Used Car Price Prediction and Recommender System

Trang Khong

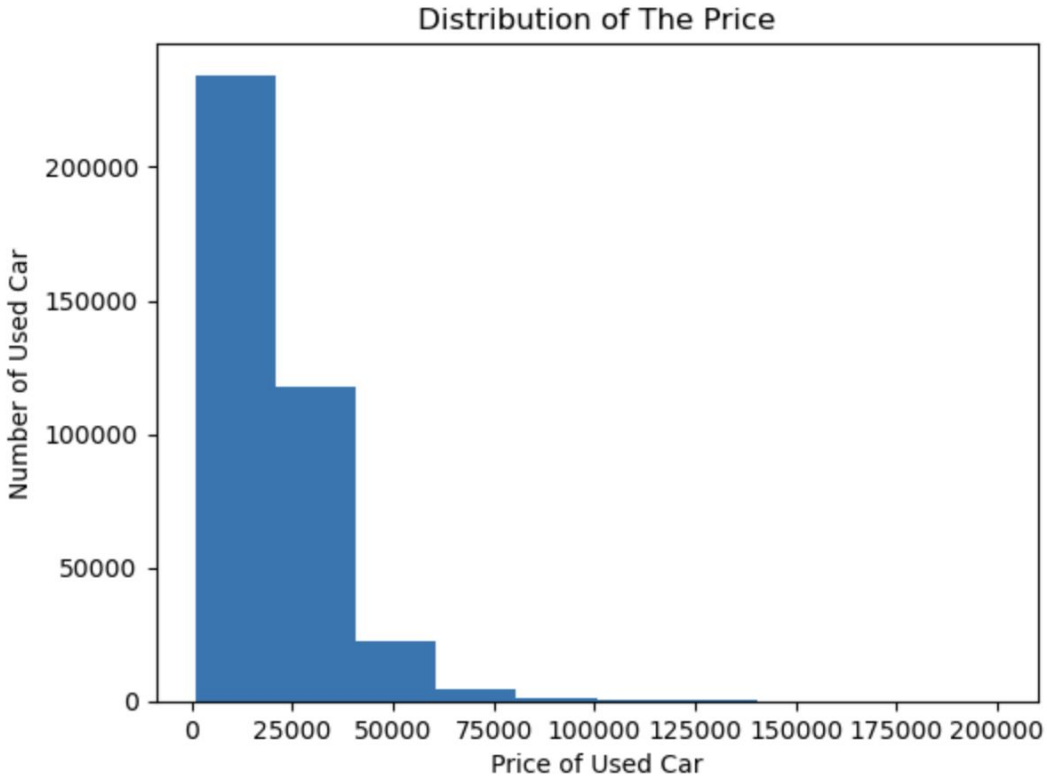
Problem Statement

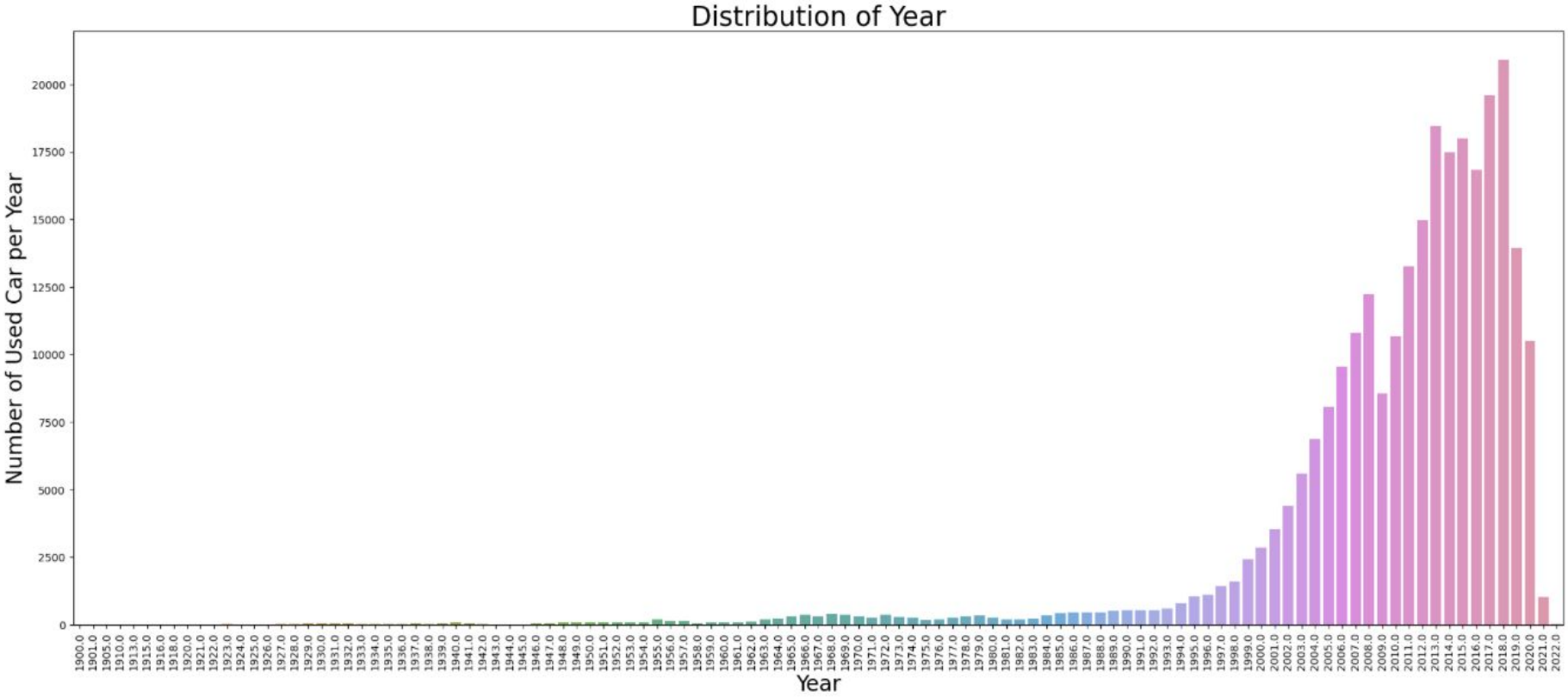
- The price of used cars has increased after the coronavirus pandemic due to the issue of supply chain.
- Making a decision on buying and selling a car is challenging.
- The purpose of this project is to build a model to predict the price of used cars and a used car recommender system.
- Those models will help sellers when they are planning to sell their car and buyers to have an idea when they are looking for a used car.

