##### DATA COLLECTION

**A PROJECT REPORT**

###### ***Submitted by***

**Anjeleen Tirkey (20MIP10048)**

**Devarakonda Anuhya (20MIP10024)**

**Parthiv J P (20MIP10036)**

**Sidharth E S (20MIP10015)**

**Sundar N (20MIP10003)**

*in partial fulfillment for the award of the degree*

*of*

**M.Tech. (Integrated) - COMPUTER SCIENCE & ENGINEERING**

*in*

# SPECIALIZATION IN COMPUTATIONAL AND DATA SCIENCES

****

**SCHOOL OF COMPUTING SCIENCE AND ENGINEERING**

**VIT BHOPAL UNIVERSITY**

**KOTHRIKALAN, SEHORE**

**MADHYA PRADESH - 466114**

Dec 2021

**VIT BHOPAL UNIVERSITY, KOTHRIKALAN, SEHORE**

**MADHYA PRADESH – 466114**

**BONAFIDE CERTIFICATE**

Certified that this project report titled **“DATA COLLECTION”** is the bonafide work of **“Anjeleen Tirkey (20MIP10048), Devarakonda Anuhya (20MIP10024), Parthiv J P (20MIP10036), Sidharth E S (20MIP10015), Sundar N (20MIP10003)”** who carried out the project work under my supervision. Certified further that to the best of my knowledge the work reported at this time does not form part of any other project/research work based on which a degree or award was conferred on an earlier occasion on this or any other candidate.

**PROGRAM CHAIR PROJECT GUIDE**

Dr Pon Harshavardhan Dr Pon Harshavardhan

Associate Professor Associate Professor

School of Computer Science School of Computer Science

and Engineering and Engineering

VIT BHOPAL UNIVERSITY VIT BHOPAL UNIVERSITY

The Project Exhibition I Examination is held on December 26th, 2021.

**ACKNOWLEDGEMENT**

First and foremost I would like to thank the Lord Almighty for His presence and immense blessings throughout the project work.

I wish to express my heartfelt gratitude to Dr Pon Harshavardhanan, Division-Head of Data Science Division, School of Computer Science and Engineering for much of his valuable support encouragement in carrying out this work.

I would like to thank my internal guide Dr Pon Harshavardhan, for continually guiding and actively participating in my project, giving valuable suggestions to complete the project work.

I would like to thank all the technical and teaching staff of the School of Computer Science and Engineering, who extended directly or indirectly all support.

Last, but not least, I am deeply grateful to my parents and my batch mates who have been the greatest support while I worked day and night for the project to make it a success.

**LIST OF FIGURES AND GRAPHS**

| **FIGURE NO.** | **TITLE** | **PAGE NO.** |
| --- | --- | --- |
| **1** | AQI Description | 8 |
| **2** | Overall Architecture Diagram | 9 |
| **3** | Overall AQI System | 13 |
| **4** | Sample data collection - Bhopal | 15 |
| **5** | Sample data collection - Mandideep | 15 |

**ABSTRACT**

As the demand for Data Analysis, interpretation, prediction increases so does the need for clean and organized data. One of the key components of data analysis is reliable and accurate data. The collection of required data is vital for data analysis.

In this report, the data of the Air Quality Index(AQI) is taken to meet specifications.

We initially collect all the required data from a local server or an online website (any authentic source), then store it into a database.

This includes getting the data from the Air Quality Index monitoring stations and storing it so that it can further be processed and used for analysis.

To collect the data related to the Air Quality Index (AQI) from different air quality monitoring stations across Madhya Pradesh.

The air quality index (AQI) is an index for reporting air quality on a daily basis. It is a measure of how air pollution affects one's health within a short period. The purpose of the AQI is to help people know how the local air quality impacts their health. The Environmental Protection Agency (EPA) calculates the AQI for five major air pollutants, for which national air quality standards have been established to safeguard public health.

* Ground-level ozone
* Particle pollution/particulate matter (PM2.5/pm10)
* Carbon Monoxide
* Sulfur dioxide
* Nitrogen dioxide

The higher the AQI value, the greater the level of air pollution and the greater the health concerns. The concept of AQI has been widely used in many developed countries over the last three decades. AQI quickly disseminates air quality information in real-time.

