UCI Adult Income Dataset - Exploratory and Descriptive Analysis

Chance UWUMUKIZA

2025-06-25

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 pandas as pd
import numpy as np
import os
import plotly.express as px
```

```python
import plotly.io as pio
pio.renderers.default = 'notebook'
```

1. 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.

```python
# Get working directory
current_dir = os.getcwd()

# Go one directory 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)
```

2. 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 = pd.read_csv(adult_data_filename)
adult_df.head(10)
```

```
C:\Users\USER\anaconda3\lib\site-packages\IPython\core\formatters.py:342: FutureWarning:

In future versions `DataFrame.to_latex` is expected to utilise the base implementation of `S
```

	age	workclass	fnlwgt	education_num	marital_status	relationship	race	sex
0	39	government	77516	13	single	not-in-family	white	male
1	50	self-employed	83311	13	married	husband	white	male
2	38	private	215646	9	divorced or separated	not-in-family	white	male
3	53	private	234721	7	married	husband	black	male
4	28	private	338409	13	married	wife	black	female
5	37	private	284582	14	married	wife	white	female
6	49	private	160187	5	divorced or separated	not-in-family	black	female
7	52	self-employed	209642	9	married	husband	white	male
8	31	private	45781	14	single	not-in-family	white	female
9	42	private	159449	13	married	husband	white	male

3. Dataset Dimensions and Data Types

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
```

```
(32515, 16)
```

Table 1: Overview of dataset columns, their data types, and the count of missing values in each column.

```
adult_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32515 entries, 0 to 32514
Data columns (total 16 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   age                32515 non-null  int64
 1   workclass          32515 non-null  object
 2   fnlwgt             32515 non-null  int64
 3   education_num      32515 non-null  int64
 4   marital_status     32515 non-null  object
 5   relationship       32515 non-null  object
 6   race               32515 non-null  object
 7   sex                32515 non-null  object
 8   capital_gain       32515 non-null  int64
 9   capital_loss       32515 non-null  int64
 10  hours_per_week     32515 non-null  int64
 11  income             32515 non-null  object
 12  education_level    32515 non-null  object
 13  occupation_grouped 32515 non-null  object
 14  native_region      32515 non-null  object
 15  age_group          32515 non-null  object
dtypes: int64(6), object(10)
memory usage: 4.0+ MB
```

4.Summary Statistics: Numerical Variables

Table 2: Summary statistics for numerical variables in the dataset, including count, mean, standard deviation, min, and quartile values

```python
adult_df.describe()
```

```
C:\Users\USER\anaconda3\lib\site-packages\IPython\core\formatters.py:342: FutureWarning:

