EBUS3030 Assignment 2

Stavros Karmaniolos c3160280@uon.edu.au
Jay Rovacsek c3146220@uon.edu.au
Jacob Litherland c3263482@uon.edu.au
Edward Lonsdale c3252144@uon.edu.au

September 21, 2018

Contents

1 Assignment Overview & Requirements

Business Intelligence - EBUS3030
Assignment 2

Assignment Outcomes
This assignment requires multiple outputs to be created to exhibit your understanding of business intelligence/data analysis through an example 'real world' question that is comparable to what you may be asked of you as you become an IT professional.

Key outcomes to be delivered are: Data Modelling/ETL to get the data in a usable format, Output of your analysis, Report summarising your findings and a presentation to the class of your work. The presentation is expected to concentrate more on your findings/recommendations as if it were a situation where you are presenting the response to the CEO.

Assignment Question
The CEO of 'BIA Inc' had been speaking to the Sales Executive and had heard about some recent work you had completed and thought you might be able to assist them with a problem they have.

"I've heard that you helped the Sales Exec recently with understanding more about our Newcastle site. I would now like your help to get a handle on the whole business. As you are aware, the Newcastle office is just 1 of 10 sites we have across the country. Unfortunately, sales in some items have dropped across the country in recent years and we are currently running at a loss.

We need to consider consolidating our company offices. We need to reduce costs for the longevity of the company as a whole. I need you to get some numbers together around the performance of our 10 offices, so that I can factor this information into any decision regarding which office (or offices) we might consider closing.

I would like a summary of recent numbers and some trend analysis as well please. It would be great if you could also project sales for the next 12 months for each office as well. It would be helpful if you could indicate the 3 most popular and 3 least popular items in each of our stores, as well as the worst performing items for the company as a whole.

As you can imagine, this is a very sensitive topic so, as part of your response I want you to provide the justification as to which office we may close. Our decision will upset some people and I want to make sure we have all the background information on hand. If you can provide a Ranking of offices based on your analysis that would be wonderful.

…. I believe you started to bring together a data store of this information from the Newcastle Office, can you expand that and load all of the sales information for all offices and complete your analysis.
"

Assignment Deliverables

Using the data file provided in Excel and notes about the data (*Assignment 2 – data.xlsx*), you are required to complete the following elements as part of the assignment.

- Data Model/Data Load Process
  - Provide an overview of the data model & ETL process completed to get the data ready for analysis
  - Ensure you record any assumptions you have made as part of this component and your reasoning behind the assumption.
- Analysis including any predictive work undertaken
  - Provide the SQL and raw output of your base analysis
  - Provide workings of the predictive work you completed for the trending & prediction on future sales.
  - Ensure you record any assumptions you have made as part of this component and your reasoning behind the assumption (this includes answers to any relevant questions put to management (your lecturer) during workshops.

- Executive Summary & Presentation in response to business question.
    - Provide an Executive report and Dashboard
        - Your Executive report should include a short Executive brief/summary that presents a clear concise response back to the CEO question about possible downsizing of operations including evidence/justification.
        - A dashboard should allow quick comparisons between the sites to be undertaken as well as contain at least one element of 'predictive' analysis
        - Present the material as if it is to be consumed in a formal boardroom meeting
            - All members of the team need to participate in a (15 – 20 minute) presentation to be given as part of the lab in Week 11.
            - Please be aware that any of the company Executive may ask questions as you present your findings.

NB: As part of your response, you should also specifically include any assumptions/external information you have made/used throughout the process as well as any quality processes/checking you have completed or limitations you discovered.

Breakup of assignment Marks (total course mark for assignment = Assignment Part B submission (28% + Presentation Two (7%) = 35%.

Assignment Component	Percentage Allocation
Data Model/ETL	10%
Core Analysis	50%
Executive Summary/Evidence	25%
Dashboard	15%
	100%

Key Documents Required & Format

You are required to upload all files in a single zip file (including any presentation items for the team delivery within the lab) via blackboard to the Assignment Two drop folder by 12 noon, Thursday 25th October. You will also be required to submit a paper copy of your report at the beginning of the presentation workshop (make sure this is printed well before the workshop and has a group Assessment Cover sheet signed by ALL team members).

NB: Only 1 load per team only but it should contain all of the deliverable items.

Your data model should include a printout of an ER diagram using the notation described in lectures. It should also include a printout of your SQL schema showing Primary and foreign keys, as well as all attributes.

