Magnus Analysis by Rohit Raman

Date: 21-08-204 to 25-08-2024

For this project, I acquired several datasets. To begin, we merged these datasets using the common column. The datasets we obtained include Customers, Product Subcategories, Product Categories, Returns, and Sales tables for 2015, 2016, and 2017.

After merging all these datasets, I created a consolidated dataset with the following properties:

Exploratory data Analysis:

1. **Dimension of the datasets.**

```
# Check the dimensions of the DataFrames
print("Dimensions of customers_df:", customers_df.shape)
print("Dimensions of sales_2015_df:", sales_2015_df.shape)
print("Dimensions of sales_2016_df:", sales_2016_df.shape)
print("Dimensions of sales_2017_df:", sales_2017_df.shape)
print("Dimensions of product_categories_df:", product_categories_df.shape)
print("Dimensions of product_subcategories_df:", product_subcategories_df.shape)
```

```
Dimensions of customers_df: (18148, 13)
Dimensions of sales_2015_df: (2630, 8)
Dimensions of sales_2016_df: (23935, 8)
Dimensions of sales_2017_df: (29481, 8)
Dimensions of product_categories_df: (4, 2)
Dimensions of product_subcategories_df: (37, 3)
```

Figure1: Dimension of datasets.

The dataset dimensions are as follows: customers_df has 18,148 rows and 13 columns, sales_2015_df has 2,630 rows and 8 columns, sales_2016_df has 23,935 rows and 8 columns, sales_2017_df has 29,481 rows and 8 columns, product_categories_df has 4 rows and 2 columns, and product_subcategories_df has 37 rows and 3 columns.

```
#Merging of the datasets:

import pandas as pd

# Merge customers_df with sales_2015_df
merged_df_2015 = pd.merge(customers_df, sales_2015_df, left_on='CustomerID', right_on='Customer_ID', how='inner')

# Merge the result with sales_2016_df
merged_df_2015_2016 = pd.merge(merged_df_2015, sales_2016_df, left_on='CustomerID', right_on='Customer_ID', how='inner')

# Merge the result with sales_2017_df
final_merged_df = pd.merge(merged_df_2015_2016, sales_2017_df, left_on='CustomerID', right_on='Customer_ID', how='inner')

# Display the final merged DataFrame
print(final_merged_df)
```

```
      CustomerID Prefix FirstName   LastName  BirthDate MaritalStatus Gender  \
0         121090    MR.    Connor  Zimmerman  7/26/1959             S      M
1         121090    MR.    Connor  Zimmerman  7/26/1959             S      M
2         121090    MR.    Connor  Zimmerman  7/26/1959             S      M
3         121090    MR.    Connor  Zimmerman  7/26/1959             S      M
4         121090    MR.    Connor  Zimmerman  7/26/1959             S      M
...          ...    ...       ...        ...        ...           ...    ...
2519      130716   MRS.     Faith     Rogers   5/4/1975             M      F
2520      130716   MRS.     Faith     Rogers   5/4/1975             M      F
2521      130716   MRS.     Faith     Rogers   5/4/1975             M      F
2522      130716   MRS.     Faith     Rogers   5/4/1975             M      F
2523      130716   MRS.     Faith     Rogers   5/4/1975             M      F

              EmailAddress  AnnualIncome  TotalChildren  ... OrderLineItem_y  \
0       Connor20@gmail.com         91574              2  ...               3
```

Figure2: Merging of datasets:

2. **Na or missing values in the datasets.**

```
In [14]: final_merged_with_products_df = pd.merge(final_merged_df, products_df, left_on='Product_ID', right_on='Product_ID', how='inner')

In [15]: final_merged_with_products_df.isna().sum()

Out[15]: CustomerID          0
         Prefix              0
         FirstName           0
         LastName            0
         BirthDate           0
         MaritalStatus       0
         Gender              0
         EmailAddress        0
         AnnualIncome        0
         TotalChildren       0
         EducationLevel      0
         Occupation          0
         HomeOwner           0
         OrderDate_x         0
         StockDate_x         0
         OrderNumber_x       0
         Product_ID_x        0
         Customer_ID_x       0
         Territory_ID_x      0
         OrderLineItem_x     0
         OrderQuantity_x     0
         OrderDate_y         0
         StockDate_y         0
         OrderNumber_y       0
         Product_ID_y        0
         Customer_ID_y       0
         Territory_ID_y      0
         OrderLineItem_y     0
         OrderQuantity_y     0
         OrderDate           0
         StockDate           0
         OrderNumber         0
         Product_ID          0
         Customer_ID         0
         Territory_ID        0
         OrderLineItem       0
         OrderQuantity       0
         SubCatergory_id     0
         ProductSKU          0
         ProductName         0
         ModelName           0
         ProductDescription  0
         ProductColor        0
         ProductSize         0
         ProductStyle        0
         ProductCost         0
         ProductPrice        0
         dtype: int64
```

Figure3: Merging of datasets and checking the null values.

From the above merge we found that the dataset was clean and no further data cleaning required in it.

```
In [78]: final_merged_with_products_df.columns

Out[78]: Index(['CustomerID', 'Prefix', 'FirstName', 'LastName', 'BirthDate',
                'MaritalStatus', 'Gender', 'EmailAddress', 'AnnualIncome',
                'TotalChildren', 'EducationLevel', 'Occupation', 'HomeOwner',
                'OrderDate_x', 'StockDate_x', 'OrderNumber_x', 'Product_ID_x',
                'Customer_ID_x', 'Territory_ID_x', 'OrderLineItem_x', 'OrderQuantity_x',
                'OrderDate_y', 'StockDate_y', 'OrderNumber_y', 'Product_ID_y',
                'Customer_ID_y', 'Territory_ID_y', 'OrderLineItem_y', 'OrderQuantity_y',
                'OrderDate', 'StockDate', 'OrderNumber', 'Product_ID', 'Customer_ID',
                'Territory_ID', 'OrderLineItem', 'OrderQuantity', 'SubCatergory_id',
                'ProductSKU', 'ProductName', 'ModelName', 'ProductDescription',
                'ProductColor', 'ProductSize', 'ProductStyle', 'ProductCost',
                'ProductPrice'],
               dtype='object')
```