In future versions `DataFrame.to_latex` is expected to utilise the base implementation of `St
```

	age	fnlwgt	education_num	capital_gain	capital_loss	hours_per_week
count	32515.000000	3.251500e+04	32515.000000	32515.000000	32515.000000	32515.000000
mean	38.590374	1.897912e+05	10.081593	1079.173428	87.427341	40.441089
std	13.638535	1.055766e+05	2.571943	7390.403187	403.231777	12.349830
min	17.000000	1.228500e+04	1.000000	0.000000	0.000000	1.000000
25%	28.000000	1.178300e+05	9.000000	0.000000	0.000000	40.000000
50%	37.000000	1.783560e+05	10.000000	0.000000	0.000000	40.000000
75%	48.000000	2.370475e+05	12.000000	0.000000	0.000000	45.000000
max	90.000000	1.484705e+06	16.000000	99999.000000	4356.000000	99.000000

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 socioeconomic necessity.

5. Summary Statistics: Categorical Variables

`workclass`

The private sector dominates, employing ~69.7% of the population. The government sector (13.4%) and self-employment (11.2%) also make up substantial portions of the workforce. A small fraction is labeled as "unknown" (5.6%), which may correspond to missing or ambiguous data entries. Tiny proportions are voluntary (0.04%) or unemployed (0.02%), possibly underreported or underrepresented in the sample.

`marital_status`

Married individuals make up the largest group (46.1%), followed by those who are single (32.8%) and divorced or separated (18.1%). Widowed individuals represent a small minority (~3.1%).

`relationship`

The majority are labeled as "male spouse" (40.5%) or "single" (36.1%). Smaller categories include children (15.6%), female spouses (4.8%), and extended relatives (3.0%). The dominance of `male spouse` reflects the dataset's gendered structure and may point to traditional family roles. The relative scarcity of "female spouse" roles suggests potential gender imbalances in how income-earning is reported within households.

`race`

The dataset is overwhelmingly composed of White individuals (~85.4%). Other racial groups include Black (9.6%), Asian or Pacific Islander (3.2%), American Indian or Eskimo (1.0%), and Other (0.8%). The racial imbalance limits the generalizability of models trained on this data. Smaller racial groups may suffer from limited statistical power, affecting fairness and performance in predictive modeling.

`sex`

Males constitute 66.9% of the dataset, with females making up the remaining 33.1%. This male-skewed distribution could be due to sampling (e.g., primary earners in households), workforce participation patterns, or reporting biases.

`education_level`

Secondary-school graduates form the largest educational group (~32%), highlighting the central role of high school completion in the labor force. Tertiary education holders — those with university or equivalent degrees — account for nearly 25% of the population, representing a substantial segment with advanced qualifications. A notable 22.4% have attended some college without necessarily earning a degree, suggesting that partial post-secondary education is common, yet may not always translate into formal certification. The remaining 20% are distributed among those with only secondary education (9.4%), associate degrees (7.5%), primary school (3.5%), and a very small group with only preschool education (0.15%). It is ecident that the education distribution is skewed toward mid- to high-level education, with relatively few individuals having only basic schooling. This reflects a dataset that largely captures working-age adults in formal labor, which may underrepresent the least-educated populations.

`occupation_grouped`

White-collar occupations are the most prevalent (~51%), followed by blue-collar, service, and unknown. Smaller categories include military, which is marginal. Essentially, slightly over half of individuals in the dataset work in professional, managerial, sales, clerical, or tech-support roles. This suggests the dataset is heavily weighted toward professional and administrative occupations. Nearly a third of the population works in manual labor or skilled trade positions (craft, transport, machine operation, farming, etc.). This indicates a significant segment engaged in physically intensive or technical labor.

`native_region`

The vast majority of individuals are from North America (~92.3%). Smaller proportions are from Central America, Asia, Europe, South America, and a generic Other category. The heavy concentration of North American individuals reflects the U.S. focus of the dataset.

`age_group`

The largest groups are 26–35 and 36–45, followed by 46–60. These three age groups represent about 73% of the dataset. Very few individuals are under 18 or above 75, consistent with the dataset's focus on the working-age population.

```
adult_df.describe(include='object')
```

```
C:\Users\USER\anaconda3\lib\site-packages\IPython\core\formatters.py:342: FutureWarning:

In future versions `DataFrame.to_latex` is expected to utilise the base implementation of `S
```

	workclass	marital_status	relationship	race	sex	income	education_level	oc
count	32515	32515	32515	32515	32515	32515	32515	32
unique	7	4	6	5	2	2	7	5
top	private	married	husband	white	male	<=50k	secondary-school graduate	wl
freq	22652	14984	13178	27773	21760	24679	10485	16

```
adult_df['workclass'].value_counts(normalize=True).rename_axis('unique values')
```

```
C:\Users\USER\anaconda3\lib\site-packages\IPython\core\formatters.py:342: FutureWarning:

In future versions `DataFrame.to_latex` is expected to utilise the base implementation of `S
```

unique values	workclass
private	0.696663
self-employed	0.112440
government	0.069414
local-gov	0.064370
unknown	0.056466
voluntary	0.000431
unemployed	0.000215

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

```
C:\Users\USER\anaconda3\lib\site-packages\IPython\core\formatters.py:342: FutureWarning:

In future versions `DataFrame.to_latex` is expected to utilise the base implementation of `St
```

	marital_status
married	0.460833
single	0.327664
divorced or separated	0.180963
widowed	0.030540

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

```
C:\Users\USER\anaconda3\lib\site-packages\IPython\core\formatters.py:342: FutureWarning:

In future versions `DataFrame.to_latex` is expected to utilise the base implementation of `St
```

	relationship
husband	0.405290
not-in-family	0.254775
own-child	0.155590
unmarried	0.105951
wife	0.048224
other-relative	0.030171

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

```
C:\Users\USER\anaconda3\lib\site-packages\IPython\core\formatters.py:342: FutureWarning:

In future versions `DataFrame.to_latex` is expected to utilise the base implementation of `St
```

	race
white	0.854160
black	0.096017
asian or pacific islander	0.031924
american indian or eskimo	0.009565
other	0.008335

6. 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.

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

```
C:\Users\USER\anaconda3\lib\site-packages\IPython\core\formatters.py:342: FutureWarning:

In future versions `DataFrame.to_latex` is expected to utilise the base implementation of `St
```

	income	total
0	<=50k	24679
1	>50k	7836

