A Project/Dissertation Review-ETE Report on

HOUSE PRICE ANALYSIS AND PREDICTIONS

Submitted in partial fulfilment of the requirement for the award of the degree of

BTech. In CSE with AIML

GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

Under The Supervision of

Mr Shubham Kumar
associate Professor

Submitted By
Ihechiluru Winner Chukwuemeka
19SCSE1180115

Blessing Joseph Akwesie
19SCSE1180116

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING GALGOTIAS UNIVERSITY, GREATER NOIDA, INDIA.

SEPTEMBER, 2021

Abstract-

Big Data technology is playing a very important role in education, but as we know big data has many advantages and also has some disadvantages. Educational resources in colleges and universities, building a complete educational big data analysis platform. This, the college and universities have to satisfy many conditions, as providing broad sets of different type tasks, including group discussions, oral speeches, essays with more than one possible correct opinion, developing complex skills of their students; collecting information about courses, student's activities and progress, alumni skills and online also provide online education. Big data are not the only to develop the quality of education. Many colleges and small universities provide private educational programs for small groups. Moreover, they over students more direct conversations with lecturers. This educational strategy definitely has its own advantages.

INTRODUCTION

Houses are one of human life's most essential needs, along with other fundamental needs such as food, water, and much more. Demand for houses grew rapidly over the years as people's living standards improved.

While there are people who make their house as an investment and property, yet most people around the world are buying a house as their shelter or as their livelihood.

What is House Price Prediction?

House price prediction is forecasting the future prices and trends of homes and houses by evaluating various factors like its characteristics, demand, seasonal trends, other influencing commodities prices (eg building materials, loans), offers from numerous suppliers etc.

Means of House Prediction

1. **Descriptive analytics:** Descriptive analytics rely on statistical methods that include data collection, analysis, interpretation, presentation of findings. It essentially answers the question of what happened?

   We used descriptive analytics to transform our raw observations into knowledge and insights that we'll be sharing during this project.

2. **Predictive analytics:** Predictive analytics is about analyzing current and

historical data to forecast the possibility of future events, outcomes or values.

Here, we’ll be using data mining (identification of patterns in data) and machine learning(linear regression) to build systems that can find and understand patterns in data, learn from them and also predict future outcomes.

Importance of House Price Prediction

Businesses and entrepreneurs can use information about future prices to find and properly define the best time to into a housing market, to adjust the prices of their houses(if for sale) and also know the optimal time to sell off a property

According to research, housing markets have a positive impact on a country's currency, which is an important national economy scale. Homeowners will purchase goods such as furniture and household equipment for their homes, and home builders or contractors will purchase raw materials to build houses to satisfy house demand, which is an indication of the economic wave effect created by the new house supply.

Besides that, consumers have the capital to make a large investment, and the construction industry is in good condition as can be seen through a country's high level of house supply.

Data Collection, Preparation and Preprocessing

To be able to analyze and properly predict the trends we had to use a lot of data. Now, for those of you who might be new to the world of Data Science, **data is information that has been translated into a form that is efficient for movement or processing.**

Data collection is the process of gathering and measuring information on targeted variables in an established system, which then enables one to answer relevant questions and evaluate outcomes. Most raw data (including ours) don’t usually come all sparkly and clean, you would have to clean and transform the data for analysis and processing, this process is known as **DATA CLEANING.**

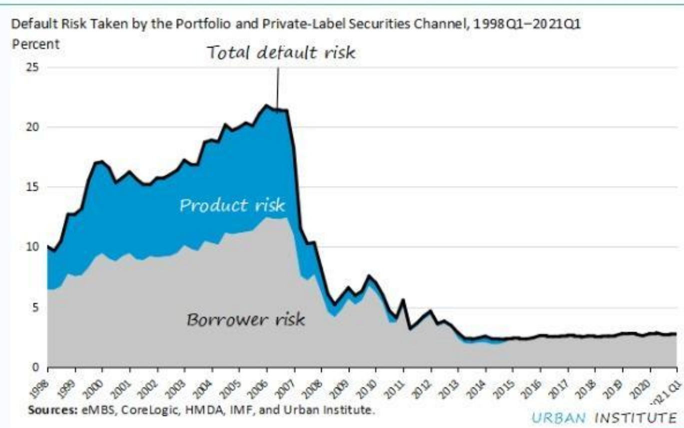
Data Cleaning:

As previously stated, the data we used for our analysis didn't come clean. We had to add, remove, transform and edit a lot of things to make the data fit for analysis and also to ensure that the results from our analysis are correct and accurate.

What's a better way to explain what we did with our data than to show what it looked like before as well as after we had it cleaned

1 PP market HCAI Updated to 2021Q1
2 See FAQ at http://www.urban.org/policy-centers/housing-finance-policy-center/projects/housing-credit-availability-index/faq
3

4	Year	Quarter	highlight	Total Risk	Borrower Risk	Product Risk
5	1998	1	.	10.024	6.438	3.586
6	1998	2	.	9.668	6.431	3.237
7	1998	3	.	10.491	6.725	3.765
8	1998	4	.	12.708	7.763	4.945
9	1999	1	.	12.709	7.560	5.149
10	1999	2	.	13.179	7.634	5.545
11	1999	3	.	15.512	8.263	7.249
12	1999	4	.	16.970	9.134	7.836
13	2000	1	.	17.087	9.477	7.610
14	2000	2	.	16.594	9.011	7.583
15	2000	3	.	15.357	8.792	6.565
16	2000	4	.	15.779	9.232	6.548
17	2001	1	17.773	16.253	9.476	6.777
18	2001	2	17.773	15.664	8.954	6.710
19	2001	3	17.773	15.213	8.892	6.321
20	2001	4	17.773	15.188	9.224	5.964
21	2002	1	17.773	15.739	9.129	6.610
22	2002	2	17.773	15.743	9.198	6.546
23	2002	3	17.773	16.048	9.246	6.801
24	2002	4	17.773	16.404	9.508	6.895
25	2003	1	17.773	17.231	10.153	7.078
26	2003	2	17.773	16.863	9.841	7.022
27	2003	3	17.773	16.945	9.643	7.301
28	2003	4	.	18.715	10.483	8.232
29	2004	1	.	18.896	10.328	8.567
30	2004	2	.	18.759	10.176	8.583
31	2004	3	.	20.181	11.185	8.996
32	2004	4	.	19.687	11.057	8.630
33	2005	1	.	19.950	11.161	8.789
34	2005	2	.	20.327	11.249	9.078
35	2005	3	.	20.090	11.351	8.740
36	2005	4	.	21.088	11.932	9.156
37	2006	1	.	21.774	12.484	9.290
38	2006	2	.	21.437	12.345	9.093
39	2006	3	.	21.362	12.318	9.044
40	2006	4	.	21.321	12.449	8.872
41	2007	1	.	18.249	10.972	7.278
42	2007	2	.	11.542	7.602	3.940
43	2007	3	.	10.264	7.202	3.062
44	2007	4	.	10.363	7.726	2.637
45	2008	1	.	8.207	6.217	1.991
46	2008	2	.	6.064	4.574	1.490
47	2008	3	.	5.199	4.139	1.060
48	2008	4	.	5.897	4.785	1.113



