Analyzing Retail Sustainability: AdventureWorks Inc. Through Time Series and Market Basket Analysis

Mark R. Hanson

Department of Data Science, University of Wisconsin – Eau Claire

DS785: Capstone

Dr. Tracy Bibelnieks

April 28, 2024

**Abstract**

AdventureWorks Inc., a bike retail company, operates in a competitive market and faces the challenges of sustaining its niche position amidst evolving consumer preferences and the saturation of online retailers like Amazon. This project aims to explore strategies for ensuring the company’s sustainability by analyzing customer segments, sales, performance, profitability, and market trends. The introduction provides an overview of AdventureWorks’ business landscape, highlighting its annual revenue over three years, and the competitive market dominated by both niche retailers and very large retailers like Walmart and Target. Through a review of top retailers and bike brands, along with an analysis of customer segments, including Champions, Loyal Customers, and At-Risk Customers, potential growth avenues are identified. The project’s objectives encompass identifying key sales trends, analyzing profitability, conducting time series analysis for sales prediction, segmenting customers based on purchasing behavior, and performing Market Basket Analysis to uncover accessory purchase patterns. These objectives will be pursued using Python-based methods such as Exploratory Data Analysis, ARIMA modeling, and Market Basket Analysis. Limitations, including the availability of three years of data and the need for data cleaning, are acknowledged. The conclusion emphasizes ongoing data analysis and the importance of uncovering essential accessories and crafting effective marketing strategies to support AdventureWorks' business model. Through this project, AdventureWorks aims to leverage data-driven insights to enhance its market position, optimize its product offerings, and adapt to the dynamic retail landscape effectively.

Keywords: Data, Analysis, Customer Segmentation, Market Basket, Revenue, Python, Trend Analysis, Power BI.

**Table of Contents**

Abstract …………………………………………………………………………………………...2

Table of Contents …………………………………………………………………………………3

List of Tables .…………………………………………………………………………………….5

List of Figures……………………………………………………………………………………..6

Chapter 1: Introduction …………………………………………………………………………...7

Background ……………………………………………………………………….………7

Call to Action ……………………………………………………………………..……..10

Customer Segment Considerations………………………………………………………10

Overview of Methods ……………………………………………………………….…..10

Project Objectives………………………………………………………………………..11

Assumptions and Limitations……………………………………………………………12

Conclusion……………………………………………………………………………….12

Chapter 2: Literature Review ……………………………………………………………………13

Introduction……………………………………………………………………………….13

` Techniques of Data Mining…………………………………………………………….....14

Market Basket Analysis in a Retail Environment………………………………………...14

Conclusion………………………………………………………………………………..20

Chapter 3: Methodology ………………………………………………………………………...21

Introduction ...........................................………………………………………………..21

Data Collection………………………………………………………………………….22

Data Preprocessing…….………………………………………………………………. 22

Statistical and Machine Learning Models……………………………………………….26

Time Series……………………………...…………………………………..…………...27

Market Basket……………………………………………………………………………28

Customer Segmentation………………………………………………………………….30

Choose the Models ………………………………………………………………………31

Conclusion ………………………………………………………………………………31

Chapter 4: Results of Methodology………………………………………………………...........32

Introduction………………………………………………………………………………32

Trend Analysis…………………………………………………………………………...33

Profitability Analysis …………………………………………………………………....34

Time Series Forecasting………………………………………………………………….36

Market Basket Analysis………………………………………………………………….40

Customer Segmentation………………………………………………………………….43

Conclusion...……………………………………………………………………………..46

Chapter 5: Conclusion …………………………………………………………………………..47

Introduction ……………………………………………………………………………...47

Summarizing……………………………………………………………………………..47

References ……………………………………………………………………………………….49

Appendix A……………………………………………………………………………………....54

Appendix B………………………………………………………………………………………55

**List of Tables**

Table 1: AIC and BIC results based on different selection of parameters……………………….37

Table 2: Predictive forecasted profits for AdventureWorks……………………………………..39

Table 3: Comparing results of the Antecedent, Consequent, Support, Confidence, Lift………..42

Table 4: Occupation is the selection, comparing a few combinations purchases………………..56

**List of Figures**

Figure 1: Top ten models for road bikes, mountain bikes, and touring bikes……………………..8

Figure 2: Rules for support, confidence, lift……………………………………………………..18

Figure 3: Extract to show the rules in a colorful diagram………………………………………..18

Figure 4: Total sales for road bikes per age group……………………………………………….24

Figure 5: Bike & accessory distribution…………………………………………………………25

Figure 6: Bike sales by quarter…………………………………………………………………..26

Figure 7: Monthly sales trend……………………………………………………………………33

Figure 8: Revenue vs Profit for each color bike…………………………………………………34

Figure 9: Road bike sales per age group…………………………………………………………35

Figure 10: Running AutoRegressive Integrated Moving Average………...…………………….36

Figure 11: Graphed results of the ARIMA model using 4, 4, 4 as the parameters………………38

Figure 12: Line chart showing the profit forecast………………………………………………..40

Figure 13: Market Basket Analysis using the Apriori Algorithm………………………………..41

Figure 14: Sum of profit by bike color…………………………………………………………..44

Figure 15: Cluster chart by marital status………………………………………………………..45

Figure 16: Cluster scatter plot by Age vs Total Children………………………………………..55

Chapter 1: Introduction

**Background**

AdventureWorks Inc. is a bike retailer, selling bikes and various accessories to individuals and families. The competition in the retail industry is typically strong, especially in the bike industry. There are so many niche businesses that set their business models based on the regions they reside. In Wisconsin, there are many trails, and many state parks, so road bikes are a specific consideration that retailers like Trek, Sun Prairie Cycle, and many others will focus on. There is a consideration of families, so retailers must also compete with the mega-retailers that sell bikes as a side business – places like Walmart and Target stand out in this analysis.

Top Ten Retailers Based on Data Through 2022 per Cycling Weekly (N.D.)

* Walmart
* Amazon
* Dick’s Sporting Goods
* Recreational Equipment Incorporated (REI)
* Trek Bicycle Corporation
* Specialized Bicycle Components
* Giant Bicycles
* Academy Sports - Outdoors
* Target
* Performance Bicycle

Top Ten Bike Brands Based on Data Through 2022

* Trek
* Specialized
* Giant
* Cannondale
* Schwinn
* Scott
* Bianchi
* Santa Cruz
* Jamis
* Electra

Figure 1 shows the top ten brands of bikes per the three primary categories of biking models by brand name. This group of bikes is the target for any bike retailer to know if they want to offer these models to their customer base, or if they want to compare their sales of these models against the top sellers. This information is important for truly knowing the market and AdventureWorks is aware.

**Figure 1**

*Top Ten for Road Bikes, Mountain Bikes, and Touring Bikes*

|  |  |  |
| --- | --- | --- |
| **Mountain Bikes** | **Touring Bikes** | **Road Bikes** |
| Trex Fuel EX | Surly Long-Haul Trucker | Trek Domane |
| Specialized Stumpjumper | Kona Sutra | Specialized Roubaix |
| Giant Trance | Trek 520 | Giant TCR Advanced |
| Santa Cruz Hightower | Raleigh Sojourn | Cannondale SuperSix EVO |
| Cannondale Habit | Co-Motion Americano | Pinarello Dogma |
| Yeti SB130 | Salsa Marrakesh | Scott Addict |
| Rocky Mountain Instinct | Brother Cycles Big Bro | Bianchi Oltre XR4 |
| Scott Spark | Fuki Touring | Cervelo R5 |
| Kona Process | Thorn Nomad Mk2 | Fuji Transonic |
| Niner Rip 9 | Vivente World Randonneur | Colnago C64 |

*Note: this figure shows models of bikes that are sold worldwide by any distributor and comes from Cycling Weekly.*

**Call to Action**

AdventureWorks is a bike retailer with an annual revenue of around eight million dollars. In an industry projected to average around nine hundred & eighty-five thousand through 2022, this has been considered a solid revenue. AdventureWorks aims to maintain its niche market and compete with online retailers. Therefore, the call to action is to find ways to ensure the company’s sustainability by understanding its customer base in the ever-changing landscape of retailer models and competing with online markets like Amazon.

