Data mart:  
A diagram of a company

Description automatically generated

The structure of the data mart is being designed in a star schema approach. The design rationale is to place a strong emphasis on simplicity and performance to allow for ease of use while maintaining high performance. This design would allow data analysts to obtain business insights into Bits&Bobs overall performance based on different locations and to give feasible solutions to Jessica Smotherington based on the analysis.

Fact table (Sales):

The fact table (sales) is designed to include the most appropriate attributes to derive meaningful insights into Bits & Bob's overall performance to answer Jessica Smotherington questions. The fact table (sales) consists of:

Primary and Foreign keys:

* salesId: the salesId serves as a surrogate primary key and is a system-generated identifier with no business meaning like the natural keys. This would allow data analysts to identify each row with ease. (Surrogate primary key)
* customerId: Customer who is associated with the sale transaction row. (Foreign key relationship with customer dimension table)
* date: The date of purchase in the sales transaction row. (Foreign key relationship with Time dimension table)
* officeId: The location ID of where the sales transaction row took place. (Foreign key relationship with Office dimension table)
* staffId: The staff who handled the sales transaction in the row. (Foreign key relationship with Staff dimension table)
* itemId: The item ID sold in each sales transaction row. (Foreign key relationship with Item dimension table)

Metrics:

* receiptId: We have decided to place receiptId in the metric area and not a dimension table since there is no additional information associated with the receiptId, unlike the previous data mart.
* Transaction row: The transaction row metric would show the multiple rows associated with each receipt ID.
* Item price: The item price would show the price tag associated with the Item ID.
* Item quantity: The quantity would show how many items are purchased in each transaction row.
* Row total: The row total would show the total cost of each transaction row.
* Discounted row total: The discounted row total would show whether there is a 5% discount on the row total when five or more unique items are purchased in each receipt and if not, it would just place the row total amount.

Dimension table (Customer):

The customer dimension tables contain a compilation of information regarding the customer, which enables the data analysts to review additional information about the particular customer based on the primary key of the table.

The customer dimension table consists of:

* customerId – Primary key
* first name of the customer
* surname of the customer

Dimension table (Time):

The time dimension table contains a compilation of information regarding the date, which enables the data analysts to review additional information about the specific date based on the primary key of the table.

The time dimension table consists of:

* date – primary key
* day
* month
* year

Dimension table (Office):

The office dimension table contains a compilation of information regarding the different locations in the organisation, which enables the data analysts to review additional information about the office based on the primary key of the table.

The office dimension table consists of:

* officeId – primary key
* location

Dimension table (Staff):

The staff dimension table contains a compilation of information regarding the staff in the organisation, which enables the data analysts to review additional information about the particular staff based on the primary key of the table.

The staff dimension table consists of:

* staffId – primary key
* first name of the staff
* surname of the staff

Dimension table (Item):

The item dimension table contains a compilation of information regarding the variety of items sold by the organisation, which enables the data analysts to review additional information about the item based on the primary key of the table.

The item dimension table consists of:

* itemId – primary key
* description

Data anomalies detection:

Clean: Check whether there a null or empty values in the customer first name or surname.

A screenshot of a computer code

Description automatically generated

Clean: Check whether all unique items have multiple item descriptions.

A screen shot of a computer code

Description automatically generated

Clean: Check whether the staff ID contains more than one staff first name and surname

A screenshot of a computer code

Description automatically generated

Clean: Check whether the customer ID contains more than one customer first name and surname

A computer screen shot of a code

Description automatically generated

Clean: Check whether the item price multiple by quantity is equivalent to the row total.

A close-up of a computer screen

Description automatically generated

Clean: Check whether the office id shares a different office location.

A computer code with text

Description automatically generated

The first anomaly detected (Row: 109678): Row 109678 has a null value in the row total.