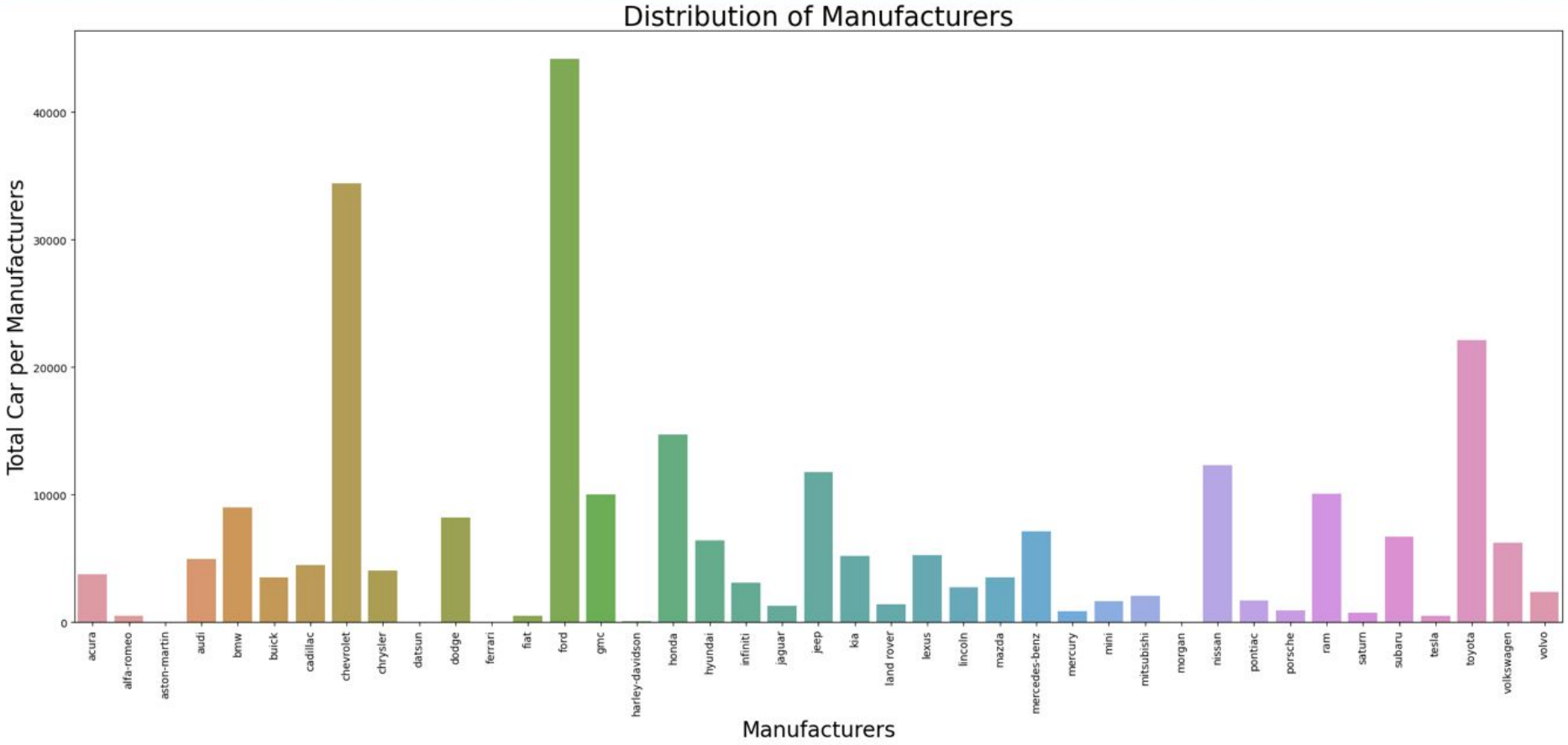
Analyzed Data

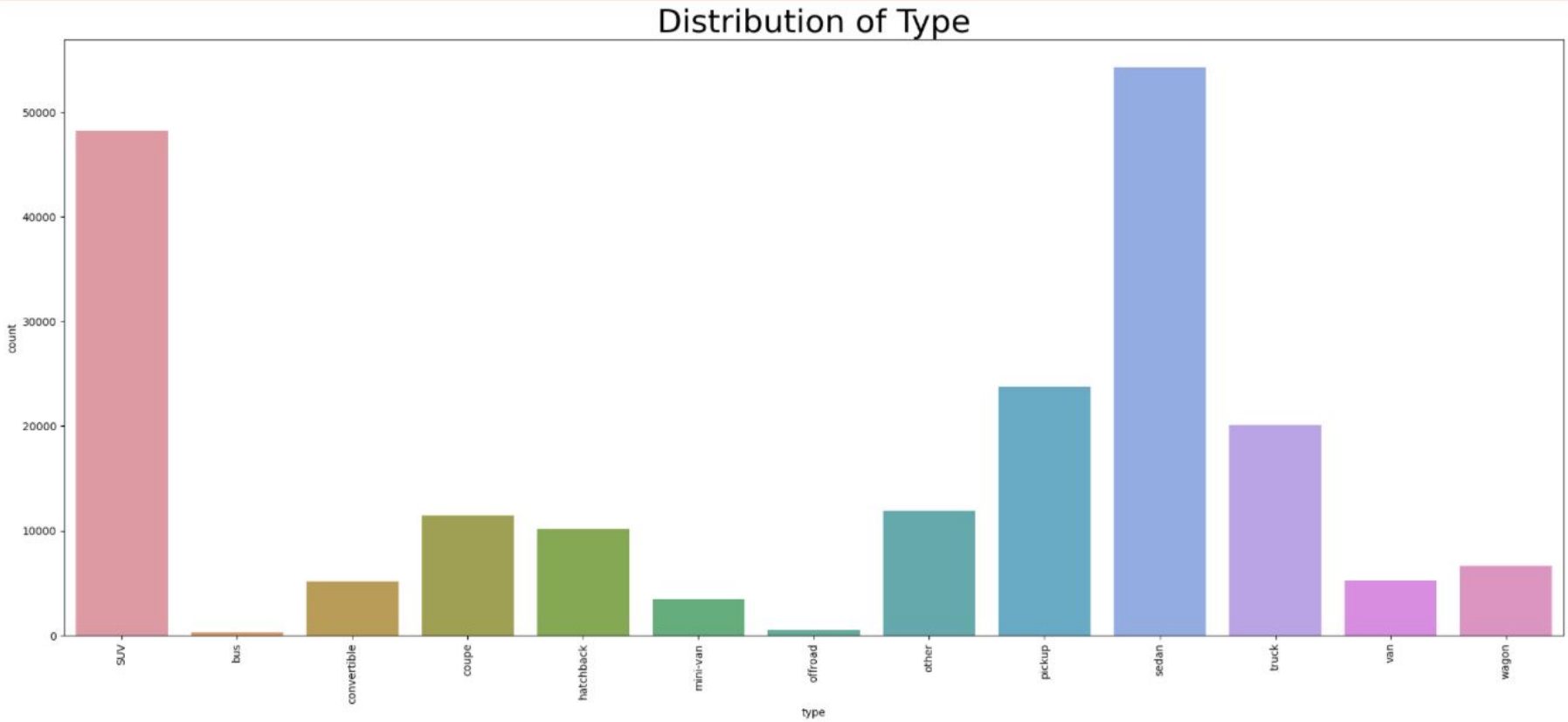
- Data was obtained from Kaggle, collected by Austin Reese from Craigslist.
- The data have 26 columns.
- After cleaning, the data have 14 columns.
- The clean data will be use for imputing and modeling.
- List of features are applied to predict the price of used car are: year, manufacturers, condition, cylinders, fuel, odometer, title status, transmission, drive, size, type, color, state.

