



TARTU ÜLIKOOL



Andmete visualiseerimise töötuba



Diana Pilvar
Andmehaldur
Arvutiteaduse instituut

Koolituste koordinaator
ELIXIR Eesti

Uku Raudvere
Teaduslik programmeerija
Arvutiteaduse instituut

Tehniline koordinaator
ELIXIR Eesti



Eesmärk: õppida andmeid kiirelt visualiseerima Google Colab abiga.

Õpiväljundid:

- Oskab üles laadida andmeid ja neist ülevaadet teha
- Teab, kuidas teha jooniseid, ning suudab otsida erandeid ja vigu andmestikust



Kui tihti kasutate tekstiroboteid?

**Kas olete varem proovinud jooniseid
teha tekstirobotitega?**

Google Colab



- Colaboratory on Google tasuta pilveteenus Jupyter Notebook keskkonnale
 - Colab on nagu Google Docs Pythonis
 - ühes kohas kood, tekst, pildid, HTML
- Ei ole vaja seadistada
- **Tasuta** ligipääs graafikaprotsessorile (GPU & TPU)
- Kerge jagada tulemusi, Notebookid on Google Drive kaustas
- Tasuta versiooni limiidid
 - Programm võib joosta max 12h
 - 12,5 GB RAM, kui on vaja rohkem siis krahh
- Colabis on Google Gemini tekstirobot



Pay As You Go

11,47 € for 100 compute units

52,39 € for 500 compute units

You currently have 0 compute units.

Compute units expire after 90 days.
Purchase more as you need them.

- ✓ No subscription required.
Only pay for what you use.
- ✓ Faster GPUs
Upgrade to more powerful GPUs.

Recommended

Colab Pro

11,47 € per month

Colab Pro for Education

Not available in your country

- ✓ 100 compute units per month
Compute units expire after 90 days.
Purchase more as you need them.
- ✓ Faster GPUs
Upgrade to more powerful GPUs.
- ✓ More memory
Access our highest memory machines.

Colab Pro+

52,39 € per month

🔥 Limited-time offer of an additional 100 compute units, totalling **600 per month**.

All of the benefits of Pro, plus:

- ✓ An additional ~~400~~ 500 compute units per month
Compute units expire after 90 days.
Purchase more as you need them.
- ✓ Faster GPUs
Priority access to upgrade to more powerful premium GPUs.
- ✓ Background execution
With compute units, your actively running notebook will continue running for up to 24 hours, even if you close your browser.

Colab Enterprise

Pay for what you use.

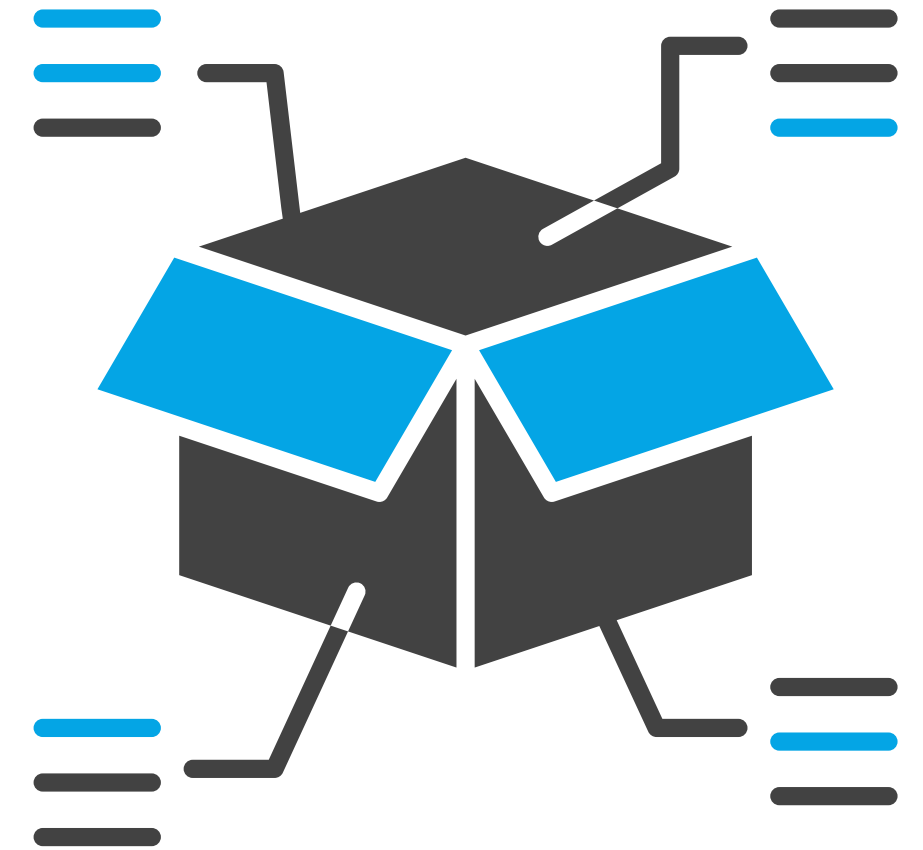
- ✓ Integrated
Tightly integrated with Google Cloud services like BigQuery and Vertex AI.
- ✓ Enterprise notebook storage
Replace your usage of Google Drive notebooks with GCP notebooks, stored and shared within your cloud console.
- ✓ Productive.
Generative AI powered code completion and generation.

