Шувалов В.М., Николаев А.А., Кузнецов М.А.

Отчёт по полусеместровому проекту «Petrov&Boshirov»

1. Преобразование данных

Исходные данные предоставлены в разных форматах (csv, json, xls, yaml), а для обработки данных была выбрана библиотека pandas для Python. Поэтому сначала необходимо преобразовать все данные к одному виду, удобному для анализа в pandas.

Для этого были разработаны подпрограммы, переводящие данные из различных форматов в формат csv (ExcelParser.py, JsonParser.py, XMLParser.py, YAMLParser.py). Данные подпрограммы были запущены по отдельности и с их помощью созданы файлы с данными формата csv.

2. Загрузка и обработка данных

Далее полученные данные были загружены в виде датафрейма pandas. В первую очередь использовались данные из файлов BoardingData.csv и PointzAggregator-AirlinesData.xml в которых находится информация о посадочных талонах пассажиров и о бонусных картах соответственно.

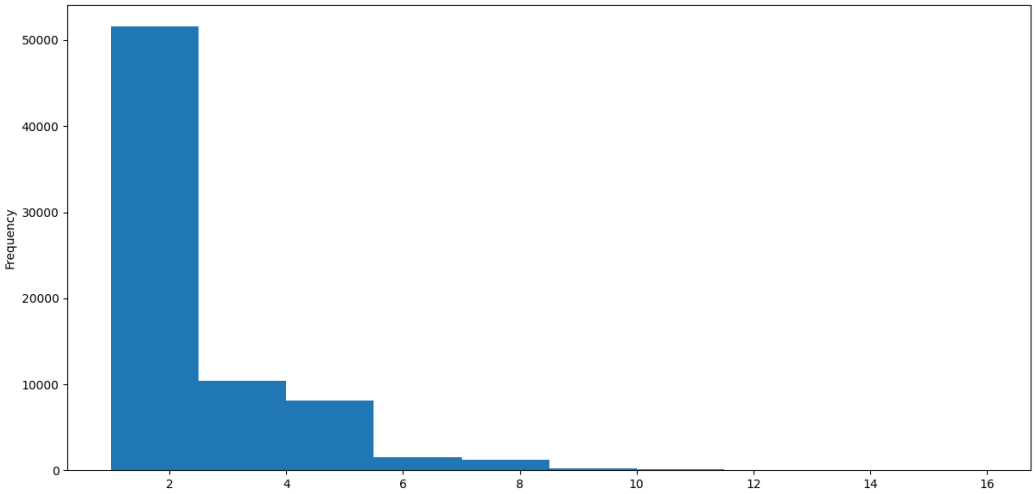
Пассажиры идентифицировались по номеру документа (паспорта). К информации из посадных талонов была добавлена информация об использовании бонусных карт этим пассажиром на этом рейсе. Сначала была получена общая информация о полётах – статистические данные кол-ва пассажиров и кол-ва полётов у каждого пассажира, а также кол-во полетов в определённые города:

```python
csv_data.rename(columns={'PassengerFirstName': 'First name', 'PassengerLastName': 'Second name', 'PassengerSecondName': 'Middle name'}, inplace=True)
xml_data.rename(columns={'Code': 'FlightNumber', 'Date': 'FlightDate'}, inplace=True)

boarding_data_with_bonus = csv_data.merge(xml_data[['First name', 'Second name', 'Card number', 'FlightNumber', 'FlightDate']], on=['First name', 'Second
#boarding_data_with_bonus = boarding_data_with_bonus.merge(json_data[['NickName', 'First name', 'Second name', 'Card number', 'FlightNumber', 'FlightDate'
#boarding_data_with_bonus = boarding_data_with_bonus.merge(json_data[['NickName', 'First name', 'Second name', 'Card number', 'FlightNumber', 'FlightDate'
#print(boarding_data_with_bonus['Card number'].isna().sum())
#print(boarding_data_with_bonus['NickName'].isna().sum())
#print(boarding_data_with_bonus.head(30))

#print(boarding_data_with_bonus['PassengerDocument'].unique())

unique_docs = boarding_data_with_bonus['PassengerDocument'].value_counts()
#print(unique_docs)
people_list = pd.DataFrame(columns=['First name', 'Second name', 'PassengerDocument', 'Number of flights', 'Number of bonus use', 'Percent of bonus use'])
people_list = people_list.astype({"Number of flights": "float", "Number of bonus use": "float", "Percent of bonus use": "float"})
#print(people_list.info())
pre_stats = unique_docs.describe()
unique_docs.plot(kind='hist')
plt.show()
print('######################################################################')
print(pre_stats)
print('######################################################################')
print(csv_data['Destination'].value_counts())
```




```
count    73263.000000
mean         2.117672
std          1.567211
min          1.000000
25%          1.000000
50%          1.000000
75%          3.000000
max         16.000000
```

