Project: Visualizing Movie Data

Step 1: Data Cleanup and Attribute Selection

- Clean up any missing information and choose the most important attributes you will explore further in your visualizations.

- List out the attributes (or variables) you plan to dive further with your visualizations. You should explore no more than 8 attributes.

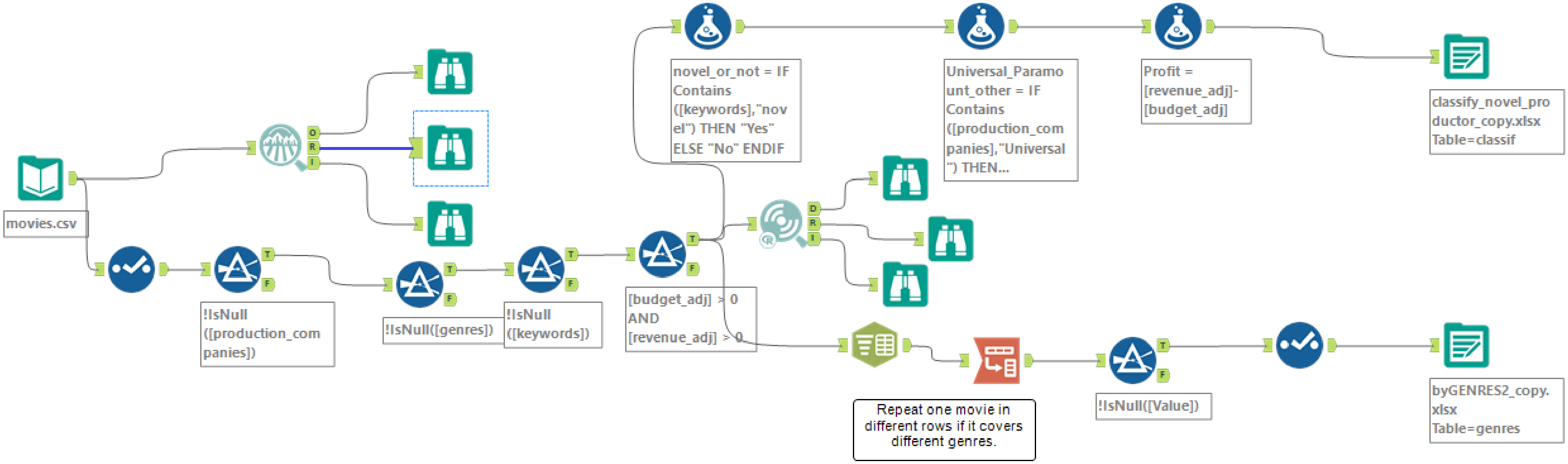
  The following attributes were selected to explore:
  - popularity
  - production companies
  - keywords
  - genres
  - budget_adj
  - revenue_adj
  - release date
  - and the calculation of profit as revenue_adj minus budge_adj.

  Adj means adjusted by inflation, in 2010 US dollars.

  Data was cleaned in Alteryx with the workflow shown in Figure 1.
  - First, data in the movies.csv file **Field Summary** to know which attributes had missing values, highlighting keywords, was explored with 13.7% of missing values and production companies with 9.5%.
  - Secondly, with the **Select** tool numeric attributes were changed to int16 or float types. Concretely, population, budget, budget_adj, revenue, revenue_adj and runtime were changed to float, whereas id, id_imbd and released_year were changed to int16.
  - Several **Filter** tools were concatenated to only get the records without null values in production_companies, genres and keywords, and selecting the movies with a budget_adj and revenue_adj greater than zero.
  - At this point, a **Formula** tool was used to get "Yes" or "No" in terms of a movie based on a novel, according to the keywords.
  - Another **Formula** tool allowed to classify the movies into Universal", "Paramount" or "Other" production company.

- Then, Profit was calculated from revenue_adj and budget_adj with a **Formula** Tool.
- Finally, the data was saved in a csv file.

