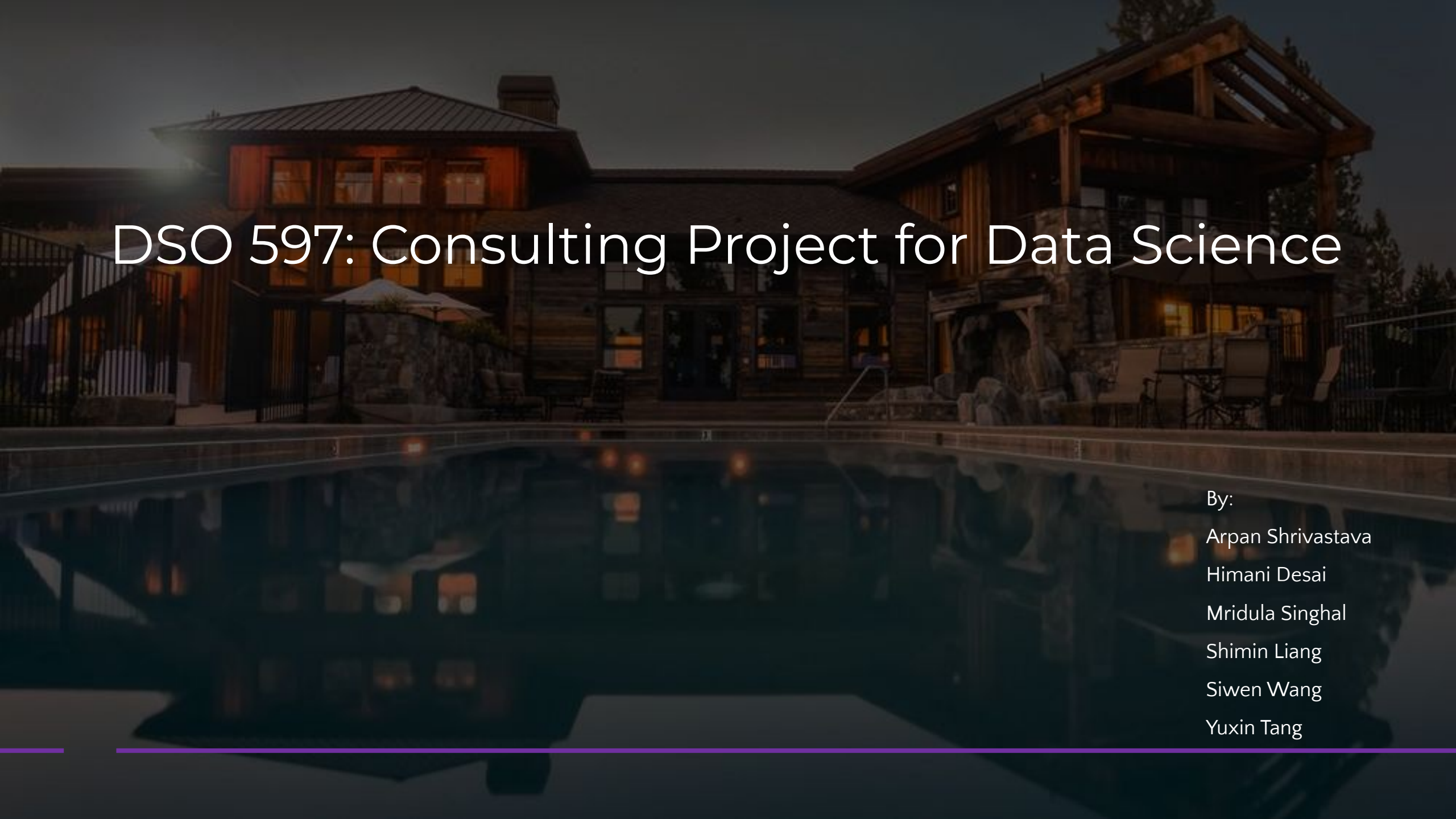


DSO 597: Consulting Project for Data Science

By:

Arpan Shrivastava

Himani Desai

Mridula Singhal

Shimin Liang

Siwen Wang

Yuxin Tang

Project Summary

In this project, we have conducted applied real estate data science research to identify relevant demographic, social, economic, and financial data sources to access and retrieve data from Census, Kaggle, FEMA, Kroger, Starbucks and Walmart to perform exploratory, numerical and visual data analysis. We have further conducted predictive and prescriptive modeling using Multiple Regression and integrated data visualization graphs to predict and formulate a recommendation on where Fondo Atlas and Centro.ia should invest to maximize returns.

