TEAM INFORMATION (1 point)

Team #: 68

Team Members:

1. Nihar Reddy Moramganti, GTID: 903848828
    a. Nihar Moramganti is an aspiring data analyst who has a passion for statistics, data visualization, and analytics. He earned his bachelor's degree in Business Analytics at San Jose State University. He has two years of experience in Compensation Analytics where he worked as a Compensation Analyst at Palo Alto Networks. He is actively seeking job opportunities as a role of a data analyst where he will be able to utilize his skills and contribute to a wide range of analytical projects.
2. Yuchen Hu, GTID: 903844369
    a. Yuchen Hu is a business intelligence engineer currently working in Prime Video. He earned his bachelor's degree in Mechanical Engineering from the University of Adelaide. He had several years of experience in analytics across several industries. Currently is actively involved in an analytics project focused on churn prevention.
3. Michael Cardenas, GTID: 903835761
    a. Michael currently works as a data analyst for a utility company. He earned his bachelor's degree in Economics and Accounting at the University of California, Santa Barbara. He possesses over 7 years of experience in analytics in various capacities ranging from financial analysis to operational analysis.
4. Linh Nguyen Dang , GTID: 903757931
    a. Linh works as a full-time Data Analytics Developer for a healthcare company that performs patient risk stratification of chronic kidney disease. He earned his bachelor and master degrees in Healthcare Management and Business, respectively. Previous analytics projects include operational reporting, developing end to end data pipelines, and large-scale patient data analysis.

OBJECTIVE/PROBLEM (5 points)

Project Title: *Analyzing the Impact of Macroeconomic Factors on Real Estate: A Study of CPI, Interest Rates, and Home Inventory*

Background Information on chosen project topic:

Being a first-time home buyer and completing that purchase is one of the best moments in a person's life. For many people, especially recent graduates and newly married couples, looking into making their first ever housing purchase, they often struggle deciding when is the right time to buy as the market might be too expensive at that given moment. There are many who work for years to save up enough money to buy their first home, but ultimately, the market dictates when most have an opportunity to buy. In addition, during the COVID-19 pandemic, housing prices went up exponentially, and many people who wanted to buy a house had to defer, not knowing how long unfavorable market conditions would persist.

Problem Statement (clear and concise statement explaining purpose of your analysis and investigation):

We want to find which variables have the most influence on housing prices so we can better predict housing prices. One of the reasons we want to look into this is because there are many drivers that affect housing market prices. With housing prices being at their highest during the COVID-19 pandemic, we want to explore how these drivers influenced housing prices during that time, and also how these factors played pre-pandemic as well. Based on housing data from 2016-2022, we want to build a model that would be the best fit for determining the most influential variables that will predict future housing prices. These factors will also play a role in investment opportunities for people to invest in property based on the housing market trends.

State your Primary Research Question (RQ): *Which market variables have the most influence in determining housing prices?*

Add some possible Supporting Research Questions (2-4 RQs that support problem statement):

1. *How can home buyers utilize these variables to make necessary investments into other housing properties?*
2. *Which laws can Congress pass that can impact affordable housing in this country?*
3. *How does the change in interest rates by the Federal Reserve affect housing prices?*

Business Justification:

The first point we want to address is helping consumers better predict housing prices so they will be more informed on the proper time to purchase a home, and understand how much they will need to save. The second point is in regards to investment opportunities, we know the goal of most home-buyers isn't to just purchase a home, but to build equity via property appreciation over time. Based on the current price of the house and the features it contains, there can be a lot of equity made on the property if they better understand housing market trends to buy and sell at the right times. Lastly, this can be looked at at a national policy level where the President or Congress is able to pass laws that affect home inventory. This project will be able to help us understand these factors as this has massive real-life implications.

DATASET/PLAN FOR DATA (4 points)

Data Sources (links, attachments, etc.):

1. https://files.zillowstatic.com/research/public_csvs/zhvi/County_zhvi_uc_sfrcondo_tier_0.33_0.67_sm_sa_month.csv?t=1687066277
2. https://fred.stlouisfed.org/series/MORTGAGE30US
3. https://data.bls.gov/timeseries/CUUR0000SA0
4. https://fred.stlouisfed.org/series/ACTLISCOUUS
5. https://www2.census.gov/programs-surveys/popest/datasets/2010-2020/counties/totals/
   https://www2.census.gov/programs-surveys/popest/datasets/2020-2022/counties/totals/

Data Description (describe each of your data sources, include screenshots of a few rows of data):

