**Table of Contents**

[Executive Summary 3](#_Toc11310)

[Introduction 4](#_Toc11311)

[Method 4](#_Toc11312)

[Data Analysis Result 5](#_Toc11313)

[Total Sales through each month of the year 5](#_Toc11314)

[Sales per year 6](#_Toc11315)

[Sales per state 7](#_Toc11316)

[Profit per state 7](#_Toc11317)

[Sales by segment 8](#_Toc11318)

[Sales per category 9](#_Toc11319)

[Shipping mode distribution 10](#_Toc11320)

[Top 5 customers 11](#_Toc11321)

[Linear regression of sales vs Profit in California 12](#_Toc11322)

[Predicted Sales for the next 12 months 14](#_Toc11323)

# Executive Summary

This report provides a comprehensive analysis of our historical sales data to support strategic decision-making and enhance business performance. We employed various data analytics techniques to uncover valuable insights into sales trends, key drivers, and future performance. Some key findings include:

**Seasonality**: Sales exhibiting seasonal patterns, with peaks in March and October. Targeted marketing campaigns during these periods could boost sales

**Yearly Trends:** Sales peaked in 2015, followed by a significant drop in 2016. Understanding the reasons behind these variations can inform future strategies

**Geographic Distribution:** California, New York, and Pennsylvania have the highest sales volume, with New York and California showing the highest profits. Texas, Illinois, Ohio and Pennsylvania show the lowest profit. State specific marketing campaigns tailored to regional needs may improve profitability.

**Customer Segmentation:** The home office segment and the technology category each respectively generate the highest average sales. Optimising inventory levels and product offerings can prevent stockouts and capitalise on customer needs

**Shipping Preferences:** Standard Class shipping is the most popular option. Evaluating alternative delivery options could enhance the customer experience.

**Top Customers:** Identified the top 5 customers, allowing utilisation of personalised communication strategies to promote customer satisfaction, loyalty and potentially higher sales.

**Sales-Profit Relationship (California):** A weak positive correlation suggests that higher sales generally lead to higher profits.

**Sales Forecasting (Next 12 months):** A weak positive correlation indicates a general upward trend in sales across all categories for the next year, with a slightly steeper increase for furniture sales.

This data analysis provides valuable insights to inform strategic decision-making across various aspects of the business. By leveraging these findings, we can optimise inventory management target marketing efforts effectively, and anticipate future demand to improve overall sales performance and profitability.

# Introduction

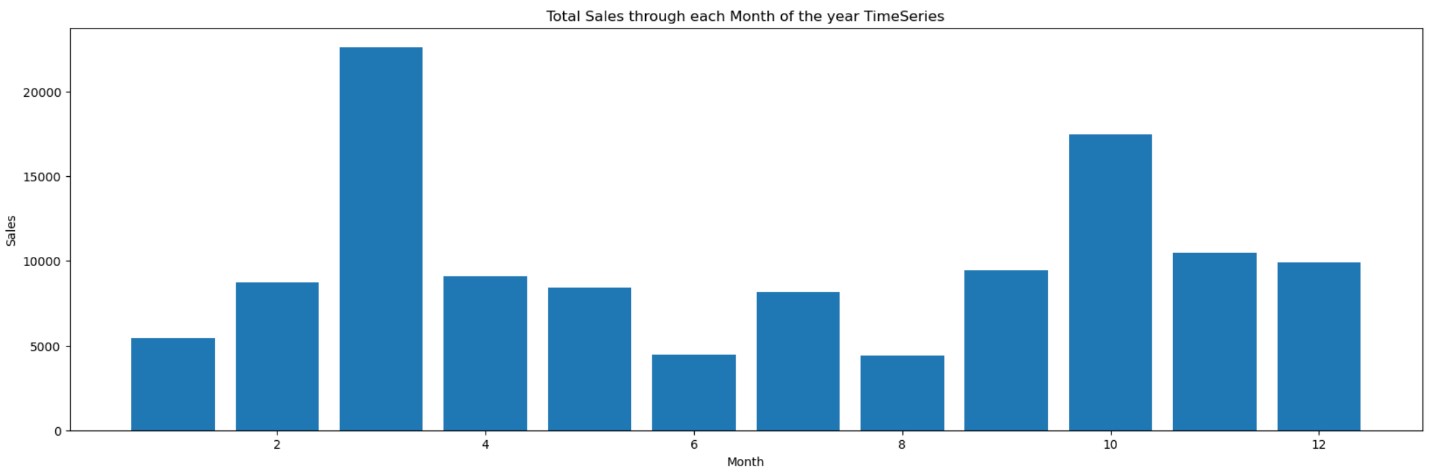
This report provides a comprehensive analysis of our sales data to support strategic decisionmaking and enhance business performance. The dataset includes information of sales, categories, months, states, shipping methods, and other relevant variables. The data types were primarily numerical (sales figures, month numbers, etc.), and categorical (category names, state names, shipping methods). We utilized various analytical techniques to understand historical sales trends, identify key drivers of sales, and predict future performance. The analysis includes statistical summaries, exploratory visualizations, and regression models to determine sale to profitability in the state of California, and to forecast sales for each item category for the next 12 months. Key insights include sales distribution across states, profitability by region, annual growth patterns, and the impact of shipping modes on sales. The goals of this analysis are to understand historical trends and patterns, to develop models to forecast sales and profit, and to identify key drivers of sales. These goals are well-suited for the provided dataset, as the data includes historical sales figures, allowing us to analyse trends and build models for forecasting, and the presence of information on categories, months, states and shipping methods enables us to explore how these factors influence sales. This structured approach enables us to plan inventory and production levels, develop targeted market campaigns based on identified key drivers, and to make data-driven decisions. By setting clear goals and leveraging the strengths of the data, this analysis provides valuable insight and allows the business to anticipate future demand effectively.

