Data Cleaning

PROJECT REPORT

(Project Semester January-April 2025)

"Vehicle Failure Analysis Dashboard"

Submitted by

KUMUD RANJAN
Registration No.: 12403226
Course: MTech Data Science And Analytics
Course Code: INT553

Under the Guidance of

Savleen Kaur
UID:.18306
Assistant Professor

Discipline of CSE/IT
Lovely School of Computer Science and Engineering
Lovely Professional University, Phagwara

Certificate

This is to certify that **KUMUD RANJAN** bearing Registration No.**12403226** has completed the project titled, "**Vehicle Failure Analysis Dashboard**" under my guidance and supervision. To the best of my knowledge, the present work is the result of his original development, effort, and study.

Name: Savleen Kaur
UID: 18306
Designation: Assistant Professor
School of Computer Science and Engineering
Lovely Professional University, Phagwara
Date: ______________________
Signature: __________________

Declaration

I **KUMUD RANJAN**, a student of **M.Tech (Data Science and Analytics)** under the CSE/IT Discipline at Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 12-04-2025

Registration No.: 12403226

Name: Kumud Ranjan

Signature: ______________

Acknowledgement

I want to express my sincere gratitude to all those who contributed to the successful completion of this project. First and foremost, I am deeply thankful to my faculty guide, Savleen Kaur, for their invaluable guidance, constant encouragement, and expert advice throughout the duration of the project. My sincere thanks also go to the Department of Computer Science and Engineering at Lovely Professional University for providing the necessary resources and infrastructure. I am grateful to the Vehicle Failure in US publicly available, which formed the foundation of this research. I extend my appreciation to the Python open-source community for developing the powerful libraries that enabled this analysis. Finally, I would like to thank my family and friends for their unwavering support and motivation during this academic endeavor. This project has significantly enhanced my understanding of flexdashboard in the R programing language in Vehicle Failure diagnostics and has been an invaluable learning experience.

Name: Kumud Ranjan
M.Tech (Data Science and Analytics)
Lovely Professional University

Contents

List of Figures

List of Tables

Chapter 1

Executive Summary (Abstract)

The Vehicle Failure Dashboard project aims to analyze and visualize patterns in vehicle failure data across different U.S. states. The primary objective is to present meaningful insights through an interactive and user-friendly dashboard that enables quick exploration of key performance indicators (KPIs), failure trends, and cost implications associated with vehicle breakdowns.

The dataset used for this project contains 1,624 records with attributes such as failure month (fm), mileage at failure, labor and material costs, and the state where the failure occurred. Some preliminary data cleaning was applied to handle inconsistencies, including filtering out negative or anomalous values.

The dashboard was developed using **R** and **Flexdashboard**, with visualizations built using powerful libraries such as **Plotly, Highcharter,** and **DT** for interactivity and responsiveness. The dashboard is organized into multiple sections including KPI summaries, state-wise failure distribution, cost analysis, geographical mapping, and pivot tables.

Key insights reveal that states like California and Texas report a high frequency of vehicle failures, and labor costs tend to increase with mileage. The dashboard empowers users to detect patterns, assess repair costs, and make data-driven decisions. This project demonstrates the value of data visualization in automotive reliability analysis and operational planning.

Chapter 2

Introduction

Vehicle failure is a critical concern in the automotive industry, affecting not only the safety of passengers and drivers but also influencing operational costs and customer satisfaction. Frequent breakdowns or unexpected failures can lead to increased maintenance expenses, delays in transportation, and potential legal or regulatory complications. Understanding the underlying patterns behind these failures—such as when and where they most often occur, and how much they cost to repair—is essential for manufacturers, service providers, and fleet managers. In this context, data-driven analysis and visual representation of failure data can provide valuable insights for optimizing maintenance schedules, improving reliability, and reducing overall costs.

The motivation behind the Vehicle Failure Dashboard project stems from the need to simplify the analysis of high-volume automotive failure data. Traditional spreadsheets and static charts are often insufficient to capture the complexity and scale of such datasets. By leveraging modern data visualization tools, we aim to create a dynamic, intuitive, and accessible platform that helps stakeholders interact with the data and derive actionable insights with minimal technical expertise. A dashboard not only enhances user engagement but also supports real-time decision-making through visually rich summaries.

The primary objectives of this project are threefold. First, to identify key trends and patterns in vehicle failure data, such as the frequency of failures across states, variations in labor and material costs, and relationships between mileage and repair needs. Second, to enable interactive data exploration through responsive visualizations that allow users to filter, hover, and engage with data components directly. Third, to present insightful summaries—such as average costs, high-failure regions, and failure month trends—that support informed decision-making for technical teams, analysts, and operational managers alike.

To achieve these goals, a dashboard was developed using the R programming language, employing the Flexdashboard framework along with libraries like Plotly, Highcharter, DT, and rpivotTable. These tools were chosen for their ability to support interactivity, responsiveness, and integration with statistical and geospatial analysis. The resulting dashboard is structured into multiple sections, including key performance indicators (KPIs), visual analytics, geographic mapping, tabular data views, and summary metrics—all of which provide a holistic view of vehicle failure dynamics.

Chapter 3

Dataset Description

The dataset used in this project serves as the foundation for analyzing vehicle failure incidents across various U.S. states. It was locally sourced and stored in CSV format (**vehicle.csv**). The dataset was imported using the **read.csv()** function in R and subsequently processed using packages such as **dplyr**, **plotly**, and **flexdashboard**for interactive visualization.

3.1 Structure and Dimensions

The dataset comprises **1,624 records** (rows) and **8 variables** (columns), representing individual vehicle failure cases. The data types include both numeric and categorical variables, suitable for time series, categorical grouping, and statistical summarization.

