Airline Quality

One of the sources we need to scrap was the Airline Quality website. We retrieve the comments related to airports and fly experience. In this report, we're going to explain in a first time which data we scrap and the format of these data. In a second time, we're going to develop our scraping method and how we automate the code with a robot and we're going to finish with some statistics.

I Format of scraped data :

The table below summary all data we scraped and give their type and a description of what is the column.

Variable name	Type	Description
Data_Source	string	Source of scrapping
Date_Review	date	Date of comment
Review	string	Comment in anglais
Title	string	Vidéo title
Date_Flown	date	Date of fly
Description	string	video description
Cleanliness	int	Overall on terminal cleanliness
Food_And_Beverages	int	Overall on Food and Beverages
Wifi_And_Connectivity	int	Overall on airport connection
Cabin_Staff_Service	int	Overall on airport staff
Overall_Customer_Rating	int	Global overall
Recommended	string	Recommandation of user
Airport	string	Name of airport concerning the fly
Type_Of_Traveller	string	Situation of user (couple, single)
Queuing_Times	int	Overall on Queuing Times in airport
Terminal_Seating	int	Overall on terminal seating
Terminal_Signs	int	Overall on terminal signs
Airport_Shopping	int	Overall on airport shopping
Experience_At_Airport	string	Depart or arrival in airport

II Scraping method:

We scrape out data in several steps :

- For each airport, select comments with description and overalls (according to date).
- For each scraped comment, we will retrieve and save the different variables in a dataFrame.
- DataFrame will be exported in a json file.

In order to retrieve the comments, you have to go through Beautiful Soup, so that you can to retrieve the page soup. In the soup, we have all tags and information pages. In order to retrieve the data we go through the htlm code of the page .

We made also a Robot which collects last data. To use it, you need just to give the number of days you want to have (fixed to 7 days) and if you execute the code, you're going to retrieve only the one week old comments. With a command, you can automate the process on OSIRIM (read the READ.txt files to know how).

III Statistics:

The final dataset counts 38002 rows X 63 columns (18 columns not NA) and weigh 82,5 Mo.

The following statistics shows for 8 numeric columns the mean, the median, the min, the max and the ecart-type of each of these columns.

All the statistics below are made on all comments about airports without date limitation (we don't take only one week old comments).

"""Cleanliness
moy 3.139530630905212
med 3.0
min 1
max 5
ecart-type 1.375485603444257

Food_And_Beverages
moy 2.4437046004842613
med 2.0
min 1
max 5
ecart-type 1.347665124592081

Wifi_And_Connectivity
moy 2.5715202967824786
med 2.0
min 1
max 5
ecart-type 1.472260229069088

Cabin_Staff_Service
moy 2.278249406802004
med 2.0
min 1
max 5
ecart-type 1.4610990139330031

Queuing_Times
moy 2.473302891933029
med 2.0
min 1
max 5
ecart-type 1.53215247994542

Terminal_Seating
moy 2.4315887645380734

med 2.0
min 1
max 5
ecart-type 1.3953449991820368

Terminal_Signs
moy 2.78709294269029
med 3.0
min 1
max 5
ecart-type 1.4267573846439106

Airport_Shopping
moy 2.6824801412180053
med 3.0
min 1
max 5
ecart-type 1.363486640956753
"""