Beuth University of Applied Sciences Berlin – Data Science

Project 1 for Data Visualization ST 2019

Color Measurements

Authors: Florian Becker
Rupali Sharma
Kerstin Wagner

Github: https://github.com/SharmaRupali/DataViz

Date: 26/05/19

Table of Contents

I. Abstract

Color cards when printed may show significant variations depending on the printers used. Here, the goal is to evaluate color cards determining skin color in comparison to a master card from Douglas. ΔE distances and cosine.similarity have been used as different methods to make comparisons and later conclude the best method to use along with a range of visualizations (histograms, boxplots, violin plots, and density plots) in such scenarios. Statistics and visualizations in different aspects show the boxplots to be the most readable along with a wider range of variations in cosine.similarity in comparison to other plots and ΔE distances.

II. Introduction

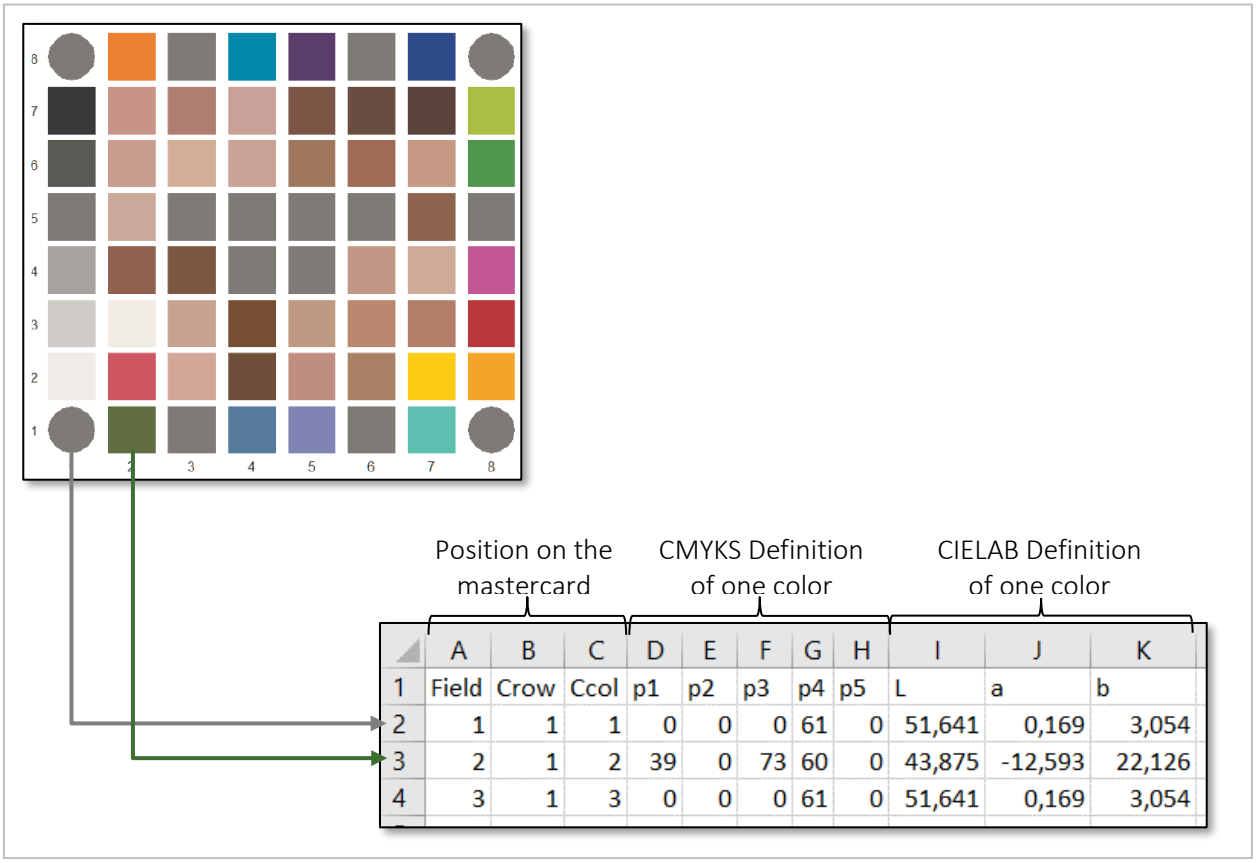
Our first project in Data Visualization is about evaluating printed color cards in comparison to a master color card. The cards are from Douglas for determining skin color.

We have two datasets:

- MasterColorCard.csv (MCC) contains intended colors for color card production
- LabMeasurements-Color-Card.csv (LMCC) contains the measurements for 13 sheets with each having 42 color cards printed on, in 7 rows and 6 columns