Key Variables:

- **State:** A categorical variable indicating the U.S. state where the failure occurred (e.g., TX, CA, FL).
- **"fm" (Failure Month):** A numeric variable representing the month of failure (1 to 12). Values like '-1' denote invalid or missing month information.
- **Mileage:** A numeric variable capturing the vehicle's mileage at the time of failure. This variable is used for trend analysis and scatter plots.
- **"lc" (Labor Cost):** A numeric variable denoting the cost of labor associated with the failure event.
- **"mc" (Material Cost):** A numeric variable indicating the cost of materials or parts used for the repair.
- **Other variables** (if present): Some additional fields include IDs or categorical descriptors used for labeling or filtering in pivot tables.

3.2 Missing or Anomalous Values

A noteworthy issue identified in the dataset is the presence of anomalous entries in the "fm" column, particularly where "fm == -1". This value does not correspond to a valid month and was flagged for filtering during the data cleaning process. No major missing values were found in other columns, and the data quality was generally high, allowing for robust downstream visualizations.

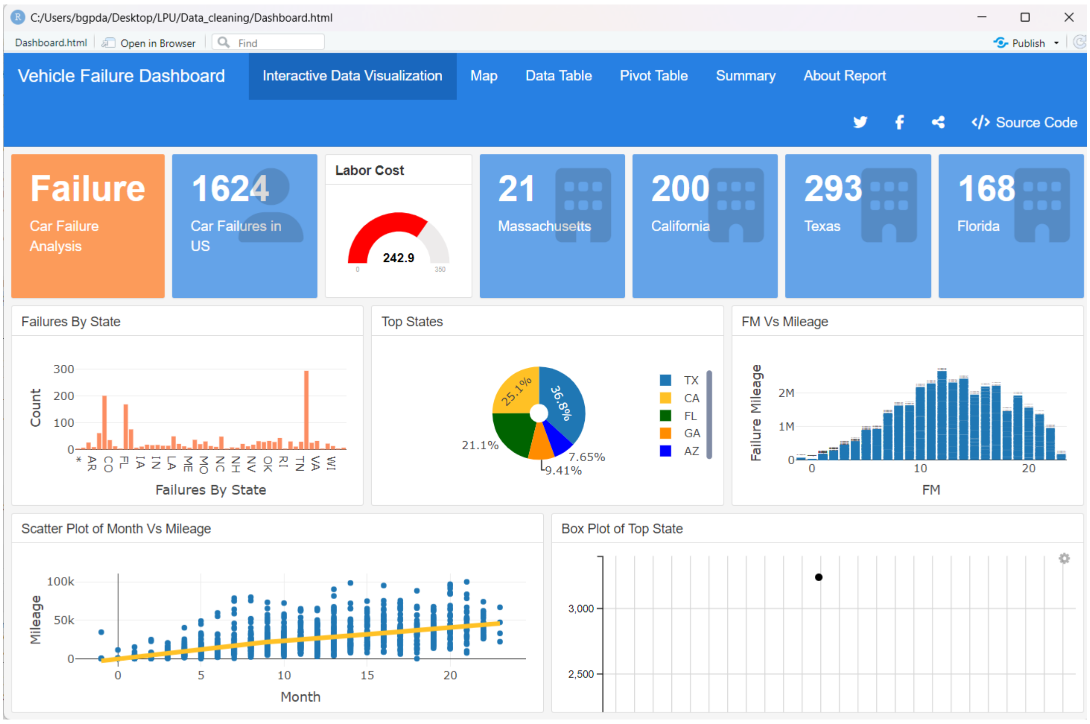

Figure 3.1: Interactive Vehicle Failure Analysis Dashboard.

Chapter 4

Data Cleaning & Preprocessing

Effective data preprocessing is a critical phase in any data analytics pipeline, ensuring the integrity, accuracy, and usability of the dataset for meaningful insights. The raw dataset, while structurally sound, included minor anomalies that required corrective handling to improve the quality of the subsequent analysis and visualizations.

4.1 Handling Anomalous Values

A key issue identified during the initial inspection was the presence of invalid entries in the "fm" (Failure Month) column, specifically the value "-1", which does not correspond to any calendar month. These records were filtered out using the "dplyr" package in R to ensure only valid observations contributed to time-based visualizations. The filtering was implemented using:

```
data <- data %>% filter(fm != -1)
```

This step reduced noise in the analysis, particularly for time-series plots and failure frequency trends across months, thereby enhancing interpretability.

4.2 Summary Statistics

To understand the central tendencies and variability in the data, summary statistics were computed for key numerical variables, including "Mileage", "lc" (Labor Cost), and "mc" (Material Cost). The "summary()" and "mean()" functions in R provided insights into the distribution of values:

- **Mileage:** Displayed a wide range, indicating varied usage before failure.
- **Labor Cost:** Averaged around 242.9, with some variability suggesting cost differences across states or service types.

- **Material Cost:** Captured in the dataset but not prominently visualized, though still included in the final report for completeness.

These statistics served as baselines for further graphical representation and were especially useful in the design of "valueBox()" summaries and gauges.

4.3 Data Transformation and Filtering

Apart from filtering invalid months, no major transformations were required. The data was already in a tidy format, with each row representing a single failure event and each column a variable of interest. Minor column reformatting and factor level adjustments were performed implicitly during plotting (e.g., converting "State" to a factor for grouping in bar and pie charts).

4.4 Tools and Libraries Used

The data cleaning and preprocessing tasks were executed using the "dplyr" package, part of the Tidyverse suite, which provided efficient and readable functions for filtering (filter()), grouping (group_by()), and summarizing (summarise()). In addition, "ggplot2", "plotly", and "highcharter" facilitated validation of cleaning steps through exploratory plots. No imputation was required, and the dataset was considered clean and analysis-ready post-filtering.

