Project Outline:

Clay

(Slide 1) Greeting

Introduction

- Whether it's the scenic views, bustling nightlife, or eccentric people, there is an allure that is causing people to flock in large numbers to Austin. You might get different answers depending on who you ask.

(Slide 2)

- **Background:** For most of us living or are familiar with the Austin metropolitan area, we can all see the rapid changes that are taking place throughout the region in terms of development and real estate appreciation. It has taken the city by storm. Will this trend continue? If so, which areas are experiencing the most rapid growth and development.
- The inspiration for the project came from the tableau unit of the Bootcamp when we were visualizing the types of housing units available in the Austin metro area. I created a dashboard to discover which zip codes had the most single-family homes for those wanting to start families in the Austin area.

(Slide 3)

- For this project, we wanted to determine which areas in the Austin metro area are poised to appreciate the most in value when compared with the rest. We believe this data is valuable for residential real estate investors such as home flippers, those involved in the home-sharing marketplace, or even those seeking to buy their first home with the intention of selling it later in their life.

Main Body

- **Takeaway:** We want to determine which zip codes in Austin have the most potential and a relatively low barrier of entry.
- How do we determine which areas are candidates?
    - We used permits issued by zip code over time to help us determine areas of interest
    - Constant development in specific regions is an indicator that properties in those regions will appreciate
    - We will attempt to use supporting metrics such as distance from schools, average school ratings, etc

Josue

- **Analysis / EDA:**
    - (Slide 4) Tools used: Excel, Pandas, Matplotlib, Numpy
    - (Slide 5) During the analysis phase of the project, we decided to traverse through a variety of data sets that contained permit data, house listing data, housing market data from the greater Austin area. Considering how sprawled out the Austin metro has become in the past few years, we decided to focus our analysis on land under the jurisdiction of Travis county. For example, San Marcos was considered part of the Austin Metro but is in Hays county!

- (Slide 6) : The first dataset we explored was a table that contained thousands of home listings from Zillow. The data was acquired from Kaggle and features home listings from 2020. Each row pertains to a different listing and the listing's attributes such as home type, location data, pricing information, and home features. The first task in our analysis was checking the number of columns and their data types. The dataset contained 46 columns with a variety of data types. We decided to condense the dataset into eight columns that contain pricing and location information for our analysis. The next step was addressing null values, if any, and dropping duplicate data. We then aggregated data by zip code to find the average home value of each zipcode from the listing data.
- (Slide 7) We also applied the same methodology to determine the average number of price changes from each zip code since we are interested in determining which areas are susceptible to price changes. We combined both outputs into a new DataFrame.

Kelly

- Slide 8:
  - The Austin permit data was sourced from data.austintexas.gov. ~~After opening the CSV file in Jupyter Notebook,~~ we used Pandas to condense the 68 column dataframe down to 5 columns that could explain each permit succinctly with relevant data such as Zip Code and Calendar Year Issued. ~~We dropped rows with null values, reordered columns with zip codes as the first column, and removed the index column before exporting the CSV file. This initial cleaning of the data gave us 1,994,987 permits for our analysis, down from 2,033,655 in the original dataset.~~

  - ~~Some of the challenges we faced with this particular dataset were due to the size of the data as well as a majority of 'string' data types within the dataframe, which was hard to manipulate within rows using Pandas. To quickly filter the condensed permit data, we used Excel which allowed us to focus on permits issued in the past five years, and narrow down to 15 zip codes with the highest number of permits issued in that timeframe.~~

  - ~~After importing that data into a data frame, we summarized it with the describe method, then used the melt function to convert the data from a wide to a long format. We determined the average number of permits issued per zip code and sorted the data in descending order. We chose to focus our analysis on the 15 zip codes with the most permits issued, since we predict those areas are undergoing the most rapid development and, ultimately, the most appreciation in home values. We will use these zip codes to determine which homes in the Zillow listing data are potentially lucrative investments.~~

- Slide 9 : The home value data was cleaned and formatted to show the last five years worth of home listings
- ~~features time-series data that reflects the typical home value from a given zip code across time. Zillow uses the 'Zillow Home Value Index (ZHIV)' to measure the typical home value, which is NOT the same as the median home value. Each row is a different zip code with location and pricing data every month from 1996 until the end of June 2021. Zillow has this data publicly available on their website, along with their methodology on how they derive their ZHVI. The original dataset contained over 300 zip codes across the nation, so our first task was filtering the data with the 'iloc' method to only contain the zip codes within Travis county. Next, we needed to address the wide format of the dataset, where each column was a different month. We used the pandas' melt function to convert the data frame from a wide to a long format. The resulting data frame was condensed further to include only the home values the past five years since home values have drastically appreciated during this time frame.~~
- (Slide 10) Using the permit data, where we acquired the top fifteen zip codes that have the most issued building permit, we explored the price trends of each respective zip code with the help of visualizations created with matplotlib.
- (Slides 11) This simple bar graph shows the number of permits issued per zip code. They range from around 100 permits to over 1,000 permits, which is obviously a very broad range.

