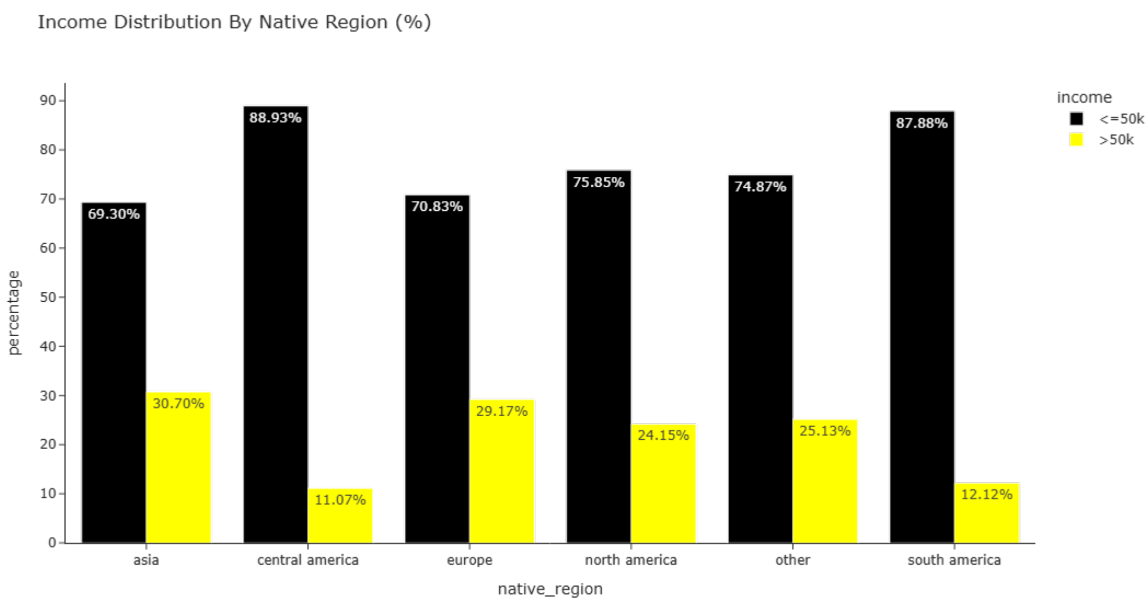

Income Distribution By Native Region (%)
income
<=50k
>50k
percentage
90
80
70
60
50
40
30
20
10
0
69.30%
30.70%
88.93%
11.07%
70.83%
29.17%
75.85%
24.15%
74.87%
25.13%
87.88%
12.12%
asia
central america
europe
north america
other
south america
native_region

