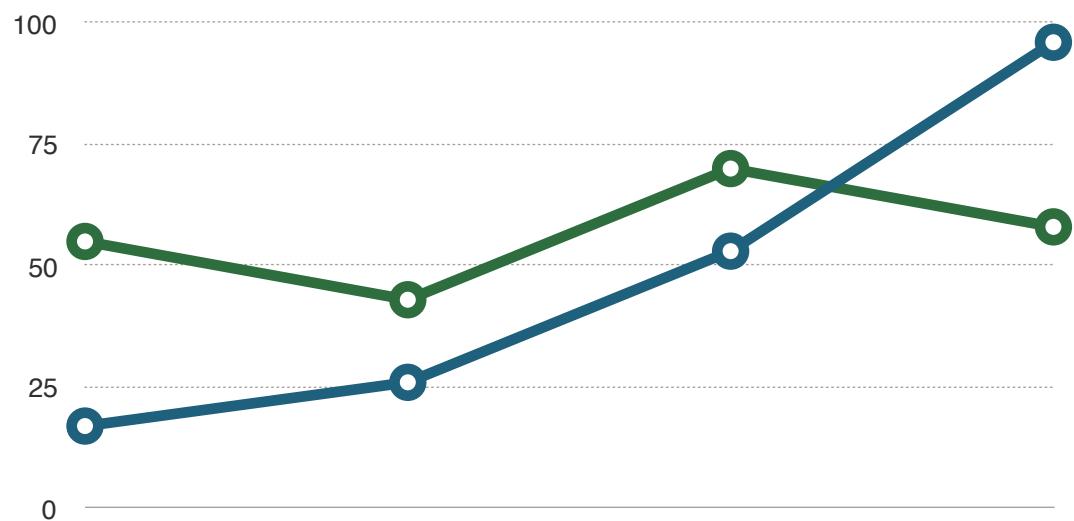


# Расчётное задание по курсу “Статистические методы анализа данных”



Выполнил:

Степан Морозов, ИВТ-11МО

Преподаватель:

Спиридонова Е. М., доцент д.э.н.

24 декабря 2019 года

---

## Вступление

В рамках курса “Статистические методы анализа данных” была поставлена задача провести статистический анализ открытых данных. В рамках этой задачи были выделены следующие подзадачи:

- Найти открытые данные (далее - “датасет”);
- Провести анализ данных;
- Подвести и визуализировать итоги;
- Подготовить статью, описывающую ход работы.

Мною были подготовлены три датасета:

- Список зарегистрированных имен для собак за 2012-2017 г. в Торонто (более 50 тыс. записей);
- Список авиакатастроф за 1908-2009 г. в мире (более 5 тыс. записей);
- Список моих музыкальных предпочтений с сервиса [last.fm](#) за последний год (более 15 тыс. записей).

Соответственно, были проведены три различных исследования по подготовленным датасетам.

Датасеты, рабочая тетрадь и статья доступны по адресу [github.com/morsstepan/data-analysis](https://github.com/morsstepan/data-analysis) в удаленном репозитории.

---

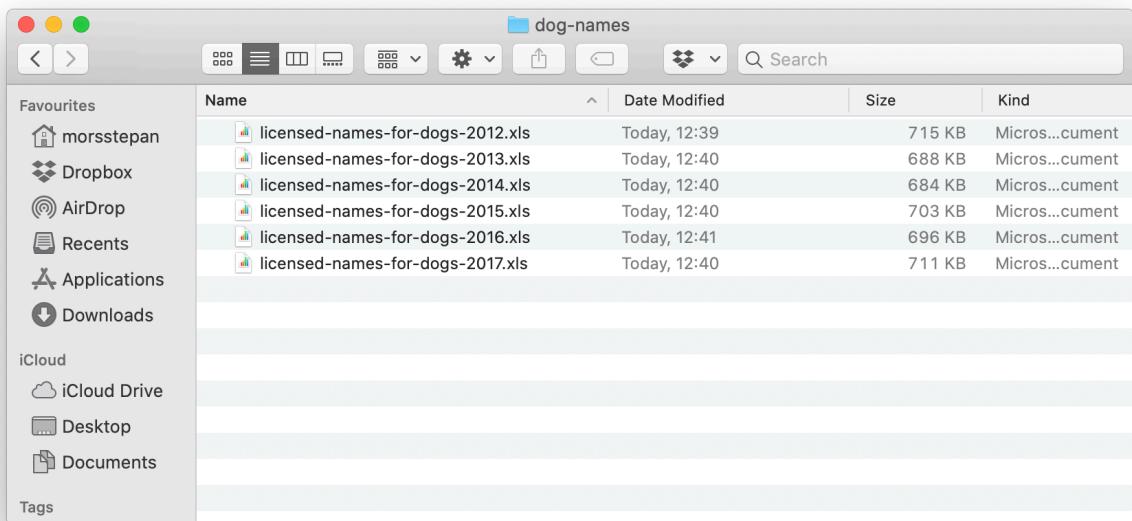
## Используемые инструменты

Во время исследования были следующие инструменты:

- Google Dataset Search — для поиска бесплатных датасетов;
- Программное обеспечение Tableau Desktop 2019.4 — для выполнения вычислений и построения диаграмм;

# Анализ наиболее популярных собачьих кличек в Торонто

Были загружены данные за 2012-2017 г. в формате .xls с официального сайта города



Торонто с открытыми данными <https://open.toronto.ca/dataset/licensed-dog-and-cat-names/>.

Данные выглядят следующим образом:

Names for DOGs Licenced From January 01, 2017 To December 31, 2017				
CHARLIE	728			
BELLA	558			
MAX	519			
MOLLY	444			
BUDDY	424			
BAILEY	396			
COCO	395			
MAGGIE	394			

В первой колонке имя, во второй колонке — количество регистраций собак с этим именем. Список довольно обширный, суммарное количество строк — более 50 тыс.

Чтобы приступить к анализу данных, их необходимо загрузить в программное обеспечение Tableau Desktop. Вот так это выглядит стартовый экран этой программы.

Tableau - Regional [Read-Only] - Tableau license expires in 14 days

## Connect

- Search for Data
- Tableau Server
- To a File
  - Microsoft Excel
  - Text file
  - JSON file
  - PDF file
  - Spatial file
  - Statistical file
  - More...
- To a Server
  - Microsoft SQL Server
  - MySQL
  - Oracle
  - Amazon Redshift
  - More...

## Open

Most Popular Dog Na...

Regional

## Discover

[Open a Workbook](#)

- Training
- Getting Started
- Connecting to Data
- Visual Analytics
- Understanding Tableau
- More training videos...

[Resources](#)

- Get Tableau Prep
- Blog - Nominations are open for 2020 Zen Masters!
- Forums

[More Samples](#)

Craving some viz inspiration?

Explore stunning examples from Tableau Public with [Viz of the Day](#) →

Нам необходимо создать новую рабочую область и импортировать данные для анализа. Загруженные в программе данные выглядят следующим образом:

Tableau - Most Popular Dog Names - Tableau license expires in 14 days

Sheet1 (licensed-names-for-dogs-2012)

Connection:  Live  Extract

Filters: 0 | Add

Names for Dogs Lice...	null
CHARLIE	674
MAX	660
BUDDY	534
MOLLY	506
BELLA	489
BAILEY	456
MAGGIE	446
LUCY	404
DAISY	399
LUCKY	398
COCO	379
TOBY	373
ROCKY	347
LOLA	323
RILEY	256
JACK	253

Data Source: Case 1: Most Popular Dog Name... Case 2: Most Popular Dog Name... Most Popular Dog Names in T... Air Plane Crashes

Так как данные уже агрегированные (В датасете напротив клички собаки - количество регистраций. Нет необходимости вычислять это самостоятельно), то совокупное значение, в моем случае количество регистраций кличек, уже будет собственной мерой. Тогда мы действуем следующий образом:

1. Переносим измерение в текст на панели меток.
2. Переносим меру в размер на панели меток (это должно автоматически использовать расчет суммы).
3. Меняем тип диаграммы на «Text» (это может занять пару минут).
4. Переносим меру в Color на панели Marks (это тоже займет некоторое время для вычисления). Мера должна быть снова автоматически суммирована.

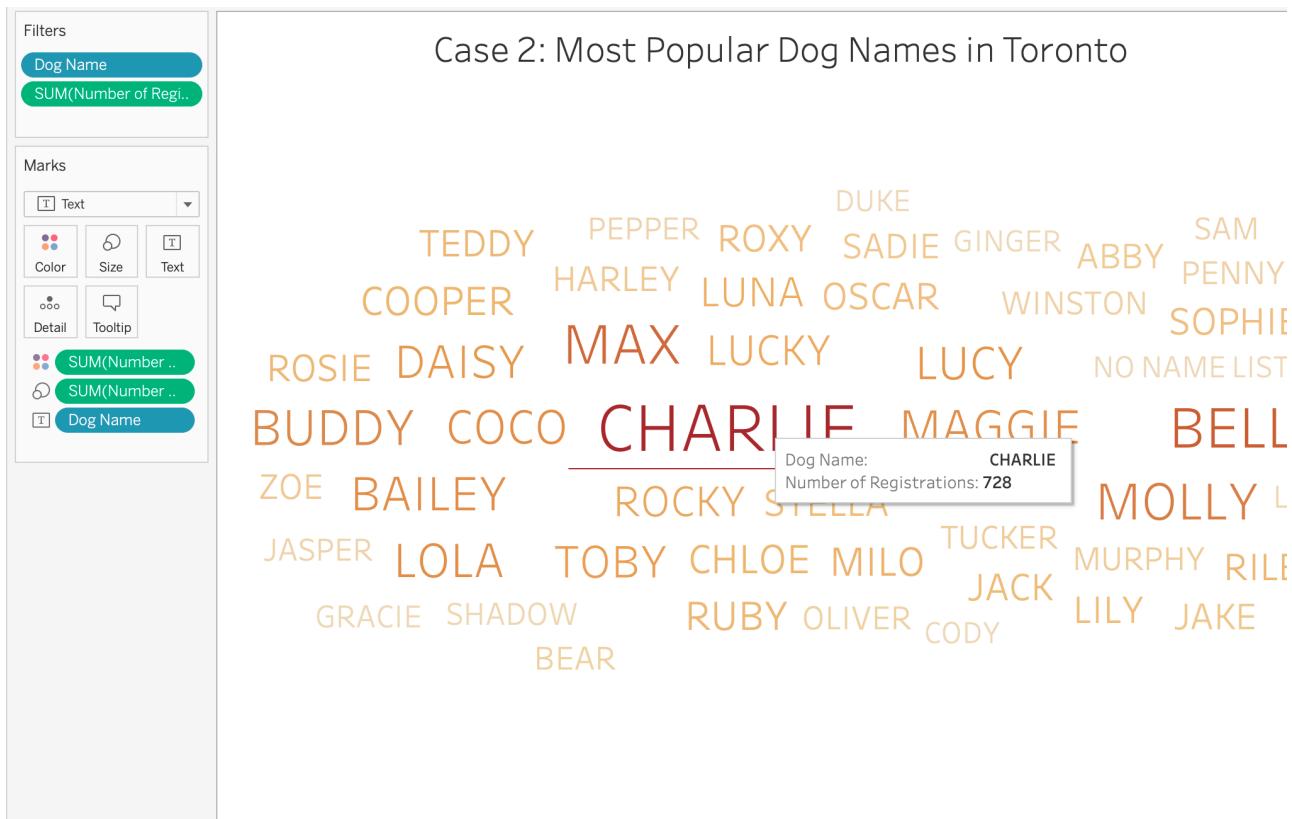
Dog Name	Number of Registrations
CHARLIE	102
LUCY	98
MOLLY	95
MAX	92
OLIVER	88
DAISY	85
ROCKY	82
TOBY	78
COOPER	75
RUBY	72
SOPHIE	70
LUCKY	68
CHLOE	65
MILO	62
PEPPER	60
Winston	58
GRACIE	55
GINGER	52
DUKE	50
DEXTER	48
PENNY	45
STELLA	42
JASPER	40
LOLA	38
CODY	35
JACK	32
TUCKER	30
ZOE	28
SHADOW	25
ABBY	22
JAKE	20
LILY	18
NO NAME LISTED	15
WINSTON	12
LEO	10
ROSIE	8
GINGER	6
GRACIE	5
PEPPER	4
DUKE	3
SADIE	2
PEPPERMINT	1

На этом все. Остается немного поправить настройки в Tableau, чтобы получить красивую диаграмму.

