# Machine Learning Analysis on AVC and Salary Datasets

Dragomir Andrei

May 25, 2024

**Abstract**

This document provides a comprehensive analysis of two machine learning models applied to the given datasets. The analysis includes data preprocessing, model training, evaluation, and comparison of results.

# Contents

# 1 Introduction

In this analysis, we explore the Logistic Regression and the Multy Layer Perceptron models on the AVC and Salary datasets. The goal is to predict the target variable based on the features provided in the datasets.

# Dataset Description

# 2 Medical and Lifestyle Information Dataset

This dataset contains medical values and relevant lifestyle information for 5110 individuals. The dataset is intended to be used for predicting whether a person is likely to have a cerebrovascular accident (stroke) or not. The target attribute is *cerebrovascular_accident*, which has binary values indicating the presence (1) or absence (0) of a stroke. The classification task is binary.

The attributes in the dataset are as follows:

Table 1: Attributes of the Medical and Lifestyle Information Dataset

| **Attribute** | **Type** | **Description** |
|---|---|---|
| mean_blood_sugar_level | numeric | The average blood sugar level over the observation period. |
| cardiovascular_issues | categorical | Indicates if the subject has a history of cardiovascular issues. Possible values: 0, 1. |
| job_category | categorical | The domain in which the person works. Possible values: child, entrepreneurial, N_work_history, private_sector, public_sector. |
| body_mass_indicator | numeric | The body mass index indicating if the person is underweight, normal weight, overweight, or obese. |
| sex | categorical | The gender of the person. Possible values: F, M. |
| tobacco_usage | categorical | Indicator for tobacco use, present or past. Possible values: ex-smoker, smoker, non-smoker. |
| high_blood_pressure | categorical | Binary attribute indicating if a person suffers from high blood pressure. Possible values: 0, 1. |
| married | categorical | Binary attribute indicating if the person has ever been married. Possible values: Y, N. |
| living_area | categorical | Type of area where the person has lived most of their life. Possible values: City, Countryside. |
| years_old | numeric | The age of the person in years. |
| chaotic_sleep | categorical | Binary attribute for an irregular sleep schedule. Possible values: 0, 1. |
| analysis_results | numeric | Results of the person's medical tests, which may include various measurements and health indicators. |
| biological_age_index | numeric | An index estimating the biological age of a person based on various factors such as lifestyle and health status. |
| cerebrovascular_accident | categorical | Binary indicator of whether the person has had a stroke. Possible values: 0, 1. |

# 3 Employee Information Dataset

This dataset contains personal, educational, and professional information of various employees. The objective of this dataset is the binary classification of employees into categories of earning above or below $50K per year. The classification task is binary.

The attributes in the dataset are as follows:

Table 2: Attributes of the Employee Information Dataset

| **Attribute** | **Type** | **Description** |
|---|---|---|
| fnl | numeric | Socio-economic characteristic of the population from which the individual comes. |
| hpw | numeric | Number of hours worked per week. |
| relation | categorical | The type of relationship in which the individual is involved. |
| gain | numeric | Capital gain. |
| country | categorical | Country of origin. |
| job | categorical | The individual's occupation. |
| edu_int | numeric | Number of years of education. |
| years | numeric | Age of the individual. |
| loss | numeric | Capital loss. |
| work_type | categorical | Type of occupation. |
| partner | categorical | Type of partner the individual has. |
| edu | categorical | Type of education of the individual. |
| gender | categorical | Gender of the individual. |
| race | categorical | Race of the individual. |
| prod | numeric | Capital production. |
| gtype | categorical | Type of work contract. |

# 4 Attribute Type Identification

Before utilizing a machine learning model for a dataset, it is crucial to identify the types of features in the dataset and their values. Understanding the nature of the attributes helps in selecting appropriate preprocessing techniques and models. The key distinctions among the types of attributes in the provided datasets are as follows:

## 4.1 Continuous Numerical Attributes

Continuous numerical attributes are features that can take any value within a given range. These attributes are measured on a continuous scale and can be divided into finer increments. Examples from the datasets include:

- **mean_blood_sugar_level**: Average blood sugar level measured over the observation period.
- **body_mass_indicator**: Body mass index indicating whether a person is underweight, normal weight, overweight, or obese.
- **years_old**: Age of the person in years.

- **analysis_results**: Results of medical tests.
- **biological_age_index**: An index estimating the biological age of a person based on various factors.
- **fnl**: Socio-economic characteristic of the population the individual comes from.
- **hpw**: Number of hours worked per week.
- **gain**: Capital gain.
- **edu_int**: Number of years of education.
- **years**: Age of the individual.
- **loss**: Capital loss.
- **prod**: Capital production.

## 4.2 Discrete Attributes

Discrete attributes are features that take on a finite number of distinct values. These values are often categorical and can be counted in whole numbers. Examples from the datasets include:

- **cardiovascular_issues**: Whether the subject has a history of cardiovascular issues (0 or 1).
- **sex**: Gender of the person (F or M).
- **tobacco_usage**: Indicator for smokers, either past or present (ex-smoker, smoker, non-smoker).
- **high_blood_pressure**: Indicator if a person has high blood pressure (0 or 1).
- **married**: Whether the person has ever been married (Y or N).
- **living_area**: Type of area where the person has lived most of their life (City or Countryside).
- **chaotic_sleep**: Indicator for an irregular sleep schedule (0 or 1).
- **cerebrovascular_accident**: Indicator if the person has had a stroke (0 or 1).
- **relation**: Type of relationship the individual is involved in.
- **country**: Country of origin.
- **job**: Job of the individual.
- **work_type**: Type of job.
- **partner**: Type of partner the individual has.
- **edu**: Type of education of the individual.

- **gender**: Gender of the individual.
- **race**: Race of the individual.
- **gtype**: Type of work contract.
- **job_category**: Domain in which the person works.

Recognizing these distinctions helps in choosing the right methods for handling the data during preprocessing. For instance, continuous numerical attributes might require normalization or standardization, discrete attributes might need encoding, and ordinal attributes might need ordinal encoding to maintain the order information. Properly identifying and classifying the attributes ensures that the machine learning models can effectively learn from the data and make accurate predictions.

# 5 Numeric Attributes Analysis

In this section, we analyze the numeric attributes of the two datasets provided. The analysis includes the number of non-missing values, mean value, standard deviation, minimum value, 25th percentile, 50th percentile (median), 75th percentile, and maximum value for each numeric attribute.

## 5.1 Healthcare Dataset

Table 3: Statistics of Numeric Attributes in the Healthcare Dataset

| Attribute | No-miss | Mean | Std Dev | Min | 25th Pctl | Mid | 75th Pctl | Max |
|---|---|---|---|---|---|---|---|---|
| mean_blood_sugar_level | 5110 | 106.15 | 45.28 | 55.12 | 77.25 | 91.89 | 114.09 | 271.74 |
| body_mass_indicator | 4909 | 28.89 | 7.85 | 10.30 | 23.50 | 28.10 | 33.10 | 97.60 |
| years_old | 5110 | 46.57 | 26.59 | 0.08 | 26.00 | 47.00 | 63.75 | 134.00 |
| analysis_results | 4599 | 323.52 | 101.58 | 104.83 | 254.65 | 301.03 | 362.82 | 756.81 |
| biological_age_index | 5110 | 134.78 | 50.40 | -15.11 | 96.71 | 136.37 | 172.51 | 266.99 |

**Comments:**

- The **mean_blood_sugar_level** has a mean of 106.15 with a standard deviation of 45.28, indicating a wide range of values.
- The **body_mass_indicator** shows a mean of 28.89, suggesting that on average, individuals fall into the overweight category.
- The **years_old** attribute ranges from 0.08 to 134 years, with a median of 47 years.
- The **analysis_results** attribute has significant variability, as indicated by its standard deviation of 101.58.
- The **biological_age_index** has negative values, which may need to be investigated further for data correctness.

## 5.2 Employee Dataset

Table 4: Statistics of Numeric Attributes in the Employee Dataset

| Attribute | No-miss | Mean | Std Dev | Min | 25th Pctl | Mid | 75th Pctl | Max |
|---|---|---|---|---|---|---|---|---|
| fnl | 9999 | 190352.9 | 106070.8 | 19214 | 118282.5 | 178472 | 237311 | 1455435 |
| hpw | 9199 | 40.42 | 12.52 | 1.0 | 40.0 | 40.0 | 45.0 | 99.0 |
| gain | 9999 | 979.85 | 7003.80 | 0.0 | 0.0 | 0.0 | 0.0 | 99999.0 |
| edu_int | 9999 | 14.26 | 24.77 | 1.0 | 9.0 | 10.0 | 13.0 | 206.0 |
| years | 9999 | 38.65 | 13.75 | 17.0 | 28.0 | 37.0 | 48.0 | 90.0 |
| loss | 9999 | 84.11 | 394.04 | 0.0 | 0.0 | 0.0 | 0.0 | 3770.0 |
| prod | 9999 | 2014.93 | 14007.60 | -28.0 | 42.0 | 57.0 | 77.0 | 200125.0 |

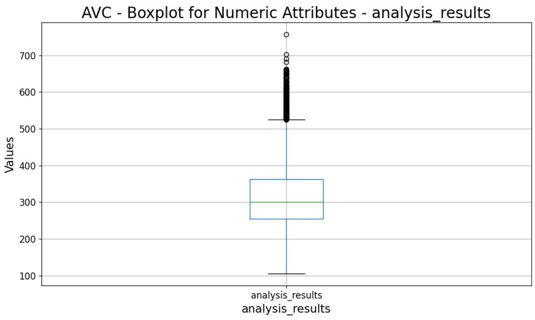
**Comments:**

- The **fnl** attribute has a high mean value and a wide range, indicating significant differences in the socio-economic status of individuals.
- The **hpw** (hours per week) attribute shows that most individuals work around 40 hours per week.
- The **gain** and **loss** attributes have a high standard deviation, indicating that only a few individuals have large capital gains or losses.
- The **edu_int** (years of education) attribute shows that the majority of individuals have around 10 to 13 years of education.
- The **years** attribute shows an average age of around 38.65 years, with a minimum of 17 and a maximum of 90 years.
- The **prod** attribute has a high standard deviation, indicating variability in capital production.

