Cairo University

Faculty of Computers & AI

Predictive Analysis & Forecasting

Adventure Works Dataset: Analysis, Regression & Forecasting

Mohamed Ayman Matbouly (20206063)
Abdelrahman Aggour (20207014)
Ahmed Waleed (20207002)
Assem Ihab (20206122)
Hussein Hazem (20207004)

Table of contents:

Figures & Illustrations:

Section 1: Introduction

Brief description of data:

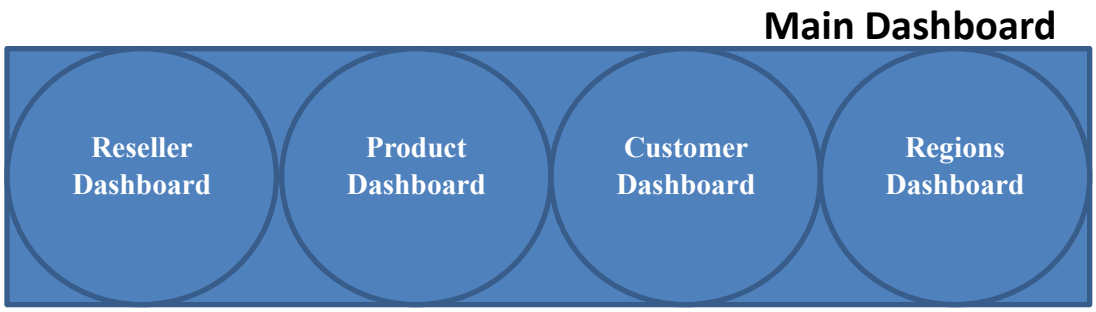
Adventure Works is a company that produces and sells bicycles, components, and accessories. They have physical stores and an online store, sell to individuals, retailers, and wholesalers, and prioritize quality and customer service. Note: our data is from 2017 until 2021.

Objectives:

To present and visualize Adventure Works' performance to top managerial levels. Pivot tables and data models were used to create the dashboard.

Methodologies used:

Many graph types like maps, bar charts, and cross tabulation etc. were used, depending on which was more appropriate. Five dashboards were created:



Obstacles faced:

Some of the Data was not cleaned, like the discount column, so we had to calculate it manually by subtracting the sales amount from the extended amount. Additionally, due to the intricate nature of the data, we encountered challenges in identifying all the relevant KPIs (key performance indicators) that could effectively contribute to informed decision-making.

Section 2: Theme

In this section, we introduce to you the five dashboards created for this project. Each dashboard focuses on a specific dimension of the dataset, enabling in-depth analysis and exploration of KPIs within that dimension. The five dashboards developed are as follows:

1.Dashboard for the Product Dimension:

Description: The product dimension dashboard aims to provide insights into the performance and characteristics of different products within the data set. It includes visualizations such as product sales, Demand, profitability. This dashboard allows us to identify top-selling products, analyse product performance over time, and make data-driven decisions related to product strategies.

2.Dashboard for the Region Dimension:

Description: The region dimension dashboard focuses on analysing the data set based on geographical regions. It provides visualizations of sales, revenue, customer distribution. This dashboard helps in understanding regional sales performance, identifying regions with growth potential, and optimizing distribution strategies based on regional trends and patterns.

3.Dashboard for the Reseller Dimension:

Description: The reseller dimension dashboard is designed to explore the performance and relationship with various resellers within the dataset. It includes visualizations of reseller sales, revenue contribution, customer satisfaction, and profitability. This dashboard allows us to evaluate the performance of different resellers, identify top-performing resellers, and develop strategies to improve reseller partnerships and engagement.

4.Dashboard for the Customer Dimension:

Description: The customer dimension dashboard focuses on analyzing customer-related metrics and behaviors. It includes visualizations of customer segmentation, lifetime value. This dashboard enables us to understand customer preferences, target high-value segments, and enhance customer experience based on insights derived from the data.

5.Dashboard for Overall Performance:

Description: The overall performance dashboard provides a comprehensive view of the data set, combining key metrics from different dimensions. It includes visualizations such as overall sales, revenue trends, profitability analysis, and market performance. This dashboard serves as a summary of the entire data set, allowing us to assess the overall health and performance of the business, identify areas of improvement, and make strategic decisions.

These five dashboards provide a comprehensive analysis of the dataset from various dimensions, enabling a holistic understanding of the business and facilitating data-driven decision-making. Through these dashboards, we gain insights into product performance, regional dynamics, reseller relationships, customer behavior, and overall business performance.

