|  |  |
| --- | --- |
|  | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное автономное 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

**ОТЧЕТ**

***ПО РУБЕЖНЫЙ КОНТРОЛЬ №1***

***ПО ДИСЦ******ИПЛИНЕ***

***«МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ»***

***ВАРИАНТ 16***

Студент \_\_ИУ5И-22М\_\_ **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_****Сюй Хаоюй\_**

(Группа) (Подпись, дата) (И.О.Фамилия)

Преподаватель **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_Ю.Е.Гапанюк\_\_**

(Подпись, дата) (И.О.Фамилия)

2025 г.

**ВВЕДЕНИЕ**

Для студентов групп ИУ5-21М, ИУ5-22М, ИУ5-23М, ИУ5-24М, ИУ5-25М номер варианта = номер в списке группы.

Для студентов групп ИУ5И-21М, ИУ5И-22М, ИУ5И-23М, ИУ5И-24М, ИУ5И-25М номер варианта = 15 + номер в списке группы.

Для студентов групп ИУ5-25МВ номер варианта = 20 + номер в списке группы.

Дополнительные требования по группам:

* Для студентов групп ИУ5-21М, ИУ5И-21М - для пары произвольных колонок данных построить график "Диаграмма рассеяния".
* Для студентов групп ИУ5-22М, ИУ5И-22М - для произвольной колонки данных построить гистограмму.
* Для студентов групп ИУ5-23М, ИУ5И-23М - для произвольной колонки данных построить график "Ящик с усами (boxplot)".
* Для студентов группы ИУ5-24М, ИУ5И-24М - для произвольной колонки данных построить график "Скрипичная диаграмма (violin plot)".
* Для студентов группы ИУ5-25М, ИУ5И-25М, ИУ5-25МВ - для произвольной колонки данных построить парные диаграммы (pairplot).

Каждая задача предполагает использование набора данных. Набор данных выбирается Вами произвольно с учетом следующих условий:

* Вы можете использовать один набор данных для решения всех задач, или решать каждую задачу на своем наборе данных.
* Набор данных должен отличаться от набора данных, который использовался в лекции для решения рассматриваемой задачи.
* Вы можете выбрать произвольный набор данных (например тот, который Вы использовали в лабораторных работах) или создать собственный набор данных (что актуально для некоторых задач, например, для задач удаления псевдоконстантных или повторяющихся признаков).
* Выбранный или созданный Вами набор данных должен удовлетворять условиям поставленной задачи. Например, если решается задача устранения пропусков, то набор данных должен содержать пропуски.

Полученные варианты:

* Номер варианта = 15 + 1 =16
* Номер задачи №1: 16

Задача №16 - Для набора данных проведите нормализацию для одного (произвольного) числового признака с использованием преобразования Бокса-Кокса (Box-Cox transformation).

* Номер задачи №2: 36

Задача №36 - Для набора данных проведите процедуру отбора признаков (feature selection). Используйте класс SelectKBest для 5 лучших признаков, и метод, основанный на взаимной информации.

Дополнительные требования по группам:

* Для студентов групп ИУ5-22М, ИУ5И-22М - Для произвольной колонки данных построить гистограмму.

**ХОД ВЫПОЛНЕНИЯ РАБОТЫ**

# Часть 1. Текстовое описание набора данных

Набор данных № 1: bmw.csv

Этот набор данных собирает цены на подержанные автомобили BMW в Великобритании и используется для анализа влияния различных факторов на цены на подержанные автомобили.

Набор данных содержит информацию о цене, трансмиссии, пробеге, типе топлива, дорожном налоге, расходе миль на галлон (mpg) и объеме двигателя.

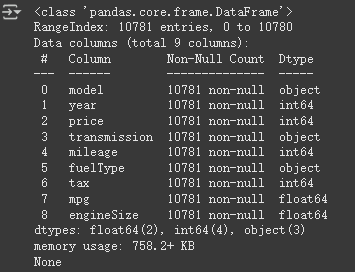


Рисунок 1: Информация о наборе данных (bmw.csv)

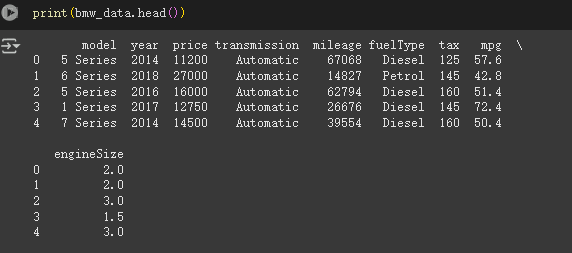


Рисунок 2: Первые 5 строк набора данных (bmw.csv)

Набор данных № 2: Car\_Features.csv

Набор данных об автомобилях с такими характеристиками, как марка, модель, год выпуска, двигатель и другие свойства автомобиля, используемые для прогнозирования его цены.