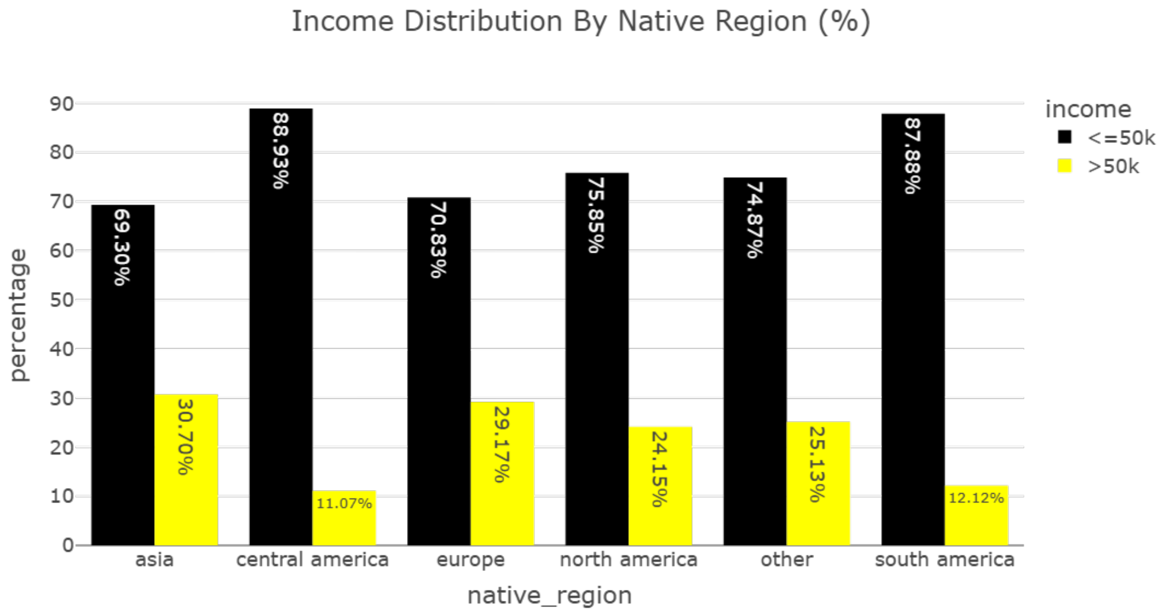

Income Distribution By Native Region (%)
income
<=50k
>50k
percentage
90
80
70
60
50
40
30
20
10
0
69.30%
30.70%
88.93%
11.07%
70.83%
29.17%
75.85%
24.15%
74.87%
25.13%
87.88%
12.12%
asia
central america
europe
north america
other
south america
native_region

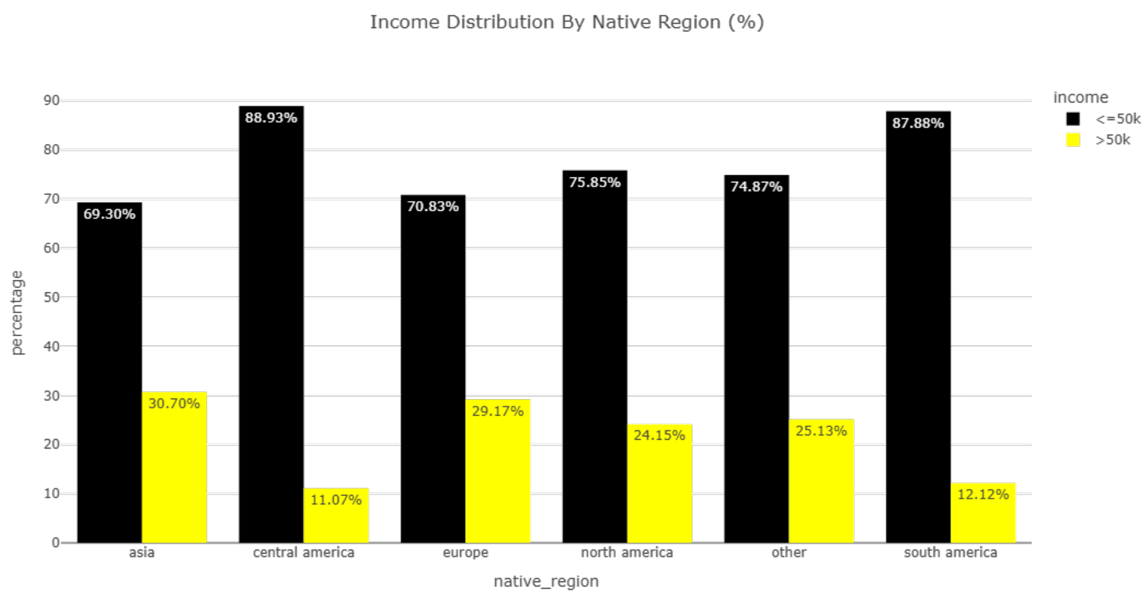

Income Distribution By Native Region (%)
income
<=50k
>50k
percentage
native_region
asia
central america
europe
north america
other
south america
69.30%
30.70%
88.93%
11.07%
70.83%
29.17%
75.85%
24.15%
74.87%
25.13%
87.88%
12.12%

