Sentiment Analysis : Digging into Airbnb Data

Jayant Kumar
CS - Graduate
Steven Institute of Technology

Malika Thakur
CS - Graduate
Steven Institute of Technology

Team Name : Opinion Pickers

Abstract:

Airbnb is the main and quickly developing option in contrast to the customary inn organizations. It gathers a great deal of information about their hosts and their properties, including point by point sight seers' audits. Curiously, this information is openly accessible from the Inside Airbnb site, so it is normal for an information researcher to delve into, pose inquiries and attempt to respond to them. In this project, various parameters have been taken care viz. reviews (positive or negative) about the host, distinction of super hosts from the ordinary hosts, renting prices factor on daily basis and prediction of daily renting price based on the city.

1. Introduction:

Today the travel industry has become coordinated piece of our life and inside recent years, this industry has picked up a lot of prominence. Individuals nowadays need to see world at least costs. Lodgings are one of the principal concerns while making arrangements for a get-away, work excursions and so forth Airbnb is a notable organization which has given a stage to hosts and guests to convey. They give a stage where explorer try not to need to stress over genuineness of host and the spot. Airbnb is the world's biggest home sharing organization and has more than 800,000 postings in excess of 34,000 urban communities and 190 nations. Whole site is in light of the rating framework, which is a both way measure. Host who offers the spot likewise gets rating on various boundaries and visitors are likewise evaluated based up their conduct. This audit framework permits clients to choose place dependent on their financial plan and necessities.

Surveys assumes a significant job in catching individuals' feelings and conclusions about a wide scope of items and administrations. Surveys gives a stage which permits individuals to share or communicate their fair-minded sentiments. Thoughtfulness regarding the suppositions and input which visitors give about the items and administrations through audits is a basic factor to the achievement of the hosts in the commercial center. This project moreover centers around understanding what visitors enjoyed and disdained which might be one of the pioneer supporters of the income produced by the organization and furthermore the host.

2. Data set Analysis:

Context

Melbourne was announced as 6th on the list of top ten cities for users globally in 2016 and has been one of the top cities for listings globally since then. As part of the InsideAirbnb initiative, this dataset describes the listing activity of homestays in Melbourne, VIC, Australia. The dataset was compiled on 07 Dec 2018.

Content

The following Airbnb activity is included in this Melbourne Airbnb dataset:

- **Listings:** detailed listings data including full descriptions and average review score.
- **Calendar**: detailed calendar data for listings, including listing id and the price and availability for that day.
- **Reviews**, detailed review data for listings including unique id for each reviewer and detailed comments.
- **Listings-Summary:** summary information and metrics for listings (good for visualizations).
- **Reviews-Summary:** summary Review data and Listing ID (to facilitate time-based analytics and visualizations linked to a listing).
- **Neighborhoods:** neighborhood list for geo filter. Sourced from city or open-source GIS files.

Inspiration

- How is Airbnb really being used in and affecting your neighborhoods?
- Can you describe the vibe of each Melbourne neighborhood using listing descriptions?
- What are the busiest times of the year to visit Melbourne? By how much do prices spike?

- Is there a general upward trend of both new Airbnb listings and total Airbnb visitors to Melbourne?

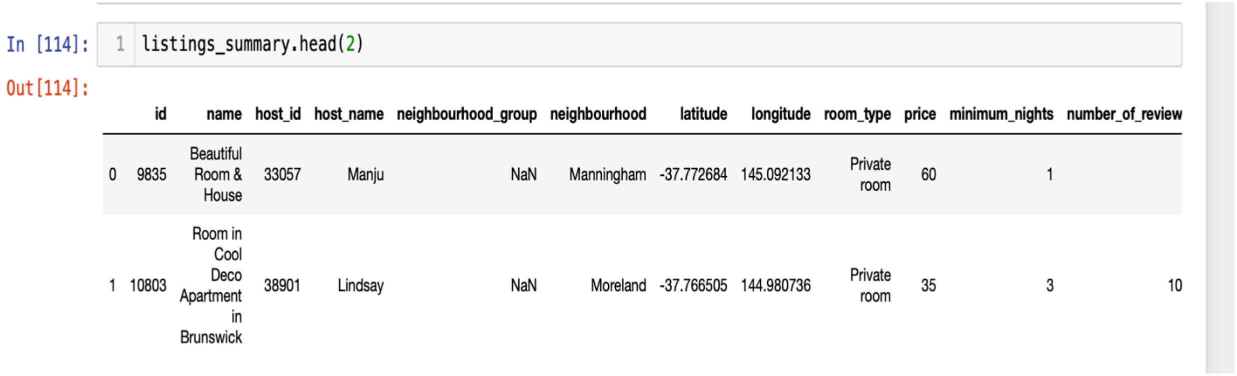
Acknowledgements

This dataset is part of InsideAirbnb, and the original source can be found: http://insideairbnb.com/

2.1 Airbnb property mapping data:

"listings_summary_dec18.csv" contains the property_id, name of each property, host_id, host_name, neighborhood reviews, location, room type and daily rental price of the property.

The snippet of data set is shown below:

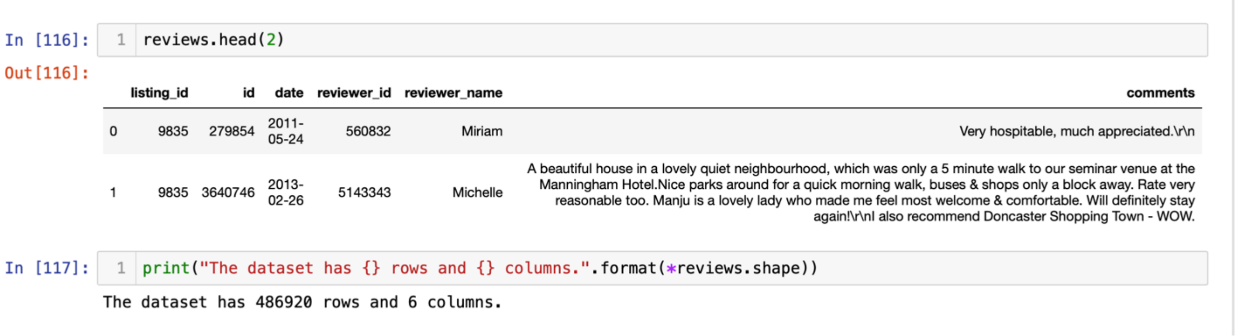


```
In [114]: listings_summary.head(2)
Out[114]:
```

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_review
0	9835	Beautiful Room & House	33057	Manju	NaN	Manningham	-37.772684	145.092133	Private room	60	1	
1	10803	Room in Cool Deco Apartment in Brunswick	38901	Lindsay	NaN	Moreland	-37.766505	144.980736	Private room	35	3	10

2.2 User review data:

"reviews_dec18.csv" contains listing_id on Airbnb site, review ID, date of review posted, Guestid, Guest_name and comments.



