AUG 2024

MMA 860: FINAL REPORT

EXPLORATION

HOW CAN WE ACCELERATE SALES GROWTH?

ANALYTICS

TOP DATA-DRIVEN TACTICS: OUR ANALYSIS

IMPACT

HARNESSING POWER BI AND STRATEGIC INSIGHTS

Table of Contents:

Executive Summary

In the sprawling digital marketplace of the modern age, e-commerce stands as a towering beacon of convenience and connectivity. In the midst of the expansion of social commerce, lack of brand loyalty, and the emergence of niche competitors, Chow Market Place (a fictional E-commerce company made for this report) finds itself at a critical juncture. A juncture where it must decide the subsequent step in its progression. At present, they are dealing with a problem of sluggish growth in sales on its platform. The fundamental nature of this issue underscores its regrettable impact, a domino effect that starts from a lack of seller interest to remain part of Chown Market Place's ecosystem and ends with a drop in customer interest due to a scarcity of quality products at reasonable prices.

To tackle this quintessential problem of stagnant growth in sales, we plan to leverage the concepts we learned in MMA860 and design a solution that the organization can adhere to moving forward. We have selected 5 activities that will allow us to unearth crucial insights, develop a model, and create a dashboard that will generate perpetual benefits for Chown Market Place. At the end of our exercise, we realized that the root cause of Chown Market Place's stagnant sales is the underutilization of promotional discounts to the customers of different. This issue is more evident when we compare the sales of product categories and establish impact variables in our mode.

Our three-fold recommendation to Chown Market Place includes (1) Future focused data collection needed to develop competitive and customer-centric competencies (2) Strategic deployment of promotional initiatives to a targeted customer segment and (3) Leveraging the business dashboard to track the performance of important metrics such as sales by geography/product/customer segment to spot trends at an early stage.

Introduction

About the Business:

Chown Market Place is an established player in the e-commerce industry. It serves its customers in multiple locations, operating in 147 countries. The company sells products that are in 4 main categories namely, Fashion, Electronics, Auto C Accessories, and Home C Furniture. As a result of their global presence, our team is assuming that the company exercising dynamic pricing based on distinct geographic locations.

Problems Statement: Stagnation of average growth rate in Sales

Indicators of the problem -

a) **The average sales growth rate** for the company over the year was 0.35% indicating stagnation (Figure 11 in appendix).
b) **Geographic performance discrepancies** were identified, suggesting targeted expansion efforts in underperforming regions such as Central Asia (Sri Lanka C Nepal), North Asia (Mongolia C Hong Kong), and the Caribbean (excluding Cuba and the Dominican Republic).

c) **Products (such as "Curtains" and "Keyboards) performance discrepancies**," which together accounted for 2.43% of total orders, significantly impacted the overall sales due to their low average selling prices (Curtains - $34 and Keyboard - $33) compared to the mean sales price of $156.43.

Dataset: https://www.kaggle.com/code/abixnishandh/e-commerce-data-for-regression-analysis/input?select=E-Commerce+data+for+regression.xlsx

Technical summary

The dataset we are using is synthetic data that was pulled from Kaggle to mimic an E-commerce company's sales. There are two main components to the data set. The first is sales information such as product type, shipping mode, and order Id. The second is customer information such as name, segment, location, and market selection.

One assumption was made when interpreting this dataset as no description was provided, the value in the 'sales' column is total sales and not the sale price per product. This is because a common practice in multinational business is to use dynamic pricing (different prices for different geographies) hence we see that the total sale column shows the same amount for the distinct number of units sold.

We chose to do an OLS linear regression as sales is a continuous variable that we knew could be predicted using the linear regression model. The building and testing of the model was entirely done in Python, additionally, data exploration was also conducted in Python.

To demonstrate the relation among data more intuitively, we build a dashboard to help the management team track sales performance and customer behavior while discovering potential business opportunities. The data source comes from Excel and the dashboard will be built using Power BI. The dashboard consists of four filters: date, discount, customer value group (high, medium, low), and high sales product(true and false) to demonstrate the performance of different customer groups and products in each time series. Maintain the consistency of the data (no change in data types and column names in Excel source) and refresh data to ensure that the dashboard reflects the most up-to-date information for the management team.

Execution Activities

Significant Data Exploration

Management: Once the data is imported into Python, we conduct an in-depth analysis of the dataset to check for data issues such as missing values, incompatible datatypes, etc., and to understand underlying correlations to uncover key patterns and insights. For example, does the customer's region have an impact on the amount that they spend? Analysis such as these will be further accompanied by graphs generated from Python, showing different combinations of variables and their relationship trends with sales. We started with macro-level columns such as Region, Product Category, and Segment.