**TABLE OF CONTENTS**

| **CHAPTER NO.** | **TITLE** | **PAGE NO.** |
| --- | --- | --- |
|  | List of Figures and Graphs  Abstract | iv  v |
| 1 | **CHAPTER-1:**  **PROJECT DESCRIPTION AND OUTLINE** Introduction 1.2 Motivation for the work  1.3 Problem Statement  1.4 Objective of the work  1.5 Organization of the project  1.6 Summary | 1  .  .  . |
| 2 | **CHAPTER-2:**  **RELATED WORK INVESTIGATION**  2.1 Introduction  2.2 Approaches/Methods  2.2.1 Approaches/Methods -1  2.2.2 Approaches/Methods -2  2.2.3 Approaches/Methods -3  2.3 Pros and cons of the stated Approaches/Methods  2.4 Observations from investigation  2.5 Summary |  |
| 3 | **CHAPTER-3:**  **REQUIREMENT ARTIFACTS AND METHODOLOGY**  3.1 Introduction  3.2 Software requirements  3.3 Methodology  3.3.1 Operationalization  3.3.2 Sampling  3.3.3 Standardizing Procedures  3.4 Summary |  |
| 4 | **CHAPTER-4:**  **IMPLEMENTATION AND PROJECT OUTCOME**  4.1 Outline  4.2 Data Exploration  4.2.1 Understand the variables  4.2.2 Detect any Outliers  4.2.3 Examine Patterns and Relationships  4.3 Summary |  |
| 5 | **CHAPTER-5:**  **CONCLUSION**  5.1Outline  5.2 Enhancements  5.3 Conclusion |  |
|  | References |  |

**Chapter - 1**

**PROJECT DESCRIPTION AND OUTLINE**

**1.1 Introduction -**

The goal of this project is to collect the data related to the Air Quality Index (AQI) from different air quality monitoring stations across Madhya Pradesh.

Air quality index (AQI) is used by government agencies to communicate to the public how polluted the air currently is or how polluted it is forecast to become

AQI calculations focus on major air pollutants, including particulate matter, ground-level ozone, sulfur dioxide (SO2), nitrogen dioxide (NO2), and carbon monoxide (CO). Particulate matter and ozone pollutants pose the highest risks to human health and the environment.

Data is said to be unclean if it is missing attribute, attribute values, contain outliers and duplicate or wrong data. The presence of any of these will degrade the quality of the results.

Data cleaning is crucial in any data analysis process as it directly impacts the success rate of the project. This reduces the complexity of the data under analysis as data in the real world is unclean.

Thus, before using that data for the purpose we want, we need it to be as organized and clean as possible.

Data collection is the procedure of collecting, measuring and analyzing accurate insights for research using standard validated techniques. A researcher can evaluate their hypothesis based on collected data. In most cases, data collection is the primary and most important step for research, irrespective of the field of research. The approach of data collection is different for different fields of study, depending on the required information.

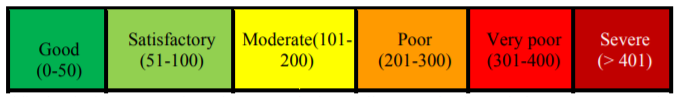
The most critical objective of data collection is ensuring that information-rich and reliable data is collected for statistical analysis so that data-driven decisions can be made for research.

**1.2 Motivation for the work -**

Our motivation to work was mainly due to the increasing rise in pollution in our areas. Air pollution levels in most of the urban areas have been a matter of serious concern. It is

the right of the people to know the quality of the air they breathe. However, the data generated through National Ambient Air Monitoring Network are reported in a form that may not be easily understood by a common person, and therefore, the present system of air quality information does not facilitate people's participation in air quality improvement efforts.

We mainly focused on the different levels of pollution in various areas such as residential areas or industrial areas. or, and Severe. The AQI considers eight pollutants Pradesh short-term (up to 24-hourly averaging period) standards are prescribed, however, AQI can be calculated if monitoring data are available for a minimum of three pollutants of which one should necessarily be PM 2.5 or PM 10. Based on the measured ambient concentrations, corresponding standards and likely health impact, a sub-index is calculated for each of these pollutants. The worst sub-index reflects overall AQI.



