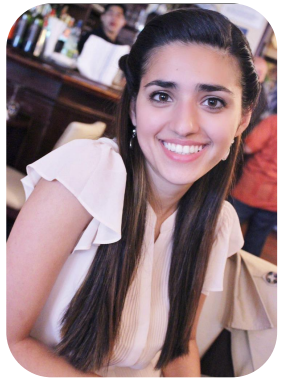
Comparative Trends in Used Car Market

Introductions:

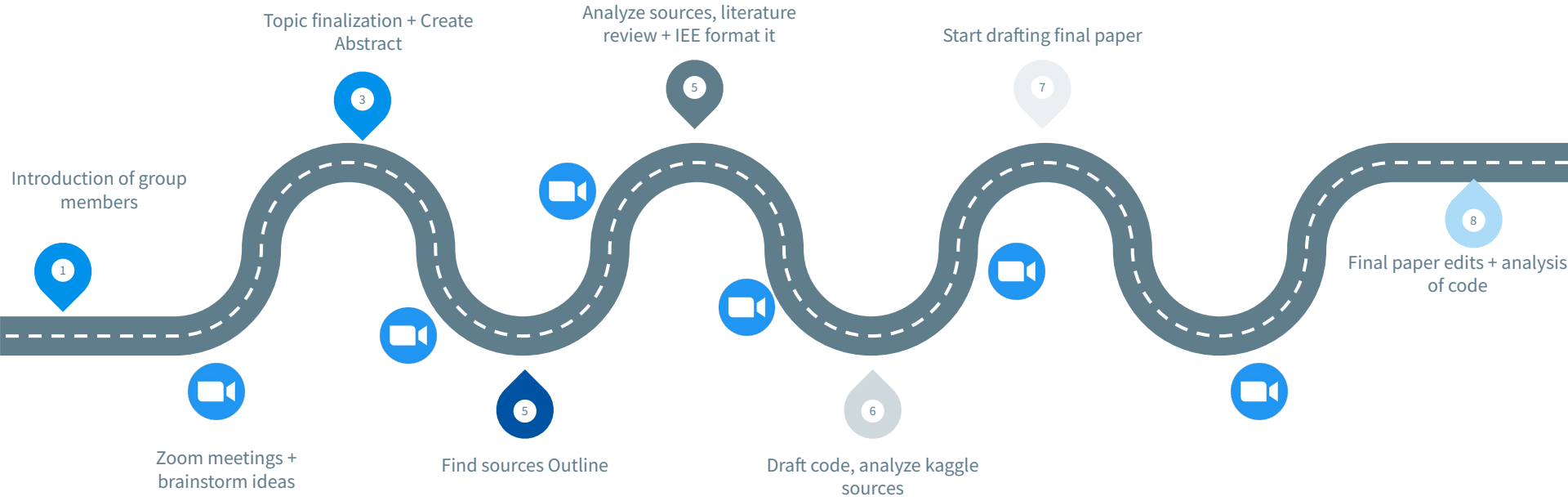
Maham Rasheda → Project Manager and Research analyst
Background: Bangladesh

Sumaiya Gulshan → Research and developer frontend
Background: Bangladesh



Gozde → Product and DevOps analyst
Background: Turkey

Roadmap


Introduction of group members
1
Zoom meetings + brainstorm ideas
Topic finalization + Create Abstract
3
Find sources Outline
5
Analyze sources, literature review + IEE format it
5
Draft code, analyze kaggle sources
6
Start drafting final paper
7
Final paper edits + analysis of code
8

Goal of the study:

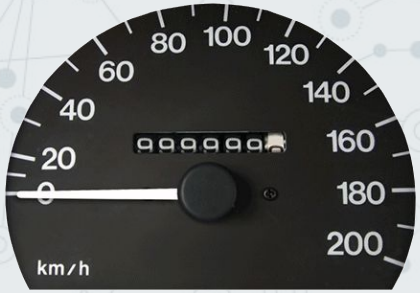
The present study compares the different car models in the used car market , specifically in the US. As a group, we will write a thorough paper with supporting evidence demonstrating the most common color, popular model, odometer readings, price ranges, and condition of a used car. Data information will be presented and cleaned from a Craigslist dataset. This paper will also go into great detail about our procedure for producing and cleaning our data.

- Since the late 1800s, used cars have been the preference of many Americans due to the reliable prices.

- Most often used cars are brought in outlets with independent car sellers or dealerships.

General Background:

- For a used car, the seller must enclose the car's age, condition, mileage, wheel drive, past owner, history, interior, and speed. It's shown that vehicles with less mileage will even sell for higher prices.



Factors to consider when buying a car:

- When examining the mileage of a used car, a buyer will typically inquire with the seller as to how many miles it has been driven.
- The used car must be in good condition for consumers to buy it or believe the cost is reasonable.
- Additionally, the model of used cars should be displayed in the event that customers have a preference.
- There are three types of car transmissions; four-wheel drive, forward wheel drive, and rear-wheel drive. With all of these key factors in mind, there is a benefit to buying a used car.