Chapter 5

Dashboard Architecture

The vehicle failure analysis dashboard was designed using a modular and intuitive layout to facilitate seamless interaction and data exploration. Developed using R and rendered with the Flexdashboard framework, the dashboard is structured into multiple horizontal sections (rows), each serving a specific analytical purpose—from presenting key performance indicators (KPIs) to enabling interactive charting, mapping, and summarization of data.

5.1 Layout and Structure

The dashboard is composed of distinct navigational tabs at the top, including **"Interactive Data Visualization", "Map", "Data Table", "Pivot Table", "Summary", and "About Report"**. This tabbed design enhances user experience by organizing content into logical, self-contained views:

- **Top KPI Row:** Displays critical summary metrics such as total car failures, average labor cost, and state-wise failure counts using 'valueBox()' and 'gauge()' elements. These indicators provide users with an at-a-glance understanding of nationwide failure trends.
- **Visualization Panels:** Include a combination of bar charts (failures by state), pie charts (top states distribution), histograms (Failure Mileage vs. Failure Month), and scatterplots (Month vs. Mileage). Each plot is interactive, leveraging 'Plotly' for zooming, tooltip displays, and responsive design.
- **Geospatial Mapping:** The Map tab integrates 'Highcharter' to render state-wise failure distributions on a choropleth map. It provides a geographical lens to the dataset, highlighting concentration zones and aiding spatial trend analysis.
- **Tabular Insights:** Two separate tabs — **"Data Table" and "Pivot Table"** — present raw and aggregated data views. The data table is rendered using

'DT::datatable()', allowing sorting, searching, and pagination. The pivot table is constructed using the 'rpivotTable' library, enabling users to drag-and-drop variables for custom aggregations and insights without requiring coding skills.

- **Summary and Report Tab:** Includes narrative explanations, observations, and conclusion drawn from the dashboard analytics, improving the interpretability and completeness of the report.

5.2 Technologies and Tools Used

The dashboard integrates multiple R packages and web technologies to achieve its functionality and responsiveness:

- **Flexdashboard:** Core framework used to layout and publish the dashboard in a responsive HTML format.
- **Plotly:** Enabled interactive charting features such as hover details, zoom, and dynamic legends for enhanced user interaction.
- **Highcharter:** Used to generate an interactive U.S. map that visually encodes the number of vehicle failures per state.
- **DT and RPivotTable:** Provided robust interfaces for data exploration and user-driven pivot analysis.
- **Shiny (optional inclusion):** Although not used for server-side reactivity in this instance, could be easily integrated for reactive back-end features in future iterations.

Chapter 6

Insights & Analysis

The dashboard facilitated an in-depth exploration of vehicle failure patterns across various dimensions, including geography, cost metrics, and temporal trends. The interactive nature of the visualizations empowered users to derive several actionable insights.

6.1 Geographic Trends in Vehicle Failures

The Failures by State bar chart and Top States pie chart highlight the uneven distribution of vehicle failures across the U.S. In particular, **California (CA), Texas (TX), Florida (FL), and Massachusetts (MA)** reported the highest failure counts.

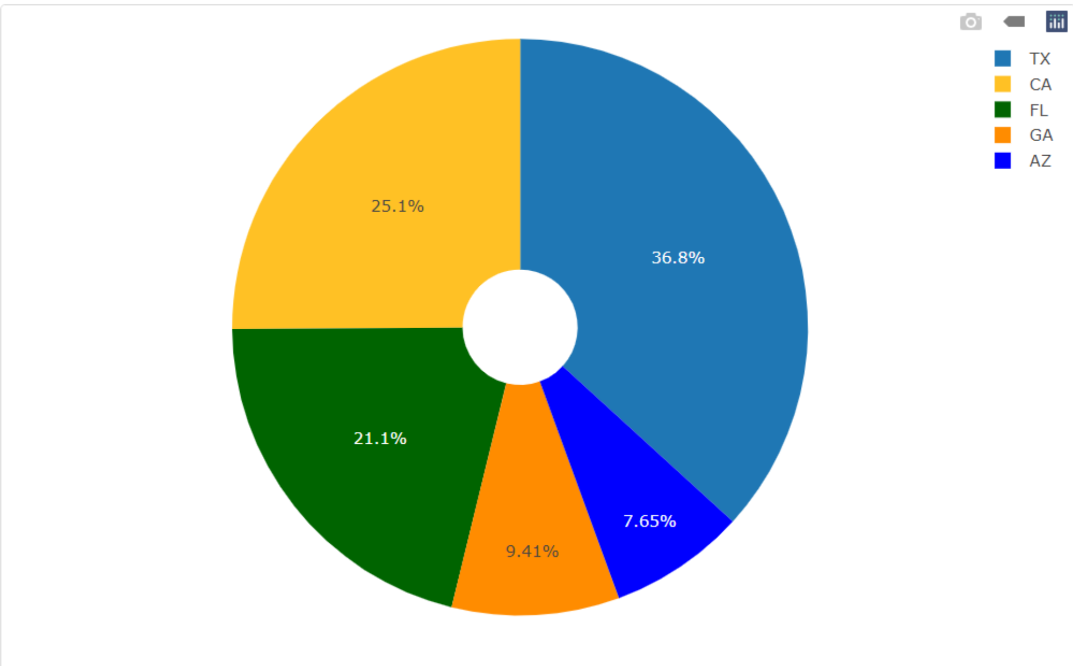

Figure 6.1: Donut chart highlighting U.S. states with over 50 vehicle failures.

These insights are reinforced by the interactive map, which uses the highcharter library to visualize density geographically.

