|  |  |
| --- | --- |
|  | **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\_\_\_\_\_\_\_\_\_\_\_

Отчет по лабораторной работе № **1**

**«Разведочный анализ данных. Исследование и визуализация данных.»**

по дисциплине «Технологии машинного обучения»

Студент ИУ5-62Б  М.С. Вольвач

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

Преподаватель  Ю.Е. Гапанюк

(Подпись, дата) (И.О.Фамилия)

**Москва**

**2024**

Цель работы:

Изучение различных методов визуализация данных.

Задание:

* Выбрать набор данных (датасет).
* Создать ноутбук, который содержит следующие разделы:
  + Текстовое описание выбранного Вами набора данных.
  + Основные характеристики датасета.
  + Визуальное исследование датасета.
  + Информация о корреляции признаков.

Описание набора данных:

Этот набор данных содержит различные характеристики, связанные с автомобилями, включая год выпуска, цену продажи, пройденный километраж, тип топлива, тип продавца, тип трансмиссии, количество предыдущих владельцев, пробег и технические характеристики двигателя. Эти атрибуты дают ценное представление о факторах, влияющих на цены автомобилей, и могут быть использованы для разработки прогнозных моделей для оценки цены продажи автомобилей.

Код программы:

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

df = pd.read\_csv("./cardekho.csv")

# Visualize the distribution of numerical features

plt.figure(figsize=(12, 6))

plt.subplot(2, 3, 1)

sns.histplot(df['year'], bins=20, kde=True)

plt.title('Год выпуска')

plt.subplot(2, 3, 2)

sns.histplot(df['selling\_price'], bins=20, kde=True)

plt.title('Цена продажи')

plt.subplot(2, 3, 3)

sns.histplot(df['km\_driven'], bins=20, kde=True)

plt.title('Пройдено километров')

plt.subplot(2, 3, 4)

sns.countplot(data=df, x='owner')

plt.title('Владельцы')

plt.subplot(2, 3, 5)

sns.histplot(df['mileage(km/ltr/kg)'], bins=20, kde=True)

plt.title('Расход топлива')

plt.subplot(2, 3, 6)

sns.histplot(df['engine'], bins=20, kde=True)

plt.title('Двигатель')

plt.tight\_layout()

plt.show()

plt.figure(figsize=(12, 4))

# Fuel

plt.subplot(1, 3, 1)

sns.countplot(data=df, x='fuel')

plt.title('Тип топлива')

# Seller\_Type

plt.subplot(1, 3, 2)

sns.countplot(data=df, x='seller\_type')

plt.title('Тип продавца')

# Transmission

plt.subplot(1, 3, 3)

sns.countplot(data=df, x='transmission')

plt.title('Тип трансмиссии')

plt.tight\_layout()

plt.show()

df\_numeric = df.drop(columns=['name', 'fuel', 'seller\_type', 'transmission', 'engine', "owner", 'mileage(km/ltr/kg)', 'engine', 'max\_power', 'seats'])

# Создаем корреляционную матрицу

corr\_matrix = df\_numeric.corr()

# Визуализируем корреляционную матрицу

plt.figure(figsize=(10, 8))

sns.heatmap(corr\_matrix, annot=True, cmap='coolwarm', fmt=".2f")

plt.title('Correlation Matrix')

plt.show()

Результат работы программы



