CIS 5367.251 MACHINE LEARNING PRESENTATION 1

- Bezawit Tafesse
- Jonathan Cook
- Babacar Tall
- Harshal Shelare
- Sasha Munir

ALGOPROD CONSULTING, LLP

CONTENTS

- Introduction
- Business scenario
- Project Specification
- Data Collection
- Data Cleaning and Transformation
- Data Set values
- Visualization
- Conclusion

BUSINESS SCENARIO (BT)

- We're a team of data scientist consultants, based in San Marcos, TX.
- One of our clients runs a real-estate firm:
  - They provide relocation services to Texas, with locations in multiple cities.
  - Going through restructuring and they want to add the newest technologies to optimize housing options for their customers.

REAL ESTATE INDUSTRY

- Potential home buyers often consider:
  - Square feet of the home
  - Number of stories
  - Number of rooms
- Multiple variables such as average sales prices, crime reported as well as number of schools and hospitals located in the area are of much importance on choosing a home
- We have considered these variables to help build a model to estimate a fair housing price in a given region or area,especially the safety
  - Our project would help our client to optimize the process of providing crucial information that would assist their customers in deciding where to buy a house.

PROJECT SPECIFICATION (JC)

- Our focus is on the ideal location to buy a home, based on the pricing trends and crime rates over the last few decades.
  - Our project is a machine learning model which predicts the value of a home based on the crime rate of the region it is located.
  - It is useful for both home buyers and sellers.
  - Our model will have transparency about the safety of the home's neighborhood relative to the price.

DATA COLLECTION

- We have carefully vetted the source for our training data set.
- Our data is updated and uses data collected as recently as last month.
- We chose two different data sources and merged them to get the best values:
  - One source includes data about crime, while the other has data about pricing.
- Our data is collected for the metropolitan areas of Texas:
  - Dallas, Austin, Houston, Laredo and many more areas

DATA COLLECTION

- The first data was from an official source that has collective data of real estate sales over the past years for each city: Texas A&M University's Texas Real Estate Research Center
  - https://www.recenter.tamu.edu/data/housing-activity/#!/activity/State/Texas
  - This includes data about housing activity trends in Texas since January of 1990.
- The second data set was crime rates of cities for each month over the last few decades: FBI's Crime Data Explorer
  - https://crime-data-explorer.app.cloud.gov/pages/explorer/crime/crime-trend
  - Data about crime trends and statistics in Texas since 1990.

DATA CLEANING AND TRANSFORMATION

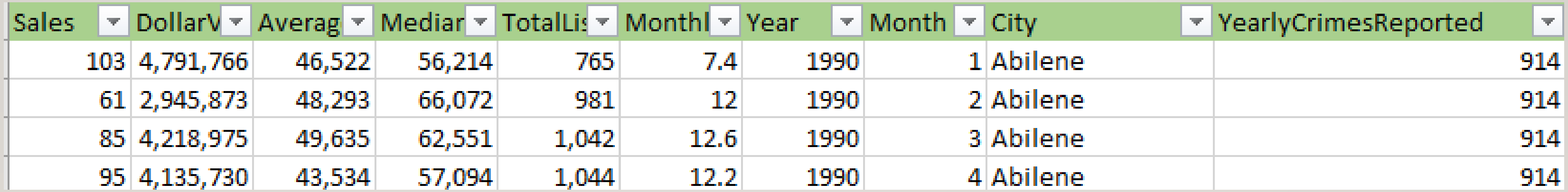
- We merged the two data sets to get a collective data of crime rate and home sales for each city over a span of time.
- We will use this data and parameters to train our model.
- This will give us a good model that enables the prediction of house price based on the city and location of the house.
- Our data has various parameters such as the month, crime rate, city, price, number of sales per month etc..

DATA CLEANING AND TRANSFORMATION

- Eliminated any duplicates or null values that were not useful.
- Filled the null values where needed with aggregated values from that column.
- Transformed the data to make it usable in spark. For example, converted the object datatypes to int, float, str datatypes.