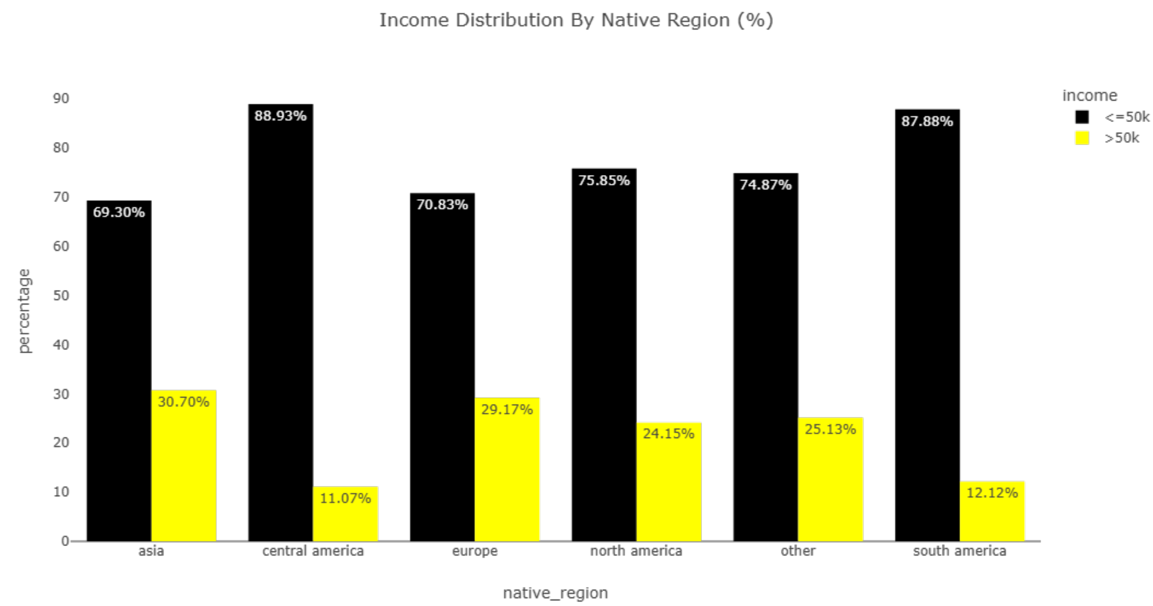

Income Distribution By Native Region (%)
income
<=50k
>50k
percentage
native_region
asia
central america
europe
north america
other
south america
69.30%
30.70%
88.93%
11.07%
70.83%
29.17%
75.85%
24.15%
74.87%
25.13%
87.88%
12.12%


Income Distribution By Native Region (%)
income
<=50k
>50k
percentage
native_region
asia
central america
europe
north america
other
south america
69.30%
30.70%
88.93%
11.07%
70.83%
29.17%
75.85%
24.15%
74.87%
25.13%
87.88%
12.12%

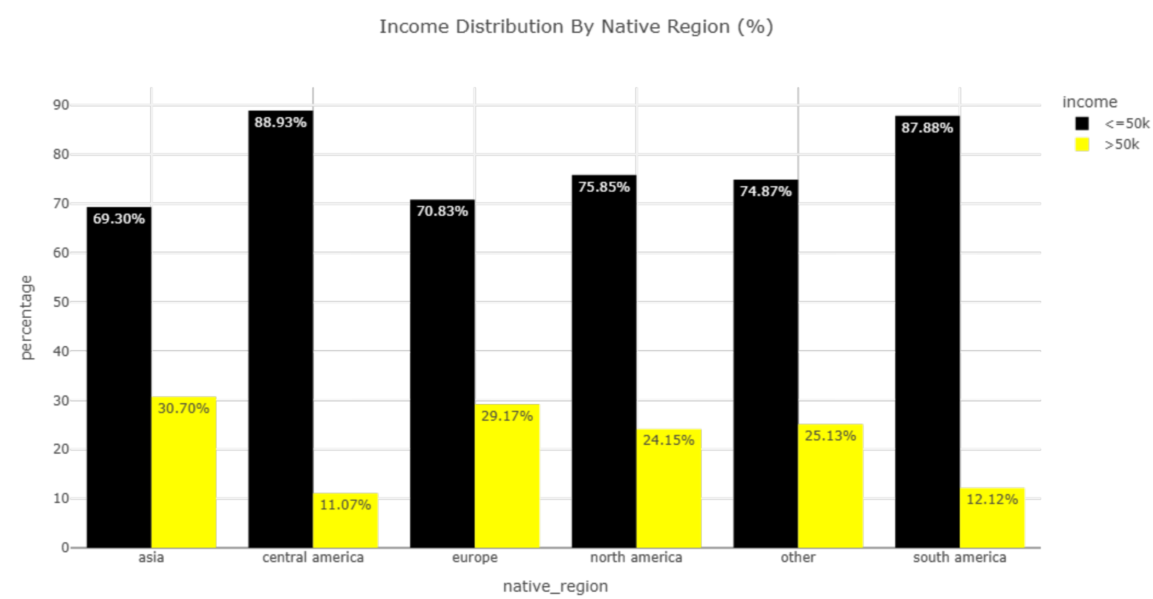

Income Distribution By Native Region (%)
income
<=50k
>50k
percentage
native_region
asia
central america
europe
north america
other
south america
69.30%
30.70%
88.93%
11.07%
70.83%
29.17%
75.85%
24.15%
74.87%
25.13%
87.88%
12.12%


