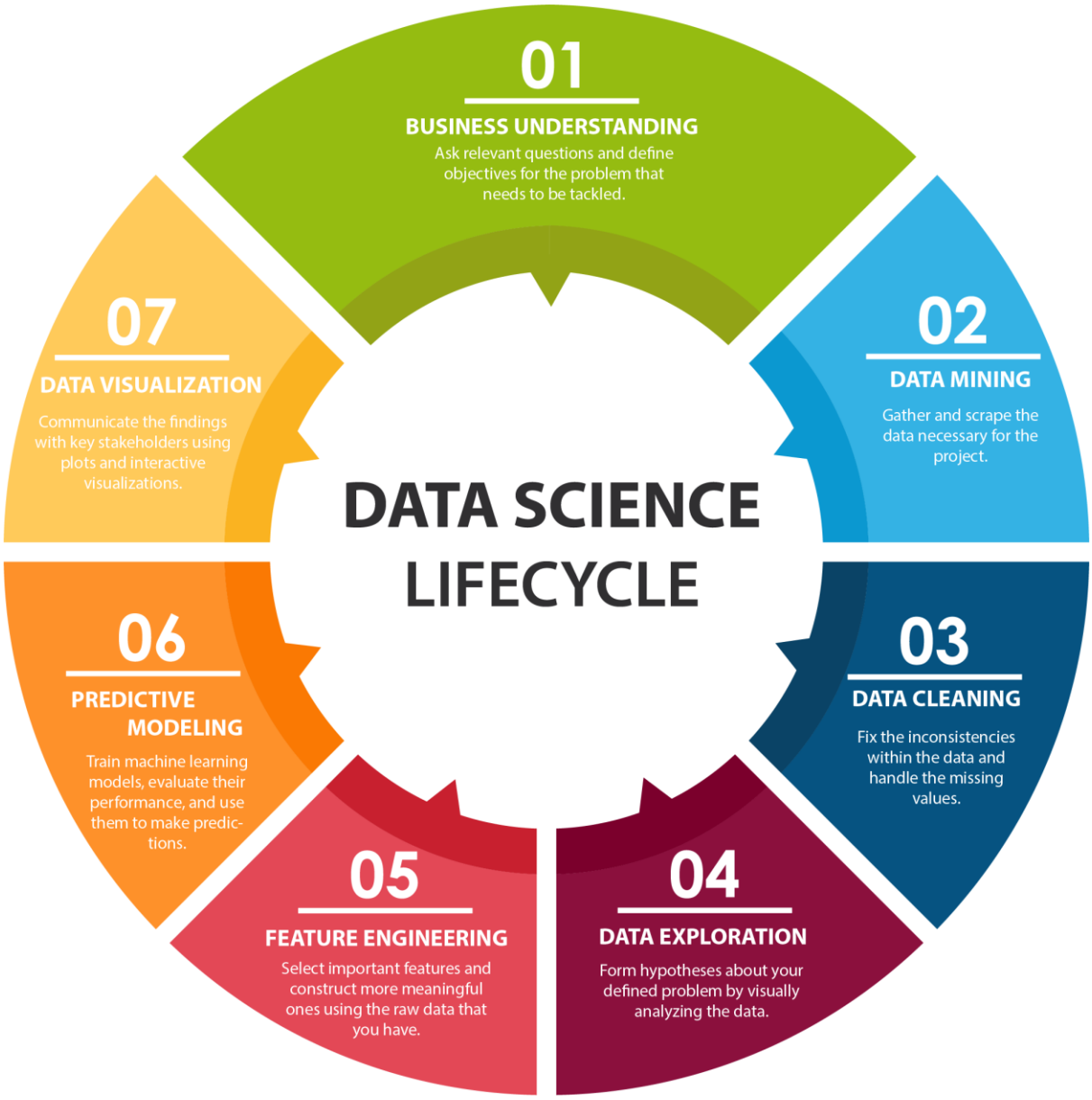
Capstone Project Assignment
Report - Week 02

1. Introduction

Business Understanding, for this assignment is collecting and categorizing data in order to evaluate a market analysis for restaurants in the city of Munich, Germany. Following this analysis, it will be possible to make a business decision on what kind of restaurant to open in a particular part of town. Most of this assignment is focused on the data lifecycle: collecting, preparing and processing data in order to categorize, visualize and deploy it. We consider the following data science lifecycle diagram:



Working with data is a significant part of data science, according to studies it is 80% of it, the rest is business understanding and evaluating business decisions in order to

present it to the relevant stakeholders. Stakeholders for this particular assignment could be startup investors that are considering opening a restaurant in the city of Munich. The data scientist could be a consultancy researching an evaluating the restaurant market in order to present the results to the startup investors.
So, after considering business understanding we move to data mining which is considered in the next part.

