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.

# 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 | capital\_num | capital\_loss | hour\_per\_week | income | education\_level | occupation\_group | native\_region | age\_group |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 39 | state-gov | 77516 | 13 | single | single | white | male | 2174 | 0 | 40 | <=50k | tertiary | white collar | north america | 36-45 |
| 1 | 50 | self-employment | 83311 | 13 | married | male-spouse | white | male | 0 | 0 | 13 | <=50k | tertiary | white collar | north america | 46-60 |
| 2 | 38 | private | 215646 | 9 | divorced or separated | single | white | male | 0 | 0 | 40 | <=50k | secondary-school graduate | blue collar | north america | 36-45 |
| 3 | 53 | private | 234721 | 7 | married | male-spouse | black | male | 0 | 0 | 40 | <=50k | secondary | blue collar | north america | 46-60 |
| 4 | 28 | private | 338409 | 13 | married | female-spouse | black | female | 0 | 0 | 40 | <=50k | tertiary | white collar | central america | 26-35 |
| 5 | 37 | private | 284582 | 14 | married | female-spouse | white | female | 0 | 0 | 40 | <=50k | tertiary | white collar | north america | 36-45 |
| 6 | 49 | private | 160187 | 5 | divorced or separated | single | black | female | 0 | 0 | 16 | <=50k | secondary | service | central america | 46-60 |
| 7 | 52 | self-employment | 209642 | 9 | married | male-spouse | white | male | 0 | 0 | 45 | >50k | secondary-school graduate | white collar | north america | 46-60 |
| 8 | 31 | private | 45781 | 14 | single | single | white | female | 14084 | 0 | 50 | >50k | tertiary | white collar | north america | 26-35 |
| 9 | 42 | private | 159449 | 13 | married | male-spouse | white | male | 5178 | 0 | 40 | >50k | tertiary | white collar | north america | 36-45 |

## 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

adult\_df.describe(include="object")

|  | workclass | marital\_status | relationship | race | sex | income | education\_level | occupation\_group | native\_region | age\_group |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| count | 32514 | 32514 | 32514 | 32514 | 32514 | 32514 | 32514 | 32514 | 32514 | 32514 |
| unique | 8 | 4 | 5 | 5 | 2 | 2 | 8 | 5 | 6 | 7 |
| top | private | married | male-spouse | white | male | <=50k | secondary-school graduate | white collar | north america | 26-35 |
| freq | 22650 | 14984 | 13178 | 27772 | 21758 | 24678 | 10484 | 16533 | 30019 | 8501 |

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

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

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

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

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

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.

adult\_df\_Income = adult\_df.groupby('income').size().reset\_index(name='total')  
adult\_df\_Income

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

fig = px.pie(adult\_df\_Income, names='income',values='total', title='Overall Income Distribution',color\_discrete\_sequence=['darkblue','skyblue'])  
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\_by\_age').sort\_values(['age\_group','income'])  
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 |

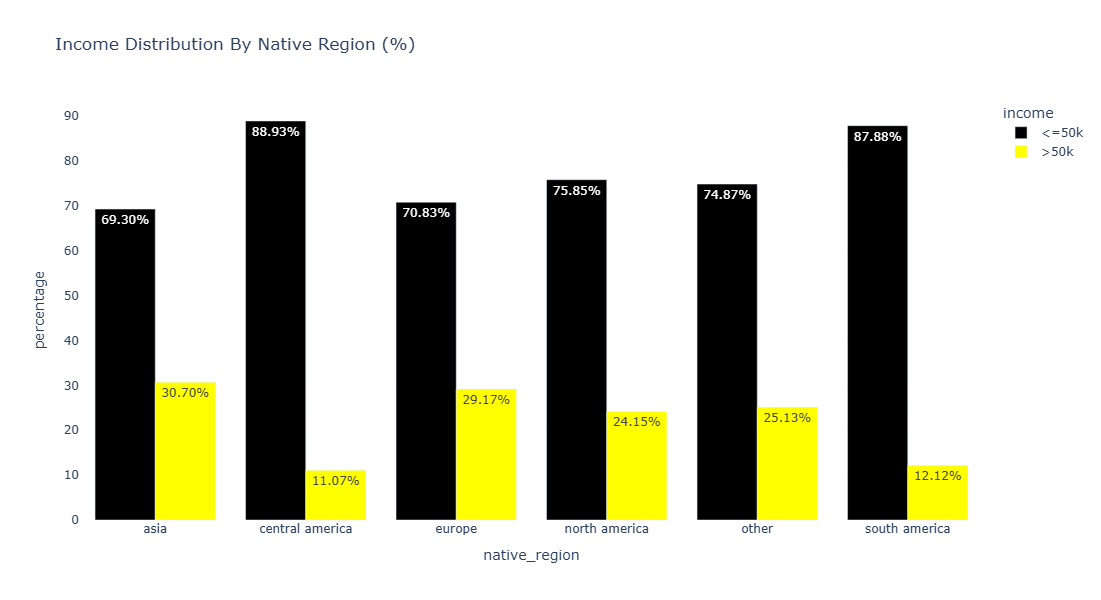
total\_per\_group = adult\_df\_Income\_age.groupby('age\_group')['total\_by\_age'].transform('sum')  
adult\_df\_Income\_age['percentage'] = (adult\_df\_Income\_age['total\_by\_age']/total\_per\_group)\*100  
adult\_df\_Income\_age

