

PROFESSIONAL TRAINING REPORT
Sathyabama Institute of Science and Technology
(Deemed to be University)

Submitted in partial fulfilment of the requirements for the award
of Bachelor of Technology Degree in Information Technology

By

Sagar Kumar

(40120076)



DEPARTMENT OF INFORMATION TECHNOLOGY

SCHOOL OF COMPUTING

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

JEPPIAAR NAGAR, RAJIV GANDHI SALAI,

CHENNAI, TAMILNADU – 600119

April 2023



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE

(www.sathyabama.com)

DEPARTMENT OF INFORMATION TECHNOLOGY

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of Sagar Kumar (Reg. No. 40120076) who carried out the project as a "ETL Processing and Analytics for Client's Historical Data" under our supervision from January 2023 to April 2024.

R.S. 25/4/23

Internal Guide

Dr. R Subhashini, M.E., Ph.D.

R.S. 25/4/23

Head of the Department

Dr. R Subhashini, M.E., Ph.D.

29/4/2023

db

Internal Examiner

[Signature]

External Examiner

DECLARATION

I **Sagar Kumar (Reg. No. 40120076)** hereby declare that the Project Report entitled “**ETL Processing and Analytics for Client’s Historical Data**” is done by me under the guidance of **Dr. R Subhashini, M.E., Ph.D.**, and is submitted in partial fulfilment of the requirements for the award of Bachelor of Technology degree in Information Technology.

DATE: 29/04/2023

A handwritten signature in black ink that reads "Sagar Kumar". The signature is written in a cursive style with a large, stylized 'S' at the beginning.

PLACE: Chennai

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to the **Board of Management** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T Sasikala, M.E., Ph.D.**, Dean, School of Computing **Dr. R Subhashini, M.E., Ph.D.**, Head of the Department of Information Technology for providing me with necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. R Subhashini, M.E., Ph.D.** for her valuable guidance, suggestions and constant encouragement that paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Information Technology** who were helpful in many ways for the completion of the project.

CERTIFICATE OF COMPLETION



TO WHOMSOEVER IT MAY CONCERN

This is to certify that **Mr.Sagar Kumar** has done his internship as Data Analyst at NYL10 Software private Limited, Chennai, from 11/01/2023 to 31/03/2023.

He has worked on a project on Data Analytics. This project was aimed at the purpose of learning and understanding the concept of data analytics,structure of the data.He also worked with extraction, transformation and loading.

He was very Proactive,very effectively delivering an excellent outcome as expected by the company.

We wish him all the best for his upcoming career.

Sathish

A handwritten signature in blue ink, appearing to be "Sathish", written over a light blue circular stamp.

Chief Technology Officer

NYL10 Software private Limited

NYL10 Softwares Private Limited
CIN-U72501TN2018PTC124718
Estate Main Rd, Industrial Estate, Perungudi, Chennai, Tamil Nadu 600096

ABSTRACT

Any company generates large amounts of data in multiple forms. But these data are sometimes very tough to manage and sometimes make no sense at all. A large part of this data remains in the archives. In a majority of cases if the company is small or is not related to the IT industry then the data generated by these firms are not much used.

Using this data and incorporating the extracted knowledge into the business can change the whole dynamics of the concerned business. These data can be used to boost the performance of the company in terms of overall profit and customer relations. It can help to draw attention towards the other parts of the business which had been overlooked.

The work mentioned in this report has been done as an internship with a Consulting firm based in Chennai. The major part of the work was to learn about the usability of different types of data and the scope it brought with it. The work also gave a little exposure to ETL or Extract, Transform & Load processing of the data. The details of the work are mentioned in the different sections of the report.

TABLE OF CONTENT

Chapter No.	Title	Page No.
1	Introduction	1
2	Problem Statement	3
	2.2 Objectives	4
3	Literature Survey	5
4	Technological Overview	7
	4.1 Libraries used	7
	4.2 Tools Used	12
5	Architecture & Methodology	15
	5.1 Data Extraction	16
	5.2 Transformation and Loading	17
	5.3 Diagnostic System	18
6	Results & Observation	21
	6.1 Mechanical Data	21
	6.2 Electrical Data	22
	6.3 Investigation Report Data	26
7	References	33

List of Images

Sl. No	Figure Number	Figure Description
1	5.1	Basic Architecture
2	5.2	Process for Extraction of Data
3	5.3	Process for the transformation of the extracted Data
4	5.4	The architecture of the Diagnostic system
5	6.1	A chart demonstrating the number of different components that are failing
6	6.2	A chart demonstrating the maximum failures occurring in different periods
7	6.3	A chart demonstrating the equipment failure in different groups
8	6.4	A chart demonstrating the electrical issues that occurred in different types of vessels
9	6.5	A chart demonstrating the count of electrical issues occurring at either sea or docks
10	6.6	A chart demonstrating the presence of an electrician on the vessels
11	6.7	A chart demonstrating the seriousness of problems occurring
12	6.8	A chart demonstrating the occurrence of the issue in vessels of different ages
13	6.9	A chart demonstrating the occurrence of the issue in vessels built by different manufactures
14	6.10	Chart demonstrating the occurrence of issues in different wind speeds

15	6.11	Chart demonstrating the performance of different vessels based on age
16	6.12	A chart demonstrating the occurrence of issues in terms of the Beaufort Scale
17	6.13	A chart demonstrating the occurrence of issues in different seasons

List of Tables

Sl. No	Table Number	Table Description
1	4.1	List of modules and technologies used

Chapter 1

INTRODUCTION

The work mentioned in this report has been done at NYL TECHNOLOGY, Chennai from January 2023 to March 2024. Due to company's policy, all the details of client and asserts have been changed. The names used in this work is only for the sole purpose of academic presentation. This work is part of a bigger workflow that the company had already been working on for one of their clients. The work included mostly two parts, the first one was ETL and the second one was extracting a knowledge base from the collected data that can be used in the business.

The client was in the cargo transportation business which involved a lot of marine vessels. The job nature forced the company to generate a variety of data which was either present as PDF document reports or Excel sheets full of textual and image data. The data had been recorded over 6 to 7 years manually. Since there was less and sometimes no standard to store the data, a big part of the collected data was random and made no sense. The client had three important data:

- 1) ***Mechanical Failure Data*** – These are the records of mechanical components that failed on a particular vessel.
- 2) ***Electrical Failure Data*** – These are the records of electrical components that failed on a particular vessel.
- 3) ***Incident Report Data*** – These are the records of investigations and other related problems that had occurred with a vessel.

The client was more involved in operations so had no time to look at the data and think of the possibilities of using the historical data. But with the increasing popularity and need of the time it decided to give it a try.

The ETL phase in this work involved understanding the data, anomaly, nature, and quality of data. And knowledge extraction involved understanding the business and business needs.

Chapter 2

PROBLEM STATEMENT

This work involved multiple problem statements but two of the most important were:

1. Collecting Data and Storing the Data
2. Improve business based on the needs.

One of the tough jobs is to convert heterogeneous unstructured data into structured data. The data provided for this job was not collected with a fixed standard. Since the data was collected by different people and placed in the same place the chances of irregularity were too much. The data was collected manually so human errors and poor data quality were sure to be there.

Another thing with business needs is that they can be very dynamic and change frequently. As time goes on the same business can be working with something else. This versatility doesn't guarantee that the business knowledge generated today will be valid tomorrow. Another problem with real-life business is the gap in business understanding by the engineers. The engineering team can spend time working on something which would turn out to be useless to the actual business. So the engineers need to look at the work and criticize it as the business owner themselves.

2.2. OBJECTIVES

The nature of the work brought a lot of objectives to be completed. Some of the main objectives are mentioned below.

- Extracting heterogeneous data from multiple data sources.
- Transforming the data to a workable and proper format.
- Store this data in a persistent and efficient format for future use by any party.
- Standardize the data and improve data quality.
- Create a collection rule which can be used to collect data in the future.
- Increase the usefulness of the collected data.
- Create a module that can easily be merged into the company's ongoing product.
- Understand the business logic and generate a useful knowledge base.
- Provide a diagnostic system for the client to access the knowledge base and the collected data.
- To provide maximum information with the least amount of data.
- To improve user experience and make data discovery fast

Chapter 3

LITERATURE SURVEY

Extract, Transform, Load (ETL) is a process used to extract data from various sources, transform it into a format suitable for analysis, and load it into a data warehouse or other analytical system. ETL is a crucial step in the data analytics process and can have a significant impact on the quality and accuracy of insights derived from data.

