# AIRBNB DATA CLEANING, PROCESSING, AND ADVANCED ANALYSIS

Shubham Raj – G23AI2028
Bhavesh Arora – G23AI2126
Niraj Bhagchandani – G23AI2087
Jai Singh Khushvah – G23AI2018
Paras Panda – G23AI2117
Jatin Shrivas – G23AI2094

## CAPSTONE PROJECT

This Report Presented for the PG Diploma in Data Engineering

Department of Computer Science IIT Jodhpur

# CONTENTS

# PROJECT BACKGROUND

**Onboarding and Initial Data Handling**

1. Understand, clean, and augment the Airbnb dataset, and perform exploratory data analysis (EDA)

**Data Ingestion and Optimization**

2. Develop efficient data ingestion pipelines and optimize data storage.

**Data Transformation**

3. Transform data to meet business requirements and enable advanced analytics

**Data Warehousing, Reporting, and Visualization**

4. Understand, clean, and augment the Airbnb dataset, and perform exploratory data analysis (EDA)

# KEY DELIVERABLES

01 

Monthly Progress Reports : Detailed reports outlining progress on each problem statement, including methodologies, challenges, and preliminary results.

02 

Final Data Warehouse : A fully populated and optimized data warehouse schema.

03 

Predictive Models : Trained and validated machine learning models with performance metrics.

04 

BI Dashboards: Interactive dashboards for exploring data and insights.

05 

Comprehensive Final Report : A detailed report summarizing the entire project, including methodologies, findings, and recommendations.

# Comprehensive Overview of Airbnb Datasets

## Calendar.csv

- **Number of Rows:** 1,393,570
- **Number of Columns:** 7
- **Key Column Names:**
  - Listing_id,
  - Date,
  - Available
  - Price
  - Minimum nights
  - Maximum nights
  - Available units

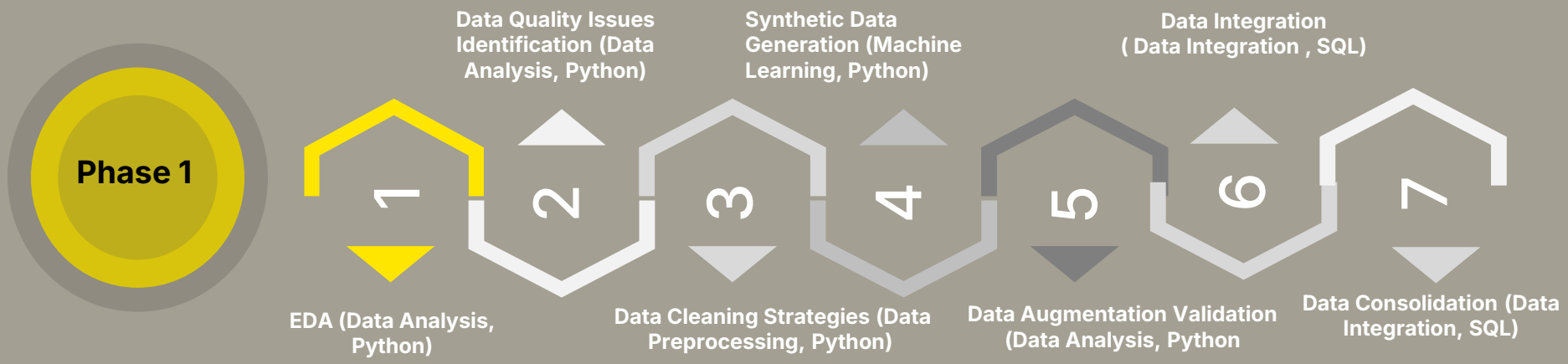

## Listings.csv

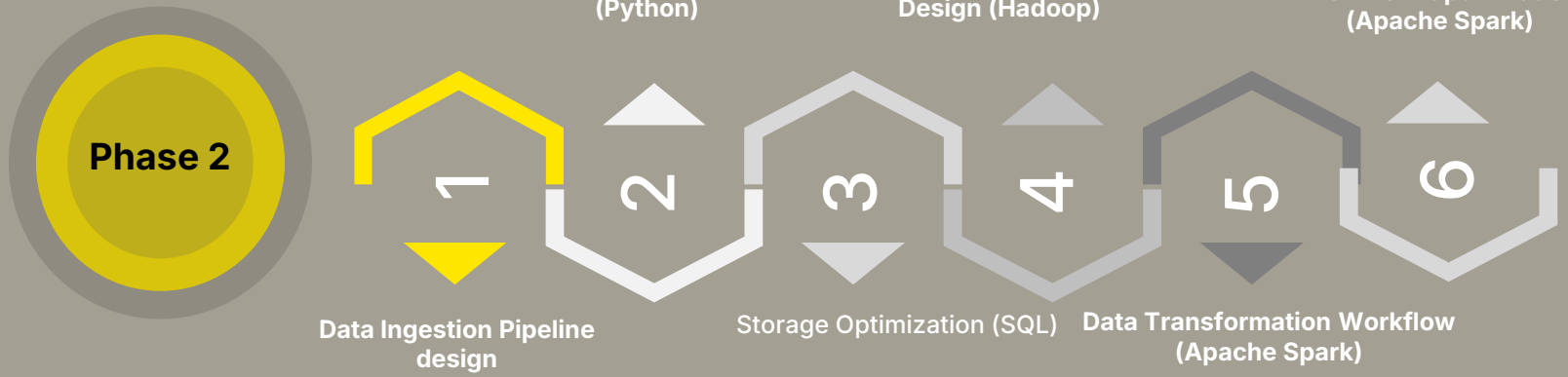
- **Number of Rows:** 3,818
- **Number of Columns:** 92
- **Key Column Names:**
  - id
  - listing_url
  - scrape_id,
  - Last_scraped
  - Name
  - host_id
  - host_name,


## Reviews.csv

- **Number of Rows:** 84,849
- **Number of Columns:** 7
- **Key Column Names:**
  - listing_id
  - id
  - Date
  - reviewer_id
  - reviewer_name
  - Comments
  - sentiment score

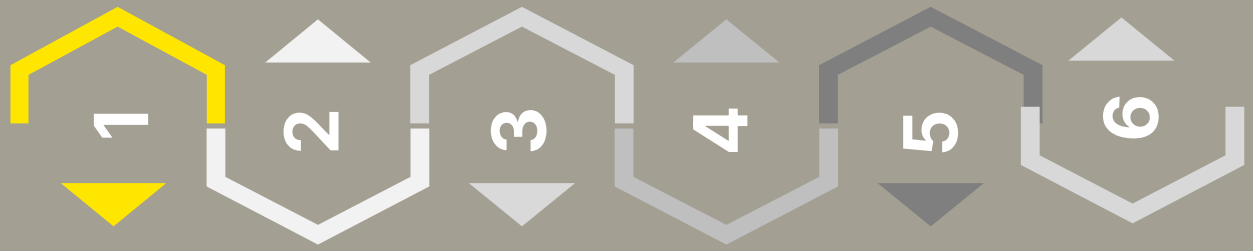
# DIFFERENT PHASES OF PROJECT



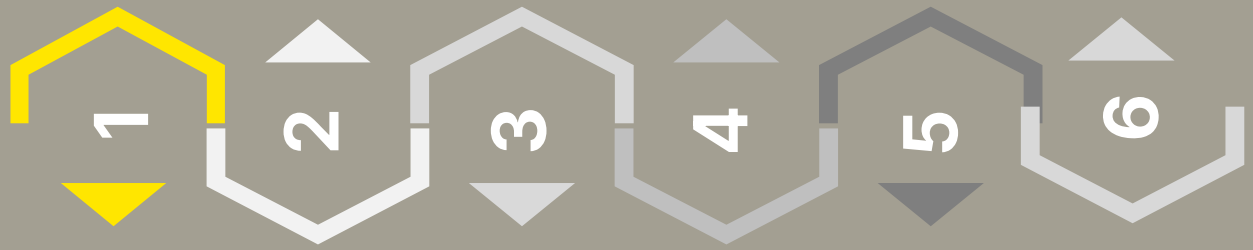
# DIFFERENT PHASES OF PROJECT

## Phase 3




1. Feature Engineering (ML, Python)
2. Feature Validation (Data Analysis, Python)
3. Data Normalization (Data preprocessing, python)
4. Data Encoding (Preprocessing python)
5. Data Warehousing (Amazon RedShift)
6. Warehouse Schema Design (Amazon Redshift)

## Phase 4




1. Predictive Modelling (ML, Python)
2. Model Validation (Data Analysis, Python)
3. Business Intelligence Reporting (Data visualization, PowerBI)
4. Report Optimization (Data visualization , PowerBi)
5. Optimization Analysis (Data Analysis, Python)
6. Revenue Analysis (Data Analysis, Python)

# Some Useful Links

GitHub Links:

https://github.com/shubham14p3/IITJ-AIRBNB-DATA-CLEANING-PROCESSING-AND-ADVANCED-ANALYSIS-CAPSTONE-PROJECT/tree/main/backend

Front End Link:

http://3.222.77.245:5173/

Back End Link:

http://44.198.64.142:5000/

Phase 01

# Onboarding and Initial Data Handling

# PROBLEM STATEMENT & PROPOSED SOLUTIONS

## 1 Comprehensive Data Profiling and Cleaning

Conduct EDA on calendar.csv, listings.csv, and reviews.csv. Identify and visualize data quality issues.

### Instructions

- Perform EDA to identify missing values, outliers, and anomalies using visualizations like heatmaps and box plots.
- Implement data cleaning strategies, including imputing missing values, treating outliers, and standardizing data formats.

### Solution

- **EDA Insights :** Analysed datasets with summaries, heatmaps for missing data, and box plots for outliers.
- **Key Findings :** Missing values in critical columns like price and review_scores. Outliers in price and availability_365.
- **Data Cleaning :** Imputed missing values (median for numeric, mode for categorical). Capped outliers at the 95$^{th}$ percentile. Standardized date and currency formats.
- **Documentation :** Captured key processes, code snapshots, and a summary table for cleaned data.

## 2 Data Augmentation and Synthesis

Develop techniques to augment the dataset by creating synthetic data.

### Instructions

- Develop machine learning models (e.g., GANs or SMOTE) to predict missing values and generate synthetic data points.
- Validate the augmented data using statistical methods to ensure robustness.

### Solution

- **Data Augmentation :** Used SMOTE to address class imbalance in categorical variables and GANs to synthesize data for continuous columns, mitigating the impact of missing values.
- **Validation :** Conducted statistical tests (e.g., Chi-square and t-tests) to verify the synthetic data's distribution and cross-validated to ensure alignment with original patterns.
- **Outcome:** Successfully filled gaps in listings.csv and reviews.csv with synthetic data, enhancing dataset robustness for analysis

## 3 Data Integration and Consolidation

Integrate calendar.csv, listings.csv, and reviews.csv into a consolidated dataset.

### Instructions

- Merge the datasets, resolve inconsistencies such as duplicate records, and remove redundant information.
- Optimize the consolidated dataset for subsequent analysis by standardizing data types and normalizing values.

### Solution

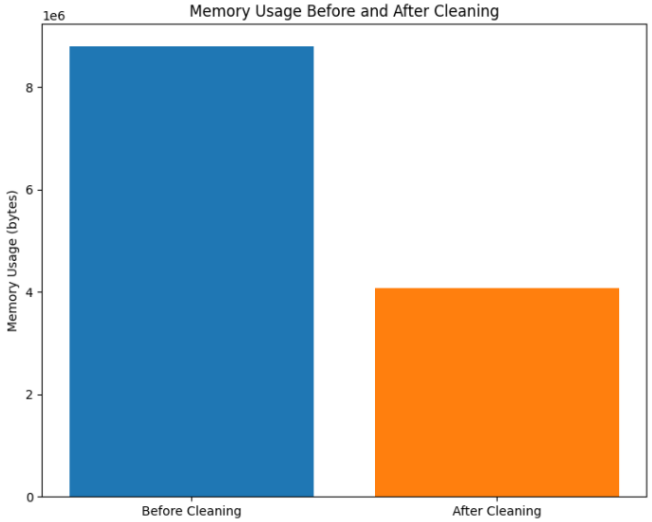
- **Merging Process :** Identified primary keys (e.g., listing_id), resolved duplicates, and addressed data type discrepancies during dataset merging.
- **Data Standardization :** Ensured consistent formats for dates and numeric columns and normalized values to align with expected ranges for smooth analysis.

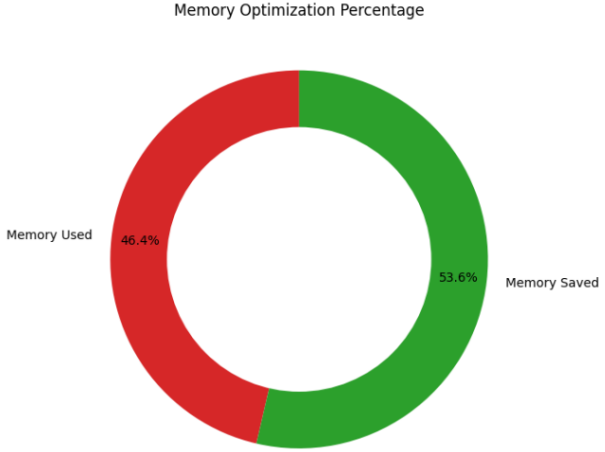
# Listings Dataset

Changes in Missing Values:

| | Missing Values | Percentage |
|---|---|---|
| id | 0 | 0.0 |
| listing_url | 0 | 0.0 |
| name | 0 | 0.0 |
| host_id | 0 | 0.0 |
| host_name | 0 | 0.0 |
| host_since | 0 | 0.0 |
| neighbourhood | 0 | 0.0 |
| latitude | 0 | 0.0 |
| longitude | 0 | 0.0 |
| number_of_reviews | 0 | 0.0 |
| review_scores_rating | 0 | 0.0 |
| instant_bookable | 0 | 0.0 |
| room_type_id | 0 | 0.0 |
| cancellation_policy_id | 0 | 0.0 |

Memory Usage Optimization:
Memory reduced by: 4718414 bytes
Memory optimized by: 53.64%



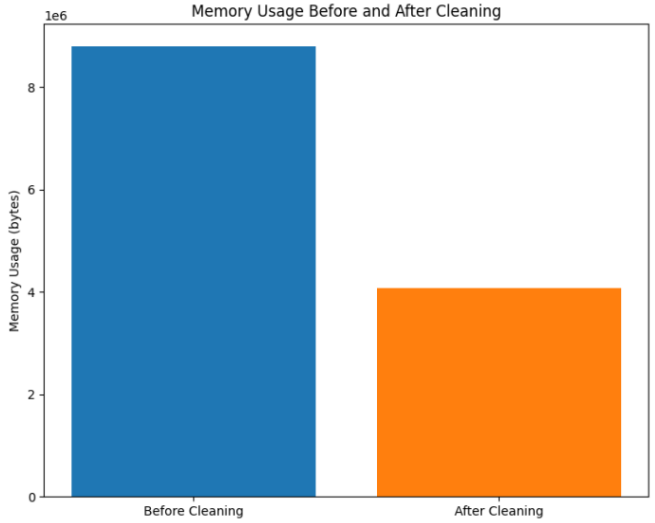

53% Memory Saved

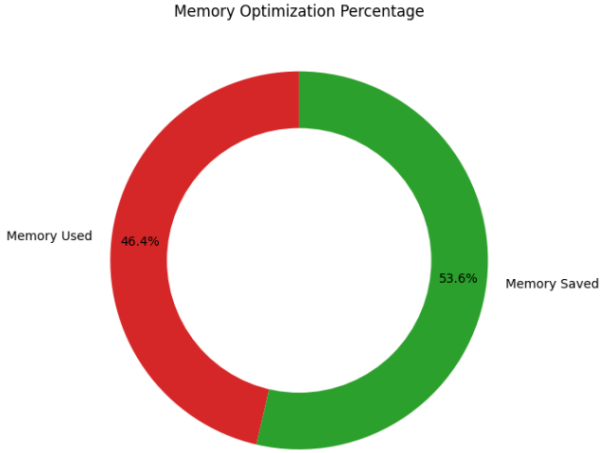
# Reviews Dataset

Changes in Missing Values:

| | Missing Values | Percentage |
|---|---|---|
| id | 0 | 0.0 |
| listing_url | 0 | 0.0 |
| name | 0 | 0.0 |
| host_id | 0 | 0.0 |
| host_name | 0 | 0.0 |
| host_since | 0 | 0.0 |
| neighbourhood | 0 | 0.0 |
| latitude | 0 | 0.0 |
| longitude | 0 | 0.0 |
| number_of_reviews | 0 | 0.0 |
| review_scores_rating | 0 | 0.0 |
| instant_bookable | 0 | 0.0 |
| room_type_id | 0 | 0.0 |
| cancellation_policy_id | 0 | 0.0 |

Memory Usage Optimization:
Memory reduced by: 4718414 bytes
Memory optimized by: 53.64%



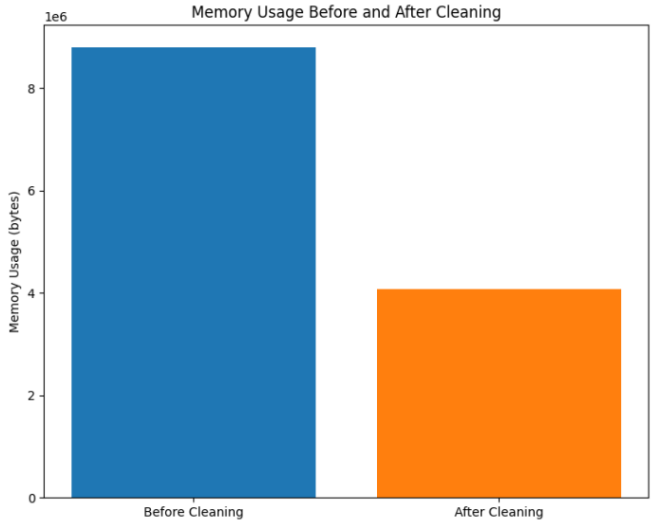

53% Memory Saved

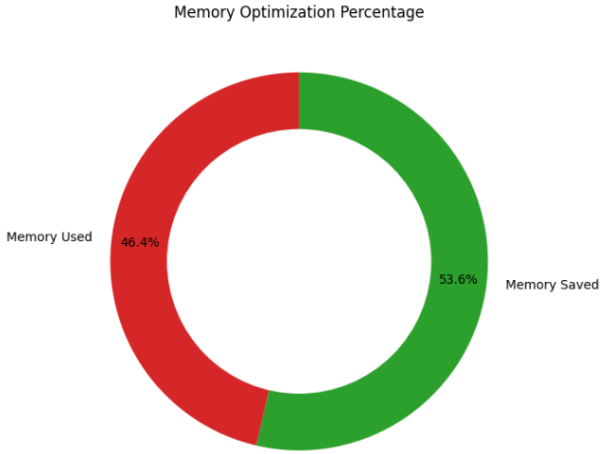
# Calendar Dataset

Changes in Missing Values:

| | Missing Values | Percentage |
|---|---|---|
| id | 0 | 0.0 |
| listing_url | 0 | 0.0 |
| name | 0 | 0.0 |
| host_id | 0 | 0.0 |
| host_name | 0 | 0.0 |
| host_since | 0 | 0.0 |
| neighbourhood | 0 | 0.0 |
| latitude | 0 | 0.0 |
| longitude | 0 | 0.0 |
| number_of_reviews | 0 | 0.0 |
| review_scores_rating | 0 | 0.0 |
| instant_bookable | 0 | 0.0 |
| room_type_id | 0 | 0.0 |
| cancellation_policy_id | 0 | 0.0 |

Memory Usage Optimization:
Memory reduced by: 4718414 bytes
Memory optimized by: 53.64%




53% Memory Saved

# 02

# Phase

## Data Ingestion and Optimization

# PROBLEM STATEMENT & PROPOSED SOLUTIONS

## 4

### Scalable Data Ingestion Pipeline

Design and implement a scalable data ingestion pipeline using Python, Pandas, and SQL.

