Table of Contents

[Data Analysis 1](#_Toc102104033)

[Description of the Data 1](#_Toc102104034)

[Research Questions 4](#_Toc102104035)

[Descriptive Statistics 5](#_Toc102104036)

[Data Visualisation – Histogram 10](#_Toc102104037)

[Predictive Analysis 14](#_Toc102104038)

[Multiple Linear Regression 14](#_Toc102104039)

[Decision Tree Algorithm 14](#_Toc102104040)

[Interpretation of Results 15](#_Toc102104041)

[Recommendation for Performance Improvement 16](#_Toc102104042)

[References 18](#_Toc102104043)

[Appendix 18](#_Toc102104044)

# Data Analysis

The focus of this section is the analysis of the data obtained on the financial activities of Adventure Works. Adventure Works Cycle is a multinational company dealing in the manufacturing of bicycles and their components. The data is made available on Kaggle, so it is from there that we have downloaded it for the purpose of this analysis.

## Description of the Data

The Adventure works data is a database consisting of several datasets. For this reason, the AdventureWorks\_Sales and AdventureWorks\_Territories datasets in the database were joined in RStudio to form a large single dataset which was used for the analysis. To achieve this, the merge function was used to perform a left join on the datasets. The new dataset obtained from the join performed was named AdventureWorksData. The dataset consists of 30 columns and 60398 rows. The columns in the dataset were described below:

|  |  |  |  |
| --- | --- | --- | --- |
| **S/N** | **Column Name** | **Column Description** | **Data type** |
| 1 | ProductKey | ID of the product | Integer |
| 2 | OrderDateKey | ID of the order date | Integer |
| 3 | DueDateKey | ID of the due date | Integer |
| 4 | ShipDateKey | ID of the ship date | Integer |
| 5 | CustomerKey | ID of the customer | Integer |
| 6 | PromotionKey | ID of the promotion | Integer |
| 7 | CurrencyKey | ID of the currency | Integer |
| 8 | SalesTerritoryKey | ID of the sales territory | Integer |
| 9 | SalesOrderNumber | Number of sales order | Integer |
| 10 | SalesOrderLineNumber | Number of sales order line | Integer |
| 11 | RevisionNumber | Numbr of revison | Integer |
| 12 | OrderQuantity | Quantity of order | Integer |
| 13 | UnitPrice | The unit price of an item sold | Number |
| 14 | ExtendedAmount | The extended amount of the product sold | Number |
| 15 | UnitPriceDiscountPct | Percentage of unit price discount | Integer |
| 16 | DiscountAmount | Amount of discount | Integer |
| 17 | ProductStandardCost | Standard cost of product | Number |
| 18 | TotalProductCost | Cost of total product | Number |
| 19 | SalesAmount | Amount of sales | Number |
| 20 | TaxAmt | Amount of tax | Number |
| 21 | Freight | Cost of freight | Number |
| 22 | CarrierTrackingNumber | Number of carrier tracking | Character |
| 23 | CustomerPONumber | Ciustomer purchase order number | Character |
| 24 | OrderDate | Date of order | Character |
| 25 | DueDate | The due date | Character |
| 26 | ShipDate | The date the item was shipped | Character |
| 27 | SalesTerritoryAlternateKey | The alternate key of sales territory | Integer |
| 28 | SalesTerritoryRegion | The region of sales territory | Integer |
| 29 | SalesTerritoryCountry | Country of sales territory | Character |
| 30 | SalesTerritoryGroup | Group of sales territory | Character |

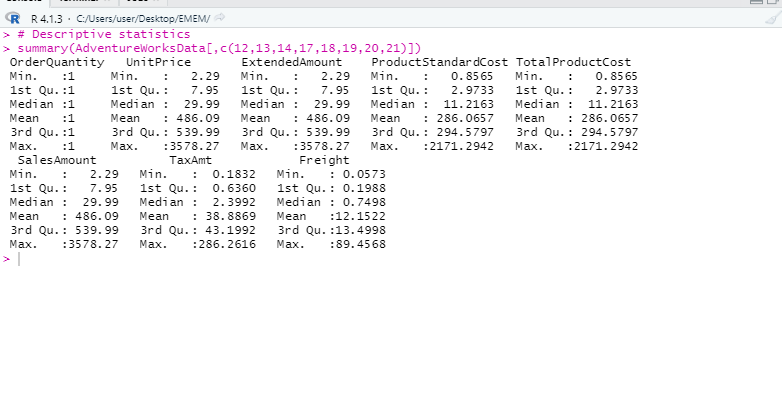
The data type of each column was obtained by using the str() function in RStudio.

