Student Housing Matchmaker Final Report

Brandon Kim, Cristal Martinez, Drew Vranicar, Manav Kohli, Tyler Yuen

Data

Cleaned data can be found in our project GitHub at the following link:
https://github.com/cmart71/Student_Housing_Matchmaker
Most cleaning was done in the Mid-Progress Report*
Along with the cleaned data is our Mid-Progress report, and CSV data files. Additional code for other visualizations or ML/Statistical Analyses can be found in the Code folder.

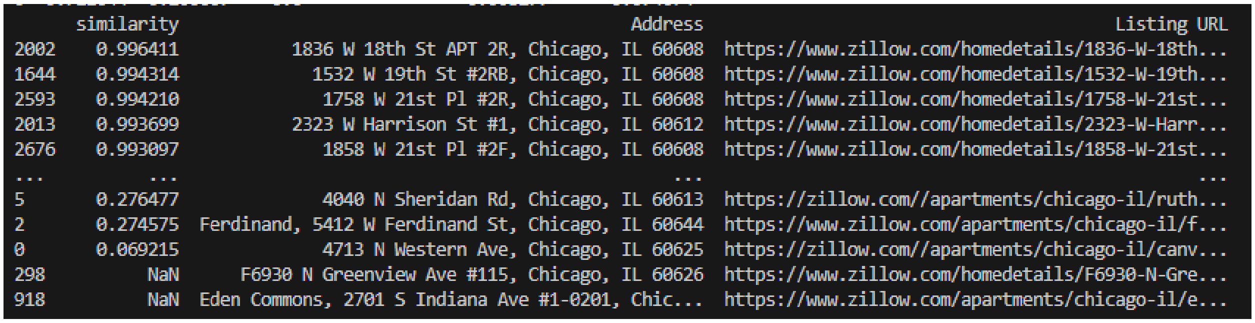
Machine Learning / Statistical Analysis

Machine Learning / Statistical Analysis 1 (Tyler Yuen)

For my analysis, I focused on utilizing an unsupervised learning technique, similar to clustering, to find similarities between apartment listings and a student's input preferences, specifically revolving around monthly rental price, the number of bedrooms and bathrooms, the distance between the listing and the school, and nearby Divvy stations. In order to accomplish this, I had to build a feature vector for the apartment listings along with a student's preference vector. Afterwards, I would normalize this apartment listings vector using a Min-Max normalization to scale the values from 0 to 1. From there, I calculated the cosine similarity between the student's preferences and the normalized apartment listing vector, which resulted in a sorted output of apartments and similarity scores, presenting the student with their best apartment matches based on their input. This is the foundation for our recommendation system.

For the data, I used our cleaned Chicago rentals dataframe as our baseline dataframe to work with. I only included the price, address, listing URL, latitude, longitude, beds, and bath columns as the other columns were not necessary. With the trimmed dataframe, I wanted to include a column for the distance from a specific school to each of the apartments. I utilized the Haversine formula (as seen in our progress report) to find and add columns for the mile distance between the school and the apartment listings, as well as for the number of nearby Divvy stations. To clarify "nearby", I scaled anything within a distance of 0.5 miles to be nearby. Using the sklearn library, I was able to create a MinMaxScaler and transform our apartment listings dataframe. Since this is a recommendation system, I needed to create a students preference dataframe to mimic what a student would input in terms of ideal apartments so I made one with a preference of monthly rental price at $1400, 1 bedroom, 1 bathroom, a distance of 1 mile to the university, and 2 nearby divvy stations. After clarifying our students' preferences, I used our same scalar to transform our students' dataframe, and I then performed cosine similarity over the two normalized dataframes to get the output we see below.

Some inferences that I can draw from this analysis are that this method is a very effective recommendation method due to how well a student's preference is able to be met. If you were to compare the preferences that you would input with the Listing URL, you would find that you are given a very reasonable apartment that meets your needs. Along with that, it provides ample flexibility and can be expanded to meet even more preference criteria and different schools just by adding or removing features to the vector. With enough time and more resources, this tool could be fully imagined and feasible.



```
      similarity                                        Address                                        Listing URL
2002    0.996411           1836 W 18th St APT 2R, Chicago, IL 60608  https://www.zillow.com/homedetails/1836-W-18th...
1644    0.994314             1532 W 19th St #2RB, Chicago, IL 60608  https://www.zillow.com/homedetails/1532-W-19th...
2593    0.994210              1758 W 21st Pl #2R, Chicago, IL 60608  https://www.zillow.com/homedetails/1758-W-21st...
2013    0.993699           2323 W Harrison St #1, Chicago, IL 60612  https://www.zillow.com/homedetails/2323-W-Harr...
2676    0.993097              1858 W 21st Pl #2F, Chicago, IL 60608  https://www.zillow.com/homedetails/1858-W-21st...
...          ...                                            ...                                                ...
5       0.276477              4040 N Sheridan Rd, Chicago, IL 60613  https://zillow.com//apartments/chicago-il/ruth...
2       0.274575  Ferdinand, 5412 W Ferdinand St, Chicago, IL 60644  https://www.zillow.com/apartments/chicago-il/f...
0       0.069215              4713 N Western Ave, Chicago, IL 60625  https://zillow.com//apartments/chicago-il/canv...
298          NaN      F6930 N Greenview Ave #115, Chicago, IL 60626  https://www.zillow.com/homedetails/F6930-N-Gre...
918          NaN  Eden Commons, 2701 S Indiana Ave #1-0201, Chic...  https://www.zillow.com/apartments/chicago-il/e...
```

(Example of the output which shows the highest match at the top based on what the student input for Price, Beds, Baths, Distance, and Nearby Bike Stations)

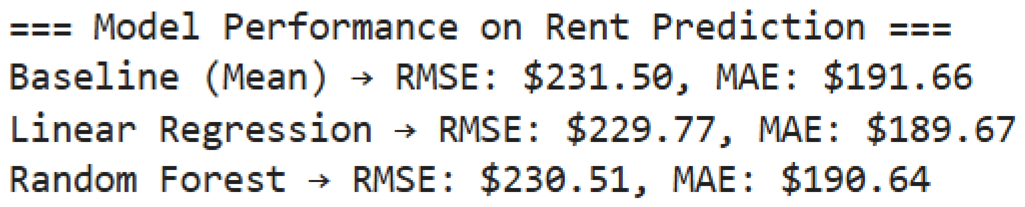
Machine Learning / Statistical Analysis 2 (Brandon Kim)

The Machine Learning analysis I did was based on predicting the prices of rentals based on the number of colleges within a 2 mile radius of that rental.

The first thing I did was to calculate the number of colleges within a 2 mile radius of every rental. I did this by applying the Haversine formula on every rental location with every college in the dataset for Chicago colleges. I chose a 2 mile radius based on information gathered from the first visualization(credit to Tyler) that 2 miles seemed to represent the nearby colleges.

After obtaining the number of colleges within 2 miles for every rental, I then used mean, linear regression, and random forest models to predict the rental price based on the number of colleges near the rental.