Based on the data reviewed from Xutiantony (2021), there are possible customer segments that can be analyzed in greater depth to determine which one of these segments leads to short-term and long-term growth. This analysis can also assist in shaping new marketing strategies and tactics. The potential benefits of identifying and analyzing customer segments include optimizing advertising campaigns, designing targeted promotions, and improving retention rates.

***Customer Segment Considerations***

1. Champions – this group will purchase the latest models that AdventureWorks will sell.
2. Loyal Customers – spends money more than once, responds to promotions.
3. Potential Loyalists – recent customers, spend good money, need to retain.
4. Recent Customers – Bought recently, possible first-time customers.
5. Promising – recent shoppers, low spend.
6. Customers needing attention – customers that spent good money in the past, but not recently.
7. About to sleep – customers that did not spend good money and did not purchase recently.
8. At Risk – customers that spent good money a long time ago, but not recently.
9. Can’t Lose Them – Customers that spent a lot of money a long time ago, need to win them back.
10. Hibernation – low spenders, low number of orders, and a long time ago, may show interest based on promotions.
11. Lost – low spenders, low number of orders, no interest based on promotions.

Classifying the customers in this format can help put focus on how to spend on promotions, loyalty programs, ad campaigns, and a host of other marketing ideas.

**Overview of Methods**

The dataset was found in GitHub and is in Kaggle. The primary method for doing the Exploratory Data Analysis (EDA) is Python. A Power PI editor query was created to explore the data and prepare dashboards that can be used for presenting any or all the customer segment programs, and other Key Performance Metrics as part of the journey of growing the business. These dashboards can help in visualizing the data and identifying trends, allowing for better decision-making and more effective implementation of marketing strategies.

A few methods will be used for the analysis, either as part of the EDA or for creating and solving the objectives. A Time Series analysis is one such type of trend that will be investigated. Python models that can be used include ARIMA, Bayes Theorem, Sklearn and Prophet. Depending on the initial results from the above analyses, other types of models may also be used.

Customer segmentation was previously discussed, and Python can be used to extract useful data, including age, region, marital status, and other categories.

Market Basket Analysis (MBA) is one of the objectives that will be reviewed for its effectiveness in analyzing the data. The retailer knows that getting additional business, and getting additional revenue from customers involves offering what the customer needs, beyond just a bicycle. Examples would include reviewing accessories like helmets, gloves, and other bike-related products.

**Project Objectives**

* Identify and analyze overall sales performance to identify key trends and patterns.
* Analyze profitability on overall sales for products to identify the profit margin and sales volume. What is the price point for the products that can create the higher revenues?
* Conduct a time series analysis to predict future sales. There are many tools to do a time series study. One method is called AutoRegressive Integrated Moving Average (ARIMA) model. Seasonality is also a measure that can be explored if there is a data supporting a seasonal sales trend.
* Segment customers based on purchasing behavior and use this information to tailor marketing strategies.
* Perform Market Basket Analysis (MBA) to find the most common accessories purchased together on one transaction.

The methods for reviewing these objectives will be Python, using specific programs like SKLearn, ARIBA, Linear and Multiple Regression techniques, and finally Market Basket Analysis. The objectives as stated may change in scope and language based on what the data brings to light.

**Assumptions and Limitations**

The data that is pulled from Github has only three years of data. This is enough to analyze as identified in the objectives, but not fully developed to do true customer segmentation and true revenue projections. There are also bike models that are not defined as indigenous to AdventureWorks or if they partner with specific manufacturers. Different models are similar in a few aspects that it may require a higher level of cleaning the data to determine if these models are treated the same or kept separate.

Understanding the data, having some in-depth analysis from the Python query, and reviewing visualizations necessary to put together a strong picture of the performance, can either improve some of the assumptions or identify new opportunities that can shed light on the prospects for AdventureWorks.

**Conclusion**

Exploratory Data Analysis is in progress, but what has been identified so far is the sales of the three primary types of bikes, the demographics for the customers purchasing the bikes, and the cost associated with selling the bikes. There are many accessories frequently purchased with a bike. The goal shall be uncovering the accessories that are required and which accessories may be removed and establishing a marketing strategy to support the business model going forward.

**Chapter 2: Literature Review**

**Introduction**

M. Hamdami Santoso defines “data mining” as a technique for finding patterns or finding interesting information in large, or voluminous amounts of data. The purpose of uncovering patterns is to provide meaning or direction for decision support (November 2021). The more often-used association rule discovery method used for data mining is the Apriori Algorithm. Without going into detail on how to set up the Apriori Algorithm, the premise is to find frequently occurring sets of items from transaction data and perform calculations to find the minimum value of support and confidence. This minimum value is known as the Association Rule. The results when running the Apriori algorithm is to find sets of items purchased together that meet minimum confidence, i.e. bikes and helmets to tie it back to the project for AdventureWorks. The values derived can be drilled deeper to provide data on models, colors, or any other characteristics to define marketing strategies. Another name for this process is “A Knowledge Discovery and Data (KDD) process” (Kaur, M., Kang, S., 2016).

How exactly can any company properly analyze future sales, in the short term, and more importantly the long term? This is where having specific statistical analysis tools comes to fruition. A basic name that is labeled for these tools is data mining. This is a generic catch-all term to use the data in a business to break down the quantitative, and even qualitative data that is gathered, stored, and categorized.

Per M. H. Saurker, the onset of this key data mining concept called Market Basket Analysis is demonstrated in the techniques shown below (2014).

Techniques of Data Mining

1. Classification: In classification, first examine the features of the newly presented object and assign it to a predefined class. For example, classify the credit applicants as low, medium, or high risk,
2. Association: The main goal of the association is to establish the relationship between items that exist in the market. The typical examples of association modeling are Market basket Analysis and cross-selling programs. The tools used for association rule mining are the Apriori algorithm and the Weka tool kit.
3. Prediction: In this functionality, the prediction of some unknown or missing attribute values is based on other information. For example: Forecast the sale value for next week based on available data.
4. Clustering: In this, Data Mining organizes data into meaningful sub-groups (clusters) such that points within the group are like each other, and as different as possible from the points in the other groups. It is an unsupervised classification. An effective dynamic unsupervised clustering algorithmic approach for market basket analysis has been proposed by Verma et al.
5. Outlier Analysis: In this, Data Mining is done to identify and explain exceptions. For example, in the case of Market Basket Data Analysis, an outlier can be some transaction that happens unusually (Saurker et al. 2014).

**Market Basket Analysis in a Retail Environment**

Retailers in all types of industries have the challenge of how to grow a business, like AdventureWorks, the primary focus of this case study. Many different articles propose various methods. Some are truly valid, some innovative, and some are just not worth discussing further. Before reviewing the Market Basket Analysis aspect, review methods of bike retailers for all sizes, and in any market must try to maintain or grow the primary focus of selling bikes (Seitz 2019)

The first idea according to Saurker (2014), the service department should be expanded to build relationships with customers. In the case of AdventureWorks, a service department can do more than just repair a flat tire, replace a chain, or add a light. The service department can equip bike riders with the tools and enhancements for their specific terrain. For example, if a bike rider lives in a rural area, the service team can outfit their bike with the appropriate tires, frame, seat, spokes, and more.

