Boston Housing Market

By Kamal Jeter

Data Analytics – The What and The Why

What is Data Analytics?

- Data analytics is the collecting, cleaning, analyzing, and interpreting of raw data to uncover insights.

Why does it matter?

- Data analysis is apart of our daily lives.
- Utilities companies use data analysis to detect leaks and outages.
- If we were to receive an alien signal, it would be because of a data analyst!

Applying Data Analysis To The Boston Housing Market

- Objectives to apply data analysis

- Explore
- Prepare
- Visualize
- Model
- Summarize

Boston Housing Business Questions

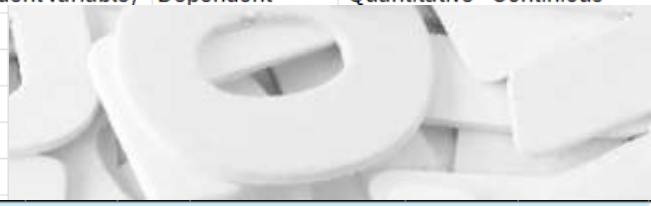
- Do the average number of rooms a house has influence the median price of homes?
- Is crime a significant factor that influences home prices?

Comprehensive Data Summary

- We begin with setting up a data dictionary and information sheet.
- These sheets are explain and share information like the variables we'll be including, the types of variables they are, the number of rows and columns in the dataset, and more.
- We include the summary statistics to help us understand what would be a large amount of data. Easier to see what to look for.

Variable ID	Variable Name	Variable Description	Variable Type: 1. Independent 2. Dependent	Variable Data Type 1. Qualitative 2. Quantitative - Discrete 3. - Quantitative - Continious
1	CRIM	per capita crime rate by town	Independent	Quantitative - Continious
2	ZN	proportion of residential land zoned for lots over 25,000 sq.ft	Independent	Quantitative - Continious
3	INDUS	proportion of non-retail business acres per town	Independent	Quantitative - Continious
4	CHAS	Charles River dummy variable (1 if tract bounds river; 0 otherwise)	Independent	Quantitative - Discrete
5	NOX	nitric oxides concentration (parts per 10 million)	Independent	Quantitative - Continious
6	RM	average number of rooms per dwelling	Independent	Quantitative - Continious
7	AGE	proportion of owner-occupied units built prior to 1940	Independent	Quantitative - Continious
8	DIS	weighted distances to five Boston employment centres	Independent	Quantitative - Continious
9	RAD	index of accessibility to radial highway	Independent	Quantitative - Discrete
10	TAX	full-value property-tax rate per $10,000	Independent	Quantitative - Discrete
11	PTRATIO	pupil-teacher ratio by town	Independent	Quantitative - Continious
12	LSTAT	% lower status of the population	Independent	Quantitative - Continious
13	MEDV	Median value of owner-occupied homes in 1000s (Your dependent variable)	Dependent	Quantitative - Continious

Dataset was received on October 9th, 2025
Data within the dataset was collected in the 1970s
The dataset was received by download on Brightspace
The format of the dataset was received by CSV
The size of the raw dataset is 31.1 KB



Dataset Completion

While there are no missing values or empty cells in this datas
there is a variable missing which represented the
proportion of black residents by town (B).

Dataset Coverage

Number of Rows: 506
Number of Columns: 13

Intended Use / Purpose

Purpose of Dataset: To analyze various attributes of
houses in different areas around Boston, Massachussets fro
the 1970s.

Summary Statistics	Mean	Median	Mode	Standard Deviation	Minimum	Maxiumum
CRIM	3.61	0.2565	0.02	8.593041351	0.00632	88.9762
ZN	11.4	0	0	23.29939569	0	100
INDUS	11.1	9.69	18.1	6.853570583	0.46	27.74
CHAS	0.07	0	0	0.253742935	0	1
NOX	0.55	0.538	0.54	0.115763115	0.385	0.871
RM	6.28	6.2085	5.71	0.701922514	3.561	8.78
AGE	68.6	77.5	100	28.12103257	2.9	100
DIS	3.8	3.2075	3.5	2.103628356	1.1296	12.1265
RAD	9.55	5	24	8.698651118	1	24
TAX	408	330	666	168.370495	187	711
PTRATIO	18.5	19.05	20.2	2.162805191	12.6	22
LSTAT	12.7	11.36	8.05	7.134001637	1.73	37.97
MEDV	22.5	21.2	50	9.188011545	5	50

All values computed from 506 observations

Data Cleaning

- Consistent data is crucial!
- Check for missing values, consistent data types, duplicates, and formatting.
- Use conditional formatting to highlight outliers.

