

Рубежный контроль №1

Группа	ФИО	Номер в списке	Задание 1	Задание 2	Задание 3
ИУ5- 24М	Леонид Перлин	9	Для набора данных проводите устранение пропусков для одного (произвольного) числового признака с использованием метода заполнения "хвостом распределения".	Для набора данных проводите удаление константных и псевдоконстантных признаков.	Для произвольной колонки данных построить график "Скрипичная диаграмма (violin plot)"

§0. Настройка

```
In [78]: pip install plotly tabulate tqdm numpy pandas matplotlib
```

```
Requirement already satisfied: plotly in /Users/blacksnow/opt/anaconda3/envs/snowflakes/lib/python3.10/site-packages (5.13.1)
Requirement already satisfied: tabulate in /Users/blacksnow/opt/anaconda3/envs/snowflakes/lib/python3.10/site-packages (0.9.0)
Requirement already satisfied: tqdm in /Users/blacksnow/opt/anaconda3/envs/snowflakes/lib/python3.10/site-packages (4.64.1)
Requirement already satisfied: numpy in /Users/blacksnow/opt/anaconda3/envs/snowflakes/lib/python3.10/site-packages (1.23.5)
Requirement already satisfied: pandas in /Users/blacksnow/opt/anaconda3/envs/snowflakes/lib/python3.10/site-packages (1.5.3)
Requirement already satisfied: matplotlib in /Users/blacksnow/opt/anaconda3/envs/snowflakes/lib/python3.10/site-packages (3.6.2)
Requirement already satisfied: tenacity>=6.2.0 in /Users/blacksnow/opt/anaconda3/envs/snowflakes/lib/python3.10/site-packages (from plotly) (8.2.2)
Requirement already satisfied: python-dateutil>=2.8.1 in /Users/blacksnow/opt/anaconda3/envs/snowflakes/lib/python3.10/site-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /Users/blacksnow/opt/anaconda3/envs/snowflakes/lib/python3.10/site-packages (from pandas) (2022.7.1)
Requirement already satisfied: pillow>=6.2.0 in /Users/blacksnow/opt/anaconda3/envs/snowflakes/lib/python3.10/site-packages (from matplotlib) (9.3.0)
Requirement already satisfied: cycler>=0.10 in /Users/blacksnow/opt/anaconda3/envs/snowflakes/lib/python3.10/site-packages (from matplotlib) (0.11.0)
Requirement already satisfied: pyparsing>=2.2.1 in /Users/blacksnow/opt/anaconda3/envs/snowflakes/lib/python3.10/site-packages (from matplotlib) (3.0.9)
Requirement already satisfied: contourpy>=1.0.1 in /Users/blacksnow/opt/anaconda3/envs/snowflakes/lib/python3.10/site-packages (from matplotlib) (1.0.5)
Requirement already satisfied: packaging>=20.0 in /Users/blacksnow/opt/anaconda3/envs/snowflakes/lib/python3.10/site-packages (from matplotlib) (23.0)
Requirement already satisfied: fonttools>=4.22.0 in /Users/blacksnow/opt/anaconda3/envs/snowflakes/lib/python3.10/site-packages (from matplotlib) (4.25.0)
Requirement already satisfied: kiwisolver>=1.0.1 in /Users/blacksnow/opt/anaconda3/envs/snowflakes/lib/python3.10/site-packages (from matplotlib) (1.4.4)
Requirement already satisfied: six>=1.5 in /Users/blacksnow/opt/anaconda3/envs/snowflakes/lib/python3.10/site-packages (from python-dateutil>=2.8.1->pandas) (1.16.0)
Note: you may need to restart the kernel to use updated packages.
```

```
In [79]: from tabulate import tabulate
import pandas as pd
import numpy as np
from tqdm import tqdm
import plotly.express as px
from sklearn.impute import MissingIndicator, SimpleImputer
import seaborn as sns
import os
import plotly.io as pio
pio.renderers.default = 'notebook'

%matplotlib inline
```

Папка для графиков

```
In [80]: if not os.path.exists("images"):
    os.mkdir("images")
```

§1. Задание, вариант и набор данных

```
In [81]: in_group_list_num = 9
group_name = 'ИУ5-24М'
name = 'Леонид Перлин'
task_1_num = 9
task_2_num = 29

task_1_text = '''
Для набора данных проведите устранение пропусков для одного (произвольного)
'''

task_2_text = '''
Для набора данных проведите удаление константных и псевдоконстантных признаков
'''

task_3_text = '''
Для произвольной колонки данных построить график "Скрипичная диаграмма (violin plot)"
'''

print(tabulate([[group_name, name,in_group_list_num,task_1_text,task_2_text,
| Группа | ФИО | Номер в списке | Задание 1 | Задание 2 | Задание 3 |
```

Группа	ФИО	Номер в списке	Задание 1	Задание 2	Задание 3
ИУ5-24М Леонид Перлин 9 Для набора данных проведите устранение пропусков для одного (произвольного) числового признака с исполь- зованием метода заполнения "хвостом распределения". Для набора данных про- ведите удаление константных и псевдоконстантных признаков. Для произволь- ной колонки данных построить график "Скрипичная диаграмма (violin plot)"					

Первый датасет (Задание №3)

[kaggle](#) Spotify топ чарты в мире из Kaggle

Второй датасет (Задания №1 и №2)

`df_bike` взят из приложения "Здоровье" с моего телефона и представляет собой набор данных с велопоездки

```
In [82]: df_bike = pd.read_csv('./data/route_2023-03-03_4.16pm/track_points.csv')
df_bike.head()
```

Out[82]:

	X	Y	track_fid	track_seg_id	track_seg_point_id	ele	t
0	37.685089	55.765370	0	0	0	146.670151	2023/03 12:57:19
1	37.685092	55.765369	0	0	0	146.682946	2023/03 12:57:20
2	37.685095	55.765369	0	0	0	146.693712	2023/03 12:57:21
3	37.685099	55.765368	0	0	0	146.700226	2023/03 12:57:22
4	37.685102	55.765368	0	0	0	146.705792	2023/03 12:57:23

5 rows × 32 columns

In [83]:

```
df = pd.read_csv('data/charts.csv')
df.head()
```

Out[83]:

	title	rank	date	artist	url	regi
0	Chantaje (feat. Maluma)	1	2017- 01-01	Shakira	https://open.spotify.com/track/6mlCuAdrwEjh6Y6...	Argent
1	Vente Pa' Ca (feat. Maluma)	2	2017- 01-01	Ricky Martin	https://open.spotify.com/track/7DM4BPaS7uoffFul...	Argent
2	Reggaetón Lento (Bailemos)	3	2017- 01-01	CNCO	https://open.spotify.com/track/3AEZUABDXNtecAO...	Argent
3	Safari	4	2017- 01-01	J Balvin, Pharrell Williams, BIA, Sky	https://open.spotify.com/track/6rQSrBHf7HIZjt...	Argent
4	Shaky Shaky	5	2017- 01-01	Daddy Yankee	https://open.spotify.com/track/58IL315gMSTD37D...	Argent

In [84]:

```
df.shape
```

Out[84]:

```
(26173514, 9)
```

Количество пропусков

In [85]:

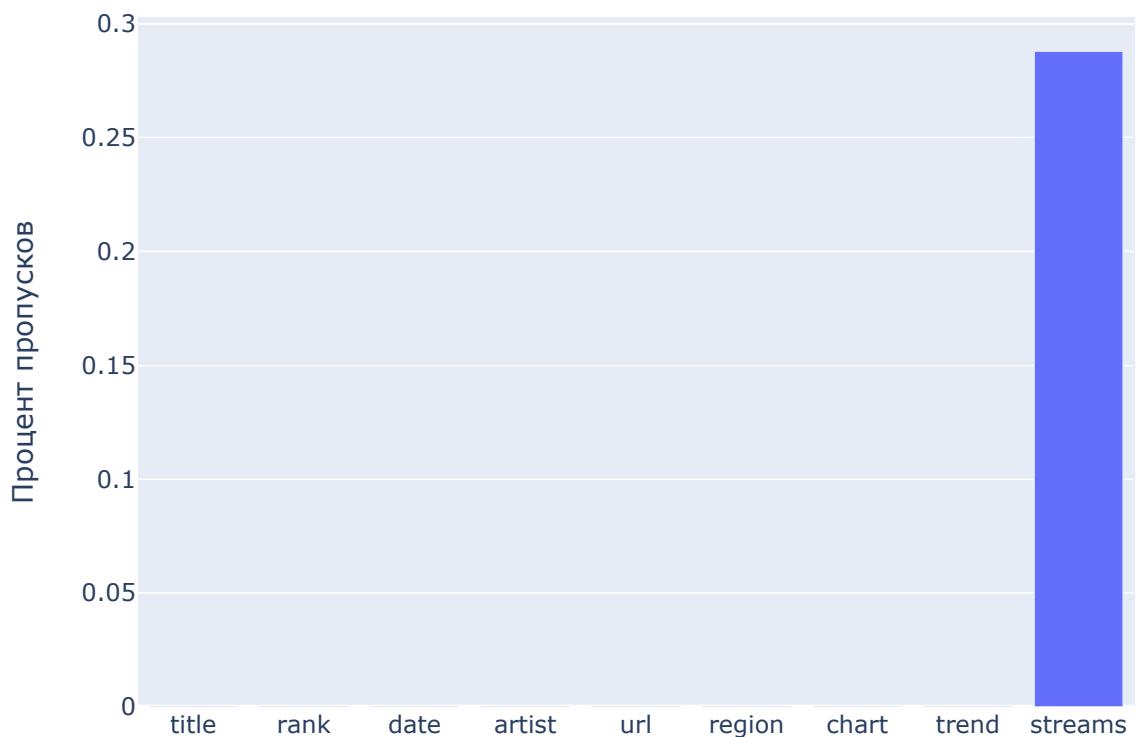
```
def analyze(df=df):
    table_headers = ['Название столбца', 'Процент пропусков', 'Кол-во уникальных значений']
    table = np.empty([3, len(df.columns)], dtype=np.dtype(object))
    table[0] = [col for col in tqdm(df.columns)]
    table[1] = [df[col].isna().sum() / df[col].notna().sum() for col in tqdm(df.columns)]
    table[2] = [df[col].nunique() for col in tqdm(df.columns)]
    table = table.transpose()
    print(tabulate(table, tablefmt='github', headers=table_headers))
```

```
df_stats = pd.DataFrame(table, columns=table_headers)
# df_stats
fig = px.bar(df_stats, x="Название столбца", y='Процент пропусков', title='Статистика пропусков')
fig.show()
```

In [86]: `analyze()`

100%	[00:00<00:00, 136277.03it/s]	
100%	[00:11<00:00, 1.31s/it]	
100%	[00:11<00:00, 1.24s/it]	
Название столбца	Процент пропусков	Кол-во уникальных значений
title	4.20272e-07	164806
rank	0	200
date	0	1826
artist	6.87719e-07	96156
url	0	217704
region	0	70
chart	0	2
trend	0	4
streams	0.287946	788013

Процент пропусков в датасете



Датасет содержит пропуски в большом количестве строк в колонке `[streams]` -

количество прослушиваний песни

§2. Устранение пропусков для числового признака (хвост распределения)

In [87]: `analyze(df_bike)`

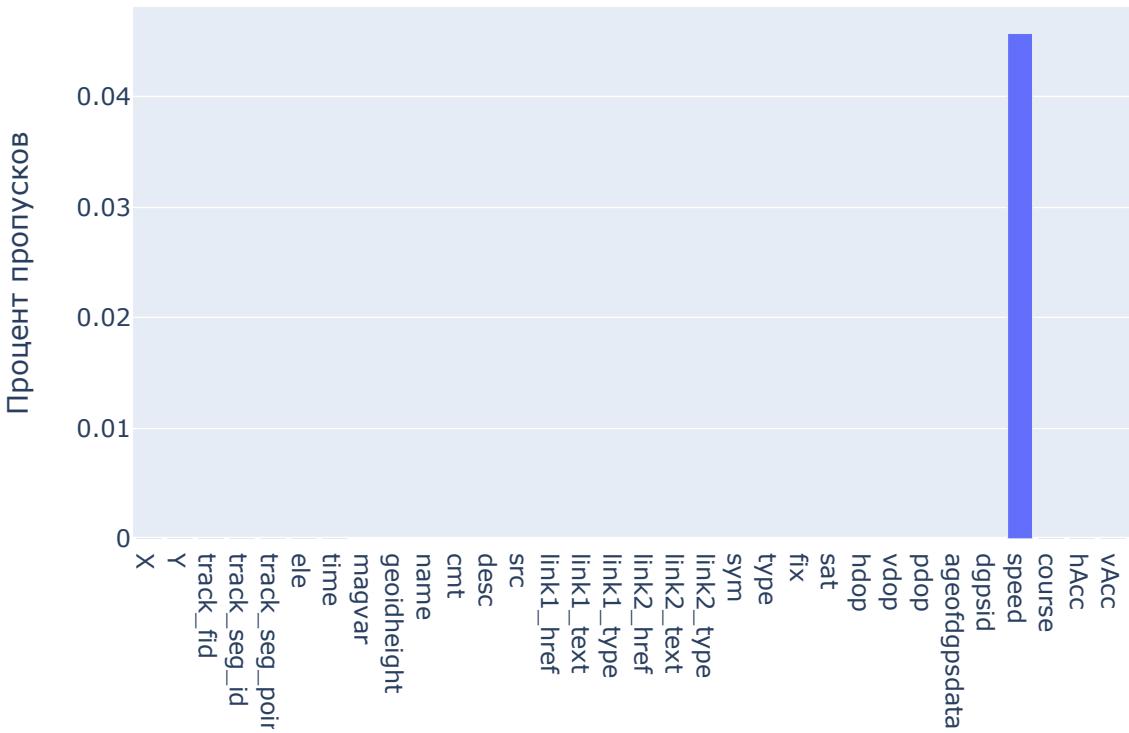
```
100%|██████████| 32/32 [00:00<00:00, 333046.47it/s]
 0%|          | 0/32 [00:00<?, ?it/s]/var/folders/g8/xtt465_57r53fy239yttr
rrm0000gn/T/ipykernel_14914/4023454359.py:5: RuntimeWarning:
```

divide by zero encountered in long_scalars

```
100%|██████████| 32/32 [00:00<00:00, 3707.57it/s]
100%|██████████| 32/32 [00:00<00:00, 6214.93it/s]
```

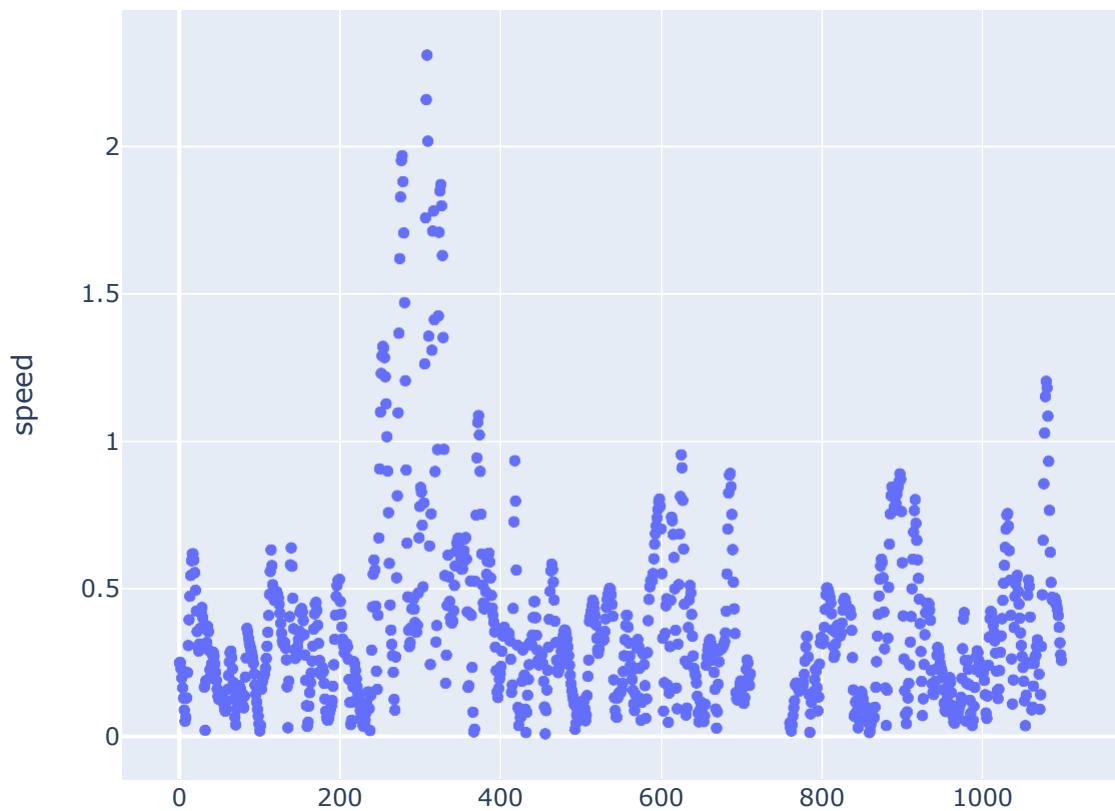
Название столбца	Процент пропусков	Кол-во уникальных значений
X	0	396
Y	0	105
track_fid	0	1
track_seg_id	0	1
track_seg_point_id	0	1098
ele	0	1098
time	0	1089
magvar	inf	0
geoidheight	inf	0
name	inf	0
cmt	inf	0
desc	inf	0
src	inf	0
link1_href	inf	0
link1_text	inf	0
link1_type	inf	0
link2_href	inf	0
link2_text	inf	0
link2_type	inf	0
sym	inf	0
type	inf	0
fix	inf	0
sat	inf	0
hdop	inf	0
vdop	inf	0
pdop	inf	0
ageofdgpsdata	inf	0
dgpsid	inf	0
speed	0.0457143	1049
course	0	1098
hAcc	0	1093
vAcc	0	1092

Процент пропусков в датасете



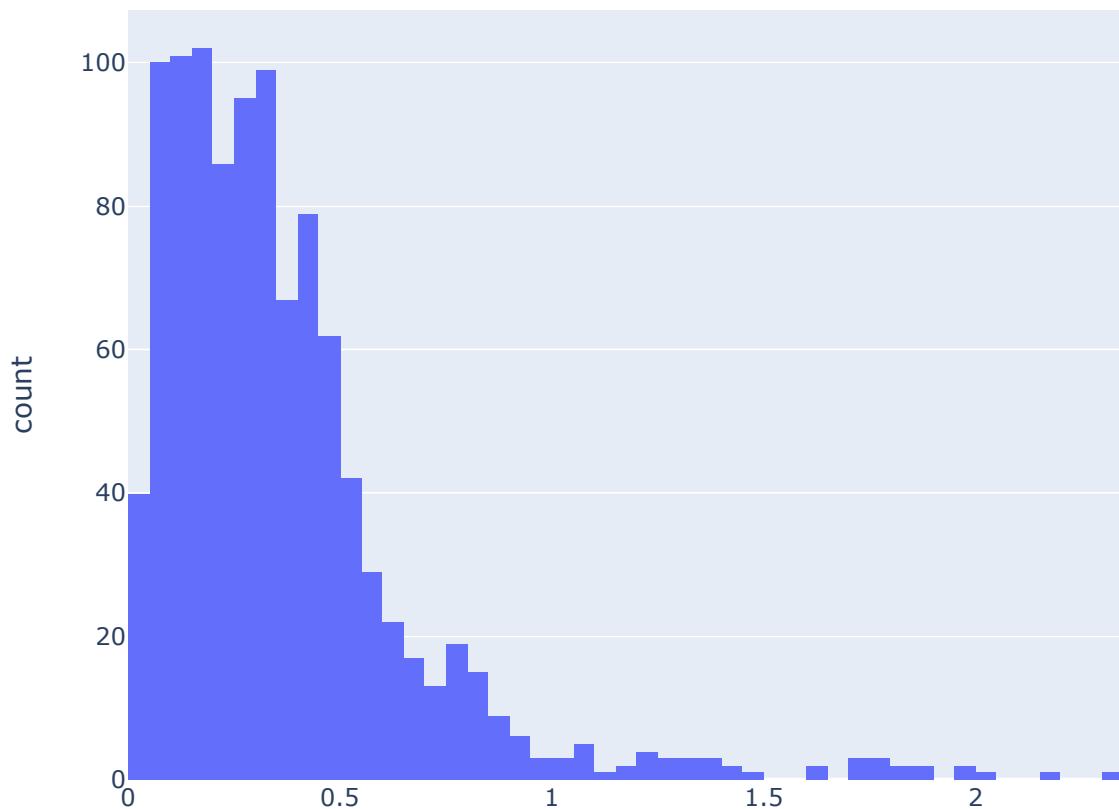
В колонке speed пропущено 4% значений

```
In [88]: fig = px.scatter(df_bike, y='speed')
fig.show()
```



Пропуск данных в промежутке от 700 до 750 строчки

```
In [89]: fig = px.histogram(df_bike,x='speed')
fig.show()
```



```
In [90]: def research_impute_numeric_column(dataset, num_column, const_value=None):
    strategy_params = ['mean', 'median', 'most_frequent', 'constant']
    strategy_params_names = ['Среднее', 'Медиана', 'Мода']
    strategy_params_names.append('Константа = ' + str(const_value))

    original_temp_data = dataset[[num_column]].values
    size = original_temp_data.shape[0]
    original_data = original_temp_data.reshape((size,))

    new_df = pd.DataFrame({'Исходные данные':original_data})

    for i in range(len(strategy_params)):
        strategy = strategy_params[i]
        col_name = strategy_params_names[i]
        if (strategy != 'constant') or (strategy == 'constant' and const_value):
            if strategy == 'constant':
                temp_data, _, _ = impute_column(dataset, num_column, strategy)
            else:
                temp_data, _, _ = impute_column(dataset, num_column, strategy)
        new_df[col_name] = temp_data

    sns.kdeplot(data=new_df)

def impute_column(dataset, column, strategy_param, fill_value_param=None):
```

```

"""
Заполнение пропусков в одном признаке
"""

temp_data = dataset[[column]].values
size = temp_data.shape[0]

indicator = MissingIndicator()
mask_missing_values_only = indicator.fit_transform(temp_data)

imputer = SimpleImputer(strategy=strategy_param,
                         fill_value=fill_value_param)
all_data = imputer.fit_transform(temp_data)

missed_data = temp_data[mask_missing_values_only]
filled_data = all_data[mask_missing_values_only]

return all_data.reshape((size,)), filled_data, missed_data

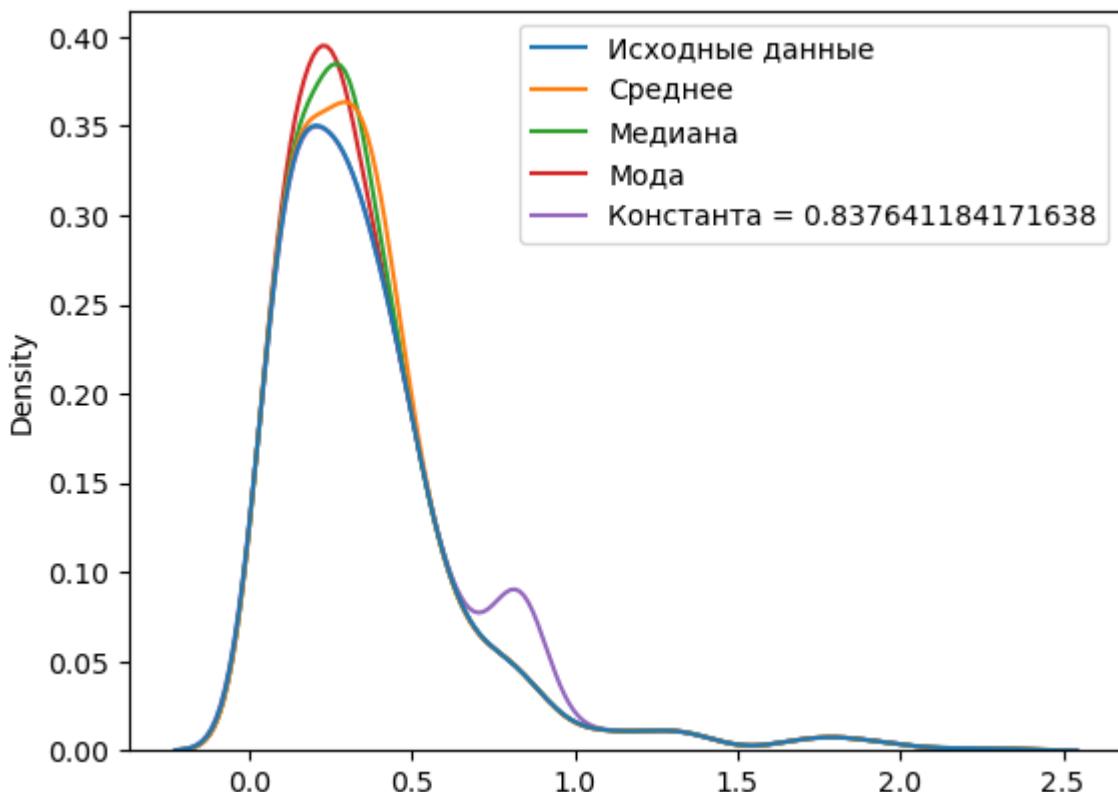
```

```

In [91]: k_param = 1.5
bikeiqr = df_bike['speed'].mean() + k_param*df_bike['speed'].std()
bikeiqr

research_impute_numeric_column(df_bike, 'speed', bikeiqr)