Colab Enterprise

- Notebooke saab jagada organisatsioonis koos juurdepääsuõiguste detailse kontrolliga.
- Kliendi hallatavad krüpteerimisvõtmed (Customer-Managed Encryption Keys): võimaldavad teil oma andmeid jõudeolekus krüpteerida, kasutades enda krüpteerimisvõtmeid, et tagada parem turvalisus ja vastavus nõuetele.
- Arvutusvõimsuse pakkumine vastavalt vajadusele ja sulgemine jõudeolekul, et optimeerida ressursside kasutust ja kulusid.
- Konfigureeritavad käituskeskkonnad kindlate masinatüüpide, kiirendite (GPU/TPU) ja kettaruumiga
- Integratsioon Vertex AI ja BigQueryga.
- Notebookide ajastatud käivitamine.

Omadused

- Sobib andmete avastamiseks, masinõppeks, andmeanalüüsiks, hariduslikel eesmärkidel
- Käitamise ajalugu
- GitHub-iga linkimine
- Ette-installeeritud teegid
 - Pandas ja NumPy andmeanalüüsiks
 - Matplotlib ja Seaborn andmete visualiseerimiseks
 - joon-, sektor- ja tulpdiagrammid
 - Histogrammid
 - Hajusgraafik (*scatter plot*)
 - alagraafikute loomine
 - 3D- ja interaktiivsed graafikud
 - *heatmap*



Gemini Colabis

- Koodi genereerimine loomulikus keeles käskluste abil (eesti, inglise)
- Veaseletus
 - Mis on valesti ja miks
 - Veaparandus
- Silumine (*Debugging*)
- Gemini chat
 - Tuutor
 - Küsi ülevaadet Pythoni funktsioonidest, teekidest, programmeerimisest
 - Vaata soovitatud jooniseid
 - Kommenteerib koodi, et sa teaks iga rea funktsiooni

Gemini Colabis

- Koodi genereerimine loomulikus keeles käskluste abil (eesti, inglise)
- Veaseletus
 - Mis on valesti ja miks
 - Veaparandus
- Silumine (*Debugging*)
- Gemini chat
 - Tuutor
 - Kõnelema funktsioonidest, teekidest, programmeerimisest
 - Vahetult soovitatud jooniseid
 - Kommenteerib koodi, et sa teaks iga rea funktsiooni

**KOOD TAGAB
REPRODUTSEERITAVUSE!**

Privaatsus

- Google kogub
 - viipasid (*prompte*),
 - koodi,
 - tulemust,
 - funktsioonide kasutussagedus
 - tagasiside
- Ülevaatajad võivad lugeda, annoteerida ja töödelda teilt kogutud andmeid
- Andmeid säilitatakse kuni 18 kuud!
- Kustutada pole võimalik



Privaatsus

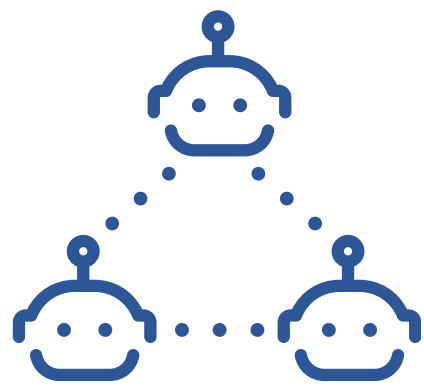
- Google kogub
 - viipasid (*prompte*),
 - koodi,
 - tulemust,
 - funktsioonide kasutussagedus
 - tagasiside
- Ülevaatajad võivad lugeda, annotateerida ja töödelda teilt kogutud andmeid
- Andmeid säilitatakse kuni 18 kuud!
- Kustutada pole võimalik