Another innovative way a service team can interact with customers is via apps like Strava. These apps help bike riders, especially less experienced riders, understand the topography, of their geographical area, and the equipment they may need, and foster a powerful personal relationship with the service department and company.

The second idea is to offer fittings (hint, this could become part of the MBA). More experienced riders don’t want to have to purchase fittings, try to put these fittings on their current bike, and hope the fittings accomplish what they are intended. Having a professional (sales and/or service team) understand the fittings, what they do, and how they work can solidify a relationship. If the market is not truly that open for carrying fittings like this, another opportunity is for the bike retailer to partner with such a partner and refer business to each other.

A third idea, which may be one of those reach opportunities is to offer rentals. When Seitz wrote this article, this would be a reach (2019). Today, in 2024, there is a proliferation of e-bikes. This may be the perfect opportunity for a bike retailer to feel the market for e-bikes by renting, and then determining if the business model can expand to offer this type of bike. Also, the area where a bike retailer may be situated is a tourist area. This can provide a new revenue stream, by providing a rental option for tourists that want to enjoy their time without having to figure out a plan to bring their bikes.

A fourth idea is to consider accepting selling ‘used’ and/or encouraging trade-ins. If this idea is to be considered, a plan to bring in used bikes would have to be part of the decision-making process. Buying bikes from other retailers, or from distributors may be one way. Another is to buy directly from your customers. Rather than expecting the customer to buy a new bike, and throw the bike away, take advantage of your service department, repair the bike, and choose to sell it, or rent. The financial advantage for the retailer is allowing a financially challenged customer to have an opportunity to make a purchase, perhaps establishing the beginnings of a long-term relationship that would not normally happen.

Other ideas in no order are partnering with online influencers, optimizing the layout for both the store that people visit as well as the online store, collaborating with the local business community to promote each other’s business, and partnering with schools. All these ideas in full, or in part, can grow a presence in the community, in the area, and in the industry (Waltz 2023). Focusing on that specific new trend known as online influencers, if one of the employees is not that person, search for and find that person. The reach of TikTok and Instagram is just too great to ignore per the Market Basket Analysis in Data Mining link in geeksforgeeks.org, (2023)

The bike industry sells bikes and accessories that are specific to that industry. A person looking for a bike is most likely going to look for items that add, enhance, improve, or sustain the experience for the user. Bikers are not typically going to purchase online randomly. Bikes need to be seen, tested, measured, and experienced for a user to make the decision to spend money that can be quite high. That means a bike retailer has to offer more than just bikes.

Examples of the most common accessories are helmets, water bottles, bike locks, bike pumps, pouches, clothing, and gloves. That is a lot, but there is more. There are reflectors, repair kits, mirrors, pads, shoes, odometers, horns, travel guides, bike chain oil, and probably several other items. This is where market basket analysis becomes a data tool that any bike retailer will want to understand and utilize.

The basic definition is analyzing the combination of products that are purchased together. The most well-known, and oft-used application of this concept is in grocery stores. But it is not limited to that industry. The analysis works with this common association rule [IF] -> [THEN]. The antecedent is the IF and the consequent is THEN. For a bike retailer, the bike is the antecedent, and any accessory or accessories is the consequent. The retailer must find a way to calculate the Consequent. The common formula, known as an algorithm, is the Apriori Algorithm. Spreadsheets are used, but with a lot of data, spreadsheets are limited so having the means to use Data Science applications like Python or SQL would be preferable.

The three components in the Apriori Algorithm are support, confidence, and lift. Each component has a basic math concept shown as:

Support = freq(A, B) / N – writing this equation, it says the number of transactions for items A & B divided by total transactions N. The answer is going to be a percentage.

Confidence = freq(A, B) / freq(A) – writing this equation, it says combined transactions A & B divided by individual transactions. The answer is going to be a percentage

Lift = confidence percentage/support percentage. This is where the resulting answer provides the true insight. The answer when below 1 will mean combinations are not common. The answer when above 1 and going even higher adds confidence the combinations are more common. Li (2017)

**Figure 2**

*Rules for support, confidence, lift*

A diagram of a problem

Description automatically generated

*Note:**Visual for showing the diagram of support, confidence, lift*

**Figure** 3

*Extract to show the rules in a colorful diagram*

A row of shopping carts with different colored bags

Description automatically generated

Note *A visual look at the rules using ratios rather than a figure*

The math in the figure is simple but it gets more difficult with so many accessories and so many transactions.

There are three primary types of Market Basket Analysis (MBA) per Cecil Hermina in the journal Market Basket Analysis for a Supermarket (2014).

1. A descriptive MBA is looking for customer behavior by using the data from the three methods. There are a couple of uses for this type of data, such as store layout – placing accessories right next to the bikes they are most often associated with – and promotions – buy a bike, get common accessories at a discounted rate.
2. Predictive MBA is looking at predicting sales based on trends the data is showing. The data can be in a period such as a quarter, a length of time such as a year, or during a promotional period. This kind of data can be used to determine how many accessories to carry, not carry, or have extra material stored somewhere or easily attainable.
3. Differential MBA is looking at comparing different data results to draw comparisons and use that data for decision-making. Different customers buy different accessories, but over the same period so understanding those comparisons is vital. A specific example to help put this analysis into practice is a retailer that operates in an area that has competitive races along with tourists. Accessories are varied for these groups, but the time of season promotes the element of being prepared to offer more of these options and how to position them in a store layout and pricing model.

The results of the analysis will predict customer behaviors and offer opportunities to capitalize on these behaviors. The benefits are becoming clear, improved customer relationships, pricing models, and sales that can increase year over year.

**Conclusion**

Market Basket Analysis (MBA) will be a useful process to help with the study of a retail business model. Bike retailers have the added challenge of having a wide array of accessories, and a potential for a diverse customer base that will make the data important to find patterns.

There are other techniques beyond just using Apriori, but this algorithm is the most-often used tool that the company will examine to be able to give them the best information that will help them grow and help them compete against the bigger retailers in the bike industry.

**Chapter 3: Methodology**

**Introduction**

The overall scope of this capstone project aims to outline the systematic approach that is being taken to analyze the bike company AdventureWorks. The scope is a vital lifeline for AdventureWorks because there are so many competitors in this space. It is not a high-margin field, and AdventureWorks cannot set its price as the market dictates what customers are willing to pay for bikes, for accessories, for clothing, for advice. The customers will continue to remain with AdventureWorks for specific reasons such as customer service experience, available products at the proper price points, service, quality, warranty, and loyalty.

Five specific paths have been determined to be primary objectives for this market basket analysis, shown in the list below:

1. Identify and analyze overall sales performance to identify key trends and patterns.
2. Analyze profitability on overall sales for products to identify the profit margin and sales volume.
3. Conduct a time series analysis to predict future sales.
4. Segment customers based on purchasing behavior and use this information to tailor marketing strategies.
5. Perform Market Basket Analysis to find the most common accessories purchased together on one transaction.

AdventureWorks can now start going into the methodologies of how to evaluate the topics and find out what has been uncovered.

**Data Collection**

The dataset for this project was sourced from GitHub, as noted in Chapter 1. The dataset is comprised of three years of sales data, including transaction details specific to product types, product cost, product sales, customer demographics, regional demographics, and accessory items. The various data types are going to be analyzed in combinations of detail to generate a more targeted analysis of trends for sales of bikes, and accessories, and predictions for growing the company.