**AQI Description**

It is hoped that the AQI will provide more meaningful air quality information to the people, ensure everyone's participation to strengthen efforts being made for improving air quality in urban areas.

**1.3 Problem Statement -**

Collecting scattered data from an authentic source(s) and arranging it into meaningful datasets, which can be used for exploratory data analysis.

With technological advancements, a vast amount of data on ambient air quality is generated and used to establish the quality of air in different areas. One way to describe air quality is to report the concentrations of all pollutants with acceptable levels (standards). As the number of sampling stations and pollution parameters (and their sampling frequencies) increase, such descriptions of air quality tend to become confusing even for the scientific and technical community.

As for the general public, they usually will not be satisfied with raw data, time series plots, statistical analyses, and other complex findings of air quality. The result is that people tend to lose interest and can neither appreciate the state of air quality nor the pollution mitigation efforts by regulatory agencies. Since awareness of daily levels of urban air pollution is important to those who suffer from illnesses caused by exposure to air pollution, the issue of air quality communication should be addressed effectively. Further, the success of a nation to improve air quality depends on the support of its citizens who are well-informed about local and national air pollution problems and the progress of mitigation efforts.

**1.4 Objective of the work -**

The project aims to achieve the following:

(i) To collect the data related to the Air Quality Index (AQI) from different air quality monitoring stations across Madhya Pradesh.

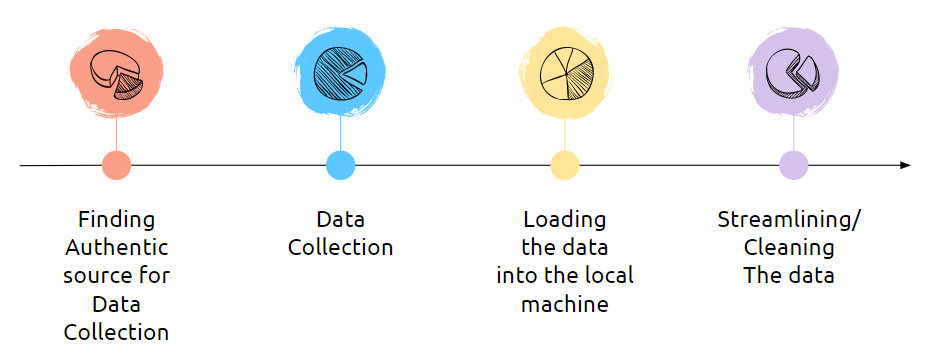
(ii) Further to understand the health impacts of air pollution exposure clean the data; this would help for analysis.

The overall objective of the project can be stated as under -

To adopt an Air Quality Index (AQI) based on national air quality standards, health impacts and monitoring programme which represent perceivable air quality for the general public in easy to understand terms and assist in data interpretation and decision-making processes related to pollution mitigation measures.

**1.5 Organization of the project -**

The following flowchart depicts the Organization of the project



**Overall Architecture Diagram**

The organization of the project is following the traditional data analysis procedure which is discussed earlier in the methodology section. This Organization of the project will be explained in detail in the future chapters.

**1.6 Summary -**

In most cases, data collection is the primary and most important step for research, irrespective of the field of research. The most critical objective of data collection is ensuring that information-rich and reliable data is collected for statistical analysis so that data-driven decisions can be made for research. The fact of the matter is that every organization has problems and inefficiencies. Due to the ever-changing nature of the business climate and society as a whole, it is nearly impossible to perfect how an organization is run. That being said, access to good data will ensure that you’re able to identify significant problems early on and take action to solve them.

It is proposed that for continuous air quality stations, AQI is reported in near real-time for as many parameters as possible. For manual stations, the daily AQI is reported with a lag of one week to ensure manual data are scrutinized and available for AQI. A web-based AQI dissemination system is developed for quick, simple and elegant looking responses to an AQI query. The other features of the website include reporting of pollutants responsible for index, pollutants exceeding the standards and health effects.

Combining all the knowledge gained we look forward to seeing a huge data collected and cleaned for the analysis.

