**DATA SCIENCE MINOR PROJECT REPORT**

CONTENTS OF THE REPORT

* Cover page
* Declaration
* Certificate
* Acknowledgement
* Table of Content

1. Introduction
2. Source of dataset
3. EDA process
4. Analysis on dataset (for each analysis)
   1. Introduction
   2. General Description
   3. Specific Requirements, functions and formulas
   4. Analysis results
   5. Visualization
5. Conclusion
6. Future scope
7. References

**Other Conventions**

i.

ii.

iii. iv.

v.

vi. vii. viii. ix.

x.

Please note the case of letters in the cover page: The 3rd line is 16 pt bold and other lines are 12 pt. The page is centred. Department and Institute names are bold.

All the matter contained in the report should be typed in MS word (1.5 spacing) Times New Roman, 12 pt or equivalent with other software.

Figures and tables may be inserted in the text as they appear or may be appended in order. Table of Content shall be in well hyperlinked

List of figures and tables shall be maintained with captions in MS word. List of references shall be appended at the end.

References shall be in IEEE format

Total Number of pages with A4 size paper shall be minimum 30 pages and maximum 80 pages. Hard copy of report must be available with each student on the day of evaluation.

In addition to Hard copy of reports e-copy shall also be submitted. An e-copy of the report shall be

submitted by the student to respective teacher on their emails.

**INT 375: PYTHON PROGRAMMING PROJECT REPORT**

**(Project Semester January-April 2025)**

***(“Netflix Dataset of 2024 Users”)***

**Submitted by: Nareddy Vinuthna**

**Registration No: 12304062**

Programme and Section: B.TECH(CSE), K23GS

Course Code: INT375

Under the Guidance of

**Mam Gargi Sharma**

Discipline of CSE/IT

Lovely School of Computer Science Engineering

Lovely Professional University, Phagwara

**CERTIFICATE**



This is to certify that I **Nareddy Vinuthna** bearing Registration no. **12304062** has completed INT375 project titled, “Netflix Dataset of 2024 Users” under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

Signature and Name of the Supervisor Designation of the Supervisor

School of Computer Science Lovely Professional University Phagwara, Punjab.

Date: 11 April, 2025

**DECLARATION**

I, Nareddy Vinuthna student of B.Tech under CSE/IT Discipline at, Lovely

Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and

is genuine.

Date: 5 April, 2025 Registration No : 12304062

Signature:

Student’s Name: Nareddy Vinuthna

# □ Table of Content

1. Introduction
2. Source of dataset
3. EDA process
4. Analysis on dataset (for each analysis)
   1. Introduction
   2. General Description
   3. Specific Requirements, functions and formulas
   4. Analysis results
   5. Visualization
5. Conclusion
6. Future scope
7. References
8. Introduction

This project focuses on analyzing detailed election results across India with a goal to uncover insights into voting patterns, demographics, party performance, and voter behavior. Exploratory Data Analysis (EDA) is applied using Python tools such as NumPy, Pandas, Matplotlib, and Seaborn to study trends, relationships, and anomalies.

1. Source of Dataset

The dataset includes metadata about entertainment titles, covering information such as title names, main and sub-genres, release years, maturity ratings, original audio languages, and recommendation links. It offers a structured view of content distribution across genres and maturity levels. The data appears to be curated from streaming platform archives or media databases and is suitable for exploratory data analysis and visualization.

Dataset source link: <https://old.datahub.io/dataset>





The dataset used in this project is titled

‘Netflix Dataset 2024 Users’ was sourced from the Datahub



1. EDA Process (Exploratory Data Analysis)

This is the most technical part, where you showcase how the data was cleaned, structured, and summarized.

Steps Taken:

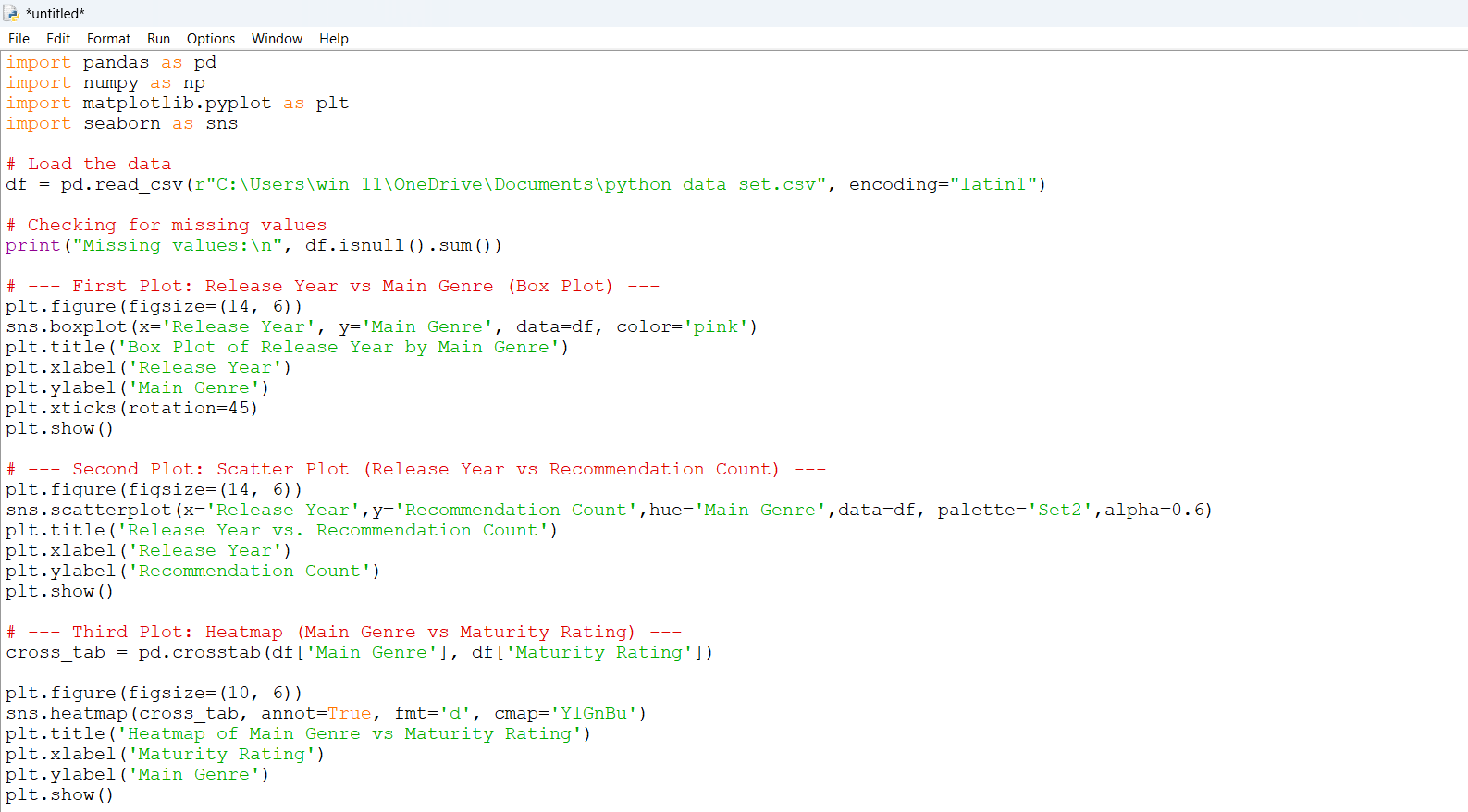
* + Import libraries: numpy, pandas, matplotlib.pyplot, seaborn
  + Read the dataset using pd.read\_csv()
  + Clean the dataset: drop empty rows, handle missing values, standardize column names
  + Descriptive statistics: mean, median, std, min-max for numeric fields
  + Prepare dataset for analysis with groupby(), merge(), melt(), etc. Purpose of EDA:
  + Understand the data structure and range
  + Detect anomalies or inconsistencies
  + Prepare the dataset for visualization and deeper analysis