# Method

This project employed a comprehensive suite of data analytical methods to extract valuable insights from our sales data. These methods include: statistical analysis to compute key metrics such as averages and means, and exploratory data analysis involving bar charts, pie charts, and time series plots in order to visualise sales patterns across different dimensions such as state, year, and shipping mode. Additionally, based on the initial exploration and data characteristics, we performed regression analyses to forecast profit for the number of sales in the state of California, as well as to forecast sales for the next 12 months. The models were trained on historical sales data to identify relationships between sales and potentially relevant factors such as seasonality and past sales trends. The analysis was conducted using primary libraries such as pandas, numpy, and matplotlib, to create clear and informative visualisations of the data and analysis results. We also delved deeper, incorporating sklearn for modelling and plotly for interactive visualizations to further enhance visual communication and exploration of data. The chosen methods were strategically selected to align with the goals of the analysis and the characteristics of the sales data. The statistical analysis provided a solid foundation, the exploratory data analysis helped us to visually explore the data and identify patterns, and the regression analysis with appropriate visualisations allowed us to forecast future sales and effectively communicate the results.

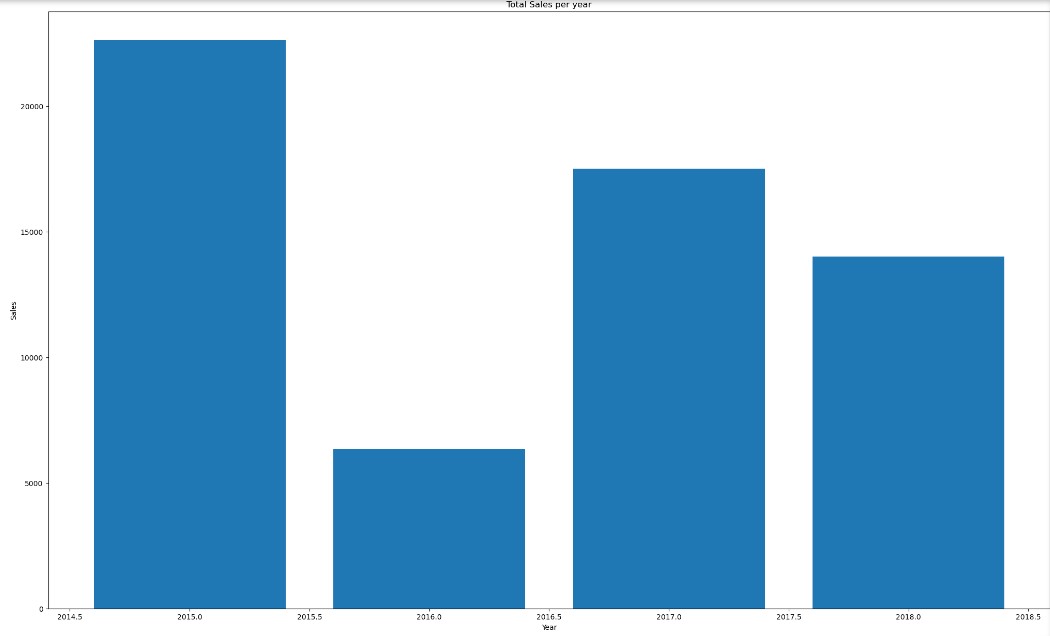
# Data Analysis Results

# Total sales through each month of the year



Using a timeseries bar chart, total sales was grouped by the month in which it was completed. This was done in order to see if there was any seasonal preference for sales, so that the company is able to make adjustments to their sales or marketing techniques in the future. The datapoints used were sales and month. The timeseries bar chart above reveals that there is a large spike in sales in month 3 (March), as well as a slightly smaller spike in month 10 (October). The company can leverage this information to improve sales by focussing on marketing around March and October, allocating additional budget to fund advertising campaigns. In anticipation for these peak months, the company can also adjust their production or inventory management strategies in order to meet the surge in sales and avoid being out of stock, which could lead to loss of sales and customer dissatisfaction.

## Sales per year



Using a bar chart we were able to distinguish the years with the most sales. The datapoints used were sales and year. The chart clearly shows that 2015 was the highest year for sales, surpassing 20,000 sales followed by a much drastically lower 2016 of around 6,000 sales. 2017 and 2018, the sales evened out to represent the average sales between the 4 years quite accurately. Using this information, the company can analyse the factors behind the surge in sales in 2015, and try to understand why that was the case, in order to replicate that success. In the same way, the company can use this information to try and decipher why there was such a large drop of sales in 2016, which could help the company implement corrective measures. The sales in 2017 and 2018 appear to be more stable, so the company can use these sales as a more accurate benchmark to represent average sales performance. It can be used to evaluate the effectiveness of future sales and marketing initiatives.