```



```

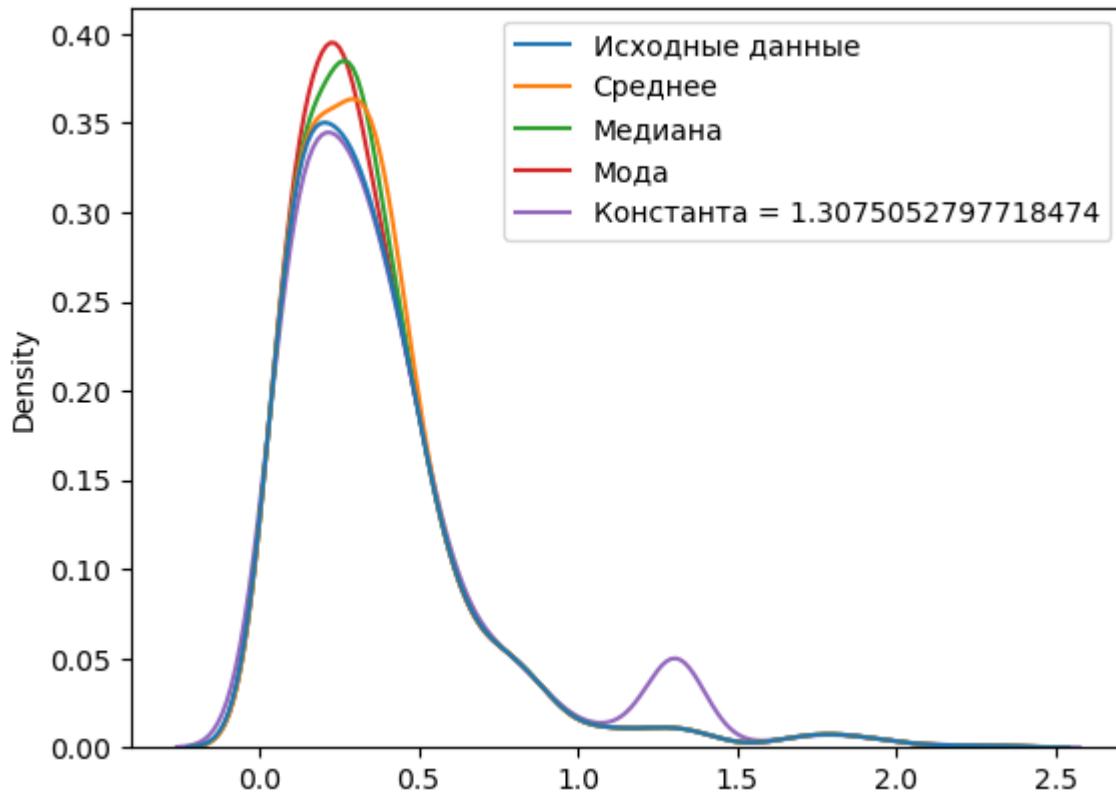
In [92]: IQR = df_bike['speed'].quantile(0.75) - df_bike['speed'].quantile(0.25)
speed_ev1 = df_bike['speed'].quantile(0.75) + k_param*IQR
print('IQR={}, extreme_value={}'.format(IQR, speed_ev1))

```

```
IQR=0.3019427500000004, extreme_value=0.9147161250000001
```

```
In [93]: k_param = 3
bikeiqr = df_bike['speed'].mean() + k_param*df_bike['speed'].std()
bikeiqr

research_impute_numeric_column(df_bike, 'speed', bikeiqr)
```



§3. Удаление константных и псевдоконстантных признаков

```
In [94]: print(df_bike['X'].max() - df_bike['X'].min())
print(df_bike['Y'].max() - df_bike['Y'].min())
df_bike = df_bike.dropna(axis=1, how='all')
df_bike.head()
```

0.0005450000000024602
0.00010600000000238197

Out[94]:

	X	Y	track_fid	track_seg_id	track_seg_point_id	ele	t
0	37.685089	55.765370	0	0		0	146.670151 2023/03 12:57:19
1	37.685092	55.765369	0	0		1	146.682946 2023/03 12:57:20
2	37.685095	55.765369	0	0		2	146.693712 2023/03 12:57:21
3	37.685099	55.765368	0	0		3	146.700226 2023/03 12:57:22
4	37.685102	55.765368	0	0		4	146.705792 2023/03 12:57:23

In [95]:

```
from sklearn.preprocessing import OrdinalEncoder
ord_enc = OrdinalEncoder()
df_bike["time"] = ord_enc.fit_transform(df_bike[["time"]])
df_bike.head()
```

Out[95]:

	X	Y	track_fid	track_seg_id	track_seg_point_id	ele	time
0	37.685089	55.765370	0	0		0	146.670151 0.0 0.
1	37.685092	55.765369	0	0		1	146.682946 1.0 0.
2	37.685095	55.765369	0	0		2	146.693712 2.0 0.
3	37.685099	55.765368	0	0		3	146.700226 3.0 0
4	37.685102	55.765368	0	0		4	146.705792 4.0 0.

In [96]:

```
from sklearn.feature_selection import VarianceThreshold
from sklearn.preprocessing import OrdinalEncoder
ord_enc = OrdinalEncoder()

var_tresh = VarianceThreshold(threshold=0)
var_tresh.fit(df_bike)
var_tresh.variances_
```

Out[96]:

```
array([1.22733494e-08, 4.97191848e-10, 0.00000000e+00, 0.00000000e+00,
       1.09700000e+03, 2.09481995e+01, 1.08800000e+03, 9.80275596e-02,
       3.54440562e+02, 1.00281684e-01, 9.92613973e-03])
```

In [97]:

```
df_bike.shape
```

Out[97]:

```
(1098, 11)
```

In [98]:

```
var_tresh.transform(df_bike).shape
```

Out[98]:

```
(1098, 9)
```

In [99]:

```
df_bike.tail()
```

Out[99]:

	X	Y	track_fid	track_seg_id	track_seg_point_id	ele	tim
1093	37.684871	55.765338	0	0		1093	151.253197 1084
1094	37.684865	55.765337	0	0		1094	150.856628 1085
1095	37.684860	55.765336	0	0		1095	150.463774 1086
1096	37.684856	55.765335	0	0		1096	150.075060 1087
1097	37.684852	55.765334	0	0		1097	149.689613 1088

§4. График "Скрипичная диаграмма"

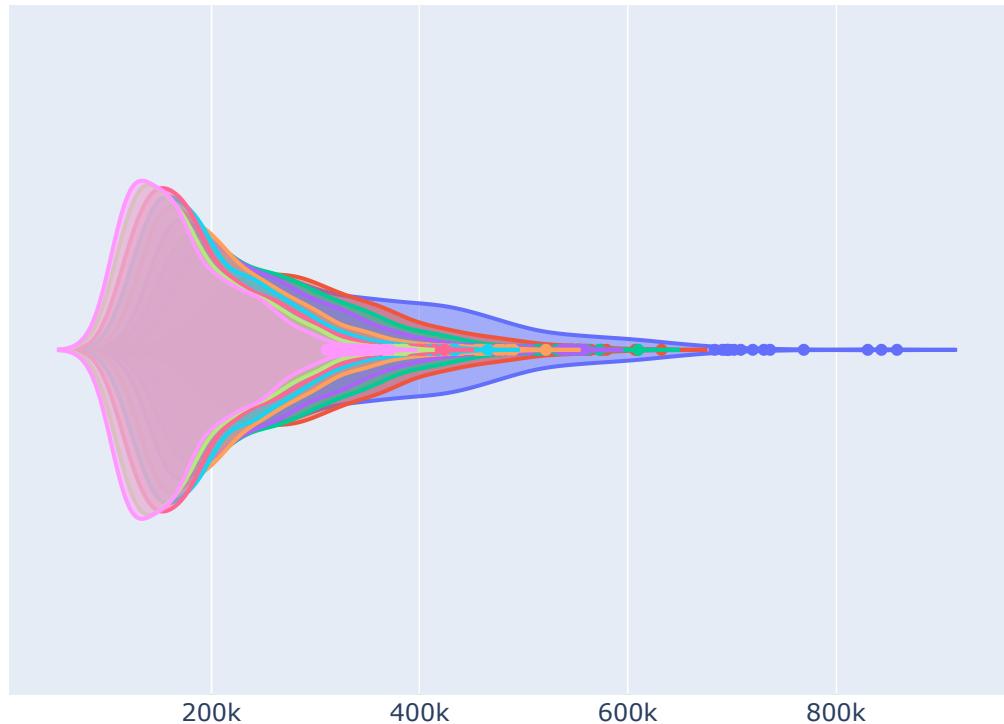
In [100...]

```
for region in df['region'].unique():
    print(region)
    region_df = df[df['region']==region]
    region_df = region_df[region_df['rank']<10]

    fig = px.violin(region_df,x='streams',color='rank',violinmode='overlay',
    fig.show()
```