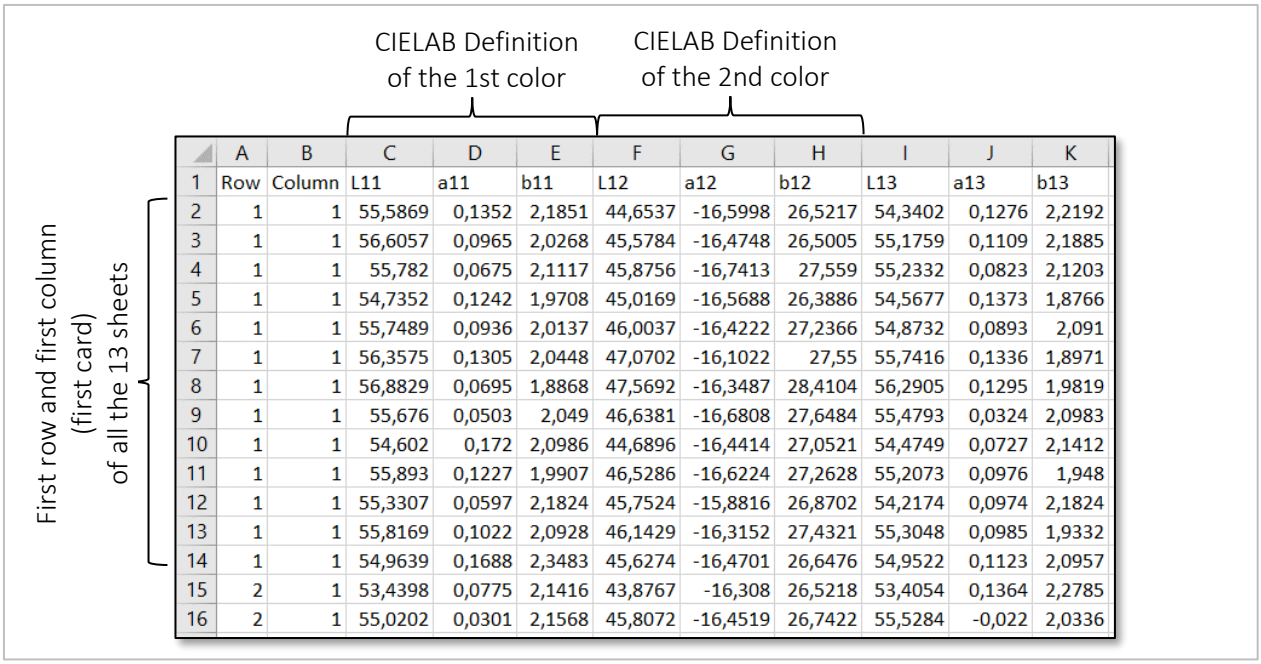
Figure 1 and Figure 2 show the structure of the datasets for a better understanding of the data:

- The MCC gives us the colors in two color spaces: CMYK including a special skin color for the print and calculated CIELAB for comparing the measurements with the LMCC.
- The LMCC only uses the CIELAB color definition.



	A	B	C	D	E	F	G	H	I	J	K
1	Field	Crow	Ccol	p1	p2	p3	p4	p5	L	a	b
2	1	1	1	0	0	0	61	0	51,641	0,169	3,054
3	2	1	2	39	0	73	60	0	43,875	-12,593	22,126
4	3	1	3	0	0	0	61	0	51,641	0,169	3,054

Figure 1: Structure of MasterColorCard.csv



	A	B	C	D	E	F	G	H	I	J	K
1	Row	Column	L11	a11	b11	L12	a12	b12	L13	a13	b13
2	1	1	55,5869	0,1352	2,1851	44,6537	-16,5998	26,5217	54,3402	0,1276	2,2192
3	1	1	56,6057	0,0965	2,0268	45,5784	-16,4748	26,5005	55,1759	0,1109	2,1885
4	1	1	55,782	0,0675	2,1117	45,8756	-16,7413	27,559	55,2332	0,0823	2,1203
5	1	1	54,7352	0,1242	1,9708	45,0169	-16,5688	26,3886	54,5677	0,1373	1,8766
6	1	1	55,7489	0,0936	2,0137	46,0037	-16,4222	27,2366	54,8732	0,0893	2,091
7	1	1	56,3575	0,1305	2,0448	47,0702	-16,1022	27,55	55,7416	0,1336	1,8971
8	1	1	56,8829	0,0695	1,8868	47,5692	-16,3487	28,4104	56,2905	0,1295	1,9819
9	1	1	55,676	0,0503	2,049	46,6381	-16,6808	27,6484	55,4793	0,0324	2,0983
10	1	1	54,602	0,172	2,0986	44,6896	-16,4414	27,0521	54,4749	0,0727	2,1412
11	1	1	55,893	0,1227	1,9907	46,5286	-16,6224	27,2628	55,2073	0,0976	1,948
12	1	1	55,3307	0,0597	2,1824	45,7524	-15,8816	26,8702	54,2174	0,0974	2,1824
13	1	1	55,8169	0,1022	2,0928	46,1429	-16,3152	27,4321	55,3048	0,0985	1,9332
14	1	1	54,9639	0,1688	2,3483	45,6274	-16,4701	26,6476	54,9522	0,1123	2,0957
15	2	1	53,4398	0,0775	2,1416	43,8767	-16,308	26,5218	53,4054	0,1364	2,2785
16	2	1	55,0202	0,0301	2,1568	45,8072	-16,4519	26,7422	55,5284	-0,022	2,0336

Figure 2: Structure of LabMeasurements-Color-Card.csv

The first aspect that we want to cover in the project, is the likeness between colors, cards and sheets with the MCC as given baseline, for example:

- Are the distances and similarities nearly the same for cards on one sheet?
- Are the distances and similarities nearly the same for one color on every sheet?
- Are any patterns recognisable?
- Are any anomalies visible?

The second aspect is to get a better understanding of different calculations and visualizations of the likeness: what works good and is easily accessible and what is hardly readable and interpretable?

- We calculate the distance by ΔE and the similarity by cosine similarity. For more details about the algorithms we have chosen, see “III.A Algorithms”, p. 4).
- The calculated results are visualized in histograms, boxplots, density and violin plots (see “IV. Results”, p. 6)

III. Materials and Methods

A. Algorithms

1. Delta E 2000

The CIELAB color definition allows to calculate perceptual differences. “ΔE is a metric for understanding how the human eye perceives color difference” [1]. Because Delta E 2000 is the current state-of-the-art, we used it for calculation of distances.

