

**ASSIGNMENT: Individual Report 2500 words**

**BMD0004: Managing Big Data**



Quoc Khanh Bach (u2260074)-University of Huddersfield

Table of Contents

[Executive Summary 1](#_Toc158337226)

[1. Introduction 1](#_Toc158337227)

[2. Data Brief 2](#_Toc158337228)

[3. Data Preparation: 3](#_Toc158337229)

[4. Analysing key metrics and Data insights 7](#_Toc158337230)

[4.1 *Product categories’ distribution*: 7](#_Toc158337231)

[*4.2* *Total profit by product categories:* 8](#_Toc158337232)

[4.3 *Total sales by markets*: 9](#_Toc158337233)

[*4.4* *Customer demographic:* 10](#_Toc158337234)

[4.5 *Sales forecast*: 11](#_Toc158337235)

[5. Conclusion 13](#_Toc158337236)

[6. References 14](#_Toc158337237)

# Executive Summary

Big Data Analyticsrefers to the evaluation and investigation of data generated through the operation of organisation, industries or even government globally aiming at producing meaningful insights from those data to serve their objectives. In today’s business world, Big Data Analytics is becoming inevitadble across industries ranging from banking, health care, supply chain and so on (Shan et al., 2019). Thus, to further address the importance of Big Data Analytics, especially in the retailing sector, this paper will conduct a systematic analysis on the dataset of a bike retailing company which is collected from Kaggle.com website. Initially, the importance of Big Data to organisational performance in general and the selected case and industry in particular will further be explained. Then, a brief description of the chosen dataset will be provided along with the explanation of key metrics utilised to analyse the data. After that, the data cleansing process will be conducted adopting SQL server management studio tool to get rid of any problematic error in the dataset which could affect the analysis and calculation. Next, cycling products data will be visualised implementing the Tableau tool and explained based on developed measurements. Lastly, strategic insights on the visualised data will be provided to suggest operational strategies and decisions enabling the bike retailing firm to enhance its over performance and profitability.

# Introduction

In the contemporary business world, the rapid advancement of digitalisation and modern technologies has been posing challenges for global organisations to be more productive and sustainable (Shan et al., 2019). This phenomenon could be directly solved by adopting Big Data Analytics concept which could provide organisations critical business insight based on their historical data, allowing them to promote performance whilst adapting with unpredictable nature of the modern business environment (Lutfi et al., 2022). Although the notion of Big Data Analytics has been around for decades attracting the attention of numerous scholars globally, it is yet to be fully utilised by international enterprises, especially small and medium ones (SMEs). In instance, it was found that 80% of SMEs tried and failed to use big data analytics to enhance their operations. This is the result of a number of difficulties encountered when implementing big data analytics, including high implementation costs and a shortage of qualified data professionals, among other issues (Lutfi et al., 2022). Big Data analytics is still, however, the most beneficial and sustainable solution to any business issue. In particular, this idea improves business operations and performance in a number of ways, such as reducing operating expenses, offering thorough market knowledge, encouraging innovation, and many more (Lutfi et al., 2022). During the course of this paper, to further examine the importance and application of Big Data Analytics in retailing industry specifically, the Big Data of a bike’s products retailing company will be analysed supported by two popular data analytic tools which are SQL server management studio and Tableau.

Among all practices of Big Data Analytics in retailing industry, the concept of retail analytics is the most remarkable. In simple definition, retail analytics are the various stages of data analysis performed on sales data in the retail industry, including pattern analysis, decision making, and data collecting (Rooderkerk et al., 2022; Mohammed, 2023). Retail analytics come in four primary categories: prescriptive, predictive, diagnostic, and descriptive. Each type caters to a particular business need or circumstance. Businesses in this sector can gain vital insights regarding their products, consumers, and market by implementing the suitable type of retail analytics. These insights can then be used to create accurate and adaptable business plans that support any performance improvement (Rooderkerk et al., 2022).

