Визуализация данных “Schengen Visa Stats 2017/2018”

Курс “Прикладные задачи анализа данных 2019”

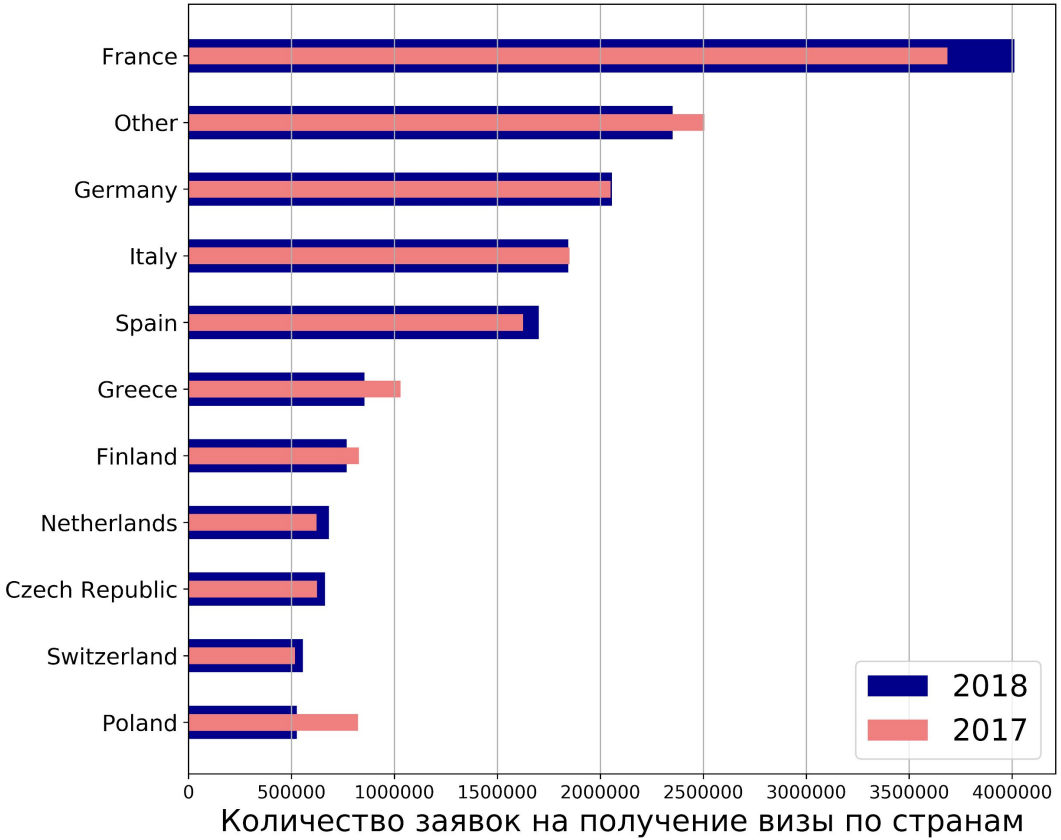
Шамшиев Мамат

Данные

Датасет с Kaggle, содержащий информацию о заявках на получение Шенгенской визы за 2017-2018 года.

