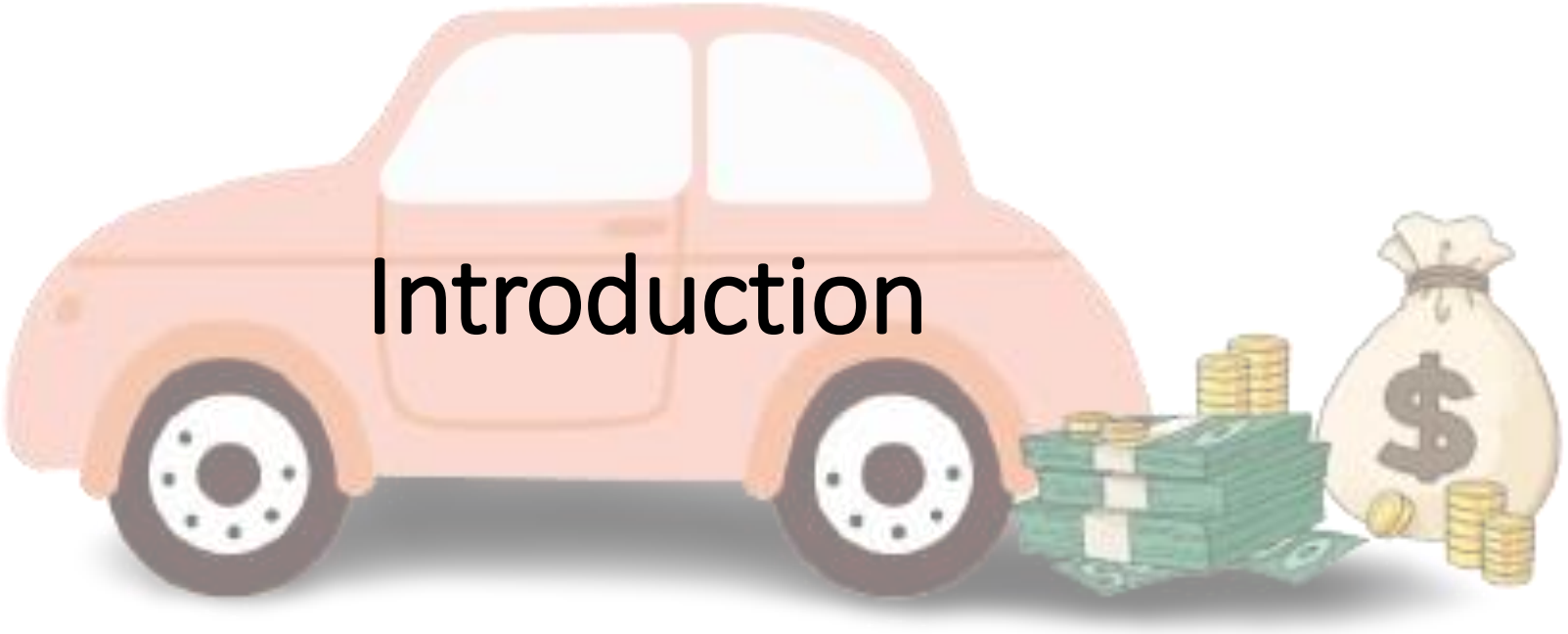
중고차 가격 예측하기

이소희, 신경훈, 김준구, 윤아람, 정은서, 안은정

Contents

- ✓ Introduction
- ✓ Data processing
- ✓ Modeling
- ✓ Results
- ✓어느 지역에 팔아야 할까?
- ✓ Modeling Conclusion
- ✓Conclusion