A number of studies have been conducted on ETL and its role in data analytics. For example, Katal, Wazid, and Goudar (2013) conducted a study on ETL tools and their capabilities. They found that ETL tools provide a variety of functionalities, such as data profiling, data cleansing, and data validation, which can significantly improve data quality.

In addition to ETL tools, a number of studies have focused on the use of ETL frameworks and methodologies. For example, Lin, Chen, and Liu (2018) proposed a framework for ETL process design and implementation based on a service-oriented architecture. They found that their framework can significantly improve the scalability, flexibility, and reusability of ETL processes.

Another important aspect of ETL is data integration, which involves combining data from multiple sources. A study by El-Sappagh, Elmogy, and Saddik (2018) focused on data integration techniques for ETL processes. They found that semantic integration, which involves mapping data from different sources based on their meaning, can significantly improve the accuracy and efficiency of ETL processes.

Data analytics is another important area related to ETL. A study by Davenport and Patil (2012) highlighted the importance of data analytics in organizations and the need for skilled professionals in this area. They found that organizations that invest in data analytics and hire skilled data analysts can gain significant competitive advantages.

Overall, the literature on Basic ETL and Data Analytics highlights the importance of ETL in data analytics processes and the need for effective ETL tools, frameworks, and methodologies. It also emphasizes the importance of data integration and the role of skilled data analysts in organizations.

Chapter 4

TECHNOLOGICAL OVERVIEW

This whole work is based on Python 4.9.0. For working on different modules, different Python libraries and tools have been brought to use. The list of the libraries and tools used are as follows:

Table 4.1 List of Modules Used

S. no.	Name	Type	Purpose
1	Pandas	P Y T H O N L I B R A R Y	For processing data
2	NumPy		
3	OpenCV		For working with images
4	PyTesseract		For handling OCR
5	Pillow		For handling and adding features to image
6	FuzzyWuzzy		For supporting fuzzy match
7	Streamlit		For Making a Diagnostic system
8	PyMuPDF		For working with PDF document
9	Matplotlib		For graphical representation
10	Jupyter Notebook	IDE	For writing and executing the code
11	Google OCR	Tool	For OCR
12	AWS CodePipeline	Tool	For Automating the workflow
13	AWS EC2	Tool	For hosting

4.1. LIBRARIES USED

4.1.1. *Pandas*

It is a widely-used data analysis and manipulation library for Python. It provides a lot of functions and methods that expedite the data analysis and pre-processing steps. It also provides fast, flexible, and expressive data structures working with relational or labelled or both easy and intuitive. Considered a fundamental high-level building block in performing practical, real-world data analysis in Python. Has powerful tools like Dataframe and Series for analyzing

Command for Installation:

```
pip install pandas
```

4.1.2. *NumPy*

It stands for Numerical Python, which is a library consisting of multidimensional array objects and a collection of countless routines for processing those arrays. Using this mathematical and logical operations on arrays can be performed. The difference in using NumPy from pandas is, it works on numerical data whereas pandas on tabular data.

Command for Installation:

```
pip install numpy
```

4.1.3. *OpenCv*

OpenCV, or Open-Source Computer Vision, is a popular computer vision library that provides developers with a wide range of image and video processing algorithms. It was initially developed by Intel in the late 1990s but has since become a widely used open-source library. OpenCV supports multiple programming languages, making it a versatile tool for computer

vision tasks. The library provides a variety of functionalities, such as image and video input/output, image filtering, object detection and tracking, and much more. OpenCV is an excellent choice for developers looking to build applications that analyze visual data, such as video surveillance systems, robotics, and augmented reality.

Command for Installation:

```
pip install Opencv-python
```

4.1.4. *PyTesseract*

PyTesseract is a Python wrapper for Tesseract-OCR, an open-source optical character recognition (OCR) engine that can read and recognize text from images. Tesseract-OCR was initially developed by Hewlett-Packard in the 1980s and was later released as an open-source project in 2006. PyTesseract simplifies the use of Tesseract-OCR by providing an easy-to-use Python interface. It supports various image formats and can be customized to improve OCR accuracy. PyTesseract is a great tool for developers looking to extract text from images, such as scanned documents, receipts, and business cards.

Command for Installation:

```
pip install pytesseract
```

4.1.5. *Pillow*

Pillow is a Python Imaging Library that adds support for opening, manipulating, and saving various image file formats. Pillow is a fork of the Python Imaging Library (PIL), which was created by Fredrik Lundh in 1996. Pillow provides a wide range of image processing capabilities, including resizing, cropping,

rotating, filtering, and more. Pillow is an excellent choice for developers who need to perform basic image processing tasks quickly and efficiently. It also provides support for advanced features such as image enhancement, image segmentation, and colour manipulation.

Command for Installation:

```
pip install pillow
```

4.1.6. *FuzzyWuzzy*

FuzzyWuzzy is a Python library that provides advanced string matching and fuzzy search capabilities. It was developed by SeatGeek, a popular ticket search engine, to help them match ticket listings with events accurately. FuzzyWuzzy uses the Levenshtein distance algorithm to calculate the difference between two strings. It can be used to compare and match strings that are not exactly the same, such as miss-spelt words or variations in formatting. FuzzyWuzzy is especially useful for data cleaning and normalization tasks where accurate string matching is essential.

Command for Installation:

```
pip install fuzzywuzzy
```

4.1.7. *Streamlit*

Streamlit is an open-source Python library that simplifies the development of data science and machine learning applications. It was created in 2019 by Adrien Treuille and Amanda Kelly to address the challenges they faced in building data science applications. Streamlit provides an intuitive interface that allows

developers to create interactive and visually appealing applications quickly. It supports real-time updates, making it easy to build applications that respond to changes in data. Streamlit is an excellent choice for developers looking to build data-driven applications without having to write complex front-end code.

Command for Installation:

```
pip install streamlit
```

4.1.8. *PyMuPDF*

PyMuPDF is a Python wrapper for the MuPDF library, a lightweight PDF and XPS viewer that provides a wide range of PDF manipulation capabilities. PyMuPDF can be used to extract text and images from PDFs, add watermarks, merge and split PDFs, and much more. It is a great tool for developers who need to work with PDFs programmatically, such as those working on document management systems or digital archives.

Command for Installation:

```
pip install pymupdf
```

4.1.9. *Matplotlib*

Matplotlib is used to create 2D graphs and plots by using Python scripts. It comes with a module named Pyplot which makes things easy for plotting by providing a feature to control line styles, font properties, formatting axes, etc. It supports a very wide variety of graphs and plots namely - histograms, bar charts, power spectra, error charts, etc. It is used along with NumPy to provide an environment that is an effective open-source alternative for MatLab. It can also be used with graphics toolkits like PyQt and wxPython.

Command for Installation:

```
pip install matplotlib
```

4.1.10. *Jupyter Notebook*

Jupyter Notebook is an open-source web application that you can use to create and share documents that contain live code, equations, visualizations, and text. Jupyter Notebook is maintained by the people at Project Jupyter.

Jupyter Notebooks are a spin-off project from the IPython project, which used to have an IPython Notebook project itself. The name, Jupyter, comes from the core supported programming languages that it supports: Julia, Python, and R. Jupyter ships with the IPython kernel, which allows us to write our programs in Python, but there are currently over 100 other kernels that we can also use.

Command for Installation:

```
pip install notebook
```

4.2. TOOLS USED

4.2.1. *Google OCR*

Google OCR is a cloud-based optical character recognition (OCR) service that enables developers to extract text from images and convert them into a machine-readable format. It is a part of Google Cloud Vision, a suite of cloud-based vision products that use artificial intelligence to understand images and video. Google OCR uses machine learning algorithms to recognize text in images, making it highly accurate and efficient. The service supports a wide range of image formats and languages, making it suitable for developers worldwide. Google OCR is an excellent tool for developers who need to extract text

from images for a variety of use cases, such as document digitization, automated data entry, and mobile app development.

4.2.2. AWS CodePipeline

AWS CodePipeline is a continuous delivery service provided by Amazon Web Services (AWS) that enables developers to build, test, and deploy code changes quickly and reliably. It is a fully managed service that automates the build, test, and deployment process, making it easier for developers to release new features and updates to their applications. AWS CodePipeline supports multiple programming languages and integrates with various AWS services, such as AWS CodeBuild and AWS CodeDeploy. The service provides a visual pipeline view that allows developers to track the progress of their code changes as they move through the different stages of the pipeline. AWS CodePipeline is an excellent tool for developers who want to streamline their software development process and improve code quality.