2. Data

Data mining is one of the main components of the data science lifecycle. Consider the following diagram:

This Capstone Project Assignment, starts with collecting data for Restaurants located in Munich, Germany. It streams data using the Yelp API - url='https://api.yelp.com/v3/businesses/search' - which collects max. 50 entries.
Data is then parsed, stored and optimized in DataFrames.
This Data Set allows us to parse and categorize restaurants: name, address, latitude, longitude, review counts, rating and obviously restaurant genres - ie. italian, bavarian, spanisch, tapas, vietnam, sushi, pan-asian, mexican, cocktail bar, bistro, japanese, chinese kebab, etc.
There are many Business Search APIs on the web: Google, Foursquare, Yahoo..., I choose Yelp API since it is quite straight forward implementing it into Python v3.6 - Version that is used in this assignment. Choosing the API is quite essential, since it specifies, which data types are streamed (request obj, json...) but also data parameters that are stored (Name, Address, coordinates, category, reviews...).
Not all APIs offer the same feature set as Yelp. The only limitation to the Yelp API is the maximum number of search entries which are limited to 50.

The analytic approach is to categorize data and later process the database with data science tools to extract information and highlight certain patterns.
This analytic approach is applied in data science since it connects business understanding with data requirements and data processing, which is the core of data science since it leads to data visualization, evaluation and deployment.
Once these final stages of the data lifecycle, are reached, we can evaluate results in order to influence business decisions.

Following the Data Lifecyle in Data Minning, described in our lecture, we consider the diagram "From Understanding to Approach" mentioned in Module 3, we considered Business Understanding and Analytic Approach in the previous section.

Now we shift to the "Data Chain" – Data Requirements/Collection/Understanding/Preparation/Modeling. All of the Python source code and data processing is included in the Jupyter Note Book found in Part 02, of this assignment.

2.1 Data Requirements, comprises applying software tools, to stream data from the web, store and process it in Python, we do this via the Yelp API, accessing - url='https://api.yelp.com/v3/businesses/search', we specify which parameters we stream – params = {'term':'restaurant','location':'Munich', 'limit':'50'}

2.2 Data Collection is implemented by, declaring a request object to store the Yelp API request Data - req=requests.get(url, params=params, headers=headers)

2.3 Data Preparation - is implemented by parsing the request object as text and json data types - JSON request data is often multi nested and quite complex to store in 2-D arrays, therefore a number of tools have been developed to flatten JSON data. Python offers many JSON flattening/normalization tools, which fall into Data Preparation, (data = json.loads(req.text)) – is a command to flatten data, since it is a multi nested JSON data text type. Since we are dealing with multi nested JSON objects there is also multi flattening which is applied:

1. df2 = pd.DataFrame.from_dict(req.json()['businesses']) - using this command we store JSON data in the df2 dataframe - columns coordinates, address and categories are nested ,see Part2 Notebook
2. d1 = pd.json_normalize(data["businesses"]) - we normalize data, d1 contains the normalized json dataframe - coordinates and address columns are flattened, the categories column is still nested
3. pd.DataFrame(d3["cat"].apply(pd.Series)) - is applied to flatten the categories=cat column - from here we can extract the restaurants category type - sushi, vietnamese, pan-asian, tapas, mexican, bavarian, italian..., see Part2 Notebook

As you can see the Data Preparation stage can be quite complex and it is a significant component of data science, often it entails normalizing and filtering data to make it easier to store and process with data science tools.
Now, we have flattened data stored in Data Frames, we can select the columns to consider and display: Name, Address, Phone Nr, Categories, Rating, Review Count, Latitude, Longitude. All other columns are deleted from DataFrame.

Data Evaluation, Visualization and Deployment are covered later in this Assignment, see the Jupyter Notebook.