Introduction

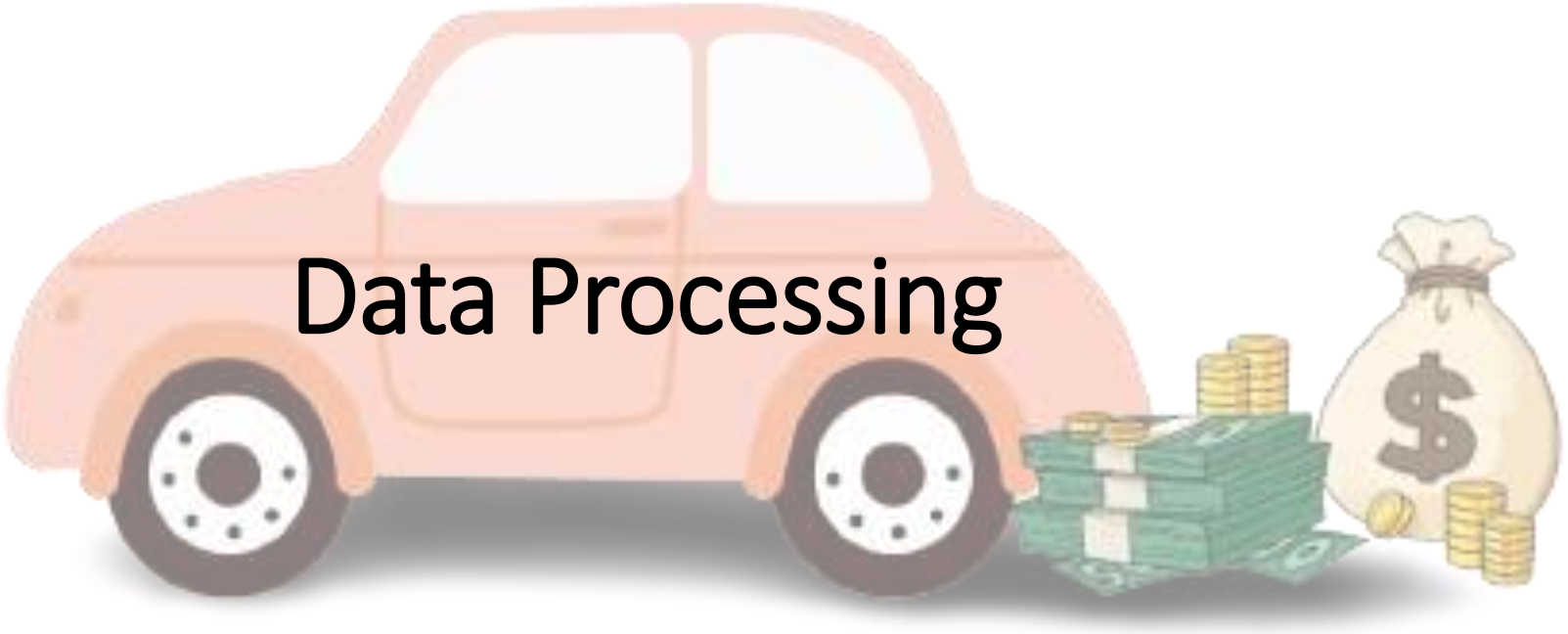


Data

- Kaggle에서 주최한 "인도 중고차 예측하기" 데이터셋의 train data를 바탕으로 중고차 가격 예측하기
- https://www.kaggle.com/datasets/avikasliwal/used-cars-price-prediction
- **6019** rows and **14** columns

Name	Location	Year	Kilometers Driven	Fuel Type	Transmission	Owner Type	Mileage	Engine	Power	Seats	New Price	Price
Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	First	26.6 km/kg	998 CC	58.16 bhp	5		1.75
Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	19.67 kmpl	1582 CC	126.2 bhp	5		12.5

Data Processing



Data Processing

Name	Location	Year	Kilometers Driven	Fuel Type	Transmission	Owner Type	Mileage	Engine	Power	Seats	New Price	Price
Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	First	26.6 km/kg	998 CC	58.16 bhp	5		1.75
Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	19.67 kmpl	1582 CC	126.2 bhp	5		12.5

Name 앞에 브랜드명만 남기기

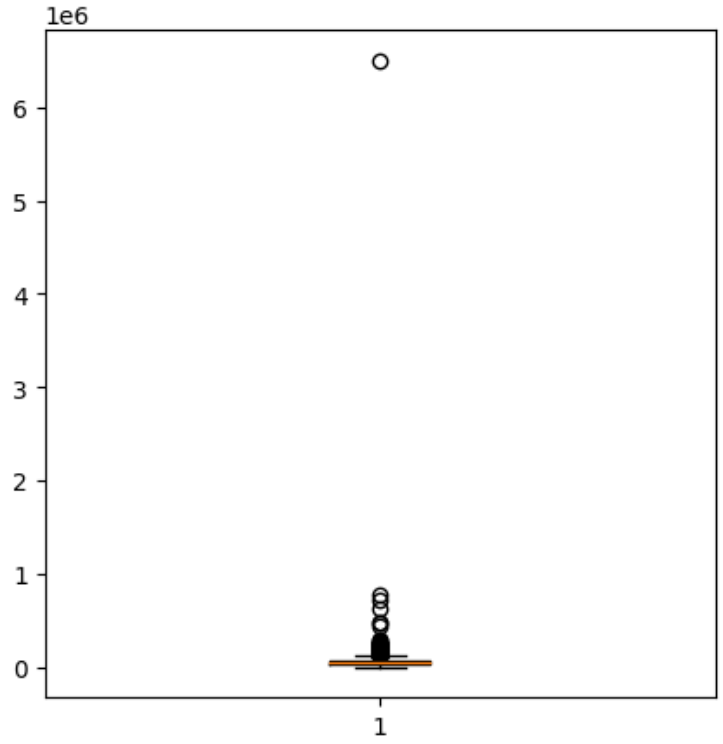
Fuel Type = 'CNG', 'LPG' -> Mileage 단위 km/kg
Fuel Type = 'Diesel', 'Petrol' -> Mileage 단위 kmpl
'CNG', 'LPG'에 각각 1.64, 1.3을 곱해서 kmpl로 단위 맞춰준 후 단위 제거

