FINAL PROJECT

Presentation by Kairbayeva Aiaulym,Berdibek Gulaiym,Zhumagul Azel

LIST OF CONTENT

- Project objectives and motivation.
- Data collection process.
- Key findings, visualizations, and insights.
- Challenges faced and how they were addressed.

PROJECT OBJECTIVE

The objective of this project is to analyze the car market in Kazakhstan using data collected from the Kolesa website. We aim to organize the data and find useful insights, such as popular car types, price ranges, and preferences in different cities. Additionally, we will create models to predict car prices and demand for certain types of cars in various regions. This will help stakeholders make better decisions based on the data.

PROJECT MOTIVATION

The project aims to understand the car market in Kazakhstan. Kolesa.kz is a popular website for buying and selling cars, and analyzing its data gives useful information about prices, car types, and locations.

KEY MOTIVATIONS

1.Understand the Market:
To see trends in car prices, types, and availability in different cities.

2.Help Buyers and Sellers:
To give clear information about fair prices and popular cars.

3.Support Businesses:
Car dealers and companies can use this data to improve their services and target the right customers.

4.Learn Data Analysis:
This project is a great way to practice web scraping and data analysis skills.

5.Expand to Other Topics:
The methods used here can also work for other websites or markets.

The project turns data into useful insights and helps improve decision-making for everyone involved.

DATA COLLECTION PROCESS

The data collection process involves web scraping from Kolesa.kz, using Python and its libraries.

The main steps and methods are as follows:

1. TOOLS AND LIBRARIES

- **requests:** Used to send HTTP requests and get the HTML content of the website.

- **BeautifulSoup:** Used for parsing and navigating the HTML structure to extract specific data elements.

- **pandas:** Used to store the extracted data in a structured format, such as a DataFrame, and save it as a CSV file.

- **random:** Used to introduce random delays between requests to mimic human-like browsing behavior.

- **time:** Used to pause the script between requests, reducing the risk of overloading the server.

DATA COLLECTION PROCESS

2. DATA PARSING PROCESS

1.Target Website:
The car listings were accessed from the main pages of Kolesa.kz, with each page containing multiple car advertisements.

2.HTML Structure Analysis:
The website's HTML structure was inspected to locate key data points like car titles, prices, descriptions, and locations. CSS selectors were used to extract these details.

3.Fetching the Data:
The requests library was used to send GET requests to the website. Random User-Agent headers were added to simulate requests from different browsers, avoiding detection and blocking.

DATA COLLECTION PROCESS

2. DATA PARSING PROCESS

4.Parsing the HTML:
Using BeautifulSoup, the fetched HTML was parsed, and specific elements (title, price) were extracted using CSS selectors and tags.

5.Storing the Data:
The extracted data was stored in a structured format using the pandas library. Each data point (title, price) was saved as a column in a DataFrame.

6.Saving as CSV:
The collected data was appended to a CSV file after each page scrape. This ensured all pages were combined into a single dataset.

DATA COLLECTION PROCESS

3. AUTOMATED PAGINATION

The scraper was designed to navigate through multiple pages (up to 100) by dynamically modifying the page number in the URL. This allowed for the collection of data from a wide range of car listings.

