


SALES PRICE FOR PROPERTIES IN



DATA ANALYTICS | Karim
November, 2022

- 
- Note: My coworker ([Ahmed Gaber](#)) and I decided to solve the same problem, but we each have our own way of cleaning, and we also have different research questions.

GOAL

- For this project, I'm going to build a machine learning model that could predict the property sales price for properties in NYC. I'm dividing my project into the following sections:
 - EDA
 - Data Preparation
 - Modelling

EDA, FEATURE ENGINEERING, AND AUTOML

- We will apply learned concepts from Day 3-4 lectures to perform 'data exploration', 'feature engineering', and 'autoML' on a 'House Sale' dataset.
 - The goal of this assignment is to analyze 3 years' "(2018-2020) House Sales" data provided by "New York City" (NYC) government and build regression model to predict house price.
 - NYC has five boroughs, i.e., "Bronx", "Brooklyn", "Manhattan", "Queens" and "Staten Island".
 - Sales of houses in each borough has been provided.

EXPLORATORY DATA ANALYSIS (EDA)

- The EDA is divided in the following steps:
 - Basic data exploration
 - Identification of variables and data types
 - Missing values
 - Univariate Analysis
 - Bivariate Analysis
 - Correlation matrix
 - Conclusion

Note :

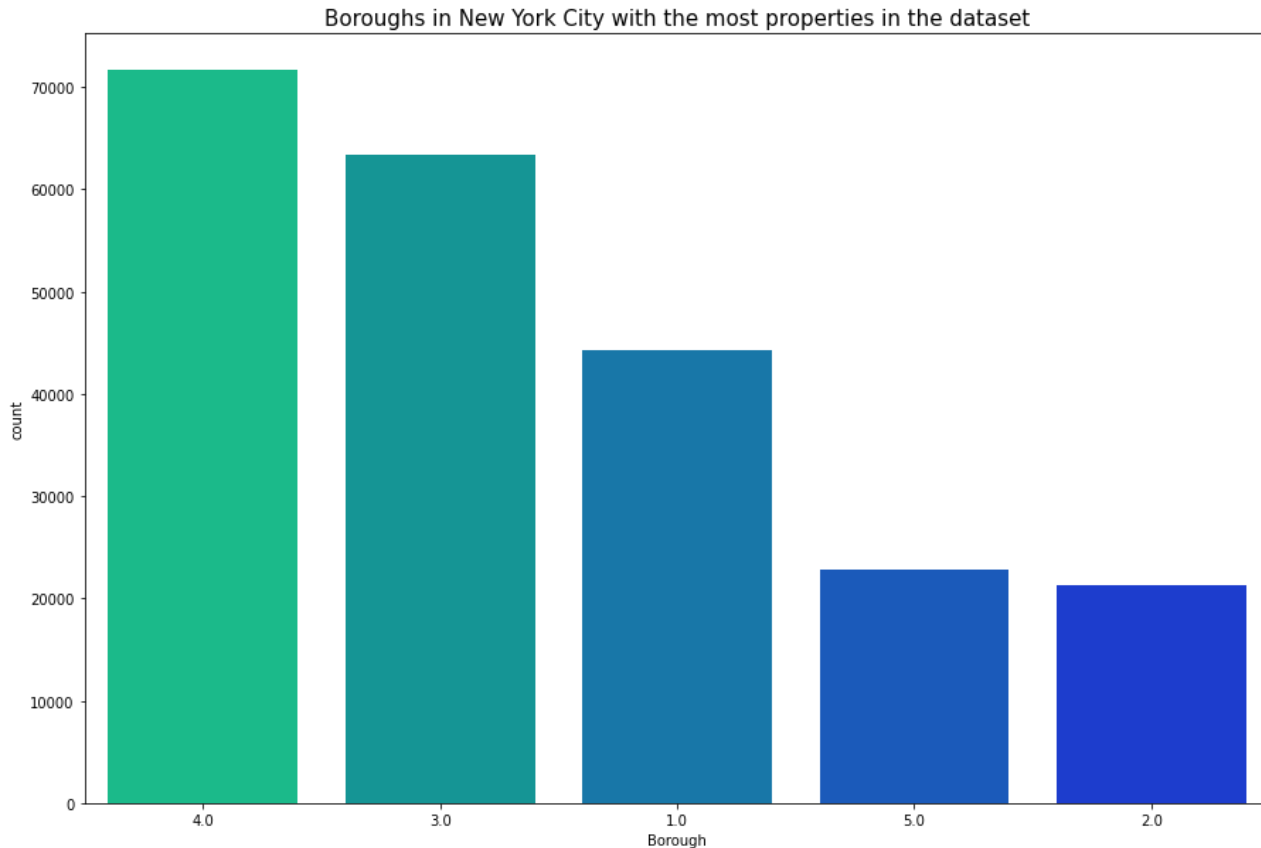
- The data size before EDA was 277,803 rows x 21 columns.
- And after EDA it was 223,426 rows x 19 columns.



