**Group Project**

**Insights and Analysis on Hospitality expenses in Canada**

Group number- 2

Aayushi Singhvi

Deepak Dalal

Girika Jawrani

Harshrajsinh Chavda

Sanjana Upender

Suyash Srivastav

Business Analytics, Conestoga College

INFO8076 – 23F – Sec3- SQL and Data Analysis

Professor: Jey Kumaresan

Assignment Due Date: December 11th, 2023

Table of Contents

|  |  |
| --- | --- |
| **Description** | **Page No.** |
| Introduction | 3 |
| Goal of the project | 3 |
| Data source and cleaning | 3 |
| Relational schema/ERD | 5 |
| SQL Queries | 7 |
| Conclusion | 22 |

**Introduction**

A large volume of data is available at free hand to a lot of individuals and organisations, which is an essential asset for them to gain benefit by forecasting trends, taking strategical decisions, etc. Organising and structuring the data is an important step in this process. This project focuses on creating a relational database to effectively manage and analyze data on hospitality expenses of of Ontario, Canada. The guidelines and principles regulating hospitality are delineated in the Policies for Ministers' Offices and the Treasury Board’s Directive on Travel, Hospitality, Conference, and Event Expenditures. This dataset amalgamates all Hospitality Expense reports submitted by federal institutions.

The data set has information about various hospitality events and their types organised from a period over 20 years rom 2003 to 2023.

**Goal of the project**

The main goal of this project is to understand procurement and organization of data, its analysis and working with SQL. This project will explore the design, implementation, and utilization of a relational database system. It will cover the entire data management lifecycle, including data source identification, schema design, data population, SQL query application, and the creation of user-friendly views for end users.

The objective of this project is to acquire insights and conduct an analysis of hospitality events in Canada as reported by federal institutions. SQL is employed to populate tables from the dataset, utilizing an Entity-Relationship Diagram (ERD) for database modeling. Subsequently, SQL queries are executed to generate outputs for the analysis of expenditure, event type, event date, and to assess how the organization of events has evolved in the post-COVID era compared to the pre-COVID era. The ultimate aim is to present these insights to the Finance Minister of Canada.

**Data source and cleaning**

* *Data source and acquisition:*

This data set was fetched https://open.canada.ca and is called Proactive Disclosure - Hospitality Expenses and the publisher of the same is Treasury Board of Canada Secretariat.

* *Data format:*

The data set was available in the .csv format and was edited in the same format to later be exported to pgadmin4 app.

* *Data cleaning:*

The data set was cleaned in Excel using basic functions and power query tool to remove errors. Originally there were 20 columns in the set namely:

ref\_number, disclosure\_group, title\_en, title\_fr name, description\_en description\_fr, start\_date, end\_date, location\_en, location\_fr, vendor\_en, vendor\_fr, employee\_attendees, guest\_attendees, total, additional\_comments\_en, additional\_comments\_fr, owner\_org and owner\_org\_title.

The columns titled title\_fr, description\_fr, location\_fr, and vendor\_fr were excluded from the dataset as they contained French characters and were essentially translations of their respective English-language counterparts.

Eliminated null values in the "disclosure\_group" column to ensure the proper establishment of hierarchy and table relationships.

* *Data transformation:*

The “vendors” column and the column for “additional comments” both had over a thousand blank entries, posing a potential challenge for analysis. In order to smoothly run queries without subsequent errors caused by these null values, these two columns were voluntarily dropped before creating tables.

We opted to modify the designation from SLE to SOE. While the website indicates that SLE stands for Senior Officer or Employees, we believe the abbreviation SOE serves the purpose more effectively.

Deleted wrong datatype entries from location\_en column.

**Relational Schema/ERD**

**Description of tables**: Out of the given dataset, 4 tables were populated, which are as follows:

1. Table name: **event\_details**

Table purpose: This table was created to store all information related to the hospitality events.

Table columns: there are 9 columns in this table-

* event\_id (Primary key): A unique identifier for each event.
* ref\_number: This column has the reference number for each corresponding event.
* officer\_title: this column has information come out the officer and his title who conducted the event.
* event\_description: This column has all the information about the type of event and its description.
* start\_date: The date when the events began.
* end\_date: The date when the event ended.
* event\_location: The location or the venue where the event was organised or will the organised.
* total\_cost: The over all cost in conducting and organising the events.
* disclosure\_id (Foreign key): Identity number for the disclosure group.

