PROJECT REPORT

Title

Loan Default Risk Analysis and Credit Risk Assessment Using Data Analytics

Sector

Banking and Financial Services (Loan & Credit Risk Analytics)

Team details

Team G-7

Swati Ghosh Enrollment: 2401010471

Raghav Kaushal Enrollment: 2401010366

Divyanshu Raj Enrollment: 2401010160

Farhana Pervin Enrollment: 2401010164

Shreyansh Agrawal Enrollment: 2401010447

Aditya Rao Enrollment: 2401010036

Institute Name:

Rishihood University, Newton School of Technology

Faculty: Aayushi vashishth

Executive Summary

Problem

Loan default represents a major risk to profitability and portfolio stability in the banking and financial services sector. As borrower profiles become more diverse and competitive pressures increase, financial institutions face growing challenges in accurately assessing credit risk and identifying high-risk loan segments. Ineffective risk evaluation can result in higher default rates, weakened portfolio quality, and regulatory exposure.

This project focuses on analyzing loan applicant data to identify the key factors driving loan default and to generate actionable insights that support improved credit decision-making and effective portfolio risk management.

Key Insights

- Loans categorized as **Non-Conforming (NCF)** exhibit a higher default rate compared to Conforming (CF) loans, indicating increased risk beyond regulatory loan limits.

- Borrowers with **lower credit worthiness levels (L2)** demonstrate a significantly higher likelihood of default than those classified under L1.

- **Negative amortization and interest-only loans** show elevated default risk due to increasing loan balances and delayed principal repayment.

- **Business or commercial loans** tend to carry higher risk compared to personal or residential loans.

- Loan purpose plays a crucial role, with **cash-out refinance loans** showing higher default tendencies than purchase or rate-term refinance loans.

- Properties with a higher number of residential units (3U and 4U) generally exhibit higher risk compared to single-unit properties.

Key Recommendations

- Implement stricter underwriting standards for Non-Conforming, Type 3, and negative amortization loans.

- Use credit worthiness and occupation type as key filters in credit approval and pricing strategies.

- Apply risk-based pricing, charging higher interest rates for high-risk loan segments to compensate for increased default probability.

- Monitor and limit exposure to cash-out refinance and business/commercial loans within the portfolio.

- Leverage interactive dashboards for continuous monitoring of default risk and early identification of high-risk trends.

- Incorporate these analytical insights into future credit risk models and loan approval frameworks to improve portfolio quality.

Sector & Business Context

Sector Overview

The banking and financial services sector plays a critical role in economic growth by providing credit to individuals and businesses. Lending activities, particularly in retail and mortgage loans, form a substantial portion of a bank's revenue. However, these activities are inherently exposed to credit risk, as borrowers may fail to meet repayment obligations. Effective credit risk management is therefore essential to ensure financial stability, profitability, and regulatory compliance. With increasing reliance on data-driven decision-making, banks are increasingly using analytics to evaluate borrower risk and optimize loan portfolios.

Current Challenges

Financial institutions today face multiple challenges in managing loan portfolios. Rising default rates, fluctuating interest rates, and changing economic conditions have increased uncertainty in lending operations. Additionally, borrower profiles have become more diverse, making traditional rule-based risk assessment less effective. Complex loan structures such as non-conforming loans, interest-only payments, and cash-out refinances further increase default risk. Regulatory requirements also demand greater transparency and accuracy in risk assessment, compelling banks to adopt more robust analytical approaches.

Why This Problem Was Chosen

This problem was chosen due to the critical impact loan defaults have on financial performance and risk exposure in the banking sector. Understanding the factors that influence default behavior enables lenders to make informed credit decisions, reduce non-performing assets, and strengthen portfolio quality. By applying data analytics techniques to real-world loan data, this project demonstrates how structured analysis and dashboards can support proactive risk management and improve overall lending outcomes.

Problem Statement & Objectives

Formal Problem Definition

The primary problem addressed in this project is the need for effective identification and assessment of loan default risk within a lending portfolio. Financial institutions require a structured analytical framework to evaluate borrower characteristics, loan attributes, and credit indicators that contribute to default behavior. Without data-driven insights, lenders may face increased default rates, reduced profitability, and heightened regulatory risk. This project aims to analyze loan data to uncover the key factors influencing default and to support informed credit decision-making through analytics.