- On the other side, after the filters, a **Text to Column** and **Transpose** tools were placed to get one row for each movie with different genres, so that it has a different row for each genre. The genres were splitted by "|" in 5 terms so, some rows had a null value because the movie covered less than 5 genres. Afterwards, null values were removed, "Values" was renamed to "Genres_film" with **Select** and the data were stored in another csv file.



movies.csv

!IsNull([production_companies])

!IsNull([genres])

!IsNull([keywords])

[budget_adj] > 0 AND [revenue_adj] > 0

novel_or_not = IF Contains([keywords],"novel") THEN "Yes" ELSE "No" ENDIF

Universal_Paramount_other = IF Contains([production_companies],"Universal") THEN...

Profit = [revenue_adj]-[budget_adj]

classify_novel_productor_copy.xlsx Table=classif

Repeat one movie in different rows if it covers different genres.

!IsNull([Value])

byGENRES2_copy.xlsx Table=genres

Figure 1. Movie data cleaning workflow with Alteryx. The popularity, budget, revenue, profit, genres, production companies, novel-based keywords and release date from the movies without null values was saved in two csv files.

Step 2: Tableau Visualizations

- Please make sure you follow the rubric and include Tableau Dashboards, Stories, and the appropriate visualizations (small multiples, scatter plot, bar chart, etc..) your reviewer expects your visualizations to contain. Remember: You need one Dashboard for every question (Q1-Q4) and in addition, you also need one Story, pertaining to a question of your choosing.
- Attach your visualizations as Tableau Workbooks in a zip file along with this report.

IMPORTANT: Please upload the workbooks to **Tableau Public** to allow reviewers to access your workbooks. Note that simply saving your file as a ".twbx" is not enough to allow all reviewers to access. Instructions on how to do this.

The dashboards and story are published in Tableau Public. The links are provided in the next section. This is a link to download the workbook: https://public.tableau.com/workbooks/P5_UPerezRamirez.twb

Step 3: Questions

- Answer the following questions. Refer to your online visualizations to back up your answers:
  - **Question 1: How have movie genres changed over time?**
    In general, the number of movies in each genre has exponentially increased from 1984 as indicated in the top line plot of Dashboard Q1 (Figure 2). Drama, comedy and thriller are the genres most represented in the movies. From 1961 to 1967 animation genre highlighted in revenue, and 1973 was excellent for horror movies as shown the bottom figure in Dashboard Q1.

