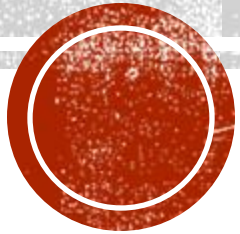
SHIFT CAR ACQUISITION PRICE MODELING



SUHAS CHOWDARY J

BACKGROUND

- Shift is a leading end-to-end auto ecommerce platform transforming the used car industry with a technology-driven, hassle-free customer experience.
- There are three stages in a shift's vehicle journey:
  - Acquisition - Shift pays a price to acquire the vehicle from the user
  - Listing - Shift lists the vehicle on the website to sell
  - Final sale - Shift finally sell the vehicle to customer
- Shift is looking for a modeling approach for acquisition price - the price at which the car is acquired from the user.

TABLE OF CONTENTS

1. GOAL

- Build a **data science/ML model** to predict the **acquisition price** of the car.

- If the selling price was greater than acquisition price, SHIFT makes money on the cars.
- If the selling price was less than acquisition price, SHIFT ends up losing money.
- Ensure the profits are maximized and Losses are decreased.

2. EXPLORATORY DATA ANALYSIS

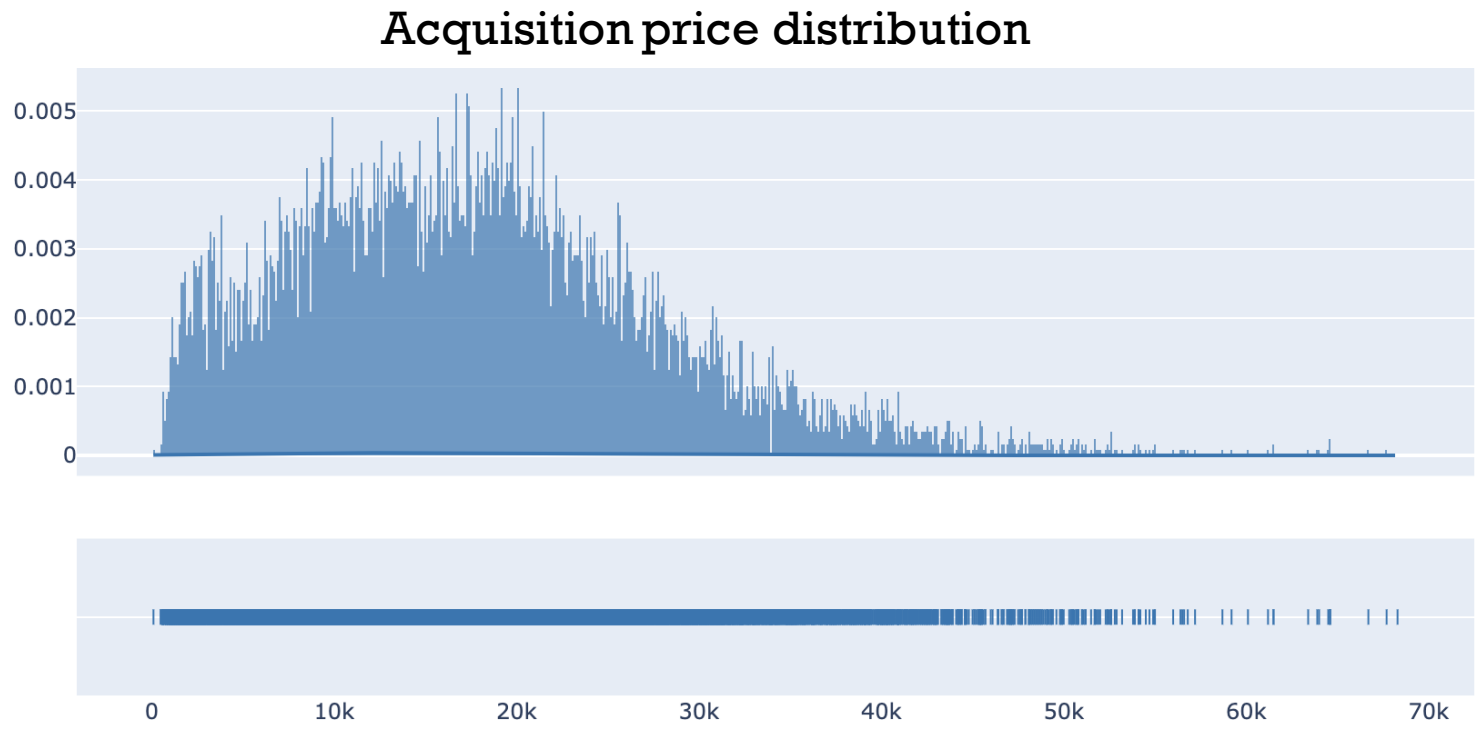
Columns overview

1. Seller id - unique identifier for a vehicle year - vehicle year
2. Make - vehicle make
3. Model - vehicle model
4. Trim - vehicle trim
5. Body type - vehicle body type
6. Engine type - vehicle engine type
7. Mileage - mileage on the vehicle at the time of acquiring it
8. Accidents - number of accidents on vehicle
9. Owners - number of owners for this vehicle
10. Seller region - which region vehicle was acquired in
11. Appraised date - date of acquisition
12. Acquistion price - price at which the vehicle was acquired for
13. Listing date - date when vehicle was listed on our shift website
14. List price - price at which the vehicle was listed for sale on our website
15. Sell date - date when vehicle was sold
16. Final sale price - price at which the vehicle was finally sold

