Is there a relationship between **house prices** and **walk score**?

<u>CS 797Q</u>
Applied and Practical Data Science

By:
- Purva Natoo - R644J946
- Curtis Martin - E252H926
- Francisco Javier Rafful Garfias - Z367M789
- Sriram Srinivasan - E334W844
- Vijay Ram - F448J755
- Sai Chandana Kondamadugula - N897P533

<u>Group #: 5</u>

It's Walkable

Introduction

FOCUS
Correlation between neighborhood's walkability and house prices.

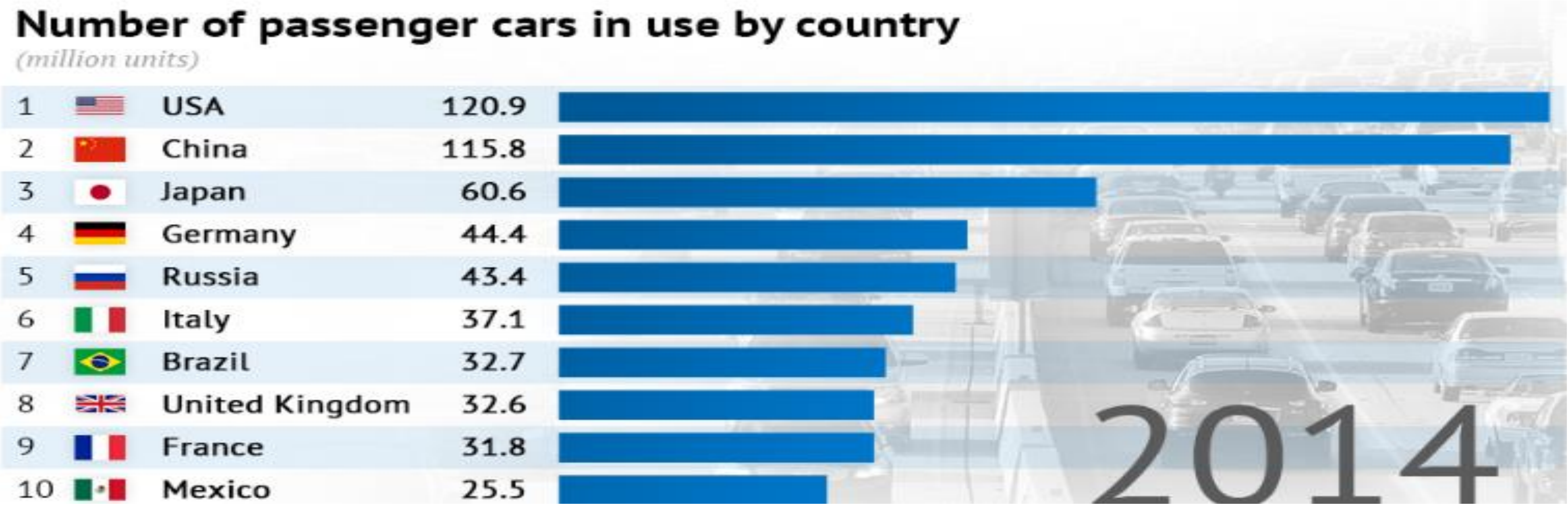
WALKABILITY
Walking distance from that address to a variety of key services. A higher walk score is a better walk score.

KEY POINTS IN HOUSING MARKET
Crime, Housing supply, Air Pollution, etc.

BENEFITS
Health, Economic, Social, and Environmental benefits.

It's Walkable

Background



- USA is primarily a car-based nation, impossible to survive if one doesn't live in big metro cities such as Philadelphia, NYC, Chicago.
- We wanted to see if having amenities such as grocery store, post office, etc. within a reasonable walking distance help drive up/down housing prices.

Research Question

An increase in neighborhood´s walkability will lead to an increase in house prices in absence of other factors.