1. Zillow Home Value Index by County: this data reflects monthly typical value for homes in the 35th to 65th percentile range on county level between January 2000 and May 2023. (Appendix Image 1)
2. 30-Year Fixed Rate Mortgage Average in the United States: This data provides weekly level data since April 1971. (Appendix Image 2)
3. CPI (Consumer Price Index) for All Urban consumers which is available on monthly basis with data from 1913, this data is typically used as a proxy for cost of living. (Appendix image 3)
4. Housing Inventory: Active listing count in the United States: This dataset counts the average active single-family and condo/townhouse listing for each month from 2016 (Appendix image 4)
5. This provides population estimates by county on yearly level since 2010 (Appendix image 5)

The independent variables are CPI (dataset 3), mortgage rate (dataset 2), population estimates by county (dataset 5), and active listings (dataset 4). The dependent variable is the home value index by county (dataset 1). We believe those existing variables will be enough, therefore, no new variables are planned to be created at this stage. And we believe CPI and Mortgage rate will have a major impact on the overall price, however, population growth in each county will also play an important role in determining various growth rates across the country.

APPROACH/METHODOLOGY (8 points)

Our approach starts with an exploratory analysis of the various data sources we were able to pull together. The aim is to gain insights into the data and identify potential features that may influence housing prices. Pulling together data from the vast amount of sources found online will require a significant amount of cleaning and pre-processing. We expect to

have to normalize various data sets, as well as one-hot encode categorical variables such as states, counties, and property types. Some features that we anticipate having a significant impact on housing prices are macroeconomic factors, demographic indicators, and historic pricing. Once the variables have been selected, we will split the data into training, validation, and testing sets. From there we will fit multiple models using various methodologies such as linear regression, decision trees, and random forest regressions. To optimize our model, we will use established procedures such as Bayesian optimization, grid search or randomized search. Since each approach to optimization carries its own benefits and drawbacks, we will decide on how to proceed once we have completed the variable selection phase. To evaluate different models, we will use AUC and R-squared to compare model performance. We will also use cross-validation to ensure that our model generalizes well to unseen data. The model with the best overall performance will then be used in our final analysis.

We expect our analysis to produce results in line with expert opinion. Additionally, the final model will have a reasonably low mean absolute error while exhibiting a relatively high R-squared value which should indicate a strong ability to accurately predict housing prices. Given such a model, we will be able to predict housing prices across various regions in order to identify potential investment opportunities.

Our analysis directly impacts various stakeholders in the real estate industry such as individuals, investors and even government bodies. The goal of the analysis is to provide these stakeholders with the knowledge to make better informed, strategic decisions regarding the real estate market. This will ultimately lead to benefits such as increased profitability, risk mitigation and improved operational efficiency.

PROJECT TIMELINE/PLANNING (2 points)

From project proposal to the final project deliverable, the project is completed over 4 weeks. Below are the major milestones we have identified. Deliverable details are outlined in the project guideline.

- *W0, 6/21 - **DELIVERABLE** - Project proposal (this document).*
- *W1, 6/28 - Explore, clean, and join data sets. Refine our objective and determine potential approaches. Determine project progress report and video proposal responsibilities.*
- *W2, 7/5 - **DELIVERABLES** - Project progress report and video proposal presentation. Determine project final report responsibilities.*
- *W3, 7/12 - First draft of project. Request TA feedback on areas of strength and weaknesses in our report.*
- *W4, 7/19 - Complete final draft of project report. Assign presentation responsibilities.*
- *W4, 7/20 - **DELIVERABLE** - Project final report.*
- *W4, 7/22 - Complete project video presentation.*
- *W4, 7/23 - **DELIVERABLE** - Project video presentation + slides.*

Appendix (any preliminary figures or charts that you would like to include):

