|  |  |
| --- | --- |
| Gerb-BMSTU_01 | 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 (ИУ) |

|  |  |
| --- | --- |
| КАФЕДРА | Система обработки информации и управления |

|  |  |
| --- | --- |
| дисциплина | Методы машинного обучения |

|  |
| --- |
| отчет по лабораторной работе № 2 |

|  |
| --- |
| Обработка признаков (часть 1) |
| *название лабораторной работы* |

|  |  |
| --- | --- |
| Группа | ИУ5-14М |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Студент | 28.03.2022 |  |  |  | Молева А. А. |
|  | *дата выполнения работы* |  | *подпись* |  | *фамилия, и.о.* |

|  |  |  |  |
| --- | --- | --- | --- |
| Преподаватель |  |  | Гапанюк Ю. Е. |
|  | *подпись* |  | *фамилия, и.о.* |

Москва, 2022 г.

# **Цель работы**

# Цель лабораторной работы: изучение продвинутых способов предварительной обработки данных для дальнейшего формирования моделей.

# **Задание**

1. Выбрать набор данных (датасет), содержащий категориальные и числовые признаки и пропуски в данных. Для выполнения следующих пунктов можно использовать несколько различных наборов данных (один для обработки пропусков, другой для категориальных признаков и т.д.) Просьба не использовать датасет, на котором данная задача решалась в лекции.
2. Для выбранного датасета (датасетов) на основе материалов лекций решить следующие задачи:
   1. устранение пропусков в данных;
   2. кодирование категориальных признаков;
   3. нормализацию числовых признаков.

# **Текст программы**

import pandas as pd

df = pd.read\_csv('kamyr-digester.csv')

df.info()

#Удаление пустых строк

df1 = df.dropna(axis=0)

df1.isnull().sum()

#Simpleimputer

from numpy import nan

from numpy import isnan

from pandas import read\_csv

from sklearn.impute import SimpleImputer

imputer = SimpleImputer(missing\_values=nan, strategy='mean')

dfSimpleImputer = pd.DataFrame(imputer.fit\_transform(df.iloc[:, 1:]), columns=df.columns[1:])

dfSimpleImputer['Observation'] = df['Observation']

dfSimpleImputer

dfSimpleImputer.isnull().sum()

#KNNImputer

from sklearn.impute import KNNImputer

knnimputer = KNNImputer(

    n\_neighbors=5,

    weights='distance',

    metric='nan\_euclidean',

    add\_indicator=False,

)

knnimpute\_hdata\_imputed\_temp = knnimputer.fit\_transform(df.iloc[:, 1:])

knnimpute\_hdata\_imputed = pd.DataFrame(knnimpute\_hdata\_imputed\_temp, columns=df.columns[1:])

knnimpute\_hdata\_imputed.head()

knnimpute\_hdata\_imputed

knnimpute\_hdata\_imputed.isnull().sum()

#Категориальные признаки

! kaggle competitions download -c titanic

dfCat = pd.read\_csv('train.csv')

dfCat

dfCat.info()

from sklearn.preprocessing import OneHotEncoder

#Разбиение на две колонки

pd.get\_dummies(dfCat[['Sex']]).head()

#Кодирование массива

from sklearn.preprocessing import LabelEncoder

le = LabelEncoder()

cat\_enc\_le = le.fit\_transform(dfCat['Name'])

dfCat['Name'].unique()

import numpy as np

np.unique(cat\_enc\_le)

le.inverse\_transform([0, 1, 2, 3])

#Нормализация

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import scipy.stats as stats

def diagnostic\_plots(df, variable):

    plt.figure(figsize=(15,6))

    # гистограмма

    plt.subplot(1, 2, 1)

    df[variable].hist(bins=30)

    ## Q-Q plot

    plt.subplot(1, 2, 2)

    stats.probplot(df[variable], dist="norm", plot=plt)

    plt.show()

#Логарифмическое преобразование

df['ChipRate\_log'] = np.log(df['ChipRate'])

diagnostic\_plots(df, 'ChipRate\_log')

#Обратное преобразование

df['ChipRate\_reciprocal'] = 1 / (df['ChipRate'])

