DATA ANALYTICS REPORT

Customer Color Preference Analysis

Analysis of Independence between Age Groups and Color Preferences
Using Chi-Square Test of Independence

Course Information

Institution: Manipal Institute of Technology

Department: School of Computer Engineering

Course: Data Analytics

Subject Code: CSS 2103

Academic Year: 2024-2025

Section: DSE-A

Group Information

Group No: 9

Team Members:

Roll No.	Name
20	Mithil
22	Nikhil
41	Kowshik
42	Akshay

Executive Summary

This comprehensive report investigates the relationship between customer age groups and color preferences using statistical hypothesis testing. The study analyzes survey data from 500 respondents across 34 countries to determine whether color preference is independent of age group demographics.

Research Question

Is there a statistically significant relationship between age group and color preference among customers?

Methodology

Chi-Square Test of Independence with $\alpha = 0.05$ significance level

Key Findings

- No statistically significant relationship exists between age and color preference ($\chi^2 = 10.7135$, $p = 0.5536$)
- Blue is the most preferred color across all age groups (33.2% overall)
- Yellow is the least preferred color across all demographics (11%)
- Effect size is negligible (Cramér's V = 0.0845)
- Color preferences are remarkably consistent across age groups

Business Implications

- Marketing strategies should not differentiate color choices based on age demographics
- Product design can use universal color preferences rather than age-targeted palettes
- Blue remains the safest choice for broad appeal across all customer segments

1. INTRODUCTION

1.1 Background

Color preference is a critical factor in marketing, product design, branding, and consumer behavior. Understanding whether color preferences vary across demographic segments—particularly age groups—can inform strategic business decisions. Previous research has shown mixed results regarding age-related color preferences, with some studies suggesting preferences change with maturity while others indicate universal aesthetic principles.

1.2 Objectives

The primary objectives of this analysis are:

1. To determine whether color preference is independent of age group
2. To identify overall color preference patterns across the sample
3. To quantify the strength of any relationship between age and color choice
4. To provide actionable insights for marketing and product design strategies

1.3 Dataset Description

Source: Customer Survey Responses (responses.csv)

Sample Size: 500 respondents

Geographic Coverage: 34 countries worldwide

Variables Collected: Country, State/Territory, Age, Gender, Favorite Color (hex code)

1.4 Scope and Limitations

Scope:

- Analysis focuses exclusively on the relationship between age and color preference
- Color classification into four primary categories: Red, Blue, Green, Yellow
- Statistical testing at $\alpha = 0.05$ significance level

Limitations:

- Self-reported data without behavioral validation
- Cultural confounding due to global geographic diversity
- Simplification of 500 unique colors into 4 categories
- Unbalanced age group distribution (56% are 45+)
- No context provided (e.g., color preference for what purpose?)

2. METHODOLOGY

2.1 Data Collection

The dataset was obtained from a global customer survey collecting demographic information and color preferences. Respondents were asked to select their favorite color from a spectrum, recorded as hexadecimal color codes (e.g., #7248ad).

2.2 Data Preprocessing

2.2.1 Missing Value Treatment

Variable	Missing Count	Missing %	Action Taken
Country	0	0%	Retained
State or Territory	492	98.4%	Dropped
Age in years	0	0%	Retained
Gender	139	27.8%	Retained (not analyzed)
Favorite color	0	0%	Retained

Rationale: The "State or Territory" variable was removed due to excessive missing data (98.4%), rendering it unsuitable for meaningful analysis.

2.2.2 Color Classification System

The dataset contained 500 unique hexadecimal color codes, which were algorithmically classified into four primary color categories to enable statistical analysis.

