

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ  
Федеральное государственное автономное образовательное учреждение высшего образования  
Национальный исследовательский технологический университет «МИСИС»  
Институт информационных технологий и автоматизированных систем управления  
Кафедра Бизнес-информатики и систем управления производством

**Практическая работа №6**

по дисциплине «Статистические методы анализа данных в принятии решений»  
на тему «Виды и анализ распределений данных»

Направление подготовки  
38.03.05 Бизнес-информатика  
Семестр 4

Выполнил:

Сычиков Владимир Андреевич

\_\_\_\_\_  
(ФИО студента)

ББИ-23-6

\_\_\_\_\_  
(№ группы)

31.03.2025

\_\_\_\_\_  
(дата сдачи)

Подпись:

\_\_\_\_\_

Проверил:

\_\_\_\_\_  
(ФИО преподавателя)

\_\_\_\_\_  
(оценка)

\_\_\_\_\_  
(дата проверки)

Подпись:

\_\_\_\_\_

Москва – 2025

## Ход работы:

Для начала работы я импортировал все необходимые библиотеки(pandas, numpy) с помощью команды `import`, а также задал элиасы для удобства обращения к библиотеке. Далее я подгрузил выборку и записал ее в датафрейм `df`. Также я выделил нужные мне строки из выборки, по заданию.

```
[1] import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
from statsmodels.stats.outliers_influence import variance_inflation_factor

df = pd.read_excel('data.xlsx')
```

Рисунок 1 - Импорт библиотек

На рисунке 2 реализуется процесс построения тепловой карты корреляционной матрицы данных, загруженных из файла "data.xlsx". Данные считываются с пропуском второй строки (`skiprows=[1]`) и использованием первой строки в качестве заголовков столбцов (`header=0`).

Тепловая карта (heatmap) визуализирует корреляции между переменными с помощью цветовой шкалы (`cmap='coolwarm'`), где значения близкие к 1 выделены теплыми оттенками (красный), а близкие к -1 — холодными (синий). Нулевые значения корреляции соответствуют центру шкалы (`center=0`). Аннотации (`annot=True`) отображают числовые значения коэффициентов корреляции непосредственно на карте.

График позволяет быстро оценить силу и направление взаимосвязей между переменными: высокие абсолютные значения корреляции указывают на тесную линейную зависимость, а близкие к нулю — на ее отсутствие.

```
1 data = pd.read_excel("data.xlsx", header=0, skiprows=[1])
сек. corr_matrix = data.corr()
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', center=0)
plt.title('Корреляционная матрица')
plt.show()
```

