Final Capstone Projection

Objective:

As jobs trends towards being more remote, people have had the opportunities to work from pretty much anywhere in the world as long as they can manage their time. With this influx of remote workers another industry has begun to capitalize on this shift. That industry being hospitality, more specifically airbnb listings. Many people look to add additional income to their primary source of income through means of short-term rentals. However, there is a daunting fear of this not panning out for the individual who decides to get into short-term leasing contracts with those seeing a place to stay. If there are not enough bookings within a month many times the host may take a loss in revenue due to monthly ongoing property expenses, rent not being the least of them. This project is aiming to create a regression model that can predict the booking percentage of an airbnb listing given a set of features that can be found on airdna.

Data Wrangling:

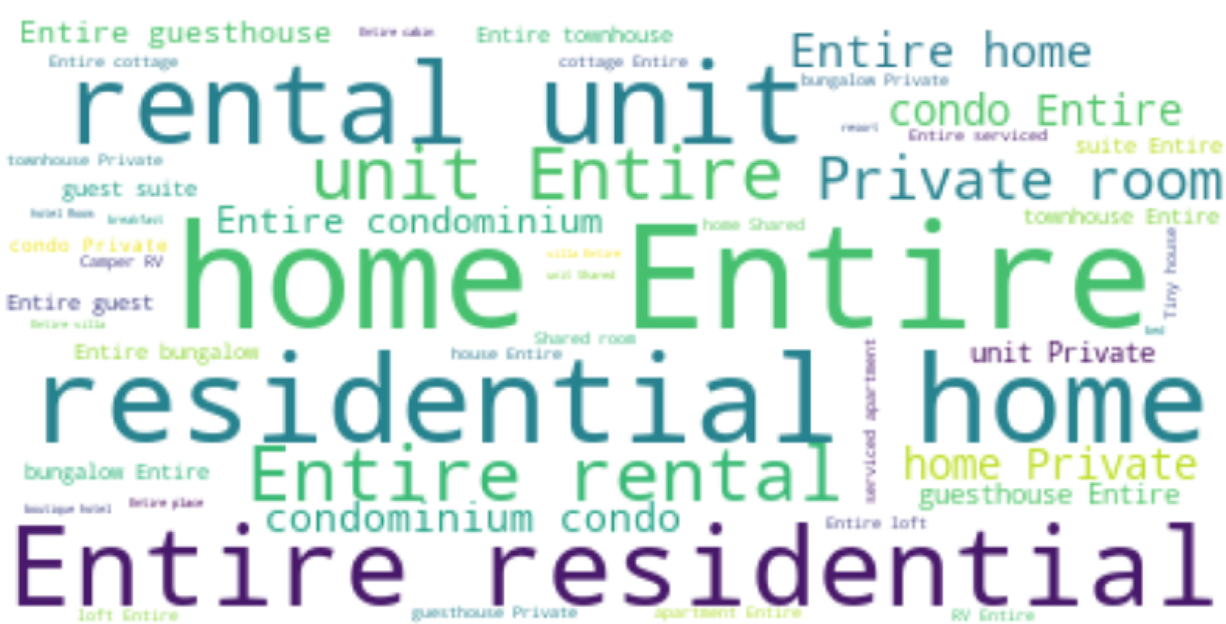
The data pulled from airdna is centered around the Austin area and consists of over ~12K records for 2022. These initial fields for this data source are listed below.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11972 entries, 0 to 11971
Data columns (total 74 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   id                       11972 non-null  int64
 1   listing_url              11972 non-null  object
 2   scrape_id                11972 non-null  int64
 3   last_scraped             11972 non-null  object
 4   name                     11972 non-null  object
 5   description              11808 non-null  object
 6   neighborhood_overview    7059 non-null   object
 7   picture_url              11971 non-null  object
 8   host_id                  11972 non-null  int64
 9   host_url                 11972 non-null  object
 10  host_name                11969 non-null  object
 11  host_since               11969 non-null  object
 12  host_location            11954 non-null  object
 13  host_about               7293 non-null   object
 14  host_response_time       8523 non-null   object
 15  host_response_rate       8523 non-null   object
 16  host_acceptance_rate     9110 non-null   object
```

17	host_is_superhost	11969 non-null	object
18	host_thumbnail_url	11969 non-null	object
19	host_picture_url	11969 non-null	object
20	host_neighbourhood	10254 non-null	object
21	host_listings_count	11969 non-null	float64
22	host_total_listings_count	11969 non-null	float64
23	host_verifications	11972 non-null	object
24	host_has_profile_pic	11969 non-null	object
25	host_identity_verified	11969 non-null	object
26	neighbourhood	7059 non-null	object
27	neighbourhood_cleansed	11972 non-null	int64
28	neighbourhood_group_cleansed	0 non-null	float64
29	latitude	11972 non-null	float64
30	longitude	11972 non-null	float64
31	property_type	11972 non-null	object
32	room_type	11972 non-null	object
33	accommodates	11972 non-null	int64
34	bathrooms	0 non-null	float64
35	bathrooms_text	11956 non-null	object
36	bedrooms	11261 non-null	float64
37	beds	11822 non-null	float64
38	amenities	11972 non-null	object
39	price	11972 non-null	object
40	minimum_nights	11972 non-null	int64
41	maximum_nights	11972 non-null	int64
42	minimum_minimum_nights	11971 non-null	float64
43	maximum_minimum_nights	11971 non-null	float64
44	minimum_maximum_nights	11971 non-null	float64
45	maximum_maximum_nights	11971 non-null	float64
46	minimum_nights_avg_ntm	11971 non-null	float64
47	maximum_nights_avg_ntm	11971 non-null	float64
48	calendar_updated	0 non-null	float64
49	has_availability	11972 non-null	object
50	availability_30	11972 non-null	int64
51	availability_60	11972 non-null	int64
52	availability_90	11972 non-null	int64
53	availability_365	11972 non-null	int64
54	calendar_last_scraped	11972 non-null	object
55	number_of_reviews	11972 non-null	int64
56	number_of_reviews_ltm	11972 non-null	int64
57	number_of_reviews_l30d	11972 non-null	int64
58	first_review	9026 non-null	object
59	last_review	9026 non-null	object
60	review_scores_rating	9026 non-null	float64

