INFO 5871 INTRO TO DATA SCIENCE

Color Crafting Data Analysis

Keke Wu

Color in Visualization

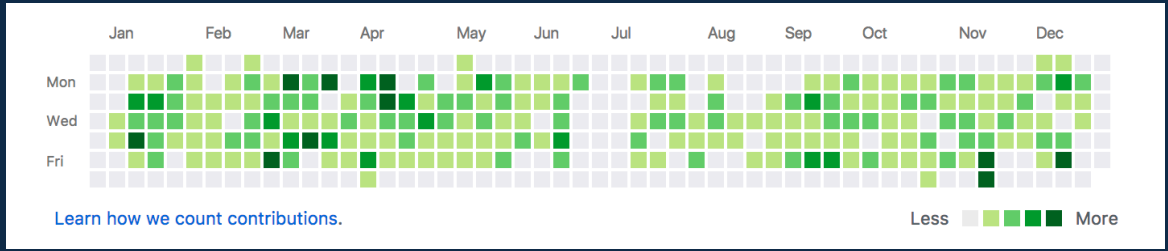
Color in Visualization

Encode Numerical Values

Distinguish Categories

Color in Visualization

Encode Numerical Values



Color in Visualization

Distinguish Categories

Color Crafting

Automate the Construction of Designer Quality Color Ramps

Kmeans Model	Designer Crafted Ramps
Bayesian Model	Linear Approach Generated Ramps