4.2.3. AWS EC2

AWS EC2, or Amazon Elastic Compute Cloud, is a cloud-based virtual machine service provided by AWS that enables developers to run applications and services in the cloud. It provides scalable computing capacity in the cloud and supports a wide range of operating systems, programming languages, and applications. AWS EC2 allows developers to choose the instance type that best suits their application's needs, such as compute-optimized instances for high-performance applications and memory-optimized instances for memory-intensive applications.

The service provides features such as load balancing, auto-scaling, and security groups, making it easy for developers to manage their applications in the cloud. AWS EC2 is an excellent tool for developers who want to build and deploy applications in the cloud quickly and easily.

Chapter 5

METHODOLOGY & ARCHITECTURE

Our work starts with collecting all the data sources. Once the sources are collected, they are sorted and grouped based on time and

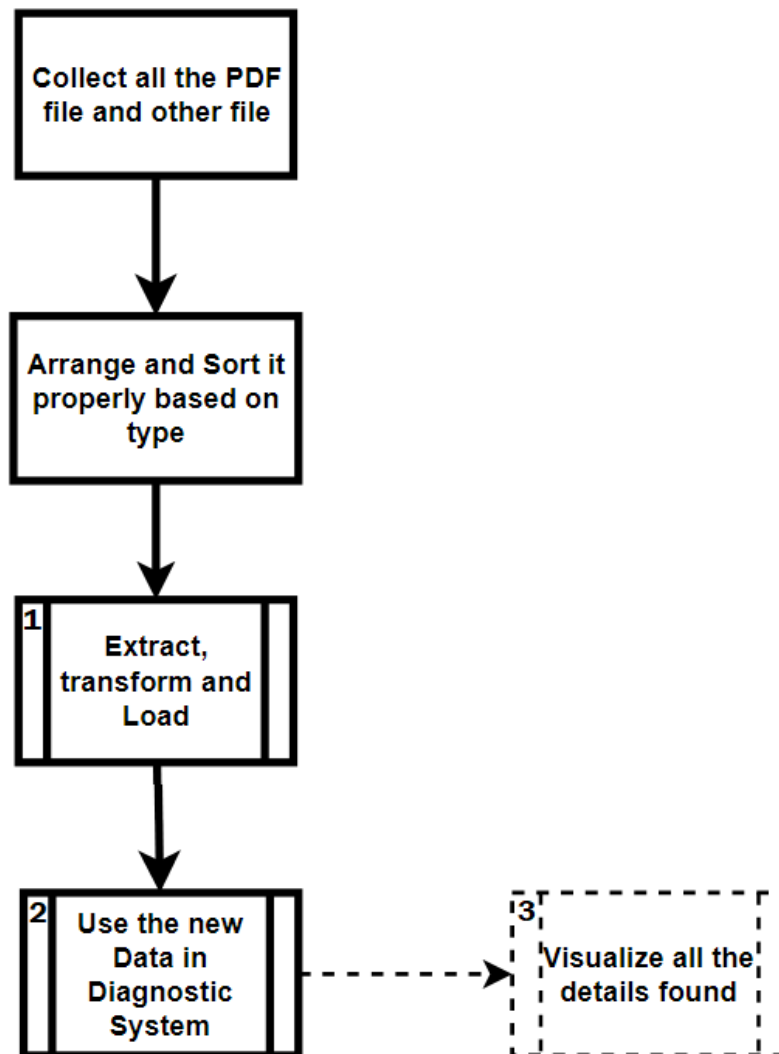


Fig. 5.1 Basic Architecture

date of creation, arrival, and type of data source. Once the source raw data is categorized, it is pushed to the next step. The raw data is processed with ETL steps mentioned in the next steps. The processed data is used to make a diagnostic system. The details from the processed data are also sent for visualization process where different

types of charts are used to explain the data pattern and relation between the business operations and data generated by the business.

6.1. DATA EXTRACTION

Since a major part of the raw data was present as an image or a PDF file containing multiple pages. The text data and tabular data had to be extracted in different ways as mentioned in Fig 6.2.

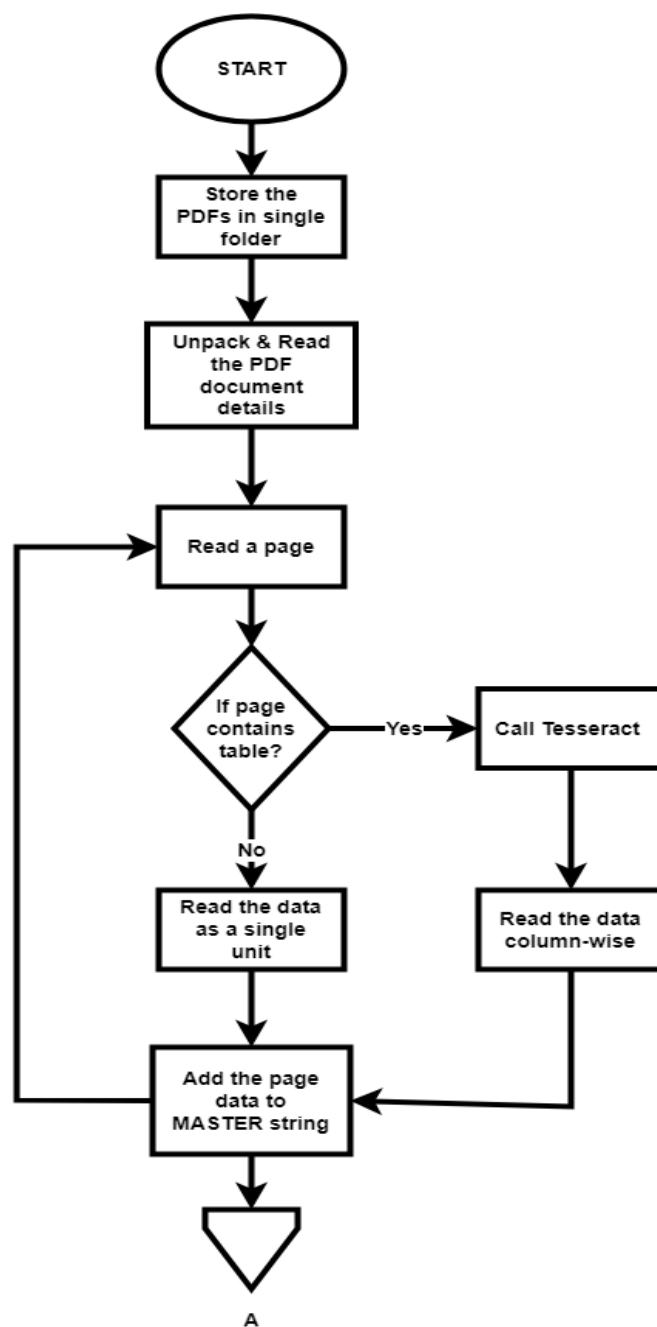


Fig. 5.2 Process for Extraction of Data

The extraction step is the first step in the ETL process, where data is extracted from various sources such as databases, spreadsheets, and web services. The extracted data is often raw and unstructured, requiring further processing and transformation before being loaded into the target system. Data quality is crucial during this step, as poor-quality data can negatively impact the results of subsequent ETL steps.

6.2. TRANSFORMATION AND LOADING

The extracted data had to be standardized and brought into a proper format. Since the data had been extracted from multiple formats so the was not proper. In this process empty values were replaced by some other values, few values like date and other details were brought in standard format as per the transformation rule as per Fig 6.4.

In general, the transformation step involves a variety of operations, such as data cleaning, data validation, data enrichment, data aggregation, and data normalization. The primary goal of the transformation step is to ensure that the data is accurate, complete, and consistent with the target system's schema. Some common transformation techniques used in the ETL process:

- a) **Data Cleaning:** Removing unwanted or duplicate data, correcting errors, and filling in missing values.
- b) **Data Validation:** Checking data for accuracy and completeness, identifying any data that does not conform to the expected format, and removing or fixing it.
- c) **Data Enrichment:** This involves adding additional data from external sources to enrich the existing data set.
- d) **Data Aggregation:** This involves combining data from multiple sources into a single data set.