Your presentation is worth 7% of the course mark. It should simulate presenting the report to management. You should time your presentation to be between 12-15 minutes with 5-10 minutes for questions. You presentation should include a demonstration of your dashboard, results and recommendations from your analysis. Your presentation marks will contain components for organisation, comprehension of presented results, and timing.

2 Executive Summary

This report was created for the head sales executives of BIA Inc for the purpose of determining which sales staff member would be considered the best performer based on the data set provede by the firm. As no specific metrics or measurement requirements were provided, an analysis was performed by our team. This analysis included extracting, cleaning and loading the data using SQL scripts and a supplementary Python script which assisted in preparing the data provided by the firm for analysis and query in SQL Server Management Studio (SSMS).

The findings of our analysis resulted in ranking high achieving staff members determined by a number of key metrics selected by the team. Firstly, we ranked the sales officers by total number of sales and, from this group, identified the top staff member of sales, Mr Daniel Baker. Mr Baker had made the largest number of sales in the 12 months of data supplied with a total of 700 sales. Placed immediately after Mr Baker, with 18 fewer sales is Ms Kaitlyn Ortiz (682), then Ms Michelle Miller (676), followed by Ms Stephanie Watson (664) and Mr Evan Hill (664).

The next metric was total items sold. In this relation, the best sales officer is Ms Kaitlyn Ortiz, the second place sales officer in the first metric. In the 12 months of data supplied, it can be observed that of the 682 total sales, Ms Ortiz sold 4217 items with approximately six (6.18) items per sale. The second place sales officer, Mr Daniel Baker, sold 4212 items, only five items less than Ms Ortiz, and also sold approximately six (6.02) items per sale. The third placed employee in this relation is Ms Michelle Miller who resulted in 4414 total sales and approximately six (6.13) items per sale.

The third key metric is discounted sales ratio. As stated in the business rules document provided by the firm, any sale with five or more row items would be eligible for a 15% discount to the total sale. We identified this as an important factor because understanding how many items were discounted may offer insight into sales methods and techniques applied by sales officers which can then be used to enhance future performance. From the results, we aggregated the data by total percentage of sales which were discounted. Mr Robert Wood (84.57%) discounted the greatest share of his sales of all sales staff members. After Mr Wood comes Mr Dylan Hall (83.41%), then Ms Lauren Martin (83.31%), Mr Jordan Turner (82.74%), Mr Noah Brooks (82.72%) and Mr Daniel Baker (81.39%).

The final metric considered was total sales value per staff member. The purpose of this report is to find the "best" salesperson, therefore, we can assume that, along with other metrics, the firm would be interested to know which sales officer generates the most revenue. After examining the data, we can state that Ms Michelle Miller is the sales officer that generates the most value with a total of $78,572.22 over 12 months. This equates to $4,725.10 more than her nearest competitor, Mr Daniel Baker who generated $72,764.89 who was followed by Mr Noah Brooks with $71,699.67 and Ms Amber Hill with $70,514.68.

After carefully considering the aforementioned key metrics and reviewing the results, we can conclude that Ms Michelle Miller should be considered the most valuable sales office at BIA Inc. Our findings showed clearly that Ms Michelle Miller achieved the highest sales value when compared with other sales officer by a significant margin ($4,725.10). While Ms Miller was not ranked first in total number of sales or total items sold, she did rank highly in both relations. Ms Miller did however score poorly in her discounted sales ratio. This could indicate that Ms Miller is not as effective at 'upselling' as her fellow sales officers and this could be an area for improvement. We can assert that Ms Miller should be considered for the reward (and possible cash prize) suggested in the original document outlining the firm requirements. If for any reason Ms Miller should not be applicable or eligible for the discount, we would recommend Ms Kaitlyn Ortiz as the alternative choice due to her high ranking in all metrics discussed in this summary.

2.1 Datamart Business Rules

The following business rules were provided to be used in the context of this assignment:

- At BIA all customers interacts are in an online environment, all orders are electronic.
- Returning customers can provide POI information via the web interface and look up their record and that will flow with the sale.
- The sales associate can complete the order form/sale for the client.
- Each sale will have a receipt number/id.
- A receipt can have many line items.
- Each line item can only be for a single item, but the customer can purchase multiples of the same item.
- Where a customer has multiple line items, any sale with 5 or more row items (containing at least five (5) different items) is provided a 5% discount.
- The system automatically handles the total for the sale by looking up the item, then multiplying the costs per item by number purchased, and then should store this final field total as a record in the system (but should also be able to see clearly sales that were provided a discount.
- Item prices can change at any point, and the price the customer pays is the amount listed for the item on the sale date. We need to keep a record of all item prices historically so that we can determine what the store item price was at any particular past date.
- Only one (1) BIA sales assistant can be attributed to any receipt.
- Customers may visit multiple stores for purchases (ie they are not locked to a particular store). As a result, all customer records are replicated across all stores, so they do not need to be re-recorded at a store by store level.

