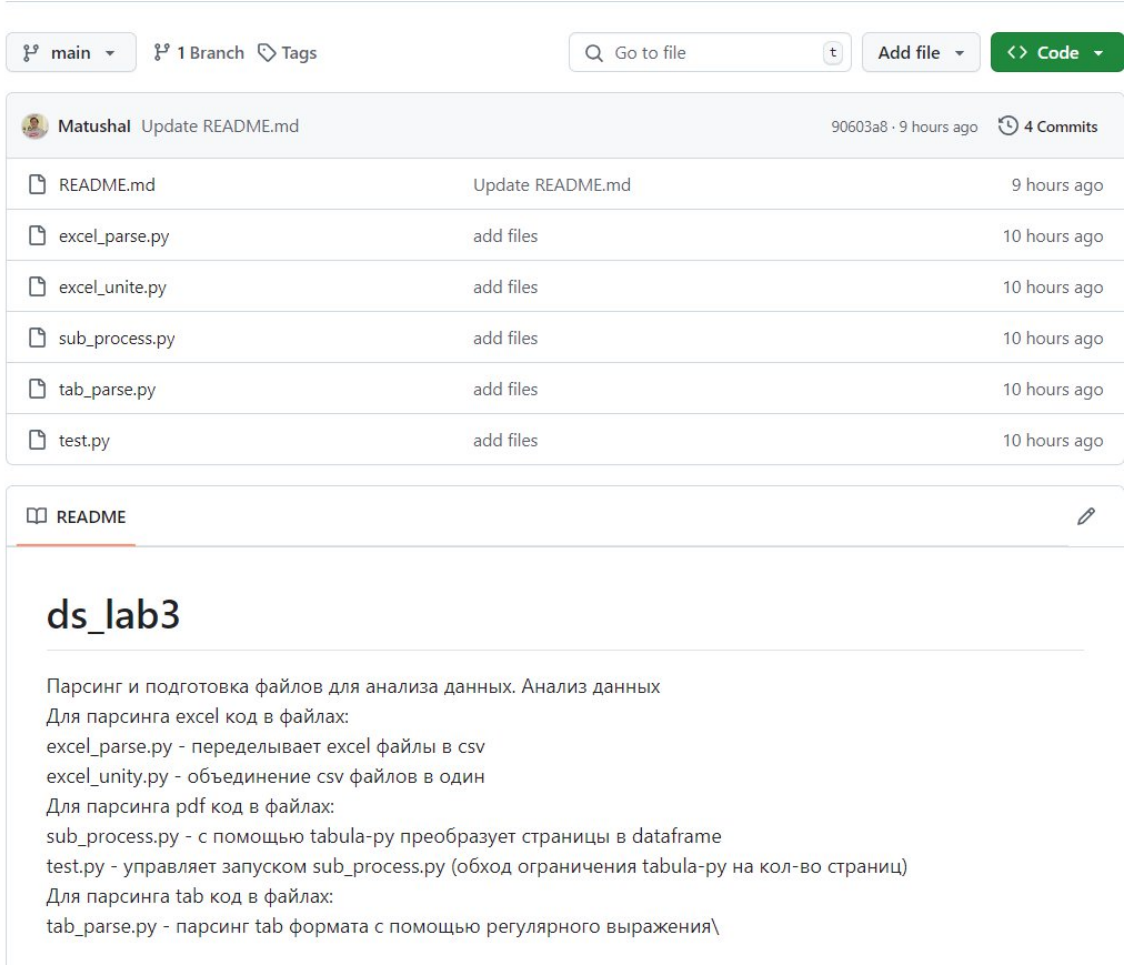
Catch the Spies

Галимьянов Матвей С22-501
Смирнов Илья С22-501
Скалин Иван Б22-513
Дистанов Марат Б22-513

Преподаватель - Дюмин А.А

Репозиторий с материалами



https://github.com/Matushal/ds_lab3

Работа с данными

1. Сделать обзор исходных данных
2. Перевести данные в один формат
3. Очистить данные
4. Подготовить таблицы для выполнения запросов
5. Провести анализ
6. Визуализировать

Исходные данные:

Название	Формат	Тип информации
YourBoardingPassDotAero	*zip(xlsx)*	Посадочные талоны пассажиров
SkyTeam_Exchange	*yaml*	О статусах полетов
Skyteam_Timetable	*pdf*	Расписание полетов в периоде 01 NOV 2018 – 30 JAN 2019
Sirena-export-fixed	*tab*	Данные из сервиса брони билетов
PointzAggregator-AirlinesData	*xml*	Сведения об участниках лояльности авиакомпании
FrequentFlyerForum-Profiles	*json*	Сведения об участниках лояльности авиакомпании
BoardingData	*csv*	Данные о посадке на самолет

Исходные данные:

- разных форматов;
- из разных источников;
- большие;
- «грязные»;
- требуют комплексного анализа

Парсинг данных

Стек технологий:

- **pandas,**
- **tabula-py,**
- **re**

осуществлялся перевод в формат .csv с необходимыми преобразованиями;

```python
import pandas as pd
csv_files = ["skyteam_part1.csv","skyteam_part2.csv","skyteam_part3.csv","skyteam_part4.csv"]
combined_df = pd.concat([pd.read_csv("skyteam_part1.csv", sep=','),
                         pd.read_csv("skyteam_part2.csv", sep=','),
                         pd.read_csv("skyteam_part3.csv", sep=','),
                         pd.read_csv("skyteam_part4.csv", sep=',')])
combined_df.head(1000)
days_names = ['SundayDeparture', 'MondayDeparture', 'TuesdayDeparture', 'WednesdayDeparture',
       'ThursdayDeparture', 'FridayDeparture', 'SaturdayDeparture']
combined_df[days_names] = False
for i, day in enumerate(days_names):
  combined_df[day] = combined_df['1'].str.contains(str(i + 1))
```

```python
combined_df[['Hours', 'Minutes']] = combined_df['6'].str.extract(r'(\d+)H(\d+)M')

combined_df['FlightTimeMinutes'] = combined_df['Hours'].fillna(0).astype(int) * 60 +
combined_df['Minutes'].fillna(0).astype(int)

combined_df.drop(columns=['Hours', 'Minutes'], inplace=True)
combined_df.to_csv("SKYTEAM.csv", index=False)
```

переход от формата 1H25M - времени полета к времени полета в минутах и выгрузка в формате csv

Таблица SKYTEAM расписание перелётов за период 01 NOV 2018 - 30 JAN 2019

Departure	Arrival	DepartureTime	ArrivalTime	FlightNumber	AircraftNumber	SundayDeparture	MondayDeparture	TuesdayDeparture	WednesdayDeparture	ThursdayDeparture	FridayDeparture	SaturdayDeparture	StartOfFlightPeriod	EndOfFlightPeriod	FlightTimeMinutes
AAL	AMS	06:00	07:25	KL1328	73W	True	True	True	True	True	True	True	01 Nov	31 Jan	85
AAL	AMS	12:10	13:35	KL1334	73W	True	True	True	True	True	True	True	01 Nov	31 Jan	85
AAL	AMS	18:15	19:35	KL1336	EQV	True	True	True	True	True	True	True	01 Nov	23 Dec	80
AAL	AMS	18:15	19:35	KL1336	EQV	True	False	True	True	True	True	True	26 Dec	06 Jan	80
AAL	AMS	18:15	19:35	KL1336	EQV	True	True	True	True	True	True	True	07 Jan	31 Jan	80
AAL	AMS	06:00	07:25	KL1328	73W	True	True	True	True	True	True	True	01 Nov	31 Jan	85
AAL	AMS	12:10	13:35	KL1334	73W	True	True	True	True	True	True	True	01 Nov	31 Jan	85