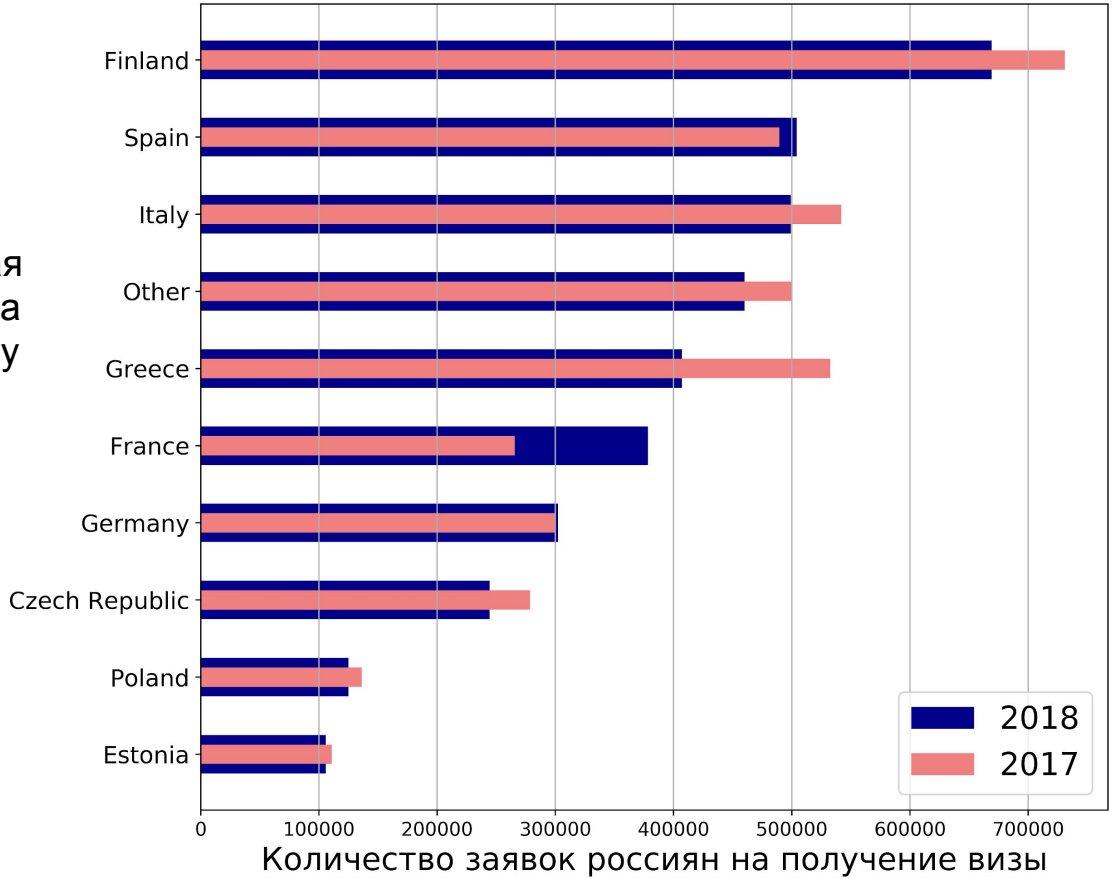
	schengen_state	consulate_country	consulate_city	num_applications	uniform_visas_issued	MEVs_issued	MEVs_share	LTVs_issued	uniform_rejected	r
0	Austria	ALBANIA	TIRANA	62	32	10	31.3%	21	9	
1	Austria	ALGERIA	ALGIERS	2481	1658	1461	88.1%	1	822	
2	Austria	ARGENTINA	BUENOS AIRES	16	16	16	100.0%	NaN	NaN	
3	Austria	AUSTRALIA	CANBERRA	2776	2653	989	37.3%	1	122	
4	Austria	AZERBAIJAN	BAKU	1976	1895	1769	93.4%	2	79	
5	Austria	BOSNIA AND HERZEGOVINA	SARAJEVO	616	615	615	100.0%	NaN	1	
6	Austria	BRAZIL	BRASILIA	37	37	36	97.3%	NaN	NaN	
7	Austria	BULGARIA	SOFIA	217	199	162	81.4%	13	5	
8	Austria	CANADA	OTTAWA	686	659	107	16.2%	1	26	
9	Austria	CHILE	SANTIAGO DE CHILE	22	22	17	77.3%	NaN	NaN	

В какие страны поступает больше всего заявок?



