Data Science Tools and Techniques

****

**ASSIGNMENT 1**

**Course Instructor: Dr. Mateen Yaqoob**

**Section: MS DS A**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Submitted By:**

**Valeena Afzal   
25I-8023**

Table of Contents

[Introduction 3](#_Toc208671698)

[Task 1: Data Preprocessing 3](#_Toc208671699)

[Issues Identified 3](#_Toc208671700)

[Solutions Applied 4](#_Toc208671701)

[Task 2: Exploratory Data Analysis (EDA) 5](#_Toc208671702)

[Before Preprocessing 5](#_Toc208671703)

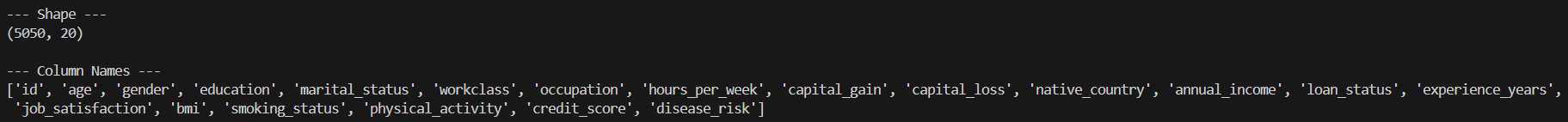
[After Preprocessing 6](#_Toc208671704)

[Key Insights 8](#_Toc208671705)

[Conclusion 8](#_Toc208671706)

# Introduction

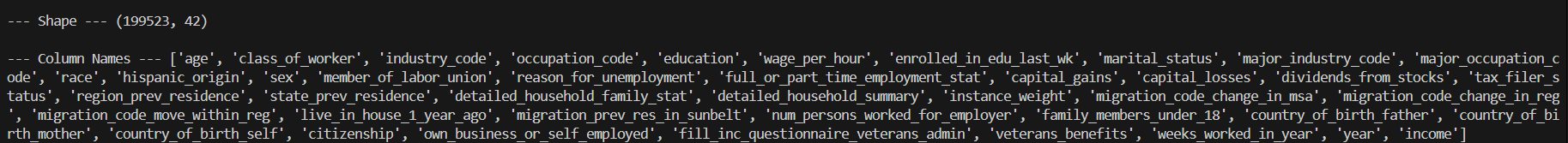
1. Attached Dataset was adult data set having 20 columns and 5050 rows.

