Housing Affordability for Indiana State using federal dataset.

STAT 46700-001 TOPICS IN DATA SCIENCE XLST

Team Members:
Adusumilli Sree Sai
Andrea George
Dharani Meka

PURDUE UNIVERSITY NORTHWEST

CONTENTS:-

1. Motivation of the Project
2. Dataset Details
3. Handling missing values
4. Raw Dataset
5. Analysis
6.Significance of dataset features
7. Conclusion

MOTIVATION:-

This project aims to address the critical issue of housing affordability in Indiana by leveraging comprehensive federal datasets spanning 12 years (2010-2021) obtained from the Census Bureau.

Through the analysis of the percentage of income spent on housing, calculated as the median yearly housing cost divided by the median income, we seek to gain insights into housing affordability trends over time, as well as regional disparities and demographic variations within the state.

DATASET DETAILS:-

0:Households below poverty line B17017_002E

1. Distribution of household income by income bracket B19001_001E

2.Real estate taxes:B25102_008E

3.Unemployed population DP03_0109E

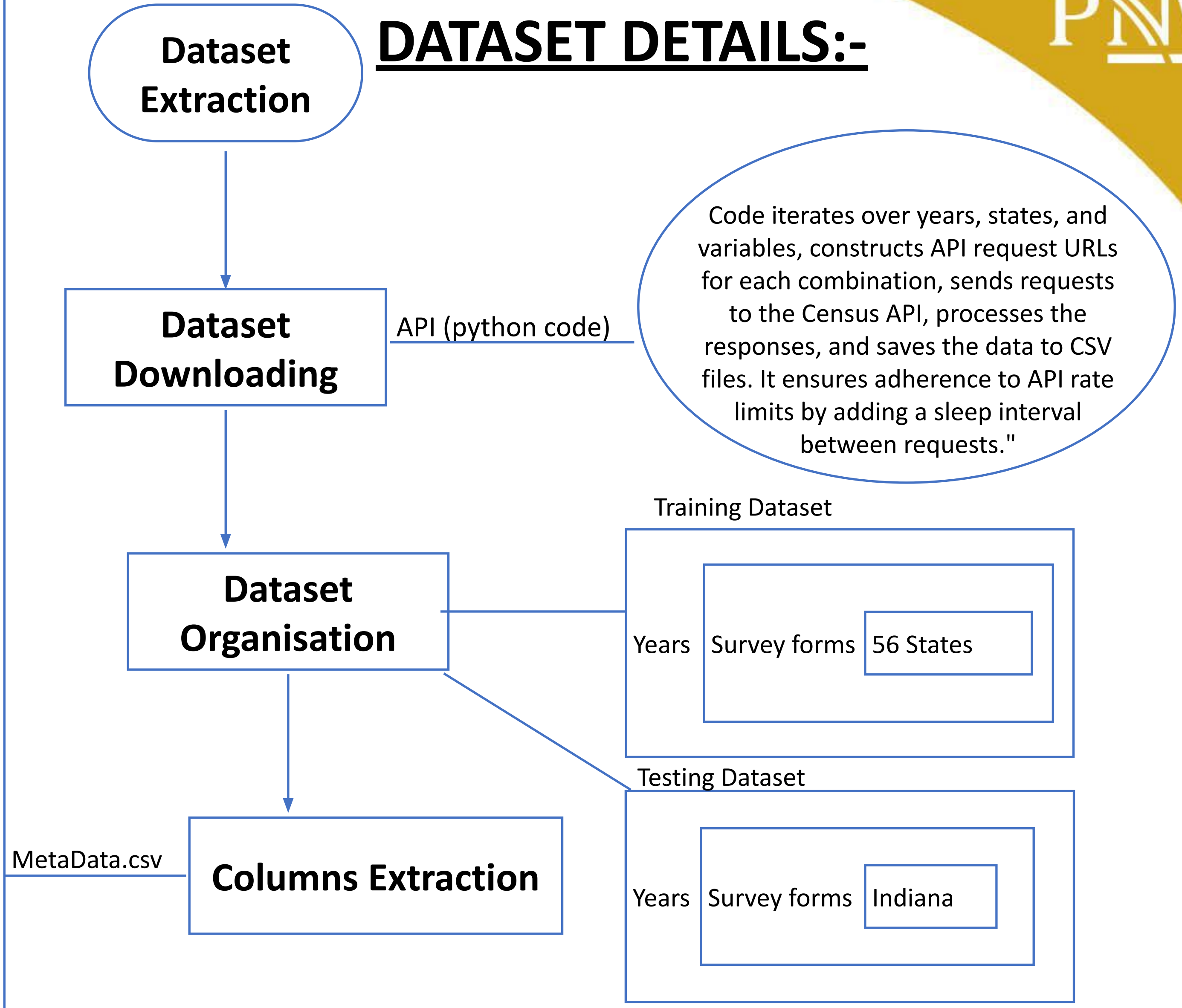
4.Family Income by single earner household S1903_C02_016E