## Research Questions

Asking the right questions gives an analysis a direction. The research questions for this analysis were presented below:

* What is the minimum and maximum unit price of the items sold?
* What is the average sales amount?
* What is the average amount spent on freight?
* What is the total revenue made in each territory region?
* In which country was most quantity of items sold?
* What is the total product cost incurred in each sales territory country?
* What impact do product standard cost, total product cost, tax amount and freight have on sales amount?
* How can Adventure Works Cycle boost its revenue?
* How can Adventure Works Cycle boost make more profit?

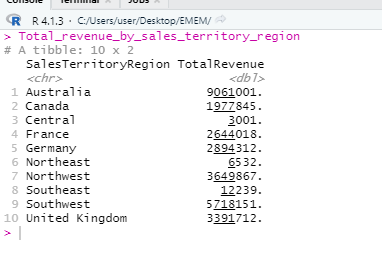
## Descriptive Statistics



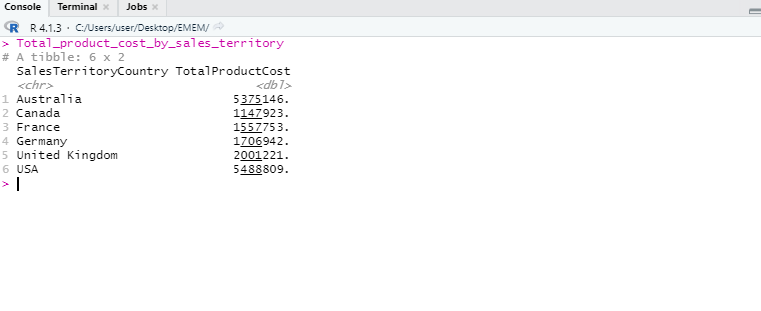
**Figure 1: Descriptive Statistics of Key Variables**

The descriptive statistics results were discussed in this section. Only key variables that are numeric were included in the statistics. The results of the descriptive statistics were presented in Figure 1. The result revealed for instance that the average tax amount paid was 38.89. It also revealed that the maximum freight paid was 89.46.

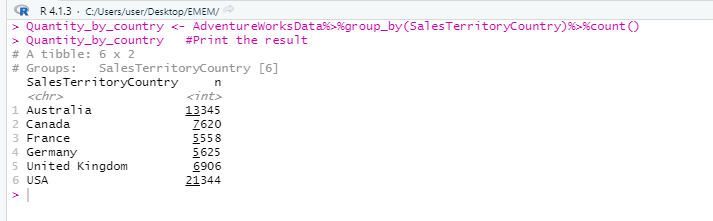
Also, the descriptive statistics enabled us to answer a few of our research questions. The minimum unit price of the items sold was $2.29, and the maximum unit price of the item was 3578.27. The average sales amount was $486.09. The average amount spent on freight was $12.15.