For the results are different interpretations possible. Figure 3 shows two: a more wide-ranged overview on the left [1], and one focussed on the printing quality on the right (errors between ΔE 2 to 4 are considered accurate implementations) [2].

Lots of relevant factors: for example, the sample size or the experience of the observer, but also the time somebody needs to recognize a difference is to be considered [2].

Delta E	Perception
<= 1.0	Not perceptible by human eyes.
1 - 2	Perceptible through close observation.
2 - 10	Perceptible at a glance.
11 - 49	Colors are more similar than opposite
100	Colors are exact opposite

Color Difference	Impact
< 0.2	Not visible
0.2 - 1.0	Very low
1 - 3	Low
3 - 6	Medium
> 6	High

Figure 3: Two color difference ratings

2. Cosine similarity

The cosine similarity calculates the angle between two vectors, the result ranges from −1 (exactly opposite) to 1 (exactly the same).

B. Datasets

We decide to shrink the data we are working with and focus on the skin colors in the middle of the card and to leave the border color patches out, because they seem irrelevant for determining a skin color.

The four grey patches in the middle of the master card represent the hole for the customer's skin – that's why they are also not relevant.

Figure 4 shows the remaining 32 of 64 colors that we have used for the comparisons.

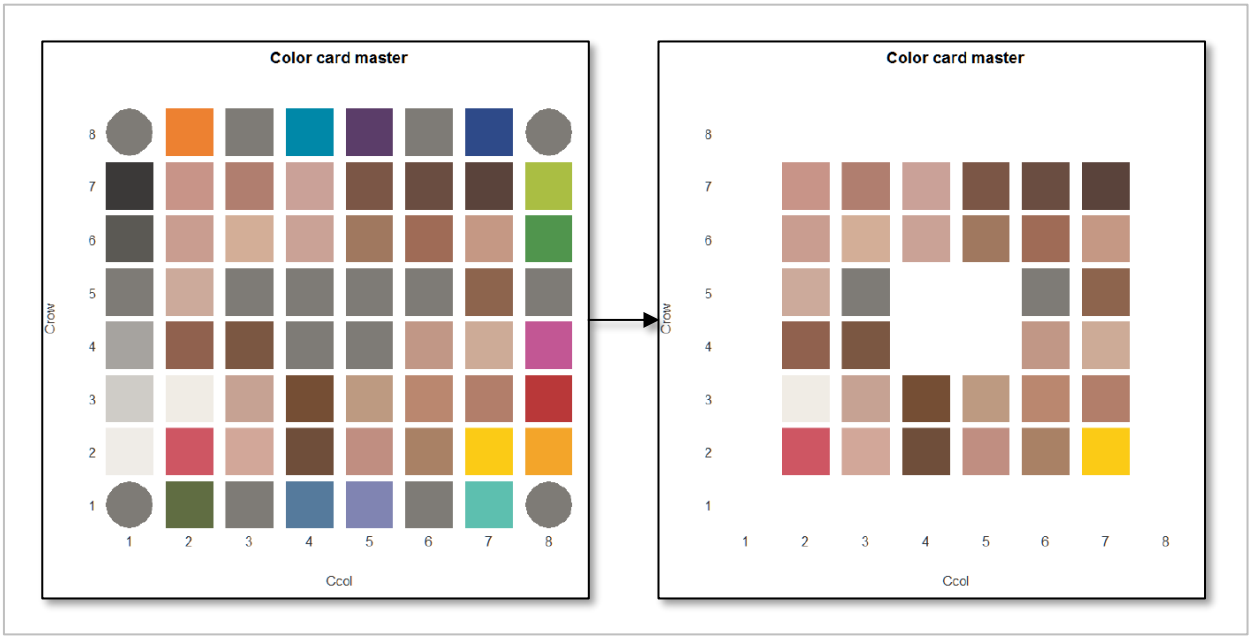


Figure 4: All colours vs. tested colors

C. R Packages

We use some additional packages for the project:

Package	Used function	Reason
data.table	fread	Fast import of the csv files with automatic controls detection, e.g. separator [3]
colorscience	deltaE2000	Calculates the distance between two colors in CIELAB [4]

tcR	cosine.similarity	Calculates the similarity between two vectors [5]
vioplot	vioplot	Produces violin plot(s) of the given grouped values with enhanced annotation

D. Procedure

The procedure for distances and similarities is the same:

- Create the necessary data.frames with calculated distances / similarities.
- Plot and save the different charts as png in the folder "Images".
- Use the colors of MCC for plotting color patch relevant charts.
- Use the same colors for the sheets in the charts.

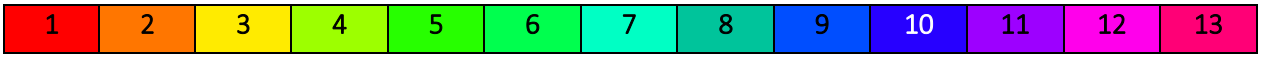


Figure 5: Color legend for sheets

IV. Results

For describing the deviation, we choose minimum, maximum, mean, median and number of outliers.

Section A shows these statistics for different data groups, which can be found partially in the charts.

A. Statistics

1. For all sheets by colors

	Distance					Similarity				
	Min	**Max**	**Mean**	**Median**	**Out.**	**Min**	**Max**	**Mean**	**Median**	**Out.**
All	0.047	26.759	2.476	1.383845	549	1.092	2.638	2.009	2.036915	0

Table 1: For all sheets by colors

Noticeable is the wide range of distances. Interesting is also the high number of outliers for the distance, but no outliers for the similarity.