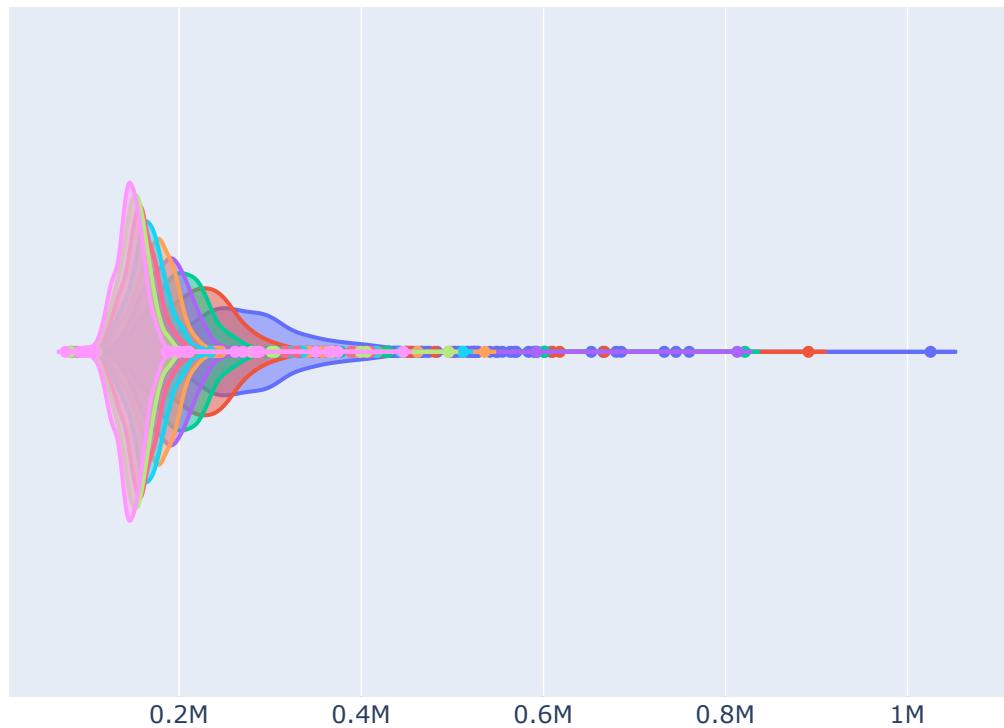
Argentina

Количество прослушиваний песен из топ-10 в Argentina



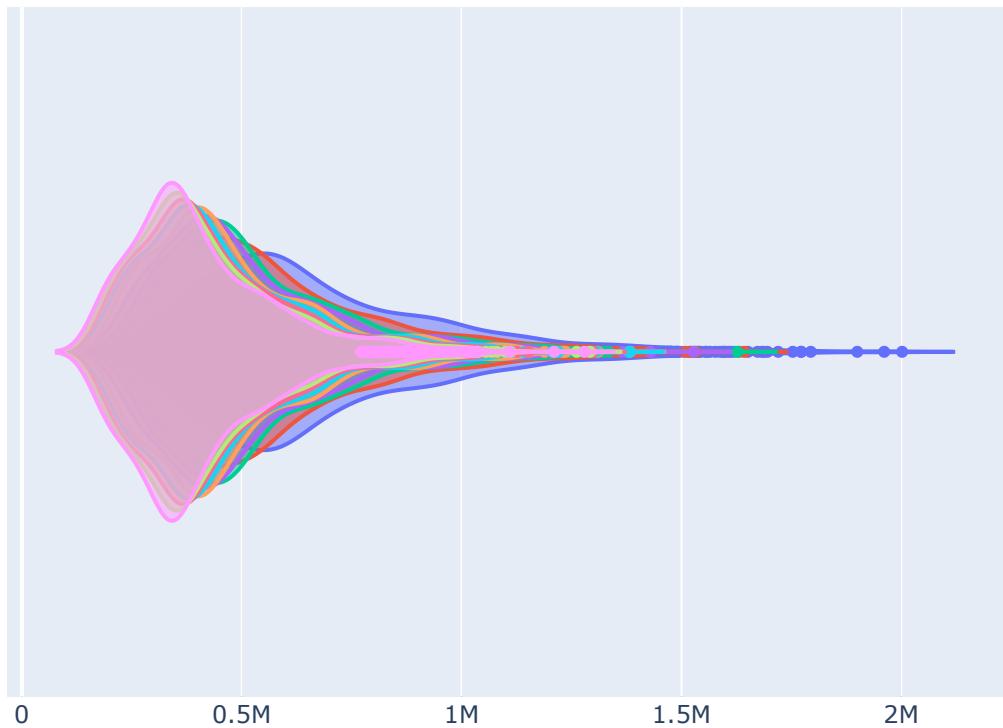
Australia

Количество прослушиваний песен из топ-10 в Australia



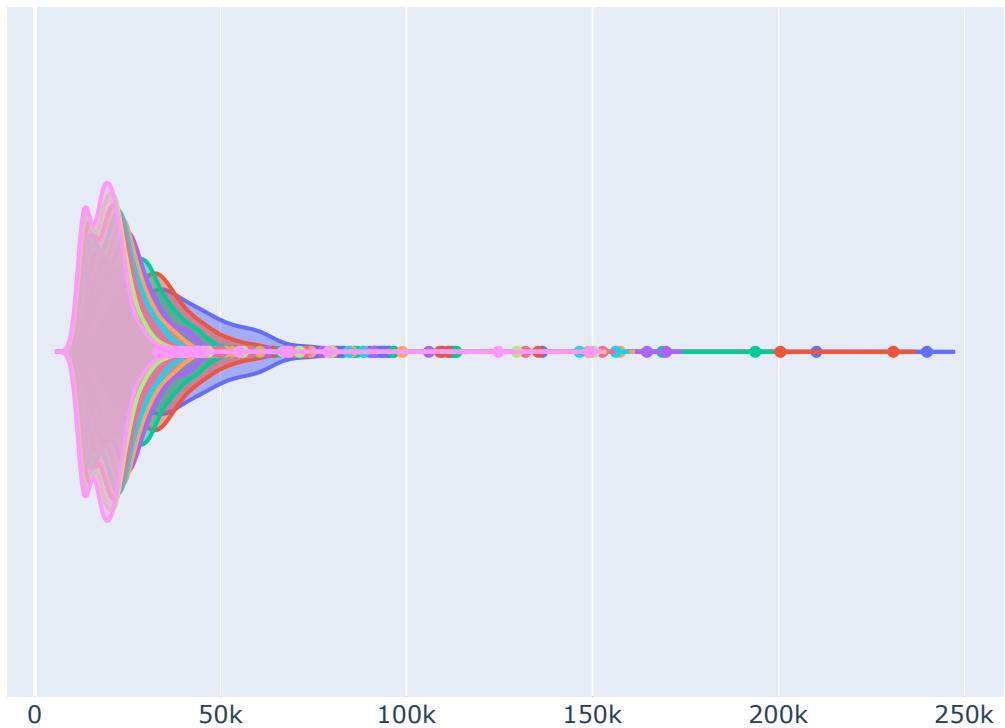
Brazil

Количество прослушиваний песен из топ-10 в Brazil



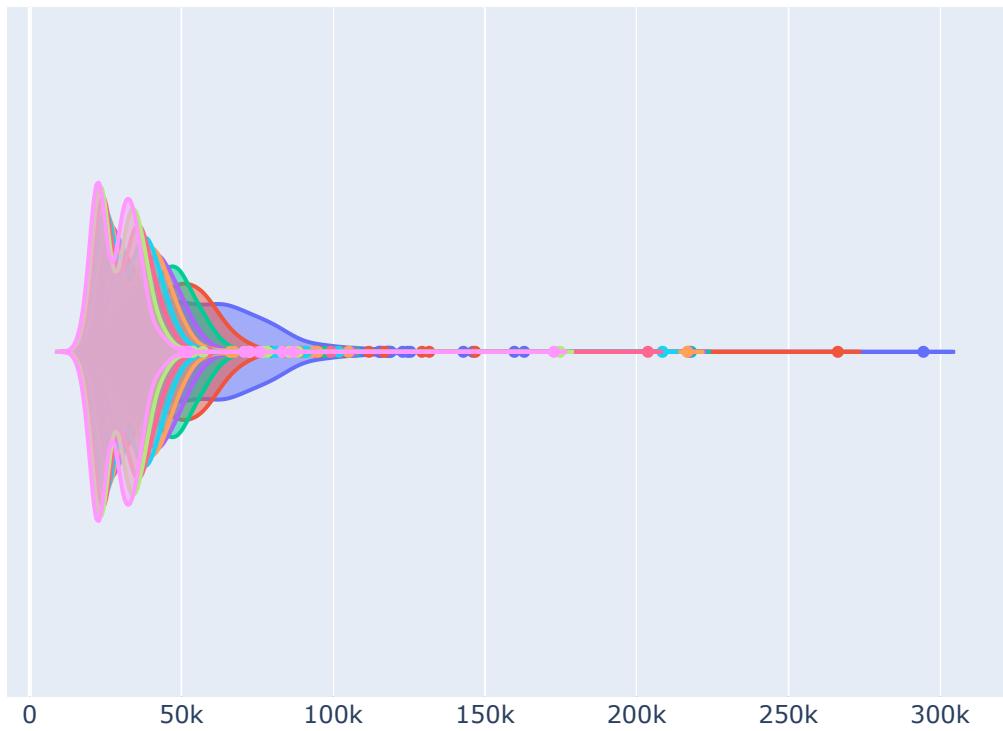
Austria

Количество прослушиваний песен из топ-10 в Austria



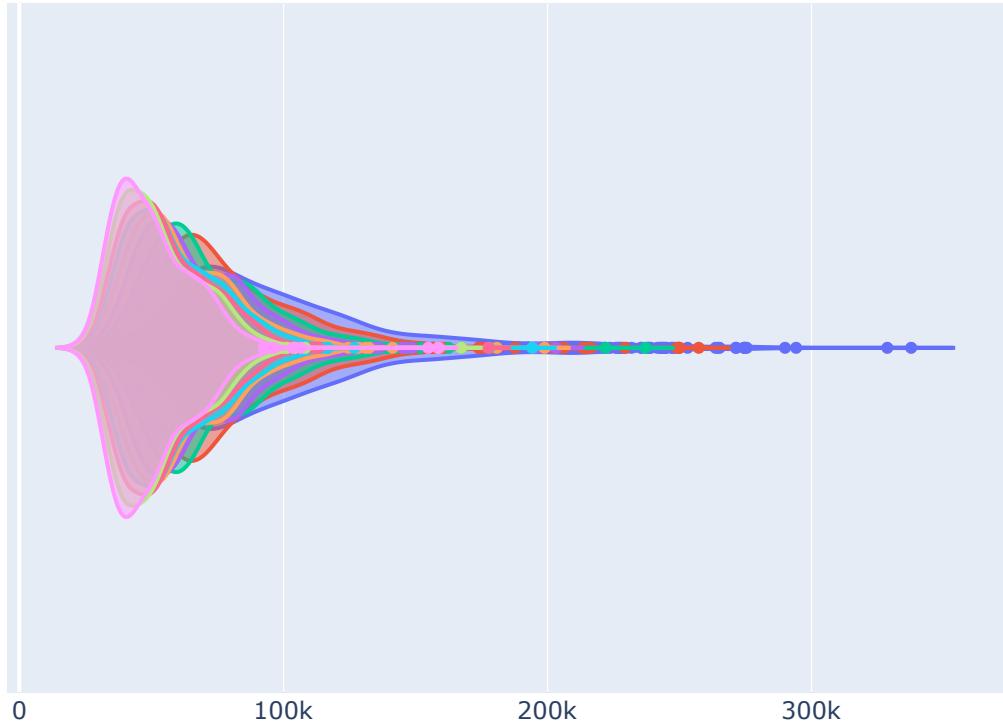
Belgium

Количество прослушиваний песен из топ-10 в Belgium



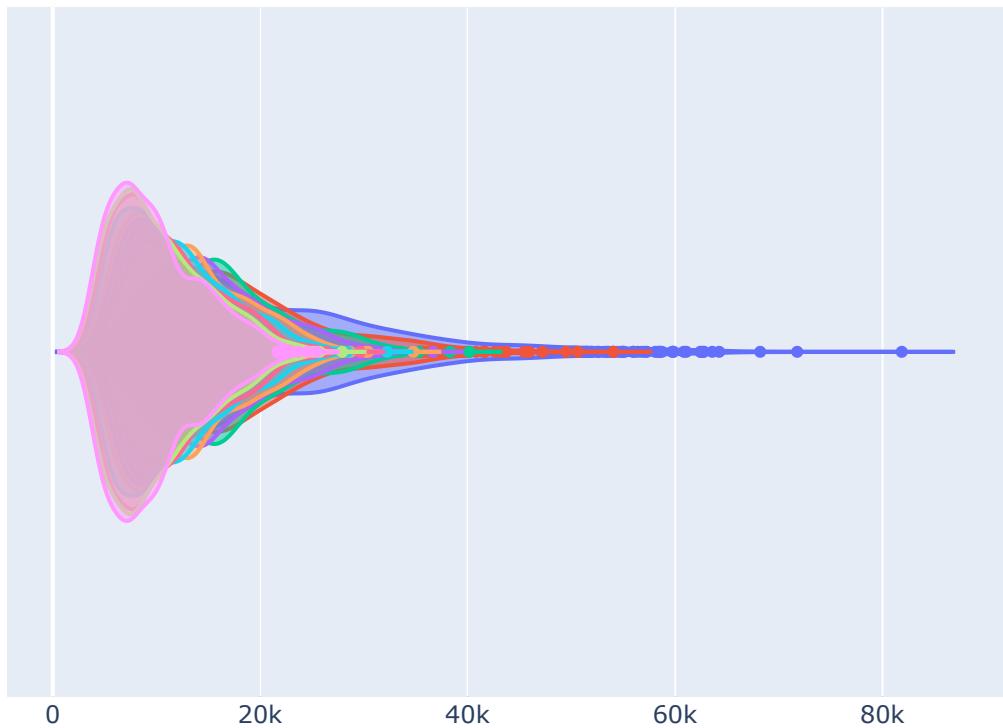
Colombia

Количество прослушиваний песен из топ-10 в Colombia



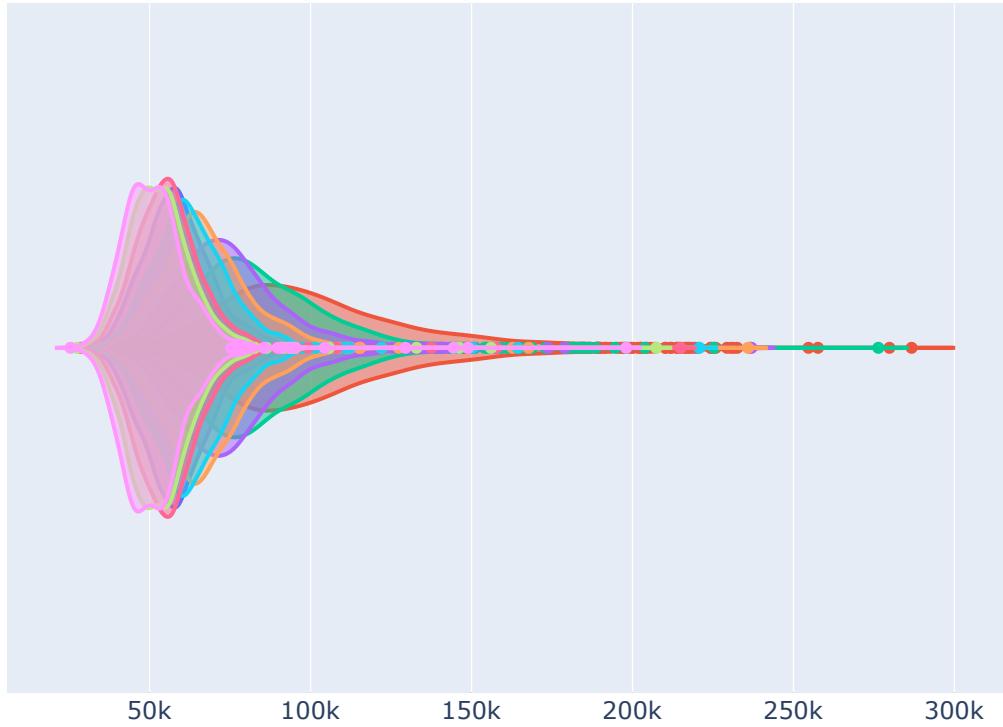
Bolivia

Количество прослушиваний песен из топ-10 в Bolivia



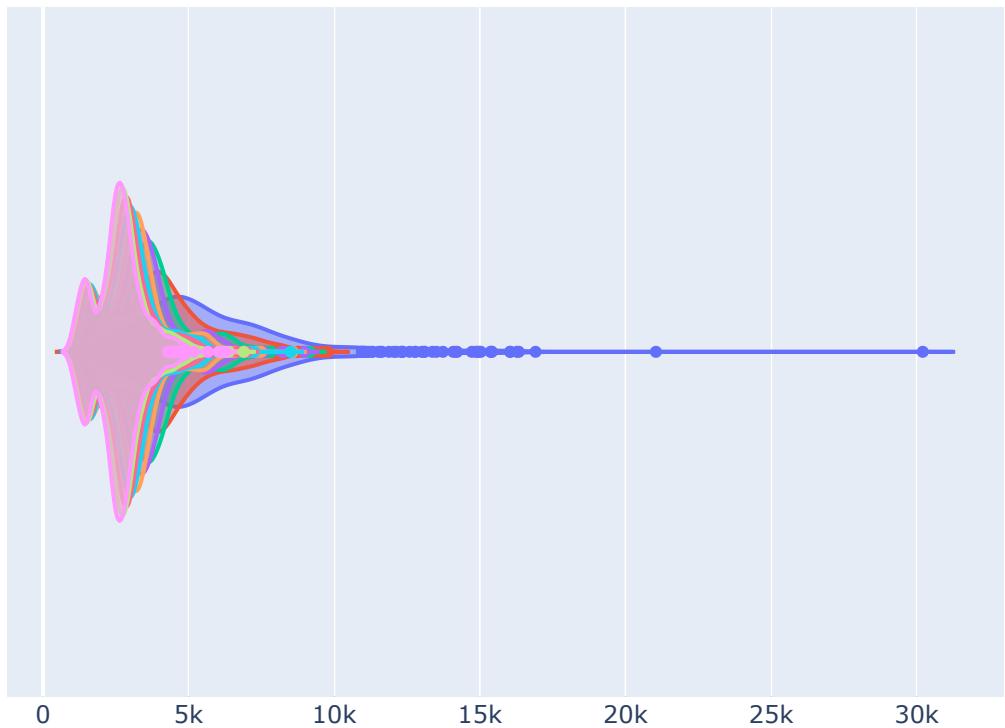
Denmark

Количество прослушиваний песен из топ-10 в Denmark



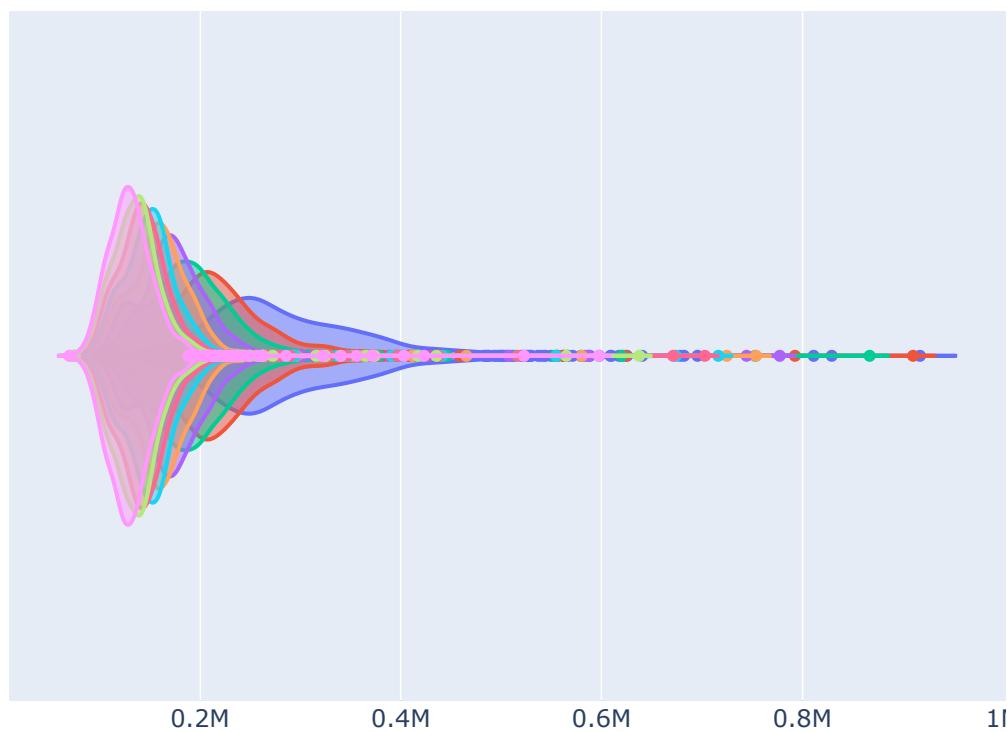
Bulgaria

Количество прослушиваний песен из топ-10 в Bulgaria



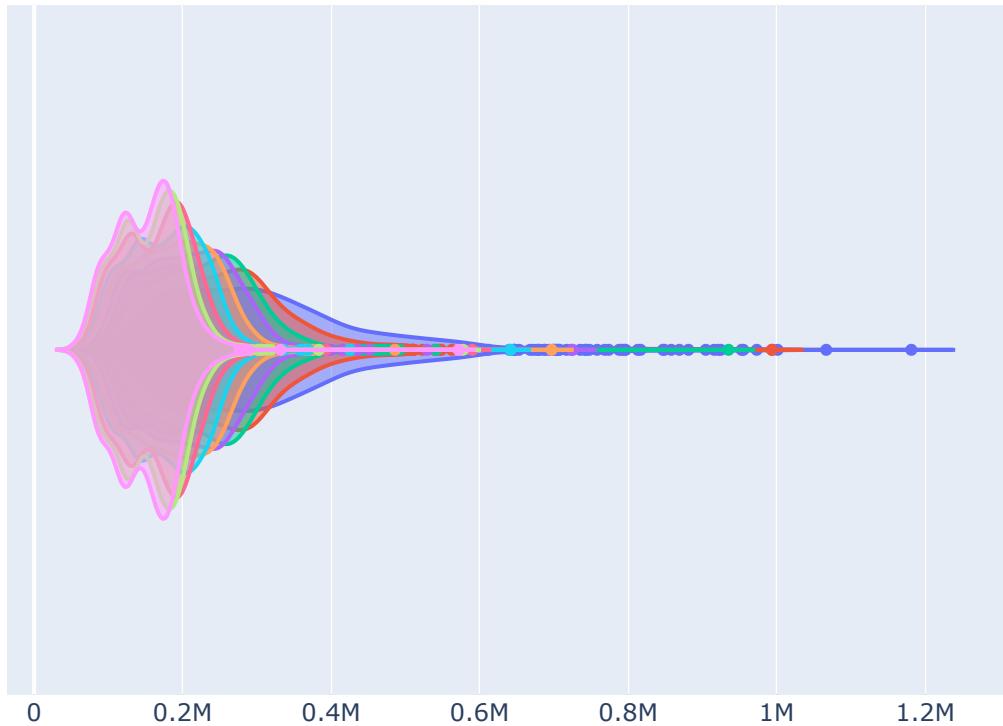
Canada

Количество прослушиваний песен из топ-10 в Canada



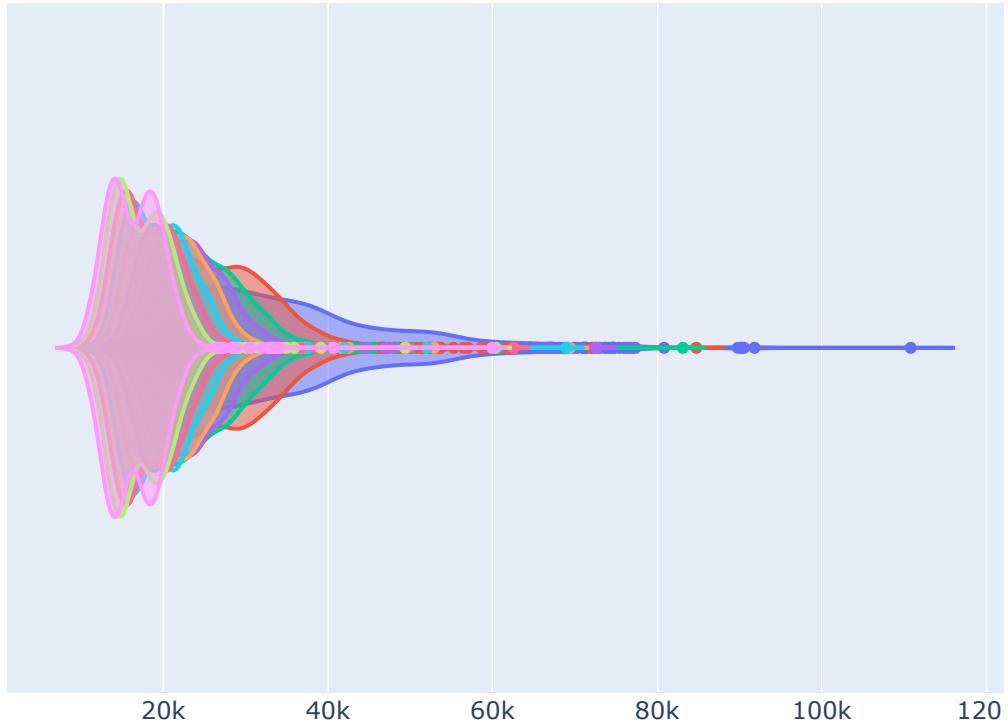
Chile

Количество прослушиваний песен из топ-10 в Chile



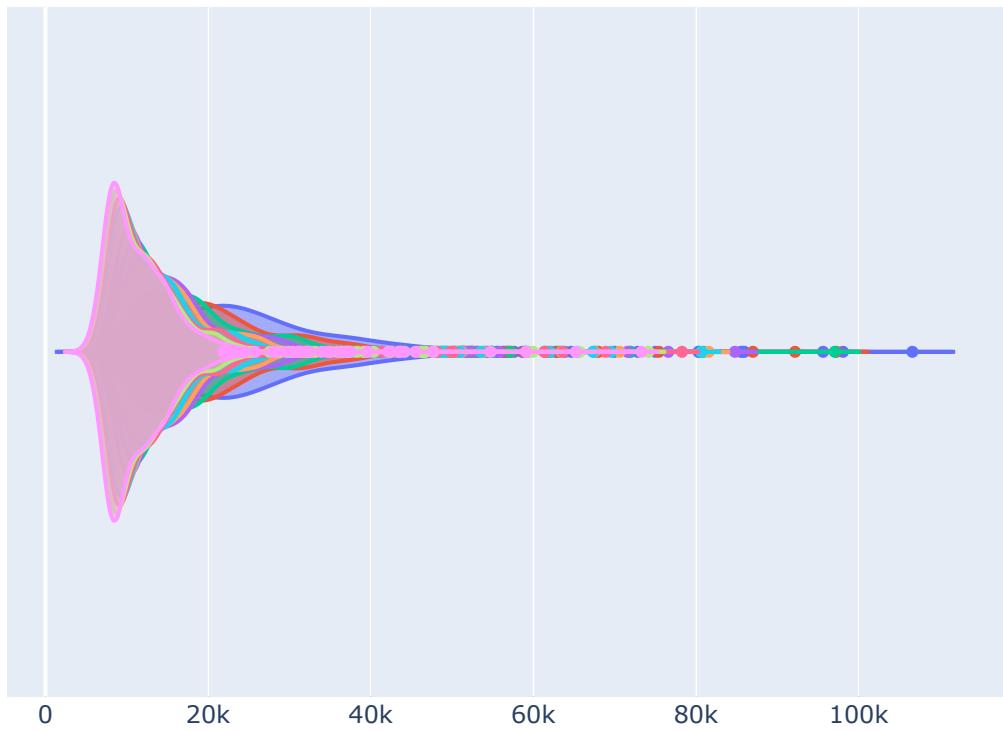
Costa Rica

