Business Understanding

Business Objectives

Every Business builds on the customer requirements and demand which directly means that Business is a direct relationship with the customer. To improve business growth, you always need to understand customer feedback. So, I choose some relevant topic for my data mining project named **Sentiment analysis on Europe Hotel Reviews.**

Objectives

There are two main objectives are first is to Improve Business and earn more revenue and second is to find out areas which are highly needed to improve like Rooms, Staff Services, Breakfast, cleaning and many more.

Success Criteria

For success first, we need to understand the data by visualizing it and find out some insights from them to explain a visual in the report.

Data Visualization Goals

- Hotel's Location in a map.
- Top 20 rated hotels in Europe.
- Forecast to predict the number of visitors.
- Trend by Date, Month, Quarter and Year
- Popular Country Visitors visiting Europe.
- Top 50 words appearing in positive reviews.
- Top 50 words appearing in negative reviews.
- Top 20 Tags.
- Visitor Nationality.

- ## Data Understanding

Initial Data Report

This Dataset is belonging to the most popular site booking.com which do the booking and get the online feedback from the customer, so they can improve in their suggestions and recommendation to the new customers. So, they publish their dataset on Kaggle to get more insights into their dataset. This dataset is available on Kaggle by name of 515k Europe Hotel Reviews. This dataset reviews collected between two years.

Describe Data Report

Dataset is in good form. Dataset is an unstructured data which means its without a label. It consists of 17 features out of which are very useful for findings like Rating, Positive and Negative Review column. Dataset more of text type data which means we must do text mining with that.

Explore Data

The dataset consists of 515k rows with 17 columns. Dataset has some missing values and duplicate values. Dataset is already pre-processed with removing punctuation, Unicode and white spaces. Dataset have not one column of Review which mostly present in the other datasets while Versa in one row it has two columns one is Positive Review and second is Negative Review which makes them different and unique from another dataset's that's why me and my team partner choose this Dataset to get a good knowledge of text mining and analysis.

Data Challenges

- Unlabelled data.
- Manually created Target variable label name.
- Use of text mining for TF-IDF Document (Term Frequency Inverse frequency Document).
- Used Stop words, Stemming functions.

- # Data Preparation

➢ Dataset Description

This dataset contains 515,000 customer reviews and scoring of 1493 luxury hotels across Europe.

The CSV file contains 17 fields. The description of each field is as below:

- ❖ Hotel_Address: Address of the hotel.
- ❖ Review_Date: Date when the reviewer posted the corresponding review.
- ❖ Average_Score: Average Score of the hotel, calculated based on the latest comment in the last year.
- ❖ Hotel_Name: Name of Hotel
- ❖ Reviewer_Nationality: Nationality of Reviewer
- ❖ Negative_Review: Negative Review the reviewer gave to the hotel. If the reviewer does not give the negative review, then it should be: 'No Negative'
- ❖ Review_Total_Negative_Word_Counts: Total number of words in the negative review.
- ❖ Positive_Review: Positive Review the reviewer gave to the hotel. If the reviewer does not give the negative review, then it should be: 'No Positive'
- ❖ Review_Total_Positive_Word_Counts: Total number of words in the positive review.
- ❖ Reviewer_Score: Score the reviewer has given to the hotel, based on his/her experience
- ❖ Total_Number_of_Reviews_Reviewer_Has_Given: Number of Reviews the reviewers have given in the past.
- ❖ Total_Number_of_Reviews: Total number of valid reviews the hotel has.
- ❖ Tags: Tags reviewer gave the hotel.
- ❖ days_since_review Duration between the review date and scrape date.
- ❖ Additional_Number_of_Scoring: There are also some guests who just made scoring on the service rather than a review. This number indicates how many valid scores without review in there.
- ❖ lat: Latitude of the hotel
- ❖ lng: longitude of the hotel

➢ Select Data

we select all variables leaving these variables count of Positive, Negative and reviews. But our more focus on Positive and Negative review more than others. Like other variables Tags and Review Date is crucial as well to find the trend of customer's in a time of year.

- **Clean Data**

  In the cleaning process, we first remove all those Neutral Reviews like No Positive, No Negative, Nothing, Nothing at all and Null from Reviews. Then Remove the Duplicate row's Data.