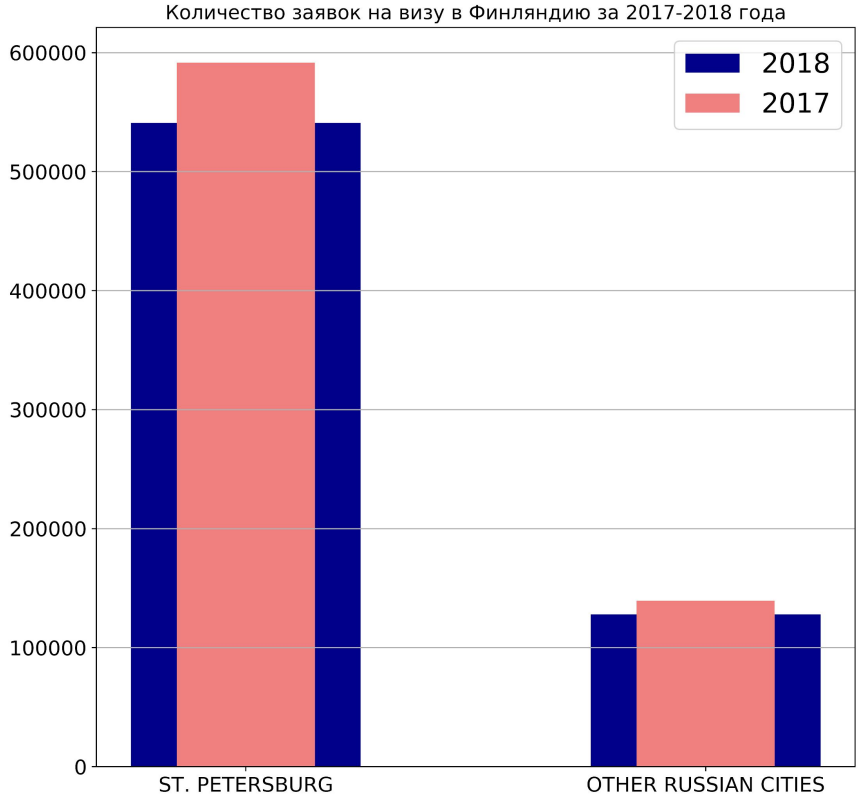
А куда чаще едут россияне?

Финляндия - самая популярная страна Шенгенской зоны у россиян...



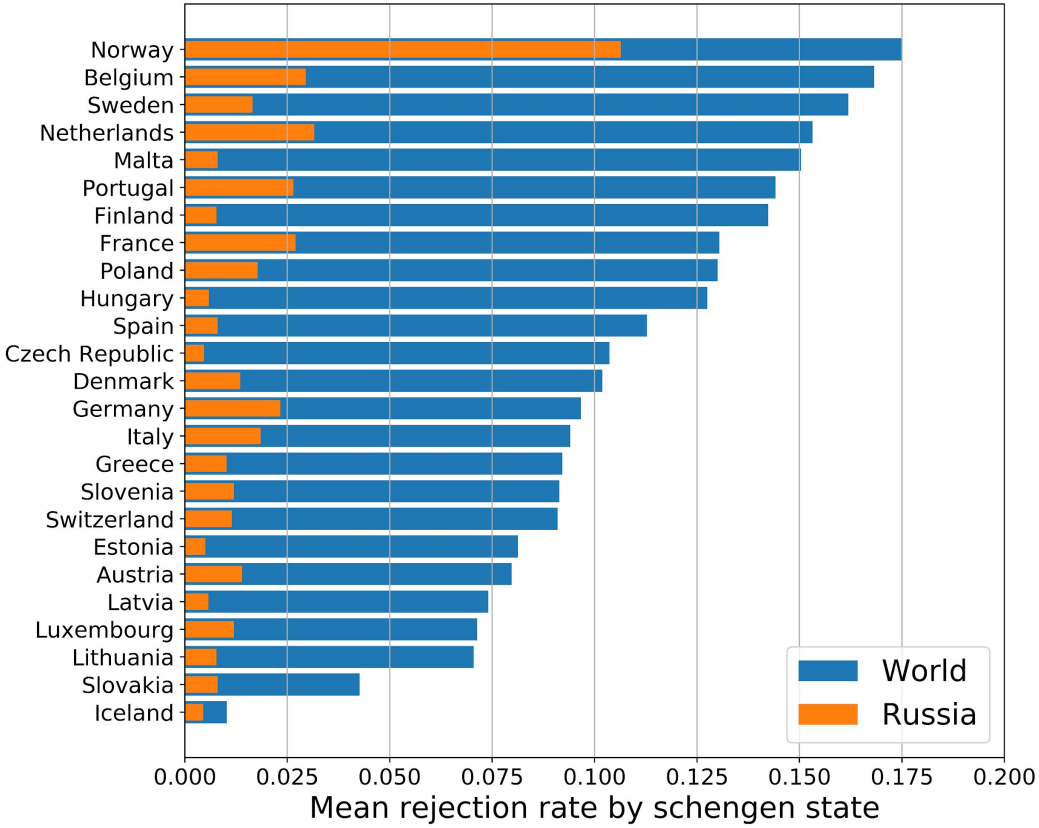
А куда чаще едут россияне?