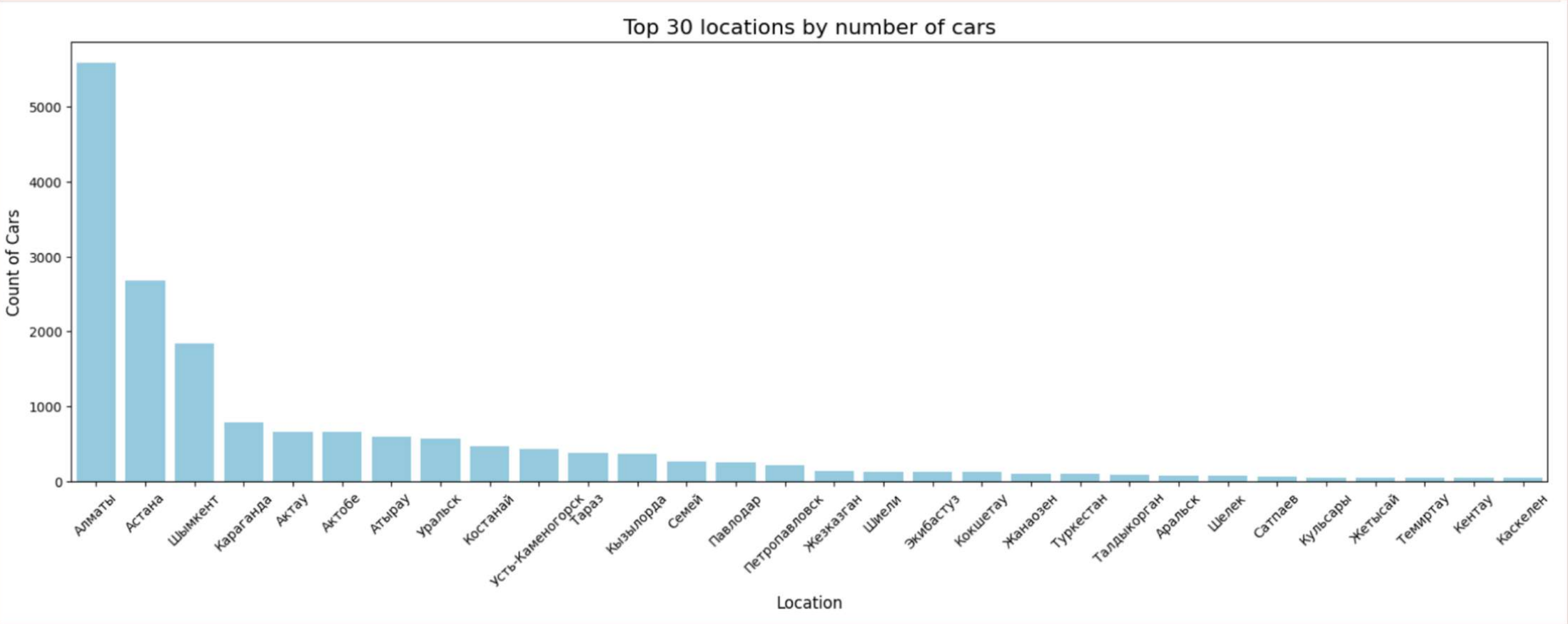
4. ETHICAL CONSIDERATIONS

Randomized delays (10–20 seconds) were introduced between requests using the time and random libraries to simulate human behavior and prevent server overload.

VISUALIZATIONS

Bar plot:

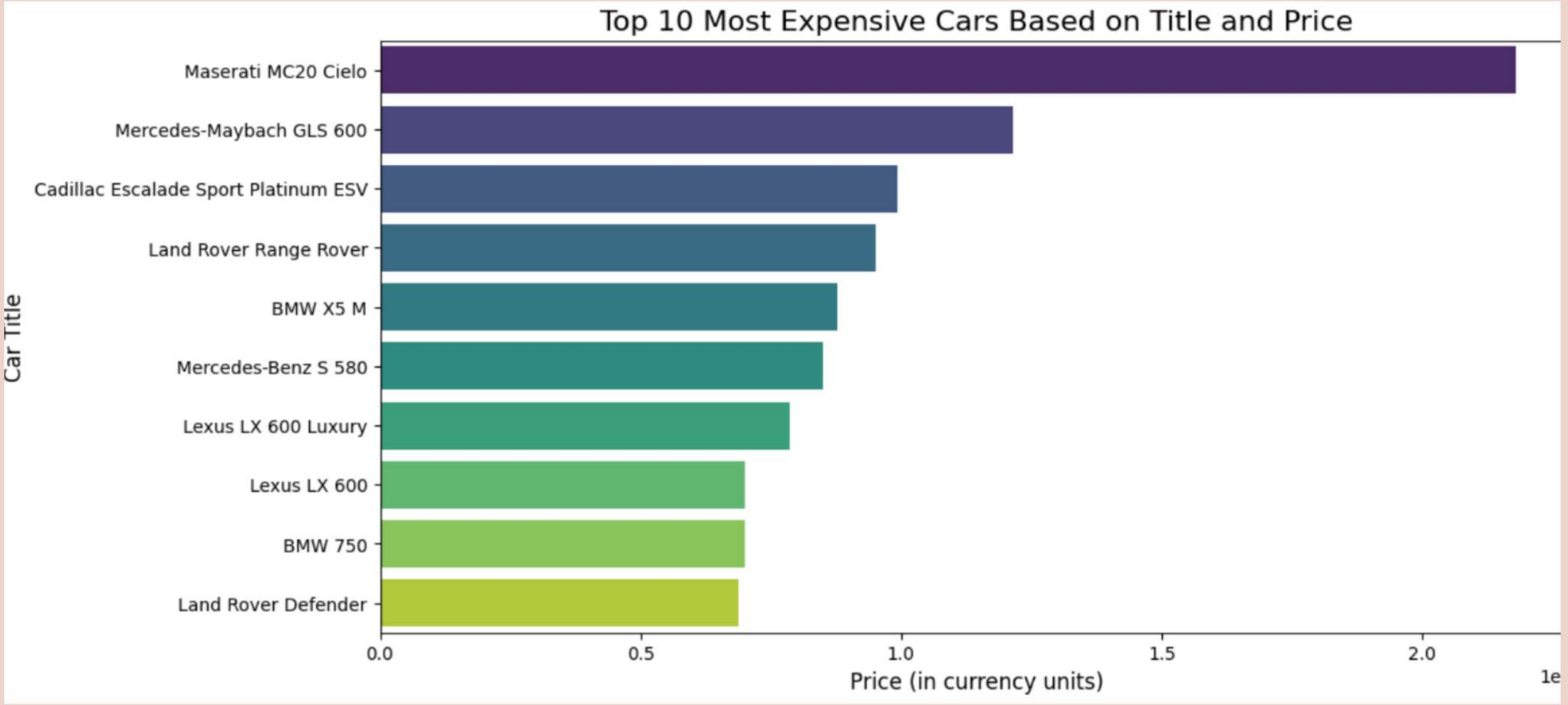
Displays the number of car listings per location, highlighting urban centers as hubs for car sales.



VISUALIZATIONS

Bar Chart of Most Expensive Cars:

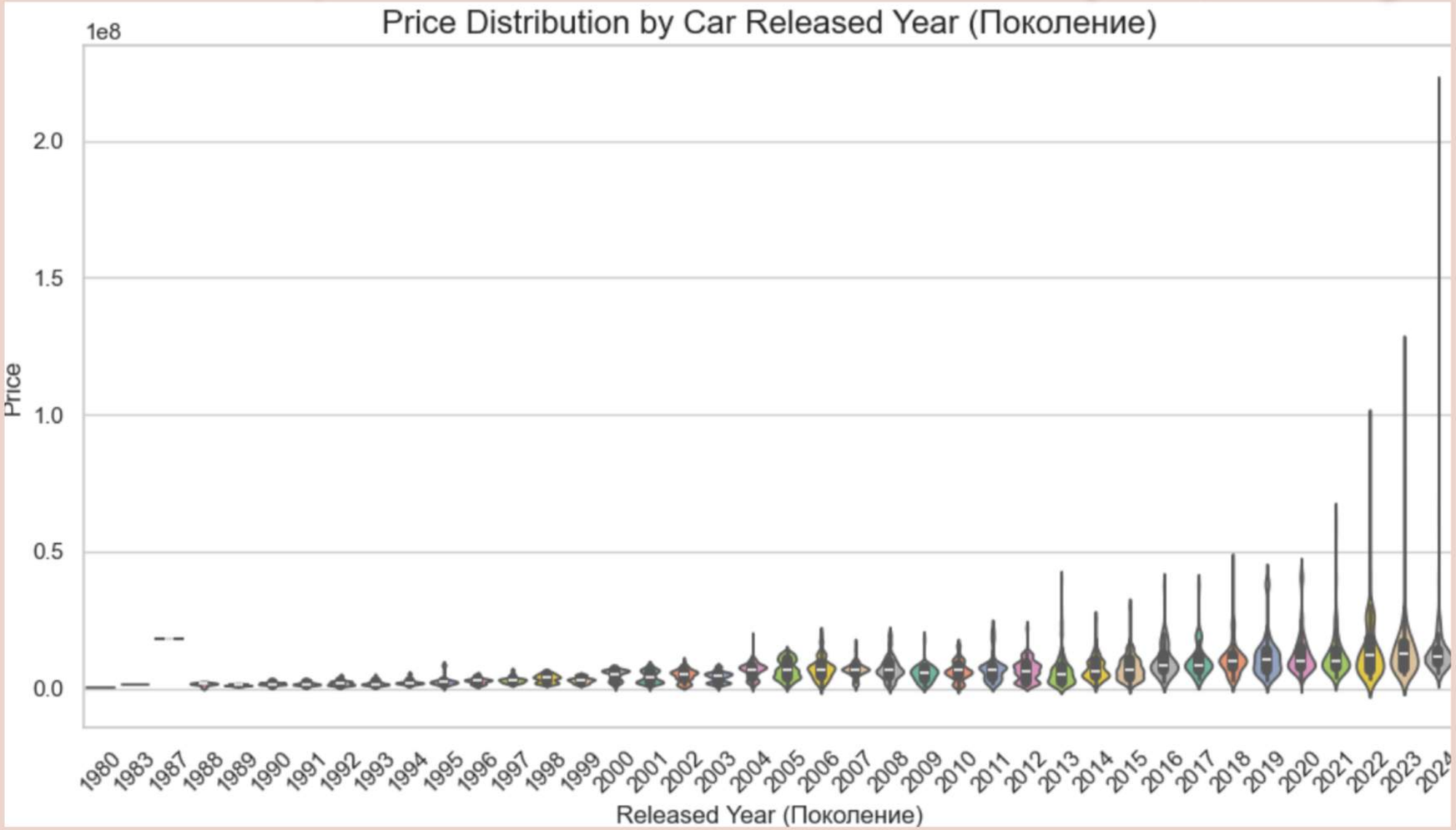
Shows the top 10 highest-priced cars, emphasizing luxury brands and rare models.



VISUALIZATIONS

Violin Plot for Price vs. Release Year:

Visualizes the price distribution across car release years, showing how age impacts valuation.



