# Overview / Reasoning

* This was a borrowed idea from websites to analyse the second-hand car market.

# Project Objectives

1. Explore and analyse the second-hand car market.
2. Predicting the price of second-hand cars:
   1. By City
   2. By State
   3. Overall Prediction

# Project Workflow

* Refer File ‘***Proposal\_and\_Plan.xlsx***’

# Data Sources

* <https://www.carwale.com/used/>
* Datasets are stored [here](https://drive.google.com/drive/folders/1mZXTZtJbI3IXK7gCxKuQrrCRJ_6vTYGP?usp=sharing) due to size constraints. Both Raw and Processed Data accounts to around **6 GB**.

# Data Process, Storage and Management

A close-up of a diagram

Description automatically generated

# Glossary of Terms

* NA

# Tools/ Resources in Play

* PYTHON
* MS EXCEL

# Challenges of the Project

* Scrapping Data – took 2 continuous days to extract the complete data.
  + Challenges were, power-cuts, internet disruptions, requests being denied.
  + 2 laptops were used to extract the data in batches.
* Cleaning data

# Data Collection

* Required Data is extracted from <https://www.carwale.com/used/>.
* Extraction is done in 2 parts:
  + Extraction of data for all the cities - named as “**Dataset A**”.
  + Extraction of data for all cities with major focus on - CHENNAI, HYDERABAD, BANGALORE, DELHI, MUMBAI – named as “**Dataset B**”.
* Due to the huge volume of data, extraction is done in 2 parts.

# Pre-processing

* Main Scope of pre-processing is to clean and format data for EDA.
* Pre-processing is done in 3 stages as below:
  + STAGE-I Dataset A and Dataset B
    - Libraries Used - ***PANDAS***
    - Preliminary data checkup and consolidation of chucks into a single file for Dataset A and Dataset B
  + STAGE-II Dataset A and Dataset B
    - Libraries Used – ***PANDAS*** for Dataset A
    - Libraries Used – ***POLARS*** for Dataset B
  + STAGE-III Dataset-A and Dataset-B
    - Libraries Used – ***POLARS***
    - Combining Datasets, A and B
    - Final cleaning and formatting
    - Storing the data into a single file.

# EDA - Analysis - II

* The most no. of cars available for sales in the second hand market is of brand **Maruti Suzuki**, followed by **Hyundai** and then **Honda**
* Other brands contribute to around 60% of the second hand market including high end cars.
* **There are highest no. of cars being sold were made in 2018.**
* **This might give us a hint that may be every 5 years, most people sell their old car to buy new ones.**
* **But that might not be the case as we do not have enought evidence.**
* Cars listed under Maharashtra is the highest compared to other states, followed by Karnataka, Delhi, Uttar Pradesh and Tamil Nadu.
* Around 22% of the total Cars listed are from Maharashtra.
* Mumbai city has the highest no. of cars listed
* **Most of the cars listed for selling are Petrol cars and have single owners.**
* **Maximum cars listed are of Manual Transmission.**
* **Many of the cars being sold are non-commercial registered vehicles.**
* Average Engine Size of the cars list is 1364 cc. Since the distribution is right skewed, Median is used as a measure of Average.
* Most cars listed for sales have a moderate engine size.
* Whereas the average bhp (Max Power) is 89 bhp, which is considered as low.
* Many cars being sold have low bhp.
* We can expect the average mileage of 18.9 kmpl with a deviation of around (+ or - 5) kmpl.
* It can be seen that the distribution of Price is heavily skewed to the right (positive).
* Huge no. of outliers could be seen and signifies that there are cars with very high prices.
* The average price of the cars is Rs. 625000 as we treat cars with value 50 lakhs and above as outliers.
* Below is the breakdown of the distribution of prices omitting the outliers.
* Most of the cars listed fall in the range 3.9 lakhs to 10 Lakhs.
* The box plot suggests that very few no. of cars fall beyond te 50 lakhs category.
* Most of the cars beyond 50 Lakh rupees are luxury brands.
* Aston Martin Vantage V8 F1 Edition is priced at the highest, whereas Fiat Palio 1.2 EL is priced at the lowest.
* Maharashtra has a wide range of Car Brands to select from followed by Tamil Nadu, Karnataka and Kerala.

# Prediction

* Analysis – II – Overall Prediction
  + After trying different regression models, RIDGE REGRESSION was the best model for prediction of prices.
* Analysis – III – Prediction (cars made after 2015)
  + After trying different regression models, RIDGE REGRESSION was the best model for prediction of prices.

# Execution

* Present Findings as a Concluding Table with insights and predictions.

# References

-NA-