3. Methodology

For the methodology section we will consider the Programming Language used in this assignment and throughout the IBM / Coursera - Course - Data Science – Python V3.6
Python Data Science tools, are implemented in Python using the various libraires, there is an extensive amount of libraries available on the web. Libraries have to be manually installed using the ! pip command

```
!pip install -U numpy
!pip install -U pandas
!pip install -U scipy
!pip install -U scikit-learn
! pip install plotly
!pip install beautifulsoup4
```

Lets look at the essential Python library packages for this assignment:

1. Pandas is to apply on data frames in order to implement data manipulation and analysis.
2. Beautifulsoup is to parse data from html and xml files, mainly Text and JSON data, used for WebScraping and Data Wrangling
3. Plotly is used for datay analytics and visualization tools, including interactive maps
4. SciPi used for scientific computing, optimization, linear algebra, integration, interpolation, special functions, FFT, signal and image processing, ODE solvers and other tasks common in science and engineering.

Once Data is cleaned, flattened and prepared it is stored in Dataframes, we can apply visualization tools using the various libraries

4. Results

A Pie Chart, displaying Restaurant Categories with relevant percentages is displayed:

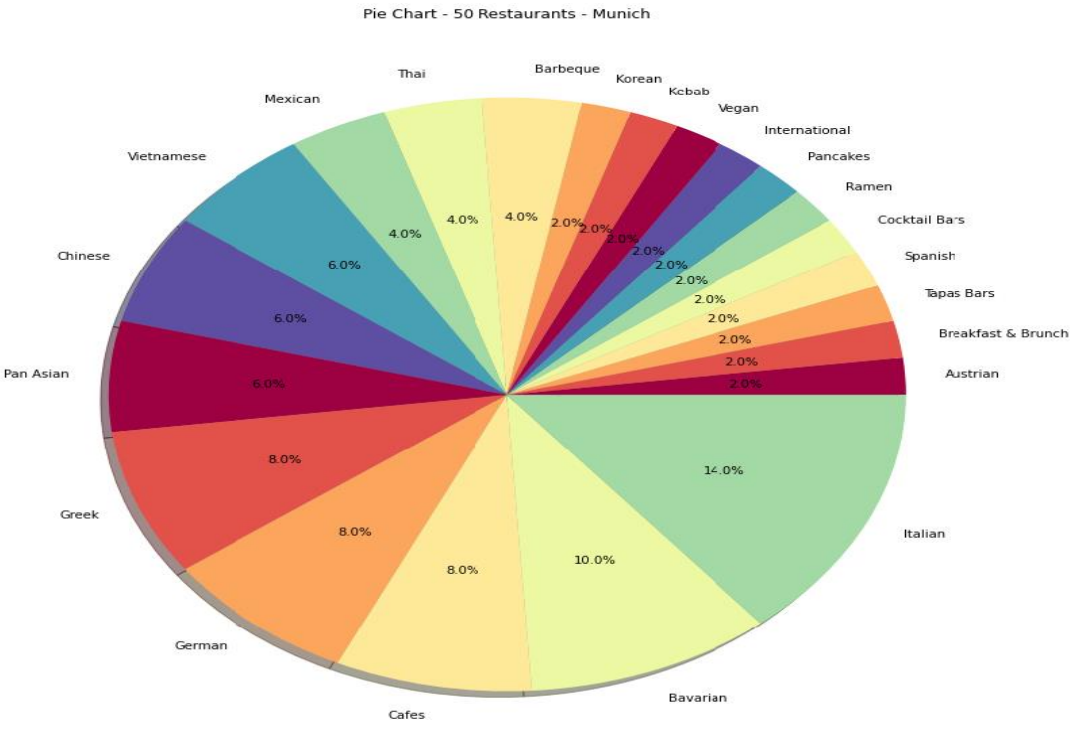


Fig. 1 – Pie Chart

A Histogram, Rating VS Number of Restaurants in visualized

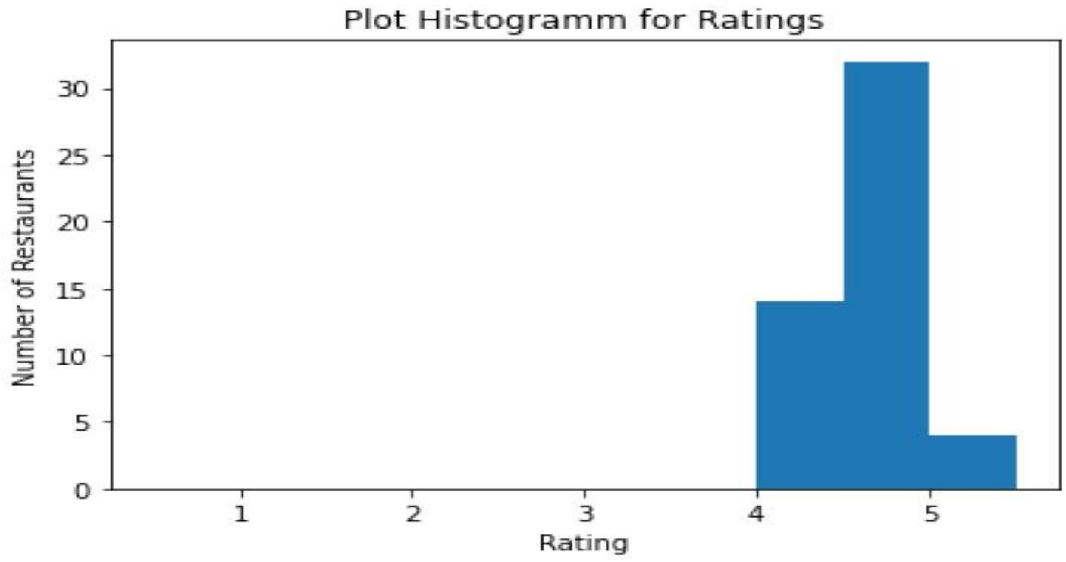


Fig. 2 – Rating Histogramm

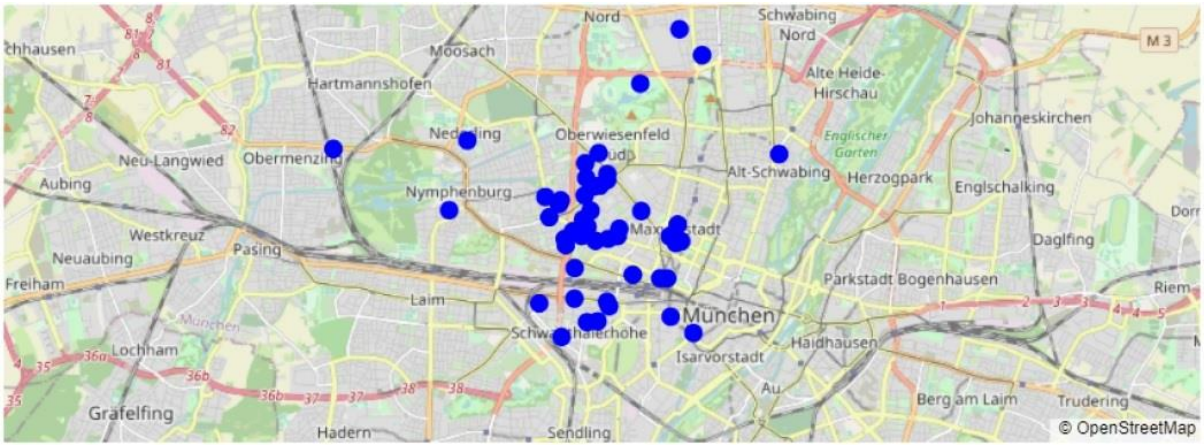


Fig. 3 - Interactive Map

5. Discussion

Looking at the pie graph in Fig.1, we can estimate that there is room for more Vegan/Vegeterian, Thai, Tapas, International, Latin American, Austrian, Pan-Asian, Vietnamese and Chinese Restaurants.

