

MSc. in Data Analytics
Data warehousing and Business Intelligence Project
Data Warehouse for Zomato
Name: Mandar Vaidya
Student no. – X17153409
National College of Ireland

Basic Requirement:

Food is the necessity of life. Restaurant search, ordering food online industry is increasing at a rapid rate. In that, Zomato has become one of the major firms not only in India, but in the world. To track which restaurant, cuisine, area food etc. and checking customer reviews needs to be updated every time. This data is vital for the success for that firm. Today, data is playing a vital role in predictive analytics, we can use that data so that which items are trending and then we will be able to make some decisions out of it.

Idea and Approach:

I wanted to do a project on food applications (Zomato) because of the curiosity to understand how some factors play a major role in creating this massively useful app. In Zomato Data warehouse, we will be using same concepts which are used in DWBI to generate results to get few business queries. In this project, we will be creating a Data Warehouse for Zomato. Based on this we will decide what factors are contributing to such application. With the help of Business Query tool, we will decide, which country has which cuisine demand etc and few other queries through Tableau.

Introduction:

According to (Inmon, 2005), a data warehouse is “subject oriented, integrated, non-volatile, time variant collection of data for management’s decision making”. The paper explains the complete design and implementation of the data warehouse using SSIS for ETL, SSAS for building the cube and Power BI for visualization of case studies. Data Warehousing and Business Intelligence (DWBI) is growing in today's data world. It is the best method for decision making and creating some trending outputs from the previously present data. Data Warehouse is a system used for reporting and data analysis, and is considered a core component of business intelligence. It is a relevant data on which we perform operations and statistics and based on those stats, reports and decisions are generated.

Tools and Technologies used:

Programming Languages:

- R for sentiment analysis on reviews

Programming API's:

- Zomato API for customer reviews

Tools:

- Microsoft Excel for basic operations
- Statistical Package for the Social Sciences (SPSS) for statistical Analysis

Database Management:

- SQL Server Integration Services (SSIS) for ETL
- SQL Server Analysis Services (SSAS) for building and cube development
- Tableau for Business Queries

The steps followed in our DW design are:

- 1) Source data collection
- 2) Developing the data warehouse through Star Schema Approach
- 3) Design for ETL
- 4) Developing ETL
- 5) Cube development
- 6) Loading the cleaned data in Data warehouse
- 7) Business Queries through Tableau

Data Sources:

- Kaggle: Structured (.csv) file was downloaded. (structured data).
- Zomato: Reviews (.csv) were extracted by using R programming language for various restaurants (unstructured data).
- Mockaroo: Semi-structured data (.csv) file was created because less data-sets were available.

Architecture and Design:

The most popular approaches for building a data warehouse are Inmon and Kimball's approach. Kimball's approach is bottom-up and Inmon's approach is top-down design. For generating Zomato Analysis Kimball's approach was used. As, Kimball emphasizes on creating report and analysis first and then data warehouse is created. This approach follows where smaller databases are merged up and located into the data warehouse. Benefits/Reasons for selecting this approach.

Reasons for selecting Kimball's approach:

- Time taken for designing the data is less.
- Focus is on process of building the data warehouse
- Model focuses on dimensions which are the main part in data analysis.
- All the data has not to be linked. Due to such presence, Kimball approach is widely used.
- Architecture is mostly used in slow changing time frame systems

Star Schema: The star schema architecture is the simplest data warehouse schema. It is called a star schema because the diagram resembles a star, with points radiating from a centre. The centre of the star consists of fact table and the points of the star are the dimension tables. Usually the fact tables in a star schema are in third normal form(3NF) whereas dimensional tables are de-normalized.

Fact Table: A fact table typically has two types of columns: foreign keys to dimension tables and measures those that contain numeric facts. A fact table can contain fact's data on detail or aggregated level.

Dimension Table: A dimension is a structure usually composed of one or more hierarchies that categorizes data. If a dimension hasn't got a hierarchies and levels it is called flat dimension or list. The primary keys of each of the dimension tables are part of the composite primary key of the fact table. Dimensional attributes help to describe the dimensional value. They are normally descriptive, textual values. Dimension tables are generally small then fact table.

Benefits of Star Schema: Star schema has numerous capabilities which prove its efficiency in building data warehouse.

- It's a simple structure and easy to understand.
- It is great query effective because foreign key is present in less quantity. Each dimension table has only one-dimension table connected to it in star schema.
- Normalized data is already stored in dimensions so need of normalization.
- It is most commonly used in data warehouse implementations as it is widely supported by many business tools.
- Data navigation is faster because data can be traversed with connected nature of fact and dimension tables.