Methodology

Dataset

- Publicly available dataset from Kaggle called “Philadelphia Real Estate”.
- Has data for real estates in Philadelphia.
- Has features like Zillow estimate(price), crime rates, school scores, walk scores, etc., for each real estate property.
- Dataset is 57KB in size.

kaggle

Philadelphia Real Estate

Sample dataset of Philadelphia real estate for analysis

```
Data columns (total 30 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Address              575 non-null    object
 1   Zillow Address       575 non-null    object
 2   Sale Date            575 non-null    object
 3   Opening Bid          575 non-null    float64
 4   Sale_Bid_Price       575 non-null    object
 5   Book/Writ            575 non-null    object
 6   OPA                  575 non-null    float64
 7   Postal Code          575 non-null    float64
 8   Attorney             575 non-null    object
 9   Ward                 575 non-null    float64
 10  Seller               575 non-null    object
 11  Buyer                575 non-null    object
 12  Sheriff_Cost         575 non-null    float64
 13  Advertising          575 non-null    float64
 14  Other                575 non-null    float64
 15  Record Deed          575 non-null    float64
 16  Water                575 non-null    float64
 17  PGW                  575 non-null    float64
 18  Avg_Walk_Score       575 non-null    float64
 19  Violent_Crime_Rate   575 non-null    float64
 20  School_Score         575 non-null    float64
 21  Zillow_Estimate      575 non-null    object
 22  Rent_Estimate        575 non-null    object
 23  Tax_Assessment       575 non-null    object
 24  Year_Built           575 non-null    float64
 25  SqFt                 575 non-null    float64
 26  Bathrooms            575 non-null    object
 27  Bedrooms             575 non-null    object
 28  PropType             575 non-null    object
 29  Average comps        575 non-null    object
```

