# Data Exploration Questions

* How many rows and columns are there in the dataset?

Rows: 48842 and columns: 15

**Df.info()**

* What are the data types of each column (numeric, categorical, object)?

age int64

workclass object

fnlwgt int64

education object

educational-num int64

marital-status object

occupation object

relationship object

race object

gender object

capital-gain int64

capital-loss int64

hours-per-week int64

native-country object

income object

**Df.info()**

**Df.dtypes**

* How many unique values exist in each categorical column?

age 74

workclass 9

fnlwgt 28523

education 16

educational-num 16

marital-status 7

occupation 15

relationship 6

race 5

gender 2

capital-gain 123

capital-loss 99

hours-per-week 96

native-country 42

income 2

**df.nunique()**

* What is the distribution of the target variable income (<=50K vs >50K)?

income

<=50K 76.071823

>50K 23.928177

df['income'].value\_counts(normalize = True) \* 100

* What are the summary statistics (mean, median, std, min, max, quartiles) of numerical columns like age, fnlwgt, educational-num, hours-per-week, capital-gain, capital-loss?

|  | **age** | **fnlwgt** | **educational-num** | **capital-gain** | **capital-loss** | **hours-per-week** |
| --- | --- | --- | --- | --- | --- | --- |
| **count** | 48842.000000 | 4.884200e+04 | 48842.000000 | 48842.000000 | 48842.000000 | 48842.000000 |
| **mean** | 38.643585 | 1.896641e+05 | 10.078089 | 1079.067626 | 87.502314 | 40.422382 |
| **std** | 13.710510 | 1.056040e+05 | 2.570973 | 7452.019058 | 403.004552 | 12.391444 |
| **min** | 17.000000 | 1.228500e+04 | 1.000000 | 0.000000 | 0.000000 | 1.000000 |
| **25%** | 28.000000 | 1.175505e+05 | 9.000000 | 0.000000 | 0.000000 | 40.000000 |
| **50%** | 37.000000 | 1.781445e+05 | 10.000000 | 0.000000 | 0.000000 | 40.000000 |
| **75%** | 48.000000 | 2.376420e+05 | 12.000000 | 0.000000 | 0.000000 | 45.000000 |
| **max** | 90.000000 | 1.490400e+06 | 16.000000 | 99999.000000 | 4356.000000 | 99.000000 |

**Df.describe()**

* What is the age distribution of individuals?

age

17 595

18 862

19 1053

20 1113

21 1096

...

86 1

87 3

88 6

89 2

90 55

**df['age'].value\_counts().sort\_index()**

* What is the distribution of hours-per-week? Are there extreme values (like people working 0 or 100 hours)?

count 48842.000000

mean 40.422382

std 12.391444

min 1.000000

25% 40.000000

50% 40.000000

75% 45.000000

max 99.000000

**df[‘hours-per-week’].describe()**

1

**df['hours-per-week'].min()**

99

**df['hours-per-week'].max()**

* Which individuals have the highest capital-gain and capital-loss?

|  | **age** | **workclass** | **education** | **capital-gain** |
| --- | --- | --- | --- | --- |
| **16030** | 36 | Private | Bachelors | 99999 |
| **14450** | 40 | Self-emp-inc | Doctorate | 99999 |
| **14624** | 48 | Self-emp-not-inc | Doctorate | 99999 |
| **23853** | 63 | Private | HS-grad | 99999 |
| **31108** | 33 | Private | Bachelors | 99999 |

|  | **age** | **workclass** | **education** | **capital-loss** |
| --- | --- | --- | --- | --- |
| **25244** | 90 | ? | HS-grad | 4356 |
| **31037** | 82 | Private | HS-grad | 4356 |
| **47062** | 66 | ? | Some-college | 4356 |
| **40083** | 41 | Private | Some-college | 3900 |
| **36697** | 54 | Private | 7th-8th | 3900 |

**df[['age', 'workclass', 'education', 'capital-gain']].sort\_values('capital-gain', ascending=False).head(5)**

**df[['age', 'workclass', 'education', 'capital-loss']].sort\_values('capital-loss', ascending=False).head(5)**

* Is fnlwgt column meaningful for prediction, or just a sampling weight?

No fnlwgt is not a meaningful column for prediction it’s just a sampling weight I will drop it in future.

* What are the unique values and their counts in workclass?

workclass

Private 33906

Self-emp-not-inc 3862

Local-gov 3136

? 2799

State-gov 1981

Self-emp-inc 1695

Federal-gov 1432

Without-pay 21

Never-worked 10

**df['workclass'].value\_counts()**

* How is education distributed? Which education levels are most common?

