**SIG743 End Term Assessment**

|  |  |
| --- | --- |
| Unit Code & Name | SIG742 - Modern Data Science |
| Assignment | End Term |
| Submission Date | October 05, 2025 |
| Team Members | Ankur Luthra (Student Id: **S225697234**)  Frederick Abbey (Student Id :)  Mmusi Dithotse (Student Id :) |

Table of Contents

[Background 4](#_Toc210491033)

[1. Load and explore the dataset 4](#_Toc210491034)

[Data Overview 5](#_Toc210491035)

[Structure of the data 5](#_Toc210491036)

[Missing Value Check 5](#_Toc210491037)

[Duplicate Values 5](#_Toc210491038)

[Invalid Values 5](#_Toc210491039)

[Statistical Summary 5](#_Toc210491040)

[Part 1 5](#_Toc210491041)

[Answer 1.1 5](#_Toc210491042)

[Answer 1.2 5](#_Toc210491043)

[Answer 1.3 5](#_Toc210491044)

[Answer 1.4 5](#_Toc210491045)

[Answer 1.5 5](#_Toc210491046)

[Answer 1.6 6](#_Toc210491047)

[Answer 1.6.1 6](#_Toc210491048)

[Answer 1.6.2 6](#_Toc210491049)

[Answer 1.7 6](#_Toc210491050)

[Answer 1.7.1 6](#_Toc210491051)

[Answer 1.7.2 6](#_Toc210491052)

[Answer 1.8 6](#_Toc210491053)

[Answer 1.8.1 6](#_Toc210491054)

[Answer 1.8.2 6](#_Toc210491055)

[Answer 1.8.3 6](#_Toc210491056)

[Part 2 6](#_Toc210491057)

[Answer 2.1 6](#_Toc210491058)

[Answer 2.2 6](#_Toc210491059)

[Answer 2.3 6](#_Toc210491060)

**Table of Figures**

[Figure 1 Structure of Data 5](#_Toc210491002)

# Background

The project focuses on analysing businesses and their customer reviews from Google Maps. The business dataset captures attributes such as location, category, ratings, and operational details, providing a structural overview of the business landscape. The review dataset contains user-generated ratings, text, and timestamps, enabling exploration of temporal trends in customer engagement and satisfaction. By linking businesses and reviews via unique identifiers, the datasets support forecasting future review activity, identifying patterns and seasonality, and applying advanced predictive models such as ARIMA, SARIMA, and LSTM for time series analysis.

# Load and explore the dataset

**Shape of the Data:**

* **Business dataset**: 12,774 rows × 15 columns
* **Review dataset**: 521,515 rows × 8 columns

**Business Dataset Columns:**

* name: Business name
* address: Business address
* gmap\_id: Unique Google Maps identifier
* description: Text description of the business
* latitude / longitude: Geolocation coordinates
* category: Business type/category
* avg\_rating: Average rating of the business
* num\_of\_reviews: Total number of reviews
* price: Price range of the business
* hours: Business operating hours
* MISC: Miscellaneous notes
* state: State or region of the business
* relative\_results: Related business results
* url: Google Maps URL for the business

**Review Dataset Columns:**

* user\_id: Unique identifier of the reviewer
* name: Name of the business reviewed
* time: Timestamp of the review
* rating: Rating given by the user
* text: Review text content
* pics: Photos uploaded with the review
* resp: Business responses to reviews
* gmap\_id: Unique Google Maps identifier linking to business

## Data Overview

### Structure of the data

|  |  |
| --- | --- |
| Figure 1 Structure of Data |  |

### Missing Value Check

### Duplicate Values

### Invalid Values

### Statistical Summary

# Part 1

## Answer 1.1

## Answer 1.2

## Answer 1.3

## Answer 1.4

## Answer 1.5

## Answer 1.6

### Answer 1.6.1

### Answer 1.6.2

## Answer 1.7

### Answer 1.7.1

### Answer 1.7.2

## Answer 1.8

### Answer 1.8.1

### Answer 1.8.2

### Answer 1.8.3

# Part 2

## Answer 2.1

## Answer 2.2

## Answer 2.3