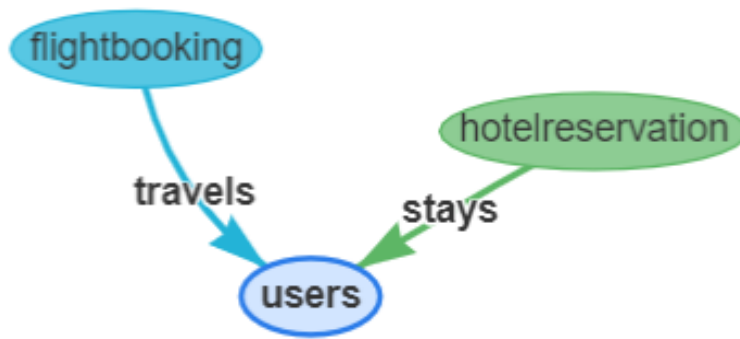


MCGILL DESAUTELS
Individual Project Part II (Neo4j)

AN ANALYSIS OF CORPORATE TRAVEL IN BRAZIL
2019 and 2020

***Please Note:** This document contains the same analysis as the one performed in Part 1. The only difference is that Part 1 uses SQL, while Part 2 uses Neo4j. Please see Part 1 for full analysis of results obtained from the queries.



1. .

- a. Total amount of money spent on business travel over the last 2 years, and the amount of money spent in 2019 vs 2020.

```

MATCH (n:flightbooking)
RETURN '2019 & 2020' AS type, sum(n.flightprice)
UNION
MATCH (n:flightbooking) WHERE n.flightdate CONTAINS '/2019'
RETURN '2019 Total' AS type, sum(n.flightprice)
UNION
MATCH (n:flightbooking) WHERE n.flightdate CONTAINS '/2020'
RETURN '2020 Total' AS type, sum(n.flightprice)
  
```

type	sum(n.flightprice)
"2019 & 2020"	141824265.07999897
"2019 Total"	34124281.03000124
"2020 Total"	107699984.04999153

- b. Total amount of money spent on hotel accommodation over the last 2 years, and the amount of money spent in 2019 vs 2020.

```

MATCH (n:hotelreservation)
RETURN '2019 & 2020' AS Year, sum(n.totalpaid) AS GrandTotal, sum(n.days) as TotalDays
UNION
MATCH (n:hotelreservation) WHERE n.checkindate CONTAINS '/2019'
RETURN '2019 Total' AS Year, sum(n.totalpaid) AS GrandTotal, sum(n.days) AS TotalDays
UNION
MATCH (n:hotelreservation) WHERE n.checkindate CONTAINS '/2020'
RETURN '2020 Total' AS Year, sum(n.totalpaid) AS GrandTotal, sum(n.days) AS TotalDays
  
```

Year	GrandTotal	TotalDays
"2019 & 2020"	11909789.620000899	55543
"2019 Total"	2858892.140000046	13361
"2020 Total"	9050897.480000226	42182

- 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.

```
MATCH (n:flightbooking)
RETURN sum(n.distance) AS totaldistance, sum(n.flighttime) AS totaltime
```

totaldistance	totaltime
81057797.3899981	210611.3800000272

```
MATCH (n:users) - [:travels]- (m:flightbooking)
RETURN n.company, sum(m.distance), sum(m.flighttime)
```

n.company	sum(m.distance)	sum(m.flighttime)
"4You"	25207425.300001353	65505.82000001894
"Monsters CYA"	10289941.71999996	26726.760000001654
"Wonka Company"	11834299.42999948	30741.739999995305
"Acme Factory"	18986103.299999673	49335.44999999363
"Umbrella LTDA"	14740027.639998777	38301.60999999377

- d. How many employees travelled on business in 2019 vs in 2020

```
MATCH (m:users) - [:travels] - (n:flightbooking) WHERE n.flightdate
CONTAINS '/2019' RETURN '2019' AS Year, count(distinct m.usercode)
UNION
MATCH (m:users) - [:travels] - (n:flightbooking) WHERE n.flightdate
CONTAINS '/2020' RETURN '2020' AS Year, count(distinct m.usercode)
```