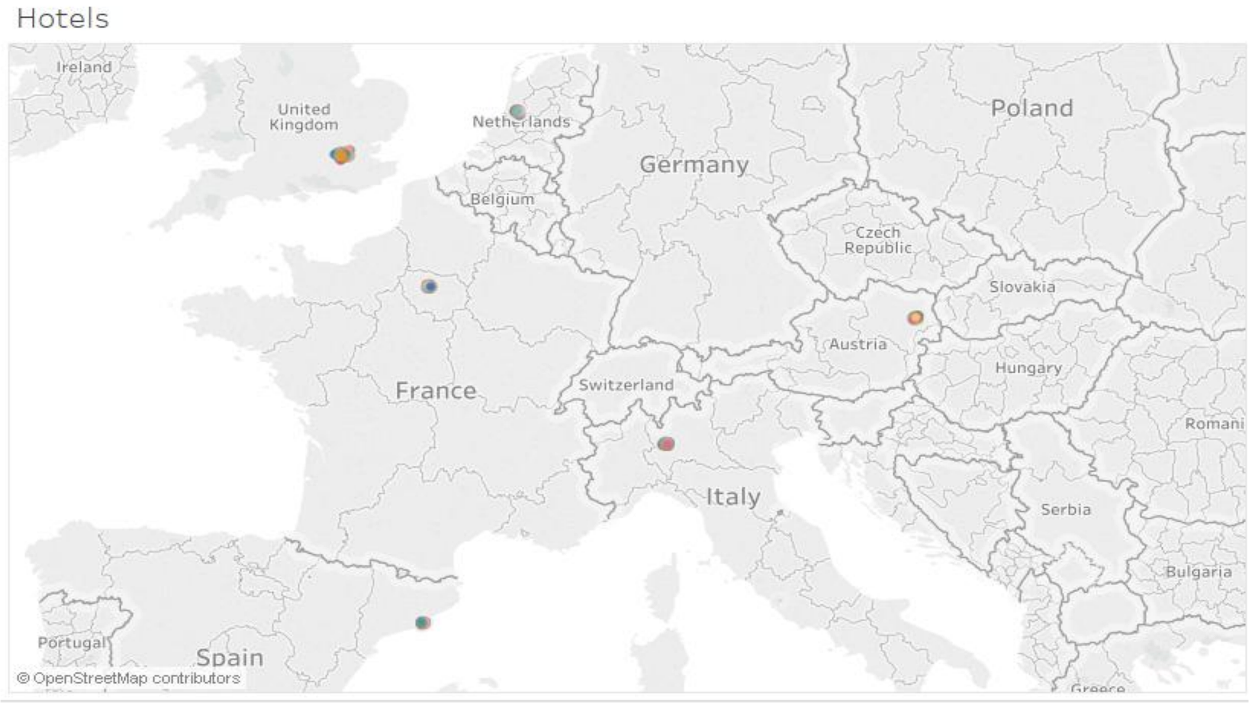
- **Tools and Languages**

  We used one tool and two languages to visualize our data. The Tool we used Tableau for the specific reason is that it handles the big data easily and its portability and user interface is so good. The languages we choose Python for Word cloud and Seaborn and R for Word cloud-only.

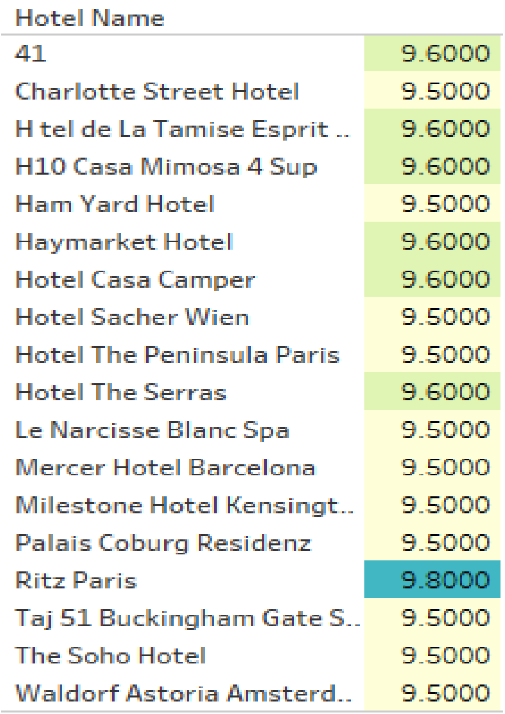
Visualization

For visualization, we try serval plots to find good insight because we are doing sentiment analysis on hotel reviews. So, our most Target on Text Mining. We visualize our data into two forms one is in the form of Word cloud and other in the form of Graphs which visual the map, box-whisker and man more also showing the comparison.

The first graph is a map which displays the locations of a Hotels. From this simple chart, we can conclude that this data belongs to some specific countries of Europe's Hotels Data. If you can look over the graph, then you can find out that you can simply conclude that Hotels are in the famous visiting countries of Europe.

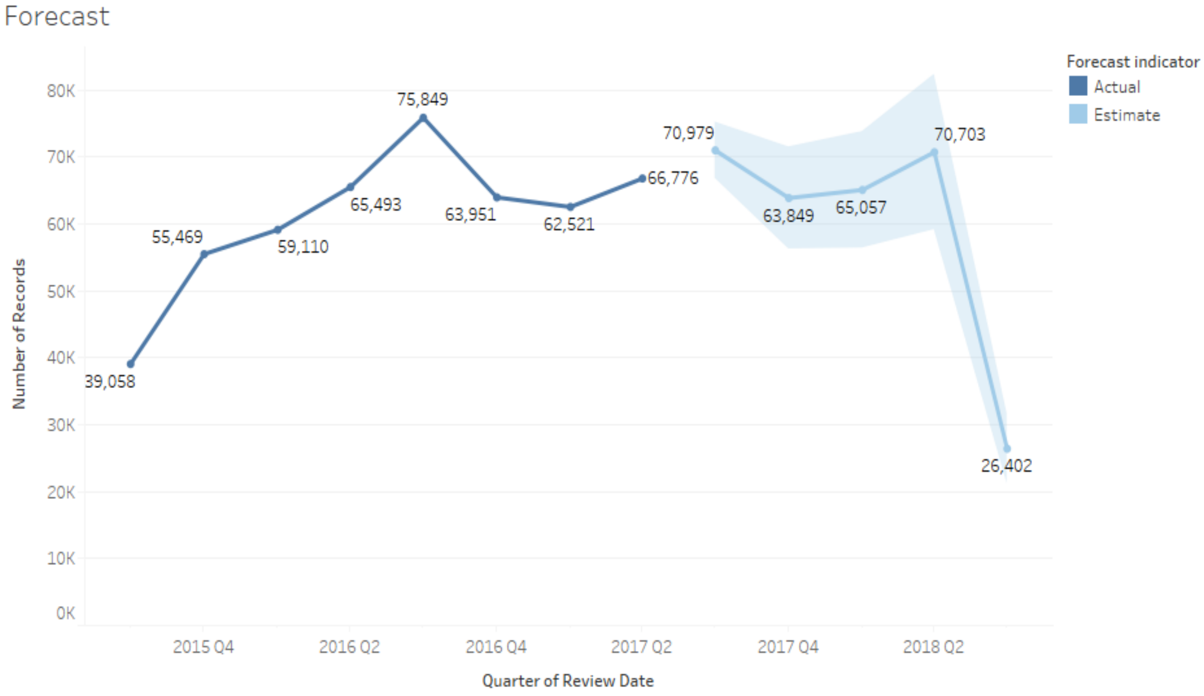


Second, I visual my data using a tabular form which shows the rating of Hotels between 9.5 and 9.8 and It also describe the top 20 rated hotels according to the Data The highest rated Hotel Name is **Ritz Paris** which is highlighted in blue colour to easily identify. The tabular form is the most simple approach to highlight the facts and figures.

