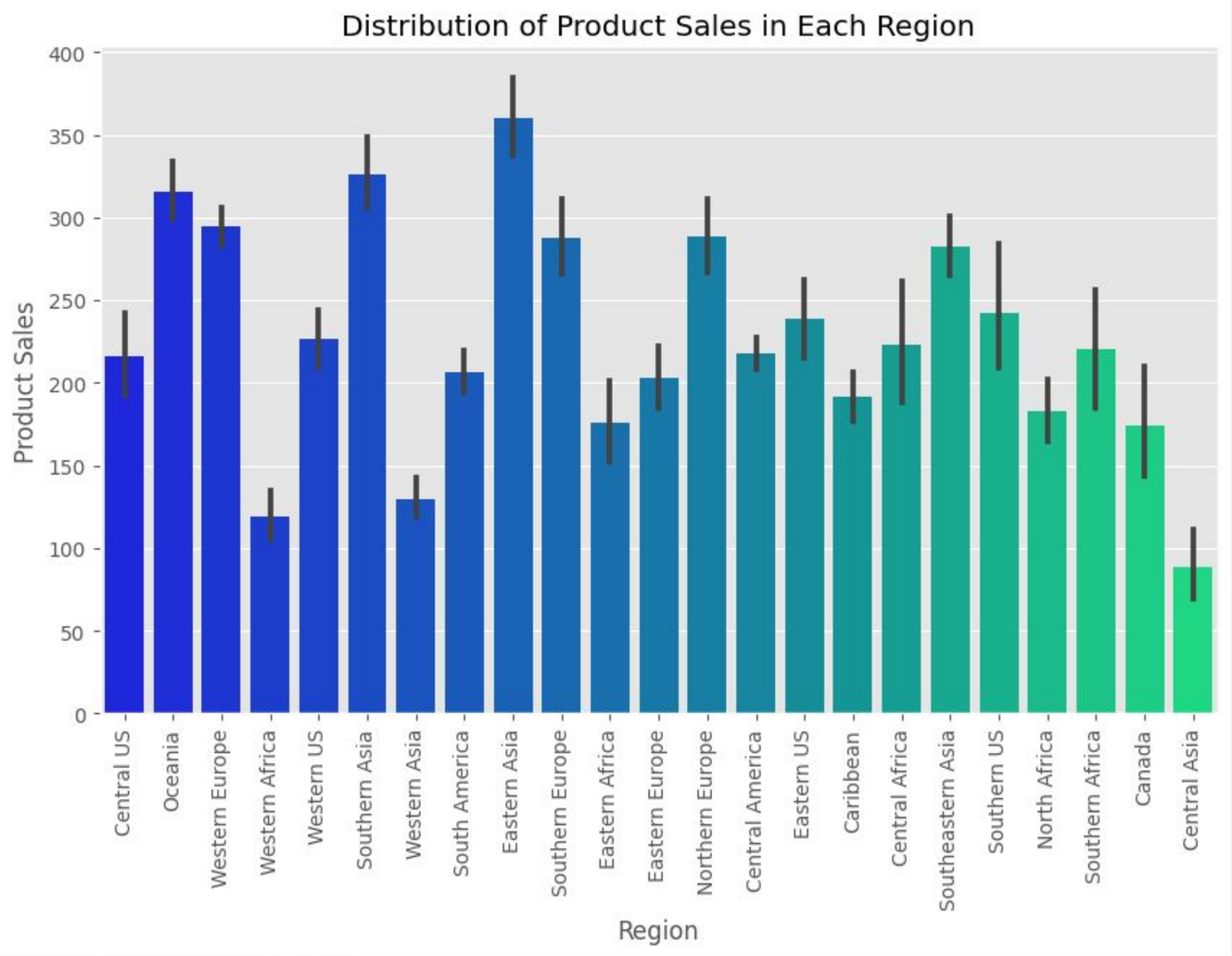
Product Performance by Region

Using Anova (Analysis of Variance)

by Aditya Virgiansyah

Background

In light of the dynamic nature of consumer preferences across diverse geographic regions, this project aims to employ Analysis of Variance (ANOVA) to systematically assess and compare the mean performance of our product in various regions, providing valuable insights for tailored marketing strategies and informed business decision-making.



