CSE 5544 Lab 2/4 (15/60 pts):

Exploratory Data Analysis

Assigned 2/7 Due 2/22 8:59pm ~~2/21 8:59pm~~

Learning Goal

Your last homework is about drawing basic elements and visual language (encoding, mark, attributes, items, channel) without any task specifications.

Please note that graduate students have more work than undergraduates.

In this assignment, we'll perform an exploratory analysis to better understand the shape & structure of the data, investigate

- (1) how to build a task taxonomy to begin with, and
- (2) to use visualizations can help answer initial questions, and
- (3) develop preliminary insights & hypotheses.

As a result, you will describe what your tasks, exploratory analysis, and observations are.

Your final submission will take the form of

- (1) a task list –
    - At least two elementary level for graduate students | one elementary for undergraduate students
    - At least two synoptic level tasks for graduate students | one elementary for undergraduate students
    - Must address at least two of the overall, partial, and individual items in your tasks.
- (2) Google colab python (for non-CS major) or D3 (mandatory for CS major) implementation;
    - Like the 1st assignment, there is not really right or wrong answer here. During the semester, you will get familiar with all kinds of visualization methods.
- (3) additional plots and a pdf file describing your observations to convey key insights gained during your analyses (graduate students ONLY)

Please note that graduate students have more work than undergraduate students.

Step 1: Data Analysis

In this assignment you will be working on some eye-tracking data - treat these eye-tracking data like a map. You will perform an **exploratory analysis** (location unknown and target unknown) of the dataset. In general, the best way to perform exploratory data analysis is through trial and error.

- Eye-tracking data we recently collected of pathologists' observing cancer images: https://docs.google.com/spreadsheets/d/1PlUVRGmUwNYKddBzyLGo-IwGIgqb8mWPWwWoWFIcfqk/edit?usp=sharing

Note: for the table columns not listed here, there is no need to consider them largely because they are perhaps not the first thing you would look at. So focus on the most important thing first.

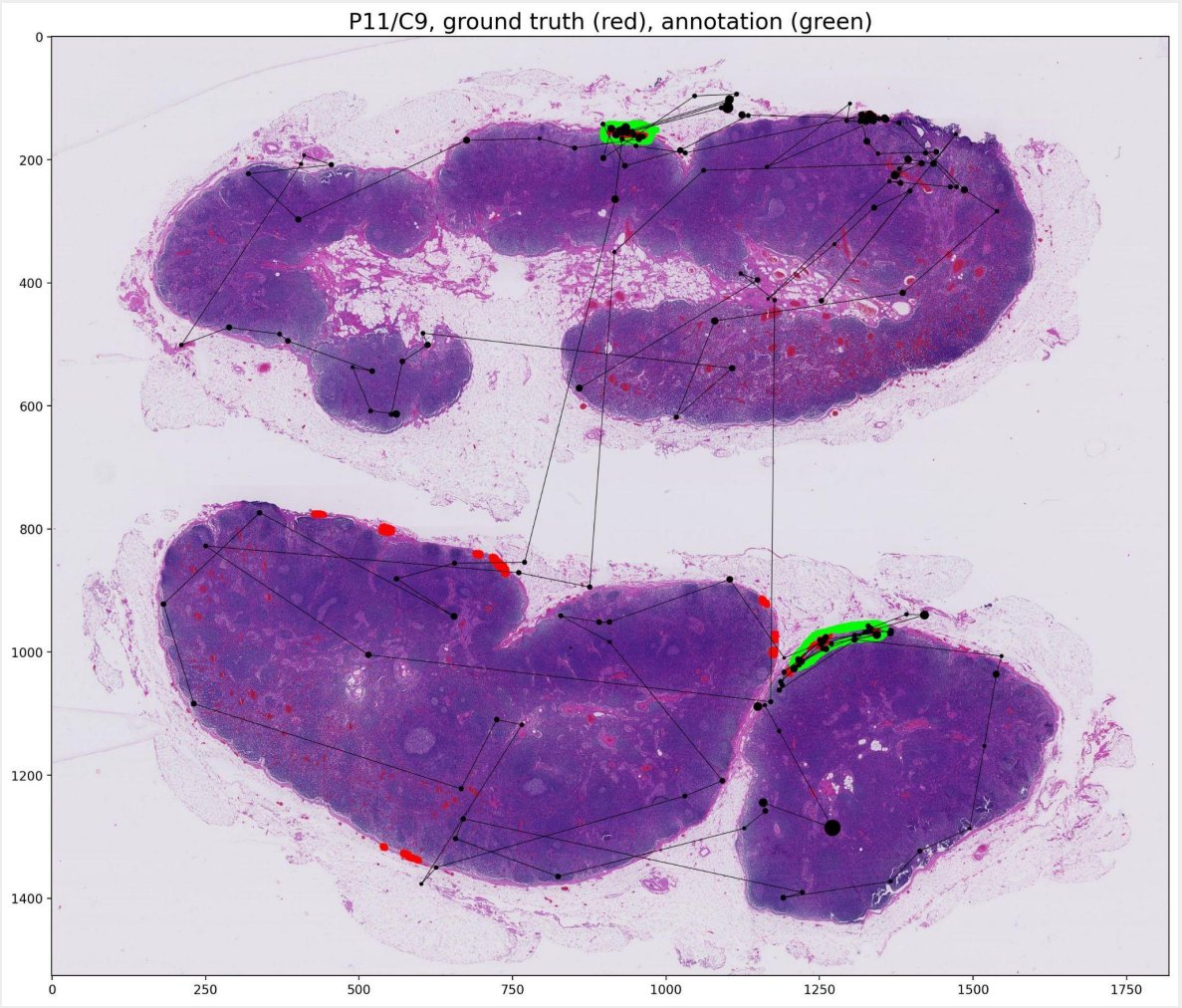


Figure 1. The green regions are ground truth cancer regions. The lines shows the visual scan path from a single pathology. The dot shows the fixation points.