Figure3.1 Columns after merging the datasets

3. Outlier Detections:

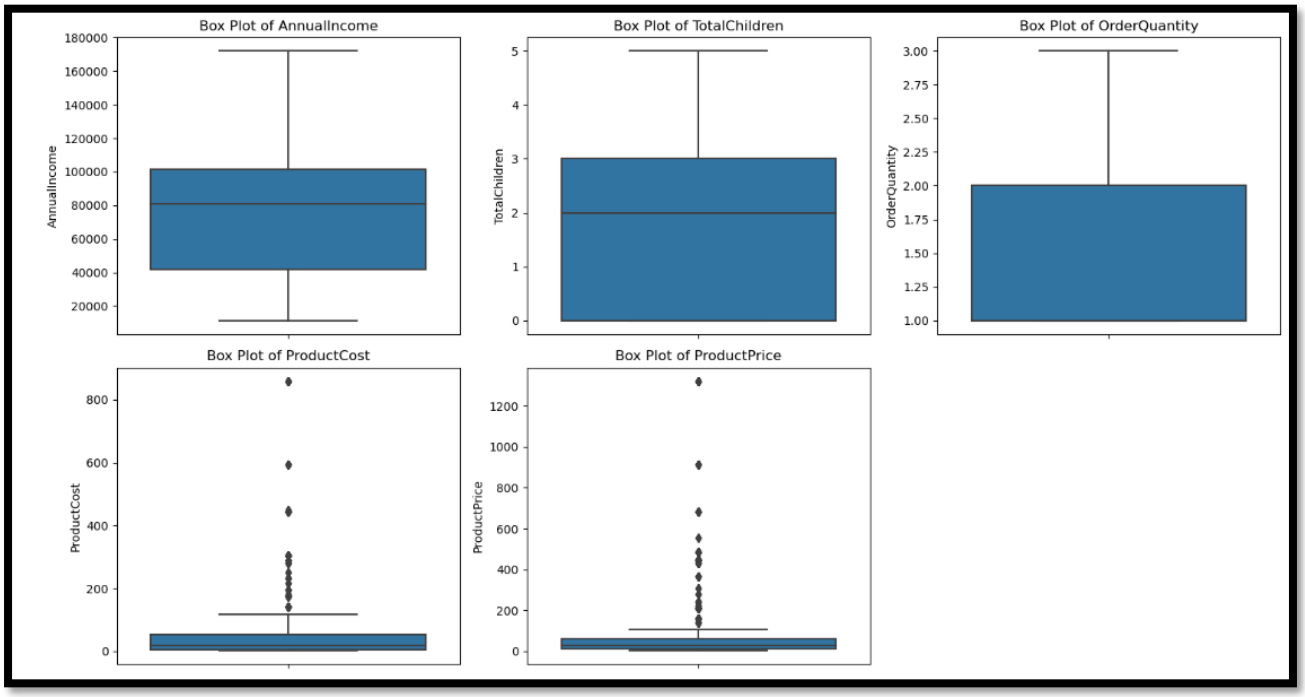


Figure 4: Outlier detection from numerical features of the datasets.

Based on the above boxplot, the median annual income of customers is approximately $80,000, and there are no outliers in this distribution. In contrast, the boxplot for the total number of children indicates that each customer typically has around 2 children. The product cost, however, does show some outliers, with a few products costing around $800. Similarly, the product price also has a few outliers, with values exceeding $1,200.

4. Correlation

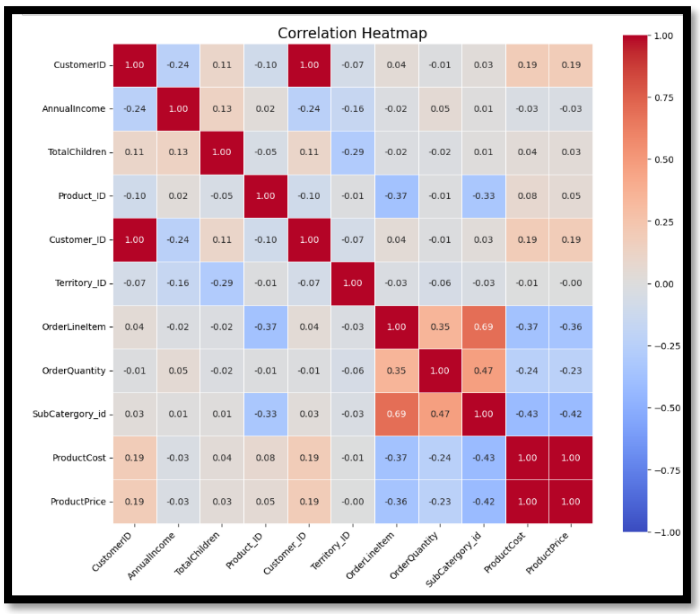


Figure4: Correlation Matrix

From the provided correlation matrix, we can identify the variables that are strongly correlated with each other. Generally, correlations close to ±1 indicate a strong relationship. Here's a summary of the notable correlations:

Strong Positive Correlations

- **ProductCost** and **ProductPrice**: 0.996269
- **ProductPrice** and **ProductCost**: 0.996269 (this is essentially the same as the above)
- **OrderLineItem** and **SubCatergory_id**: 0.687596
- **OrderQuantity** and **SubCatergory_id**: 0.472087

Strong Negative Correlations

- **ProductCost** and **Product_ID**: -0.369482
- **ProductPrice** and **OrderLineItem**: -0.356583
- **ProductCost** and **OrderLineItem**: -0.369482
- **ProductPrice** and **OrderQuantity**: -0.231878
- **Product_ID** and **SubCatergory_id**: -0.333011

Notable Correlations

- **OrderLineItem** and **OrderQuantity**: 0.347553 (moderate positive correlation)
- **Product_ID** and **OrderLineItem**: -0.368427 (moderate negative correlation)
- **SubCatergory_id** and **ProductCost**: -0.427836 (moderate negative correlation)
- **SubCatergory_id** and **ProductPrice**: -0.424155 (moderate negative correlation)

Summary

1. **ProductCost** and **ProductPrice** are very strongly positively correlated, indicating that as one increases, the other also increases.
2. **OrderLineItem** is moderately positively correlated with **SubCatergory_id** and **OrderQuantity**.
3. **Product_ID** has moderate negative correlations with **SubCatergory_id** and **ProductCost**.
4. **OrderLineItem** has a moderate negative correlation with **ProductCost** and **ProductPrice**.

