DATA & DISTRIBUTED SYSTEMS FOR ANALYTICS

Individual Project

AN ANALYSIS OF CORPORATE TRAVEL IN BRAZIL

2019 and 2020

PART 1

OVERVIEW

A leading expense management company in Latin America has contacted me to help analyze corporate travel data in Brazil in 2019 and 2020 using SQL to facilitate the process. I am committed to identifying patterns in corporate travel activity pertaining to all stakeholders involved (employees, employers, travel agencies, and hotels). The insights uncovered through this analysis can be shared with the different stakeholders to drive their business decisions.

DATA ACQUISITION

To undergo this process, I collected data from Kaggle consisting of 148,397 flight trips and 22,255 hotel reservations from 1335 employees who work at the 5 biggest corporations in Brazil.

DATA PREPARATION

The original dataset was very large, and included information on travel in 2017, 2018, and even predictive data for 2021, 2022, and 2023. My goal was to analyze historical data and generate my own insights, therefore, I decided to DROP all observations except those relating to 2019 and 2020. I was selective and kept relevant data only. An example of the SQL query I used is:

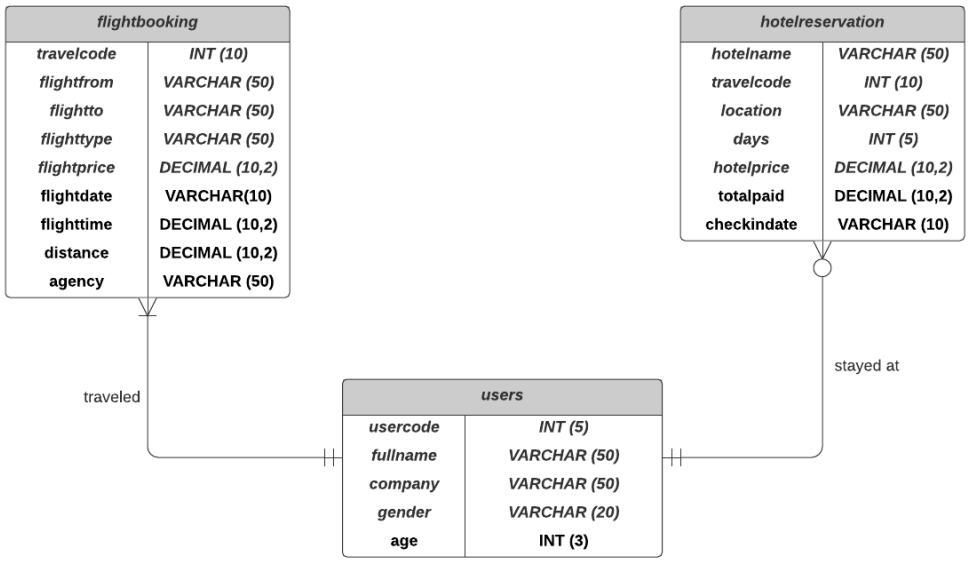
```
DELETE FROM hotel
WHERE checkindate REGEXP '/2022'
```

To get the data ready for analysis, I had to change the column names and edited the format of certain data types. For example, the way the DATE was entered (01/01/2019), phpMyAdmin was unable to read it as a DATE. So, I changed the variable type to VARCHAR(10). It still contained the same information, but I was able to efficiently manipulate SQL to fit my needs.

I chose this dataset because as an avid traveller myself, and someone who is hoping to get a job which allows me travel frequently, it is interesting to gain an understanding of how much corporations spend on corporate travel every year.

PART 2

ERD



ASSUMPTIONS

1. Each user can have many flight bookings (could have completed multiple business trips) and must have at least one flight booking on file to be in the database.
2. Each flight booking must be related to only one user (mandatory)
3. Each user can have many hotel reservations (could have completed multiple overnight business trips), but does not have to have a hotel reservation (e.g. went on a one-day business trip).
4. Each hotel reservation must be related to only one user (mandatory)

DATA DICTIONARY

Description of Entities

Entity Name	Description	Aliases	Occurrence
users	Contains information needed to identify individuals/employees who went on business trips	employee	One user may make multiple flight bookings (mandatory) and may have multiple hotel reservations (optional)
flightbooking	Contains information needed to describe the flight that was taken by the user	flight booking	One flight booking may have been made by one user only (mandatory)

hotelreservation	Contains information needed to describe the hotel stay taken by the user during the business trip	hotel reservation	On hotel reservation may have been made by one user only (mandatory)