We have divided the project into four parts:

- **PART 1:-**
  We predict rent per square foot using multiple factors on zip code level data

- **PART 2:-**
  We predict rent per square foot using property ranking and age

- **PART 3:-**
  We collect data of total population and unemployment of recent years on zip code level and calculate the population growth rate and unemployment rate to reflect the demand of rent.

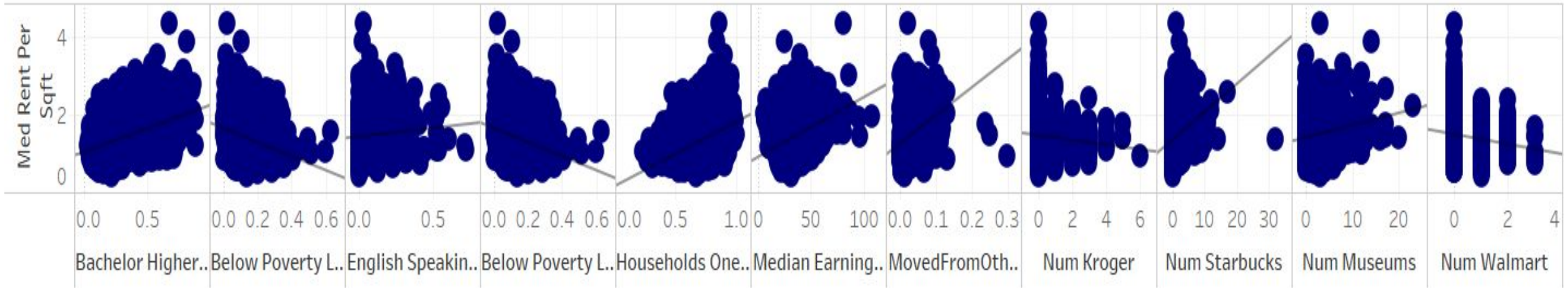
- **PART 4:-**
  Final analysis for investment by using 2 methods to give the properties which would give the highest return on investment.

Variables Used in Part 1

For each Zip Code:

- **MED_RENT_PER_SQFT**: median rent per sqft
- **Had_Birth_35to50_pct**: what percentage of females who had birth in 2019 are between age 35 and 50 years\
- **NeverMarriedMales_20to34_pct**：what percentage of males with age from 20 to 34 years are never married\
- **Median_Earnings_thousand**: median household income
- **English_Speaking_Households_pct**：what percenatage of households are English speaking households
- **MovedFromOtherStates_25YearsOver_pct**： what percentage of people with age 25 years and over are moved from other states
- **Households_OneOrMore_Computer_pct**: what percentage of households possess one or more desktops or laptops
- **Num_Costco**: number of Costcos in it
- **Num_Walmart**: number of Walmarts in it
- **Num_Kroger**: number of Krogers in it
- **Num_Starbucks**: number of Starbucks in it
- **Num_Museums**: number of museums, aquariums, zoos, etc. in it

Correlation Analysis of Independent Variables with Median Rent Per Sqft



Bachelor Higher pct, Below Poverty Level Families pct, English Speaking Households pct, Below Poverty Level Families pct, Households OneOrMore Computer pct, Median Earnings thousand, MovedFromOtherStates 25YearsOver pct, Num Kroger, Num Starbucks, Num Museums and Num Walmart vs. Med Rent Per Sqft.