**df['education'].value\_counts()**

* Does educational-num map properly to education categories (e.g., HS-grad → 9)?

| **education** | **educational-num** |
| --- | --- |
| Preschool | 1 |
| 1st-4th | 2 |
| 5th-6th | 3 |
| 7th-8th | 4 |
| 9th | 5 |
| 10th | 6 |
| 11th | 7 |
| 12th | 8 |
| HS-grad | 9 |
| Some-college | 10 |
| Assoc-voc | 11 |
| Assoc-acdm | 12 |
| Bachelors | 13 |
| Masters | 14 |
| Prof-school | 15 |
| Doctorate | 16 |

**df[['education', 'educational-num']].drop\_duplicates().sort\_values('educational-num')**

* What are the common occupations (occupation) and how are they distributed?

**df['occupation'].value\_counts()**

* How are marital statuses distributed in marital-status?

**df['marital-status'].value\_counts()**

* What are the proportions of genders in the dataset?

**df['gender'].value\_counts(normalize=True) \* 100**

* What is the race distribution in the dataset?

**df['race'].value\_counts()**

* Which countries are represented in native-country, and which are most frequent?

**df['native-country'].value\_counts()**

* How does income (<=50K vs >50K) vary with **education level**?

| **income** | **<=50K** | **>50K** |
| --- | --- | --- |
| **education** |  |  |
| **10th** | 1302 | 87 |
| **11th** | 1720 | 92 |
| **12th** | 609 | 48 |
| **1st-4th** | 239 | 8 |
| **5th-6th** | 482 | 27 |
| **7th-8th** | 893 | 62 |
| **9th** | 715 | 41 |
| **Assoc-acdm** | 1188 | 413 |
| **Assoc-voc** | 1539 | 522 |
| **Bachelors** | 4712 | 3313 |
| **Doctorate** | 163 | 431 |
| **HS-grad** | 13281 | 2503 |
| **Masters** | 1198 | 1459 |
| **Preschool** | 82 | 1 |
| **Prof-school** | 217 | 617 |
| **Some-college** | 8815 | 2063 |

**pd.crosstab(df['education'], df['income'])**

* What is the relationship between **age** and income category?

**df.groupby('income')['age'].sum()**

**df.groupby('income')['age'].describe()**

**pd.crosstab(pd.cut(df['age'], bins=[15,25,35,45,55,65,90]), df['income'])**

* Do certain occupations correlate with higher income?

| **income** | **<=50K** | **>50K** |
| --- | --- | --- |
| **occupation** |  |  |
| **?** | 90.566038 | 9.433962 |
| **Adm-clerical** | 86.312600 | 13.687400 |
| **Armed-Forces** | 66.666667 | 33.333333 |
| **Craft-repair** | 77.372382 | 22.627618 |
| **Exec-managerial** | 52.218206 | 47.781794 |
| **Farming-fishing** | 88.389262 | 11.610738 |
| **Handlers-cleaners** | 93.339768 | 6.660232 |
| **Machine-op-inspct** | 87.690271 | 12.309729 |
| **Other-service** | 95.856185 | 4.143815 |
| **Priv-house-serv** | 98.760331 | 1.239669 |
| **Prof-specialty** | 54.893065 | 45.106935 |
| **Protective-serv** | 68.667345 | 31.332655 |
| **Sales** | 73.201308 | 26.798692 |
| **Tech-support** | 70.954357 | 29.045643 |
| **Transport-moving** | 79.575372 | 20.424628 |

**pd.crosstab(df['occupation'], df['income'], normalize='index') \* 100**

* Does gender affect income level distribution?

**pd.crosstab(df['gender'],df['income'], normalize = True) \* 100**

* How does race distribution differ across income categories?

**pd.crosstab(df['race'], df['income'], normalize='index') \* 100**

* What is the effect of hours-per-week on income (>50K vs <=50K)?

**df.groupby('income')['hours-per-week'].describe()**

* Does capital-gain strongly influence income prediction?

**df.groupby('income')['capital-gain'].count()**

* How many total observations (rows) and features (columns) are in the full dataset?

**df.shape**

* What are the data types (e.g., integer, continuous, categorical/object) of each column?

**Df.dtypes**

* What is the distribution of the target variable, income? (i.e., What percentage of individuals earn >$50K versus ≤$50K?)