1. Table name: **disclosure\_group\_details**

Table purpose: This table provides a centralized repository for details about disclosure groups, facilitating a structured classification system.

Table columns: there are 3 columns in this table-

* Disclosure\_id: Identity number for the disclosure group.
* Disclosure\_group (Primary key): A short name or code representing the disclosure group.
* Disclosure\_group\_full\_name: This column stores the full names for each disclosure group.

1. Table name: **attendees\_details**

Table purpose: this table stores all information related to the attendees of the events.

Table columns: there are 5 columns in this table-

* attendees\_register (Primary key): Unique identification number for each attendee.
* event\_officer: this column stores information about officers who conducted the events.
* employee\_attendees: This column has information about number of employees who attended the events.
* guest\_attendees: this column saves information about number of guests who attended the events.
* event\_id: Unique identification number corresponding to each event.

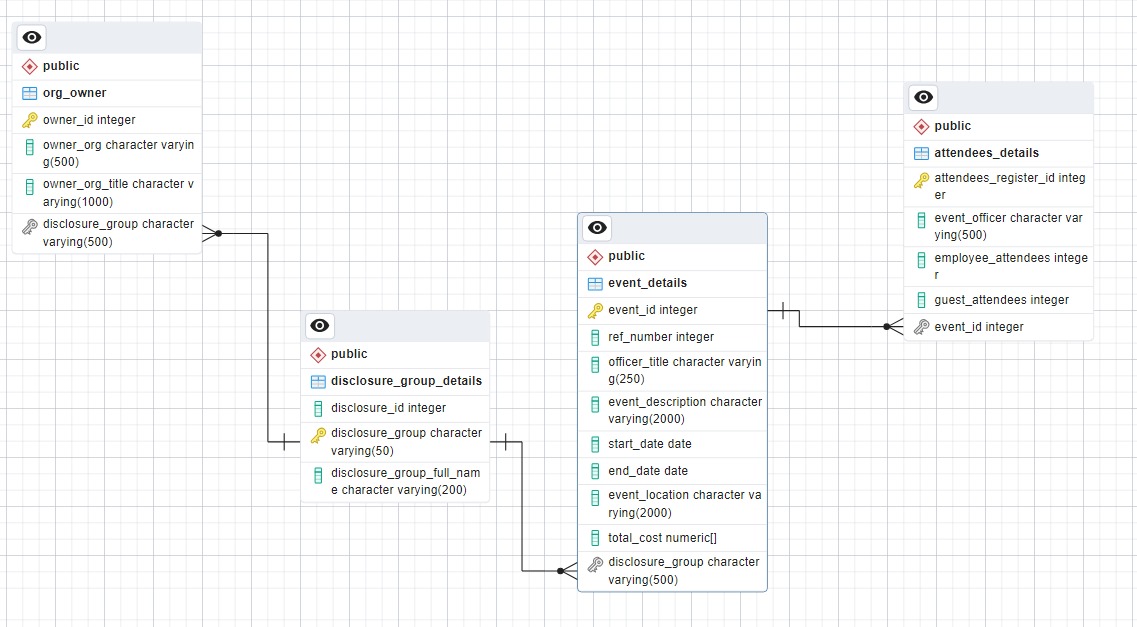
1. Table name: **org\_owner**

Table purpose: this table sores information about organisations who conducted the hospitality events.

Table columns: there are 4 columns in this table-

* owner\_id (Primary Key): A unique identifier for each owner.
* owner\_org: this column has values about the owner of each organization.
* owner\_org\_title: this column has information about the title of the organization.
* disclosure\_id (Foreign key): Identity number for the disclosure group.

**Entity Relationship Diagram ERD**



ERD diagram of the four tables

**Relationhips:**

The ERD script establishes several relationships between the tables in the relational database. The "event\_details" table serves as a central entity, relating to other tables through foreign key references. The "disclosure\_id" field in both the "event\_details" and "org\_owner" tables forms a connection with the "disclosure\_group\_details" table, indicating a relationship between events, organizational owners, and disclosure groups. Additionally, the "attendees\_details" table is linked to the "event\_details" table through the "event\_id" foreign key, representing a connection between event details and attendee information. These relationships ensure data consistency and integrity, allowing for a structured representation of events, organizational ownership, disclosure groups, and attendee details within the database.

