Color Crafting: Automate the Generation of Designer Quality Color Ramps

1. Data Description and Questions

In this project, I used the datasets we collected for Color Crafting, which was an ongoing research effort in the CU VisuaLab to automate the generation of high quality color ramps for effective visualization design. We collected a corpus of 222 designer crafted color ramps from reliable resources including Tableau, ColorBrewer, and ColorLouver.com. We constructed two machine learning models, namely the K-means and Bayesian models to mimic the shape of these collected ramps. We evaluated the ramps generated by our models in terms of their effectiveness of supporting numerical value mapping and positive aesthetics, in comparison with those hand-crafted by designers and generated by traditional linear approach.

a. Data Description

To collect the data, we designed a quantitative web-based experiment through Amazon Mechanical Turk and recruited 35 professional designers from the US, UK, India, and China through social media, online forums, and interest groups. Each participant went through 39 trials of a combined variety of independent variables that we used to design this experiment.

In particular, we applied the ramps generated from different models (K-means, Bayesian, Designer, Linear) to 3 visualization types, i.e. scatterplot, heatmap, and choropleth map, and measured the numerical mapping accuracy and aesthetics for each of them(Fig 1).
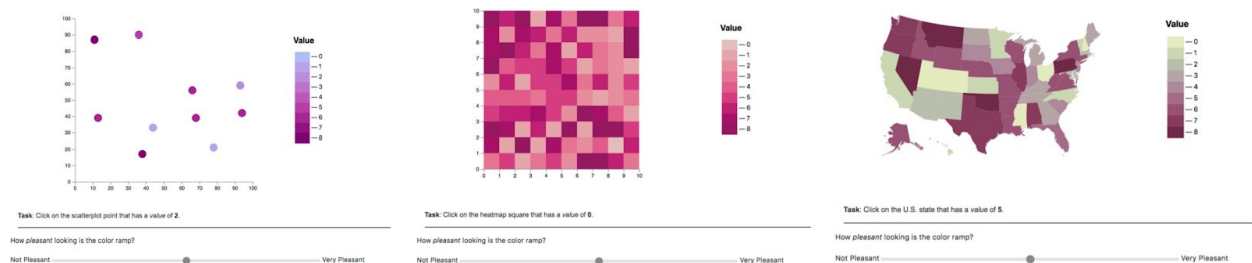
Fig 1. Screenshot of Our Web-Based Experiment: In each trial, the participant will be given a specific value and be asked to click on the colored point / square / state that represents that value from the above visualization. At the bottom there is a slider with a range of -100 to 100 for the user to decide on the aesthetics for a given ramp.

In the database, we had three tables to collect user ip address data, response data, and demographics data, which could all be connected by an auto generated user id (Fig 2).

#	Name	Type	Collation	Attributes	Null	Default	Comments	Extra	Action
1	user_id	int(15)			No	None			Change ⊖ Drop ▽ More
2	age	int(15)			No	None			Change ⊖ Drop ▽ More
3	occupation	varchar(100)	utf8_general_ci		No	None			Change ⊖ Drop ▽ More
4	gender	varchar(100)	utf8_general_ci		No	None			Change ⊖ Drop ▽ More
5	normal_vision	varchar(100)	utf8_general_ci		No	None			Change ⊖ Drop ▽ More
6	cdv	varchar(100)	utf8_general_ci		No	None			Change ⊖ Drop ▽ More
7	hours_per_day_on_computer	int(15)			No	None			Change ⊖ Drop ▽ More
8	education	varchar(100)	utf8_general_ci		No	None			Change ⊖ Drop ▽ More
9	display_type	varchar(100)	utf8_general_ci		No	None			Change ⊖ Drop ▽ More
10	display_size_inches	int(15)			No	None			Change ⊖ Drop ▽ More
11	design_experience	int(15)			No	None			Change ⊖ Drop ▽ More
12	comments	varchar(1000)	utf8_general_ci		No	None			Change ⊖ Drop ▽ More

