BUSINESS INTELLIGENCE FINAL PROJECT - SUBMITTED TO PROFESSOR NICOLAS PRAT

SHAHMIR KAZI - B00771255
ESSEC BUSINESS SCHOOL – CENTRALESUPELEC
MASTER IN DATA SCIENCES AND BUSINESS ANALYTICS

INTRODUCTION

The project is intended as a briefing to the Product Management and Marketing divisions of Airbnb USA, to present granular analysis and present insights with relevant action points.

Problems addressed and decisions supported

Currently the team does not insights for the State of New York. We will hence delve into the below two broad areas to provide actionable insights:

1. Profitability
2. Consumer Satisfaction
3. Sales and Marketing policy

We will therefore assess on a State, City and Neighborhood level.

The decisions required include the below:

1. Which Cities and particular neighborhoods are performing well in terms of listing and which lag behind?

2. How is Price actually influenced?
3. How to incentivize the Sales team?

This is important as Price is directly proportional to Profit Margin, due to Airbnb earning a commission on the price of the listing.

4. Should Sales incentives and Product Management vary for Hosts vs Super Hosts?

5. Some hosts earn more than others, but what is/are the factor(s) behind it?

6. How to analyze the reviews that we have on the Airbnb platforms?

7. How to train the Sales team and prepare the product pitch for pitching to Hosts/ Super Hosts?

8. Based on current listings, have an insight to prepare City wise budget for next year after comparing our listings with the retail audit routinely conducted by the Marketing research division?

Ex-ante justification for the choice of tools
For this analysis, Tableau and Python will be used.

Tableau will be used extensively for data visualization and deriving insights from the visuals. It will be the final product of this project in the form of Tableau Dashboards and Story.

It will also allow us to perform feature engineering on the data to enhance analytics. The aim is to follow a data discovery process along with a top down approach explained above.

Other alternatives are only Python and R but they have limited capabilities for intuitive and **interactive** visualizations and Microsoft Power BI was not considered for this task.

Python will be used for data preparation and cleaning, and the output will be forwarded to Tableau for analysis mentioned earlier.

In some cases, Python will be used solely for tasks that require enhanced algorithm complexity or that required custom solutions compared to inbuilt Tableau functionality.

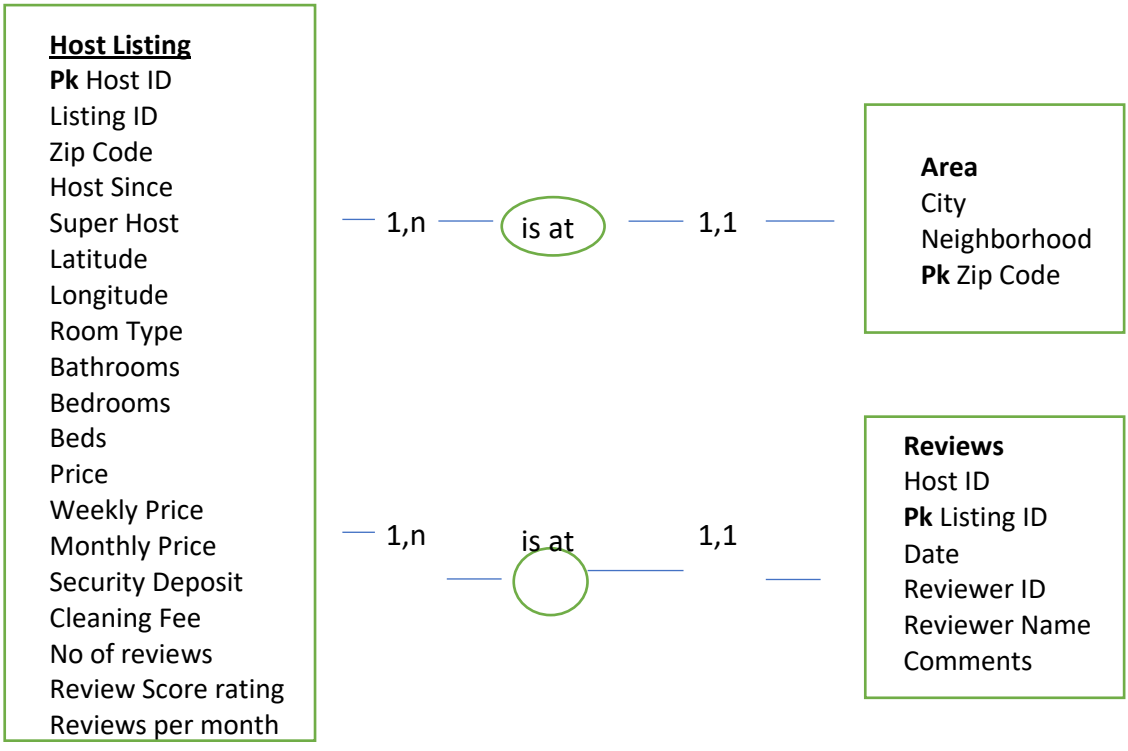
This is in line with our aim of performing Explainable AI which is imperative given the abundant digital exhaust available in today's day and age.