2. For all sheets by card

	Distance					Similarity				
	Min	**Max**	**Mean**	**Median**	**Out.**	**Min**	**Max**	**Mean**	**Median**	**Out.**
All	2.207	2.970	2.476	2.482172	3	1.989	2.027	2.009	2.009369	6

Table 2: For all sheets by card

If the colors are clustered before, different values result for most of them.

3. For each sheet by card

	Distance					Similarity				
Sh.	**Min**	**Max**	**Mean**	**Median**	**Out.**	**Min**	**Max**	**Mean**	**Median**	**Out.**
1	2.256	2.603	2.432	2.42836	0	2	2.017	2.009	2.00949	0
2	2.283	2.719	2.533	2.55824	3	1.992	2.012	2.002	2.00277	0
3	2.284	2.749	2.456	2.42202	0	2.004	2.019	2.012	2.01245	0
4	2.31	2.6	2.451	2.4422	0	2.003	2.02	2.012	2.01164	0
5	2.237	2.603	2.429	2.42427	0	1.999	2.019	2.01	2.00976	0
6	2.242	2.784	2.499	2.50957	0	1.989	2.006	1.999	1.99975	0

	Distance					Similarity				
Sh.	**Min**	**Max**	**Mean**	**Median**	**Out.**	**Min**	**Max**	**Mean**	**Median**	**Out.**
7	2.238	2.693	2.464	2.4692	0	1.998	2.015	2.006	2.00731	0
8	2.254	2.767	2.526	2.54369	0	1.998	2.014	2.006	2.00711	0
9	2.3	2.97	2.52	2.50883	0	2.006	2.027	2.015	2.01633	0
10	2.29	2.623	2.479	2.48299	0	1.998	2.022	2.01	2.0094	0
11	2.316	2.699	2.499	2.5047	0	2.002	2.023	2.014	2.01329	0
12	2.326	2.671	2.465	2.47758	0	1.997	2.016	2.007	2.00683	0
13	2.207	2.68	2.441	2.4664	0	2.003	2.02	2.012	2.0132	0

Table 3: For each sheets

Are the cards clustered by sheet, the values are only slightly different than before. The three distance outliers can be found again; the six similarity outliers disappear.

4. For each colors

	Distance					Similarity				
Col.	Min	Max	Mean	Median	Out.	Min	Max	Mean	Median	Out.
1	0.256	3.11	1.271	0.6215896	1	2.576	2.625	2.599	2.599173	0
2	0.216	1.998	0.66	3.565043	9	1.719	1.766	1.744	1.743961	1
3	2.622	4.773	3.567	0.6839234	5	2.485	2.525	2.503	2.501948	0
4	0.24	2.27	0.748	1.512918	11	1.957	2.018	1.991	1.992313	3
5	0.719	2.661	1.481	1.055471	1	2.086	2.133	2.113	2.114288	2
6	0.393	1.837	1.046	25.31076	3	2.056	2.081	2.066	2.065985	4
7	23.855	26.759	25.294	1.234916	6	1.371	1.398	1.386	1.386039	7
8	0.579	3.186	1.223	4.505533	4	1.667	1.722	1.692	1.691421	5
9	3.538	5.583	4.503	0.8432311	9	2.58	2.638	2.61	2.608857	0
10	0.098	2.226	0.825	1.229698	2	1.799	1.847	1.826	1.826865	3
11	0.562	2.237	1.238	1.811967	4	2.156	2.22	2.193	2.193381	3
12	1.044	3.01	1.849	2.562715	3	2.199	2.267	2.233	2.234101	3
13	1.437	4.658	2.63	3.031098	6	2.462	2.511	2.488	2.488132	0
14	2.31	4.2	3.047	0.52415	10	2.42	2.461	2.441	2.440897	0
15	0.047	2.145	0.591	0.6040576	8	1.838	1.896	1.868	1.867933	0
16	0.069	1.535	0.643	1.037628	1	1.646	1.688	1.669	1.669991	2
17	0.528	2.006	1.077	1.936186	5	1.582	1.627	1.607	1.608216	15
18	0.229	5.185	2.018	2.018438	0	1.092	1.127	1.108	1.107621	2
19	0.285	4.976	2.067	2.332186	0	1.092	1.124	1.107	1.107537	0
20	1.517	4.587	2.416	0.8583082	4	2.375	2.43	2.402	2.402575	0
21	0.208	1.978	0.887	1.208014	2	1.794	1.843	1.821	1.821359	4
22	0.842	1.866	1.226	0.8678517	2	1.686	1.729	1.706	1.706272	1
23	0.412	1.823	0.927	1.545201	2	1.681	1.723	1.705	1.70529	2
24	0.825	3.21	1.59	2.303437	7	2.146	2.197	2.176	2.176179	2

	Distance					Similarity				
Col.	Min	Max	Mean	Median	Out.	Min	Max	Mean	Median	Out.
25	1.187	4.615	2.427	1.014791	1	2.44	2.492	2.468	2.468607	0
26	0.596	2.905	1.094	0.7054563	29	1.878	1.954	1.92	1.920033	1
27	0.237	1.805	0.784	1.269947	12	1.932	1.982	1.957	1.957615	0
28	0.418	2.689	1.31	0.968093	11	2.112	2.178	2.147	2.14874	4
29	0.435	1.948	1.032	2.911953	0	1.671	1.711	1.692	1.692088	0
30	2.09	4.155	2.935	3.11869	7	2.414	2.455	2.436	2.436329	2
31	2.217	4.396	3.115	3.715516	3	2.319	2.371	2.348	2.349743	0
32	3.131	4.387	3.729	0.6215896	3	2.238	2.279	2.26	2.260672	6

Table 4: For each color

Regarding single colors are bigger differences visible, e.g. the already mentioned wide range of distances.

