SRM Institute of Science and Technology

Summer Internship Report-2024

Intern Name: Keshwanth G P
Reg. No.: RA2332014010076
Course: M.Sc. Applied Data Science
Company: Cognifyz Technologies
Position: Data Analyst Intern
Duration: May 2024 – June 2024
Location: Remote Work

1. Introduction

1.1 Overview of the Internship

During my internship at **Cognifyz Technologies**, I worked as a Data Analyst, focusing on extracting meaningful insights from restaurant data to help businesses optimize their strategies. My role involved tasks such as data cleaning, exploratory data analysis, visualization, and correlation studies. The goal was to uncover trends related to restaurant pricing, service offerings, and customer preferences, and to provide data-driven recommendations.

Throughout the internship, I gained valuable hands-on experience with advanced data analysis tools like **Python**, **pandas**, **Matplotlib**, and **Seaborn**. Working in a professional environment allowed me to manage real-world datasets effectively and apply analytical techniques to solve practical business problems. This experience significantly improved both my technical skills and my ability to interpret data for actionable insights.

The project included several key analyses, such as identifying the most popular cuisines, analyzing the distribution of restaurants across cities, examining price ranges and their correlation with services like online delivery and table booking, and understanding customer behaviour through ratings and reviews. These tasks helped me contribute meaningfully to the ongoing projects at Cognifyz Technologies and deepen my understanding of data-driven decision-making in a business context.

1.2 Company Profile

Cognifyz Technologies is a prominent data science company specializing in artificial intelligence (AI), machine learning (ML), and advanced business analytics. The company offers a range of services, including predictive analytics, fraud detection, and recommendation systems, enabling businesses to make data-driven decisions in real-time.

With a focus on innovation, Cognifyz empowers clients by providing customized data solutions that optimize operations and improve customer experiences. The company is also committed to nurturing new talent through internships and training programs, giving young professionals hands-on experience with the latest tools and techniques in data analytics.

Cognifyz's mission is to help businesses harness the power of data for smarter decision-making, solidifying its position as a key player in the fast-growing data analytics industry.

2. Project Description

2.1 Problem Statement

The primary task was to analyze restaurant data to uncover trends and patterns that could help businesses optimize their services. The dataset included variables like price range, cuisines, city location, customer ratings, and service offerings (online delivery and table booking).

The project aimed to answer key business questions, such as:

- What are the most popular cuisines, and do certain combinations impact ratings?
- How do price range and city location influence the availability of services like online delivery and table booking?
- Is there a correlation between customer engagement (votes/reviews) and restaurant success?
- How does pricing impact the likelihood of offering certain services?

The goal was to provide data-driven insights to help restaurants make informed decisions on pricing, service offerings, and customer engagement strategies.

2.2 Objectives

The key objectives of the project were to:

- **Analyze Popular Cuisines:** Identify the most common cuisines and calculate the percentage of restaurants offering them.
- **City-wise Distribution and Ratings:** Determine the city with the most restaurants, calculate average ratings, and find the city with the highest-rated restaurants.
- **Price Range Distribution:** Visualize and analyze the distribution of restaurants by price range and calculate the percentage in each category.
- **Correlation Between Pricing and Services:** Understand the relationship between price range and the likelihood of offering online delivery or table bookings.
- **Customer Engagement:** Analyze customer ratings and votes to determine the most common rating ranges and the factors contributing to higher engagement.
- **Cuisine Combination Impact:** Explore how offering multiple cuisines affects customer satisfaction and ratings.

Each objective was aimed at helping restaurants better understand market trends and improve their service offerings.

2.3 Methodology

The project followed a structured approach to analyze the restaurant dataset, uncovering patterns and relationships between key variables. Below is a summary of the methodology:

1. Data Collection and Preparation

The dataset, containing information on restaurant names, cuisine types, locations, price ranges, ratings, and services, was imported into **Google Colab**. I handled missing values by using median imputation and removed duplicates to ensure data accuracy. Standardization of variables, such as price range categories, was applied to prepare the data for analysis.