Таблица Sirena-export-fixed

	FullName	BirthDate	DepartDate	DepartTime	ArrivalDate	ArrivalTime	Flight	From	To	E-Ticket	Document	Meal	Class	LoyaltyP	Service
0	ОЗЕРОВ ИЛЬДАР ДАНИИЛОВИЧ	1999-05-15	2017-05-30	00:05	2017-05-30	08:05	SU1306	SVO	OVB	7360415302044672	9375 053270		J	SU 38116280	Go2See
1	КОЛОСОВ САМИР ТАМЕРЛАНОВИЧ	N/A	2017-12-27	02:15	2017-12-27	04:40	SU1323	MMK	SVO	7398421117936516	2244 645520	KSML	Y		
2	ИГНАТОВА СНЕЖАНА КОНСТАНТИНОВНА	N/A	2017-09-19	06:40	2017-09-19	07:45	SU1481	KJA	SVO	5174973140468001	8115 961316		Y		KupiBilet
3	ЖАРОВ ПЛАТОН АЛЬБЕРТОВИЧ	1999-05-02	2017-03-18	22:10	2017-03-19	01:05	SU1180	SVO	VOG	5274206497242737	98 6865148		J	FB 884556993	Travelgenio
4	НИКОЛЬСКИЙ НИКОЛАЙ ИГОРЕВИЧ	1990-12-26	2017-03-18	22:10	2017-03-19	01:05	SU1180	SVO	VOG	6247422701565929	4396 926588		Y	SU 183142068	OZON.travel
5	ГЛУШКОВ КОНСТАНТИН ИЛЬИЧ	N/A	2017-03-12	11:45	2017-03-12	12:25	SU6284	UUS	SVO	5874178506968181	4788 422492		Y	FB 553284496	KupiBilet
6	КАПУСТИН АРТЁМ ЭДУАРДОВИЧ	1982-10-24	2017-03-12	11:45	2017-03-12	12:25	SU6284	UUS	SVO	7467749130398378	0058 142289	VLML	Y		KupiBilet
7	ЕРШОВА ЛЮБОВЬ ЗАХАРОВНА	N/A	2017-03-12	11:45	2017-03-12	12:25	SU6284	UUS	SVO	2183161939566868	0776 380126		Y		Aeroflot

Таблица BoardingData

PassengerFirstName	PassengerSecondName	PassengerLastName	PassengerSex	PassengerBirthDate	PassengerDocument	BookingCode	TicketNumber	Baggage	FlightDate	FlightTime	FlightNumber	CodeShare	Destination
SAVELII	VIKTOROVICH	RUSANOV	Male	03/10/1983	2879 096860	FRNINO	6625956945991971	Transit	2017-03-22	06:05	SU1369	Own	Moscow
LEV	MARKOVICH	ISAEV	Male	12/13/1975	1788 173211	Not presented	1643715499224676	Registered	2017-03-18	22:10	SU1180	Own	Volgograd
NIKOLAI	I.	NIKOLSKII	Male	12/26/1990	4396 926588	VWNYGF	6247422701565929	Transit	2017-03-18	22:10	SU1180	Own	Volgograd
ANATOLII	PETROVICH	SHILOV	Male	05/24/1997	2595 919752	WQFFUE	Not presented	Registered	2017-03-18	22:10	SU1180	Own	Volgograd
MIROSLAVA	VIACHESLAVOVNA	SEMENOVA	Female	01/31/1976	6775 516990	Not presented	Not presented	Registered	2017-03-12	11:45	SU6284	Own	Moscow
ARTEM	EDUARDOVICH	KAPUSTIN	Male	10/24/1982	0058 142289	JJADFB	7467749130398378	Transit	2017-03-12	11:45	SU6284	Own	Moscow
ZARINA	E.	TITOVA	Female	12/22/2000	6600 251370	Not presented	5954073786122008	None	2017-07-29	14:15	SU1281	Own	Moscow
SAMIR	GORDEEVICH	BARSUKOV	Male	02/21/1995	0078 271703	WGKZTB	2264717979478322	Registered	2017-07-29	14:15	SU1281	Own	Moscow

Таблица YourBoardingPassDotAero

	Flight	Name	Sex	Departure	Arrival	Date	Time	PNR	ETicket	LP	LP Number	Class
0	DL3539	GOLOVANOV DMITRII	GOLOVANOV DMITRII	MSP	BOI	2017-11-26	11:30	JCMKZW	762361762268625		0	Y
1	DL3903	YULIYA A LOGINOVA	YULIYA A LOGINOVA	JFK	ORD	2017-11-26	12:15	KAUNTH	7495554002459386	SU	535648188	J
2	DL4168	VLADIMIR ANTONOV	VLADIMIR ANTONOV	ELP	SLC	2017-11-26	08:00	OAMNCL	9280991095130410		0	Y
3	VN3504	MARKELOV MIROSLAV	MARKELOV MIROSLAV	CAN	SGN	2017-11-26	11:55	XENNWQ	7304106044464688		0	Y
4	SU6016	UTKINA AGATA B	UTKINA AGATA B	VKO	LED	2017-11-26	09:40	SHEZOP	3810434426712787	KE	407778420	Y
5	DL3402	FEDOR EVDOKIMOV	FEDOR EVDOKIMOV	IND	DTW	2017-11-26	08:00	MNDCQQ	9725668050676075		0	Y

