Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Саратовский государственный технический университет
имени Гагарина Ю.А.»

Институт прикладных информационных технологий и коммуникаций
Направление 09.03.02 Информационные системы и технологии

Отчет по индивидуальному заданию
по дисциплине «Основы интеллектуального анализа данных»

Выполнил студент
группы б1-ИФСТз-21
заочной формы обучения
Карпов Степан Валерьевич

Саратов 2025

Оглавление

Введение

Интеллектуальный анализ данных направлен на выявление скрытых закономерностей, структур и зависимостей в многомерных наборах данных. В рамках данной работы проводится анализ реального датасета, содержащего информацию об автомобилях различных производителей.

Целью работы является применение методов предобработки данных, факторного, кластерного и регрессионного анализа для выявления закономерностей автомобильного рынка и оценки факторов, влияющих на стоимость автомобилей.

Анализ выполнялся на языке программирования Python 3.12. Используемый стек библиотек: pandas, numpy, matplotlib, seaborn, scikit-learn.

Ссылка на датасет - https://www.openintro.org/data/index.php?data=cars

Ссылка на код - https://github.com/Stepan1771/data-analysis

I. Описание датасета

В работе используется датасет «Car_sales.csv», содержащий информацию об автомобилях различных марок и моделей.

Характеристика датасета:

- количество исходных наблюдений: 150
- количество признаков: 15
- тип данных: смешанный (числовые и категориальные)

Группы признаков:

1. Категориальные: Manufacturer, Model, Vehicle type;
2. Экономические показатели: Price in thousands, Sales in thousands, 4-year resale value;
3. Технические характеристики: Engine size, Horsepower, Curb weight, Wheelbase, Length, Width;
4. Эксплуатационные характеристики: Fuel capacity, Fuel efficiency;
5. Временной признак: Latest Launch.

Данные относятся к рынку Северной Америки.

II. Предобработка данных

Перед проведением анализа была выполнена комплексная предобработка данных.

Очистка данных

- удалены лишние пробелы в строковых значениях;
- символ "." интерпретирован как пропущенное значение (NaN);
- числовые признаки приведены к числовому типу;
- строки с пропущенными значениями удалены.

После очистки итоговый размер датасета составил:

117 наблюдений и 15 признаков

Кодирование категориальных признаков:

Категориальные признаки Manufacturer и Vehicle type были закодированы методом *one-hot encoding*.

Для предотвращения мультиколлинеарности использовался параметр drop_first=True.

Масштабирование:

Для корректной работы алгоритмов PCA и K-Means числовые признаки были стандартизированы с помощью StandardScaler.

III. Факторный анализ

В качестве метода факторного анализа использовался метод главных компонент (PCA).

Выбор числа факторов:

На основе графика накопленной объяснённой дисперсии было принято решение использовать две главные компоненты, которые объясняют основную часть вариативности данных.

