CASE STUDY DETAILED INSTRUCTIONS

‘What's the Next Big Thing?’: Netflix and Analyzing Data for Insights

N.B.: Before starting this assignment, read the case study, ‘What's the Next Big Thing?’: Netflix and Analyzing Data for Insights.

Your assignment

Put yourself in Zach Joel’s shoes. You’re hoping to glean high-level insights into the key questions you’ve agreed to explore for Netflix’s head of content marketing. Consider online self-service tools from Netflix’s digital platform partners – such as Google and Facebook – and open-source datasets from IMDb.

Your objective is to find interesting patterns that offer insights into your key questions. The output from this exercise should be a collection of exploratory, data-driven charts and tables that identify areas for further exploration and refinement as your data story takes shape.

Follow these steps to complete your assignment:

Collect & clean IMDb data

1. Download the seven files found at IMDb’s facilitated download site here
2. Unzip files by clicking on them
3. For simplicity’s sake, leave the files in your “Downloads” folder

Analyze IMDb data

4. Launch R and use the script found here to conduct a thorough analysis of the data
5. After completing the analysis, conduct further analysis of csv files you’ve exported to your desktop (in the last step of the R script) using one or more tools:
    a. Spreadsheet tools like Excel or Google Sheets
    b. Presentation tools like Powerpoint or Google Slides
    c. Dataviz tools like Tableau
6. Produce a series of data-based, exploratory data visualizations that reveal some interesting patterns in the IMDb data

Use Google Trends to identify key areas of consumer interest (and geographic relevance)

7. Open a web browser and navigate to trends.google.com
8. Enter a search term or topic related to your insights in the query box (labeled, "Enter a search term or topic") and hit enter
    a. Remember these important rules of search on Google Trends with example queries related to the topic of *tennis shoes*
        i. *Tennis shoes*: Searches will include tennis AND shoes, in any order, and possibly with other search terms
        ii. *"Tennis shoes:"*:Searches that include tennis shoes – that is, the exact search terms inside the quotation marks
        iii. *Tennis - shoes*: Google Trends includes only searches for tennis – not for shoes
        iv. *Tennis* + *shoes*: Google Trends includes tennis OR shoes, in any order, and possibly with other search terms
        v. *Sneakers* + *sneekers* + *sneakerss*: Google Trend searches include alternative spellings and aren't case-sensitive
    b. Note that Google Trends provides access to a largely unfiltered sample of Google search requests that's anonymized (no one is personally identified), categorized (determining the topic for a search query), and aggregated (grouped together)
9. Interpret the results of your query
    a. Google Trends will return a "search index" for the term(s) you're exploring – not a number that's the actual *search volume* for term(s)
        i. For example, "100" on the Google Trends report doesn't mean only 100 searches were conducted on that day for the term(s) you're exploring
        ii. There could have been hundreds, millions, or even billions of searches for that term(s) on that day, but Google keeps those figures confidential
    b. Google Trends normalizes search data to simplify comparisons between terms
    c. Results are normalized to the time and location of a query using this process:
        i. Each data point is divided by the total searches of the geography and time range it represents to compare relative popularity (otherwise, places with the most search volume would always rank highest)
        ii. The resulting numbers are then scaled on a range of 0 to 100 based on a topic's proportion to all searches on all topics
        iii. Different regions that show the same search interest for a term don't always have the same total search volumes

d. Google Trends data reflects searches people make on Google every day, but it can also reflect irregular search activity, such as automated searches or queries that may be associated with attempts to spam search results
    i. While Google has mechanisms to detect and filter irregular activity, Google Trends may retain these searches as a security measure. Filtering them from Google Trends would help those issuing such queries to understand how Google has identified them, making it harder to keep such activity filtered out from other Google Search products where high-fidelity search data is critical
    ii. Given this, those relying on Google Trends data should understand it's not a perfect mirror of search activity

e. Google Trends filters out searches, including:
    i. Those that few people make: Trends only shows data for popular terms, so search terms with low volume appear as “0”
    ii. Duplicate searches: Trends eliminates repeated searches from the same person over a short period
    iii. Special characters: Trends filters out queries with apostrophes and other special characters

f. Note that if you enter a search term using non-Latin characters, you only see data from countries or regions that use those characters
    i. For example, if you enter “ねこ,” the Japanese characters for “cat,” you don’t see much data for the U.S. since many people in the U.S. use “cat” as their search term
    ii. To compare searches for Japanese characters for “cat” to searches for “cat”in English, search for both terms by combining them with a “+” sign, like “ねこ + cat”