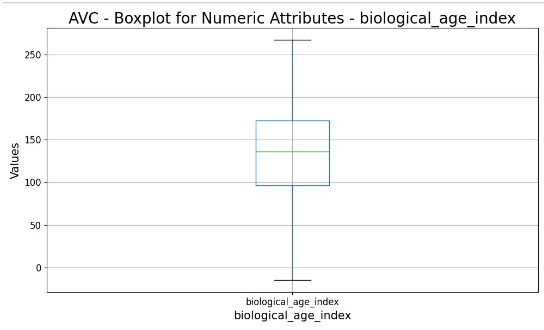
# 6 Numeric Ranges for Attributes

In this section, we present the boxplots for the numeric attributes in the datasets to visualize the distribution and identify any potential outliers.

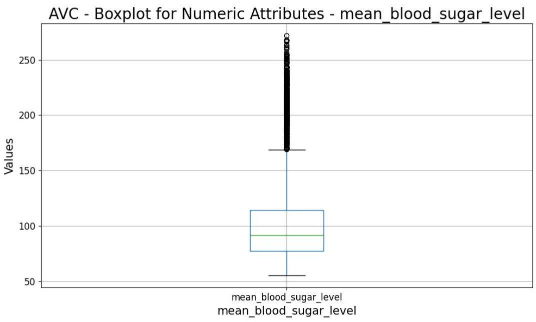
## 6.1 Healthcare Dataset



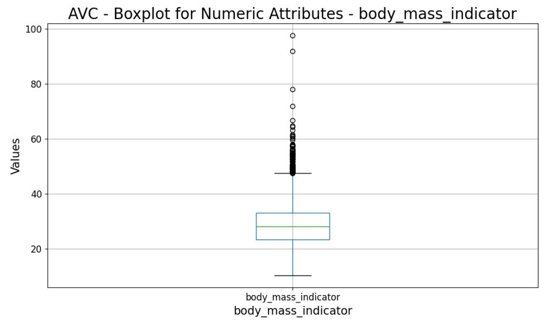

(a) Analysis Results



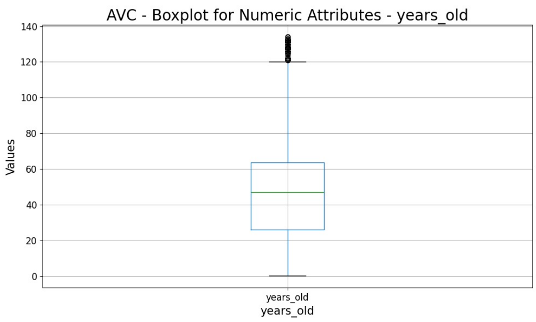

(b) Biological Age Index




(c) Mean Blood Sugar Level



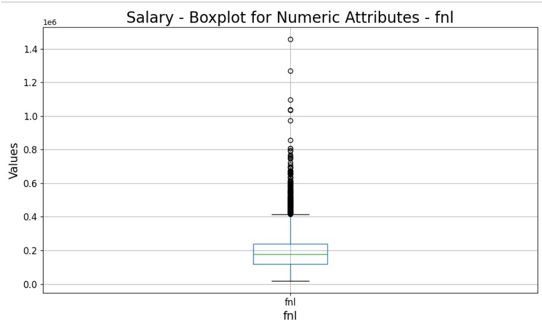

(d) Body Mass Indicator



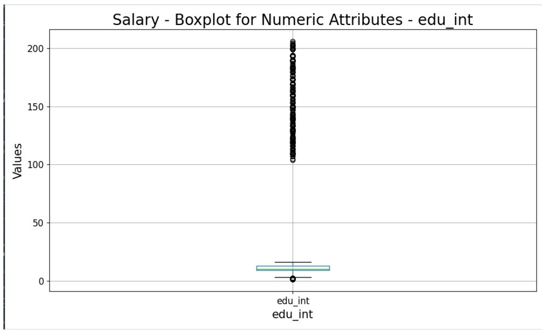

(e) Years Old

Figure 1: Boxplots for Healthcare Dataset Numeric Attributes

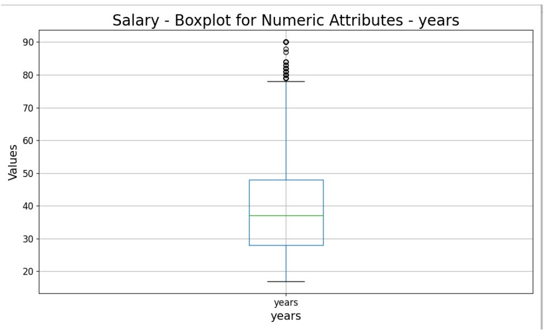
## 6.2 Employee Dataset



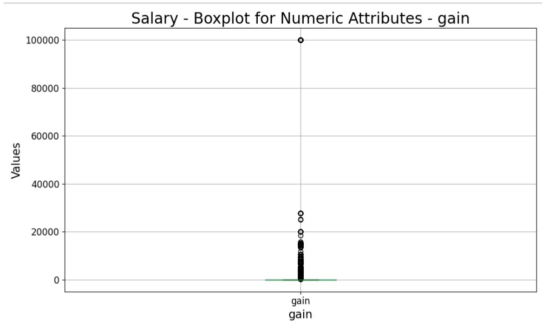

(a) FNL



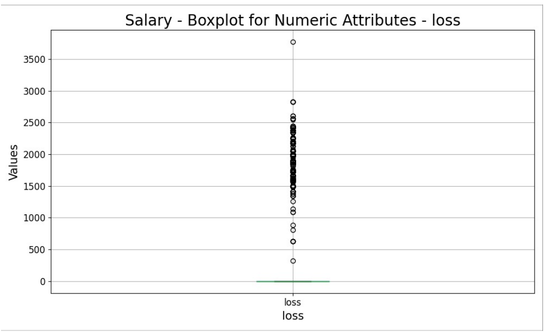

(b) Education (Years)



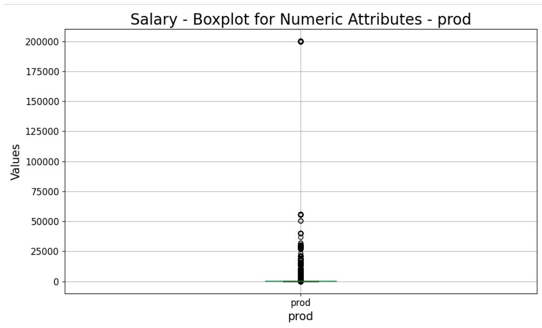

(c) Years




(d) Gain




(e) Loss




(f) Prod

Figure 2: Boxplots for Employee Dataset Numeric Attributes

# 7 Categorical Attributes Analysis

In this section, we analyze the categorical attributes of the datasets. The tables below present the attributes along with the number of non-missing values and the number of unique values for each attribute.

## 7.1 Healthcare Dataset

Table 5: Categorical Attributes - Healthcare Dataset

| **Attribute** | **Number of Non-Missing Values** | **Number of Unique Values** |
|---|---|---|
| cardiovascular_issues | 5110 | 2 |
| job_category | 5110 | 5 |
| sex | 5110 | 2 |
| tobacco_usage | 5110 | 4 |
| high_blood_pressure | 5110 | 2 |
| married | 4599 | 2 |
| living_area | 5110 | 2 |
| chaotic_sleep | 5110 | 2 |
| cerebrovascular_accident | 5110 | 2 |

## 7.2 Employee Dataset

Table 6: Categorical Attributes - Employee Dataset

| **Attribute** | **Number of Non-Missing Values** | **Number of Unique Values** |
|---|---|---|
| relation | 9999 | 6 |
| country | 9999 | 41 |
| job | 9999 | 14 |
| work_type | 9999 | 9 |
| partner | 9999 | 7 |
| edu | 9999 | 16 |
| gender | 9199 | 2 |
| race | 9999 | 5 |
| gtype | 9999 | 2 |
| money | 9999 | 2 |

### 7.3 Comments

The healthcare dataset includes several categorical attributes such as 'cardiovascular_issues', 'job_category', 'sex', and 'tobacco_usage'. Each of these attributes has a varying number of unique values, indicating different levels of categorization. For instance, 'tobacco_usage' has four unique values, while 'sex' has only two. It is also noted that the 'married' attribute has fewer non-missing values compared to others, which might require special handling during data preprocessing.

The employee dataset, on the other hand, contains a wider range of categorical attributes with a larger variety of unique values, especially for 'country' and 'edu', which have 41 and 16 unique values, respectively. Attributes like 'gender' and 'gtype' are binary, which simplifies their processing. The variety in the number of unique values across different attributes in both datasets indicates the complexity and diversity of the data, which must be carefully considered during the feature engineering and model building stages.