RESEARCH QUESTIONS:

- Hypothesis: Manhattan is the most densely populated of New York City's 5 boroughs source
- What is the most real estate properties in New York City for each Boroughs ?

RESEARCH QUESTIONS CONT.:

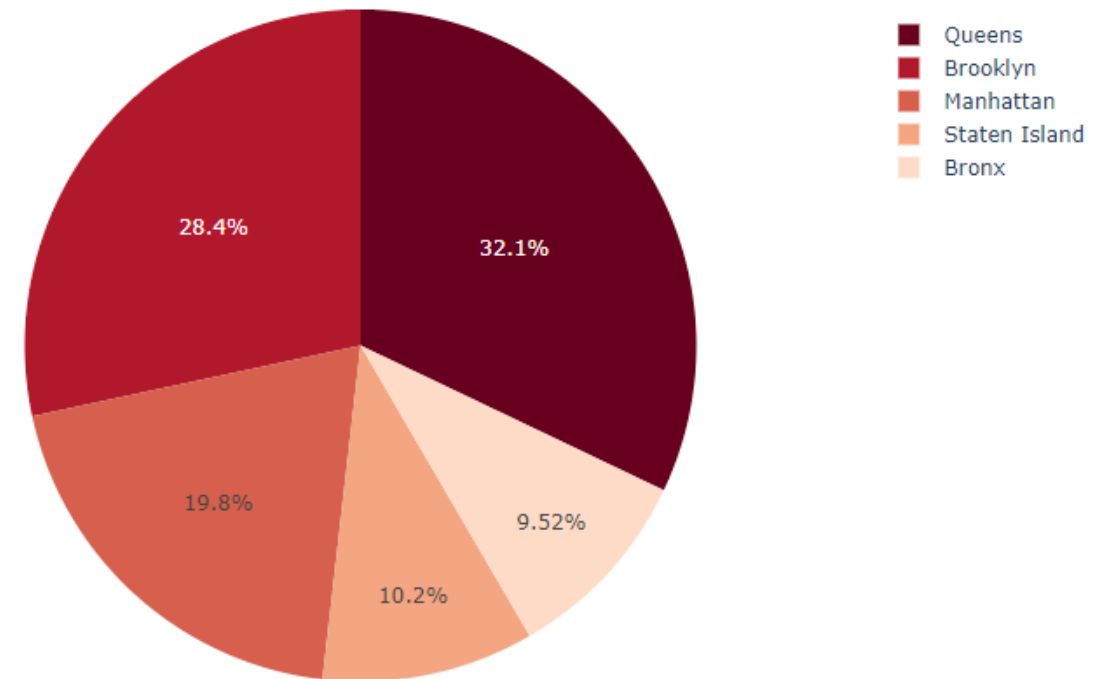


- Ans: Although Manhattan is the most densely populated neighborhood in New York City, Queens in New York City with the most properties in this dataset.
- More than 32% just for Queens Borough , Which is more than Manhattan's percentage ~19%
- This will be useful for governments because they will think that the neighborhoods with more real estate have more population density, which is not true, and therefore they will focus their efforts to provide public services to the most crowded neighborhoods
- Note : BOROUGH -> A digit code for the borough the property is located in; in order these are Manhattan (1), Bronx (2), Brooklyn (3), Queens (4), and Staten Island (5).