DATA MODELING AND PREPARATION

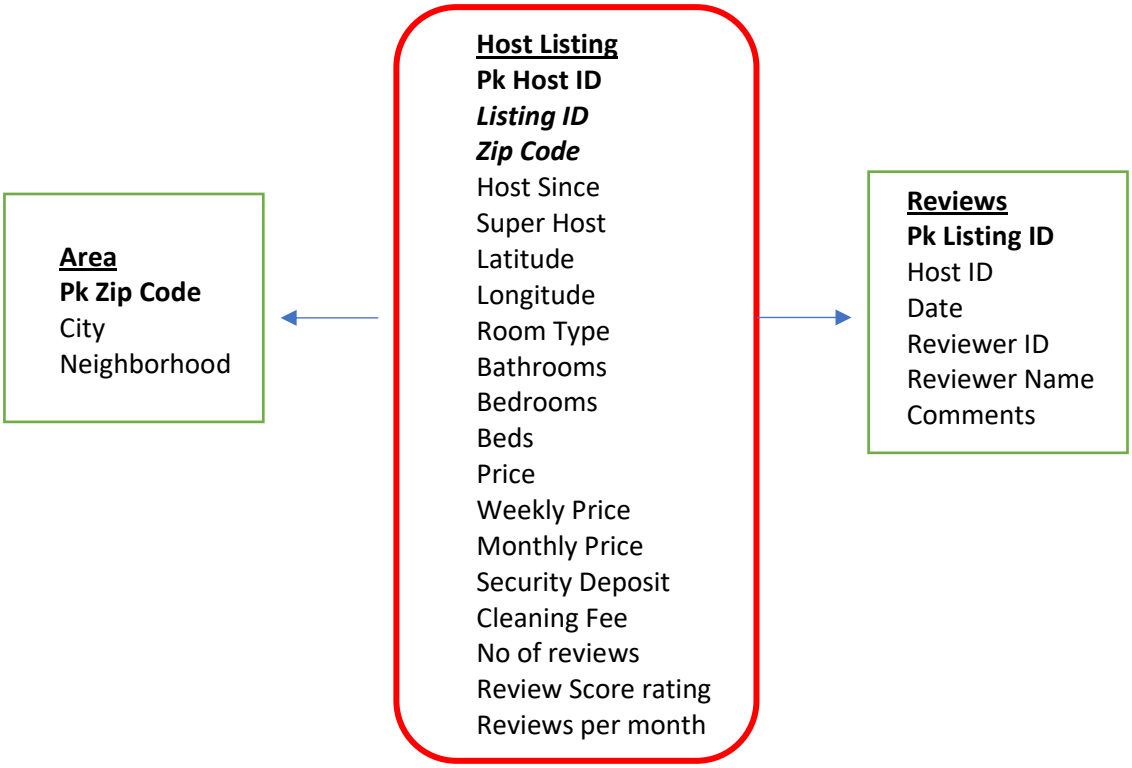
Modeling
The data we have at hand consists of mostly two broad areas:

- Hosts, Listings and Price
- Reviews

Conceptual Modeling - ER schema
(Illustration not conduced in JMerise as Mac doesn't support Win'Design and JMerise is in French)



Logical Mapping - Snowflake Schema



Data Preparation

For data preparation Python was used with the Jupyter IDE. The file ***"airbnb_bi_project_nyc_homes_rawdata.csv"*** was taken as input and the below actions were performed, which are shown in the submitted Python notebook:

Initial state = 5 column s x 96 rows

- ***Is.na()*** shows missing values and they are accounted for.

