Final Listing Price Prediction for Private Used Car Sellers

Group 2 - Ji-Soo Kim, Tae-Yoon Kim, Jun-Beom Lee, Jin-Joo Yang, Sung-Hyun Kim

0. Recap - Problem statement

Target Audience

Used-car owners

those who want to sell their cars as a used car with a reasonable & profitable price

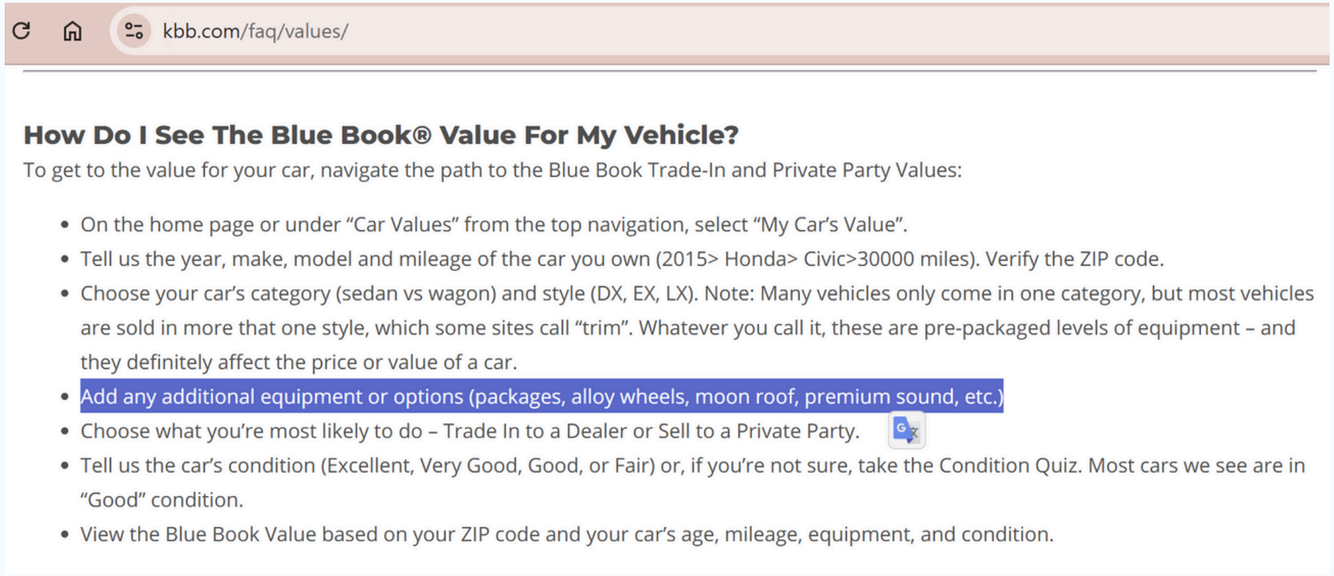
Used-car Dealer

those who want to afford trust to customers

0. Intro - Background research

Used-car Dealer

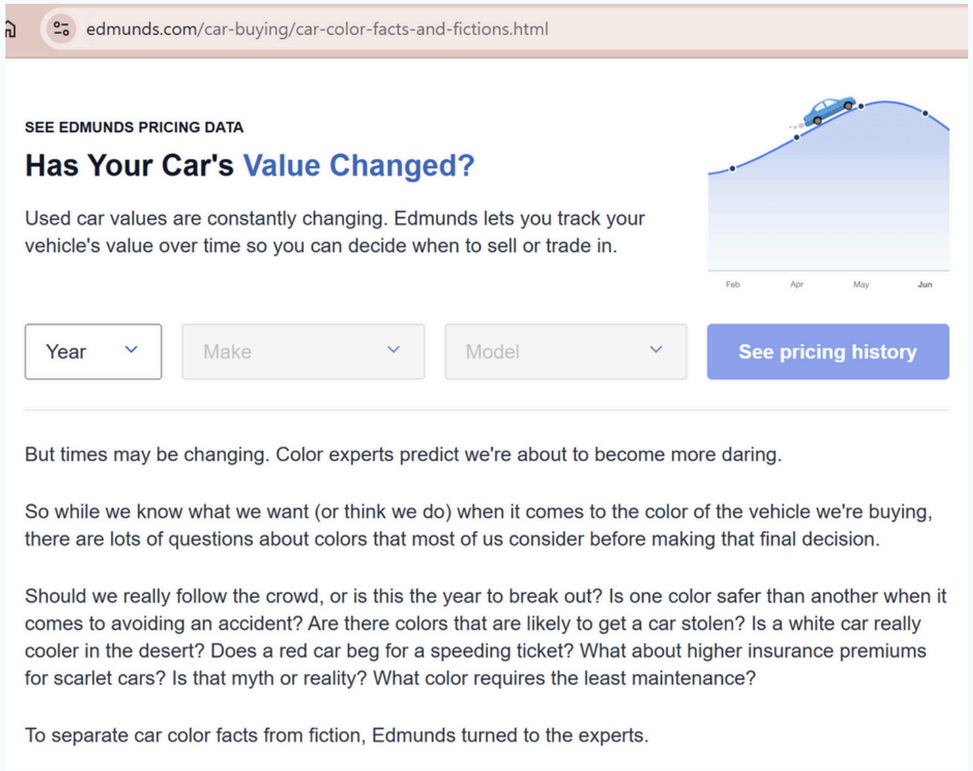
Can build trust with customers using reasonable price



