KLE Society's
KLE Technological University, Hubballi.

A Minor Project Report

on

Production and Freight Analysis Using Power BI for Dana

submitted in partial fulfillment of the requirement for the degree of

Bachelor of Engineering

in

Computer Science and Engineering

Submitted by

Sonali Kabadi	01FE18BCS214
Ayazahemad Shaikh	01FE18BCS058
Ankita Patil	01FE18BCS039
Akshay Kulkarni	01FE18BCS026

Under the guidance of
Ms.C V Soumya, Asst. Prof

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

Hubballi – 580 031

2020 -21

KLE Society's
KLE Technological University, Hubballi.

2020 - 2021

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that Minor Project “**Production and Freight Analysis Using Power BI for Dana**“ is a bonafied work carried out by the student team comprising of **Sonali Kabadi - 01FE18BCS214, Ankita Patil - 01FE18BCS039, Akshay - Kulkarni 01FE18BCS026, Ayazahemad Shaikh - 01FE18BCS058** for partial fulfillment of completion of sixth semester B.E. in Computer Science and Engineering during the academic year 2020-21.

Guide	Head, SoCSE
Ms.C V Soumya	Dr. Meena S. M

Viva -Voce:

Name of the Examiners	Signature with date
1.	
2.	

Acknowledgement

We would like to thank our faculty and management for their professional guidance towards the completion of the project work. We take this opportunity to thank **Dr. Ashok Shettar**, Vice-Chancellor, **Dr. N.H Ayachit**, Registrar, and **Dr. P.G Tewari**, Dean Academics, KLE Technological University, Hubballi, for their vision and support.

We also take this opportunity to thank **Dr. Meena S. M**, Professor and Head, SoCSE for having provided us direction and facilitated for enhancement of skills and academic growth.

We thank our guide **Ms.C V Soumya**, Assistant Professor, SoCSE for the constant guidance during interaction and reviews.

We extend our acknowledgement to the reviewers for critical suggestions and inputs. We also thank Project Co-ordinator **Dr. Sujatha C**. and faculty in-charges for their support during the course of completion.

We express gratitude to our beloved parents for constant encouragement and support.

Abstract

Dana Incorporated is an American supplier of axles, driveshaft, transmissions and electrodynamics, thermal, sealing, and digital equipment for conventional, hybrid, and electric-powered vehicles. The company has been analysing its production using traditional methods. We aim to build a dashboard which can make the analysis more interactive and detailed using Power BI, along with Visual interpretation of the raw data with the help of statistical graphics, plots, information graphics and other tools. Data analytical tool such as Power BI is used to provide visual interpretation of the data with the help of statistical graph. This process simplifies the complexity of the data analysis and provides user friendly interface to understand the business statistics. Production and Loss Analysis, Freight Prediction and Freight Analysis will provide insights on how the company performs over a period of time, along with reasons for the loss incurred. To achieve this, we have implemented few Machine Learning models. For the Freight analysis, the number, of trips took by transporter to deliver the goods to particular customers and the cost for the same is calculated and visualized.

The client can add data of each month and refresh the dashboards to visualize and take further decisions. This visual representation provides key metric to make on the go decisions and the model predicts the supply chain cost.

We have build a prototype to achieve the above mentioned objectives.

Keywords*: Business Intelligence, Dashboard, Freight, Power BI*

Contents

List of Figures

1 INTRODUCTION

Dana Incorporated is an American supplier of axles, driveshaft, transmissions and electrodynamics, thermal, sealing, and digital equipment for conventional, hybrid, and electric-powered vehicles. The company's products and services are aimed at the light vehicle, commercial vehicle, and off-highway equipment markets. Found in 1904, it is based in Maumee, Ohio, the company employs nearly 36,000 people in 33 countries. In 2019, Dana generated sales of 8.6 billion dollars. The company is included in the Fortune 500. The company has been analysing it's production using traditional methods. We aim to build a dashboard which can make the analysis more interactive and detailed using Power BI.

1.1 Motivation

Data is one of the most valuable assets owned by any business. Managing and analysing the data is a key challenge faced by all modern businesses. Since the data is scalable, it becomes very difficult to interpret the data at any given level. In order to simplify this complexity, we aim to present the data more user friendly using data analytical tool known as Power BI

The company wanted an automized prototype to analyse its day-to-day data such as production and fright so that at any given level of management or department they could easily analyse the reasons for loss and take necessary actions. Also this prototype which is dynamic in nature helps the company to track the information of freight information which aids in avoiding any possible discrepancies.

Our approach allows the user to understand their business statistics more accurately. The visualizations that are focused on interpreting the production of the company can easily be drilled down to minute details available from the overall business data.

1.2 Problem Statement

To create a business dashboard using Power BI for visualization of the Production and Freight data generated by the company for 24 hours monitoring.This prototype created is atomised and plays a important role in smart manufacturing pipeline.

1.3 Market Survey

Power BI is a very strong data visualization tool. Different tools like tableau, QlikView, SAP Analytics Cloud, Amazon QuickSight perform similar functionalities to that of the Power BI. Compared with these tools, the main advantage of Power BI is that of expense. All the tools mentioned above have higher expenses thus slowing the reach to Return on Investment(ROI). For small businesses, Power BI is the optimal tool because of the lower expense. The cost can be a major hurdle for most small businesses who want to digitalize to make the work more efficient but they are not affected much by it. Because it is a want and not a need, the cost they are likely to pay must be lower, and as Power BI is less expensive it is preferred.

1.4 Applications

For manufacturing companies, using data shrewdly can mean gaining a decisive competitive edge. In the automotive industry companies usually look to optimise every small detail from production of a part to the delivery of that part. Global uncertainties, especially price wars, and COVID-19 pandemic are resulting in new operational and technological challenges as well as potential opportunities for the automotive sector. Automotive companies have been adopting digital technologies across their processes from product design, procurement, production, supply chain, all the way to sales and marketing.

Various losses occurring in a manufacturing plant have direct bearing on reduced productivity and increased costs. Though some losses like asset failure can be measured using traditional methods, advance systems are required for root cause analysis. Then there are many minuscule losses which are very difficult to measure hence are difficult to keep track of. Their frequency of occurrence can be high and hence their cumulative effect significant. Business Intelligence (BI) solutions can help identify these losses and make the manufacturing unit more accountable in order to increase efficiency.

Smart manufacturing is strongly correlated with the digitalization of all manufacturing activities. This increases the amount of data available to drive productivity and profit through data-driven decision making. The industries are moving at a rapid pace and to keep up they need to digitize records and use the data more efficiently. Giving operations teams the ability to centrally analyse general ledger, production costs can go a long way in making the output of a plant more efficient by the hour.

Automotive companies are always looking for ways to cut costs and maximize spending. This requires extensive analyses. But more analyses are a hard-sell for busy logistics managers. The freight professionals are so preoccupied immediate needs that it's often difficult for them to step back and make thorough analysis of freight costs. Freight Analysis provides an explanation for extra costs which will help the business with time and money in the long run.

Machine learning has a handful of strengths that outperform the way things traditionally get done in the freight industry like assessing a massive amount of information and providing a quick perspective of things. A simple regression model can be used to make a prediction of the freight cost for a particular type of shipment to a particular type of place. The analysis of this cost helps in making crucial changes in price for the benefit of the company.

1.5 Objectives

Create a dashboard for effective and easy visualisation of:

- **Production and Loss analysis**

  We aim to analyse the efficiency i.e Operation Equipment Efficiency for the production of Plants and Value Stream for planned work and actual work done, and visualize the loss occurred per day, per week, per month and year with reason.

- **Freight Prediction**

  We aim to predict the transportation loss using Linear Regression from one location to another based on the factors such as distance, vehicle capacity, terrain, fuel cost. The same to be visualized in dashboard.

- **Freight Analysis**

  We aim to generate virtualization based on freight data produced by another team working under this company. This data needs to filtered out based on the hierarchy level such as customer,location.

1.6 Scope of the project

The production and loss analysis dashboard will help the production manager analyse trends of production over a period of time, it will also help to identify losses and the departments responsible for these losses, thereby increasing accountability.

The freight analysis dashboard will help verify the accuracy of bills submitted, it will also help the company in getting to know how much cost is incurred for freight for each of their customer.

The freight prediction dashboard helps the company to calculate freight price using Machine Learning models running in the background based on features such as distance of freight, transit hours,vehicle capacity.