- After cleaning this, 21 columns are chosen which are relevant for our analysis. Columns such as jurisdiction_names and licence are dropped as they provide no additional information for the 8 questions we were are seeking answers for.

- Renaming of some columns
- Dtype inspection
- Changing zipcode to numeric and mistakes corrected manually
- Cleaning the dataset by ***drop.na()***

The results are exported as ***"airbnb_bi_project_nyc_homes_cleaneddata.csv"***

Feature Engineering

These steps are completed in Tableau for ease of use.

- Hierarchy is created with: City → Neighborhood → Zip Code

- One field created named “Total Size”, for showing Total Area in the absence of square meters measurements. Total Size = Rooms + Bedrooms + Bathrooms.
- Bins for Total Size were created at interval of 1.
- Bins for Price were created at internal of 20
- Bins for Review Score Rating were created at interval of 1

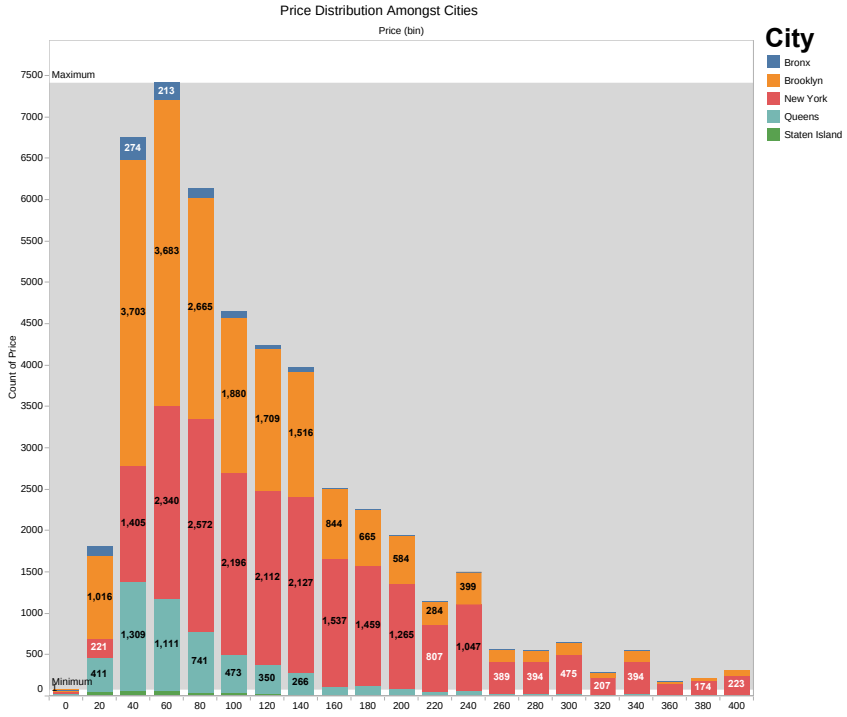
APPLICATION

Answers to the 8 questions posed by the PM team will be solved:

1. Important Locations

Homes can be found between $40-60-80 within Brooklyn, but as the price point increases, New York takes the lion’s share. The other two cities within New York State are not that popular/highly visited.

Analysis is done with Count of Price and Price (bins of 20) on the axis.



Count of Price for each Price (bin). Color shows details about City. The marks are labeled by count of Price. The view is filtered on Price (bin), which keeps 21 of 121 members.

The below tree map shows the listings and follows our **Funnel Approach**, as we have already highlighted the popular cities and now will highlight the popular neighborhoods within those cities.