Description of Attributes

Entity Name	Attributes	Description	Data Type	Nulls	Multi-valued	Derived	Default
users	**usercode**	unique id for each user	INT(5)	NO	NO	NO	NONE
	fullname	full name of user	VARCHAR(50)	NO	NO	NO	NONE
	company	name of company that the user works for	VARCHAR(50)	NO	NO	NO	NONE
	gender	gender of the user	VARCHAR(20)	YES	NO	NO	NONE
	age	age of the user	INT(5)	YES	NO	NO	NONE
flightbooking	**travelcode**	unique id of each flightbooking	INT(10)	NO	NO	NO	NONE
	flightfrom	city where the flight departs from	VARCHAR(50)	NO	NO	NO	NONE
	flightto	city where the flight lands	VARCHAR(50)	NO	NO	NO	NONE
	flighttype	seat class of the booking (economy, first class, premium)	VARCHAR(50)	NO	NO	NO	NONE
	flightprice	price of the flight	DECIMAL(10,2)	NO	NO	NO	NONE
	flightdate	date when the flight departs the departure city	VARCHAR(10)	NO	NO	NO	NONE
	flighttime	duration of the flight from takeoff to landing	DECIMAL(10,2)	NO	NO	NO	NONE
	distance	distance between the departure city and the arrival city	DECIMAL(10,2)	NO	NO	NO	NONE
	agency	travel agency that booked the flight for the user	VARCHAR(50)	NO	NO	NO	NONE

hotelreservation	**hotelname**	name of the hotel	VARCHAR(50)	NO	NO	NO	NONE
	travelcode	unique id associated with the booking	INT(10)	NO	NO	NO	NONE
	location	city where the hotel is located	VARCHAR(50)	NO	NO	NO	NONE
	days	number of nights that the user will be staying at the hotel	INT(5)	NO	NO	NO	NONE
	hotelprice	price of the hotel room per night	DECIMAL(10,2)	NO	NO	NO	NONE
	totalpaid	total price paid by the user for the duration of their stay at the hotel	DECIMAL(10,2)	NO	NO	NO	NONE
	checkindate	date when the user checks in to the hotel	VARCHAR(50)	NO	NO	NO	NONE

RELATIONAL SCHEMA

users(**usercode**, name, company, gender, age)
PRIMARY KEY: usercode

flightbooking(**travelcode**, **travelfrom**, travelto, flighttype, flightprice, flightdate, flighttime, distance, agency, *usercode*)
PRIMARY KEY: travelcode, travelfrom
FOREIGN KEY: usercode

hotelreservation(**hotelname, travelcode**, location, days, hotelprice, totalpaid, checkindate, *usercode*)
PRIMARY KEY: hotelname, travelcode
FOREIGN KEY: usercode

CORRESPONDING TABLES TO STORE THIS DATASET IN MYSQL

CREATE STATEMENTS

```
CREATE TABLE users (
usercode INT(5) NOT NULL,
fullname VARCHAR(50) NOT NULL,
company VARCHAR (50) NOT NULL,
gender VARCHAR(20) NOT NULL,
age INT(3) DEFAULT NULL,
PRIMARY KEY (usercode)
);


CREATE TABLE flightbooking (
travelcode INT(10) NOT NULL,
flightfrom VARCHAR(50) NOT NULL,
flightto VARCHAR (50) NOT NULL,
flighttype VARCHAR(50) NOT NULL,
flightprice DECIMAL(10,2) NOT NULL,
flightdate VARCHAR (50) NOT NULL,
flighttime DECIMAL(10,2) NOT NULL,
distance DECIMAL(10,2) NOT NULL,
agency VARCHAR(50) DEFAULT NULL,
usercode INT(5) NOT NULL,
PRIMARY KEY (travelcode, flightfrom),
FOREIGN KEY (usercode) REFERENCES users (usercode)
);

CREATE TABLE hotelreservation (
hotelname VARCHAR(50) NOT NULL,
travelcode INT(10) NOT NULL,
location VARCHAR(50) NOT NULL,
days INT(5) NOT NULL,
hotelprice DECIMAL(10,2) NOT NULL,
totalpaid DECIMAL(10,2) NOT NULL,
checkindate VARCHAR(50) NOT NULL,
usercode INT(5) NOT NULL,
PRIMARY KEY (hotelname, travelcode),
FOREIGN KEY (usercode) REFERENCES users (usercode)
);
```

INSERT STATEMENTS

*Please see attached SQL file for INSERT statements including VALUES. Sample code:

```
INSERT INTO users (
  usercode, fullname, company, gender, age)
VALUES (….
);

INSERT INTO flightbooking(
  travelcode, travelfrom, traveltoto, flighttype, flightprice,
flightdate, distance, agency, usercode)
VALUES (….
);

INSERT INTO hotelreservation(
  hotelname, travelcode, location, days, hotelprice, totalpaid,
checkindate, usercode)
VALUES (….
);
```

PART 3

QUERIES

I want to start off my analysis by getting a summary of some descriptive insights about my dataset. All the queries used to obtain this preliminary information is quite simple and similar (i.e. repetitive), however, it is essential when starting such an analytical project, to gain an overview of your data:

1)

a) Total amount of money spent on <u>business travel</u> over the last 2 years, and the amount of money spent in 2019 vs 2020.

```
SELECT '2019 & 2020' AS Year, SUM(flightprice) AS GrandTotal FROM
flightbooking
UNION
SELECT '2019' AS Year, SUM(flightprice) AS 2019Total FROM flightbooking
WHERE flightdate regexp '/2019'
UNION
SELECT '2020' AS Year, sum(flightprice) AS 2020Total FROM flightbooking
WHERE flightdate regexp '/2020';
```