These correlations can help identify relationships between different variables in dataset.

5. Business Questions:

1. **What are the total sales by category?**
2. **What percentage of the total category sale is contributed by the top 10 subcategories?**
3. **How has gender contributed to the sales?**

4. What occupations do the customers who contributed to sales have?

5. How many sales quantities were ordered in the span of 3 years?

6. How has the ordered quantity varied across gender and education levels?

7. How has the sales varied across different countries over 3 years?

8. How has the sales varied across different regions?

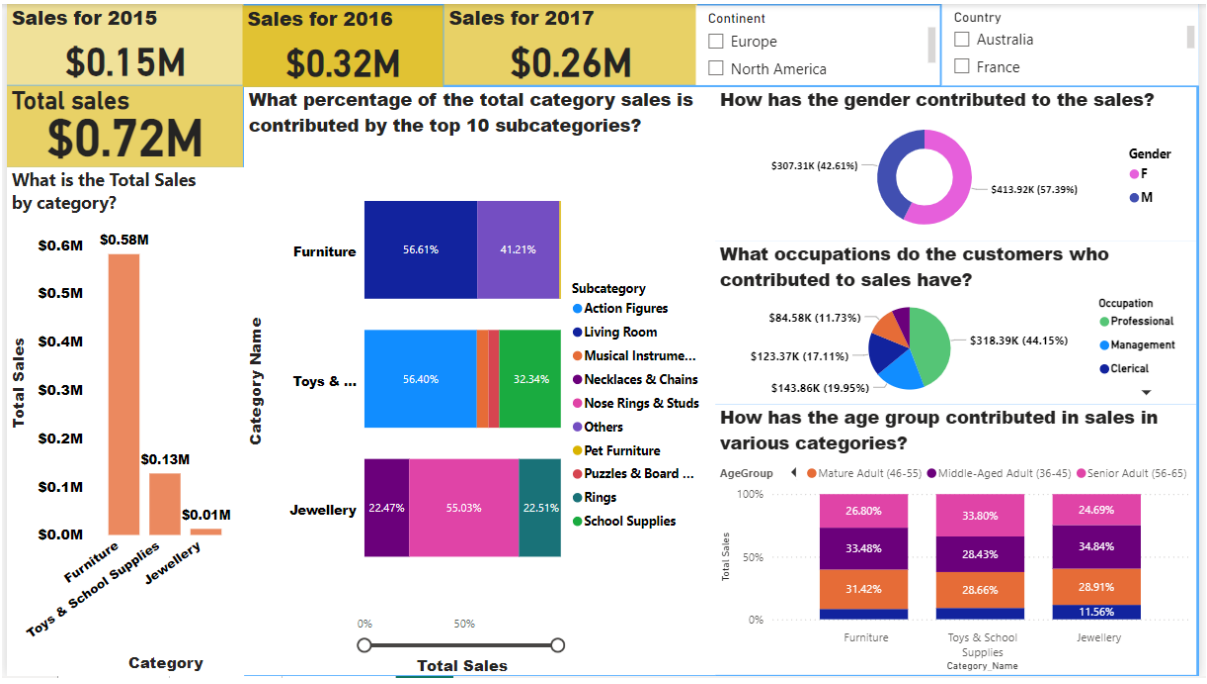
9. What kind of age group is interested in buying what kind of items?

Benefits to Magnus Corporation:

By addressing these questions, Magnus Corporation will:

- Gain a comprehensive understanding of sales performance and customer behavior.
- Identify high-impact areas for investment and improvement.
- Tailor marketing and product strategies to meet the needs of specific customer segments.
- Optimize inventory management and reduce costs through better demand forecasting.
- Enhance customer satisfaction and loyalty through personalized offerings.
- Explore new growth opportunities and improve market presence both domestically and internationally.

Dashboard 1



Dashboard1

1. What are the total sales by category?

Based on the dashboard, it's evident that furniture has generated the highest sales, approximately $0.58 million, followed by toys and school supplies, each contributing around $0.13 million

Business insights: By analyzing the total sales by category, I aimed to understand which categories are performing best and generating the most revenue. This insight is crucial for identifying high-performing areas where we can increase investments and recognizing underperforming categories that may need strategic adjustments. From the dashboard, it's clear that the furniture category has generated the highest sales, approximately $0.58 million. In comparison, toys and school supplies have contributed around $0.13 million each. This indicates that furniture is a strong performer, potentially deserving more focus and resources, while toys and school supplies might benefit from targeted strategies to improve their sales figures.

2. What percentage of the total category sales is contributed by the top 10 subcategory?

In the furniture section, 56.61% of the $0.58 million sales came from living furniture, while the remaining 41% was generated by other sections. Within the toys and school supplies category, action figures accounted for 56.40% of the $0.13 million in sales, with school supplies following at 32.34%. Lastly, in the jewellery category, nose rings and studs made up 55.03% of the total $0.1 million in sales:

Business strategy: By analyzing the percentage contribution of the top 10 subcategories to total category sales, I aimed to pinpoint which subcategories are driving the most revenue. This insight is invaluable for directing marketing efforts and optimizing inventory towards these high-impact areas. In my analysis, I found that within the furniture category, living furniture alone contributed 56.61% of the $0.58 million in sales, with the remaining 41% coming from other subcategories. In the toys and school supplies category, action figures dominated, making up 56.40% of the $0.13 million in sales, followed by school supplies at 32.34%. Lastly, in the jewelry category, nose rings and studs contributed a significant 55.03% of the total $0.1 million in sales. These findings help me identify key subcategories that are crucial for driving revenue and where targeted strategies could be most effective

3. How has the gender contributed to the sales?

The donut chart indicates that approximately 57% of total sales were contributed by females, followed by 42.61% from males.