Очистка данных

- **приведение к единому виду;**
- **удаление NaN, пустых полей и полей, имеющих смысл пустых**

ANATOLII	PETROVICH	SHILOV
MIROSLAVA	VIACHESLAVOVNA	SEMENOVA
ARTEM	EDUARDOVICH	KAPUSTIN

```
AgentInfo
ОЗЕРОВ ИЛЬДАР ДАНИИЛОВИЧ
ZBQSPY736041530204467 2  9375 053270 N/A
Go2See
КОЛОСОВ САМИР ТАМЕРЛАНОВИЧ
MHPBBX7398421117936516  2244 645520 N/A   KS
284903754
ИГНАТОВА СНЕЖАНА КОНСТАНТИНОВНА
REDLVB5174973140468001  8115 961316 N/A
KupiBilet
ЖАРОВ ПЛАТОН АЛЬБЕРТОВИЧ
NSJNGQ52742064972427 37  98 6865148  N/A
```

Схема БД

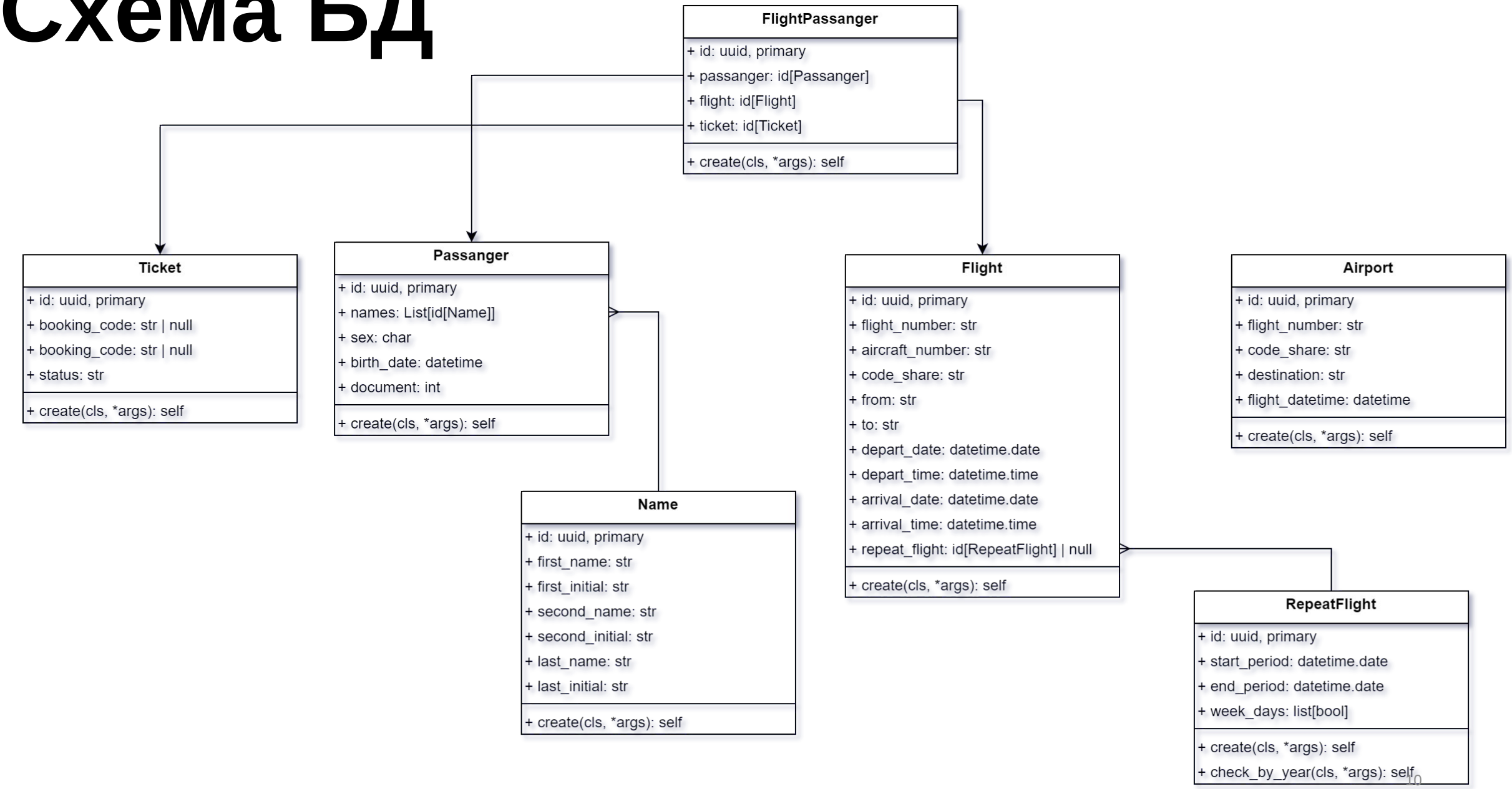

FlightPassanger
+ id: uuid, primary
+ passanger: id[Passanger]
+ flight: id[Flight]
+ ticket: id[Ticket]
+ create(cls, *args): self
Ticket
+ id: uuid, primary
+ booking_code: str | null
+ booking_code: str | null
+ status: str
+ create(cls, *args): self
Passanger
+ id: uuid, primary
+ names: List[id[Name]]
+ sex: char
+ birth_date: datetime
+ document: int
+ create(cls, *args): self
Name
+ id: uuid, primary
+ first_name: str
+ first_initial: str
+ second_name: str
+ second_initial: str
+ last_name: str
+ last_initial: str
+ create(cls, *args): self
Flight
+ id: uuid, primary
+ flight_number: str
+ aircraft_number: str
+ code_share: str
+ from: str
+ to: str
+ depart_date: datetime.date
+ depart_time: datetime.time
+ arrival_date: datetime.date
+ arrival_time: datetime.time
+ repeat_flight: id[RepeatFlight] | null
+ create(cls, *args): self
Airport
+ id: uuid, primary
+ flight_number: str
+ code_share: str
+ destination: str