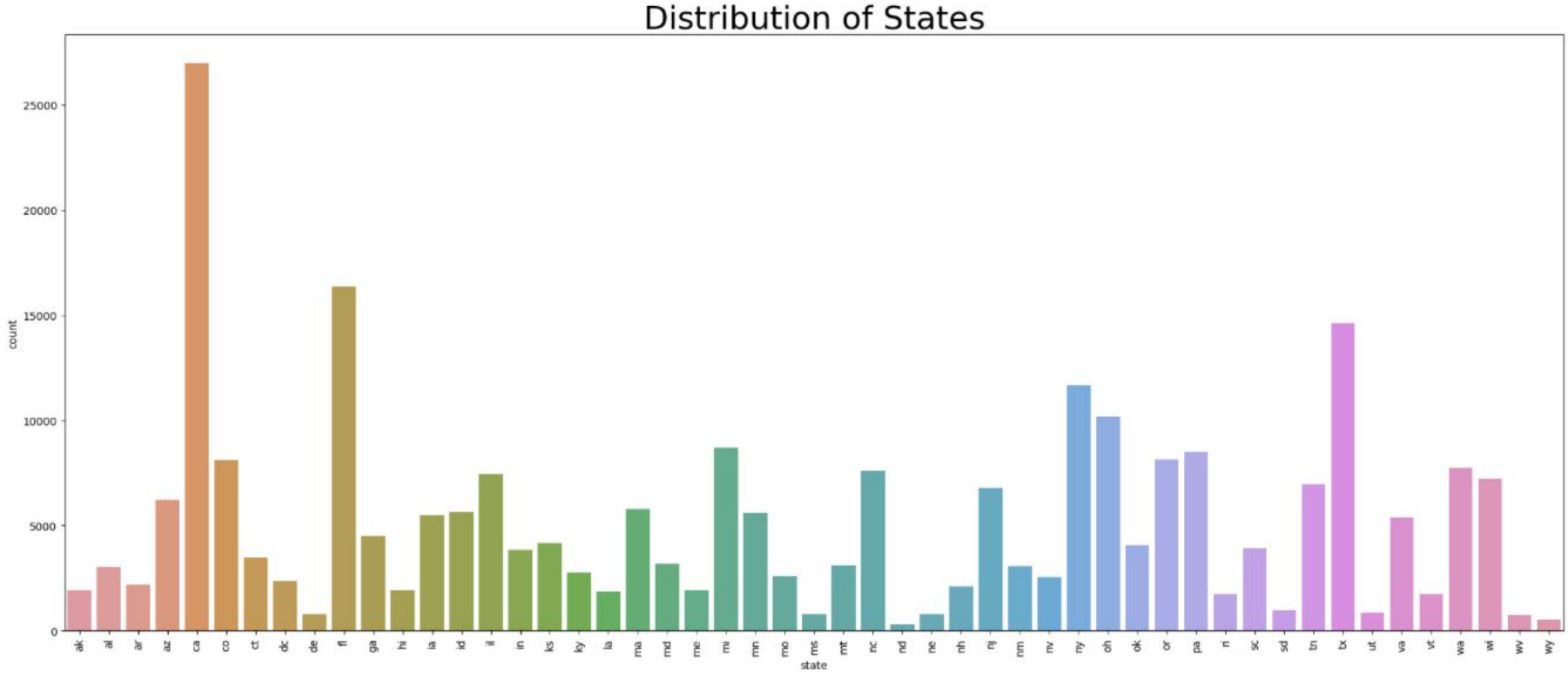
Distribution of Price




Distribution of Year
Number of Used Car per Year
Year
0
2500
5000
7500
10000
12500
15000
17500
20000
1900.0
1901.0
1905.0
1910.0
1913.0
1915.0
1916.0
1918.0
1920.0
1921.0
1922.0
1923.0
1924.0
1925.0
1926.0
1927.0
1928.0
1929.0
1930.0
1931.0
1932.0
1933.0
1934.0
1935.0
1936.0
1937.0
1938.0
1939.0
1940.0
1941.0
1942.0
1943.0
1944.0
1945.0
1946.0
1947.0
1948.0
1949.0
1950.0
1951.0
1952.0
1953.0
1954.0
1955.0
1956.0
1957.0
1958.0
1959.0
1960.0
1961.0
1962.0
1963.0
1964.0
1965.0
1966.0
1967.0
1968.0
1969.0
1970.0
1971.0
1972.0
1973.0
1974.0
1975.0
1976.0
1977.0
1978.0
1979.0
1980.0
1981.0
1982.0
1983.0
1984.0
1985.0
1986.0
1987.0
1988.0
1989.0
1990.0
1991.0
1992.0
1993.0
1994.0
1995.0
1996.0
1997.0
1998.0
1999.0
2000.0
2001.0
2002.0
2003.0
2004.0
2005.0
2006.0
2007.0
2008.0
2009.0
2010.0
2011.0
2012.0
2013.0
2014.0
2015.0
2016.0
2017.0
2018.0
2019.0
2020.0
2021.0
2022.0


