Airbnb Analysis
Location-Based Recommendation

Outline

- Project plan
- Objective
- Data Description
- Exploratory Data Analysis
- Analysis and Report
- Summary & Recommendations

Project plan

- Objective
- Data source
- Data Mining Techniques

Project plan

Given that most predictors or recommend systems seldom take the security into consideration to help tourists to choose a house from Airbnb, we decided to mine the combined data of Airbnb and crime data around LA.

Objectives

- Which affects the price most?
- Find out the differences between each city in neighbourhoods around LA

Project plan

Data source

- Detailed listings data for Los Angeles
  - House listing dataset with 9 columns of information on 18,624 incidents
  - http://insideairbnb.com/get-the-data.html
- Detailed review data for listings in Los Angeles
  - Crime dataset with 3 columns of information on 179,448 incidents
  - http://insideairbnb.com/get-the-data.html
- The crime data reflects incidents of crime in the City of Los Angeles
  - Crime dataset with 4 columns of information on 162,314 incidents
  - https://data.lacity.org/A-Safe-City/Crime-Data-from-2020-to-Present/2nrs-mtv8

Project plan

Planned Data Mining Techniques

- Use clustering techniques such as Kmeans to identify similar conditions of Airbnb housing
- Use classification methods such as Decision Tree to identify factors most influencing the price and ratings
- Use the modeling techniques to develop a recommender system that can provide a good choice for users

Objective

- Factor that affects price
- Neighbourhoods similarity

Objective

Factor that affects price

Use data mining techniques to analyze and visualize the factor that affects the price of the housing most.

Neighbourhoods similarity

After EDA, we will do clustering to find out the difference between each cluster of Neighbourhood.

Data Description

- Raw dataset
- Listings dataset
- Reviews dataset
- Crimes dataset
- Final jointed dataset

Data Description

Raw data – listings dataset

Housing ID & name

Neighbourhood

LAR
&
LON

Room type

Price

Availability_365

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	availability_365
2	109	Amazing bri	521	Paolo	Other Cities	Culver City	33.98209	-118.38494	Entire home/apt	115	30	2	5/15/16	0.02	1	207
3	2708	Beautiful Fu	3008	Chas.	City of Los Angeles	Hollywood	34.09768	-118.34602	Private room	75	30	27	10/6/20	0.35	2	335
4	2732	Zen Life at t	3041	Yoga Priestess	Other Cities	Santa Monica	34.00475	-118.48127	Private room	155	1	21	12/27/19	0.18	2	365
5	2864	* Beautiful I	3207	Bernadine	Other Cities	Bellflower	33.87619	-118.11397	Entire home/apt	50	30	0			1	0
6	5729	Zen Room w	9171	Sanni	City of Los Angeles	Del Rey	33.9875	-118.432	Private room	70	30	230	4/11/20	1.69	4	358
7	5843	Artist Oasis i	9171	Sanni	City of Los Angeles	Del Rey	33.9875	-118.432	Entire home/apt	135	30	128	8/22/20	1.12	4	97
8	6931	Beau Furn R	3008	Chas.	City of Los Angeles	Hollywood	34.09521	-118.34801	Private room	73	30	22	9/23/20	0.16	2	344
9	7874	Sunny and P	21700	Henry	Other Cities	Bellflower	33.8761	-118.11509	Private room	55	1	12	10/27/19	0.55	3	147
10	7992	Quiet,Walka	22363	Tom	City of Los Angeles	Atwater Village	34.11543	-118.2609	Entire home/apt	89	30	241	10/16/20	2.2	2	39
11	8770	Cozy Guest I	26996	Lillian	City of Los Angeles	Venice	33.99399	-118.45637	Entire home/apt	122	3	401	2/24/20	3.02	1	200
12	9140	City Place Lc	28350	Wendell	Other Cities	Long Beach	33.77206	-118.18893	Private room	80	2	393	11/25/19	4.23	1	0
13	9376	Bright Apt, v	30319	Cristina	City of Los Angeles	Venice	33.99638	-118.47734	Private room	85	30	47	2/21/20	0.35	2	333
14	9545	Burnham Be	31306	Wendy	Other Cities	Redondo Beach	33.83823	-118.38569	Private room	50	1	137	2/21/20	1.01	4	288
15	10760	CASAMIGOS	38596	Debra	City of Los Angeles	Mid-Wilshire	34.05437	-118.35641	Private room	74	30	45	6/16/19	0.34	1	87
16	11374	Budget hote	42220	Searockinn	Other Cities	Gardena	33.90348	-118.29269	Private room	85	1	33	9/19/20	0.28	4	361
17	11511	Craftsman D	40884	Suzan And Michael	City of Los Angeles	Hollywood Hills	34.11479	-118.32289	Entire home/apt	195	30	7	8/13/15	0.05	2	295
18	11877	Le petit bun	30484	Pascalou	City of Los Angeles	Venice	33.99753	-118.47226	Entire home/apt	70	30	45	11/12/19	0.41	2	5
19	12320	1930's Spani	47757	Lori	City of Los Angeles	Mid-Wilshire	34.05864	-118.34352	Entire home/apt	155	30	10	1/20/20	0.38	1	0
20	13776	Burnham Be	31306	Wendy	Other Cities	Redondo Beach	33.83847	-118.38522	Entire home/apt	175	2	172	7/25/20	1.31	4	333
21	14098	Glamour Cal	55411	HankBubbie	City of Los Angeles	Hollywood Hills	34.11959	-118.32056	Entire home/apt	284	5	27	2/24/20	0.23	7	0
22	14107	ARCHITECTU	55422	Jane	City of Los Angeles	Venice	33.99025	-118.45374	Entire home/apt	425	31	30	7/25/18	0.24	1	365
23	14124	Burnham Be	31306	Wendy	Other Cities	Redondo Beach	33.83984	-118.3871	Entire home/apt	175	2	161	9/26/20	1.27	4	303
24	14337	Beautiful Ro	56327	Celia	Other Cities	Torrance	33.84042	-118.32062	Private room	95	2	1	3/28/10	0.01	2	365
25	15089	*****Mode	59169	Josh	City of Los Angeles	Mid-City	34.04244	-118.35222	Entire home/apt	123	3	70	8/1/20	0.52	1	343
26	15333	The Enchant	60057	Georgia	City of Los Angeles	Valley Village	34.16701	-118.41118	Private room	438	2	4	6/1/20	0.08	1	328
27	18041	Bohemian h	69546	Lisa	City of Los Angeles	Venice	33.99179	-118.45056	Entire home/apt	282	2	3	8/27/20	0.25	3	0
28	18067	Bohemian h	69546	Lisa	City of Los Angeles	Venice	33.99065	-118.44962	Entire home/apt	214	2	51	9/12/20	0.48	3	0
29	19887	Eco-friendly	75052	Janet	City of Los Angeles	Silver Lake	34.08362	-118.28309	Private room	75	30	253	8/1/20	1.95	2	365
30	20585	Private Stud	77857	Barbara	City of Los Angeles	Venice	33.98012	-118.4649	Entire home/apt	124	3	484	10/4/20	3.75	1	41
31	20786	Mondrian-Ir	55411	HankBubbie	City of Los Angeles	Hollywood Hills	34.11939	-118.32044	Entire home/apt	199	5	33	10/18/20	0.32	7	53
32	22355	Experience /	55411	HankBubbie	City of Los Angeles	Hollywood Hills	34.11821	-118.32178	Entire home/apt	171	4	44	10/3/20	0.35	7	63
33	23363	Silver Lake fi	91335	David	City of Los Angeles	Silver Lake	34.10312	-118.25778	Entire home/apt	110	3	171	10/6/20	1.54	1	37

Data Description

Raw data - rating dataset

Housing ID Date Rating

	listing_id	date	reviewer_id	reviewer_na	rating	comments
2	2708	3/17/20	117823684	Nathan	3	Location is
3	2708	6/26/20	275399037	Tom	4	Charles was
4	2708	7/31/20	54526949	Jonathan	4	This here is
5	2708	10/6/20	354881527	Haji	4	Top notch spot to stay within walking distance to most everything in West Hollywood.
6	5729	2/17/20	267484570	Scott	2	Number one place to stay. Look nowhere else.
7	5729	2/22/20	102157384	Linda	3	Sanni‚Äôs place is a true oasis ! The experience was very seamless and I could not have been more happy. The floating bed is by far the most unic
8	5729	2/24/20	332215882	Christian	3	We stayed in the ZEN home and it was incredible! The floating bed was so unique and relaxing! The whole area that is shared between homes fe
9	5729	2/27/20	7821387	Jade	4	My mother and I stayed at Sanni's place for two nights while we explored Venice. We had a lovely stay! The space was charming and comfortable
10	5729	3/1/20	62036196	Lucy	3	The space is configured to feel almost like a meditation retreat! I had pretty much all the amenities I needed, which was super convenient. It was
11	5729	3/6/20	141205016	Kayla	4	Sanni‚Äôs place was amazing - a true gem! The Zen room is exactly as advertised. Beautiful, relaxing, safe environment to stay during your visit t
12	5729	3/9/20	262472084	Matthew	3	Beautiful place, a pretty solid location, and an amazing sleep thanks to the bed!
13	5729	4/2/20	279652779	Clarise	2	A must stay!
14	5729	4/11/20	120134826	Emily	4	Sanni‚Äôs place has the perfect touch of outdoor ‚Äúglamping‚Äù spirit and indoor boho, which makes for a unique and fun stay! The hanging bed
15	5843	1/7/20	172057893	Hilda	4	Super neat place. Felt very safe & satisfied while staying here. Hosts were very nice & easy to communicate with.
16	5843	2/17/20	86918818	Richard	3	What a restful place, we (family of 4) stayed in the main house and it had all of the amenities we could need for our time in LA. Close to grocery
17	5843	2/20/20	9552135	Hugo	4	Funky place, quite close to Venice. We didn‚Äôt meet Sanni nor Helene but the communication was smooth and everything was as described.
18	5843	2/24/20	111090022	Monica	3	This house has a lovely garden. Other than dining room chairs there is no couch or comfy chairs to sit on and there is no TV. The house was clean
19	5843	3/3/20	66681786	Michael	4	Great place
20	5843	3/16/20	24021597	Alex	3	Wonderful experience at the Circle, lots of sunlight in the house and surrounded by beautiful plants. Sanni and Helene were very hospitable. Our o
21	5843	3/31/20	9406013	Matt	3	I love this place. The property is so cozy and has such unique character -- and it's a great value.
22	5843	8/22/20	50491424	Avishai	3	Loved The Circle and Sanni's Place, and Helene's hospitality. Chill atmosphere in the Circle Complex, well lit, spacious house, fully equipped kitche
23	6931	3/30/20	273035613	Eric	3	Had a great experience at this Airbnb super clean chas was super friendly. Always answering all my questions and leading me the right way. My t
24	6931	7/1/20	301611732	Ritesh	3	Sparkling
25	6931	7/31/20	91260372	Eoin	4	The best
26	6931	9/23/20	5608763	Benjamin	3	The room is fantastic! The balcony is such a great part of the experience - I spent so much time out there. Everything is tidy and very convenient, a
27	7992	1/4/20	146232401	Jennifer	4	This space is perfect for its central location and walkable access to the Atwater neighborhood. The interior space is clean, but not fussy, and well
28	7992	1/6/20	13817261	Davonte	3	Great location, specious and private quarters. I would highly recommend this space for last minute stays.
29	7992	1/10/20	158771353	Kelly	3	Tom and Heidi were wonderful hosts. Their space was conveniently located and walking distance to many shops and restaurants in Atwater Villag
30	7992	1/13/20	115146	Chester	4	Listing is as described. Pleasant. Good location. Quiet street. Def recommend. Tom was very attentive and even parked my car when needed!
31	7992	1/18/20	1142686	Jonny	4	Very nice spot. Great check in and commutation throughout. A pleasant stay in Atwater.
32	7992	6/17/20	120592783	Maria	5	Perfect