Objective

The objective of this project is to utilize Analysis of Variance (ANOVA) to compare the mean performance of the product across different regions

Analysis Stage

Data Preparation

Data Cleaning and Explore Data

EDA

Univariate Analysis and Data Visualization

Hypothesis Testing

Statistic Test (Anova)

Analysis Process and Result

Access the notebook

Exploratory Data Analysis

	Row_ID	Order_ID	Order_Date	Ship_Date	Ship_Mode	Customer_ID	Customer_Name	Segment	Postal_Code	City	...	Product_ID	Category	Sub-Category	Product_Name	Sales	Quantity	Discount
0	40098	CA-2014-AB10015140-41954	2014-11-11	2014-11-13	First Class	AB-100151402	Aaron Bergman	Consumer	73120.0	Oklahoma City	...	TEC-PH-5816	Technology	Phones	Samsung Convoy 3	221.980	2	0.0
1	26341	IN-2014-JR162107-41675	2014-02-05	2014-02-07	Second Class	JR-162107	Justin Ritter	Corporate	NaN	Wollongong	...	FUR-CH-5379	Furniture	Chairs	Novimex Executive Leather Armchair, Black	3709.395	9	0.1
2	25330	IN-2014-CR127307-41929	2014-10-17	2014-10-18	First Class	CR-127307	Craig Reiter	Consumer	NaN	Brisbane	...	TEC-PH-5356	Technology	Phones	Nokia Smart Phone, with Caller ID	5175.171	9	0.1
3	13524	ES-2014-KM1637548-41667	2014-01-28	2014-01-30	First Class	KM-1637548	Katherine Murray	Home Office	NaN	Berlin	...	TEC-PH-5267	Technology	Phones	Motorola Smart Phone, Cordless	2892.510	5	0.1
4	47221	SG-2014-RH9495111-41948	2014-11-05	2014-11-06	Same Day	RH-9495111	Rick Hansen	Consumer	NaN	Dakar	...	TEC-CO-6011	Technology	Copiers	Sharp Wireless Fax, High-Speed	2832.960	8	0.0

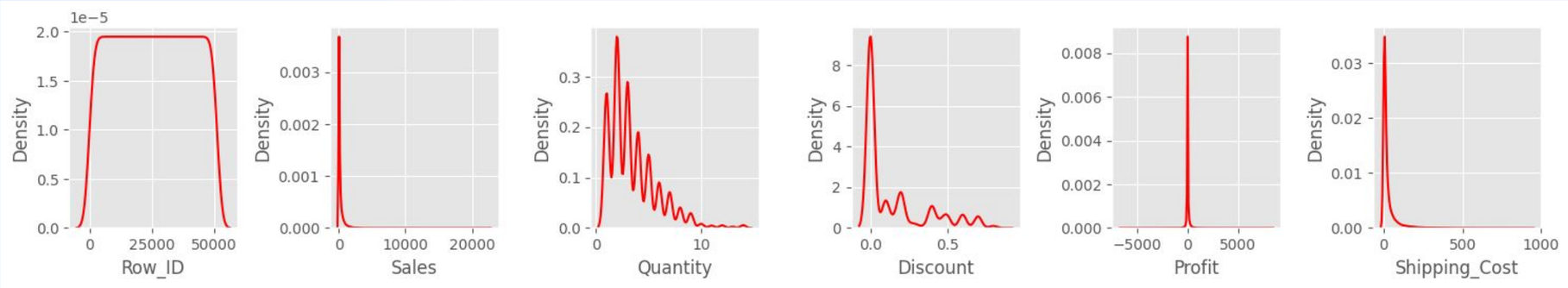
5 rows × 24 columns

There’s 24 columns with the composition of:

- **17 object cololumns:** (Order_ID, Order_Date, Ship_Date, Ship_Mode, Customer_ID, Customer_Name, Segment, City, State, Country, Region, Market, Product_ID, Category, Sub-Category, Product_Name, Order_Priority
- **2 intenger columns:** (Row_ID, Quantity)
- **5 float columns:** (Postal_Code, Sales, Discount, Profit, Shipping_Cost)