Designing the Data Warehouse:

For the Zomato data warehouse, we have decided 3 dimensions and they are: Dim Restaurant, Dim Sentiment, Dim Cuisine. Below is the detailed explanation of the dimensions defined in the warehouse:

- DimRestaurant: DimRestaurant is the dimension which consist of the restaurant data which has restaurant id, restaurant name with country code, locality, longitude, latitude, price range, table booking, online delivery, rating colour, votes etc. With votes, price range we can keep an eye on which locality has good restaurants.

| MOLAP.DWBI Proje...bo.Dim_Restaurant | | | |
|--------------------------------------|--------------|-------------------------------------|--|
| Column Name | Data Type | Allow Nulls | |
| Rest_ID | int | <input type="checkbox"/> | |
| [Restaurant ID] | float | <input checked="" type="checkbox"/> | |
| [Restaurant Name] | varchar(500) | <input checked="" type="checkbox"/> | |
| [Country Code] | float | <input checked="" type="checkbox"/> | |
| City | varchar(500) | <input checked="" type="checkbox"/> | |
| Locality | varchar(500) | <input checked="" type="checkbox"/> | |
| [Locality Verbose] | varchar(500) | <input checked="" type="checkbox"/> | |
| Longitude | float | <input checked="" type="checkbox"/> | |
| Latitude | float | <input checked="" type="checkbox"/> | |
| [Average Cost for two] | float | <input checked="" type="checkbox"/> | |
| Currency | varchar(500) | <input checked="" type="checkbox"/> | |
| [Has Table booking] | varchar(500) | <input checked="" type="checkbox"/> | |
| [Has Online delivery] | varchar(500) | <input checked="" type="checkbox"/> | |
| [Is delivering now] | varchar(500) | <input checked="" type="checkbox"/> | |
| [Switch to order menu] | varchar(500) | <input checked="" type="checkbox"/> | |
| [Price range] | float | <input checked="" type="checkbox"/> | |
| [Aggregate rating] | float | <input checked="" type="checkbox"/> | |
| [Rating color] | varchar(500) | <input checked="" type="checkbox"/> | |
| [Rating text] | varchar(500) | <input checked="" type="checkbox"/> | |
| Votes | float | <input checked="" type="checkbox"/> | |
| COUNTRY_NAME | varchar(50) | <input checked="" type="checkbox"/> | |
| | | <input type="checkbox"/> | |

- DimSentiment: DimSentiment is the dimension which consists of restaurant id with reviews has the main thing and standard deviation(SD) and average sentiment of reviews.

| MOLAP.DWBI Proje...bo.Dim_Sentiment | | | |
|-------------------------------------|-----------------|--------------|-------------------------------------|
| | Column Name | Data Type | Allow Nulls |
| ▶ | Sentiment_ID | int | <input type="checkbox"/> |
| | [Restaurant ID] | float | <input checked="" type="checkbox"/> |
| | Reviews | varchar(500) | <input checked="" type="checkbox"/> |
| | element_id | float | <input checked="" type="checkbox"/> |
| | word_count | float | <input checked="" type="checkbox"/> |
| | sd | float | <input checked="" type="checkbox"/> |
| | ave_sentiment | float | <input checked="" type="checkbox"/> |
| | | | <input type="checkbox"/> |

- DimCuisine: DimCuisine is the dimension which consist of country code, cuisines which is popular in that restaurant and their revenue in millions.

| MOLAP.DWBI Project - dbo.Dim_Cuisine | | | |
|--------------------------------------|-----------------------|--------------|-------------------------------------|
| | Column Name | Data Type | Allow Nulls |
| ▶ | Cuisine_ID | int | <input type="checkbox"/> |
| | [Restaurant ID] | float | <input checked="" type="checkbox"/> |
| | [Country Code] | float | <input checked="" type="checkbox"/> |
| | Cuisines | varchar(500) | <input checked="" type="checkbox"/> |
| | [Revenue in Millions] | float | <input checked="" type="checkbox"/> |
| | | | <input type="checkbox"/> |

- Fact Table: For our Zomato analysis data warehouse we have created one fact table which was connected to different dimensions with surrogate key relationships.
- Star Schema: To connect our fact tables and dimension tables Star Schema is used as below.

Extract Transform and Load (ETL):

For generating data warehouse ETL is the starting step. It involves extracting data from various sources, transforming the data and then loading it into database. To provide a business value, it is very important that data is clean and appropriate. ETL helps to achieve the process of preparing the data for analysis. SSIS is used for ETL process creation with the help of SSIS toolbox created during process.

Data Extraction: In this project, multiple sources of data were used. Out of which, one is structured, one is unstructured and one is semi structured. Structured data was chosen from kaggle.com, as it contains all the information regarding the restaurants, ratings, country code etc. Unstructured data was extracted from Zomato website by using R programming language by using various libraries like httr, Json, Jsonlite, curl etc. Sentiment analysis were done on restaurant reviews by using SentimentR, data. Table package. Semi structured data was chosen from Mockaroo to create some fields as limited datasets were available.

Data Cleansing: Microsoft Excel was used during cleaning process to remove special characters and identify the missing values. Data formatting, converting json to csv format, removing duplicates while "Kutools" for deleting special characters.

Data Transformation: Transforming the data into meaningful data is the basic step for transformation step in ETL. Data transformation like clustering and then formatting the data is representable format. Apart from clustering and formatting, data conversion was used in SSIS many cases where the data type did not match.

Data Loading: Transformed data should be loaded into the data warehouse hence some relationships are decided before loading the data into warehouse and then data is loaded into the data warehouse. In this part of the project, all the dimension tables, fact table and cube is loaded via SSIS (SQL Server Integration Services). From extraction to loading the entire process is automated via SSIS workflow.