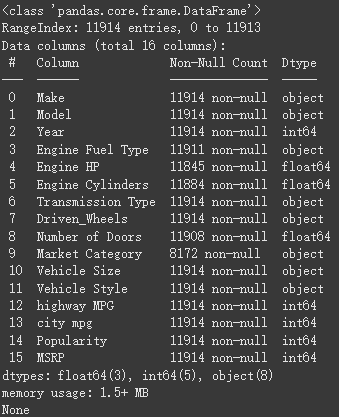


Рисунок 3: Информация о наборе данных (Car\_Features.csv)

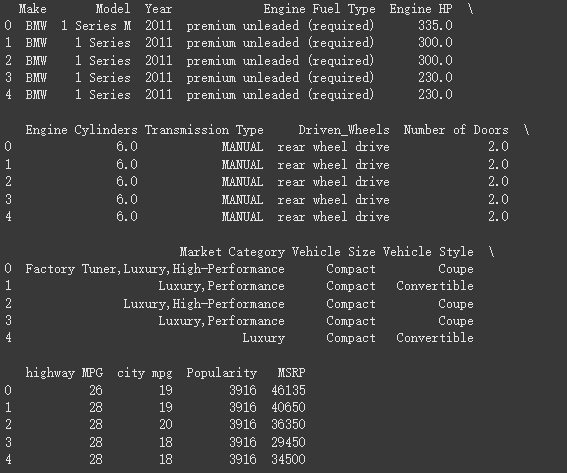


Рисунок 4: Первые 5 строк набора данных (Car\_Features.csv)

# Часть 2. Задача №16

Задача №16 - Для набора данных проведите нормализацию для одного (произвольного) числового признака с использованием преобразования Бокса-Кокса (Box-Cox transformation).

Используя набор данных № 1: bmw.csv

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import scipy.stats as stats

# 加载数据集

data = pd.read\_csv('bmw.csv')

# 定义一个函数来绘制诊断图（直方图和Q-Q图）

def diagnostic\_plots(df, variable):

    plt.figure(figsize=(15, 6))

    # 直方图

    plt.subplot(1, 2, 1)

    df[variable].hist(bins=30, edgecolor='black', alpha=0.7)

    plt.title(f'Histogram of {variable}')

    # Q-Q图

    plt.subplot(1, 2, 2)

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

    plt.title(f'Q-Q Plot of {variable}')

    plt.show()

# 对原始的price列进行诊断

diagnostic\_plots(data, 'price')

# 应用Box-Cox变换

data['price\_boxcox'], param = stats.boxcox(data['price'])

print(f'Optimal λ value for Box-Cox transformation: {param}')

# 对变换后的price\_boxcox列进行诊断

diagnostic\_plots(data, 'price\_boxcox')

