Project Report for Data Science

Business Intelligence
Primary Topic: Data Processing and Visualization (DPV)
Course: 2020-2A – Group: 43 – Submission Date: 2021-04-13

Joël Ledelay
University of Twente
j.ledelay@ student.utwente.nl

Sneha Ramesh
University of Twente
s.ramesh-1@student.utwente.nl

ABSTRACT

Business Intelligence (BI) refers to the suite of strategies and technologies that are used by organizations for the analysis of business information (data). BI helps us to identify new opportunities and implement strategies effectively. BI enables organizations take effective and informed decisions. For the processes that are involved in BI, firstly, data has to be collected. This can be from internal or external systems/ sources. It then has to be prepared for analysis for running queries against the extracted data. Visualizations are created using special tools to help users understand the data better and make relevant changes to the processes/ decisions.

This project aims to derive insights from the **SalesProd_Database**. We attempt to understand trends based on the current data and extrapolate these trends to help the organization better their processes.

The dataset is multi-dimensional, in the sense that we have information regarding the Employees, Products, Vendors and Sales. If analyzed properly, we will not only be able to unearth current trends but will also be able to provide suggestions for changes that will help the organization further their current successes and minimize their losses. The process will involve data extraction, data cleansing and processing, loading into the database and finally, data visualization.

The tools that will help us achieve the steps listed above are *MS access* (raw data), *R* (data cleansing and processing), *PostgreSQL* server (the server that houses the cleansed and processed data) and *Tableau* (for data visualization). The outcome of our analytics is a set of dashboards, which will clearly elucidate our KPIs (defined below).

KEYWORDS

Business Intelligence, insights, trends, KPIs, dashboards

1 INTRODUCTION

BI aims to drive better (business) decisions. The benefits of these may be multifold: increased revenue, increased operational efficiency and a competitive advantage over other business players [1].

BI data can contain the following classes of information: historical information and real-time data. Historical data is what is available in the organization's database, possibly accumulated over a long period of time. Real time data is gathered from source systems as and when it is generated (sensors, detectors etc.). Before the data is used in BI applications, the data must be collated, enhanced, and sanitized using data integration tools to make sure that the information being analyzed has integrity, accuracy and consistency.

It is also important for every organization to define Key Performance Indicators (KPIs). KPIs are definitive goals, which can be measured. These help the organization determine how effectively they are achieving their primary business targets. High-level KPIs usually stress upon the overall performance of the company, while low-level KPIs address processes of each division/ department within the organization such as sales, marketing HR etc. The KPIs for our case have been listed in *section 2.4.1*.

The process employed for the data processing and the development of the dashboards have been explained below in detail. The list of steps for performing the BI analytics have been listed below.

1. Extraction (E)
2. Transformation (T)
3. Loading (L)
4. Analysis (Visualization)

2 BACKGROUND - ETL PROCESS

Extract, Transform, Load (abbreviated as ETL) is the process of transferring the data from a source system (such as a database) into a destination system (such as an analytics tool). The destination system may have features that represents the data in a different manner, as compared to the source system(s).

A good ETL system is expected to perform the following functions effectively: data extraction from the source system; data quality, integrity and consistency checks; joining/ combination of data tables from different sources (if necessary); delivery of data in a format that is ready to present to stakeholders. Using this, application developers can build tools that end users can utilize to make decisions [2]. *Figure 1* depicts the ETL process. The process is applied to our BI case and each step is explained in detail in the subsequent sections.

It is crucial that each step is performed with care. Flawed or improper input-data produces flawed or improper output(s) and/ or interpretations. This is known as *Garbage In, Garbage Out* (GIGO). "Garbage in – garbage out" is an idiomatic way of saying that poor-quality data used as input can cause unreliable data output. Performing any amount of visualization on incorrect data will be of little or no use to the organization. It would lead to incorrect interpretations/ analysis and could even cause catastrophic consequences to the company.

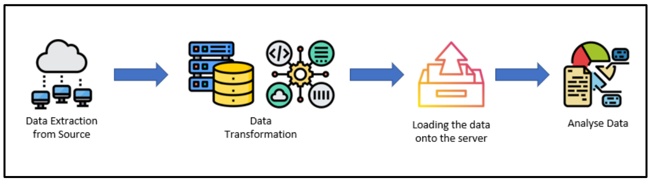
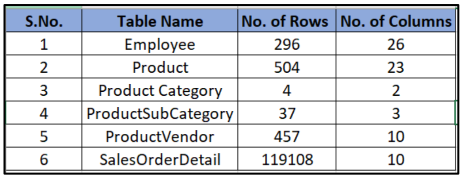


Figure 1: ETL and Analysis Process

2.1 Data Extraction (E)

The ETL process starts off with the data extraction from the source system. In our case, we extracted the data provided to us for the Data Science project from *Canvas*. The dataset used for this project is **SalesProd_Database**. The file is of the extension *.accdb*. This file opens with the help of the Microsoft Access Database.

A fundamental part of data extraction necessitates data validation to confirm that the data obtained from the source system(s) has the appropriate/ anticipated values under each domain. These are also referred to as columns/ attributes. The dataset contained the following tables with the number of records and columns, respectively. For this, we first extracted the data imported it into *R* for data transformation and cleansing. We then ensured that the rows and columns from the source data have been imported correctly for processing.



S.No.	Table Name	No. of Rows	No. of Columns
1	Employee	296	26
2	Product	504	23
3	Product Category	4	2
4	ProductSubCategory	37	3
5	ProductVendor	457	10
6	SalesOrderDetail	119108	10

Table 1:Dataset details (Raw Data)

2.2 Data Transformation (T)

In this stage, a sequence of rules or semantics are applied to the extracted data to prepare it for loading into the destination system.

A vital function of data transformation is data cleansing (or) data cleaning , which aims to pass only "correct"/ "sanitized" data to the destination system. This is vital to maintain consistency in data types, handle null values appropriately, handle fields with special/ unrecognizable characters etc. Also, the tables are joined so that all the necessary fields are available and accessible in the relevant tables for the data visualization. Some fields (profit, sales amount, average revenue per unit, net profit margin for instance) require minor computations using the existing fields. These values are computed accordingly and are added as columns to the final table that must be visualized.

The more time we spend on the cleansing, the better our output will be. This is because we would have a thorough understanding of the dataset. As mentioned earlier more meaningful data leads to better analysis.

Our data transformations are performed using *R*. The **SalesProd_Database** is first loaded into the *R* engine. *Note: We must have the Microsoft Access Database Engine installed to load the extracted data.*

The Open Database Connectivity (*ODBC*) driver makes use of the interface by Microsoft that allows engines such as R to retrieve data from Database Management Systems (*DBMS*). This is achieved by querying using SQL. ODBC allows multiple systems to work together (i.e., interoperability). This simply means that a single application such as *R* can access different databases/ data warehouses. The user can include the ODBC database drivers on their engine to connect the application to their preferred DBMS (in our case, Microsoft Access Database and *PostgreSQL* server).