...подавляющее большинство из которых - жители Санкт-Петербурга.



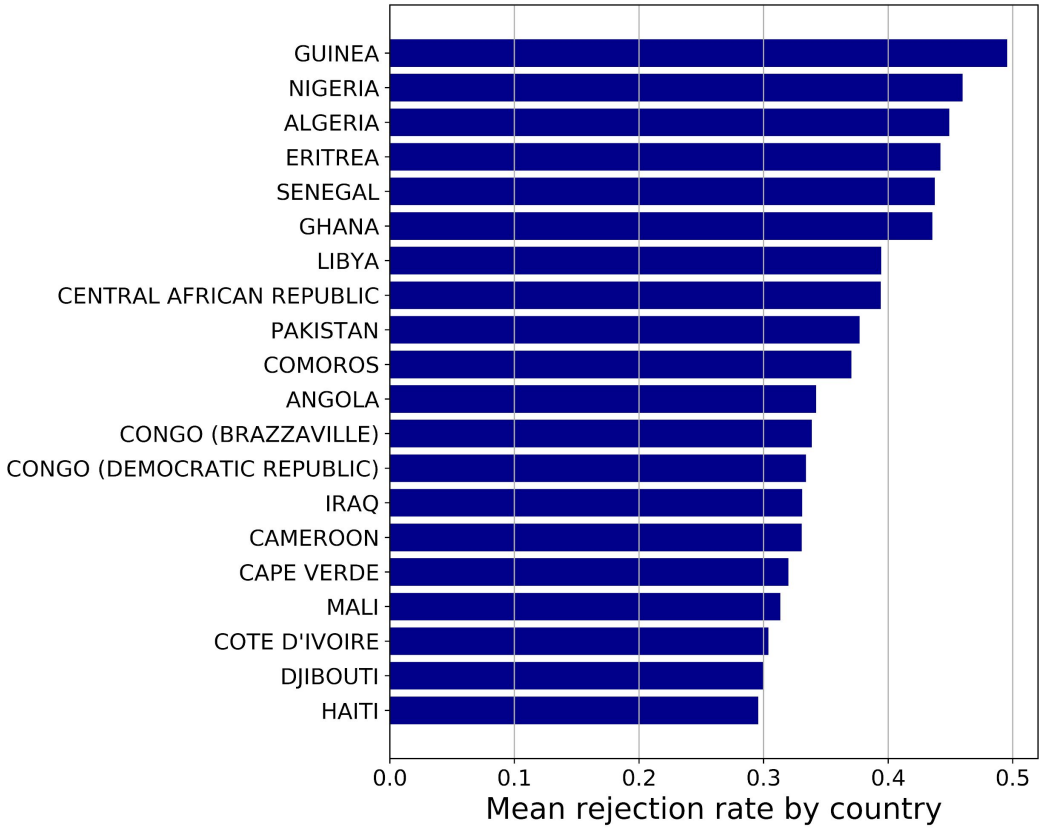
Какие страны чаще всего отказывают?

Россиянам отказывают значительно реже, чем в среднем по миру.