Used-car owners

1. Can predict their used car market price(value) based on sale price prediction

2. Fair Trade Tools for used car owners (Prevent dealer fraud)



1. Feature Preprocessing - 'major options'

<value counts of each options>

	Option	Count
11	Bluetooth	114149
12	Backup Camera	113149
5	Alloy Wheels	107720
10	Heated Seats	66047
3	Navigation System	60560
1	Sunroof/Moonroof	50015
0	Leather Seats	49957
18	Remote Start	45689
7	Blind Spot Monitoring	36890
14	CarPlay	28110
8	Parking Sensors	24677
44	Android Auto	23317
6	Third Row Seating	18526
35	Steel Wheels	16653
4	Adaptive Cruise Control	16598
9	Premium Package	9890
19	Quick Order Package	9888
2	Power Package	7306
29	Tow Package	6803
15	Convenience Package	6503
38	Multi Zone Climate Control	5642
22	Premium Wheels	5605
21	Technology Package	4939
23	Heat Package	4497
51	Chrome Wheels	4091
24	Cold Weather Package	4001
43	Appearance Package	3962
17	Preferred Package	3962
40	Suspension Package	3622
20	Sport Package	3363
13	SE Package	2939
37	Adaptive Suspension	2566
25	LE Package	2521
47	Trailer Package	2264
33	Memory Package	2227
34	Comfort Package	2043
54	Luxury Package	1911
55	Off Road Package	1903
36	Driver Assistance Package	1566

extract only the 'packages'

`unique_filtered_packages`

	make_name	model_name	filtered_packages
6	Hyundai	Elantra	Sport Package
6	Hyundai	Elantra	Audio Package
6	Hyundai	Elantra	Heat Package
6	Hyundai	Elantra	Premium Package
7	Chevrolet	Malibu	Driver Confidence Package
...	...	...	...
1573288	Volkswagen	Jetta Hybrid	Premium Audio Package
1573288	Volkswagen	Jetta Hybrid	Heat Package
1573288	Volkswagen	Jetta Hybrid	Audio Package
1573352	GMC	Yukon Hybrid	Comfort Package
1573352	GMC	Yukon Hybrid	Convenience Package

11732 rows × 3 columns

1. Feature Preprocessing - ‘major options’

Binary encoding

	major_options	\
2	[Alloy Wheels, Bluetooth, Backup Camera, Heate...	
5	[Leather Seats, Sunroof/Moonroof, Navigation S...	
9	[Leather Seats, Navigation System, Adaptive Cr...	
10	[Leather Seats]	
12	[Sunroof/Moonroof, Alloy Wheels, Bluetooth]	
23	[Leather Seats, Navigation System, Adaptive Cr...	
36	[Sport Package, Sunroof/Moonroof, Adaptive Cru...	
38	[Driver Confidence Package, Power Package, Pre...	
40	[Leather Seats, Sunroof/Moonroof, Navigation S...	
41	[Power Package, Preferred Package, Third Row S...	

	binary_encoded
2	00000000000000001000000100100000000000000000...
5	00000000000010001000000100100000000000000000...
9	00000000000010001000000101100000000000000000...
10	00000000000000000000000000000000000000000000...
12	00000000000000001000000001000000000000000000...
23	00000000000010001000000101100000000000000000...
36	00000000000010000000101011010000000000000000...
38	00000000000000000010000101101000000100000100...
40	00000000000000001000000100100000000000000000...
41	00000000000000000010000101101000000100000000...

Frequency Counts

	packages	package_score
38	[Driver Confidence Package, Power Package, Pre...	6
40	[]	0
41	[Power Package, Preferred Package, Heat Packag...	5
45	[Power Package, Preferred Package, Technology ...	7
47	[Power Package, Preferred Package, Heat Packag...	6

binary_encoded	feature_1	feature_2	feature_3	feature_4	feature_5	...
"1010110"	1	0	1	0	1	...
"0001011"	0	0	0	1	0	...

1. Feature Preprocessing - 'interior_color'

Interior_color vs listing_color

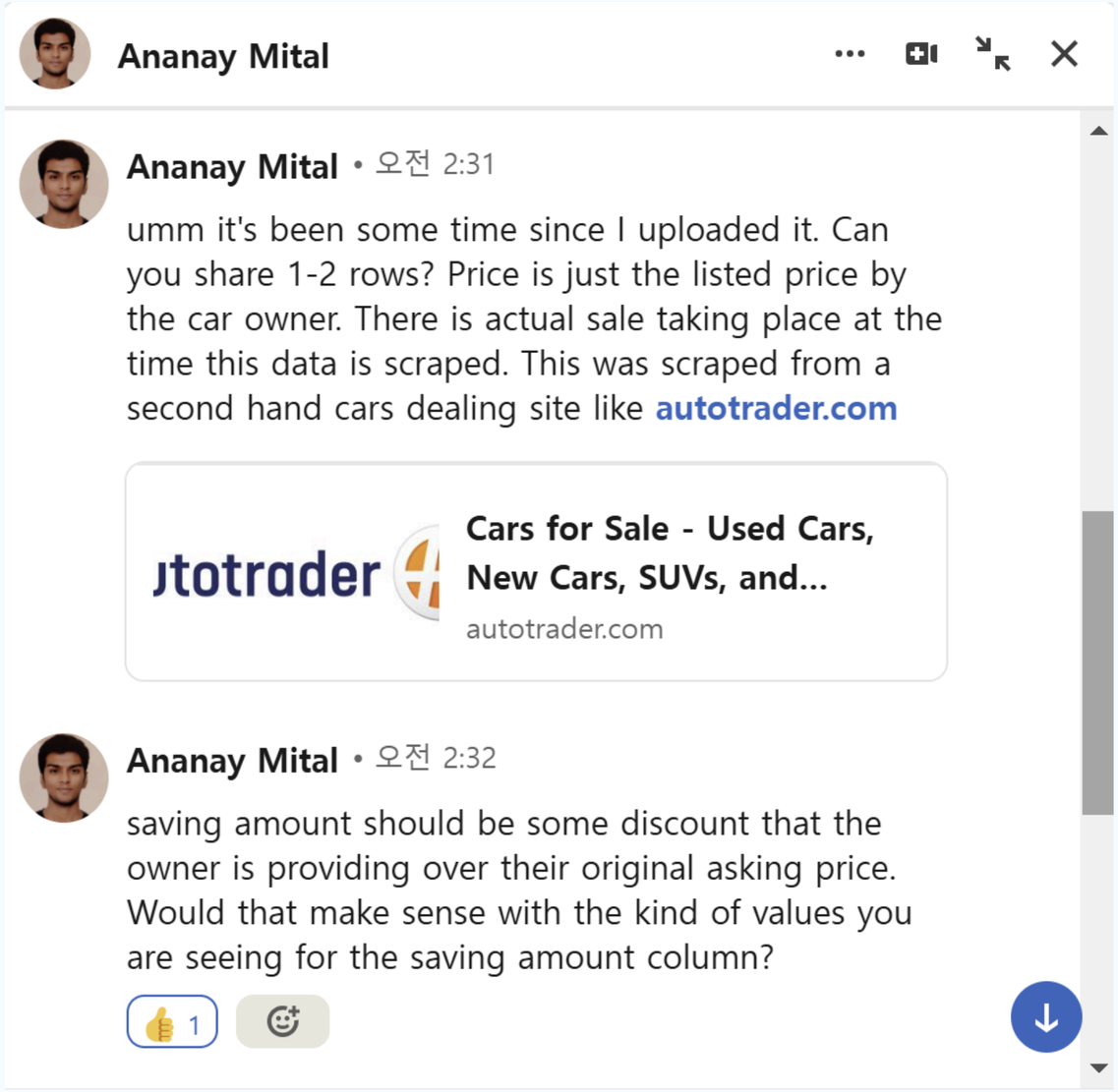
interior_color	
Black	871393
Gray	195908
Jet Black	186195
Black (Ebony)	142839
Black (Charcoal)	112051
...	
Brown (Cappuccino w/Heated Lincoln Soft Touch Front Seats)	1
Circuit Red Nuluxe[nuluxe] With Dark Gray Streamli	1
Nut Brown/ Black Leather	1
Black/Orange w/Fabric Seat Trim (FD)	1
Brown (Espresso/IV/Tan/Esp/IV/IV)	1

