Capstone Project-02

Sentiment Analysis of Airline reviews

What is Sentiment Analysis ?

- Sentiment analysis identifies the emotional tone behind a body of text.
- It's used in social media monitoring, allowing businesses to gain insights about how customers feel about certain topics, and detect urgent issues in real time before they spiral out of control.
- Example:-Sentiment analysis Tweets about a famous entity or event , reviews of hotels, airlines, products etc.

Problem Statement

- "Determining the overall sentiment of British airline reviews regarding flight experiences in order to identify areas of improvement for airlines to enhance customer satisfaction."
- The sentiment can be classified as positive, negative or neutral.

Objective

- The objective of this sentiment analysis would be to gain insight into the perceptions and experiences of customers who have flown with a the airline, specifically in regards to their flight experience.
- By identifying areas of improvement through analyzing customer reviews, the airline can take steps to address any issues and improve customer satisfaction.

Table of Content

1.Overview of the Data

- Data was collected from "Skytrax" website using beautiful soap library.
- Skytrax is a Airline and Airport customer review site for airlines and airports across the world.
- This Website was used for web scraping flight reviews of "British Airways" only.
- Total reviews collected were “3460”.
- Shape of our dataset is (3460 , 3)
- Titles of the columns are “reviews” , “stars” , “country”.