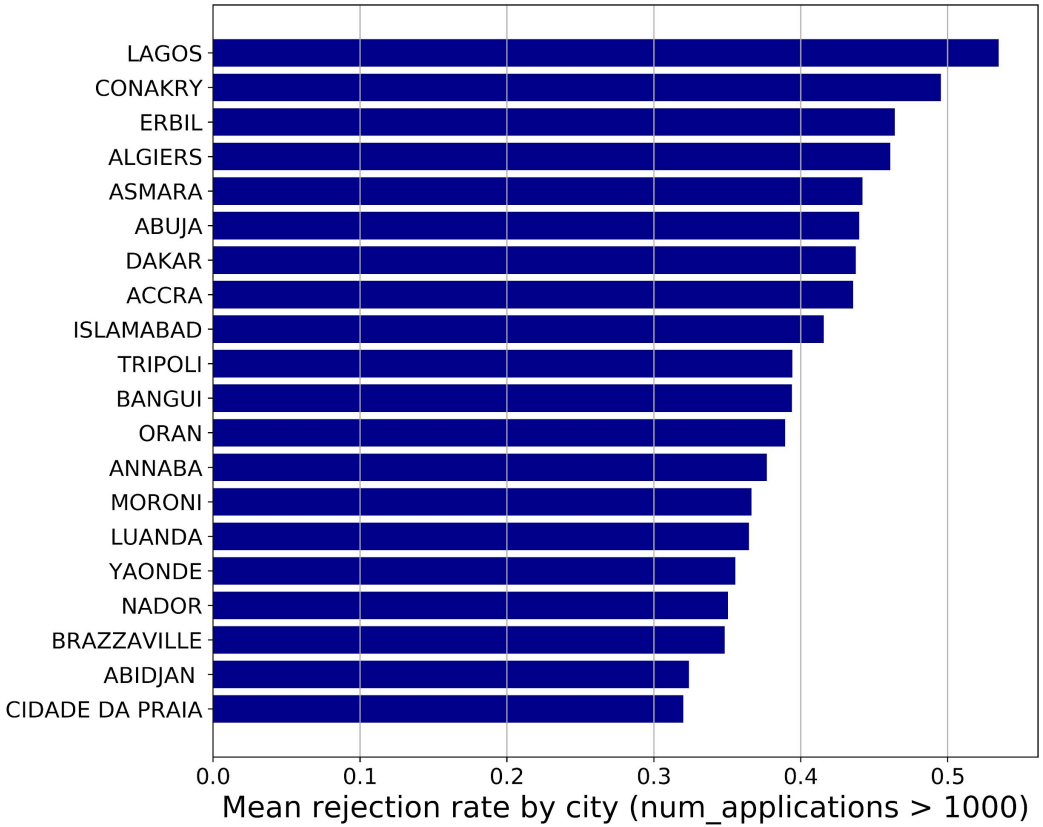
Консульства, находящиеся в каких странах, чаще всего отказывают?

Все страны находятся в Африке.



Консульства каких городов чаще всего отказывают?

Почти все города находятся в Африке.



О найденных проблемах

Данные внутри одного и того же столбца могут быть записаны в разном формате:

num_applications	uniform_visas_issued	MEVs_issued	ME
62	32	10	
2481	1658	1461	
16	16	16	
2776	2653	989	
1976	1895	1769	

num_applications	uniform_visas_issued	MEVs_issued	M
1,405	1,394	1,380	
3,187	3,116	3,109	
2,675	2,662	2,624	
1,049	993	571	
3,643	3,355	408	

О найденных проблемах

Большое количество пропусков:

	Number of NaNs
schengen_state	0
consulate_country	0
consulate_city	0
num_applications	0
year	0
uniform_visas_issued	78
MEVs_issued	270
MEVs_share	272
uniform_rejected	725
rejection_rate	725
LTVs_issued	1676

Некоторые обозначения:

- num_applications - количество заявок;
- uniform_visas_issued - количество одобренных стандартных виз;
- MEVs_issued - количество одобренных стандартных виз на многократный въезд;
- uniform_rejected - количество отклоненных заявок;
- LTVs_issued - количество одобренных виз с ограниченной территорией действия.