Distribution of Manufacturers
Total Car per Manufacturers
0
10000
20000
30000
40000
acura
alfa-romeo
aston-martin
audi
bmw
buick
cadillac
chevrolet
chrysler
datsun
dodge
ferrari
fiat
ford
gmc
harley-davidson
honda
hyundai
infiniti
jaguar
jeep
kia
land rover
lexus
lincoln
mazda
mercedes-benz
mercury
mini
mitsubishi
morgan
nissan
pontiac
porsche
ram
saturn
subaru
tesla
toyota
volkswagen
volvo
Manufacturers

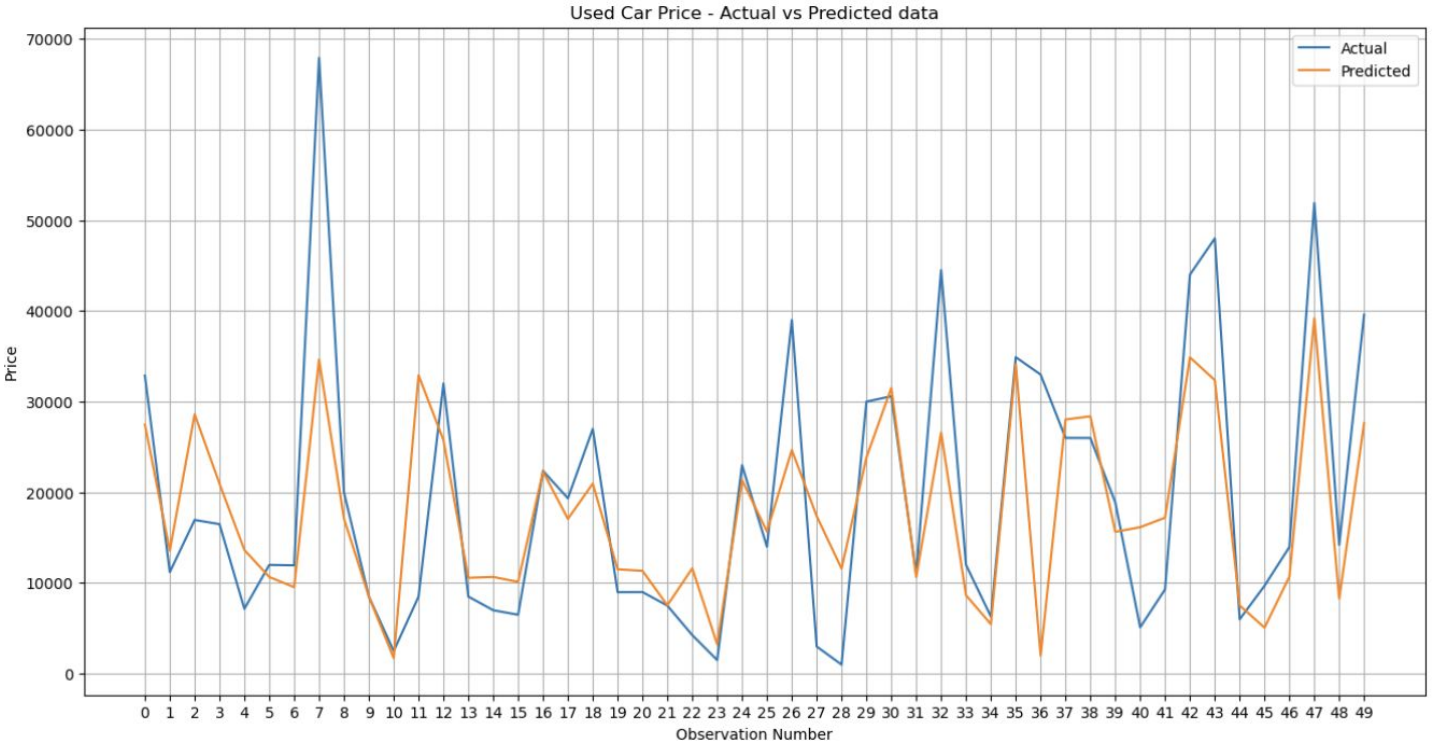

Distribution of Type
count
50000
40000
30000
20000
10000
0
SUV
bus
convertible
coupe
hatchback
mini-van
offroad
other
pickup
sedan
truck
van
wagon
type