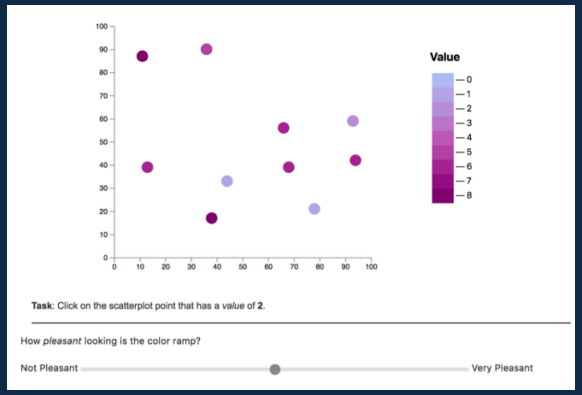
High Quality Color Ramp

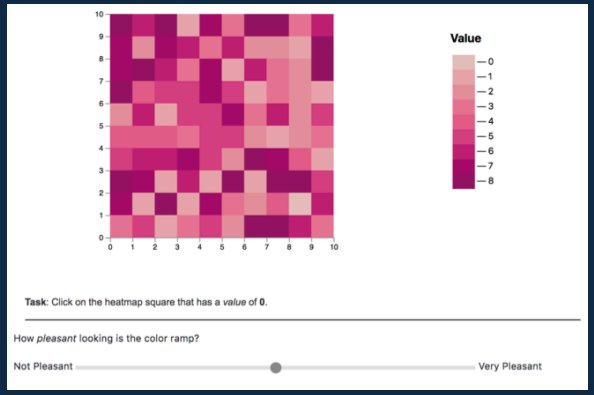
Distinguishable

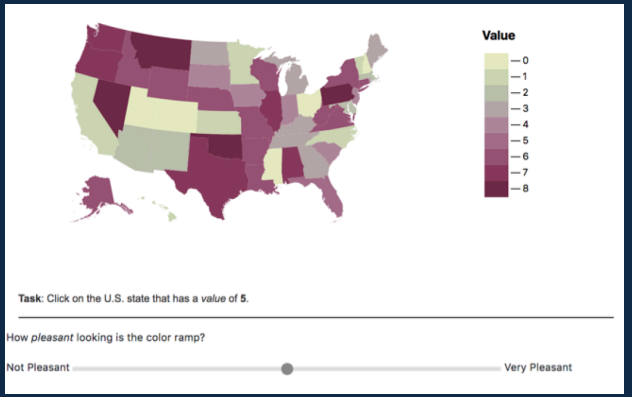
Pleasant Looking

Color Crafting

Automate the Construction of Designer Quality Color Ramps







DATA DESCRIPTION

1	user_id	vis_type	model_type	color_ramp_index	target_value	selected_value	aesthetic_rating	trial_number
2	2	choropleth	engagement	0	8	8	1	0
3	2	heatmap	designer	1	6	6	0	1
4	2	scatter	designer	8	5	3	-20	2
5	2	choropleth	linear	21	8	8	27	3
6	2	choropleth	designer	11	6	6	24	4
7	2	scatter	bayesian	24	4	3	37	5
8	2	scatter	kmeans	14	1	0	-26	6
9	2	choropleth	kmeans	18	4	4	-21	7
10	2	scatter	linear	6	8	7	44	8
11	2	heatmap	engagement	0	8	8	53	9
12	2	heatmap	bayesian	0	2	3	-16	10
13	2	scatter	bayesian	10	1	0	0	11
14	2	choropleth	bayesian	13	3	4	42	12
15	2	heatmap	designer	16	2	3	30	13
16	2	heatmap	bayesian	22	2	3	30	14
17	2	choropleth	linear	5	5	6	34	15

Response Data

DATA DESCRIPTION

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	

Demographics Data

QUESTIONS

What is the numerical mapping accuracy for each model?

What is the average aesthetic rating for each model?

What are the important factors that may influence designers' aesthetic rating?

DATA CLEANING

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)
```

DATA CLEANING

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
39    39
42    39
43    39
44    39
45    39
46    39
47    78
49    39
50    39
52    39
53    39
54    39
55    39
dtype: int64
```

DATA CLEANING

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)
```

DATA CLEANING

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')
```

DATA ANALYSIS & VISUALIZATION

1. Accuracy & Aesthetics by Model Type

```
In [3]: grouped = cc_df.groupby('model_type')
        bayesian = grouped.get_group('bayesian')
        kmeans = grouped.get_group('kmeans')
        designer = grouped.get_group('designer')
        linear = grouped.get_group('linear')

In [4]: bayesian_diff = bayesian['selected_value'] - bayesian['target_value']
        bayesian_diff.value_counts('0')

Out[4]:  0    0.526455
         1    0.246032
        -1    0.174603
        -2    0.015873
         2    0.013228
         3    0.007937
        -3    0.002646
         7    0.002646
         6    0.002646
         5    0.002646
         4    0.002646
        -4    0.002646
        dtype: float64
```

DATA ANALYSIS & VISUALIZATION

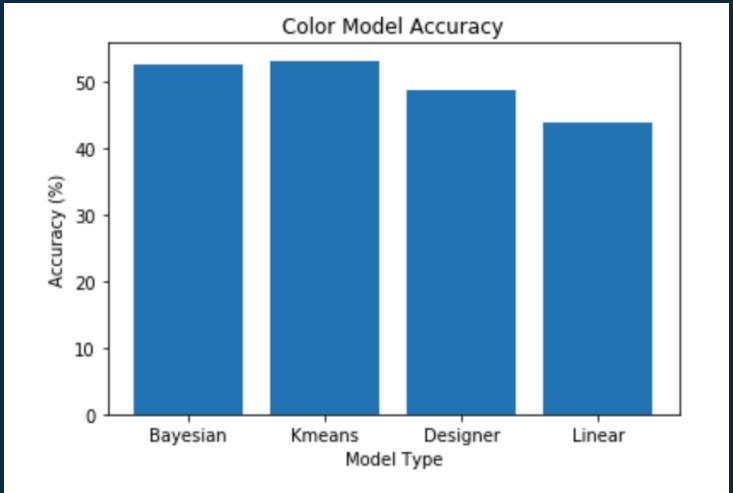
1. Accuracy & Aesthetics by Model Type

In [8]:
```
bayesian_aes = bayesian['aesthetic_rating'].mean()
bayesian['aesthetic_rating'].describe()
```