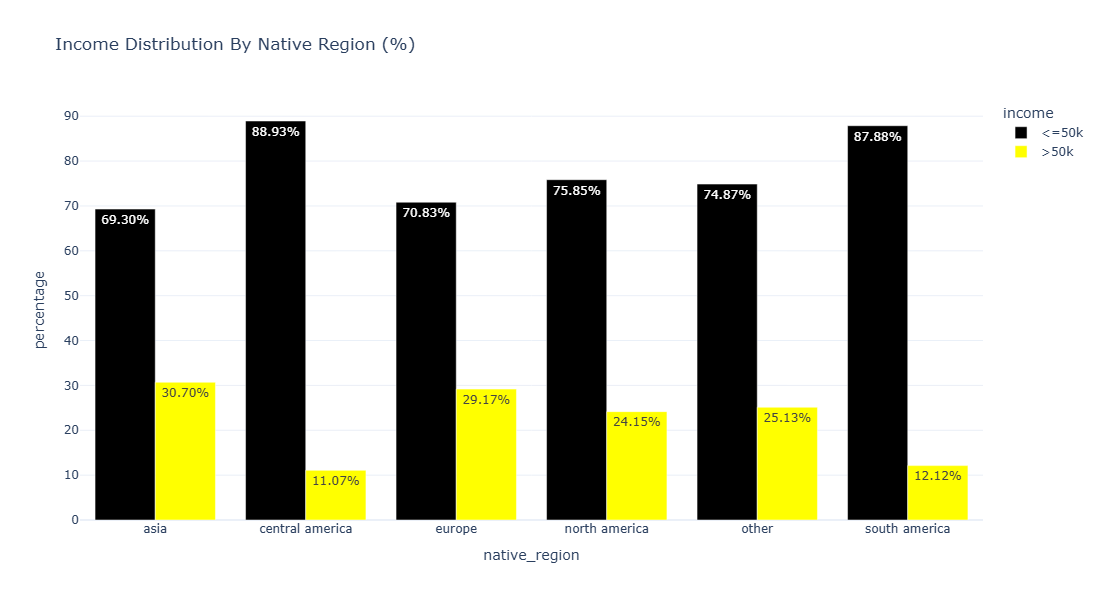
|  | age\_group | income | total\_by\_age | percentage |
| --- | --- | --- | --- | --- |
| 0 | 18-25 | <=50k | 5334 | 97.907489 |
| 1 | 18-25 | >50k | 114 | 2.092511 |
| 2 | 26-35 | <=50k | 6910 | 81.284555 |
| 3 | 26-35 | >50k | 1591 | 18.715445 |
| 4 | 36-45 | <=50k | 5230 | 65.366829 |
| 5 | 36-45 | >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 |

