Petrov and Boshirov...

Ветлин Владислав, Пушкарев Кирилл, Сабуров Вильдан,
Талалай Михаил

Ноябрь 19, 2021

Описание задачи

- Имеются данные о полетах людей, список зарегистрированных на форуме, список получивших бонусы, расписание, билеты
- Данные в разных форматах
- Нужно по этим данным вычислить шпионов

Аналитическая задача

- Нет ответов, то есть нет данных о шпионах
- Задача поиска аномалий. Нужно найти нетипичное поведение и выделить несколько признаков.
- Нужно сформировать индекс подозрительности, приняв во внимание признаки
- По некому порогу определить людей, которых мы считаем за шпионов

Разделение задачи

Поскольку все данные в разных форматах, а файлов очень много, то задачу можно разбить на две большие части:

- Обработка данных
- Поиск шпионов на обработанных данных

Приведение файлов к csv

Самая сложная часть - pdf. Отметим основные моменты:

- Нужно аккуратно выбирать парсер, иначе он неправильно выставит переводы строк
- Двойные таблицы

Ещё довольно напросто с xls

Простые файлы

- Yaml, Xml. Для этих файлов существуют библиотеки в python, поэтому тут всё не очень сложно. Правда yaml достаточно большой, поэтому пришлось разбить на части

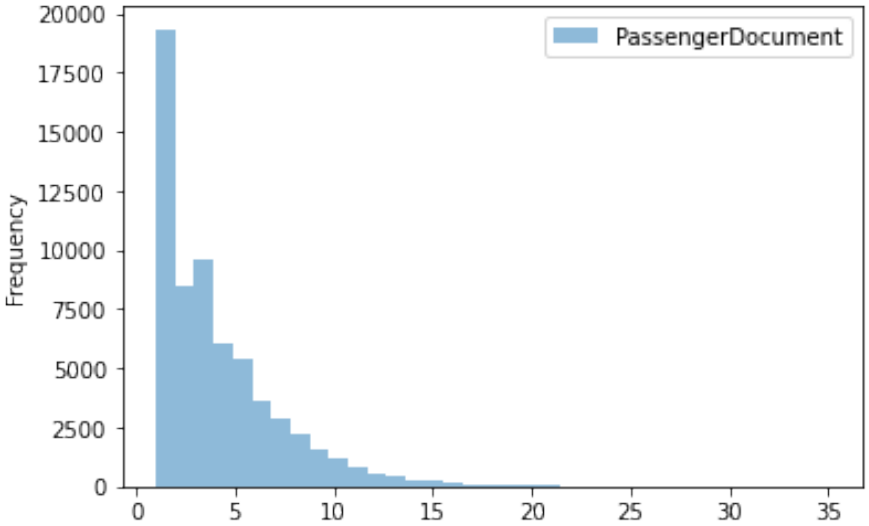
- Json, csv, tab. Это тоже простые в обработке файлы. Важно было сделать транслитерацию, чтобы их объединить

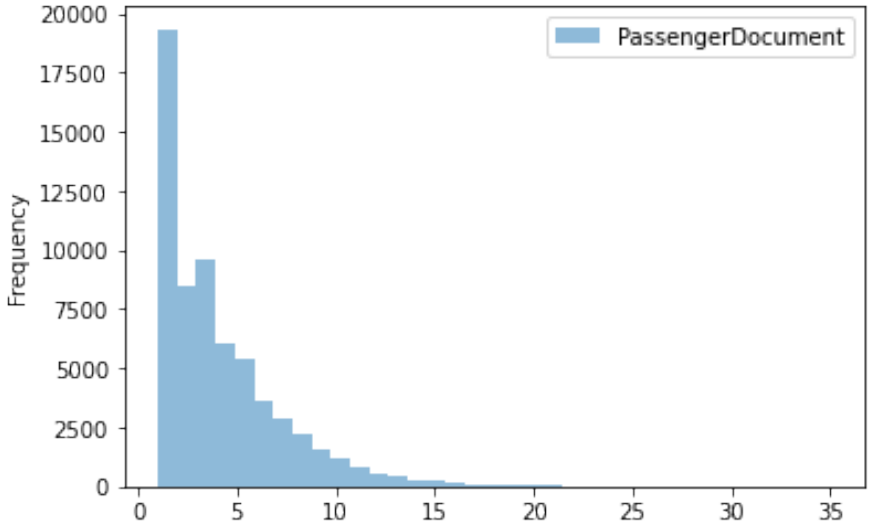
Обработка пропусков

Данные довольно грязные, в них есть несостыковки, поэтому надо обрабатывать пропуски и исправлять несостыковки. Есть 2 способа.