**df['income'].value\_counts(normalize = True) \* 100**

* What are the minimum, maximum, mean, and median values for the numerical columns like age, fnlwgt, capital-gain, capital-loss, and hours-per-week?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **age** | **fnlwgt** | **capital-gain** | **capital-loss** | **hours-per-week** |
| **min** | 17.000000 | 1.228500e+04 | 0.000000 | 0.000000 | 1.000000 |
| **max** | 90.000000 | 1.490400e+06 | 99999.000000 | 4356.000000 | 99.000000 |
| **mean** | 38.643585 | 1.896641e+05 | 1079.067626 | 87.502314 | 40.422382 |
| **median** | 37.000000 | 1.781445e+05 | 0.000000 | 0.000000 | 40.000000 |

**df[['age','fnlwgt','capital-gain','capital-loss','hours-per-week']].agg(['min','max','mean','median'])**

* Are there any potential outliers in the continuous numerical features (e.g., very high capital-gain or very low/high hours-per-week)?

Outliers detected using IQR:

{'age': 216, 'fnlwgt': 1453, 'educational-num': 1794, 'capital-gain': 4035, 'capital-loss': 2282, 'hours-per-week': 13496}

**numeric\_cols = df.select\_dtypes(include='number').columns**

**outliers\_iqr = {}**

**for col in numeric\_cols:**

**Q1 = df[col].quantile(0.25)**

**Q3 = df[col].quantile(0.75)**

**IQR = Q3 - Q1**

**lower\_bound = Q1 - 1.5 \* IQR**

**upper\_bound = Q3 + 1.5 \* IQR**

**outliers = df[(df[col] < lower\_bound) | (df[col] > upper\_bound)]**

**outliers\_iqr[col] = len(outliers)**

**print("Outliers detected using IQR:")**

**print(outliers\_iqr**

* For the categorical features (workclass, education, marital-status, occupation, relationship, race, gender, native-country), what are the top 5 most frequent categories in each column?

**categorical\_cols = ['workclass','education','marital-status','occupation','relationship','race','gender','native-country']**

**for col in categorical\_cols:**

**print(f"\nTop 5 in {col}:")**

**print(df[col].value\_counts().head(5))**

* How many unique categories are there in the high-cardinality feature native-country?

**df.nunique()**

* How does education level (e.g., 'Bachelors' vs 'HS-grad') relate to the individual's income?

**pd.crosstab(df['education'], df['income'], normalize='index') \* 100**

* What is the average age for people in the >$50K income group compared to the ≤$50K group?

**df.groupby('income')['age'].mean()**

* Is there a difference in the distribuation of hours-per-week worked between the >$50K and ≤$50K income groups?

**df.groupby('income')['hours-per-week'].describe()**

* Which occupation categories have the highest proportion of individuals earning >$50K?

**pd.crosstab(df['occupation'], df['income'], normalize='index') \* 100**

* How does the gender distribution vary across different workclass types?

**pd.crosstab(df['workclass'], df['gender'], normalize='index') \* 100**

# Data Cleaning Questions

* Which columns contain missing values (represented by '?') and what is the exact count or percentage of missing values in each of those columns (workclass, occupation, native-country)?

**(df == '?').sum()**

**(df == '?').sum() / len(df) \* 100**

**df = df[(df['workclass'] != '?') &**

**(df['occupation'] != '?') &**

**(df['native-country'] != '?')]**

* Should the missing values in categorical columns like workclass and occupation be imputed with the mode, a generic category ('Unknown'), or dropped, and what justification supports that choice?

**for col in ['workclass',** 'occupation'**, 'native-country']:**

**df[col] = df[col].replace('?', 'Unknown')**

**It should be dropped because imputing the data with mode or Unknown will risk of introducing noise or artificial data**

* Are there any inconsistencies in the naming or formatting of categories (e.g., are 'United-States' and 'United States' considered the same in the full dataset for native-country)?

**df['native-country'].unique()**

**df['native-country'] = df['native-country'].str.replace('-', ' ')**

**country\_map = {**

**'Columbia': 'Colombia',**

**'Trinadad&Tobago': 'Trinidad & Tobago',**

**'Hong': 'Hong Kong',**

**'Holand Netherlands': 'Netherlands',**

**# 'South': 'South Korea' # only if you know what it is**

**}**

**df['native-country'] = df['native-country'].replace(country\_map)**

* Are there any redundant or highly correlated features that should be considered for removal (e.g., do education and education-num essentially represent the same information)?

**df = df.drop(columns=['education']) # or drop 'educational-num'**

* Do the numerical columns contain any infeasible zero values (e.g., a person with age 0) or values that are clearly used as placeholders for missing data (e.g., capital-gain=99999)?