Data Description

Raw data - crime dataset

Data

Crime description

LAT and LON

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	DR_NO	Date Rptd	DATE OCC	TIME OCC	AREA	AREA NAME	Rpt Dist No	Part 1-2	Crm Cd	Crm Cd Desc	Mocodes	Vict Age	Vict Sex	Vict Descent
2	10304468	1/8/20 0:00	1/8/20 0:00	2230	3	Southwest	377	2	624	BATTERY - SI	0444 0913	36	F	B
3	190101086	1/2/20 0:00	1/1/20 0:00	330	1	Central	163	2	624	BATTERY - SI	0416 1822 1	25	M	H
4	201418201	10/3/20 0:00	9/29/20 0:00	1830	14	Pacific	1454	1	420	THEFT FROM	1300 0344 1	63	M	H
5	191501505	1/1/20 0:00	1/1/20 0:00	1730	15	N Hollywood	1543	2	745	VANDALISM	0329 1402	76	F	W
6	191921269	1/1/20 0:00	1/1/20 0:00	415	19	Mission	1998	2	740	VANDALISM	329	31	X	X
7	200100501	1/2/20 0:00	1/1/20 0:00	30	1	Central	163	1	121	RAPE, FORCI	0413 1822 1	25	F	H
8	200100502	1/2/20 0:00	1/2/20 0:00	1315	1	Central	161	1	442	SHOPLIFTING	1402 2004 0	23	M	H
9	200100504	1/4/20 0:00	1/4/20 0:00	40	1	Central	155	2	946	OTHER MISC	1402 0392	0	X	X
10	200100507	1/4/20 0:00	1/4/20 0:00	200	1	Central	101	1	341	THEFT-GRAN	1822 0344 1	23	M	B
11	201312148	6/12/20 0:00	6/11/20 0:00	2000	13	Newton	1383	2	740	VANDALISM	0329 1609	31	F	H
12	200100509	1/4/20 0:00	1/4/20 0:00	2200	1	Central	192	1	330	BURGLARY F	1822 1414 0	29	M	A
13	200100510	1/5/20 0:00	1/5/20 0:00	955	1	Central	111	2	930	CRIMINAL TH	0421 0906	35	M	O
14	200100514	1/5/20 0:00	1/5/20 0:00	1355	1	Central	162	1	341	THEFT-GRAN	1822 0344 2	41	M	A
15	200100515	1/7/20 0:00	1/7/20 0:00	1638	1	Central	162	1	648	ARSON	1402 1501 2	0	X	X
16	200100520	1/8/20 0:00	1/8/20 0:00	1805	1	Central	128	1	442	SHOPLIFTING	0325 1402 0	24	F	H
17	200911119	6/16/20 0:00	6/15/20 0:00	2300	9	Van Nuys	906	1	330	BURGLARY F	344	37	M	O
18	201405970	2/1/20 0:00	2/1/20 0:00	1658	14	Pacific	1494	1	440	THEFT PLAIN	1822 0344	39	M	O
19	200410047	6/22/20 0:00	2/1/20 0:00	1	4	Hollenbeck	429	1	510	VEHICLE - STOLEN		0		
20	201212066	5/3/20 0:00	5/3/20 0:00	1235	12	77th Street	1252	2	437	RESISTING A	1309	0	X	X
21	200100535	1/14/20 0:00	1/14/20 0:00	1330	1	Central	152	1	210	ROBBERY	0416 0411 0	66	M	B
22	201106871	3/4/20 0:00	3/2/20 0:00	2130	11	Northeast	1107	2	745	VANDALISM	MISDEAMEA	50	M	H
23	200100538	1/14/20 0:00	1/14/20 0:00	1730	1	Central	162	1	341	THEFT-GRAN	0344 1822 2	31	M	H
24	200100543	1/15/20 0:00	1/15/20 0:00	1445	1	Central	162	1	442	SHOPLIFTING	0325 1402 0	27	M	B
25	200312493	6/10/20 0:00	5/25/20 0:00	1500	3	Southwest	374	2	668	EMBEZZLEMENT, GRAND T		0		
26	200100546	1/15/20 0:00	1/15/20 0:00	700	1	Central	166	1	230	ASSAULT WI	0416 0913 2	62	M	A
27	200506936	3/9/20 0:00	3/8/20 0:00	1830	5	Harbor	506	1	510	VEHICLE - STOLEN		0		
28	201312939	6/26/20 0:00	6/26/20 0:00	200	13	Newton	1309	1	510	VEHICLE - STOLEN		0		
29	201913337	8/16/20 0:00	8/15/20 0:00	1530	19	Mission	1917	1	510	VEHICLE - STOLEN		0		
30	200100552	1/19/20 0:00	1/19/20 0:00	2000	1	Central	111	1	230	ASSAULT WI	2004 0305 0	71	M	W
31	201212259	5/5/20 0:00	5/5/20 0:00	932	12	77th Street	1239	1	343	SHOPLIFTING	0315 0325 0	0	X	X
32	200100556	1/20/20 0:00	1/20/20 0:00	400	1	Central	141	1	121	RAPE, FORCI	1414 1402 1	19	F	B
33	200100559	1/23/20 0:00	1/23/20 0:00	600	1	Central	111	1	310	BURGLARY	1402 1822 1	51	M	W
34	201810632	5/4/20 0:00	5/4/20 0:00	1530	18	Southeast	1804	2	624	BATTERY - SI	0444 1300 1	18	F	W
35	200911049	6/15/20 0:00	6/9/20 0:00	1025	9	Van Nuys	963	1	440	THEFT PLAIN	0394 0344	35	F	W
36	201513294	7/29/20 0:00	7/29/20 0:00	2010	15	N Hollywood	1591	2	930	CRIMINAL TH	0443 2004	57	M	H
37	200100568	1/27/20 0:00	1/27/20 0:00	1500	1	Central	166	2	930	CRIMINAL TH	1402 0910 0	69	M	B
38	200615122	9/9/20 0:00	9/7/20 0:00	100	6	Hollywood	647	1	815	SEXUAL PENE	1804 1402	21	F	W

	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
1	Premis Cd	Premis Desc	Weapon Used Cd	Weapon Desc	Status	Status Desc	Crm Cd 1	Crm Cd 2	Crm Cd 3	Crm Cd 4	LOCATION	Cross Street	LAT	LON
2	501	SINGLE FAMI	400	STRONG-ARM	AO	Adult Other	624				1100 W 39TH		34.0141	-118.2978
3	102	SIDEWALK	500	UNKNOWN WI	IC	Invest Cont	624				700 S HILL		34.0459	-118.2545
4	103	ALLEY			IC	Invest Cont	420				4700 LA VILLA MARINA		33.9813	-118.435
5	502	MULTI-UNIT DWELLING (APARTMENT, DUPLEX,			IC	Invest Cont	745	998			5400 CORTEEN		34.1685	-118.4019
6	409	BEAUTY SUPPLY STORE			IC	Invest Cont	740				14400 TITUS		34.2198	-118.4468
7	735	NIGHT CLUB	500	UNKNOWN WI	IC	Invest Cont	121	998			700 S BROADWAY		34.0452	-118.2534
8	404	DEPARTMENT STORE			IC	Invest Cont	442	998			700 S FIGUEROA		34.0483	-118.2631
9	726	POLICE FACILITY			IC	Invest Cont	946	998			200 E 6TH		34.0448	-118.2474
10	502	MULTI-UNIT DWELLING (APARTMENT, DUPLEX,			IC	Invest Cont	341	998			700 BERNARD		34.0677	-118.2398
11	501	SINGLE FAMILY DWELLING			IC	Invest Cont	740				200 W 61ST		33.9842	-118.2765
12	101	STREET	306	ROCK/THROWI	IC	Invest Cont	330				15TH	OLIVE	34.0359	-118.2648
13	108	PARKING LOT	511	VERBAL THREA	IC	Invest Cont	930				800 N ALAMEDA		34.0615	-118.2412
14	503	HOTEL			AA	Adult Arrest	341				800 S OLIVE		34.0452	-118.2569
15	404	DEPARTMEN	500	UNKNOWN WI	IC	Invest Cont	648	998			700 W 7TH		34.048	-118.2577
16	252	COFFEE SHOP (STARBUCKS, COFFEE BEAN, PEET			IC	Invest Cont	442				100 S LOS ANGELES		34.0515	-118.2424
17	101	STREET			IC	Invest Cont	330				14200 LEADWELL		34.2041	-118.4425
18	212	TRANSPORTATION FACILITY (AIRPORT)			AO	Adult Other	440				300 WORLD		33.944	-118.4073
19	101	STREET			IC	Invest Cont	510				5400 SHELLEY		34.0912	-118.1624
20	101	STREET	400	STRONG-ARM	AA	Adult Arrest	437				FLORENCE	VAN NESS	33.9746	-118.3177
21	103	ALLEY	204	FOLDING KNIF	IC	Invest Cont	210				7TH	HILL	34.0463	-118.255
22	101	STREET			IC	Invest Cont	745				6100 DELPHI		34.1241	-118.1878
23	404	DEPARTMENT STORE			IC	Invest Cont	341				700 W 7TH		34.048	-118.2577
24	404	DEPARTMENT STORE			IC	Invest Cont	442	998			700 W 7TH		34.048	-118.2577
25	203	OTHER BUSINESS			IC	Invest Cont	668				3800 3RD		34.0183	-118.3204
26	502	MULTI-UNIT	500	UNKNOWN WI	AO	Adult Other	230				600 SAN JULIAN		34.0428	-118.2461
27	101	STREET			IC	Invest Cont	510				WESTERN	221ST	33.8268	-118.309
28	108	PARKING LOT			IC	Invest Cont	510				2100 E 25TH		34.0241	-118.232
29	101	STREET			IC	Invest Cont	510				13000 DRONFIELD		34.3107	-118.4409
30	148	PUBLIC RESTI	500	UNKNOWN WI	AA	Adult Arrest	230				ALAMEDA	LOS ANGELE	34.0578	-118.2371
31	402	MARKET			IC	Invest Cont	343				5900 S FIGUEROA		33.9874	-118.2827
32	503	HOTEL	400	STRONG-ARM	IC	Invest Cont	121	998			300 S FIGUEROA		34.0542	-118.2566
33	503	HOTEL			AO	Adult Other	310				700 N MAIN		34.0583	-118.2378
34	102	SIDEWALK	400	STRONG-ARM	IC	Invest Cont	624				MANCHESTE	AVALON	33.9602	-118.2651
35	119	PORCH, RESIDENTIAL			IC	Invest Cont	440				KESTER	MAGNOLIA	34.1649	-118.4574
36	727	SHOPPING M	511	VERBAL THREA	IC	Invest Cont	930				12200 W VENTURA		34.1432	-118.3986
37	801	MTA BUS	500	UNKNOWN WI	IC	Invest Cont	930	998			6TH	SAN JULIAN	34.0428	-118.2461
38	203	OTHER BUSINESS			IC	Invest Cont	815	998			6000 W SUNSET		34.0998	-118.3211

Data Description

Raw data - crime dataset

Data Crime description LAT and LON

	A	B	C	D	E	F	G	H	I
1	DR_NO	Date Rptd	DATE OCC	TIME OCC	AREA	AREA NAME	Rpt Dist No	Part 1-2	Crm Cd
2	10304468	1/8/20 0:00	1/8/20 0:00	2230	3	Southwest	377	2	624
3	190101086	1/2/20 0:00	1/1/20 0:00	330	1	Central	163	2	624
4	201418201	10/3/20 0:00	9/29/20 0:00	1830	14	Pacific	1454	1	420
5	191501505	1/1/20 0:00	1/1/20 0:00	1730	15	N Hollywood	1543	2	745
6	191921269	1/1/20 0:00	1/1/20 0:00	415	19	Mission	1998	2	740
7	200100501	1/2/20 0:00	1/1/20 0:00	30	1	Central	163	1	121
8	200100502	1/2/20 0:00	1/2/20 0:00	1315	1	Central	161	1	442
9	200100504	1/4/20 0:00	1/4/20 0:00	40	1	Central	155	2	946
10	200100507	1/4/20 0:00	1/4/20 0:00	200	1	Central	101	1	341
11	201312148	6/12/20 0:00	6/11/20 0:00	2000	13	Newton	1383	2	740
12	200100509	1/4/20 0:00	1/4/20 0:00	2200	1	Central	192	1	330
13	200100510	1/5/20 0:00	1/5/20 0:00	955	1	Central	111	2	930
14	200100514	1/5/20 0:00	1/5/20 0:00	1355	1	Central	162	1	341
15	200100515	1/7/20 0:00	1/7/20 0:00	1638	1	Central	162	1	648
16	200100520	1/8/20 0:00	1/8/20 0:00	1805	1	Central	128	1	442

	J	K	L	M	N	O	P	Q	R
1	Crm Cd Desc	Mocodes	Vict Age	Vict Sex	Vict Descent	Premis Cd	Premis Desc	Weapon Used Cd	Weapon Desc
2	BATTERY - SI	0444 0913	36	F	B	501	SINGLE FAMI	400	STRONG-ARM
3	BATTERY - SI	0416 1822 1	25	M	H	102	SIDEWALK	500	UNKNOWN WI
4	THEFT FROM	1300 0344 1	63	M	H	103	ALLEY		
5	VANDALISM	0329 1402	76	F	W	502	MULTI-UNIT DWELLING (APARTMENT, DUPLEX,		
6	VANDALISM	329	31	X	X	409	BEAUTY SUPPLY STORE		
7	RAPE, FORCI	0413 1822 1	25	F	H	735	NIGHT CLUB	500	UNKNOWN WI
8	SHOPLIFTING	1402 2004 0	23	M	H	404	DEPARTMENT STORE		
9	OTHER MISC	1402 0392	0	X	X	726	POLICE FACILITY		
10	THEFT-GRAN	1822 0344 1	23	M	B	502	MULTI-UNIT DWELLING (APARTMENT, DUPLEX,		
11	VANDALISM	0329 1609	31	F	H	501	SINGLE FAMILY DWELLING		
12	BURGLARY F	1822 1414 0	29	M	A	101	STREET	306	ROCK/THROWI
13	CRIMINAL TH	0421 0906	35	M	O	108	PARKING LOT	511	VERBAL THREA
14	THEFT-GRAN	1822 0344 2	41	M	A	503	HOTEL		
15	ARSON	1402 1501 2	0	X	X	404	DEPARTMEN	500	UNKNOWN WI
16	SHOPLIFTING	0325 1402 0	24	F	H	252	COFFEE SHOP (STARBUCKS, COFFEE BEAN, PEET		

	S	T	U	V	W	X	Y	Z	AA	AB
1	Status	Status Desc	Crm Cd 1	Crm Cd 2	Crm Cd 3	Crm Cd 4	LOCATION	Cross Street	LAT	LON
2	AO	Adult Other	624				1100 W 39TH		34.0141	-118.2978
3	IC	Invest Cont	624				700 S HILL		34.0459	-118.2545
4	IC	Invest Cont	420				4700 LA VILLA MARINA		33.9813	-118.435
5	IC	Invest Cont	745	998			5400 CORTEEN		34.1685	-118.4019
6	IC	Invest Cont	740				14400 TITUS		34.2198	-118.4468
7	IC	Invest Cont	121	998			700 S BROADWAY		34.0452	-118.2534
8	IC	Invest Cont	442	998			700 S FIGUEROA		34.0483	-118.2631
9	IC	Invest Cont	946	998			200 E 6TH		34.0448	-118.2474
10	IC	Invest Cont	341	998			700 BERNARD		34.0677	-118.2398
11	IC	Invest Cont	740				200 W 61ST		33.9842	-118.2765
12	IC	Invest Cont	330				15TH	OLIVE	34.0359	-118.2648
13	IC	Invest Cont	930				800 N ALAMEDA		34.0615	-118.2412
14	AA	Adult Arrest	341				800 S OLIVE		34.0452	-118.2569
15	IC	Invest Cont	648	998			700 W 7TH		34.048	-118.2577
16	IC	Invest Cont	442				100 S LOS ANGELES		34.0515	-118.2424

Since we have to merge this raw crime data with airbnb housing data and we also want to include more information of the crime, we classify the crime into five level(0~4). The bigger number represents the higher severity of the crime incident. When grouping the locations by latitude and longitude, we add two new columns which are crime level and crime incidents count. The former is the average crime level of the location and the latter is the total criminal incidents happened at this location.