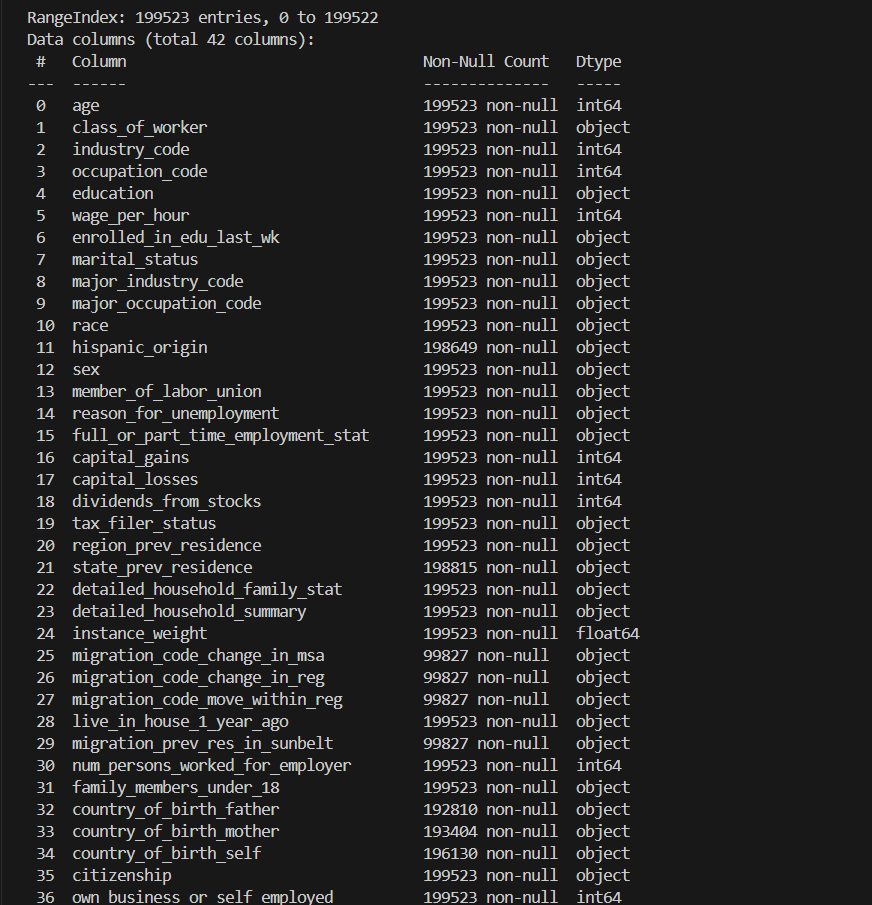


A screenshot of a computer

AI-generated content may be incorrect.

1. Census dataset





# Task 1: Data Preprocessing

## Issues Identified

1. The Attached Dataset has the following issues:

* Missing values (unknown, blanks).
* Inconsistent formatting (e.g., US, U.S., united states).
* Duplicates
* Outliers in numeric columns (capital\_gain, hours\_per\_week).
* High-cardinality categorical variables (native\_country, occupation).

A screenshot of a computer

AI-generated content may be incorrect.

A black screen with white text

AI-generated content may be incorrect.A screenshot of a computer screen

AI-generated content may be incorrect.

1. The census dataset Census Income classification – predicting whether income is above/below 50k that’s why income has two values.

* Missing values (unknown, not in universe)
* High Dimensionality

A screenshot of a computer

AI-generated content may be incorrect.

A black screen with white text

AI-generated content may be incorrect.

A screen shot of a computer screen

AI-generated content may be incorrect.

A screen shot of a computer

AI-generated content may be incorrect.

## Solutions Applied

Following solution are applied to the datasets:

Replaced missing values: median (numeric), mode (categorical).

* Standardized text (lowercase, stripped spaces, removed punctuation).
* Normalized country names.
* Outlier handling: IQR capping.
* Encoding strategies:
  + Binary Encoding (2 categories → 0/1).
  + Label Encoding (>3 categories).
  + One-Hot Encoding (small categorical variables).
* Removed duplicates and irrelevant columns (id).

# Task 2: Exploratory Data Analysis (EDA)

## Before Preprocessing

The attached dataset had following features before:

* Many missing values, inconsistent categories.
* Skewed distributions with extreme outliers.
* Correlation heatmap noisy due to unclean data.
* Some categorical columns had mixed formatting (e.g., USA).

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A graph with red and blue squares

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.