**Data Preprocessing**

The dataset has multiple subsets that were joined together to create over 56,000 rows of sales data, with 35 different columns or variables to be selected for more detailed analysis. The columns included product codes or SKUs, cost, revenue, order quantity, product codes, and demographic data like customer names, marital status, income, location, emails, and additional details The list shows a few of the columns AdventureWorks tracks.

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 OrderDate 56046 non-null object

1 StockDate 56046 non-null object

2 OrderNumber 56046 non-null object

3 ProductKey 56046 non-null int64

4 CustomerKey 56046 non-null int64

5 TerritoryKey 56046 non-null int64

6 OrderLineItem 56046 non-null int64

7 OrderQuantity 56046 non-null int64

8 ProductSubcategoryKey 56046 non-null int64

9 ProductSKU 56046 non-null object

10 ProductName 56046 non-null object

11 ModelName 56046 non-null object

12 ProductDescription 56046 non-null object

13 ProductColor 29168 non-null object

14 ProductSize 56046 non-null object

15 ProductStyle 56046 non-null object

16 ProductCost 56046 non-null float64

17 ProductPrice 56046 non-null float64

18 SubcategoryName 56046 non-null object

19 ProductCategoryKey 56046 non-null int64

20 Region 56046 non-null object

21 Country 56046 non-null object

22 Continent 56046 non-null object

23 CategoryName 56046 non-null object

24 FirstName 56030 non-null object

25 LastName 56030 non-null object

26 BirthDate 56030 non-null object

27 MaritalStatus 56030 non-null object

28 Gender 55650 non-null object

29 EmailAddress 56030 non-null object

30 AnnualIncome 56030 non-null float64

31 TotalChildren 56030 non-null float64

32 EducationLevel 56030 non-null object

33 Occupation 56030 non-null object

34 HomeOwner 56030 non-null object

dtypes: float64(4), int64(7), object(24)

The dataset did not require much cleaning. Some duplicate entries were removed, and data types were changed specifically to running analysis in either the Python program or more specifically with doing joins in Power BI. There also is feature engineering which is the process of creating quarterly and monthly buckets, and an interesting study of reviewing a specific day for any insights.

In addition, feature engineering also aggregated sales data such as average sales, gross sales, frequency, counts, and combinations. Graphing techniques are also used to show a visual performance metric. The next few graphs show an analysis of AdventureWorks revenue in a few interesting categories that are going to be the focus of analyzing the objectives.

One graph shows the age groups for bike purchases at AdventureWorks. This specific bar graph shows that one of the categories of bikes that AdventureWorks sells is Road bikes. Road bikes typically are the type of bike a customer purchases that chooses to ride on paved or gravel surfaces, and usually for distance or off the main roads. Based on the top-selling age group of 56-65, this represents the age where the older customer segment is looking for an alternative to driving, or to find a more entertaining form of exercise. There is not much in terms of sales for the younger age group, which can mean, that perhaps the older age group is buying this type of bike for their children or other factors that AdventureWorks would be wise to fully understand.

**Figure 4**

*Total Sales for Road Bikes per Age Group*

A graph with different colored bars

Description automatically generated

*Note:* the age groups are in 10-year buckets, and the sales are in millions. The key takeaways are the age groups that are customers, and the age groups that are not yet customers*.*

The next figure shows a sales breakdown of bikes, accessories, and clothing by total sales. The bar graph is showing quantity of sales, not the price, so that is why accessories are much higher. It is typical for a customer to buy a bike, but also buy many different types of accessories that AdventureWorks is offering.

**Figure 5**

*Bike & accessory distribution*

A bar chart with different colored rectangles

Description automatically generated

*Note:* the bar chart represents the three primary categories and the quantity of units sold*.*

This last figure 6 will show quarterly sales and the trend that is growth. The bar graph shows quarters for ease of comparison, but the bar chart could easily be broken out to show months or weeks and get a more visceral picture of the sales trends for AdventureWorks.

**Figure 6**

*Bike Sales by Quarter*

A graph of sales by quarter

Description automatically generated

*Note:* the bar chart is showing revenue by quarter in chronological order

*Statistical and Machine Learning models:*The study of AdventureWorks is based on the objectives listed earlier. The list will briefly outline the models that are going to be reviewed:

1. Time series analysis: the prediction of future performance will be done using the ARIMA (AutoRegressive Integrated Moving Average) model.
2. Market Basket Analysis (MBA) is the true data science approach for AdventureWorks. There is an algorithm known as Apriori that is part of the mlxtend library that has been applied to this point to start producing accessory purchases and spending. The results so far have produced results that requires further exploration. The results of this analysis can improve or expand on the strategies AdventureWorks wishes to implement for marketing programs to grow the business.
3. Customer Segmentation: there is a Python library called SKLEARN that has been used for customer segmentation analysis. This library has attempted to determine the efficacy of this analysis and discarded if the data is too narrow or not fully developed. The overall importance of doing this analysis is related to understanding the demographics and tailoring marketing strategies around this. The previous chapter discussed valid reasons for employing strategies to recruit, retain, and grow the customer base and that is the goal of exploring clustering algorithms.

**Time Series**

AdventureWorks wants to continue to see growth in sales, by quantity as well as revenue. Time series does a forecasting process that helps to put some numbers into the calculation and then is analyzed for predicting future trends. Time Series analysis is a complex and not straightforward as it involves discovering hidden factors and rooting out what is called noise. But, most importantly, forecasting requires AdventureWorks to know the factors that influence the current state of its target variables (Farhad Malik, Dec. 26, 2023).

One methodology is regression analysis which is a technique requiring formulating a mathematical equation that can be used in providing a value as close to the actual observed value as possible (Farhad Malik, Dec. 26, 2023). When the values are shown in a scatter plot, and re-occurring patters are observed, this indicates the existence of seasonality. AdventureWorks can expect to see seasonality it the data.

ARIMA and SARIMA are linear regression models that Adventures has used to predict future revenue. Briefly explaining what the properties mean, the AR property of the (AR)IMA model and the property is defined as P. ARIMA models consider past time points that can impact current and future points. ARIMA uses lag observations of the time series to forecast observations and applies a weight to each point.

AR(I)MA is integrated. If a trend exists then the time series is considered non-stationery and seasonality is present (Farhad Malik, Dec. 26, 2023). The property applied is defined as D.

ARI(MA) is moving average. Error terms predict current and future observation points from the previous points. The MA removes random movements Farhad Malik, Dec.26, 2023). The property applied is defined as Q.

Taking all of these property classifications into consideration is part of the modeling process and will be shown for executing the linear regression model with the results explained in the next section.

**Market Basket**

Chonny wrote an excellent blog in the online publication Towards Data Science that details the Apriori algorithm. The famous story that Chonny references in his blog is the “beer and diaper” story. Data Scientists discovered customers who purchase diapers will also buy beer. This point emphasized that there are interesting association rules hidden in our data sets (2023).

Concepts of Apriori that are most often used are Support, Confidence, and Lift.

Support is a fraction of transactions that contain an itemset – the support of item N is defined as the number of transactions containing N divided by the total number of transactions. Another name for N is Antecedent.

Confidence measures how often items in Y appear in transactions that contain X. Another way to state this is the likelihood that item Y is bought if item X is bought. The calculation is the number of transactions containing X and Y divided by the number of transactions containing X. Another name for the output of this formula is Consequent.

Lift represents how much more likely the Consequent is to be purchased when the Antecedent is purchased. The calculation is the confidence divided by the support of the consequent.

The Market Basket Analysis will be performed on the AdventureWorks dataset. There are limitations however for the Apriori Algorithm. The size of the dataset is the first limitation. AdventureWorks has three years of data at this point, but over time the dataset will increase when new sales data is entered. The second limitation is time. The transactions are re-counted each time new transactions are added to the dataset.