47.4% of the variance for median rent per sqft of a zip code can be explained by the chosen 11 variables, each representing a characteristic of this zip code

OLS Regression Results

Dep. Variable:	MED_RENT_PER_SQFT	R-squared:	0.474
Model:	OLS	Adj. R-squared:	0.472
Method:	Least Squares	F-statistic:	217.5
Date:	Mon, 21 Mar 2022	Prob (F-statistic):	0.00
Time:	13:44:04	Log-Likelihood:	-724.78
No. Observations:	2662	AIC:	1474.
Df Residuals:	2650	BIC:	1544.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.8722	0.071	-12.270	0.000	-1.012	-0.733
Had_Birth_35to50_pct	0.3060	0.044	7.009	0.000	0.220	0.392
NeverMarriedMales_20to34_pct	1.0701	0.066	16.296	0.000	0.941	1.199
Median_Earnings_thousand	0.0085	0.001	10.199	0.000	0.007	0.010
English_Speaking_Households_pct	1.7933	0.100	17.850	0.000	1.596	1.990
MovedFromOtherStates_25YearsOver_pct	3.6883	0.347	10.638	0.000	3.008	4.368
Households_OneOrMore_Computer_pct	1.2869	0.075	17.186	0.000	1.140	1.434
Num_Costco	0.0835	0.020	4.105	0.000	0.044	0.123
Num_Walmart	-0.0689	0.012	-5.627	0.000	-0.093	-0.045
Num_Kroger	-0.0932	0.010	-9.038	0.000	-0.113	-0.073
Num_Starbucks	0.0172	0.004	4.663	0.000	0.010	0.024
Num_Museums	0.0196	0.003	6.559	0.000	0.014	0.025

Omnibus:	345.020	Durbin-Watson:	1.171
Prob(Omnibus):	0.000	Jarque-Bera (JB):	906.954
Skew:	0.714	Prob(JB):	1.14e-197
Kurtosis:	5.477	Cond. No.	2.11e+03

RENT PREDICTIONS FOR ZIP CODES

We used our model to predict the median rent per square foot for a certain zip code.

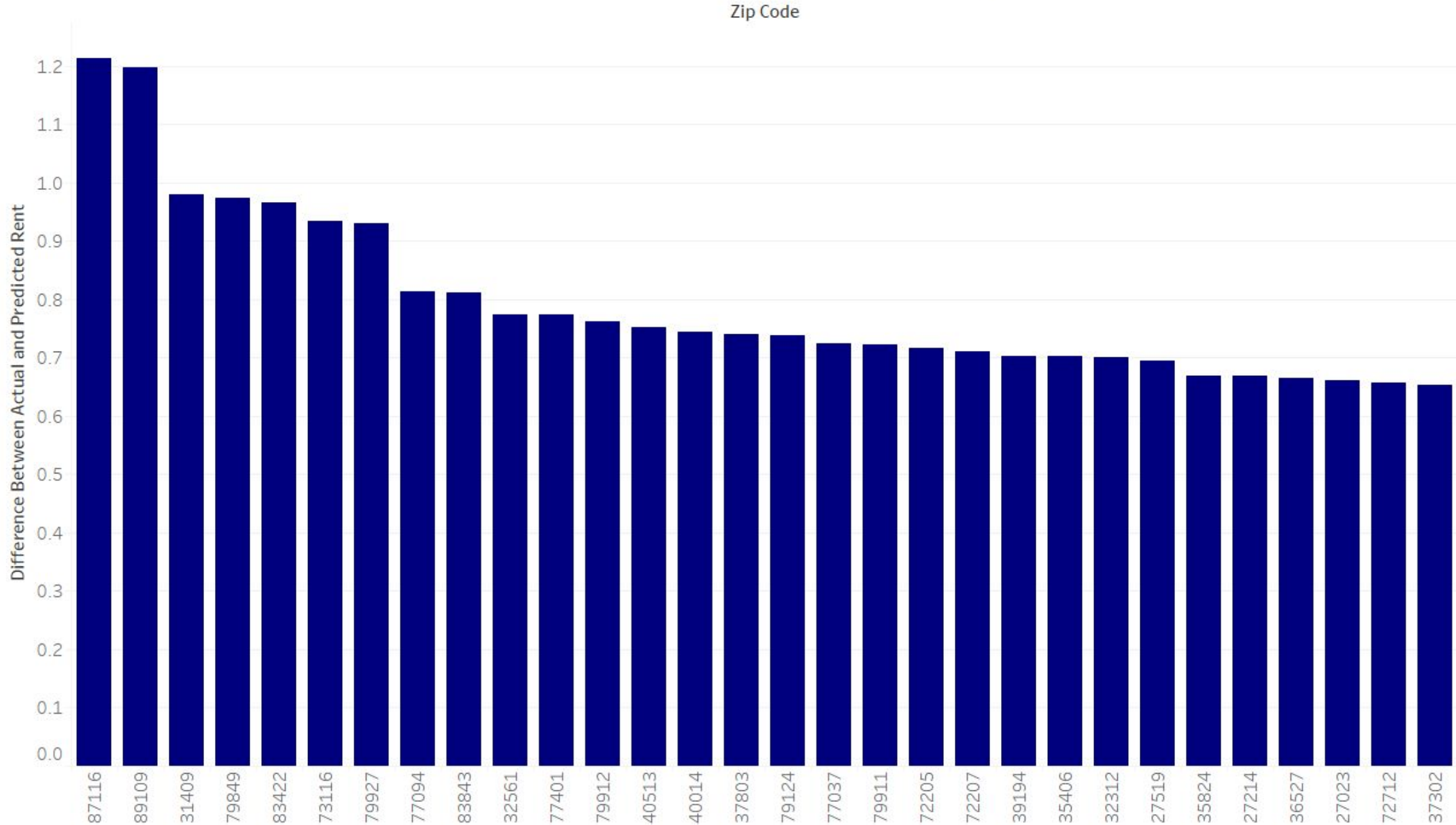
We the subtracted it with the actual median rent per square foot of the zipcode.

