**SQL AND DATA ANALYTICS**

**INFO8076-Fall 2023-Sec3**

**GROUP PROJECT**

**Member’s Name**

**Vats Sanghvi**

**Siddhesh Otari**

**Samuel**

**Rishikesh Raut**

**Arpit Patel**

**Saksham Magoo**

**Abhishek Mishra**

**Introduction**

This SQL project is to extract and organize information about sports facilities in Canada, specifically focusing on their locations. This involves writing SQL queries to extract data from a database that contains information about the different sports facilities in Canada, including their names, addresses, and geographic coordinates. The extracted data can then be used to create maps and visualizations that depict the locations of sports facilities across the province. By applying SQL queries to this data, the project can provide valuable insights into the distribution and accessibility of sports facilities in Canada, which can inform urban planning and development, sports tourism, and community engagement initiatives.

**Goal of the project:**

The goal of the project is to showcase data related to sports facilities in Canada. This involves gathering information about the different sports facilities locations across the country. The data could be presented in various formats, such as tables, charts, and maps, to help users easily understand and analyze the information. I have used tables to present the data. The project is useful for individuals looking to find sports facilities in Canada, as well as for organizations seeking to evaluate the availability and accessibility of such facilities. Additionally, the project could provide insights into potential gaps in sports facility provision and inform policy decisions related to sports infrastructure development.

**Dataset**

The dataset contains information related to Proactive Travel Disclosures by federal Institutions. The columns in different tables are defined as follows:

Doc\_Info

1. ref\_id (Primary Key) – A ref\_id is a unique identifier assigned to a doc\_info.
2. Disclosure\_group.
3. Title.
4. Name.
5. Purpose.

Travel\_Info

1. info\_id (Primary Key) – Travel information can be fetched from info\_id.
2. Destination.
3. Start\_Date.
4. End\_Date.
5. Ref\_id (Foreign Key)

Cost

1. Cost\_ID (Primary Key) – All the cost related details for every group can be retrieved .
2. Airfare.
3. Other\_Transport.
4. Lodging.
5. Meals.
6. Other\_Expenses.
7. Total.
8. Ref\_id ( Foreign Key ).

Organization

1. Org\_ID (Primary Key) .
2. Org\_Owner\_Title.
3. Org\_Owner.
4. Ref\_id ( Foreign Key ).

**Entity Relationship Diagram or Schema:**

A screenshot of a graph

Description automatically generated

**Creation of the database**

1. Right Click "Databases" >"Create" > "Database"

A screenshot of a computer

Description automatically generated

1. In the "Create - Database" dialog box, insert a name for database in the " Travel Expenses".
2. Click the "Save" button to create database.

A screenshot of a computer

Description automatically generated

**Creation of the tables And Insertion of the data**

A screenshot of a computer

Description automatically generated

**Query -**

BEGIN;

CREATE TABLE IF NOT EXISTS public.cost

(

cost\_id integer NOT NULL,

airfare numeric(20, 0),

other\_transport numeric(20, 0),

lodging numeric(20, 0),

meals numeric(20, 0),

other\_expenses numeric(20, 0),

total numeric(20, 0),

ref\_id integer,

CONSTRAINT cost\_pkey PRIMARY KEY (cost\_id)

);

CREATE TABLE IF NOT EXISTS public.doc\_info

(

ref\_id integer NOT NULL,

disclosure\_group character varying(50) COLLATE pg\_catalog."default",

title character varying(150) COLLATE pg\_catalog."default",

name character varying(50) COLLATE pg\_catalog."default",

purpose character varying(500) COLLATE pg\_catalog."default",

CONSTRAINT doc\_info\_pkey PRIMARY KEY (ref\_id)

);

CREATE TABLE IF NOT EXISTS public.organization

(

org\_id integer NOT NULL,

org\_owner\_title character varying(50) COLLATE pg\_catalog."default",

org\_owner character varying(50) COLLATE pg\_catalog."default",

ref\_id integer,

CONSTRAINT organization\_pkey PRIMARY KEY (org\_id)

);

CREATE TABLE IF NOT EXISTS public.travel\_info

(

info\_id integer NOT NULL,

destination character varying(150) COLLATE pg\_catalog."default",

start\_date date,

end\_date date,

ref\_id integer,

CONSTRAINT travel\_info\_pkey PRIMARY KEY (info\_id)

);

ALTER TABLE IF EXISTS public.cost

ADD CONSTRAINT cost\_ref\_id\_fkey FOREIGN KEY (ref\_id)

REFERENCES public.doc\_info (ref\_id) MATCH SIMPLE

ON UPDATE NO ACTION

ON DELETE NO ACTION

NOT VALID;

ALTER TABLE IF EXISTS public.organization

ADD CONSTRAINT organization\_ref\_id\_fkey FOREIGN KEY (ref\_id)