10. Refine your search to add context and deepen your insights
    a. Add a new term (or terms) to compare to your original term
        i. Inside the Topics box, click + Add term and add another search term or terms
        ii. To remove a term, hover over the search term box and click X
    b. Change the geographic location of your results in one of two ways:
        i. Change the geographic scope of all your results by selecting a new location from the drop-down list below the search box (the location will default to your present geography)
        ii. Change the geographic scope of a particular term by clicking on the three dots

on right side of the search box,select, “Change filters,” and then choose a different location for that term
c. Change the period for your results in one of two ways:
   i. Change the period of your results by selecting a new time frame from the drop-down list below the search box (it will default to the past 12 months)
   ii. Change the geographic scope of a term by clicking on the three dots to the right side of the search box, select, “Change filters,” and then choose a different period for that term
d. Depending on your needs, define your search words as terms or topics by clicking on the term and making a selection
   i. Terms: Search terms show matches for all terms in your query, in the language given
      1. If you search the term,“banana,” results include terms like “banana” or “banana sandwich”
      2. If you specify “banana sandwich,” results include searches for “banana sandwich,” as well as “banana for lunch” and “peanut butter sandwich”
   ii. Topics: Topics are a group of terms that share the same concept in any language. Topics display below search terms.
      1. If you search the topic, “London,” your search includes results for topics such as “Capital of the UK” and “Londres,” which is “London” in Spanish
e. If you’re using Trends to search for a word that has multiple meanings, you can filter your results to a certain category to get data for the right version of the word
   i. For example, if you search for “jaguar,” you can add a category to indicate if you mean the animal or the car manufacturer
   ii. Under the search box, click “All categories,” then choose a category
f. You can filter your results to a certain search platform to get data for that platform
   i. Under the search box, click “Web Search,” then choose a platform
      1. Image Search
      2. News Search
      3. Google Shopping Search
      4. YouTube Search

11. Explore results by region to gain geographic insights
   a. When you search for a term in Google Trends, you’ll see a map that shows areas where

your term is popular

        i. Darker shades indicate where your term has a higher probability of being searched

    b. If you compare search terms, you'll see a map of the world shaded according to the term's popularity

        i. The color intensity represents the percentage of searches for the leading search term in a particular region
        ii. Search term popularity is relative to the total number of Google searches performed at a specific time, in a specific location

    c. Hover over a region to get details on search volume in that region

        i. To the right of the map, you'll also see a list of regions or cities ranked according to the term's popularity

    d. To view a metro, which are geographical areas that generally correspond to metropolitan areas, follow the steps below (N.B.: Currently, Google Trends only provides metros for some countries):

        i. In the Regional interest section, click the U.S. on the map
        ii. Click a state on the map
        iii. Click a metro area on the map
        iv. If a region on the map isn't highlighted, it doesn't mean there's no interest. Google Trends data is adjusted, so the term may be used in that region, but it's more popular in other regions.

12. Find related searches to expand your results

    a. When you search for a term in Trends, you'll see searches related to your term in the related searches sections at the bottom of the page
    b. If you're comparing multiple search terms, locations, or time ranges, you can see related top searches by selecting the tab for your term

        i. Top searches: Top searches are terms that are most frequently searched with the term you entered in the same search session, within the chosen category, country, or region. If you didn't enter a search term, top searches overall are shown.
        ii. Rising searches: Rising searches are terms that were searched for with the keyword you entered (or overall searches, if no keyword was entered), which had the most significant growth in volume in the requested period. For each rising search term, you'll see a percentage of the term's growth compared to the

previous period. If you see “Breakout” instead of a percentage, it means that the search term grew by more than 5,000%.