**Chapter - 2**

**RELATED WORK INVESTIGATION**

**2.1 Introduction -**

Data Collection is the process of collecting, measuring and analyzing different types of information using a set of standard validated techniques. The main objective of data collection is to gather information-rich and reliable data and analyze them to make critical business decisions. Once the data is collected, it goes through a rigorous process of [data cleaning](https://www.simplilearn.com/data-cleaning-why-and-how-to-get-started-article) and [data processing](https://www.simplilearn.com/what-is-data-processing-article) to make this data truly useful for businesses. The data collection provides the basis for reliability estimations. Thus, a good data collection procedure is crucial to ensure that the reliability estimate is trustworthy. A prediction is never better than the data on which it is based. Thus, it is important to ensure the quality of the data collection.

Quality of data collection involves:

* Collection Consistency
* Completeness
* Measurement System Consistency

**2.2 Approaches/Methods -**

**2.2.1 Approach/Method 1 -**

**Observations -** Observation allows researchers to experience a specific aspect of social life and get a firsthand look at a trend, institution, or behaviour. Participant observation involves the researcher joining a sample of individuals without interfering with that group’s normal activities to document their routine behaviour or observe them in a natural context. Observational research is a type of descriptive research that differs from most other forms of data gathering in that the researcher’s goal is not to manipulate the variables being observed.

**2.2.2 Approach/Method 2 -**

**Surveys -** The survey method of data collection is a type of descriptive research, and is likely the most common of the major methods. The benefits of this method include its low cost and its large sample size. Surveys are an efficient way of collecting information from a large sample and are easy to administer compared with an experiment. Surveys are also an excellent way to measure a wide variety of unobservable data, such as stated preferences, traits, beliefs, behaviours, and factual information. It is also relatively simple to use statistical techniques to determine validity, reliability, and statistical significance.

**2.2.3 Approach/Method 3 -**

**Libraries -** Most researchers donate several copies of their academic research to libraries. We can collect important and authentic information based on different research contexts. Libraries also serve as a storehouse for business directories, annual reports and other similar documents that help businesses in their research.

**2.3 Pros and Cons of stated Approaches/Methods -**

**Pros -**

* Ease of data collection – the data already exists and no additional effort is needed.
* No need for searching and motivating respondents to participate.
* Allows us to track progress. Helps us understand the history behind an event and track changes over some time.
* Easy to analyze and present with different [data visualization types](http://intellspot.com/data-visualization-types/).

**Cons -**

* Information may be out of date or inapplicable.
* The process of evaluating documents and records can be time-consuming.
* Can be an incomplete data collection method because the researcher has less control over the results.
* Expensive method. It requires a high cost, effort, and plenty of time.
* Can take a lot of time if the observer has to wait for a particular event to happen.

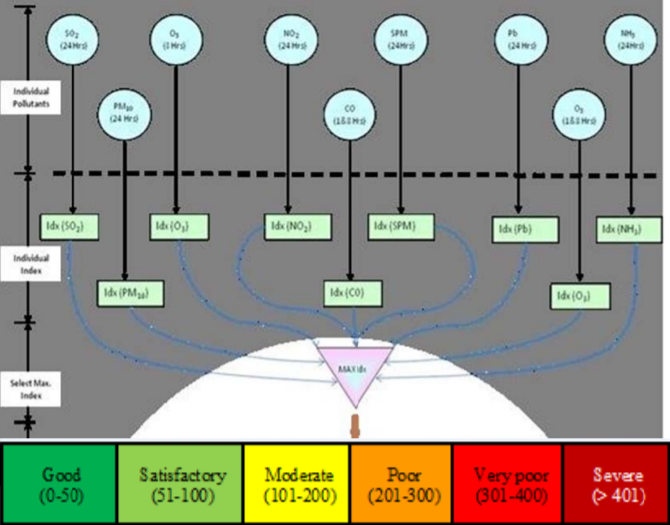
**2.4 Observation from the Investigation**

Practically, there isn’t one single best data collection method or technique. Each method has its advantages and disadvantages. Our choice depends on the type of insights we need and which pros/cons are important for our research. Determining the goals of our data collection is the first step in successful research. Consider how we can align the research to the specific user needs and requirements. Define which collection method would fit best in with the lifestyle or working style of the respondents.

