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

Отчёт к лабораторным работам по курсу «Методы машинного обучения»

Рубежный контроль №1

Выполнил:
студент(ка) группы ИУ5И-23М Ян Цзиньцзы
подпись, дата

Проверил:
к.т.н., доц., Гапанюк Ю.Е.
подпись, дата

Москва, 2023 г.

Номер варианта=21
Номер задания1=5
Номер задания2=23

1.1.1. Задача №5.

Для набора данных проведите кодирование одного (произвольного) категориального признака с использованием метода "one-hot encoding".

Code:

```
import pandas as pd

# Load the insurance.csv data into a DataFrame called "data"
data = pd.read_csv("D:\cem2\MachineLearning\lab1\insurance.csv")

# Use pd.get_dummies() to perform one-hot encoding on the "region" feature
one_hot_encoded = pd.get_dummies(data, columns=["region"])

# Print the resulting DataFrame to the console
print(one_hot_encoded)
```

Result:

```
D:\App\Python3\python.exe D:\App\Python3\Pyproject\PK1.py
      age     sex     bmi  ...  region_northwest region_southeast  region_southwest
0      19  female  27.900  ...                 0                0                 1
1      18    male  33.770  ...                 0                1                 0
2      28    male  33.000  ...                 0                1                 0
3      33    male  22.705  ...                 1                0                 0
4      32    male  28.880  ...                 1                0                 0
...   ...     ...     ...  ...               ...              ...               ...
1333   50    male  30.970  ...                 1                0                 0
1334   18  female  31.920  ...                 0                0                 0
1335   18  female  36.850  ...                 0                1                 0
1336   21  female  25.800  ...                 0                0                 1
1337   61  female  29.070  ...                 1                0                 0

[1338 rows x 10 columns]
```

1.1.2. Задача №23.

Для набора данных для одного (произвольного) числового признака проведите обнаружение и удаление выбросов на основе правила трех сигм.

Code:

```
import pandas as pd
import numpy as np

# Load the insurance.csv data into a DataFrame called "data"
data = pd.read_csv("D:\cem2\MachineLearning\lab1\insurance.csv")

# Calculate the mean and standard deviation of the "bmi" feature
mean_bmi = np.mean(data["bmi"])
std_bmi = np.std(data["bmi"])
```

```python
# Define the lower and upper bounds for outlier detection
lower_bound = mean_bmi - 3 * std_bmi
upper_bound = mean_bmi + 3 * std_bmi

# Identify outliers in the "bmi" feature
outliers = data[(data["bmi"] < lower_bound) | (data["bmi"] > upper_bound)]

# Remove outliers from the dataset
clean_data = data[(data["bmi"] >= lower_bound) & (data["bmi"] <= upper_bound)]

# Print the number of outliers and the cleaned dataset to the console
print("Number of outliers:", len(outliers))
print("Cleaned dataset:\\n", clean_data)
```

Result:

```
D:\App\Python3\python.exe D:\App\Python3\Pyproject\pk1.1.py
Number of outliers: 4
Cleaned dataset:\n       age     sex     bmi  children smoker     region      charges
0      19  female  27.900         0    yes  southwest  16884.92400
1      18    male  33.770         1     no  southeast   1725.55230
2      28    male  33.000         3     no  southeast   4449.46200
3      33    male  22.705         0     no  northwest  21984.47061
4      32    male  28.880         0     no  northwest   3866.85520
...   ...     ...     ...       ...    ...        ...          ...
1333   50    male  30.970         3     no  northwest  10600.54830
1334   18  female  31.920         0     no  northeast   2205.98080
1335   18  female  36.850         0     no  southeast   1629.83350
1336   21  female  25.800         0     no  southwest   2007.94500
1337   61  female  29.070         0    yes  northwest  29141.36030

[1334 rows x 7 columns]
```

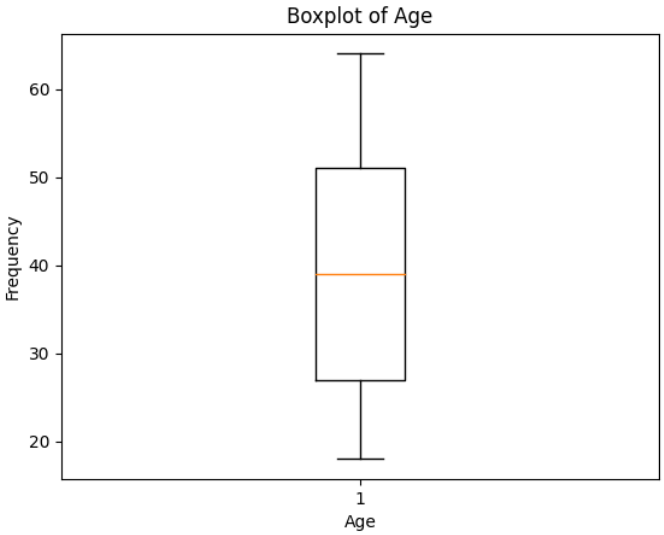
Дополнительные требования

для произвольной колонки данных построить график "Ящик с усами (boxplot)".

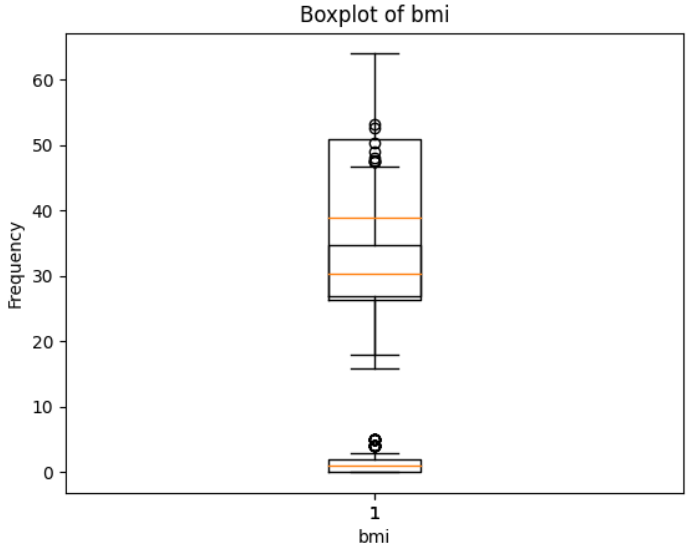
Code:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# Load the insurance.csv data into a DataFrame called "data"
data = pd.read_csv("D:\cem2\MachineLearning\lab1\insurance.csv")

# Create a boxplot for the "age" column
plt.boxplot(data["age"])

# Add a title and axis labels to the plot
plt.title("Boxplot of Age")
plt.xlabel("Age")
plt.ylabel("Frequency")

# Create a boxplot with a mustache for the "children" column
plt.boxplot(data["children"], whis=[5, 95])

# Add a title and axis labels to the plot
plt.title("Box with a Mustache of Children")
plt.xlabel("Children")
plt.ylabel("Frequency")

# Create a boxplot for the "age" column
plt.boxplot(data["bmi"])

# Add a title and axis labels to the plot
plt.title("Boxplot of bmi")
plt.xlabel("bmi")
plt.ylabel("Frequency")

# Display the plot
plt.show()
```

Result:


Boxplot of Age
Frequency
60
50
40
30
20
1
Age


Box with a Mustache of Children
Frequency
60
50
40
30
20
10
0
1
Children


Boxplot of bmi
Frequency
60
50
40
30
20
10
0
1
bmi