```
 61  review_scores_accuracy                  8954 non-null   float64
 62  review_scores_cleanliness               8954 non-null   float64
 63  review_scores_checkin                   8953 non-null   float64
 64  review_scores_communication             8954 non-null   float64
 65  review_scores_location                  8952 non-null   float64
 66  review_scores_value                     8952 non-null   float64
 67  license                                 0 non-null      float64
 68  instant_bookable                        11972 non-null  object
 69  calculated_host_listings_count          11972 non-null  int64
 70  calculated_host_listings_count_entire_homes   11972 non-null  int64
 71  calculated_host_listings_count_private_rooms  11972 non-null  int64
 72  calculated_host_listings_count_shared_rooms   11972 non-null  int64
 73  reviews_per_month                       9026 non-null   float64
dtypes: float64(24), int64(18), object(32)
memory usage: 6.8+ MB
```

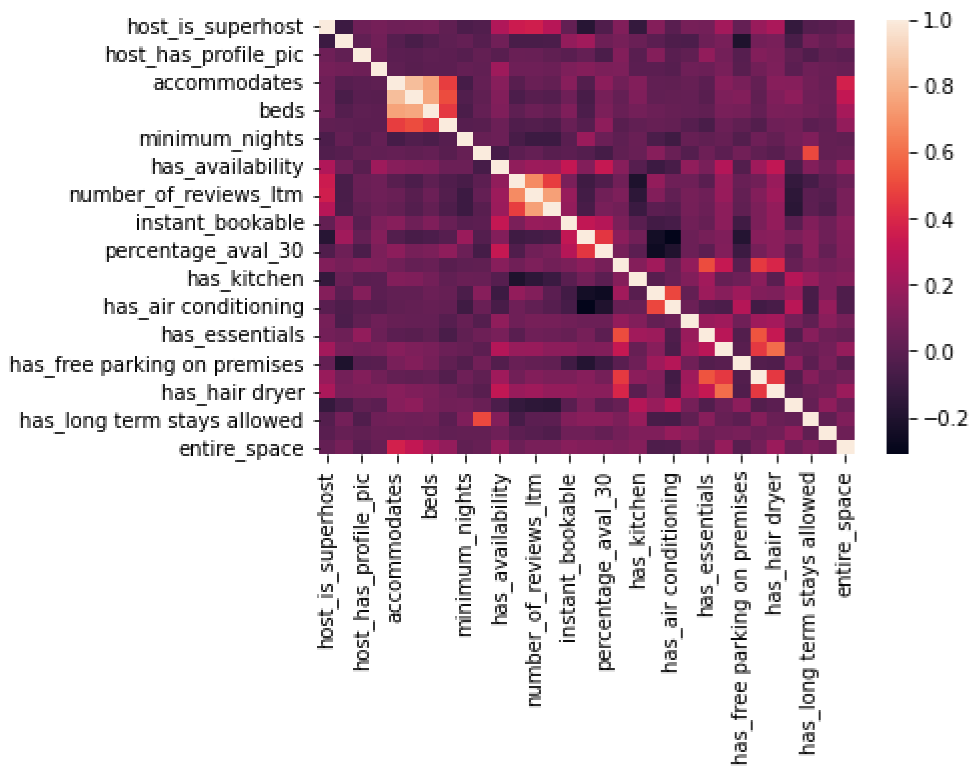
However, we can’t use all these fields in the machine learning model (object/string) especially in their current form. What was decided next was to further analyze some of the object data-type columns to glean some insight about airbnb’s in Austin. A visualization that was applied was word clouds which provided visual cues on the frequency of certain words used (i.e. property_type)



From looking at this data it was pretty clear that popular residences usually offered their entire residence versus a private space such as a room or section of the property.

After identify the valuable information from the string feature types the next step was to get the top 10 of the most frequent terms used for the features (Property type, Name, Description, etc.) and create a categorical representation of them using one hot encoding.

Next step was to identify if there were any strong linearity with the data features, specifically between our target variable and any of the independent variables (not including location variables).



Interestingly it appears that amenities and how easy it is to book may be good indicators about a listing's potential booking capacity.

Modeling