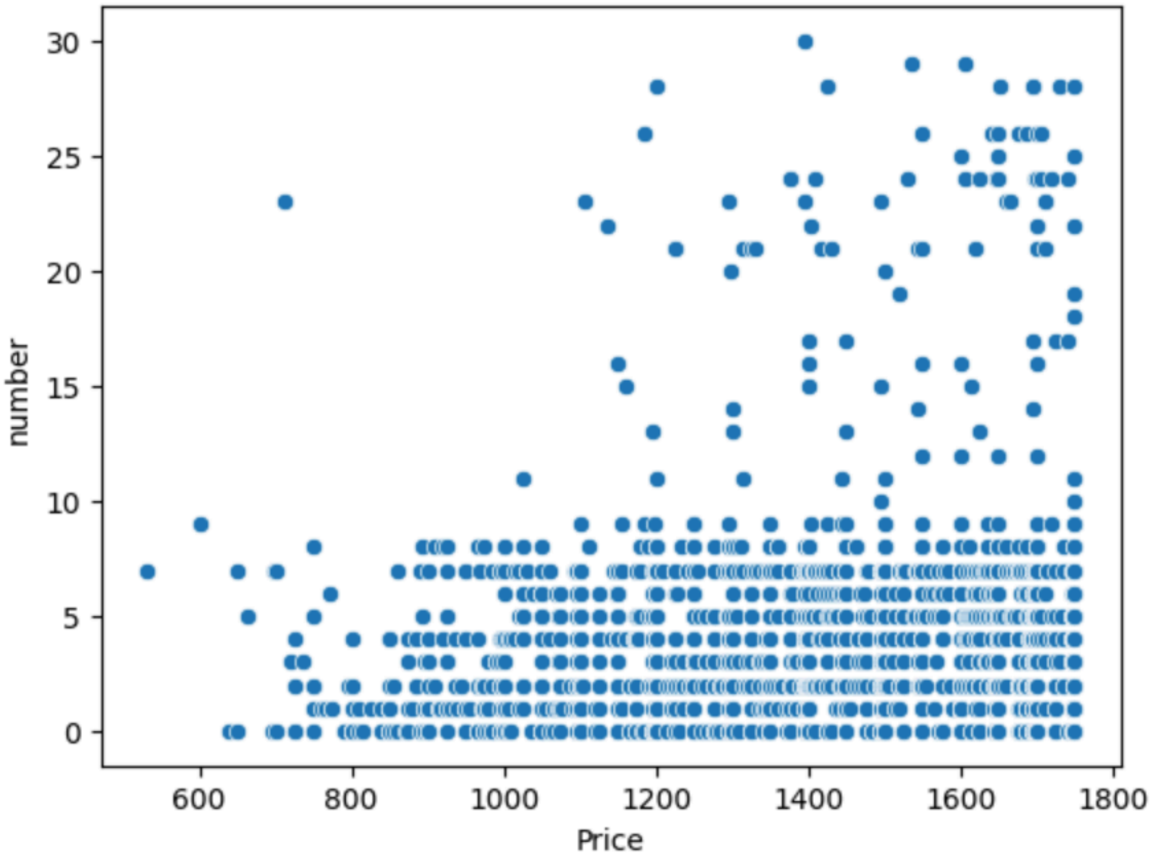
Here are the results:



```
=== Model Performance on Rent Prediction ===
Baseline (Mean) → RMSE: $231.50, MAE: $191.66
Linear Regression → RMSE: $229.77, MAE: $189.67
Random Forest → RMSE: $230.51, MAE: $190.64
```

The models didn't do much better than the baseline mean only model, which indicates that the number of colleges within a 2 mile radius of a rental isn't a good predictor of the price.

Here's a graph plotting the price vs number of colleges nearby:

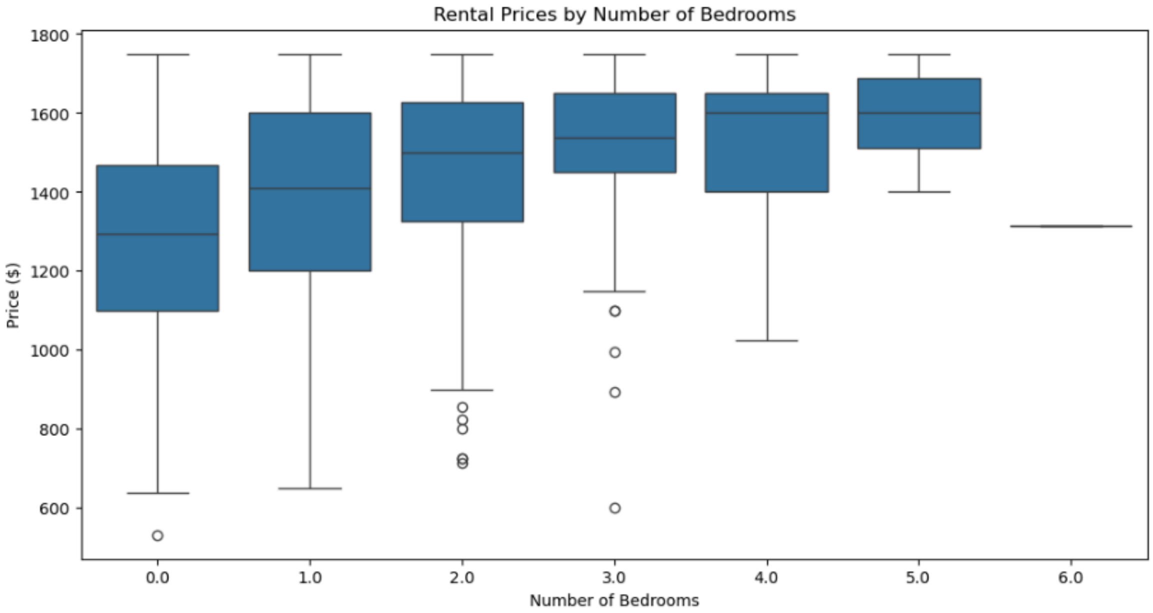


While there seems to be rentals with higher prices with a higher number of colleges nearby, there are also a lot of rentals in general that have a wide range of prices and under 10 colleges nearby. There seems to be a correlation, but not strong enough to come to a meaningful conclusion and certainly not strong enough to predict prices.

Machine Learning / Statistical Analysis 3 (Cristal Martinez)

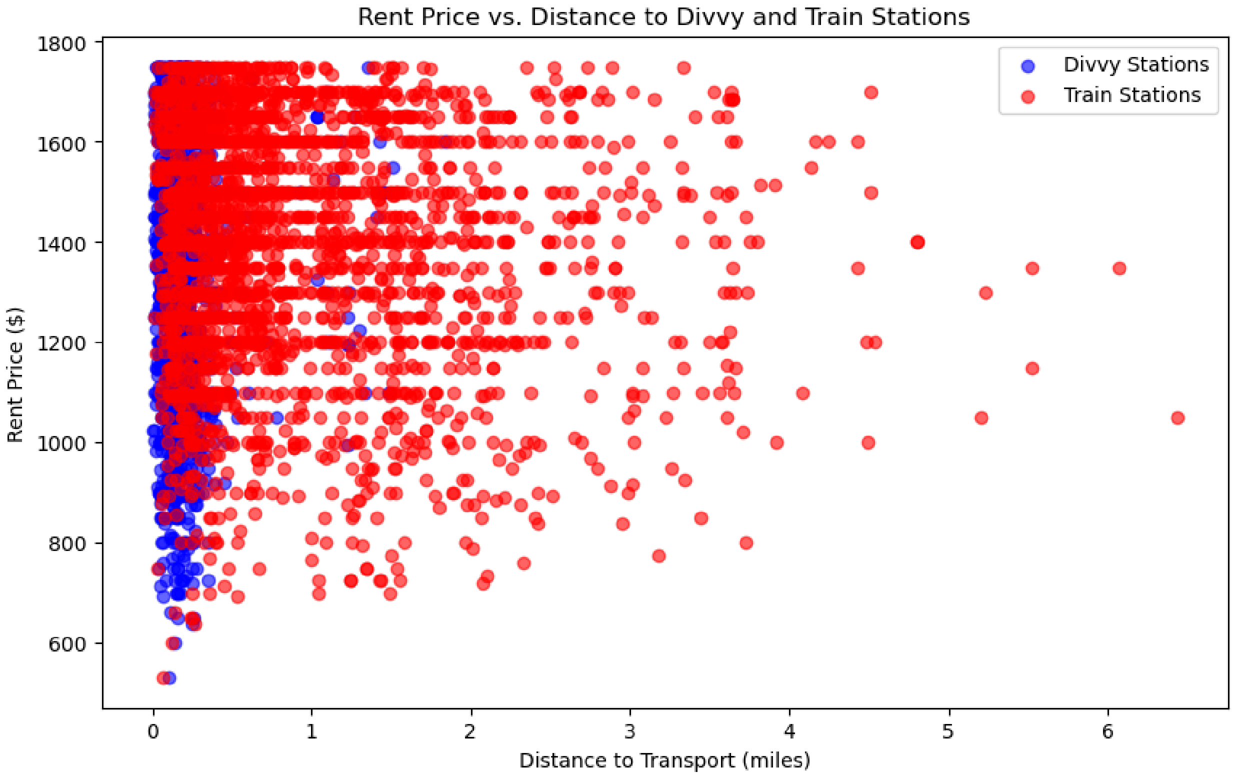
While it is common for the price of an apartment to increase as the number of bedrooms increase, I wanted to know what the median rental price for apartments in Chicago were based on the number of bedrooms. This is why I decided to use a boxplot to analyze the price alongside the number of bedrooms. When looking at the median price, it rises as the bed count increases, especially between 0 (studio), 1, 2, and 3 beds. Larger apartments tend to be more expensive, and there are fewer low priced options as the amount of bedrooms increase. Studio apartments have a wider price range when looking at the graph below. Rents for studios span from $500 up to

$1750. This suggests that 1–3 bed units are most common. Outliers exist within 2 bed and 3 bed listings and show that there are some listings that are cheaper than typical.