2. Exploratory Data Analysis (EDA)

I performed **descriptive statistics** and used visualizations like histograms, bar charts, and box plots to understand the distribution of key variables (price ranges, ratings, votes). This step revealed trends in restaurant pricing and service offerings, helping to frame further analysis.

3. Correlation Analysis

To study the relationships between variables, I computed the **Pearson correlation coefficient**, examining connections such as the relationship between price range and services (online delivery, table booking). **Heatmaps** and **scatter plots** created using **Seaborn** visually displayed the strength of these relationships, making it easier to interpret the findings.

4. Data Visualization

I used **Matplotlib** and **Seaborn** to create visual representations of the findings. **Bar charts** and **pie charts** highlighted the distribution of cuisines and services, while **scatter plots** depicted correlations between price ranges and service offerings. These visualizations effectively conveyed the insights to stakeholders.

5. Geographic and Sentiment Analysis

Using the restaurant's **latitude and longitude** data, I plotted their locations to identify clusters in different cities. Additionally, I conducted basic sentiment analysis on customer reviews to understand common themes in positive and negative feedback.

3. Tools & Technologies Used

3.1 Python and Pandas

I used Python extensively for scripting and analysis. **Pandas** was crucial for data manipulation, allowing me to clean, transform, and organize the dataset efficiently. The ability to perform data operations like grouping, filtering, and aggregation made pandas essential for the project.

3.2 NumPy

NumPy was used for numerical computations, handling arrays, and performing statistical calculations such as mean, median, and standard deviations. It complemented pandas in efficiently managing numeric data within the dataset.

3.3 Matplotlib & Seaborn

Both **Matplotlib** and **Seaborn** were used to visualize data. With Matplotlib, I created bar charts, histograms, and scatter plots to display data distributions. Seaborn, on the other hand, was used to generate correlation heatmaps, helping to illustrate the relationships between variables like price range and service offerings.

3.4 Google Colab

I utilized **Google Colab** as my development environment for writing and running Python code. Its cloud-based platform allowed me to work efficiently and collaborate with mentors and peers. Additionally, Colab's ability to integrate with Python libraries made it an ideal tool for this project.

4. Data Preparation

4.1 Data Collection

The dataset contained a variety of parameters related to restaurants, including cuisine types, city locations, price ranges, customer ratings, votes, and service offerings such as online delivery and table booking. The data provided a comprehensive overview of the restaurant industry, allowing for a wide range of analyses.

4.2 Data Cleaning and Preprocessing

To ensure the dataset was suitable for analysis, several preprocessing steps were required. Missing data points were handled by imputing median values where appropriate, and categorical variables such as cuisine types were encoded for further analysis. Duplicates were removed to avoid skewing the results. I also standardized certain fields to ensure consistency across variables.

4.3 Data Transformation

For specific tasks, transformations were applied to better understand the data. For example, I created new columns to group restaurants by city and price range, which allowed for easier analysis of location-based trends. These transformations were instrumental in uncovering meaningful insights.

5. Analysis & Findings

5.1 Cuisine Analysis

One of the tasks in the project was to analyze the most common cuisines offered by the restaurants in the dataset. By grouping the data by cuisine types and calculating the frequency of each cuisine, I was able to identify the top three most popular cuisines: North Indian, Chinese, and Fast Food. These cuisines made up a significant portion of the dataset, indicating their widespread popularity in the market.

I used pandas to group the data and create a frequency table for cuisines. A bar chart was created using Matplotlib to visualize the distribution, highlighting the dominance of these three cuisines.

Key Insight:

- North Indian, Chinese, and Fast Food accounted for over 50% of all cuisines in the dataset, showing their prevalence in the restaurant market.