В данных довольно много пропусков...

Попробуем разобраться:

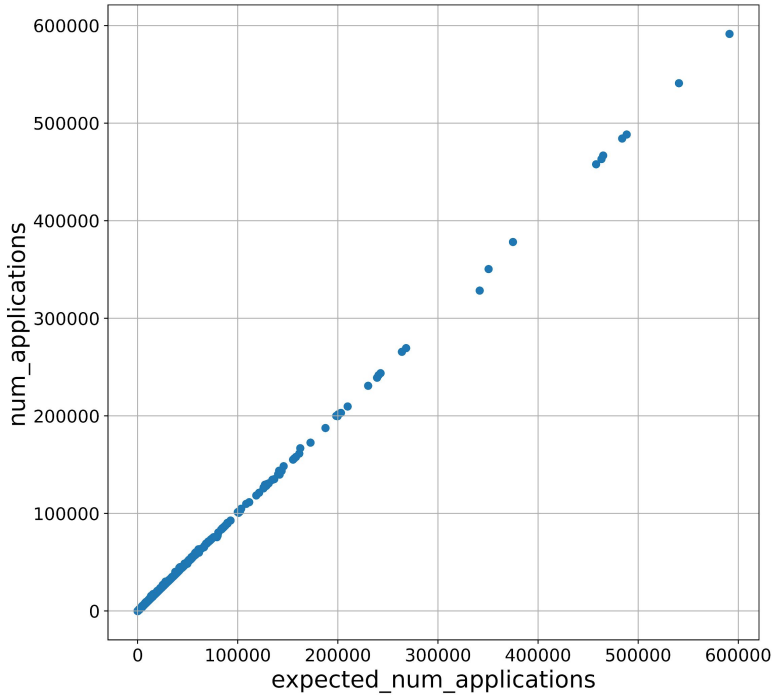
consulate_city	num_applications	uniform_visas_issued	MEVs_issued	MEVs_share	LTVs_issued	uniform_rejected	r
TIRANA	62	32	10	31.3%	21	9	
ALGIERS	2481	1658	1461	88.1%	1	822	
BUENOS AIRES	16	16	16	100.0%	NaN	NaN	
CANBERRA	2776	2653	989	37.3%	1	122	
BAKU	1976	1895	1769	93.4%	2	79	
SARAJEVO	616	615	615	100.0%	NaN	1	
BRASILIA	37	37	36	97.3%	NaN	NaN	
SOFIA	217	199	162	81.4%	13	5	
OTTAWA	686	659	107	16.2%	1	26	
SANTIAGO DE CHILE	22	22	17	77.3%	NaN	NaN	

По логике вещей, на каждую заявку должно быть вынесено решение. То есть первый признак должен быть равен сумме трех остальных:

- num_applications - количество заявок;
- uniform_visas_issued - количество одобренных стандартных виз;
- uniform_rejected - количество отклоненных заявок;
- LTVs_issued - количество одобренных виз с ограниченной территорией действия.

В данных довольно много пропусков...

Проверим гипотезу на основе данных без пропусков:

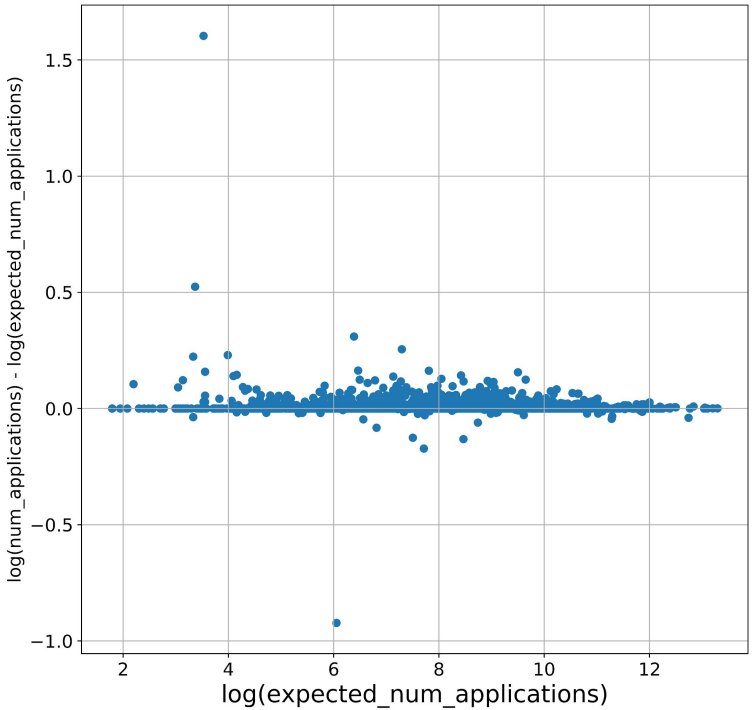


expected_num_applications = uniform_visas_issued + uniform_rejected + LTVs_issued

Как бы сделать график более наглядным...

В данных довольно много пропусков...

Проверим гипотезу на основе данных без пропусков:

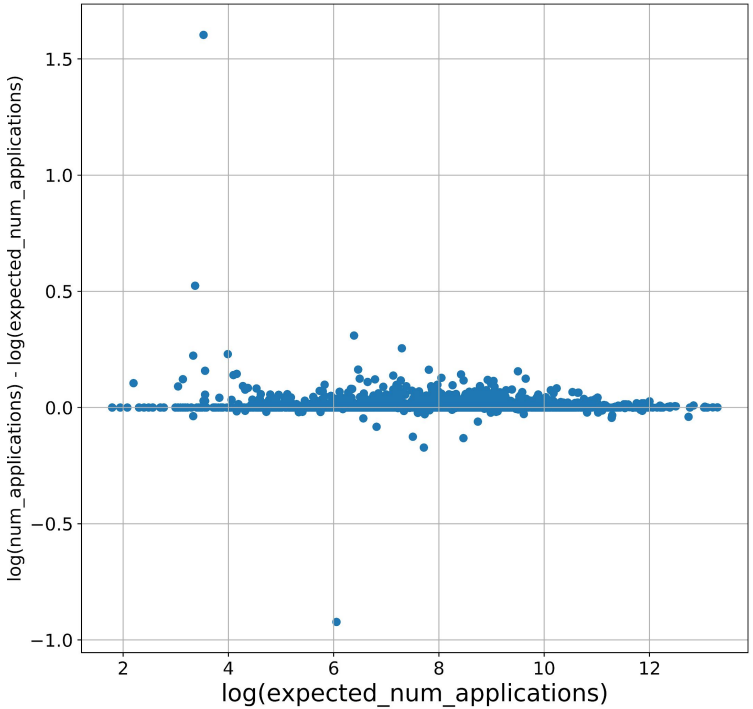


Прологарифмируем и повернем!

Для большинства объектов гипотеза выполняется. Однако существуют и явные выбросы, причем в обе стороны.

В данных довольно много пропусков...

Проверим гипотезу на основе данных без пропусков:



Возможные объяснения:

Количество заявок может быть больше количества вынесенных решений в случае, если, например, некоторые заявки находились в процессе рассмотрения в момент сбора данных.

Обратная ситуация может наблюдаться, если, наоборот, в начале 2017 года были приняты решения по заявкам, поданным ещё в 2016.

В данных довольно много пропусков...

Проверив гипотезу, попробуем уменьшить количество пропусков:

consulate_city	num_applications	uniform_visas_issued	MEVs_issued	MEVs_share	LTVs_issued	uniform_rejected
TIRANA	62	32	10	31.3%	21	9
ALGIERS	2481	1658	1461	88.1%	1	822
BUENOS AIRES	16	16	16	100.0%	NaN	NaN
CANBERRA	2776	2653	989	37.3%	1	122
BAKU	1976	1895	1769	93.4%	2	79
SARAJEVO	616	615	615	100.0%	NaN	1
BRASILIA	37	37	36	97.3%	NaN	NaN
SOFIA	217	199	162	81.4%	13	5
OTTAWA	686	659	107	16.2%	1	26
SANTIAGO DE CHILE	22	22	17	77.3%	NaN	NaN

Сразу замечаем, что если все заявки были одобрены, то в столбцах LTVs_issued и uniform_rejected стоят NaN.