## 5

### Advanced Storage Optimization Techniques

Implement storage optimization techniques such as data partitioning, indexing, and compression.

## Instructions

- Incorporate error handling, logging, and monitoring features for reliability.
- Optimize the pipeline for performance, including batch processing and real-time data ingestion.

- Apply storage optimization in a cloud-based data storage solution like AWS RDS or Google Big Query.

## Solution

- **Pipeline Development:** Designed a scalable data ingestion pipeline using Python, Pandas, and SQL for efficient data handling and storage.
- **Error Handling and Logging:** Incorporated robust error handling and logging mechanisms to enhance reliability and troubleshoot issues effectively.
- **Performance Optimization:** Implemented batch processing for large datasets and real-time ingestion features to support dynamic data workflows.
- **Monitoring and Alerts:** Enabled proactive monitoring with dashboards or alerts to track pipeline health and ensure smooth

- **Data Partitioning:** Implemented partitioning techniques in cloud-based storage like AWS RDS or Google BigQuery to organize data for faster query performance.
- **Indexing:** Applied indexing on critical fields to enhance data retrieval speeds and support high-performance queries.
- **Data Compression:** Enabled compression for large datasets to reduce storage costs while maintaining data integrity and accessibility.
- **Cloud-Based Optimization:** Leveraged tools in AWS RDS or BigQuery for seamless integration of these optimizations within a scalable cloud

# 03

# Phase

## Data Transformation

# PROBLEM STATEMENT & PROPOSED SOLUTIONS

## 6 Complex Data Transformation Workflows

Design data transformation workflows using tools like Apache Spark or Apache Airflow.

### Instructions

- Handle data aggregation, filtering, and enrichment processes to prepare for advanced analytics.
- Ensure the workflows are scalable and efficient.

### Solution

- **Data Aggregation:** Use Spark's groupBy and agg functions to compute metrics (e.g., total sales, average values).
- **Data Filtering:** Eliminate unwanted or erroneous records using the filter function
- **Partitioning and Caching:** Ensure datasets are partitioned by frequently used keys (e.g., date or region).Caching: Cache intermediate results to reduce redundant computations Tion.
- **Validation:** Perform checks to ensure data accuracy and completeness using test cases or by comparing against

## 7 Feature Engineering for Machine Learning

Develop techniques for feature engineering to extract meaningful features from the dataset.

### Instructions

- Extract temporal features from calendar.csv and textual features from reviews.csv.
- Validate the effectiveness of these features for machine learning models.

### Solution

- **Feature Engineering :** Create and preprocess new features, encode categories, and handle missing data.
- **Model Training & Evaluation :** Train models and assess performance using metrics like accuracy and precision.
- **Impact Analysis :** Compare model performance before and after feature engineering and identify key features driving improvements.

## 8 Advanced Data Normalization and Encoding

Implement data normalization techniques for machine learning model preparation.

### Instructions

- Apply normalization techniques such as min-max scaling and Z-score normalization.
- Use encoding methods like one-hot encoding, label encoding, and embeddings for categorical variables.

### Solution

- **Normalization & Encoding:** Apply Scikit-Learn's Standard Scaler and OneHotEncoder to scale numerical features and encode categorical variables.
- **Model Performance:** Train models using the preprocessed data and measure key metrics like accuracy, precision, or recall.

# 04

# Phase

## Data Warehousing, Reporting, and Visualization

# PROBLEM STATEMENT & PROPOSED SOLUTIONS

## 9

### Enterprise Data Warehousing

Design an enterprise-level data warehouse schema using a cloud-based solution.

## 10

### Advanced Business Intelligence Reporting

Create advanced business intelligence reports and dashboards using tools like Power BI or Tableau.

## Instructions

- Create star and snowflake schemas to support complex querying and analytics.

- Focus on key business metrics such as occupancy rates, revenue per listing, and customer satisfaction..

## Solution

- **Schema Design:** Create an efficient data warehouse schema using Redshift or BigQuery.
- **ETL Process:** Implement ETL pipelines for data extraction, transformation, and loading.Data
- **Quality & Performance:** Ensure data quality and optimize query performance within the data warehouse.

- **Dashboard Creation:** Build interactive dashboards using tools like Tableau or Power BI.
- **Data Integration:** Integrate and aggregate data for real-time reporting.
- **Insights & Optimization:** Derive actionable insights and optimize visualizations for clarity.

# PROBLEM STATEMENT & PROPOSED SOLUTIONS

## 11

### Predictive Analytics and Machine Learning

Develop predictive models using machine learning algorithms.

## 12

### Optimization and Revenue Analysis

Perform advanced data analysis to identify optimization areas and potential revenue growth.

## Instructions

- Use regression, classification, and clustering techniques to derive insights from the data.

- Use techniques like A/B testing, cohort analysis, and time series analysis to identify optimization areas.

## Solution

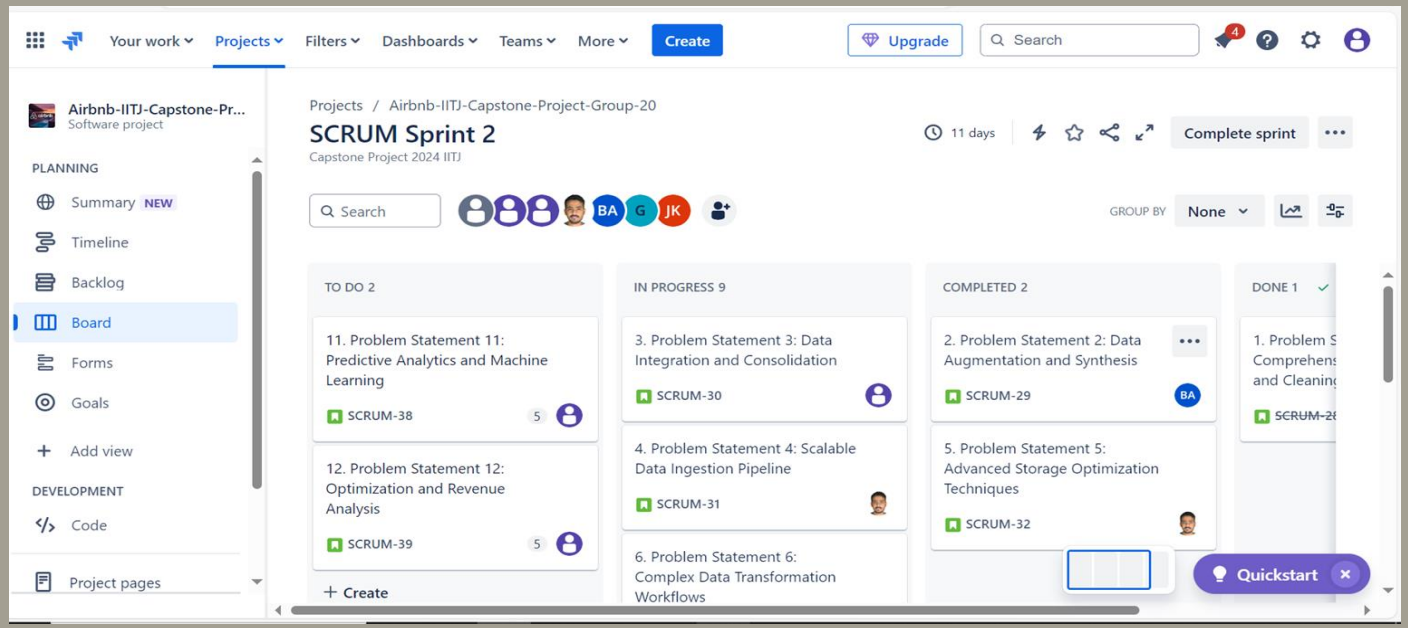
- **Model Building:** Develop and train predictive models using appropriate algorithms.
- **Model Validation:** Validate model performance through testing and cross-validation.
- **Iterative Improvement:** Continuously refine the model based on performance metrics and document the development process and results.

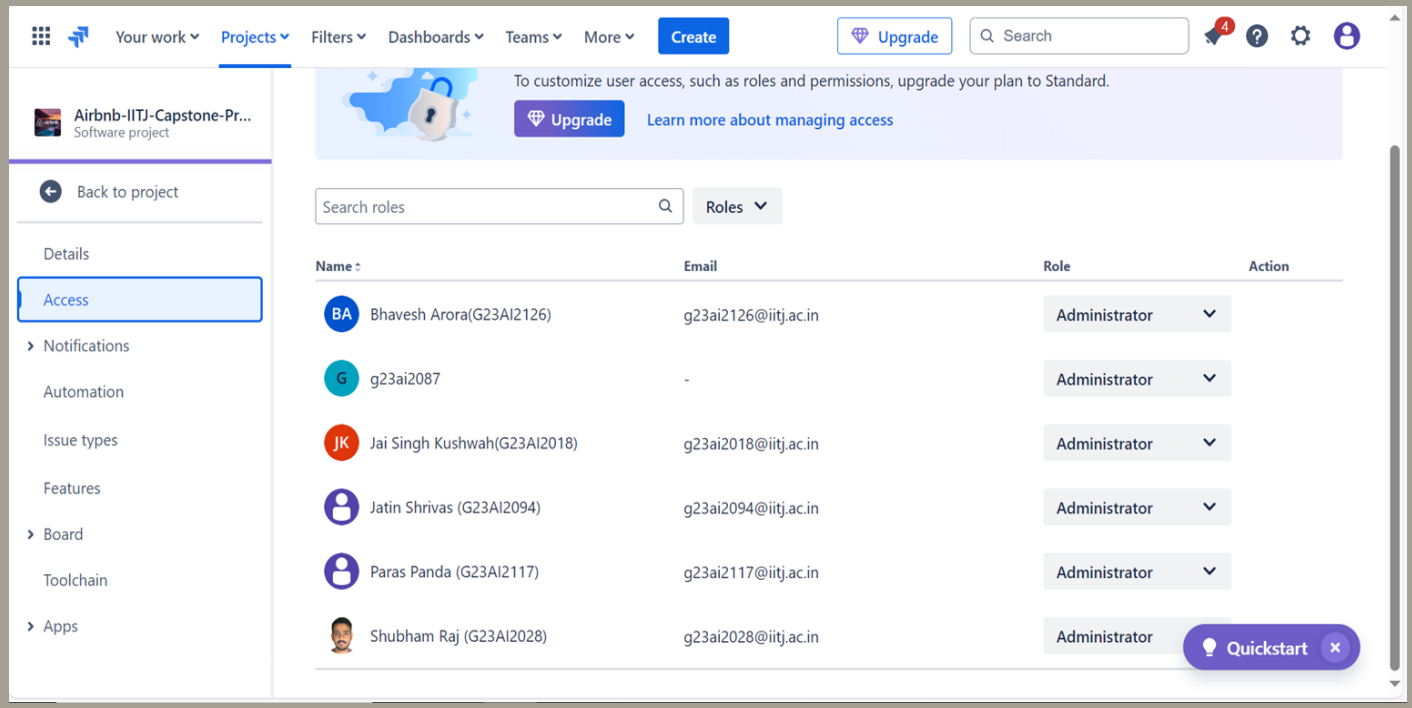
- **Data Analysis:** Analyze data to identify trends, patterns, and key insights.
- **Actionable Recommendations:** Provide recommendations for optimization and revenue growth based on analysis.
- **Documentation:** Document findings and the impact of recommendations for future reference.

# SCREENSHOTS

- JIRA Project Allocation and Tracking
- Github project Upload and Contribution
- Screenshots from Google cloud and AWS (RDS)