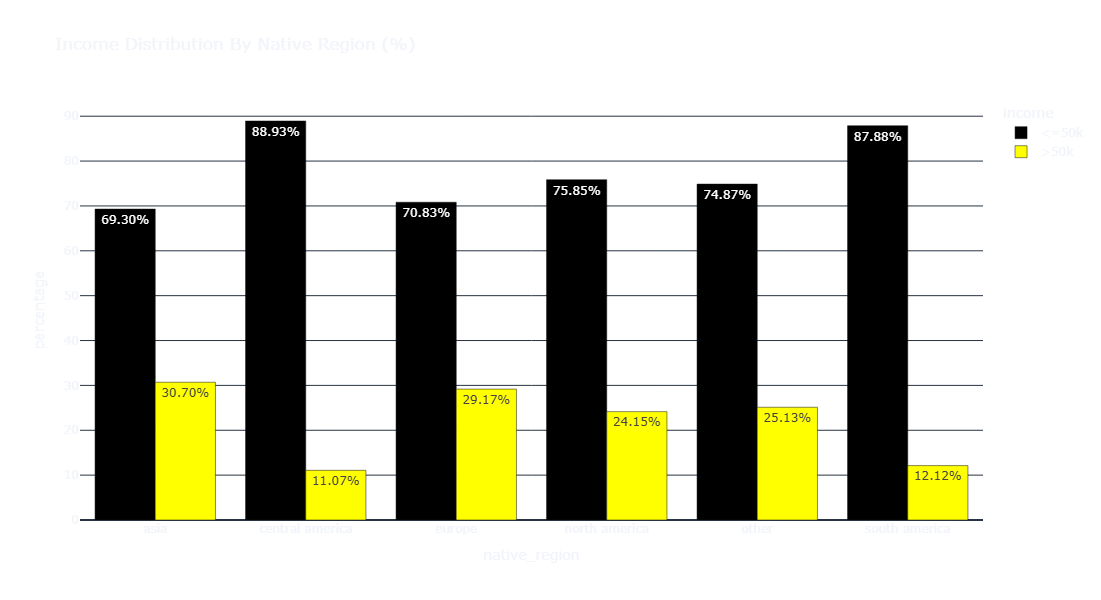
fig = px.bar(  
 adult\_df\_Income\_age,  
 x = 'age\_group',  
 y = 'percentage',  
 color = 'income',  
 title='Income Distribution by Age Group(%)',  
 barmode='group',  
 color\_discrete\_sequence=px.colors.sequential.RdBu,  
 text='percentage'  
)  
fig.update\_traces(texttemplate = '%{text:.2f}%')  
fig.show()

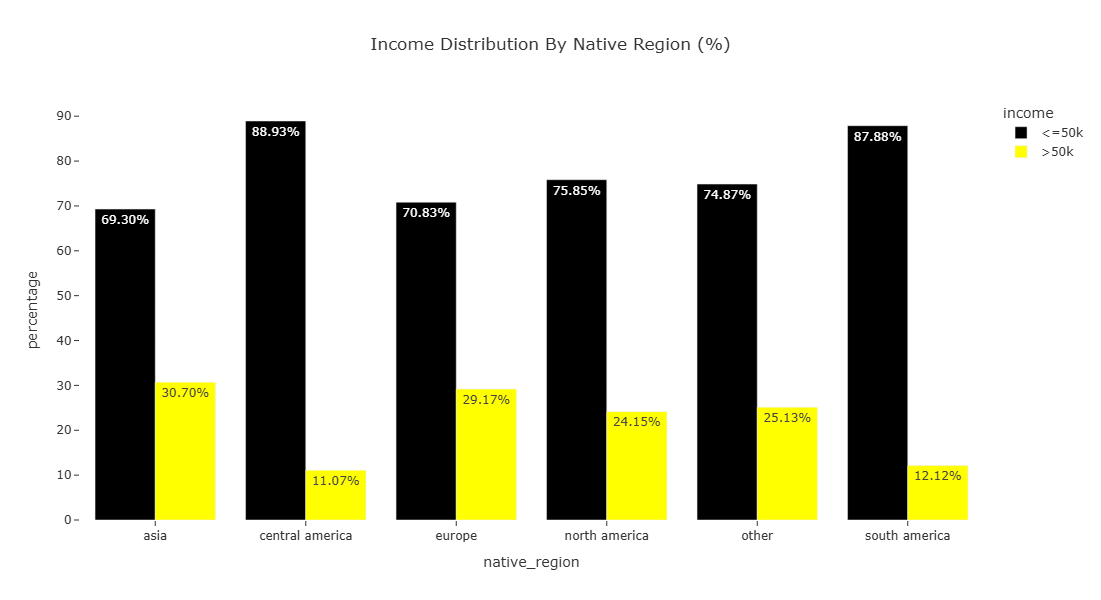
Unable to display output for mime type(s): application/vnd.plotly.v1+json, text/html