5.2 City Analysis

Another key task was to analyze the number of restaurants in different cities. I grouped the data by city and calculated the number of restaurants in each city using pandas. The city with the highest number of restaurants was Delhi, followed by Mumbai and Bangalore. To further enrich this analysis, I calculated the average restaurant ratings for each city.

Using Seaborn, I generated a bar chart that displayed the number of restaurants per city, with an overlay of the average ratings. This gave a clear picture of the density of restaurants and their quality across different cities.

Key Insight:

- Delhi had the most restaurants, but Mumbai had the highest average restaurant ratings, indicating a potential focus on quality over quantity in Mumbai.

5.3 Price Range and Service Offerings

A significant part of the analysis focused on the relationship between price range and services such as online delivery and table bookings. I used correlation analysis to study how restaurant price ranges correlate with their service offerings.

- Price Range vs. Online Delivery: I found a negative correlation of -0.29, indicating that lower-priced restaurants are more likely to offer online delivery services.
- Price Range vs. Table Booking: There was a strong positive correlation of 0.93, showing that higher-priced restaurants tend to offer table booking services more frequently.

These findings were presented using heatmaps and scatter plots created with Seaborn, clearly illustrating the trends between price range and service offerings.

Key Insight:

- Higher-end restaurants are more likely to provide table booking services, while online delivery is predominantly offered by lower-priced establishments.

5.4 Restaurant Ratings and Votes

For this task, I analyzed the distribution of restaurant ratings and the average number of votes received by each restaurant. Using histograms and box plots, I visualized the distribution of aggregate ratings and discovered that most restaurants fall within the 3.5 to 4.0 rating range.

Additionally, I analyzed the number of customer votes for each restaurant. Restaurants with higher ratings generally had more votes, indicating a correlation between customer engagement and restaurant quality.

Key Insight:

- Restaurants in the 3.5 to 4.0 rating range were the most common, and those with higher ratings tended to receive more votes, highlighting the importance of customer satisfaction in driving engagement.

5.5 Cuisine Combinations and Their Ratings

An interesting part of the analysis was identifying common combinations of cuisines offered by restaurants and determining whether certain combinations were associated with higher ratings. By grouping the data on restaurants that

offered multiple cuisines, I was able to identify which cuisine combinations were most frequently offered together.

For this task, I created a combination matrix to identify the co-occurrence of cuisines such as **North Indian** and **Chinese**, and **Fast Food** with **South Indian**. Using **Seaborn**, I visualized the distribution of these cuisine combinations and compared their ratings.

Key Insight:

- Restaurants offering combinations of **North Indian** and **Chinese** cuisines tended to have higher average ratings than those offering single cuisines or other combinations, suggesting that diverse menu offerings can positively impact customer satisfaction.

5.6 Geographic Analysis of Restaurant Locations

The geographic distribution of restaurants was also a key component of the analysis. By using the **latitude** and **longitude** data from the dataset, I plotted the location of each restaurant on a map to identify any patterns or clusters of restaurants in specific areas.

Using **Seaborn** and **Matplotlib**, I created a scatter plot of restaurant locations. This visual analysis allowed me to spot clusters of high-rated restaurants in certain cities like **Mumbai** and **Delhi**, where there was a dense concentration of restaurants in upscale neighborhoods.

Key Insight:

- High-end restaurants (with higher price ranges and ratings) were often clustered in the central or affluent parts of cities, such as **South Mumbai** and **Central Delhi**, while lower-priced restaurants were spread more evenly across all regions.

5.7 Restaurant Chains and Their Popularity

I examined whether there were any popular restaurant chains in the dataset and analyzed their ratings and votes. By filtering out individual restaurants and grouping them based on chain affiliations, I was able to analyze the performance of restaurant chains in terms of both ratings and votes.

I found that well-known chains, such as **Domino's** and **KFC**, had higher numbers of votes but did not necessarily have the highest ratings. Conversely, smaller or local chains had fewer votes but higher customer ratings on average.

