FROM DATA TO DECISIONS: ANALYZING AND VISUALIZING OPEN DATASETS USING PYTHON AND POWER BI

Submitted by

HANIFA IBRAHIM
AMNA BIBI

Supervised by

DR. AZHAR RAUF

Bs Computer Science

Session 2021–2025

Department of Computer Science
University of Peshawar

Project Approval

This is to certify that this project is approved and recommended as a partial fulfillment for the degree B.S. Computer Science from the Department of Computer Science, University of Peshawar.

External Examiner: _____

Internal Examiner: _____

Chairman: _____

ACKNOWLEDGMENT

First and foremost, we are profoundly grateful to Almighty Allah for granting us the strength, patience, and determination to complete this project successfully.

We would like to extend our deepest gratitude to our supervisor, **Dr.Azhar Rauf**, for their invaluable guidance, encouragement, and continuous support throughout the course of this work. Their expertise and insightful feedback greatly contributed to the development and completion of this project.

We are also thankful to **Dr.Rehman Ali** and, the faculty and staff of the **Department of Computer Science, University of Peshawar**, for providing a stimulating academic environment and for their constant encouragement.

Our sincere appreciation goes to our friends and colleagues, whose cooperation, constructive discussions, and moral support have been instrumental during this journey.

Lastly, we are deeply indebted to our parents and family, whose unconditional love, prayers, and sacrifices have always been our source of motivation. Without their unwavering support, this achievement would not have been possible.

Hanifa Ibrahim

Amna Bibi

Contents

List of Figures

ABSTRACT

Data plays a pivotal role in the functioning of large businesses, with executive management mostly relying on archival data to make informed decisions. However, unprocessed data is mostly unsuitable for direct use, It needs ample preprocessing to obtain meaningful Key Performance Indicators(KPIs) and practical takeaways. This project, titled ”From Data to Decisions: Analyzing and Visualizing Open Datasets Using Python and Power BI,” focuses on showcasing how open datasets can be modified into decision support systems through blending datascience and exploratory data analysis. Python is used for cleaning, preprocessing and transformation while Power BI is used to create interactive dashboards that functionally communicate insights. This project highlights the importance of data-informed decision making and the potential to combine programming with business intelligence tools to support organizational strategies going forward.

Chapter 1

INTRODUCTION

With every moment and each click we make, along with the activation of various devices, an enormous amount of data is produced. Whenever someone visits a website associated with an organization it creates a lot of data that is beneficial for that organization. Although some data is beneficial, much of the data we have created is disorganized and makes it difficult to extract insights from the information.

Data was originally recorded only on paper so we were limited to extracting insights and seeing trends. But some people who were interested in what data previously collected could provide began digging through data to analyze it and see what to predict for better informed decision making. Even though these analyses were done manually and were very difficult to do, they provided the initial conceptual framework for data science and data analytics. This was just the beginning of our effort to find better ways to to process, analyze and get insights from data.

Data analytics is the process of examining datasets with the aim of uncovering insights and making more effective and confident conclusions based on the information they have. This analysis is performed using specific tools and software. To illustrate how open datasets can be adapted and insights can be derived, we present our project titled **FROM DATA TO DECISIONS: ANALYZING AND VISUALIZING OPEN DATASETS USING PYTHON AND POWER BI.** This project integrates programming with business intelligence tools, enabling businesses and organizations to identify new trends.

1.1 Problem Statement

”Raw, open datasets are not directly usable for decision-making unless they undergo systematic preprocessing, transformation, and visualization.”

Today, organizations are increasingly relying upon archival and real-time data to make strategic decisions in a data-driven world. However, most raw datasets and particularly open datasets are unstructured, and inconsistent and cannot be used for decision making without proper preprocessing. If not properly preprocessed, the datasets will fail to produce Key Performance Indicators (KPIs) and usable information and insights. The limited ability of organizations to take advantage of open data as a decision-support resource is concerning. Therefore, there is a need for a systematic method that incorporates data science methodologies and visualization tools to support decision-making processes by creating trustworthy and interactive decision-support systems from raw data.

1.2 Research Objective

The research objective for our thesis is as following:

1. To recognize problems surrounding raw open datasets as a basis for decision-making in a business context.
2. To implement techniques using Python for cleaning, preprocessing, and transforming open datasets into structured, analyzable datasets.
3. To create interactive dashboards in Power BI that convey insights and KPIs from the processed datasets.
4. To combine data science methods and business intelligence tools to build a decision support system.
5. To assess the possibility of the respective approach to make data informed and strategic decisions.

1.3 Scope of the Study

The scope of this thesis is designed to further explore the transformation of open datasets into decision-making systems, through the application of data science methods, and through the use of business intelligence applications. To fulfill this scope, the study is limited to the following aspects:

- **Data Source:** The study utilizes publicly available open datasets that we have selected to ensure the data will be open, accessible, transparent, and reproducible processes. The study will not use proprietary or sensitive organizational data.
- **Data Preparation and transformation:** The project uses Python for data cleaning, preparation, transformation, and exploratory data analysis (EDA) with the emphasis on substituting unstructured and inconsistent forms of data with meaningful examples.
- **Visualization and dashboard creation:** The project uses Power BI to create and share interactive dashboards that convey Key Performance Indicators (KPIs) and insights in functional ways.
- **Decision support focus:** The study focuses on using processed data to allow users to make informed, evidence-based decisions by examining trends, patterns, and anomalies that will inform decisions that connect with the educational organization's strategy.
- **Illustrative Value:** The thesis demonstrates a generalized form of converting raw data to actionable insights. The work is best described as illustrative as it does not propose to be exhaustive and does not claim to represent every possible application of an organization's business and every dataset.

1.4 Tools Used

The project has required several different pieces of software to assist in the data preprocessing, analysis, and visualization aspects. The following tools were utilized in the project:

JupyterLab: JupyterLab was used as an interactive development environment in which to program in Python. It has the capability to perform data preprocessing, cleaning, transformation, and exploratory data analysis (EDA), step by step. What makes JupyterLab of particular benefit is the interaction of code, output, and documentation all from the same document (in this case, a Jupyter Notebook). The functionality makes it ideal for iterative and experimental approaches while creating reproducible analyses.

Python: Python was used as the programming language that implemented the techniques from data science. The major tasks performed were: Data cleaning and preprocessing Converting raw datasets from file formats into structured datasets Performing exploratory data analysis (EDA) Preparing datasets for visualization in Power BI

Power BI: Power BI was used as the business intelligence tool for visualization and dashboard building. Power BI allowed the creation of interactive dashboards that displayed Key Performance Indicators (KPIs), trends, and insights to provide meaningful information to the users. By connecting the datasets processed in Python with Power BI, it was possible to develop a dynamic decision-support system.

Chapter 2

LITERATURE REVIEW

2.1 Introduction

Data driven decision making is a key function in the current enterprise, where the ability to extract information from quantitative data usually informs organizations' strategic planning for efficiency in operations. Our current study exists in the context of data science, exploratory data analysis, and data visualization, all of which provide the theoretical and empirical basis for the study. This chapter reviews the relevant literature relating to data science, the exploration of data using Python, descriptive statistics, and visualization using Power BI.

2.2 Data Science

Data science derives its meaning from the interaction of the fields of statistics, programming, and knowledge of the area of study [1]. As organizations have identified their desire to rely on structured and unstructured data in making decisions, the field of data science has become increasingly important. Apparently, it was Provost and Fawcett who identified the field of data science to be the field that enables the practical transition from raw data to actionable knowledge [2]. The use of machine learning and the interoperability of business intelligence have clearly advanced the ability of organizations to make better predictions, operations, and performance decisions in direct synergy with the machine learning capabilities of data science [3].

2.3 Exploration Pipeline of Data with Python

Exploration of data commonly refers to cleaning, preprocessing, and analyzing datasets to identify patterns and then preparing those datasets for modeling [4]. Python has gained wide popularity as a programming language for data exploration because of its ease of use and because there is a robust ecosystem surrounding data science. For example, there are libraries developed for Python called ”Pandas,” ”NumPy,” and ”Matplotlib,” which streamline data manipulation, statistical analysis, and data visualization [5]. McKinney highlights that exploratory data analysis (EDA) is based on the acceptance of the importance of using Python as the statistical backbone of decision making [6]. Through systematic execution of data exploration pipelines, organizations can effectively address missing values, identify anomalies, and develop descriptive summaries to aid in downstream analytics.

2.4 Descriptive Statistics in Organization Data

Descriptive statistics provide summary information and describe the important features of a dataset, giving users a first understanding of data before they conduct further exploration [7]. Average, median, standard deviation, and correlation are important descriptive statistics used to understand organizational data and derive KPIs. Furthermore, Tukey emphasized exploratory data analysis inclusive of descriptive statistics is a form of generalization when it aids in hypotheses generation and decision support [8]. Descriptive statistics are common in business use cases to find trends, measure performance, and to help managers produce summaries of large datasets that are concise and interpretable [9].

2.5 Data Visualization through Power BI

Data visualization is the graphical representation of data that turns larger datasets into useful, meaningful information for action [10]. Within visualization programs, Microsoft Power BI is one of the leading platforms that allows the con-

struction of interactive dashboards for organizations to apply in decision making. It allows embedding from data sources and incorporate live to real time reports. Reports can also scale to develop customized KPIs [11]. In Gartner's ranking of BI tools, Power BI is consistently ranked one of the leading tools among several market variables such as access, usability, scale, and affordability [12]. Studies suggest, interactive dashboards promote managerial decision-making as well as improved communication across multiple levels in the organization [13].

2.6 Gap Analysis

The reviewed literature explains the significance of data science, exploratory data analysis, descriptive statistics, and visualization in converting raw data into useful insights. Python is an effective data preprocessing and exploring tool [5], [6], while Power BI is widely used as a platform for interactive dashboards and decision making purposes [11], [12]. However, the majority of studies develop and present either the technical aspect of data science methods or the graphical user interface of the business intelligence visualization. The research gap clearly exists in combining a Python-based data-exploration pipeline with Power BI dashboards, to create an integrated decision-support system from open datasets. Previous works have reference to the specific capabilities of either [2], [10], [13], yet little has demonstrated a combined approach to provide organizations an informative data preprocessing and interactive visualization environment conducive to decision making. This thesis develops the research gap with the framework "From Data to Decisions: Exploring and Visualizing Open Datasets Using Python and Power BI", where the data exploration is completed with Python then visualized through the Power BI dashboards. The advantage of creating an integrated experience means that decision makers have access to verified, cleaned data and an intuitive and interactive visual interface, thus creating a connected way forward from raw data to actionable organizational plans and strategies.

Chapter 3

METHODOLOGY OF RESEARCH

3.1 Introduction

This chapter gives the research methodology taken in the project—From Data to Decisions: Analyzing & Visualizing Open Datasets using Python and Power BI. The methodology describes the systematic approach taken to data gathering, preprocessing, analyzing and finally, visualizing data to turn open datasets into decision-support systems.

3.2 Research Design

This research follows a design science methodology that is, creating and evaluating an artifact to help solve a practical problem. This research created an artifact as a decision support framework including Python based preprocessing feeding Power BI dashboards. The broad research was organized into five phases:

3.2.1 Data Collection

To complete the project, open datasets were acquired from publicly accessible repositories. Open data was selected for the following reasons: it is open for access, it adds transparency regarding dataset provenance, and it shows that the proposed framework can be used for datasets from different domains. Datasets are generally produced in a combination of structured and unstructured formats that require some preprocessing before analysis.

3.2.2 Data Preparation (Python)

Data preparation was performed with Python (via Anaconda Jupyter Lab) given its various libraries for data cleaning and transformation processes. The below steps were performed:

- Handling Missing Records: Identifying & Imputing missing records.
- Removing Duplicate Records: Verifying uniqueness and reliability of records.
- Data Transformation: Normalizing and re-organizing fields for consistency.
- Feature Extraction: Producing new attributes available for analysis.
- Libraries such as Pandas, NumPy, and Matplotlib were primarily utilized during this phase.