Most Popular Dog Names in Toronto



На этой диаграмме видно, что жители Торонто предпочитают кличку “Чарли”. Всего было зарегистрировано 728 собак с такой кличкой. Диаграмма динамическая, её можно рассматривать почти с “любой стороны” и ранжировать так, как хочется.



А что, если данные сложнее? Если они не агрегированы заранее? Например, в таком виде:

Name		Registration Date
Charlie		2016-01-01
Charlie		2016-05-23

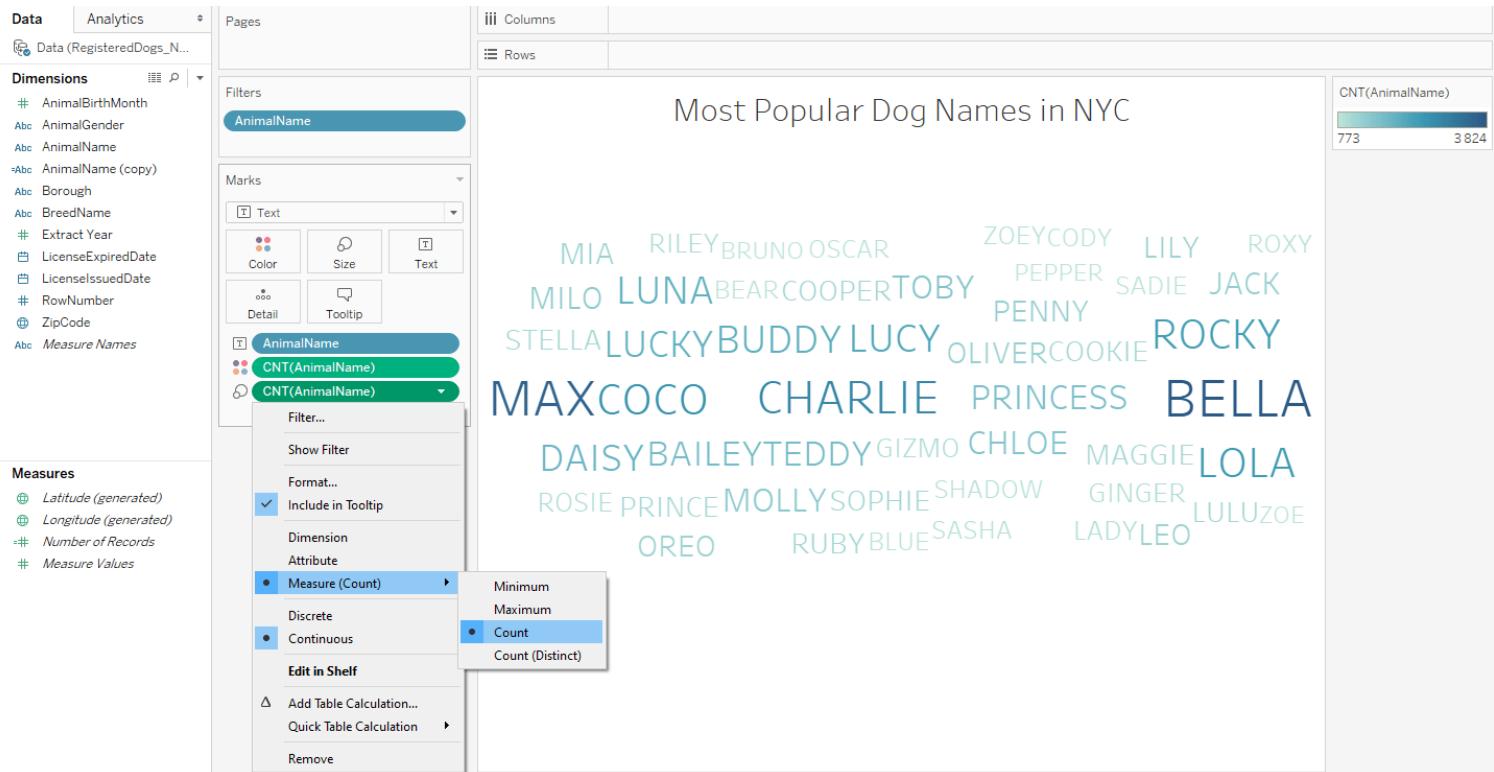
Такой датасет будет миллионы строк, т. к. каждая регистрация собаки вынесена на новую строчку, когда в предыдущем варианте правительство Канады значительно обличило нам задачу. В датасете города Нью-Йорка ситуация иная.

В таком случае это будет означать, что для каждой отдельной собаки у меня нас один ряд, даже если они имеют одно и то же имя, например два Чарли составляют два ряда.

Чтобы ранжировать эти данные нужно выполнить следующее:

1. Переносим измерение в “Text” на панели меток.

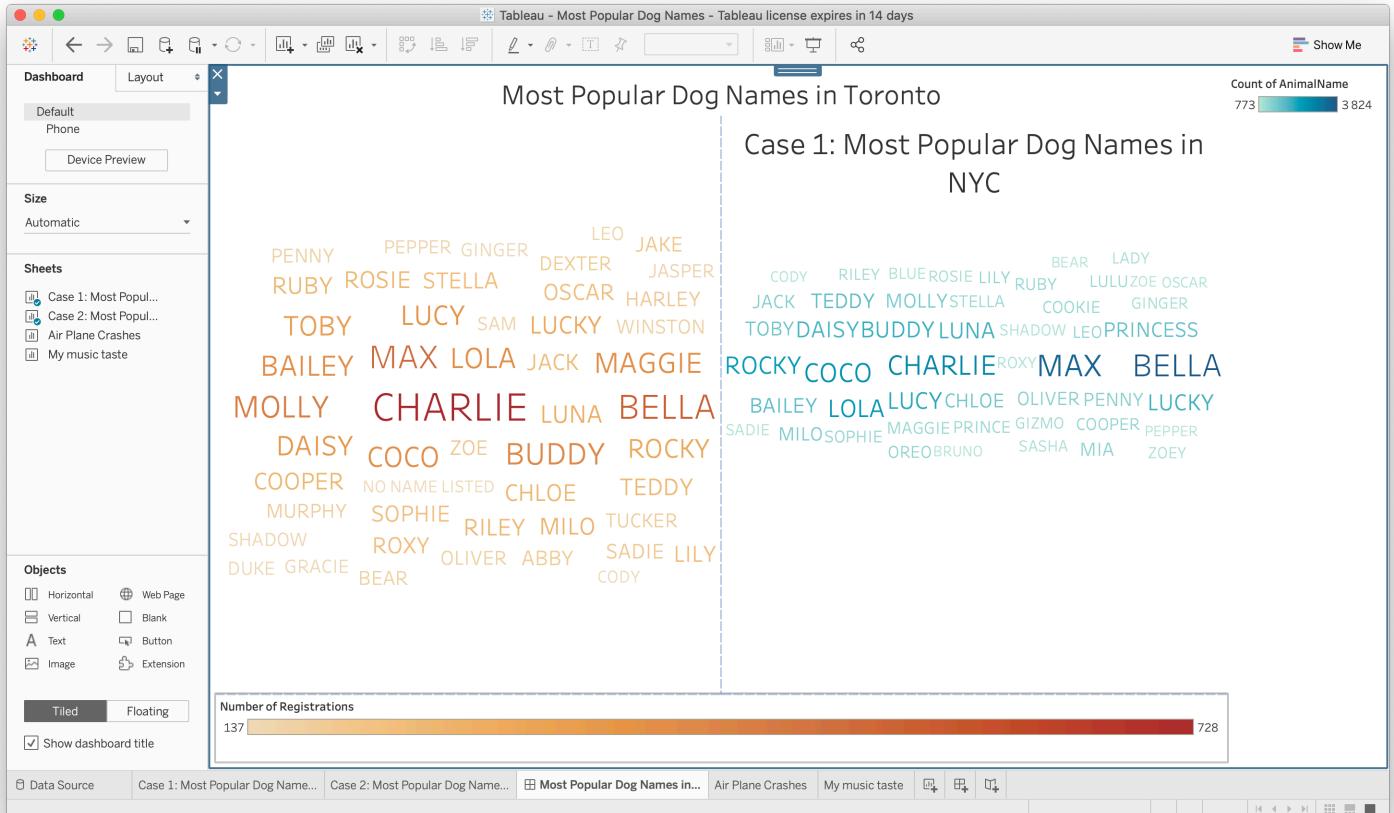
- Переносим меру в "Size" на панели меток. Нажать на перетаскиваемое поле и выберите "Measure">"Count".
- Меняем тип диаграммы на «Text» (это может занять некоторое время для расчета, в моем случае это заняло приблизительно 10 минут).
- Переносим меру в "Color" на панели меток. Выбираем "Measure">"Count".



И получаем диаграмму, на которой видно, что жители Нью-Йорка предпочитают называть своих четвероногих друзей Максами и Беллами:



Для сравнения расположим эти диаграммы рядом:



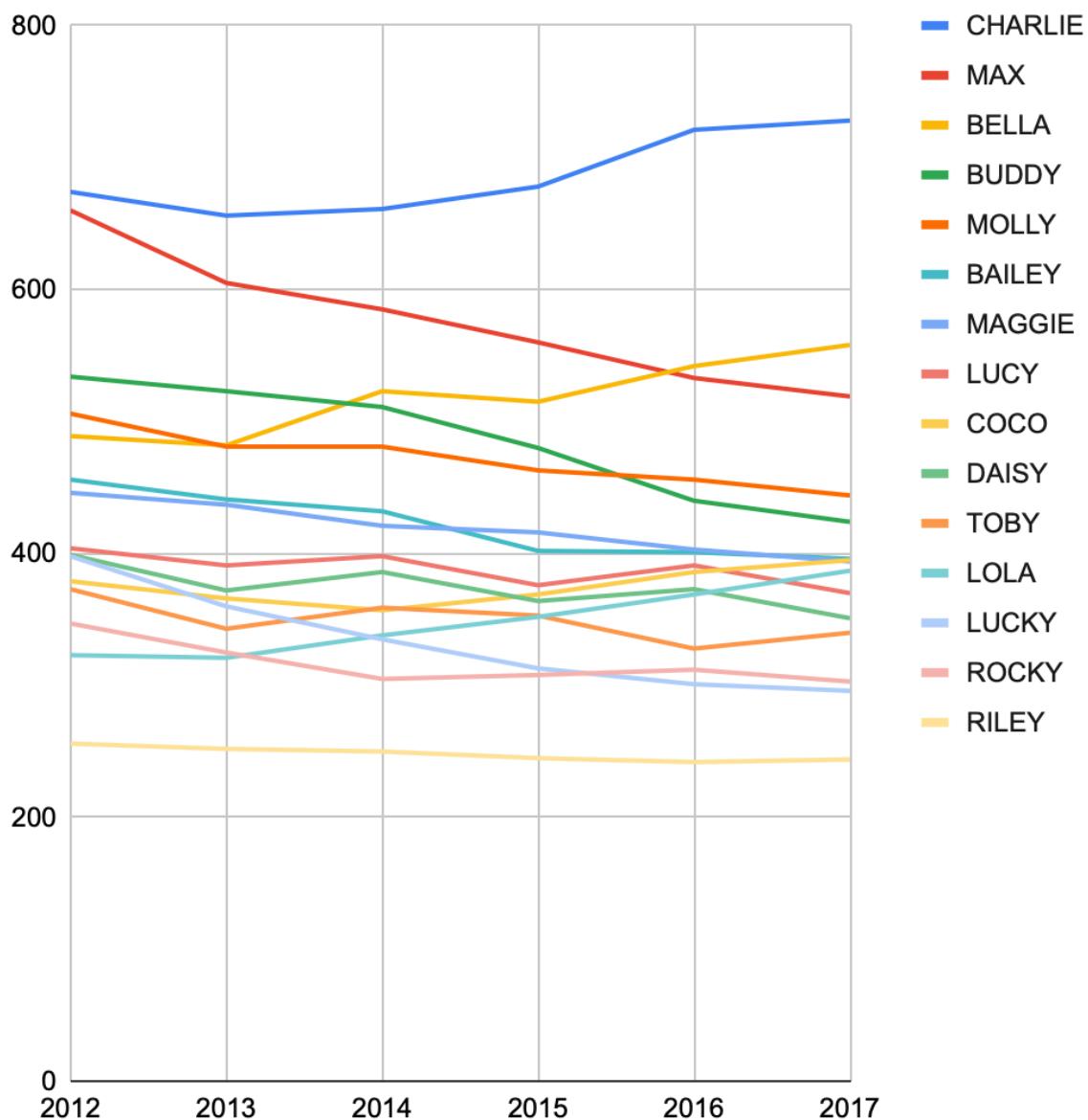
Для полноты картины вернёмся к данным по славному городу Торонто и объединим датасеты всех лет вместе.