B. Likeness based on cards

Base: mean of distances / similarity over all colors of one card

Red line in charts: mean of distances / similarity over cards (see Table 2, p.6)

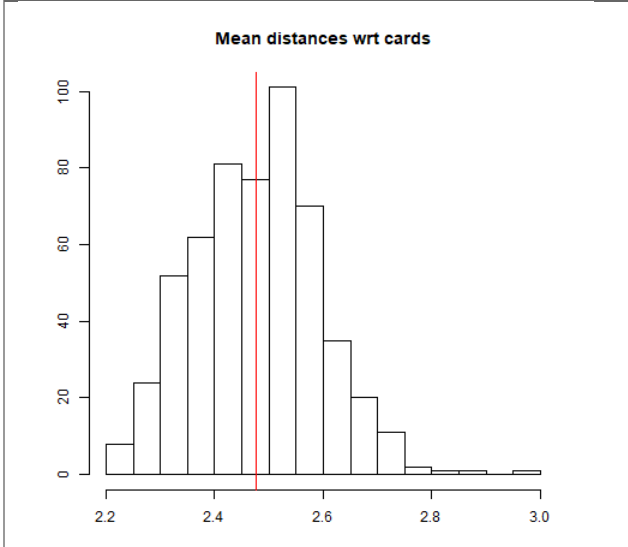
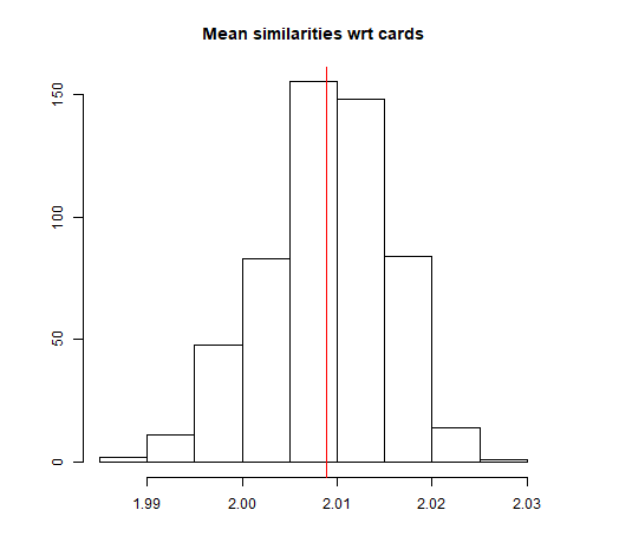
Histogram: contains a combined distribution over all 13 sheets with the distribution of the mean distances of each card.

Box, violin and density: contain 13 colors for 13 sheets where each box/violin/density_line represents the distribution of the mean distances (mean of distances over all colors) of each card on the sheet from the master card: 42 points (for 42 cards) for each sheet.

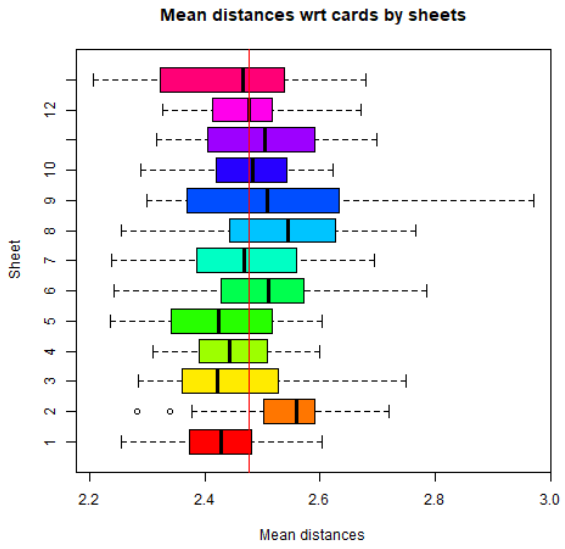
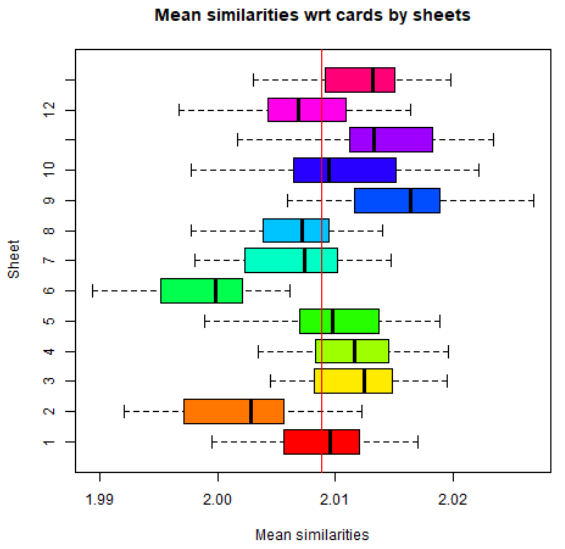
The similarity plots are much more varied than the distance ones.

The violin plots have similar readability in terms of box region and means to the boxplots, but the outliers are not visible because the violin includes them, whereas in the boxplots, the outliers are clearly visible, so between these two types of blots, boxplots are better for outlier detection, otherwise they work similarly.