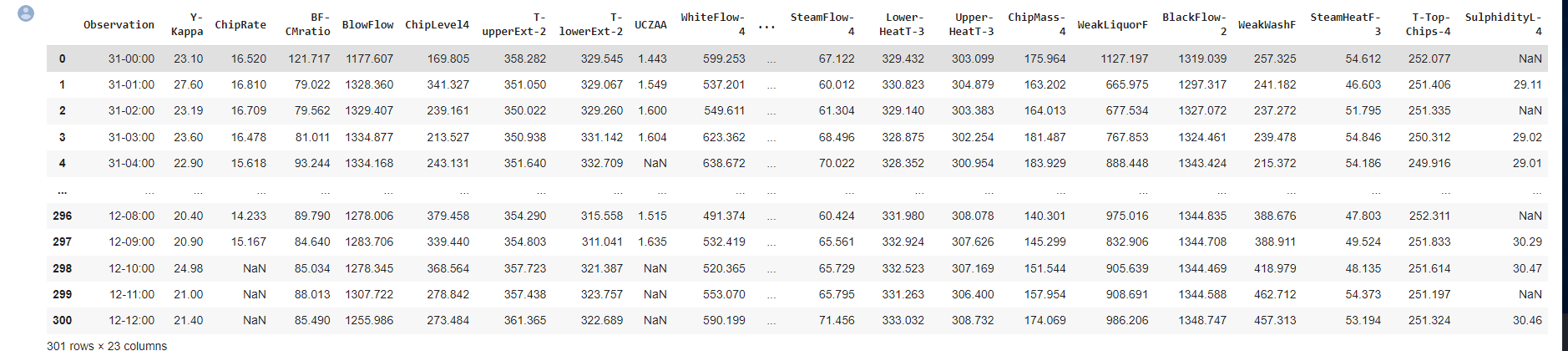
diagnostic\_plots(df, 'ChipRate\_reciprocal')

#Квадратный корень

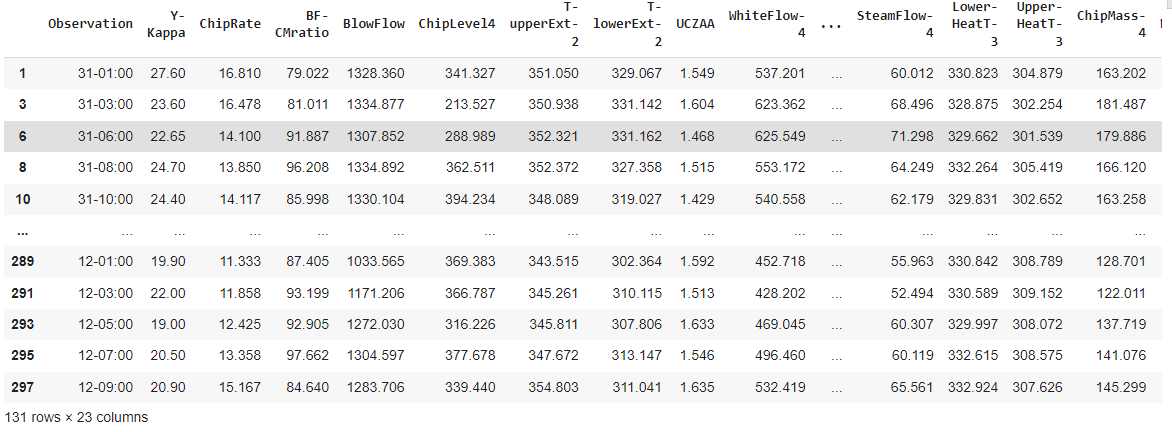
df['ChipRate\_sqr'] = df['ChipRate']\*\*(1/2)

diagnostic\_plots(df, 'ChipRate\_sqr')