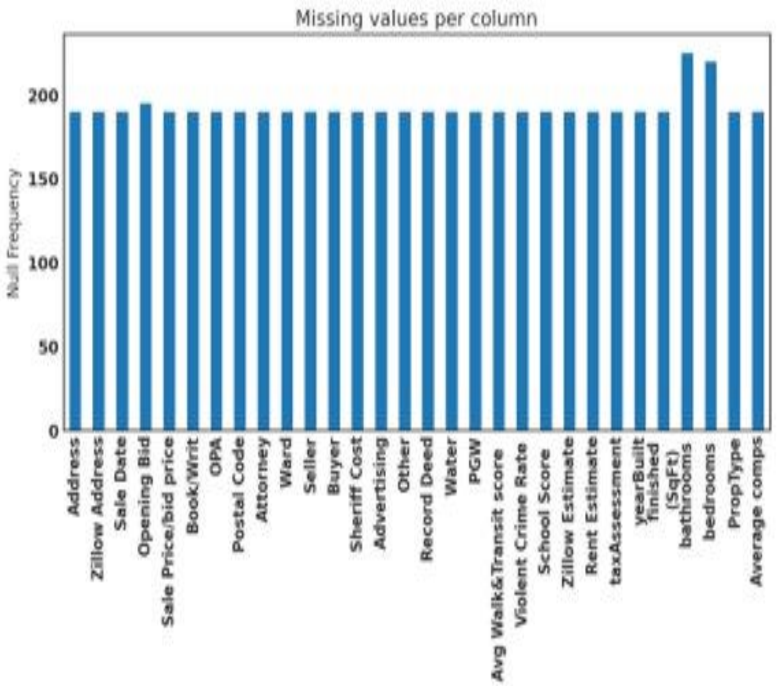
It's Walkable

Methodology

Data Cleaning

```
Data columns (total 30 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Address                615 non-null    object
 1   Zillow Address         615 non-null    object
 2   Sale Date              615 non-null    object
 3   Opening Bid            610 non-null    float64
 4   Sale Price/bid price   615 non-null    object
 5   Book/Writ              615 non-null    object
 6   OPA                    615 non-null    float64
 7   Postal Code            615 non-null    float64
 8   Attorney               615 non-null    object
 9   Ward                   615 non-null    float64
 10  Seller                 615 non-null    object
 11  Buyer                  615 non-null    object
 12  Sheriff Cost           615 non-null    float64
 13  Advertising            615 non-null    float64
 14  Other                  615 non-null    float64
 15  Record Deed            615 non-null    float64
 16  Water                  615 non-null    float64
 17  PGW                    615 non-null    float64
 18   Avg Walk&Transit score  615 non-null  float64
 19   Violent Crime Rate     615 non-null   float64