RESEARCH QUESTIONS CONT.:

- Now let's start making our Pie chart for the first 6 highest values int the BOROUGH column.

BOROUGH



It seems that Queens has the most properties of all the boroughs in New York City in this dataset.

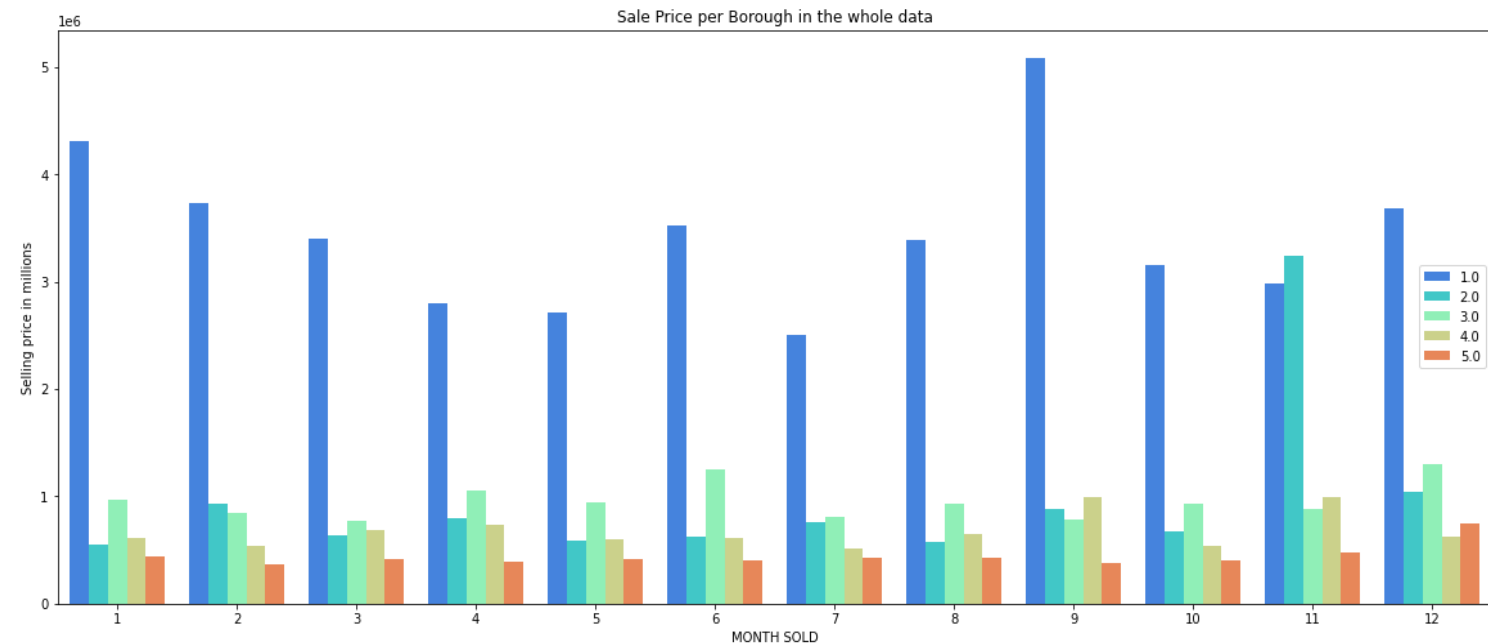
RESEARCH QUESTIONS CONT.:

- What is the distribution of real estate prices for each BOROUGH during the months of the year, and what is the number of units sold for each month as well?
- The importance of this question for the buyer so that he can know the right time to buy the best property at an appropriate price.
- It is obvious to see that the more units sold during a particular month of the year, the lower the real estate prices in that month also, according to the theory of supply and demand.