**Palun ärge sisestage tundlikke
või konfidentsiaalseid andmeid!**



Colab Gemini Agent



- Laadi üles andmestik(ud)
- Kirjelda, mida oma andmetest teada saada soovid
- Agent genereerib:
 - Koodi
 - Selgitused
 - Joonised
 - Päringu töötlemise ajal saad jälgida selle samm-sammulist arutluskäiku, planeerimist ja täitmist.
- Andmete puhastamine: Tegeleb puuduvate väärtuste, erindite ja vastuoludega !
- Uuriv andmeanalüüs (EDA): Genereerib kirjeldavat statistikat, jooniseid ja ülevaateid
- Statistiline analüüs



TARTU ÜLIKOOL

Demo

(avan colabi)

colab.google.com

(kirjutan viiba)

Gemini ✕



I need you to analyse and describe this dataset for me:

<https://raw.githubusercontent.com/vega/vega-datasets/master/data/movies.json>

(Gemini vastab)



I can help with that! I'll load the dataset from the provided URL using pandas and display the first few rows and the data types so we can start understanding the data.

(kirjutab koodi ja käitab)

```
import pandas as pd

url = "https://raw.githubusercontent.com/vega/vega-datasets/master/data/movies.json"
df = pd.read_json(url)

display(df.head())
display(df.info())
```

	Title	US Gross	Worldwide Gross	US DVD Sales	Production Budget	Release Date	MPAA Rating	Running Time min	Distributor	Source	Major Genre	Creative Type	Director	Rotten Tomatoes Rating
0	The Land Girls	146083.0	146083.0	NaN	8000000.0	Jun 12 1998	R	NaN	Gramercy	None	None	None	None	NaN
1	First Love, Last Rites	10876.0	10876.0	NaN	300000.0	Aug 07 1998	R	NaN	Strand	None	Drama	None	None	NaN
2	I Married a Strange Person	203134.0	203134.0	NaN	250000.0	Aug 28 1998	None	NaN	Lionsgate	None	Comedy	None	None	NaN
3	Let's Talk About Sex	373615.0	373615.0	NaN	300000.0	Sep 11 1998	None	NaN	Fine Line	None	Comedy	None	None	13.0
4	Slam	1009819.0	1087521.0	NaN	1000000.0	Oct 09 1998	R	NaN	Trimark	Original Screenplay	Drama	Contemporary Fiction	None	62.0

(Gemini vastab)

Now that the data is loaded, I can perform a more detailed analysis. I will calculate descriptive statistics for numerical columns, check for missing values, and look at the unique values in some of the categorical columns to get a better sense of the data's characteristics. Based on the analysis I've performed so far, here's a description of the dataset:

- **The dataset contains information about movies**, including financial data (US Gross, Worldwide Gross, Production Budget, US DVD Sales), release information (Release Date), ratings (MPAA Rating, Rotten Tomatoes Rating, IMDB Rating, IMDB Votes), and other attributes like Running Time, Distributor, Source, Major Genre, Creative Type, and Director.
- **There are a significant number of missing values in several columns**, particularly in 'US DVD Sales', 'Running Time min', 'Director', 'Rotten Tomatoes Rating', 'IMDB Rating', and 'IMDB Votes'. These missing values will need to be handled depending on the type of analysis you want to perform.
- The descriptive statistics provide insights into the distribution of numerical columns like gross revenues, production budget, and ratings. For example, **there's a wide range in US Gross and Worldwide Gross**, indicating the presence of blockbuster movies alongside lower-grossing films.
- The unique values in categorical columns like 'MPAA Rating', 'Major Genre', 'Creative Type', and 'Distributor' show the different categories present in the dataset. 'MPAA Rating' has standard ratings like R, PG, PG-13, G, and also 'Not Rated' and 'None'. 'Major Genre' covers common film genres, and 'Distributor' lists various movie distribution companies.

What would you like to explore next? We could:

- Visualize some of the data distributions (e.g., box office revenue, ratings).
- Explore the relationships between different variables (e.g., production budget vs. worldwide gross).
- Handle the missing values in a specific way (e.g., imputation or removal).
- Analyze the data based on categorical variables like genre or MPAA rating.
- Focus on analyzing a specific subset of the data.