Machine Learning / Statistical Analysis 4 (Drew Vranicar)

Students often struggle to find housing that is both affordable and conveniently located near public transportation. I wanted to know if the distance to a nearby Divvy station or CTA train stop had any effect on how much an apartment rents for. To explore this, I looked at how rent price changes based on how far an apartment is from the nearest Divvy or train station. I created two new features: one measuring the distance to the closest Divvy station, and another to the closest train stop. These distances were calculated using the Haversine formula, which helps estimate how far two locations are from each other on a map. I then used a scatter plot to see if there was a visible trend between rental price and transport proximity. The goal was to build a simple regression model to find out if being closer to a Divvy or train station is linked to higher rent prices and whether students should prioritize transport access when looking for apartments.

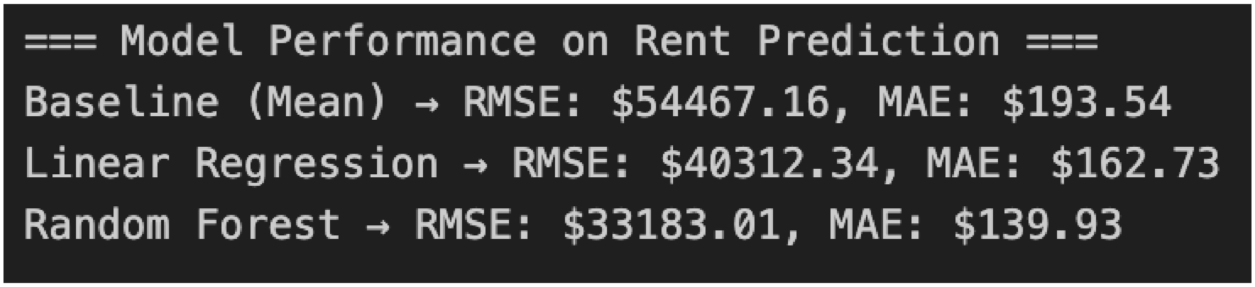


The plot shows wide variability in price regardless of distance, supporting the finding that transport alone is not enough to explain rent.

Machine Learning / Statistical Analysis 5 (Manav Kohli)

Hypothesis: A student's personal preferences, such as the Location of apartment listings and the number of bedrooms/bathrooms, can be accurately calculated and used to predict the prices of listings in Chicago.

For my Machine Learning Analysis, I used four features (Latitude, Longitude, Beds, Baths) to prove the aforementioned hypothesis. I started by ensuring all the numeric columns were in float and dropped outliers. I then proceeded to select the features and targets and do a train/test split. I then used 3 models (Baseline, Linear Regression, and Random Forest) to predict the rent, and evaluated each on how they had performed. The results were as follows:



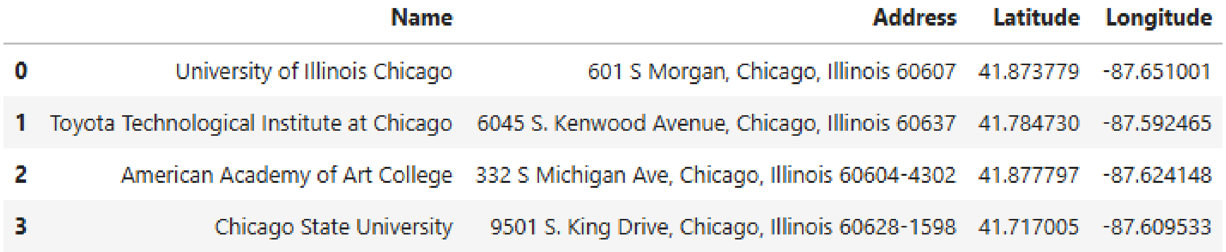
```
=== Model Performance on Rent Prediction ===
Baseline (Mean) → RMSE: $54467.16, MAE: $193.54
Linear Regression → RMSE: $40312.34, MAE: $162.73
Random Forest → RMSE: $33183.01, MAE: $139.93
```

A reasonable conclusion to draw from this is that Random Forest currently performs best and is likely benefiting from the non-linear effects of location and features. However, high RMSE suggests either outliers or unit scaling issues, which should be examined next.

Visualization

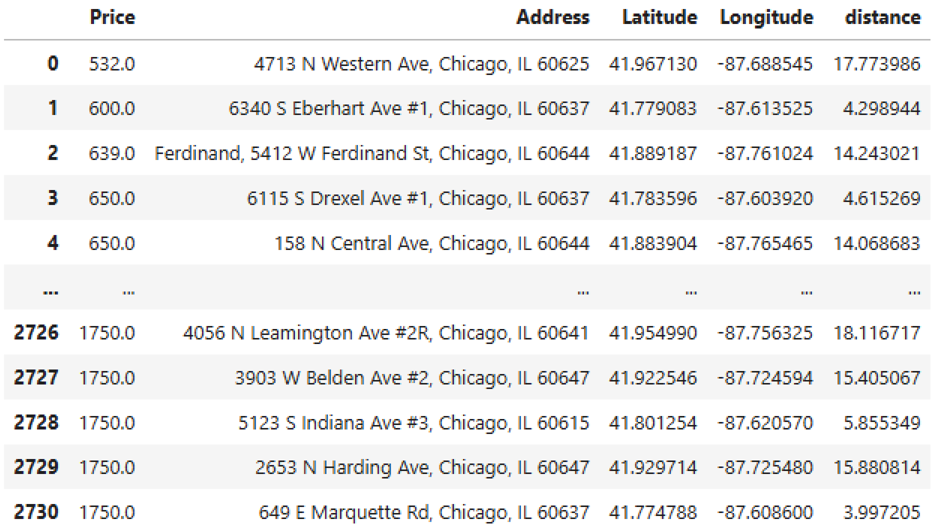
Visualization 1 (Tyler Yuen)

This visualization was based on the hypothesis in which apartments that are closer to the school in question are higher in price in comparison to apartments that are farther away from the school. In order to accomplish this, the haversine formula was used to get the distance in miles from each apartment to each university using their respective latitude and longitudes. I had four sample schools to test with.