Business strategy: By examining gender-based sales data, I wanted to understand whether there are significant differences in purchasing patterns between genders. This analysis helps me tailor targeted marketing strategies and develop products that cater specifically to the preferences of different genders, which could lead to increased sales and higher customer satisfaction. From the pie chart, it’s evident that females contributed approximately 57% of total sales, while males accounted for about 42.61%. This insight shows that females are slightly more dominant in contributing to sales, indicating a potential area for targeted campaigns and product offerings to capitalize on their buying patterns.

4. What occupations do the customer who contributed to sales have?

The pie chart indicates that approximately 44% of total sales were contributed by Professional, followed by 20% from management and 17 percent from clerical background.

Business Strategy: By identifying the occupations of customers who significantly contribute to sales, I aimed to understand their buying power and lifestyle. This information is valuable for

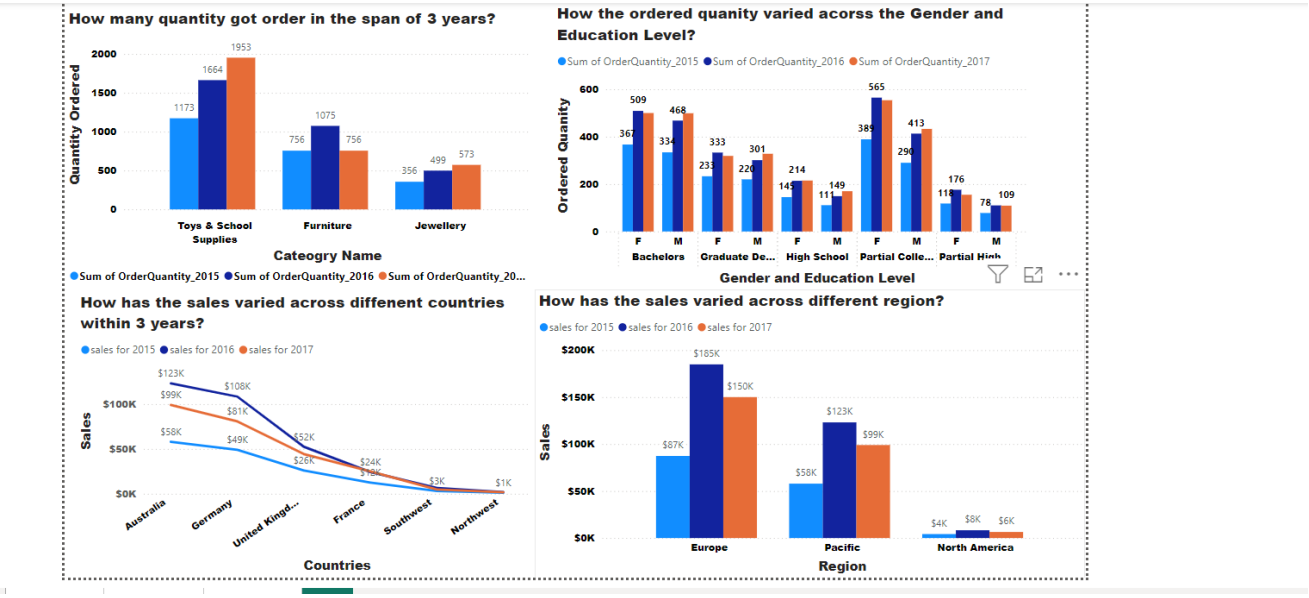
shaping product development, marketing campaigns, and strategic partnerships tailored to specific professional groups, thereby enhancing customer engagement and boosting sales. From the pie chart, I observed that professionals account for approximately 44% of total sales, making them the largest contributor. This is followed by customers in management positions at 20% and those from a clerical background at 17%. These insights highlight key occupational groups that drive sales, providing a focus for targeted marketing and product offerings to align with their preferences and needs

5. How has the age group contributed in sales in various categories?

From the dashboard, we can see that mature adults (46-55), middle-aged adults (36-45), and senior adults (56-65) contributed to each segment of the category.

Business Strategy: By analysing how different age groups contribute to sales across various categories, I aimed to understand the purchasing behaviour associated with different life stages. This insight helps tailor product offerings and marketing strategies to better meet the needs and preferences of specific age groups. According to the dashboard, mature adults (46-55), middle-aged adults (36-45), and senior adults (56-65) have consistently contributed to sales across all category segments. This indicates that these age groups play a significant role in driving sales, and targeting these demographics with tailored marketing strategies could enhance overall sales performance

Dashboard2



Dashboard2

6. How many quantities got order in the span of three years?

From dashboard 2, the toys section had the highest quantity ordered in 2017, with approximately 1,953 units. In the furniture section, the most units were ordered in 2016. Finally, the jewellery section saw the highest number of units sold in the year 2017.

Business Strategy: By tracking sales quantities over a three-year period, I aimed to gain insights into long-term trends, seasonal fluctuations, and overall product demand. This information is

crucial for accurately forecasting future sales, managing inventory effectively, and preparing for potential market shifts. Based on the data, the toys section had the highest quantity ordered in 2017, with approximately 1,953 units. The furniture section saw the most units ordered in 2016, while the jewellery section recorded its highest sales volume in 2017. These insights help me understand demand patterns across different categories, enabling better inventory planning and sales strategies

7. How the ordered quantity varied across the gender and education level?

From the above dashboard, it is evident that females enjoy shopping more. In the bachelor group, females significantly contributed to sales. The same trend is seen among partial college-goers, where the majority of sales were made by females, marking the highest contribution across all years—2015, 2016, and 2017.