Classification Algorithm:

- Convert hex code to RGB values (Red, Green, Blue channels: 0-255)
- Apply classification logic:
- • If R > 150 AND G > 150 AND B < 150 → Yellow
- • Else if R > max(G, B) → Red
- • Else if B > max(R, G) → Blue
- • Else if G > max(R, B) → Green
- • Else → Assign to dominant channel

Color Distribution Results:

Color	Count	Percentage
Blue	166	33.2%
Red	147	29.4%
Green	132	26.4%
Yellow	55	11.0%
Total	**500**	**100%**

2.2.3 Age Group Categorization

Ages were categorized into five meaningful demographic groups based on life stages and career phases:

Age Group	Age Range	Description	Count	Percentage
Under 18	< 18 years	Children and adolescents	52	10.4%
18-25	18-25 years	Young adults / College age	38	7.6%
26-35	26-35 years	Early career professionals	76	15.2%
36-45	36-45 years	Mid-career professionals	53	10.6%
45+	> 45 years	Mature adults and seniors	281	56.2%
Total			**500**	**100%**

Note: The 45+ group is the largest segment, representing over half of the sample.

2.3 Statistical Analysis Method

2.3.1 Chi-Square Test of Independence

The Chi-Square (χ^2) Test of Independence is a non-parametric statistical test used to determine whether two categorical variables are independent or associated.

Hypotheses:

H_0 (Null Hypothesis): Color preference is independent of age group (no relationship exists)

H_1 (Alternative Hypothesis): Color preference is dependent on age group (a relationship exists)

Test Assumptions:

- ✓ Independence of observations: Each respondent provides one response
- ✓ Categorical variables: Both age group and color are categorical
- ⚠ Expected frequency ≥ 5: Most cells meet this criterion (validated below)
- ✓ Adequate sample size: n = 500 is sufficient

Significance Level: $\alpha = 0.05$ (95% confidence)

Decision Rule:

- If p-value < 0.05 → Reject H_0 (significant relationship)
- If p-value ≥ 0.05 → Fail to reject H_0 (no significant relationship)

2.3.2 Effect Size Measurement

Cramér's V is calculated to measure the strength of association between variables, regardless of statistical significance.

Formula:

$V = \sqrt{(\chi^2 / (n \times \min(r-1, c-1)))}$

Where:

- χ^2 = Chi-square statistic
- n = Total sample size
- r = Number of rows (age groups)
- c = Number of columns (colors)

Interpretation:

Cramér's V	Association Strength
0.00 - 0.10	Negligible
0.10 - 0.30	Weak
0.30 - 0.50	Moderate
0.50 - 1.00	Strong

3. DESCRIPTIVE STATISTICS

3.1 Sample Demographics

3.1.1 Age Distribution

Statistic	Value
Count	500
Mean	49.4 years
Standard Deviation	23.1 years
Minimum	8 years
25th Percentile	30 years
Median (50th)	50 years
75th Percentile	70 years
Maximum	89 years

Interpretation: The sample has a mean age of 49.4 years with substantial variation (SD = 23.1), indicating representation across the entire age spectrum from children to seniors.

3.1.2 Geographic Distribution

Total Countries Represented: 34

Countries Included:

Argentina, Australia, Austria, Belgium, Brazil, Canada, Chile, Denmark, Finland, France, Germany, Greece, India, Ireland, Israel, Italy, Japan, Mexico, Netherlands, New Zealand, Norway, Poland, Portugal, Saudi Arabia, Singapore, South Korea, Spain, Sweden, Switzerland, Thailand, United Arab Emirates, United Kingdom, United States, Vietnam

Interpretation: The sample demonstrates strong global diversity, though this may introduce cultural confounding in color preferences.

3.1.3 Gender Distribution

Gender	Count	Percentage
Not Specified	139	27.8%
Other	134	26.8%
Female	114	22.8%
Male	113	22.6%

Interpretation: The sample shows balanced gender representation with inclusive categorization. Gender was not included in the primary analysis.