5. Total Occupied housing units: S2503_C01_001E

6. Median income of occupied housing units S2503_C01_013E

7. Yearly housing costs S2503_C01_028E

8.Distribution of Properties by Property Value S2506_C01_002E



Dataset Extraction

Dataset Downloading

API (python code)

Code iterates over years, states, and variables, constructs API request URLs for each combination, sends requests to the Census API, processes the responses, and saves the data to CSV files. It ensures adherence to API rate limits by adding a sleep interval between requests."

Dataset Organisation

Training Dataset

Years

Survey forms

56 States

Testing Dataset

Years

Survey forms

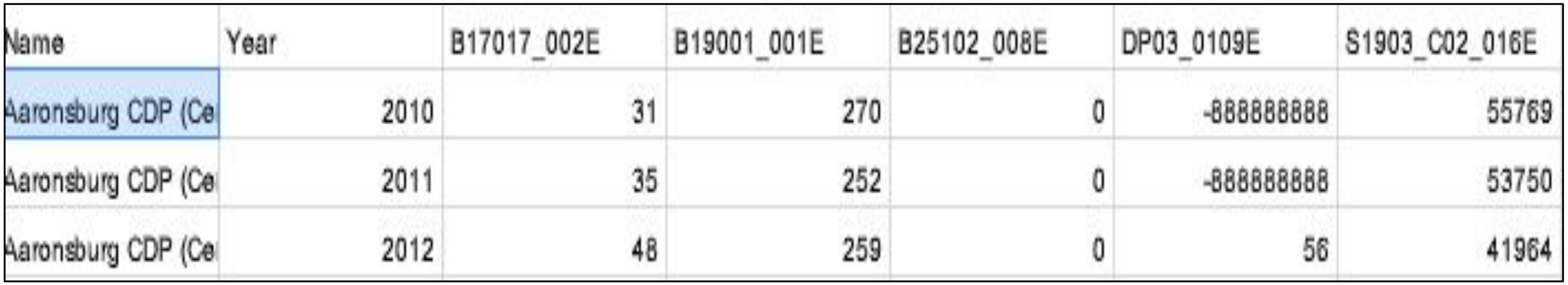
Indiana

MetaData.csv

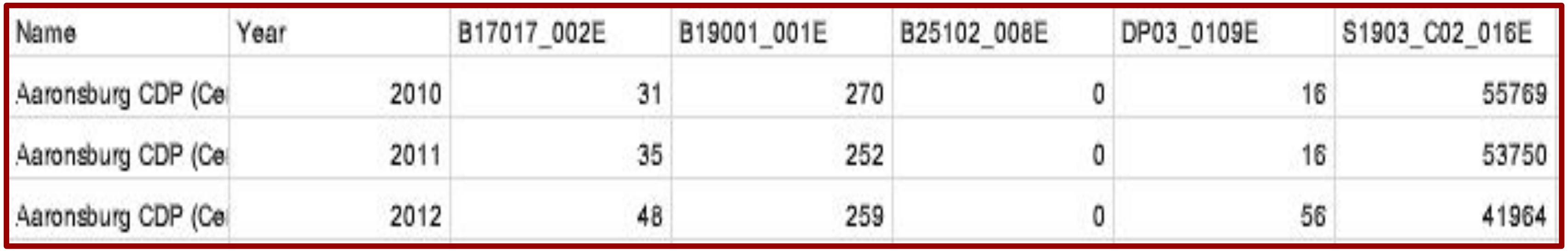
Columns Extraction

HANDLING MISSING DATASET VALUES:-

- After going through the website https://data.census.gov, it is observed that the missing values have been represented by the Jam codes which looks like −666666666.0+66E.
- Missing values have been handled by replacing the Jam codes with the median value of that particular column data.

