California Real Estate Market Analytics

Jennifer Nghi M. Nguyen (010872316)

College of Professional and Global Education, San Jose State University

DATA 230: Data Visualization

Andrew H. Bond

December 12, 2021

Abstract

California is known as a hot state for real estate markets with high home median price ranges. Also, since the pandemic, lots of aspects of the economy are getting hit hard and there are lots of severe economic consequences. Therefore, this project aims to deliver visualizations which encapsulate the effects of pandemic on the California real estate market hotness and inventory, fast selling market vs slow selling market and market performance dashboard for 2020 vs 2021. In conclusion, the market in post-pandemic appears to be much cooler, selling demand is slower, buying demand remains not so much affected. Also, fast selling markets mostly focus in the Bay area and Southern California with dominant price ranges of $300K - $1M; whereas, slow selling markets mostly focus in Southern California with dominant price ranges of $300K - $600K and $1M+. Also, in California, in 2021, the median price increases, the hotness decreases in comparison to 2020.

1. Problems and Project Objectives

Since the pandemic, the real estate market in California appears to be much slower; however, the data show that the buyers still hardly make a bargain or acquire good home deals. According to a recent report by the Los Angeles Times in 2021, Los Angeles home prices this July have increased 12.6% from last year. For example, in Orange county, home median price has risen to 12.5% and in San Bernardino county, home median price has risen to 22.4%.

In order to obtain more understanding about effects of pandemic has on California real estate market , this project aims to introduce compare and contrast visualizations to break down the market performance in the pre and post pandemic. Plus, we also have visualizations to compare and contrast factors of the fast selling market and slow selling market. Also, we also provide a dashboard to see housing data performances by zip code in comparison to last year's.

The main objectives is to view real estate market in California in four perspectives:

- How has the pandemic affected market hotness?
- How has the pandemic affected market inventory?

- Fast selling market vs slow selling market
- California market performance 2020 vs 2021

Benefited users: these visualizations will benefit anyone who is interested in the real estate market such as investors, buyers, sellers, and real estate professionals.

2. Project Scopes

The scope of this project only focuses on building four visualizations using Tableau representing main objectives listed above.

This project doesn't cover:

- Constantly update data points.

3. Project Deliverables

FInal deliverables will include:

- Presentation: https://docs.google.com/presentation/d/19YXSSFeuPr_gICimFPR9mB70dcKSUGdu/edit?usp=sharing&ouid=101508950611960214124&rtpof=true&sd=true
- Github:
  - https://github.com/jennifernghi/Nghi_Nguyen_data_230_term_project or
  - https://github.com/sjsu-data230/Nghi_Nguyen_data_230_term_project
- Final report
- Tableau Dashboard Visualization
- Live Demo (Tableau stories + ppt): https://public.tableau.com/views/Californiarealestatemarketstory/Californiarealestatemarketstory?:language=en-US&:display_count=n&:origin=viz_share_link

4. Data Collection

Dataset

The dataset chosen for this project are Realtor's zip code hotness and inventory datasets (https://www.realtor.com/research/data/)

Key Definitions (from Realtor):

Table 1 is key data definitions used for this projects:

Table 1

Key Data Definitions

Column	**Definition**
Median Listing Price	The median listing price within the specified geography during the specified month.
Active Listing Count	The count of active listings within the specified geography during the specified month. The active listing count tracks the number of for sale properties on the market, excluding pending listings where a pending status is available.
Days on Market	The median number of days property listings spend on the market within the specified geography during the specified month. Time spent on the market is defined as the time between the initial listing of a property and either its closing date or the date it is taken off the market.
New Listing Count	The count of new listings added to the market within the specified geography.
Price Increase Count	The count of listings which have had their price increased within the specified geography.
Price Decrease Count	The count of listings which have had their price reduced within the specified geography.
Pending Listing Count	The count of pending listings within the specified geography during the specified month.
Total Listing Count	The total of both active listings and pending listings within the specified geography during the specified month.
Hotness Score	The hotness score is an equally-weighted composite metric of a geography's supply score and demand score.

5. Data Management

Schema Design

Figure 1 shows the data schema design for this project. Star schema is applied in this case with dimension tables for postal code data and time data namely postal_code_dim and time_dim respectively. This structure will help with filtering, slicing and dicing with data with respect to state, postal code and time related filters.