Univariate Analysis

Numerical



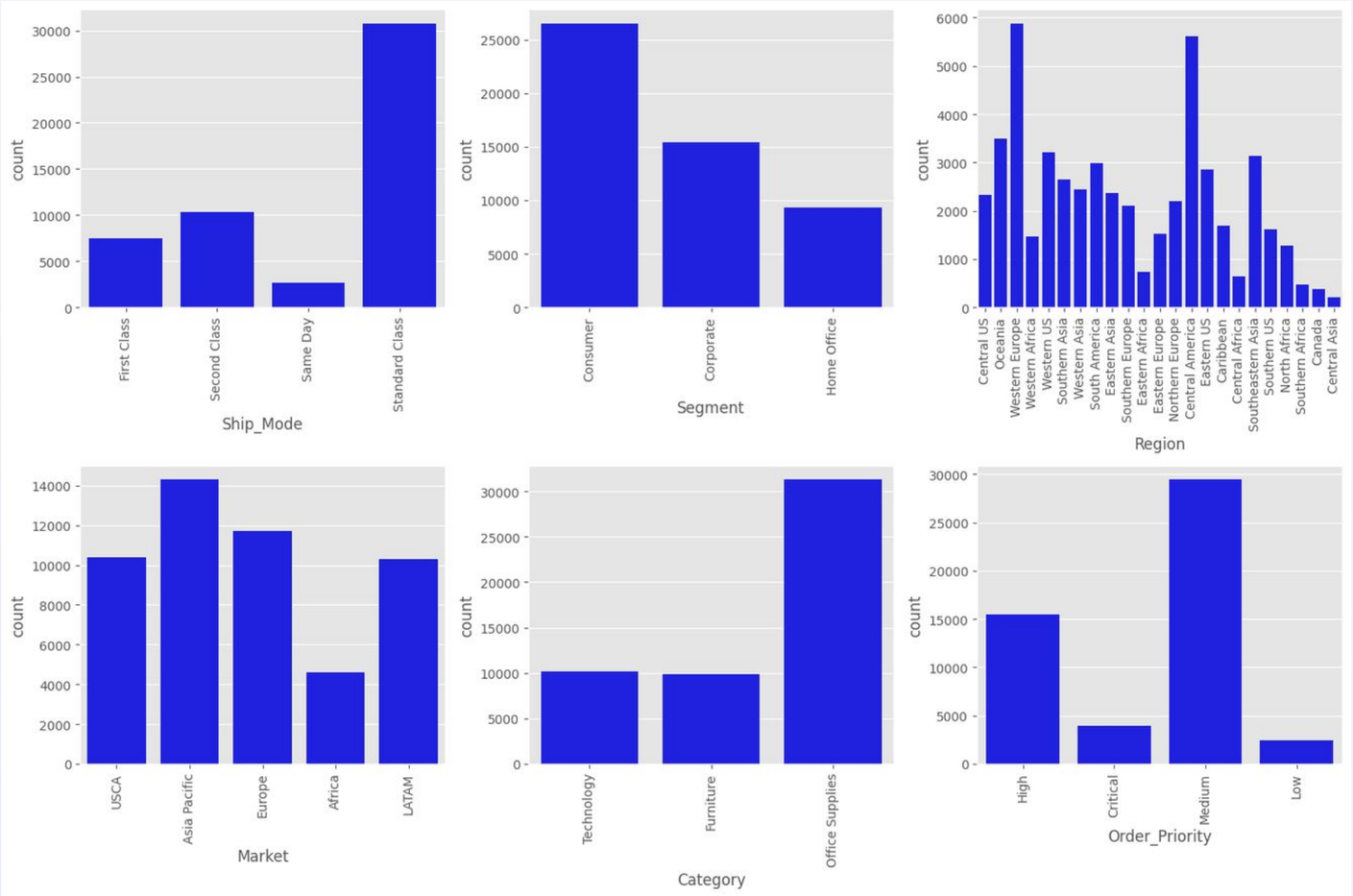
KDE Plot Analysis: It can be seen that Sales, Quantity, Discount, and Shipping_Cost has a positively skewed distribution.