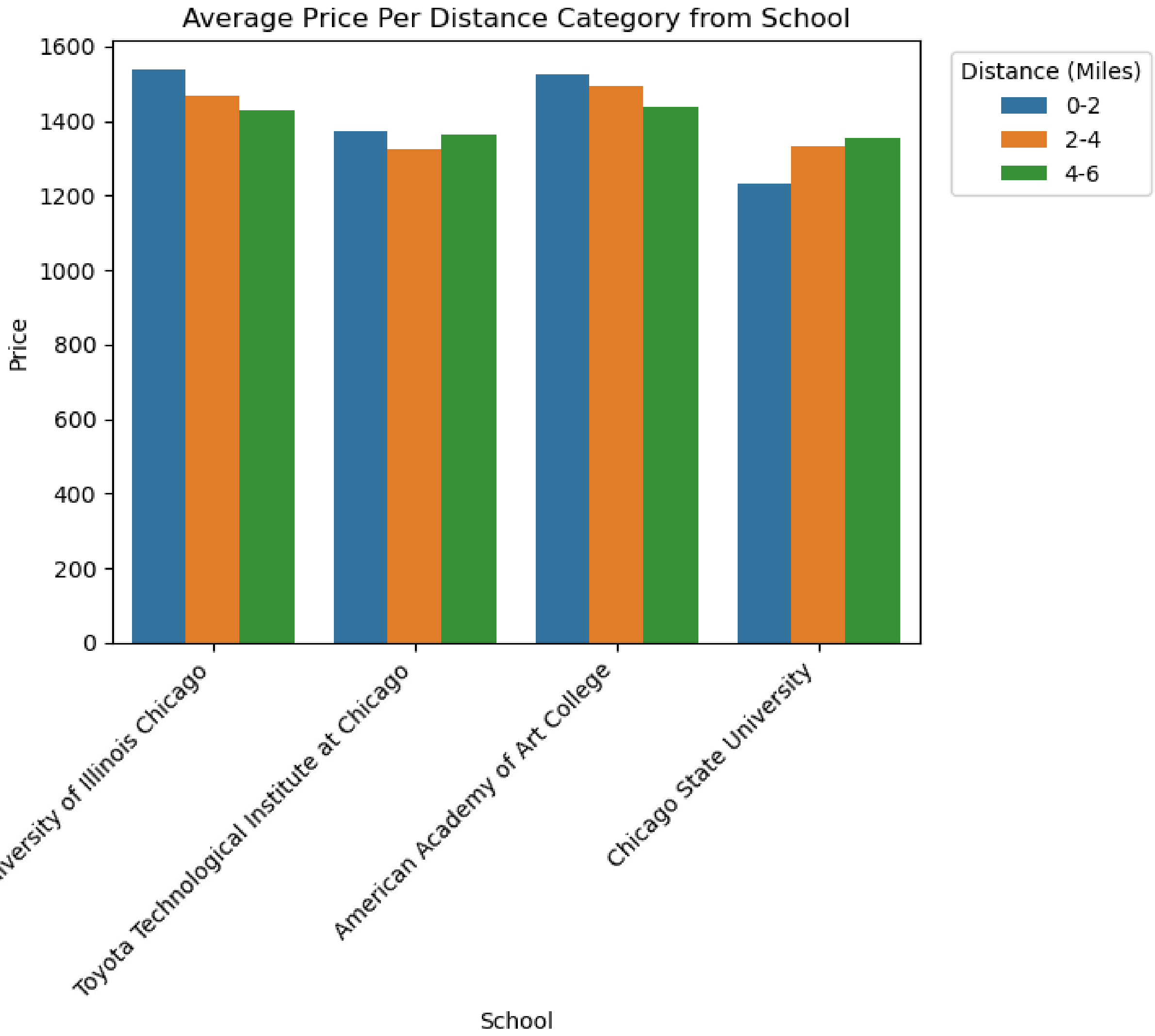
	Name	Address	Latitude	Longitude
0	University of Illinois Chicago	601 S Morgan, Chicago, Illinois 60607	41.873779	-87.651001
1	Toyota Technological Institute at Chicago	6045 S. Kenwood Avenue, Chicago, Illinois 60637	41.784730	-87.592465
2	American Academy of Art College	332 S Michigan Ave, Chicago, Illinois 60604-4302	41.877797	-87.624148
3	Chicago State University	9501 S. King Drive, Chicago, Illinois 60628-1598	41.717005	-87.609533

I then added those respective distances to a new column in the apartments listings dataframe.



	Price	Address	Latitude	Longitude	distance
0	532.0	4713 N Western Ave, Chicago, IL 60625	41.967130	-87.688545	17.773986
1	600.0	6340 S Eberhart Ave #1, Chicago, IL 60637	41.779083	-87.613525	4.298944
2	639.0	Ferdinand, 5412 W Ferdinand St, Chicago, IL 60644	41.889187	-87.761024	14.243021
3	650.0	6115 S Drexel Ave #1, Chicago, IL 60637	41.783596	-87.603920	4.615269
4	650.0	158 N Central Ave, Chicago, IL 60644	41.883904	-87.765465	14.068683
...	...	...	...	...	...
2726	1750.0	4056 N Leamington Ave #2R, Chicago, IL 60641	41.954990	-87.756325	18.116717
2727	1750.0	3903 W Belden Ave #2, Chicago, IL 60647	41.922546	-87.724594	15.405067
2728	1750.0	5123 S Indiana Ave #3, Chicago, IL 60615	41.801254	-87.620570	5.855349
2729	1750.0	2653 N Harding Ave, Chicago, IL 60647	41.929714	-87.725480	15.880814
2730	1750.0	649 E Marquette Rd, Chicago, IL 60637	41.774788	-87.608600	3.997205

With my data set, I determined average prices based on the proximities for each listing and our sampled schools. I arranged a new dataframe to show the schools and the average price of apartment listings in a short, medium, and long distance range (correlating to 0-2, 2-4, and 4-6 miles respectively).

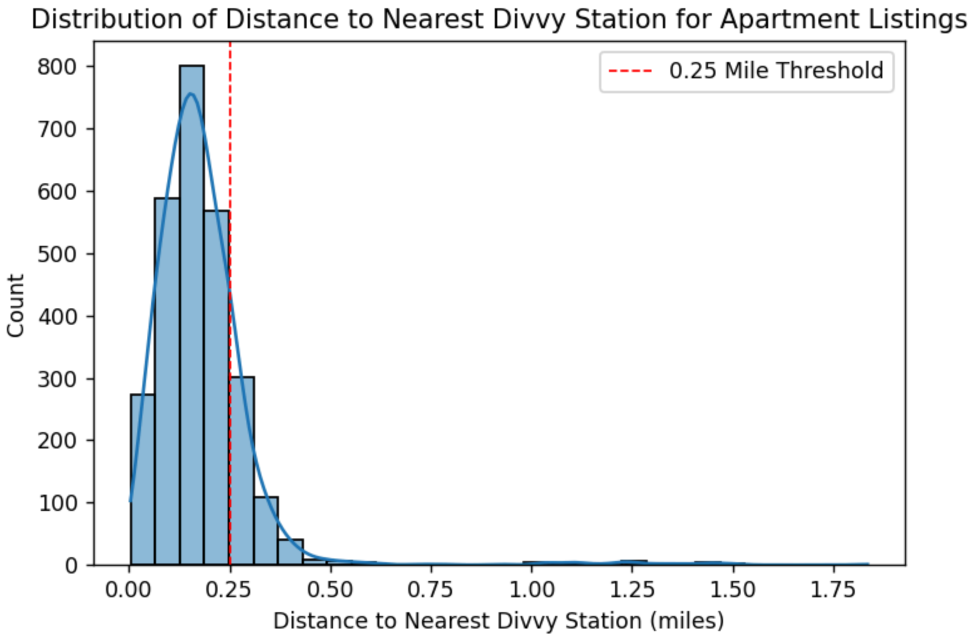


From here, we are presented with this grouped bar chart which shows that the 2-4 mile range is most likely the ideal candidate for cheaper locations. From our grouped barplot, we can see that there is a significant drop in price the farther you get away from campus for the University of Illinois Chicago and Columbia College Chicago. Even for Chicago State City Colleges of Chicago - Harry S Truman College, one can see that the 2-4 mile medium range apartment listings are a little bit cheaper than apartment listings in the 0-2 mile range. Although it does have a contrast in that the 4-6 mile range is significantly higher, thus being an outlier. Lastly, Chicago State University is the biggest outlier in that the cheaper apartment listings are actually closer to their campus as opposed to the farther apartment listings.

While it now seems to be made known that a 2-4 mile distance away from the school will in most cases be a lower price than a 0-2 mile distance, it is still far in terms of a commuting standpoint for a student. This is useful information to take into consideration as a fully polished recommendation system will use these 2-4 mile distanced apartment listings as notable mentions but not exactly top recommendations.