```
fig = px.pie(adult_df_income,
             names='income',
             values='total',
             title='Overall Income Distribution',
             color_discrete_sequence=px.colors.sequential.RdBu)

fig.update_layout(template="presentation",
                  paper_bgcolor="rgba(0,0,0,0)",
                  plot_bgcolor="rgba(0,0,0,0)")

fig.show()
fig.write_image(os.path.join(results_dir, 'income_distribution_pie_chart.jpg'))
```

```
fig.write_image(os.path.join(results_dir, 'income_distribution_pie_chart.png'))
fig.write_html(os.path.join(results_dir, 'income_distribution_pie_chart.html'))
```

```
Unable to display output for mime type(s): text/html
```

```
Unable to display output for mime type(s): text/html
```

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.

7. Income by Age Group

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

```
C:\Users\USER\anaconda3\lib\site-packages\IPython\core\formatters.py:342: FutureWarning:

In future versions `DataFrame.to_latex` is expected to utilise the base implementation of `S
```

	age_group	income	total_by_age
0	18-25	<=50k	5333
1	18-25	>50k	114
2	26-35	<=50k	6910
3	26-35	>50k	1591
4	36-45	<=50k	5232
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_age_group = adult_df_income_age.groupby('age_group')['total_by_age'].transform('su
total_per_age_group
```

```
C:\Users\USER\anaconda3\lib\site-packages\IPython\core\formatters.py:342: FutureWarning:

In future versions `DataFrame.to_latex` is expected to utilise the base implementation of `St
```

	total_by_age
0	5447
1	5447
2	8501
3	8501
4	8003
5	8003
6	7288
7	7288
8	2091
9	2091
10	240
11	240
12	945

```
adult_df_income_age['percentage'] = (adult_df_income_age['total_by_age'] / total_per_age_grou
adult_df_income_age
```

```
C:\Users\USER\anaconda3\lib\site-packages\IPython\core\formatters.py:342: FutureWarning:

In future versions `DataFrame.to_latex` is expected to utilise the base implementation of `St
```

	age_group	income	total_by_age	percentage
0	18-25	<=50k	5333	97.907105
1	18-25	>50k	114	2.092895
2	26-35	<=50k	6910	81.284555
3	26-35	>50k	1591	18.715445
4	36-45	<=50k	5232	65.375484
5	36-45	>50k	2771	34.624516
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

```
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) * 1
adult_df_income_age
```

```
C:\Users\USER\anaconda3\lib\site-packages\IPython\core\formatters.py:342: FutureWarning:

In future versions `DataFrame.to_latex` is expected to utilise the base implementation of `St
```

	age_group	income	total_by_age	percentage
0	18-25	<=50k	5333	97.907105
1	18-25	>50k	114	2.092895
2	26-35	<=50k	6910	81.284555
3	26-35	>50k	1591	18.715445
4	36-45	<=50k	5232	65.375484
5	36-45	>50k	2771	34.624516
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',
    height = 500,
    color_discrete_sequence=px.colors.sequential.RdBu,
    text = 'percentage'
)

fig.update_traces(texttemplate='%{text:.2f}%', textposition='outside')
fig.update_layout(template="presentation", xaxis_title='Age Group',
                  yaxis_title='Percentage of population', legend_title=dict(text='Income Leve
                  paper_bgcolor = "rgba(0, 0, 0, 0)", plot_bgcolor = "rgba(0, 0, 0, 0)")
```

```
fig.show()
fig.write_image(os.path.join(results_dir, 'income_distribution_by_agegroup_bar_plot.jpg'))
fig.write_image(os.path.join(results_dir, 'income_distribution_by_agegroup_bar_plot.png'))
fig.write_html(os.path.join(results_dir, 'income_distribution_by_agegroup_bar_plot.html'))
```

```
Unable to display output for mime type(s): text/html
```

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.

```
themes = ["plotly", "plotly_white", "plotly_dark", "ggplot2", "seaborn", "simple_white", "pr
for theme in themes:
  fig.update_layout(template=theme)

  fig.show()
```

```
Unable to display output for mime type(s): text/html
```

```
Unable to display output for mime type(s): text/html
```

```
Unable to display output for mime type(s): text/html
```

```
Unable to display output for mime type(s): text/html
```

```
Unable to display output for mime type(s): text/html
```

```
Unable to display output for mime type(s): text/html
```

```
Unable to display output for mime type(s): text/html
```

```
Unable to display output for mime type(s): text/html
```

```
Unable to display output for mime type(s): text/html
```

```
Unable to display output for mime type(s): text/html
```

8. Income by Native Region

```
adult_df_income_reg = adult_df.groupby(['native_region', 'income']).size().reset_index(name=
adult_df_income_reg
```

```
C:\Users\USER\anaconda3\lib\site-packages\IPython\core\formatters.py:342: FutureWarning:

In future versions `DataFrame.to_latex` is expected to utilise the base implementation of `S
```

	native_region	income	total_by_region
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	436
9	other	>50k	146
10	south america	<=50k	174
11	south america	>50k	24

```
total_per_native_region = adult_df_income_reg.groupby('native_region')['total_by_region'].tra
adult_df_income_reg['percentage'] = (adult_df_income_reg['total_by_region']/total_per_native_
adult_df_income_reg
```

```
C:\Users\USER\anaconda3\lib\site-packages\IPython\core\formatters.py:342: FutureWarning:

In future versions `DataFrame.to_latex` is expected to utilise the base implementation of `St
```

	native_region	income	total_by_region	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	436	74.914089
9	other	>50k	146	25.085911
10	south america	<=50k	174	87.878788
11	south america	>50k	24	12.121212

```
fig = px.bar(
    adult_df_income_reg,
    x = 'native_region',
    y = 'percentage',
    color = 'income',
    title = 'Income Distribution by native region (%)',
    barmode = 'group',
    height= 540,
    color_discrete_sequence=px.colors.sequential.RdBu,
    text = 'percentage'
)
fig.update_traces(texttemplate='%{text:.2f}%', textposition = 'outside')
fig.update_layout(template = "presentation",
                  xaxis_title = 'native_region',
                  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.show()
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'))
```

```
Unable to display output for mime type(s): text/html
```

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.

9. Income by Race

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

```
C:\Users\USER\anaconda3\lib\site-packages\IPython\core\formatters.py:342: FutureWarning:

In future versions `DataFrame.to_latex` is expected to utilise the base implementation of `St
```

	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	20661
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_
adult_df_income_race
```

```
C:\Users\USER\anaconda3\lib\site-packages\IPython\core\formatters.py:342: FutureWarning:

In future versions `DataFrame.to_latex` is expected to utilise the base implementation of `St
```

	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	20661	74.392395
9	white	>50k	7112	25.607605

```
fig = px.bar(
    adult_df_income_race,
    x = 'race',
    y = 'percentage',
    color = 'income',
    title = 'Income Distribution by race (%)',
    barmode = 'group',
    height= 540,
    width= 1200,
    color_discrete_sequence=px.colors.sequential.RdBu,
    text = 'percentage'
)
fig.update_traces(texttemplate='%{text:.2f}%', textposition = 'outside')
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.show()
```

```
Unable to display output for mime type(s): text/html
```

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.

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.

10. Income Distribution by Education Level and Ocuupation grouped

```
adult_df_income_edu_occ = (adult_df.groupby(['education_level', 'occupation_grouped', 'income
                           .size().reset_index(name='total').sort_values('total', ascending =
adult_df_income_edu_occ
```

```
C:\Users\USER\anaconda3\lib\site-packages\IPython\core\formatters.py:342: FutureWarning:

In future versions `DataFrame.to_latex` is expected to utilise the base implementation of `St
```

	education_level	occupation_grouped	income	total
33	secondary-school graduate	blue collar	<=50k	3977
62	tertiary	white collar	>50k	3545
61	tertiary	white collar	<=50k	3369
51	some college	white collar	<=50k	3004
41	secondary-school graduate	white collar	<=50k	2900
43	some college	blue collar	<=50k	1503
23	secondary	blue collar	<=50k	1349
37	secondary-school graduate	service	<=50k	1276
8	associate	white collar	<=50k	1015
52	some college	white collar	>50k	858
34	secondary-school graduate	blue collar	>50k	796
47	some college	service	<=50k	769
42	secondary-school graduate	white collar	>50k	731
27	secondary	service	<=50k	642
14	primary	blue collar	<=50k	634
31	secondary	white collar	<=50k	552
39	secondary-school graduate	unknown	<=50k	487
0	associate	blue collar	<=50k	482
49	some college	unknown	<=50k	481
9	associate	white collar	>50k	397
44	some college	blue collar	>50k	397
53	tertiary	blue collar	<=50k	375
29	secondary	unknown	<=50k	307
17	primary	service	<=50k	232
4	associate	service	<=50k	184
54	tertiary	blue collar	>50k	183
57	tertiary	service	<=50k	180
59	tertiary	unknown	<=50k	172
35	secondary-school graduate	military	<=50k	172
1	associate	blue collar	>50k	166
45	some college	military	<=50k	135
24	secondary	blue collar	>50k	116
19	primary	unknown	<=50k	111
21	primary	white collar	<=50k	93
6	associate	unknown	<=50k	89
60	tertiary	unknown	>50k	82
46	some college	military	>50k	69
56	tertiary	military	>50k	65
55	tertiary	military	<=50k	53
38	secondary-school graduate	service	>50k	53
2	associate	military	<=50k	53
32	secondary	white collar	>50k	49
36	secondary-school graduate	military	>50k	47
40	secondary-school graduate	unknown	>50k	46
15	primary	blue collar	>50k	40
50	some college	unknown	>50k	35
58	tertiary	service	>50k	33
3	associate	military	>50k	29
48	some college	service	>50k	26
10	preschool	blue collar	<=50k	25
25	secondary	military	<=50k	22
7	associate	unknown	>50k	19