# 8 Categorical Distribution for Attributes

## 8.1 Healthcare Dataset

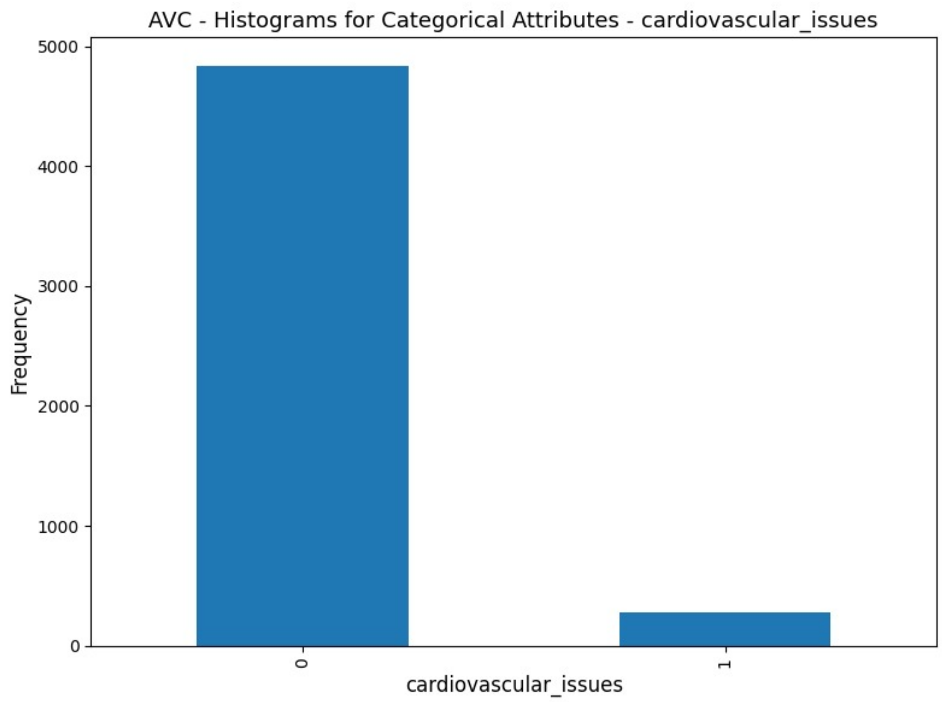



Figure 3: Cardiovascular Issues

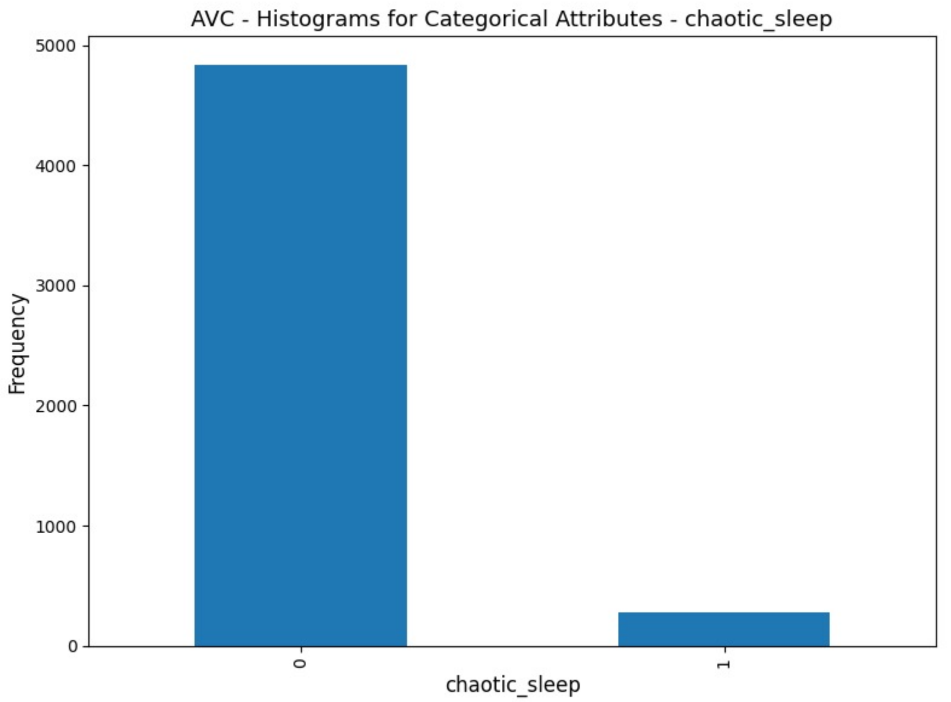



Figure 4: Chaotic Sleep

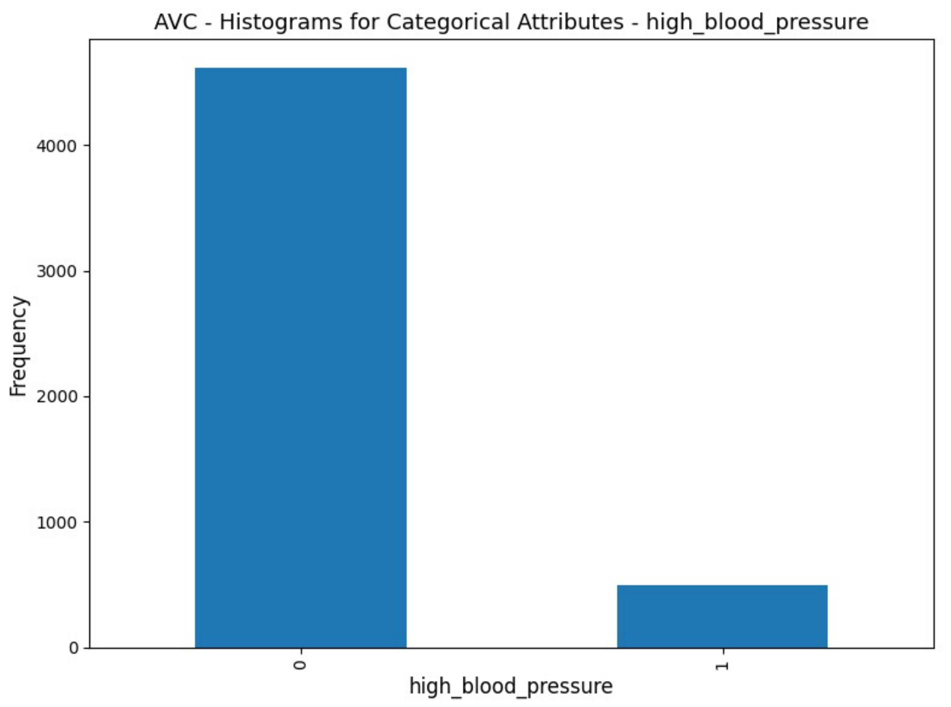



Figure 5: High Blood Pressure

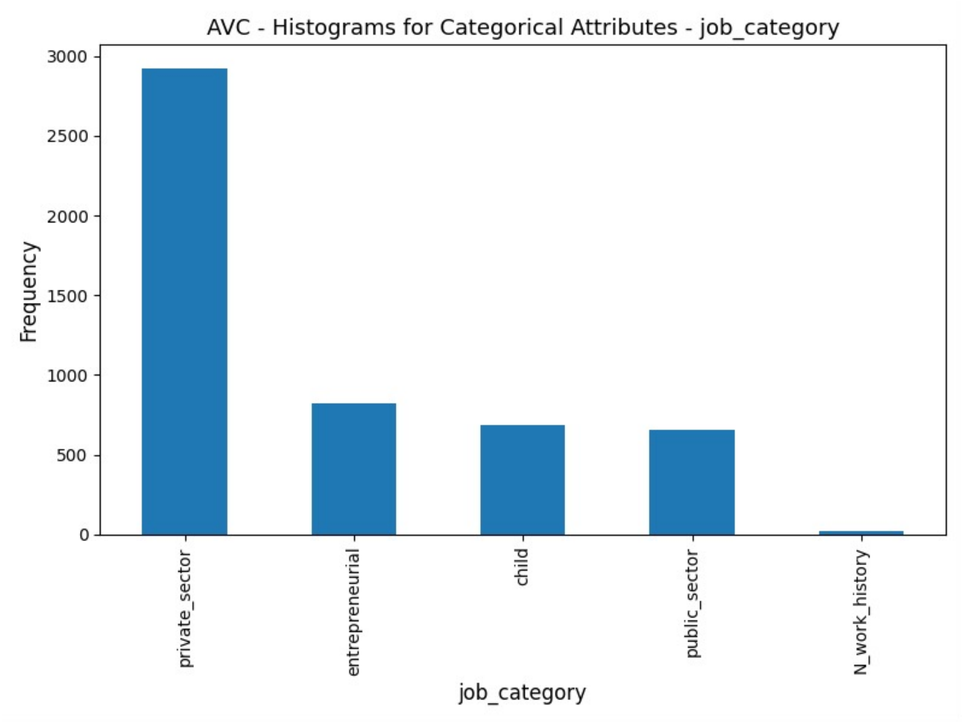



Figure 6: Job Category