Количество прослушиваний песен из топ-10 в Costa Rica



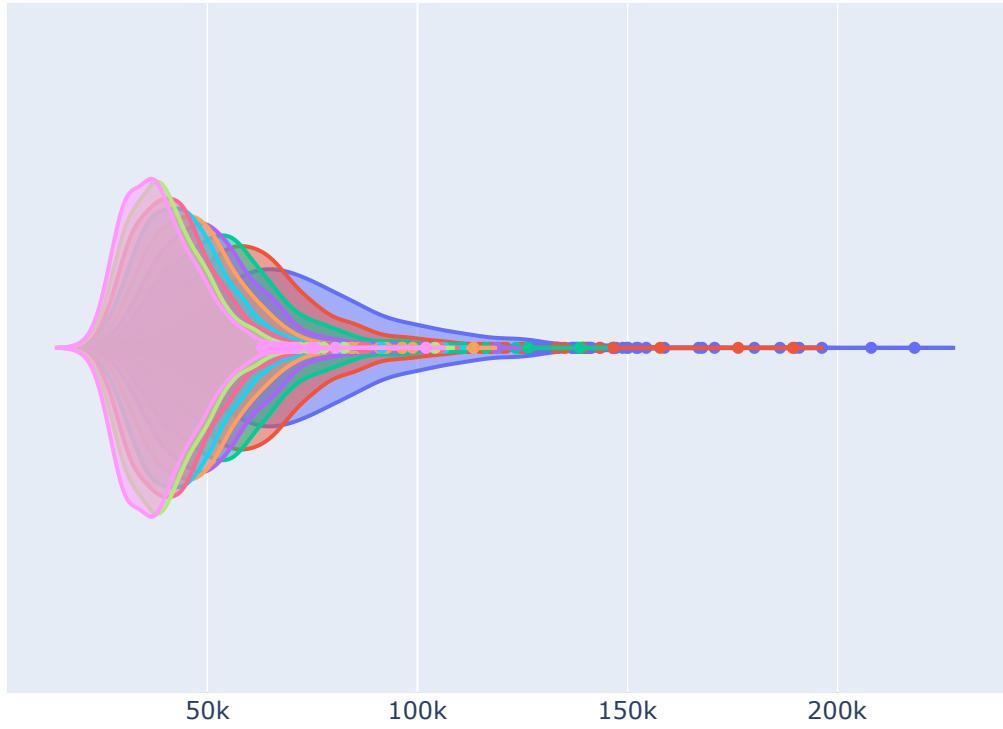
Czech Republic

Количество прослушиваний песен из топ-10 в Czech Republic



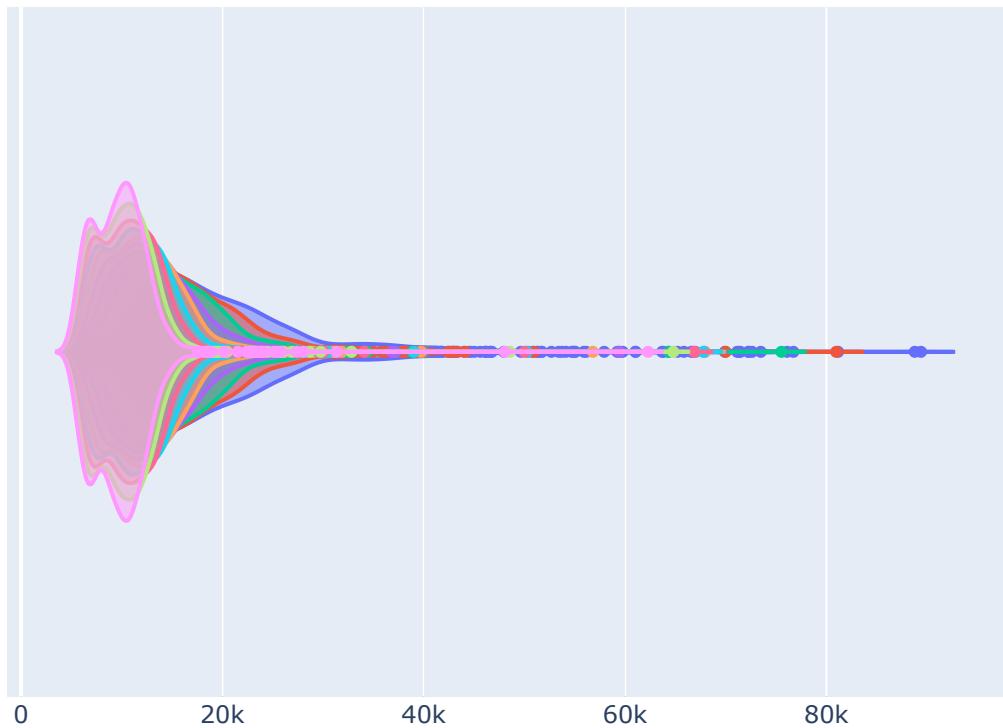
Finland

Количество прослушиваний песен из топ-10 в Finland



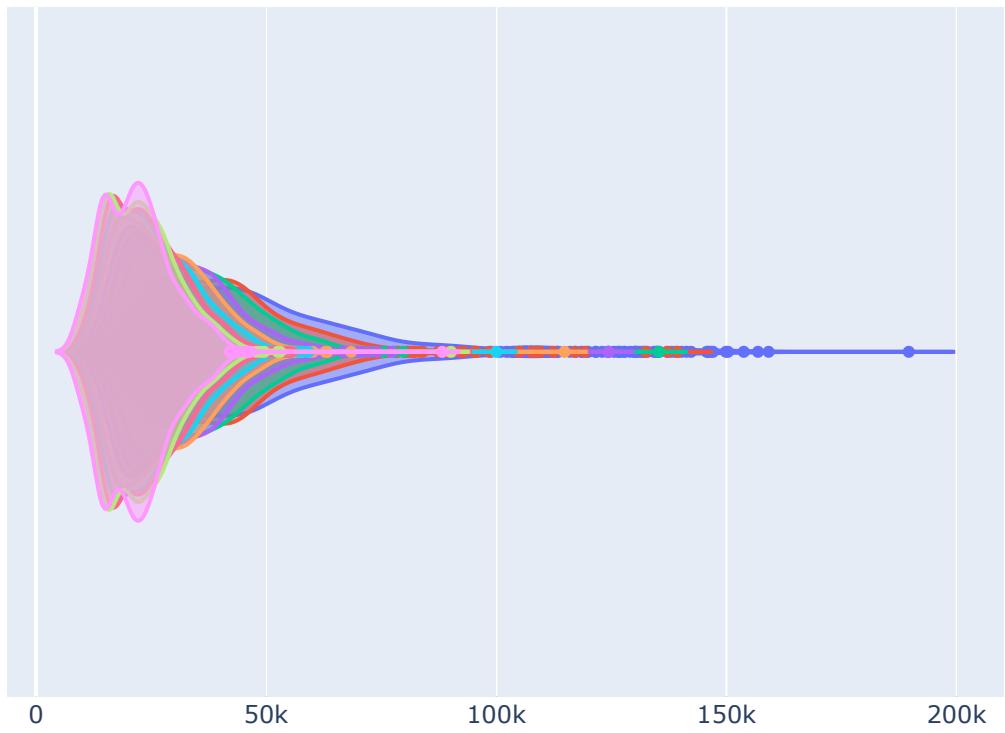
Dominican Republic

Количество прослушиваний песен из топ-10 в Dominican Rep



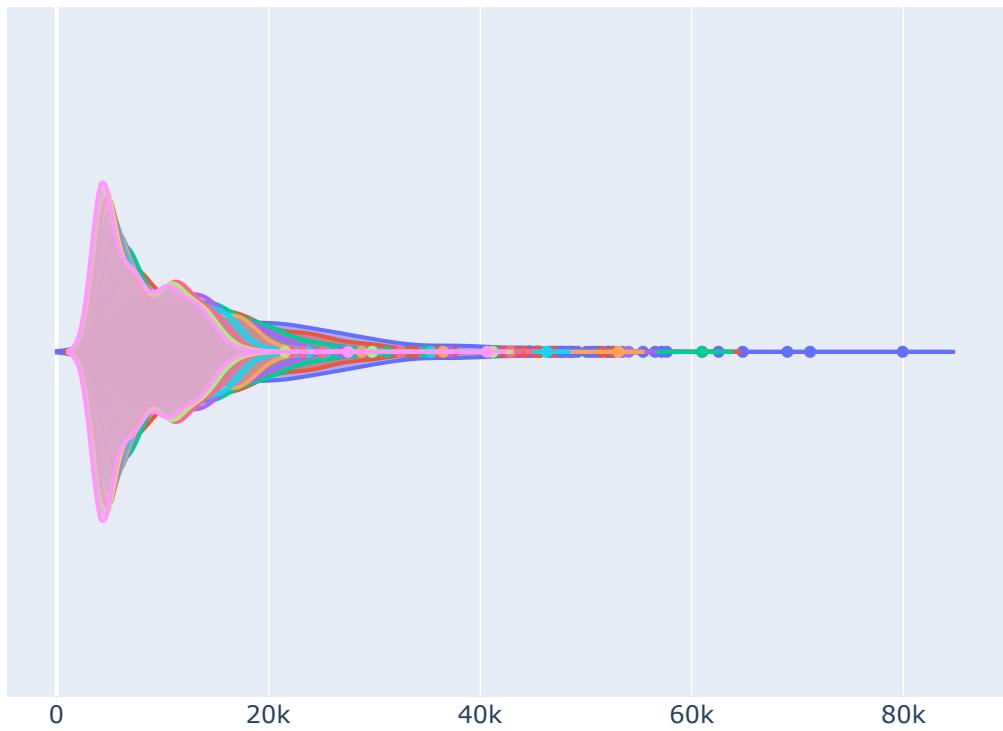
Ecuador

Количество прослушиваний песен из топ-10 в Ecuador



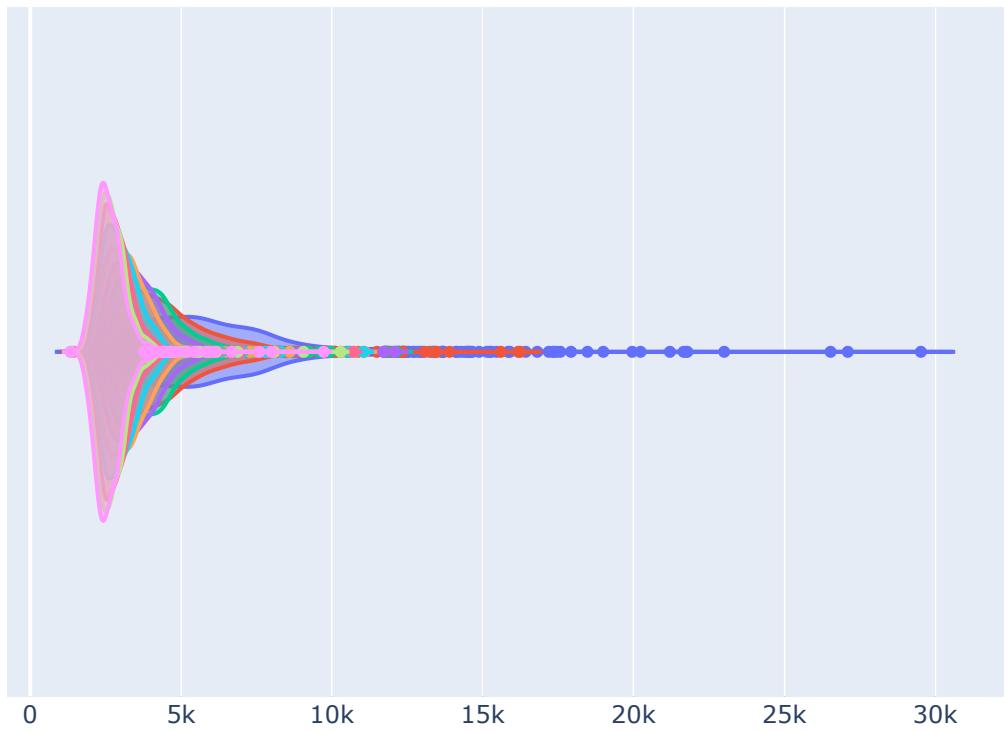
El Salvador

Количество прослушиваний песен из топ-10 в El Salvador



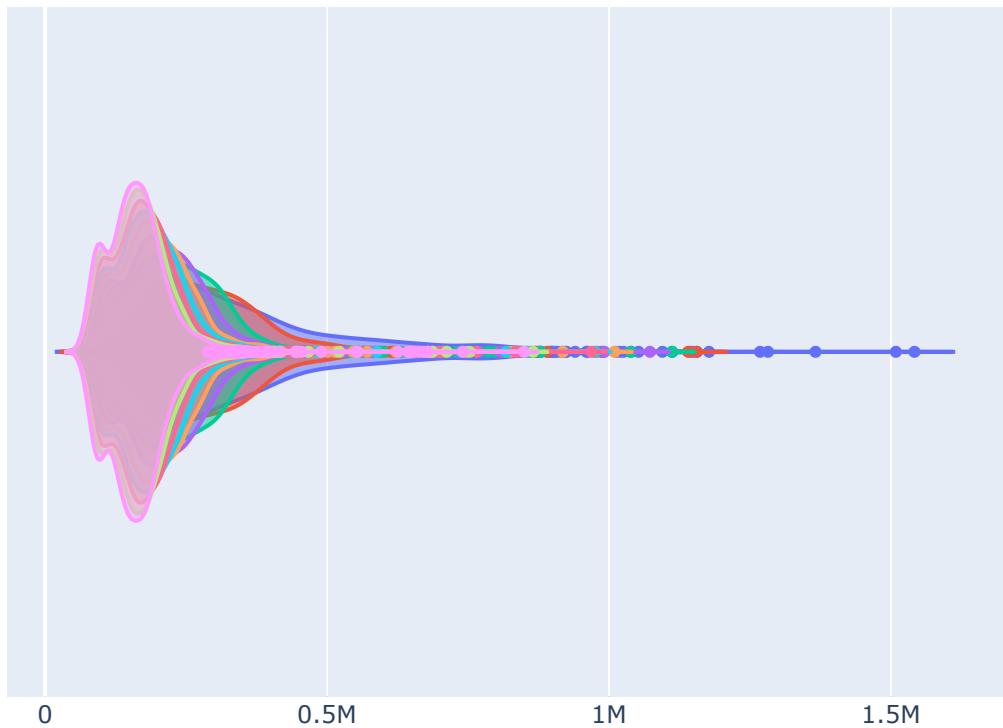
Estonia

Количество прослушиваний песен из топ-10 в Estonia



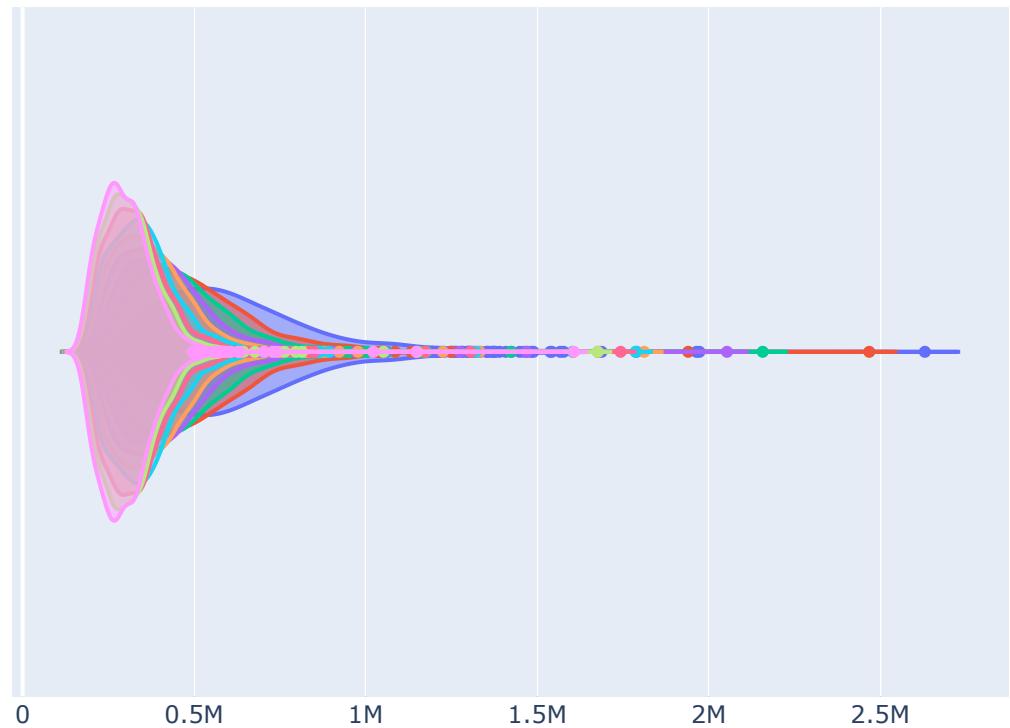
France

Количество прослушиваний песен из топ-10 в France



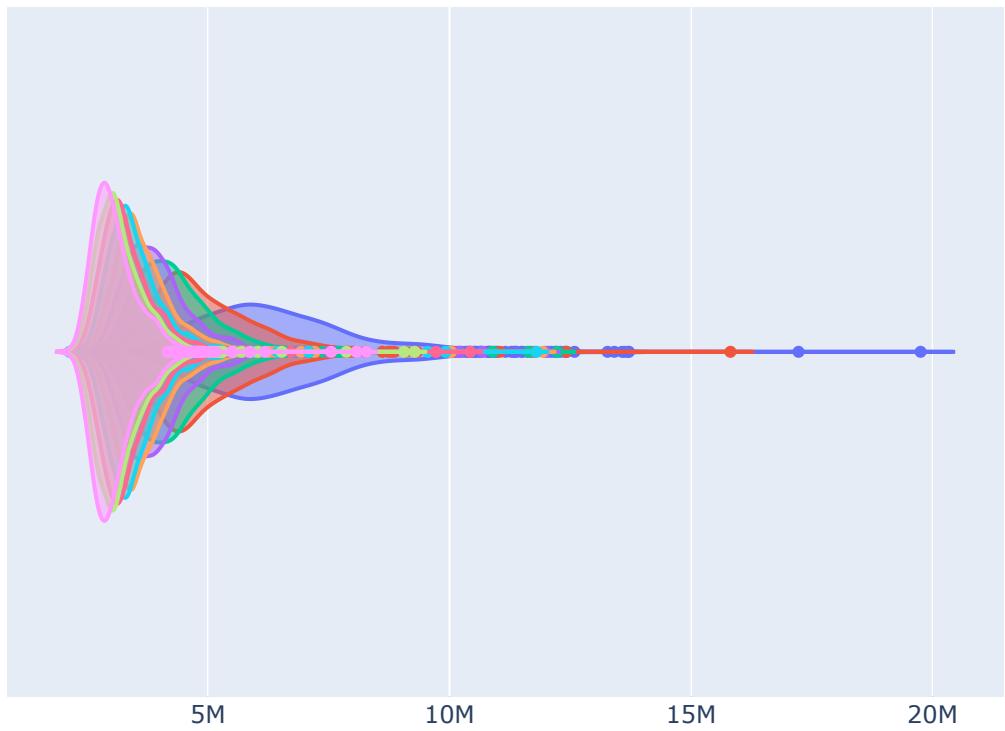
Germany

Количество прослушиваний песен из топ-10 в Germany



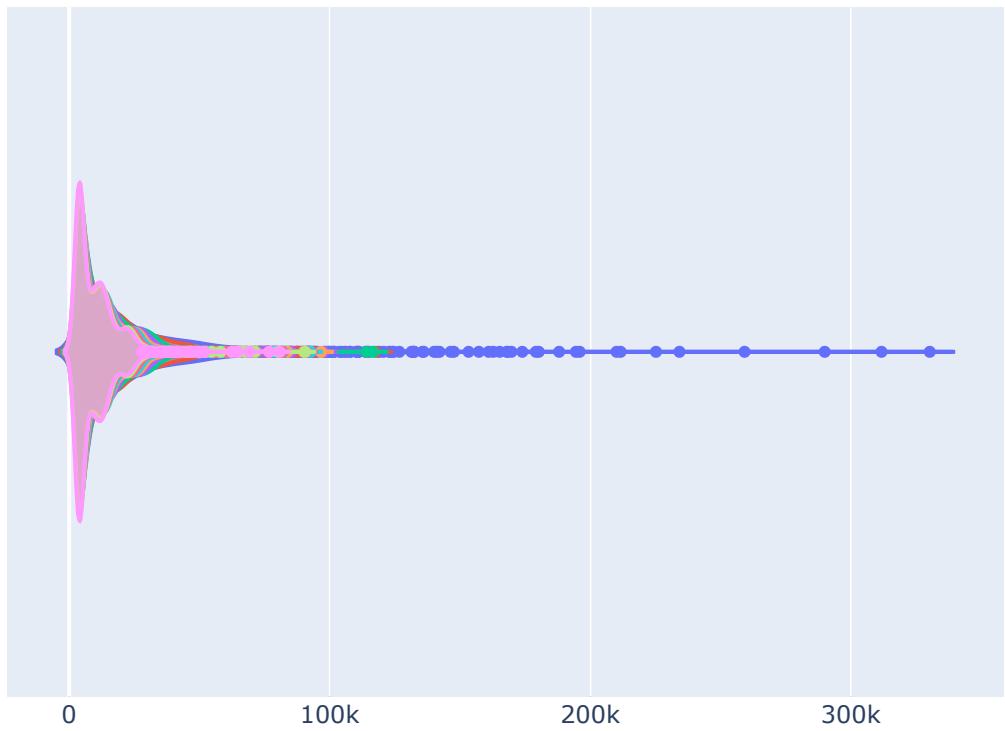
Global

Количество прослушиваний песен из топ-10 в Global



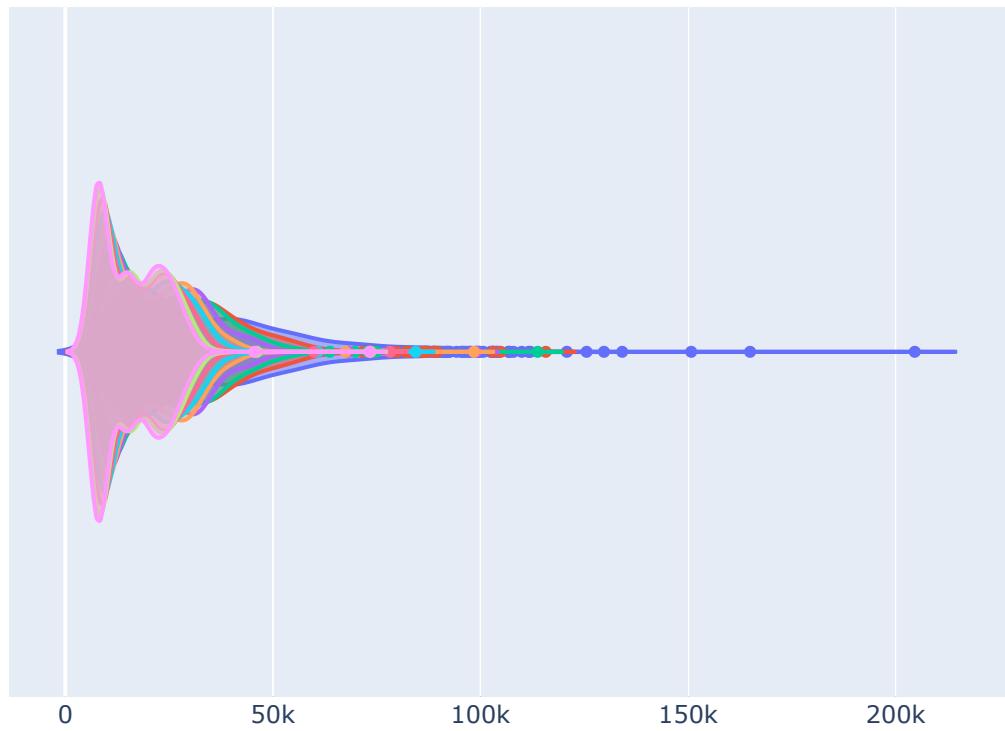
Greece

Количество прослушиваний песен из топ-10 в Greece



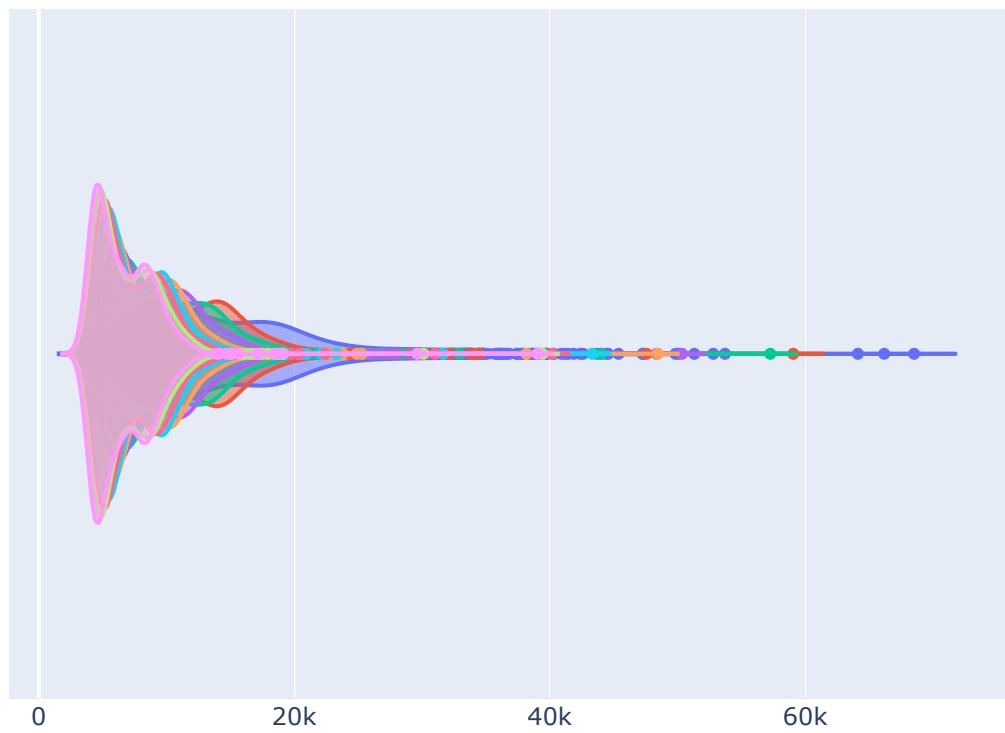
Guatemala

Количество прослушиваний песен из топ-10 в Guatemala



