ARTIFICIAL INTELLIGENCE

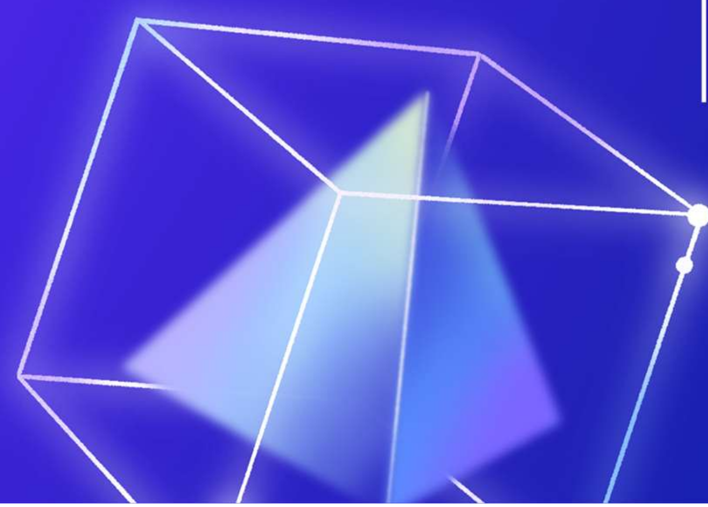
Predicting House Prices

BY: LIAM, VLAD, SEBASTIAN, FABIAN

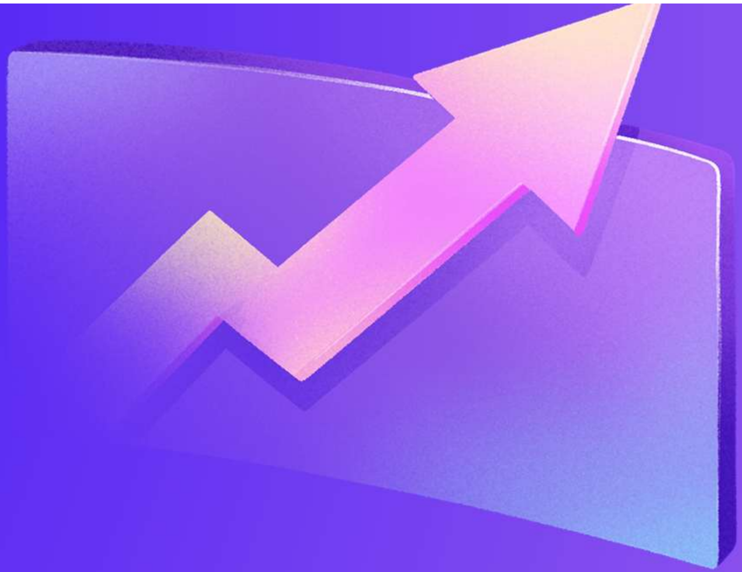
Calculate the Adjusted R^2

TABLE OF CONTENTS

- Problem Identification 0 1
- Data Collection and Preparation 02
- Exploratory Data Analysis (EDA) 03
- Model Selection and Design 04
- Model Training and Evaluation 05
- Conclusion and Ethics 06



Problem Identification



1.) Problem Definition:

- Current State: Real estate transactions pose challenges in accurately estimating house prices, impacting decision-making for buyers and sellers.
- Ideal State with AI: Create an AI solution for predicting house prices, providing users with an accurate and accessible tool for streamlined decision-making in real estate transactions.

2.) Target Audience and Stakeholders:

- Target Audience: Prospective homebuyers and sellers in King County, USA.
- Stakeholders: Real estate agents, property appraisers, and financial institutions.

Problem Identification

3.) Scope and Constraints:

- **Scope:** Focus on house sales in King County, USA. Utilize regression techniques considering key features (location, size, amenities, market trends).
- **Constraints:** Dataset availability and quality considerations. Ethical handling of data.