 20   School Score           615 non-null   float64
 21  Zillow Estimate        615 non-null    object
 22  Rent Estimate          615 non-null    object
 23  taxAssessment          615 non-null    object
 24  yearBuilt              615 non-null    float64
 25  finished
(SqFt)          615 non-null     float64
 26   bathrooms             615 non-null    object
 27   bedrooms              615 non-null    object
 28  PropType               615 non-null    object
 29  Average comps          615 non-null    object
```

Checked datatype of all variables



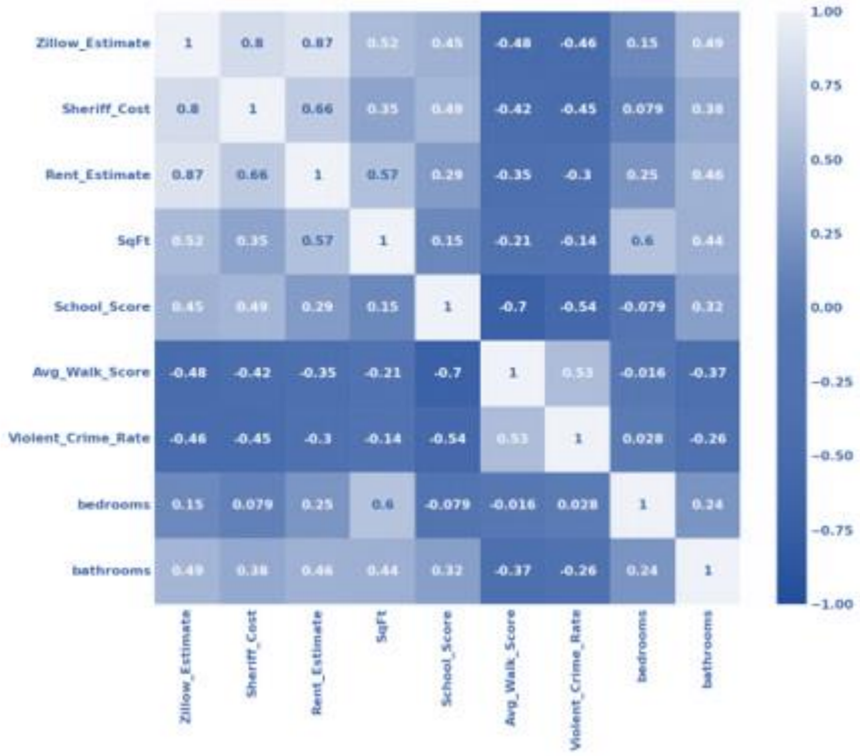
Checked for missing values

	Zillow Estimate	Rent Estimate
0	228,663.00	1,500.00
1	141,579.00	1,200.00
2	186,172.00	1,500.00
3	168,557.00	1,400.00
4	135,045.00	1,350.00
5	133,871.00	1,200.00
6	192,442.00	1,350.00
7	155,873.00	1,500.00
8	142,309.00	1,150.00
9	96,713.00	1,050.00
10	103,671.00	1,150.00
11	130,241.00	1,100.00
12	72,893.00	1,100.00
13	109,067.00	1,100.00
14	116,843.00	1,200.00
15	86,070.00	1,000.00
16	83,138.00	1,300.00
17	108,167.00	1,030.00
18	77,548.00	1,100.00
19	126,130.00	1,400.00
21	155,795.00	1,500.00
22	54,047.00	1,000.00
23	71,843.00	1,050.00
24	73,195.00	1,250.00
25	100,904.00	1,150.00

