Final Project
CS 327E: Elements of Databases

Group: FinalPush

(Che-Wei (Joanne) Chou, Tharit Tangkijwanichakul)

Dataset: Airbnb & Zillow - Short and Long Term Rent in U.S cities

Dataset we are working with

- Short-term rental price (daily) from Airbnb. The data mostly covers the the year of 2020-2021. The listing is categorized by room_type (hotel, private, shared, entire apartment)

- Long-term trend of real estate market (inventory, long-term rent contract price, home value)

- We are working on 3 different cities of Boston, LA, Austin which covers West, East and Central U.S.A

Questions we seek to answer

- (SQL1) For different cities, which city makes most sense to rent a house as opposed to buying one's own home?
- (SQL2) Does the increase in long-term housing rental price affect the popularity of short-term rent (Airbnb service)?
- (SQL3) Does the booming in real estate (home buying) affect the supply of Airbnb listings i.e. buying for commercial rent (NOT for residential purpose)?

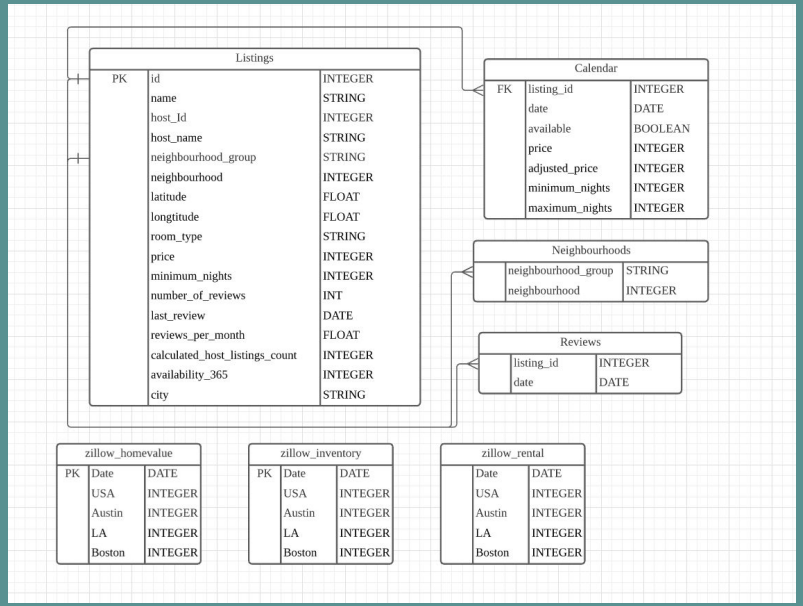
Overview of raw data

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	id	name	host_id	host_nam	neighbour	neighbour	latitude	longitude	room_typ	price	minimum_	number_c	last_revie
2	3781	HARBORSI	4804	Frank		East Bosto	42.36413	-71.0299	Entire hon	125	29	18	########
3	6695	$99 Specia	8229	Terry		Roxbury	42.32994	-71.0935	Entire hon	169	29	115	########
4	10813	Back Bay A	38997	Michelle		Back Bay	42.35061	-71.0879	Entire hon	70	29	5	########
5	10986	North End	38997	Michelle		North End	42.36352	-71.0508	Entire hon	73	29	2	########
6	13247	Back Bay s	51637	Susan		Back Bay	42.35164	-71.0875	Entire hon	75	91	0	
7	16384	Small Roo	23078	Eric		Beacon Hi	42.3581	-71.0713	Private roo	50	91	0	
8	18711	The Dorse	71783	Lance		Dorcheste	42.32212	-71.061	Entire hon	129	32	52	########
9	22195	Copley Ho	85130	Copley		Back Bay	42.34558	-71.0793	Private roo	114	1	28	########
10	22354	COPLEY SC	85770	Robert		South End	42.34496	-71.0749	Private roo	148	29	316	########