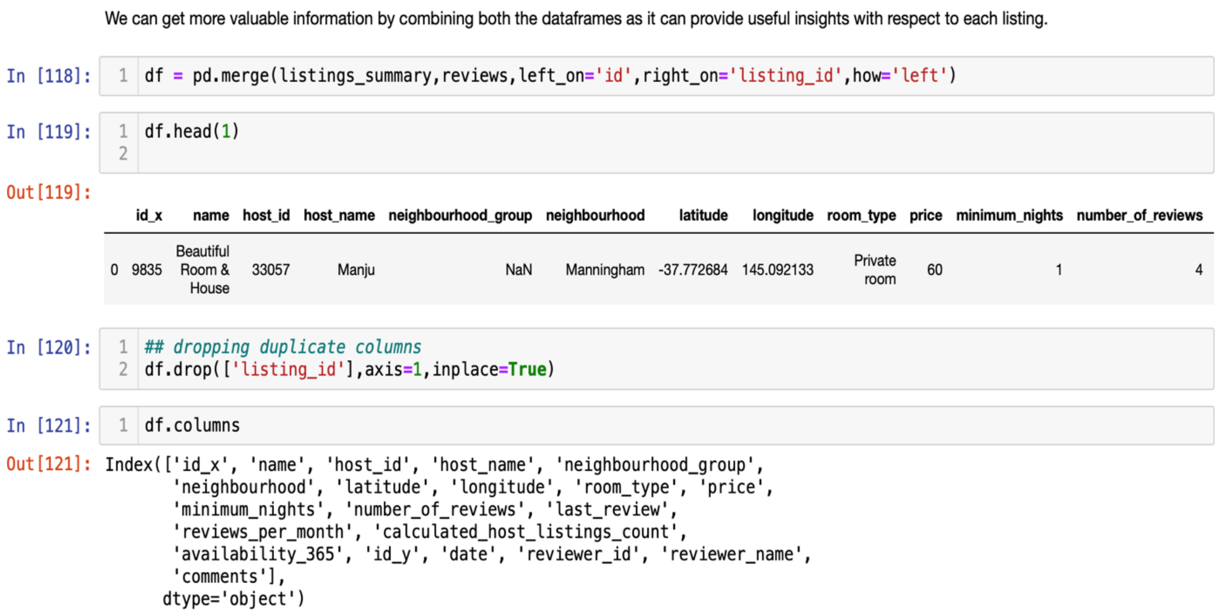
```
In [116]: reviews.head(2)
Out[116]:
```

	listing_id	id	date	reviewer_id	reviewer_name	comments
0	9835	279854	2011-05-24	560832	Miriam	Very hospitable, much appreciated.\r\n
1	9835	3640746	2013-02-26	5143343	Michelle	A beautiful house in a lovely quiet neighbourhood, which was only a 5 minute walk to our seminar venue at the Manningham Hotel.Nice parks around for a quick morning walk, buses & shops only a block away. Rate very reasonable too. Manju is a lovely lady who made me feel most welcome & comfortable. Will definitely stay again!\r\nI also recommend Doncaster Shopping Town - WOW.

```
In [117]: print("The dataset has {} rows and {} columns.".format(*reviews.shape))
The dataset has 486920 rows and 6 columns.
```

2.3 Data Exploration:

Code snippet:



We can get more valuable information by combining both the dataframes as it can provide useful insights with respect to each listing.

In [118]:
```
df = pd.merge(listings_summary,reviews,left_on='id',right_on='listing_id',how='left')
```

In [119]:
```
df.head(1)

```

Out[119]:

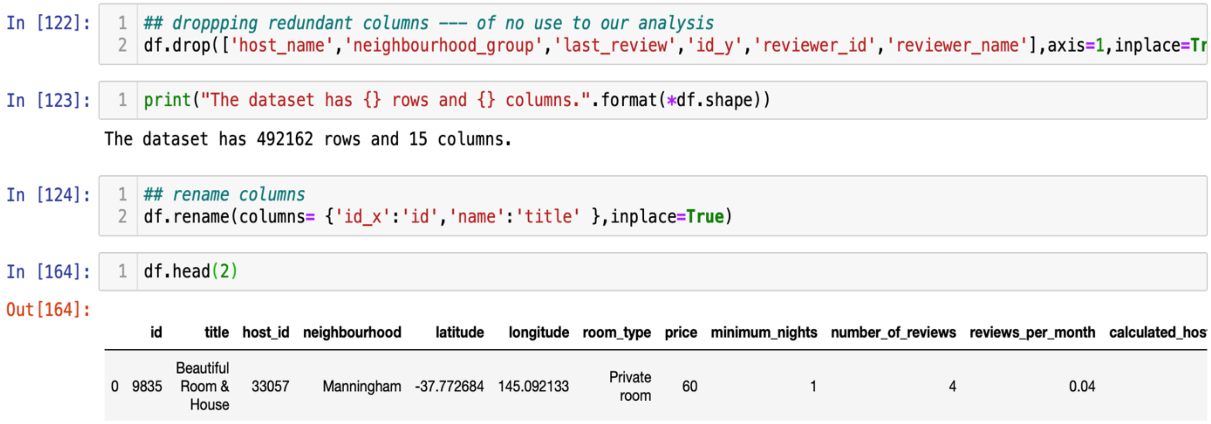
	id_x	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews
0	9835	Beautiful Room & House	33057	Manju	NaN	Manningham	-37.772684	145.092133	Private room	60	1	4

In [120]:
```
## dropping duplicate columns
df.drop(['listing_id'],axis=1,inplace=True)
```

In [121]:
```
df.columns
```

Out[121]:
```
Index(['id_x', 'name', 'host_id', 'host_name', 'neighbourhood_group',
       'neighbourhood', 'latitude', 'longitude', 'room_type', 'price',
       'minimum_nights', 'number_of_reviews', 'last_review',
       'reviews_per_month', 'calculated_host_listings_count',
       'availability_365', 'id_y', 'date', 'reviewer_id', 'reviewer_name',
       'comments'],
      dtype='object')
```

Removing the redundant data columns from the Dataset:



In [122]:
```
## droppping redundant columns --- of no use to our analysis
df.drop(['host_name','neighbourhood_group','last_review','id_y','reviewer_id','reviewer_name'],axis=1,inplace=Tr
```

In [123]:
```
print("The dataset has {} rows and {} columns.".format(*df.shape))
```

```
The dataset has 492162 rows and 15 columns.
```