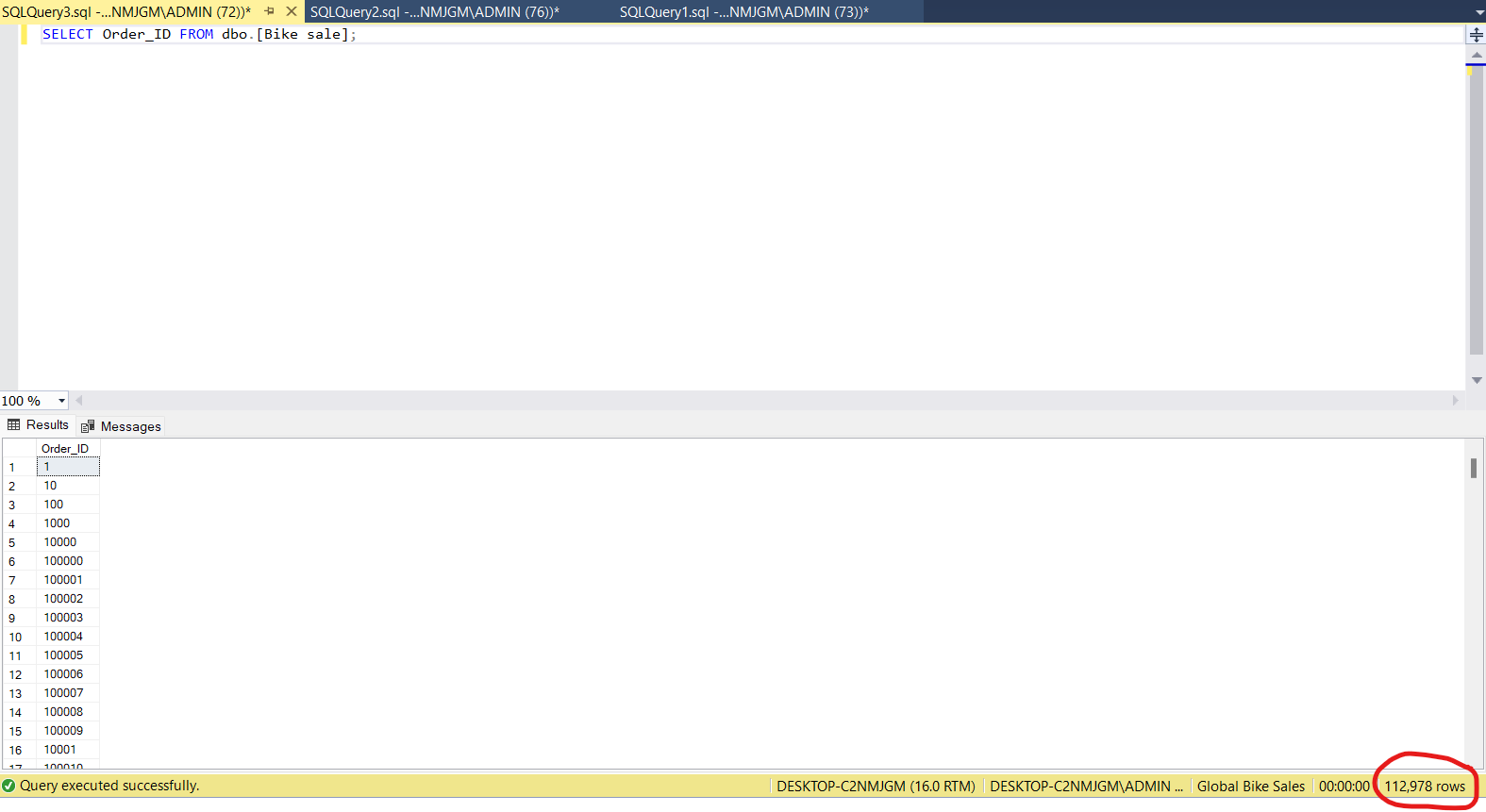
# Data Brief

In this report, the dataset of the mentioned bike retailer will be gathered via Kaggle.com, a platform that offers scientists working in machine learning and data-related sectors a competitive setting to hone their data analytics abilities (Shah, 2020). In detail, the chosen dataset consists of more than 100,000 data rows including information regarding customers’ expenditure on different cycling products in six developed countries around the world from 2011 to 2016 (Shah, 2020). Moreover, this dataset includes several variables such as *order quantity*, *customer age group*, *revenue* and *profit* which will be adopted to develop key performance metrics that help producing critical strategic business insights to promote business’s performance. In detail, these metrics includes *product categories’ distribution*, *total profit by product categories*, *total sales by market* and *customer demographic*.

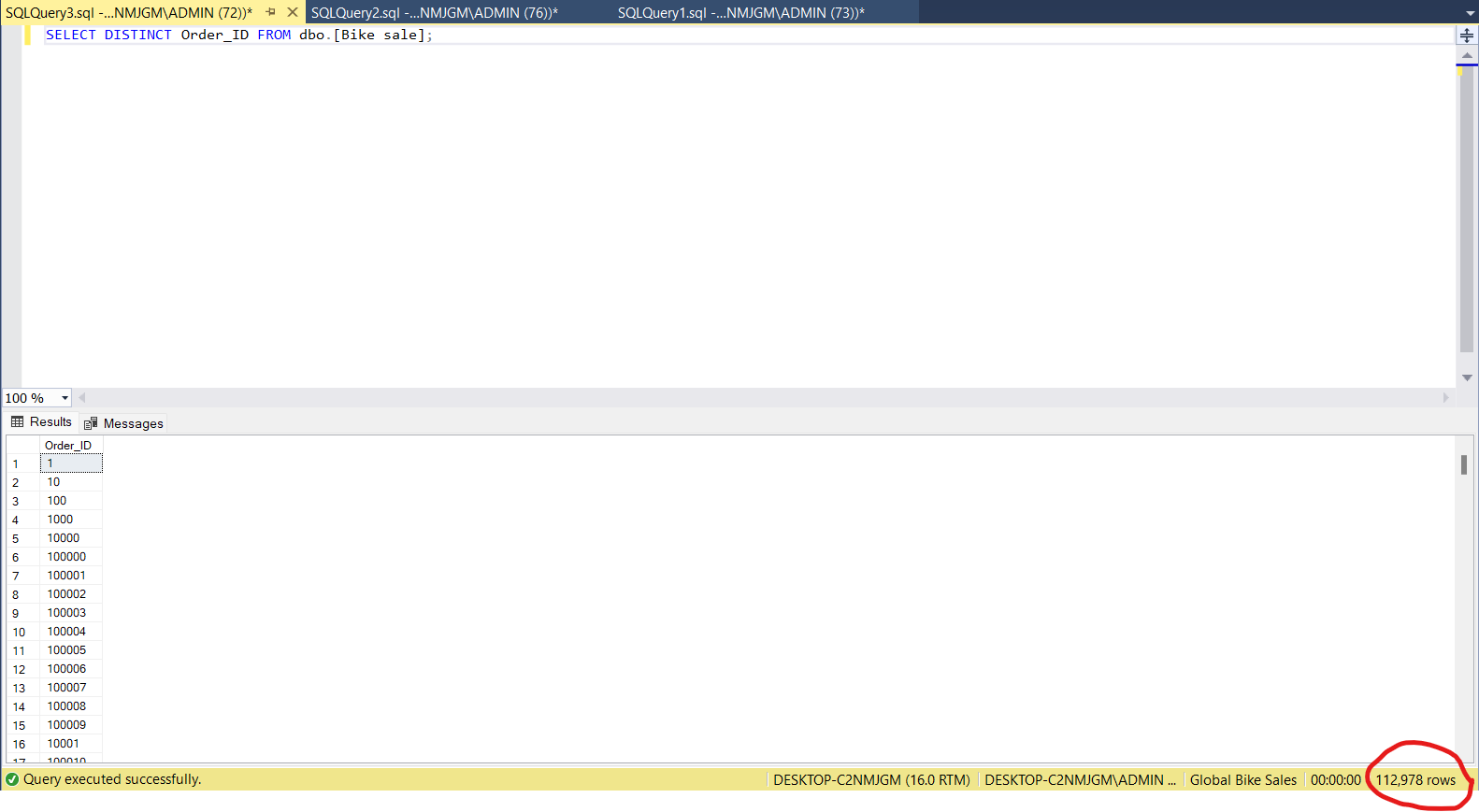
By adopting the developed metrics, this paper aim to address the following criteria. Firstly, evaluating the popularity of each product category based on the *product categories’ distribution* measurement to adopt suitable strategies that fulfil the market demand. Secondly, identifying the most profitable products in each of the cycling category throughout the given time period so as to provide strategic decision on the development and promotion of each product. Furthermore, the *total sales by market* metric will be applied to highlight the most significant market to be focused on by the bike retailer. Besides, adequate product segmentation is expected to be established by analysing the *customer demographic*. Lastly, a forecast on the total revenue of the bike retailing business in 2 following years will be conducted based on the available historical data in the dataset. However, before the raw data are analysed and visualised, they will firstly be cleaned in the so-called data cleansing process using SQL Server Management Studio indicated in the next section.