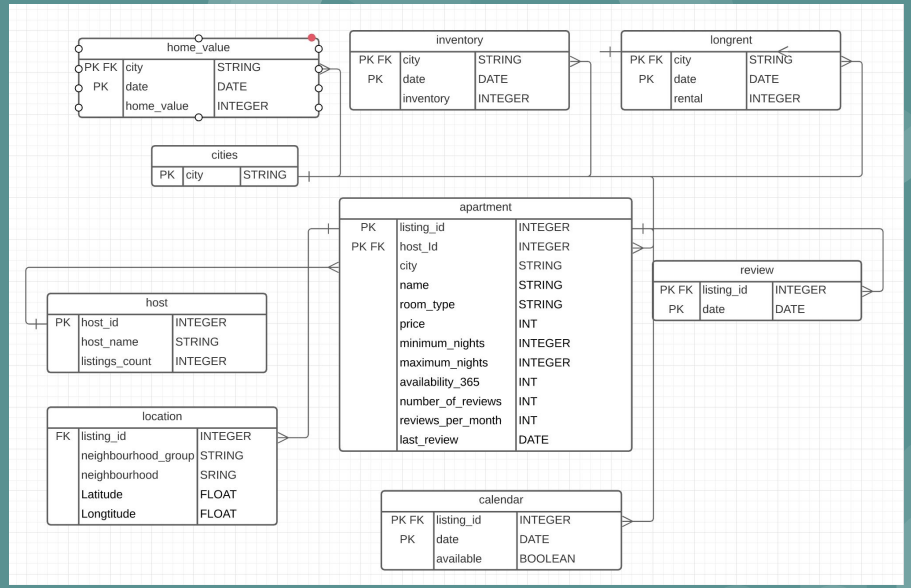
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	RegionID	RegionNa	SizeRank	2014-01	2014-02	2014-03	2014-04	2014-05	2014-06	2014-07	2014-08	2014-09	2014-10
2	102001	United Sta	0	1364	1371	1378	1386	1393	1400	1407	1414	1421	1429
3	394913	New York,	1	2388	2400	2412	2424	2436	2448	2460	2471	2483	2495
4	753899	Los Angele	2	1800	1817	1834	1851	1867	1884	1900	1916	1933	1949
5	394463	Chicago, Il	3	1493	1499	1505	1510	1516	1521	1526	1531	1536	1542
6	394514	Dallas-For	4	1175	1183	1190	1197	1204	1212	1219	1226	1233	1240
7	394974	Philadelph	5	1332	1336	1341	1345	1350	1354	1358	1363	1367	1372
8	394692	Houston, T	6	1213	1224	1236	1247	1258	1269	1280	1291	1301	1312
9	395209	Washingto	7	1826	1835	1843	1852	1859	1867	1875	1883	1890	1898
10	394856	Miami-For	8	1544	1552	1561	1569	1577	1586	1594	1603	1611	1619

ERD - Staging vs. Modeled tables

Staging (for one city)

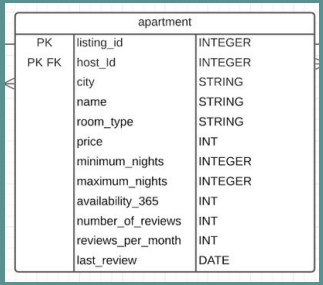


Modeled



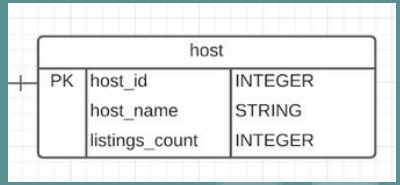
Beam pipeline: Apartment table

```python
class GroupHostListing(beam.DoFn):

    def process(self, element):
    # Rule: group Apartment entity with the same (host_id, listing_id) - PK
        return [((host_id,listing_id), record)]

class MakeUniqueApartment(beam.DoFn):
    # Make the unique apartment according to our numerical rules below
    def process(self, element):
        # Rules:
        # For the same (listing_id, host_id):
        # 1) sum the avalibility 365 (the data implies the host has many room in a unit)
        # -- the result may exceed 365
        # 2) sum the reviews_per_month, number_of_reviews
        # 3) min night = min (all of duplicated)
        # 4) max night = max (all of duplicated)
        # 5) Average the price across each duplicate
        return [record]
def run():
    ## Work starts here
    sql = 'SELECT .. '
    bq_source = ReadFromBigQuery()
    query_results = p |beam.io.Read(bq_source)
    # group by (host_id, listing_id)
    apartment_pcoll = query_results | beam.ParDo(GroupHostListing())
    grouped_apartment_pcoll = apartment_pcoll |beam.GroupByKey()
    # Make unique (host_id, listing_id)
    unique_apartment_pcoll = grouped_apartment_pcoll |beam.ParDo(MakeUniqueApartment())
    # write result to bq
```