**2.5 Summary -**

Data collection methods allow us to build strategies based on insights instead of opinions. Generally, we collect quantitative data through sample surveys, experiments and observational studies. We obtain qualitative data through focus groups, in-depth interviews and case studies. Discuss each of these data collection methods and examine their advantages and disadvantages. A researcher can evaluate their hypothesis based on collected data. In most cases, data collection is the primary and most important step for research, irrespective of the field of research. The approach of data collection is different for different fields of study, depending on the required information. The most critical objective of data collection is ensuring that information-rich and reliable data is collected for [statistical analysis](https://www.questionpro.com/market-research.html#Statistical_Analysis_Techniques_for_Market_Analysis) so that data-driven decisions can be made for research.

Overall AQI system -



**Overall AQI System**

**Chapter - 3**

**REQUIREMENT ARTIFACTS AND METHODOLOGY**

**3.1 Introduction -**

Big data analytics helps users collect and analyze large-sized data sets that have a varied mix of content. This analysis delivers insights into the content through the exploration of data patterns. These insights are used by business owners to make informed decisions that are driven by data. Big data is defined as data sets that are larger in volume than basic databases and their handling architecture. Simply put, big data is information that is beyond the scale handled by a spreadsheet like Microsoft Excel. Big data includes the process of storage, processing, and visualization of information. To draw insights, businesses need to carefully select big data tools and create a suitable environment around the information.

**3.2 Software Requirements -**

* Firefox/Chrome
* Adobe Flash Player
* Adobe Reader
* DB Browser for SQLite

**3.3 Methodology -**

Based on the data we want to collect, decide which method is best suited for our research. [Experimental](https://www.scribbr.com/methodology/experimental-design/) research is primarily a quantitative method. Interviews/focus groups and [ethnography](https://www.scribbr.com/methodology/ethnography/) are qualitative methods. [Surveys](https://www.scribbr.com/methodology/survey-research/), observations, archival research and secondary data collection can be quantitative or qualitative methods. Carefully consider what method we will use to gather data that helps us directly answer our research questions. When we know which method(s) we are using, we need to plan exactly how we will implement them.