# Data Preparation:

Data preparation, also know as data cleansing, is amongst the most significant procedure in Big Data Analytics implementation (Ridzuan and Wan Zainon, 2019; Hosseinzadeh et al., 2023). In detail, raw data regularly consists of errors which need to be addressed before data are analysed in order to avoid miscalculation. Besides, most popular issues include *duplicated* and *null* values which is also addressed in the dataset of this study. Initially, *duplicated* data are harmful to the analysis and calculation of Big Data due to the waste of data resource consumption since multiple efforts are conducted to interpret similar information producing similar results (Ridzuan and Wan Zainon, 2019). To clean duplicated value, this study applied simple SQL queries to investigate duplicated value in the primary key of the dataset as could be observed in *figure 1* and *figure 2*.



*Figure 1: Filtering Order ID rows:*

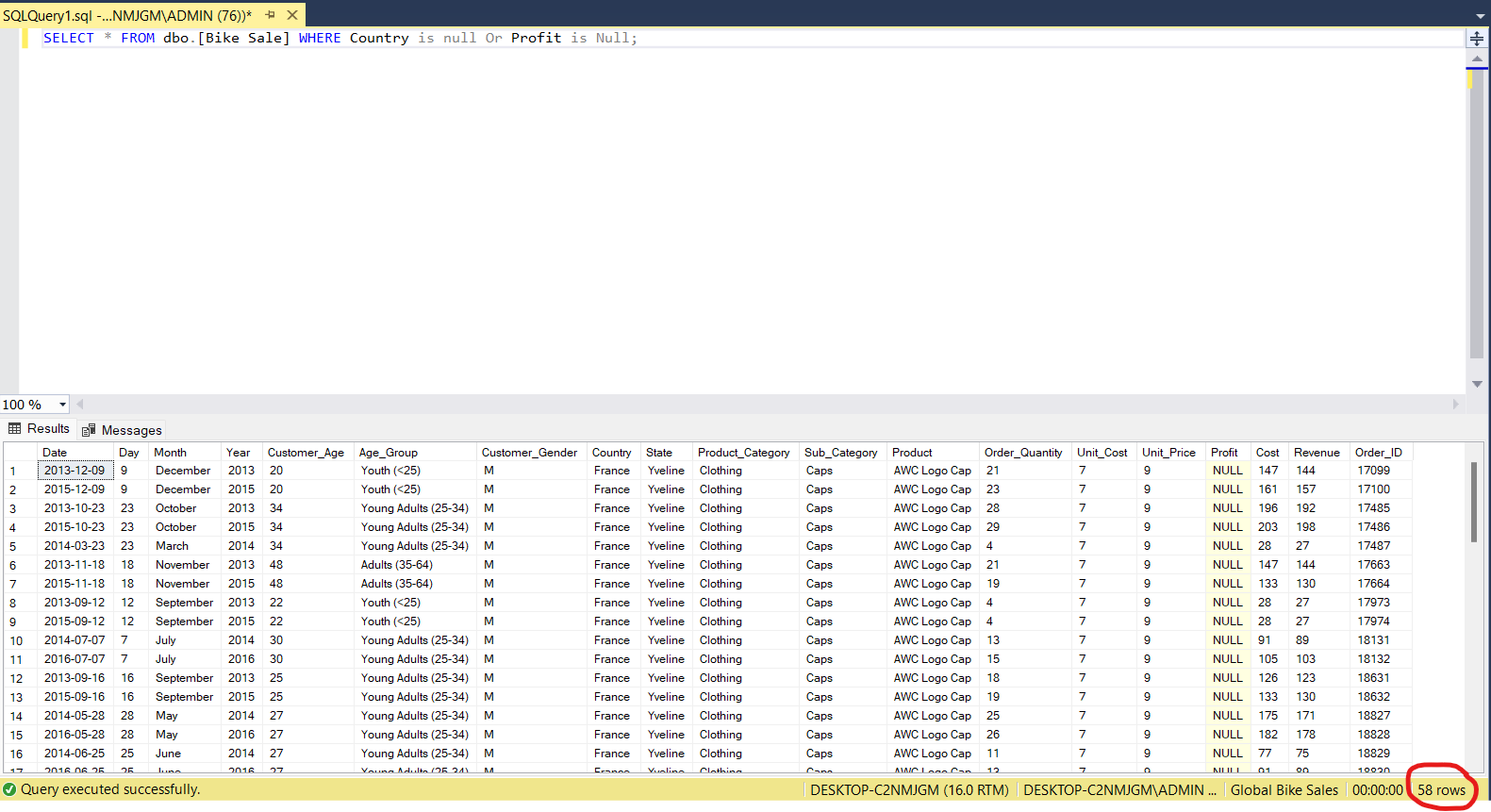


*Figure 2: Finding duplicate Order ID rows:*

To be specific, the primary key refers to data columns that could only include unique value and cannot be duplicated (Larson et al., 2015) which is the Order ID column in the bike retailing dataset. Thus, all of the data in the Order ID column are firstly listed using the “SELECT” query in *figure 1*. After that, to check if there are any duplicated value, the “SELECT DISTINCT” query is adopted in *figure 2*. The result after conducting two previous queries indicates that there is no duplicated value in the Order ID column as the number of rows listed in both cases are exactly the same at 112,978 rows (*figure 1* and *figure 2*). In other words, duplicate values in the chosen dataset are already cleaned.

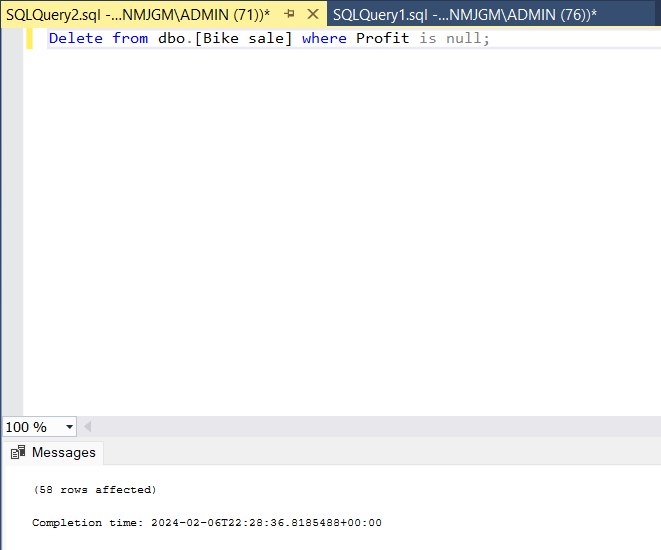
Additionally, *null* value is also a problematic issue in Big Data Analytics as it directly affects the accuracy of data when being processed. In explanation, null is defined as data fields that include missing values which often occur and could not be avoided during the collection process of Big Data (Ridzuan and Wan Zainon, 2019). There are two main approaches to deal with null data. The first solution is to remove the entire missing data areas such as a whole data row or column. Despite being straightforward and simple, this method is not highly recommended since it poses other risks such as losing essential data that could be used in future analysis leading to inadequate business insights (Hosseinzadeh et al., 2023). Therefore, another solution is suggested which is to logically replace a null value with non-null value such as random values or previous values in the same row or column and so on (Hosseinzadeh et al., 2023). Comparatively, imputing null values strategy is less risky than the deletion method when handling null issue in Big Data Analytics, especially if a substantial volume of null value exists in the dataset that may all be disposed of and not recoverable for future use. However, eliminating null values might be more preferable if they represent only a small percentage of the dataset. This will lessen the likelihood of incorrect data calculations arising from the addition of erroneous and random values to the dataset (Hosseinzadeh et al., 2023).