2. EXPLORATORY DATA ANALYSIS

b. Number of rows and acquisition price distribution

- Total number of rows= 639,401
- Number of unique cars = **12,000**

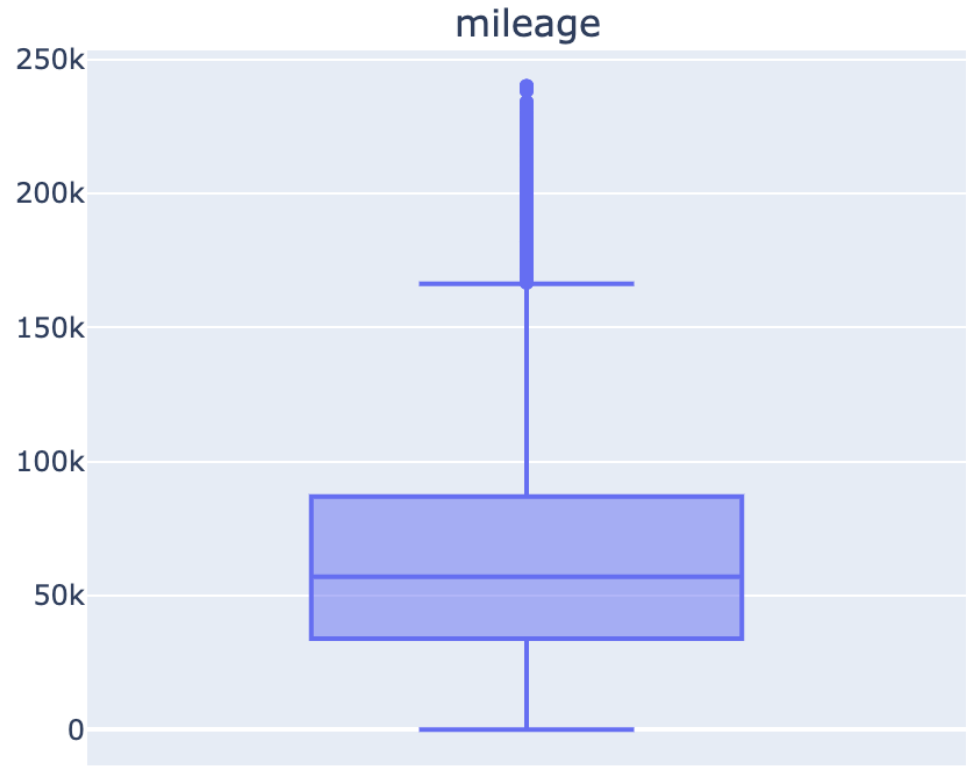


Average acquisition price of the car is ~ 17,805 dollars

2. EXPLORATORY DATA ANALYSIS

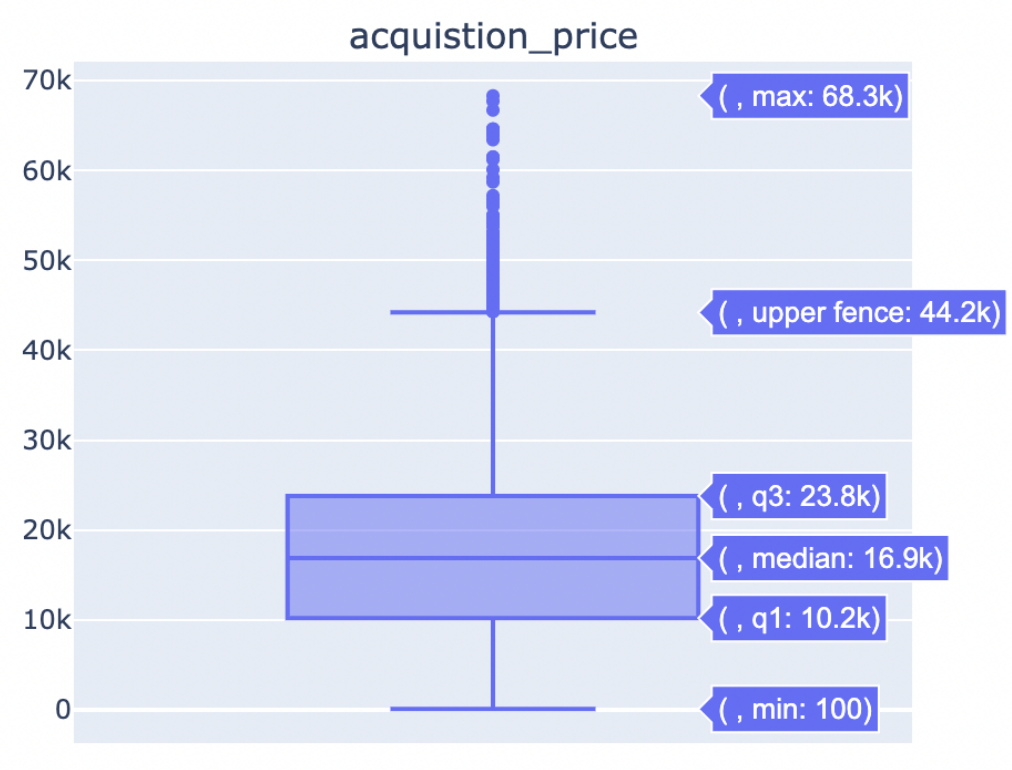
c. Box plots of numerical columns