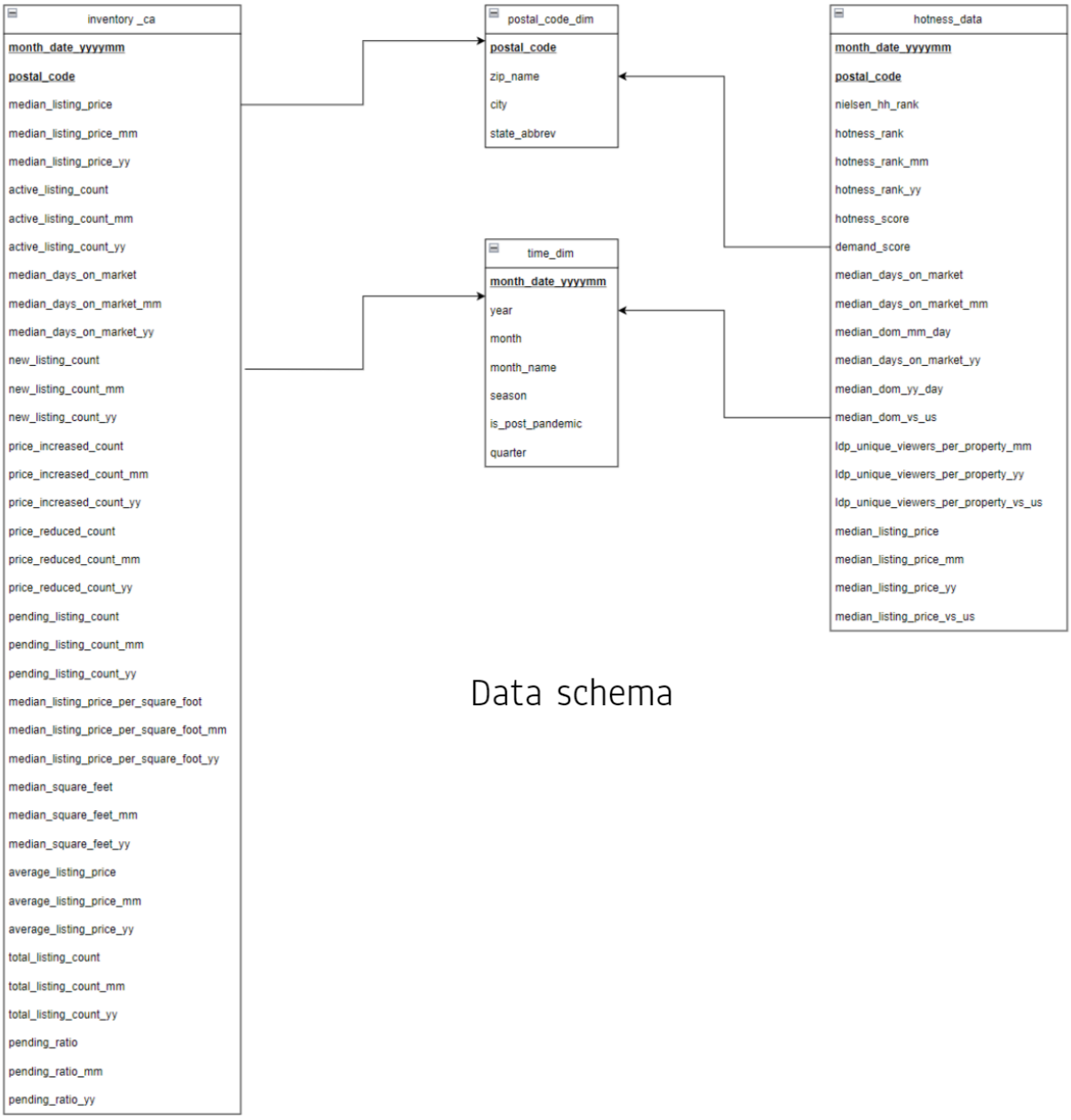
Table hotness_data is the zip code hotness data from Realtor and table inventory_ca is the zip code inventory data from Realtor. These two tables' records can be uniquely identified using combination primary key of month_date_yyyymm and postal_code

Table postal_code_dim's records can be uniquely identified using postal_code primary key.

Table time_dim's records can be uniquely identified using month_date_yyyymm primary key.

Figure 1

Data Schema with Star Schema Design

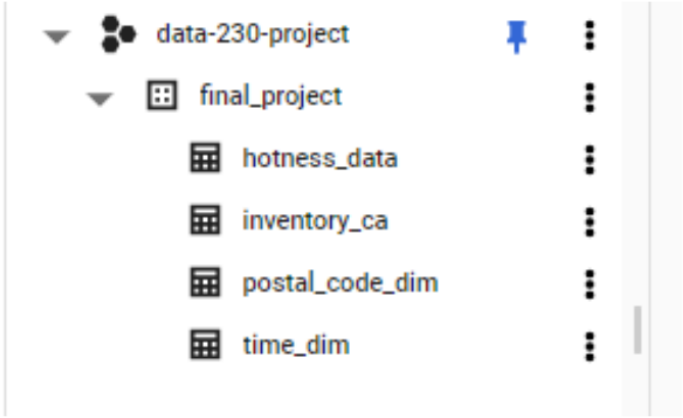


Storage

Data is extracted from Realtor as CSV and stored in GCP Bigquery for data accessing and management purposes. Figure 2 shows data stored in GCP Bigquery with tables specified in Figure 1.

Figure 2

GCP Bigquery Storage for Data Access



Bigquery for data storage and accessing

Sample Datasets

Figure 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12 show sample datasets of tables defined in Figure 1.