Рисунок 2 - Построение тепловой карты корреляционной матрицы данных

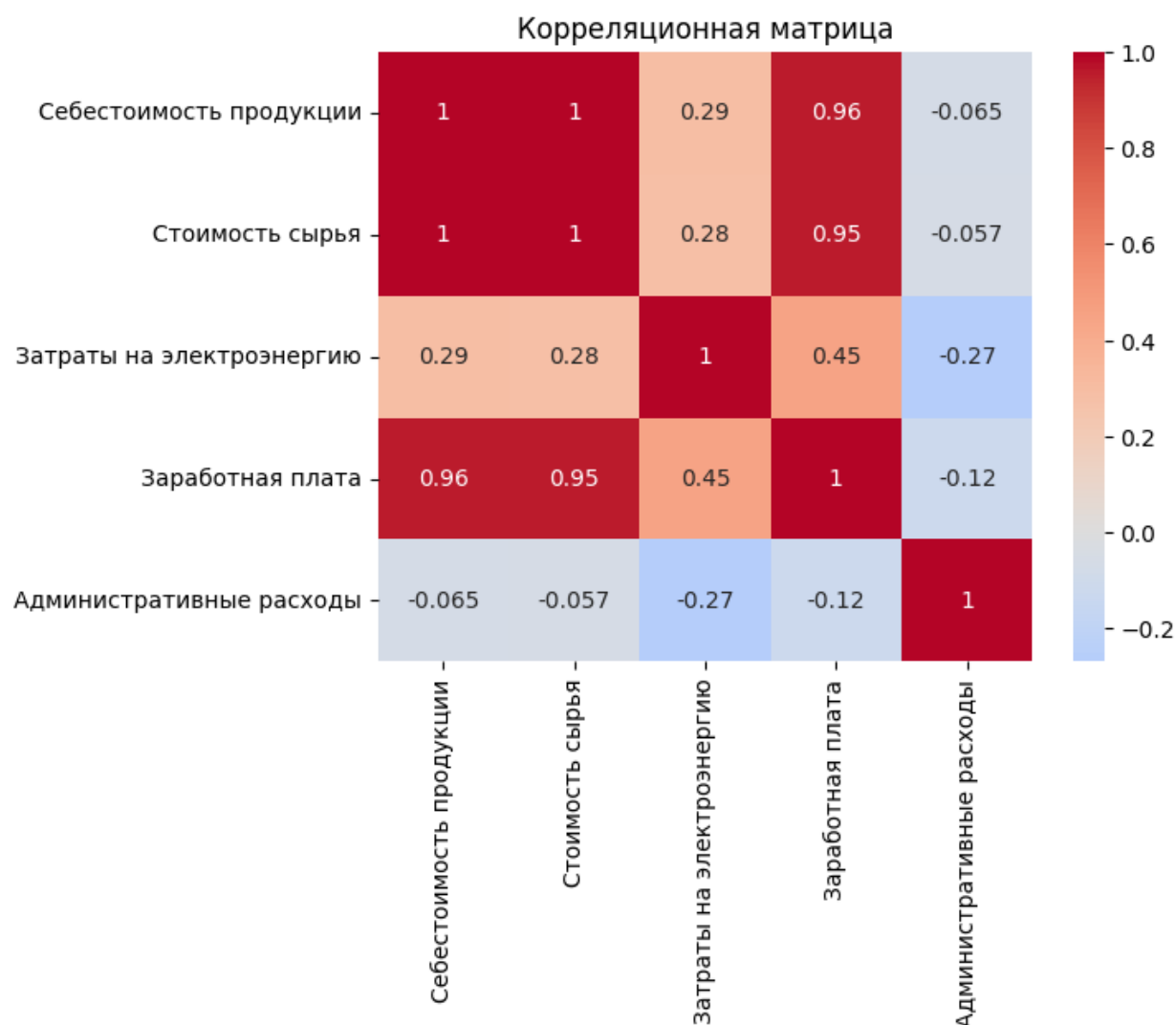


Рисунок 3 - Тепловая карта корреляционной матрицы данных

Сильная положительная корреляция (близко к 1):

- Себестоимость продукции и стоимость сырья (0.96) — почти линейная зависимость, что объясняется тем, что сырье — ключевая составляющая себестоимости.
- Себестоимость продукции и заработная плата (0.96), а также стоимость сырья и заработная плата (0.95) — высокая корреляция, возможно, зарплата зависит от объемов производства, например, в формулу заработной платы заложено KPI.

Умеренная положительная корреляция (0.28–0.45):

- Затраты на электроэнергию слабо связаны с другими факторами, кроме заработной платы (0.45) — возможно, энергопотребление растет при увеличении штата.

Слабая/отсутствующая корреляция (близко к 0):

- Административные расходы почти не зависят от других показателей (значения около -0.2...-0.06), что говорит об их фиксированном характере.

На рисунке 4 представлен процесс расчета показателей инфляции дисперсии (Variance Inflation Factor, VIF), который выполняется для оценки степени

мультиколлинеарности между независимыми переменными в наборе данных. Анализ проводится с использованием библиотек pandas для работы с данными и statsmodels для статистического моделирования.