While doing this analysis, we found no significant issues with our data other than having to change some of the column names. In addition to this, we identified some key patterns such as

1. Fashion products are likely to sell more, and Home C Furniture products are likely to sell less.
2. Customer region C Product Segment does not have an impact on sales.

In conclusion, we determined the data was in working condition and we could move forward with creating dummy variables for columns based on the patterns that were identified. We have also provided graphs to identify the month-on-month sales growth rate for additional insights into the business.

Technical: Once the data is imported, numerous tests must be performed in the initial data exploration phase to make sure our data is in a workable format. These tests can be categorized under two steps:

Step 1: Data cleanup

We began with the function data.isnull().sum() to check for null values. Then we used data.dtypes to make sure that all the columns that has numbers were either as float or integers for which we found no issues. The only concern was that we would have to create multiple dummy variables to test different columns for significance on our dependent variable. We then used the data.describe() function to get a better understanding of our data set. Finally, we used the data.columns function to see all the columns we would be working with to determine what we would need to create as dummy variable in our next step which is feature engineering.

Step 2: Data Analysis

We used group by function to better understand the relationship between average sales and certain variables to get a better understanding of our data. We looked at 4 different segments which are regions (figure 2 in appendix), product categories (figure 3 in appendix), segment (figure 4 in appendix), and months (figure 5 in appendix.), and how they compared to average sales. From these visualizations, we found that there was little to no variance in average sales amongst the different regions, segments, and months of the year. However, there was variance in average sales by product category where fashion had the highest average sales, Home C Furniture had the lowest average sales, and Auto/Electronics had the same average sales.

Feature Engineering

Management Description:

Dummy variables definition: a dummy variable is a binary variable (0 or 1) used to include categorical data. It represents the presence (1) or absence (0) of a category, allowing the model to estimate different intercepts for each category.

The second step in creating our model is feature engineering. This is where we essentially create new data that will benefit our regression either by fixing any issues we encounter or simply altering existing data so

that it can be used in regression. In this case, we added new columns which contained dummy variables so that they can be used in our regression. This is because our model cannot read words such as city names, months, shipping methods, or product categories and can only use mathematical operators.

Dummy variable columns were then created for all values where we found patterns in our data exploration step to see if they would have an impact on our regression model for sales. To be specific we created dummy variables for product categories as that was the only distinction we found through data exploration and all other attributes had similar average sales with, the exception of the product category (specifically home C Furniture, and Fashion) [figure 3 in appendix]. Additionally, we found how they would interact with our sales model.

There are some variables we explicitly chose not to include in our model such as ship mode, profit, and order priority due to the phenomenon of correlation vs causation. To expand, shipping is likely to increase because of the type of the product and might not predict the Sales. In addition, order priority is likely to be dependent on the value of the order.

Technical Description: In the data exploration phase, we determined that the only variance amongst average sales was with the product categories however because they were objects, we could not include them in our regression. We proceeded to create dummy variables for these categories specifically for Fashion, Home Furniture, and Electronics. Using the df_with_dummies function in Python we prepared our model to read these dummy variables.

Predictive Modelling

Management Description:

Alpha definition: a cutoff point used in hypothesis testing to help decide if the results of an experiment are meaningful. If the results have less than a 5% chance (alpha of 0.05) of happening by random luck, we say they are significant and likely real. It's like setting a rule that says, "Only believe this result if it's very unlikely to be just a fluke or if the test result is below 5%"

Our final model has 96% accuracy in predicting sales using the independent variables: discount, low-sale products, high-sale products, and shipping cost. We selected these independent variables to be in our model by performing tests on them to see if they were below our alpha of 0.05 which they all passed. We chose a small alpha of 0.05 as sales is the main contributor to any e-commerce business as without it the business would fail so we wanted a low probability of choosing independent variables that would not help us in predicting sales.

Technical Description: Including all the existing columns as well as the dummy variables we created into our model we went through numerous tests to determine what was relevant to our model. We first performed individual hypothesis test on the p-value of each individual independent variable to determine significance. If any independent variable's p-value was greater than 0.05 by a significant amount they would be removed from the model as we would embrace the null that they are not significant and if the p-

values were less than 0.05 we would reject the null and embrace the alternative that these independent variables have an impact on the dependent variable which is sales.