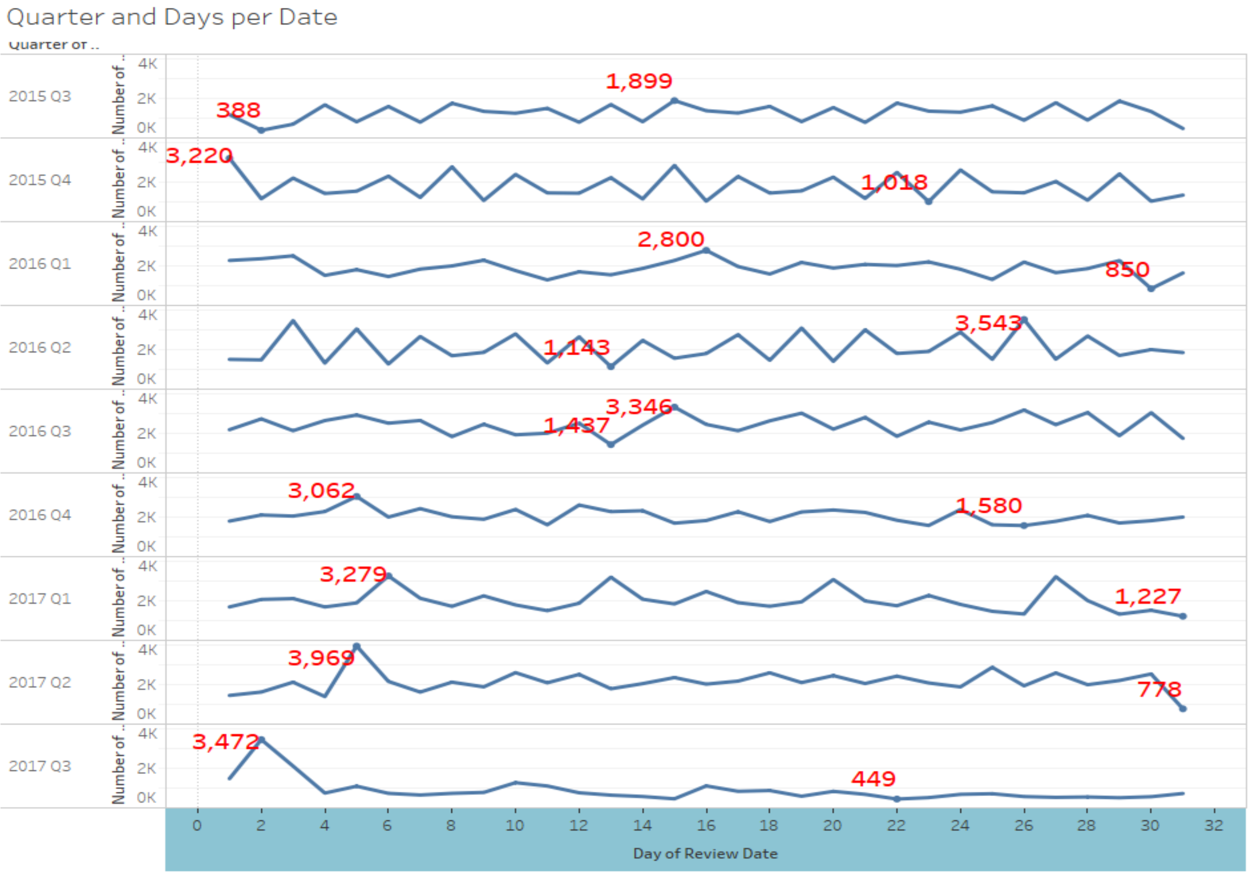


Hotel Name	
41	9.6000
Charlotte Street Hotel	9.5000
H tel de La Tamise Esprit ..	9.6000
H10 Casa Mimosa 4 Sup	9.6000
Ham Yard Hotel	9.5000
Haymarket Hotel	9.6000
Hotel Casa Camper	9.6000
Hotel Sacher Wien	9.5000
Hotel The Peninsula Paris	9.5000
Hotel The Serras	9.6000
Le Narcisse Blanc Spa	9.5000
Mercer Hotel Barcelona	9.5000
Milestone Hotel Kensingt..	9.5000
Palais Coburg Residenz	9.5000
Ritz Paris	9.8000
Taj 51 Buckingham Gate S..	9.5000
The Soho Hotel	9.5000
Waldorf Astoria Amsterd..	9.5000

The third chart is a time series graph which explains the trend. Like for any Business a businessman should know at which quarter time of a year you have more customers are coming or visiting the city, so you can early prepare like hiring more staff to provide good services to the visitors, like refurnish, Colour's and foodstuff, So, the visitors can enjoy their trip and gives you good feedback and recommend their friend's. Below chart shows that we have a high number of visitor's during 2016 Q3 (July – Sept). it also forecasts for the next two quarters based on previous data.

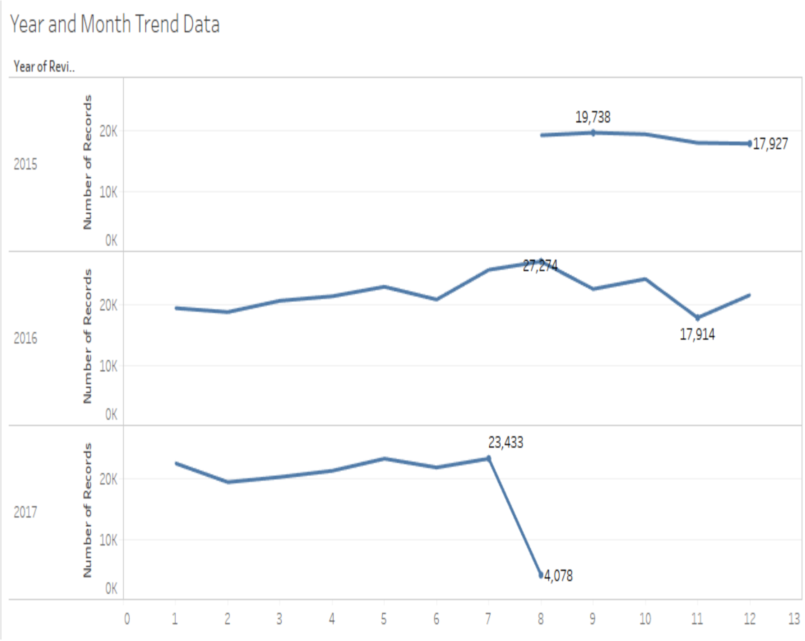


The trend of sum of Number of Records (actual & forecast) for Review Date Quarter. Color shows details about Forecast indicator. The marks are labeled by sum of Number of Records (actual & forecast) .