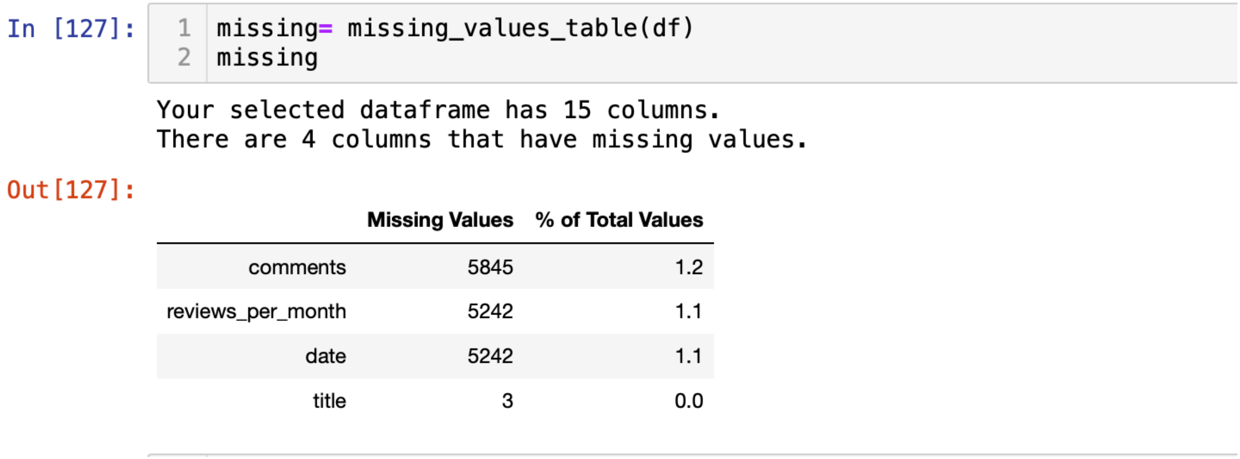
In [124]:
```
## rename columns
df.rename(columns= {'id_x':'id','name':'title' },inplace=True)
```

In [164]:
```
df.head(2)
```

Out[164]:

	id	title	host_id	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_hos
0	9835	Beautiful Room & House	33057	Manningham	-37.772684	145.092133	Private room	60	1	4	0.04	

Checking the missing values:



In [127]:

```
missing= missing_values_table(df)
missing
```

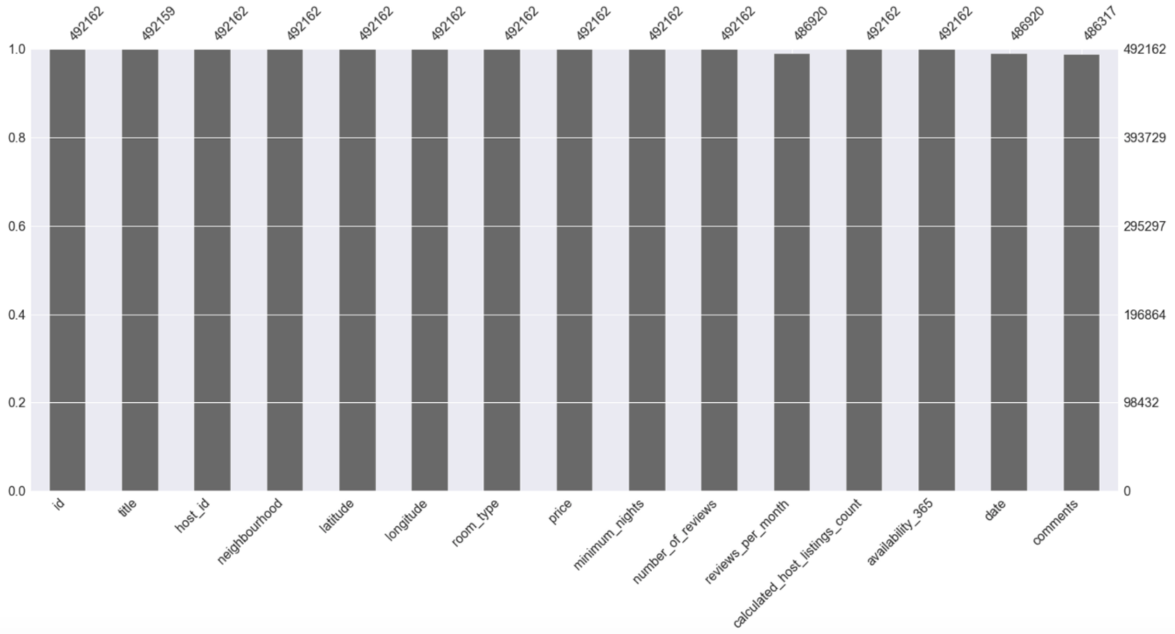
```
Your selected dataframe has 15 columns.
There are 4 columns that have missing values.
```

Out[127]:

	Missing Values	% of Total Values
comments	5845	1.2
reviews_per_month	5242	1.1
date	5242	1.1
title	3	0.0

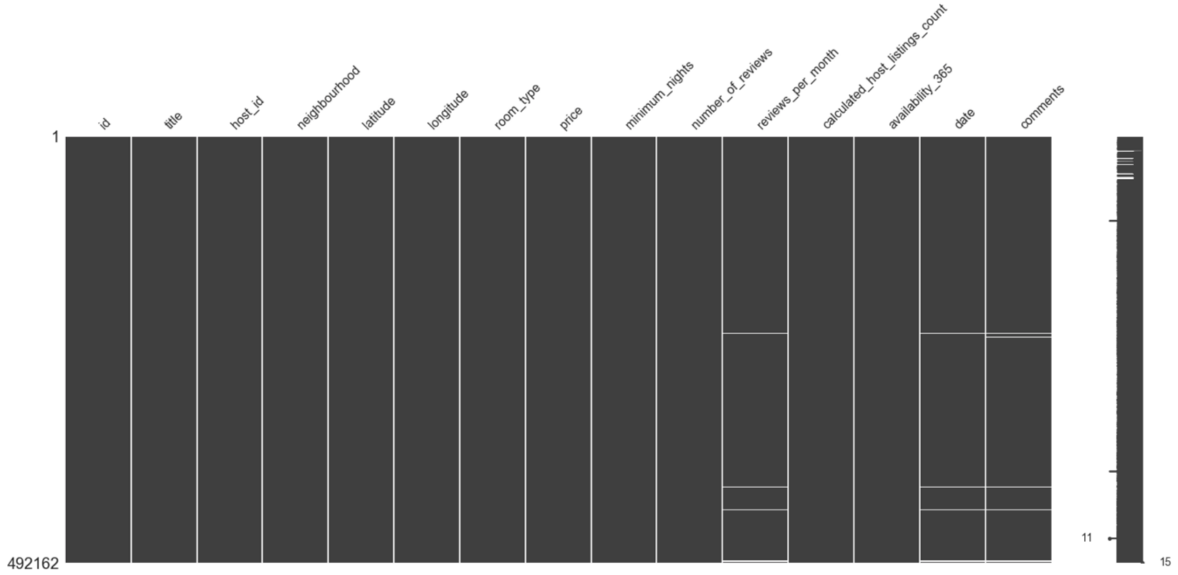
Graphical outline of the fulfillment of the dataset:

We can see that remarks, reviews_per_month,date and title segments have missing qualities. Next, it would bode well to discover the areas of the missing information.