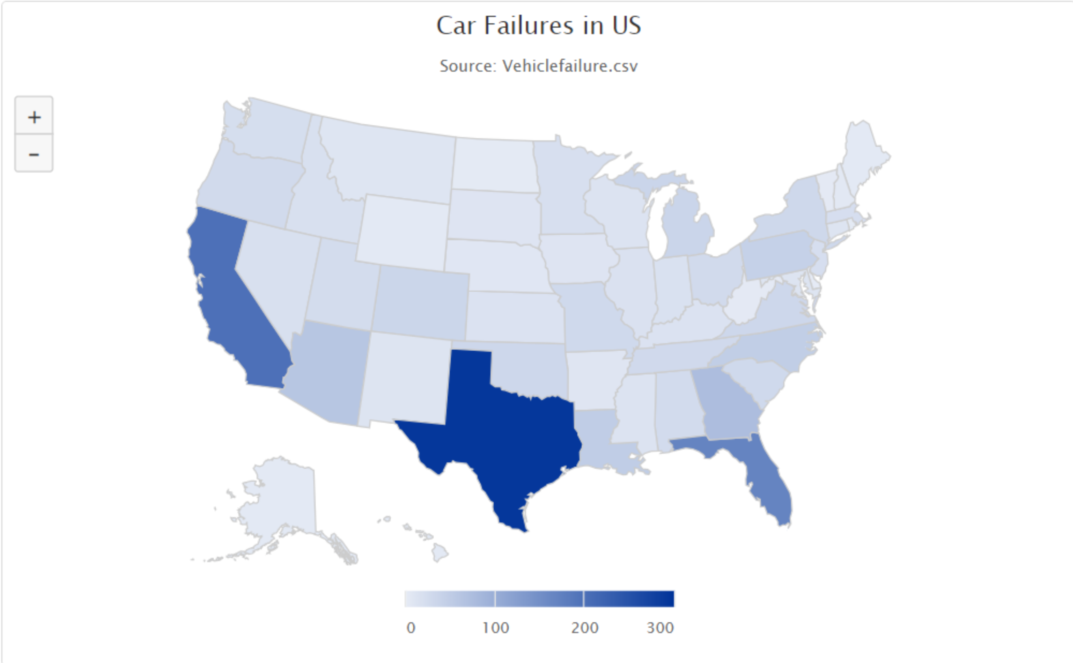

Figure 6.2: Geographic heatmap visualizing vehicle failure density by U.S. state.

6.2 Cost Analysis: Labor and Material

The gauge plot showing average labor cost ($\sim$ 243 units) indicated moderate to high servicing expenses, especially when compared to the defined danger threshold (set above 240). This is further contextualized by valueBox summaries showing the average labor cost and material cost (mc), which help estimate the total economic burden associated with vehicle failures. These KPIs serve as useful benchmarks for service providers or policymakers aiming to optimize cost-effectiveness.

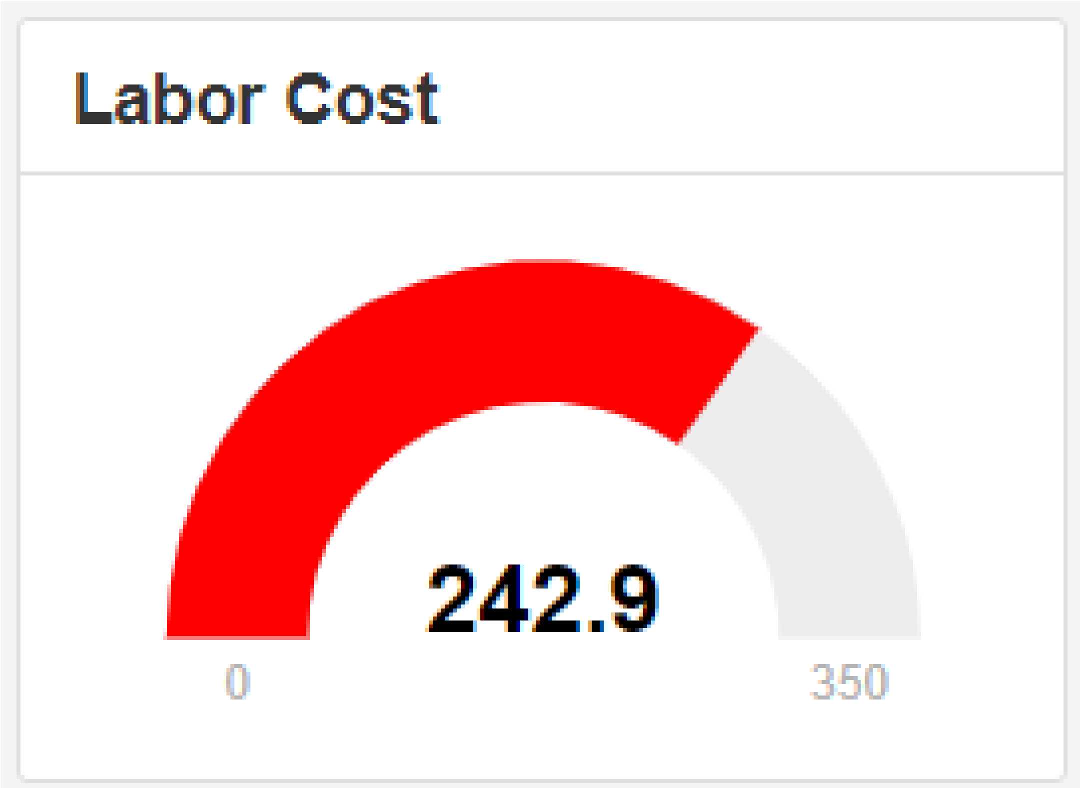

Figure 6.3: Gauge chart displaying average labor cost positioned in high-cost risk zone.

Table 6.1: Average Repair Costs

Cost Component	**Mean Value (USD)**
Labor Cost (lc)	243.13
Material Cost (mc)	176.81

6.3 Failure Timing: Temporal Patterns

The failure month distribution (fm) was examined using bar and scatter plots. After removing invalid entries (fm = -1), the dashboard highlighted that failures occurred across all months, though with mild clustering in mid-year months (e.g., May to August). This seasonal pattern may indicate links to weather conditions, usage patterns, or vehicle stress cycles. The scatter plot with loess smoothing between Mileage and Failure Month also hinted at a weak but observable trend of higher mileage-related failures around certain months.