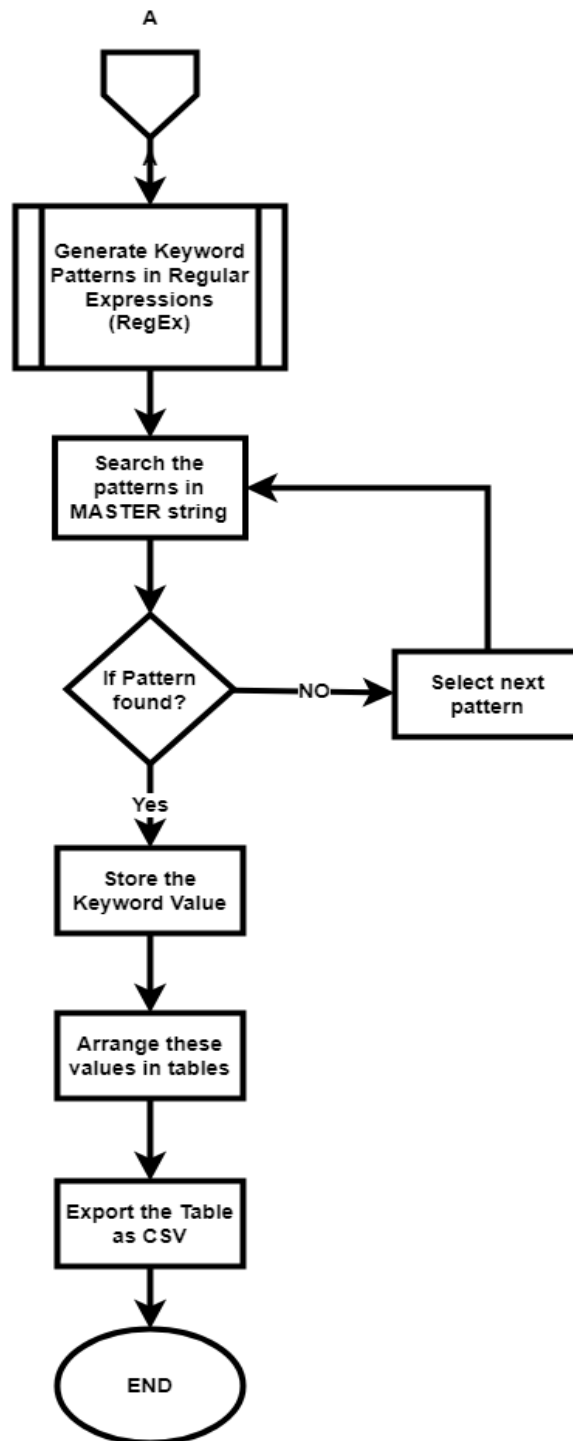


Fig. 5.3 Process for transformation of the extracted Data

- e) **Data Normalization:** This involves organizing data into a consistent format to facilitate easy analysis and querying.

6.4. DIAGNOSTIC SYSTEM

The data from the previous step is used to create a diagnostic

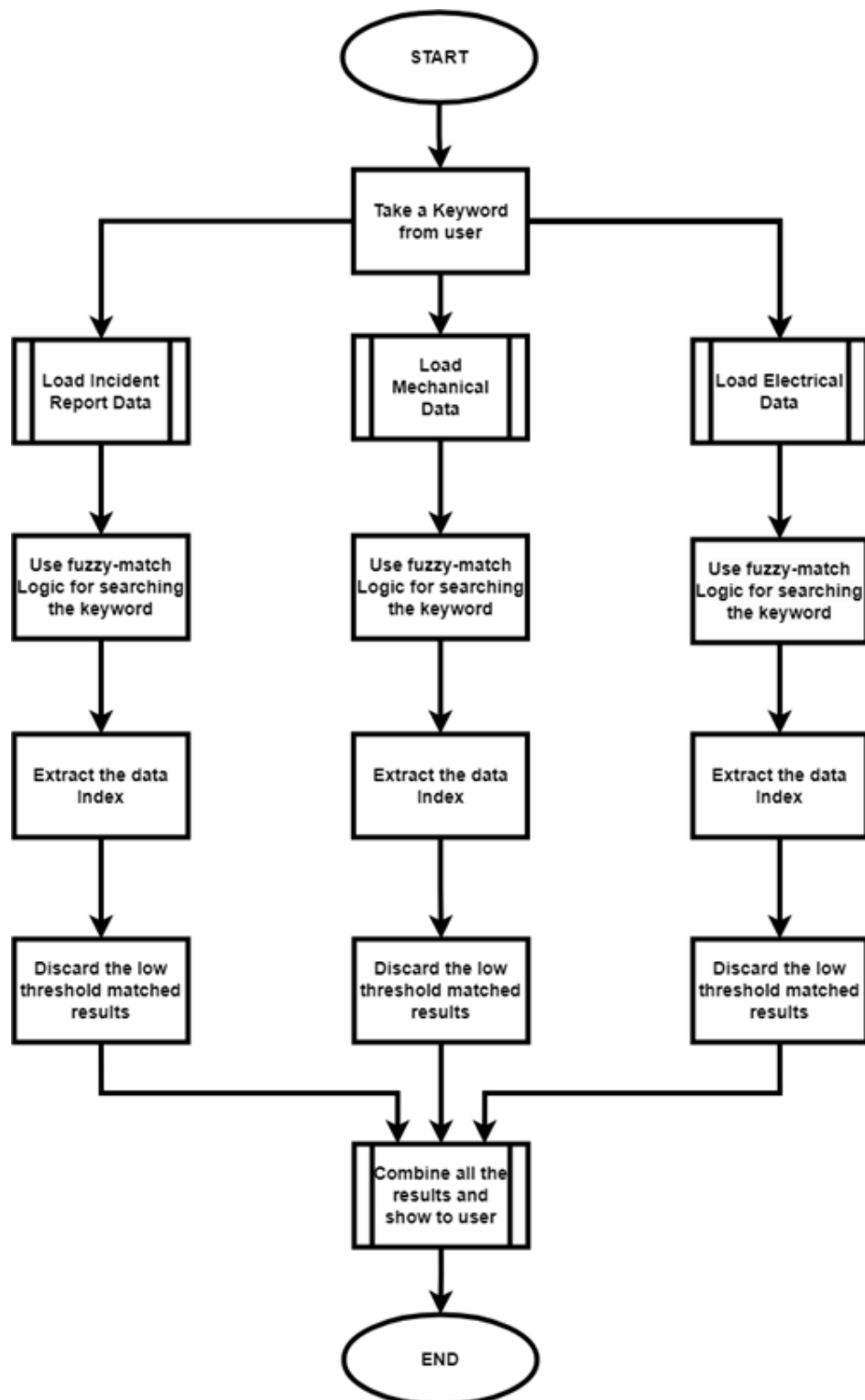


Fig. 5.4 Architecture of Diagnostic system

system that provides features like data searching and monitoring. This system can further be used by the client to have a summarized view of

all the data. It is also supposed to work as a reference for any similar future issue. This system mostly talks about types of mechanical and electrical failures that had occurred with any vessel on the voyage. It also contains details about any other incidents that may have happened on the vessel.

Chapter 6

RESULTS & DISCUSSION

On analysing the final data, the following observations were made.

6.1. MECHANICAL DATA

1. Bigger components tend to fail more than smaller ones.

Suspected Reason: Unclear

Possible Solution: More attention should be paid to these parts/Equipment.