**df['capital-gain'] = df['capital-gain'].replace(99999, np.nan)**

* Are there missing values in the dataset?

**df.isnull().sum()**

* How are missing values represented (e.g., "?", "NaN", blanks)?

**(df == '?').sum()**

**df.isna().sum()**

**(df == '').sum()**

* How many rows have missing values in workclass and occupation?

**Non of them have missing rows**

**workclass\_missing = (df['workclass'] == '?').sum()**

**occupation\_missing = (df['occupation'] == '?').sum()**

**print(workclass\_missing, occupation\_missing)**

* Should we drop rows with "?" values, or replace them with "Unknown"?

**We Should drop them**

* Are there duplicate rows? If so, how many, and should we drop them?

**Number of duplicate rows: 47**

**No we should not drop them because 2 people can have the same everythingg**

**df.duplicated()**

**duplicate\_count = df.duplicated().sum()**

**print("Number of duplicate rows:", duplicate\_count)**

* Are all categorical values consistent (e.g., "United-States" vs "United States")?

**for col in df.select\_dtypes(include='object').columns:**

**print(f"\nUnique values in {col}:")**

**print(df[col].unique())**

**Yes. After cleaning, all categorical values now appear consistent with no spelling inconsistencies.**

* Are there any outliers in numerical columns (age, hours-per-week, capital-gain, capital-loss)?

**numeric\_cols = ['age', 'hours-per-week', 'capital-gain', 'capital-loss']**

**for col in numeric\_cols:**

**Q1 = df[col].quantile(0.25)**

**Q3 = df[col].quantile(0.75)**

**IQR = Q3 - Q1**

**lower = Q1 - 1.5 \* IQR**

**upper = Q3 + 1.5 \* IQR**

**outliers = df[(df[col] < lower) | (df[col] > upper)]**

**print(f"{col}: {len(outliers)} outliers")**

**age: 269 outliers**

**hours-per-week: 11899 outliers**

**capital-gain: 3561 outliers**

**capital-loss: 2140 outliers**

* Are all ages realistic (e.g., >0, <100)?

**df['age'].between(0, 100).all()**

* Should we treat fnlwgt column as relevant, or remove it since it may not help modeling?

**We will remove it   
df = df.drop(columns=['fnlwgt'])**

* Are there spelling inconsistencies in categorical data (e.g., Self-emp-not-inc vs Self-emp-inc)?

**No there are no spelling inconsistencies**

**for col in df.select\_dtypes(include='object').columns:**

**print(f"\n{col}:")**

**print(df[col].unique())**

# Data Transformation

* How can categorical variables (like workclass, education, occupation) be encoded — label encoding or one-hot encoding?

**workclass, occupation, marital-status, relationship, race, gender, native-country, income → One-Hot Encoding (because they’re categories without order).**

**df\_encoded = pd.get\_dummies(df, columns=['workclass', 'occupation'])**

* Should we convert gender to binary (0 = Male, 1 = Female) or keep categorical?

**We can convert It to binary or we can also keep it categorical depending on which model are we going to use**

**df['gender\_binary'] = df['gender'].map({'Male': 0, 'Female': 1})**

* Should we group rare native-country values into an “Other” category?

**Yes, group rare countries into “Other” before encoding. It will simplify your model and reduce the number of columns a lot.**

**country\_counts = df['native-country'].value\_counts(normalize=True)**

**rare\_countries = country\_counts[country\_counts < 0.01].index**

**df['native-country-grouped'] = df['native-country'].replace(rare\_countries, 'Other')**

**df['native-country-grouped'].value\_counts().head(15)**

* Should we transform education into educational-num for modeling and drop one of them?

**df\_model = df.drop(columns=['education'])**

**df\_model = df.drop(columns=['educational-num'])**

* Can we create new features, such as **age groups** (young, middle-aged, senior)?

**bins = [0, 25, 45, 65, 100] # edges of the groups**

**labels = ['Young', 'Middle-aged', 'Senior', 'Elder']**

**df['age\_group'] = pd.cut(df['age'], bins=bins, labels=labels)**

* Can we create a binary variable for **capital-gain/loss presence** instead of raw values (e.g., 0 = no gain, 1 = has gain)?

**# 1 if capital-gain > 0 else 0**

**df['has\_capital\_gain'] = (df['capital-gain'] > 0).astype(int)**

**# 1 if capital-loss > 0 else 0**

**df['has\_capital\_loss'] = (df['capital-loss'] > 0).astype(int**)

* Should we scale numerical columns (age, hours-per-week, capital-gain, capital-loss) using normalization or standardization?