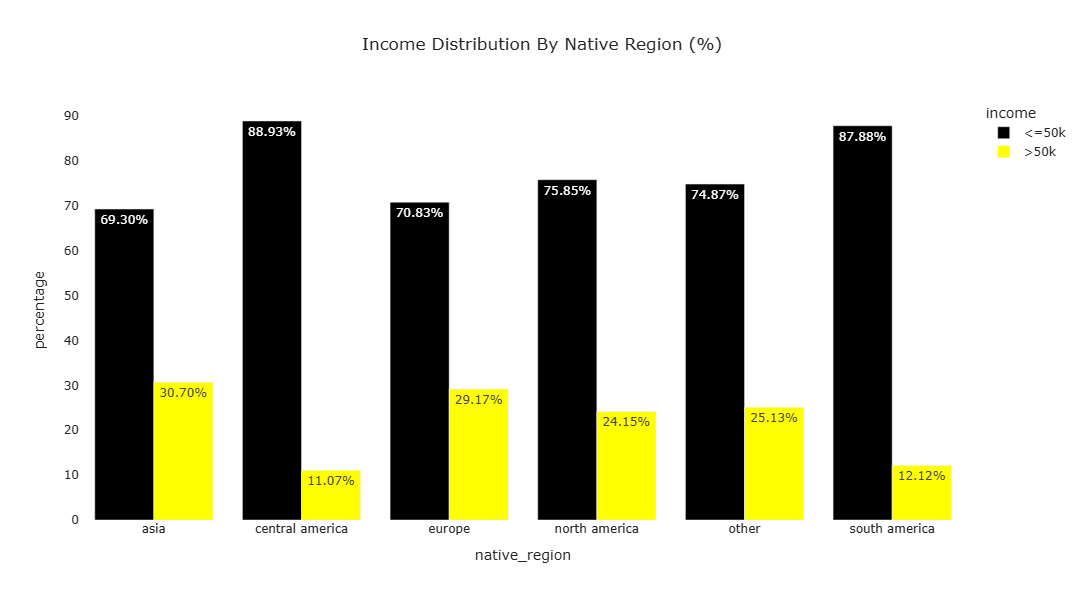
themes = ["plotly", "plotly\_white", "plotly\_dark", "ggplot2", "seaborn", "simple\_white", "presentation", "xgridoff", "ygridoff", "gridon", "none"]  
  
for theme in themes:  
 fig.update\_layout(template=theme)  
  
 fig.show()

