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)**

Kaggle Dataset

1. **Machine Learning Model- David (4 mins)**

We chose to use a Linear Regression model for our analysis because the gross income variable we were targeting is a continuous variable, meaning it cannot be sorted into categories. We trained and tested our data on three different linear regression models, and the model that performed the best was the Ridge Regression model. Ridge regression is a linear regression model that better estimates variable coefficients when multiple variables are highly correlated, but still independent of each other.

There are limits to a linear regression model. First, these models are heavily influenced by the presence of outliers in a dataset, and this can impact the fit of the regression line as well as the final accuracy score. Also, linear regression assumes that there is a linear relationship between the target variable and each independent feature variable, and if that is not the case, it can lower the accuracy of the model.

When analyzing our dataset, we noticed that our target variable, gross income, had a strong positive skew, meaning most of the data concentrated around a smaller value while the higher values trail off in a “tail” towards the positive axis. In order for our Ridge Regression model to be accurate, we had to transform our target variable to give it a more “normal” distribution. We did this by using the square root of the gross income while training and testing our linear regression model.

Our model produced an r2 score of 0.58 and a mean squared error of approximately 22 million. The r2 score is commonly used to describe the accuracy of a linear regression model, and it shows how well a linear model fits a set of data. In this case, our model shows 58% of the total variance in the dependent variable was predictable from the independent variables. The mean squared error describes how close, on average, our regression line is to the original target variables. While an error of 22 million may seem alarming, the gross income variable that is being measured is commonly in the hundred-million range.

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