Year	Grand Total
2019 & 2020	141824265.08
2019	34124281.03
2020	107699984.05

Overall, over R$ 141,824,265.08 (Brazilian Real) was spent on corporate travel in 2019 and 2020 by the top 5 corporations. There was a huge increase in spending from 2019 (R$ 34,124,281.03) to 2020 (R$ 107,699,984.05) – over 3 time more money was spent. Although this data does not explicitely provide information pertaining to why that occurred, the corporations can use this insights and discuss internally to find out why they spent more money on corporate travel in 2020.

b) Total amount of money spent on <u>hotel accommodation</u> over the last 2 years, and the amount of money spent in 2019 vs 2020.

```
SELECT '2019 & 2020' AS Year, SUM(totalpaid) AS GrandTotal, SUM(days)
AS TotalDays FROM hotelreservation
UNION
select '2019'AS Year, SUM(totalpaid) AS 2019Total, SUM(days) AS
2019Days FROM hotelreservation WHERE checkindate REGEXP '/2019'
UNION
SELECT '2020' AS Year, SUM(totalpaid) AS 2020Total, SUM(days) AS
2020Days FROM hotelreservation WHERE checkindate REGEXP '/2020';
```

Year	Grand Total	TotalDays
2019 & 2020	11909789.62	55543
2019	2858892.14	13361
2020	9050897.48	42182

Overall, over R$ 11909789.62 was spent on hotels in 2019 and 2020 by the top 5 corporations. There was a huge increase in spending from 2019 (R$ 2,858,892.14) to 2020 (R$ 9,050,897.48) – over 3 time more money was spent. Similarly, the number of days spent in hotels more than tripled from 13,361 days to 42,182 days. This data can be used by the corporations to decide if they should be sending their employees on more one-day trips rather than over-night trips, if they want to cut costs.

c) Total distance travelled by flights and the total time spent travelling over the last 2 years. Additionally, find the breakdown of distance and time travelled by each company over the last 2 years.

```
SELECT SUM(distance) AS TotalDistance, SUM(flighttime) TotalTime
FROM flightbooking;

SELECT users.company, SUM(flightbooking.distance) AS TotalDistance,
SUM(flightbooking.flighttime) TotalTime
FROM flightbooking
JOIN users
ON users.usercode = flightbooking.usercode
GROUP BY users.company;
```

TotalDistance	TotalTime
81057797.39	210611.38

company	TotalDistance	TotalTime
4You	25207425.30	65505.82
Acme Factory	18986103.30	49335.45
Monsters CYA	10289941.72	26726.76
Umbrella LTDA	14740027.64	38301.61
Wonka Company	11834299.43	30741.74

Overall, 81,057,797.39km was covered by air travel by all 5 corporations. 4You had the highest distance, and Wonka Company had the least. If these companies are seeking to cut down on their carbon footprint, they could use these figures as starting point. Additionally, a total of 210611.38 hours was spent travelling, which is a loss of productive time for employees and the company. These corporations could use this information to figure out how to cut down on travelling so that they can maximize employee time for more productive work.

d) How many employees travelled on business in 2019 vs in 2020

```
SELECT '2019' AS YEAR, COUNT(DISTINCT usercode) AS 2019Total
FROM flightbooking
WHERE flightdate REGEXP '/2019'
UNION
SELECT '2020' AS YEAR, COUNT(DISTINCT usercode) AS 2020Total
FROM flightbooking
WHERE flightdate REGEXP '/2020';
```

YEAR	2019Total
2019	1335
2020	1233

Less employees travelled in 2020 than in 2019. However, according to Query #1a, more money was spent on travel in 2020 than in 2019. There could be several reasons for this: flight prices increased; employees flew more firstclass than economy class; etc… This is an interesting insight which can be explored further with market research.

****

Now that we have a basic overview of the corporate travel landscape pertaining to our Brazilian companies, let us take a closer look at the details of the data set.

2) What is the destination that most employees (users) travelled to for business?

```
SELECT DISTINCT flightto, COUNT(*)
FROM
    (SELECT flightto
    FROM flightbooking) AS travelcount
GROUP BY flightto
ORDER BY COUNT(*) DESC
```

flightto	count(*) ▼ 1
Florianopolis (SC)	30991
Aracaju (SE)	20369
Campo Grande (MS)	18827
Brasilia (DF)	16711
Recife (PE)	16545
Natal (RN)	13221
Sao Paulo (SP)	13005
Salvador (BH)	9446
Rio de Janeiro (RJ)	9282

This shows that most people flew to the city of Florianopolis (SC) which could mean that it is a business hub. The corporations could use this information when deciding whether to open a new branch in that city. If they have a branch there, it will limit travel and that means they could take advantage of the business opportunities in that city. Furthermore, as a potential business hub, the hotels could use this information as well to decide where to open new locations.