Average mileage per car at shift is ~66k miles



2. EXPLORATORY DATA ANALYSIS

c. Box plots of numerical columns

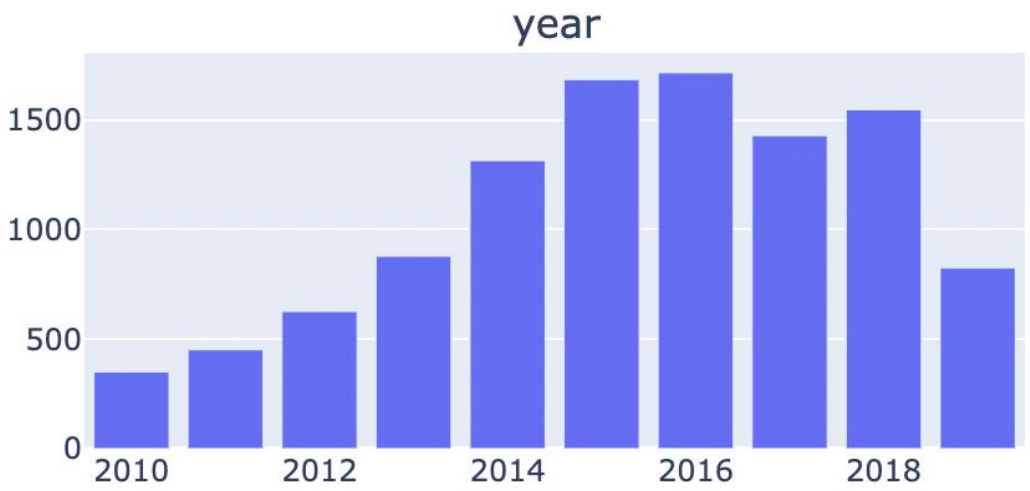


Mean acquisition price = 17.8k

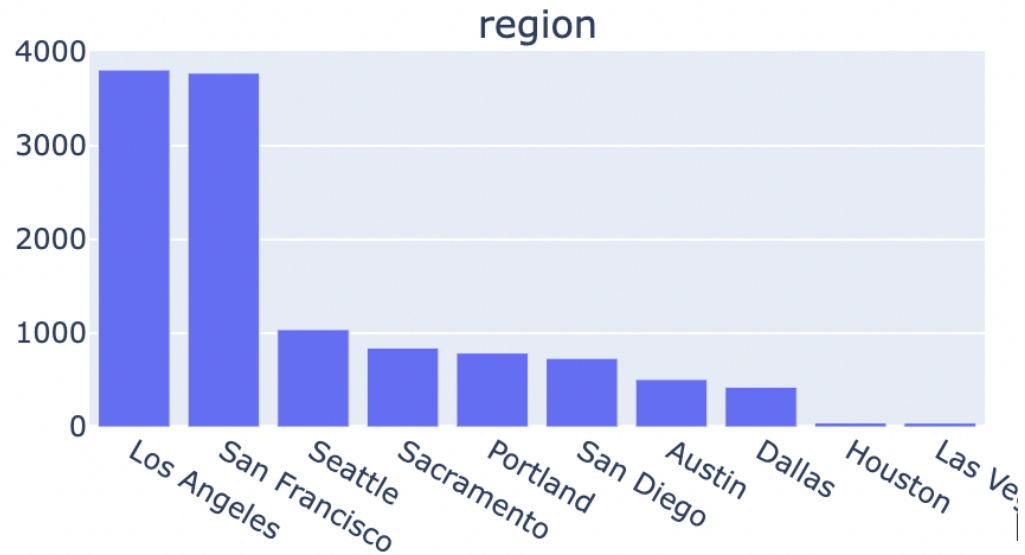
Mean sale price= 19k

2. EXPLORATORY DATA ANALYSIS

d. Year of manufacture and region



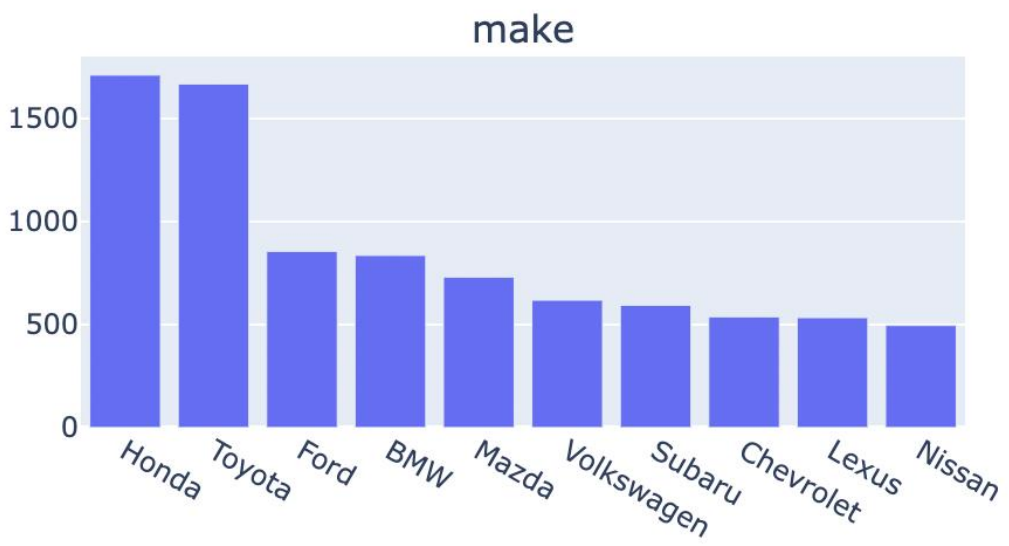