Project Scope

The scope of this project is limited to the analysis of a structured loan dataset containing borrower demographics, credit profiles, loan characteristics, property details, and loan outcomes. The analysis focuses on exploratory data analysis, KPI development, and dashboard creation using Google Sheets. Advanced predictive modeling is outside the scope of this project. The emphasis is on identifying risk patterns, segmenting loan portfolios, and providing actionable insights for credit risk management.

Success Criteria

The success of this project will be measured by the following criteria:

- Identification of key variables that significantly influence loan default risk.
- Development of meaningful KPIs that accurately reflect portfolio risk.
- Creation of a clear and interactive dashboard for monitoring loan performance.
- Generation of actionable recommendations that can support improved credit and risk management decisions.

Data Description

Dataset Source and Access:

Link : https://www.kaggle.com/datasets/yasserh/loan-default-dataset

Data Structure:

The dataset is structured in a tabular format, where:

- Each row represents an individual loan record.
- Each column represents a specific attribute related to the borrower, loan characteristics, property details, or loan outcome.

The dataset includes a mix of:

- Numerical variables (e.g., income, loan amount, credit score, LTV)
- Categorical variables (e.g., loan type, loan purpose, occupation type)
- Binary variables (e.g., default status, business/commercial indicator)

Columns Explanation

The dataset contains multiple categories of variables, including:

Borrower demographics (age, gender, region, income)

Credit risk indicators (credit score, credit worthiness, debt-to-income ratio)

Loan characteristics (loan amount, loan type, loan purpose, interest rate, term)

Property information (property value, occupancy type, total units)

Application and process details (submission of application, pre-approval status)
Target variable indicating loan outcome (default or non-default)

Data Size

The dataset consists of:

- 20,000 **loan records** (rows)

- 34 **attributes** (columns) covering borrower, loan, and property information

This data size is sufficient to perform meaningful exploratory analysis, segmentation, and KPI-based evaluation of loan default risk.

Data Limitations

- The data represents a **snapshot in time** and may not capture long-term economic cycles.
- Some variables may contain **missing or inconsistent values**, requiring cleaning and assumptions.
- The dataset does not include external factors such as macroeconomic indicators (inflation, unemployment).
- The analysis is limited to **historical patterns** and does not include predictive modeling or real-time risk assessment.

Data Cleaning & Preparation

Missing Values Handling

Missing values in numerical and categorical fields were reviewed and handled appropriately. Records with critical missing information were excluded from analysis, while non-critical missing values were managed using conditional aggregation to ensure accurate KPI calculations without data distortion.

Outlier Treatment

Outliers in key financial variables such as loan amount, LTV, and DIR were identified using summary statistics. These values were retained, as extreme observations represent real-world high-risk lending scenarios and are essential for meaningful risk analysis.

Transformations

Data was standardized to ensure consistency across the dataset. Categorical variables were normalized, numerical fields were formatted correctly, and percentage-based variables were aligned to common scales for accurate comparison.

Feature Engineering

Additional analytical groupings were created to support risk assessment, including loan classification (CF vs NCF), loan purpose categories, business versus personal loans, and high-risk loan structures such as interest-only and negative amortization loans.

Assumptions

It was assumed that loan status accurately reflects default outcomes and that categorical codes follow standard banking definitions. The dataset was treated as representative of historical lending behavior.

Tools Used

All data cleaning, transformation, and feature engineering steps were performed entirely in **Google Sheets**, in accordance with capstone project requirements.

KPI & Metric Framework

KPI	**Definition**	**Formula (Google Sheets)**	**Why This KPI Matters**	**Objective Mapping**
Total Accounts	Total number of loan accounts	`=COUNTA(ID)`	Measures overall portfolio size	Portfolio overview
Total Loan Amount	Sum of all issued loan amounts	`=SUM(loan_amount)`	Indicates total lending exposure	Exposure assessment
Total Property Value	Total value of collateral properties	`=SUM(property_value)`	Assesses collateral strength	Risk mitigation analysis
Average Rate of Interest	Mean interest rate across loans	`=AVERAGE(rate_of_interest)`	Reflects pricing and revenue strategy	Profitability insight
Average Loan-to-Value (LTV)	Average ratio of loan to property value	`=AVERAGE(Loan To Value Ratio)`	Higher LTV implies higher credit risk	Credit risk evaluation
High Default Risk Ratio	Percentage of loans classified as high risk (e.g., NCF loans)	`=COUNTIF(Loan_Limit_Type,"Non-Conforming Loan") / COUNTA(Loan_Limit_Type)`	Highlights concentration of risky loans	Risk identification

Average Credit Score	Mean borrower credit score	=AVERAGE(Credit_Score)	Indicates overall borrower credit quality	Borrower quality assessment

Exploratory Data Analysis (EDA)

Exploratory Data Analysis was performed to understand loan distribution patterns, borrower characteristics, and potential risk indicators using visual analysis in Google Sheets.