4. CONTINGENCY TABLE ANALYSIS

4.1 Observed Frequencies

The contingency table below displays the actual count of respondents in each Age Group × Color combination:

Age Group	Blue	Green	Red	Yellow	Row Total
Under 18	18	10	21	3	**52**
18-25	16	10	9	3	**38**
26-35	29	16	22	9	**76**
36-45	20	14	14	5	**53**
45+	83	82	81	35	**281**
Col Total	**166**	**132**	**147**	**55**	**500**

4.2 Row Percentages (Within Age Group)

This table shows what percentage of each age group prefers each color:

Age Group	Blue	Green	Red	Yellow
Under 18	34.62%	19.23%	40.38%	5.77%
18-25	42.11%	26.32%	23.68%	7.89%
26-35	38.16%	21.05%	28.95%	11.84%
36-45	37.74%	26.42%	26.42%	9.43%
45+	29.54%	29.18%	28.83%	12.46%

Key Insights:

- Under 18: Red is most preferred (40.38%), diverging from other groups
- 18-25: Blue leads significantly (42.11%)
- 26-35: Blue preference remains high (38.16%)
- 36-45: Balanced distribution across Blue, Green, Red
- 45+: Nearly equal preference for Blue, Green, Red (~29% each)
- Yellow: Consistently least preferred across ALL age groups (6-12%)

5. CHI-SQUARE TEST RESULTS

5.1 Test Statistics

Metric	Value
Chi-Square Statistic (χ^2)	10.7135
P-value	0.5536
Degrees of Freedom	12
Significance Level (α)	0.05
Sample Size (n)	500

Degrees of Freedom Calculation:

df = (r - 1) × (c - 1) = (5 - 1) × (4 - 1) = 4 × 3 = 12

5.2 Expected Frequencies

Expected frequencies represent what we would observe if color preference were completely independent of age:

Age Group	Blue	Green	Red	Yellow
Under 18	17.26	13.73	15.29	5.72
18-25	12.62	10.03	11.17	4.18
26-35	25.23	20.06	22.34	8.36
36-45	17.60	13.99	15.58	5.83
45+	93.29	74.18	82.61	30.91

Formula: E(i,j) = (Row Total × Column Total) / Grand Total

Example:

E(18-25, Blue) = (38 × 166) / 500 = 12.62

5.3 Assumption Validation

Chi-Square Test Assumption: All expected frequencies should be ≥ 5

Minimum Expected Frequency: 4.18 (Age 18-25, Yellow)

Assessment:

- ⚠️ One cell (18-25 × Yellow) has expected frequency < 5
- ✓ This represents 1 out of 20 cells (5%) — within acceptable limits
- ✓ Test results remain reasonably reliable
- ✓ All other 19 cells meet the ≥5 criterion

5.4 Statistical Decision

Decision Rule:

- If p-value < 0.05 → Reject H_0
- If p-value ≥ 0.05 → Fail to reject H_0

Our Result: p-value = 0.5536 which is >> 0.05

Decision: FAIL TO REJECT H_0

Conclusion: There is NO statistically significant relationship between age group and color preference. The observed differences in the contingency table can be attributed to random sampling variation rather than a true underlying relationship.

With a p-value of 0.5536 (well above the 0.05 threshold), we have insufficient evidence to conclude that color preference varies systematically by age group. If we were to repeat this study 100 times, we would expect to see differences this large or larger in 55 of those studies purely by chance, even if no true relationship exists.

5.5 Effect Size Analysis

Cramér's V Calculation:

$V = \sqrt{(\chi^2 / (n \times \min(r-1, c-1)))}$

$V = \sqrt{(10.7135 / (500 \times 3))}$

$V = \sqrt{(10.7135 / 1500)}$

$V = \sqrt{0.00714}$

$V = 0.0845$

Interpretation: Negligible effect ($V < 0.10$)

What This Means:

- Even if a relationship exists, it is extremely weak
- Age explains less than 1% of the variation in color preference
- Other factors (culture, personal taste, context, trends) play far larger roles
- The association is so small it has no practical significance

6. FINDINGS AND DISCUSSION

6.1 Primary Research Question

Question: Is there a relationship between age group and color preference?

Answer: **NO.** Our chi-square test of independence found no statistically significant relationship between age group and color preference (χ^2 = 10.7135, p = 0.5536).