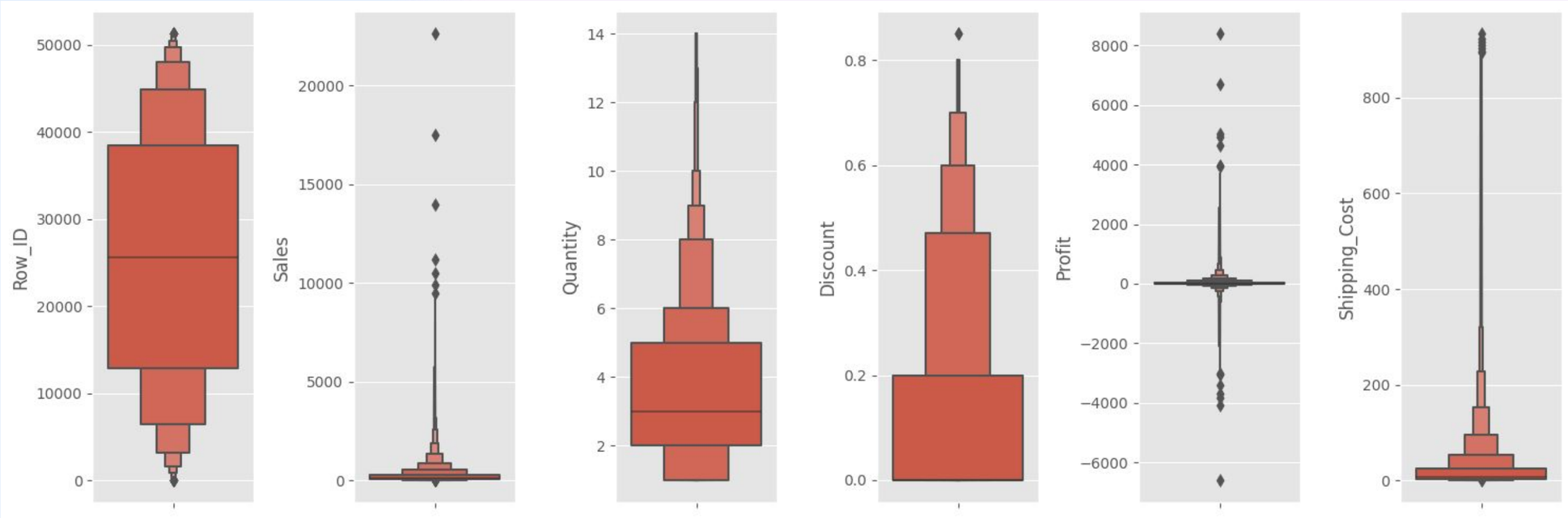
Univariate Analysis

Categorical

observation result:

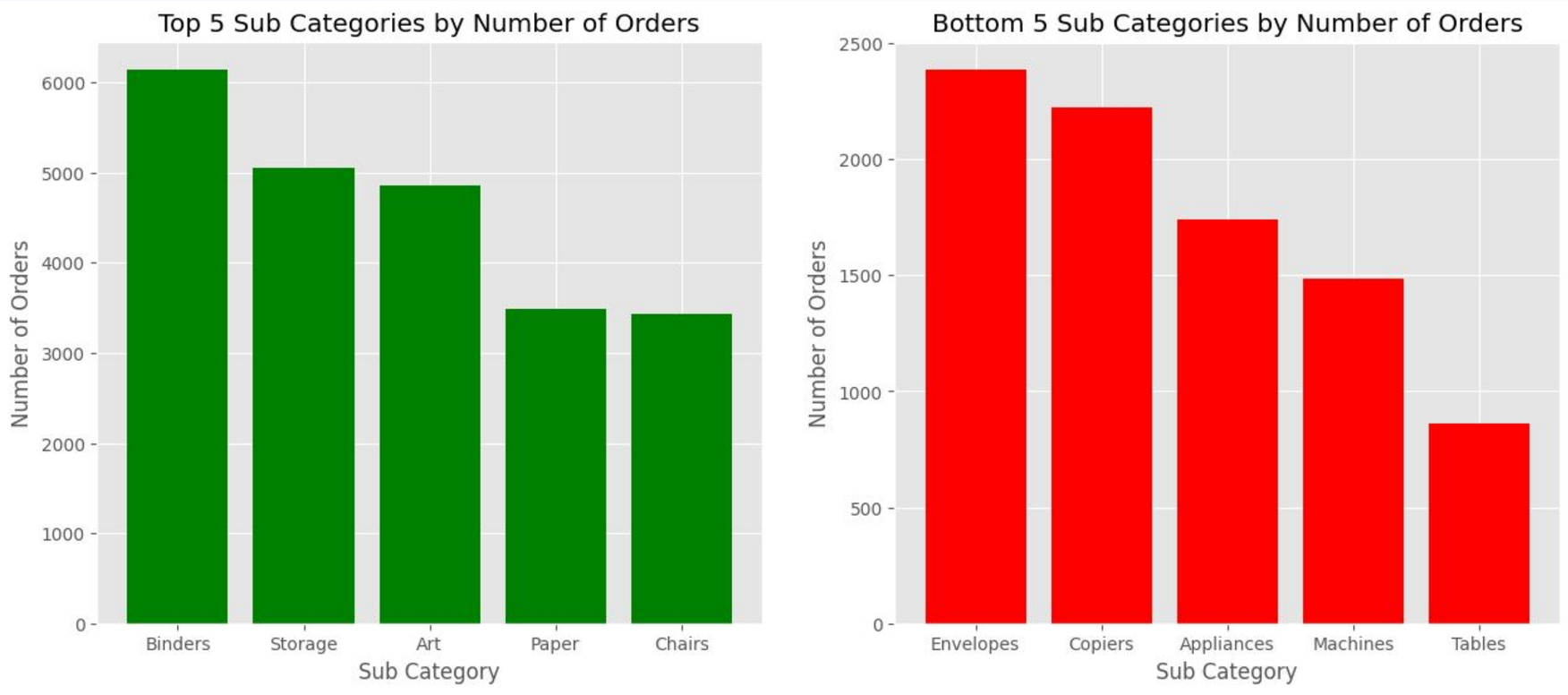
- From the plot calculations, it turns out that the customer composition mostly uses standard class for the shipping mode of their goods.
- Western Europe is the largest customer buyer with a gap that tends to be far from other regions except for Central America.
- The most ordered category are office supplies.





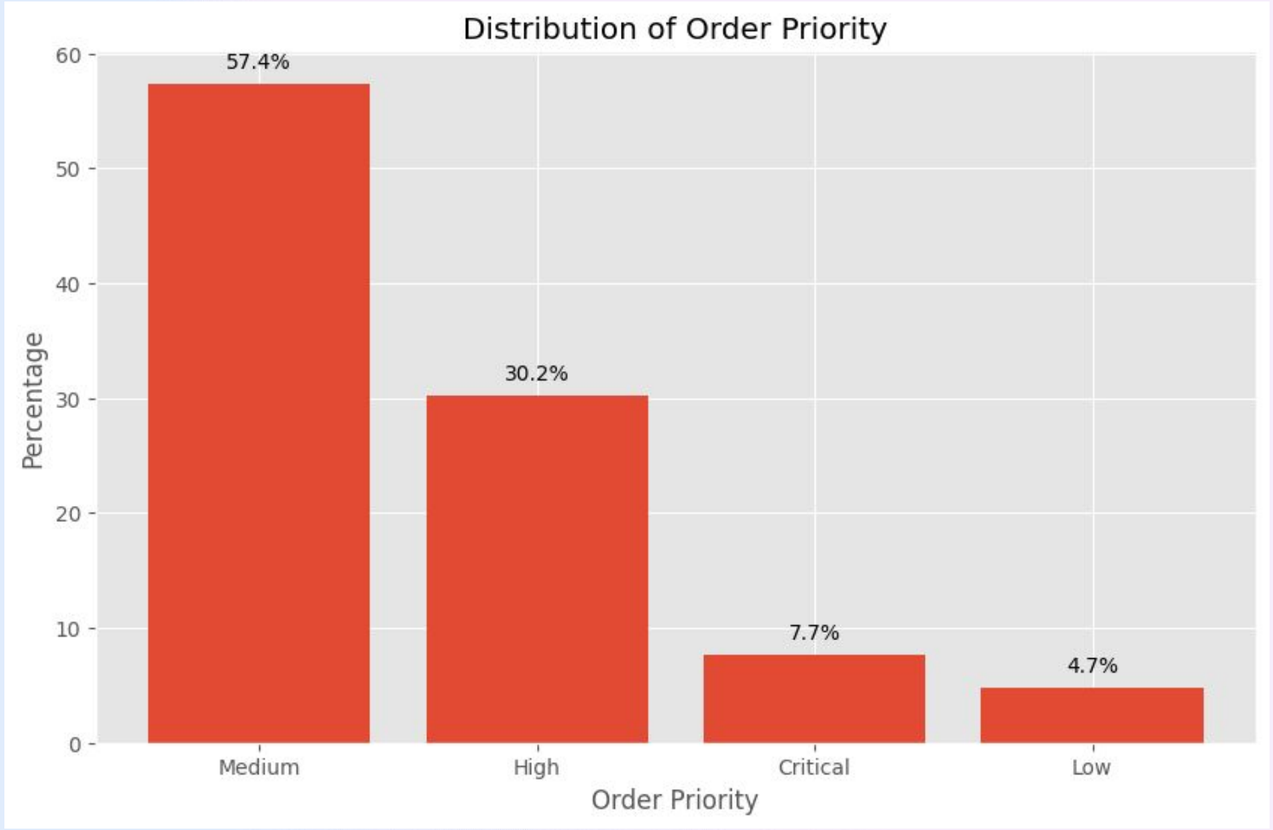
although there are outliers in the boxplot, In many businesses, it's normal to have occasional spikes or dips due to factors such as seasonal trends, promotions, or special events. These fluctuations may result in outliers, but they are a natural part of business operations.