The screenshot shows a Google Sheets document titled "dognames". The spreadsheet contains data for 27 dog names across 9 years (2012-2017). The columns are labeled A through I, representing the years. The data shows the following counts:

	A	B	C	D	E	F	G	H	I
1		2012	2013	2014	2015	2016	2017	Total Pop	Change
2	CHARLIE	674	656	661	678	721	728	4118	2916
3	MAX	660	605	585	560	533	519	3462	19881
4	BELLA	489	482	523	515	542	558	3109	4761
5	BUDDY	534	523	511	480	440	424	2912	12100
6	MOLLY	506	481	481	463	456	444	2831	3844
7	BAILEY	456	441	432	402	401	396	2528	3600
8	MAGGIE	446	437	421	416	403	394	2517	2704
9	LUCY	404	391	398	376	391	370	2330	1156
10	COCO	379	366	357	369	386	395	2252	256
11	DAISY	399	372	386	364	373	351	2245	2304
12	TOBY	373	343	359	353	328	340	2096	1089
13	LOLA	323	321	338	352	369	387	2090	4096
14	LUCKY	398	360	335	313	301	296	2003	10404
15	ROCKY	347	325	305	308	312	303	1900	1936
16	RILEY	256	252	250	245	242	244	1489	144
17	JACK	253	240	238	242	237	236	1446	289
18	CHLOE	243	224	246	235	230	251	1429	64
19	ROXY	246	235	229	235	229	225	1399	441
20	TEDDY	224	217	204	215	217	244	1321	400
21	RUBY	192	203	208	216	238	250	1307	3364
22	COOPER	183	206	209	218	236	252	1304	4761
23	LILY	213	199	195	213	213	233	1266	400
24	JAKE	246	215	198	197	188	204	1248	1764
25	STELLA	202	188	201	210	214	232	1247	900
26	SOPHIE	185	208	194	203	219	227	1236	1764
27	OSCAR	213	180	197	198	222	225	1235	144

Сделаем диаграмму и назовем её “Топ собачьих имен по версии жителей Торонто за 2012-2017”.

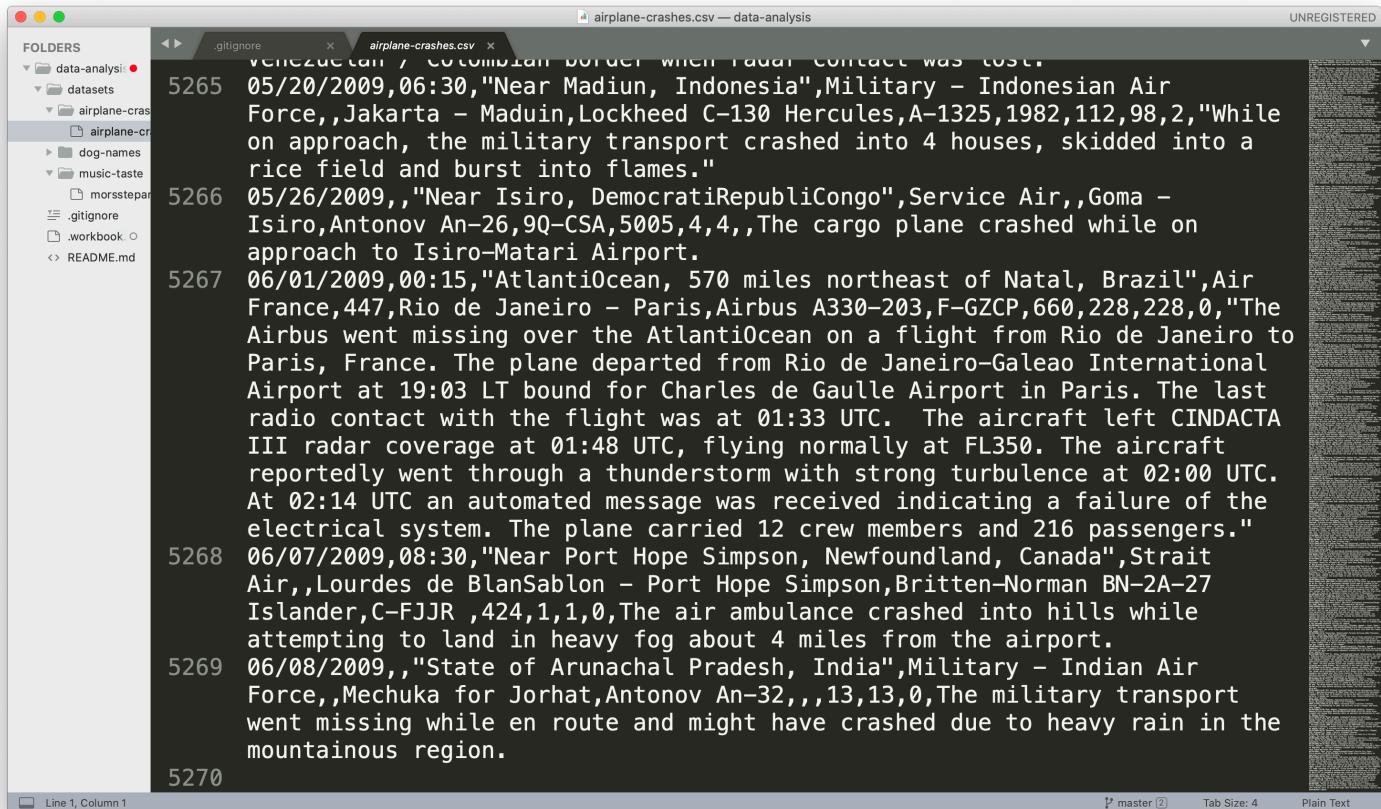
2012, 2013, 2014, 2015, 2016...



Оценим график зависимостей. Кличка “Чарли” несомненный лидер. И, кажется, жители Торонто очень любят британскую актрису Мэгги Смит, раз начали называть собак “Мэгги” вместо “Люси” (популярность имени “Люси” упала, а “Мэгги” — возросла).

# Анализ авиакатастроф с 1908 по 2009

Датасет с авиакатастрофами содержит в себе более 5 тыс. записей. Загружен датасет с ресурса <https://opendata.socrata.com/Government/Airplane-Crashes-and-Fatalities-Since-1908/q2te-8cvq>



```
airplane-crashes.csv — data-analysis
UNREGISTERED

FOLDERS
└── data-analysis
    ├── .gitignore
    └── datasets
        └── airplane-crashes
            └── airplane-crashes.csv
    ├── dog-names
    ├── music-taste
    └── morsstepan
        └── .gitignore
    └── .workbook.o
    └── README.md

5265 05/20/2009,06:30,"Near Madiun, Indonesia",Military – Indonesian Air Force,,Jakarta – Maduin,Lockheed C-130 Hercules,A-1325,1982,112,98,2,"While on approach, the military transport crashed into 4 houses, skidded into a rice field and burst into flames."
5266 05/26/2009,,,"Near Isiro, Democratic Republic Congo",Service Air,,Goma – Isiro, Antonov An-26,9Q-CSA,5005,4,4,,The cargo plane crashed while on approach to Isiro-Matari Airport.
5267 06/01/2009,00:15,"Atlantic Ocean, 570 miles northeast of Natal, Brazil",Air France,447,Rio de Janeiro – Paris,Airbus A330-203,F-GZCP,660,228,228,0,"The Airbus went missing over the Atlantic Ocean on a flight from Rio de Janeiro to Paris, France. The plane departed from Rio de Janeiro-Galeao International Airport at 19:03 LT bound for Charles de Gaulle Airport in Paris. The last radio contact with the flight was at 01:33 UTC. The aircraft left CINDACTA III radar coverage at 01:48 UTC, flying normally at FL350. The aircraft reportedly went through a thunderstorm with strong turbulence at 02:00 UTC. At 02:14 UTC an automated message was received indicating a failure of the electrical system. The plane carried 12 crew members and 216 passengers."
5268 06/07/2009,08:30,"Near Port Hope Simpson, Newfoundland, Canada",Strait Air,,Lourdes de BlanSablone – Port Hope Simpson,Britten-Norman BN-2A-27 Islander,C-FJJR ,424,1,1,0,The air ambulance crashed into hills while attempting to land in heavy fog about 4 miles from the airport.
5269 06/08/2009,,,"State of Arunachal Pradesh, India",Military – Indian Air Force,,Mechuka for Jorhat, Antonov An-32,,,13,13,0,The military transport went missing while en route and might have crashed due to heavy rain in the mountainous region.
5270
```

Данные записаны в следующем формате:

Airplane_Crash... Date	Airplane_Crashes_and_Fata... Time	Airplane_Crashes_and_Fataliti... Location	Airplane_Crashes_and_Fataliti... Operator	Airplane_Crash... Flight #	Airplane_Crashes_a... Route	Airplane_Crashes_and_Fataliti... Type	Airplane_Crashes_and_F... Registration	Airplane_Cras... cn/ln
05.01.1949	null	Caravelas Bay, Brazil	British South Americ...	null	null	Avro 685 York 1	G-AHEX	1301
06.01.1949	30.12.1899 07:20:00	Brandywine, Maryland	Coastal Cargo	null	Raleigh, NC - ...	Douglas C-47A	NC53210	13777
11.01.1949	null	Near Pelotas, Brazil	Viacao Aerea Gaucha...	null	Porto Alegre ...	Lockheed 18 Lodestar	PP-SAC	null
15.01.1949	null	Ras-el-Tin, Egypt	Pan African Air Charter	null	null	Douglas C-54 Skymas...	ZS-AYB	19584
16.01.1949	null	Balihal Pass, India	Dalmia Jain Airways	null	null	Douglas DC-3	VT-CDZ	14145/255...