#	Name	Type	Collation	Attributes	Null	Default	Comments	Extra	Action
1	user_id	int(15)			No	None			Change ⊖ Drop ▽ More
2	vis_type	varchar(100)	utf8_general_ci		No	None			Change ⊖ Drop ▽ More
3	model_type	varchar(100)	utf8_general_ci		No	None			Change ⊖ Drop ▽ More
4	color_ramp_index	int(15)			No	None			Change ⊖ Drop ▽ More
5	target_value	int(15)			No	None			Change ⊖ Drop ▽ More
6	selected_value	int(15)			No	None			Change ⊖ Drop ▽ More
7	aesthetic_rating	int(15)			No	None			Change ⊖ Drop ▽ More
8	trial_number	int(15)			No	None			Change ⊖ Drop ▽ More

#	Name	Type	Collation	Attributes	Null	Default	Comments	Extra	Action
1	user_id 🔑	int(15)			No	None		AUTO_INCREMENT	Change ⊖ Drop ▽ More
2	turk_id	varchar(1000)	utf8_general_ci		Yes	NULL			Change ⊖ Drop ▽ More
3	ip_address	varchar(1000)	utf8_general_ci		Yes	NULL			Change ⊖ Drop ▽ More

Fig 2. Screenshot of Our Database Table Structures: From top to bottom are our tables for user demographic data, user response data, and user ip address data, with "user_id" being used to identify a specific user across different tables.

After checking on the user ip data and making sure there was no repetitive user, I exported the user response table and demographic table as two csv files (Fig 3). As shown in the screenshots, we had both numerical data such as target_value, selected_value, and aesthetic_rating, and categorical data like vis_type, model_type, education level, display type etc.

b. Questions
Given the goal of this study is to investigate the accuracy and aesthetics of each model, my initial questions are:
- What is the numerical mapping accuracy for each of these models?
- What is the average aesthetic rating for each of these models?

user_id	vis_type	model_type	color_ramp_index	target_value	selected_value	aesthetic_rating	trial_number
2	choropleth	engagement	0	8	8	1	0
2	heatmap	designer	1	6	6	0	1
2	scatter	designer	8	5	3	-20	2
2	choropleth	linear	21	8	8	27	3
2	choropleth	designer	11	6	6	24	4
2	scatter	bayesian	24	4	3	37	5
2	scatter	kmeans	14	1	0	-26	6
2	choropleth	kmeans	18	4	4	-21	7
2	scatter	linear	6	8	7	44	8
2	heatmap	engagement	0	8	8	53	9
2	heatmap	bayesian	0	2	3	-16	10
2	scatter	bayesian	10	1	0	0	11
2	choropleth	bayesian	13	3	4	42	12
2	heatmap	designer	16	2	3	30	13
2	heatmap	bayesian	22	2	3	30	14
2	choropleth	linear	5	5	6	34	15

user_id	age	occupation	gender	normal_vision	cdv	hours_per_day_on_computer	education	display_type	display_size_inches	design_experience	comments
2	30	Service Designer	Male	yes	no		6 graduateStudies	phone	4	8	
3	57	Web Developer	Female	yes	no		8 highSchool	desktop	17	10	
5	28	UI Designer	Male	yes	no		10 someCollege	laptop	13	7	Use both hue and tint
6	40	Design	Male	yes	no		8 collegeGraduate	laptop	24	16	kerouan@gmail.com
7	35	Graphic Design/Marl	Male	yes	no		10 collegeGraduate	desktop	27	15	john@gawley.com
8	35		Female	yes	no		12 graduateStudies	phone	6	0	
9	24	Interactive and Visua	Female	yes	no		8 collegeGraduate	laptop	15	4	I based my answers on two
10	32	Product Designer	Female	yes	no		5 collegeGraduate	desktop	26	11	
11	31	Analyst	Female	yes	no		10 collegeGraduate	laptop	15	6	
12	27		Female	no	no		9 collegeGraduate	desktop	27	3	
13	29		Male	yes	no		9 someCollege	desktop	-1	2	
14	25		Female	yes	no		8 someCollege	desktop	0	2	
15	20		Male	yes	no		5 collegeGraduate	laptop	13	0	it so cool
16	27		Male	yes	no		10 graduateStudies	desktop	27	3	
17	24		Male	yes	no		10 collegeGraduate	desktop	1	2	
19	28	UI	Female	yes	no		12 collegeGraduate	laptop	15	1	
20	28		Male	yes	no		12 collegeGraduate	desktop	23	2	
21	22	beijing	Male	yes	yes		3 collegeGraduate	laptop	13	2	
22	23		Female	yes	no		9 collegeGraduate	desktop	-1	3	
23	28	designer	Male	yes	no		10 collegeGraduate	desktop	8	5	
24	23	Student	Male	yes	no		6 someCollege	laptop	15	6	

Fig 3. Screenshot of the CSV Files: User response(Top), User demographics(Bottom).