**Figure 2: Total Revenue Made in Each Territory**

****

**Figure 3: Total Product Cost Incurred in Each Country**

****

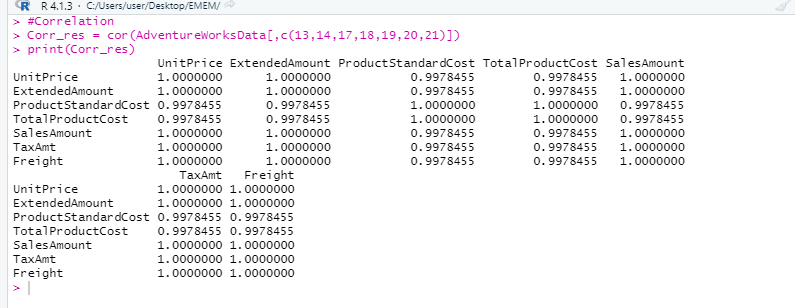
**Figure 4: Quantity Sold in Each Country**

As regards the total revenue made in each sales territory region, it was observed that $90,610,001 was made in Australia, $1,977,845 was made as revenue in Canada, $3,001 was made in Central Region, $2,644,018 was made in France, $2,89,4312 was made in Germany, $6,532 was made in Northeast Region; $3,649,867 was made in Northwest Region, $12,239 was made in Southeast Region, $5,718,151 was made in Southwest Region, and $3,391,712 was made in the United Kingdom.

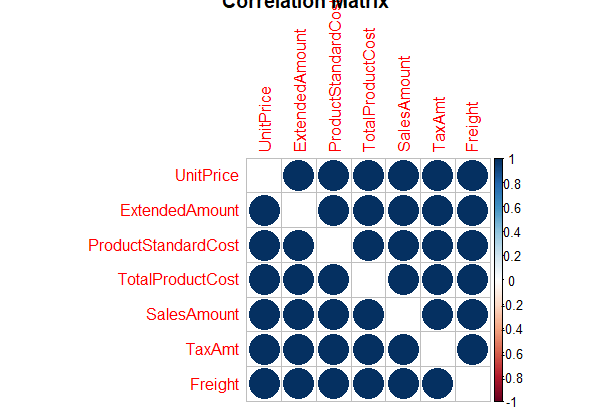
Concerning the total product cost incurred in each country, $5,375,146 was incurred in Australia; $1,147,923 was made in Canada, $1,557,753 was incurred in France, $1706942 was incurred in Germany, $2,001,221 was incurred in the United Kingdom and 5,488,809 was incurred in the USA.

From Figure 4, it was found that most quantity of items was sold in Australia. This explains why most of the revenue is coming from the country.

**Correlation Analysis**



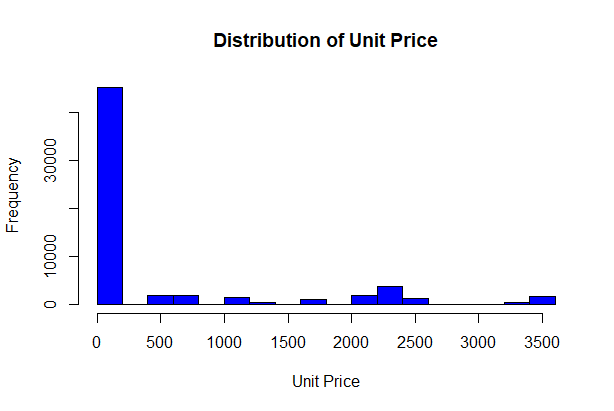
**Figure 5: Correlation Analysis Result**

****

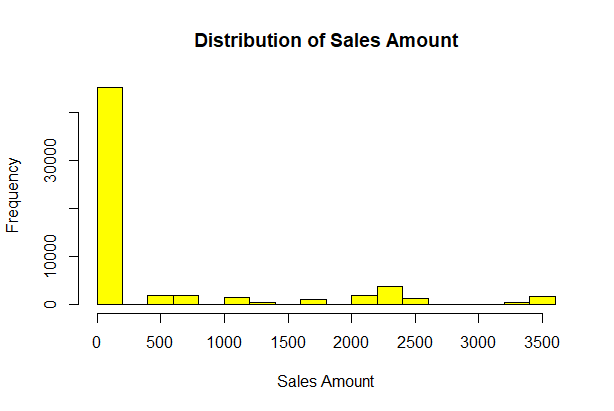
**Figure 6: Correlation Matrix**

The correlation among the variables was conducted with the result presented in Figure 2. Also, it was presented visually with a figure made possible by the corrplot() function. From the result of the correlation analysis, it was observed that there is a high correlation among the variables of study.

