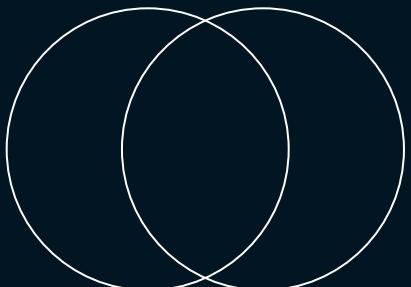


Final Listing Price Prediction for Private Used Car Sellers

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)

kbb.com/faq/values/

How Do I See The Blue Book® Value For My Vehicle?

To get to the value for your car, navigate the path to the Blue Book Trade-In and Private Party Values:

- On the home page or under "Car Values" from the top navigation, select "My Car's Value".
- Tell us the year, make, model and mileage of the car you own (2015> Honda> Civic>30000 miles). Verify the ZIP code.
- Choose your car's category (sedan vs wagon) and style (DX, EX, LX). Note: Many vehicles only come in one category, but most vehicles are sold in more than one style, which some sites call "trim". Whatever you call it, these are pre-packaged levels of equipment – and they definitely affect the price or value of a car.
- Add any additional equipment or options (packages, alloy wheels, moon roof, premium sound, etc.)
- Choose what you're most likely to do – Trade In to a Dealer or Sell to a Private Party.
- Tell us the car's condition (Excellent, Very Good, Good, or Fair) or, if you're not sure, take the Condition Quiz. Most cars we see are in "Good" condition.
- View the Blue Book Value based on your ZIP code and your car's age, mileage, equipment, and condition.

edmunds.com/car-buying/car-color-facts-and-fictions.html

SEE EDMUNDS PRICING DATA

Has Your Car's Value Changed?

Used car values are constantly changing. Edmunds lets you track your vehicle's value over time so you can decide when to sell or trade in.

Feb Apr May Jun

Year Make Model See pricing history

But times may be changing. Color experts predict we're about to become more daring.

So while we know what we want (or think we do) when it comes to the color of the vehicle we're buying, there are lots of questions about colors that most of us consider before making that final decision.

Should we really follow the crowd, or is this the year to break out? Is one color safer than another when it comes to avoiding an accident? Are there colors that are likely to get a car stolen? Is a white car really cooler in the desert? Does a red car beg for a speeding ticket? What about higher insurance premiums for scarlet cars? Is that myth or reality? What color requires the least maintenance?

To separate car color facts from fiction, Edmunds turned to the experts.