Income Distribution By Native Region (%)
income
<=50k
>50k
percentage
90
80
70
60
50
40
30
20
10
0
69.30%
30.70%
88.93%
11.07%
70.83%
29.17%
75.85%
24.15%
74.87%
25.13%
87.88%
12.12%
asia
central america
europe
north america
other
south america
native_region

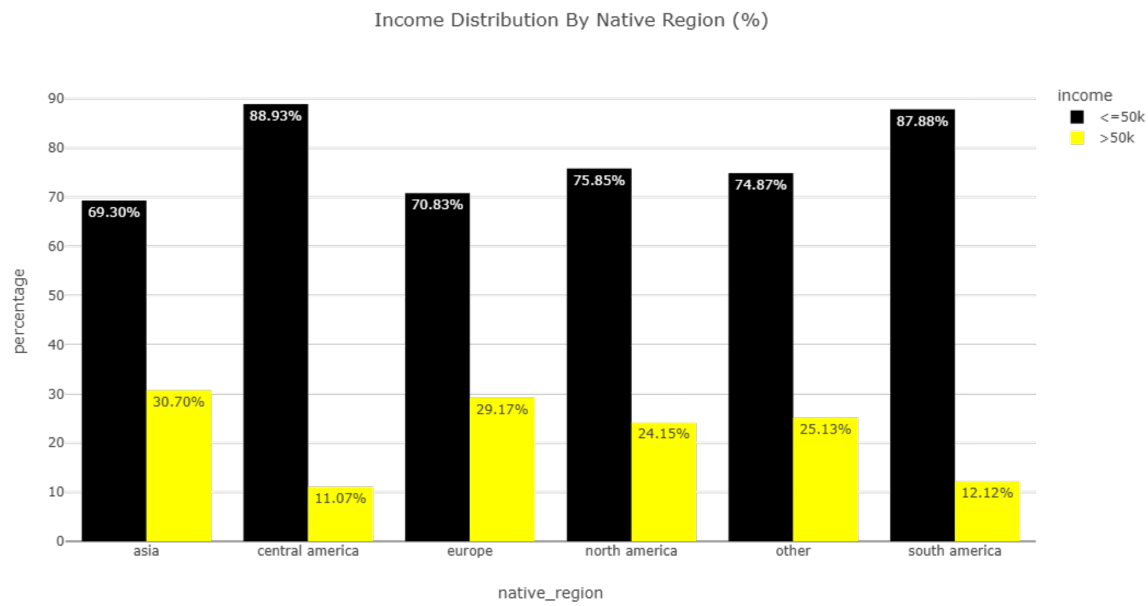

Income Distribution By Native Region (%)
income
<=50k
>50k
percentage
90
80
70
60
50
40
30
20
10
0
69.30%
30.70%
88.93%
11.07%
70.83%
29.17%
75.85%
24.15%
74.87%
25.13%
87.88%
12.12%
asia
central america
europe
north america
other
south america
native_region



Income Distribution By Native Region (%)
income
<=50k
>50k
percentage
native_region
asia
central america
europe
north america
other
south america
69.30%
30.70%
88.93%
11.07%
70.83%
29.17%
75.85%
24.15%
74.87%
25.13%
87.88%
12.12%



Income Distribution By Native Region (%)
income
<=50k
>50k
percentage
native_region
asia
central america
europe
north america
other
south america
69.30%
30.70%
88.93%
11.07%
70.83%
29.17%
75.85%
24.15%
74.87%
25.13%
87.88%
12.12%




```
adult_df_income_native_region = adult_df.groupby(['native_region', 'income']).size().reset_i
adult_df_income_native_region
```

	native_region	income	total_income_distr
0	asia	<=50k	465
1	asia	>50k	206
2	central america	<=50k	466
3	central america	>50k	58
4	europe	<=50k	369
5	europe	>50k	152
6	north america	<=50k	22769
7	north america	>50k	7250
8	other	<=50k	435
9	other	>50k	146
10	south america	<=50k	174
11	south america	>50k	24

Asia (30.7%) and Europe (29.2%) have the highest proportions of high-income earners. This suggests these immigrant groups might be better integrated into high-paying professional roles, or may represent a more skilled migrant profile in the dataset. Central America (11.1%) and South America (12.1%) have the lowest proportions of >50K earners. With 24.2% of North

Americans earning >50K, this serves as a middle-ground baseline. Interestingly, both Asian and European groups outperform the native-born population proportionally in high-income brackets. The 'Other' group sits around 25.1%, close to North America's rate. This likely reflects a diverse mix of regions not explicitly listed.