Two Tone colors?
-> process as a separate value

reference?

listing_color	
WHITE	666564
BLACK	587999
UNKNOWN	399905
SILVER	384779
GRAY	377442
RED	252917
BLUE	249758
GREEN	24074
BROWN	22611
ORANGE	11631
GOLD	10297
TEAL	5453
YELLOW	5003
PURPLE	1468
PINK	139

1. Feature Preprocessing - ‘price’ & ‘savings_amount’ ?



Price : Price ‘posted’ on used car site (sold price X)

Savings_amount : A reduced price (if the price is lowered) compared to the original price that was initially posted

1. Feature Preprocessing - ‘price’ & ‘savings_amount’ ?

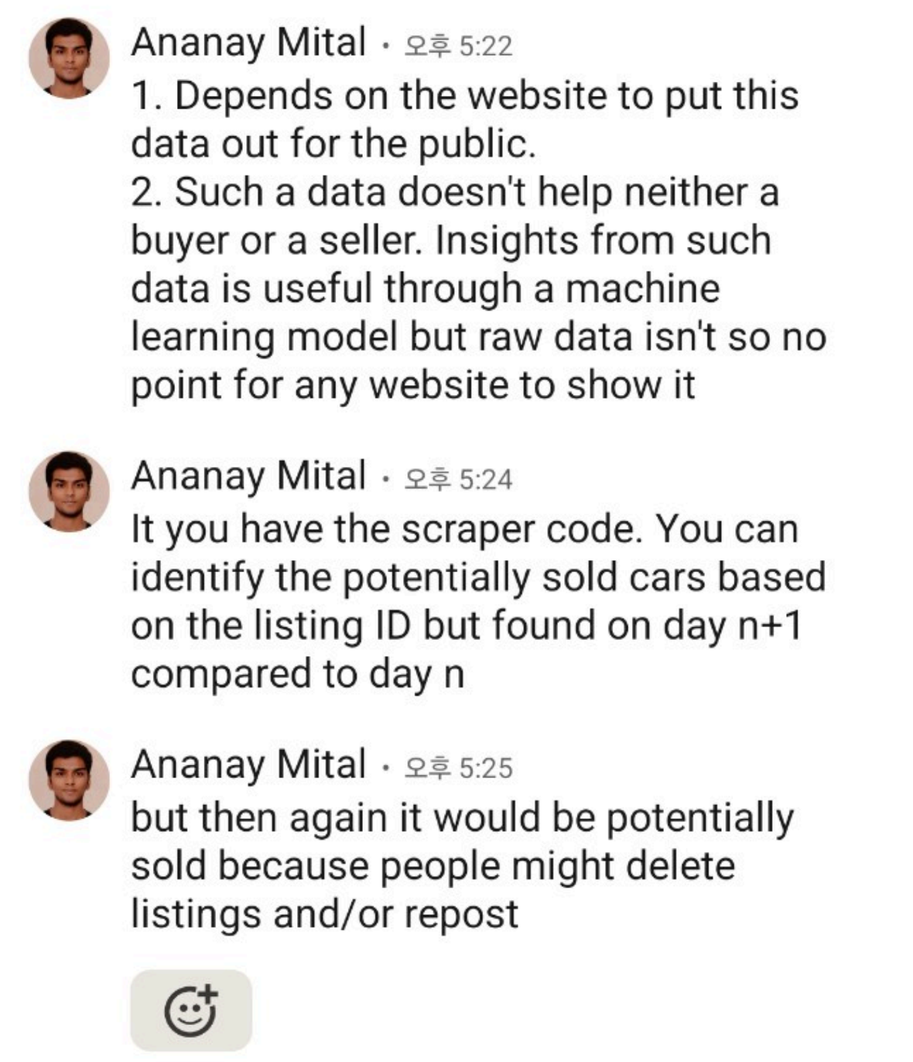
	price	savings_amount	price2
38	14639	1749	12890
40	32000	1861	30139
41	23723	3500	20223
45	22422	2416	20006
47	29424	2254	27170
...	...	...	...
3000033	5371	453	4918
3000034	40993	2220	38773
3000035	17998	381	17617
3000038	26998	849	26149
3000039	19900	1203	18697

[1094520 rows x 3 columns]

price2 = price - savings_amount

Replace ‘price’ feature with price2
=> so that the price can be predicted as close to the **'selling price'** as possible.

2. Our specific topic?



Using scraping code : Can check whether a specific vehicle existed on a specific date (n days) based on the list ID, and if the ID disappears on the next day (n+1 days), the vehicle is likely to have been sold.