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  000000000000000010000010010000000000000000...
5  000000000010001000001001000000000000000000...
9  00000000001000100000101100000000000000000...
10  000000000000000000000000000000000000000000...
12  000000000000000010000000010000000000000000...
23  000000000001000100000010110000000000000000...
36  000000000001000000001010110100000000000000...
38  000000000000000010000101101000001000001000...
40  000000000000000010000001001000000000000000...
41  00000000000000001000010110100000100000000...
```

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’ ?

Ananay Mital • 오전 2:31
umm it's been some time since I uploaded it. Can you share 1-2 rows? Price is just the listed price by the car owner. There is actual sale taking place at the time this data is scraped. This was scraped from a second hand cars dealing site like [autotrader.com](#)

autotrader Cars for Sale - Used Cars, New Cars, SUVs, and... autotrader.com

Ananay Mital • 오전 2:32
saving amount should be some discount that the owner is providing over their original asking price. Would that make sense with the kind of values you are seeing for the saving amount column?

1

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]

$$\text{price2} = \text{price} - \text{savings_amount}$$

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

2. Our specific topic?



Ananay Mital · 오후 5:22

1. Depends on the website to put this data out for the public.
2. Such a data doesn't help neither a buyer or a seller. Insights from such data is useful through a machine learning model but raw data isn't so no point for any website to show it



Ananay Mital · 오후 5:24

If you have the scraper code. You can identify the potentially sold cars based on the listing ID but found on day n+1 compared to day n



Ananay Mital · 오후 5:25

but then again it would be potentially sold because people might delete listings and/or repost



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...



3. Data preprocessing - Missing value

vehicle_damage_category	100.000000	wheel_system	4.891001
combine_fuel_economy	100.000000	mileage	4.812836
is_certified	100.000000	trim_name	3.876415
bed	99.347742	trimId	3.860849
cabin	97.882262	engine_type	3.352655
is_oemcpo	95.487993	engine_cylinders	3.352655
is_cpo	93.903481	fuel_type	2.757430
bed_height	85.696924	description	2.596665
bed_length	85.696924	transmission	2.139471
owner_count	50.566426	transmission_display	2.139471
fleet	47.552533	exterior_color	1.665144
theft_title	47.552533	seller_rating	1.362382
isCab	47.552533	body_type	0.451427
has_accidents	47.552533	sp_id	0.003200
frame_damaged	47.552533	sp_name	0.000000
salvage	47.552533	vin	0.000000
franchise_make	19.087579	savings_amount	0.000000
torque	17.259537	price	0.000000
highway_fuel_economy	16.375948	model_name	0.000000
city_fuel_economy	16.375948	make_name	0.000000
power	16.047319	longitude	0.000000
interior_color	12.799363	listing_id	0.000000
main_picture_url	12.302936	listing_color	0.000000
major_options	6.668178	listed_date	0.000000
engine_displacement	5.746123	latitude	0.000000
horsepower	5.746123	is_new	0.000000
back_legroom	5.308896	franchise_dealer	0.000000
wheelbase	5.308896	dealer_zip	0.000000
maximum_seating	5.308896	daysonmarket	0.000000
width	5.308896	city	0.000000
length	5.308896	year	0.000000
height	5.308896		
front_legroom	5.308896	dtype: float64	
fuel_tank_volume	5.308896		
wheel_system_display	4.891001		

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

```
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]

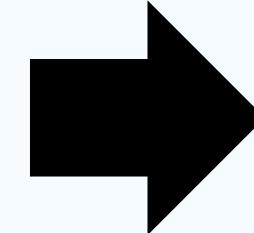
# check whether the rows including missing values are all same among all features
all_equal = all(missing_indices[0] == indices for indices in missing_indices)

