**МИНИСТЕРСТВО ТРАНСПОРТА РОССИЙСКОЙ ФЕДЕРАЦИИ**

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ**

**УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ**

**«РОССИЙСКИЙ УНИВЕРСИТЕТ ТРАНСПОРТА»**

**(РУТ (МИИТ)**

Институт/факультет «Академия “Высшая Инженерная Школа”, АВИШ»

Специальность/Направление подготовки Информатика и вычислительная техника

**КУРСОВАЯ РАБОТА**

**по дисциплине:** Технологии хранения больших данных

|  |  |
| --- | --- |
| **на тему:** | Анализ мирового автомобильного рынка и выявление зависимостей между техническими характеристиками автомобилей с использованием методов Big Data |

|  |  |
| --- | --- |
| **Студент группы ШАД-211** | **(Лозовой В.С )** |
| **Научный руководитель** | **(Зуев И.А. )** |

**Москва 2024 г.**

**Содержание**

**Основные термины/сокращения**

**Введение**

**1. Подготовка и обработка данных**

**1.1. Источник данных и автоматическое чтение**

**1.2. Очистка и предобработка данных**

**1.3. Создание числовых признаков**

**2. Проверка гипотез и визуализация результатов**

**2.1. Анализ распределения гибридных и электрических автомобилей странам**

**2.2. Исследование зависимости динамики автомобиля от количества мест**

**2.3. Корреляционный анализ цены и максимальной скорости**

**2.4. Выявление стран-лидеров по производству скоростных автомобилей**

**2.5. Анализ ценового распределения гибридных автомобилей**

**Заключение**

**Список использованных источников и программ**

**Приложение**

1. **Основные термины/сокращения**

**DataFrame** — основная структура данных в библиотеке **Pandas**, представляющая собой таблицу с индексированными строками и столбцами для обработки и анализа данных.

**ETL (Extract, Transform, Load)** — процесс извлечения, преобразования и загрузки данных.

**Гибридный автомобиль** — автомобиль, использующий для привода ведущих колёс как двигатель внутреннего сгорания, так и электродвигатель.

**Корреляция Пирсона** — статистическая мера, отражающая степень линейной зависимости между двумя переменными.

**Big Data** — большие объемы данных, которые сложно обрабатывать традиционными методами из-за их размера, скорости поступления и разнообразия.

**Pandas** — библиотека **Python** для обработки и анализа данных.

**Seaborn** — библиотека **Python** для построения статистических графиков на основе **Matplotlib.**

**Гипотеза** — научное предположение, выдвигаемое для объяснения каких-либо явлений и требующее проверки на опыте.

1. **Введение**

Современный мировой автомобильный рынок характеризуется не только ростом объёмов производства, но и значительным усложнением структуры. Появление новых типов двигателей (гибридных, электрических), ужесточение экологических норм и изменение потребительских предпочтений формируют сложную, многомерную картину, для анализа которой требуются современные методы обработки данных.

Автомобиль сегодня — это сложный продукт, описываемый десятками параметров: от технических характеристик (мощность, разгон, тип топлива) до экономических (цена, страна производства). Анализ взаимосвязей между этими параметрами позволяет выявить ключевые тенденции рынка, стратегии производителей и потенциальные ниши для развития. Традиционные методы анализа часто не справляются с большими и разнородными наборами данных, что обуславливает необходимость применения подходов Big Data и автоматизированной аналитики.

**Цель работы** — провести комплексный анализ набора данных об автомобилях, проверить ключевые гипотезы о взаимосвязях между их характеристиками и визуализировать полученные результаты с использованием современных инструментов анализа данных на Python.

**Задачи работы:**

Реализовать надёжный механизм чтения и первичной обработки сырых данных из CSV-файла.

Провести очистку данных, стандартизацию и создание производных числовых признаков.

Проверить серию гипотез, касающихся распределения типов автомобилей по странам, взаимосвязи их характеристик и ценовой политики.

Визуализировать полученные результаты для наглядного представления выявленных закономерностей.

Для достижения поставленной цели используются следующие инструменты: Pandas — для обработки и анализа табличных данных, Seaborn и Matplotlib — для визуализации, Scipy — для статистического анализа.