Figure 3

Table hotness_data - Part I

month_date_yyyymm	postal_code	zip_name	nielsen_hh_rank	hotness_rank	hotness_rank_mm	hotness_rank_yy	hotness_score	supply_score	demand_score
202001	2476	arlington, ma	6242	28	7244	84	97.69775	99.35986224	96.03563808
202001	95124	san jose, ca	1017	32	157	1944	97.62288	99.33365777	95.91210272
202001	94502	alameda, ca	7870	5	528	108	99.124022	99.24755737	99.00048665
202001	94117	san francisco, ca	745	2	397	32	99.427245	99.24755737	99.60693295
202001	94598	walnut creek, ca	3784	1	58	91	99.483398	99.24755737	99.71923782
202001	95054	santa clara, ca	4894	349	334	2440	92.372628	99.06412608	85.68112904
202001	94121	san francisco, ca	855	8	189	55	98.996743	99.06412608	98.92936024
202001	94703	berkeley, ca	4811	3	60	184	99.273762	99.06412608	99.4833976
202001	94044	pacifica, ca	1899	9	99	212	98.940591	98.95556471	98.92561674
202001	94066	san bruno, ca	1471	23	86	465	98.053382	98.94059072	97.16617377

Figure 4

Table hotness_data - Part II

median_days_on_market	median_days_on_market_mm	median_dom_mm_day	median_days_on_market_yy	median_dom_yy_day	median_dom_vs_us	ldp_unique_viewers_per_property_mm
5.5	-0.968660969	-170.0	-0.877777778	-39.5	-80.0	0.329411765
6.0	-0.857142857	-36.0	-0.7	-14.0	-79.5	0.555555556
6.5	-0.888888889	-52.0	-0.857142857	-39.0	-79.0	-0.399280576
6.5	-0.883928571	-49.5	-0.1875	-1.5	-79.0	0.36196319
6.5	-0.845238095	-35.5	-0.835443038	-33.0	-79.0	0.769784173
7.5	-0.872881356	-51.5	-0.84375	-40.5	-78.0	-0.234042553
7.5	-0.845360825	-41.0	-0.727272727	-20.0	-78.0	0.547169811
7.5	-0.807692308	-31.5	-0.810126582	-32.0	-78.0	1.217391304
8.0	-0.802469136	-32.5	-0.728813559	-21.5	-77.5	1.0
8.5	-0.795180723	-33.0	-0.753623188	-26.0	-77.0	0.465116279

Figure 5

Table hotness_data - Part III

ldp_unique_viewers_per_property_yy	ldp_unique_viewers_per_property_vs_us	median_listing_price	median_listing_price_mm	median_listing_price_yy	median_listing_price_vs_us
-0.130769231	3.1389	924000.0	-0.013874066	0.321888412	3.0801
2.393939394	3.1111	1223500.0	0.013250518	-0.058121632	4.0784
0.176056338	4.6389	949000.0	0.137207909	0.193710692	3.1634
1.581395349	6.1667	1245500.0	-0.176256614	-0.09385231	4.1517
1.342857143	6.8333	1097500.0	0.189701897	0.310447761	3.6584
1.25	2.0	843944.0	-0.350312548	-0.155961468	2.8132
0.802197802	4.5556	1898975.0	0.085438697	0.270217391	6.33
1.615384615	5.6667	1200000.0	-0.094339623	0.230769231	4.0001
1.484848485	4.5556	1049000.0	-0.040695016	0.079218107	3.4967
1.333333333	3.5	1189000.0	0.081400637	0.324053452	3.9634

Figure 6

Table inventory_ca - Part I

month_date_yyyymm	postal_code	zip_name	median_listing_price	median_listing_price_mm	median_listing_price_yy	active_listing_count	active_listing_count_mm
202110	94306	palo alto, ca	2999888	-0.0549	0.1111	28	0.0
202110	92324	colton, ca	435000	0.0308	*null*	43	0.1622
202110	94526	danville, ca	1649000	-0.0006	0.2685	27	0.1739
202110	93555	ridgecrest, ca	275000	0.0581	*null*	106	0.2471
202110	90059	los angeles, ca	524995	0.0437	*null*	36	0.0
202110	91901	alpine, ca	869000	0.0052	0.0931	32	-0.0857
202110	95247	murphys, ca	619000	-0.0221	0.2405	37	0.0
202110	95482	ukiah, ca	615000	0.0425	-0.0752	47	0.2703
202110	93465	templeton, ca	1675000	0.4732	0.3673	13	0.0833
202110	94503	american canyon, ca	789000	-0.0652	0.2624	18	-0.1

Figure 7

Table inventory_ca - Part II

active_listing_count_yy	median_days_on_market	median_days_on_market_mm	median_days_on_market_yy	new_listing_count	new_listing_count_mm	new_listing_count_yy	price_increased_count
0.0769	39	0.258064516	0.444444444	24	0.2	0.2	4
null	32	-0.111111111	*null*	40	0.4286	0.4286	4
-0.4	25	0.136363636	-0.074074074	56	0.4	-0.1765	0
null	37	-0.108433735	*null*	56	0.4	0.4	0
0.44	63	0.726027397	-0.3	20	0.25	0.0	4
-0.0588	49	0.272727273	*null*	20	0.6667	-0.2857	0
-0.2292	44	-0.137254902	0.047619048	16	-0.2	1.0	0
0.3824	47	0.0	0.044444444	20	0.25	-0.1667	0
-0.5667	41	-0.364341085	-0.568421053	8	1.0	0.0	0
null	41	-0.035294118	0.108108108	12	0.0	0.0	0