Visualization 2 (Tyler Yuen)

For my second visualization, I decided to look into one of the commute options that students may prefer, that being Divvy Bicycle Stations. Divvy bikes are probably one of Chicago's most useful commuter options for those who don't have a car, moreso for students. Therefore, it was a useful idea to consider for our recommendation system because a student may be looking for apartment listings where they are right next to a divvy bike station. Utilizing the cleaned Chicago apartment listing rentals data as well as the Divvy bicycle station data, I performed calculations on each apartment row using the haversine formula to find the approximate distance between the apartment and the closest divvy station. Then using seaborn, I plotted the result in a histogram which gave me the following.



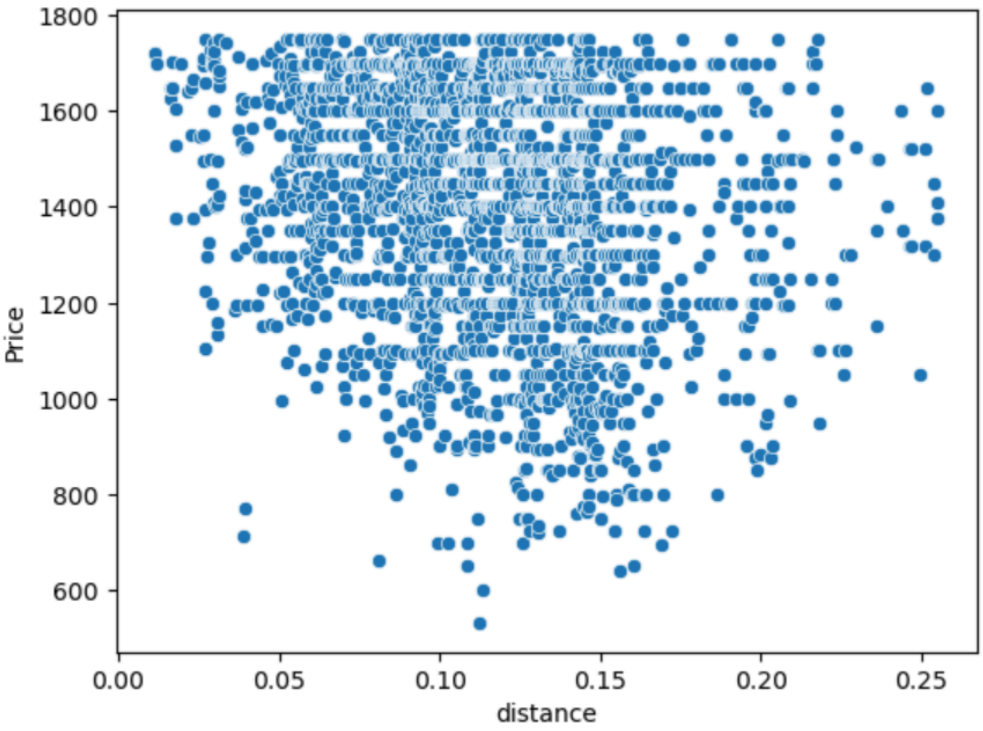
I had originally hypothesized that most divvy stations would be found at a 0.25 mile threshold (given the red line), but to my surprise, the closest station on average for about 800 apartment listings seems to be 0.125 miles away, a lot closer to the listings than I would have expected. This visualization is important because it helped as a reference as to how our recommendation tool should act. Examples could be that if a student had a preference for bicycling, it'll be more inclined to provide some of the 800 apartment listings that have a divvy bike station 0.125 miles (or less).

Visualization 3 (Brandon Kim)

Hypothesis 3: Apartments closer to Navy Pier tend to increase in price (Brandon Kim)

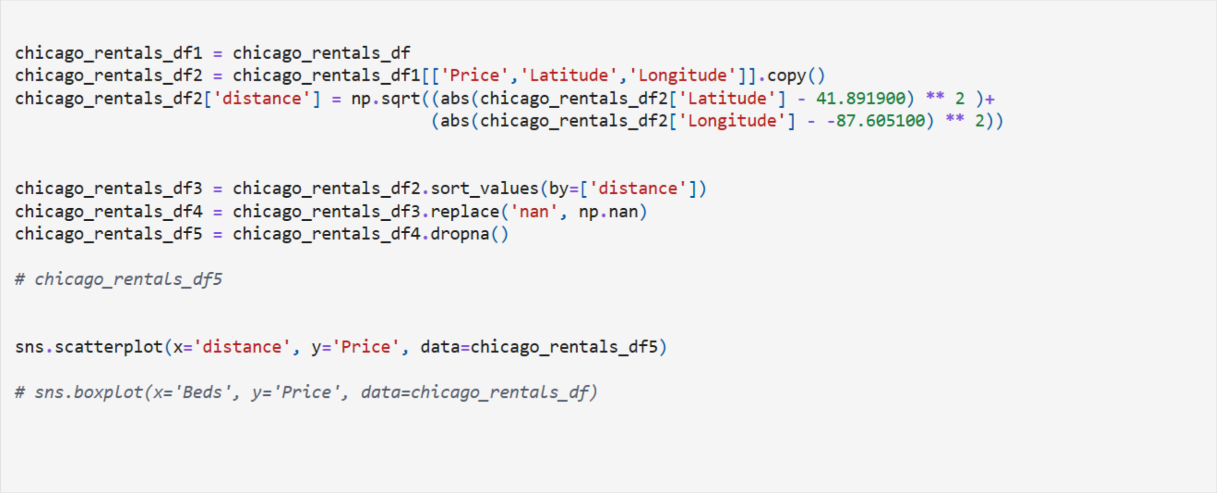
The coordinates for Navy Pier are 41.891900° latitude, -87.605100° longitude, according to Google. To calculate the "distance", I just used the distance formula on the coordinate values between the apartment coordinate values and Navy Pier's coordinates. While they aren't the most accurate I found the relative distance to be about the same even without accounting for the curvature of the Earth because all of the distances are relatively close to each other.

Here's a graph of the distance from the apartments to Navy Pier versus the price:



While there doesn't seem to be a strong correlation between distance from Navy Pier and the price of an apartment, we may be missing some information. For example, the price could be varied because of the many different types of apartments and what comes in each apartment(like rooms and how much sq feet), and these other features aren't accounted for in this graph.

Here’s the code for the visualization as in the Progress Report (Jupyter Notebook):



```
chicago_rentals_df1 = chicago_rentals_df
chicago_rentals_df2 = chicago_rentals_df1[['Price','Latitude','Longitude']].copy()
chicago_rentals_df2['distance'] = np.sqrt((abs(chicago_rentals_df2['Latitude'] - 41.891900) ** 2 )+
                                          (abs(chicago_rentals_df2['Longitude'] - -87.605100) ** 2))


chicago_rentals_df3 = chicago_rentals_df2.sort_values(by=['distance'])
chicago_rentals_df4 = chicago_rentals_df3.replace('nan', np.nan)
chicago_rentals_df5 = chicago_rentals_df4.dropna()

# chicago_rentals_df5


sns.scatterplot(x='distance', y='Price', data=chicago_rentals_df5)

# sns.boxplot(x='Beds', y='Price', data=chicago_rentals_df)
```

Visualization 4 (Brandon Kim)