**Scale using Standardization (Z-score) because your dataset contains skewed distributions and outliers. If using models sensitive to skewness, log-transform capital features first.**

* What is the most appropriate encoding technique for each categorical column (e.g., One-Hot Encoding for workclass and marital-status, or Target Encoding for native-country)?

 **One-Hot Encoding:** workclass, marital-status, occupation (if low-dim model), relationship, race.

 **Ordinal Encoding:** education (or map to education-num).

 **Binary Encoding (0/1):** sex, income.

 **Target/Frequency Encoding:** native-country (because of many categories)

* How should the high-cardinality feature native-country be handled? (e.g., group all non-USA countries into a single 'Other' category to reduce dimensionality).

**The most practical approach is to group all non-USA countries into “Other” (since U.S. is overwhelmingly dominant). If you want to squeeze more predictive power, use frequency or target encoding, but only with proper safeguards against overfitting.**

* Should the numerical features (e.g., age, hours-per-week) be scaled (e.g., using StandardScaler or MinMaxScaler)?

**Yes, scale the numerical features. Use StandardScaler, but apply a log-transform first for skewed ones (capital-gain & capital-loss).**

* Can an ordinal encoding be applied to the education column, mapping the educational categories to a proper numerical scale (e.g., 'Preschool'=1, 'Bachelors'=13, 'Doctorate'=16) instead of using the existing education-num?

**Yes, you can apply ordinal encoding to education, but since education-num already exists, it’s smarter to reuse that column. Custom mapping is only necessary if you want more control over the ordering.**

* Should new features be engineered, such as a capital-diff feature (capital-gain - capital-loss) or an is-married binary flag derived from the marital-status column?

**Yes — engineering** capital-diff **and** is-married **is recommended, as they capture useful information in a simpler, more predictive form. You can also explore** bucketing features **(age, education, hours) for even stronger signal.**

**df["capital-diff"] = df["capital-gain"] - df["capital-loss"]**

**df["is-married"] = df["marital-status"].apply(lambda x: 1 if "Married" in x else 0)**

**df[["capital-gain", "capital-loss", "capital-diff", "marital-status", "is-married"]].head(10)**

# Visualization

## Univariate Visualizations

|  |  |  |
| --- | --- | --- |
| **Variable Type** | **Chart Type** | **Purpose** |
| Numerical (age, fnlwgt, hours-per-week) | Histogram | To visualize the distribution shape, central tendency, and spread. |
| Numerical (age, fnlwgt, hours-per-week) | Box Plot | To identify potential outliers and visualize quartiles and range. |
| Categorical (workclass, education, occupation) | Bar Chart / Count Plot | To show the frequency or count of each category. |

## Bivariat and Multivariate Analysis

|  |  |  |
| --- | --- | --- |
| **Variables & Goal** | **Chart Type** | **Purpose** |
| Categorical vs. Target (education vs. income) | Stacked Bar Chart / Percentage Bar Chart | To see the proportion of >$50K within each education level. |
| Numerical vs. Target (age vs. income) | Box Plot or Violin Plot | To compare the distribution of age across the two income groups. |
| Categorical vs. Numerical (workclass vs. hours-per-week) | Box Plot or Violin Plot | To compare the median hours-per-week for different employment types. |
| Numerical vs. Numerical (capital-gain vs. capital-loss) | Scatter Plot | To check for a relationship or pattern between the two capital features. |
| Correlation (All numerical features) | Heatmap (of the correlation matrix) | To visualize the linear correlation coefficients between numerical variables. |
| Relationship/Gender vs. Target | Heatmap (of counts/proportions) | To analyze the joint distribution of relationship and gender with the income target. |

**Univariate Analysis**

* Histogram of age distribution.

**plt.hist(df['age'], bins=10, color='skyblue', edgecolor='black')**

**plt.xlabel('Age')**

**plt.ylabel('Frequency')**

**plt.title('Histogram of Age')**

**plt.show()**

* Histogram of hours-per-week distribution.

**plt.hist(df['hours-per-week'], bins=10, color='skyblue', edgecolor='black')**

**plt.xlabel('Age')**

**plt.ylabel('Frequency')**

**plt.title('Histogram of Hours per Week')**

**plt.show()**

* Bar chart of education levels frequency.

**counts = df['education'].value\_counts()**

**plt.bar(counts.index, counts.values, color='orange')**

**plt.xlabel('Gender')**

**plt.ylabel('Count')**

**plt.title('Bar Chart of Education Frquency')**

**plt.xticks(rotation=50)**

**plt.show()**

* Bar chart of workclass categories.