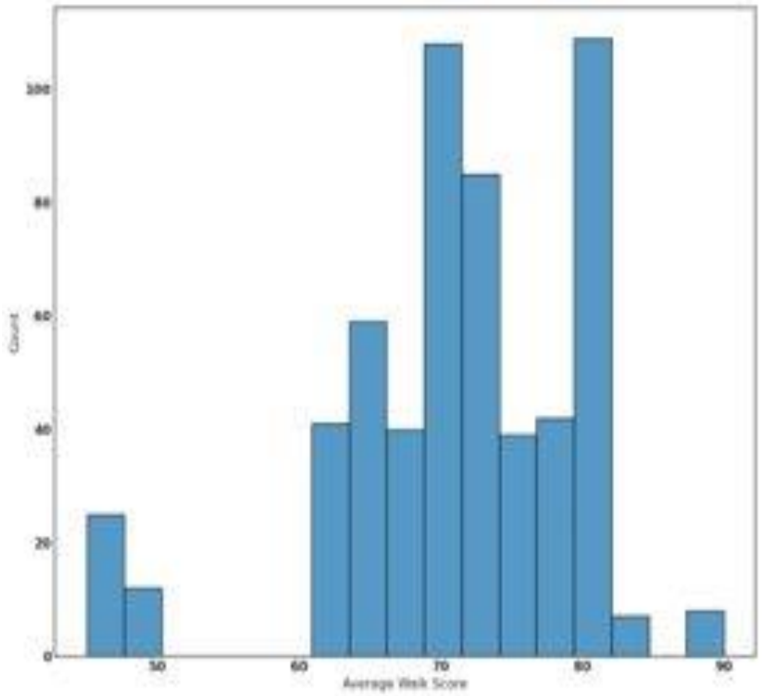
Removed invalid characters

Methodology

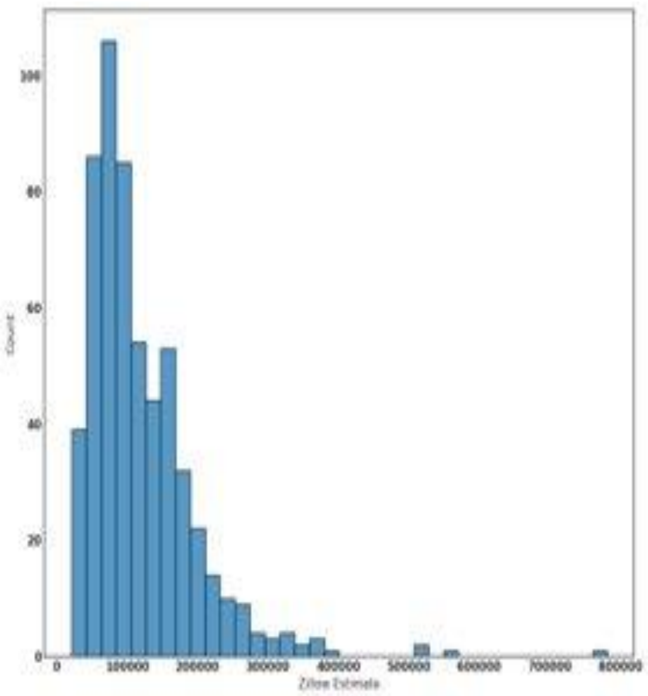
Data Visualization



Heat Map



Distribution of independent variable



Distribution of dependent variable

Methodology

Feature Engineering

Binning –
Average Walk Score, Violent Crime Rate

```
data['Walk_Class'] = pd.qcut(data['Avg_Walk_Score'],q=5,labels=['Car Dependent','Somewhat Car Dependent','Somewhat Walkable','Walkable','Walkers Paradise'])

data['Crime_Class'] = pd.qcut(data['Violent_Crime_Rate'],q=4, labels=['Low crime','Medium crime','High crime','Extreme crime'])
```