Figure 8

Table inventory_ca - Part III

price_increased_count_mm	price_increased_count_yy	price_reduced_count	price_reduced_count_mm	price_reduced_count_yy	pending_listing_count	pending_listing_count_mm
null	*null*	4	*null*	-0.5	28	0.1667
null	*null*	4	-0.5	0.0	65	0.0656
null	*null*	4	0.0	-0.6667	51	0.0408
null	*null*	32	0.1429	0.6	67	0.0308
null	*null*	8	*null*	*null*	37	-0.26
null	*null*	4	-0.5	-0.5	49	0.0208
null	*null*	8	0.0	-0.6	18	0.0588
null	*null*	16	3.0	0.3333	30	0.0
null	*null*	4	*null*	0.0	15	-0.1176
null	*null*	4	0.0	*null*	12	-0.2

Figure 9

Table inventory_ca - Part IV

pending_listing_count_yy	median_listing_price_per_square_foot	median_listing_price_per_square_foot_mm	median_listing_price_per_square_foot_yy	median_square_feet	median_square_feet_mm
0.4737	1401.36601	0.0532	0.1301	2076	0.0538
-0.1875	306.8181818	0.0369	0.2898	1419	0.0157
-0.3929	714.4714038	-0.0159	0.1943	2003	-0.0603
-0.3093	168.8178528	-0.0037	0.1711	1603	0.0085
-0.0513	445.5357143	0.0198	0.1203	1140	0.0179
-0.125	370.5489614	0.01	0.1549	2224	-0.05
-0.1818	269.9718045	0.0	0.0721	2343	0.0121
-0.1667	340.6809619	0.0	0.0482	1824	0.0562
-0.2857	493.3114035	0.0091	0.1337	2634	0.218
-0.5	346.5542885	0.0025	0.2002	2082	-0.2059

Figure 10

Table inventory_ca - Part V

median_square_feet_yy	average_listing_price	average_listing_price_mm	average_listing_price_yy	total_listing_count	total_listing_count_mm	total_listing_count_yy	pending_ratio	pending_ratio_mm	pending_ratio_yy
0.0458	3794781.25	-0.0162	0.2318	56	0.0769	0.2444	1.0	0.1429	0.2692
-0.0515	477056.4237	0.0019	0.0985	108	0.102	0.0189	1.511627907	-0.137	-1.5653
-0.0384	1873850.5	-0.0229	0.2084	78	0.0833	-0.3953	1.888888889	-0.2415	0.0222
-0.0856	284606.2727	0.0309	0.096	173	0.1533	0.1234	0.632075472	-0.1326	-1.0697
0.1004	513348.0698	-0.0041	0.1933	73	-0.1512	0.1406	1.027777778	-0.3611	-0.5322
-0.1544	1056087.5	0.031	-0.013	81	-0.0241	-0.1	1.53125	0.1598	-0.1158
0.1463	838791.5122	-0.0033	0.0587	55	0.0185	-0.2143	0.486486486	0.027	0.0282
-0.0172	1342781.444	-0.0966	-0.1973	77	0.1493	0.1	0.638297872	-0.1725	-0.4205
0.1422	2693564.286	0.3533	0.5416	28	-0.0345	-0.451	1.153846154	-0.2628	0.4538
-0.0536	1339081.25	-0.0578	0.9936	30	-0.1429	-0.1667	0.666666667	-0.0833	-1.3333

Figure 11

Table postal_code_dim

state_abbrev	city	zip_name	postal_code
AK	homer	homer, ak	99603
AK	kenai	kenai, ak	99611
AK	sitka	sitka, ak	99835
AK	haines	haines, ak	99827
AK	juneau	juneau, ak	99801
AK	kodiak	kodiak, ak	99615
AK	palmer	palmer, ak	99645
AK	willow	willow, ak	99688
AK	chugiak	chugiak, ak	99567
AK	wasilla	wasilla, ak	99623

Figure 12

Table time_dim

month_date_yyyymm	year	month	month_name	season	is_post_pandemic	quarter
201901	2019	01	January	Winter	0	Q1
201801	2018	01	January	Winter	0	Q1
201802	2018	02	Febuary	Winter	0	Q1
201902	2019	02	Febuary	Winter	0	Q1
201903	2019	03	March	Spring	0	Q1
201803	2018	03	March	Spring	0	Q1
201804	2018	04	April	Spring	0	Q2
201904	2019	04	April	Spring	0	Q2
201805	2018	05	May	Spring	0	Q2
201905	2019	05	May	Spring	0	Q2