This way we got the potential increase or decrease in the rent per square foot for each property in a zipcode.

The following table demonstrates the zip rank variable we created. The zip rank indicates the potential for rent growth. Lower the rank, higher the potential.

	Zip Code	MED_RENT_PER_SQFT	Zip_Prediction	Zip_Diff	Zip_Rank
2572	87116	0.926792	2.140326	1.213533	1
2612	89109	1.347579	2.544902	1.197323	2
462	31409	1.489480	2.469451	0.979971	3
2174	79849	1.100953	2.073549	0.972596	4
2338	83422	0.857306	1.823393	0.966087	5

Difference Between Actual and Predicted Rents For Top 30 Zipcodes



Average of Diff for each Zip Code.

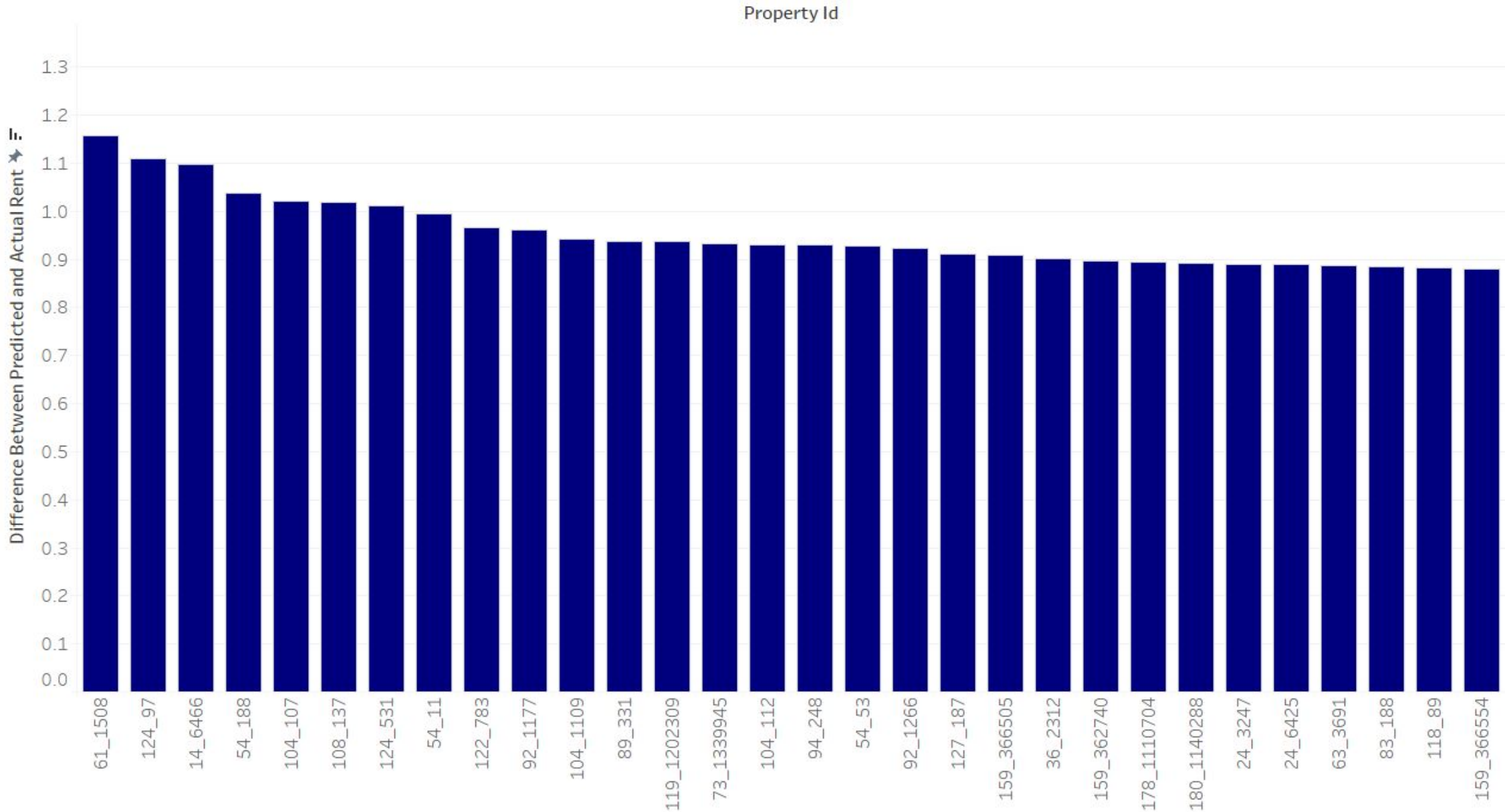
Variables Considered at Property ID Level Used in Part 2

- PROPERTY_IMPRATING
- PROPERTY_LOCRATING
- PROPERTY_YEAR

Regression Summary at Property Level for Part 2

Looking at our model which used property level data to predict the rent per square feet, we can explain about 29.1% of the data.

Top 30 Properties by Property Location Rating, Improvement Rating and Property Age



Average of Prop Diff for each Property Id. Details are shown for Zip Code. The data is filtered on Prop Rank, which has multiple members selected.