To generate some analysis of market baskets, the code follows a premise of finding all the orders over the period and putting those orders into groups of 0 and 1. The reason is that for orders that have only one item purchased, there is no second item or multiple items, and there is no market basket analysis. So, the output classifies those order numbers as 0. When the output does identify multiple items, one or more, purchased together, then the output is going to assign a value of 1. The final output for all orders is either a 0 or a 1 shown in Figure Then, the code is written to try now only to classify the orders with a 1 and put together the Antecedent, identified as the first item on the order, and the Consequent, the second item on the order. This analysis is going to be done and the results will be shared in the following section.

**Customer Segmentation**

The data models that have been developed and initially tested are going to analyze the data specific to the objectives outlined. The idea is to validate the models and provide the proper analysis that can be used for the next steps of what AdventureWorks can do to grow its business and compete with its many competitors. The list for the evaluation metric process is shown below:

1. Accuracy – measuring the proportion of true positives vs the true negatives in the AdventureWorks dataset.
2. Precision – measuring the ratio of true positives against the sum of true positives and false positives.
3. Recall – The ratio of true positives against the sum of true positives and false negatives.
4. F1 score: measuring the results of the precision and recall metrics. These scores in theory should be aligned.
5. AUC – ROC: the acronym stands for Area Under the Receiver Operation Characteristic Curve. The AUC – ROC is a classification technique used to evaluate the metrics from a zero () to one (1) scale and the higher the value along this axis, this will be considered the ideal classification process.

All these metrics listed are part of the SKLEARN library. The evaluation is done after the various customer segmentation features are created.

**Choose the models**

At this point, the models are being developed and analyzed. Nothing has been ruled out, and the list of classification programs may grow. The process of determining the results, analyzing the data, and producing a convincing story based on the data is part of the next step of this process.

**Conclusion**

The methodology process is a time-consuming process that requires a full understanding of the data, and a deep perspective of what the data shows after choosing the various customer segmentation methods and running market basket analysis on multiple variables. The approach is starting to provide results that start to paint a picture of growth, of a deep portfolio of accessories that are aligned with bikes, and a customer segmentation group that can be targeted for more robust marketing strategies. The findings, as the analysis goes into an even deeper portrayal of AdventureWorks, can only offer more valuable insights that will be utilized for taking AdventureWorks into the next product cycle.

**Chapter 4: Results of Methodology**

**Introduction**

The overall scope of this capstone project aims to outline the systematic approach that is being taken to analyze the bike company AdventureWorks. The scope is a vital lifeline for AdventureWorks because there are so many competitors in this space. It is not a high-margin field, and AdventureWorks cannot set its price as the market dictates what customers are willing to pay for bikes, for accessories, for clothing, for advice. The customers will continue to remain with AdventureWorks for specific reasons such as customer service experience, available products at the proper price points, service, quality, warranty, and loyalty.

Five specific paths have been determined to be primary objectives for this market basket analysis, shown in the list below:

1. Identify and analyze overall sales performance to identify key trends and patterns.
2. Analyze profitability on overall sales for products to identify the profit margin and sales volume.
3. Conduct a time series analysis to predict future sales.
4. Segment customers based on purchasing behavior and use this information to tailor marketing strategies.
5. Perform Market Basket Analysis to find the most common accessories purchased with specific bike models and predict future accessory activity.

In this chapter, the results of the analysis conducted on AdventureWorks Inc. will be presented on the objectives. The analysis includes insights gained from exploring overall sales performance, time series forecasting, customer segmentation, and market basket analysis.

**Trend Analysis**

The analysis of AdventureWorks overall sales performance revealed several key trends and patterns. Monthly sales data was examined to identify noticeable fluctuations or growth trends. Figure 7 illustrates the monthly sales trend over the three years

**Figure 7**

*Monthly sales trend*

A graph with a red line and blue dots

Description automatically generated

*Note:* The trend line has a vertical slope, and AdventureWorks actual sales represented by the blue point are quite close to this trend line.

As shown in Figure 7, AdventureWorks expects the sales trend to continue. This trend suggests AdventureWorks has started approaching a strategy that has attracted more customers, and perhaps started to recognize at this point the key metrics that have proven successful in the past.

**Profitability Analysis**

The profitability analysis aimed to identify similarities between profit margin and sales volume for AdventureWorks products. By examining the relationship between product cost, revenue, and profit margin, insights were gained into the pricing strategies that could maximize revenue. Figure 2 displays the product cost and revenue specific to bike color.

**Figure 8**

*Revenue vs Profit for Each Color Bike*

A graph of sales and revenue

Description automatically generated with medium confidence

*Note:* The bar chart was created in Power BI and shows a clear signal the colors are most popular – black followed by red.

Figure 8 represents what was mentioned earlier, that recognizing bike color impacts sales. The black models and the red models are the top two colors, and there is very low interest in bikes that are white in color. This information is important for several business initiatives, such as inventory, pricing, and product placement.

Figure 9 also is a representation of having a strong customer data portfolio. Knowing the ages of customers who purchase specific bikes also aids in developing a strategy for sales, advertising, product placement, and inventory.

**Figure 9**

*Road Bike Sales per Age Group*

A graph with different colored bars

Description automatically generated

*Note: road bikes have the most sales in the 56 to 65 age group*

Further analysis is important to continue to identify the optimal price points for maximizing revenue yet maintaining competitive pricing with the competition.

**Time Series Forecasting**

The time series analysis focused on predicting future sales using the ARIMA model and the SARIMA model. While limited by the dataset’s three-year timeframe, the analysis provided insights into potential sales trends. Table 1 showcases the forecasted sales trend using the ARIMA model. The most important points that Figure 10 shows are Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). This information is important to recognize the value for AIC and BIC. These values are measures of the model’s fit and complexity. When comparing AIC and BIC for different models, the lower the result, the better the model is for forecasting.

The values produced are similar based on the model. The variables, also identified as parameters, used in this model are 1, 1, 1. These parameters are selectively chosen and represent the order of the autoregressive, differencing, and moving average components, respectively (Farhad Malik, Dec. 26, 2023). The parameters are changed to compare the models. The results of these parameter changes are shown in Figure 10, listed only by the AIC and BIC values and not the full data.

**Figure 10**

*Running AutoRegressive Integrated Moving Average (ARIMA) simplified*

ARIMAX Results

====================================================================

Dep. Variable: Profit No. Observations: 30

Model: ARIMA(1, 1, 1) – P, D, Q assigned numerical values

AIC 730.157

BIC 734.259

Sample: 01-31-2015 HQIC 731.442 06-30-2017

*Note:* The most important takeaway is showing the AIC value 730.157 and BIC value 734.259.

AdventureWorks is trying to find the lowest of these scores for determining the best model. Additional queries are not shown as a figure, but the results are produced in Table 1 that is showing the results of AIC and BIC over four specifically selected parameters to compare head-to-head and make the selection of the model that appears to be the best.

**Table 1**

*AIC and BIC results based on different selection of parameters*

Model (P, D, Q) AIC BIC Rank

1, 1, 1 730.157 734.259 3

2, 1, 2 732.228 739.064 4

4, 2, 4 715.979 727.969 2

4, 4, 4 681.013 692.335 1

*Note:* Comparing models using AIC and BIC and different selections of parameters using many iterations based on the parameters selected and making the final choice

The first result is using parameters 1, 1, 1. The second result is using parameters 2, 1, 2. The third result is using parameters 4, 2, 4. The fourth result is using parameters 4, 4, 4. The last grouping of parameters produces the lowest of output, so this is the model that is best suited for doing an ARIMA analysis. The results can be graphed and are shown in Figure 11.