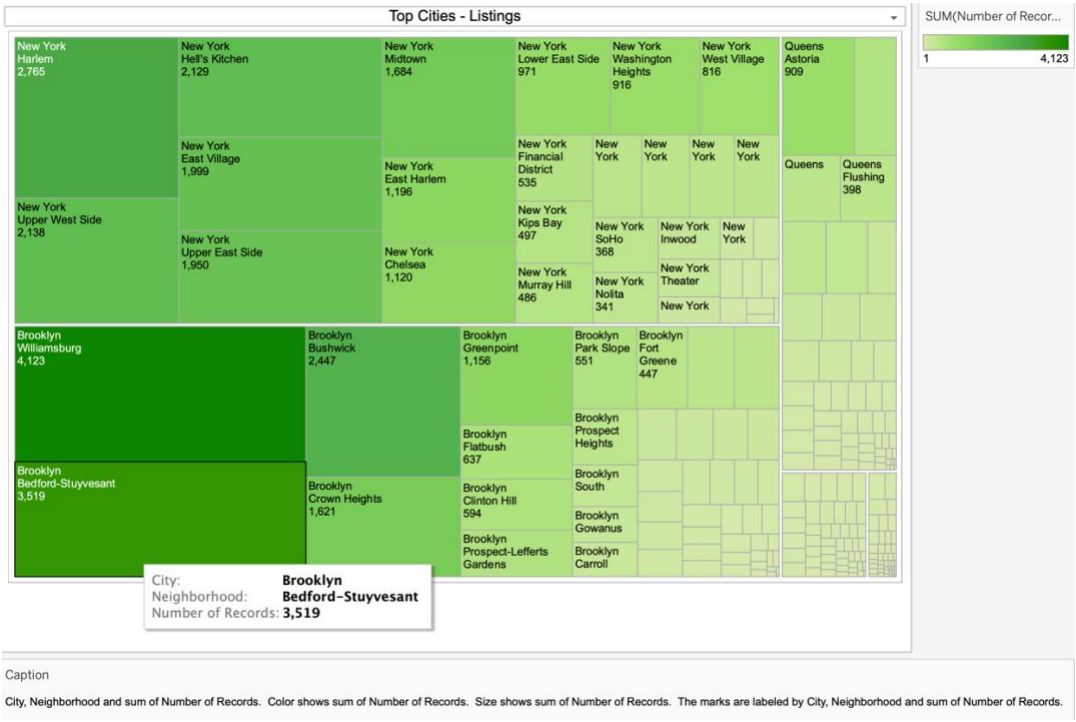
As price points are established, listings provide a good benchmark for neighborhoods.
We see that 5 neighborhoods in New York and 4 in Brooklyn deserve the highest attention from the PM and Marketing team. They are the **growth driver** for Airbnb in New York State.

9 Neighborhoods with top listings are mentioned below for focusing policy upon:
New York

- Harlem – 2,765
- Upper West Side – 2,138
- Hell's Kitchen – 2,129 (the team suspects it may be due to Gordon Ramsay but data is not provided for this hypothesis)
- East Village – 1,999
- Upper East Side – 1,950

Brooklyn

- Williamsburg – 4,123
- Bedford – 3,519
- Bushwick – 2,447
- Crown heights – 1,621