Key Insight:

- Popular international chains like **Domino's** and **KFC** attracted more customer engagement (votes), but local chains such as **Biryani Blues** received higher average ratings, suggesting a preference for locally-flavoured cuisine.

5.8 Restaurant Reviews: Positive and Negative Sentiment Analysis

To deepen the understanding of customer preferences, I conducted an analysis of customer reviews to identify common positive and negative keywords associated with restaurant ratings. Using basic natural language processing (NLP) techniques, I extracted frequent terms from customer reviews, identifying keywords like "delicious," "fast service," "affordable," as positive terms, and "slow," "expensive," "poor hygiene," as negative terms.

This analysis was complemented by word clouds and sentiment analysis to visualize the distribution of positive and negative sentiments.

Key Insight:

- Positive reviews frequently mentioned "delicious" and "fast service," while negative reviews often cited "slow service" and "expensive" as key complaints. This suggests that service speed and pricing are crucial factors in shaping customer satisfaction.

5.9 Votes Analysis: Correlation Between Votes and Ratings

One of the final tasks was to analyze the correlation between the number of votes a restaurant received and its overall rating. I used **pandas** to group restaurants by the number of votes they received and then calculated the average rating for each group.

Scatter plots and regression analysis showed a positive correlation between votes and ratings, indicating that highly rated restaurants tend to receive more customer engagement in the form of votes.

Key Insight:

- Restaurants with higher ratings (above **4.0**) consistently received more votes, suggesting that customer satisfaction drives greater engagement and word-of-mouth recommendations.

5.10 Price Range vs. Online Delivery and Table Booking

Expanding on the earlier correlation analysis, I examined the relationship between the price range of restaurants and their likelihood of offering online delivery and table booking services. Using **pandas** and **Seaborn**, I performed a

correlation analysis and visualized the findings through heatmaps and scatter plots.

As discussed earlier, higher-priced restaurants are more likely to offer table booking services, while lower-priced restaurants frequently provide online delivery. This dual-service offering indicates that restaurants adjust their service models based on their pricing strategy to cater to different customer segments.

Key Insight:

- The correlation analysis revealed a **strong positive correlation** (0.93) between higher prices and table bookings, and a **negative correlation** (-0.29) between higher prices and online delivery, reinforcing the idea that luxury restaurants focus more on exclusive, in-person experiences, while lower-priced restaurants emphasize convenience through delivery services.

6. Challenges & Solutions

I encountered several challenges that required problem-solving and creative approaches to ensure accurate analysis. Below are the key challenges and the solutions implemented:

6.1 Data Quality Issues

One of the main challenges was dealing with missing values and inconsistent data across several important fields, such as customer ratings and service offerings. These gaps in the dataset had the potential to skew the analysis if not handled properly.

Solution:
I addressed missing values by using **median imputation** for numeric fields like ratings and votes, and **mode imputation** for categorical variables like cuisine type. Inconsistent entries were standardized through data preprocessing in **pandas**, ensuring uniformity across the dataset.

6.2 Large Dataset Management

Managing a large dataset with multiple attributes and categories posed challenges in terms of memory and computation, especially when running more complex calculations like correlation analysis.

Solution:
I optimized data operations by using **pandas** for efficient data handling and **NumPy** for numerical computations. By breaking down complex operations

into smaller, manageable steps, I ensured that the analysis was performed efficiently without overwhelming system resources.

6.3 Visualization Complexity

Another challenge was representing complex data relationships in a clear and visually effective way. Certain analyses, such as correlation between multiple variables or cuisine combinations, required advanced visualizations.

Solution:
I leveraged **Matplotlib** and **Seaborn** to create advanced visualizations like **heatmaps** and **scatter plots**, which made it easier to represent correlations and patterns. I also used **Seaborn's** intuitive plotting functions to create clean, insightful visuals that conveyed the data trends clearly.

