**SQL and Data Analytics**

**INFO8076 - Fall 2023 - Section 3**

Siddhesh Sanjay Otari

**Proactive Travel Disclosures by federal Institutions**

**Introduction**

I have created this Project for the Final-Term submission for the Business Analytics Course, and I have discussed the goals for this project below.

**Goal of the project**

* My goal of this project is to retrieve the Travel expense of various federal institutions by using the various queries below.
* After going through the data, I can analyze the travel expenses and according to that information I can do cost optimization in various areas and further I can increase efficiency.
* However, I can also find some patterns through which I can reduce the travel costs by compromising the effectiveness of the travel.
* Using historical travel expense data, I can forecast and according to the conditions budget projections for future budget can be decided. This can also aid financial planning and resource allocation.

**Data Source and how it was cleaned.**

* I took this data from **opencanada.ca**
* URL - [Proactive Disclosure - Travel Expenses - Open Government Portal (canada.ca)](https://open.canada.ca/data/en/dataset/009f9a49-c2d9-4d29-a6d4-1a228da335ce)
* Data Cleaning
  + Starting with this process I removed the French language columns as Ill as French values from the whole dataset so that it becomes understandable with for a larger audience and moving forward, I removed blank rows, duplicate values, replaced null strings with Nan values and then changed Nan values with appropriate values.
  + In addition to that I removed all the small things like comma, delimiter, special characters, and many more.
  + Finally, I removed the trailing white spaces in our data set and our dataset was cleaned and then I added primary key id columns for all four tables.

**Data Processing**

* Here I have created 4 tables namely doc\_info, travel\_info, cost, and organization.
* In doc\_info table there are 5 columns namely - ref\_id (PK), disclosure\_group, title, name, purpose.
* In travel\_info table there are 5 columns namely - info\_id(PK), destination, start\_date, end\_date, ref\_id(FK).
* In cost table there are 8 columns namely - cost\_id(PK), airfare, other\_transport, lodging, meals, other\_expenses, total, ref\_id(FK).
* In organization table there are 4 columns namely - org\_id(PK), org\_owner\_title, org\_owner, ref\_id(FK).
* Below is our ERD Diagram for further elaboration.

A screenshot of a computer

Description automatically generated

**Analysis and Insights**

* **Query 1**
  + Give a query for 10 highest titles by total expense.
* 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
  + Here I have got the 10 highest titles with their total expenses with the Ministers having the highest expense folloId by other titles.
  + From the data I can ask the particular titles to reduce their expenses.
* Visualization
* Insights
  + Here I have kept the expenses on y-axis and the title on x-axis and I can the bar chart as given above, and I have compared the expenses of 10 highest titles.

**Query 2**

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

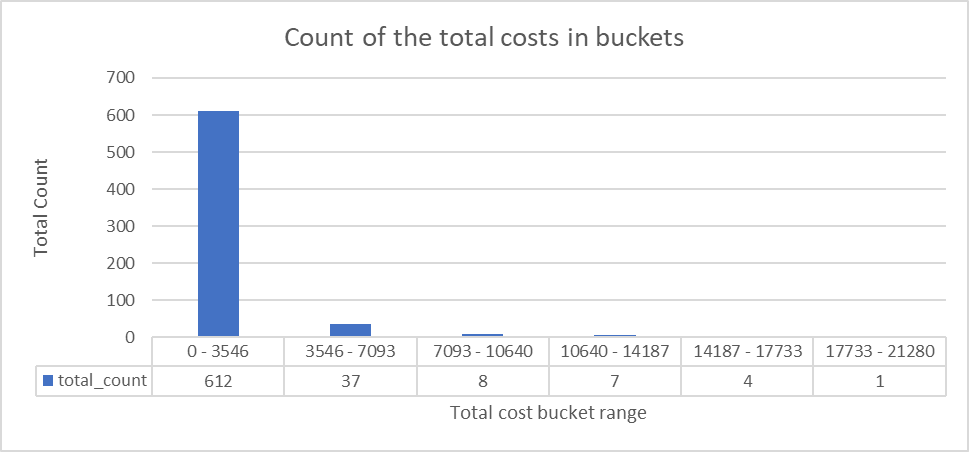
total\_count

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



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

**Query 3**

* + To find the count of the starting months and rank them (highest to loIst).
* 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
  + 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.
  + I also can get insights into trips per month on which I can further find the reason for highest trip in that particular month.
* Visualization

A graph with blue line and white text

Description automatically generated

* Insights
  + Through this graph, I 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.