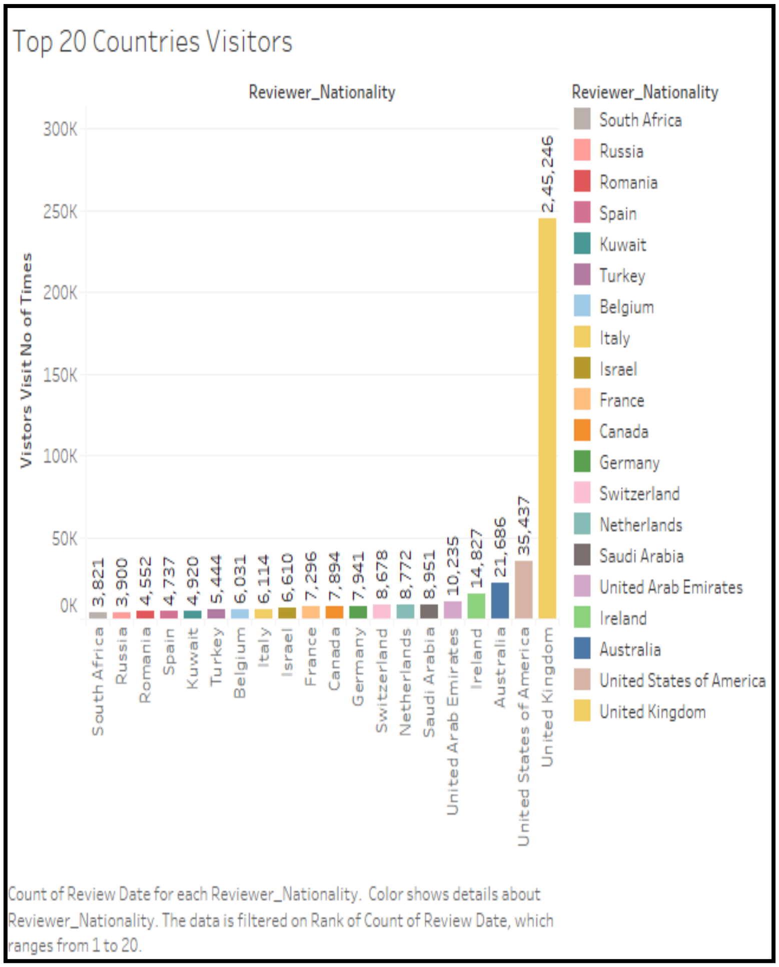
The trend of sum of Number of Records for Review Date Day broken down by Review Date Quarter. The marks are labeled by sum of Number of Records.

The above time series chart explains the relation between the specific quarter and date. In each quarter of the time, it shows two results one is a maximum no of visitors which shows that on date of all months in a quarter we have most bookings and while Versa we have a less booking.



Another time-series graph which shows the The relation between month and year about the trend of bookings. As I have two years of data, so you notice that Aug 2015 to August 2017 that's why after August 2017 it is falling. So, I have just a one-year seasonal trend. So, finding from this graph is not such good.

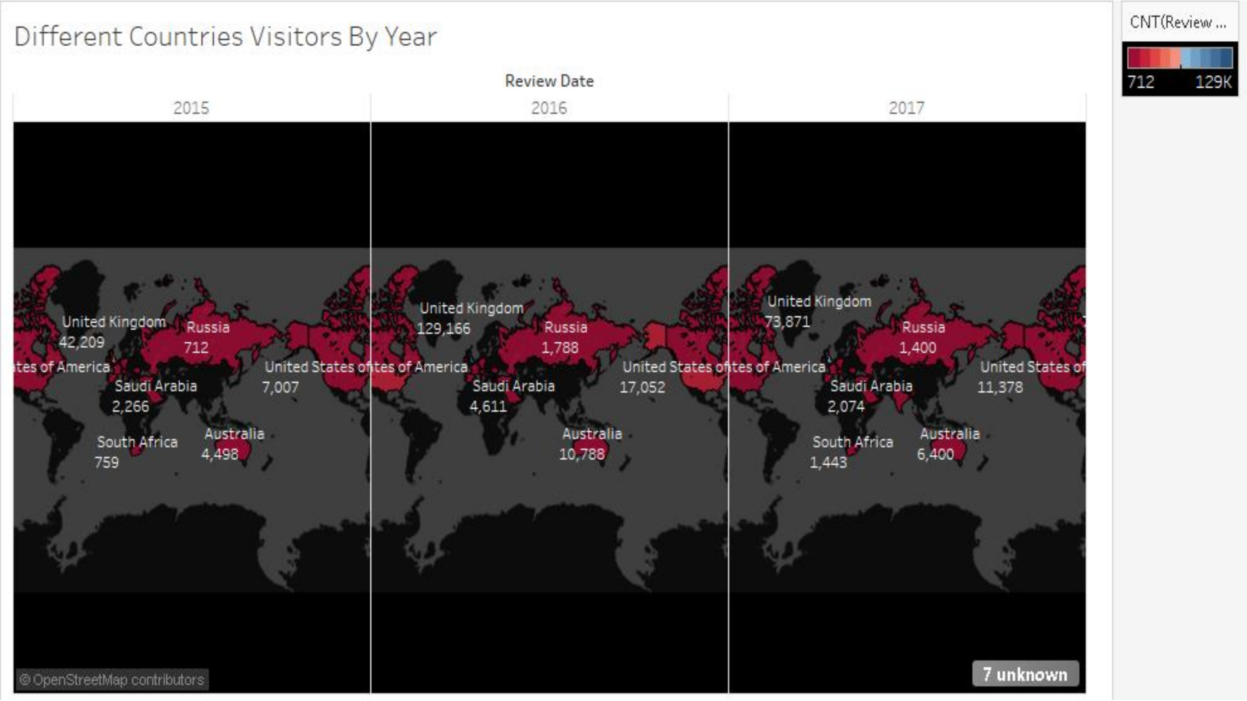
The left graph is known as a bar chart which helps me to visualize categorical variable reviewer Nationality, so I can get the top 20 Countries whose visitors travelled to Europe between 2015 -2107 more. In the left graph, you can identify that the United Kingdom people's more like to visit the other countries of Europe.