6. Visualizations and use cases

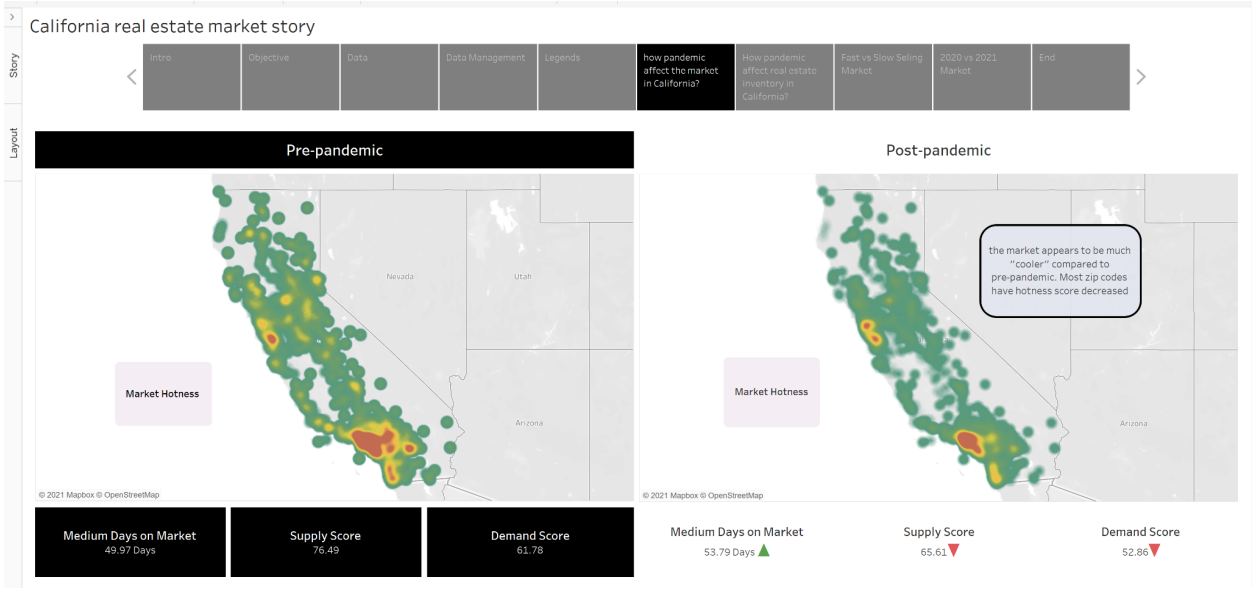
Visualization Legends

- $: listing price <= $300,000 - low range
- $$: $300,000 < listing price <= $600,000 - medium range
- $$$: $600,000 < listing price <= $1,000,000 - high range
- $$$$: listing price > $1,000,000 - luxury range
- Pre-pandemic: before Jan 2020
- Post- pandemic: On and after Jan 2020
- Fast selling market: properties stay on market for less than 60 days
- Slow selling market: properties stay on market for more than 100 days

Visualization 1: How has the pandemic affected market hotness in California?

Figure 13

Visualization 1: How pandemic affects the Market Hotness in California?



This story shows the contrast of market hotness pre and post pandemic in California. The density map indicates the market hotness using hotness_score in hotnes_data table.

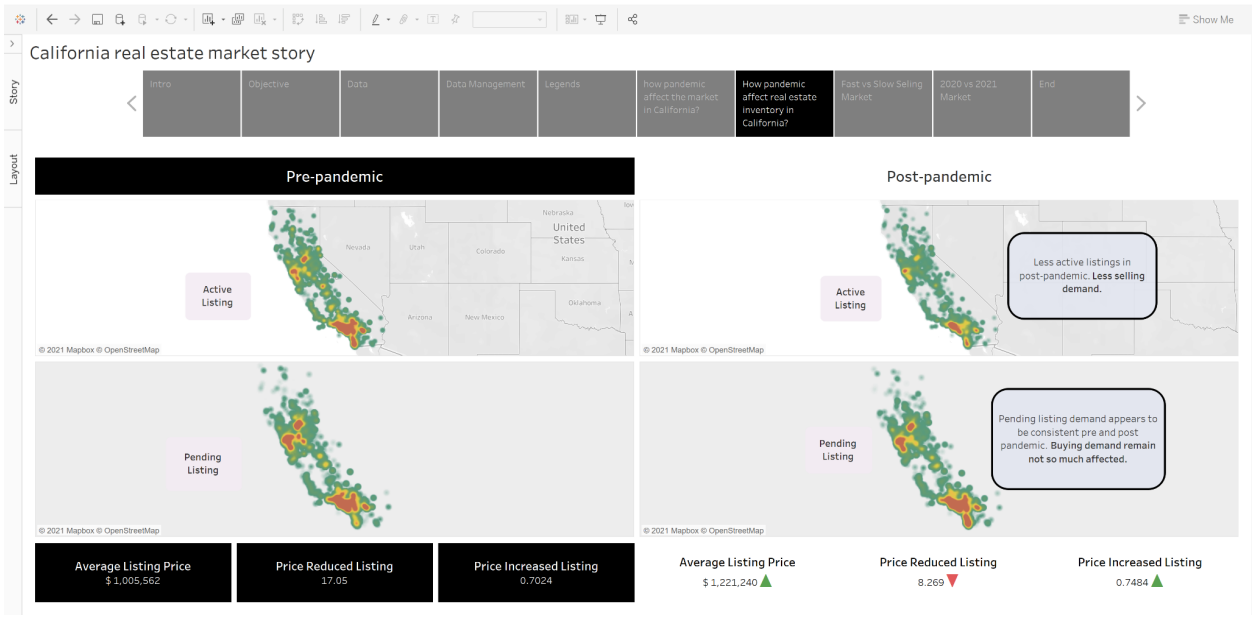
We can see that in post-pandemic (after January 2020) the market appears to be much cooler compared to pre-pandemic (before January 2020). The most affected regions are northern california and southern California (around Los Angeles). The Bay area still remains hot in general.