sub-category with most purchased and least popular



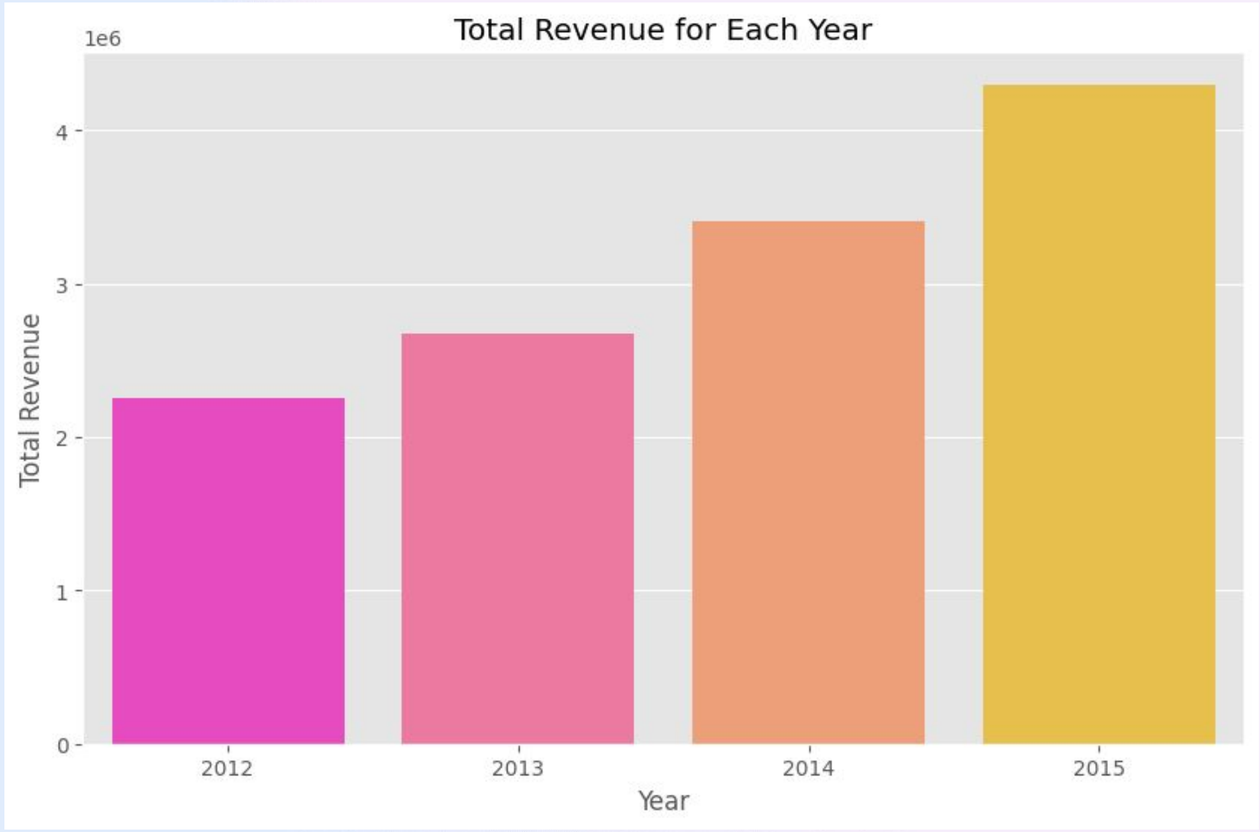
the top-selling sub-categories include organizational products like binders and storage, creative items in the art category, essential office supplies such as paper, and furniture represented by chairs these insights can inform inventory management and marketing strategies for these product categories and in the other side bottom 5 sub-categories represent products with lower sales volumes, and businesses may need to assess market demand and adjust inventory strategies accordingly for these items.