Staging Arena:

Staging databases are loaded first because these databases are used for transformation purposes. Then, Data is cleaned into these databases. In our case, we changed the data types of the fields and increased their limit. As well as, we merged different sources of data into single staging table for creating dimension from it.

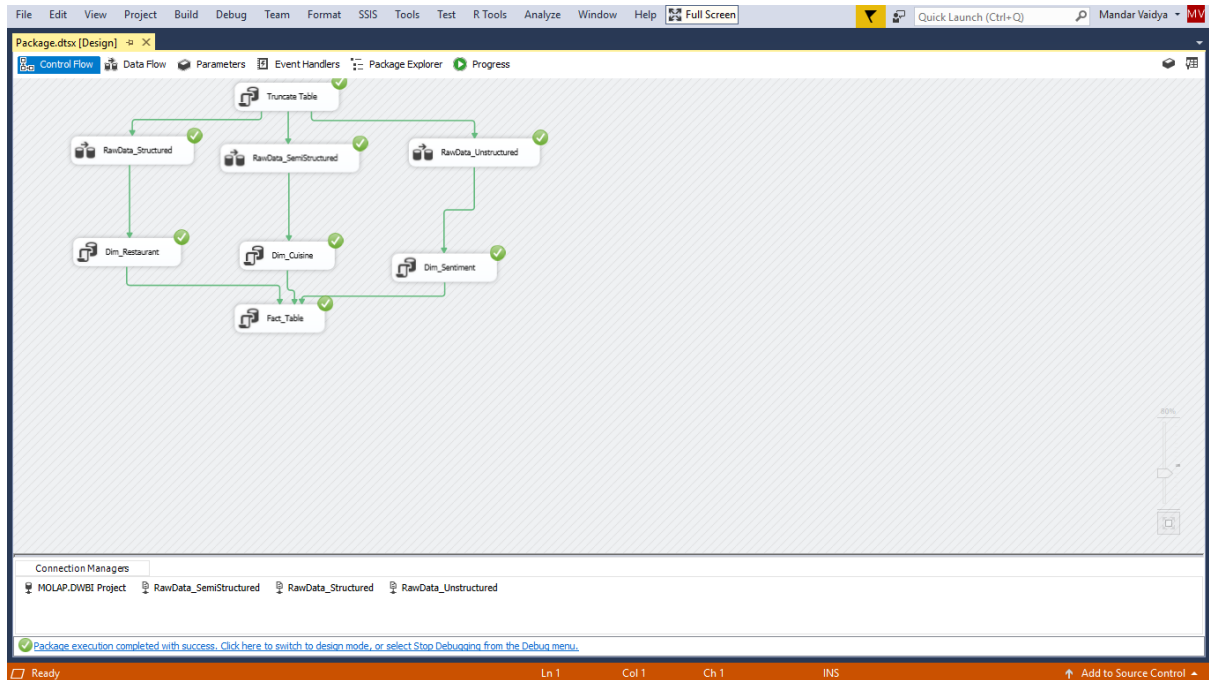
Staging tables which has cleaned data then to the next step which is loading the dimension tables. Dimension tables are the one of the main factors to go further. So, with the help of staging tables, factors are used in analysis of the data. Those factors are known as dimension factors. Mapping of the columns while loading dimension tables is very important and it should be done carefully. Otherwise ETL flow might get terminated due to such mistakes.

Loading Fact Table: After loading the dimension tables, next is loading the massive important step fact table. Fact table has the primary keys of dimension tables and some measures from the dimension tables which are used in analysis with some aggregation rules. With the help of (Joins) we have populated different dimension tables into fact table.

| MOLAP.DWBI-X17153...- dbo.Fact_Table | | | |
|--------------------------------------|------------------------|--------------|-------------------------------------|
| | Column Name | Data Type | Allow Nulls |
| ▶ | Cuisine_ID | int | <input type="checkbox"/> |
| | Rest_ID | int | <input type="checkbox"/> |
| | Sentiment_ID | int | <input type="checkbox"/> |
| | [Restaurant ID] | varchar(500) | <input checked="" type="checkbox"/> |
| | [Country Code] | varchar(500) | <input checked="" type="checkbox"/> |
| | [Revenue in Millions] | varchar(500) | <input checked="" type="checkbox"/> |
| | Longitude | varchar(500) | <input checked="" type="checkbox"/> |
| | Latitude | varchar(500) | <input checked="" type="checkbox"/> |
| | [Average Cost for two] | varchar(500) | <input checked="" type="checkbox"/> |
| | [Price range] | varchar(500) | <input checked="" type="checkbox"/> |
| | [Aggregate rating] | varchar(500) | <input checked="" type="checkbox"/> |
| | Votes | varchar(500) | <input checked="" type="checkbox"/> |
| | sd | varchar(500) | <input checked="" type="checkbox"/> |
| | ave_sentiment | varchar(500) | <input checked="" type="checkbox"/> |

Control Flow:

In this flow, the data is flown from different data flow tasks in SSIS. In truncate table, all the staging tables are truncated so that all the duplicate values will not generated after multiple runs. It is one of the most important step in Data Warehousing. After that excel files are loaded into the staging tables. Staging tables are different in sources hence for easy creation of dimensions we have merged them in the databases control flow. After that dimension were loaded into it. After that dimension were loaded into it. After dimensions loading phase fact table is loaded from all the dimensions.