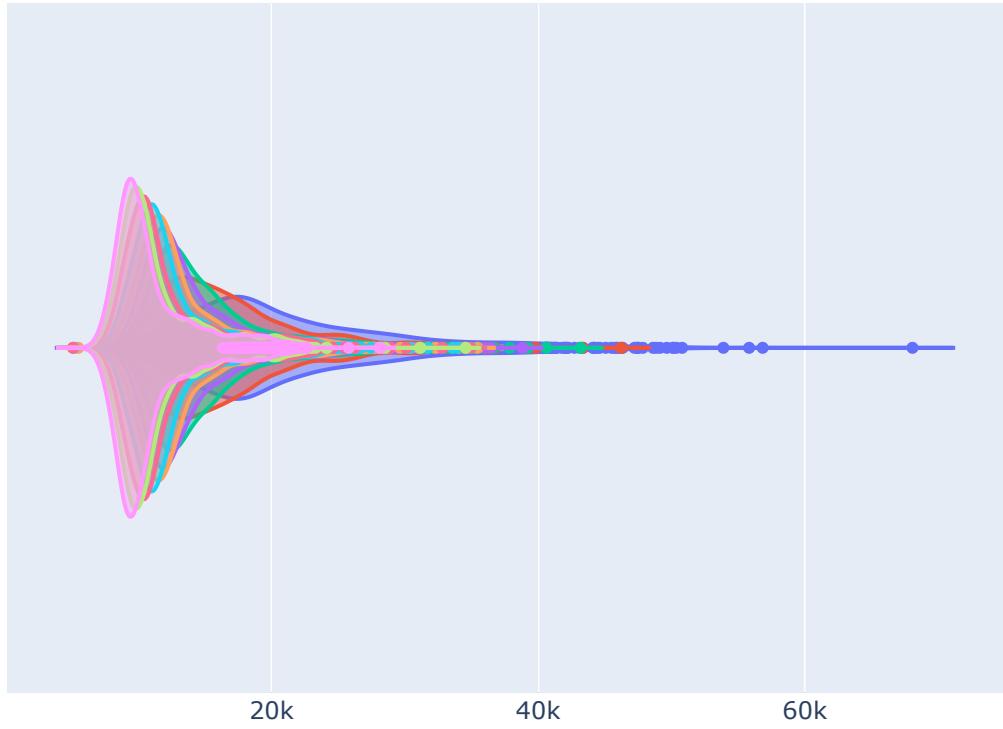
Honduras

Количество прослушиваний песен из топ-10 в Honduras



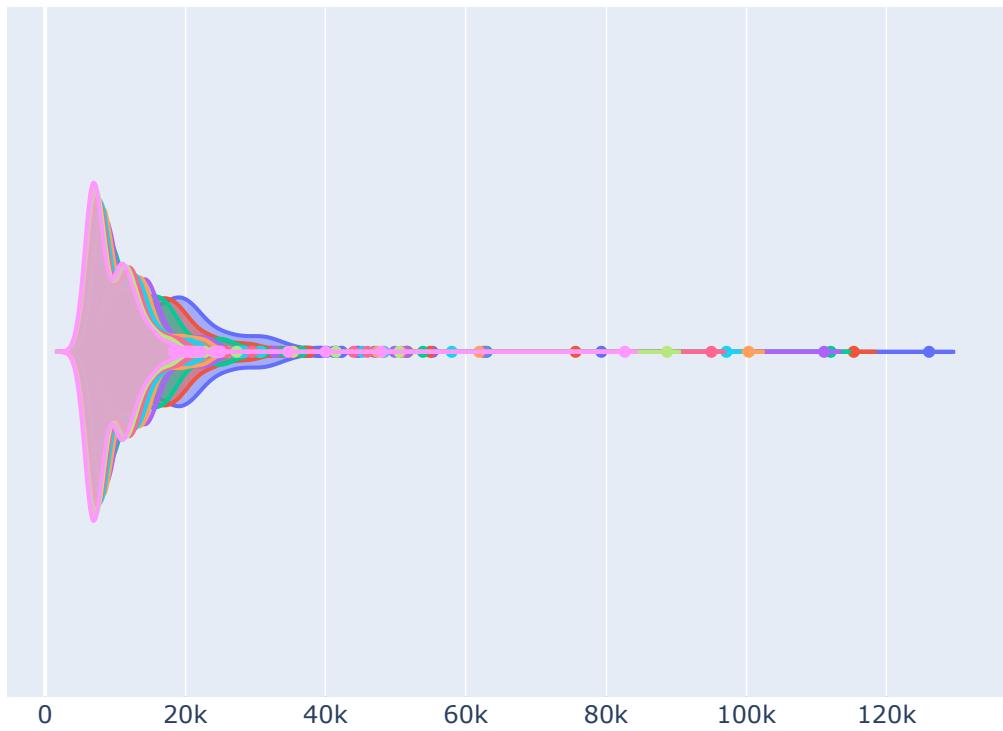
Hong Kong

Количество прослушиваний песен из топ-10 в Hong Kong



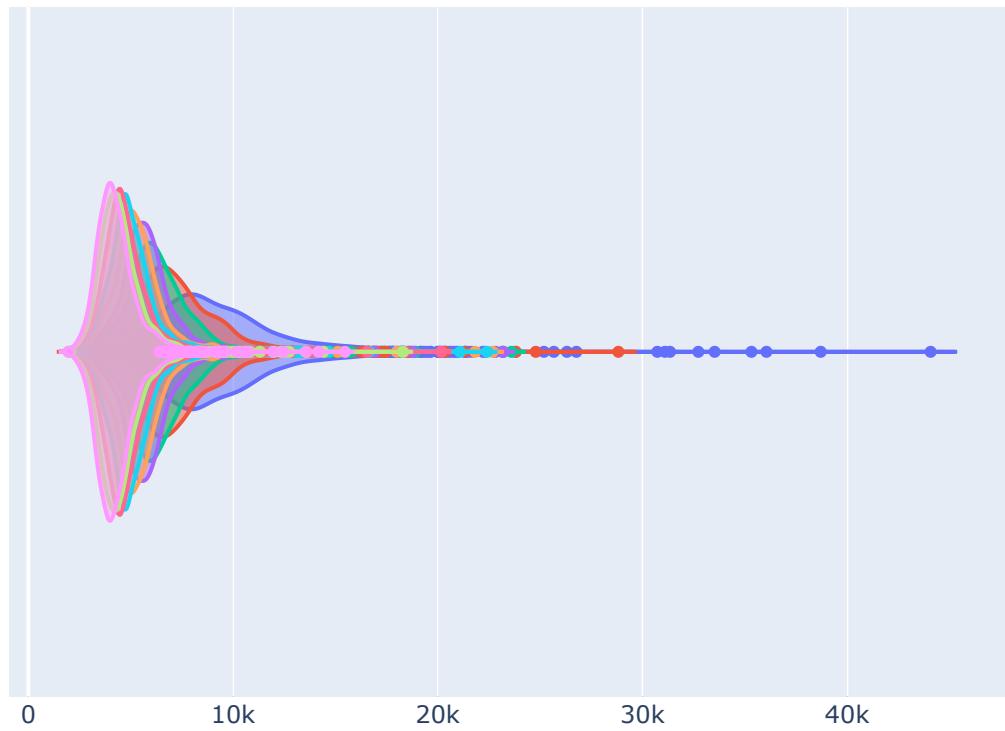
Hungary

Количество прослушиваний песен из топ-10 в Hungary



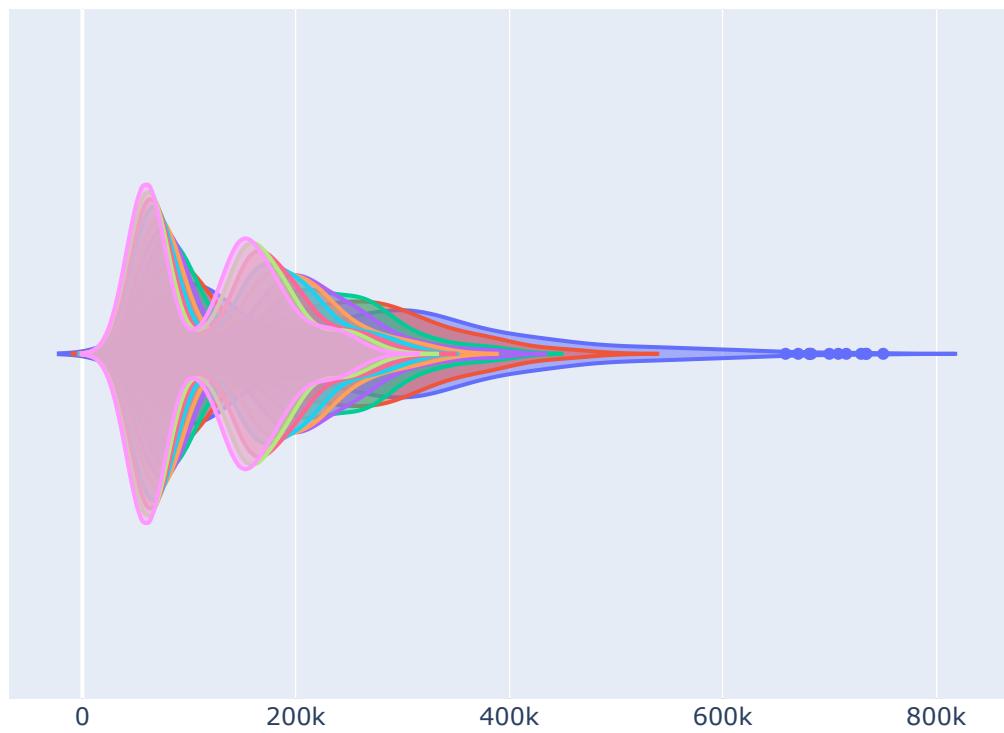
Iceland

Количество прослушиваний песен из топ-10 в Iceland



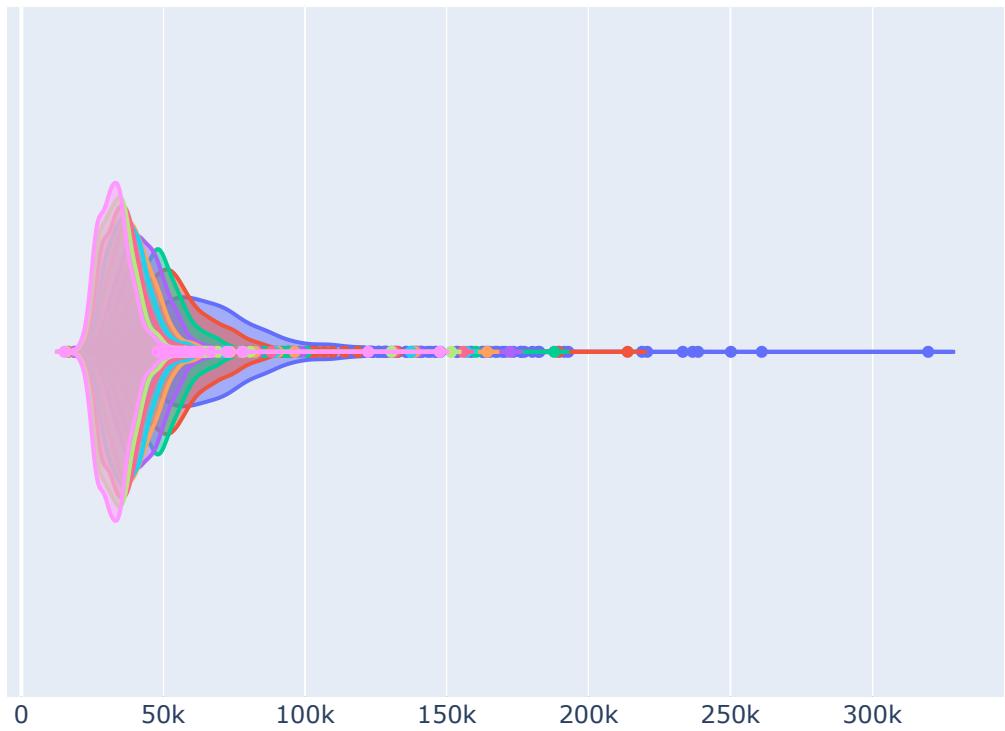
Indonesia

Количество прослушиваний песен из топ-10 в Indonesia



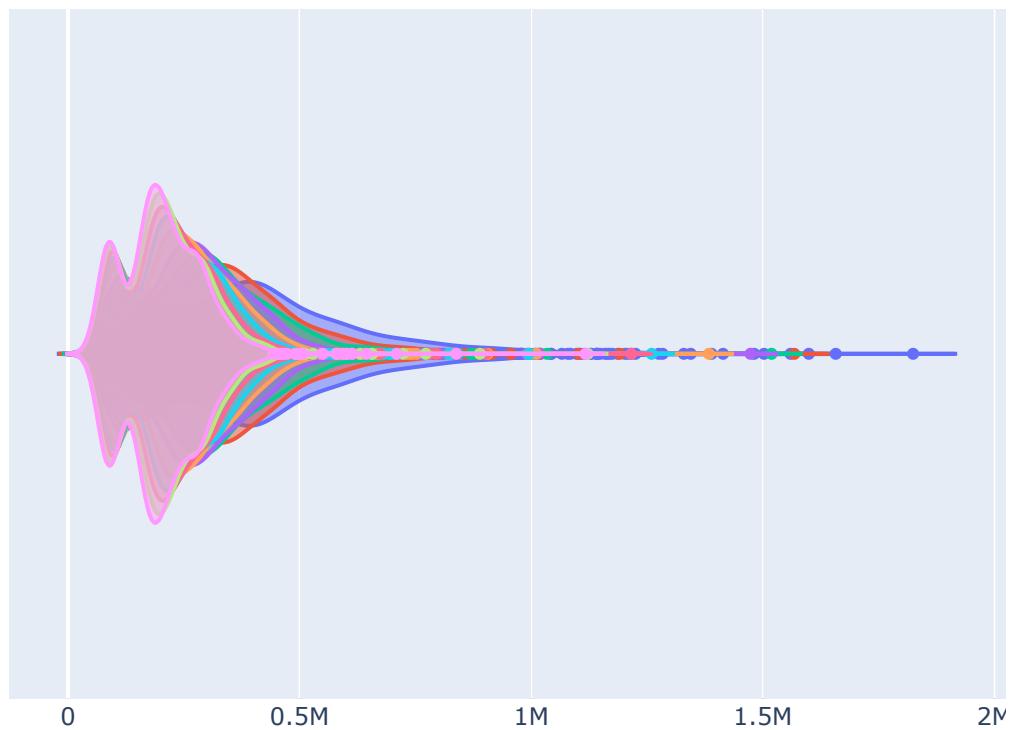
Ireland

Количество прослушиваний песен из топ-10 в Ireland



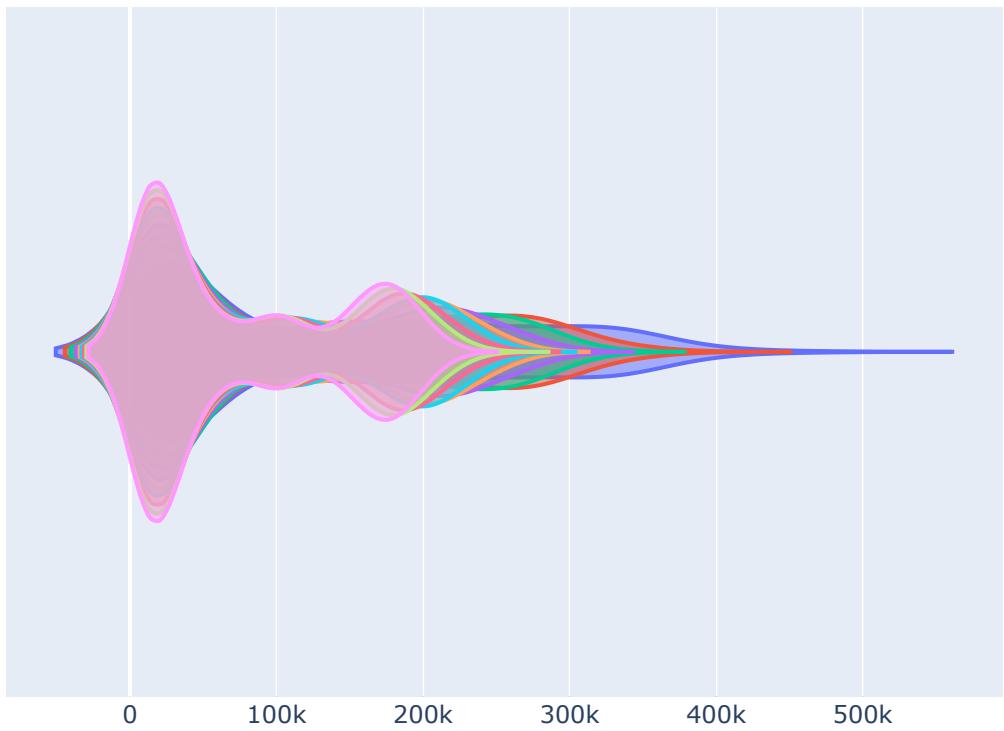
Italy

Количество прослушиваний песен из топ-10 в Italy



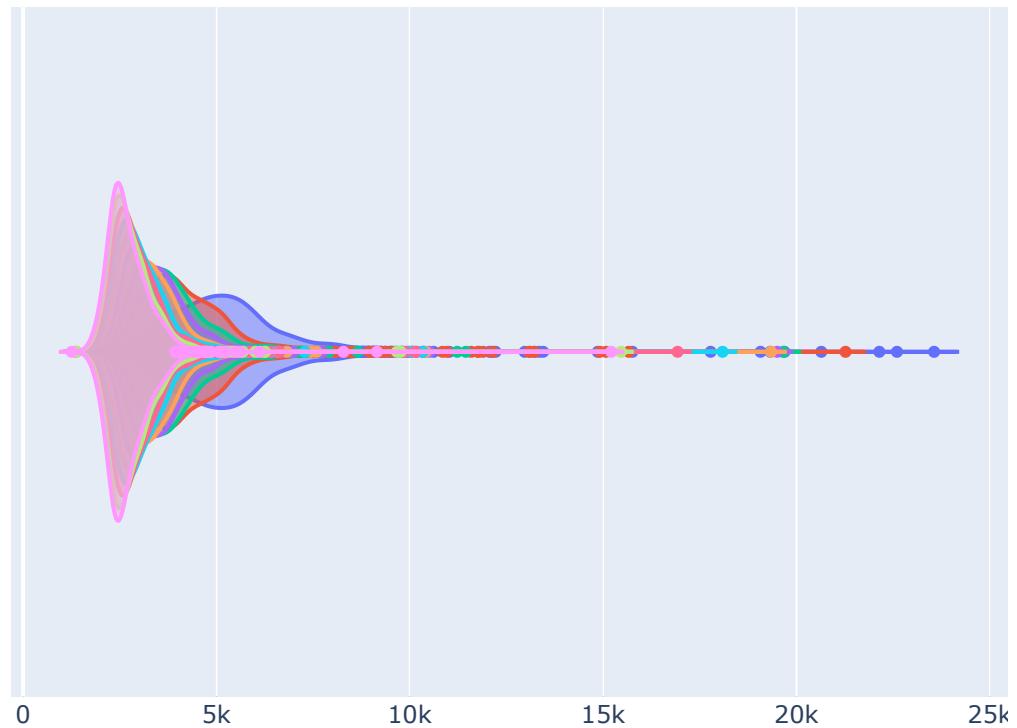
Japan

Количество прослушиваний песен из топ-10 в Japan



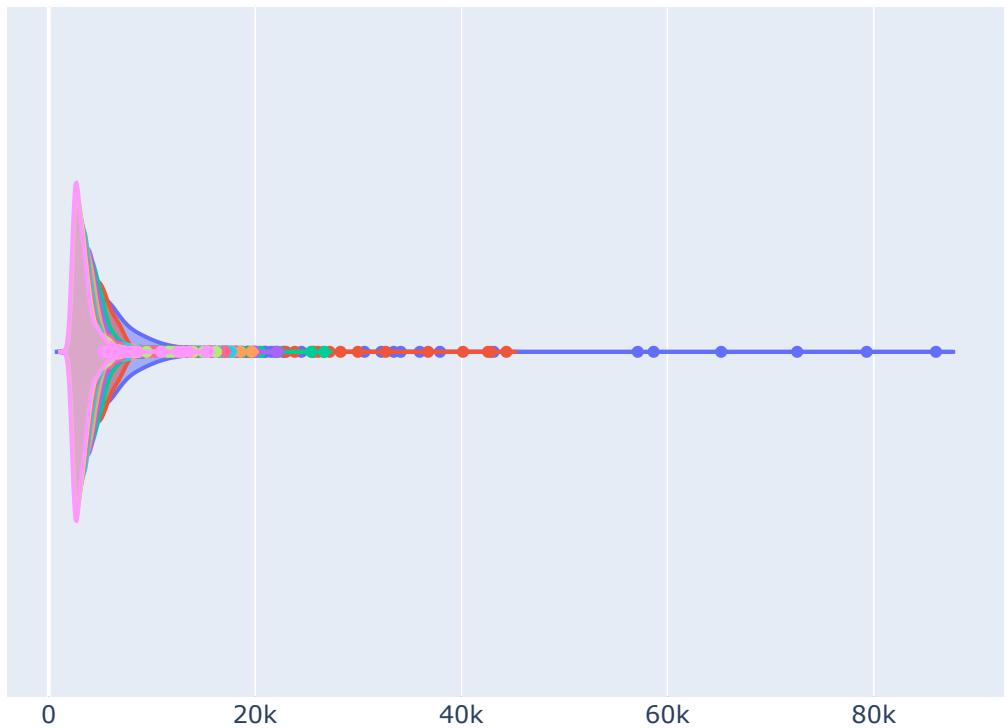
Latvia

Количество прослушиваний песен из топ-10 в Latvia



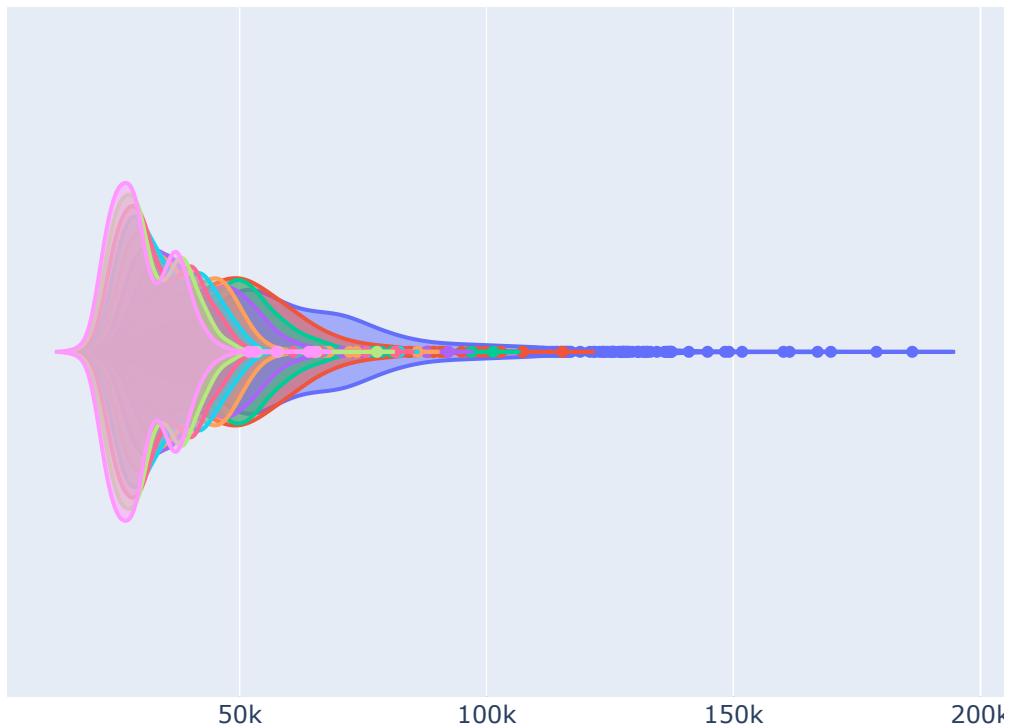
Lithuania

Количество прослушиваний песен из топ-10 в Lithuania



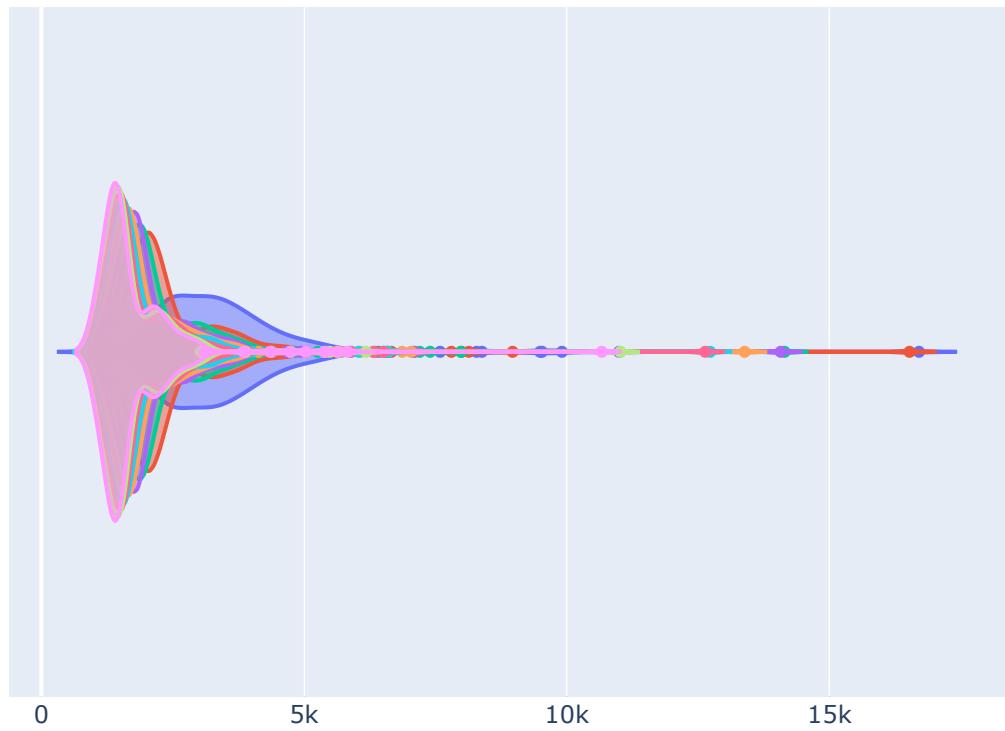
Malaysia

Количество прослушиваний песен из топ-10 в Malaysia



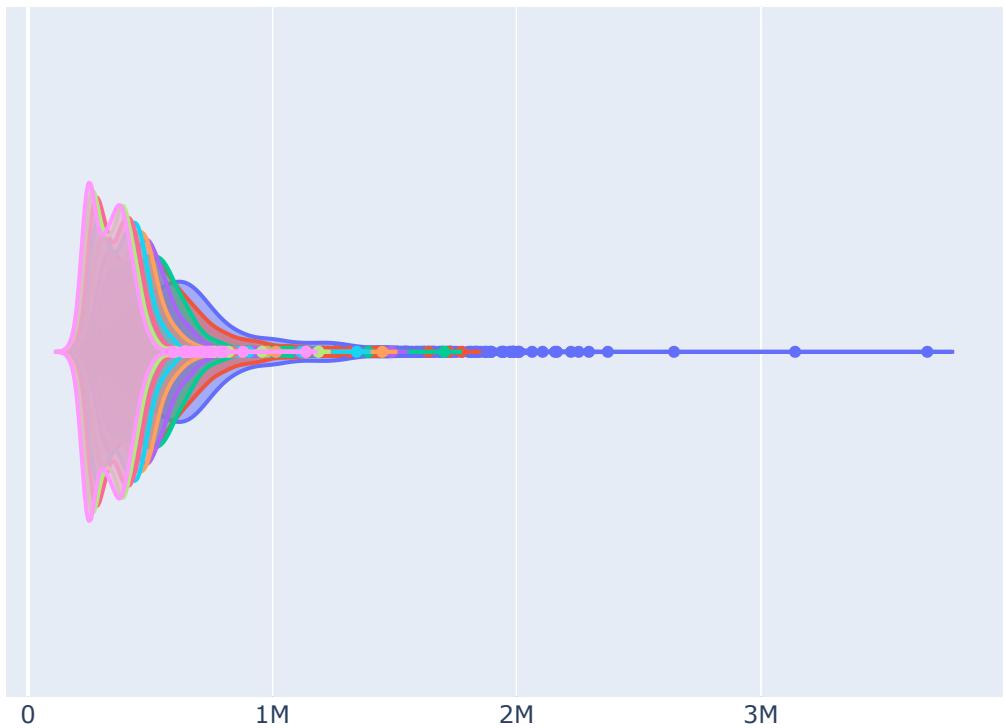
Luxembourg

Количество прослушиваний песен из топ-10 в Luxembourg