In this study, the “SELECT” and “ISNULL” queries are firstly applied to each of the data column to identify columns that consist of null value. The result demonstrates that, within over 100,000 rows of data analysed, there are only 58 rows in the Profit column that include null values as could be seen in *figure 3*.

****

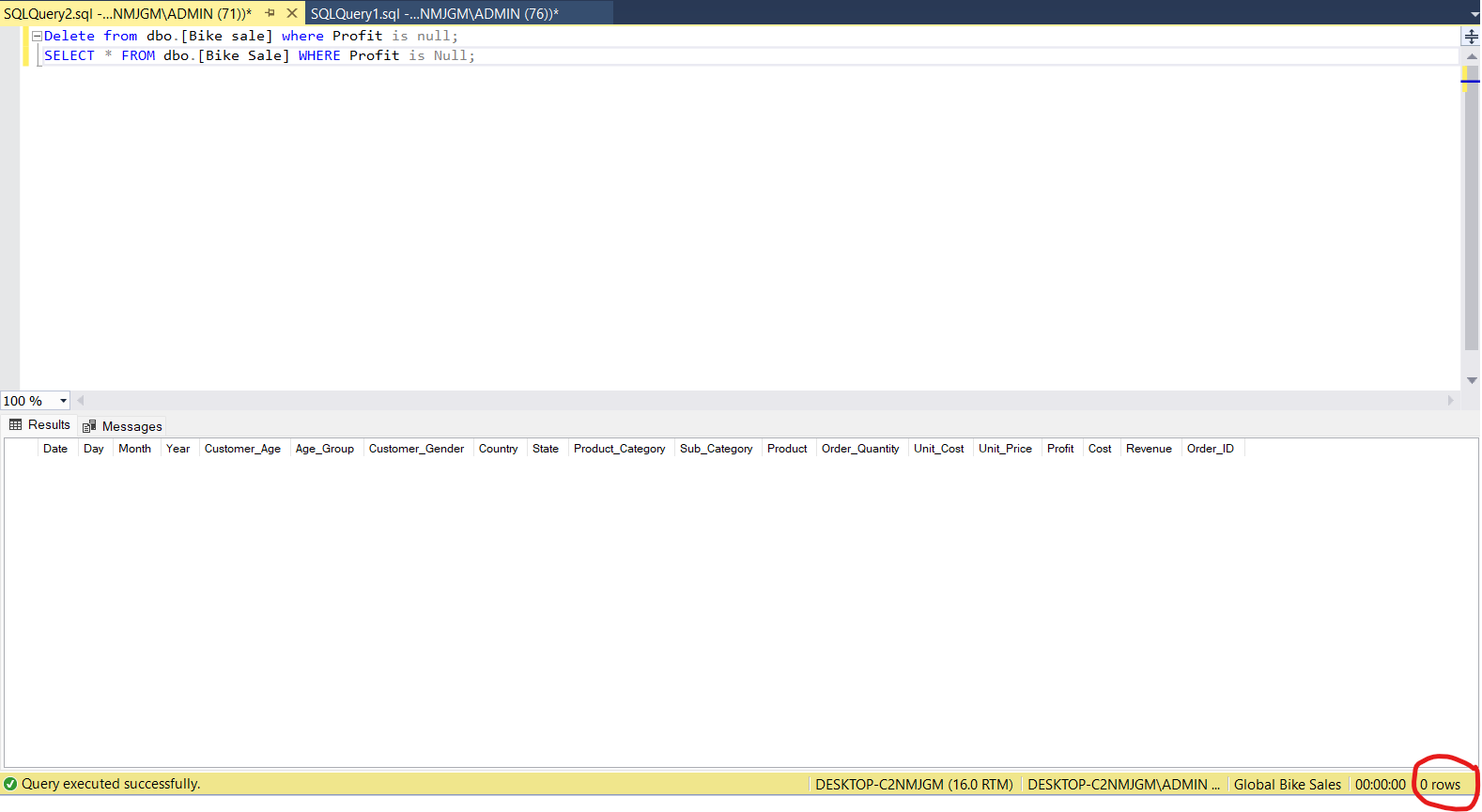
*Figure 3: Null values investigation*

Since the number of null values found in this case is insignificant comparing to the data size, the deletion method is adopted in this study to clean these null data using “DELETE” and “ISNULL” queries (*figure 4*).

****

*Figure 4: Null value deletion*

Lastly, the “SELECT” and “ISNULL” queries are applied again to the profit column to ensure that the data is cleaned (*figure 5*).