**Figure 11**

*Graphed results of the ARIMA model using 4, 4, 4 as the parameters*

A group of graphs and diagrams

Description automatically generated

The standardized residual for P is shown on the line graph, centered around the 0 plot, but most often higher (or lower) than 0.05.. The higher the p-value, the interpretation is the residuals are approximately normally distributed. This means that this model meets the assumptions AdventureWorks is looking at for predicting future profitability.

There is also the seasonality ARIMA or SARIMA. This output improves the overall result that is generated based on using the parameters. However, caution is needed for reviewing seasonality because the model may choose to not consider some of the categorical variables that are included with the ARIMA model. This model is being considered and for the time being it is not included for the next twelve months of forecasted profits.

Table 2 shows the next twelve months of forecasted profits based on the data that is available. As with any business model, the sales and profits don’t happen just because the forecast shows a positive trend. The marketing plan must be aligned with all of the other pieces of data that are available to create multiple scenarios that are key to successfully meeting the forecast.

**Table 2**

*Predictive forecasted profits for AdeventureWorks*

*Forecast over the next 12 months*

ARIMA Forecast:

Future Date Predicted Profits

2017-07-31 910,702.36

2017-08-31 920,028.37

2017-09-30 1,024,860.15

2017-10-31 1,072,865.27

2017-11-30 1,146,563.15

2017-12-31 1,372,944.78

2018-01-31 1,274,850.37

2018-02-28 1,316,362.30

2018-03-31 1,392,553.00

2018-04-30 1,443,141.98

2018-05-31 1,576,324.81

2018-06-30 1,620,309.58

*Note:* The results of the ARIMA method show a positive trend based on a monthly output.

This forecast can be graphed using a line chart as well as shown in Figure 12.

**Figure 12**

*Line chart showing the profit forecast*

A line graph with numbers and a line

Description automatically generated

**Market Basket Analysis**

Market basket analysis focused on identifying common accessory purchases associated with various bike models. The Apriori algorithm is the primary method for extracting the accessory purchases from the dataset. The accessory purchases were explored in the process of trying to determine the most popular accessories as well as the least popular.

The information that is shown in Figure 13 is a highlight of what the Apriori algorithm has produced so far. This data shows accessories that are purchased with bikes. The values are between 0 and 1. The results that show accessories that are trending towards 1 mean that the number is a percentage of sales that include the accessory. The number 60 is the product key and the accessory is a tire. There is a strong association between a tire being purchased with a bike. The water bottle is another example of an accessory that has a strong sales presence for the customer when purchasing a bike.

**Figure 13**

*Market Basket Analysis using the Apriori Algorithm*

A screenshot of a computer

Description automatically generated

*Note*: There is a lot of information that is produced. The antecedent is the first product, and the consequent are the products purchased at the same time. The data is showing the support, confidence, and lift of the top ten parts. The higher the value, the more likely AdventureWorks will predict which accessories are purchased together. The output shows the product key and the product name. Tires and tubes are frequently purchased together. Bottle cages and water bottles are frequently purchased together.

There are many ways to analyze Market Basket Analysis (MBA). One of the ways to compare MBA against variables like the country. The reason to try this is to determine if the same products are sold together in the UK, Australia, or Canada. AdventureWorks has run the analysis, and the output is shown in Table 3.

**Table 3**

*Comparing results of the Antecedent, Consequent, Support, Confidence, Lift*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Country | Total combos | Antecedent | Consequent | Support | Confidence | Lift |
| United States | 19,811 | Mountain Bottle Cage | 30 oz water bottle | 0.077156 | 0.872562 | 5.004145 |
| Australia | 12,489 | Road Bottle Cage | 30 oz water bottle | 0.070627 | 0.862903 | 6.953715 |
| Canada | 6,875 | Mountain Bottle Cage | 30 oz water bottle | 0.091601 | 0.847095 | 4.657481 |
| United Kingdom | 6,423 | Road Bottle Cage | 30 oz water bottle | 0.074341 | 0.949309 | 5.718553 |
| Germany | 5,289 | Mountain Bottle Cage | 30 oz water bottle | 0.061029 | 0.786517 | 4.82425 |
| France | 5,239 | Road Bottle Cage | 30 oz water bottle | 0.066955 | 0.901163 | 6.321794 |

*Note:* The value used for executing the analysis for min support is 0.06.

Canada is the only country that had a higher value for support, which is 0.09. Germany has a lower value, which is .06. The values are in the table and tell a story. The values in Support are the percentage of times this item is in a transaction. That means, 6 to 9 percent of the time, the two flavors of bottle cages and the 30 oz water bottle are in total transactions. The values in Confidence range from .78 to .90. What the values show is that 78 percent of all transactions that have the cage will also have the 30 oz water bottle. Germany has a lower percentage than France, which has the highest at over 90 percent. Lift is another predictive element that can be interpreted as the higher the value, the stronger the association. The county of France has a Lift of 6.32. This is interpreted as the purchase of the road bottle cage increases the probability the 30 oz water bottle will be purchased by a factor of 6.32 compared to the expected probability. There are other specific conditions in the Appendix B showing additional MBA analyses.

**Customer Segmentation**

Customer segmentation analysis aimed to identify distinct customer segments based on purchasing behavior. Various clustering algorithms were applied to categorize customers into meaningful groups. The age group and the bike color were the most obvious segments that were uncovered using the analysis. Most of the sales for AdventureWorks occurred in the US so that data is not interesting from a customer segmentation. Figure 14 shown below gives a clear picture of the color that by a wide margin is the top revenue and profit maker for AdventureWorks. Over the three years, the black color represents nearly double the sales and profit of the rest of the colors.

**Figure 14**

*Sum of Profit by Bike Color*

A screenshot of a graph

Description automatically generated

*Note:* Profit by bike color with the top-selling color by a wide margin is black. Multi-colored and white colored bikes really do not sell*.*

Another interesting customer segmentation output is shown in the cluster below. Figure 17 shows the count of individuals with different marital statuses within each cluster. There are two bars with five clusters. The 0 represents being married and the 1 represents not having a married status. The height of each cluster represents the number of customers in that cluster. The clusters can be changed based on the characteristics that are considered important. For this specific example, marital status, gender, education, occupation, and homeowner are selected, in order of 0 to 4. The legend identifies the color of each feature.

The information generated in this bar chart shown in Figure 15 can be further used for marketing strategies. For example, the customer who is married and is a homeowner has the higher number of individuals in that cluster. AdventureWorks can use that information in setting forth a marketing strategy tailored to that segment. There are other clusters shown in appendix.

**Figure 15**

*Marital Status Distribution by Clusters*

A colorful graph with black border

Description automatically generated

*Note:* The two bars represent married with the 0 and not married with the 1. There are five clusters based on features selected (Marital Status, Gender, Education Level, Occupation, Homeowner). The height of the bars represents the number of AdventureWorks customers who are married vs the number of customers who are not married. AdventureWorks can gain additional insight from this cluster bar graph and use that information appropriately.

**Conclusion**

The analysis conducted on AdventureWorks has yielded valuable insights into many key performance indicators. Examples include sales performance, future sales trends, profitability, customer segmentation, and market basket analysis. AdventureWorks can use these insights to be used for preparing strategic decision-making initiatives, and optimize marketing strategies that will drive growth, giving confidence to compete against the bigger, more established bike retail giants.

AdventureWorks has the tools necessary to innovate, enhance the customer experience, win on opportunities in specific markets, and expand its sales and marketing experience into new markets, especially internationally.

**Chapter 5: Conclusion**

**Introduction**

Based on the analysis conducted for AdventureWorks Inc. there is a plethora of valuable insights. By addressing the objectives, AdventureWorks has unlocked valuable information that can inform strategic decision-making that leads to growth in a highly competitive market.

