Big Data Retake

Panov Evgenii

e.panov@innopolis.university

Project description

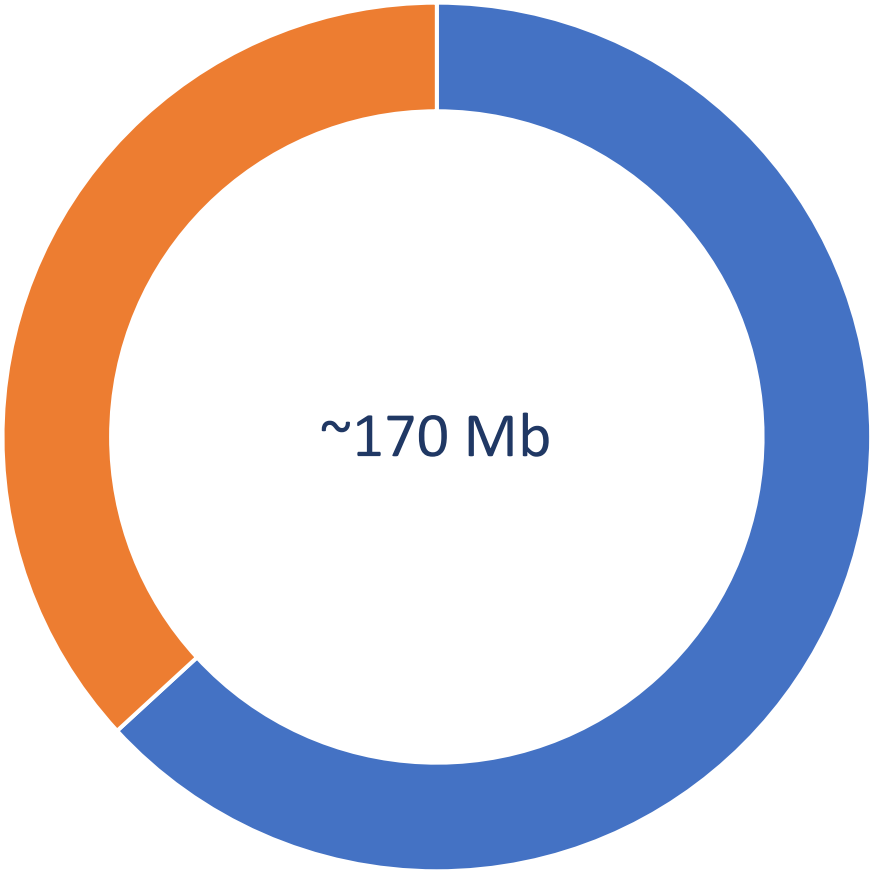
GOAL - Predict the popularity of Spotify tracks based on some features