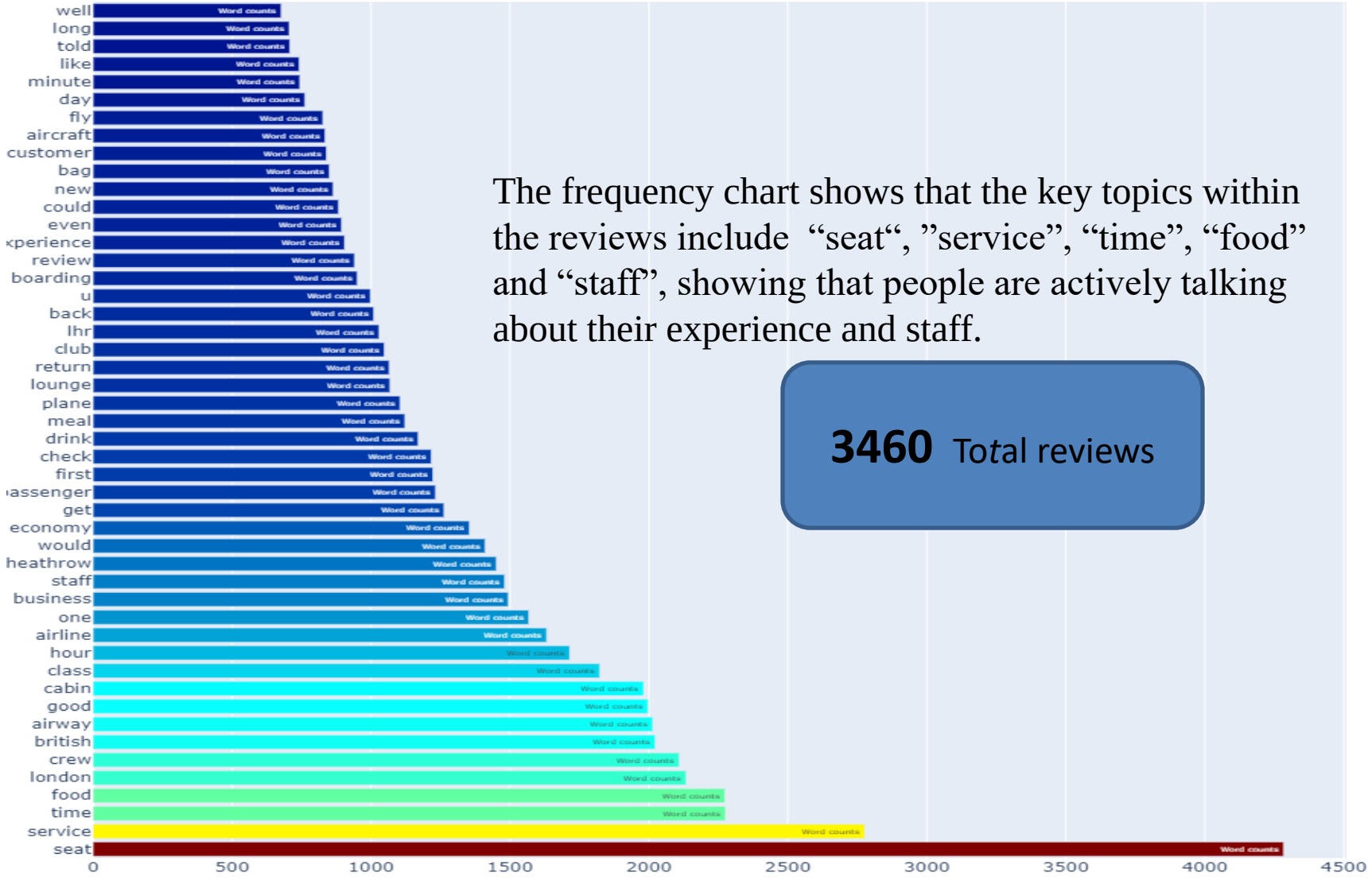
2.Data Pre-processing

- Cleaning the Text
  - Removing special characters , numbers , symbols using regular expressions
  - Converting all the text into lower case using lower() function
  - Splitting the sentences into words using word tokenization
  - Removed the stop words using stop words from nltk library
