Project: SQL - Travel

Path: start2impact University Master Data Science Course 1 SQL

Author: Giacomo Abramo

PRELIMINARY PHASE

- For the development of this project I used the DBMS MySQL.
- I immediately created a database using a name (“travel”) consistent with the topic under analysis.
- Subsequently, after making sure I understood the purpose of the project, I viewed all the tables necessary for the analysis; I uploaded only these tables by selecting and manipulating the structure of the columns useful for the purpose.

ANALYSIS OF TABLE “AMERICAN_DEATHS_ABROAD _10_09_to_06_16”

- In relation to this table, I wanted to investigate the values assumed by the "cause_of_death" variable in order to verify whether the cause of death could be associated with the dangerousness of the country. Among the various values that this variable assumes, the value "suicide" appears. In my opinion rows associated with this value should not be considered as the suicide of a person shouldn’t depend on the state of danger of the country in absence of additional information.

```
USE TRAVEL;
SELECT DISTINCT CAUSE_OF_DEATH
FROM AMERICAN_DEATHS_ABROAD
```

- Subsequently I counted the deaths for each country excluding from the count the lines in which the "cause_of_death" variable assumes the value "suicide". The query indicates that the 10 countries with the highest number of American deaths are: Mexico (1503), Haiti (221), Costa Rica (154), Philippines (154), Thailand (140), Dominican Republic (137), Jamaica (123), Bahamas (104), Afghanistan (101) and Honduras (94). I exported the resulting table from the following query naming it "Analysis_death_abroad".

```
SELECT
COUNTRY,
count(CAUSE_OF_DEATH) AS COUNT_COD
FROM AMERICAN_DEATHS_ABROAD
WHERE CAUSE_OF_DEATH <> "suicide"
GROUP BY COUNTRY
ORDER BY COUNT_COD DESC
```

- I repeated the same analysis of the previous step by discriminating not only for "country" but also for "date".
- In reference to “Mexico” in the years between 2009 and 2015 there were more or less the same deaths while in 2016 and 2009 the number of deaths was lower. In this regard it should be useful to know if the data for 2009 and 2016 are complete, i.e. if 2009 starts in January and if 2016 ends in December given that these years are the extreme values. The data seems to support my doubt as there is no data at the beginning of 2009 and at the end of 2016.
- In reference to Haiti we note a large number of American deaths in 2010. This number may have been dictated by the 2010 earthquake (https://it.wikipedia.org/wiki/Terremoto_di_Haiti_del_2010). So this country could also be considered dangerous due to the risk of natural disasters. There were far fewer deaths in 2011 (perhaps far fewer Americans traveled to Haiti following the previous year's crash). With regard to 2009 and 2016 there are the same problems already encountered for Mexico
- With reference to all the other countries, I did not find very different numbers when comparing the various years except for the usual problem linked to 2009 and 2016. Perhaps Honduras deserves further study as in 2016, despite this year being incomplete, there were more or less the same deaths as in other years

```
SELECT
COUNTRY,
right(DATE,2) AS YEAR,
count(CAUSE_OF_DEATH) AS COUNT_COD
FROM AMERICAN_DEATHS_ABROAD
WHERE CAUSE_OF_DEATH <> "suicide" AND COUNTRY IN ("Mexico", "Haiti", "Costa Rica", "Philippines", "Thailand",
"Dominican Republic", "Jamaica", "Bahamas", "Afghanistan", "Honduras")
GROUP BY COUNTRY, YEAR
ORDER BY CASE
WHEN COUNTRY = 'Mexico' then 1
WHEN COUNTRY = "Haiti" then 2
WHEN COUNTRY = 'Costa Rica' then 3
WHEN COUNTRY = 'Philippines' then 4
WHEN COUNTRY = 'Thailand' then 5
WHEN COUNTRY = 'Dominican Republic' then 6
WHEN COUNTRY = "Jamaica" then 7
WHEN COUNTRY = 'Bahamas' then 8
WHEN COUNTRY = 'Afghanistan' then 9
WHEN COUNTRY = 'Honduras' then 10
END ASC, YEAR DESC
```

ANALYSIS OF TABLE “BTSORIGINUS_10_09_to_06_16”

- After having analyzed the table linked to the deaths of Americans abroad, I have analyzed the table linked to the travels of Americans abroad. The 10 most visited countries are: Canada (41915), Mexico (35826), United Kingdom (10047), Japan (9151), Germany (8524), Colombia (7314), Brazil (6585), Dominican Republic (6345), The Bahamas (5393), South Korea (5242). I exported the resulting table from the following query naming it "Analysis_visit_abroad".

```
SELECT
COD.COUNTRY_NAME,
count(COUNTRY_NAME) AS COUNT_VISIT
FROM BTSORIGINUS ORI
LEFT JOIN COUNTRY_CODES COD
ON ORI.DEST_COUNTRY = COD.CODE
GROUP BY COD.COUNTRY_NAME
ORDER BY COUNT_VISIT DESC
```

JOINT ANALYSIS OF THE TWO TABLES

- Since full outer join is not supported in MySQL I used the following shortcut; the full outer join procedure was performed in order to check for problems in the data related to the different strings in the two tables but actually referring to the same country (for example I changed in the "Analysis_of visit_abroad" csv file "The Bahamas" ->"Bahamas" in order to match the line in this csv with the corresponding line of the "Analysis_of_death abroad" csv file).

```
SELECT
*
FROM ANALYSIS_DEATH_ABROAD DEATH
LEFT JOIN
ANALYSIS_VISIT_ABROAD VISIT
ON DEATH.COUNTRY = VISIT.COUNTRY_NAME
UNION
SELECT
*
FROM ANALYSIS_DEATH_ABROAD DEATH
RIGHT JOIN
ANALYSIS_VISIT_ABROAD VISIT
ON DEATH.COUNTRY = VISIT.COUNTRY_NAME
```

- After carrying out the above checks, I redid the join between the two tables with the inner method as I wanted to take into consideration only the data in common between the two tables. In order to consider relevant data I have also inserted a where filter. Based on the results obtained, the 10 most dangerous countries (in which the Death/Visit ratio is highest) are: Thailand, Vietnam, Afghanistan, Egypt, Grece, Philippines, Haiti, Pakistan, New Zealand and Nigeria. The 10 less dangerous countries (in which the Death/Visit ratio is lowest) are instead: Canada, Japan, United Kingdom, Brazil, Colombia, Taiwan, France, Germany, Panama and Cuba.

```
SELECT
*,
DEATH.COUNT_COD / VISIT.COUNT_VISIT AS DEATH_VISIT
FROM ANALYSIS_DEATH_ABROAD DEATH
INNER JOIN
ANALYSIS_VISIT_ABROAD VISIT
ON DEATH.COUNTRY = VISIT.COUNTRY_NAME
WHERE DEATH.COUNT_COD > 20 AND VISIT.COUNT_VISIT > 20
ORDER BY DEATH_VISIT DESC
```