---

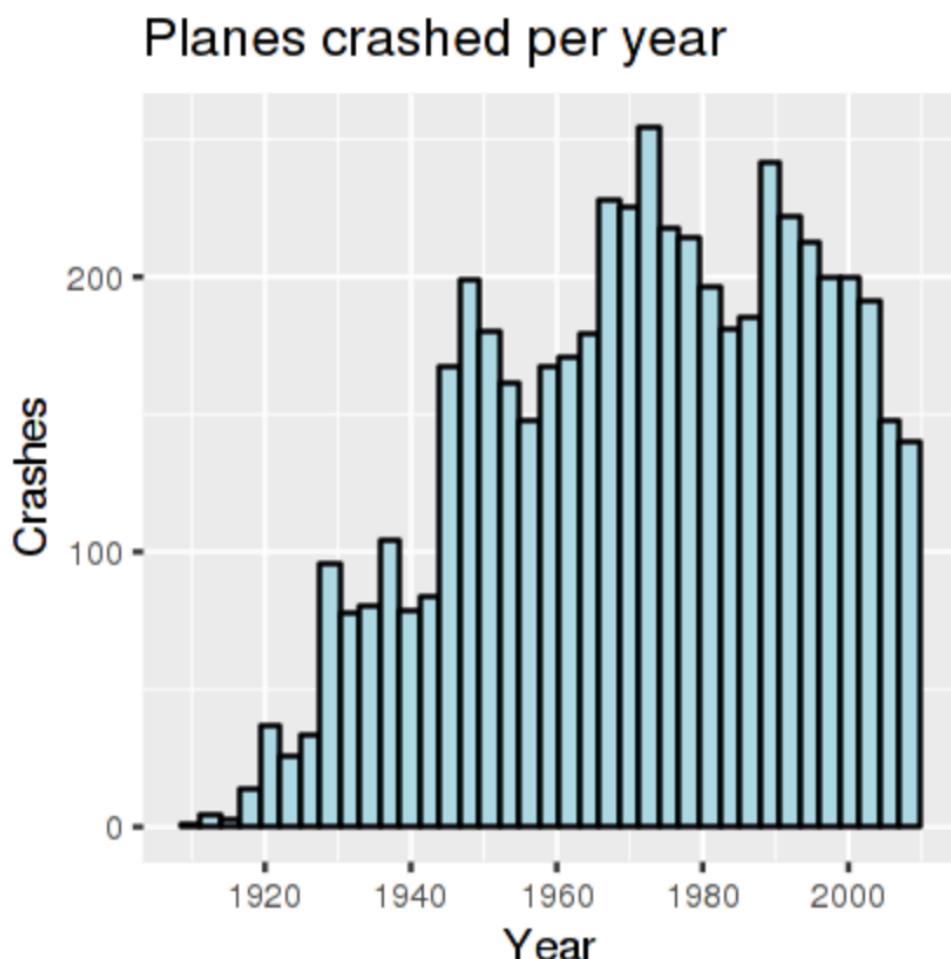
Для анализа данных будем использовать Tableau Desktop.

Так как этот анализ данных более трудоемкий, чем ранжирование собачьих кличек, то можно сформулировать вопросы, на которые необходимо узнать ответы с помощью анализа:

1. Сколько самолетов разбивалось ежегодно? Сколько людей было на борту?  
Сколько людей выжило? Сколько людей погибло?
2. Наибольшее количество аварий по перевозчику и типу самолетов.
3. Поле «Summary» содержит информацию о сбоях. Хочется узнать причины сбоя и классифицировать их по разным кластерам: огонь, бой, погода (для «Пробелов» категории данных может быть НЕИЗВЕСТНО). Количество разбившихся самолетов и количество смертей из каждой категории.

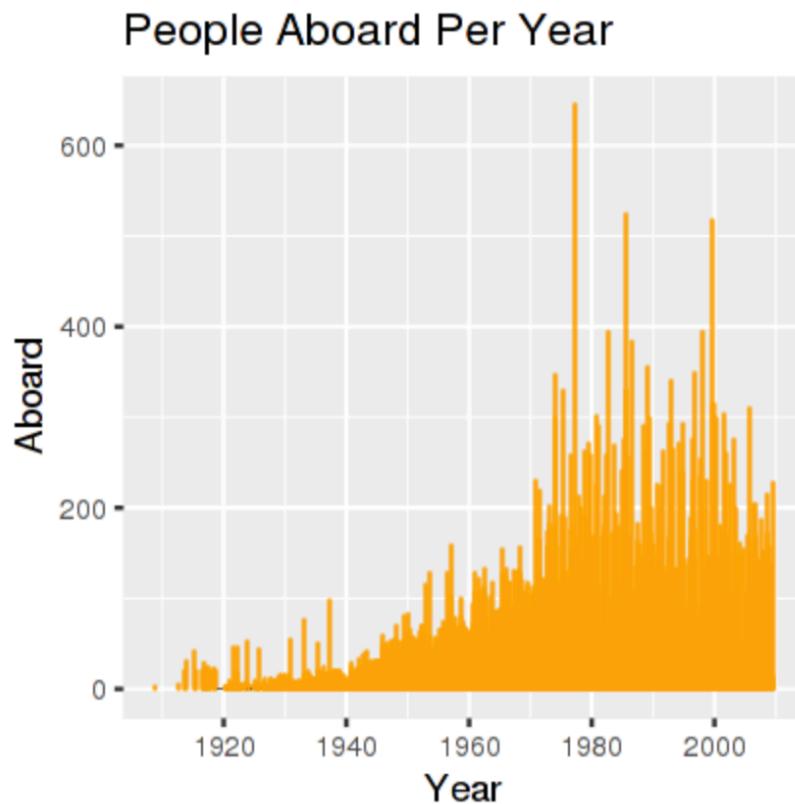
Ответом на **вопрос 1.** может послужить следующий набор диаграмм, составленный при вычислении.

1. Количество авиакатастроф (в год):

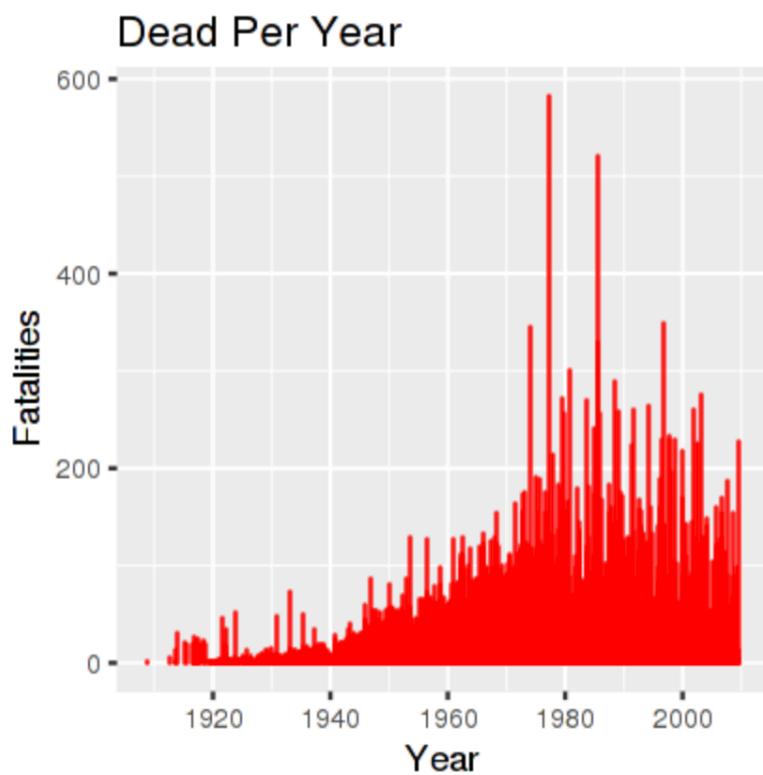


- 2.

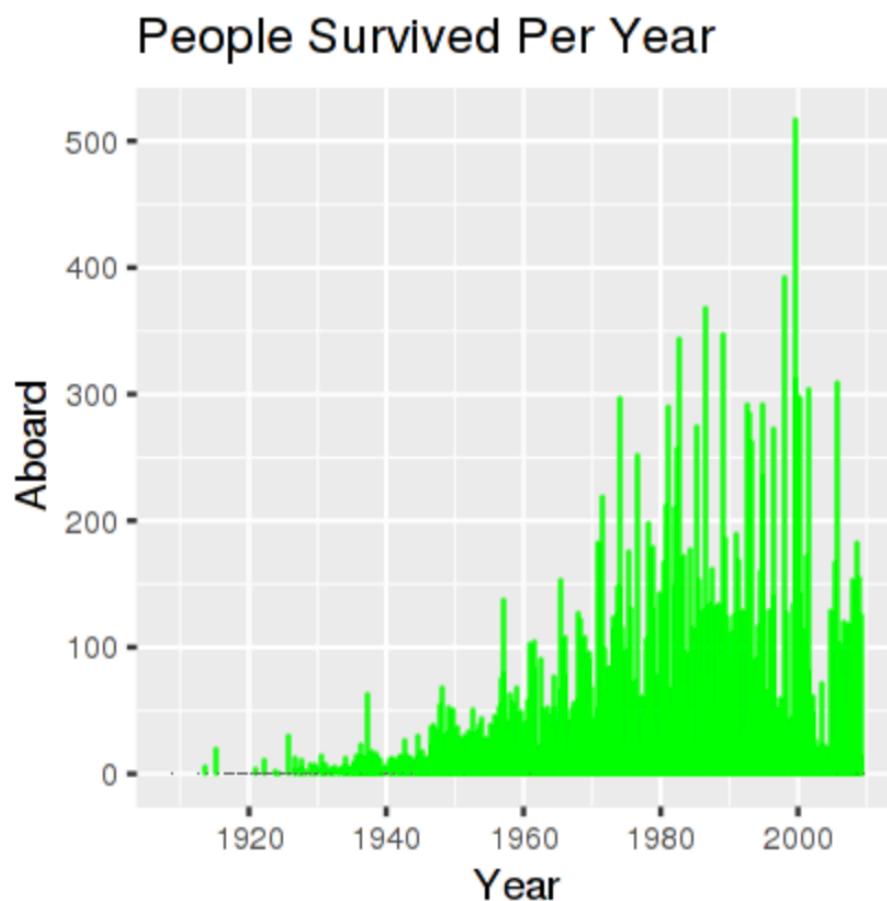
Количество человек на борту (в год):



3. Число погибших (в год):



4. Число выживших (в год):



Ответ на *вопрос 2.* :

- Наибольшее количество аварий по перевозчику - *Аэрофлот* с 179 вылетами.
- Наибольшее количество аварий по типу самолета - *Douglas DC-3* с 334 вылетами

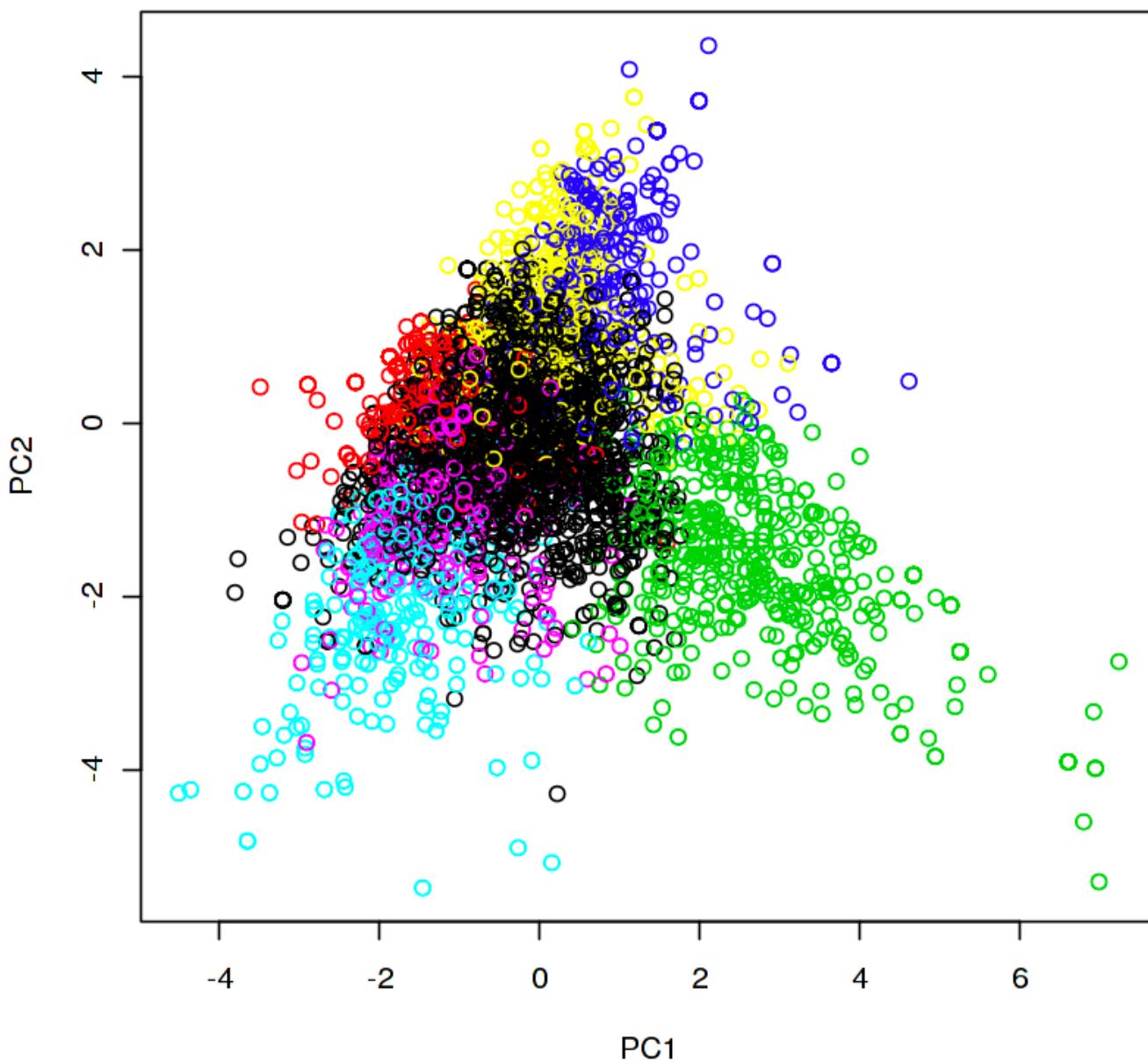
Для получения ответа на вопрос 3. использовался метод кластеризации k-средних на матрице, с помощью корпуса текстов (<https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D1%80%D0%BF%D1%83%D1%81%D1%82%D0%B5%D0%BA%D1%81%D1%82%D0%BE%D0%B2>), созданного с помощью анализа текста (простой текст, без знаков препинания, и. т.д.).