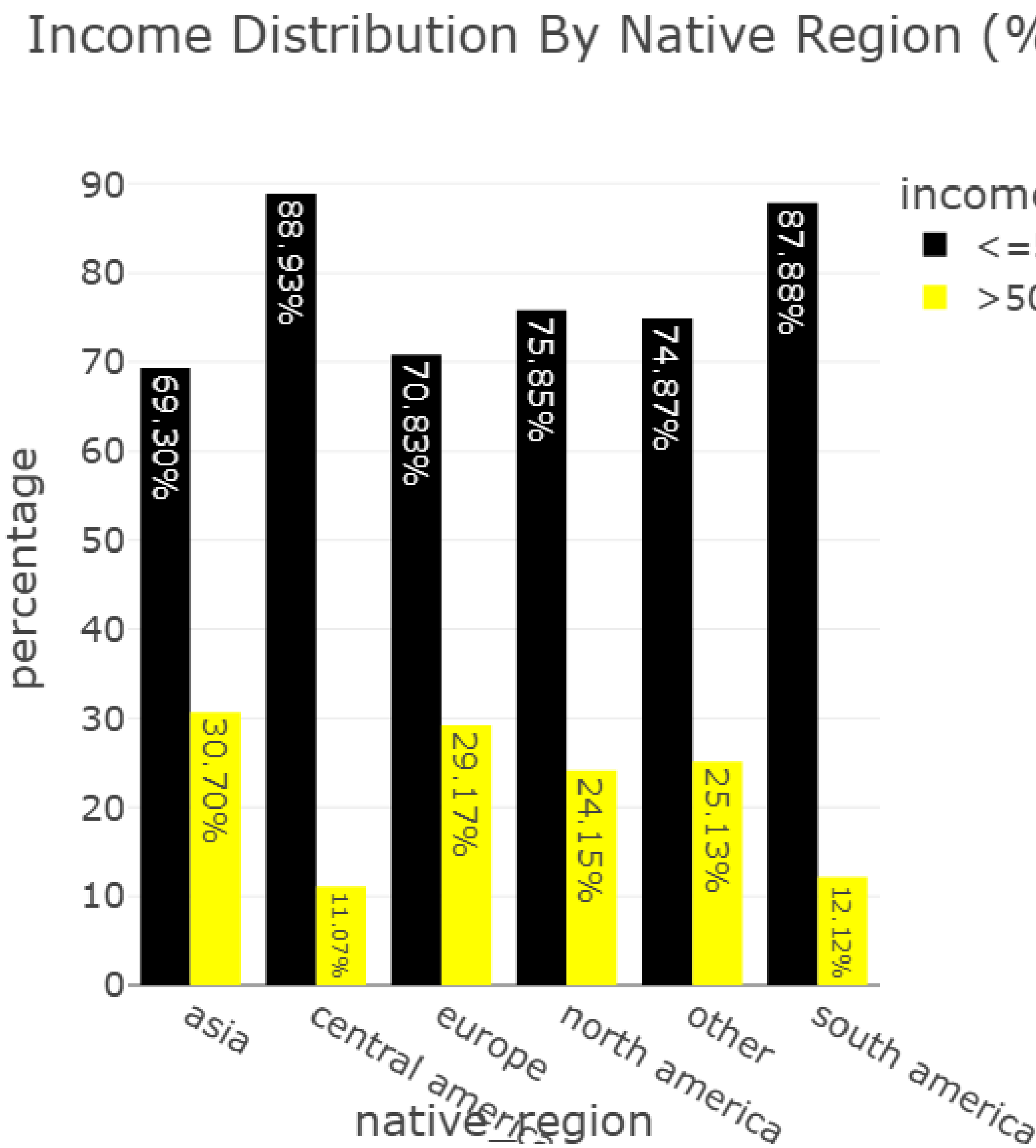
```
adult_df_income_native_region = adult_df.groupby(['native_region', 'income']).size().reset_i
adult_df_income_native_region
```

	native_region	income	total_income_distr
0	asia	<=50k	465
1	asia	>50k	206
2	central america	<=50k	466
3	central america	>50k	58
4	europe	<=50k	369
5	europe	>50k	152
6	north america	<=50k	22769
7	north america	>50k	7250
8	other	<=50k	435
9	other	>50k	146
10	south america	<=50k	174
11	south america	>50k	24

```
total_per_region = adult_df_income_native_region.groupby('native_region')['total_income_dist
adult_df_income_native_region['percentage'] = (adult_df_income_native_region['total_income_d
adult_df_income_native_region
```