Finding Outliers With Quartiles

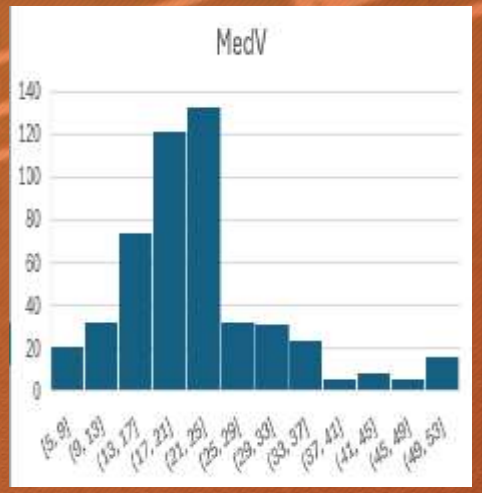
Q1	Q3	IQR	Lower Bound	U
0.082045	3.677083	3.595038	0	
0	12.5	12.5	0	
5.19	18.1	12.91	0	
0	0	0	0	
0.449	0.624	0.175	0.1865	
5.8855	6.6235	0.738	4.7785	
45.025	94.075	49.05	0	
2.100175	5.188425	3.08825	0	
4	24	20	0	
279	666	387	0	
17.4	20.2	2.8	13.2	
6.95	16.955	10.005	0	
17.025	25	7.975	5.0625	

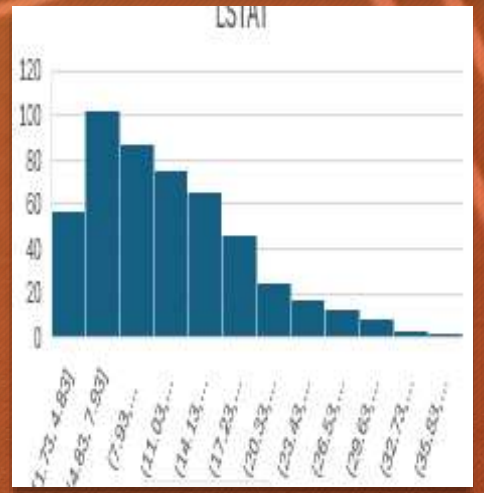
Notes

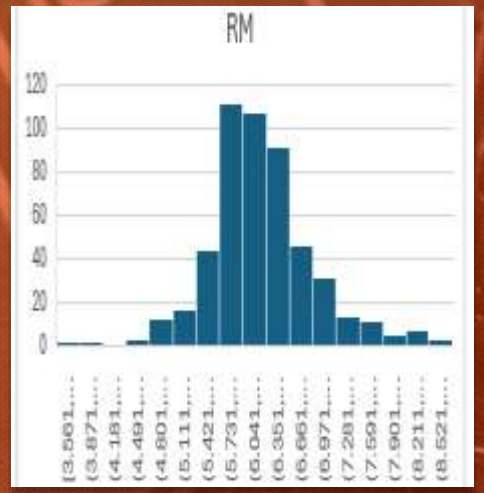
1.CHAS should be excluded because it's a binary and not continuous.
2.Some variables had lower bounds less than 0, but because these are all non negative variables,the MAX function was used to prevent non existent negative values.
3. There are no duplicates in the dataset
4. Outliers were confirmed using the IQR method and visually confirmed with conditional formatting. No outliers were deleted due to them being real data points showing real life variance, and are not considered errors.
5. Outliers for each variable were counted using the COUNTIFS function. The only variables that yielded outliers were MEDV, LSTAT, PTARATIO, RM, ZN, and CRIM. This may hint that some of these variables may have a significant influence on the Boston housing market.

