

RUSSIAN SPY: Airlines data analysis



Состав команды

- | Козунов Артем
- | Богомолов Виктор
- | Геворгян Сона
- | Михеева Анастасия
- | Налимов Никита

ЦЕЛЬ ПРОЕКТА

Разработать методику поиска
шпионов с использованием
данных о полетах.



План работы:

- | Парсинг файлов в разном формате
- | Объединение и подготовка данных
- | Проверка гипотез
- | Выводы и future work

Парсинг данных

Исходные форматы:

- .csv
- .json
- .xml
- .tab
- .pdf
- .yaml
- .xlsx



Требуется:

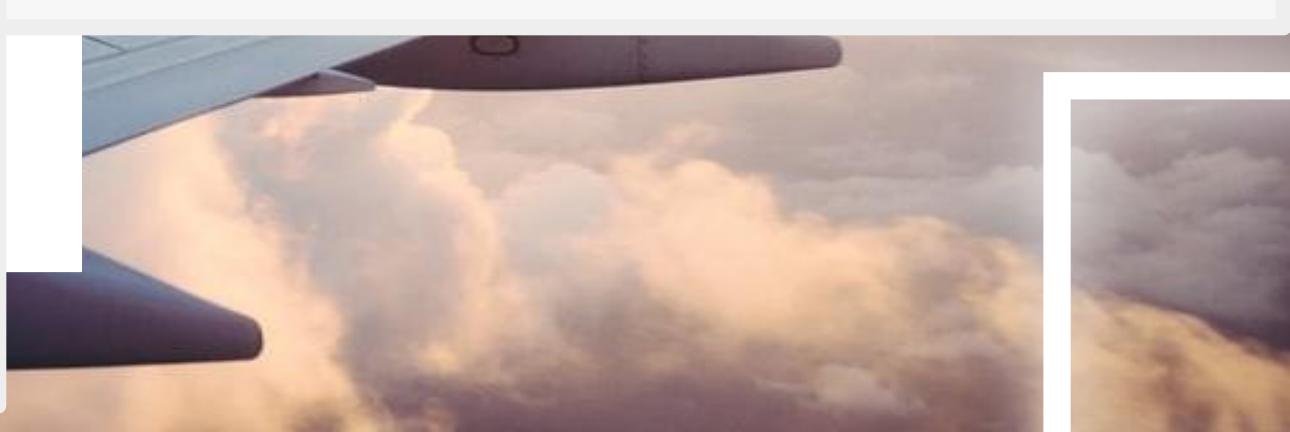
Привести все
файлы к общему
формату .csv

Пример парсинга

Для .xlsx файла

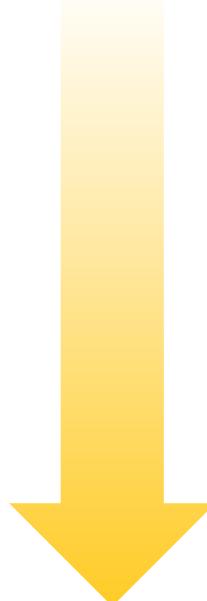
	A	B	C	D	E	F	G	H	I
1	BOARDING PASS					SEQUENCE:		32	
2									
3	MRS	LIDIYAZHDANOVA					Y		
4									
5	SU5436		VLADIVOSTOK				SEOUL		
6									
7	GATE	N/A		VVO		->		REA	
8									
9	2017-01-01	/	11:25		Operated by Some Other Airline				
10									
11	Boarding is ended 20 minutes before departure time				SEAT		N/A		
12									
13	PNR:	DYELAO		E-TICKET	7100246988860481				
14									
15									

```
import os
import openpyxl
import pandas as pd
directory = '/Users/a18289972/Desktop/Технологии БД/airlines-main/data/YourBoardingPassDotAero'
id_ = 0
data = []
for filename in os.listdir(directory):
    if filename.endswith('.xlsx'):
        xlsx_file = f'{directory}/{filename}'
        xlsx_file_obj = openpyxl.load_workbook(xlsx_file)
        for sheet in xlsx_file_obj.worksheets:
            sheet = xlsx_file_obj[sheet.title]
            data.append([])
            data[id_].append(sheet['H1'].value)
            data[id_].append(sheet['A3'].value)
            data[id_].append(sheet['B3'].value)
            data[id_].append(sheet['H3'].value)
            data[id_].append(sheet['A5'].value)
            data[id_].append(sheet['F3'].value)
            data[id_].append(sheet['B7'].value)
            data[id_].append(sheet['D5'].value)
            data[id_].append(sheet['D7'].value)
            data[id_].append(sheet['H5'].value)
            data[id_].append(sheet['H7'].value)
            data[id_].append(sheet['A9'].value)
            data[id_].append(sheet['C9'].value)
            data[id_].append(sheet['E9'].value)
            data[id_].append(sheet['A11'].value)
            data[id_].append(sheet['H11'].value)
            data[id_].append(sheet['B13'].value)
            data[id_].append(sheet['E13'].value)
            id_ += 1
columns = ['Sequence', 'Title', 'Name', 'Class', 'FlightNumber', 'BoardNumber',
           'Gate', 'From', 'FromCode', 'To', 'ToCode', 'Date',
           'Time', 'Operated', 'BoardingEnded', 'Seat', 'PNR', 'ETicket']
df = pd.DataFrame(data, columns=columns)
print(df.head())
df.to_csv(r'/Users/a18289972/Desktop/Технологии БД/airlines-main/data_csv/YourBoardingPassDotAero.csv',
          index_label='id')
```