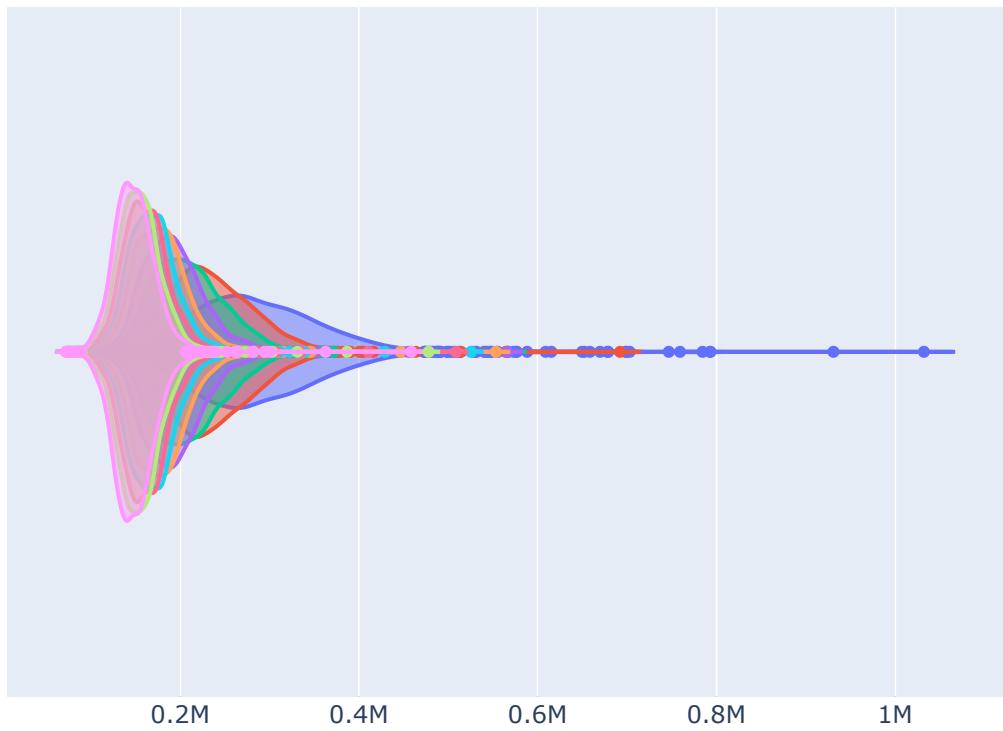
Mexico

Количество прослушиваний песен из топ-10 в Mexico



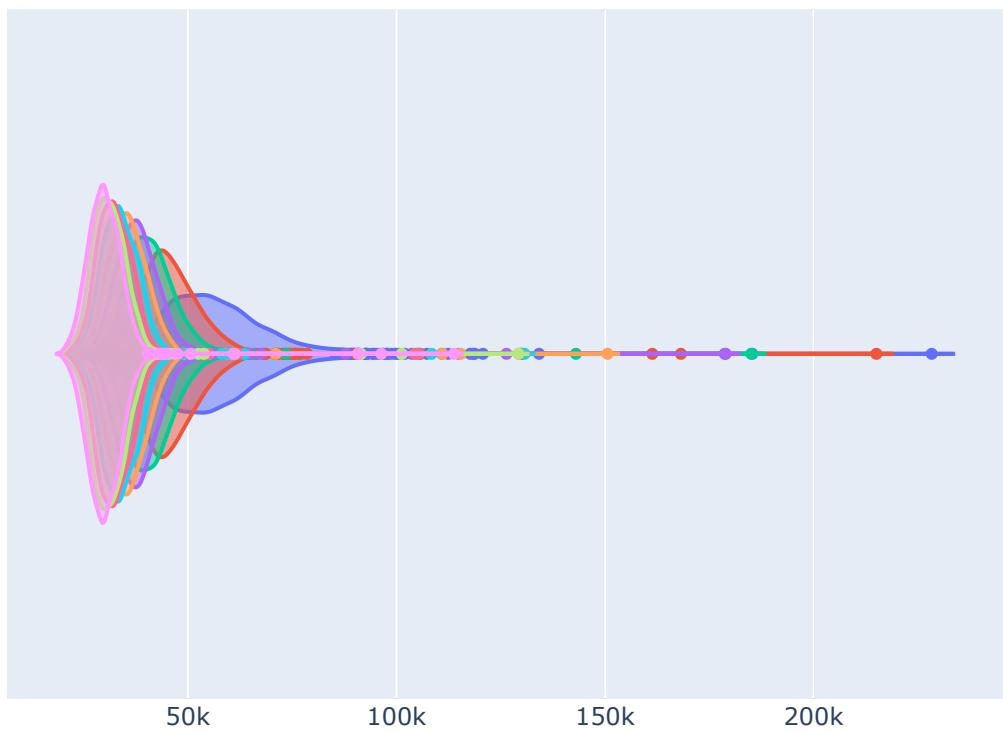
Netherlands

Количество прослушиваний песен из топ-10 в Netherlands



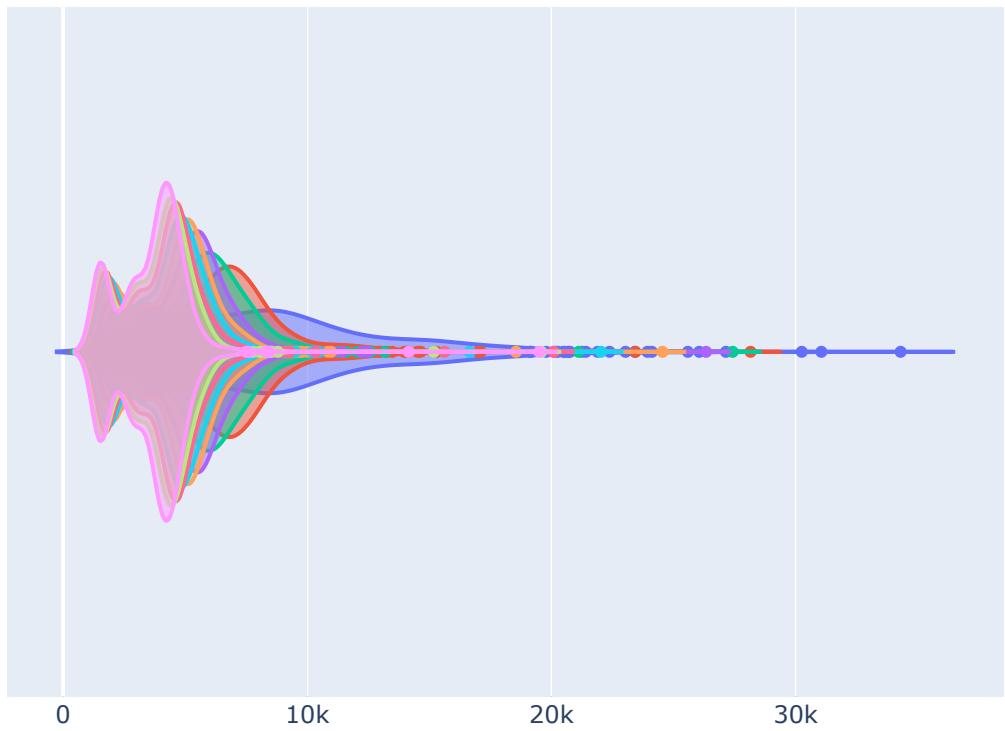
New Zealand

Количество прослушиваний песен из топ-10 в New Zealand



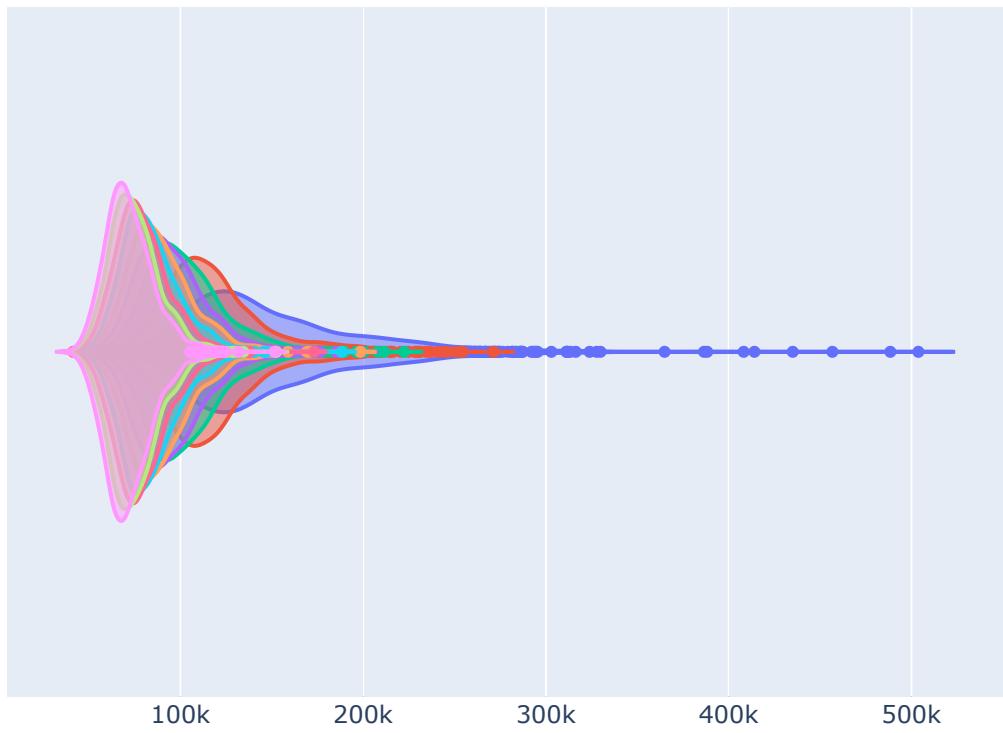
Nicaragua

Количество прослушиваний песен из топ-10 в Nicaragua



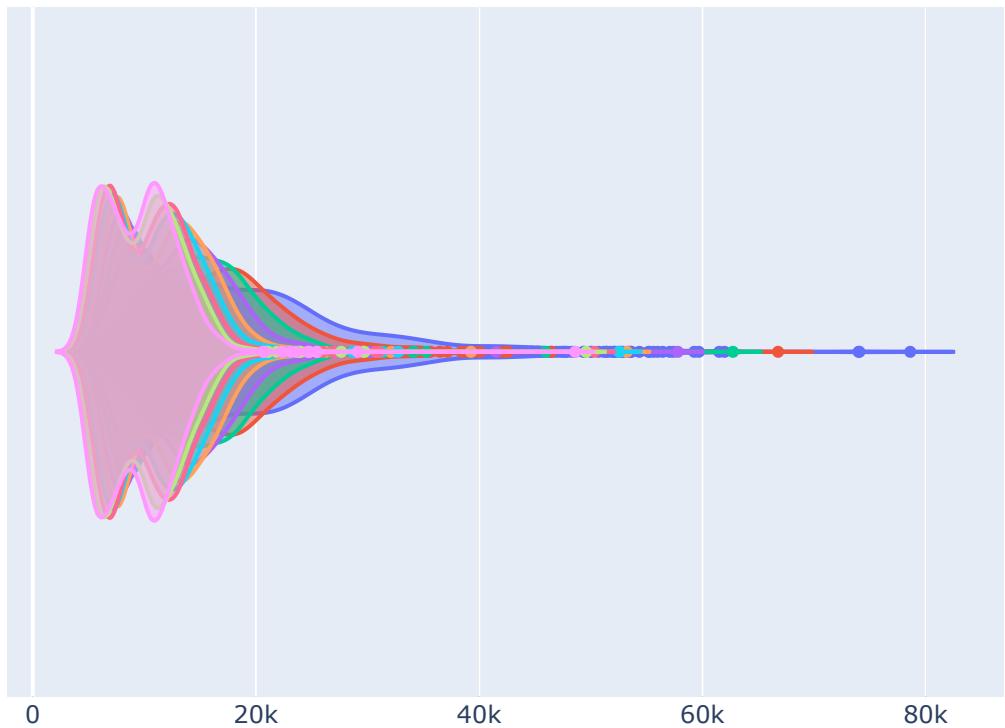
Norway

Количество прослушиваний песен из топ-10 в Norway



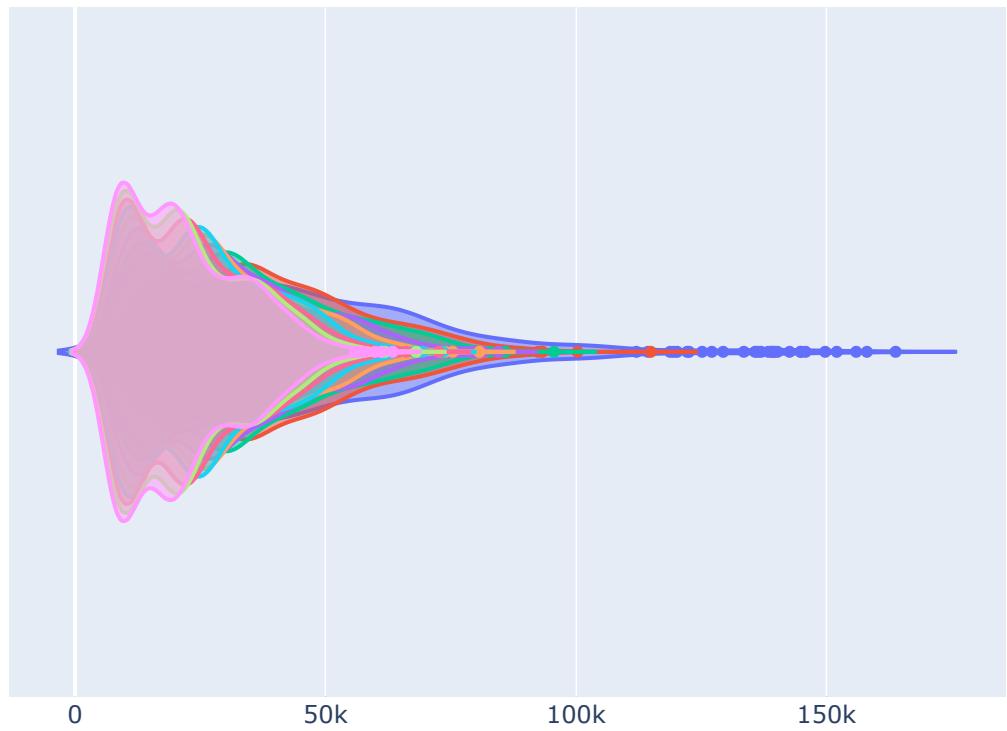
Panama

Количество прослушиваний песен из топ-10 в Panama



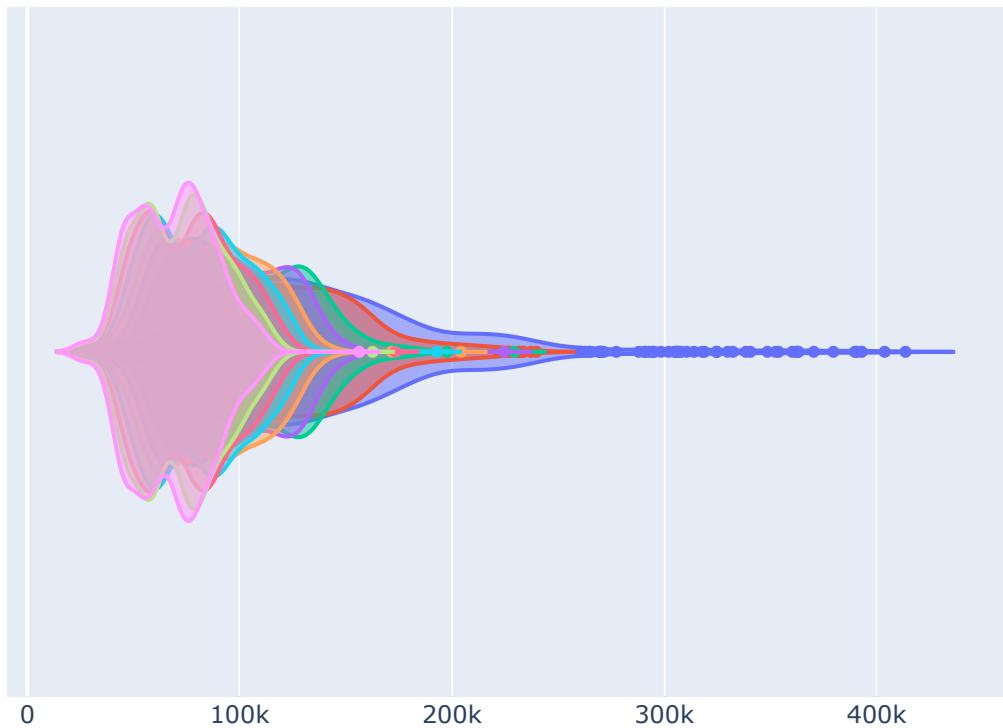
Paraguay

Количество прослушиваний песен из топ-10 в Paraguay



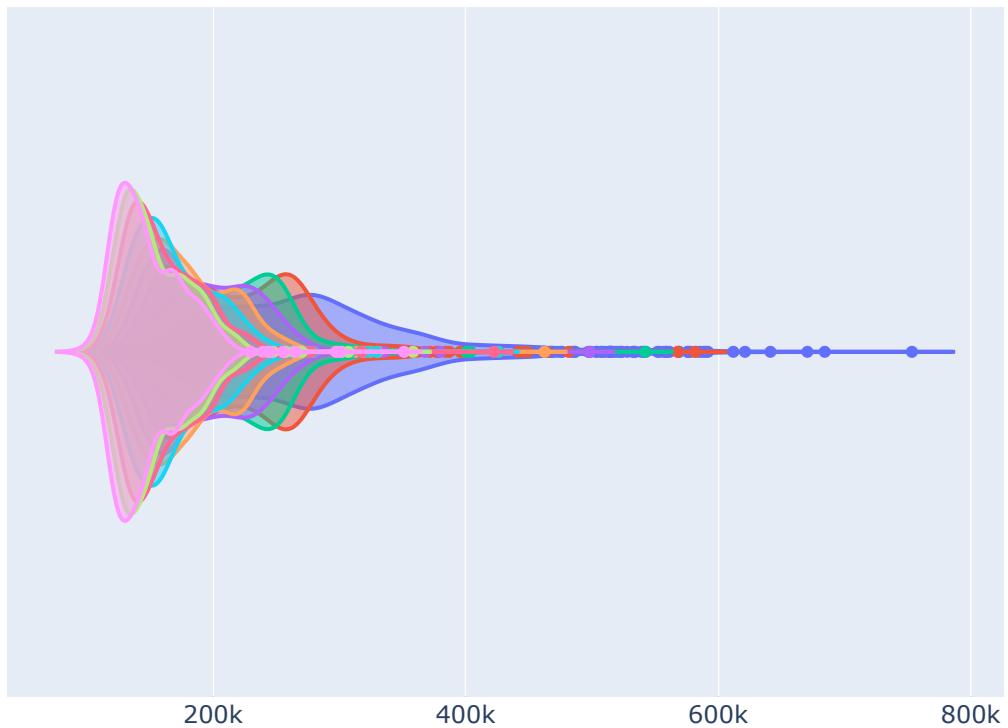
Peru

Количество прослушиваний песен из топ-10 в Peru



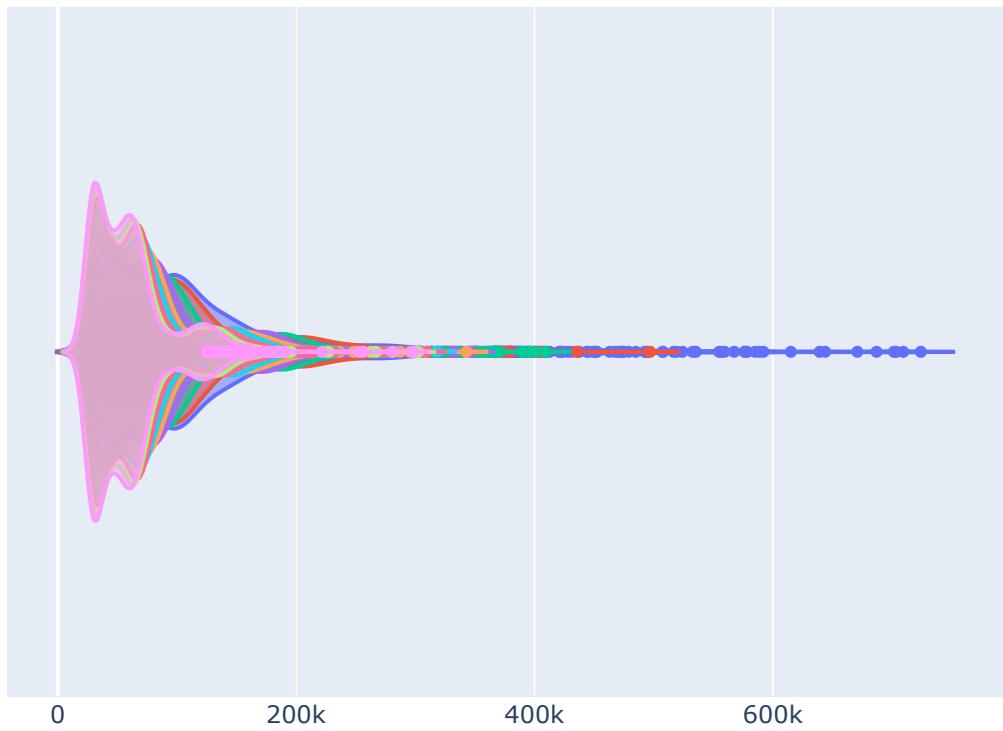
Philippines

Количество прослушиваний песен из топ-10 в Philippines



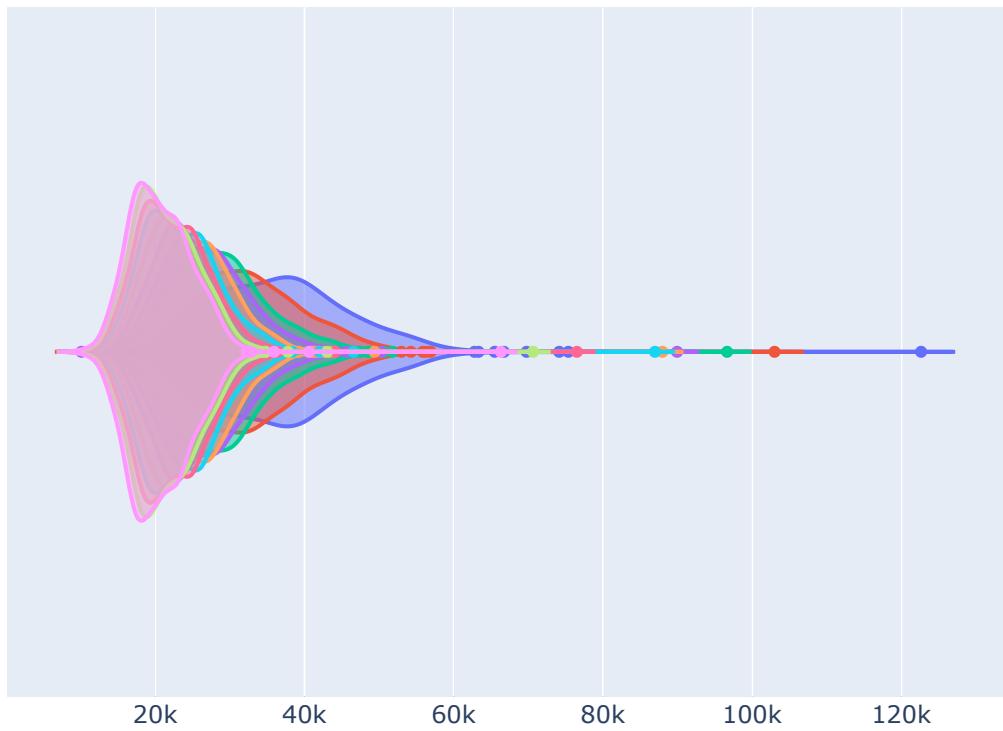
Poland

Количество прослушиваний песен из топ-10 в Poland



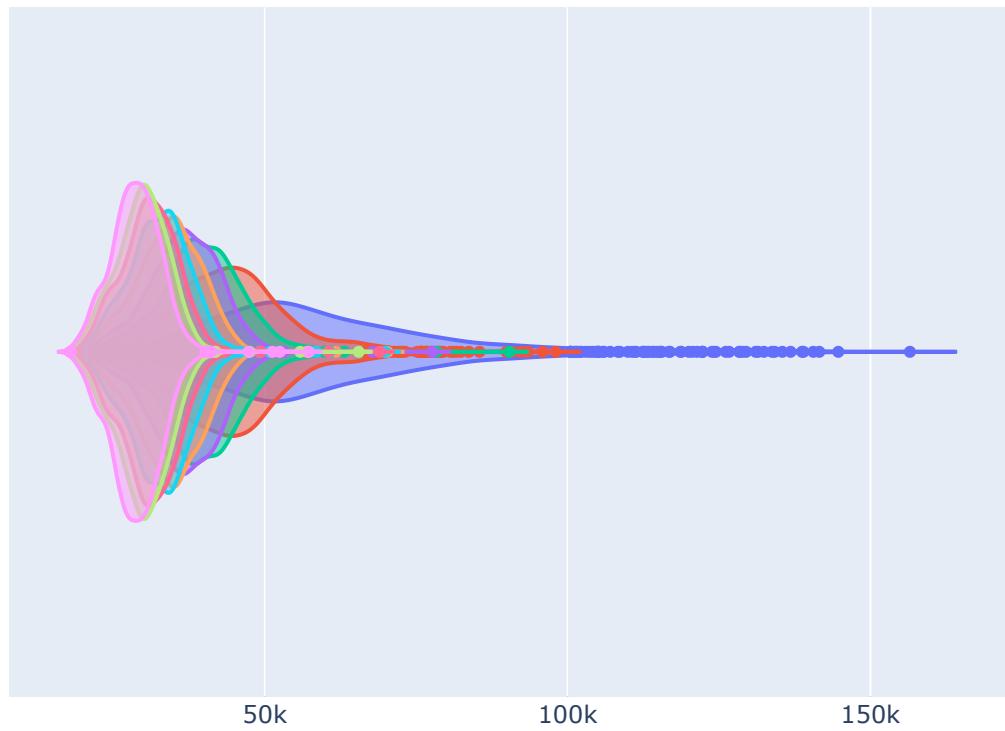
Portugal

Количество прослушиваний песен из топ-10 в Portugal



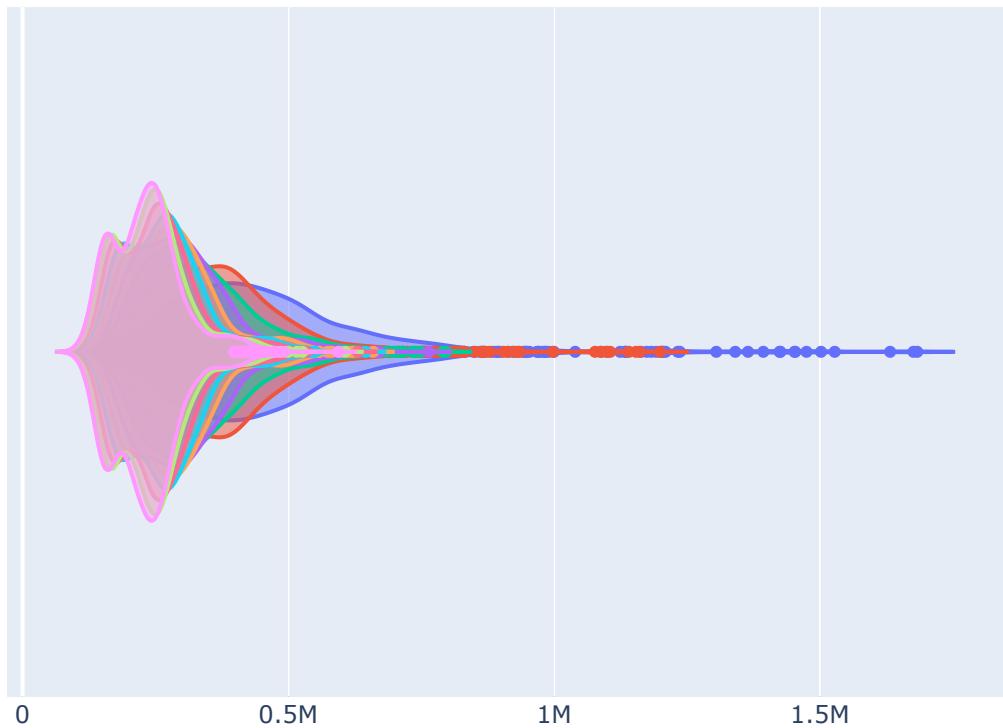