Given the nature of our objective we used linear regression and variants of linear regression to best fit the data. One thing to note when applying various metrics to measure the predictive power of our models we decided to go with r2_score to measure models ability to handle variance within our test data, and mean absolute error to keep track of on average how off are our predictions.

To ensure that we are moving in the right direction and can say that we at least perform better than a model predicting the average of the data we a dummy regressor was created and the following were the metrics associated with this mode (test data)l:

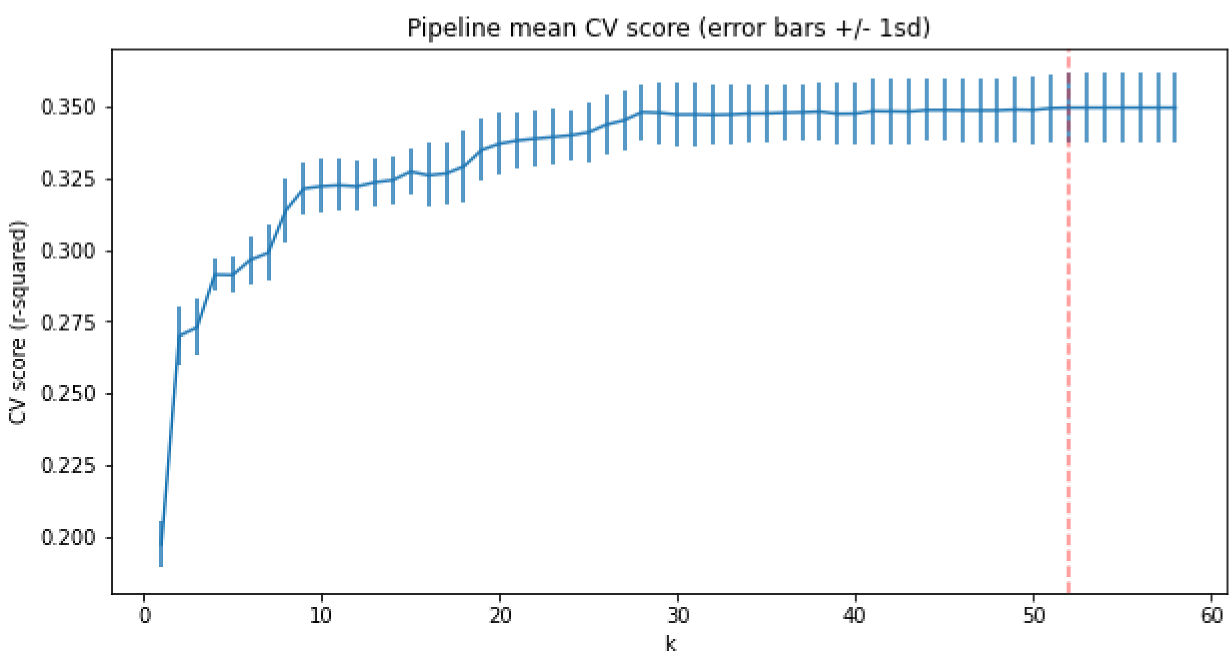
R2 score: 9.53773203222763E-05
MAE: 0.27242263977780000

One thing to note is the negative score of the r2 score which indicates that the residual errors from our predictions are way off and there are possibly more changes we can make to improve our model.

The next model that was tried was a vanilla Linear Regression model that utilized all 58 features in the training data. While providing more promising results, there could be a better optimization of the model's predictive power and performance. Namely, applying regularization to the features using hyper-parameter optimization via Ridge or Laso regression. After applying regularzation of the independent variables it was concluded that Ridge had the best overall r2 score and mae.

Metrics

After performing hyper-parameter optimization it was concluded that 52 out of the 58 features provided the most significant predictive power for our model regarding variance.



The following table provides the results of these experiments with the best model being in bold:

Model Name	R2 Score (train)	R2 Score (test)	MAE (test)	MAE (train)	Features
Dummy Regression	0	-9.53773203222763E-05	0.28	0.27	58
Linear Regression	0.36	0.37	0.2	0.2	58
Laso Regression	0.36	0.37	0.2	0.2	52
Ridge Regression	0.36	0.37	0.2	0.2	52

Conclusion

After performing analysis on the various models it is concluded that **Ridge Regression** provides the best predictive power for our data. Although the scores can be better, given the nature of this project, scores within these ranges are acceptable for a social science project, although some improvements can be made.