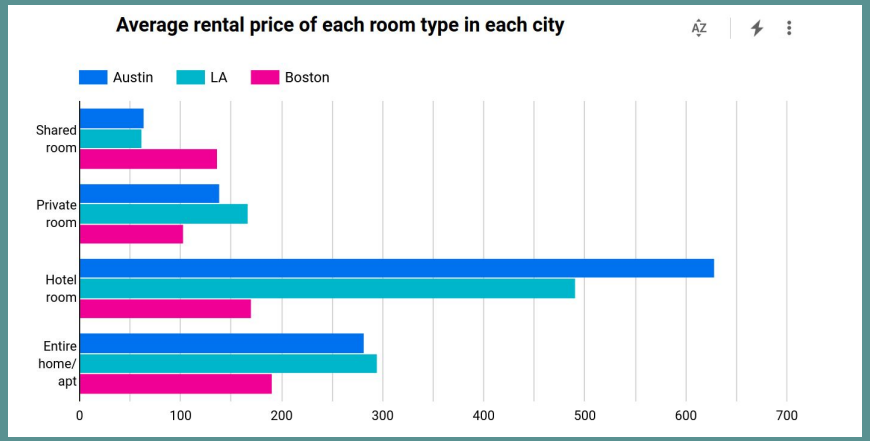


Beam pipeline: host table

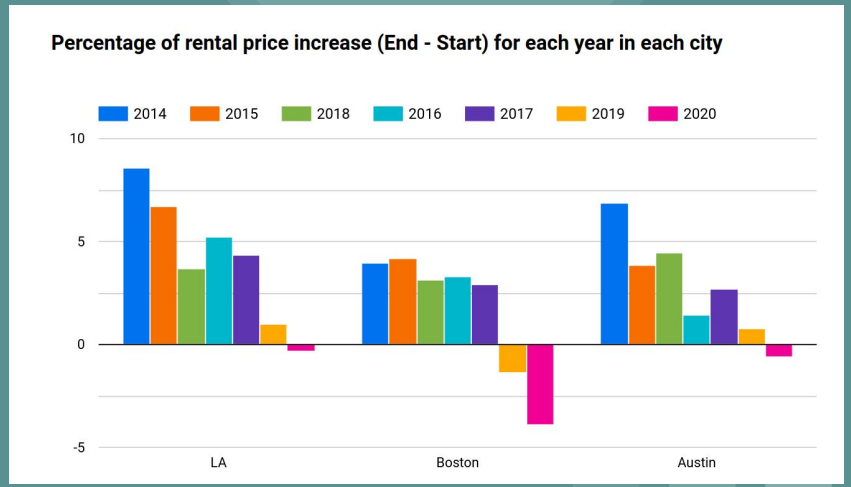
```python
class GroupHostListing(beam.DoFn):
    def process(self, element):
        # group host entity with the same (host_id) - PK of the table
        return [((host_id), record)]
class MakeUniqueHost(beam.DoFn):
    def process(self, element):
        # Rule:
        # For the same (host_id), sum the listings_count
        return [record]
def run():
    ## Work starts here
    sql = 'SELECT ...'
    query_results = p |beam.io.Read(bq_source)
    # group by (host_id)
    host_pcoll = query_results | beam.ParDo(GroupHostListing())
    grouped_host_pcoll = host_pcoll | beam.GroupByKey()
    # Make unique (host_id)
    unique_host_pcoll = grouped_host_pcoll | beam.ParDo(MakeUniqueHost())
    # write result to bq
```