distribution of order priority



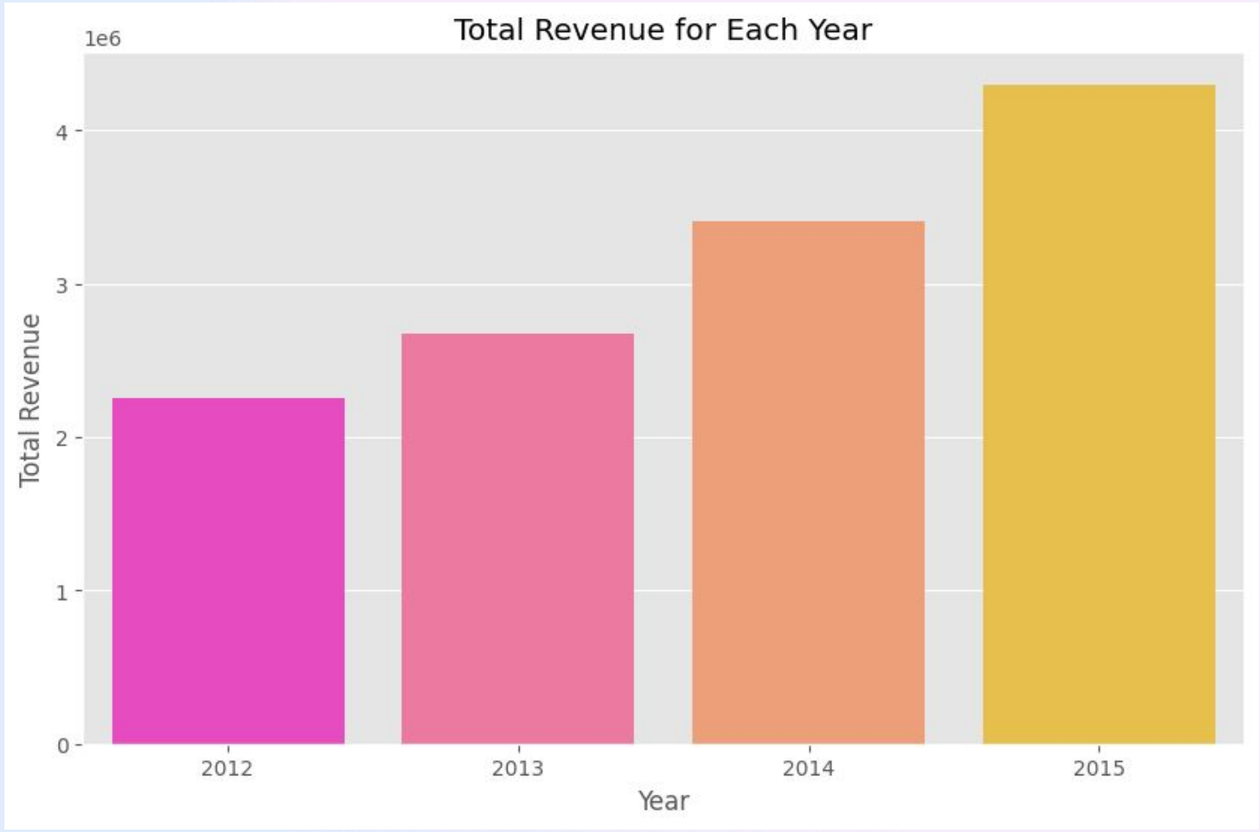
Around 57.4% of order priority is medium