Power BI is a powerful Business Intelligence tool. Other prominent tools in the market include Tableau, MicroStrategy, QlickView, QlickStart. PowerBI is a highly efficient and user-friendly tool for powerful data analysis and visualisations. It allows the user to amalgamate data from many sources, create interactive dashboards, validate data, create didactic reports and deploy these reports for users to access it remotely. Power BI offers many business intelligence and data analysis tools. Below are a few reasons why you should use Power BI for data analysis and visualisation in a firm.

- Easy-to-use:The end-user is not expected to have any technical skills, to visualise and update the dashboard.
- Interactive visual dashboards:User can drill down a particular visual to gain more information.
- One time deployment and remote access-The dashboard built can be deployed once and can be accessed remotely, which essential in these dire situation due to covid-19

Having in-depth knowledge and hands-on experience in Power BI will help immensely with implementing large-scale project. A career in Power BI is exciting and fast-paced.Power BI's interactive community is growing at a great pace thereby making it a great place to network and also increase the ease of use of Power BI

2 REQUIREMENT ANALYSIS

After interaction with the client, the Software Requirement Specification(SRS) is documented. This will help define the requirements and expectations of the client from the software being developed. This document can also include changes to existing product to match client satisfaction and match the customer demand.

2.1 Functional Requirements

These are the Functional Requirements which are product features that developers must implement to enable users to accomplish their tasks.

- User shall be able to interpret efficiency in the hierarchical manner.
- User shall be able to interpret loss details in the hierarchical manner.
- User shall be able to identify the reason for loss in the hierarchical manner.
- User shall be able to predict the transportation cost.

2.2 Non Functional Requirements

These Non Functional Requirements (NFR's) define system attributes such as security, maintainability, scalability. They serve as constraints or restrictions on the design of the system across the different backlogs.

- Should show graphs with 100 % accuracy
- ML model should predict with at 75-85 % accuracy
- The reason should be displayed only if the loss time is more than 15 minutes.

2.3 Hardware Requirements

Below are the requirements one system should have in order to have business intelligent software, Power BI

- Android Studio supports Microsoft® Windows® 7/8/10 (64-bit)
- Memory (RAM): At least 4 GB available, 8 GB or more recommended.
- 2 GB of available disk space minimum.

2.4 Software Requirements

Below mentioned software are required in order to implement the problem statement.

- Android Studio
  Android Studio is the official Integrated Development Environment (IDE) for Android app development, where applications can be developed for all Android devices.

3 SYSTEM DESIGN

Systems design is a method of defining components of a system like it's architecture, interface etc based on the requirements specified in the Software Requirement Specification(SRS). In the process systems are developed to match specific user requirements.

3.1 Architectural Framework

An architecture framework is an encapsulation of a minimum set of practices and requirements for artifacts that describe a system's architecture. Models are representations of how objects in a system fit structurally in and behave as part of the system.

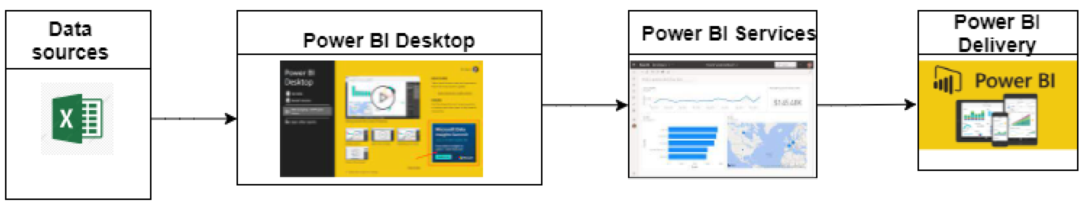

Fig. 1: Architectural Design

The data source is the important component of Power BI.We can import data files from our system,cloud based online data sources.Some of the commonly used data sources in Power BI are:

- Excel
- Text/CSV
- XML
- JSON
- Oracle Database
- IBM DB2 Database
- MySQL Database
- PostgreSQL Database

Power BI Desktop

Power BI Desktop is a client-side. Power BI desktop-based software is loaded with tools and functionalities to connect to data sources, transform data, data modeling and creating reports.

Dashboards

Dashboards are a wonderful way to monitor your business and see all of your most important metrics at a glance.Dashboard contains only the highlights of that story. Readers can view related reports for the details.

3.2 Proposed System

Proposed System comprising of hardware and software is diagrammatically represented in figure 2.

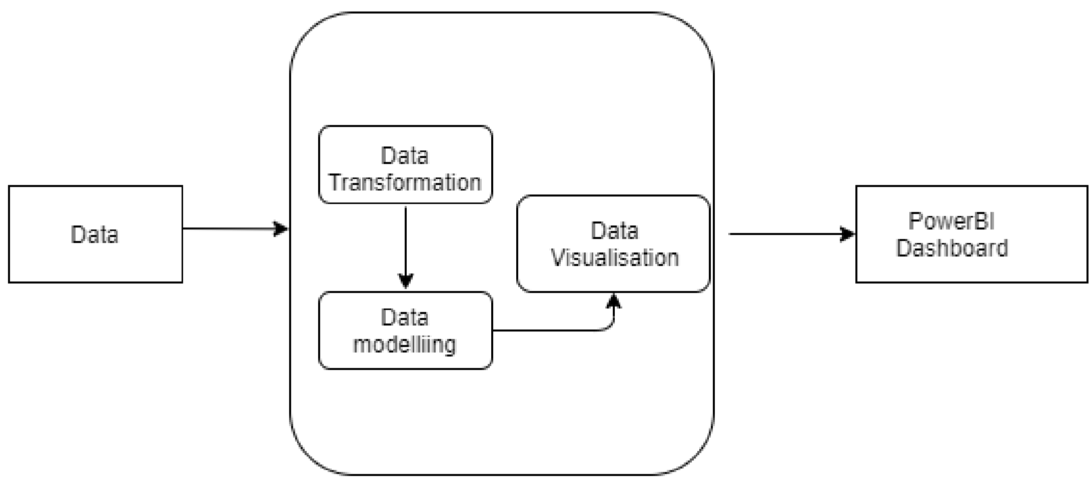

Fig. 2: Proposed System

To create the dashboard, the raw data cannot be directly used to create the visuals. The ETL process has to be carried out, i.e, Extract, Transform and Load. We have received data in excel format which has to be uploaded into Power BI by necessary transformations that include:

- Removing null columns and rows
- Attribute reconstruction - Create new attributes from existing attribute.
- Providing the cardinality between the tables.

Once the transformation is done, the data is loaded into the Power BI. The required visuals are created to give proper inference and finally building the report.

3.3 Flow Chart

Flow chart comprising of flow of data that depicts the computation of algorithm is diagrammatically represented in figure 3.

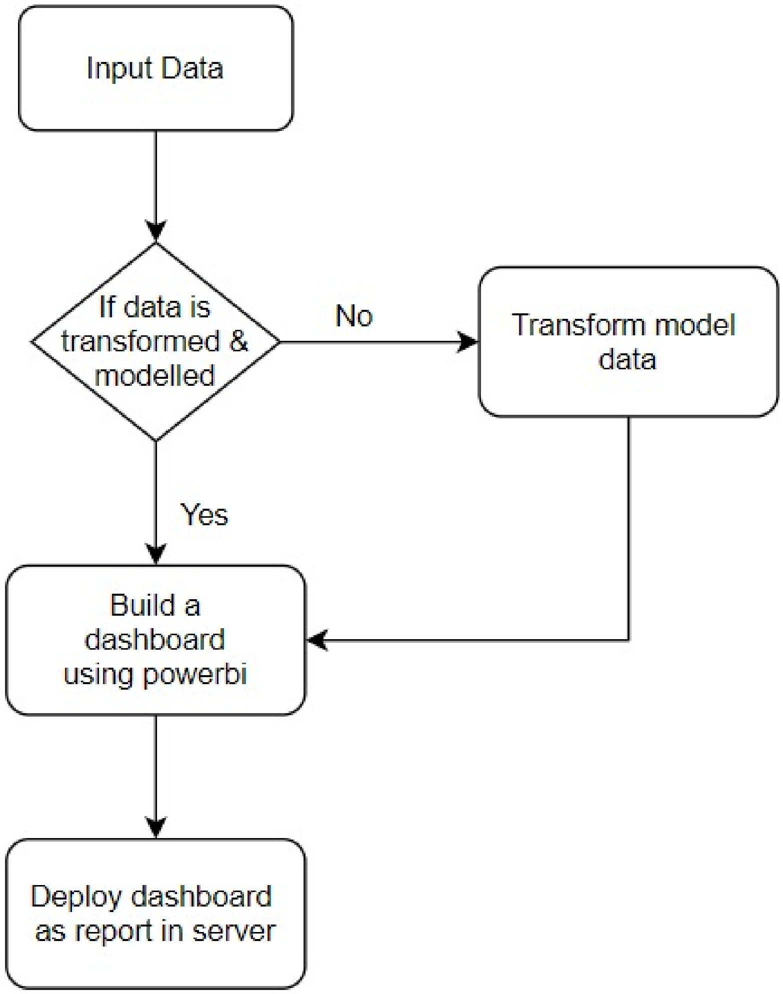

Fig. 3: Flow Chart

As depicted in the flow chart, the data is taken as input and checks whether the data is transformed. If it is, then visuals are created else we follow ETL. Different Visuals created helps build the dashboard that is deployed as a server to the industry organization.