Majority of cars sold at shift are Manufactured between 2015 and 2018.



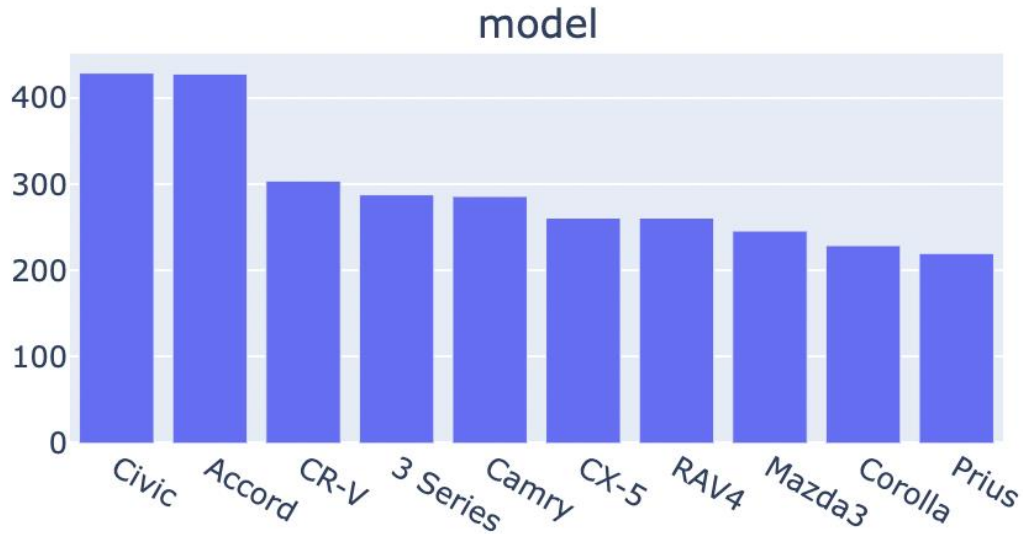
Major regions of business for shift are Los Angeles, SFO and Seattle

2. EXPLORATORY DATA ANALYSIS

e. Most frequently sold makes and models



'Honda' cars are most sold cars at Shift followed by Toyota.



Honda Civic and accord are the leading models.