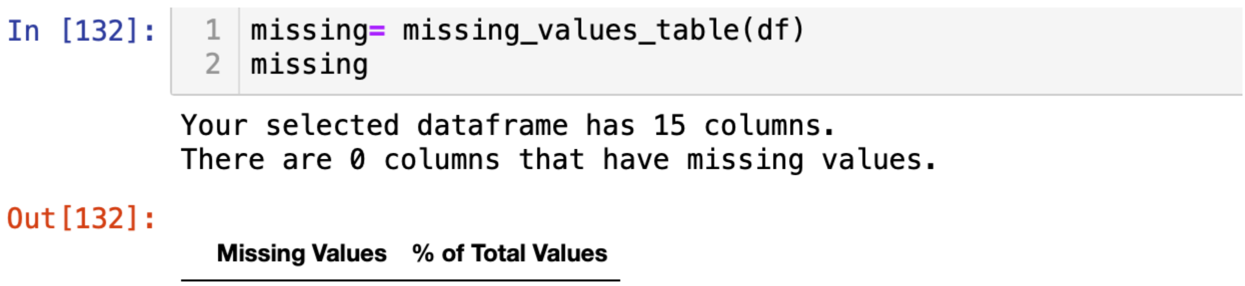


Correlation Matrix:

We can these all-missing qualities are profoundly corresponded for all these 4 segments. From above perception it is very obvious these missing qualities has a place with similar perceptions. Since the rate is quite low, we are going it until further notice.



- **No missing values in final dataset:**



In [132]:
```
missing= missing_values_table(df)
missing
```

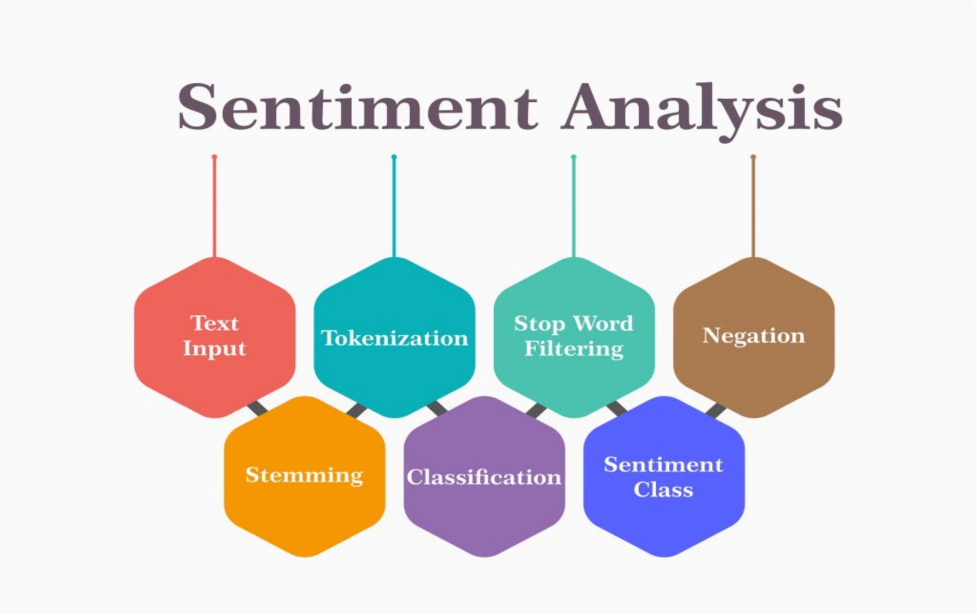
```
Your selected dataframe has 15 columns.
There are 0 columns that have missing values.
```

Out[132]:

	Missing Values	% of Total Values

3 Sentiment Analysis : Review Level

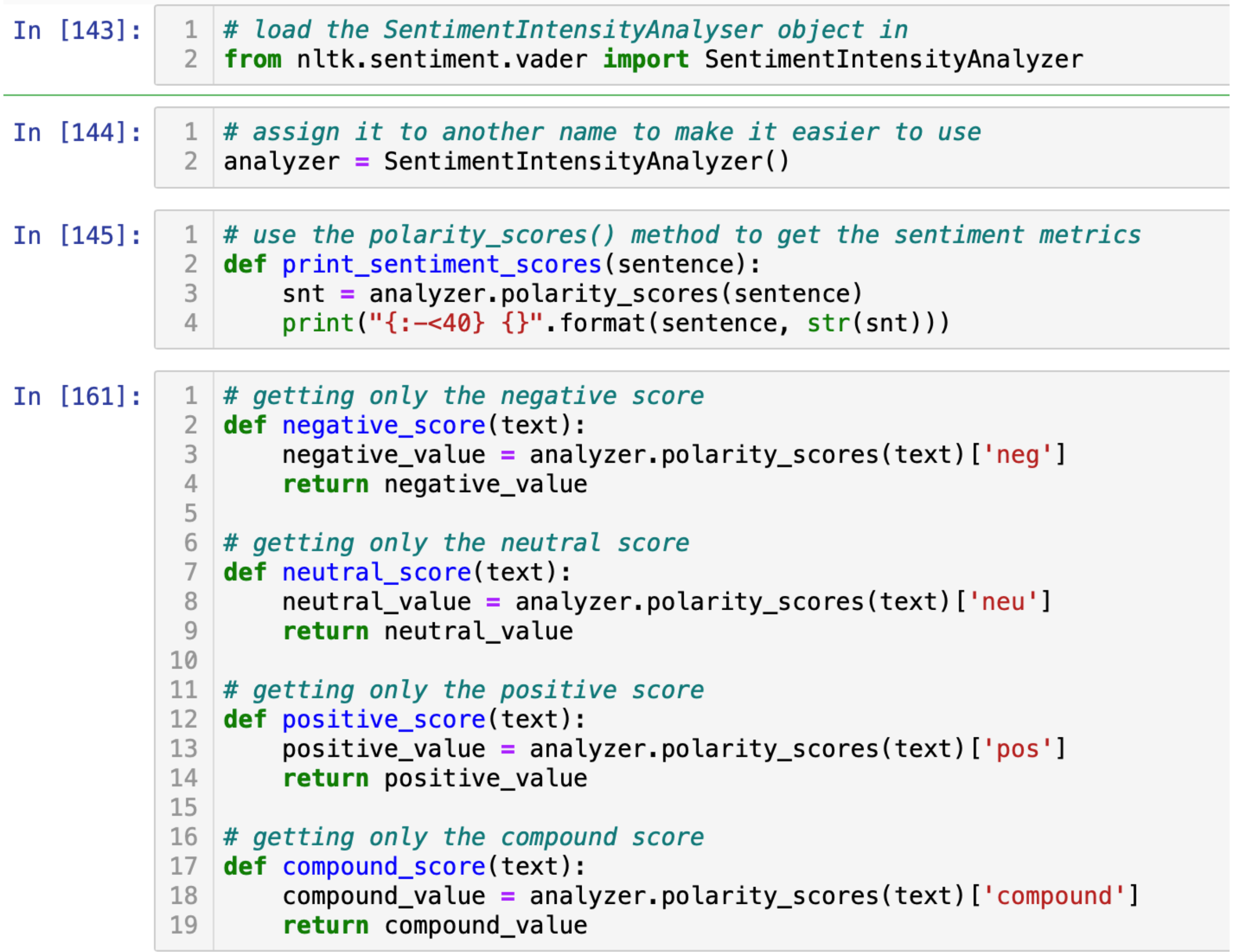
Review Level Sentiment Analysis is a typical procedure used to characterize the feeling of an assortment of text. Programmed classifiers are utilized to order the feeling of text portions dependent on characterized rules. Opinion investigation is a continuous exploration challenge and feeling classifiers are utilized to improve a few errands identified with Natural Langue Processing. Notion classifiers fluctuate with respect to the sort of results they create. A few classifiers produce a mathematical worth that speaks to the opinion while others make a notion class. For notion classes, a few classifiers utilize paired grouping (positive or negative) while others utilize an order framework where five classes are utilized, and each class speaks to a star rating . At the point when a classifier can't unmistakably group the content as one or the other positive or negative, a few models utilize a second-rate class (i.e., "impartial"). A few assessment examination techniques right now exist and are utilized practically speaking and exploration. In this paper, Vader Sentiment is utilized to group the suppositions of surveys and sentences. Vader Sentiment is a standard based model and as indicated by its creators, it has accomplished F1 scores of 0.96 when assessed via web-based media text, 0.63 when assessed on surveys from Amazon.com, 0.61 when assessed on film audits, and 0.55 when assessed on New York Times publications (Hutto and Gilbert 2014a). While the creators of the model gave an assessment, it is as yet imperative to research the model's exhibition when utilized on new datasets.