Section 2.1: Product Dimension

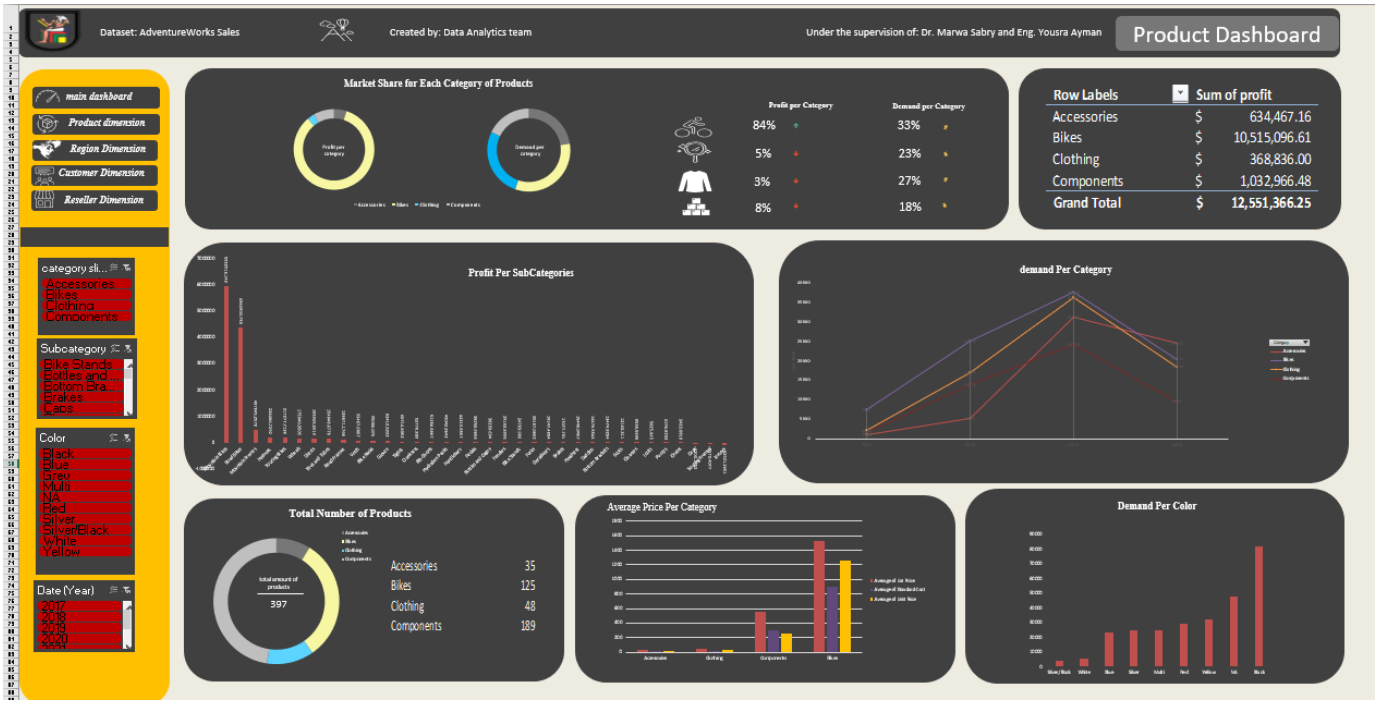


Figure 1.1

The product dimension dashboard provides insights into the performance and characteristics of different products within the data set. It includes visualizations that help analyse demand, profitability, category distribution, subcategory profitability, price analysis, and colour preferences. The following KPIs are displayed in the dashboard:

1. Pie Charts Showing Demand and Profit per Category:

The dashboard includes two pie charts that showcase the distribution of demand and profit across different product categories. These visualizations provide a quick overview of the contribution of each category to the overall demand and profitability.

2. Percentage of Profit and Demand per Category using Conditional Formatting:

To enhance data visibility, the percentage of profit and demand per category is displayed using conditional formatting. This enables easy identification of categories that have a higher or lower percentage of profit and demand compared to others.

3. Table Showing Profit per Category in Dollars:

A table is included in the dashboard that displays the profit generated by each product category in dollars. This allows for a detailed analysis of the profitability of each category and helps identify categories that contribute significantly to the overall profit.

4. Bar Chart Showing Profit per Subcategory:

The dashboard includes a bar chart that visualizes the profitability of different product subcategories. This chart allows for a comparison of the profitability of various subcategories, aiding in identifying the most profitable subcategories and potential areas for improvement.

5. Line Chart Comparing Demand per Each Category Across Time:

A line chart is included to showcase the demand trends for each product category over time. This visualization helps identify seasonality or trends in demand for different categories, facilitating strategic planning and inventory management.

6. Pie Chart Showing the Number of Products per Category:

The dashboard features a pie chart that illustrates the distribution of products across different categories. This visualization provides insights into the relative size of each category in terms of the number of products available.

7. Cross Tabulation Showing Average Price per Category:

A cross tabulation table is used to present the average price per category. This allows for a comparison of the price ranges across different product categories and helps identify any pricing patterns or disparities.

8. Bar Chart Comparing Demand per Colour:

The dashboard includes a bar chart that compares the demand for products based on different colours. This visualization aids in understanding customer colour preferences and identifying which colours are more popular among buyers.

These visualizations and KPIs in the product dimension dashboard provide valuable insights into the demand, profitability, category distribution, subcategory performance, pricing analysis, and colour preferences within the data set. They enable informed decision-making and strategic planning related to product assortment, pricing strategies, and inventory management.

Section 2.2: Region Dimension

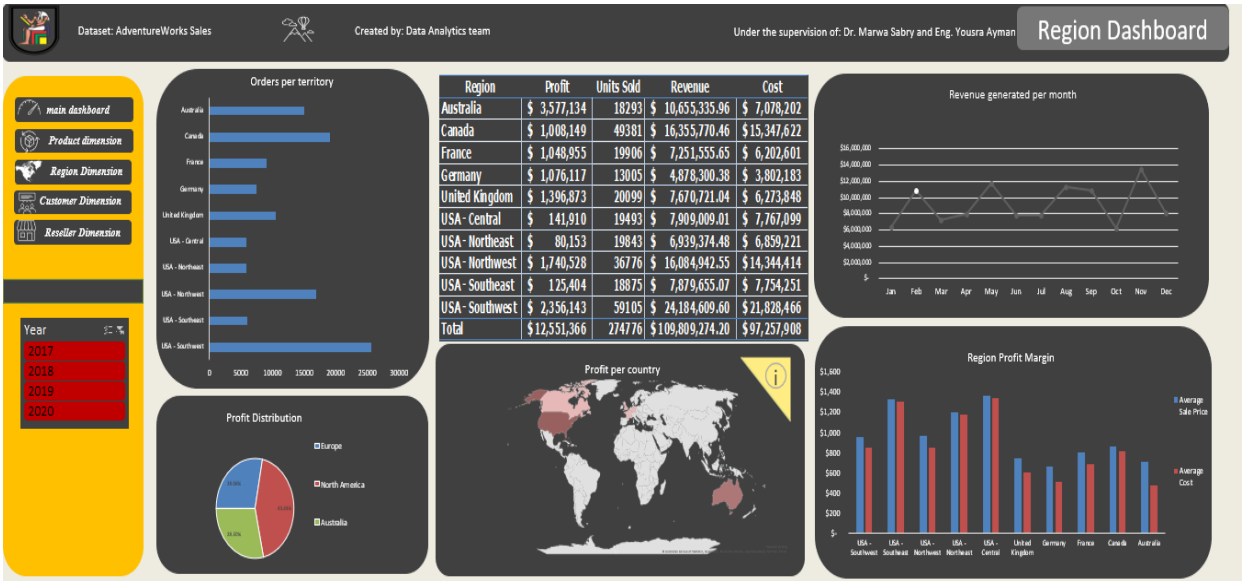


Figure 1.2

The region dashboard was centred around displaying the most (and least) successful regions and territories within the dataset. It mainly concerns the revenue and how profitable each region is throughout the years. Its main KPIs were shown in the charts displayed in the figure above.

The higher part contained the main data table, containing some important statistics about each territory, a bar chart to signify the territories who managed to deliver the most orders,

as well as a line chart, outlining a crucial KPI, used to determine the profit trend of each territory.

The lower part consisted of a pie chart, aiming to simply convey where the most profits are generated, supported by a heatmap to specify those places.

And lastly a multicolumn chart, used to determine how healthy the margins are for each territory, conveying the profit % smoothly.