### 3.3.1 Operationalization -

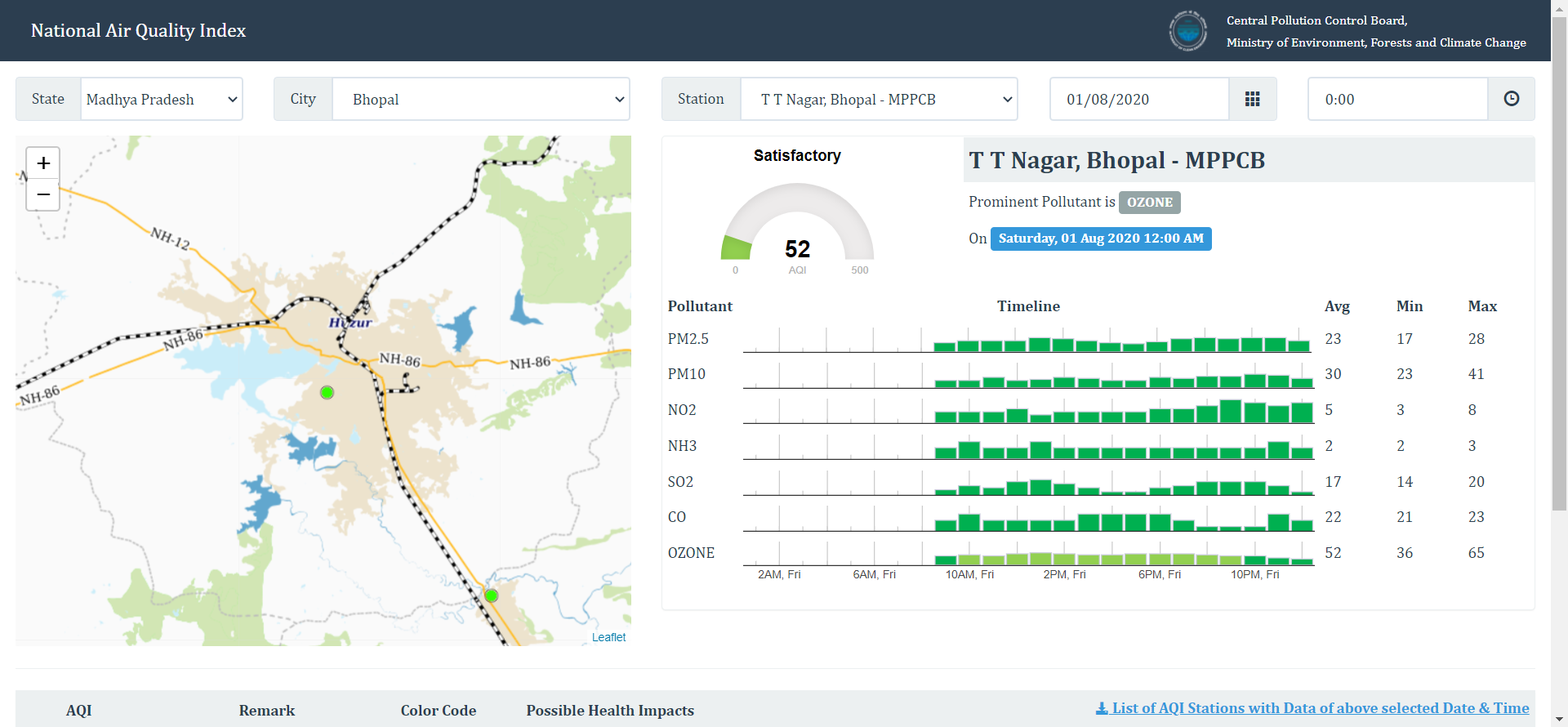
Sometimes variables can be measured directly. However, often we’ll be interested in collecting data on more abstract concepts or variables that can’t be directly observed.

[Operationalization](https://www.scribbr.com/dissertation/operationalization/) means turning abstract conceptual ideas into measurable observations. When planning how we will collect data, we need to translate the conceptual definition of what we want to study into the operational definition of what we will measure.