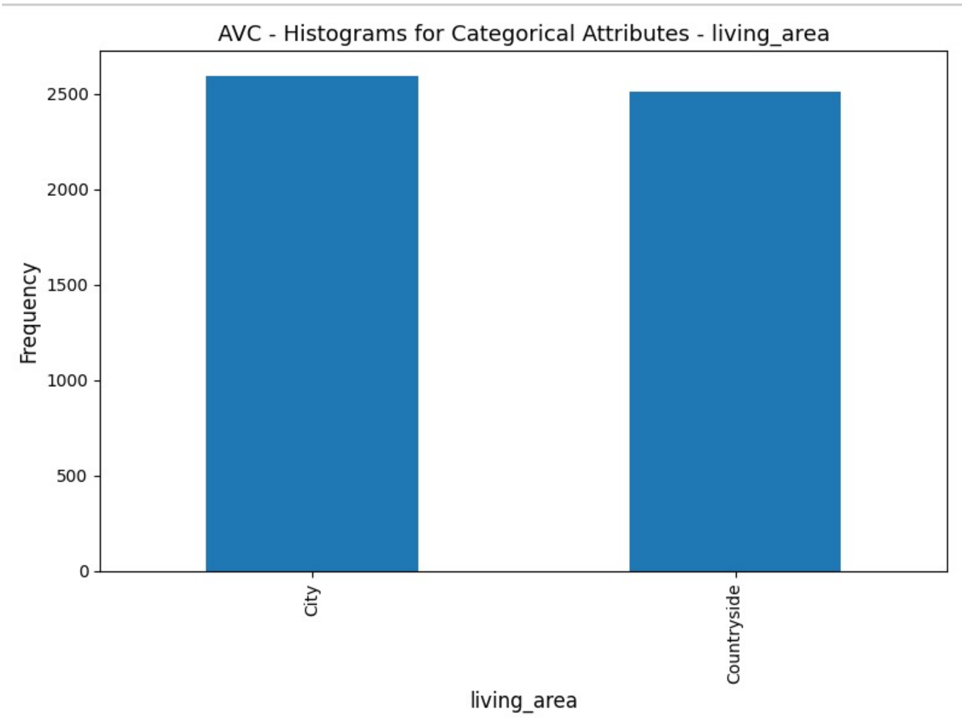



Figure 7: Living Area

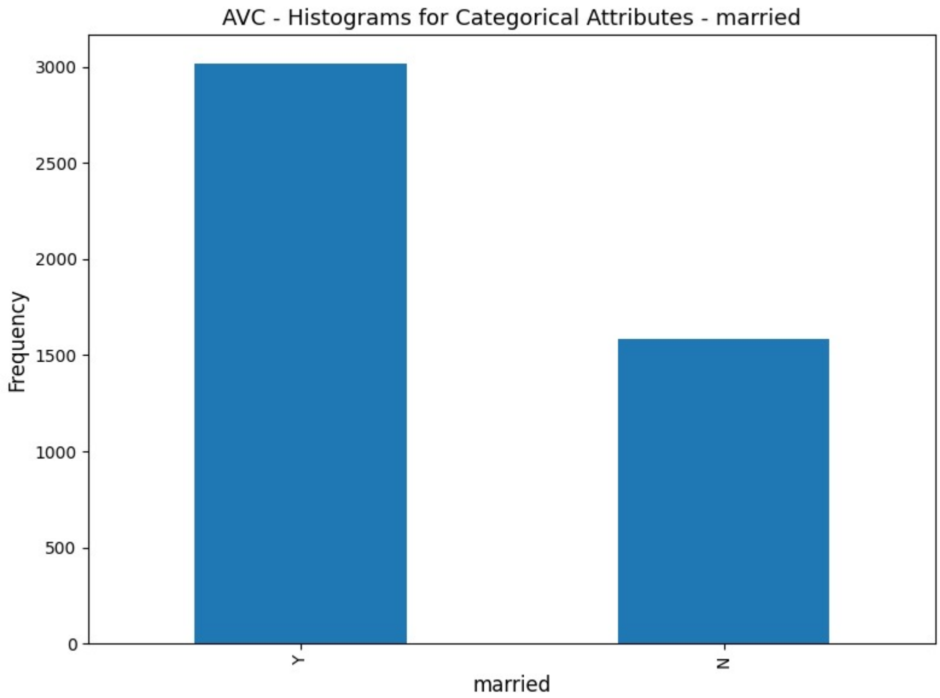



Figure 8: Married

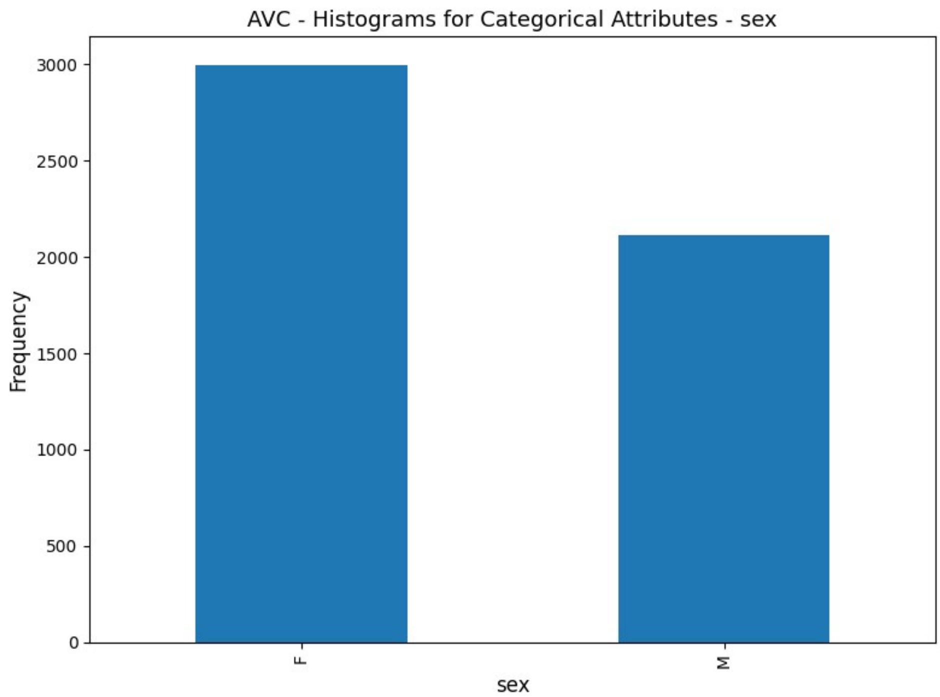



Figure 9: Sex

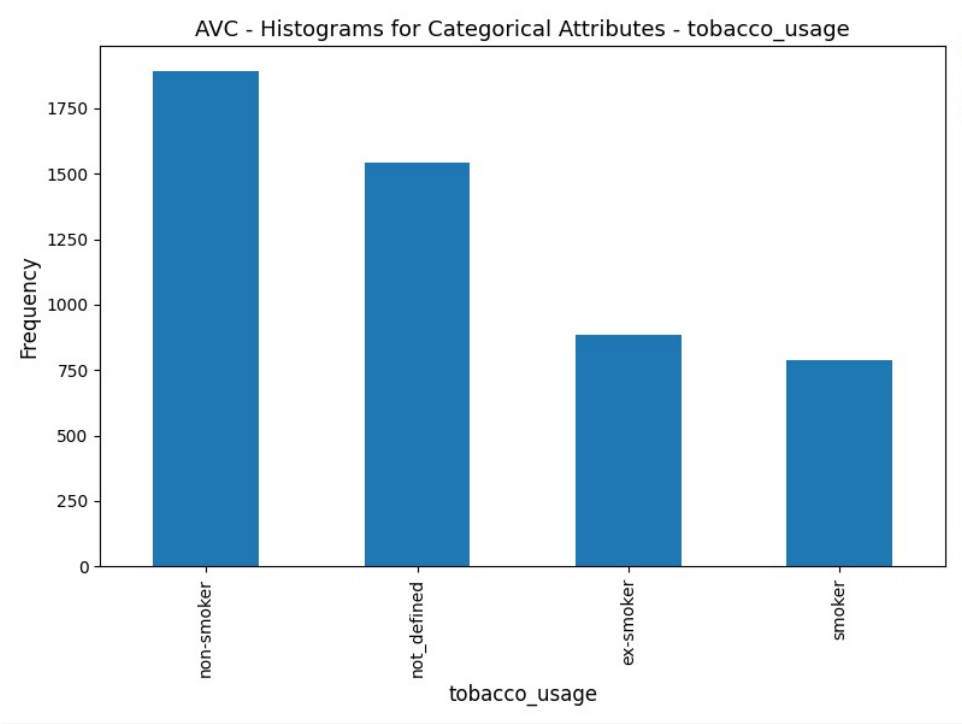



Figure 10: Tobacco Usage

## 8.2 Salary Dataset

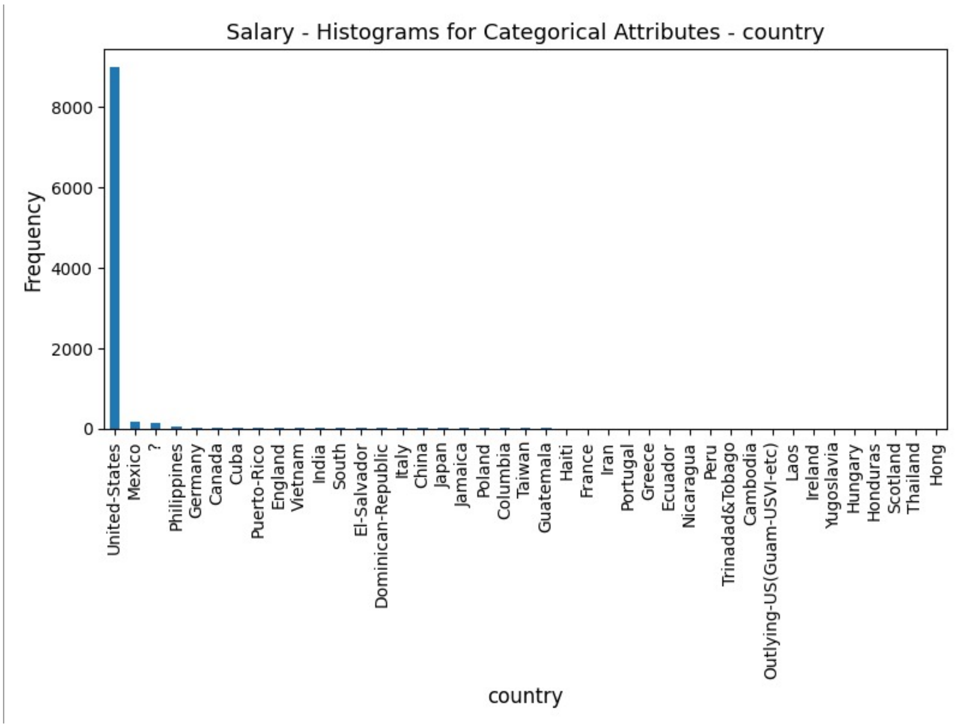



Figure 11: Country