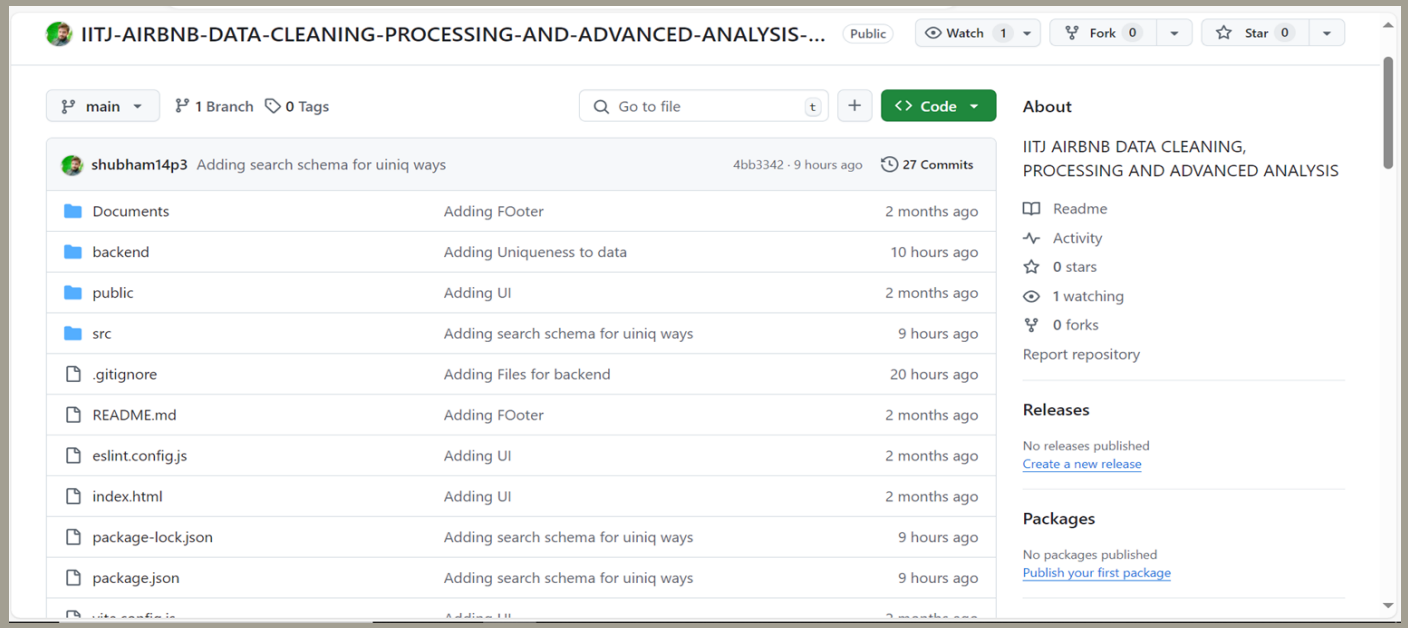
# JIRA PROJECT ALLOCATION AND TRACKING

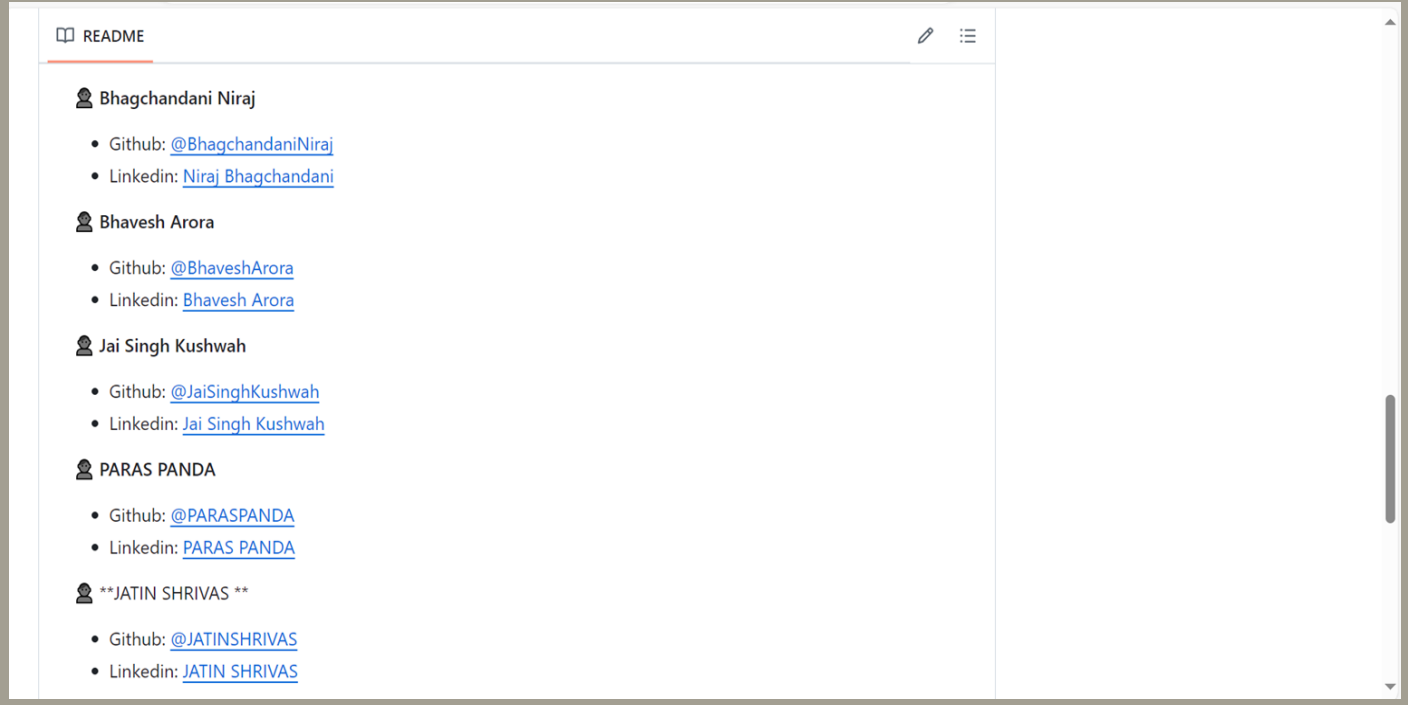


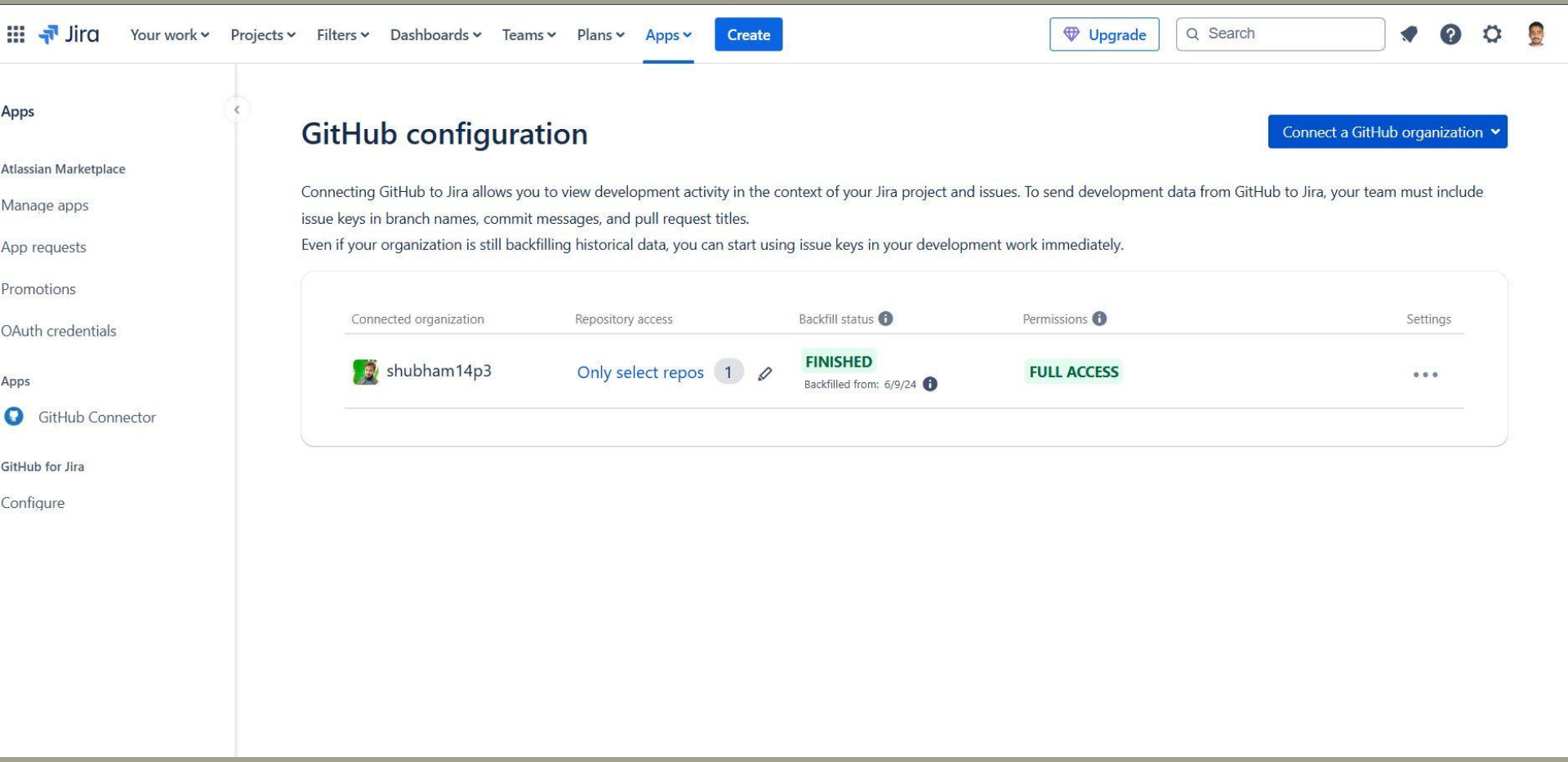

**Jira URL:**
https://iitj-team-data-crunchers.atlassian.net/jira/software/projects/SCRUM/boards/1

# GITHUB PROJECT UPLOAD AND CONTRIBUTION

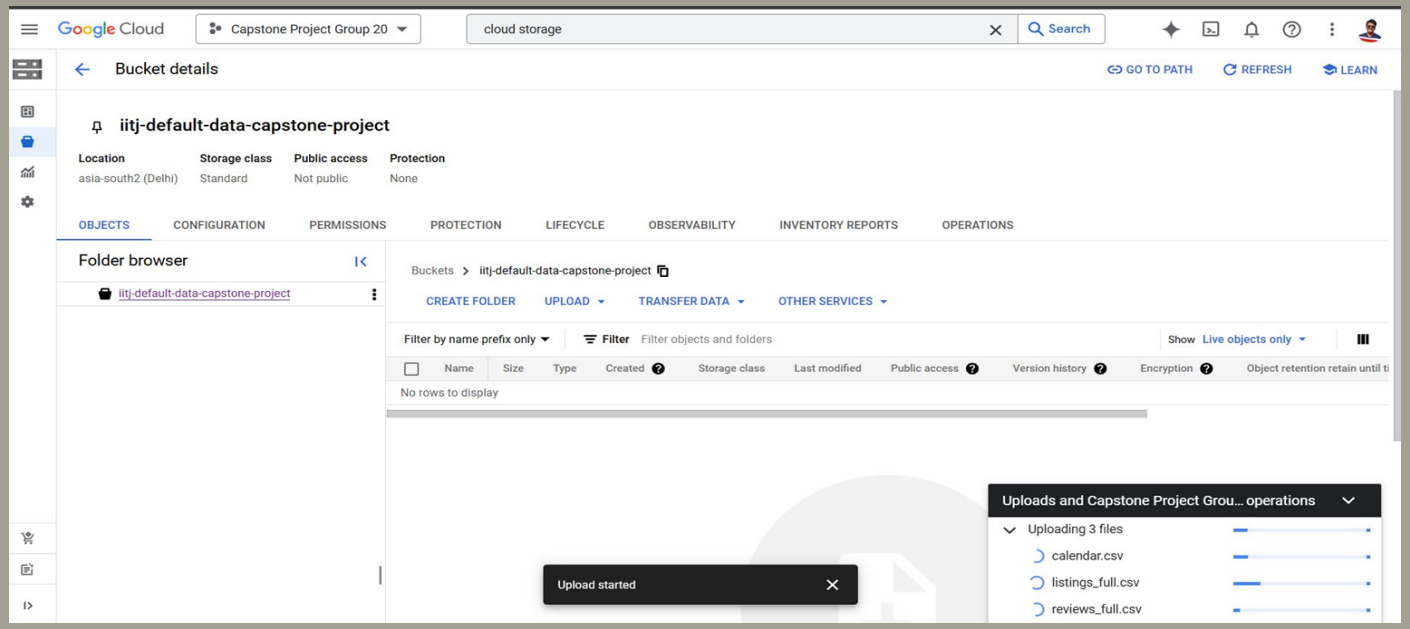


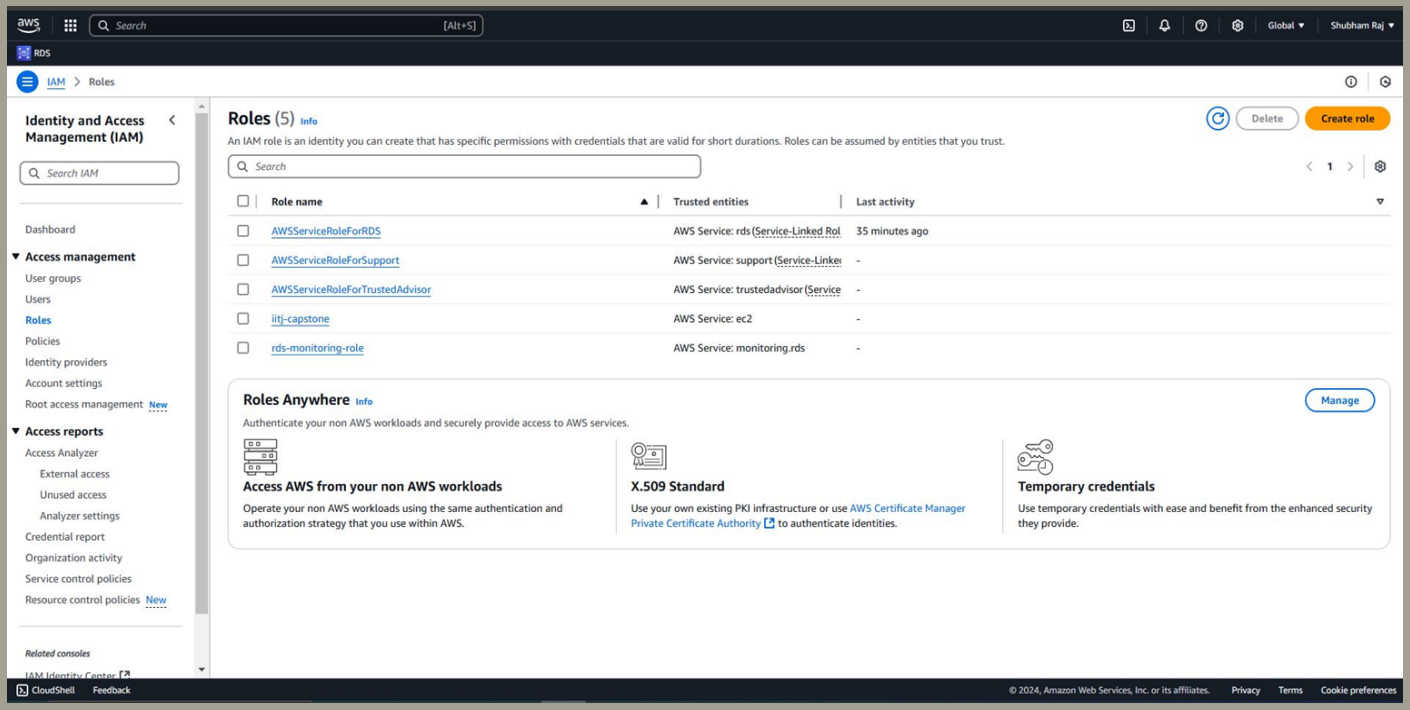




**GITHUB URL:**
https://github.com/shubham14p3/IITJ-AIRBNB-DATA-CLEANING-PROCESSING-AND-ADVANCED-ANALYSIS-CAPSTONE-PROJECT

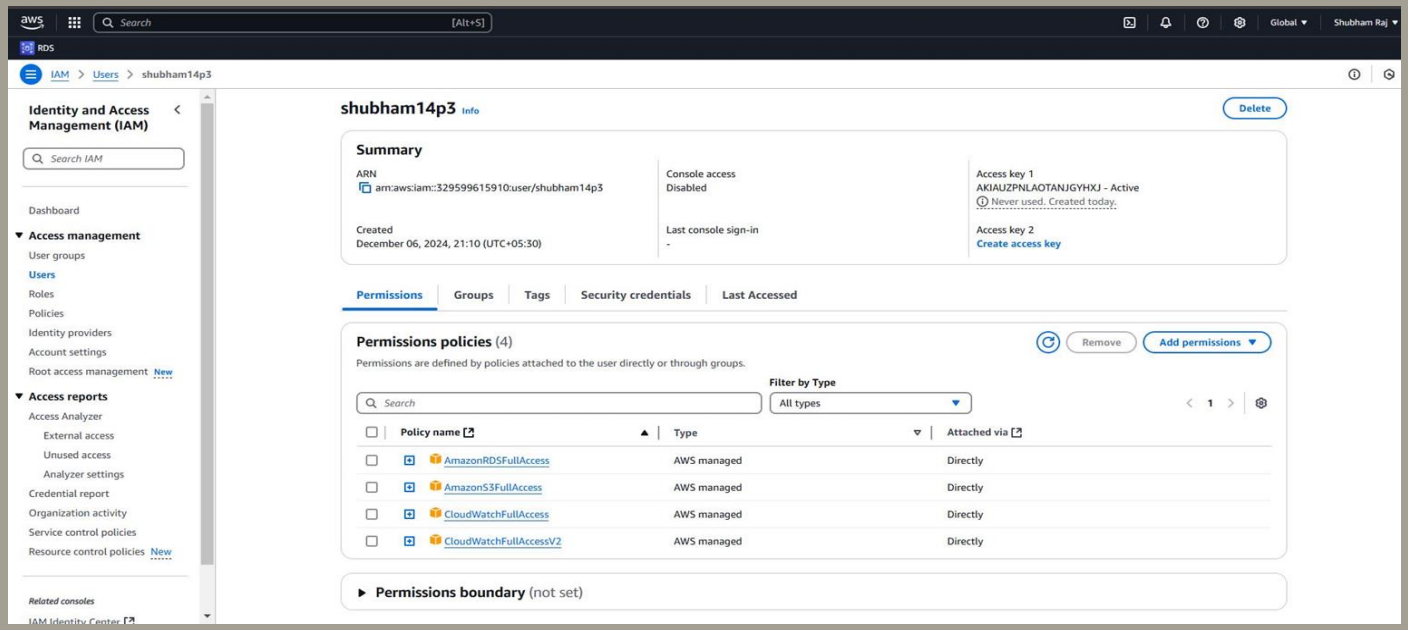

Jira
Your work
Projects
Filters
Dashboards
Teams
Plans
Apps
Create
Upgrade
Search
Apps
Atlassian Marketplace
Manage apps
App requests
Promotions
OAuth credentials
Apps
GitHub Connector
GitHub for Jira
Configure
GitHub configuration
Connect a GitHub organization
Connecting GitHub to Jira allows you to view development activity in the context of your Jira project and issues. To send development data from GitHub to Jira, your team must include issue keys in branch names, commit messages, and pull request titles.
Even if your organization is still backfilling historical data, you can start using issue keys in your development work immediately.
Connected organization
Repository access
Backfill status
Permissions
Settings
shubham14p3
Only select repos
1
FINISHED
Backfilled from: 6/9/24
FULL ACCESS

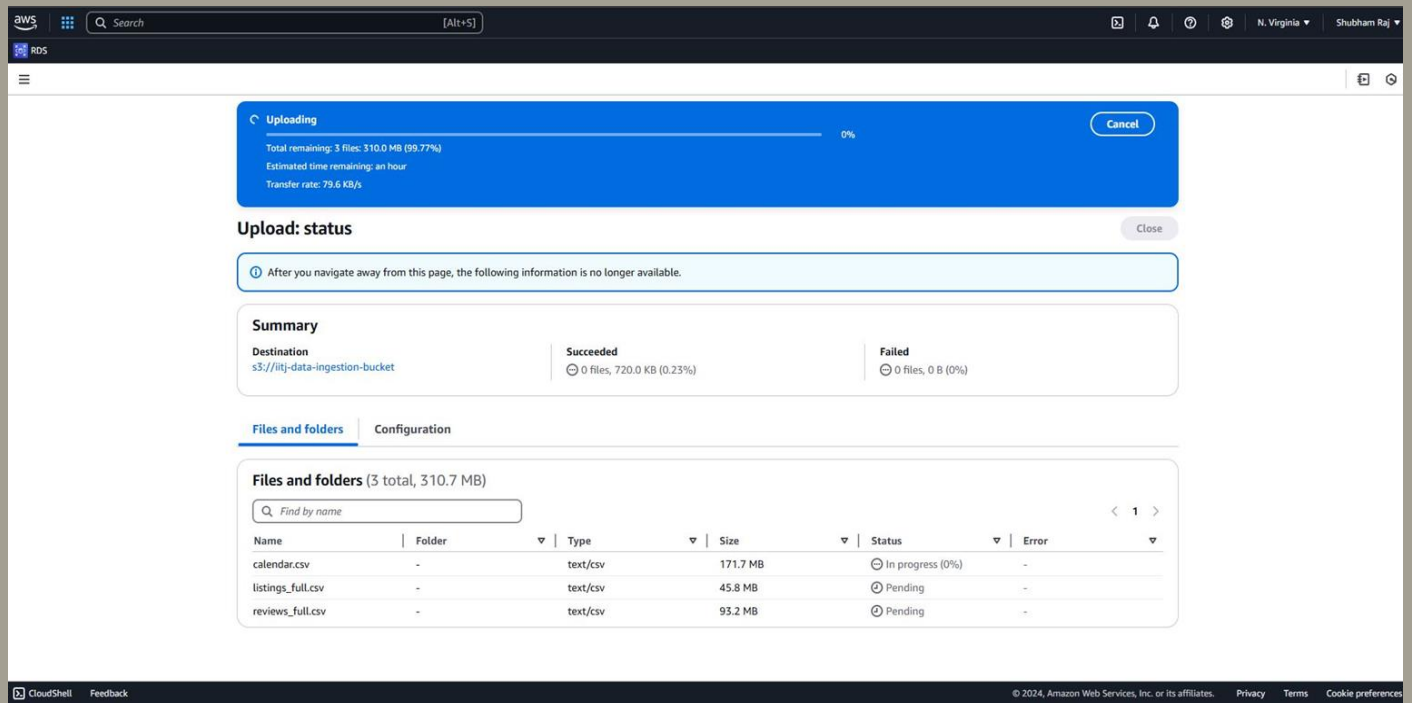
# SCREENSHOTS FROM GOOGLE CLOUD AND AWS (RDS)