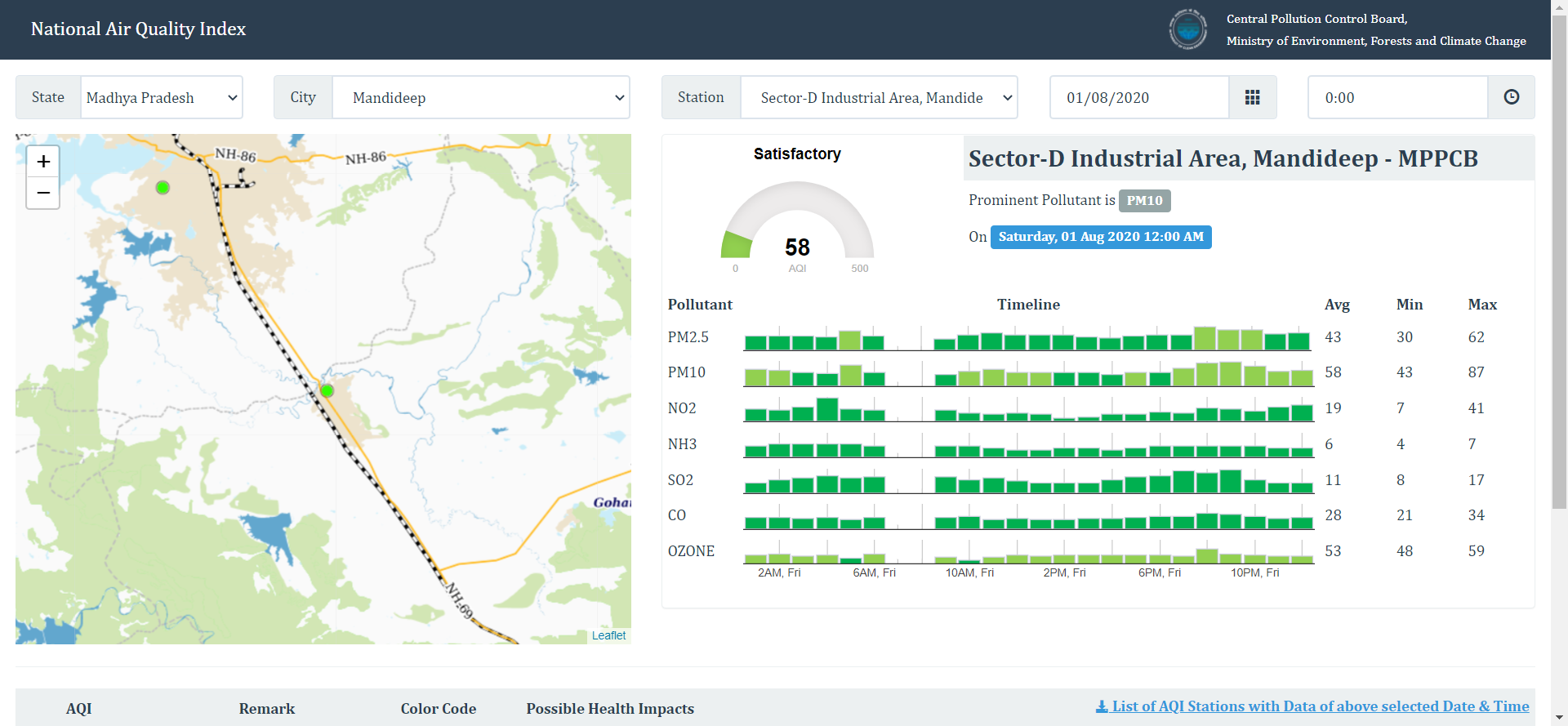
### 3.3.2 Sampling -

We may need to develop a [sampling](https://www.scribbr.com/methodology/sampling-methods/) plan to obtain data systematically. This involves defining a [population](https://www.scribbr.com/methodology/population-vs-sample/), the group we want to conclude about, and a sample, the group we will collect data from. The sampling method will determine how we recruit participants or obtain measurements for our study. To decide on a sampling method we will need to consider factors like the required sample size, accessibility of the sample, and timeframe of the data collection.

Here is a sample of our data collection website from the 2 different stations of Madhya Pradesh; **Bhopal -**



**Mandideep -**



### 3.3.3 Standardizing procedures -

If multiple researchers are involved, write a detailed manual to standardize data collection procedures in our study. This means laying out specific step-by-step instructions so that everyone in the research team consistently collects data – for example, by conducting experiments under the same conditions and using objective criteria to record and categorize observations. This helps ensure the [reliability](https://www.scribbr.com/methodology/types-of-reliability/) of our data, and we can also use it to replicate the study in the future.

**3.4 Summary -**

The challenge of collecting software engineering data is to make sure that the collected data can provide useful information for the project, process, and quality management and, at the same time, that the data collection process will not be a burden on development teams. Therefore, it is important to consider carefully what data to collect. The data must be based on well-defined metrics and models, which are used to drive improvements. Therefore, the goals of the data collection should be established and the questions of interest should be defined before any data is collected. Data classification schemes to be used and the level of precision must be carefully specified. The collection form or template and data fields should be pretested. The amount of data to be collected and the number of metrics to be used need not be overwhelming. It is more important than the information extracted from the data be focused, accurate, and useful than that it be plentiful. Without being metrics-driven, the over-collection of data could be wasteful. Overcollection of data is quite common when people start to measure software without an a priori specification of purpose, objectives, profound versus trivial issues, and metrics and models.

**Chapter - 4**

**IMPLEMENTATION AND PROJECT OUTCOME**

**4.1 Outline -**

An air quality index is a scale used to show how polluted the air is, along with the risks associated with each rating. An AQI is calculated using established standards based on medical research for the acceptable levels of major air pollutants.

Air quality indexes serve two main purposes:

1. To inform the public about air quality in a comprehensible manner so that they may take action to protect their health
2. To help countries develop and assess policies for better air quality

There are many different air quality indexes used by various government agencies out there. However, none of the AQIs reviewed by us included all relevant air quality parameters for all relevant timeframes.