Слияние и очистка

- Поиск схожих полей
- Приведение их к общему виду: форматирование
- Маппинг таблиц по совпадающим полям
- Заполнение пустых значений информацией из других таблиц (при наличии)
- Удаление пустых столбцов и строк



Итоги предобработки

- Получили общий файл в формате .csv
- По возможности заполнили пропуски
- Собрали максимальное количество информации

Данные пассажиров

- ФИО
- Дата рождения
- Номер документа
- Пол
- Номер бронирования
- Номер билета
- Тип питания
- Номер бонусной карты
- Программа лояльности
- Статус в программе
- Количество багажа
- Класс обслуживания

Данные о рейсе

- Номер рейса
- Дата и время вылета
- Дата и время прибытия
- Страна, город и аэропорт вылета
- Страна, город и аэропорт прибытия
- Статус рейса

Last_Name	First_Name	PassengerMiddleName	BirthDate	Passenger_Sex	PassengerDocument
NIKOLSKII	NIKOLAY	IGOREVICH	1990-12-26 00:00:00	Male	4396 926588
RUMIANTSEV	EGOR	EVGENEVICH	1972-04-01 00:00:00	Male	7536 277407
CHERNIAEV	DENIS	ARTEMICH	1993-10-25 00:00:00	Male	6793 521613
KOLTSOV	ARTUR	FILIPPOVICH	1976-12-15 00:00:00	Male	0383 434647
LUKIN	RAMIL	ADELEVICH	2000-05-12 00:00:00	Male	4816 776333

NumberFlight	TicketNumber	Fare	TravelClass	Meal	ProgrammNumber	bonusprogramm	Status	Baggage
SU1180	6247422701565929.0	YFLXPG		Y	SU 183142068	Aeroflot Bonus	Elite+	
SU1217	9764492390857976.0	AFLXWG		A	SU 441106611		NaN	Basic
SU1461	9078250945121430.0	YGRPKV		Y NLML	FB 447978679	Flying Blue	Basic	
SU1392	982769590416415.0	YRSTD A		Y HNML	DT 179843508		NaN	Elite
SU1204	9626077660344096.0	YRSTBJ		Y	KE 56301046		NaN	Basic

AirportDeparture	DepartureCountry	DepartureCity	DepartDate	DepartureTime	ArrivalAirport	ArrivalCountry	CityArrival	ArrivalDate	ArrivalTime
SVO	Russian Federation	MOSCOW	2017-03-18	22:10	VOG	Russian Federation	Volgograd	2017-03-19	01:05
KUF	Russian Federation	SAMARA	2017-01-08	06:05	SVO	Russian Federation	Moscow	2017-01-08	06:55
OVB	Russian Federation	NOVOSIBIRSK	2017-01-10	18:30	SVO	Russian Federation	Moscow	2017-01-10	19:00
SVO	Russian Federation	MOSCOW	2017-02-27	11:50	PEE	Russian Federation	Perm	2017-02-27	16:00
SVO	Russian Federation	MOSCOW	2017-02-27	17:00	PEE	Russian Federation	Perm	2017-02-27	21:10



**ТЕПЕРЬ РАССМОТРИМ
ГИПОТЕЗЫ**

Частые полеты в 1 страну

- Шпионы могут работать в одной стране
- Таких случаев мало: 2 человека летали 5 раз
- Учитываем полеты > 3 раз

```
In [52]: country_count2[(country_count2['count'] > 3)]
```

```
Out[52]:
```

Last_Name	First_Name	PassengerDocument	AirportDeparture	count
DAVYDOVA	IRINA	8008 568039	SVO	5
DIAKOVA	MARINA	0931 650989	SVO	4
EFIMOV	VLADIMIR	0746 284653	SVO	4
EFREMOVA	KAROLINA	7404 215541	SVO	4
FEDOTOV	ROMAN	9433 215493	SVO	4
FEDOTOVA	KSENIIA	6573 323697	SVO	4
HUDIAKOV	MATVEI	9105 895403	KHV	4
KOCHERGIN	MATVEI	9170 338460	SVO	4
LAPINA	OLESIA	1064 183397	SVO	4
MASLOV	AMIR	8816 864475	SVO	5
PETROV	ARSEN	2891 687634	SVO	4
SOKOLOV	TIHON	7366 309115	SVO	4
SOKOLOVA	NATALIA	7814 876912	KHV	4
VESNIAKOVA	ALIIA	7540 684400	KHV	4
VLASOVA	NELLI	2441 166340	SVO	4
ZHAROVA	ALINA	4887 850705	KHV	4