**Query 4**

* + To find the total cost incurred by each disclosure group and find the total sum of them.
* 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
  + This query gives the sum of total costs incurred by each disclosure group and its total sum of those groups together.
  + Here I can further analyze why the disclosure group (SLE) has more expenses than the other groups.
* Visualization
* Insights
  + 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, I can find the total cost incurred by every disclosure group and its individual cost incurrences as Ill.

**Query 5**

* + 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.
* 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 |

* Insights
  + 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.
  + Also, I can allot the budget to departments for the next financial year on this previous data.
* Visualization
* Insights
  + 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.

**Query 6**

* + Which locations have the highest expense per day.
* 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 |

* Insights
  + 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 I can reduce the travel expenses in cities with highest expense.

**Query 7**

* Which organization has travelled the most.
* 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

ORDER BY trip\_count DESC;

* 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 |

* Insights
  + The provided query retrieves the organization owners based on the number of trips they have taken. The output includes the organization owner title and the count of trips for each owner.
  + I have also analyzed that their huge difference is the highest and loIst organization trips and further I can analyze the reason for his huge difference.

**Query 8**

* + Write a SQL query to retrieve information about travelers who have visited Regina.
* 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 |

* Insights
  + 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.

**Query 9**

* + To find the number of people traveling to Boston with their total expense and purpose?
* ----SELECT

CONCAT(DI.name, ' - ', TI.destination) AS traveler\_dest,

CONCAT ( TI.start\_date, ' to ', TI.end\_date) AS travel\_dates,

DI.purpose,

CONCAT(C.airfare, '+', C.lodging, '+', C.meals,'+',C.other\_transport,'+',C.other\_expenses) AS expenses,

C.total AS total\_expense

FROM cost C

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

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

WHERE TI. destination = 'Boston'

* Output

|  |  |  |  |
| --- | --- | --- | --- |
| **traveler\_dest** | **travel\_dates** | **purpose** | **total\_expense** |
| Guy Gallant - Boston | 2021-11-18 to 2021-11-19 | Minister to attend Seafood Expo North America attend bilateral meetings with provincial and state legislators | 3353 |
| Lawrence MacAulay - Boston | 2022-05-25 to 2022-05-25 | the Seafood Expo North America bilateral meetings with provincial and state legislators and stakeholders and deliver a keynote speech to the new England Canada Business Council | 3897 |
| Matthew Mann - Boston | 2022-09-18 to 2022-09-21 | Minister to attend Seafood Expo North America attend bilateral meetings with provincial and state legislators deliver a keynote speech to the new England Canada Business Council | 3095 |
| Lawrence MacAulay - Boston | 2016-08-17 to 2016-08-17 | Ministerial Mission to Boston Seafood Expo North America | 2214 |
| Virginia Adamson - Boston | 2017-03-01 to 2017-03-03 | a meeting of the Association of Labour Relations Agencies | 3348 |
| Ginette Brazeau - Boston | 2017-03-07 to 2017-03-10 | a meeting of the Association of Labour Relations Agencies | 1587 |
| Brazeau or Ginette - Boston | 2017-03-17 to 2017-03-21 | a meeting of the Association of Labour Relations Agencies | 3181 |
| Guilbert or Sylvie - Boston | 2017-05-19 to 2017-05-26 | a conference of the Association of Labour Relations Agencies ALRA | 3930 |
| Ebbs or Catherine - Boston | 2017-04-19 to 2017-04-20 | the conference of the Association of Labour Relations Agencies | 2589 |
| Adamson or Virginia - Boston | 2017-09-24 to 2017-09-25 | the 2018 Conference of Labour Board Chairs and Administrators | 3455 |

* Insights
  + I can see many people are traveling to Boston for a variety of events, including expo, conferences, seminars, meetings, and competition. Each event has a different objective behind them.
  + Each trip's total costs also vary significantly. The trips to Boston due to expo and meeting are significantly expensive as compared to other trips. Each trip's cost is in the range of 3000.
  + It can be due to many reasons, perhaps because they involve longer stays, more luxurious lodging, or other costs associated with the trip's objectives.

**Query 10**

* + Find out the Average cost of organizations.
* 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 |

* Insights
  + The query looks at organizations that often spend a lot of money and the average spending of every owner. Based on the Average spending I can make further contract with owners.
  + If an organization is loIr 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, I have analyzed the data and here I can reduce the travel cost of the department or any title where I can get the numbers for each member.
* I have also 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 Ire analyzed, which can be useful insight for authorities to further reduce the expenses.
* Due to this project, I was able to 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.