2. Data Preparation, Exploration, and Cleaning

A couple of things to note about the data: we used engagement check (see in the model_type column) across the study, which were just some overly easy tasks to make sure the participants were paying attention to the study and giving effective results. If they failed one of these checks, their results would be removed. Each of the 35 participants were presented with 39 different trials and everyone only did this study once.

a. Data Preparation & Exploration

Looking at the response data, I found the vis_type, model_type, selected_value, and target_value, aesthetic_rating can be useful variables to help answer my questions, while the demographic data might be used to gain some insights about what might influence designers' aesthetic judgements.

b.Cleaning

In JupyterLab, I imported these CSV files as dataframes and merged them on "user_id", which gave me a dataframe with 1716 rows and 19 columns. Then I grouped the dataframe by "user_id" and dropped the duplicates to make sure each participant finished all the trials and had done it only once. Then I checked on the "normal_vision"

column to make sure there weren't any color blindness issues that might influence our analyses. After that I also filtered the data frame by "engagement" so that each participant passed the engagement check and gave the effective results (Fig 4).

Import Datasets

Make dataframes for both csv files and drop the duplicates in demographics by user_id .

```
In [2]: response_df = pd.read_csv('response.csv')
        demographics_df = pd.read_csv('demographics.csv')
        demographics_df = demographics_df.drop_duplicates(subset='user_id',keep='last')
        response_df.shape
```

```
Out[2]: (1716, 8)
```

1. Merge dataframes on "user_id".

```
In [3]: cc_df = pd.merge(response_df, demographics_df, on='user_id')
        print(cc_df.shape)
```

```
        (1716, 19)
```

2. Check on the data with the "groupby" method on 'user_id', there should be 39 trials for each participant.

```
In [4]: cc_df.groupby('user_id').size()
```

```
Out[4]: user_id
        2     39
        3     39
        5     39
        6     39
        7     39
        8     39
        9     39
        10    39
        11    39
        12    39
        13    39
        14    39
        15    39
        16    39
        17    39
```

3. Something went wrong for the subject with id 47, who seemed to have done this study twice, drop their second result.

```
In [5]: user47 = cc_df[cc_df['user_id']==47]
        user47 = user47.drop_duplicates(subset='trial_number',keep='first')
```

```
In [6]: df = cc_df[cc_df["user_id"] != 47]
        df.shape
```

```
Out[6]: (1638, 19)
```

```
In [7]: cc_df = pd.concat([df, user47], ignore_index=True)
        print(cc_df.shape)
        cc_df.head()
```

```
        (1677, 19)
```

4. Check on the normal_vision column, drop "no" if there is any.

```
In [8]: cc_df[cc_df['normal_vision']=='no']
```

5. Drop entries for engagement check

```
In [10]: cc_df = cc_df[cc_df.model_type != 'engagement']
```

```
In [11]: cc_df.shape
```

```
Out[11]: (1512, 19)
```

Export Dataframe as CSV

```
In [12]: cc_df.to_csv('cc.csv')
```

Fig 4. Screenshot of the Data Cleaning Process: After data cleaning and preprocessing, I had a new dataframe which had 1512 rows and 19 columns.

3. **Data Analysis**

Based off of the questions I have, I investigated the data from both quantitative perspective and qualitative perspectives and made visualizations out of my results.