Also, housing average median days on market increases, average supply score and average demand score decrease. This indicates that properties are taking longer to close and both housing supply and demand needs are both decreasing.

Visualization 2: How has the pandemic affected market inventory in California?

Figure 14

Visualization 2: How Pandemic Affects The Market Inventory in California?



This story shows the contrast of market inventory pre and post pandemic in California. The density map indicates the market inventory with respect to active listings and pending listings using active_listing_count and pending_listing_count in the inventory_ca table.

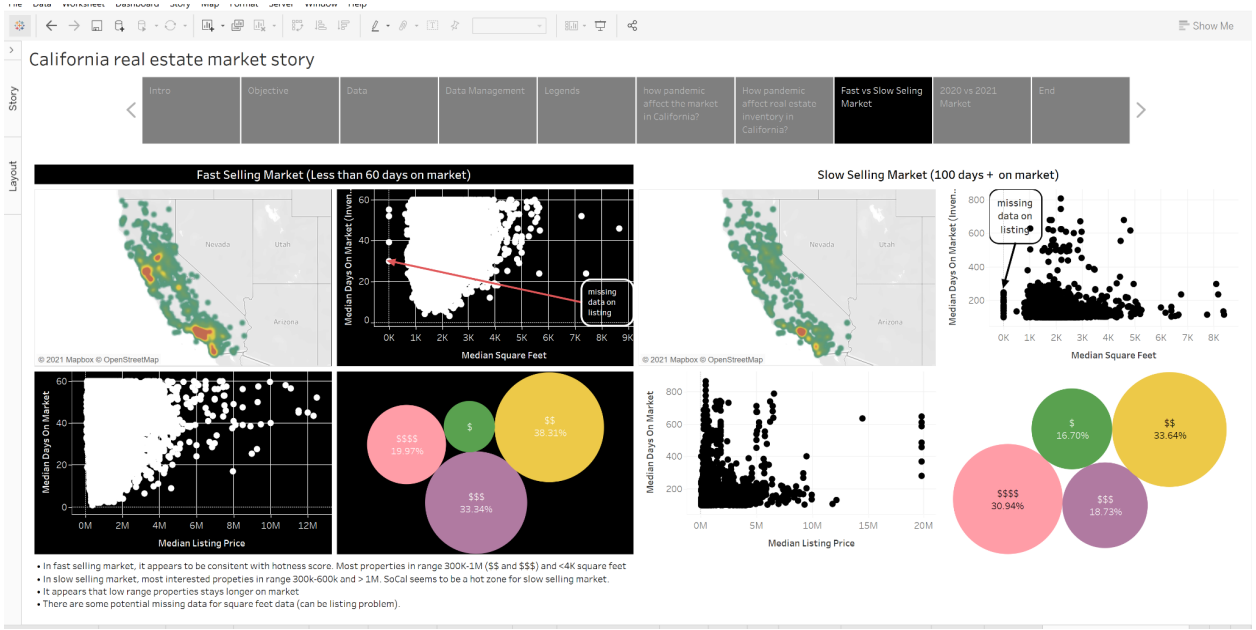
We can see that in post-pandemic (after January 2020), there are less active listings than in pre-pandemic which indicates selling demands are slowing down. However, for pending listings, there are not many changes which can indicate that buying demand in post-pandemic remains unchanged in comparison to pre-pandemic.

Also, in general, the average listing price increases, the average number of properties with price reduced are decreasing and the average number of properties with price increased are increasing.

Visualization 3: Fast selling market vs slow selling market

Figure 15

Visualization 3: Fast Selling Market vs Slow Selling Market.



This story shows the comparison and contrast between the fast selling market (less than 60 days on market) and slow selling market (more than 100 days on market).

In the fast selling market, we can see most properties are focused in the Bay area and southern California (around Los Angeles and San Diego). Most homes in this market have $$ ($300K - $600K) and $$$ ($600K - $1M) price ranges. Most homes have median square feet ranging from 1,000 - 4,000 square feet.

