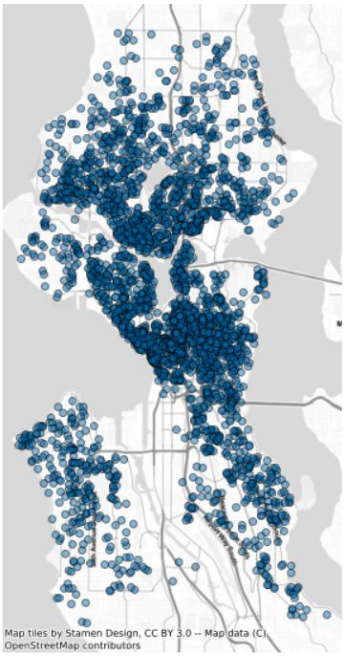
PSTAT 134 Final Project Proposal

Group Members:
Meilin Shi, Pedro Aristizabal, Jensen Newlin, Ao Yu Hsiao.

Proposed data set:
https://www.kaggle.com/airbnb/seattle
The data we will use is called “Seattle Airbnb Open Data” from Kaggle. The data is divided into 3 data sets, describing the different available **Airbnb listings** across Seattle since 2008. In them, we see the different **prices** for the listings, how these may change over different times of the year, as well as **customer reviews and descriptions** of the different listings. Additionally, there is also supplementary data regarding the neighborhoods in which they are located, such as the availability of public transport near the listings, landmarks to sightsee and experience close by, as well as public information of the hosts for each listing.
On the right is a visualization of our data set using geopandas package. Each point is an entry in Airbnb listings, total number around 3800.



Project Outline & Goals

1. Clustering of airbnbs in the city
    - Geographical location of Airbnbs
    - Where are the most popular destinations?
2. Checking the “vibe” of different neighborhoods
    - Keyword extraction from neighborhood overviews
    - Popular topics in reviews per neighborhood
3. How do prices change over the year? By neighborhood?
    - Finding average price changes

Overall Goals:
The group wants to analyze this data set in a way that could help people identify the Airbnb listing to stay in, based on their budget and preferences.

Questions to Consider:

- How will the rise in popularity for the Airbnb app from 2008 to today affect our research?
- What preferences will we focus on when finding an ideal listing?

Potential ideas:

- How does price change as time passes
- Comparing reviews between neighborhood
- Text mining to determine the quality of Airbnb
- When to visit? Best time of the year?
- How do clusters of Airbnb qualities relate to neighborhoods

Introduction

In recent years, with the popularity of Airbnb increases, Airbnb listings review has been used to evaluate perceptions of people toward neighborhoods. This kind of work can help people decide travel destinations or place to live. With the listing information, the numeric ratings and user reviews, we can study the perceptions of people toward the neighborhoods. In this project, we use the *Seattle Airbnb Open Data*, available on Kaggle (https://www.kaggle.com/airbnb/seattle). We extract the top keywords from five selected neighborhoods in Seattle from listings overview and user review. We also want to find the most popular destinations of Seattle and when is the best time to visit Seattle by looking at the booking rate and price changes over the years. To extract keywords, we use bigrams within NLTK and spaCy model for Named Entity Recognition in neighborhood overview, TF-IDF and topic modeling for user reviews. The dataset we used is divided into 3 parts, describing the different available Airbnb listings across Seattle since 2008. This dataset includes 3818 unique Airbnb listings in Seattle and their user reviews. This work can also be applied to neighborhood reviews in other cities for future study.

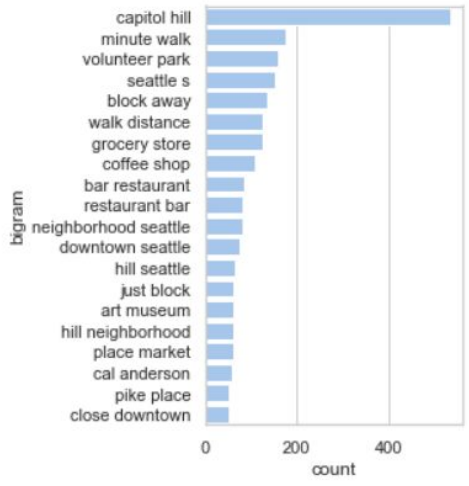
Listings Overview

To provide an overview of our data and study area, below is a visualization of our total Airbnb listings.