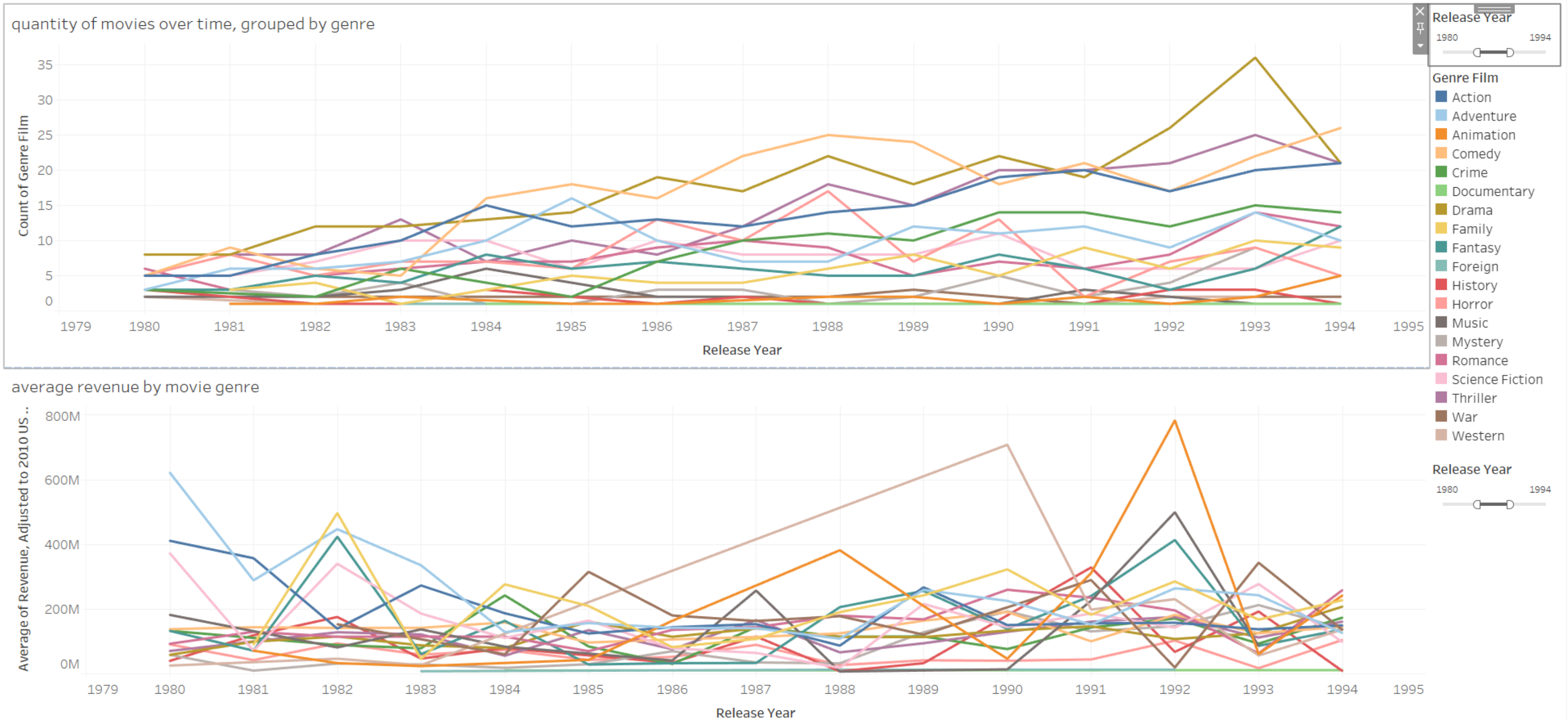


Figure 2. Dashboard Q1 to study the evolution of genres over time. Top: quantity of movies over years, grouped by genre. Bottom: average revenue, adjusted to 2010 US dollars, group by genders and over years.

- **Question 2: How do the attributes differ between Universal Pictures and Paramount Pictures?**
  Universal Pictures have generally worked better than Paramount Pictures over years in terms of profit and popularity. Dashboard Q2 and Figure 3 cover this question. In the top graphic we can see a red circumference, Jurassic Word, a successful movie from Universal with a great popularity-budget relationship and with a huge profit as indicated by the size of the circumference. Mad Max: Fury Road is another case of successful movie, from a company other than Paramount or Universal. Interstellar movie from Paramount was a remarkable movie as well. Over the years, other production companies summed a greater revenue and profit than Paramount and Universal. The trend was practically constant for Universal and Paramount companies, whereas probably more companies (other) have appeared from 1990. In terms of profit Universal (approx. 44 billions dollars) had a bigger profit than Paramount (around 38 billion dollars).