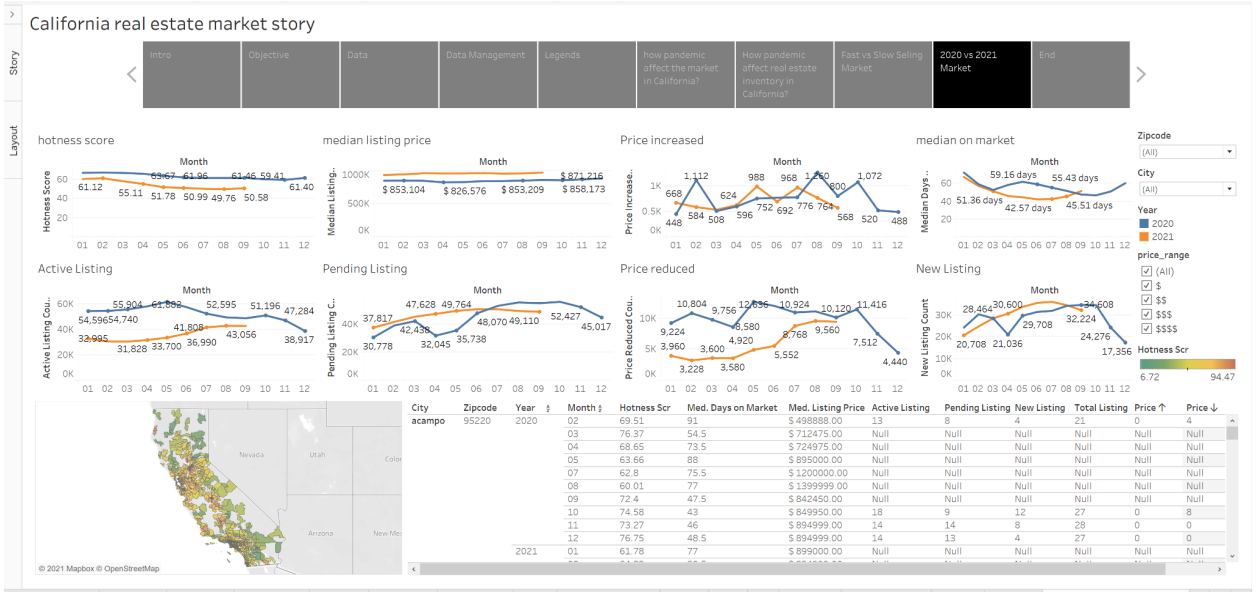
In the slow selling market, we can see most properties are focused in southern California (around Los Angeles). Most homes in this market have $$ ($300K - $600K) and $$$$ ($1M+) price ranges. Most homes have median square feet ranging from 1,000 - 3,000 square feet.

We can see there is some missing data for median square feet which could be a listing problem to begin with. Also, it appears that low range homes tend to stay longer on the market.

Visualization 4: California market performance in 2020 vs 2021

Figure 16

Visualization 4: Overall 2020 vs 2021 performance dashboard



This dashboard shows performance comparisons in 2020 vs 2021 (up to September).

Following measures are included:

- Hotness score by month 2020 vs 2021.
- Median listing price by month 2020 vs 2021.
- Price increased properties by month 2020 vs 2021.
- Price reduced properties by month 2020 vs 2021.
- Median days on market by month 2020 vs 2021.
- Active listing by month 2020 vs 2021.
- Pending listing by month 2020 vs 2021.
- New listing by month 2020 vs 2021.
- Median listing price by zip code map
- Detailed table which encapsulated all data above as shown in Figure 17.

Filters: data can be filtered using city, zip codes, price range and year filters. Also, this dashboard is interactive, data can be changed if zip code(s) is selected on the map or in the table.

Overall for California in 2021 (compared to 2020), the hotness decreases, the median price listing increases, properties' median days on market decreases.

This dashboard will help any user of interest to gain insights of current market trends and whether they should do next given the data.

Figure 17

Detailed look for table in Figure 16

City	Zipcode	Year	Month	Hotness Scr	Med. Days on Market	Med. Listing Price	Active Listing	Pending Listing	New Listing	Total Listing	Price ↑	Price ↓
acampo	95220	2020	02	69.51	91	$ 498888.00	13	8	4	21	0	4
			03	76.37	54.5	$ 712475.00	Null	Null	Null	Null	Null	Null
			04	68.65	73.5	$ 724975.00	Null	Null	Null	Null	Null	Null
			05	63.66	88	$ 895000.00	Null	Null	Null	Null	Null	Null
			07	62.8	75.5	$ 1200000.00	Null	Null	Null	Null	Null	Null
			08	60.01	77	$ 1399999.00	Null	Null	Null	Null	Null	Null
			09	72.4	47.5	$ 842450.00	Null	Null	Null	Null	Null	Null
			10	74.58	43	$ 849950.00	18	9	12	27	0	8
			11	73.27	46	$ 894999.00	14	14	8	28	0	0
			12	76.75	48.5	$ 894999.00	14	13	4	27	0	0
		2021	01	61.78	77	$ 899000.00	Null	Null	Null	Null	Null	Null
			02	64.89	80.5	$ 894999.00	Null	Null	Null	Null	Null	Null
			03	53.67	98	$ 1071999.50	Null	Null	Null	Null	Null	Null
			04	65.78	40.5	$ 922499.50	13	14	12	27	0	0
			05	73.39	34	$ 829000.00	Null	Null	Null	Null	Null	Null
			06	52.76	58.5	$ 957500.00	Null	Null	Null	Null	Null	Null
			07	51.03	55	$ 774950.00	Null	Null	Null	Null	Null	Null
			08	66.33	36	$ 836999.50	Null	Null	Null	Null	Null	Null
			09	59.84	47	$ 846999.50	13	3	4	16	0	0
acton	93510	2020	01	52.18	89.5	$ 594450.00	56	22	12	78	0	8
			02	46.04	95	$ 599900.00	43	34	16	77	0	12
			03	51	64	$ 647000.00	39	32	20	71	0	8
			04	66.03	44.5	$ 678700.00	37	28	12	65	0	0
			05	65.33	52	$ 665000.00	41	32	24	73	0	8
			06	56.39	61	$ 657950.00	33	45	8	78	0	12
			07	47.42	87.5	$ 664999.50	23	45	16	68	0	12
			08	69.11	51	$ 599750.00	21	48	20	69	0	4
			09	65.84	36.5	$ 802475.00	21	51	20	72	0	4
			10	68.8	32	$ 720000.00	20	54	28	74	0	4
			11	63.19	39.5	$ 719500.00	25	50	20	75	0	8
			12	61.64	55	$ 730000.00	14	44	0	58	0	0

Figure 18 shows another scenario by setting a city filter with San Jose. The dashboard now reflects all changes and focuses only on San Jose data.