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

```
C:\Users\USER\anaconda3\lib\site-packages\IPython\core\formatters.py:342: FutureWarning:

In future versions `DataFrame.to_latex` is expected to utilise the base implementation of `St
```

	education_level	occupation_grouped	income	total	edu_occ
33	secondary-school graduate	blue collar	<=50k	3977	secondary-school graduate | blue col
62	tertiary	white collar	>50k	3545	tertiary | white collar
61	tertiary	white collar	<=50k	3369	tertiary | white collar
51	some college	white collar	<=50k	3004	some college | white collar
41	secondary-school graduate	white collar	<=50k	2900	secondary-school graduate | white co
43	some college	blue collar	<=50k	1503	some college | blue collar
23	secondary	blue collar	<=50k	1349	secondary | blue collar
37	secondary-school graduate	service	<=50k	1276	secondary-school graduate | service
8	associate	white collar	<=50k	1015	associate | white collar
52	some college	white collar	>50k	858	some college | white collar
34	secondary-school graduate	blue collar	>50k	796	secondary-school graduate | blue col
47	some college	service	<=50k	769	some college | service
42	secondary-school graduate	white collar	>50k	731	secondary-school graduate | white co
27	secondary	service	<=50k	642	secondary | service
14	primary	blue collar	<=50k	634	primary | blue collar
31	secondary	white collar	<=50k	552	secondary | white collar
39	secondary-school graduate	unknown	<=50k	487	secondary-school graduate | unknow
0	associate	blue collar	<=50k	482	associate | blue collar
49	some college	unknown	<=50k	481	some college | unknown
9	associate	white collar	>50k	397	associate | white collar
44	some college	blue collar	>50k	397	some college | blue collar
53	tertiary	blue collar	<=50k	375	tertiary | blue collar
29	secondary	unknown	<=50k	307	secondary | unknown
17	primary	service	<=50k	232	primary | service
4	associate	service	<=50k	184	associate | service
54	tertiary	blue collar	>50k	183	tertiary | blue collar
57	tertiary	service	<=50k	180	tertiary | service
59	tertiary	unknown	<=50k	172	tertiary | unknown
35	secondary-school graduate	military	<=50k	172	secondary-school graduate | military
1	associate	blue collar	>50k	166	associate | blue collar
45	some college	military	<=50k	135	some college | military
24	secondary	blue collar	>50k	116	secondary | blue collar
19	primary	unknown	<=50k	111	primary | unknown
21	primary	white collar	<=50k	93	primary | white collar
6	associate	unknown	<=50k	89	associate | unknown
60	tertiary	unknown	>50k	82	tertiary | unknown
46	some college	military	>50k	69	some college | military
56	tertiary	military	>50k	65	tertiary | military
55	tertiary	military	<=50k	53	tertiary | military
38	secondary-school graduate	service	>50k	53	secondary-school graduate | service
2	associate	military	<=50k	53	associate | military
32	secondary	white collar	>50k	49	secondary | white collar
36	secondary-school graduate	military	>50k	47	secondary-school graduate | military
40	secondary-school graduate	unknown	>50k	46	secondary-school graduate | unknow
15	primary	blue collar	>50k	40	primary | blue collar
50	some college	unknown	>50k	35	some college | unknown
58	tertiary	service	>50k	33	tertiary | service
3	associate	military	>50k	29	associate | military
48	some college	service	>50k	26	some college | service
10	preschool	blue collar	<=50k	25	preschool | blue collar
25	secondary	military	<=50k	22	secondary | military
7	associate	unknown	>50k	19	associate | unknown

```
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.update_traces(textposition = 'inside')
fig.show()
fig.write_image(os.path.join(results_dir,'income_distribution_by_eduandocc_bar_plot.jpg'))
fig.write_image(os.path.join(results_dir,'income_distribution_by_eduandocc_bar_plot.png'))
fig.write_html(os.path.join(results_dir,'income_distribution_by_eduandocc_bar_plot.html'))
```

```
Unable to display output for mime type(s): text/html
```

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.

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.