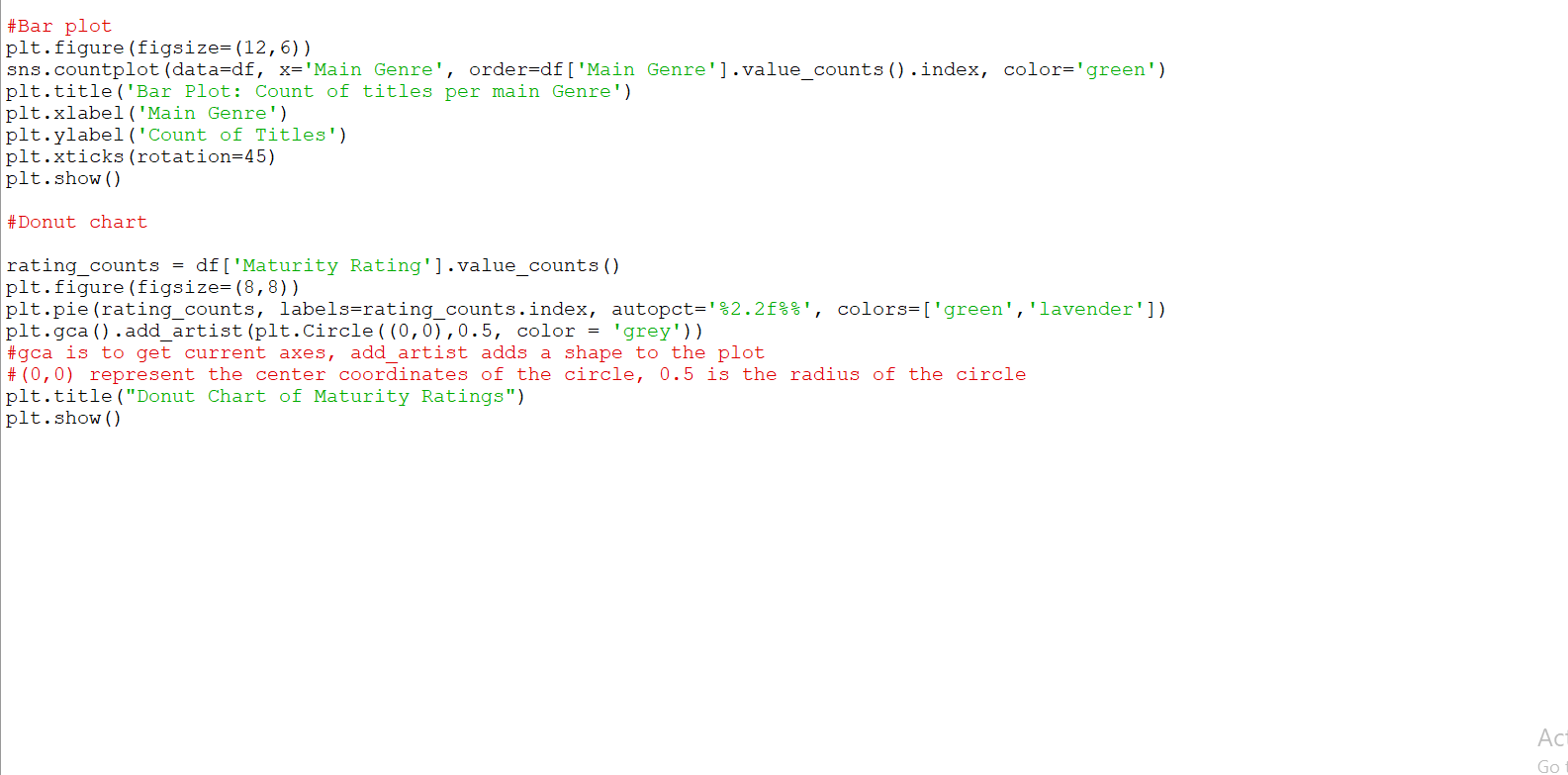
1. Analysis on Dataset

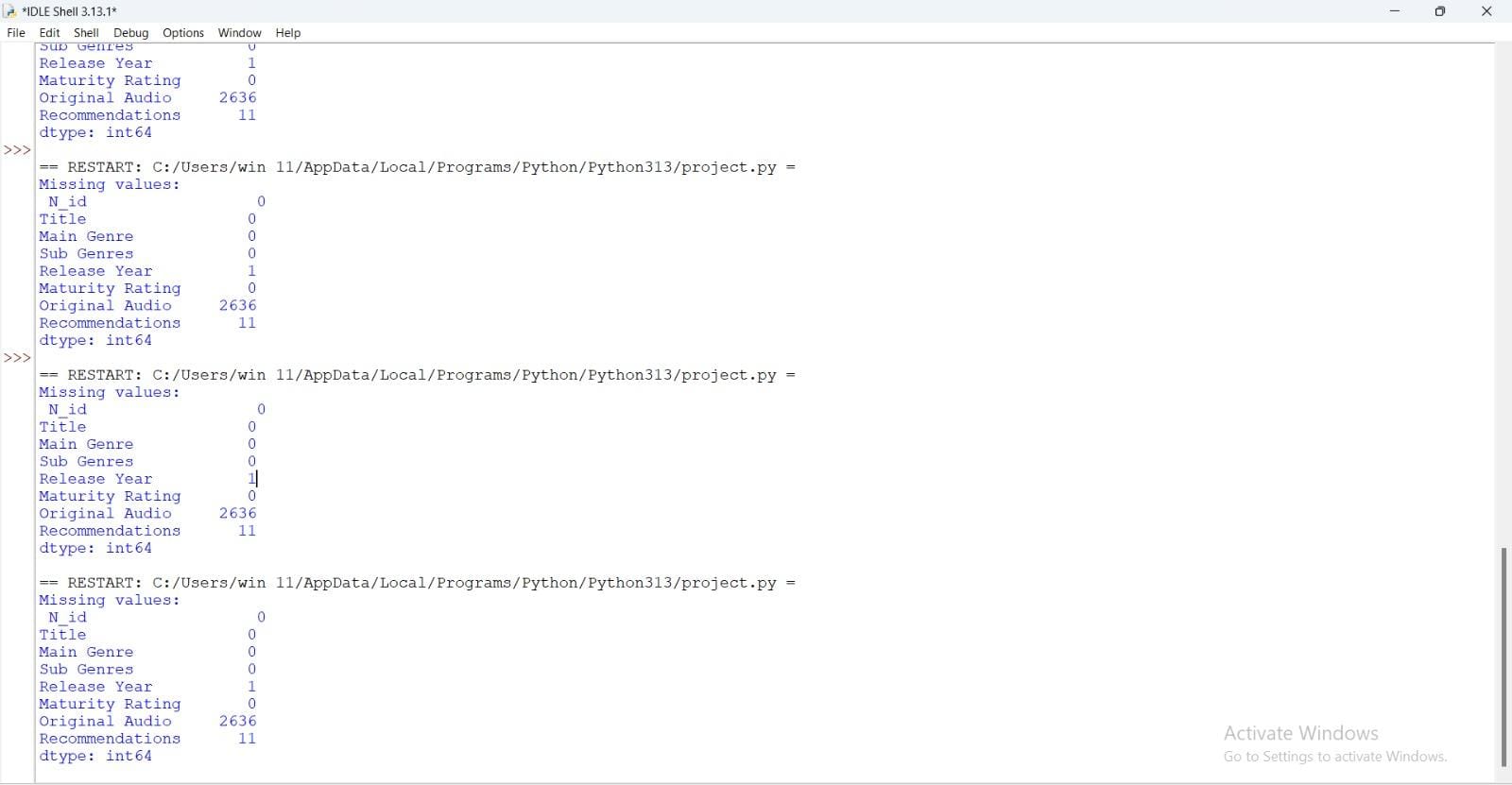
This part is broken into m u l t i p l e analytical question,seach with a detailed breakdown.