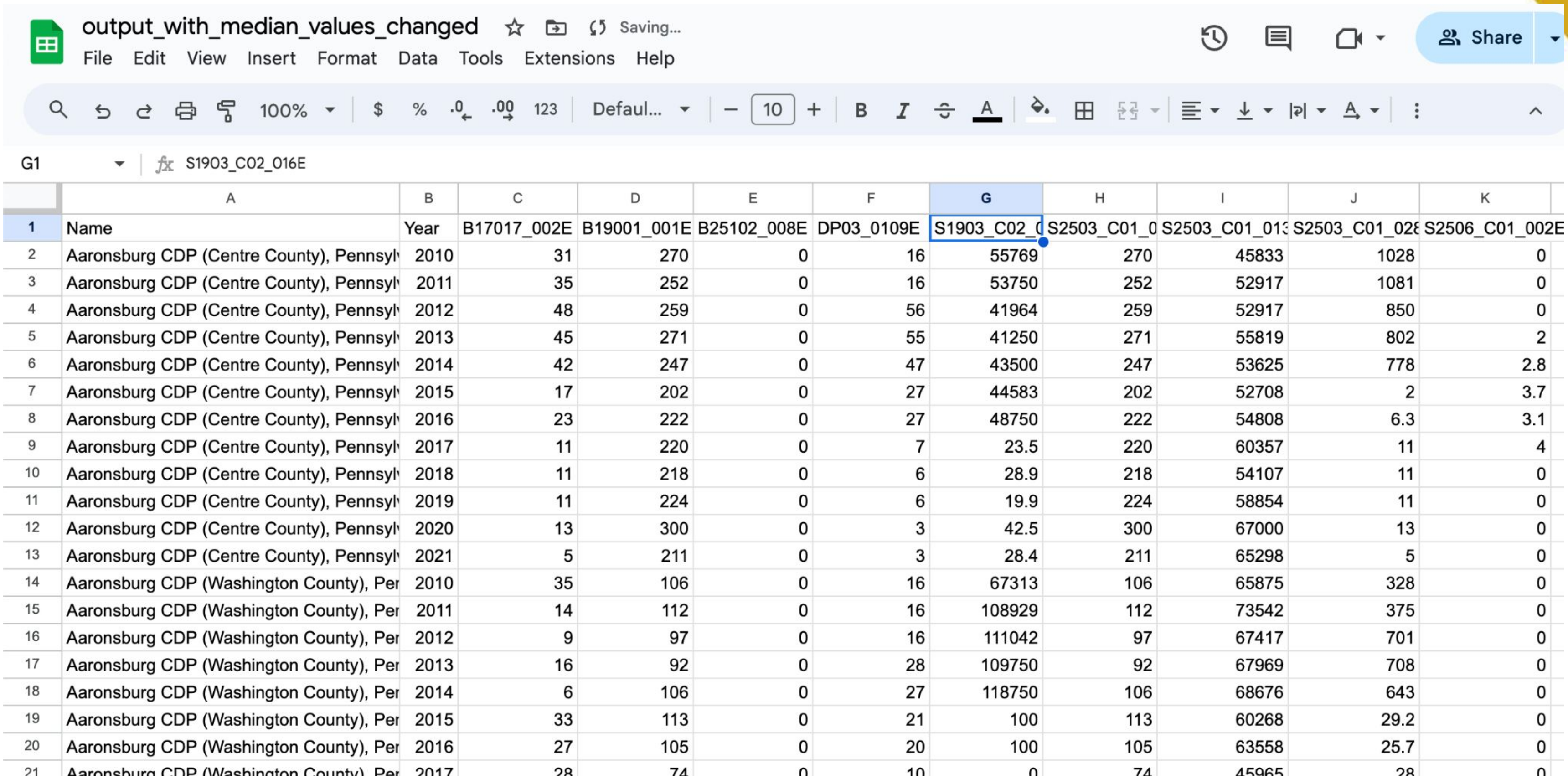


Name	Year	B17017_002E	B19001_001E	B25102_008E	DP03_0109E	S1903_C02_016E
Aaronsburg CDP (Ce	2010	31	270	0	-888888888	55769
Aaronsburg CDP (Ce	2011	35	252	0	-888888888	53750
Aaronsburg CDP (Ce	2012	48	259	0	56	41964



Name	Year	B17017_002E	B19001_001E	B25102_008E	DP03_0109E	S1903_C02_016E
Aaronsburg CDP (Ce	2010	31	270	0	16	55769
Aaronsburg CDP (Ce	2011	35	252	0	16	53750
Aaronsburg CDP (Ce	2012	48	259	0	56	41964