# 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	

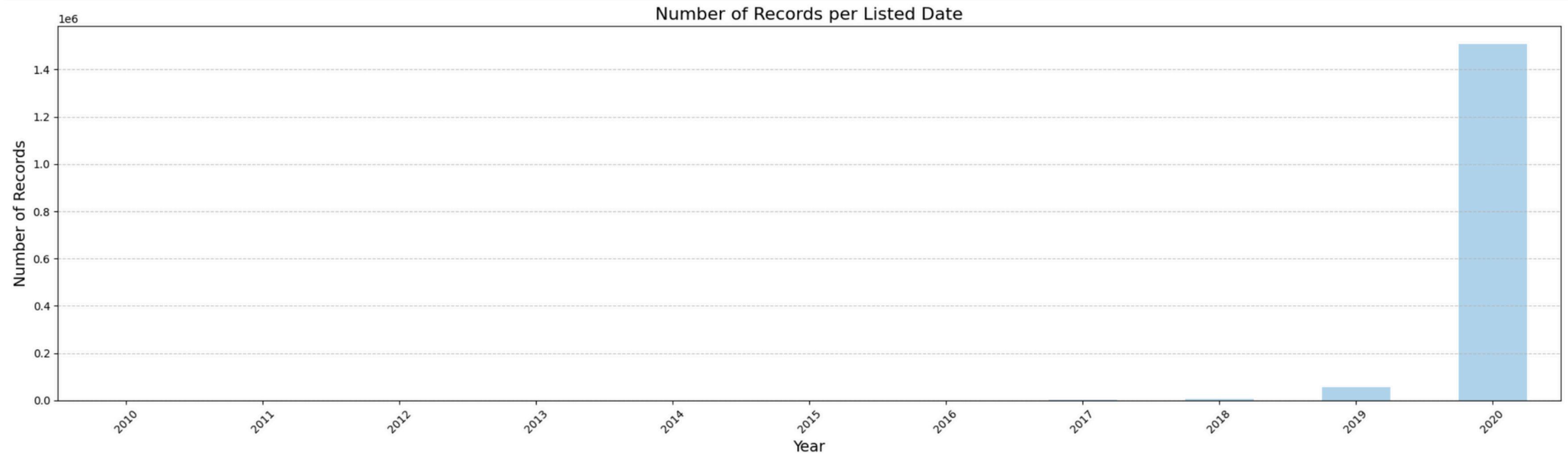


1,500,000 data samples

	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	