A screenshot of a computer

Description automatically generated

Query:

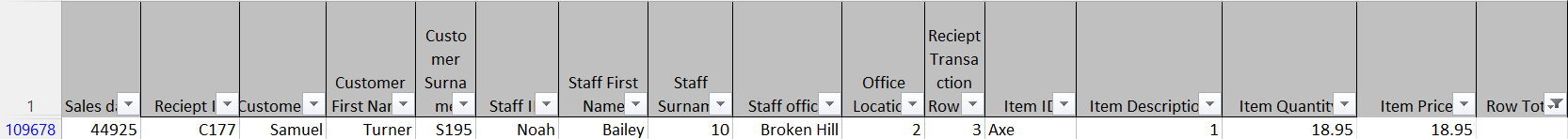
A close-up of a text

Description automatically generated

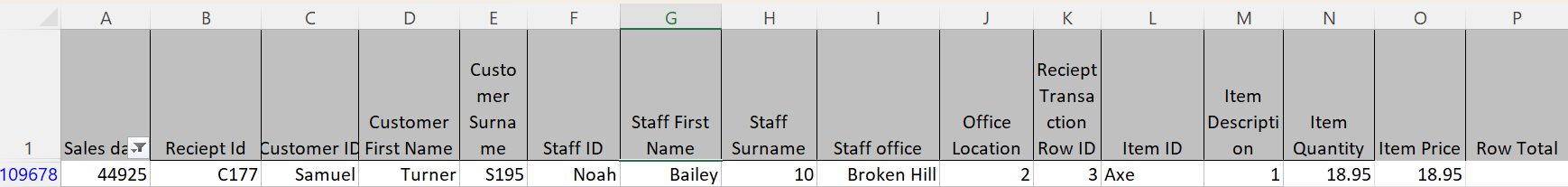
Results:

A close up of a white grid

Description automatically generated



The second detected (row: 109678): Under the sales date column, there is an invalid date format detected.

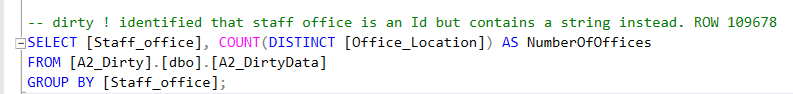


A screenshot of a computer

Description automatically generated

Third anomaly detected (row: 109678): Under the staff office column, it is supposed to be an office ID but instead it is a string named “Broken Hill” and under the office location, it is supposed to have the name of the office location, not the number “2”. Furthermore, “Broken Hill” seems to belong to the staff office ID “10.”

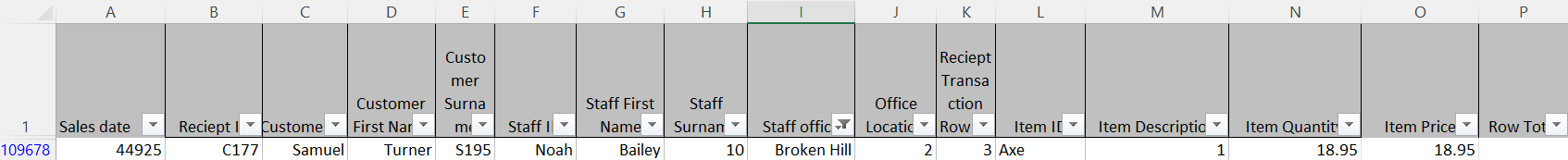
Query:



Results:

A screenshot of a computer

Description automatically generated



Fourth anomaly detected: (Row: 109678):

Under the transaction row, it starts with 3 instead of 1.

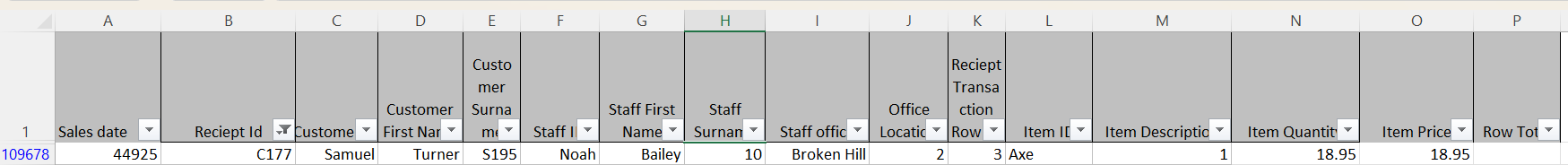
Query:

A screen shot of a computer

Description automatically generated

Results:  
A screenshot of a computer screen

Description automatically generated



A screenshot of a computer

Description automatically generated

Fifth anomaly detected (receipt ID:104312 and 118551):

A screenshot of a computer

Description automatically generated

Receipt ID: 104312 shares two different customer IDs. The customer names are Brandon Bennet and Justin Gray.