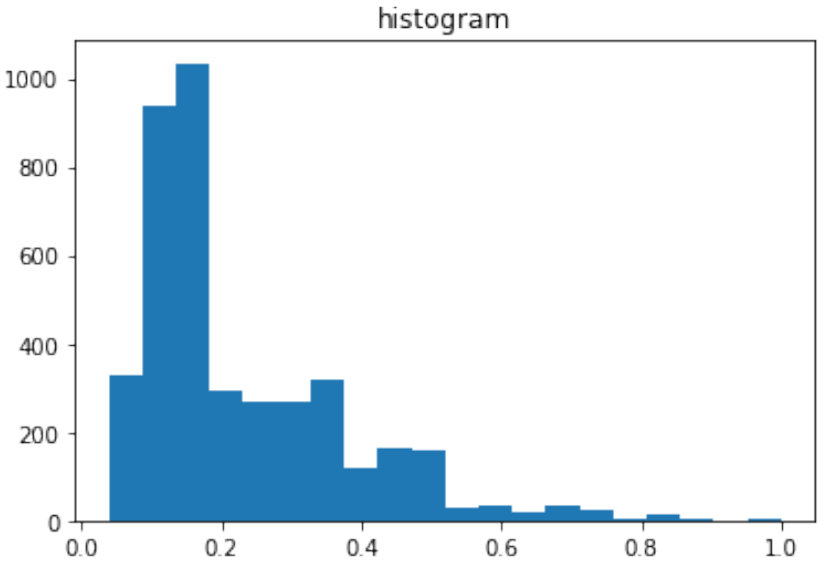
- С помощью расписания. Здесь просто смотрим совпадение основных полей
- С помощью билетов. Здесь проблема в несовпадении имён (во-первых учесть порядок, во-вторых разное написание)

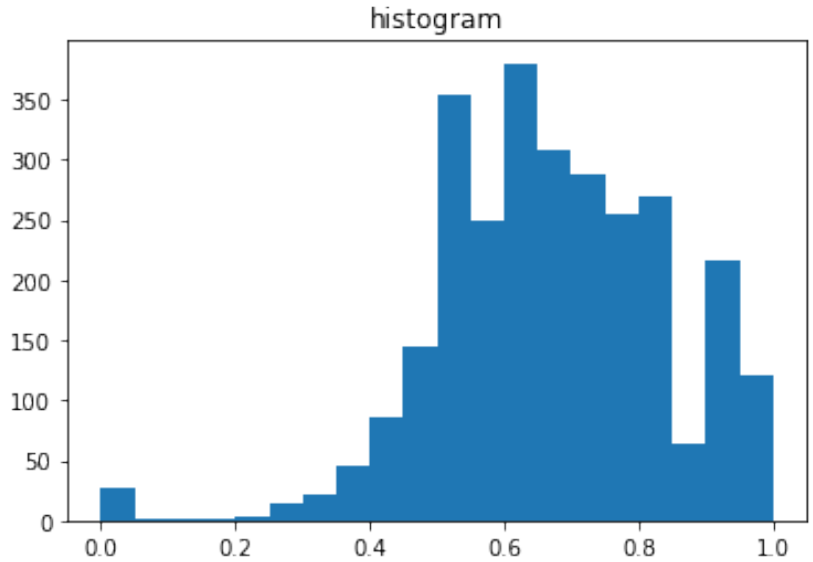
Простые гистограммы





Гистограммы с аномалиями





Критерии подозрительности

- Летает много, но по одному направлению
- Летает много, но не получает бонусы
- Летает много, но не зарегистрирован на форуме
- Часто пользуется заменой рейсов
- Летает в одну сторону
- Летает дорогим классом

Результаты

	A	B
1	Name	Score
2	SERGEI ERMILOV	1.7000000000000002
3	VLAD MALTSEV	1.6
4	DMITRII ANTONOV	1.6
5	LIANA PLOTNIKOVA	1.5
6	SAVVA DEGTIAREV	1.5
7	ARINA KLIMOVA	1.5
8	IAN VORONOV	1.5
9	SVIATOSLAV VOROBEV	1.5
10	IAROSLAV BESSONOV	1.5
11	ZAHAR MURATOV	1.5
12	GERMAN HOHLOV	1.5
13	LEV NIKITIN	1.5
14	SAVELII LAZAREV	1.5
15	ALINA RUSANOVA	1.5
16	STEFANIIA EREMINA	1.5
17	DEMID OSIPOV	1.5
18	EKATERINA MARKOVA	1.5
19	ADEL SIZOV	1.5
20	NELLI LVOVA	1.5
21	MARGARITA GLUHOVA	1.5
22	AIDAR ROZHKOV	1.5
23	SVIATOSLAV PAVLOV	1.5
24	ANDREI SHILOV	1.5
25	IANA KAPUSTINA	1.5
26	IGOR KOLTSOV	1.5
27	MIHAIL AFANASEV	1.5
28	SNEZHANA BELOUSOVA	1.5
29	ANATOLII KALININ	1.5
30	GEORGII ZHURAVLEV	1.5
31	OLGA ELIZAROVA	1.5
32	ZAHAR FILIMONOV	1.5
33	KSENIIA RAKOVA	1.5
34	ALISA BONDAREVA	1.5
35	NAZAR SYCHEV	1.5
36	VASILINA BLOHINA	1.5
37	EKATERINA SEMENOVA	1.2
38	ALBERT MININ	1.2