Assignment 3 Part A

Declaration: I have read, and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at http://www.tcd.ie/calendar. I have also completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at http://tcdie.libguides.com/plagiarism/ready-steady-write.

i. New York City Weather in 1980 – by Edward Tufte

What?

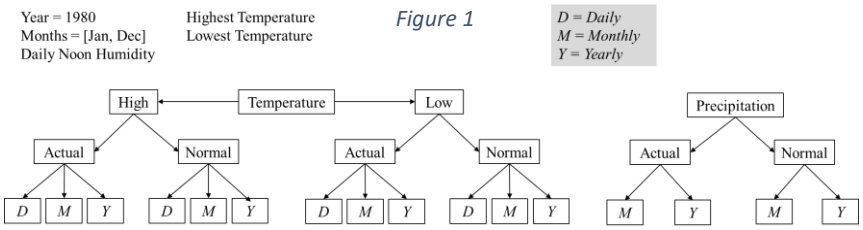

Figure 1

#	Month	Precipitation 1980 "	Precipitation Normal "
1	Jan	1.8	2.1
2	Feb	1.5	2.1
...	...	...	...
25	Dec	1.1	2.3

Table 2

#	Day	Month	Temperature High 1980 °F	Temperature Low 1980 °F	Temperature High Normal °F	Temperature Low Normal °F	Noon Humidity 1980 %
1	1	Jan	42	32	39	27	40
2	2	Jan	41	30	39	27	45
...	...	...	...	...	...	...	...
366	31	Dec	32	14	39	27	28

Table 1

Figure 1 (created in MS PowerPoint) identifies all data that this chart is displaying. The time-series dataset using which this graph was plotted is likely to have comprised 2 tables as given (created on MS Excel), where each row item of Table 1 is comprised of attributes "Day", "Month", "Temperature High 1980 °F", "Temperature Low 1980 °F", "Temperature High Normal °F", "Temperature Low Normal °F" and "Noon Humidity 1980 %" while that of Table 2 is composed of attributes "Month", "Precipitation 1980 " " and "Precipitation Normal " ". Temporal attributes day and month are ordinal here since data is ordered based on them but no differences between years/months (time interval computation) is relevant here. All temperature related attributes and humidity are continuous and quantitative with an interval measurement level (no absolute 0) while precipitation related ones, although continuous and quantitative, is a measurement (absolute 0 = no precipitation). Highest and lowest temperatures are derived from high/low temperature values (compute max/min). For temperature and humidity data, it's plausible for monthly and yearly values to have been derived by averaging daily ones. Precipitation for the entire year may also have been derived by aggregating (summing up) monthly precipitation values.

Why?

Upon targeting all data, users can discover/present basic weather conditions across various months in the year 1980 in New York City. They may lookup/compare approximate temperature, humidity, and precipitation values using the y axes. Subplots with yearly data and monthly plot divisions allow users to summarize values at per month/year scales. Targeting temperature and precipitation allows comparison of 1980 values not just across months but also against normal values. Temperature attributes also capture clear trends and correlation (between pairwise combinations of actual/normal high/low values) across months that users may explore/compare. This visualization aimed at all average users (not just analysts) is explanatory since it clearly shows how temperatures in 1980 were largely normal except few extremes.

How?

An area range plot displays high and low 1980 temperatures (area between high and low temperature filled in) with multiple trend lines plotted over it capturing normal temperature values. Humidity values are also visualized via an area chart. Precipitation and annual temperature values are represented using a multiple column bar plot with categories encoded using a combination of brightness and pattern with size (height) of bars, bottom-aligned at 0, encoding precipitation in inches or temperature in °F respectively. The visualization is partitioned horizontally to encode months by position. Vertical partitions separate temperature, precipitation, and humidity subplots. Vertical position also encodes magnitude of every datapoint for each attribute. All attributes are ordered in time (from January to December). Single values like highest/lowest temperature extremes pop out via callouts.

Discussion

Temperature/humidity and precipitation related quantitative attributes were aptly encoded with sufficient accuracy using position and size (bar height) respectively. Position from left to right is also effectively employed to capture passage of time (month order). Association of values to 1980 v/s normal data is efficiently presented using clearly discriminable brightness (shades of grey with no hue) and texture (normal = diagonal lines pattern while 1980 = solid pattern). Redundant use of brightness and texture to encode these categories aid in improving distinguishability of normal trend lines on the temperature plot wherein colour (higher brightness/lighter shade) allows for clear visibility although lines are too narrow for texture to be visible. Since there are only 2 categories, this redundancy does not lend itself to added confusion. All axes, subplots, and callouts are labelled in sufficient detail. All channels show good separation and do not interfere with each other. Only criticism can be that an area chart may have been unnecessary to represent non-accumulative (adding up does not give cumulative % here) % values related to humidity. A simple line chart would have sufficed. That said, the shaded area plot is aesthetically pleasing (subjective). Heterogeneous parallelism can be observed here wherein isomorphism and juxta positioning of normal over actual temperature trends highlights connections. Overall, this visualization effectively captures the dataset using clever combinations/arrangement of standard idioms.

ii. Music, Google and Books – by Federica Fragapane

What?

Figure 2 (created in MS PowerPoint) identifies all data that this chart displays. The dataset using which this graph was plotted is likely to have comprised 2 tables where each row item is comprised of attributes corresponding to columns of Tables 3 and 4 (created on MS Excel). Here, attributes related to name of "Artist/Group", "Country", "Continent", and "Most Interested Country" are qualitative with no order (nominal categorical). The temporal attribute "First Release Year", though represented using numbers, is categorical and ordinal in nature (computing differences between years is not too useful here). Count related attributes like "No. of Albums/Books/Artists that a country is first in terms of interest for" have discrete quantitative values. The "no. of artists that a country is first in terms of interest for" attribute is likely derived from data corresponding to columns "Artist/Group" and "Most Interested Country" in Table 3 (compute artist count per country).

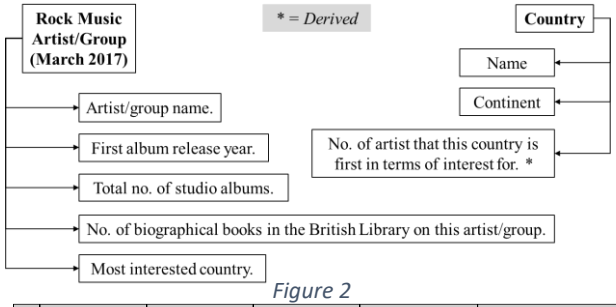

Figure 2

#	Artist/Group	No. of Books	Fist Album Release Year	No. of Albums	Most Interested Country
1	Elvis Presley	97	1956	23	United Kingdom
2	Chuck Berry	6	1956	27	Uruguay
...	...	...	...	...	...
40	Foo Fighters	5	1995	8	New Zealand

Table 3

#	Country	Continent
1	Canada	North & Central America
2	United States	North & Central America
...	...	...
12	New Zealand	Oceania

Table 4

Why?

This visualization is exploratory since all data is presented as is, for users to draw conclusions from. Users are not shepherded towards any conclusion in particular. Via this plot, users may explore/present popularity of rock music artists or groups as indicated by books written about them as well as popularity of rock music in various countries. The visualization is engaging with pleasing colours and an interesting take on a lollipop chart. Moreover, the topic of visualization is related to rock music which many find interesting. Thus, users may also enjoy this visualization. Users can lookup Table 3 attribute values related to any available artist of choice. They may also browse connections (dependencies) between countries and artists in terms of interest. New information like continent with largest interest in rock music etc, can be derived from continent to artist connections. Extreme values like artist with highest/lowest no. of books written about them can easily be identified from size of circles. Users may also use this visualization to identify/derive values that can be leveraged to summarize data (e.g. mean no. of artists that each continent is interested in, max books written about an artist, etc). Users can also compare no. of books published v/s no. of albums to derive a general understanding of impact of songs of an artist/group.

How?

The position channel is used to aptly encode the ordinal feature "year of 1st album release" (sorted horizontally in increasing order from left to right) and the quantitative one "no. of albums" (sorted vertically in increasing order from bottom to top) related to each artist. Size of country parabolas and circles is used to convey discrete quantitative count of artists that a particular nation had the highest interest for and no. of books about an artist respectively, quite well. Each continent is associated with a particular colour (encoded using hue channel). Colour coded (as per continent) solid lines connect artists to the country most interested in them. Horizontal dotted lines from each artist intercept the y axis at the tick corresponding to no. of albums under that artist's belt. Overall, this visualization is a creative blend of elements from the lollipop chart, node-link idiom and bubble chart. Parabola height encoding no. of artists per country is also reminiscent of a bar plot (here, parabolas instead of bars).

Discussion

All graphical elements are well labelled, and a clear, descriptive key is provided. This makes it easy to understand unconventional design choices made (parabola, curved lollipop chart with varying sized bubbles). Superposition of bigger bubbles over smaller ones is handled by adding transparency to make sizes more discernible. Use of size and hue together can cause slight interference (darker circles appear bigger). Also, the lighter hued circles are slightly less striking. These effects are, however, not too overpowering. Grouping countries by their continents (ordinal) through both sorted bottom-aligned position on x axis and hue of parabolas/lines/circles helps make the visualization more readable and alleviate difficulties that criss-crossing lines introduce in tracing connections between artist bubble and country parabola while at the same time also encoding continent attribute values. Small black dots inside bubbles that can easily be traced to the y-axis curtsy of dotted lines and the dark lines dividing the parabolas into parts equal to the no. of artists they represent, allows this visualization to encode said attributes with high accuracy. Differences in 2D area is generally hard to visually perceive, but by labelling each bubble with the no. of books it represents, the visualization makes it possible for the reader to precisely ascertain values that circle encodes while also being able to visually pick out biggest and smallest (extreme) values more intuitively based on bubble size. The creative use of non-standard idioms is fully justified here given that this single plot manages to effectively encode a high degree (7) of attributes in an attractive and engaging fashion.

iii. Growing Family – by Nathan Yau

What?

This visualization plots no. of births corresponding to 1000 mothers and their ages at the time of birth of each of their children. The overall percentage of births per age group is also displayed, as is the no. of mothers with y no. of total births by the time she gets to age x. All 4 main attributes "No. of Births", "Age of Mother", "Percentage of Births by Age Group" and "Count of mothers for a given no. of total kids by a certain age" are quantitative in nature but their measurement levels vary. Count related attributes are discrete while the age of mothers and the percent of births by age group is at the "measurement" data level (absolute 0). The dataset used for this visualization is likely a set of 1000 mothers where each item is a list of ages of the mother at the time of birth of each of her children sorted by birth order. Attributes "no. of births", "percent of births by age group" and "total counts for a given age and no. of kids" may be derived from said dataset.

Why?

The plot along with supporting text is explanatory in nature since it clearly states how most women surveyed gave birth to their first child around age 20 with the distribution shifting older for more children. The visualization also displays a wide variation in no. of children although, most people tend to have 2 to 3 children. Users can discover percentage of births per age group. Parents, to-be parents, and women in general are likely to find the results intriguing and hence may enjoy this visualization. Targeting the entire dataset, users can explore/present how many children most mothers tend to have and can easily spot outliers (like the annotated 12 kids by age 30 case). Targeting just the no. of births and age reveals the distribution of when women tend to have their n^{th} (first, second, …, 12^{th}) child. Users can lookup features like how common it is to have had a certain no. of kids at a certain age, locate others like oldest/youngest age at which to have first child and browse through general birthing trends. The plot also allows for comparison of features like popularity of having different no. of kids.

How?

The position channel is used to encode age (horizontally increasing from left to right) and no. of births (vertically increasing from top to bottom). Size of circles encode no. of mothers with y no. of children in total by age x. Motion is used to encode timeline (ages at which she had each of her children in increasing birth order) of each mother. Vertical position is also used to separate the births and ages plot from the plot with percentages. Percent of births by age group is displayed by means of a table while the birthing timelines of each of the 1000 mothers is visualised as an animated scatter plot where each mother is represented by a black dot that travels along the x axis starting from the top left corner of the plot and moving towards the right, corresponding to increase in age of the mother and drops down one unit along the y axis every time the mother gives birth (at a particular age). In few cases, the dot was observed to drop by more than one unit in a single fall (not staggered). This is expected to correspond to birth of multiple children in one birth (e.g. twins, triplets).

Discussion

The use of an unconventional scatter plot with static as well as moving dots is justified here, since displaying both total no. of women who had x no. of total children by age y while taking data from all surveyed women into account as well as no. of children had by each woman (one at a time) at different points in their lives, is a complex task given that different women have varying no. of children (wide range [0, 12]) at varying ages. That is to say, this plot manages to capture both dataset wide and item wide attribute values in a single innovative plot. The colour differences among moving dots and static ones as well as the temporary size increase of moving dots upon y axis drop along with temporary trailing lines that they trace, make them easy to follow visually to count no. of kids each woman (black dot) had, at what age they had each one, and how many of them they had together at a time. The quantitative attributes age and no. of kids were aptly encoded using position. Encoding of no. of mothers with y total children by age x through area of green circles while useful to view general distributions and extreme values, is not accurate as its not possible to gauge precise counts from area of circles. All channels are used together with great separability. Although there is some superposition of green dots, this is compensated for by making them translucent. This plot boasts minimal clutter which is a significant achievement considering that it plots data corresponding to a fairly large dataset with 1000 timelines.