3.4 Design Principles

Software designs are a bridge between the planning phase where requirements are defined and the implementation phase where the requirements are realised. There a few principles which help in structuring the design which can help developers while creating the software.

3.4.1 Design Principle 1

The figure 4 shows the general flow of data throughout the Production and Loss Analysis. The production data given by the industry was unorganized as it was the data that was first taken down manually in a ledger.Since it's real world data, it's unorganised and messy The discrepancies in the data were identified and removed using the Power Query tool. Then attribute reconstruction, transformation, and modeling of tables using basic database topics such as relationships were created among tables to perform dynamic functions for the visualization.

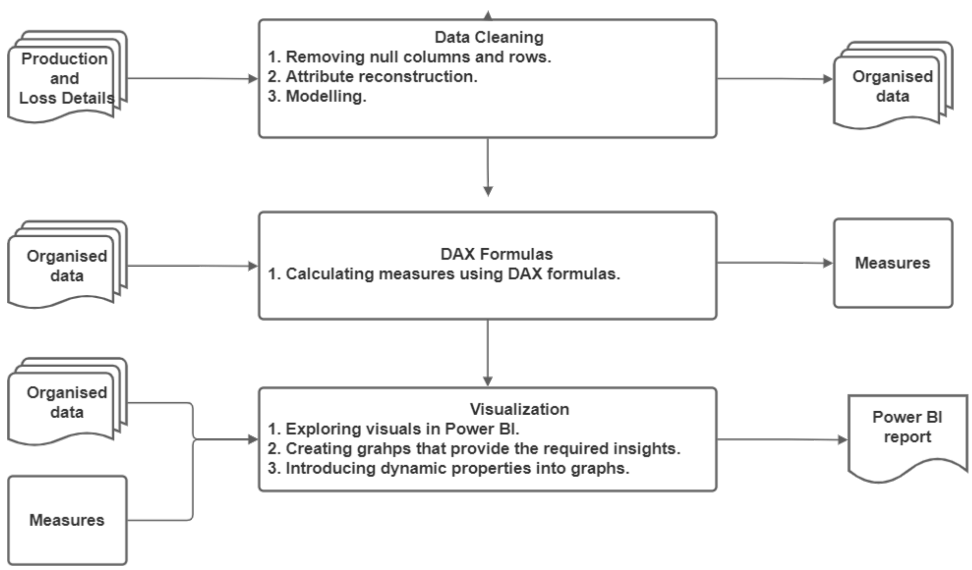

Fig. 4: Detailed Design for Production and Loss Analysis

Data Analysis Expressions(DAX) formulas were applied to the organized data to get meaningful measures out of it to perform the specified task such as calculating Operating Equipment Efficiency(OEE). The measures obtained help in creating powerful visualization which provides meaningful insights to the management brass of the client, helping them take apt decisions based on these insights. The Power BI also has a provision for dynamic graphs which help you drill down methodically to provide more insight into the data.

3.4.2 Design Principle 2

The figure 5 shows the general flow of data throughout the Freight Prediction.Data mining was carried out on the data provided.The data consisted of sizeable number of attributes, in which most of them were not required for the prediction, thus only required columns were selected using correlation matrix. Outliers from the data were removed using a box plot.Comparison of machine learning models was done based on the requirements provided by the company.

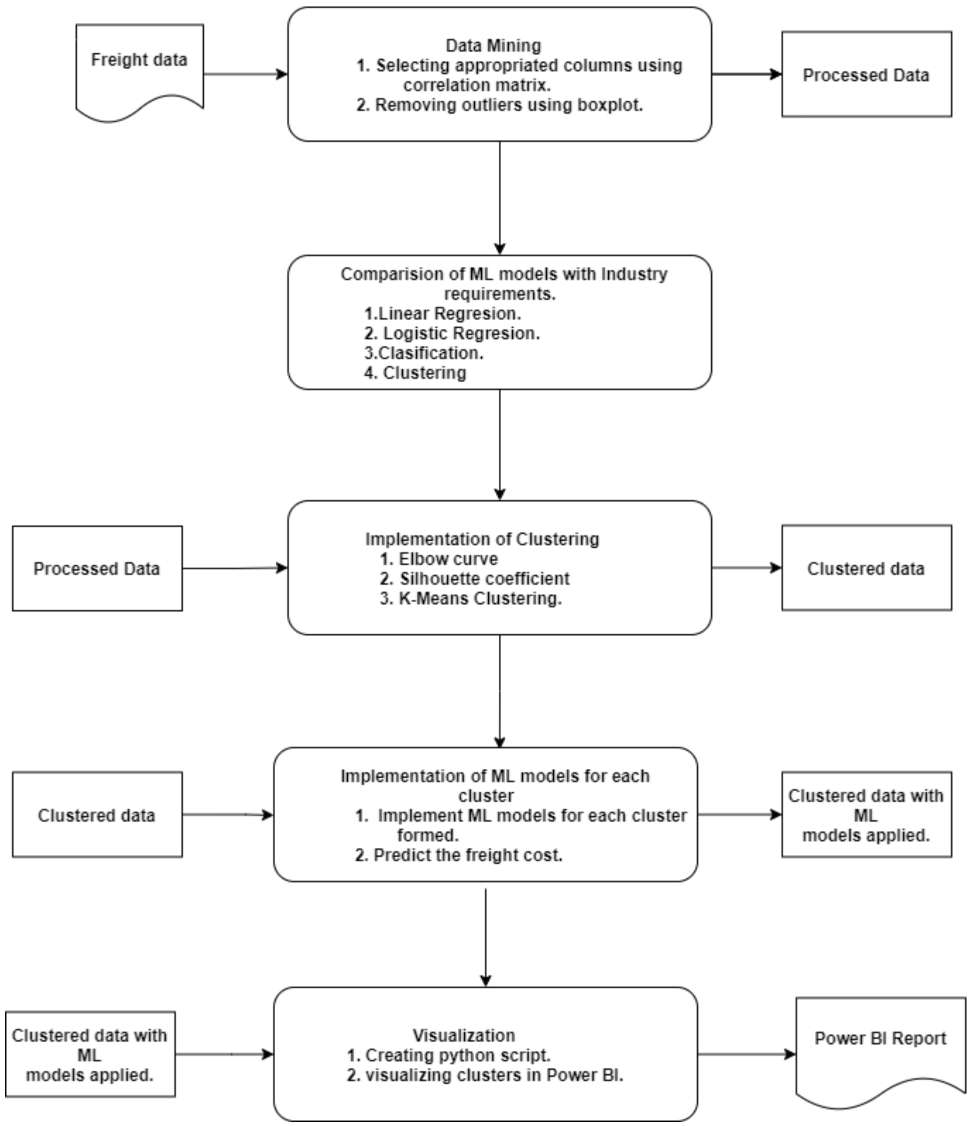

Fig. 5: Detailed Design for Freight Prediction

Clustering was opted for this specific task. Using elbow curve, number of clusters were selected for the prepossessed data and K-means clustering is used to create the clusters.This developed algorithm is used in Power BI to create a report for prediction of price range for the freight, and in turn a visual representation is done to infer the results.

3.4.3 Design Principle 3

The figure 6 shows the general flow of data throughout the Freight Analysis. The DAX formulas are calculated on freight data that shows the number of deliveries done to customers, and cost agreed to pay for transporter for delivering planned and extra trips.