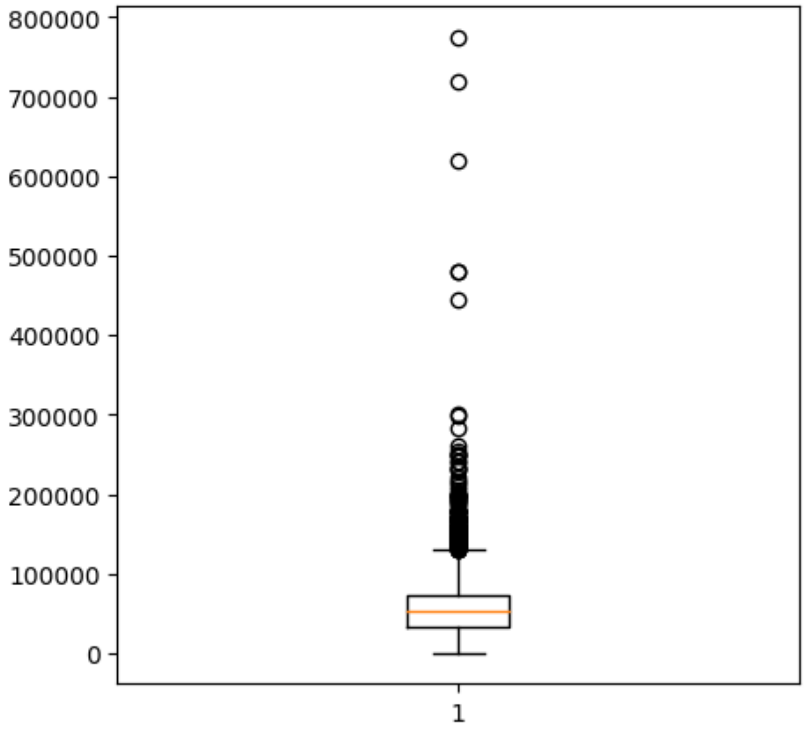
동일 모델의 신차 가격
예측 모델에 필요 없으므로 drop

Data Processing

- ‘Driven’ 이상치 데이터 제거
- ‘Nan’ 값 갖는 데이터 전부 제거

(자세한 전 처리 과정은 마지막 페이지 증빙자료 참조)





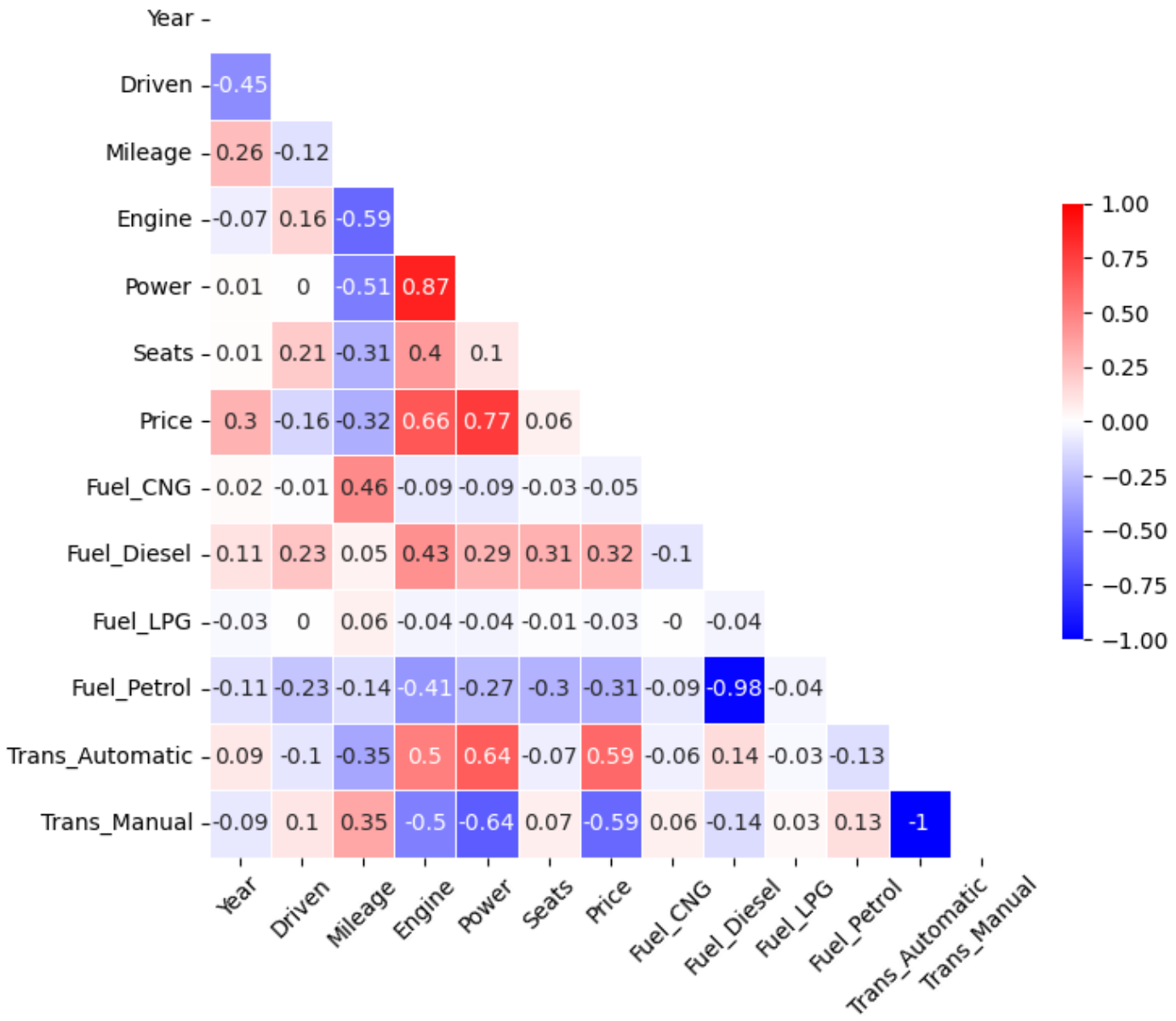
Data Processing

- Column명 바꾸고 One-Hot Encoding 한 후 테이블
- **5872** rows and **57** columns

Year	Driven	Mileage	Engine	Power	Seats	Price	Brand Ambassador	Brand Audi	Brand BMW	Brand Bentley	Brand Chevrolet	Brand Datsun	Brand Fiat	Brand Force	Brand Ford	Brand Honda	...
2010	72000	43.624	998	58.16	5	1.75	0	0	0	0	0	0	0	0	0	0	...
2015	41000	19.67	1582	126.2	5	12.5	0	0	0	0	0	0	0	0	0	0	...

