Отчет по проекту

«Анализ Dataset продажи поддержанных автомобилей»

Команда

Мирошниченко А.С; Сагайдак В. В.

1. Постановка задачи

Цель данного проекта — проанализировать данные о продажах подержанных автомобилей в России. Исследование направлено на выявление сезонных и региональных изменений продаж, а также на визуализацию результатов для более глубокого понимания рынка подержанных автомобилей.

Для достижения этой цели необходимо использовать данные о продаже автомобилей, собранные в течение нескольких месяцев. Эти данные включают различные показатели, такие как год выпуска машины, тип топлива, цвет машины, цена продажи, город продажи, марка машины и другие.

Основные шаги в проекте:

1. Сбор и предобработка данных.
2. Статистический анализ временных рядов.
3. Визуализация данных для выявления трендов и аномалий.
4. Сравнительный анализ различных марок и моделей автомобилей.

Таким образом, конечной целью проекта является полное понимание текущих трендов на рынке подержанных автомобилей выявление ключевых факторов, влияющих на ценообразование и объемы продаж.

2. Описание датасета

Датасет(https://www.kaggle.com/datasets/ekibee/car-sales-information/data) представляет собой реальный набор данных, собранный на основе продаж машин, собранных роботом, с российских сайтов по продаже машин.

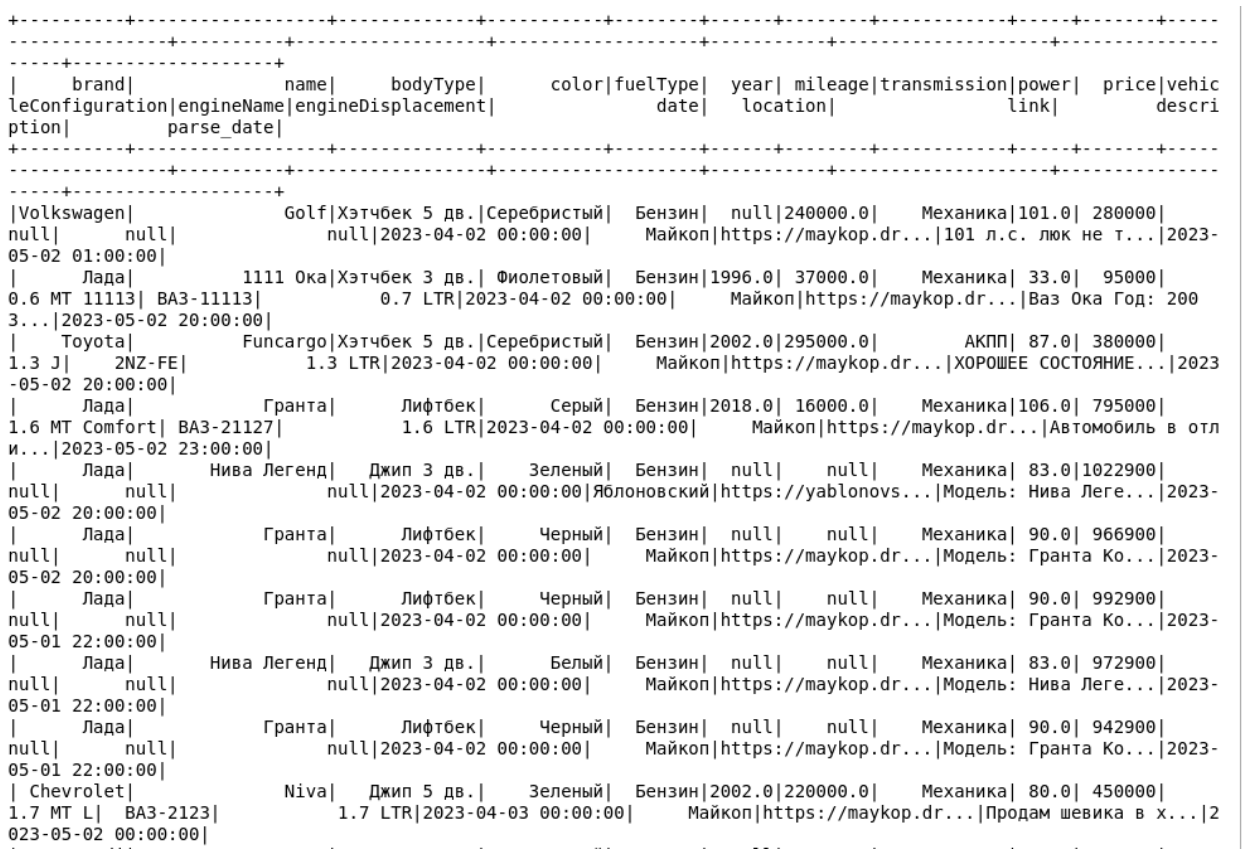
Общая информация:

В выбранном датасете количество записей: 1.294.757 . В нём хранятся данные о продаже автомобилей за период с 27.04.2023 по 21.06.2023

Данные взяты с популярных веб-сайтов для продажи автомобилей в России. Поля Dataset :

1. brand – содержит информацию о марке автомобиля
2. Name - содержит информацию о моделе автомобиля
3. Color - содержит информацию о цвете автомобиля
4. bodyType - содержит информацию о типе кузова автомобиля
5. fuelType - содержит информацию о типе топлива автомобиля
6. Year - содержит информацию о годе выпуска автомобиля
7. Mileage - содержит информацию о пробеге автомобиля
8. Transmission - содержит информацию о типе передачи автомобиля
9. Power - содержит информацию о количестве лошадиных сил
10. Price - содержит информацию о цене автомобиля в рублях
11. vehicleConfiguration - содержит информацию о конфигурации автомобиля
12. engineName - содержит информацию о названии двигателя автомобиля
13. engineDisplacement - содержит информацию о рабочем объеме двигателя автомобиля
14. Date - содержит информацию о дате продажи автомобиля
15. Location - содержит информацию о названии города в котором продали автомобиль
16. Link - содержит информацию о ссылке на продажу автомобиля
17. Description - содержит информацию об описании автомобиля
18. parse_date - содержит информацию о внесении информации об автомобиле в Dataset

Пример нескольких строк из датасета для наглядности:



```
+----------+-----------------+-------------+-----------+--------+------+--------+------------+-----+-------+-----
---------------+----------+-----------------+-------------------+-----------+--------------------+--------------
-----+-------------------+
|     brand|             name|     bodyType|      color|fuelType|  year| mileage|transmission|power|  price|vehic
leConfiguration|engineName|engineDisplacement|               date|   location|                link|         descri
ption|         parse_date|
+----------+-----------------+-------------+-----------+--------+------+--------+------------+-----+-------+-----
--------------+----------+-----------------+-------------------+-----------+--------------------+--------------
-----+-------------------+
|Volkswagen|             Golf|Хэтчбек 5 дв.|Серебристый|  Бензин|  null|240000.0|    Механика|101.0| 280000|
null|      null|             null|2023-04-02 00:00:00|     Майкоп|https://maykop.dr...|101 л.с. люк не т...|2023-
05-02 01:00:00|
|      Лада|          1111 Ока|Хэтчбек 3 дв.| Фиолетовый|  Бензин|1996.0| 37000.0|    Механика| 33.0|  95000|
0.6 MT 11113| ВАЗ-11113|          0.7 LTR|2023-04-02 00:00:00|     Майкоп|https://maykop.dr...|Ваз Ока Год: 200
3...|2023-05-02 20:00:00|
|    Toyota|          Funcargo|Хэтчбек 5 дв.|Серебристый|  Бензин|2002.0|295000.0|        АКПП| 87.0| 380000|
1.3 J|    2NZ-FE|          1.3 LTR|2023-04-02 00:00:00|     Майкоп|https://maykop.dr...|ХОРОШЕЕ СОСТОЯНИЕ...|2023
-05-02 20:00:00|
|      Лада|           Гранта|      Лифтбек|      Серый|  Бензин|2018.0| 16000.0|    Механика|106.0| 795000|
1.6 MT Comfort| ВАЗ-21127|          1.6 LTR|2023-04-02 00:00:00|     Майкоп|https://maykop.dr...|Автомобиль в отл
и...|2023-05-02 23:00:00|
|      Лада|       Нива Легенд|   Джип 3 дв.|    Зеленый|  Бензин|  null|    null|    Механика| 83.0|1022900|
null|      null|             null|2023-04-02 00:00:00|Яблоновский|https://yablonovs...|Модель: Нива Леге...|2023-
05-02 20:00:00|
|      Лада|           Гранта|      Лифтбек|     Черный|  Бензин|  null|    null|    Механика| 90.0| 966900|
null|      null|             null|2023-04-02 00:00:00|     Майкоп|https://maykop.dr...|Модель: Гранта Ко...|2023-
05-02 20:00:00|
|      Лада|           Гранта|      Лифтбек|     Черный|  Бензин|  null|    null|    Механика| 90.0| 992900|
null|      null|             null|2023-04-02 00:00:00|     Майкоп|https://maykop.dr...|Модель: Гранта Ко...|2023-
05-01 22:00:00|
|      Лада|       Нива Легенд|   Джип 3 дв.|      Белый|  Бензин|  null|    null|    Механика| 83.0| 972900|
null|      null|             null|2023-04-02 00:00:00|     Майкоп|https://maykop.dr...|Модель: Нива Леге...|2023-
05-01 22:00:00|
|      Лада|           Гранта|      Лифтбек|     Черный|  Бензин|  null|    null|    Механика| 90.0| 942900|
null|      null|             null|2023-04-02 00:00:00|     Майкоп|https://maykop.dr...|Модель: Гранта Ко...|2023-
05-01 22:00:00|
| Chevrolet|             Niva|   Джип 5 дв.|    Зеленый|  Бензин|2002.0|220000.0|    Механика| 80.0| 450000|
1.7 MT L|  ВАЗ-2123|           1.7 LTR|2023-04-03 00:00:00|     Майкоп|https://maykop.dr...|Продам шевика в х...|2
023-05-02 00:00:00|
```