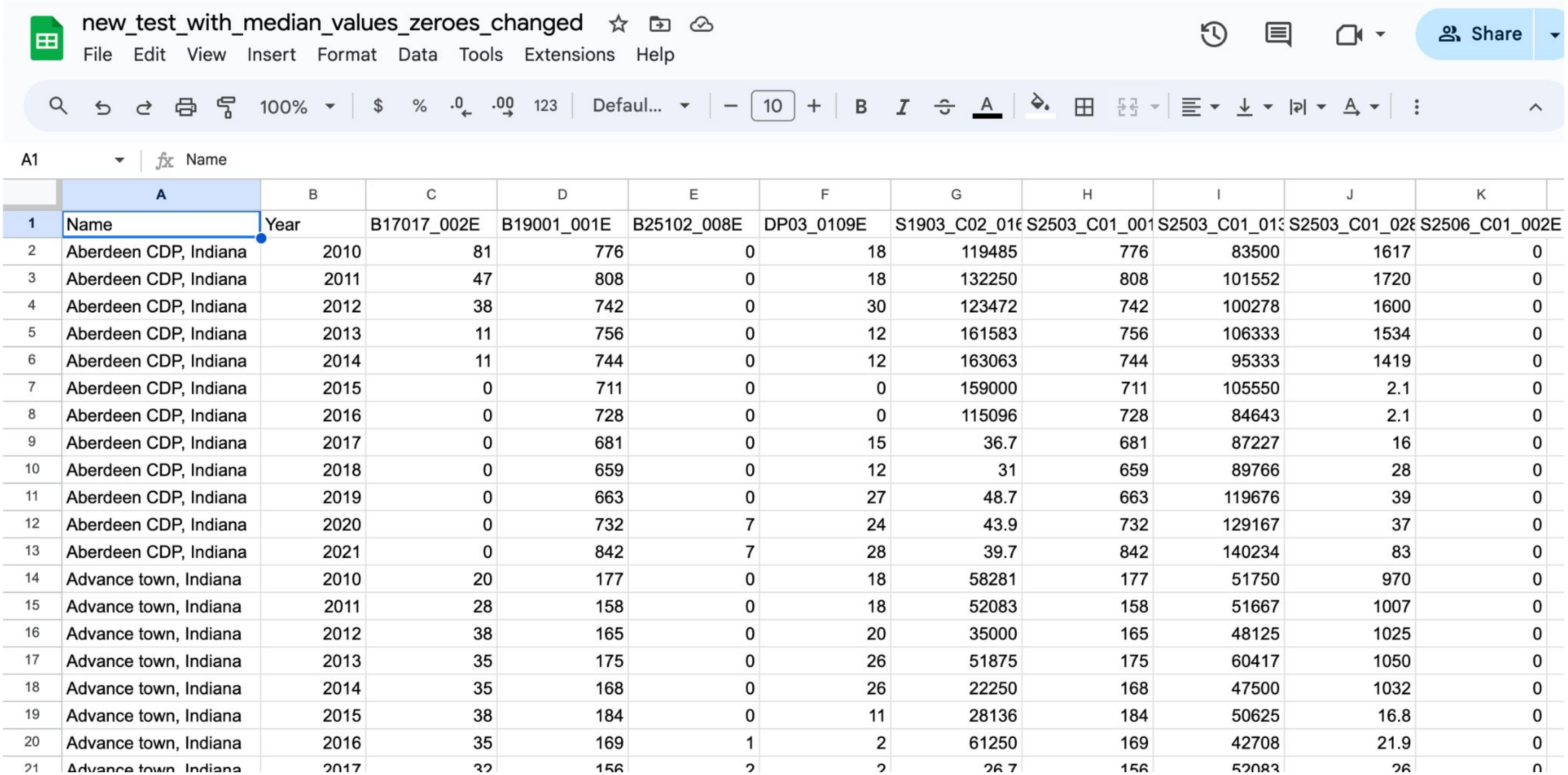
FINAL RAW DATASET :-



G1 *fx* S1903_C02_016E

	A	B	C	D	E	F	G	H	I	J	K
1	Name	Year	B17017_002E	B19001_001E	B25102_008E	DP03_0109E	S1903_C02_0	S2503_C01_0	S2503_C01_013	S2503_C01_028	S2506_C01_002E
2	Aaronsburg CDP (Centre County), Pennsyl	2010	31	270	0	16	55769	270	45833	1028	0
3	Aaronsburg CDP (Centre County), Pennsyl	2011	35	252	0	16	53750	252	52917	1081	0
4	Aaronsburg CDP (Centre County), Pennsyl	2012	48	259	0	56	41964	259	52917	850	0
5	Aaronsburg CDP (Centre County), Pennsyl	2013	45	271	0	55	41250	271	55819	802	2
6	Aaronsburg CDP (Centre County), Pennsyl	2014	42	247	0	47	43500	247	53625	778	2.8
7	Aaronsburg CDP (Centre County), Pennsyl	2015	17	202	0	27	44583	202	52708	2	3.7
8	Aaronsburg CDP (Centre County), Pennsyl	2016	23	222	0	27	48750	222	54808	6.3	3.1
9	Aaronsburg CDP (Centre County), Pennsyl	2017	11	220	0	7	23.5	220	60357	11	4
10	Aaronsburg CDP (Centre County), Pennsyl	2018	11	218	0	6	28.9	218	54107	11	0