Singapore

Количество прослушиваний песен из топ-10 в Singapore



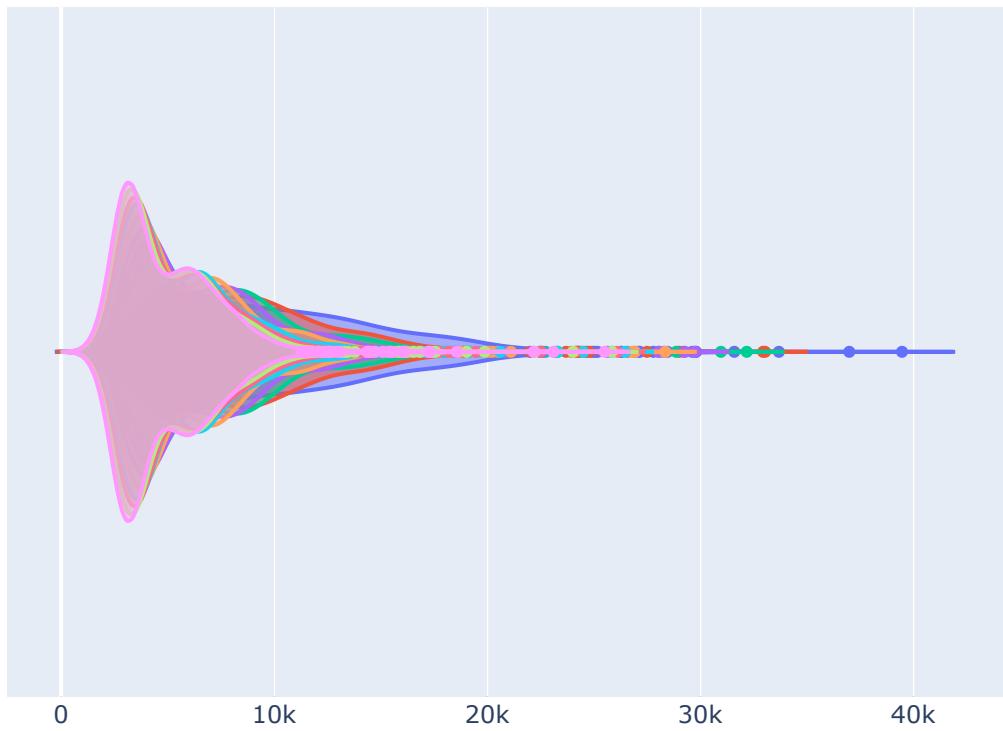
Spain

Количество прослушиваний песен из топ-10 в Spain



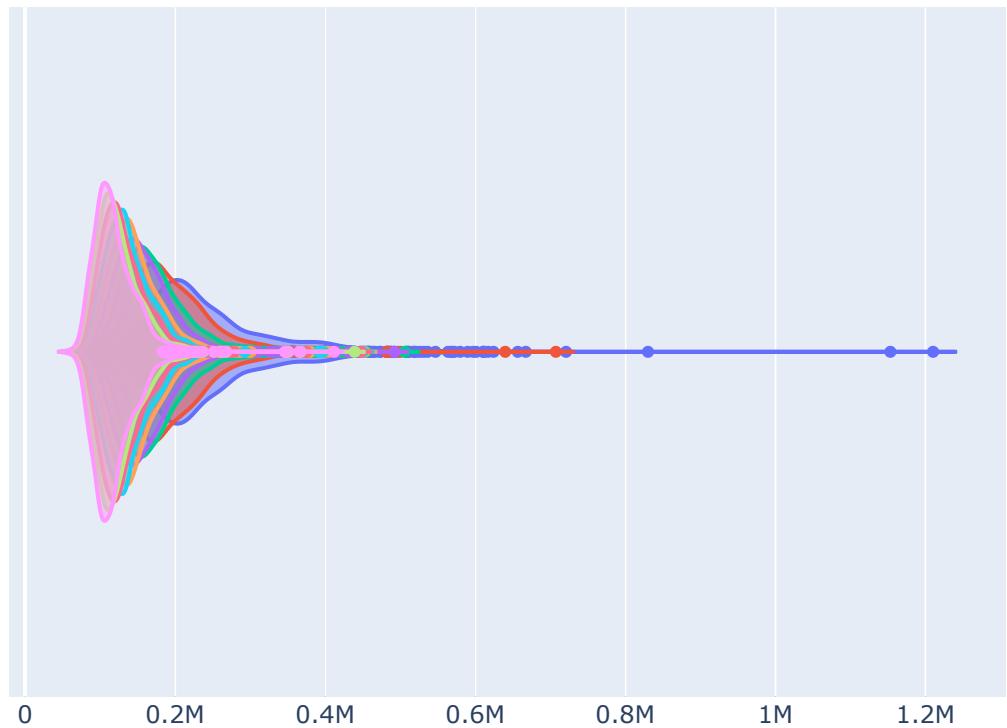
Slovakia

Количество прослушиваний песен из топ-10 в Slovakia



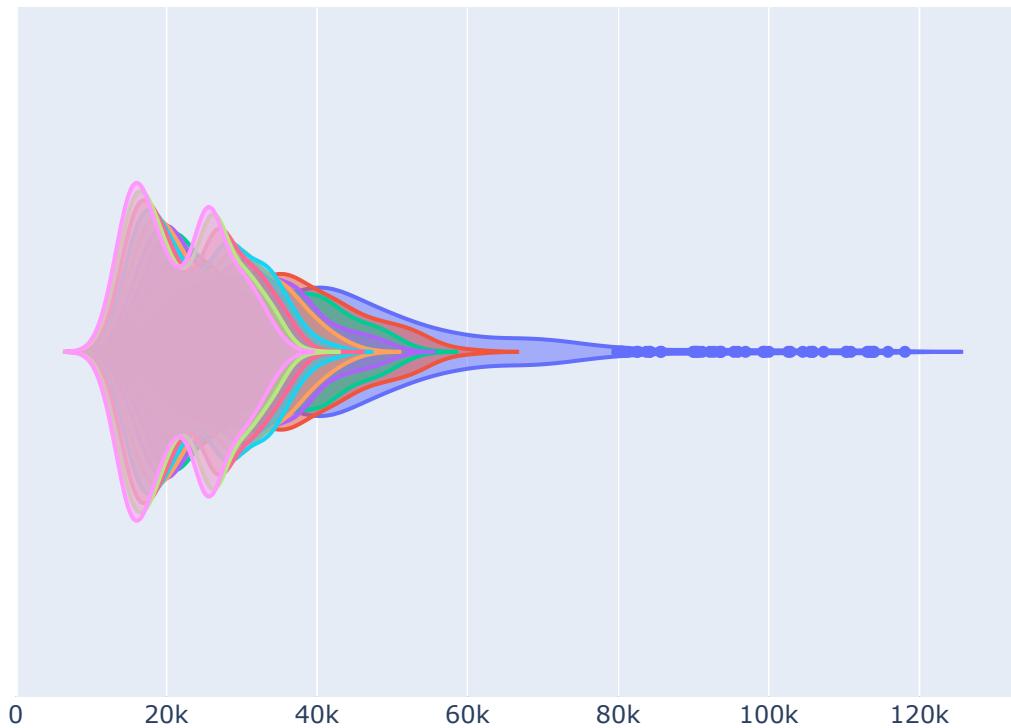
Sweden

Количество прослушиваний песен из топ-10 в Sweden



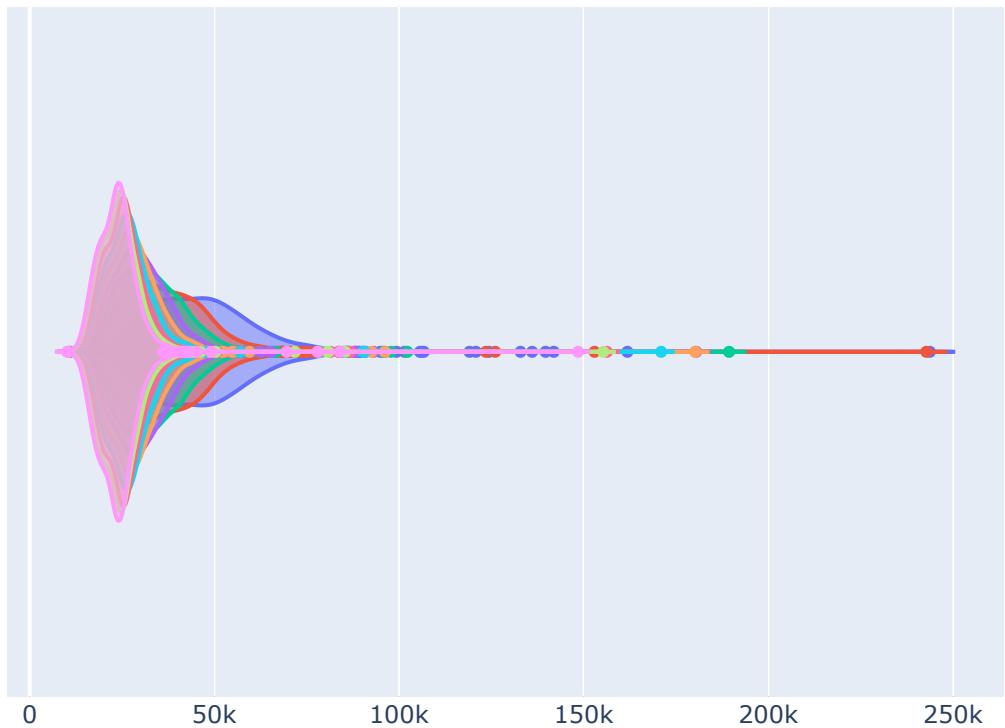
Taiwan

Количество прослушиваний песен из топ-10 в Taiwan



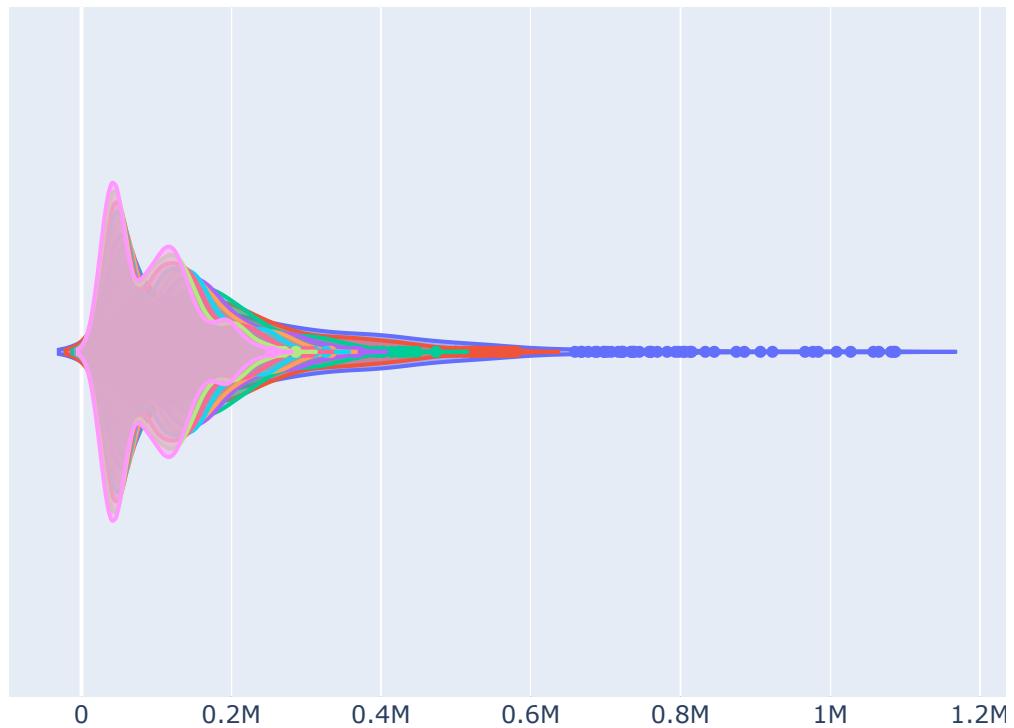
Switzerland

Количество прослушиваний песен из топ-10 в Switzerland



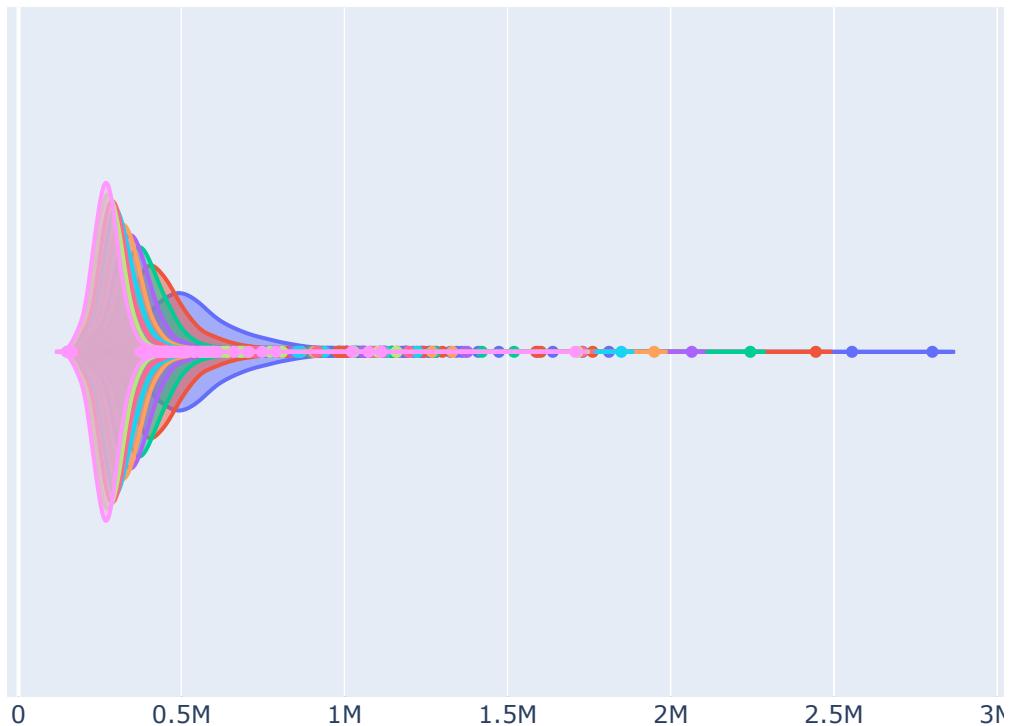
Turkey

Количество прослушиваний песен из топ-10 в Turkey



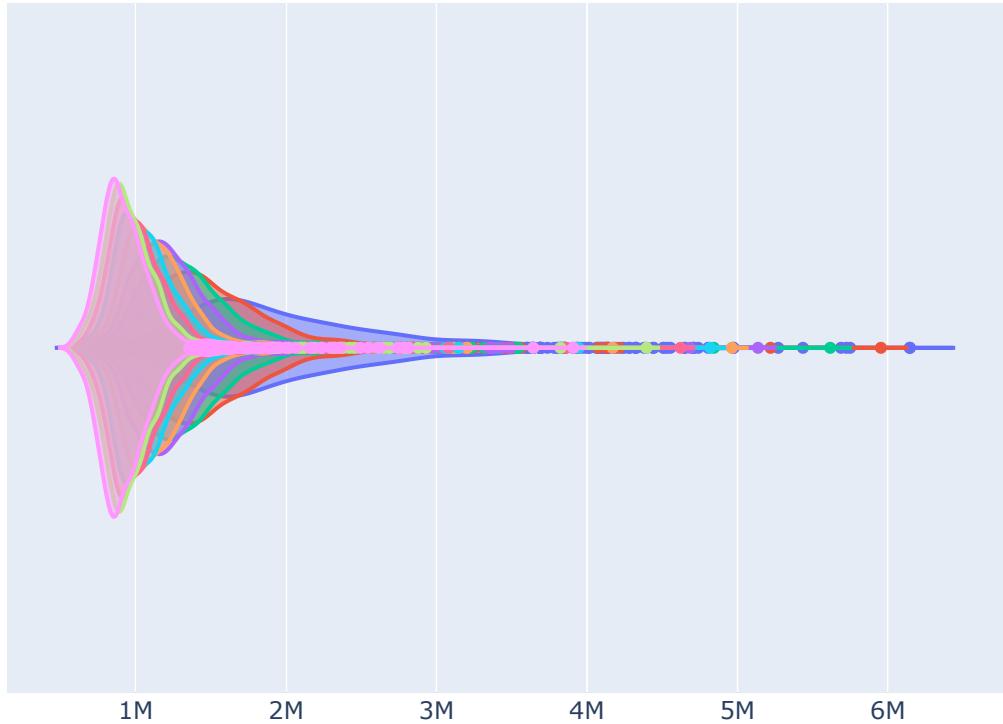
United Kingdom

Количество прослушиваний песен из топ-10 в United Kingdom



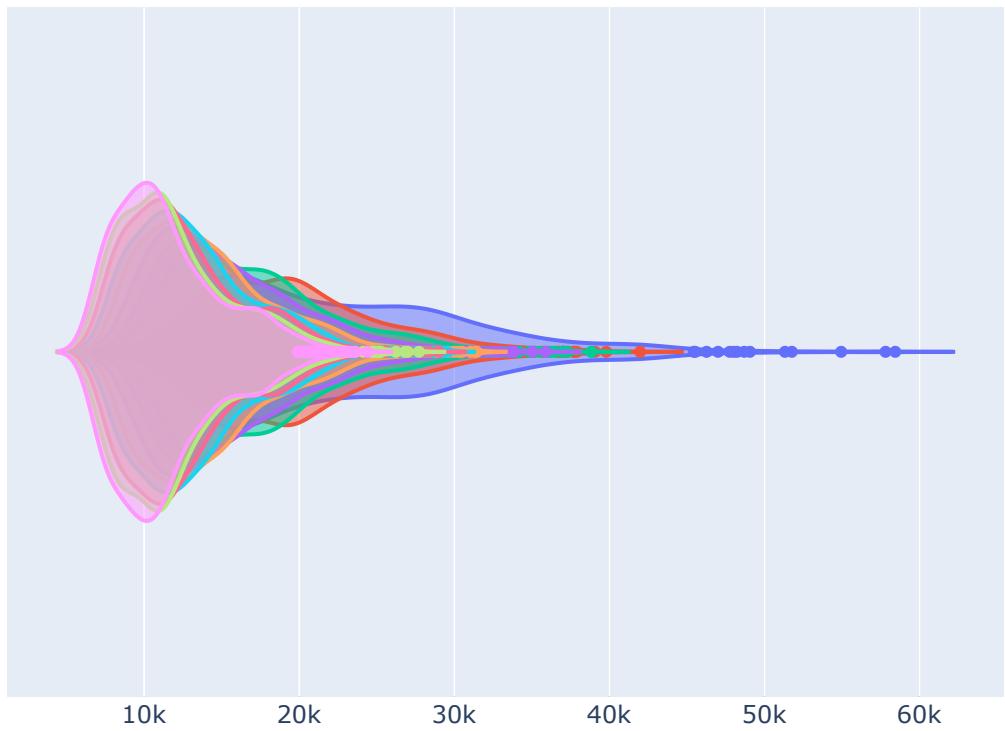
United States

Количество прослушиваний песен из топ-10 в United States



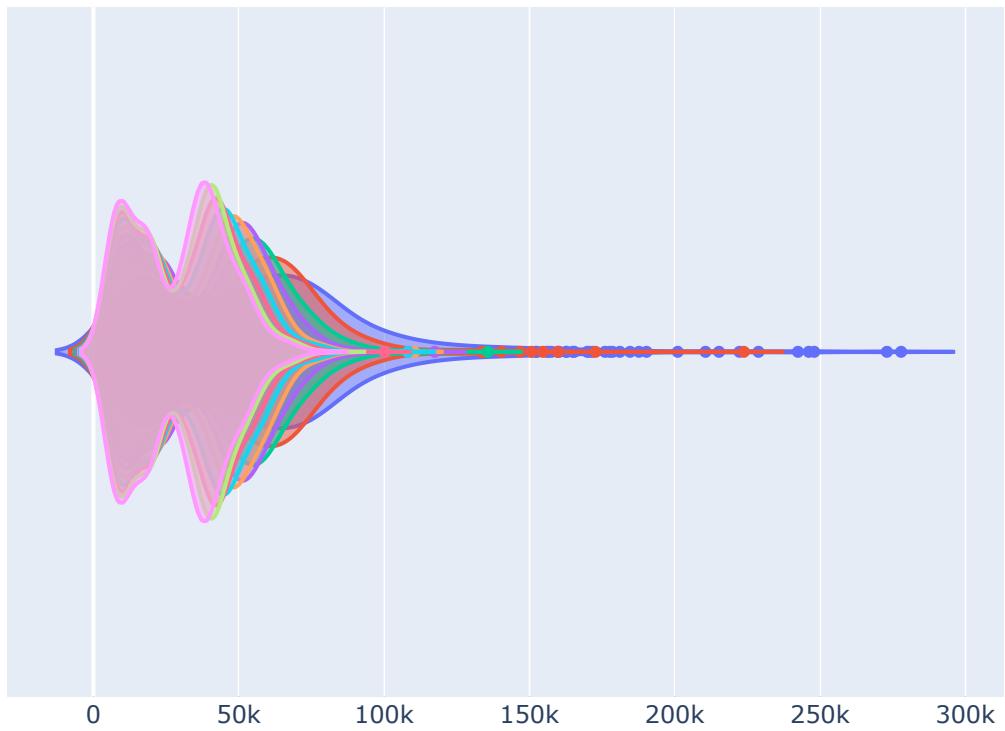
Uruguay

Количество прослушиваний песен из топ-10 в Uruguay



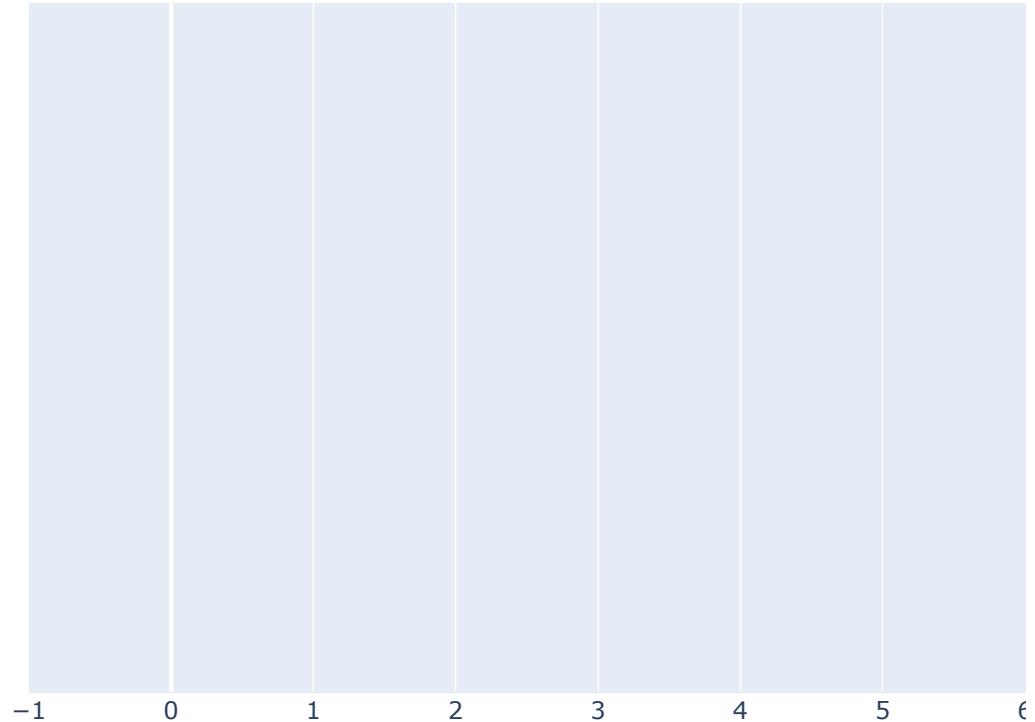
Thailand

Количество прослушиваний песен из топ-10 в Thailand



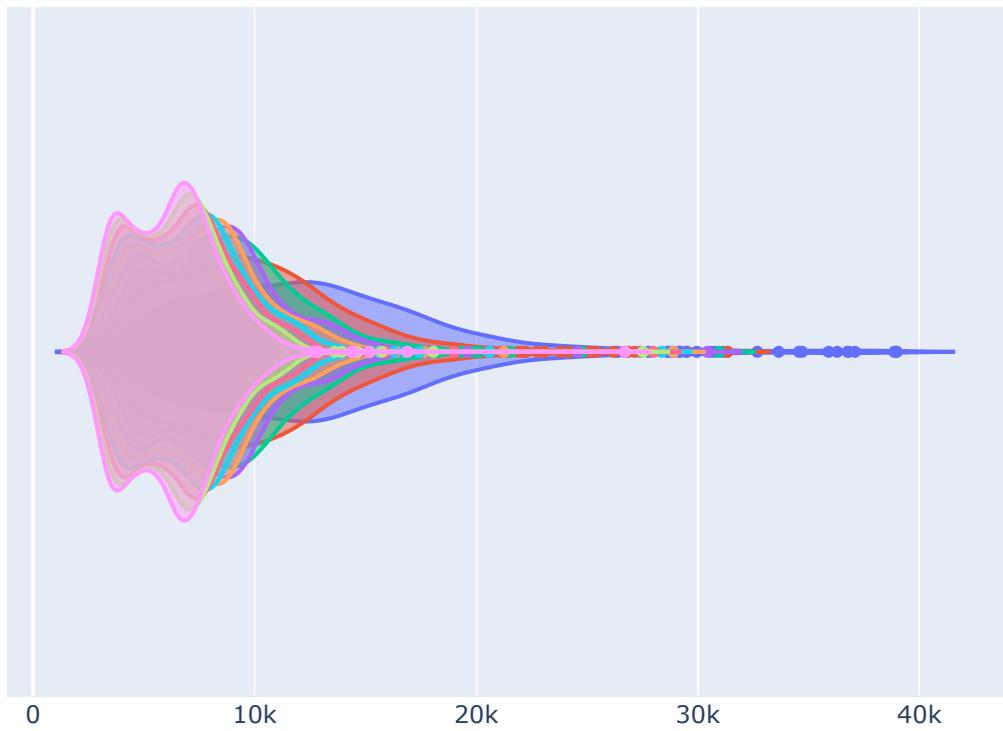
Andorra

Количество прослушиваний песен из топ-10 в Andorra