One Hot Encoding –
Average Walk Score, Violent Crime Rate

Crime_Class_Low crime	Crime_Class_Medium crime	Crime_Class_High crime	Crime_Class_Extreme crime
0	0	0	1
1	0	0	0
0	0	1	0
0	0	1	0
0	0	1	0

MinMax Scaling of Data–
Normalized the data

```
minmax = MinMaxScaler()
minmax.fit(X_train)
X_train_minmax=minmax.transform(X_train)
X_test_minmax=minmax.transform(X_test)
```

It's Walkable

Methodology

Statistical Analysis

OLS Regression Results

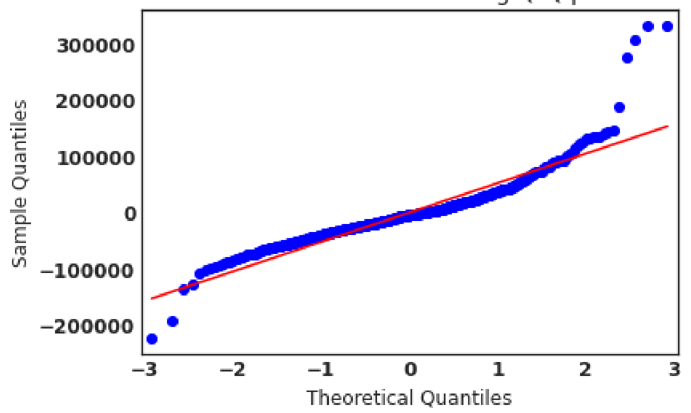
Dep. Variable:	Zillow_Estimate	R-squared:	0.550
Model:	OLS	Adj. R-squared:	0.538
Method:	Least Squares	F-statistic:	48.83
Date:	Mon, 21 Nov 2022	Prob (F-statistic):	1.92e-87
Time:	06:13:03	Log-Likelihood:	-7064.0
No. Observations:	575	AIC:	1.416e+04
Df Residuals:	560	BIC:	1.422e+04
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.425e+04	1.52e+04	3.572	0.000	2.44e+04	8.41e+04
C(Walk_Class)[T.Somewhat Car Dependent]	-1120.6954	9960.661	-0.113	0.910	-2.07e+04	1.84e+04
C(Walk_Class)[T.Somewhat Walkable]	-1.83e+04	1.36e+04	-1.350	0.178	-4.49e+04	8324.746
C(Walk_Class)[T.Walkable]	-3.338e+04	1.42e+04	-2.353	0.019	-6.12e+04	-5521.323
C(Walk_Class)[T.Walkers Paradise]	-1.028e+04	1.09e+04	-0.946	0.345	-3.16e+04	1.11e+04
C(Crime_Class)[T.Medium crime]	-2.42e+04	8333.481	-2.904	0.004	-4.06e+04	-7830.472
C(Crime_Class)[T.High crime]	-3.058e+04	1.2e+04	-2.555	0.011	-5.41e+04	-7074.725
C(Crime_Class)[T.Extreme crime]	-3.638e+04	1.14e+04	-3.204	0.001	-5.87e+04	-1.41e+04
C(PropType)[T.MultiFamily2To4]	-8.658e+04	2.53e+04	-3.424	0.001	-1.36e+05	-3.69e+04
C(PropType)[T.SingleFamily]	1.14e+04	6449.677	1.768	0.078	-1266.802	2.41e+04
C(PropType)[T.Townhouse]	713.4828	5504.375	0.130	0.897	-1.01e+04	1.15e+04
SqFt	88.3989	7.624	11.595	0.000	73.424	103.374
Bathrooms	2.248e+04	3933.623	5.714	0.000	1.48e+04	3.02e+04
Bedrooms	-1.946e+04	4354.235	-4.469	0.000	-2.8e+04	-1.09e+04
School_Score	884.9719	266.184	3.325	0.001	362.131	1407.813