Data Description

Listings data

Name	Label	Description
id	Accommodation ID	Unique listing id for accommodation
name	Name of listing Airbnb	Unique name of each listing Airbnb
neighbourhood	Location of surrounding area	Cities around Los Angeles
neighbourhood_group	Group of neighbourhood in LA	Groups of cities in LA
latitude	Latitude of accommodation	The latitudinal positions of housing
longitude	Longitude of accommodation	The longitude positions of housing

Data Description

Listings data (cont.)

Name	Label	Description
room_type	Room type of accommodation	A three-type categorical variable that are: • 0: Entire home/apt • 1: Private room • 2: Hotel room • 3: Shard room
availability_365	Total available days in one year for the housing	An Airbnb host can setup a calendar for their listing so that it is only available for a few days or weeks a year.
price	Price of accommodation	The price of accommodation per night

Data Description

Reviews data

Name	Label	Description
listing_id	Accommodation ID	Unique listing id for accommodation
date	The comment date	The date when the housing property was commented on Airbnb
score	The average score of accommodation	The score of all visitors for each accommodation range from 0 to 5.

Data Description

Crime data

Name	Label	Description
LAT	Latitude of location	The latitudinal positions for incidents of crime
LON	Longitude of location	The longitude positions for incidents of crime
crime_level	Incidents severity level	A four-level categorical variable that represent the severity level of incidents 0: lowest crime severity ~ 4: highest crime severity (The Crime_preprocessing R script has describe how we classify them)
crime_incident_count	Calculate number of times crime incidents happened	Total criminal incidents happened in 2020 at each location by latitude and longitude

Data Description

Final jointed data

- Firstly, we have removed the missing data of rows in each dataset.
- The id in listings data and the listing_id in review data are the same which represent the unique id of each housing. We merge the listings data and reviews data using these key values. Since one housing may several rating, we average those ratings on the with the id of housing.
- Afterwards, we merge it with crime data based on the latitude and longitude. One thing needs to be noticed that we round the latitude and longitude to 2 decimal places in both data so that we can using these variables to merge them. Moreover, one location may have several crime incidents, so we average those crime level with same latitude and longitude. We also add a new attribute which represents the total incidents count in each location.
- After all three dataset be jointed, the output data frame contains 8576 rows and 12 columns.

Exploratory Data Analysis (EDA)

- Dataset Summary
- Data Manipulation
- Univariate Analysis
- Bivariate Analysis
- Discussion and Conclusions

EDA - Dataset Summary

In the final dataset, some attributes make no difference for the analysis of our objective at first glance such as the names of person and the date. We remove them and tried to understand furtherly the remain variables and make sure all of that are appropriate for the analysis.

Initial data survey

Variable	VariableType	MissingValues	n	mean	sd	median	se	min	max	range
availability_365	integer	0	8576	216.32	122.07	235.5	1.32	1	365	364
avg_rating	numeric	0	8576	3.52	0.46	3.54	0.01	0	5	5
crime_incidents_count	integer	0	8576	302.86	347.38	192	3.75	1	1790	1789
crime_level	numeric	0	8576	2.22	0.2	2.23	0	1	4	3
id	integer	0	8576	26076432.06	14122953.66	27776148	152504.67	2708	45789909	45787201
latitude	numeric	0	8576	34.07	0.07	34.07	0	33.71	34.32	0.61
longitude	numeric	0	8576	-118.37	0.09	-118.36	0	-118.66	-118.16	0.5
name	character	0	8576							
neighbourhood	character	0	8576							
neighbourhood_group	character	0	8576							
price	integer	0	8576	190.77	261.88	120	2.83	10	4000	3990
room_type	character	0	8576							

EDA - Dataset Summary

Initial Observations

- No missing values in the final dataset
- The variables id and name appear to be the unique keys, so we won't use it for analysis but we may use it for visualization
- The variables latitude and longitude have little connection to the analysis, but we can do the visualization based on them
- 8 numeric variables
  - Categories: id, latitude, longitude
  - Measures: availability_365, crime_incidents_count, crime_level, avg_rating
- 4 character variables
  - Categories: name, room_type, neighbourhood, neighbourhood_group

EDA - Data Manipulation

Convert character attributes to factors

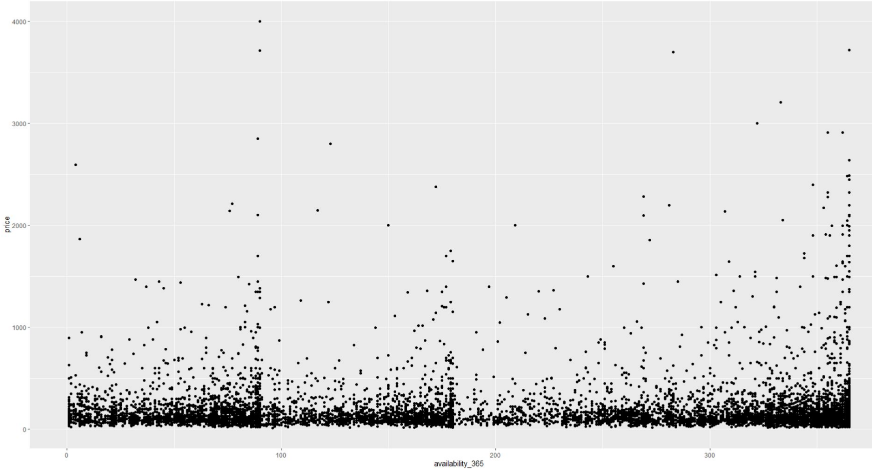
- There are 4 unique character values in room_type attribute
- There are 3 unique character values in neighbourhood_group attribute
- There are 141 unique character values in neighbourhood attribute

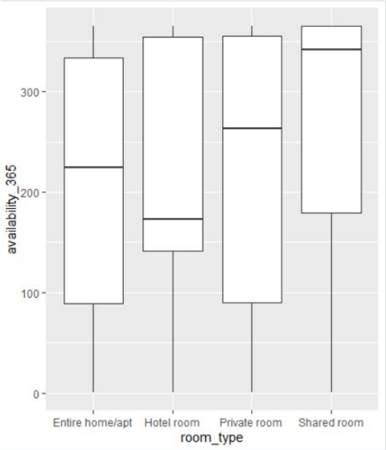
room_type	neighbourhood_group	neighbourhood
Entire home/apt Private room Shared room Hotel room	City of Los Angeles Other Cities Unincorporated Areas	Venice Hollywood Downtown Hollywood Hills : :

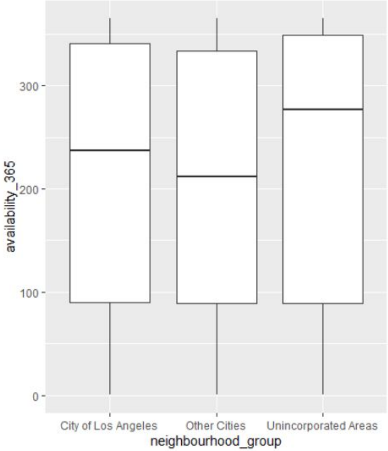
EDA - Discussion

Remove availability_365

Attribute definition: Days of availability for rental in one calendar year.







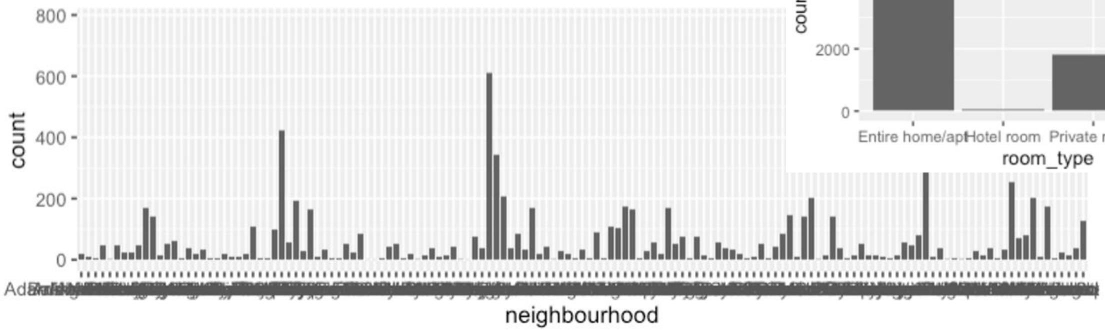
- Although by looking at the boxplot on the right side, availability is varied from room type and neighborhood group, there is no obvious relation observed in the scatter plot on the left.
- This value is randomly determined by the hosts and logically no closely connected with price so we decide to not include this column

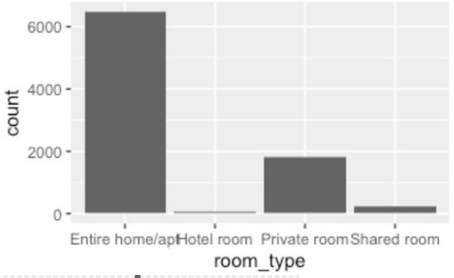
EDA - Univariate Analysis

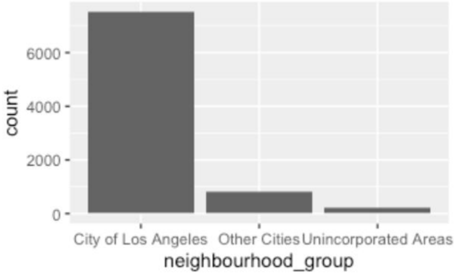
Univariate Summary of Factors

```
> colnames(final_data %>% select_if((is.character)))
[1] "name"                    "neighbourhood_group" "neighbourhood"        "room_type"
```

The variable name appears to be the unique keys, so we won't use it for analysis but we may use it for the label of visualization







EDA - Univariate Analysis

Explore Numeric Attributes

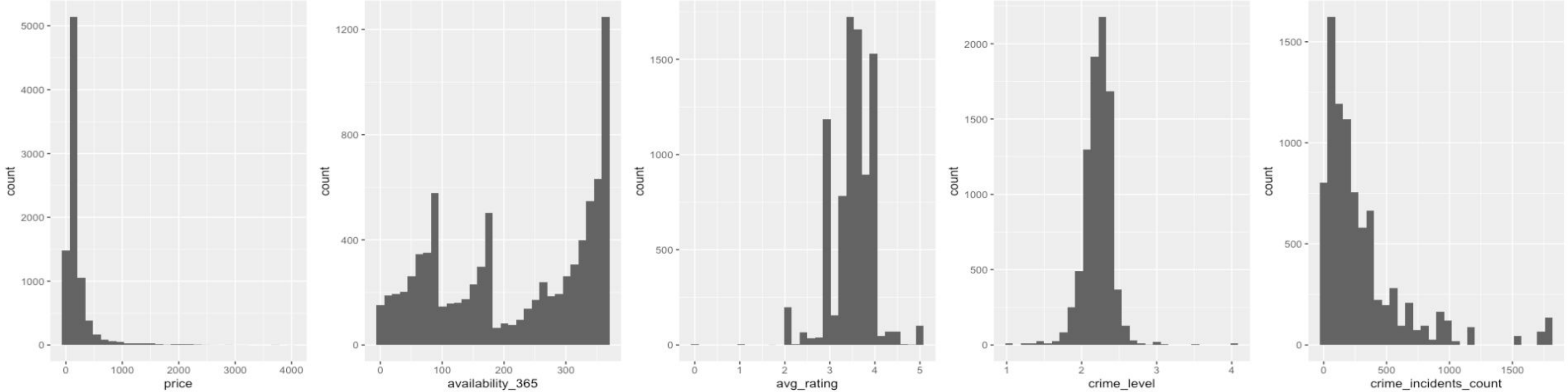
```
> colnames(final_data %>% select_if((is.numeric)))
[1] "id"                "latitude"          "longitude"         "price"
[5] "availability_365"  "avg_rating"        "crime_level"       "crime_incidents_count"
```

- The variable id appears to be the unique keys, it could be useful in the future if we want to join to other information, but currently is not useful for our analysis and modeling purposes. We will still keep it since we may use it for the label of visualization as well.

- The variables latitude and longitude appear to be meaningless for the analysis and modeling, but we will still keep them so that we can use them for visualization.