The only dependent variables that passed our hypothesis test were discount, shipping cost, fashion category, and home furniture category. From there we checked our r-squared which was high enough to ensure that our model can predict sales with a high probability of 86% based on our independent variables and that our model passes the F-test which is done as our alpha is set at 0.05 and our F-statistic is 8.198e+04. The next step was to train our model and ensure there was no overfitting for which we trained our model at 75% from which we got an r-squared of 86.5% and 86.3% to confirm that there was no overfitting with our model.

The final step was to run tests on our regression model from which we found there were issues with heteroskedasticity and outliers. Heteroskedasticity was first discovered by doing a visual plot (figure 6) where we saw patterns that could point to heteroskedasticity, but we then confirmed doing the Breuch-Pagan Test where our p-value was less than 0.05 confirming that the model had heteroskedasticity. The second issue with outliers was discovered using the influence plot (figure 7).

To address the issue with heteroskedasticity we did further analysis on our data set. We discovered there was high variance of sales for different products, so we created two clusters to address this issue. The first cluster is for high sale product and the second cluster for low-sale products. We then created dummy variables for these clusters so they could be included in our model. To address the issue with outliers we realized that the data that consisted of outliers was less than 5% and determined not to be important to the dataset so it was deleted from the dataset. When re-performing the tests we confirmed that the issues with heteroskedasticity and outliers were fixed (figure 10) so we could move forward with our final tests using a density plot to determine if our data was normal (figure 8) and a Q-Q plot (figure 9) to confirm that our model had no issues which both passed.

With the changes, our final model was built using discounts, low-sale products, high-sale products, and shipping costs as our main predictors of sales. Our R-squared increased to 96% and our F-statistic stayed very low at 3.807e+05 passing our hypothesis test that these independent variables predict our dependent variable. The Fashion category and Home Furniture category were removed from the model because the additional complexity shown by the adjusted R-squared in the model was not justified by the value these variables were providing. Finally, we did train and test sets on our data set to confirm that there was no overfitting within our model, and since our R-squared were almost identical at 0.9685 and 0.9697 we concluded that there was no overfitting.

Hypothesis Testing

Management Description: A hypothesis test is a test that is performed to see if certain assumptions hold against data. We first start by stating the null hypothesis which is that our assumption is not true, we then set an alpha which is to what probability we are willing to accept our assumption. In our model we set a couple assumptions and performed hypothesis tests on them to determine if they were true or not. Here

we are trying to determine whether the predictor variables are statistically significant in explaining the variability in the dependent variable.

Our assumptions were that regions, months and segments would have an impact of sales. To test this, we set the null hypothesis where high sales, low sales, regions, months and segment were set equal to zero or that they would not have any significance in predict sales.

When performing the hypothesis test, we determined that the dummy variables that were created for high sales and low sales were both statistically significant and belonged to our model as the results were below our set alpha of 0.05. We then performed individual hypothesis tests on dummy variables for regions, months, and segments to see if any passed our hypothesis test. Unfortunately, in every case, the test failed meaning that we embraced the null and rejected the alternative that these values were in any way significant to predicting sales.

Technical Description: Hypothesis tests were performed on our dummy variables for high sales and low sales to determine if they had any significance where our null hypothesis was that low_sale_products=0 and high_sale_products=0. We set out alpha for 0.05 as we wanted to sure that our test would be accurate but not leave slight room for error. Our p-value when performing the wald test on our hypothesis was 4.89e-04 meaning that we reject the null and embrace that these values have at least some significance in predicting and hence belong in our model.

Further hypothesis tests were conducted on individual months, individual segments, and individual regions to see if any of them have any impact on predicting sales. For individual months, the p-value was 0.1983 which is significantly above our alpha of 0.05 so we embrace the null that months do not have an impact on sales. For individual regions, the p-value was 0.1525 which is significantly above our alpha of 0.05 so we embrace the null that months do not have an impact on sales. Finally, for individual segments, the p-value was 0.2147 which is significantly above our alpha of 0.05 so we embrace the null that months do not have an impact on sales.

Executive Dashboard Development

Management Description:

Purpose: The dashboard helps transform complex data into an accessible and visual format. It enables the management team to quickly gain business insights and provide suggestions. By interacting with the dashboard, the management team could discover the correlation among the features.