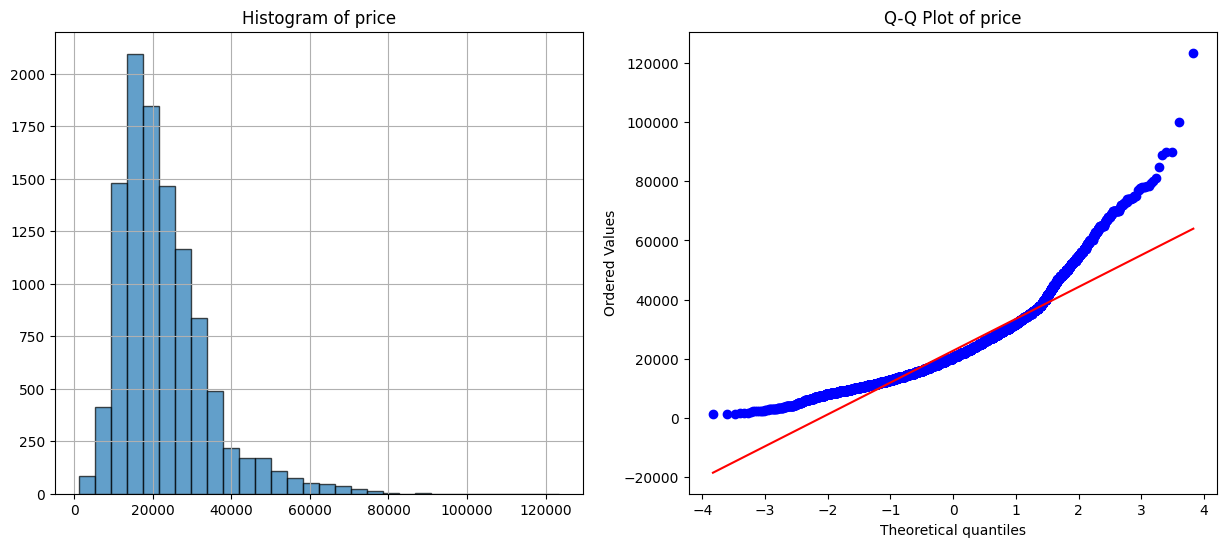


Рисунок 5: Гистограмма и график Q-Q перед преобразованием данных

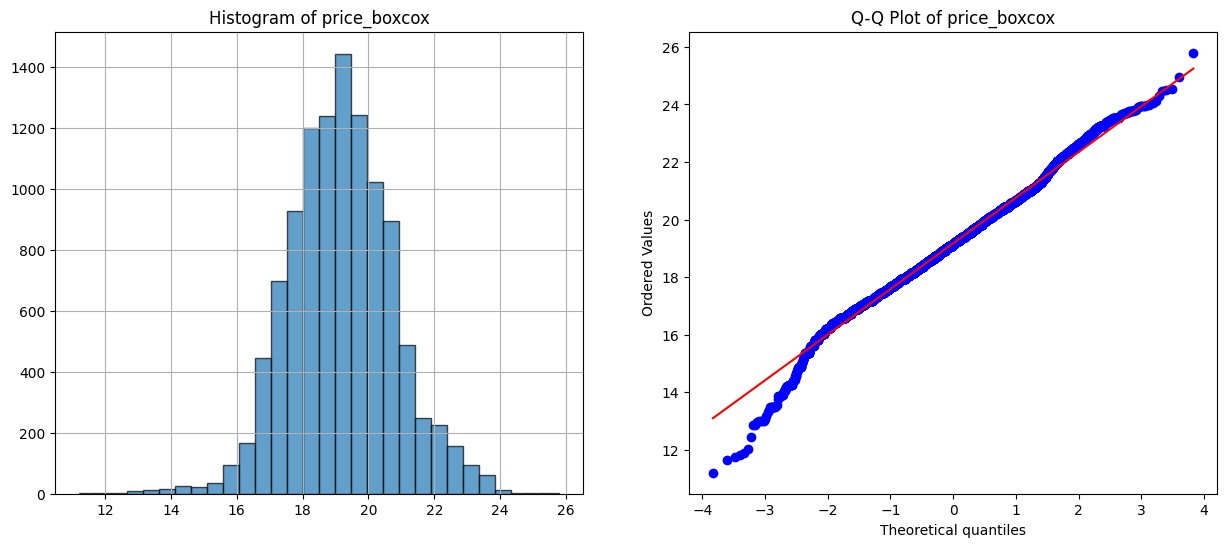


Рисунок 6: Гистограмма и график Q-Q после преобразования данных

# Часть 3. Задача №36

Задача №36 - Для набора данных проведите процедуру отбора признаков (feature selection). Используйте класс SelectKBest для 5 лучших признаков, и метод, основанный на взаимной информации.

Используя набор данных № 2: Car\_Features.csv

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from sklearn.feature\_selection import SelectKBest, mutual\_info\_regression, f\_regression

from sklearn.impute import SimpleImputer

# 加载数据集

Car\_Features\_data = pd.read\_csv('Car\_Features.csv')

df = pd.read\_csv('Car\_Features.csv')

# 查看数据集的结构

print(df.info())

# 检查缺失值

print(df.isnull().sum())

# 填充缺失值

imputer = SimpleImputer(strategy='median')

numeric\_columns = df.select\_dtypes(include=[np.number]).columns.tolist()

df[numeric\_columns] = imputer.fit\_transform(df[numeric\_columns])

# 假设我们要预测的目标变量是 'MSRP'，其他数值型列为特征

X = df[numeric\_columns].drop('MSRP', axis=1)

y = df['MSRP']

# 使用互信息方法选择5个最佳特征

selector\_mutual\_info = SelectKBest(score\_func=mutual\_info\_regression, k=5)

X\_new\_mutual\_info = selector\_mutual\_info.fit\_transform(X, y)

# 获取选中的特征名称

selected\_features\_mutual\_info = X.columns[selector\_mutual\_info.get\_support()]

print("\nSelected features using mutual information:")

print(selected\_features\_mutual\_info.tolist())

# 可视化特征分数

def plot\_feature\_scores(selector, title):

    scores = selector.scores\_

    features = X.columns

    plt.figure(figsize=(10, 6))

    plt.barh(features, scores)

    plt.xlabel('Feature Scores')

    plt.title(title)

    plt.gca().invert\_yaxis()

plot\_feature\_scores(selector\_mutual\_info, 'Feature Scores using Mutual Information')