a. Quantitative Analysis

I grouped the dataframe by model_type and calculated the accuracy rate and average aesthetic rating for each of them. The result was very encouraging and it confirmed the robustness of our models as in Fig 5: Color ramps generated by our K-means model and Bayesian model outperformed design and linear ramps in numerical mapping tasks and performed at least as well as designer ramps in supporting positive aesthetic judgements.
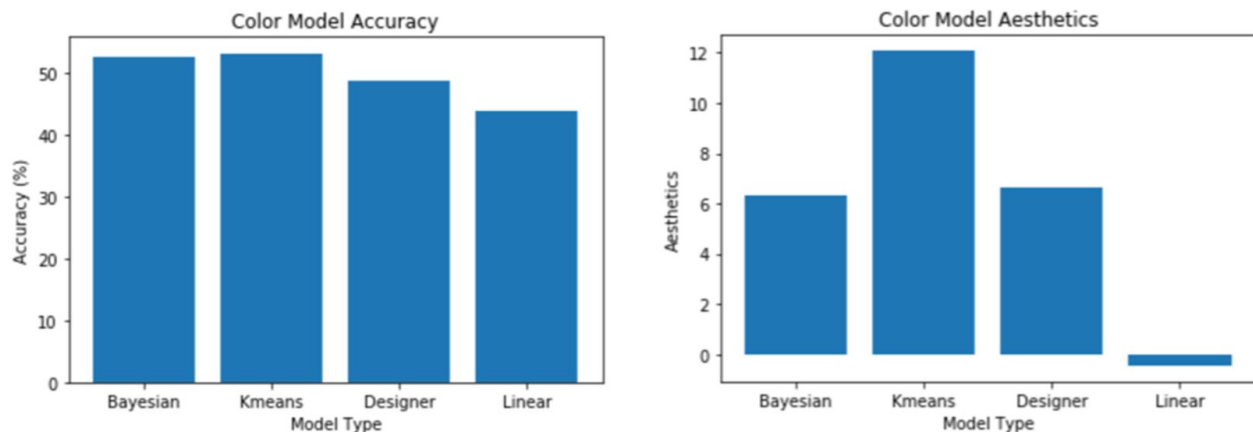
Fig 5. Visualizations of the accuracy and aesthetics for each color model:
Accuracy: K-means > Bayesian > Designer > Linear
Aesthetics: K-means > Designer > Bayesian > Linear

Then I was interested in further investigating the designer demographics data to see if there was any interesting pattern that may inform us some principles for effective color usage. So I ran a pairplot of the demographics data as in Fig 6, which didn't seem to show a clear correlation between any of these variables other than an obvious positive relationship between age and years of design experience. So I did a multiple regression with aesthetic_rating as the dependent variable and variables in other columns as the independent variables. I ran this multiple regression with the backward p-value selection approach and plotted the residuals with a histogram and qq plot (Fig 7) for condition check. And the result was interesting as the most important factors seemed to be "display_size", "model_type", and "vis_type". I was not very convinced about the result of display_size and I found there was one user who finished the study on their phone, which may have caused the significant difference in aesthetic judgement, given that the same color may render and look differently on different displays of different sizes.
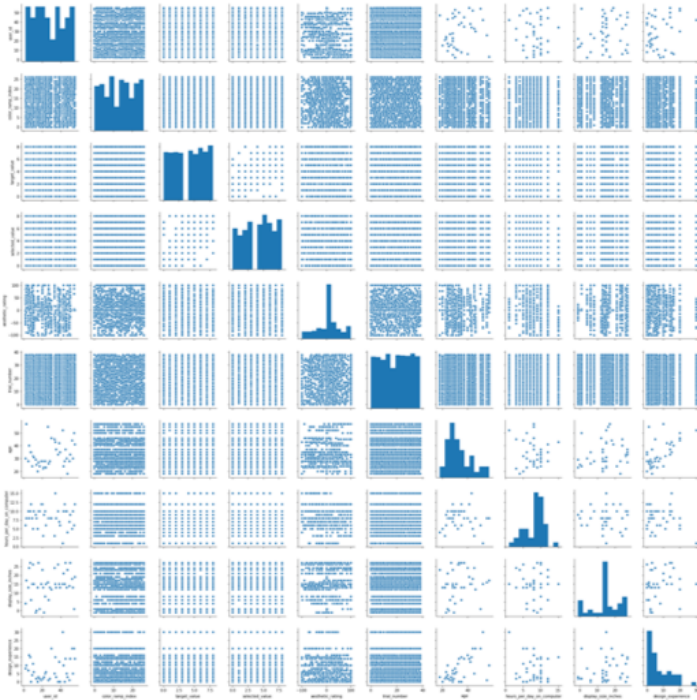
**Fig 6. Pairplot of the Demographics Data**
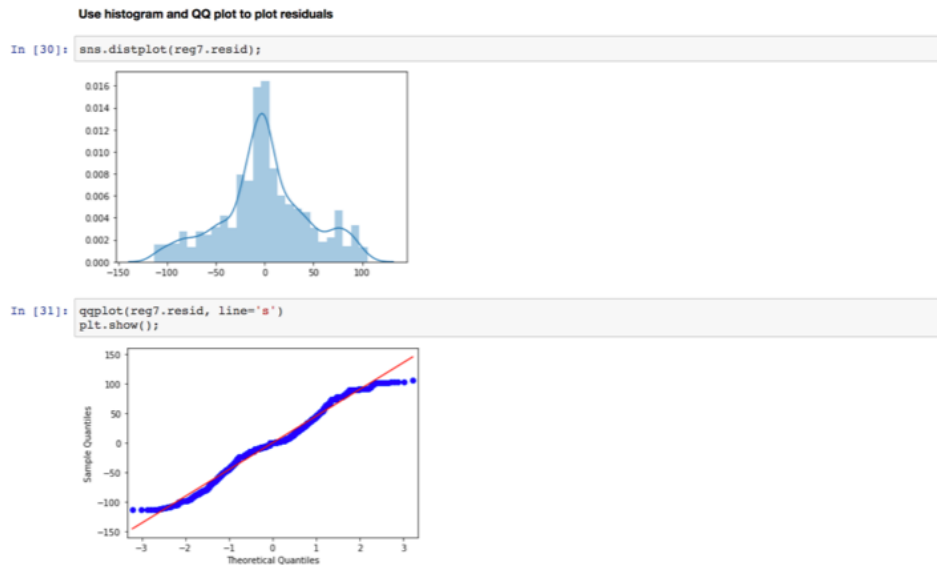
Fig 7. Histogram and QQ Plot for residuals: The data seemed to have been normally distributed, however the R-squared wasn't very good which might indicate this actually is not a perfect model.