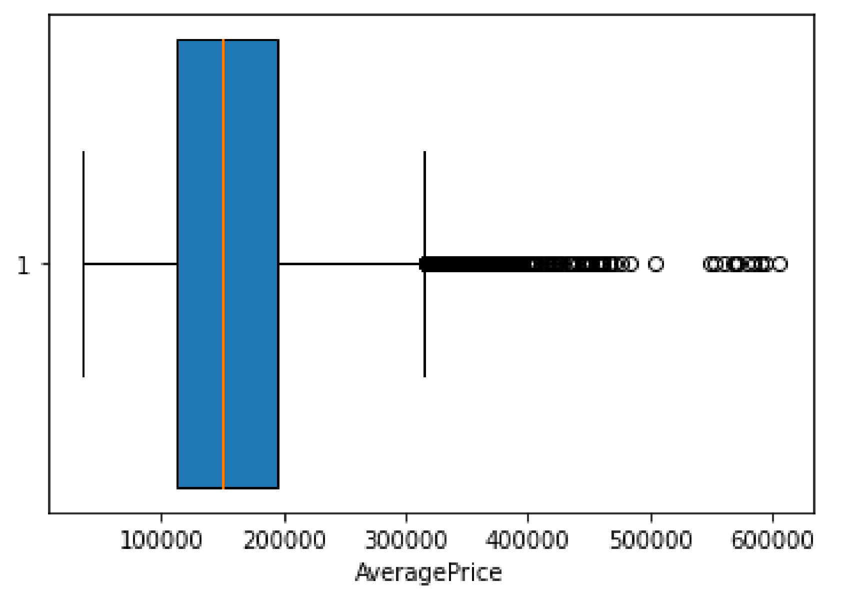
DATA SET VALUES (TB)

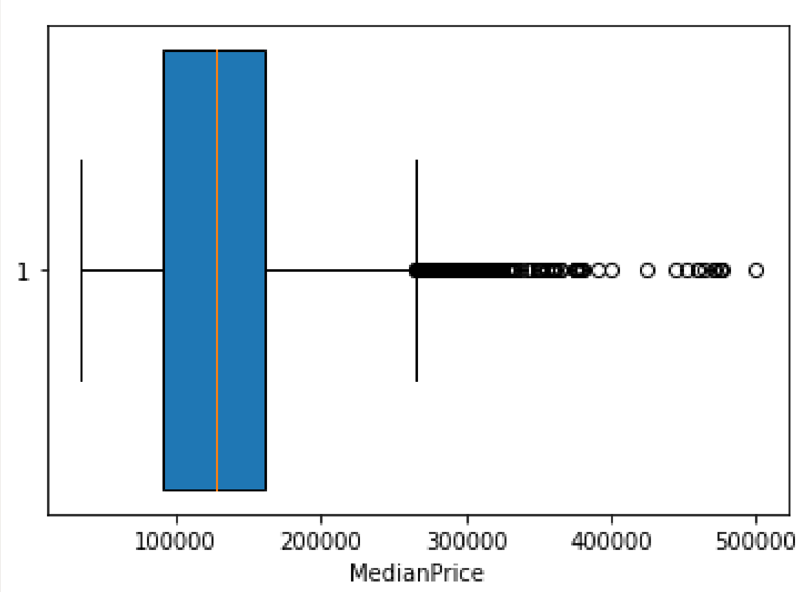
- Our housing data table looks like this after cleaning up. We will use the following variables in our analysis (refer table):
- Price, crime numbers reported, number of hospitals and schools are our input parameters, which will help to predict the recommended city to purchase a home with a fair price.



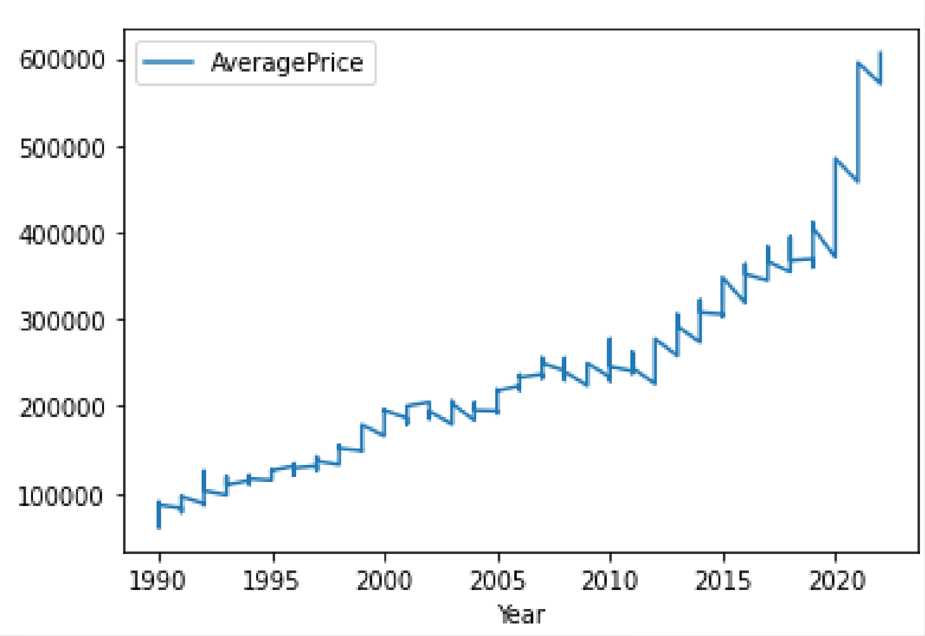
Sales	DollarV	Averag	Mediar	TotalLi	Monthl	Year	Month	City	YearlyCrimesReported
103	4,791,766	46,522	56,214	765	7.4	1990	1	Abilene	914
61	2,945,873	48,293	66,072	981	12	1990	2	Abilene	914
85	4,218,975	49,635	62,551	1,042	12.6	1990	3	Abilene	914
95	4,135,730	43,534	57,094	1,044	12.2	1990	4	Abilene	914