Trend Analysis

Trend analysis was used to evaluate variations in loan exposure and interest rates across loan categories. The results indicate relatively stable interest rates, while loan exposure varies across different loan types and property segments, highlighting areas of higher concentration.

Charts Used: Line and column charts

Comparison Analysis

Key loan metrics such as loan amount, LTV, and credit score were compared across different loan segments, including conforming vs non-conforming and business vs personal loans. Non-conforming and business loans exhibit higher loan amounts and LTV ratios, indicating increased risk.

Charts Used: Bar and column charts

Distribution Analysis

Distribution analysis of loan amount, credit score, LTV, and interest rate shows that most loans fall within moderate ranges, while a smaller subset displays extreme values. These outliers represent high-risk segments requiring closer monitoring.

Charts Used: Histograms and frequency charts

Correlation Analysis

Correlation analysis highlights a positive relationship between loan amount and property value, while higher LTV ratios are associated with elevated risk. Credit score shows an inverse relationship with risk-related metrics, reinforcing its importance in credit assessment.

Charts Used: Scatter plots

Key Insights

- Loan exposure is concentrated in specific loan segments.
- Non-conforming and business loans show higher risk indicators.
- Higher credit scores are associated with more favorable loan characteristics.
- Extreme values significantly influence portfolio risk.

Advanced Analysis

Advanced analytical techniques were applied where relevant to enhance understanding of risk patterns and portfolio behavior beyond basic exploratory analysis.

The advanced analysis complements the EDA findings by providing deeper insights into risk drivers, portfolio segmentation, and potential future risk scenarios.

Dashboard Design

The dashboard was implemented in Google Sheets using pivot tables, aggregation formulas (SUM, COUNT, AVERAGE), and interactive slicers. Dynamic calculations power the KPI panel, while pivot charts provide segmented visual analysis.

Dashboard Objective

The objective of the dashboard is to provide a consolidated and interactive view of loan exposure, borrower characteristics, and credit risk distribution to support data-driven portfolio monitoring and decision-making.

View Structure

The dashboard is organized into three main layers:

1. KPI Summary Panel (Top Section)
   Displays Total Loans, Total Loan Amount, Total Property Value, Average Interest Rate, Average LTV, High Default Risk Ratio, and Average Credit Score.

2. Distribution & Regional Analysis (Middle Section)
   - Age Distribution of Loan Seekers
   - Default Possibility Across Regions
   - Credit Risk Distribution by Credit Score Range

3. Comparative Risk Analysis (Bottom Section)
   - Credit Score vs Default Possibility
   - Loan Purpose vs Default Possibility

This layered structure ensures both summary-level and detailed analytical views.