EDA - Univariate Analysis

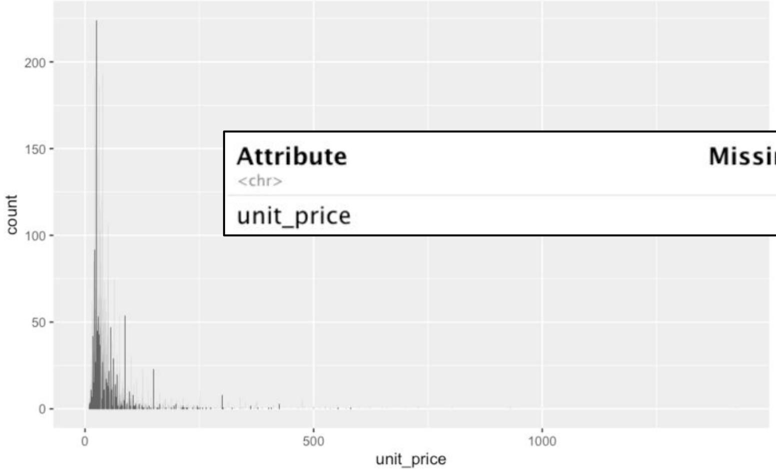
Explore Numeric Attributes



Attribute <chr>	Missing Values <int>	Unique Values <int>	Mean <dbl>	Min <dbl>	Max <dbl>	SD <dbl>
price	0	712	190.769007	10	4000	261.8774621
availability_365	0	365	216.315182	1	365	122.0673375
avg_rating	0	471	3.523852	0	5	0.4638161
crime_level	0	664	2.224057	1	4	0.1982071
crime_incidents_count	0	344	302.859492	1	1790	347.3810503

EDA - Bivariate Analysis (Measures)

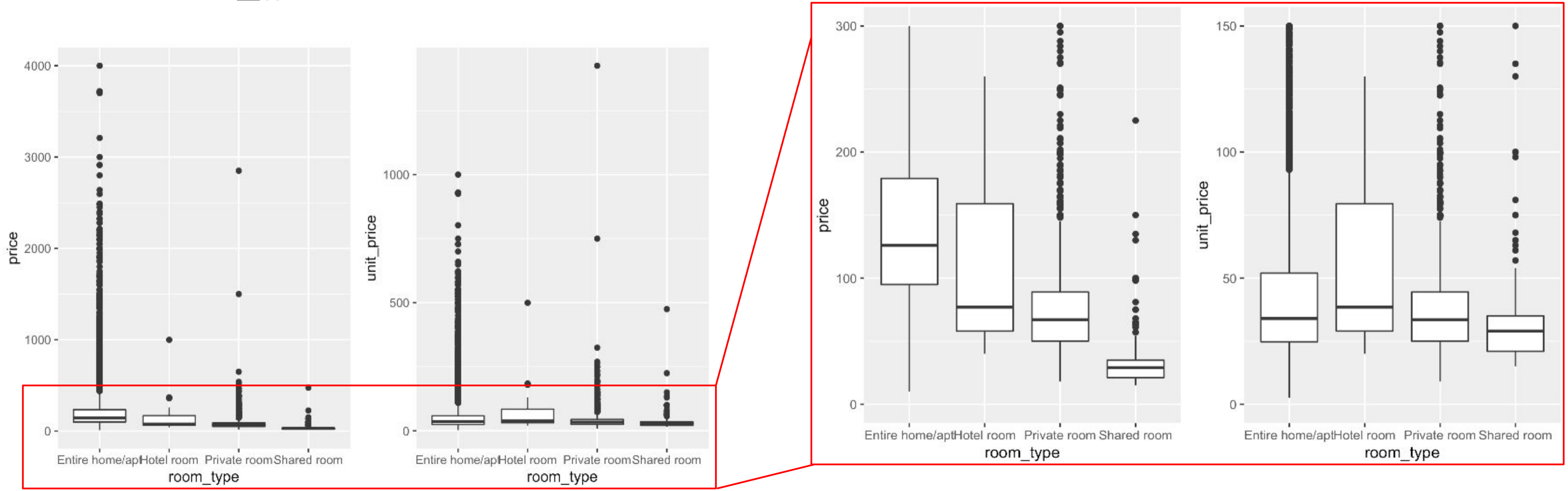
- Since we focused on finding the factor that affect the price most, we plot the relation between price and each other attributes.
- Before we do the bivariate analysis on the price, we do some data manipulation on the price.
- We convert the the price into <u>unit price</u> for each person based on the room type that can accommodate how many people.



Attribute <chr>	Missing Values <int>	Unique Values <int>	Mean <dbl>	Min <dbl>	Max <dbl>	SD <dbl>
unit_price	0	716	52.902781	2.5	1425	66.7841040

EDA - Bivariate Analysis (Measures)

Price vs unit_ price



- Unit_price appears to be more reasonable for each persons' payment
- Unit_price are more suitable for our analysis

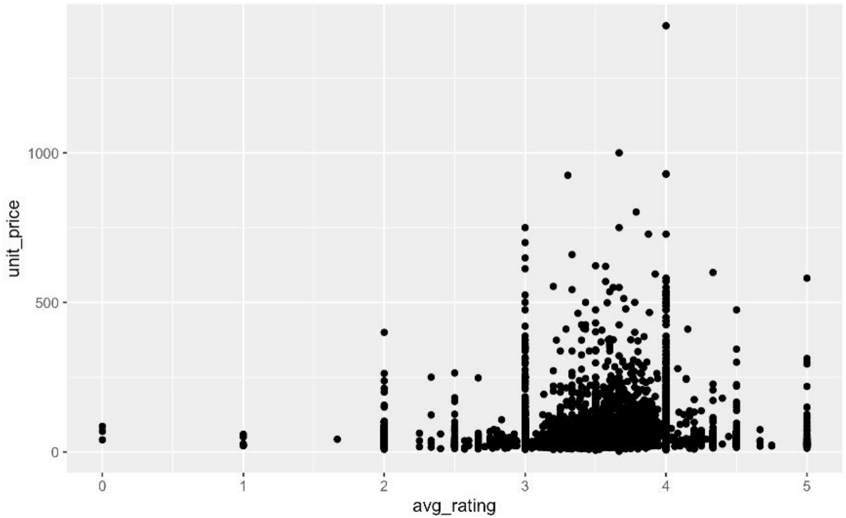
EDA - Bivariate Analysis (Measures)

Correlation matrix

All attributes has low correlation between each others

EDA - Bivariate Analysis (Measures)

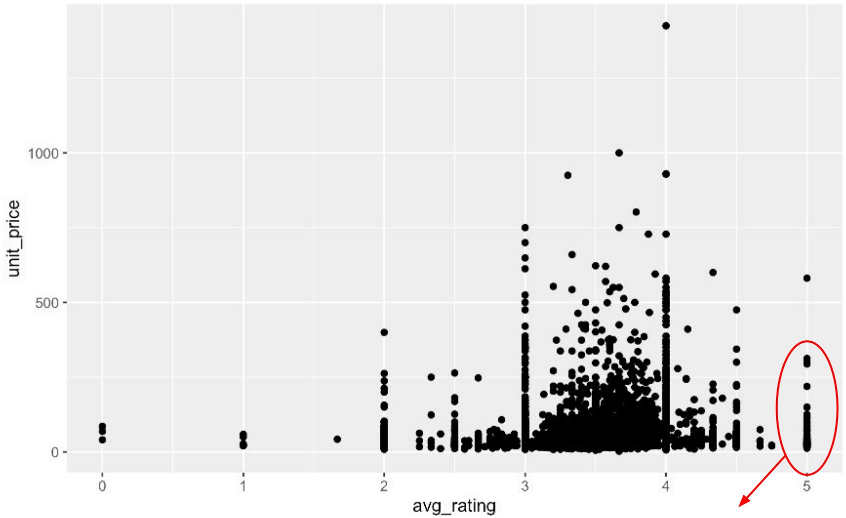
Unit price vs average rating



- There are some housings only be rated a few times, so the distribution seems to be a little discrete even after we averaged the ratings.
- At least, we can tell by the figure that housings with low rating must below $500.

EDA - Bivariate Analysis (Measures)

Explore the data points with high rating but low price

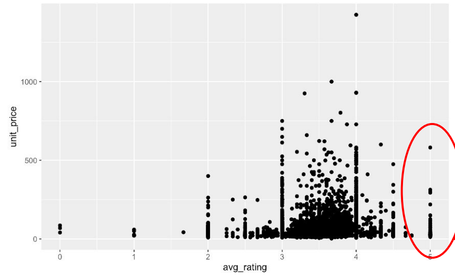


High rating but low price

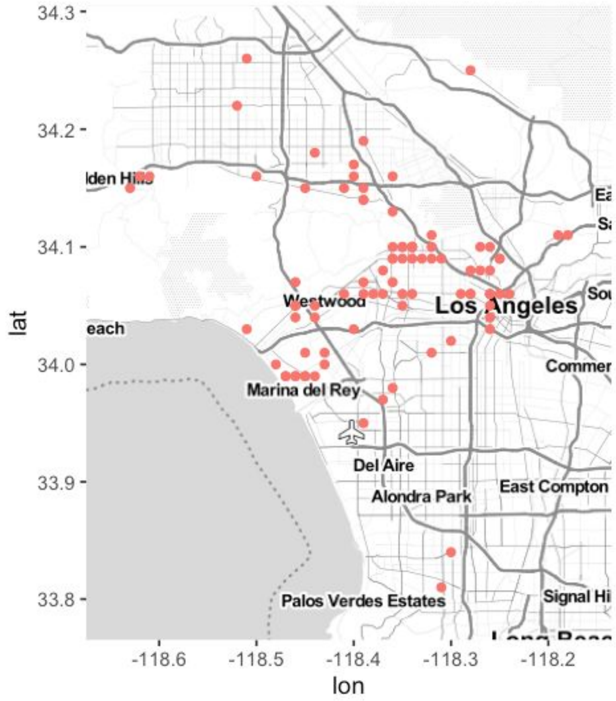
Attribute <chr>	Missing Values <int>	Unique Values <int>	Mean <dbl>	Min <dbl>	Max <dbl>	SD <dbl>
id	0	89	2.807381e+07	131557.00	45642741.00	1.570252e+07
latitude	0	28	3.407326e+01	33.81	34.26	7.194972e-02
longitude	0	32	-1.183709e+02	-118.63	-118.18	9.026024e-02
price	0	65	1.404045e+02	21.00	399.00	9.424963e+01
availability_365	0	62	2.429101e+02	7.00	365.00	1.193053e+02
avg_rating	0	1	5.000000e+00	5.00	5.00	0.000000e+00
crime_level	0	74	2.231017e+00	1.75	3.00	1.684830e-01
crime_incidents_count	0	72	3.258989e+02	10.00	1790.00	3.130327e+02
unit_price	0	64	3.995787e+01	12.25	99.75	2.267512e+01

EDA - Bivariate Analysis (Measures)

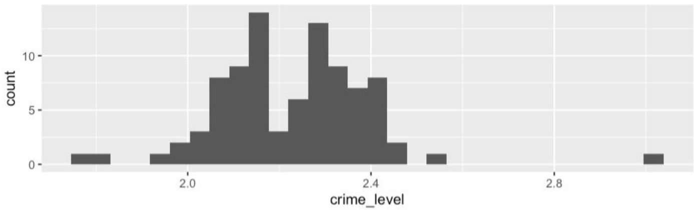
Explore the data points with high rating but low price

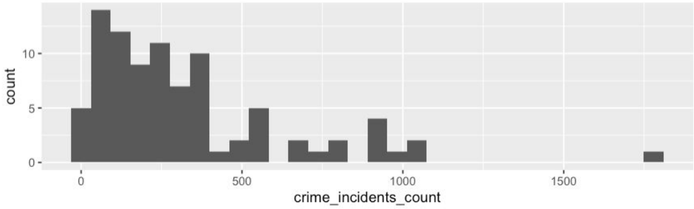


Most housings are around Hollywood and Westwood.



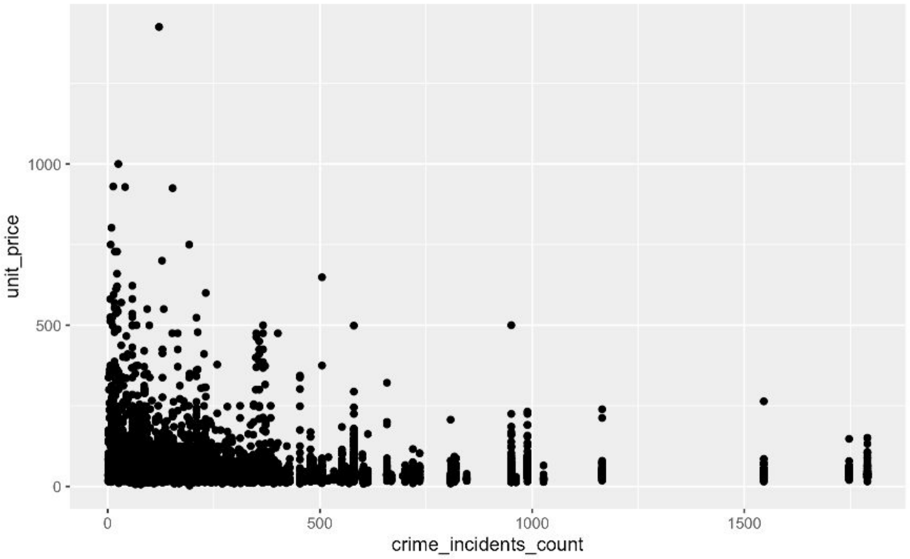
Crime level and criminal incidents seem to be lower





EDA - Bivariate Analysis (Measures)

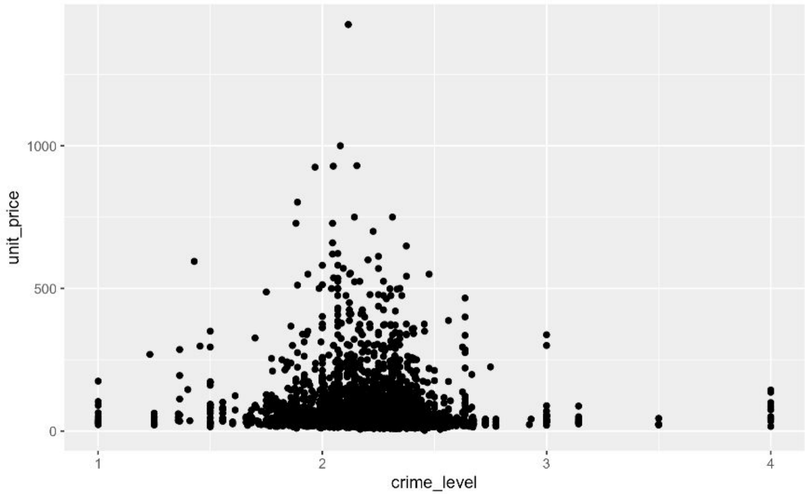
Unit price vs crime incidents count



- Most of the locations only happen a few criminal incidents (less than 500).
- Moreover, housings with high unit price have the tendency of low criminal incidents.