11	Aaronsburg CDP (Centre County), Pennsyl	2019	11	224	0	6	19.9	224	58854	11	0
12	Aaronsburg CDP (Centre County), Pennsyl	2020	13	300	0	3	42.5	300	67000	13	0
13	Aaronsburg CDP (Centre County), Pennsyl	2021	5	211	0	3	28.4	211	65298	5	0
14	Aaronsburg CDP (Washington County), Per	2010	35	106	0	16	67313	106	65875	328	0
15	Aaronsburg CDP (Washington County), Per	2011	14	112	0	16	108929	112	73542	375	0
16	Aaronsburg CDP (Washington County), Per	2012	9	97	0	16	111042	97	67417	701	0
17	Aaronsburg CDP (Washington County), Per	2013	16	92	0	28	109750	92	67969	708	0
18	Aaronsburg CDP (Washington County), Per	2014	6	106	0	27	118750	106	68676	643	0
19	Aaronsburg CDP (Washington County), Per	2015	33	113	0	21	100	113	60268	29.2	0
20	Aaronsburg CDP (Washington County), Per	2016	27	105	0	20	100	105	63558	25.7	0
21	Aaronsburg CDP (Washington County), Per	2017	28	74	0	10	0	74	45965	28	0

Testing dataset(Indiana)