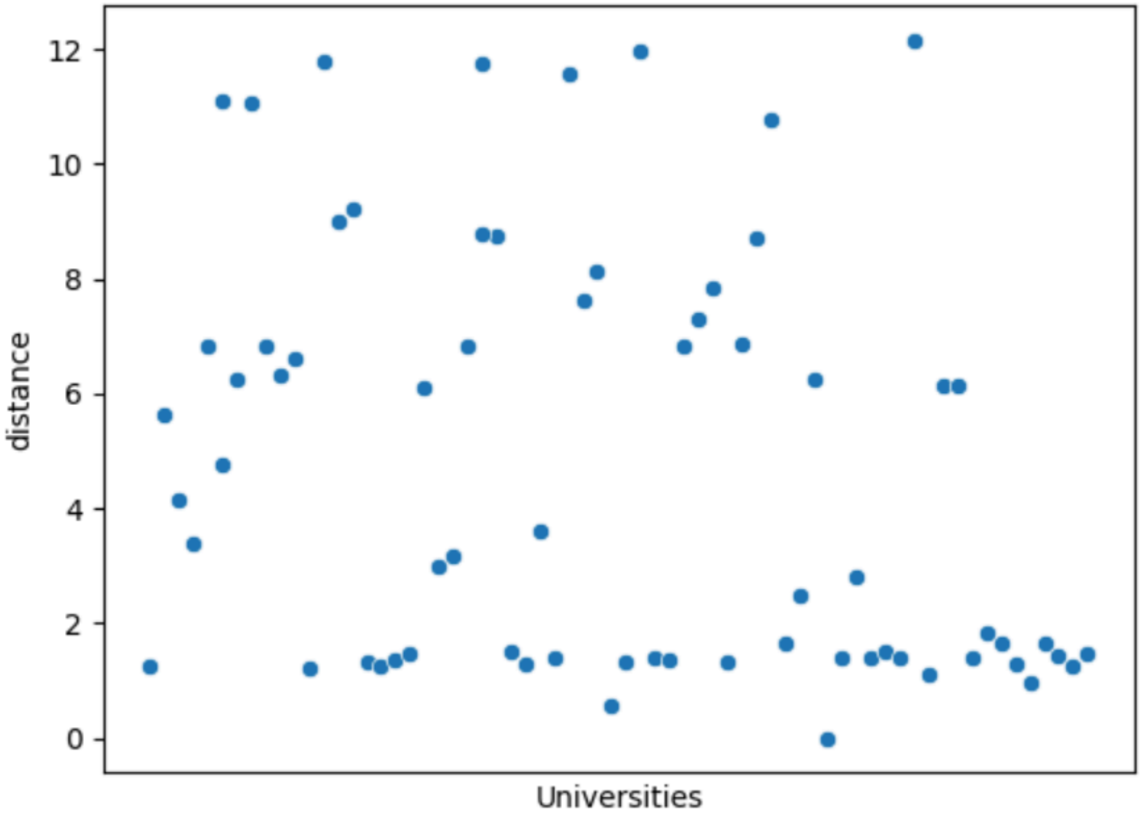
Hypothesis 4: Most Universities in Chicago are grouped near UIC (Brandon Kim)

The reason I thought this was the case was because many universities want to be located near the center of Chicago where there are the most people, apartments, and other places for human consumption. It would make sense that universities would group near the most populated parts of Chicago.

The coordinates for UIC are 41.873779° latitude, -87.651001 longitude, according to our dataset

To calculate the distance to UIC from every other university, this time I used the Haversine formula used/implemented by Tyler to add in a column next to every university, the distance in miles from UIC. This takes into account the curvature of the Earth leading to more accurate distances.

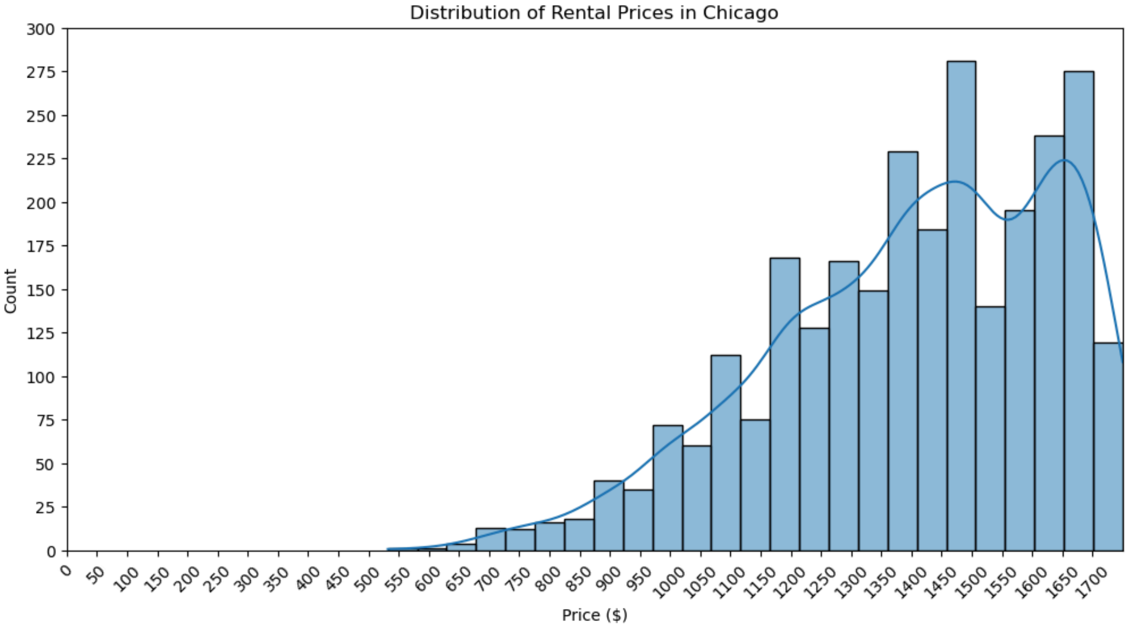
Here’s the graph showing the relationship between the Universities and the distances from UIC:



There seems to be more universities closer to UIC, and as the distance increases, the number of universities also seems to decrease. However, there aren’t enough universities close enough to UIC to conclude that most universities are grouped together in Chicago.

Visualization 5 (Cristal Martinez)

This visualization was based on comparing the different price distributions available within our apartments dataframe, chicago_rentals_df. Based on the 2,731 apartments that were extracted from Zillow, most rental listings were listed between $1475.00 and $1650.00. With a mean of $1403.00, the rental market at the moment contains multiple apartment listings that are at the upper end of a typical student's budget. The minimum price a current listing can go for is $532.00, however these low cost options are very limited. For higher priced rentals, a student may consider splitting the cost with roommates.



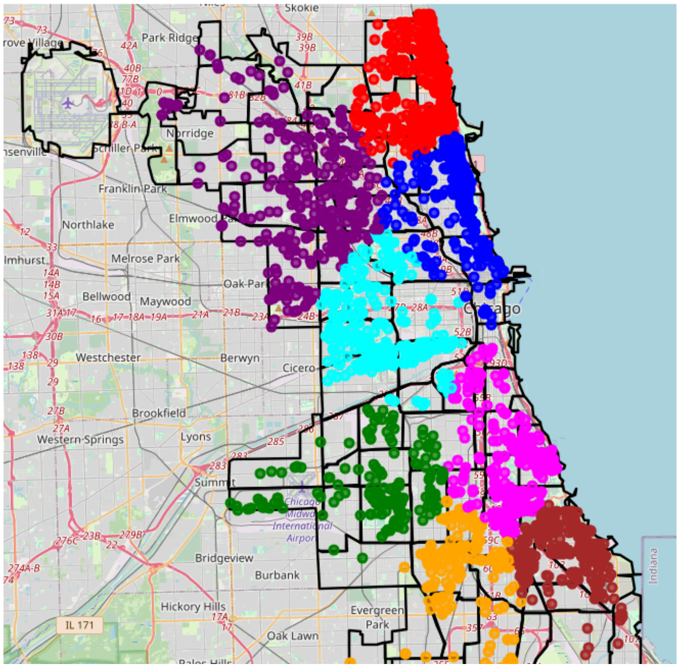
Visualization 6 (Cristal Martinez)

It was important for us to have an apartment database that is not clustered within one area of Chicago, but rather have apartment listings throughout the city. Even though the address was provided for the apartment listings, the latitude and longitude values offered the most accurate way to map and visualize the apartment locations throughout Chicago. The results from k-means clustering were similar to well known community areas within the city. For example, the map shows:

1. Cluster 0 (red): This cluster predominantly covers apartments located in the far north side of Chicago. These include neighborhoods such as Edgewater, Uptown, Lincoln Square, and Rogers Park.
2. Cluster 6 (blue): This cluster predominantly covers apartments within the lakefront areas of Chicago. These include neighborhoods such as Lakeview and Lincoln Park.
3. Cluster 3 (purple): This cluster predominantly covers apartments located in the north-west side neighborhoods of Chicago. These include neighborhoods such as Humboldt Park, Irving Park, and Austin.
4. Cluster 5 (cyan): This clusters span central Chicago. These include neighborhoods such as portions of Near West Side and Pilsen.