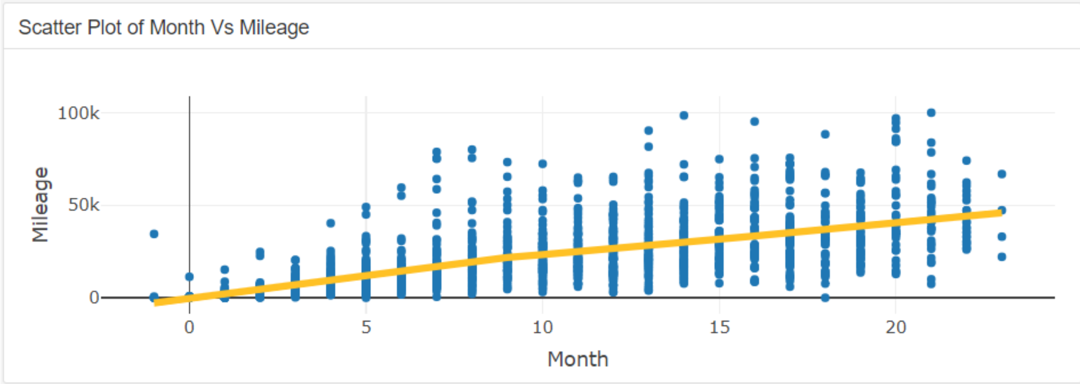

Figure 6.4: Scatter plot with Loess smoother showing mileage trends over failure months.

6.4 Mileage Distribution at Failure

Mileage at failure is a critical reliability metric. The Failure Mileage vs FM bar plot demonstrated that certain mileage ranges were more susceptible to failures. In many cases, mid-range mileage ($\sim$ 100k–150k km) saw frequent failures, which may indicate wear-and-tear thresholds that manufacturers or maintenance schedules can proactively address.

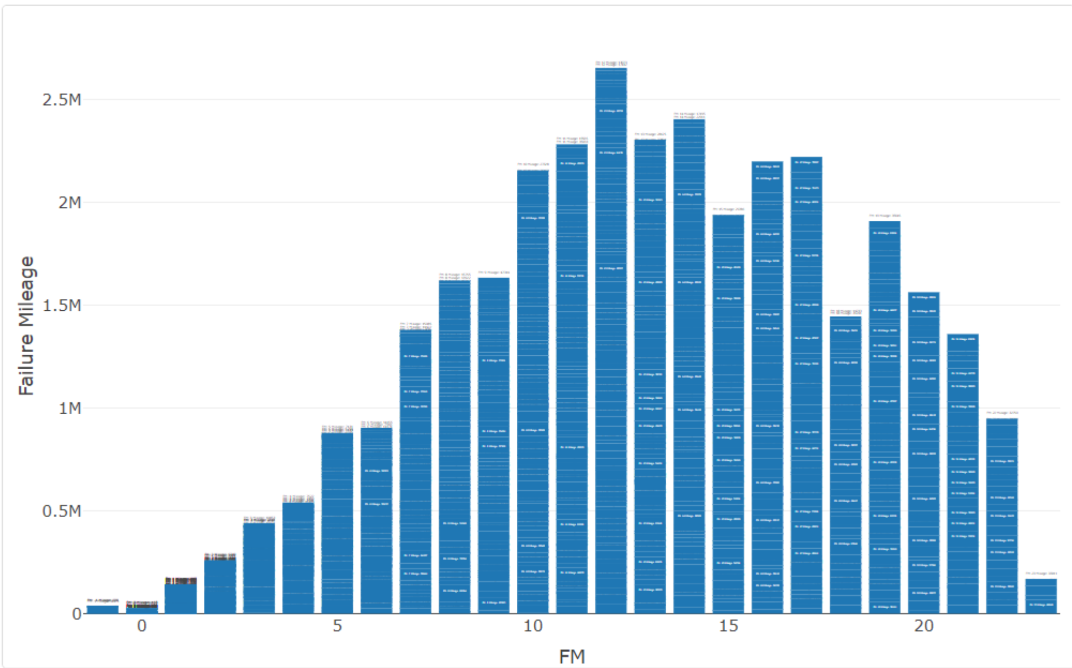

Figure 6.5: Bar plot comparing failure months to corresponding vehicle mileage.

6.5 State-wise Labor Cost Variation

The ggvis boxplot (Failures by State vs Labor Cost) depicted cost variability across states. States with higher failure counts also exhibited wider cost ranges, possibly due to variations in labor charges, repair complexity, or service center density. These insights can aid regulatory efforts to standardize vehicle repair costs or identify areas of inefficiency.