Description: Our dashboard is divided into two parts: one part shows the sales distribution of Chown Market Place, while the other part presents the customer profile of Chown Market Place. The dashboard could be filtered by date, discount rate, product feature (high sale vs low sale product), and customer value groups (high, medium, low) to view the sales distribution and customer distribution. We display data from five perspectives on the sales dashboard: Monthly Sales Trend, Product Category, Sales Product, Sales Region, and Consumer type; and five perspectives on the customer profile dashboard: Total Customer by Product Category, Total Customer by Product, Average Shipping Cost by Product Category, Geography of Customers and Total Customers by Segment.

Observations:

1) Chown Market achieved the highest sales in December and the lowest sales in February. It can be interpreted because of the seasonal response since people tend to buy more on Christmas and Boxing Day and stop buying after New Year
2) Profit margin and average discount are highly correlated to Sales since they have the same trend
3) In terms of Product Categories, Fashion has the highest sales with a total of $5.1 million, among which T-shirts have the highest sales at $681,000
4) Most of our customers placed their orders from the Central Market Place. However, the Canadian market has the lowest sales the business could focus on marketing and advertising to increase sales in Canada in the future
5) Team Chown Market has a 50,500-customer base, most of them are individual consumers and they are in North America
6) Asia has huge population potential but performed sub-ideal in the timeframe under consideration. In the future, the management team should pay attention to the Asian market, especially China, India, and Indonesia by increasing marketing efforts, expanding the customer base, or providing some culture-relative products

Technical Description

Data Preparation: Before building the dashboard, we used Python to add three new columns: Customer Value, Low Sales Products, and High Sales Products. The high-value customer showcases customer spending in the top 25 percentile of the total, the low-value customer is in the last 25 percentile of the total and the spending amount in between is placed as medium value customer. Similarly, the low-sales products are those with average sales below $100 for all products whereas high-sales products are those with average sales above $200. These three columns will divide customer groups and products more specifically and help with the data insights.

Future Use: For the sustainable use of the dashboard, we need to ensure the consistency of the column names. By adding the new data to the Excel and refreshing the dashboard, the most up-to-date data will be shown.

Recommendation

Recommendation 1: Future Driven Strategy

Understanding your customer base is crucial for driving sales in e-commerce. **Collecting precise customer data allows for building comprehensive profiles and better product targeting, increasing sales.**

Our initial sales model showed that the existing customer data (location, customer ID, segment) is very limited and lacks predictive power. To improve accuracy and create detailed profiles, we recommend collecting more data, such as gender, occupation, browsing history, and engagement metrics like time spent on the site and device type. Richer data enables more accurate models and personalized marketing, boosting conversion rates. It helps identify profitable customer segments as well as those not contributing to sales, revealing underlying issues, and informing engagement strategies.

Analyzing this data also spots market trends and shifts in preferences, allowing timely adjustments in marketing and product strategies. Implementing these recommendations will enhance engagement and sales, providing a competitive edge in e-commerce.

Recommendation 2: Strategic Business Recommendation

Promotional initiatives such as customer discounts and loyalty programs in high growth segment of the base will provide a tactical advantage to Chown Market Place, especially in markets that comprise of price-sensitive consumers. As part of our strategic recommendation, we advise Chown Market Place to exercise promotional efforts in the form of an average annual price discount of 5% and loyalty programs targeted at Middle and Low-Value Customers residing in India, China, and Indonesia. To optimize the full capacity of this recommendation, Chown Market Place can boost its sales by exercising high discount rates in the first 6 months of the year.

Overall, as our model highlighted discounts to have a strong positive impact on sales and Chown Market Place requiring more customer-centric data, we believe this recommendation will address both requirements effectively and efficiently.

Recommendation 3: Leverage our Power BI Dashboard to conduct performance evaluation

An effective organization requires an established system to conduct pre-and post-benchmark analysis for its projects. **Given the scope of the problem we are trying to solve in this project, we recommend Chown Market Place capitalize on the Power BI dashboard that our team has put together**. The company must pay special focus on comparing sales performance across key segments using the "Sales by Quarter and Segment" section, which will help identify any shifts after implementing new strategies. The "Sum of Profit, Sum of Sales, and Average of Discount by Month" chart is essential for assessing monthly trends and pinpointing periods where profit margins indicate the impact of these initiatives. Additionally, the "Sales by Region" visualization can be used to evaluate geographical performance, highlighting regions that exceeded or missed benchmarks. By examining the "Sales by Product Category" section, you can identify which product categories contributed most to changes in overall sales. Finally, reviewing customer metrics from the “Chown Market Customer Profile Dashboard" will offer insights into how customer behavior and demographics have shifted, revealing the effectiveness of customer-focused strategies.