VISUALIZATIONS

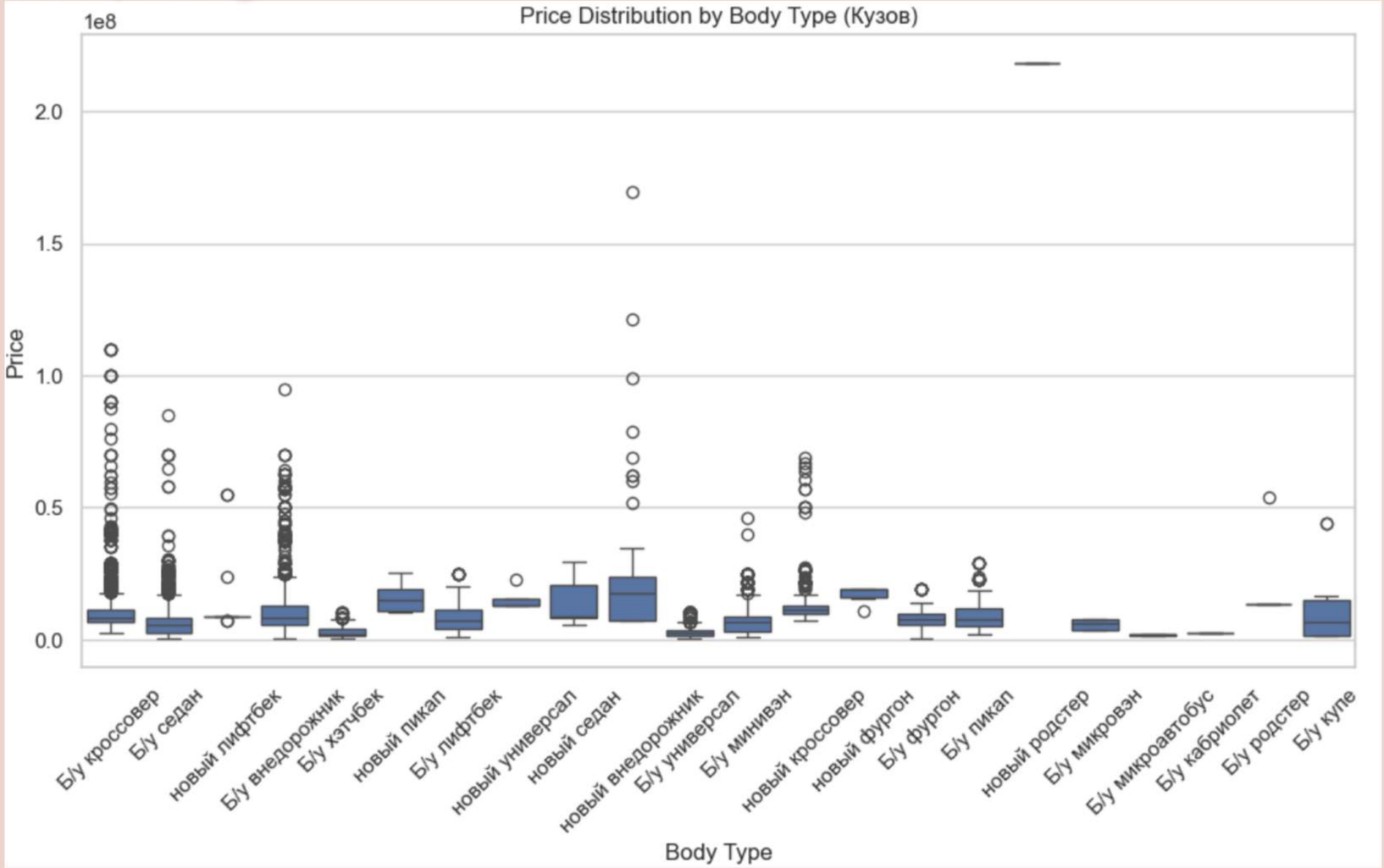
Heatmap for Correlation Analysis:

Illustrates relationships between price, engine capacity, and mileage, confirming expected trends.

VISUALIZATIONS

Boxplot for Price by Body Type:

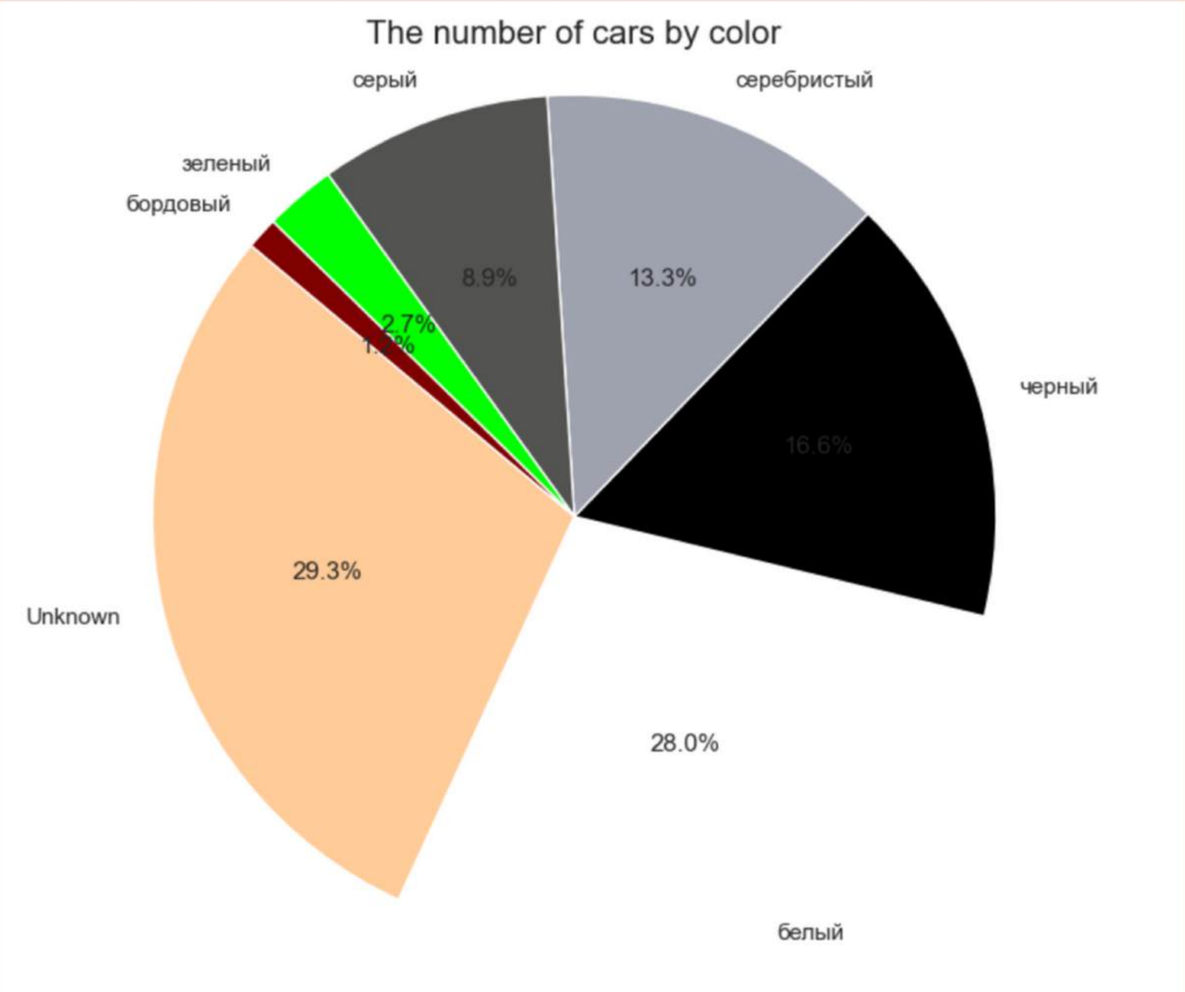
Compares the price ranges of different car body types, highlighting diversity in sedan and SUV pricing.



VISUALIZATIONS

Pie Chart for Car Colors:

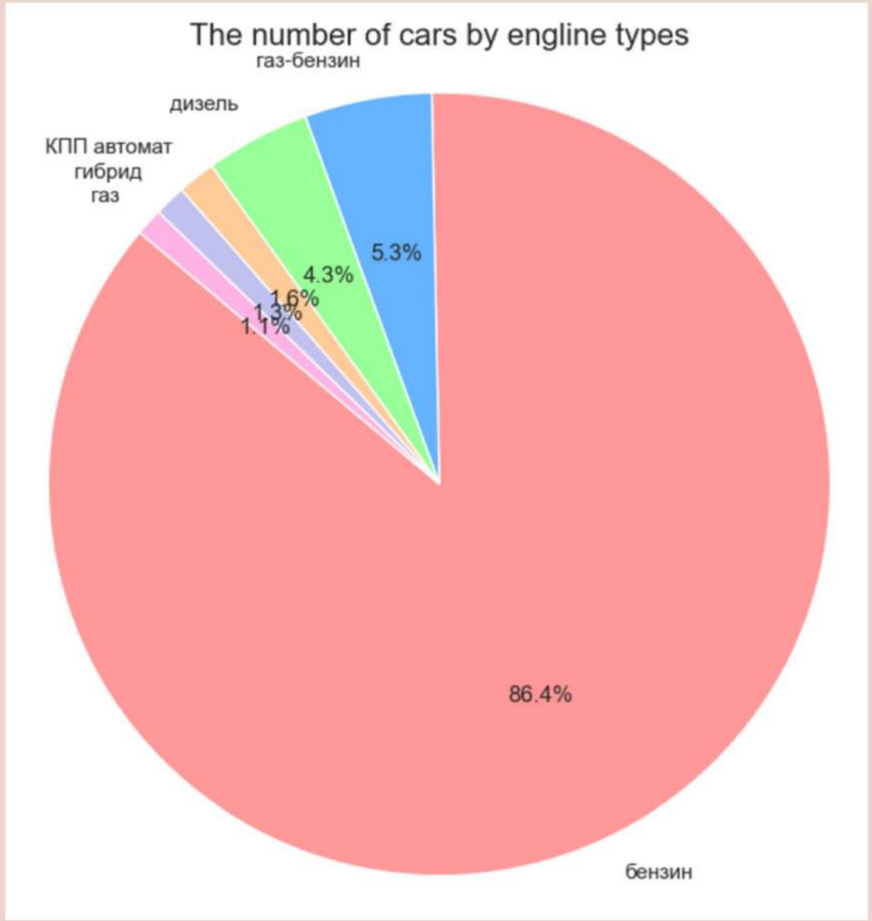
Shows the percentage of listings by color, with white and black being the most common.



VISUALIZATIONS

Pie Chart for Engine Type:

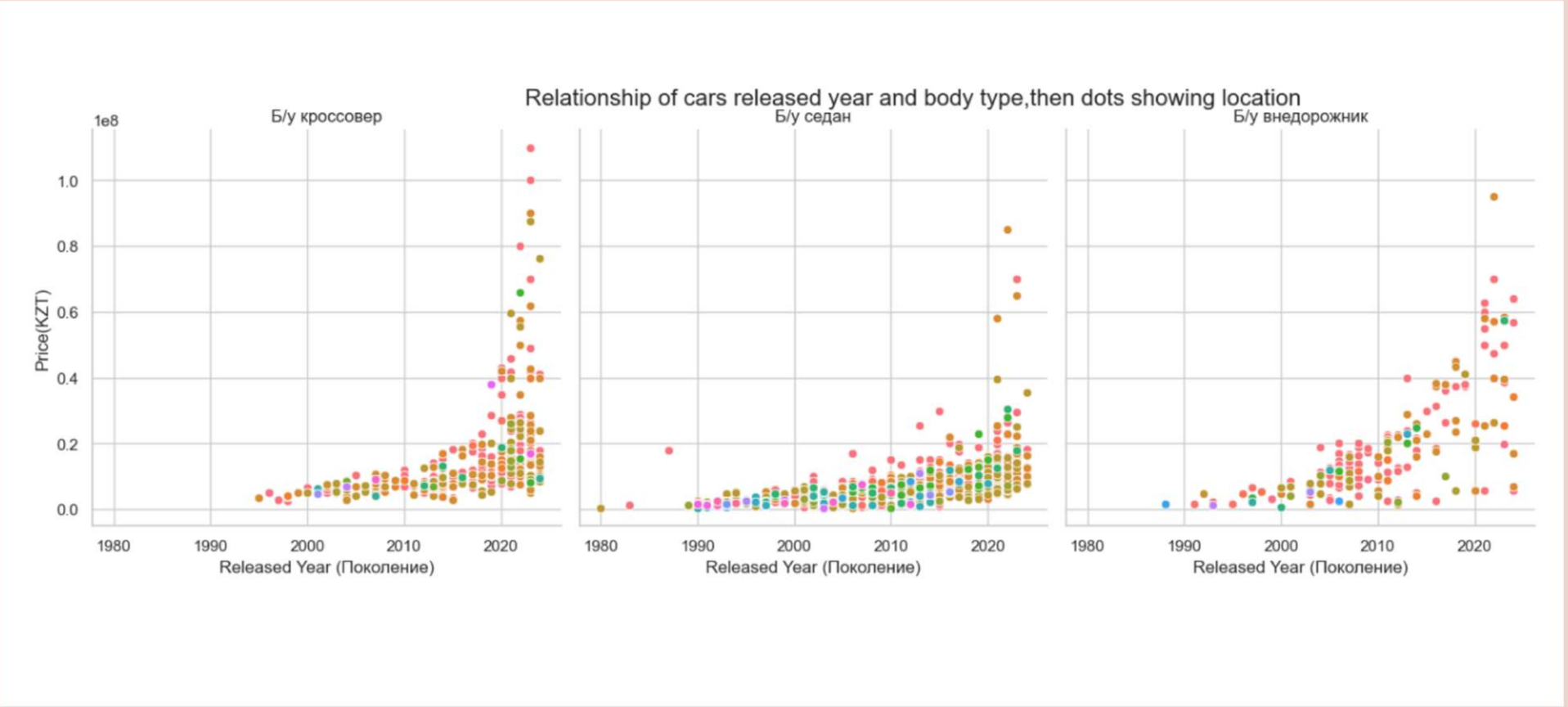
Represents the distribution of listings by engine type, with a dominance of gasoline-powered vehicles.