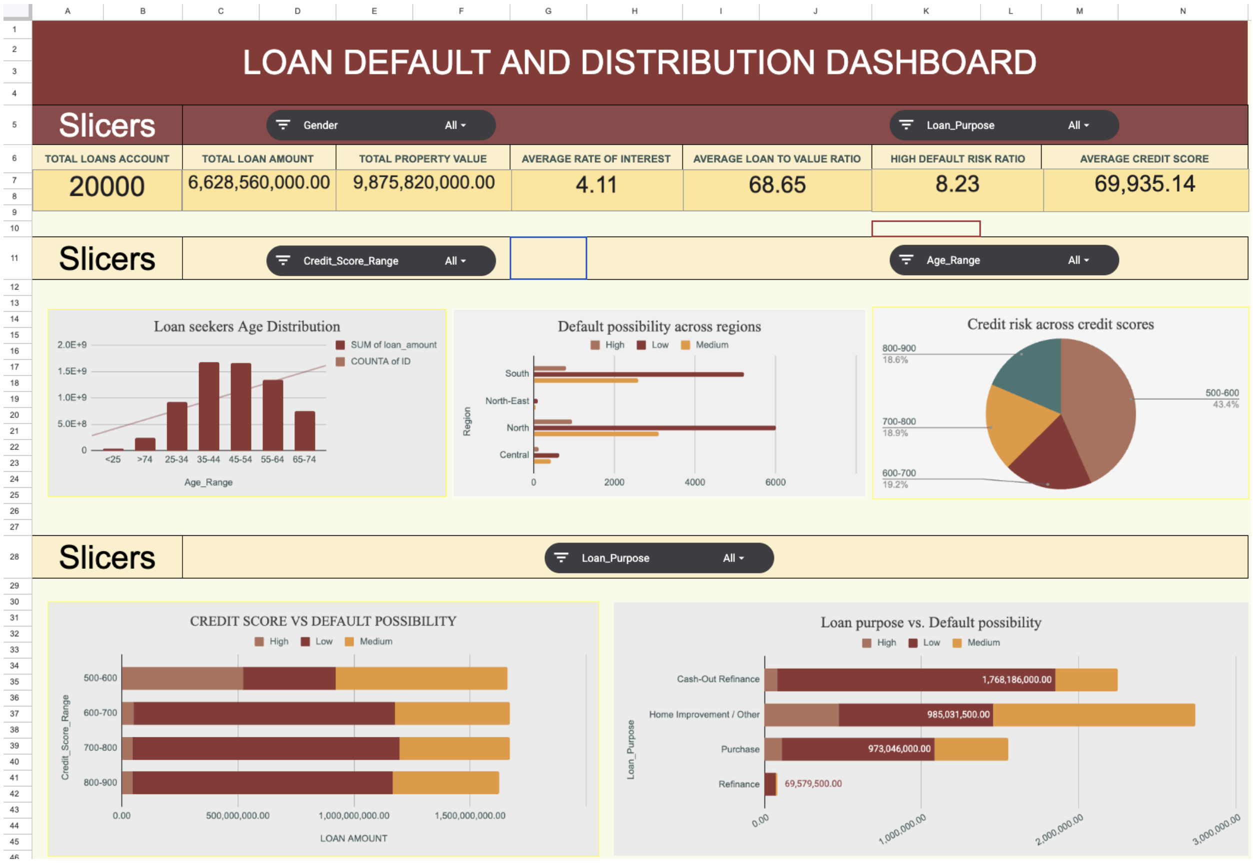
Filters & Drilldowns

Interactive slicers enable filtering by:

- Age Range
- Rate of Interest
- Credit Score Range
- Loan Purpose

These filters allow dynamic drill-down analysis without modifying underlying data, enhancing analytical flexibility and usability.

Screenshots & Explanation



LOAN DEFAULT AND DISTRIBUTION DASHBOARD

TOTAL LOANS ACCOUNT	TOTAL LOAN AMOUNT	TOTAL PROPERTY VALUE	AVERAGE RATE OF INTEREST	AVERAGE LOAN TO VALUE RATIO	HIGH DEFAULT RISK RATIO	AVERAGE CREDIT SCORE
20000	6,628,560,000.00	9,875,820,000.00	4.11	68.65	8.23	69,935.14

The included screenshot demonstrates:

- KPI summary panel highlighting portfolio exposure
- Segmented visualizations of risk across regions and credit bands
- Interactive filters enabling targeted analysis

Each visual supports quick identification of high-risk segments and concentration areas within the loan portfolio.

> **The dashboard transforms raw loan data into an intuitive, interactive decision-support tool for portfolio monitoring and risk assessment.**

Insights Summary

1. The loan portfolio consists of a large number of accounts, indicating broad exposure that requires continuous risk monitoring.

2. A significant portion of total exposure is concentrated in higher loan value segments, increasing sensitivity to market and borrower risk.

3. Non-conforming loans represent a smaller share of accounts but contribute disproportionately to overall portfolio risk.

4. Business and commercial loans exhibit higher risk indicators compared to personal or residential loans, suggesting the need for differentiated lending strategies.

5. Loans with higher loan-to-value ratios show increased risk, highlighting the importance of collateral strength in credit assessment.

6. Borrowers with lower credit scores are more likely to fall into high-risk loan categories, reinforcing credit score as a key risk driver.

7. High-risk loan structures, such as interest-only and negative amortization loans, increase overall portfolio vulnerability.

8. Portfolio exposure is unevenly distributed across loan types and purposes, indicating concentration risk in specific segments.