3.1 Implementation of Sentiment analysis using Sentiment Intensity Analyzer from nltk:

If we can filter out the positive and negative comments, we can compare what makes a listing likeable among for the tourist with respect to the listings that are not liked.

- **For this we will using VADER package**



In [143]:

```
# load the SentimentIntensityAnalyser object in
from nltk.sentiment.vader import SentimentIntensityAnalyzer
```

In [144]:

```
# assign it to another name to make it easier to use
analyzer = SentimentIntensityAnalyzer()
```

In [145]:

```
# use the polarity_scores() method to get the sentiment metrics
def print_sentiment_scores(sentence):
    snt = analyzer.polarity_scores(sentence)
    print("{:-<40} {}".format(sentence, str(snt)))
```

In [161]:

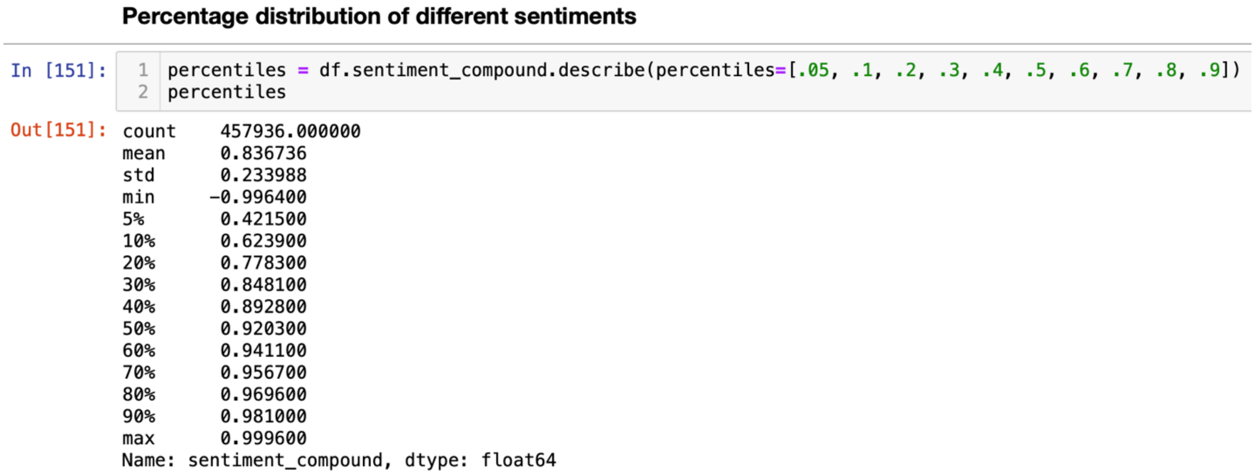
```
# getting only the negative score
def negative_score(text):
    negative_value = analyzer.polarity_scores(text)['neg']
    return negative_value

# getting only the neutral score
def neutral_score(text):
    neutral_value = analyzer.polarity_scores(text)['neu']
    return neutral_value

# getting only the positive score
def positive_score(text):
    positive_value = analyzer.polarity_scores(text)['pos']
    return positive_value

# getting only the compound score
def compound_score(text):
    compound_value = analyzer.polarity_scores(text)['compound']
    return compound_value
```