**SQL QUERIES**

1. **Question:**

Find year wise highest total cost of event after 2017.

**Query:**

SELECT

EXTRACT(YEAR FROM e.start\_date) AS event\_year,

MAX(e.total\_cost) AS max\_total\_cost

FROM

event\_details e

GROUP BY

event\_year

HAVING

EXTRACT(YEAR FROM e.start\_date) > 2017

ORDER BY

event\_year;

**Output table:**

|  |  |
| --- | --- |
| event\_year | max\_total\_cost |
| 2018 | 4633.89 |
| 2019 | 18368.25 |
| 2020 | 11467.12 |
| 2021 | 2071.88 |
| 2022 | 7456.02 |
| 2023 | 8980.71 |

**Insights:**

This query returns the total cost in each year with the focus on filtering latest or last 5 years. This shows 2019, is the most expensive year out of all, whereas the costs really dropped during 2021 due to very less events taking place and finally rising the cost of events gradually every year showing the progress or more and more events taking place after everything went normal.

1. **Question:**

Provide breakdown of the number of events and their total costs, categorized into pre-COVID and post-COVID periods to understand the impact of the COVID timeline on both the frequency and financial aspects of our events?

**Query:**

WITH event\_time\_buckets

AS

(SELECT event\_id, total\_cost,

CASE

WHEN start\_date < '2020-03-01' THEN 'pre-covid'

WHEN start\_date >= '2020-04-01' AND start\_date <= '2022-04-30' THEN 'during-covid' WHEN start\_date > '2022-05-01' THEN 'post-covid'

ELSE NULL

END AS time\_bucket

FROM event\_details )

SELECT time\_bucket,

COUNT(event\_id) AS event\_count,

ROUND(SUM(CAST(total\_cost AS numeric)), 2)

AS total\_cost\_in\_bucket FROM event\_time\_buckets

WHERE time\_bucket IS NOT NULL

GROUP BY time\_bucket

ORDER BY time\_bucket;

**Output table:**

|  |  |  |
| --- | --- | --- |
| time\_bucket | event\_count | total\_cost\_in\_bucket |
| during-covid | 18 | 8156.50 |
| post-covid | 62 | 91050.87 |
| pre-covid | 1228 | 425967.64 |

**Insights:**

The data reveals a substantial pre-covid event activity with high associated costs, followed by a pandemic-induced dip in event count and costs during covid. Post-covid, there's a notable resurgence in both events and financial investment, indicating a potential recovery or adaptation to the new normal.

Pre-Covid Insights (time\_bucket: pre-covid):

Event Count: There are a significant number of events (1228) categorized as pre-covid, indicating a robust level of event activity before the Covid-19 pandemic.

Total Cost: The total cost of events in the pre-covid period is substantial (425967.64), suggesting that a considerable financial investment was made in organizing these events.

During Covid Insights (time\_bucket: during-covid):

Event Count: The count of events during the Covid-19 pandemic is lower (18) compared to the pre-covid period. This could be attributed to restrictions and safety concerns during the pandemic.

Total Cost: The total cost during the Covid-19 pandemic is comparatively lower (8156.5), indicating potential budget constraints or changes in event scale or format to adapt to pandemic conditions.

Post-Covid Insights (time\_bucket: post-covid):

Event Count: The number of events post-covid is higher (62) compared to the during-covid period, suggesting a potential rebound or adaptation to the new normal.

Total Cost: The total cost of events post-covid is the highest (91050.87), indicating a notable financial investment and potential recovery or growth in event spending post-pandemic.

1. **Question:**

Events with high employee attendance (where the count of attendees is more than 400) that have already taken place?