- (Slide 12) Similarly to the permit bar graph, the typical home value by zip code bar graph shows a very broad range from 200k dollars to 1 million dollars.
- (Slides 13) With so many factors affecting home values, we looked at a few variables within our 15 zip codes. All of these were made in tableau and are included in our dashboard which is linked at the end of our presentation, but cut for time. (READ SLIDE)
- (Slide 14) (READ SLIDE)
- (Slide 15) This bar graph shows the number of permits by zip code with property tax rates which hints at a possible correlation.
- (Slide 16) This is my favorite chart we created, (READ SLIDE)

Josue

- **Database: (**Slides 17, 18, 19)

Steven

- **Machine Learning: (Slide 20)**
  - Using the potential zip codes, we iterated through a data set that contained home listings from Zillow to create a new possible investment candidate column. Using a similar approach with future datasets, we can swiftly determine candidates in the Austin area.
  - (Slide 21) We ran 2 models: The first was a random forest classifier
  - The preliminary data included columns that describe the amount of permits, latest price, number of price changes, year built, zipcode of homes, etc.

    - After connecting to the database, we printed out the header for each column to see all of the features available, then chose the features that we believed would have the highest correlation to increased home prices per zip code.
    - The data was split into training and test data using the train_test_split function. We used the 75% to 25% default split.
    - We grouped our data into two categories: candidate or non-candidate for investing. The benefit of this model is that it can be used to predict our binary outcome.
    - Our machine learning model will attempt to predict the candidate column of the Zillow data frame. The candidate column displays a one if the property is located in a rapidly growing region of the city. Otherwise, the column will display a zero for properties that are not considered to be in rapidly growing areas. Using scikit-learn, our first model uses a decision tree to predict the outcome of each row of our data frame. Considering most of the columns from the Zillow data are categorical, the first step was encoding the columns that contained categorical variables. From there, we followed the supervised learning workflow of separating the target and features, splitting the data, fitting the model, and then assessing the model performance. Our provisional model appeared to have symptoms of overfitting, considering we achieved an accuracy score of 100 percent.
    - (Slide 22) Because of the overfitting in the decision tree classifier model, we decided to use a Logistic Regression Model. The main benefit of using this model is that it allows us to predict a binary outcome - candidate or non-candidate. The variables were narrowed down to include columns that describe the number of permits, latest price, number of price changes, year built, zipcode of homes, etc. The solver "SAGA" was used to quickly run a large data set. the precision to predict candidate and non candidates was about the same at 57% and 58% respectively
        - the sensitivity was much higher on predicting candidates at 82% than non candidates at 28%
    -

Josue

Conclusion (Slide 23)

- Zip codes of interest:
    - 78744, 78704, 78745, 78723, 78702, 78660, 78757, 78731, 78703, 78748, 78701, 78759, 78758, 78747, 78741
    - We narrowed down the potential zip codes to 15. Those 15 zip codes had the highest number of permits issued within the past five years.
        - shows images of the zip codes that are not that profound. Most people already know about these bustling areas so the barrier to entry is high
    - Although there are areas that appear to be no-brainers for those familiar with Austin, such as the zip code for downtown Austin and the zip code encompassing Zilker/Bouldin creek, we also discovered elusive zip codes:

- (Slide 24)
  - 78745: South Austin (Specifically, the area southeast of the greenbelt, including Sunset Valley)
- (Slide 25)
  - 78744: Southeast Austin (Area encompassing McKinney Falls and adjacent to the Airport)
  - 78747: South of McKinney Falls near Mustang Ridge
- (Slide 26)
  - 78741: East Riverside & Montopolis
  - 78723: Windsor Park and University Hills
  - 78757: North Shoal Creek

Clay

- Addressing the elephant in the room:
  - The data and insights we gathered also creates a new dilemma that the city government will need to address; affordable housing and transportation. With specific areas seeing unprecedented property value appreciation, we can expect population displacement. For example, consider the working-class individuals who work near the city center and its surrounding areas. Many will likely move further away from where they work to afford their living expenses, which will inadvertently create an affordable housing shortage and longer commutes that will stress transportation networks.
  - There are many ramifications of rapid development that can be stand-alone projects on their own.
  - The more you dig into the data, the more questions that arise