RENT PREDICTIONS FOR INDIVIDUAL PROPERTIES

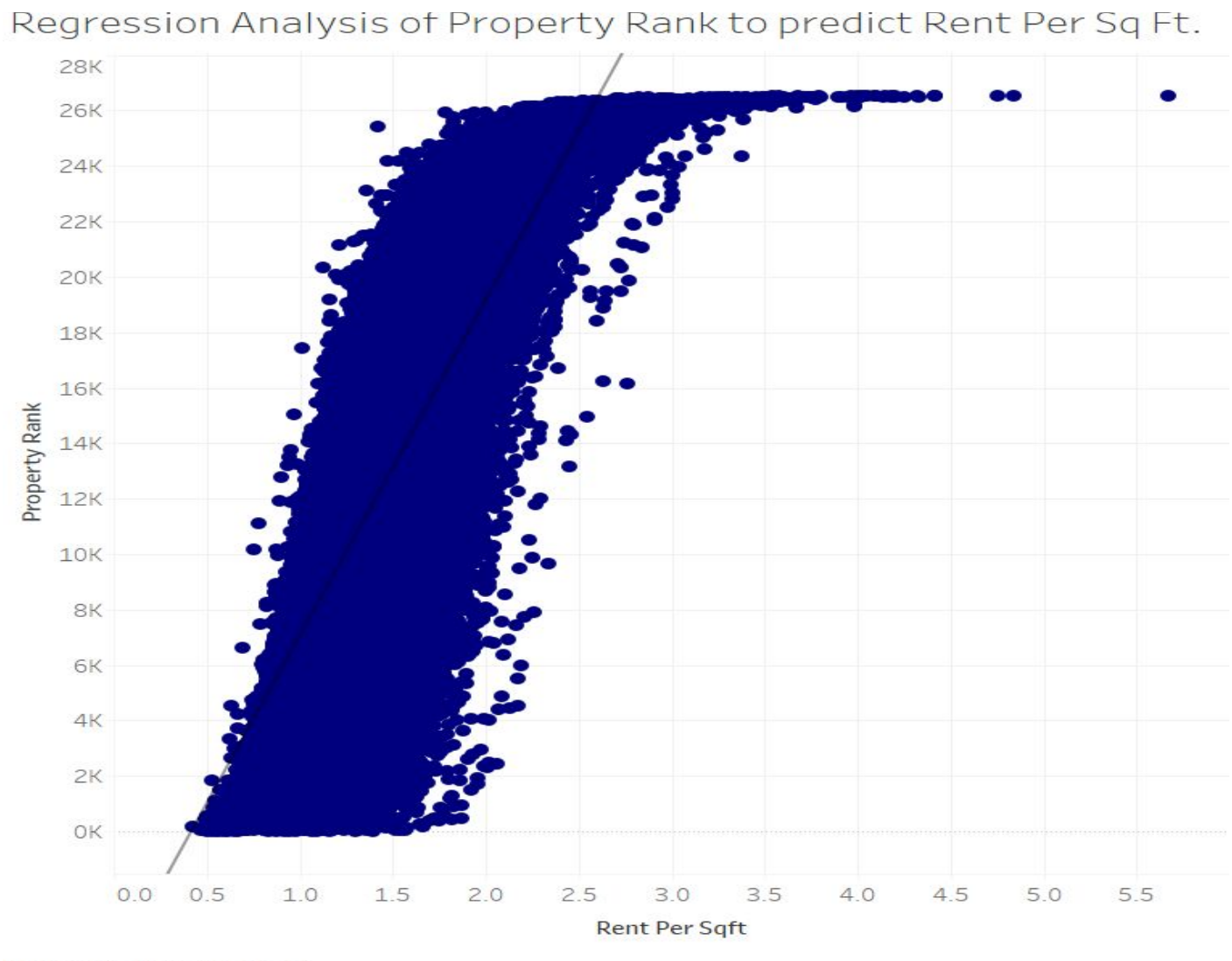
We used our model to predict the rent per square foot for individual properties.

We then subtracted it with the actual rent per square foot of the property.

This way we got the potential increase or decrease in the rent per square foot for each property.

The following table demonstrates the property rank variable we created. The property rank indicates the potential for rent growth. Lower the rank, higher the potential.

	PROPERTY_ID	Zip Code	PROPERTY_SQFT	RENT_PER_SQFT	Prof_Prediction	Prop_Diff	Prop_Rank
10883	61_1508	78258	480460	1.295508	2.451608	1.156099	1
5723	124_97	30809	260404	1.087664	2.196034	1.108370	2
538	14_6466	89149	112404	1.162734	2.260548	1.097814	3
4794	54_188	72211	571835	1.392445	2.431280	1.038834	4
1330	104_107	37803	53450	0.564079	1.584505	1.020426	5


Regression Analysis of Property Rank to predict Rent Per Sq Ft.
Property Rank
28K
26K
24K
22K
20K
18K
16K
14K
12K
10K
8K
6K
4K
2K
0K
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
Rent Per Sqft

PART 3

We merged the previous two parts and utilized the zip code rank along with property rank to combine these results with factors that affect housing demand namely ‘Population Growth Rate’ and ‘Unemployment Rate’ to select the ideal properties for investment

Metrics used for Part 3

- PROPERTY_ID
- Zip Code
- PROPERTY_SQFT
- RENT_PER_SQFT
- Prop_Prediction
- Prop_Diff
- Prop_Rank
- MED_RENT_PER_SQFT
- Zip_Prediction
- Zip_Diff
- Zip_Rank
- Population_Avg_Yearly_Growth
- Population_Rank
- UnemploymentRate_Avg_Yearly_Growth
- Unemployment_Rank
- Dollar

Summary for Part 3

We collected the data of total population and average unemployment from the years 2016 to 2020 on zip code level and calculated the population growth rate and unemployment rate to reflect the demand of rent.