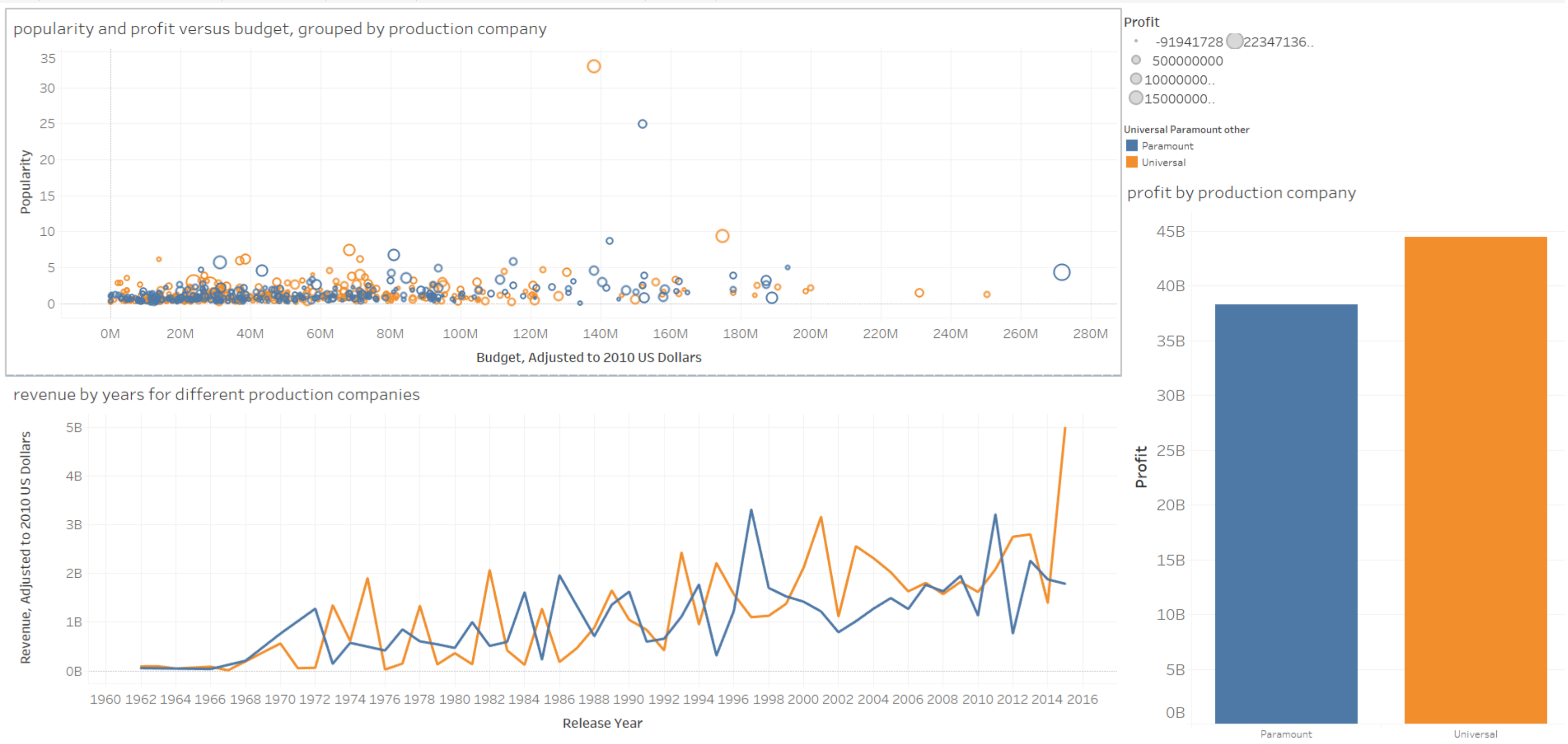
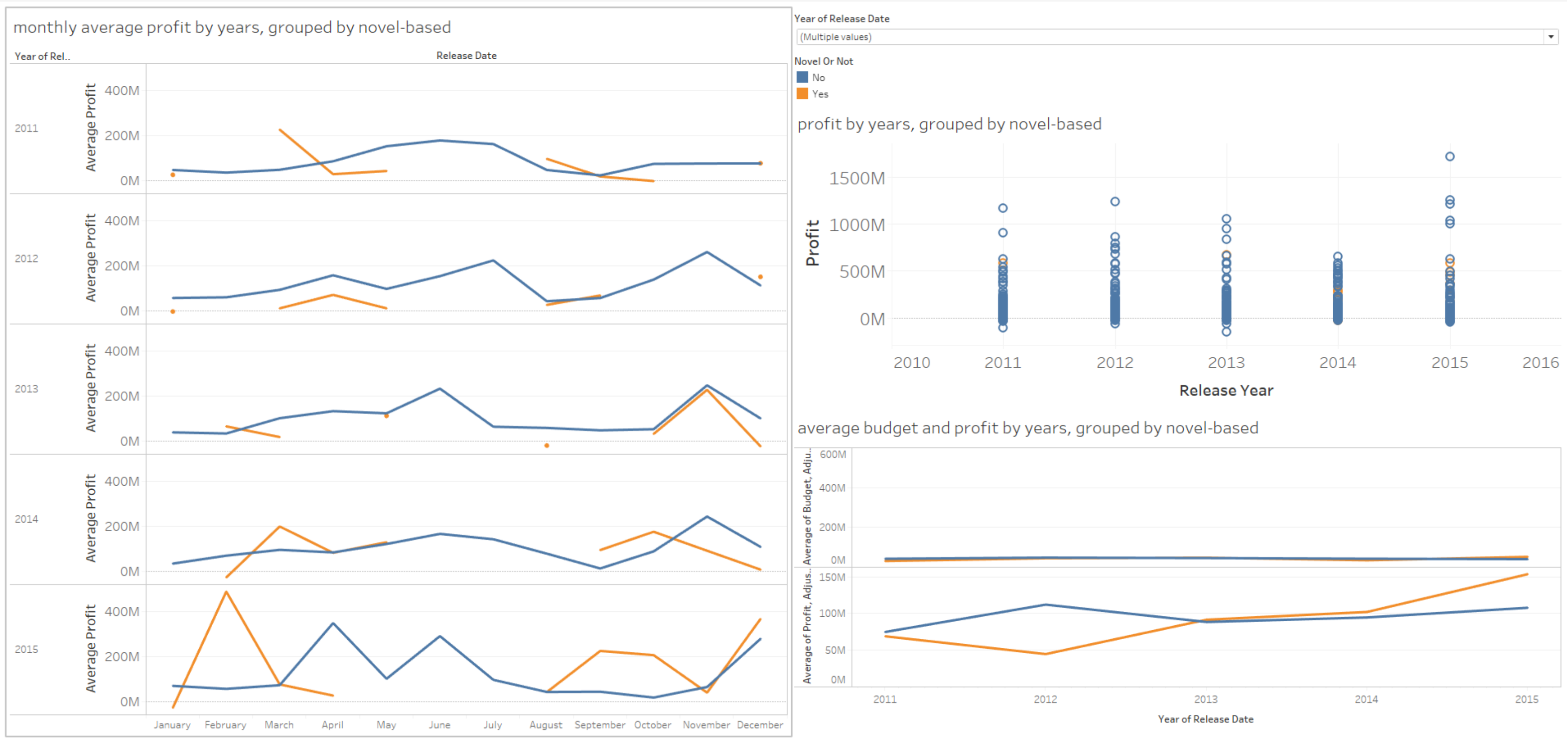