  - TrialID: trial id
    - a pathologist examines a single slide is called one trial. Each pathologist did 60 trials.
  - Username: participantsID; there are 11 experts named p1-p11.
    - Their experience levels are here:
      - Resident: P2, 5, 10
      - [0, 5] years in service: P1, 4, 9
      - [6, 10] years in service: p3, 6, 11

        - >10: P7, 8
    - Image_id: there are 60 images from C1-C60. 32 times smaller along each dimension is stored in the image directory: https://drive.google.com/drive/folders/1zYv5oFzcQ8-1cYO-pBrFpPBNVC7MPReI?usp=sharing.
    - imgW | imgH: the size of the whole-slide image represented by the pixel resolution. The version we provide you is 32 times smaller along each of the x / y axis. So a 32x32 pixel image becomes a 1x1 pixel display.
    - ImageZoomR: this is the zoom level during the navigation. The pathologist can zoom in and out of the huge image. The range of this data is (0, 40]
    - CVisibleArea: this shows which part of the image is visible in the current view. There are 4 variables to construct the window.
    - InNavigator: This flags if the participant looks insight the navigator view (yes if true) or not (false). ~~this is not useful gaze - so it is false~~
    - relativeTime: when does that action in that row occur in ms (millisecond) since this trial begins.
    - inGT: fixation in ground truth tumor region
    - inAnnotation: fixation inside the annotated area (viewer annotated region)
    - errorType: for that moment (not trial). TN means the fixation is not in any tumor area and that fixation point also does not contain tumor. TP: means the fixation is in the tumor area defined by GT. TP: fixation is looking at the right tumor region (== inGT).
        - Recognition error: I look but don't see. In the table, this represents that single row.
        - Decision error: this is also a type of recognition error but the difference between recognition and decision is the length of fixation - longer than 1000ms, a recognition error becomes a decision error.
        - Detection error: the participant failed to look at the tumor region.

One idea: these errors should not occur in the TN cases. You could use visualization to detect errors in data. These errors are often associated with a person's expertise. Recognition and decision errors could occur in the TP cases because in that moment (row), the error occurred. The participant could still make the correct final decision (because there may exist multiple tumor areas.)

    - lesionInView: cancer region (GT) on the visible area (CVisibleArea).
        - So you may have noticed that 0s are used for TN cases. They should be N/A. In data science, N/A is different from 0.
    - tumorInView: total numbers of tumor areas. Each area is defined by the connected pixels.
    - tumorIndicesArea
        - List of all visible tumors and their respective areas. Entries are in the format $x_1;a_1,x_2;a_2,...,x_n;a_n$ where x_i is a tumor index and a_i is the corresponding area.
    - GazeEventDuration: for how long the viewer look at that location of (CFixationPointX, CFixationPointY) at the zoom level of imageZoomR.

These notations are applicable to overall diagnosis. A file Dr. C will provide (you can choose not to use it.)

    - TN: true negative (no tumor, the answer is also no tumor): TP: true positive (has tumor, the answer also has tumor), FN: false negative (GT has tumor, but reported no tumor); FP: false positive (GT no tumor, but reported has tumor). In cancer diagnosis, FP is largely fine, although it gives stress to the patient and patient's family members. Often, hospitals would do another round of checks to make sure the tumor exists. FPs are often getting fixed later.

If you have used the eye-tracking dataset from the last assignment, you will have already somewhat done a proportion of this assignment. If so, try something different.

You should

- seek to gain an overview of the shape & structure of your dataset – this question will ask about the overall dataset. For example, Show an overview of all 11 pathologists' visual scan trajectories from a single image using a line plot.
    - What variables does the dataset contain?
    - What would you show in your visualization?
    - How do the scan paths distributed?
    - Are there any notable data quality issues?
- Are there any surprising relationships among the 11 participants' variables?
    - The participants have four different experience levels. Would the experience be related to fixations?
    - Any outliers? An outlier is a data point that is significantly different from other observations.

Next, be sure to also perform "sanity checks" for patterns you expect to see. Sanity checks mean checking the data to make sure they are correctly shown. The plots generally match your expectations. If not, explain why.

Step 2: Exploratory Analysis

Prior to this analysis, write down an initial set of at least four | two questions you'd like to investigate related to someone looking for cancers. In the data, we captured the fixations (https://en.wikipedia.org/wiki/Fixation_(visual)) , and experience levels (by years in service: resident, [0-5], (6, 10], >10 years in service). **These questions should be composed using different scales.**

In this phase, you should investigate your initial questions, as well as any new questions that arise during your exploration. You can even look for newer data to facilitate your data exploration (because data science problem-solving is an iterative process) – certainly, this is not mandatory.

For each question, start by creating a visualization that might provide a useful answer. Then refine the visualization (by adding additional variables, changing sorting or axis scales, filtering or subsetting data, etc.) Adding images to the data etc.

To develop better perspectives, explore unexpected observations, or sanity check your assumptions. You should repeat this process for each of your questions, but feel free to revise your questions or branch off to explore new questions if the data warrants.

Write down every stage of the data processing - The lecture code example might be useful to let you see the messiness of real-world data exploration problems - Please feel free to use those code examples to explore the data.

Step 3: Submission

Your final submission should take the form of a document report – similar to a slide show or comic book – that consists of 3-5 (or more if you wish) captioned visualizations detailing your most important insights. Like lab 1, please also submit your code. Please also post your visualization and your report to Piazza so other people can learn from your work. Like 1st assignment, please do not post code.

Your "insights" can include important insights, such as experienced people tend to be fast; they scan less regions etc. It can also be some observations, e.g., this person's scan path is a bit ad-hoc; compared to other people. You may not validate it computationally, but your visualization techniques will lead to observations that later let you begin to think about computational solutions – this is a purpose of visual exploration.

To help you begin, just grab the data and plot some simple charts – say, line charts or scatter plots to let you see what might be interesting initially. Then you can refine your visualization methods by turning those line charts to the starglyph for example – then you will see that shape may let you put more data on paper. Try not to delete any charts you draw – the process is also important to show what might not work well. When you post, only post your best ones.

To help you gauge the final product of this assignment, see the example in the "Help" section for a similar report analyzing data about movies | motion pictures. The report has been annotated and graded this example to help you calibrate for the breadth and depth of exploration one can do. This example below has many charts - you are free to create as many as you wish. Creating different types of charts may let you generate different insights. You may focus on a smaller set (3-5) which gives you an interesting story you like to tell others about pathologists' viewing process. So pick and choose carefully.

Each visualization image should be a screenshot exported (Google colab code provided to save image file and your images will be stored to your google drive folder you can download). In your report, please caption these figures accompanied with a title and descriptive caption (1-4 sentences long) describing the insight(s) learned from that view. Provide sufficient detail for each caption such that anyone could read through your report and understand what you've learned. You are free, but not required, to annotate your images to draw attention to specific features of the data. You may perform highlighting within the visualization tool itself, or draw annotations on the exported image.

Name your python | D3 code project to file to <Your lastName.dotNumber>.lab2 code|doc. You should consider two different phases of exploration. The end of your report should include a brief summary of main lessons learned.