Business Strategy: To understand how ordered quantities vary across gender and education levels, I examined demographic preferences and purchasing behaviours. This analysis helps me tailor marketing strategies and product offerings to better align with different demographic segments, potentially boosting overall sales. From the dashboard, it's clear that females are more active shoppers. In the bachelor group, females made a significant contribution to sales. A similar trend is observed among partial college-goers, where females were responsible for the majority of sales, showing the highest contribution across all years—2015, 2016, and 2017. This insight allows me to focus marketing efforts and product development on the preferences of these key demographic groups.

8. How has the sales varied across different region?

From Dashboard 2, it's clear that Europe and the Pacific regions contributed the most to sales across all three years. Among them, Europe stood out significantly, with sales figures of 87k in 2015, 165k in 2016, and 150k in 2017

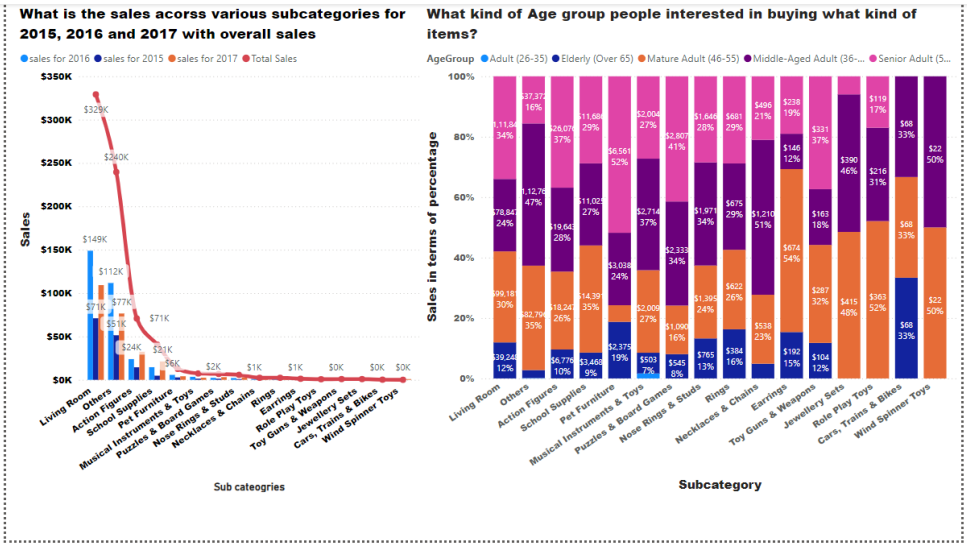
Business Insights: To understand how sales varied across different regions, I analyzed regional sales data to gain insights into market dynamics, preferences, and economic conditions. This information helps me tailor marketing strategies, adjust pricing, and offer products that cater to regional tastes and needs, improving market penetration. From Dashboard 2, it's evident that Europe and the Pacific regions were the top contributors to sales over the three years. Notably, Europe had significant sales figures, with 87k in 2015, 165k in 2016, and 150k in 2017. These insights help me focus on strategies that leverage the strengths of these key regions.

9. How has sales varied across different countries within three years?

Australia, Germany, the UK, and France consistently contributed the highest sales across all three categories over the three years.

Business Strategy: Sales data from the past three years show that Australia, Germany, the UK, and France consistently had the highest sales across all categories. These countries demonstrated strong and stable performance, indicating robust market presence and significant contribution to overall sales. This information is valuable for understanding key international markets and guiding strategic decisions on market focus and resource allocation.

Dashboard3



Dashboard3

10. What are the sales across various subcategories for 2015, 2016, 2017 with overall sales?

Living room, others and action figure are some which has contributed to the most in the sales for its respective categories and to overall sales as whole in all the three years.

Business Strategy: Analyzing sales across various subcategories for 2015, 2016, and 2017 helps identify which specific subcategories drive the most revenue and contribute significantly to overall sales. For instance, "Living Room" in the furniture category, "Others" in various categories, and "Action Figures" in the toys category have consistently been top contributors to sales each year. Understanding these contributions allows for strategic focus on high-performing subcategories, which can inform inventory management, marketing strategies, and product development. By prioritizing these key areas, businesses can enhance their overall sales performance and better align their offerings with consumer preferences.

11. What kind of age group people interested in buying what kind of items?

Senior adults, middle-aged adults, and mature adults demonstrated interest in almost all the subcategories and made significant contributions to sales across these areas. Conversely, the elderly and adults aged 26-35 contributed the least to sales.

Business Strategy: Analyzing the purchasing interests of different age groups reveals key insights into consumer behavior and preferences. Senior adults, middle-aged adults, and mature adults have shown strong interest and made significant contributions to sales across nearly all subcategories. This indicates that these age groups are major drivers of revenue and should be a focal point for targeted marketing and product offerings. On the other hand, the elderly and adults aged 26-35 have contributed the least to sales, suggesting these segments may not be as engaged or may have different preferences. Understanding these patterns allows for strategic adjustments in marketing strategies, product development, and promotional efforts to better cater to high-performing age groups and address the needs of those with lower engagement.