VISUALIZATIONS

Scatterplot of Body Types by Location and Year:

Explores the relationship between car release year, body type, and location, showing how trends differ geographically.



KEY FINDINGS

1.Top Locations for Listings:
The top 30 locations with the most car listings were identified. Urban centers like Almaty and Astana dominated the dataset, showing high activity in metropolitan areas.

2.Most Expensive Cars:
The top 10 most expensive cars included luxury brands and high-end models, showcasing the upper range of the market. These vehicles had significantly higher prices than the average listing.

3.Price Distribution by Release Year:
Prices were generally higher for newer cars.

4.Correlation Analysis:
A heatmap showed strong correlations between price, engine capacity, and mileage. Cars with larger engines typically commanded higher prices, while higher mileage resulted in lower valuations.

KEY FINDINGS

5.Price by Body Type:
SUVs and sedans had wider price ranges compared to other body types, reflecting their popularity and variety in the market.

6.Car Color Distribution:
White and black cars were the most common, followed by shades of silver and grey. Bright colors were less frequent, showing consumer preference for neutral tones.

7.Engine Types:
Gasoline engines dominated the listings, with fewer options for hybrid or electric cars, indicating the market's reliance on traditional fuel types.

8.Top Body Types by Location:
Sedans, SUVs, and hatchbacks were the most popular body types. The scatterplot showed that pricing trends for these types varied significantly by location.

INSIGHTS

1.Demand in Cities
Most listings are from big cities, showing high demand. Dealers should focus on popular models and body types in these areas.

2.Luxury Cars
Expensive cars appeal to a small audience. Promotions for high-end cars can work well in cities like Almaty.

3.Buyer Preferences
White and black cars are most popular due to practicality.
Gasoline engines dominate; hybrid and electric options are still rare.

4.Pricing Tips
Newer cars and SUVs get higher prices.
Mileage and engine size strongly affect value.

5.Growth Opportunities
Few hybrid and electric cars are available. This market could grow with more options.

CHALLENGES FACED AND HOW THEY WERE ADDRESSED

1.Website Blocking:
Problem: Kolesa blocked automated scraping.
Solution: Added random User-Agent headers to mimic human activity.

2.Slow Responses:
Problem: Some requests took too long or failed.
Solution: Set a timeout of 30 seconds to avoid waiting too long.

3.Data Issues:
Problem: Extracted data had unwanted characters and missing values.
Solution: Cleaned data with regular expressions and replaced missing values with NaN.

4.Risk of IP Blocking:
Problem: Repeated requests from the same IP risked blocking.
Solution: Added a 10–20 second delay between requests.

5.Outliers:
Problem: Extreme values skewed the analysis.
Solution: Detected and handled outliers using the Z-score method.