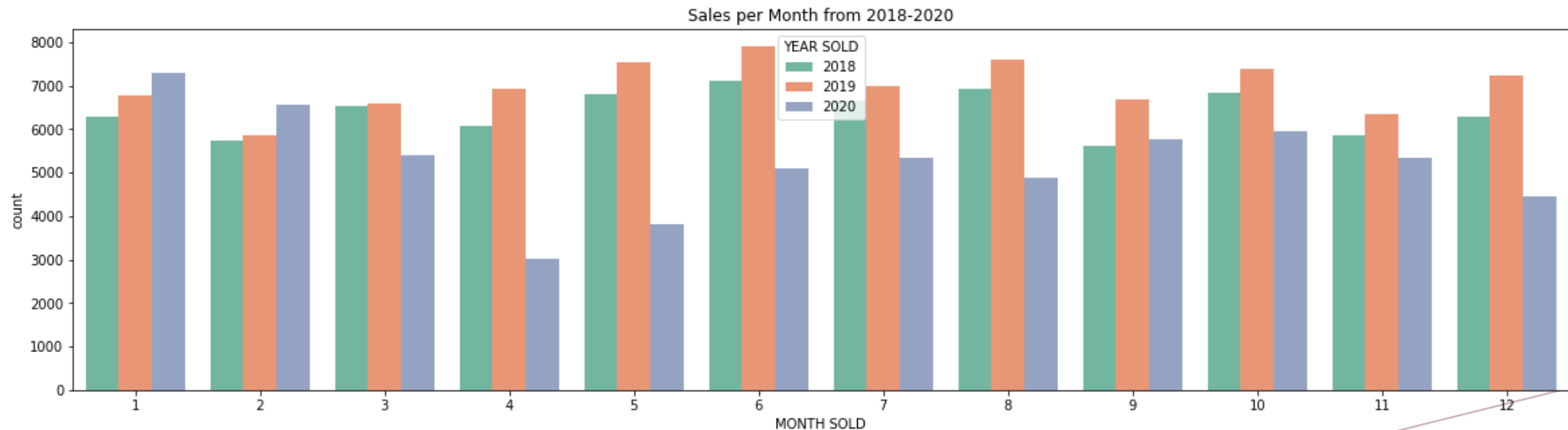
RESEARCH QUESTIONS CONT.:

- It seems that the most expensive property sold in all data was in the city of Manhattan.
- And its value was approximately \$ 5 millions ,and that was in September .



RESEARCH QUESTIONS CONT.:

- It seems that in June 2019, the largest number of properties were sold, which is more than 700 properties.



DATA EXPLORATION VIA STATISTICAL TEST

- Is there a difference between the Average Land square feet in each Borough and Average Gross square feet in each Borough ?
- This will be useful for people who are looking for cheap prices for real estate without looking at the difference between the average ratio for Gross square feet and Land square feet.
- Null hypothesis: The Average Land Square feet for each Borough will be equal to the Average Gross square feet .
- Alternative hypothesis: The Average Land Square feet for each Borough will be different to the Average Gross square feet .

AVERAGE LAND SQUARE FEET IN EACH BOROUGH



- To plot average land square feet in each borough I am using bubble chart, In this type of chart the size of the bubble indicates the magnitude of the variable.

