

# Московский Государственный Технический Университет имени Н.Э.Баумана Факультет Информатика и системы управления

# Кафедра ИУ-5 «Системы обработки информации и управления»

## Отчёт по рубежному контролю № 1

#### По дисципление

«Методы Машинного Обучения»

Группа ИУ5И-22М

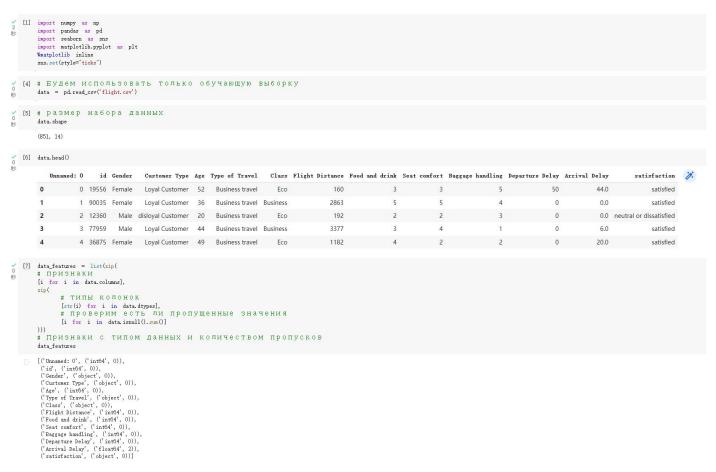
Лю Чжинань

## Номер варианта: 21

#### Номер задачи №1: 5

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

#### Загрузить данные



Метод "one-hot encoding"



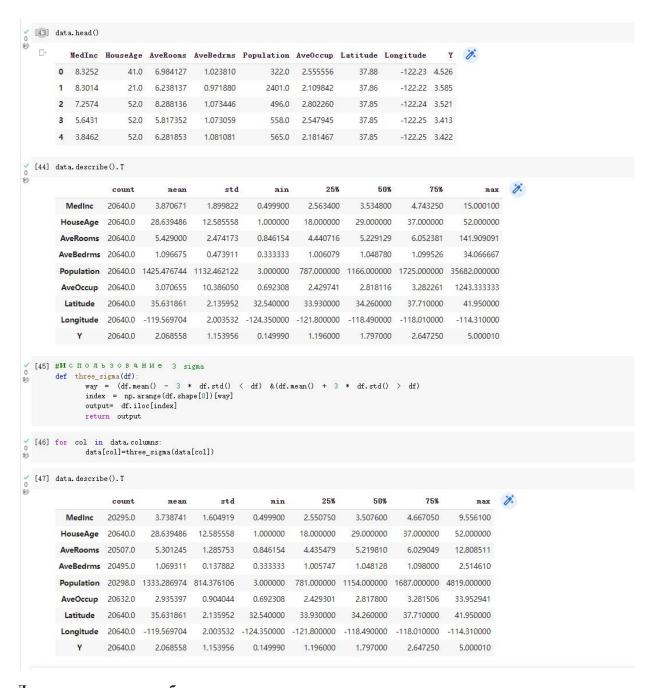
#### Номер задачи №2: 23

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

#### Загрузить данные

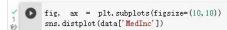
```
import numpy as np
      import pandas as pd
       import seaborn as sns
      import matplotlib.pyplot as plt
      from sklearn.datasets import fetch_california_housing
      import scipy.stats as stats
      from sklearn.svm import SVR
      from sklearn.linear_model import LinearRegression
      from sklearn.neighbors import KNeighborsRegressor
       from sklearn.tree import DecisionTreeRegressor
      from sklearn.ensemble import RandomForestRegressor
      from sklearn.ensemble import GradientBoostingRegressor
       from sklearn.metrics import mean_squared_error
      from sklearn.model_selection import train_test_split
      from IPython.display import Image
      %matplotlib inline
      sns.set(style="ticks")
 [25] x_col_list = ['AveRooms', 'HouseAge', 'MedInc']
 [26] housing = fetch_california_housing()
      data = pd.DataFrame(housing.data,
                                              columns=housing.feature_names)
       data['Y'] = housing.target
       data. shape
       (20640, 9)
[27] def diagnostic_plots(df, variable, title):
             fig, ax = plt.subplots(figsize=(10,7))
              # гистограмма
             plt.subplot(2, 2, 1)
             df[variable].hist(bins=30)
             ## Q-Q plot
             plt.subplot(2, 2, 2)
             stats.probplot(df[variable], dist="norm", plot=plt)
              # ящик с усами
             plt.subplot(2, 2, 3)
              sns.violinplot(x=df[variable])
              # ящик с усами
             plt.subplot(2, 2, 4)
              sns.boxplot(x=df[variable])
             fig.suptitle(title)
              plt.show()
```

Использование правила трех сигм



#### Дополнительные требования по группам:

Для студентов групп ИУ5-22M, ИУ5И-22M - для произвольной колонки данных построить гистограмму.



C+ <ipython-input-21-46ea8e7e1a1b>: 2: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either 'displot' (a figure-level function with similar flexibility) or 'histplot' (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see <a href="https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751">https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751</a>

sns.distplot(data['MedInc'])
<Axes: xlabel='MedInc', ylabel='Density'>