- **Percentage distribution of different sentiments:**

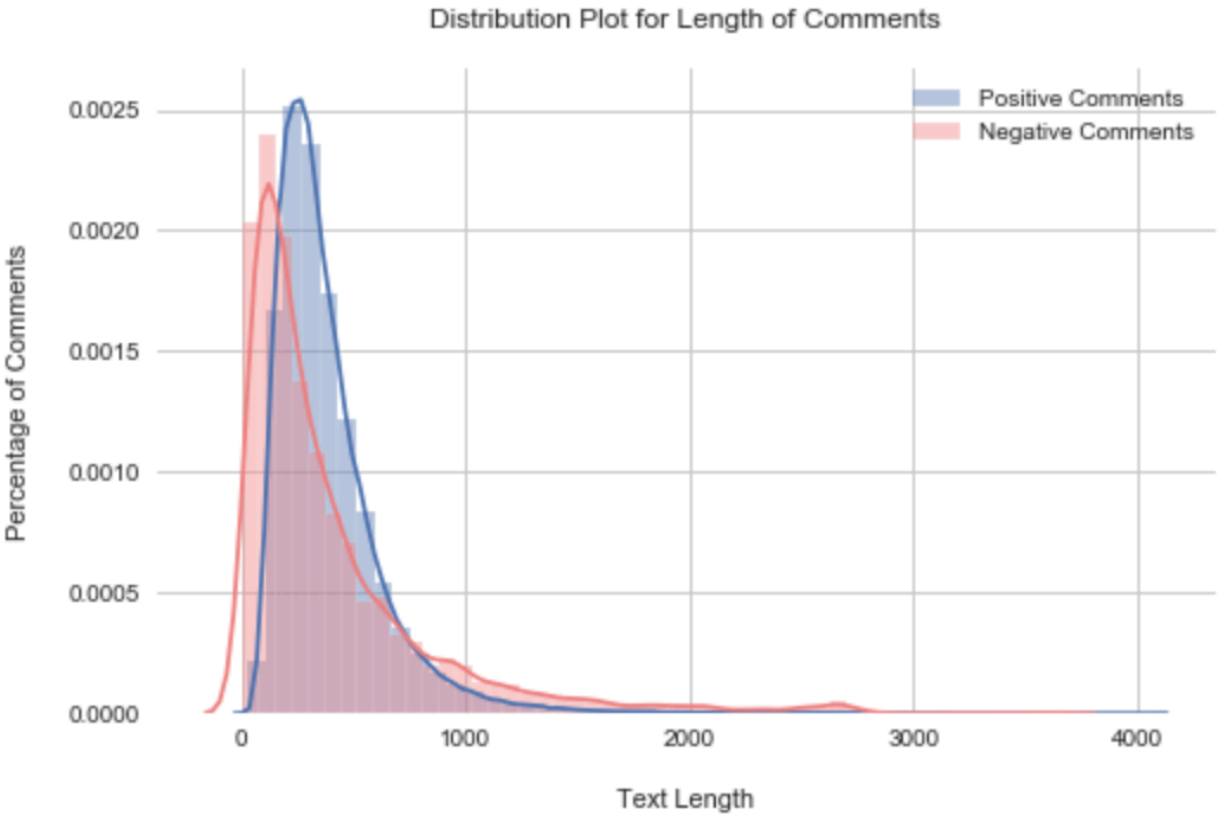


Percentage distribution of different sentiments

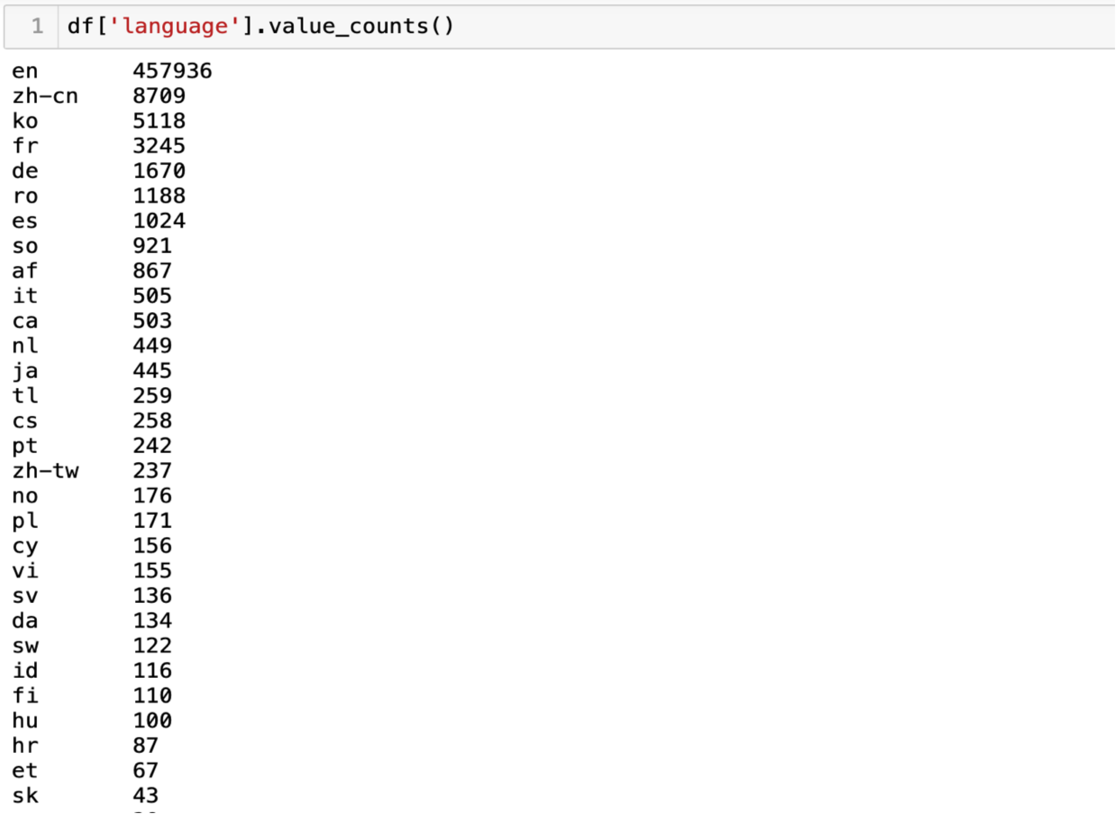
In [151]:
```
percentiles = df.sentiment_compound.describe(percentiles=[.05, .1, .2, .3, .4, .5, .6, .7, .8, .9])
percentiles
```

Out[151]:
```
count    457936.000000
mean          0.836736
std           0.233988
min          -0.996400
5%            0.421500
10%           0.623900
20%           0.778300
30%           0.848100
40%           0.892800
50%           0.920300
60%           0.941100
70%           0.956700
80%           0.969600
90%           0.981000
max           0.999600
Name: sentiment_compound, dtype: float64
```

- **Now let's check how long are the positive comments and negative comments. Is there any difference in their length w.r.t people opinions?**

