CALIFORNIA HOUSE VALUE FORECAST

TOM CHENG

DATAS SCIENCE CAREER TRACK CAPSTONE, MAY 20TH 2020 COHORT

WHAT'S FORECAST?

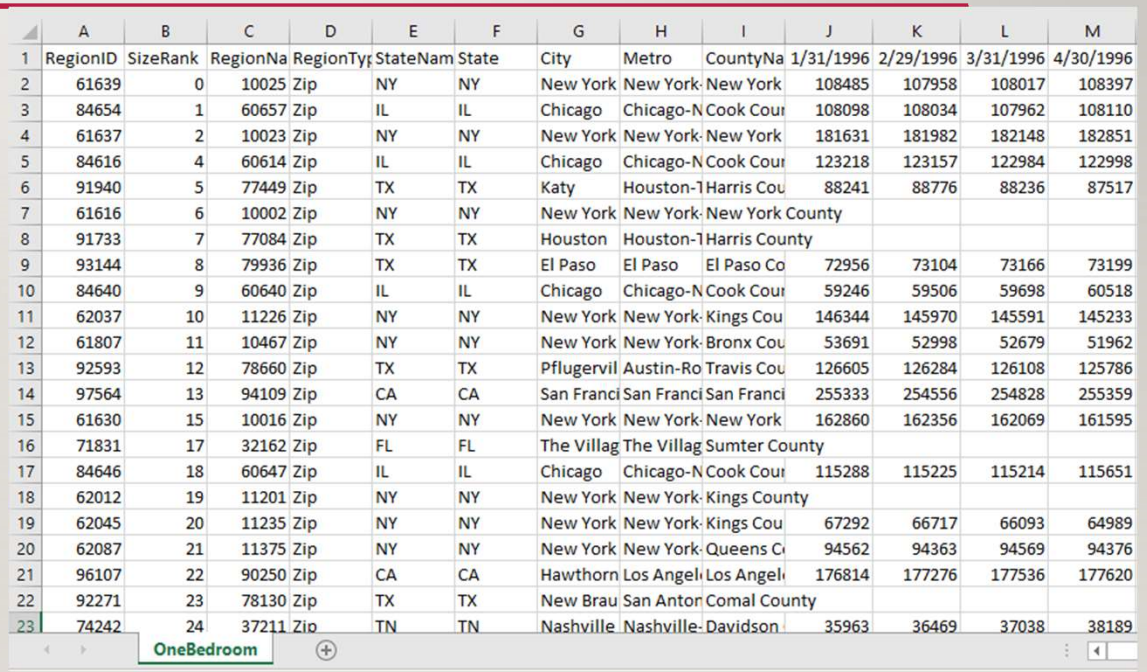
- Use **common sense and logics** to find **factors** that may influence sales, in real estates: **Time, Bedrooms, Bathrooms, and Zip-code of the estate are the main factors.**

WHY HOUSE VALUE FORECAST?

House Value Forecast helps both **real estate investors** and **home buyers** to determine if it's a good **time**, **place**, or **choice** for their investment.

WHERE MY DATA CAME FROM?