3.2.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to find trends, distributions, and correlations in the dataset. Descriptive statistics such as mean, median, variance, correlation coefficients were completed. Preliminary plotting was conducted using Python visualization libraries (Matplotlib, Seaborn), based on the preliminary plot, a dashboard in Power BI was designed.

3.2.4 Data Visualization (Power BI)

The cleaned and prepared data was imported into Microsoft Power BI to develop interactive dashboards. The dashboards included:

1. Key Performance indicators (KPIs) from the data.
2. Trends and patterns established during EDA.
3. Map and categorical visualizations.
4. Comparative patterns over differently defined time periods or categories.

5. Power BI was chosen for the user-friendliness of the dashboard, scalability, and the ability to connect with outputs from Python.

3.2.5 Assessment of the Framework

The application of the purposes mentioned above resulted in the evaluation of the framework against the following criteria:

- Accuracy of PreProcessing: marked by the understanding of a data's integrity and correctness after its transformation.
- Clarity of Visualization: assessed the dashboards for interpretability and usability.
- Decision Support: if the framework in and of itself could help organisations develop actionable insights that relate to strategic decision making.

3.3 Conclusion

This chapter concludes by outlining the methodology for the project, a hybrid methodology of Python preprocessing of publicly available data sets and creation of dashboards based off of this transformed data. The next chapter will dive deeper into the implementation of the methodology specified in Chapter-2.

Chapter 4

IMPLEMENTATION

4.1 Introduction

This chapter shows and explains how the framework **"From Data to Decisions: Analyzing and Visualizing Open Datasets using Python and Power BI"** was implemented. It provides a detailed summary of the practical aspects of the framework for processing and analyzing the dataset in Python with Anaconda JupyterLab, and building interactive dashboards in Power BI.

4.2 Environment of the Implementation

The following environment was used in the implementation:

Hardware: Intel i7 processor, 8GB RAM, Windows 10 operating system.

Software: Anaconda Distribution - for Python library management.

JupyterLab: for interactive execution of Python notebooks.

Python (v3.x): derivation, statistics and visualization.

Microsoft Power BI Desktop: for building dashboards.

This makes a stable platform for both data science workflows and business intelligence visualization

4.3 Dataset Description

The project utilized three open datasets from distinct public repositories. Each dataset represented a unique real world domain, allowing for evaluation of the proposed framework across diverse sources of data.

4.3.1 Dataset 1: Water Main Breaks

The first dataset, named *Water Main Breaks*, was acquired from the Kitchener GeoHub open data portal [14]. It contains historical data on water main breaks in the City of Kitchener.
The dataset consists of fields such as:

- Break ID: a unique identifier for each recorded incident.
- Break Date: the date when the water main break occurred.
- Location: geographic coordinates and street level information.
- Cause: the reported cause of the break (e.g., material failure, external damage, seasonal changes).
- Pipe Material: type of pipe (cast iron, PVC, etc.).

4.3.2 Dataset 2: Chicago Crimes

The second dataset, named *Chicago Crimes*, was acquired from Kaggle open data portal [15] It contains 22 years of data on crimes occurring in the city of Chicago.
The dataset consists of fields such as:

- ID: a unique identifier for each reported crime.
- Date: timestamp of when the crime occurred.
- Primary Type: classification of the crime (e.g., theft, assault, narcotics).
- Location: geographic coordinates and community area where the incident took place.

- Arrest/Domestic: indicators showing whether an arrest was made and whether the crime was domestic in nature.

4.3.3 Dataset 3: Student Depression Analysis

The third dataset, named *Student Depression Analysis* was acquired from Kaggle open data portal [16]
The dataset consists of fields such as:

- ID: a unique identifier for each student.
- Age: the age of the student.
- Gender: gender of the student (e.g., male, female, non-binary).
- CGPA: current cumulative grade point average.
- Sleep Duration: average hours of sleep per day.
- Academic Pressure: level of stress due to academic workload.
- Study Satisfaction: self-reported satisfaction with study habits and outcomes.
- Depression Status: indicator of whether the student is experiencing depression (Yes/No).

4.4 Data Preprocessing and Cleaning

Raw data will often contain inconsistencies, missing values, duplicates, and unstructured data characteristics that would need to be addressed before any serious analysis can be performed. Thus, preprocessing was performed using Python (specifically Anaconda Jupyter Lab) to guarantee the data was clean, reliable, and suitable for visualization in Power BI. The preprocessing workflow included the following:

- **Data Import:** Each dataset (Water Main Breaks, Chicago Crimes, and Student Depression) was imported into Python with the Pandas library.

- **Dealing With Missing Values:** Null or missing values were checked using df.isnull(), removing incomplete records if so or taking as imputation the mean or median value if the property in question allowed it.
- **Removing Duplicates:** Duplicate records were found and dropped using df.drop_duplicates() to ensure consistency in the data.
- **Date and Time Transformation:** Date fields were transformed to appropriate datetime formats to better illustrate temporal analysis of the datasets. For instance, through time fields, year, month, and day could be extracted from timestamp-related columns for trend visualization.
- **Categorical Encoding:** Non-numeric categorical types of values (e.g., gender, crime type, levels of academic pressure) were converted to numeric types to enable analysis and visualization.
- **Normalization and Transformation:** When necessary, the selected classes or attributes were adjusted or converted to enable precise comparisons of comparable attributes.
- **Clean Data Export:** After the data was cleaned and transformed, the processed datasets were exported to a CSV format that allowed them to be imported into Power BI for visualization.

The preprocessing pipeline ensured the data integrity and usability of the datasets which allowed for a robust descriptive statistical analysis in Python, and meaningful visualizations in Power BI.

4.4.1 Advanced Cleaning Techniques

The Chicago Crime dataset, which contains around 7.8 million rows of records, was processed through an initial data processing pipeline that was developed for memory management and processing at scale. First, we implemented a chunk-based pipeline based on the Pandas and Dask framework that read the very large source CSV file in 500,000-row chunks, downcasting data types (for example:

converting to Int16 and float32) for more optimal memory usage, and then serializing each chunk directly to the more memory efficient Parquet format using the PyArrow engine. After checking that the total row count matched for verification, we proceeded to the cleaning step, where we loaded the entire Parquet dataset retrieved from the preprocessing pipeline, removed zero duplicates, and completed a substantial amount of imputation, including using mode for categorical columns, and median for general numerical columns, especially focusing on missing geographic coordinates. Finally, we took advantage of the date column to create some important time-based features such as: Month, Hour, and DayOfWeek, and saved a final crime$_c leaned.parquet file that we could perform more th$
$series modeling using libraries like Prophet.$

4.5 Exploratory Data Analysis

Prior to developing the dashboard, exploratory data analysis (EDA) was conducted on the cleaned datasets in Python using Jupyter Lab. This was a critical part of understanding the data structure, recognizing patterns, and identifying outliers prior to visualizing the data. The following libraries were used for the analysis: Pandas, NumPy, Matplotlib, and Seaborn.

The EDA process included the following steps:

- **Descriptive Statistics:** Basic summary measures (mean, median, mode, variance, and standard deviation) were calculated to summarize the datasets. These measures were useful for getting a sense of central tendency and range.
- **Quantile Analysis:** Quartiles (25th, 50th, and 75th percentiles) were recognized as ways to understand the data and to determine outliers. Quartiles, for example, helped to identify whether crime counts or academic pressure levels were significantly above the average.
- **Skewness and Kurtosis:** Skewness was calculated to measure the symmetry of the distribution. Kurtosis indicated whether or not the data had

heavy tails and/or extreme outliers. These were especially useful for attributes of interest for CGPA, sleep hours, and frequency of crimes.

- **Frequency Distributions:** For categorical attributes (crime type, gender, severity of academic pressure), frequency counts were calculated to determine what categories dominated the data.
- **Histograms:** Histograms were prepared to show an overall distribution of data, whereas boxplots were produced to demonstrate variability in the data and highlight outliers.
- **Correlation Analysis:** Correlation matrices and heatmaps were developed to examine possible relationships between numerical variables. For example, for the student dataset we looked at the correlation between CGPA, amount of sleep, and depression.
- **Logistic Regression (Preliminary Modeling):** A simple logistic regression model was used on the student depression dataset to explore whether academic pressure, CGPA, and amount of sleep could predict depression status.
- **Trend Analysis:** For the temporal datasets of incidents such as crimes in Chicago, we grouped records by year and month for seasonal patterns, as well as long-term trends in the frequency of incidents.

The results of this analysis informed our decisions about the most relevant metrics and attributes for the dashboards. This helped us inform a meaningful selection of indicators and solid visualizations in Power BI.

4.6 Data Visualizations and Dashboard Creation

After preprocessing and cleaning, the datasets were imported into Power BI to create interactive dashboards that convert the data into actionable insights. Power BI was selected due to the software's powerful data modeling capabilities, the breadth of available visualizations, and the ability to create dynamic dashboards supporting decision-making.

4.6.1 Workflow for visualization

It involved the following steps:

1. **Data Import:** Clean datasets exported from Python were imported into Power BI in CSV format. Where applicable, relationships between datasets were established.
2. **Data Modeling:** Measures and calculated fields were created using DAX (Data Analysis Expressions) to identify additional Key Performance Indicators (KPIs), including, but not limited to, arrest rate, average academic pressure, and incidents/month.
3. **Dashboard Design:** Visualizations were arranged and organized around thematic dashboards, organized as three domain areas:

Urban Infrastructure (Water Main Breaks): visualizations across the dashboards describing the frequency and locations of breaks in conjunction with seasonal patterns.

Public Safety (Chicago Crimes): interactive charts across the dashboards depicting crime types, arrests by crime type, geographic ”hotspots,” and crime trends within selected timeframes (monthly/yearly).

Student Well-Being (Depression Dataset): dashboards depicting the correlation of age, CGPA, hours of sleep, academic pressure, and student mental health.

4.6.2 Visualization Techniques:

Various types of visuals were used, including:

- Clustered column charts for frequency analysis.
- Line charts for temporal trends.
- Pie/donut charts for distribution analysis.
- Maps for geospatial insights.

- Card visuals of KPIs (e.g., total crime, total breaks, percentage of students depressed).
- Interactivity:Slicers and filters were included in the dashboards to allow users to data explore interactively by year, by category, and by location.

The dashboards function as decision support systems by summarizing complex datasets into clear and interactive visuals. This demonstrates how to combine the organizational power of Python for data preprocessing with the visualization power of Power BI, as an end to end data analysis pipeline.

4.7 Conclusion

In this chapter, the datasets went through a structured process of cleaning, preprocessing, and exploratory data analysis (EDA). The raw data was transformed into a usable form by addressing missing values, duplicate records, and standardizing categorical or temporal features, to ensure uniformity and consistency throughout the datasets.

In addition, exploratory data analysis (EDA) was conducted to gain insight into the nature and the distribution of data. By leveraging descriptive statistics, quantile analysis, skewness, kurtosis, histograms, and regression, patterns were identified, and abnormal observations were detected. These processes highlighted important characteristics of the datasets and helped choose the methods available to analyze the data in detail.

After the preprocessing stage and EDA stage is complete, the data is ready for analysis. Chapter 6 will focus on Analytics and Visuals and data will be used for deriving insights through statistical exploration, visualizations, and dashboards.

Chapter 5

ANALYTICS & VISUALS

5.1 Introduction

After cleaning, preprocessing and exploring the data in the prior chapter, the next step is to translate processed datasets into valuable insights. Analytics and visuals are core components of this transition, as it is through these, that raw numbers and statistical outputs become interpretable patterns that may drive decision making.

This chapter provides the analytics and visualizations based on the three datasets selected:
(1) Water Main Breaks,
(2) Chicago Crime Records, and
(3) Student Depression Dataset.
In the development of the analytics, exploratory data analysis (EDA) was first applied in Python to provide a summary and interpretation of the datasets through descriptive statistics, quantiles, skewness, kurtosis, regression models, and visualizations. Then, Power BI dashboards were developed to visualize the processed datasets in a more interactive, decision-making supportive format.

The following sections describes the work done in each of these datasets and details the analytics and visuals.