Distribution of States
count
25000
20000
15000
10000
5000
0
state
ak
al
ar
az
ca
co
ct
dc
de
fl
ga
hi
ia
id
il
in
ks
ky
la
ma
md
me
mi
mn
mo
ms
mt
nc
nd
ne
nh
nj
nm
nv
ny
oh
ok
or
pa
ri
sc
sd
tn
tx
ut
va
vt
wa
wi
wv
wy

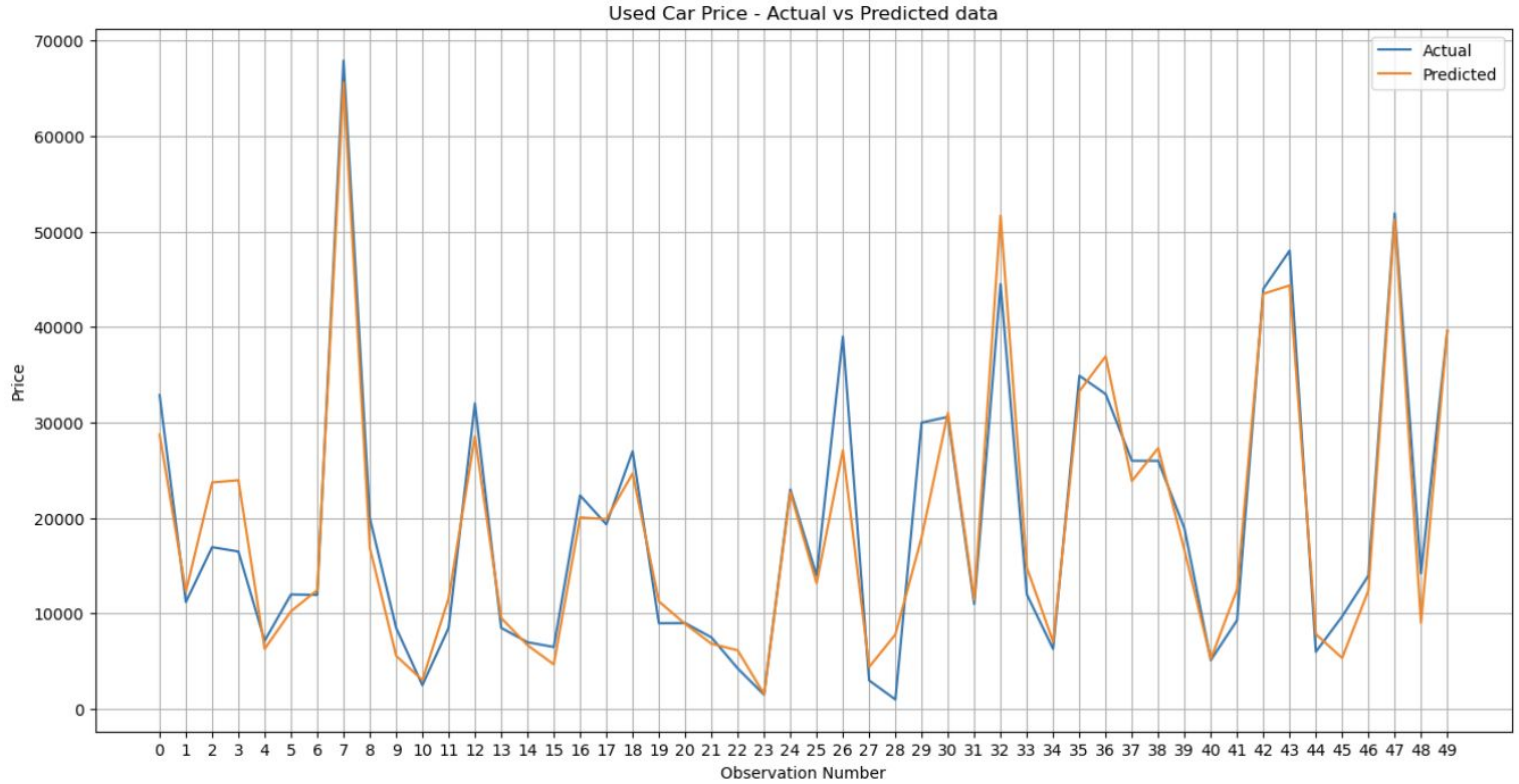
Imputing Method

- KNN imputer
- Mean and mode imputer base on different range of price
  - 1000 to 15000
  - 15000 to 50000
  - 50000 to 200000
- Different price of range give different mean and mode
- All models were compared to a baseline model.
- The used car price is evaluated with R Squared states how much of the variability in the data are explained by model, and Root Mean Squared Error (RMSE) represents (approximately) average difference between model predicted values and the actual values.