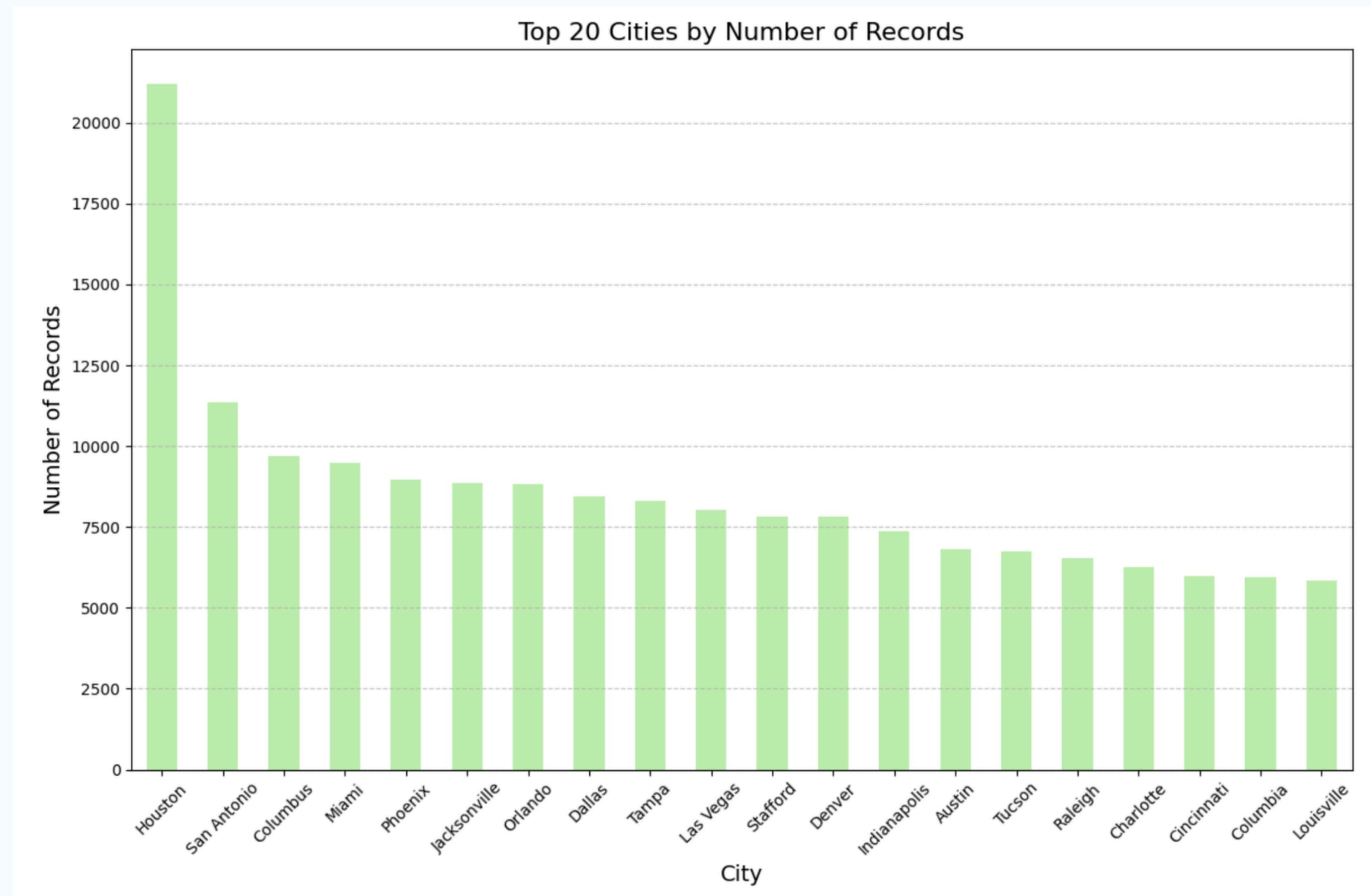
1,300,000 data samples

3. Data preprocessing - Listed Date per Year



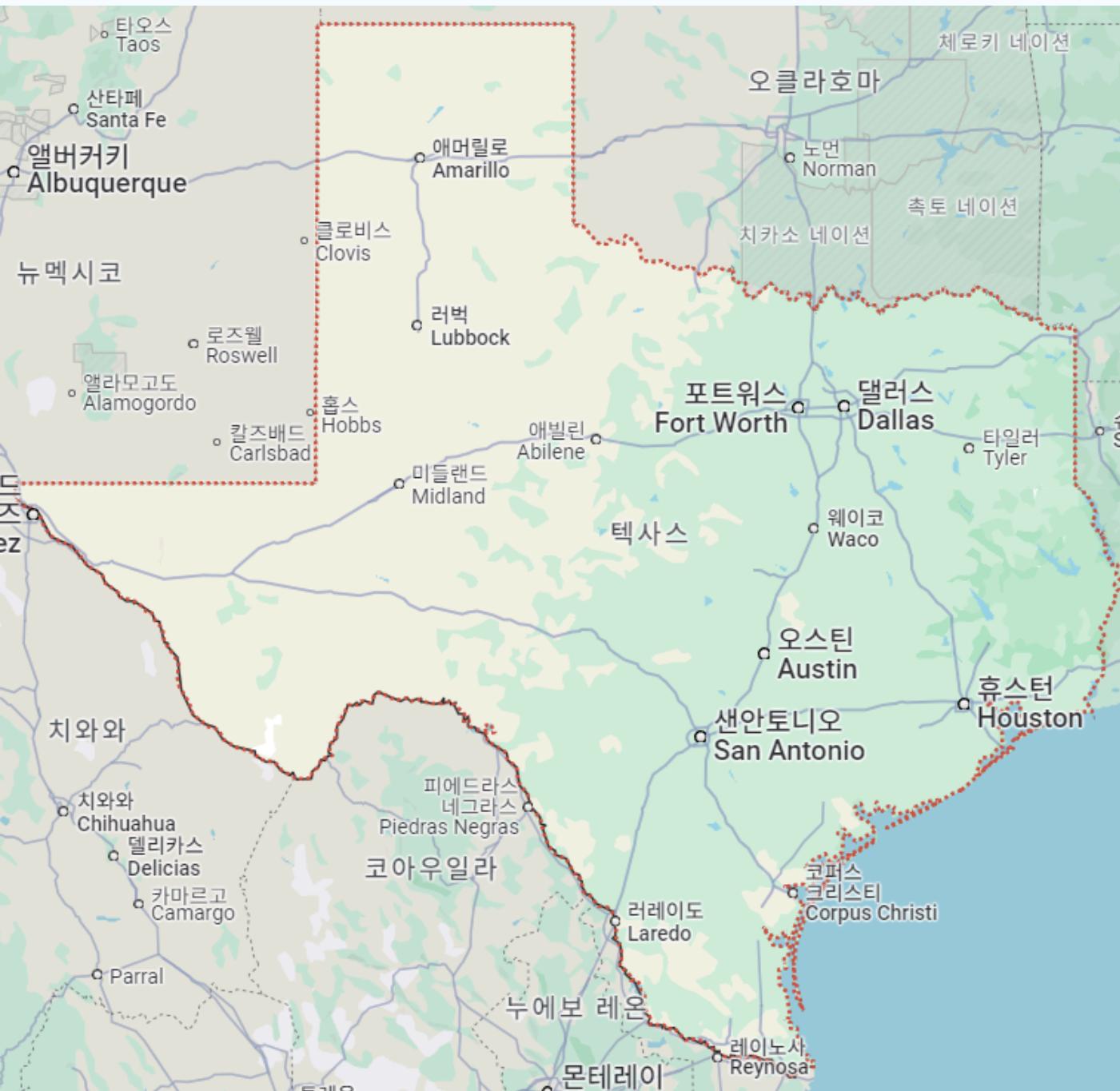
Our range of Year :