5.2 DATASET 1: WATER MAIN BREAKS ANALYSIS

The dataset was obtained from the City of Kitchener's Open Data Portal. The "Water Main Breaks" dataset offers real-time records of pipe break incidents

throughout the city. It includes detailed information such as the type of break, location, material, repair method, date of occurrence, and more. [14]

5.2.1 Exploratory Data Analysis

In this chapter, we provide a detailed presentation of the results of the Exploratory Data Analysis (EDA) and the data visualization component of this study. This chapter follows the proposed methodology, as discussed in Chapter 4, and was conducted using Python in a Jupyter Notebook environment, exploring descriptive statistics, data distributions, and correlations in the data to reveal important relationships in the data.

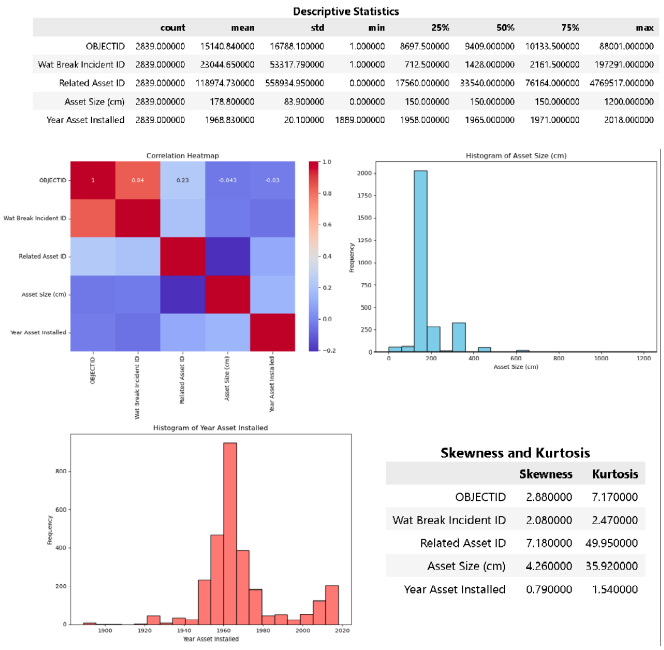

Figure 5.1: Composite Dashboard: Descriptive Statistics, Correlation, Heatmap, Skewness & Kurtosis

Descriptive Statistics

The descriptive analysis provides a summary of the complete dataset containing **2,839** entries and 15 variables. The dataset has both continuous and text data and has no missing values, making it complete. The dataset has date, cause, material and the type of repair, which makes it a complete dataset to investigate the topic presented.

a: Summary of Statistics

Insights

- The most common ”Asset Size” is 150 cm, based on the median.
- The maximum size is 1200 cm, suggesting some unusually large values.
- Most assets are old, with an average year of installation is 1968 and a median of 1965.

Statistical Summary of Categorical Columns This table provides key statistics for the 'Year Asset Installed' column. A major insight is that the infrastructure is, on average, very old. The mean installation year is approximately 1969, and the median is 1965. This indicates that half of all the assets in the dataset were installed before 1965. The fact that a significant portion of the network is over 50 years old is a crucial finding, as older assets are generally more prone to failures and breaks.

b: Correlation Heatmap

Insights

The water breaks heatmap highlights the frequency and distribution of breaks across different times and conditions, making it easy to spot patterns. Darker areas show where breaks happen most often, while lighter areas indicate fewer occurrences. This helps identify peak times or situations when water breaks are more common, pointing to possible underlying causes such as workload, scheduling, or environmental factors that may need attention.

c: Distributional Analysis: Histograms

Insight

Figure 5.1c shows the histograms and that most assets are relatively small, with a high concentration below 200cm and very few exceeding that range. This suggests that the dataset is mainly made up of smaller assets.

Regarding installation year, most assets were installed between the 1940s and

1970s, with a notable peak in the 1960s. Installations decreased afterward but experienced a smaller resurgence in the 2000s and 2010s.
These patterns imply that the asset base is mostly old and may need maintenance or replacement, while the recent additions point to some modernization in the system.

d: Distributional Analysis: Skewness and Kurtosis

Insight

- The Asset Size (cm) distribution, with a *skewness of 4.26* and *kurtosis of 35.92*, is highly *positively skewed* and extremely peaked, indicating that most assets are clustered at smaller sizes with a few very large outliers.
- In contrast, the Year Asset Installed distribution, with a *skewness of 0.79* and *kurtosis of 1.54*, is moderately *right-skewed* and slightly more peaked than a normal distribution suggesting that most assets were installed around a central time period, with some leaning toward more recent years.

5.2.2 Advanced Visualization and Insights (Power BI)

Dashboard Overview

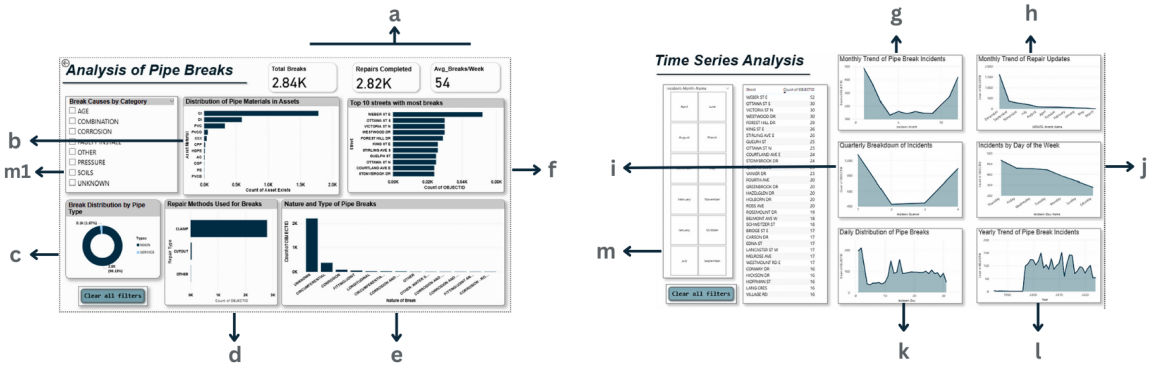

Figure 5.2: Water Main Breaks Dashboard

This Power BI dashboard provides an in-depth analysis of 485 water main break incidents, focusing on location, frequency, causes, materials, and time-based

patterns. It includes breakdowns by pipe material, type of break, repair method, geography, and temporal trends (year, month, day).

a. Key Metrics - KPI Cards

Observation: Total Breaks: 2.84k

- Repairs Completed: 2.82k
- Average Breaks per Week: 54

Insights:

The KPIs provided give key insights into the performance and frequency of pipe break incidents and how they are resolved. A total of **2.84K** breaks have been reported, with **2.82K** successfully repaired, showing a high repair success rate and effective maintenance response. Additionally, an average of **54 breaks** per week emphasizes the ongoing issue and highlights the need for continuous monitoring and proactive maintenance strategies to decrease break frequency over time.

b. Distribution of Pipe Materials

Observation:

displays the count of assets by material type shows that the most commonly used pipe materials are:

- Cast Iron (CI)
- Ductile Iron (DI)
- PVC and PVCO

Insights:

The use of older materials like Cast Iron may correlate with higher break rates. Gradually phase out older materials like CI in favor of modern, durable alternatives like PVC or PE.

c. Break Distribution by Pipe Type

Observation:

Pie chart reveals that:

- 98.13% of break incidents occurred in the main pipes
- Only 1.87% were in service lines

Insights:

Major infrastructure issues are concentrated in the main distribution network, highlighting the need for focused maintenance efforts on these critical pipelines.

d. Repair Methods Used

Observation:

The most common repair methods include:

1. Clamp
2. Cutout
3. Other minor fixes

Insights:

Most breaks are often fixed through simple repairs, suggesting that many failures could be manageable if identified early. However, this also indicates a lack of long term solutions, which may result in recurring problems

e. Nature and Cause of Breaks

Observation:

- Break causes are categorized as:
- Corrosion
- Circumferential and Longitudinal Cracks
- Fitting/Joint Failures
- Unknown Causes

Insights:

Corrosion is a leading cause of pipe failure. A significant number of cases are marked as **"Unknown"**, indicating a need to improve data collection during field inspections.

f. Top Streets with Most Breaks

Observation:

The streets with the highest number of reported breaks include:

- Weber Street East (highest)
- Stirling Avenue South
- Queens Boulevard
- Courtland Avenue East

Insights:

These streets consistently appear in the top ranks for incidents. This suggests a localized infrastructure problem and points to areas needing urgent repair or pipe replacement.

Temporal Analysis of Breaks

g. Monthly Trends Break incidents increase noticeably towards the end of the year, particularly in November and December.

i. Quaterly Trends Most incidents occur in Q1 and Q4, with Q4 showing the highest frequency.

j. Weekday Analysis The highest number of breaks are reported on:

- Thursdays
- Fridays

l. Yearly Trends There is a clear upward trend in break incidents since the year 2000, peaking after 2010, which may reflect aging infrastructure and increased stress on the network.

Insights:

Time based patterns suggest that weather, water pressure variations, and increased weekday usage may contribute to higher break incidents. The spike at year end may relate to freezing conditions or system overload.

Filters Available

To enhance interactivity and allow for more focused analysis, the Power BI dashboard includes several slicers and filters. These enable users to dynamically adjust the visuals based on specific attributes or time frames. The available filters include:

m1. Break causes by category:

Observation:

- 219 breaks were caused by pipe age, and 217 of them were repaired.
- 137 breaks were due to a combination of factors, and all were repaired.
- 111 breaks happened due to corrosion, with 109 repaired.
- 9 breaks were caused by faulty installation, and all were repaired.
- 23 breaks occurred due to pressure, all were repaired.
- 20 breaks were related to soil conditions, all were repaired.
- 2.29K breaks were due to other reasons, with 2.27K repaired.
- 35 breaks had an unknown cause, and all were repaired

Insights:

This study indicates, pipe breaks mainly occur due to three factors:

1. The age of the pipes
2. Corrosion
3. Other unidentified issues.

Despite different causes, repairs are typically done quickly and well, reflecting a robust maintenance program. However, the volume of breaks caused by old pipes and corrosion indicates a need for inspections and appropriate replacements to prevent breaks in the future.

m. Break Incident Month:

Observations:

- 485 pipe breaks were recorded in January.
- 368 pipe breaks were recorded in February.
- 224 pipe breaks were recorded in March.
- 129 pipe breaks were recorded in April.
- 157 pipe breaks were recorded in May.
- 140 pipe breaks were recorded in June.
- 156 pipe breaks were recorded in July.
- 144 pipe breaks were recorded in August.
- 142 pipe breaks were recorded in September.
- 207 pipe breaks were recorded in October.
- 269 pipe breaks were recorded in November.
- 418 pipe breaks were recorded in December.

Insights:

The data clearly shows that pipe breaks in Kitchener peak during the winter months.January and December experience the highest number of breaks, with 485 and 418 incidents, respectively. In contrast, the fewest breaks occur in the spring and early summer, particularly in April (129) and June (140). This seasonal pattern suggests that cold temperatures, ground frost, and possibly freeze thaw cycles significantly contribute to the stresses placed on pipeline infrastructure.

5.3 Understanding the Seasonal Influence on Pipe Breaks:

Kitchener's climate is classified as humid continental, with cold, snowy winters. Average high temperatures range from about $-3\,^{\circ}\mathrm{C}$ in January to around $-2\,^{\circ}\mathrm{C}$ in February, with average lows between $-11\,^{\circ}\mathrm{C}$ and $-10\,^{\circ}\mathrm{C}$ [17–19]. Snowfall is substantial during winter—Kitchener sees roughly 44 cm in January and 30 cm in February [18, 19]. Persistent frost and low temperatures can deepen ground freezing well below the typical pipe depth, causing soil shifts that stress or break pipes [20].

City officials note that extended periods of $-20\,^{\circ}\mathrm{C}$ to $-30\,^{\circ}\mathrm{C}$ lasting 10 to 20 days are especially damaging, as they drive frost deeper into the ground and increase the likelihood of main breaks [20]. In response, they often recommend letting taps drip—especially basement taps—to keep water flowing and reduce the risk of freezing [20–22].