Year	count(distinct m.usercode)
"2019"	1335
"2020"	1233

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

```
MATCH (n:flightbooking)
RETURN n.flightto, count(*)
ORDER BY count(*) DESC
```

n.flightto	count(*)
"Florianopolis (SC)"	30991
"Aracaju (SE)"	20369
"Campo Grande (MS)"	19827
"Brasilia (DF)"	16711
"Recife (PE)"	16545
"Natal (RN)"	13221
"Sao Paulo (SP)"	13005
"Salvador (BA)"	9446
"Rio de Janeiro (RJ)"	9282

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

```
MATCH (n:users) - [:travels] - (m:flightbooking)
RETURN n.gender, count(*)
```

n.gender	count(*)
"male"	49393
"female"	50156
""	48848

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

```
MATCH (n:hotelreservation)
RETURN n.hotelname, n.location, count(*), avg(n.hotelprice) AS avg_price,
sum(n.totalpaid) AS revenue
ORDER BY count(*) DESC
```

n.hotelname	n.location	count(*)	avg_price	revenue
"Hotel K"	"Salvador (BA)"	2808	263.41	1863362.3399999544
"Hotel CB"	"Rio de Janeiro (RJ)"	2785	165.89	1144833.0299999597
"Hotel AF"	"Sao Paulo (SP)"	2701	139.1	946992.8000000006
"Hotel BD"	"Natal (RN)"	2624	242.88	1585520.6399999974
"Hotel AU"	"Recife (PE)"	2474	312.83	1957377.3099999966
"Hotel BP"	"Brasilia (DF)"	2451	247.62	1513205.8200000208
"Hotel BW"	"Campo Grande (MS)"	2376	60.39	355213.98000000895
"Hotel Z"	"Aracaju (SE)"	2241	208.04	1154413.9600000107
"Hotel A"	"Florianopolis (SC)"	1785	313.02	1388869.7400000034

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

```

MATCH (n:users)
WHERE NOT EXISTS ((n) - [:stays] - ())
RETURN (n.usercode), (n.fullname)

```

(n.usercode)	(n.fullname)
33	"Isla Turck"
42	"Robert Collins"
128	"Alexander Carter"
298	"Gary Schwab"
315	"Alfred Alkinson"
361	"Curtis Sexton"
364	"Kristina Schultze"
382	"Arthur Mikinola"
483	"Robert White"
527	"Cima Gonzalez"

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

```

MATCH (n:hotelreservation)
CALL {MATCH (n:hotelreservation) RETURN max(n.hotelprice)}
RETURN 'Max' AS ranking, (n.hotelname), (n.location), (n.hotelprice) LIMIT 1
UNION
MATCH (n:hotelreservation)
CALL {MATCH (n:hotelreservation) RETURN min(n.hotelprice)}
RETURN 'Min' AS ranking, (n.hotelname), (n.location), (n.hotelprice) ORDER BY
(n.hotelprice) ASC LIMIT 1

```

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

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

a. Which company has the most business flight activity?

```
MATCH (n:users) - [:travels] - (m:flightbooking)
RETURN n.company, count(*)
ORDER BY count(*) DESC
```

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

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)?

```
MATCH (n:users) - [:travels] - (m:flightbooking)
RETURN n.usercode, n.fullname, n.company, count(*)
ORDER BY count(*) DESC
```

n.usercode	n.fullname	n.company	count(*)
0	"Roy Braun"	"4You"	133
2	"Wilma Mcinnis"	"4You"	133
3	"Paula Daniel"	"4You"	133
4	"Patricia Carson"	"4You"	133
6	"Jesse Decelle"	"4You"	133
8	"Jack Sabo"	"4You"	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:

```
MATCH (n:users) - [:travels] - (m:flightbooking)
WITH n, count(m.travelcode) as cnt
WHERE cnt = 133
RETURN n.company, count(*)
ORDER BY count(*) DESC
```

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

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

```
MATCH (n:users) - [:travels] - (m:flightbooking)
RETURN n.company, m.flighttype, count(m.flighttype), avg(m.flightprice),
sum(m.flightprice)
```

n.company	m.flighttype	count(m.flighttype)	avg(m.flightprice)	sum(m.flightprice)
"fyou"	"economic"	14189	636.0401021918357	9024773.00999774
"fyou"	"firstClass"	21631	1143.006404234686	24724371.53000013
"fyou"	"premium"	14396	884.0260702986967	12726439.379999861
"Monsters CVA"	"economic"	6273	615.6450023911999	3861941.100000027
"Monsters CVA"	"firstClass"	9249	1099.3030976321774	10167916.800000163
"Monsters CVA"	"premium"	6307	848.1686962636755	5425735.1500000525
"Worka Company"	"premium"	7465	853.4211386470187	6370788.800000033
"Worka Company"	"firstClass"	11030	1095.0085140525818	12084561.909999963
"Worka Company"	"economic"	7540	610.5837188328889	4603801.240000036
"Acme Factory"	"firstClass"	11916	1299.751831890591	15487842.629999637