В следующей таблице для каждого кластера суммируется количество аварий и смертей:

Таблица кластеров		
Кластер 1	258 аварий	6368 смертей
Кластер 2	500 аварий	9408 смертей

Кластер 3	211 аварий	3513 смертей
Кластер 4	1014 аварий	14790 смертей
Кластер 5	2749 аварий	58826 смертей
Кластер 6	195 аварий	4439 смертей
Кластер 7	341 аварий	8135 смертей

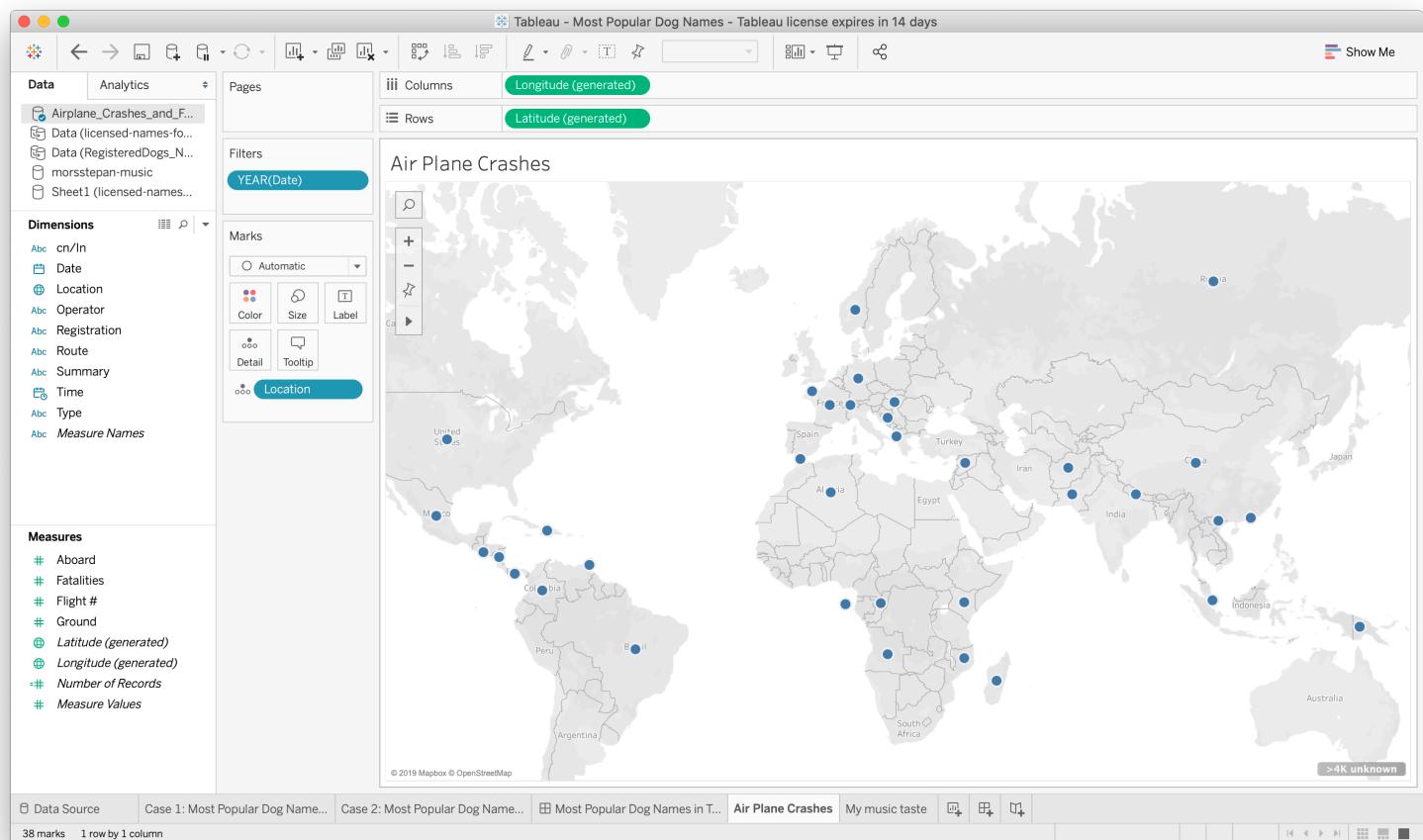
На следующей диаграмме можно увидеть выделенные кластеры:



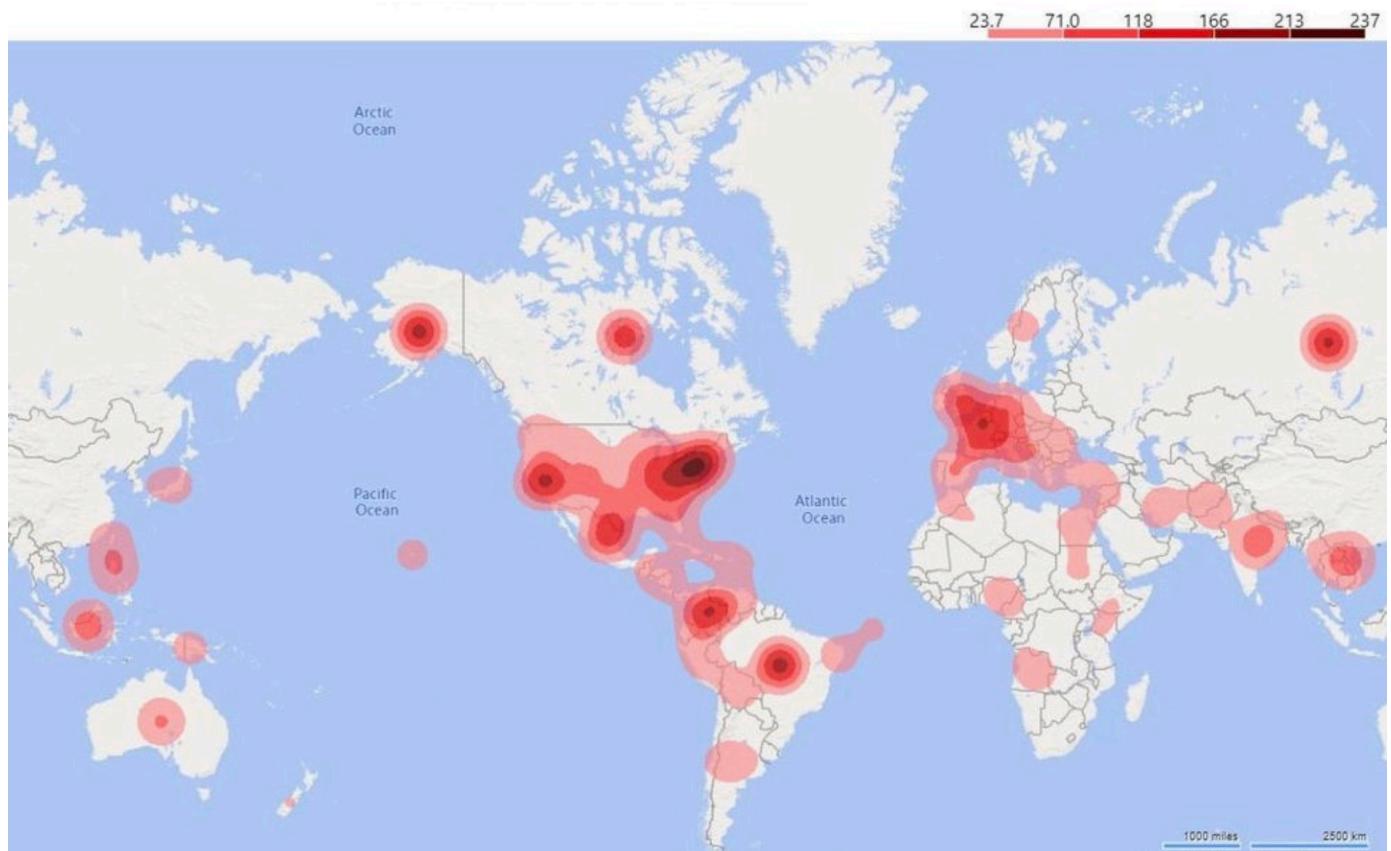
Обобщим для каждого кластера наиболее употребляемые слова и определим причины сбоя:

- **Кластер 1 (258)**: aircraft, crashed, plane, shortly, taking. No many information about this cluster can be deducted using Text Analysis
- **Кластер 2 (500)**: aircraft, airport, altitude, crashed, crew, due, engine, failed, failure, fire, flight, landing, lost, pilot, plane, runway, takeoff, taking. Engine failure on the runway after landing or takeoff
- **Кластер 3 (211)**: aircraft, crashed, fog Crash caused by fog
- **Кластер 4 (1014)**: aircraft, airport, attempting, cargo, crashed, fire, land, landing, miles, pilot, plane, route, runway, struck, takeoff Struck a cargo during landing or takeoff
- **Кластер 5 (2749)**: accident, aircraft, airport, altitude, approach, attempting, cargo, conditions, control, crashed, crew, due, engine, failed, failure, feet, fire, flight, flying, fog, ground, killed, land, landing, lost, low, miles, mountain, pilot. plane, poor, route, runway, short, shortly, struck, takeoff, taking, weather Struck a cargo due to engine failure or bad weather conditions mainly fog
- **Кластер 6 (195)**: aircraft, crashed, engine, failure, fire, flight, left, pilot, plane, runway Engine failure on the runway
- **Кластер 7 (341)**: accident, aircraft, altitude, cargo, control, crashed, crew, due, engine, failure, flight, landing, loss, lost, pilot, plane, takeoff, engine failure

Теперь отметим на карте наиболее частые места, где происходили катастрофы.



Сделаем карту более интересной, добавив цвета и эффект “повторной катастрофы”.



Из этой карты можно сделать вывод, что наибольшее количество катастроф пришлось на Европу и Америку. Причем европейская часть материка здесь скорее всего из-за Второй Мировой войны, т.к. датасет охватывал и эти случаи.

# Анализ моих музыкальных предпочтений

Я выгрузил данные из стримингового сервиса, который я использую для прослушивания. Датасет выглядит следующим образом:

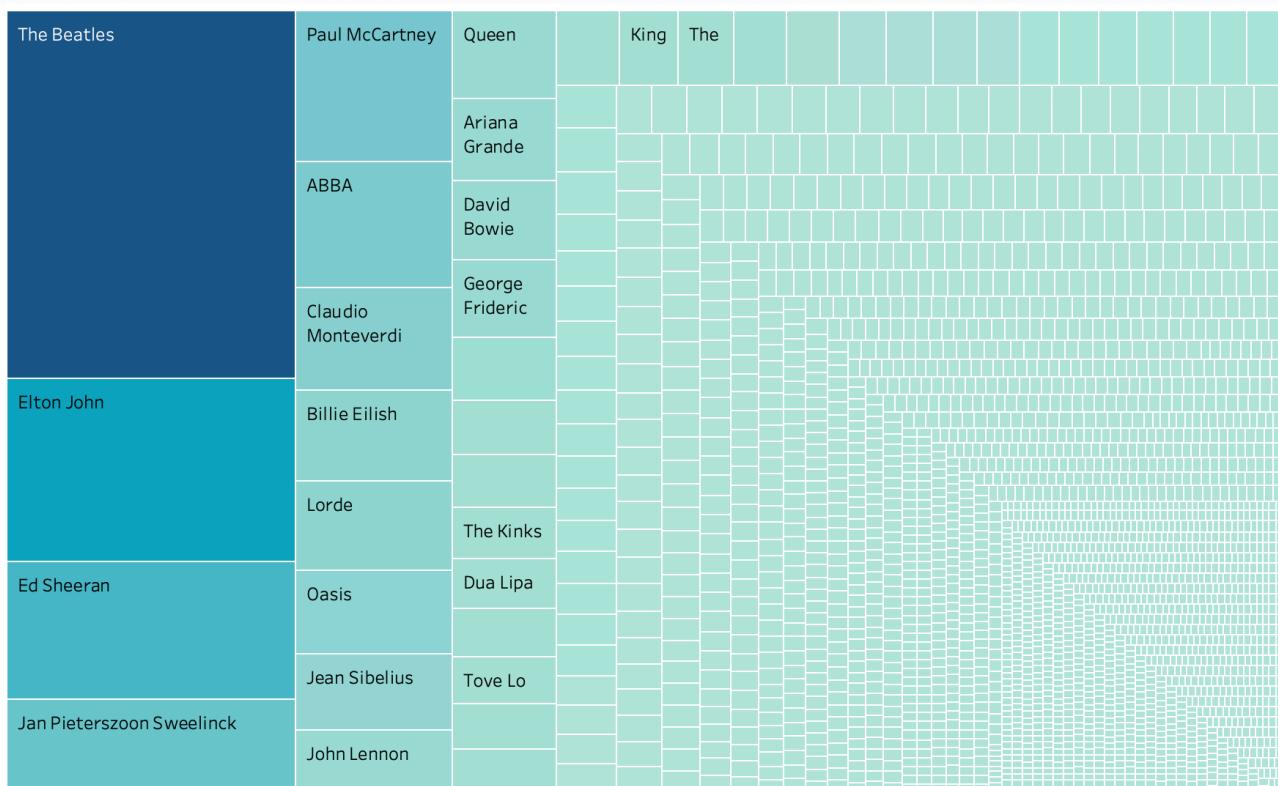
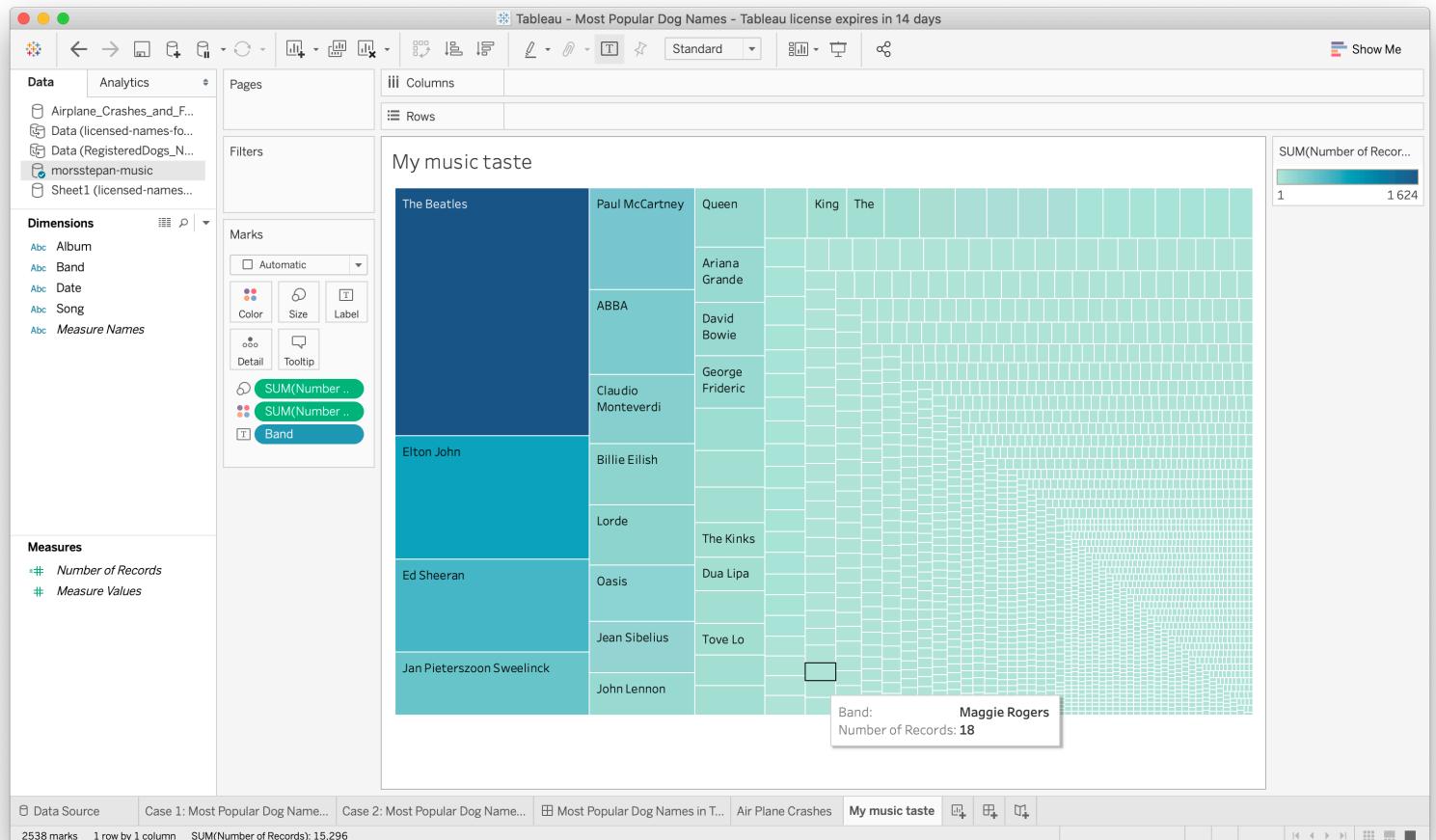
15291	The Rolling Stones	,Let It Bleed,Gimme Shelter,04 May 2019 07:09
15292	Supertramp	,Breakfast In America (Deluxe Edition),Breakfast In America – Remastered,04 May 2019 07:06
15293	Grand Funk Railroad	,Grand Funk (Red Album) [Remastered],Got This Thing On The Move – Remastered,04 May 2019 06:59
15294	Queen	,The Works (2011 Remaster),Radio Ga Ga – Remastered,04 May 2019 06:54
15295	The Animals	,The Singles Plus,The House of the Rising Sun,03 May 2019 18:08
15296	The Smiths	,The Queen Is Dead,There Is a Light That Never Goes Out – 2011 Remaster,03 May 2019 18:05

Импортируем данные в Tableau.

The screenshot shows the Tableau Data Source interface. On the left, under 'Connections', there is one connection named 'morsstepan-music'. Under 'Files', there is one file named 'morsstepan-music.csv'. The main area displays the contents of 'morsstepan-music.csv' in a grid format. The columns are labeled: Band, Album, Song, and Date. The data includes entries like Ariana Grande's 'Santa Tell Me', Lady Antebellum's 'On This Winter's Night', and Michael Bublé's 'Christmas (Deluxe Sp...)'. The 'Date' column shows various recording dates from 2019.

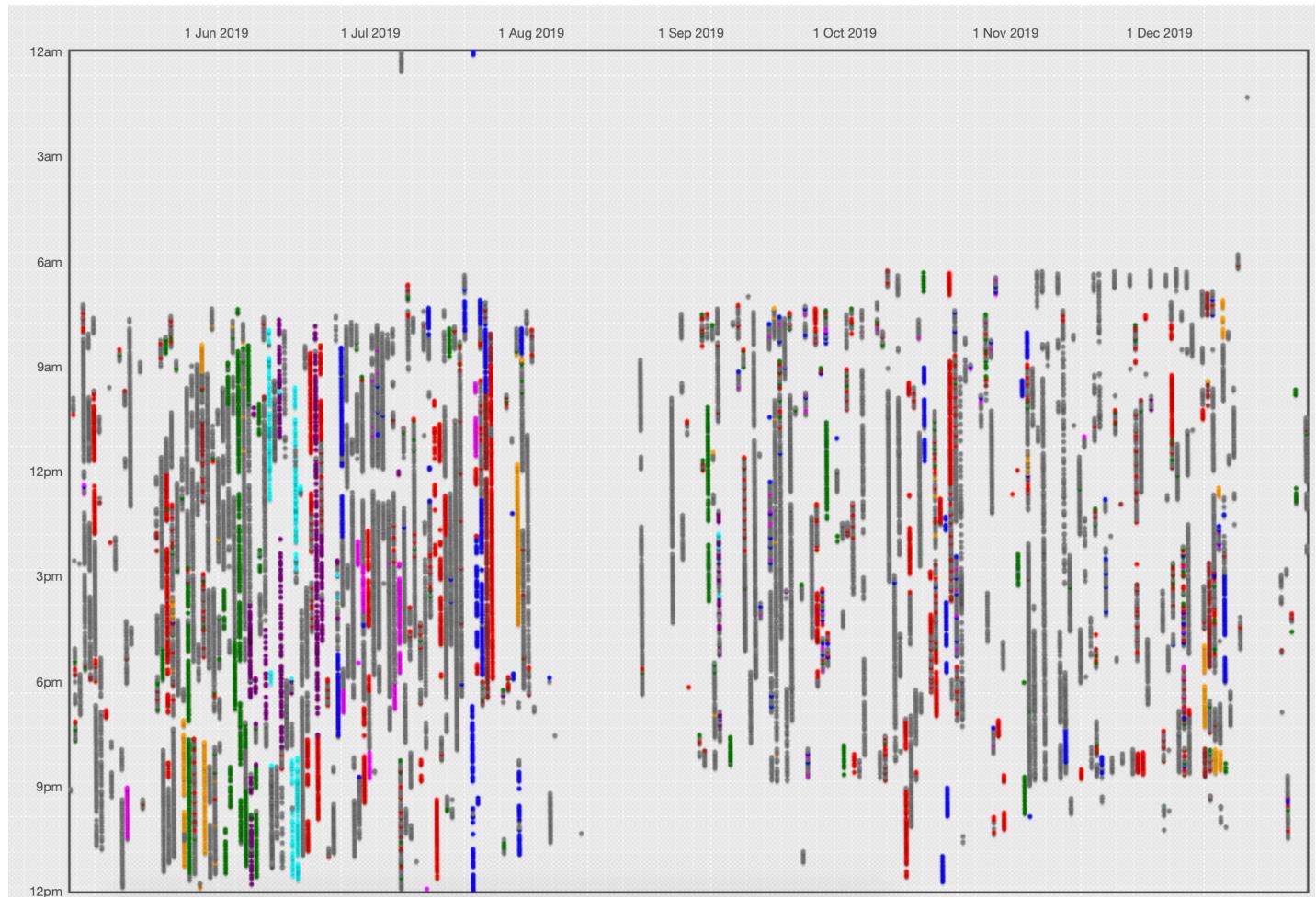
Band	Album	Song	Date
Ariana Grande	Santa Tell Me	Santa Tell Me	null
Lady Antebellum	On This Winter's Night	A Holly Jolly Christmas	29 Dec 2019 08:34
The Big Moon	Carol Of The Bells	Carol Of The Bells	29 Dec 2019 08:31
Darlene Love	Christmas Voices	All Alone On Christmas	29 Dec 2019 08:27
Michael Bublé	Christmas (Deluxe Sp...)	All I Want for Christ...	29 Dec 2019 08:24
Carpenters	Christmas Collection	Sleigh Ride	29 Dec 2019 08:21
She & Him	Christmas Party	Let It Snow	29 Dec 2019 08:19
Eagles	Please Come Home F...	Please Come Home f...	29 Dec 2019 08:16
Eartha Kitt	The Very Best of Eart...	Santa Baby (with He...	29 Dec 2019 08:12
Bing Crosby	Christmas Is A Comin'	Here Comes Santa Cl...	29 Dec 2019 08:09
Idina Menzel	Christmas Wishes	Baby It's Cold Outsid...	29 Dec 2019 08:07
Coldplay	Christmas Lights	Christmas Lights	29 Dec 2019 08:03
Bing Crosby	Bing Crosby Rediscov...	White Christmas	29 Dec 2019 08:00
Elton John	Caribou (Remastered)	Step Into Christmas	29 Dec 2019 07:55
Nat King Cole	The Christmas Song	The Christmas Song (...)	29 Dec 2019 07:52