  - Converting the words to their base forms using lemmatization.

- Checking for null values
  - Null values were dropped

3.EDA

Top 50 Word frequencies in the dataset

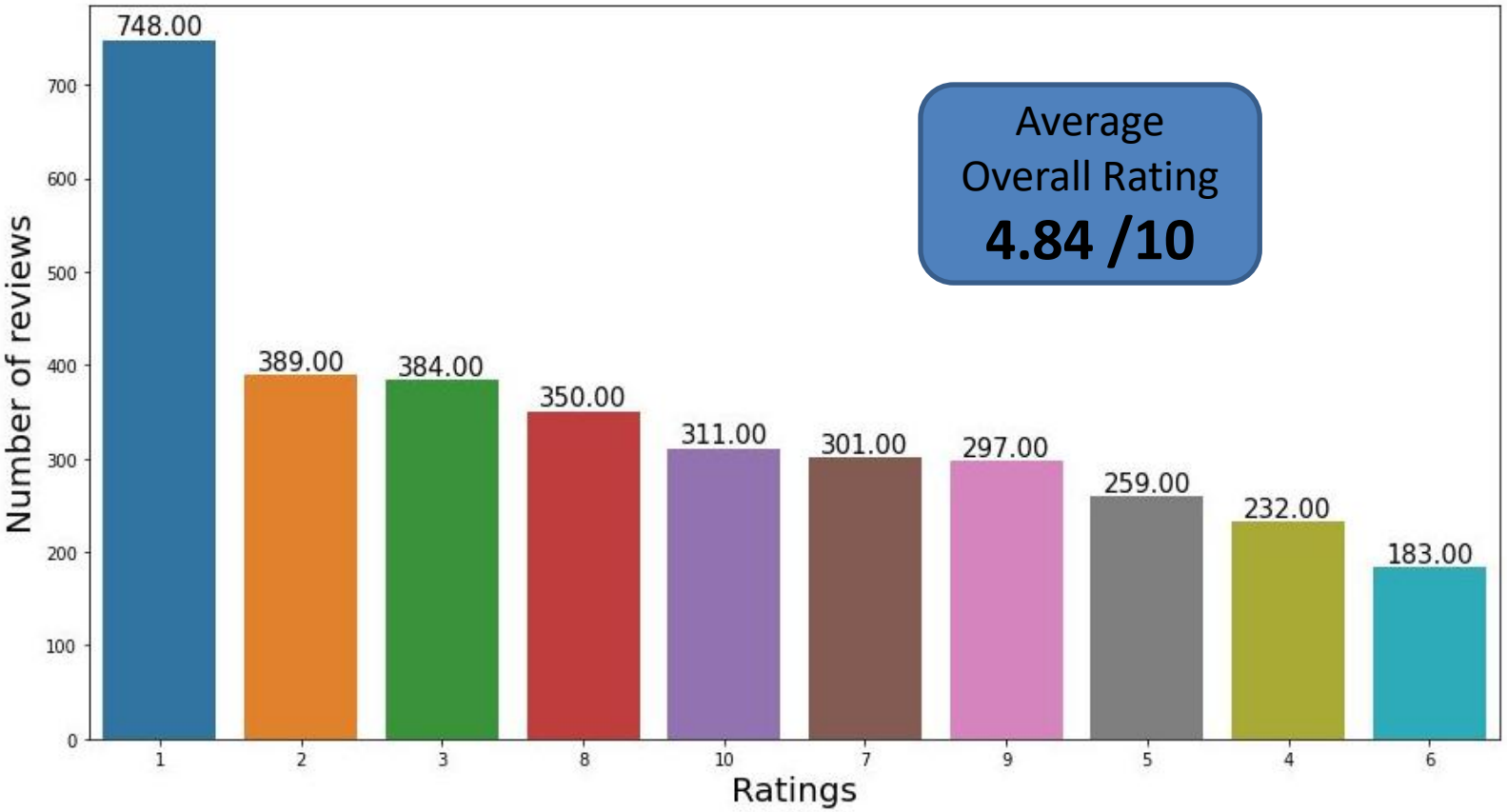


The frequency chart shows that the key topics within the reviews include "seat", "service", "time", "food" and "staff", showing that people are actively talking about their experience and staff.