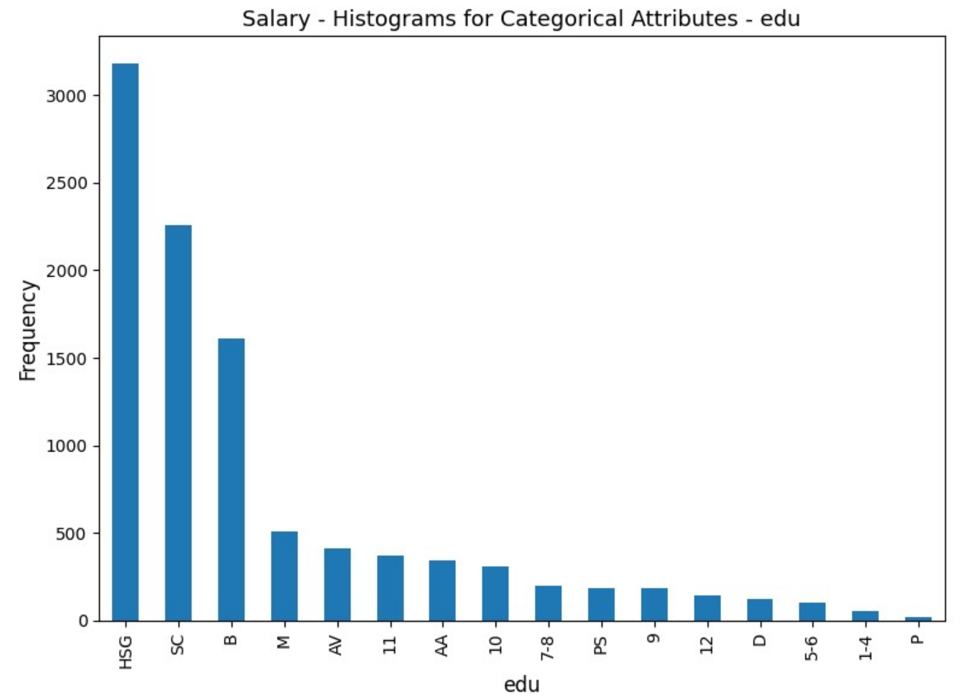



Figure 12: Education

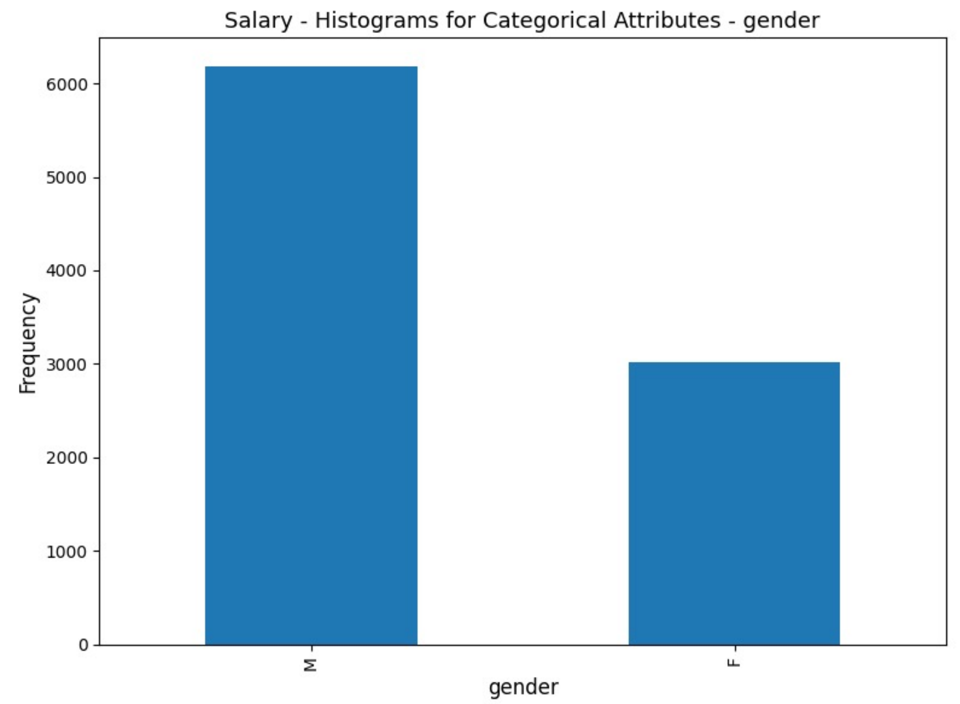



Figure 13: Gender

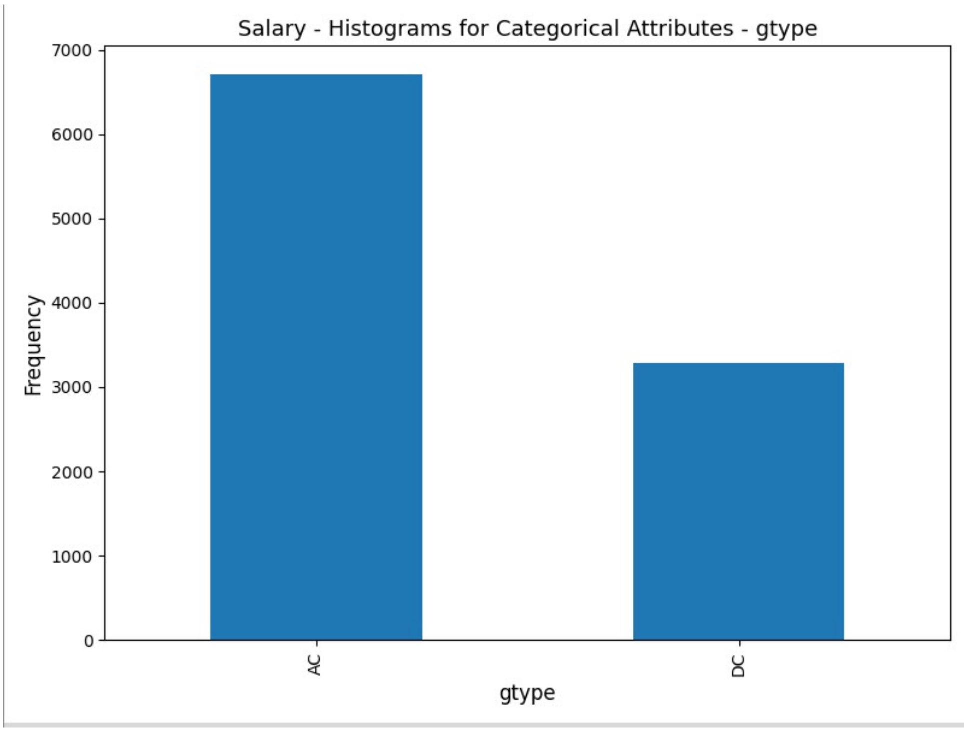



Figure 14: Job Type

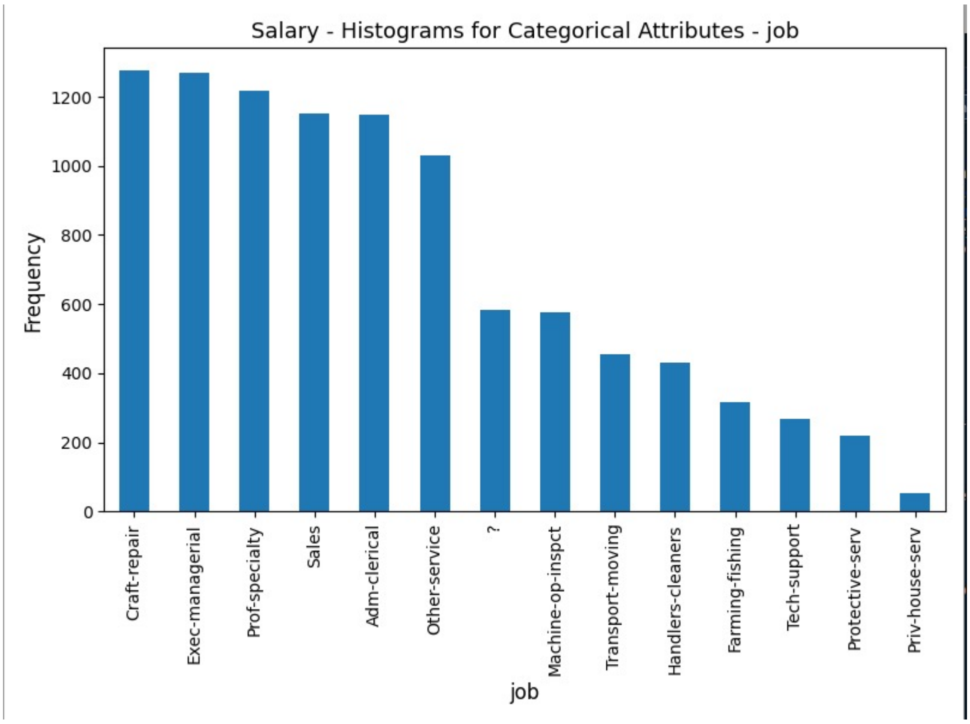



Figure 15: Job

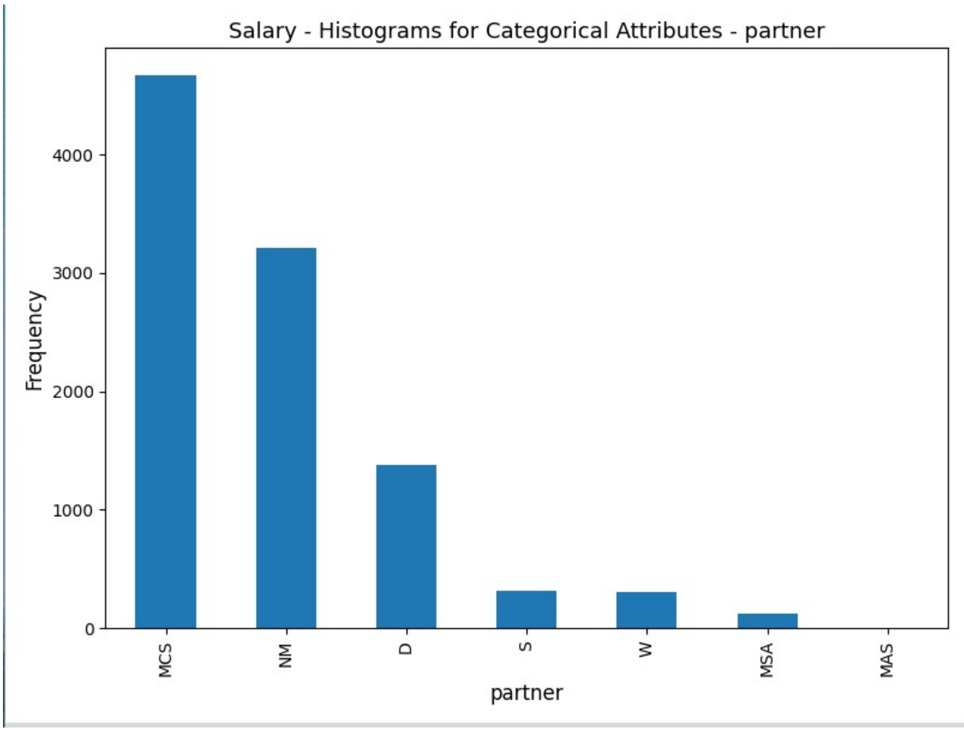



Figure 16: Partner

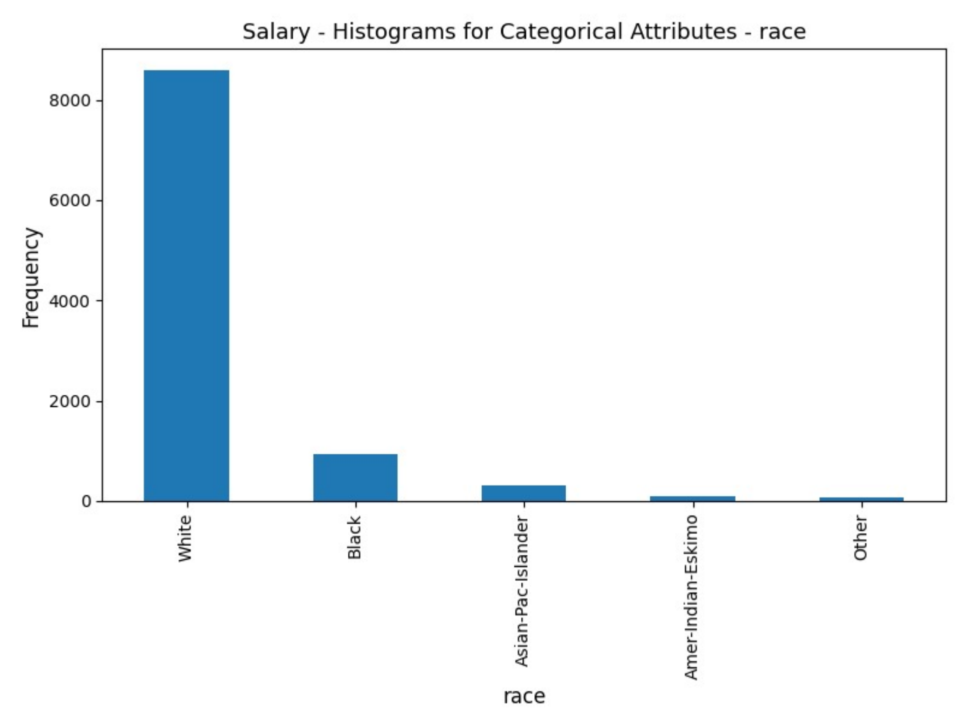