New Privately-Owned Houses Sold and For Sale

(Thousands of Units. Detail may not add to total because of rounding.)

Table 1a - Seasonally adjusted

Period	Sold during period[1]					For sale at end of period					Months' supply[2]	Median sales price ($)	Average sales price ($)
	United States	North-east	Mid-west	South	West	United States	North-east	Mid-west	South	West			
2020													
August	977	46	96	578	257	286	X	X	X	X	3.5	X	X
September	971	38	97	564	272	286	X	X	X	X	3.5	X	X
October	969	41	113	540	275	284	X	X	X	X	3.5	X	X
November	865	34	96	523	212	290	X	X	X	X	4.0	X	X
December	943	41	112	553	237	299	X	X	X	X	3.8	X	X
2021													
January	993	47	124	575	247	302	X	X	X	X	3.6	X	X
February	823	40	104	465	214	306	X	X	X	X	4.5	X	X
March	873	47	109	550	167	305	X	X	X	X	4.2	X	X
April	796	41	98	476	181	317	X	X	X	X	4.8	X	X
May (r)	733	40	92	412	189	331	X	X	X	X	5.4	X	X
June (r)	685	28	86	391	180	349	X	X	X	X	6.1	X	X
July (r)	729	23	74	420	212	366	X	X	X	X	6.0	X	X
August (p)	**740**	**29**	**51**	**445**	**215**	**378**	**X**	**X**	**X**	**X**	**6.1**	**X**	**X**
Average RSE (%)[3]	*10*	*23*	*24*	*13*	*14*	*5*	*X*	*X*	*X*	*X*	*10*	*X*	*X*
Percent Change[4]													
Aug. 2021 from Jul. 2021	***1.5%***	***26.1%***	***-31.1%***	***6.0%***	***1.4%***	***3.3%***	***X***	***X***	***X***	***X***	***1.7%***	***X***	***X***
90 percent confidence interval[5]	*± 15.1*	*± 73.4*	*± 23.5*	*± 19.2*	*± 32.0*	*± 2.0*	*X*	*X*	*X*	*X*	*± 18.5*	*X*	*X*
Aug. 2021 from Aug. 2020	***-24.3%***	***-37.0%***	***-46.9%***	***-23.0%***	***-16.3%***	***32.2%***	***X***	***X***	***X***	***X***	***74.3%***	***X***	***X***
90 percent confidence interval[5]	*± 19.1*	*± 23.2*	*± 21.7*	*± 27.0*	*± 32.4*	*± 10.1*	*X*	*X*	*X*	*X*	*± 53.9*	*X*	*X*

Table 1. Rental and Homeowner Vacancy Rates for the United States: 1965 to 2021
(in percent)

Year	Rental Vacancy Rates				Homeowner Vacancy Rates			
	First Quarter	Second Quarter	Third Quarter	Fourth Quarter	First Quarter	Second Quarter	Third Quarter	Fourth Quarter
2021	6.8	6.2			0.9	0.9		
2020	6.6	5.7	6.4	6.5	1.1	0.9	0.9	1.0
2019	7.0	6.8	6.8	6.4	1.4	1.3	1.4	1.4
2018	7.0	6.8	7.1	6.6	1.5	1.5	1.6	1.5
2017	7.0	7.3	7.5	6.9	1.7	1.5	1.6	1.6
2016	7.0	6.7	6.8	6.9	1.7	1.7	1.8	1.8
2015	7.1	6.8	7.3	7.0	1.9	1.8	1.9	1.9
2014	8.3	7.5	7.4	7.0	2.0	1.9	1.8	1.9
2013	8.6	8.2	8.3	8.2	2.1	1.9	1.9	2.1
2012	8.8	8.6	8.6	8.7	2.2	2.1	1.9	1.9
2011	9.7	9.2	9.8	9.4	2.6	2.5	2.4	2.3
2010	10.6	10.6	10.3	9.4	2.6	2.5	2.5	2.7
2009	10.1	10.6	11.1	10.7	2.7	2.5	2.6	2.7
2008	10.1	10.0	9.9	10.1	2.9	2.8	2.8	2.9
2007	10.1	9.5	9.8	9.6	2.8	2.6	2.7	2.8
2006	9.5	9.6	9.9	9.8	2.1	2.2	2.5	2.7
2005	10.1	9.8	9.9	9.6	1.8	1.8	1.9	2.0
2004	10.4	10.2	10.1	10.0	1.7	1.7	1.7	1.8
2003	9.4	9.6	9.9	10.2	1.7	1.7	1.9	1.8
2002[r1]	9.1	8.4	9.0	9.3	1.7	1.7	1.7	1.7
2002	9.1	8.5	9.1	9.4	1.7	1.7	1.7	1.7
2001	8.2	8.3	8.4	8.8	1.5	1.8	1.9	1.8
2000	7.9	8.0	8.2	7.8	1.6	1.5	1.6	1.6
1999	8.2	8.1	8.2	7.9	1.8	1.6	1.6	1.6
1998	7.7	8.0	8.2	7.8	1.7	1.7	1.7	1.8
1997	7.5	7.9	7.9	7.7	1.7	1.6	1.5	1.7
1996	7.9	7.8	8.0	7.7	1.6	1.5	1.7	1.7
1995	7.4	7.7	7.7	7.7	1.5	1.6	1.5	1.6
1994	7.5	7.4	7.2	7.4	1.4	1.4	1.4	1.6
1993[r2]	7.8	7.6	7.0	6.9	1.4	1.4	1.4	1.4
1993	7.9	7.6	7.1	6.9	1.4	1.4	1.4	1.4
1992	7.4	7.7	7.3	7.1	1.5	1.6	1.6	1.5
1991	7.5	7.3	7.6	7.3	1.7	1.8	1.8	1.6
1990	7.5	7.0	7.2	7.2	1.7	1.7	1.7	1.7