Deploying the Cube:

Multidimensional representation of the data can be carried out with the help of the data cubes in SSAS. In SAAS, we can analyse the data based on the measures we have chosen and based on the attributes of the dimensions. After deploying the cube, detailed analysis and reporting part of the data warehouse starts where Business Intelligence or Business queries are carried out.

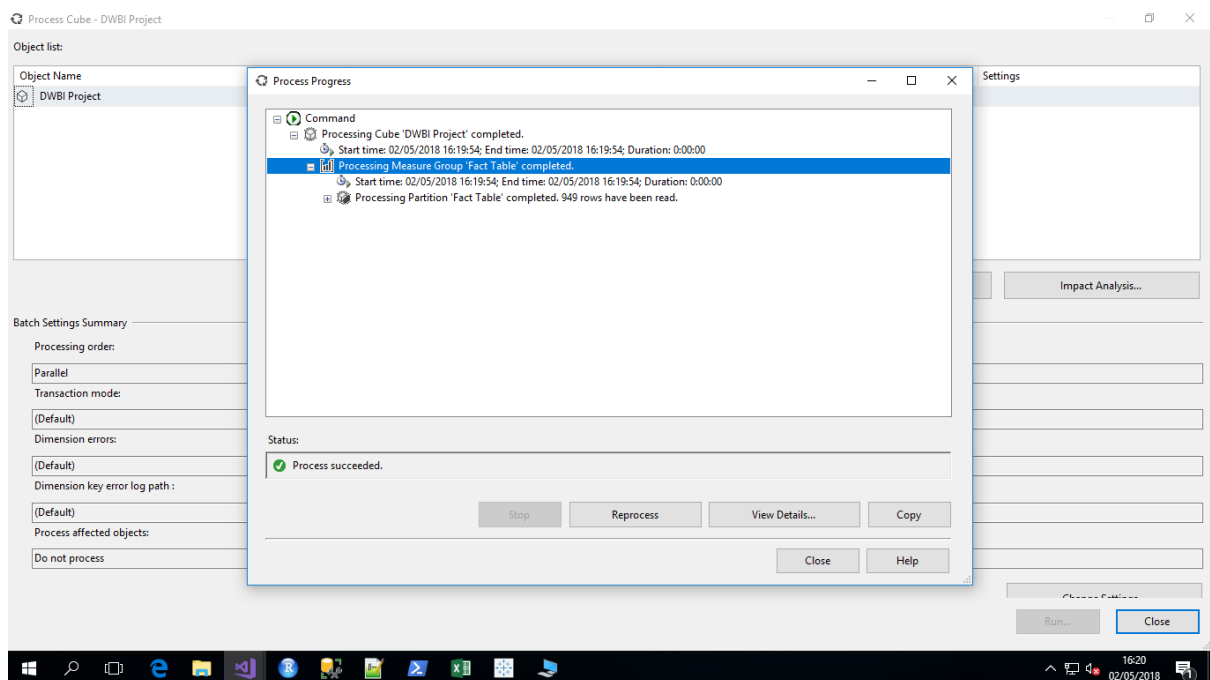
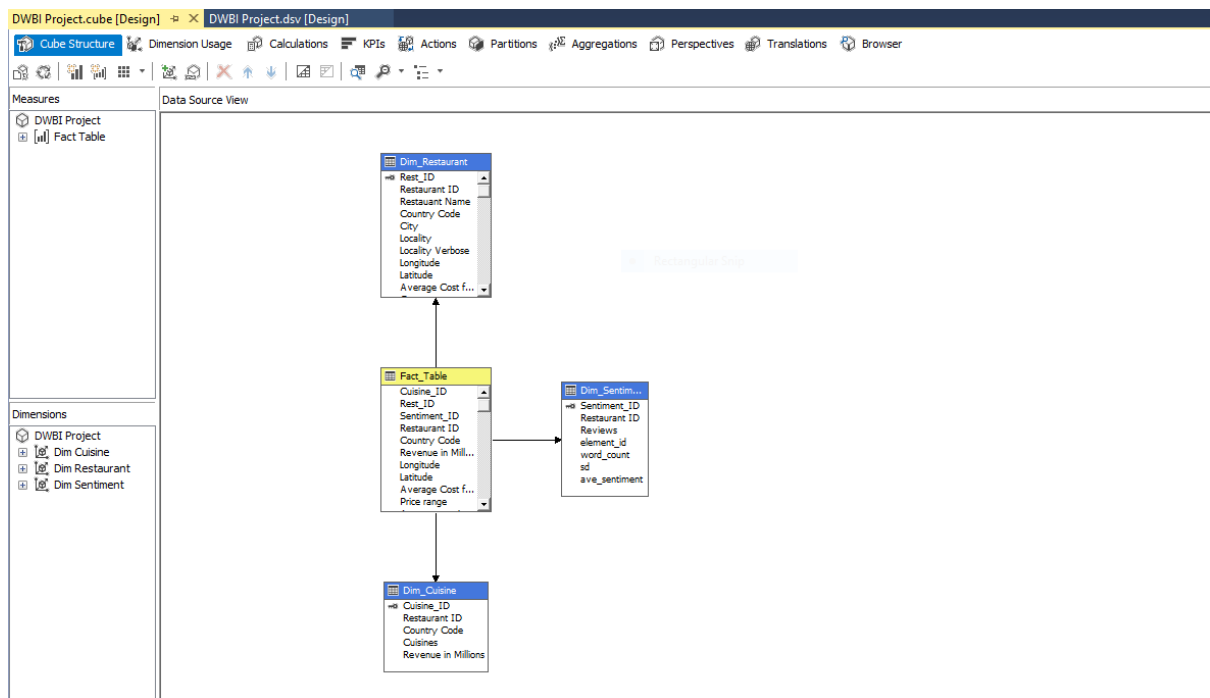
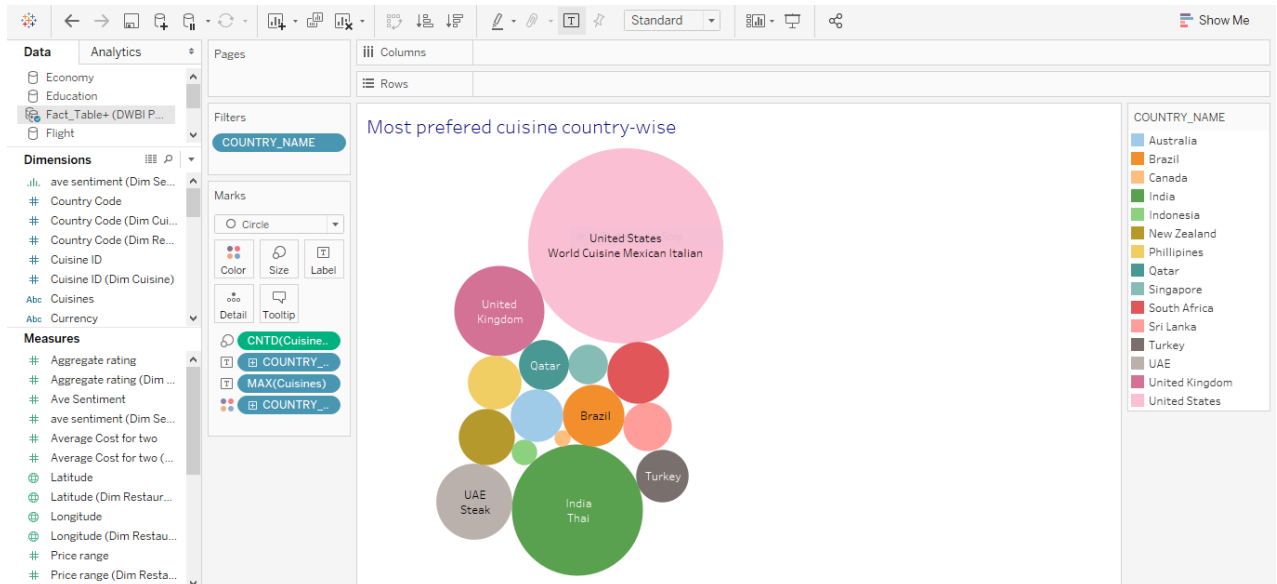