This analysis highlights that winter weather significantly affects pipe break ratios. It reinforces our observation that sustained extreme cold and deep frost are key factors contributing to pipe failures in Kitchener. To reduce break rates during winter, it is crucial to implement mitigation strategies such as proper insulation, methods to prevent dripping, infrastructure upgrades, and monitoring ground frost.

5.4 Recommendations:

1. Replace High-Risk Pipe Materials

- Prioritize replacement of Cast Iron and other outdated materials.
- Focus initial upgrades on streets with repeated failures.

2. Strengthen Preventive Maintenance

- Monitor pipes in Weber St E, Stirling Ave S, and Queens Blvd proactively.
- Use pressure sensors and flow monitors to predict stress points.

3. Improve Field Data Collection

- Train staff to record detailed and accurate break causes.
- Introduce digital data entry tools for real-time updates.

4. Focus on Main Line Rehabilitation

- Allocate funds to upgrade main distribution lines, as they represent 98% of failures.

5. Prepare for Seasonal Peaks

- Increase inspection and maintenance teams in Q4 and Q1, especially before winter.

6. Reevaluate Repair Methods

- Review the success rates of clamp and cutout methods.
- Where failures recur, prioritize full pipe replacement over temporary repairs.

5.5 Conclusion:

The application of Python for data cleaning, combined with Power BI for data visualization, presented advantages for using a data driven methodology to explore infrastructure data. The data driven approach identified important trends, indicating that aging materials, seasonal stresses, and location of hotspots were the main drivers of water pipe breaks.

The results generated analysis can assist utility managers in preparing decisions about preventive maintenance, infrastructure expenditure, and policy adjustments. Ultimately, this will improve the safety and reliability of the water distribution system to the community.

5.6 DATASET 2: STUDENT DEPRESSION ANALYSIS

The Student Depression dataset, sourced from Kaggle [16], captures approximately **28,000** surveys collected from students in India. It records various factors such as academic pressure, financial strain, sleeping patterns, CGPA, study satisfaction, degree type, gender, age, dietary practices, family history of mental illness, and if the student suffers from depression. All of these factors allow us to glimpse how several academic and personal factors affect a student's well-being.

5.6.1 Exploratory Data Analysis

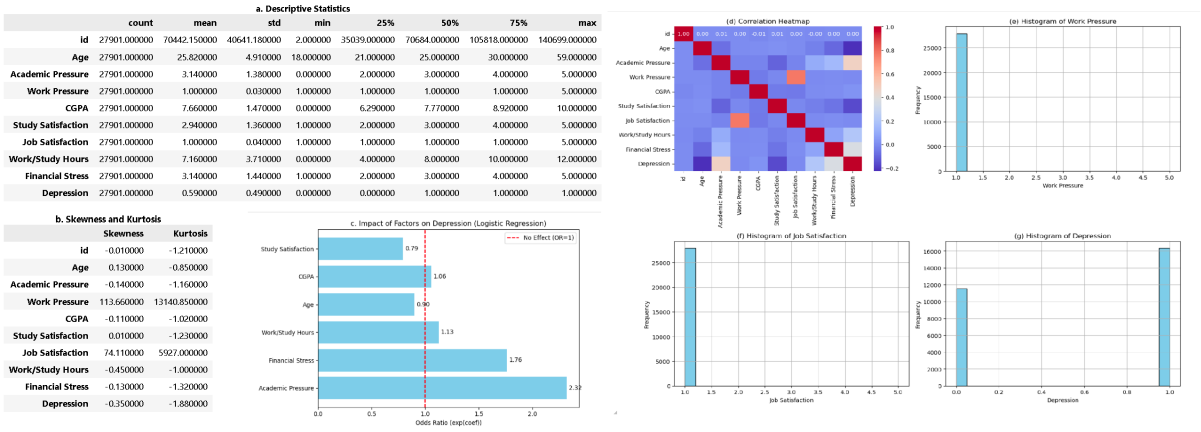

Figure 5.3: Composite Dashboard: Descriptive Statistics, Skewness & Kurtosis, Regression Findings, Correlation Heatmap and Histograms

a. DESCRIPTIVE STATISTICS

The initial analysis of the dataset included descriptive statistics for any numerical attributes. The dataset has 27,901 entries and 25 columns and provides a pretty good snapshot of the student population surveyed.

Figure 5.4a shows the results of the key statistics as follows:

Age: The average age of students is approximately *25.8 years* with a standard deviation of 4.9. The data indicate the majority of students are between *18-30* years of age.

Academic Pressure: The mean score for academic pressure is *3.14 out of 5*, which suggests that students are feeling moderate to high academic pressure overall.

CGPA: The average CGPA is *7.66* with a standard deviation of *1.47*, which shows that the majority of students have relatively good grades.

Study/Work Hours: Students reported an average of *7.16* hours spent working and studying, with a standard deviation of *3.71*, which indicates a fair bit of time spent working and studying.

Depression Label: The dataset is imbalanced with *16,336* students (58.55%) reporting depression symptoms and *11,565* students (41.45%) reporting no symptoms. As shown in figure 2.

b. DISTRIBUTIONAL ANALYSIS: SKEWNESS & KURTOSIS

Figure 5.4b shows skewness & Kurtosis
Skewness measures the asymmetry of data distribution, showing whether values are more spread out on the left or right side of the mean.
Kurtosis measures the "tailedness" of a distribution, indicating whether data have heavy or light tails compared to a normal distribution.

Insights:

1. **Age :** *Skewness = 0.13, Kurtosis = -0.85*

- Slightly right skewed but very close to symmetric, confirms that majority of students are young.
- Flatter than a normal distribution, with fewer extreme values.

2. **Academic Pressure :** *Skewness = -0.14, Kurtosis = -1.16*

- Slight left skew, almost balanced.

- Flatter distribution, meaning responses are spread across the scale, indicating a wide range of perceived stress levels among students.

3. Work Pressure : *Skewness = 113.66, Kurtosis = 13140.85*

- Extremely right skewed with very high kurtosis.
- Data is highly concentrated at the lowest values, with a few extreme cases pulling the distribution far to the right.

4. CGPA *Skewness = -0.11, Kurtosis = -1.02*

- Almost symmetric with a slight left skew.
- Flatter than normal, values are spread without heavy tails.

5. Study Satisfaction : *Skewness = 0.01, Kurtosis = -1.23*

- Perfectly symmetric.
- Flatter than normal, spread evenly across categories.

6. Job Satisfaction : *Skewness = 74.11, Kurtosis = 5927.00*

- Extremely right-skewed with very high kurtosis.
- Almost all students report very low job satisfaction, with very few extreme responses creating long tails.

7. Work/Study Hours : *Skewness = -0.45, Kurtosis = -1.00*

- Moderate left skew $\rightarrow$ more students are on the higher side of study hours.
- Flatter than normal, responses spread across the range.

8. Financial Stress : *Skewness = -0.13, Kurtosis = -1.32*

- Nearly symmetric with slight left skew.
- Flatter distribution, students' responses cover the whole range.

9. Depression : *Skewness = -0.35, Kurtosis = -1.88*

- Some left skew → more students reporting depression (value 1).
- Very flat distribution because data is concentrated in two categories (0 and 1) rather than
- spread.

c. LOGISTIC REGRESSION

Logistic regression is used for binary outcomes, such as predicting whether a student is depressed (1) or not (0). It helps identify factors like academic pressure, financial stress, study hours, CGPA, study satisfaction, and age that are significantly linked to the likelihood of depression.

The model provides coefficients, p-values, and odds ratios, showing the strength and direction of these relationships.

Insights:

1. Academic Pressure : *OR = 2.32, p ¡ 0.001*

Students with higher academic pressure are more than twice as likely to experience depression compared to those with lower academic pressure. This is the strongest predictor.

2. Financial Stress : *OR = 1.76, p ¡ 0.001*

Students experiencing financial stress are 1.76 times more likely to be depressed, showing a strong positive effect.

3. Work/Study Hours : *OR = 1.13, p ¡ 0.001*

Longer work or study hours slightly increase the chances of depression.

4. CGPA : *OR = 1.06, p ¡ 0.001*

Higher CGPA is weakly but positively associated with depression, suggesting that academic achievement alone does not protect students from mental health challenges.

5. **Age :** *OR = 0.90, p ¡ 0.001*

Older students are slightly less likely to experience depression compared to younger ones.

6. **Study Satisfaction :** *OR = 0.79, p ¡ 0.001*

Higher study satisfaction reduces the likelihood of depression, indicating a protective effect.

d. CORRELATION ANALYSIS

A correlation heatmap was created to visualize the linear relationships between the numerical variables. This was important for identifying potential predictors of depression. The heatmap displayed some important correlations:

Insights:

1. **Depression & Academic Pressure**

Correlation: Positive

Students who experience higher academic pressure are more likely to report symptoms of depression. This may be attributed to high expectations from themselves, parents, or faculty, as well as fear of failure and constant comparisons with peers. Competitive environments, such as merit based scholarships and limited seating, can also contribute to this issue.

2. **Depression & Study Satisfaction**

Correlation: Negative

Students who are more satisfied with their studies generally experience lower levels of depression. Higher satisfaction may lead students to feel more in control of their education and enjoy the learning process. They often benefit from better support from faculty and engage in more interesting coursework. Additionally, a sense of achievement helps protect against feelings of low mood.

3. **Depression & Work/Study Hours**

Correlation: Positive

Longer study/work hours are linked with higher depression.
Long work hours leave little time for self-care, rest, or socializing. Students often struggle to balance their studies with part-time jobs due to financial stress. This can lead to burnout from an unbalanced life.

4. Depression & CGPA

Correlation: Weak Positive

CGPA doesn't show a strong relation with depression in this dataset. CGPA had a weak but present inverse relationship with depression level.
Mental health issues can impact all students, including those who perform well academically. Some students may maintain their grades while hiding certain problems, such as family issues or the constant pressure to achieve high marks.

5. Academic Pressure & Study Satisfaction

Correlation: Strong Negative

Students experiencing greater academic pressure often report lower satisfaction with their studies. Strategies to reduce this pressure, such as a balanced curriculum and flexible deadlines, can enhance satisfaction.

6. Academic Pressure & Financial Stress

Correlation: Moderate Positive

Students who face financial stress often feel more academic pressure. They may not have access to important resources like tutoring, textbooks, and technology. This lack of resources makes their situation even more stressful.

7. Study Satisfaction & Job Satisfaction

Correlation: Moderate Positive

Students who are satisfied with their study life are also more likely to feel content with their job or work life. A sense of stability or success in one area often influences other parts of life.

8. Sleep Duration & Study Satisfaction

Correlation: Weak to Moderate Positive

Students who sleep more tend to be more satisfied with their studies. Well rested students may have better focus, motivation, and engagement in their academic life.

9. Work/Study Hours & Academic Pressure

Correlation: Moderate to High Positive

The longer students engage in study or work, the more academic pressure they report. This may indicate overwork or poor time management.

10. Financial Stress & Job Satisfaction

Correlation: Moderate Negative

Students experiencing financial stress are likely to have lower job satisfaction, possibly due to low paying or overly demanding jobs.

11. CGPA & Study Satisfaction

Correlation: Weak Positive

Students who have higher CGPAs typically report greater satisfaction with their studies.

e/f. DISTRIBUTIONAL ANALYSIS: HISTOGRAMS

Insights:

1. Age

Most students are between 18–25 years old. Very few are above 30. This shows the dataset mainly represents young students.

2. Academic Pressure

Values range from 1 to 5. Many students report medium to high levels of academic pressure. Academic stress is common among students.

3. Work Pressure

Most students are at the lowest level (1). Very few have higher work pressure. This suggests that students generally do not have much job related stress.

4. CGPA

The majority of students have a CGPA between 6 and 10, with fewer students at very low CGPA levels. Academically, many students are performing reasonably well.

5. Study Satisfaction

Responses are spread across the scale, but many students fall around 3 and 4. This shows moderate levels of satisfaction with studies.

6. Job Satisfaction

Almost all values are at 1 (very low). This may be because most students do not have jobs while studying.

7. Work/Study Hours

Most students study between 4 and 10 hours daily, with some going up to 12 hours. This indicates a significant time commitment to studies.