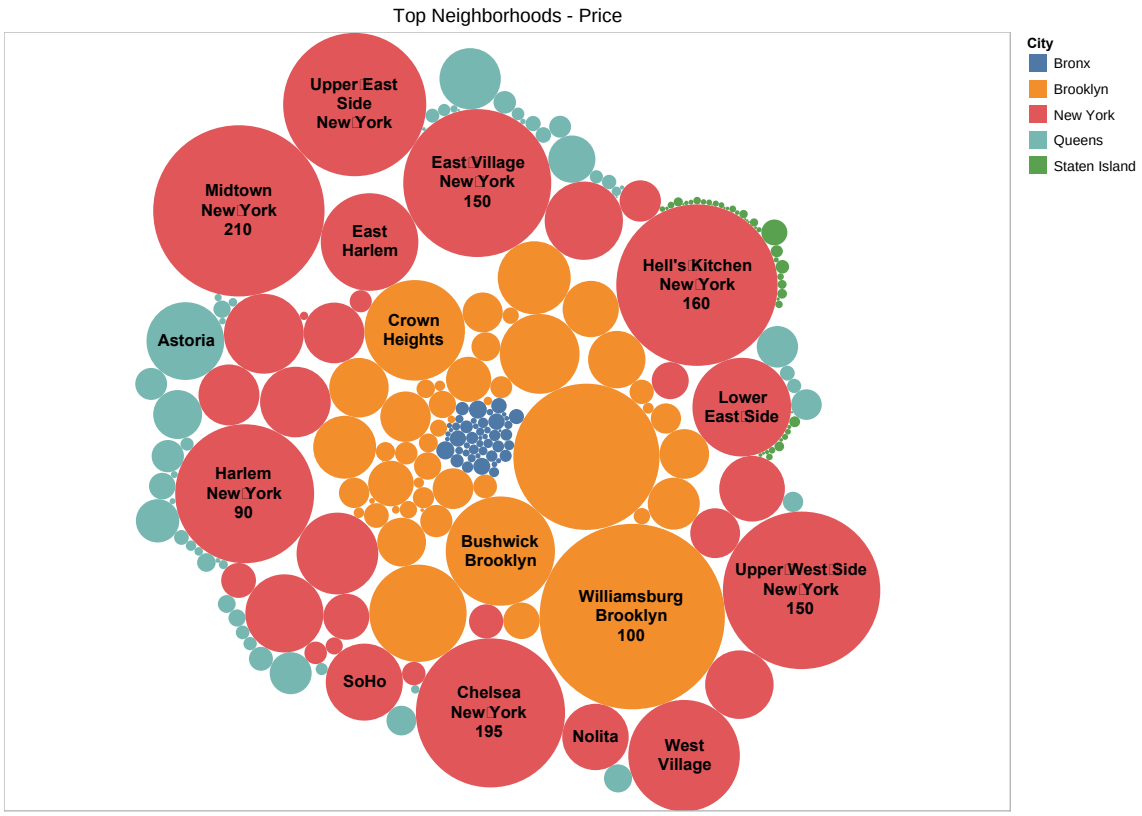


2. **How is Price actually influenced?**
3. **How to incentivize the Sales team?**

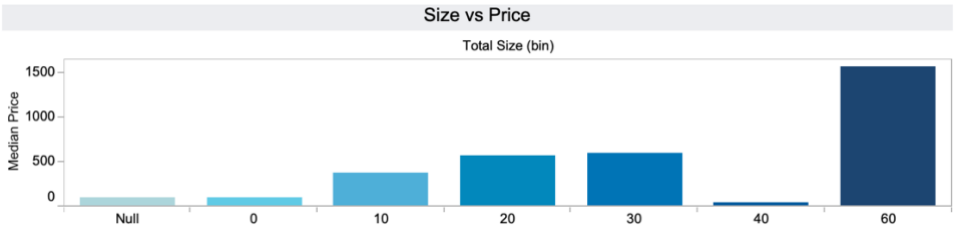
The bar graph below shows that apparently Price increases as Total Size of the property increases. When regressed, the R^2 was found to be 23% which shows a really low correlation.

Therefore, the packed bubbles map below shows how price drastically increases in Midtown and Upper East Side of New York compared to WilliamsBurg in Brooklyn.
Unless specifically mentioned, price refers to Median Price

The insight from this is to increase high margin properties such as resorts and suites/commercial spaces to increase margin. **SEC B+ and above** are already the consumer segment being targeted and after knowing that Airbnb price (**profits) are dependent on location**, the Sales and Marketing teams must align efforts accordingly.



Neighborhood, City and median of Price. Color shows details about City. Size shows sum of Price. The marks are labeled by Neighborhood, City and median of Price.