VISUALIZATION - BOXPLOTS

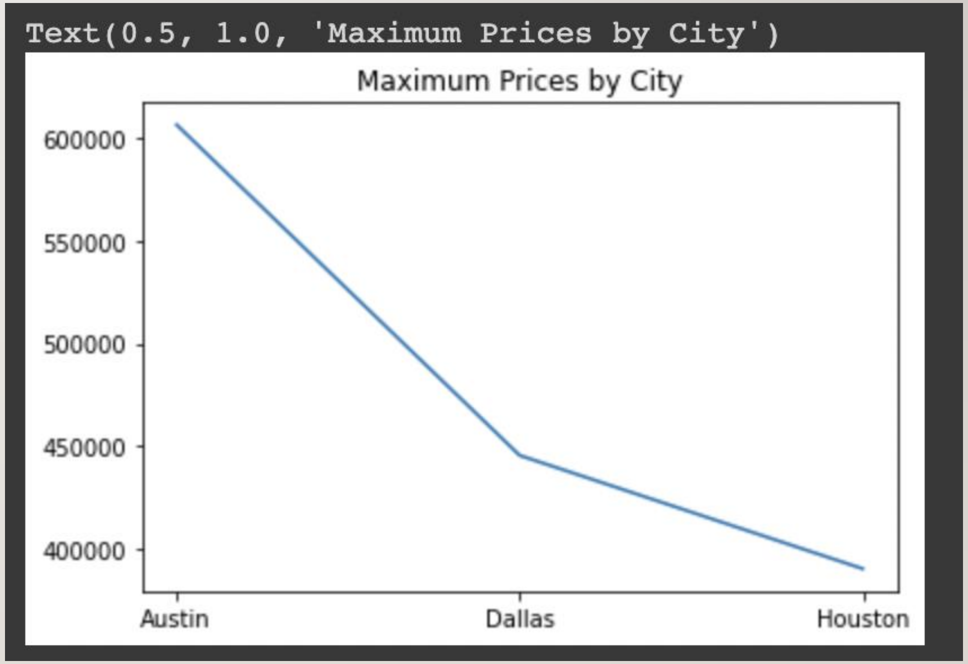




VISUALIZATION (SM)