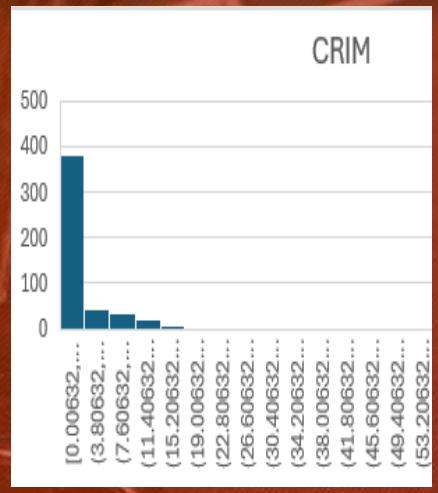
Number of ZN outliers	68
Number of INDUS outliers	0
CHAS is exculuded	excluded
Number of NOX outliers	0
Number of RM outliers	30
Number of AGE outliers	0
Number of DIS outliers	5
Number of RAD outliers	0
Number of Tax outliers	0

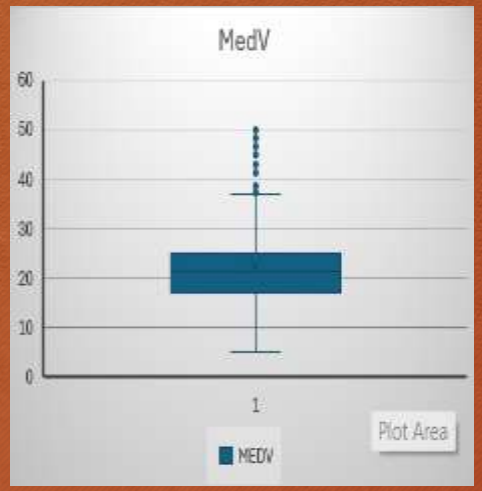
Univariate Analysis

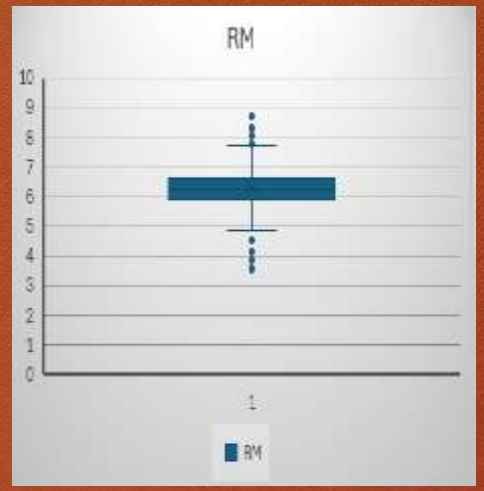