13. You can export Trends data to see a comprehensive list of search data as CSV files:
    a. In the top right of the chart, click Download
    b. Open the file using a spreadsheet application, like Google Sheets

Use Facebook Audience Insights to size audiences and collect demographic details

14. Navigate to www.facebook.com/business/insights/tools/audience-insights
15. Click on the “Go to Audience Insights” button
    a. You’ll need a Facebook account to use Audience Insights. Facebook will ask you to log in to your account if you’re not already logged in.
    b. Facebook may also ask you to “Choose an Audience to Start”
        i. If asked, choose “Everyone on Facebook”
16. Create an audience relevant to your IMDb analysis using the filters on the left side of the Audience Insights report
    a. Make sure to deselect your default location (usually your home country) so that you do not limit your analysis
        i. Netflix is, after all, a global operation and understanding where trends exist (and where they don’t) is very valuable to the company
    b. Important topics like movie genres and actors can be easily found by simply typing them into the “Interest” search box
    c. Use the additional selections to bring more texture to the interest trends you’ve identified (N.B.: Try to limit your selections inputs that clarify the interest so that the audience demographics arise organically from the Facebook data)
        i. Location (country, region, or city)
        ii. Age (or age range)
        iii. Gender
        iv. Interests
        v. People connected to your page (or not connected to your page)
        vi. Languages spoken
        vii. Relationship status (Single, In a relationship, Engaged, Married)
        viii. Education levels (high school, college, grad school)
        ix. Work (job titles)
        x. Market segments (multicultural affinity)

        xi. Parents (based on children’s ages)
        xii. Politics (in the U.S. only, which includes Very Conservative, Conservative, Moderate, Liberal, Very Liberal)
        xiii. Life events (Away from family, Away from hometown, Long-distance relationship, New job, New relationship, Recently moved, Upcoming birthday)

17. Note that Facebook will create a demographic report of the audience you create (visualized in blue) and compare that to “All people on Facebook” (visualized in grey) for context
    a. Facebook estimates how many people fit your audience criteria. This is an estimate of the size of the audience that’s eligible to see an ad on Facebook. It’s based on your targeting criteria, ad placements, and how many people were shown ads on Facebook apps and services in the past 30 days.
    b. This isn’t an estimate of how many people will actually see your ad, and the number may change over time. It’s not designed to match population or census estimates.
18. Take note of the Page Likes (categories and pages), Locations (Top Cities, Top Countries, Top Languages), Activities (“lifetime” and “last 30 days” counts of Pages Liked, Comments, Posts Liked, Posts Shared, Promotions Redeemed, Ads Clicked) and Devices Used for this audience
19. Use this information to bring depth to your understanding of the audiences Netflix should consider and related interests that Netflix should consider

Use TweetReach to sample consumer sentiment

20. Open a web browser and navigate to tweetreach.com
21. Click on “GET STARTED”
22. Enter a hashtag, account, or keyword you want to explore in the search box (labeled, “Enter a hashtag, account or keyword”)
    a. Think of titles, properties, actors, directors, and topic areas that are related to what you know about Netflix’s opportunities and consumers
23. Interpret the results of your query
    a. Note that you must sign in with Twitter to run your snapshot report
        i. TweetReach will only use this information to retrieve Tweets that match your search query and won’t post anything without your permission
        ii. You can avoid having to sign into Twitter each time by creating a free account with TweetReach
    b. The Twitter snapshots provide insight into metrics such as reach, impressions, top Tweets and contributors for up to 100 Tweets from the past few days

c. Examine the metrics in your snapshot report to gain insights:

   i. Reach: This number represents the maximum number of unique Twitter accounts that received tweets about your search query during this period, based on our robust reach algorithm. Think of reach as the size of your maximum unique potential audience.

   ii. Exposure: Exposure is the number of overall potential impressions that tweets generated in this report. That's the total number of times tweets were delivered to timelines, including repeats. And since replies are only delivered to common followers' timelines, we calculate them as a single impression. The exposure bar graph breaks down how many tweets were sent to users with that many followers.

   iii. Activity: The activity section provides details about the tweets in this report, including:

      1. Total number of tweets analyzed
      2. Total number of unique contributors (people who posted the tweets in this report)
      3. Duration of the period the report covers
      4. Graphical timeline showing tweet volume during the report period
      5. Tweet type breakdown, including how many retweets and replies are included in the report

   iv. Top Contributors: This section shows you the top three contributors – participants whose tweets appear in this report. You'll see the highest contributor for each of three influence dimensions (sometimes, the same person may show up in more than one category):

      1. Highest Exposure – the participant whose tweets generated the most impressions
      2. Most Retweeted – the participant who received the most retweets
      3. Most Mentioned –the participant who was @ mentioned the most times in this report

   v. Top Tweets: This shows the three most retweeted tweets in this report, showing retweet counts for each tweet. This retweet count includes new-style automatic retweets and old-style manual retweets that start with "RT @username." For a tweet to show up as retweeted in this section, the original tweet must be included in the report.

vi. Contributors: A complete list of all contributors (participants) in this report, including how many tweets they posted, how many retweets they received and how many impressions their tweets generated. This list is ordered by impressions.

vii. Tweets Timeline: A full list of all tweets in this report, in reverse chronological order (newest first). This list includes time stamps, as well as start and end times.

d. Understanding reach vs. exposure

i. Reach is the total number of potential unique Twitter users that received tweets about the search term. Exposure is the total number of times tweets about the search term were delivered to Twitter users. TweetReach calls each receipt of a tweet a potential impression.

ii. Reach provides an understanding of the overall effect of your message or campaign. A high reach indicates that a broad base of users found your message interesting and spread it to their followers. It often means that multiple unrelated people learned about your campaign from sources outside Twitter. Conversely, a lower reach means your message is likely only being shared among a smaller group of people who may be more interrelated (e.g., people in the same geographic area).

iii. A high reach will often be combined with a high exposure. Be careful if you notice your campaign has a low reach and a high exposure. It indicates you may have a core group of users who are trying to spread your message by tweeting repeatedly, but that your campaign is failing to take off beyond those users' followers. A high exposure among a small group of people may mean they feel "bombarded" by your message. You may want to alter your message or seek other ways to get more Twitter users involved to avoid oversaturating a small group. For more on this, check out our blog post about the reach:exposure ratio and how to interpret it.

24. Refine your search to deepen your insights

a. Twitter supports a number of advanced search operators and filters that allow you to customize your search query and find the exact tweets you're looking for.

b. Here are a few of the best Twitter search operators and how to use them (with examples):

i. Find one keyword OR another: First, Twitter doesn't require an AND or +

operator to search for multiple keywords. So don't include them. Just type multiple keywords in your query and Twitter will return tweets that include those terms. For example: *social media metrics*.

   1. Sometimes, however, you might want to find tweets that include one keyword or another keyword. Use the OR operator to separate those terms and your report will include tweets that mention one or the other.
   2. You can also chain together multiple keywords to create a more complex query. The OR operator will attach to the word that immediately precedes it, very much like order of operations in algebra. For example, *social media metrics OR analytics* will find tweets that mention social media metrics or social media analytics, because the OR links to the metrics and analytics terms.

ii. @Username queries: There are several ways to learn more about the reach of tweets from a particular Twitter account, depending on the type of information you're looking for.

   1. Tweets to, from and about an account - tweetreachapp: Run a report for a username but don't include the @ symbol. This will return all mentions and tweets (including retweets and replies), as well as all tweets from that Twitter account. This is the most comprehensive set of reach stats for a specific Twitter account.
   2. Tweets to and about an account – @tweetreachapp: Run a report for a username and include the @ symbol. This will return all mentions of an account, but not any tweets from that account. This report will let you know how many people are talking about a certain Twitter account, and the ways they're talking about it (including all retweets, replies, and mentions).
   3. Tweets to an account – to:tweetreachapp: Run a report using the to: operator and a username. Do not use the @ symbol. This report will return only direct replies to that account (where the username is the first word in the tweet). This report is useful for learning more about how people talk to that account.
   4. Tweets from an account – from:tweetreachapp: Run a report using the from: operator and a username. Do not use the @ symbol. This report will return only tweets from that account. This report is useful for

measuring the reach of an individual Twitter account, and for learning more about the kinds of tweets that account is posting.

      iii. Date filters: You can filter your search results to a particular period by adding the since: and until: operators to your search query. Use these date filters to narrow your results. And since you can access up to 1,500 tweets per query, if you run a report for each day of a campaign using date filters, you can find more total tweets. For example, *social media since:2011-04-18* or *@mashable until:2011-04-17*

         1. You can use one or both filters in a query. These dates are inclusive and correspond to the UTC time zone. And no matter what, snapshot reports can only go back five days, so you can't use these filters to access older tweets.

      iv. Exclusions: You can exclude certain keywords from your search by adding a minus sign (-) before the keyword. This will filter out all tweets that include that keyword. This is particularly useful if your company/brand/client/product has a common name and you'd like to exclude mentions of others with that name. For example, *hilton -paris*.