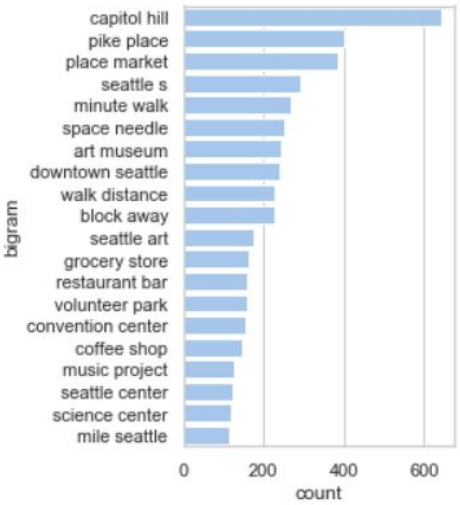
Data

The neighborhood data frame consists of 3818 rows, which means there are 3818 unique listings. We select the "id", "latitude", "longitude", "neighborhood_overview", "neighbourhood", and "neighbourhood_group_cleansed" columns from the full dataset. To visualize the listings by neighborhood, we decided to use the "neighborhood_group_cleansed" column, because it has the least categories, so that we can get a better display on the map.

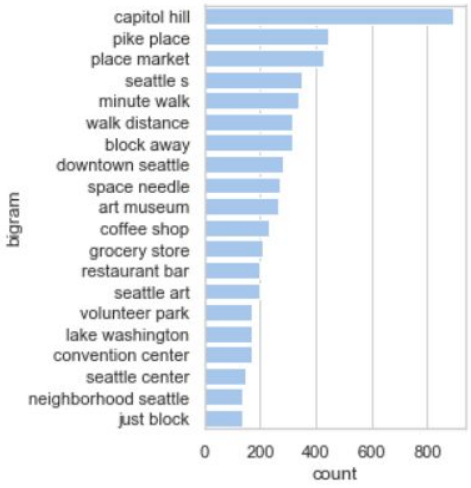
Capitol Hill:



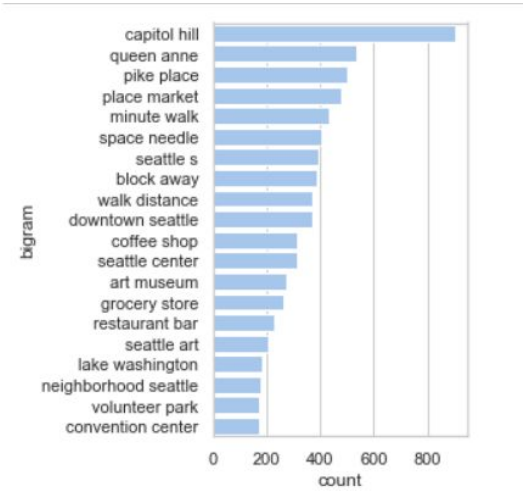
Downtown:



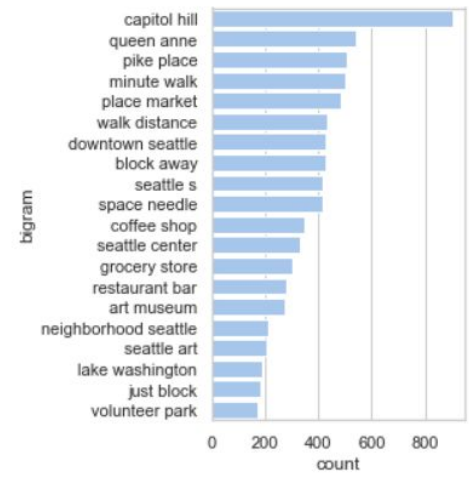
Central Area:



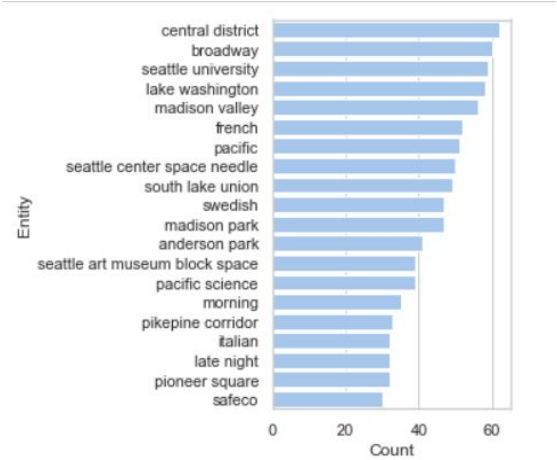
Queen Anne:



Ballard:



Capitol Hill NER:



Bi-gram

Bi-gram is a special case of the N-gram Model for n = 2,

to show a sequence of two adjacent words from tokens.