Cleaned Data

	C	D	E
1	k	_risk	risk
2	11.236	8.836	2.400
3	11.790	9.040	2.750
4	13.209	9.471	3.737
5	14.332	10.160	4.172
6	14.764	10.372	4.393
7	14.154	9.896	4.257
8	13.524	9.870	3.654
9	13.412	9.898	3.514
10	12.966	9.530	3.436
11	12.511	9.168	3.343
12	12.272	9.086	3.186
13	12.120	9.076	3.044
14	12.211	9.147	3.064
15	12.334	9.295	3.039
16	12.163	9.006	3.157
17	12.014	8.896	3.118
18	12.258	9.041	3.217
19	12.204	8.924	3.280
20	12.131	8.732	3.399
21	13.444	9.387	4.058
22	14.270	9.588	4.682
23	14.390	9.457	4.933
24	15.572	10.042	5.530
25	15.516	9.980	5.536
26	15.672	9.977	5.694
27	15.849	9.861	5.988
28	15.563	9.749	5.814
29	16.370	10.212	6.158

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	sqft_above	sqft_basem	yr_built	yr_renovate	street	city	statezip	country	
2	02-05-2014 00:00	313000	3	1.5	1340	7912	1.5	0	0	3	1340	0	1955	2005	18810 Dens	Shoreline	WA 98133	USA	
3	02-05-2014 00:00	2384000	5	2.5	3650	9050	2	0	4	5	3370	280	1921	0	709 W Blain	Seattle	WA 98119	USA	
4	02-05-2014 00:00	342000	3	2	1930	11947	1	0	0	4	1930	0	1966	0	26206-2621	Kent	WA 98042	USA	
5	02-05-2014 00:00	420000	3	2.25	2000	8030	1	0	0	4	1000	1000	1963	0	857 170th P	Bellevue	WA 98008	USA	
6	02-05-2014 00:00	550000	4	2.5	1940	10500	1	0	0	4	1140	800	1976	1992	9105 170th	Redmond	WA 98052	USA	
7	02-05-2014 00:00	490000	2	1	880	6380	1	0	0	3	880	0	1938	1994	522 NE 88th	Seattle	WA 98115	USA	
8	02-05-2014 00:00	335000	2	2	1350	2560	1	0	0	3	1350	0	1976	0	2616 174th	Redmond	WA 98052	USA	
9	02-05-2014 00:00	482000	4	2.5	2710	35868	2	0	0	3	2710	0	1989	0	23762 SE 25	Maple Valle	WA 98038	USA	
10	02-05-2014 00:00	452500	3	2.5	2430	88426	1	0	0	4	1570	860	1985	0	46611-4662	North Bend	WA 98045	USA	
11	02-05-2014 00:00	640000	4	2	1520	6200	1.5	0	0	3	1520	0	1945	2010	6811 55th A	Seattle	WA 98115	USA	
12	02-05-2014 00:00	463000	3	1.75	1710	7320	1	0	0	3	1710	0	1948	1994	Burke-Gilma	Lake Forest	WA 98155	USA	
13	02-05-2014 00:00	1400000	4	2.5	2920	4000	1.5	0	0	5	1910	1010	1909	1988	3838-4098	Seattle	WA 98105	USA	
14	02-05-2014 00:00	588500	3	1.75	2330	14892	1	0	0	3	1970	360	1980	0	1833 220th	Sammamish	WA 98074	USA	
15	02-05-2014 00:00	365000	3	1	1090	6435	1	0	0	4	1090	0	1955	2009	2504 SW Po	Seattle	WA 98106	USA	
16	02-05-2014 00:00	1200000	5	2.75	2910	9480	1.5	0	0	3	2910	0	1939	1969	3534 46th A	Seattle	WA 98105	USA	
17	02-05-2014 00:00	242500	3	1.5	1200	9720	1	0	0	4	1200	0	1965	0	14034 SE 20	Kent	WA 98042	USA	
18	02-05-2014 00:00	419000	3	1.5	1570	6700	1	0	0	4	1570	0	1956	0	15424 SE 9t	Bellevue	WA 98007	USA	
19	02-05-2014 00:00	367500	4	3	3110	7231	2	0	0	3	3110	0	1997	0	11224 SE 30	Auburn	WA 98092	USA	
20	02-05-2014 00:00	257950	3	1.75	1370	5858	1	0	0	3	1370	0	1987	2000	1605 S 245t	Des Moines	WA 98198	USA	
21	02-05-2014 00:00	275000	3	1.5	1180	10277	1	0	0	3	1180	0	1983	2009	12425 415th	North Bend	WA 98045	USA	
22	02-05-2014 00:00	750000	3	1.75	2240	10578	2	0	0	5	1550	690	1923	0	3225 NE 92	Seattle	WA 98115	USA	
23	02-05-2014 00:00	435000	4	1	1450	8800	1	0	0	4	1450	0	1954	1979	3922 154th	Bellevue	WA 98006	USA	
24	02-05-2014 00:00	626000	3	2.25	1750	1572	2.5	0	0	3	1470	280	2005	0	3140 Frankl	Seattle	WA 98102	USA	
25	02-05-2014 00:00	612500	4	2.5	2730	12261	2	0	0	3	2730	0	1991	0	10212 NE 1	Bothell	WA 98011	USA	
26	02-05-2014 00:00	495000	4	1.75	1600	6380	1	0	0	3	1130	470	1959	1989	2021 NE 10	Seattle	WA 98125	USA	
27	02-05-2014 00:00	285000	3	2.5	2090	10834	1	0	0	4	1360	730	1987	0	27736 23rd	Federal Wa	WA 98003	USA	
28	02-05-2014 00:00	615000	3	1.75	2360	7291	1	0	0	4	1360	1000	1948	0	8436-8438	Seattle	WA 98136	USA	
29	02-05-2014 00:00	698000	4	2.25	2200	11250	1.5	0	0	5	1300	900	1920	0	1036 4th St	Kirkland	WA 98033	USA	
30	02-05-2014 00:00	675000	5	2.5	2820	67518	2	0	0	3	2820	0	1979	2014	23525 SE 32	Issaquah	WA 98029	USA	