Help

Assignment 2 Method

Where do questions come from?

Remember each real-word problem starts with a question. (See the lecture on Feb 7th on task analysis. Watch the Hans Rosling's video again to see how he has used animation to support what he likes to express). Often the questions are related to what we learned in class Lec04.TM3.taskAbstraction | taskExamples.

What we learned in the lecture was that if the data have been collected, there are a finite set of questions you can answer.

(0) data attributes

N, Q, C (see lecture slides)

Q>N>C
So what is quantitative is likely ordered and qualitative
What is ordered is also qualitative
What is qualitative is neither quantitative nor ordered, but is arbitrarily reorder able.

Any questions you can ask about N, C, you can also ask Q. The reverse is not true.
Any questions you can ask about C, you can also ask N. The reverse is not true.
See slide #17 from Lec03.Foundation(2).

There are a set of questions you can answer in related to the number of items: single, partial or overall. You can also ask relationship questions or associate where (location) to what (event) or when (time).

(1) proportion of the data

- Single data point (near the average, extremes - largest or smallest)
- A fraction of all data point (half of the data … in the earlier year, the data show…)
- All data points (the overall trend of the data…)

(2) aggregation level (related to sampling statistics, e.g., mean, standard deviation, grouping by region etc.)

Where do visualizations come from?

- Lecture materials
- Try to change the visual marks (e.g., line styles) and channels (e.g., color, texture, shape)
- Try to sort the data
- Try to change the axis or even rotate your chart
- Try to add trend-lines
- See examples for inspirations and storytelling strategies for news media:
    - https://www.nytimes.com/interactive/2022/12/28/us/2022-year-in-graphics.html
    - https://www.nytimes.com/interactive/2021/12/29/us/2021-year-in-graphics.html
    - https://www.nytimes.com/interactive/2020/12/30/us/2020-year-in-graphics.html

Other notes:

- Schneiderman's mantra: Overview first (to show all data), zoom and filter (see parts), detailed on demand (item-wise query)
- We haven't learned interaction. To create a zoom or a filter effect, try zooming or subsetting in a static view in your story.

Starter code if you wish: (D3 and python to load an image)

- https://drive.google.com/drive/folders/1KT65v9a6xRnax7niubeGeNXG4ncGLCIt?usp=sharing
- It is not mandatory to use it.

Assignment 2 Example Answer (based on the movie data)

This is an example exploratory analysis report intended to give you a sense of the scope of Assignment 2. This particular report concerns a dataset of motion pictures. Note that this analysis is ***far from perfect***. There are interesting follow-up questions the analysis does not pursue, and also fundamental data quality and sample selection (filtering) concern that are ignored. As you read through, ask yourself what you might do differently!

Dataset: Motion Pictures Data

This dataset contains statistics for a sample of 3,201 movies collected in 2010. The data includes movie titles, genres, and box office gross revenues, as well as audience (IMDB) and critic (Rotten Tomatoes) ratings. The dataset combines data from multiple sources: Rotten Tomatoes, The Numbers, and IMDB.

The data is available online as movies.csv.

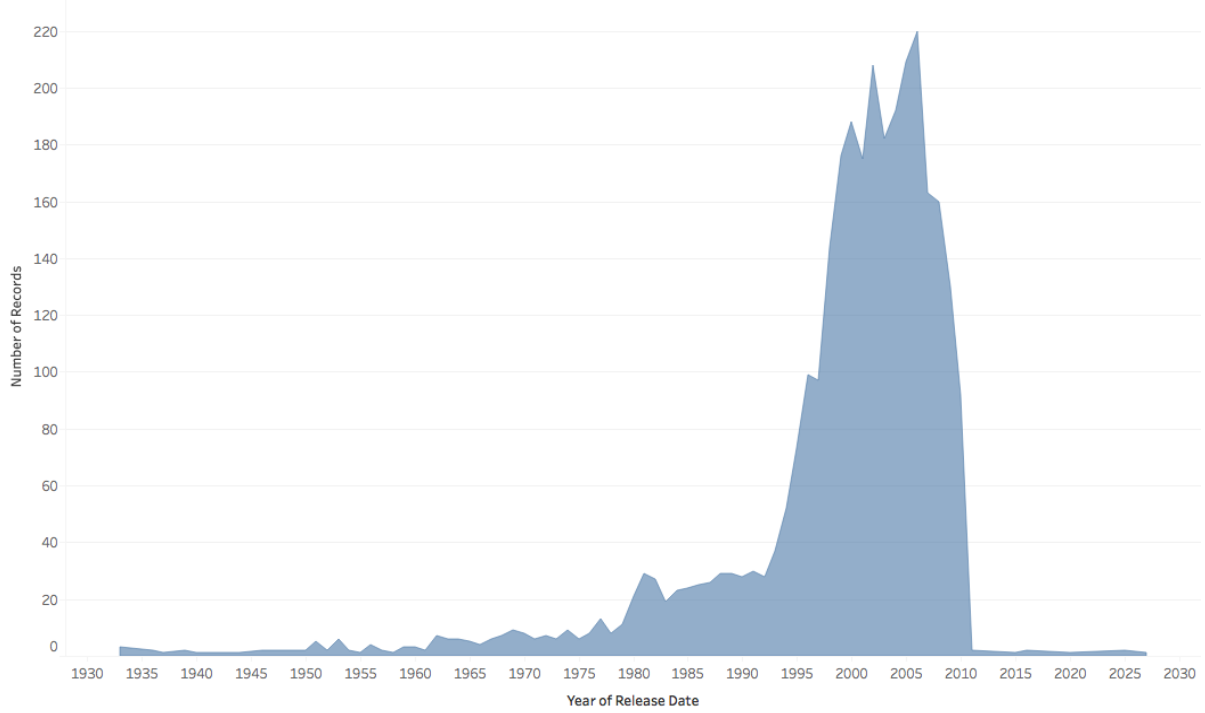
Initial Analysis Questions

1. **Overarching Question:** What factors drive the box office success of motion pictures?
2. How do films fare in the U.S. vs. worldwide markets?
3. Do producers strategize release dates for certain kinds of films?
4. How do fan favorites and critical darlings relate to ticket sales?