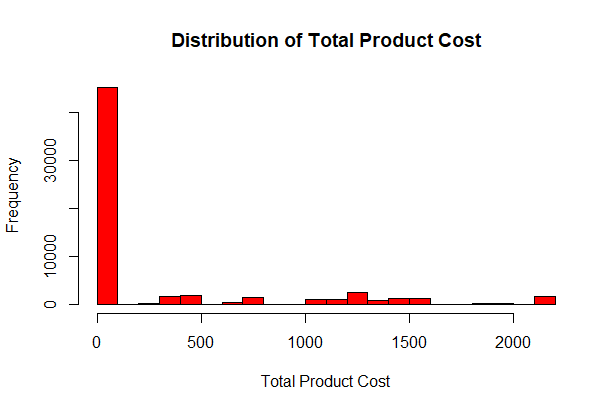
## Data Visualisation – Histogram

****

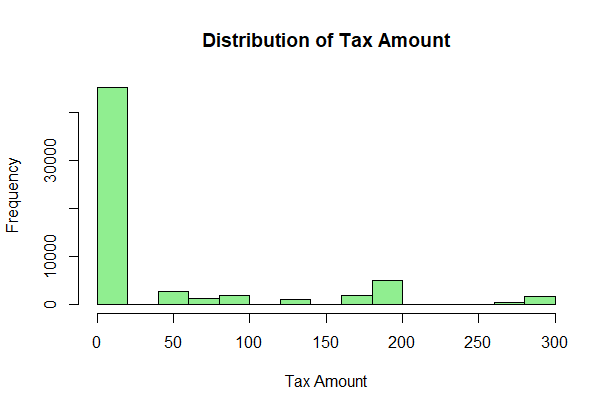
**Figure 7: Histogram showing Distribution of Unit Price**

****

**Figure8: Histogram showing Distribution of Sales Amount**

****

**Figure 9: Histogram Showing Distribution of Total Product Cost**

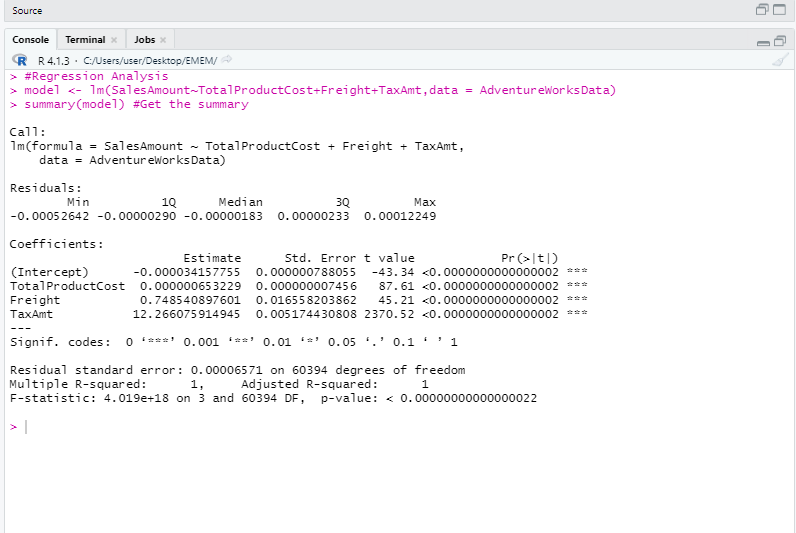
****

**Figure 10: Histogram Showing Distribution of Tax Amount**

The histogram generated for unit price, sales amount, total product cost and tax amount revealed that each of the variable has most of its values concentrated on the left side of the distribution. In other words, the variables have values that are not normally distributed.