With these considerations in mind, the following report was created to outline the discovery, creation and polish to satisfy the assignment requirements.

3 Data Model

The below data model is only a suggestion and is still subject to change into the future. A full create script can be found in the appendix

It must be noted that the structure of this data model is less than efficient, and it would be expected in a datamart situation that only at lower levels of data would this schema remain responsive in the manner it is now, as the outline suggests the datamart is not necessarily the most suitable design for future use, however suits very well currently.

It would be expected that only at extremely large data sets would this model prove a bad design. In such cases a model more representative of the snowflake or star schema would be heavily advised.

An EER diagram of the suggested data model:

4 Data Load Process (ETL/ELT)

Initial import of the data supplied in the xlsx file generated a very basic table that allowed us to analyze the data for potential outliers, confirm the business requirements of the data and then create tables from which the data model was derived.
The Imported table structure was as follows:

A decision to leave this initial import table as default was made to allow easy reference to the initially supplied excel data file.

In the following sections of Quality Assurance Processes, Assumptions and Reasoning and Base Analysis we intend to clarify the reasoning behind leaving the imported data in the default table suggested by SSMS.

4.1 Quality Assurance Processes

A number of queries were written to look for data which did not adhere to the spec outlined in business requirements and to ensure data was ”clean” before entry. The first instance of potential issues were encountered with a basic python script which checked validity of column data, it was found that cells starting at B13777 to the end of file in the originally supplied excel file were formula values and not static values, this would not have caused an issue with importing into SSMS however certainly broke the script temporarily.

After clarifying the issues with the aforementioned cells with Peter, a data file without the offending formula was supplied and used for the remainder of the assignment.

Discrepancies with some cell formating were noted in Reciept_Id column in the raw data presented to us in the excel file. After testing both an unmodified and a modified version of the excel file it was recognised that these cells did not impact the import of data into SSMS. The offending cells in question were: B13776 - B13865.
The next potential issue encountered was not until a suggested schema structure was complete and data was being scripted to be added to the new schema for analysis. The issue encountered was that receipt number 52136 seemed to be an incorrect entry, this was discovered when running the import query for the new schema:

```sql
INSERT INTO Receipt(ReceiptId,  ReceiptCustomerId,ReceiptStaffId)
SELECT DISTINCT(Reciept_Id),Customer_ID,Staff_ID
FROM Assignment1Data
ORDER BY Reciept_Id
```

Which resulted in the error:

```
Violation of PRIMARY KEY constraint 'PK_Receipt'. Cannot insert duplicate key in object
'dbo.Receipt'. The duplicate key value is (52136).
```

Leading us to recognise that either one of the entries could be incorrect, therefore best to investigate both records of the customer Id against the rest of the database:

```sql
SELECT * FROM Assignment1Data
WHERE Customer_ID='C32'
AND Staff_ID='S15'
AND Sale_Date='2017-11-12 00:00:00.0000000';

SELECT * FROM Assignment1Data
WHERE Customer_ID='C13'
AND Staff_ID='S4'
AND Sale_Date='2017-12-30 00:00:00.0000000';
```

When both queries were performed it was apparent that the data associated with C32 was the likely broken record and modification of the data occurred:

```sql
UPDATE Assignment1Data
SET Reciept_Id=51585,
Reciept_Transaction_Row_ID=(
    SELECT MAX(Reciept_Transaction_Row_ID)+1
    FROM Assignment1Data
    WHERE Reciept_Id=51585)
WHERE Customer_ID='C32'
AND Staff_ID='S15'
AND Sale_Date='2017-11-12 00:00:00.0000000'
AND Item_ID='14';
```