- All of my data came from **ZHVI,** which can be downloaded from https://www.zillow.com/research/data/
- **Zillow Home Value Index (ZHVI)**: A smoothed, seasonally adjusted measure of the typical home value and market changes across a given region and housing type. It reflects the typical value for homes in the 35th to 65th percentile range.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	RegionID	SizeRank	RegionNa	RegionTyp	StateNam	State	City	Metro	CountyNa	1/31/1996	2/29/1996	3/31/1996	4/30/1996
2	61639	0	10025	Zip	NY	NY	New York	New York-	New York	108485	107958	108017	108397
3	84654	1	60657	Zip	IL	IL	Chicago	Chicago-N	Cook Cour	108098	108034	107962	108110
4	61637	2	10023	Zip	NY	NY	New York	New York-	New York	181631	181982	182148	182851
5	84616	4	60614	Zip	IL	IL	Chicago	Chicago-N	Cook Cour	123218	123157	122984	122998
6	91940	5	77449	Zip	TX	TX	Katy	Houston-T	Harris Cou	88241	88776	88236	87517
7	61616	6	10002	Zip	NY	NY	New York	New York-	New York County				
8	91733	7	77084	Zip	TX	TX	Houston	Houston-T	Harris County				
9	93144	8	79936	Zip	TX	TX	El Paso	El Paso	El Paso Co	72956	73104	73166	73199
10	84640	9	60640	Zip	IL	IL	Chicago	Chicago-N	Cook Cour	59246	59506	59698	60518
11	62037	10	11226	Zip	NY	NY	New York	New York-	Kings Cou	146344	145970	145591	145233
12	61807	11	10467	Zip	NY	NY	New York	New York-	Bronx Cou	53691	52998	52679	51962
13	92593	12	78660	Zip	TX	TX	Pflugervil	Austin-Ro	Travis Cou	126605	126284	126108	125786
14	97564	13	94109	Zip	CA	CA	San Franci	San Franci	San Franci	255333	254556	254828	255359
15	61630	15	10016	Zip	NY	NY	New York	New York-	New York	162860	162356	162069	161595
16	71831	17	32162	Zip	FL	FL	The Villag	The Villag	Sumter County				
17	84646	18	60647	Zip	IL	IL	Chicago	Chicago-N	Cook Cour	115288	115225	115214	115651
18	62012	19	11201	Zip	NY	NY	New York	New York-	Kings County				
19	62045	20	11235	Zip	NY	NY	New York	New York-	Kings Cou	67292	66717	66093	64989
20	62087	21	11375	Zip	NY	NY	New York	New York-	Queens Co	94562	94363	94569	94376
21	96107	22	90250	Zip	CA	CA	Hawthorn	Los Angel	Los Angel	176814	177276	177536	177620
22	92271	23	78130	Zip	TX	TX	New Brau	San Anton	Comal County				
23	74242	24	37211	Zip	TN	TN	Nashville	Nashville-	Davidson	35963	36469	37038	38189

OneBedroom

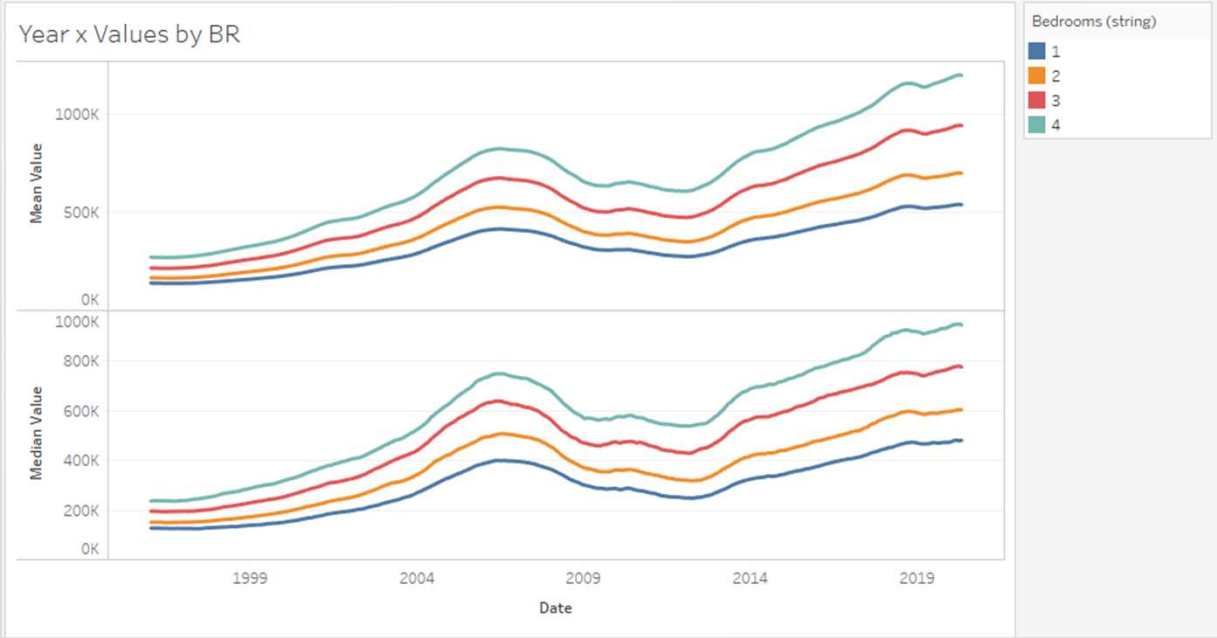
WHAT DID I DO TO THE DATA?