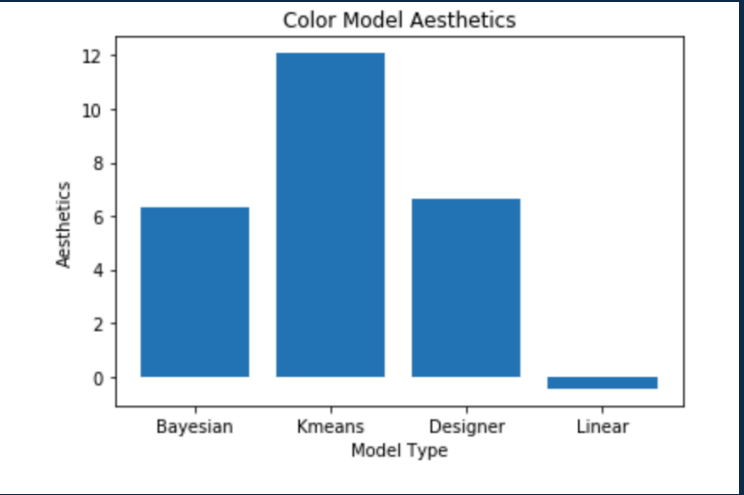
Out[8]:
```
count    378.000000
mean       6.314815
std       43.410982
min     -100.000000
25%      -15.000000
50%        0.000000
75%       30.000000
max      100.000000
Name: aesthetic_rating, dtype: float64
```