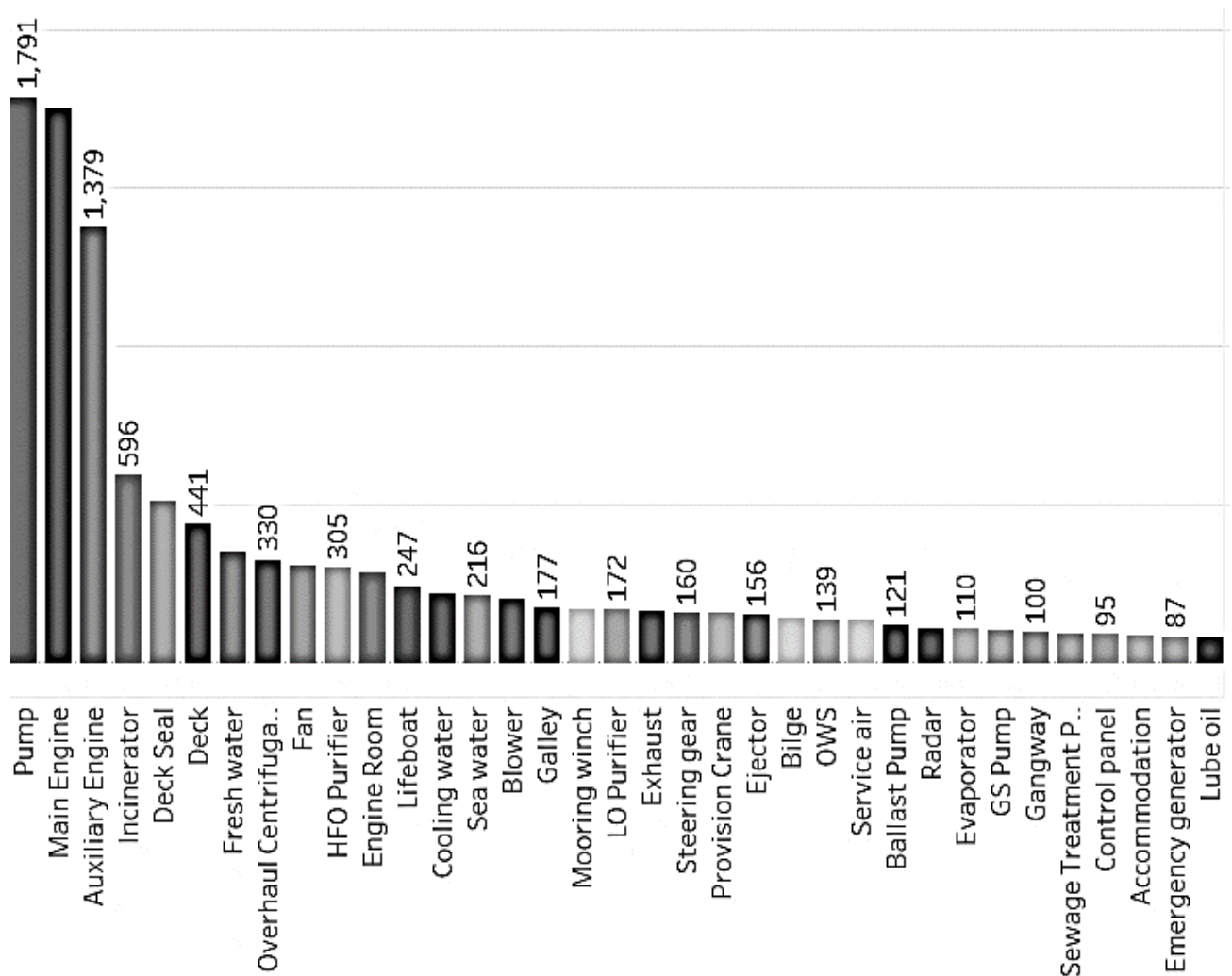


Fig. 6.1 Chart demonstrating the number of different components that are failing

2. Machines failed more during the covid period.

Suspected Reason: There was very less crew and resources available to manage the vessels.

Possible Solution: Make preparations for any such future events.

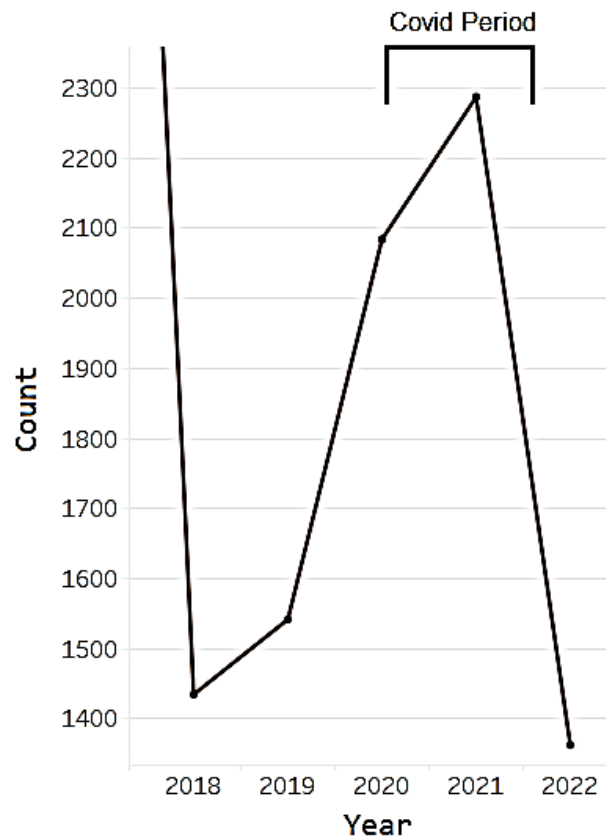


Fig. 6.2 Chart demonstrating the maximum failures occurring in different periods

4. All Groups have the almost same amount of mechanical failure.

Suspected Reason: All groups have been facing the same types of issues in the voyage.

Possible Solution: Unclear

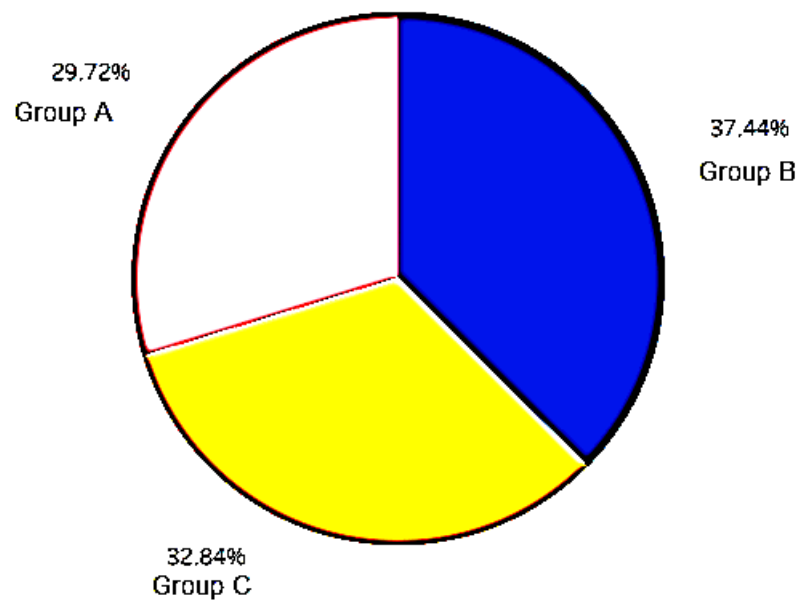


Fig. 6.3 Chart demonstrating the equipment failure in different groups

6.2. ELECTRICAL DATA

1. Oil Tankers have more electrical problems.

Suspected Reason: Oil tankers use more equipment than any other vessel.

Possible Solution: Put more effort into maintenance of such vessels.

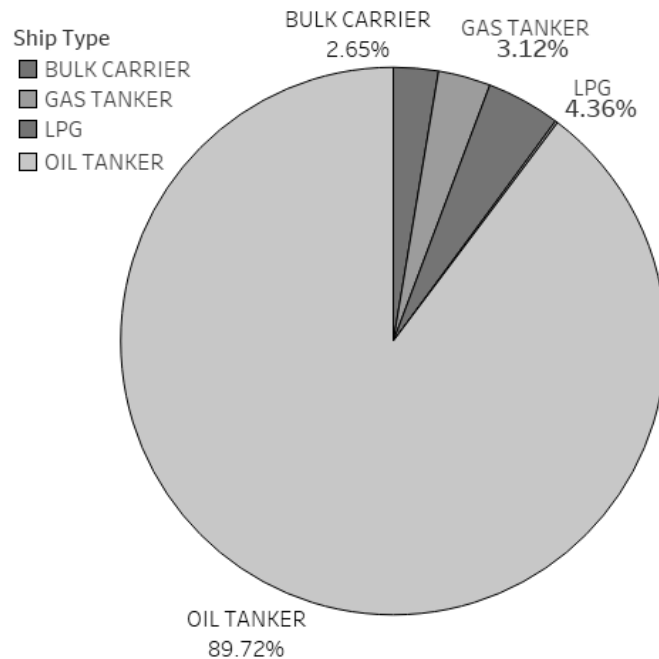


Fig. 6.4 Chart demonstrating the electrical issues occurred in different types of vessels

2. More electrical components fail when vessels are on voyage.

Suspected Reason: Components face rough conditions when at the sea.

Possible Solution: Proper and regular maintenance.