- All data must have estimates within the same time frame: Jan 1996 to Dec 2020
- California Zip code, and County
- Specified number of Bedrooms

Zipcode	County	Date	Value	Bedrooms
90004	Los Angeles County	1/31/1996	141542	1
90004	Los Angeles County	1/31/1996	175380	2
90004	Los Angeles County	1/31/1996	124140	3
90004	Los Angeles County	1/31/1996	137079	4
90007	Los Angeles County	1/31/1996	90619	1
90007	Los Angeles County	1/31/1996	121384	2
90007	Los Angeles County	1/31/1996	136689	3
90007	Los Angeles County	1/31/1996	146297	4
90012	Los Angeles County	1/31/1996	105719	1
90012	Los Angeles County	1/31/1996	123579	2
90012	Los Angeles County	1/31/1996	198392	3
90012	Los Angeles County	1/31/1996	196627	4
90016	Los Angeles County	1/31/1996	107440	1
90016	Los Angeles County	1/31/1996	144677	2
90016	Los Angeles County	1/31/1996	162227	3
90016	Los Angeles County	1/31/1996	185197	4

EXPLORATORY DATA ANALYSIS (EDA)
YEAR X AVG. VALUES COLORED BY NUMBER OF BEDROOMS.

- Same trend for all 8 lines
- More bedrooms means more value.
- Obvious Mean > Median; this means huge outliers.

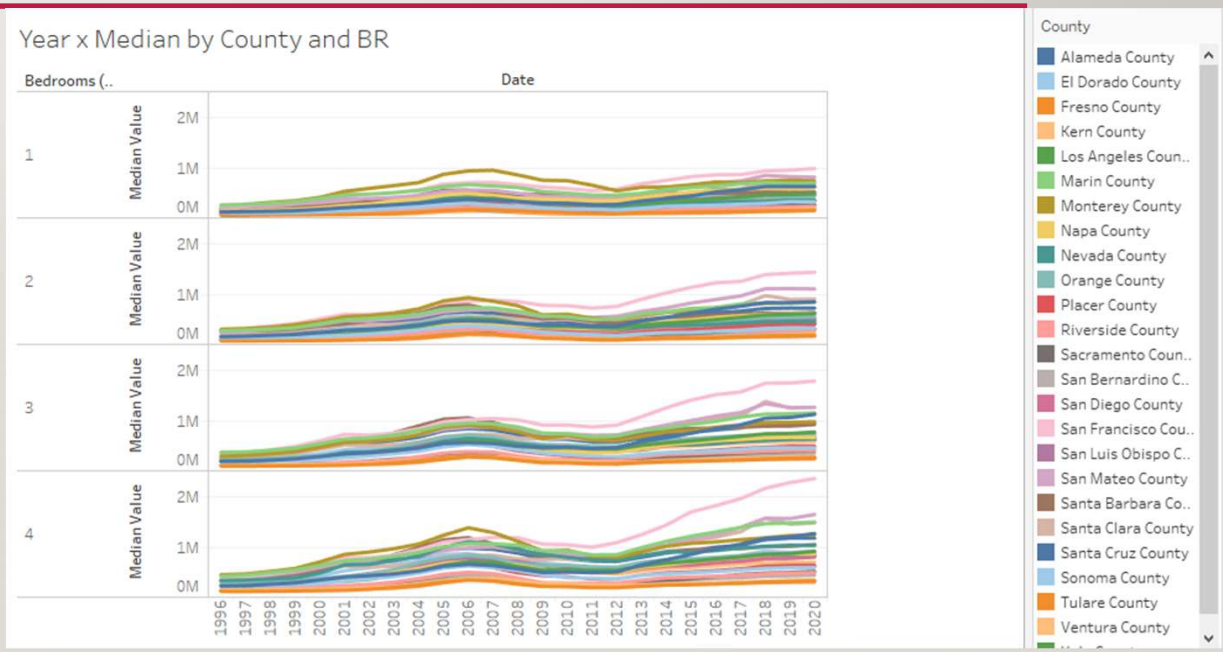


EDA CONT.
MONTH X VALUES BY MONTH

- Month of year doesn't seem to matter, so we can just focus on other factors.

