# Final Project: Preliminary Plan

**Team members:** [Macon Barker](mailto:barkjm23@wfu.edu)**,** [Hewei Shen](mailto:shenh23@wfu.edu)

**You should review the data and create a brief outline describing how you plan to use the data for your analysis (e.g., which variables will you drop, how will you handle missing values, what additional variables might you need to create, how will you treat categorical variables, which variables will serve as your target variables and input variables to answer the questions posed above)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Data Type** | **Missing %** | **How to Use** | **Why** |
| id | **Int** | **0** | **Drop** | **Not useful** |
| listing\_URL | **String** | **0** | **Drop** | **Not useful** |
| name | **String** | **0** | **Drop** | **Not useful** |
| summary | **Text** | **7.9** | **Drop** | **Redundant** |
| space | **Text** | **23.3** | **Word cloud after clustering** |  |
| description | **Text** | **0.05** | **Word cloud after clustering** |  |
| experiences\_offered | **String** | **0** | **Drop** | **Not useful** |
| neighborhood\_overview | **Text** | **38.8** | **Drop** | **Too many missing** |
| notes | **Text** | **58.7** | **Drop** | **Too many missing** |
| transit | **Text** | **40.2** | **Drop** | **Too many missing** |
| host\_id | **Int** | **0** | **Drop** | **Not useful** |
| host\_name | **String** | **0.26** | **Drop** | **Not useful** |
| host\_since | **Date** | **0.26** | **Drop** | **Not useful** |
| host\_location | **String** | **0.43** | **Drop** | **Not useful** |
| host\_about | **Text** | **31.9** | **Drop** | **Too many missing** |
| host\_response\_time | **String** | **28.4** | **Drop** | **Too many missing** |
| host\_response\_rate | **Float** | **28.4** | **Drop** | **Too many missing** |
| host\_is\_superhost | **Binary** | **0.26** | **Fillna with 0** |  |
| host\_listings\_count | **Int** | **0.26** | **Fillna with median(1)** |  |
| host\_has\_profile\_pic | **Binary** | **0.26** | **Fillna with 0** |  |
| host\_identity\_verified | **Binary** | **0.26** | **Fillna with 0** |  |
| neighbourhood | **String** | **17.7** | **Exploratory first, then group and put into the model if any important**  **Fillna with ‘Unknown’** |  |
| city | **String** | **0** | **Drop** | **Not useful** |
| property\_type | **Category** | **0** | **House V.S. Apartment V.S. Others** |  |
| room\_type | **Category** | **0** |  |  |
| accommodations | **Int** | **0** |  |  |
| bathrooms | **Float** | **0.79** | **Fillna with median(1)** |  |
| bedrooms | **Int** | **0.1** | **Fillna with median(1)** |  |
| beds | **Int** | **0.39** | **Fillna with median(2)** |  |
| bed\_type | **Category** | **0** |  |  |
| amenities | **Text** | **0** | **Word cloud after clustering** |  |
| square\_feet | **Float** | **94.8** | **Drop** | **Too many missing** |
| price | **Float** | **0** |  | **Target variable for linear reg** |
| weekly\_price | **Float** | **61.8** | **Drop** | **Too many missing** |
| security\_deposit | **Float** | **52.5** | **Fillna with 0** |  |
| cleaning\_fee | **Float** | **38.5** | **Fillna with 0** |  |
| guests\_included | **Int** | **0** |  |  |
| extra\_people | **Float** | **0** |  |  |
| minimum\_nights | **Int** | **0** |  |  |
| has\_availability | **Binary** | **0** | **Drop** | **Not useful** |
| availability\_30 | **Int** | **0** |  |  |
| availability\_60 | **Int** | **0** |  |  |
| availability\_90 | **Int** | **0** |  |  |
| availability\_365 | **Int** | **0** |  |  |
| number\_of\_reviews | **Int** | **0** |  |  |
| review\_scores\_rating | **Float** | **35.1** | **Exploratory only for notNa** |  |
| review\_scores\_accuracy | **Float** | **35.3** | **Exploratory only for notNa** |  |
| review\_scores\_cleanliness | **Float** | **35.3** | **Exploratory only for notNa** |  |
| review\_scores\_checkin | **Float** | **35.3** | **Exploratory only for notNa** |  |
| review\_scores\_communication | **Float** | **35.3** | **Exploratory only for notNa** |  |
| review\_scores\_location | **Float** | **35.3** | **Exploratory only for notNa** |  |
| review\_scores\_value | **Float** | **35.3** | **Exploratory only for notNa** |  |
| instant\_bookable | **Binary** | **0** |  |  |
| cancellation\_policty | **Category** | **0** |  |  |
| desccription\_filled | **Category** |  |  | **Count how many text columns were filled (Not NA) out of a total of eight.**  **‘Complete’ if n = 8,**  **‘Mostly filled’ if n >=4,**  **‘Mostly blank’ if n <4** |
| availability\_90\_pct | **Float** |  |  | **Availability / 90.** |
| booked | **Binary** |  |  | **Target variable for DT/Log reg**  **1 if availability\_90\_pct < 40%** |
| neighbourhood\_2 | **Category** |  | **Fillna with ‘Others’** | **Group the neighborhood column according price and count** |

**You should create a preliminary analysis plan highlighting the analytical techniques that you plan to use to address the questions required for the report**

Questions

* What are the key factors driving the nightly rental price for units in the Austin area?

– **Linear Regression**

– **Y: price**

– **X:** host\_is\_superhost, host\_listings\_count, host\_has\_profile\_pic, host\_identity\_verified, neighbourhood, property\_type, room\_type, accommodates, bathrooms, bedrooms, beds, bed\_type, amenities, security\_deposit, cleaning\_fee, guests\_included, extra\_people, minimum\_nights, availability\_30, availability\_60, availability\_90, availability\_365, number\_of\_reviews, review\_scores\_rating, review\_scores\_accuracy, review\_scores\_cleanliness, review\_scores\_checkin, review\_scores\_communication, review\_scores\_location, review\_scores\_value, instant\_bookable, cancellation\_policy

* What are the key factors driving the probability of keeping a unit booked in the Austin area? To answer this question, you can classify a unit as “booked” if the 90-day availability for the property is less than 40%.

– **Classification Tree/Log Regression**

– **Y: booked**

– **X:** host\_is\_superhost, host\_listings\_count, host\_has\_profile\_pic, host\_identity\_verified, neighbourhood, property\_type, room\_type, accommodates, bathrooms, bedrooms, beds, bed\_type, amenities, price, security\_deposit, cleaning\_fee, guests\_included, extra\_people, minimum\_nights, availability\_30, availability\_60, availability\_90, availability\_365, number\_of\_reviews, review\_scores\_rating, review\_scores\_accuracy, review\_scores\_cleanliness, review\_scores\_checkin, review\_scores\_communication, review\_scores\_location, review\_scores\_value, instant\_bookable, cancellation\_policy

* How important are guest reviews and ratings? Is there anything the host could do to improve those ratings?

- **Exploratory Analysis?**

* + **Are we looking at reviews and ratings as it impacts the price?**
* Segment the Airbnb properties in the Austin area into different groups. Provide a profile of each segment. Which segments command the highest nightly rates? Which segments get the best guest ratings? Which segments have the highest percentage of “booked” units?

- **Clustering**