****

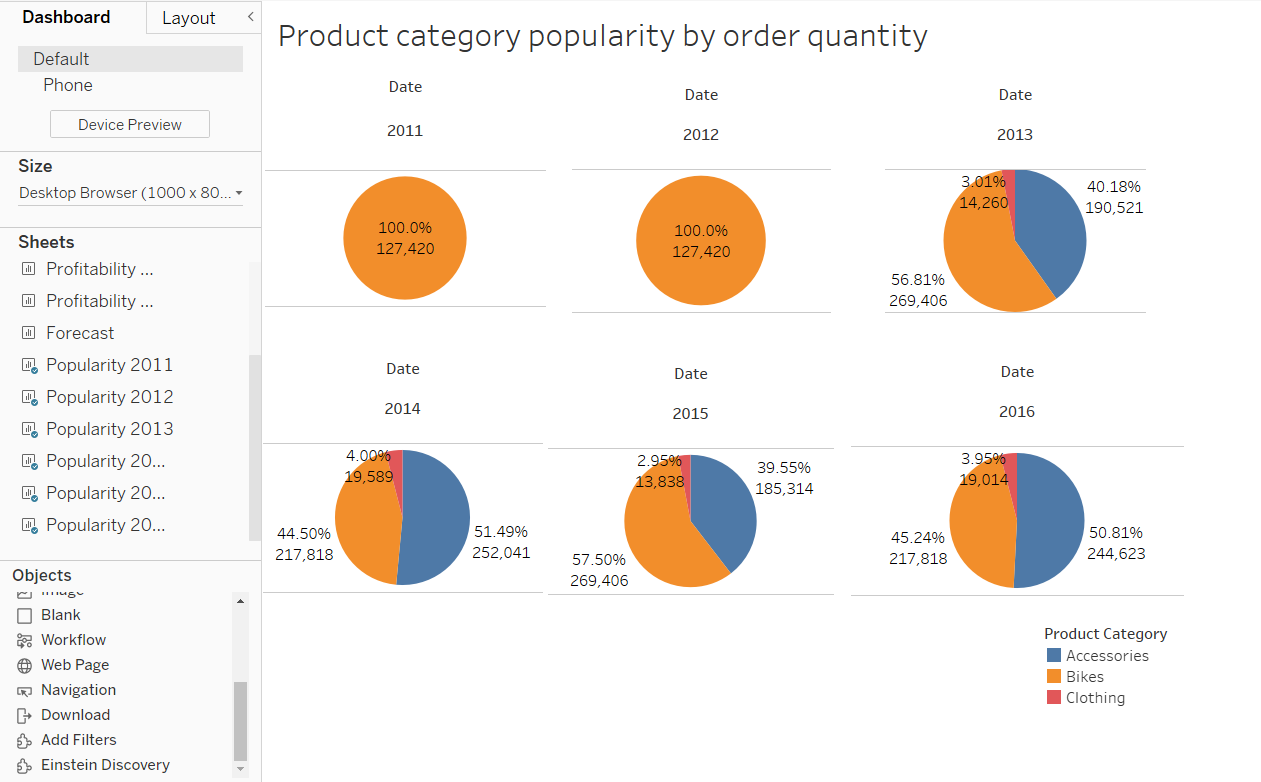
*Figure 5: Result evaluation after handling null values*

After the data cleansing process is performed successfully, cleaned data of global bike retailing are analysed and visualised using four key measurements mentioned in the data brief section.

# Analysing key metrics and Data insights

## *Product categories’ distribution*:

The first measurement compares the total demand or order quantity of three main product categories in the selected dataset including accessories, bikes and clothing in each year from 2011 to 2016. In other words, this metric provides information about the popularity of each category in the given period to support operational decisions such as pricing strategies or production quotas to fit the market demand and maximise revenue (Shen et al., 2014; Tahat, 2023).

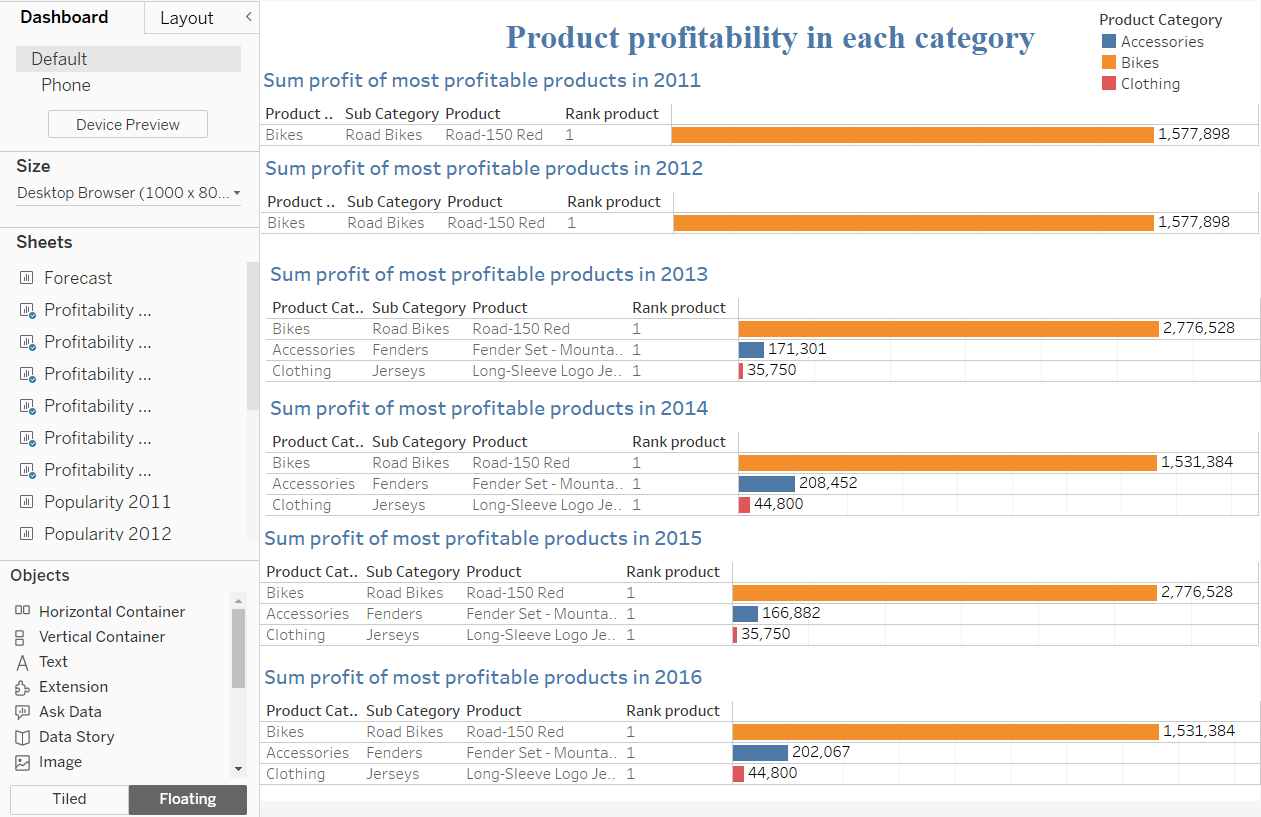


*Figure 6: The annual distribution of 3 product categories*

**Data insights:** Figure 6 shows six Tableau-generated pie charts that demonstrate the total order quantity for each of the three product categories over the course of six years, from 2011 to 2016. Since there was no information available for the other two cycling product categories in 2011 and 2012, all orders were placed under the bikes category during those years. This might indicate that the bike retailing company only offered bikes at that period. In later years, as could be observed, bikes and accessories were the most popular segments whose order quantities alternatively account for 45% to 55% of the total order during the period from 2013 to 2016. To utilise the analysed statistics, the bike retailing company could adjust the production of bikes and accessories as well as their prices to take advantages of the high demand in these categories (Shen et al., 2014).