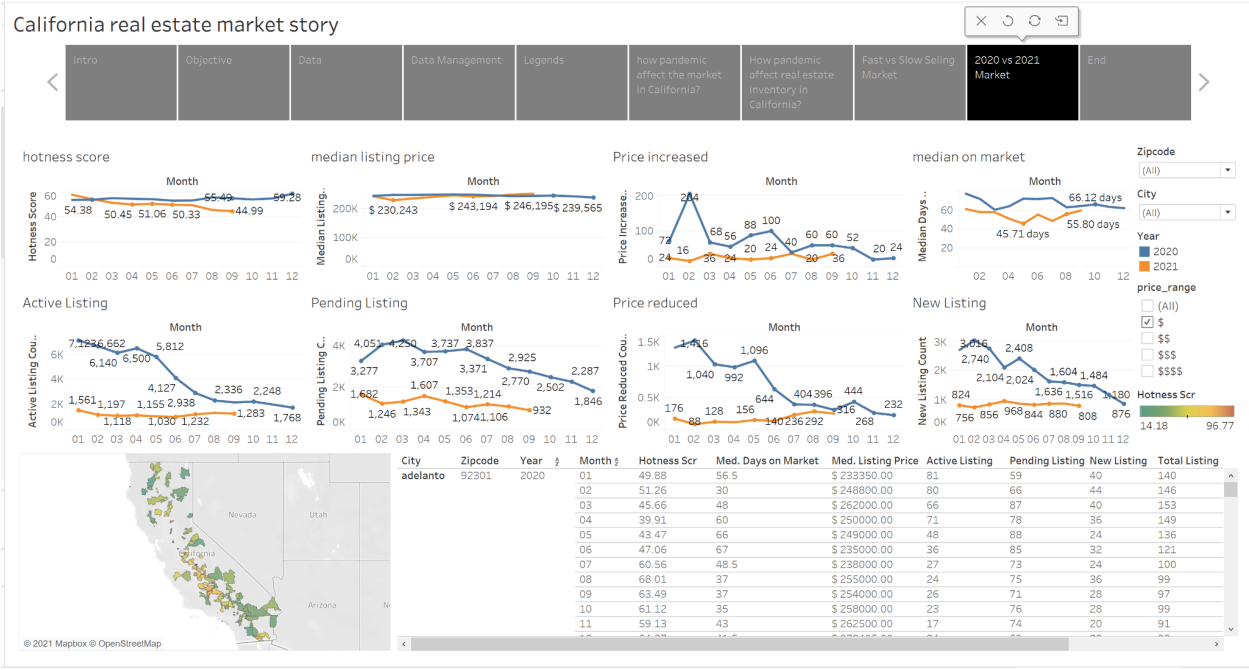
Figure 18

Visualization 4: 2020 vs 2021 performance dashboard - San Jose city

Figure 19 shows a scenario for only the $ (<=$300K) price range filter. The dashboard reflects the change of data. We can see that most of these properties are located along the center regions of California.

Figure 19

Visualization 4: 2020 vs 2021 performance dashboard - low price range homes (<=$300K)



7. Conclusion

With these Tableau visualizations, we have successfully delivered what promised in the project objectives. From here, benefited users can use these visualizations and obtain analytics about the effects of pandemic has had on the real estate market performance in California. Also, they can see more analytics for the fast vs slow selling market. Plus, users can explore California real estate market performance in 2020 vs 2021 with various filters and therefore they can gain more insights about the current market and make beneficial decisions.

California market hotness in post pandemic:

- California Market hotness decreases in general.
- Both supply and demand decrease
- Properties's median days on market getting longer

California market inventory in post pandemic:

- Selling demand decreases

- Buying demand remains not so much affected,
- Average listing price increases.
- Properties with reduced prices decrease.
- Properties with increased prices increase.

California fast selling market:

- Mostly focused in the Bay area and southern California.
- Mostly ranging from $300K - $1M.
- Mostly ranging from 1000 - 4000 square feet.

California slow selling market:

- Mostly focused in southern California.
- Mostly ranging from $300K - $600K and $1M+
- Mostly ranging from 1000 - 3000 square feet.

California market performances in 2021 vs 2020

- In overall, the market hotness decreases, median listing price increases, median days of market decreases.
- Less active listings and more pending listings.

References

1. Realtor's zipcode hotness and inventory datasets https://www.realtor.com/research/data/

2.https://www.latimes.com/business/story/2021-09-20/southern-california-home-prices-slightly-cooling-august-2021