1. Data

Datasets

Artists
1.1m items
62 Mb

Information about artists and their followers



Tracks
586k items
106 Mb

Information about tracks – release date, popularity, etc…

Artists schema

id	followers	genres	name	popularity
text	float	text	text	integer

Tracks schema

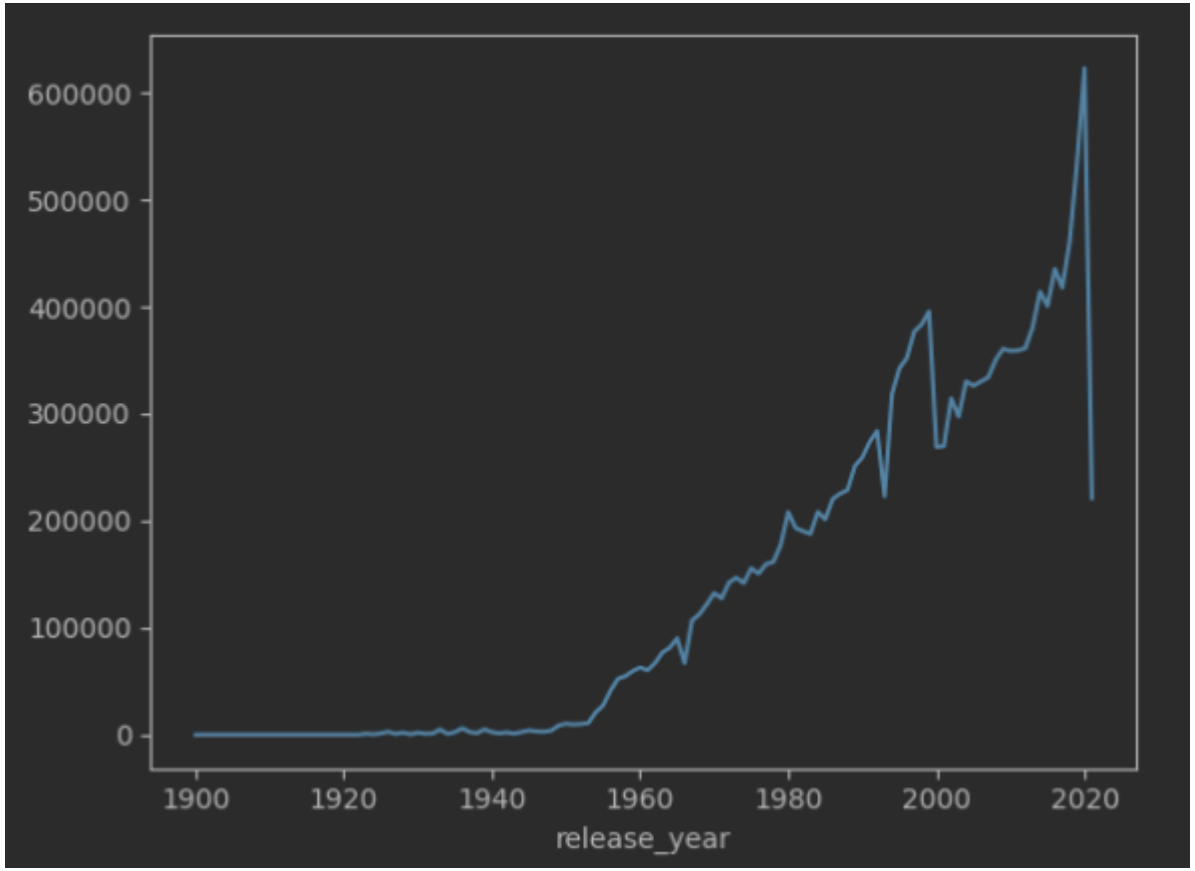
id	name	popularity	duration	explicit	artists	id_artists	release_date	danceability
text	text	integer	integer	integer	text	text	date	float

Tracks schema (cont.)

energy	loudness	speechiness	tempo	valence	time_signature
float	float	float	float	float	integer