## *Total profit by product categories:*

Whilst the previous metrics analyses the demand for the entire product category, this second indicator provides information regarding the profitability of specific products. The purpose of this measurement is to investigate most profitable products so as to adopting suitable business strategies such as marketing campaigns, research and development activities. Through these strategies, retailers could promote customer’s engagement with their best products to maximise profit (Mohammed, 2023; Tahat, 2023).

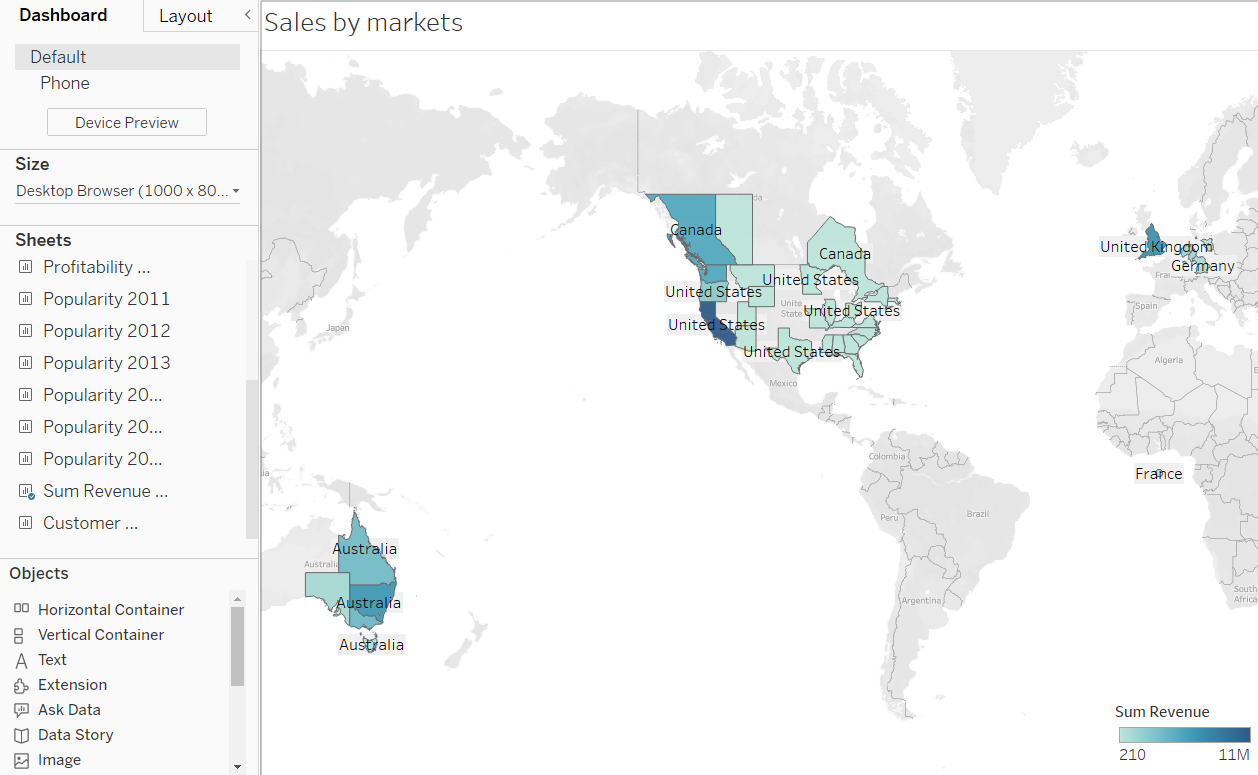


*Figure 7: Most profitable products by each category*

**Data insights:** *Figure 7* illustrated total profits of three best products amongst 23 goods in all three categories of the bike retailing dataset. To be specific, six bar charts indicate that, within the entire period, Road-150 Red generated the highest profit for the bike selling firm in bikes segment at over 1.5 to 2.7 million profit per year. Meanwhile, Fender set-Mountain and Long-Sleeve Logo Jerseys were the most profitable one in accessories and clothing categories, respectively. Based on these data, the bike company could develop a marketing or promotion campaign for the three aforementioned products to utilise their profitability and boost the company’s performance (Mohammed, 2023).

## *Total sales by markets*:

In this metric, the total revenue of all cycling products in various parts of six countries is visualised as could be seen in *figure 8*. The expected outcome is to identify greatest markets to promote whilst making decision on dealing with worst markets such as withdrawing or producing better business strategies (Rooderkerk et al., 2022).

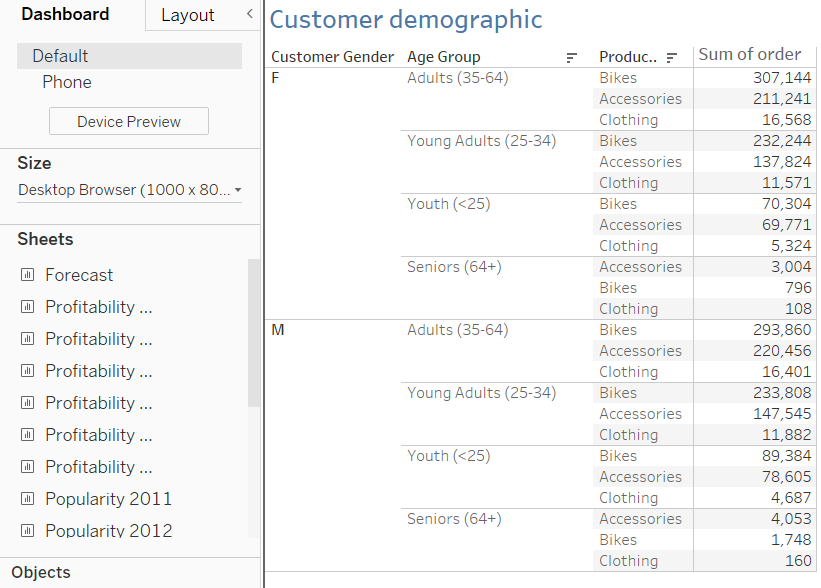


*Figure 8: Total sales by markets*

**Data insights**: The bike retailing company’s total sales across its operational zones varied significantly, as observed by the map presented in *Figure 8*. In particular, the market in Western America performed the best in terms of revenue, at around 11 million currency units. In contrast, Middle East regions of America and parts of Canada fared poorly, generating less than 210 currency units after six years. As a result, the bike retailing company could enhance its performance by promoting the Western America region. For instance, a marketing campaign aligning the company products with the interest and culture of people in this area could be launched. On the other hand, to mitigate the impact of low-performance market such as East America, the company could withdraw from this region to reallocate resource to other promising areas.