Appendix

Figure 1: Final model

OLS Regression Results

Dep. Variable:	Sales	R-squared:	0.969
Model:	OLS	Adj. R-squared:	0.969
Method:	Least Squares	F-statistic:	3.807e+05
Date:	Sun, 04 Aug 2024	Prob (F-statistic):	0.00
Time:	16:53:54	Log-Likelihood:	-1.8787e+05
No. Observations:	48943	AIC:	3.758e+05
Df Residuals:	48938	BIC:	3.758e+05
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	87.5569	0.195	448.079	0.000	87.174	87.940
Discount	371.1164	3.742	99.169	0.000	363.782	378.451
low_sale_products	-45.1477	0.155	-290.644	0.000	-45.452	-44.843
high_sale_products	18.4646	0.221	83.635	0.000	18.032	18.897
Shipping_Cost	8.6534	0.024	361.650	0.000	8.607	8.700

Omnibus:	4341.558	Durbin-Watson:	2.033
Prob(Omnibus):	0.000	Jarque-Bera (JB):	7013.610
Skew:	0.664	Prob(JB):	0.00
Kurtosis:	4.295	Cond. No.	660.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

R-squared increased and the dummy variables low_sale_products and high_sale_products are clearly significant. Removing 'Product_Category_Fashion', 'Product_Category_Home_Furniture' because the additional complexity in the model is not justified by the value these variables are providing.

