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:

OLS Regression Results

Dep. Variable:	Purchase	R-squared:	0.000
Model:	OLS	Adj. R-squared:	0.000
Method:	Least Squares	F-statistic:	126.0
Date:	Thu, 14 Mar 2024	Prob (F-statistic):	1.95e-55
Time:	12:19:15	Log-Likelihood:	-5.3393e+06
No. Observations:	537577	AIC:	1.068e+07
Df Residuals:	537574	BIC:	1.068e+07
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	9169.9715	14.416	636.088	0.000	9141.716	9198.227
Occupation	15.9980	1.042	15.360	0.000	13.957	18.039
Stay_In_Current_City_Years	18.5975	5.268	3.530	0.000	8.272	28.923

Omnibus:	33489.149	Durbin-Watson:	1.668
Prob(Omnibus):	0.000	Jarque-Bera (JB):	37477.021
Skew:	0.624	Prob(JB):	0.00
Kurtosis:	2.658	Cond. No.	23.0

Рис. 1. Результаты анализа

Рис. 2. Столбчатая диаграмма

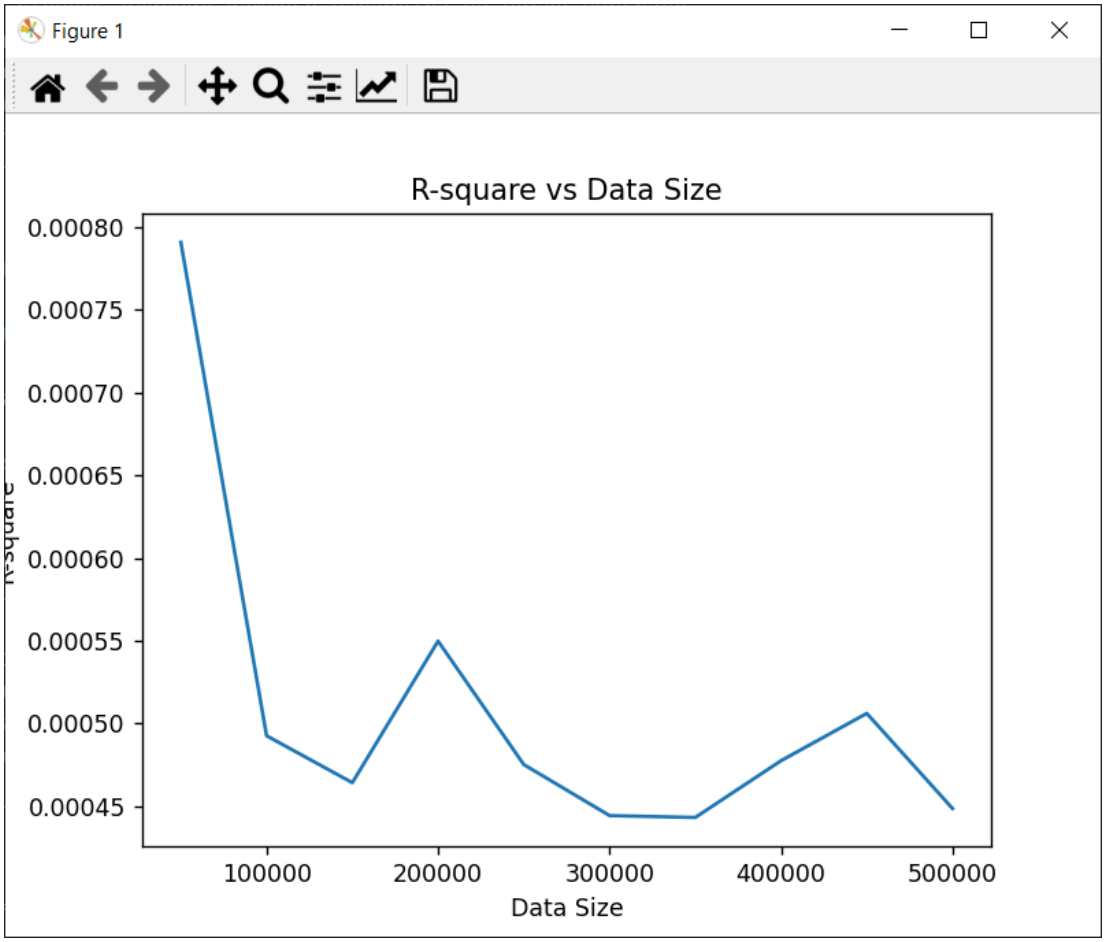

Рис. 3. График

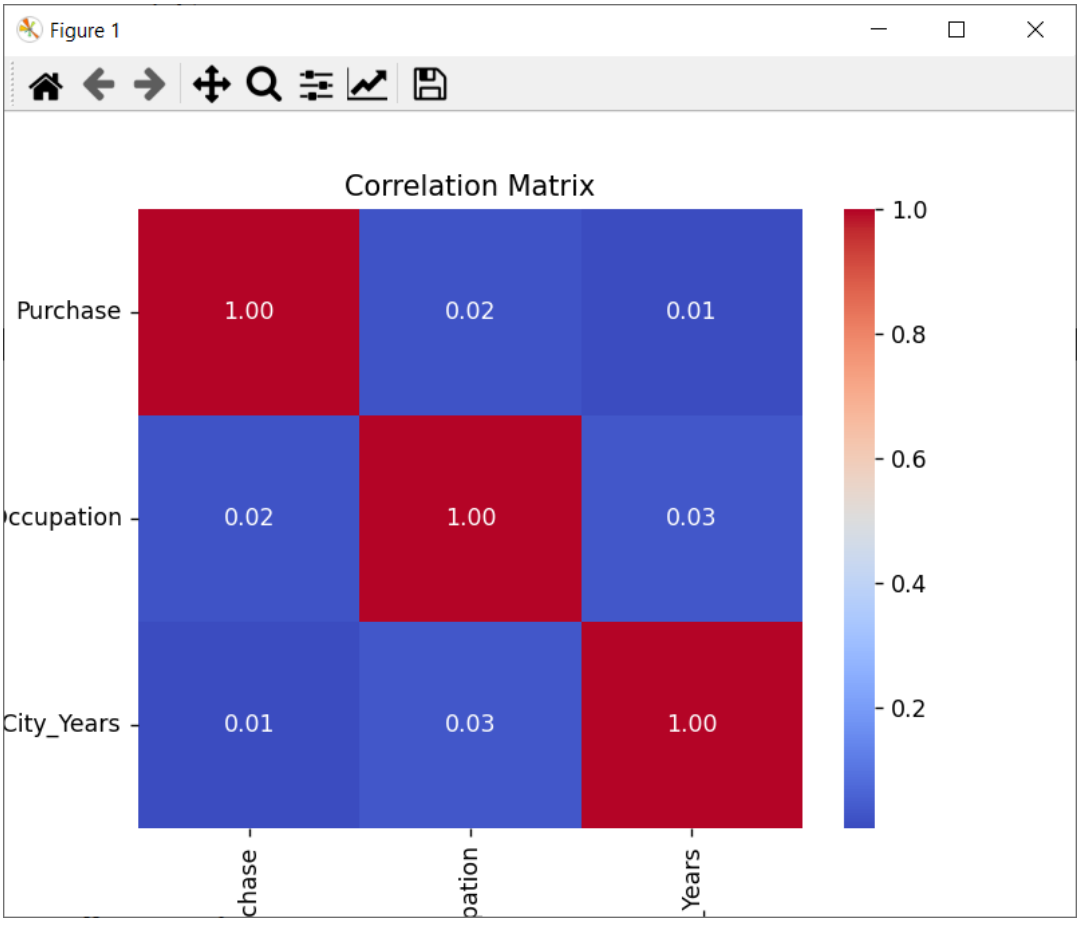

Рис. 4. Корреляционная матрица

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.