Московский государственный технический университет имени Н. Э. Баумана

Факультет «Информатика и системы управления»

Кафедра ИУ5

Отчёт по
Рубежному контролю № 1
по курсу Технологии машинного обучения.

Подготовила:

Калашникова Анастасия

Группа ИУ5-64Б

Москва
2021

Рубежный контроль № 1

Выполнила: Калашникова Анастасия

Группа: ИУ5-64

Вариант: 4

Выполнение работы

Импорт библиотек

In [12]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
```

Загрузка данных

Набор данных 4го варианта - https://www.kaggle.com/carlolepelaars/toy-dataset

In [2]:
```
data = pd.read_csv('toy_dataset.csv', sep=",")
```

In [3]:
```
# Первые 5 строк датасета
data.head()
```

Out[3]:

	Number	City	Gender	Age	Income	Illness
0	1	Dallas	Male	41	40367.0	No
1	2	Dallas	Male	54	45084.0	No
2	3	Dallas	Male	42	52483.0	No
3	4	Dallas	Male	40	40941.0	No
4	5	Dallas	Male	46	50289.0	No

In [4]:
```
# Размер датасета
data.shape
```

Out[4]:
```
(150000, 6)
```

In [5]:
```
# Проверим есть ли пропущенные значения
data.isnull().sum()
```

Out[5]:
```
Number     0
City       0
Gender     0
Age        0
Income     0
Illness    0
dtype: int64
```

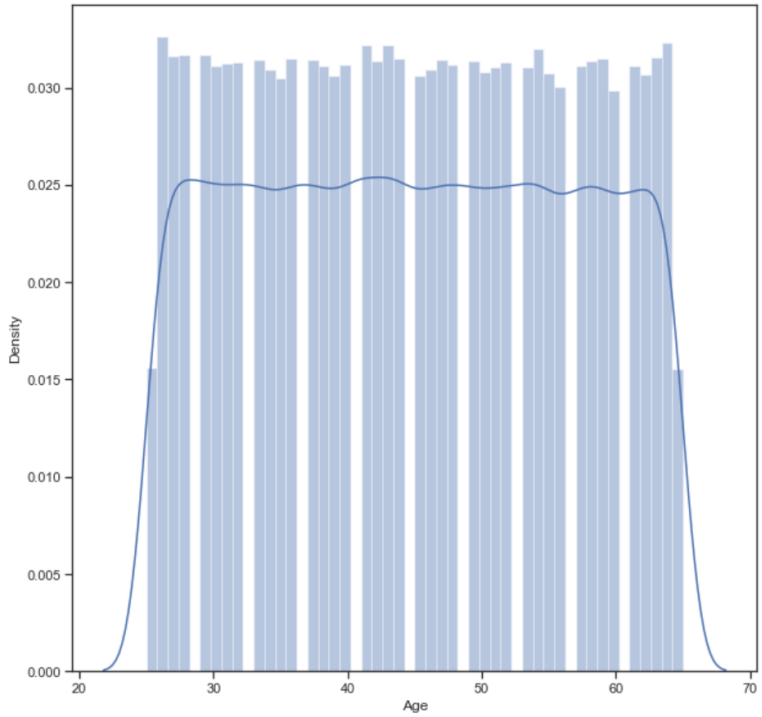
После проверки на пустые значения видно, что нет ни одного пропуска в данных. Поэтому удалять ни строки, ни столбцы не нужно

Построим некоторые диаграммы для набора данных

In [6]:
```
fig, ax = plt.subplots(figsize=(10,10))
sns.distplot(data['Age'])
```

```
/Users/kalashnikova/opt/anaconda3/lib/python3.8/site-packages/seaborn/distributions.py:2551: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

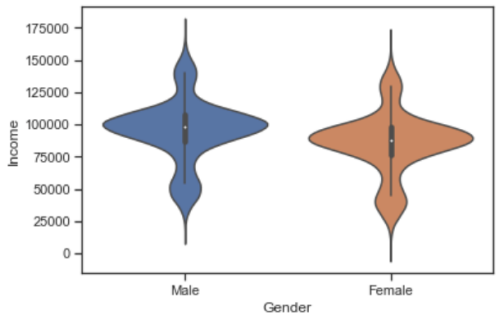
Out[6]:
```
<AxesSubplot:xlabel='Age', ylabel='Density'>
```



Из гистограммы видно, что в датасете собраны возраста от 25 до 65.

In [8]:
```
# Распределение параметра дохода по половому признаку с помощью violin plot
sns.violinplot(x='Gender', y='Income', data=data)
```

Out[8]: <AxesSubplot:xlabel='Gender', ylabel='Income'>



In [9]:
```
# Корреляционная матрица
data.corr()
```

Out[9]:

	Number	Age	Income
Number	1.000000	-0.003448	0.410460
Age	-0.003448	1.000000	-0.001318
Income	0.410460	-0.001318	1.000000

In [11]:
```
sns.heatmap(data.corr(), annot = True, cmap="YlGnBu", fmt='.3f', linewidths=.8)
```

Out[11]: <AxesSubplot:>

Выводы на основе корреляционной матрицы

- Параметр Income слабо коррелирует с параметром Number (0.41)
- Параметр Age отрицательно коррелирован с параметрами Number (-0.003) и Income (-0.001). Значит, когда он возрастает, они убывают и наоборот.
- Так как все числовые признаки слабо или очень слабо коррелируют, то их все нужно исключить из модели, потому что они могут ухудшить ее качество. Значит, не получится построить модель машинного обучения.