Discoveries & Insights

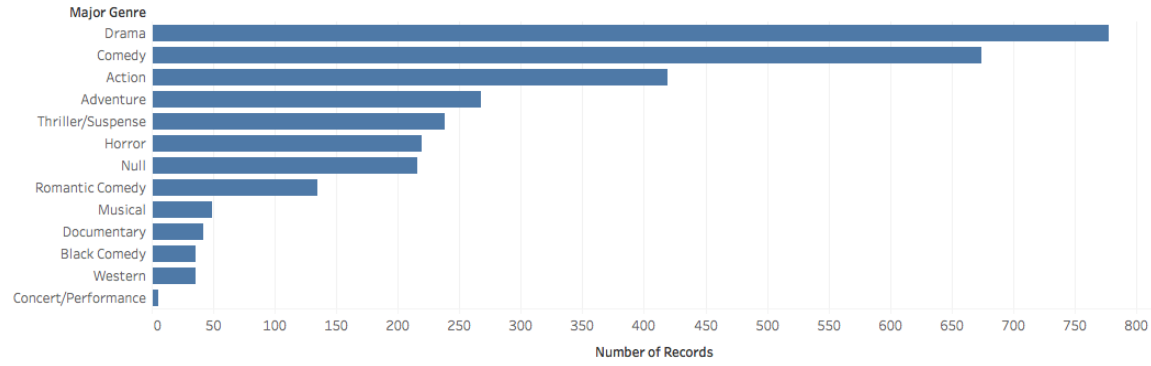
Our analysis starts with plots of individual variables to assess distributions and data quality. As we progress, we build up multi-dimensional views for our analysis questions.

Summary of Release Date



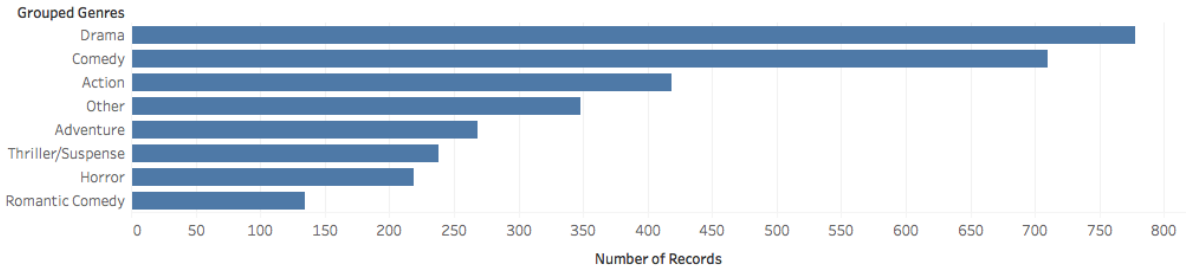
This chart shows the count of movies for each year. The dataset has an uneven distribution, with a long tail of historical films and larger numbers of more recent films (1999-2009). We also see films from the future! These are most likely errors or films in production. As a result, all subsequent views include a filter to remove films with release dates that are either null or have years after 2010 (the publication date of this dataset.)

Summary of Major Genre



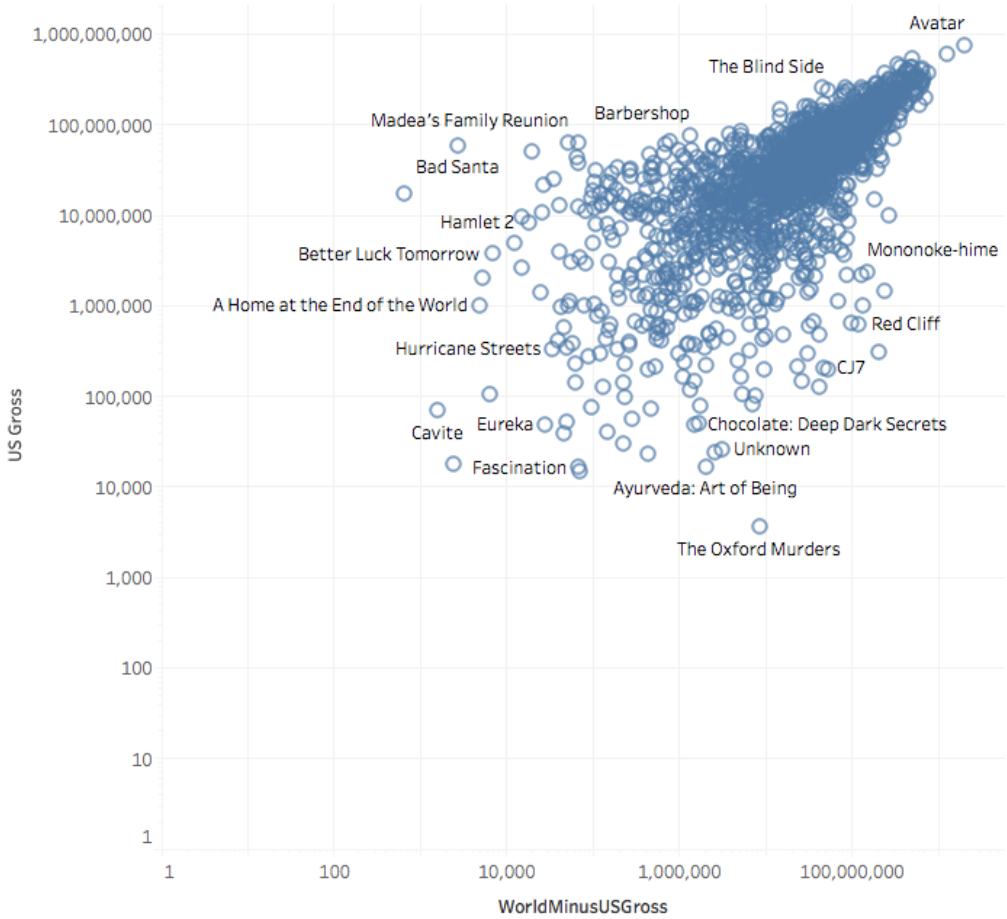
Next, we look at the number of movies by genre, sorted by count. We see that Drama is the most prevalent, followed by Comedy, Action and Adventure. Some of the categories have many fewer films (less than 100) and so due to sampling error may not be as reliable for assessing overall trends. We also see a "Null" category for (presumably) uncategorized films.

Summary of Grouped Genres



In order to simplify subsequent analyses and ensure a minimum sample size (> 100) per genre, we adjust the genre taxonomy by grouping them together. Here, Comedy and Black Comedy have been merged together, while the Null genre and all other groups with less than 100 films have been combined into a new "Other" category.