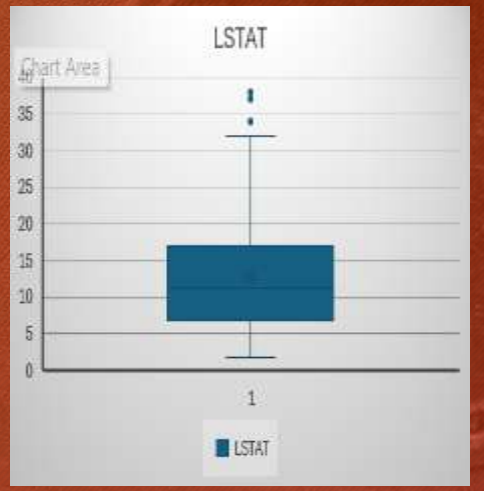


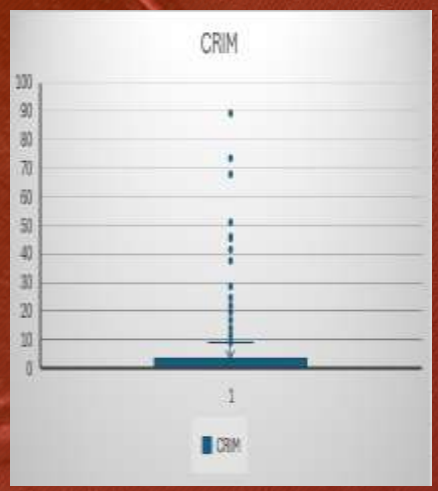












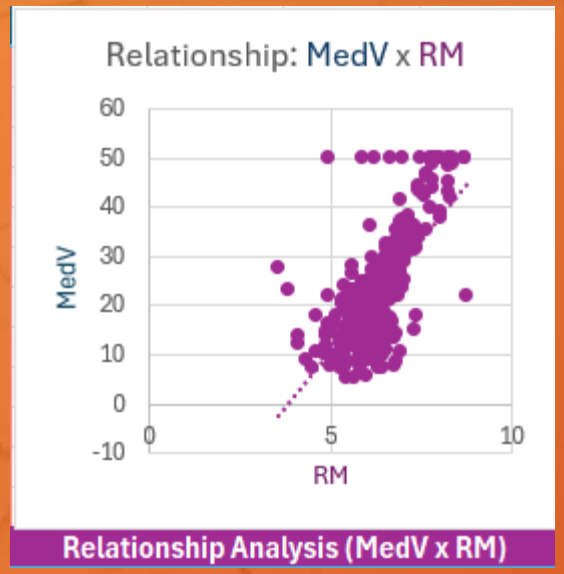
These charts are the result of univariate analysis.

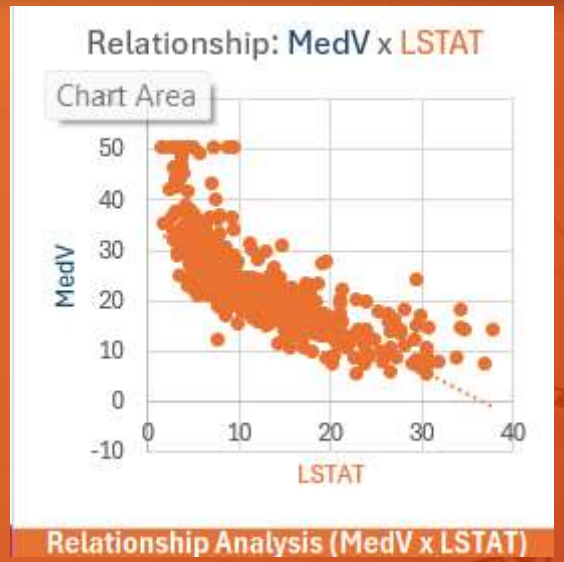
We have 4 variables, the top row being histograms, and the bottom row being box and whisker plots.

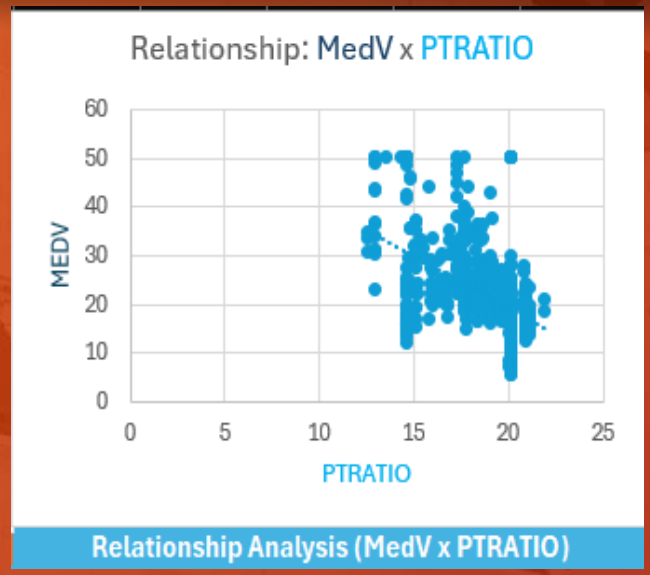
Histograms show skew, boxplots show outliers.