Histogram Rating. Fig. 2, displays that there are no ratings below 4,0 - most of them are 4,5 (64%), followed by 4,0 (28%) and 5,0 (8%). A Startup restaurant with a Rating of 5, could place it into the more “premium” segmentation, reflecting in higher dining prices that could be charged.

The Interactive Chart, Fig.3, visualizes categories with relevant percentages. Plotly library/ StreetMap have been used. The Popup Menu displays static data (Name, Address, Coordinates, Category…) it also displays the geographic location in the city Map, this is useful to asses where the might be demand for a particular Restaurant

category. The interactive Map displays a centered distribution of the Restaurants, there should be more demand for international cuisine in the suburbs East/West/North/South

6. Conclusion

This assignment gives a complete overview over the Data Science Lifecycle - Business Understanding, Analytic Approach initially.
After that, Data Mining is evaluated: Data Requirements, Collection, Understanding and Data Preparation are assessed and implemented. Here we have Software Engineering constraints and limitations that depend on the various projects and technologies. In this assignment I chose to work with the Yelp API, its limitation of 50 Search Results is obviously a constraint for this project, but it was chosen also to limit the length of this assignment. But, we have access with Python and its libraries to excellent Data Science Tools, for visualization, scientific evaluation end estimation. Another API with more search results could improve data science evaluation quality and accuracy.
Obviously more analysis, CPU Power and Memory and data processing should be applied.