Figure 2: Average sales by region

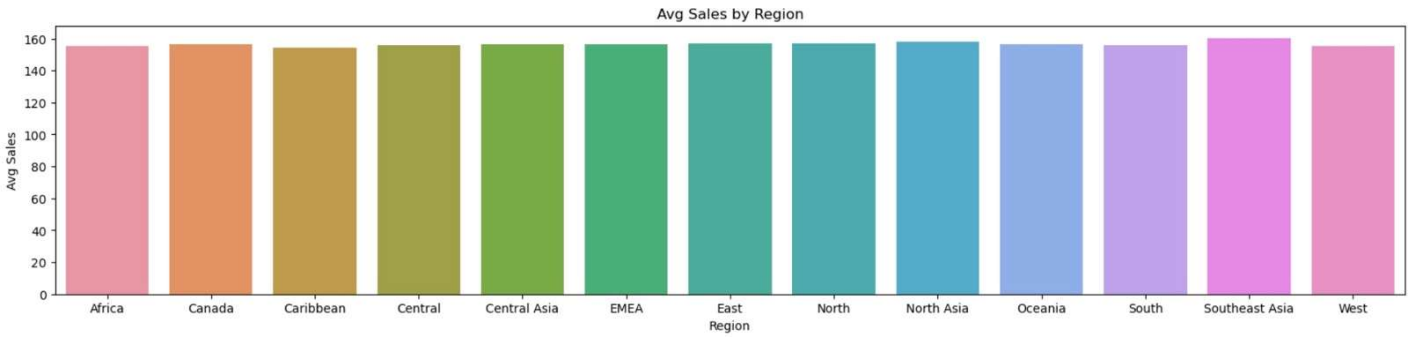

Figure 3: Average sales by product category

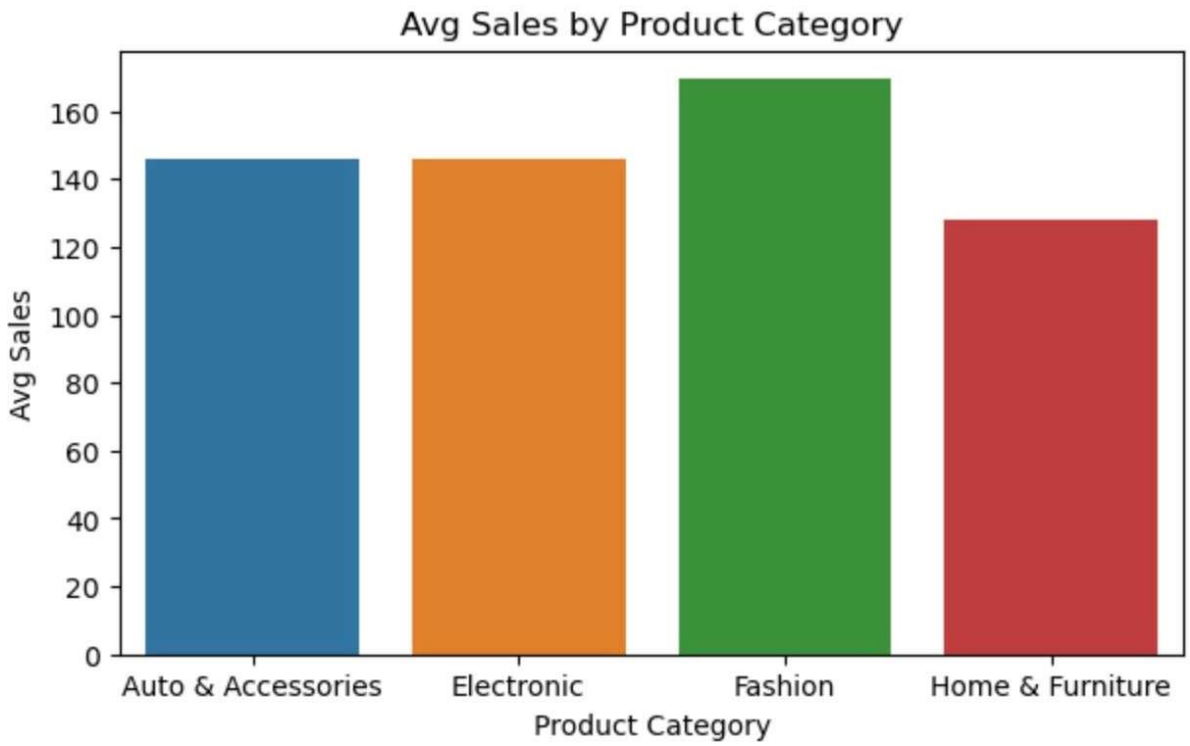

Figure 4: Average sales by Segment

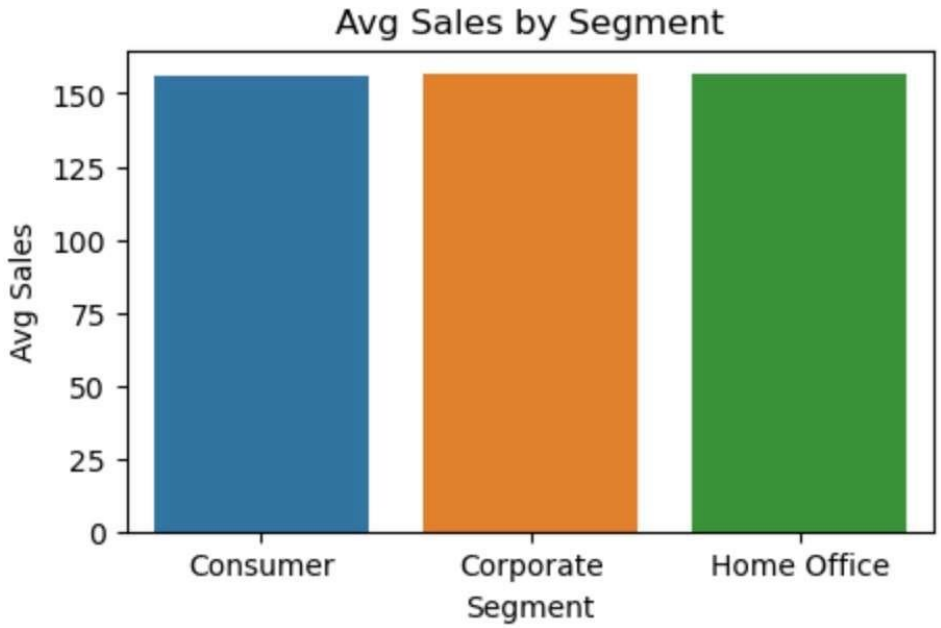