By examining the bi-gram, we can find out which two words

commonly co-occur in the overviews of selected

neighborhoods. Here we show the top 20 bigrams

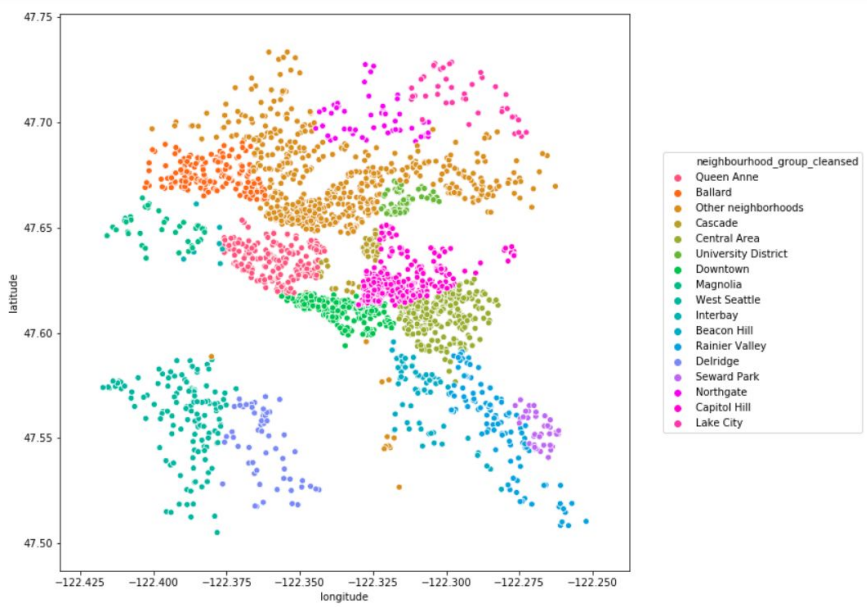
for each neighborhood.

spaCy (https://spacy.io/)
The spaCy model is designed for Named Entity Recognition in Natural Language Processing. It can detect Parts-of-Speech (PoS) Tagging and provide entity text and entity labels. In this project, we select the entity labels of:

- 'ORG' for Companies, agencies, institutions, etc.
- 'LOC' for Non-GPE locations, mountain ranges, bodies of water.
- 'FAC' for Buildings, airports, highways, bridges, etc.
- 'GPE' for Countries, cities, states.

Here we include the visualization of NER in the Capitol Hill neighborhood. We can also compare this with the bigram method.

Listings color coded by neighborhood district



Capitol Hill WordCloud:

THESE ARE THE NEIGHBORHOODS WE ARE WORKING ON GUYS

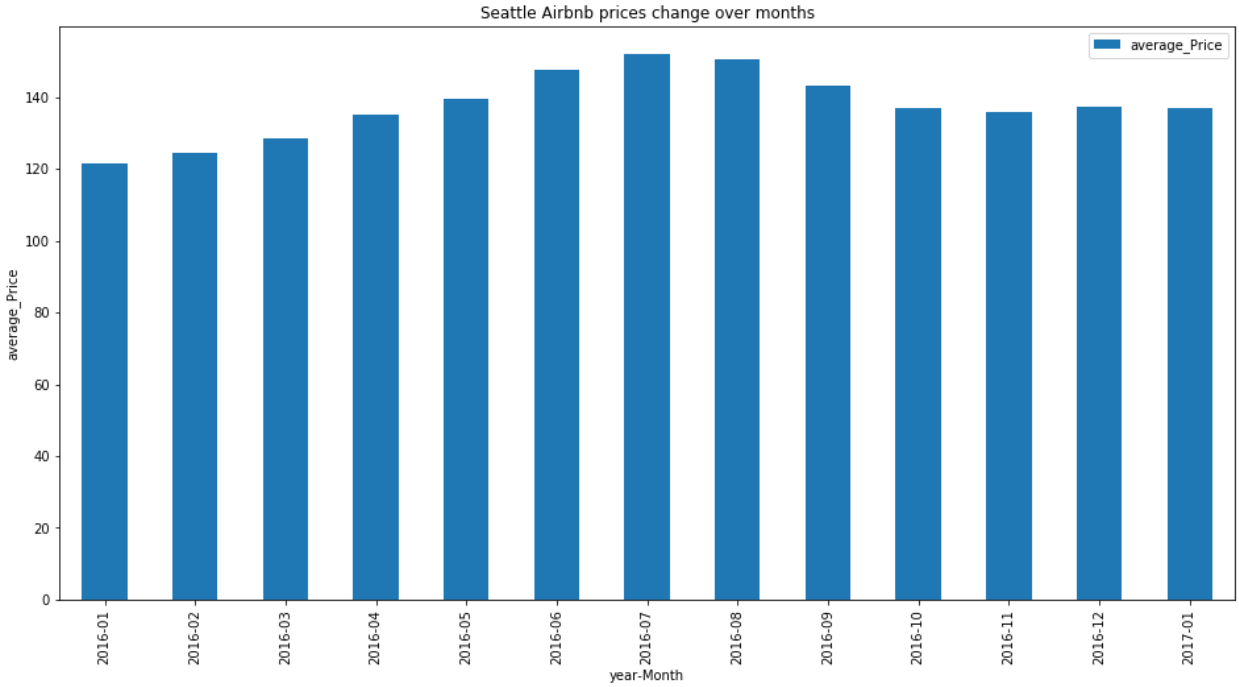
Queen_Anne = neighborhood[neighborhood['neighbourhood_group_cleansed']=='Queen Anne']