1	DATE	home_ownership	PERMIT	supply_ratio	new_houses	placements_US	AvgPrice_US	house_price_index	new_sales
2	1999-01-01	66.7	1732.0	3.9	875.0	357	40500	93.212	591.0
3	1999-04-01	66.7	1720.0	4.0	848.0	348	39700	93.675	464.0
4	1999-07-01	66.8	1665.0	4.1	863.0	351	40100	94.221	461.0
5	1999-10-01	66.9	1600.0	3.9	918.0	376	41300	94.789	605.0
6	2000-01-01	67.1	1640.0	4.0	888.0	354	41200	95.3489999999999	586.0
7	2000-04-01	67.3	1702.0	3.9	923.0	378	41600	95.979	526.0
8	2000-07-01	67.5	1682.0	4.0	900.0	388	42700	96.596	665.0
9	2000-10-01	67.5	1671.0	4.0	893.0	385	41900	97.2239999999999	570.0
10	2001-01-01	67.6	1551.0	4.5	826.0	395	41600	97.869	590.0
11	2001-04-01	67.8	1649.0	4.2	872.0	361	43400	98.529	567.0
12	2001-07-01	67.9	1672.0	4.3	863.0	385	42200	99.16	579.0
13	2001-10-01	67.9	1683.0	4.3	873.0	410	42400	99.8489999999999	514.0
14	2002-01-01	67.9	1727.0	4.3	873.0	378	42900	100.552	549.0
15	2002-04-01	67.8	1692.0	4.3	856.0	409	42700	101.339	609.0
16	2002-07-01	67.9	1651.0	4.3	900.0	388	41600	102.127	562.0
17	2002-10-01	68.2	1597.0	4.4	841.0	370	42000	102.922	559.0
18	2003-01-01	68.1	1543.0	4.4	857.0	333	42100	103.677	523.0
19	2003-04-01	68.2	1572.0	4.8	793.0	320	43300	104.423999999999	580.0
20	2003-07-01	68.3	1542.0	4.1	887.0	304	43100	105.054	575.0
21	2003-10-01	68.5	1552.0	4.4	848.0	317	43100	105.767999999999	582.0
22	2004-01-01	68.7	1570.0	4.0	912.0	322	43800	106.537	590.0

To explain the entire process:

Firstly, we formatted the dates on each dataset, taking only data between 1999 and now. Some of these dates were mixed up so we had to sort the data in ascending order (This part was done using MS Excel).

Then we stripped each data set of variables(columns) we didn't need. Leaving only metrics that were necessary for our analysis (This part was done using both MS Excel and Python (Drop Function)).

Thereafter, we merged all the data sets from all the data sources into one dataset (this was done using python (merge and concat functions) and also Excel)

After that, We removed all the null values or rows with a lot of empty values. We did this using the dropna function in python.

We also labelled and re-labelled some of the columns to better explain their functionalities.

Data Preparation For ML model

-	bedrooms	bathroom	sqft_living	sqft_lot	floors	waterfront	view	condition	sqft_above	sqft_baser	yr_built	yr_renova	city	price
2014-05-0	3	1.5	1340	7912	1.5	0	0	3	1340	0	1955	2005	Shoreline	313000
2014-05-0	5	2.5	3650	9050	2	0	4	5	3370	280	1921	0	Seattle	2384000
2014-05-0	3	2	1930	11947	1	0	0	4	1930	0	1966	0	Kent	342000
2014-05-0	3	2.25	2000	8030	1	0	0	4	1000	1000	1963	0	Bellevue	420000
2014-05-0	4	2.5	1940	10500	1	0	0	4	1140	800	1976	1992	Redmond	550000
2014-05-0	2	1	880	6380	1	0	0	3	880	0	1938	1994	Seattle	490000
2014-05-0	2	2	1350	2560	1	0	0	3	1350	0	1976	0	Redmond	335000
2014-05-0	4	2.5	2710	35868	2	0	0	3	2710	0	1989	0	Maple Vall	482000
2014-05-0	3	2.5	2430	88426	1	0	0	4	1570	860	1985	0	North Ben	452500
2014-05-0	4	2	1520	6200	1.5	0	0	3	1520	0	1945	2010	Seattle	640000
2014-05-0	3	1.75	1710	7320	1	0	0	3	1710	0	1948	1994	Lake Fores	463000
2014-05-0	4	2.5	2920	4000	1.5	0	0	5	1910	1010	1909	1988	Seattle	1400000
2014-05-0	3	1.75	2330	14892	1	0	0	3	1970	360	1980	0	Sammami	588500
2014-05-0	3	1	1090	6435	1	0	0	4	1090	0	1955	2009	Seattle	365000
2014-05-0	5	2.75	2910	9480	1.5	0	0	3	2910	0	1939	1969	Seattle	1200000
2014-05-0	3	1.5	1200	9720	1	0	0	4	1200	0	1965	0	Kent	242500
2014-05-0	3	1.5	1570	6700	1	0	0	4	1570	0	1956	0	Bellevue	419000
2014-05-0	4	3	3110	7231	2	0	0	3	3110	0	1997	0	Auburn	367500
2014-05-0	3	1.75	1370	5858	1	0	0	3	1370	0	1987	2000	Des Moine	257950
2014-05-0	3	1.5	1180	10277	1	0	0	3	1180	0	1983	2009	North Ben	275000
2014-05-0	3	1.75	2240	10578	2	0	0	5	1550	690	1923	0	Seattle	750000
2014-05-0	4	1	1450	8800	1	0	0	4	1450	0	1954	1979	Bellevue	435000
2014-05-0	3	2.25	1750	1572	2.5	0	0	3	1470	280	2005	0	Seattle	626000