Bivariate Analysis and Conditional Formatting

RM Correlation	0.69536
LSTAT Correlation	-0.73766
PTRATIO Correlation	-0.50779







PTRATIO	LSTAT	MEDV
20.2	30.59	5
20.2	22.98	5
20.2	26.77	5.6
20.2	29.97	6.3
20.2	36.98	7
20.1	23.97	7
20.2	30.81	7.2
20.2	20.32	7.2
20.2	29.05	7.2
20.2	31.99	7.4
20.2	25.79	7.5
20.1	29.68	8.1
20.2	19.77	8.3
14.4	2.97	50
14.7	2.88	50
17.4	4.63	50
13	5.12	50
13	7.44	50
13.6	3.16	50
20.2	3.26	50
20.2	3.73	50
20.2	2.96	50
20.2	9.53	50
20.2	8.88	50

- Bivariate analysis looks at two variables to discover their relationship.
- Use of trendlines and CORREL function can make relationships clear.
- Conditional formatting can show the relationship between variables

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	LSTAT	MEDV
CRIM	1												
ZN	-0.20047	1											
INDUS	0.40658	-0.53383	1										
CHAS	-0.05589	-0.0427	0.06294	1									
NOX	0.42097	-0.5166	0.76365	0.0912	1								
RM	-0.21925	0.31199	-0.39168	0.09125	-0.30219	1							
AGE	0.35273	-0.56954	0.64478	0.08652	0.73147	-0.24026	1						
DIS	-0.37967	0.66441	-0.70803	-0.09918	-0.76923	0.20525	-0.74788	1					
RAD	0.62551	-0.31195	0.59513	-0.00737	0.61144	-0.20985	0.45602	-0.49459	1				
TAX	0.58276	-0.31456	0.72076	-0.03559	0.66802	-0.29205	0.50646	-0.53443	0.91023	1			
PTRATIO	0.28995	-0.39168	0.38325	-0.12152	0.18893	-0.3555	0.26152	-0.23247	0.46474	0.46085	1		
LSTAT	0.45562	-0.41299	0.6038	-0.05393	0.59088	-0.61381	0.60234	-0.497	0.48868	0.54399	0.37404	1	
MEDV	-0.3883	0.36045	-0.48373	0.17526	-0.42732	0.69536	-0.37695	0.24993	-0.38163	-0.46854	-0.50779	-0.73766	1

Correlation Matrix Helps With Insights

The correlation matrix shows the r value between any two values at a glance. From here we can see RM has the strongest positive correlation with MEDV, while LSTAT has the strongest negative correlation

At this point, one can conclude hypothesize that the Boston housing market is postvely influence by RM, and negatively influenced by PTRATIO and LSTAT

Simple Linear Regression

- What is regression in the first place?
- Regression is a method to understand the relationship between a dependent and independent vairiable
- We can learn how “strong” the relationship is, and the direction it points to.
- Y = b0 + b1x + e
- The difference between r and r^2

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.695359947
R Square	0.483525456
Adjusted R Square	0.482500705
Standard Error	6.61615975
Observations	506

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	20654.41622	20654.41622	471.84674	2.48723E-74
Residual	504	22061.8792	43.77356983		
Total	505	42716.29542			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-34.6706208	2.649802993	-13.0842258	6.9502E-34	-39.87664103	-29.4646005	-39.876641	-29.4646005
RM	9.102108981	0.41902656	21.72203351	2.4872E-74	8.27885504	9.925362923	8.27885504	9.925362923

Multiple Linear Regression

- Simple regression includes one independent variable, while multiple includes more than one.
- Y = b0 + b1x + b2x + b3x… +e
- We focus on adjusted r square value here because it takes into account the number of predictors and sample size
- - F statistic, coefficients, and significance
- - With a model like this we can move on to solving for our hypothesis

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.825628454
R Square	0.681662343
Adjusted R Square	0.678478967
Standard Error	5.21501781
Observations	506