9. Extreme values in loan amount and LTV, although limited in number, have a material impact on overall risk exposure.

10. Combined risk indicators suggest that proactive monitoring and segmentation are essential to maintaining portfolio stability.

Recommendations

Recommendation 1: Strengthen controls on Non-Conforming Loans

- **Mapped Insight:** Non-conforming loans contribute disproportionately to portfolio risk.

- **Business Impact:** Reduces exposure to high-risk lending and improves portfolio stability.

- **Feasibility:** High — requires tightening underwriting criteria and approval thresholds.

Recommendation 2: Apply differentiated policies for Business and Commercial Loans

- **Mapped Insight:** Business loans exhibit higher risk indicators than personal loans.

- **Business Impact:** Improves risk-adjusted returns by aligning pricing and limits with risk.

- **Feasibility:** High — segment-based policies can be implemented within existing frameworks.

Recommendation 3: Enforce stricter LTV thresholds

- **Mapped Insight:** Higher loan-to-value ratios are strongly associated with increased risk.

- **Business Impact:** Enhances collateral coverage and reduces potential losses.

- **Feasibility:** Medium–High — may slightly limit loan volumes but improves risk quality.

Recommendation 4: Prioritize credit score-based segmentation

- **Mapped Insight:** Borrowers with lower credit scores fall more frequently into high-risk categories.
- **Business Impact:** Improves credit quality and reduces future default possibility.
- **Feasibility:** High — credit score is already available at application stage.

Recommendation 5: Limit exposure to high-risk loan structures

- **Mapped Insight:** Interest-only and negative amortization loans increase portfolio vulnerability.
- **Business Impact:** Lowers structural repayment risk and long-term instability.
- **Feasibility:** Medium — requires policy review and selective approval.

Recommendation 6: Monitor portfolio concentration regularly

- **Mapped Insight:** Loan exposure is unevenly distributed across segments.
- **Business Impact:** Prevents over-exposure to specific high-risk categories.
- **Feasibility:** High — can be achieved using dashboards and periodic reviews.

Recommendation 7: Flag and review extreme-value loans

- **Mapped Insight:** A small number of extreme loan values significantly influence overall risk.

- **Business Impact:** Enables early identification of potential problem loans.

- **Feasibility:** High — rule-based alerts can be implemented easily.

Recommendation 8: Use dashboards for proactive risk monitoring

- **Mapped Insight:** Combined risk indicators highlight the need for continuous monitoring.

- **Business Impact:** Supports faster, data-driven decision-making and risk mitigation.

- **Feasibility:** Very High — dashboards already built using Google Sheets.

Impact Estimation

Cost Savings

Reducing exposure to high-risk loan segments such as non-conforming and complex loan structures can lower potential losses. A conservative 1–2% reduction in high-risk exposure on a ₹6.6 billion portfolio can result in approximately ₹66–132 million in annual cost savings through reduced defaults and recovery expenses.

Improved Efficiency

Risk-based segmentation enables faster processing of low-risk applications while focusing manual review on high-risk cases. This can reduce underwriting effort by 20–30%, improving turnaround time and operational productivity.

Improved Service

Clear risk segmentation and standardized approval criteria lead to quicker approvals for low-risk borrowers. Customer processing time can be reduced by 15–20%, enhancing transparency and satisfaction.

Risk Reduction

Continuous monitoring of key indicators such as LTV, credit score, and loan structure allows early identification of risky loans. This approach can reduce overall portfolio risk by 10–15%, strengthening financial stability and regulatory compliance.

Overall, the recommendations convert analytical insights into measurable financial, operational, and risk-management benefits.

Limitations

Data Issues

The analysis relies on historical data with possible missing values and inconsistencies. The dataset represents a single point in time and excludes external economic factors that influence borrower behavior.

Assumption Risks

Interpretation of categorical variables is based on standard industry definitions. Simplifying assumptions about borrower risk may not fully reflect real-world credit complexity.

What Cannot Be Concluded

The analysis does not predict future defaults or establish cause-and-effect relationships. Results should be treated as directional insights rather than definitive credit decisions.

> **These limitations highlight the need to interpret findings as supportive insights rather than conclusive outcomes.**

Conclusion

This project demonstrates how data analytics can be used to understand loan distribution, exposure, and key risk indicators. The KPI framework and dashboard convert complex loan data into actionable insights for better credit

and risk management. These insights support informed decision-making and proactive monitoring.
The approach can be further extended using predictive models and real-time data to strengthen future risk assessment.