The next issue arose when again, attempting to run the aforementioned query to import into the new Receipt table, this time not one stray record was found, but a complete collision on the ReceiptId of 52137, this time as neither record seemed to have records that were correct, it was decided to move one to the maximum ReceiptId + 1:

```sql
UPDATE Assignment1Data
SET Reciept_Id=(
    SELECT MAX(Reciept_Id)+1
    FROM Assignment1Data)
WHERE Customer_ID='C27'
AND Staff_ID='S4'
AND Sale_Date='2017-12-30 00:00:00.0000000';
```

The same issue was replicated on ReceiptId 52138, resolved via:

```sql
UPDATE Assignment1Data
SET Reciept_Id=(
    SELECT MAX(Reciept_Id)+1
    FROM Assignment1Data)
WHERE Customer_ID='C30'
AND Staff_ID='S19'
AND Sale_Date='2017-05-16 00:00:00.0000000';
```

At this point we recognised the broken data likely continued for a while, and evaluated our hypothesis by looking at the original excel file. It turned out that data with ReceiptId from 52137-52145 was all broken in the same manner. The following query shows this well:

```sql
SELECT Reciept_Id, Customer_ID,Staff_ID
FROM Assignment1Data
WHERE Reciept_Id BETWEEN 52137 AND 52150
GROUP BY Reciept_Id, Customer_ID,Staff_ID
ORDER BY Reciept_Id;
```

In order to clean this data we looked at a number of potential methods, with an emphasis on avoiding effort in the task if possible but not breaking the data further, which to this point just appeared to be a collision of a number of receipts.
We knew a structure such as a CTE [3] would allow us to easily split distinct records which shared a receiptId and filter by a value such as row number.

```
WITH CTE AS
(
    SELECT ROW_NUMBER() OVER (ORDER BY Reciept_Id) AS RowNumber,
                Reciept_Id,
                Customer_ID,
                Staff_ID
    FROM  Assignment1Data
    WHERE Reciept_Id BETWEEN 52137 AND 52150
    GROUP BY Reciept_Id,  Customer_ID,Staff_ID
)
SELECT Reciept_Id,Customer_ID,Staff_ID FROM CTE WHERE (RowNumber % 2 = 0)
```

Results of the above query yielded:

Reciept_Id	Customer_Id	Staff_Id
52137	C59	S2
52138	C30	S19
52139	C31	S20
52140	C52	S10
52141	C42	S7
52142	C47	S6
52143	C8	S13
52144	C50	S4
52145	C40	S15
52146	C38	S5
52147	C9	S19
52148	C43	S16
52149	C45	S11
52150	C57	S7

Whereas the original result without a modulo comparison on the row would have yielded a much different result, the raw table supplied in the appendix

With this known, and additional section was added to the python script to generate update statements that would be easy to add to the current migrations.sql script we were prototyping.
The generated update statements appeared as:

```sql
-- Auto-generated query to fix error of type: Staff.Id Mismatch
-- Resolved error identified by UUID: dcf16fba08c63ecc85556c385204d9524ec359cf
UPDATE Assignment1Data
SET Reciept_Id=(
SELECT MAX(Reciept_Id)+1
FROM Assignment1Data)
WHERE Reciept_Id=52136
AND Customer_Id = 'C13' AND Staff_Id = 'S4'
GO
```

Determining now potential entries that broke further rules was our next objective. We pursued the idea that entries of receipts could potentially have duplicate items recorded against the ReceiptItem table. A simple script was generated to check our assumptions of this:

```sql
-- Verify that no receipt has duplicate ItemIds and all are unique per order
SELECT *
FROM
(
    SELECT [ReceiptItem].[ReceiptId],
    COUNT([ReceiptItem].[ReceiptId]) AS 'ItemCount',
    COUNT(DISTINCT [ReceiptItem].[ItemId]) AS 'ItemIdCount'
    FROM [ReceiptItem]
    GROUP BY [ReceiptItem].[ReceiptId]) AS SubQuery
WHERE [SubQuery].[ItemIdCount] != [SubQuery].[ItemCount]
ORDER BY [SubQuery].[ReceiptId]
GO
```

This query returned a result of 912 rows out of the total 2514, which we believed was a large amount given the issues identified earlier numbered in only the teens, however on manual inspection of a number of the reported issue records, it was apparent this figure was actually correct.

Given the large task associated with the entries, an additional module was written for generation of SQL in python which resulted in two queries for each duplicate item entry per receipt, the first query updating the total of one of the records to reflect the real item quantity, the later dropping the non-altered entry after the first had been completed.

