Министерство образования Республики Беларусь
Учреждение образования
«Брестский Государственный технический университет»
Кафедра ИИТ

Отчёт по лабораторной работе №1

Выполнил:
Студент 3 курса
Группы АС-65
Романюк Д. А.
Проверил:
Крощенко А. А.

Брест 2025

Цель работы: Получить практические навыки работы с данными с использованием библиотек Pandas для манипуляции и Matplotlib для визуализации. Научиться выполнять основные шаги предварительной обработки данных, такие как очистка, нормализация и работа с различными типами признаков.

Вариант 5

Задание1. Загрузите данные и выведите первые 10 строк.

```
import pandas as pd
df = pd.read_csv("adult.data", header=None, sep=",", engine="python")

df.columns = [
    "age", "workclass", "fnlwgt", "education", "education-num",
    "marital-status", "occupation", "relationship", "race", "sex",
    "capital-gain", "capital-loss", "hours-per-week", "native-country",
"income"
]

print(df.head(10))
```

```
   age          workclass  fnlwgt   education  education-num  \
0   39          State-gov   77516   Bachelors             13
1   50   Self-emp-not-inc   83311   Bachelors             13
2   38            Private  215646     HS-grad              9
3   53            Private  234721        11th              7
4   28            Private  338409   Bachelors             13
5   37            Private  284582     Masters             14
6   49            Private  160187         9th              5
7   52   Self-emp-not-inc  209642     HS-grad              9
8   31            Private   45781     Masters             14
9   42            Private  159449   Bachelors             13

          marital-status         occupation    relationship    race  \
0          Never-married       Adm-clerical   Not-in-family   White
1     Married-civ-spouse    Exec-managerial         Husband   White
2               Divorced  Handlers-cleaners   Not-in-family   White
3     Married-civ-spouse  Handlers-cleaners         Husband   Black
4     Married-civ-spouse     Prof-specialty            Wife   Black
5     Married-civ-spouse    Exec-managerial            Wife   White
6   Married-spouse-absent      Other-service   Not-in-family   Black
7     Married-civ-spouse    Exec-managerial         Husband   White
8          Never-married     Prof-specialty   Not-in-family   White
9     Married-civ-spouse    Exec-managerial         Husband   White

       sex  capital-gain  capital-loss  hours-per-week  native-country  income
0     Male          2174             0              40   United-States   <=50K
1     Male             0             0              13   United-States   <=50K
2     Male             0             0              40   United-States   <=50K
3     Male             0             0              40   United-States   <=50K
4   Female             0             0              40            Cuba   <=50K
5   Female             0             0              40   United-States   <=50K
6   Female             0             0              16         Jamaica   <=50K
7     Male             0             0              45   United-States    >50K
8   Female         14084             0              50   United-States    >50K
9     Male          5178             0              40   United-States    >50K
```

Задание 2. Проанализируйте столбец workclass. Найдите и замените значения ? на наиболее часто встречающееся значение в этом столбце.

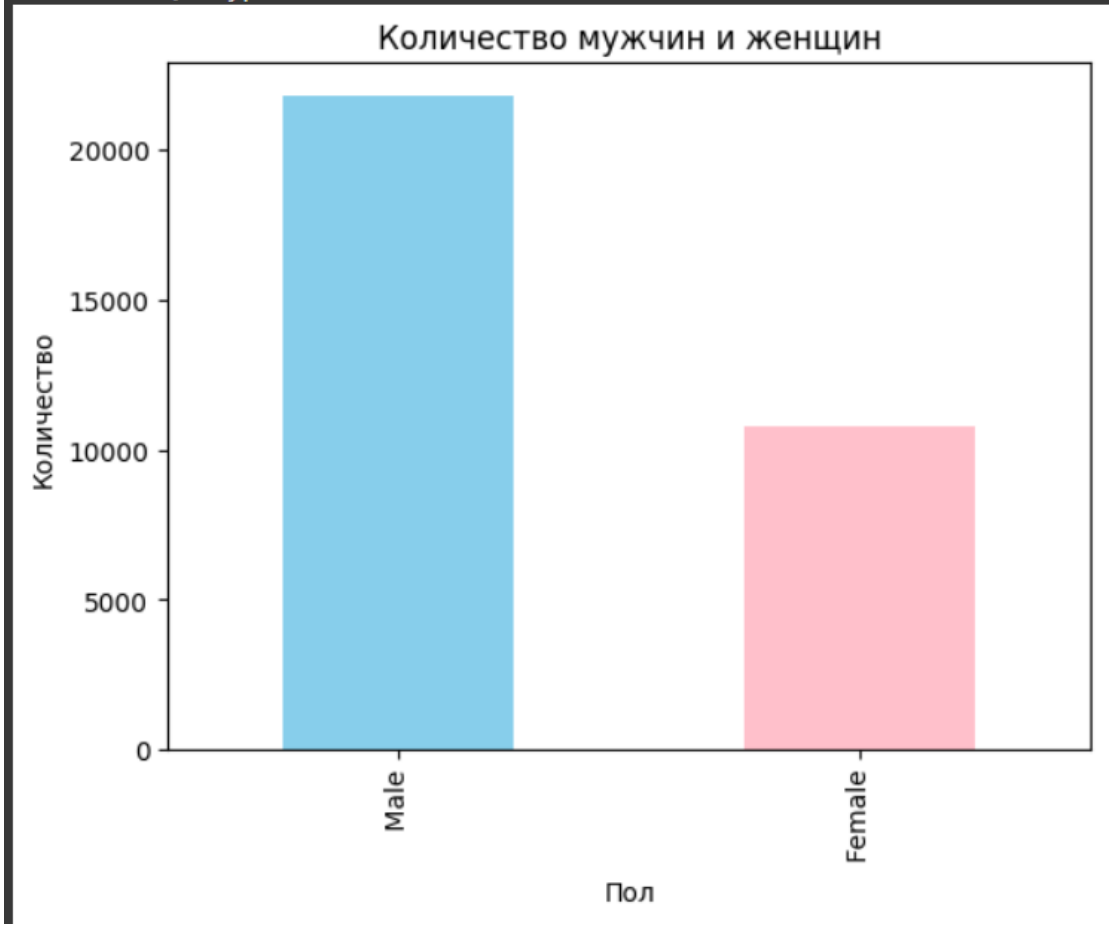
```
print(df['workclass'].value_counts())

most_common = df['workclass'].mode()[0]
print("Наиболее частое значение:", most_common)

df['workclass'] = df['workclass'].replace('?', most_common)

print(df['workclass'].value_counts())
```

```
workclass
Private             22696
Self-emp-not-inc     2541
Local-gov            2093
?                    1836
State-gov            1298
Self-emp-inc         1116
Federal-gov           960
Without-pay            14
Never-worked            7
Name: count, dtype: int64
Наиболее частое значение: Private
workclass
Private             24532
Self-emp-not-inc     2541
Local-gov            2093
State-gov            1298
Self-emp-inc         1116
Federal-gov           960
Without-pay            14
Never-worked            7
Name: count, dtype: int64
```