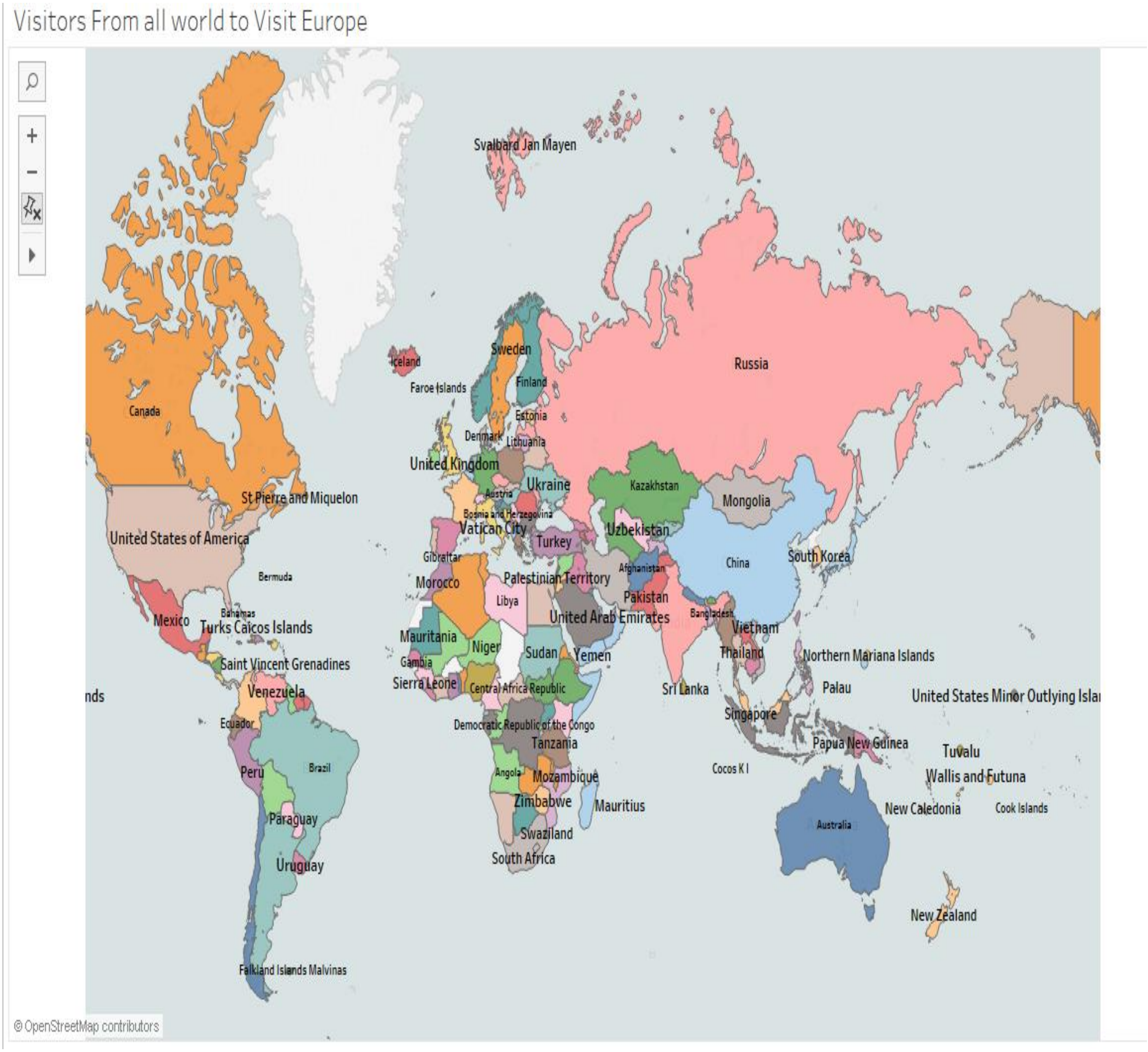


Count of Review Date for each Reviewer_Nationality. Color shows details about Reviewer_Nationality. The data is filtered on Rank of Count of Review Date, which ranges from 1 to 20.

The Below chart is again a map which displays left graph by dividing them into a year by which can help us to identify that nowadays more people are coming to travel and each year the numbers are increases. Two reasons maintenance of visiting points, good environment and Lovely people's around all Europe.