However, since people can simply delete or re-upload a vehicle listing, **we cannot be sure that it was sold** simply because the listID disappears.

=> Hard to figure out whether the car is sold or not...😭

2. Our specific topic?

‘Final listed’ Price Prediction of Private Sales Vehicles with Used Car Datas

Because the ‘price’ feature is not the **“sold”** price

it’s the **“listed price”** on the website…

~~CPI ratio~~

~~whether cars are sold~~

3. Data preprocessing - Missing value

vehicle_damage_category	100.000000
combine_fuel_economy	100.000000
is_certified	100.000000
bed	99.347742
cabin	97.882262
is_oemcpo	95.487993
is_cpo	93.903481
bed_height	85.696924
bed_length	85.696924
owner_count	50.566426
fleet	47.552533
theft_title	47.552533
isCab	47.552533
has_accidents	47.552533
frame_damaged	47.552533
salvage	47.552533
franchise_make	19.087579
torque	17.259537
highway_fuel_economy	16.375948
city_fuel_economy	16.375948
power	16.047319
interior_color	12.799363
main_picture_url	12.302936
major_options	6.668178
engine_displacement	5.746123
horsepower	5.746123
back_legroom	5.308896
wheelbase	5.308896
maximum_seating	5.308896
width	5.308896
length	5.308896
height	5.308896
front_legroom	5.308896
fuel_tank_volume	5.308896
wheel_system_display	4.891001
wheel_system	4.891001
mileage	4.812836
trim_name	3.876415
trimId	3.860849
engine_type	3.352655
engine_cylinders	3.352655
fuel_type	2.757430
description	2.596665
transmission	2.139471
transmission_display	2.139471
exterior_color	1.665144
seller_rating	1.362382
body_type	0.451427
sp_id	0.003200
sp_name	0.000000
vin	0.000000
savings_amount	0.000000
price	0.000000
model_name	0.000000
make_name	0.000000
longitude	0.000000
listing_id	0.000000
listing_color	0.000000
listed_date	0.000000
latitude	0.000000
is_new	0.000000
franchise_dealer	0.000000
dealer_zip	0.000000
daysonmarket	0.000000
city	0.000000
year	0.000000

dtype: float64

Missing value ratio of each features

85% up : delete

85% down : consider separately

3. Data preprocessing - Missing value

bed_height	85.696924
bed_length	85.696924
owner_count	50.566426
fleet	47.552533
theft_title	47.552533
isCab	47.552533
has_accidents	47.552533
frame_damaged	47.552533
salvage	47.552533
franchise_make	19.087579
torque	17.259537
highway_fuel_economy	16.375948

same ratio of missing values

```python
columns_to_check = ['fleet', 'theft_title','isCab','has_accidents','frame_damaged','salvage']

# extract the indexes of the data samples including missing values of each features
missing_indices = [set(df[df[col].isna()].index) for col in columns_to_check]

# check whether the rows including missing values are all same among all features
all_equal = all(missing_indices[0] == indices for indices in missing_indices)

# outcome
if all_equal:
    print("Rows including missing values are all the same.")
else:
    print("Rows including missing values are not same.")
```

```
Rows including missing values are all the same.
```

3. Data preprocessing - Missing value

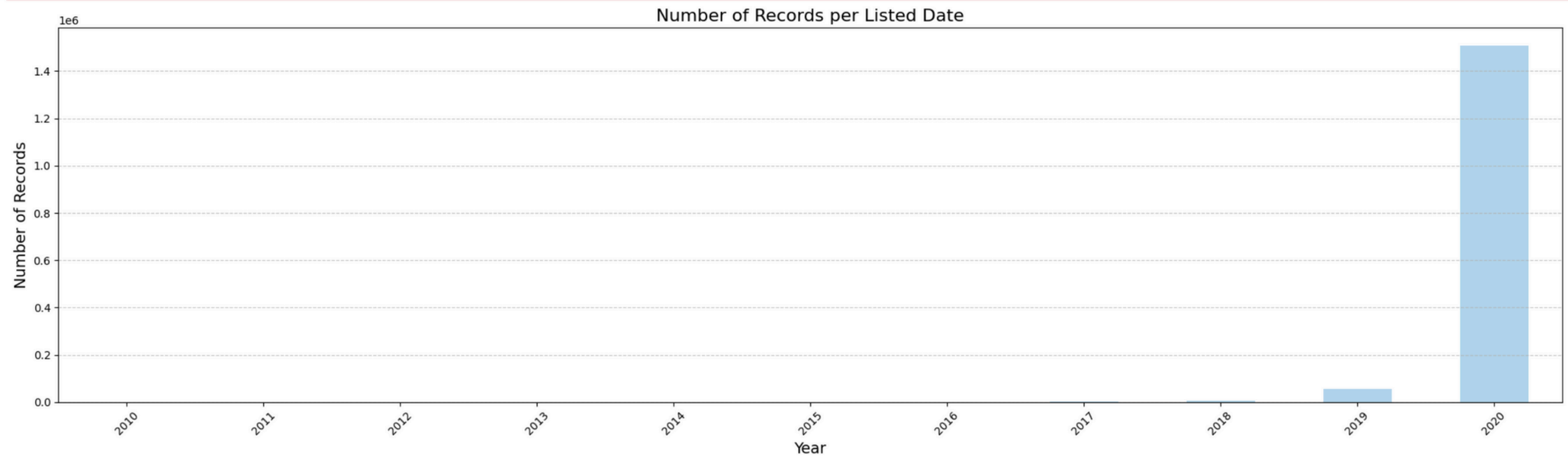
	Missing Values	Percentage (%)
city_fuel_economy	246997	15.697848
highway_fuel_economy	246997	15.697848
interior_color	243832	15.496697
torque	201806	12.825742
owner_count	90418	5.746499
major_options	70971	4.510549
maximum_seating	67164	4.268595
engine_displacement	63925	4.062741
horsepower	63925	4.062741
wheel_system	50942	3.237609
trim_name	44536	2.830477
exterior_color	35323	2.244947
fuel_type	32573	2.070171
transmission	26791	1.702697
mileage	19390	1.232328
body_type	1862	0.118339
price	0	0.000000
salvage	0	0.000000
model_name	0	0.000000
theft_title	0	0.000000
savings_amount	0	0.000000
listed_date	0	0.000000
make_name	0	0.000000
listing_color	0	0.000000
city	0	0.000000
isCab	0	0.000000
has_accidents	0	0.000000
franchise_dealer	0	0.000000
frame_damaged	0	0.000000
fleet	0	0.000000
engine_type	0	0.000000
dealer_zip	0	0.000000
daysonmarket	0	0.000000
year	0	0.000000