3) **What is the gender distribution between men and women? Does one gender group take more business trips than the other?**

```
SELECT
    gender,
    COUNT(*)
FROM
    users,
    flightbooking
WHERE
    users.usercode = flightbooking.usercode
GROUP BY
    gender
```

gender	count(*)
	48848
female	50156
male	49393

Interestingly, more women took business trips than men. Could this mean that more women hold senior positions which require business travel and important meetings? We will need to explore further!

4) **Which hotel received the most reservations and how much did each hotel make in revenue from these reservations?**

```
SELECT
    hotelname,
    location,
    COUNT(*) AS COUNT,
    AVG(hotelprice) AS avg_price,
    SUM(totalpaid) AS revenue
FROM
    hotelreservation
GROUP BY
    hotelname,
    location
ORDER BY
    COUNT(*)
DESC
```

hotelname	location	count	avg_price	revenue
Hotel K	Salvador (BH)	2808	263.410000	1863362.34
Hotel CB	Rio de Janeiro (RJ)	2785	165.990000	1144833.03
Hotel AF	Sao Paulo (SP)	2701	139.100000	946992.80
Hotel BD	Natal (RN)	2624	242.880000	1585520.64
Hotel AU	Recife (PE)	2474	312.830000	1957377.31
Hotel BP	Brasilia (DF)	2451	247.620000	1513205.82
Hotel BW	Campo Grande (MS)	2376	60.390000	355213.98
Hotel Z	Aracaju (SE)	2241	208.040000	1154413.96
Hotel A	Florianopolis (SC)	1795	313.020000	1388869.74

We can see that although Hotel K had the most reservations (2808), it did not make the most amount of money. The most amount of money was made by Hotel AU (R$ 1,957,377.31). This is because Hotel AU has a higher average price per night. Therefore, Hotel K could use this insight and decide to explore the possibility of increasing their price per night. However, they would need to do more market research on the hotel prices in their location (Salvador (BH)).

5) **Which users did not stay in hotels during their business travels?**

```
SELECT
    usercode,
    fullname
FROM
    users
WHERE NOT EXISTS
    (
    SELECT
        usercode
    FROM
        hotelreservation
    WHERE
        users.usercode = hotelreservation.usercode
)
```

usercode	fullname
33	Ida Turzak
42	Robert Collins
128	Alexander Carter
298	Gary Schwab
315	Alfred Atkinson
361	Curtis Sexton
364	Kristina Schuetz
382	Arthur Mckinnis
385	Jeffery Gill
463	Robert White
527	Elma Gonzales
588	Mary Parnell
595	Maureen Burns
647	Edna Ortiz
670	Patrick Obrien
720	Tina Heath
845	Alex Branhan
863	Ronald Markus
883	Laura Webb
893	Jessica Godina
983	Daniel Marin

A total of 25 employees did not have hotel bookings.

6) **Which hotel offers the cheapest nightly rate, and which hotel offers the most expensive nightly rate?**

```
SELECT DISTINCT 'Max' AS ranking, hotelname, location, hotelprice
FROM hotelreservation
WHERE hotelprice IN (
    SELECT MAX(hotelprice)
    FROM hotelreservation
)
UNION
SELECT DISTINCT 'Min' AS ranking, hotelname, location, hotelprice
FROM hotelreservation
WHERE hotelprice IN (
    SELECT MIN(hotelprice)
    FROM hotelreservation
)
```

ranking	hotelname	location	hotelprice
Max	Hotel A	Florianopolis (SC)	313.02
Min	Hotel BW	Campo Grande (MS)	60.39

The price difference between the max hotel rate and the minimum one is quite large (R$ 252.63). Of course, this is relative to the average hotel prices of their respective cities.

7) ****THIS IS A THREE-PART QUERY**:***

a) Which company has the most business flight activity?

```
SELECT company, COUNT(*)
FROM users, flightbooking
WHERE users.usercode = flightbooking.usercode
GROUP BY company
ORDER BY COUNT(*) DESC
```

company	count(*) 1
4You	50216
Acme Factory	28266
Wonka Company	26035
Umbrella LTDA	21961
Monsters CYA	21919

4YOU booked 50,216 flights for its employees in the last 2 years, which is twice as many as the other companies.