Next, the following standard set of data cleansing steps are performed:

- Filtering – choosing only selected columns to load.
- Data-type Conversion/ Check – For instance, all date fields must be in one of the date formats and not in any other format.
- Null value validation – blanks must be appropriately labelled, wherever necessary.
- Joining Tables - Data flow validation from the staging area to the intermediate tables.
- Adding computed columns – Performing minor arithmetic operations on the data fields and adding the resulting columns into the final table.

Note: We have listed the steps that have been used in this project to cleanse the data set. However, this is not an exhaustive list of all the data cleansing activities.

2.2.1 Star Schema of the data

The star schema is a visual representation of the dataset. The star schema renders the business data into fact tables and dimension tables. Facts tables house all the quantifiable data about a business. Dimension tables contain the descriptive attributes or fields related to the fact tables. So, to sum up, every star schema comprises fact and dimension tables. The star schema of our dataset is given in *figure 2*.

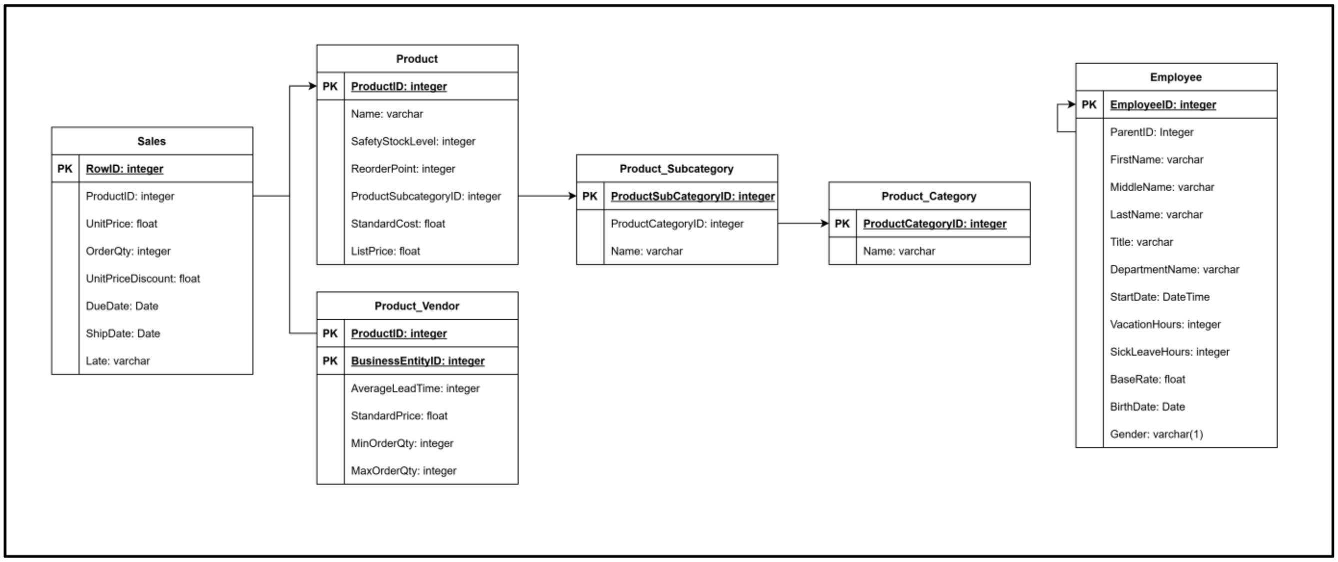


Figure 2: Star Schema of the dataset

- **Fact table (*sales*)** - Fact tables usually have numeric values and the foreign keys that link to the dimensional data tables where all the descriptive (qualitative) information is stored. They usually have many records and can also be referred to as a "transaction table".
- **Dimension tables:** Dimension tables generally have relatively fewer rows as compared to fact tables. However, but each record may have a very large number of attributes/ columns to describe the fact data.

Star schemas are denormalized. This means that the standard rules of normalization applied to transactional relational databases are not considered strictly during star-schema design and implementation. The advantages of denormalization are simple queries, simple business logic, fast aggregations etc.

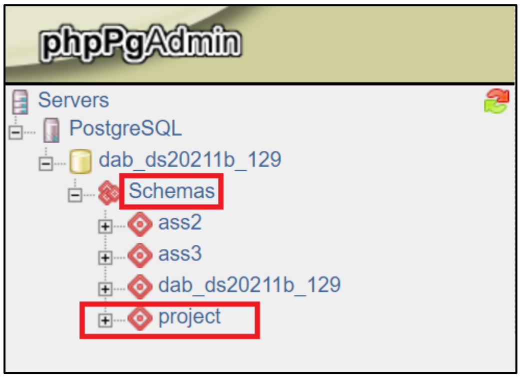
2.3 Data Loading (L)

This step loads the data into the server (*PostgreSQL server*). This is where the data tables rest. Since this phase communicates with a database, the conditions and constraints specified in the database schema provide the overall data quality execution of the ETL process.

Once the data is prepared, we load it into the *PostgreSQL* server with the help of *R*. Our tables rest in the *PostgreSQL 10.16* running on server *bronto.ewi.utwente.nl*. This server also runs a web-based database tool, called **PhpPgAdmin,** from where we can easily access, view and modify our tables. We have specific credentials provided to us by the university to access the database. The screenshots of the server and the tables resting in it have been displayed in *figure 3 and figure 4,* respectively.

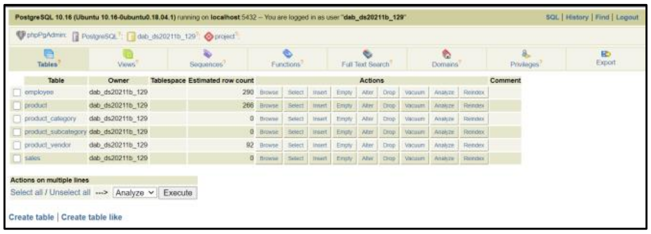
There are two types of tables in the database: Fact Tables and Dimension Tables (as explained in the previous section). Once the data is loaded into fact and dimension tables, we utilize this to improve performance for BI data by creating aggregates.

The *R* source code used for data processing (transformation) and loading can be found in *Appendix A*. We used *RStudio* for running the source code.