*Summary of Findings:*

The objectives are restated.

# Objectives:

1. Identify and analyze overall sales performance to identify key trends and patterns.
2. Analyze profitability on overall sales for products to identify the profit margin and sales volume.
3. Conduct a time series analysis to predict future sales.
4. Segment customers based on purchasing behavior and use this information to tailor marketing strategies.
5. Perform Market Basket Analysis to find the most common accessories that are purchased together

**Summarizing**

The examination of overall sales performance over the three-year period has revealed consistent growth trends. This indicates that AdventureWorks has been successful in attracting more customers. The key metric is the recognition of understanding the importance of bike colors and having those colors available to drive sales.

Moreover, the profitability analysis has shed light on the profit margin and sales volume, particularly emphasizing the importance of certain bike colors in driving revenue. This understanding is pivotal for inventory management, pricing strategies, and product placement.

The time series forecasting using ARIMA and SARIMA models has provided valuable insights into future sales trends. AdventureWorks can use the data accumulated to start doing modeling for the next few years knowing where the customer segmentation strengths and weaknesses are and creating strategies to take advantage of these trends.

Additionally, customer segmentation analysis has identified other categories that have not yet been developed, such as children, and marital status. This segmentation enables a more direct target marketing strategy tailored to these demographics, which will maximize sales opportunities and grow customer loyalty.

Finally, the market basket analysis has uncovered common accessory purchases, offering insights into consumer preferences, and purchasing patterns. This knowledge empowers AdventureWorks to optimize its product offerings and enhance the customer experience.

In conclusion, the findings from the analysis conducted prepare AdventureWorks with invaluable tools and insights to create strategic decision-making and maximize its competitive position in the bike retail market. By leveraging these insights, AdventureWorks can innovate, improve the customer experience, and expand into new territories with the confidence that there is a path for increasing sales, customers, and revenue in the bike retail landscape.

References

Harrison, J., Kjelander, S., Komolov, P., Loza, L., & Shah, R. (2012, December 10). *The Bike Kitchen Business Plan*. pp. 1-50. <https://open.library.ubc.ca/media/stream/pdf/66428/1.0108519/1>

Gupta, N., Moro, M., Ayala, K. A., & Sadler, B. (2020, Fall). *Price Optimization for Revenue Maximization at Scale*. 1-32. <https://scholar.smu.edu/cgi/viewcontent.cgi?article=1167&context=datasciencereview>

Nawaz, I, Shaukat & K, Zaheer, S, (2015). *Association Rule Mining: An Application Perspective*. pp 29-38. Research Gate. <https://www.researchgate.net/profile/Kamran-Shaukat-Dar/publication/284721728_Association_Rule_Mining_An_Application_Perspective/links/59564727aca272fbb37d21d4/Association-Rule-Mining-An-Application-Perspective.pdf>

Kabir. H. (2016). *Data Mining Framework for Generating Sales Decision-Making Information Using Association Rules*. pp. 1-8. <https://pdfs.semanticscholar.org/f6c1/04146de021a8a4cc1624cc9d2ce196c706a7.pdf>

Purdue Online Writing Lab (n.d.). General Writing FAQ’s. Purdue Online Writing Lab. <https://owl.purdue.edu/owl/general_writing/general_wrirting_faqs.html>

Baker, C. (n.d.). Customer Retention Part 1. Nbda. <https://nbda.com/customerretentionpart1/>

Liu, D., Yang, Y., Yao, E., Shao, X., Zhao, F., & Zhu, Y. (2023, February 3). *Short-Term Forecasting of Dockless Bike-Sharing Demand with the Built Environment and Weather.* pp, 1-50. <https://www.hindawi.com/journals/jat/2023/7407748/>

Rodriguez, C.P., Aracil, J.P., Fister, D. Lopez, R.T., & Sanz, S.S. (2023, January). *Bike Sharing and Cable Car Demand Forecasting using Machine Learning and Deep Learning Multivariate Time Series Approaches*. pp. 1-22. <https://www.sciencedirect.com/science/article/pii/S0957417423027665>

Chau, D.C.K., Ngai, E.W.T., & Xiu, Li., (2009, March). *Application of Data Mining Techniques in Customer Relationship Management: A Literature Review and Classification*. pp. 2592-2602. <https://www.sciencedirect.com/science/article/abs/pii/S0957417408001243#preview-section-abstract>

Hilage, T.A., & Dr. Kulkarni, R. V. (2011, November). *Application of Data Mining Techniques to a Selection Business Organization with Special Reference to Buying Behavior*. pp. 1-13. <https://airccse.org/journal/ijdms/papers/3411ijdms13.pdf>

Jitkar, B.D., Kulkarni, R.V., & Raarone, A.A.(2012, February). *Association Rule – Extracting Knowledge Using Market Basket Analysis*. pp. 19-27. <http://www.kresttechnology.com/krest-academic-projects/krest-mtech-projects/CSE----/mtech-dotnet-abstracts-bpapers/Data%20Mining/37)%20Association%20Rule%20Extracting/37)%20Association%20Rule%20Extracting.pdf>

Aguimis, H., Forcum, L.E., & Joo, H. (2013, November). *Using Market Basket Analysis in Management Research*. pp. 1-26. <https://www.hermanaguinis.com/pdf/JOMMBA.pdf>

Bhagat, P., Bhudade, V., Khaparde, A., & Saurker, A. (2014. April). *A Review Paper on Various Data Mining Techniques*. pp. 98-101. <https://www.semanticscholar.org/paper/A-Review-Paper-on-Various-Data-Mining-Techniques-Saurkar-Bhujade/883325ac20546fe6b73660a400cd8c157d8f6e3c>

Meenakshi (2014, May). *A Review on Association Rule Mining*. pp 1-5. <http://www.ijarse.com/images/fullpdf/1524646540_Paper_2_ij.pdf>

Kaur, M., Kang, S. (2016, May 5). Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining. pp. 1-8. Science Direct. <https://www.sciencedirect.com/science/article/pii/S1877050916305208>

Nunan, J. (2016, December 1). Customer Retention Strategies for Bike Shops Which Won’t Cost a Cent. LinkedIn. <https://www.linkedin.com/pulse/customer-retention-strategies-bike-shops-which-wont-cost-nunan/>

Li, S. (2017, September 24). A Gentle Introduction on Market Basket Analysis – Association Rules. 1-12, Towards Data Science. <https://towardsdatascience.com/a-gentle-introduction-on-market-basket-analysis-association-rules-fa4b986a40ce>

Li, S. (2017, September 24). A Gentle Introduction on Market Basket Analysis – Association Rules. pp 1-8. Towards Data Science. <https://towardsdatascience.com/a-gentle-introduction-on-market-basket-analysis-association-rules-fa4b986a40ce>

Seitz, M. (2019, June 24). 4 Sales Strategies Worth Trying in Your Bike Shop. pp 1-5. Science Direct. <https://thebikecooperative.com/bike-shop-management/4-ways-to-improve-bike-sales-with-a-new-strategy/>

Untold Data Science. (2020, March 3). *Market Basket Analysis in Python*[Video]. YouTube. <https://www.youtube.com/watch?v=4QIWJVVWJdQ>

Untold Data Science (2020, March 10). *Association Rule Analysis / Market Basket analysis[Video].* YouTube. <https://www.youtube.com/watch?v=4OORU17o1GY>

Santoso, M.H. (2021, November). *Application of Association Rule Method Using Apriori Algorithm to Find Sales Patterns: Case Study of Indomaret Tanjung Anom*. pp 1-13, [https://jurnal.itscience.org/index.php/brilliance/article/view/1228/885.](https://mkmanu.wordpress.com/2014/09/11/what-is-causing-customers-to-churn-attrition-analysis-using-r/)