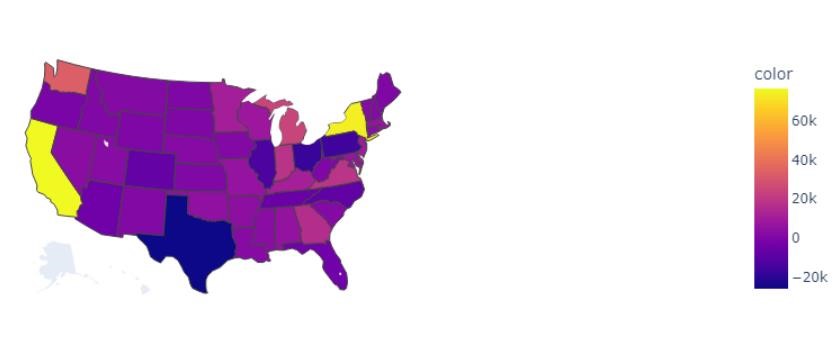
## Sales per State

A map of the united states

Description automatically generated

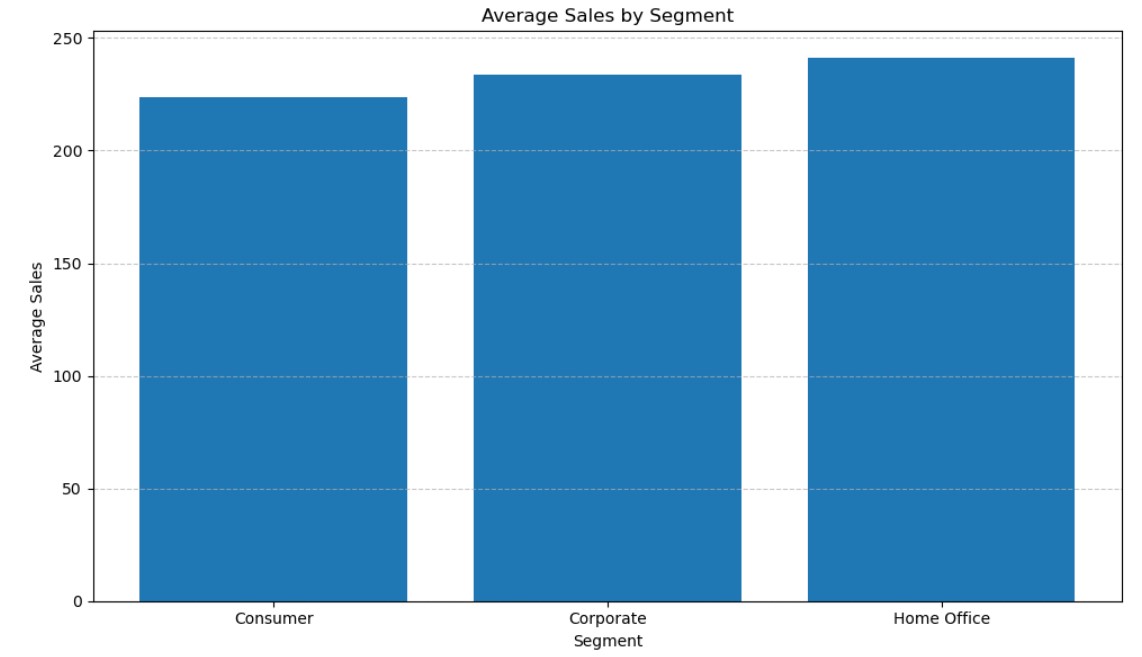
Using a choropleth map/heat map, we were able to see the total number of sales per state. The data points used were sales and state. The map represents the highest sales in dark green, and the lowest sales in red, with the colour range represented in the legend. The states with the most sales appears to be California followed by New York then Pennsylvania and Seattle represented by the green shading. What is noticeable is that most states had below 100k sales particularly in Northern and Central America. The company can use this insight to improve their sales strategy. They can focus on high volume states, replicating their successful strategies, and investing in customer retention. They could also look to expand into the low sale-volume areas around the rest of the US, identifying the reasons for low sales, and tailoring their marketing campaigns towards each respective region.