REFERENCES public.doc\_info (ref\_id) MATCH SIMPLE

ON UPDATE NO ACTION

ON DELETE NO ACTION;

ALTER TABLE IF EXISTS public.travel\_info

ADD CONSTRAINT travel\_info\_ref\_id\_fkey FOREIGN KEY (ref\_id)

REFERENCES public.doc\_info (ref\_id) MATCH SIMPLE

ON UPDATE NO ACTION

ON DELETE NO ACTION

NOT VALID;

END;

**Choose a file**

A screenshot of a computer

Description automatically generated

**In Options, Enable the “Header” And Click “OK”** A screenshot of a computer

Description automatically generated

**SQL Functions**

1. Give a query for 10 highest titles by total expense.

**Query** - select salary\_rank, title, total

from(

select title,total,

dense\_rank() over (order by total desc) as "salary\_rank"

from (

select d.title as title, sum(c.total) as total

from doc\_info d

inner join cost c

on d.ref\_id = c.ref\_id

group by title

) as title\_ranking

) as top\_10

where salary\_rank <= 10;

**Output**

|  |  |  |
| --- | --- | --- |
| **salary\_rank** | **title** | **total\_expense** |
| 1 | Minister | 215342 |
| 2 | Chief Executive Officer | 144068 |
| 3 | Director of Communications | 119040 |
| 4 | Chairperson Canada Industrial Relations Board | 113842 |
| 5 | Executive Director Secretariat to the Canadian Cultural Property Export Review Board | 108241 |
| 6 | Driver | 85490 |
| 7 | Executive Director and General Counsel Secretariat to the Canada Industrial Relations Board | 60189 |
| 8 | Chairperson Canada Industrial Relations Board | 55575 |
| 9 | Chief of Staff | 49798 |
| 10 | Chairperson Canadian International Trade Tribunal | 46101 |

**Insights for query**

Here we have got the 10 highest titles with there total expenses with the Ministers have the highest expense followed by other titles.

**Visualization**

**Insights of graph**

Here we have kept the expenses on y-axis and the title on x-axis and we can the bar chart as given above, and we have compared the expenses of 10 highest titles.

1. Calculate the range of the total cost incurred from travel and distribute them in five buckets

**Query** -

with min\_max as(

select min(total) as min\_total, max(total) as max\_total from cost

),

histogram as (

select width\_bucket(total,min\_total,max\_total,5) as bucket, count(cost\_id) as total\_count

from cost c inner join min\_max m

on 1=1

group by bucket

order by bucket

)

select concat(low\_value,' - ',high\_value) as bucket\_range,

total\_count

from (select bucket,

floor(((max\_total - min\_total)/5) \* (bucket - 1)) as low\_value,

floor(((max\_total - min\_total)/5) \* bucket) as high\_value,

total\_count

from histogram inner join min\_max on 1=1) as outer\_query

**Output**

|  |  |
| --- | --- |
| **bucket\_range** | **total\_count** |
| 0 - 3546 | 612 |
| 3546 - 7093 | 37 |
| 7093 - 10640 | 8 |
| 10640 - 14187 | 7 |
| 14187 - 17734 | 4 |
| 17734 - 21280 | 1 |

**Insights for query**

This query gives the count of the total costs incurred in travel added in a specific range of buckets, which are currency (CAD) in data format.

We have taken the total cost, added them in six different buckets in which they fall in, and sorted them in a descending order.

**Visualization**

A graph of numbers and a number of items

Description automatically generated with medium confidence

**Insights for graph**

Through this graph, we found that most of the travel costs falls between 0 - 3546 CAD with a count of 612 and only one travel cost falls between 17733 - 21280 CAD, which is the highest range of buckets.

1. To find the count of the starting months and rank them ( highest to lowest ).

**Query** -

select start\_month, total\_count,

rank () over (order by total\_count desc) as ranking

from (

select

case

when extract(month from start\_date) = 1 then 'January'

when extract(month from start\_date) = 2 then 'February'

when extract(month from start\_date) = 3 then 'March'

when extract(month from start\_date) = 4 then 'April'

when extract(month from start\_date) = 5 then 'May'

when extract(month from start\_date) = 6 then 'June'

when extract(month from start\_date) = 7 then 'July'

when extract(month from start\_date) = 8 then 'August'

when extract(month from start\_date) = 9 then 'September'

when extract(month from start\_date) = 10 then 'October'

when extract(month from start\_date) = 11 then 'November'

when extract(month from start\_date) = 12 then 'December'

end as start\_month,

count(info\_id) as total\_count

from travel\_info

group by start\_month

order by start\_month

) as rank\_outer

**Output**

|  |  |  |
| --- | --- | --- |
| **start\_month** | **total\_count** | **ranking** |
| September | 87 | 1 |
| March | 76 | 2 |
| May | 74 | 3 |
| October | 71 | 4 |
| February | 60 | 5 |
| November | 60 | 5 |
| April | 49 | 7 |
| June | 44 | 8 |
| July | 40 | 9 |
| January | 40 | 9 |
| December | 36 | 11 |
| August | 32 | 12 |

**Insights for query**

This query gives the findings of the total count of the months of the start date and ranks them accordingly.