OUTPUT:

Selected features using mutual information:

['Year', 'Engine HP', 'highway MPG', 'city mpg', 'Popularity']

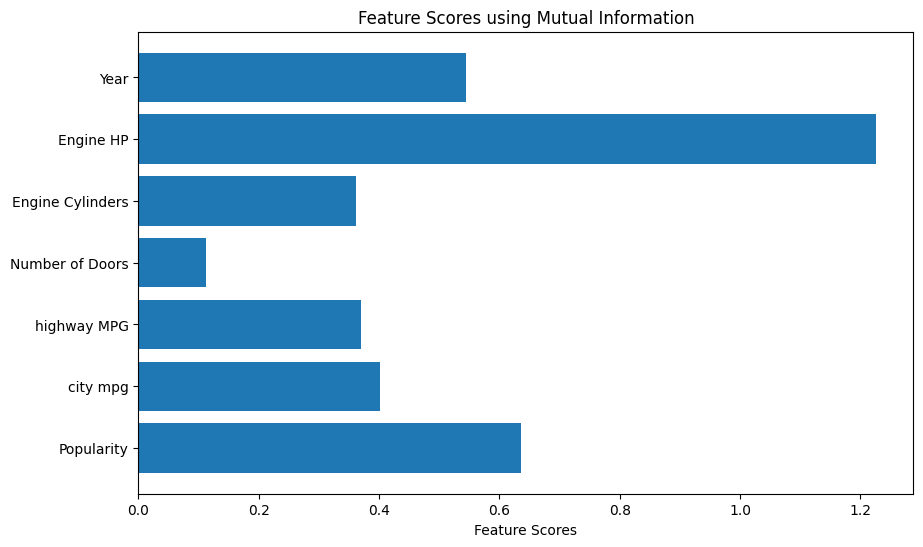


Рисунок 7: Результаты с использованием методов А и В

# Часть 4. Дополнительные требования

Для произвольной колонки данных построить гистограмму.

Используя набор данных № 1: bmw.csv

import pandas as pd

import matplotlib.pyplot as plt

data = pd.read\_csv('bmw.csv')

data['price'].hist(bins=30, edgecolor='black', alpha=0.7)

plt.title('Histogram of BMW Prices')

plt.xlabel('Price')

plt.ylabel('Frequency')

plt.show()

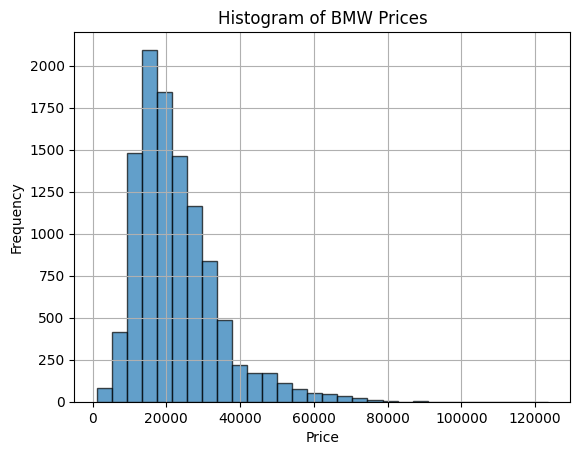


Рисунок 8: Гистограмма столбца Price

**ЗАКЛЮЧЕНИЕ**

В ходе выполнения рубежного контроля №1 по дисциплине «Методы машинного обучения» была проведена комплексная работа по обработке и анализу данных из двух различных наборов, связанных с автомобилями. В рамках задачи №16 была успешно применена нормализация данных с использованием преобразования Бокса-Кокса к числовому признаку «price» из набора данных bmw.csv. Это позволило преобразовать распределение цен на подержанные автомобили BMW к нормальному виду, что является важным шагом при подготовке данных для многих алгоритмов машинного обучения, чувствительных к масштабу и распределению признаков.

Для задачи №36 была выполнена процедура отбора признаков на наборе данных Car\_Features.csv с применением класса SelectKBest и метода, основанного на взаимной информации. В результате были выявлены пять наиболее значимых признаков для прогнозирования цены автомобиля, что демонстрирует эффективность использования данных методов в задачах предсказания и позволяет упростить модель, исключив менее значимые признаки, что может привести к улучшению её производительности и интерпретируемости.

Дополнительным требованием для группы было построение гистограммы для произвольной колонки данных, что было выполнено на примере столбца «price» из набора данных bmw.csv. Гистограмма наглядно представила распределение цен, что помогает визуально оценить форму распределения и presence возможных выбросов или аномалий в данных.