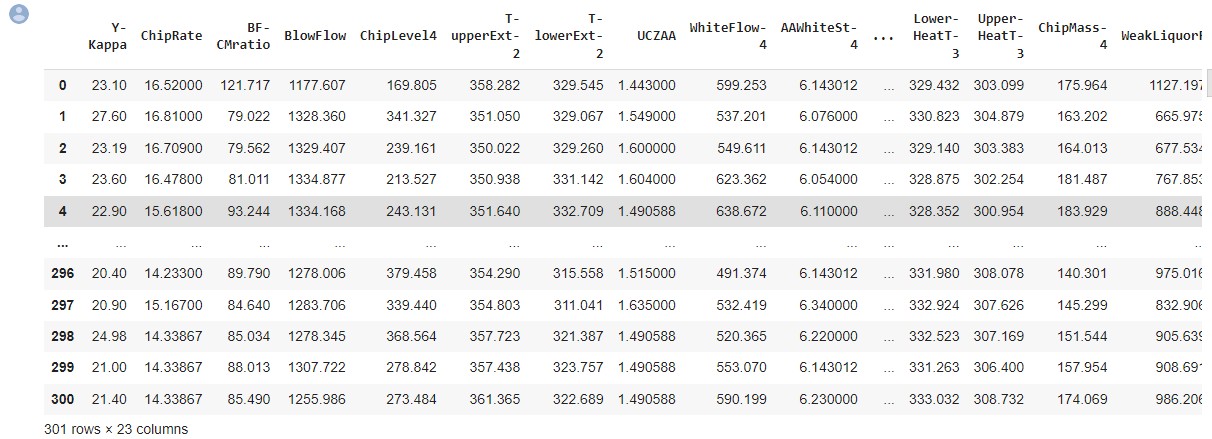
**Экранные формы**

****

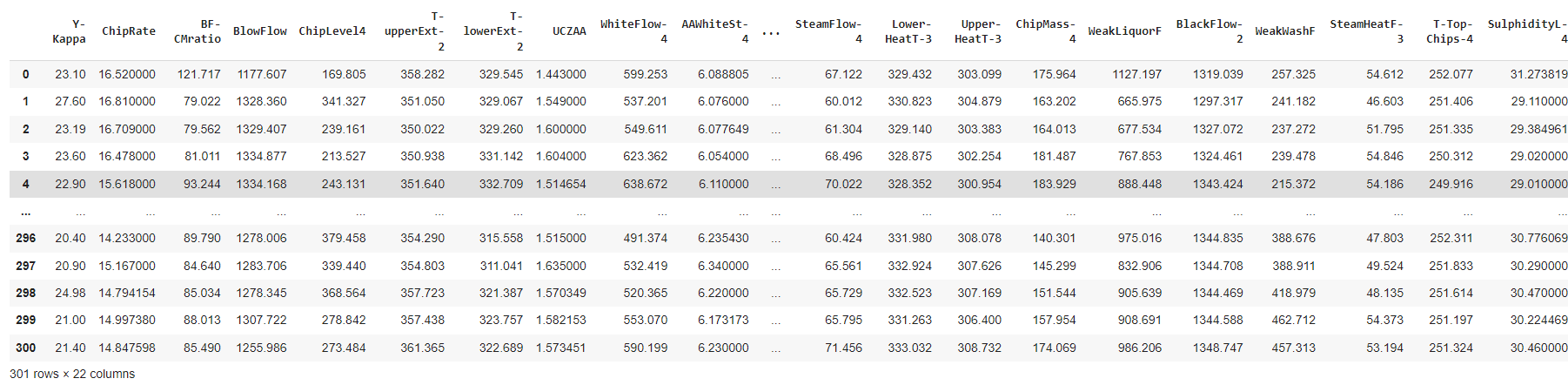
**Рисунок 1 – Датасет 1**

****

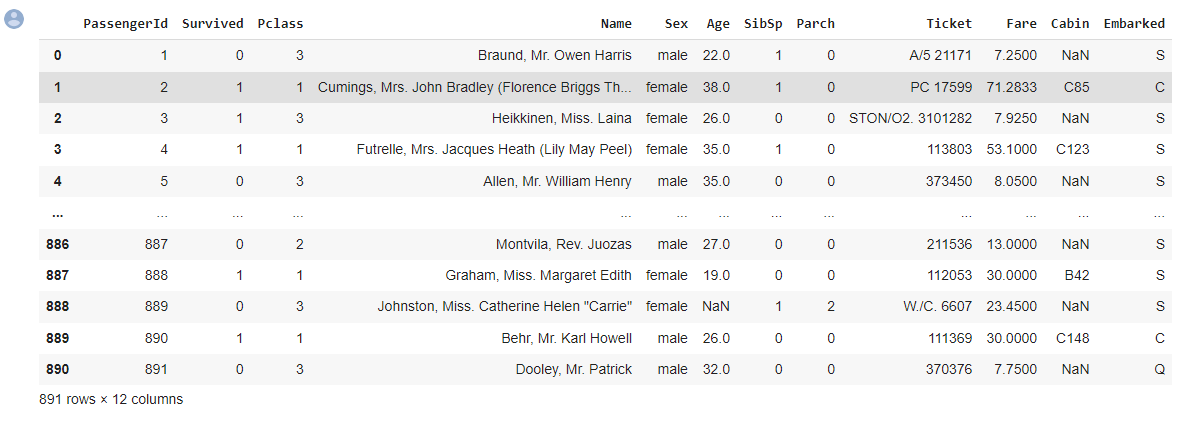
**Рисунок 2 – Удаление пустых строк**

****

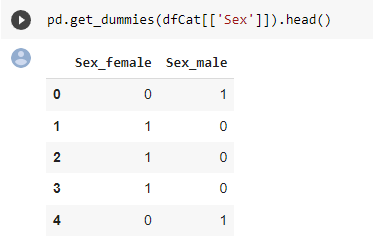
**Рисунок 3 – Замена средним значением**

