**DELIVERABLES + TIMELINES**

1. **Five-minute Presentation** - during Lesson 6
   * Business needs as per your interpretation of the scenario;
   * Data selected from the original file;
   * Cleaning methods used to remove erroneous data;
   * Format: Google Slides or PDF (Keynote/PPT need to be exported); presentation will be given in small groups.
2. **Excel Workbook** - due by beginning of Lesson 7
   * Include cleansed listing data, with requested data points included (as per Requirements section);
   * Results of analysis are presented in separate worksheets, formatted and (if applicable to data) visualized;
   * Include exploratory efforts using PivotTables, visualizations, and statistical review.
   * Format: Excel file with multiple worksheets:
     + Sheet 1: Cleansed listing data, with requested data points
     + Other Sheets (as needed): Results of analysis and exploratory efforts.
3. **Data Handling** **Summary** - due by beginning of Lesson 7
   * In a separate file, summarize your approach to cleansing data. Don’t include every step but do include the rationale behind your inclusion/exclusion of data and outliers;
   * Format: Separate text file or worksheet.

*\*A detailed description of each deliverable can be found below in the “Requirements” section*

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**PROMPT**

You are doing work for a client that wishes to invest in an AirBnB hotel in Amsterdam. Before they decide to invest, they would like clear data about the AirBnB performance in that specific market: What property types receive the most positive reviews? What neighbourhoods host the most listings? How much revenue successful hosts generate? From this data, your client will be able to make a more informed decision about whether to invest and what neighborhood to invest in.

To complete this analysis, consider the analytics workflow: you have been provided with scraped data captured by a web program that pulled listing information from the AirBnB website; this data may contain unformatted data points with duplicate entries. You will want to cleanse and format the data prior to performing exploratory analysis, which will help you better understand the available data and build some business context.

You will be presenting your progress to your client’s consultants. They will want to hear the cleaning techniques you used and the metrics/insights you plan on including in the final deliverable, the data handling summary. Keep in mind, they wish to understand the local market and its potential to host a profitable AirBnB hotel - demand is key, as is the price point. You will only have five minutes of their time, so limit yourself to less than five content slides.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**REQUIREMENTS**

**1. Five-minute Presentation:**

* Include business needs as per your interpretation of the scenario;
* Identify the data selected from the original file;
* Describe cleaning methods used to remove erroneous data;

**2. Excel Workbook:**

* Calculate host account age
  + Combine the ‘host\_since’ column with the ‘host\_since\_anniversary’ column to get a full date;
  + Calculate time passed since that day to today.
* Remove duplicate data
  + Remove listings without any reviews and duplicate rows where the bot may have re-recorded listing data
* Standardize entry of ‘State’ and ‘City’ column *(North Holland vs. Noord-Holland)*
  + “Find and Replace” can be used to find all of the values of a specified text value (e.g. “Rd.”) and replace them with a different text value (e.g. “Road”). This is the same to Word or Google Docs in which you can find and replace words. For example, if you wanted to replace all of the 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 to 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.
* Estimate revenue per listing
  + Assume each booking always has 2 guests, unless the listing accommodates only one;
  + The booking is always for the minimum number of days allowed;
  + Only half of the bookings generate a review;
  + The extra person charge is per night (column name ‘extra\_people’)
  + **Format**: have a column that calculates daily revenue (account for number of guests accommodated, number of guests included in the price, extra charge for additional people - using nested IF statements); another column would then calculates revenue per booking; finally, multiply that by the number of total stays the listings has had.
* Build several Pivot Tables in order to quickly explore the data at a high-level
  + First PivotTable should contain host name, total revenue, number of listings (make sure to exclude listings with no bookings)
  + Second PivotTable should contain neighborhood name, total revenue, number of listings (make sure to exclude listings with no bookings)
  + Additional PivotTables are welcomed - and useful to you to better understand the data
  + **Format**: Currency/thousands separators should be used as appropriate.
* Visualize insightful data for easier comprehension in support of your argument
  + **Format**: copy/paste the results of a PivotTable and adjust to only show relevant findings; create chart using most applicable visualization mode
  + Your client is not willing to go into untested markets - consider limiting the data you visualize to only the top 10 best performers
* Summarize relevant statistics for various data points (max, median, mode)
  + Identify listing/host that generates the most revenue
    - xy% rating with entire 2bd apt under $200/night
  + Provide a profile of Airbnb activity in Amsterdam
    - xy% are 2bdr apartments, cd% use airbeds, st% of all listings are in X neighbourhood;
  + Identify top-performing hosts and listings;
  + Generate list of prospect hosts to target;
* Organize your workbook appropriately

**3. Data Handling** **Summary:**

* Summarize your approach to cleansing data.
  + Explain the rationale behind your inclusion/exclusion of data and outliers;

**4. BONUS: Debate for/against case made by articles using same dataset**

* Using visualizations and data used in [this article](http://insideairbnb.com/behind.html), show how author either misrepresented the findings, or provided valuable valid analysis.
* Create dynamic revenue calculation that allows to change the assumption we made (eg: only a quarter of booking leave a review; or each booking is only one ;

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**SUGGESTED WAYS TO GET STARTED**

* Look to the [Analytics Workflow Document](https://drive.google.com/open?id=0BwdiPd53UHiSWjVxeUZLRUliUWc) to guide you through each step.
* Look to these guiding questions and information to get an idea of other questions you should answer.
* Review [existing articles](https://www.google.com/search?q=insideairbnb.com#q=insideairbnb.com&tbm=nws) - on Airbnb, the short-term rental industry, from sources that quote insideairbnb.com, etc. to see how others approached the data.
* Starting off with raw, uncleansed data, use the Analytical Mindset to structure your approach to the analysis; understand the data and how it relates to the business objectives

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**RESOURCES**

* Review Specific Class Decks:
  + Format, Clean, Manipulate data (Lesson 2 & 3)
  + Dynamic Data Referencing (Lesson 4)
  + Dynamic Data Aggregation (Lesson 5)
* Data Mining [Framework](https://decisionstats.files.wordpress.com/2011/10/12345.png) and [Analytics Workflow Document](https://drive.google.com/open?id=0BwdiPd53UHiSWjVxeUZLRUliUWc)
* [Data Cleaning Walk-Through](http://schoolofdata.org/handbook/courses/data-cleaning-witchs-brew/)
* [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)
* Data Source [disclaimers](http://insideairbnb.com/about.html#disclaimers)
* [Scraped data](http://#)

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**EVALUATION**

* The Excel worksheet containing the cleaned data will be evaluated using the requirements section above as a rubric.
* The Presentation will be reviewed by your peers and graded by instructors as a pass/fail.