3460 Total reviews

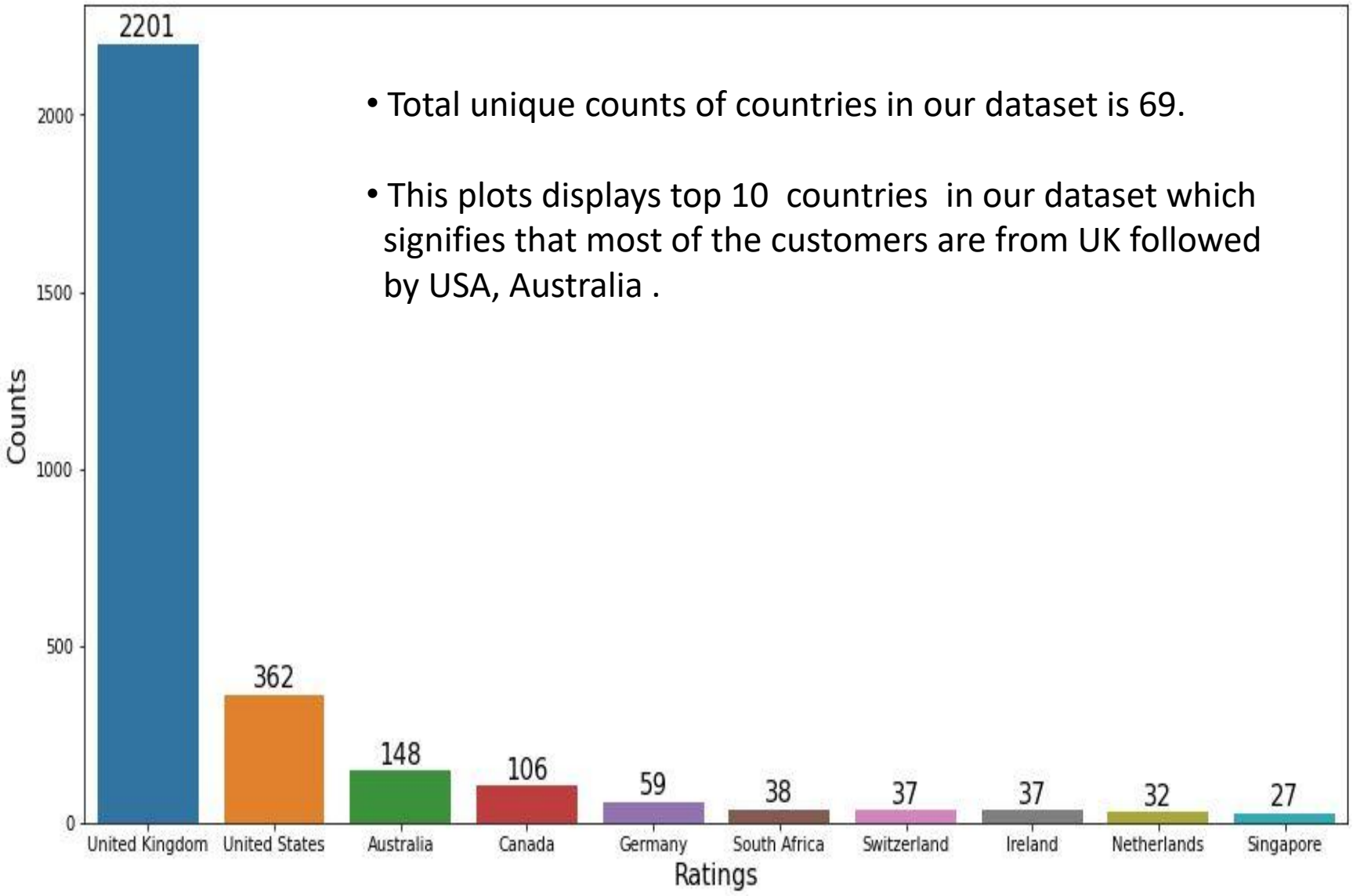
WordCloud of our dataset

Distribution of Ratings



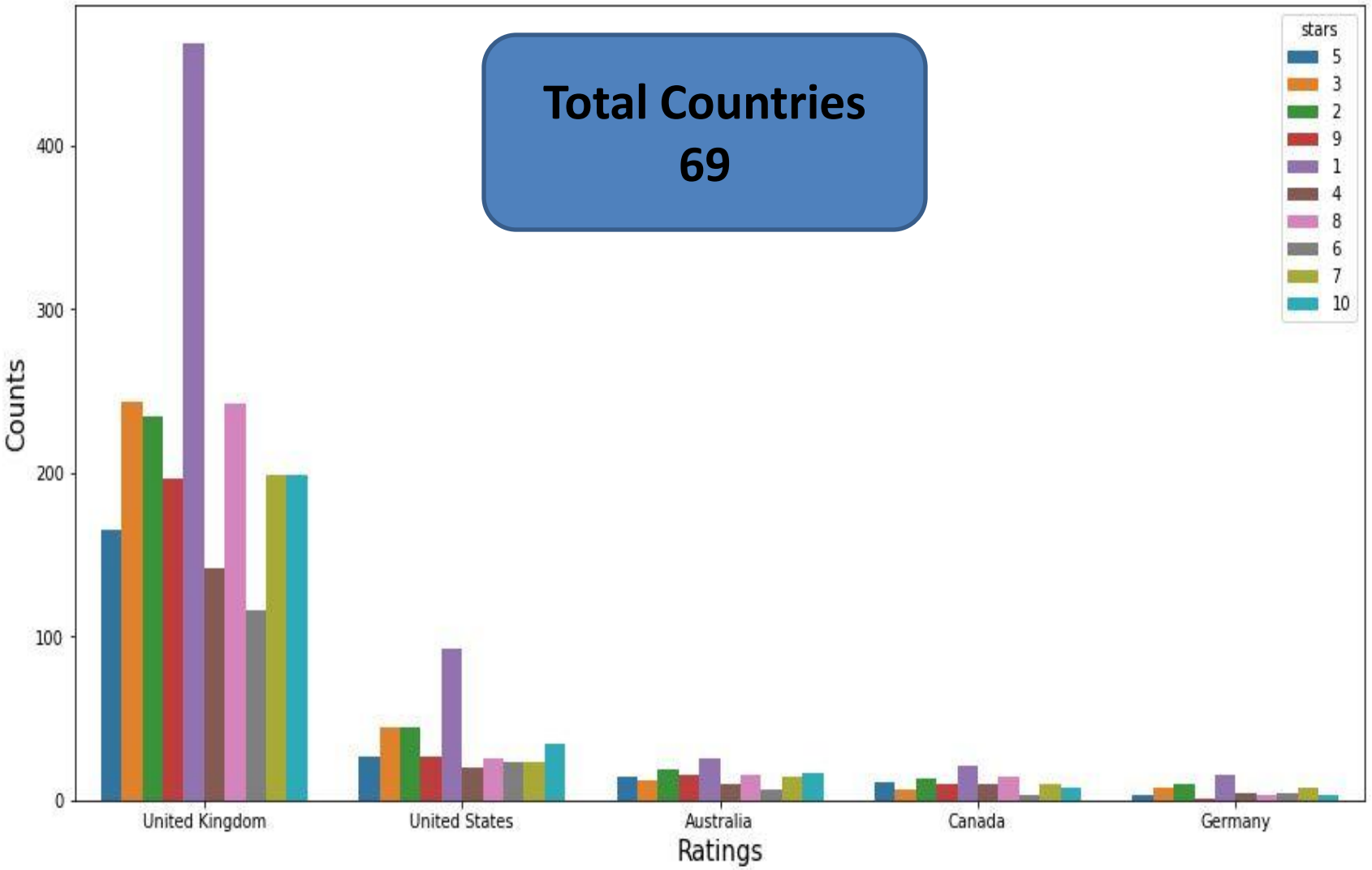
- Total no. of reviews are = 3454
- 58 % of the customers have given 5 or less than 5 rating
- Only 9% have given 10 stars