EDA - Bivariate Analysis (Measures)

Unit price vs crime level



- Similarity, based on the figure in previous slide, the distribution on the right figure seems to be a little discrete even after we averaged them by locations
- It seems that no matter how much the price is, the crime level for surrounding locations of each housing concentrated around 2.2 to 2.3

EDA - Bivariate Analysis (Categories)

room_ type vs neighbourhood_ group

room_type	Entire home/apt	Hotel room	Private room	Shared room	Sum
neighbourhood_group					
City of Los Angeles	5704	57	1544	221	7526
Other Cities	606	1	210	6	823
Unincorporated Areas	172	0	46	9	227
Sum	6482	58	1800	236	8576

- Since we used inner join to merge the dataset only, these housings are those people who choose to stay and leave the ratings in 2020.
- It seems reasonable that visitors have a preference for the entire housing due to the COVID-19.
- Actually, we have counted the room type before we merged them and most of the housings are Entire home/apt and private room around LA. Thus, it may also be the reason for the result of such distribution.

EDA - Bivariate Analysis (Categories vs Measures)

- Category/measure bivariate analysis generally involved looking at measure distribution by different category values. Side-by-side boxplots are a good way to visualize this.

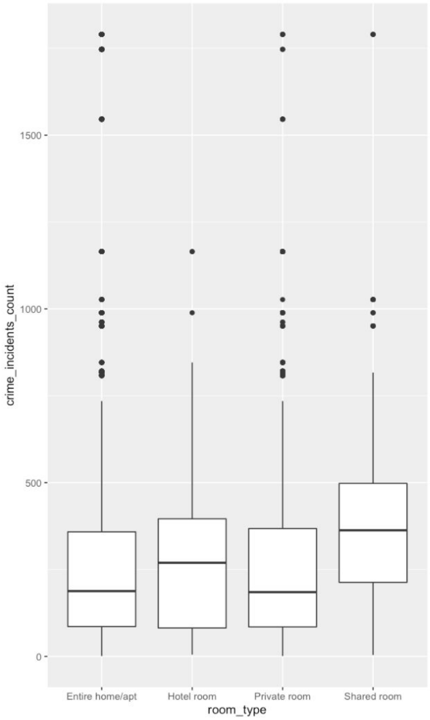
- Since we have generally verified the measures and categories individually, we are primarily looking to improve our understanding and to identify relationships we didn't expect.

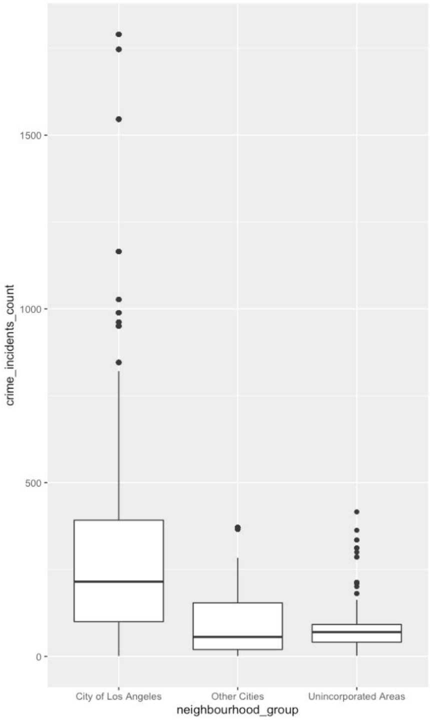
- We are down to 7 measures (not counting the employee number which is just retained as a unique identifier) and 16 categories

- Since it is feasible, we will create one boxplot for each of the 7 measures plotted against each category

EDA - Bivariate Analysis (Categories vs Measures)

Crime incident count by category

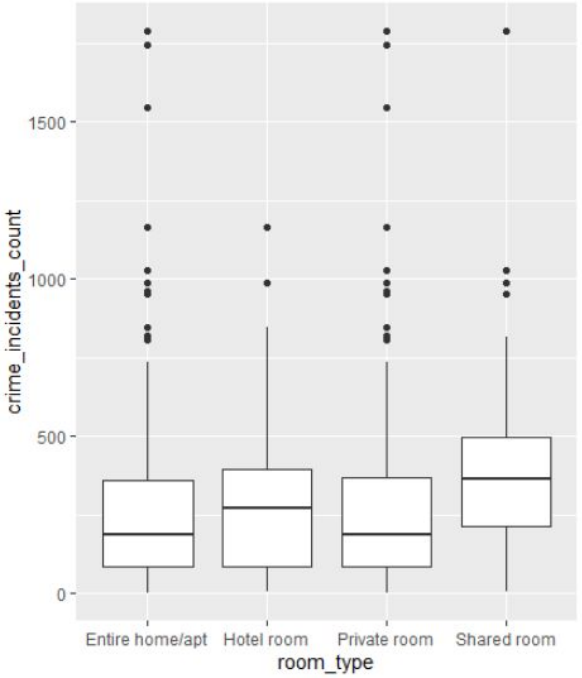




- There are big differences on crime numbers in each group.
- It looks like city of los angeles are more likely have more crime than other areas.
- To make more interpretation on relationship between room type and crime count. We will look at table in the next slide.

EDA - Bivariate Analysis (Categories vs Measures)

Crime incident count by category (cont.)

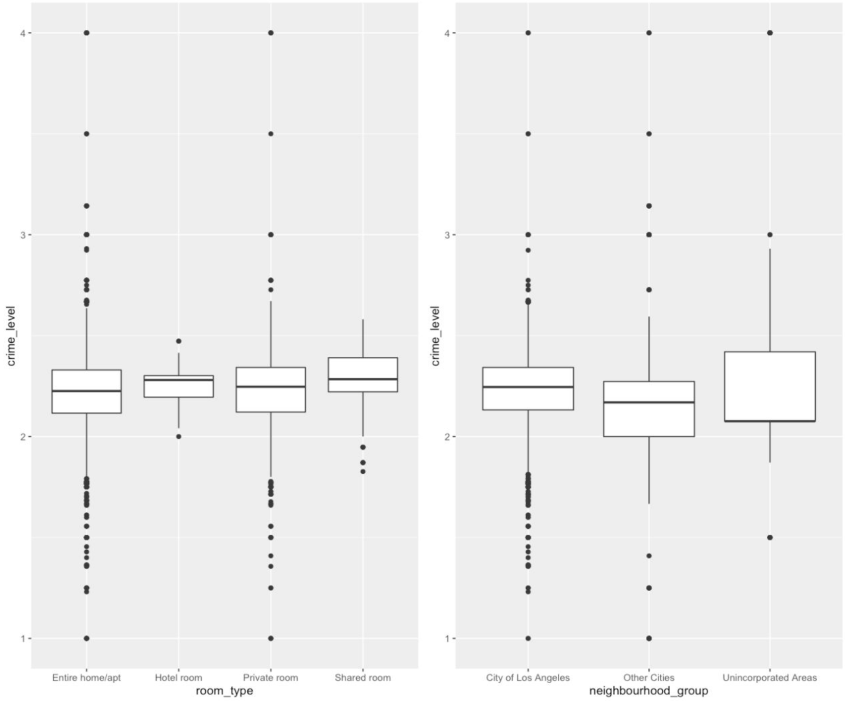


	Entire home/apt	Hotel room	Private room	Shared room
City of Los Angeles	5704	57	1544	221
Other Cities	606	1	210	6
Unincorporated Areas	172	0	46	9

- Combining table above and box plot on left, we can tell that shared room is mostly located in city of los angeles and has the greatest average number of crime incidents.
- This can also be one of the reason shared rooms intend to have lower price

EDA - Bivariate Analysis (Categories vs Measures)

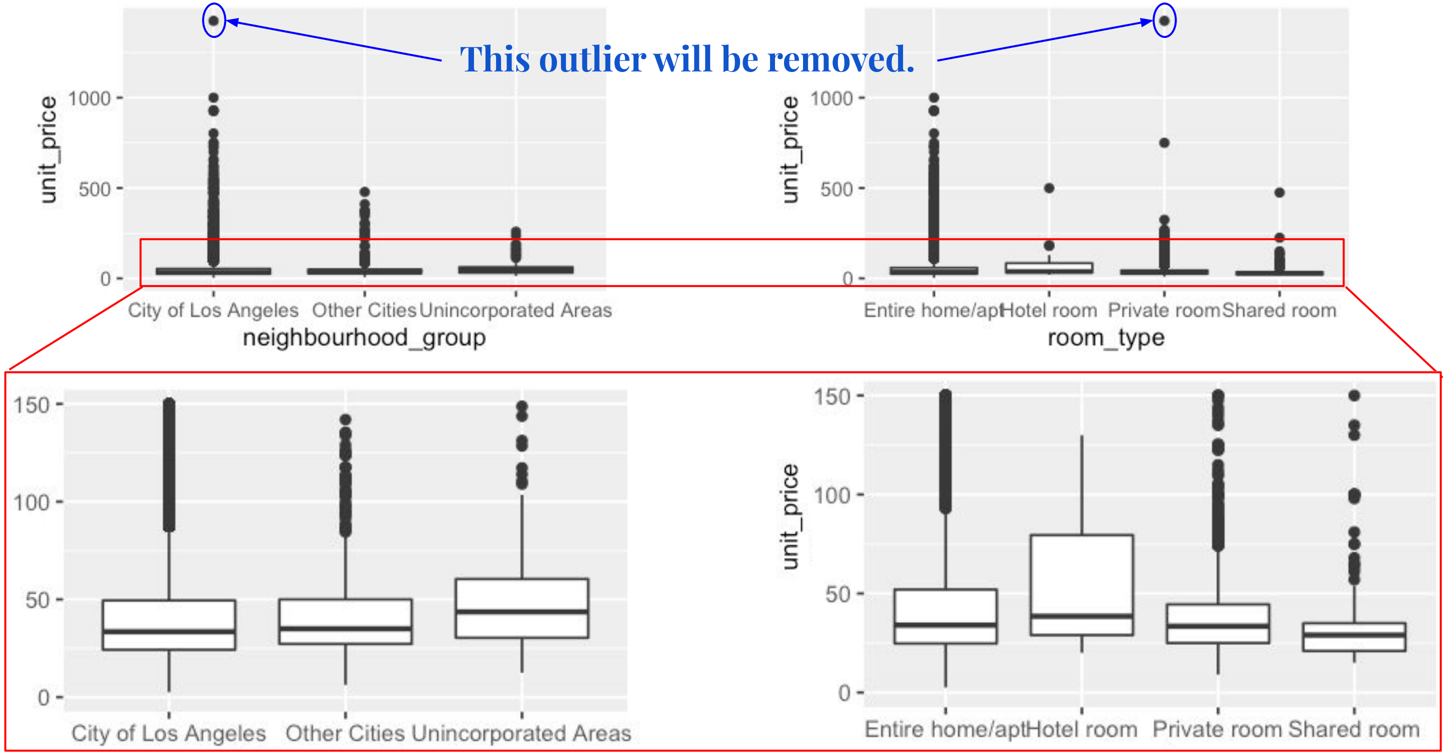
Crime level by category



- Average crime level seems to have no much difference depend on room type
- Unincorporated areas have lower average crime level than other two areas