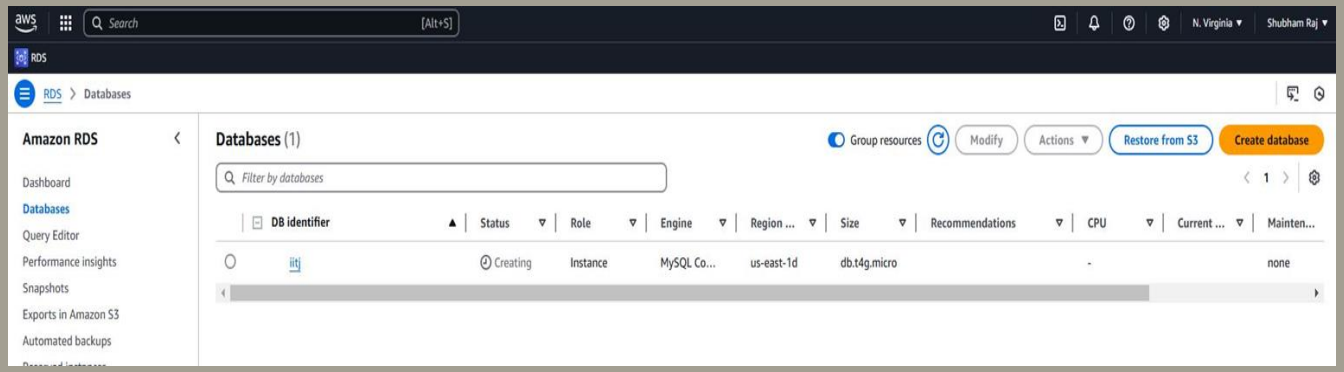


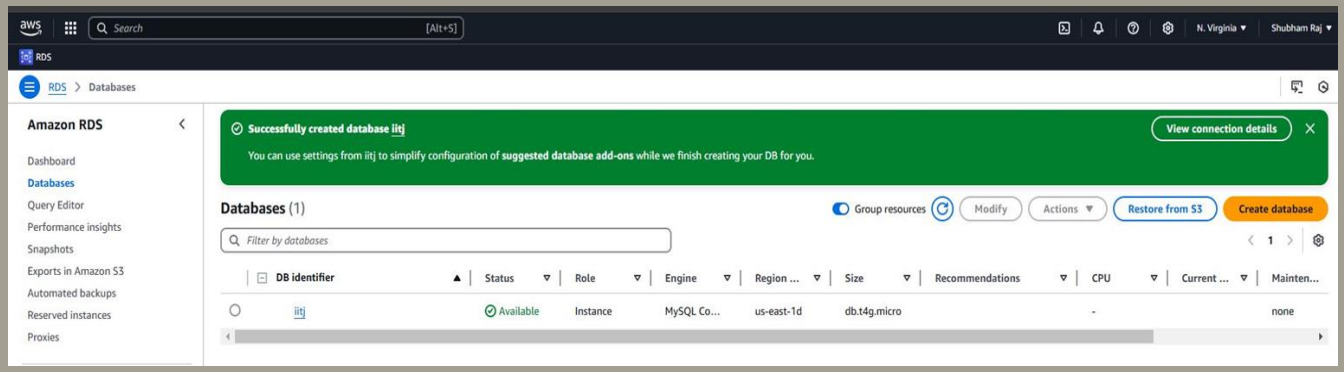



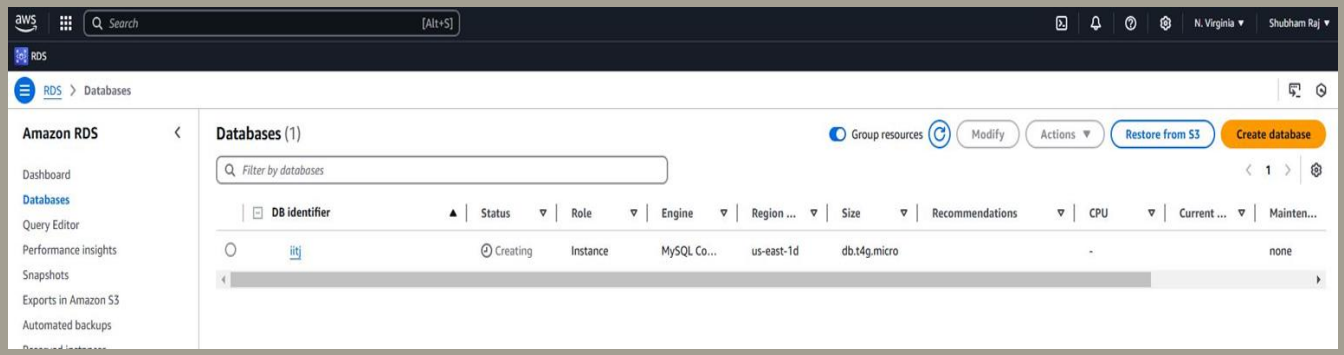
# SCREENSHOTS FROM GOOGLE CLOUD AND AWS (RDS)

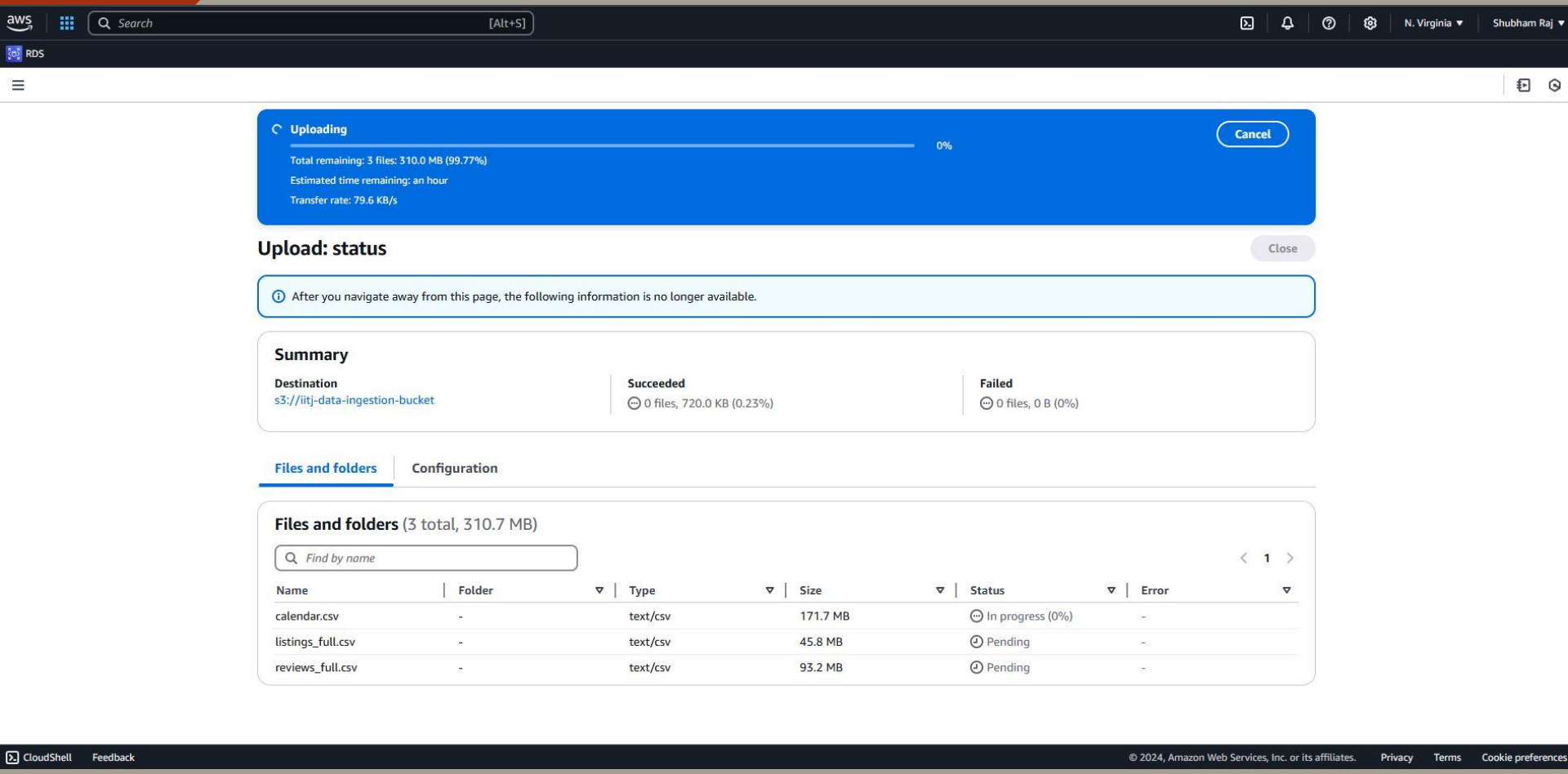


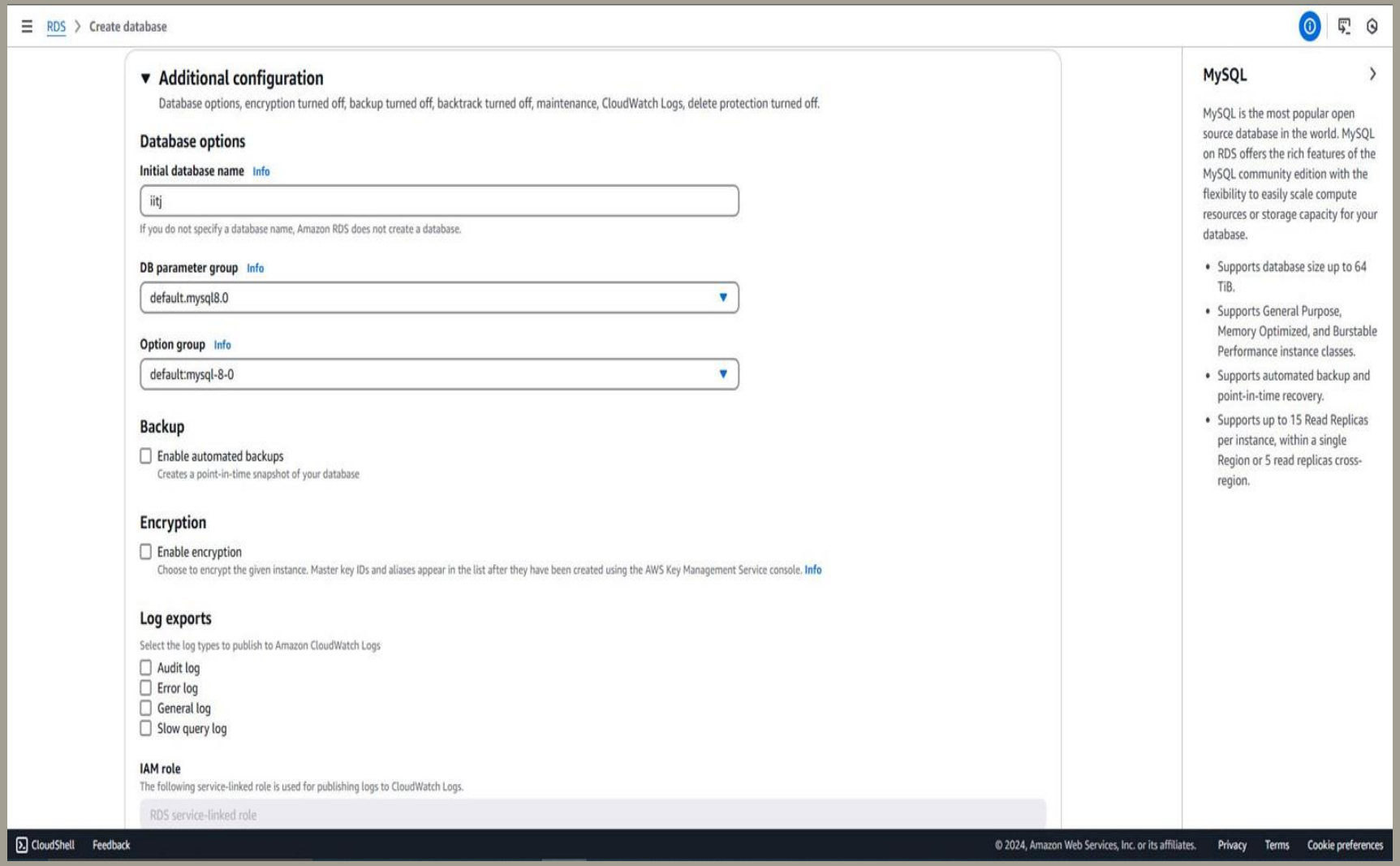

Search
[Alt+S]
N. Virginia
Shubham Raj
RDS
RDS > Databases
Amazon RDS
Dashboard
Databases
Query Editor
Performance insights
Snapshots
Exports in Amazon S3
Automated backups
Databases (1)
Group resources
Modify
Actions
Restore from S3
Create database
Filter by databases
DB identifier
Status
Role
Engine
Region ...
Size
Recommendations
CPU
Current ...
Mainten...
iitj
Creating
Instance
MySQL Co...
us-east-1d
db.t4g.micro
-
none


Search
[Alt+S]
N. Virginia
Shubham Raj
RDS
RDS > Databases
Amazon RDS
Dashboard
Databases
Query Editor
Performance insights
Snapshots
Exports in Amazon S3
Automated backups
Reserved instances
Proxies
Successfully created database iitj
View connection details
You can use settings from iitj to simplify configuration of suggested database add-ons while we finish creating your DB for you.
Databases (1)
Group resources
Modify
Actions
Restore from S3
Create database
Filter by databases
DB identifier
Status
Role
Engine
Region ...
Size
Recommendations
CPU
Current ...
Mainten...
iitj
Available
Instance
MySQL Co...
us-east-1d
db.t4g.micro
-
none


Search
[Alt+S]
N. Virginia
Shubham Raj
RDS
RDS > Databases
Amazon RDS
Dashboard
Databases
Query Editor
Performance insights
Snapshots
Exports in Amazon S3
Automated backups
Databases (1)
Group resources
Modify
Actions
Restore from S3
Create database
Filter by databases
DB identifier
Status
Role
Engine
Region ...
Size
Recommendations
CPU
Current ...
Mainten...
iitj
Creating
Instance
MySQL Co...
us-east-1d
db.t4g.micro
-
none


Search
[Alt+S]
N. Virginia
Shubham Raj
RDS
Uploading
0%
Cancel
Total remaining: 3 files: 310.0 MB (99.77%)
Estimated time remaining: an hour
Transfer rate: 79.6 KB/s
Upload: status
Close
After you navigate away from this page, the following information is no longer available.
Summary
Destination
s3://iitj-data-ingestion-bucket
Succeeded
0 files, 720.0 KB (0.23%)
Failed
0 files, 0 B (0%)
Files and folders
Configuration
Files and folders (3 total, 310.7 MB)
Find by name
1
Name
Folder
Type
Size
Status
Error
calendar.csv
-
text/csv
171.7 MB
In progress (0%)
-
listings_full.csv
-
text/csv
45.8 MB
Pending
-
reviews_full.csv
-
text/csv
93.2 MB
Pending
-
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

# SCREENSHOTS FROM GOOGLE CLOUD AND AWS (RDS)



## Additional configuration

Database options, encryption turned off, backup turned off, backtrack turned off, maintenance, CloudWatch Logs, delete protection turned off.

### Database options

**Initial database name** Info

iitj

If you do not specify a database name, Amazon RDS does not create a database.