8. Financial Stress

Values are spread across the scale, with many students reporting higher levels (3–5). This shows that financial stress is a major issue for students.

9. Depression

The data is mostly split between 0 and 1 (0 = no depression, 1 = depression). A large portion of students fall into the depression category. Depression is a significant concern in this dataset.

5.6.2 Advanced Visualization and Insights (Power BI)

Dashboard Overview

The dashboard presents a comprehensive analysis of depression among 28,000 surveyed students, using various academic, personal, and demographic variables. The visuals explore correlations between depression and factors such as academic pressure, financial stress, sleep duration, satisfaction with studies,

and degree programs. Interactive slicers (filters) allow for breakdowns by city, gender, profession, and sleep duration.

Figure 5.4: Student Depression Dashboard

a. KPI Cards (Top Metrics)

Total Students Surveyed: 28,000

- Average Age: 26 years
- Depression Ratio: 59%
- Average Work/Study Hours: 7.16 hours/day
- Suicidal Thoughts: 63%

Insights:

Most students (59%) report experiencing depression, and 63% have had suicidal thoughts, highlighting serious mental health issues. The average workload exceeds 7 hours, which could add to mental stress.

b. Depression by Degree

Observation:

- B.Tech, B.Sc., and M.Tec.h students report higher levels of depression
- Students in BHM(Bachelor of Hotel Management), BCA(Bachelor of Computer Applications), and B.Com(Bachelor of Commerce) show comparatively lower depression

Insight:
Technical degrees correlate with higher depression rates, likely due to heavy workloads or competitive environments.

c. Depression by Financial Stress

Observation:

- Strongly Agree (Financial Stress): 5.5K depressed
- Depression decreases as stress decreases

Insight:
Financial stress is a strong predictor of depression among students.

d. Impact of Sleep Duration on Depression

Observation:

- Students with ¡5 hrs sleep show high depression levels (5.4K)
- Students with ¿7 hrs sleep have relatively lower depression counts
- Highest ”No Depression” counts among those sleeping 7–8 hours

Insight:
There is a clear connection between sleep duration and depression. Students who get less than 5 hours of sleep exhibit the highest levels of depression, while those who sleep more tend to have better mental health.

e. Depression by Academic Pressure

Observation:

- Strongly Agree: 5.4K students are depressed
- Agree: 3.9K
- Neutral: 4.5K
- Disagree/Strongly Disagree: Less than 1.6K combined

Insight:
There is a positive correlation between academic pressure and depression. Students who strongly agree that academic pressure is high tend to experience higher levels of depression.

f. Depression by Study Satisfaction Level

Observation:

- Depression is highest in students who are dissatisfied (Strongly Disagree: 3.9K, Disagree: 3.8K)
- It decreases with satisfaction: only 2.1K for ”Strongly Agree”

Insight:
Increased study satisfaction is linked to reduced depression levels, suggesting that student engagement and academic satisfaction have a protective effect.

g. Depression Label (Yes/No)

Observation:

- Yes: 58.58%
- No: 41.42%

Insight:
Over half of the surveyed students are experiencing symptoms of depression.

h. Filters Available

Use of Filters These features make the dashboard interactive, enabling granular analysis, such as exploring depression among males and females.
Filters allow users to analyze mental health trends by:

1. Age:

   Approximately 14,000 students in the dataset fall within the age group of 18 to 25 years. Among them, 67% are experiencing symptoms of depression, and around 68% have reported having suicidal thoughts. This group also faces high academic and financial stress, and very low study satisfaction, which likely harms their mental health. The pressure of studies and money during this time can be very hard on them. About 16,000 students in the 25 to 40 age group are represented in the dataset. Within this group, around 51% show signs of depression, and 60% report suicidal thoughts. This age group also experiences the highest levels of financial stress, likely due to responsibilities like supporting families, managing debt, and balancing work with education.

2. City:

   Hyderabad reports the highest percentage of students with suicidal thoughts at 68%, accompanied by a depression rate of 67%. Indore follows with 62% for suicidal thoughts and 60% for depression, while Agra shows 64% and 53%, respectively. Also, students in other major cities experience depression and suicidal thoughts, typically between 53% and 60%. The common factor among these students is sleeping less than 5 hours per night and facing significant academic and financial stress, contributing to their mental health challenges.

3. Gender:

   Among the surveyed students, about 12,000 are female. Of these, 63% reported suicidal thoughts and 58% showed signs of depression. Female students face high financial stress, limited sleep, and significant academic pressure, suggesting a vulnerability to mental health challenges due to

these combined factors. Among the 16,000 male students, 63% reported suicidal thoughts and 59% showed signs of depression. They also face high financial stress, less sleep, and low study satisfaction, which likely contribute to their poor mental health.

4. Family History of Mental Illness:
   In a study of 14,000 students without a family history of mental illness, 62% reported suicidal thoughts and 56% showed signs of a family history, 65% had suicidal thoughts and 61% showed signs of depression, suggesting a link between family history and mental health issues.

5. Dietary Habits:
   Students who consume unhealthy food have a higher depression rate (71%) and higher Suicidal thoughts (70%), while students with healthy dietary habits have a lower depression rate (45%) and lower Suicidal thoughts (56%).

5.7 Recommendations

- **Mental Health Resources:** Launch awareness campaigns, peer support programs, and provide access to on-campus mental health resources.
- **Academic Reforms:** Institutions should consider curriculum moderation, offer academic counseling, and reduce overburdening schedules.
- **Financial Aid:** Scholarships and budgeting education to ease financial anxiety.
- **Sleep Education:** Promote awareness around healthy sleep habits.
- **Healthy Diet:** A healthy diet has a significant impact on students' mental health, so it is essential to encourage students to adopt healthy eating habits.
- **Degree-specific Interventions:** Address challenges specific to technical degree holders.

- **Study Satisfaction:** Improve teaching quality, mentoring, course flexibility, and promote feedback systems to enhance satisfaction.
- **Sleep Duration:** Promote healthy sleep hygiene, discourage late-night academic loads, and encourage mindfulness/sleep workshops.

5.8 Conclusion

The dashboard provides a clear overview of the mental health crisis impacting students today. It highlights the need for structured institutional responses, including both academic and psychological support. Key interventions can significantly improve student well-being and academic performance.

5.9 DATASET 3: CHICAGO CRIME DATA ANALYSIS

The Chicago Crimes dataset sourced from Kaggle [15] is a raw dataset contains over 8 million records. The dataset is big data, containing millions of records, which cannot be processed efficiently in traditional tools like Microsoft Excel. Even pandas (Python library) struggled with memory limitations when attempting to load the full dataset at once. Therefore, advanced data engineering approaches were applied for efficient processing. An interactive dashboard was created that provides an extensive overview of crimes across the city. It includes maps, crime trends across time and location, and KPI's demonstrating different metrics.

5.9.1 Exploratory Data Analysis

a. Descriptive Statistics

	count	mean	min	25%	50%	75%	max	std
ID	7842937.000000	7071242.243789	634.000000	3814349.000000	7072857.000000	10260821.000000	13142449.000000	3526350.447696
Date	7842937	2010-07-17 10:44:42.238712832	2001-01-01 00:00:00	2005-02-08 19:07:24	2009-08-07 09:30:00	2015-08-14 19:50:00	2023-07-09 23:59:00	nan
Beat	7842937.000000	1185.691965	111.000000	621.000000	1034.000000	1731.000000	2535.000000	703.207610
District	7842890.000000	11.294907	1.000000	6.000000	10.000000	17.000000	31.000000	6.923590
Ward	7728088.000000	22.756783	1.000000	10.000000	24.000000	34.000000	50.000000	13.854398
Community Area	7225461.000000	37.475506	0.000000	23.000000	32.000000	57.000000	77.000000	21.505226
X Coordinate	7754915.000000	1164604.000000	0.000000	1152984.000000	1166430.000000	1176483.000000	1205119.000000	7721.062508
Y Coordinate	7754915.000000	1885792.875000	0.000000	1893078.000000	1898736.000000	1909283.000000	1951652.000000	31579.066250
Year	7842937.000000	2010.040908	2001.000000	2005.000000	2010.000000	2015.000000	2023.000000	6.336775
Latitude	7754915.000000	41.842924	36.619446	41.768726	41.855430	41.906496	42.022911	0.451843
Longitude	7754915.000000	-87.671432	-91.686569	-87.713654	-87.665817	-87.628181	-87.524529	3.991130

b. Descriptive Statistics for Categorical Columns

	Unique Values	Most Frequent	Frequency	Missing
Block	63068	100XX W OHARE ST	16538	0
Primary Type	36	THEFT	1654618	0
Description	546	SIMPLE	920254	0
Location Description	217	STREET	2044066	10730
FBI Code	26	06	1654618	0

c. Skewness and Kurtosis

	Skewness	Kurtosis
ID	0.030000	-1.260000
Beat	0.390000	-0.960000
District	0.430000	-0.870000
Ward	0.170000	-1.060000
Community Area	0.180000	-1.100000
X Coordinate	-5.290000	347.560000
Y Coordinate	-3.080000	176.250000
Year	0.360000	-1.010000
Latitude	-3.140000	181.020000
Longitude	-4.590000	284.070000

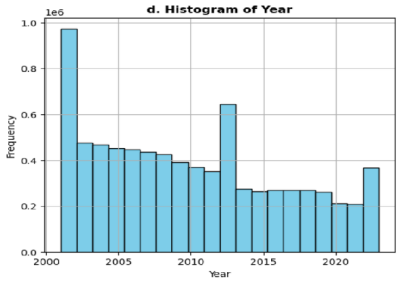

Figure 5.5: Composite Dashboard: Descriptive Statistics, Skewness & Kurtosis and histograms

DESCRIPTIVE STATISTICS:

a. NUMERICAL ATTRIBUTES:

The initial analysis of the dataset included descriptive statistics for any nu-

merical attributes. The dataset has over 7 Million entries and 22 columns and provides a pretty good snapshot of the Chicago Crime.

The results of the key statistics of numerical columns are as below:

ID: These are unique identifiers, with all 7,842,937 values unique.

Administrative Codes: District, Ward, and Community Area are categorical numerical codes with 24, 50, and 78 unique values, respectively.

Spatial attributes: (X and Y coordinates, Latitude, Longitude) cover a broad geographic range, with latitude values ranging from 36.61° to 42.02° and longitude ranging from –91.68° to -87.52°, which captures the geographical extent of the city.

Year: values range from 2001 to 2023, which captures 20 years of data.

Temporal Attributes: Time related features (Month, Day, Hour) capture seasonal and hourly variations in time, from 12 unique months, 31 unique days, and 24 hours.

Completeness: The complete absence of missing values from all numerical variables demonstrates the data has value and is trustworthy for the analysis.

b. CATEGORICAL ATTRIBUTES:

The results of the key attributes are as follows:

Block: This attribute refers to the anonymized block level location of each incident. There are a total of 63,068 unique values, with the most common block being "*100XX W OHARE ST*", with a total of 16,538 occurrences. This shows that crime incidents are spread out over many blocks in urban spaces, and some blocks have more incidents occurring.

Primary Type: This variable sorts all offenses into 36 unique crime types. *Theft* is the most frequent crime, with 1,654,618 incidents, which leads to the crime type being the most prevalent crime type in the dataset.

Description: This variable provides more detail within the primary crime type, with 546 unique values present. The most common description is "*Simple*", which occurs 920,254 times.

Location Description: This variable tells whether the incident occurred at a certain type of place. It consists of 217 types of places, with the most common location being "*Street*" with 2,054,796 occurrences, suggesting that public streets were the most common area for incidents to occur.

FBI Code: This variable categorizers crimes using standard FBI classification codes. There are 26 unique FBI codes that occurred within the research, with the most common being "06", in reference to theft (1,654,618).

DayOfWeek: This variable represents the day that an incident occurred. As expected there would be seven unique values for the days, with *Friday* being the most common day associated with 1,178,197 records.

c. DISTRIBUTIONAL ANALYSIS: SKEWNESS & KURTOSIS

Insights:

ID, Beat, District, Ward, Community Area, and Year:
These variables exhibit very *low skewness* values, revealing their symmetrical distribution (values from 0.03 to 0.43). This aligns with the previous findings that indicate these variables' distributions are uniform-like or multi-modal.