Section 2.3: Region Dimension

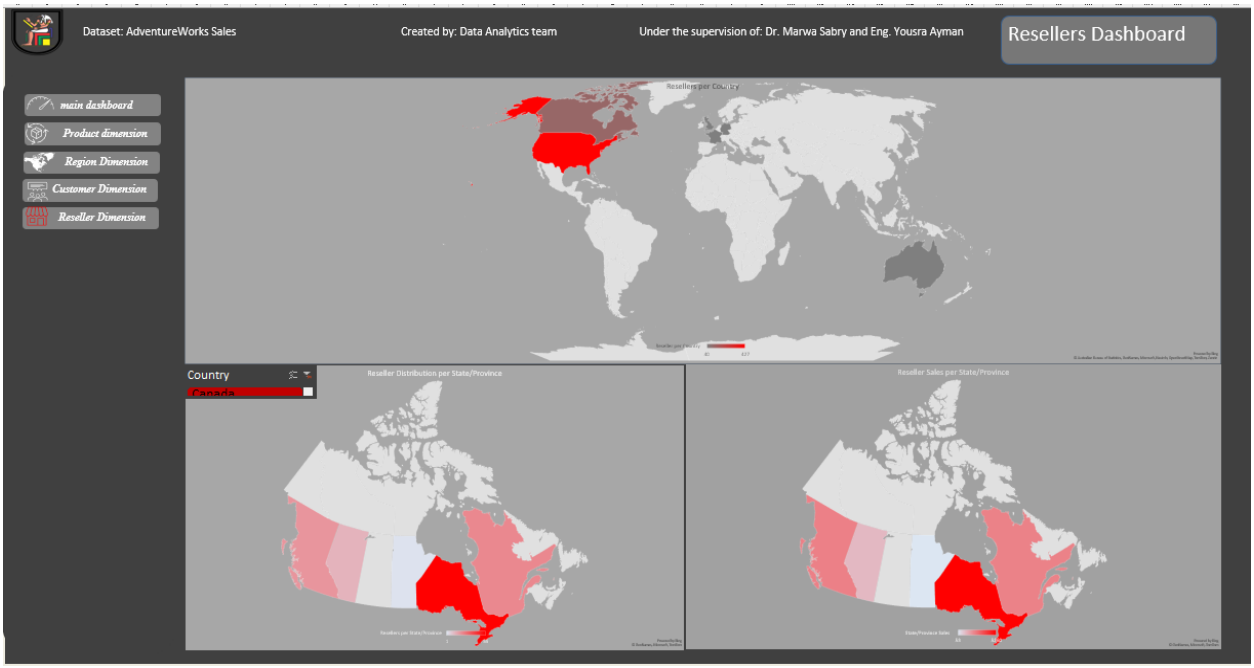


Figure 1.3

The reseller dashboard consists of three charts, one of which shows seller concentration around the world, while the remaining two are made to work in tandem.

The bottom two charts server to compare reseller concentration with reseller sales in a certain state/province, which might help with identifying poorly performing resellers and choosing the best course of action after further analysis.

Heatmaps were chosen as they were the perfect fit to deliver the KPI efficiently and concisely to the observer, as it tracks a geographical distribution of sales and resellers.

Section 2.4: Region Dimension

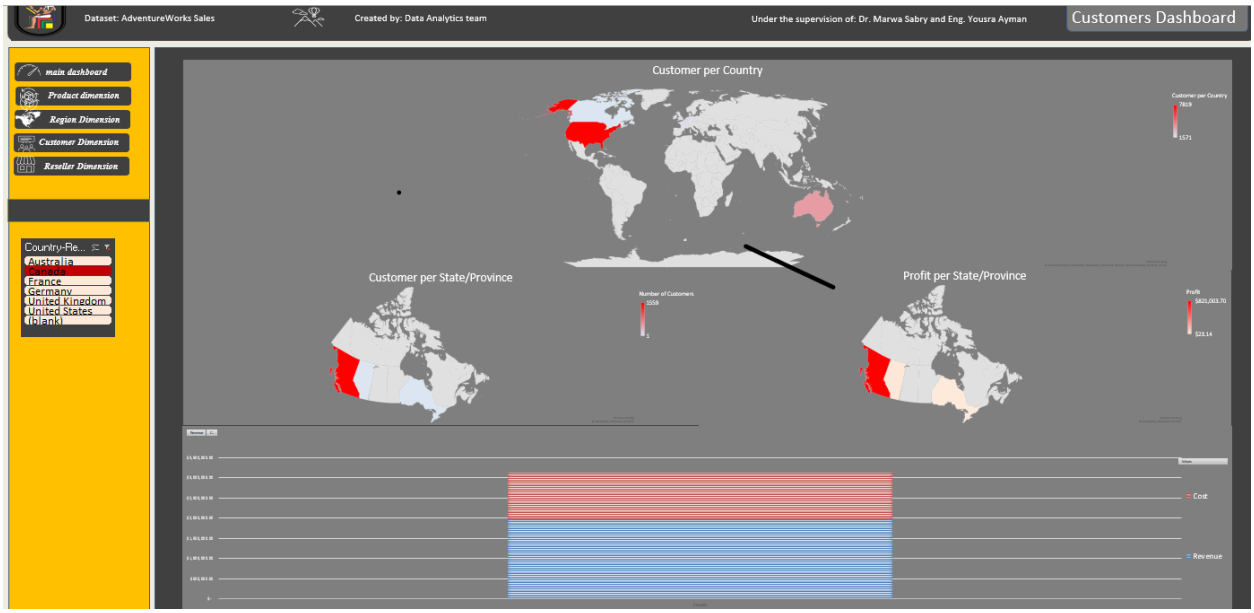


Figure 1.4

The customer dashboard consists of 3 primary heatmaps, and a supporting stacked column chart.

The first heatmap is a global heatmap showing the density of the customers across the various countries.

The two following heatmaps aim to find if there is discrepancy between the concentration of customers across a certain region and the profits in that region.

The supporting column chart simply shows the losses incurred in comparison to the revenue generated in the chosen region.