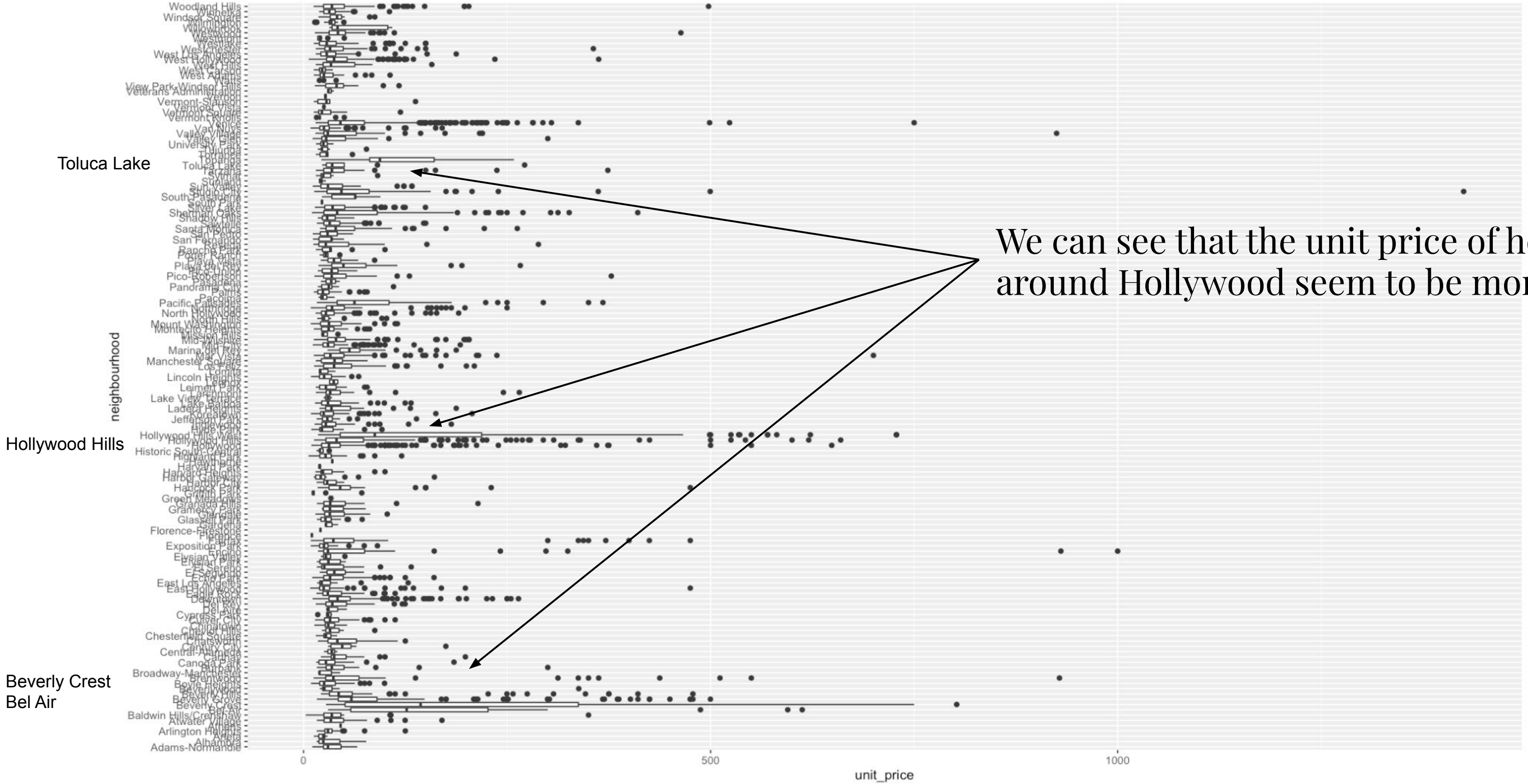
EDA - Bivariate Analysis (Categories vs Measures)

unit_price vs neighbourhood_group ***unit_price vs room_type***



EDA - Bivariate Analysis (Categories vs Measures)

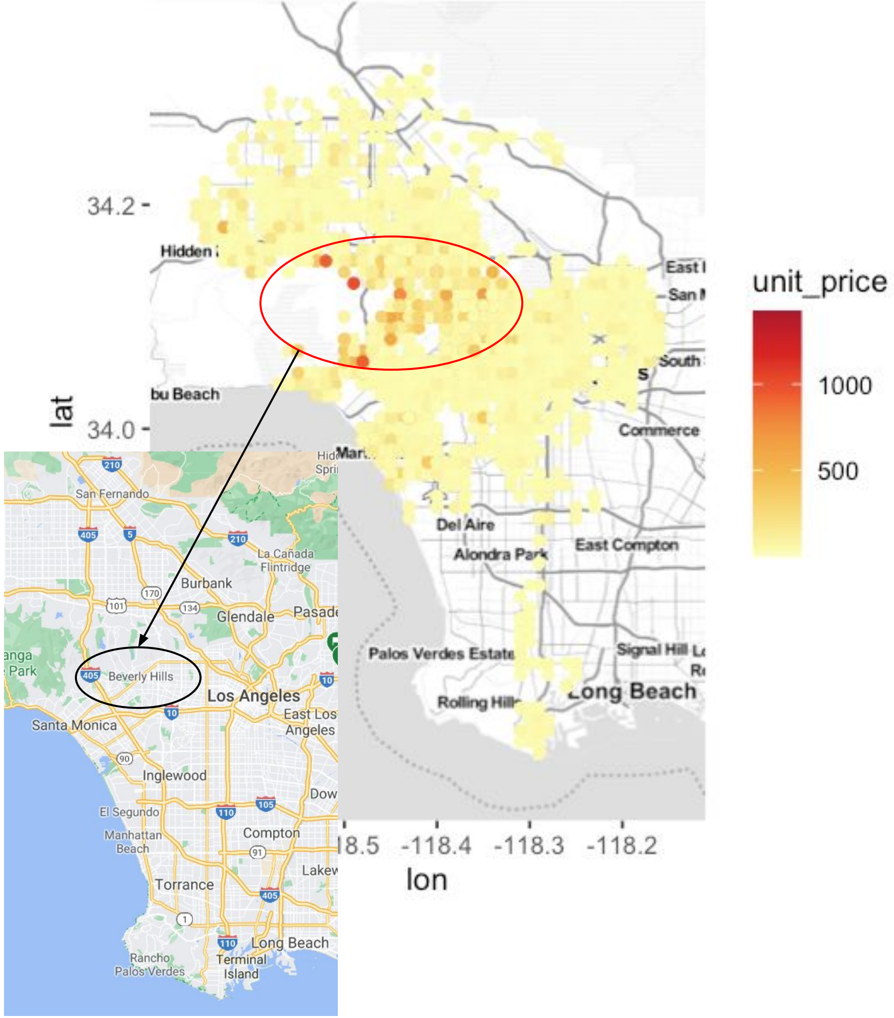
Unit_price vs neighbourhood



We can see that the unit price of housings around Hollywood seem to be more expensive

Discussion and Conclusions

- There are some data points seem to be outliers, but the prices of Airbnb housing are certainly affected by the location of it.
- The housing around Beverly Hills are indeed more expensive.
- Therefore, we will still take them into consideration when doing neighbourhood clustering except the one which is way more expensive.



Analysis & Report

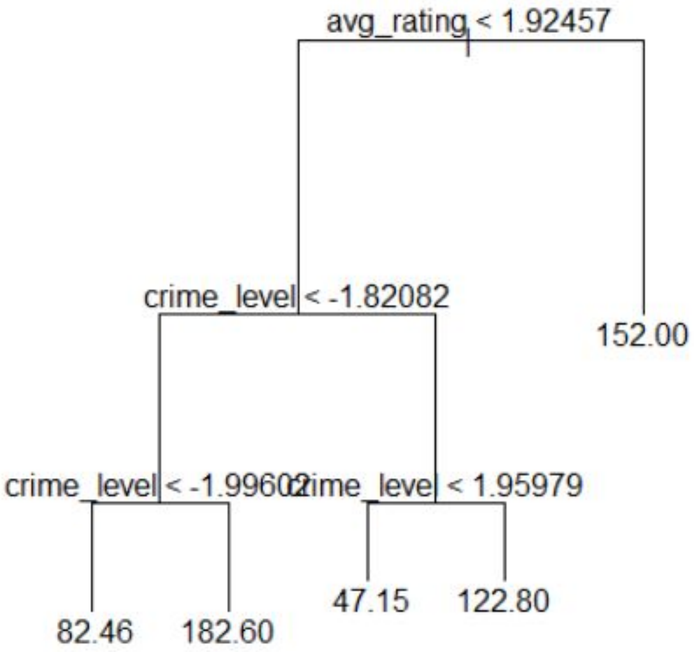
- Decision Tree
- K-means

Object 1

- Which affects the price most?

Decision Tree

Set unit price as target



From Tree graph on the left we can tell that the most important factor affecting unit price is average rating

Target: unit price

Conclusion for Object 1

- According to the bivariate analysis for unit price, we can conclude that price are more expensive around Beverly Hill area.
- Hotel room have more variance compared to the other room type. However, the unit price for different room type has no significantly differences. Maybe it is because these Airbnb housing are all located close to Los Angeles area.
- According to the decision tree, average rating may highly affected the price of housings and crime level is the following factor that affect price.
- In conclusion, based on all above analysis, we still believe that the location may be the most factor that affect the price of Airbnb housings.

Object 2

- Find out the differences between each city in neighbourhoods around LA

Analysis & Report (neighbourhood clustering)

General Guidelines

- Adding a few new columns for counting the total number of each room type of housings, based on our second objective which is finding the difference between the cluster of neighbourhood
- Clustering neighborhoods using k-means to find that the housings at which neighbourhoods have similar conditions and use decision tree to find that the important factors when classifying the neighbourhoods.
- Using decision tree to determine the factor importance of each cluster.

Analysis & Report (neighbourhood clustering)

General Guidelines

- Adding a few new columns for counting the total number of each room type of housings, based on our second objective which is finding the difference between the cluster of neighbourhood
- Clustering neighborhoods using k-means to find that the housings at which neighbourhoods have similar conditions and use decision tree to find that the important factors when classifying the neighbourhoods.
- Using decision tree to determine the factor importance of each cluster.

Analysis & Report (neighbourhood clustering)

Neighbourhood data

neighbourhood <chr>	shared_room_count <dbl>	private_room_count <dbl>	Entire_home_apt_count <dbl>	Hotel_room_count <dbl>	avg_unit_price <dbl>	avg_crime_level <dbl>	avg_crime_incidents_count <dbl>	avg_rating <dbl>	latitude <dbl>
Adams-Normandie	0	11	6	0	33.63235	2.365274	371.647059	3.461588	34.03176
Alhambra	0	1	7	0	35.09375	2.278540	68.250000	3.549750	34.08500
Arleta	0	4	2	0	22.33333	2.397202	130.000000	3.595833	34.24500
Arlington Heights	22	16	11	0	32.44388	2.268912	315.510204	3.277816	34.04735
Athens	0	2	0	0	45.50000	2.397790	181.000000	3.552000	33.92000
Atwater Village	0	4	42	0	43.61413	2.297493	76.108696	3.526087	34.11739
Baldwin Hills/Crenshaw	0	11	13	0	51.11458	2.336621	176.791667	3.420125	34.01625
Bel-Air	0	3	19	0	177.56818	1.813052	16.909091	3.582227	34.09318
Beverly Crest	0	6	39	0	218.16667	2.062383	11.488889	3.517756	34.10778
Beverly Grove	3	24	138	2	100.10928	2.134936	260.227545	3.588204	34.07485
Beverly Hills	0	35	105	0	65.90000	2.103414	90.607143	3.577979	34.06729
Beverlywood	0	2	10	0	54.06250	2.015913	122.833333	3.626750	34.04333
Boyle Heights	7	11	33	0	31.89216	2.438070	272.784314	3.409353	34.04549
Brentwood	0	12	50	0	97.29032	2.050419	133.145161	3.501806	34.05677
Broadway-Manchester	0	5	2	0	27.35714	2.390299	527.571429	3.494000	33.95000
Burbank	0	10	28	0	46.63816	2.367023	30.421053	3.606342	34.16079
Canoga Park	0	11	8	0	39.21053	2.268822	150.473684	3.502263	34.20789
Carthay	0	12	22	0	48.15441	2.232318	181.647059	3.528735	34.05853
Central-Alameda	1	0	2	0	33.75000	2.502337	241.666667	4.000000	34.00333
Century City	0	4	3	0	62.28571	2.076084	184.285714	2.996143	34.05429
Chatsworth	0	9	8	0	52.33824	2.266777	44.941176	3.739588	34.26294
Chesterfield Square	0	2	7	0	29.05556	2.478201	304.666667	3.586889	33.98556
Cheviot Hills	1	2	8	0	38.43182	1.985246	161.363636	3.438364	34.03455
Chinatown	0	5	13	0	32.22222	2.265785	514.777778	3.495444	34.06167
Culver City	0	34	73	0	37.70561	2.177557	115.747664	3.535598	34.01280
Cypress Park	0	1	4	0	28.95000	2.463085	105.000000	3.461800	34.09800
Del Aire	0	1	2	0	37.08333	2.222222	29.333333	3.681000	33.93000
Del Rey	0	23	78	0	42.56436	2.198970	144.584158	3.555485	33.99129
Downtown	6	57	353	5	50.70012	2.275808	1118.888361	3.499299	34.04615
Eagle Rock	0	8	49	1	38.46552	2.221667	77.827586	3.574328	34.13414
East Hollywood	12	56	116	8	32.24870	2.352458	449.093750	3.394380	34.08953
East Los Angeles	0	4	25	0	30.25862	2.432196	49.551724	3.491483	34.04897
Echo Park	0	36	129	0	38.73333	2.245091	207.933333	3.600455	34.08188
El Segundo	0	3	8	0	42.06818	2.015186	11.000000	3.503182	33.92909
El Sereno	1	10	21	0	37.27344	2.466261	84.656250	3.544781	34.08031
Elysian Park	0	1	5	0	32.66667	2.038717	68.500000	3.329167	34.07833
Elysian Valley	0	1	6	0	30.67857	2.263726	124.285714	3.635714	34.09429
Encino	3	5	45	0	93.30660	2.132270	139.358491	3.605226	34.16226
Exposition Park	0	9	15	0	29.73958	2.375848	344.416667	3.572958	34.01750
Fairfax	0	16	70	0	71.83430	2.233058	292.686047	3.539244	34.08221
Florence	0	0	1	0	10.25000	2.344418	421.000000	4.000000	33.99000
Florence-Firestone	0	0	1	0	20.75000	2.507463	201.000000	3.917000	33.99000
Gardena	0	3	0	0	32.50000	2.364583	96.000000	3.410333	33.90000

1–43 of 141 rows | 1–10 of 12 columns

Previous 1 2 3 4 Next

Analysis & Report (neighbourhood clustering)

Dataset Summary

Variable	VariableType	MissingValues	n	mean	sd	median	se	min	max	range
avg_crime_incidents_count	numeric	0	141	187.65	177.53	133.15	14.95	3	1118.89	1115.89
avg_crime_level	numeric	0	141	2.26	0.17	2.27	0.01	1.76	2.93	1.18
avg_rating	numeric	0	141	3.52	0.16	3.53	0.01	2.82	4	1.18
avg_unit_price	numeric	0	141	44.19	26.65	37	2.24	10.25	218.17	207.92
Entire_home_apt_count	numeric	0	141	45.97	87.3	18	7.35	0	701	701
Hotel_room_count	numeric	0	141	0.41	1.81	0	0.15	0	15	15
housing_count	integer	0	141	60.82	102.65	28	8.64	1	783	782
latitude	numeric	0	141	34.06	0.11	34.06	0.01	33.73	34.3	0.58
longitude	numeric	0	141	-118.35	0.1	-118.35	0.01	-118.63	-118.16	0.47
neighbourhood	character	0	141							
private_room_count	numeric	0	141	12.76	16.27	7	1.37	0	100	100
shared_room_count	numeric	0	141	1.67	4.29	0	0.36	0	30	30