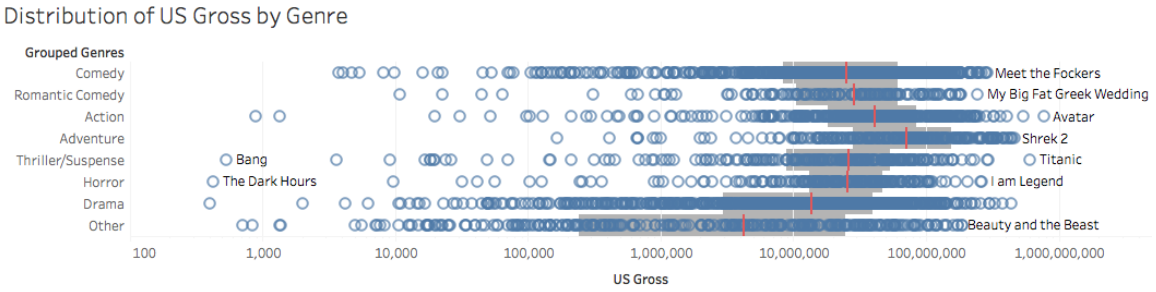
US vs WW



We now turn to looking at measures of a films' box office success. Our dataset contains both US Gross and Worldwide Gross fields. As the worldwide results should contain the US results, we first create a new variable that subtracts the US figures from the worldwide result.

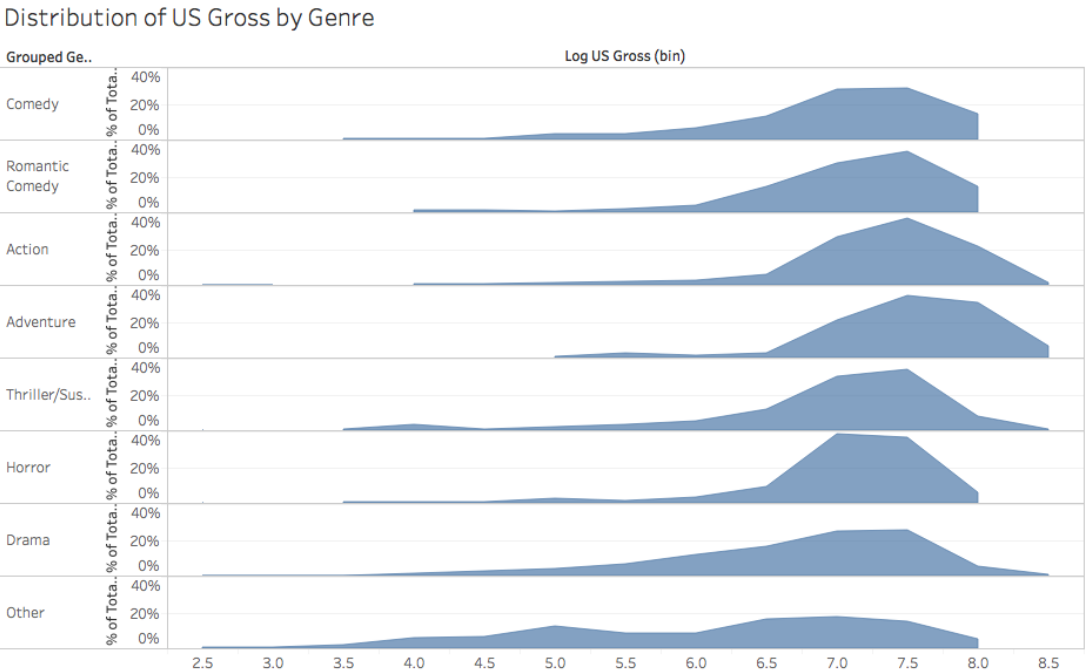
This plot shows a scatter plot (with log base 10 scaled axes) comparing those two. We can see that by and large US gross and Worldwide (minus US) gross follow similar patterns. Based on this high level of association, we will largely focus on US gross in subsequent plots.

On the edges of the distributions, we find films that are either popular in the US but less so elsewhere (or vice versa). High-grossing films in the US (but not elsewhere) include the Barbershop films and Bad Santa. High-grossing films outside the US unsurprisingly consist of foreign-made films. And a few films are not popular anywhere... :)



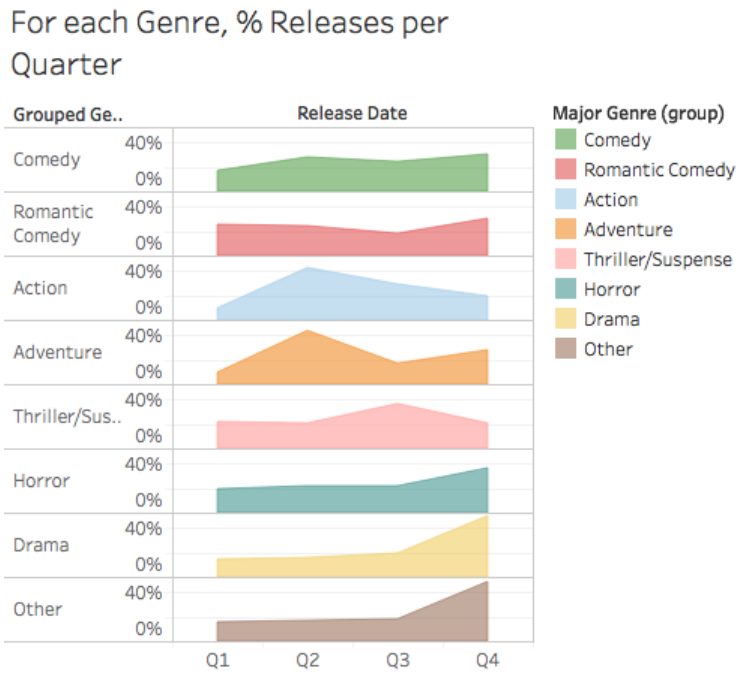
This log-scaled dot plot shows the distribution of US gross by genre. Each film is included, along with reference lines for the interquartile range (middle 50% of the data, in grey) and median values (in red). We see that, on median, Adventure films have the greatest revenue, followed by Action films. Next, comedies (both romantic and otherwise), Thrillers, and Horror films show similar central tendencies.

We also note differences in the tails of the distribution, as some genres have a larger proportion of films with very low US gross. For example, the Other category – which includes documentaries and concert films – fits this pattern.