- 9) How many employees are in each age range?

```
MATCH (n:users)
RETURN count(*), (case
when (n.age>=0 AND n.age <=25) THEN '0_25'
when (n.age>=26 AND n.age<=35) THEN '26_35'
when (n.age>=36 AND n.age<=45) THEN '36_45'
when (n.age>=46 AND n.age<=55) THEN '46_55'
when (n.age>=56 AND n.age<=65) THEN '56_65'
when n.age>=66 THEN 'over_65'
END) AS age_range
ORDER BY age_range
```

count(*)	age_range
141	"0_25"
312	"26_35"
297	"36_45"
299	"46_55"
286	"56_65"

- 10) Average length of stay per age group

```
MATCH (n:users) - [:stays] - (m:hotelreservation)
RETURN avg(m.days), (case
when (n.age>=0 AND n.age <=25) THEN '0_25'
when (n.age>=26 AND n.age<=35) THEN '26_35'
when (n.age>=36 AND n.age<=45) THEN '36_45'
when (n.age>=46 AND n.age<=55) THEN '46_55'
when (n.age>=56 AND n.age<=65) THEN '56_65'
when n.age>=66 THEN 'over_65'
END) AS age_range
ORDER BY age_range
```

avg(m.days)	age_range
2.5116082735331298	"0_25"
2.5075640031031785	"26_35"
2.47938668974461	"36_45"
2.497606701236535	"46_55"
2.4901960784313792	"56_65"

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

```
MATCH (n:users) - [:travels] - (m:flightbooking)
WHERE n.ageRange = '0_25'
RETURN n.ageRange, m.flightfrom, count(*)
ORDER BY count(*) DESC
```

n.ageRange	m.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

12) On what months are most flights done? And are flight prices more expensive during those months compared to other months?

```
match (n:flightbooking)
RETURN count(*), (case
when (n.flightdate STARTS WITH '01/') then 'january'
when (n.flightdate STARTS WITH '02/') then 'february'
when (n.flightdate STARTS WITH '03/') then 'march'
when (n.flightdate STARTS WITH '04/') then 'april'
when (n.flightdate STARTS WITH '05/') then 'may'
when (n.flightdate STARTS WITH '06/') then 'june'
when (n.flightdate STARTS WITH '07/') then 'july'
when (n.flightdate STARTS WITH '08/') then 'august'
when (n.flightdate STARTS WITH '09/') then 'september'
when (n.flightdate STARTS WITH '10/') then 'october'
when (n.flightdate STARTS WITH '11/') then 'november'
when (n.flightdate STARTS WITH '12/') then 'december'
END) AS months, avg(n.flightprice)
ORDER BY count(*) DESC
```

count(*)	months	avg(n.flightprice)
20893	"october"	948.129403149379
18813	"december"	954.3382581193869
18729	"november"	956.5341150088099
11326	"january"	953.4312652304426
10624	"september"	952.8933565512061
10321	"april"	952.8945121596748
10028	"march"	963.5592241723156

- 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).

```
MATCH (n:flightbooking) - [:travels] -> (m:users) <- [:stays] -
(p:hotelreservation)
RETURN (case
when (n.distance >= 0 AND n.distance <= 200) THEN '0_200'
when (n.distance >= 201 AND n.distance <= 400) THEN '200_400'
when (n.distance >= 401 AND n.distance <= 600) THEN '401_600'
when (n.distance >= 601 AND n.distance <= 800) THEN '601_800'
when (n.distance >= 801 AND n.distance <= 1000) THEN '801_1000'
END) as averagedistance, avg(p.days)
ORDER BY averagedistance
```

averagedistance	avg(p.days)
"0_200"	2.4991940792030576
"200_400"	2.5012505168262256
"401_600"	2.4924005139228425
"601_800"	2.4932826212563355
"801_1000"	2.491946191830189

- 14) How many employees are there per company?

```
MATCH (n:users)
RETURN n.company, count(*)
```

n.company	count(*)
"4You"	452
"Monsters CYA"	195
"Wonka Company"	235
"Acme Factory"	259
"Umbrella LTDA"	194