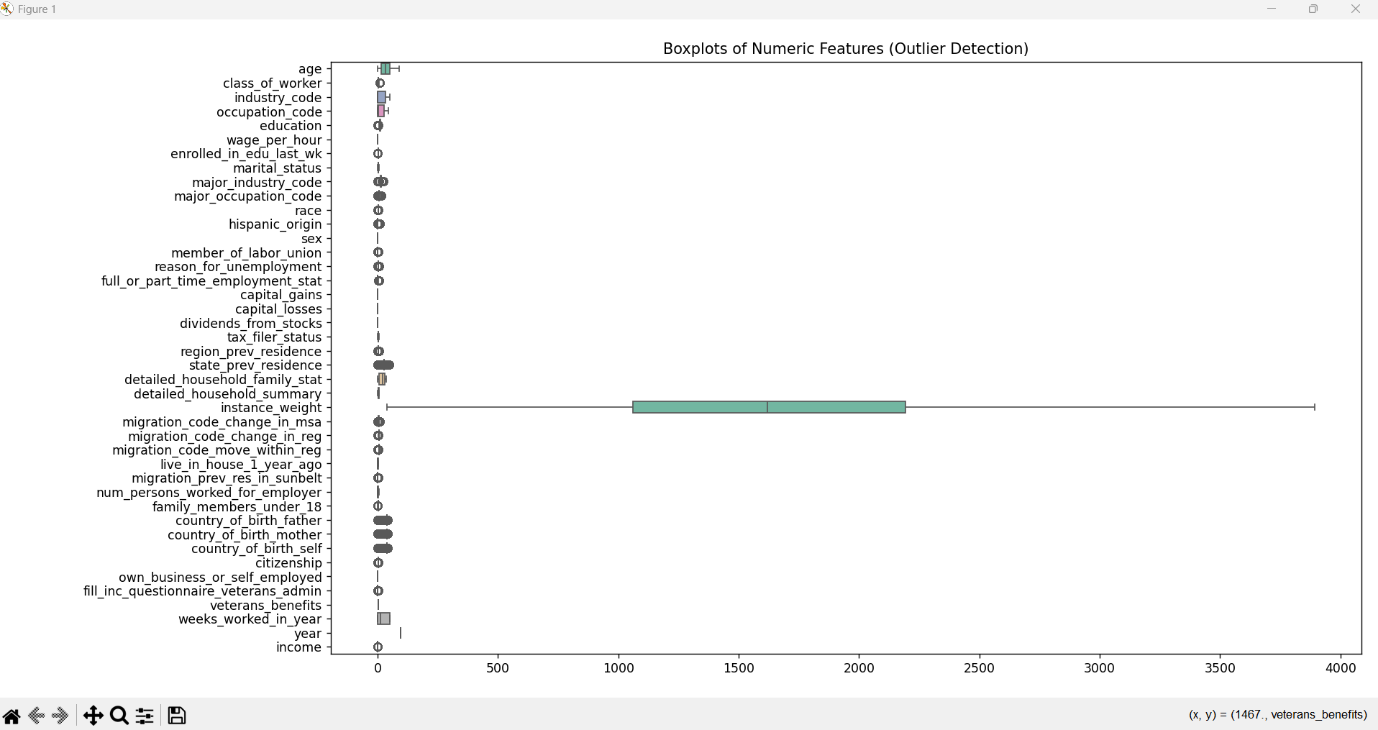
## After Preprocessing

The attached has following properties after cleaning:

* Cleaned dataset i.e. no missing values.
* Standardized categories improved clarity (e.g., only “united\_states”).
* Outliers reduced i.e. boxplots more interpretable.
* Correlation heatmap more reliable.
* Encoding made dataset ML-ready (numeric format).
* Duplicates removed.