Heat map

- 주요 변수들만 표시



Machine Learning

다중 공선성 판단

- 회귀 분석에서 하나의 feature(예측 변수)가 다른 feature와의 상관 관계가 높으면, 회귀 분석 시 부정적인 영향을 미칠 수 있기 때문에, 모델링 하기 전에 먼저 다중 공선성의 존재 여부를 확인 필요
- 보통 다중 공선성을 판단할 때 VIF(Variance Inflation Factors)값 이용
- 일반적으로, VIF > 10인 feature들은 다른 변수와의 상관관계가 높아, 다중 공선성이 존재하는 것으로 판단

	features	VIF Factor
0	Year	2.006301
1	Driven	1.692319
2	Mileage	4.254951
3	Engine	11.765154
4	Power	10.590181
5	Seats	2.540213

제거

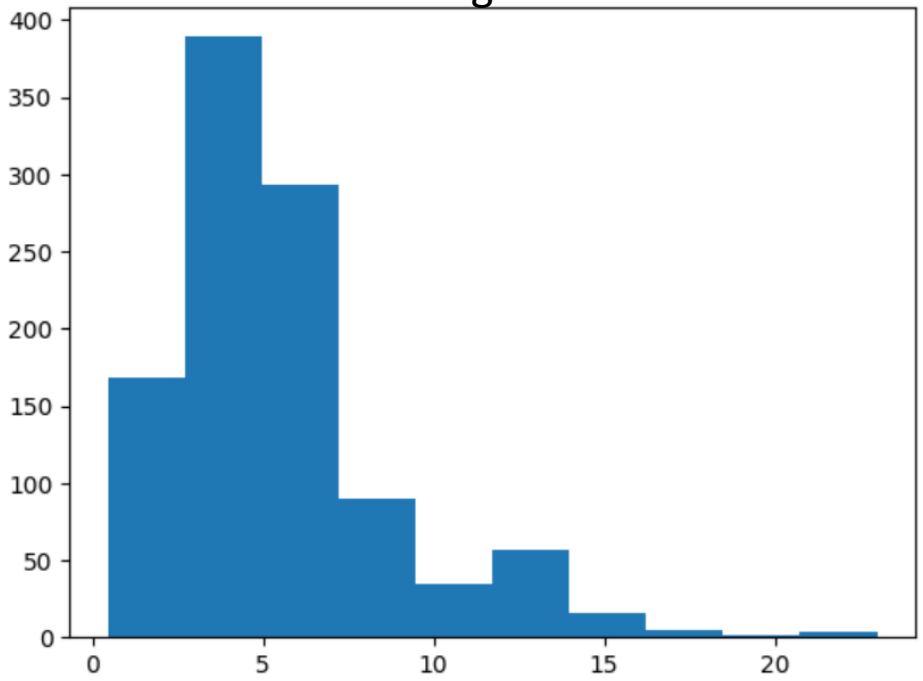
	features	VIF Factor
0	Year	1.981956
1	Driven	1.689667
2	Mileage	3.881672
3	Power	4.031312
4	Seats	2.355936

왜도, 첨도 확인

- y의 왜도가 3.32, 첨도가 19로 y의 분포가 고르지 않으므로 모델 정확도를 올리기 위하여 log를 사용한다.