Analysis & Report (neighbourhood clustering)

Summary for Numeric Attributes

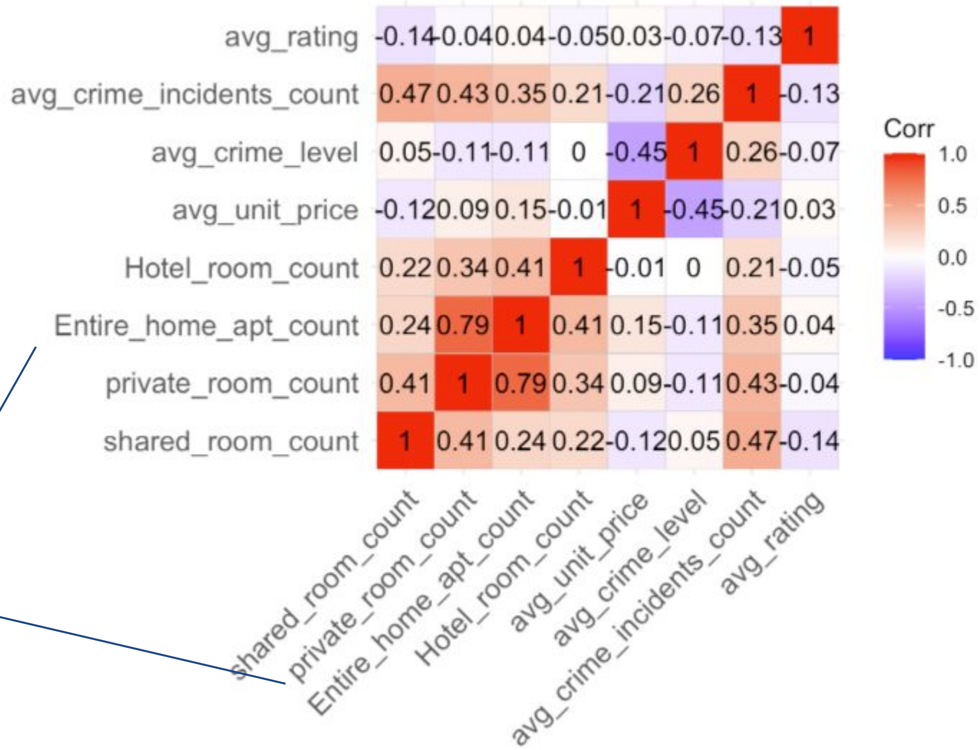
Attribute <chr>	Missing Values <int>	Unique Values <int>	Mean <dbl>	Min <dbl>	Max <dbl>	SD <dbl>
avg_unit_price	0	140	44.1892292	10.250000	218.166667	26.6507907
avg_rating	0	139	3.5191650	2.820500	4.000000	0.1571248
avg_crime_level	0	141	2.2554408	1.757919	2.933333	0.1746000
avg_crime_incidents_count	0	140	187.6457626	3.000000	1118.888361	177.5324483
shared_room_count	0	17	1.6737589	0.000000	30.000000	4.2903487
private_room_count	0	38	12.7588652	0.000000	100.000000	16.2691209
Entire_home_apt_count	0	67	45.9716312	0.000000	701.000000	87.3016890
Hotel_room_count	0	8	0.4113475	0.000000	15.000000	1.8129334

- 8 variables
  - Categories:
    avg_unit_price, avg_rating, avg_crime_level, avg_crime_incidents_count
  - Measures:
    shared_room_count, private_room_count, entire_home_apt_count, hotel_room_count

Analysis & Report

Correlation matrix

Still no significant high correlation between any two attributes

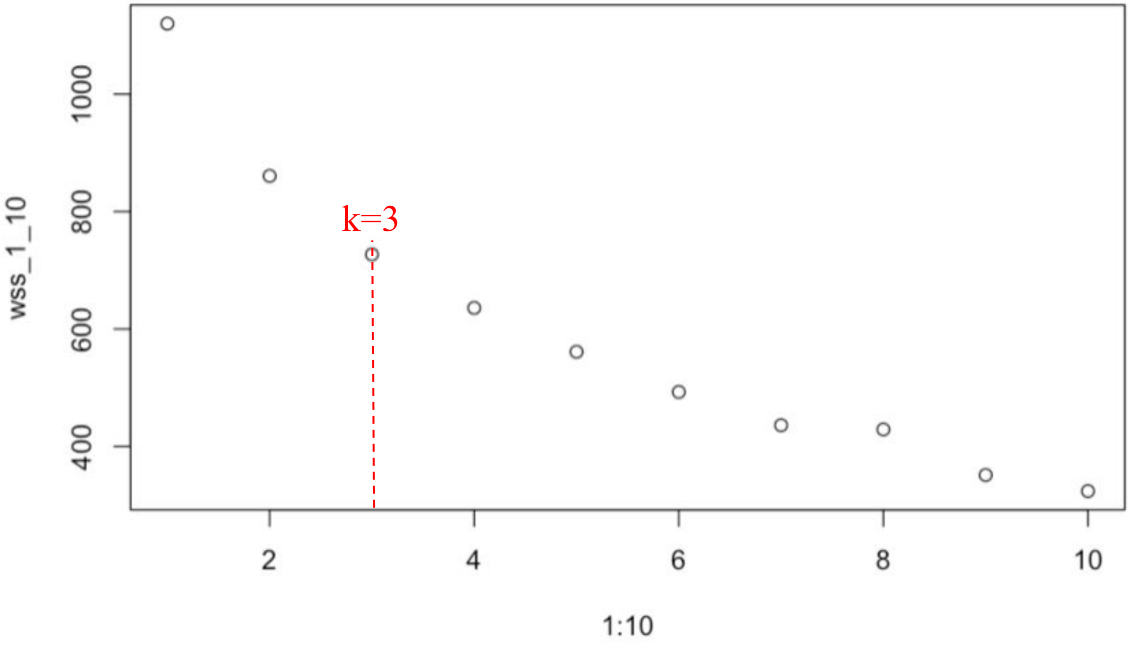


It's reasonable that entire housing and private room have a little close relation, since such places are supposed to have many housings.

Analysis & Report

Elbow diagram

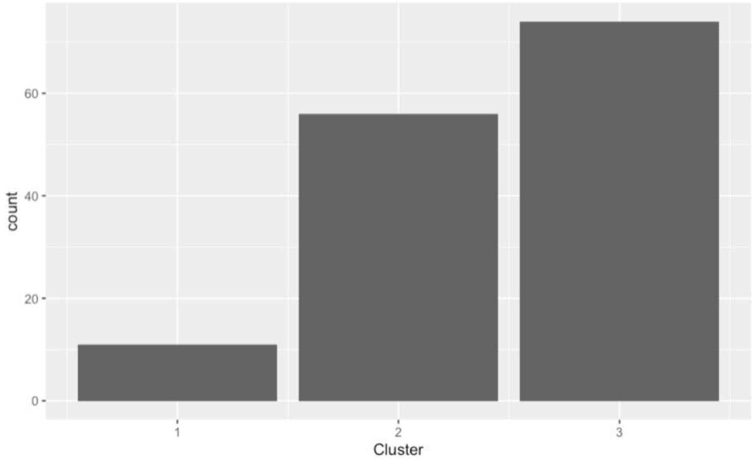
The elbow diagram shows that the optimal number of cluster is 3.

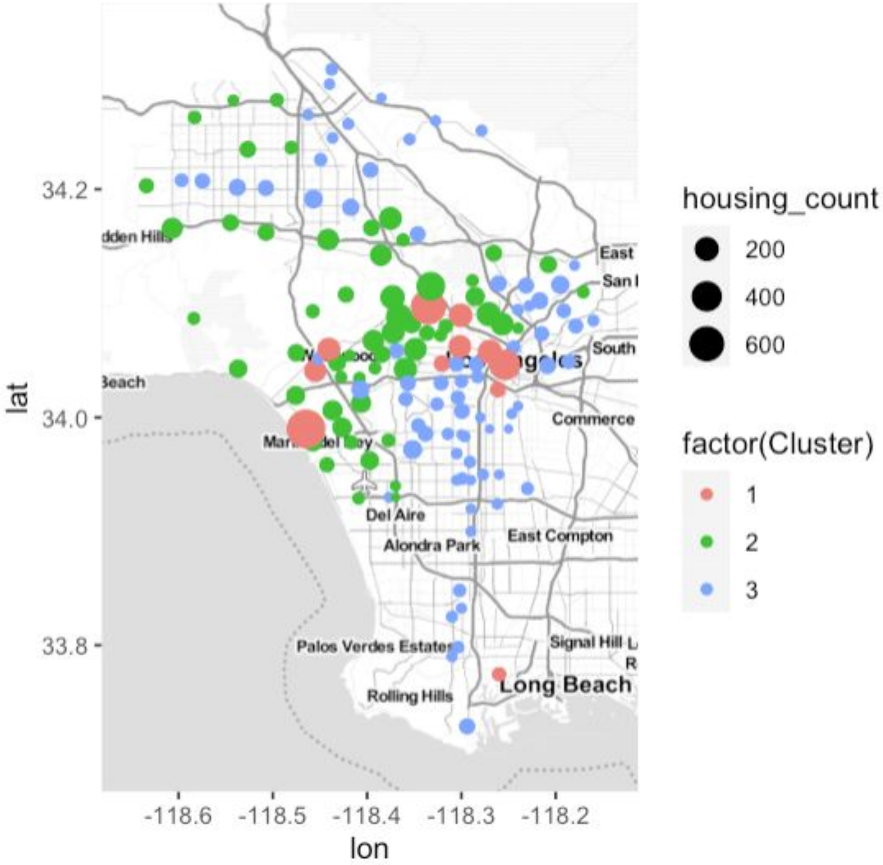


Analysis & Report

Clustering result

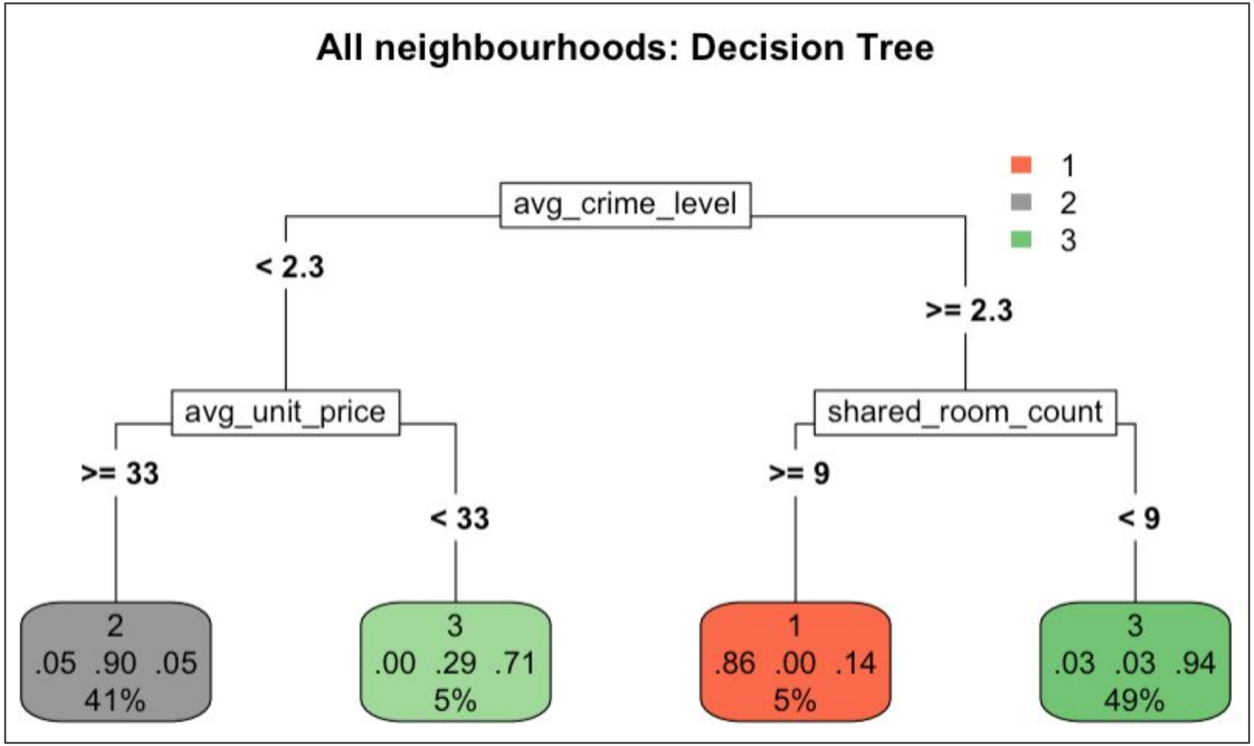
Each point represent a neighbourhood





Analysis & Report

DT for clustering result



Looks like average crime level is the most important factor for clustering.

Looking at the detailed numerical summary for each cluster in later slides can also prove this.