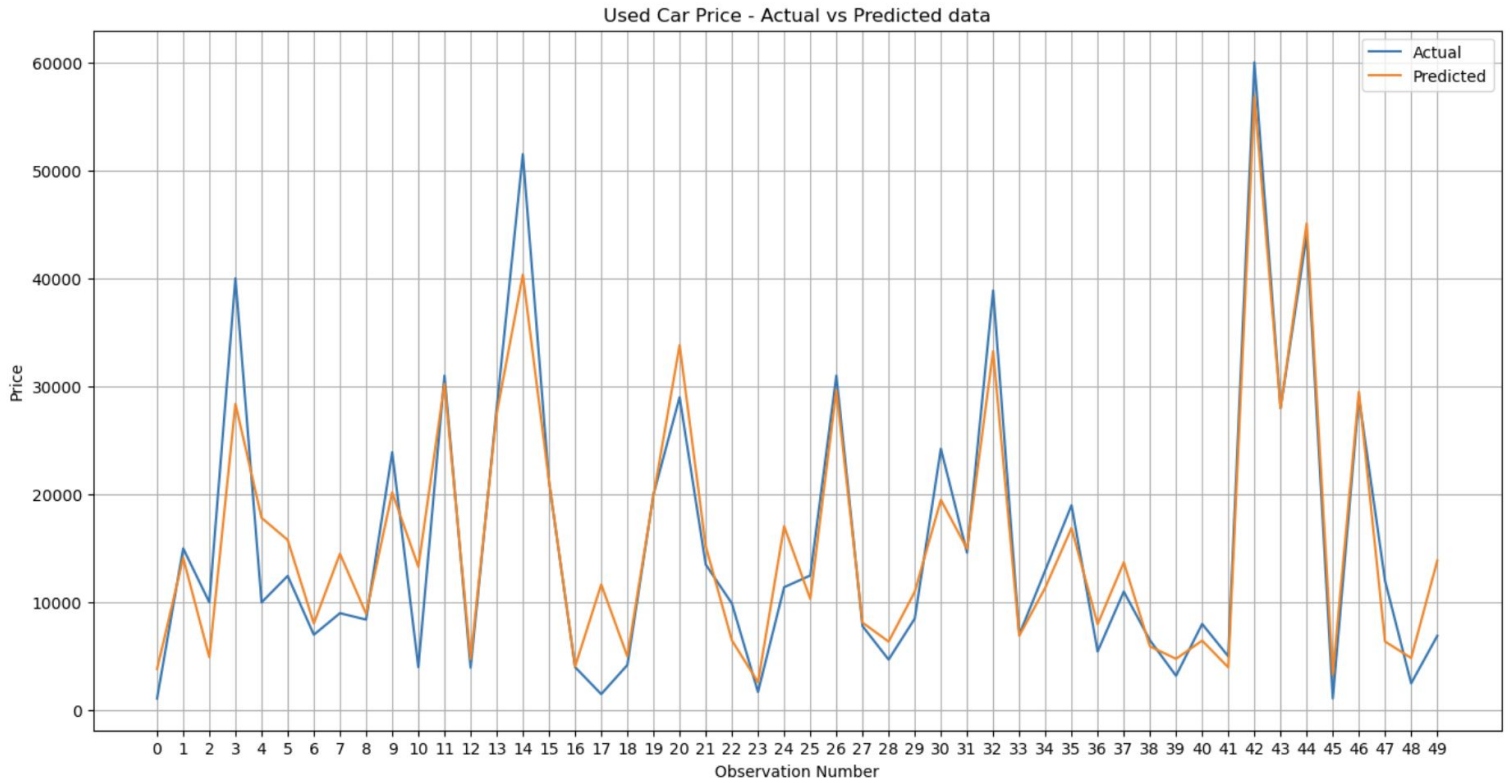
Linear Regression with Mean Imputing



Random Forest Regressor with Mean Imputer



Random Forest Regressor with KNN Imputer



Model Results with Two Different Imputer

Name of Model	Train score	Test score (R2)	RMSE
Baseline (Null model)	0.00	0.00	14504.270
Linear Regression (mean imputer)	0.514	0.521	10081.260
RandomForest Regressor (mean imputer)	0.982	0.874	5172.700
ExtraTrees Regresssor (mean imputer)	0.952	0.874	5161.382
AdaBoost Regressor (mean imputer)	0.961	0.872	5217.493
GradientBoosting Regressor (mean imputer)	0.777	0.775	6913.230
RandomForest Regressor(pipeline)	0.982	0.874	5167.861
ExtraTrees Regressor(pipeline)	0.999	0.873	5179.484
RandomForest Regressor (KNN imputer)	0.925	0.836	5853.053
ExtraTrees Regresssor (KNN imputer)	0.941	0.840	5781.103
AdaBoost Regressor (KNN imputer)	0.953	0.836	5867.411