**1. Подготовка и обработка данных**

**1.1 Источник данных и автоматическое чтение**

В качестве источника данных используется файл Cars.csv, содержащий информацию о 1395 автомобилях с различными техническими характеристиками. Файл имеет нестандартизированный формат, что требует реализации специального механизма для его чтения. Для автоматического определения разделителя поля (запятая, точка с запятой, табуляция и др.) была разработана и применена функция read\_csv\_with\_unknown\_delimiter. Данная функция анализирует выборку файла, подсчитывает встречаемость потенциальных разделителей и загружает данные с использованием наиболее вероятного из них, что делает процесс импорта данных устойчивым к изменениям формата.

**1.2 Очистка и предобработка данных**

После успешной загрузки данных в DataFrame была проведена их очистка. Названия столбцов были стандартизированы: удалены лишние пробелы и символы переноса строк, что обеспечивает корректный доступ к данным по именам колонок. Были выявлены и обработаны пропущенные значения (NaN) в ключевых для анализа полях.

**1.3 Создание числовых признаков**

Исходные данные содержали ключевые параметры (такие как мощность, цена, время разгона) в строковом формате, часто с дополнительными символами («$», «km/h», «sec», «hp»). Для проведения количественного анализа была разработана и применена функция to\_number, которая:

Удаляет все нечисловые символы.

Заменяет запятые для корректного преобразования в float.

Обрабатывает диапазоны значений, выбирая первое число.

Преобразует результат в числовой формат.

На основе этой функции были созданы новые числовые столбцы: HorsePower\_num, Speed\_num, Accel\_num, Seats\_num, Price\_num, которые стали основой для последующего статистического анализа и визуализации.

Таким образом, рассмотренные методы и инструменты позволили преобразовать исходные «сырые» данные в очищенный, структурированный и готовый к анализу набор данных.

**2. Проверка гипотез и визуализация результатов**

**2.1. Анализ распределения гибридных и электрических автомобилей по** **странам**

**Гипотеза:** Страны с высоким уровнем авторазвития (Германия, Япония) производят больше электромобилей и гибридов, но их доля в общем объёме производства остаётся низкой.

Методика проверки: Из общего набора данных был выделен сабсет автомобилей, у которых в типе топлива встречаются подстроки «hybrid» или «electric». Для этого сабсета было построено столбчатое отображение количества моделей по странам-производителям.

Результат: Визуализация показала, что Германия и Япония действительно являются лидерами по количеству моделей гибридных и электрических автомобилей. Однако при сопоставлении с общим количеством автомобилей из этих стран становится ясно, что доля «зелёных» моделей в их портфолио всё ещё умеренна. Гипотеза подтверждена частично.

**2.2. Исследование зависимости динамики автомобиля от количества мест**

**Гипотеза:** Существует обратная зависимость между количеством мест в автомобиле и его динамическими характеристиками (временем разгона, мощностью).

**Методика проверки:** Для проверки гипотезы была построена диаграмма рассеяния, где по оси X отложено количество мест (Seats\_num), по оси Y — время разгона до 100 км/ч (Accel\_num). Для наглядности точки были окрашены и масштабированы в соответствии с мощностью автомобиля (HorsePower\_num).

**Результат:** На графике чётко прослеживается тенденция: автомобили с меньшим количеством мест (спортивные купе, суперкары) демонстрируют лучшее время разгона и, как правило, более высокую мощность. С увеличением числа мест (внедорожники, минивэны) динамика ухудшается. Гипотеза подтверждена.

**2.3. Корреляционный анализ цены и максимальной скорости**

**Гипотеза:** Цена автомобиля слабо коррелирует с его максимальной скоростью, но сильно зависит от бренда и эксклюзивности.

**Методика проверки**: Для проверки гипотезы была построена диаграмма рассеяния «Цена vs Максимальная скорость». Дополнительно был рассчитан коэффициент корреляции Пирсона между этими двумя числовыми признаками.

**Результат:** Визуально на графике не наблюдается чёткой линейной зависимости. Это подтверждается расчётным значением коэффициента корреляции Пирсона, которое составило 0.322, что указывает на слабую линейную связь. Высокая цена многих автомобилей обусловлена не столько их скоростными качествами, сколько престижем марки, уникальностью и другими факторами. Гипотеза подтверждена.