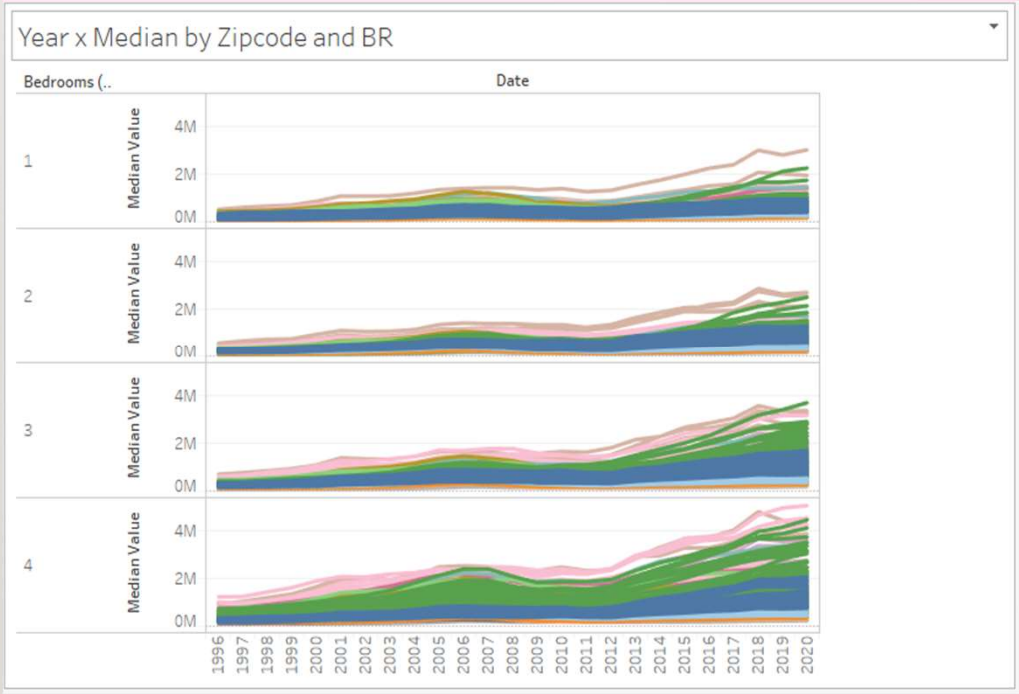
EDA CONT.
YEAR X MEDIAN BY COUNTY AND NUMBER OF BEDROOMS

- Much more variation than number of bedrooms.
- San Francisco seems to rise much more as the number of bedrooms increases more than other county.



EDA CONT.
YEAR X MEDIAN BY ZIP CODE AND NUMBER OF BEDROOMS

- This plot shows the outliers, so the forecast should definitely include the zip code of the property.



MACHINE LEARNING MODELING

- Type: Time Series Forecast
- Tools: pandas, numpy, matplotlib, seaborn, datetime
- Methods/Models:
  - Autoregressive Integrated Moving Average (**ARIMA**)
  - Vector Autoregressive Integrated Moving Average (**VARMA**)
  - Holt Winter's Exponential Smoothing (**HWES**)

MODELING STEPS

- Data Preprocessing: time series format, no missing values, difference the data then $1/5^{th}$ root the value to make it stationary.
- Split to train/test, 1996 Jan-2015 Dec for training and 2016 Jan-2020 Dec for testing.
- Use grid search to find best hyperparameters for each model
- Use MAE, grid search time, and fitting time to evaluate the model's performance