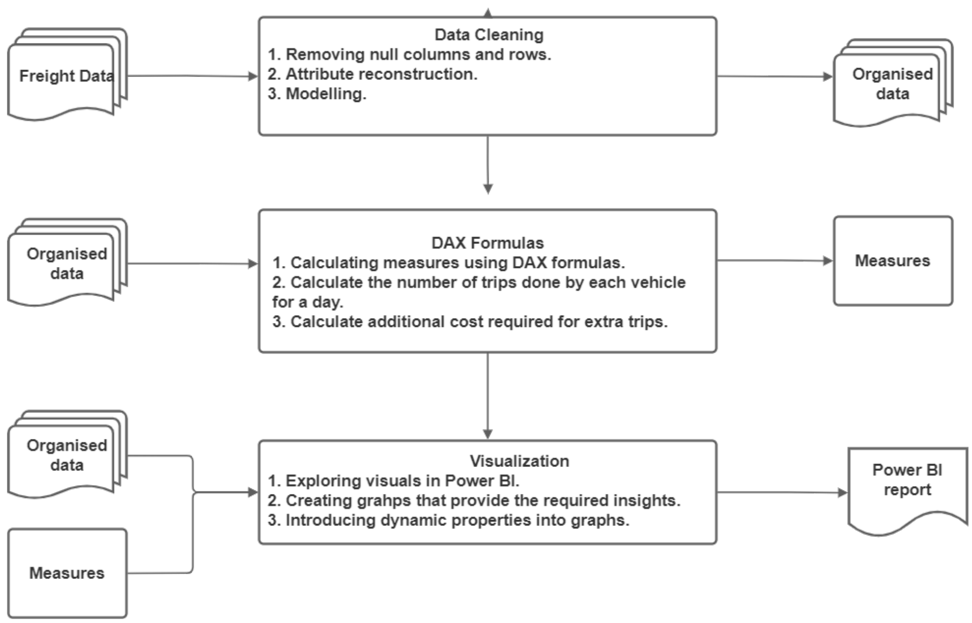

Fig. 6: Detailed Design for Freight Analysis

4 DATASET DETAILS

4.1 Dataset of Production data

The production data is the excel file that contains the details of production of the industry.Below are the attributes of the production file.Value Stream(VSM) is primary key.

- Date
- Part Number - Part number is the part of a vehicle that is produced.
- Trim Number - Trim number is same as that of part number.
- VSM - VSM is value stream that is different collection of process where we get output.
- Type - This attribute indicates whether the information in that respective row is planned to produce or actually produced.
- Quantity - Quantity is the attribute that indicates the number of part numbers either planned to produce or produced.
- Equity Quantity Factor - This is time in minutes required to produce a part number.
- Equity Quantity - This is time in minutes required to produce given number of quantity of the part number. we can also say it a product of quantity and equity quantity factor.

4.2 Loss Details

The loss details contains the amount of time lost in producing the part numbers along with the reason for cause.VSM is primary key.

- Date
- VSM - VSM is value stream that is different collection of process where we get output.
- Time Lost In Min - It is time in minutes where employee didn't work.
- Loss Number - This stays the category of the loss.
- Reason - this attribute says the reason for the occurred loss.

4.3 Freight Prediction Data

This data has information of the freight price.

- Source - Place from where the transportation begins.
- Destination - Place where the goods has to be delivered.
- Vehicle Type - This attribute specifies what type of vehicle is used for transporting certain amount of goods.
- Vehicle Capacity - This attributes says how much amount of goods can the vehicle withstand for transportation.
- Distance in KM - It is the distance between source and destination in kilometers.
- Revised Hrs - It is the time taken to transport the goods from source to destination.
- Diesel Price - It us the diesel price for a km for a particular vehicle type.
- Price - It is cost to carry out the transportation from source and destination.

4.4 Freight Prediction Analysis

4.4.1 Agreement

This file contains the information of the customer with the agreed cost for transporting goods to the customer.

- Code - It is the customer id or vendor id.
- Cost - It is the price decided to pay for the transporter for delivering the goods from source to destination.

4.4.2 GO

The above mentioned file contains the details of what materials are transported by what type of vehicle o what date.

- Created on - It is the date at which the vehicle starting its journey to transport the goods to the customer.
- Customer - It is the customer id, where goods are delivered.
- Customer Name - It is the customer name.

- Vendor - It is the vendor id, where goods are delivered.
- Name - It is the name of the vendor.
- Veh Registration No - It is the registration number of the vehicle.
- LR Number - It is the lorry receipt generated on created on.
-

4.4.3 Transporter bill

It contains the data of the cost of trip done by the transporter. Here Date and Vehicle number are the primary keys.

- Date
- LR No - It is the lorry receipt generated on created on. number.
- Vehicle No - It is the number of the vehicle.
- Cost - It is the freight cost.

5 IMPLEMENTATION

This chapter gives a brief description about implementation details of the system by describing each component with its code skeleton in terms of algorithm.

5.1 Production and Loss analysis

Below are the DAX formulas that are used to create measures for Production and Loss analysis to create visuals.

1. OEE = DIVIDE(CALCULATE (SUM ('Production data consolidated' [Eq Qty]), FILTER (' Production data consolidated ' , ' Production data consolidated ' [Type] = "Actual")), CALCULATE(SUM('Production data consolidated'[Eq Qty]), FILTER (' Production data consolidated ' , ' Production data consolidated'[Type] = "Plan")) ,0)
   - OEE stands for Operational Equipment Efficiency that gives the efficiency of the production.
   - With the help of this measure we can calculate the efficiency of each plant, efficiency of production for year, quarter, month, day.
2. Planned Quantity = var PQ = CALCULATE(sum (' Production data consolidated'[QTY]), ' Production data consolidated ' [Type] = " Plan ") return IF (ISBLANK(PQ) , 0, PQ)
   - Planned Quantity is the measures that gives us total quantity of products that was planned to manufacture for given period of time.
3. Actual Quantity = var AQ = CALCULATE (sum (' Production data consolidated ' [QTY]), ' Production data consolidated ' [Type] = " Actual ") return IF (ISBLANK(AQ) , 0 , AQ)
   - Actual Quantity is the total number of quantities that was manufactured for given time period.
4. planned minutes = CALCULATE (SUM (' Production data consolidated ' [Eq Qty]) , ' Production data consolidated ' [Type] = " Plan ")
   - Planned minutes is the measure that says the total time the employees had to work to produce the number of quantity of products that was planned.
5. actual minutes = CALCULATE (SUM (' Production data consolidated ' [Eq Qty]), ' Production data consolidated ' [Type] = " Actual ")
   - Actual minutes is measure that gives the total time employees took to produce planned number of quantity of products.

6. Planned M/c Utilization = DIVIDE ([planned minutes] , [Planned max minutes] * [Count of unique days] , 0)

    - Planned machine utilization is a measure that gives the efficiency of time utilization in the working hours.

7. Loss = CALCULATE (SUM (' Production data consolidated ' [Eq Qty]) , ' Production data consolidated ' [Type] = " Plan ") - CALCULATE (SUM (' Production data consolidated ' [Eq Qty]) , ' Production data consolidated ' [Type] = " Actual ")

    - Loss is a measure showing the total amount of time the employees didn't work.

8. Loss in % (loss table) = divide (' Loss details consolidated ' [Loss table] , [planned minutes] , 0)

9. Time Loss in % = CALCULATE(([Loss]/[planned minutes]))

    - The above 8th and 9th measures are percentage of wastage of time done during the process of manufacturing the products.

10. Count of unique days = COUNT('Production data'[Date])

    - The above measure gives the number of working days.

5.2 Freight Prediction

Below algorithm shows the implementation details of clustering the freight data for freight cost prediction

Algorithm 1: K-Means Clustering for Freight price

initialization;
while *True do* **do**
  instructions;
  for i = 1 to m ... **do**
    $c^(i)$= index 1 to k of cluster centroid closest to $X^(i)$
  end for
  for k = 1 to K **do**
    $u_(k)$= Average of points assigned to cluster k
  end for
end

K-means clustering is an unsupervised learning algorithm, which is used when you have unlabeled data. The aim is to find clusters in the data based on attributes which will impact the price of freight, with the number of clusters denoted by the variable K. The value K is determined using elbow curve. The algorithm repetitively assigns each data point to one of the clusters. Data are clustered based on feature similarity. This will label the data with each data point being assigned to a cluster.