Figure 3. Dashboard Q2 to compare Universal and Paramount production companies. Blue: Paramount, orange: Universal. Top: popularity and profit versus budget, grouped by Universal or Paramount companies. Bottom left: revenue adjusted to 2010 US dollars, for Universal and Paramount companies. Bottom right: profit by production company.

- **Question 3: How have movies based on novels performed relative to movies not based on novels?**
  This is covered in Figure 4 and Dashboard Q3. My hypothesis was that movies based on novels are more widely known and therefore would generate more profit than the movies not based on novels, but that is not always the case, as reflected in the graphs. Perhaps it depends on how successful and well-known the novel is. I was expecting that the movies released in the summer and November-December (Christmas) would give more revenue but that is not always the case as shown in the left graph. On the right bottom graph we can see that the average budget and profit trends are very similar for both types of movies.


Year of Release Date: (Multiple values)

Novel Or Not: No, Yes

Figure 4. Dashboard Q3: performance comparison on novel-based movies versus movies not based on novels. Left graph: monthly average profit by years, grouped by novel-based. Top right: Profit by years. Bottom right: average budget and profit by years.

- What is your additional question that you proposed? What is the answer? How did you come up with this question?

  This question is addressed in Figure 5 and Dashboard Q4. The additional question that I proposed is which are the key ingredients for succeeding with a movie. The answer is that budget is not the crucial part as indicated in the area chart on the left side. It seems that good ideas and movie genre are crucial factors. The most successful genres are adventure, science fiction, fantasy, animation, action and family as indicated in the right bar chart.