**DB parameter group** Info

default.mysql8.0

**Option group** Info

default:mysql-8-0

### Backup

Enable automated backups
Creates a point-in-time snapshot of your database

### Encryption

Enable encryption
Choose to encrypt the given instance. Master key IDs and aliases appear in the list after they have been created using the AWS Key Management Service console. Info

### Log exports

Select the log types to publish to Amazon CloudWatch Logs

- Audit log
- Error log
- General log
- Slow query log

**IAM role**

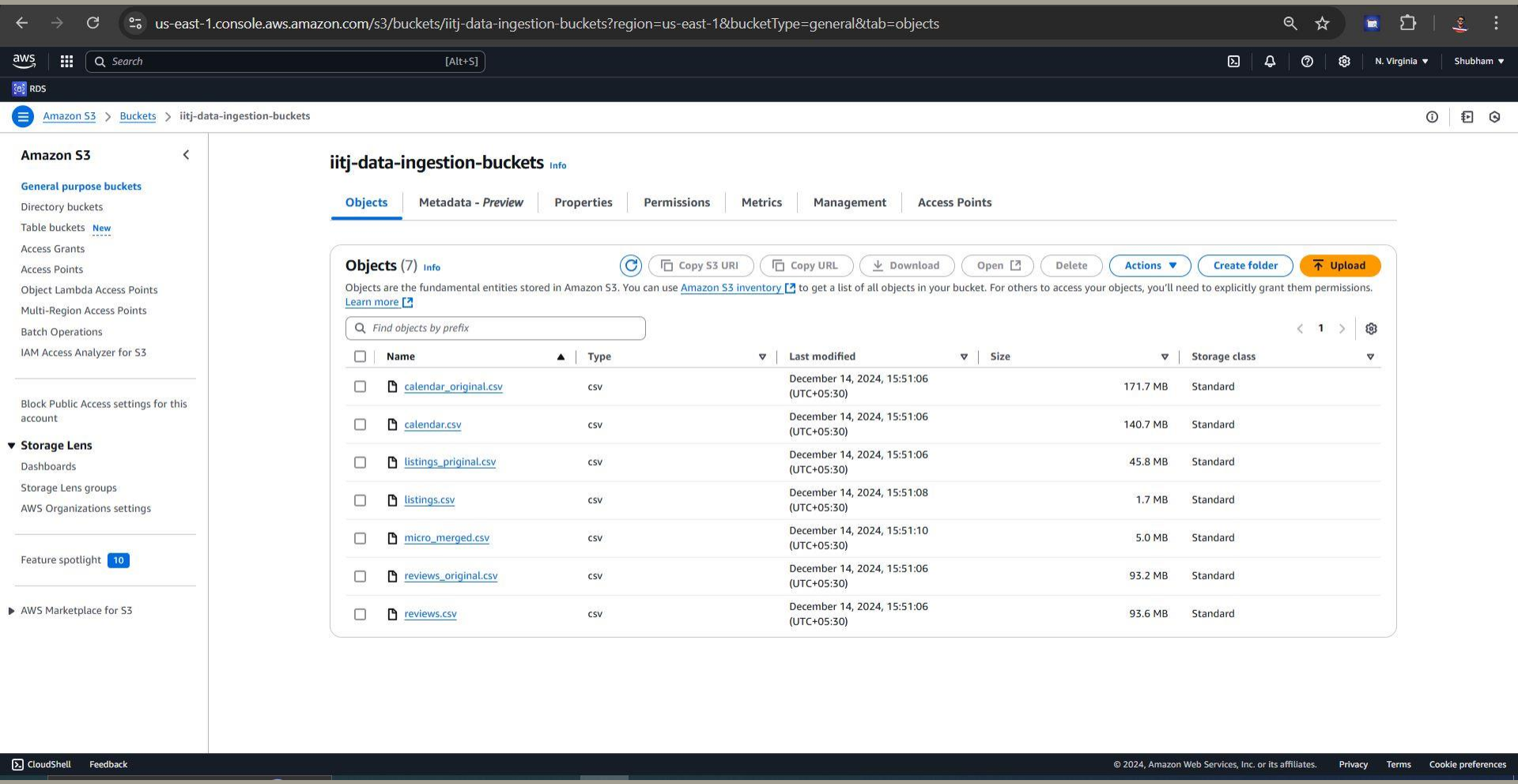
The following service-linked role is used for publishing logs to CloudWatch Logs.

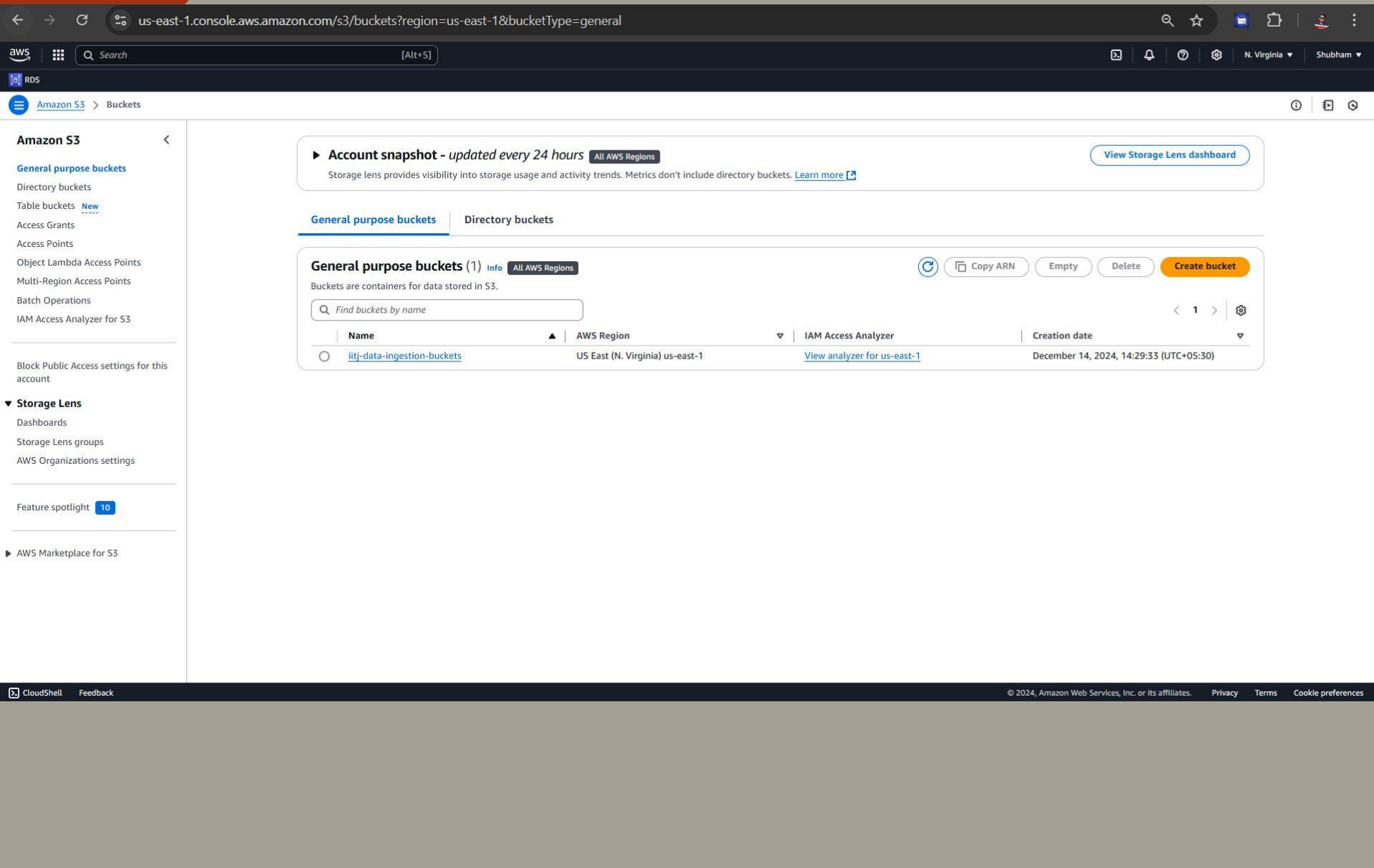
RDS service-linked role

### MySQL

MySQL is the most popular open source database in the world. MySQL on RDS offers the rich features of the MySQL community edition with the flexibility to easily scale compute resources or storage capacity for your database.

- Supports database size up to 64 TiB.
- Supports General Purpose, Memory Optimized, and Burstable Performance instance classes.
- Supports automated backup and point-in-time recovery.
- Supports up to 15 Read Replicas per instance, within a single Region or 5 read replicas cross-region.


us-east-1.console.aws.amazon.com/s3/buckets/iitj-data-ingestion-buckets?region=us-east-1&bucketType=general&tab=objects
Search [Alt+S]
N. Virginia
Shubham
RDS
Amazon S3 > Buckets > iitj-data-ingestion-buckets
Amazon S3
General purpose buckets
Directory buckets
Table buckets New
Access Grants
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
IAM Access Analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
Storage Lens groups
AWS Organizations settings
Feature spotlight 10
AWS Marketplace for S3
iitj-data-ingestion-buckets Info
Objects
Metadata - Preview
Properties
Permissions
Metrics
Management
Access Points
Objects (7) Info
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions.
Learn more
Find objects by prefix
1
Name
Type
Last modified
Size
Storage class
calendar_original.csv csv December 14, 2024, 15:51:06 (UTC+05:30) 171.7 MB Standard
calendar.csv csv December 14, 2024, 15:51:06 (UTC+05:30) 140.7 MB Standard
listings_priginal.csv csv December 14, 2024, 15:51:06 (UTC+05:30) 45.8 MB Standard
listings.csv csv December 14, 2024, 15:51:08 (UTC+05:30) 1.7 MB Standard
micro_merged.csv csv December 14, 2024, 15:51:10 (UTC+05:30) 5.0 MB Standard
reviews_original.csv csv December 14, 2024, 15:51:06 (UTC+05:30) 93.2 MB Standard
reviews.csv csv December 14, 2024, 15:51:06 (UTC+05:30) 93.6 MB Standard
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


us-east-1.console.aws.amazon.com/s3/buckets?region=us-east-1&bucketType=general
Search
[Alt+S]
N. Virginia
Shubham
RDS
Amazon S3
Buckets
Amazon S3
General purpose buckets
Directory buckets
Table buckets
New
Access Grants
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
IAM Access Analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
Storage Lens groups
AWS Organizations settings
Feature spotlight
10
AWS Marketplace for S3
Account snapshot - updated every 24 hours
All AWS Regions
Storage lens provides visibility into storage usage and activity trends. Metrics don't include directory buckets.
Learn more
View Storage Lens dashboard
General purpose buckets
Directory buckets
General purpose buckets (1)
Info
All AWS Regions
Buckets are containers for data stored in S3.
Copy ARN
Empty
Delete
Create bucket
Find buckets by name
1
Name
AWS Region
IAM Access Analyzer
Creation date
iitj-data-ingestion-buckets
US East (N. Virginia) us-east-1
View analyzer for us-east-1
December 14, 2024, 14:29:33 (UTC+05:30)
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

# SCREENSHOTS FROM WEBSITE HOSTED

## Availability Over Time

- This line chart shows the availability of listings over time.
- X-axis: Availability (days), Y-axis: Count of listings.
- Helps identify patterns and trends in booking availability.
- Useful for seasonal analysis.


DOWNLOAD CLOSE

## Price vs. Review Scores

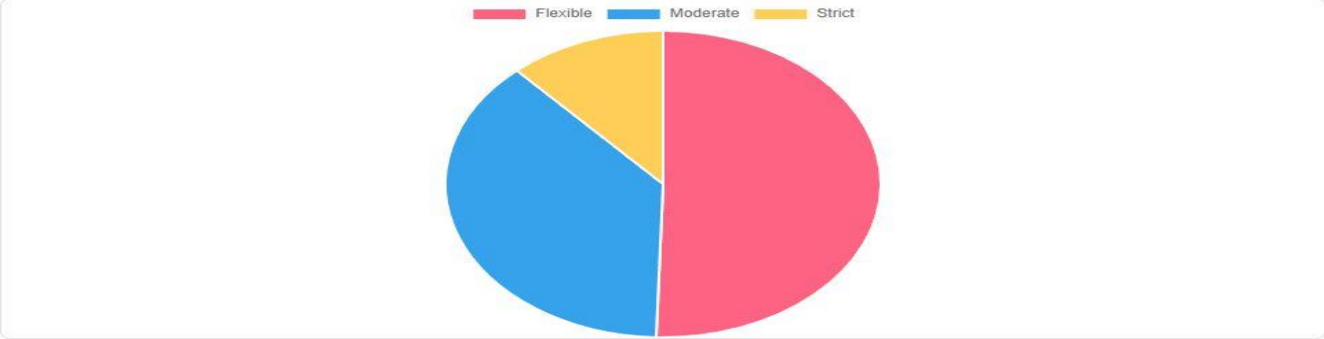
- This scatter plot compares the price of a listing against its review scores.
- X-axis: Price, Y-axis: Review Scores.
- Helps analyze the relationship between pricing and customer satisfaction.
- Useful for performance and pricing analysis.


DOWNLOAD CLOSE

# SCREENSHOTS FROM WEBSITE HOSTED

## Cancellation Policies Distribution

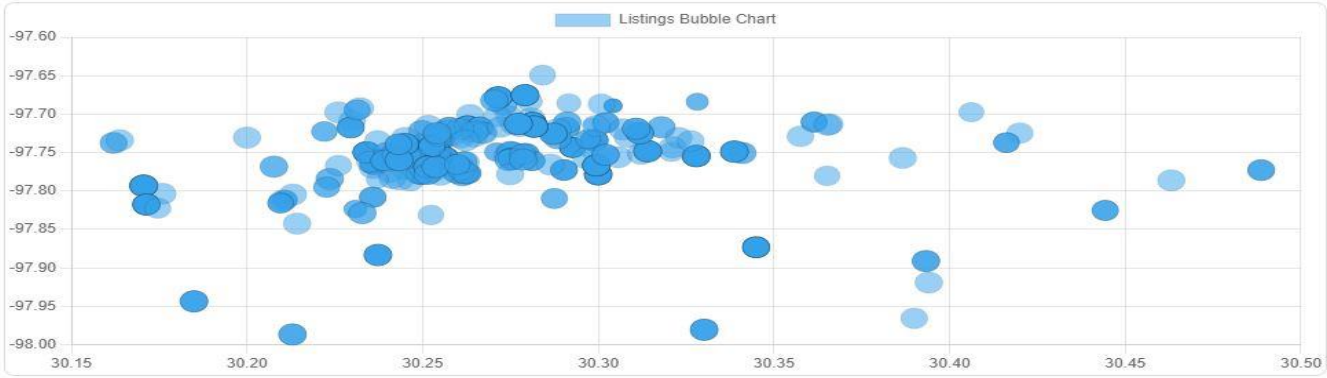
- This pie chart shows the proportion of cancellation policies.
- Policies include Flexible, Moderate, and Strict.
- Helps analyze flexibility offered by hosts.
- Useful for understanding host preferences.




DOWNLOAD CLOSE

## Listings Bubble Chart

- This chart maps the geographical distribution of listings.
- X-axis: Latitude, Y-axis: Longitude.
- Bubble size corresponds to review scores.
- Helps identify clustering of listings.



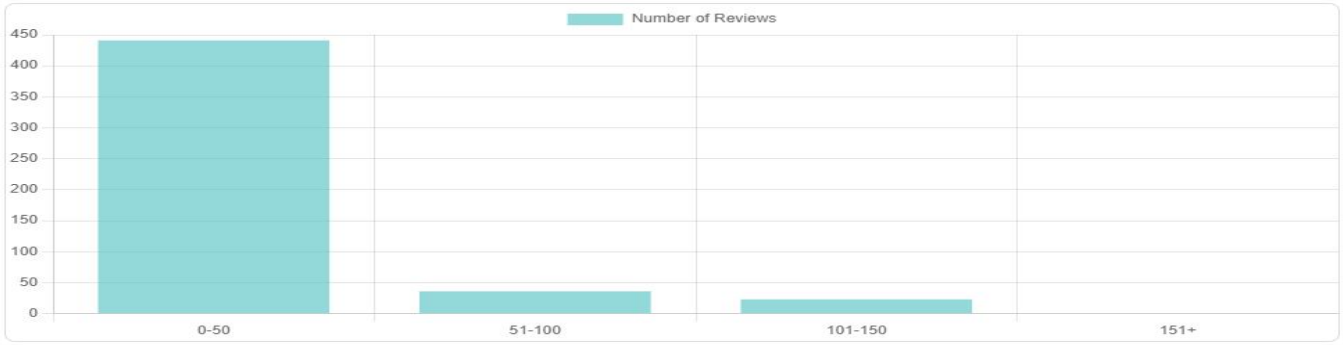

DOWNLOAD CLOSE

# SCREENSHOTS FROM WEBSITE HOSTED

# SCREENSHOTS FROM WEBSITE HOSTED

## Number of Reviews Distribution

- This bar chart shows the frequency of review counts.
- X-axis: Review count buckets (e.g., 0-50, 51-100).
- Y-axis: Count of listings in each range.
- Helps identify highly reviewed properties.