3. Ход работы

3.1 Почему мы выбрали этот датасет

Мы выбрали датасет продаж подержанных автомобилей по следующим причинам:

Актуальность и практическая значимость:

Рынок подержанных автомобилей является важной частью автомобильной индустрии. Анализ данных о продажах подержанных автомобилей позволяет выявить текущие тренды и предпочтения потребителей, что может быть полезно как для автопроизводителей, так и для потенциальных покупателей.

Обширность данных:

Датасет содержит разнообразную информацию, включая марки и модели автомобилей, типы топлива, цвета, цены, даты продаж и регионы. Это позволяет провести всесторонний анализ и получить ценную информацию о различных аспектах рынка подержанных автомобилей.

Возможности для анализа:

Наличие временных данных позволяет изучить динамику изменения цен и объемов продаж. Мы можем выявить сезонные и долгосрочные тренды, а

также провести корреляционный и регрессионный анализ для определения взаимосвязей между различными признаками.

Интересные гипотезы:

Датасет позволяет проверить несколько интересных гипотез, таких как влияние времени и объема продаж на цену автомобилей, предпочтение определенных брендов и типов топлива, а также предпочтение цветов автомобилей. Проверка этих гипотез может дать полезные инсайты и рекомендации для участников рынка.

3.2 Гипотеза

Мы выдвинули следующие гипотезы на основе общего понимания рыночных тенденций и поведения потребителей на рынке подержанных автомобилей.

Во-первых, цена на автомобили увеличивается с течением времени и с количеством продаж. Исторически цены на подержанные автомобили могут расти по мере увеличения популярности моделей и спроса на них. Влияние инфляции также способствует повышению цен со временем. Рост количества продаж часто свидетельствует о повышенном спросе, что обычно ведет к увеличению цен.

Во-вторых, каждый месяц самой продаваемой маркой автомобиля является одна и та же марка. На рынке подержанных автомобилей часто наблюдается стабильное лидерство определенных брендов, пользующихся доверием и популярностью у покупателей благодаря своей надежности и стоимости владения. Например, такие бренды, как Toyota и Lada, часто занимают лидирующие позиции в продажах.

В-третьих, бензин является доминирующим видом топлива на рынке. Традиционно автомобили с бензиновыми двигателями занимают значительную долю на автомобильном рынке, включая рынок подержанных автомобилей. Это обусловлено их большей распространенностью и популярностью среди потребителей.