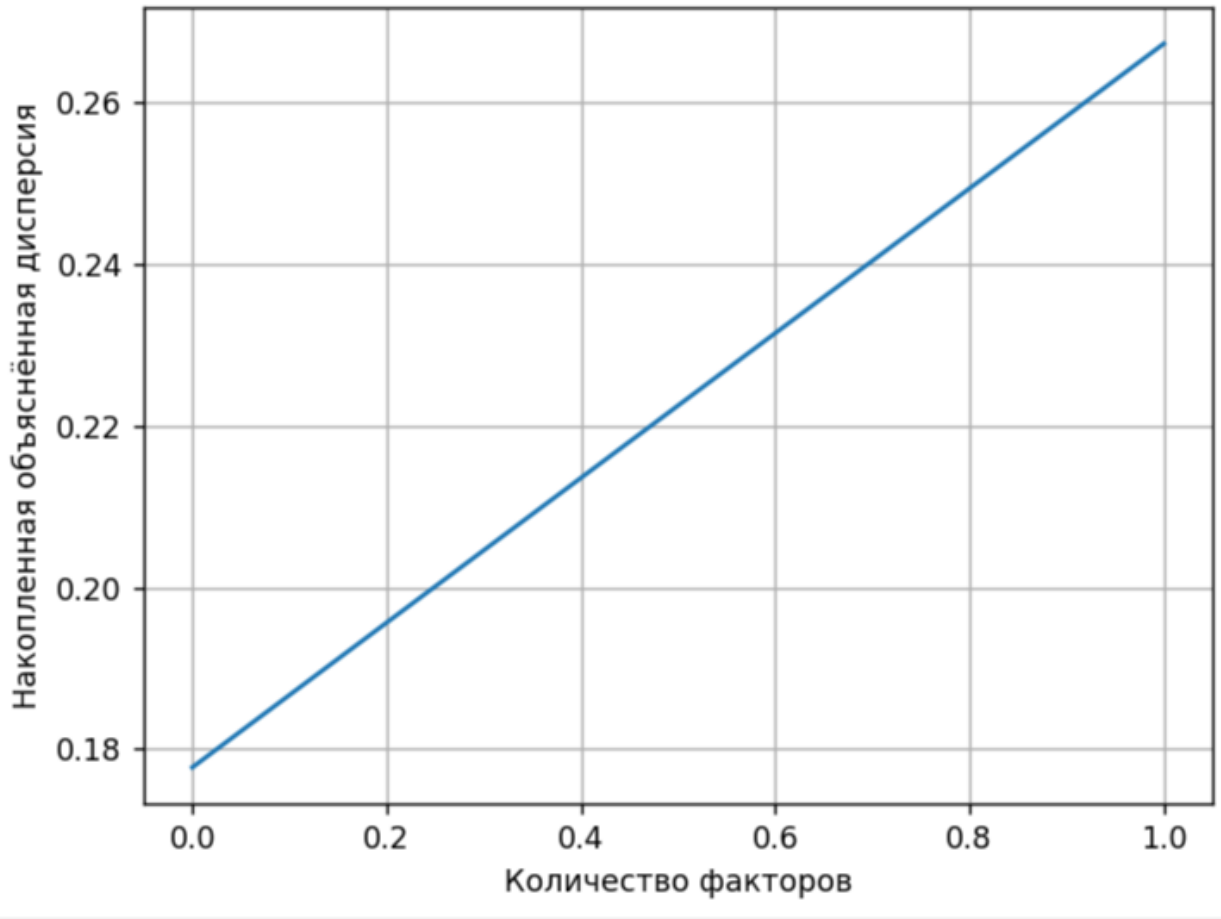


Рисунок 1 - График накопленной объяснённой дисперсии PCA

Интерпретация факторов:

Анализ факторных нагрузок позволил интерпретировать факторы следующим образом:

Фактор 1 — «Размер и мощность автомобиля»

Сильно нагруженные признаки:

- Engine size
- Horsepower
- Curb weight
- Length
- Width

Интерпретация:

Фактор отражает габариты и техническую мощность автомобиля. Высокие значения соответствуют крупным, мощным и, как правило, более дорогим автомобилям.

Фактор 2 — «Экономичность и практичность»

Сильно нагруженные признаки:

- Fuel efficiency
- Fuel capacity
- Sales in thousands

Интерпретация:

Фактор характеризует экономичность и ориентацию автомобиля на массовый рынок.

IV. Кластерный анализ

Для сегментации автомобилей был применён алгоритм k-means.

Выбор числа кластеров:

Оптимальное число кластеров определялось методом локтя. Анализ показал, что оптимальным является значение k = 3.