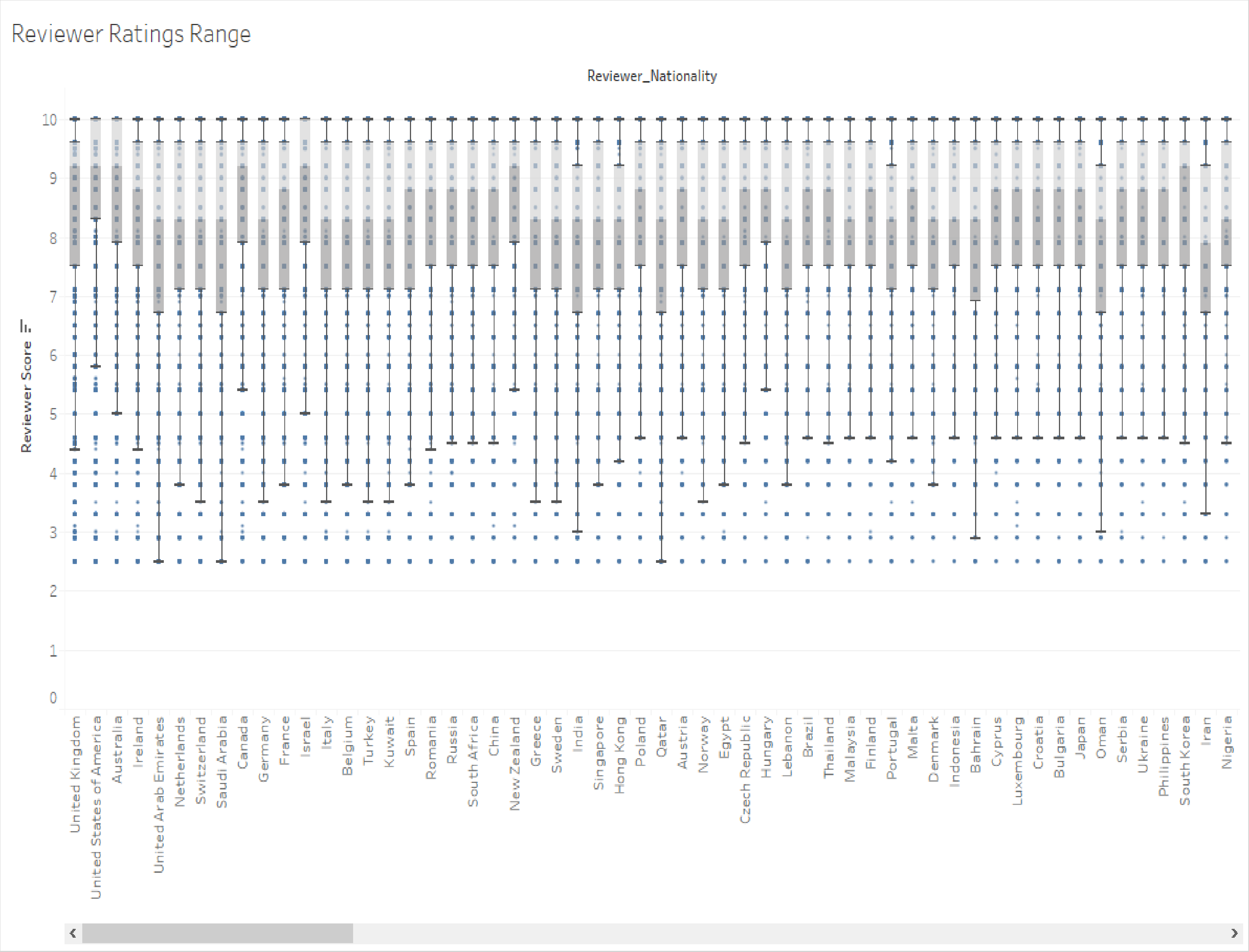
Extra work





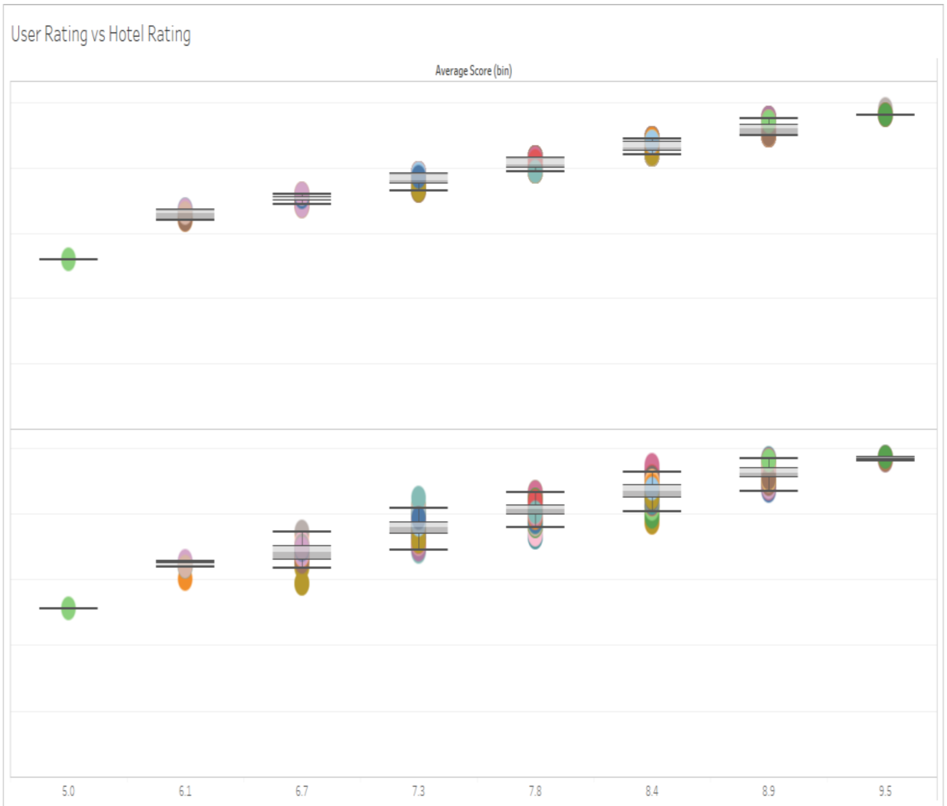
Visitors From all world to Visit Europe

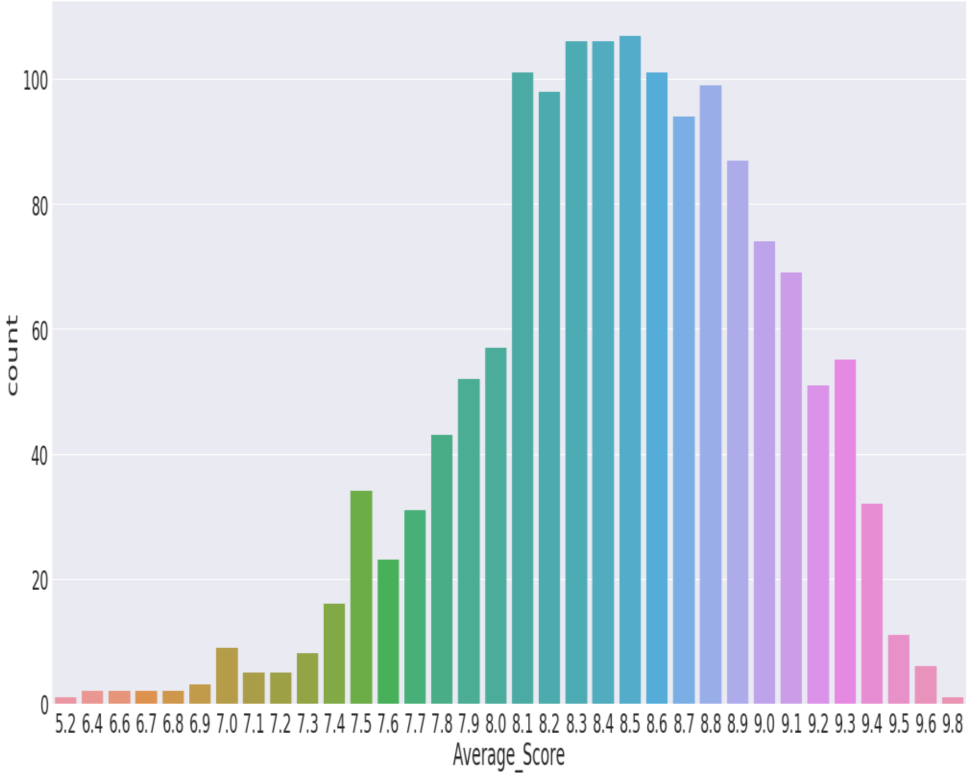
A map shows that people from all countries around the world are more likely to visit the Europe Countries which also include the Europe countries as well.