H1 house_price_index

DATE	home_ow	PERMIT	supply_rat	new_hous	placement	new_sales	house_pri	AvgPrice_US
1999-01-0	66.7	1732	3.9	875	357	591	93.212	40500
1999-04-0	66.7	1720	4	848	348	464	93.675	39700
1999-07-0	66.8	1665	4.1	863	351	461	94.221	40100
1999-10-0	66.9	1600	3.9	918	376	605	94.789	41300
2000-01-0	67.1	1640	4	888	354	586	95.349	41200
2000-04-0	67.3	1702	3.9	923	378	526	95.979	41600
2000-07-0	67.5	1682	4	900	388	665	96.596	42700
2000-10-0	67.5	1671	4	893	385	570	97.224	41900
2001-01-0	67.6	1551	4.5	826	395	590	97.869	41600
2001-04-0	67.8	1649	4.2	872	361	567	98.529	43400
2001-07-0	67.9	1672	4.3	863	385	579	99.16	42200
2001-10-0	67.9	1683	4.3	873	410	514	99.849	42400
2002-01-0	67.9	1727	4.3	873	378	549	100.552	42900
2002-04-0	67.8	1692	4.3	856	409	609	101.339	42700
2002-07-0	67.9	1651	4.3	900	388	562	102.127	41600
2002-10-0	68.2	1597	4.4	841	370	559	102.922	42000
2003-01-0	68.1	1543	4.4	857	333	523	103.677	42100
2003-04-0	68.2	1572	4.8	793	320	580	104.424	43300
2003-07-0	68.3	1542	4.1	887	304	575	105.054	43100
2003-10-0	68.5	1552	4.4	848	317	582	105.768	43100
2004-01-0	68.7	1570	4	912	322	590	106.537	43800
2004-04-0	69.4	1577	4	933	312	583	107.381	45000
2004-07-0	68.9	1614	4.2	880	320	548	108.3	45300

First, we removed the features (columns) that weren't important for our model and then converted the columns with string type values to int as the model doesn't read string type entries.

Here's what our data looks like after :

45]

	DATE	home_ownership	PERMIT	supply_ratio	new_houses	placements_US	new_sales	house_price_index	AvgPrice_US
0	0	66.7	1732	3.9	875	357	591	93.212	40500
1	1	66.7	1720	4.0	848	348	464	93.675	39700
2	2	66.8	1665	4.1	863	351	461	94.221	40100
3	3	66.9	1600	3.9	918	376	605	94.789	41300
4	4	67.1	1640	4.0	888	354	586	95.349	41200
...	...	...	...	...	...	...	...	...	...
85	85	68.1	2141	6.1	1061	130	373	183.288	61200
86	86	67.3	2118	5.9	1116	119	389	184.364	63600
87	87	65.7	1998	6.3	1123	117	445	184.330	59500
88	88	65.6	1905	6.2	1086	124	466	184.157	62100
89	89	65.6	1867	6.3	1074	125	485	183.509	63100

90 rows × 9 columns

To explain the entire process:

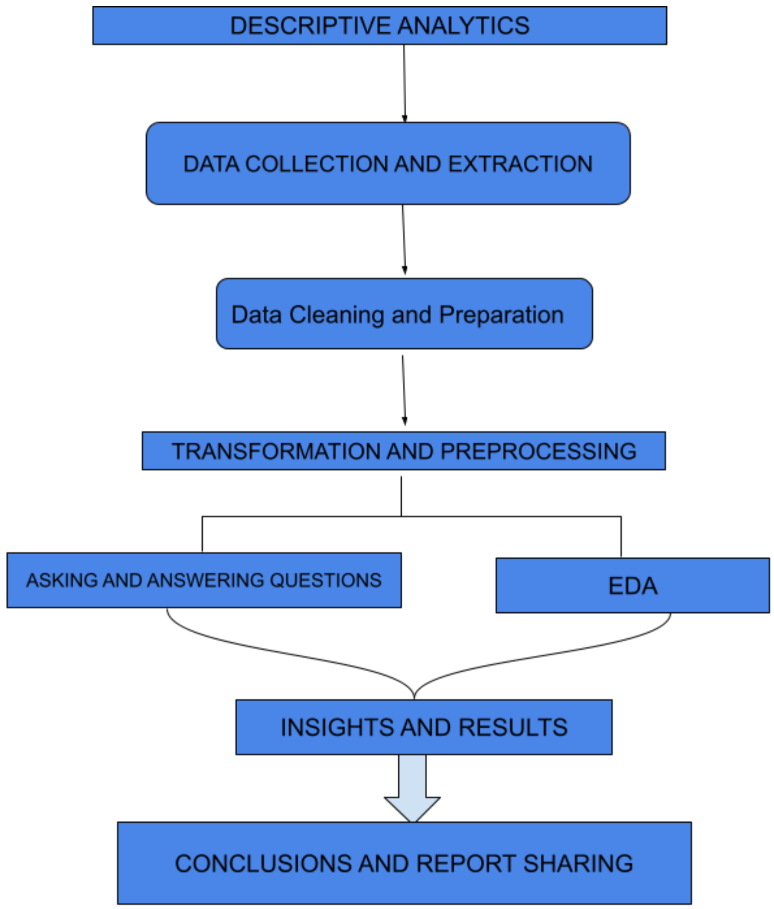
Firstly, we formatted the dates on each dataset, taking only data between 1999 and now. Some of these dates were mixed up so we had to sort the data in ascending order (This part was done using MS Excel).

