Analysis of Color Preference by Month in Online Clickstream Data

Background

This analysis examines clickstream data from an online clothing store specializing in maternity wear, collected over five months in 2008 (April through August). The dataset contains **165,474 instances** and **14 features**, including session ID, product color, product category, and month of each session. The objective of this study is to determine whether there is a statistically significant association between **product color** and the **month** in which users viewed items, potentially reflecting seasonal trends in color preferences.

Dataset Overview

- **Data Scope**: Clickstream data from an e-shop specializing in maternity clothing.
- **Variables of Interest**: `month` (representing the month of the session) and `colour` (product color viewed).
- **Total Instances**: 165,474
- **Duration**: April to August 2008
- **Features**: Categorical (e.g., product category, color, country of origin), Integer (price, session ID), and Binary (e.g., whether price is above category average).
- **No Missing Values**: The dataset is complete with no missing entries.

Statistical Methodology

A **chi-square test of independence** was employed to examine the relationship between **month** and **product color** viewed by users. This test is suitable for categorical data and allows us to determine if a significant association exists between two categorical variables—in this case, month and color.

- **Chi-square Test Statistic**: 1,444.90
- **Degrees of Freedom (DF)**: 52, calculated as (R−1)(C−1)(R-1)(C-1)(R−1)(C−1) with R=5R = 5R=5 months and C=14C = 14C=14 color categories.
- **P-value**: The p-value was effectively zero, indicating a highly significant association.

Interpretation

The chi-square test results show a **significant association** between the month and product color viewed, suggesting that users' color preferences vary across months. This could indicate a tendency for users to favor specific colors during certain times of the year, such as lighter colors in warmer months or different tones that align with seasonal trends.

Limitations
While these findings suggest a possible trend in **seasonal color preferences**, several limitations must be considered:

1. **External Influences**: Monthly variations in color preference could be influenced by external factors such as **advertising** campaigns or site-driven promotions that emphasize particular colors at different times.
2. **Limited Time Frame**: Data are limited to a single year (2008) and span only five months. Seasonal trends observed within this period may not generalize to other years or other months not represented in this dataset.
3. **Website Structure**: Changes in the website structure or product layout over time could affect user navigation and browsing patterns, potentially skewing monthly color preferences.
4. **Product Availability**: Inventory constraints, including variations in product availability for different colors and categories, could bias the apparent preferences if certain items were more prominently available or featured.

Implications
Despite these limitations, the significant relationship between month and product color preference highlights seasonal variations in user behavior that could help optimize inventory management, targeted marketing, and user experience. By aligning offerings with potential seasonal preferences, the e-shop can enhance user engagement and improve the browsing experience.

Conclusion
The observed association between month and product color preference suggests that user behavior may reflect seasonal color trends. However, caution should be exercised in generalizing these findings due to potential influences from advertisements, product availability, and the limited time span of the data. These insights nonetheless provide valuable guidance for refining the e-shop's approach to data-driven decision-making and customer engagement.

Observed Months/Color	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Totals
4	2502	8564	8126	4772	365	5120	2077	690	4174	1059	1622	2570	1826	4732	48199
5	1902	5933	5923	3314	288	3926	1550	499	3132	784	1245	2004	1457	3697	35654
6	1621	5493	5330	3035	268	3468	1514	489	2724	760	1171	1845	1312	3212	32242
7	1337	6692	7142	3747	485	3564	1308	697	2569	717	924	1751	1205	3093	35231

8	423	3082	2738	1649	261	1398	427	289	932	259	330	660	495	1205	14148
Totals	7785	29764	29259	16517	1667	17476	6876	2664	13531	3579	5292	8830	6295	15939	165474

Expected Months/Color	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total
4	2267.602252	8669.609945	8522.514359	4811.045137	485.561073	5090.381111	2002.830197	775.965626	3941.28787	1042.485351	1541.44523	2571.988167	1833.597453	4642.686229	48199
5	1677.401827	6413.126268	6304.316001	3558.849837	359.1816116	3765.48161	1481.543348	574.0010878	2915.468738	771.1523623	1140.245404	1902.563666	1356.357676	3434.310562	35654
6	1516.878603	5799.405876	5701.008485	3218.276672	324.8088159	3405.134293	1339.763298	519.0705972	2636.46556	697.3549802	1031.126727	1720.493008	1226.557586	3105.655499	32242
7	1657.501088	6337.040768	6229.521429	3516.627549	354.9202715	3720.807837	1463.966279	567.1911237	2880.87954	762.0033903	1126.717502	1879.9916	1340.265812	3393.565811	35231
8	665.6162297	2544.817143	2501.639726	1412.200805	142.528228	1494.195148	587.8968781	227.7715653	1156.898292	306.003916	452.4651365	754.9635592	538.2214729	1362.781899	14148
Totals	7785	29764	29259	16517	1667	17476	6876	2664	13531	3579	5292	8830	6295	15939	165474

Pearson Residuals	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total
4	24.22925112	1.286500835	18.44803424	0.3168797382	29.93438548	0.1723404501	2.746693006	9.523732249	13.74041614	0.261618665	4.209731772	0.00153686916 8	0.0314798091	1.718175504	48199
5	30.072901	35.94521972	23.06386495	16.84573546	14.10657358	6.842724534	3.163129301	9.799917262	16.08173214	0.2140456315	9.623827748	5.408139619	7.46770388	20.09304033	35654
6	7.147088258	16.18865158	24.14437658	10.43736816	9.935818871	1.160628855	22.65954614	1.742038213	2.906269053	5.627547823	18.97393602	9.010202886	5.951947286	3.641470521	32242
7	61.97338151	19.8824753	133.6566783	15.09158009	47.67475159	6.608429885	16.61614782	29.70840633	33.76359411	2.657868929	36.47266126	8.850482546	13.65165012	26.62090902	35231
8	88.43329279	113.3933816	22.33182444	39.70671775	98.47565607	6.192970577	44.03460256	16.45912741	43.71969616	7.220064861	33.14666358	11.94505016	3.470868054	18.26787379	14148
Totals	7785	29764	29259	16517	1667	17476	6876	2664	13531	3579	5292	8830	6295	15939	165474

Chi-Square Test Statistic	
1444.903028	
DF=(R-1)(C-1) =4*13	52
Significance:	Use python because I can never remember

	the excel codes there's too many
from scipy.stats import chi2	
>>> from scipy.stats import chi2	
>>> p_value = 1-chi2.cdf(1444.903028,52)	
>>> p_value	
0	
>>> print(f"p-value: {p_value:.2e}")	
p-value: 0.00e+00	
>>> chi_square_statistic=1444.903028	
>>> degrees_of_freedom=52	
>>> log_p_value = np.log10(1 - chi2.cdf(chi_square_statistic, degrees_of_freedom))	
Traceback (most recent call last):	
File "<stdin>", line 1, in <module>	
NameError: name 'np' is not defined	
>>> import numpy as np	

>>> log_p_value = np.log10(1 - chi2.cdf(chi_square_statistic, degrees_of_freedom))
<stdin>:1: RuntimeWarning: divide by zero encountered in log10
>>> log_p_value = np.log10(p_value + 1e-100)
>>>
>>> print(f"log(p-value): {log_p_value:.2f}")
log(p-value): -100.00
So, it's basically 0