	A	B	C	D	E	F	G	H	I	J	K
1	Name	Year	B17017_002E	B19001_001E	B25102_008E	DP03_0109E	S1903_C02_016	S2503_C01_001	S2503_C01_013	S2503_C01_028	S2506_C01_002E
2	Aberdeen CDP, Indiana	2010	81	776	0	18	119485	776	83500	1617	0
3	Aberdeen CDP, Indiana	2011	47	808	0	18	132250	808	101552	1720	0
4	Aberdeen CDP, Indiana	2012	38	742	0	30	123472	742	100278	1600	0
5	Aberdeen CDP, Indiana	2013	11	756	0	12	161583	756	106333	1534	0
6	Aberdeen CDP, Indiana	2014	11	744	0	12	163063	744	95333	1419	0
7	Aberdeen CDP, Indiana	2015	0	711	0	0	159000	711	105550	2.1	0
8	Aberdeen CDP, Indiana	2016	0	728	0	0	115096	728	84643	2.1	0
9	Aberdeen CDP, Indiana	2017	0	681	0	15	36.7	681	87227	16	0
10	Aberdeen CDP, Indiana	2018	0	659	0	12	31	659	89766	28	0
11	Aberdeen CDP, Indiana	2019	0	663	0	27	48.7	663	119676	39	0
12	Aberdeen CDP, Indiana	2020	0	732	7	24	43.9	732	129167	37	0
13	Aberdeen CDP, Indiana	2021	0	842	7	28	39.7	842	140234	83	0
14	Advance town, Indiana	2010	20	177	0	18	58281	177	51750	970	0
15	Advance town, Indiana	2011	28	158	0	18	52083	158	51667	1007	0
16	Advance town, Indiana	2012	38	165	0	20	35000	165	48125	1025	0
17	Advance town, Indiana	2013	35	175	0	26	51875	175	60417	1050	0
18	Advance town, Indiana	2014	35	168	0	26	22250	168	47500	1032	0
19	Advance town, Indiana	2015	38	184	0	11	28136	184	50625	16.8	0
20	Advance town, Indiana	2016	35	169	1	2	61250	169	42708	21.9	0
21	Advance town, Indiana	2017	32	156	2	2	26.7	156	52083	26	0

ANALYSIS OF NEURAL NETWORK MODELS:-

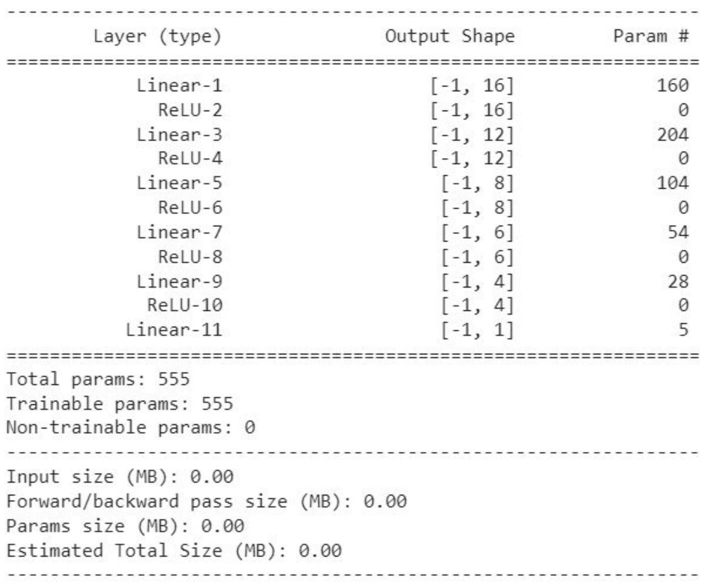
METRIC —---------------- MODEL No.	Mean Squared Error(MSE)	Root Mean Squared Error (RMSE)	Mean Absolute Error (MAE)	Coefficient of Determination (r2 score)

Hidden Layers: 3
batch_size=16
learning_rate=0.001
n_epochs=50
	11.046147	3.3235745	1.160193	0.99161837912