This allows you to find and analyze the clusters formed, instead of basing the clusters on preconceived notations. Centroid of each cluster is a collection of values that define the data points in that cluster.

ALGORITHM

The k-means clustering algorithm uses an repetitive procedure with refinement after each step to arrive at the final result. The algorithm inputs are the number of clusters N and the data set. The algorithms start with initial estimates for the N centroids, which are randomly decided. The algorithm then iterates between two steps:

- Each centroid defines one of the clusters. In this step, each data point is assigned to its nearest centroid, based on the squared Euclidean distance.
- Centroid update step: In this step, the centroids are recomputed. This is done by taking the mean of all data points assigned to that centroid's cluster.

The algorithm iterates between the two steps until a stopping criterion is met which could be either of no data points change clusters, the sum of the distances is minimized, or some maximum number of iterations is reached. This algorithm is guaranteed to converge to a result. The result may be a local optimum (i.e. not necessarily the best possible outcome), meaning that assessing more than one run of the algorithm with randomized starting centroids will give a better result

Below algorithm shows the implementation details Multiple Linear Regression for a cluster of freight cost prediction

Algorithm 2: Multiple Linear Regression

initialization;
θ = some number
α = small number (0.001)
while *True do* **do**
  instructions;
  $\theta_(i) = \theta_(i)$ - $\alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^(i)) - y^(i)) x_j^(i)$
  update θ_j for j =0 .....n
end

Multiple Linear Regression(MLR)

Multiple Linear Regression attempts to model the relationship between Distance in KM, Vehicle capacity, Transit hours and Freight Price by fitting a linear equation to the Freight data. The steps to perform multiple linear Regression are similar to that of linear Regression and the difference Lies in the hypothesis function. We can use it to find out which feature has the highest impact on the Freight Price and how other features relate to each other based on coefficient values b_0, b_1, b_2. Here : $\mathrm{Y} = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3$ Y = Dependent variable and x_1, x_2, x_3 = multiple independent variables

ALGORITHM

Step 1: Data Pre-Processing

- Importing the Freight Data.
- Encoding the Categorical Data.
- Avoiding the Categorical Variable Trap.
- Splitting the Data set into Training Set and Test Set in Power BI.

Step 2: Fitting Multiple Linear Regression to the Training data. Step 3: Predicting the Test data results in Power BI by user input.

Below algorithm shows the implementation details K-Nearest Neighbour for a cluster of freight cost prediction

Algorithm 3: K-Nearest Neighbour

instructions;
for i = 1 to n **do** Visited[i] = false
Initialize the list path with s
Visited[s] = true
Current = s
end for
for i = 2 to n **do**
    Find the lowest element in row current and unmarked column j containing the element
    Current = j
    Visited[j] = true
    Add j to the end of Path
end for
Add s to the end of list path
return path

Nearest neighbor is a special case of the k-nearest neighbor class. Where k value is 1 (k = 1). In this case, a new data point target class will be assigned to the 1st closest neighbor. How to choose the value of K? Selecting the value of K in K-nearest neighbor is the most critical problem. A small value of K means that noise will have a higher influence on the result i.e., the probability of over-fitting is very high. A large value of K makes it computationally expensive and destroys the basic idea behind KNN (that points that are near might have similar classes). A simple approach to select k is k = $n^(1/2)$. To optimize the results, we can use Cross-Validation. Using the cross-validation technique, we can test the KNN algorithm with different values of K. The model which gives good accuracy can be considered to be an optimal choice. K-value with the lowest Root Mean Squared Error(RMSE) score, which was 2 for this data was selected It depends on individual cases, at times the best process is to run through each possible value of k and test our result.

Condensed Nearest Neighbour Data Reduction Rule: Working on a big data set can be an expensive task. Using the condensed nearest neighbor rule, we can clean our data and can sort the important observations out of it. This process can reduce the execution time of the machine learning algorithm. But there is a chance of accuracy reduction. The steps to condense is to divide data points into these:

- Outliers: Observations that lie at an abnormal distance from all the data points. Most of these are extreme values. Removing these observations will increase the accuracy of the model.
- Prototypes: Minimum points in the training set required to recognize non-outlier points.
- Absorbed points: These are points that are correctly identified to be non-outlier points

K-means clustering was done on the freight data selecting the distance, transit hours, and vehicle capacity as the attributes. These attributes were selected based on a correlation matrix which showed that the price of freight depended largely on these three attributes. Based on an elbow curve, the number of clusters to be formed was finalized as three. The three clusters formed to put the data which had similar values for the attributes selected.

Predictive algorithms like Multiple Linear Regression(MLR) and K-Nearest neighbors (KNN) were applied to individual clusters to predict the freight cost. The clustering helped to group data to give certain linearity to the data. This helps when applying predictive algorithms. The price was taken as the dependent variable y and vehicle capacity, transit hours, and distance were the multiple independent variables. Using the linear regression function in python the intercept and coefficient for each independent variable are calculated and the linear regression model is applied when a predict function is called. The important thing is that the model.predict() function is applied on a data frame containing values for the multi dependent variables and this returns an array containing the price for each tuple.

The KNN was also applied in a similar way, where the selection of K was done using the RMSE values. The optimal value for K was found to be 2. Data would contain only attributes and the target variable. This data is divided into test and train data and the model is built. Then when the predict function is called on a data frame that contains only the attributes and not the price. This would return the data frame adding a column for the price calculated.

5.3 Freight Analysis

Below are the DAX formulas that are used to create measures for Freight Analysis to create visuals.

1. Additional Cost = Verification[Number of Additional trips] * 2088
   - Additional Cost is a column that gives the additional price that has to be paid to the transporter for delivering the goods for more than one customer during one trip.
2. Total Cost = sumx (Verification, Verification [Agreement.Cost] + Verification [Additional Cost])
   - Total Cost is total price calculated for all the trips made by transporter.
3. Difference = SUMX(Verification, Verification[Transport Bill.Cost] - Verification[Total Agreement Cost])
   - Difference is a measure which is additional amount that is added to agreement cost which in-turn gives the total cost of transportation.

6 RESULTS AND DISCUSSIONS

6.1 Results of Production and Loss analysis

The aim for production and loss analysis was to analyse the efficiency for production of plants and value stream for planned work and actual work done. The visuals are created using the measures created using DAX formulas and below are the inference as the result of analysis.

Operational Equipment Efficiency(OEE) is a metric that identifies the percentage of planned production quantity that is truly productive and OEE score of 100% represents perfect production: manufacturing only good parts, as fast as possible, with no downtime. Here in figure 7, there are two plants named DTA and EOU with 87.95% and 73.46% OEE respectively,based on the calculations performed by DAX formula mentioned before.

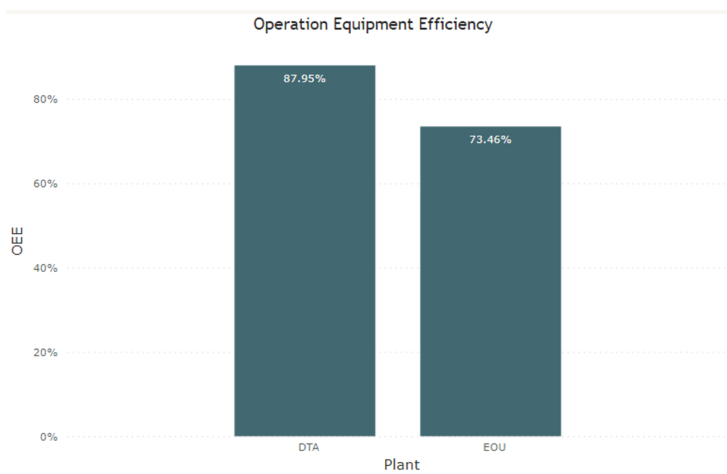

Fig. 7: Efficiency of DTA and EOU plants

In figure 8, in DTA, VSM Cincinati has highest OEE with 93.83% and VSM EFD-3 with 73.43% for production data of 2021.