A screenshot of a computer

Description automatically generated

Receipt ID: 118551 shares two different customer IDs. The customer names are Christina Ross and Mia Clark.

A screenshot of a computer

Description automatically generated

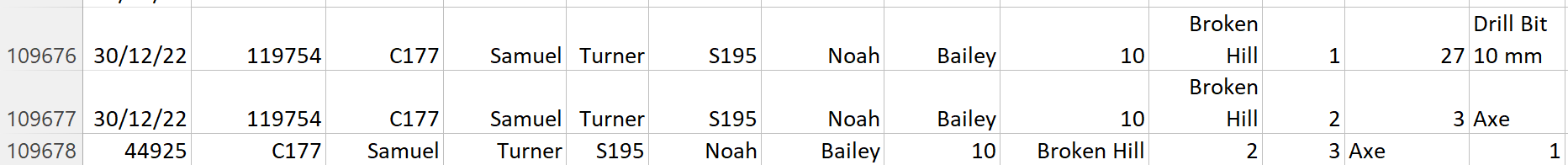
Total anomalies summary and solutions:

Row: 109678 has multiple errors apparently because the values are shifted by 1 due to it missing a sales date, and receipt IDs: 104312 and 118551 share more than one customer ID.

Solutions to fix the erroneous row 109678 anomalies:

1. Omit the entire row.
2. Setting the row 109678 sales date and receipt ID that follows the receipt ID: 119754 and treating it as another transaction row in that receipt since the values of row 109678 have the same customer, staff, and office as receipt: 119754.
3. Placing the latest date in the row and shifting the values down.

We have decided to choose option two as we do not want to omit any row that might affect the analysis. The erroneous row might be insignificant, but we are aiming for precision. We are assuming that the erroneous row 109678 is associated with the previous row as the important values inside row 109677 are almost identical to row 109678. This way, we set row 109678 sales date and receipt ID to follow receipt ID 119754 to become the third transaction row for it, afterwards we will shift the values of each column by one as it appears that the values are shifted by one. This would fix the cause of all errors in row 109678.



Solutions to fix the receipt ID belonging to two customers.

1. Omit the row where the receipt ID appears in another customer ID.
2. Supplement an alphabet to the duplicated receipt ID.

We have decided to go with option 2 as previously mentioned we are aiming for precision. Supplementing an alphabet to the duplicated receipt ID would allow each customer to have a unique receipt ID while avoiding any loss of data that can significantly impact the analysis.

ETL:

This is the data flow of the ETL in SSIS. The files would be imported as a CSV file and will flow through the conditional split where it checks for any empty values and splits them between empty row total and no empty row total flow. The empty row total flow will join with no empty row total flow after it is cleaned to go through another script to fix the duplicated receipt IDS. Finally, the data would be passed to the SQL as “Clean Data.”

A screenshot of a computer program

Description automatically generatedA screenshot of a computer

Description automatically generated

Conditional split to check whether the row total is empty or not empty.

A screenshot of a computer

Description automatically generated

Cleaning script for the erroneous row 109678 to change the structure of its values and to change its receipt ID to follow receipt ID 119754 and date.

A screen shot of a computer program

Description automatically generated

Remove duplication of receipt ID script to add a “D” to duplicated receipt IDs for rows: 104312 and 118551.

A computer screen with text

Description automatically generated

A computer screen shot of a program code

Description automatically generated

Datamart implementation:

After the data is cleaned, the implementation of the data mart is created and populated.

Time dimension table:

A screenshot of a computer

Description automatically generated

Customer dimension table:

A screenshot of a computer

Description automatically generated

Office dimension table:

A computer code with text

Description automatically generated

Staff dimension table:

A computer code with text

Description automatically generated with medium confidence

Item dimension table:  
A computer code with text

Description automatically generated with medium confidence

Fact table (Sales):  
A computer screen shot of a computer code

Description automatically generated

A screenshot of a computer program

Description automatically generated

SQL-generated database diagram

A screenshot of a computer

Description automatically generated