4.) Desired Outcome and Success Criteria:

- **Desired Outcome:** Develop an accurate and user-friendly AI model for predicting house prices.

Data Collection and Preparation

Data Set

- This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015.

- Source:
  - Kaggle
  - Kingcounty Web Site

- There is no mention of the data collection method, but in our case we only visited publicly available databases.

- The database contains the location, price year, # of bathrooms, # of bedrooms, # of floors etc..

kc_house_data.csv (2.52 MB)

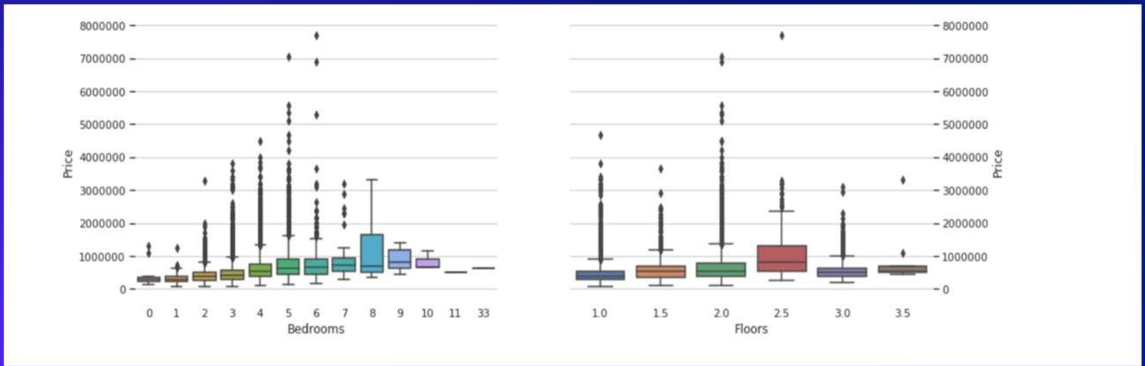
Detail Compact Column

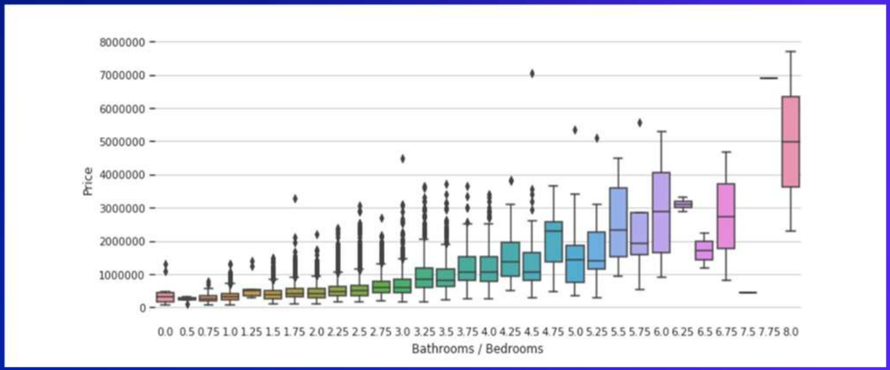
10 of 21

id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view
1.00m 9.90b	372 unique values	75k 7.7m	0 33	0 8	290 13.5k	520 1.65m	1 3.5	0 1	0 4
7129300520	20141013T000000	221900	3	1	1180	5650	1	0	0
6414100192	20141209T000000	538000	3	2.25	2570	7242	2	0	0
5631500400	20150225T000000	180000	2	1	770	10000	1	0	0
2487200875	20141209T000000	604000	4	3	1960	5000	1	0	0
1954400510	20150218T000000	510000	3	2	1680	8080	1	0	0
7237550310	20140512T000000	1.225e+006	4	4.5	5420	101930	1	0	0
1321400060	20140627T000000	257500	3	2.25	1715	6819	2	0	0
2008000270	20150115T000000	291850	3	1.5	1060	9711	1	0	0
2414600126	20150415T000000	229500	3	1	1780	7470	1	0	0
3793500160	20150312T000000	323000	3	2.5	1890	6560	2	0	0
1736800520	20150403T000000	662500	3	2.5	3560	9796	1	0	0
9212900260	20140527T000000	468000	2	1	1160	6000	1	0	0
