**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**

1. **Final Presentation**
2. **Project Deliverables**

**1.**  **Final Presentation**

* **Due**: Delivered on Thursday, May 2nd.
* **Format**: Google Slides or PowerPoint
* **Time**: 6 minutes minimum and 8 minutes maximum (6-8 minutes) with additional time for audience Q&A.
* **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 chosen sub-prompt(s).
  + Present your findings to the class and your recommendation to the investor.

**2.** **Jupyter Notebook(s)**

* **Due**: By 11 a.m. on Thursday, May 2nd.
* **Format**: A Jupyter notebook or notebooks containing all your code, calculations, and analysis.
* **Description**:
  + Include list of cleaning steps with the requested data points.
  + Results of analysis in code is presented and visualized.
  + Include exploratory efforts using various Pandas built-in methods of exploration such as Groupbys/PivotTables, visualizations, and statistical review.
  + <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.groupby.html>
  + <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.groupby.html>
  + <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.Series.value_counts.html>

**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 carefully review, clean, and format the data prior to performing exploratory analysis — this will help you better understand the available data and build some business insights to inform your business recommendation.

**PROMPTS**

As a class, we’ll address the main prompt: **Should our investor invest in an Airbnb in Washington, D.C.? If so, in which neighborhood should they invest?**

Choose one (or more) of the following sub-prompts to help guide your analysis:

* Subprompt 1: Host revenue — How much revenue do successful hosts generate?
* Subprompt 2: Property reviews — Which property types receive the most positive reviews?
* Subprompt 3: Neighborhood popularity — Which neighborhoods host the most listings?
* Subprompt 4: Neighborhood sentiment — Which neighborhoods receive the most positive reviews?

**GETTING STARTED: PROJECT STEPS**

1. Project Step 1: Preparing our data set.
   1. Regardless of the sub-prompt(s) 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 chosen prompt(s).
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
* Review for possible duplicates and erroneous data.
  + Remove listings without any reviews and possible duplicate rows where the bot may have re-recorded listing data.
  + **The “id” (col A) is the ID you should use to look for possible duplicates**. The “host\_id” column 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?
  + Are all the columns useful to your analysis?

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

**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:
  + The “price” column: The price for a one-night stay.
  + The “guests\_included” column: This is how many guests are included in the price.
  + The “extra\_people” column: The extra cost per person if you go above the number in the “guests\_included” column.
  + The “accommodates” column: 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).
* **A proxy number is figure that can be used to \*represent\* the value of a calculation based on a condition.**
* Step 2: Calculate estimated daily revenue for two guests, unless the accommodation explicitly accepts only one. Use the following logic:
  + 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 coding logic. Another column would then calculate the revenue per booking. Finally, multiply that by the number of total stays for the listings.
* Use Pandas built-in methods such Groupbys or Pivot Tables in Pandas in order to quickly explore the data at a high-level:
  + Groupbys and/or PivotTables should contain the host name, total revenue, and number of listings (make sure to exclude listings with no bookings).
  + Additional Groupbys and/or PivotTables are welcome and useful for gaining a better understanding of the data.

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

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

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

* Build several Groupbys and/or PivotTables in order to quickly explore the data at a high level:
  + Groupbys and/or PivotTables should contain the neighborhood (maybe host name as well) and the number of listings (make sure to exclude listings with no bookings).
  + Additional Groupbys and/or PivotTables are welcome and useful for gaining a better understanding of the data.

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

* Build several Groupbys and/or PivotTables in order to quickly explore the data at a high level:
  + Groupbys and/or PivotTables should contain the neighborhood name, number of listings (make sure to exclude listings with no bookings), and average rating.
  + Additional Groupbys and/or PivotTables are welcome and useful for gaining a better understanding of the data.

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

* Visualize insightful data for easier comprehension in support of your argument.
  + Create visualizations in Python that best fits your findings.
  + What if your client is unwilling to go into untested markets? — Consider limiting the data you visualize to the top 10 best performers **or** consider making the case for a relatively untested area of the D.C. market.
* Summarize relevant statistics for various data points (max, median, mode).
  + Identify the listing/host that generates the most revenue.
  + Provide a profile of Airbnb activity in Washington, D.C.
    - EXAMPLES: percentage that are two-bedroom apartments, percentage that use airbeds, percentage of all listings are in certain neighborhoods, etc, etc, etc.
  + Identify the top-performing hosts and listings.
  + Generate list of prospective hosts to target.
* Organize your team’s Jupyter notebook or notebooks appropriately. **THIS IS KEY.**
* Present of your findings to the class:
  + Each team will present, each student in the team will speak.
  + In addition to your conclusions, **briefly** discuss the cleaning techniques you used or include a cleaning techniques slide as an appendix to your presentation.
  + While others are presenting, take notes on the important points made by each team.

**PROJECT 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**

* Use the [Analytics Workflow](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).
* Look at some of these data source [disclaimers](http://insideairbnb.com/about.html#disclaimers).

**REQUIREMENTS & EVALUATION**

* Your presentation will be reviewed by peers and graded by instructors as a pass/fail.