3-9-2019 ~ 2-9-2020

3. Data preprocessing - Top 20 Cities



3. Data preprocessing - city : “Texas”

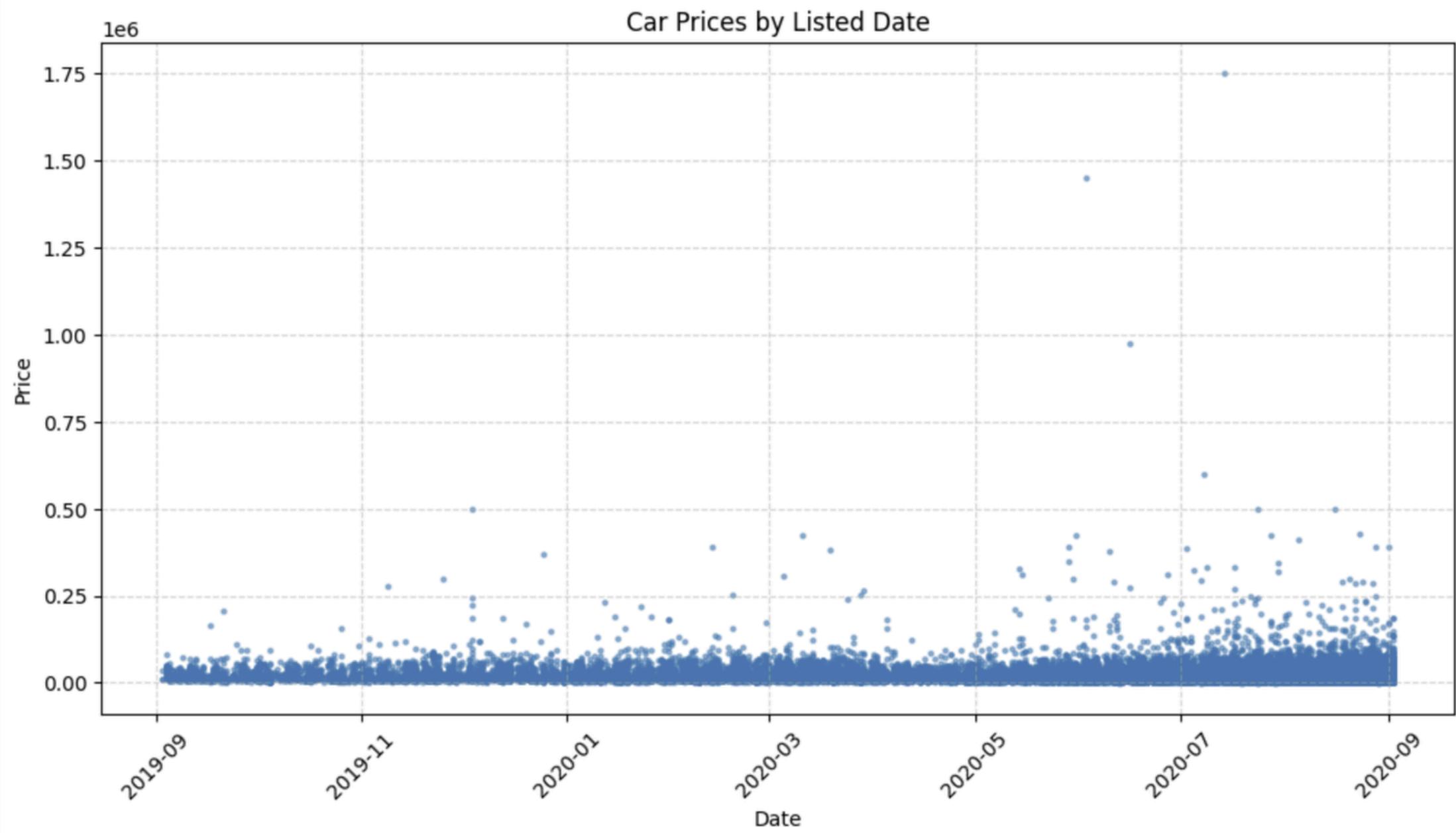
(latitude, longitude)