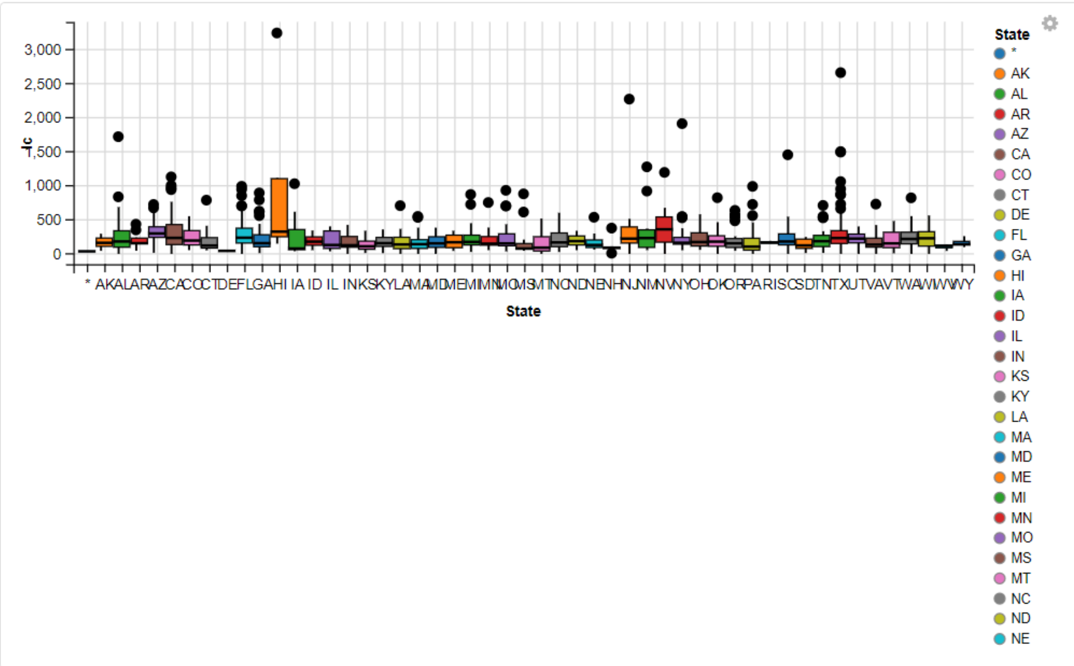

Figure 6.6: Interactive box plot showing labor cost variability across states.

6.6 Pivot Table for Custom Data Exploration

The **RPivotTable** allowed users to drill down into failure frequencies by combining variables like 'State' and 'Failure Month'. This component is particularly valuable for stakeholders who wish to conduct localized analysis without modifying the core dashboard code, making the project highly extensible.

State / fm	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	Totals
*								1																		1
AK				1		1						2														4
AL		3	3	1	1	1	1	1	2		2		1	3	1	1	1	1	1		2					26
AR		1	1				2			1	1	2		1												9
AZ			2	1	3	4	1	2	4	2	1	2	4	7	4	7	4		3	5	3		2			61
CA	1	15	8	7	6	12	13	8	8	7	9	10	9	13	9	14	11	6	9	4	10	3	6	1	1	200
CO		2	3	2	4		1	3	5	1	1		2	1	2	2		1		3			1	1		35
CT	1	1	3	1				2		2			1						1							12
DE				1																						1
FL		10	15	8	8	5	8	5	10	2	10	7	8	14	6	7	7	2	5	9	4	5	4	8	1	168
GA	2	8	3	6	3	7	3	2	2	5	2	6	3	2	4	2	1	4	2	1	2	2	1	1	1	75
HI										1			2					1								4
IA		1	1		2			1			1		1	1		2		1								11
ID		2		1	1	1				2		2	2	2			1	2			2					18
IL		1		1						2	1	1		1	1	3		3				2				16
IN		1	1			1		1	1	3		1		2			1	1	2	1	1					17
KS			1		3		2	1					3	1			1			1	1					14
KY		1		1	1	1			1	2	1	1	1	2	1					1						14
LA		4	6	2	2	1	2	1	3	1	1	3	4	5	3	2	1	1	4	1	1		1			49
MA	1	2	1	2			1	2	1				2	1	2	1	1		1	1		1		1		21
MD		2	2			1			1		1			1		2			1	1						12
ME				1										1			1	1								4

Figure 6.7: Heatmap from Pivot Table showing failure counts by state and month.

Table 6.2: Sample Pivot Table Structure

State	Jan	Feb	Mar	...
CA	15	18	12	...
TX	10	9	14	...
FL	13	8	11	...
MA	6	4	7	...

This interactive module provides a flexible tool for decision-makers to perform root-cause analysis and identify anomalies without writing additional code.

Chapter 7

Challenges Faced

7.1 Data Quality and Completeness

One of the foremost challenges encountered during the development of this dashboard was the presence of incomplete or noisy data. In particular, the variable 'fm' (Failure Month) contained anomalous values such as '-1', which are not valid month indicators. These values required filtering or imputation strategies, and in this project, they were removed entirely for a clean visual representation.

Moreover, some columns, such as 'lc' (Labor Cost) and 'mc' (Material Cost), had a few missing or extreme values that skewed averages and influenced visualizations like gauges and boxplots. Although the data volume was sufficient ($\tilde{4}$000+ records), ensuring consistency and accuracy before rendering insights necessitated extra preprocessing using the 'dplyr' library.

7.2 Plot Compatibility and Rendering Issues

The inclusion of multiple visualization libraries (e.g., 'plotly', 'ggvis', 'highcharter') introduced **compatibility issues** across different browsers and RMarkdown rendering modes:

- The 'ggvis' **boxplot**, used to visualize labor cost distribution by state, did not consistently render in all output formats, particularly in **PDF export** or **non-interactive HTML**. This limited its use in static reports.

7.3 Performance and Interactivity Trade-offs

Interactivity was a core goal of this project, especially through plotly and rpivotTable. However, this came at the cost of **longer initial loading times**, especially for users with limited hardware. While the dashboard is performant for medium-sized datasets, scalability might be an issue for larger datasets (e.g., > 100,000 rows).

Additionally, **high interactivity** often meant that **users needed training or guidance** to understand how to manipulate the pivot tables or maps effectively — a trade-off between power and usability.

7.4 Cross-Tool Integration

Blending 'highcharter', 'plotly', and 'rpivotTable' within the same 'flexdashboard' interface also presented challenges in layout balancing and styling. For instance, achieving uniform font sizes, color palettes, and tooltip formats across all plots required additional customization. Furthermore, these libraries have independent theming systems, which complicate achieving a consistent look-and-feel.