1,500,000 data samples

	Missing Values	Percentage (%)
interior_color	192683	14.526238
torque	129741	9.781085
owner_count	71963	5.425241
major_options	46695	3.520304
exterior_color	26862	2.025108
mileage	14944	1.126618
maximum_seating	10500	0.791588
transmission	10372	0.781938
engine_displacement	8488	0.639904
horsepower	8488	0.639904
wheel_system	818	0.061668
fuel_type	16	0.001206
body_type	9	0.000679
price	0	0.000000
make_name	0	0.000000
savings_amount	0	0.000000
model_name	0	0.000000
theft_title	0	0.000000
trim_name	0	0.000000
salvage	0	0.000000
listed_date	0	0.000000
listing_color	0	0.000000
city	0	0.000000
isCab	0	0.000000
highway_fuel_economy	0	0.000000
has_accidents	0	0.000000
franchise_dealer	0	0.000000
frame_damaged	0	0.000000
fleet	0	0.000000
engine_type	0	0.000000
dealer_zip	0	0.000000
daysonmarket	0	0.000000
city_fuel_economy	0	0.000000
year	0	0.000000

1,300,000 data samples

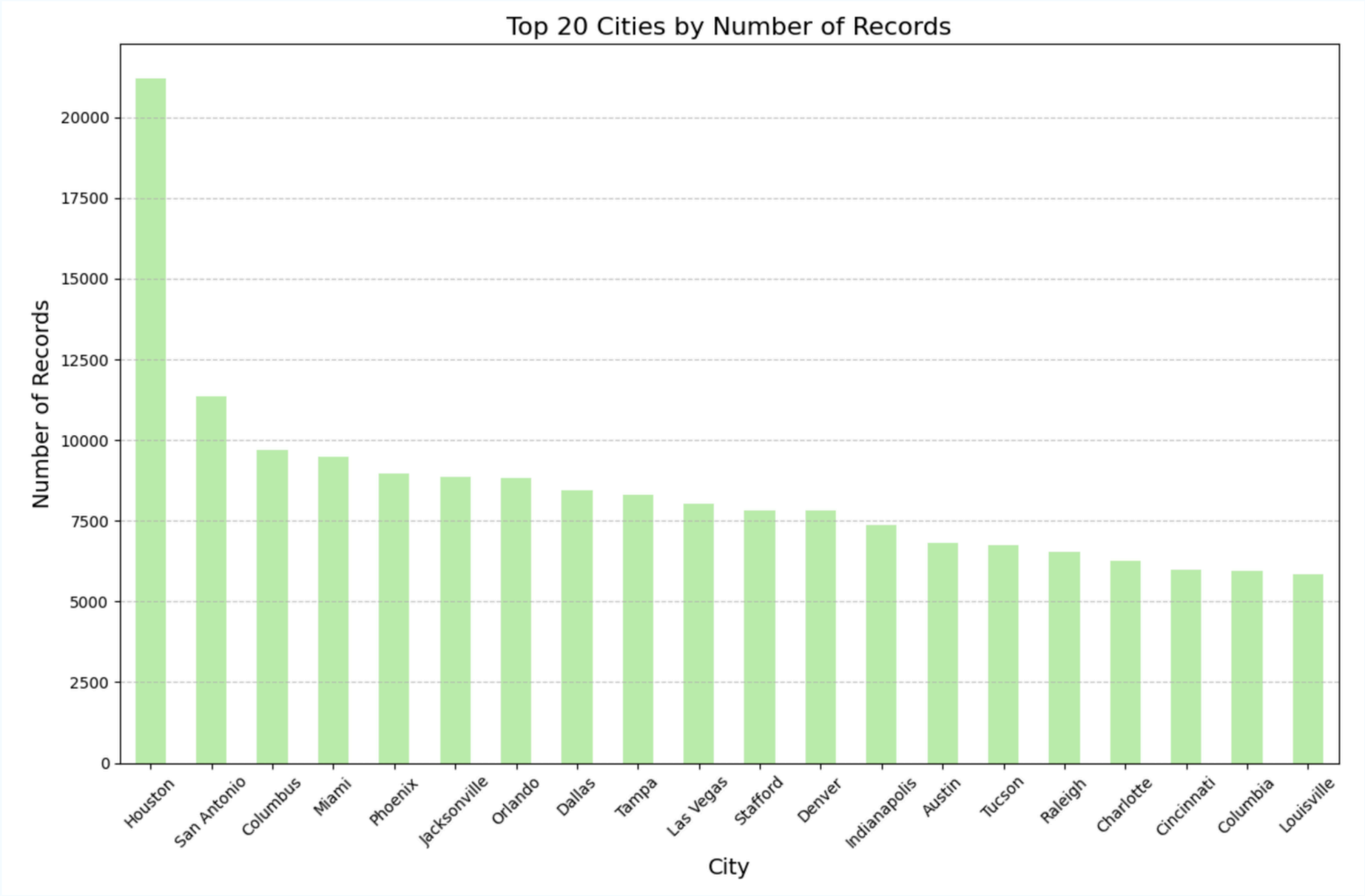
3. Data preprocessing - Listed Date per Year