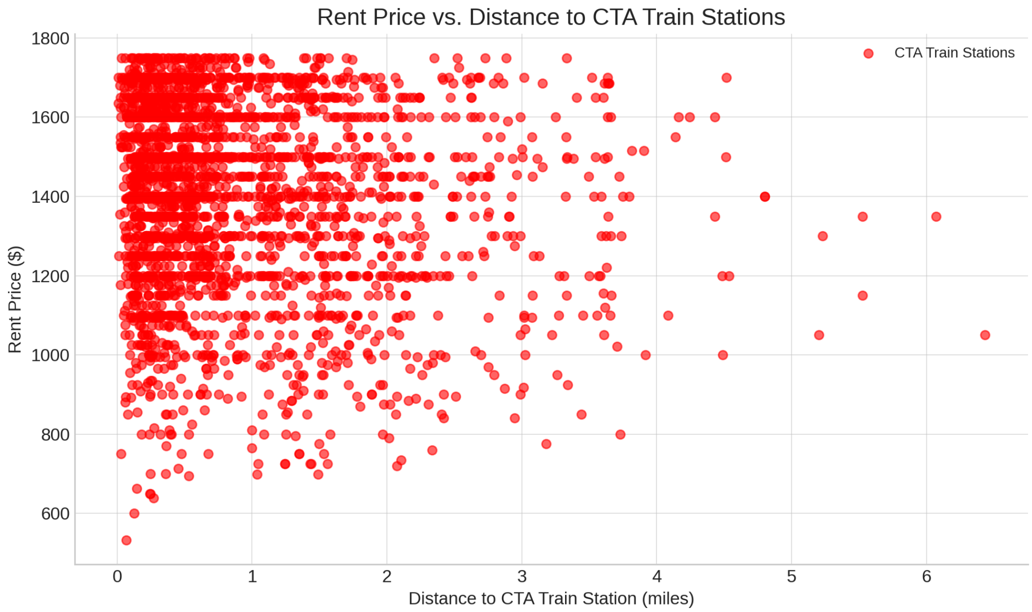
5. Cluster 1 (magenta): This cluster covers apartments in the Hyde Park and mid south regions. Close to the University of Chicago.
6. Cluster 2 (green): This cluster primarily covers apartments located in the south-west neighborhoods of Chicago. These include neighborhoods such as Chicago Lawn, West Lawn, and Ashburn.
7. Cluster 4 (orange): This cluster predominantly covers the far south Side of Chicago. These include neighborhoods such as Greater Grand Crossing, Chatham, Auburn Gresham, and Washington Heights.
8. Cluster 7 (brown): This cluster predominantly covers the south-east side of Chicago. These include neighborhoods such as South Shore, Greater Grand Crossing, and portions of Catham.

Ultimately, the map displays a well distributed amount of apartments throughout popular neighborhoods within the far north side, Lakefront, north-west side, central Chicago, mid-south, south-west, far south side, and southeast side regions. By knowing that the apartments are spaced throughout the city of Chicago's neighborhoods, we were confident to continue using our apartment database, as it captured a diverse range of apartment options that are not clustered within one area of Chicago.



Visualization 7 (Drew Vranicar)

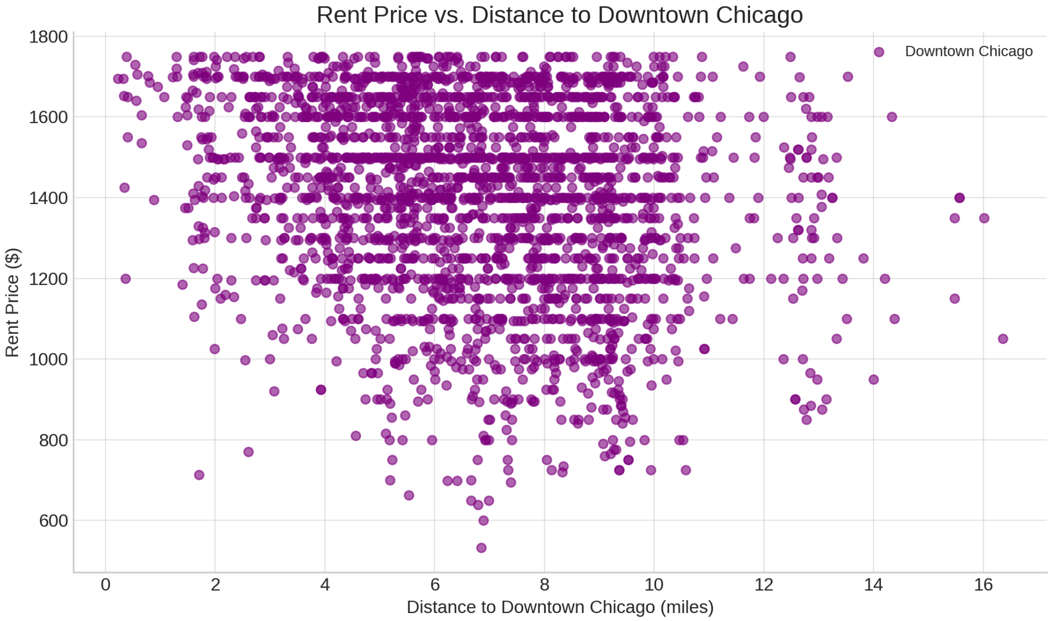
Hypothesis: Apartments closer to CTA train stations tend to be more expensive.



This scatterplot shows a concentration of higher rent prices within approximately 1 mile of CTA train stations. As the distance from train stations increases, rent prices tend to decline, although the trend is not perfectly linear. This suggests a general preference or premium placed on apartments that offer convenient access to public transportation. Proximity to CTA stations likely appeals to commuters who rely on trains for daily travel, which increases demand and thus rent near these locations. Overall, the data supports the hypothesis that apartments closer to CTA train stations tend to be more expensive, although other factors may also influence rent values.

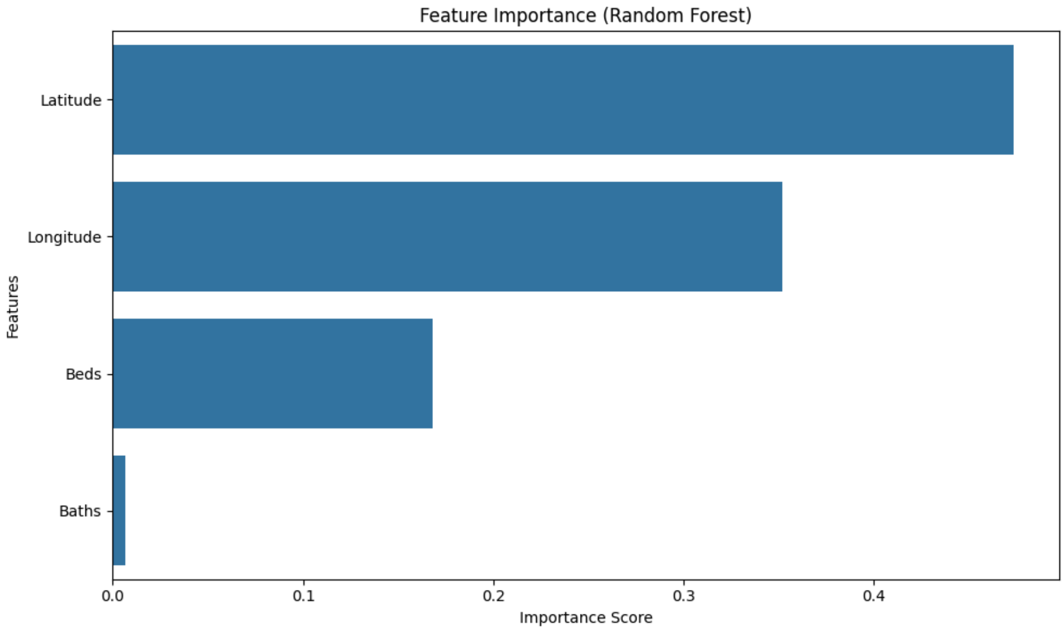
Visualization 8 (Drew Vranicar)

Hypothesis: Apartments closer to downtown Chicago (Loop area) tend to have higher rent.