Initial data exploration

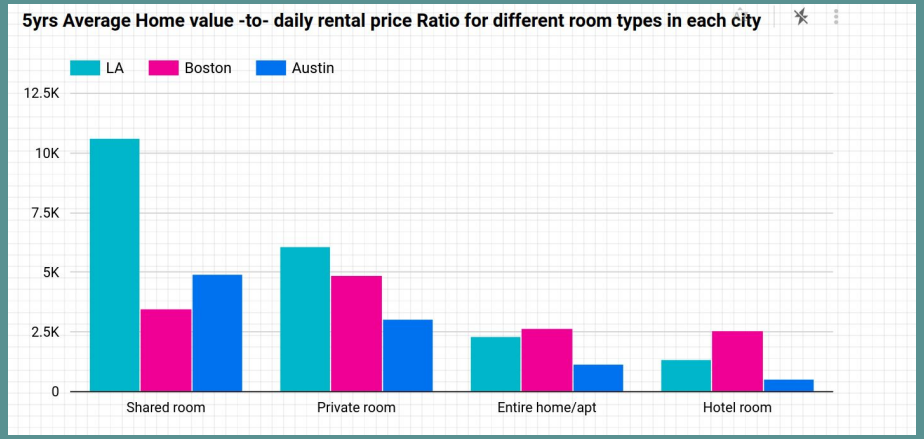


- Austin Hotel is the most expensive which is unexpected
- Austin room price in on-par with LA in many respects
- Boston rooms are clearly cheaper than others except in shared room
- Room price does not differ by large margin from city-to-city except the hotel room



- Boston % increase is relatively steady over time but it drops the most (by wide margin) in 2020
- Austin and LA experiences significant rental price increases. But during the downturn, they both drop only marginally
- What did happen in 2019-2020 that cause the downturn in housing market (short-term rent)?

SQL#1: For different cities, which city makes most sense to rent a house as opposed to buying one's own home (high multiple = value)




```sql
#CREATE VIEW reports.Homeval_Aptprice AS
SELECT home_value, city1 as city,AVG_Apt_price_day,Type_room,
(home_value/AVG_Apt_price_day) as Homeval_to_apt_ratio FROM
(
# Home value
(SELECT AVG(home.home_value) as home_value, home.city as city1
FROM datamart.home_value as home WHERE extract (year from home.date) > 2015
GROUP BY city1)
# END: Home value
JOIN
# Apt by room
(SELECT AVG(apt.price) as AVG_Apt_price_day, apt.city as city2,
apt.room_type as Type_room FROM datamart.apartment_Dataflow as apt
GROUP BY Type_room, city2)
# END: Apt by room
ON city1=city2
)
ORDER BY city1
```

- LA shared room is the best value (10k+ multiples) by wide margin (high home value and low rent)
- Generally for all three cities, it is most reasonable to rent Hotel Shared room > Private > Apartment > Hotel (ordered by high->low multiple)
- Austin rental price is expensive. Across 3 cities, buying home in Austin makes the most sense.

Row	home_value	city	AVG_Apt_price_day	Type_room	Homeval_to_apt_ratio
1	338552.01612903224	Austin	68.89473684210527	Shared room	4914.0475985115445
2	338552.01612903224	Austin	112.04755309325951	Private room	3021.50298496254
3	338552.01612903224	Austin	297.38432554634556	Entire home/apt	1138.4326174792657
4	338552.01612903224	Austin	647.2	Hotel room	523.1026207185294
5	471465.79032258055	Boston	178.1923696937134	Entire home/apt	2645.8247967236825
6	471465.79032258055	Boston	96.2467043314502	Private room	4898.513602075841

SQL#2: Does the increase in long-term housing rental price affect the popularity of short-term rent (Airbnb service) ?