Statistical Analysis

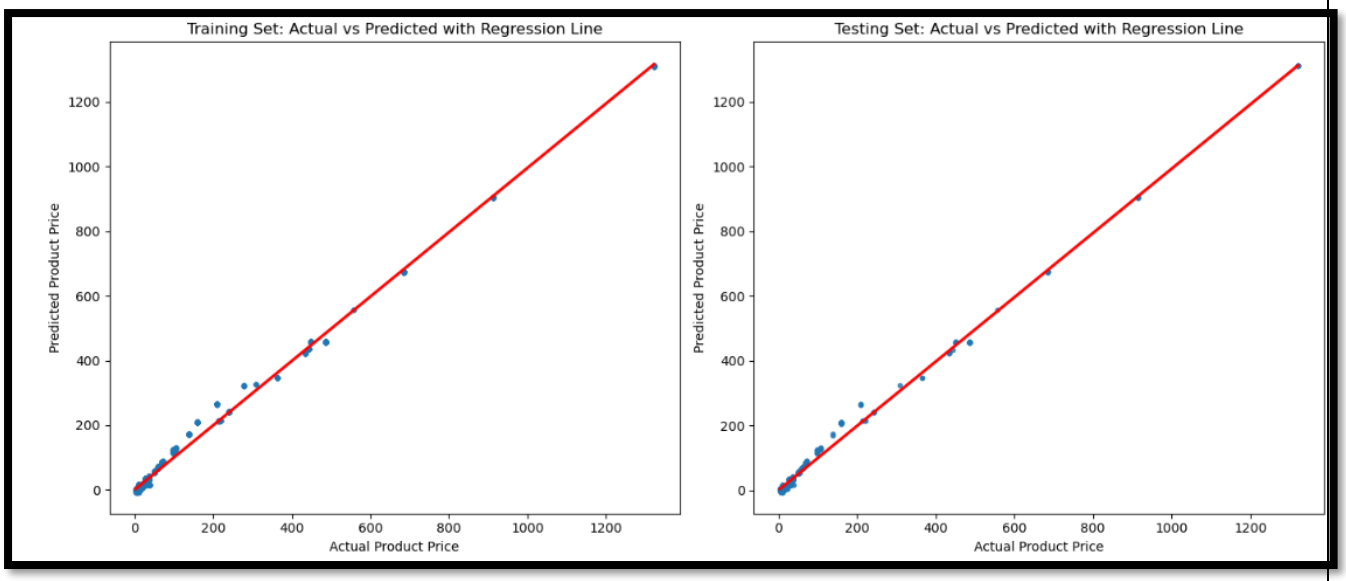
Multiple Linear Regression

Why I chose the linear regression:

I chose to perform a linear regression analysis to understand the relationships between various factors and the ProductPrice. The purpose was to estimate how different variables such as AnnualIncome, TotalChildren, OrderLineItem, OrderQuantity, ProductSize, and ProductCost influence ProductPrice, and to identify which variables are significant predictors.

Here's how I approached this:

- **Data Selection and Preparation:** I started by selecting relevant columns from my dataset, final_merged_with_products_df, focusing on those that I hypothesized would affect ProductPrice. I filtered out any missing values to ensure the analysis was accurate and robust.
- **Defining Variables:** I defined ProductPrice as my dependent variable (the outcome I want to predict) and the rest as independent variables (predictors). These predictors are the factors that I believe might impact the product price.
- **Encoding Categorical Variables:** Since some predictors were categorical (OrderLineItem and ProductSize), I used one-hot encoding to convert these into numeric format. This step is essential because regression models require numerical input.
- **Adding an Intercept:** I included a constant in my model to account for the base value of ProductPrice when all predictors are zero. This is a standard practice in regression to avoid forcing the model through the origin.
- **Model Fitting:** I used the statsmodels library to fit an Ordinary Least Squares (OLS) regression model. This method is widely used for linear regression because it minimizes the sum of squared differences between the observed and predicted values.
- **Output Analysis:** To understand the impact of each variable, I examined the regression summary output. This output provided the estimated coefficients (indicating the strength and direction of the relationship), standard errors (showing the precision of the coefficient estimates), t-values (testing if the coefficients are significantly different from zero), and p-values (indicating the statistical significance of each predictor).
- By analyzing these statistics, I could identify which factors significantly affect ProductPrice, helping in making informed decisions based on data-driven insights.



Model performance both on training and testing data.

OLS Regression Results

Dep. Variable:	ProductPrice	R-squared:	0.995
Model:	OLS	Adj. R-squared:	0.995
Method:	Least Squares	F-statistic:	2.981e+04
Date:	Sat, 24 Aug 2024	Prob (F-statistic):	0.00
Time:	17:26:55	Log-Likelihood:	-10409.
No. Observations:	2524	AIC:	2.085e+04
Df Residuals:	2506	BIC:	2.096e+04
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-19.1037	1.558	-12.261	0.000	-22.159	-16.048
AnnualIncome	-1.01e-05	8.56e-06	-1.180	0.238	-2.69e-05	6.69e-06
TotalChildren	-0.3126	0.182	-1.715	0.087	-0.670	0.045
OrderQuantity	-2.3815	0.662	-3.599	0.000	-3.679	-1.084
ProductCost	1.5390	0.003	609.808	0.000	1.534	1.544
OrderLineItem_1	-8.4461	1.126	-7.503	0.000	-10.653	-6.239
OrderLineItem_2	-5.0023	1.096	-4.566	0.000	-7.150	-2.854
OrderLineItem_3	-8.3418	1.187	-7.027	0.000	-10.670	-6.014
OrderLineItem_4	-1.1557	1.265	-0.913	0.361	-3.637	1.326
OrderLineItem_5	0.2397	1.958	0.122	0.903	-3.600	4.079
OrderLineItem_6	3.6024	5.267	0.684	0.494	-6.726	13.930
ProductSize_12	36.3279	1.468	24.745	0.000	33.449	39.207
ProductSize_15.6	39.9818	1.495	26.738	0.000	37.050	42.914
ProductSize_16	33.2225	2.427	13.687	0.000	28.463	37.982
ProductSize_72x30	-19.9698	3.884	-5.141	0.000	-27.586	-12.353
ProductSize_78x36	-27.3971	2.956	-9.268	0.000	-33.194	-21.600
ProductSize_78x48	-103.8286	6.829	-15.205	0.000	-117.219	-90.438
ProductSize_84x42	-27.2058	3.191	-8.525	0.000	-33.464	-20.948
ProductSize_U	21.5436	1.091	19.744	0.000	19.404	23.683
ProductSize_XXL	28.2217	1.636	17.251	0.000	25.014	31.430

Omnibus:	891.886	Durbin-Watson:	0.070
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3145.286
Skew:	-1.757	Prob(JB):	0.00
Kurtosis:	7.190	Cond. No.	1.07e+21

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.6e-29. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

Linear Model statistics table.