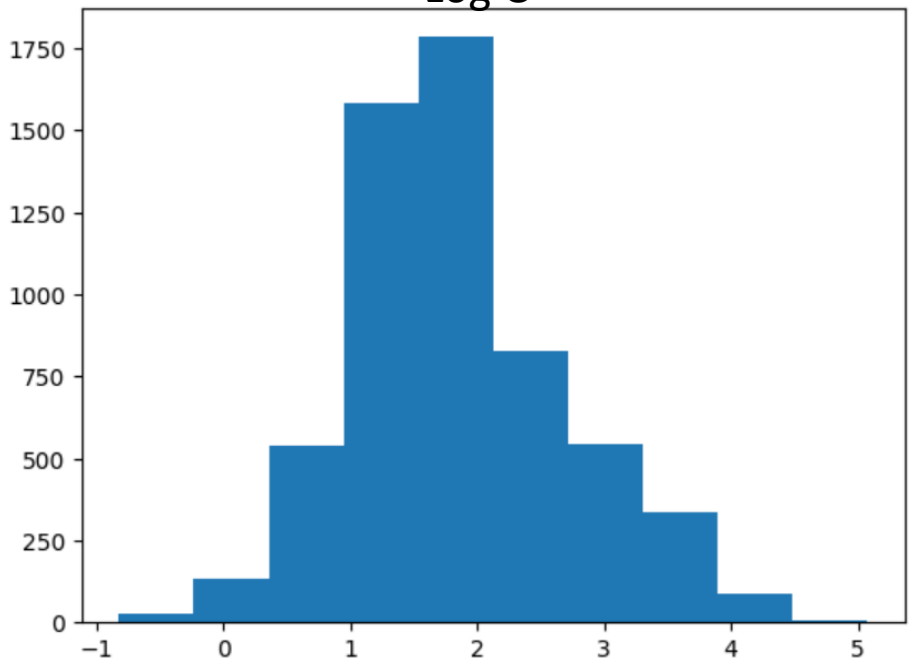
왜도	첨도
3.32	19

Log X



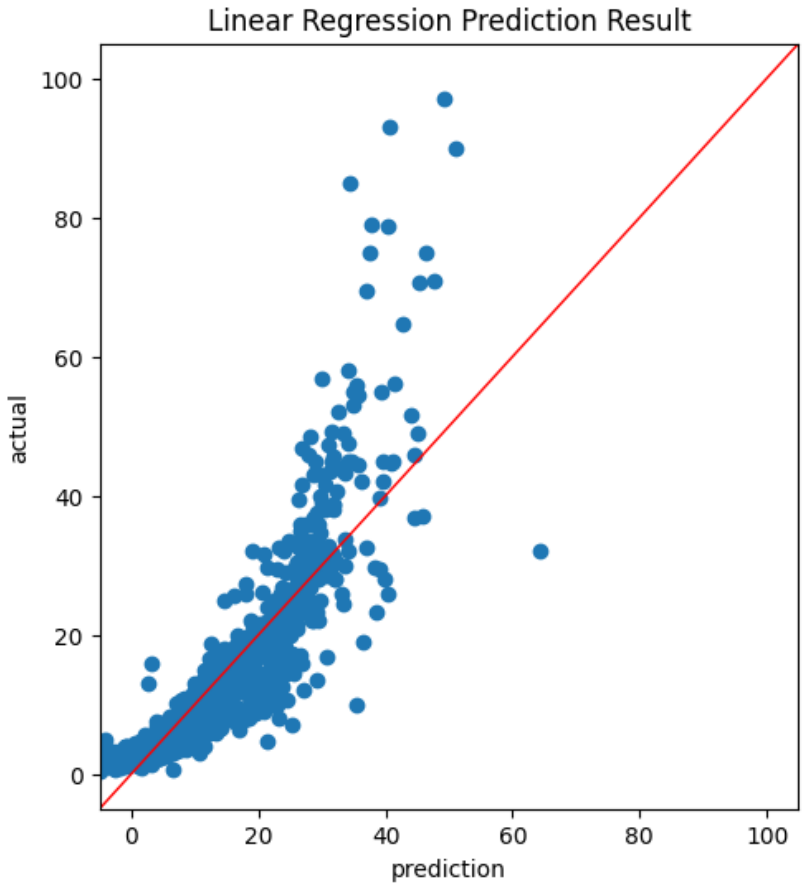
왜도	첨도
0.4	3.18

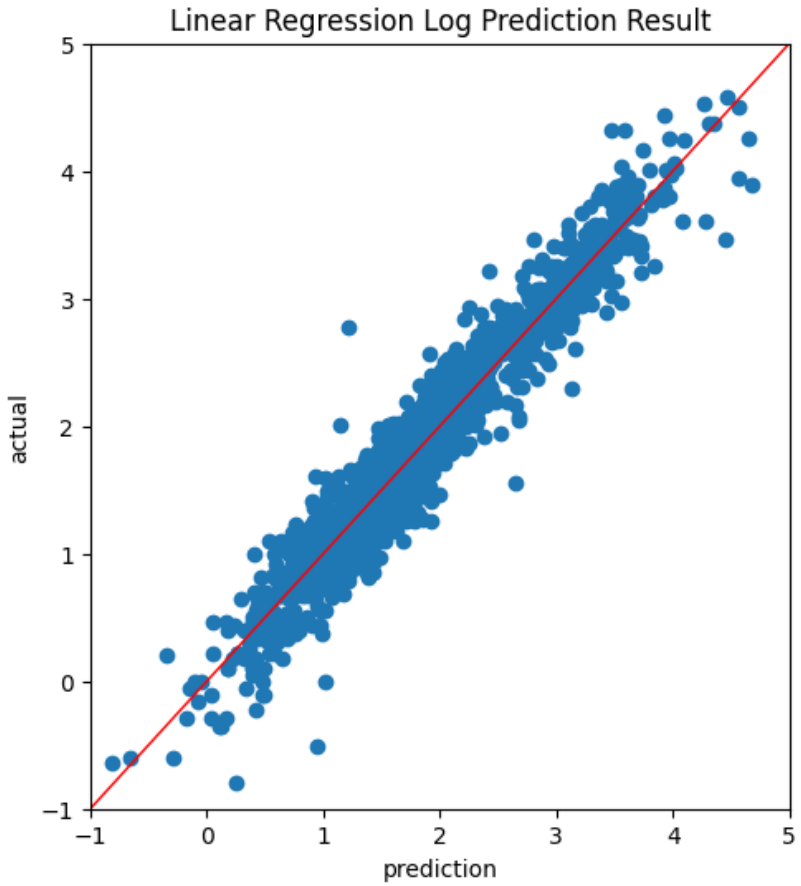
Log O



Results

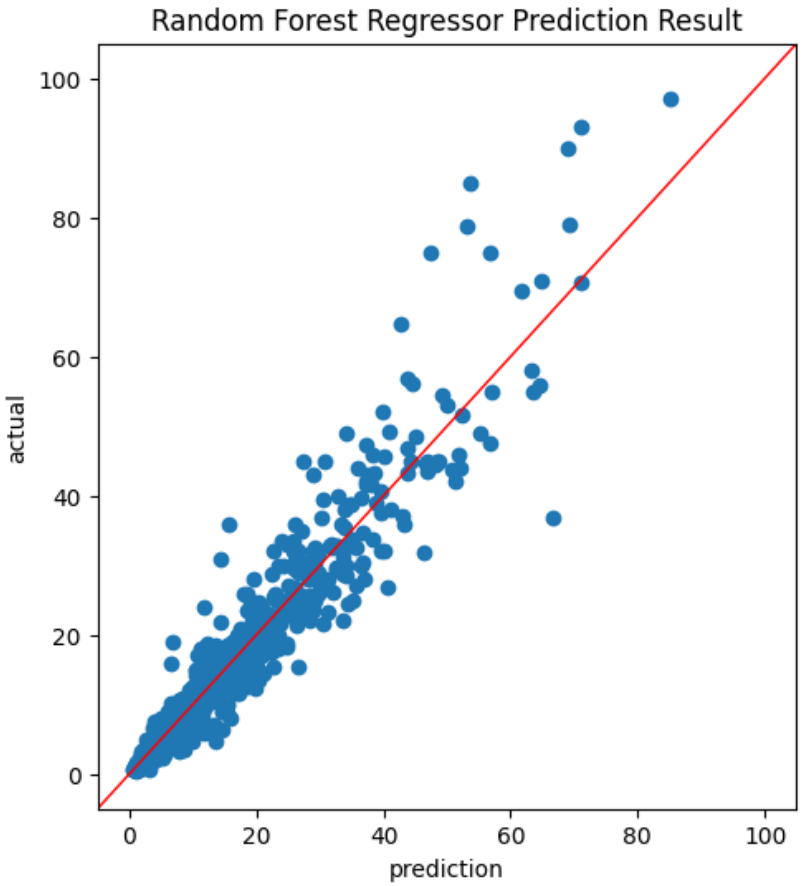