The script was as follows:

```sql
-- Auto-generated query to fix error of type: Item.Id Duplicate
-- Resolved error identified by UUID: 8b34383524a00eb2097c1c22f870ef2ad104b6b8
UPDATE Assignment1Data
SET [Item_Quantity]=(
SELECT SUM([Item_Quantity])
FROM Assignment1Data
WHERE Reciept_Id=52316
AND Item_ID = 8)
WHERE Reciept_Id=52316
AND Item_ID = 8
AND Item_Quantity = 10
GO

-- Auto-generated query to fix error of type: Item.Id Duplicate
-- Resolved error identified by UUID: 8b34383524a00eb2097c1c22f870ef2ad104b6b8
DELETE FROM Assignment1Data
WHERE Reciept_Id=52316
AND Item_ID = 8
AND Item_Quantity < (
    SELECT MAX([Item_Quantity])
    FROM Assignment1Data
    WHERE Reciept_Id=52316
    AND Item_ID = 8
)
GO
```

Having now cleaned what we believed to be all discrepancies, we could finally start to look at evaluating data, our analysis outlined in base analysis

4.2 Assumptions and Reasoning

4.2.1 Item Table

An assumption of the ItemId never needing to be larger than a smallint was followed, as a basic query into the maximum range within the test data suggested that the maximum Id that currently existed was 30:

```sql
-- Some basic queries for us to determine potential outlier data:
-- What is the max of each column where datatype is int?
SELECT MAX(Item_ID) AS 'Max Item_ID'
FROM Assignment1Data;
```

With the results:

ItemDescription underwent some size optimisation, as the max data length that currently existed within the supplied data was 52, and we are to assume that into the future more items may be added, a value of 255 should allow for a varied range of descriptions.
SQL queried to determine to above assumption:

```sql
-- Determine current max varchar used in Item_Description
SELECT MAX(DATALENGTH(Item_Description))
FROM Assignment1Data;
```

We do recognise the requirements for optimisation may not require such measures, and acknowledge that a varchar(max)/text datatype would also be reasonable.

4.2.2 Price Table

The price table was designed to hold historical data as required by the business rules, an effective range can be used here to determine item pricing for time frames, current items having no end date or an end date as some point in time into the future.

Accuracy on the pricing was important, we decided to use a decimal(19,5) structure to ensure no problems should arise at any point with calculation of totals. [1]
Another notable feature of the price table is the relationships with both item and receiptItem, which allows the receiptItem table to point at a price value that can either be current or historical in nature.

4.2.3 ReceiptItem

The receipt item table acts as a line-item style associative entity, the quantity and historical priceId used at time of transaction can allow an item's price to be updated and still maintain historical pricing associated with the receipt.

As noted above, to facilitate historical value lookup, this table also holds relationships with the price table.

4.2.4 Receipt

The receipt table acts as a meta-table in this instance, other tables associate with this table with the receiptId field. Due to this it made it extremely easy to use a number of joins/inner joins to determine some of the metrics outlined in the base analysis.

4.2.5 Staff

Staff was left in a non-normalised state to ensure efficiency of queries into the future, normalising the table further would yield little value to the business based on the requirements. The office table is referenced by the staff table. This is merely to satisfy the assumption that, while the only office to exist was Newcastle in this setting, the requirement of more offices into the future is a possibility and the required join would be little impact on speed of queries in a datamart.

4.2.6 Customer

Customer, just like staff could be normalised further requiring more joins and potentially causing a performance issue into the future, for simplicity we kept only the supplied data in mind, and assumed no more data would be required by the datamart into the future.

5 Base Analysis

5.1 Notes on Analysis

As a group, we emphasised identifying and avoiding bad data as our top priority for the analysis outlined in this report. One of the major steps taken in this process included the removal of duplicate items on receipts and consolidating them into single line items. The reasoning behind this is that we discovered a number of receipts that included redundant entries which showed double or, in rare cases, triple the number of items expected which would falsely inflate the cost of items on a receipt.