b) Which users did the most travel? Which company do these users work for? Did the top traveller come from the top company (see 4a)?

```
SELECT
    users.usercode,
    users.fullname,
    company,
    COUNT(*) AS flight_count
FROM
    users,
    flightbooking
WHERE
    users.usercode = flightbooking.usercode
GROUP BY
    flightbooking.usercode
ORDER BY
    flight_count
DESC
```

usercode	fullname	company	count(*) 1
1101	Kristin Gaiser	Acme Factory	133
896	Josephine Sigmon	Acme Factory	133
1105	Cynthia Hambleton	Acme Factory	133
900	Darin Lopez	Acme Factory	133
904	Rebecca Beach	Acme Factory	133
1317	Goldie Pankratz	Umbrella LTDA	133
1113	Travis Cordell	Acme Factory	133
908	Mary Worley	Acme Factory	133
1321	Cynthia Starkey	Umbrella LTDA	133
912	Nicole Gray	Acme Factory	133
1325	Doris Wendt	Umbrella LTDA	133
1121	Carrie Weir	Acme Factory	133
920	Vincent Reyna	Acme Factory	133
1333	Misty Littlefield	Umbrella LTDA	133
924	Marie Witt	Acme Factory	133
928	Sonia Walsh	Acme Factory	133
1141	Anthony Doggett	Acme Factory	133
936	Mary Everts	Acme Factory	133
1145	Leonor Blanchard	Acme Factory	133
940	Ronald Sullivan	Acme Factory	133

It looks like there are multiple employees who took 133 flights (which is the top # of flights) – it is necessary to write another query to determine which company has the most top travellers:

```
SELECT companys.company, count(companys.company)
FROM (SELECT users.usercode, users.fullname, company, count(*) as
counter
from users
INNER JOIN flightbooking ON users.usercode = flightbooking.usercode
group by flightbooking.usercode
order by count(*), company) AS companys
WHERE companys.counter = '133'
GROUP BY companys.company
```

company	count(companys.company)
4You	306
Acme Factory	160
Monsters CYA	141
Umbrella LTDA	137
Wonka Company	155

We can see that most of the top travellers work at 4You. This company should look into getting travel discounts or loyalty points if they have not already – as there could be some substantial savings to be done!

8) **Let's take a look at the different classes of flights (economic, first class, premium) that each company purchased.**

```
SELECT users.company, flightbooking.flighttype,
count(flighttype), avg(flightprice), sum(flightprice)
from users, flightbooking
WHERE users.usercode = flightbooking.usercode
GROUP BY users.company, flightbooking.flighttype
ORDER BY company, flighttype
```

company ▲ 1	flighttype ▲ 2	count(flighttype)	avg(flightprice)	sum(flightprice)
4You	economic	14189	636.040102	9024773.01
4You	firstClass	21631	1143.006404	24724371.53
4You	premium	14396	884.026075	12726439.38
Acme Factory	economic	8223	722.624363	5942140.14
Acme Factory	firstClass	11916	1299.751832	15487842.83
Acme Factory	premium	8127	1015.911441	8256312.28
Monsters CYA	economic	6273	615.645002	3861941.10
Monsters CYA	firstClass	9249	1099.353098	10167916.80
Monsters CYA	premium	6397	848.168696	5425735.15
Umbrella LTDA	economic	6292	724.678595	4559677.72
Umbrella LTDA	firstClass	9444	1297.097248	12249786.41
Umbrella LTDA	premium	6225	1018.181009	6338176.78
Wonka Company	economic	7540	610.583719	4603801.24
Wonka Company	firstClass	11030	1095.608514	12084561.91
Wonka Company	premium	7465	853.421139	6370788.80

Several insights can be gleaned from this table. For instance, we see that 4You purchased more first-class flights for its employees (21,631) than any other flight type, AND overall, 4You was the company that bought the most first-class flights (21,631) out of all the other companies. On average, Umbrella LTDA and Acme Factory spent considerably more on all their flight types compared to the other companies. Therefore, they should perhaps look into finding a better travel agency to work with which gets them cheaper flight deals.

9) How many employees are in each age range?

```
SELECT
  COUNT(case when age >= 0 and age <= 25 then users.usercode end)
0_25,
  COUNT(case when age >= 26 and age <= 35 then users.usercode
end) 26_35,
  COUNT(case when age >= 36 and age <= 45 then users.usercode
end) 36_45,
  COUNT(case when age >= 46 and age <= 55 then users.usercode
end) 46_55,
  COUNT(case when age >= 56 and age <= 65 then users.usercode
end) 56_65,
  COUNT(case when age >= 66 then users.usercode end) over_66
FROM users
```

0_25	26_35	36_45	46_55	56_65	over_66
141	312	297	299	286	0

Most people who travel are in the older age groups, 46-55 and 36-45, and the younger employees (0-25) travel the least. This makes sense because usually, more traveling opportunities are given to senior employees and older people with more experience usually occupy these senior roles. On the other hand, younger people are more likely to be in entry-level positions which largely require less travel. ***Assumption:*** *There are no employees over 65, as that is the retirement age in Brazil. My data validates this assumption.*

10) Average length of stay per age group

```
ALTER TABLE users
ADD COLUMN age_range varchar(10) DEFAULT NULL AFTER age;

UPDATE users SET age_range = CASE
  when age >= 0 and age <= 25 then '0_25'
  when age >= 26 and age <= 35 then '26_35'
  when age >= 36 and age <= 45 then '36_45'
  when age >= 46 and age <= 55 then '46_55'
  when age >= 56 and age <= 65 then '56_65'
  when age >= 66 then 'over_66' END;

SELECT users.age_range, avg(days)
FROM users, hotelreservation
WHERE users.usercode = hotelreservation.usercode
GROUP BY age_range;
```

age_range	avg(days)
0_25	2.5116
26_35	2.5076
36_45	2.4794
46_55	2.4976
56_65	2.4902

There is no correlation between age_range and average days spent on business travel.