⇨ Target: cluster

Analysis & Report

Clustering Based on Numeric Demographic Attributes

- Cluster 1

Attribute <chr>	Missing Values <int>	Unique Values <int>	Mean <dbl>	Min <dbl>	Max <dbl>	SD <dbl>
shared_room_count	0	11	12.363636	0.000000	30.000000	8.66340265
private_room_count	0	11	42.909091	4.000000	100.000000	32.10437523
Entire_home_apt_count	0	11	196.090909	4.000000	701.000000	225.33550743
Hotel_room_count	0	6	4.454545	0.000000	15.000000	4.96716491
avg_unit_price	0	11	39.023689	21.723684	63.780013	11.43198371
avg_crime_level	0	11	2.280154	2.156374	2.371429	0.05972505
avg_crime_incidents_count	0	11	525.188866	103.360000	1118.888361	336.95410358
avg_rating	0	11	3.441701	3.277816	3.570078	0.08618687

- Cluster 2

Attribute <chr>	Missing Values <int>	Unique Values <int>	Mean <dbl>	Min <dbl>	Max <dbl>	SD <dbl>
shared_room_count	0	7	0.9642857	0.000000	12.000000	2.03571998
private_room_count	0	28	14.7321429	0.000000	54.000000	13.65615555
Entire_home_apt_count	0	44	58.6071429	0.000000	304.000000	61.63460615
Hotel_room_count	0	4	0.1607143	0.000000	3.000000	0.56493880
avg_unit_price	0	55	60.3219077	17.854167	218.166667	35.34840017
avg_crime_level	0	56	2.0962760	1.757919	2.289968	0.11367910
avg_crime_incidents_count	0	56	118.8611297	3.769231	292.686047	76.51909508
avg_rating	0	56	3.5454023	2.996143	4.000000	0.12641473

- Cluster 3

Attribute <chr>	Missing Values <int>	Unique Values <int>	Mean <dbl>	Min <dbl>	Max <dbl>	SD <dbl>
shared_room_count	0	9	0.6216216	0.000000	9.000000	1.78043171
private_room_count	0	20	6.7837838	0.000000	29.000000	7.05413562
Entire_home_apt_count	0	31	14.0945946	0.000000	60.000000	15.83144860
Hotel_room_count	0	1	0.0000000	0.000000	0.000000	0.00000000
avg_unit_price	0	74	32.7485664	10.250000	64.150000	8.15358222
avg_crime_level	0	74	2.3722163	2.122482	2.933333	0.12202629
avg_crime_incidents_count	0	73	189.5236720	3.000000	547.000000	140.28194074
avg_rating	0	73	3.5108247	2.820500	4.000000	0.18107084

Analysis & Report

Clustering Based on Numeric Demographic Attributes

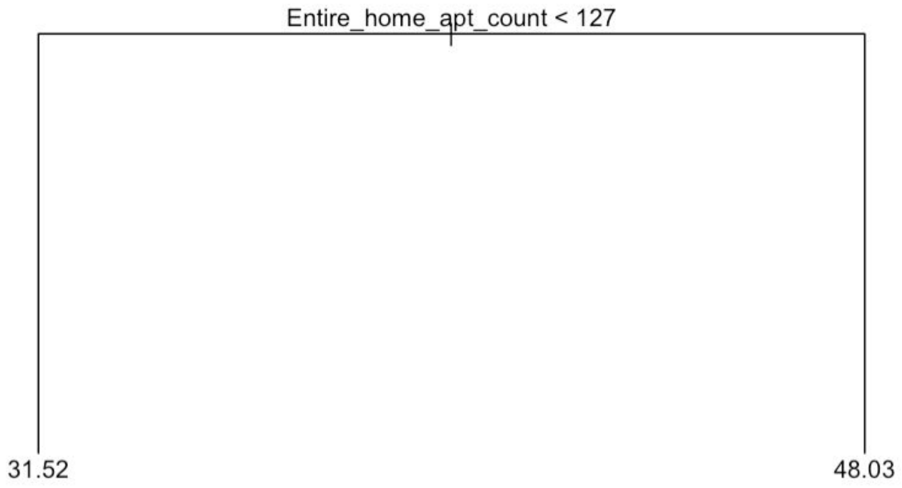
- Cluster 1

Attribute <chr>	Missing Values <int>	Unique Values <int>	Mean <dbl>	Min <dbl>	Max <dbl>	SD <dbl>
shared_room_count	0	11	12.363636	0.000000	30.000000	8.66340265
private_room_count	0	11	42.909091	4.000000	100.000000	32.10437523
Entire_home_apt_count	0	11	196.090909	4.000000	701.000000	225.33550743
Hotel_room_count	0	6	4.454545	0.000000	15.000000	4.96716491
avg_unit_price	0	11	39.023689	21.723684	63.780013	11.43198371
avg_crime_level	0	11	2.280154	2.156374	2.371429	0.05972505
avg_crime_incidents_count	0	11	525.188866	103.360000	1118.888361	336.95410358
avg_rating	0	11	3.441701	3.277816	3.570078	0.08618687

- High entire home apartment rented
- High private room rented
- Medium average unit price
- High average crime incidents happened

Analysis & Report

Decision Tree for each cluster (cluster 1)



The most important factor for cluster 1 is entire home apartment.

⇨ Target: unit price

Analysis & Report

Clustering Based on Numeric Demographic Attributes

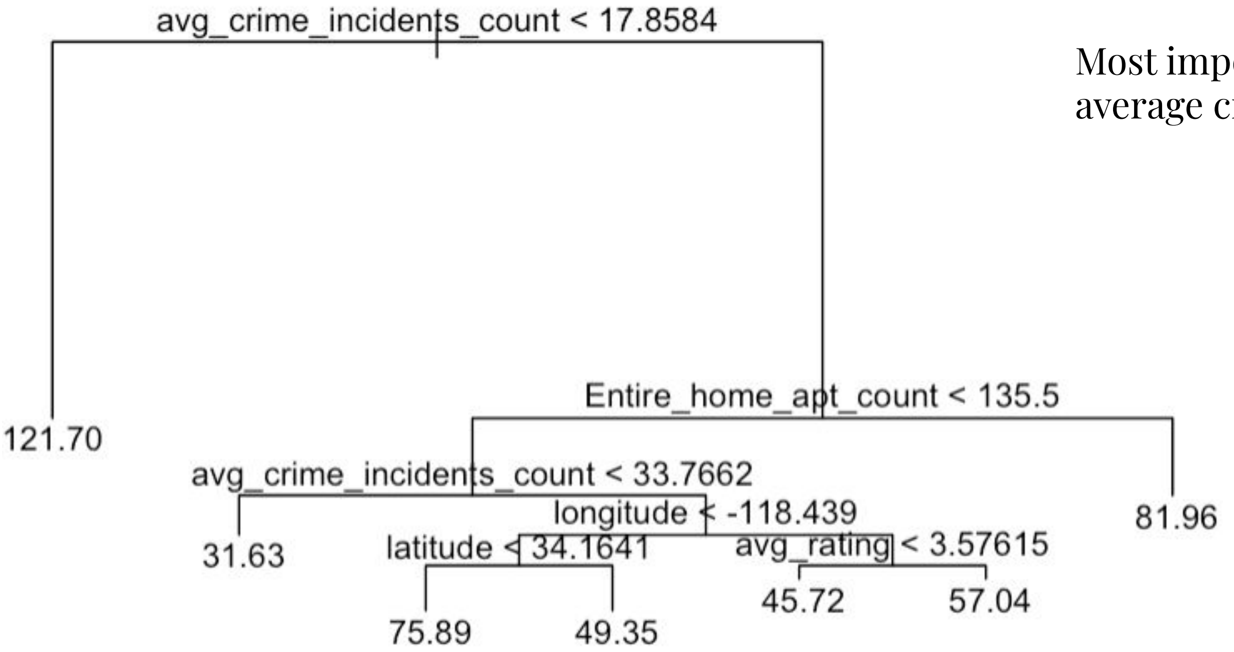
- Cluster 2

Attribute <chr>	Missing Values <int>	Unique Values <int>	Mean <dbl>	Min <dbl>	Max <dbl>	SD <dbl>
shared_room_count	0	7	0.9642857	0.000000	12.000000	2.03571998
private_room_count	0	28	14.7321429	0.000000	54.000000	13.65615555
Entire_home_apt_count	0	44	58.6071429	0.000000	304.000000	61.63460615
Hotel_room_count	0	4	0.1607143	0.000000	3.000000	0.56493880
avg_unit_price	0	55	60.3219077	17.854167	218.166667	35.34840017
avg_crime_level	0	56	2.0962760	1.757919	2.289968	0.11367910
avg_crime_incidents_count	0	56	118.8611297	3.769231	292.686047	76.51909508
avg_rating	0	56	3.5454023	2.996143	4.000000	0.12641473

- Medium entire home apartment rented
- Medium private room apartment rented
- Low average crime incidents happened
- High average unit price

Analysis & Report

Decision Tree for each cluster (cluster 2)



Most important factor of cluster 2 is average crime incidents count.

Target: unit price

Analysis & Report

Clustering Based on Numeric Demographic Attributes

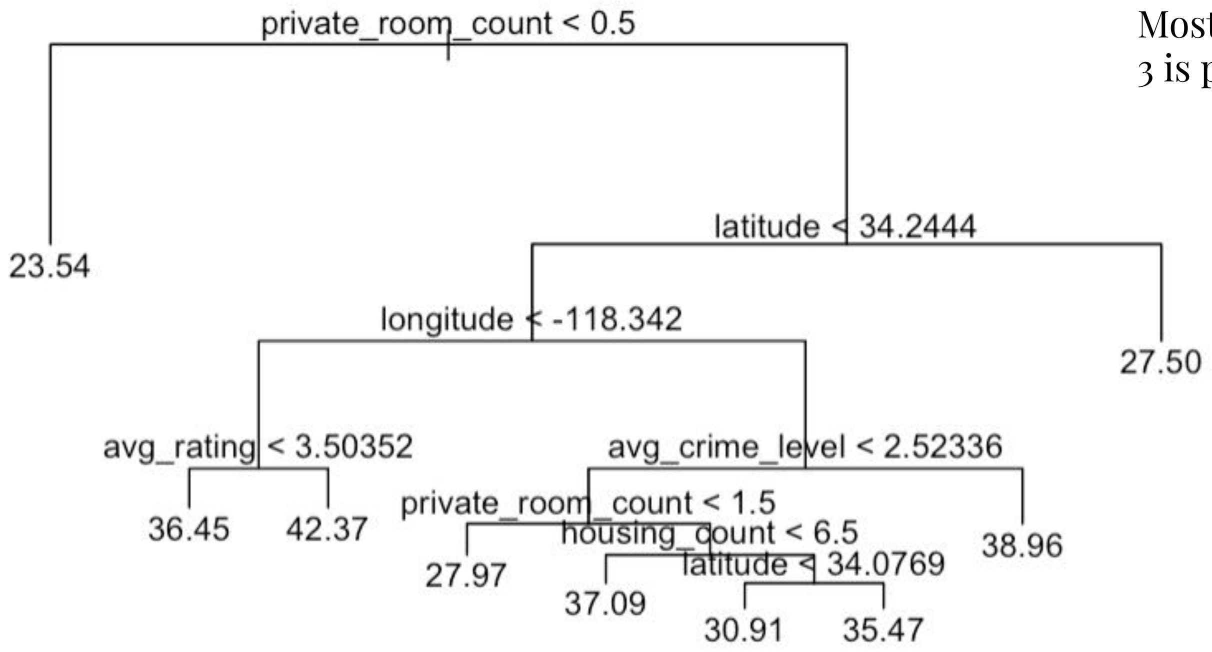
- Cluster 3

Attribute <chr>	Missing Values <int>	Unique Values <int>	Mean <dbl>	Min <dbl>	Max <dbl>	SD <dbl>
shared_room_count	0	9	0.6216216	0.000000	9.000000	1.78043171
private_room_count	0	20	6.7837838	0.000000	29.000000	7.05413562
Entire_home_apt_count	0	31	14.0945946	0.000000	60.000000	15.83144860
Hotel_room_count	0	1	0.0000000	0.000000	0.000000	0.00000000
avg_unit_price	0	74	32.7485664	10.250000	64.150000	8.15358222
avg_crime_level	0	74	2.3722163	2.122482	2.933333	0.12202629
avg_crime_incidents_count	0	73	189.5236720	3.000000	547.000000	140.28194074
avg_rating	0	73	3.5108247	2.820500	4.000000	0.18107084

- Low entire home apartment rented
- Low private room rented
- Medium average unit price
- Medium average crime incidents happened

Analysis & Report

Decision Tree for each cluster (cluster 3)



Most important factor of cluster 3 is private room count.

Target: unit price

Conclusion for Object2

- Housings around Hollywood, Marina Del Rey and DTLA are mostly clustered into Cluster 1. They have the highest count of entire housings be rented in total and highest criminal incidents.
- There are another group of housings surrounding Beverly Hills and Central La are clustered into Cluster 2. These housings are around popular location such as Universal Studio and Hollywood mountain and happened low criminal incidents which may be one of the reasons why they have highest unit price.
- The rest of locations may not be convenient spots to those famous locations are likely to be classified in Cluster 3. We can see that in this cluster there are low number of housings be rented.

Summary
&
Recommendations

Summary and Recommendations

General Observations

- Since we do the analysis specific on year 2020, there are limited sample data, the recommendation and analysis may not be general enough.
- In addition, most of the ratings are higher than 3 which means people who left comments and rating have a tendency of preferring the housings. However, those who dislike the housings probably won't even leave any comments or rating. Thus, the average rating of the housing seems to not be an important attribute for the analysis.

Summary and Recommendations

Recommendations

- If someone is considering where to buy the housing for leasing to visitors around LA, we will recommend that the housings around Hollywood and Beverly Hills are better choices.
  - The housings at such places with lowest criminal incidents which may also lead to take a low risk of getting damage of housings.
  - The unit price at these locations are compared to be highest which means the owner can earn more money here.
  - In the bivariate analysis, we can see that even though the price of housings at these places are higher, they received high rating.