You’re working for a client who wishes to invest in an Airbnb property in Washington, D.C to rent out to guests. 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**

1. [**Presentation**](#bookmark=id.gjdgxs)
2. [**Workbook**](#bookmark=id.30j0zll)

**1.** **Five-Minute Presentation**

* **Due**: Thursday, March 19
* **Format**: PowerPoint, Google Slides or 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.
  + Present your findings to the class.

**2.** **Excel Workbook**

* **Due**: Thursday, March 24
* **Format**: A spreadsheet worksheet 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: Clean listing data with the requested data points.
    - Sheet 2: Summary of data cleansing.
    - Other sheets (as utilized): Results of analysis and exploratory efforts.

**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.

Interested in exploring *more current data* from Airbnb? You can find the csv’s for many cities on this website: http://insideairbnb.com/get-the-data.html The listings.csv.gz link under each city is the dataset that aligns with the given prompts. Unzip the listings.csv.gz file.

Consider including *secondary data source(s)* that can validate your data-driven recommendation or bring additional insight. (i.e., Zillow data for pricing, Legal constraints for individual markets.) Include citations for validating resources in your presentation.

**PROMPT**

**Should our investor invest in an Airbnb in Washington, D.C.? If so, in which neighborhood should they invest?**

Choose from the following sub-prompts to help guide your analysis:

* Sub-Prompt 1: Host revenue — How much revenue do successful hosts generate?
* Sub-Prompt 2: Property reviews — Which property types receive the most positive reviews?
* Sub-Prompt 3: Neighborhood popularity — Which neighborhoods host the most listings?
* **Sub-Prompt 4: Neighborhood sentiment — Which neighborhoods receive the most positive reviews? *And why? What else makes those neighborhoods stand out? Nightlife? Density?***

**GETTING STARTED: PROJECT STEPS**

1. Project Step 1: Preparing our data set.
   1. Regardless of the sub-prompt you chose, you must have clean data before beginning your analysis. Follow these steps to prepare the Airbnb data set for analysis.
2. Project Step 2: Data exploration and analysis.
   1. Leveraging the clean Airbnb data set, use this step to answer your assigned prompt.
3. Project Step 3: Visualize, summarize, and present.
   1. Polish your work and findings from your data exploration and analysis by distilling your insights with the steps provided.

**PROJECT STEP 1: PREPARING OUR DATA SET (ALL PROMPTS)**

* Data Cleaning
* Remove duplicate and erroneous data (keep in mind, there may not be any).
  + Remove listings without any reviews and duplicate rows where the bot may have re-recorded listing data.
  + The “id” (col A) is the ID you should use to find duplicates. The “host\_id” (col O) is the ID of the host, but hosts can have more than one listing.
  + Are descriptions of the property and summaries of the neighborhood going to help your analysis?
* Standardize the entry of “State”, “City”, and “Neighborhood”.
  + “Find and Replace” can be used to find all values of a specified text value (e.g., “Rd.”) and replace them with a different text value (e.g., “Road”). This is the same as in Microsoft Word or Google Docs, in which you can find and replace words. For example, if you wanted to replace all occurrences of the name “John Smith” with the name “Jane Smith,” you could use the “Find and Replace” functionality.
  + Alternatively, you could identify the values in the data and translate them into a standard format of your choice. Then, by adding a new temporary column, you could use VLOOKUP to translate entries to the spelling of your choice.

**PROJECT STEP 2: DATA EXPLORATION AND ANALYSIS (BY SUB-PROMPT)**

**SUB-PROMPT 1: How much revenue do successful hosts make?**

* Estimate 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.
* Columns you’ll use to do this include:
  + Column BD, “price:” The price for a one-night stay.
  + Column BI, “guests\_included:” This is how many guests are included in the price.
  + Column BJ, “extra\_people:” The extra cost per person if you go above the number of “guests\_included” in column BI.
  + Column AW, “accommodates:” How many people the property can accommodate.
* Step 1: Calculate a proxy number of stays for each listing by assuming that 50 percent of customers who stayed left a review (use the data in the “number\_of\_reviews” column). 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 daily revenue for two guests, unless the accommodation explicitly accepts only one. Use the following logic to create a nested IF statement:
  + Determine if “guests\_included > 1” (i.e., whether the price quoted already includes two or more people). This price represents daily revenue under the assumptions.
  + If “guest\_included = 1” and “accommodates = 1,” then the listing is only for one person. This price represents daily revenue under the assumptions.
  + If “guest\_included = 1” but “accommodates > 1”, then the price listed is only for one person but the property can accommodate two or more, so you need to add in the additional price for another person to get the daily revenue for two people. This additional price is contained in column V.
* Step 3: Multiply the estimated daily revenue by the minimum number of nights to get an estimated revenue per booking.
* Step 4: Calculate an estimated total revenue for each listing by multiplying the estimated revenue per booking by the estimated number of stays.
* Format: Have a column that calculates daily revenue — account for number of guests accommodated, number of guests included in the price, and extra charge for additional people using nested IF statements. Another column would then calculate the revenue per booking. Finally, multiply that by the number of total stays for the listings.
* Build several 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.

**SUB-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.

**SUB-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 host 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.

**SUB-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, 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.

**PROJECT STEP 3: VISUALIZE, SUMMARIZE, AND PRESENT (ALL PROMPTS)**

* Visualize insightful data for easier comprehension in support of your argument.
  + Format: Copy and paste the results of your PivotTable(s) and adjust them to only show the relevant findings. Create a graph or chart using the format that best fits your findings’ needs.
  + Your client is unwilling to go into untested markets — consider limiting the data you visualize to the top 10 best performers.
* Summarize relevant statistics for various data points (max, median, mode).
  + Identify the listing that generates the most revenue.
    - xy percent rating with an entire two-bedroom apartment under $200/night.
  + Provide a profile of Airbnb activity in Washington, D.C.
    - xy percent are two-bedroom apartments, cd percent use airbeds, st percent of all listings are in X neighborhood, etc.
  + Identify the top-performing hosts and listings.
* Organize your workbook appropriately.
* Present your findings to the class:
  + Each student will present.
  + Each presentation should last five minutes.
  + Unless you think there is something particularly notable, avoid discussing the cleaning techniques you used in the presentation (but put those notes in your Excel workbook).
  + While others are presenting, take notes on the important points.

**PROJECT STEP 4: RECOMMENDATION (ALL PROMPTS)**

* Finally, discuss your overall recommendation. Review the original prompt: “Should our investor invest in an Airbnb 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 and price point are key.

**RESOURCES**

* Use the [Analytics Wor kflow](https://docs.google.com/document/d/1h5BDZLdhDNHFD9_OqEBvZ-6yAgN38i07ZWd3de2pw9A/edit?usp=sharing) to guide you through each step.
* 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).

**REQUIREMENTS & EVALUATION**

* For a complete list of all project requirements, please see the grading guide.
* Your Excel worksheet (containing clean data) will be evaluated using the requirements in the grading guide.
* Your presentation will be reviewed by peers and graded by instructors as a pass/fail.

**RUBRIC**

* For all requirements, project deliverables will be evaluated using a simple point scale.
* In addition to numeric feedback, your instructor will provide written comments.

|  |  |
| --- | --- |
| **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 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.