AirBnB Pre-Work Project

An Inquiry into the Washington D.C. Housing Market

Table of Contents

[PROJECT OVERVIEW 3](#_Toc527544727)

[DELIVERABLES AND TIMELINE 3](#_Toc527544728)

[Presentation 3](#_Toc527544729)

[Excel Workbook 3](#_Toc527544730)

[Data Set 3](#_Toc527544731)

[Prompts 4](#_Toc527544732)

[Project Steps 4](#_Toc527544733)

[Step 1: Data Preparation (All Prompts) 4](#_Toc527544734)

[Step 2: Data Exploration (by Prompt) 5](#_Toc527544735)

[Step 3: Visualize, Summarize, and Present (All Prompts) 7](#_Toc527544736)

[Step 4: Recommendation (All Prompts) 8](#_Toc527544737)

[RESOURCES 8](#_Toc527544738)

[GRADING GUIDE 9](#_Toc527544739)

[RUBRIC 9](#_Toc527544740)

# PROJECT OVERVIEW

You’re working for a client who wishes to invest in an Airbnb property in Washington, D.C. Before your client decides to invest, they’d like clean data about Airbnb performance in D.C.’s neighborhoods that supports a clear recommendation for an investment in a specific market.

# DELIVERABLES AND TIMELINE

## Presentation

* Due: {Monday September 10th, 2o18}
* Format: PDF (export any Keynote or PPT files to PDF)
* Description:
  + Include all relevant prompts & sub-prompts you chose
  + Reference any data selected from the original file
  + Describe any cleaning methods used to remove erroneous data
  + List recommendations based on your sub-prompt
  + A “headline” that summarizes your key takeaway/recommendation. You will pair and share this headline on the first day of class

## Excel Workbook

* Due: {Monday September 10th, 2o18}
* Format: An Excel file containing your calculations
* Description:
  + Include clean listing data with the requested data points
  + Results of analysis are presented in separate worksheets, formatted, and (if applicable) visualized
  + Include exploratory efforts using PivotTables, visualizations, and statistical review
  + Format: Excel file with multiple worksheets:
    - Sheet 1, ‘cleaned\_data’: A copy of the original .csv data with columns properly formatted (dates, currencies, etc.). Any modifications outside of format should be applied as derived formulas, located *to the right* of the original columns, with descriptive titles applied.
    - Sheet 2, ‘prompt\_main’: A brief response to the [main prompt](#_Prompts). This should include pivot tables and charts as needed.
    - Other sheets (as needed): Responses to sub-prompts, as needed, and titled appropriately.

## Data Set

You’ve been provided with scraped data captured by a web program with listing information from the Airbnb website. This data may contain unformatted data points with duplicate entries. You’ll want to clean and format the data prior to performing exploratory analysis — this will help you better understand the available data and build some business context.

## Prompts

* Main Prompt (mandatory):
  + Should our investor invest in an Airbnb property in Washington, D.C.? If so, in which neighborhood should they invest?
* Sub-Prompt (pick one):
  + Prompt 1: Host revenue — How much revenue do successful hosts generate?
  + Prompt 2: Property reviews — Which property types receive the most positive reviews?
  + Prompt 3: Neighborhood popularity — Which neighborhoods host the most listings?
  + Prompt 4: Neighborhood sentiment — Which neighborhoods receive the most positive reviews?

## Project Steps

### Step 1: Data Preparation (All Prompts)

* Data Cleaning
* Remove duplicate and erroneous data.
  + Ensure that the *id* column doesn’t have any duplicates. This field is considered a [primary key](https://en.wikipedia.org/wiki/Primary_key) and should be unique within the table (sheet). If it isn’t unique, please make it so. If merging multiple records (rows), state your reasoning about how to conduct the merge.
  + Note: The *id* column is the ID for the table and represents a listing. The *host\_id* column is the ID of the host. Hosts can have more than one listing.
  + Are descriptions of the property and summaries (column *description* and *summary*, respectively) of the neighborhood going to help your analysis?
* Standardize the *neighbourhood* entry.
  + Write a formula, or series or formulas that seek to reduce the number of redundant/duplicate values within the *neighbourhood* column.
  + Your solution should minimize human manual involvement – using a formula allows your analysis to scale to other datasets.
  + Balance accuracy and time – a timely approximate answer beats a late exact one.

