**GSCM 521**

Abdi Alamir

**Analysis Assignment 2 - Part A**

**Data Cleaning with Power Query in Power BI:**

Each file was loaded individually for processing and transformation in Power Query:

**Cleaning Bills data:**

**1- Removal of Unnecessary Columns:**

* Excluded the initial column, primarily serving as an index, to enhance data clarity.
* Excluded “Row Expiration date” and “Current row indicator” as they don’t have an added value for the analysis.

**2- Date Variable Reformatting**

Ensured the correct formatting of date variables such as “Row Effective Date”

**3- Remove Duplicate rows:**

-Eliminated any duplicate entries ensuring data integrity.

**4-Checking for Missing Data**

Missing entries are observed in the variable payment reference, so I added a new column where I replaced the blanks with zeros following the assumption mentioned below.

No missing entries are observed within the other variables.

**Cleaning Customer data:**

**1-Removal of Unnecessary Columns:**

* Excluded the initial column, primarily serving as an index, to enhance data clarity.
* Excluded any duplicate variables, such as Doctor Key and NIF as they are the same as “customer\_ref\_id”.
* Excluded “Row Expiration date” and “Current row indicator” as they don’t have an added value for our analysis.

**2- Date Variable Reformatting**

**-**Ensured the correct formatting of date variables such as “Row Effective Date”

**3- Removed Duplicate rows:**

-Eliminated any duplicate entries ensuring data integrity.

**Cleaning Doctors data:**

**1-Removal of Unnecessary Columns:**

* Excluded the initial column, primarily serving as an index, to enhance data clarity.
* Excluded any duplicate variables, such as Doctor Key and NIF as they are the same as “doctor\_ref\_id”.
* Excluded “Row Expiration date” and “Current row indicator” as they don’t have an added value for my analysis.

**2- Date Variable Reformatting**

**-**Ensured the correct formatting of date variables such as “Row Effective Date”

**3- Removed Duplicate rows:**

-Eliminated any duplicate entries ensuring data integrity.

**Cleaning Payment Methods data:**

No cleaning is required for this small amount of data, I have only excluded the initial column, primarily serving as an index, to enhance data clarity.

**Cleaning Stores data:**

**1-Removal of Unnecessary Columns:**

* Excluded the initial column, primarily serving as an index, to enhance data clarity.
* Excluded any duplicate variables, such as Doctor Key and NIF as they are the same as “store\_ref\_id”.
* Excluded “Row Expiration date” and “Current row indicator” as they don’t have any added value for the analysis.

**2- Date Variable Reformatting**

**-**Ensured the correct formatting of date variables such as “Row Effective Date”

**3- Removed Duplicate rows:**

-Eliminated any duplicate entries ensuring data integrity.

**Assumption:**

I will assume that if the payment method is not specified that it is null in the Bills data in “payment\_ref\_id”, and it is considered as Cash.

**Individual Data Exploration:**

**1-Exploring Bills Data:**

A graph of blue bars

Description automatically generated with medium confidence

A graph of blue rectangular bars

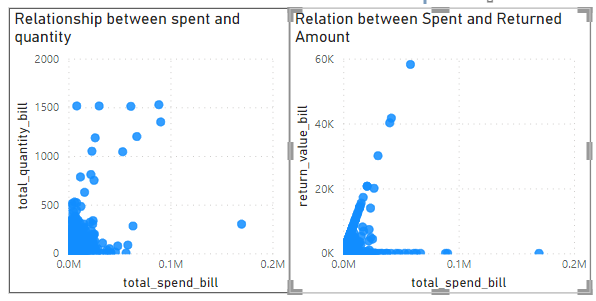
Description automatically generated with medium confidence

A graph showing a number of different times

Description automatically generated with medium confidence

A graph with different colored bars

Description automatically generated



**2- Exploring Customers Data:**

A graph with numbers and a black rectangle

Description automatically generated

**3- Exploring Doctors Data:**

A graph with blue lines and black text

Description automatically generated

**4- Exploring Stores Data:**

A graph of sales

Description automatically generated

**Data Linkage:**

**1. Bills Data (bill\_fact.csv):** This dataset contains records of transactions, including the IDs of doctors, stores, customers, and payment references. It serves as the central dataset where transactions between doctors, stores, and customers are recorded.

**2. Customers Data (dim\_customer.csv):** This dataset contains information about customers, which includes attributes like customer ID, name, contact information, and demographic details.

**3. Doctors Data (dim\_doctor.csv):** This dataset contains information about doctors, such as Doctor ID, name, specialization, contact details, and possibly other professional information.

**4. Store Data (dim\_store.csv):** This dataset provides information about stores or locations where transactions take place. It may include store ID, name, address, contact details, and other relevant information.

**5. Payment Methods Data (dim\_payment\_method.csv):** This dataset contains information about payment methods accepted or used in transactions. It includes payment method ID, and description.

**-Joining Bills Data with Customers Data:** I can use the `customer\_ref\_id ` in the bills data to link with the customers data (`customer\_ref\_id `). This linkage allows us to enrich transaction records with customer information, such as demographics or contact details.

**- Joining Bills Data with Doctor Data:** Similarly, I can use the `doctor\_ref\_id` in the bills data to link with the doctor’s data (`doctor\_ref\_id`). This linkage helps in understanding which doctors are associated with which transactions.

**- Joining Bills Data with Store Data:** I can use the `store\_ref\_id` in the bills data to link with the store data (`store\_ref\_id` ). This linkage helps in understanding which stores the transactions took place in.

**- Linking Payment Methods Data:** The new created column, `pay\_ref\_id` in the bills data can be used to link with the payment methods data (`payment\_ref\_id`, providing details about the payment methods used in each transaction.

A screenshot of a computer

Description automatically generated

**Business Questions:**

**1. Business Question: What is the geographical distribution of our sales and which regions contribute the most to our revenue?**

- Create interactive maps or bar charts showing sales performance by region.

- Use slicers to allow stakeholders to filter sales data by different time periods (e.g., month, quarter, year) to identify seasonal trends or changes in regional performance.

**2. Business Question: How do different payment methods compare in terms of usage and total spend?**

- Create visualizations such as pie charts or bar graphs to illustrate the distribution of payment methods used in transactions.

- Calculate the total spend associated with each payment method.

- Provide insights into trends in payment method usage over time and identify any correlations between specific payment methods and customer spending behavior.

**3. Business Question: Who are our top-performing doctors in terms of the number of drugs prescribed and total spend generated?**

- Analyze the number of drugs prescribed and total spend associated with each doctor.

- Create visualizations such as tables or bar charts ranking doctors based on prescription volume and revenue generated.

- Implement drill-through functionality to allow stakeholders to explore detailed transaction data associated with each doctor, including patient demographics and types of drugs prescribed.