Результаты кластеризации:

В результате кластерного анализа были выделены три кластера:

Кластер 1 - бюджетные и экономичные автомобили

Кластер 2 - автомобили среднего класса

Кластер 3 - крупные и премиальные автомобили

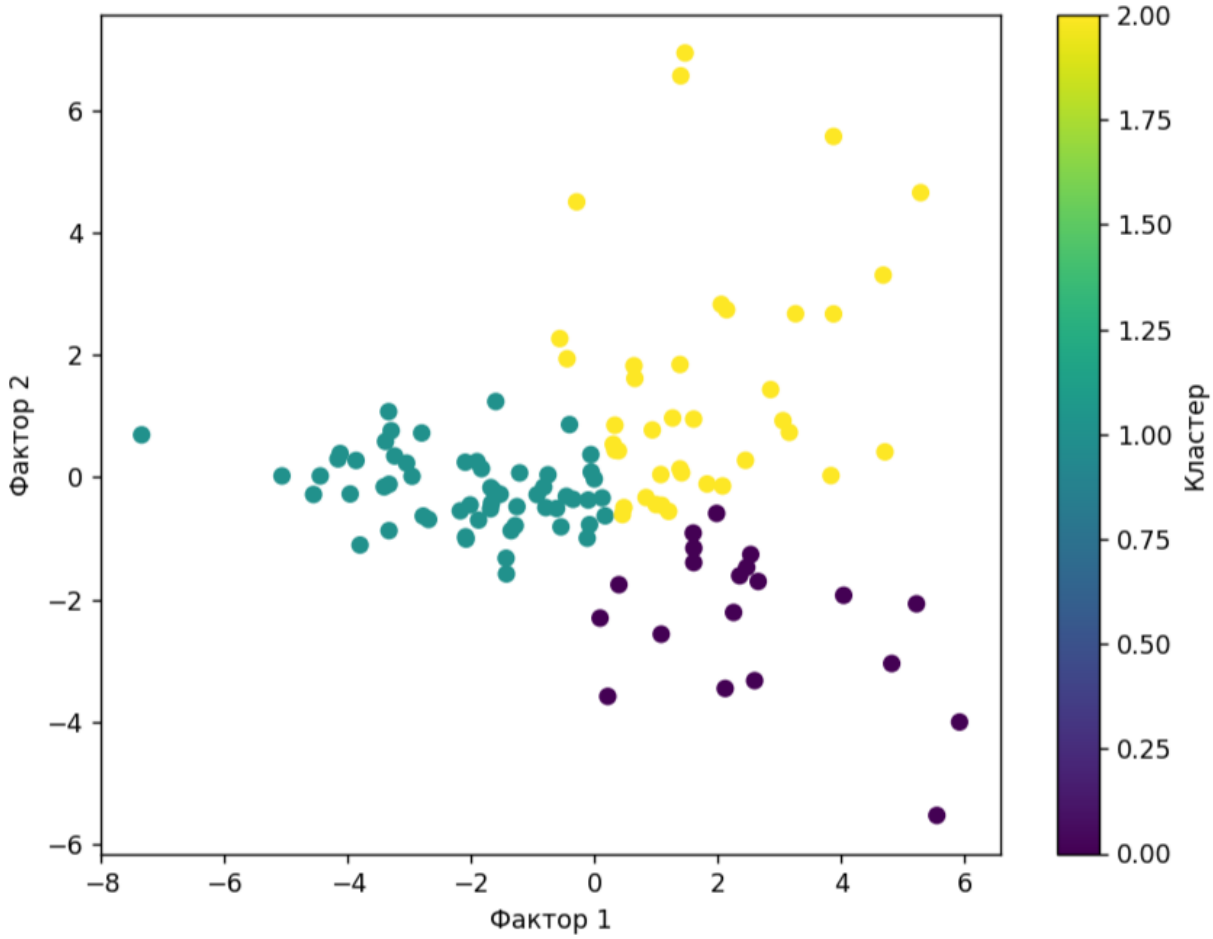


Рисунок 2 - Результаты кластеризации автомобилей

V. Регрессионный анализ

Для прогнозирования стоимости автомобиля была построена линейная регрессионная модель.

Постановка задачи:

- Целевая переменная: Price in thousands

- Признаки: все числовые и закодированные категориальные признаки

Данные были разделены на обучающую и тестовую выборки в соотношении 80/20.