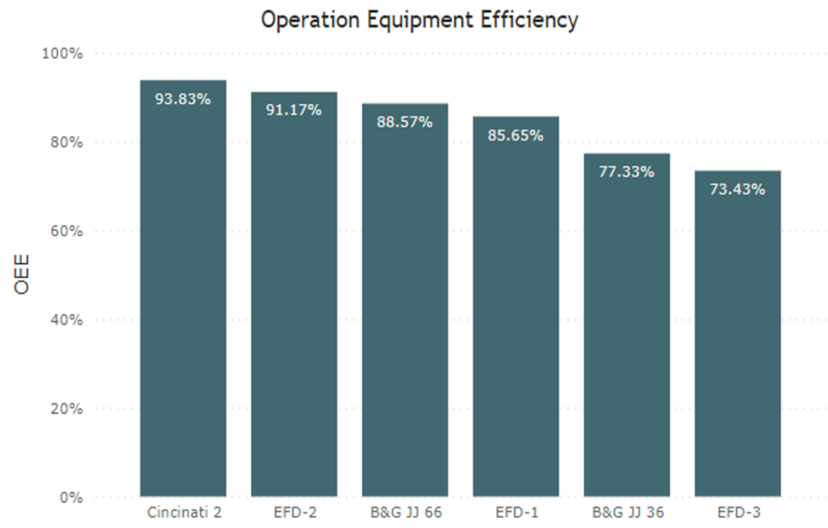

Fig. 8: VSM with highest and lowest OEE in DTA plant

In figure 9, in plant EOU, VSM Water fall AMS-1 with 88.62% is highest OEE and VSM EOU GM-AMS has lowest OEE with 69.4% for production data of 2021.

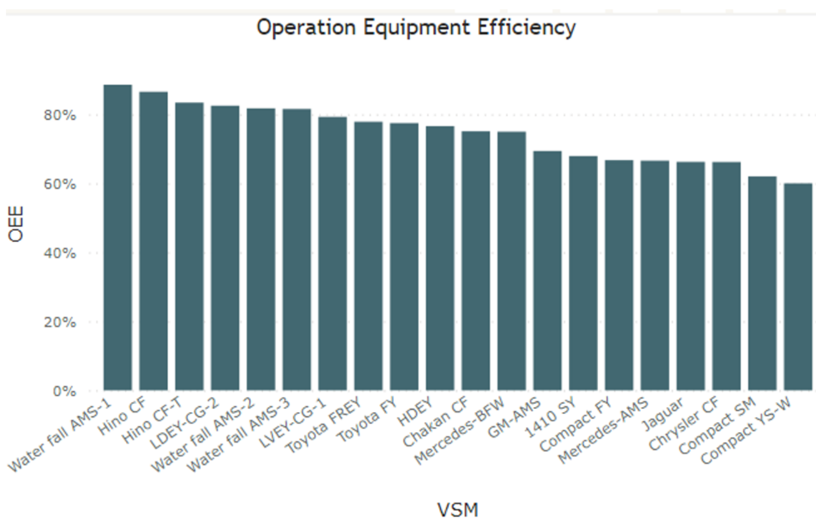

Fig. 9: VSM with highest and lowest OEE in EOU plant

In figure 10,is the Pareto chart with minutes lost during production process based on department, where SCGM having highest time loss with 186400 minutes and Tolling with least i.e, 26180 minutes.

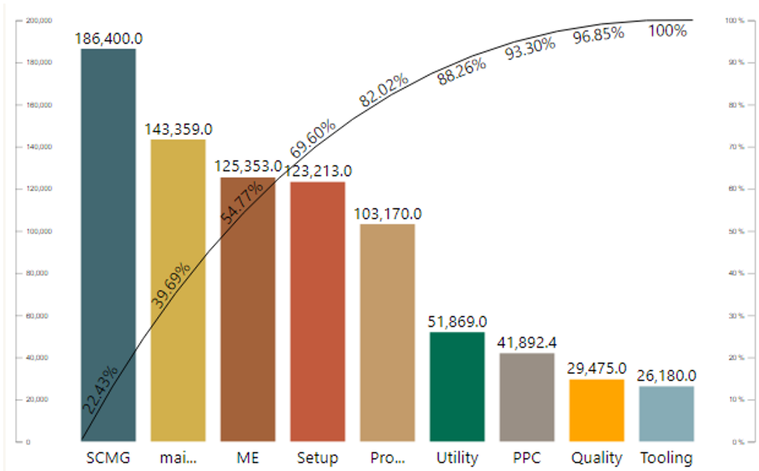

Fig. 10: Loss of time in each department

In figure 11, we see SCMG department has 22.81% loss of time stating more loss of working minutes and tooling department has 4.1% loss of time stating least loss of working minutes

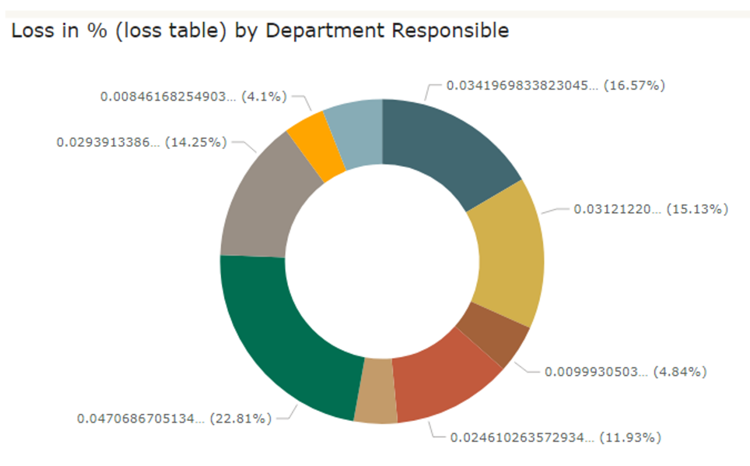

Fig. 11: Loss of time in percentage

6.2 Freight Prediction

Freight prediction is the method of developing models for forecasting freight price for existing and new routes.

The prediction of freight cost is done using metrics such as vehicle capacity, distance in kilometer and transit hours, which showed maximum correlation with price.Higher the correlation lighter the color. The correlation is on a scale of 0 to 1, with 0 indicating least correlation or dependency and 1 indicating highest. Figure 12 shows the heatmap for freight data indicating that price is heavily dependent on Distance in KM, Vehicle capacity and Transit hours

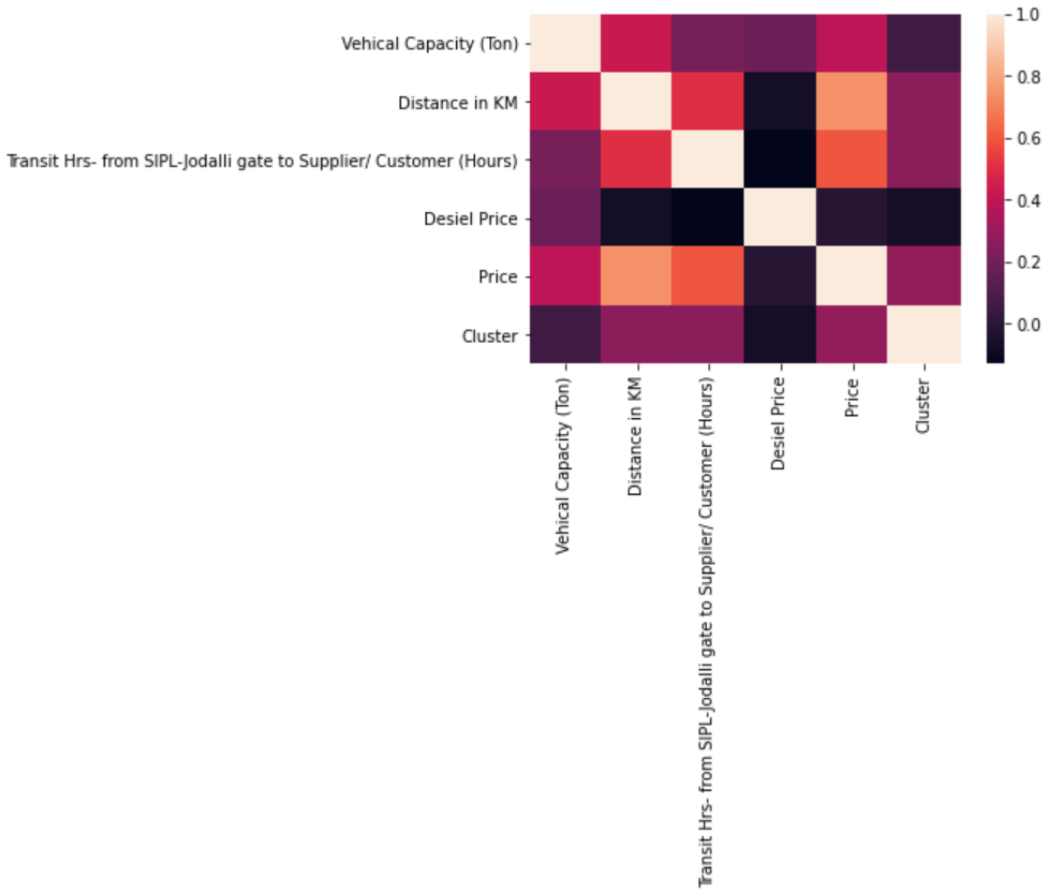

Fig. 12: Heatmap for Freight data

The figure 13, shows the elbow curve based on which the k value was chosen for K-Means clustering.