Section 2.5: Main Dashboard

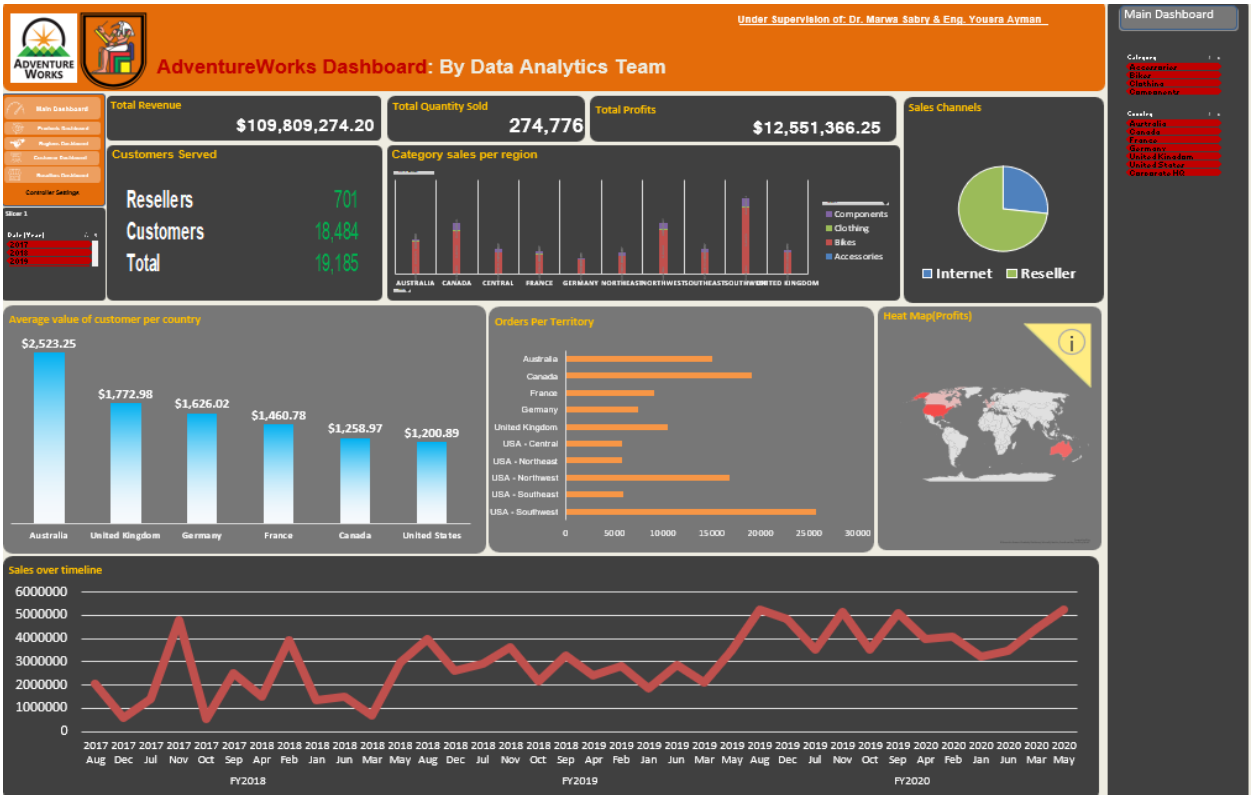


Figure 1.5

The main dashboard is the heart of the project. It contains the indispensable KPIs and conveys them in a variety of ways.

On the absolute top next to the navigation bar are some of the crucial numbers. These numbers represent the company lifetime revenue, the quantity of units sold, and the profits generated from those sales.

Next to them are a stacked column chart for each region and their respective sales per each category, and a pie chart highlighting both channels' share of profit.

Underneath them are a variety of KPIs. The first of which showcases the average value of each customer in a country (how much a customer pays in average for each country).

The second is a bar chart representing the orders per each territory, while to the right is a heatmap for the profits signifying the hottest (most profitable) countries.

Last but certainly not least, the major line chart represents the total sales over the dataset timeline, highlighting the rises and the falls all the time.

Best Fitted Regression Model

ProductKey	Category	Demand	profit	DemandLog	Standard Cost2	List Price2	Average Unit Price
210	Components	0	0	0	$ 868.63	$ 1,431.50	0
211	Components	0	0	0	$ 868.63	$ 1,431.50	0
212	Accessories	564	4601.5068	2.752048448	$ 12.03	$ 33.64	20.1865
213	Accessories	1493	8311.1981	3.174350597	$ 13.88	$ 33.64	19.79699306
214	Accessories	4209	62275.6219	3.624282096	$ 13.09	$ 34.99	33.0905039
215	Accessories	600	4881.3988	2.778874472	$ 12.03	$ 33.64	20.18315224
216	Accessories	1635	9175.9729	3.213783299	$ 13.88	$ 33.64	19.82235235
217	Accessories	4297	60672.7791	3.633266411	$ 13.09	$ 34.99	32.75154809
218	Clothing	1107	2300.6843	3.04453976	$ 3.40	$ 9.50	5.656037234
219	Clothing	90	207.333	1.959041392	$ 3.40	$ 9.50	5.7
220	Accessories	661	5381.2058	2.820857989	$ 12.03	$ 33.64	20.18337023
221	Accessories	1739	9661.0737	3.240549248	$ 13.88	$ 33.64	19.83830188
222	Accessories	4343	61445.9721	3.637889817	$ 13.09	$ 34.99	32.8040686
223	Clothing	985	-538.5748	2.993876915	$ 5.71	$ 8.64	5.182028621
224	Clothing	2304	-350.432	3.36267093	$ 5.23	$ 8.64	5.163226697
225	Clothing	5022	-314.1885	3.700963178	$ 6.92	$ 8.99	8.358121736
226	Clothing	0	0	0	$ 31.72	$ 48.07	0
227	Clothing	0	0	0	$ 29.08	$ 48.07	0
228	Clothing	429	4932.5133	2.633468456	$ 38.49	$ 49.99	49.99
229	Clothing	437	-1260.308	2.641474111	$ 31.72	$ 48.07	28.8404
230	Clothing	1238	-501.0348	3.093071306	$ 29.08	$ 48.07	28.80355192
231	Clothing	1961	-8338.3119	3.292699003	$ 38.49	$ 49.99	41.69882988
232	Clothing	991	-3039.3858	2.996511672	$ 31.72	$ 48.07	28.81441689
233	Clothing	2346	-2057.1996	3.37051309	$ 29.08	$ 48.07	28.7036906

Table 2.1

As illustrated in Table 6.1 and based on our expertise, these are the variables that we have deemed the best fit out of the available dimensions to predict the profit of a certain product. They are: Category, Demand Per each product, Profit, the Log of the Demand variable (we added this variable to visualize the data only as it was a little bit skewed to the right), the cost of producing this product, the List price assigned to

this product and the average unit price, which is calculated from the sales fact table by averaging the selling price of this products among all invoices).

We first need to build a correlation matrix between all variables to see whether there is multicollinearity. Below table 6.2 shows that there is multicollinearity, as expected, between the standard cost and the list price and the average unit price.