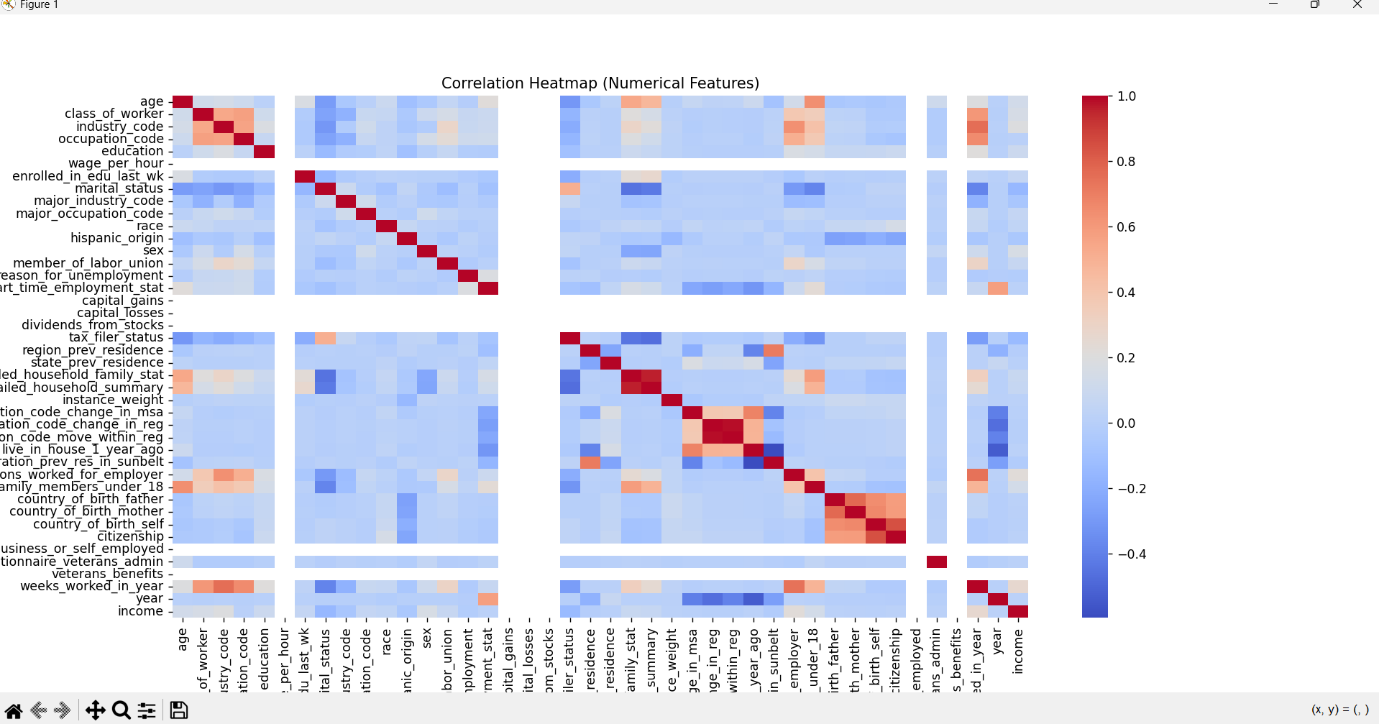
A screenshot of a graph

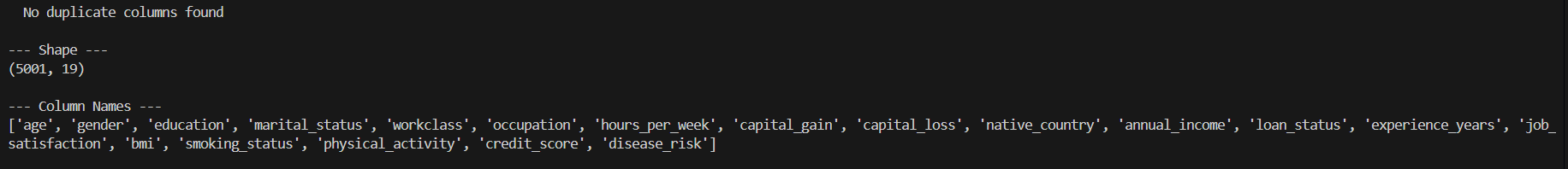
AI-generated content may be incorrect.



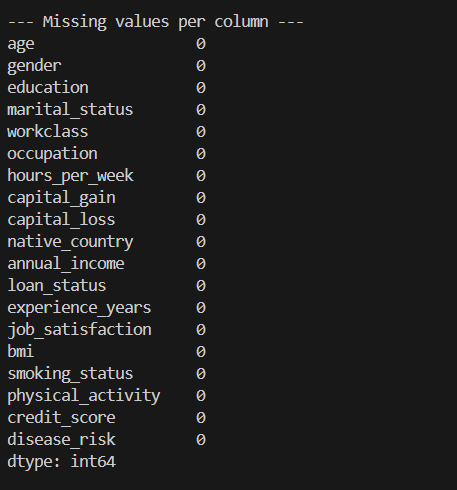
A graph of a graph

AI-generated content may be incorrect.



A screenshot of a computer

AI-generated content may be incorrect.



# Key Insights

* EDA before preprocessing allowed to find out what type values are there in categorical columns that helped to define one format.
* It helped to see duplicated rows and null and non-non null counts before and after Preprocessing.
* It helped to choose best encoding method for categorical columns depending on type of value in them.

# Conclusion

Assignment demonstrated that preprocessing is not just a preparatory step but a core part of the analysis. Clean data ensures Accuracy, Consistency by using standardized categories to prevent misleading fragmentation, efficiency as duplicates and irrelevant features are removed, reducing computational overhead and Interpretability as clean and encoded features make both analysis and modelling easier to explain.

In summary, assignment reinforced that high-quality data is the foundation of effective data science. Without preprocessing, EDA would highlight misleading trends, and any models built later would be biased or inaccurate. By carefully handling missing values, outliers, and categorical encoding, we prepared the dataset for deeper analysis and reliable machine learning tasks.