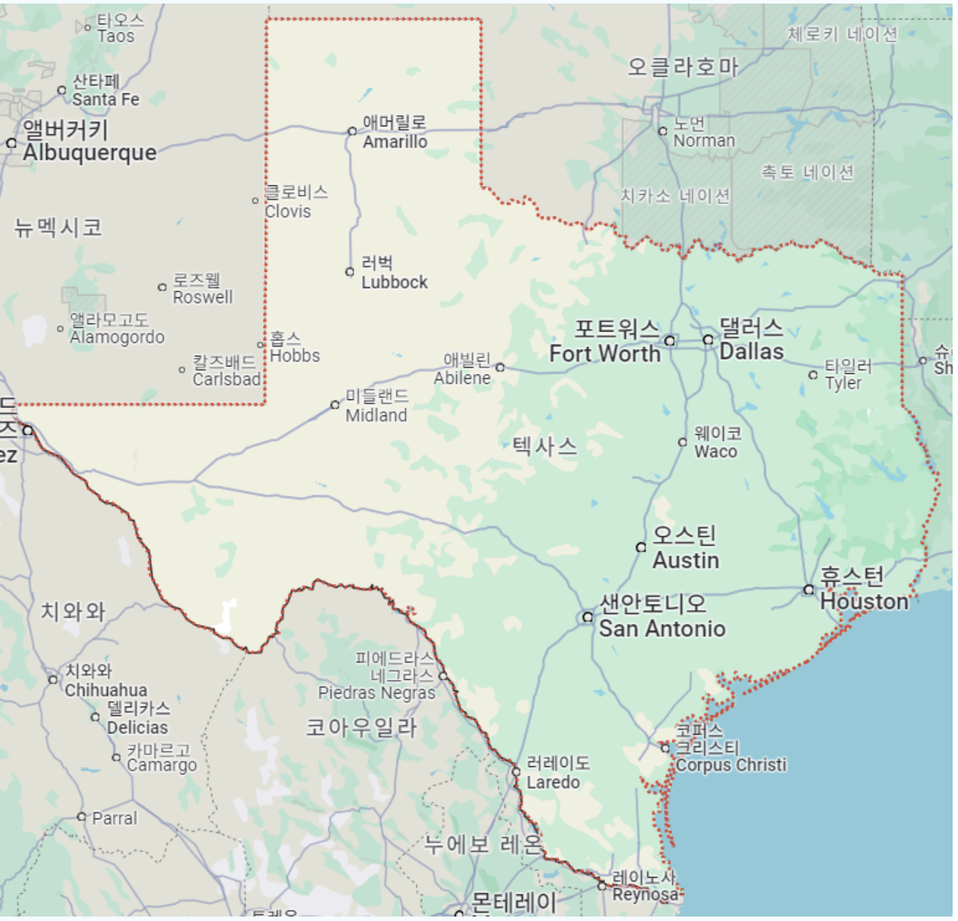
Our range of Year :

