**INFO 3400 Course Project Phase I**

Isaiah Erb, Jason Koff

**Introduction:**

|  |  |
| --- | --- |
| **Introduction** | **Dashboard** |
| 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. We aim to use text mining techniques, such as keyword and sentiment analysis, to analyze the reviews and answer the explorative questions listed below. |  |

**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 |

**Text Preprocessing:**

|  |  |
| --- | --- |
| **Text Preprocessing Process** | **Word Cloud:** |
| We removed special characters, uppercase letters, null comments, and the default English stopwords included in the text preprocessing node in Alteryx. To remove the special characters, we used multiple Regex nodes, replacing them with an apostrophe or just a blank space. We also removed specific reviews using a filter node and the comment ID for ones that were in a different language or were nonsensical reviews that consisted of all special characters. |  |

**Explorative Questions:**

|  |  |  |
| --- | --- | --- |
| **Questions** | **Findings** | **Figures** |
| **What are the top keywords mentioned in the textual reviews?** | The top five keywords were ‘great’, ‘place’, ‘stay’, location, and ‘clean’. They all had over 3,000 appearances in Airbnb reviews. Keywords and their values are shown in the figure below. The keywords that appear most frequently indicate positive sentiment, honing in on the property aspects and location.  \*Note - Alteryx skips over a few keywords in the graph to make space, clicking on the individual bars in the graph itself to indicate the exact keywords |  |
| **How do the top keywords vary among property and room types?** | Regarding the top keywords by property type, we concluded that the majority were associated with houses, with apartments next, then townhouses, guest suites, guesthouses, and condominiums making up the least. The keywords seemed to stay the same, with positive descriptors about the property and location. We found that ‘neighborhood’ was only mentioned for houses, indicating that reviewers were concerned with the location of houses more than the other property types.  Regarding property type, most of the keywords were associated with entire homes/apartments rather than private rooms, which logically makes sense as there are more entire homes/apartments listed on Airbnb in Denver than private rooms. The keywords stayed consistent with the other two figures, with the top six being split between the two room types. We found that ‘home’, ‘comfortable’, and ‘host’ were also shared among the two room types. |  |
| **The overall sentiment of all textual reviews? Sentiment by topic?** | The sentiment of all the textual reviews combined was overwhelmingly positive, with 11,483 words yielding positive sentiment scores, 702 neutral, and only 56 negatives. This indicates that, for the most part, Airbnb guests that leave reviews are most likely to leave ones that positively describe the property, location, and host.  Regarding sentiment by topic, we found that nearly all of the reviews were positive among the three topics, the two leading topics with around 4,000 positive scores and the other two following closely behind both with over 3,500 positive scores. Topics one and three had similar negative figures, each about 250-300. |  |
| **How does sentiment vary across different neighbor-**  **hoods?** | None of the neighborhoods had more negative words associated with them than positive ones. The most negative sentiment scores came from Five Points, the most frequently reviewed neighborhood with the most positive reviews. Other popular neighborhoods included Highland and Capitol Hill, with 905 and 518 sentiment scores, respectively. |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Topic** | **Topic Name Choice Explanation** | **Top 5 Key-**  **words** | **Reccomendation** | **Figure** |
| **Intertopic Distance Map** | The intertopic distance map is ideal, having no overlapping. |  |  |  |
| **Topic 1: Property Details (26.8%)** | ‘Property Details’ seemed the most fitting name for this topic since words like ‘stay’, ‘space’, ‘spot’, ‘house’, and ‘cozy’ all describe the guest’s description of the property itself. | great, place, location, stay, clean | With such a positive sentiment seen in all 3 of the topics we created, it is hard to determine recommendations. There is a possibility for location and communication for being negative so it is recommended to advise the bookee of any recent crimes in the community and to communicate efficiently. |  |
| **Topic 2: Guest Experience (36.7%)** | ‘Guest Experience’ seemed fitting for topic two, as words like ‘clean’, ‘host’, ‘quick’, ‘amazing’, ‘comfortable’, and ‘excellent’ apply to the guest’s time staying at the Airbnb listing and how they enjoyed their stay. | stay, place, denver, home, recommend | With such a positive sentiment seen in all 3 of the topics we created, it is hard to determine recommendations.While topic two had the most positive sentiment of the three, there could be some improvement at the host level, such as improving the cleanliness of the property and the atmosphere of the house. |  |
| **Topic 3: Neighbor-**  **hood Details (36.5%)** | Finally, ‘Neighborhood Details’ seemed fitting for topic three because the words ‘downtown’, ‘neighborhood’, ‘denver’, ‘parking’, ‘area’, ‘night’,, ‘restaurants’, and ‘coffee’ all apply to the area around the Airbnb, not the property itself. | downtown, neighborhood, nice, restaurants, easy | With such a positive sentiment seen in all 3 of the topics we created, it is hard to determine recommendations. Since restaurants is listed very highly it is advisable to prepare a list of restaurants at different price points for the guest to enjoy. |  |