	native_region	income	total_income_distr	percentage
0	asia	<=50k	465	69.299553
1	asia	>50k	206	30.700447
2	central america	<=50k	466	88.931298
3	central america	>50k	58	11.068702
4	europe	<=50k	369	70.825336
5	europe	>50k	152	29.174664
6	north america	<=50k	22769	75.848629
7	north america	>50k	7250	24.151371
8	other	<=50k	435	74.870912
9	other	>50k	146	25.129088
10	south america	<=50k	174	87.878788
11	south america	>50k	24	12.121212

```
import plotly.express as px

fig = px.bar(
    adult_df_income_native_region,
    x='native_region',
    y='percentage',
    color='income',
    title='Income Distribution By Native Region (%)',
    barmode='group',
    color_discrete_sequence=['black', 'yellow'],
    text='percentage',
    width=700,
    height=600,
)
fig.update_traces(texttemplate='%{text:.2f}%')
fig.update_layout(template= 'presentation',paper_bgcolor= "rgba(0,0,0,0)",plot_bgcolor = "rgb
fig.write_image(os.path.join(results_dir,'income_distribution_bar_plot.jpg'))
fig.write_image(os.path.join(results_dir,'income_distribution_bar_plot.png'))
fig.write_html(os.path.join(results_dir,'income_distribution_bar_plot.html'))
fig.show()
```



Asian or Pacific Islander (26.6%) and White (25.6%) populations have the highest proportions of >50K earners. Asians/Pacific Islanders marginally outperform Whites, a pattern often attributed to occupational concentration in high-paying sectors like technology and medicine. On the other hand, American Indian or Eskimo (11.6%), Black (12.4%), and Other (9.2%) groups show significantly lower rates of high-income earners. These figures reflect long-standing economic disparities rooted in historical exclusion, occupational segregation, and systemic inequality.

```
adult_df_income_race = adult_df.groupby(['race', 'income']).size().reset_index(name='total_in
adult_df_income_race
```

	race	income	total_income_distr
0	american indian or eskimo	<=50k	275
1	american indian or eskimo	>50k	36
2	asian or pacific islander	<=50k	762
3	asian or pacific islander	>50k	276
4	black	<=50k	2735
5	black	>50k	387
6	other	<=50k	246
7	other	>50k	25
8	white	<=50k	20660
9	white	>50k	7112

```
total_per_race= adult_df_income_race.groupby('race')['total_income_distr'].transform('sum')
adult_df_income_race['percentage'] = (adult_df_income_race['total_income_distr']/total_per_ra
adult_df_income_race
```

	race	income	total_income_distr	percentage
0	american indian or eskimo	<=50k	275	88.424437
1	american indian or eskimo	>50k	36	11.575563
2	asian or pacific islander	<=50k	762	73.410405
3	asian or pacific islander	>50k	276	26.589595
4	black	<=50k	2735	87.604100
5	black	>50k	387	12.395900
6	other	<=50k	246	90.774908
7	other	>50k	25	9.225092
8	white	<=50k	20660	74.391473
9	white	>50k	7112	25.608527

```
fig=px.bar(adult_df_income_race,
           x='race',
           y='percentage',
           color='income',
           title='Income Distribution by Race',
           color_discrete_sequence=["black","yellow"],
           barmode='group',
```

```
            text='percentage'


)
fig.update_layout(template="presentation",
                  xaxis_title='Race',
                   yaxis_title='Percentage of population',
                   legend_title=dict(text='Income Level'),
                  paper_bgcolor="rgba(0,0,0,0)",plot_bgcolor=("rgba(0,0,0,0)"))
fig.update_traces(texttemplate='%{text:.2f}%',textposition='outside')
fig.show()
fig.write_image(os.path.join(results_dir,'income_distribution-Race-bar_chart.jpg'))
fig.write_image(os.path.join(results_dir,'income_distribution-Race_bar_chart.png'))
fig.write_html(os.path.join(results_dir,'income_distribution_Race_bar_chart.html'))
```

```
Unable to display output for mime type(s): application/vnd.plotly.v1+json, text/html
```

The stark differences in high-income proportions:

- **Between Whites and Blacks:** 25.6% vs 12.4% — slightly over double the proportion.
- **Between Asians and Others:** 26.6% vs 9.2% — nearly triple.

