1. **Introduction- Patrick (3 mins)**

We chose movies as our subject of exploration for the machine learning project & final. Imagine yourself as a production office executive and needing information on new releases to determine the gross amount of revenue your company would make for a particular movie. This is the crystal ball that we created with our project. It is also fun to explore new releases as we head into the summer blockbuster season. The type of regression models we consider using include: Linear Regression, Lasso, and Ridge Regression. We will also have a dashboard with a user interface that allows users to enter details regarding a pending release and the machine learning model will predict the movie's gross revenue. Now to talk more about our data exploration phase, here is Kylie Hicks.

1. **Data Exploration, Database, & Cleaning- Kylie (3 mins)**

The main movie dataset was found on [Kaggle](https://www.kaggle.com/datasets/danielgrijalvas/movies). The dataset was originally scraped from the IMDB website. IMBD is an online database that contains information related to movies, TV series, video games, as well as streaming content. This particular dataset caught our eyes because it is robust and we could seemingly draw direct conclusions from linear regression models.

We obtained additional datasets containing the birthdates, number of nominations, and number of awards won by the lead actor/actress listed for each movie in the movie dataset. These secondary datasets were obtained by webscraping Wikipedia pages. Wikipedia is an online encyclopedia allowing free public access. Information provided by volunteers and contributors through open collaboration. The dataset containing the birthdates was obtained by a script that visited each actors' individual Wikipedia page. The second dataset, containing the Academy Award nominations and awards won for leading and supporting roles since 1927, was obtained by scraping [Academy Award Nominations](https://en.wikipedia.org/wiki/List_of_actors_with_Academy_Award_nominations#List_of_actors).

Initial cleaning of the movie dataset from Kaggle included removing null values. The released column contained both the date of the movie's release and the country of release. To obtain the release date only, we used the str.replace method and RegEx, then converted the date to the month only using pandas dt.month. Since the dataset contains monetary information in the budget and gross revenue columns spanning movies released from 1980 to 2020, we wanted to account for inflation. To do this, we used the cpi.inflate from a python library that adjusts U.S. dollars for inflation using the Consumer Price Index.

When pulling data tables from our database, we merged the movies table with the released\_dayofweek to obtain the day of week each movie was released. Additionally, we merged the actors\_bday with the actor\_awards table to essentially create a profile for each movie's starring actor. Since the information scraped from Wikipedia's Academy Awards included all actors who received a nomination or award since 1921, any actor who did not have a value in the actor\_awards table is reasoned to have neither been nominated nor awarded and their values were filled with 0's accordingly. These two dataframes were then merged to create the final dataset including the cleaned information for movies and actors. Upon investigation, we removed movies with the following ratings: NC-17, Not Rated, TV-MA, Unrated, and X. These movies made up a small portion of the sample and were likely to skew our results since we were focusing on theater releases. Since some columns had numerous unique values, to prepare for the machine learning, we binned the following data: genre, director, writer, star, and company. Additionally, the purpose of obtaining the birthdate of each actor was to find their age at the time of the movie's release. We used their birthdate and the release date to find the age. The following information in our data were categorical: rating, genre, director, writer, star, company, released\_month, and weekday. We used Sklearn's OneHotEncoder to encode these data columns so we could feed them to our model.

We have chosen to use a PostgreSQL database for this project that will be created and maintained locally. Our database contains four tables: movies, actor\_awards, actor\_bday, and released\_dayofweek. The movies table contains the cleaned dataset obtained from Kaggle. The actor\_awards table consists of the number of Academy Award nominations and awards. This table can be merged with the movies table to determine the prominence of the movie's leading actor. The actor\_bdays table contains the actors' birthdates, this is used to determine the age of the leading actor when the movie was released. Lastly, the released\_dayofweek table contains the day of week that each movie from the movies table was released.

1. **Machine Learning Model- David (4 mins)**
   1. **Describe linear regression**
   2. **Explain Ridge model**
   3. **Talk about limitations of the model**
   4. **Share results/accuracy score**

1. **Conclusion & Demo- Kaiya (5 mins)**
2. Describe Dash Plotly
3. Model demo:
   1. Avatar 2
   2. Top Gun