## Profit per State



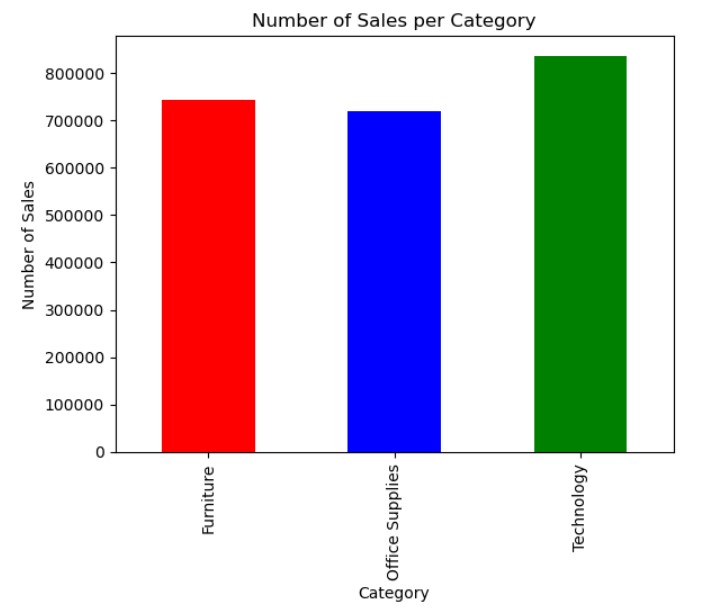
Using a choropleth map/heat map, we were able to see the profit per state. The data points used were profit and state. The map represents the highest profit in yellow, and the lowest profit/highest lost in dark blue, with the colour range represented in the legend. The states with the highest profit appear to be New York and California in yellow. This is followed by a distinguishable orange in the top left-hand corner in Washington State. The graph has also identified the states with the lowest profit: Texas, Illinois, Ohio and Pennsylvania. The company can use this valuable geographical insight in order to improve their sales strategy. They could increase their marketing efforts in the states that show the highest profits, and also analyse or try to identify the reasons why some states have the lowest profits/greatest losses. The company could develop state-specific marketing campaigns in order to cater to the specific needs or preferences of customers in those regions.

