**GERMAN CREDIT DATA DATASHEET**

**The German Credit dataset is a popular dataset used for binary classification tasks, such as predicting whether a person has good or bad credit based on various attributes. Here’s a datasheet summarizing the key information about the dataset:**

**German Credit Dataset - UCI Repository**

**Source: UCI Machine Learning Repository  
Dataset URL: German Credit Dataset - UCI Repository**

**1. Dataset Overview**

* **Purpose: The dataset is used for binary classification, where the goal is to predict whether an individual has "Good" or "Bad" credit based on several demographic and financial attributes.**
* **Number of Instances: 1,000 instances (rows)**
* **Number of Attributes: 20 attributes (features)**
* **Class Distribution:**
  + **Good Credit (0): 700 instances (70%)**
  + **Bad Credit (1): 300 instances (30%)**

**2. Attribute Information**

**The dataset consists of 20 features, divided into categorical and numerical types. The target variable is binary, with two classes: Good (1) and Bad (0) credit.**

**Categorical Attributes:**

1. **Checking Account Status**
   * **Values: no checking, little checking, moderate checking, rich checking**
2. **Credit History**
   * **Values: critical/other existing credit, existing credits paid back duly, delayed previously, critical/other existing credit**
3. **Purpose**
   * **Values: new car, used car, furniture/equipment, radio/tv, domestic appliance, education, retirement pension, other**
4. **Current Balance Savings**
   * **Values: little savings, moderate savings, rich savings**
5. **Employment**
   * **Values: unemp/unemp highly qualified, emp longer than 1 year, emp 1-4 years, emp 4-7 years, emp 7+ years**
6. **Personal Status and Sex**
   * **Values: male single, female div/dep/mar, male div/sep**
7. **Other Parties**
   * **Values: none, guarantor, party**
8. **Residential Status**
   * **Values: own residence, for free, rent**
9. **Property-Magnitude**
   * **Values: real estate, life insurance, car, no known property**
10. **Other Payment Plans**
    * **Values: none, bank, stores**
11. **Housing**
    * **Values: own, for free, rent**
12. **Existing Credit**
    * **Values: none, bank**
13. **Job**
    * **Values: skilled, high qualif/self emp/mgmt, unemp/unemp highly qualified, others**
14. **People Liable for Maintenance**
    * **Values: none, 1, 2, 3**
15. **Telephone**
    * **Values: none, yes**
16. **Foreign Worker**
    * **Values: yes, no**

**Numerical Attributes:**

1. **Age**
   * **Type: Numeric (continuous, age of the individual)**
2. **Debt Duration**
   * **Type: Numeric (duration in months)**
3. **Credit Amount**
   * **Type: Numeric (credit amount requested in Deutsche Marks)**
4. **Average Credit Balance**
   * **Type: Numeric (average credit balance in the past 6 months)**

**3. Target Attribute**

* **Target: Credit Approval (Good or Bad)**
  + **Bad Credit (0): Represents individuals with poor credit, i.e., high risk of defaulting on payments.**
  + **Good Credit (1): Represents individuals with good credit, i.e., low risk of defaulting on payments.**

**4. Data Preprocessing Considerations**

* **Handling Missing Values: The dataset is generally well-structured but may contain some missing values in certain categorical features.**
* **Encoding: Categorical features should be encoded using techniques like one-hot encoding or label encoding for machine learning models.**
* **Feature Scaling: Numerical features, such as Age, Credit Amount, and Debt Duration, may need normalization or standardization depending on the model used.**

**5. Evaluation Metrics**

* **Accuracy: The proportion of correct predictions (good or bad credit).**
* **Precision: The proportion of true positive predictions (good credit) among all predicted good credit instances.**
* **Recall: The proportion of true positive predictions (good credit) among all actual good credit instances.**
* **F1-Score: The harmonic mean of precision and recall.**
* **ROC-AUC: The area under the Receiver Operating Characteristic curve, indicating the model’s ability to discriminate between the two classes.**