(Mina küsin)



I need an overview of the release dates of the movies

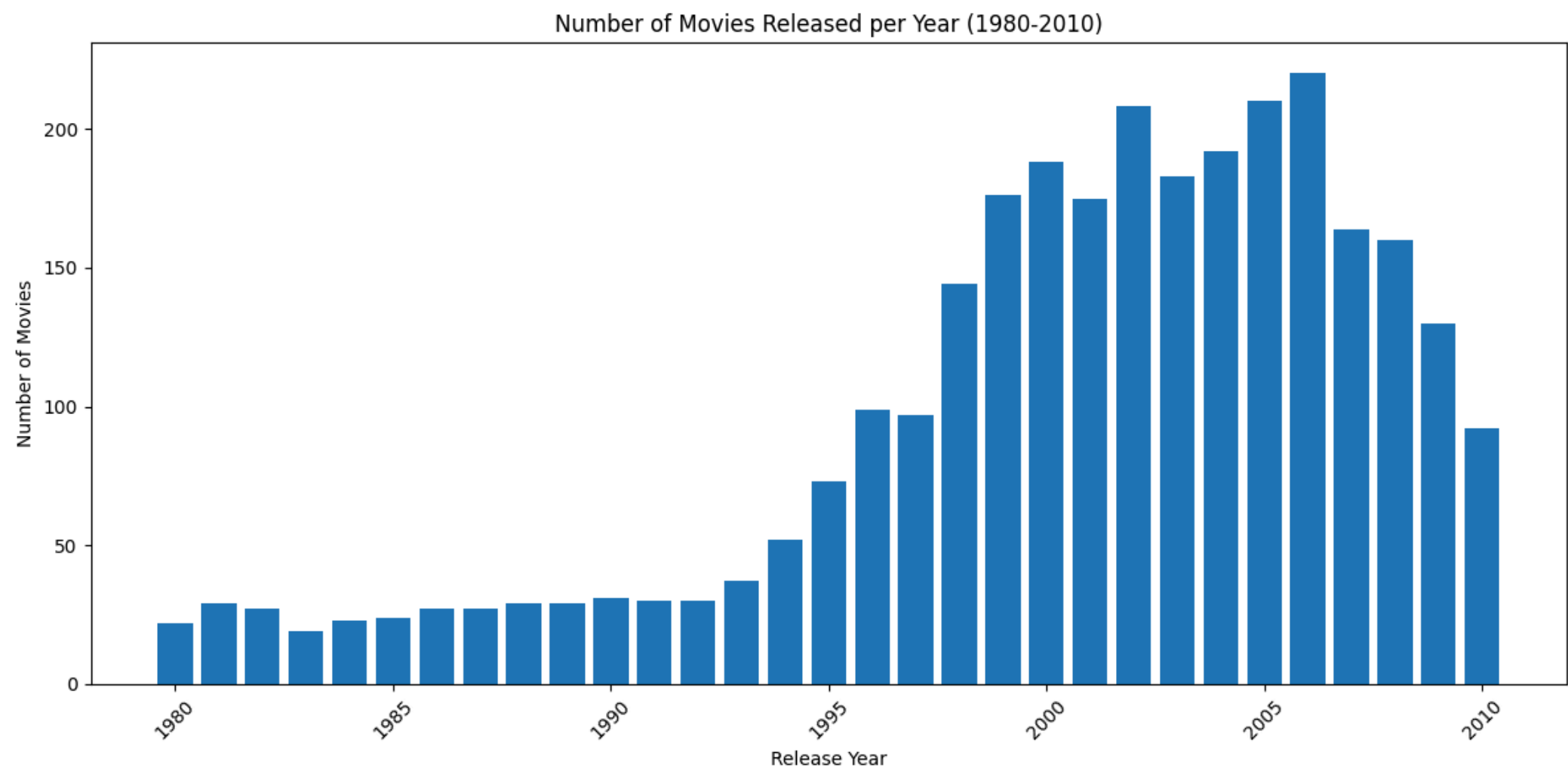
(Gemini kirjutab koodi ja käitab)

```
df['Release Date'] = pd.to_datetime(df['Release Date'], errors='coerce')  
  
display("Release Date Range:")  
display(f"Earliest Release Date: {df['Release Date'].min()}")  
display(f"Latest Release Date: {df['Release Date'].max()}")
```