11) In which city are youngest employees most likely to be located?

```
SELECT
    users.age_range,
    flightbooking.flightfrom,
    COUNT(*)
FROM
    users,
    flightbooking
WHERE
    users.usercode = flightbooking.usercode
GROUP BY
    users.age_range,
    flightbooking.flightfrom
HAVING
    users.age_range = '0_25'
ORDER BY count(*) DESC
```

age_range	flightfrom	COUNT(*)
0_25	Florianopolis (SC)	2952
0_25	Aracaju (SE)	2169
0_25	Brasilia (DF)	1907
0_25	Recife (PE)	1868
0_25	Campo Grande (MS)	1798
0_25	Natal (RN)	1382
0_25	Sao Paulo (SP)	1304
0_25	Rio de Janeiro (RJ)	993
0_25	Salvador (BH)	954

Assumption: The departure city is where the office is located.
0-25 year olds are mostly located in Florianopolis(SC). It is common for younger people to go live in bigger cities in search for opportunities, so companies could centre their recruitment efforts on locations where young people are likely to be located – for entry level jobs.

12) On what months are most flights done?

```
ALTER TABLE flightbooking
ADD COLUMN month varchar(10) DEFAULT NULL AFTER flightdate;

UPDATE flightbooking SET month = CASE
  when flightdate REGEXP '^01/' then 'january'
  when flightdate REGEXP '^02/' then 'february'
  when flightdate REGEXP '^03/' then 'march'
  when flightdate REGEXP '^04/' then 'april'
  when flightdate REGEXP '^05/' then 'may'
  when flightdate REGEXP '^06/' then 'june'
  when flightdate REGEXP '^07/' then 'july'
  when flightdate REGEXP '^08/' then 'august'
  when flightdate REGEXP '^09/' then 'september'
  when flightdate REGEXP '^10/' then 'october'
  when flightdate REGEXP '^11/' then 'november'
  when flightdate REGEXP '^12/' then 'december'
  END;

SELECT month, count(*)
FROM flightbooking
```

```
GROUP BY month
ORDER BY count(*) DESC;

SELECT month, avg(flightprice)
FROM flightbooking
GROUP BY month
ORDER BY avg(flightprice) DESC
```

month	count(*) 1
october	20893
december	18813
november	18729
january	11326
september	10624
april	10321
march	10028
february	9902
july	9847
may	9793
august	9071
june	9050

2- Most travel occurs towards the end of the year (October, November, December)

month	avg(flightprice) 1
may	969.684986
march	963.559224
june	960.183883
november	956.534115
july	956.369857
august	956.008738
february	954.502952
december	954.338258
january	953.431265
april	952.894512
september	952.893357
october	948.129403

1 - Interestingly, there is no correlation between month and flight price – travel agencies and travel sites should use this information to adjust their prices and increase revenue

13) Is there a correlation between distance and days. We are trying to find out if people tend to stay in hotels for more days if they are travelling long distances, to avoid doing 2 long trips back-to-back).

```
ALTER TABLE flightbooking
ADD COLUMN avg_distance VARCHAR(20) DEFAULT NULL AFTER distance;

UPDATE flightbooking SET avg_distance = CASE
  when distance >= 0 and distance <= 200 then '0_200'
  when distance >= 201 and distance <= 400 then '201_400'
  when distance >= 401 and distance <= 600 then '401_600'
  when distance >= 601 and distance <= 800 then '601_800'
  when distance >= 801 and distance <= 1000 then '801_1000'
  END;

SELECT flightbooking.avg_distance, avg(days)
FROM flightbooking, hotelreservation
WHERE flightbooking.usercode = hotelreservation.usercode
GROUP BY avg_distance;
```

The output shows that there is no correlation between average distance travelled and average days spent in a hotel, because there is very little variation in the average days spent in hotel.

avg_distance	avg(days)
0_200	2.4992
201_400	2.5013
401_600	2.4924
601_800	2.4933
801_1000	2.4919

****

Let us take a deeper dive into company statistics

14) How many employees are there per company?

```
SELECT
    company,
    COUNT(usercode) AS Number_of_Employees
FROM
    users
GROUP BY
    Company
```

4You has the greatest number of employees in the database, which would explain why they ranked first in a lot of our queries (e.g. amount of money spent on flight, number of flights taken in the last 2 years, etc…)

company	Number_of_Employees
4You	452
Acme Factory	259
Monsters CYA	195
Umbrella LTDA	194
Wonka Company	235