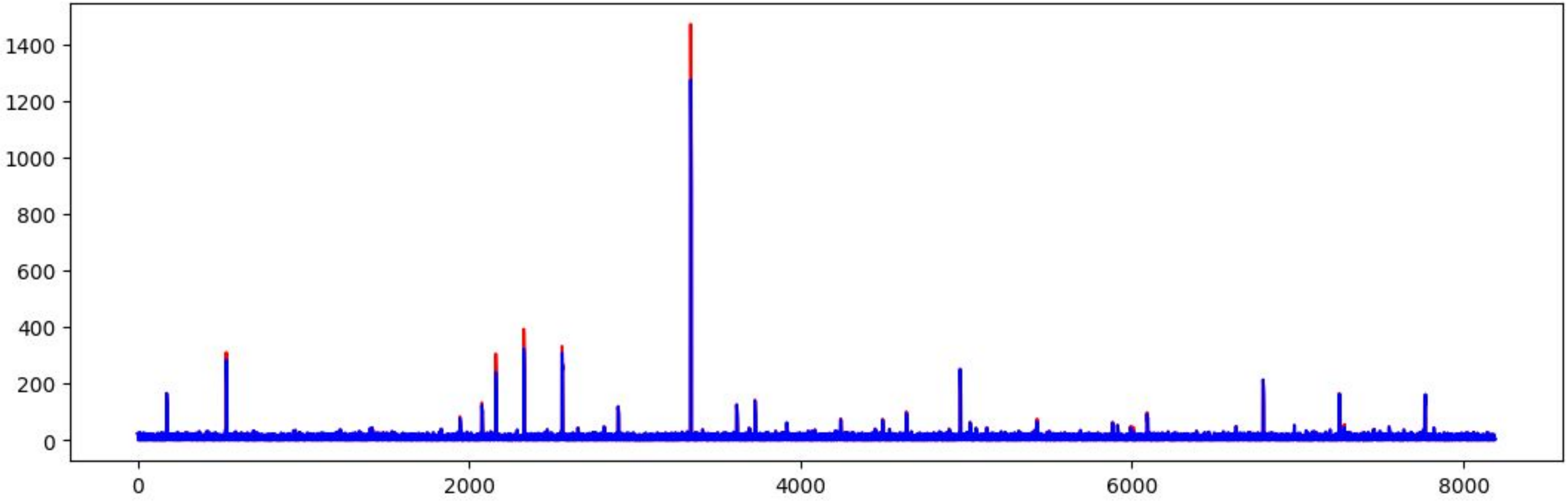
Hidden Layers: 5
batch_size=16
learning_rate=0.001
n_epochs=50
	3.6760151	1.9172937	0.95963496	0.9972107046

batch_size=16
learningrate=0.0005
n_epochs=50
	37.385357	6.1143565	1.0551189	0.97163265309

batch_size=16
learning_rate=0.001
n_epochs=70
	7.2646923	2.6953094	0.87021846	0.99448768034



```
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Linear-1                   [-1, 16]             160
              ReLU-2                   [-1, 16]               0
            Linear-3                   [-1, 12]             204
              ReLU-4                   [-1, 12]               0
            Linear-5                    [-1, 8]             104
              ReLU-6                    [-1, 8]               0
            Linear-7                    [-1, 6]              54
              ReLU-8                    [-1, 6]               0
            Linear-9                    [-1, 4]              28
             ReLU-10                    [-1, 4]               0
           Linear-11                    [-1, 1]               5
================================================================
Total params: 555
Trainable params: 555
Non-trainable params: 0
----------------------------------------------------------------
Input size (MB): 0.00
Forward/backward pass size (MB): 0.00
Params size (MB): 0.00
Estimated Total Size (MB): 0.00
----------------------------------------------------------------
```



ANALYSIS OF DECISION TREE MODELS :-

METRIC —--------------- MODEL No.	Train Root Mean Squared Error	Test Root Mean Squared Error	Feature Importance

