|  |  |
| --- | --- |
| **Gerb-BMSTU_01** | **Министерство науки и высшего образования Российской Федерации**  Калужский филиал  федерального государственного бюджетного  образовательного учреждения высшего образования  ***«Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»***  ***(КФ МГТУ им. Н.Э. Баумана)*** |

**ФАКУЛЬТЕТ** ***ИУК «Информатика и управление»***

**КАФЕДРА** \_\_***ИУК4 «Программное обеспечение ЭВМ, информационные технологии»***

**ЛАБОРАТОРНАЯ РАБОТА №2**

**«Факторный анализ данных. Корреляция»**

**ДИСЦИПЛИНА: «Технологии анализа данных»**

|  |  |  |
| --- | --- | --- |
| Выполнил: студент гр. ИУК4-82Б | | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ( Карельский М.К. )  (Подпись) |
| Проверил: | | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ( Ерохин И.И. )  (Подпись) |
| Дата сдачи (защиты):  Результаты сдачи (защиты): | | |
|  | - Балльная оценка:  - Оценка: | |

Калуга, 2024

**Цель:** формирование практических навыков проведения факторного анализа и обнаружения корреляции между параметрами.

**Задачи:**

1. Ознакомиться с понятием факторный анализ и корреляция.
2. Изучить средства языка Python для выполнения факторного анализа

**Вариант 5**

Считать данные из CSV файла в структуру DataFrame. Провести факторный анализ зависимости параметра Purchase от 2 характеристик: Occupation и Stay\_In\_Current\_City\_Years. Вывести результаты (summary) анализа и интерпретировать их. Построить столбчатую диаграмму, отражающую степень влияния каждого параметра на число Purchase. Повторить анализ для данных разного размера. Построить график зависимости коэффициента детерминации (R-square) от размера набора данных. Сделать выводы. Построить и визуализировать корреляционную матрицу для всех параметров (Purchase, Occupation, Stay\_In\_Current\_City\_Years). Сделать выводы о наличии связей между параметрами.

**Листинг:**

import pandas as pd

import numpy as np

import statsmodels.api as sm

import matplotlib.pyplot as plt

import seaborn as sns

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

data['Stay\_In\_Current\_City\_Years'] = data['Stay\_In\_Current\_City\_Years'].replace('4+', 4).astype(int)

X = data[['Occupation', 'Stay\_In\_Current\_City\_Years']]

y = data['Purchase']

X = sm.add\_constant(X)

model = sm.OLS(y, X).fit()

print(model.summary())

coefficients = model.params.drop('const')

coefficients.plot(kind='bar')

plt.title('Influence of factors on Purchase')

plt.xlabel('Factors')

plt.ylabel('Coefficient')

plt.show()

r\_square\_values = []

data\_sizes = range(50000, len(data), 50000)

for size in data\_sizes:

    sample\_data = data.sample(size)

    X\_sample = sample\_data[['Occupation', 'Stay\_In\_Current\_City\_Years']]

    y\_sample = sample\_data['Purchase']

    X\_sample = sm.add\_constant(X\_sample)

    model\_sample = sm.OLS(y\_sample, X\_sample).fit()

    r\_square\_values.append(model\_sample.rsquared)

plt.plot(data\_sizes, r\_square\_values)

plt.title('R-square vs Data Size')

plt.xlabel('Data Size')

plt.ylabel('R-square')

plt.show()

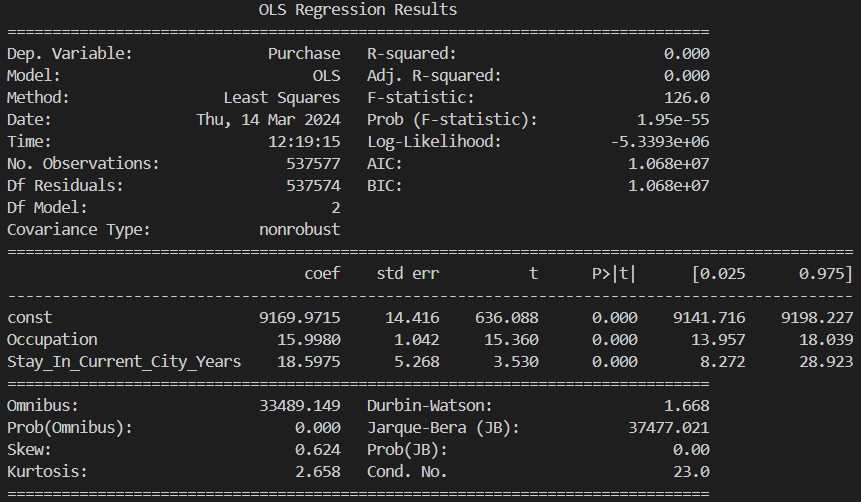
correlation\_matrix = data[['Purchase', 'Occupation', 'Stay\_In\_Current\_City\_Years']].corr()

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

plt.title('Correlation Matrix')

plt.show()