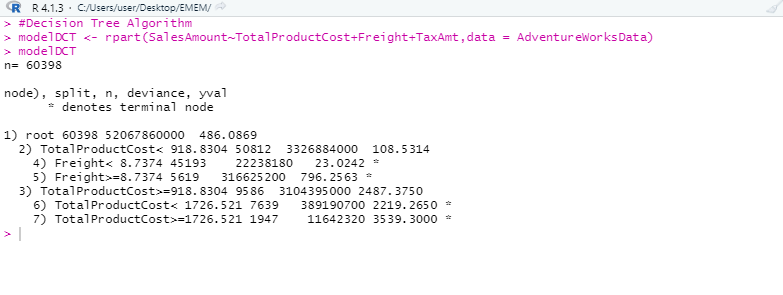
## Predictive Analysis

### Multiple Linear Regression

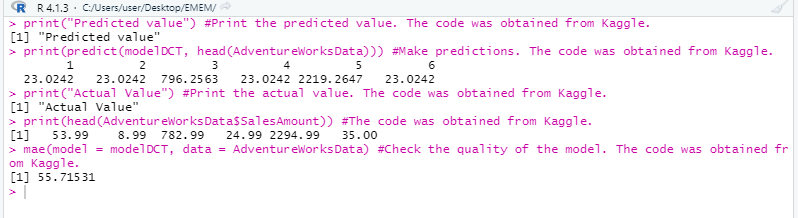
****

**Figure 11: Multiple Linear Regression**

### Decision Tree Algorithm



**Figure 12: Decision Tree Regression Result**

****

**Figure 13: Predicted Values and Test of Model Quality**

## Interpretation of Results

The result of the linear regression model conducted was presented in Figure 11. According to the result, a unit increase in total product cost will increase sales amount by 0.00001 units, holding all other variables constant. A unit increase in freight will increase sales amount by 0.7485, holding all other variables constant. A unit increase in tax amount will lead to a 12.266 units increase in the sales amount, holding all other variables constant. The constant term in the model was -0.00003. It is the value of sales amount when all other variables are held constant. The p-value of total product cost was 0.000, a value less than 0.05. This means total product cost was significant to the model. The p-value of freight was 0.000, a value less than 0.05. This showed that freight was statistically significant to the model. The p-value of tax amount was 0.000, a value less than 0.005. This implied that tax amount was statistically significant to the model. The p-value corresponding to the F-statistic was 0.000, a value less than 0.05. This showed that overall; the variables were statistically significant to the model.

The Decision Tree algorithm was employed to make predictions on our dataset and check the quality of the model. We predicted the sales amount for the first five samples in the data, and then compared with the actual values. It was revealed that in some cases, the predicted sales amount was lower than the actual sales amount, and in some other cases, the predicted sales amount was greater than the actual sales amount.

There are many metrics for summarizing model quality. In most (though not necessarily all) applications, the relevant measure of model quality is predictive accuracy (Tatman, 2018). The mean absolute error was used to test the quality of the model. This was made possible with the use of **modelr** library. The value obtained for the mean absolute error (MAE) was 55.7, indicating a high quality for the Decision Tree model.

## Recommendation for Performance Improvement

Adventure Works Cycle is a large company with operations in different parts of the world. It is important that the company assess how it is performing in the regions where it operates, and also examine its performance in relation to competitors that also operate in the same industry as Adventure Works Cycle. The following recommendations are made for performance improvement:

* Adventure Works Cycle should adopt cost minimization techniques that will ensure it is able to lower its production cost in the necessary regions.
* The company should engage in sales promotion. This will help attract more customers to the company. The company will also be able to boost its sales revenue.
* The company should consider engaging in wide publicity in those areas where its sales is low.
* The company should partner with shipping companies that will provide freight for them at low costs. This will help them reduce their overhead, and consequently increase their profit.
* Adventure Works Cycle can expand its operations to other regions where they have not been using its products.
* The company should aim for converting visitors to its website to potential customers by offering irresistible offers that will get them fascinated to its products. This can be done by offering discount to customers who buy in bulk; promotion strategy and reward on referral.

# **References**

Tatman, R., 2018. *Welcome to Data Science in R: Melbourne Housing Snapshot.* [Online]   
Available at: https://www.kaggle.com/code/rtatman/welcome-to-data-science-in-r  
[Accessed 29 April 2022].

## Appendix