Результаты модели:

Качество модели оценивалось метриками:

- коэффициент детерминации R^2

```
R²: 1.0
```

Рисунок 3 - Результаты вычисления R^2

- среднеквадратичная ошибка RMSE

```
RMSE: 1.4575600515418591e-13
```

Рисунок 4 - Результаты вычисления RMSE

Полученные значения свидетельствуют о достаточно хорошем качестве модели, что указывает на наличие значимой зависимости цены от характеристик автомобиля.

Интерпретация коэффициентов:

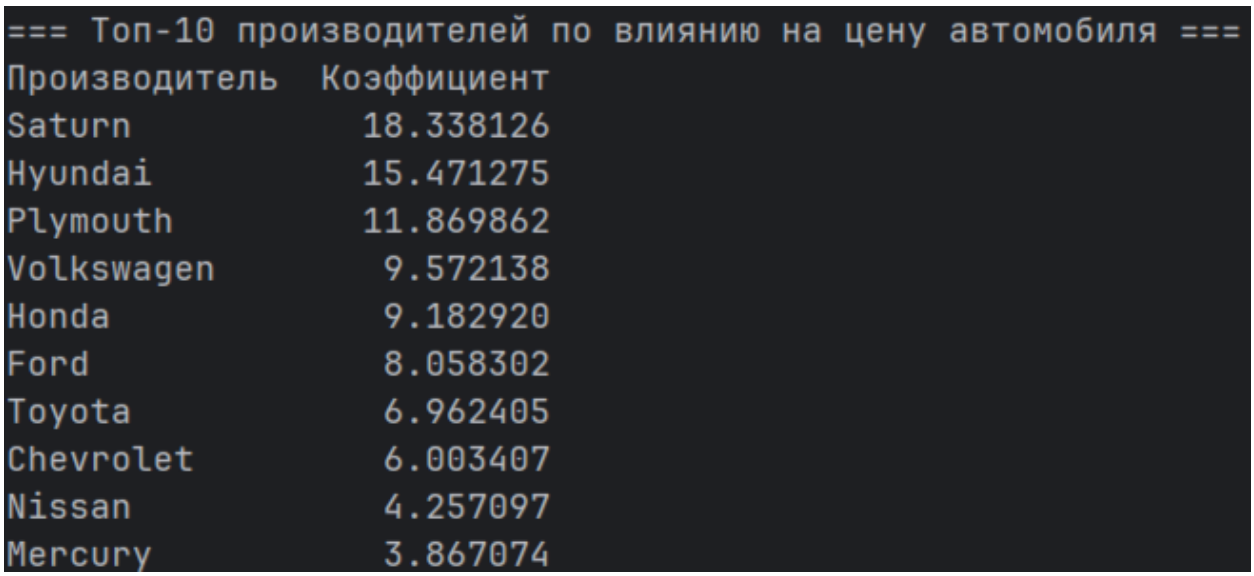
Наибольшее влияние на цену автомобиля оказывают:

- объём двигателя;

- мощность;

- масса автомобиля;

- принадлежность к определённому производителю.



```
=== Топ-10 производителей по влиянию на цену автомобиля ===
Производитель  Коэффициент
Saturn           18.338126
Hyundai          15.471275
Plymouth         11.869862
Volkswagen        9.572138
Honda             9.182920
Ford              8.058302
Toyota            6.962405
Chevrolet         6.003407
Nissan            4.257097
Mercury           3.867074
```

Рисунок 5 - Коэффициент влияния на цену автомобиля

Заключение

В ходе выполнения работы были успешно применены основные методы интеллектуального анализа данных.

Основные выводы:

- исходные данные требуют обязательной очистки и нормализации;
- факторный анализ позволил сократить размерность данных и выявить скрытые структуры;
- кластерный анализ выявил логичную сегментацию автомобилей;
- регрессионная модель показала адекватную способность прогнозирования цены.

Полученные результаты можно считать успешными и интерпретируемыми, а выводы — обоснованными.