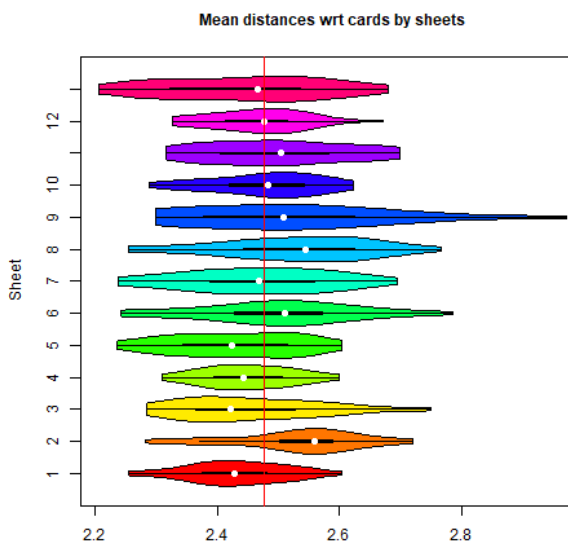
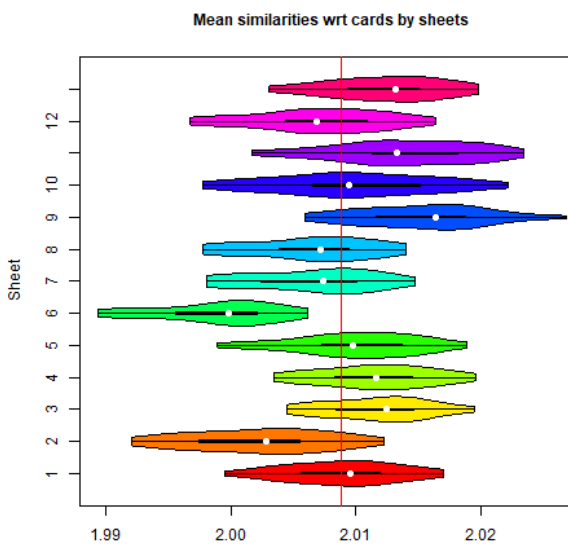
1. Histogram

Distance	Similarity
	

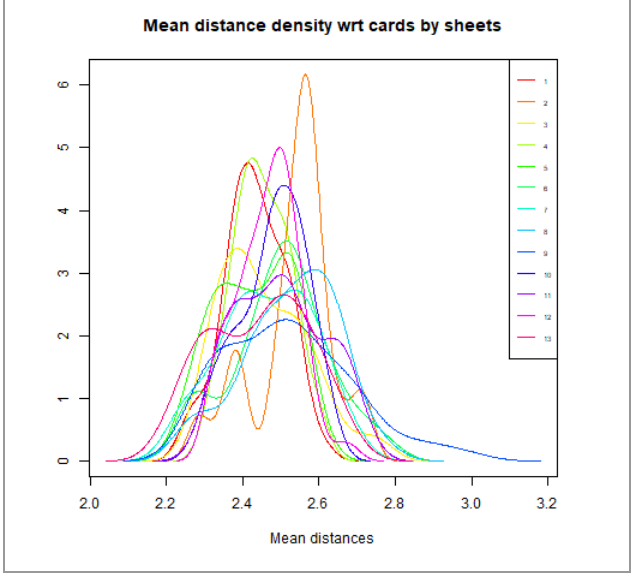
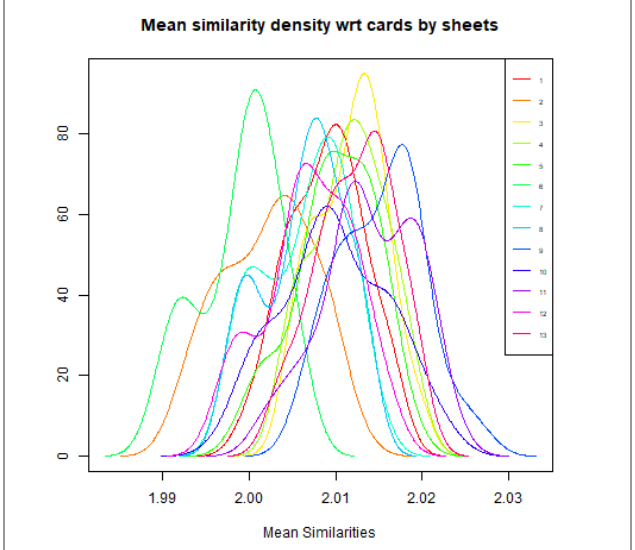
2. Boxplot

Distance	Similarity
	

3. Violin Plot

Distance	Similarity
	

4. Density plot

Distance	Similarity
	

C. Likeness based on colors

Base: distance / similarity of one color

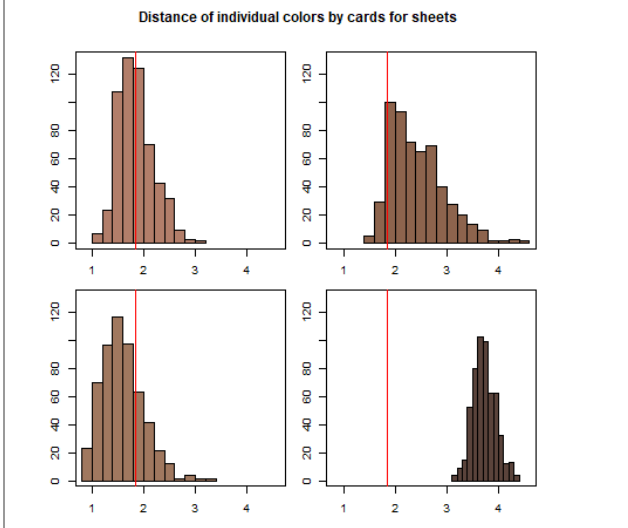
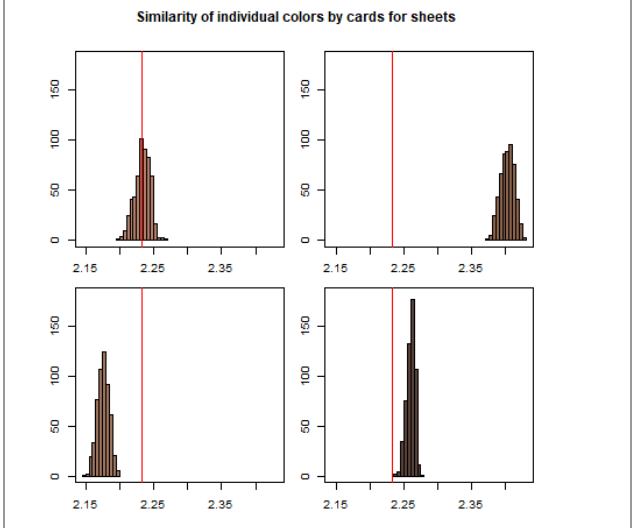
Red line in charts: mean of distances / similarity over the colors shown in the plots