These variables exhibit a *negative kurtosis* (values from -0.87 to -1.26), indicating they are platykurtic distributions; distributions that are flatter (less tall) at the peak and have lighter tails than a normal distribution.

X Coordinate, Y Coordinate, Latitude, and Longitude:

These variables exhibit very high *negative skewness* values (values from -3.08 to -5.29), confirming they are very left-skewed (most data points lie toward the right side of the distribution, with a long left-handed tail). This is consistent with the histogram plotting of each variable, which indicated the majority of observations were clustered at high values relative to the distribution.

These variables exhibit extremely *high positive kurtosis* (values from 176.25 to 347.56), indicating they are distributions that have very tall and sharp peaks and heavy tails with a larger number of outliers (equivalent to having heavier tails). This also lines up with the identified histogram plots, indicating the data was heavily packed within one or two bins around the higher number distribution, creating the sharp peak and heavy-tailed distribution.

d. DISTRIBUTIONAL ANALYSIS: HISTOGRAMS

Insights:

Year: The year variable shows a multi modal distribution with peaks and dips in various periods, indicating that data collecting was not standardized over time and instead spiked at certain points. There are clear clusters in the early 2000s, again around 2010, and a final cluster in the 2020s. This pattern likely coincides with separate data collection periods or specials projects that were implemented at those times.

e. CRIME FORECAST

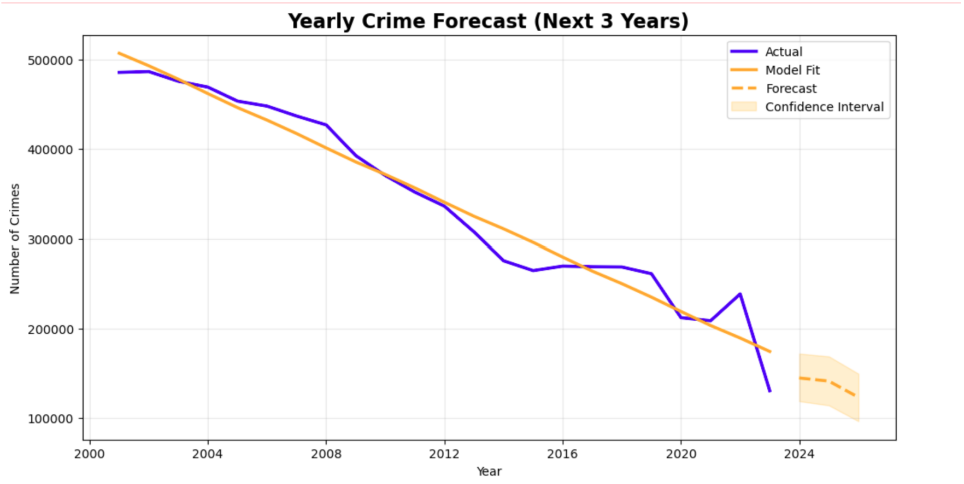

Figure 5.6: Crime Forecast

Insights:

Downward Trend: The visual shows a clear downward trend in crime rate There are some short term spikes and drops in crime trend around 2015-2022.

Covid Era: Between 2019 and 2021 the crime trend captures a drop. this drop can be attributed to the emergency situation in the city during which shops were closed, streets were empty and people stayed at home.

Future Prediction: For the next 3 years the crime rate is predicted to be decreasing but at a slower rate and by the end of 2024 we can see that it may fall to around 120,000 cases.

5.9.2 Advanced Visualization and Insights (Power BI)

Dashboard Overview

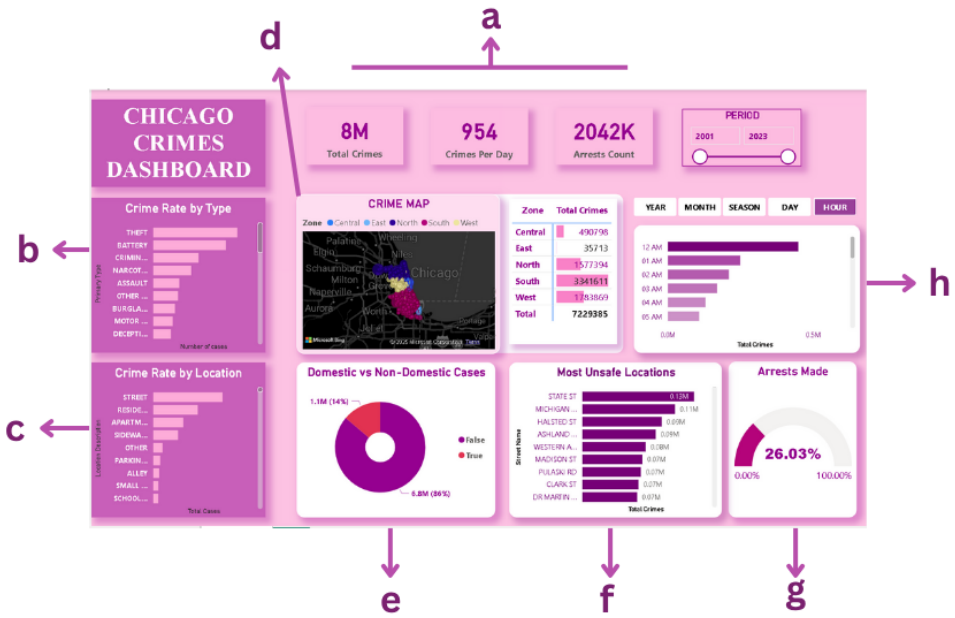

Figure 5.7: Crime Dashboard

This dashboard analyzes over 8 million reported crimes in Chicago (2001–2023). Due to the dataset's size, we used chunk-based processing and Parquet storage for efficient handling before creating visuals in Power BI. It highlights crime types, hotspots, time patterns, arrests, and case resolutions, offering a clear view of the city's crime trends. Key metrics like total crimes, arrests, and average daily crime rate provide a quick snapshot, while interactive visuals reveal where, when, and how crimes happen. Overall, it serves as a crime intelligence tool to support law enforcement, policymakers, and researchers in improving public safety.

a. KPI Cards (Top Metrics)

Observation:

- Total crimes reported 8 Million
- Average number of crimes happening on a daily basis is 954 crimes per day.

- Crimes that resulted in arrests among the 8 Million cases are about 2 Million
- The timespan of this study is 22 years(2001-2023)

Insights:

Everyday on average 950+ people are affected by the crimes in the city. Many of these crimes are not even resolved as indicated by the arrests count which is way lower than the total number of crimes. This puts the citizen's safety at stake. It also indicates the low efficiency of police in providing safety to the citizens.

b. Crime Rate by Type:

Observation:

Theft is the most common type of crime in Chicago with almost 1.6 Million reported cases in the past 22 years followed by Battery and Criminal Damage Crimes like sex offense, public peace violation and liquor law violation are in low comparison to the others but still it has an impact.

Insights:

We see that property related like theft and violent offenses are city's most frequent issues. So the law enforcement should make these areas a priority and come up with strategies to eradicate these crimes and make the community a safe space for everyone.

c. Crime Rate by type of Location:

Observation:

Public spaces like Streets are the most common locations for crimes to occur in Chicago with over 2 Million cases reported followed by residential areas like Homes and Apartments while parks, offices, gas stations shows a lower frequency of crimes.

Insights:

We see that public spaces and residential areas dominate crime occurrences this can be due to less security in such areas. so, security measures should be improved in such areas like installing surveillance cameras, police patrolling etc.

d. Top Crime Zones:

Observation:

The city of Chicago is divided into 5 zones i-e North, South, East, West And central. South zone shows the highest concentration of crimes i-e about 46% of all crimes are reported in the South zone of the city making it the most unsafe zone of the city, followed by west zone. While East is shown to have the lowest crime rate.

Insights:

We see that South is the most unsafe zone of the city. According to statistics South and West zone struggles with violent crimes, it can be due to a variety of reasons like poverty, lack of resources, unemployement and certain ethnic groups residing in these zones etc. on the other hand East zone is shows to have the lowest crime rates. Now this is because the East of Chicago is mainly Lake Michigan. We also see that Central Chicago has the lowest crime rates. It is due to the fact that this zone is the city's business and financial hub, with tourist locations etc so the goverment invest heavily in this zone thus keeping it much safer than other parts of the city.

e. Domestic Vs Non Domestic Cases:

Observation:

Out of the 8 Million reported cases 1.1 Million are classified as Domestic crimes while 6.8 Million crimes are classified as Non-Domestic in nature, which makes up about 86% of the cases.

Insights:

We see that majority of the crimes are non-domestic and according to our earlier insights we saw that most of the crimes occur in public spaces like streets, making them the most vulnerable. While only 14% of the cases are classified as Domestic, its still a significant ratio which shows that homes and apartments are still not entirely safe. If we co-relate with the Location chart, we see that Non-domestic cases align with public spaces like streets, sidewalks and parking areas. Domestic cases align with residences, apartments etc. Domestic cases often goes unreported which is why we see a relatively less number of these sorts of crimes reported, maybe the numbers are much higher than this.

f. Most Unsafe Locations:

Observation:

The bar chart shows Top 10 most unsafe locations of Chicago as indicated by the number of crimes occurred in them. State Street tops the chart with almost 133k crimes reported in the past 22 years followed by Michigan Avenue and Halsted Street

Insights:

While State street appears to be the most unsafe location, but if we co-relate it with our zone chart we can see that State Street is in the Central zone which was said to be the most safe zone yet State street reports about 16 crimes per day on average, while the top locations in the other zones on average reports 8 cases daily. It can be due to a variety of factors:

1. As central zone is the downtown of the city so crimes are reported more but these crimes are generally non-violent in nature like theft which we can see in our Crime Type chart.
2. As central Chicago is a tourist hub so its packed with tourists and has a busy night life which means more opportunities for crimes like theft.
3. Generally, If we analyze in tourist centric places people reports more crimes like pick pockets, bag stolen etc while in places with little to no

tourism many such crimes goes unreported. This makes it seem like the central areas have more crimes.

g. Arrests Made:

Observation:

Out of the 8 million reported cases only 26% of the cases led to an arrest.

Insights:

Only 26% of the crimes in the data set led to an arrest, meaning that most of the cases are unresolved at the initial stage. It points towards crimes resolution difficulties and provides directions for targeted reforms in policing strategies. If we compare the arrest rate between domestic and non-domestic crimes we see that non-domestic crimes have an arrest rate of 18% only, while non-domestic crimes have an arrest rate of 27%. Low arrest rate can make the community feel less secure as citizens may think that offenders are less likely to be caught. While it may give a boost to the criminals as they might think they can get away with their crimes.

h.Temporal Trends:

1. Yearly Crime Analysis:

Observation:

The line chart shows the total crimes occurring each year. From 2001-2004 the crime rate was at a peak with 0.5 million cases occurring per year. From 2005 we see a gradual decline in the crime rate till 2015 with cases dropping to less than 0.3 million cases per year. From 2016 to 2019 we see that the crime rate is relatively stable with no such peaks and dips. Around 2020 there is a noticeable drop in crime rates likely due to Covid 19 pandemic. Post pandemic the crime rate slightly rose again.

Insights:

In the past 2 decades the crime rate has significantly dropped to almost half

which shows that their have been improvements in law enforcing strategies. The decline in Covid era can be due to the fact that lockdown was imposed and people stayed home thus non-domestic crimes were reduced to a great extent.

2. Monthly Crime Analysis:

Observation:

July and August reports the highest number of crimes almost reaching 0.7 million. February and December reports the lowest number of crimes.

Insights:

A peak in crime rates during July and August can be due to the warmer season during which people tend to be more outside their houses. More people in public spaces increases the opportunities for more crimes.While February and December are much colder having heavy snowfall which makes city life less mobile thus crime rate drops. But keep in mind that this drop might only be in non-domestic crimes.