0114101516	20140528T000000	310000	3	1	1430	19901	1.5	0	0
6054650070	20141007T000000	400000	3	1.75	1370	9680	1	0	0
1175000570	20150312T000000	530000	5	2	1810	4850	1.5	0	0
9297300055	20150124T000000	650000	4	3	2950	5000	2	0	3
1875500060	20140731T000000	395000	3	2	1890	14040	2	0	0
6865200140	20140529T000000	485000	4	1	1600	4300	1.5	0	0
0016000397	20141205T000000	189000	2	1	1200	9850	1	0	0
7983200060	20150424T000000	230000	3	1	1250	9774	1	0	0
6300500875	20140514T000000	385000	4	1.75	1620	4980	1	0	0

Exploratory Data Analysis (EDA)

Descriptive statistical analysis plays a fundamental role in understanding and summarizing the main features of our dataset. In this case analyzing the correlation of different aspects for buying a house like: price, location, number of rooms or even bathrooms.
This can help to recognize the tendency, dispersion, and shape of the data.





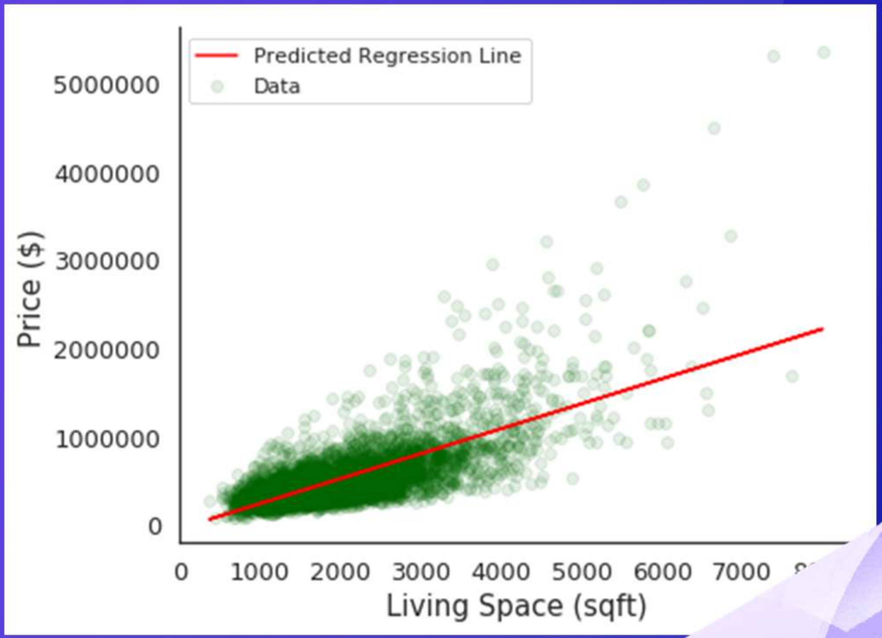
Drawing charts and examining the data before applying a model is a very good practice because we may detect some possible outliers or decide to do normalization. This is not a must but get know the data is always good.

Outliers: a person or thing differing from all other members of a particular group or set.

Model Selection and Design

<u>The appropriate AI model</u>

- **Linear Regression:** Suitable for predicting a continuous target variable, like house prices.
- In the dataset we chose, the living area (soft) appeared to be the most crucial factor with relation to price.