## *Customer demographic:*

Analysing customer demographic is one of the best options to enhance customer engagement in the retailing industry. In this measurement, two basic consumer demographic information which are gender and age group are adopted to support product segmenting decisions (Ho et al., 2023).

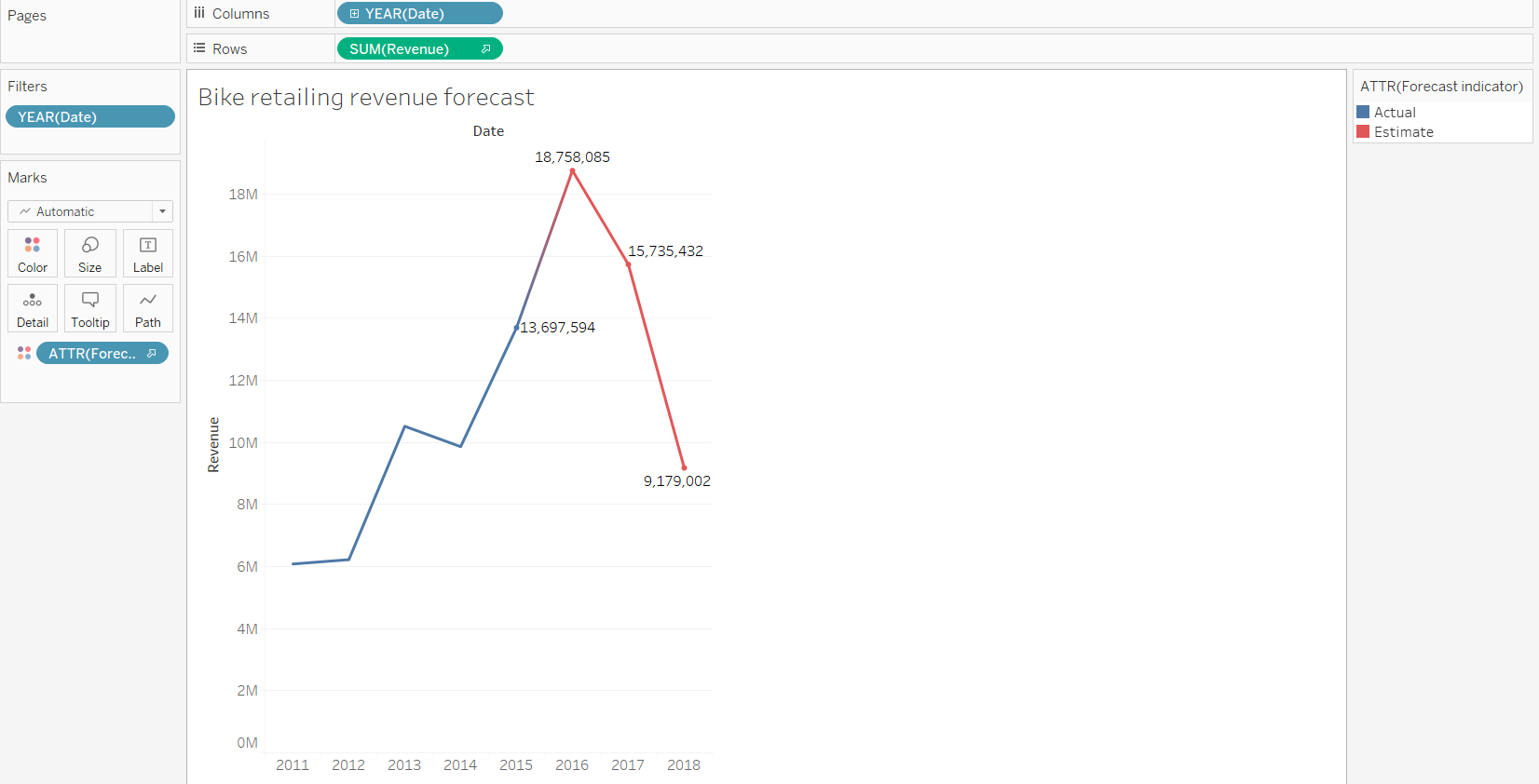


*Figure 9: Customer demographic by total demand for each product category*

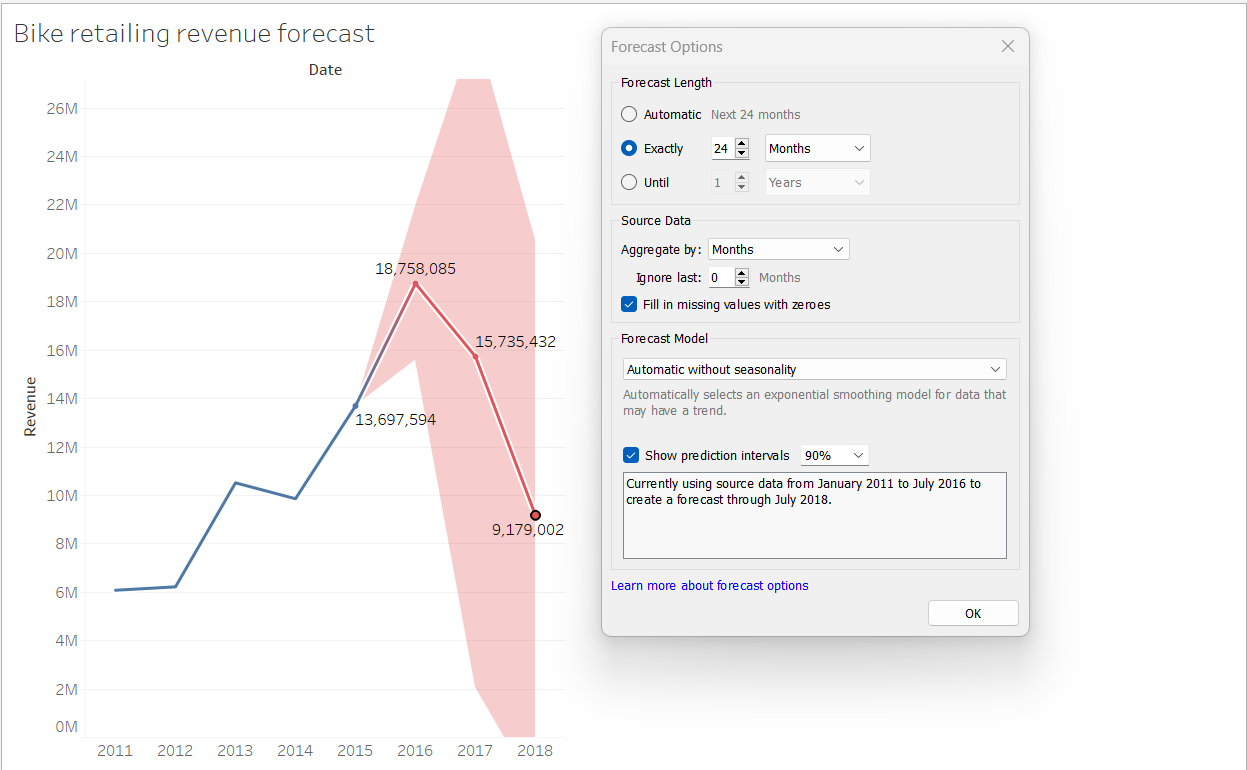
**Data insights:** The table in *figure 9* demonstrates that male and female adults in the 35-64 age group are the most important customers of the bike retailing firm who had considerably high demand for all cycling products throughout the observed period. Besides, senior males and females are the least potential customers whose demand for these products is relatively low in all categories. Thus, it is the most optimal for the bike retailing company to focus on product segmentation that serve the interest of buyers in adults age group in order to promote its performance.