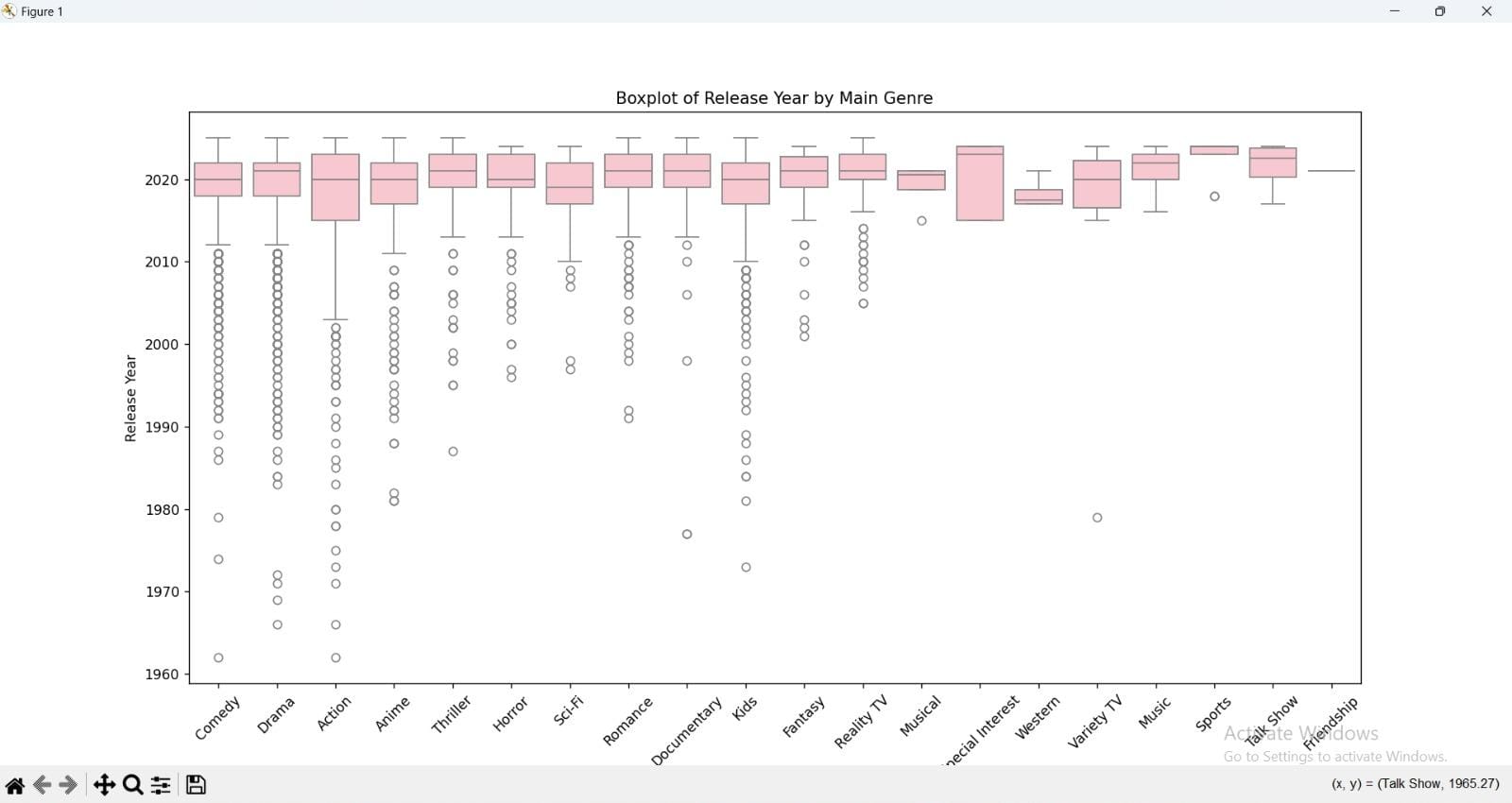




The first graph is BOX PLOT of Release Year by Main Genre

The X-axis is Genre

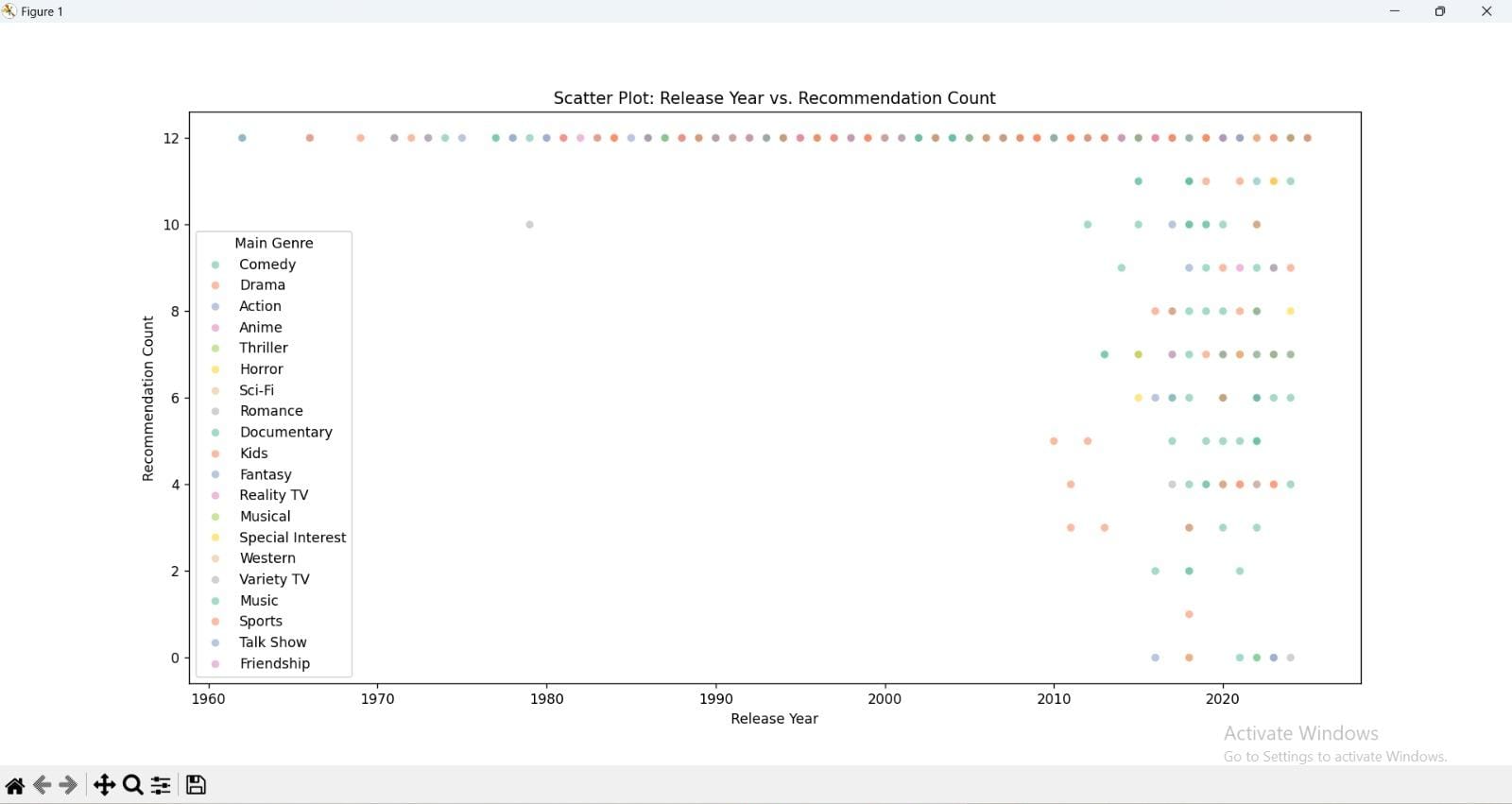
The Y-axis is Release Year



The second graph is Scatter Plot of Release Year vs. Recommendation Count

The X-axis is Release Year