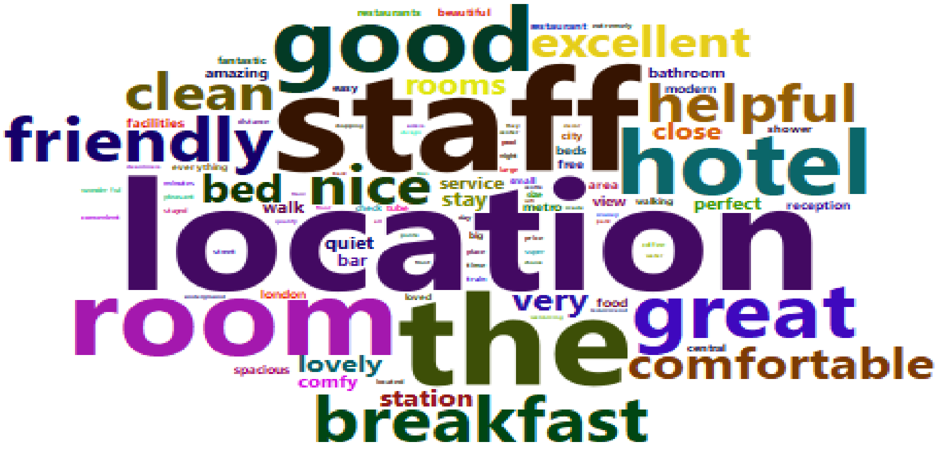
The above graph is called a box-whisker plot. This plot helps me to find the great insight about reviewer’s countries people that in each country there are some people’s who gave good ratings and while Versa some give bad rating. We can see the box and the outlier’s in the graph. We can’t reject these outlier’s it’s the best solution for us to improve our performance what’s the reason they give a low rating. This plot helps me a lot.

This right plot is a combination of scatter plot +box plot. In this plot, I will compare the Hotel Actual rating vs the median of Visitor’s Rating. Above is a Hotel rating and below is a user Ratings. You easily notice the difference between both that user rating varies more compare to the average one.

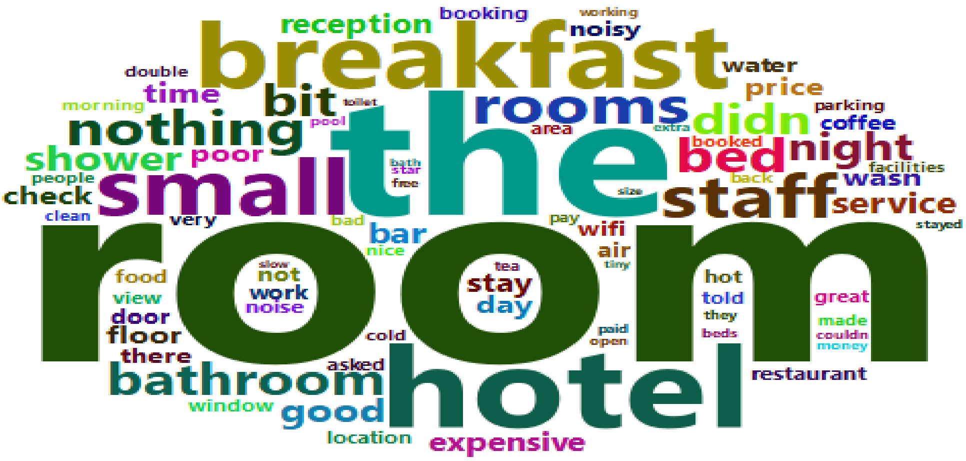




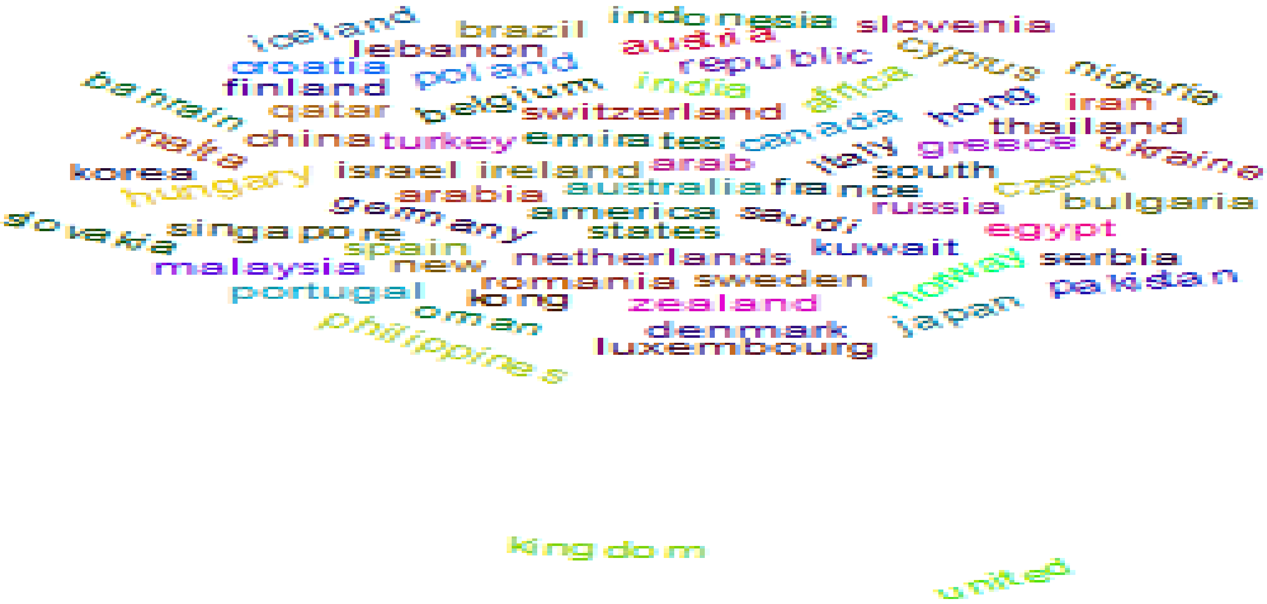
The left one is a simple Histogram plot of all Hotel Ratings which display ‘s that most hotels have a good rating. The height of the rating between 8.1 -9.0 range of rating. This range explains to us that the hotel owners know their business well and likely to maintain their services good so their rating increases.