## *Sales forecast*:

*Figure 10* provides a forecast on the total sales of the bike retailing firm in two following years 2017 and 2018. As could be observed, after reaching its peak in 2016 at 18,758,085 million, the total revenue of this retailer is predicted to significantly decrease in the next two years to nearly 16 million in 2017 and 9 million in 2018. Based on past data, prediction interval of 90% as in *figure 11* suggests that there is a 90% possibility that this decline would occur in the future. Besides, Tableau employed exponential smoothing, a time series forecasting technique, to produce this forecast (Tableau, n.d.). Before evaluating the forecast’s accuracy, it is essential to comprehend the two types of time series forecasting approaches which are moving average (MA) and exponential smoothing (ES) (Zhou et al., 2013). In particular, the MA technique is applied to datasets with randomness that have the potential to yield simple short-term forecasts devoid of patterns like seasonality and trend. On the other hand, ES is adopted when there are patterns in the forecasted data (Zhou et al., 2013). Thus, the application of ES method to predict the future revenue of the bike retailing company is suitable in this case since an upward trend is observable in the graph (*Figure 10*). Nevertheless, the accuracy of the forecast is uncertain since it depends on the calculation of forecasting error such as mean absolute deviation (MAD) or mean squared error (MSE) (Zhou et al., 2013).



*Figure 10: Sales forecast*

**

*Figure 11: Prediction intervals*

# Conclusion

Overall, through the analysis of a global bike retailer’s dataset, this study has highlighted the application of Big Data Analytics and its importance to the retail sector. The paper offered multiple strategic insights on the performance of the chosen firm from 2011 to 2016 and suggested several business strategies that could be implemented to improve such performance in the future by analysing the data based on four key metrics developed, namely *product categories’ distribution*, *total profit by product categories*, *total sales by market*, and *customer demographic*. Finally, a projection is produced on the company's revenue growth over the next two years, whose accuracy necessitates the review of further data and formulation.

# References

Ho, T., Nguyen, S., Nguyen, H., Nguyen, N., Man, D., & Le, T. (2023). An extended RFM model for customer behaviour and demographic analysis in retail industry. *Business Systems Research, 14*(1), 26-53. <https://doi.org/10.2478/bsrj-2023-0002>

Hosseinzadeh, M., Azhir, E., Ahmed, O. H., Ghafour, M. Y., Ahmed, S. H., Rahmani, A. M., & Vo, B. (2023). Data cleansing mechanisms and approaches for big data analytics: A systematic study. *Journal of Ambient Intelligence and Humanized Computing, 14*(1), 99-111. <https://doi.org/10.1007/s12652-021-03590-2>

Larson, P., Birka, A., Hanson, E. N., Huang, W., Nowakiewicz, M., & Papadimos, V. (2015). Real-time analytical processing with SQL server. *Proceedings of the VLDB Endowment, 8*(12), 1740-1751. <https://doi.org/10.14778/2824032.2824071>

Lutfi, A., Alsyouf, A., Almaiah, M. A., Alrawad, M., Abdo, A. A. K., Al-Khasawneh, A. L., Ibrahim, N., & Saad, M. (2022). Factors influencing the adoption of big data analytics in the digital transformation era: Case study of jordanian SMEs. *Sustainability (Basel, Switzerland), 14(*3), 1802. <https://doi.org/10.3390/su14031802>

Mohammed, F. C. (2023). Retail analytics and firm profitability: How analytics can help retail companies be successful. *Business & IT (Praha, on-Line), XIII*(1), 193-202. <https://doi.org/10.14311/bit.2023.01.21>

Ridzuan, F., & Wan Zainon, W.M.N. (2019). A review on data cleansing methods for big data. *Procedia Computer Science, 161*, 731-738. <https://doi.org/10.1016/j.procs.2019.11.177>

Rooderkerk, R. P., DeHoratius, N., & Musalem, A. (2022). The past, present, and future of retail analytics: Insights from a survey of academic research and interviews with practitioners. *Production and Operations Management, 31*(10), 3727-3748. <https://doi.org/10.1111/poms.13811>

Shah, S. (2020). *Bike sales in Europe*. Kaggle. <https://www.kaggle.com/datasets/sadiqshah/bike-sales-in-europe>

Shan, S., Luo, Y., Zhou, Y., & Wei, Y. (2019). Big data analysis adaptation and enterprises' competitive advantages: The perspective of dynamic capability and resource-based theories. *Technology Analysis & Strategic Management, 31*(4), 406-420. <https://doi.org/10.1080/09537325.2018.1516866>

Shen, W., Duenyas, I., & Kapuscinski, R. (2014). Optimal pricing, production, and inventory for new product diffusion under supply constraints. *Manufacturing & Service Operations Management, 16*(1), 28-45. <https://doi.org/10.1287/msom.2013.0447>

Tableau. n.d. *How Forecasting Works in Tableau*. Tableau. <https://help.tableau.com/current/pro/desktop/en-us/forecast_how_it_works.htm#SnippetTab>

Tahat, I. (2023). Pricing strategy and firms profitability. *Socioeconomic Challenges, 7*(4), 124-132. <https://doi.org/10.61093/sec.7(4).124-132.2023>

Zhou, S. X., Wu, D., & Li, C. (2013). Research on business intelligent forecasting method with time series. *Information Technology Journal, 12*(22), 7009-7009. <https://doi.org/10.3923/itj.2013.7009.7013>