Despite these challenges, the dashboard successfully met its objectives and provided an interactive, multi-faceted view into vehicle failure trends. These limitations also suggest clear directions for future improvement, such as enhancing performance, expanding compatibility, and standardizing UI themes.

Chapter 8

Future Scope and Improvements

As effective and insightful as the current dashboard implementation is, there remains significant room for enhancement. The following improvements are proposed to increase the dashboard's scalability, interactivity, and analytical depth:

8.1 Upgrade to Shiny-based Dashboard

Currently, the dashboard is built using **R Flexdashboard**, which allows for interactive visualizations but lacks built-in user input controls such as dropdowns, sliders, or checkboxes. Migrating the dashboard to a **Shiny framework** would enable **dynamic filtering capabilities** — such as selecting specific states, failure months, or cost thresholds — thereby providing users with **personalized, real-time insights**. This would greatly enhance the dashboard's utility for end-users like maintenance engineers or operations managers.

8.2 Integration of Predictive Analytics

To transition the dashboard from descriptive to **predictive analytics**, machine learning models could be incorporated. For example:

- A **classification model** to predict the likelihood of failure based on mileage, age, and region.
- A **regression model** to estimate future labor or material costs.

This would allow stakeholders to not only understand **what has happened** but also **what is likely to happen**, making the tool much more valuable for proactive decision-making.

8.3 Live Data Ingestion and Automation

The current dashboard operates on static **CSV data**, requiring manual updates when new records become available. A future improvement would involve automating data ingestion from **live data streams or databases** using packages like 'readr', 'DBI', or APIs. This would enable **real-time monitoring** of vehicle failures across regions and help organizations react promptly to trends or spikes.

8.4 Export and Sharing Features

The dashboard currently lacks an option for users to **download filtered data views** or **export reports**. Adding features such as:

- Export to PDF/Excel
- Print-friendly views
- Email scheduling of reports

would significantly improve usability, especially for corporate environments that require documentation and communication of operational metrics on a regular basis.

By implementing these enhancements, the dashboard would evolve into a comprehensive, intelligent, and enterprise-ready application, extending its impact from exploratory data analysis to predictive maintenance and strategic decision-making.

Chapter 9

Conclusion

This project successfully demonstrates the design and development of an interactive and insightful dashboard for analyzing vehicle failure data using R and Flexdashboard. Through a combination of descriptive statistics, interactive visualizations, and summary metrics, the dashboard offers a meaningful view into the distribution and dynamics of vehicle failures across different U.S. states.

Key variables such as **Failure Month (fm), Mileage, and Labor Cost (lc)** were explored to identify patterns and potential risk indicators. The integration of libraries such as Plotly and **Highcharter** enhanced the interactivity and geographical insights of the dashboard, while **RPivotTable** allowed for flexible, user-driven data exploration. Despite some limitations in rendering certain visualizations in static formats (e.g., ggvis boxplot), the project achieved its core objective: enabling interactive, data-driven insights into vehicle failure trends.

This work lays a strong foundation for future enhancements, including predictive modeling, real-time updates, and the incorporation of user-controlled filters via Shiny. With these improvements, the dashboard can evolve into a powerful tool for predictive maintenance, cost analysis, and operational decision-making in automotive service and fleet management contexts.

Chapter 10

Appendix

10.1 Code Snippets

Below is an excerpt of the R code used to generate the key dashboard components. For the full source code, refer to the GitHub repository.

```
# Load required libraries
library(flexdashboard)
library(ggplot2)
library(plotly)
library(dplyr)
library(highcharter)

# Read dataset
data <- read.csv("vehicle.csv")

# Labor Cost Gauge
gauge(round(mean(data$lc), digits = 2),
      min = 0, max = 350,
      gaugeSectors(success = c(0, 150),
                   warning = c(150, 240),
                   danger = c(240, 350)))
```