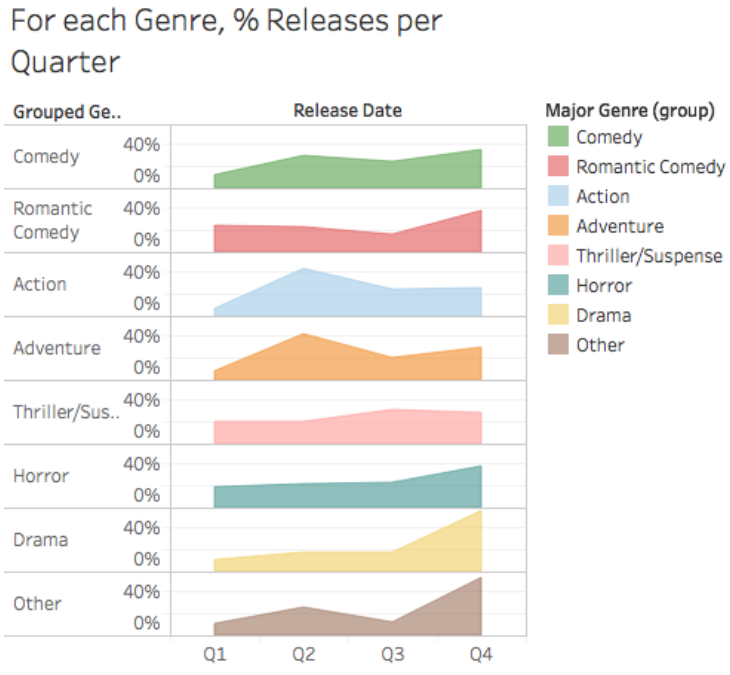


This ridge-plot (python: https://seaborn.pydata.org/examples/kde_ridgeplot or D3: https://d3-graph-gallery.com/ridgeline.html) shows an alternative view of the previous chart, but showing a normalized histogram to give a better sense of the shape of the per-genre probability

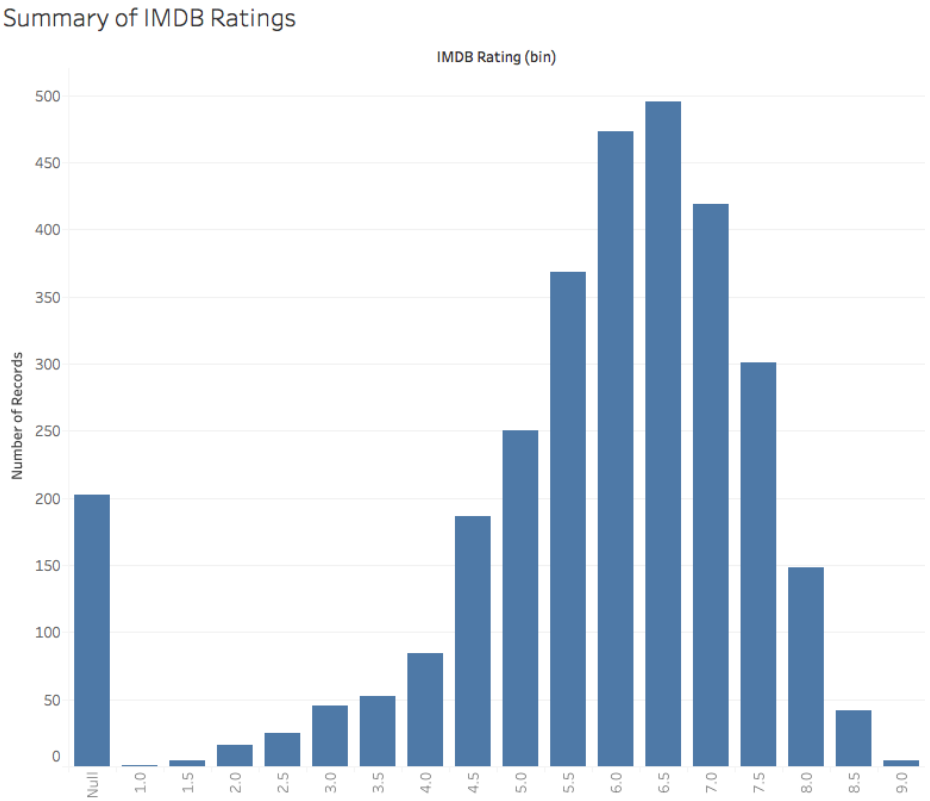
densities for US revenue. The y-axis shows the total percentage of records (within that genre) occurring within a given bin.



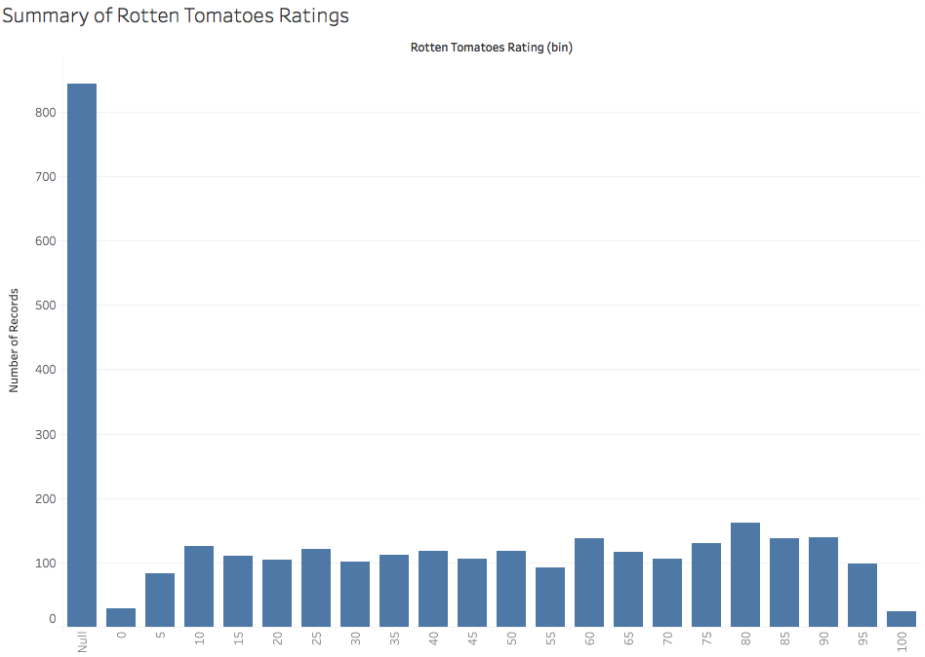
We now move on to examine seasonal patterns. This ridge-plot shows the percentage of US gross for films released in each financial quarter, subdivided by genre. The genres have been sorted by peak release season. We see that Comedy proceeds are largely steady throughout the year. Action and Adventure films peak in the spring and early summer, whereas Thriller/Suspense films peak in the late summer. Horror movies peak in Q4 (perhaps around Halloween?) and Dramas also peak in Q4 (released in time for Oscar consideration?). The Other category also peaks in Q4, perhaps with musicals or concert films targeting a holiday release?