Moscow	88022
Khabarovsk	11142
Vladivostok	7243
Yuzhno	4371
Sochi	2740
Yekaterinburg	2306
Kazan	2134
Krasnodar	1849
Ufa	1836
Rostov	1749
Petropavlovsk	1683
Samara	1574
Novosibirsk	1526
Irkutsk	1513
Krasnoyarsk	1496
Chelyabinsk	1432
Perm	1428
Volgograd	1383
Nizhnekamsk	1365
Kaliningrad	1184
Okha	1062
Murmansk	938

Была выдвинута гипотеза о том, что потенциальные шпионы совершают много перелётов за определённый период. Поэтому из списка пассажиров были выбраны люди с наибольшим количеством полётов. Создан новый датафрейм, содержащий основные данные о пассажире, количество его полётов, количество использований бонусных карт и процент полётов с использованием карты.

```
for psg in boarding_data_with_bonus['PassengerDocument'].unique():
    if unique_docs[psg] > pre_stats.at['max'] * 0.75:
        buf_df = boarding_data_with_bonus[boarding_data_with_bonus['PassengerDocument'] == psg]
        cards = buf_df['Card number'].notna().sum()
        buf_df = buf_df.reset_index()
        new_row = {'First name': buf_df.at[0, 'First name'], 'Second name': buf_df.at[0, 'Second name'], 'PassengerDocument': psg, 'Numb
        people_list = people_list.append(new_row, ignore_index=True)
        #print('---Done---', (i / len(boarding_data_with_bonus['PassengerDocument'].unique())) * 100, '%')
        #i += 1
    else:
        pass
```

	First name	Second name	PassengerDocument	Number of flights	Number of bonus use	Percent of bonus use
0	AMIR	MASLOV	8816 864475	14.0	5.0	0.357143
1	EMILIIA	KUPRIIANOVA	2956 495871	13.0	3.0	0.230769
2	VERA	KOLPAKOVA	0185 106233	13.0	0.0	0.000000
3	DAMIR	NIKONOV	9154 853684	13.0	5.0	0.384615
4	INNA	LUKIANOVA	2788 100077	13.0	0.0	0.000000
5	SVIATOGOR	CHERNOV	6116 437997	13.0	0.0	0.000000
6	SNEZHANA	FETISOVA	7255 946452	13.0	5.0	0.384615
7	MIKHAIL	CHEBOTAREV	6937 933652	14.0	4.0	0.285714
8	MARINA	DIAKOVA	0931 650989	13.0	0.0	0.000000
9	KARINA	NAUMOVA	5293 500602	14.0	0.0	0.000000
10	TIKHON	GONCHAROV	7126 372174	15.0	3.0	0.200000
11	MATVEI	DOROFEEV	5462 105040	13.0	12.0	0.923077
12	AMALIIA	KOZLOVSKAIA	6435 081994	13.0	0.0	0.000000
13	MAIIA	BARANOVA	1548 113497	16.0	0.0	0.000000
14	TIKHON	ROZHKOV	7933 752359	13.0	5.0	0.384615
15	ALEKSANDRA	EREMEEVA	2294 770067	13.0	0.0	0.000000

Далее была выдвинута следующая гипотеза: шпионы не будут часто использовать бонусную карту чтобы не оставлять о себе лишнюю информацию. Получена статистика использования карт оставшимися пассажирами.