Servers
PostgreSQL
dab_ds20211b_129
Schemas
ass2
ass3
dab_ds20211b_129
project

Figure 3: "Project" schema within the PostgreSQL server


SQL | History | Find | Logout
phpPgAdmin: PostgreSQL dab_ds20211b_129 project
Tables
Views
Sequences
Functions
Full Text Search
Domains
Privileges
Export

PostgreSQL 10.16 (Ubuntu 10.16-0ubuntu0.18.04.1) running on **localhost** 5432 – You are logged in as user "**dab_ds20211b_129**"

	Table	Owner	Tablespace	Estimated row count	Actions									Comment
□	employee	dab_ds20211b_129		290	Browse	Select	Insert	Empty	Alter	Drop	Vacuum	Analyze	Reindex	
□	product	dab_ds20211b_129		266	Browse	Select	Insert	Empty	Alter	Drop	Vacuum	Analyze	Reindex	
□	product_category	dab_ds20211b_129		0	Browse	Select	Insert	Empty	Alter	Drop	Vacuum	Analyze	Reindex	
□	product_subcategory	dab_ds20211b_129		0	Browse	Select	Insert	Empty	Alter	Drop	Vacuum	Analyze	Reindex	
□	product_vendor	dab_ds20211b_129		92	Browse	Select	Insert	Empty	Alter	Drop	Vacuum	Analyze	Reindex	
□	sales	dab_ds20211b_129		0	Browse	Select	Insert	Empty	Alter	Drop	Vacuum	Analyze	Reindex	

Actions on multiple lines
Select all / Unselect all ---> Analyze Execute

Create table | Create table like

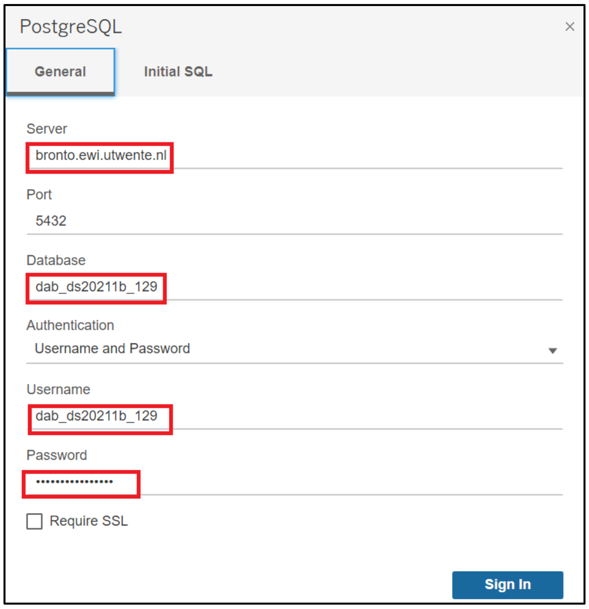
Figure 4: Tables within the database

2.4 Data Visualization

Data visualization is the representation of information and data using charts and graphs. These visual elements provide an accessible way to see and understand trends, outliers, patterns and models in data [3]. Visualization also provides a rapid and useful method to convey information in a comprehensive manner using visual stimuli. This can improve the identification of factors that impact buyer actions; identify zones that need to be upgraded or need more attention; make data more meaningful for participants; identify when and where to position particular products and forecast sales volumes and patterns. However, the scope of data visualization is not limited to the points mentioned above.

For visualizing our data, we make use of *Tableau 2020.4*. As mentioned in the previous stages, th**e** data is extracted, transformed, and loaded into the *PostgreSQL server*. A connection is established between Tableau and the *PostgreSQL* server.

*Note: Tableau can establish a connection with a variety of data sources. Since our data rests on PostgreSQL server, in the project schema***,** *we choose PostgreSQL under the connection options in Tableau. We then must specify the credentials – server name, username and password – to access the database***.** *See Figure 5 for the same.*



PostgreSQL

Server: bronto.ewi.utwente.nl

Port: 5432

Database: dab_ds20211b_129

Authentication: Username and Password

Username: dab_ds20211b_129

Password: ••••••••••••••••

□ Require SSL

Figure 5: Establishing the database connection with Tableau.

Next, we proceed to load the tables in *Tableau* with the star schema as a guide. The connections between the tables will have to be specified, i.e., the references are the primary key/ foreign keys for the tables. *Figure 6* shows the tables that have been loaded successfully in Tableau.

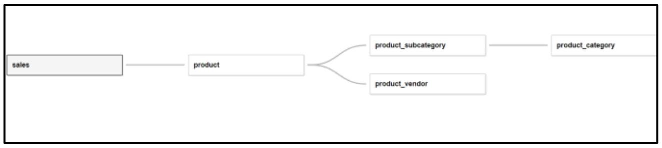


Figure 6: Tables and their relationships in Tableau

Following this, the data is visualized using charts, graphs, and other visualization tools. Colors and shapes help us accentuate our ideas better and aid in identifying trends and patterns effectively.

The dashboards will be constructed with a view of these KPIs. Our KPIs have been defined in the following section. We expect to find some interesting trends and patterns apart from the intended results.

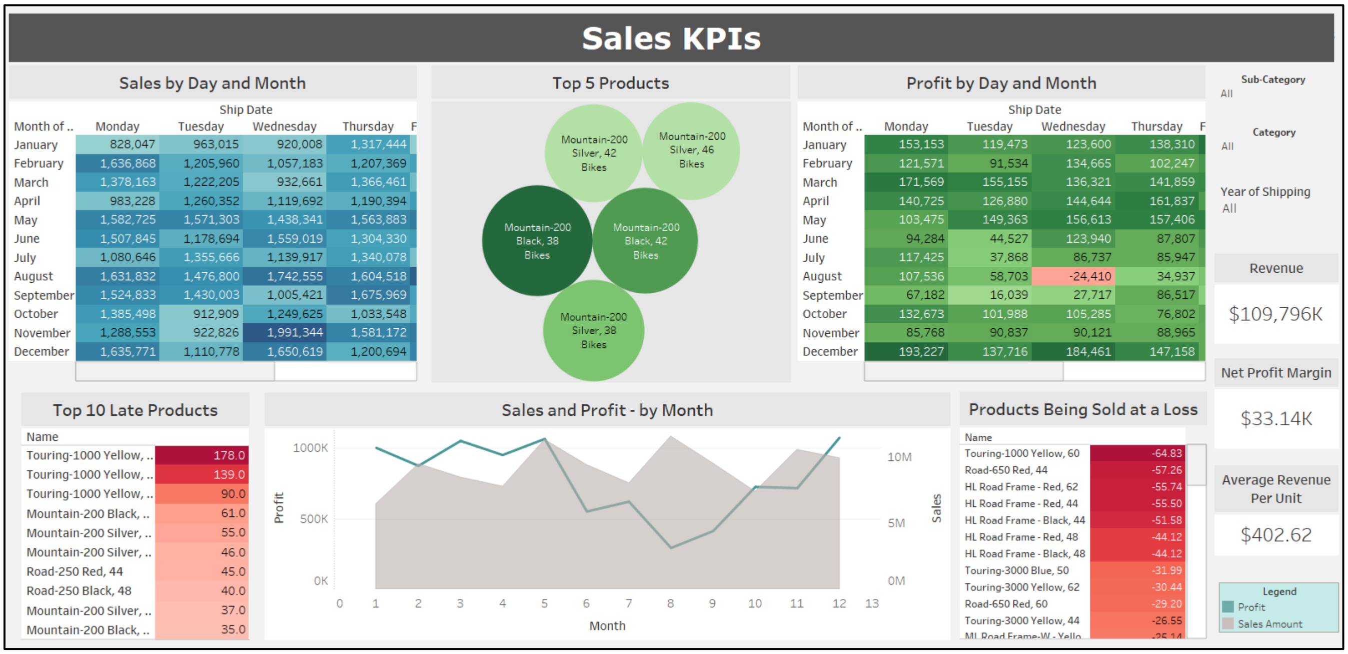
2.4.1 Key Performance Indicators (KPIs)

We have developed our Key Performance Indicators (KPIs) based on the **'Balanced Scorecard Perspective**' given in the BI project course guide. The main questions or indicators of the Balanced Scorecard Perspective have been listed below.