Four colors have been selected (37, 57, 65, 77): four plots on one representing the plots for resp. colors

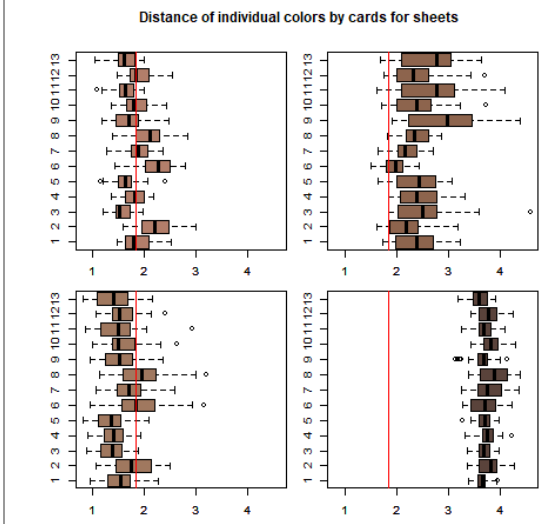
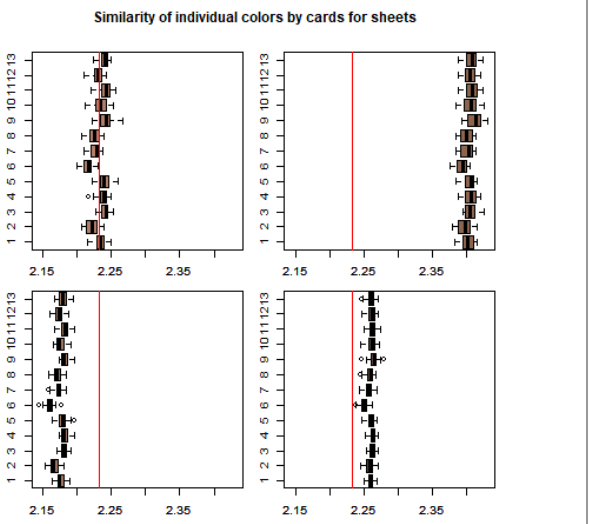
Histogram: each box represents the distribution of the distances of resp. color on each card on the sheet from the resp. color on the master card: 42 points (for 42 cards) for each sheet

Box, violin and density: contain for each sheet the distance/similarity between those colors. Each card makes for a point in the box/violin/density plot.

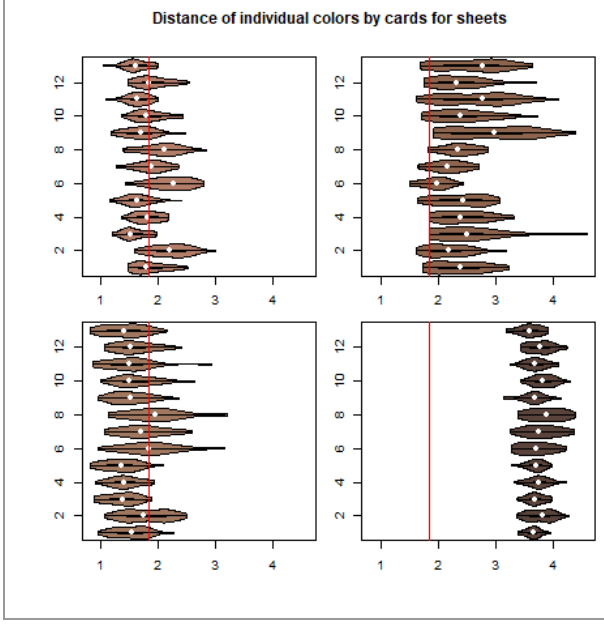
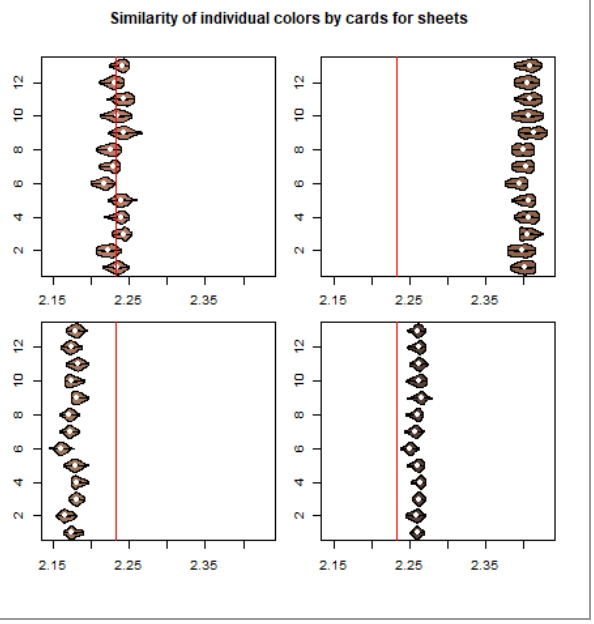
1. Histogram