These disparities are consistent with well-documented wage gaps and underrepresentation of marginalized groups in higher-paying roles.

```
adult_df_income_edu_occ = adult_df.groupby(['education_level', 'occupation_group', 'income']
adult_df_income_edu_occ
```

	education_level	occupation_group	income	total
42	secondary-school graduate	blue collar	<=50k	3976
71	tertiary	white collar	>50k	3545
70	tertiary	white collar	<=50k	3369
60	some-college	white collar	<=50k	3004
50	secondary-school graduate	white collar	<=50k	2900
...	...	...	...	...
39	secondary	unknown	>50k	3
35	secondary	military	>50k	2
16	high school	unknown	>50k	2
14	high school	service	>50k	1
27	primary	service	>50k	1

From the bar chart, we can pick out the largest groups per income-level. We see that secondary-school graduates working a blue collar job occupy the largest group in the dataset (3976). This reflects a common socio-economic profile: individuals with basic schooling in manual or technical trades predominantly earning lower incomes. The largest high-income group are tertiary-educated individuals in white collar roles. This highlights the strong earning advantage conferred by higher education and skilled jobs.

```
adult_df_income_edu_occ['edu_occ']= (adult_df_income_edu_occ['education_level']+" | "
                                   +adult_df_income_edu_occ['occupation_group'])
adult_df_income_edu_occ
```

	education_level	occupation_group	income	total	edu_occ
42	secondary-school graduate	blue collar	<=50k	3976	secondary-school graduate | blue collar
71	tertiary	white collar	>50k	3545	tertiary | white collar
70	tertiary	white collar	<=50k	3369	tertiary | white collar
60	some-college	white collar	<=50k	3004	some-college | white collar
50	secondary-school graduate	white collar	<=50k	2900	secondary-school graduate | white collar
...	...	...	...	...	...
39	secondary	unknown	>50k	3	secondary | unknown
35	secondary	military	>50k	2	secondary | military
16	high school	unknown	>50k	2	high school | unknown
14	high school	service	>50k	1	high school | service
27	primary	service	>50k	1	primary | service

```
adult_df_income_edu_occ.head(15)
```

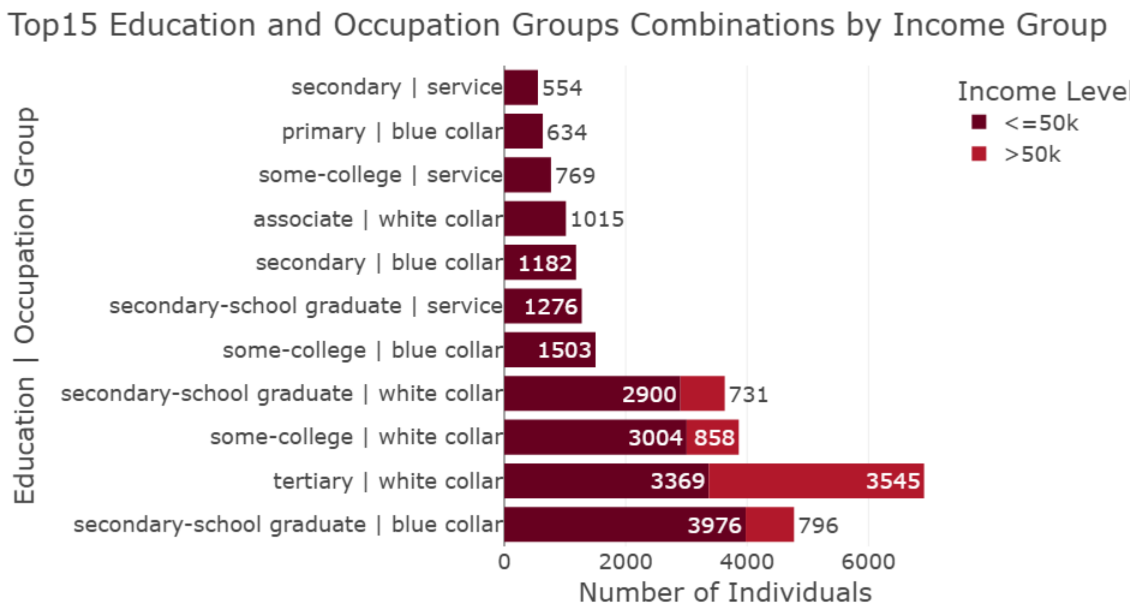
	education_level	occupation_group	income	total	edu_occ
42	secondary-school graduate	blue collar	<=50k	3976	secondary-school graduate | blue collar
71	tertiary	white collar	>50k	3545	tertiary | white collar
70	tertiary	white collar	<=50k	3369	tertiary | white collar
60	some-college	white collar	<=50k	3004	some-college | white collar
50	secondary-school graduate	white collar	<=50k	2900	secondary-school graduate | white collar
52	some-college	blue collar	<=50k	1503	some-college | blue collar
46	secondary-school graduate	service	<=50k	1276	secondary-school graduate | service
32	secondary	blue collar	<=50k	1182	secondary | blue collar
8	associate	white collar	<=50k	1015	associate | white collar
61	some-college	white collar	>50k	858	some-college | white collar
43	secondary-school graduate	blue collar	>50k	796	secondary-school graduate | blue collar
56	some-college	service	<=50k	769	some-college | service