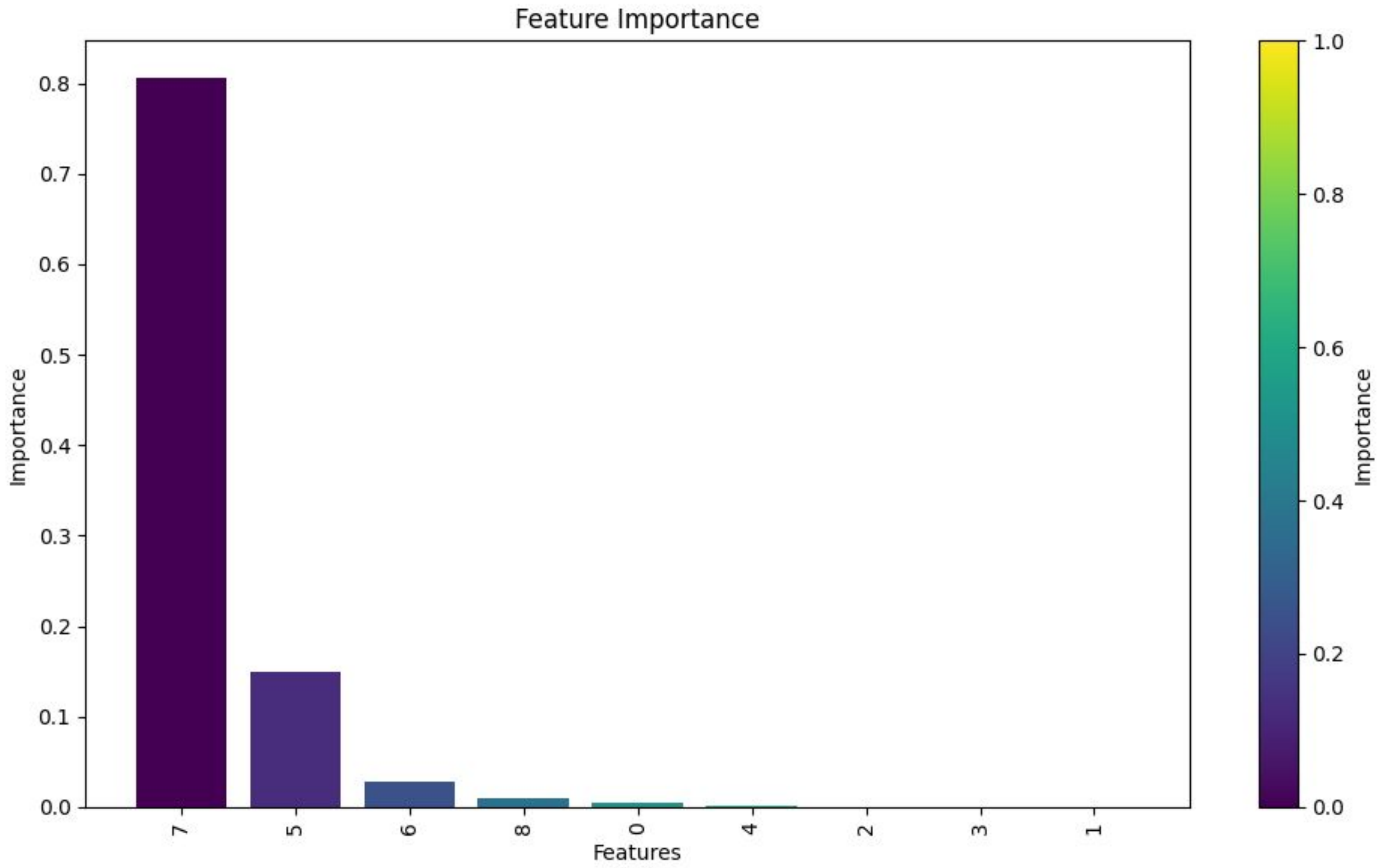
max_depth = 10
min_samples_split = 2
	Train RMSE: 3.288714787896192	Test RMSE: 7.088595040258605	Yearly Housing Costs Importance: 0.8065762944919934

max_depth = 12
min_samples_split = 2
	Train RMSE: 1.9832226405116682	Test RMSE: 6.287172930543692	Yearly Housing Costs Importance = 0.79959185853688510.00916565935642405

max_depth=10,
min_samples_split=2
k=5
	Average Train RMSE: 7.190047611364362	Average Test RMSE: 4.764804232894837	

max_depth=12,
min_samples_split=2
k=7
	Average Train RMSE: 6.500499552270875	Average Test RMSE: 4.0195625029625	

Significance of the Dataset Features



CONCLUSION:-

- Through the implementation of both a deep learning regression model and a decision tree model, we have effectively addressed the task of predicting housing affordability.

- Our deep learning network demonstrates high accuracy, as evidenced by its high R2 score, indicating strong predictive capability across the given nine features.

- Additionally, the decision tree model showcases superior performance in terms of root mean square error, highlighting its efficiency in capturing complex relationships within the dataset.

Thank You