+ flight_datetime: datetime
+ create(cls, *args): self
RepeatFlight
+ id: uuid, primary
+ start_period: datetime.date
+ end_period: datetime.date
+ week_days: list[bool]
+ create(cls, *args): self
+ check_by_year(cls, *args): self

Таблица рейсов

Откуда - Куда - Номер рейса

- файл `Skyteam_Timetable.pdf` содержит расписание перелетов за период ноябрь 2018 - январь 2019;
- т.к. в других источниках описываются другие периоды времени, полезными являются только номера рейсов и места перелетов:

SKYTEAM_clear

Departure	Arrival	FlightNumber
AAL	AMS	KL1328
AAL	AMS	KL1334
AAL	AMS	KL1336
AES	AMS	KL1326
ABA	SVO	SU1479
ABZ	AMS	KL1440
ABZ	AMS	KL1442
ABZ	AMS	KL1444
ABZ	AMS	KL1448
ABZ	CDG	AF1473
ABZ	CDG	KQ3747
ABZ	CDG	AF1273
ABZ	CDG	KQ3749
ABR	MSP	DL7365

Подозрительные перелеты:

пассажиры, которые часто летают в разные страны за короткий период времени

пассажиры с несколькими документами на одно имя или с одинаковыми паспортными данными

пассажиры, которые часто покупают билеты в последний момент или меняют свои планы, могут быть под подозрением

Выводы

Крайне необходимо:

1. Тщательнее подходить к выбору команды (накладки в расписаниях);
2. Уметь координировать работу группы и декомпозировать задачу;
3. Уметь рассчитывать время выполнения подзадач

Поиск шпионов:

1. Рассмотрены шаги цикла обработки данных
2. Подготовлены .csv таблицы с данными о пассажирах и рейсах

Дальнейшая аналитика:

1. Проверить соответствие посадочных талонов расписанию;
2. Найти подозрительные цепочки перемещений через посадочные талоны;