Отчет по лабораторной работе №3. Савва Даниил

В данной лабораторной работе для исследования были использованы:

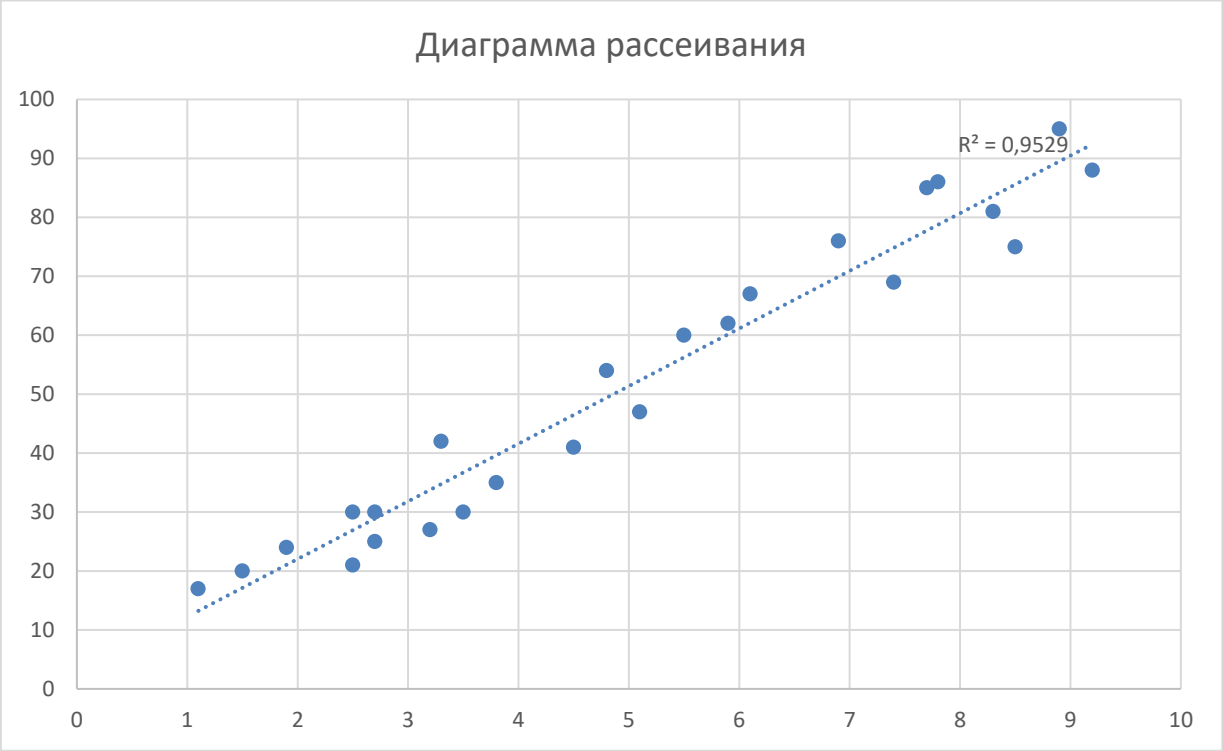
- данные зависимости количества учебных часов и полученных оценок студентами,

- статистика проданных автомобилей
https://www.kaggle.com/datasets/thedevastator/uncovering-factors-that-affect-used-car-prices.

1. В данный части лабораторной работы происходит исследование предположительной зависимости количество потраченных на учебы часов и полученных оценок студентами (файл student_scores.xlsx). Расчеты проводились в Excel.

Полученных коэффициент корреляции равен 0,976191, что означает очень сильную положительную корреляцию, а значит очень сильную зависимость данных друг от друга.

Была построена диаграмма рассеивания, проведена линия аппроксимации и вычислена величина достоверности аппроксимации R^2, который равен 0.9529.



2. Для части 2.1, сделали случайную выборку. Для части 2.2, стратифицированной выборки, была выбрана фильтрация по дате регистрации проданного автомобиля.

3. Данная часть работы была сделана с использованием python, код из main.py

```
import pandas as pd
from pandas import DataFrame
```

```python
import numpy
import matplotlib.pyplot as plt
from scipy import stats

data_general: DataFrame = pd.read_csv('autos.csv')
data_random: DataFrame = data_general.sample(frac=1/3, random_state=100)
data_stratified: DataFrame = data_general.groupby('yearOfRegistration',
group_keys=False).apply(lambda x: x.sample(frac=1/3, random_state=100))

COLUMN_NAME: str = "price"

# Среднего значения по выборкам
print(f"Среднее значение (генеральная совокупность):
{data_general[COLUMN_NAME].mean():.2f}")
print(f"Среднее значение (случайная выборка): {data_random[COLUMN_NAME].mean():.2f}")
print(f"Среднее значение (стратифицированная выборка):
{data_stratified[COLUMN_NAME].mean():.2f}")


def confidence_interval(data: DataFrame, confidence: float) -> tuple:
    """Расчет доверительного интервала"""
    length: int = len(data)
    mean: float = numpy.mean(data)
    sem = stats.sem(data)
    margin = sem * stats.t.ppf((1 + confidence) / 2, length - 1)
    return mean - margin, mean + margin

# Доверительные интервалы для случайной выборки
print("\n")
print(f"Доверительный интервал - случайная выборка 90%:
{confidence_interval(data=data_random[COLUMN_NAME], confidence=0.90)}")
print(f"Доверительный интервал - случайная выборка 95%:
{confidence_interval(data=data_random[COLUMN_NAME], confidence=0.95)}")
print(f"Доверительный интервал - случайная выборка 99%:
{confidence_interval(data=data_random[COLUMN_NAME], confidence=0.99)}")

# Доверительные интервалы для стратифицированной выборки
print("\n")
print(f"Доверительный интервал - стратифицированная выборка 90%:
{confidence_interval(data=data_stratified[COLUMN_NAME], confidence=0.90)}")
print(f"Доверительный интервал - стратифицированная выборка 95%:
{confidence_interval(data=data_stratified[COLUMN_NAME], confidence=0.95)}")
print(f"Доверительный интервал - стратифицированная выборка 99%:
{confidence_interval(data=data_stratified[COLUMN_NAME], confidence=0.99)}")
```

Среднее значение (генеральная совокупность): 17295.14

Среднее значение (случайная выборка): 10759.67

Среднее значение (стратифицированная выборка): 8547.26

Доверительный интервал - случайная выборка 90%: (8013.152120474991, 13506.17811215826)

Доверительный интервал - случайная выборка 95%: (7486.985523071335, 14032.344709561918)

Доверительный интервал - случайная выборка 99%: (6458.610810464762, 15060.719422168491)

Доверительный интервал - стратифицированная выборка 90%: (6595.928704574953, 10498.58854460135)

Доверительный интервал - стратифицированная выборка 95%: (6222.100280212627, 10872.416968963676)

Доверительный интервал - стратифицированная выборка 99%: (5491.46534298661, 11603.051906189694)