# Prediction of Used Car Price

By Song Gao, Navdeep Rana, Yutong Duan and Xuzhi Tang

# Goal of project and Stakeholders

- **Goal:** The goal of this project is to use various machine learning models to predict the price of a used car based on its given features.
- **Stakeholders:**
  - Used Car Dealers
  - Car Owners Looking to Sell
  - Online Car Marketplace Platforms
  - Insurance Companies
  - Bank and Financial Institutions

# Dataset and Data Description

- We have used the “Used Car Prediction Dataset" obtained from the kaggle websitehttps://www.kaggle.com/datasets/taeefnajib/used-car-price-prediction-dataset. This data is publicly available for use under the CC BY 4.0 license.

- There are 4009 instances where each row represents a unique vehicle listing with the following features: brand & model, model year, mileage, fuel type, engine, transmission, exterior and interior colors, accident history, clean title and price.

# Data Cleaning

- Convert the numerical feature to correct format: model year, mileage, and price.
- Extract information from engine column by keywords regular expression search.
- Drop the electric cars entries, due to lack of instances.
- Investigate the missing values
    - By exploratory data analysis
    - By searching model information on Google
    - Drop instances with many missing values

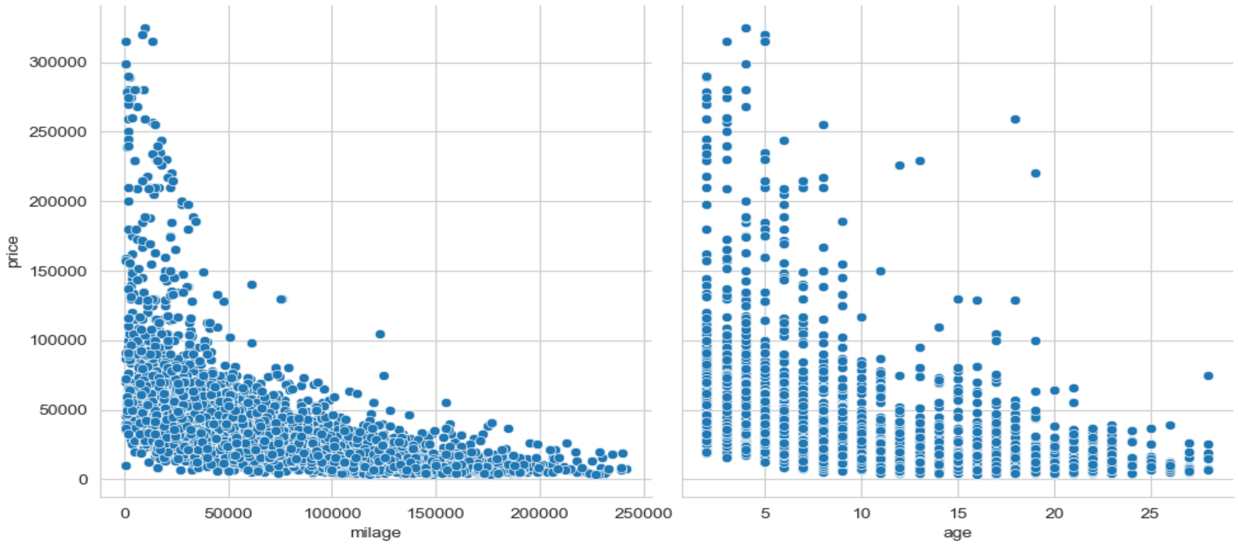
# Feature Engineering

We want to add the following features:

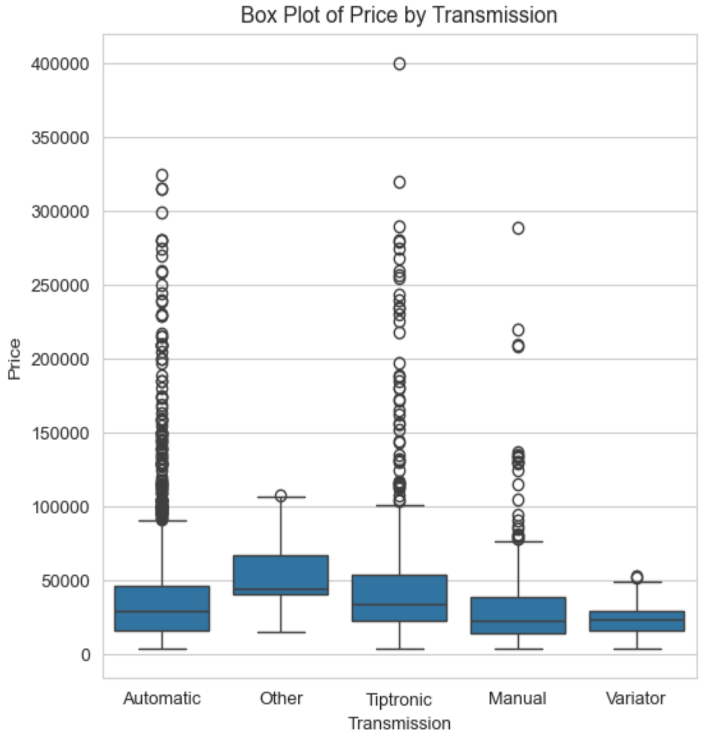
- Logarithm of price presents a better distribution than the original price.
- Categorize brand & model into Luxury and Economy.
- One-hot encoding the categorical features.
    - Clean title, accident history, colors, fuel type, transmission type, etc.
- Create various combinations of numerical features. For example: mileage per year, square root of mileage, average mileage by age groups, etc.

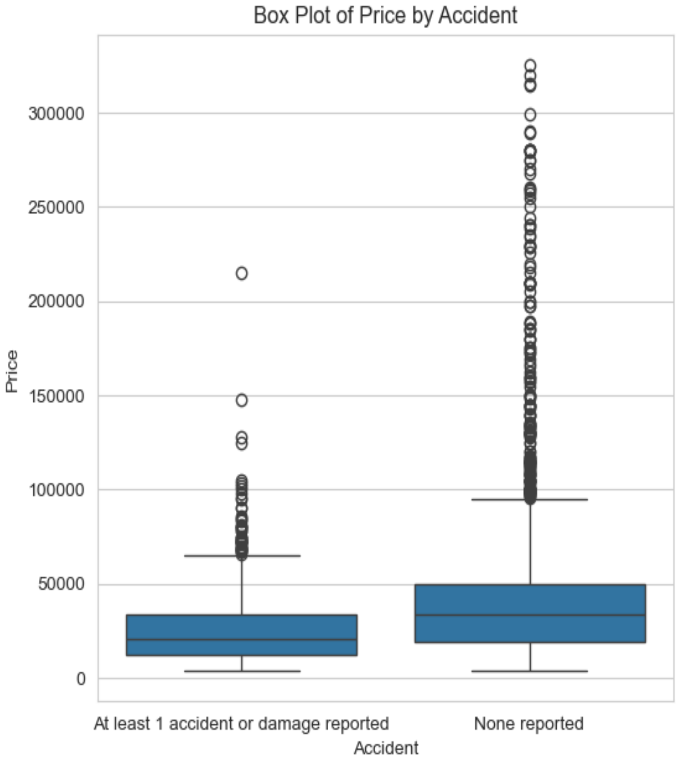
# Exploratory Analysis

Relation of price with mileage and Age of Vehicle:



# Effect of Type of Transmission and Accident history on Price:



## Data Pipeline

Some models require comparable data scales (e.g.  KNN), we have used standard scaler on the dataset before applying regression techniques.

## Proposed Models

- Linear Regression
- Polynomial Regression
- K-Nearest Neighbors (KNN)
- Random Forest Regressor
- XGBoost Regressor

# Models performance

| Model | MAE | RMSE | $R^2$ | MAPE | SMAPE |
|---|---|---|---|---|---|
| Linear Regression | $10389 | $19366 | 0.77 | 27.0% | 25.9% |
| Polynomial Regression | $7795 | $14418 | 0.85 | 21.8% | 21.2% |
| KNN Regression | $9006 | $18007 | 0.84 | 24.9% | 22.9% |
| Random Forest | $8407 | $16058 | 0.87 | 21.6% | 20.5% |
| XGBoost | $6945 | $11254 | 0.90 | 22.2% | 20.6% |

# Further Research

- Creating interaction terms, using polynomial features --Dive deeper into feature engineering.
- Introducing advanced time-series methods.
- Try other machine learning models -- testing gradient boosting alternatives.

# Thank You!

We would like to thank the Erdõs Institute for hosting the Data Science Bootcamp. In particular, we appreciate  lectures and support from  Steven Gubkin and Alec Clott.