Tableau Analysis:

Tableau is a visualization tool that allows you to connect volumes of data and visualize in detail. It brings spreadsheets, databases, big data sources to create reports and analyse them between any objects.

Case 1: Most preferred cuisine country-wise:

To check the above scenario, we will label country name and maximum cuisines and distinct count of cuisines as size and insert country names in colour in a packed bubble chart. Country name will be our dimension here.



Analysis: From the stats and analysis above, we can say that United States prefers 3 types of cuisines; world cuisine, Mexican and Italian with the maximum number of distinct cuisines 290; whereas in people in India prefer Thai the most who came 2nd with a total number of 131 distinct cuisines.

| Country Name | Number of Distinct Cuisines | Maximum Cuisines |
|----------------|-----------------------------|----------------------------------|
| Australia | 21 | Seafood, Asian Grill, Sushi |
| Brazil | 29 | South Indian, Brazilian, Chinese |
| Canada | 2 | Mexican |
| India | 131 | Thai |
| Indonesia | 5 | Seafood |
| New Zealand | 24 | Southern |
| Philippines | 22 | North Indian, Chinese, Mughlai |
| Qatar | 19 | Southern |
| Singapore | 12 | Seafood |
| South Africa | 19 | Turkish, Arabian, Middle Eastern |
| Sri Lanka | 18 | South Indian |
| Turkey | 21 | Pizza, Cafe, Italian |
| UAE | 44 | Steak |
| United Kingdom | 62 | World Fusion, Fast Food |
| United States | 290 | World Cuisine, Mexican, Italian |

Case 2: Relation between aggregate rating and average sentiment based on restaurant features

To check the above scenario, we have taken aggregate rating as dimension in columns and maximum average sentiment and has table booking features in rows whereas in detail we will take country name and has online delivery in colour mark in a building line chart.