**counts = df['workclass'].value\_counts()**

**plt.bar(counts.index, counts.values, color='orange')**

**plt.xlabel('Work Class')**

**plt.ylabel('Count')**

**plt.title('Bar Chart of Work Class Categories')**

**plt.xticks(rotation=50)**

**plt.show()**

* Bar chart of occupation categories.

**counts = df['occupation'].value\_counts()**

**plt.bar(counts.index, counts.values, color='orange')**

**plt.xlabel('Occupation')**

**plt.ylabel('Count')**

**plt.title('Bar Chart of Occupation Categories')**

**plt.xticks(rotation=50)**

**plt.show()**

* Pie chart or bar chart of gender distribution.

**counts = df['gender'].value\_counts()**

**plt.pie(counts, labels=counts.index, autopct='%1.1f%%', colors=['skyblue','orange'])**

**plt.title('Pie Chart of Gender')**

**plt.show()**

* Pie chart of race distribution.

**counts = df['race'].value\_counts()**

**plt.pie(counts, labels=counts.index, autopct='%1.1f%%', colors=['skyblue','orange'])**

**plt.title('Pie Chart of race')**

**plt.show()**

* Bar chart of native-country (top 10 most frequent countries).

**top10\_countries = df['native-country'].value\_counts().nlargest(10)**

**plt.bar(top10\_countries.index, top10\_countries.values, color='skyblue')**

**plt.xticks(rotation=45)**

**plt.xlabel('Country')**

**plt.ylabel('Count')**

**plt.title('Top 10 Native Countries')**

**plt.show()**

* Histogram of capital-gain and capital-loss values.

**plt.hist(df['capital-gain'], bins=10, color='skyblue', edgecolor='black')**

**plt.xlabel('Capital Gain')**

**plt.ylabel('Frequency')**

**plt.title('Histogram of Capitla Gain')**

**plt.show()**

**plt.hist(df['capital-loss'], bins=10, color='skyblue', edgecolor='black')**

**plt.xlabel('Capital** Loss')

**plt.ylabel('Frequency')**

**plt.title('Histogram of Capital Loss')**

**plt.show()**

* Countplot of income distribution (<=50K vs >50K).

**import seaborn as sns**

**sns.countplot(x='income', data=df, palette='Set2')**

**plt.title('Countplot of Income')**

**plt.show()**

**Bivariate Analysis with Target (income)**

* Bar plot: income vs education level.
* Box plot: age vs income category.
* Bar plot: occupation vs income category.
* Bar plot: workclass vs income category.
* Grouped bar chart: gender vs income.
* Grouped bar chart: race vs income.
* Box plot: hours-per-week vs income.
* Scatter plot: age vs capital-gain, colored by income.
* Violin plot: educational-num vs income.

**Multivariate Analysis**

* Heatmap of correlations among numerical features (age, hours-per-week, capital-gain, capital-loss, educational-num).
* Pair plot of numerical features colored by income category.
* Stacked bar chart: marital-status vs gender vs income.
* Box plot: hours-per-week grouped by education and colored by income.
* Stacked bar chart: workclass vs race vs income.
* Scatter plot: capital-gain vs capital-loss with income categories.

POWERBI

# Part 1: Data Transformation (Power Query)

### **Basic Cleaning**

**Q1. Are there missing values in** workclass **and** occupation **(shown as “?”)? How to handle them?**  
-> In Power Query → Use **Replace Values** → Replace “?” with “Unknown” or create a filter to remofindve rows with missing values.

**Q2. Do categorical values have extra spaces (like “ Private” vs “Private”)?**  
-> In Power Query → Select the column → **Transform → Format → Trim & Clean**.

**Q3. Are all column names readable and user-friendly?**  
-> Rename columns in Power Query to proper names (e.g., educational-num → Education Level Num).

**Q4. Should** fnlwgt **be kept or removed since it is a sample weight?**  
-> Decide: If not needed, **Remove Columns** in Power Query.

### **Data Type Corrections**

**Q5. Are all columns using correct data types (text, number, date, categorical)?**  
-> In Power Query → Use **Detect Data Type** or manually set:

* age, hours-per-week, capital-gain/loss → Whole Number.
* income, gender, race, workclass, etc. → Text.

**Q6. Should we change** income **column into a binary categorical field (<=50K / >50K)?**  
-> In Power Query → Use **Replace Values** to standardize categories, or create a **Conditional Column**.