Reasons of Used Car is Brought:

- Buying a used car may be a better investment, especially for people who have never owned a car before.
- They have less depreciation
- Consumers on a monthly basis are also saving money since car insurance for a used car costs less than a new car.
- Furthermore, used cars are also more available in the car markets, as manufacturers are having difficulty producing new cars.

Methodology:

- For this project, we've decided to use the Kaggle dataset that scrapes Craigslist for car information from 2021.
- We first gathered the dataset, and then we downloaded the dataset, which was a csv file but we viewed it as an excel file.
- Then we uploaded it as a CSV file into a Jupyter notebook. Followed by importing the library's Seaborn, Pandas, and Numpys to clean, analyze, and visualize the imported dataset.

Database Implementation Process

- To begin, we wanted to become acquainted with the first five rows of data. This will allow us to see how the data looks after we import it into a Jupyter notebook. We discovered 26 columns in total, including: id, Url, region, price, model, manufacturer, number of cylinders, size, type, color, imageurl, country, state, latitude, longitude, and hosting date. As soon as we had finished our final observations, we could finish Part 1 of the project, which involved importing the vehicles.csv data into the Jupyter notebook.

- Part 2 of our project was motivated by our agreement that the data needed to be cleaned before we could better understand it.

- In order to clean explanatory variables and deal with missing data, we moved on to part 3 of our project.Once the unnecessary columns had been removed, we examined the remaining columns. Given the size of the dataset, we chose to discard the values that are not numbers (NaN) rather than replacing them.

- We proceeded with the data aggregation portion of our project after observing that there were no longer any empty rows or missing values. By looking for anomalies in the odometer readings that a used car should have, we were able to accomplish this.

Dataset Discussion

According to our research, sedans are the most popular kind of vehicle. SUVs were the second most popular vehicle, followed by trucks. Based on our data, the most popular car color was white, followed by black, and then silver. In addition, cleaning our data revealed that 4687 of the vehicles were in fair condition and 53063 were in good condition. According to our charts, the most popular car manufacturers were Ford, Chevrolet, Toyota, and Honda. And most models were found within the years 2000–2020.

Our dataset:

jupyter Comparative Trends in Used Car Market Last Checkpoint: 3 minutes ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Importing the Libraries

In [1]:
```
#First we import the library

import pandas as pd
import numpy as np
import warnings
warnings.filterwarnings("ignore")

import seaborn as sns
import matplotlib.pyplot as plt
```

Part 2: Data Cleaning

In [4]:
```
# DATA CLEANING

#drop columns that won't be used - id, url, region-url, county

vehicles_df = vehicles_df.drop(['id', 'url', 'region_url','county'], axis=1)

pd.set_option('display.max_columns', 500)

# drop rows with more than 19/22 columns missing.

vehicles_df = vehicles_df[vehicles_df.isnull().sum(axis=1) < 19]
```

In [6]:
```
vehicles_df.isnull().sum() #look at each column that is useful but has many null values.
#There are no null region values. we will use that as a base for cleaning and replacing missing/wrong odometer and 0 pri
```

Out[6]:

region	0
price	0
year	1137
manufacturer	17578
model	5209
condition	174036
cylinders	177610
fuel	2945
odometer	4332
title_status	8174
transmission	2488
VIN	160974
drive	130499
size	306293
type	92790
paint_color	130135
image_url	0
description	2
state	0
lat	6481
long	6481
posting_date	0

dtype: int64

In [7]:
```
vehicles_df.describe(include = 'all') #see the composition of each column
```

Out[7]:

	region	price	year	manufacturer	model	condition	cylinders	fuel	odometer	title_status	transmission
count	426812	4.268120e+05	425675.000000	409254	421603	252776	249202	423867	4.224800e+05	418638	424324
unique	404	NaN	NaN	42	29649	6	8	5	NaN	6	3
top	columbus	NaN	NaN	ford	f-150	good	6 cylinders	gas	NaN	clean	automatic
freq	3608	NaN	NaN	70985	8009	121456	94169	356209	NaN	405117	336524
mean	NaN	7.520976e+04	2011.235191	NaN	NaN	NaN	NaN	NaN	9.804333e+04	NaN	NaN
std	NaN	1.218325e+07	9.452120	NaN	NaN	NaN	NaN	NaN	2.138816e+05	NaN	NaN
min	NaN	0.000000e+00	1900.000000	NaN	NaN	NaN	NaN	NaN	0.000000e+00	NaN	NaN
25%	NaN	5.900000e+03	2008.000000	NaN	NaN	NaN	NaN	NaN	5.770400e+04	NaN	NaN
50%	NaN	1.395000e+04	2013.000000	NaN	NaN	NaN	NaN	NaN	8.554800e+04	NaN	NaN
75%	NaN	2.648900e+04	2017.000000	NaN	NaN	NaN	NaN	NaN	1.335425e+05	NaN	NaN
max	NaN	3.736929e+09	2022.000000	NaN	NaN	NaN	NaN	NaN	1.000000e+07	NaN	NaN