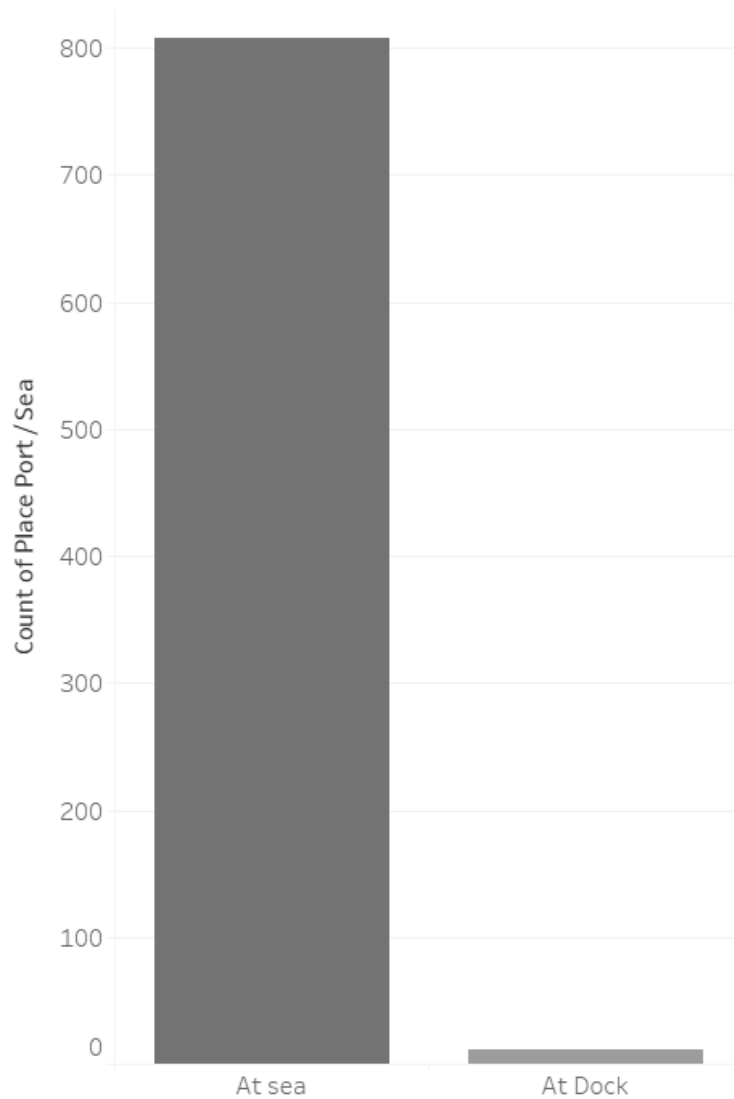


Fig. 6.5 Chart demonstrating the count of electrical issues occurring at either sea or docks

4. Majority of the time an electrician was not present on the vessel.

Suspected Reason: Unawareness of this pattern.

Possible Solution: Always Keep at least one electrician onboard.

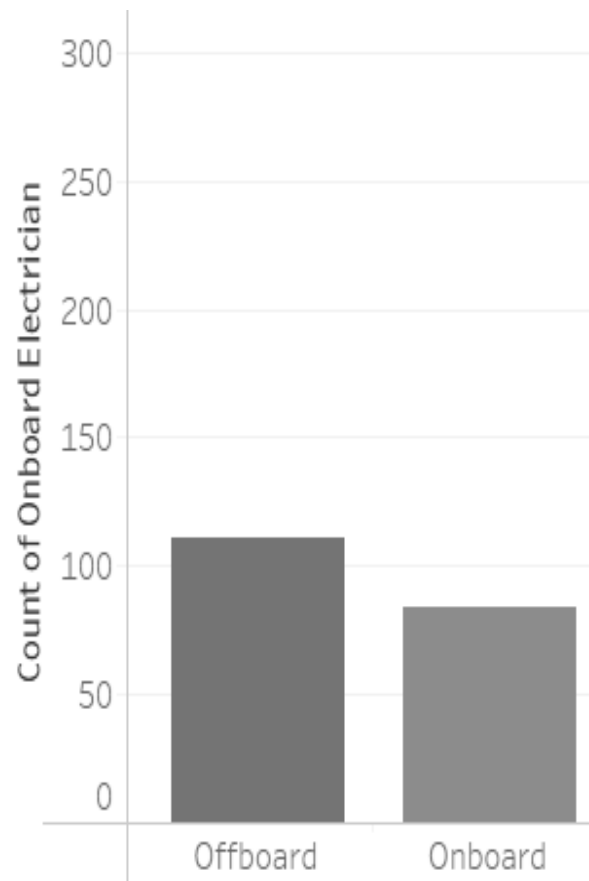


Fig. 6.6 Chart demonstrating the presence of an electrician on the vessels

6.3. INVESTIGATION REPORT DATA

1. Less number of problems with higher risk levels.

Suspected Reason: Proper training and maintenance by crew members.

Possible Solution: Not required

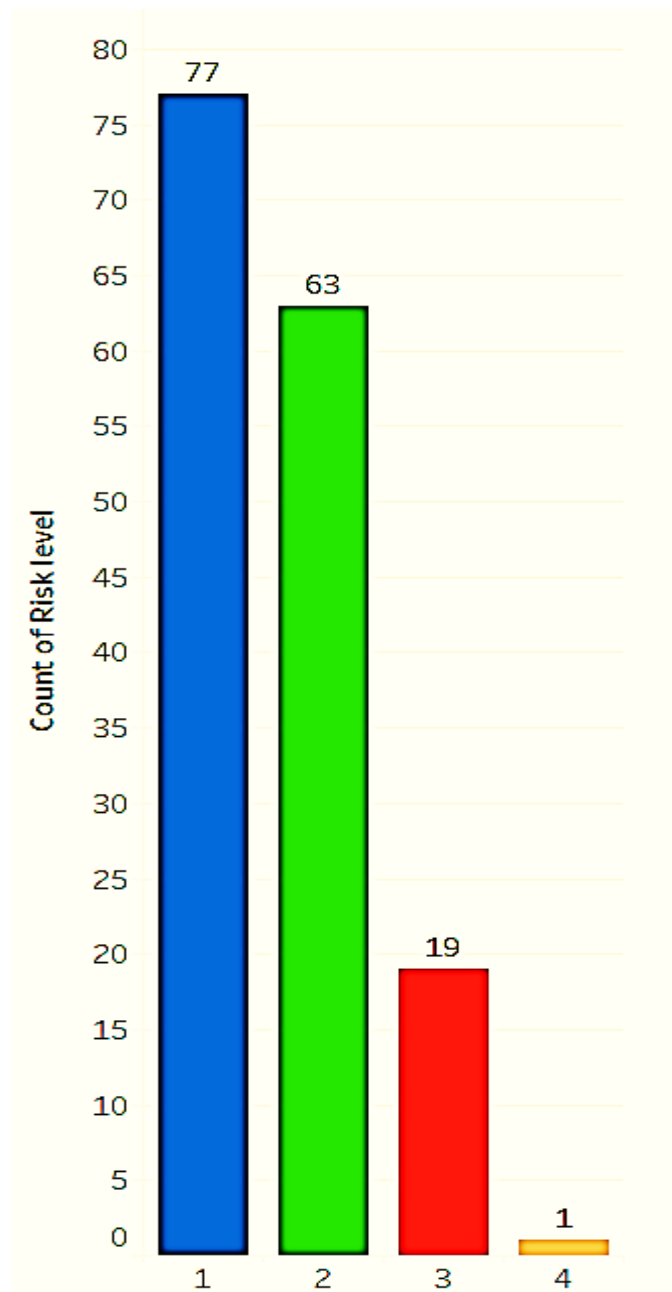


Fig. 6.7 Chart demonstrating the seriousness of problems occurring

2. Mid Aged (10-12 years) vessels have more problems

Suspected Reason: Unclear

Possible Solution: Unclear

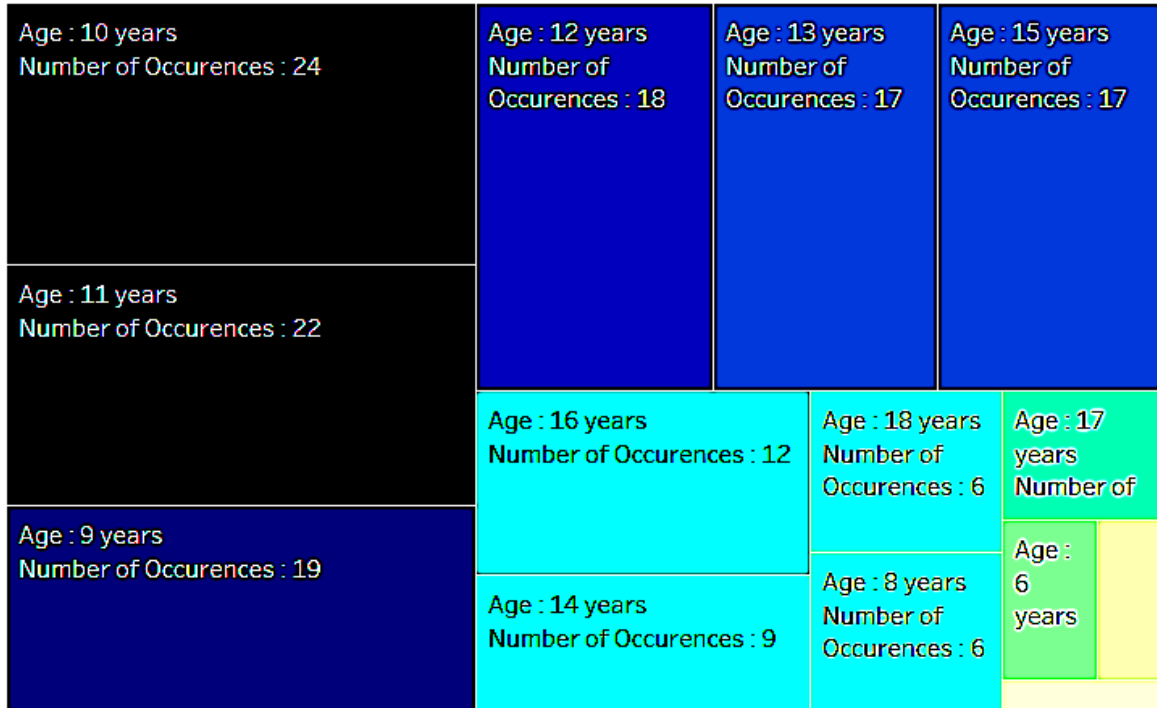


Fig. 6.8 Chart demonstrating the occurrence of issue in vessels of different ages