Попутчики

Люди, не имеющие родственных связей, которые летают вместе, могут быть шпионами

Таких обнаружено двое

PassengerLastName	PassengerFirstName	PassengerSecondName	PassengerDocument	FlightNumber	FlightDate	FlightTime	Destination	
35287	GORELOV	LEV	S.	7473 879825	SU1145	2017-01-10	08:00	Moscow
35289	LEONOV	ANDREI	MAKSIMOVICH	2027 454777	SU1145	2017-01-10	08:00	Moscow
72454	LEONOV	ANDREI	MAKSIMOVICH	2027 454777	SU1383	2017-01-04	06:40	Moscow
72475	GORELOV	LEV	SAVVOVICH	7473 879825	SU1383	2017-01-04	06:40	Moscow
PassengerLastName	PassengerFirstName	PassengerSecondName	PassengerDocument	FlightNumber	FlightDate	FlightTime	Destination	
34047	LEONOV	ANDREI	MAKSIMOVICH	2027 454777	SU1144	2017-01-08	17:15	Anapa
151810	GORELOV	LEV	SAVVOVICH	7473 879825	SU1144	2017-01-05	17:15	Anapa

Несколько билетов в 1 день

- Способ запутать возможных наблюдателей
- Возможны полеты по разным документам
- Резкая смена маршрута

	Passengers	time
0	GLADKOVA REGINA M.	2017-02-14
1	PANTELEEV GRIGORII A.	2017-01-01
2	NOVIKOVA NATASHA V.	2017-03-02
3	BASOVA VITALINA A.	2017-01-18
4	BORODINA IANA A.	2017-03-08
5	BULATOVA LILIIA D.	2017-01-02
6	MUKHINA NATALIIA SAVELEVNA	2017-01-09
7	ZAVIALOVA OLGA I.	2017-01-23

Нет данных по документам

- Билеты приобретены по особым каналам
- Часть информации по полетам может быть удалена по запросу органов
- У 3 человек постоянно нет паспортных данных

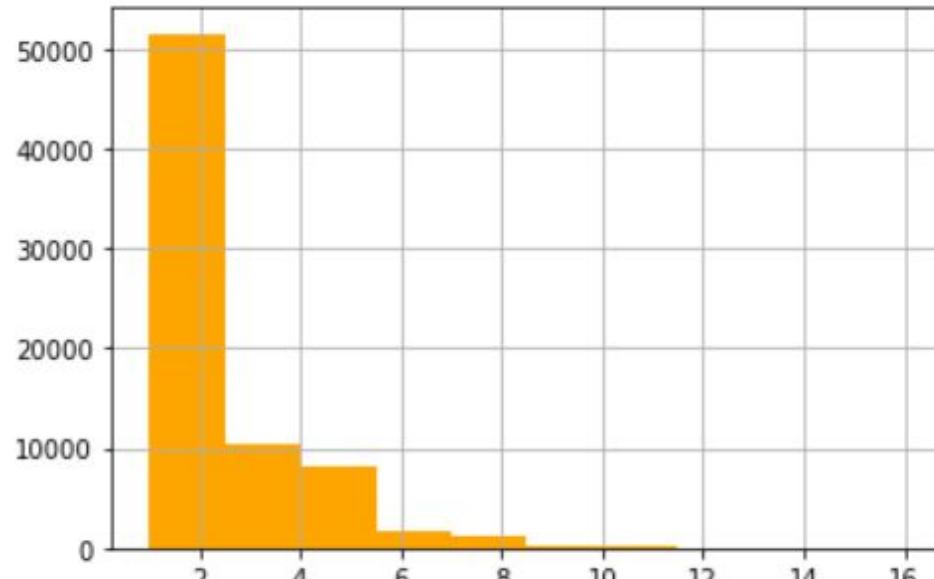
```
Ввод [76]: empty_values_data.value_counts()
```

```
Out[76]: 3    1492
          4    1481
          5    1432
          2    1252
          6    1204
          7    987
          8    801
          1    791
          9    528
         10   386
         11   262
         12   160
         13   95
         14   63
         15   38
         16   24
         17    9
         18    8
         19    6
         21    2
         20    2
         23    1
```

```
Name: First__Name, dtype: int64
```

Нет карты лояльности при частых полетах

- Политика запрещает использовать бонусные карты
- Человек не оформляет покупки самостоятельно
- Таких нашлось <100 человек



PassengerDocument	NumberFlight	Last_Name	First_Name	PassengerMiddleName
23913	1548 113497	16	BARANOVA	MAIIA ANATOLEVNA
23919	1548 113497	16	BARANOVA	MAIIA A.
81601	5293 500602	14	NAUMOVA	KARINA VLADISLAVOVNA
81604	5293 500602	14	NAUMOVA	KARINA V.
109710	7126 372174	15	GONCHAROV	TIHON ROSTISLAVOVICH

Дополнительные гипотезы

Признаками шпионов также могут быть:

- Поездки в страны, в которых происходят особые события
- Поездки без багажа
- Запутанные маршруты
- Преследование определенных пассажиров
- Несколько паспортов
- Распространенное имя

СПАСИБО ЗА
ВНИМАНИЕ