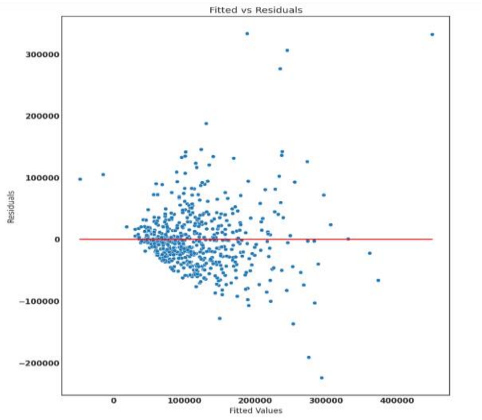
Omnibus:	245.031	Durbin-Watson:	1.947
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2242.030
Skew:	1.632	Prob(JB):	0.00
Kurtosis:	12.106	Cond. No.	1.64e+04

Normality

Residuals distribution using Q-Q plots



Linearity



Homoscedasticity

Goldfeld Test or Beusch wagon Test

```
[ ] import statsmodels.stats.api as sms
    sms.het_goldfeldquandt(reg.resid,reg.model.exog)

(1.4877164996587802, 0.0005462695282261776, 'increasing')
```

Multicollinearity

	Feature	VIF
0	Zillow_Estimate	11.613353
1	SqFt	25.344044
2	Bathrooms	7.119481
3	Bedrooms	21.152648
4	School_Score	3.325857
5	Sheriff_Cost	15.168404

- Used OLS Regression to understand statistical significance of various features.
- Statistical Analysis suggested use of Lasso Regression as a machine learning model.

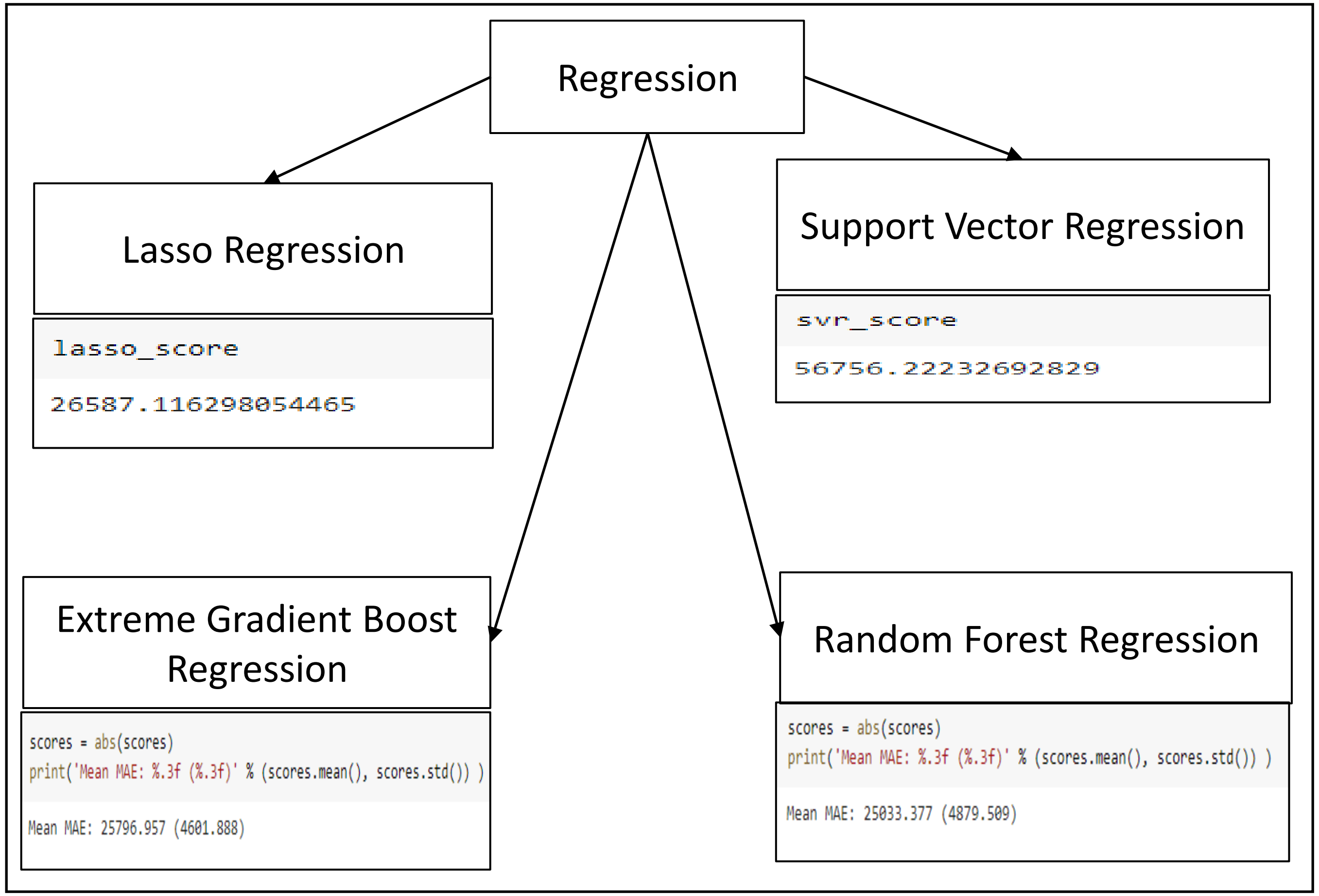
- Performed various tests to check if linearity assumptions hold.
- All assumptions except multicollinearity assumption hold true for the used data.