## Sales by Segment



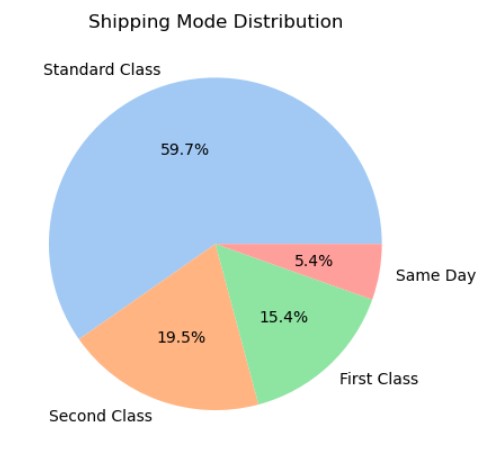
Using a bar chart, we were able to recognise the average number of sales for each segment. The data points used were average sale amount, and customer segment. The bar chart shows that the home office segment generated the largest average sales, followed by the corporate segment, and finally the consumer segment. This offers valuable insight for the company into their customer base and respective buying habits. The home office segment generated the largest average sales, indicating that customers tend to purchase more in one go, and the corporate segment represents a lower average sale – this may indicate that businesses are making more frequent, smaller purchases. By using this information, the company can tailor their product offerings and marketing. For example, they can consider offering product bundles that would fit into the home office segment, increasing potential for cross-selling and increasing the number of sales.

## Sales per Category



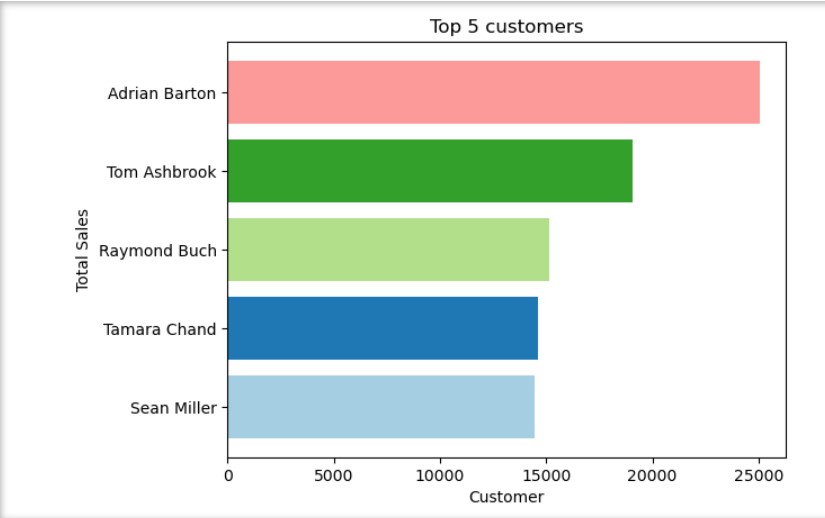
Using a bar chart, we were able to recognise the total number of sales for each item category. The data points used were sales and item category. The chart shows that the company has as fairly even distribution on product category sold, with a slight emphasis on technology, followed by furniture, and then office supplies. The company can use this information to guide their future direction. As there is a larger proportion of technology sales, the company may want to shift their focus to selling more technology products. The company can also use this information to ensure optimal inventory levels for the high-sales categories to avoid being out of stock and to capitalise on their customer demands.

## Shipping Mode Distribution



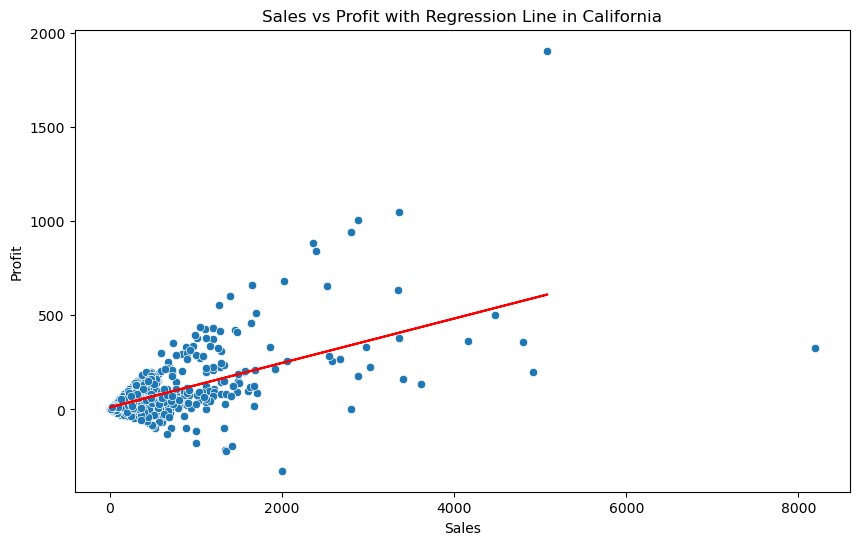
Using a pie chart we represented the shipping mode distribution. From this chart, we can see that Standard Class shipping is the most popular choice by customers at 59.7%, followed by Second Class (19.5%), and First Class (15.4%). Same Day shipping is the least popular, contributing only 5.4% of the shipping mode distribution. This pie chart provides valuable insight into customer preferences and delivery needs. The company can use this information to cater to the customers’ needs. The popularity of Standard Class indicates that a majority of customers prioritise affordability over speed. It may indicate that the company should evaluate their other classes of delivery options to see if anything can be done to improve the overall customer experience.