	Demand	*profit*	*DemandLog*	*Standard Cost2*	*List Price2*	*Average Unit Cost*
Demand	1					
profit	0.12134427	1				
DemandLog	0.603181923	0.20289703	1			
Standard Cost2	-0.134385578	0.549409193	0.05328675	1		
List Price2	-0.139299939	0.567047569	0.04683288	0.996852984	1	
Average Unit Cost	-0.040343802	0.676342414	0.255156137	0.939856443	0.936810331	1

Table 2.2

Based on the correlation matrix above we decided to only conclude Average Unit cost column in our model, since it has the highest correlation among all the multicollinear variable with the profit. Also, we Included the Demand of the variable. In addition to this, we decided to one hot encode the category column since we concluded from that dashboard that, below there is statistics for the regression model we ran on the variables we just mentioned.

Regression Statistics		Standard deviation of Profit	89577.69952
Multiple R	0.70111528		
R Square	0.491562636		
Adjusted R Square	0.485060879		
Standard Error	64280.29311		
Observations	397		

Table 2.3

From these statistics illustrated in Table 6.3 we were able to decide that the regression model explains some of variability that happens in real life to a certain extent (49%) and since the standard Error of the residuals is less than the standard deviation of the profit variable so we could say that resulted numbers from this model are credible. Table 6.4

features some Statistics that could help us in evaluating the performance of the regression model such as the ANOVA table and the Individual variable statistics table.

ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	5	1.56197E+12	3.12395E+11	75.6045894	2.67578E-55			
Residual	391	1.61559E+12	4131956083					
Total	396	3.17757E+12						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-7759.040493	13771.34093	-0.563419389	0.5734723	-34834.18092	19316.09993	-34834.18092	19316.09993
Componenets	-15718.88978	13963.0647	-1.12574783	0.260962777	-43170.96866	11733.1891	-43170.96866	11733.1891
Clothing	-8132.413971	14332.52551	-0.567409698	0.57076149	-36310.87115	20046.04321	-36310.87115	20046.04321
Bikes	-40188.64958	15417.21286	-2.606738971	0.00949045	-70499.65597	-9877.643183	-70499.65597	-9877.643183
Demand	13.22033106	4.783965949	2.763466798	0.005989436	3.814816342	22.62584578	3.814816342	22.62584578
Average Unit Price	107.6264933	6.884108389	15.63404979	3.82797E-43	94.09199424	121.1609924	94.09199424	121.1609924

Table 2.4

As illustrated in Figure 2.1a and 2.2b the data is heteroscedastic so we could apply transform it to make the regression model based on valid assumptions, but since this is out of our course's scope we will work with our model as it is, and as illustrated in Figure 2.3c that the Errors are normally distributed, so the second assumption holds true.

Fig. 2.1a

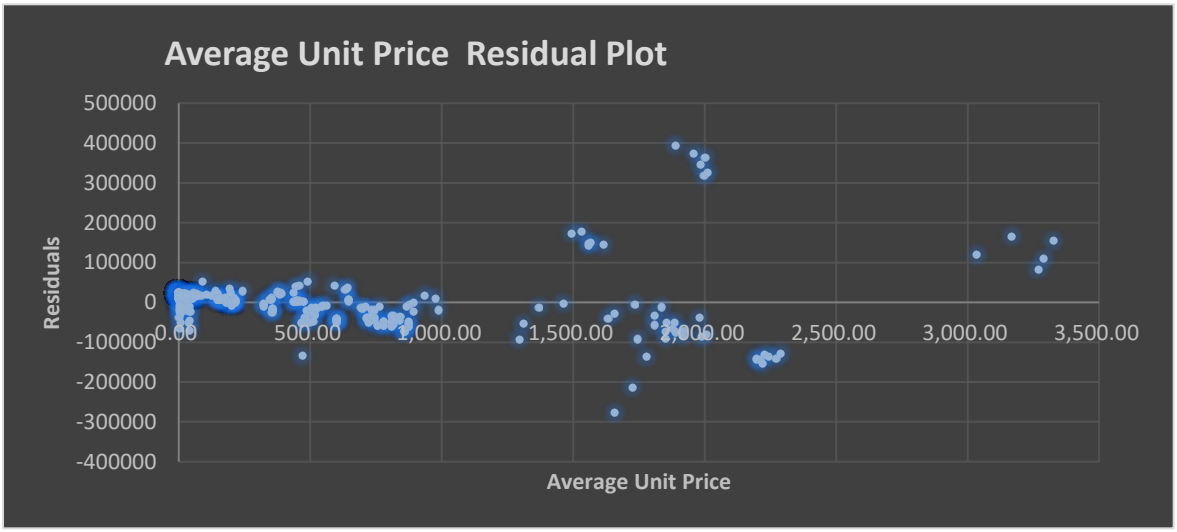


Fig. 2.1b

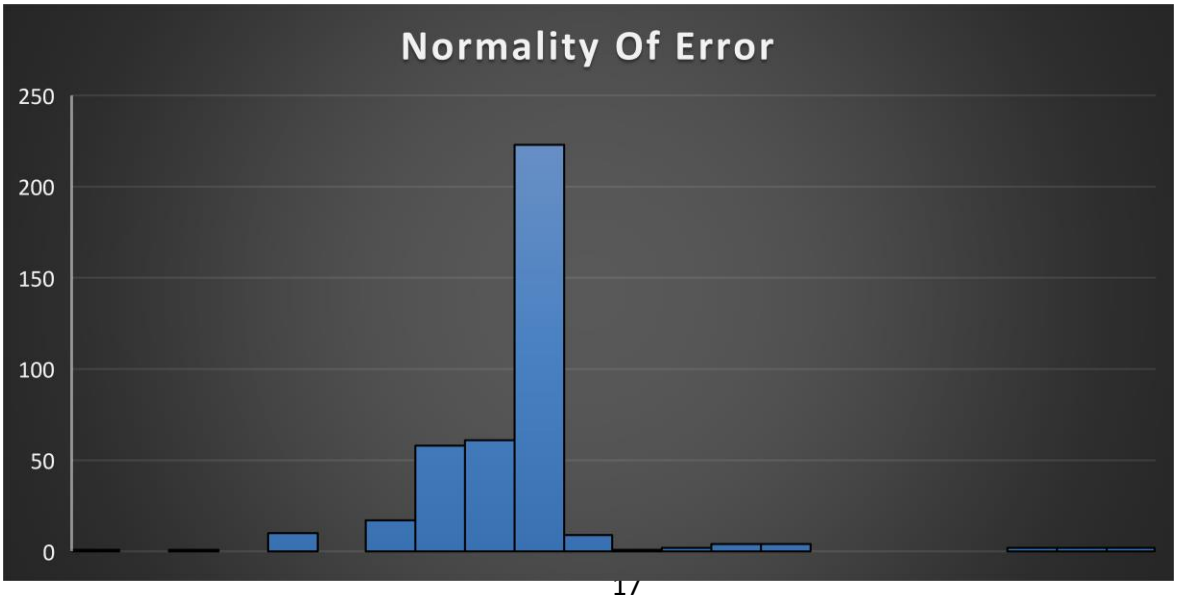


Fig. 2.1c

Time Series Analysis

Due to the nature of our data, it was deemed suitable to do time-series decomposition and forecasting, which enables us to forecast future trends, as well as spikes and dips. Through it very interesting insights were of note, the most important of them being the presence of seasonality in company sales, as illustrated in figure 3.1:

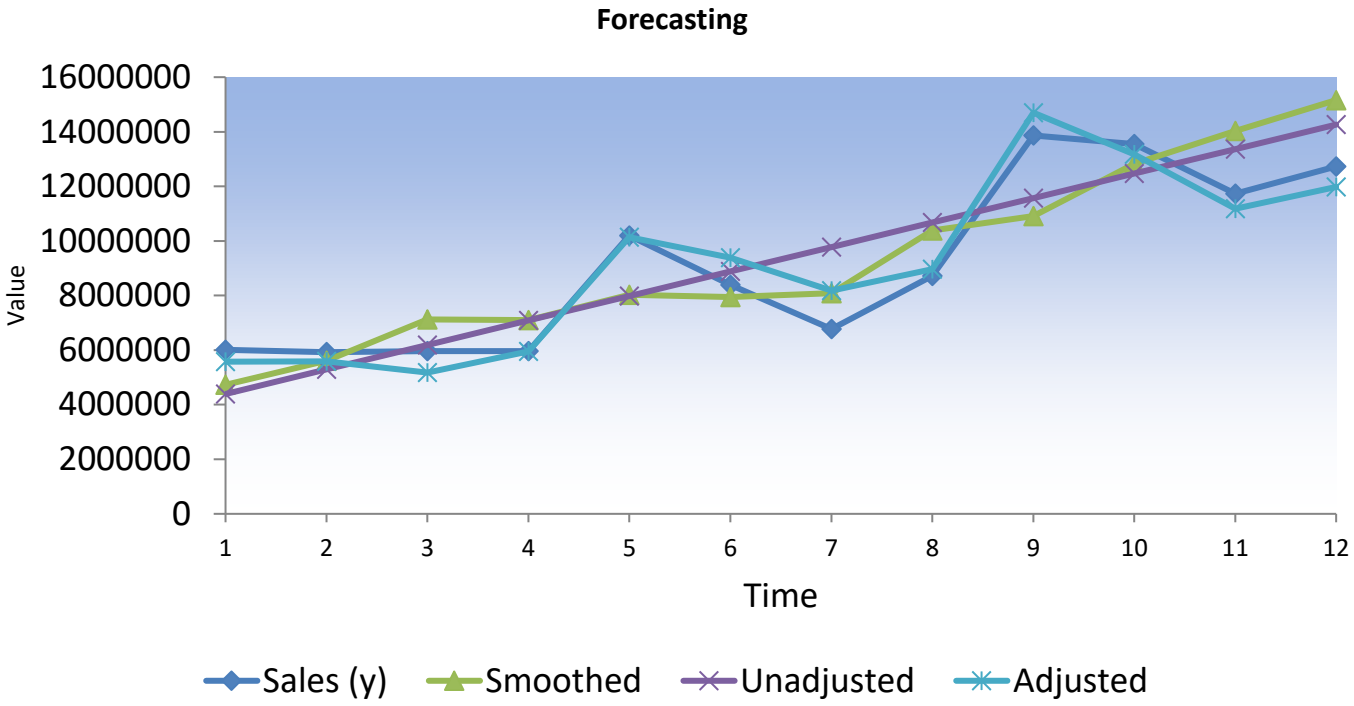


Fig. 3.1

Forecasts

Period	Unadjusted	Seasonal	Adjusted
13	15159064	1.270335	19257088
14	16056359	1.056595	16965064
15	16953653.9	0.836719	14185448
16	17850948.9	0.83993	14993544

Table 3.1

As we can see from the forecast produced in table 3.1, sales are expected to rise sharply in the upcoming quarter due to a seasonal spike, then gradually die down.

It is noticed that the company sales follow an upwards trend, most likely due to an increase in

company exposure, as well as logistical capabilities and facilities. The most interesting feature here would be the recurring seasonal spikes, as it indicates periods of increased activity and sales, and consequently, periods of time where it would be recommended to increase production as well as other courses of actions as mentioned below.

Some recommendations inferred from the time series would be to increase bike production, as well as bike accessories, components and sporting gear nearing the end of the 2nd fiscal quarter and during the 3rd fiscal quarter due to increase of sales during the seasons, which can be attributed to consumers getting ready for outdoors activities during the summer.

There can also be placing special attention on certain types of bikes and accessories in these seasonal peaks in certain regions, such as mountain bikes in countries with many mountains and arid environments.

Such analysis allows the company to prepare early for such seasonal spikes, whether it would be increasing production early to not meet shortages in demand later or start preparing marketing strategies centered around holidays in the seasonal spike. It also prepares the company for potential dips in sales, allowing for time to prepare events such as clearance sales and the like.

We went with a simple decomposition model as it produced the least MSE and MAPE as compared with exponential smoothing and moving average:

					Total	-176776	6781523	5.65325E+12	81.66%
Average		Intercept	3494229.436			-14731.4	565126.9	4.71105E+11	06.80%
		Slope	897294.9655			Bias	MAD	MSE	MAPE
							SE	898670.0396	

Table 3.2

Total	13156463	17228136	7.03E+13	153.16%
Average	**1461829**	**1914237**	**7.81E+12**	**17.02%**
	Bias	**MAD**	**MSE**	**MAPE**
		SE	**3169044**	

Table 3.3

Total	37746786	38066663	2.27893E+14	328.63%
Average	**3145566**	**3172222**	**1.8991E+13**	**27.39%**
	Bias	**MAD**	**MSE**	**MAPE**
		SE	**4773809.633**	

Table 3.4

What-IF Analysis

One of the main concerns when it comes to running a business is budget balancing, you can have a functioning business without money after all. In Adventure Works' case, since the company sells many types of products, it must balance many variables, such as the unit price for an item, its cost, trying to drive sales for it through marketing, and so forth.

What-If analysis aims to provide a method for the above, by producing scenarios or searching for a goal, the company can optimize the many variables for the many products it sells, helping it assure profit or at least break-even.