This extracts the month of every start date and finds out the count by its occurrences.

**Visualization**

A graph with blue line and white text

Description automatically generated

**Insights for graph**

Through this graph, we can find the month of most travel occurrences and least travel occurrences. This helps the business to plan their budgets according to these output results.

1. To retrieve the names and elevations of facilities located in Alberta, ordered by latitude in descending order

**Query**

select coalesce(d.disclosure\_group, 'Total') as disclosure\_group,

round(sum(total),0)::integer as total\_expenses

from doc\_info d

inner join

cost c

on c.ref\_id = d.ref\_id

group by cube (disclosure\_group)

order by disclosure\_group

**Output**

|  |  |
| --- | --- |
| **disclosure\_group** | **total\_expenses** |
| FOVC | 28924 |
| MPSES | 755642 |
| SLE | 1112486 |
| Total | 1897052 |

**Insights for query**

This query gives the sum of total costs incurred by each disclosure group and its total sum of those groups together.

**Visualization**

A pie chart with numbers and a number of percentages

Description automatically generated

**Insights for graph**

This pie chart illustrates the cost incurred by the three groups “FOVC”, “MPSES”, and “SLE”, total cost added together as “Total” and its percentage of contribution towards its total sum.

Through these results, we can find the total cost incurred by every disclosure group and its individual cost incurrences as well.

1. Write a query through which you can find the total cost and what % of the amount the airfare cost contributes to the total cost for each organization.

**Query**

SELECT

org\_owner\_title,

total\_cost,

airfare\_percentage

FROM ( SELECT o.org\_owner\_title,

SUM(c.total) AS total\_cost,

( SUM(c.airfare) / SUM(c.total) \* 100) AS airfare\_percentage

FROM organization o

INNER JOIN cost c ON o.ref\_id = c.ref\_id

GROUP BY o.org\_owner\_title

) AS org\_costs

ORDER BY total\_cost ;

**Output**

|  |  |  |
| --- | --- | --- |
| **org\_owner\_title** | **total\_cost** | **airfare\_percentage** |
| Department of Justice Canada | 28924 | 33.22846079 |
| Accessibility Standards Canada | 270451 | 38.00466628 |
| Agriculture Food Canada | 755642 | 48.05913382 |
| Administrative Tribunals Support Service of Canada | 842035 | 33.25134941 |

**Insight for query**

By executing this query, you can gain valuable insights into the financial aspects of different organizations, particularly in terms of total costs and the proportion contributed by airfare expenses.

**Visualization**

**Insights for graph**

This graph shows that the Administrative Tribunals Support Service of Canada's airfare expenditure accounts for the largest portion of the organization's overall costs, while the Department of Justice Canada Organization's airfare expense accounts for the smallest portion.

1. Which locations has the highest expense per day.

**Query**

with ranked\_trips AS (

SELECT

ti.destination,

MIN(ti.start\_date) AS first\_trip\_start\_date,

MAX(ti.end\_date) AS last\_trip\_end\_date,

SUM(c.total) AS total\_spent,

RANK() OVER (ORDER BY SUM(c.total) DESC) AS rank\_num

FROM travel\_info ti

JOIN cost c ON ti.ref\_id = c.ref\_id

WHERE c.total > 5000 AND ti.start\_date >= '2018-01-01'

GROUP BY ti.destination

)

SELECT

destination AS "Location",

(last\_trip\_end\_date - first\_trip\_start\_date) AS "Duration (In Days)",

ROUND((total\_spent / (last\_trip\_end\_date - first\_trip\_start\_date)),2) AS "Expense Per Day (Calculated)",

total\_spent AS "Total Spent for Location"

FROM ranked\_trips

WHERE rank\_num <= 10

ORDER BY “Expense Per Day (Calculated)" DESC;

**Output**

|  |  |  |  |
| --- | --- | --- | --- |
| **Location** | **Duration (In Days)** | **Expense Per Day (Calculated)** | **Total Spent for Location** |
| Venice | 2 | 7653 | 15306 |
| Edmonton | 2 | 2895.5 | 5791 |
| Soul | 3 | 2504 | 7512 |
| Valparaiso | 3 | 2302.33 | 6907 |
| Geneve | 3 | 2052 | 6156 |
| Melbourne | 8 | 1525.38 | 12203 |
| Victoria and Vancouver | 5 | 1307 | 6535 |
| Beijing | 27 | 902.74 | 24374 |
| Hanoi | 928 | 59.12 | 54861 |
| Vancouver | 1477 | 15.97 | 23584 |

**Insight for query**

The provided query retrieves the top 10 destinations based on the total spending, considering the specified conditions. The output includes the destination, duration of trip, expense per day and the total spending across all qualifying trips for each destination.