1. Financial Perspective: How do we appear to our Shareholders?
2. Customer Perspective: How do our Customers see/ perceive Us?
3. Internal Business Process Perspective: What should we do that is Excellent?
4. Employee and Organization Innovation and Learning Perspective: Can we continue to Improve and Add Value?

For these lists of indicators, we have come up with a set of indicators that will help us define our dashboards and metrics better.

1. Financial Questions: How do we look to our Shareholders?
   a. What is our gross sales per product category/ subcategory/ overall?
   b. What is our profit per product category/ subcategory/ overall?
   c. What is our Net Profit Margin?
   d. What is our Average Revenue Per Unit?
2. Customer: How do our customers see us?
   a. What are our top 5 products?
   b. What are the products/ categories/ subcategories of products that are shipped late?
   c. What can be done to reduce late shipments?
3. Internal Business Process: What should we do that is Excellent?
   a. How can we improve our strategy based on findings from the analysis?
   b. What can we do to improve profit?
   c. Should we improve our marketing strategy towards any product in particular?



Sales KPIs

Sales by Day and Month

Month of ..	Monday	Tuesday	Wednesday	Thursday
January	828,047	963,015	920,008	1,317,444
February	1,636,868	1,205,960	1,057,183	1,207,369
March	1,378,163	1,222,205	932,661	1,366,461
April	983,228	1,260,352	1,119,692	1,190,394
May	1,582,725	1,571,303	1,438,341	1,563,883
June	1,507,845	1,178,694	1,559,019	1,304,330
July	1,080,646	1,355,666	1,139,917	1,340,078
August	1,631,832	1,476,800	1,742,555	1,604,518
September	1,524,833	1,430,003	1,005,421	1,675,969
October	1,385,498	912,909	1,249,625	1,033,548
November	1,288,553	922,826	1,991,344	1,581,172
December	1,635,771	1,110,778	1,650,619	1,200,694

Profit by Day and Month

Month of ..	Monday	Tuesday	Wednesday	Thursday
January	153,153	119,473	123,600	138,310
February	121,571	91,534	134,665	102,247
March	171,569	155,155	136,321	141,859
April	140,725	126,880	144,644	161,837
May	103,475	149,363	156,613	157,406
June	94,284	44,527	123,940	87,807
July	117,425	37,868	86,737	85,947
August	107,536	58,703	-24,410	34,937
September	67,182	16,039	27,717	86,517
October	132,673	101,988	105,285	76,802
November	85,768	90,837	90,121	88,965
December	193,227	137,716	184,461	147,158

Sub-Category: All

Category: All

Year of Shipping: All

Revenue: $109,796K

Net Profit Margin: $33.14K

Average Revenue Per Unit: $402.62

Top 10 Late Products

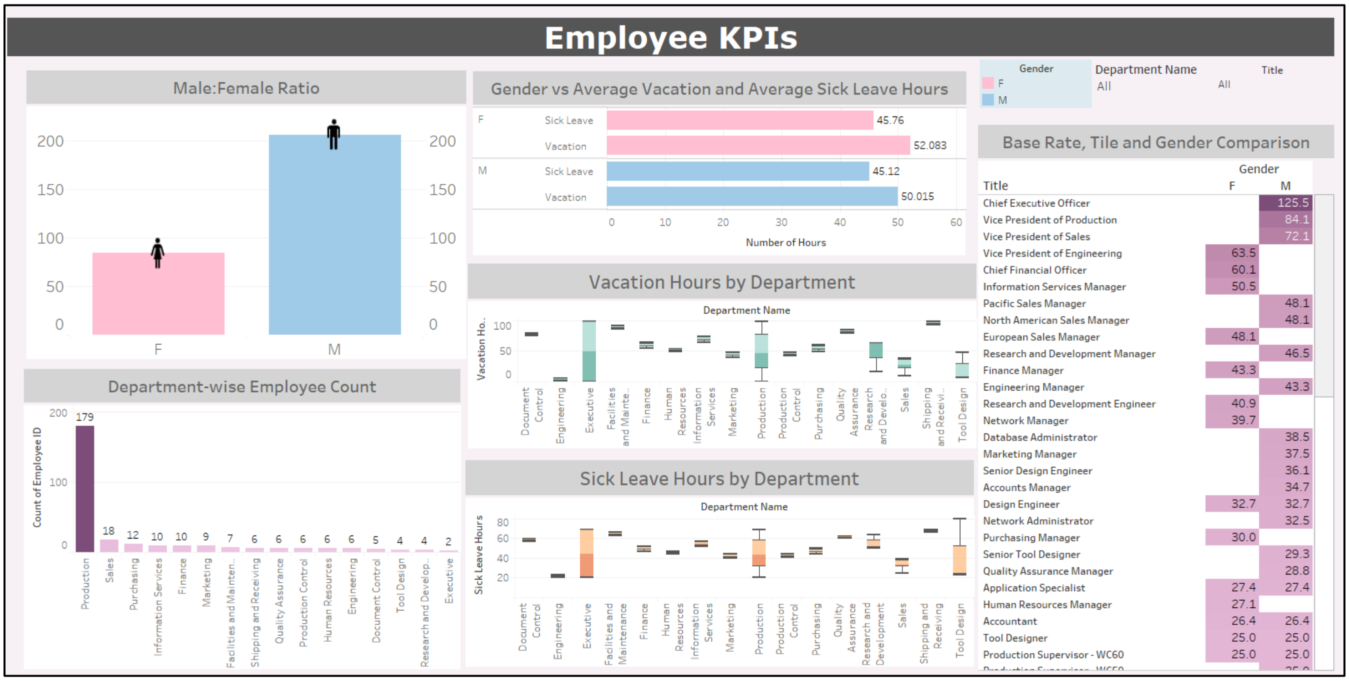
Name	
Touring-1000 Yellow, ..	178.0
Touring-1000 Yellow, ..	139.0
Touring-1000 Yellow, ..	90.0
Mountain-200 Black, ..	61.0
Mountain-200 Silver, ..	55.0
Mountain-200 Silver, ..	46.0
Road-250 Red, 44	45.0
Road-250 Black, 48	40.0
Mountain-200 Silver, ..	37.0
Mountain-200 Black, ..	35.0