b. Qualitative Analysis

Then I moved on to the qualitative side of analysis and did both univariate and bivariate investigations of the demographics data. In doing so, I wanted to see if any patterns may reveal (Fig 8).
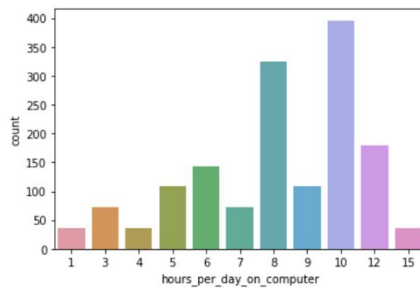
Q1: How many years of design experience do our participants have?

```
In [32]: sns.countplot(cc_df.design_experience);
```
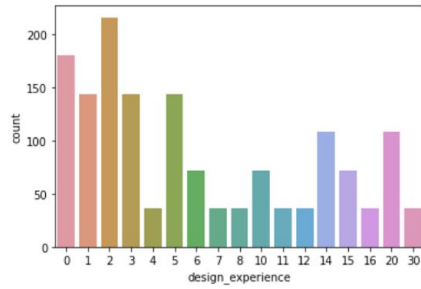

Q2: How many hours do designers commonly spend on computer everyday?

```
In [33]: sns.countplot(cc_df.hours_per_day_on_computer);
```
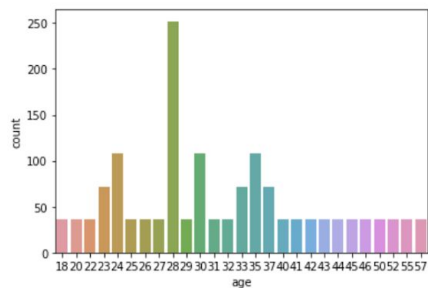

Q3: What level of education do our participants have?

```
In [34]: sns.countplot(cc_df.education);
```
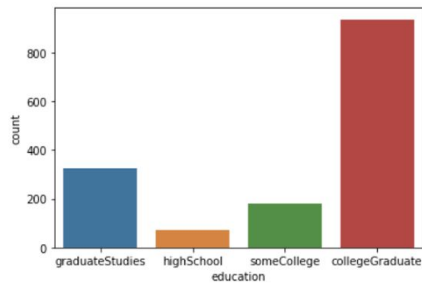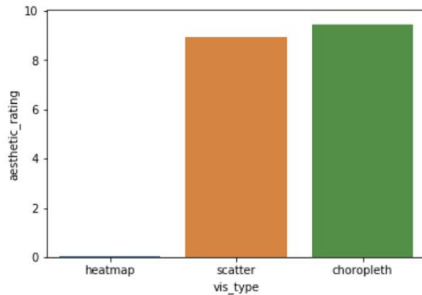