In [8]:
```
#Drop VIN, description,image URL, lat, long, posting date columns
vehicles_df.drop(['VIN', 'image_url', 'lat', 'long', 'posting_date'],axis=1, inplace=True)
```

In [20]:
```
#These are some error values that have been seen in the past
odometer_error_value_list = [1111,11111,111111,1234,12345,123456]
```

```
for i in odometer_error_value_list:
  if i in vehicles_df.odometer:
    print(i, True)
```

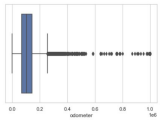
```
1111 True
11111 True
111111 True
1234 True
12345 True
123456 True
```

Part 4B: Data Visualization

In [39]:
```
#Check for outlier values
```

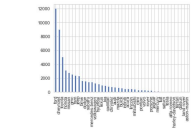
In [40]:
```
#Let's check for odometer outliers
sns.set_theme(style="whitegrid")
sns.boxplot(x=new_updated['odometer'])
```

Out[40]: <AxesSubplot:xlabel='odometer'>



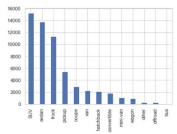
In [43]:
```
new_updated['manufacturer'].value_counts().plot(kind='bar')
#Here we can see Ford, Chevrolet and Toyota brands sold the most.
```

Out[43]: <AxesSubplot:>



In [44]:
```
new_updated['type'].value_counts().plot(kind='bar') #this shows the most sold car make was sedans followed by SUVs and t
```

Out[44]: <AxesSubplot:>



Bottlenecks:

- After much inspection, as a team we couldn’t web scrape craigslist directly because they blocked the developers/bots/third parties from extracting their data.
- Also, since we have a large dataset and tried to clean it up, we had trouble figuring out whether the code was error-free because the Jupitner notebook wouldn't produce the output as efficiently.
- Furthermore, our technological devices crashed a few times due to the data being really big.

Highlights

- Being able to analyze a large dataset and compile a smaller dataset to what we needed
- Having a collaborative space
- Since we had a big data we relied on python that enabled us to learn more about cleaning data in python
- We developed more research skills and gained more analytic and technical skills
- Working together helped highlight our individual skills

Non-Highlights

- Project was challenging due to the big dataset
- We didn’t have more accurate seasonal or yearly sales dataset to explain effects of used car market in pandemic
- Jupyter notebook crashing due to large data set led us to be delayed in analyzing results
- We struggled to work together on google collab as it didn’t have live time updates, resulting to work together on zoom through one person's jupyter notebook
- The large csv file wouldn’t open onto excel
- We lost a group member, resulted in us to do double the work
- Getting sick and losing momentum
- But we coded so a win is a win!

Future Plans + Recommendations

For future work, we will be doing more regression and predictive analysis and using data mining to check the accuracy of the results since our data had so many inaccuracies.

References:

[1]P. Jones, “Do A New Car’s Odometer Say Exactly ZERO Miles? (Checked).” https://motorandwheels.com/new-cars-say-exactly-zero-miles-on-odometer/ (accessed Dec. 04, 2022).

[2]“When Will Car Prices Drop? | J.P. Morgan Research,” *www.jpmorgan.com*. https://www.jpmorgan.com/insights/research/when-will-car-prices-drop

[3]“Buying a Car During COVID-19 Pandemic | Equifax - United States - Evo Prod,” *United States*.

[4]A. Gavazza, A. Lizzeri, and N. Roketskiy, “A Quantitative Analysis of the Used-Car Market,” *The American Economic Review*, vol. 104, no. 11, pp. 3668–3700, 2014, Accessed: Dec. 04, 2022. [Online]. Available: http://www.jstor.org/stable/43495350

[5]J. Neuburger, “Ending Data Scraping Dispute, Craigslist Reaches $31M Settlement with Instamotor,” *New Media and Technology Law Blog*, Aug. 24, 2017.

[6]I. A. Wains, “Exploring and Analyzing Used Car Data Set,” *The Startup*, Dec. 03, 2020. https://medium.com/swlh/exploring-and-analyzing-used-car-data-set-2e2bf1f24d52

Datasets:

[1]“Used Cars Dataset,” *www.kaggle.com*. https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data

[2]T. Shin, “A Machine Learning Project — Predicting Used Car Prices,” *Medium*, May 06, 2020. https://towardsdatascience.com/a-machine-learning-project-predicting-used-car-prices-efbc4d2a4998 (accessed Dec. 02, 2022).

[3]Elladuke, “Elladuke/SCAMP-FINAL-PROJECT-INDI,” *GitHub*, May 20, 2021. https://github.com/Elladuke/SCAMP-FINAL-PROJECT-INDI/blob/master/new_vehicle.csv (accessed Dec. 02, 2022).


Thaaaaaanks!