### **Feature Engineering**

**Q7. Can we group** education **levels into broader categories (Low, Medium, High)?**  
-> In Power Query → Add **Conditional Column** or use **Group By** logic.

**Q8. Can we create age groups (e.g., Young: 18–30, Middle: 31–50, Senior: 51+)?**  
-> In Power Query → Use **Conditional Column** to categorize ranges.

**Q9. Should we create a new column for “Has Capital Gain” (Yes/No)?**  
-> In Power Query → Add **Custom Column** → If [capital-gain] > 0 then “Yes” else “No”.

**Q10. Can we derive a column for** Work Hours Category **(Part-time <35, Full-time 35–50, Overtime >50)?**  
-> In Power Query → Add **Conditional Column** with those ranges.

**Q11. Should we simplify marital status into just “Married” vs “Not Married”?**  
-> In Power Query → Add **Conditional Column** grouping categories.

**Q12. Can we group rare** native-country **values into “Other”?**  
-> In Power Query → Use **Group By** or **Replace Values** for infrequent categories.

### **Data Transformation Enhancements**

**Q13. Do we need both** education **and** educational-num**?**  
-> Decide whether to **Remove Columns** or keep both.

**Q14. Should we unpivot or pivot any columns for better visualization?**  
-> If needed (e.g., analyzing capital-gain and capital-loss together) → Use **Unpivot Columns**.

**Q15. Should we merge datasets (if multiple files)?**  
-> In Power Query → Use **Append Queries** or **Merge Queries**.

# 📊 Part 2: Visualizations in Power BI

### **Univariate Visuals**

**Q1. What is the age distribution of people in the dataset?**  
-> Use **Histogram or Column Chart** on Age.

**Q2. What is the distribution of** hours-per-week**?**  
-> Use **Histogram or Column Chart**.

**Q3. Which education levels are most common?**  
-> Use **Bar Chart** for Education.

**Q4. What is the distribution of workclass categories?**  
-> Use **Bar Chart**.

**Q5. How many males vs females are in the dataset?**  
-> Use **Donut Chart** or **Bar Chart**.

**Q6. What is the race distribution?**  
-> Use **Pie Chart** or **Stacked Bar Chart**.

**Q7. Which are the top 10 native countries?**  
-> Use **Bar Chart (Top N filter)**.

**Q8. What percentage of people earn** <=50K **vs** >50K**?**  
-> Use **Donut Chart**.

### **Bivariate Visuals (Comparisons with Income)**

**Q9. How does income vary by education level?**  
-> Use **Stacked Bar Chart** (Education on X-axis, Income as legend).

**Q10. What is the average age by income group?**  
-> Use **Clustered Column Chart** (Income vs Avg Age).

**Q11. Which occupations have the highest percentage of >50K earners?**  
-> Use **Bar Chart** (Occupation vs Income).

**Q12. How does gender affect income distribution?**  
-> Use **Stacked Bar Chart** (Gender vs Income).

**Q13. How does race affect income levels?**  
-> Use **Clustered Bar Chart** (Race vs Income).

**Q14. What is the average hours worked per week by income group?**  
-> Use **Clustered Column Chart**.

**Q15. How does workclass relate to income categories?**  
-> Use **Stacked Bar Chart**.

### **Multivariate Visuals**

**Q16. How does marital status affect income by gender?**  
-> Use **Stacked Bar Chart** (Marital Status vs Income, split by Gender).

**Q17. How do capital gains/losses affect income?**  
-> Use **Scatter Plot** (Capital Gain vs Capital Loss, colored by Income).

**Q18. How does education and hours-per-week together affect income?**  
-> Use **Matrix Table or Clustered Column Chart** (Education on X-axis, Income as legend, Avg Hours as value).

**Q19. What is the correlation between numerical columns (Age, Hours, Capital Gain/Loss, Education Num)?**  
-> Create a **Table or Heatmap** using correlation (done via Power Query or custom calculation).

**Q20. Which factors (age group, work hours, education) together drive income category?**  
-> Use **Stacked Column Chart** or **Treemap**.

### **Interactive Visuals / Filtering**

**Q21. Can we build a slicer for filtering by gender, race, and workclass?**  
-> Add **Slicers** for categorical fields.