Попробуем составить самую простую диаграмму, которая отобразит наиболее часто прослушиваемых мною исполнителей.



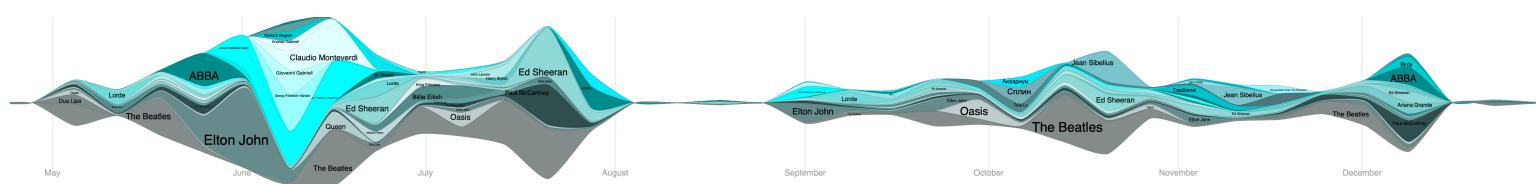
Здесь видно, что больше всего времени я потратил на прослушивание группы The Beatles.

Однако это самая простая выборка. Посмотрим, какие другие варианты отображения этой диаграммы есть. Например, с зависимостью от времени.

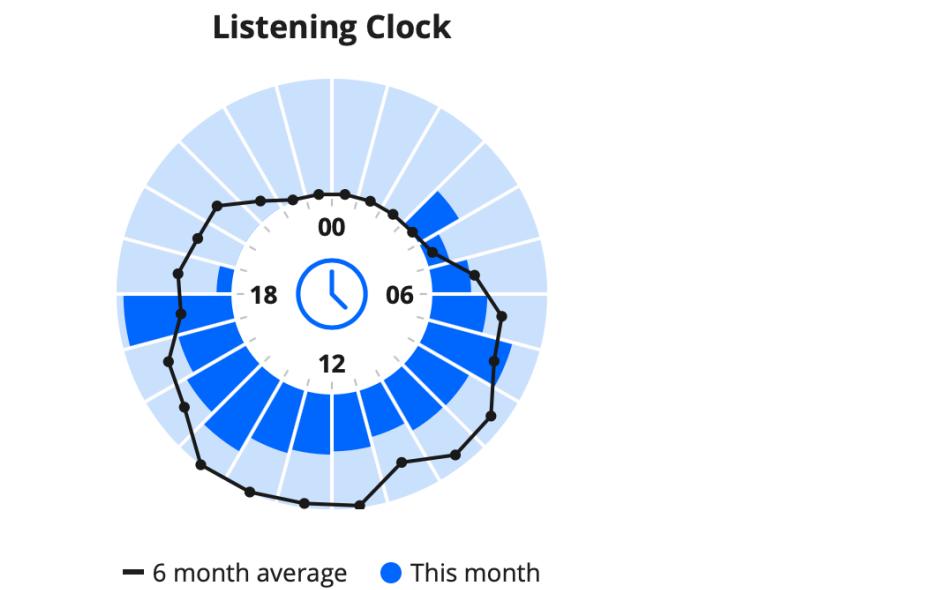


#### Artist Legend

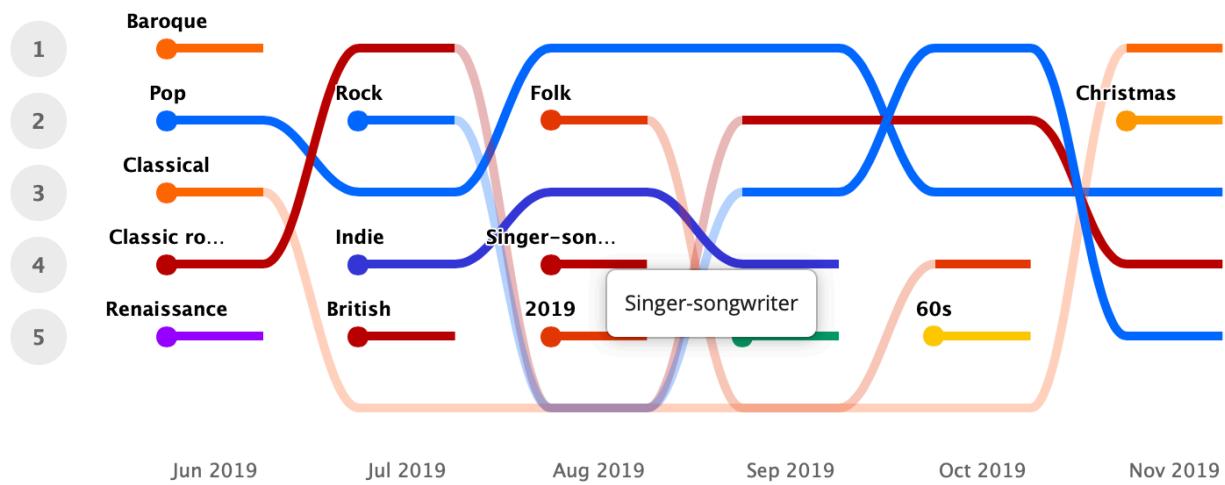
- The Beatles (1624)
- Elton John (811)
- Ed Sheeran (606)
- Jan Pieterszoon Sweelinck (414)
- Paul McCartney (363)
- ABBA (305)
- Claudio Monteverdi (248)
- Billie Eilish (219)
- [Other Artists]



Построим график по времени, когда я чаще всего слушал музыку в течение дня.



Построим график изменения жанров, чтобы отследить, какая музыка была предпочтительнее для меня в определенный период.



Ну и построим обычную таблицу с самыми прослушиваемыми песнями:

1. Elton John - Rocket Man (I Think It's Going To Be A Long, Long Time) (39 plays)
2. Elton John - Tiny Dancer (36 plays)
3. The Beatles - Here Comes The Sun - Remastered 2009 (36 plays)
4. The Beatles - Let It Be - Remastered 2009 (33 plays)
5. The Beatles - Yesterday - Remastered 2009 (31 plays)

- 
6. John Lennon - Imagine - Remastered (29 plays)
  7. The Beatles - Strawberry Fields Forever - Remastered 2009 (27 plays)
  8. Ed Sheeran - Shape of You (25 plays)
  9. Elton John - Don't Go Breaking My Heart (25 plays)
  10. The Beatles - In My Life - Remastered 2009 (25 plays)

На основе полученных графиков можно сделать вывод, что я люблю слушать музыку :). А если точнее, то весной я слушал Элтона Джона, что может быть связано с выходом фильма "Rocketman", летом — Эда Ширана, т.к. посетил его концерт в июле в Москве. А уже с ноября начал слушать рождественскую музыку.

"Часы прослушивания" показывают, что я слушаю музыку почти все время. Особенно явно заметен прирост в 7-8 утра, когда я еду на учебу/работу.

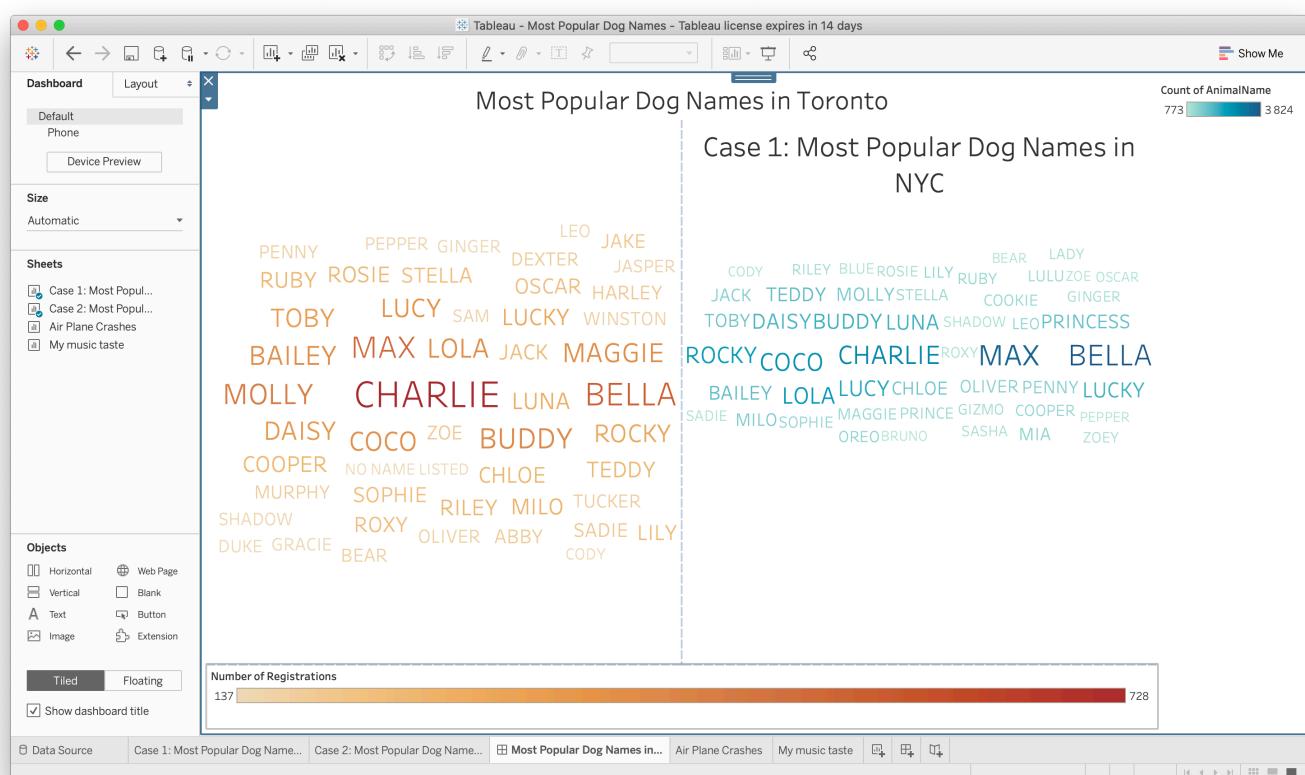
# Результаты

Были проведены три независимых друг от друга исследования:

- Анализ наиболее популярных собачьих кличек в Торонто;
- Анализ авиакатастроф с 1908 по 2009 г.
- Анализ моих музыкальных предпочтений

Выводы были приведены в каждой главе соответственно. Ниже продублированы наиболее интересные диаграммы для каждого из трех исследований.

