**DATA CLEANING AND TRANSFORMATION**

**Summary Table:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Datasets | Accuracy | Completeness | Consistency | Currency | Relevancy | Validity |
| **Customers** |  |  |  |  | * New/Old customer Column Created |  |
| **Logistics** |  | * Contact email   Blanks   * Phone number   Blanks |  |  | * Air Freight Column   Created |  |
| **Products** | * Inaccurate Price\_per\_unit | * Cost   Blanks   * Price\_Per\_Unit   Blanks |  |  |  |  |
| **Transactions** | * Products\_profit column Missing * Total\_Profit   Missing |  |  |  | * Month and Monshort Column Created * Quarterly Column created * PreJuly/Postjuly   Column created | * Transactions date   Wrong format in some columns |
| **Sales People** |  |  |  |  | * Age Category Column Created |  |

Below are more in depth descriptions of the data cleanings done and methods of mitigation used. Recommendations and explanations have also been included to avoid further data quality issues in future. Following recommendations can improve the accuracy of the data and can be used to influence business decisions in the future. The Accuracy and completeness issues cleaning and transformation is done in excel.

**ACCURACY ISSUES:**

* **Price\_Per\_Unit was inaccurate for “Products” dataset**
* **Products\_PROFIT, Total\_Profit columns missing for “Transactions” dataset.**

**Mitigation**: The value was too high as 7000000.but most of the other products price under that category were only 7pounds.So,it must have been a type error. So value changed to 7.

**Recommendation:** *Products\_Profit and Total\_Profit column, allowing for more comprehensible data and easier to check for errors.*

*Create Products\_Profit and Total\_Profit Column in “Transactions” table to check accuracy of sales.*

***COMPLETENESS:***

* ***Blanks in Contact\_email and contact\_number in “Logistics” dataset.***
* ***Blanks in Cost and Price\_per\_unit in “Products” dataset.***

***Mitigation:*** *Blanks in contact\_email or contact\_number does not affect our analysis much. Leave it as it is.*

*Blanks in cost and Price\_per\_unit needs to be fixed by taking the mean/average of that particular category of product and filled up.*

***Recommendation****:*

*Blanks are treated as incomplete data and can skew further analysis results. The addition of not able to submit the form with blanks as an option will allow to have more complete data and will result in more accurate analysis.*

***The following Columns were created in Power Bi using Power Query:***

***RELEVANCY:***

* ***New/Old Customers Column created in “Customers” dataset.***
* ***Air Freight Column created in “Logistics” dataset.***
* ***Month and MonShort columns are created in “Transactions” dataset.***
* ***Quarterly Column and PreJuly/PostJuly Column Created in “Transactions” dataset.***
* ***Age category column was created in “Sales People” dataset.***

***Mitigation:***New/Old Customers were created to get the count of the new and old customers.

Air freight column is created for checking the freights which are sent by air and not by air.

Month, MonShort, Quarterly Columns are created for monthly and quarterly analysis.

PreJuly/Post July columns are created for checking the reduction of air transport Postjuly period as per the requirement of the management.

***Recommendation:***All these columns are needed and very useful for data analysis of our visuals.

***VALIDITY:***

* ***Format of some of the rows of Transactions date column was wrong in “Transactions” dataset.***
* ***Mitigation:*** *Split the date column by delimiter and merge the columns according to “DDMMMYYYY” format and change the format to short date format.*

***Recommendation:*** *Set up columns so that formats such as “DDMMYYYY” are already in place while entering new data.*

*Allowable values will make data to be interpreted more easily. Formatting all dates in one format consistently will increase readability. This will reflect positively on speed and accuracy of analysis for business decisions.*

***MEASURES CREATED:***

*1) Sales*

*2) Expenditure*

*3) Income*

*4) Revenue*

***TABLES CREATED:***

*1)KPI Calculator*

*2) Profitable\_Products\_Index*

*3) Customer\_Index*

*4) Sales\_Revenue\_Index*

*All the above tables are created to calculate and find out the best 5 ranking Sales Agents for the management to appoint Sales Head at their own discretion.*

*Sales\_Revenue\_Index Parameter is calculated based on amount of sales and Revenue of each Sales Agent. Its maximum points is 55points with weightage of 42 %( total 130 points).*

*Profitable\_Products\_Index Parameter is calculated based on the maximum percentage of revenue of each sales Agent under Most Profitable Products parameter. Its maximum points is 50 with weightage of 38 % (total 130 points).*

*Customer\_Index Parameter is calculated based on retaining Old customers and acquiring New customers. Its maximum points is 25with weightage of 20 %*

*( total 130 points).*

*KPI\_Calculator Parameter is the Key Performance Index which measures total 3 parameters.*

*Then we created two Powerbi Dashboards with required visuals as per the management, one for the Shareholders point of view and the other for the Stakeholders point of view.*