**Q22. Can we build a slicer for filtering by income categories?**  
-> Add **Slicer** for Income.

**Q23. Can we allow users to filter by** native-country**?**  
-> Add **Dropdown Slicer**.

------------------------------------------------------------------------------------------------------------------------

| Step (Question) | How to Perform in Power Query Editor | Rationale |
| --- | --- | --- |
| 1. Identify Missing Values | Filter each categorical column (workclass, occupation, native-country) to show only rows containing the value ? (the missing value placeholder). | Pinpoint all missing data points for treatment. |
| 2. Clean Categorical Data | Right-click the header of a categorical column (e.g., workclass), select Replace Values, and replace ? with a specific value like Unknown or the Mode (most frequent value, if calculated manually first). | Prepare the data by explicitly labeling missing categories. |
| 3. Clean Numerical Data | For columns like capital-gain or capital-loss, if the full dataset shows a placeholder like 99999 for missing data, use Replace Values to replace it with 0 or the Median for a more accurate representation. | Handle potential numerical outliers or placeholders. |
| 4. Correct Data Types | Select columns like age, fnlwgt, education-num, and hours-per-week. Go to the Transform tab, and change the data type to Whole Number or Decimal Number as appropriate. | Ensure Power BI treats each feature correctly (e.g., as a number for calculations). |
| 5. Simplify High-Cardinality Column | Select the native-country column. Right-click, select Replace Values. Replace the majority value United-States with USA. (Bonus): Group the remaining countries into Other manually or using Conditional Column logic. | Reduce the number of distinct values for better visualization performance. |
| 6. Create a Simple Binary Feature | Select the gender column. Go to the Add Column tab and use Conditional Column. Create a new column, Is\_Male, where the value is 1 if gender equals Male and 0 otherwise. | Create a numerical feature from a binary categorical one for simple analysis. |
| 7. Simplify the Target Column | Select the income column. Use Replace Values to remove the period (.) at the end of the values (e.g., change <=50K. to <=50K). | Standardize and clean the final target output labels. |
| 8. Create Age Groups (Binning) | Select the age column. Go to the Transform tab and select Group By. Choose a custom grouping logic to create bins (e.g., 18-30, 31-45, 46-60, 60+). | Convert a continuous numerical feature into a categorical feature for easier slicing and filtering in visualizations. |
| 9. Extract Value from Text | Select the education column. Use Extract (under the Transform tab) to pull out a specific portion of the text (though simple cleaning is likely better for this column). (Alternative): Use Column From Examples to quickly generate a simplified version of the education level. | Practice manipulating string (text) data. |

📈 Power BI Visualizations (Desktop Practice)

These visualizations cover fundamental chart types and interaction techniques in Power BI, using the cleaned dataset.

1. Analysis of the Target Variable (income)

| Visualization | Fields to Use | Question Answered |
| --- | --- | --- |
| Donut Chart or Pie Chart | Legend: income; Values: Count of income (or any field). | What is the overall distribution of income (class imbalance)? |
| Card | Fields: Count of all rows (or count of ≤$50K rows). | What is the total number of individuals in the dataset? |

2. Income Distribution by Categorical Features

| Visualization | Fields to Use | Question Answered |
| --- | --- | --- |
| Stacked Bar Chart | Y-axis: workclass; X-axis: Count of income; Legend: income. | Which employment types (workclass) have the highest volume of people, and what proportion of them are high-earners (>$50K)? |
| 100% Stacked Bar Chart | Y-axis: education; X-axis: Count of income; Legend: income. | What percentage of people at each education level (e.g., Bachelors) earn >$50K vs. ≤$50K? |
| Treemap | Group: occupation; Values: Count of income. Color Saturation: Average of Is\_Male (for a quick gender comparison). | Which occupations are the largest contributors to the dataset, and how does gender play a role within them? |
| Slicer | Field: marital-status. | How does selecting a marital status (e.g., Married-civ-spouse) filter and affect *all other charts* on the report page? |

3. Analysis of Numerical Features

| Visualization | Fields to Use | Question Answered |
| --- | --- | --- |
| Clustered Column Chart | X-axis: Age Group (the binned column); Y-axis: Average of hours-per-week. | Which age group works the highest average number of hours per week? |
| Scatter Plot | X-axis: age; Y-axis: hours-per-week; Legend: income. | Is there a visible relationship between age and hours worked, and how does income influence that pattern? |
| Line Chart | X-axis: education-num (treated as a continuous axis); Y-axis: Count of income (only >$50K filtered). | How does the count of high-earners change as the number of education years increases? |

4. Geographical and Miscellaneous

| Visualization | Fields to Use | Question Answered |
| --- | --- | --- |
| Table or Matrix | Rows: native-country; Values: Count of income (for both categories). | Provide a detailed breakdown of income counts by country. |
| Funnel Chart | Group: education-num; Values: Count of income (filter for only >$50K). | Visualize the drop-off in the number of high-earners as the years of education decrease. |