6.2 Key Statistical Findings Summary

Metric	Value	Interpretation
Sample Size	500 responses	Adequate for robust analysis
Chi-Square Statistic (χ^2)	10.7135	Low value indicates similarity
P-value	0.5536	Far above 0.05 threshold
Statistical Decision	Fail to reject H_0	No evidence of relationship
Cramér's V	0.0845	Negligible effect size
Degrees of Freedom	12	(5-1) × (4-1)
Standardized Residuals	All within ±2	No significant cell-level deviations
Minimum Expected Freq.	4.18	Borderline acceptable (1 cell)

6.3 What This Means in Practice

6.3.1 Age-Independent Color Preferences

✅ Universal Appeal: A 25-year-old is just as likely to prefer blue as a 65-year-old

✅ Random Variation: Observed differences (e.g., 42% vs 30% blue preference) are within expected random sampling variation

✅ Marketing Strategy: Age-targeted campaigns should NOT rely on color as a differentiating factor

✅ Product Design: Color choices can be based on overall preferences, not age-specific palettes

6.3.2 Universal Color Patterns

Color	Overall Rank	Consistency Across Ages
Blue	1st (33.2%)	Most popular in 4/5 groups
Red	2nd (29.4%)	Strong second in most groups
Green	3rd (26.4%)	Mid-range appeal everywhere
Yellow	4th (11.0%)	Least popular in ALL groups

Insight: These patterns hold regardless of age, suggesting fundamental aesthetic preferences may be more universal than age-specific.

6.4 Business and Marketing Implications

6.4.1 Product Design

✓ Use universal color preferences (Blue first, Yellow last) regardless of target age

✓ Test colors with representative samples rather than assuming age preferences

✓ Consider context (clothing vs. home decor) more than demographics

6.4.2 Marketing and Advertising

✓ Age-segmented campaigns should differentiate on messaging, not color schemes

✓ Brand colors can appeal broadly across age demographics

✓ A/B testing based on age groups for color choice is likely unnecessary

6.4.3 Retail and E-commerce

✓ Website design: One color scheme works for all ages

✓ Packaging: Focus on shelf presence and brand identity over age targeting

✓ Seasonal collections: Trends matter more than customer age

6.5 Limitations of This Study

Limitation	Impact	Future Direction
Geographic diversity	Cultural confounding may obscure age effects	Analyze within single countries
Color simplification	500 unique colors → 4 categories loses nuance	Use more granular color categories
Self-reported data	No behavioral validation (actual purchases)	Link preferences to behavior
Unbalanced age groups	56% are 45+, only 7.6% are 18-25	Stratified sampling for balance
No context provided	Color preference for what purpose?	Specify context (clothing, cars, etc.)
Cross-sectional design	Cannot detect preference changes over time	Longitudinal study tracking individuals

6.6 Future Research Directions

Recommended Studies:

1. Within-Culture Analysis: Test the age-color relationship within individual countries to control for cultural factors

2. Context-Specific Preferences: Separate studies for clothing, home decor, technology products, food packaging

3. Granular Color Categories: Expand from 4 to 10-12 categories including light vs. dark, pastel vs. neon

4. Interaction Effects: Investigate Gender × Age × Color three-way interactions

5. Longitudinal Design: Track the same individuals over time to see if preferences change with aging

6. Behavioral Validation: Link stated preferences to actual purchase behavior and eye-tracking studies

7. CONCLUSIONS AND RECOMMENDATIONS

7.1 Primary Conclusion

With high statistical confidence (p = 0.5536, well above the 0.05 threshold), we conclude that age does NOT meaningfully influence color preference in this dataset. The similarity in color distributions across age groups is remarkable and suggests that fundamental aesthetic preferences may be more universal than age-specific.