1) Сравнение популярных собачьих кличек в Торонто и в Нью-Йорке:



2012, 2013, 2014, 2015, 2016...

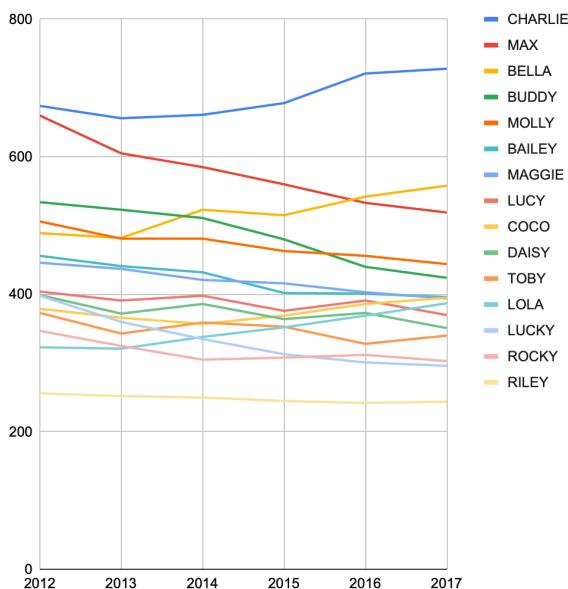
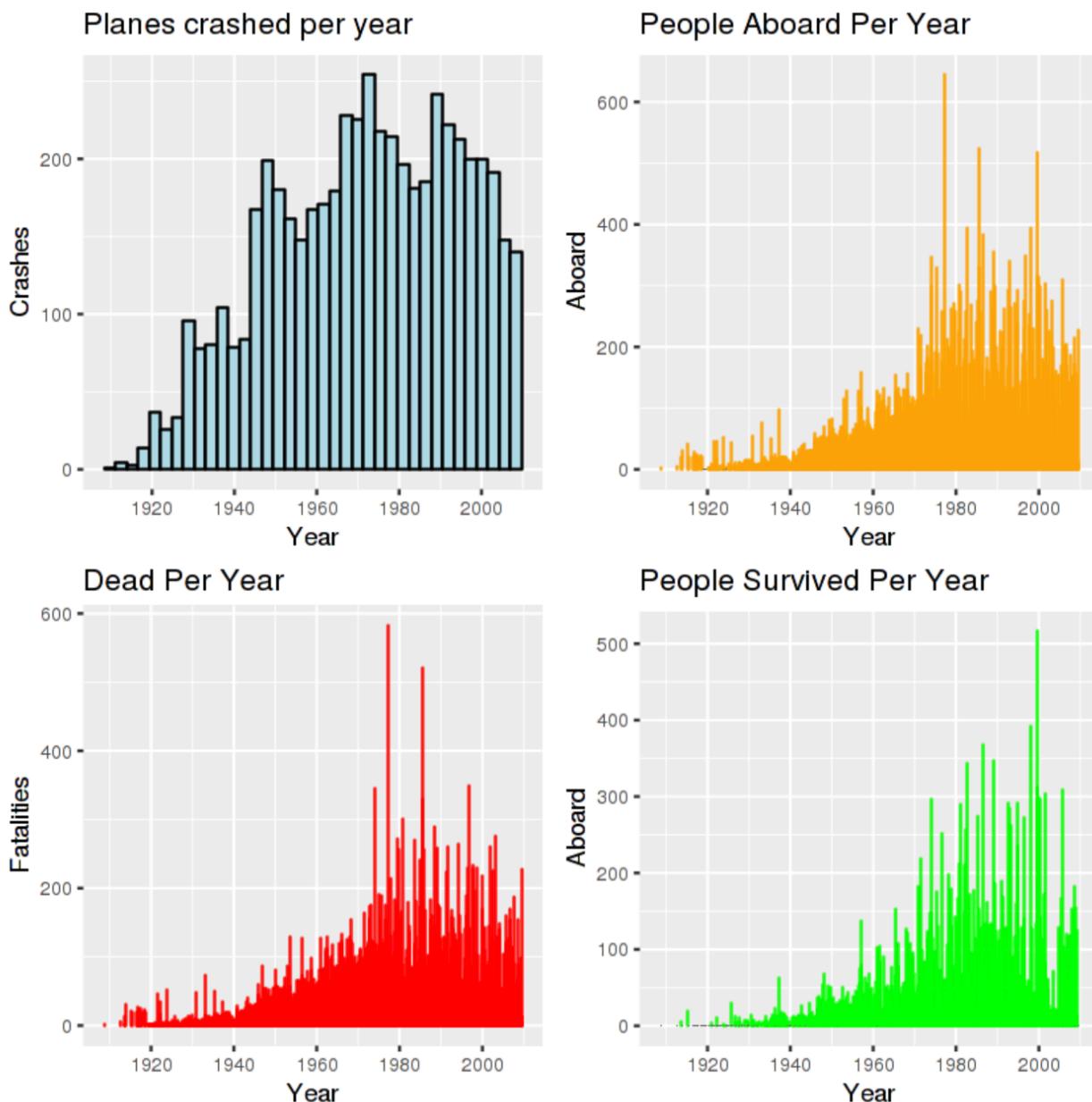
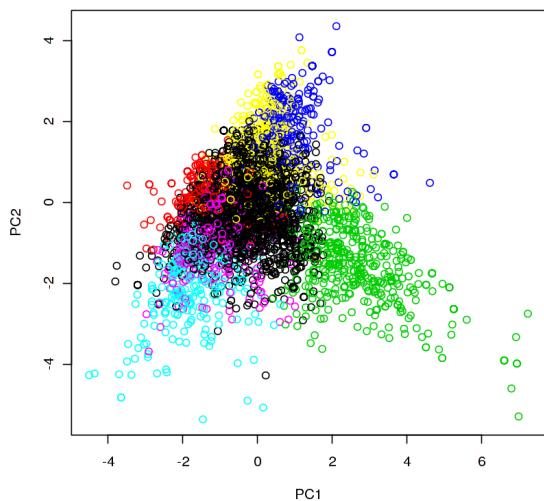


График изменения предпочтений жителей Торонто в выборе кличек своим питомцам.

2) Диаграммы авиакатастроф

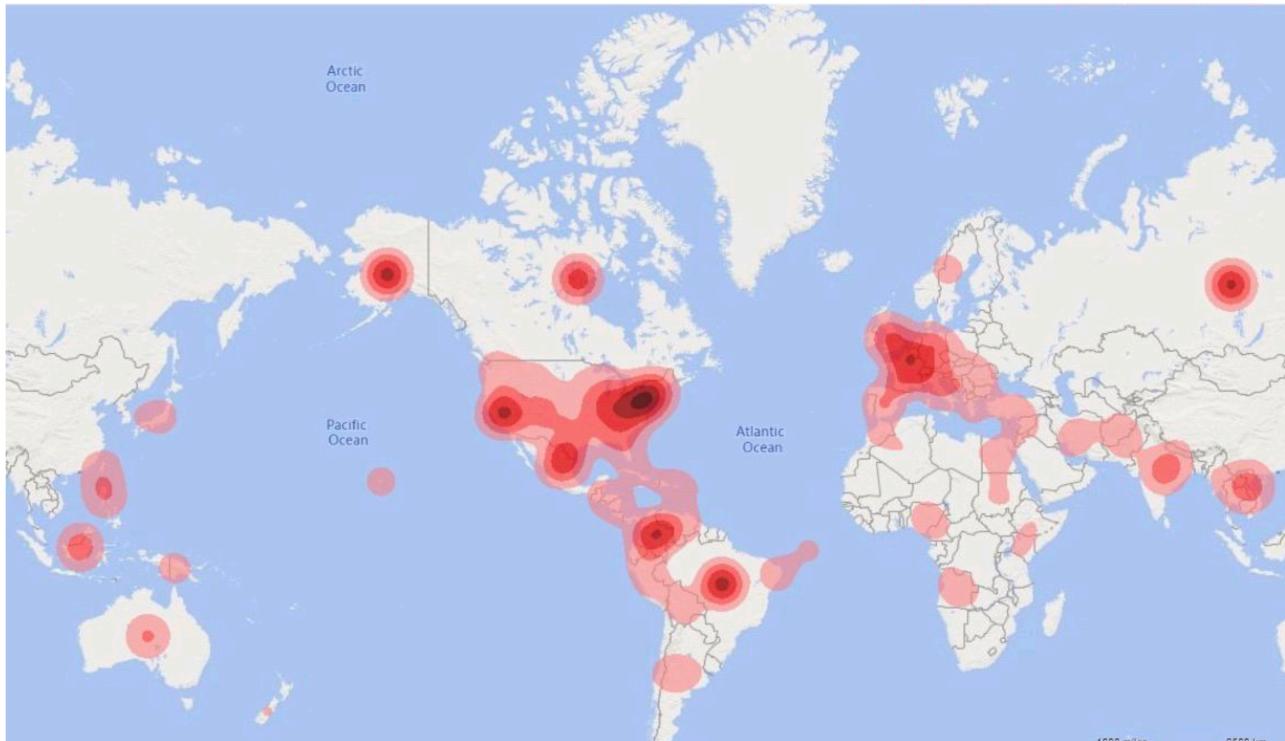


7

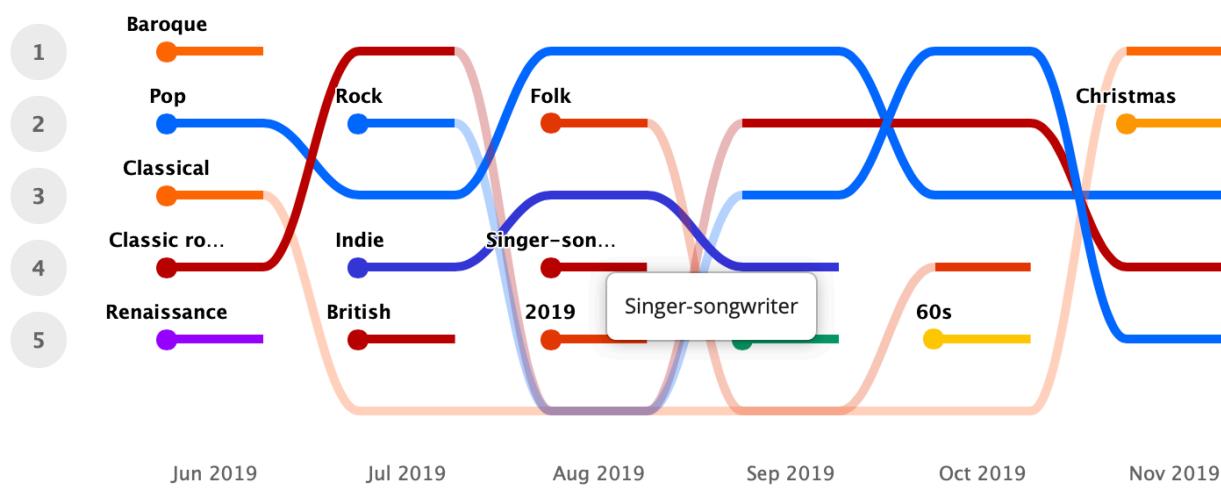


Разбиение на кластеры (причины авиакатастроф)

## Тепловая карта авиакатастроф



3) Мои музыкальные предпочтения в течение полугода.



---

## Информационные источники

1. <https://open.toronto.ca/dataset/licensed-dog-and-cat-names/> - датасеты для первого исследования
2. <https://opendata.socrata.com/Government/Airplane-Crashes-and-Fatalities-Since-1908/q2te-8cvq> - датасет с авиакатастрофами
3. <http://tableau.com> - сайт используемого ПО
4. <https://github.com/morsstepan/data-analysis> - мой репозиторий с датасетами, статьей и рабочей тетрадью Tableau.