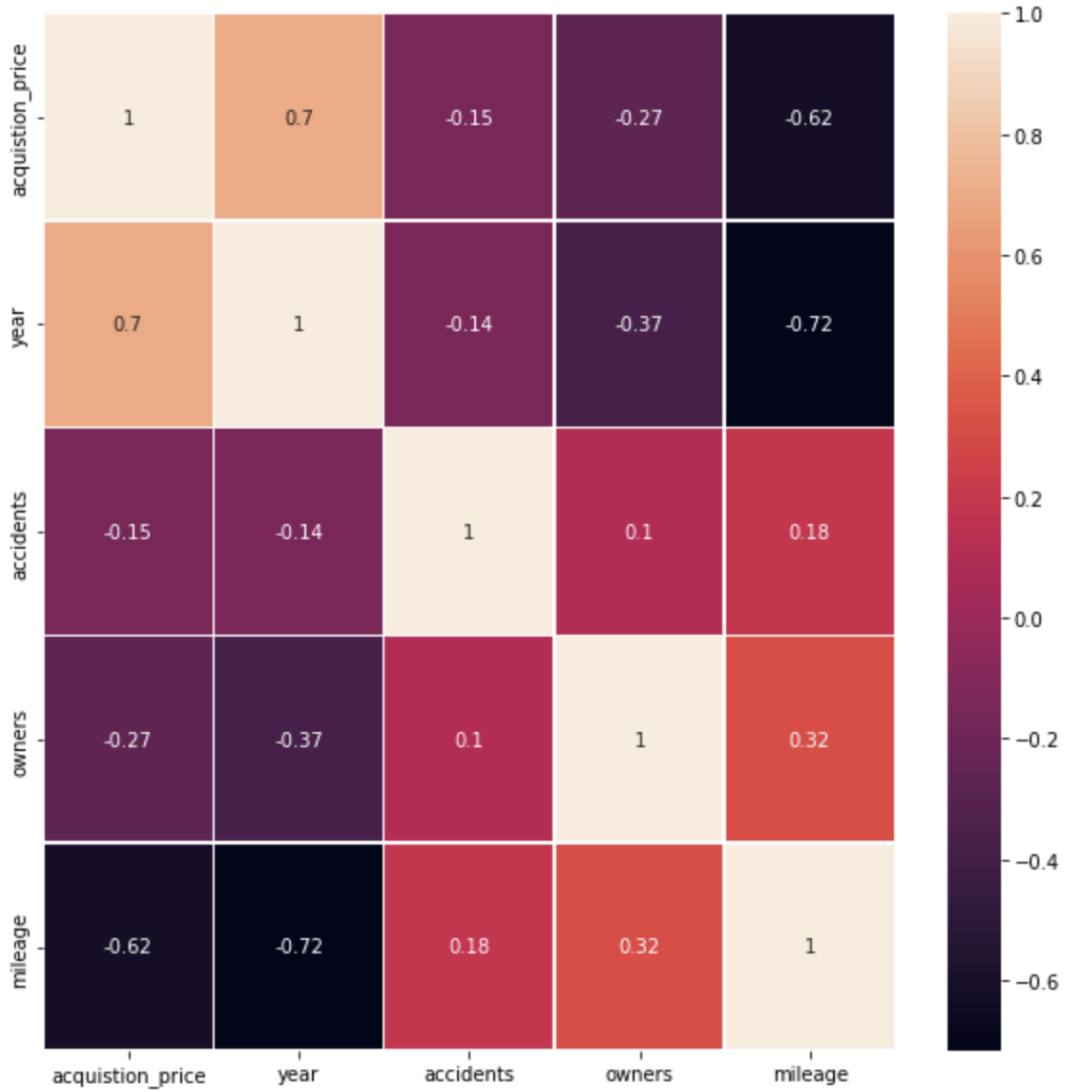
2. EXPLORATORY DATA ANALYSIS

f. Null/Missing values

	feature	number of missing values	Percentage of missing values(%)
0	accidents	1273	10.603
1	owners	964	8.029
2	region	8	0.067
3	final_sale_price	4	0.033