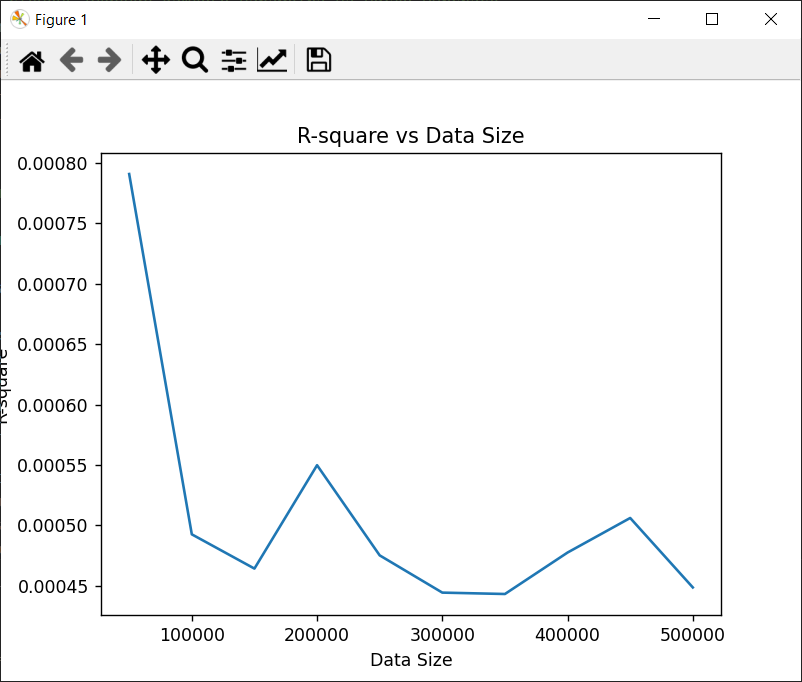
**Результат:**



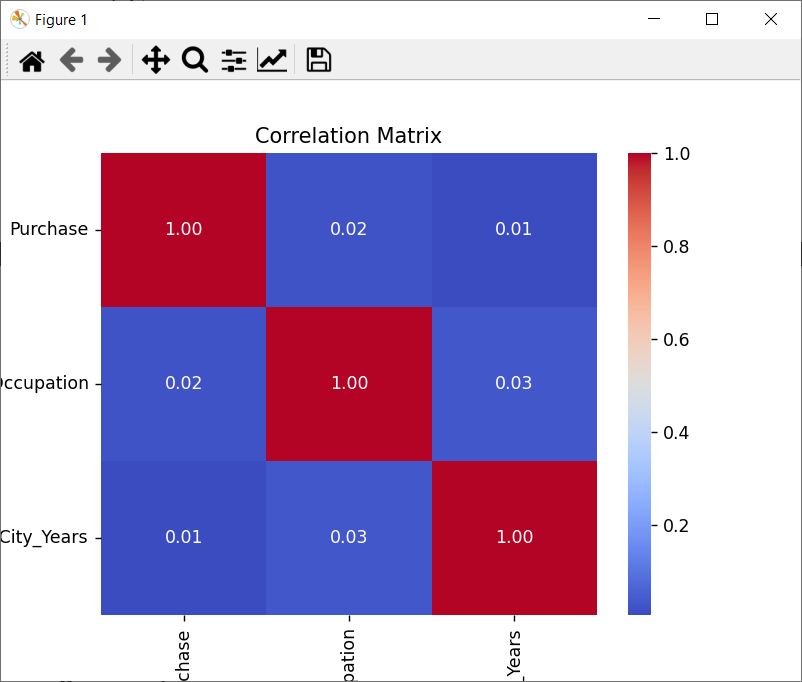
**Рис. 1.** Результаты анализа



**Рис. 2.** Столбчатая диаграмма



**Рис. 3.** График



**Рис. 4.** Корреляционная матрица

Результаты анализа показывают, что модель линейной регрессии имеет низкое значение коэффициента детерминации, близкое к нулю. Это означает, что объясняющие переменные не объясняют значительной части вариации в зависимой переменной:

* Значение R-squared равно 0, что означает, что модель не объясняет никакой доли изменчивости зависимой переменной Purchase.
* P-values для всех коэффициентов модели намного меньше 0.05, что говорит о статистической значимости коэффициентов. То есть, существует статистически значимая связь между Occupation, Stay\_In\_Current\_City\_Years и Purchase.
* Коэффициенты для Occupation и Stay\_In\_Current\_City\_Years равны примерно 16 и 18.6 соответственно. Это означает, что при увеличении Occupation на единицу, ожидается увеличение Purchase на 16 единиц, а при увеличении Stay\_In\_Current\_City\_Years на единицу, ожидается увеличение Purchase на 18.6 единиц.
* Промежутки доверительных интервалов для коэффициентов не содержат нуля, что подтверждает статистическую значимость этих коэффициентов.

В целом, модель показывает статистически значимую, но очень слабую связь между Occupation, Stay\_In\_Current\_City\_Years и Purchase.

**Вывод:** в ходе выполнения лабораторной работы были получены практические навыки проведения факторного анализа и обнаружения корреляции между параметрами.