Xutiantony (2021, December). AdventureWorks: Analysis Techniques and RFM Model. Medium. <https://medium.com/@xutiantony/adventure-works-analysis-techniques-and-rfm-model-56f6aa8ecf08>

Beckner, M., Jackson, R., & Steidel, K. (2022, October). *Improving Marketing Efficiency in the Retail Bicycle Industry through Geospatial Segmentation*. pp. 1-6. <https://www.ijtef.com/vol13/741-F707.pdf>

Hermina, C.I. (2022, November). *Market Basket Analysis for a Supermarket. pp 1-9. Research Gate.* <https://www.researchgate.net/publication/365489098_MARKET_BASKET_ANALYSIS_FOR_A_SUPERMARKET>

Market Basket Analysis in Data Mining (2023, February 15). Geeksforgeeks. <https://www.geeksforgeeks.org/market-basket-analysis-in-data-mining/>

Waltz, A. (2023, May 28). How to Grow Your Bike Shop in 10 Ways. pp. 1-8. <https://www.podium.com/article/how-to-grow-your-bike-shop/>

panData (2024, February 6). Mastering Time Series Analysis from Scratch: A Data Scientist’s Roadmap. Level Up GitConnected. 1-23, <https://levelup.gitconnected.com/mastering-time-series-analysis-from-scratch-a-data-scientists-roadmap-437ff673df72>

Warepam, R. (2024, March 5). How to Perform Statistical Analysis Using Python: The Ultimate guide. Medium. <https://medium.com/illumination/how-to-perform-statistical-analysis-using-python-the-ultimate-guide-9458ae0ace1c>

Desmaris, B. (2024, April 12). In Machine Learning, Can Good Predictive Models Also be Interpretable? pp 1-3. [In Machine Learning, Can Good Predictive Models also be Interpretable? | Statistical Horizons](https://statisticalhorizons.com/in-machine-learning-can-good-predictive-models-also-be-interpretable/)

Cycling Weekly (n.d.). <https://www.cyclingweekly.com/news>

Chonyy (2020, October 25). Apriori – Association Rule Mining In-Depth Explanation and Python Implementation. Towards Data Science. [Apriori — Association Rule Mining In-depth Explanation and Python Implementation | by Chonyy | Towards Data Science](https://towardsdatascience.com/apriori-association-rule-mining-explanation-and-python-implementation-290b42afdfc6)

Chonyy (2020, October 30). FP Growth – Frequent Pattern Generation in Data Mining with Python Implementation. Towards Data Science. [FP Growth — Frequent Pattern Generation in Data Mining with Python Implementation | by Chonyy | Towards Data Science](https://towardsdatascience.com/fp-growth-frequent-pattern-generation-in-data-mining-with-python-implementation-244e561ab1c3)

Kane, J.V. (2024, April 17). Correlation vs. Regression: a Key Difference that Many Analysts Miss. Medium. <https://medium.com/the-stata-gallery/correlation-vs-regression-a-key-difference-that-many-analysts-miss-3770c9b368d9>

Dabbura, I. (2018, September 7). K-means Clustering: Algorithm, Application, Evaluations Methods, and Drawbacks. Medium. [K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks | by Imad Dabbura | Towards Data Science](https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a)

Malik, F. (2023, December 26). Time Series Analysis in Python. Medium. <https://medium.com/fintechexplained/learn-time-series-analysis-in-python-97708c06aba2>

**Appendix A**

https://[www.linkedin.com/advice](http://www.linkedin.com/advice)

[**https://github.com/mrhanson1/Capstone**](https://github.com/mrhanson1/Capstone)

**Appendix B**

**Figure 16**

*Cluster showing Age vs Total Children*

A white background with many colored dots

Description automatically generated

*Note:* The scatter plot using k-means clustering shows the relationship between the age of individuals and the number of children the individuals have. Each point represents an individual in the AdventureWorks dataset. The features are selected and represented in order by the numbers 0 – 4. The features are Marital Status, Gender, Education Level, Occupation, and Homeowner. The range of -1 to 2 for the X axis and -1 to 5 for the Y axis indicates standardized values after normalization.

**Table 4**

*Occupation is the selection, comparing a few combination purchases*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Occupation as category - top few combos for each occupation** | | | | | |
| Professional | Antecendent | Consequent | Support | Confidence | Lift |
| 17,750 | HL Mountain Tire | Mountain Tire Tube | 0.044055 | 0.70935 | 5.722472 |
|  | Mountain Tire Tube | 30 oz water bottle | 0.070627 | 0.862903 | 6.953715 |
|  | Mountain Bottle Cage | 30 oz water bottle | 0.091601 | 0.847095 | 4.657481 |
|  | 30 oz watter bottle | 30 oz water bottle | 0.074341 | 0.949309 | 5.718553 |
|  | Road Bottle Cage | 30 oz water bottle | 0.061029 | 0.786517 | 4.82425 |
|  | 30 oz watter bottle | Road Bottle Cage | 0.054784 | 0.340125 | 5.465463 |
| Skilled Manual | Antecendent | Consequent | Support | Confidence | Lift |
| 13,137 | Mountain Bottle Cage | 30 oz water bottle | 0.068537 | 0.853556 | 5.195523 |
|  | 30 oz watter bottle | Mountain Bottle Cage | 0.068537 | 0.417178 | 5.195523 |
|  | Road Bottle Cage | 30 oz water bottle | 0.055434 | 0.861619 | 5.244598 |
|  | 30 oz watter bottle | Road Bottle Cage | 0.055434 | 0.337423 | 5.244598 |
| Management | Antecendent | Consequent | Support | Confidence | Lift |
| 9,791 | HL Mountain Tire | Mountain Tire Tube | 0.047956 | 0.7 | 5.493369 |
|  | Mountain Tire Tube | HL Mountain Tire | 0.047956 | 0.376344 | 5.493369 |
|  | Mountain Bottle Cage | 30 oz water bottle | 0.069879 | 0.838656 | 5.148894 |
|  | 30 oz watter bottle | Mountain Bottle Cage | 0.069879 | 0.429173 | 5.148894 |
|  | Road Bottle Cage | 30 oz water bottle | 0.052523 | 0.898438 | 5.517893 |
|  | 30 oz watter bottle | Road Bottle Cage | 0.052523 | 0.322581 | 5.517893 |
| Clerical | Antecendent | Consequent | Support | Confidence | Lift |
| 8,845 | Mountain Bottle Cage | 30 oz water bottle | 0.052141 | 0.799228 | 5.288224 |
|  | 30 oz watter bottle | Mountain Bottle Cage | 0.052141 | 0.345 | 5.288224 |
|  | Road Bottle Cage | 30 oz water bottle | 0.074811 | 0.894578 | 5.919127 |
|  | 30 oz watter bottle | Road Bottle Cage | 0.074811 | 0.495 | 5.919127 |
|  | Touring Tire | Touring Tire Tube | 0.041562 | 0.901639 | 11.62178 |
|  | Touring Tire Tube | Touring Tire | 0.041562 | 0.535714 | 11.62178 |
| Manual | Antecendent | Consequent | Support | Confidence | Lift |
| 6,507 | Road Bottle Cage | 30 oz water bottle | 0.06478 | 0.931373 | 6.566624 |
|  | 30 oz watter bottle | Road Bottle Cage | 0.06478 | 0.456731 | 6.566624 |
|  | Tourting Tire | Touring Tire Tube | 0.051483 | 0.848315 | 8.266135 |
|  | Touring Tire Tube | Touring Tire | 0.051483 | 0.501661 | 8.266135 |