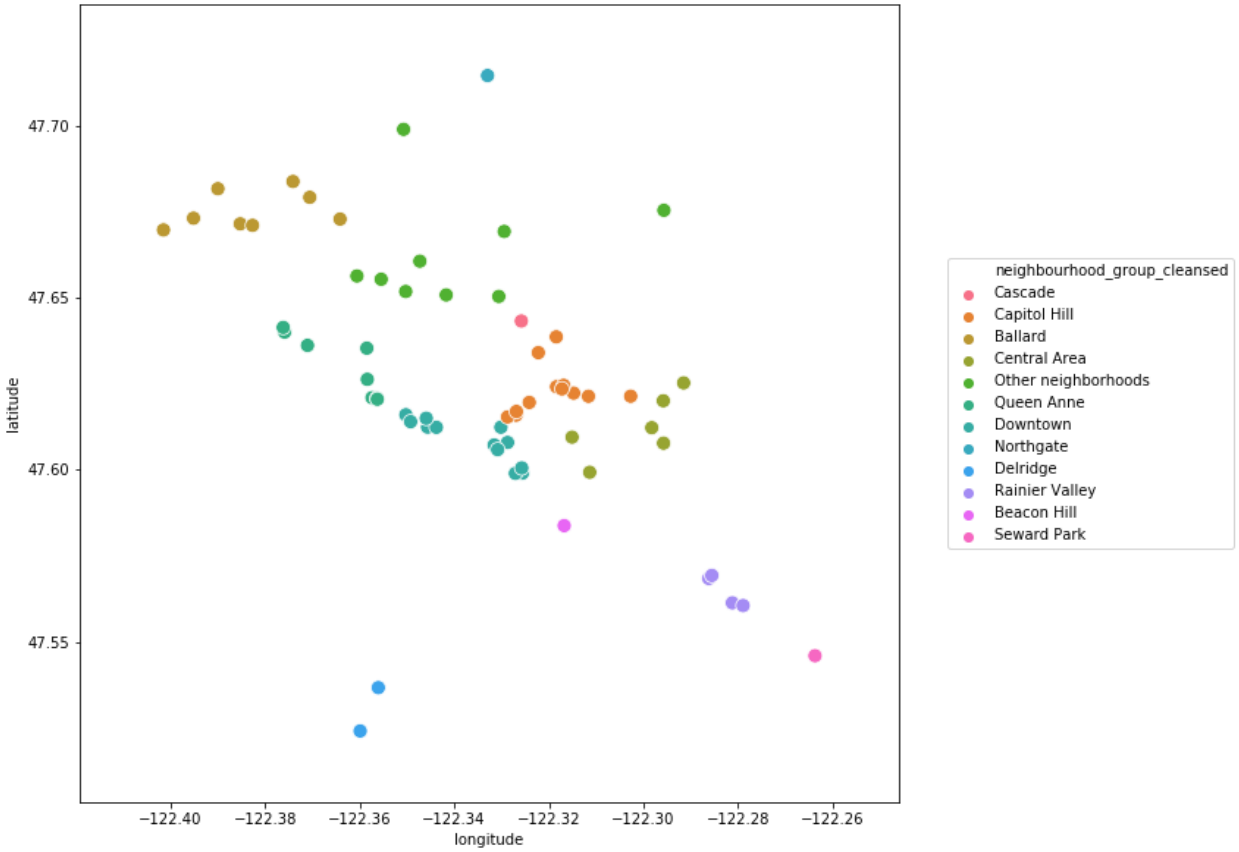
Ballard = neighborhood[neighborhood['neighbourhood_group_cleansed']=='Ballard']