**Query:**

SELECT ed.event\_description, ad.employee\_attendees

FROM event\_details ed

JOIN attendees\_details ad ON ed.event\_id = ad.event\_id

WHERE ed.start\_date < CURRENT\_DATE

GROUP BY ed.event\_description, ad.employee\_attendees

HAVING SUM(ad.employee\_attendees) > 400;

**Output table:**

|  |  |  |
| --- | --- | --- |
| event\_description | event\_location | employee\_attendees |
| Refreshments muffins Financial and Administrative Services Staff meeting | Marche Marcello Phase IV Gatineau | 600 |
| Meals refreshments All staff meeting | Marche frais Escale Restaurant Phase IV Gatineau | 740 |
| Morning refreshments for a staff meeting | Costco Ottawa | 1791 |
| Blueprint 2020 | Ottawa ON | 809 |
| Morning and afternoon refreshments for an all-staff retreat | Palais des Congrs Gatineau | 471 |
| National Public Service Week Barbecue | "Moussette Park | 500 |
| Refreshments during the 5th Anniversary Celebration of Service Canada | 361 Boulevard de Lucerne | 525 |
| Refreshments Reception for the Innovation Information and Technology Branch employees for the National Public Service Week | Gatineau Quebec J9H 7K9" | 500 |

**Insights:**

The query retrieves events with high employee attendance (more than 400 employees). Here we are using only the past data to infer this because we just want to analyse the events that have already occurred in the past. Costco Ottawa has a large-scale gathering with attendance close to 1800 employees. With this information in place the specific Department can plan optimized event planning based on location-specific requirements in the future.

1. **Question:**

Generate a Quarterly Trend Report for events, including the number of events and total costs per quarter for the years 2016 and 2022, categorizing the events by quarter (Q1, Q2, Q3, Q4) for understanding the distribution of events and associated costs across quarters in the specified years.

**Query:**

SELECT EXTRACT(YEAR FROM start\_date) AS event\_year, CASE WHEN EXTRACT(MONTH FROM start\_date) BETWEEN 1 AND 3 THEN 'Q1' WHEN EXTRACT(MONTH FROM start\_date) BETWEEN 4 AND 6 THEN 'Q2' WHEN EXTRACT(MONTH FROM start\_date) BETWEEN 7 AND 9 THEN 'Q3' WHEN EXTRACT(MONTH FROM start\_date) BETWEEN 10 AND 12 THEN 'Q4' END AS event\_quarter, COUNT(event\_id) AS num\_events, ROUND(SUM(CAST(total\_cost AS numeric)), 2) AS total\_costFROM public.event\_detailsWHERE EXTRACT(YEAR FROM start\_date) IN (2016, 2022)GROUP BY event\_year, event\_quarterORDER BY event\_year, event\_quarter;

**Output table:**

|  |  |  |  |
| --- | --- | --- | --- |
| event\_year | event\_quarter | num\_events | total\_cost |
| 2016 | Q1 | 6 | 1610.43 |
| 2016 | Q2 | 10 | 4290.98 |
| 2016 | Q3 | 14 | 8413.48 |
| 2016 | Q4 | 27 | 11810.49 |
| 2022 | Q1 | 4 | 1045.07 |
| 2022 | Q2 | 9 | 10898.54 |
| 2022 | Q3 | 5 | 9238.73 |
| 2022 | Q4 | 10 | 22448.55 |

**Insights:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sum of num\_events** | **Column Labels** | |  |
| **Row Labels** | **2016** | **2022** | **Grand Total** |
| Q1 | 6 | 4 | 10 |
| Q2 | 10 | 9 | 19 |
| Q3 | 14 | 5 | 19 |
| Q4 | 27 | 10 | 37 |
| **Grand Total** | **57** | **28** | **85** |

This is a quarterly breakdown of events for the years 2016 and 2022, along with a grand total. Quarter 4 (Q4) in both years contributed significantly to the overall event count, indicating a potential seasonality or increased activity during that period. Overall, 85 events were held, with 57 occurring in 2016 and 28 in 2022, showcasing a notable difference in event frequency between the two years with a gap of 5 years.

1. **Question:**

List all the event locations with ‘high and very high’ cost categories.

**Query:**

WITH CTE as (

SELECT event\_location,

(case when total\_cost between 1 and 4999 then 'low'

when total\_cost between 5000 and 9999 then 'medium'

when total\_cost between 10000 and 14999 then 'high'

when total\_cost between 15000 and 20000 then 'very high'

End) as cost\_category

from event\_details)

select event\_location, cost\_category

from CTE

where cost\_category like '%high'

order by cost\_category desc

**Output table:**

|  |  |
| --- | --- |
| event\_location | cost\_category |
| Gatineau Quebec | very high |
| Wilson House Chelsea QC | high |
| Victoria British Columbia | high |

**Insights:**

We have categorized total costs for events into 4 categories ‘very high, high, medium, and low’. The above query displays the event locations falling into ‘very high and high’ categories. The top 3 locations contribute 8% of the total costs. We can dive further into what’s causing this hike in costs or prefer not to hold events at the mentioned places.

