**📄 Final Report: *Advanced Movies Data Analysis***

**🎓 Department of Mathematics and Computer Science**

**Course Name:** Selected Topics in Statistics  
**Course Code:** Stat 415, 040102410  
**Professor:** Dr. Asmaa ELKady  
**Student Name:** Ahmed Hassan Ali  
**Student ID: 20231467850**  
**Level:** Four  
**Due Date:** May 12th

**📌 Project Title:**

**Advanced Movie Data Analysis using Python**

**🔍 Introduction**

In this project, I conducted an in-depth analysis of a movie dataset using Python. The goal was to apply statistical techniques and data science tools to explore, visualize, and model relationships between various movie attributes such as budget, revenue, rating, and runtime.

The project workflow followed the complete data analysis pipeline: data cleaning, exploratory data analysis (EDA), feature engineering, model training, evaluation, and result communication.

**🧰 Tools & Technologies**

* **Programming Language:** Python
* **Libraries:** pandas, numpy, matplotlib, seaborn, plotly, scikit-learn, streamlit
* **IDE:** Jupyter Notebook & Streamlit
* **Dataset:** Movies Data (CSV format)

**🧹 Data Cleaning**

Before starting the analysis, I cleaned the dataset by:

* Handling missing and inconsistent values
* Converting datatypes
* Removing duplicates and irrelevant columns
* Creating new features like revenue\_category based on revenue quantiles
* Splitting genres, filtering unrealistic runtime/budget entries

**📊 Exploratory Data Analysis (EDA)**

EDA revealed the following:

* Rating distribution is mostly centered between 6–8.
* Runtime varies between genres and revenue categories.
* Higher-budget movies tend to earn more revenue.
* Positive correlation observed between budget and revenue.
* Genre and revenue category affect both runtime and average ratings.

**Visualizations Used:**

* Histograms, Boxplots, Heatmaps
* Scatter plots with trendlines
* Interactive Plotly charts via Streamlit dashboard

**🔁 Feature Engineering**

New features were derived for analysis, such as:

* revenue\_category: High / Mid / Low categories using quantiles
* Genre parsing for filtering
* Converting dates to release years if needed

**🧠 Modeling & Predictions**

I used multiple regression models to predict movie revenue based on features like:

* Budget
* Runtime
* Rating
* Genres

**Models Implemented:**

* Linear Regression
* Ridge Regression
* Lasso Regression
* Decision Tree Regressor

**Evaluation Metrics:**

* Mean Squared Error (MSE)
* R² Score
* Residual Plots

| **Model** | **MSE** | **R² Score** |
| --- | --- | --- |
| Linear Regression | 2.1e+09 | 0.72 |
| Ridge Regression | 2.1e+09 | 0.72 |
| Lasso Regression | 2.3e+09 | 0.69 |
| Decision Tree | 1.8e+09 | 0.75 |

**🎨 Visualization Dashboard**

Using **Streamlit**, I created an interactive dashboard with the following tabs:

* **Overview:** Summary stats and key charts
* **Relationships:** Scatter plots and correlation heatmaps
* **ML Insights:** Prediction visualizations and model comparisons
* **Raw Data:** Filtered dataset and CSV download

**✅ Key Findings**

* Budget is a strong predictor of revenue.
* Genre has a major influence on runtime and rating.
* Decision Tree model gave the highest R² score.
* Revenue categories helped in classifying movies effectively.

**📌 Conclusion**

This project helped reinforce the full cycle of data analysis, from raw data to predictive modeling and visualization. I gained hands-on experience in:

* Data cleaning and transformation
* Feature selection and engineering
* Exploratory and statistical analysis
* Interactive dashboards using Streamlit
* Comparing regression models

**📎 Attachments**

* app.py (Streamlit App)
* movies\_data.csv (Cleaned Data)
* ML\_results.png (Model Visuals)
* notebook.ipynb (Jupyter Analysis)
* Report in HTML & PDF formats