Задание 3. Определите, сколько в наборе данных мужчин и женщин. Визуализируйте результат.

```
import matplotlib.pyplot as plt

gender_counts = df['sex'].value_counts()
print(gender_counts)

gender_counts.plot(kind='bar', color=['skyblue', 'pink'])
plt.title("Количество мужчин и женщин")
plt.xlabel("Пол")
plt.ylabel("Количество")
plt.show()
```

```
Male      21790
Female    10771
Name: count, dtype: int64
```



Задание 4. Преобразуйте категориальный признак race в числовой формат.

```
df['race_encoded'] = pd.factorize(df['race'])[0]
```

```
print(df[['race', 'race_encoded']].head(10))
```

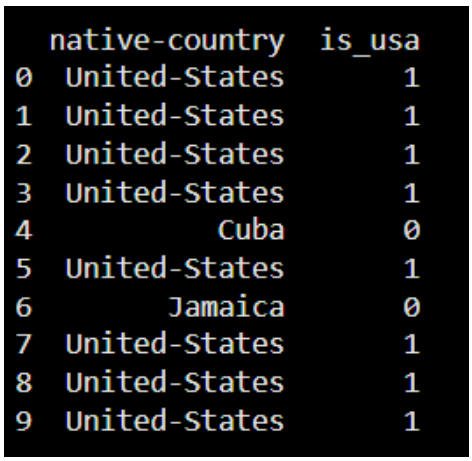
```
     race  race_encoded
0   White             0
1   White             0
2   White             0
3   Black             1
4   Black             1
5   White             0
6   Black             1
7   White             0
8   White             0
9   White             0
```

Задание 5. Постройте гистограмму распределения возраста (age) для двух групп: тех, кто зарабатывает >50K, и тех, кто зарабатывает <=50K.

```
df = df.apply(lambda x: x.str.strip() if x.dtype == "object" else x)
import matplotlib.pyplot as plt

plt.figure(figsize=(8,5))

df[df['income'] == '<=50K']['age'].hist(bins=30, alpha=0.5, label='<=50K')
df[df['income'] == '>50K']['age'].hist(bins=30, alpha=0.5, label='>50K')

plt.legend()
plt.xlabel("Возраст")
plt.ylabel("Количество")
plt.title("Распределение возраста по группам дохода")
plt.show()
```

Задание 6. Создайте новый бинарный признак is_usa на основе столбца native-country.

```
df['is_usa'] = df['native-country'].apply(lambda x: 1 if x == 'United-States' else 0)

print(df[['native-country', 'is_usa']].head(10))
```



	native-country	is_usa
0	United-States	1
1	United-States	1
2	United-States	1
3	United-States	1
4	Cuba	0
5	United-States	1
6	Jamaica	0
7	United-States	1
8	United-States	1
9	United-States	1

Вывод: Я получил практические навыки работы с данными с использованием библиотек Pandas для манипуляции и Matplotlib для визуализации. Научился выполнять основные шаги предварительной обработки данных, такие как очистка, нормализация и работа с различными типами признаков.