В-четвертых, водители не предпочитают автомобили с яркими и броскими цветами. На рынке подержанных автомобилей чаще всего популярны автомобили нейтральных цветов, таких как черный, белый, серый и серебристый, поскольку они считаются более универсальными и привлекательными для широкой аудитории. Яркие и броские цвета могут ограничить круг потенциальных покупателей.

Наконец, между средней ценой и количеством продаж в городах наблюдается сильная линейная зависимость. В более крупных и экономически развитых городах спрос на автомобили выше, что может приводить к более высоким средним ценам. Высокий спрос часто коррелирует с большими объемами продаж, что позволяет предположить наличие линейной зависимости между ценой и количеством продаж.

3.3 Предобработка данных:

Столбец с датой имел тип данных string. Для дальнейшей качественной обработки мы должны перевести в date формат и извлечь отдельно год и месяц. Также отфильтровали пустые значения в колонках типа топлива, года выпуска и пробега

```
# Преобразование столбца date в формат даты
df = df.withColumn("date", to_date(col("date")))

# Извлечение года и месяца из даты
df = df.withColumn("year", year(col("date"))).withColumn("month", month(col("date")))
```

```
df = df.filter(col("fuelType").isNotNull()) \
.filter(col("year").isNotNull()) \
.filter(col("mileage").isNotNull())
```

3.4 Формирование датафреймов

На этом этапе мы создавали датафреймы с информацией, которая в дальнейшем была использована для нахождения корреляций и визуализации тенденций и ситуаций на рынке

```
# Создание временного DataFrame для хранения топ-5 автомобилей за каждый месяц
top_5_cars_pd = pd.DataFrame(columns=["month", "brand", "count"])

# Получение списка уникальных месяцев
months = df.select("month").distinct().orderBy("month").collect()

# Поиск топ-5 автомобилей за каждый месяц
for month_row in months:
    month_num = month_row["month"]

    # Формирование подписи для диаграммы (номер месяца)
    month_label = f"{month_num}"

    # Фильтрация данных для текущего месяца
    top_5_cars_month = df.filter(col("month") == month_num) \
                         .groupBy("brand").count() \
                         .orderBy(desc("count")).limit(5) \
                         .withColumn("month", lit(month_label))

    top_5_cars_pd = top_5_cars_pd.append(top_5_cars_month.toPandas())
```

```
# Вычисление количества продаж по городам
sales_per_city = df.groupBy("location").agg(count("*").alias("sales_count"))
# Фильтрация городов с количеством продаж от 10 до 1000
filtered_cities = sales_per_city.filter((col("sales_count") > 10) & (col("sales_count") < 1000)).select("location")
# Фильтрация данных для выбранных городов
filtered_data = df.filter(col("location").isin(filtered_cities))
# Вычисление средней цены по месяцам для отфильтрованных городов
average_price_per_month = filtered_data.groupBy("year", "month").agg(avg("price").alias("average_price"))
# Преобразование в pandas DataFrame для построения графиков
average_price_per_month_pd = average_price_per_month.toPandas()
# Отсортируем данные по месяцам для правильного порядка на графике
average_price_per_month_pd['year_month'] = average_price_per_month_pd['year'].astype(str) + '-' + average_price_per
average_price_per_month_pd = average_price_per_month_pd.sort_values(by=['year_month'])
```

3.5 Нахождение корреляций и визуализация данных

На этом этапе мы анализировали полученные датафреймы и искали корреляции между признаками для подтверждения гипотез, и визуализировали полученный результат

```
correlation = df_avgprice_count.stat.corr("average_price", "city_count")
```

```
print(correlation)
```

```
0.34493273214323333
```