**Can we accurately use textual reviews to predict the price per day?**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Training - Validation Split** | **Variables Included** | **Adj R-Squared** | **MSE** | **Variables with Significant p-values (less than 0.05)** |
| **Linear Regression Without Text** | 60 Training: 40 Validation | 18 Variables included: host\_since, host\_responce\_time, host\_is\_superhost, host\_identity\_verified, neighborhood, property\_type, room\_type, accommodates, bathrooms, bedrooms, bed\_type, cleaning\_fee, minimum\_nights, availability\_365, instant\_bookable, cancellation\_policy, calculated\_host\_listings | 0.204 | 24079.574 | At the 0.05 significance level, we determined that host\_since, neighbourhood Villa Park, Room\_TypePrivate room, Room\_TypeShared room, bathrooms, bedrooms, beds, cleaning\_fee, minimum\_nights, availability\_365 are significant predictor variables. Both neighborhood and Roo,\_Type are categorical variables that held some significant predictors. |
| **Linear Regression With Text** | 60 Training: 40 Validation | 28 Variables included: host\_since, host\_responce\_time, host\_is\_superhost, host\_identity\_verified, neighborhood, property\_type, room\_type, accommodates, bathrooms, bedrooms, bed\_type, cleaning\_fee, minimum\_nights, availability\_365, instant\_bookable, cancellation\_policy, calculated\_host\_listings, compound\_sentiment\_score, comment\_length, review\_scores\_rating, review\_scores\_accuracy, review\_scores\_cleanliness, review\_scores\_checkin, review\_scores\_communication, review\_scores\_location, review\_scores\_value, reviews\_per\_month | 0.255 | 23846.122 | At the 0.05 significance level, we determined that host\_since, neighbourhoodVilla Park, Room\_TypePrivate room, Room\_TypeShared room, bathrooms, bedrooms, beds, cleaning\_fee, minimum\_nights, availability\_365, review\_scores\_accuracy, review\_scores\_checkin, review\_scores\_location, review\_scores\_value, reviews\_per\_month, calculated\_host\_listings\_count |
| **Interpretation** | The linear regression model with text was the better predictor of the target variable price, however, the models preformed similarly. In comparison to the linear regression model without text, it has a higher Adj R-Sqaured and a higher Mean Sqaured Error. At an Adj R-Sqaured of 0.252, the linear regression model with text is not a fantastic predictor but does hold predictive capabilities.  To improve the model in the future, we would like to use AirBnB reviews with negative sentiment. Having more negative sentiment could allow the model to predict lower prices more accurately. |  |  |  |  |

\*Note: for model comparison, we just used the R-Squared and RMSE values outputted in the interactive section of the Linear Regression node. Our Union and Model Comparison nodes weren’t working, and we couldn’t figure out what the issue was.



**Conclusion:**

After preprocessing our text using the methods described above, the top keywords observed in our textual data were 'great', 'place', 'stay', location, and 'clean'. All indicated relatively positive sentiment, equally reflected by our sentiment analysis, which yielded 11,483 positive scores. After reviewing keywords by property and room types, we found that houses were the most frequent property type, and entire houses/apartments were the most frequent room type. The keywords in the top 50 stayed relatively the same among the property and room type graphs. Additionally, we performed a sentiment analysis by neighborhood, with Five Points being the most frequent neighborhood, then Highland and Capitol Hill following shortly after. Sentiment for these neighborhoods was overwhelmingly positive; however, words with negative scores were associated with each neighborhood. Next, we split the keywords into three topics: Property Details, Guest Experience, and Neighborhood Details. These titles were derived from some of the main keywords associated with each topic. There are recommendations in the report for Airbnb business decisions based on the sentiment associated with the keywords in each topic. The recommendations primarily focus on the host being communicative, kind, and knowledgeable of local activities, such as restaurants, coffee shops, etc. Finally, we used two different linear regression models, one using the textual data (including sentiment and comment length) and one without the textual data. According to each model's R-Squared and RMSE values, we concluded that the linear regression with text was a better predictor than the linear regression without text. However, both models' performances were similar.