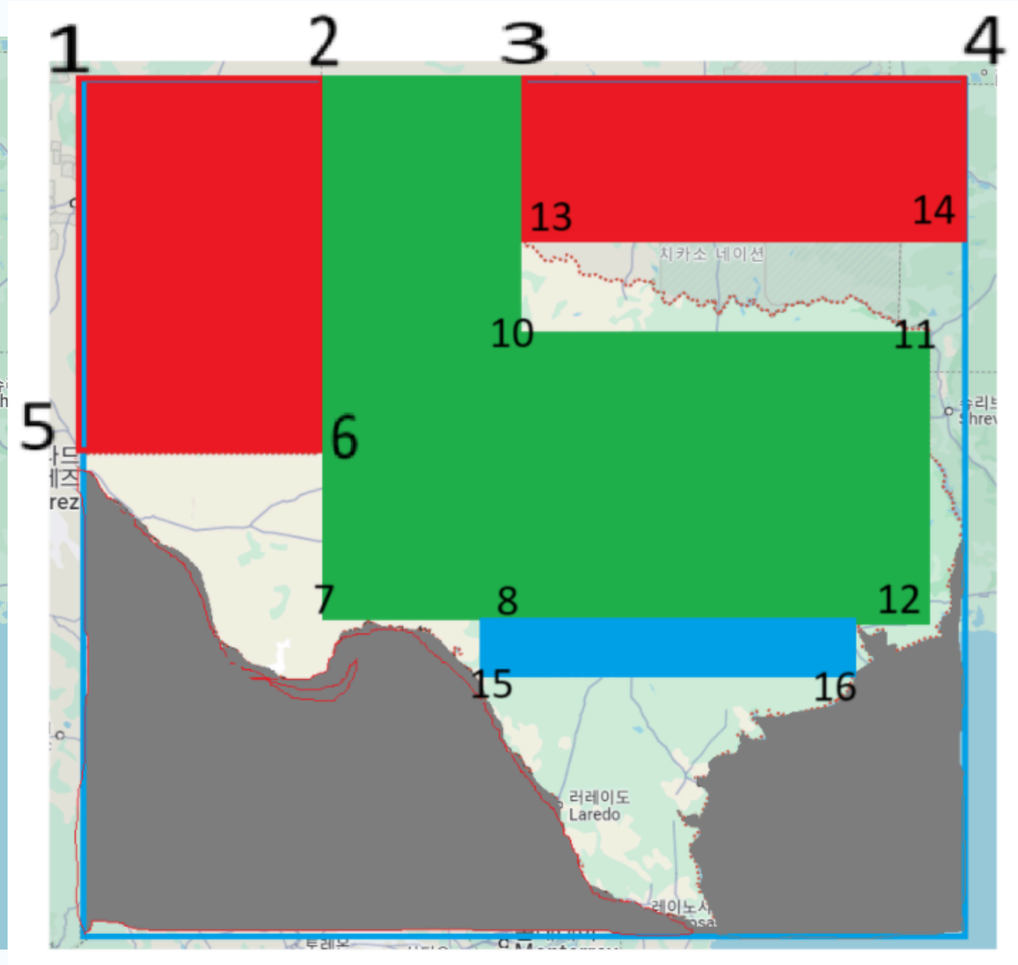
3-9-2019 ~ 2-9-2020

3. Data preprocessing - Top 20 Cities



3. Data preprocessing - city : “Texas”



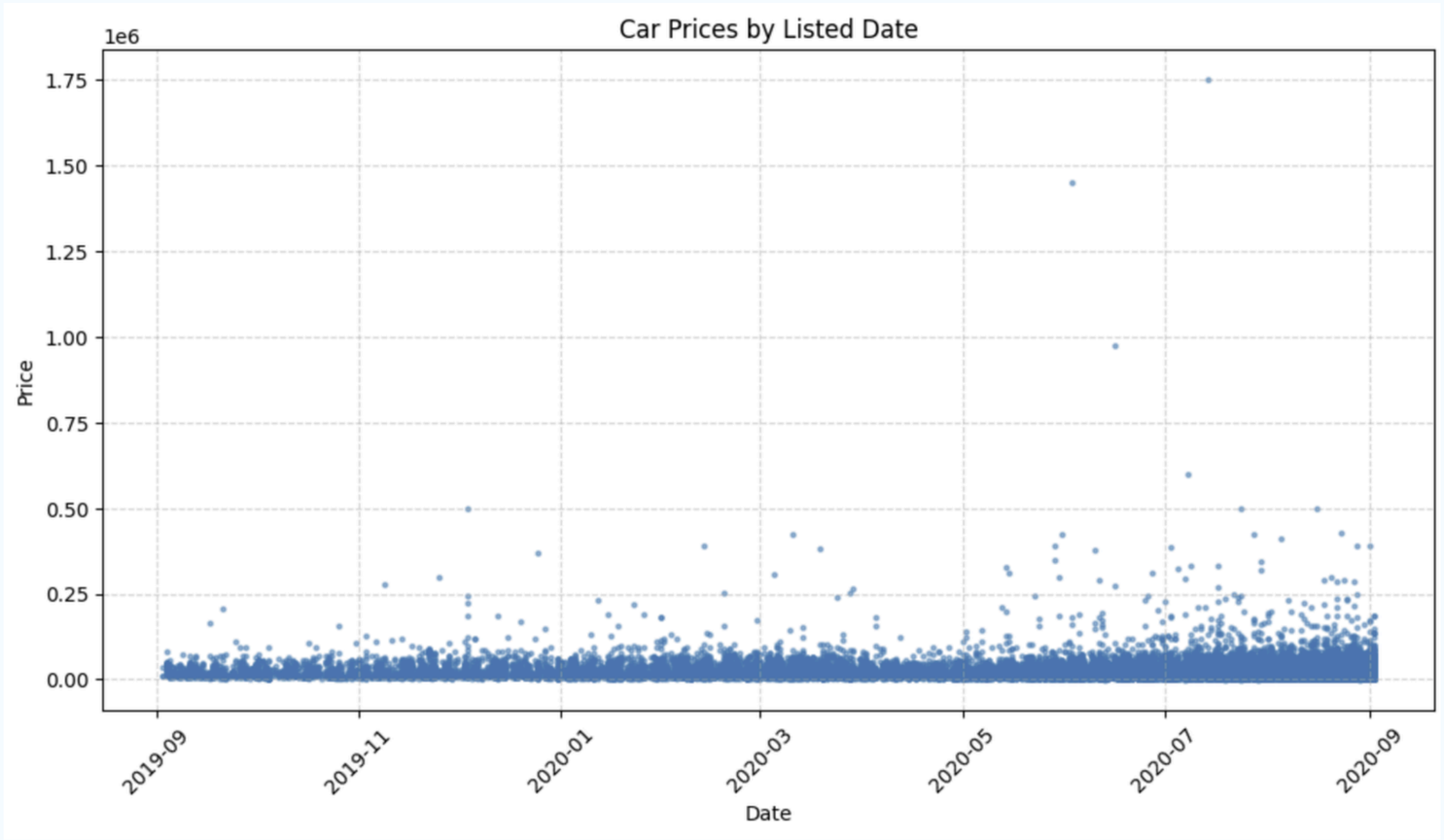


(latitude, longitude)

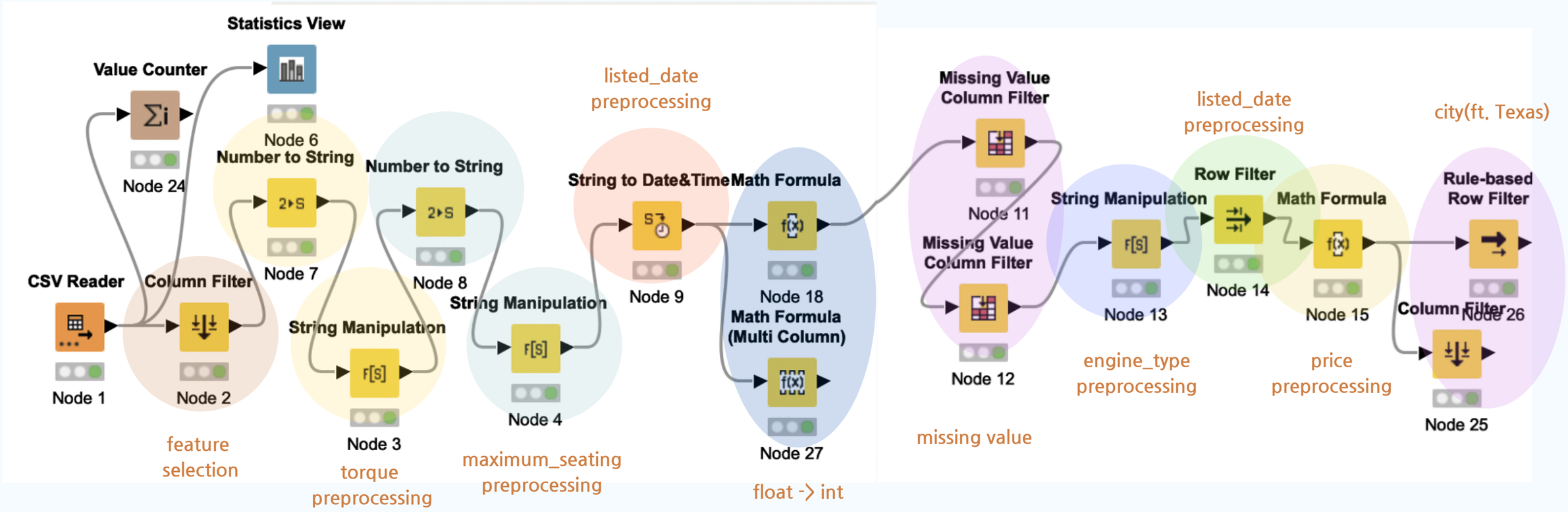
1: (36.49, -106.59)
2: (36.49, -103.02)
3: (36.49, -100.01)
4: (36.49, -93.5)
5: (32,-106.59)
6: (32, -103.05)
7: (29.88, -103.05)
8: (29.88, -100.01)
10: (33.55, -100.01)
11: (33.55, -94.05)
12: (29.88, -94.05)
13: (34.58, -100)
14: (34.58, -93.5)
15: (29.2, -100.01)
16: (29.2, -95.2)

