UDACITY DATA VISUALIZATION NANODEGREE

Mid Term Project

Kay Sun

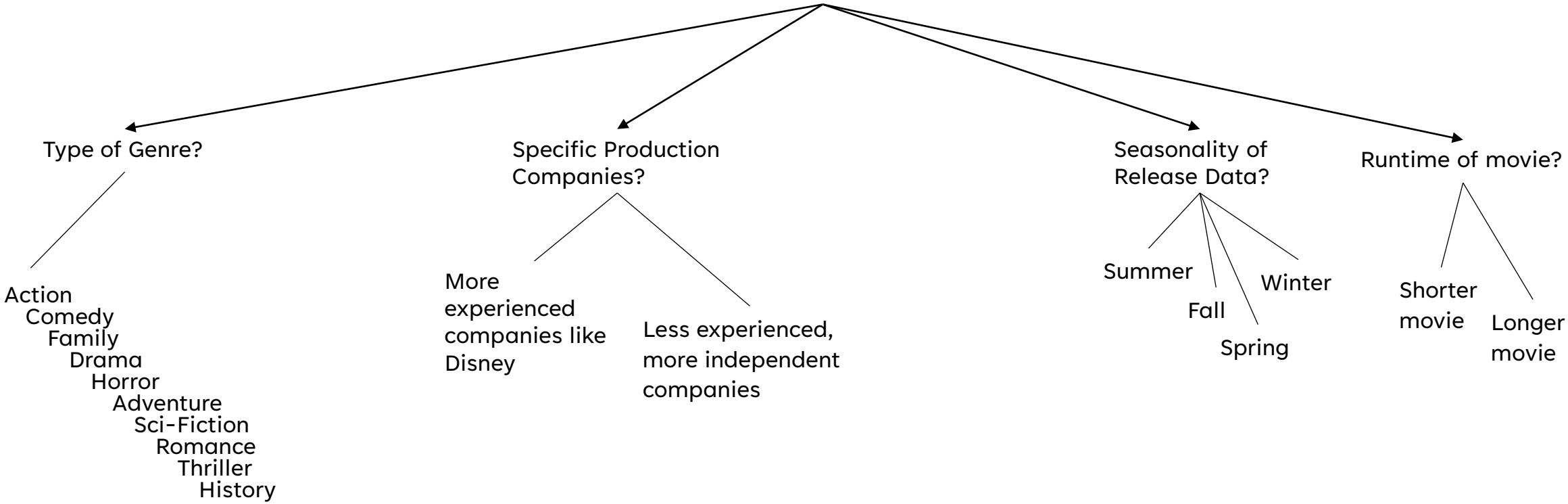
October 20 2023

PROBLEM STATEMENT

What are the **key movie features** the **top 5% movies** have compared to the **bottom 95% movies** when considering **profits**?

ISSUE TREE

What are the **key movie features** the **top 5% movies** have compared to the **bottom 95% movies** when considering **profits**?



SYNTHESIS

GENRE

Drama, Comedy, and Romance are most common genres to both Top and Bottom movies.

PRODUCTION COMPANY

Productive production companies make many movies.

Some are top grossing or most are not. Hit or miss.

RELEASE DATE

Q2 and Q4 are most profitable for Top Movies.

All quarters are loss for Bottom Movies

MOVIE RUNTIME

Median runtime of 100 mins similar for Top and Bottom Movies.

ANALYSIS OVERVIEW

What drives the profitability for the Top 5% of movies?

Profit = Revenue - Budget

GENRE

Compare profits by genres for Top and Bottom Movies to find the most common genres and most/least profitable.

PRODUCTION COMPANY

Compare profits by production companies for Top and Bottom Movies to find the most common genres and most/least profitable.

RELEASE DATE

Compare profits by quarter for Top and Bottom Movies to find the most profitable release date.

MOVIE RUNTIME

Compare profits by runtime for Top and Bottom Movies to find any correlation.

LIMITATIONS AND BIASES

Limitation

- Raw data is messy, have to be parsed to separate out like genre and production companies.

Data Collection

- Popularity and voter ratings were sourced from 1 website – GroupLens.
- Potential errors as dataset was assembled as part of an education coursework and not an official release with quality checks.

Data Processing

- Missing data for some fields.
- Mislabel data for some fields.

Data Insights

Profit defined by Revenue – Budget may not be present the P/E relationship. There could be extraneous factors like tax incentives, insurance payouts, etc.

NEXT STEPS

1. Perform actual data analysis.
2. Impute missing values
3. Remove outliers.
4. Expand to include other datasets.
5. Expand to include analysis of other features and multi-variate analysis for combination of features.