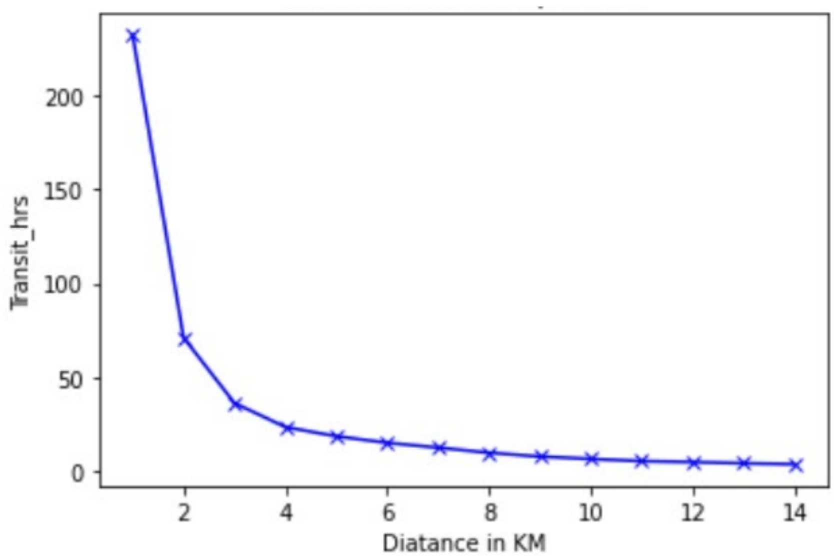

Fig. 13: Result of Elbow Curve

The figure 14, shows the results of K-Means clustering.

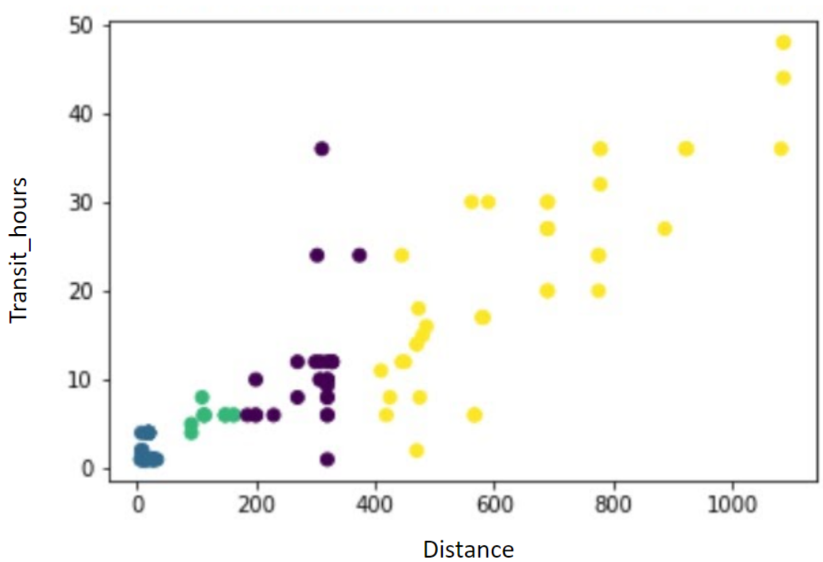

Fig. 14: Clusters for Freight Data

The figure 15, shows the working of Linear Regression in Power BI.The two lines in the graph corresponds to actual and predicted values. Due to lack of amount of data from the client, the model has over-fitted. User will give the input for the dependent variables - Distance in KM, Transit hours and Vehicle capacity.With the help of the model the price for the corresponding inputs is predicted.The involved integration of Python and Power BI which is one of the most powerful feature of Power BI.

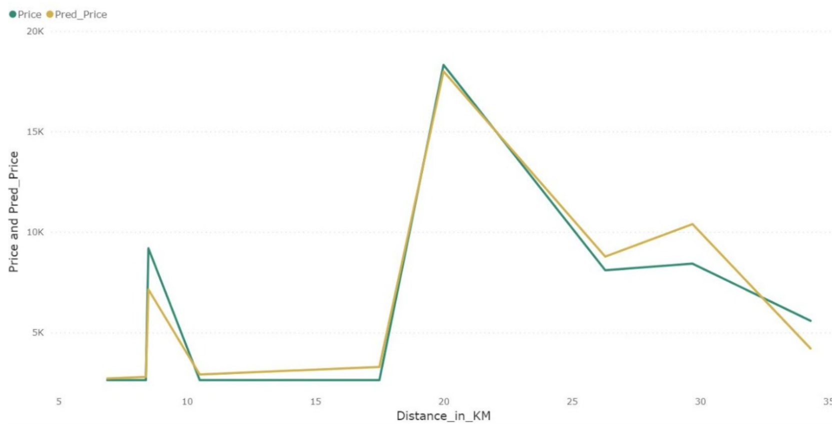

Fig. 15: Actual and Predicted values for Multiple Linear Regression

6.3 Freight Analysis

In Freight Analysis, the transportation details are analysed such as number of trips made by the transporter, total cost for the trip and identifying any extra cost given.In this attributes such as agreement cost and total transporter cost is used to identify the reason for the extra amount paid to a particular transporter.

In figure 16, the graph represents the comparison of the agreed cost by the company and the transporter cost which is the result of of agreement cost and the cost of additional trips.

Difference	Customer	Total Agreement Cost	Customer Name	Transport Bill.Cost	Year
26,928.00	0000300015	249,324.00	Dana Italia Srl	276252	2021
6,000.00	0000300020	172,216.00	Spicer Nordiska Kardan Ab	178216	2021
4,500.00	0000300020	129,162.00	Spicer Nordiska Kardan Ab	133662	2021
4,500.00	0000300020	129,162.00	Spicer Nordiska Kardan Ab	133662	2021
16,704.00	C1004	107,232.00	DANA ANAND India Pvt.Ltd. -SATARA	123936	2021
28,616.00	C1004	93,828.00	DANA ANAND India Pvt.Ltd. -SATARA	122444	2021
6,756.00	C1004	80,424.00	DANA ANAND India Pvt.Ltd. -SATARA	87180	2021
3,717.00	C1004	46,476.00	DANA ANAND India Pvt.Ltd. -SATARA	50193	2021
92,819.00		**1,163,038.00**		**1255857**	

Fig. 16: Details of Agreement and Transporter Bill

In figure 17, the table consists of the details of customer and their agreed cost of freight and if there are any additional trips then the transporter will charge certain amount of additional money in turns which gives transport bill cost.

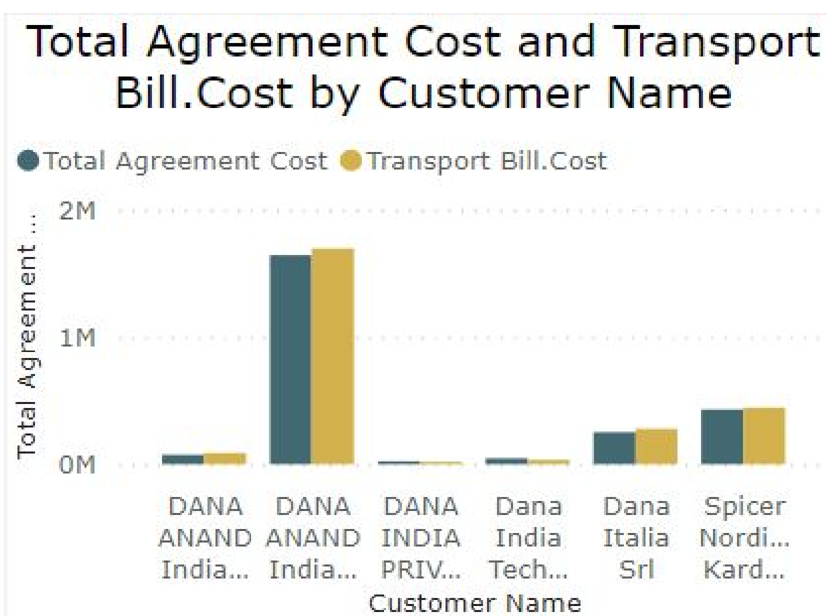

Fig. 17: Comparison of Agreement and Transporter Bill Cost

In figure 18, the graph gives the count of the trips made for each individual customer, this consists of the original and additional trips.