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	29118.09004	5823.618007	214.131859	8.91E-122
Residual	500	13598.20538	27.19641076		
Total	505	42716.29542			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	17.46891477	4.116286743	4.243852739	2.6193E-05	9.381564549	25.55626499	9.381564549	25.55626499
CRIM	-0.06198012	0.031548912	-1.96457241	0.05001713	-0.123964894	4.65102E-06	-0.12396489	4.65102E-06
NOX	-1.31901846	2.552647452	-0.51672567	0.60557615	-6.334255545	3.696218629	-6.33425555	3.696218629
RM	4.63324063	0.428409886	10.81497133	1.2227E-24	3.79153523	5.474946031	3.79153523	5.474946031
PTRATIO	-0.89353765	0.11914528	-7.49956397	2.9426E-13	-1.127624743	-0.65945055	-1.12762474	-0.65945055
LSTAT	-0.52224493	0.051294261	-10.1813521	2.9865E-22	-0.623023779	-0.42146608	-0.62302378	-0.42146608

Hypothesis

- A null hypothesis is saying an independent variable has no statistically significant relationship with the dependent variable.

- An alternative hypothesis says there is a statistically significant relationship

- For my hypothesis I chose RM and CRIM.

- - Next step is to check the p-values for those predictors in the model. Any value > or = 0.05 is not statistically significant.

Statistical Decisions

- P-value (RM) = 1.22E-24

This model rejected the null hypothesis, indicating that there's a statistically significant relationship between RM and MedV.

- P-value (CRIM) = 0.05

This model failed to reject the null hypothesis, showing there is no statistically significant relationship between CRIM and MedV

- These results suggest that home values in the Boston area are strongly driven by a homes RM, PTRATIO, and LSTAT. While the crime rate appears related, it may not be that strong of a factor once other elements are considered.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.825628454
R Square	0.681662343
Adjusted R Square	0.678478967
Standard Error	5.21501781
Observations	506

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	29118.09004	5823.618007	214.131859	8.9144E-122
Residual	500	13598.20538	27.19641076		
Total	505	42716.29542			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	17.46891477	4.116286743	4.243852739	2.6193E-05	9.381564549	25.55626499	9.381564549	25.55626499
CRIM	-0.06198012	0.031548912	-1.96457241	0.05001713	-0.123964894	4.65102E-06	-0.12396489	4.65102E-06
NOX	-1.31901846	2.552647452	-0.51672567	0.60557615	-6.334255545	3.696218629	-6.33425555	3.696218629
RM	4.63324063	0.428409886	10.81497133	1.2227E-24	3.79153523	5.474946031	3.79153523	5.474946031
PTRATIO	-0.89353765	0.11914528	-7.49956397	2.9426E-13	-1.127624743	-0.65945055	-1.12762474	-0.659450554
LSTAT	-0.52224493	0.051294261	-10.1813521	2.9865E-22	-0.623023779	-0.42146608	-0.62302378	-0.421466076

Summary of Findings

- *From Data Preparation*
  -There were 231 outliers out of 6,578 data points (3.51%)
  -No missing values

- *From Data Visualization*
  - MedV, CRIM, and LSTAT all had a right skewed distribution, while RM maintained a normal distribution. Most neighborhoods are normal, a few have very high house prices, crime, and lower income populations.
- *From Regression Analysis*
  -Simple linear regression between RM and showed positive relationship
  -Multiple regression model displayed PTRATIO, LSTAT, and RM were the only statistically significant predictors in this model.

  - This would lead one to believe that factors like the average income, or the attention the average student receives are stronger influences on homes in Boston rather than crime, or environmental variables.

Actionable Recommendations

- *For City Planners and Policy Makers*
  -Investing in smaller class sizes, tutoring, and free programs that gurantee employment could improve housing prices over time.
- *For Real Estate Investors*
  -Agents should focus on advertising multi room housing for higher income clients. Investors should also be mindul of demographics, such as low-income population of the town, and pupil to teacher ratio of the schools in the area.
- *For Homebuyers*
  -Homebuyers looking to invest should consider towns where the schools in the area put an emphasis on attention to the student, and towns where lower-income residents are evidently escaping poverty overtime
- *For Social Programs and Equity Planning*
  -City programs that support low-income residents, such as workforce development, can help balance these imbalances and be a great benefit to the housing market.

“Data is the new oil. It’s valuable, but if unrefined, it cannot really be used.”

Clive Humby

Whether it’s housing markets, customer behavior, or discovering alien signals, data analysis can make a world of impact, and I hope this project gave you a glimpse as to how

Thank you!