7.2 Supporting Evidence

✅ Chi-square test: No significant association (p = 0.5536)

✅ Effect size: Negligible association (Cramér's V = 0.0845)

✅ Residual analysis: All cells within normal variation (±2 range)

✅ Visual patterns: Consistent color distributions across all age groups

✅ Row percentages: Similar preference breakdowns within each age group

✅ Column percentages: Similar age distributions within each color

7.3 Recommendations

For Marketing Professionals:

1. Abandon age-based color targeting in favor of universal appeal

2. Focus differentiation on messaging, channels, and product features—not color

3. Leverage Blue as a safe, broadly appealing choice across all demographics

4. Avoid Yellow in mass-market products (only 11% preference)

5. Test your specific context: These findings may not apply to all product categories

For Product Designers:

1. Choose colors based on brand identity and shelf presence, not target age

2. Consider context (formal vs. casual, professional vs. playful) over demographics

3. A/B test colors with representative samples across ages

4. Monitor cultural context for global products

For Researchers:

1. Replicate within single cultures to eliminate geographic confounding

2. Add context specificity: Study color preference for specific purposes

3. Include behavioral measures: Link preferences to actual choices

4. Explore interactions: Gender, income, and culture may moderate effects

7.4 Final Statement

This analysis provides strong evidence that color preference is largely age-independent in a global, diverse sample. While individual exceptions and context-specific patterns may exist, businesses and marketers should not assume that age determines color choice. Instead, focus on universal aesthetic principles, cultural context, and product-specific testing to optimize color strategies.

The data speaks clearly: whether you're 18 or 80, blue is blue, and age doesn't change what you see or what you prefer.

8. TECHNICAL APPENDIX

8.1 Chi-Square Test Formula

The chi-square statistic measures the discrepancy between observed and expected frequencies:

$$\chi^2 = \Sigma\,[(O - E)^2 / E]$$

Where:

- O = Observed frequency in each cell
- E = Expected frequency in each cell under independence assumption
- Σ = Sum across all cells in the contingency table

Interpretation:

- Larger χ^2 values → Greater deviation from independence
- Smaller χ^2 values → Closer to what independence predicts
- Our result: $\chi^2 = 10.7135$ (relatively small for 12 degrees of freedom)

8.2 Degrees of Freedom

$$df = (r - 1) \times (c - 1)$$

Where:

- r = Number of rows (age groups) = 5
- c = Number of columns (colors) = 4

Calculation:

$df = (5 - 1) \times (4 - 1) = 4 \times 3 = 12$

Purpose:

Determines the shape of the chi-square distribution used to calculate the p-value

8.3 Cramér's V Formula

$$V = \sqrt{[\chi^2 / (n \times \min(r-1, c-1))]}$$

Where:

- χ^2 = Chi-square statistic = 10.7135
- n = Total sample size = 500
- r = Number of rows = 5
- c = Number of columns = 4
- min(r−1, c−1) = min(4, 3) = 3

Calculation:

$V = \sqrt{[10.7135 / (500 \times 3)]}$

$V = \sqrt{[10.7135 / 1500]}$

$V = \sqrt{0.00714}$

V = 0.0845

Our Result:

V = 0.0845 = Negligible association

9. REFERENCES

1. Myatt, G. J., & Johnson, W. P. (2014). Making Sense of Data: A Practical Guide to Exploratory Data Analysis and Data Mining (2nd ed.). John Wiley & Sons.

2. Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann Publishers.

3. Kumar, U. D. (2021). Business Analytics: The Science of Data-Driven Decision Making (2nd ed.). Wiley Publications.

4. Grus, J. (2019). Data Science from Scratch: First Principles with Python. O'Reilly Media.

5. Maheshwari, A. (2021). Data Analytics: A Comprehensive Guide to Data Analysis and Decision-Making. Wiley Publications.

6. McHugh, M. L. (2013). "The Chi-square test of independence." Biochemia Medica, 23(2), 143-149.

7. NPTEL. Introduction to Data Analytics. Retrieved from https://archive.nptel.ac.in/courses/110/106/110106072/

8. NPTEL. Data Analytics with Python. Retrieved from https://onlinecourses.nptel.ac.in/noc21_cs45/preview

9. SciPy Documentation. scipy.stats.chi2_contingency. Retrieved from https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2_contingency.html

Report Submitted: November 2, 2025

Group 9 | Section: DSE-A | Course: Data Analytics (CSS 2103)
School of Computer Engineering | Manipal Institute of Technology

Team Members:

Roll No. 20 - Mithil | Roll No. 22 - Nikhil | Roll No. 41 - Kowshik | Roll No. 42 - Akshay

END OF REPORT