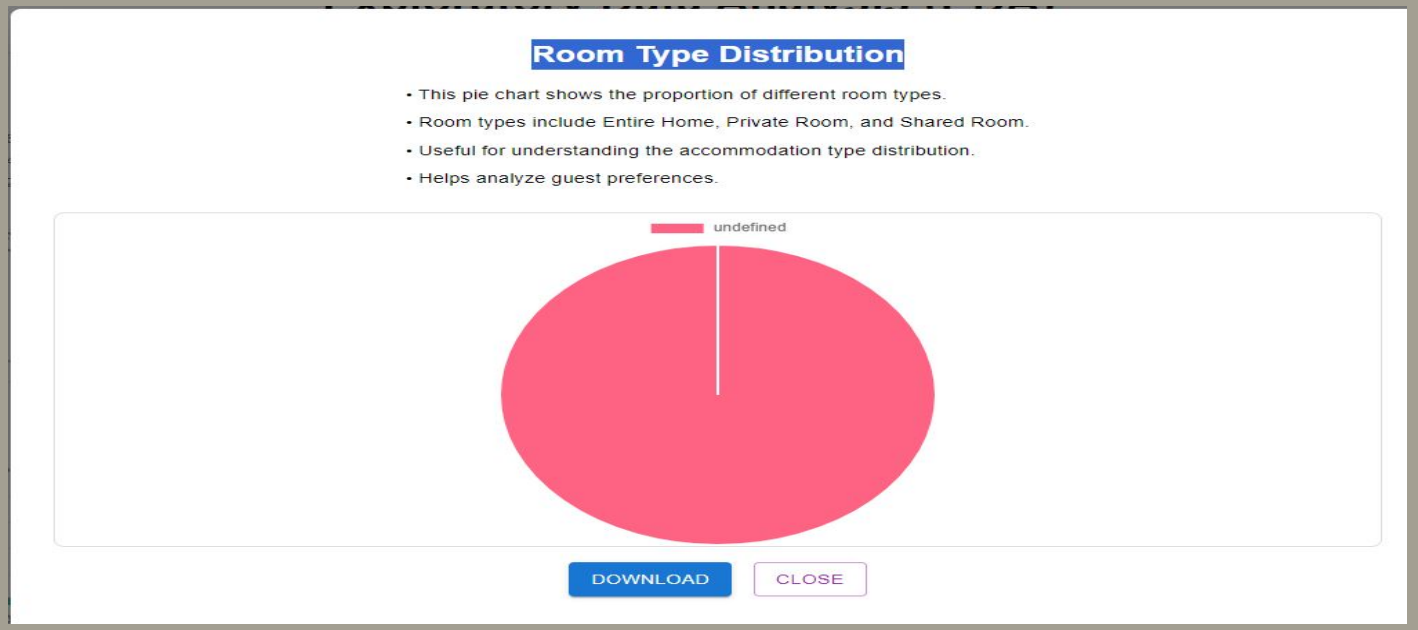

DOWNLOAD CLOSE

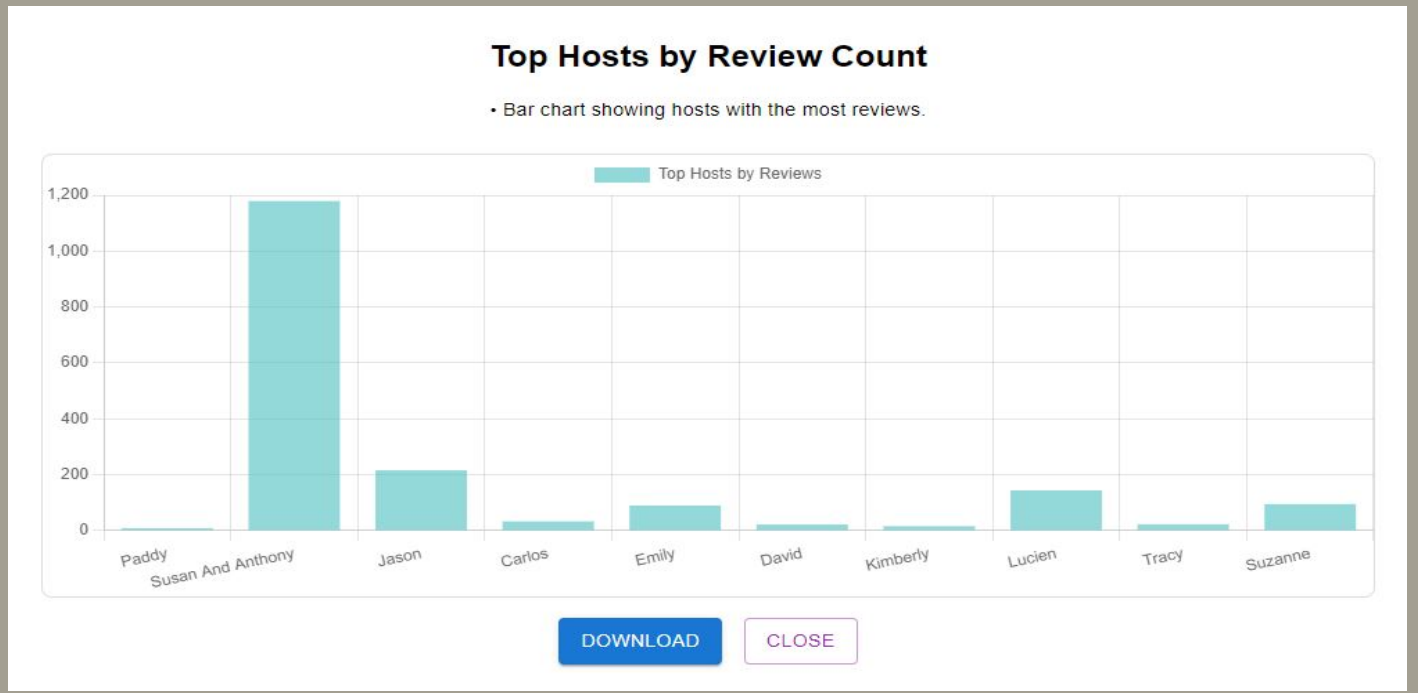
## Price Distribution

- This histogram shows the distribution of prices.
- X-axis: Price ranges (e.g., 0-50, 51-100).
- Y-axis: Count of listings in each range.
- Helps identify common pricing trends.

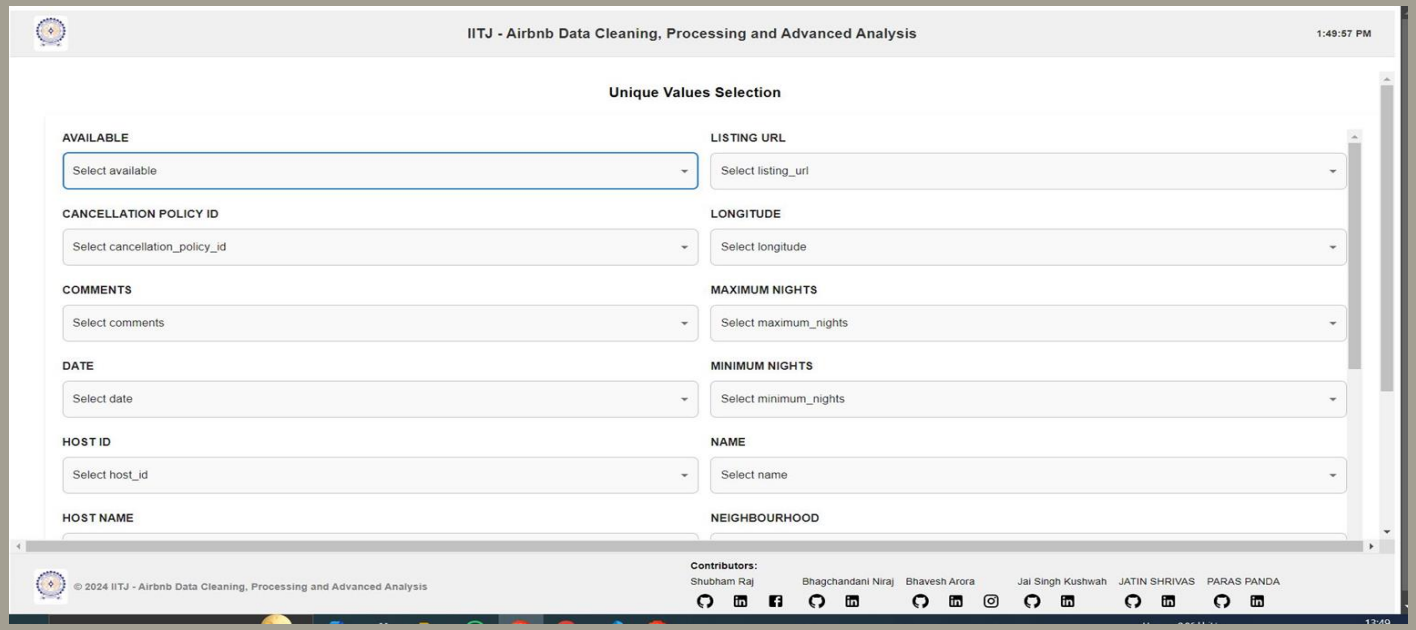

DOWNLOAD CLOSE

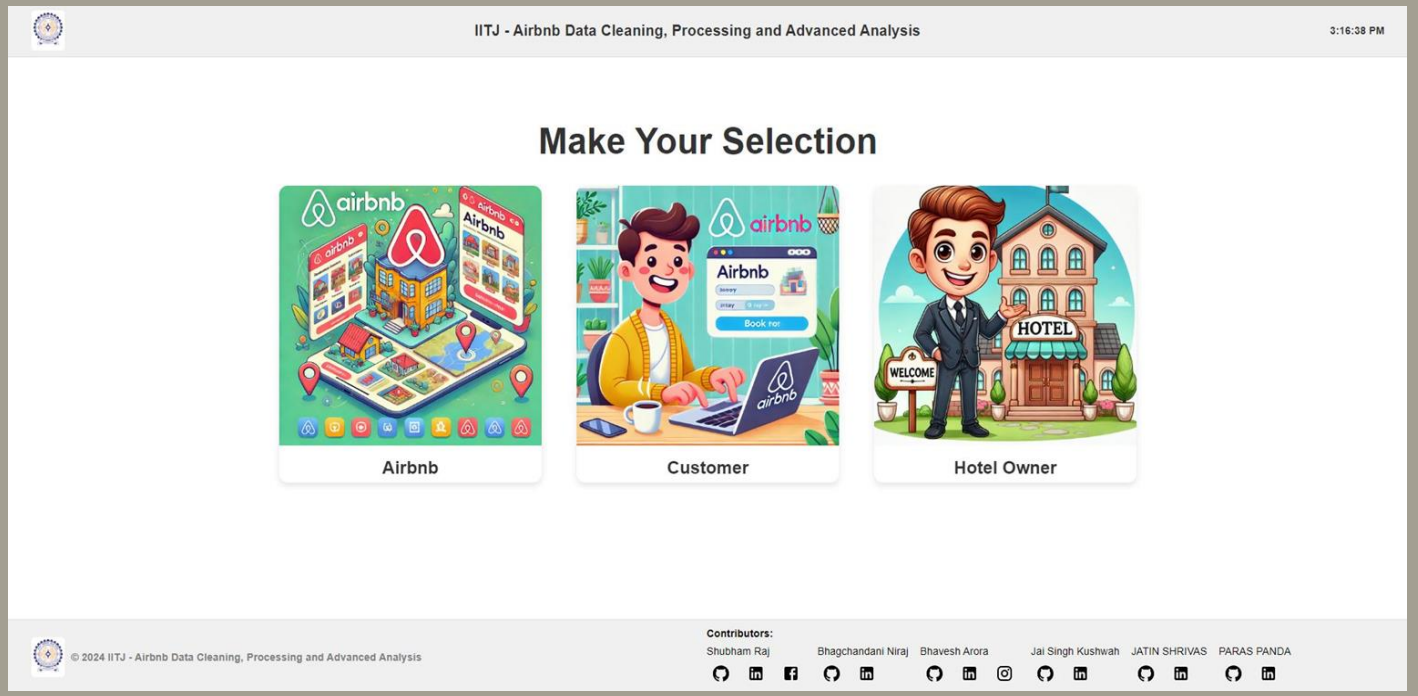
# SCREENSHOTS FROM WEBSITE HOSTED



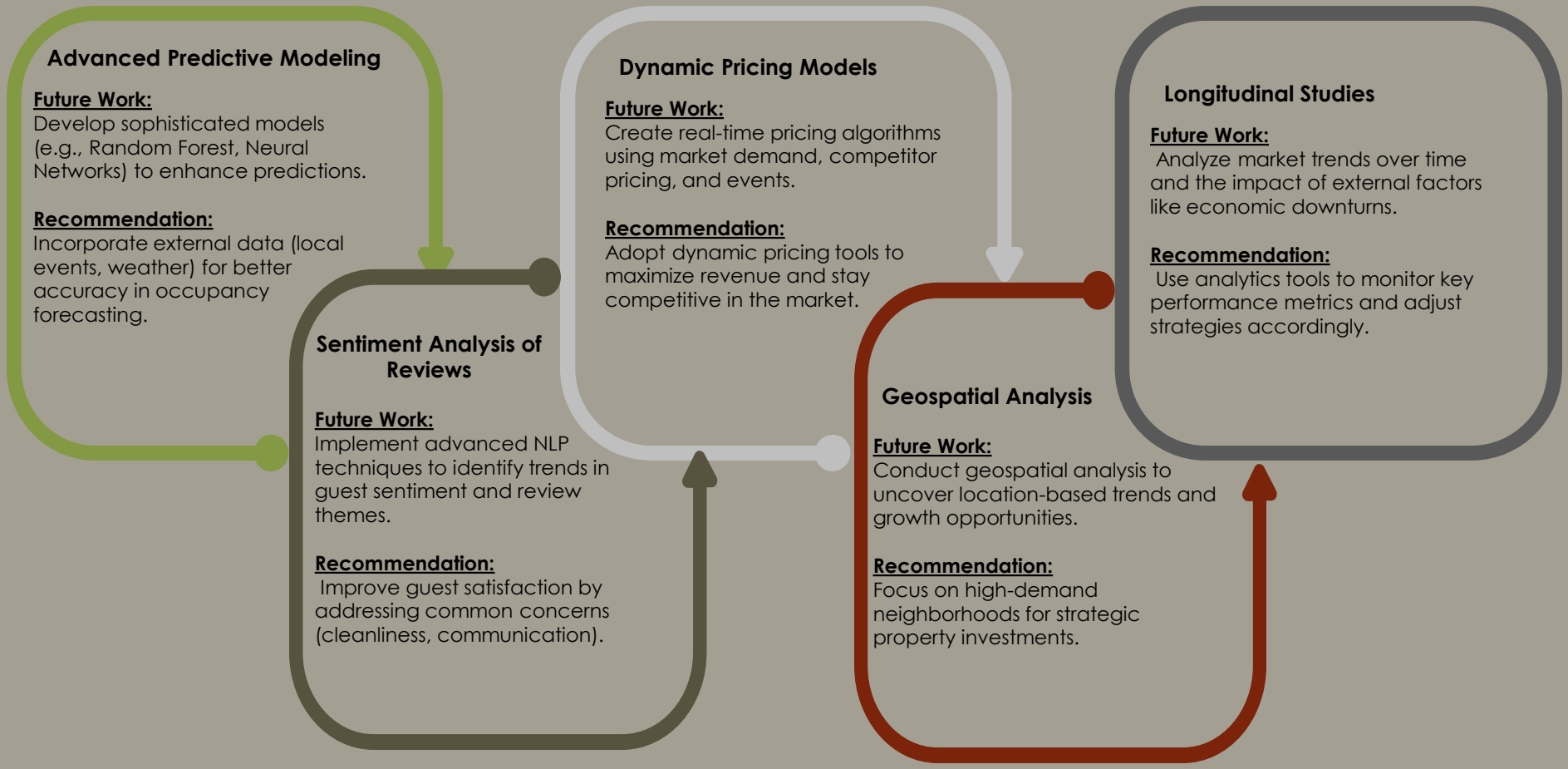
# SCREENSHOTS FROM WEBSITE HOSTED



# FUTURE WORK RECOMMENDATION & CONCLUSION

# FUTURE WORK AND RECOMMENDATION

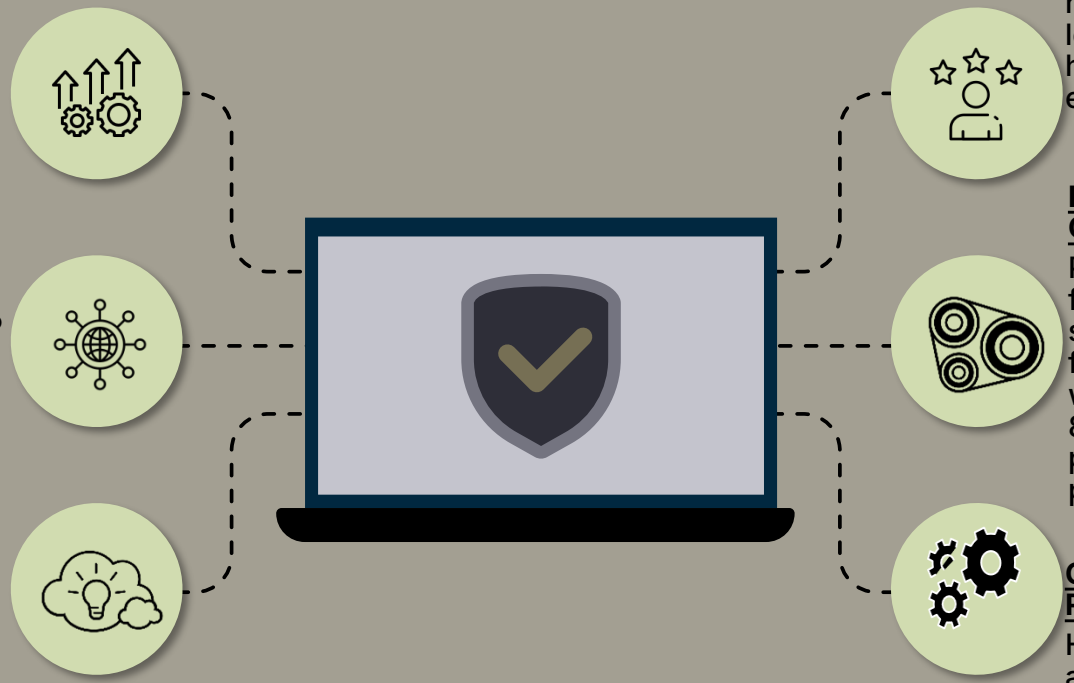


# CONCLUSION

**Dynamic Pricing Optimization:**
Data-driven pricing models have shown a significant correlation between pricing and occupancy rates. By analysing market demand and competitor pricing, hosts can optimize prices and improve occupancy by up to 15-20%.

**Guest Sentiment Analysis:**
Sentiment analysis of reviews reveals that listings with positive sentiment scores have 30% higher occupancy rates. Hosts can leverage NLP techniques to identify and address common guest concerns to improve ratings.