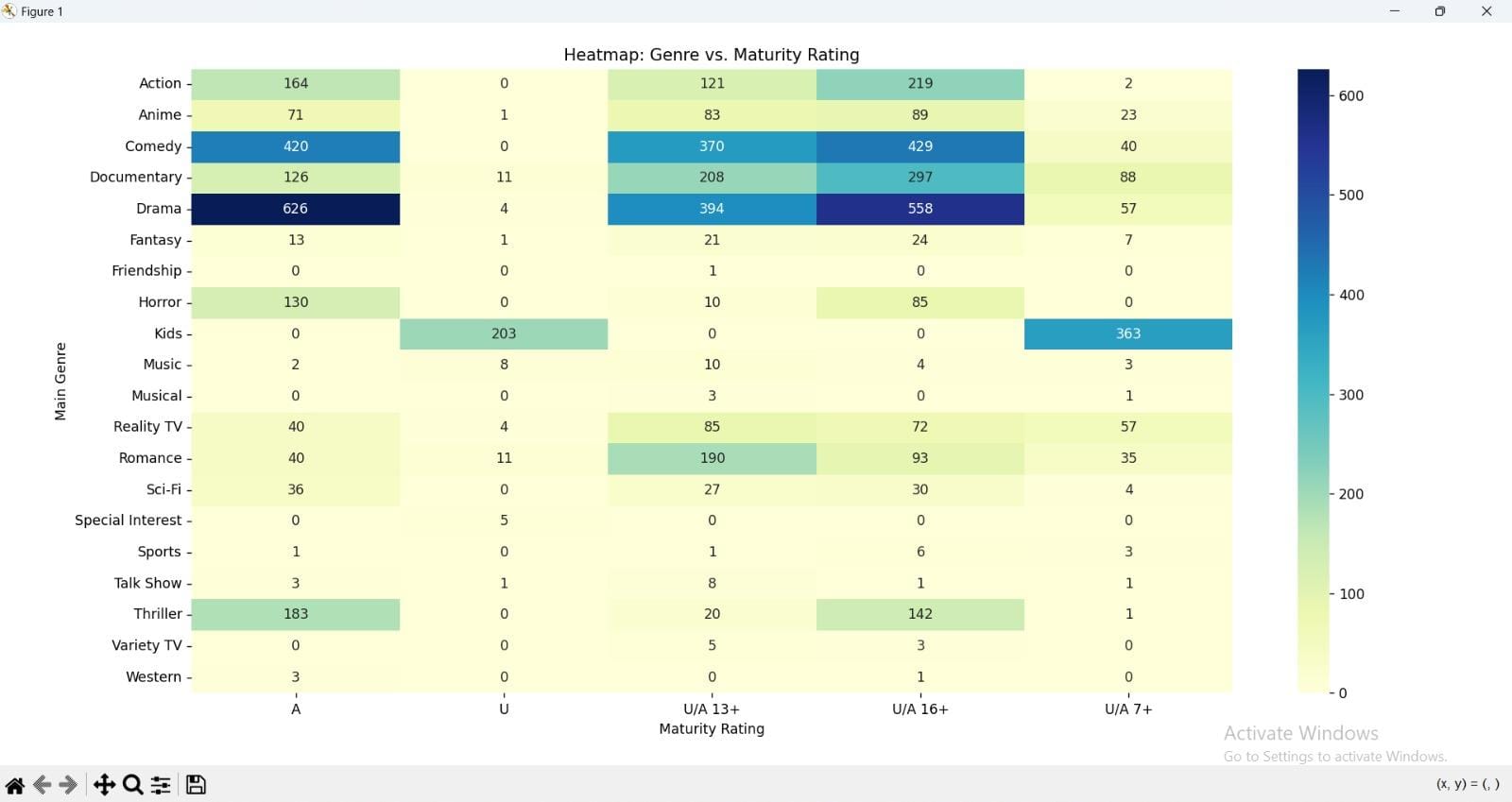
The Y-axis is The Recommendation Count



The third graph is HEAT MAP of Genre vs. Maturity Rating

The X-axis is Maturity Rating

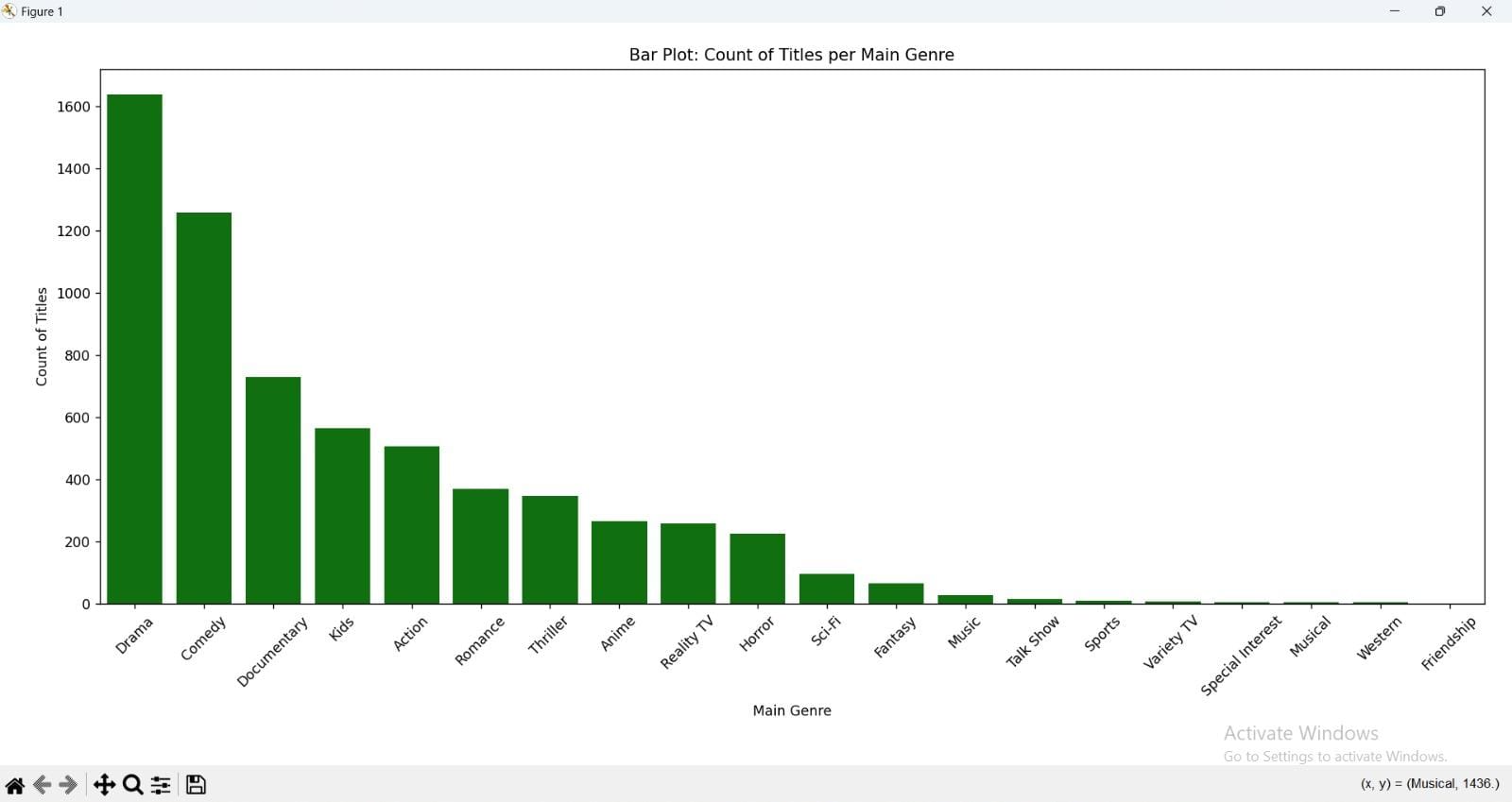
The Y-axis is Main Genre

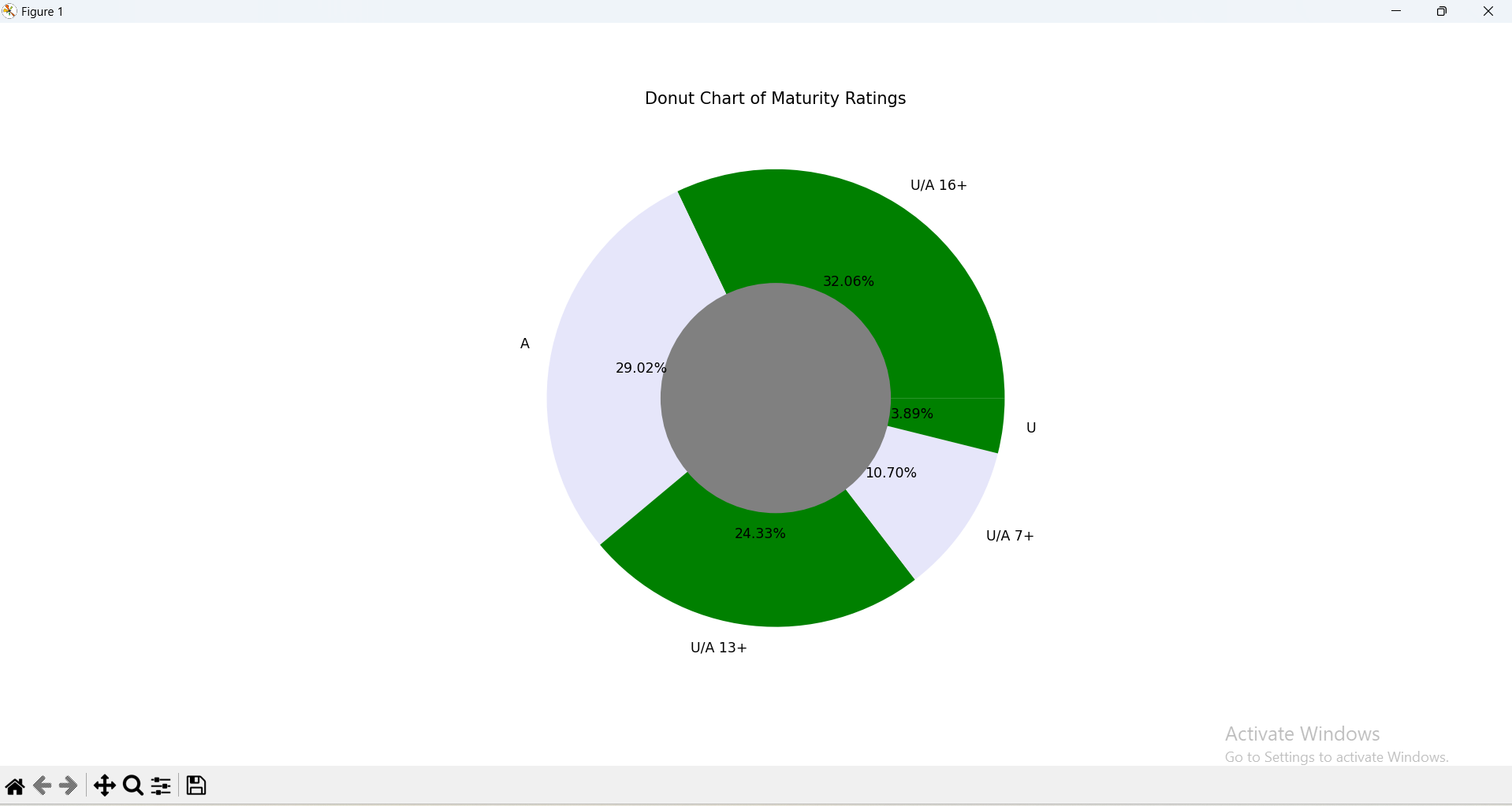


The fourth graph is BAR PLOT of Count of Titles per Main Genre

The X-axis is about Main Genre

The Y-axis is the Count of Titles



The fifth graph is Donut Chart about Maturity Ratings

**Key Observations**

**1. Genre-wise Distribution of Content**

From the bar plot representing the number of titles in each main genre, it is evident that certain genres dominate the platform. Genres like *Drama*, *Comedy*, and *Action* appear most frequently, suggesting these categories are the most popular or most produced. Niche genres such as *Documentary* or *Historical* have significantly fewer titles, indicating limited production or a smaller target audience.