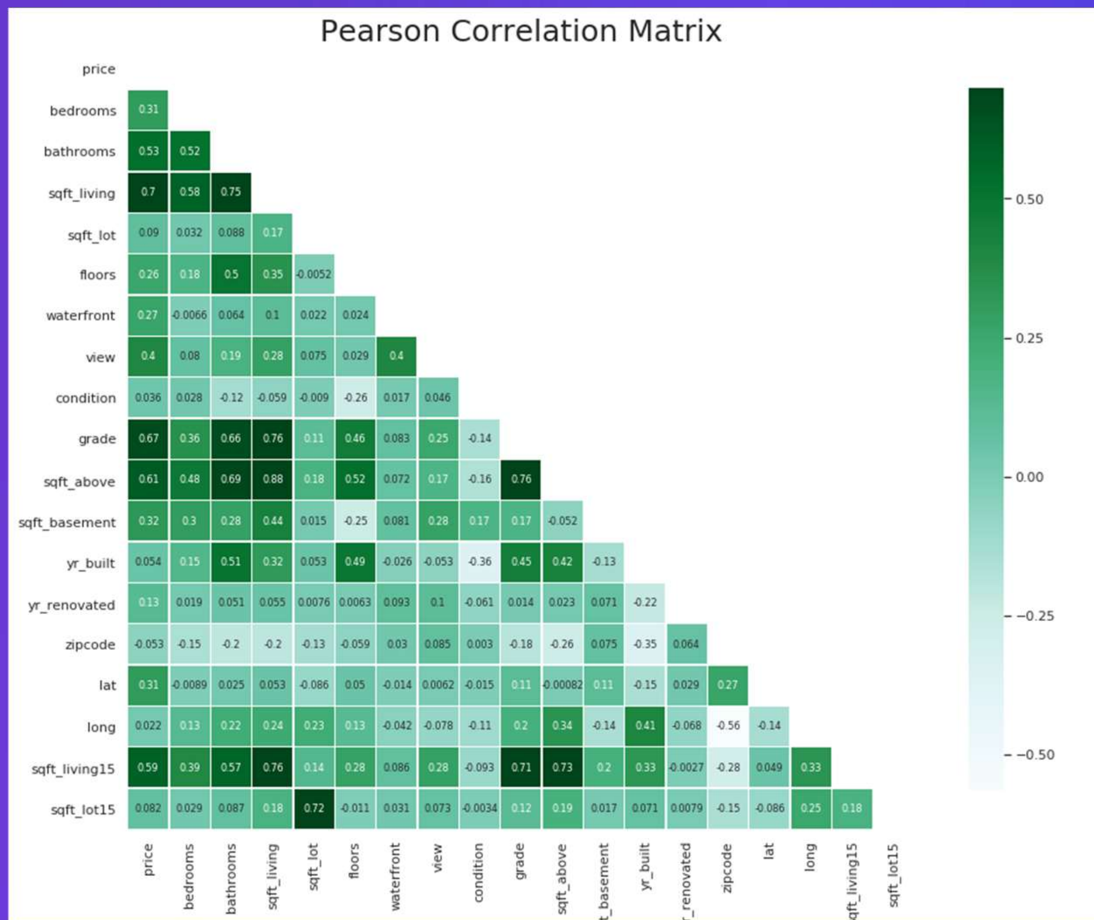


<u>Architecture and parameters of the chosen model</u>

- To predict house prices the variable has to be the price.
- In the example we chose, they decided to use living area (soft) as a feature to examine the relationship between price and living area.

Nevertheless, we could have used other features, such as the number of bedrooms and bathrooms, or even the location, to establish this relationship.