4. Vessels built by KZ manufacturer are more prone to issues when compared to other manufactures' vessels.

Suspected Reason: That manufacturer is not providing quality product.

Possible Solution: Internal enquiry; Change of manufacturer.

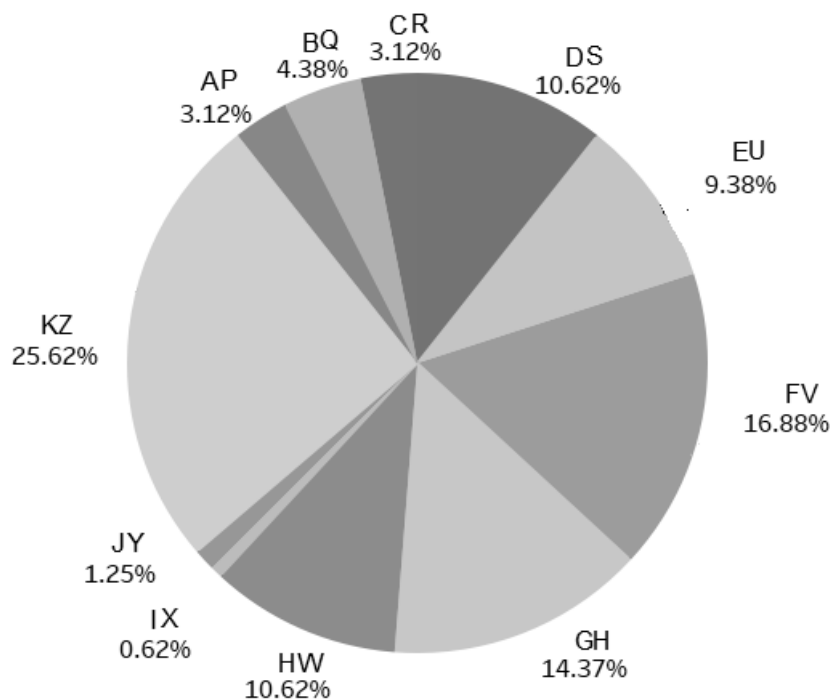


Fig. 6.9 Chart demonstrating the occuracne of issue in vessels built by different manufactures

6. More issues occur when there are moderate climatic conditions.

Suspected Reason: Carelessness and leniency of crew members.

Possible Solution: Crew needs to be more altern and serious.

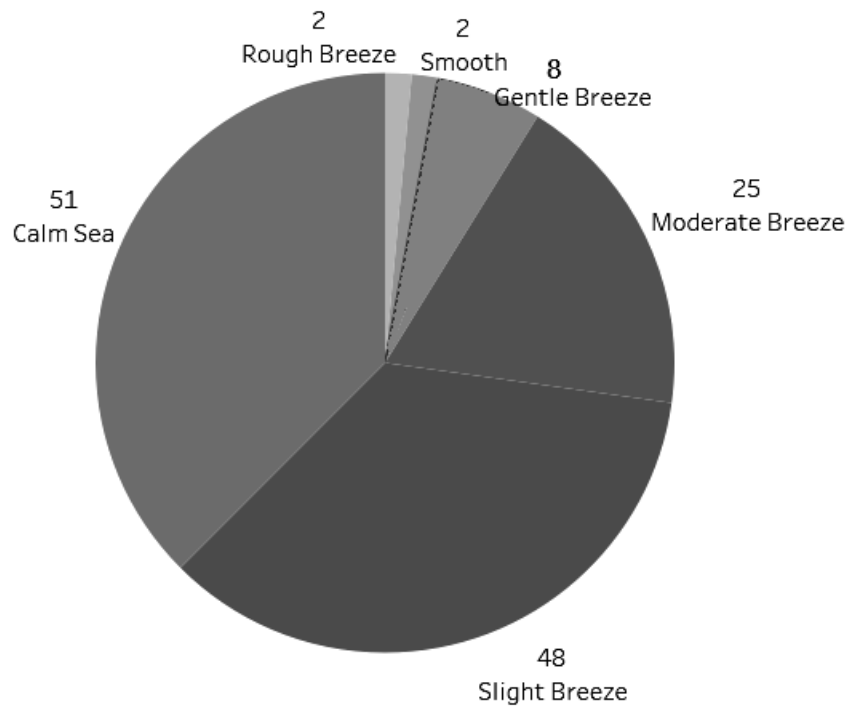


Fig. 6.10 Chart demonstrating the occurance of issues in different wind speeds

6. Old ships have the lowest risk levels and new ships have higher risk levels.

Suspected Reason: Unclear

Possible Solution: unclear

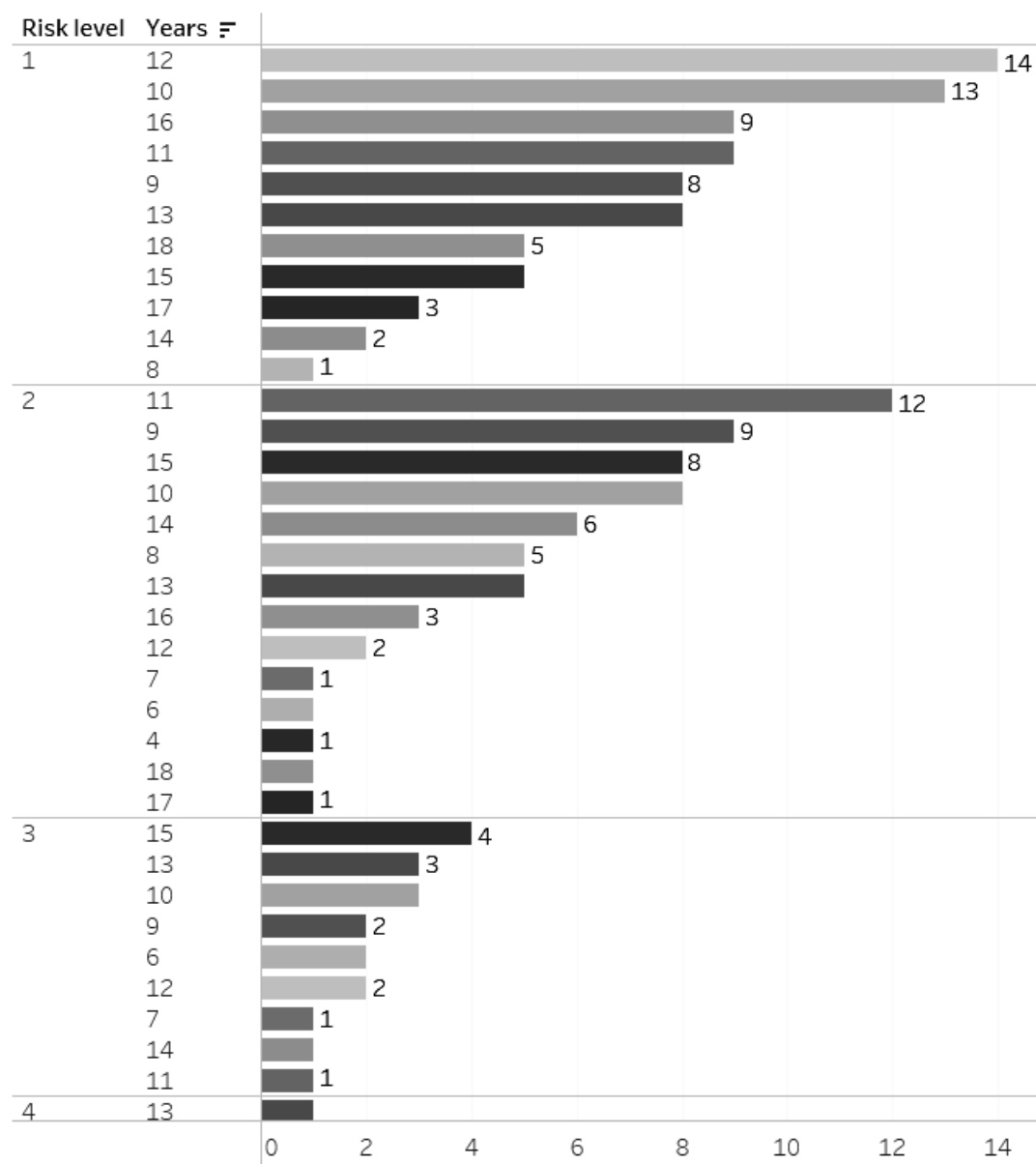


Fig. 6.11 Chart demonstrating the performance of different vessels based on age

6. Most problems occur with a Beaufort scale value of 2 to 3 (moderate to mild climatic conditions) and with a lower risk level.