5.2 Raw Results

A number of metrics were considered to satisfy the request related to the best salesperson, as we are not certain if this is determined by a specific metric or a set of metrics we included a number of analyzed points for the project:

- Total receipts attributed to a staff member
- Total items sold by a staff member
- Ratio of discounted sales to normal sales for each staff member
- Total sale value per staff member
- Average sale value per staff member
- Average item value per staff member

5.2.1 Total Number of Sales

The total number of sales per staff member were considered with the following sql query:

```sql
-- Sales count per staff member (Receipt Count)
SELECT COUNT(*) AS 'Sales Count', s.StaffId,s.StaffFirstName,s.StaffSurname
FROM Receipt r
INNER JOIN ReceiptItem ri ON r.ReceiptId = ri.ReceiptId
INNER JOIN Item i ON i.ItemId = ri.ItemId
INNER JOIN Price p ON p.PriceId = ri.PriceId
INNER JOIN Staff s ON s.StaffId = r.ReceiptStaffId
GROUP BY s.StaffId,s.StaffFirstName,s.StaffSurname
ORDER BY 'Sales Count' DESC;
```

Leading to a range of 700 to 478, the top five staff were:

Sales Count	StaffId	StaffFirstName	StaffSurname
700	S17	Daniel	Baker
682	S19	Kaitlyn	Ortiz
676	S8	Michelle	Miller
664	S5	Stephanie	Watson
664	S6	Evan	Hill

5.2.2 Total Items Sold

The total items attributed to each staff member were considered also, determined by the query:

```sql
-- Item count per staff member
SELECT SUM(ri.ReceiptItemQuantity) AS 'Item Count', s.StaffId,s.StaffFirstName,s.StaffSurname
FROM Receipt r
INNER JOIN ReceiptItem ri ON r.ReceiptId = ri.ReceiptId
INNER JOIN Staff s ON s.StaffId = r.ReceiptStaffId
GROUP BY s.StaffId,s.StaffFirstName,s.StaffSurname
ORDER BY 'Item Count' DESC;
```

Yielding a range of 4217 to 2813, with the top five staff members in this analysis:

Item Count	StaffId	StaffFirstName	StaffSurname
4217	S19	Kaitlyn	Ortiz
4212	S17	Daniel	Baker
4144	S8	Michelle	Miller
4052	S1	Lauren	Martin
4036	S5	Stephanie	Watson

5.2.3 Discounted Sales Ratio

Consideration of the number of sales made by each staff member was also made, the following query yielding the results we required:

```sql
-- Sales metrics for discounted and standard sales per staff member
SELECT s.StaffId,s.StaffFirstName,s.StaffSurname,
SUM(SubQuery.[Discounted Sales]) AS 'Discounted Sales',
SUM(SubQuery.[Standard Sales]) AS 'Standard Sales'
FROM (
    SELECT CAST(
        CASE
        WHEN COUNT(ri.[ReceiptItemQuantity]) >= 5
            THEN 1
        ELSE 0
        END AS int) AS 'Discounted Sales',
    CAST(
        CASE
        WHEN COUNT(ri.[ReceiptItemQuantity]) >= 5
            THEN 0
        ELSE 1
    END AS int) AS 'Standard Sales',
    r.ReceiptId
    FROM Receipt r
    INNER JOIN ReceiptItem ri ON r.ReceiptId = ri.ReceiptId
    INNER JOIN Item i ON i.ItemId = ri.ItemId
    INNER JOIN Price p ON p.PriceId = ri.PriceId
    GROUP BY r.ReceiptId
) AS SubQuery
INNER JOIN Receipt r ON SubQuery.ReceiptId = r.ReceiptId
INNER JOIN ReceiptItem ri ON r.ReceiptId = ri.ReceiptId
INNER JOIN Staff s ON s.StaffId = r.ReceiptStaffId
GROUP BY s.StaffId,s.StaffFirstName,s.StaffSurname
```

Results from the query yielded:

StaffId	StaffFirstName	StaffSurname	Discounted Sales	Standard Sales	Discount Rate (%)
S4	Robert	Wood	518	115	81.83%
S14	Noah	Brooks	533	119	81.75%
S1	Lauren	Martin	533	126	80.88%
S16	Jordan	Turner	520	123	80.87%
S15	Bailey	Green	500	124	80.13%
S13	Molly	Carter	527	131	80.09%
S17	Daniel	Baker	556	144	79.43%
S20	Dylan	Hall	505	132	79.28%
S6	Evan	Hill	524	140	78.92%
S10	Jonathan	Jenkins	454	123	78.68%
S5	Stephanie	Watson	520	144	78.31%
S18	Megan	James	508	142	78.15%
S19	Kaitlyn	Ortiz	531	151	77.86%
S7	Molly	Jackson	474	141	77.07%
S9	Mélissa	Garcia	489	147	76.89%
S8	Michelle	Miller	509	167	75.30%
S12	Leah	Harris	356	122	74.48%
S11	Gavin	Thompson	395	137	74.25%
S2	Joseph	Reed	447	160	73.64%
S3	Amber	Hill	396	168	70.21%