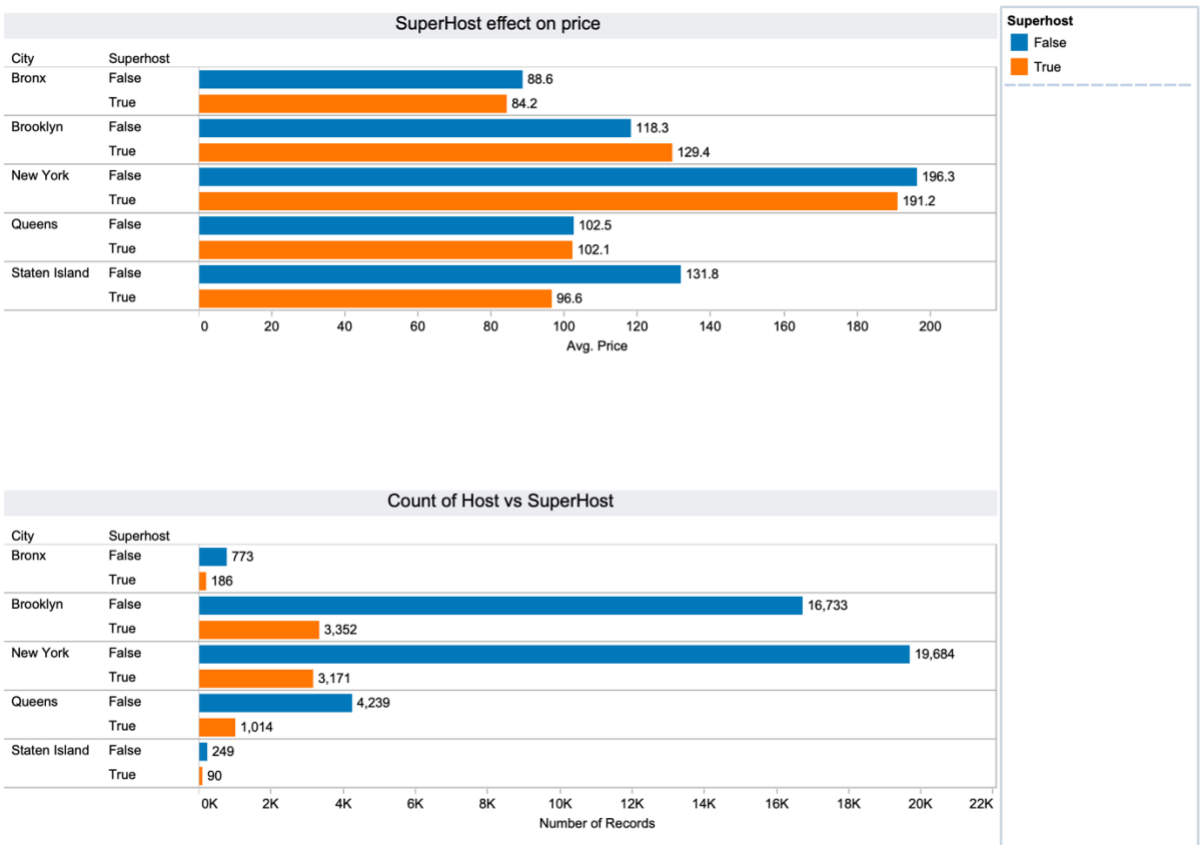
4. Should Sales incentives and Products vary for Hosts vs Super Hosts?

The first graph shows that Super Hosts don not earn a higher margin for Airbnb than the regular hosts., In fact, as shown below, for some sections Regular Hosts earn more.

Furthermore, Regular Hosts, as shown in the second graph, form the major chunk of Airbnb business as their count is 5x compared to Super Hosts in New York and Brooklyn, the major two cities we have decided to focus upon after the first two illustrations.

Hence, to protect **and** grow business, Regular and Super Hosts must be given the same incentives. Super Hosts are already benefitting from an **implied preference** given to them through the consumer side as consumer perceives Super Host as more credible (also confirmed later through Sentiment Analysis).

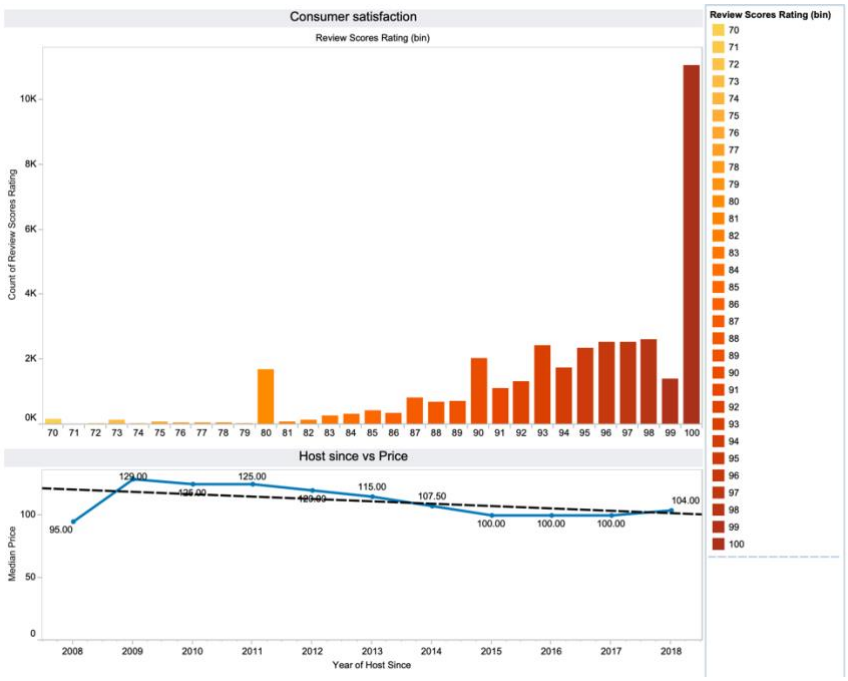
Another learning is that as the major chunk is new listings, they are usually **price sensitive** and would require focused efforts to avoid attrition towards Booking or Expedia etc.