We merged all the three parts to select ideal properties for investment. We calculated the estimated dollars we can gain for each property

*Estimated dollars = ((Predicted rent per sqft from part 1 + Predicted rent per sqft from part 2) / 2 - Actual rent per sqft) * sqft*

PART 4:

We came up with two methods to get the five properties that we recommend.

1. Set the threshold of the rank of zipcode in part 1 and the rank of property in part 2 until we get only 5 properties: Zip_Rank<30, Prop_Rank<163

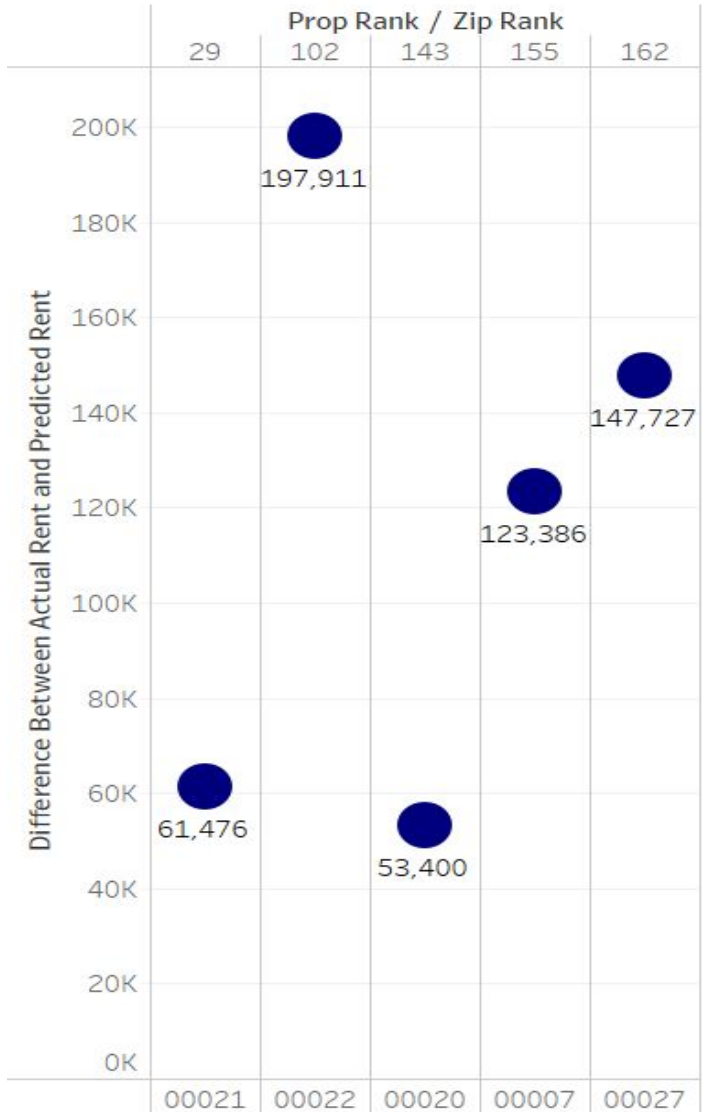
2. Set thresholds for Zip_Rank and Prop_Rank to be top 10%, select 5 properties with highest 'Dollar'

Finally we agree that the second one is more reasonable.