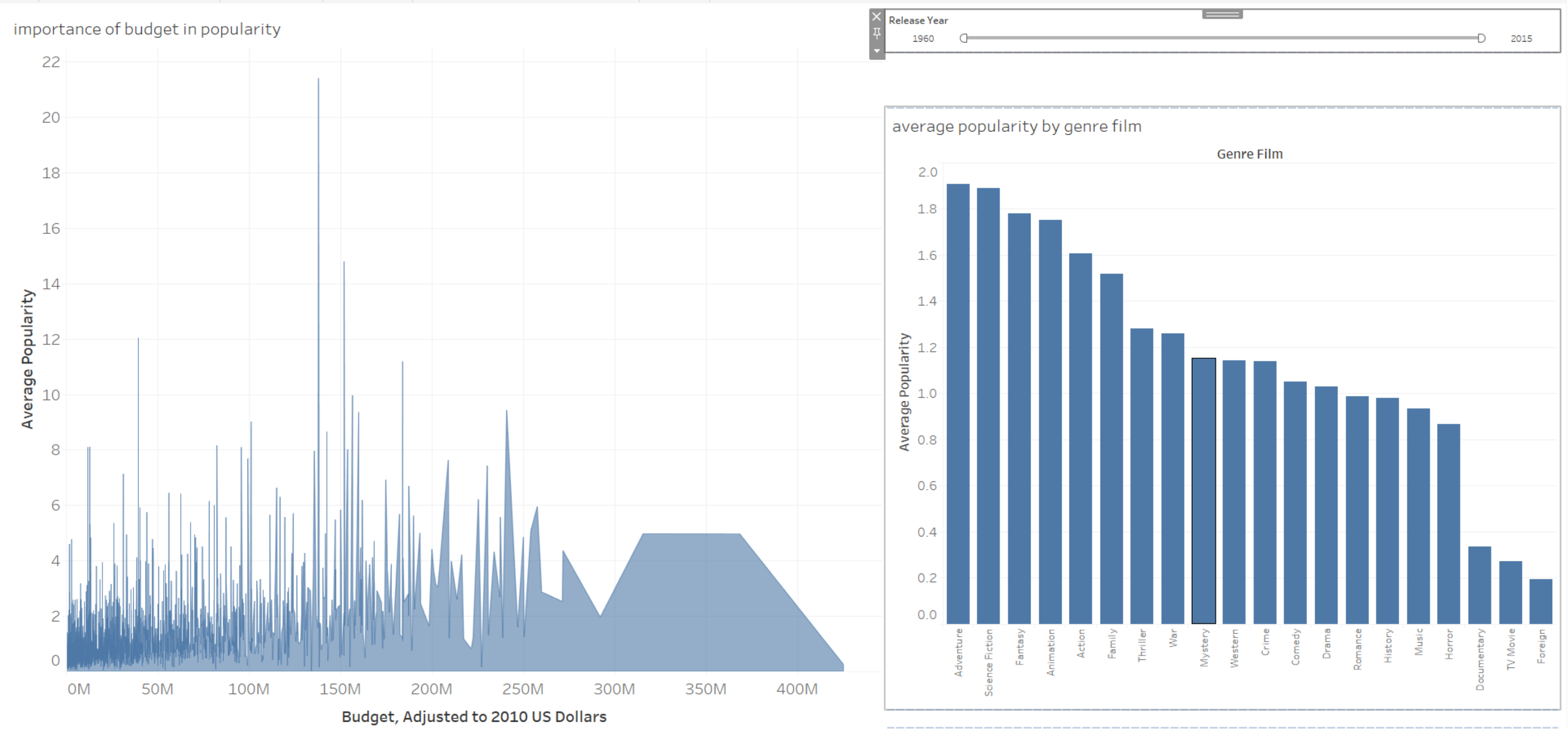