Figure 5: Average sales by month

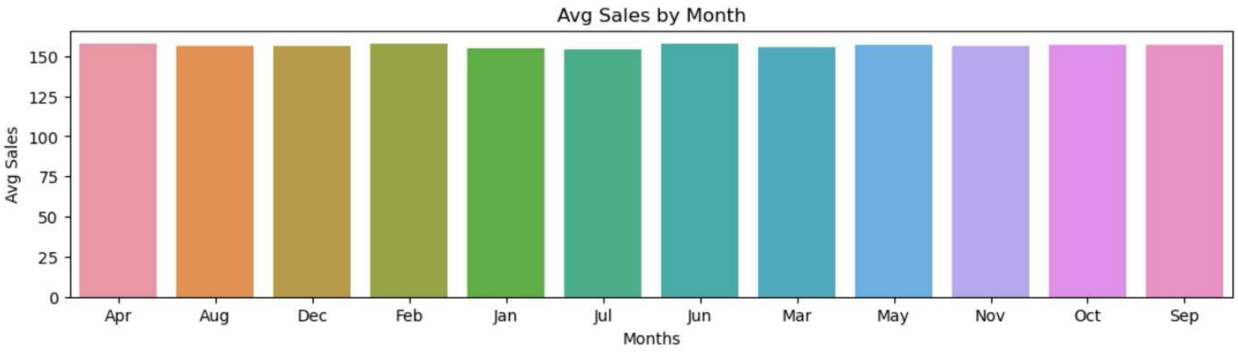

Figure 6: Heteroskedasticity

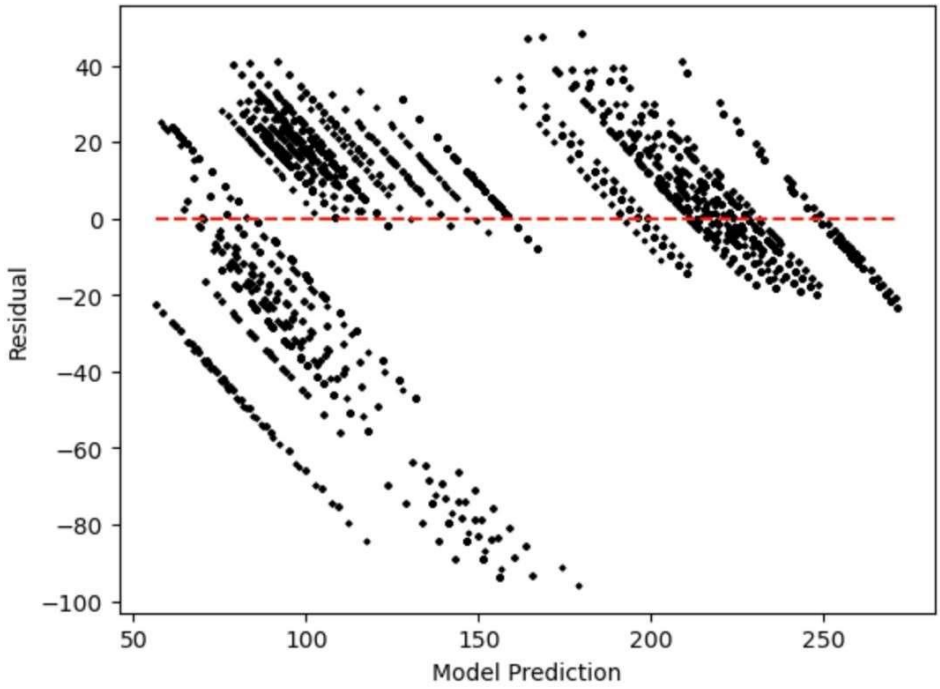

Figure 7: Cooks plot for outliers

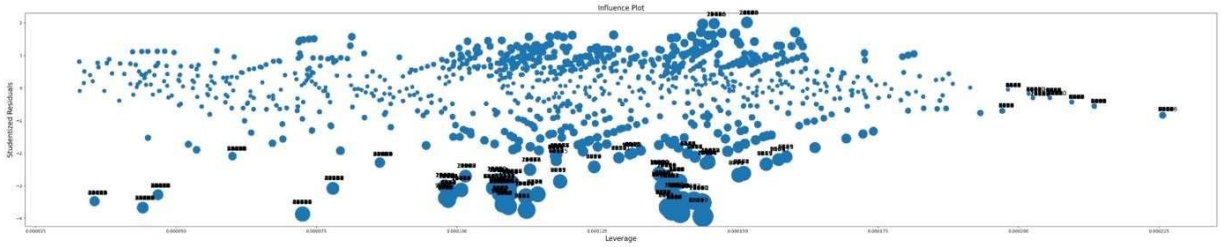