Понимаем, что на самом деле это нули

В данных довольно много пропусков...

Проверив гипотезу, попробуем уменьшить количество пропусков:

consulate_city	num_applications	uniform_visas_issued	MEVs_issued	MEVs_share	LTVs_issued	uniform_rejected
TIRANA	62	32	10	31.3%	21	9
ALGIERS	2481	1658	1461	88.1%	1	822
BUENOS AIRES	16	16	16	100.0%	NaN	NaN
CANBERRA	2776	2653	989	37.3%	1	122
BAKU	1976	1895	1769	93.4%	2	79
SARAJEVO	616	615	615	100.0%	NaN	1
BRASILIA	37	37	36	97.3%	NaN	NaN
SOFIA	217	199	162	81.4%	13	5
OTTAWA	686	659	107	16.2%	1	26
SANTIAGO DE CHILE	22	22	17	77.3%	NaN	NaN

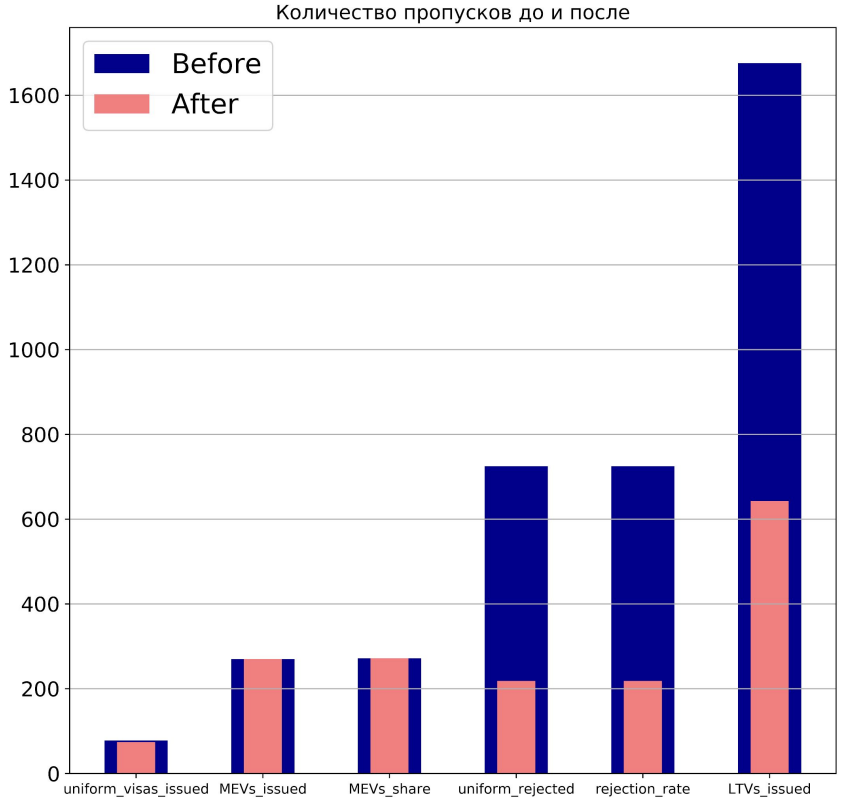
Также замечаем ситуации, когда количество заявок совпадает с суммой двух (из трёх) колонок.

Значит, опять же, в оставшейся колонке вместо пропуска ставим нуль.

В данных довольно много пропусков...

Вывод:

Бывают задачи, в которых можно значительно сократить количество пропусков, просто посмотрев на данные и воспользовавшись здравым смыслом.



Ссылки

Датасет: https://www.kaggle.com/ma7555/schengen-visa-stats

Код: Notebook