5. **Some hosts earn more than others, but what is/are the factor(s) behind it?**

This is a key learning for the commercial teams as it shows that older hosts, rather than Super Hosts, earn more for the company in terms of unit price. Hence, loyalty programs or consumer promotions targeted at retention will have a higher ROI per listing when targeted towards older hosts.

Also, we learn that most reviews are rated very highly by all occupants. Hence a deep dive is required to uncover reasons of consumer delight or dissatisfaction. A learning about over serving consumers is discussed in the Conclusion part.



6. **How to analyze the reviews that we have on the Airbnb platforms?**
7. **How to train the Sales team and prepare the product pitch for pitching to Hosts/ Super Hosts?**

For making sense of reviews, advanced Sentiment Analysis is used in Python using TextBlob and VaderSentiment. The code is shown below to highlight how the findings are uncovered, also because this report is intended as a guide and explanation to the workings.

Negative, Positive and Neutral reviews are found using each model and then taking an average of the two models.

Hence, we receive:

- 82,819 Positive reviews
- 682 Negative Reviews
- 534 Neutral Reviews

This is great news for all the teams as it shows that Customer Service, Quality Assurance and all cross functional teams have succeeded in their goal to serve the customer.

To drill down further, uni, bi and tri gram analysis was conducted to get the most used words and phrases, in negative and positive comments both. Hence, we learn that in both cases, the consumer perception and experience is shaped by the following:

1. Walking distance
2. Clean place
3. Near Downtown
4. Like Home
5. Friendly Host

Therefore, the PM and Sales team has to focus on New York and Brooklyn and that too in Downtown and City centers, ensuring the places are kept clean and regular trainings for the hosts is arranged to make sure positive experiences are created.

```
review = review[review['automated_posting']==False]
print('Actual reviews:', review.shape[0])
```

```
Actual reviews: 84035
```

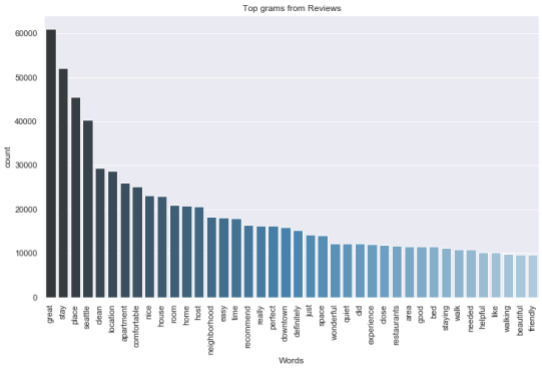
```
def getTBSubjectivity(row):
    return TextBlob(row).sentiment.subjectivity

def getTBPolarity(row):
    return TextBlob(row).sentiment.polarity

def getVDPolarity(row):
    return analyzer.polarity_scores(row)['compound']

analyzer = SentimentIntensityAnalyzer()
review['review subjectivity Textblob'] = review['comments'].apply(getTBSubjectivity)
review['review polarity Textblob'] = review['comments'].apply(getTBPolarity)
review['review polarity Vader'] = review['comments'].apply(getVDPolarity)
review['Average Polarity'] = (review['review polarity Vader'] + review['review polarity Textblob']) / 2

review['review Sentiment Textblob'] = np.where(review['review polarity Textblob']>= 0.01, 1, (np.where(review['review p

review['review Sentiment Vader'] = np.where(review['review polarity Vader']>= 0.05, 1, (np.where(review['review polarit

print(review['review Sentiment Textblob'].value_counts())
print(review['review Sentiment Vader'].value_counts())
```

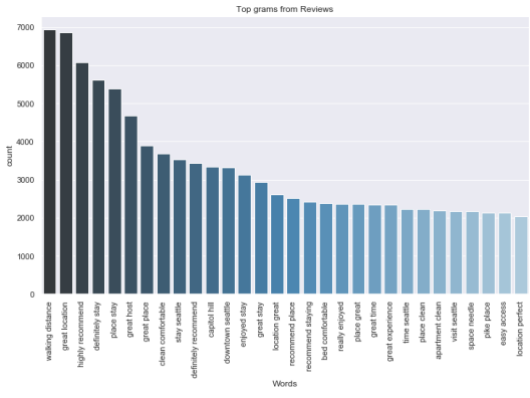
```
 1    82711
 0      961
-1      363
Name: review Sentiment Textblob, dtype: int64
 1    82524
 0      791
-1      720
Name: review Sentiment Vader, dtype: int64
```

