Shahjalal Sikder
John Jakobsen
Uzmabanu Kapadia

Preliminary Analysis

How many days will it take a property to sell?

Data Cleaning Code

Code for cleaning and processing your data. Include a data dictionary for your transformed dataset.
https://github.com/jerikjakobsen/Data-Science-Project

The final cleaned csv file is located here.

Original Data Set for property sales:
https://www.dolthub.com/repositories/dolthub/us-housing-prices/query/main?active=Schemas&q=SHOW+CREATE+TABLE+%60sales%60
Original Data Set for Trulia Listings:
https://www.kaggle.com/datasets/promptcloud/real-estate-data-from-trulia

Please note that the original data set that was crossed with the trulia data set to get the sale dates for each of the listings was 50GB so we were unable to upload it to github. However the code for cleaning and processing that dataset is located here. After filtering the sale data set we merged it with the trulia data set to get the days the listing was on trulia along with the date the property was sold.

State - State the house sold is; String
Address - Address of the house sold; String
City - City of the house sold; String
Sale_price - Price the house sold at; Int
Sale_date - Date the house sold at; Object (Date)
Sqr Ft - Square Feet of the property (In Squared Feet); Int
Longitude - Longitude of the property; Float
Latitude - Latitude of the property; Float
Lot Size - Size of the Lot (In Squared Feet); Float
Beds - Number of bedrooms in the house; Int
Bath - Number of baths in the house; Int
Year Built - Year the property was built; Int

Days On Trulia - Number of days the property was on Trulia before it sold; Int
Days on Market - Number of days the property was on the Market before it sold; Int
Zipcode - Zipcode of the property; Int
Listing_date - Date the property was initially listed on Trulia; Object (Date)

Cleaning the Property Data

To clean the property data we dropped the rows where the columns of interest were unavailable, then we lowercased all the columns that had strings as their dtype, we renamed the Address, City and State to match with the Trulia data set. Lastly we converted the data types of the remaining columns to their appropriate types. The only columns we kept from this data set was State, City, Address, sale_datetime and sale_price.

Cleaning the Trulia Listing Data

To clean the Trulia listing data we removed the columns we didnt want to use, lowercased all the columns that has a string dtype, we then cleaned the address column by removing the state, zip code and city from them (Code for that can be found here). We converted the sqr feet column to int by removing 'sqr feet', we converted all Lot Size to int and to feet. We then calculated when a listing was posted by subtracting days on trulia from crawl Timestamp.

Merging the Two Data Sets

To merge the two datasets was pretty simple, we merged them on the State, Address and City columns, then since there were multiple sale dates for houses we removed all the rows that had a sale date before the listing date, then to further remove duplicates, we sorted by sale date for each house and removed all but the earliest sale dated rows. To calculate days on market we simply subtracted the listing_date column from the sale_date column.

Exploratory Analysis

Describe what work you have done so far and **include the code.** This may include descriptive statistics, graphs and charts, and preliminary models.

At this point of the project, we explored multiple datasets regarding real estate. After cleaning out our data and filtering through usable data points with pandas, we found several correlations between the various data sets: relationship between square feet and house price, location and house price, and also the names of the cities and states that make up the data that we have.
More specifically, after the process of cleaning out the code, we dove into creating scripts that can display these correlations, which include the CSV files in the github,

which illustrates house prices in certain cities and states along with specific characteristics such as size, number of baths and bedrooms, the address, the date it as put on sale, the price, etc…
The Sales Price, Bath, Sqr Ft Columns have notably higher correlations with Days on Market than any other column with Sales Price having a correlation score of 0.18, Bath having a correlation score of 0.16 and Sqr Ft having a Correlation score of 0.17.