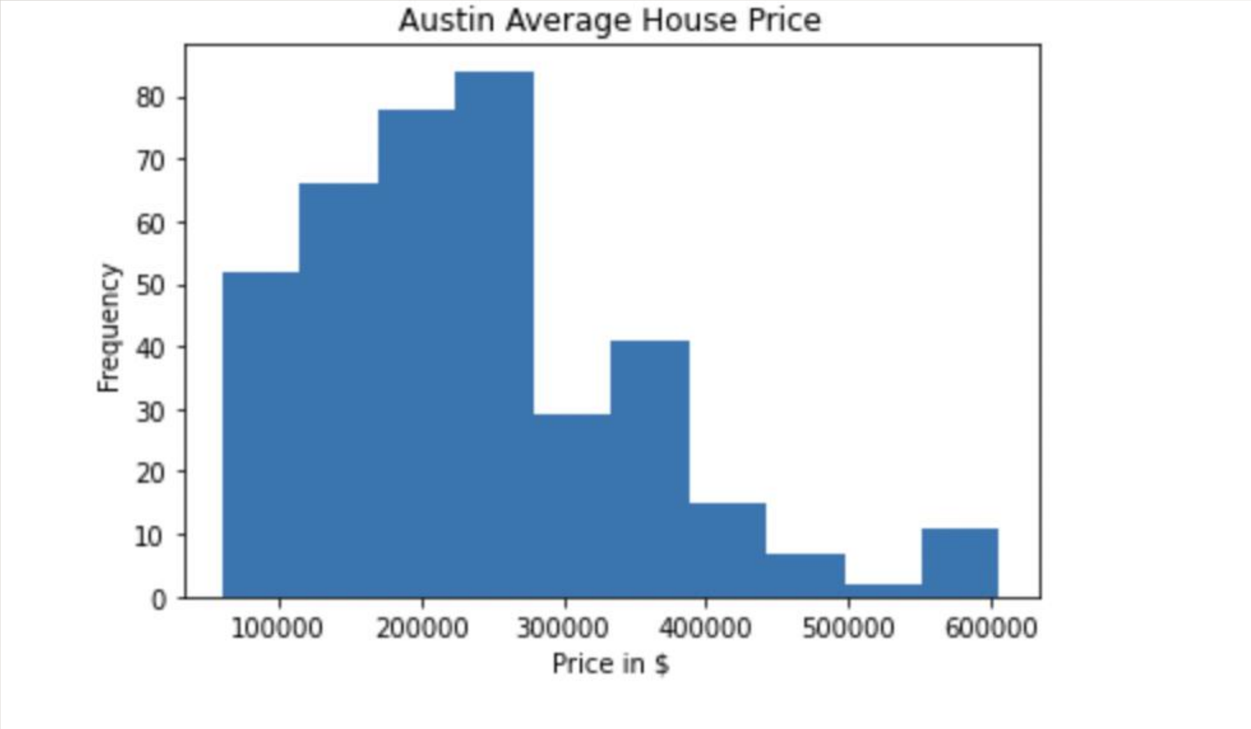
Trend for Austin Average Prices

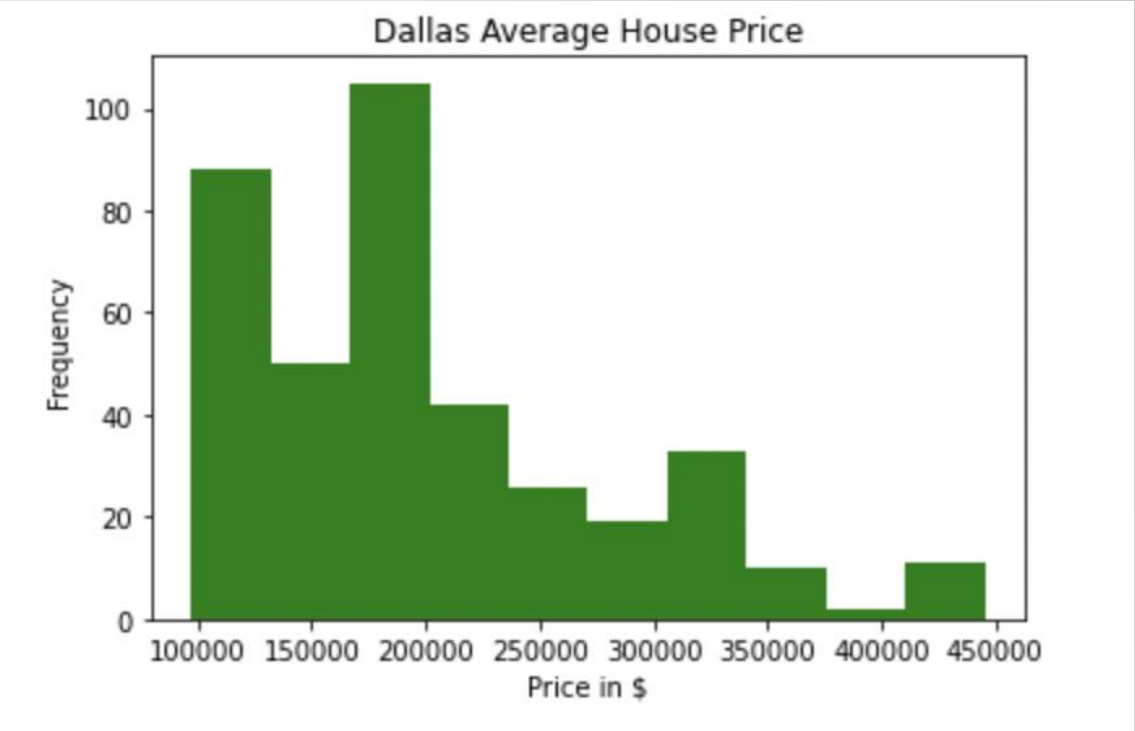


```
Text(0.5, 1.0, 'Maximum Prices by City')
```

Most Expensive City

VISUALIZATION - HISTOGRAMS





CONCLUSION

- Our project enables buyers and sellers to have a predictive value of their homes based on the crime rate of the location,number of schools, education and health services
- Buyers will be able to know the safety of the area they are committing to buy a home in.
- Sellers will be able to know which areas are more valuable (which areas the customers will be more likely to want to move to).

THANK YOU