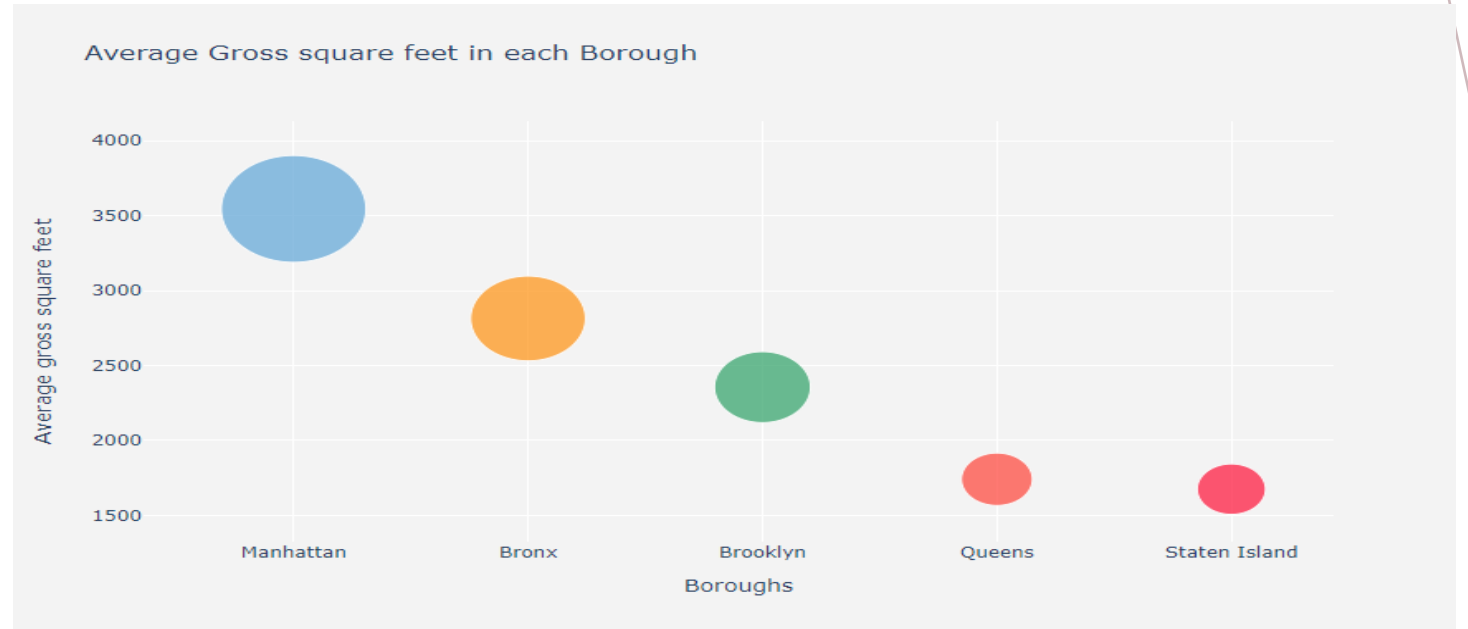
Manhattan price mean	: 2423.697
Bronx price mean	: 4962.092
Brooklyn price mean	: 2634.649
Queens price mean	: 3917.403
Staten_Island price mean	: 8304.309

THE ANOVA FOR STATISTICAL HYPOTHESIS TESTING

- With ANOVA, you get to discover obvious differences between the means of your independent features. Upon getting a clearer picture of the differences.
- you can understand how each of them connects to your dependent variable. You can see what are the influencing factors for the relationship.
- $\text{stat}=47.237, p=0.000$
- Probably different distributions

AVERAGE GROSS SQUARE FEET IN EACH BOROUGH

- Manhattan : 3546.027
 - Bronx : 2814.955
 - Brooklyn : 2356.357
 - Queens : 1741.704
 - Staten_Island : 1675.586
-
- $\text{stat}=57.698, p=0.000$
 - Probably different distributions