Suspected reason: Attentiveness of crew members.

Possible Solution: Not required

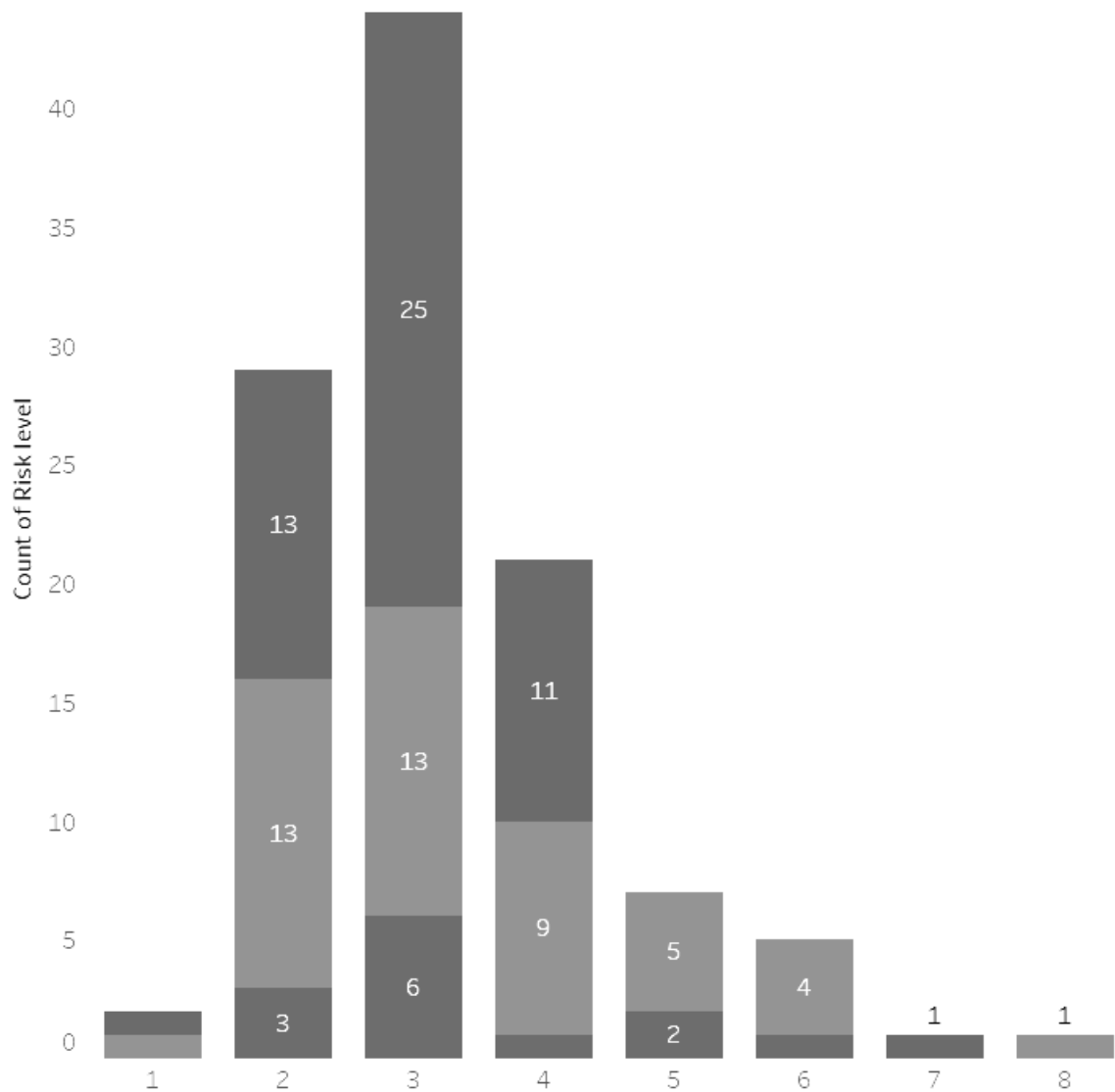


Fig. 6.12 Chart demonstrating occurrence of issues with respect to Beaufort Scale

7. More issues occur in the winter and summer seasons.

Suspected Reason: Harsh climatic condition

Possible Solution: Unclear

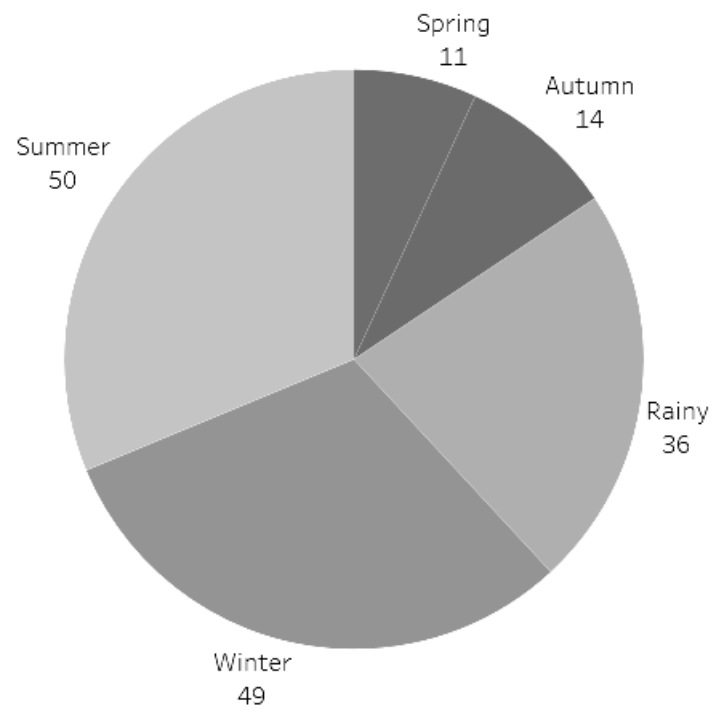


Fig. 6.13 Chart demonstrating occurrence of issues in different seasons

Chapter 7

REFERENCES

1. Katal, A., Wazid, M., & Goudar, R. H. (2013). Big data: Issues, challenges, tools and good practices. In Proceedings of the 6th International Conference on Contemporary Computing (IC3) (pp. 404-409). IEEE.
2. Lin, Y., Chen, T., & Liu, C. (2018). Design and implementation of ETL process based on service-oriented architecture. *Journal of Network and Computer Applications*, 105, 116-124.
3. El-Sappagh, S., Elmogy, M., & Saddik, A. E. (2018). Semantic data integration in big data: A systematic literature review. *Journal of Big Data*, 5(1), 1-26.
4. Davenport, T. H., & Patil, D. J. (2012). Data scientist: The sexiest job of the 21st century. *Harvard Business Review*, 90(10), 70-76.
5. Li, W., Wang, K., & Tang, H. (2018). Research on OCR technology based on deep learning. *Journal of Physics: Conference Series*, 1079(3), 032059. doi: 10.1088/1742-6596/1079/3/032059
6. Dubey, A., & Singh, R. (2018). Fuzzy matching technique for record linkage. *International Journal of Computer Applications*, 179(43), 12-16. doi: 10.5120/ijca2018918025
7. Ravikumar, M., & Shanthi, M. (2016). An overview of OCR techniques in document image processing. *International Journal of Advanced Research in Computer and Communication Engineering*, 5(3), 435-439.