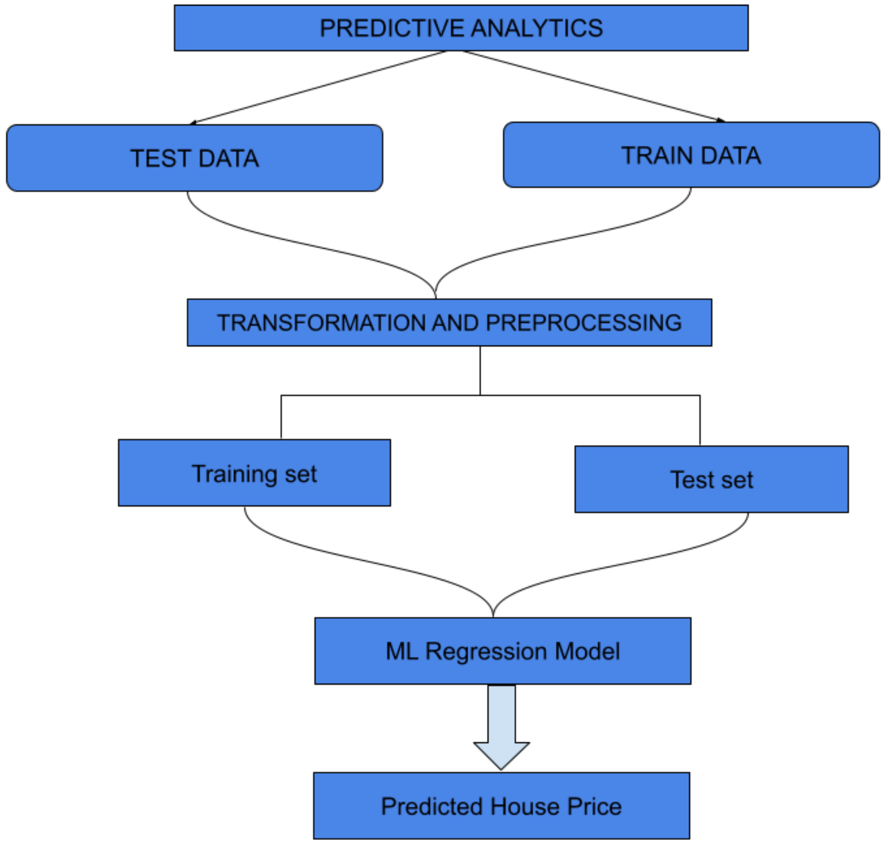
Then we stripped each data set of variables(columns) we didn't need. Leaving only metrics that were necessary for our analysis (This part was done using both MS Excel and Python (Drop Function)).

Thereafter, we merged all the data sets from all the data sources into one dataset (this was done using python (merge and concat functions) and also Excel)

After that, We removed all the null values or rows with a lot of empty values. We did this using the dropna function in python.

Architecture Diagram





A. **Descriptive Analytics:** We've gathered and collected data that was useful for our research. Then cleaned the data to ensure it's ready for processing and would give accurate results. Thereafter transformed and analysed the data by asking different questions and trying to seek insights using our data. Finally, concluded our findings and then organized them in the right report format for sharing and presentation.
B. **Predictive Analytics:** First divide our data set into Training data and Test data. Train data to get our ML model to learn the concepts and Test Data to ensure our model understands the concepts. Then normalized, preprocessed and transformed it to make it ready for our model. Finally developed a Regression Algorithm to give an estimated house price based on the test data.

Literature Survey

Using K-Nearest Neighbours for Stock Price Prediction:

In most communities in India, there has been a study where they used the k-nearest neighbours (KNN) algorithm to predict stock prices. In this study, they expressed the stock prediction problem as a similarity-based classification, and they represented the historical stock data as well as test data by vectors.

The authors listed the steps of predicting the closing price of the stock market using KNN as follows:

The number of nearest neighbours is chosen

The distance between the new record and the training data is computed

Training data is sorted according to the calculated distance

Majority voting is applied to the classes of the k nearest neighbours to determine the predicted value of the new record.

Stock Market Prediction Using Bayesian-Regularized Neural Networks:

In a study done by Ticknor (2013), he used a Bayesian regularized artificial neural network to predict the future operation of the financial market. Specifically, he built a model to predict future stock prices. The input of the model is previous stock statistics in addition to some financial technical data. The output of the model is the next-day closing price of the corresponding stocks.

The model proposed in the study is built using a Bayesian regularized neural network. The weights of this type of network are given a probabilistic nature. This allows the network to penalize very complex models (with many hidden layers) in an automatic manner. This in turn will reduce the overfitting of the model.

The model consists of a feedforward neural network that has three layers: an input layer, one hidden layer, and an output layer. The author chose the number of neurons in the hidden layer based on experimental methods. The input data of the model is normalized to be between -1 and 1, and this operation is reversed for the output so the predicted price appears in the appropriate scale Correlation Between Variables:

We want to see how the dataset variables are correlated with each other and how predictor variables are correlated with the target variable. For example, we would like to see how Lot Area and SalePrice are correlated: Do they increase and decrease together (positive correlation)? Does one of them increase when the other decrease or vice versa (negative correlation)? Or are they not correlated?

Correlation is represented as a value between -1 and +1 where +1 denotes the highest positive correlation, -1 denotes the highest negative correlation, and 0 denotes that there is no correlation.

Existing Problems:

1. The process of buying and selling homes and houses are usually stressful.
2. The risk factors and problems faced in real estate investments and in-home purchases are really high.
3. There are usually uncertainties as to when is the optimal time to either buy or sell properties at what price?
4. These problems faced by suppliers and consumers of housing properties sometimes result in real estate agents being trusted with the communication

between buyers and sellers and also all the legal documentation and transfer which in turn means extra charges and fees.

Tools and Technologies Used:

1. Python: Python is a general-purpose language, meaning it can be used to create a variety of different programs and isn't specialized for any specific problems. Python has become a staple in data science, allowing data analysts and other professionals to use the language to conduct complex statistical calculations, create data visualizations, build machine learning algorithms, manipulate and analyze data, and complete other data-related tasks. Python also has a number of libraries that enable coders to write programs for data analysis and machine learning more quickly and efficiently, like TensorFlow and Keras.

2. Excel: Microsoft Excel is a spreadsheet program. That means it's used to create grids of text, numbers and formulas specifying calculations. In our case, we used it for editing, cleaning, labelling and formatting our data (See Data Cleaning section above).

3. Tableau: Business intelligence and analytics use Tableau as a visualized platform for the intentions of helping people watch, observe, understand, and make decisions with a variety of data. Any type of graphs, plots, and charts can be made easily in it without the need for any programming.