15) What is the gender distribution in each company (% male and % female)

```
SELECT
    u1.company,
    u2.gender,
    COUNT(u2.gender),
    (
        COUNT(u2.gender) * 100 / totalemployees.totale
    ) AS percent
FROM
    users u1,
    users u2,
    (
    SELECT
        company,
        COUNT(gender) AS totale
    FROM
        users
    GROUP BY
        company
```

```
) AS totalemployees
WHERE
    u1.usercode = u2.usercode AND u1.company = totalemployees.company
GROUP BY
    u1.company,
    u2.gender
```

company	gender	Count(u2.gender)	percent
4You		164	36.2031
4You	female	151	33.3333
4You	male	138	30.4636
Acme Factory		83	31.8008
Acme Factory	female	85	32.5670
Acme Factory	male	93	35.6322
Monsters CYA		70	35.8974
Monsters CYA	female	64	32.8205
Monsters CYA	male	61	31.2821
Umbrella LTDA		55	28.3505
Umbrella LTDA	female	69	35.5670
Umbrella LTDA	male	70	36.0825
Wonka Company		68	28.6920
Wonka Company	female	79	33.3333
Wonka Company	male	90	37.9747

Overall, it is great to see that the gender distribution is fairly even between men and women (usually 30% each). 4You and Monsters CYA have slightly more females than males, while the other companies have slightly more males than females.

16) Age distribution at the company

```
SELECT
    u1.company,
    u2.age_range,
    COUNT(u2.gender) AS COUNT,
    (
        COUNT(u2.age_range) * 100 / totalemployees.totale
    ) AS percent
FROM
    users u1,
    users u2,
    (
    SELECT
        company,
        COUNT(age_range) AS totale
    FROM
        users
    GROUP BY
        company
) AS totalemployees
WHERE
    u1.usercode = u2.usercode AND u1.company = totalemployees.company
GROUP BY
    u1.company,
    u2.age_range
```

All companies have more employees in the older age ranges. Acme Factory has the least amount of young people (7.7%) while Umbrella LTDA has the greatest number of oldest employees (56_65) – 25.3%. There

company	age_range	COUNT	percent
4You	0_25	58	12.8319
4You	26_35	90	19.9115
4You	36_45	102	22.5664
4You	46_55	107	23.6726
4You	56_65	95	21.0177
Acme Factory	0_25	20	7.7220
Acme Factory	26_35	68	26.2548
Acme Factory	36_45	67	25.8687
Acme Factory	46_55	55	21.2355
Acme Factory	56_65	49	18.9189
Monsters CYA	0_25	17	8.7179
Monsters CYA	26_35	47	24.1026
Monsters CYA	36_45	39	20.0000
Monsters CYA	46_55	48	24.6154
Monsters CYA	56_65	44	22.5641
Umbrella LTDA	0_25	22	11.3402
Umbrella LTDA	26_35	50	25.7732
Umbrella LTDA	36_45	40	20.6186
Umbrella LTDA	46_55	33	17.0103
Umbrella LTDA	56_65	49	25.2577
Wonka Company	0_25	24	10.2128
Wonka Company	26_35	57	24.2553
Wonka Company	36_45	49	20.8511
Wonka Company	46_55	56	23.8298
Wonka Company	56_65	49	20.8511

17) Let's have a breakdown of the Total amount of money that each company spent on flights and on hotels separately, as well as the Grand Sum they spent on both flights and hotels over the last 2 years.

```
SELECT
    users.company,
    SUM(flightbooking.flightprice) AS TotalFlights,
    SUM(hotelreservation.totalpaid) AS TotalHotel,
    (
        SUM(hotelreservation.totalpaid) +
SUM(flightbooking.flightprice)
    ) AS TotalSpend
FROM
    users,
    flightbooking,
    hotelreservation
WHERE
    users.usercode = flightbooking.usercode AND users.usercode =
hotelreservation.usercode
GROUP BY
    users.company
ORDER BY
    TotalSpend
```

company	TotalFlights	TotalHotel	TotalSpend ▲ 1
Monsters CYA	363664181.10	240564838.58	604229019.68
Umbrella LTDA	443677762.65	215186783.69	658864546.34
Wonka Company	416937519.32	257619201.74	674556721.06
Acme Factory	549607602.79	265745129.90	815352732.69
4You	876181651.45	504083330.53	1380264981.98

As expected, based on other queries so far, 4You scores the highest in all Total figures. They spent almost double the amount of the other companies. That gives an average spend of $R3,053,683.59 per employee. On the other hand, Monsters CYA spent the least amount in terms of Total Spend, however, their average spend per employee equates to $R3,098,610.38 which is higher than 4You.