Total Revenue each Year



Every year total revenue increase

Total Revenue each Year



Every year total revenue increase

Hypothesis Testing

Designing the Hypothesis to Test

- The confidence level was 95%, so the significance level(alpha) is 5% or 0.05
- H0: The null hypothesis states that the means (μ) of product sales across all regions (k regions) are equal or The average sales performance of a product is consistent across different regions.
- H1:There is a significant difference in the means of product sales across regions.

ANOVA Test

```
ANOVA F-statistics: 35.76168437710719
ANOVA p-value: 6.163078664079868e-151
```

The extremely low p-value (6.163078664079868e-151) is well below a commonly used significance level of 0.05 Therefore, there is strong evidence to reject the null hypothesis (Ho), which states that the means of product sales across regions are equal.

Post-Hoc Tests ANOVA (pairwise_tukeyhsd)

ANOVA Table:

	sum_sq	df	F	PR(>F)
Region	1.842801e+08	22.0	35.761684	6.163079e-151
Residual	1.200814e+10	51267.0	NaN	NaN

Tukey HSD Results:

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Canada	Caribbean	17.5901	1.0	-81.3391	116.5192	False
Canada	Central Africa	49.0827	0.9964	-63.778	161.9434	False
Canada	Central America	43.4965	0.9897	-48.8085	135.8015	False
Canada	Central Asia	-85.2992	0.9126	-233.9169	63.3185	False
Canada	Central US	41.4806	0.9969	-54.9207	137.8818	False
Canada	Eastern Africa	1.3343	1.0	-109.0354	111.704	False
Canada	Eastern Asia	185.8846	0.0	89.6303	282.1388	True
Canada	Eastern Europe	28.4767	1.0	-71.4122	128.3655	False
Canada	Eastern US	64.044	0.705	-31.0885	159.1765	False
Canada	North Africa	8.1935	1.0	-93.6453	110.0323	False
Canada	Northern Europe	114.6277	0.004	17.8579	211.3974	True
Canada	Oceania	141.2182	0.0	47.1271	235.3094	True
Canada	South America	32.2754	0.9999	-62.5919	127.1427	False
Canada	Southeastern Asia	108.3615	0.0072	13.7379	202.9851	True
Canada	Southern Africa	45.7743	0.9995	-74.1488	165.6974	False
Canada	Southern Asia	152.1006	0.0	56.5582	247.643	True
Canada	Southern Europe	113.7315	0.0048	16.6531	210.8099	True
Canada	Southern US	67.5115	0.6874	-31.8125	166.8356	False
Canada	Western Africa	-55.197	0.9417	-155.5585	45.1645	False
Canada	Western Asia	-44.3302	0.9921	-140.4031	51.7426	False
Canada	Western Europe	120.1036	0.0006	27.9327	212.2744	True

The Tukey Honestly Significant Difference (HSD) test is conducted to identify which specific regions exhibit statistically significant differences in product performance. The results are presented in a pairwise manner:

- Canada vs. Eastern Asia: Statistically significant difference (reject the null hypothesis).
- Eastern US vs. Western Asia: Statistically significant difference.
- North Africa vs. Southern Asia: Statistically significant difference (reject the null hypothesis).

Several other pairwise comparisons also show significant differences, including comparisons involving regions such as Central Asia, Eastern Asia, South America, and Western Africa.

Conclusion

- The conclusion is that there is at least one significant difference in the means of product sales across regions.
- We don't have enough statistical evidence to conclude that the average sales performance of a product is consistent across different regions.
- In other words, the analysis suggests that there are variations in product sales among different regions, and these differences are not likely due to random chance.