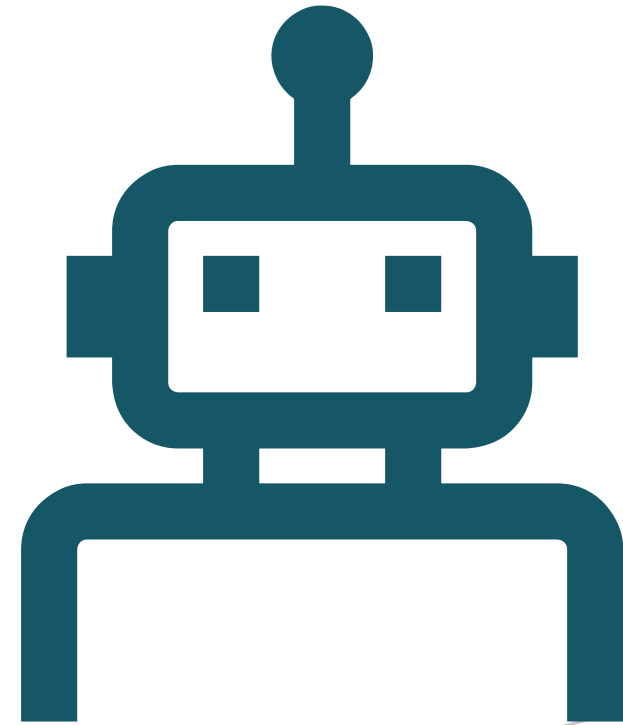


As we can see both of them are different in distributions the first one give us ~47 and the second one give us ~57 in statistics testing with ANOVA test.

So my Alternative Hypothesis is true.

MACHINE LEARNING PART

- This is what my modelling part looks like so far:
 - 1. Modelling
 - 2. Predictions and evaluation
 - 3. Feature Engineering
 - 4. Compare results



MODELLING:

- For the modelling part, I'm going to use pipelines. Pipelines are a simple way to keep your preprocessing and modelling code organized.
- These are the steps that I'm taking for building a machine learning model:
 - Splitting the data
 - Using OneHotEncoding to transform the categorical features into numerical values
 - Using ColumnTransformer to bundle the different preprocessing steps
 - Instantiate a model
 - Bundle the preprocessing and model code in a pipeline

PREDICTIONS AND EVALUATION

- Mean Absolute Error (MAE): Every residual (distance between fitted line and observation) needs to be calculated for every data point, taking only the absolute value of each. Then the average of the residuals is the MAE.
- Mean Squared Error (MSE): The MSE is just like the MAE, but squares the difference before summing them all instead of using the absolute value. MSE is more popular because it punishes larger which tend to be useful in the real world.
- R²: In regression, R² a statistical measure of how well the regression predictions approximate the real data points.

FEATURE ENGINEERING

- To improve my model's performance, I'm going to use feature engineering. Feature engineering is the process of creating new features that didn't exist before to improve my models predicting performance.
 - Mathematical transforms: Using arithmetic operations to create new numerical features.
 - Group Transforms: Groups transforms aggregate information across multiple rows grouped by some category.
- Features I'm thinking about:
 - Price per land square feet
 - Price per gross square feet
 - Median sale price per neighborhood
 - Median sale price per borough
 - Median sale price per zipcode

COMPARE RESULTS

- To compare the results, I'm taking a look at the performance on the validation data.
- For the predictions and evaluation part of this project I'm using the Random Forest Regressor.

Before Feature Selection	
Validation MAE	72.01%
Validation MSE	84.71%
Validation R2	15.18%

After Feature Selection	
Validation MAE	0.098%
Validation MSE	0.025%
Validation R2	99.75%

COMPARE RESULTS CONT.

Algorithm	Train Score	Val MAE	Val MSE	Val RMSE	Val R^2
AutoML	0.9993	0.014	0.003	0.051	0.997
Random Forest Model	0.9997	0.014	0.003	0.051	0.997
Linear Regression Model	0.091	0.771	0.918	0.958	0.085
KNN MODEL	0.804	0.243	0.310	0.557	0.691
SVR MODEL	0.114	0.649	0.901	0.949	0.109
Decision Tree Regressor MODEL	1.0	0.017	0.007	0.081	0.993
GBM MODEL	0.969	0.101	0.033	0.182	0.967
AdaBoost Regressor MODEL	0.715	0.451	0.283	0.532	0.823
XGB MODEL	0.969	0.099	0.033	0.181	0.967

SOME COMMENTS ON THE RESULTS

- The models with the highest performance were the Decision Tree Regressor and Random Forest Models with (0.997 for R^2). Knowing that they were the fastest to train.
- On the other hand, the SVR model (with 0.109 for R^2) and the linear regression model (with 0.085 for R^2) were not good and they took a lot of time training as well, unlike the other models.
- In the end, choosing the right model depends on the problem that I want to solve.



THANK YOU.