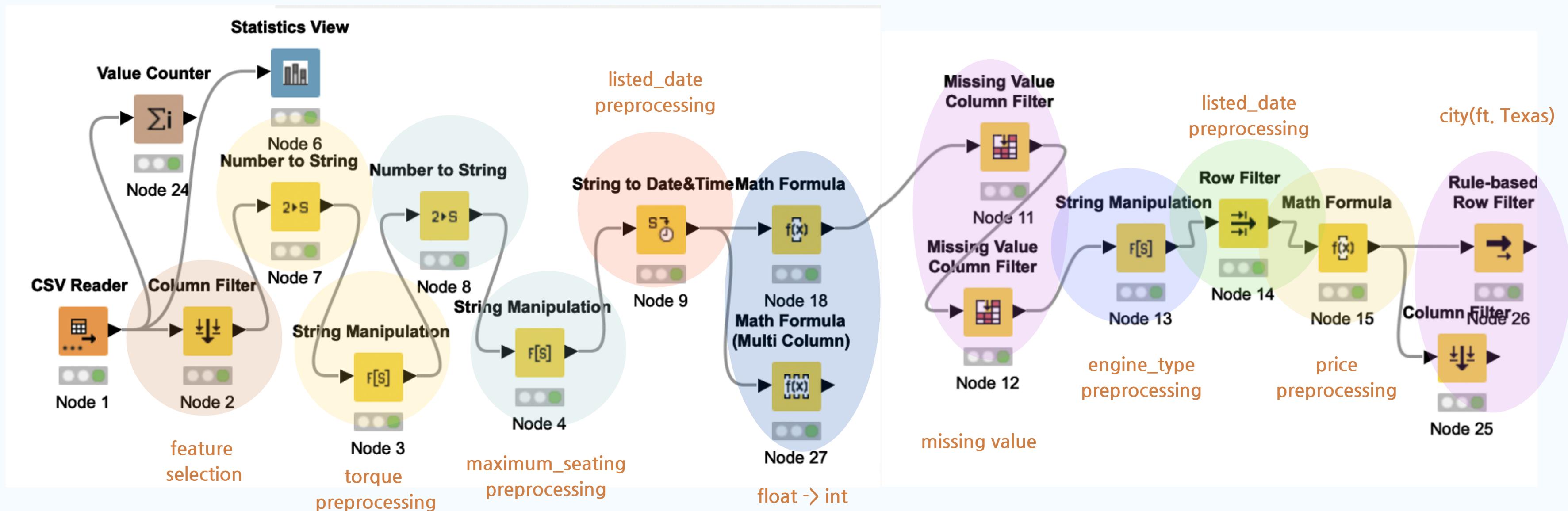
- 1: (36.49, -106.59)
- 2: (36.49, -103.02)
- 3: (36.49, -100.01)
- 4: (36.49, -93.5)
- 5: (32, -106.59)
- 6: (32, -103.05)
- 7: (29.88, -103.05)
- 8: (29.88, -100.01)
- 10: (33.55, -100.01)
- 11: (33.55, -94.05)
- 12: (29.88, -94.05)
- 13: (34.58, -100)
- 14: (34.58, -93.5)
- 15: (29.2, -100.01)
- 16: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

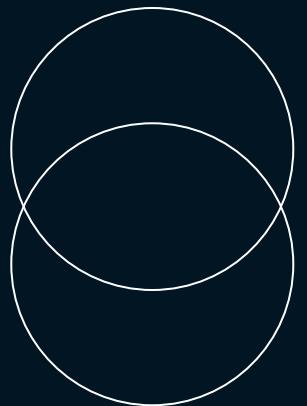
3. Data preprocessing - city : “Texas”

	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