Top 10 countries in our data

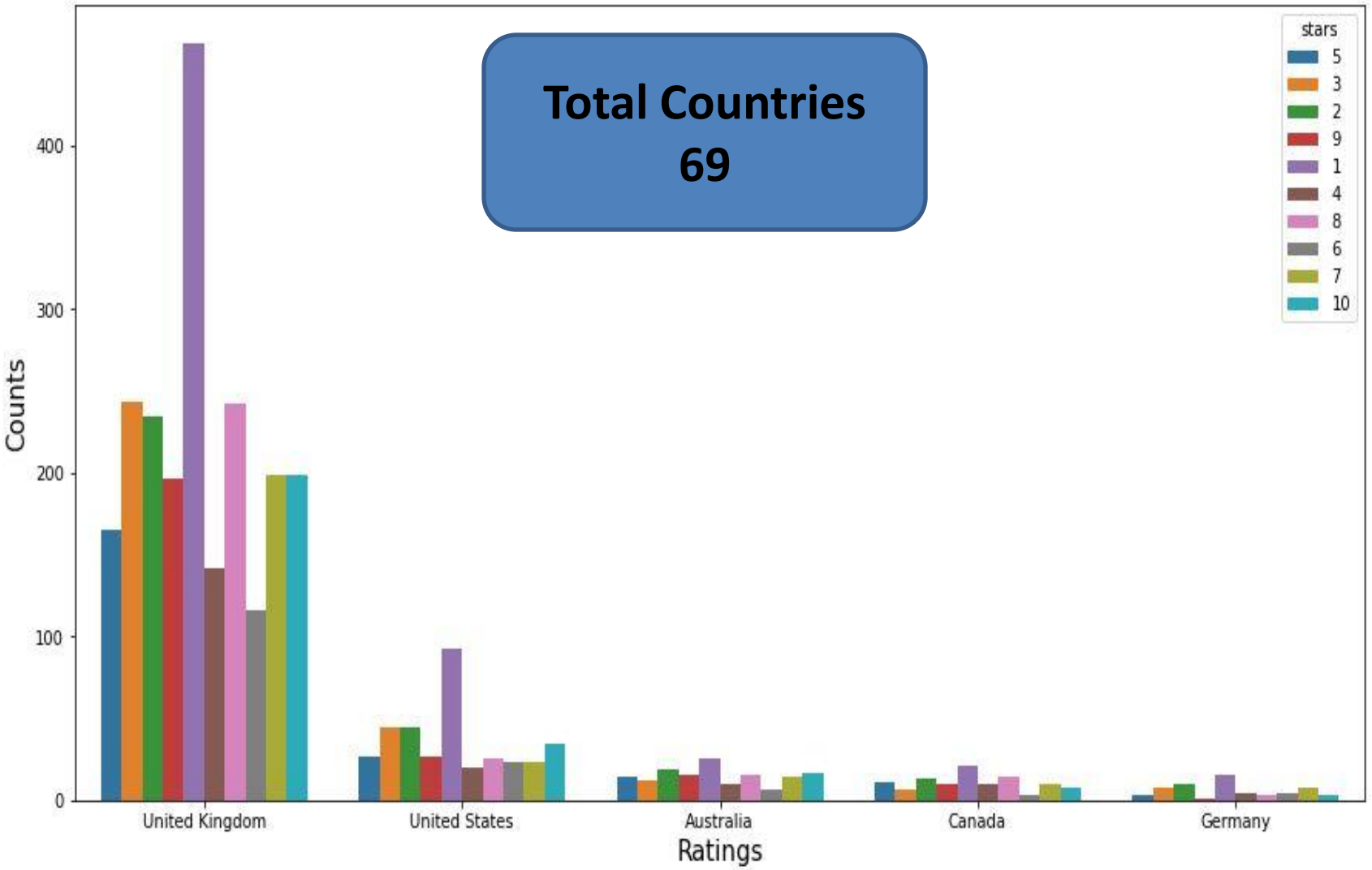


- Total unique counts of countries in our dataset is 69.
- This plots displays top 10 countries in our dataset which signifies that most of the customers are from UK followed by USA, Australia .