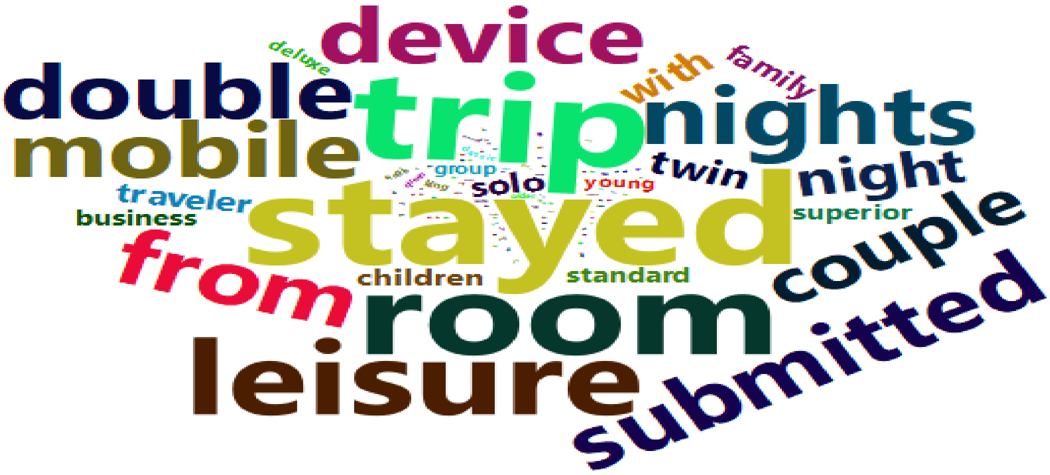
This is a type of graph where you visual the word which has the number of occurrences in your Data and it is known as a word cloud. Its mostly used in text mining for sentiment analysis. As my data is about reviews. So, this word cloud represents the top 100 Positive words which occur more in the Positive reviews.



This is another word cloud of Negative words from Negative Reviews. In this word cloud if you see the room has a high number of occurrences. But in Both Positive and negative word clouds, you see the same words reason is simple that Services, locations, rooms and hotels somewhere it's good and somewhere it is bad.



Another word cloud it is about reviewer nationality which displays the top 100 visiting countries where the United Kingdom is lead and have the high number of weights to visit Europe so that's why it's far from all other words.



This word cloud is about Tags which tells that Visitor who books their Bookings by mobile or calls or books for how many people's or a solo trip or which type of room and many more. This is unique from all charts because we didn't do such things with this variable, but this word cloud defines it well.

- **Evaluation and Deployment**

We can easily evaluate our data visualization goals from all the charts we used in this document. Visualize is one big Aspect so rather than just on facts in a table you can see on the graph or plot it explains many things it looks easier to look all on in one graph.

For Deploy purpose, I like to suggest that each Hotel should have such a system which displays all these charts, so they can improve their weak areas and services and grow their business and generate more results.

Appendix

R Script of Word Clouds

The following is the code for retrieving negative wordcloud:

```
reviews <- hr[sample(nrow(hr), 40000), ]
reviews <- reviews[reviews$Positive_Review!='No Positive',]
reviews <- reviews[reviews$Negative_Review!='No Negative',]
term_freq <- function(hr,sent){
  if(sent=='pos'){
    corpus <- Corpus(VectorSource(hr$Positive_Review))
  }else{
    corpus <- Corpus(VectorSource(hr$Negative_Review))
  }
  corpus <- tm_map(corpus, removeWords, stopwords("SMART"))
  corpus <- tm_map(corpus, removeWords, stopwords("en"))
  corpus <- tm_map(corpus, stripWhitespace)
  dtm <-TermDocumentMatrix(corpus)
  mat_dtm <- as.matrix(dtm)
  v_dtm <- sort(rowSums(mat_dtm),decreasing = TRUE)
  FreqMat <- data.frame(word = names(v_dtm), Freq = v_dtm)
  FreqMat <- FreqMat[1:100,]
  return(FreqMat)
}

wordcloud2(data = term_freq(reviews,'pos'),minRotation = 0,maxRotation=0)

wordcloud2(data = term_freq(reviews,'neg'),minRotation = 0,maxRotation = 0)
```

The following is the code for retrieving reviewer’s nationality and Tags word cloud:

```
reviews <- hr[sample(nrow(hr), 40000), ]
reviews <- reviews[reviews$Reviewer_Nationality!=' ',]
reviews1 <- hr[sample(nrow(hr), 40000), ]
reviews <- reviews[reviews$Tags!=' ',]


term_freq <- function(hr){
  corpus <- Corpus(VectorSource(hr$Reviewer_Nationality))
  dtm <-TermDocumentMatrix(corpus)
  mat_dtm <- as.matrix(dtm)
  v_dtm <- sort(rowSums(mat_dtm),decreasing = TRUE)
  FreqMat <- data.frame(word = names(v_dtm), Freq = v_dtm)
  return(FreqMat)
}

data = term_freq(reviews)
data
wordcloud2(data,size =15,fontWeight = 250)

wordcloud2(data=term_freq(reviews1), size =15,fontWeight = 250)
```