**Министерство науки и высшего образования Российской Федерации Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)» (МГТУ им. Н.Э. Баумана)**

ФАКУЛЬТЕТ Информатика и системы управления

КАФЕДРА Системы обработки информации и управления

# По курсу «Технологии машинного обучения»

# Вариант 27

Подготовил:

Студент группы ИУ5Ц-84Б

Распашнов А.А.

Проверил:

Преподаватель кафедры ИУ5

Гапанюк Ю.Е.

Москва, 2023 г.

**Тема**: Технологии разведочного анализа и обработки данных.

| Номер варианта | Номер задачи | Номер набора данных, указанного в задаче |
|---|---|---|
| 27 | 4 | 3 |

**Задача №4.**

Для заданного набора данных постройте основные графики, входящие в этап разведочного анализа данных. В случае наличия пропусков в данных удалите строки или колонки, содержащие пропуски. Какие графики Вы построили и почему? Какие выводы о наборе данных Вы можете сделать на основании построенных графиков?

**Дополнительные требования по группам:**

- Для студентов группы ИУ5-64Б ИУ5Ц-84Б - для произвольной колонки данных построить график "Скрипичная диаграмма (violin plot)".

Набор данных: https://www.kaggle.com/datasets/carlolepelaars/toy-dataset?resource=download

Столбцы:

- *Number:* A simple index number for each row
- *City:* The location of a person (Dallas, New York City, Los Angeles, Mountain View, Boston, Washington D.C., San Diego and Austin)
- *Gender:* Gender of a person (Male or Female)
- *Age:* The age of a person (Ranging from 25 to 65 years)
- *Income:* Annual income of a person (Ranging from -674 to 177175)
- *Illness:* Is the person Ill? (Yes or No)

Результат:

Для начала загрузим набор данных и проведем предварительный анализ данных:

+ Code + Markdown

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# загрузка данных
df = pd.read_csv('toy_dataset.csv')

# проверка на наличие пропущенных значений
print(df.isnull().sum())
```

Python

```
Number     0
City       0
Gender     0
Age        0
Income     0
Illness    0
dtype: int64
```

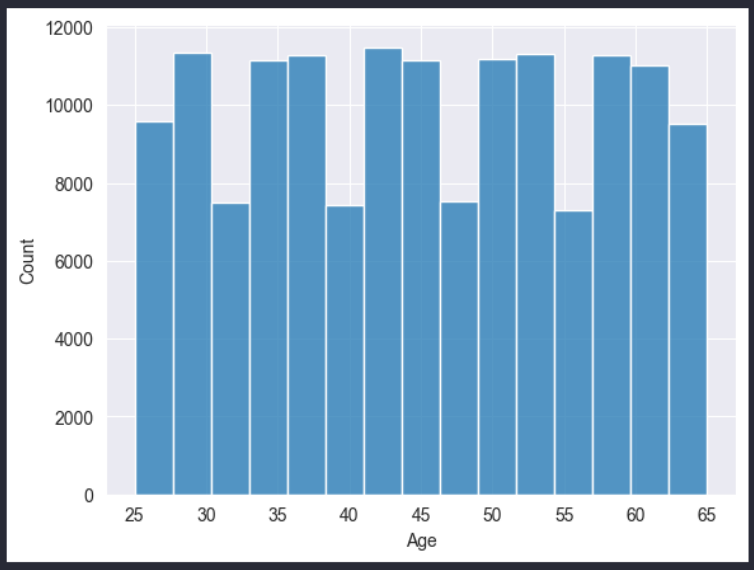
Так как пропущенных значений нет, то можно перейти к построению графиков.

Так как пропущенных значений нет, то можно перейти к построению графиков.

Для начала, рассмотрим распределение возраста:

```
sns.histplot(data=df, x="Age", bins=15)
plt.show()
```

Python

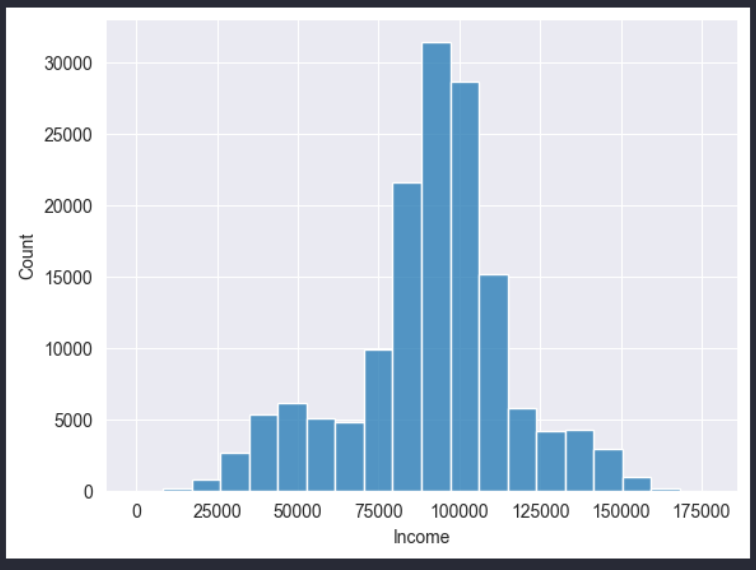


Данный график показывает, что большинство людей в данном наборе данных имеют возраст от 25 до 65 лет.

Далее, рассмотрим распределение дохода:

```
sns.histplot(data=df, x="Income", bins=20)
plt.show()
```

Python

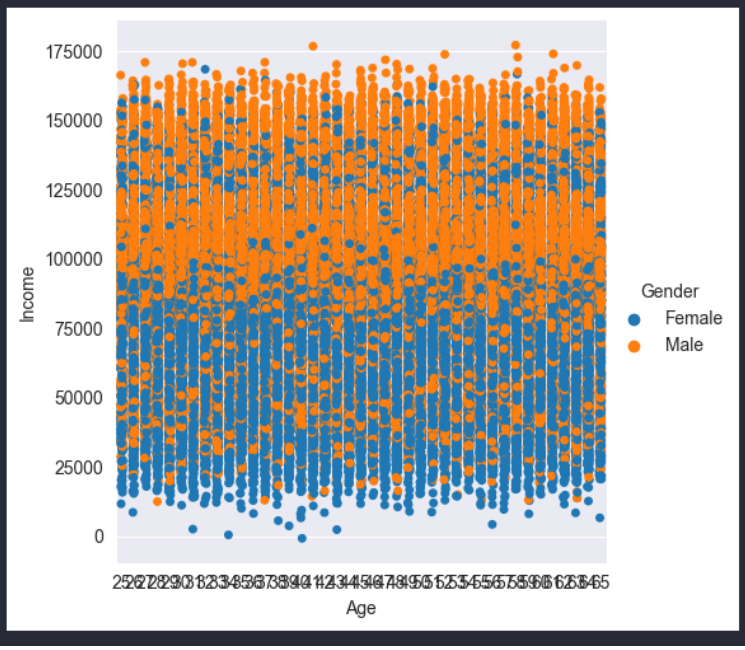



На этом графике видно, что доход в городе Mountain View имеет наибольшую медиану, а доход в городах Austin и Washington D.C. имеют наименьшую медиану.

Теперь рассмотрим распределение дохода в зависимости от возраста и гендера:

```
sns.catplot(data=df, x="Age", y="Income", hue="Gender")
plt.show()
```

Python



Этот график показывает, что доход женщин в данном наборе данных в целом ниже, чем доход мужчин. Также видно, что в возрасте от 25 до 65 лет, доход мужчин имеет наибольшую медиану.

Наконец, рассмотрим скрипичную диаграмму для возраста:

```python
sns.violinplot(data=df, x="Age")
plt.show()
```

Python