Figure 8: Density plot

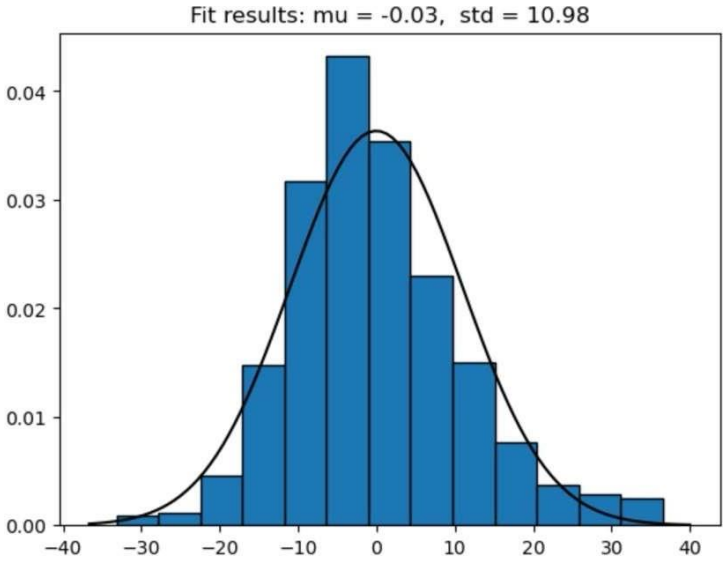

Figure 9: Q-Q plot

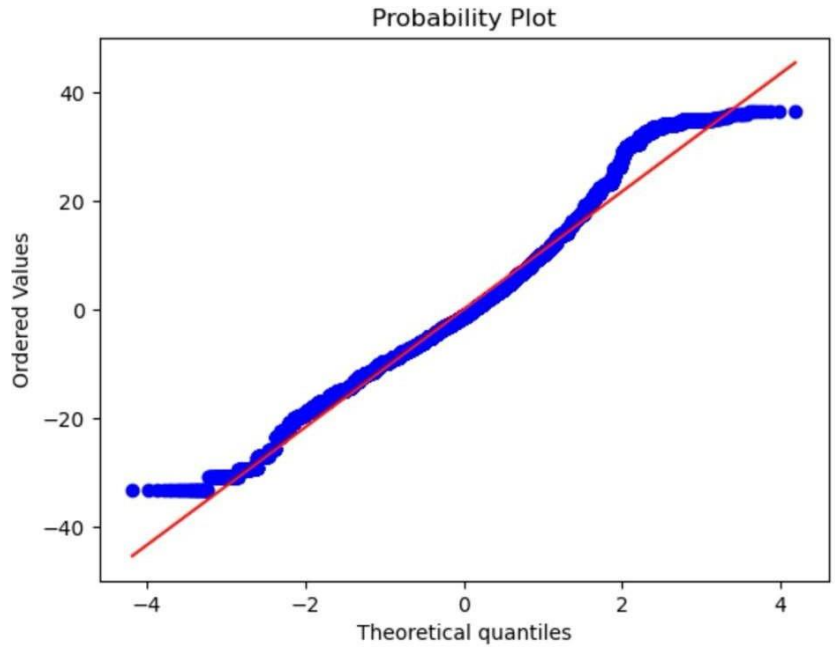

Figure 10: Heteroskedasticity fixed

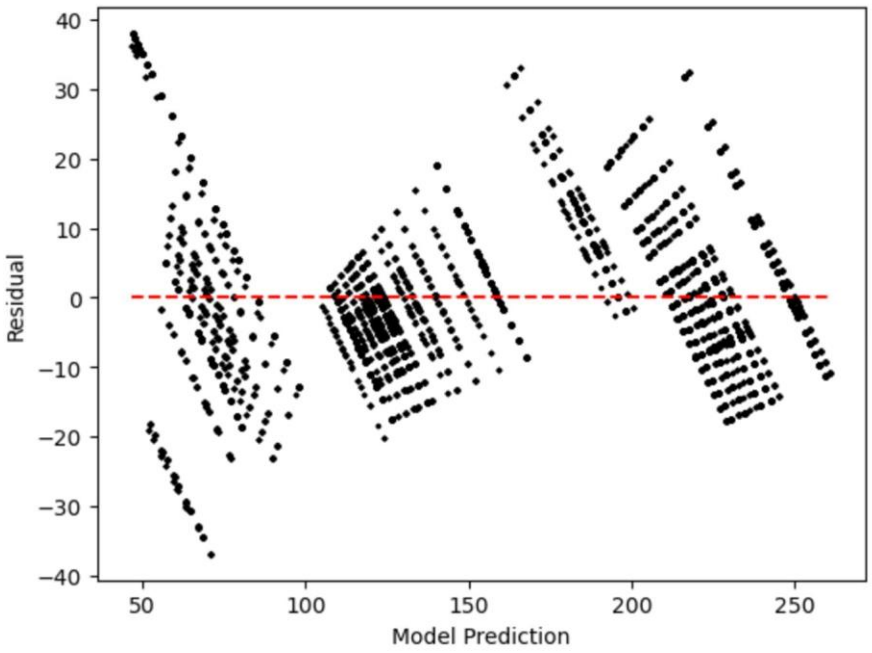

Figure 11: Monthly sales growth rate