5.2.4 Total Sales Value per Staff Member

Consideration of the total sales per staff member was considered a highly important metric to consider also, we did consider comparing the results of this to the results of a query that did not include discount to see whom would be considered the best performer if discounts were not relevant, however we also recognise this to be too speclutive in nature. The required query was as follows:

```sql
-- Sales total per staff with discounts applied ($)
SELECT CAST(
        CASE
        WHEN COUNT(ri.[ReceiptItemQuantity]) >= 5
            THEN SUM(p.[Price] * ri.[ReceiptItemQuantity]) * 0.85
        ELSE SUM(p.[Price] * ri.[ReceiptItemQuantity])
        END AS decimal(19,5)) AS 'Sales Totals',
        s.StaffId,s.StaffFirstName,s.StaffSurname
FROM Receipt r
INNER JOIN ReceiptItem ri ON r.ReceiptId = ri.ReceiptId
INNER JOIN Item i ON i.ItemId = ri.ItemId
INNER JOIN Price p ON p.PriceId = ri.PriceId
INNER JOIN Staff s ON s.StaffId = r.ReceiptStaffId
INNER JOIN Customer c ON c.CustomerId = r.ReceiptCustomerId
GROUP BY s.StaffId,s.StaffFirstName,s.StaffSurname
ORDER BY 'Sales Totals' DESC;
```

Resulting in the following results:

Sales Totals	StaffId	StaffFirstName	StaffSurname
78572.21500	S8	Michelle	Miller
73847.10750	S19	Kaitlyn	Ortiz
72764.88750	S17	Daniel	Baker
71699.66750	S14	Noah	Brooks
70514.68250	S3	Amber	Hill
69182.56250	S2	Joseph	Reed
69051.19500	S13	Molly	Carter
68831.81000	S5	Stephanie	Watson
68133.83250	S4	Robert	Wood
66267.19000	S6	Evan	Hill
65018.37000	S10	Jonathan	Jenkins
64002.06750	S1	Lauren	Martin
63760.66750	S16	Jordan	Turner
62304.70250	S18	Megan	James
61862.44750	S9	Mélissa	Garcia
61832.27250	S15	Bailey	Green
61536.85500	S20	Dylan	Hall
58996.16250	S7	Molly	Jackson
52102.66250	S11	Gavin	Thompson
50259.35250	S12	Leah	Harris

5.2.5 Average Value Per Sale

The average receipt value per staff member was another metric we considered would add value to the descision to be suggested in the executive summary. The required query to determine this metric was as follows:

```
-- Sales average per staff with discounts applied
SELECT (CAST(
        CASE
        WHEN COUNT(ri.[ReceiptItemQuantity]) >= 5
            THEN SUM(p.[Price] * ri.[ReceiptItemQuantity]) * 0.85
        ELSE SUM(p.[Price] * ri.[ReceiptItemQuantity])
        END AS decimal(19,5)) / COUNT(r.ReceiptId)) AS 'Sales Average',
        s.StaffId,s.StaffFirstName,s.StaffSurname
FROM Receipt r
INNER JOIN ReceiptItem ri ON r.ReceiptId = ri.ReceiptId
INNER JOIN Item i ON i.ItemId = ri.ItemId
INNER JOIN Price p ON p.PriceId = ri.PriceId
INNER JOIN Staff s ON s.StaffId = r.ReceiptStaffId
GROUP BY s.StaffId,s.StaffFirstName,s.StaffSurname
ORDER BY 'Sales Average' DESC;
```

Outlier: We recognise that Amber Hill (S3) has the highest Average Sale Total, this is backed up by his very high Item sales count, showing she is making more sales per Receipt on average than any other sales Officer. However She has not made as much revenue as some other employees, about $7.5k behind the most sales at $78,572. We suspected that maybe she was a new employee but after looking at her sale receipt dates we can confirm this is inacurate and that she was working within the business throughout the entirety of 2017. After learning that she is not a new employee we can rule the possibility of her being the best salesperson out. The data suggests that she has been selling higher cost items to make up the lack of transactions shes involed in compared to others. This is negligable however since she is under performing compared to others anyway.