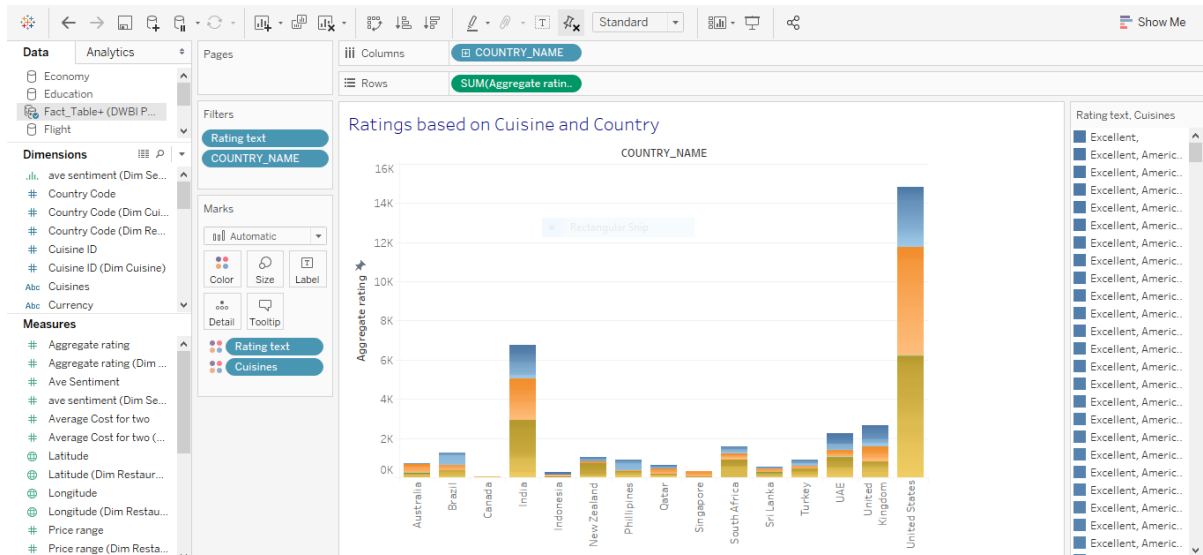
Analysis: The restaurant which has no table booking and no online delivery has the maximum average sentiment of 1.832 and maximum aggregate rating as 37.00 which is in United States.

While, the restaurant which has table booking but no online delivery has a comparatively lesser average sentiment of 0.850 but maximum aggregate rating as 47.00.

Thus, we can conclude that, restaurant features like table booking and online delivery play a major role in average sentiment and aggregate ratings in this case study.

Case 3: Ratings based on Cuisine and Country

To check and analyse the above scenario we have taken Country Name in columns which is dimension where as aggregate ratings as a measure in rows. We have filtered rating text too. As well as, marking rating text and cuisines for a better analysis in bar plot in Tableau.



Analysis: Now a day, ratings play a major role in every field whether it is negative or positive or natural it has a significant importance in this world and add to it in restaurant field it is major object to grow your business. In this data set, rating text is qualified in 5 types: Poor, Good, Very Good, Aggregate and Excellent. But, for in depth analysis we have taken only 3 texts: Good, Very Good and Excellent.

As, we say in the chart, Australia, Brazil, Qatar, Turkey etc. have almost equal ratings where as in India, UAE, United Kingdom it is slightly higher in Good and Excellent categories. But, a massive column is there in United States with Good being the major rating followed by Very good and Excellent, thus by making it clear that United States value ratings a lot.

Thus, we can conclude that, customers play a major role in giving reviews and ratings which in turn impacts the business of a restaurant.

Case 4: Average Sentiment by Restaurant ID

In this case study, we have taken country name as a dimension in Columns whereas Average Sentiment as a maximum measure. With markings in Colour denotes as Restaurant ID and distinct count as the average sentiment as a detailed major in a simple chart. In this query, we are trying to find maximum sentiments using restaurant ID.