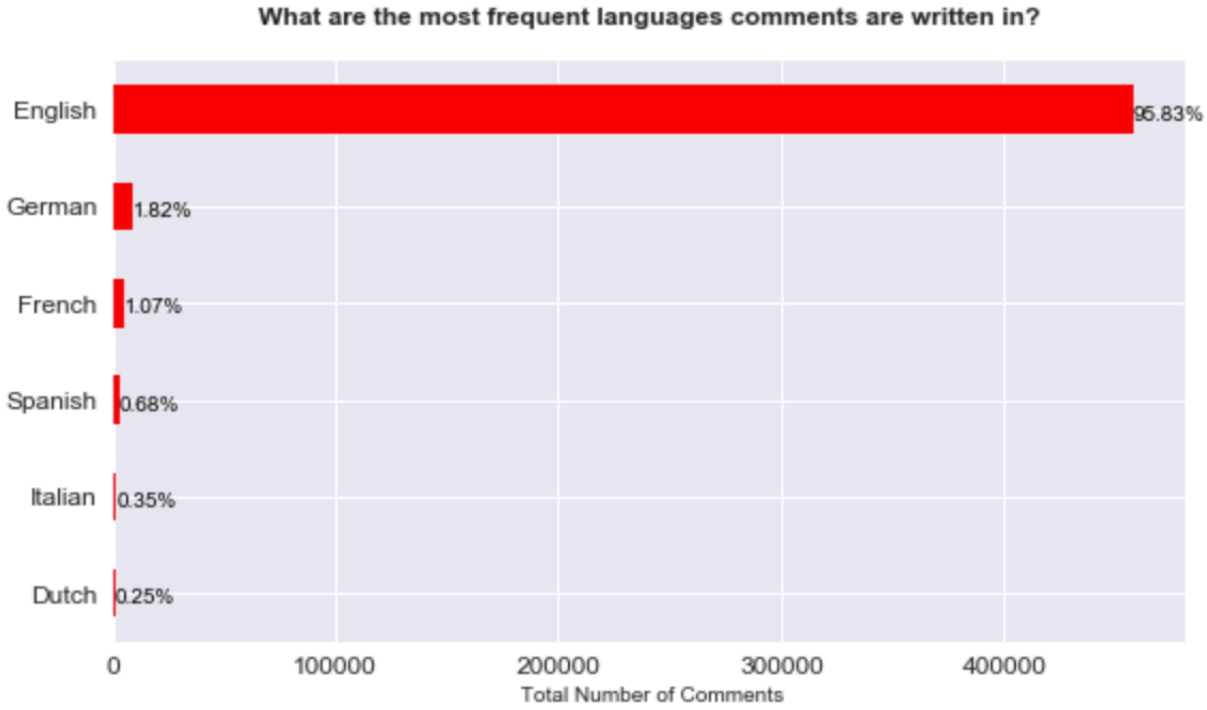


- **Most Frequent languages comments are written in:**



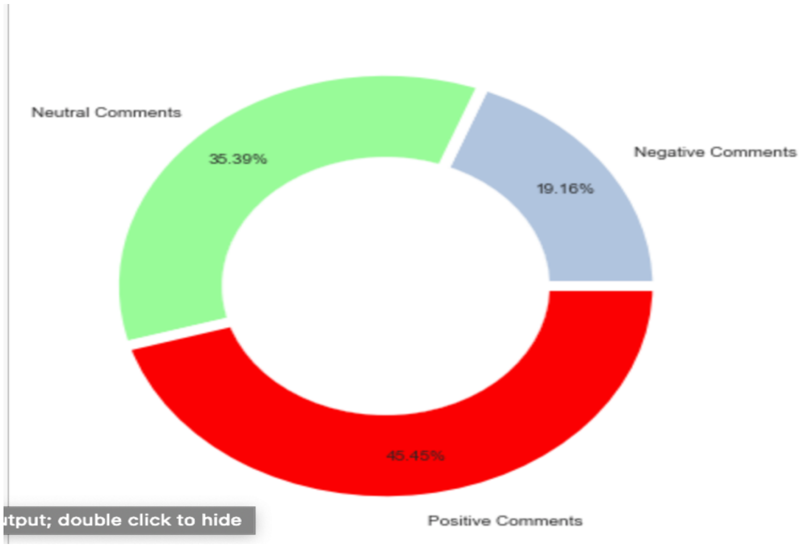
```python
df['language'].value_counts()
```

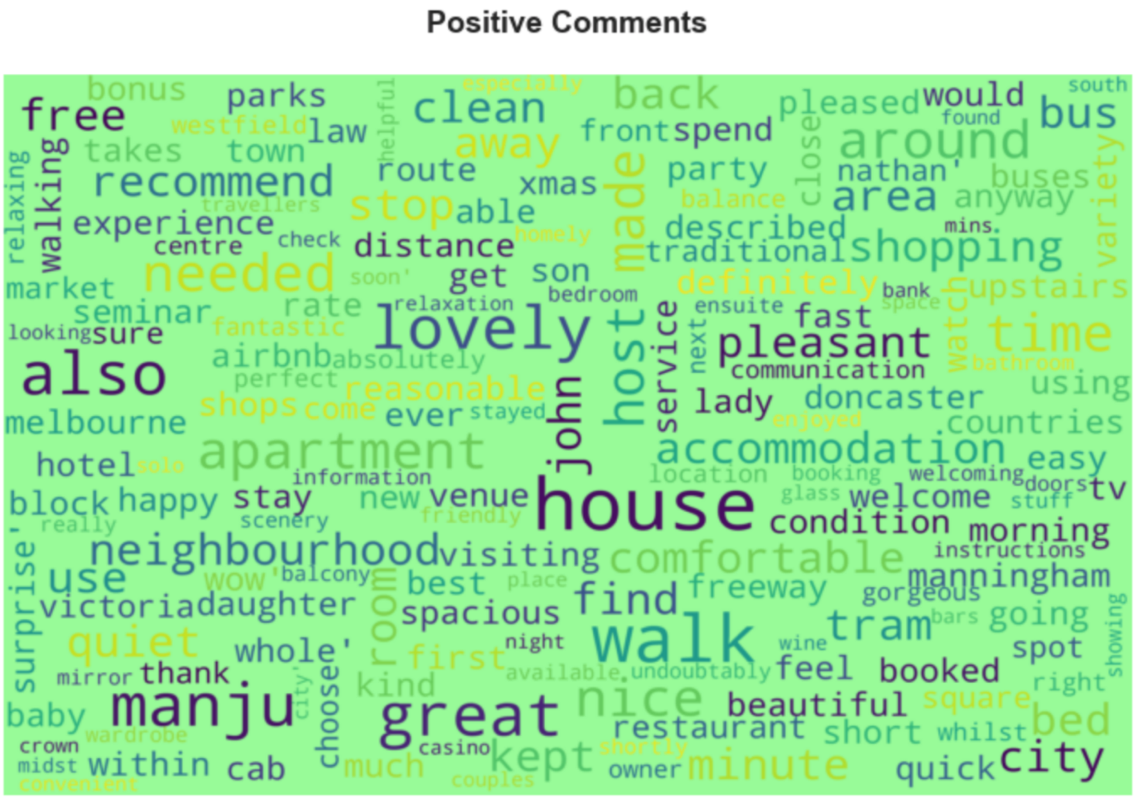
```
en        457936
zh-cn     8709
ko        5118
fr        3245
de        1670
ro        1188
es        1024
so        921
af        867
it        505
ca        503
nl        449
ja        445
tl        259
cs        258
pt        242
zh-tw     237
no        176
pl        171
cy        156
vi        155
sv        136
da        134
sw        122
id        116
fi        110
hu        100
hr        87
et        67
sk        43
```