**2.4. Выявление стран-лидеров по производству скоростных автомобилей**

**Гипотеза:** Странами-лидерами по производству высокоскоростных автомобилей являются Германия, Италия и Великобритания.

**Методика проверки:** В данные был добавлен бинарный признак FastCar, отмечающий автомобили с временем разгона менее 4 секунд или максимальной скоростью более 250 км/ч. Для этого подмножества было построено столбчатое отображение по странам-производителям.

**Результат**: График наглядно демонстрирует абсолютное доминирование европейских производителей. Германия, Италия и Великобритания производят значительно больше высокопроизводительных автомобилей, чем другие страны. Гипотеза подтверждена.

**2.5. Анализ ценового распределения гибридных автомобилей**

**Гипотеза:** На рынке гибридных автомобилей наблюдается ценовой «разрыв»: представлены либо бюджетные модели, либо гиперкары, в то время как средний сегмент выражен слабо.

**Методика проверки:** Из общего датасета было отфильтровано подмножество гибридных автомобилей. Для этого подмножества было построено гистограммное распределение с отображением ядерной оценки плотности (KDE), показывающее концентрацию моделей в различных ценовых диапазонах.

**Результат:** Гистограмма демонстрирует бимодальное распределение с двумя ярко выраженными пиками: один в области низких цен (бюджетные и среднебюджетные гибриды), другой — в области экстремально высоких цен (гибридные гиперкары, такие как Ferrari SF90 Stradale, Porsche 918 Spyder). Сегмент среднеценовых гибридов заполнен существенно меньше. Гипотеза подтверждена.

**Заключение**

В ходе выполнения работы был успешно проведён комплексный анализ большого набора данных об автомобилях, что позволило проверить ряд ключевых гипотез о структуре и тенденциях мирового авторынка. Все этапы работы — от загрузки и очистки «сырых» данных до визуализации итоговых результатов — были автоматизированы с использованием современных инструментов анализа данных на Python.

Основные результаты и выводы:

Подтверждена технологическая ориентация развитых автопроизводителей: Германия и Япония лидируют по количеству моделей гибридных и электрических автомобилей, хотя их доля в общем объёме производства пока не является доминирующей.

Выявлена чёткая инженерная зависимость: Обнаружена сильная обратная связь между комфортом (количеством мест) и динамикой автомобиля (временем разгона), что отражает фундаментальный компромисс при проектировании автомобилей.

Опровергнута прямая связь цены и скорости: Установлено, что цена автомобиля слабо коррелирует с его максимальной скоростью, что подтверждает гипотезу о преобладающем влиянии на цену таких факторов, как бренд, роскошь и эксклюзивность, а не чистой производительности.

Определены центры производства высокопроизводительных автомобилей: Подтверждено, что Европа (в лице Германии, Италии и Великобритании) остаётся глобальным центром производства скоростных и высокодинамичных автомобилей.

Обнаружен структурный разрыв на рынке гибридов: Анализ выявил поляризацию рынка гибридных автомобилей, который делится на бюджетный сегмент и сегмент гиперкаров при слабой представленности среднего ценового сегмента.

Таким образом, работа продемонстрировала эффективность применения методов Big Data и инструментов Python (Pandas, Seaborn, Matplotlib) для извлечения содержательных инсайтов из сложных, многомерных данных. Разработанный аналитический пайплайн может быть легко адаптирован для работы с другими наборами данных и решения схожих аналитических задач в автомобильной и других отраслях промышленности.

**Список использованных источников и программ**

Документация Python: <https://docs.python.org/3/>

Документация Pandas: <https://pandas.pydata.org/docs/>

Документация Seaborn: <https://seaborn.pydata.org/>

Документация Matplotlib: <https://matplotlib.org/stable/contents.html>

Документация Scipy: <https://docs.scipy.org/doc/scipy/>

**Приложение**

Исходный датасет Cars.csv

Jupyter Notebook с кодом выполненного проекта: ТХБД\_курс

Репозиторий с кодом проекта: ( https://github.com/Zlexe/THBD\_-.git )