**7. Use Cases**

* **Credit Scoring: Used to assess the creditworthiness of applicants.**
* **Risk Assessment: Helps in determining the financial risks associated with lending to certain individuals.**
* **Machine Learning Model Training: Ideal for classification algorithms like Logistic Regression, Decision Trees, Random Forests, SVM, etc.**

**Target\_Variable:**

**Target Variable Distribution:**

**1 700**

**2 300**

**Name: count, dtype: int64**

****

**[7]:**

**Attributes:**

Distribution for checking\_account:

checking\_account

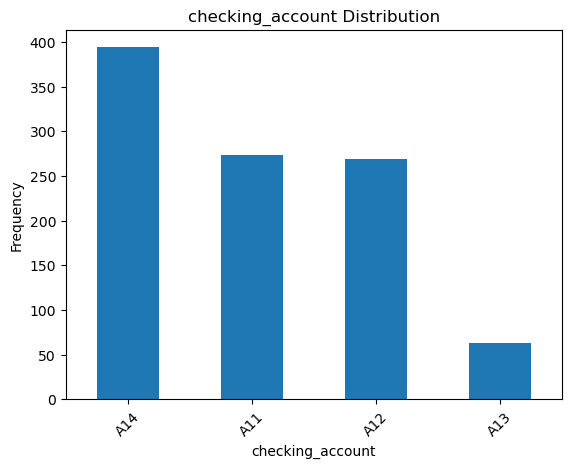
A14 394

A11 274

A12 269

A13 63

Name: count, dtype: int64



Distribution for history:

history

A32 530

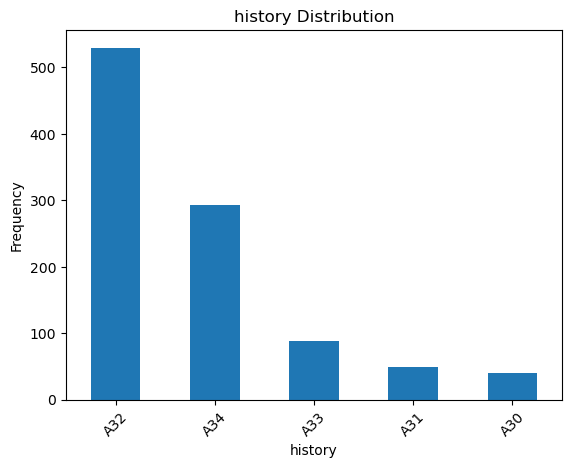
A34 293

A33 88

A31 49

A30 40

Name: count, dtype: int64



Distribution for purpose:

purpose

A43 280

A40 234

A42 181

A41 103

A49 97

A46 50

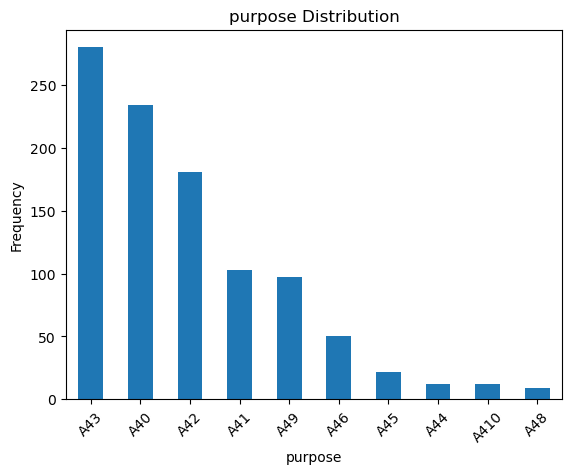
A45 22

A44 12

A410 12

A48 9

Name: count, dtype: int64



Distribution for savings\_account:

savings\_account

A61 603

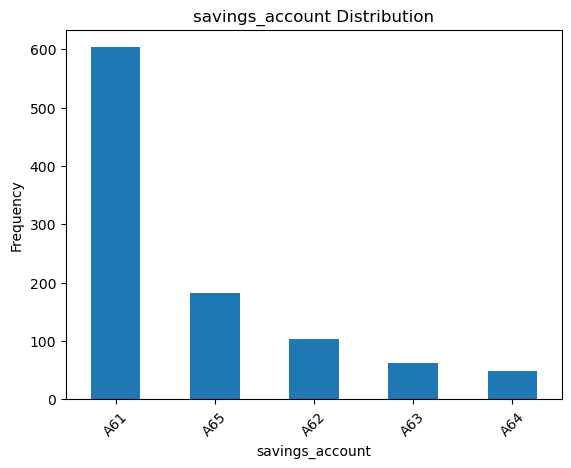
A65 183

A62 103

A63 63

A64 48

Name: count, dtype: int64



Distribution for employment:

employment

A73 339

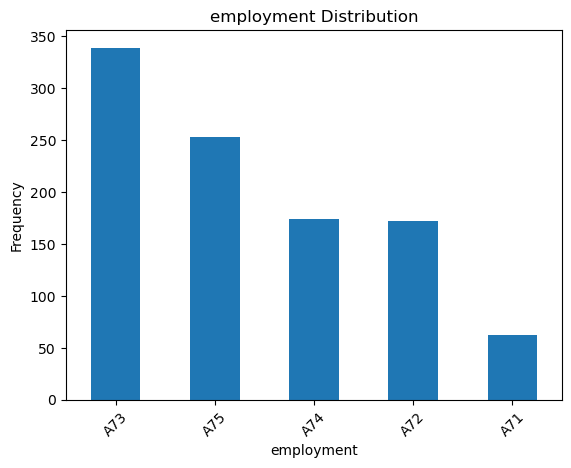
A75 253

A74 174

A72 172

A71 62

Name: count, dtype: int64



Distribution for personal\_status:

personal\_status

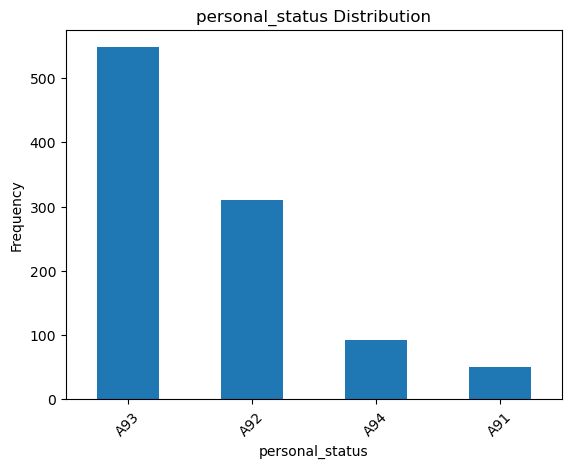
A93 548

A92 310

A94 92

A91 50

Name: count, dtype: int64



Distribution for other\_parties:

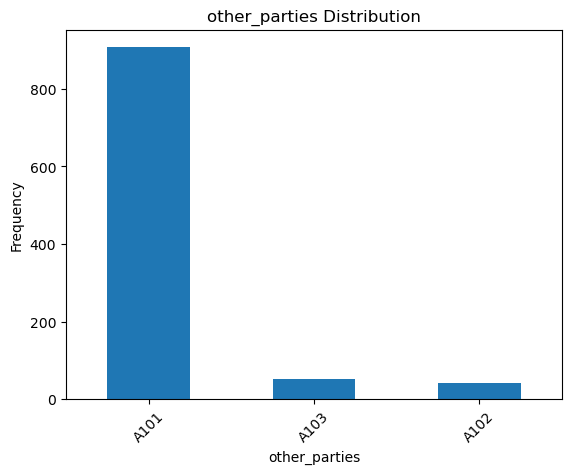
other\_parties

A101 907

A103 52

A102 41

Name: count, dtype: int64



Distribution for property\_magnitude:

property\_magnitude

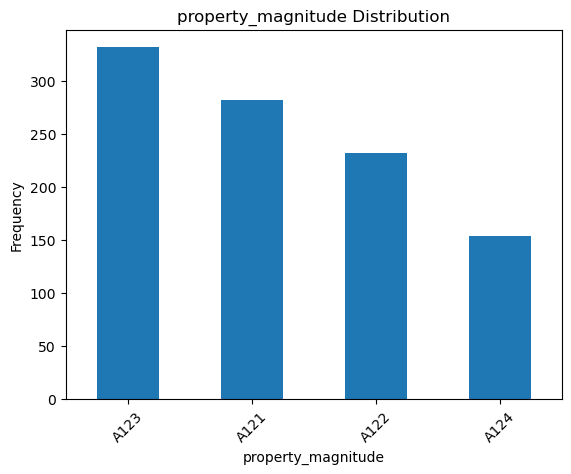
A123 332

A121 282

A122 232

A124 154

Name: count, dtype: int64



Distribution for other\_payment\_plans:

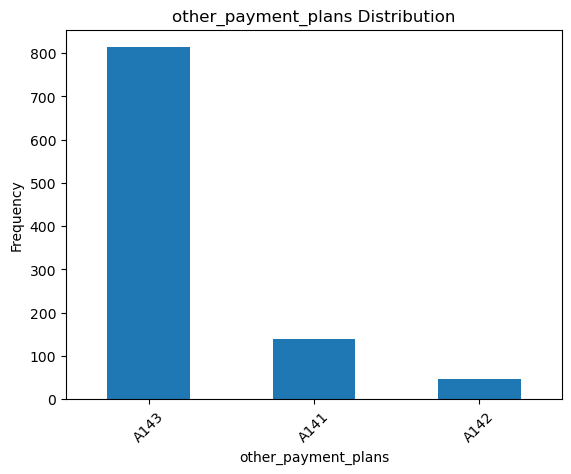
other\_payment\_plans

A143 814

A141 139

A142 47

Name: count, dtype: int64



Distribution for housing:

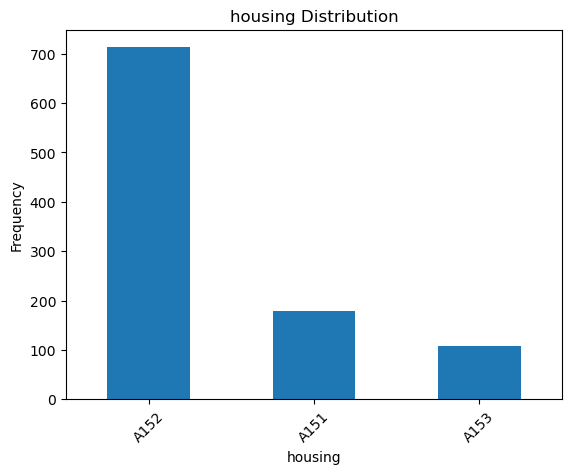
housing

A152 713

A151 179

A153 108

Name: count, dtype: int64



Distribution for job:

job

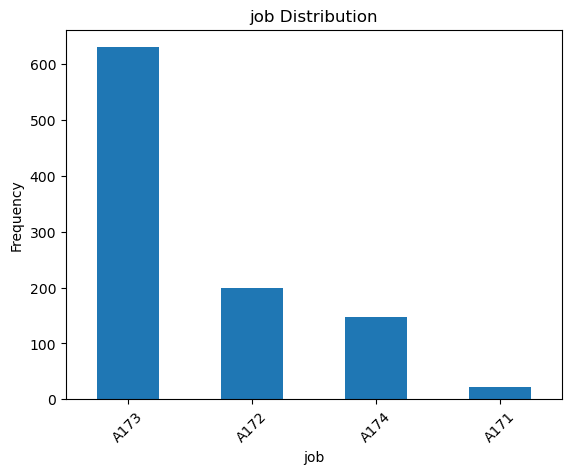
A173 630

A172 200

A174 148

A171 22

Name: count, dtype: int64



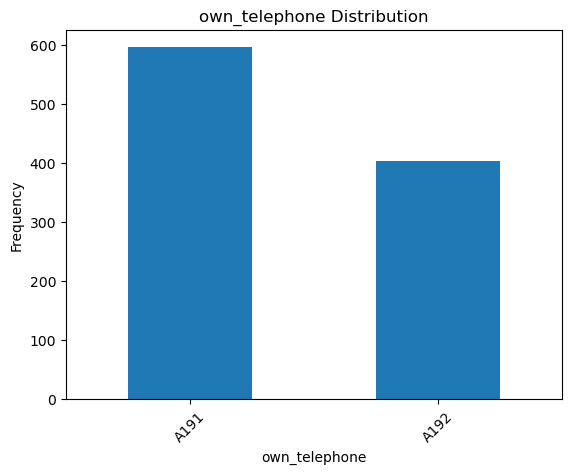
Distribution for own\_telephone:

own\_telephone

A191 596

A192 404

Name: count, dtype: int64



Distribution for foreign\_worker:

foreign\_worker

A201 963

A202 37

Name: count, dtype: int64

Summary statistics for continuous features:

duration credit\_amount location residence\_since age \

count 1000.000000 1000.000000 1000.000000 1000.000000 1000.000000

mean 20.903000 3271.258000 2.973000 2.845000 35.546000

std 12.058814 2822.736876 1.118715 1.103718 11.375469

min 4.000000 250.000000 1.000000 1.000000 19.000000

25% 12.000000 1365.500000 2.000000 2.000000 27.000000

50% 18.000000 2319.500000 3.000000 3.000000 33.000000

75% 24.000000 3972.250000 4.000000 4.000000 42.000000

max 72.000000 18424.000000 4.000000 4.000000 75.000000

existing\_credits num\_dependents

count 1000.000000 1000.000000

mean 1.407000 1.155000

std 0.577654 0.362086

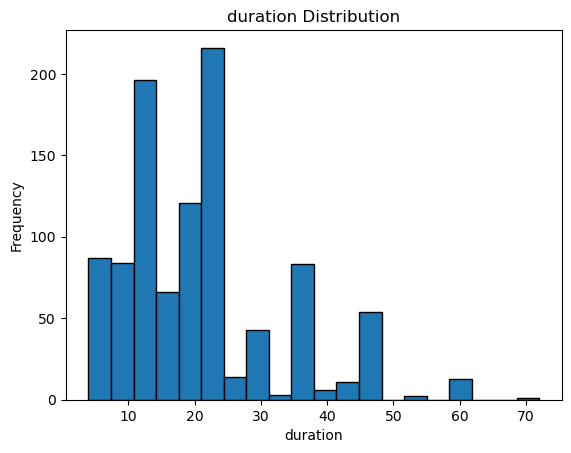
min 1.000000 1.000000

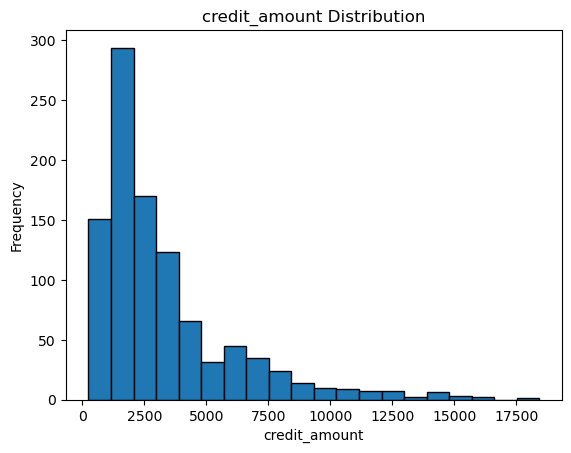
25% 1.000000 1.000000

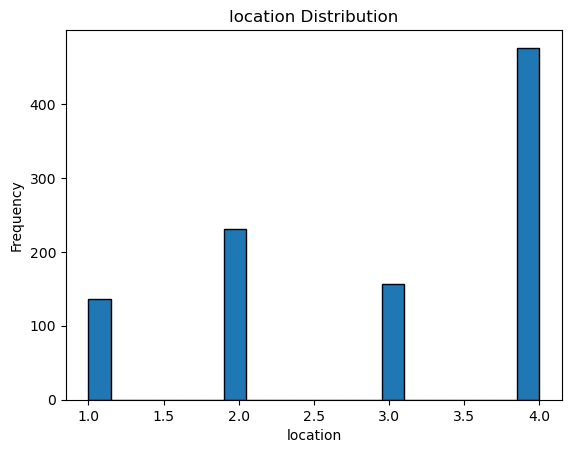
50% 1.000000 1.000000

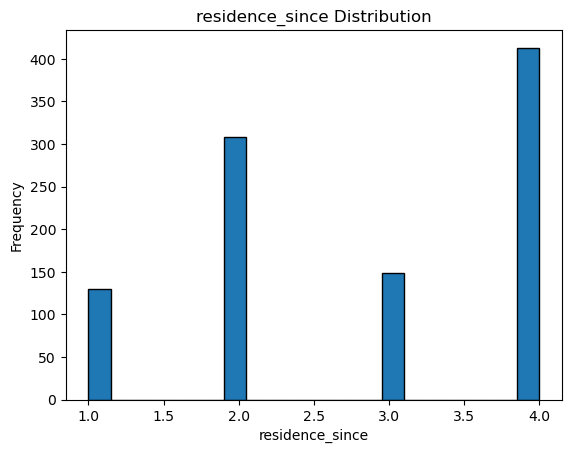
75% 2.000000 1.000000

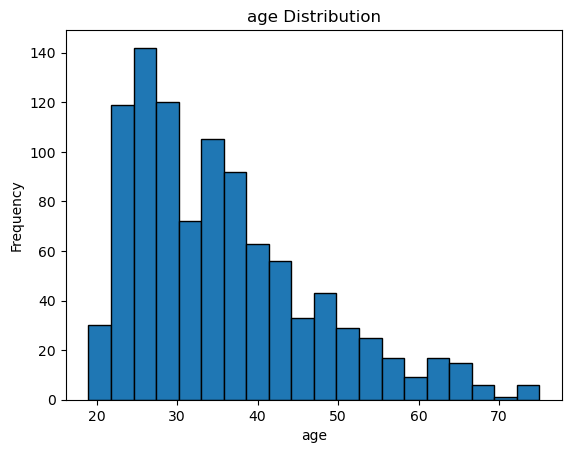
max 4.000000 2.000000

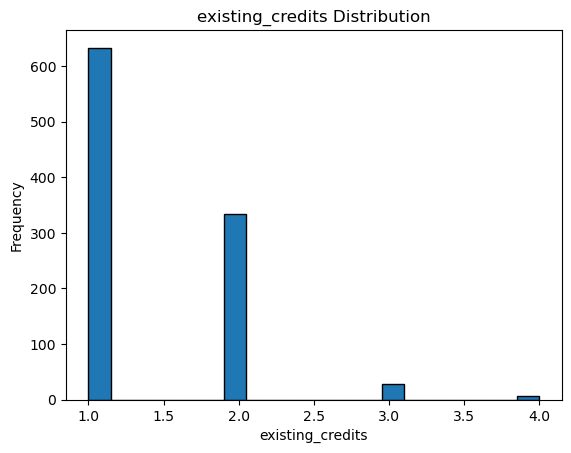


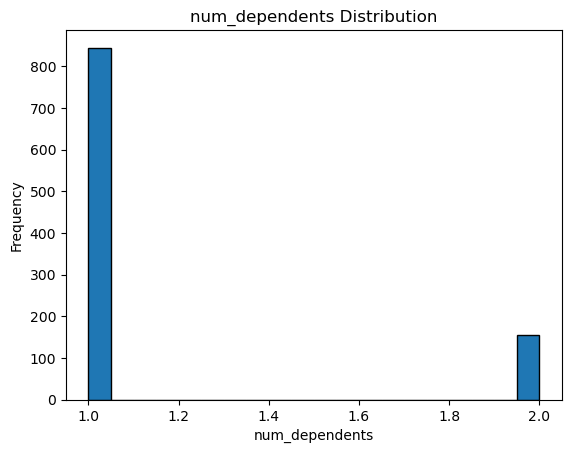












[11]:

