***Student Name:** Gayathri Girish Nair* | ***Student No:** 23340334* | ***Course:** CS7DS4 Data Visualization 2023-24*

Assignment 3 Part B

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.

What?

Attribute	Meaning	Datatype
cut	Craftsmanship/quality indicator: 'Fair' < 'Good' < 'Very Good' < 'Premium' < 'Ideal'.	Ordinal qualitative.
color	Yellowness ranging from D to Z in increasing order.	Ordinal qualitative.
clarity	Degree of absence of inclusions/blemishes: 'IF' < 'VV1' < 'VV2' < VS1 < VS2 < SI1 < SI2 < I1.	Ordinal qualitative.
carat	Weight of the diamond. One carat = 0.2 g.	Continuous Quantitative Measurement
depth	Total depth % = z/mean(x, y) = 2*(z/(x + y))	Continuous Quantitative Measurement
table	Table % = width of the top of the diamond relative to its widest point.	Continuous Quantitative Measurement
price	Price in US Dollars.	Continuous Quantitative Measurement
x	Length in millimetres.	Continuous Quantitative Measurement
y	Width in millimetres.	Continuous Quantitative Measurement
z	Depth in millimetres.	Continuous Quantitative Measurement

Table 1

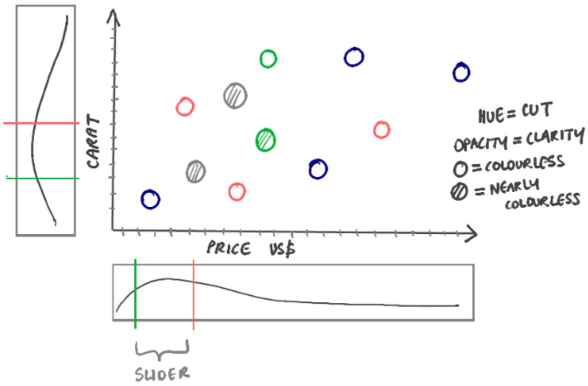
The dataset used is the diamond dataset from list of suggestions on blackboard. Table 1 (created in Excel) captures the attributes comprising this tabular dataset, their meaning, and data type. (Cullen Jewellery, 2022)**.** Here, the colour attribute's 7 possible values (D to J) can be reduced to 2 since categories D to F and G to J mean 'colourless' and 'near colourless' respectively (Gemological Institute of America, color, n.d.). Also, attribute clarity with 8 possible values can be reduced to 5 (IF, VVS, VS, SI, I1) since differences between VVS1 and 2, VS1 and 2 as well as SI1 and 2 are very small (Gemological Institute of America, clarity, n.d.).

Why?

Some useful tasks that users might perform on an exploratory visualization based on this dataset would be: **1.** Discover/present attributes w.r.t price. **2.** Explore distribution of price and carat attributes of diamonds. **3.** Browse likely correlations among attributes. **4.** Identify outliers/extremes (price, carat, clarity, cut). **5.** Lookup attribute values in particular price/carat ranges.

How?

Attributes 'x', 'y', and 'z' convey size/volume which is already captured by attribute 'carat'. Also, depth and table are useful only when shape of the diamond is known (Gian, 2015), which isn't the case here. Hence, plotting just cut, carat, clarity, colour and price removes added complexity through scrapping least meaningful attributes. A scatterplot idiom can be used with quantitative attributes price and carat encoded along x and y axis (position) respectively. Categorical attributes cut (5 categories), clarity (5 categories) and colour (2 categories) may be encoded using hue (blue < green < yellowish orange < red < purple), opacity (saturation inversely proportional to clarity) and texture (no pattern = colourless, diagonal lines = nearly colourless). Rectangles beside axes shall plot distribution of price and clarity via line plots while also acting as sliders that allow users to interact with the plot and filter it to declutter and manage high size complexity (53940 items). Dots change (motion channel) to reflect filter changes.



Discussion

The biggest challenge with this dataset is size which was managed by enabling filtered views of data. Quantitative attributes were aptly encoded using position. Hue although not always suitable for ordinal data is acceptable here since there are just 5 categories and colours may be chosen carefully to be as distinguishable as possible and convey order intuitively (e.g. cool to warm colours to indicate increase in order). Occlusion due to overlapping of dots is managed by enabling opacity which also serves to encode clarity attribute. Once again, since there are only 5 ordinal clarity categories, they would be sufficiently and distinctly captured within discriminability offered by the saturation channel. With only 2 categories for the colour attribute, texture channel here does not demand too much added mental exertion. The unconventional sliders serve to both inform about general distribution of the quantitative attributes and provide handles via which users may interact with the plot. Although removal of few attributes may have led to a slight reduction in expressiveness, this decision likely improves effectiveness as users shall not be misled/confused by presence of 4 more attributes that despite not contributing to new useful information significantly furthers complexity of the already dense visualization. The reliance on memory as users slide filters along the axes is less than ideal, but necessary here to manage/simplify large amounts of data. This incorporation of the motion channel via user interaction is what makes it possible to keep data to ink ratio fairly small. Encoding using hue, saturation and texture may contribute towards some interference that can reduce separability. But this is not expected to be too detrimental as long as hues chosen are distinct and easily visible/perceived.