	Number of flights	Number of bonus use	Percent of bonus use
count	16.000000	16.000000	16.000000
mean	13.500000	2.625000	0.196909
std	0.894427	3.344149	0.255333
min	13.000000	0.000000	0.000000
25%	13.000000	0.000000	0.000000
50%	13.000000	1.500000	0.100000
75%	14.000000	5.000000	0.364011
max	16.000000	12.000000	0.923077

Фильтрация пассажиров выполнена по нижнему квартилю процента полётов с картами (0 %).

```
bonus_stat = stats.at['25%', 'Percent of bonus use']
spies_list = people_list[people_list['Percent of bonus use'] <= bonus_stat]
```

Список оставшихся подозреваемых:

	First name	Second name	PassengerDocument
2	VERA	KOLPAKOVA	0185 106233
4	INNA	LUKIANOVA	2788 100077
5	SVIATOGOR	CHERNOV	6116 437997
8	MARINA	DIAKOVA	0931 650989
9	KARINA	NAUMOVA	5293 500602
12	AMALIIA	KOZLOVSKAIA	6435 081994
13	MAIIA	BARANOVA	1548 113497
15	ALEKSANDRA	EREMEEVA	2294 770067

Далее выдвинута гипотеза по возрасту. Возраст шпиона может быть от 21 до 50 лет. Соответственно пассажиры были отфильтрованы по данному показателю.

```
csv_data = csv_data.loc[csv_data['PassengerDocument'].isin(spies_list['PassengerDocument'].tolist())]

print('########################################################################')
print(spies_list)
print('########################################################################')
for item in csv_data['PassengerBirthDate'].unique():
    print(item)

csv_data = csv_data.loc[~csv_data['PassengerDocument'].isin(['0931 650989', '6435 081994'])]
spies_list = spies_list.loc[~spies_list['PassengerDocument'].isin(['0931 650989', '6435 081994'])]

print('########################################################################')
print(spies_list)
spies_list = spies_list.reset_index()
```

Список дат рождения и оставшихся подозреваемых:

```
04/22/1983
10/14/1993
11/14/1989
10/05/1999
01/15/1989
08/06/1997
05/22/1970
10/28/1990
##############################################
    First name Second name PassengerDocument
2         VERA   KOLPAKOVA       0185 106233
4         INNA   LUKIANOVA       2788 100077
5    SVIATOGOR     CHERNOV       6116 437997
9       KARINA     NAUMOVA       5293 500602
13       MAIIA    BARANOVA       1548 113497
15  ALEKSANDRA    EREMEEVA       2294 770067
```

Далее был получен список перелётов для всех подозреваемых. Видно что подозреваемые совершают перелёты примерно в одном регионе примерно в одни и те же даты, что наводит на мысли об их совместной шпионской деятельности.

################ EREMEEVA ################

	FlightDate	FlightNumber	Destination
43815	2017-02-26	SU1481	Moscow
30361	2017-06-05	SU5617	Vladivostok
56497	2017-06-07	SU5602	Khabarovsk
21580	2017-06-09	SU5626	Yuzhno
109546	2017-06-10	SU5621	Khabarovsk
63819	2017-08-20	SU4599	Nogliki
147825	2017-09-06	SU4600	Khabarovsk
22392	2017-09-07	SU5612	Magadan
42508	2017-09-13	SU5613	Khabarovsk
70054	2017-12-08	SU5601	Vladivostok
39583	2017-12-14	SU4130	Yuzhno
31806	2017-12-19	SU5625	Khabarovsk
123819	2017-12-27	SU5684	Petropavlovsk

################ CHERNOV ################

	FlightDate	FlightNumber	Destination
106433	2017-01-07	SU1459	Moscow
27530	2017-03-18	SU1730	Petropavlovsk
7837	2017-03-25	SU4599	Nogliki
18208	2017-04-12	SU4600	Khabarovsk
61957	2017-05-02	SU5684	Petropavlovsk
18810	2017-05-04	SU5685	Khabarovsk
4252	2017-05-16	SU5612	Magadan
143544	2017-05-23	SU5613	Khabarovsk
129528	2017-07-10	SU5612	Magadan
67829	2017-07-20	SU5613	Khabarovsk
18784	2017-07-21	SU5624	Yuzhno
14363	2017-08-05	SU5621	Khabarovsk
69298	2017-08-31	SU6288	Moscow