The code for correlations.py:

```python
import pandas as pd
import sys
import os
filePath = sys.path[0] + "/CSVFiles/FilteredListing+SaleData.csv"
data = pd.read_csv(filePath, sep=",")
print(data)

data.corr().to_csv("correlations.csv")
```

	Days on Market	sale_price
sale_price	0.18510349045313862	1.0
Longitude	0.009133669431530598	-0.3055267866242391
Latitude	-0.0009753684934820287	0.011794391903366324
Lot Size	0.02945362006294566	-0.016748064347063097
Beds	0.13231863460040733	0.34087843158655323
Bath	0.1607873947371871	0.5707179063523402
Year Built	-0.0991110394102611	-0.03599385530772069
Days On Trulia	0.4545536735853784	0.11407513546973143
Zipcode	-0.03844938406831873	0.27589456896764647
Days on Market	1.0	0.18510349045313862

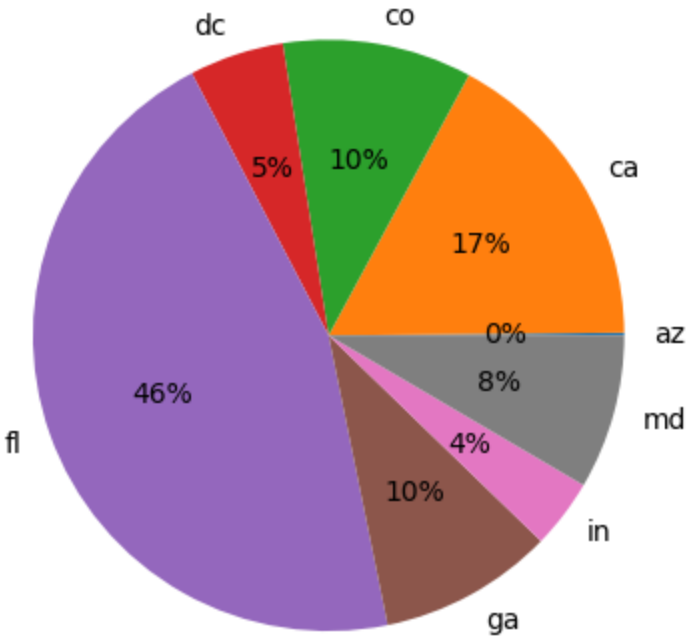
The code for state_pie.py:

```python
import pandas as pd
import os

filteredDF = pd.read_csv(os.path.join('..', 'CSVFiles', 'FilteredListing+SaleData.csv'), index_col=0)

filteredDF.groupby(["State"]).sum().plot(kind='pie', y= len(filteredDF), autopct='%1.0f%%' )

```



Along with finding the correlations, we use pandas to display how many of the states make up our dataset through a visual representation. This pie chart can help us see the percentage of houses that make up the total market. This can help analyze which state has the most houses on sale, possibly affecting the sell time in that area.

Challenges

Describe any challenges you've encountered so far. **Let me know if there's anything you need help with!**

Throughout the process so far, there have been some bumps in the road, as well as some processes that did not take that much time at all. For example, some of us had trouble using pandas on our own machines, where the terminal would display import errors and make it very difficult to run the code. Finding a solution was very tedious for this due to the ambiguity of the errors that we had and no concrete solution found to help us set up the environment.

Also since we had to change our question we needed to find a new data set that supported the question we had, How long will it take a house to sell on Trulia? We ended up having to use two data sets to solve this which worked out pretty well.

Future Work

Describe what work you are planning to complete for the final analysis.

We plan to further dissect the data to find more relevant ways for analyzing the data. This could include looking at correlations of prices, location and the amount of days it stayed on trulia to help get a sense of how the houses sell in different areas, how long do they stay on market usually in an area and if the prices play a big role in this or not. We also plan to figure out our model we want to use based on our analysis as a next step.

Contributions

John Jakobsen - Worked on the Data Cleaning Code and helped with Analysis
Uzma Kapadia - Worked on the preliminary analysis and analysis code
Shahjalal Sikder- Worked on the preliminary analysis and determining correlations