7. Reflections & Learnings

7.1 Technical Skills

During this internship, I enhanced my technical skills, particularly in data analysis using tools like **Python**, **pandas**, and **Matplotlib**. I learned how to clean and manipulate datasets effectively, handle missing data, and perform correlation analysis to uncover key insights. I also gained experience with **Seaborn**, which improved my ability to create clear and meaningful visualizations.

Beyond these tools, the internship gave me deeper insight into how data analysis is applied in real-world business contexts, allowing me to bridge the gap between academic learning and practical application.

7.2 Professional Development

In addition to the technical skills, I developed valuable soft skills, particularly in problem-solving, time management, and communication. Managing large datasets required a keen eye for detail and the ability to troubleshoot issues quickly. Presenting my findings to team members also helped me develop skills in communicating complex data analysis results clearly and effectively.

This experience taught me the importance of working efficiently under deadlines while ensuring accuracy, which will be invaluable in future projects and roles.

7.3 Problem-Solving and Adaptability

Throughout the project, I encountered challenges that required flexibility and adaptability, such as dealing with inconsistent data and creating clear

visualizations for complex relationships. These problem-solving opportunities taught me to think critically, experiment with different approaches, and ultimately find the best solutions for the task at hand.

8. Conclusion & Recommendations

The analysis performed during the internship provided valuable insights into the restaurant industry. The project revealed important trends and patterns in customer preferences, restaurant pricing, and service offerings, all of which can help businesses optimize their operations and better serve their customers.

8.1 Summary of Findings

Key findings from the analysis include:

Popular Cuisines: North Indian, Chinese, and Fast Food are the most commonly offered cuisines.

City Analysis: Delhi has the highest number of restaurants, while Mumbai boasts higher average ratings, suggesting a focus on quality in that region.

Price Range & Services: Higher-priced restaurants are more likely to offer table bookings, while lower-priced restaurants are more inclined to provide online delivery services.

Customer Engagement: Restaurants with higher ratings tend to receive more votes, indicating a strong correlation between customer satisfaction and engagement.

8.2 Recommendations

Based on the findings, I suggest the following recommendations:

High-End Restaurants: Focus on providing table booking services and enhancing the in-restaurant experience, as higher-priced restaurants are expected to offer more premium services.

Budget Restaurants: Emphasize online delivery services to attract a broader customer base, particularly those seeking convenience.

Cuisine Diversity: Restaurants offering multiple cuisines, especially popular combinations like North Indian and Chinese, may enhance customer satisfaction and ratings.

Encouraging Feedback: Encourage customer engagement through reviews and votes, as this can boost the restaurant's visibility and reputation.

8.3 Future Work

For future research, I recommend exploring additional variables, such as customer demographics, that may influence restaurant performance. Furthermore, using machine learning models to predict trends in restaurant ratings, pricing, and service preferences could provide more advanced insights to drive business strategy.

Cognifyz
Where Data Meets Intelligence

Cognifyz Technologies

Internship Completion Certificate

Date - 14/06/2024

This is to certify that **Keshwanth G P, (Intern ID: CT1/A1/C32510)**, currently pursuing a M.Sc. Applied Data Science from The SRM University, was working as a **Data Analysis Intern** with Cognifyz Technologies from May 2024 - June 2024.

During this period, he has served as a Data Analysis Intern and has displayed remarkable dedication, sincerity, and a strong desire to learn. He has exhibited exceptional coordination skills and effective communication abilities. Moreover, his attention to detail has been truly impressive.

he has consistently approached new assignments and challenges with enthusiasm, showcasing his passion for Data Analysis. His commitment and willingness to acquire new knowledge and skills have been evident throughout his internship.

We extend our best wishes to Keshwanth G P for a successful future, and we have no doubt that he will continue to excel in the field of Data Analysis.

With Regards,
Cognifyz Technologies

Cognifyz Technologies
Cognifyz Technologies
Nagpur, India

AICTE
All India Council for Technical Education

MSME
MICRO, SMALL & MEDIUM ENTERPRISES

cognifyztechnologies@gmail.com
www.cognifyz.com