Products Being Sold at a Loss

Name	
Touring-1000 Yellow, 60	-64.83
Road-650 Red, 44	-57.26
HL Road Frame - Red, 62	-55.74
HL Road Frame - Red, 44	-55.50
HL Road Frame - Black, 44	-51.58
HL Road Frame - Red, 48	-44.12
HL Road Frame - Black, 48	-44.12
Touring-3000 Blue, 50	-31.99
Touring-3000 Yellow, 62	-30.44
Road-650 Red, 60	-29.20
Touring-3000 Yellow, 44	-26.55
ML Road Frame-W - Yello	-25.14

Ship Date
Top 5 Products
Mountain-200 Silver, 42 Bikes
Mountain-200 Silver, 46 Bikes
Mountain-200 Black, 38 Bikes
Mountain-200 Black, 42 Bikes
Mountain-200 Silver, 38 Bikes
Sales and Profit - by Month
Profit
1000K
500K
0K
Sales
10M
5M
0M
0 1 2 3 4 5 6 7 8 9 10 11 12 13
Month
Legend
Profit
Sales Amount

Figure 7: Dashboard 1 -Analysis of Sales KPIs



Employee KPIs

Gender: F, M

Department Name: All

Title: All

Base Rate, Tile and Gender Comparison

Title	Gender F	Gender M
Chief Executive Officer		125.5
Vice President of Production		84.1
Vice President of Sales		72.1
Vice President of Engineering	63.5	
Chief Financial Officer	60.1	
Information Services Manager	50.5	
Pacific Sales Manager		48.1
North American Sales Manager		48.1
European Sales Manager	48.1	
Research and Development Manager		46.5
Finance Manager	43.3	
Engineering Manager		43.3
Research and Development Engineer	40.9	
Network Manager	39.7	
Database Administrator		38.5
Marketing Manager		37.5
Senior Design Engineer		36.1
Accounts Manager		34.7
Design Engineer	32.7	32.7
Network Administrator		32.5
Purchasing Manager	30.0	
Senior Tool Designer		29.3
Quality Assurance Manager		28.8
Application Specialist	27.4	27.4
Human Resources Manager	27.1	
Accountant	26.4	26.4
Tool Designer	25.0	25.0
Production Supervisor - WC60	25.0	25.0

Figure 8: Dashboard 2 - Analysis of Employee KPIs

d. How can we improve our HR capabilities to better our organization?

4. Employee and Organization Innovation and Learning: Can we continue to Improve and Add Value?

a. What is the *male:female* ratio our organization?

b. Do men and women in the same level/ designation receive a comparable salary?

c. What are the vacation trends of men and women? What changes can be made to this to improve employee satisfaction?

2.4.2 Dashboards

We present two dashboards that help us describe the KPIs listed above. The first one describes the Sales KPIs. The second one represents the Employee related analysis for the dataset. *Figures 7 and 8* give us a glimpse into the visualizations that we have carried out. *Note: The dashboards are dynamic, i.e., the views in each dashboard can be customized based on the filters provided in each one of them. Also, we have established a "LIVE" connection with the dataset. Hence, any changes to the dataset can be refreshed and directly updated on the dashboard. The screenshots of the dashboard are static and do not allow us to apply these filters. For more details, please look at the Tableau workbook.*

3 RESULTS

We have listed some findings from the visualizations. We have also included supporting snapshots from the visuals as evidence for our findings. *Appendix B* has the screenshots of the findings.

- **Finding 1:** The Sales Price of some products is lesser than the cost incurred by the organization for procuring/ producing them. This results in quite a few losses for the company.
- **Finding 2:** Products will lowest profit margin are caps, jerseys, and road frames. These products have done badly over all 4 years (2006, 2007, 2008 and 2009).
    - In 2009, caps performed slightly better than the other years.
    - The sales of jerseys are inversely proportional to profit in 2008.
- **Finding 3:** Products with high and fluctuating profit margins have been listed below:
    - High profit Margin - Cleaners, Hydration Packs, Gloves, Vests
    - Fluctuating Profit Margin - Mountain bikes, Road bikes, touring bikes, touring frame,
- **Finding 4:** The highest revenue per unit is given by the category "Bikes", specifically "Mountain Bikes", although it has a fluctuating profit margin.
- **Finding 5:** The male to female ratio is not proportional (84:206 males to females).
- **Finding 6:** There are very few people in marketing (9) and shipping and receiving (6) as compared to production (179).
- **Finding 7:** The trends in vacation are similar across both genders.
- **Finding 8:** Base rates of people in the same level are comparable. But some roles do not have women employees at all.

4 DISCUSSION

Based on our analysis, we propose the following improvements in the organization's strategy for a better throughput.

1. The marketing strategy towards some products such as caps, jerseys and road frames needs to be improved. The organization could focus on creating target specific or seasonal campaigns to improve the sales of these products.
2. Some products incur heavy losses because the amount spent on them by the organization is higher than the price at which it is sold. For this, the organization will have to revamp their price list (or) remove these products from the store is their demands are not very high.
3. Every organization today is striving towards a gender equal work environment. As we have seen the previous section the *male:female* ratio is not proportional. There are fewer women working in many of the designations than males. The organizations should look at recruiting more women to equalize this ratio.
4. The number of people in sales and shipping are quite low as compared to those in the production division. We also see that there are quite a few products that are shipped late and quite a few products that are incurring losses. The organization should recruit more people in the sales and shipping and control division to mitigate this.

However, it should be noted that the dataset spans the years 2006, 2007, 2008 and 2009. It is highly possible that the organization has considered these effects and is already aware of these limitations, since the data is quite old. Also, the company could have made the relevant changes and could have altered their processes in order to realize better strategies. There could be some new findings that could be unearthed, if a more recent dataset is provided to us.

All the findings that we have listed and the proposed improvements to the strategy of the organization will hopefully also be applicable to similar areas of other organizations. Multidimensional data, such as this one provides immense scope and capabilities for analysis and visualization. All organizations should strive to store consistent, reliable, authentic and clean data. The better the data quality is, the better the outputs of our analysis will be.

5 CONCLUSIONS

We have performed the ETL procedure and using a variety of tools such as *MS Access, R, PostgreSQL, Excel* and *Tableau*. These have helped us unearth some interesting patterns, trends and have given us some interesting findings on the data. The visualizations help us analyze the data well and we have been able to point out some areas of improvement for the organization. These findings and suggestions might be scalable to other products and other domains as well.

Data translates to insights; insights translate to information and information ultimately translates to knowledge. BI has tremendous potential that should be made use by all organizations to understand their data. The knowledge that can be mined from these methods have capabilities that can do wonders to the functioning of the organizations. There are multiple tools that can be used for the ETL and analysis processes. The best fit will have to be evaluated considering the sector and context for which it is being used.