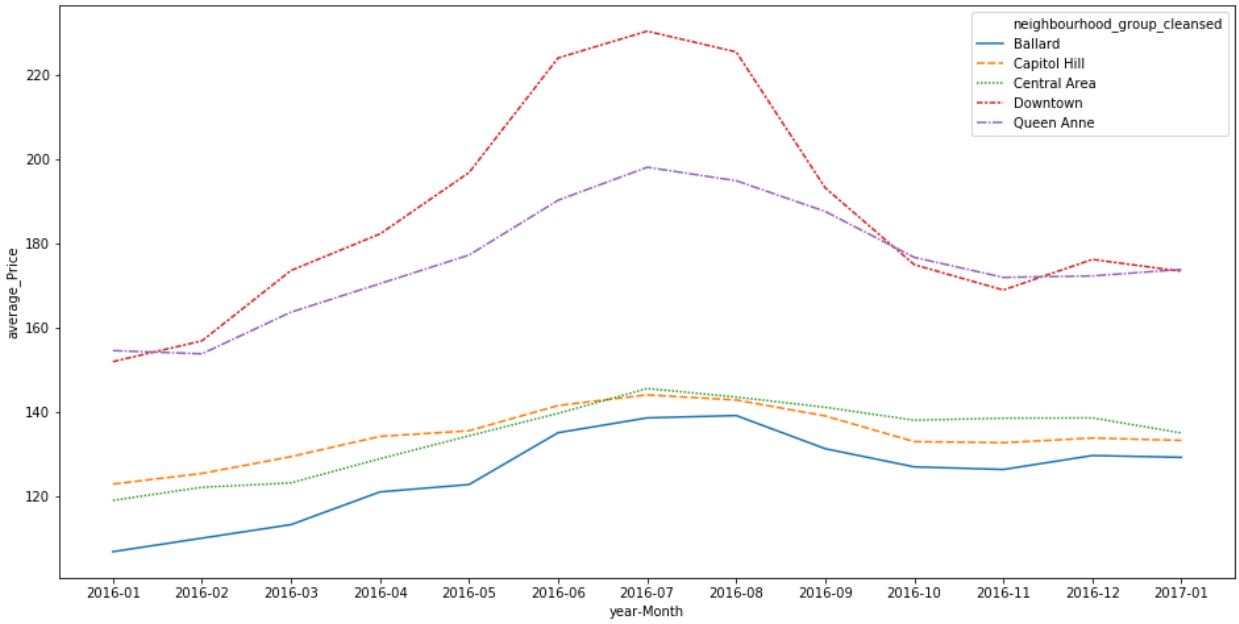
Downtown=neighborhood[neighborhood['neighbourhood_group_cleansed']=='Downtown']

Capitol_Hill = neighborhood[neighborhood['neighbourhood_group_cleansed']=='Capitol Hill']

Central_Area =neighborhood[neighborhood['neighbourhood_group_cleansed']=='Central Area']
Price change over time:







Conclusion

We used bigrams and spaCy NER model to extract keywords from the neighborhood overview provided by the Airbnb listings for the selected five neighborhoods. The NER results are more relative to what we expected: key destinations that conclude the neighborhood.

With finding the most popular Airbnbs based on the number of reviews and the average ratings, we can recommend these to people that are looking to stay in Seattle. Even if someone wants to stay in a certain neighborhood, they can see how many apartments are located there and give them a better idea of which one to pick.

By normalizing our user reviews, and then performing a TF-IDF analysis on it, we were able to find the 'weight' for the words used in the reviews for each neighborhood. With that, we performed topic modeling for each neighborhood, showing which words and topics were the most prominent features for each neighborhood in the reviewer's eyes. This can help a new user determine which airbnbs they would want to choose by checking which neighborhood to look for, based on which neighborhood featured words their personal preferences can match with.

After cleaning out NaN values and sorting the data, we calculated the average price for the neighborhoods that we are focusing on. If we are choosing the neighborhood based on budget, Ballard would be our choice due to the lowest average price throughout the year.

How can we improve?

- Some of the reviews in our data set were written with non-Western characters (reviews written in Mandarin or Korean). Could we find a way to take reviews in other languages into account in future analysis?
- How can we take the date when the review was submitted? 2016 versus 2019?
- Can our reviews analysis go deeper into seeing what features people enjoyed the most during their stay in Seattle, and their specific Airbnb?