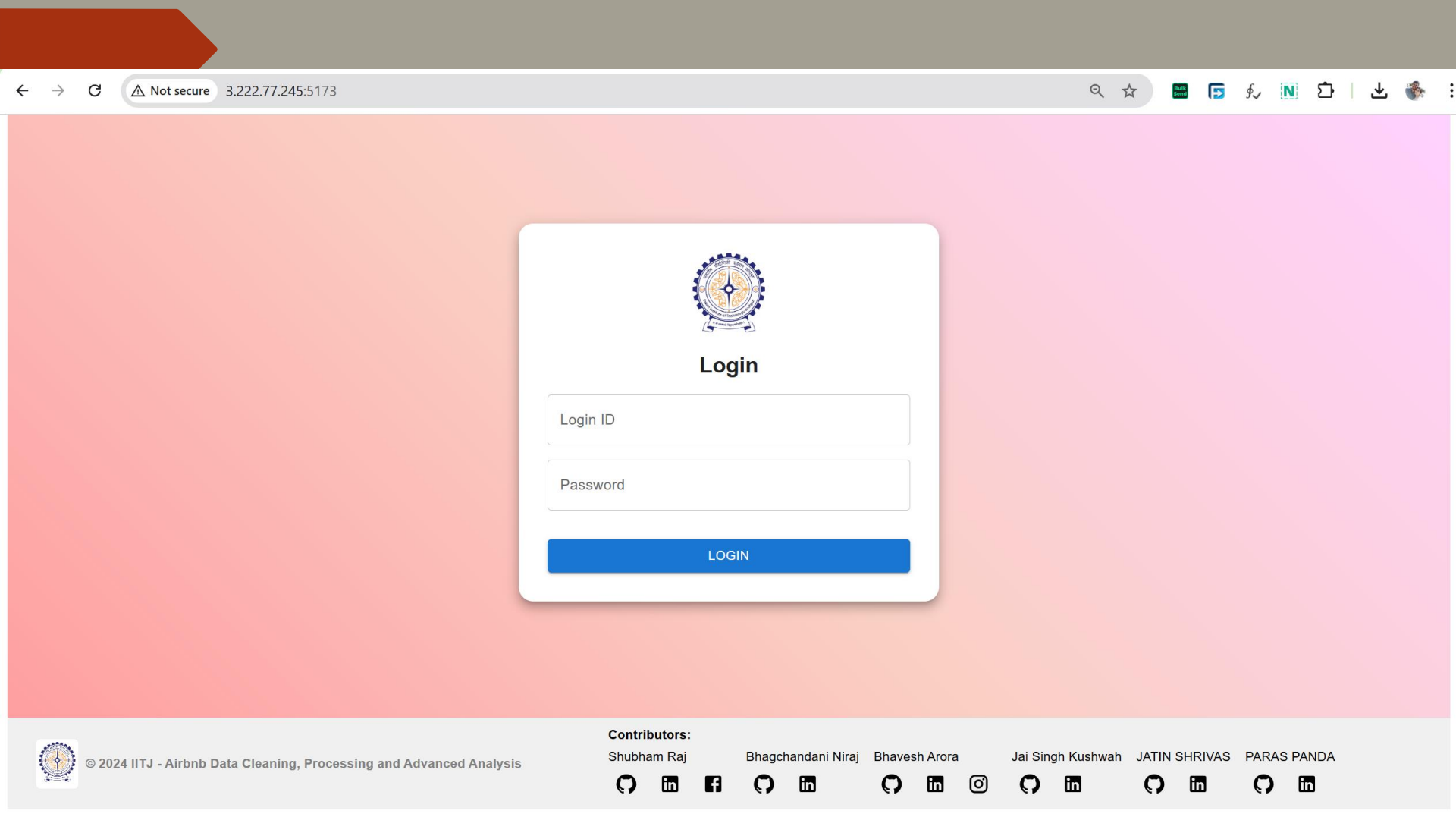
**Room Type Analysis & Segmentation:**
Data shows that entire homes have occupancy rates 25-30% higher compared to private or shared rooms, making them a preferred choice for families and larger groups.



**Geospatial Data for Location Strategy:**
Geospatial analysis indicates that certain neighbourhoods see up to 40% higher demand for rentals, highlighting the value of location-based insights for hosts looking to invest in emerging areas.

**Predictive Analytics for Occupancy Forecasting:**
Predictive models, utilizing features like price, review scores, and room type, can forecast occupancy rates with an accuracy of over 85%, enabling better planning and revenue predictions

**Continuous Data Monitoring & Real-Time Adjustments:**
Hosts using real-time analytics dashboards experience a 20-25% improvement in performance metrics, allowing for rapid adjustments to pricing and availability, driving increased bookings.


Not secure
3.222.77.245:5173
Login
Login ID
Password
LOGIN
© 2024 IITJ - Airbnb Data Cleaning, Processing and Advanced Analysis
Contributors:
Shubham Raj
Bhagchandani Niraj
Bhavesh Arora
Jai Singh Kushwah
JATIN SHRIVAS
PARAS PANDA

Not secure 3.222.77.245:5173

**Indian Institute of Technology Jodhpur**

Date: 12/17/2024

Project Report on

AIRBNB DATA CLEANING, PROCESSING, AND ADVANCED ANALYSIS

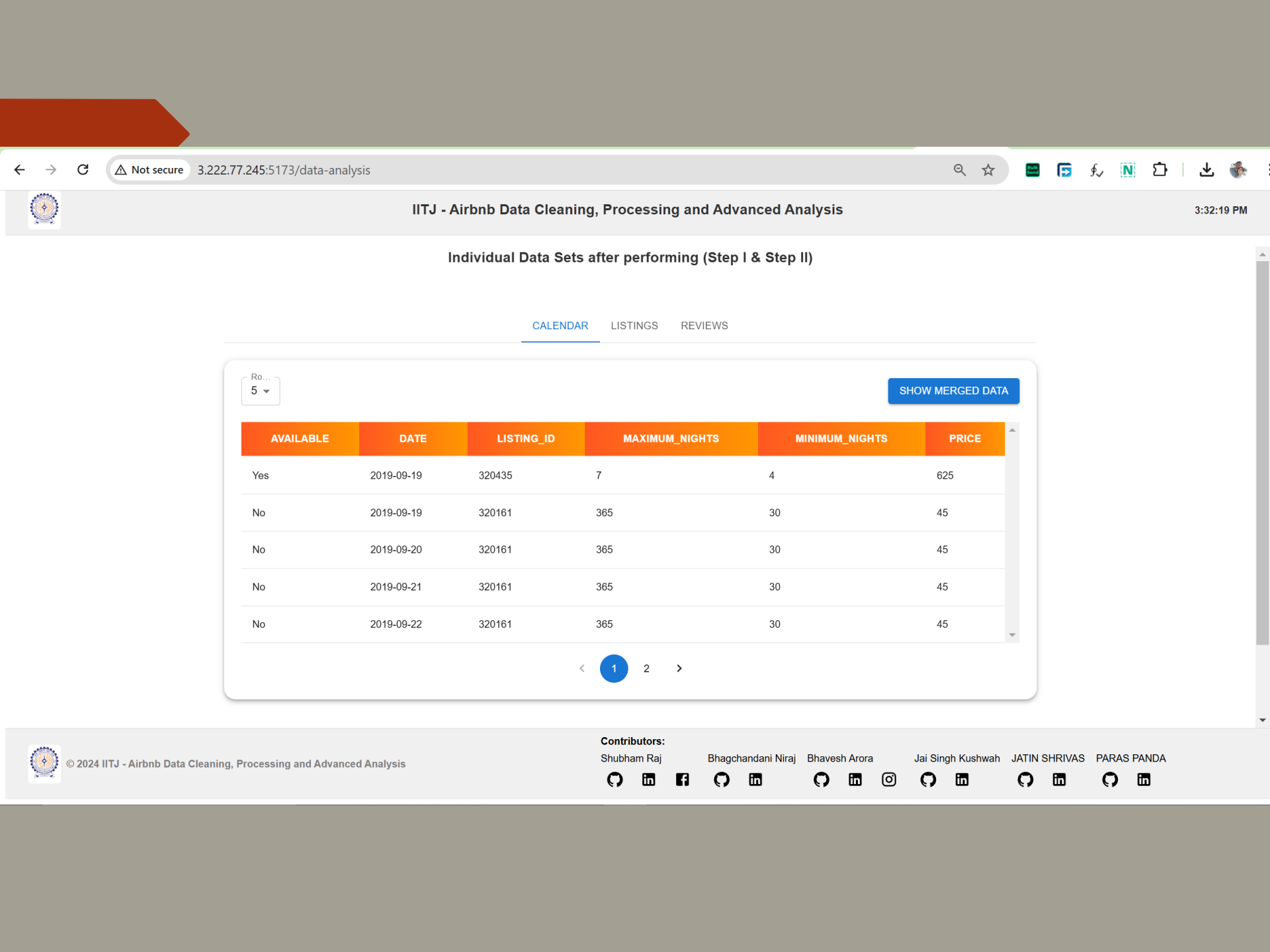
## Welcome to
# AIRBNB DATA CLEANING, PROCESSING, AND ADVANCED ANALYSIS

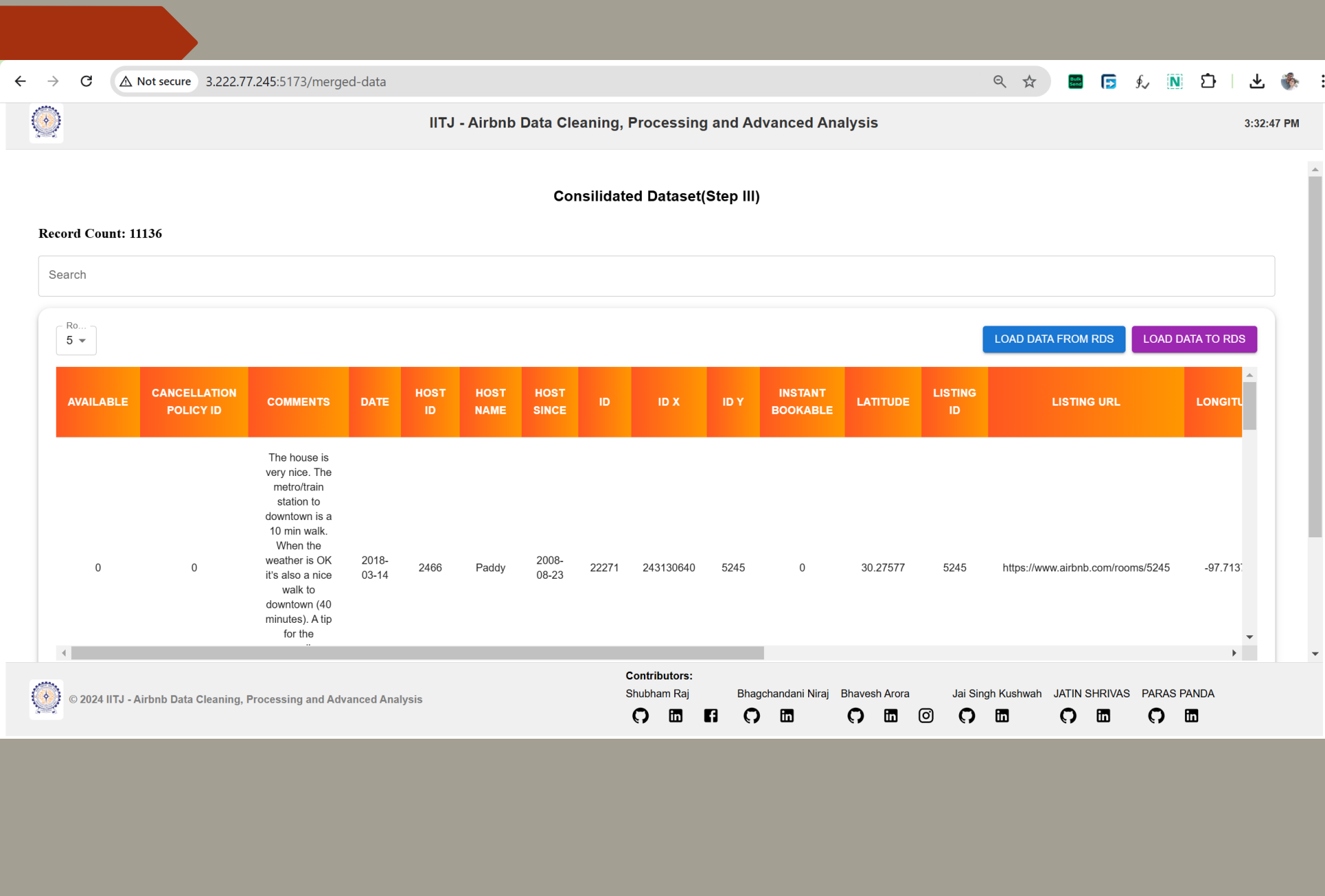
START

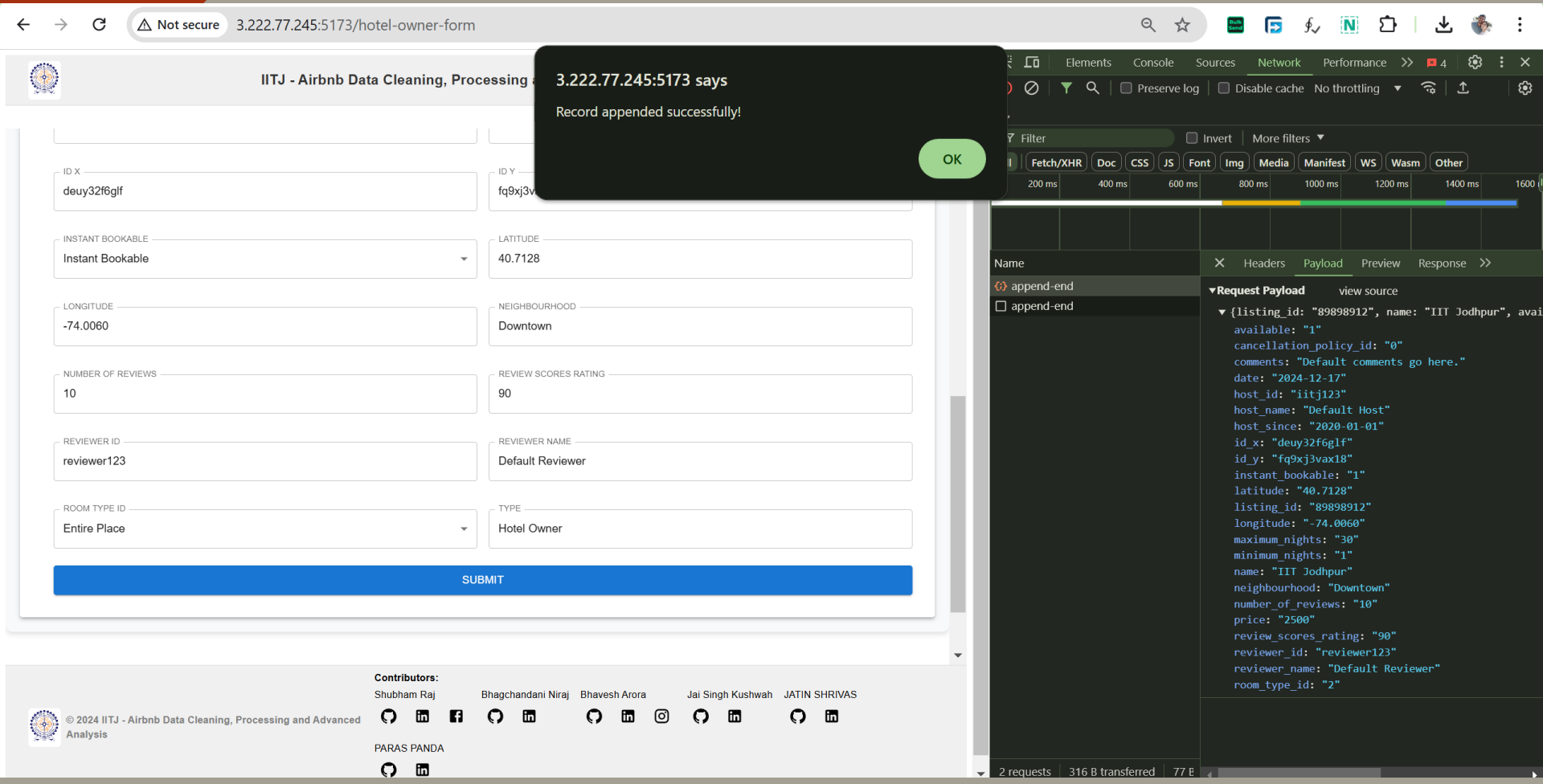


**Contributors:**

Shubham Raj Bhagchandani Niraj Bhavesh Arora Jai Singh Kushwah JATIN SHRIVAS PARAS PANDA


Not secure
3.222.77.245:5173/data-analysis
IITJ - Airbnb Data Cleaning, Processing and Advanced Analysis
3:32:19 PM
Individual Data Sets after performing (Step I & Step II)
CALENDAR
LISTINGS
REVIEWS
Ro...
5
SHOW MERGED DATA
AVAILABLE
DATE
LISTING_ID
MAXIMUM_NIGHTS
MINIMUM_NIGHTS
PRICE
Yes 2019-09-19 320435 7 4 625
No 2019-09-19 320161 365 30 45
No 2019-09-20 320161 365 30 45
No 2019-09-21 320161 365 30 45
No 2019-09-22 320161 365 30 45
1
2
© 2024 IITJ - Airbnb Data Cleaning, Processing and Advanced Analysis
Contributors:
Shubham Raj
Bhagchandani Niraj
Bhavesh Arora
Jai Singh Kushwah
JATIN SHRIVAS
PARAS PANDA


Not secure
3.222.77.245:5173/merged-data
IITJ - Airbnb Data Cleaning, Processing and Advanced Analysis
3:32:47 PM
Consilidated Dataset(Step III)
Record Count: 11136
Search
Ro...
5
LOAD DATA FROM RDS
LOAD DATA TO RDS
AVAILABLE
CANCELLATION POLICY ID
COMMENTS
DATE
HOST ID
HOST NAME
HOST SINCE
ID
ID X
ID Y
INSTANT BOOKABLE
LATITUDE
LISTING ID
LISTING URL
LONGITU
0
0
The house is very nice. The metro/train station to downtown is a 10 min walk. When the weather is OK it's also a nice walk to downtown (40 minutes). A tip for the
2018-03-14
2466
Paddy
2008-08-23
22271
243130640
5245
0
30.27577
5245
https://www.airbnb.com/rooms/5245
-97.713
© 2024 IITJ - Airbnb Data Cleaning, Processing and Advanced Analysis
Contributors:
Shubham Raj
Bhagchandani Niraj
Bhavesh Arora
Jai Singh Kushwah
JATIN SHRIVAS
PARAS PANDA