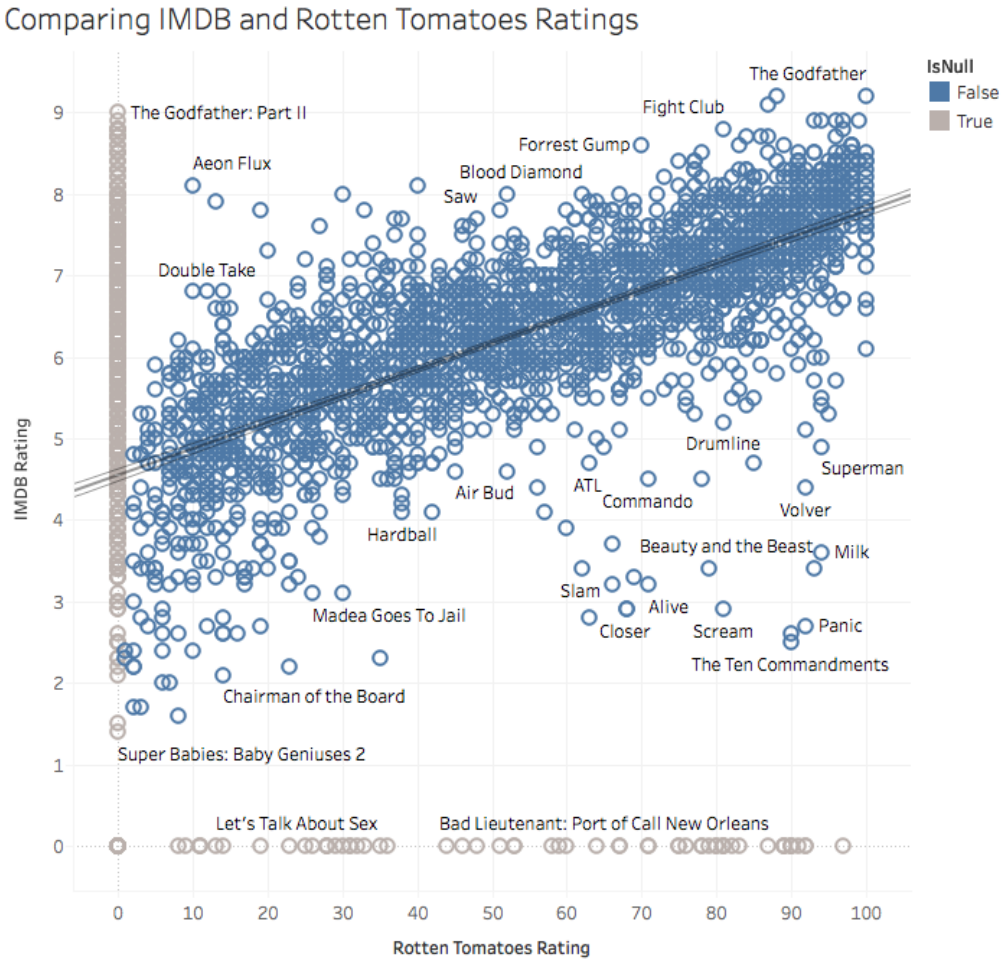
Do worldwide trends match the US trends? This plot is identical to the prior chart, but with US gross replaced by Worldwise (minus US) gross. We see very similar patterns as before.



Having investigated the relationship between ticket sales and genre, we now turn our attention to incorporating ratings. This plot shows a histogram of IMDB user ratings. We see a positively-skewed bell curve shape with a modal rating around 6.5. This is higher than neutral: is this because people, in general, like movies, or is there sample bias in the data? We also see a number of "null" ratings, for which we are missing data.

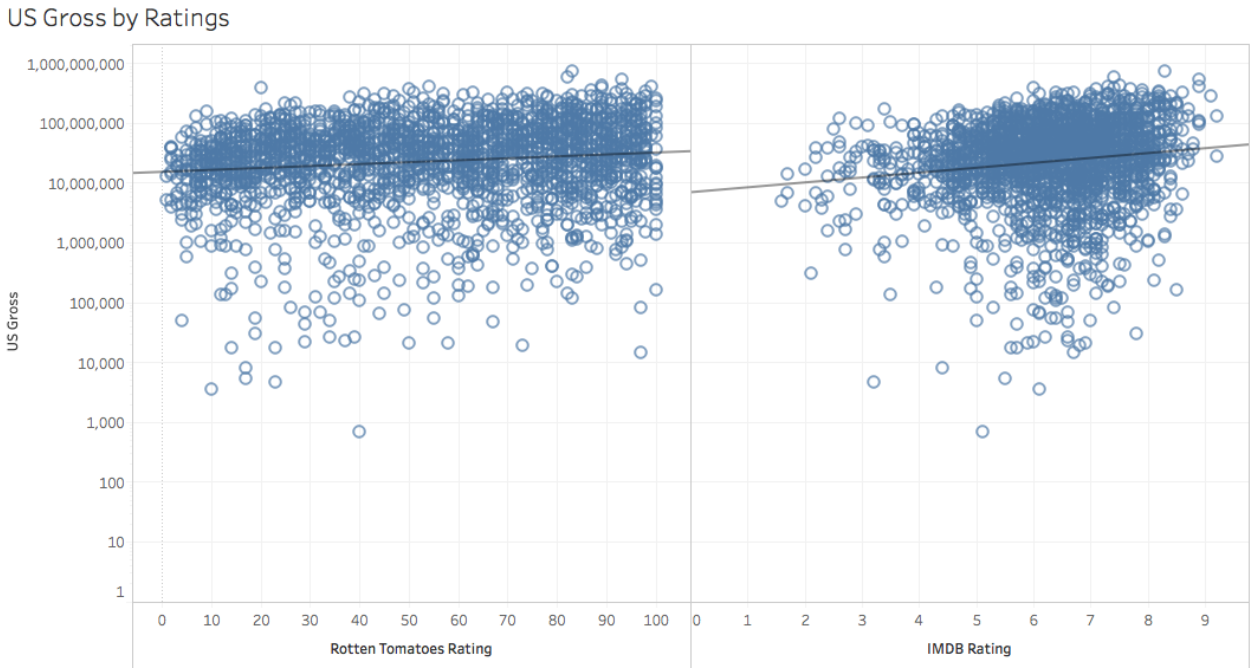


This plot shows a histogram of critic ratings from the site Rotten Tomatoes. Similar to the IMDB ratings, there are a number of "null" (missing) values. Unlike the bell-curve shape of IMDB ratings, the shape of Rotten Tomatoes ratings indicates a more uniform distribution. This discrepancy may be due to the different mechanisms: IMDB simply aggregates user ratings on a 1-10 scale, while Rotten Tomatoes shows the 0-100 percentage of thumbs-up reviews (i.e., like coin tosses) from a collection of critics.