DATA ANALYSIS & VISUALIZATION

1. Accuracy & Aesthetics by Model Type



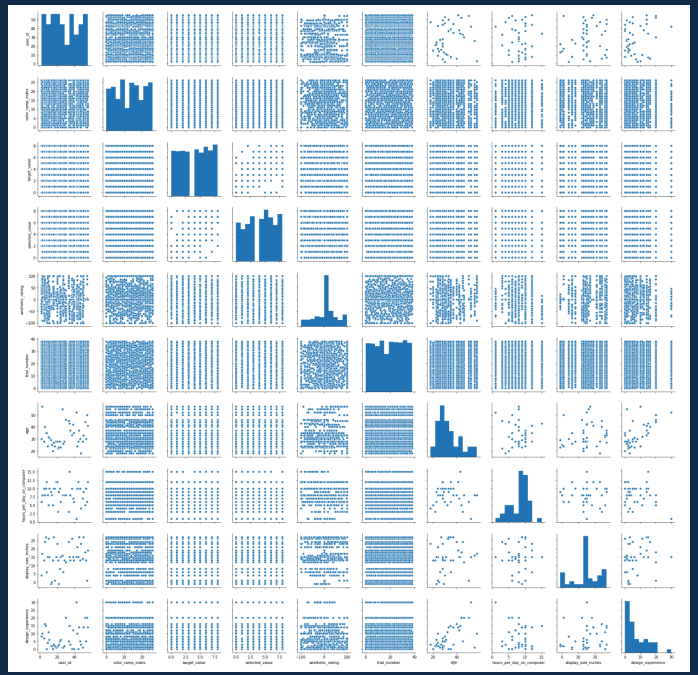
52.64%, 53.17%, 48.68%, 43.92%



6.31, 12.07, 6.61, -0.45

DATA ANALYSIS & VISUALIZATION

2. Aesthetics



DATA ANALYSIS & VISUALIZATION

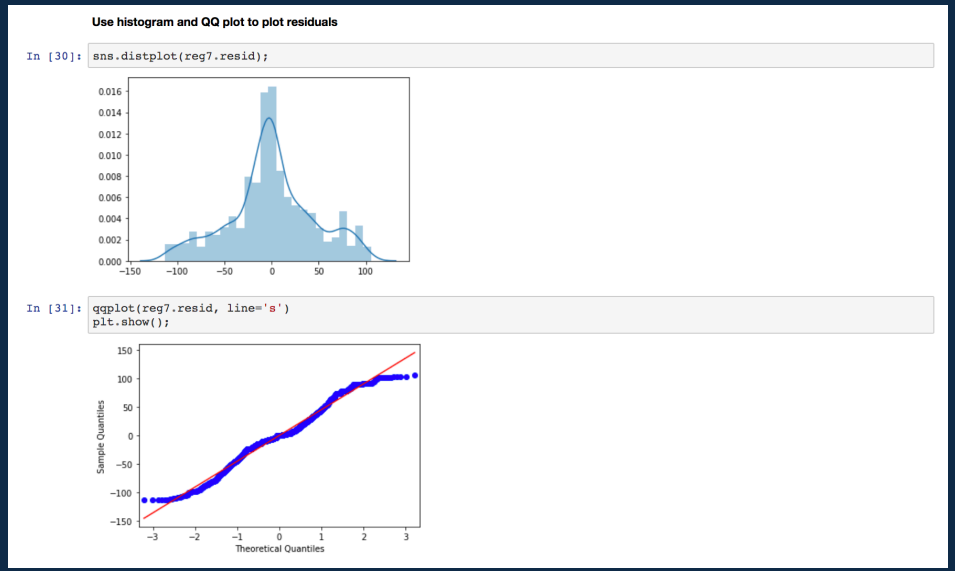
2. Aesthetics

In [17]:
```
model1 = sm.ols(formula ="aesthetic_rating ~ vis_type + model_type  + age + gender + hours_per_day_on_computer + educa
tion + display_type + display_size_inches + design_experience", data=cc_df)
reg1 = model1.fit()
print(reg1.aic)
reg1.summary()
```

```
15751.286798904113
```

DATA ANALYSIS & VISUALIZATION

2. Aesthetics



display_size

model_type

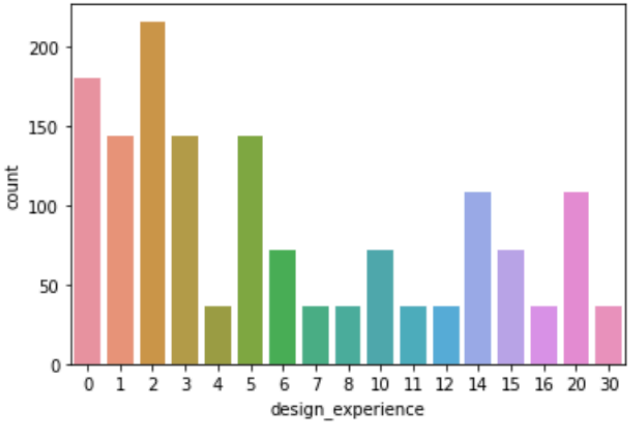
vis_type

DATA ANALYSIS & VISUALIZATION

3. Designer Demographics (Univariate)

Q1: How many years of design experience do our participants have?