**2. Temporal Trends in Content Release**

The boxplot of release years by main genre reveals distinct temporal patterns. For instance, *Classic* and *Romantic* genres tend to have a broader spread in release years, implying long-standing presence, while genres like *Sci-Fi* and *Thriller* show more recent production trends. The median release year for most genres lies between 2005 and 2015, highlighting a recent surge in digital content production, likely due to the rise of streaming platforms.

**3. Recommendations and Viewer Behavior**

The scatter plot between *Release Year* and *Recommendation Count* shows that newer titles often receive a higher number of recommendations. This may be attributed to better visibility, platform algorithms, or increased viewer engagement with recent content. Additionally, genres like *Action* and *Drama* tend to dominate higher recommendation counts, indicating strong viewer preferences or sharing behavior for these genres.

**4. Content Regulation and Audience Targeting**

The donut chart illustrating maturity ratings reveals that a majority of the content is rated either *13+* or *16+*, suggesting that platforms are targeting teenage and young adult demographics. Very few titles are rated for *All Ages* or strictly *18+*, indicating content regulation standards or a strategic focus on moderately mature audiences to maximize reach.

**5. Genre and Maturity Correlation**

The heatmap comparing genres and maturity ratings provides insights into content tone and audience targeting per genre. For example, *Comedy* and *Drama* appear across multiple rating categories, showing their wide appeal. On the other hand, *Horror* and *Thriller* are mostly concentrated in higher maturity brackets, reflecting the nature of their content. *Animation*, interestingly, is spread across both lower and higher ratings, likely due to the mix of children’s content and adult-themed animations.

**6. Multilingual and Global Nature of Content**

The dataset includes a column for *Original Audio*, which—although not visualized here—can be analyzed to highlight the linguistic diversity in content. A deeper look might show a wide range of original languages, reinforcing the idea that streaming platforms host international content and cater to multilingual audiences.

**7. Data Quality Observations**

During preprocessing, some missing values were noted, especially in the *Release Year* and *Recommendations* columns. This highlights either incomplete metadata collection or data loss during curation. Handling these missing values was necessary to ensure the accuracy of visualizations and insights.

**8. Insights for Business Strategy**

These patterns can inform platform strategies for content acquisition, recommendation algorithms, and audience engagement. For instance, increasing the visibility of older but highly recommended titles, or producing more content in high-performing genres like *Drama* or *Thriller*, could improve viewer satisfaction and retention.

**🔮 Future Scope**

**1. Incorporating Viewer Engagement Metrics**

Future versions of this dataset can be enriched with user engagement metrics such as view counts, likes, average watch time, or user ratings. This would enable more accurate and personalized analysis of content performance and audience preferences, allowing for better recommendation system tuning and marketing strategies.

**2. Sentiment and Review Analysis**

If user reviews or social media comments are available, applying Natural Language Processing (NLP) techniques can help understand audience sentiment toward various genres, titles, or maturity ratings. Sentiment trends over time could also provide deeper insights into shifting viewer interests.

**3. Time Series Forecasting**

Analyzing historical release trends across genres and regions can allow forecasting of future content production patterns. Time series models could help platforms predict peak production periods, emerging genres, and changes in maturity rating distributions.

**4. Regional and Language-Based Segmentation**

By including regional metadata and analyzing original audio/language fields, one could explore regional content preferences and linguistic diversity. This would help platforms localize their content strategies for different countries or cultural groups.

**5. Recommendation System Modeling**

The "Recommendations" column opens up opportunities to build and evaluate basic recommendation algorithms. Collaborative filtering or content-based models could be trained using recommendation count, genres, and maturity ratings to predict similar titles or improve discovery.

**6. Visual Analytics Dashboard**

A future improvement could be creating an interactive visual dashboard using tools like Plotly, Dash, or Power BI. This would allow dynamic filtering, genre comparisons, and real-time data exploration, making insights more accessible to non-technical users or stakeholders.

**7. Cross-Platform Content Analysis**

If data from multiple streaming services is combined, comparative analysis across platforms could reveal exclusivity trends, genre specialization, and competitive content strategies. This would be especially useful in market research and media consultancy.

**8. Machine Learning Applications**

The dataset can be used to train classification models (e.g., predicting maturity rating based on genre and audio), clustering algorithms (e.g., grouping titles based on similarity), or anomaly detection (e.g., spotting outliers in recommendation patterns or release years).

Reference :

**📚 References**

1. **Pandas Documentation**  
   *Pandas: Python Data Analysis Library* – https://pandas.pydata.org/
2. **Matplotlib Documentation**  
   *Matplotlib: Visualization with Python* – <https://matplotlib.org/>
3. **Seaborn Documentation**  
   *Seaborn: Statistical Data Visualization* – https://seaborn.pydata.org/
4. **Python Official Documentation**  
   *Python Language Reference, version 3.x* – <https://docs.python.org/3/>



THANK YOU