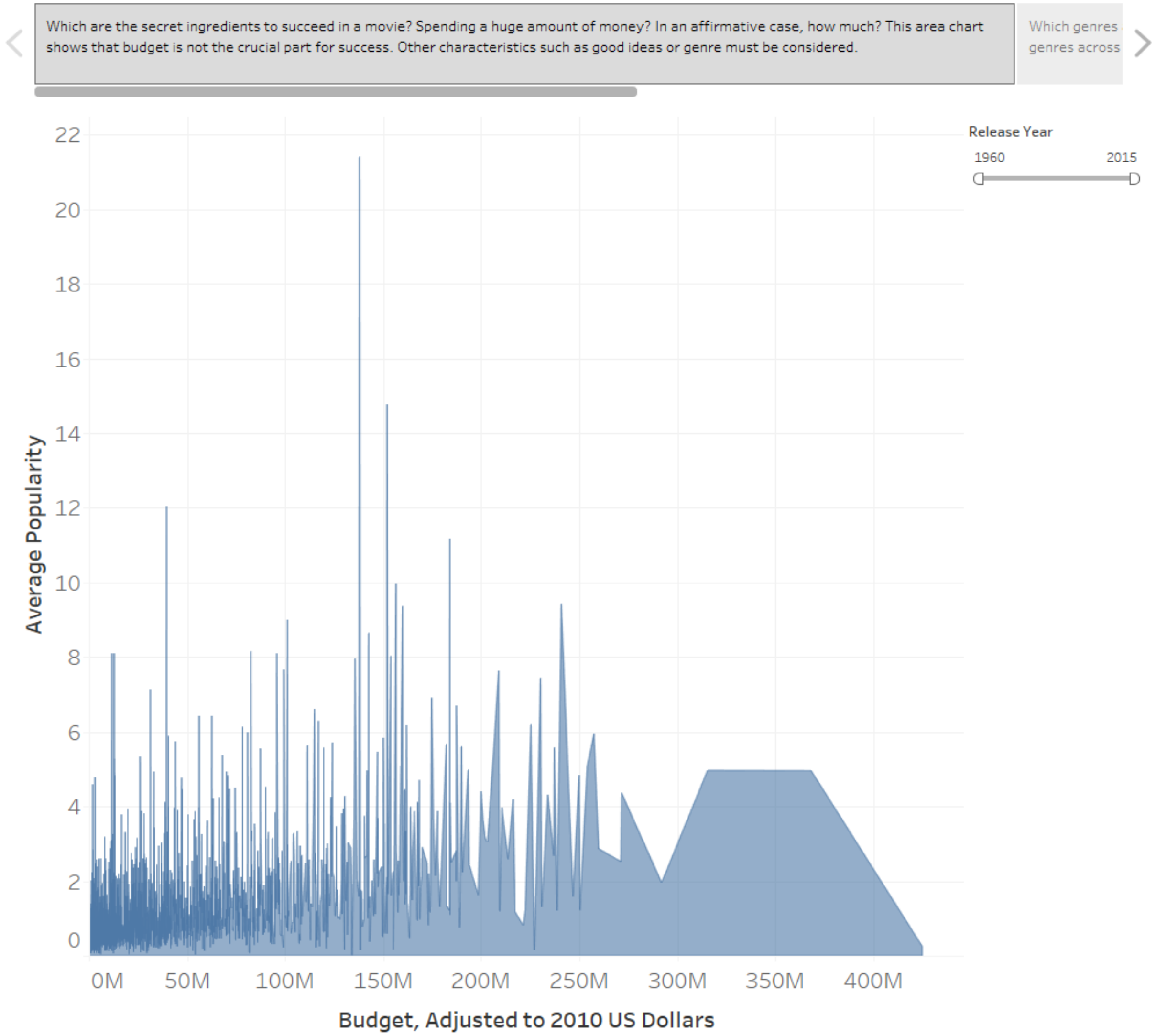