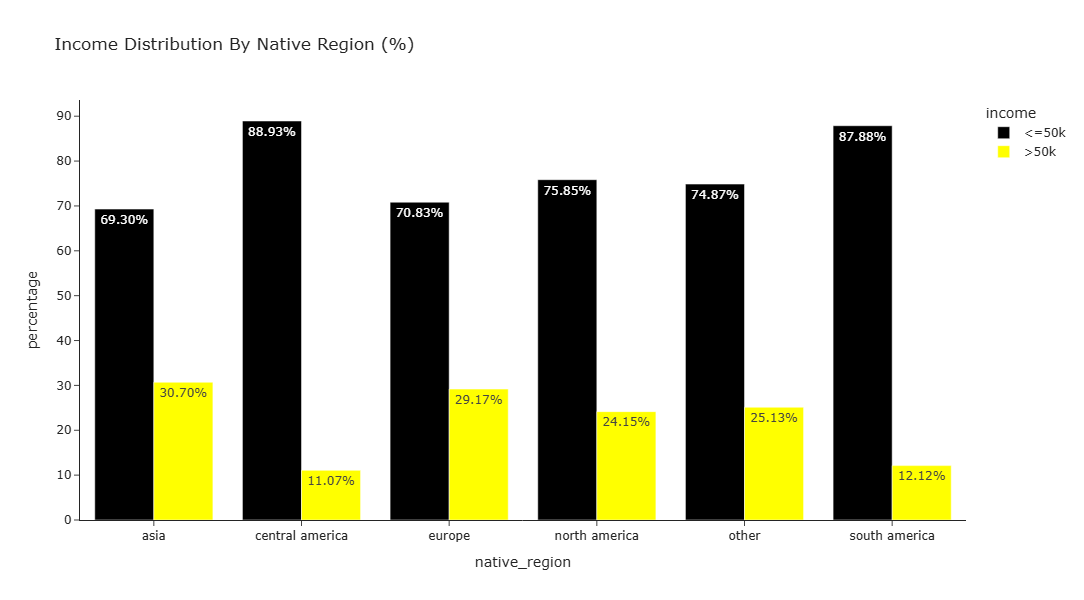


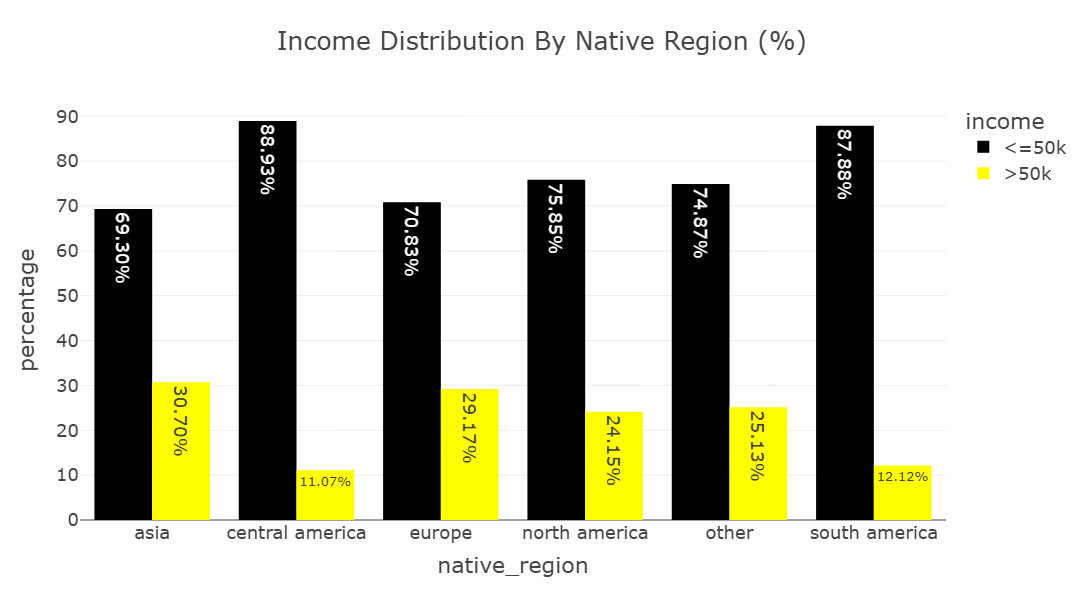


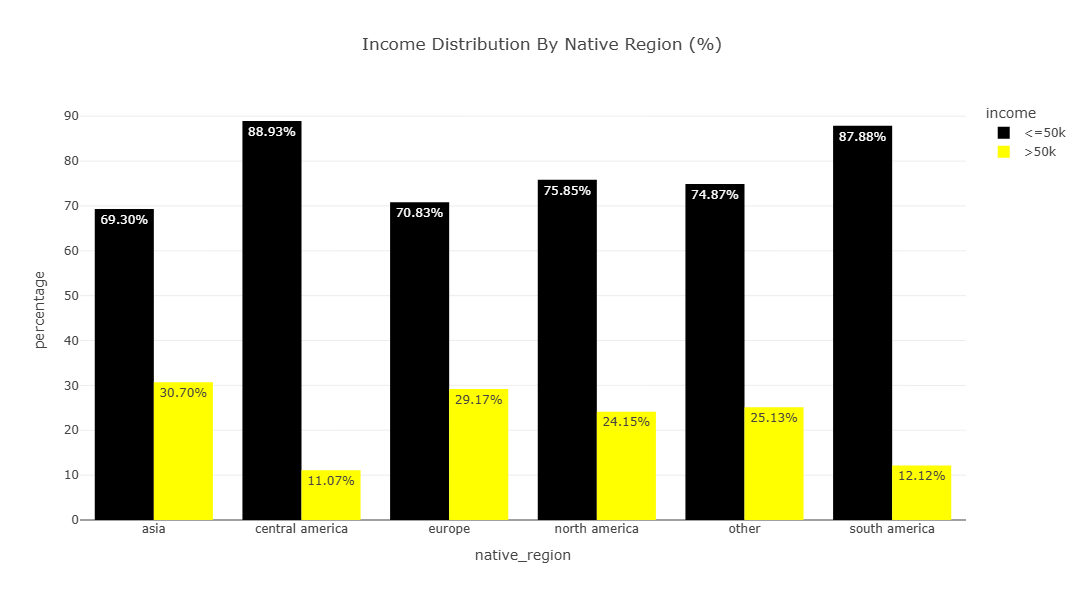


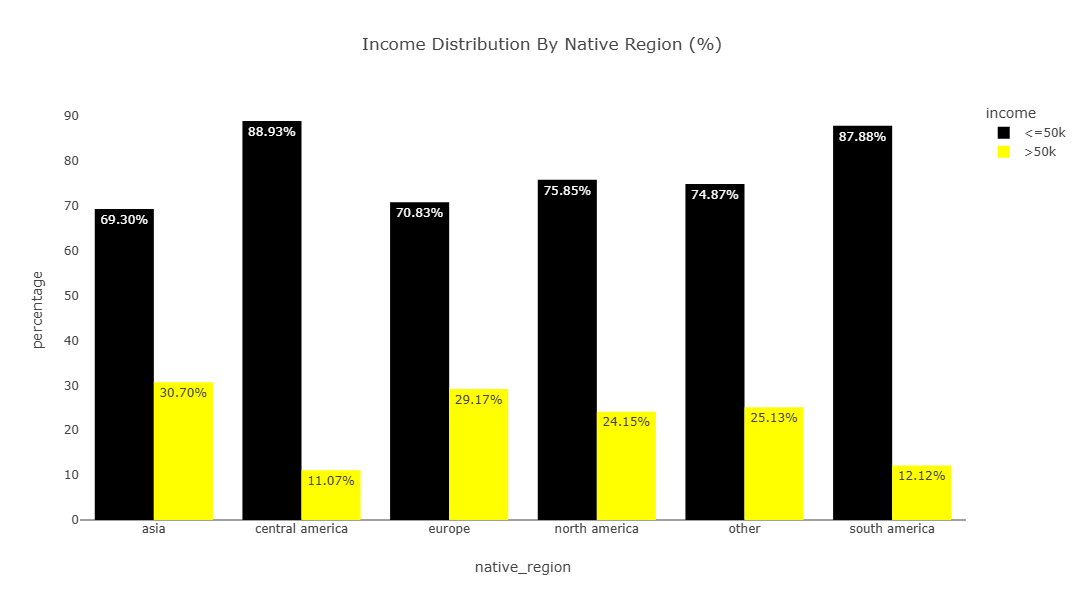


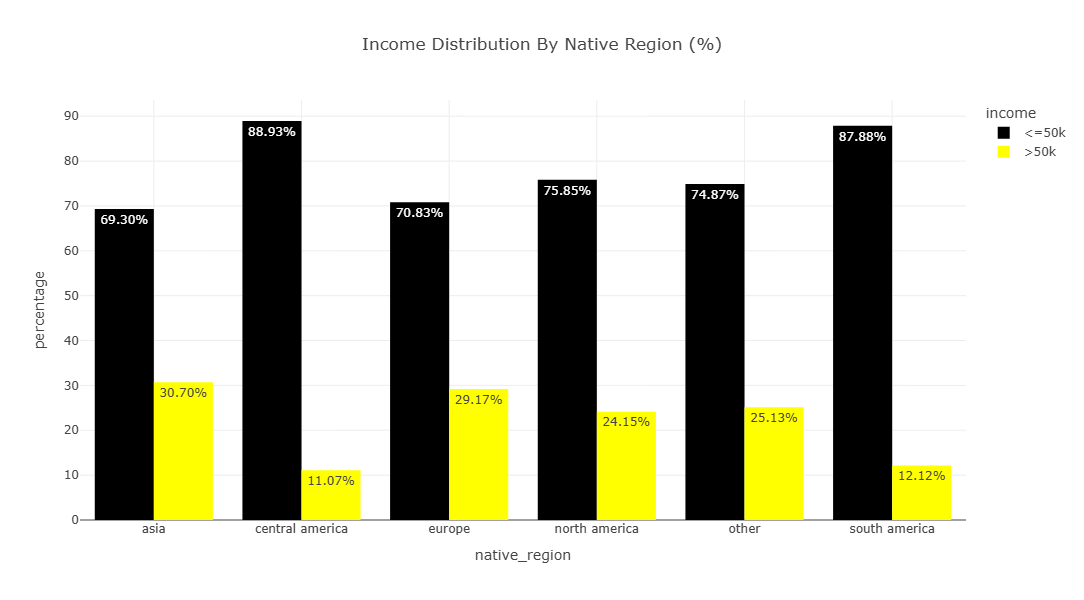














adult\_df\_income\_native\_region = adult\_df.groupby(['native\_region', 'income']).size().reset\_index(name='total\_income\_distr')  
adult\_df\_income\_native\_region

|  | 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.