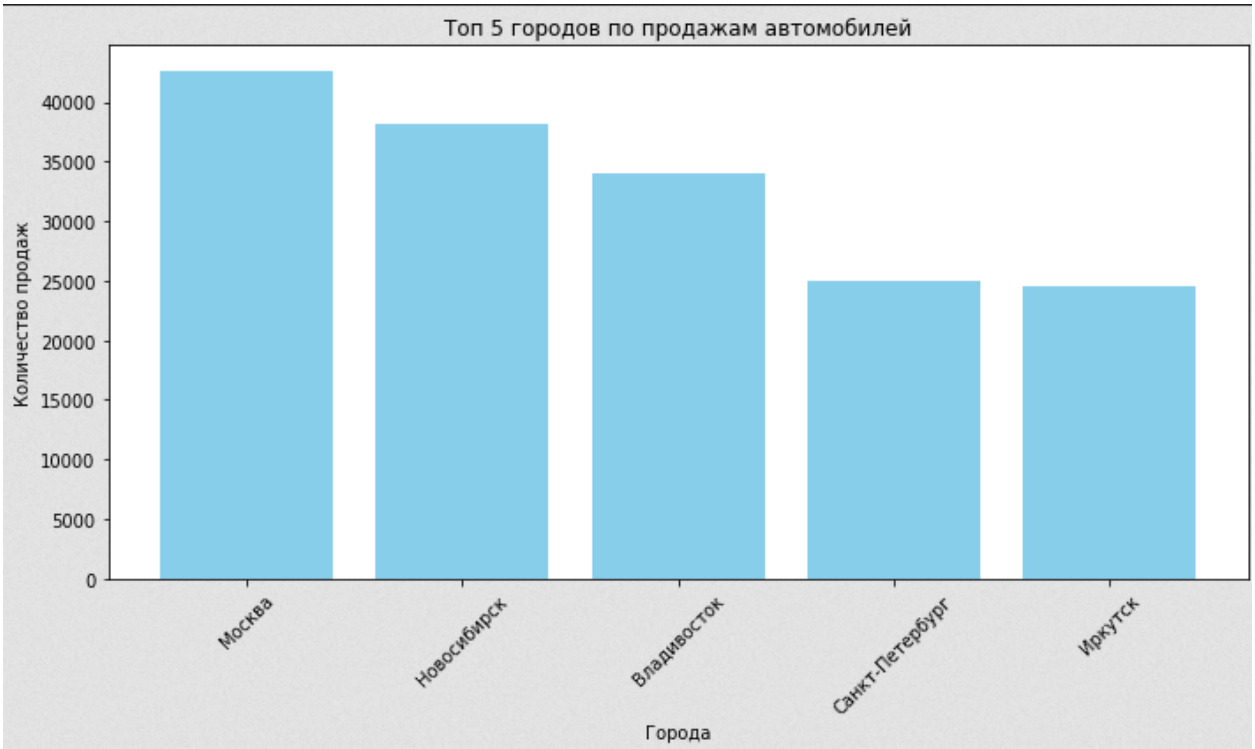
```
df_avgprice_count_mincity=df_avgprice_count.filter(df_joined["city_count"] < 1000)
correlation_mincity = df_avgprice_count_mincity.stat.corr("average_price", "city_count")
print(correlation_mincity)
```

```
0.23410412823588436
```

```
df_avgprice_count_maxcity=df_joined.filter(df_joined["city_count"] > 1000)
correlation_maxcity = df_avgprice_count_maxcity.stat.corr("average_price", "city_count")
print(correlation_maxcity)
```

```
0.369795985319722
```

```
plt.figure(figsize=(10, 6))
plt.bar(cities, counts, color='skyblue')
plt.xlabel('Города')
plt.ylabel('Количество продаж')
plt.title('Топ 5 городов по продажам автомобилей')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



4. Выводы

В ходе работы мы выбрали корректный датасет с актуальной информацией, подвергли его предобработке для упрощения работы с данными, из него сформировали датафреймы, с интересующей нас информацией, и на основе их нашли корреляцию и визуализировали информацию, касающуюся наших гипотез. В результате работы, гипотеза о том, что цена на автомобили увеличивается с течением времени и с количеством продаж оказалась верной. Второая гипотеза о том, что каждый месяц самым продаваемым брендом является одна и та же марка автомобиля, оказалась ложна. Третья гипотеза о том, что бензин является доминирующим видом топлива на рынке была подтверждена данными из датасета. Четвёртая гипотеза о том, что водители не предпочитают автомобили с яркими и бросками цветами тоже была подтверждена. И последняя гипотеза о том, что между средней ценой и количество продаж в городах сильная линейная зависимость, оказалась не точной, потому что хоть корреляция и есть, она недостаточно сильная, особенно в небольших городах.