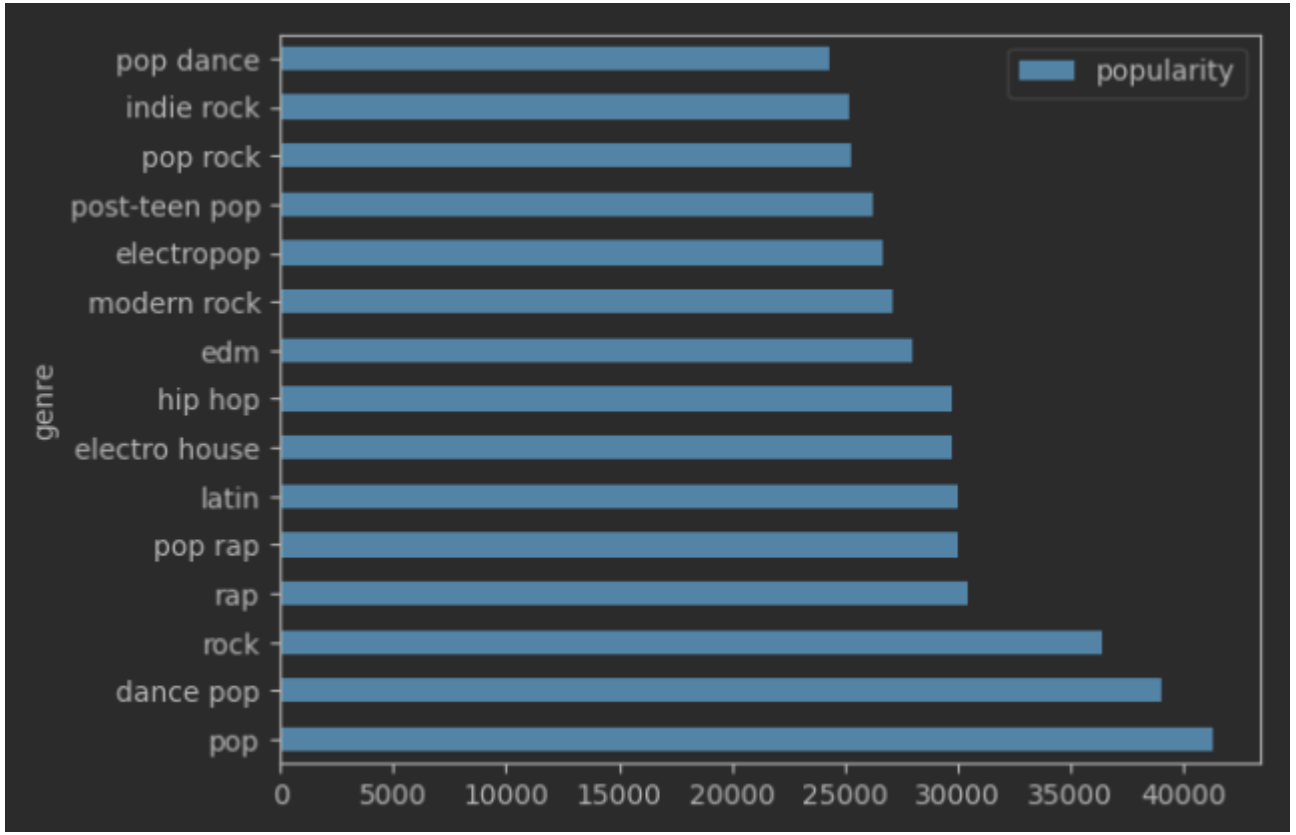
2. Data analytics

Aggregated tracks by release year



As we can see, there is more information and tracks for a later period of time

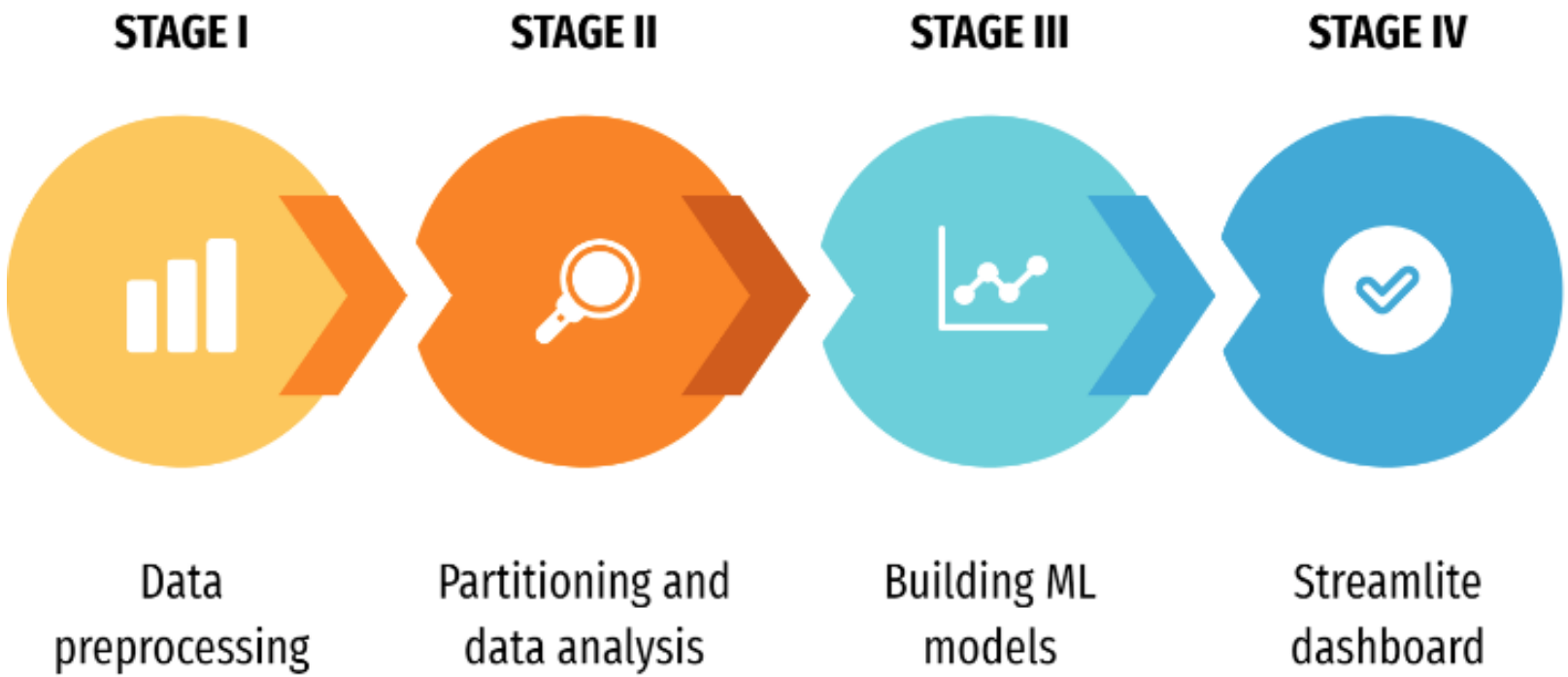
Some genre statistics



The most popular tracks are related to the pop and rap genres

3. Work process

Progress


STAGE I
STAGE II
STAGE III
STAGE IV
Data preprocessing
Partitioning and data analysis
Building ML models
Streamlite dashboard