This scatter plot compares the IMDB and Rotten Tomatoes ratings directly. We see that the two are highly correlated (an observation supported by the least-squares regression trend line). Films popular in one system tend to be popular in the other, notably including The Godfather, which has the highest score in both sets of ratings. However, we can also find bivariate outliers that are popular with users but not critics (or vice versa). For example Aeon Flux is popular among viewers but not critics. Meanwhile, Panic is rated highly by critics but unloved by viewers.

For context, we also include films that have a null score in one of the rating sets in gray along the edges of the chart.



Now we explore our final question: does the "quality" of a film (as reflected by its ratings) affect the box office ticket sales. To address this question, this visualization consists of two side-by-side scatter plots, showing the relationship between either IMDB and Rotten Tomatoes ratings and US gross. We also include trend lines, fitted with an exponential model to account for the log-scaled axis for US gross. Both the plots and the trend line indicate a positive relationship (higher ratings, higher gross on average), but with a very shallow slope.

Summary. The take-away? Better films do seem to make more money, but the effect is weak. Perhaps this is a function of a Hollywood strategy of pre-release marketing to hype up a film before word-of-mouth spreads? So while studios might strive to make great films, it seems that other, more easily controlled, factors such as genre and release schedule might lead to more reliable projections of future ticket sales.

Rubric

Component	Excellent (extra 20% pts)	Satisfactory (full points)	Poor
Breadth of Exploration	More than 4 | 2 questions were initially asked, and target substantially different portions/aspects of the data.	At least 4 | 2 questions were initially asked of the data, but there is some overlap between questions.	Fewer than 4 | 2 initial questions were posed of the data.
Depth of Exploration	Several follow-up questions were asked and yielded insights that helped to more deeply	Some follow-up questions were asked, but they did not take the analysis much deeper than the	No follow-up questions were asked after answering the initial questions.

	explore the initial questions.	initial questions.	
Data Quality	Data quality was thoroughly assessed with extensive profiling of fields and records.	Simple checks were conducted on only a handful of fields or records.	Little or no evidence that data quality was assessed.
Visualizations	More than 8 | 4 visualizations were produced, and a variety of marks and encodings were explored. All design decisions were both expressive and effective.	At least 8 | 4 visualizations were produced. The visual encodings chosen were largely effective and expressive, but some errors remain.	Several ineffective or inexpressive design choices are made. Fewer than 8 visualizations have been produced
Data Transformation	More advanced transformation such as considering image masks to compute if fixations are inside the mask areas. Other activities used to extend the dataset in interesting or useful ways	Simple transforms (e.g., sorting, filtering) were primarily used.	The raw dataset was used directly, with little to no additional transformation.
Captions	Captions richly describe the visualizations and contextualize the insight within the analysis.	Captions do a good job describing the visualizations, but could better connect prior or subsequent steps of the analysis.	Captions are missing, overly brief, or shallow in their analysis of visualizations.
Creativity & Originality	You exceeded the parameters of the assignment, with original insights or particularly creative visualizations or transformations.	You met all the parameters of the assignment.	You met most of the parameters of the assignment.
Discussion & Sharing		Post your visualization and your report to Piazza	The visualization and report are not posted to Piazza.

Final highest point one can earn: 125/100.

FAQ:

1. Do I need to draw all 11 participants' data?

No. You can pick and choose. Overall here will mean fixation and scan from the chosen set. If you can depict all participants' data clearly, we will give you 20% extra credits (because this is hard). Some ideas would be to draw say figures by expertise levels.

2. Can I draw from one image?

Yes, you can see even drawing one is challenging. A student drew the figure below to show the zoom levels of all participants viewing a single image. But it is very difficult to be scalable.

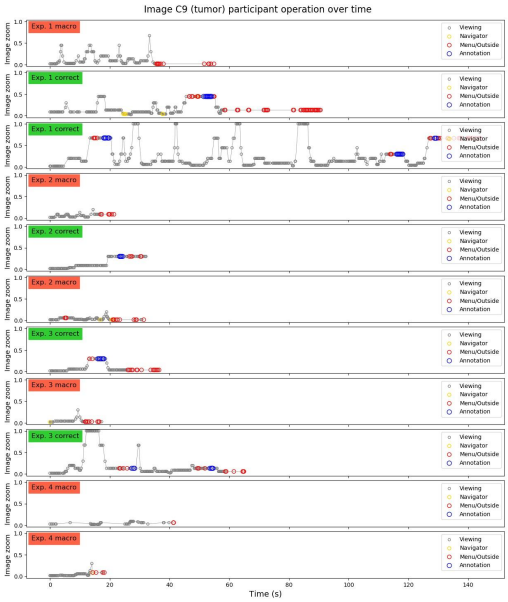


Figure 2. The horizontal axis shows the time line. The vertical axis shows zoom level. Small multiples shows all 11 participants data. The red tag shows wrong answers; The green tag shows correct answers.

3. What are other useful data I can use?

The following figure draws the ground-truth cancerous tissue region to the overall picture size. The smaller the ratio, the harder the image.

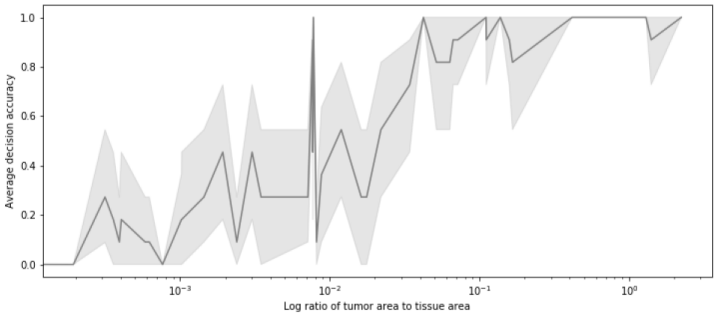


Figure 3. The horizontal axis shows all images. The vertical axis shows the ground-truth cancer region to overall picture size ratio. Small multiples shows all 11 participants data. The red tag shows wrong answers; The green tag shows correct answers.