Q4: How many hours on average do designers spend on computer everyday?

```
In [35]: sns.countplot(cc_df.age);
```
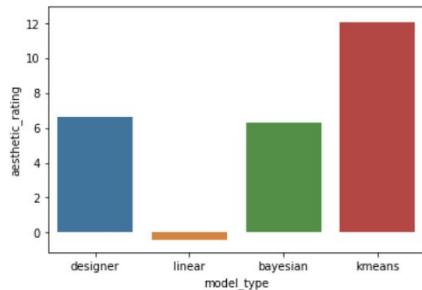
Q5: Is there a relationship between "vis_type" and "aesthetic_rating"?

```
In [36]:   ax = sns.barplot(x="vis_type", y="aesthetic_rating", data=cc_df, ci=0)
```
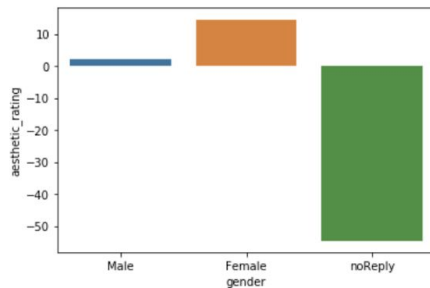

Q6: Is there a relationship between "model_type" and "aesthetic_rating"?

```
In [37]:   ax = sns.barplot(x="model_type", y="aesthetic_rating", data=cc_df, ci=0)
```


Q7: Is there a relationship between "gender" and "aesthetic_rating"?

```
In [38]:   ax = sns.barplot(x="gender", y="aesthetic_rating", data=cc_df, ci=0)
```


Fig 7. Visualizations for Demographics Information: Years of design experience, Hours per day on computer, Education Levels, Age, the relationship between Vis type and Aesthetic rating, the relationship between Model type and Aesthetic Rating.

From the visualizations above, we can easily see that
1. Most participants in our study were junior designers who had fewer than 5 years design experience.
2. Most designers spend more than 10 hours per day on their computers.
3. Most designers have graduate degrees.
4. Most designers in our study aged from 23 - 35, with a mode of 28.

5. Heatmap seemed to have significantly lower aesthetic ratings than choropleth map and scatter plot.

6. The rank of aesthetic ratings by model confirmed our computation in prior analysis: Kmeans > Designer > Bayesian > Linear.

7. Female designers tend to give higher aesthetic ratings than male ones.

4. **Conclusion**

The investigation of the Color Crafting datasets confirmed that the ramps generated by our models work pretty good in both accuracy and aesthetic judgement and shed some light on the color usage in visualization design. The accuracy score and average aesthetic rating scores that I calculated in JupyterLab aligned with the former results generated by JMP.

a. Limitations

The multiple regression model didn't seem to be a very good fit for this dataset, as it didn't yield a very good r value. Also, although the quantitative analyses confirmed the robustness of our models, the qualitative analyses could use a larger size of data and a border range of participants to gain more insights into these professionals' preferences on color aesthetic judgement. Meanwhile, given the dataset we have in this context, the most useful insights we can get are more on the descriptive statistics side, we could definitely have more inferential statistics once we have a bigger sample size.

b. Future Work

As suggested by the results, our K-means model and Bayesian model can generate color ramps that support both accurate numerical mappings as well as positive aesthetics, and are thus worth further investigation and development. Our latest plan is to incorporate these models into a color picker tool to help designers and visualization developers of all levels to readily generate color ramps of their choice.

Meanwhile from the qualitative explorations, it looks like female designers are more tolerant when giving aesthetic ratings, which might be an interesting future work topic to discover the sex differences in color aesthetic preferences. Besides, there seemed to have something with the visualization type and aesthetic, as heatmaps were rated way lower plant looking than scatter plots and choropleth maps. So moving forward, it could also be a potentially interesting direction to explore whether some visualization types are considered more visually pleasant than others and the reasons behind the scene.