Discussion

- The best performance model is Extra Trees Regressor with lowest RMSE and highest R2 score on both imputing methods.
- Linear Regression and Gradient Boosting have lowest train score and test score but the different between train and test score are small, those models are less overfit.
- The others models have train score higher than test score, those models are overfit but the test score are improved.
- Tuning hyperparameters can be apply to help improve the overfit.
- Keep the hyperparameter the same, train score, test score, and RMSE will be slightly different when apply the pipeline versus without applying the pipeline.
- Using the same model and hyperparameters, models with KNN imputer are less performance than mean imputer. However, those models with KNN imputer are less overfit than those models with mean imputer.

Price Recommender

```
price_recommender((10000, 20000),7)
```

	price	manufacturer	type	year	condition	fuel	title_status	transmission	paint_color
0	12998.0	nissan	sedan	2018.0	excellent	gas	clean	automatic	white
1	16500.0	buick	SUV	2014.0	excellent	gas	clean	automatic	white
2	16900.0	jeep	SUV	2018.0	good	gas	clean	other	white
3	16990.0	honda	sedan	2015.0	good	other	clean	other	silver
4	17337.0	toyota	hatchback	2018.0	good	gas	clean	automatic	white
5	18500.0	audi	SUV	2015.0	excellent	diesel	clean	automatic	black
6	18995.0	land rover	SUV	2012.0	good	gas	clean	automatic	black

Brand Recommender

```
brand_recommender('toyota',(10000, 20000),7)
```

	price	manufacturer	type	year	condition	fuel	title_status	transmission	paint_color
0	10995.0	toyota	SUV	2008.0	excellent	gas	clean	automatic	white
1	12987.0	toyota	sedan	2012.0	excellent	gas	clean	automatic	white
2	14998.0	toyota	SUV	2012.0	excellent	gas	clean	automatic	white
3	14999.0	toyota	SUV	2013.0	excellent	gas	clean	automatic	white
4	16999.0	toyota	truck	2010.0	good	gas	clean	manual	silver
5	17995.0	toyota	sedan	2014.0	like new	gas	clean	automatic	red
6	19500.0	toyota	SUV	2013.0	excellent	gas	clean	automatic	white

Used Car Recommender

```
all_recommender('acura',2007,'sedan',(5000, 10000),7)
```

	price	manufacturer	type	year	condition	fuel	title_status	transmission	paint_color
0	5500.0	acura	sedan	2007.0	excellent	gas	rebuilt	automatic	white
1	5950.0	acura	sedan	2007.0	excellent	gas	clean	automatic	grey
2	6900.0	acura	sedan	2007.0	excellent	gas	clean	automatic	blue
3	7500.0	acura	sedan	2007.0	excellent	gas	clean	automatic	black
4	7995.0	acura	sedan	2007.0	excellent	gas	clean	automatic	blue
5	7995.0	acura	sedan	2007.0	good	gas	clean	automatic	silver
6	8550.0	acura	sedan	2007.0	excellent	gas	clean	automatic	white

Cosine Similarity Recommender

	price	manufacturer	type	year	cylinders
10860	3000.0	nissan	SUV	2004.0	4
20084	3000.0	nissan	SUV	2004.0	4
10203	2995.0	nissan	SUV	2003.0	4
5895	3000.0	nissan	SUV	2003.0	4
32363	2800.0	nissan	SUV	2002.0	4
25463	2850.0	nissan	SUV	2007.0	4
24269	3290.0	nissan	SUV	2005.0	4
35024	3500.0	nissan	SUV	2003.0	4
19548	2195.0	nissan	SUV	2002.0	4
12262	2700.0	nissan	SUV	2010.0	4

Discussion

- Due to the limit of computer memory, the cosine similarity was not be able to run for this dataset. Limited the price of the data to limited the amount of data for cosine similarity.
- The cosine similarity works when limit the data to a small dataset.
- Brand recommender, Price recommender, and Used Car recommender can be use as a recommender system. Those recommender systems do not provide a perfect result but those recommender systems will give a list of used car base on some features that customers are looking for.

Conclusion

- The method use for data imputing has an important effect on model training and performance.
- The best performing model in this case was ExtratreesRegressor with mean imputed values.
- Cosine Similarity Recommender system will work with powerful computer or limited the data to certain size.

Bibliography

https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data

https://www.cnbc.com/2023/09/08/used-vehicle-prices-may-have-bottomed-for-2023.html#:~:text=Used%20vehicle%20prices%20have%20been,high%20prices%20amid%20resilient%20demand.

Thank you