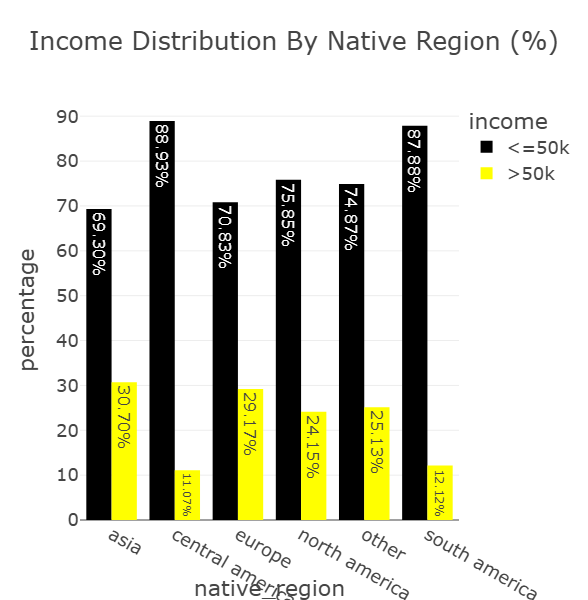
adult\_df\_income\_native\_region = adult\_df.groupby(['native\_region', 'income']).size().reset\_index(name='total\_income\_distr')  
adult\_df\_income\_native\_region

|  | 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\_distr'].transform('sum')  
adult\_df\_income\_native\_region['percentage'] = (adult\_df\_income\_native\_region['total\_income\_distr']/total\_per\_region) \* 100  
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 = "rgba(0,0,0,0)")  
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\_income\_distr')  
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\_race) \* 100  
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']).size().reset\_index(name='total').sort\_values('total', ascending= False)  
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 |

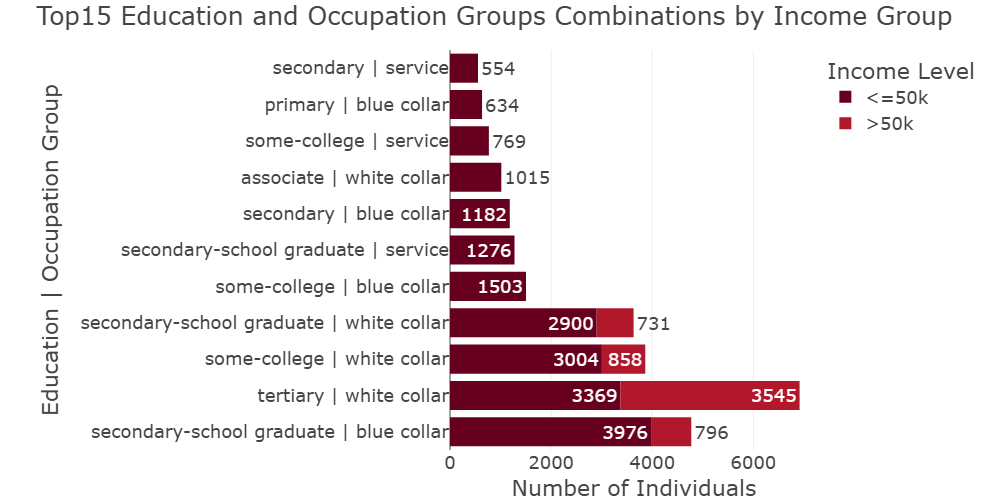
adult\_df\_income\_edu\_occ.head(15)

|  | 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 |
| 51 | secondary-school graduate | white collar | >50k | 731 | secondary-school graduate | white collar |
| 23 | primary | blue collar | <=50k | 634 | primary | blue collar |
| 36 | secondary | service | <=50k | 554 | secondary | service |

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.