3. Data preprocessing - city : “Texas”

	A	B	C	D
1	latitude	longitude	city	texas
2	35.1901	-106.659	Albuquerque	0
3	27.7626	-98.048	Alice	1
4	29.1903	-95.4048	Angleton	1
5	27.9369	-97.1482	Aransas Pass	1
6	34.1906	-97.1593	Ardmore	0
7	34.3689	-96.1389	Atoka	0
8	29.5065	-94.9838	Bacliff	1
9	28.9902	-95.9268	Bay City	1
10	29.7766	-94.9634	Baytown	1
11	28.3975	-97.7467	Beeville	1
12	32.6984	-93.7398	Benton	0
13	33.5933	-96.1774	Bonham	1
14	34.8519	-106.691	Bosque Farms	0
15	32.5657	-93.7288	Bossier City	0
16	33.5862	-97.914	Bowie	1
17	34.0115	-94.7396	Broken Bow	0
18	25.9952	-97.4846	Brownsville	1
19	34.078	-98.5563	Burkburnett	1
20	31.8961	-106.599	Canutillo	1



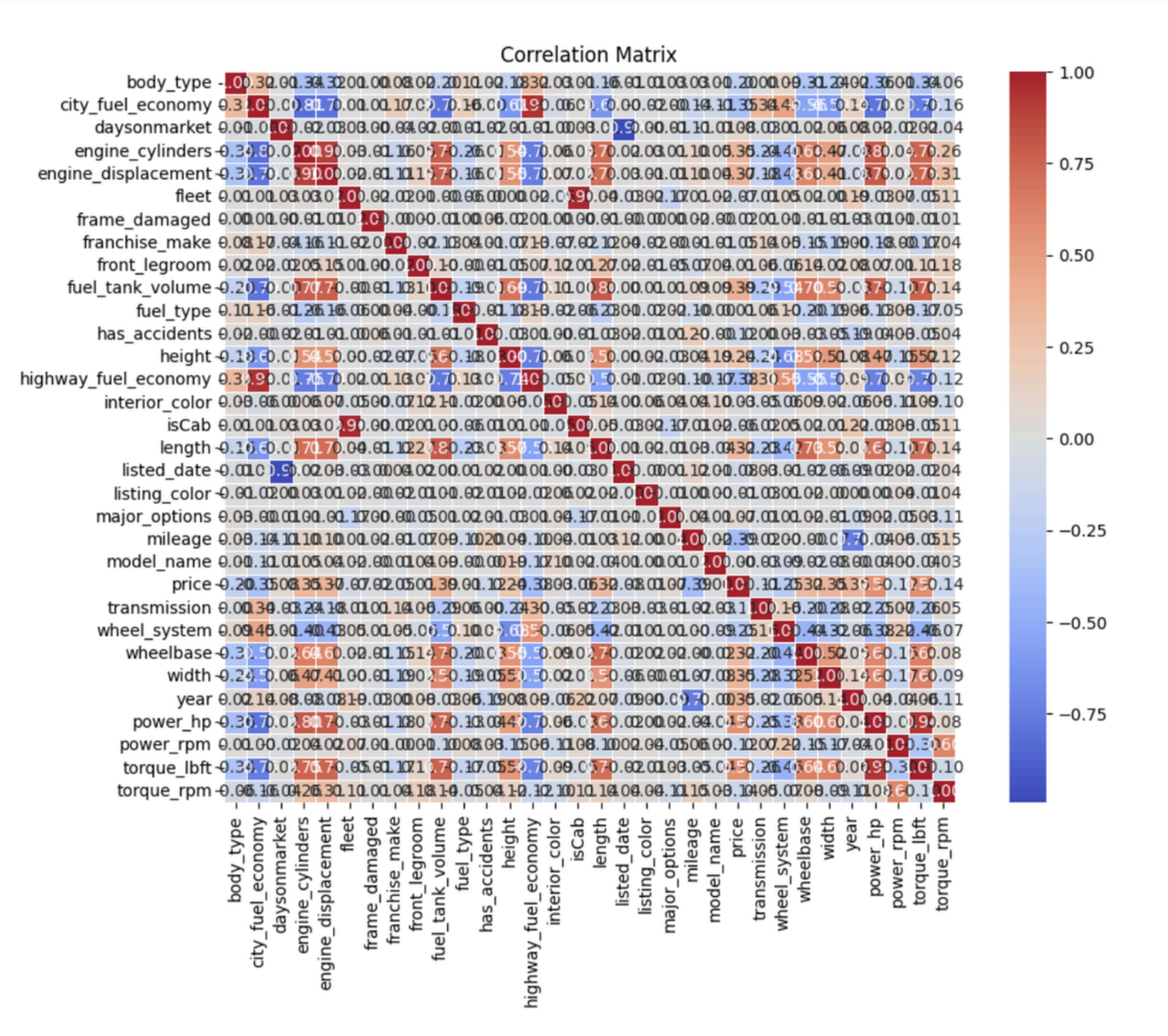
4. Brief KNIME Workflow



Q & A

Thank you for listening!

Appendix


Correlation Matrix
body_type
city_fuel_economy
daysonmarket
engine_cylinders
engine_displacement
fleet
frame_damaged
franchise_make
front_legroom
fuel_tank_volume
fuel_type
has_accidents
height
highway_fuel_economy
interior_color
isCab
length
listed_date
listing_color
major_options
mileage
model_name
price
transmission
wheel_system
wheelbase
width
year
power_hp
power_rpm
torque_lbft
torque_rpm
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75