****

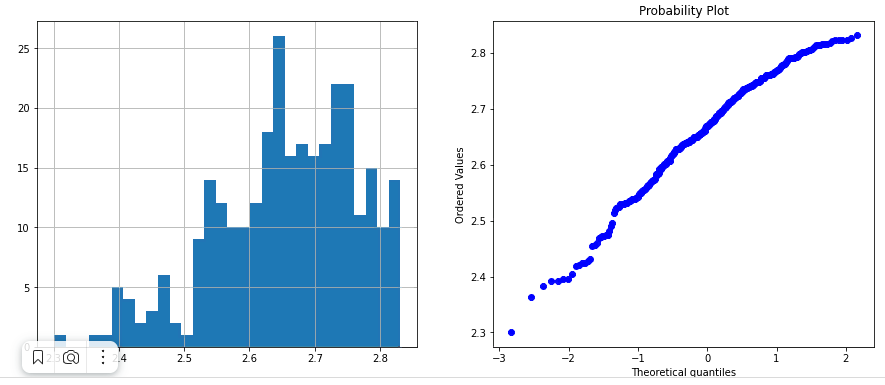
**Рисунок 4 – Заполнение по ближайшим соседям**



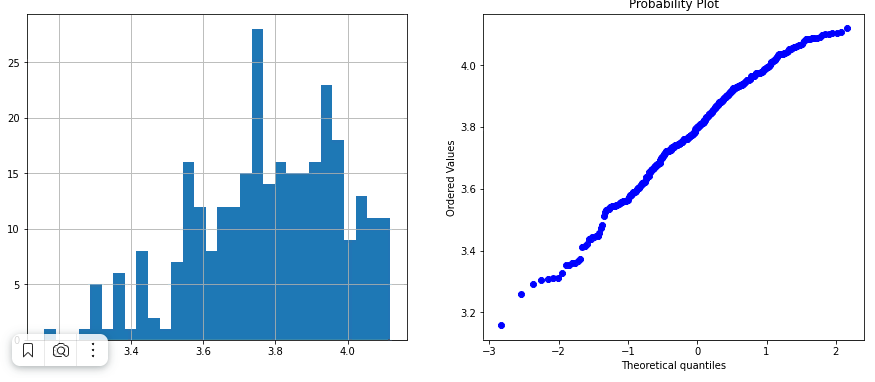
**Рисунок 5 – Датасет 2**

****

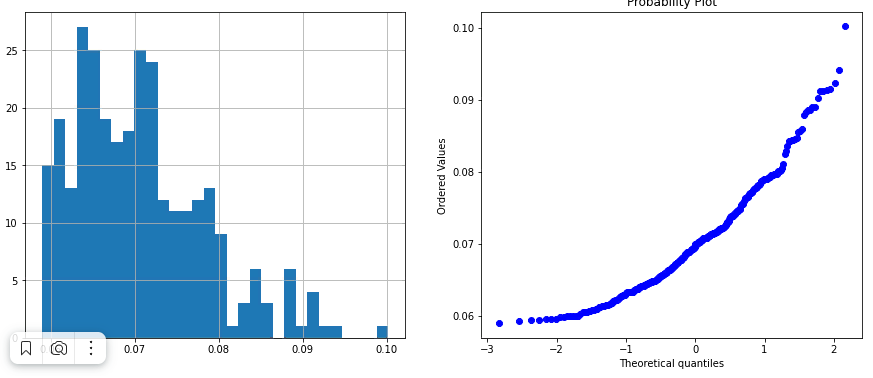
**Рисунок 6 – Разделение на две колонки**

****

**Рисунок 7 – Логарифмическое преобразование**

****

**Рисунок 8 – Обратное преобразование**

****

**Рисунок 9 – Квадратный корень**

**Выводы**

В результате проделанной работы были решены следующие задачи: устранение пропусков в данных; кодирование категориальных признаков; нормализация числовых признаков.