Linear Regression

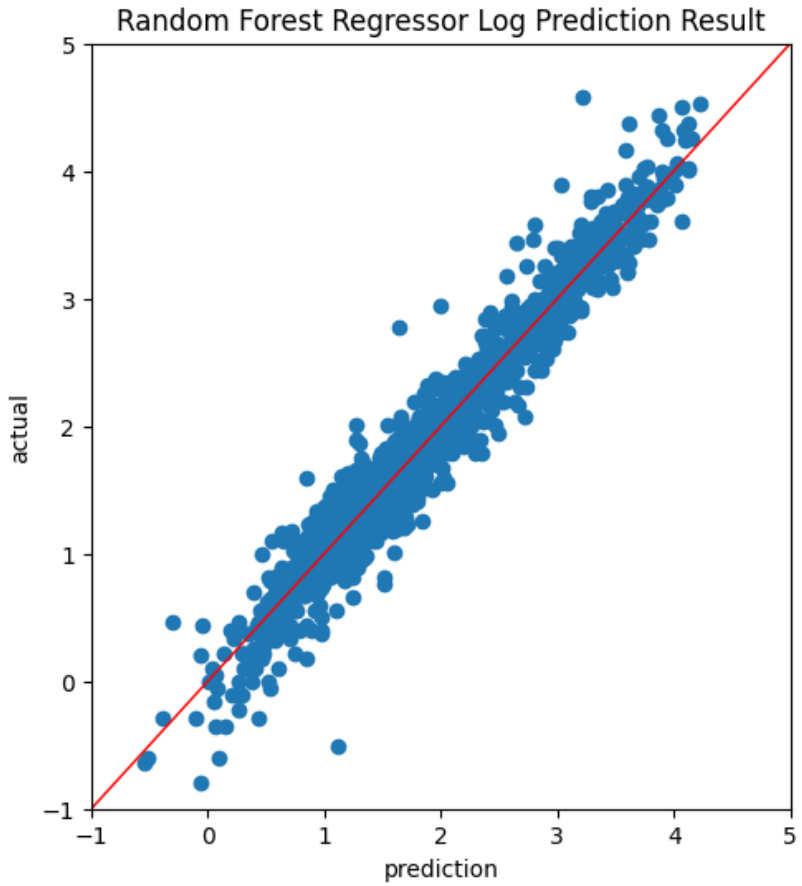




	Linear Regression	Linear Regression Log
accuracy	0.7803	0.93
r2 score	0.7790	0.92
Root Mean Square(rmse)	5.04	12.37