- If we analyze July we see that in the past 22 years July alone reported 724k crimes with an average of 1047 crimes per day.
- Streets are the most vulnerable locations as Thefts are the crime with the highest frequency during this month.
- 85% of the cases in july are Non-Domestic in nature.
- If we analyze February we see that in the past 22 years February reported 548k crimes with an average of 844 crimes per day.
- Streets and residences are the most vulnerable locations with theft and battery being the high frequency crimes.
- As people tend to stay indoors during this month due to the cold weather, crimes like battery are reported more. mAmong the reported crimes during this month 14% are domestic in which battery tops the chart with 0.63 Million cases.

3. Seasonal Crime Trends:

Observation:

The pie chart shows the distributions of crime rate across different seasons. Summers shows the highest frequency of crime with about 27% of the crimes followed by spring and Autumn. While winters have the lowest frequency of crimes about 22%.

Insights:

Seasons influence human behavior thus having an impact on the number of crimes.

During the summer months, crime rates exhibit a strong positive correlation with longer days and increased outdoor activity. As more people spend time outside and public spaces become crowded, opportunities for crimes such as theft increase significantly. More than 1 million crimes were reported during summers, averaging around 563 incidents per day, yet only 26% of these cases resulted in an arrest, highlighting a gap in law enforcement outcomes. Geographically, the South Zone recorded the highest number of crimes, with street theft emerging as the predominant type of offense in this period.

During the winter months, crime rates exhibit a strong negative co-relation with shorter days, heavy snowfall making mobility difficult thus decreased outdoor activity. People spend more time inside their homes thus decreasing street crimes. But if we analyze Domestic crimes during winters we see a huge frequency of crimes like Battery, Assault etc happening at residences and apartments.

4. Weekday Crime Trends:

Observation:

The line chart shows the crime trend across weekdays. Friday records the highest number of crimes reaching almost 1.2 Million cases. Sunday shows a major dip in crime rate showing the lowest crime rate across the week. With slight fluctuations the crime rate is mostly stable from Monday to Thursday

Insights:

The spike on Fridays may be linked to increased social activity, nightlife, and crowded public spaces before weekends, creating more opportunities for thefts and street crimes. Sundays being the lowest suggests reduced mobility, closed businesses, and fewer crowded areas, lowering crime opportunities.But on Sundays Domestic cases are higher in frequency than any other weekday. The steady rise from Monday to midweek indicates routine urban activity (work, commuting, commerce) plays a strong role in driving crime patterns.

5. Hourly Crime Trends:

Observation:

12PM and 12AM shows the highest crime rate reaching almost 0.5 Million. 3-5PM window shows a gradual decrease in crime rate. from 6-9PM the crime rate shows an increase.

Insights:

Crime tends to peak during busy midday and relatively quiet late night hours, as evidenced by the spikes at 12 PM and 12 AM. High incidents between 6 and 9 PM may be a sign of crimes involving parties, nightlife, and transportation. The middle of the day (3–4 PM) has the least activity, suggesting that this is a comparatively safer time slot. By giving resources priority during noon, night, and midnight hours and deprioritizing low-incident intervals, this data can help the police plan patrol schedules.

5.9.3 CHICAGO CRIME REPORT- COVID ERA (2020-2021):

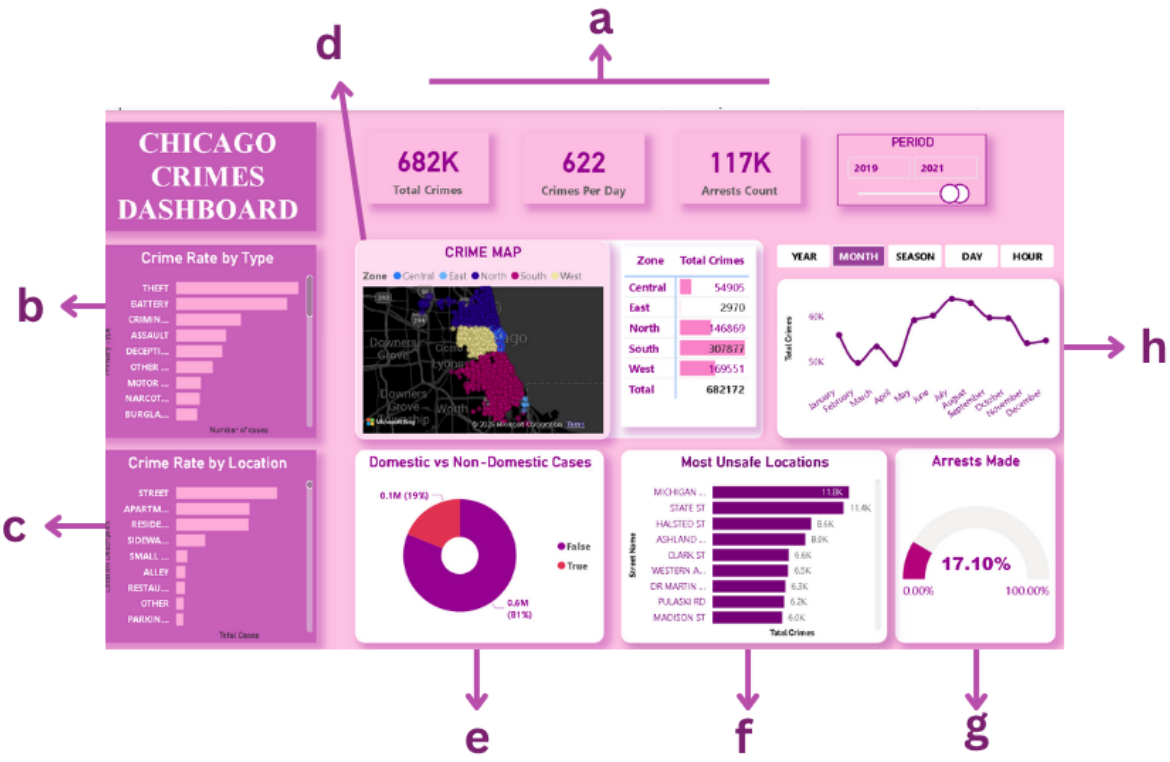

Figure 5.8: Covid Era Dashboard

Overview:

The COVID-19 pandemic led to noteworthy changes in social behavior, movement, and crime trends in Chicago. Analyzing data from 2020 to 2021 shows a clear decrease in overall crime compared to the years before the pandemic. Factors such as movement restrictions, the closing of public areas, and stay-at-home orders directly impacted both the types and frequency of criminal activities.

Key Findings:

a.Total Crime Trends: During the period from 2020 to 2021, a total of 421,000 crimes were reported, averaging 576 crimes per day. This represents a significant decline compared to previous years, with 2021 recording one of the lowest crime rates in the past two decades. The sudden drop in crime is strongly

associated with lockdowns and reduced public interactions, which limited opportunities for street crimes.

b. Crime by Location: Streets continue to be the most dangerous locations, with over 100,000 incidents reported, followed by apartments and residential areas. However, the overall number of crimes in public spaces has decreased due to limited movement. Domestic environments, such as residences and apartments, have shown a relatively higher proportion of crimes during this period, reflecting a shift from outdoor to indoor settings.

c. Crime By Type: The most prevalent offenses were theft and battery, both totaling 82,000 incidents. Other commonly reported crimes included criminal damage, with 50,000 cases, and assault, with 39,000. While public crimes like theft saw a decline compared to the years before the pandemic, there was an increase in attention toward domestic violence and assaults occurring within households.

d. Geographic Distribution: The South Zone had the most crimes about 193,000, followed by the West Zone around 103,000 and the North Zone about 92,000. Michigan Avenue, State Street, and Halsted Street were the most dangerous streets in the city.

e. Domestic Vs Non Domestic: 80% of crimes were non-domestic, while 20% were domestic in nature. This ratio indicates that despite restrictions, domestic violence and indoor crimes increased in significance, a trend reported globally during lock downs.

g. Arrests: Arrests were relatively low, with only 14.36% of reported cases resulting in arrests. This could reflect both reduced police activity during the pandemic and operational challenges in law enforcement.

h. Temporal Trends:

1.By Hour: Crimes are most frequent at midnight and midday, likely due to

nighttime gatherings and daily routines.

2.By Day: Fridays have the most crimes about 62,000, while Tuesdays have the least about 58,000.

3.By Month: Crime rates fell in April 2020 during peak lockdowns but gradually increased afterward, with a slight rise in mid-2021 as restrictions were relaxed.

Insight: The COVID-19pandemic caused a major drop in overall crime rates, especially in public spaces due to limited movement. However, domestic violence incidents increased as household tensions grew during lockdowns.

While safety improved in public areas, some urban hotspots, like Michigan Ave and State St, still reported high crime levels. The low arrest rates suggest there were gaps in law enforcement during the pandemic.

5.9.4 RECOMMENDATIONS FOR ENHANCING COMMUNITY SAFETY:

It's crucial to prioritize the safety and well-being of our communities. Here are some thoughtful recommen dations based on the crime trends observed during this time.

1. **Improve Safety in Public Spaces:** To make areas like Michigan Avenue, State Street, and Halsted Street safer, we should boost surveillance efforts. This could involve installing more CCTV cameras, improving street lighting, and increasing police patrols. Additionally, community policing programs can really make a difference. By fostering partnerships between residents and law enforcement, we can build trust and create environments that discourage street level crime. Utilizing predictive crime analytics can help us stay ahead of potential issues by pinpointing emerging hotspots and allocating resources where they are needed most.

2. **Strengthen Support for Domestic Violence Victims:** Expanding resources

for victims of domestic violence should be a priority. We need more helplines, shelters, and legal assistance to support those in distress. It's equally important to train first responders to handle these situations with sensitivity, ensuring that victims feel comfortable and safe when they report incidents. By raising awareness about reporting channels and the protections available for victims, we can encourage more people to seek help when they need it most.

3. **Enhance Law Enforcement Practices:** Our law enforcement agencies can benefit from adopting digital tools to improve evidence collection and case management. This would help increase the efficiency of investigations and improve the arrest-to-case ratio. Furthermore, enhancing coordination between police, prosecutors, and courts will facilitate quicker crime resolutions.

4. **Utilize Data for Better Decision-Making:** Continuing to use crime dashboards for real-time analysis of crime trends is essential. This approach will help us plan targeted interventions, such as increased patrols during peak crime times, like Friday nights or around midnight. Sharing aggregated crime data with researchers and community organizations can foster collaboration, leading to innovative safety initiatives that benefit everyone.

5. **Tackle the Root Causes of Crime:** Addressing the underlying social and economic factors that contribute to crime is vital. We should develop youth engagement programs, vocational training, and job opportunities in high-crime neighborhoods to help reduce theft and assault rates. Expanding access to mental health services is crucial as well. Investing in community development projects, particularly in the South and West Zones, which often see higher crime rates, is essential for fostering long-term safety and resilience.

6. **Create Crisis-Ready Crime Prevention Systems:** Establishing contingency plans for crime prevention during emergencies—whether due to pandemics, natural disasters, or civil unrest—is key to maintaining com-

munity safety. Promoting online reporting options will ensure that residents can report crimes without interruption, even during lockdowns. Lastly, it's important to provide law enforcement with training that balances public health considerations with their law enforcement responsibilities, preparing them for dual crisis management roles. By working together and implementing these strategies, we can create safer, more resilient communities that thrive even in challenging times.

5.9.5 Conclusion:

The dashboards provide a clear overview of the crimes and their impact, providing law enforcement agencies with the statistics and overview which can help them implement safety measures accordingly.

Chapter 6

CONCLUSION & FUTURE WORK

6.1 Conclusion

This study seeks to explore how open datasets, often unstructured and inconsistent, can systematically be transformed into usable decision-support systems by harnessing the data analysis capacity of Python and the data visualization capabilities of Power BI. In this case, three heterogeneous datasets—urban infrastructure (Water Main Breaks in Kitchener), student well-being (Student Depression Dataset), and public safety (Chicago Crime Dataset) were implemented as case studies.

The study showed that a hybrid approach that integrates data preprocessing, exploratory data analysis (EDA) and visualization may facilitate the link between raw data and actionable insights. The preprocessing stage in Python ensured data integrity by addressing issues such as missing values, duplicate observations, and inconsistent formats. Exploratory analysis such as descriptive statistics, correlation studies, skewness, kurtosis, and regression modeling provided insights into the domain. After this stage, Power BI dashboards allowed for interactive analysis of patterns and trends, improving the trustworthiness of results interpretation for decision-makers.