Figure 17: Race

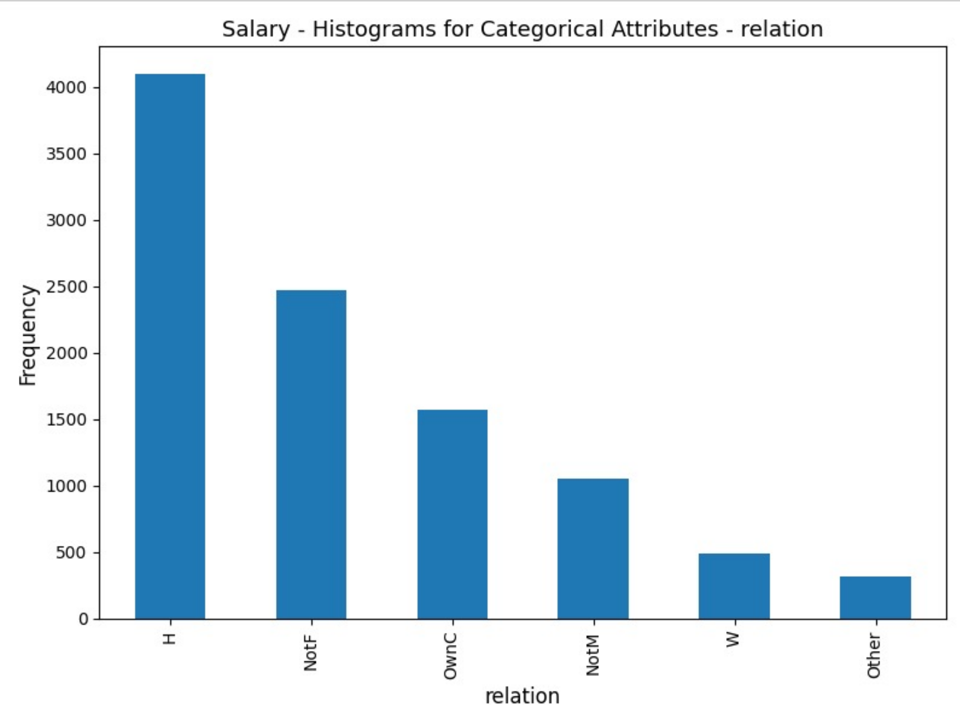



Figure 18: Relation

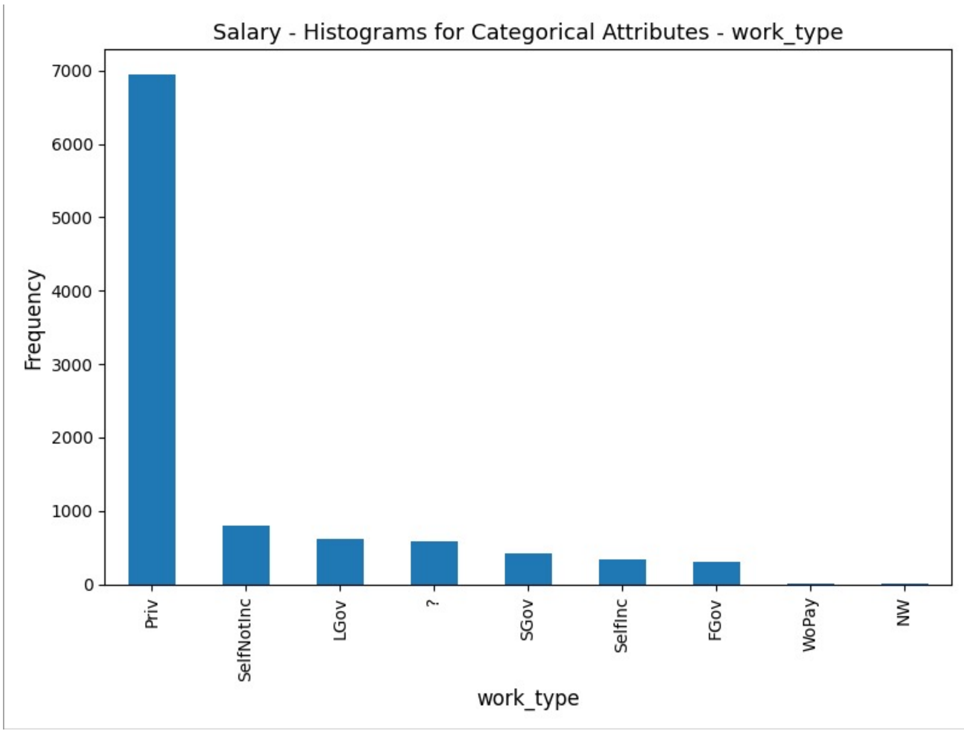



Figure 19: Work Type

# 9 Class Balance

This section presents the class balance for both the AVC and Salary datasets. The class balance is an important aspect to consider as it influences the evaluation metrics we should focus on.