**4.2 Data Exploration -**

**4.2.1 Understand the variables -** The basis for any data analysis begins with an understanding of variables. A quick read of column names is a good place to start. A closer look at data catalogues, field descriptions, and metadata can offer insight into what each field represents and help discover missing or incomplete data.

**4.2.2 Detect any Outliers -** Outliers or anomalies can derail an analysis and distort the reality of a dataset, so it’s important to identify them early on. Data visualization, numerical methods, interquartile ranges, and hypothesis testing are the most common ways of detecting outliers. A boxplot, histogram, or scatterplot, for example, makes it easy to spot points far outside the standard range, while a z-score informs how far from the mean a data point is. Once found, an analyst can investigate, adjust, omit, or ignore the outliers. No matter the choice, the decision should be noted in the analysis.

**4.2.3 Examine Patterns and Relationships -** Plotting a dataset in a variety of ways makes it easier to identify and examine the patterns and relationships among variables. A temperature exploring data from multiple stations may have information on location, parameters, and cities. To estimate the air quality index for a location, they need to decide which variables to include in their predictive model.

**4.3 Summary -**

Exploration allows for a deeper understanding of a dataset, making it easier to navigate and use the data later. The better an analyst knows the data they’re working with, the better their analysis will be. Successful exploration begins with an open mind, reveals new paths for discovery, and helps to identify and refine future analytics questions and problems. Data without a question is simply information. Asking a question of data turns it into an answer. Data with the right questions and exploration can provide a deeper understanding of how things work and even enable predictive abilities. The exploration process is also increasingly important to work with Geographic Information Systems since so much of today’s data is location-enriched.

**Chapter - 5**

**CONCLUSION**

**5.1 Outline -**

Data standards outline how common data items and demographic information should be collected. Established standards typically contain data definitions, standardized questions and accepted response options which guide consistent collection practices. Currently, there are many national and state-wide data standards that are used for collecting administrative data. These standards are not always broadly applied, and may themselves be inconsistent, and this can impact the comparability of data collections. Organizations may face several challenges in collecting consistent and quality data. To develop methods to improve data collection practices, it is necessary to first identify barriers to consistent data collection. Organizations changing data collection systems and processes also need to be aware of the need to ensure continuity of reporting using existing data items. For example, many service providers are bound by the requirements of their funding body to provide particular data fields regularly.

**5.2 Enhancement -**

Improving data collection and the quality of data holdings requires a concerted effort from an entire organization, and should begin with a top-down commitment for change. This includes identifying priority areas for improvement and barriers to improvement, adopting best practice procedures for collecting quality data, using data standards where available, ensuring IT infrastructure is kept up to date and allows for efficient and effective data collection, and providing training where needed to those collecting data to ensure confidence and consistency in data collection practices.

It is important to provide training to staff involved in the collection of data. Training should emphasize why it is important to collect data and highlight the benefits of data for operations, planning, research and evaluation. If staff understand the rationale for collecting certain information, they will feel more confident to ask for these data items and to explain why it is important. Training should include how to phrase questions, clarify answers and record responses.

**5.3 Conclusion**

The data collection process is not an end in itself: the culminating activities of evaluation are analysis, interpretation, and the presentation of findings. Data collection, analysis and interpretation is an iterative process, moving back and forth within these three stages. There’s a bit more to collecting data, however. Recording and organizing data may take different forms, depending on the kind of information we’re collecting. The way we collect our data should relate to how we’re planning to analyze and use it. Regardless of what method we decide to use, recording should be done concurrently with data collection if possible, or soon afterwards, so that nothing gets lost and memory doesn’t fade.

Once we’ve organized our data and run them through whatever statistical or other analysis we’ve planned for, it’s time to figure out what they mean for our evaluation. Probably the most common question that evaluation research is directed toward is whether the program being evaluated works or makes a difference. If our analysis gives us a clear indication that what we’re doing is accomplishing our purposes, interpretation is relatively simple: We should keep doing it while trying out ways to make it even more effective, or while aiming at other related issues as well. Whether quantitative and/or qualitative methods of gathering data are used, the analysis can be complex, or less so, depending on the methods used and the amount of data collected.