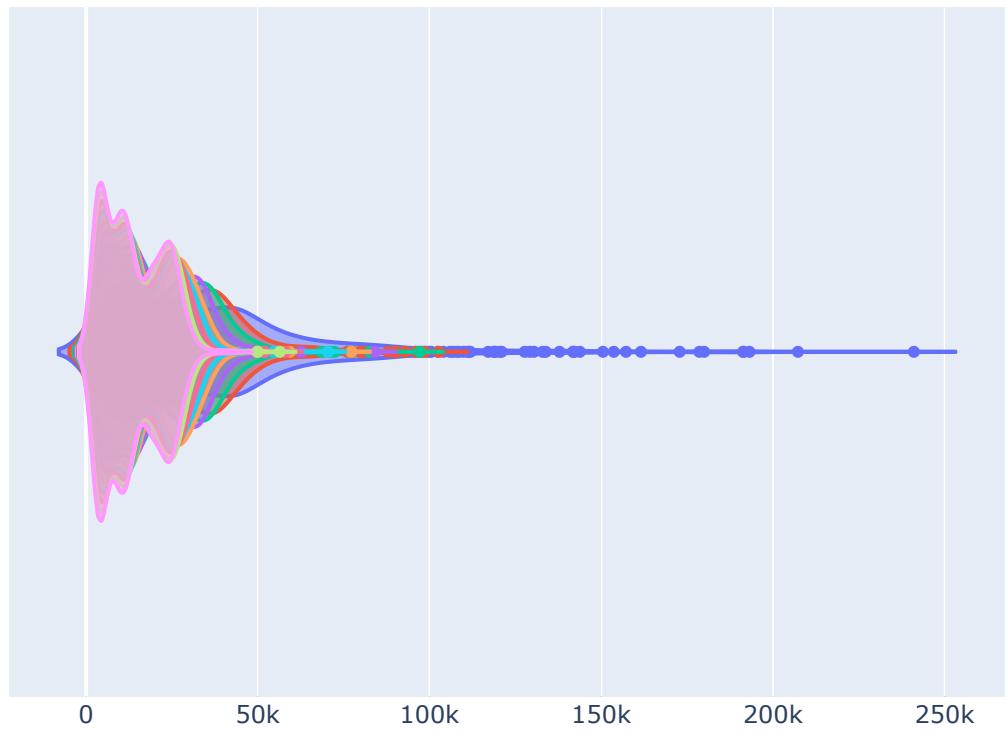
Romania

Количество прослушиваний песен из топ-10 в Romania



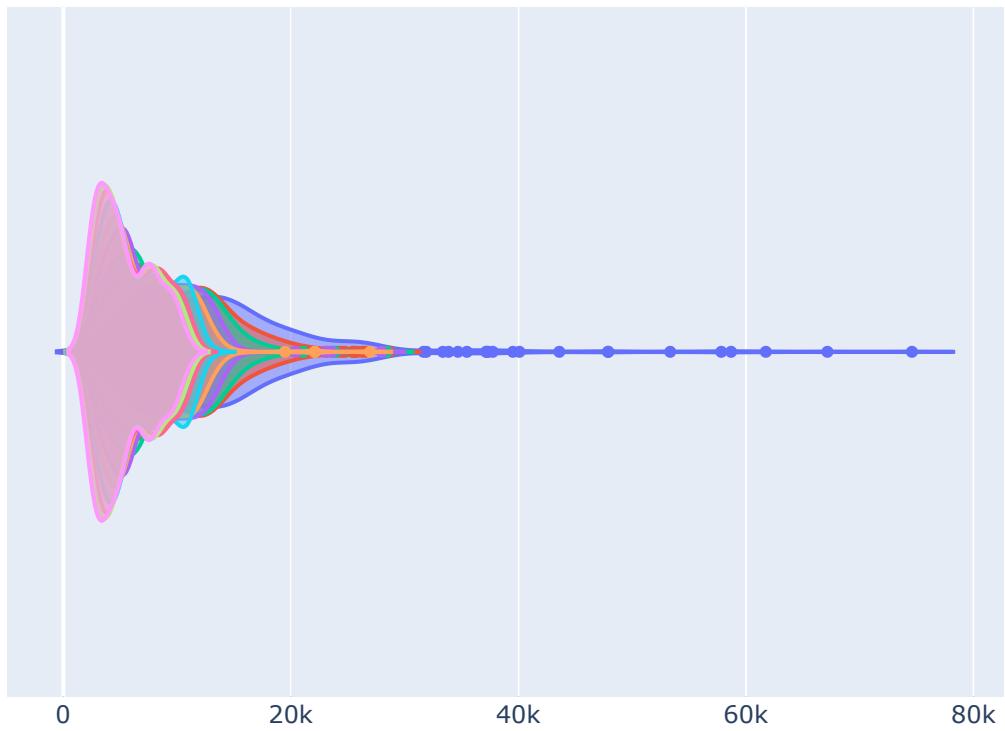
Vietnam

Количество прослушиваний песен из топ-10 в Vietnam



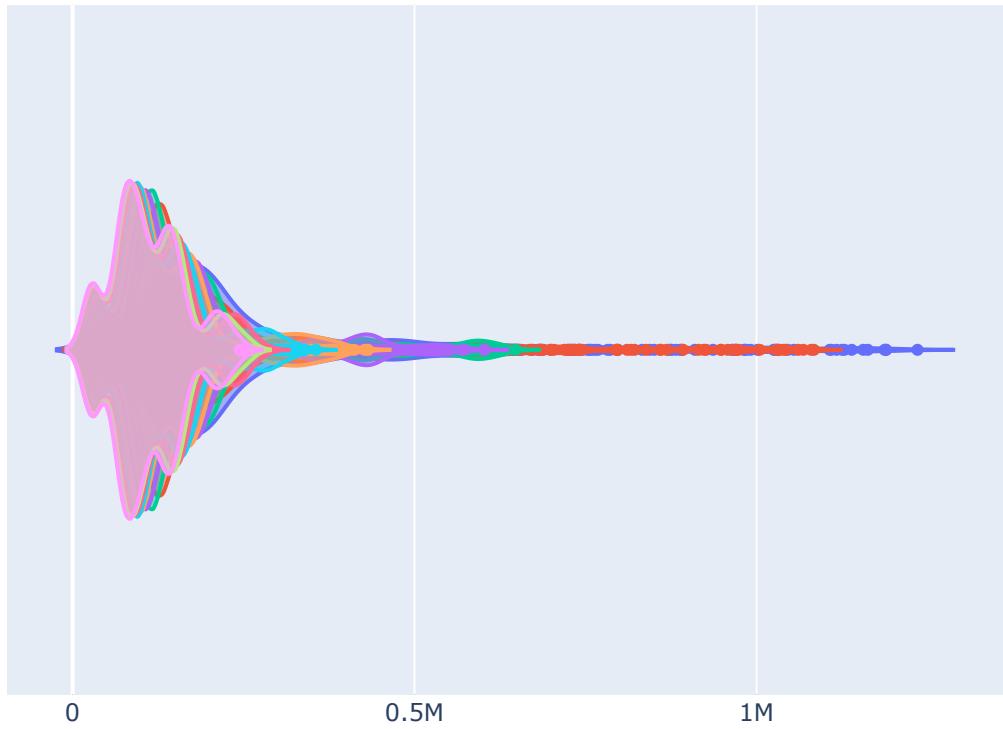
Egypt

Количество прослушиваний песен из топ-10 в Egypt



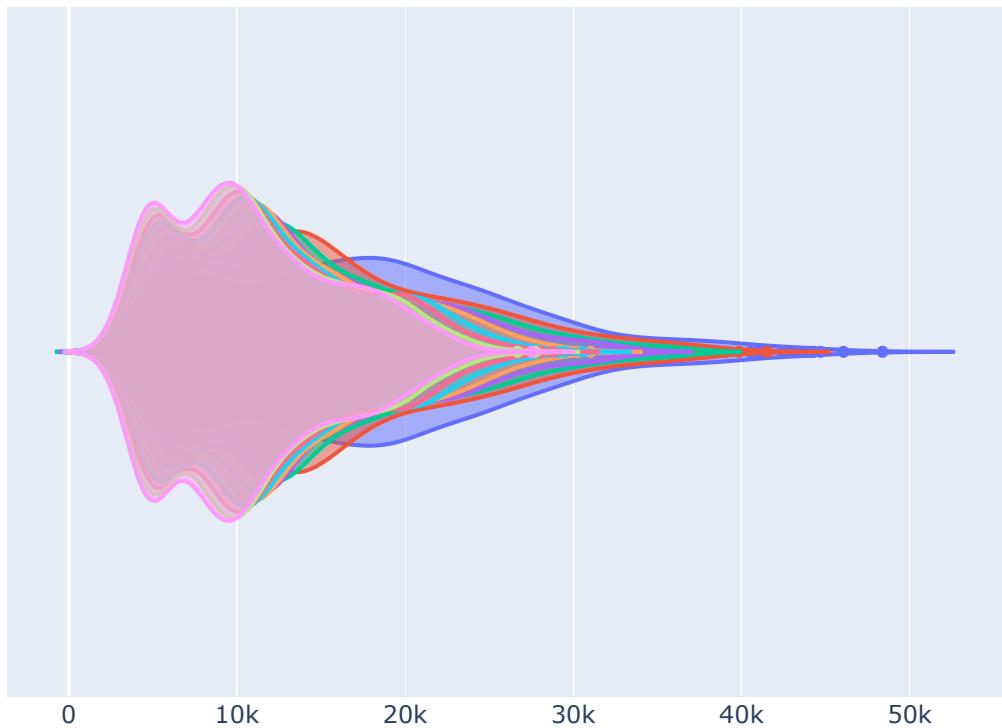
India

Количество прослушиваний песен из топ-10 в India



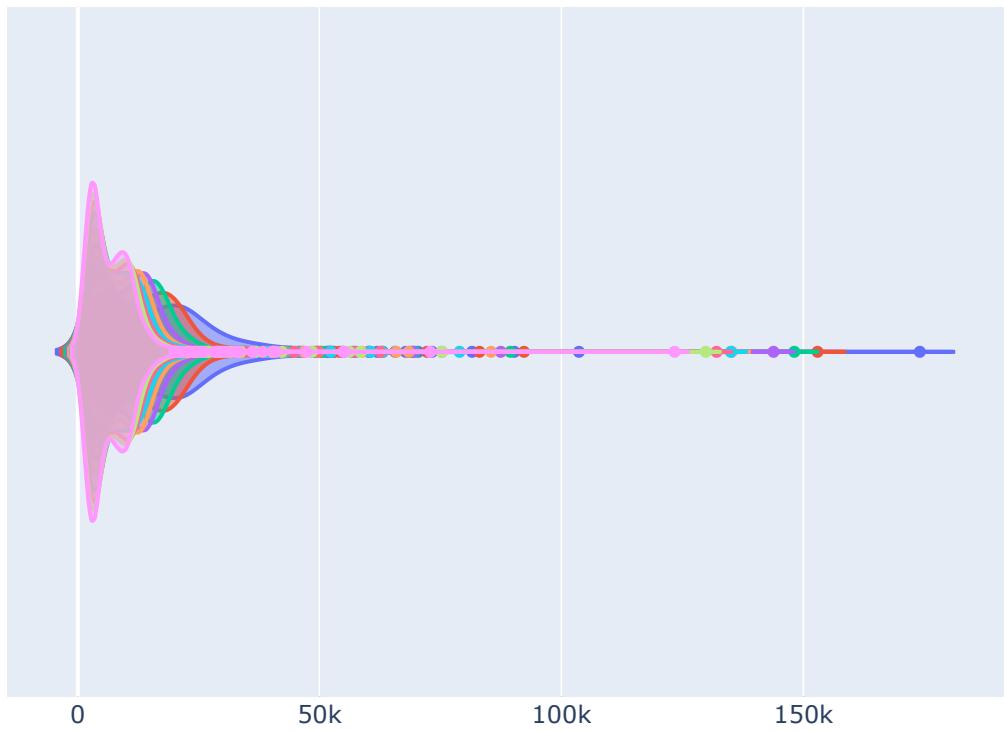
Israel

Количество прослушиваний песен из топ-10 в Israel



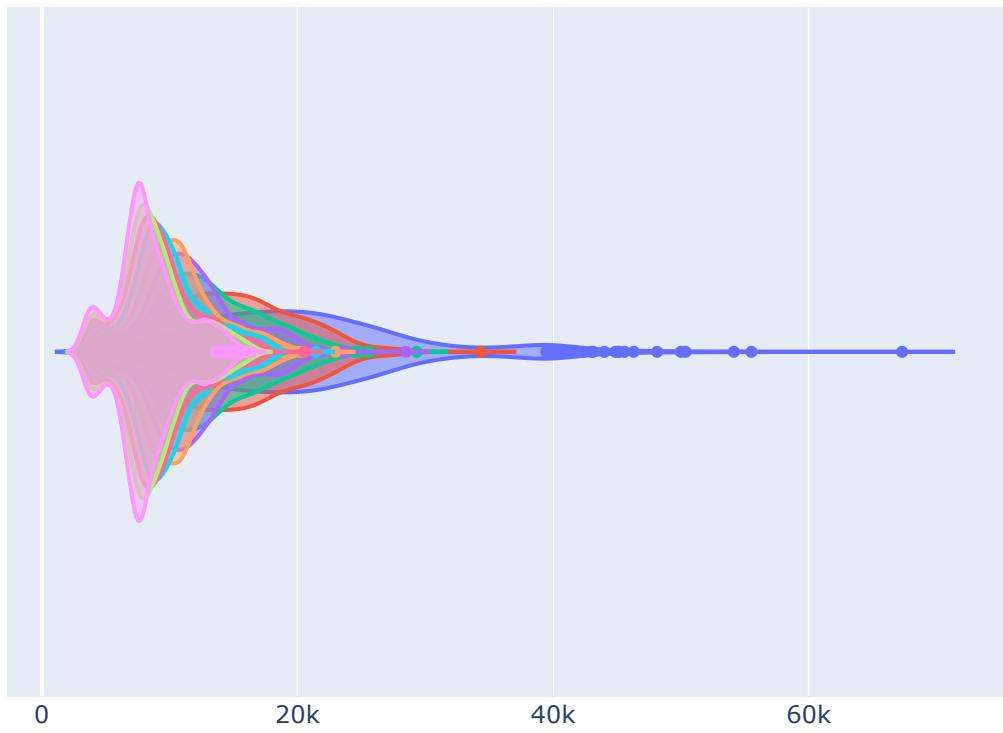
Morocco

Количество прослушиваний песен из топ-10 в Morocco



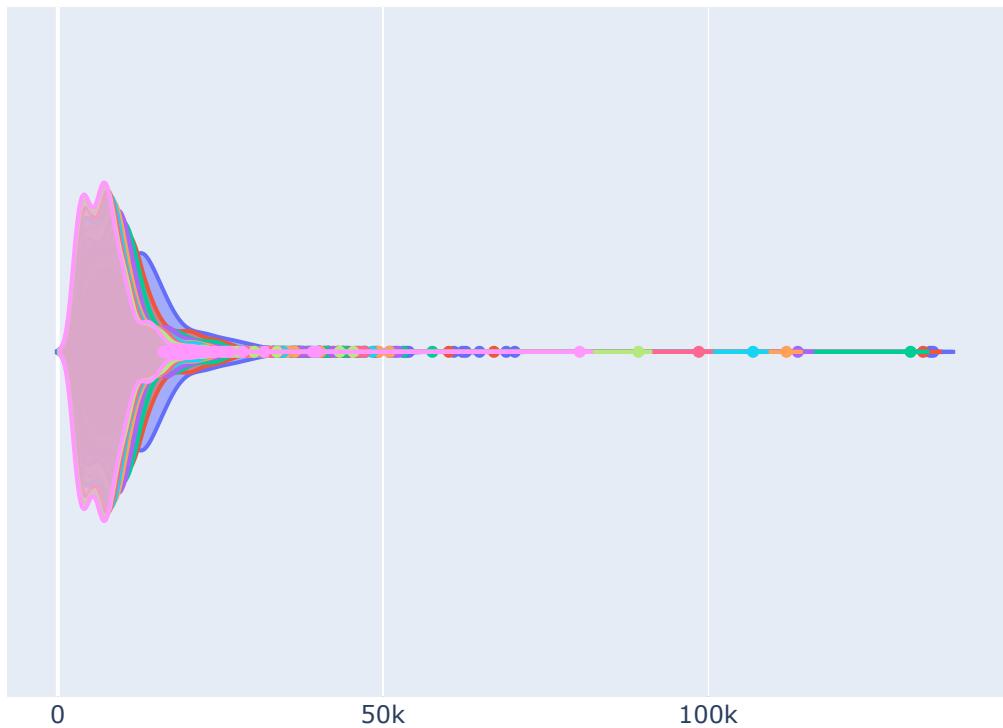
Saudi Arabia

Количество прослушиваний песен из топ-10 в Saudi Arabia



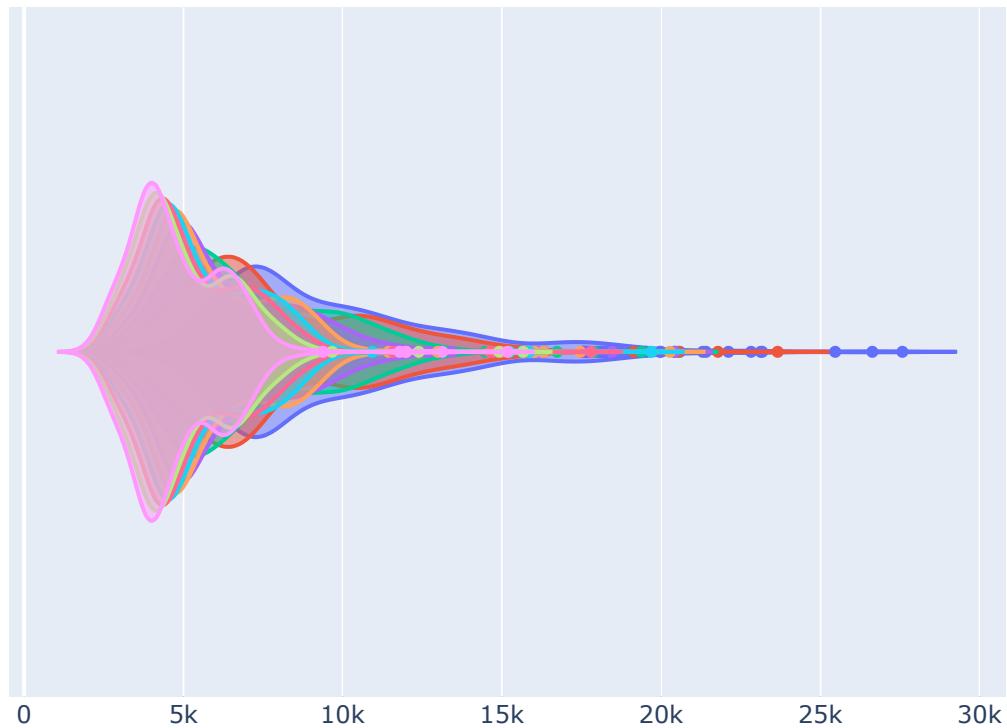
South Africa

Количество прослушиваний песен из топ-10 в South Africa



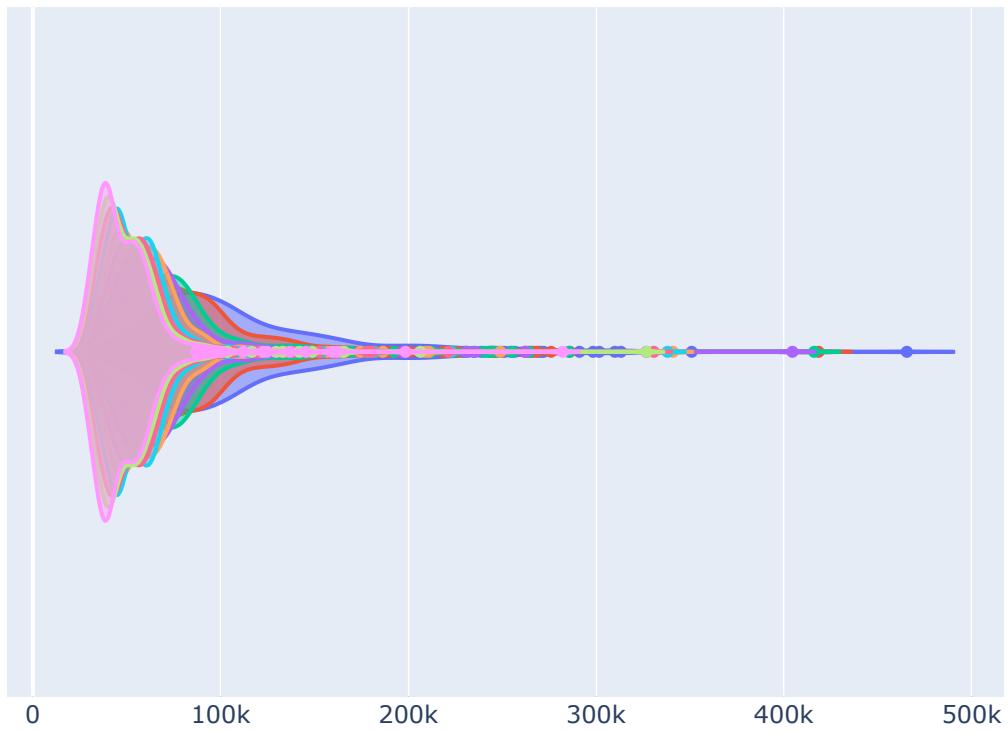
United Arab Emirates

Количество прослушиваний песен из топ-10 в United Arab Emirates



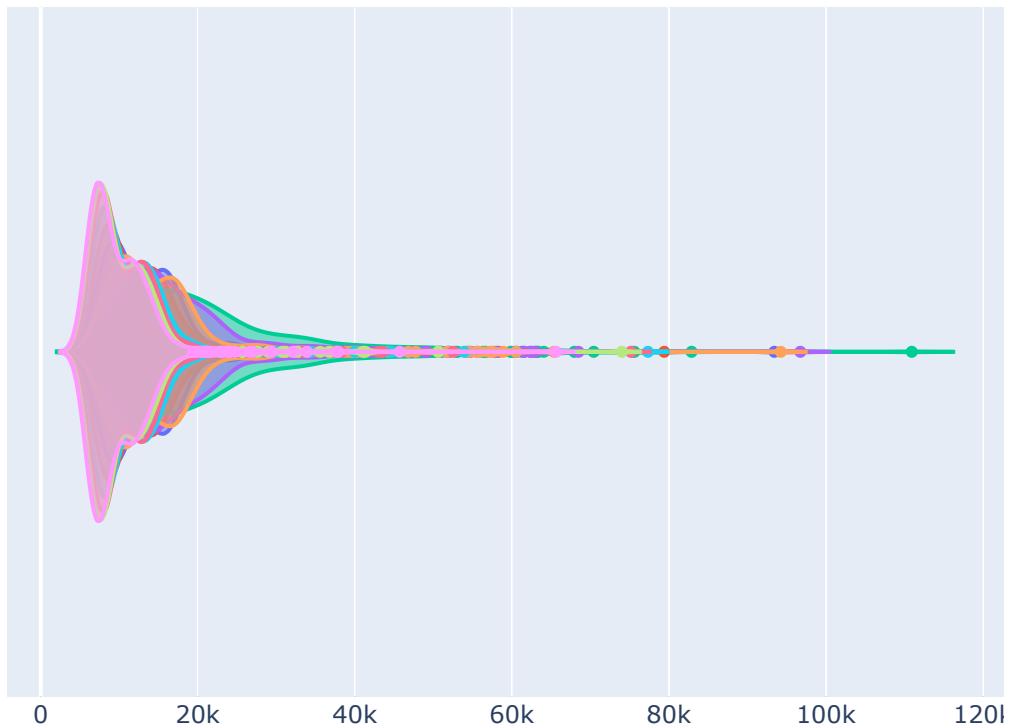
Russia

Количество прослушиваний песен из топ-10 в Russia



Ukraine

Количество прослушиваний песен из топ-10 в Ukraine



South Korea

Количество прослушиваний песен из топ-10 в South Korea