25. Consider using a tool like LIWC to conduct sentiment analysis on the Tweets that are returned in your snapshot report

   a. Download the results of your TweetReach Snapshot report into a .csv file by clicking on the "CSV" icon at the top right of the report page

   b. Open the downloaded file in a spreadsheet like Google Sheets or Microsoft Excel

   c. Find and copy the "Text" content under the header, "Tweets"

      i. Note that the "Text" content is what was actually tweeted and that the other data under the "Tweets" header -- Id, Screen_Name, Time, Retweets, Impressions, URL –is metadata created by Twitter

   d. Navigate to the LIWC homepage and paste the content into the box at the bottom of the page next to the text, "GIVE IT A TRY"

   e. In the drop-down menu labeled, "How would you classify this text (choose one):," select "Social media: Twitter, Facebook, blog" and then hit the "ANALYZE" button

      i. Note that the free LIWC report caps its analysis at 500 words. If you entered more than 500 words, only the first 500 words were analyzed. The paid-for LIWC2015 report actually produces about 90 different output dimensions.

   f. LIWC will return a report containing the following variables:

i. Word count (WC): The raw number of words within a file.

ii. Words per sentence (WPS): The mean number of words within each sentence within the file

iii. Percentage of total words: Most of the LIWC output variables are percentages of total words within a text. For example, imagine you have analyzed a blog and discover that the Positive Emotions (or posemo) number was 4.20. That means that 4.20 percent of all the words in the blog were positive emotion words.

iv. Analytical thinking (Analytic): The analytical thinking variable is a factor-analytically derived dimension based on eight function word dimensions. Originally published as the categorical-dynamic index, or CDI, the dimension captures the degree to which people use words that suggest formal, logical, and hierarchical thinking patterns. People low in analytical thinking tend to write and think using language that is more narrative ways, focusing on the here-and-now, and personal experiences. Those high in analytical thinking perform better in college and have higher college board scores. To learn more about analytical thinking, see Pennebaker, Chung, Frazee, Lavergne, and Beaver (2014).

v. Clout: Clout refers to the relative social status, confidence, or leadership that people display through their writing or talking. The algorithm was developed based on the results from a series of studies where people were interacting with one another (Kacewicz, Pennebaker, Davis, Jeon, & Graesser, 2013). Note that Clout is different from the LIWC2015 Power variable. Power or, more accurately, need for power, reflects people's attention to or awareness of relative status in a social setting. You can have a confident leader who has no interest in other people's standing in the social hierarchy.

vi. Authenticity: When people reveal themselves in an authentic or honest way, they are more personal, humble, and vulnerable. The algorithm for Authenticity was derived from a series of studies where people were induced to be honest or deceptive (Newman, Pennebaker, Berry, & Richards, 2003) as well as a summary of deception studies published in the years afterwards (Pennebaker, 2011).

vii. Emotional tone (Tone): Although LIWC2015 includes both positive emotion and negative emotion dimensions, the Tone variable puts the two dimensions into a single summary variable Cohn, Mehl, & Pennebaker, 2004). The algorithm is built so that the higher the number, the more positive the tone. Numbers below 50 suggest a more negative emotional tone.

g. Note that the LIWC report will include a comparison of “YOUR DATA” to the “AVERAGE FOR SOCIAL MEDIA: TWITTER, FACEBOOK, BLOG” to provide context to your sample

Complete your assignment

26. Pull together your collection of exploratory, data-driven charts and tables
27. Make note of the areas you’ve identified for further exploration and refinement as your data story takes shape
28. Think about additional data (and data sources) you’d like to add to bring further clarity to your analysis or expand into interesting areas