It's Walkable

Methodology

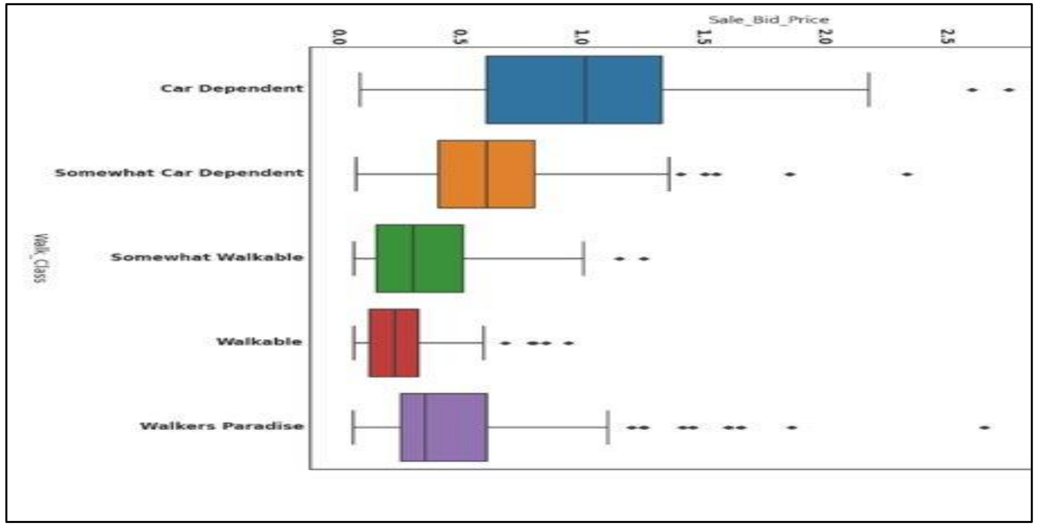
Machine Learning Model

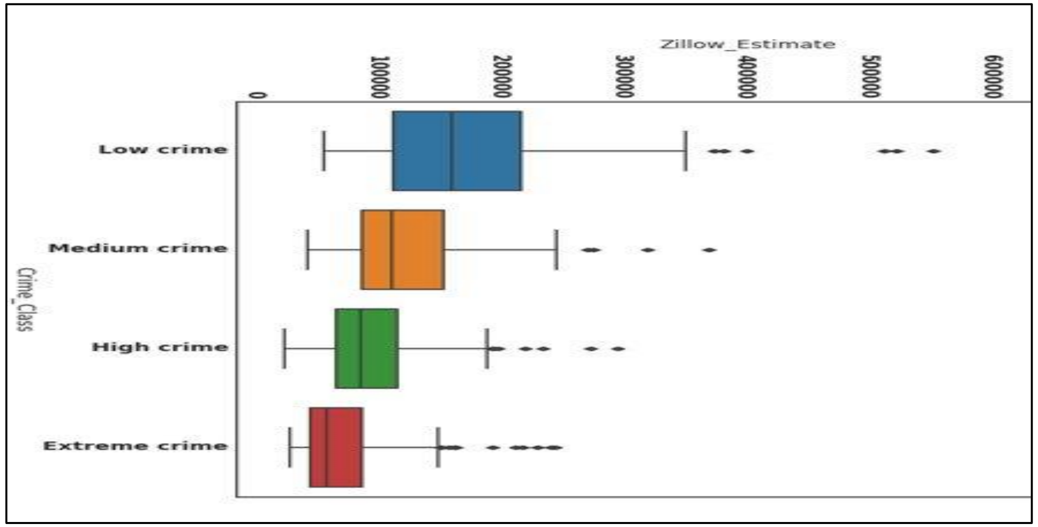
- Compared multiple regression algorithms using Mean Absolute Error as criteria.
- Used K-fold Grid Search to fine tune parameters (k=5)
- Our best model is random forest regression (max depth = 200) with a MAE of 20991 on our testing data.



Findings

- Walk Score is a statistically significant variable in the regression analysis.
- Overall, there is a negative correlation between walk score and house prices.
- For walk class with highest walk scores (walkers paradise), house prices are seen to be high.
- Negative correlation observed between crime rates in the region and house prices.
- Reject the hypothesis - "An increase in neighborhood´s walkability will lead to an increase in house prices in absence of other factors."



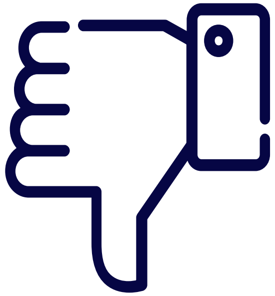


Interesting Finding!

Give preference to living in walkable areas

Live in safe crime-free areas

Recommendations



Would not recommend using model to start a real estate business. Error is around 20% of the average home price.

Look deeper into the negative relationship between walk score and other features

Gather data from other large cities and compare results.

References

- https://www.americantrails.org/resources/walking-the-walk-how-walkability-raises-home-values-in-u-s-cities
- https://nacto.org/docs/usdg/walking_the_walk_cortright.pdf
- https://conservancy.umn.edu/bitstream/handle/11299/187840/JTLU_vol10no1_pp241-261.pdf?sequence=1
- https://uppereastriver.com/why-walkability-is-so-important-for-property-investments-and-how-to-measure-that-walkability/
- https://www.forbes.com/sites/axiometrics/2016/02/15/high-walkability-may-mean-higher-rent/?sh=5268a4f94536
- https://scholarsbank.uoregon.edu/xmlui/bitstream/handle/1794/10386/SustDataAnalysis_ReportOpt.pdf?sequence=1
- https://www.redfin.com/news/how-much-does-walkability-increase-home-values/
- http://www.u.arizona.edu/~gpivo/Walkability%20Paper%20February%2010.pdf
- https://www.mdpi.com/2071-1050/12/2/593
- https://www.mehrnazamiri.com/project/2020-11-18-walkability/
- https://www.loopnet.com/learn/understanding-your-propertys-walk-score-/1309636409/
- https://www.jstor.org/stable/24860580#metadata_info_tab_contents.

It's Walkable

Thank You