PROPERTY_ID	Zip Code	Prop_Rank	Zip_Rank	Dollar
118_89	39194	29	21	61476.236345
73_361	35406	102	22	197910.693556
54_140	72207	143	20	53400.381846
89_451	79927	155	7	123386.282647
119_1088	36527	162	27	147726.931134

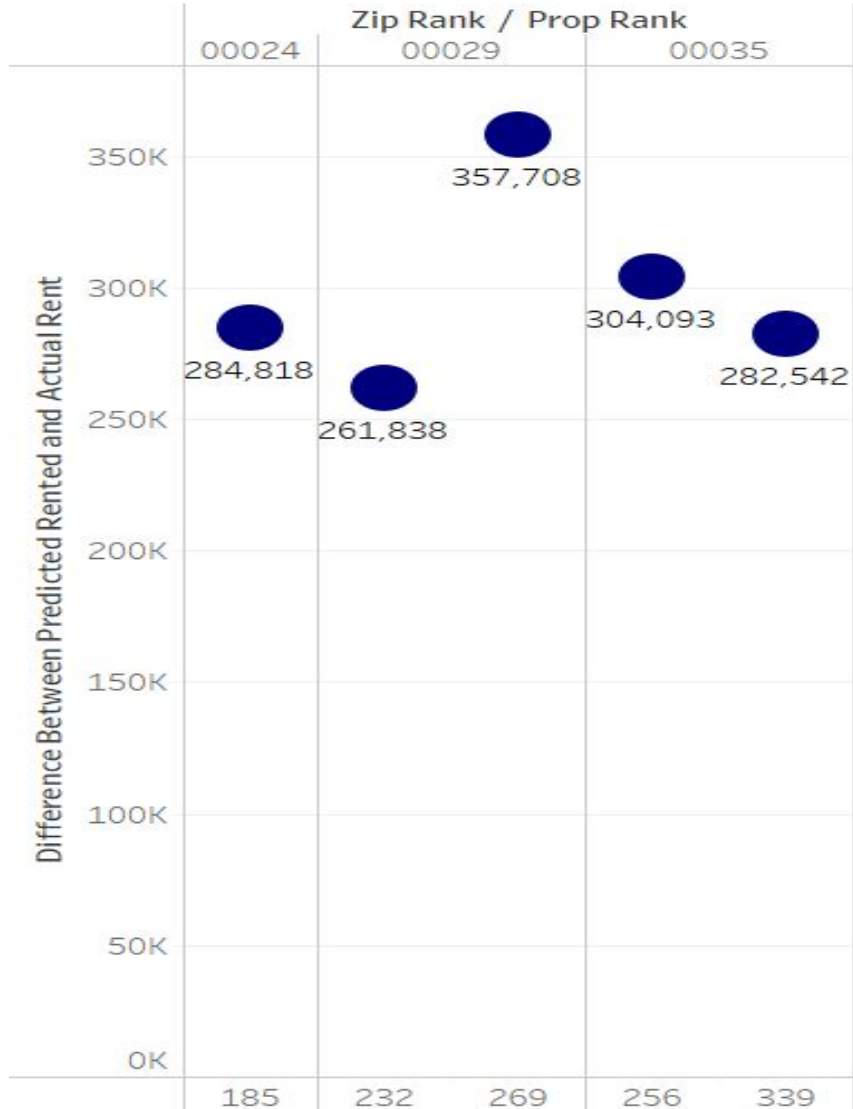
PROPERTY_ID	Zip Code	Prop_Rank	Zip_Rank	Dollar
178_1110800	72712	269	29	357707.714324
94_166	74075	256	35	304092.540451
46_1554	27519	185	24	284817.710109
94_87	74075	339	35	282542.017736
178_1110703	72712	232	29	261837.641540

Method 1 : Top 5 Properties By Zip Code Rank and Property Rank



Average of Dollar for each Zip Rank broken down by Prop Rank.

Method 2: Top 5 Properties by Zip Code Rank and Property Rank



Average of Dollar for each Prop Rank broken down by Zip Rank.

Final Results

We recommend the below 5 properties for investment

1) Bentonville, Arkansas (178_1110800)
2) Maramec, Oklahoma (94_166)
3) Cary, North Carolina (46_1554)
4) Stillwater, Oklahoma (94_87)
5) Bentonville, Arkansas (178_1110703)


Thank You