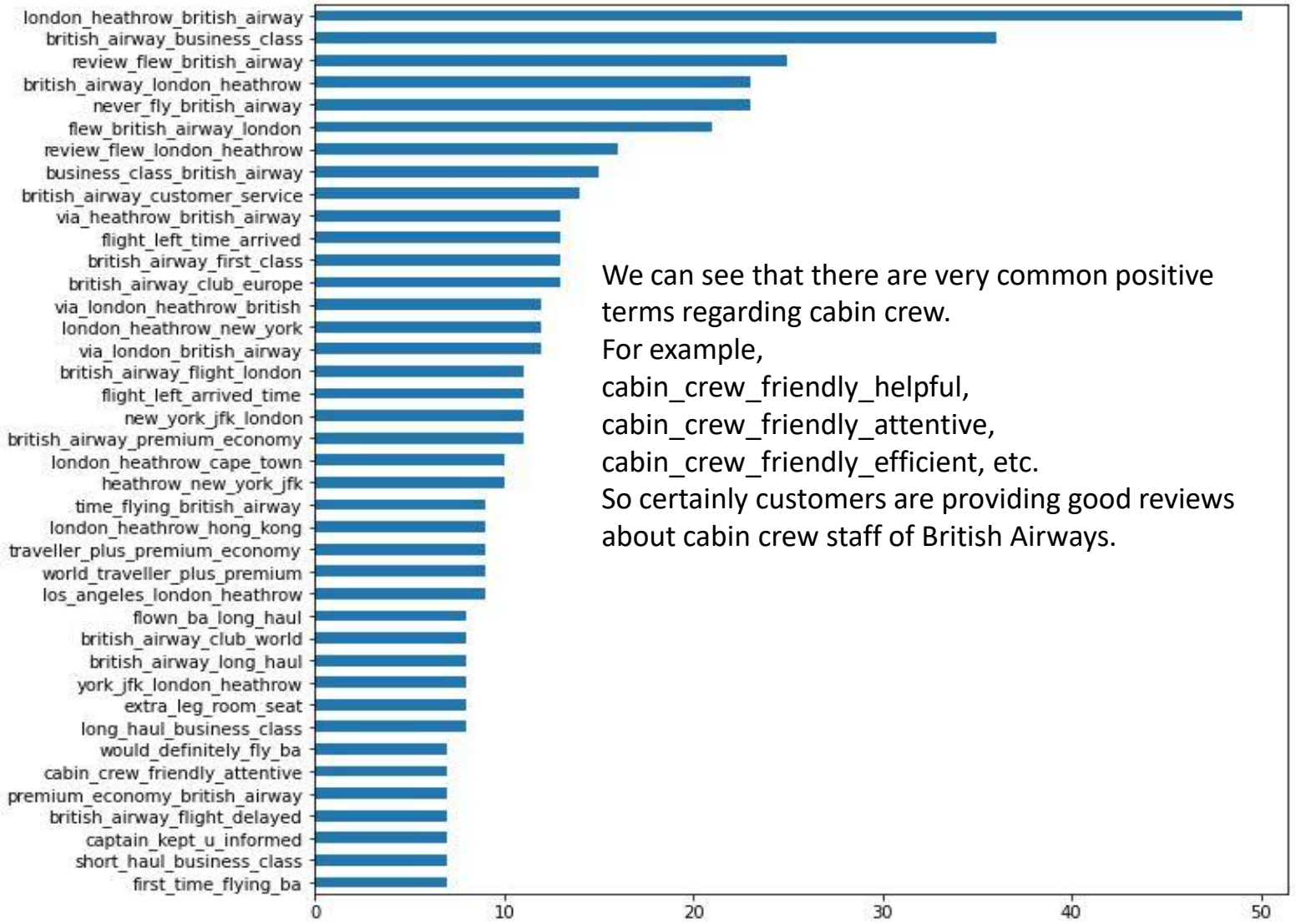
From which country most review comes



From which country most review comes



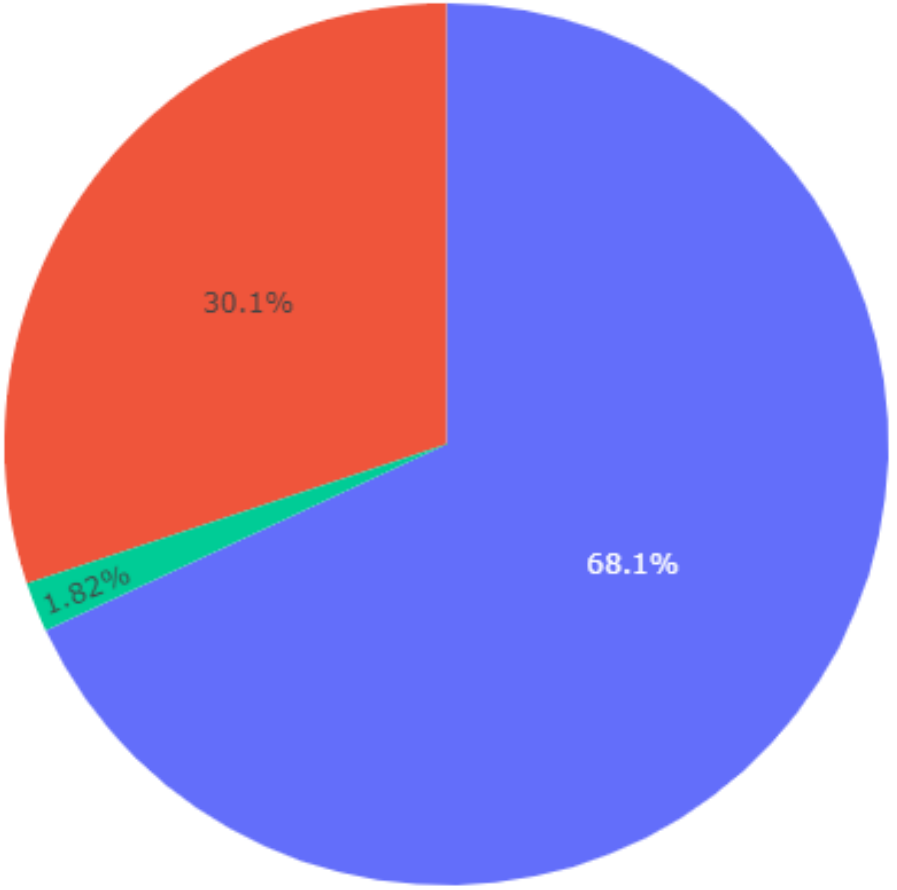
Word Frequency with N-gram



We can see that there are very common positive terms regarding cabin crew.
For example,
cabin_crew_friendly_helpful,
cabin_crew_friendly_attentive,
cabin_crew_friendly_efficient, etc.
So certainly customers are providing good reviews about cabin crew staff of British Airways.

4.Sentiment Analysis

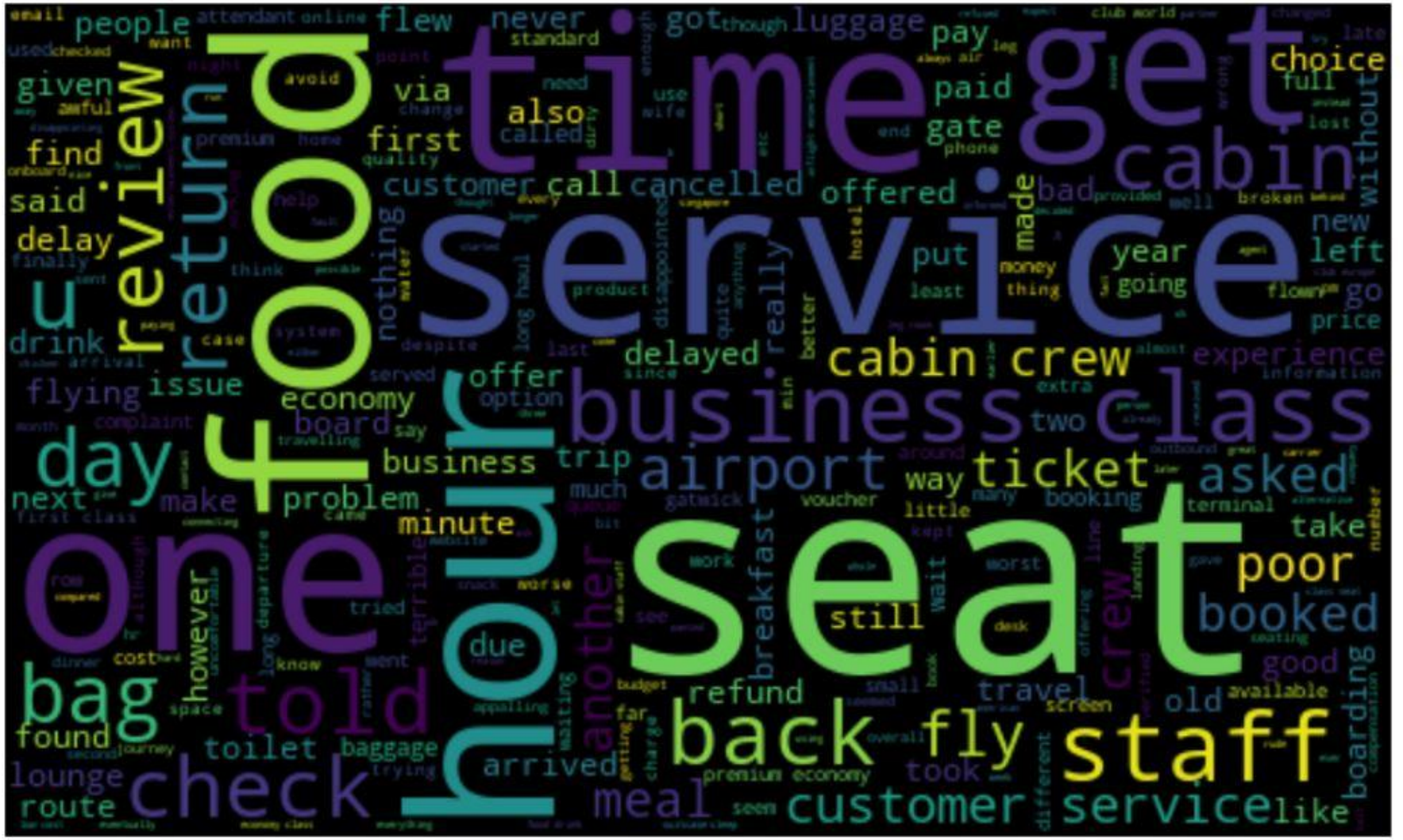
- For Sentiment Analysis, VADER Sentiment Analysis is used.
- VADER means Valence Aware Dictionary and sentiment Reasoner.
- VADER tells us how positive or negative a statement is.
- VADER sentiment analysis (well, in the Python implementation anyway) returns a sentiment score in the range -1 to 1, from most negative to most positive.
- Based on the sentiment score of a sentence it is classified into Positive ,Negative or Neutral.
- In this analysis the sentence is
- Positive >= 0.05
- Negative <= -0.05
- Else , its' Neutral