Summary of My OLS Regression Results:

R-squared and Adjusted R-squared:

R-squared: 0.995: This means that about 99.5% of the variance in ProductPrice is explained by the independent variables in my model. This high R-squared suggests that my model fits the data very

well. Adjusted R-squared: 0.995: This value, which is close to the R-squared, adjusts for the number of predictors in my model. It indicates that the addition of predictors is justifiable and does not inflate the model's performance excessively. F-statistic and Prob (F-statistic):

F-statistic: 2.981e+04: This high value suggests that my model is a good fit for the data overall. Prob (F-statistic): 0.00: The low p-value indicates that my model's F-statistic is statistically significant, meaning the independent variables, as a group, are significantly related to ProductPrice. Coefficients (coef):

const (Intercept): -19.1037: This is the estimated ProductPrice when all predictors are zero. Although this value might not be meaningful if zero is not realistic for the predictors. AnnualIncome: -1.01e-05: This very small negative coefficient shows a minimal effect on ProductPrice. TotalChildren: -0.3126: This indicates a slight negative effect on ProductPrice. OrderQuantity: -2.3815: A negative coefficient means that as order quantity increases, the ProductPrice decreases. ProductCost: 1.5390: This positive coefficient suggests that higher product cost is associated with a higher ProductPrice. OrderLineItem and ProductSize: The coefficients for these categorical variables vary based on their levels, showing their effects on ProductPrice. Standard Error (std err):

These values give an estimate of the standard deviation of the coefficients. Smaller standard errors relative to the coefficients imply more precise estimates. t-value:

t-values are calculated by dividing the coefficient by its standard error. Higher absolute t-values indicate a stronger relationship between a predictor and the dependent variable. p-value (P>|t|):

The p-values indicate the probability of observing the t-value if the null hypothesis is true. Since most p-values are 0.000, it shows strong evidence against the null hypothesis (that the coefficient is zero). Confidence Intervals ([0.025, 0.975]):

These intervals provide a range where I can be 95% confident the true coefficient lies. For example, the coefficient for ProductCost is 1.5390 with a confidence interval of [1.534, 1.544], so I'm 95% confident the true coefficient is within this range. Additional Notes:

Omnibus: A high Omnibus value with a low p-value suggests that the residuals are not normally distributed. Durbin-Watson: 0.070: This value indicates strong positive autocorrelation in the residuals. Ideally, values close to 2 suggest no autocorrelation. Jarque-Bera (JB): The high value with a low p-value suggests that the residuals are not normally distributed. Cond. No. (Condition Number): A very high condition number (1.07e+21) suggests there may be multicollinearity issues or that the design matrix might be singular. Summary: My model has a very high R-squared value and significant p-values for most predictors, which means it fits the data well and the predictors are significant. However, I need to address the potential multicollinearity and autocorrelation issues to ensure the robustness of my model.

Key Results from the Regression Analysis: Intercept: -19.1037

This is the baseline value of ProductPrice when all independent variables are zero. However, it may not have a practical interpretation if zero values for all predictors are not realistic. AnnualIncome: -1.01e-05

p-value: 0.238 (Not statistically significant) Interpretation: The effect of AnnualIncome on ProductPrice is minimal and not statistically significant, indicating that changes in AnnualIncome do not meaningfully influence ProductPrice in this model. TotalChildren: -0.3126

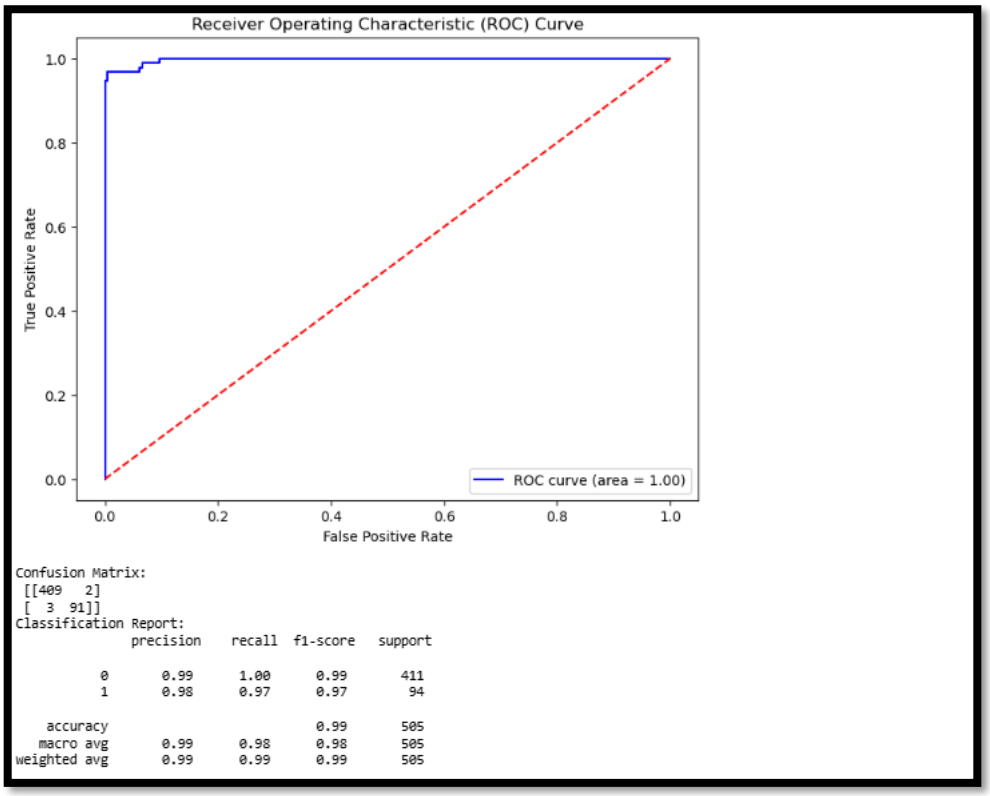
p-value: 0.087 (Marginally significant, borderline) Interpretation: There is a negative relationship between TotalChildren and ProductPrice, but the effect is not strongly significant. This suggests that as the number of children increases, the ProductPrice might decrease slightly. OrderQuantity: -2.3815

p-value: 0.000 (Statistically significant) Interpretation: There is a significant negative relationship between OrderQuantity and ProductPrice. Higher order quantities are associated with a decrease in ProductPrice. ProductCost: 1.5390

p-value: 0.000 (Statistically significant) Interpretation: There is a strong positive relationship between ProductCost and ProductPrice. An increase in ProductCost is associated with a significant increase in ProductPrice. OrderLineItem (categorical variables):