1. **Question:**

Give details of events with no attendees which cost us more than 50 dollars along with officer names.

**Query:**

SELECT ed.event\_id, ed.total\_cost, ad.event\_officer\_name, round (avg (ad.guest\_attendees))as gst\_attd, round (avg (ad.employee\_attendees))as emp\_attd

FROM public.event\_details ed

JOIN public.attendees\_details ad ON ed.event\_id = ad.event\_id

WHERE ad.guest\_attendees = 0 and ad.employee\_attendees = 0 and ed.total\_cost > 50

group by ed.event\_id, ad.event\_officer\_name

**Output table:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| event\_id | total\_cost | event\_officer\_name | gst\_attd | emp\_attd |
| 93 | 5512.5 | Alain Beaudet | 0 | 0 |
| 101 | 287.11 | Laffin Angeline | 0 | 0 |
| 107 | 85 | Buist Margaret | 0 | 0 |
| 151 | 269.8 | Derrick Cheung | 0 | 0 |
| 171 | 80.82 | Nick Leswick | 0 | 0 |
| 672 | 293.2 | Vats Nipun | 0 | 0 |
| 694 | 138 | Rear Admiral Art McDonald | 0 | 0 |
| 1107 | 129.82 | Jaton Patricia | 0 | 0 |

**Insights:**

This Query displays the events where there were zero attendees’ guests and employees, yet we have incurred expenses on these events, along with the event\_id, and officer names responsible for such events to avoid the same in the future and to create accountability. We have spent nearly 7000 dollars which could be utilized more efficiently.

1. **Question:**

Provide figures for cost and events based on disclosure groups.

**Query:**

SELECT disclosure\_id, disclosure\_group, event\_count, round (total\_cost),

RANK() OVER (ORDER BY total\_cost DESC) AS cost\_rank

FROM (SELECT dg.disclosure\_id, dg.disclosure\_group,

COUNT(ed.event\_id) AS event\_count,

SUM(ed.total\_cost) AS total\_cost

FROM public.disclosure\_group\_details dg

LEFT JOIN public.event\_details ed ON dg.disclosure\_id = ed.disclosure\_id

GROUP BY dg.disclosure\_id, dg.disclosure\_group

) AS grouped\_data;

**Output table:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| disclosure\_id | disclosure\_group | event\_count | total\_cost | cost\_rank |
| 1 | SOE | 982 | 488925 | 1 |
| 2 | MPSES | 330 | 43789 | 2 |

**Insights:**

SOE ranks higher amongst the two, accommodating a higher number of events held, and consequentially higher costs highlighting the area of work. The aim is to provide insights into the relationship between disclosure groups, and the events they're associated with, particularly focusing on the total cost of those events providing insight into potential resource allocation or investment patterns and comparing the financial impact of different disclosure groups based on their events.

1. **Question:** Find the details of the events held by 'atssc scdata’ organization in the past 6 months

**Query:**

SELECT ed.event\_id, ed.event\_description, ed.start\_date, ed.end\_date

FROM public.event\_details ed

JOIN public.org\_owner oo ON ed.disclosure\_id = oo.disclosure\_id

WHERE oo.owner\_org = 'atssc scdata'

AND ed.start\_date >= CURRENT\_DATE - INTERVAL '6 months'

AND ed.start\_date <= CURRENT\_DATE

**Output table:**

|  |  |  |  |
| --- | --- | --- | --- |
| event\_id | event\_description | start\_date | end\_date |
| 408 | Gift Exchange with the Alexis Nakota Sioux First Nation | 2023-08-11 | 2023-08-11 |
| 409 | Dinner Meeting with stakeholders | 2023-08-14 | 2023-08-14 |

**Insights:**

The query fetches details of events, including event IDs, descriptions, start dates, and end dates of one organization namely 'atssc scdata', for the past 6 months.

This query enables tracking and management of recent events organized by 'atssc scdata' within the past 6 months. It can assist in analytics, planning, or monitoring its activities, helping us determine the upcoming needs of any particular organization as well as their contribution to events, their impact, and corresponding cost to us.