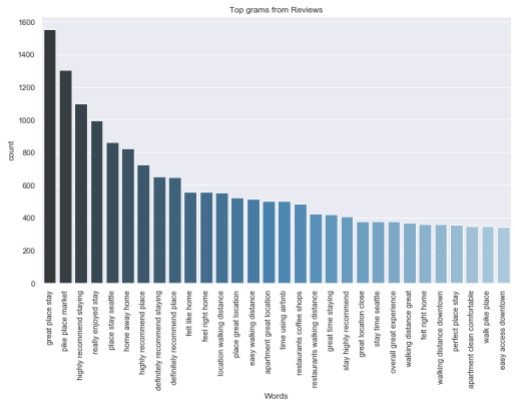
```
review['Sentiment (lexicon)'] = np.where(review['Average Polarity']>0, 'Positive',\
                                (np.where(review['Average Polarity']<0, 'Negative','Neutral')))
print('Combined sentiment from both methods:\n',review['Sentiment (lexicon)'].value_counts())
```

```
Combined sentiment from both methods:
 Positive    82819
Negative       682
Neutral        534
Name: Sentiment (lexicon), dtype: int64
```

```
#Now we count the frequency of N-grams
def getTopNGrams(corpus, gram, n):

    vec = text.CountVectorizer(ngram_range = (gram, gram), stop_words = englishStopwords)
    bag_of_words = vec.fit_transform(corpus)
    sum_words = bag_of_words.sum(axis=0)

    words_freq = []
    for word, index in vec.vocabulary_.items():
        words_freq.append((word, sum_words[0, index]))

    # Sort list of tuples in decending order
    #By the count
    #Hence also shown as True
    words_freq = sorted(words_freq, key = lambda x: x[1], reverse = True)

    # Return the first n elements
    return words_freq[:n]
```







8. **Based on current listings, have an insight to prepare City wise budget for next year after comparing our listings with the retail audit routinely conducted by the Marketing research division?**

Yes, the budget must be focused in major districts and teams must price aggressively for consumers while focusing on retention of existing and new hosts. KPIs and budgets may be constructed accordingly.

CONCLUSION

The use of Tableau and Python has been appropriate for these tasks as some tasks required custom made solutions and algorithms while others were depicted visually.

The use of Tableau for feature engineering and creating new variables is useful as is the use of Python for advanced tasks. **TabPy** was used for gauging Sentiment Polarity but due to n-gram approach and more advanced use of Python, it was not shown in the report or tableau workbook.

Additional analysis to be done for Airbnb may include:

- Check price sensitivity if prices may be increased across the board. This is based on the extremely high review scores across all listings. If Airbnb is overserving consumers and is able to raise prices, it is direct positive impact on their bottom line. Hence an optimal balance must be struck as attrition would also need to be kept in check.
- Rather than Linear Regression, XG Boost regression may be used for such large datasets for parallel processing and it will help analyze the pricing decisions due in the future.
- Aspect Based Sentiment Analysis may be used on top of Vader and TextBlob to gauge reviews that have both positive and negative lexicons.

In terms of learning, this project has helped me prepare extensively for future pitches to C level executives as telling a story with data is as important as running advanced algorithms.

Another learning is that research is different from business and it build upon by 4 years of work experience before joining this MS DSBA degree.

This project has also helped me prepare for the Tableau Desktop Specialist certification.

SOURCES

- Data set, *Inside Airbnb*
  http://insideairbnb.com/get-the-data.html

- Inspiration for dashboarding, *Tableau Public*
  https://public.tableau.com/views/AirBnBProjectDashboard/AirBnbBostonListings?:embed=y&:showVizHome=no&:display_count=y&:display_static_image=y&:bootstrapWhenNotified=true

- Analytics in Tableau, *DataCrunch*
  https://datacrunchcorp.com/tableau-predictive-analytics-linear-regression/

- TabPy documentation, *Tableau*
  https://www.tableau.com/about/blog/2020/1/python-tableau-now-10

- Sentiment Analysis with Python and Tableau, *Medium - Towards Data Science*
  https://medium.com/@liana.melissa12/sentiment-analysis-using-python-in-tableau-with-tabpy-ec492db783f3