Pie-Chart showing distribution of sentiments

Wordcloud for Positive reviews

Wordcloud for Negative reviews



5.Model Building

- Predictors and Target variables were separated and stored into x, y.
- Target variable was categorical and was converted to numerical using label encoding.
- Before we can train our model we have to convert the text present in X into numeric values.
- For that we have used TF-IDF.
- After applying TF-IDF, shape of our dataset has changed to (3454 , 11130)
- Data was split into train & test in the ration of 70:30.

6. Comparison of Models Performance

	Model	Train Accuracy	Test Accuracy
0	Naive Bayes Classifier	0.918908	0.603664
1	SVM_linear Classifier	0.962350	0.831244
2	SVM_Sigmoid Classifier	0.928010	0.822565
3	SVM_rbf Classifier	0.978486	0.793635
4	RadnomForest Classifier	1.000000	0.757956
5	GradientBoost	0.947456	0.807136
6	AdaBoost	0.719073	0.697203
7	XGBoost	1.000000	0.809065
8	LogisticRegression Classifier	0.937526	0.812922
9	DecisionTree Classifier	1.000000	0.721311

7.Conclusion

Best Model for Prediction

1.Best Model will be SVM with kernel as Linear with 83 % accuracy.
2.SVM_Linear shows maximum accuracy on test data. This can be seen in the confusion matrix also.
3.Train accuracy is also good as compared to other models.
4.Variation is also less as compared to other models.

Key Insights

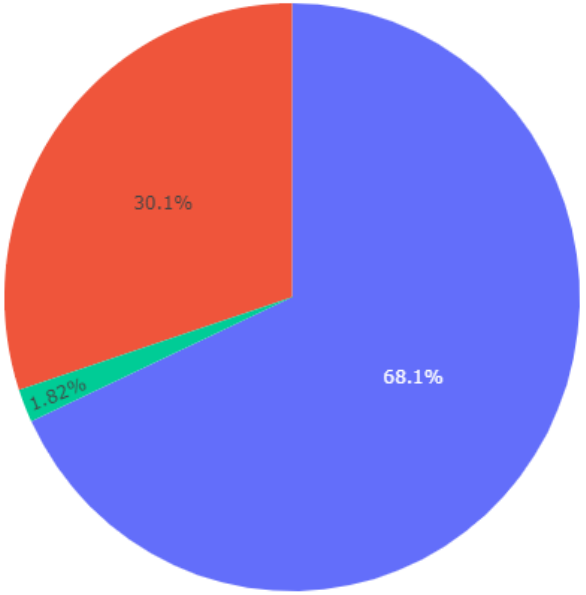
Total Countries 69

Average Overall Rating 4.84 /10

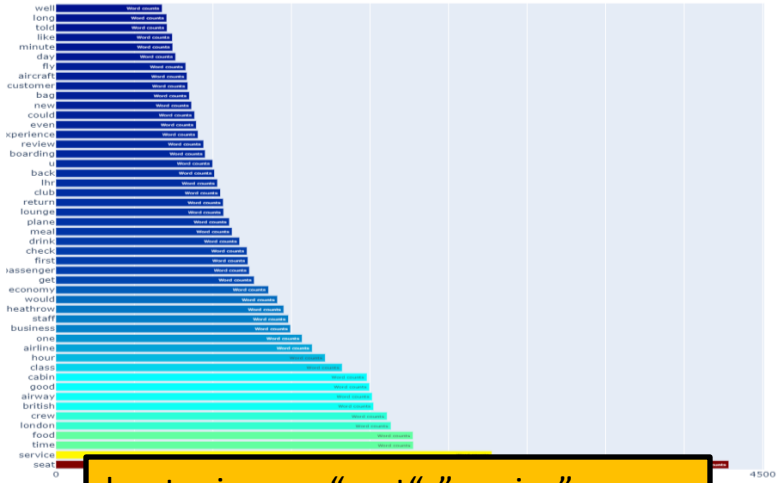
3460 Total reviews

83% Model Accuracy

Maximum comments are positive



Customers are providing good reviews about cabin crew staff of British Airways.



key topics are “seat“, ”service”, “time”, “food” and “staff”, showing that people are actively talking about their experience and staff.