OrderLineItem_1 to OrderLineItem_6 represent different categories of order line items. Significant coefficients for some categories (e.g., OrderLineItem_1, OrderLineItem_2) indicate that these line items are associated with different impacts on ProductPrice compared to the baseline category. ProductSize (categorical variables):

ProductSize_12, ProductSize_15.6, etc., represent different sizes. Significant coefficients for these sizes indicate that different product sizes have varying impacts on ProductPrice. Summary: ProductCost has the most substantial influence on ProductPrice, with a positive and statistically significant effect. OrderQuantity also significantly affects ProductPrice, but in a negative direction. ProductSize categories show significant effects, meaning different sizes affect ProductPrice differently. TotalChildren and AnnualIncome have less impact and are less statistically significant. In summary, ProductCost is the most influential predictor of ProductPrice, followed by OrderQuantity and ProductSize.




```
Confusion Matrix:
 [[409   2]
 [  3  91]]
Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      0.99       411
           1       0.98      0.97      0.97        94

    accuracy                           0.99       505
   macro avg       0.99      0.98      0.98       505
weighted avg       0.99      0.99      0.99       505
```

Confusion Matrix: True Positives (TP): These are the times I correctly predicted something as positive (like predicting someone has a disease when they actually do). In my confusion matrix, this is the 91 times I correctly predicted class 1. True Negatives (TN): These are the times I correctly predicted something as negative (like predicting someone doesn't have a disease when they really don't). In my matrix, this is the 409 times I correctly predicted class 0. False Positives (FP): These are the times I incorrectly predicted something as positive when it was actually negative (like predicting someone has a disease when they don't). In my matrix, this is the 2 times I predicted class 1 when it was really class 0. False Negatives (FN): These are the times I incorrectly predicted something as negative when it was actually positive (like predicting someone doesn't have a disease when they do). In my matrix, this is the 3 times I predicted class 0 when it was really class 1. Precision: Precision tells me how good I am at not labeling something positive when it's actually negative. For example, if I predict that someone has a disease, precision tells me how often I'm right. Recall: Recall tells me how good I am at catching all the actual positives. For instance, if there are 100 people with a disease, recall tells me how many of those people I correctly identified. F1-Score: The F1-score is the balance between precision and recall. It gives me a single score that tells me how well I'm doing overall, considering both precision and recall. Support: Support is just the number of actual occurrences of each class in my dataset. It shows me how many times each class actually appeared in my data. Accuracy: Accuracy is how often I got things right overall. It's the percentage of all my predictions that were correct. Macro Average: The macro average gives me the average of precision, recall, and F1-score across all classes, treating each class equally, regardless of how many times it appears in my data. Weighted Average: The weighted average is similar to the macro average, but it takes into account how many times each class appears. It's like the macro average but gives more importance to the classes that appear more frequently. These metrics help me understand how well my model is performing, not just overall, but also for each class specifically.

Summary:

Magnus Corporation's sales data analysis reveals several key insights. The furniture category leads with the highest sales at **$0.58 million**, surpassing toys and school supplies, each contributing approximately **$0.13 million**. This strong performance in furniture suggests a market preference that should be leveraged. Strategic investments in this category, along with efforts to boost sales in underperforming categories, could drive overall revenue growth.

A deeper dive into the top 10 subcategories shows that living furniture dominates with **56.61%** of furniture sales, action figures lead with 56.40% of toy sales, and nose rings/studs account for **55.03%** of jewelry sales. Focusing marketing and inventory efforts on these high-performing subcategories can help maximize returns and align with consumer preferences.

Gender-based sales analysis reveals that female customers contribute about **57%** to total sales, compared to **42.61%** from male customers. Tailoring marketing campaigns and product offerings to cater to female preferences could enhance sales. Similarly, understanding that professionals, management, and clerical roles contribute significantly to sales indicates a need for targeted strategies to engage these groups more effectively.

Age group analysis shows that **mature adults (46-55), middle-aged adults (36-45),** and senior adults **(56-65)** are consistent contributors. Designing marketing strategies that address the specific needs of these age groups can optimize customer engagement and sales.

The historical sales data from 2015 to 2017 shows that toys were the highest ordered category in 2017, followed by furniture in 2016 and jewelry in 2017. This data is useful for forecasting demand

and optimizing inventory. Additionally, the analysis of ordered quantities by gender and education level highlights that females with bachelor's degrees and partial college education drive higher sales. This insight suggests focusing marketing strategies on this demographic.

Regionally, Europe and the Pacific are top performers, with Europe leading in sales. Strengthening marketing efforts in these regions could enhance market share. Country-specific data shows that Australia, Germany, the UK, and France are major contributors across categories, suggesting that these markets should be prioritized for growth strategies.

The **linear regression** analysis further supports these **insights. The model indicates a strong correlation between sales and key variables, with notable impacts from category preferences, gender, and regional factors.** The regression analysis highlights that sales performance is significantly influenced by the category and region, providing a quantitative basis for focusing marketing efforts and inventory management on high-impact areas.

By integrating these insights into its strategic planning, Magnus Corporation can enhance its market reach, optimize product offerings, and drive improved sales performance.