Model Selection and Design



Model Selection and Design

This data frame includes Root Mean Squared Error (RMSE), R-squared, Adjusted R-squared and mean of the R-squared values obtained by the k-Fold Cross Validation, which are the important metrics to compare different models. Having a R-squared value closer to one and smaller RMSE means a better fit.

$$\bar{R}^2 = R^2 - \frac{k-1}{n-k}(1 - R^2)$$

Model Selection and Design

```python
#%%capture
train_data,test_data = train_test_split(df,train_size = 0.8,random_state=3)

lr = linear_model.LinearRegression()
X_train = np.array(train_data['sqft_living'], dtype=pd.Series).reshape(-1,1)
y_train = np.array(train_data['price'], dtype=pd.Series)
lr.fit(X_train,y_train)

X_test = np.array(test_data['sqft_living'], dtype=pd.Series).reshape(-1,1)
y_test = np.array(test_data['price'], dtype=pd.Series)

pred = lr.predict(X_test)
rmsesm = float(format(np.sqrt(metrics.mean_squared_error(y_test,pred)),'.3f'))
rtrsm = float(format(lr.score(X_train, y_train),'.3f'))
rtesm = float(format(lr.score(X_test, y_test),'.3f'))
cv = float(format(cross_val_score(lr,df[['sqft_living']],df['price'],cv=5).mean(),'.3f'))

print ("Average Price for Test Data: {:.3f}".format(y_test.mean()))
print('Intercept: {}'.format(lr.intercept_))
print('Coefficient: {}'.format(lr.coef_))

r = evaluation.shape[0]
evaluation.loc[r] = ['Simple Linear Regression','-',rmsesm,rtrsm,'-',rtesm,'-',cv]
evaluation
```

```
Average Price for Test Data: 539744.130
Intercept: -47235.811302901246
Coefficient: [282.2468152]
```

	Model	Details	Root Mean Squared Error (RMSE)	R-squared (training)	Adjusted R-squared (training)	R-squared (test)	Adjusted R-squared (test)	5-Fold Cross Validation
0	Simple Linear Regression	-	254289.149	0.492	-	0.496	-	0.491

DATA PROCESSING

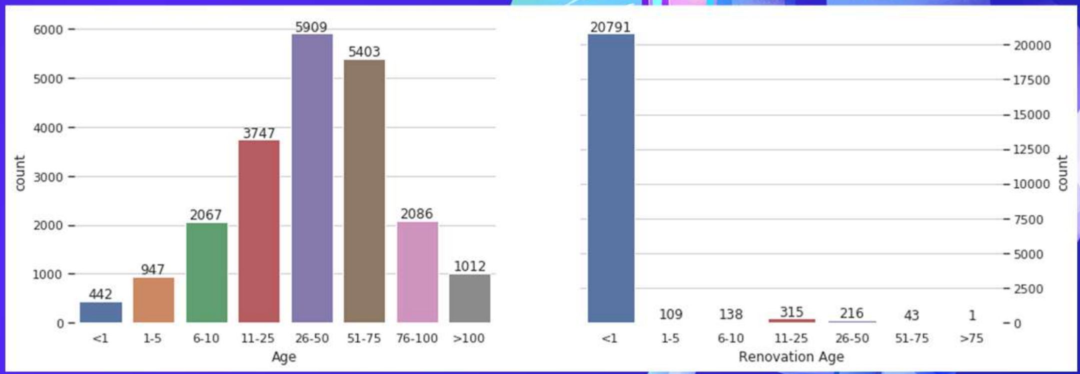
Data Preprocessing Importance:
- Modifying data before modeling enhances prediction accuracy and reliability.

Selection of "Binning" Method:
- Chose the "binning" method as the preferred approach for data transformation.

Age and Renovation Age Calculation:
- Calculated ages and renovation ages of houses at the time of sale.

Column Division into Intervals:
- The calculated ages and renovation ages were divided into intervals or ranges.

Presentation through Histograms:
- The results of the binning process are visually represented in histograms.



Primary Goal:
- The primary objective is to enhance outcomes by meticulous attention to data

details and the strategic use of processing methods.