REFERENCES

[1] Business Intelligence (BI) - https://searchbusinessanalytics.techtarget.com/

[2] ETL Database - https://www.stitchdata.com/etldatabase/etl-process/

[3] Data Visualization: A Beginner's Guide - https://www.tableau.com/learn/articles

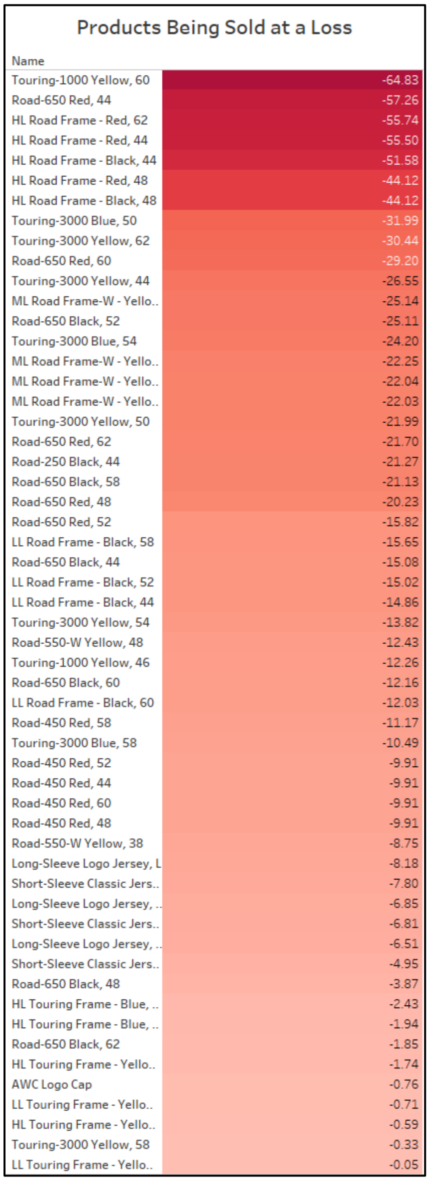
APPENDIX

A. R Source Code for Data Processing and Loading

```
library(RODBC)
library(dplyr)
library(lubridate)
library(DBI)
library(RPostgreSQL)


#Step 1 - download data set from - http://castle.ewi.utwente.nl/datasciencedata/BI/

#Step 2 - Microsoft access database engine installed - https://www.microsoft.com/en-us/download/details.aspx?id=13255 (mine is the 32-bit version)

#Step 3 - Load the data into data0
data0 <- odbcConnectAccess2007('D:/Masters/Q1/DS/Project/SalesProd_Database.accdb')

#Step 4 - load the different sheets into the respective tables
employee <- sqlFetch(data0, 'Employee') %>%
  # Select relevant rows
  select(EmployeeBusinessEntityID, ParentEmployeeBusinessEntityID,
         FirstName, MiddleName, LastName, Title, DepartmentName,
         StartDate, VacationHours, SickLeaveHours, BaseRate,
         BirthDate, Gender) %>%
  # Rename columns
  rename(EmployeeID = EmployeeBusinessEntityID,
         ParentID = ParentEmployeeBusinessEntityID) %>%
  # Filter out duplicates by grouping by EmployeeID
  group_by(EmployeeID) %>%
  # Since each employee still works at the company,
  # summarising StartDate by using the min() function is enough
  mutate(StartDate = min(StartDate),
         BirthDate = ymd(BirthDate),
         StartDate = ymd(StartDate)) %>%
  # Ungroup the data again and filter out duplicate rows
  ungroup() %>%
  distinct()
sales <- sqlFetch(data0, 'SalesOrderDetail') %>%
  # Select relevant columns
  select(ProductID, UnitPrice, OrderQty,
       UnitPriceDiscount, DueDate, ShipDate) %>%
  # Add Late column
  mutate(RowID = row_number(),
         Late = factor(ifelse(ShipDate > DueDate, "Late", "NotLate")),
         DueDate = ymd(DueDate),
         ShipDate = ymd(ShipDate))


product <- sqlFetch(data0, 'Product') %>%
  # Select relevant columns
  select(ProductID, Name, SafetyStockLevel, ReorderPoint, ProductSubcategoryID,
         StandardCost, ListPrice) %>%
  # Only keep rows for which we have Sales data
  filter(ProductID %in% sales$ProductID)

product_category <- sqlFetch(data0, 'ProductCategory')

product_subcategory <- sqlFetch(data0, 'ProductSubcategory') %>%
  filter(ProductSubcategoryID %in% product$ProductSubcategoryID)

product_vendor <- sqlFetch(data0, 'ProductVendor') %>%
  select(ProductID, BusinessEntityID, AverageLeadTime, StandardPrice,
         MinOrderQty, MaxOrderQty) %>%
  filter(ProductID %in% sales$ProductID)

close(data0)

drv <- dbDriver("PostgreSQL")
con <- dbConnect(drv, port = 5432, host = "bronto.ewi.utwente.nl",
                 dbname = "dab_ds20211b_129",
                 user = "dab_ds20211b_129",
                 password = "Mp2+Dw6Cyohz0ZiT",
                 options="-c search_path=project")
dbWriteTable(con, "employee", value = employee, overwrite = T, row.names = F)
dbWriteTable(con, "product", value = product, overwrite = T, row.names = F)
dbWriteTable(con, "product_category", value = product_category, overwrite = T, row.names = F)
dbWriteTable(con, "product_subcategory", value = product_subcategory, overwrite = T, row.names
= F)
dbWriteTable(con, "product_vendor", value = product_vendor, overwrite = T, row.names = F)
dbWriteTable(con, "sales", value = sales, overwrite = T, row.names = F)
dbListTables(con)

str(dbReadTable(con,"employee"))
str(dbReadTable(con,"product"))
str(dbReadTable(con,"product_category"))
str(dbReadTable(con,"product_subcategory"))
str(dbReadTable(con,"product_vendor"))
str(dbReadTable(con,"sales"))
dbGetQuery(con,"SELECT      table_name      FROM      information_schema.tables      WHERE
table_schema='project'")

## to get the tables from schema project

str(dbReadTable(con, c("project", "employee")))
str(dbReadTable(con, c("project", "product")))
str(dbReadTable(con, c("project", "product_category")))
str(dbReadTable(con, c("project", "product_subcategory")))
str(dbReadTable(con, c("project", "product_vendor")))
str(dbReadTable(con, c("project", "sales")))
```

B. Screenshots of the findings

- Finding 1 - The Sales Price of some products is lesser than the cost incurred by the organization for procuring/ producing them. This results in quite a few losses for the company.