Distance	Similarity
	

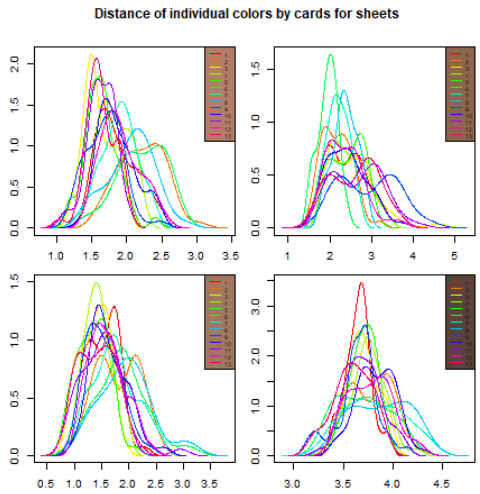
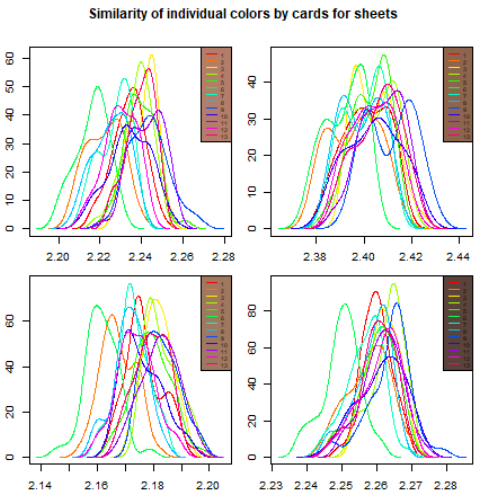
2. Boxplot

Distance	Similarity
	

3. Violin Plot

Distance	Similarity
	

4. Density plot

Distance	Similarity
	

V. Discussion

The statistics and visualizations show how different the two ways of working on likeness are.

Regarding cards on sheets and the statistics for each sheet, the differences in distances are low and the similarities vary widely – so they are not nearly the same.

Regarding colors, there's also a wide range of distances and similarities – so they are not nearly the same either.

Looking at the DeltaE ratings in "Figure 4: All colours vs. tested colors" (p.5) it can be said that all prints are accurate implementations of the master card.

Taking our aspect of visualization readability in consideration, following observations have been made.

The violin plots have similar readability in terms of box region and means to the boxplots, but the outliers are not visible because the violin includes them, whereas in the boxplots, the outliers are clearly visible, so between these two types of plots, boxplots are better for outlier detection, otherwise they work similarly.

The density plots in "Likeness based on colors" are quite overwhelming for this kind of representation of color distribution over sheets. Other plots (boxplots) are much more readable in the given context.

From all the other interesting facts about the types of plots for better readability, there's one related to rationalizing errant data. Color #7 seems to be erroneous data because of a high distance range. The median of distances shows us a better value in "Table 1: For all sheets by colors" (p. 6) compared to the mean.

From "IV.A.1 For all sheets by colors" (p. 6) we saw 549 outliers for distances. When looked into details for the same, we noticed the distance values to be [23.855,26.759] which are way higher than other colors. Hence, the conclusion for this one would be that this particular color is not a type of skin color but some other, and so we have the 546 outliers for the 546 cards resp. and 3 more from others (but that's not of our interest). On the other hand, this is not the case for similarities, where we see values similar to its "neighbors".

VI. Conclusion

As it's already been mentioned, one of our first aspects was to find the likeness between colors, cards and sheets with the MCC as given baseline, for which we used ΔE distances and cosine.similarity.

From the boxplot/violinplot in “IV.B Likeness based on cards”(p. 8) we see that cosine.similarity shows much more variation in the sheets than the ΔE distances, but in “IV.C Likeness based on colors” (p. 10) cosine.similarity has a similar behavior to the ΔE distances.

To improve the results, it would be helpful to go more in the detail redards to coherences of Delta E 2000 and cosine similarity. A scatterplot “distance vs similarity” can visualize possible relationships between these statistics, e.g. for each color.

To improve the visualization, 3D plots (only on the monitor for rotating / zooming) or overall small multiples can help to get a deeper insight.

VII. References

[1] Z. Schuessler, “Delta E 101,” 11 11 2016. [Online]. Available: http://zschuessler.github.io/DeltaE/learn/. [Accessed 24 05 2019].

[2] U. Häßler , “L*a*b* Farben vergleichen und Farbabstand,” 08 2017. [Online]. Available: https://wisotop.de/farbabstand-farben-vergleichen.php. [Accessed 25 05 2019].

[3] M. Dowle, “Package 'data.table',” 07 04 2019. [Online]. Available: https://cran.r-project.org/web/packages/data.table/data.table.pdf. [Accessed 25 05 2019].

[4] J. Gama and G. Davis, “Package ‘colorscience’,” 25 07 2018. [Online]. Available: https://cran.r-project.org/web/packages/colorscience/colorscience.pdf. [Accessed 25 05 2019].

[5] D. Adler, “Package 'vioplot',” 25 01 2019. [Online]. Available: https://cran.r-project.org/web/packages/vioplot/vioplot.pdf. [Accessed 26 05 2019].