Proposed Solution:

1. **Income:** The average income of people living in a particular area is one of the most defining factors of the prices of homes in that area. Houses around major cities and states where people tend to earn more are usually also costlier than houses in rural areas.

2. Location:

Residential houses are usually more costly where there are

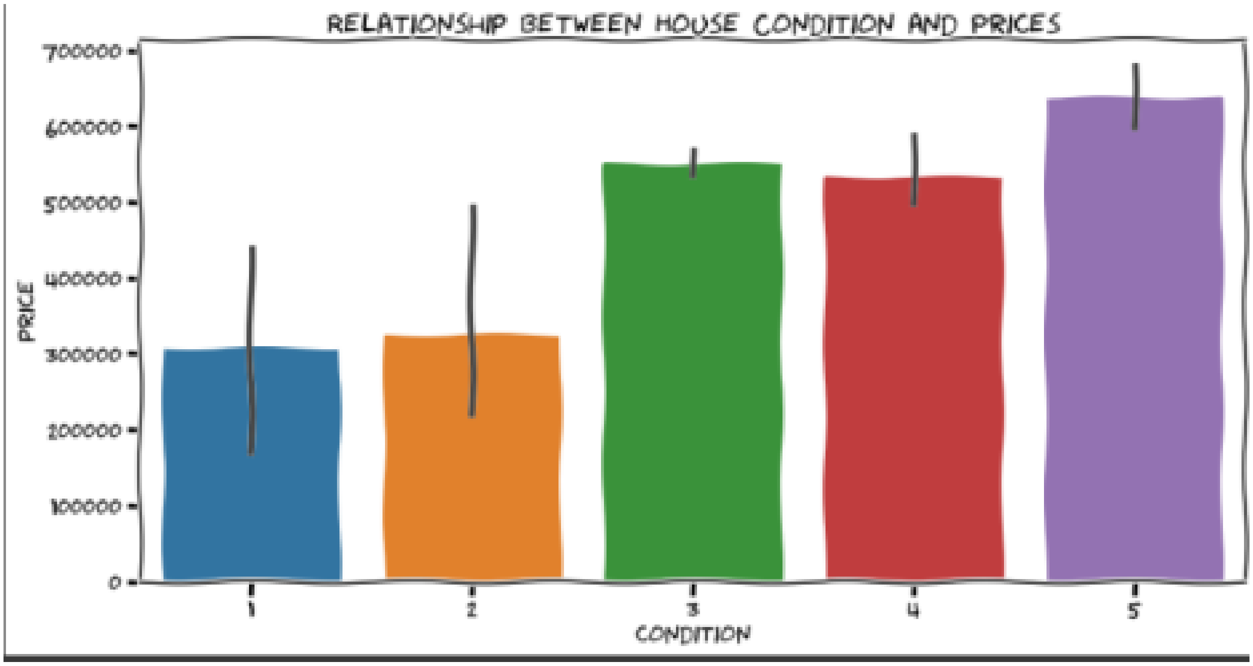
a. Better Employment opportunities.

b. More people (High population)

c. Close Malls, Entertainment centres, schools etc.

3. The general condition of the house:

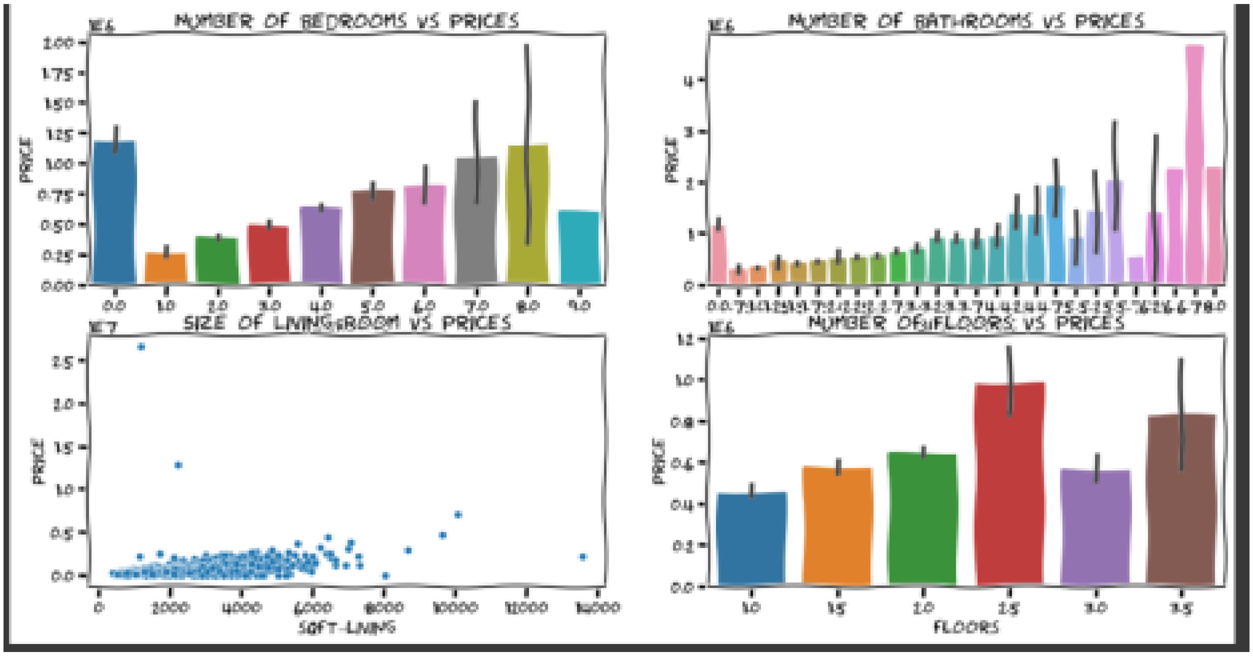
People tend to be willing to pay more for houses that are in better condition. So, little factors like the state of the bulbs, fans, toilets, kitchen and all that can go a long way in determining the value of your house.



