**INFO 3400 Course Project Phase I**

Isaiah Erb, Jason Koff

**Introduction:**

We’ve been provided with AirBnB data from 2018 and 2019. This dataset includes quantitative data for each listing and textual reviews from previous guests. Our goal is to use text mining techniques, such as keyword and sentiment analysis, to analyze the reviews and answer the explorative questions listed below.

**Graphical user interface, chart

Description automatically generated**

**Explorative Questions:**

1. What are the top keywords mentioned in the textual reviews?
2. How do the top keywords vary among property and room types?

Formula node, select node, something categorizing by property types, another one categorizing by room types

1. Overall sentiment of all textual reviews? Sentiment by topic?
2. How does sentiment vary across different neighborhoods?

Group by?

1. What topics are mentioned, and what keywords make up these topics?
2. Can we accurately use textual reviews to predict the price per day?

Linear regression

**Variables:**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Type** |
| id | ID of the listing | Continuous, integer |
| listing\_url | Url of the listing | Nominal, text string |
| name | Name of the renting house/room | Nominal, text string |
| summary | Summary of the place | Nominal, text string |
| experiences\_offered  \*(could be removed) | Whether the listing also offers other traveling experiences | Nominal, text string  \*(should be binary) |
| interaction | The host’s desired way of interaction | Nominal, text string |
| house\_rules | House rules | Nominal, text string |
| host\_id | The ID number of the host | Continuous, integer |
| host\_name | Name of the host | Nominal, text string |
| host\_since | Since when the host has been hosting | Continuous, integer  \*(should be date) |
| host\_response\_time | Host response time | Nominal, text string |
| host\_is\_super\_host | Whether or not the host is a super host | Nominal, text string  \*(should be binary) |
| host\_identity\_verified | Whether or not the identity of the host is verified | Nominal, text string  \*(should be binary) |
| neighborhood | The neighborhood the house is at | Nominal, text string  \*(should be categorical) |
| zipcode | The zip code of the listing | Continuous, integer |
| property\_type | The type of the property | Nominal, text string  \*(should be categorical) |
| room\_type | The type of the room | Nominal, text string  \*(should be categorical) |
| accomodates | Number of accommodates | Continous, integer |
| bathrooms | Number of bathrooms | Continuous, numeric |
| bedrooms | Number of bedrooms | Continuous, integer |
| beds | Number of beds | Continuous, integer |
| bed\_type | Type of the bed | Nominal, text string |
| amenities | Available amenities | Nominal, text string |
| square\_feet | Total size of the listing in square feet | Continuous, integer |
| price | The listing price per night | Continuous, decimal |
| weekly\_price | The listing price per week | Nominal, text string  \*(should be continuous, integer/decimal) |
| monthly\_price | The listing price per month | Nominal, text string  \*(should be continuous, integer/decimal) |
| cleaning\_fee | The fee for cleaning | Nominal, text string  \*(should be continuous, integer/decimal) |
| minimum\_nights | Number of minimum nights for each reservation | Continuous, integer |
| availability\_365 | Number of available days for the year | Continuous, integer |
| number\_of \_reviews | Number of reviews | Continuous, integer |
| first\_review | First time the listing was reviewed | Continuous, integer  \*(should be date) |
| last\_review | Last time the listing was reviewed | Continuous, integer  \*(should be date) |
| review\_scores\_rating | Overall rating of the listing | Continuous, integer |
| review\_scores\_accuracy | Average rating on accuracy | Continuous, integer |
| review\_scores\_cleanliness | Average rating on cleanliness | Continuous, integer |
| review\_scores\_checkin | Average rating on checkin | Continuous, integer |
| review\_scores\_communication | Average rating on communication | Continuous, integer |
| review\_scores\_location | Average rating on location | Continuous, integer |
| review\_scores\_value | Average rating on value | Continuous, integer |
| instant\_bookable | Whether or not hte listing supports instant booking | Nominal, text string  \*(should be binary) |
| cancellation\_policy | The cancellation policy of the listing | Nominal, text string |
| calculated\_host\_listings\_count | The number of house/rooms owned by the listing’s host | Continuous, integer |
| reviews\_per\_month | Average number of reviews each month | Continuous, numeric |