Data Handling Summary – AirBnB Amsterdam – Stuart King

- Created a copy of the raw data and retitled the sheet
- Searched for and removed duplicate entries using the Remove Duplicates tool in Excel (20 entries found and deleted) – I did not use the ID column to search for and remove duplicates as I interpreted each unique line as its own listing and the ID is simply a host identifier, not a listing identifier. The same host can have multiple listings/properties.
- Developed a color-coding pattern of coloring new columns with formulas in blue and new columns with copy/pasted values in orange.
- Created a combined data field by combining the host_since_year and host_since_aniversary columns to create a "host_since_full_date" column that lists the year by mm/dd/yyyy (used CONCATENATE to combined the two columns)
- Created a new column that calculated the number of days since the listings anniversary subtracting the host anniversary date from April 10, 2016 (used the DATE function as part of the formula)
- Copied the data from the city column and pasted values into a new column to be cleaned. I began cleaning the city data by using the filter tool to see which values are part of the column and then used find & replace to rename misspellings of Amsterdam. I performed a Google search to determine if any of the cities listed in the raw data were indeed cities, and if so I left those cities in the list. Also using Google I determined if any of the cities in the raw data were neighborhoods of Amsterdam, and if so I filtered the data and then renamed to Amsterdam.
- To clean the state column I added a new column and then used an IF formula to change all cell values that were not North Holland to North Holland as all the cities included in the cleaned city column are within North Holland. Copy and pasted values into another column for the final cleaned data.
- Added a column to calculate daily revenue using one nested IF function using the calculation criterion as stipulated in the Project 1 instructions. I then copy and pasted values of the computed daily revenue column to create a clean column for daily revenue.
- To calculate total revenue I added a column to determine the revenue per booking by multiplying the daily revenue by the minimum night stay requirement, then in a separate column I calculated the total number of bookings for each listing by using the assumption that only half of bookings leave reviews (multiplied the number of reviews by two), and finally created a new column that multiplies the revenue per booking by the total number of bookings to arrive at the total revenue for each listing. After performing the calculations I created a new column and copy and pasted values for a clean total revenue column.
- For blank entries in the number of bedrooms column I added a column and used an IF function based on the number of beds. If there were 2 or less beds, then the listing would be assumed to have 1 bedroom; if there were more than 4 beds, then the listing would have 3 bedrooms; and finally if there were 3 beds then the listing would be assumed to have 2 bedrooms.
- For blanks in the bed column I added a column and used an IF function based on how many people a listing could accommodate and how many bedrooms the listing had. If a cell was blank, the IF function would first check if the number of people the listing could accommodate was less than or equal to 2. If so, then a value of one would be returned. If not, and the number of people the listing could accommodate was greater than or equal to 4 AND the number of bedrooms was greater than 1, then a value of 2 would be returned; if not, a value of 1. I then created a second formula column to confirm that none of the listings had more than two times the number of beds than it had bedrooms. I copy and pasted values once these two IF formulas were written.
- To calculate an annualized revenue I performed the following:
    1. Calculated each listing's occupancy days (minimum nights x number of bookings)

2. Calculated each listing's occupancy rate using the number of days the host has been on AirBnB (this is a potential flaw in the calculation since it's possible hosts added listings at different times!) and total occupant days
3. Checked to ensure no listings had an occupancy rate above 100%
4. Calculated an annualized revenue by multiplying the daily revenue x 365 x occupancy rate

- Used CONCATENATE to combine the host name and host ID to develop a unique identifier for each host. As part of this process I first trimmed the host name field to ensure names there were not double entries of the same host.
- Created two separate pivot tables per the project's instructions, placing each table in separate tabs and labeling each tab as such.
- I then created a separate pivot (Pivot 3) to identify the highest performing neighborhoods by average annual revenue. Following this I analyzed the market share of the top ten hosts in each neighborhood. I did this by creating a separate pivot table that pulled the total annualized revenue for each host in each neighborhood. I then copy and pasted the values into a separate table and used the RANK function to identify the top ten host by annual revenue generated. In the Data Pool tab I used the INDEX and MATCH functions to pull the names of the top ten hosts for each neighborhood by rank, and then used the INDEX and MATCH functions to pull the top hosts' total annual revenue in each neighborhood. I then summed the top hosts' total annual revenue and compared this sum to the total annualized revenue for each neighborhood to arrive at a market share percentage for the top hosts.
- The market share analysis aligned with the top neighborhoods based on average annual revenue and five neighborhoods were identified to further analyze using other variables. The top five neighborhoods are: **Centrum-West, Centrum-Oost, De Baarsjes – Oud-West, Westerpark, and De Pijp – Rivierenbuurt.**
- Once the five top neighborhoods were identified I tailored my remaining analysis to filter for these neighborhoods. I created multiple more pivot tables that only calculated listings that had bookings and were located in the target neighborhoods. The following pivots were created:
    - Average annual revenue per year (to calculate the compound annual growth rate)
    - Average annual revenue per property type
    - Average annual revenue per the number of beds
    - Average annual revenue per the number of bedrooms
    - Average annual revenue per bed type
    - Average annual revenue per room type
- To aid my analysis I created a separate tab (Frequency) to compute descriptive statistic metrics. I did this by first filtering the cleaned data for the target neighborhoods and then pulling the annualized revenue for all listing in these neighborhoods into the newly created tab. From there I calculated the mean, median, mode, max, min, and count of the data. I then manually calculated the standard deviation of the data and verified it using the Excel formula for standard deviation (I used the population version of the formula). Then I calculated the frequency of listings that fell with $10,000 bins and created a histogram of the results, visually depicting the right skewness of the data and the vast concentration of observations within the $0-10,000 bin.
- I then created a summary tab into which I pulled the most relevant data that I wanted to use to create the visualizations that would be used in my presentation.