<table>
<tr><th></th><th>education_level</th><th>occupation_group</th><th>income</th><th>total</th><th>edu_occ</th></tr>
<tr><td>51</td><td>secondary-school graduate</td><td>white collar</td><td>>50k</td><td>731</td><td>secondary-school graduate | white collar</td></tr>
<tr><td>23</td><td>primary</td><td>blue collar</td><td><=50k</td><td>634</td><td>primary | blue collar</td></tr>
<tr><td>36</td><td>secondary</td><td>service</td><td><=50k</td><td>554</td><td>secondary | service</td></tr>
</table>

```
fig=px
adult_df_income_edu_occ.head(15),
```

```
(              education_level occupation_group income  total  \
 42  secondary-school graduate      blue collar  <=50k   3976
 71                   tertiary     white collar   >50k   3545
 70                   tertiary     white collar  <=50k   3369
 60               some-college     white collar  <=50k   3004
 50  secondary-school graduate     white collar  <=50k   2900
 52               some-college      blue collar  <=50k   1503
 46  secondary-school graduate          service  <=50k   1276
 32                  secondary      blue collar  <=50k   1182
 8                   associate     white collar  <=50k   1015
 61               some-college     white collar   >50k    858
 43  secondary-school graduate      blue collar   >50k    796
 56               some-college          service  <=50k    769
 51  secondary-school graduate     white collar   >50k    731
 23                    primary      blue collar  <=50k    634
 36                  secondary          service  <=50k    554

                                      edu_occ
 42   secondary-school graduate | blue collar
 71                   tertiary | white collar
 70                   tertiary | white collar
 60               some-college | white collar
 50  secondary-school graduate | white collar
 52                some-college | blue collar
 46       secondary-school graduate | service
 32                  secondary | blue collar
 8                   associate | white collar
 61               some-college | white collar
 43   secondary-school graduate | blue collar
 56                    some-college | service
 51  secondary-school graduate | white collar
 23                    primary | blue collar
 36                      secondary | service  ,)
```

```
num= 15
adult_df_combos = adult_df_income_edu_occ.head(num)
fig = px.bar(
    adult_df_combos,
    x = 'total',
    y = 'edu_occ',
    color = 'income',
    orientation = 'h',
    title = f'Top{num} Education and Occupation Groups Combinations by Income Group',
    # barmode = 'group',
    height = 500,
    width=1100,
    color_discrete_sequence=px.colors.sequential.RdBu,
    text = 'total'
)

fig.update_layout(template="presentation", xaxis_title='Number of Individuals',
                  yaxis_title='Education | Occupation Group',
                  legend_title=dict(text='Income Level'),
                 margin=dict(l=450, r=50, t= 50, b=50))
fig.write_image(os.path.join(results_dir,'income_Distribution_by_nativeregion_bar_plot.jpg'))
fig.write_image(os.path.join(results_dir,'income_Distribution_by_nativeregion_bar_plot.png'))
fig.write_html(os.path.join(results_dir,'income_Distribution_by_nativeregion_bar_plot.html'))

fig.show()
```



Some of the key patterns we can get from the dataset are:

- **Education matters, but isn't deterministic** Tertiary education combined with white-collar work offers the highest income prospects. Yet a substantial number of tertiary-educated white-collar workers earn <=50K, likely early career, part-time, or structural pay gaps.
- **Blue-collar and service work predominantly pay <=50K, regardless of education.** Even some college education doesn't guarantee high incomes in these sectors. Manual and service sector income is highly occupation-dependent (some skilled trades can break the 50K mark).
- **Some non-tertiary education groups do reach >50K** Secondary-school graduates in blue-collar and white-collar work have decent representation among >50K earners. This reflects upward mobility possible through skilled trades, tenure, or niche roles.