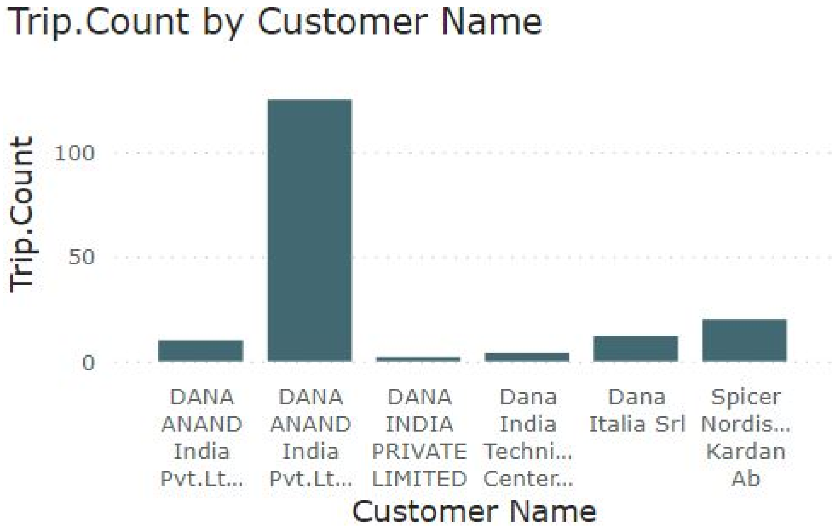

Fig. 18: Count of total trip for each customer

7 CONCLUSION

The production dashboard enables the client to keep tabs on the everyday working of the plants. Adding new data is an easy process and requiring no technical knowledge, thereby enabling workers to easily update the dashboard. These visualizations will also help the management gauge the impact of each plant during quarterly and yearly meetings. Finding minute problems and reasons due to which they occur will be uncomplicated, trouble-free, and less time-consuming.

The freight prediction dashboard will help assess whether the client is paying the right cost for a particular route, also providing an option to calculate the cost if a new route is added. The freight analysis dashboard will help verify if the right cost is being paid for a trip made by the transporter. This dashboard will also help to check trends of the cost incurred for each customer.

8 FUTURE SCOPE

In the future as the data increases, we expect the model to be performing much better while predicting the freight cost. The production and freight analysis dashboards can be deployed on cloud platforms thereby enabling remote access to users all over the world. Globally automotive companies are gradually increasing their digital capabilities in order to minimise errors in manual quality control. Our project will not only help companies to improve operational efficiency but also enhance accountability in operations.

9 References

[**1**] Minewiskan. "DAX Function Reference - DAX." DAX — Microsoft Docs. Accessed April 27, 2021. https://docs.microsoft.com/en-us/dax/dax-function-reference.

[**2**] "Microsoft Power BI." Wikipedia. March 29, 2021. Accessed April 27, 2021. https://en.wikipedia.org/wiki/Microsoft_Power_BI.

[**3**] Scardina, Jesse, and Lauren Horwitz. "What Is Microsoft Power BI? - Definition from WhatIs.com." SearchContentManagement. December 03, 2018. Accessed April 27, 2021. https://searchcontentmanagement.techtarget.com/definition/Microsoft-Power-BI.

[**4**] "Create Power BI Visuals Using Python in Power BI Desktop - Power BI." Create Power BI Visuals Using Python in Power BI Desktop - Power BI — Microsoft Docs. Accessed April 27, 2021. https://docs.microsoft.com/en-us/power-bi/connect-data/desktop-python-visuals

10 Appendix

10.1 Gantt Chart

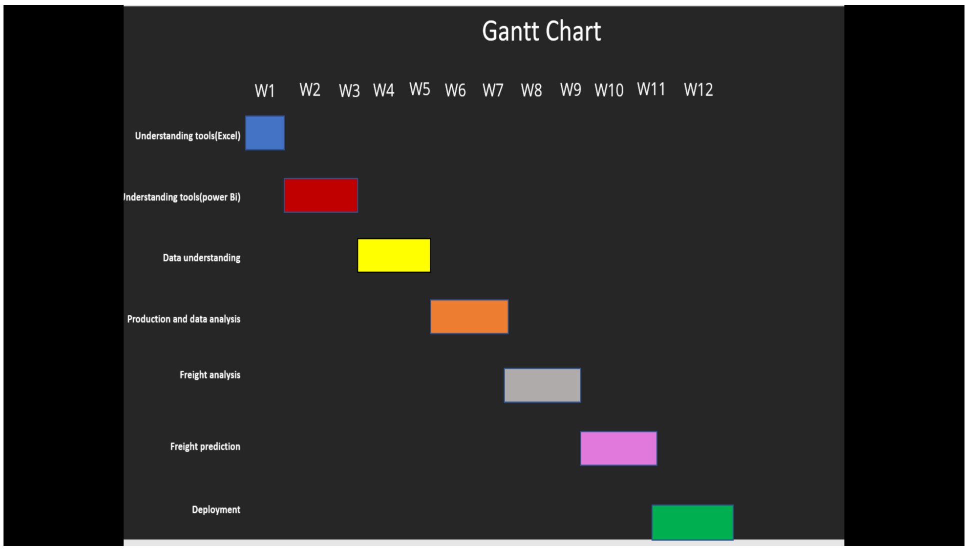

Fig. 19: Gantt Chart

Plagiarism Report

Project Title: Production and Freight Analysis Using Power BI for Dana

Team No: I6

Project Domain: Data Analytics

Category: Industry

ORIGINALITY REPORT

23%	19%	8%	16%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dataaspirant.com Internet Source	5%
2	Submitted to University of Patras Student Paper	2%
3	www.geeksforgeeks.org Internet Source	2%
4	en.wikipedia.org Internet Source	1%
5	Omid Dehzangi, Alexander Melville, Mojtaba Taherisadr. "Chapter 5 Automatic EEG Blink Detection Using Dynamic Time Warping Score Clustering", Springer Science and Business Media LLC, 2019 Publication	1%
6	www.iiconsortium.org Internet Source	1%
7	Submitted to University of Southampton Student Paper	1%
8	data-flair.training Internet Source	

		1%
9	www.futurebridge.com Internet Source	1%
10	link.springer.com Internet Source	1%
11	Submitted to Glyndwr University Student Paper	1%
12	www.mitre.org Internet Source	1%
13	www.knowledgehut.com Internet Source	1%
14	docplayer.net Internet Source	1%
15	www.datascience.com Internet Source	<1%
16	Submitted to University of Greenwich Student Paper	<1%
17	docs.microsoft.com Internet Source	<1%
18	www.coursehero.com Internet Source	<1%
19	Submitted to B.V. B College of Engineering and Technology, Hubli	<1%

Student Paper

20	Ilya Galkin, Andrei Blinov, Ievgen Verbytskyi, Denys Zinchenko. "Modular Self-Balancing Battery Charger Concept for Cost-Effective Power-Assist Wheelchairs", Energies, 2019 Publication	<1%
21	Submitted to Newcastle College Group Student Paper	<1%
22	W. O. Baker, C. S. Fuller, N. R. Pape. "Effects of Heat, Solvents and Hydrogen-bonding Agents on the Crystallinity of Cellulose Esters", Journal of the American Chemical Society, 2002 Publication	<1%
23	Timothy Oladunni, Sharad Sharma, Raymond Tiwang. "A Spatio - Temporal Hedonic House Regression Model", 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), 2017 Publication	<1%
24	www.kaggle.com Internet Source	<1%
25	Submitted to Nottingham Trent University Student Paper	<1%
26	Submitted to National College of Ireland Student Paper	<1%

27	Submitted to University College Technology Sarawak Student Paper	<1%
28	Adam Aspin. "Pro Power BI Desktop", Springer Science and Business Media LLC, 2018 Publication	<1%
29	www.kabargames.id Internet Source	<1%
30	"Pattern Recognition and Machine Intelligence", Springer Science and Business Media LLC, 2019 Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		