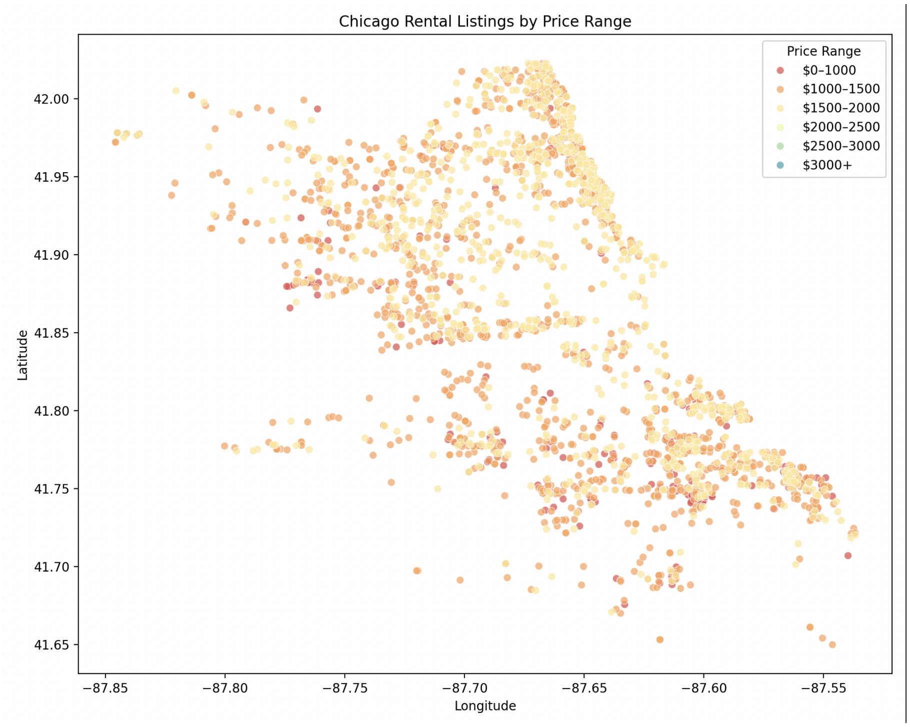
This scatterplot reveals a stronger inverse relationship between rent prices and distances to downtown Chicago. The majority of high rent apartments are clustered within 5 miles of downtown Chicago with rents gradually decreasing as distance increases. This aligns with economic and lifestyle factors with downtown areas typically offering greater access to jobs, entertainment, and amenities by making nearby housing more desirable and competitive. Overall, the data clearly supports the hypothesis that apartments closer to downtown Chicago tend to have higher rents.

Visualization 9 (Manav Kohli)



Feature importance for the Random Forest Model in Rent Prediction

Visualization 10 (Manav Kohli)



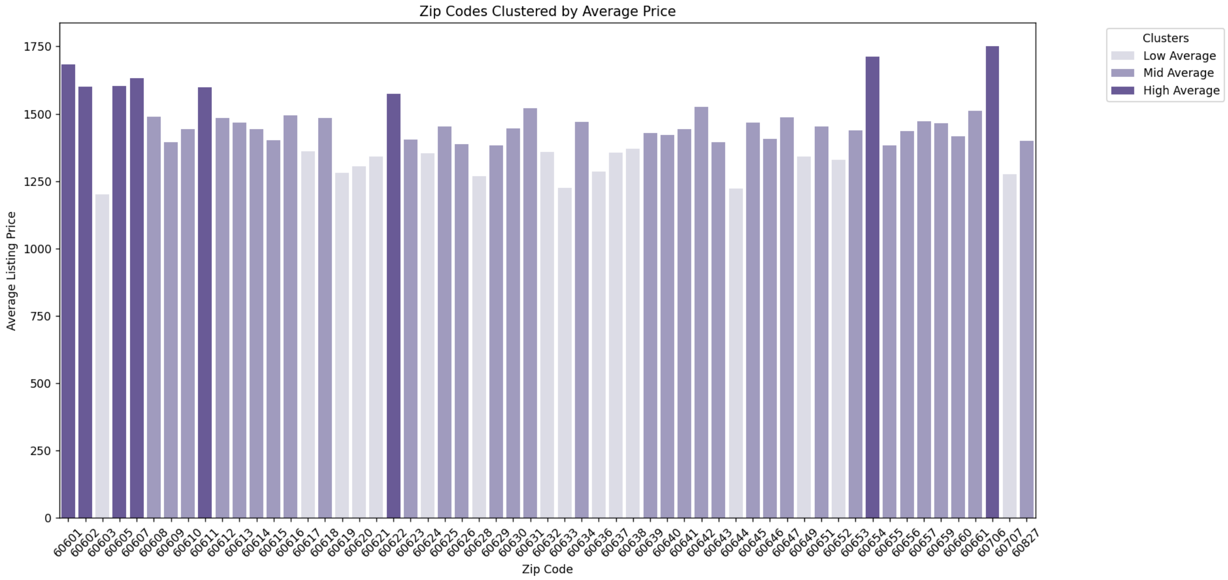
A simple scatter plot showcasing Latitude against Longitude and coloring the points according to the price points they are in

Additional Work

Extra Machine Learning / Statistical Analysis (Tyler Yuen)

For our additional work, we decided to investigate one of the questions we had earlier, “Which ZIP codes have the highest concentration of apartments?” For that, we had looked at the amount of apartment listings per zip code, but we never explored the pricing that each listing in each zip code had. This is important to our recommendation system because if a student input preferences that had a university in proximity to one of the cheaper zip code areas, then our recommendation system should in theory be more inclined to provide the cheaper apartments in the zip code (if the student has that same price range preference, or is focusing on cheapest apartments).

To accomplish this, I had decided to use Kmeans clustering in order to cluster zip codes into 3 categories, Low Average (Price), Mid Average (Price), and High Average (Price). To do this, I used the chicago rentals dataframe, as seen in our progress report, and used a StandardScaler (from sklearn) to fit our Price column. After transforming, I was able to run Kmeans over them and predict each apartment listing to a specific cluster. The result we get is seen below.



As one can see, the darker purple highlights the higher expensive areas, the mild purple highlights the mid range areas, and the light purple (almost gray) highlights the cheaper areas. We can infer that from this, if a student inputs a cheaper price range or simply sorts by cheapest price, they will most likely see more options from the light purple zip code listings.

Results

In conclusion, we were able to perform extensive data analysis on our data and have gotten close to a feasible recommendation system for students looking for reasonable housing. Through the visualizations and statistical analyses that Tyler performed, we were able to uncover various distance and preference-based methods to approximate favorable listings such as favoring apartments that fit a certain similarity score with students preferences and looking at zip code and mile range average prices. His analysis helped create the backbone of a possible recommendation system. Brandon's visualizations supported the system in supporting the idea that apartments closer to landmarks are more likely to be expensive and as a result, should not be the first choice. Cristal's visualization and analysis helped understand that for a student's best choice, it would be wise to have a smaller amount of bedrooms to lighten the price. Along with that, her visualizations on neighborhood clustering and price range analysis solidified our data and made us more confident in the data we were using and how to approach our recommendations. Drew's visualizations help us understand that living downtown and specifically by train stations will result in a higher cost which will most likely lessen the chance of those types of apartments within our system. Lastly, Manav made a rent prediction model which helps students predict what type of apartment listing (location, etc.) they should choose from when deciding in the future and a good geographic location map to support it.

Overall, the visualizations themselves explain the very core of our recommendations, showing how there is a correlation between the numbers of beds and baths to price. Also, they explain that certain neighborhoods as well as proximity to the university are more likely to influence the price as well. While a fully built tool is not something that has been accomplished due to time and money constraints (some of the tool capabilities we wanted to implement were blocked by rate limits or high pricing) we believe that through our data analysis, our main point that we could create a tool to help students find reasonable housing has been proved. With some investment and more resources, our insight itself is close to a fully polished product.