### Step 2: Data Exploration (by Prompt)

#### PROMPT 1: How much revenue do successful hosts make?

* Estimate annual revenue per listing (each row is considered a listing).
* Use the following assumptions:
  + Each booking always has two guests, unless the listing only accommodates one.
  + The booking is always for the minimum number of days allowed.
  + Only half of the bookings generate a review.
  + If the listing has a review history of fewer than 60 days, use the number of reviews within the period of its activity as its annual number of reviews. Otherwise, interpolate or extrapolate the number of reviews over its period of activity to an annual number of reviews.
  + Assume the security deposit is always returned in full upon checkout.
  + Assume cleaning fee is charged in full upon checkout.
* *Some of the columns* you may need to use include:
  + : How many people the property can accommodate, *accommodates*.
  + : How many guests are included in the price. Column *guests\_included*.
  + : The extra cost per person, *extra\_people*, if you go above the number of *guests\_included*, . [USD]
  + : The price of the listing, *per night*.
  + : The cleaning fee of the listing, this is charged *once* *per stay*.
  + : The minimum number of nights a listing can be booked for, *minimum\_nights*
  + : Number of reviews for a listing, *number\_of\_reviews*
  + : First review date, *first\_review*
  + : Last review date, *last\_review*  
      
    *Note*: variable names, in parentheses, have been arbitrarily assigned to the above columns for use in equations below.
* *Constants* you may need to use include:
  + : percentage of stays which had reviews left = 0.5 (50 percent)

*Note*: These are assumptions we’re making and not included in the dataset.

* *Outputs* you may have as results include:
  + : revenue generated by a listing for a *single* *night* (assumes 2 people staying)
  + : revenue generated by a listing for a *single booking* (assumes 2 people staying)
  + : revenue generated by a listing for *the entire lifetime of the listing* (all history of all reviews ever left; assumes 2 people staying)
  + : estimated (extrapolated or interpolated) number of stays in for one year of a listing. Note that the listing must have 60 days of review history to qualify for an extrapolation. See note in prompt statement above.
* Step 1: *Estimate* the actual number of *stays* (not nights!) **per year** for each listing by assuming that 50 percent of customers who stayed left a review (use the data in the *number\_of\_reviews* column). For example, if 10 customers leave a review, you can assume the listing had 20 *stays* in total (create a new column with a number that’s an estimate for how many stays each listing has received).
* Step 2: Calculate estimated revenue per **day** for two guests. If the accommodation explicitly accepts only one, add the additional cost *per guest* to get the total price per **day**. Note: “price” column is the **daily** price for the listing.

Example: If a listing has an *accommodates* value of 2, a *guests\_included* value of 1, an *extra\_people* value of $15.00, and a “price” of $80.00, the the price for a 2-person booking would be:

* Step 3: Multiply the estimated revenue per night by the minimum number of nights, plus the cleaning fee, to get an estimated revenue per booking (or *stay*).
* Step 4: Calculate an estimated total revenue for each listing by multiplying the estimated revenue per booking by the estimated number of stays *for a year*.
* If needed, build PivotTables in order to quickly explore the data at a high-level:
  + PivotTables should contain the host name, total revenue, and number of listings (make sure to exclude listings with no bookings).
  + Additional PivotTables are welcome and useful for gaining a better understanding of the data.
  + Format: Currency/thousands separators should be used where appropriate.

#### PROMPT 2: Which property types receive the most positive reviews?

* Build several PivotTables in order to quickly explore the data at a high level:
  + PivotTables should contain the property type, number of listings (make sure to exclude listings with no bookings), and average rating.
  + Additional PivotTables are welcome and useful for gaining a better understanding of the data.
  + Format: Currency/thousands separators should be used where appropriate.

#### PROMPT 3: Which neighborhoods host the most listings?

* Build several PivotTables in order to quickly explore the data at a high level:
  + PivotTables should contain the neighborhood name and number of listings (make sure to exclude listings with no bookings).
  + Additional PivotTables are welcome and useful for gaining a better understanding of the data.
  + Format: Currency/thousands separators should be used where appropriate.