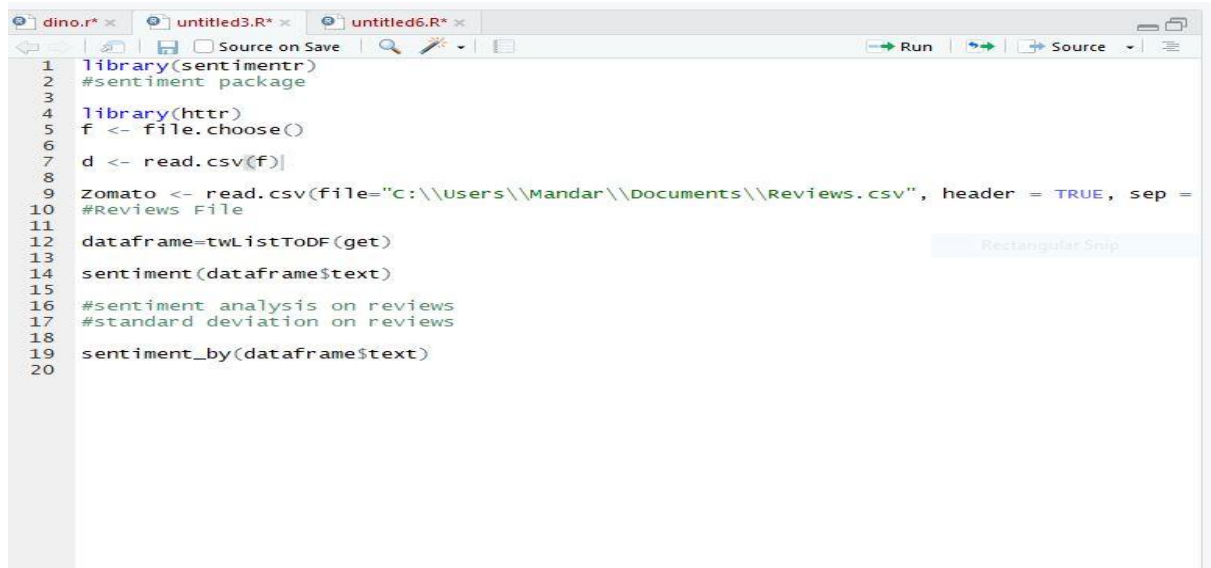
Analysis: As, we have lots of restaurants ID, that's why, comparing them with country names would be little easier.

As seen in Tableau, Restaurant ID 2800881 in United States has the maximum average sentiment of 1.832 whereas, restaurant ID 17694716 comes second with the maximum average sentiment of 1.154.

We can conclude that reviews play a big role in sentiment analysis.

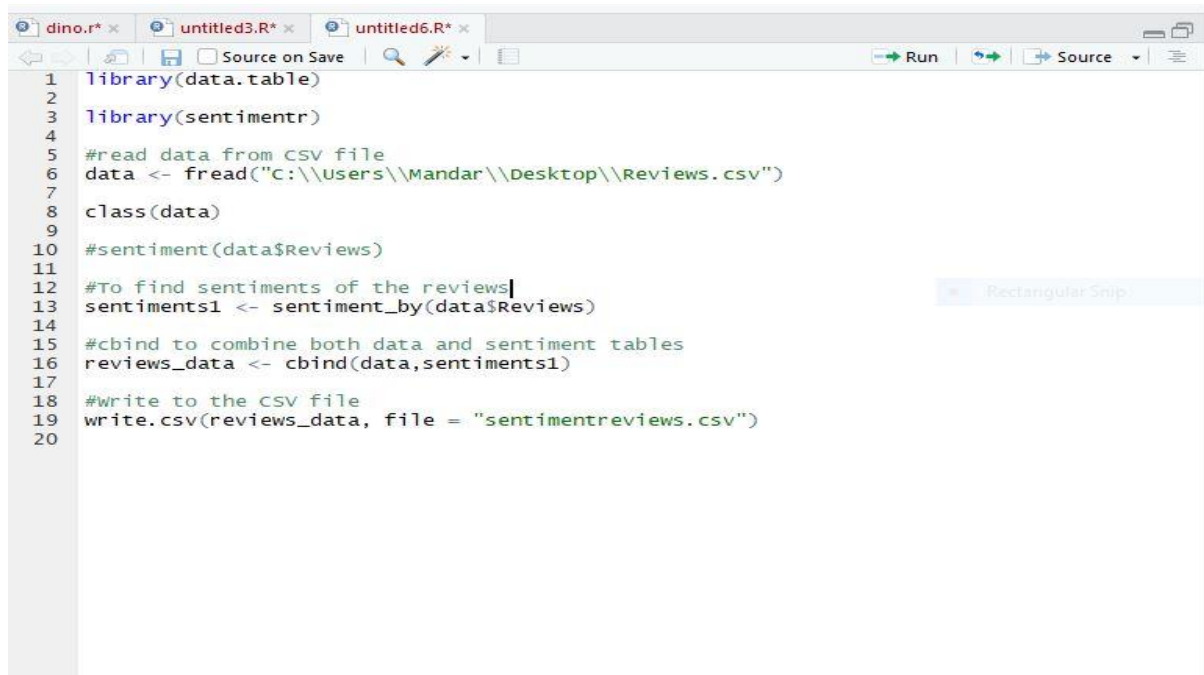
Appendix:

➤ Sentiment Analysis Code using R



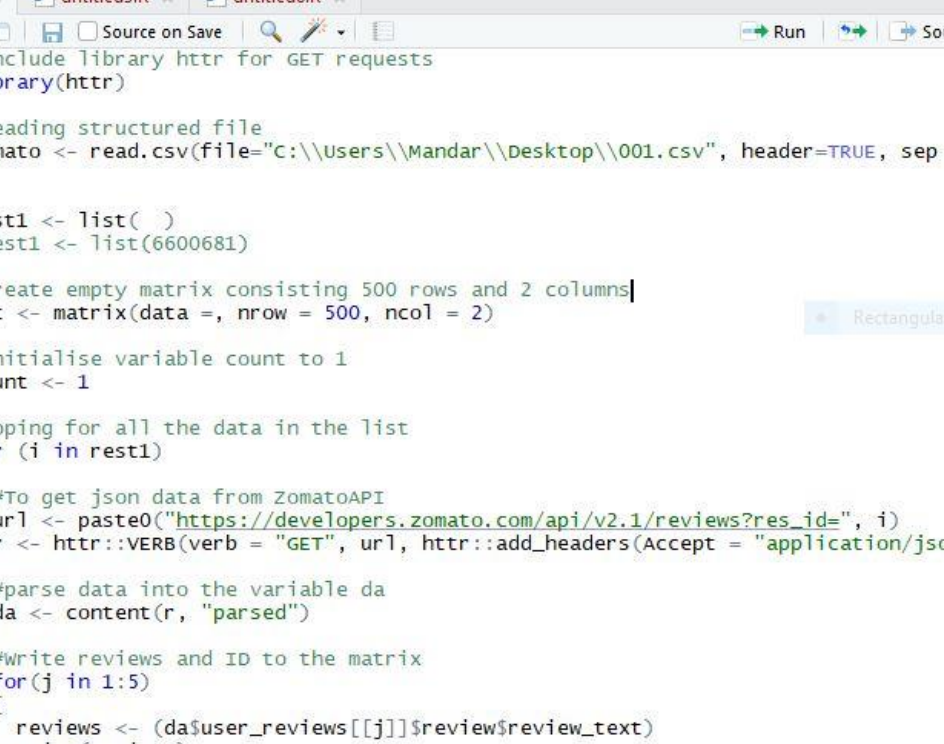
```
1 library(sentimentr)
2 #sentimentr package
3
4 library(httr)
5 f <- file.choose()
6
7 d <- read.csv(f)
8
9 Zomato <- read.csv(file="C:\\Users\\Mandar\\Documents\\Reviews.csv", header = TRUE, sep =
10 #Reviews File
11
12 dataframe=twListToDF(get)
13
14 sentiment(dataframe$text)
15
16 #sentiment analysis on reviews
17 #standard deviation on reviews
18
19 sentiment_by(dataframe$text)
20
```

➤ Sentiment Reviews Code using R



```
1 library(data.table)
2
3 library(sentimentr)
4
5 #read data from CSV file
6 data <- fread("C:\\Users\\Mandar\\Desktop\\Reviews.csv")
7
8 class(data)
9
10 #sentiment(data$Reviews)
11
12 #To find sentiments of the reviews
13 sentiments1 <- sentiment_by(data$Reviews)
14
15 #cbind to combine both data and sentiment tables
16 reviews_data <- cbind(data,sentiments1)
17
18 #Write to the CSV file
19 write.csv(reviews_data, file = "sentimentreviews.csv")
20
```

➤ Restaurant Reviews using R Code (Part I)



The screenshot shows the RStudio IDE interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. The toolbar contains icons for file operations and a 'Go to file/function' search bar. The tab bar shows three open files: dino.r, untitled3.R, and untitled6.R. The main editor window displays R code for scraping Zomato reviews. The code includes comments and uses the httr and readr packages. A 'Rectangular Snip' watermark is visible on the right side of the code editor.

```

1 #include library httr for GET requests
2 library(httr)
3
4 #reading structured file
5 zomato <- read.csv(file="c:\\Users\\Mandar\\Desktop\\001.csv", header=TRUE, sep = ",")
6
7
8 rest1 <- list( )
9 #rest1 <- list(6600681)
10
11 #create empty matrix consisting 500 rows and 2 columns
12 mat <- matrix(data = , nrow = 500, ncol = 2)
13
14 #initialise variable count to 1
15 count <- 1
16
17 #looping for all the data in the list
18 for (i in rest1)
19 {
20   #To get json data from ZomatoAPI
21   url <- paste0("https://developers.zomato.com/api/v2.1/reviews?res_id=", i)
22   r <- httr::VERB(verb = "GET", url, httr::add_headers(Accept = "application/json", `use`
23
24   #parse data into the variable da
25   da <- content(r, "parsed")
26
27   #write reviews and ID to the matrix
28   for(j in 1:5)
29   {
30     reviews <- (da$user_reviews[[j]]$review$review_text)
31     print(reviews)
32     mat[j+5*(count-1), 1] <- i
33     mat[j+5*(count-1), 2] <- reviews
34   }
35

```

➤ Restaurant Reviews using R Code (Part II)

```

30     reviews <- (da$user_reviews[[j]]$review$review_text)
31     print(reviews)
32     mat[j+5*(count-1), 1] <- i
33     mat[j+5*(count-1), 2] <- reviews
34 }
35
36 #increment the count by 1
37 count <- count+1
38 }
39
40 #write the table to the csv file
41 write.csv(mat, file = "myfile57.csv", row.names=FALSE)
42
43

```

References:

- Inman, W.H. (2005). Building the Data Warehouse. John Wiley & Sons.
- Ralph Kimball (2013). The data warehouse toolkit: Kimball Group.