With results as follows:

Sales Average	StaffId	StaffFirstName	StaffSurname
125.0260328014184397	S3	Amber	Hill
116.2310872781065088	S8	Michelle	Miller
113.9745675453047775	S2	Joseph	Reed
112.6834835355285961	S10	Jonathan	Jenkins
109.9688151840490797	S14	Noah	Brooks
108.2802162756598240	S19	Kaitlyn	Ortiz
107.6363862559241706	S4	Robert	Wood
105.1450889121338912	S12	Leah	Harris
104.9410258358662613	S13	Molly	Carter
103.9498392857142857	S17	Daniel	Baker
103.6623644578313253	S5	Stephanie	Watson
99.7999849397590361	S6	Evan	Hill
99.1612247278382581	S16	Jordan	Turner
99.0901802884615384	S15	Bailey	Green
97.9373355263157894	S11	Gavin	Thompson
97.2679992138364779	S9	Mélissa	Garcia
97.1199810318664643	S1	Lauren	Martin
96.6041679748822605	S20	Dylan	Hall
95.9287195121951219	S7	Molly	Jackson
95.8533884615384615	S18	Megan	James

We consider this to be a metric which weighs heavily in our analysis, as multiple factors would impact this result, the number of items on the sale (resulting in a lower total if discount was applied). Another consideration for this metric would be that it leans towards anyone who could sell a larger quantity of the same item, as this lends itself towards a higher receipt total.

We see that Ms Amber Hill (S3) has the highest average sale total and this is backed up by her high item sales count, indicating she is making more sales per receipt on average than any other sales office. However, Ms Hill has not made as much revenue as some other employees, approximately \$7,500 behind the sales leader who achieved \$78,572. Originally, we suspected that perhaps Ms Hill was a new employee but reviewing sale receipt dates we can confirm this was not a valid assumption and that she was working within the business throughout the entirety of 2017. After this revelation, we can rule Ms Hill out of our assessment for the best sales person. The data suggests that Ms Hill sells higher cost items which makes up for the lack of transactions she completes when compared to other staff members.

6 Conclusion and Recommendations

This report is designed to identify the best performing Sales Officer at BIA Inc by directive of the Head Sales Executive of BIA Inc. After considering multiple data points from the newly created Sales Database, using sales data from 2017 supplied to us by the Head Sales Executive of BIA Inc. We conclude that the Best Sales Officer is Mrs Michelle Miller (S8) becasue she has a sales total of $78,572.22 more than $4k more than the second highest revenue maker Mrs Kaitlyn Ortiz (S19). She does not have the Highest transaction count however because she has sold more higher value items this makes up for it. She has sold 4144 compared to the highest count of 4217. It works out to be a 73 item difference. We beleive that the in excess of $4k extra that Mrs Michelle Miller brings to the company out weighs the value of selling 73 extra items. There for we recommend Mrs Michelle Miller for the reward as the best Sales Officer at BIA Inc. In the event that Mrs Michelle Miller is not applicable we would reccomend Mrs Kaitlyn Ortiz whom has achieved a high evaluation from all the data points that we analyzed.

References

[1] Reasons against TSQL Money type: Stackoverflow User; *SQLMenace* https://stackoverflow.com/questions/582797/should-you-choose-the-money-or-decimalx-y-datatypes-in-sql-server

[2] Microsoft TSQL documentation of Decimal/Numeric types https://docs.microsoft.com/en-us/sql/t-sql/data-types/decimal-and-numeric-transact-sql?view=sql-server-2017

[3] Microsoft documentation: WITH common_table_expression (Transact-SQL) https://docs.microsoft.com/en-us/sql/t-sql/queries/with-common-table-expression-transact-sql?view=sql-server-2017

[4] Upselling - Business Dictionary http://www.businessdictionary.com/definition/upselling.html

7 Appendix

7.1 CTE Raw Results

Reciept_Id	Customer_Id	Staff_Id
52137	C27	S4
52137	C59	S2
52138	C29	S13
52138	C30	S19
52139	C3	S5
52139	C31	S20
52140	C38	S4
52140	C52	S10
52141	C24	S19
52141	C42	S7
52142	C46	S8
52142	C47	S6
52143	C51	S17
52143	C8	S13
52144	C11	S10
52144	C50	S4
52145	C21	S8
52145	C40	S15
52146	C38	S16
52146	C38	S5
52147	C40	S18
52147	C9	S19
52148	C26	S8
52148	C43	S16
52149	C10	S19
52149	C45	S11
52150	C15	S10
52150	C57	S7

7.2 Python Script