1. **Question:**

Display the details of the events where the total cost of the event is more than 15 times the average cost of all events.

**Query:**

SELECT event\_id,event\_location,total\_cost,disclosure\_id,officer\_title

FROM public.event\_details

WHERE total\_cost >= (

SELECT 15 \* AVG(total\_cost)

FROM public.event\_details);

**Output table:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| event\_id | event\_location | total\_cost | disclosure\_id | officer\_title |
| 3 | Toronto Ontario Canada | 7456.02 | 1 | Chief Executive Officer |
| 94 | Montreal Quebec Canada | 8980.71 | 1 | President and CEO |
| 117 | Toronto Ontario Canada | 7376.95 | 1 | Vice President |
| 243 | Toronto Ontario Canada | 6095.98 | 1 | President and Chief Executive Officer |
| 360 | Wilson House Chelsea QC | 10978.41 | 1 | General Director Financial Sector Policy Branch |
| 429 | Victoria British Columbia | 11467.12 | 1 | Deputy Minister of Justice and Deputy Attorney General of Canada |
| 432 | Gatineau Quebec | 18368.25 | 1 | Assistant Deputy Minister Aboriginal Affairs |
| 706 | Ottawa Conference and Event Centre Ottawa | 8567.93 | 1 | Executive Director |

**Insights:**

This query will return event details (event ID, location, cost, disclosure ID, officer title) for events that meet the specified condition of having a total cost of at least 15 times the average cost of all events.

This query aims to identify events that significantly exceed the average cost by a considerable factor (15 times higher) and might indicate outliers or events with high costs compared to others in the dataset.

In our case, the biggest outlier could be event 429, held at Victoria British Columbia costing 11,467 dollars.

1. **Question:**

Find Event Officers with the Most Attendees (Employees and Guests Combined):

**Query**:

WITH OfficerAttendeeTotals AS (

SELECT

event\_officer\_name,

SUM(employee\_attendees) + SUM(guest\_attendees) AS total\_attendees

FROM

attendees\_details

GROUP BY

event\_officer\_name

HAVING

SUM(employee\_attendees) + SUM(guest\_attendees) > 500

)

SELECT event\_officer\_name, total\_attendees FROM OfficerAttendeeTotals

ORDER BY total\_attendees DESC

**Output table:**

|  |  |
| --- | --- |
| event\_officer\_name | total\_attendees |
| James Moore | 1791 |
| Josee Duplessis | 809 |
| Goodyear Gary | 740 |
| Abdul Jalil | 600 |
| VAdm Hawco D C | 525 |

**Insights:**

This query gives a list of the top 10 event officers who could gather the most attendees in their events. It sums up both the employees and guests who attended and shows the highest total attendees for officers. It shows can James has outshined every other officer by gathering almost double the attendees for his events as compared to all other officers.

**Conclusion:**

The project on proactive disclosure of hospitality expenses involved the use of SQL queries to demonstrate cost-cutting measures. By leveraging the available data on government travel and hospitality expenses, the project aimed to proactively showcase efforts to reduce costs and enhance transparency. The use of SQL queries allowed for the extraction and analysis of relevant information, enabling the identification of areas where cost-saving initiatives were implemented.

The proactive disclosure of hospitality expenses is a crucial aspect of government transparency and accountability. Through the utilization of SQL queries, the project was able to present a comprehensive overview of hospitality expenditures, thereby fulfilling the government's commitment to proactive information disclosure. The data obtained through these queries not only highlighted the costs associated with hospitality but also demonstrated the efforts undertaken to achieve savings and ensure responsible expenditure.

Furthermore, the project's focus on cost-cutting measures aligns with the government's emphasis on fiscal responsibility and efficient resource utilization. By leveraging SQL queries to analyze hospitality expenses, the project contributed to a culture of transparency and prudent financial management within the public sector. The insights derived from the SQL queries provided a clear demonstration of the government's dedication to optimizing hospitality expenditures and implementing cost-effective measures.

In conclusion, the project on proactive disclosure of hospitality expenses, which utilized SQL queries to showcase cost-cutting measures, exemplifies the government's commitment to transparency, accountability, and responsible financial stewardship. By leveraging data-driven insights, the project effectively demonstrated the efforts to reduce costs and enhance efficiency within the realm of hospitality expenditures, thereby contributing to a culture of informed decision-making and prudent resource management.