Image 1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	RegionID	SizeRank	RegionName	RegionType	StateName	State	Metro	StateCodeFIF	MunicipalCo	31-Jan-2000	29-Feb-2000	31-Mar-2000	30-Apr-2000	31-May-2000	30-Jun-2000
2	3101	0	Los Angeles County	county	CA	CA	Los Angeles-Long Beach-Anaheim, CA	6	37	211518.8394	211745.3457	212616.4701	214336.1385	216536.2063	218648.1836
3	139	1	Cook County	county	IL	IL	Chicago-Naperville-Elgin, IL-IN-WI	17	31	136551.4357	136525.3128	136751.5897	137401.9069	138240.9084	139119.1506
4	1090	2	Harris County	county	TX	TX	Houston-The Woodlands-Sugar Land, TX	48	201	109337.9364	109308.2649	109146.6026	109078.7693	109052.177	109238.4209
5	2402	3	Maricopa County	county	AZ	AZ	Phoenix-Mesa-Chandler, AZ	4	13	145536.572	145834.2307	146227.1615	147034.8538	147925.0105	148677.4276
6	2841	4	San Diego County	county	CA	CA	San Diego-Chula Vista-Carlsbad, CA	6	73	216360.015	217210.3108	218243.8951	220584.3851	223113.1848	226306.3492
7	1286	5	Orange County	county	CA	CA	Los Angeles-Long Beach-Anaheim, CA	6	59	259417.6363	261699.817	263320.199	266636.3137	269500.7866	272533.8566
8	581	6	Kings County	county	NY	NY	New York-Newark-Jersey City, NY-NJ-PA	36	47	153127.055	153871.6796	154681.1374	156433.0386	158289.1388	160314.996
9	2964	7	Miami-Dade County	county	FL	FL	Miami-Fort Lauderdale-Pompano Beach, FL	12	86	110553.0617	110942.8162	111274.2875	111943.7874	112531.6235	113145.794
10	978	8	Dallas County	county	TX	TX	Dallas-Fort Worth-Arlington, TX	48	113	102480.0642	102544.5255	102632.45	102830.4873	103079.3374	103328.6292
11	2832	9	Riverside County	county	CA	CA	Riverside-San Bernardino-Ontario, CA	6	65	153760.4959	154108.9902	154534.3523	155507.4547	156779.3296	158345.6776
12	1347	10	Queens County	county	NY	NY	New York-Newark-Jersey City, NY-NJ-PA	36	81	136464.8708	137132.2465	137881.2046	139085.8662	140142.4692	141141.1331
13	207	11	King County	county	WA	WA	Seattle-Tacoma-Bellevue, WA	53	33	248565.8698	249569.2514	250547.316	252456.1222	254626.4424	256551.2679
14	445	12	Clark County	county	NV	NV	Las Vegas-Henderson-Paradise, NV	32	3	162153.0778	162127.6054	162394.0432	163019.5249	163840.831	164555.2453
15	3250	13	San Bernardino County	county	CA	CA	Riverside-San Bernardino-Ontario, CA	6	71	128198.2955	128935.5488	129585.8195	130843.6233	131848.6772	132724.2414

Image 2

1	DATE	MORTGAGE30US
2	1/7/00	8.15
3	1/14/00	8.18
4	1/21/00	8.26
5	1/28/00	8.25
6	2/4/00	8.25
7	2/11/00	8.36
8	2/18/00	8.38
9	2/25/00	8.31

Image 3

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	HALF1	HALF2
2000	168.8	169.8	171.2	171.3	171.5	172.4	172.8	172.8	173.7	174.0	174.1	174.0	170.8	173.6
2001	175.1	175.8	176.2	176.9	177.7	178.0	177.5	177.5	178.3	177.7	177.4	176.7	176.6	177.5
2002	177.1	177.8	178.8	179.8	179.8	179.9	180.1	180.7	181.0	181.3	181.3	180.9	178.9	180.9
2003	181.7	183.1	184.2	183.8	183.5	183.7	183.9	184.6	185.2	185.0	184.5	184.3	183.3	184.6
2004	185.2	186.2	187.4	188.0	189.1	189.7	189.4	189.5	189.9	190.9	191.0	190.3	187.6	190.2
2005	190.7	191.8	193.3	194.6	194.4	194.5	195.4	196.4	198.8	199.2	197.6	196.8	193.2	197.4

Image 4

1	DATE	ACTLISCOUUS
2	1/1/20	951225
3	2/1/20	927811
4	3/1/20	936768
5	4/1/20	941187
6	5/1/20	927807

Image 5

1	SUMLEV	REGION	DIVISION	STATE	COUNTY	STNAME	CTYNAME	CENSUS2010POP	ESTIMATESBASE2010	POPESTIMATE2010	POPESTIMATE2011
2	40	3	6	1	0	Alabama	Alabama	4779736	4780118	4785514	4799642
3	50	3	6	1	1	Alabama	Autauga County	54571	54582	54761	55229
4	50	3	6	1	3	Alabama	Baldwin County	182265	182263	183121	186579
5	50	3	6	1	5	Alabama	Barbour County	27457	27454	27325	27344
6	50	3	6	1	7	Alabama	Bibb County	22915	22904	22858	22736