Products Being Sold at a Loss

Name	
Touring-1000 Yellow, 60	-64.83
Road-650 Red, 44	-57.26
HL Road Frame - Red, 62	-55.74
HL Road Frame - Red, 44	-55.50
HL Road Frame - Black, 44	-51.58
HL Road Frame - Red, 48	-44.12
HL Road Frame - Black, 48	-44.12
Touring-3000 Blue, 50	-31.99
Touring-3000 Yellow, 62	-30.44
Road-650 Red, 60	-29.20
Touring-3000 Yellow, 44	-26.55
ML Road Frame-W - Yello..	-25.14
Road-650 Black, 52	-25.11
Touring-3000 Blue, 54	-24.20
ML Road Frame-W - Yello..	-22.25
ML Road Frame-W - Yello..	-22.04
ML Road Frame-W - Yello..	-22.03
Touring-3000 Yellow, 50	-21.99
Road-650 Red, 62	-21.70
Road-250 Black, 44	-21.27
Road-650 Black, 58	-21.13
Road-650 Red, 48	-20.23
Road-650 Red, 52	-15.82
LL Road Frame - Black, 58	-15.65
Road-650 Black, 44	-15.08
LL Road Frame - Black, 52	-15.02
LL Road Frame - Black, 44	-14.86
Touring-3000 Yellow, 54	-13.82
Road-550-W Yellow, 48	-12.43
Touring-1000 Yellow, 46	-12.26
Road-650 Black, 60	-12.16
LL Road Frame - Black, 60	-12.03
Road-450 Red, 58	-11.17
Touring-3000 Blue, 58	-10.49
Road-450 Red, 52	-9.91
Road-450 Red, 44	-9.91
Road-450 Red, 60	-9.91
Road-450 Red, 48	-9.91
Road-550-W Yellow, 38	-8.75
Long-Sleeve Logo Jersey, L	-8.18
Short-Sleeve Classic Jers..	-7.80
Long-Sleeve Logo Jersey, ..	-6.85
Short-Sleeve Classic Jers..	-6.81
Long-Sleeve Logo Jersey, ..	-6.51
Short-Sleeve Classic Jers..	-4.95
Road-650 Black, 48	-3.87
HL Touring Frame - Blue, ..	-2.43
HL Touring Frame - Blue, ..	-1.94
Road-650 Black, 62	-1.85
HL Touring Frame - Yello..	-1.74
AWC Logo Cap	-0.76
LL Touring Frame - Yello..	-0.71
HL Touring Frame - Yello..	-0.59
Touring-3000 Yellow, 58	-0.33
LL Touring Frame - Yello..	-0.05

Figure 7: Finding 1 - Products being sold at a loss

- **Finding 2: Products will lowest profit margin are caps, jerseys, and road frames. These products have done badly over all 4 years (2006, 2007, 2008 and 2009).**
  - In 2009, caps performed slightly better than the other years.
  - The sales of jerseys is inversely proportional to profit in 2008.

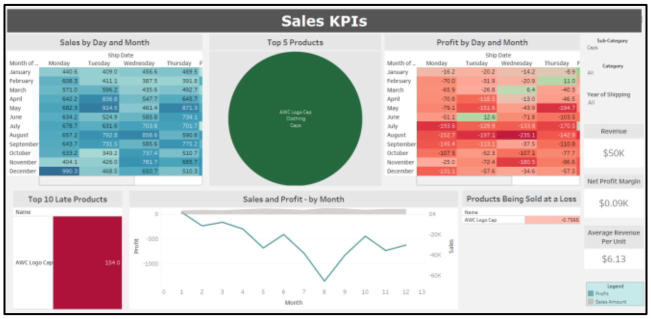


Figure 8: Finding 2a - Caps - Low profit Margin

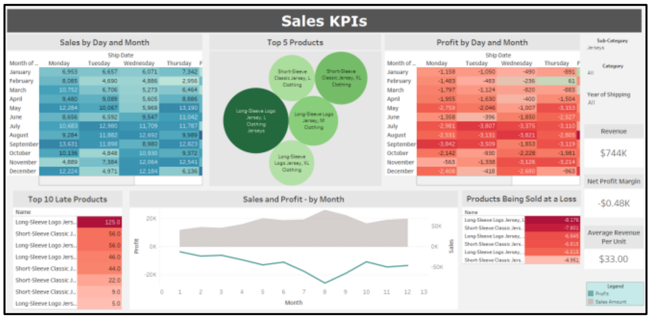


Figure 9: Finding 2b - Jerseys - Low profit Margin

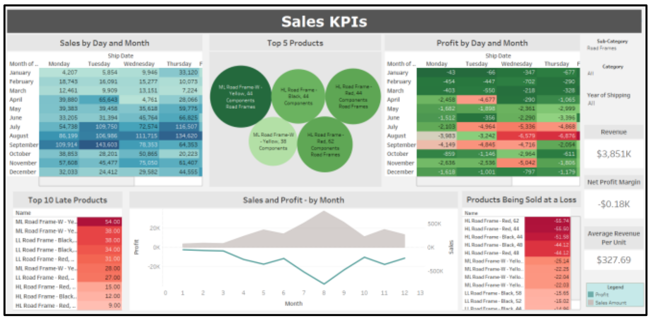


Figure 10: Finding 2c - Road Frames - Low profit Margin

Finding 3: Products with high and fluctuating profit margins have been listed below:

- High profit Margin - Cleaners, Hydration Packs, Gloves, Vests
- Fluctuating Profit Margin - Mountain bikes, Road bikes, touring bikes, touring frames

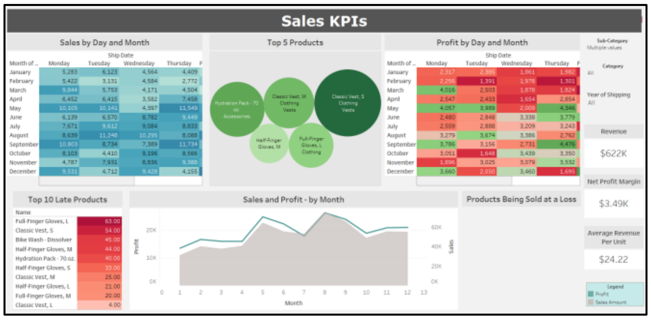


Figure 11: Finding 3a - Products with High Profit Margin – Cleaners, Hydration Packs, Gloves, Vests

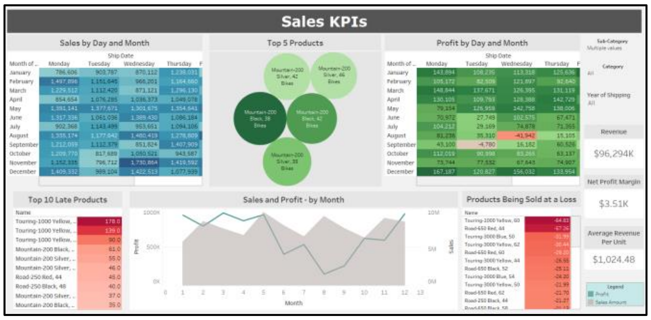


Figure 12: Finding 3b - Mountain Bikes, Road Bikes, Touring Bikes, Touring Frames

Finding 4: The highest revenue per unit is given by the category "Bikes", specifically "Mountain Bikes", although it has a fluctuating profit margin.