## 9.1 AVC Dataset

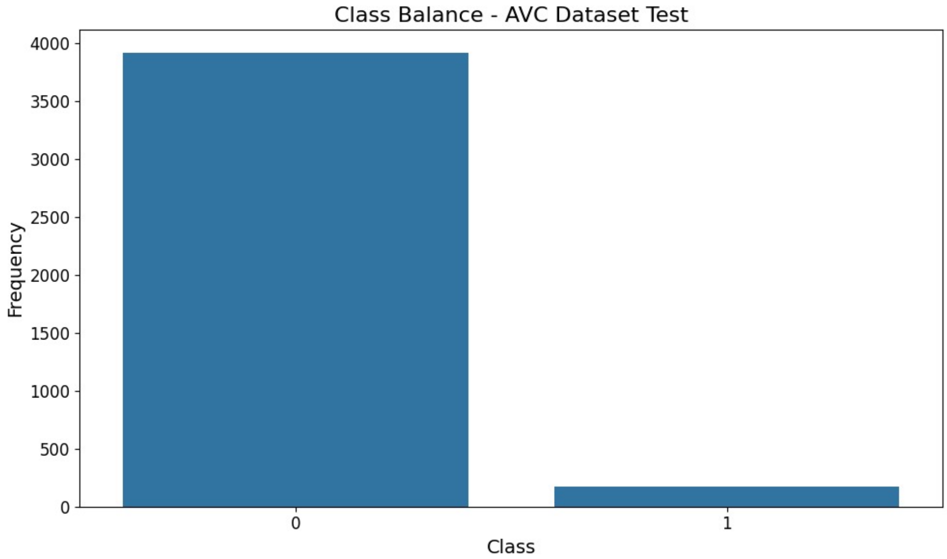



Figure 20: Class Balance - AVC Dataset Test

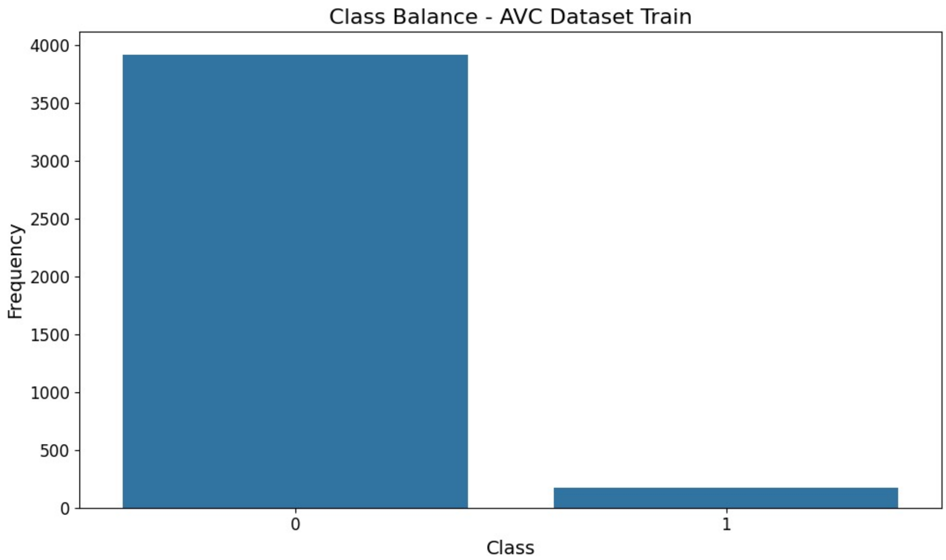



Figure 21: Class Balance - AVC Dataset Train

## 9.2 Salary Dataset

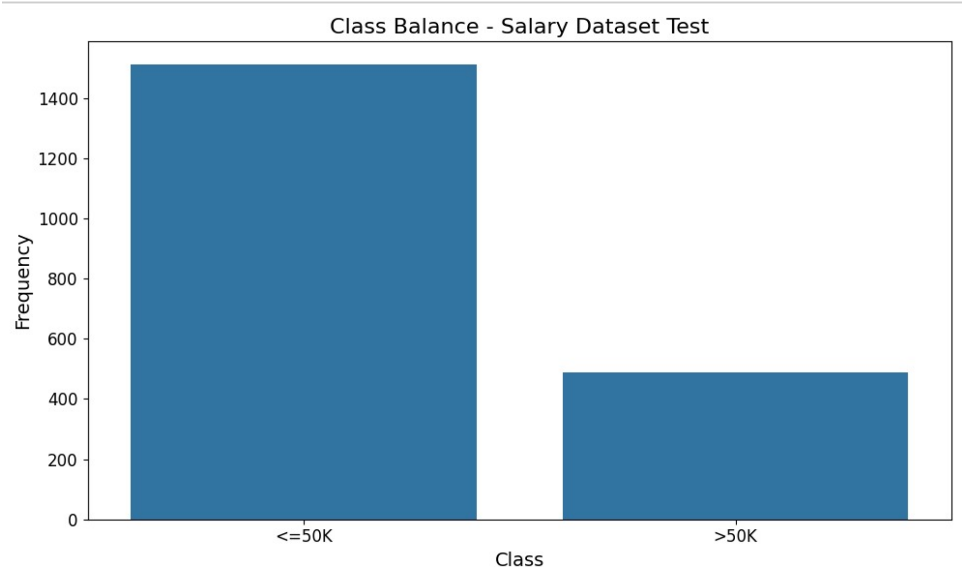



Figure 22: Class Balance - Salary Dataset Test

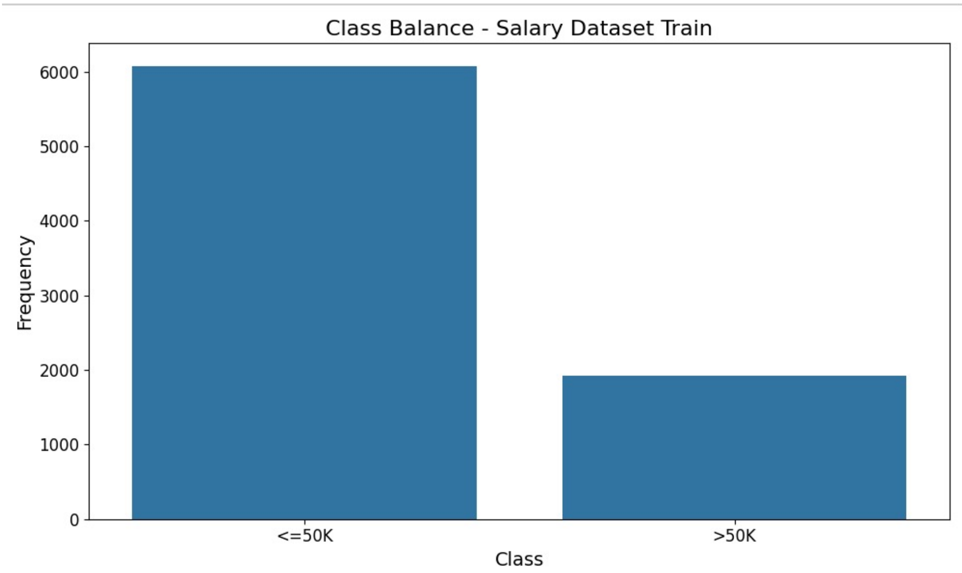



Figure 23: Class Balance - Salary Dataset Train

## 9.3 Comments

From the class balance plots above, it is evident that both datasets exhibit significant class imbalance. In the AVC dataset, the majority class (no stroke) is much more prevalent than the minority class (stroke). Similarly, in the Salary dataset, the class of individuals earning less than or equal to 50K is more common than the class of individuals earning more than 50K.

Class imbalance can have a substantial impact on model performance and evaluation. Models may become biased towards the majority class, leading to poor performance on the minority class. To address this, it is crucial to focus on evaluation metrics that provide a better understanding of performance on imbalanced datasets. Specifically, we should emphasize:

- **F1 Score**: The harmonic mean of precision and recall, which provides a balance between the two.
- **Precision**: The ability of the classifier to not label a negative sample as positive.
- **Recall**: The ability of the classifier to find all positive samples.

By prioritizing these metrics, we can better assess and compare the performance of different algorithms on these imbalanced datasets.

# 10 Attribute Correlation

This section presents the correlation matrices for both the AVC and Salary datasets. The correlation matrices are split into numerical and categorical attributes for both train and test sets.