4. Number and Sizes of the rooms:

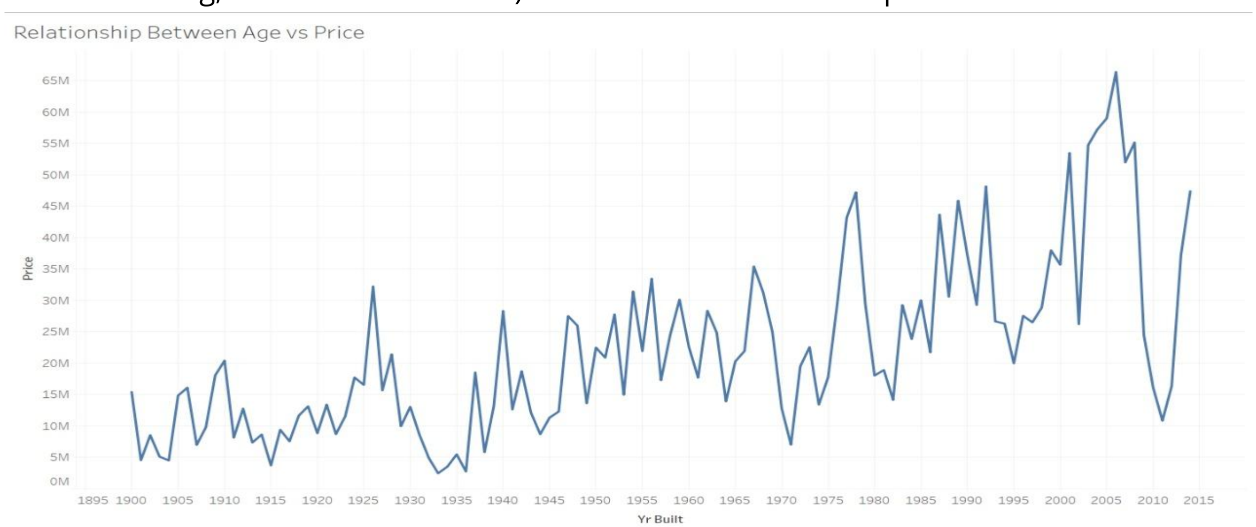
a. 1 bedroom Flats are on average, relatively the costliest flats. After that prices of residential apartments rise linearly from two to eight with its peak at eighth then falls back from 9 onwards.

b. The price of a house tends to increase with the number of bathrooms available. The sweet point is somewhere around 7 bathrooms



5. **Age of the Houses:**
   Based on our analysis, generally newer houses tend to have more value in terms of prices than older homes but then as well you have to understand that it's not exactly a one-way thing. In the sense that, the age of a house can also positively influence its value that's if it has historic values or sentiments attached to it.

   Then again, upgrades are necessary if you're going to be purchasing an older apartment (adding newer utensils and pieces of equipment like baths, air-conditioning, heaters and the likes) and this leads to the next point.



6. **Renovation:**
   When buying a house, the last thing anybody wants is to purchase a property where he'll have to spend tons of money on getting it to shape. Also, our research shows clearly that the most priced houses are the ones that were just recently innovated. So, getting in little inexpensive improvements and touch ups could improve your property resale value and ultimately makes it more attractive to buyers
   Things like:
   1. Painting your property
   2. Changing the faulty windows and doors
   3. Fixing all the faulty electronics and toiletries
   4. Renovating the kitchen

   Would go a long way in improving the market value of your property.

PRICE PREDICTION USING LINEAR REGRESSION

Linear regression is a regression model that estimates the relationship between one independent variable and one dependent variable using a straight line.

Our Machine learning model basically uses a multiple linear regression algorithm to predict the prices and house price indexes based on some input features.

Our model was fitted and tested using python sci-kit learn module and has a test accuracy score of 100%.

We are working on getting more data with better and stronger features to make our model even better and more versatile.

```
y_test
```

```
2      94.221
13    101.339
53    132.649
41    121.812
66    151.336
30    113.491
45    125.733
43    123.830
78    172.859
89    183.509
7      97.224
26    111.108
33    115.308
63    146.592
8      97.869
16    103.677
24    109.846
56    136.294
Name: house_price_index, dtype: float64
```

```
[ ] from sklearn.metrics import accuracy_score, confusion_matrix
```

```
[ ] print('Training Set accuracy score: {:.2f}'.format(lin.score(x_train, y_train)))
    print('Test set accuracy: {:.2f}'. format(lin.score(x_test, y_test)))
```

```
Training Set accuracy score: 0.99
Test set accuracy: 1.00
```

Expected Results and Output

1. House prices are relatively higher in areas with a high income earning population (eg big cities, urban areas).
2. The location where a house is situated plays a large role as to what the price of that house would be in the next few years.
3. Prices of homes tend to increase with the number of rooms available in the house. So for example purchasing a 5 bedroom duplex will probably pay you more in the long run than if you bought a 1- room apartment.
4. People are usually willing to pay more for a home if the house is in very good condition. So, it'll be a good idea to touch up or possibly renovate your house if you plan on selling or renting it out.

COSEDS 2022 Submission 2

Update informatic
Update authors
Update file

The submission has been saved!

Submission 2	
Title	House Price Analysis and Prediction
Paper:	(Dec 20, 19:28 GMT)
Author keywords	Big Data Data Science Business Analytics Machine Learning Data Analytics
Topics	Artificial Intelligence, Big Data, Business Intelligence, Data Analytics, Data Sciences, Machine Learning
Abstract	Big Data technology is playing a very important role in education, but as we know big data has many advantages and also has some disadvantages. Educational resources in colleges and universities, building a complete educational big data analysis platform. This, the college and universities have to satisfy many conditions, as providing broad sets of different type tasks, including group discussions, oral speeches, essays with more than one possible correct opinion, developing complex skills of their students; collecting information about courses, student's activities and progress, alumni skills and online also provide online education. Big data are not the only to develop the quality of education. Many colleges and small universities provide private educational programs for small groups. Moreover, they over students more direct conversations with lecturers. This educational strategy definitely has its own advantages.
Submitted	Dec 20, 19:28 GMT
Last update	Dec 20, 19:28 GMT

Authors

first name	last name	email	country	affiliation	Web page	corresponding?
Ihechiluru	Winner	ihechiluruwinner@gmail.com	India	Galgotias University		✓
Akwesie	Blessing	blessingakwesie2018@gmail.com	India	Galgotias University		✓