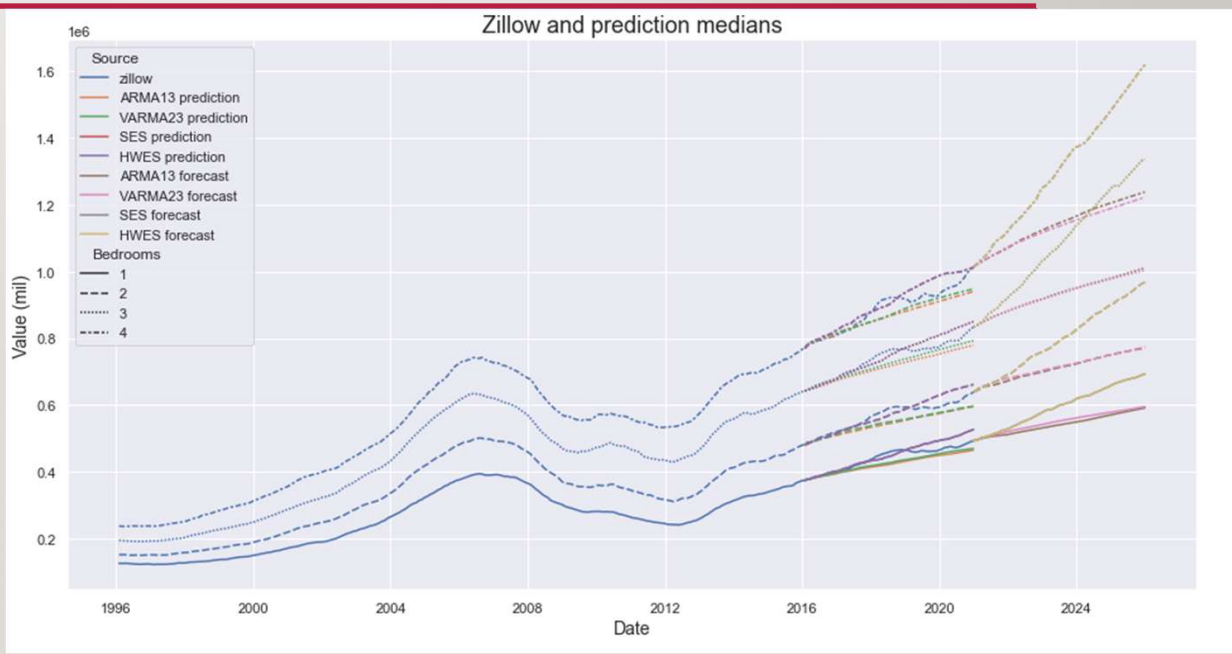
MODEL SCORES

- ARMA has the lowest (best) MAE score, but does have grid search time of 40 minutes
- VARMA's grid search was terminated after more than 5 hours. The tested p was 1,2, and q was 1~3. When p=2, q=3, the model had best MAE.
- HWES didn't need grid search, and was trained in less than a minute.

	Model	MAE	grid_search_time(m)	training_time(m)
0	ARMA13	479059.9372	40	8
1	VARMA23	479438.4738	300+	98
2	HWES	504900.4982	0	0

MODEL PRED/FC MEDIAN

- HWES's pred/test looked okay, but the forecast doesn't seem realistic. It's way too exponential, which usually isn't how housing prices will grow.
- ARMA and VARMA had very similar results, but due to time spent on modeling, ARMA is winner.



PRICE CATEGORY

Adding a pricing category

0= 0~499,999

1 = 500,000~999,999

2 = 1000,000~149,999

3 = 1,500,000~19,999,999

Etc. I find out that houses with 30+% increase in 5 years, usually

```
zillow30
```

	Zipcode	County	2015_value	Bedrooms	2020_value	Difference	Difference%	2015cat
457925	90004	Los Angeles County	658469	2	972189	313720	0.3200	1
457926	90004	Los Angeles County	714973	3	1392110	677137	0.4900	1
457927	90004	Los Angeles County	1125436	4	2324580	1199144	0.5200	2
457929	90007	Los Angeles County	416698	2	691590	274892	0.4000	0
457930	90007	Los Angeles County	474355	3	756250	281895	0.3700	0
...	...	...	...	...	...	...	...	...
459827	95822	Sacramento County	278477	4	420899	142422	0.3400	0
459828	95825	Sacramento County	89601	1	178832	89231	0.5000	0
459829	95825	Sacramento County	214142	2	316215	102073	0.3200	0
459830	95825	Sacramento County	287253	3	409144	121891	0.3000	0
459831	95825	Sacramento County	294887	4	450215	155328	0.3500	0

549 rows × 8 columns

```
zillow30['2015cat'].value_counts()
```

```
0    303
1    156
2     63
3     14
4     10
5      3
```

RECOMMENDATION

- Most of the properties were under one million
- So, 2 price categories
- Viewing the top 5 of each category by county

127 out of 137 are under 1 million!

```
arma0 = arma30[arma30['price_cat']==0]
arma1 = arma30[arma30['price_cat']==1]
```

```
arma0.County.value_counts().head(5)
```

```
8]: San Bernardino County    25
    Los Angeles County       16
    Riverside County         13
    Kern County               7
    San Diego County          7
    Name: County, dtype: int64
```

```
arma1.County.value_counts().head(5)
```

```
9]: San Diego County         14
    Los Angeles County       12
    San Bernardino County     4
    Riverside County          3
    Napa County               2
    Name: County, dtype: int64
```

THANK YOU!