#### PROMPT 4: Which neighborhoods receive the most positive reviews?

* Build several PivotTables in order to quickly explore the data at a high level:
  + PivotTables should contain the neighborhood name and rating. How might you best aggregate the rating metric here?
  + Additional PivotTables are welcome and useful for gaining a better understanding of the data.
  + Format: Currency/thousands separators should be used where appropriate.

### Step 3: Visualize, Summarize, and Present (All Prompts)

* Visualize insightful data for easier comprehension in support of your argument.
  + Format: If needed, copy and paste the results of your PivotTable(s) and adjust them to only show the relevant findings. Create a chart using the format that best fits your needs.
  + Your client is unwilling to go into untested markets — consider limiting the data you visualize to established (multiple listings) areas.
* Use your judgment and curiosity to uncover insights useful to your client. Here are some ideas:
* Histograms of:
  + Price per night
  + Minimum number of nights stay
  + Review scores
* Possible seasonal trends
* Vacancy rates by property type or location
* Property profiles and trends by neighborhood
* Top performing hosts, and their property portfolios
* Characteristics of top performing listings
* Organize your workbook appropriately.
* Present of your findings to the class:
  + Each student or team will present.
  + Each presentation should last five minutes.
  + In addition to your conclusions, discuss the cleaning techniques you used.
  + While others are presenting, take notes on the important points made by each team.

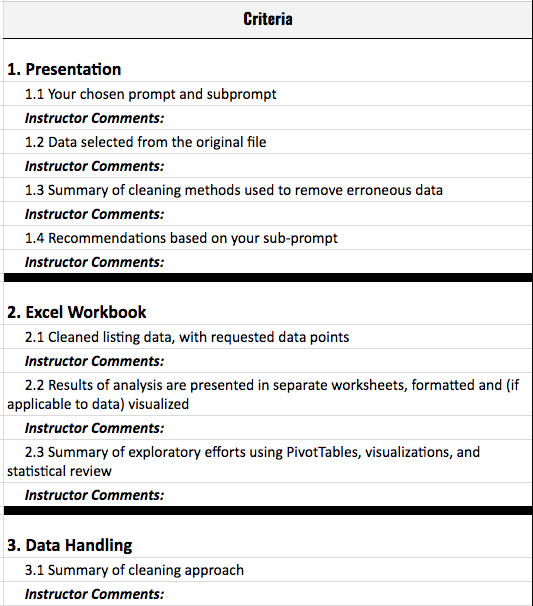
### Step 4: Recommendation (All Prompts)

* Finally, discuss your overall recommendation. Review the original prompt: “Should our investor invest in an Airbnb hotel in Washington, D.C.? If so, in which neighborhood should they invest?”
  + Create a final recommendation to the investor based on pertinent data from your analysis.
  + Keep in mind that the client wishes to understand the local market and its potential to host a profitable Airbnb property — demand is key, as is the price point.

# RESOURCES

* Review [existing articles](https://www.google.com/search?q=insideairbnb.com#q=insideairbnb.com&tbm=nws) on Airbnb or sources that quote “insideairbnb.com,” etc. to see how others approached this type of data.
* Explore this [Data Mining Framework](https://decisionstats.files.wordpress.com/2011/10/12345.png).
* Check out this [Data Cleaning Walkthrough](https://drive.google.com/file/d/1KuYvKg1IcQviv3ROatzoXs1w7Ib4rprE/view?usp=sharing).
* Here’s a handy list of [Excel keyboard shortcuts](https://support.office.com/en-us/article/Keyboard-shortcuts-in-Excel-2010-20603861-42b6-4c93-82ec-66924ea9b323?ui=en-US&rs=en-US&ad=US).
* Look at some of these data source [disclaimers](http://insideairbnb.com/about.html#disclaimers).

# GRADING GUIDE



# RUBRIC

* For all requirements, project deliverables will be evaluated using a simple point scale.
* In addition to numeric feedback, instructors will provide comments on all required portions.

|  |  |
| --- | --- |
| **Score** | **Expectations** |
| **0** | *Incomplete.* |
| **1** | *Partial credit but does not meet expectations.* |
| **2** | *Meets expectations.* |
| **3** | *Surpasses expectations.* |

**Description:**

* A “1” means you have met some but not all of the project requirements.
* A “2” means you have completely satisfied all requirements.
* A “3” indicates performance above and beyond these requirements and will not apply to most items.