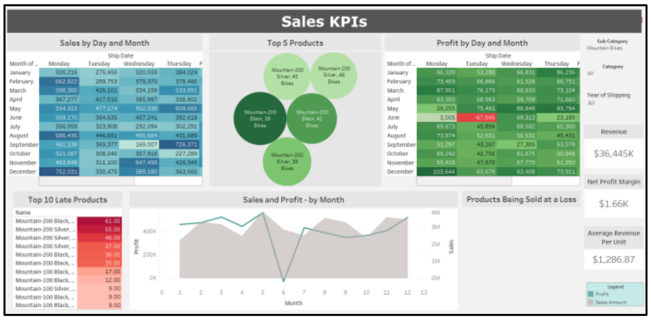


Figure 13: Finding 4 - Highest Average Revenue per Unit - Mountain Bikes

Finding 5: The male to female ratio is not proportional (84:206 males to females).

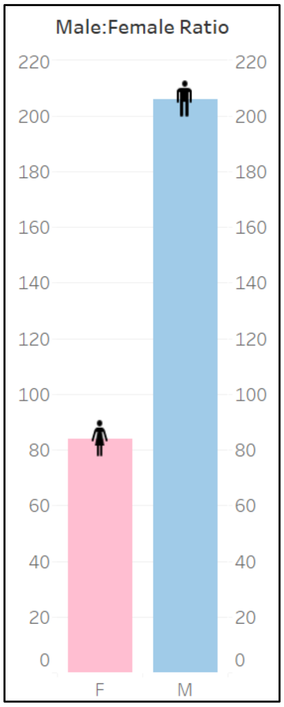


Figure 14: Disproportionate Male-Female Ratio

Finding 6: There are very few people in marketing (9) and shipping and receiving (6) as compared to production (179).

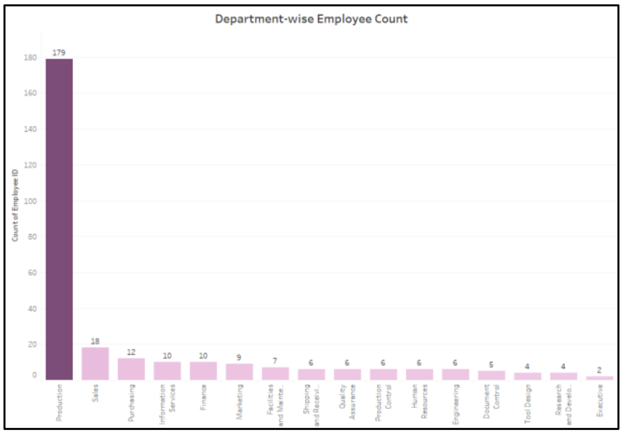


Figure 15: Finding 6 - Employee distribution across departments

- **Finding 7: The trends in vacation are similar across both genders.**

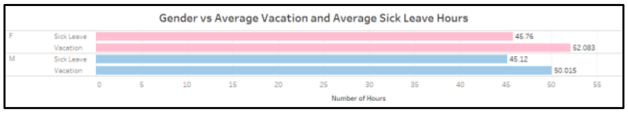
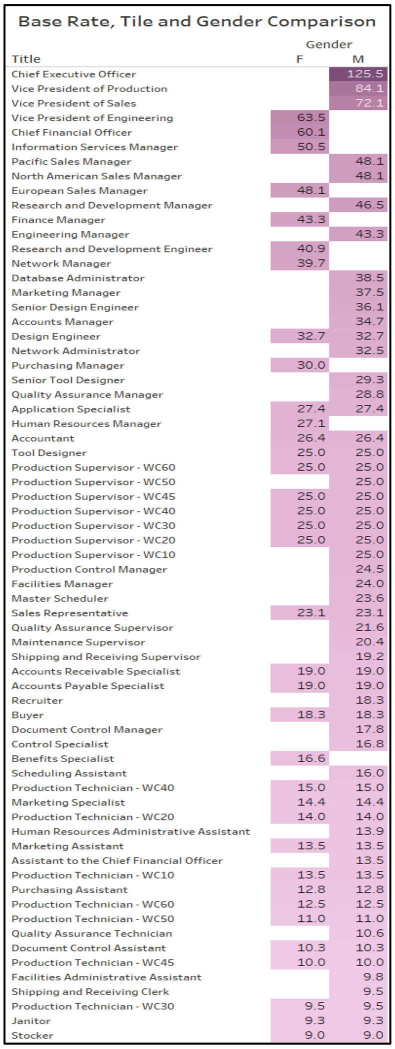


Figure 16: Finding 7 - Proportional average Vacation and Sick Leave hours for males and females

- **Finding 8: Base rates of people in the same level are comparable. But some roles do not have women employees at all.**



Base Rate, Tile and Gender Comparison

	Gender	
Title	F	M
Chief Executive Officer		125.5
Vice President of Production		84.1
Vice President of Sales		72.1
Vice President of Engineering	63.5	
Chief Financial Officer	60.1	
Information Services Manager	50.5	
Pacific Sales Manager		48.1
North American Sales Manager		48.1
European Sales Manager	48.1	
Research and Development Manager		46.5
Finance Manager	43.3	
Engineering Manager		43.3
Research and Development Engineer	40.9	
Network Manager	39.7	
Database Administrator		38.5
Marketing Manager		37.5
Senior Design Engineer		36.1
Accounts Manager		34.7
Design Engineer	32.7	32.7
Network Administrator		32.5
Purchasing Manager	30.0	
Senior Tool Designer		29.3
Quality Assurance Manager		28.8
Application Specialist	27.4	27.4
Human Resources Manager	27.1	
Accountant	26.4	26.4
Tool Designer	25.0	25.0
Production Supervisor - WC60	25.0	25.0
Production Supervisor - WC50		25.0
Production Supervisor - WC45	25.0	25.0
Production Supervisor - WC40	25.0	25.0
Production Supervisor - WC30	25.0	25.0
Production Supervisor - WC20	25.0	25.0
Production Supervisor - WC10		25.0
Production Control Manager		24.5
Facilities Manager		24.0
Master Scheduler		23.6
Sales Representative	23.1	23.1
Quality Assurance Supervisor		21.6
Maintenance Supervisor		20.4
Shipping and Receiving Supervisor		19.2
Accounts Receivable Specialist	19.0	19.0
Accounts Payable Specialist	19.0	19.0
Recruiter		18.3
Buyer	18.3	18.3
Document Control Manager		17.8
Control Specialist		16.8
Benefits Specialist	16.6	
Scheduling Assistant		16.0
Production Technician - WC40	15.0	15.0
Marketing Specialist	14.4	14.4
Production Technician - WC20	14.0	14.0
Human Resources Administrative Assistant		13.9
Marketing Assistant	13.5	13.5
Assistant to the Chief Financial Officer		13.5
Production Technician - WC10	13.5	13.5
Purchasing Assistant	12.8	12.8
Production Technician - WC60	12.5	12.5
Production Technician - WC50	11.0	11.0
Quality Assurance Technician		10.6
Document Control Assistant	10.3	10.3
Production Technician - WC45	10.0	10.0
Facilities Administrative Assistant		9.8
Shipping and Receiving Clerk		9.5
Production Technician - WC30	9.5	9.5
Janitor	9.3	9.3
Stocker	9.0	9.0

Figure 17: Finding 8 - Comparison between base rates for designations and roles where there are very few women employees.