# Data Exploration Questions

* How many rows and columns are there in the dataset?
* What are the data types of each column (numeric, categorical, object)?
* How many unique values exist in each categorical column?
* What is the distribution of the target variable income (<=50K vs >50K)?
* 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?
* What is the age distribution of individuals?
* What is the distribution of hours-per-week? Are there extreme values (like people working 0 or 100 hours)?
* Which individuals have the highest capital-gain and capital-loss?
* Is fnlwgt column meaningful for prediction, or just a sampling weight?
* What are the unique values and their counts in workclass?
* How is education distributed? Which education levels are most common?
* Does educational-num map properly to education categories (e.g., HS-grad → 9)?
* What are the common occupations (occupation) and how are they distributed?
* How are marital statuses distributed in marital-status?
* What are the proportions of genders in the dataset?
* What is the race distribution in the dataset?
* Which countries are represented in native-country, and which are most frequent?
* How does income (<=50K vs >50K) vary with **education level**?
* What is the relationship between **age** and income category?
* Do certain occupations correlate with higher income?
* Does gender affect income level distribution?
* How does race distribution differ across income categories?
* What is the effect of hours-per-week on income (>50K vs <=50K)?
* Does capital-gain strongly influence income prediction?
* How many total observations (rows) and features (columns) are in the full dataset?
* What are the data types (e.g., integer, continuous, categorical/object) of each column?
* What is the distribution of the target variable, income? (i.e., What percentage of individuals earn >$50K versus ≤$50K?)
* What are the minimum, maximum, mean, and median values for the numerical columns like age, fnlwgt, capital-gain, capital-loss, and hours-per-week?
* Are there any potential outliers in the continuous numerical features (e.g., very high capital-gain or very low/high hours-per-week)?
* 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?
* How many unique categories are there in the high-cardinality feature native-country?
* How does education level (e.g., 'Bachelors' vs 'HS-grad') relate to the individual's income?
* What is the average age for people in the >$50K income group compared to the ≤$50K group?
* Is there a difference in the distribuation of hours-per-week worked between the >$50K and ≤$50K income groups?
* Which occupation categories have the highest proportion of individuals earning >$50K?
* How does the gender distribution vary across different workclass types?

# 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)?
* 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?
* 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)?
* 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)?
* 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)?
* Are there missing values in the dataset?
* How are missing values represented (e.g., "?", "NaN", blanks)?
* How many rows have missing values in workclass and occupation?
* Should we drop rows with "?" values, or replace them with "Unknown"?
* Are there duplicate rows? If so, how many, and should we drop them?
* Are all categorical values consistent (e.g., "United-States" vs "United States")?
* Are there any outliers in numerical columns (age, hours-per-week, capital-gain, capital-loss)?
* Are all ages realistic (e.g., >0, <100)?
* Should we treat fnlwgt column as relevant, or remove it since it may not help modeling?
* Are there spelling inconsistencies in categorical data (e.g., Self-emp-not-inc vs Self-emp-inc)?

# Data Transformation

* How can categorical variables (like workclass, education, occupation) be encoded — label encoding or one-hot encoding?
* Should we convert gender to binary (0 = Male, 1 = Female) or keep categorical?
* Should we group rare native-country values into an “Other” category?
* Should we transform education into educational-num for modeling and drop one of them?
* Can we create new features, such as **age groups** (young, middle-aged, senior)?
* Can we create a binary variable for **capital-gain/loss presence** instead of raw values (e.g., 0 = no gain, 1 = has gain)?
* Should we scale numerical columns (age, hours-per-week, capital-gain, capital-loss) using normalization or standardization?
* 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)?
* 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).
* Should the numerical features (e.g., age, hours-per-week) be scaled (e.g., using StandardScaler or MinMaxScaler)?
* 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?
* 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?

# 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.
* Histogram of hours-per-week distribution.
* Bar chart of education levels frequency.
* Bar chart of workclass categories.
* Bar chart of occupation categories.
* Pie chart or bar chart of gender distribution.
* Pie chart of race distribution.
* Bar chart of native-country (top 10 most frequent countries).
* Histogram of capital-gain and capital-loss values.
* Countplot of income distribution (<=50K vs >50K).

**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. |