Random Forest Regressor



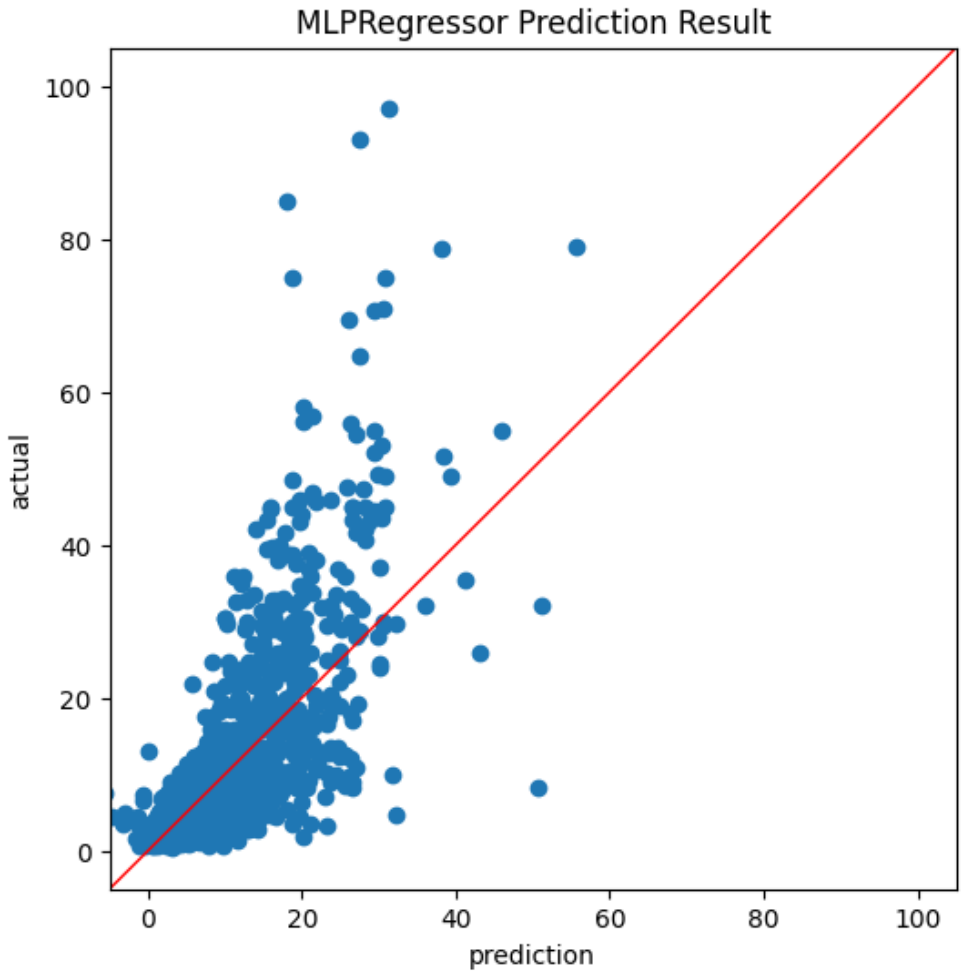


	Random Forest Regressor	Random Forest Regressor Log
accuracy	0.98	0.99
r2 score	0.923	0.88
Root Mean Square(rmse)	2.75	3.68

Decision Tree Regressor

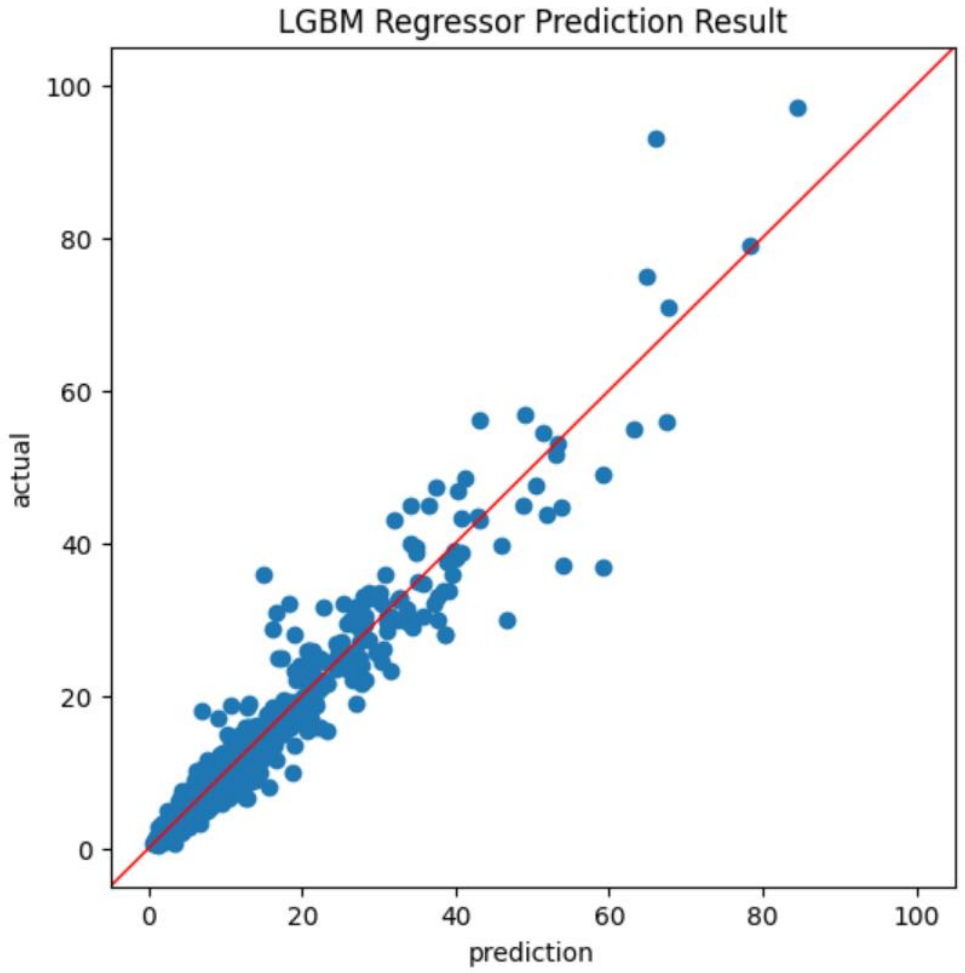
	Decision Tree Regressor
accuracy	1
r2 score	0.88
Root Mean Square(rmse)	3.68

Multi-layer Perceptron Regressor



	Multi-layer Perceptron Regressor
accuracy	0.72
r2 score	0.69
Root Mean Square(rmse)	5.83

LGBM Regressor



	Multi-layer Perceptron Regressor
accuracy	0.96
r2 score	0.94
Root Mean Square(rmse)	2.62

어느 지역에 팔아야 할까?

어느지역에 팔아야 할까?

	coef
Location_Coimbatore	-123.806340
Location_Bangalore	-124.178709
Location_Hyderabad	-124.572854
Location_Chennai	-124.705883
Location_Ahmedabad	-124.797414
Location_Kochi	-124.874754
Location_Mumbai	-124.964927
Location_Pune	-124.969929
Location_Jaipur	-124.999514
Location_Delhi	-125.329379
Location_Kolkata	-126.021372

- 브랜드 별로 지역과의 회귀계수들을 뽑아 오름차순 나열을 진행함
- 각 브랜드의 수가 100이상 인 것만 진행하였음
- Rank 1 인 지역에서 팔면 되겠다고 판단

어느지역에 팔아야 할까?