## 10.1 AVC Dataset

### 10.1.1 Numerical Attributes

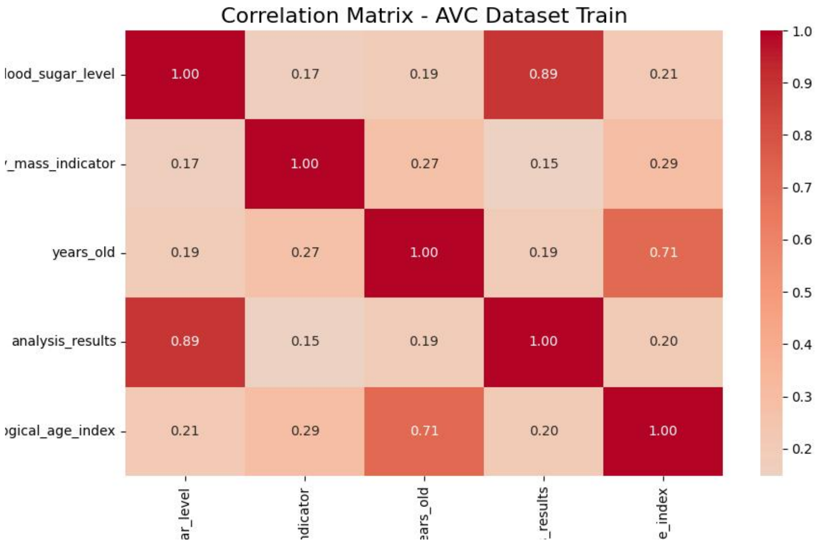



Figure 24: Correlation Matrix - Numerical Attributes - AVC Dataset Train

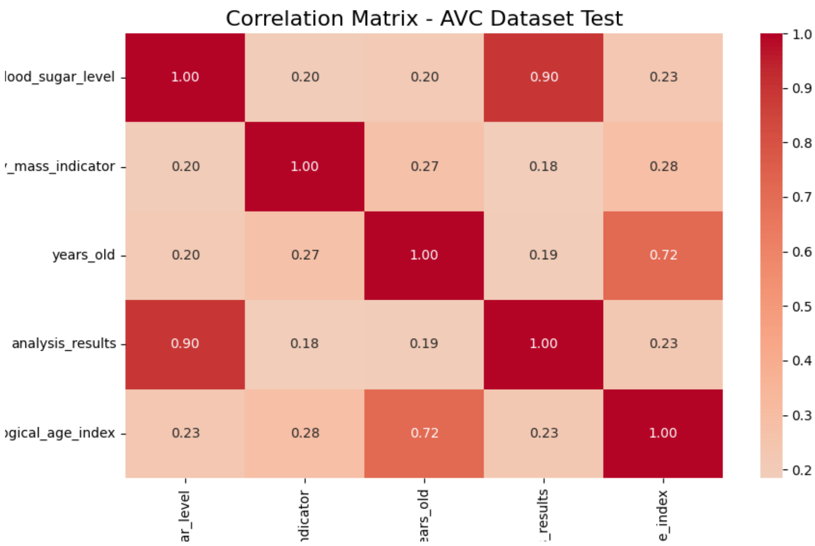



Figure 25: Correlation Matrix - Numerical Attributes - AVC Dataset Test

### 10.1.2 Categorical Attributes

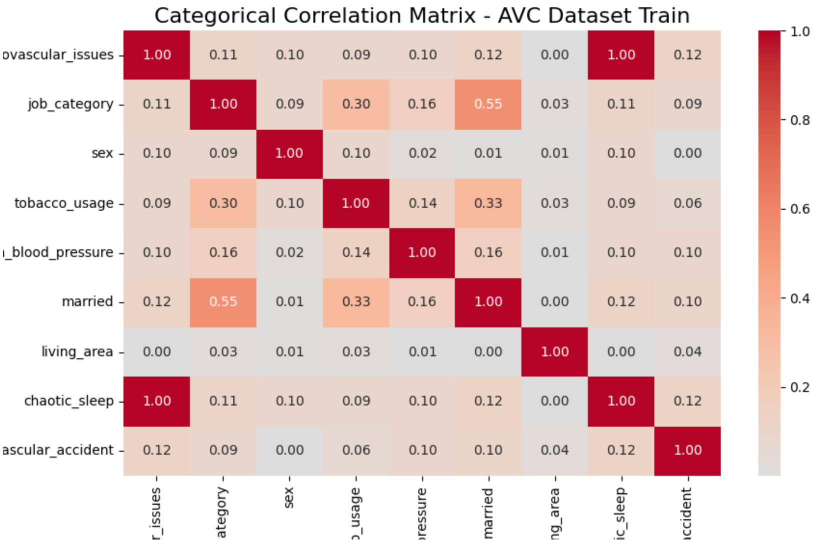



Figure 26: Categorical Correlation Matrix - AVC Dataset Train

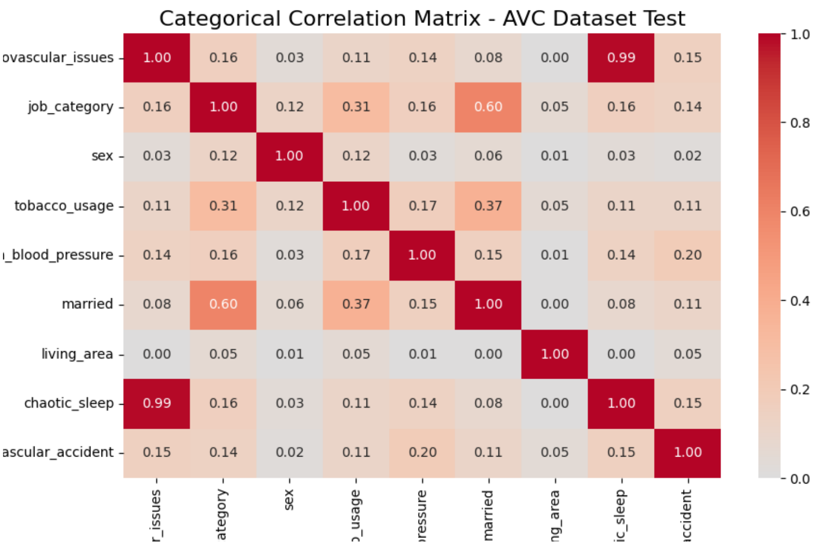



Figure 27: Categorical Correlation Matrix - AVC Dataset Test

## 10.2 Salary Dataset

### 10.2.1 Numerical Attributes

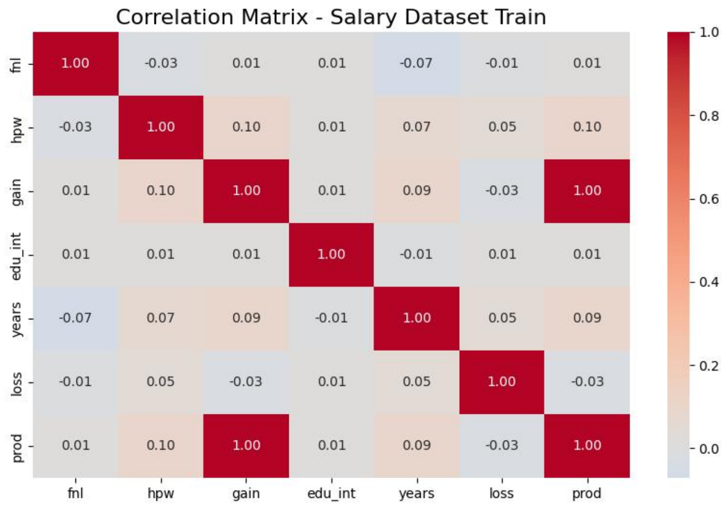



Figure 28: Correlation Matrix - Numerical Attributes - Salary Dataset Train

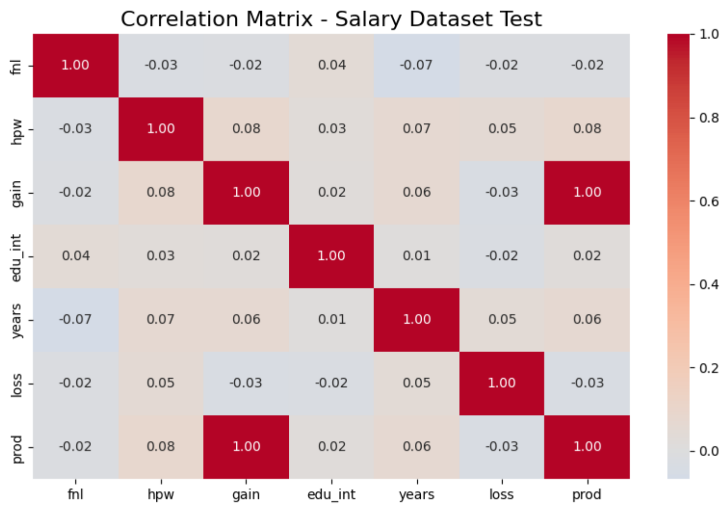



Figure 29: Correlation Matrix - Numerical Attributes - Salary Dataset Test

### 10.2.2 Categorical Attributes

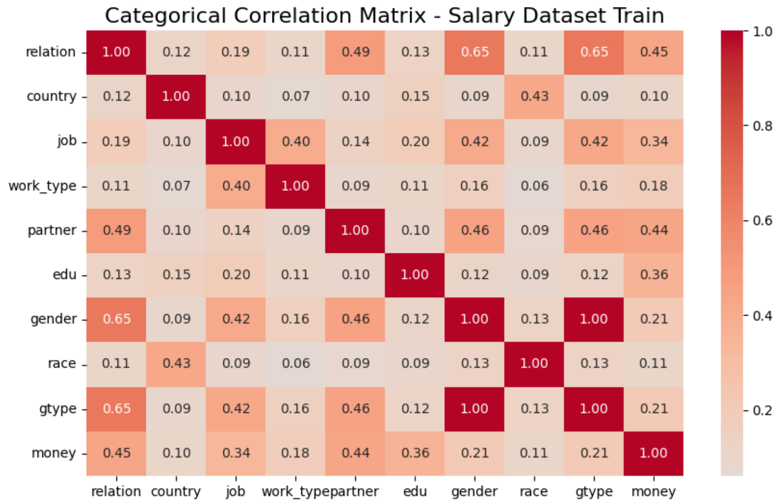



Figure 30: Categorical Correlation Matrix - Salary Dataset Train

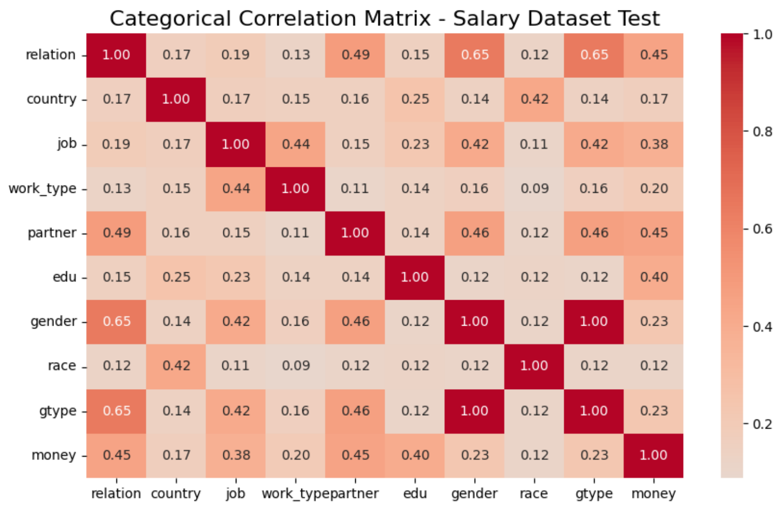



Figure 31: Categorical Correlation Matrix - Salary Dataset Test

## 10.3 Comments

The correlation matrices provide insights into the relationships between different attributes in the datasets.

### 10.3.1 AVC Dataset

**Numerical Attributes** From the numerical attribute correlation matrices of the AVC dataset, it is observed that:

- There is a moderate positive correlation between *years_old* and *biological_age_index* in both train and test sets, which indicates that as the age of the person increases, the biological age index also tends to be higher.
- *Analysis_results* show a high correlation with *blood_sugar_level*, but low with other numerical attributes, suggesting that the results of medical analyses are relatively independent of the person's age and other indicators.

**Categorical Attributes** The categorical correlation matrices for the AVC dataset indicate:

- There are some moderate correlations between categorical variables such as *married* and *job_category*, and *job_category* and *tobacco_usage*.
- These correlations suggest that there might be certain patterns or dependencies between the demographic and lifestyle attributes.
- There is a high correlation between *chaotic_sleep* and *high_blood_pressure*, indicating that individuals with irregular sleep schedules are more likely to have high blood pressure.

### 10.3.2 Salary Dataset

**Numerical Attributes** For the Salary dataset, the numerical attribute correlation matrices reveal:

- A strong positive correlation between *gain* and *prod*, indicating that individuals who gain more also tend to have higher produce of capital.
- *Hours per week (hpw)* shows a low correlation with most other numerical attributes, suggesting that the number of hours worked per week is not strongly related to other numerical factors like capital gains or losses.

**Categorical Attributes** In the categorical correlation matrices for the Salary dataset:

- Moderate correlations are observed between *work_type* and *job*, as well as between *country* and *race*.
- These correlations highlight possible demographic and professional patterns.
- There is also a high correlation between *gtype* and *gender*, indicating that the type of work contract is directly influenced by the gender.

Overall, these correlation matrices help in understanding the interdependencies between attributes, which can be useful in feature selection and in improving the performance of machine learning models.

Eliminating specific variables such as *analysis_results* and *biological_age_index* from the AVC dataset, and prod from the Salary dataset, can enhance the efficiency and performance of our machine learning models. These variables were identified as candidates for elimination due to their high correlation with other attributes, which makes them redundant, and their minimal impact on predictive accuracy. By removing these redundant or less impactful features, we reduce noise, prevent overfitting, and simplify the model, leading to improved generalization and more robust predictive performance. This streamlined approach allows the model to focus on the most significant attributes, enhancing both computational efficiency and model interpretability.