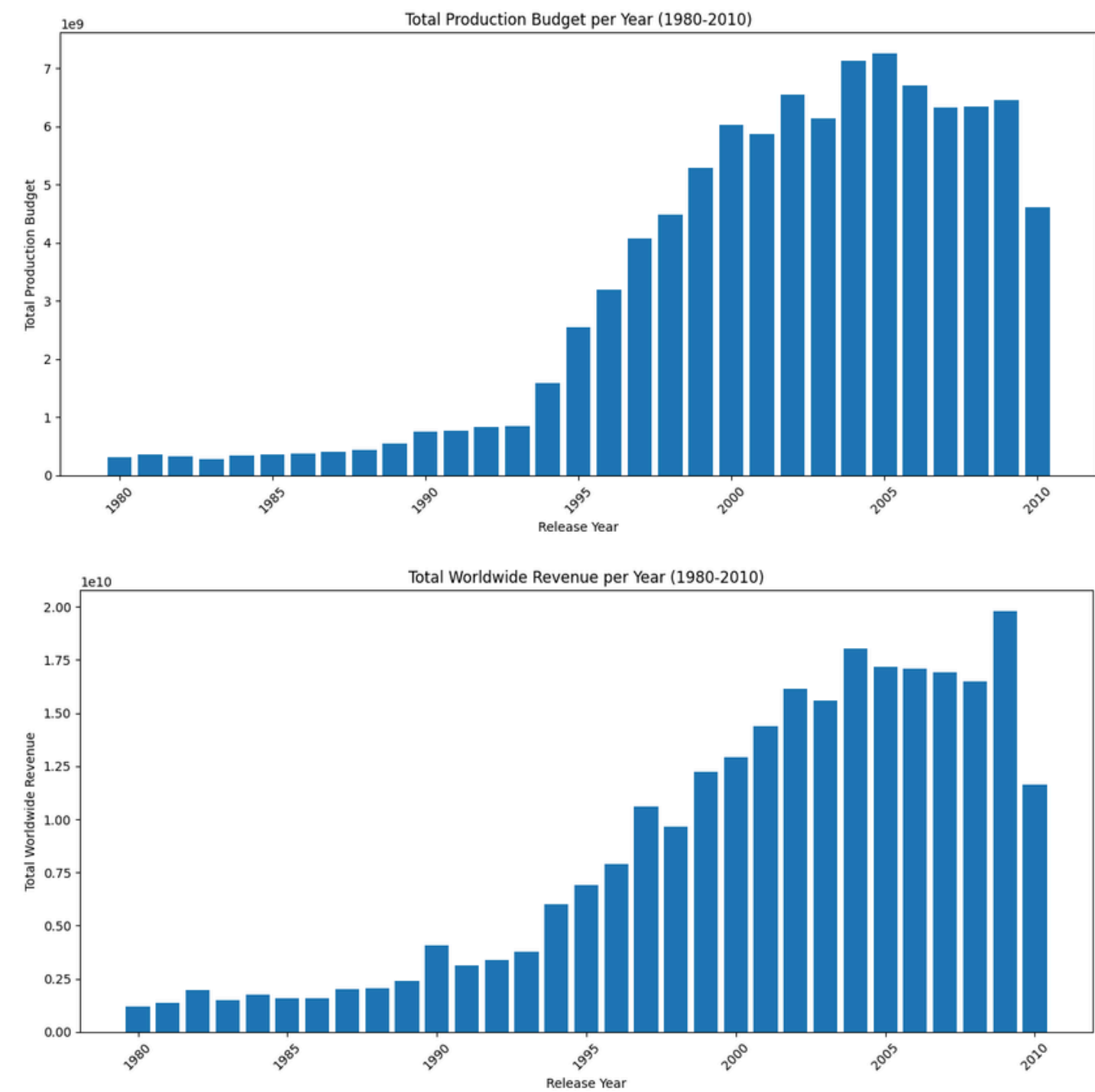
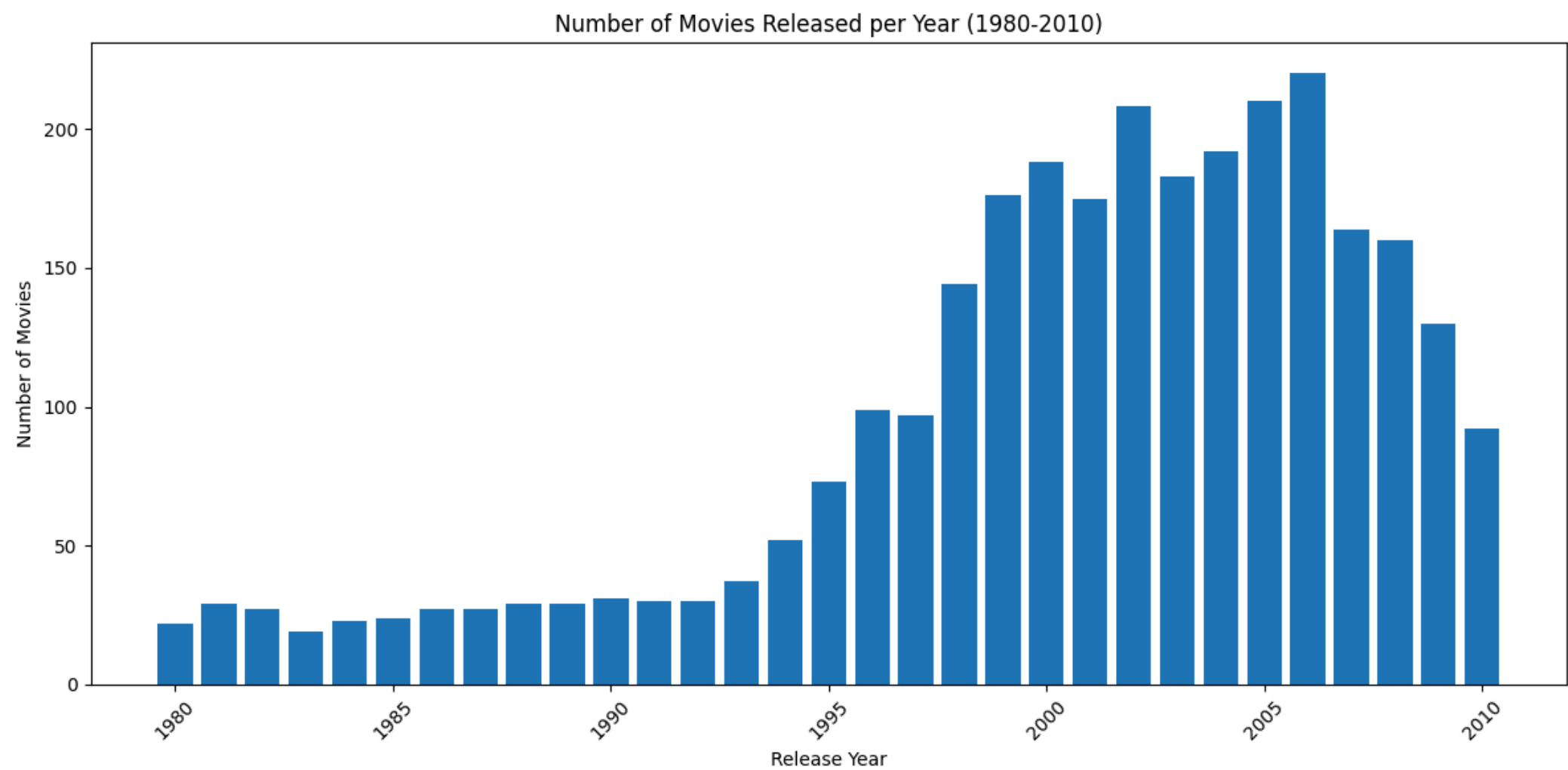
```
➞ 'Release Date Range:'  
'Earliest Release Date: 1928-12-31 00:00:00'  
'Latest Release Date: 2046-12-31 00:00:00'
```