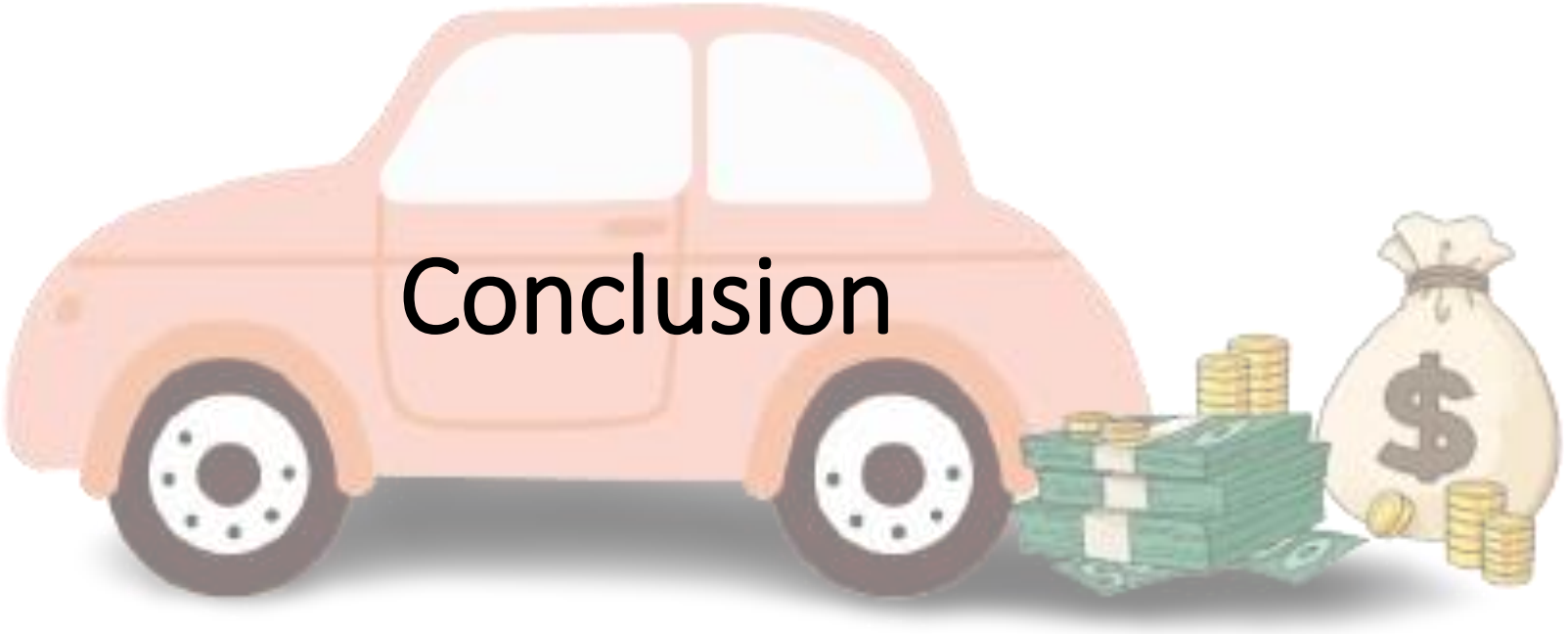
예측

Car name	Maruti	Hyundai	Honda	Toyota	Mercedes-Benz	Volkswagen	Ford	Mahindra	BMW	Audi	Tata	Skoda	Renault	Chevrolet
Stats models	Coimbatore	Coimbatore	Coimbatore	Coimbatore	Hyderabad	Coimbatore	Coimbatore	Coimbatore	Hyderabad	Hyderabad	Coimbatore	Hyderabad	Bangalore	Hyderabad
Random Forest Regressor	Hyderabad	Coimbatore	Coimbatore	Coimbatore	Mumbai	Kolkata	Coimbatore	Coimbatore	Mumbai	Kochi	Hyderabad	Hyderabad	Bangalore	Coimbatore

인도 도시별 GDP 순위

1위	2위	3위	4위	5위	6위	7위	8위	9위	10위	11위	12위
Mumbai	Delhi	Bangalore	Hyderabad	Pune	Kolkata	Ahmedabad	Chennai	Goa	Jaipur	Kochi	Coimbatore

Conclusion



Modeling Conclusion

	Train data Accuracy	Test data r-square	Root mean squared error
LinearRegression	0.78	0.77	5.04
LinearRegression log	0.93	0.92	12.37
RandomForestRegressor	0.98	0.93	2.75
RandomForestRegressor log	0.99	0.94	0.21
DecisionTreeRegressor	1.00	0.88	3.68
MLPRegressor	0.72	0.69	5.83
LGBMRegressor	0.96	0.94	2.62

- 제일 결과가 좋은 모델: Random Forest Regressor

- 이상치 데이터 제거 안하고 선형회귀 분석했을 때: Train data Accuracy < 0
- 이상치 제거 후 : Train data Accuracy = 0.78

→ 하나의 이상 값으로도 결과에 큰 영향을 미침 (데이터 전처리의 중요성)

- 같은 데이터로 다른 모델을 사용했을 때 결과 다름

→ 모델 선택의 중요성

Conclusion

Stats models

- 지역을 Coimbatore or Hyderabad 밖에 예측을 못 하는 경향이 있었음
- 비교적 낮은 가격을 형성하는 차들은 GDP 순위가 낮은 지역으로 추천하였음

Random Forest Regressor

- Stats models과 달리 좀 더 세분화하여 지역을 추천하였음

증빙 자료

감사합니다.