## Top 5 Customers



Using a horizontal bar chart, the top 5 customers with the highest total sales has been shown. The data points used are total sales, and the top 5 customers by sale. From the chart we can see that Adrian Barton is the top customer, followed by Tom Ashbrook, Raymond Buch, Tamara Chand and finally Sean Miller. This information provides valuable customer relation management insights. The company, using this information, has identified the most highvalued customers. Understanding their buying habits and preferences can inform strategies to retain their business and potentially encourage them to spend more. Personalised communication strategies for these top customers could prove useful, as it could lead to an increase in their satisfaction, loyalty, and potentially even higher sales from the top spenders.

## Linear Regression for Sales vs Profit in California



The objective of this graph is to show the relationship between Sales and Profit for the state of California. We used a linear regression to model this relationship and visualise the results. A linear regression is a suitable model for this scenario as it aims to find a linear relationship between two variables. In this case, we are looking for a straight line that best fits the data points representing sales (independent variable) and profit (dependent variable) for the state of California.

**Data preparation:**

* We filtered the dataset to the state of 'California'.
* We selected 'Sales' as the independent variable (X) and 'Profit' as the dependent variable (y)
* We split the data into training (80%) and testing (20%) sets in order to evaluate the model's performance on unseen data.
* A Linear Regression model was then created.
* Trained the model using the training data (X\_train and y\_train).
* Did a prediction for the Profit value for the test data (X\_test) using the trained model - Created a scatter plot of the actual sales vs. profit data.
* Plotted the regression line to visualise the linear relationship between sales and profit.

**The Outcome:**

The model identifies a weak positive correlation between sales and profit. This means there is generally an upward trend, where higher sales tend to lead to higher profits. The model can be used to understand the general trend of how sales affect profit in the state of California. This can be helpful for planning and setting sales targets that consider their impact on profitability. As California was identified in earlier analyses to have one of the highest sales and profit numbers, this information would be very valuable to the company. As there is only a weak correlation, it suggests that there may be other important factors affecting profit. This may prompt the company for further data analysis to explore these additional factors in order to replicate them in other states.

Overall, the linear regression model provided us with a basic understanding of the relationship between sales and profit in California. However, due to the weak correlation, its use for precise prediction is limited and we should also consider other factors that may be affecting sales and profit.

## Predicted Sales for the Next 12 Months

A graph of a graph

Description automatically generated with medium confidence

The objective of this graph is to show a prediction of sales for the next 12 months for each unique item category. We used a linear regression to model this relationship and visualise the results. A linear regression is a suitable model for this scenario as it aims to find a linear relationship between two variables. In this case, we are looking for a straight line that best fits the data points representing future month (independent variable) and sales (dependent variable) for each of the item categories.

**Data preparation:**

* We filtered the data set by category, month and total sales.
* We selected 'month' as the independent variable (X) and 'Sales' as the dependent variable (y)
* We prepared and trained the model for future predictions - A Linear Regression model was then created.
* We then did predictions for the sales value using the trained model for each of categories
* We then plotted the regression lines to visualise the linear relationship between the next 12 months and future sales **The Outcome:**

The model identifies a weak positive correlation for sales in each category for the next 12 months. This means there is generally an upward trend, where there will be higher sales in the future. The model can be used to understand the general trend of how seasonality affects the number of sales for each of the categories. This can be helpful for planning and setting sales targets for each category. It can be seen that Technology and Office supplies have a similar slope and will increase at a similar rate, however Furniture seems to have a greater slope. This indicates that in the coming future, furniture may provide a greater number of sales. As there is only a weak correlation, it suggests that there may be other important factors affecting number of sales. This should prompt the company for further data analysis to explore these additional factors in order to provide a more accurate picture of predicted sales.

Overall, the linear regression model provided us with a basic understanding of the relationship between category sales and the next 12 months. However, due to the weak correlation, its use for precise prediction is limited and we should also consider other factors that may be affecting sales numbers.