I need a barplots of: the number of movies released by year, the total budget of the industry by year the total revenue of the industry by year cap the data between 1980 and 2010

I need a barplots of: the number of movies released by year, the total budget of the industry by year the total revenue of the industry by year cap the data between 1980 and 2010



I need a barplots of: the number of movies released by year, the total budget of the industry by year the total revenue of the industry by year cap the data between 1980 and 2010



the drop-off at 2010
looks weird. Show me
the number of movies
produced between
2000 and 2020

the drop-off at 2010
looks weird. Show me
the number of movies
produced between
2000 and 2020

```
1 df_2000_2020 = df[(df['Release Year'] >= 2000) & (df['Release Year'] <= 2020)].copy()
2 movies_per_year_2000_2020 = df_2000_2020.groupby('Release Year').size().reset_index(name='Movie Count')
3
4 display("Number of movies released per year (2000-2020):")
5 display(movies_per_year_2000_2020)
6
```

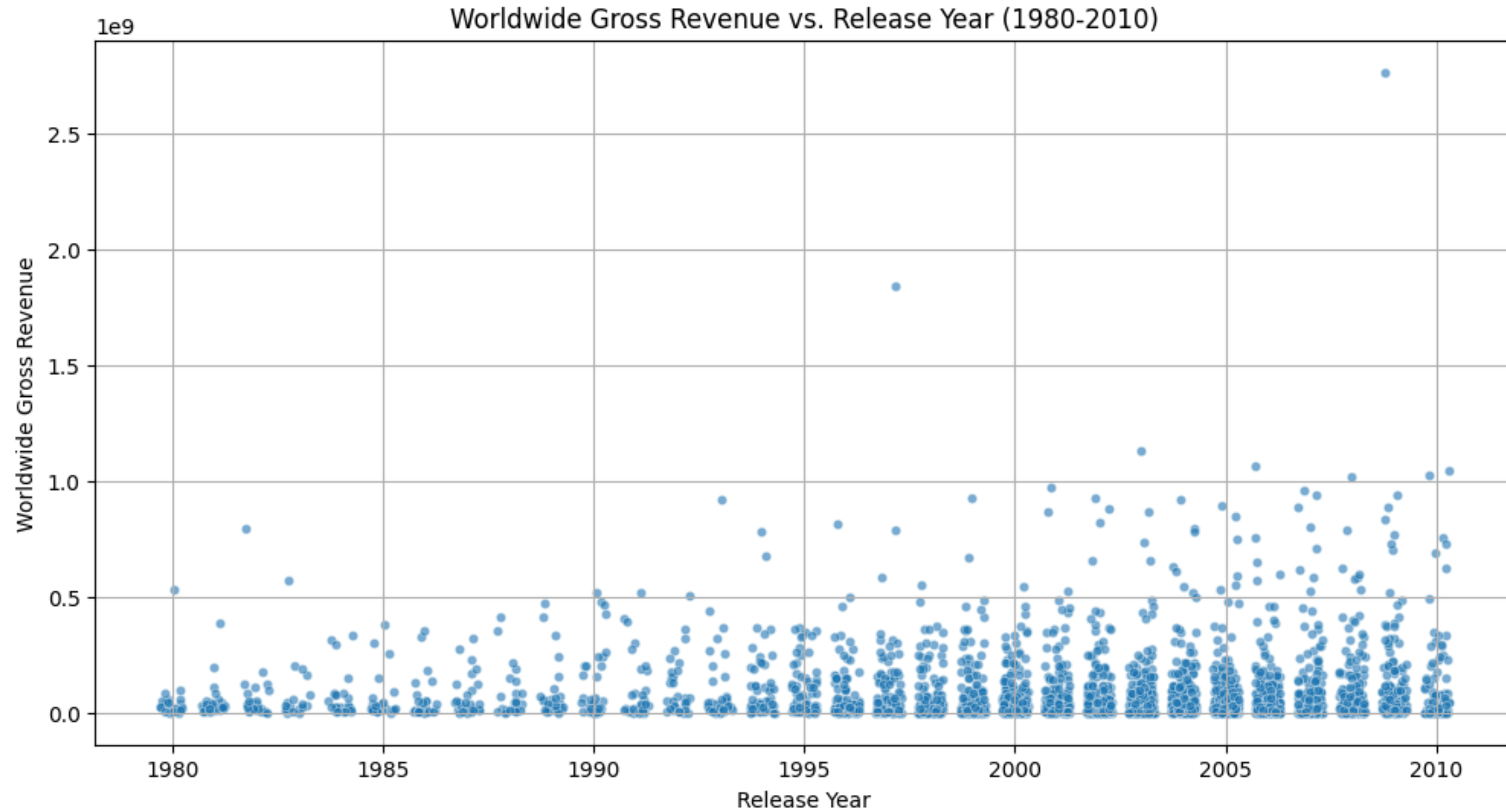
the drop-off at 2010
looks weird. Show me
the number of movies
produced between
2000 and 2020

```
1 df_2000_2020 = df[(df['Release Year'] >= 2000) & (df['Release Year'] <= 2020)].copy()
2 movies_per_year_2000_2020 = df_2000_2020.groupby('Release Year').size().reset_index(name='Movie Count')
3
4 display("Number of movies released per year (2000-2020):")
5 display(movies_per_year_2000_2020)
6
```

'Number of movies released per year (2000-2020):'

	Release Year	Movie Count
0	2000	188
1	2001	175
2	2002	208
3	2003	183
4	2004	192
5	2005	210
6	2006	220
7	2007	164
8	2008	160
9	2009	130
10	2010	92
11	2011	2
12	2015	1
13	2016	2
14	2020	1

I would like a scatterplot of movies. year to total gross revenue. with a small jitter and opacity for clarity, keep dates between 1980 and 2010





TARTU ÜLIKOOL

(veel üks)

Demo

Soovitused

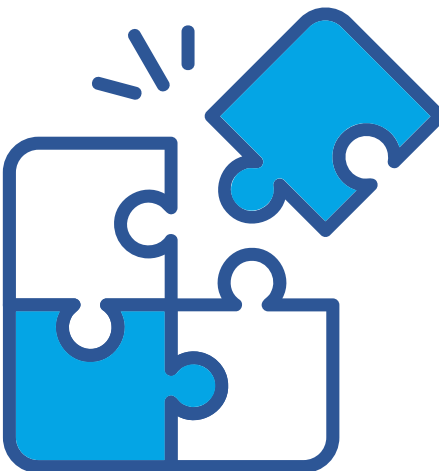


1. Alusta alati **kokkuvõtlikust statistikast**
2. **Joonised** näitavad, kas andmetes on probleeme
3. **Täpsemad küsimused** agendile annavad suurema tõenäosusega seda, mida sa tahad
 - a. Kasuta **<https://www.data-to-viz.com>**, et valida **parimad joonised sinu andmetüübile**
4. Agent võib puuduvate andmetega toiminguid, mida sa ei soovi. Ole tähelepanelik!
5. Mõnikord **võib agent eksida** lihtsate asjadega.