In section 3, the team had produced a regression model using some variables, which is perfect for running Excel's what-if analysis tool, we can produce multiple scenarios to find the minimum parameters needed to turn a profit for each of the categories included in the model.

We can also start a goal search for a specific profit margin by changing certain parameters in the model till we find the perfect combination of goals for a certain profit margin.

Scenario Summary						
	Current Values:	Original	Component	Clothing	Bikes	IncDemand,DecAvgUnitPrice
Changing Cells:						
R4	0	0	1	0	0	0
S4	0	0	0	1	0	0
T4	0	0	0	0	1	0
U4	564	564	4000	2000	200	7000
V4	20.19	20.19	50.00	28.00	1,000.00	14.00
Result Cells:						
M13	1869.828433	1869.828433	34784.71863	13562.74947	62322.86947	86290.04784

Notes: Current Values column represents values of changing cells at time Scenario Summary Report was created. Changing cells for each scenario are highlighted in gray.

Fig. 4.1

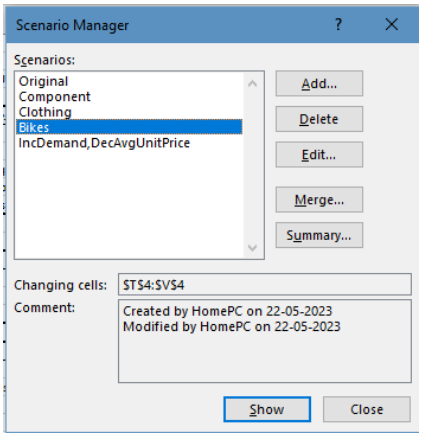
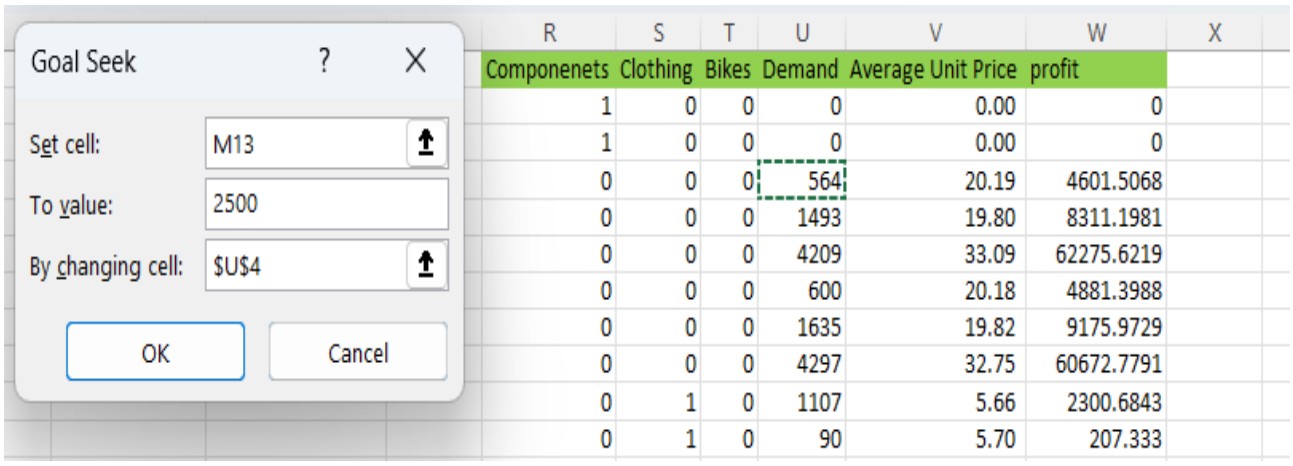
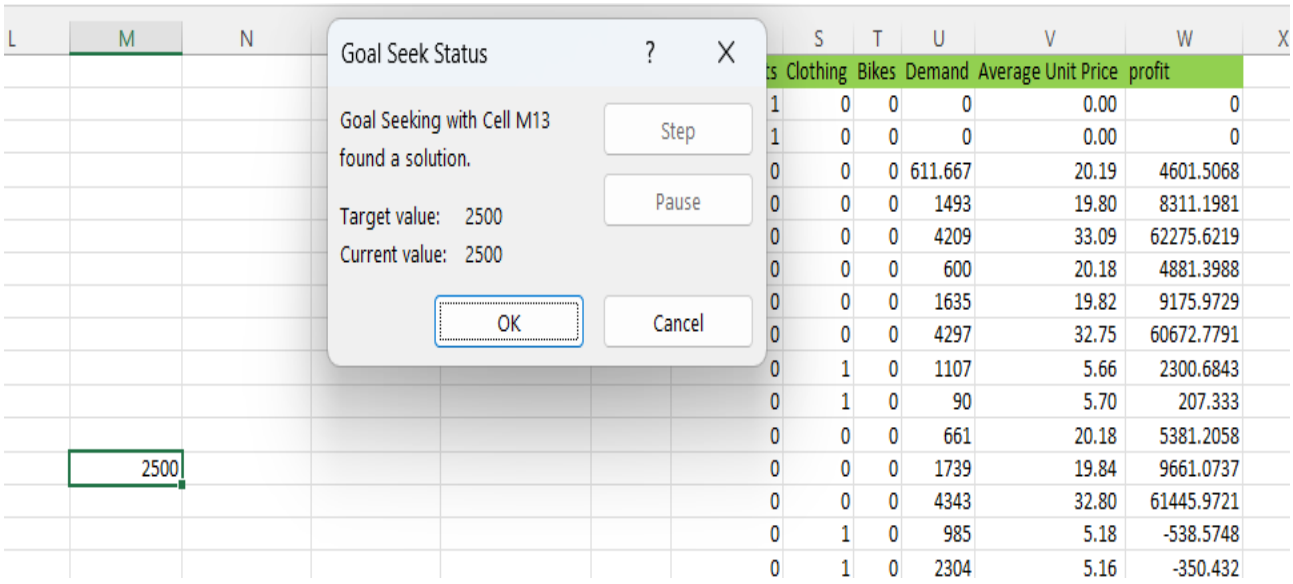