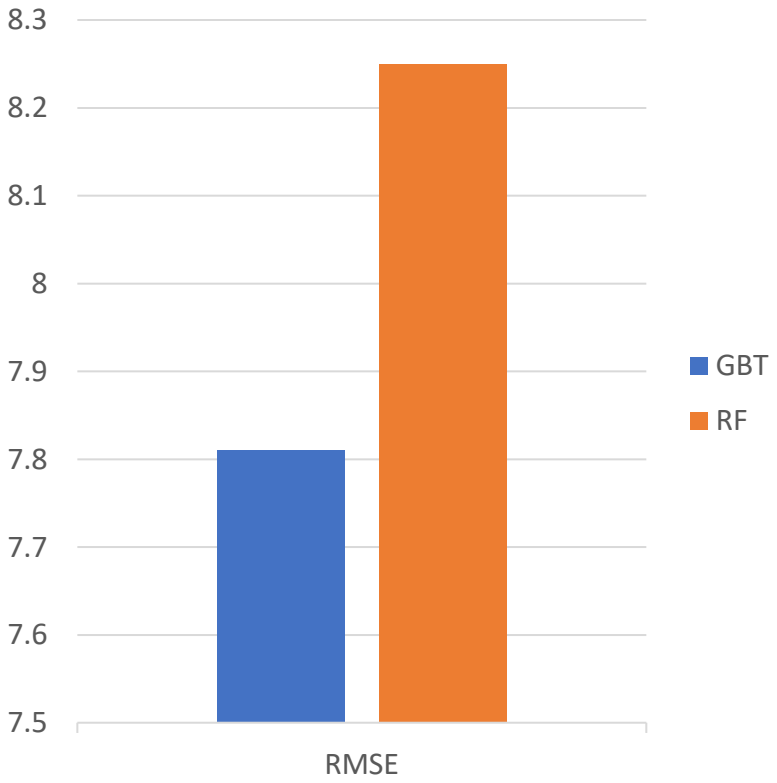
4. Challenges

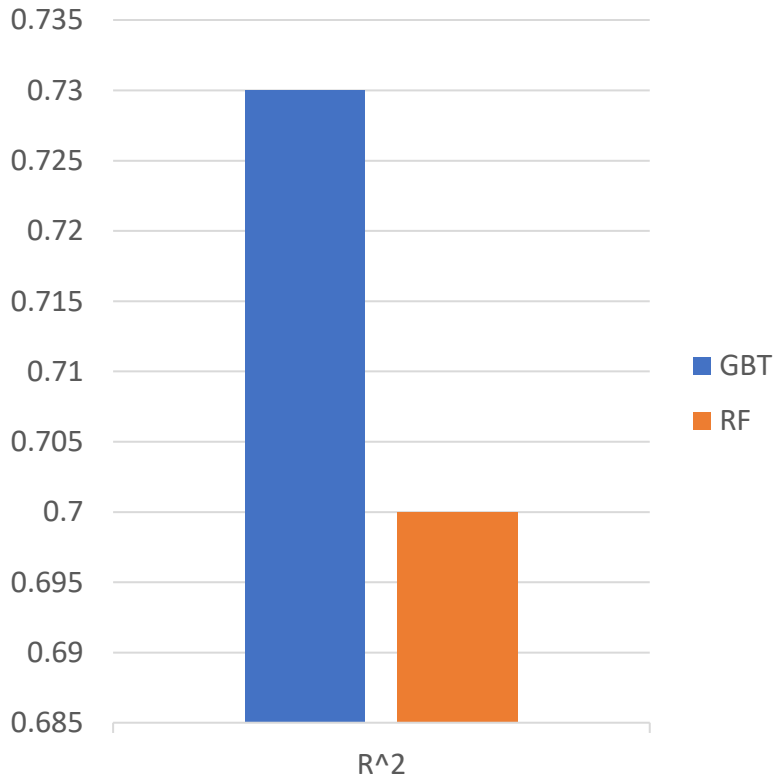
Difficulties

1. **Resources** - cluster limitations as it runs on local machine
2. **Old python** - strange errors, pip libraries dependencies difficulties
3. **New stack** - PySpark, Hive, HDFS
4. **Data** – parsing and converting between stages

5. Models performance

Model metrics





6. Conclusion

For me it was a good experience to work on this project, learn and interact with distributed file systems, see how the big data flow is being operated and aggregated. At least now I have an idea how it works and handles in large companies

Thanks for attention!