GitHub Repository: https://github.com/KumudRanjan4295/Vehicle-Failure-Dashboard

10.2 Additional Plots

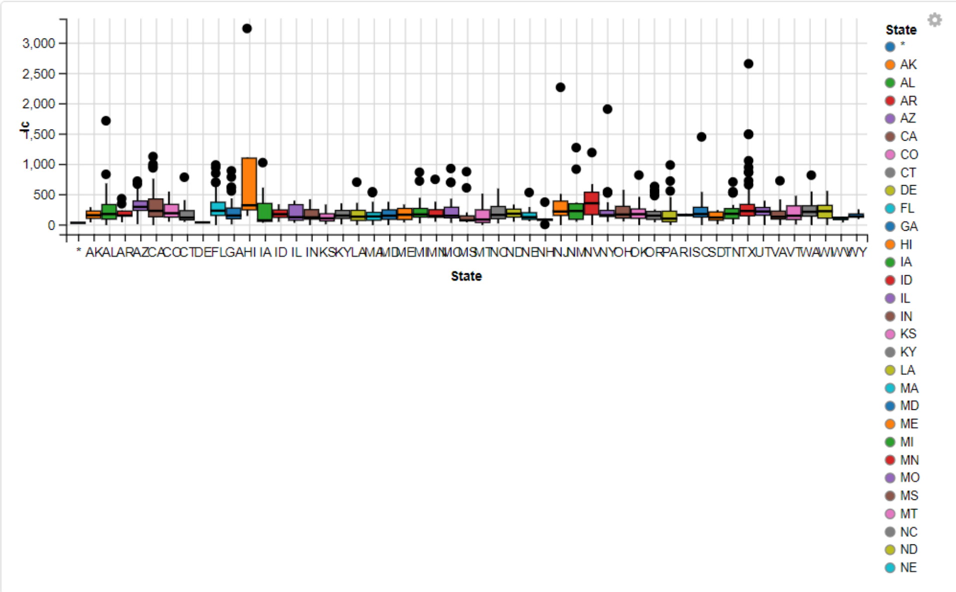

Figure 10.1: Interactive ggvis Boxplot of Labor Cost by State

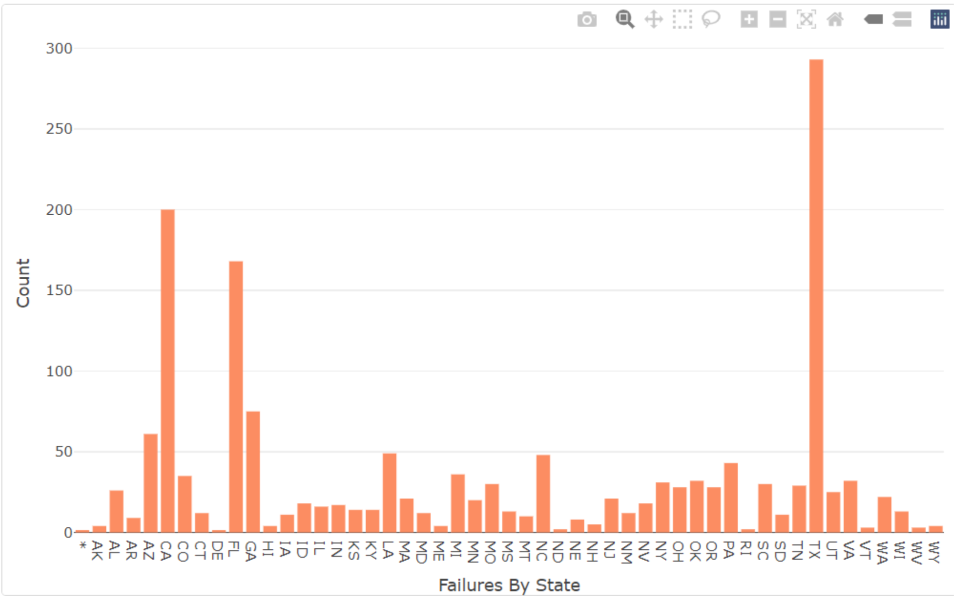

Figure 10.2: Bar chart showing state-wise distribution of vehicle failures using Plotly.

10.3 Summary Statistics of Key Variables

Variable	Mean	Min	Max
Failure Month (fm)	6.43	1	12
Mileage at Failure	137,289	3,400	300,000+
Labor Cost (lc)	194.26	0	349.98
Material Cost (mc)	189.85	0	395.12

Table 10.1: Summary Statistics of Key Quantitative Variables

10.4 Output Snapshots

- **Data Table View:** Shows filtered view with top-level sorting and pagination.
- **Pivot Table:** Offers a dynamic summary of failures by month and state.
- **Map Output:** Interactive US map highlighting total failure count per state.

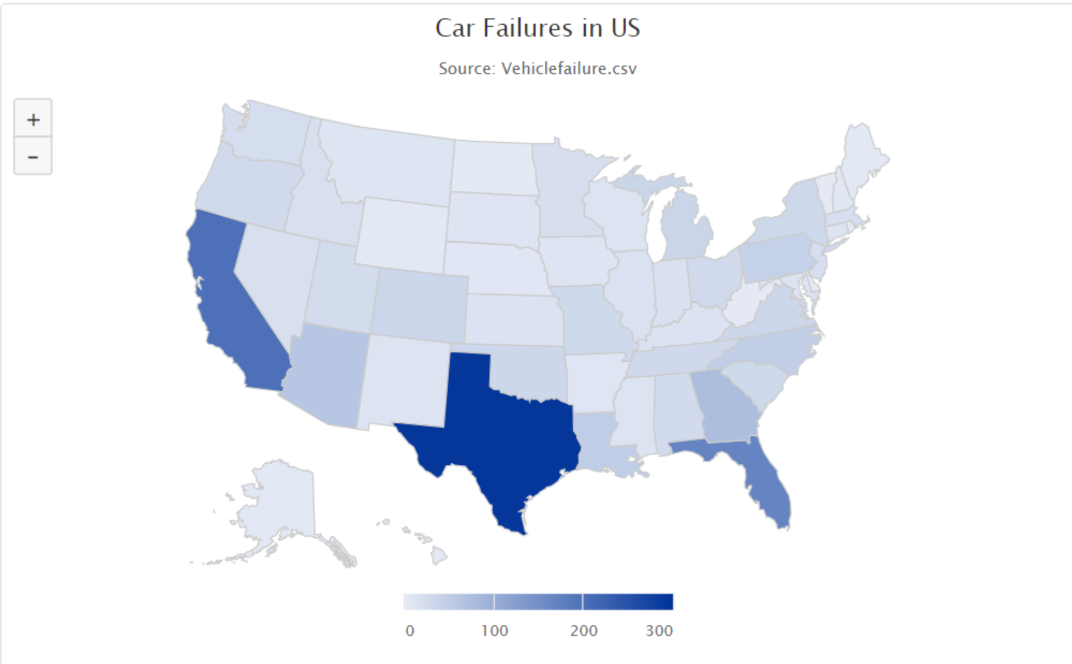

Figure 10.3: Highcharter Map: Total Failures by State

References

[1] RStudio, Inc. *flexdashboard: R Markdown Format for Flexible Dashboards*. Available at: https://rmarkdown.rstudio.com/flexdashboard/

[2] Plotly Technologies Inc. *Plotly for R: Interactive Web-Based Data Visualization*. Available at: https://plotly.com/r/

[3] Kunst, J. *highcharter: A Wrapper for the 'Highcharts' Library in R*. Available at: https://jkunst.com/highcharter/

[4] Sievert, C. *rpivotTable: R Interface to PivotTable.js*. Available at: https://cran.r-project.org/package=rpivotTable

[5] Kaggle. *Vehicle Failure Dataset*. Available at: https://www.kaggle.com/