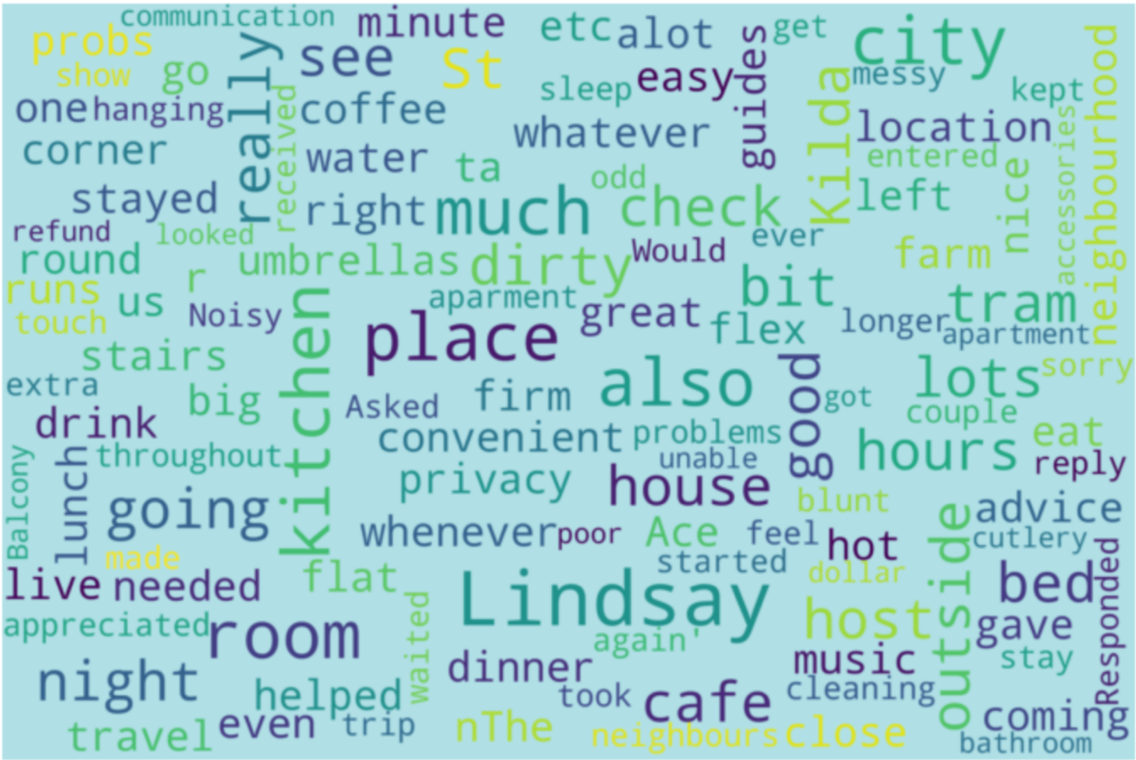
What are the most frequent languages comments are written in?

3.2 Result Analysis:





Negative Comments



3.3 Preprocessing Results and Descriptive Statistics

The implemented preprocessing steps eliminated 28,999 problematic reviews. Most of these reviews were removed because they were not written in English. The percentage of non-English reviews varies by location. Similar to the suggestion by, this observation may be utilized for predicting travel trends for cities or popular touristic destinations for international travelers. The eliminated reviews column in the table includes the sum of non-English, empty, and automatically generated reviews

- Broke down vibes of homestays through AirBnB in Melbourne, Australia. The investigation incorporates results for the suppositions by neighborhood, word cloud to show most successive words in the client surveys and top words that portray the postings.

- Summing up the usage: The information was scratched and controlled in like manner for the examination. The information was then surveyed graphically to figure out what is the overall vibe in the area. Opinion Analysis was finished utilizing faceted vertical structured presentations, word cloud, and flat bar diagrams. Likewise, a pamphlet work was utilized to find the expensive postings of Chicago.

- Synopsis/Insights: Various outcomes and investigation indicated that there is a blended vibe in Chicago neighborhood in with almost equivalent positive and negative feelings. In any case, word cloud, and bar graphs, all together show us sure words that clients much of the time partner their Airbnb remain with.

  Underneath referenced are some additional experiences that we got from the examination:

- Lovely, Nice, Quiet, Comfortable, Beautiful and Perfect - clients partner Airbnb homestays with positive words like these. This can help Airbnb in arrangement how marking and connecting with these words can additionally help them making a buzz.

4 Conclusion and Limitations

In this paper, a huge scope dataset of surveys and sentences (from Airbnb)was examined. The essential target was to utilize an opinion classifier to measure the rates of positive, nonpartisan, and negative audits and sentences. The dataset was accumulated for Melbourne, Australia. Results demonstrate that 45.45% of audits sentences are positive. These finding should help grow new bits of knowledge on the patterns and highlights that rouse and characterize the attributes of the sharing economy. Past these elucidating bits of knowledge, this paper creates and gives a bunch of solid informational indexes on surveys and single sentences for future work to expand on. This information is given upon demand. To build unwavering quality and validity of these datasets, broad assessment of the presentation of the estimation classifier was directed. While the high level of negative surveys on Airbnb is fascinating, extra investigations, conceivably from different controls, are urges to additionally investigate and interpret the drivers of this perception. Note that it can't be presumed that 55% of visitors of Airbnb have negative encounters ,

To start with, we just give introductory understanding into the bigger structure of audit slant. It is worthwhile to additionally analyze supposition of surveys and sentences to comprehend the semantic attributes of positive and negative portions.

Second, the pipeline doesn't consider the fleeting information related with the surveys to investigate patterns and moves. It is conceivable that early clients of Airbnb were very little more amicable and consequently prior surveys had higher rates of negative audits.

Third, the cycle regards the sets as one unit on a city-level. It is conceivable, nonetheless, that inside urban areas, there exist critical contrasts in the rates of negative audits, for example, when separated by different factors, for example, value, area, condo type, and host attributes.

5 Future Work

There exist numerous undeniable augmentations for this work: **First**, text mining strategies might be utilized to investigate the etymological highlights of audits in more prominent detail. By utilizing such techniques, extra inquiries regarding the sharing economy could be investigated and replied. For example, when does positive surveys incorporate negative sentences? **Second**, negative audits can be additionally concentrated to inspect visitors' inspirations driving composing negative surveys. **Third**, a similar pipeline can be utilized to distinguish the assessments of audits from extra urban areas, maybe outside the United States. At last, portioning the information inside urban communities as indicated by the sketched-out elements, for example, could be another productive development of this work.

6 References:

1. https://www.researchgate.net/publication/333114485_LargeScale_Sentiment_Analysis_on_Airbnb_Reviews_from_15_Cities

2. https://towardsdatascience.com/digging-into-airbnb-data-reviews-sentiments-superhosts-and-prices-prediction-part1-6c80ccb26c6a

3. https://medium.com/@mayrazrodriguez/sentiment-analysis-and-topic-modeling-using-airbnb-dataset-c65848b98a57

4. https://rstudio-pubs static.s3.amazonaws.com/333696_8d56164feb0d447f86a1cded4b40b57f.html

5. https://www.kaggle.com/tylerx/melbourne-airbnb-open-data