Alonso David Bermúdez Moncayo
Didi Business Intelligence Challenge
All files are either on the ZIP file or in the github repository:
https://github.com/alonsomoncayo/Challenge---Alonso-Bermudez-Moncayo

Challenge:

First by doing an exploratory analysis I found out that the column visit_date from the table "restaurants_visitors" had some missing values, but given that the description of the table that says that "visit_date" is an extract of the column "visit_datetime" from the same table; therefore, I can fill those missing value by extracting the date with a SQL function from the "visit_datetime" column.

1. **Write the SQL queries necessary to generate a list of the five restaurants that have the highest average number of visitors on holidays. The result table should also contain that average per restaurant.**
   (Also in the sql file)

```
select r.id, avg(r.reserve_visitors) as Avg_num_vist
from restaurants_visitors r
join date_info d
on d.calendar_date = r.visit_date
where d.holiday_flg = 1
group by r.id
order by avg(r.reserve_visitors) desc
limit 5;
```

2. **Use SQL to discover which day of the week there are usually more visitors on average in restaurants.**
   (Also in the sql file)

```
select d.day_of_week,avg(r.reserve_visitors)
from restaurants_visitors r
join date_info d
on d.calendar_date = r.visit_date
group by d.day_of_week
order by avg(r.reserve_visitors) desc
limit 1;
```

3. **How was the percentage of growth of the amount of visitors week over week for the last four weeks of the data? Use SQL too.**
   (Also in the sql file)
   The percentage of growth has been negative for the last four weeks!

```
select *,
lead (num_visitors) over () as Previous_num_visitors,
num_visitors/lead (num_visitors) over () -1 as Pct_change
from (
select year(visit_date) as year ,week(visit_date) as week,
sum(reserve_visitors) as num_visitors
from restaurants_visitors
group by year(visit_date), week(visit_date)
order by year(visit_date) desc, week(visit_date) desc
limit 5) as T
limit 4;
```

4. **Forecast for the next six months, after the last date of the data, the sum of visitors of all the restaurants and validate the accuracy of your forecast. You can solve this question using the tool that you prefer.**
   (The answer of this question is contained on the pdf named “Didi-Challenge.pdf” and the code on “Didi-Challenge.ipynb”)

5. **Based on the data and your ideas, plan strategies to double the total restaurant visitors in six months.**
   My strategy to double the total restaurant visitors will be focused on what the data is showing and what have we seen in the past so we can infer the future always analysing new data, with that said my first idea is to observe past trends and flag what are best holidays for the restaurants and the best days of week. Once identified those I would create incentives for customers to go eat at the restaurants; for example, I would suggest the restaurants to give special prices for certain foods or make an special event on specific days on the calendar.
   My second idea is to identify the locations where there are more demand and maybe for the restaurant with low demand open new branches or sub-branches near those that way we can attract the attention of more potential customers.
6. **Imagine that these restaurants are in your city (and not in Japan), what other data would you want to join in order of get more insights to increase the visitors?**
   The data I would like to get access to are transactions, meaning what is the average ticket of a customer what are the meals and drinks they consume, the number of customers in one check and the time a check is paid. All this is information would be important to know to understand the behaviour of a customer inside the restaurant and help each individual restaurant to maximize their products and their strategy when attending a client. Also I find very useful to

have the email address of the customers that way restaurants can create a marketing strategy and increase interest. Age of the customer and method of payment, knowing this information will allow us to give a more unique and personalized experience, e.g. if a person prefers to pay with credit card a restaurant can invest in more cashless venues through technology specially for younger people.

7. **How many channels can you think of downloading a DiDi Rides APP and how will you estimate the quality and cost of each channel?**
   Referral program. – I think is high quality because these referrals are usually by friends and family which generates trust in new users, but DiDi needs a lot of incentives.
   Social Media. – Very high quality, everyone nowadays is in social media; therefore, this channel can reach a lot of people, but certaintly there is a lot of competition but with a good marketing strategy DiDi can position its brand very well.
   Influencers. – There is a reason why they are called "influencers" it's because of their ability to connect with people a goo campaign with a well-known influencer can boost DiDi's downloading's; however, usually internet personalities are very expensive to hire.
   I think one of the better metrics to assess if a certain channel would work is the ROI (Return on Investment) this metric will tell us how well an investment was made by a company, for that we need to know what was our total value of investment subtract the costs of the investment and divide it by the total costs.

8. **We want to build up a model to predict "Possible Churn Users" for DiDi Rides APP (e.g.: no trips in the past 4 weeks). Please list all features that you can think about and the data mining or machine learning model or other methods you may use for this case.**
   My Data points will be:
   a. Frequency of trips. – how frequent does a client make a trip per month.
   b. Date of last trip. – if it exceeds 4 weeks, we can raise a flag.
   c. Funnel conversion. – what is the rate that a customer orders a trip over how many times they open the app.
   d. Age of the customer. – We might know if there are capable of purchasing their own trips or not.
   e. Time in the app. – usually if they have been longer in the app is because they are loyal customers.

   For my model I would use a classification model it can be either logistic regression, random forest, or decision trees, the reasons are:

   a. It is a binary outcome either a customer has a high probability of being a churn user or has a low probability of being a churn user.
   b. Ability to capture nonlinear relationships if using random forests or decision trees.
   c. Feature importance. – this is key because it will give us insight on wich features contribute more with the model.
   d. It reduces overfitting if using random forest.

   Definitely my first choice would be a classification model, specially a nonlinear one.