DATA PROCESSING

- Quadratic Distribution:
  - When dealing with data exhibiting a quadratic distribution, opting for a quadratic function and employing a polynomial transformation can improve accuracy.
- Hypothesis Function:
  - The hypothesis function for polynomial regression is expressed as $h_\theta(X) = \theta_0 + \theta_1 x + \theta_2 x^2 + \ldots + \theta_n x^n$
- Enhanced Model Fit:
  - The table demonstrates that implementing polynomial transformation significantly improves the model fit.
- Caution with Degree Selection:
  - Exercise caution when determining the degree of the polynomial transformation to avoid overfitting.
- Overfitting Warning:
  - Instances of overfitting are identified in certain models, where 5-fold cross-validation metrics are negative or low despite high R-squared values for the training set.

	Model	Details	Root Mean Squared Error (RMSE)	R-squared (training)	Adjusted R-squared (training)	R-squared (test)	Adjusted R-squared (test)	5-Fold Cross Validation
4	Multiple Regression-4	all features	191879.550	0.701	0.7	0.713	0.711	0.698
3	Multiple Regression-3	all features, no preprocessing	193693.989	0.698	0.697	0.708	0.707	0.695
2	Multiple Regression-2	selected features	209712.753	0.652	0.652	0.657	0.656	0.648
1	Multiple Regression-1	selected features	248514.011	0.514	0.514	0.519	0.518	0.512
0	Simple Linear Regression	-	254289.149	0.492	-	0.496	-	0.491

	Model	Details	Root Mean Squared Error (RMSE)	R-squared (training)	R-squared (test)	5-Fold Cross Validation
2	Polynomial Regression	degree=2, all features, no preprocessing	151200.970	0.830	0.822	0.813
6	Polynomial Ridge Regression	alpha=50000, degree=2, all features	159872.572	0.810	0.801	0.791
8	Polynomial Lasso Regression	alpha=50000, degree=2, all features	166020.484	0.797	0.785	0.779
7	Polynomial Lasso Regression	alpha=1, degree=2, all features	166195.984	0.807	0.785	0.778
0	Polynomial Regression	degree=2, selected features, no preprocessing	190980.547	0.730	0.716	0.714
1	Polynomial Regression	degree=3, selected features, no preprocessing	189235.269	0.749	0.721	0.595
3	Polynomial Regression	degree=3, all features, no preprocessing	186433.648	0.874	0.729	-0.927
5	Polynomial Ridge Regression	alpha=1, degree=2, all features	150177.258	0.838	0.824	-3168.943
4	Polynomial Regression	degree=2, all features	151654.993	0.840	0.821	-11230.411

- Enhancement Strategy:
  - To improve the model, the plan is to incorporate more features.
- Transition to Multiple Regression:
  - When more than one feature is used in a linear regression, it's termed as multiple regression.
- Complex Model Creation:
  - The introduction of multiple features marks the transition to more complex models.
- Multiple Regression - Feature Selection:
  - Features were determined initially by examining previous sections and were used in the first multiple linear regression.
- Prediction Definition:
  - Unlike simple regression, a specific definition is required for predictions in multiple regression, especially when conducting manual calculations.

ETHICAL IMPLICATIONS

HOW RELIABLE IS OUR PRODUCT?

NOT 100% ACCURATE

HOW SHOULD WE MARKET OUR SERVICE?

NOT FOOL PROOF

ETHICAL IMPLICATIONS

HOW TO DEAL WITH THIS ISSUE?

REQUIRE TERMS AND CONDITIONS OR CONTRACT SIGNATURE IN CASE USERS ARE DISSATISFIED OR FELT THEY WERE MISLED

ETHICAL IMPLICATIONS

CUSTOMER AWARENESS

INFORM CONSUMERS ON HOW TO USE IT

AND HOW WE ARE IMPROVING PROCESSING

PATTERNS AND PARAMETERS

CONCLUSION

THANK YOU!

RESOURCE PAGE

- https://www.kaggle.com/code/burhanykiyakoglu/predicting-house-prices
- https://www.kaggle.com/datasets/harlfoxem/housesalesprediction

House Sales in King County, USA (kaggle.com)