18) **SCENARIO: Since October, November and December are usually the busiest travel month (as determined by Query #12), Hotel AU wants to temporarily VIEW all of its upcoming reservations in those months to ensure that they are ready for the volume of people checking in.**

```
CREATE VIEW hotelau_oct_nov_dec AS SELECT
    hotelreservation.usercode,
    users.fullname,
    hotelreservation.checkindate,
    hotelreservation.days
FROM
    users
JOIN hotelreservation ON users.usercode = hotelreservation.usercode
WHERE
    hotelreservation.hotelname = 'Hotel AU' AND
hotelreservation.checkindate NOT REGEXP '/2019' AND
hotelreservation.checkindate REGEXP '^(10/|11/|12)'
ORDER BY
    hotelreservation.checkindate
```

There are a total of 448 upcoming reservations! Hotel AU will be very busy!

usercode	fullname	checkindate ▲ 1	days
247	Judy Mcmann	10/01/2020	3
264	Prince Webster	10/01/2020	2
323	Jean Obrien	10/01/2020	1
389	Mark Eisentrout	10/01/2020	3
403	Oliver Williams	10/01/2020	2
449	Valerie Gonzalez	10/01/2020	1
484	Julie Bourgeois	10/01/2020	2
500	Edward Zahner	10/01/2020	3
549	Margaret Orellana	10/01/2020	3
599	Leroy Harris	10/01/2020	1
614	Anna Zurita	10/01/2020	4
616	Jessica Kinsey	10/01/2020	2
660	Harlan Hudson	10/01/2020	2
680	Bret Dowell	10/01/2020	2
748	Bradley Rhodes	10/01/2020	2
770	Bobbie Murray	10/01/2020	1
782	Carolyn Nielsen	10/01/2020	4
786	Krista Vandyke	10/01/2020	1
830	Carlos Hernandez	10/01/2020	2
835	Phyllis Morse	10/01/2020	2
900	Darin Lopez	10/01/2020	1
908	Mary Worley	10/01/2020	1
942	Delia Rodriguez	10/01/2020	3
956	Clarence Gracely	10/01/2020	2
984	George Yanez	10/01/2020	4

19) Hotel AU knows from historical experience that business travellers who only book for 1 night are more likely to cancel. They would like to calculate how much minimum revenue they can expect from their busy months (Oct, Nov, Dec) even if the guests who booked for 1 night cancel.

```
SELECT
    SUM(
        subset.days * subset.hotelprice
    ) AS minimin_revenue
FROM
    (
    SELECT
        hotelreservation.hotelprice,
        hotelreservation.usercode,
        users.fullname,
        hotelreservation.checkindate,
        hotelreservation.days
    FROM
        users
    JOIN hotelreservation ON users.usercode = hotelreservation.usercode
    WHERE
        hotelreservation.hotelname = 'Hotel AU' AND
hotelreservation.checkindate NOT REGEXP '/2019' AND
hotelreservation.checkindate REGEXP '^(10/|11/|12)' AND
hotelreservation.days > 1
) AS subset
```

Even if all the one-night reservations are cancelled, Hotel AU can still expect to make a minimum of R$ 304,070.76 in the peak season.
***Assumption**: None of the reservations that are more than one night cancel.*

minimum_revenue
304070.76

20) SCENARIO: There has been a closure of the airport in Natal (RN). Travel agency 'FlyingDrops' wants to see all the upcoming flights that it has booked where employees are either departing from or landing at the airport in Natal (RN). With this information, they will be able to rebook their flights via another airport.

```
SELECT
    flightbooking.usercode,
    users.fullname,
    flightbooking.flighttype,
    flightbooking.flightdate
FROM
    users
JOIN flightbooking ON users.usercode = flightbooking.usercode
WHERE
    (
        flightbooking.flightto = 'Natal (RN)' OR flightbooking.flightfrom
= 'Natal (RN)'
    ) AND flightbooking.agency = 'FlyingDrops' AND(
        flightbooking.flightdate = '12/24/2020' OR
flightbooking.flightdate = '12/25/2020'
    )
```

usercode	fullname	flighttype	flightdate
20	Sally Roberts	firstClass	12/24/2020
178	Bobby Rutledge	firstClass	12/24/2020
398	Christine Spillane	firstClass	12/24/2020
563	Ashley Hernandez	firstClass	12/24/2020
619	Clifton Parsons	firstClass	12/24/2020
668	Carmen Politi	firstClass	12/25/2020
668	Carmen Politi	firstClass	12/24/2020
680	Bret Dowell	firstClass	12/24/2020
680	Bret Dowell	firstClass	12/25/2020
702	Joseph Clymer	firstClass	12/24/2020
719	Cory Ruiz	firstClass	12/24/2020
737	Paula Fleming	firstClass	12/24/2020
743	Mckenzie Dougherty	firstClass	12/24/2020
749	Ernest Reyes	firstClass	12/24/2020
760	Andrew Stoddard	firstClass	12/24/2020
783	Kelly Ross	firstClass	12/24/2020
862	Margaret Phillips	firstClass	12/24/2020
920	Vincent Reyna	firstClass	12/24/2020
920	Vincent Reyna	firstClass	12/25/2020
1141	Anthony Doggett	firstClass	12/24/2020
1157	Vilma Gehring	firstClass	12/24/2020

FlyingDrops will have to rebook a total of 21 flights. Interestingly, they are all first-class flights.