Для корректного расчета VIF к исходным данным сначала добавляется константный столбец с помощью функции `sm.add_constant(data)`. Это необходимо, так как метод VIF основан на регрессионном анализе, где константа играет важную роль. Затем в цикле последовательно вычисляются значения VIF для каждого признака (за исключением добавленной константы) с помощью функции `variance_inflation_factor`. Полученные значения сохраняются в список `vifs` для последующего анализа.

```
data_with_constant = sm.add_constant(data)
vifs = []
for i in range(1, len(data.columns)):
    vif = variance_inflation_factor(data_with_constant.values, i)
    vifs.append(vif)

print("Variance Inflation Factor")
for idx, vif in enumerate(vifs):
    print(f"VIF для {data.columns[idx]}: {round(vif, 2)}")
```

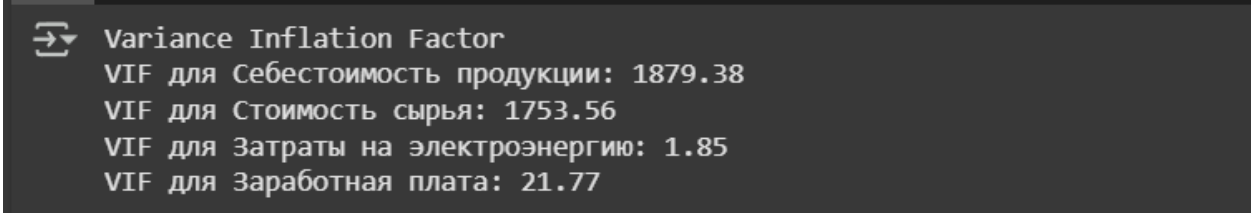


Рисунок 4 - VIF-анализ

Результаты расчета выводятся в удобочитаемом формате, где для каждого показателя указывается его название и соответствующее значение VIF, округленное до двух знаков после запятой.

Значения VIF интерпретируются следующим образом:

- $VIF < 5$  указывает на отсутствие существенной мультиколлинеарности
- VIF между 5 и 10 свидетельствует об умеренной мультиколлинеарности
- $VIF > 10$  указывает на серьезную проблему мультиколлинеарности

VIF для Себестоимости продукции, Стоимости сырья, Заработной платы выше 10, что означает очень высокую вероятность мультиколлинеарности, что нужно учитывать про аналитике корреляции и формулировании гипотез.

VIF для Затрат на электроэнергию, наоборот, не превышает 5, что означает низкую вероятность мультиколлинеарности, что может означать возможность корректной интерпретации результатов.

В данных нет нелинейных зависимостей, поэтому сразу можем отказаться от нелинейной модели.

В данных не присутствует логарифмический и экспоненциальный рост, поэтому обобщенную линейную модель не используем

Учитывая фактор мультиколлинеарность во всех столбцах кроме затрат на электроэнергию все равно можем заметить нормальную корреляцию затрат на электроэнергию с другими параметрами.

Учитывая то, что использование нелинейной и обобщенной линейной модели не обосновано и достаточно трудозатратно в нашем случае, выберем линейную модель.

Также, в дальнейшем нужно оставить в модели "Затраты на электроэнергию" как независимый значимый фактор и дополнительно проанализировать влияние административных расходов, которые в текущих данных демонстрируют слабые корреляции.

#### **Вывод:**

Проведенный анализ данных выявил серьезную проблему мультиколлинеарности между ключевыми производственными показателями, особенно между себестоимостью продукции и стоимостью сырья, что подтверждается как корреляционным анализом ( $r=1$ ), так и экстремально высокими значениями VIF (свыше 1750). При этом затраты на электроэнергию продемонстрировали относительную независимость от других факторов (VIF=1.85), что делает их ценным предиктором для моделирования. Заработная плата показала умеренную мультиколлинеарность (VIF=21.77), а административные расходы слабо коррелировали с остальными переменными. Учитывая отсутствие в данных явных нелинейных зависимостей и признаков логарифмического или экспоненциального роста, наиболее обоснованным выбором представляется классическая линейная регрессионная модель. Однако для ее корректной работы необходимо рассмотреть возможность преобразования или исключения показателя заработной платы, сохранив при этом затраты на электроэнергию как значимый независимый фактор. Дополнительного внимания заслуживает анализ роли административных расходов, которые в текущих данных демонстрируют слабые связи с другими показателями.