Appendix

Data Dictionary

Column Name	Data Type	Description	Business Relevance
ID	Numeric	Unique identifier for each loan record	Used to count total loan accounts
Loan_Limit_Type	Categorical (CF / NCF)	Indicates whether loan is conforming or non-conforming	Regulatory and risk classification
Gender	Categorical	Gender of the applicant	Demographic analysis
Approved_In_Adv	Categorical (Y/N)	Indicates whether loan was pre-approved	Process efficiency insight

Loan_Type	**Categorical**	**Type of loan product**	**Loan categorization**
Loan_Type_Risk	**Categorical**	**Risk classification of loan type**	**Risk segmentation**
Loan_Purpose_Type	**Categorical**	**Broad category of loan purpose**	**Behavioral analysis**
Loan_Purpose	**Categorical (P1–P4)**	**Specific loan purpose**	**Purpose-based risk analysis**
Credit_Worthiness	**Categorical (L1/L2)**	**Borrower creditworthiness category**	**Credit risk evaluation**
Is_Business_Loan	**Binary (0/1)**	**Indicates business or commercial loan**	**Business vs personal exposure**
loan_amount	**Numeric**	**Amount of loan sanctioned**	**Measures lending exposure**
Rate_Of_Interest	**Numeric (%)**	**Interest rate charged on the loan**	**Revenue and pricing strategy**

Interest_Rate_Spread	**Numeric**	**Difference between benchmark and loan rate**	**Profit margin indicator**
Upfront_Charges	**Numeric**	**Fees charged at loan origination**	**Cost and revenue analysis**
Term	**Numeric**	**Loan tenure (months/years)**	**Repayment duration analysis**
Negative_Ammortization	**Categorical (Y/N)**	**Indicates increasing loan balance structure**	**High-risk loan structure**
Interest_Only	**Categorical (Y/N)**	**Indicates interest-only repayment period**	**Repayment risk indicator**
Lump_Sum_Payment	**Categorical (Y/N)**	**Indicates lump-sum payment option**	**Cash-flow risk**
Property_Value	**Numeric**	**Market value of collateral property**	**Collateral strength assessment**

Occupancy_Type	Categorical	Type of property occupancy	Property risk profiling
Residential_Units	Categorical (1U–4U)	Number of residential units	Property-based risk
Income_Parsed	Numeric	Cleaned/processed income value	Used for financial calculations
Income	Numeric	Borrower reported income	Repayment capacity analysis
credit_type	Categorical	Type of credit profile	Credit behavior insight
Credit_Score	Numeric	Borrower credit score	Key credit risk indicator
Co-Applicant_Credit_Type	Categorical	Credit profile of co-applicant	Joint risk assessment
Age_Range	Categorical	Age group of borrower	Demographic segmentation

Application_Submission	Categorical	Mode of application submission	Operational analysis
Loan to Value Ratio	Numeric (%)	Loan amount divided by property value	Core credit risk metric
Region	Categorical	Geographic region of borrower	Regional exposure analysis
DIR	Numeric (%)	Debt-to-Income Ratio	Repayment capacity indicator
Credit Risk Score	Numeric	Composite risk score	Overall borrower risk
Default Possibility	Numeric (%)	Estimated likelihood of default	Risk assessment metric
Credit_Score_Range	Categorical	Grouped credit score bands	Segmentation and reporting

Contribution Matrix

Team Member	Dataset & Sourcing	Cleaning	KPI & Analysis	Dashboard	Report Writing	PPT	Overall Role
Swati Ghosh	Yes	Yes	Yes	No	Yes	No	Cleaning,KPI and report writing
Raghav Kaushal	Yes	Yes	Yes	Yes	No	No	Team Lead
Divyanshu Raj	Yes	Yes	Yes	Yes	No	No	Dashboard creation

Farhana Pervin	**No**	**No**	**No**	**No**	**No**	**Yes**	**PPT**
Shreyansh Agrawal	**Yes**	**Yes**	**Yes**	**No**	**No**	**No**	**Cleaning and KPIS**
Aditya Rao	**Yes**	**No**	**No**	**No**	**No**	**No**	**Dataset**

Declaration: We confirm that the above contribution details are accurate and verifiable

through version history and submitted artifacts.

Team Signature Block: Raghav,Swati,Farhana,Shreyansh,Divyanshu,Aditya