The results were able to identify a few important things. In the area of water infrastructure, leaks in pipes were mostly attributed to aging materials and seasonal climate stress - pointing to a need for ongoing maintenance and infrastructure replacement. In the educational area, the analysis showed that academic pressure, lack of sleep, and economic stress are significant predictors

of student suicide, it is then urgent to develop institutional supports. Lastly, crime data, while showing evidence of long-term reduction in incidents, continues to show persistent spatial and temporal hotspots that must be addressed in urban safety strategies.

Taken together, the research confirms that data driven decision-making is improved through statistical rigor supported by user-friendly, interactive visualizations. The combination of Python and Power BI, not only adds depth to the analysis of the data, but it provides a scalable and repeatable model for multi-domain applications.

6.2 Future Work

While the work accomplished its goals, there were still multiple paths for research and development extending beyond the work.

6.2.1 Predictive Analytics Integration

Subsequent research could go beyond descriptive and exploratory analysis to include predictive modeling using time-series forecasting, random forests, deep learning methods, etc. to gain a more precise sense of future projections.

6.2.2 Pipelines for Real Time Data

Additions of real-time data steaming using platforms such as Apache Kafka or cloud connectors, would make dashboards more responsive and quicken the decision-making process.

6.2.3 Scalability of Cloud Deployment

Deploying on a cloud environment such as Microsoft Azure, AWS, or Google Cloud would result in better scalable accessibility and collaborative use and make the system more suitable for organizations and larger scale.

6.2.4 Greater Interactivity and Access

Natural language processing (NLP)-based querying and AI-enhanced dashboards could increase access further, allowing non-technical users to obtain insights without substantial analysis skills.

6.2.5 Application to Other Areas

This work examined infrastructure, education, and public safety; applying the framework to fields such as health, environmental sustainability, and transportation would promote its generalizability and enhance public impact.

6.2.6 Ethical, Legal, and Privacy Implications

Future work should address ethical and legal issues inherent in the utilization of open data such as provisions for privacy, dealing with bias, and fairness in data-driven provisions. Such considerations become particularly important when analyzing data driven by sensitive domains like health or criminal justice.

6.3 Final Remarks

This thesis has presented empirical support for the proposition that combining the analytical power of Python with the ability of Power BI to visually represent analysis allows open datasets to be transformed into a decision-support system. The thesis's application of this proposition in a variety of applications illustrates the potential of data science and business intelligence tools to support evidence-based policies and strategies. Its main contribution is the thesis's demonstration that open data can become a valuable tool for improving organizational and community decision-making processes, if the data is organized in a systematic process and then effectively visualized.

Appendix A

HISTOGRAMS OF NUMERIC COLUMNS (DATASET 2 – STUDENT DEPRESSION ANALYSIS)

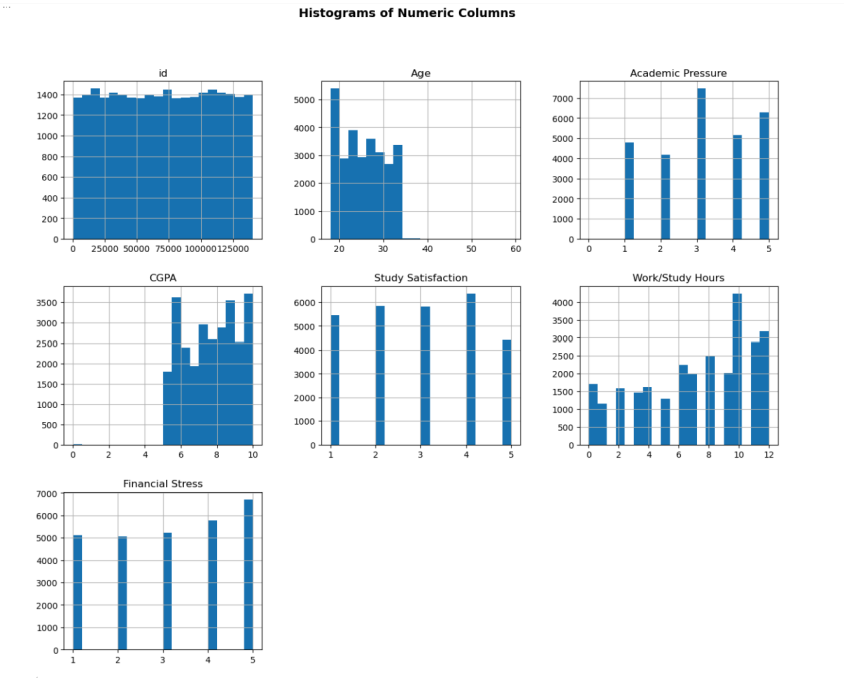

Figure A.1: Additional Histograms

The additional histograms in Appendix A show the distribution patterns of the remaining numeric variables in the dataset.

ID: The ID distribution appears uniform which confirms that the dataset is properly constructed with a uniform distribution of identifications across all records with no noticeable gaps or clustering.

Age: The age distribution, has a mild right skew, suggesting that most respondents' ages are concentrated in the younger age categories with fewer respondents in the higher age categories.

Academic Pressure: The histogram shows moderate clustering near higher levels of academic pressure, suggesting that many students are facing substantial academic demands.

CGPA: The CGPA distribution is symmetric and clustering in the mid-to-high range suggesting that a majority of students are performing adequate to good in their coursework.

Study Satisfaction: The responses regarding study satisfaction were evenly distributed, leading to variability in satisfaction with their academic experience.

Work/Study Hours: This variable reveals a mild right skew with most students reporting moderate working/studying hours and only a small proportion, dedicating a considerable amount of time.

Financial Stress: Financial stress- the distribution reveals that a considerable number of respondents have a moderate to high degree of financial stress, matching common trends revealed in student wellbeing literature.
In conclusion, these histograms provide a supplementary view of the key variables discussed in the previous chapter and serve as a general overview of the numerical characteristics of the dataset. Specifically, these supplemental figures demonstrate that there are not serious imbalances in the data, and generally highlight trends regarding the demographics, academic characteristics, and well-being measures of students.

Appendix B

HISTOGRAMS OF NUMERIC COLUMNS (DATASET 3 – CHICAGO CRIME ANALYSIS)

This appendix contains additional visualizations and supporting analysis referenced in Chapter 5 but not included in the main text for brevity.

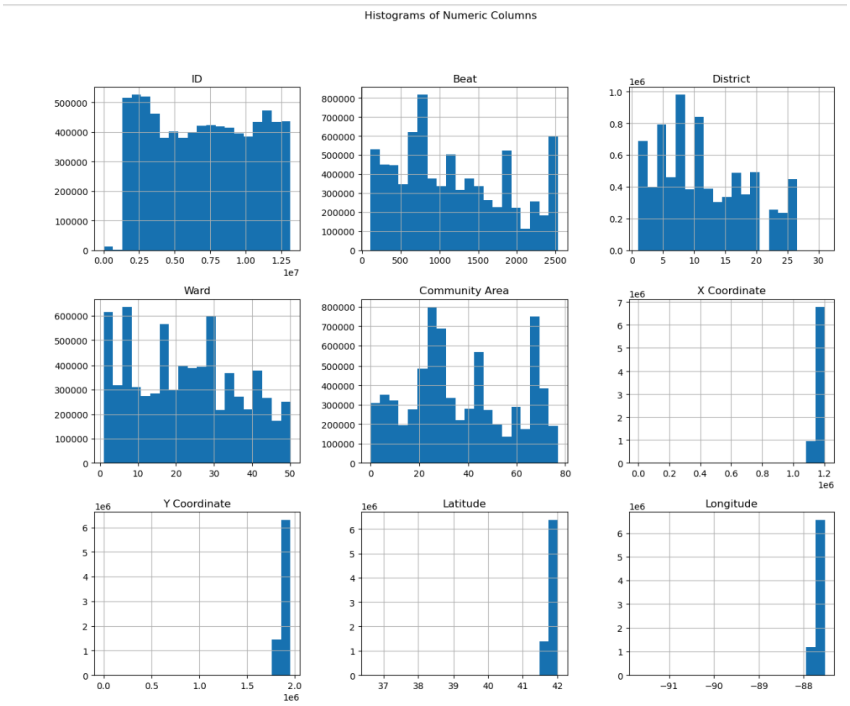

Figure B.1: Additional Histograms

As part of the analysis conducted in Dataset 3 – Chicago Crime Analysis (Section 5.9), these histograms illustrate the distribution of all numeric values from the dataset, such as ID, Beat, District, Ward, Community Area, X Coordinate, Y Coordinate, Latitude, and Longitude. They provide an illustration of the dispersion and frequency of each dispersion variable beyond what was statistically stated previously.

You can tell that ID, Beat, District, Ward, and Community Area display rela-

tively regular or mixed distributions of frequency, suggesting these are categorical labels and not continuous numeric values. Conversely, coordinate-based variables (X Coordinate, Y Coordinate, Latitude, and Longitude) have clustering, which suggests that the observations are densely concentrated in a single geographical area. Furthermore, these graphic displays reinforce the structural and spatial observations made in this analysis and serve as corroborating evidence for Chapter 5.

Bibliography

[1] C. O'Neil and R. Schutt, *Doing Data Science: Straight Talk from the Frontline*. O'Reilly Media, 2013.

[2] F. Provost and T. Fawcett, *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. O'Reilly Media, 2013.

[3] S. W. Brown, "Business intelligence and analytics: From big data to impact," *MIS Quarterly Executive*, vol. 16, no. 2, pp. 99–113, 2017.

[4] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[5] W. McKinney, *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. O'Reilly Media, 2017.

[6] J. VanderPlas, *Python Data Science Handbook: Essential Tools for Working with Data*. O'Reilly Media, 2016.

[7] A. Field, *Discovering Statistics Using IBM SPSS Statistics*. Sage Publications, 2017.

[8] J. W. Tukey, *Exploratory Data Analysis*. Addison-Wesley, 1977.

[9] R. L. Scheaffer, L. Mendenhall, and L. Ott, *Elementary Survey Sampling*. Cengage Learning, 2011.

[10] E. Tufte, *The Visual Display of Quantitative Information*. Graphics Press, 2001.

[11] Microsoft, “Power bi documentation,” https://learn.microsoft.com/en-us/power-bi/, 2023, accessed: 2025-09-12.

[12] Gartner, “Magic quadrant for analytics and business intelligence platforms,” Gartner Research, Tech. Rep., 2023.

[13] D. Few, *Information Dashboard Design: The Effective Visual Communication of Data*. Analytics Press, 2013.

[14] City of Kitchener, “Water main breaks dataset,” https://open-kitchenergis.opendata.arcgis.com/, 2025.

[15] City of Chicago, “Crimes - 2001 to present,” https://www.kaggle.com/datasets/adelanseur/crimes-2001-to-present-chicago, 2025.

[16] City of india, “Student depression analysis,” https://www.kaggle.com/datasets/hopesb/student-depression-dataset, 2025.

[17] Wikipedia. (2025) Climate of kitchener. Accessed: 2025-09-19. [Online]. Available: https://en.wikipedia.org/wiki/Kitchener,_Ontario

[18] C. Results. (2025) Kitchener on - average monthly snowfall and temperature. Accessed: 2025-09-19. [Online]. Available: https://www.currentresults.com/Weather/Canada/Ontario/Places/kitchener.php

[19] W. Atlas. (2025) Kitchener, canada - climate data. Accessed: 2025-09-19. [Online]. Available: https://www.weather-atlas.com/en/canada/kitchener-climate

[20] C. Kitchener. (2025) City news on extreme cold and water main breaks. Accessed: 2025-09-19. [Online]. Available: https://kitchener.citynews.ca/

[21] Homes and Gardens. (2025) How to prevent frozen pipes in winter. Accessed: 2025-09-19. [Online]. Available: https://www.homesandgardens.com/

[22] R. Simple. (2025) Tips to keep pipes from freezing. Accessed: 2025-09-19. [Online]. Available: https://www.realsimple.com/