Fig. 4.2



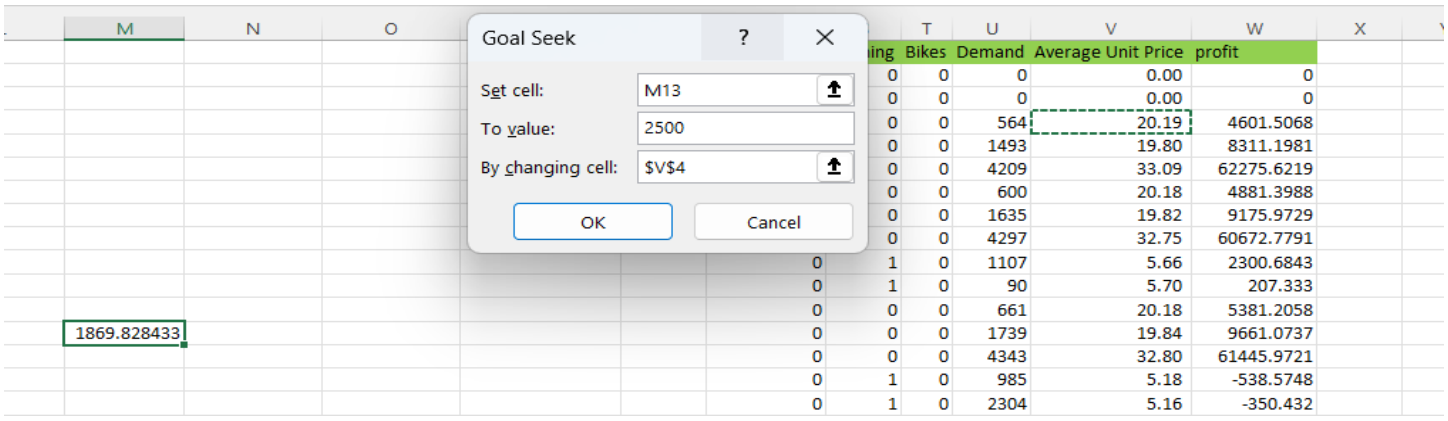
R	S	T	U	V	W	X
Componenets	Clothing	Bikes	Demand	Average Unit Price	profit	
1	0	0	0	0.00	0	
1	0	0	0	0.00	0	
0	0	0	564	20.19	4601.5068	
0	0	0	1493	19.80	8311.1981	
0	0	0	4209	33.09	62275.6219	
0	0	0	600	20.18	4881.3988	
0	0	0	1635	19.82	9175.9729	
0	0	0	4297	32.75	60672.7791	
0	1	0	1107	5.66	2300.6843	
0	1	0	90	5.70	207.333	

Fig. 4.3a



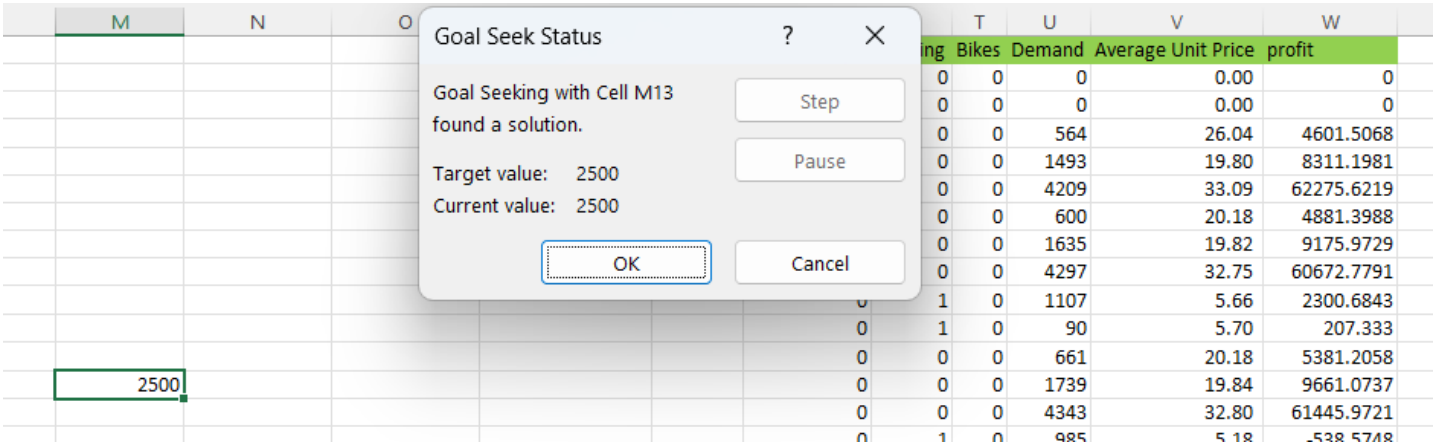
L	M	N	S	T	U	V	W	X
			Clothing	Bikes	Demand	Average Unit Price	profit	
			0	0	0	0.00	0	
			0	0	0	0.00	0	
			0	0	611.667	20.19	4601.5068	
			0	0	1493	19.80	8311.1981	
			0	0	4209	33.09	62275.6219	
			0	0	600	20.18	4881.3988	
			0	0	1635	19.82	9175.9729	
			0	0	4297	32.75	60672.7791	
			1	0	1107	5.66	2300.6843	
			1	0	90	5.70	207.333	
			0	0	661	20.18	5381.2058	
	2500		0	0	1739	19.84	9661.0737	
			0	0	4343	32.80	61445.9721	
			0	1	985	5.18	-538.5748	
			0	1	2304	5.16	-350.432	

Fig. 4.3b



ing	Bikes	Demand	Average Unit Price	profit
0	0	0	0.00	0
0	0	0	0.00	0
0	0	564	20.19	4601.5068
0	0	1493	19.80	8311.1981
0	0	4209	33.09	62275.6219
0	0	600	20.18	4881.3988
0	0	1635	19.82	9175.9729
0	0	4297	32.75	60672.7791
0	1	0	1107	5.66	2300.6843
0	1	0	90	5.70	207.333
0	0	0	661	20.18	5381.2058
0	0	0	1739	19.84	9661.0737
0	0	0	4343	32.80	61445.9721
0	1	0	985	5.18	-538.5748
0	1	0	2304	5.16	-350.432

Fig. 4.4a



ing	Bikes	Demand	Average Unit Price	profit
0	0	0	0.00	0
0	0	0	0.00	0
0	0	564	26.04	4601.5068
0	0	1493	19.80	8311.1981
0	0	4209	33.09	62275.6219
0	0	600	20.18	4881.3988
0	0	1635	19.82	9175.9729
0	0	4297	32.75	60672.7791
0	1	0	1107	5.66	2300.6843
0	1	0	90	5.70	207.333
0	0	0	661	20.18	5381.2058
0	0	0	1739	19.84	9661.0737
0	0	0	4343	32.80	61445.9721
0	1	0	985	5.18	-538.5748

Fig. 4.4b

\

Thank you for your assistance & generous supervision.

Best Wishes

Data Science Team

Mohamed Ayman Matbouly (20206063)
Abdelrahman Aggour (20207014)
Ahmed Waleed (20207002)
Assem Ihab (20206122)
Hussein Hazem (20207004)