Praktiline ülesanne

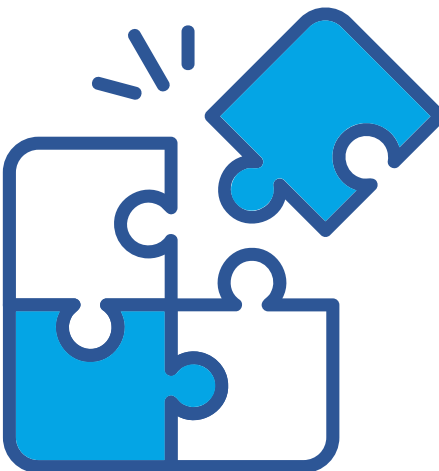
- Ülesande materjalid on siin:
 - <https://raw.githubusercontent.com/vega/vega-datasets/master/data/movies.json>
- Ava Google Colab <https://colab.research.google.com/>
 - Menüüst File - Open new notebook
- Vajuta nupule + Code
- “Start coding or generate with AI” - **Vajuta generate with AI**
 - **Lehekülje all avaneb vestlusaken**
 - Pane sinna andmete link ja anna käsklus: “loe andmed sisse”
 - Accept and run - AI koostatud käsule
 - Vaata, mis andmetega tegu on
- Küsi summeeritud statistikuid

<https://raw.githubusercontent.com/vega/vega-datasets/master/data/movies.json>



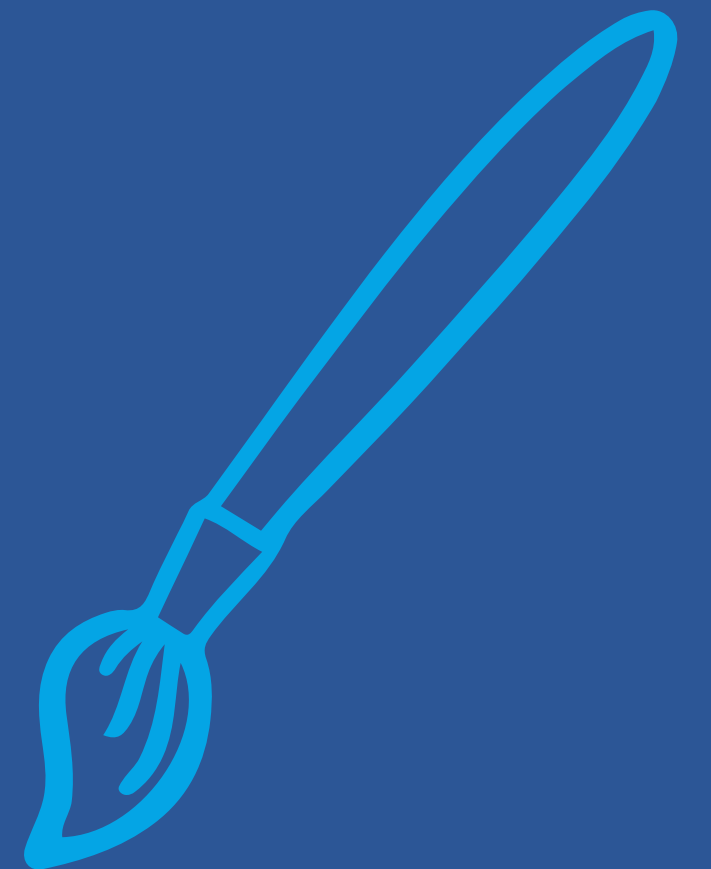
Praktiline ülesanne

- Tee tulpdiagramm filmide žanri ja sageduse kohta
 - Muuda värve
 - Liiguta diagramm horisontaali
- Tee mullidiagramm, kus
 - x teljel on USA kasum
 - y teljel Maailma kasum
 - muli värv põhižanr
 - muli suurus eelarve
- Uuri diagrammi
- Uuri anomaalia põhjust



Ära upu andmetesse, loo neist pilt.

Colab on sinu pintsel.





TARTU ÜLIKOOL

Täname tähelepanu eest!



unitartu



tartuylkool