- 15) What is the gender distribution in each company

```
MATCH (n:users)
RETURN n.company, n.gender, count(*)
ORDER BY n.company, n.gender
```

n.company	n.gender	count(*)
"4You"	""	104
"4You"	"female"	151
"4You"	"male"	137
"Acme Factory"	""	82
"Acme Factory"	"female"	85
"Acme Factory"	"male"	92
"Monsters CYA"	""	70
"Monsters CYA"	"female"	64
"Monsters CYA"	"male"	61
"Umbrella LTDA"	""	55

16) Age distribution at the company

```
MATCH (n:users)
RETURN n.company, n.ageRange, count(*)
ORDER BY n.company, n.ageRange
```

n.company	n.ageRange	count(*)
"4You"	"0_25"	58
"4You"	"26_35"	90
"4You"	"36_45"	102
"4You"	"46_55"	107
"4You"	"56_65"	95
"Acme Factory"	"0_25"	20
"Acme Factory"	"26_35"	68
"Acme Factory"	"36_45"	67
"Acme Factory"	"46_55"	55
"Acme Factory"	"56_65"	49

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.

```
MATCH (n:flightbooking) - [:travels] -> (m:users) <- [:stays] -
(p:hotelreservation)
RETURN m.company, sum(n.flightprice) as TotalFlights, sum(p.totalpaid) as
TotalHotel, sum(n.flightprice + p.totalpaid) AS TotalSpend
ORDER BY TotalSpend
```

m.company	TotalFlights	TotalHotel	TotalSpend
"Monsters CYA"	363664181.099922	240564838.57984224	604229019.679961
"Umbrella LTDA"	443677762.6498403	215186783.6896713	658864546.3398758
"Wonka Company"	416937519.3203127	257619201.73979738	674556721.0599492
"Acme Factory"	549607602.7896469	265745129.89946762	815352732.6897895
"4You"	876181651.45021	504083330.5302359	1380264981.9798698

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.

```
MATCH (n:users)-[:stays] - (m:hotelreservation)
WHERE m.hotelname = 'Hotel AU' AND m.checkindate ENDS WITH '/2020' AND
(m.checkindate STARTS WITH '10/' OR m.checkindate STARTS WITH '11/' OR
m.checkindate STARTS WITH '12/')
RETURN n.usercode, n.fullname, m.checkindate, m.days
ORDER BY m.checkindate
```

userid	username	checkindate	ndays
247	"Judy Monami"	"10/01/2020"	3
284	"Phineas Webber"	"10/01/2020"	2
323	"Jean Olivier"	"10/01/2020"	1
389	"Mark Eisenhou"	"10/01/2020"	3
403	"Oliver Williams"	"10/01/2020"	2
449	"Valerie Gonzalez"	"10/01/2020"	1
484	"Jule Bourgeois"	"10/01/2020"	2
500	"Edward Zahner"	"10/01/2020"	3
549	"Margaret Orellana"	"10/01/2020"	3
589	"Leroy Harris"	"10/01/2020"	1

- 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.

```
MATCH (n:users) - [:stays] - (m:hotelreservation)
WITH n,m
WHERE m.hotelname = 'Hotel AU' AND m.checkindate ENDS WITH '/2020' AND
(m.checkindate STARTS WITH '10/' OR m.checkindate STARTS WITH '11/' OR
m.checkindate STARTS WITH '12/') AND m.days>1
RETURN sum(m.days * m.hotelprice) as minimum_revenue
```

minimum_revenue
304070.7600000002

- 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.

```
MATCH (n:users)-[:travels] - (m:flightbooking)
WHERE m.flightto = 'Natal (RN)' AND m.agency = 'FlyingDrops' AND (m.flightdate
CONTAINS '12/24/2020' OR m.flightdate CONTAINS '12/25/2020')
RETURN n.userid, n.fullname, m.flighttype, m.flightdate
```

userid	username	flighttype	flightdate
20	"Sally Roberts"	"firstClass"	"12/24/2020"
178	"Bobby Ruffedge"	"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"
688	"Carmen Pollt"	"firstClass"	"12/25/2020"
680	"Bret Dowell"	"firstClass"	"12/25/2020"
920	"Vincent Reyna"	"firstClass"	"12/24/2020"
1141	"Anthony Doggett"	"firstClass"	"12/24/2020"