3yrs Average % increase of Long-term rent vs 2020-2021 Airbnb occupancy rate for each city

	city	AVG_Percent_Rent_Inc	occupy_rate
1.	Austin	1.54	39.2
2.	LA	1.46	23.07
3.	Boston	-0.71	19.17

```sql
#CREATE VIEW reports.Rent_Inc_Occ AS
SELECT AVG_Percent_Rent_Inc, city1 as city, occupy_rate FROM(
# Rent 3-Yar Avg Inc
SELECT AVG(percent_rent_increase) as AVG_Percent_Rent_Inc, zcity1 FROM(
# Rent Inc each Year
SELECT 100*(rent2 - rent1)/rent1 as percent_rent_increase, year1, year2, zcity1
FROM
(SELECT AVG(lr1.rental) as rent1, extract (year from lr1.date) as year1, lr1.city as zcity1
FROM datamart.longrent as lr1 GROUP BY year1, lr1.city)
JOIN
(SELECT AVG(lr2.rental) as rent2, extract (year from lr2.date) as year2, lr2.city as zcity2
FROM datamart.longrent as lr2 GROUP BY year2, lr2.city)
ON year1+1=year2 and zcity1=zcity2)
WHERE year1>2017
# End: Rent Inc each Year
GROUP BY zcity1
# END: Rent 3-Yar Avg Inc
)
JOIN
```

```sql
(
# Occupancy
SELECT 100*(tot - not_use)/tot as occupy_rate, city1 FROM (
# Empty
(SELECT count(*) as not_use, apt.city as city1 FROM datamart.apartment_Dataflow as apt
WHERE apt.availability > 0 GROUP BY apt.city)
# End: Empty
JOIN
# Total
(SELECT count(*) as tot, apt2.city as city2 FROM datamart.apartment_Dataflow as apt2
GROUP BY apt2.city)
# End: Total
ON city1=city2)
# END: Occupancy
)
ON city1=zcity1
```

Row	AVG_Percent_Rent_Inc	city	occupy_rate
1	1.4594051990712307	LA	23.066553810464278
2	1.5393652917126843	Austin	39.19774121312433
3	-0.7128184398253552	Boston	19.174839364220492

SQL#3: Does the booming in real estate (home buying) affect the supply of Airbnb listings i.e. buy for commercial rent (NOT for residential purpose)?

Comparison of average inventory from 2018-2020 and average availability (in days) of 2020-2021 Airbnb rent for each city

	city	airbnb_avg_availability	avg_change_inventory
1.	LA	206.18	-8.08
2.	Boston	422.92	-13.25
3.	Austin	386.9	-3.01

1 - 3 / 3

- Boston experiences highest demand in real estate while we also see more supply goes to Airbnb (people buy home for renting out?)
- The average availability in days is very high probably due to Covid-19

*availability > 365 because we sum days for the same listing_id (dataflow pipeline)

```
#CREATE VIEW reports.Inventory_vs_Availability AS
SELECT airbnb_avg_availability, AVG(percent_inventory_increase) AS avg_change_inventory, city FROM
(
# Avg days avail
(SELECT AVG(availability) AS airbnb_avg_availability, a.city as city0 FROM datamart.apartment_Dataflow a GROUP BY city0)
# End: Avg days avail
JOIN
# Inv Inc
(SELECT 100*(inventory2 - inventory1)/inventory1 as percent_inventory_increase, inventory1, inventory2, year1, year2, city1 AS city
FROM
(SELECT AVG(li1.inventory) as inventory1, extract (year from li1.date) as year1, li1.city as city1
FROM datamart.inventory as li1
GROUP BY year1, li1.city )
JOIN
(SELECT AVG(li2.inventory) as inventory2, extract (year from li2.date) as year2, li2.city as city2
FROM datamart.inventory as li2
GROUP BY year2, li2.city)
ON year1+1=year2 and city1=city2)
# End: Inv Inc
ON city0=city
)
```

Row	airbnb_avg_availability	avg_change_inventory	city
1	206.18495877613734	-8.08343569627527	LA
2	386.8969915295498	-3.0078832053957525	Austin
3	422.91849847818753	-13.245385741870175	Boston

Future improvements

- Historical Airbnb data (multi-year) will help us to get year-by-year comparison. Right now we can only do N-yrs average of Zillow data and compare to 2020-2021 Airbnb data.
- Some attributes in Airbnb data (reviews) are very sparse and mostly non-existent. Otherwise, we will be able to compare the popularity of housing across different city or with the price etc.
- The Airbnb data which in mostly for a year 2020-2021 is not very representative because this year we have exogenous event.