CS3300: Project 1 Write Up

1. A description of the data. Report where you got the data. Describe the variables. If you had to reformat the data or filter it in any way, provide enough details that someone could repeat your results. If you combined multiple datasets, specify how you integrated them. Mention any additional data that you used, such as shape files for maps. **Editing is important! You are not required to use every part of the dataset. Selectively choosing a subset can improve usability. Describe any criteria you used for data selection.** (10 pts)
2. A description of the mapping from data to visual elements. Describe the scales you used, such as position, color, or shape. Mention any transformations you performed, such as log scales. (10 pts)
3. The story. What does your visualization tell us? What was surprising about it? (5 pts)

#1

We got the data from Kaggle's TED Talks database. The database consists of a .csv file containing information on 2550 TED talks. Each talk has associated with it 17 variables:

- Number of user comments (number)
- Description of the talk (string)
- Duration in seconds (integer)
- Event (string)
- Film date (string)
- Languages (string)
- Main speaker (string)
- Name of the talk (string)
- Number of speakers (integer)
- Published date (string)
- Ratings depicting reactions from users (JSON object)
- Related talks (JSON object)
- Speaker occupation (string)
- Tags that describe the topics the talk addresses (array)
- Title (string)
- URL (string)
- Views (integer)

The variables we chose to use were ratings, tags, duration, and views. Before we cleaned the data for the purpose of the visualizations, we created a new CSV file using Microsoft Excel, devoid of the columns we were not using. This was done to reduce the size of the raw data being imported into our JavaScript code. This file is included in our ZIP file as ted_data_dirty.csv.

The file clean_data.html contains the code used to reformat the data obtained from the CSV file. We read the file in using d3, and used a forEach loop to go through every talk in the file. For each talk, we parsed the variables “ratings”, “tags”, “duration” and “views” and used them to construct JSON and CSV files we could use in our visualizations. These were saved in global variables outside of the forEach loop.

To extract the ratings along with their aggregate count over all talks, we parsed the ratings JSON for each talk, and constructed a global JSON object containing a separate JSON for each rating. The keys were “rating” and “count”. This JSON was sorted in descending order by the count, since we intended to represent the data using a bar graph.

To extract the tags along with the total number of talks marked with the tag, we parsed the tag array for each talk, and constructed a similar global JSON object containing a JSON for each tag. The keys were “tag” and “count”.

For the durations, we created four “bins”:

- Under 5 minutes
- 5-10 minutes
- 10-15 minutes
- Over 15 minutes

We originally had intended to have more bins, but since there were very few talks exceeding the 15 minute mark, we chose to go with four. For each talk, calculated the bin by converting the duration in seconds to minutes and taking its floor. Each bin corresponded to an index of a global array called durations. The viewcount for the talk was added to the corresponding index in the array. Finally, the array was converted to a CSV using string concatenation. The total viewcount was also calculated, and was used to convert the bin viewcount into a percentage before export.

At the bottom of the clean_data.html file, we have two functions to export data as JSON or CSV. These were written with help from Stack Overflow. We called these functions on our global variables ratingsJSON, tagsJSON, and durationsCSV to download the files to use for our project.
We chose to use ratings, tags, durations, and viewcounts since they give us valuable insight into how viewers react to TED talks, what topics appear commonly in TED talks, and what durations garner the most views.

#2

Visualization 1: Word Cloud of Popular TED Talk Tags

For the first visualization, we chose to give the viewer an idea of the most common themes in TED talks. We imported in the file tags.json, and looped through it, displaying all tags with a count of greater than 150 in an SVG element.

The **font-size** of the words was scaled on the basis of the count, with the larger words representing larger counts (i.e. more talks tagged with that word). Since the smallest count greater than 150 was 153, and since the largest was 727, we chose to divide the count by 10 to obtain a font-size:

```
font-size = count/10
```

The **X and Y positions** of the words were initially generated using a function that generates a random integer. We ran the visualization a few times until we were happy with the random placement of the words, and then hard-coded that particular set of coordinates into our file.

The **opacity** of the words is also scaled by the count. To get a good gradient from the smallest to the largest word, and ensure visibility of all words, we chose the opacity using the formula:

```
opacity = 1-count/1200
```

This ensured that the larger words were less opaque than the smaller words, allowing all text elements to be clearly visible in the word cloud.

Finally, in order to give the viewer a better idea of the scale, we added a few labels to indicate the words with the largest and smallest counts, and one in between.

Visualization 2: Bar Graph for Aggregate Count of Ratings

The second visualization is designed to give viewers an idea of the reactions TED talks evoke amongst their audience. We imported in the file ratings.json, and looped through it, creating a bar for each of the 14 ratings.

The **x-axis** simply displays each of the ratings by its index in the ratings.json file. Therefore, we used a linear scale for that takes in a value between -1 and 14 (for padding on either side of indices 0 and 13), and outputs a pixel value within the width of our SVG. The ticks for the x-axis were formatted to display the relevant text for each rating instead of numbers, and were transformed to be diagonal for clear visibility.

The **y-axis** displays the total number of people who reacted with the particular rating over all TED talks. Since these vary from approximately 50 thousand to over a million, we used a **natural log-scale** to display these. The scale takes in the raw count, and outputs a pixel value within the height of our SVG. The ticks for the y-scale were formatted to be approximately equally spaced. We used labels that were on the same order of magnitude as the data being displayed on the graph.

To add a fun visual element, we chose SVG format emoji to go with each of the 14 ratings, and added them to the top of the bars 🙂

Visualization 3: Donut Chart of Duration Viewcounts

The third and final visualization is a donut chart intended to give the viewer an idea of what TED talk duration is the most popular amongst viewers. For this, we imported the file durations.csv and looped over the 4 bins in it (under 5 minutes, 5-10 minutes, 10-15 minutes, and over 15 minutes). Each bin has an associated percentage value, which represents the percentage of total views that are within that duration range. These were mapped to the arcs of a pie layout in d3, with the help of this example.

The **arcs** and the inner and outer radius were defined using **d3.svg.arc().** The **arc length** spanned by each of the "bins" was scaled using **d3.layout.pie()**. Additionally, we scaled the **opacity** of the fill color of each arc. This was done using a linear scale that scales percentages from 0 to 75 to an opacity between 0.1 and 1. We chose these values to ensure a good gradient and maintain visibility.

#3

Both of us love watching TED talks, and so we chose to visualize TED's data to uncover surprising information. The three visualizations that we designed are trying to portray three aspects of TED, which lay the groundwork for further exploration:

1. Common themes in TED talks
2. The reactions they evoke from the audience
3. The duration of the talks

We designed three charts to showcase these three aspects of TED:

1. A word cloud showing the most popular tags amongst TED talks
2. A bar graph with a natural log y-scale, showing how popular each "rating" is
3. A donut chart showing time durations of talks along with the percentage of total views they garner

The topics of TED allows us to see the most frequent topics covered by the talks. Although the word cloud only displays the top 25 themes amongst TED talks, we can easily see that technology, science, global issues, culture, and design far outstrip other themes such as medicine and health, education, and creativity by a large margin. The word cloud inevitably reflects the cultural trend of the past decade, as the technology sector booms and the word spread about global issues. On the other hand, other important topics such as humanity, creativity, and social change are significantly less in quantity. The imbalance validates the concerns shared by contemporary scholars on the imbalance between technology and the humanities. The word cloud also reveals that a significant portion of talks are labelled "TEDx", which are independently organized TED talks in locations across the world. This indicates that people are taking interest in TED's way of reaching out to the community, and are willing to take the initiative required to organize these events.

The ratings from the audience reveal the common reaction TED talks evoke from the audience. These give us valuable insight into how TED and TEDx organizers select subjects and speakers for their talks. The most popular reactions from people include "inspiring", "informative", and "fascinating", which is a sensible finding since these are themes people commonly associate with TED talks. There is a sense amongst people that TED talks straddle a fine line between "self-help and knowledge" and "clickbait". This data further validates that ambiguity by showing how the top three reactions far outstrip those such as "courageous", "funny" and "ingenious". Further, the reaction data are mostly positive, which might be since TED provides 9 positive ratings and only 5 neutral to negative ratings.

To make the bar graph more easily decipherable, and easier to visualize, we included emoji to represent each audience rating. This allows the person viewing the chart to get a sense of what it is representing without necessarily reading the x-axis labels, as if they are on stage looking at the crowd.

The final visualization (the donut chart) displays how the duration time of the talks are distributed. The majority of views come from talks 5-15 minute time range. Although this is not surprising, since TED famously tries to limit its speakers to an 18 minute limit, it *is* fascinating to see the very small amount of talks that fall in the "over 15 minutes" range. Further, only 17% of talks fall in the "under 5 minutes" category, which is a popular lightning-talk time limit.

This data shows that a new idea can be effectively communicated to inform and inspire in less than 15 minutes. Duration is definitely a key variable for TED, and plays a large part in their success. On another note, the data also gives us insight into the “bite-sized” nature of TED talks, which follows in the footsteps of content that is popularly considered as clickbait, such as Buzzfeed listicles, Upworthy videos, and more.

In the future, we would like to dig deeper and discover the correlation between tags and reactions. Are TED talks about tech more inspiring than talks about health or education? Which category is the most persuasive? These are questions that require further analysis of the data, and can give us a better view into cultural trends.