2. EXPLORATORY DATA ANALYSIS

g. Correlation of features with acquisition price



2. EXPLORATORY DATA ANALYSIS

g. Correlation of features with acquisition price

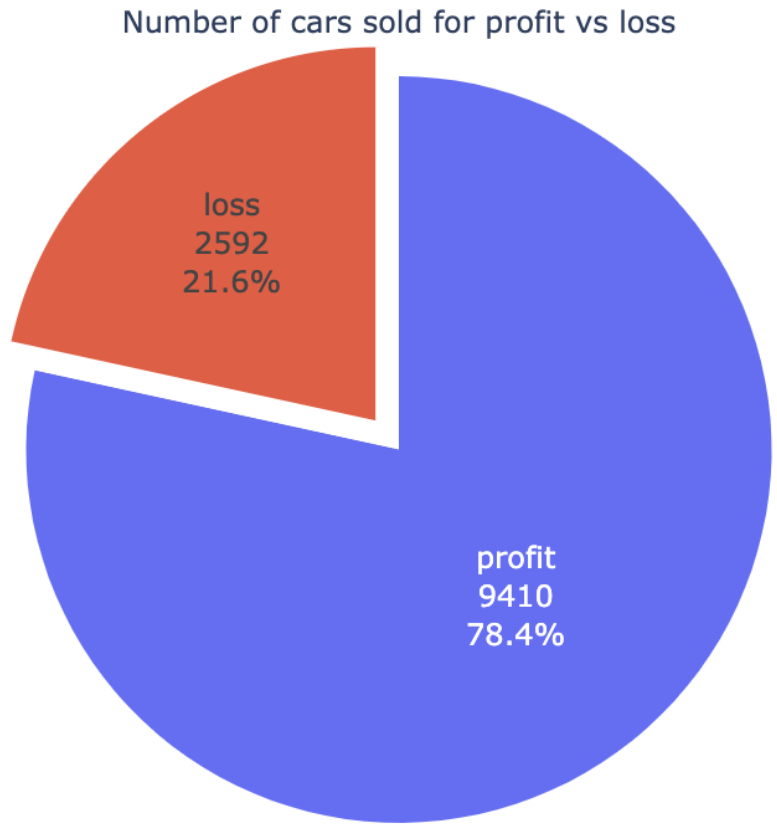
	Feature	correlation
1	year	0.703
2	accidents	-0.150
3	owners	-0.266
4	mileage	-0.619

- We can observe that the Acquistion price is correlated with the year of manufacturing. So new cars have high price.
- The Acquistion price is inversely correlated with accidents. The cars involved in accidents are sold for lower prices.
- The Acquistion price is inversely correlated with owners. If a car has multiple owners, the prices are less.
- The Acquistion price is also inversely correlated with mileage. The price of the car is less if the mileage is more.

2. EXPLORATORY DATA ANALYSIS

h. Number of cars sold for profit vs loss

- Number of cars sold for profit = 9410
- Number of cars sold for profit = 2592
- 78.4% of the cars acquired are sold for profit whereas 21.6% of cars are sold at loss.
- Total net profit made so far = ~**14million dollars**



2. EXPLORATORY DATA ANALYSIS

i. Profit/loss based on makes

- Shift made more net profit on Toyota, Honda and BMW.
- Total Net profit on Toyota = 2.3 million, on Honda = 1.7 million
- Shift loss money on GMC and Saturn.

make	total profit
Toyota	2355053.000
Honda	1732389.000
BMW	1243350.000
Mazda	932570.000
Volkswagen	834460.000
Saturn	-6900.000
GMC	-16418.000

3. PREPROCESSING

Null/missing value preprocessing

- When accidents value is null, I assume accidents = 0 as there is high probability that the car is not involved in any accident.
- When owners value is missing, I assume number of car owners = 1
- I deleted 4 rows where final sale price is null.

- Remove duplicate values
- Convert categorical variables to lower case
- Strip spaces in categorical variables
- Label encode categorical variables

4. MODEL FORMULATION

- ***Model type***- As we have to approximate the acquisition price of the car, I choose this as a **regression problem**.

- ***Evaluation metrics***- I used **coefficient of determination(r2 squared)** as base evaluation metrics. I calculated possible profits from model predictions.

- ***Train test split***- As there is no explicit testing data provided, I used **train test split** to create separate train and test data sets. Throughout modeling, I used train test split ratio of 0.2 and random state = 9 to obtain consistent results. There are 9601 train samples and 2401 test samples.