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

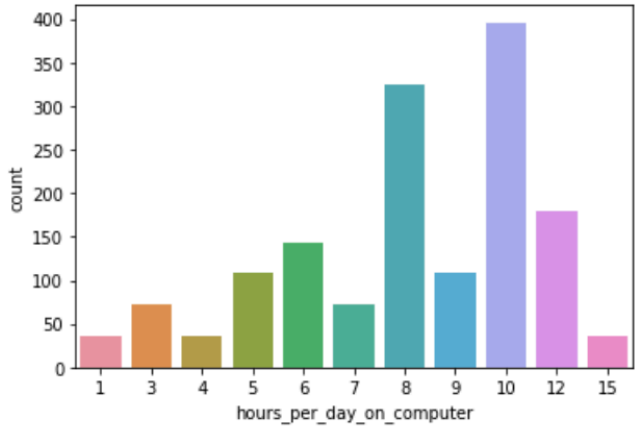


DATA ANALYSIS & VISUALIZATION

3. Designer Demographics (Univariate)

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

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

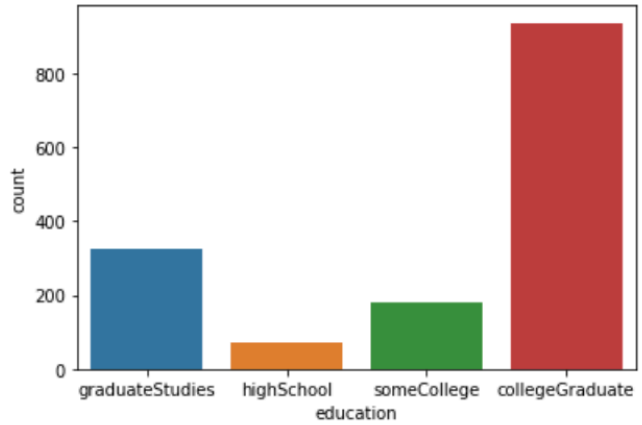


DATA ANALYSIS & VISUALIZATION

3. Designer Demographics (Univariate)

Q3: What level of education do our participants have?

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

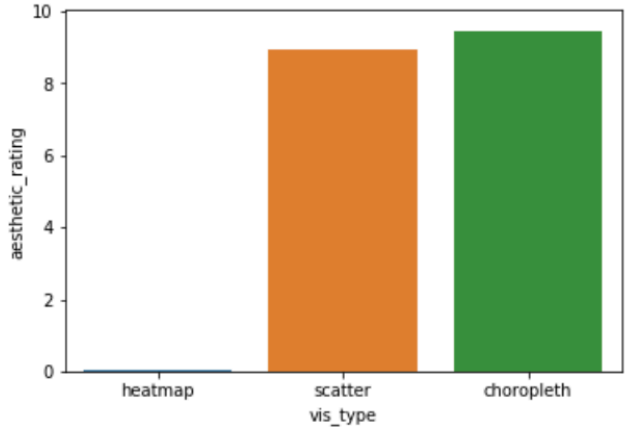


DATA ANALYSIS & VISUALIZATION

3. Designer Demographics (Bivariate)

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)
```

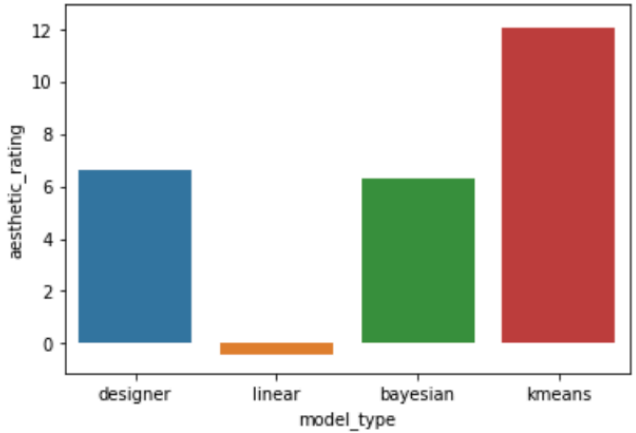


DATA ANALYSIS & VISUALIZATION

3. Designer Demographics (Bivariate)

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)
```

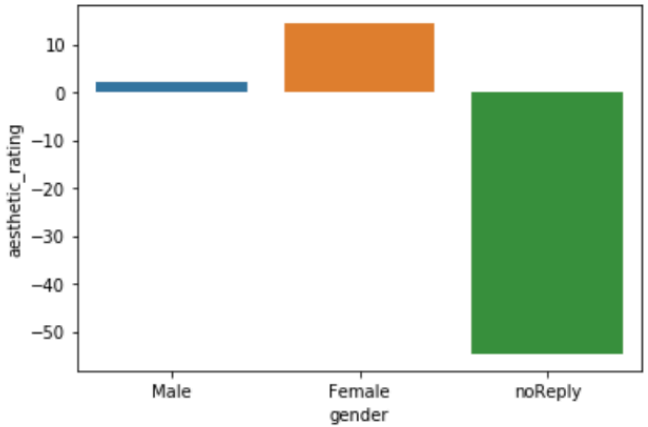


DATA ANALYSIS & VISUALIZATION

3. Designer Demographics (Bivariate)

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)
```



THANKS!