################ BARANOVA ################

	FlightDate	FlightNumber	Destination
25735	2017-01-12	SU5613	Khabarovsk
24204	2017-01-16	SU4605	Okha
132891	2017-02-19	SU4606	Khabarovsk
19436	2017-03-04	SU5640	Krasnoyarsk
104310	2017-03-05	SU1487	Moscow
15398	2017-04-04	SU6289	Magadan
84011	2017-04-08	SU5613	Khabarovsk
129271	2017-05-02	SU4606	Okha
25524	2017-05-04	SU4606	Khabarovsk
52172	2017-05-25	SU4605	Okha
73932	2017-05-29	SU4606	Khabarovsk
136632	2017-06-19	SU5620	Yuzhno
37832	2017-07-05	SU5607	Vladivostok
150317	2017-07-17	SU5602	Khabarovsk
51407	2017-07-21	SU5619	Vladivostok
58610	2017-07-26	SU6282	Moscow

################ LUKIANOVA ################

	FlightDate	FlightNumber	Destination
4170	2017-01-10	SU4606	Khabarovsk
119820	2017-02-05	SU5620	Yuzhno
93781	2017-02-08	SU5623	Khabarovsk
53641	2017-03-17	SU4605	Okha
74983	2017-03-20	SU4606	Khabarovsk
93716	2017-03-22	SU5626	Yuzhno
68371	2017-05-01	SU4131	Vladivostok
38335	2017-06-20	SU4130	Yuzhno
119873	2017-07-04	SU5611	Vladivostok
36835	2017-09-18	SU4130	Yuzhno
63325	2017-09-25	SU5627	Khabarovsk
50872	2017-09-26	SU5619	Vladivostok
133480	2017-10-17	SU5625	Khabarovsk

################ KOLPAKOVA ################

	FlightDate	FlightNumber	Destination
88103	2017-01-05	SU4600	Khabarovsk
116539	2017-01-19	SU5620	Yuzhno
54006	2017-02-23	SU5614	Petropavlovsk
2832	2017-03-09	SU5685	Khabarovsk
15886	2017-04-12	SU4605	Okha
79740	2017-05-05	SU4606	Khabarovsk
96113	2017-06-25	SU5622	Yuzhno
42458	2017-07-11	SU5621	Khabarovsk
55472	2017-07-31	SU5628	Yuzhno
149972	2017-08-17	SU5623	Khabarovsk
93372	2017-10-27	SU4601	Komsomolsk
6457	2017-12-03	SU4602	Vladivostok
145115	2018-01-01	SU4602	Vladivostok

################ NAUMOVA ################

	FlightDate	FlightNumber	Destination
43629	2017-01-25	SU5612	Magadan
116052	2017-02-19	SU5613	Khabarovsk
18767	2017-03-15	SU4606	Okha
89779	2017-03-16	SU4606	Khabarovsk
13274	2017-03-25	SU4605	Okha
99431	2017-04-01	SU4606	Khabarovsk
64867	2017-04-15	SU5624	Yuzhno
12252	2017-06-12	SU5621	Khabarovsk
139561	2017-06-22	SU4599	Nogliki
115942	2017-06-25	SU4600	Khabarovsk
20892	2017-07-09	SU5615	Vladivostok
38545	2017-10-13	SU5626	Yuzhno
45519	2017-10-17	SU5611	Vladivostok
6651	2017-11-24	SU5614	Petropavlovsk

Также получен список городов, в которые совершались перелёты подозреваемыми, в котором остались города, которые не были популярны в изначальном списке. То есть остались определенные города из одного региона, а не самые популярные города из начального списка.

```
print(csv_data['Destination'].value_counts())
```

Moscow	88022
Khabarovsk	11142
Vladivostok	7243
Yuzhno	4371
Sochi	2740
Yekaterinburg	2306
Kazan	2134
Krasnodar	1849
Ufa	1836
Rostov	1749
Petropavlovsk	1683
Samara	1574
Novosibirsk	1526
Irkutsk	1513
Krasnoyarsk	1496
Chelyabinsk	1432
Perm	1428
Volgograd	1383
Nizhnekamsk	1365
Kaliningrad	1184
Okha	1062
Murmansk	938

Khabarovsk	31
Yuzhno	13
Vladivostok	11
Okha	7
Magadan	5
Petropavlovsk	5
Moscow	5
Nogliki	3
Krasnoyarsk	1
Komsomolsk	1

В конце также вручную была просмотрена информация о международных перелётах подозреваемых и использования карт лояльности на этих перелётах.