- ***Model selection-*** My idea was to **start small** and build a linear regression model and see how it performs on the data. This serves as a baseline **model**. Later depending on the performance, I will move on to other machine learning models like Random forests and Xgb Regressor.

5. TRAINING DATA

- The moto of the project is to find the **optimal acquisition price**.
- In the dataset, there are some cars which are sold for profit and some for loss.
  - When the acquisition price is optimal? -> If the car is sold for a profit.
  - When the acquisition price is not optimal? -> If the car is sold for a loss.
- So I calculated net profit loss which calculates the net profit or loss for each car in the train data set. If net profit loss is negative, then the car is sold for a loss. If net profit loss is positive, then the car is sold for profit.
- **We don't want the cars which are sold for a loss** in our training data because the acquisition price for these is not optimal.
- So I have done the following experiments and trained the model:
  - Removed all the instances in training data where net profit loss is negative.
  - Imputed the acquisition price of training data where net profit loss is negative with sale price + average profit of the car

6. ML MODELS

- I trained the data on following models:
  - Linear Regression
  - SVM
  - Random Forest
  - Xgboost Regressor
- I fine-tuned the Xgboost model.
- Parameters tuned:

	Parameter tuned	**Values**
1	Number of estimators	100, 300, 500
2	Learning rate	0.05, 0.1
3	Maximum tree depth	5,6,8
4	Minimum child weight	3,4,5

7. MODEL EVALUATION CRITERIA

a. Model Metrics

- I used **r2 squared(coefficient of determination)** as model evaluation metrics.
- However in both training and test datasets, There are cars which are sold for both profit and loss.
- In training dataset, I removed all the instances where cars are sold for a loss.
- In test data set, As we are predicting the optimal acquisition price:
  - For the cars which are sold for profit, the predicted acquisition price should be closer to that of the true acquisition price. So r2 squared should be higher.
  - For the cars which are sold for loss, the true acquisition price is apparently wrong. In these cases, There is very high chance that shift paid more than the true value of the car. So the predicted acquisition price should be different(lower) than the true acquisition price. So r2 squared should be lower.
- As there are cars which are sold for profit and loss in test data, having high r2 square means the models is also learning data for cars which are sold at a loss. So I used net profit and loss calculation for better view of model performance.

7. MODEL EVALUATION CRITERIA

b. Total net profit or loss

- I calculated the total net profit/loss based on true acquisition price and model predicted acquisition price.

a. Final sale price: The price at which the car is sold

b. True acquisition price: The price at which the car is acquired

c. Predicted acquisition price: The acquisition price which the ML models predicted.

True net profit loss = Final sale price - True acquisition price

Predicted net profit loss = Final sale price - Predicted acquisition price

- If the Predicted net profit loss > True net profit loss: The ML models predictions of acquisition price are better than existing data as more profits can be obtained.
- If the Predicted net profit loss < True net profit loss: The ML models predictions of acquisition price are bad.

8. ML MODEL COMPARISON

S.No	Model	R2 squared	True net profit/loss(In millions)	Predicted net profit/loss(In millions)
1	Linear regression	0.78	2.83	2.46
2	SVM	0.11	2.83	-0.51
3	Random forest	0.84	2.83	2.75
4	Xgb regressor	0.80	2.83	2.77
5	Xgb regressor – fine tuning	**0.85**	2.83	**2.95**

9. CONCLUSION

- Built a regression model to predict the car acquisition price.
- Using Xgboost with parameter tuning, an r2 square of 0.85 was obtained.
- The predicted net profit on test data using Xgb is 2.95 million dollars whereas existing net profit on test data is 2.83 million dollars.

There is still lots of scope for model performance improvement. Due to the time constraints, I stopped going further. These are some of the places of improvement:

- Cold start problem: I observed many cars doesn't have enough data points. Using data augmentation techniques and external data sources can further improve acquisition price prediction.
- Concentrating on places where cars are sold for a loss may yield better results.