Here we can reduce the travel expenses in cities with highest expense.

1. Which organization has travelled the most ?

**Query**

WITH trip\_counts AS (

SELECT

o.org\_owner\_title,

COUNT(ti.info\_id) AS trip\_count,

RANK() OVER (ORDER BY COUNT(ti.info\_id) DESC) AS row\_num

FROM

organization o

JOIN doc\_info di ON o.ref\_id = di.ref\_id

JOIN travel\_info ti ON di.ref\_id = ti.ref\_id

GROUP BY o.org\_owner\_title

HAVING COUNT(ti.info\_id) > 2

)

SELECT

org\_owner\_title,

trip\_count

FROM trip\_counts

WHERE row\_num <= 10;

**Output**

|  |  |
| --- | --- |
| **org\_owner\_title** | **trip\_count** |
| Administrative Tribunals Support Service of Canada | 311 |
| Agriculture Food Canada | 252 |
| Accessibility Standards Canada | 93 |
| Department of Justice Canada | 13 |

**Insight for query**

The provided query retrieves the top 10 organization owners based on the number of trips they have taken are more than 10. The output includes the organization owner title and the count of trips for each owner.

1. Write a SQL query to retrieve information about travelers who have visited Regina

**Query**

SELECT

DI.name AS traveler\_name,

TI.destination,

COUNT(C.cost\_id) AS num\_trips,

ROUND(AVG(C.total), 2) AS average\_expenses,

SUM(C.total) AS total\_expenses

FROM cost AS C

INNER JOIN travel\_info AS TI ON C.ref\_id = TI.ref\_id

INNER JOIN doc\_info AS DI ON C.ref\_id = DI.ref\_id

WHERE TI.destination = 'Regina'

GROUP BY DI.name, TI.destination;

**Output**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **traveler\_name** | **destination** | **num\_trips** | **average\_expenses** | **total\_expenses** |
| Abed Harb | Regina | 1 | 2634 | 2634 |
| Alison Porter | Regina | 1 | 2916 | 2916 |
| Guy Gallant | Regina | 1 | 1583 | 1583 |
| Lawrence MacAulay | Regina | 3 | 3330.333333 | 9991 |
| Luc Belanger | Regina | 1 | 3465 | 3465 |
| Marie Claude Bibeau | Regina | 1 | 2904 | 2904 |
| Philip Rizcallah | Regina | 1 | 1265 | 1265 |

**Insight for query**

The insights for the above SQL query is that it provides the information about travelers who have visited Toronto. It includes their names, the destination (Regina), the number of trips they have taken, the average expenses per trip, and the total expenses for all their trips to Regina.

1. To find the number of people traveling to Boston with their total expense and purpose?

**Query**

WITH RankedOrganizations AS (

SELECT

org.org\_owner,

ROUND(AVG(cost.total), 2) AS avg\_total\_cost,

RANK() OVER (ORDER BY AVG(cost.total) DESC) AS org\_rank

FROM

public.organization AS org

JOIN public.doc\_info AS doc ON org.ref\_id = doc.ref\_id

JOIN public.travel\_info AS travel ON doc.ref\_id = travel.ref\_id

JOIN public.cost AS cost ON doc.ref\_id = cost.ref\_id

GROUP BY

org.org\_owner

HAVING

COUNT(travel.info\_id) > 1

)

SELECT

org\_owner,

avg\_total\_cost

FROM

RankedOrganizations

WHERE

org\_rank <= 10;

**Output**

|  |  |
| --- | --- |
| **org\_owner** | **avg\_total\_cost** |
| aafc aac | 2998.58 |
| casdo ocena | 2908.08 |
| atssc scdata | 2707.51 |
| jus | 2224.92 |

**Insight for query**

The query looks at organizations that often spend a lot of money and the average spending of every owner. Based on the Average spending we can make further contract with owners.

If an organization is lower on the list, it might spend different amounts each time they travel. The query helps to figure out which organizations might be good partners based on how they spend money.

**Summary and Findings**

* From the above queries and insights, we have analyzed the data and here we can reduce the travel cost of the department or any title where we can get the numbers for each member.
* We also have analyzed the expenses by each city, each month and each department where can decide the future budget allocation for each department.
* The per day expenses in each city were analyzed, which can be useful insight for authorities to further reduce the expenses.
* Due to this project, we were able get the expenses on various categories and this data can help the authorities to minimize the expense which indirectly be used as a cost effectiveness project.