Figure 5. Dashboard Q4: crucial factor for movie success. Left: area chart of budget versus average popularity. Right: average popularity by genre film.

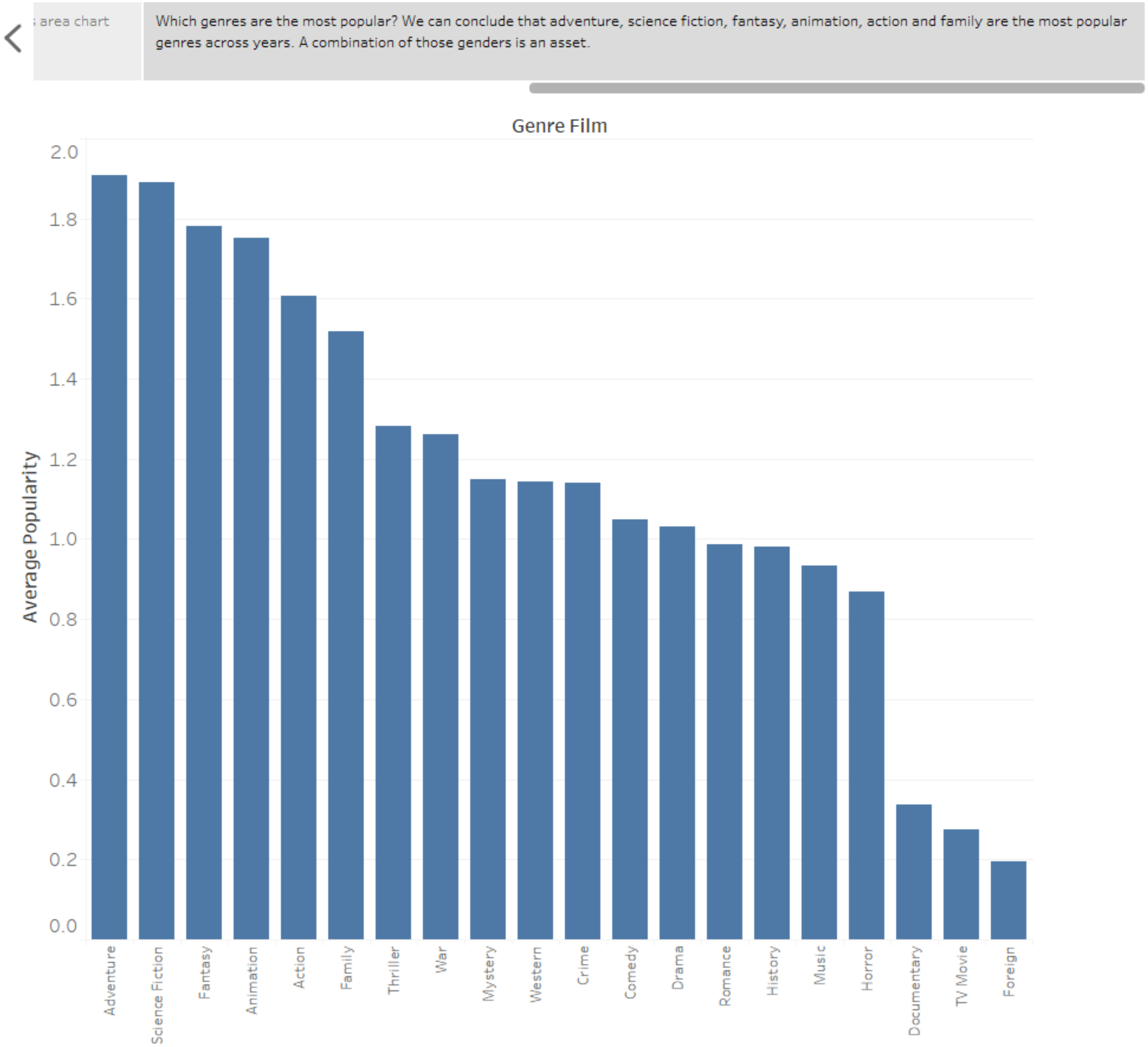
The story covers Q4, also shown in Figure 6.



Which are the secret ingredients to succeed in a movie? Spending a huge amount of money? In an affirmative case, how much? This area chart shows that budget is not the crucial part for success. Other characteristics such as good ideas or genre must be considered.

Which genres
genres across

Figure 6. Story about factors to succeed with a movie. Budget is not a key factor.



area chart

Which genres are the most popular? We can conclude that adventure, science fiction, fantasy, animation, action and family are the most popular genres across years. A combination of those genders is an asset.

Figure 6 continued. Story about factors to succeed with a movie. Genre is a key factor.