Not secure
3.222.77.245:5173/hotel-owner-form
IITJ - Airbnb Data Cleaning, Processing
3.222.77.245:5173 says
Record appended successfully!
OK
ID X
deuy32f6glf
ID Y
INSTANT BOOKABLE
Instant Bookable
LATITUDE
40.7128
LONGITUDE
-74.0060
NEIGHBOURHOOD
Downtown
NUMBER OF REVIEWS
10
REVIEW SCORES RATING
90
REVIEWER ID
reviewer123
REVIEWER NAME
Default Reviewer
ROOM TYPE ID
Entire Place
TYPE
Hotel Owner
SUBMIT
Contributors:
Shubham Raj
Bhagchandani Niraj
Bhavesh Arora
Jai Singh Kushwah
JATIN SHRIVAS
PARAS PANDA
© 2024 IITJ - Airbnb Data Cleaning, Processing and Advanced Analysis
Elements
Console
Sources
Network
Performance
Preserve log
Disable cache
No throttling
Filter
Invert
More filters
Fetch/XHR
Doc
CSS
JS
Font
Img
Media
Manifest
WS
Wasm
Other
Name
Headers
Payload
Preview
Response
append-end
append-end
Request Payload
view source
{listing_id: "89898912", name: "IIT Jodhpur", avai
available: "1"
cancellation_policy_id: "0"
comments: "Default comments go here."
date: "2024-12-17"
host_id: "iitj123"
host_name: "Default Host"
host_since: "2020-01-01"
id_x: "deuy32f6glf"
id_y: "fq9xj3vax18"
instant_bookable: "1"
latitude: "40.7128"
listing_id: "89898912"
longitude: "-74.0060"
maximum_nights: "30"
minimum_nights: "1"
name: "IIT Jodhpur"
neighbourhood: "Downtown"
number_of_reviews: "10"
price: "2500"
review_scores_rating: "90"
reviewer_id: "reviewer123"
reviewer_name: "Default Reviewer"
room_type_id: "2"
2 requests
316 B transferred
77 B

# Visualisation from RDS

Listings Per Neighborhood


EXPAND

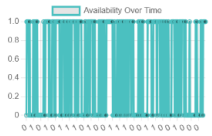
Room Type Distribution


EXPAND

Price vs. Review Scores

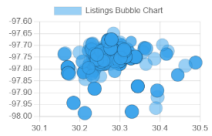

EXPAND

Availability Over Time




EXPAND

Listings Bubble Chart




EXPAND

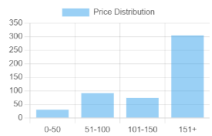
Cancellation Policies Distribution


EXPAND

Number of Reviews Distribution

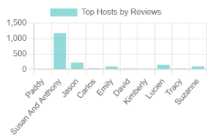

EXPAND

Price Distribution




EXPAND

Top Hosts by Review Count




EXPAND

BACK

NEXT



Contributors:
Shubham Raj
Bhagchandani Niraj
Bhavesh Arora
Jai Singh Kushwah
JATIN SHRIVAS
PARAS PANDA

## Categorical Data Distribution

Select cancellation policy id

Select longitude

**COMMENTS**

Select comments

**MAXIMUM NIGHTS**

Select maximum nights

**DATE**

Select date

**MINIMUM NIGHTS**

Select minimum nights

**HOST ID**

Select host id

**NAME**

Select name

**HOST NAME**

Select host name

**NEIGHBOURHOOD**

Select neighbourhood

**HOST SINCE**

Select host since

**NUMBER OF REVIEWS**

Select number of reviews

**ID**

Select id

**PRICE**

Select price

**ID X**

Select id x

**REVIEW SCORES RATING**

Select review scores rating

**ID Y**

Select id y

**REVIEWER ID**

Select reviewer id

**INSTANT BOOKABLE**

Select instant bookable

**REVIEWER NAME**

Select reviewer name

**LATITUDE**

Select latitude

**ROOM TYPE ID**

Select room type id

**LISTING ID**



Contributors:
Shubham Raj
Bhagchandani Niraj
Bhavesh Arora
Jai Singh Kushwah
JATIN SHRIVAS
PARAS PANDA

Not secure 3.222.77.245:5173/hotel-owner-form

# IITJ - Airbnb Data Cleaning, Processing and Advanced Analysis

3:36:02 PM

You have selected Hotel Owner Form

| | |
|---|---|
| LISTING ID: 12345 | NAME: Luxury Apartment |
| AVAILABLE: Available | PRICE: 250 |
| MINIMUM NIGHTS: 1 | MAXIMUM NIGHTS: 30 |
| CANCELLATION POLICY ID: No Policy | COMMENTS: Default comments go here. |
| DATE: 17-12-2024 | HOST ID: host123 |
| HOST NAME: Default Host | HOST SINCE: 2020-01-01 |
| ID X: tfxklq7k5e | ID Y: vh50ueyaf |
| INSTANT BOOKABLE: Not Instant Bookable | LATITUDE: 40.7128 |
| LONGITUDE: -74.0060 | NEIGHBOURHOOD: Downtown |



Contributors:
Shubham Raj Bhagchandani Niraj Bhavesh Arora Jai Singh Kushwah JATIN SHRIVAS PARAS PANDA

Not secure 3.222.77.245:5173/new-merged-data

IITJ - Airbnb Data Cleaning, Processing and Advanced Analysis 3:37:27 PM

## Consilidated Dataset (After Adding New Data)

**Record Count: 11136**

Search

Ro... 5

LOAD DATA FROM RDS | LOAD DATA TO RDS

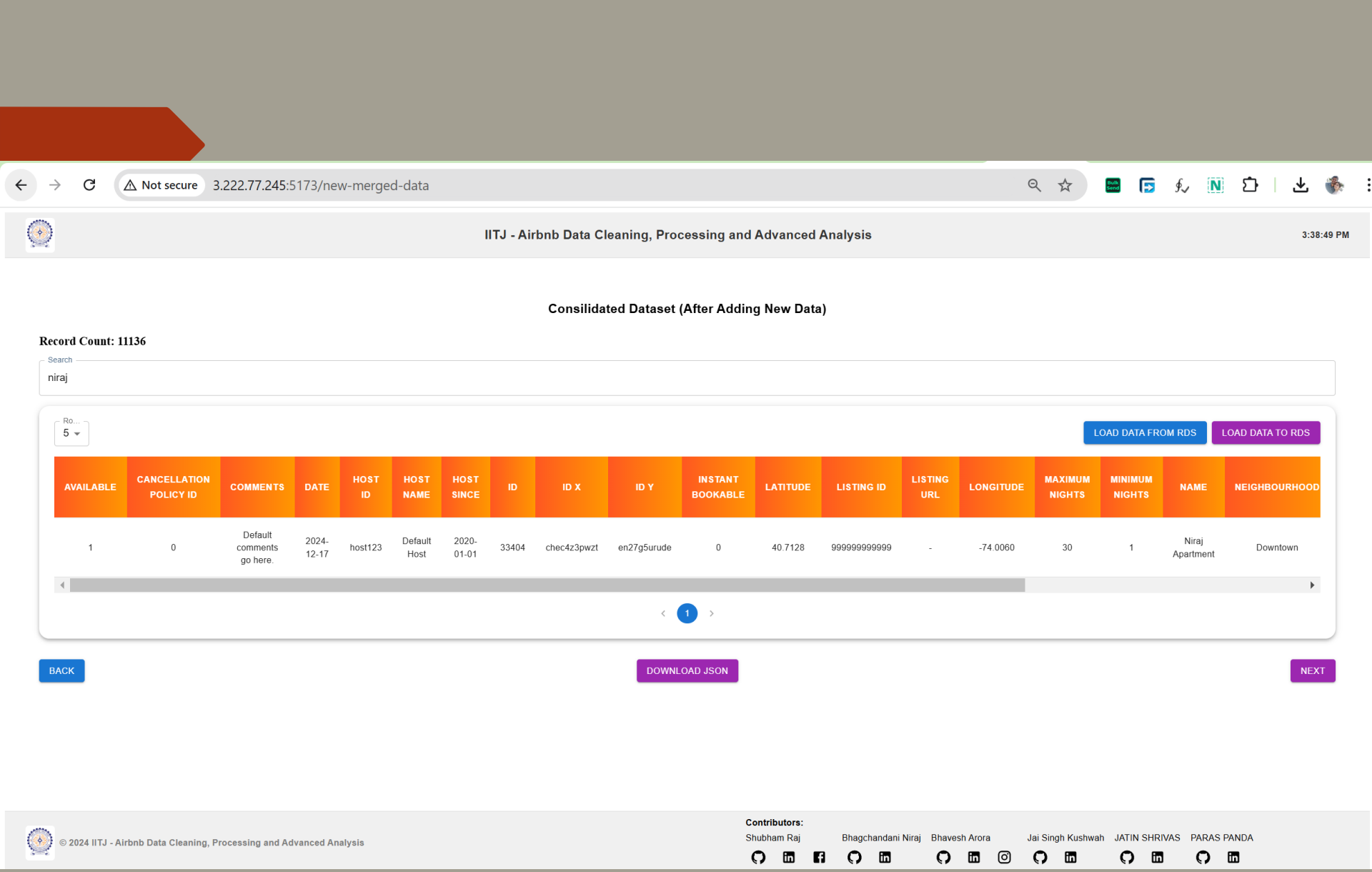
| AVAILABLE | CANCELLATION POLICY ID | COMMENTS | DATE | HOST ID | HOST NAME | HOST SINCE | ID | ID X | ID Y | INSTANT BOOKABLE | LATITUDE | LISTING ID | LISTING URL | LONGITUDE | MAXIMUM NIGHTS | MINIMUM NIGHTS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2 | We planned a last minute trip to Austin and came across Tina's place. We weren't really expecting much but a place to sleep since we would be out and about, but we loved Tina's | 2018-06-14 | 193075619 | Tina | 2018-06-02 | 33386 | 276732119 | 25621580 | 0 | 30.28441 | 25621580 | https://www.airbnb.com/rooms/25621580 | -97.75965 | 365 | 1 |

‹ 1 ... 2223 2224 2225 ... 2228 ›

BACK | DOWNLOAD JSON | NEXT



**Contributors:**
Shubham Raj Bhagchandani Niraj Bhavesh Arora Jai Singh Kushwah JATIN SHRIVAS PARAS PANDA


Not secure
3.222.77.245:5173/new-merged-data
IITJ - Airbnb Data Cleaning, Processing and Advanced Analysis
3:38:49 PM
Consilidated Dataset (After Adding New Data)
Record Count: 11136
Search
niraj
Ro...
5
LOAD DATA FROM RDS
LOAD DATA TO RDS
AVAILABLE
CANCELLATION POLICY ID
COMMENTS
DATE
HOST ID
HOST NAME
HOST SINCE
ID
ID X
ID Y
INSTANT BOOKABLE
LATITUDE
LISTING ID
LISTING URL
LONGITUDE
MAXIMUM NIGHTS
MINIMUM NIGHTS
NAME
NEIGHBOURHOOD
1
0
Default comments go here.
2024-12-17
host123
Default Host
2020-01-01
33404
chec4z3pwzt
en27g5urude
0
40.7128
999999999999
-
-74.0060
30
1
Niraj Apartment
Downtown
1
BACK
DOWNLOAD JSON
NEXT
© 2024 IITJ - Airbnb Data Cleaning, Processing and Advanced Analysis
Contributors:
Shubham Raj
Bhagchandani Niraj
Bhavesh Arora
Jai Singh Kushwah
JATIN SHRIVAS
PARAS PANDA

Not secure 3.222.77.245:5173/prediction

IITJ - Airbnb Data Cleaning, Processing and Advanced Analysis

3:39:11 PM

# DATA COOKING IN PROGRESS...

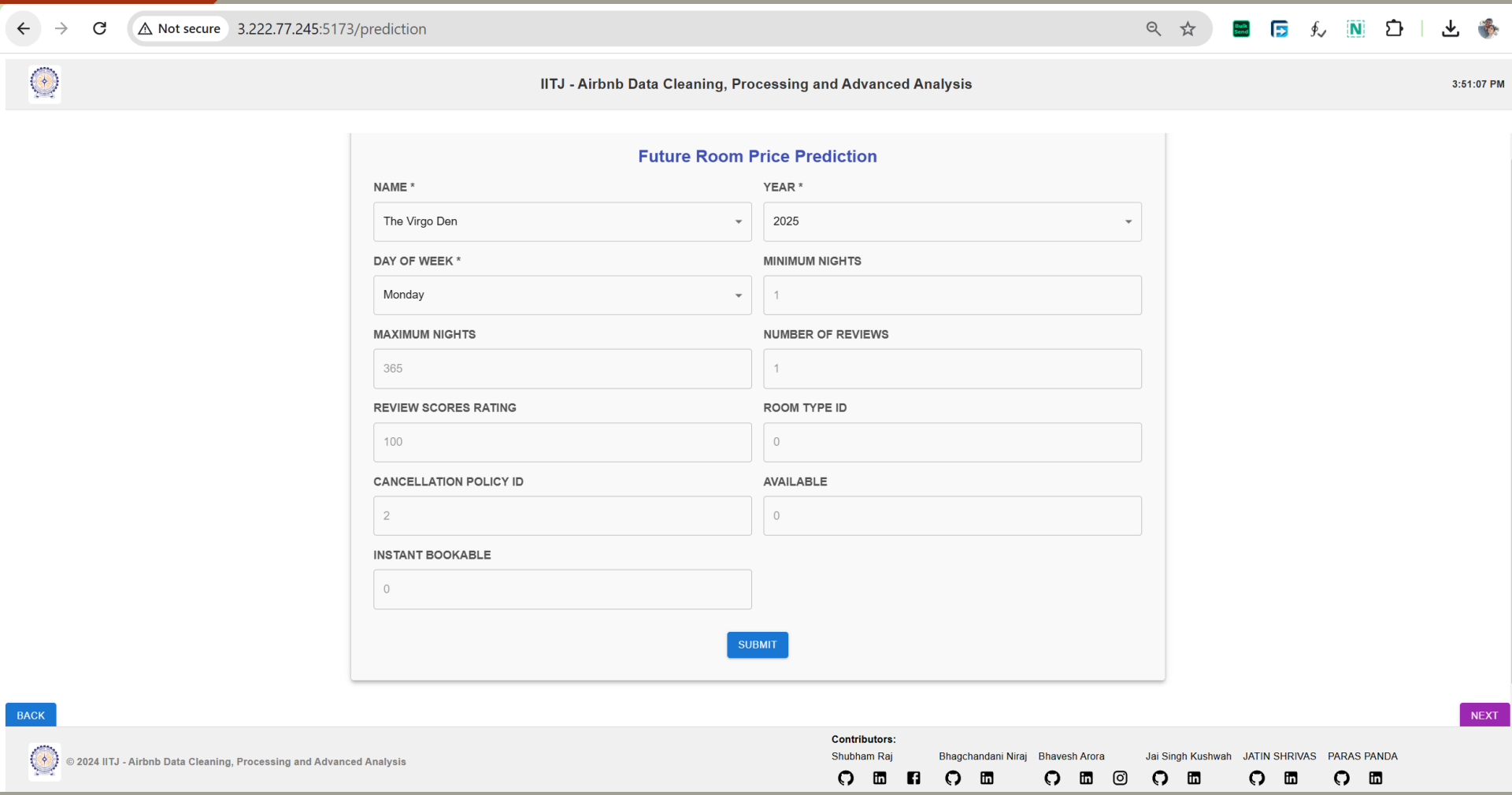
BACK

NEXT



**Contributors:**

Shubham Raj Bhagchandani Niraj Bhavesh Arora Jai Singh Kushwah JATIN SHRIVAS PARAS PANDA


Not secure
3.222.77.245:5173/prediction
IITJ - Airbnb Data Cleaning, Processing and Advanced Analysis
3:51:07 PM
Future Room Price Prediction
NAME *
The Virgo Den
YEAR *
2025
DAY OF WEEK *
Monday
MINIMUM NIGHTS
1
MAXIMUM NIGHTS
365
NUMBER OF REVIEWS
1
REVIEW SCORES RATING
100
ROOM TYPE ID
0
CANCELLATION POLICY ID
2
AVAILABLE
0
INSTANT BOOKABLE
0
SUBMIT
BACK
NEXT
© 2024 IITJ - Airbnb Data Cleaning, Processing and Advanced Analysis
Contributors:
Shubham Raj
Bhagchandani Niraj
Bhavesh Arora
Jai Singh Kushwah
JATIN SHRIVAS
PARAS PANDA

Not secure 3.222.77.245:5173/final-result

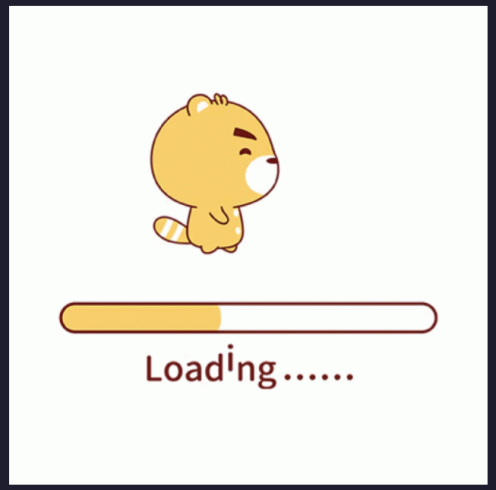
# IITJ - Airbnb Data Cleaning, Processing and Advanced Analysis

3:51:48 PM

**Filling in the data...**
Preparing input for the model

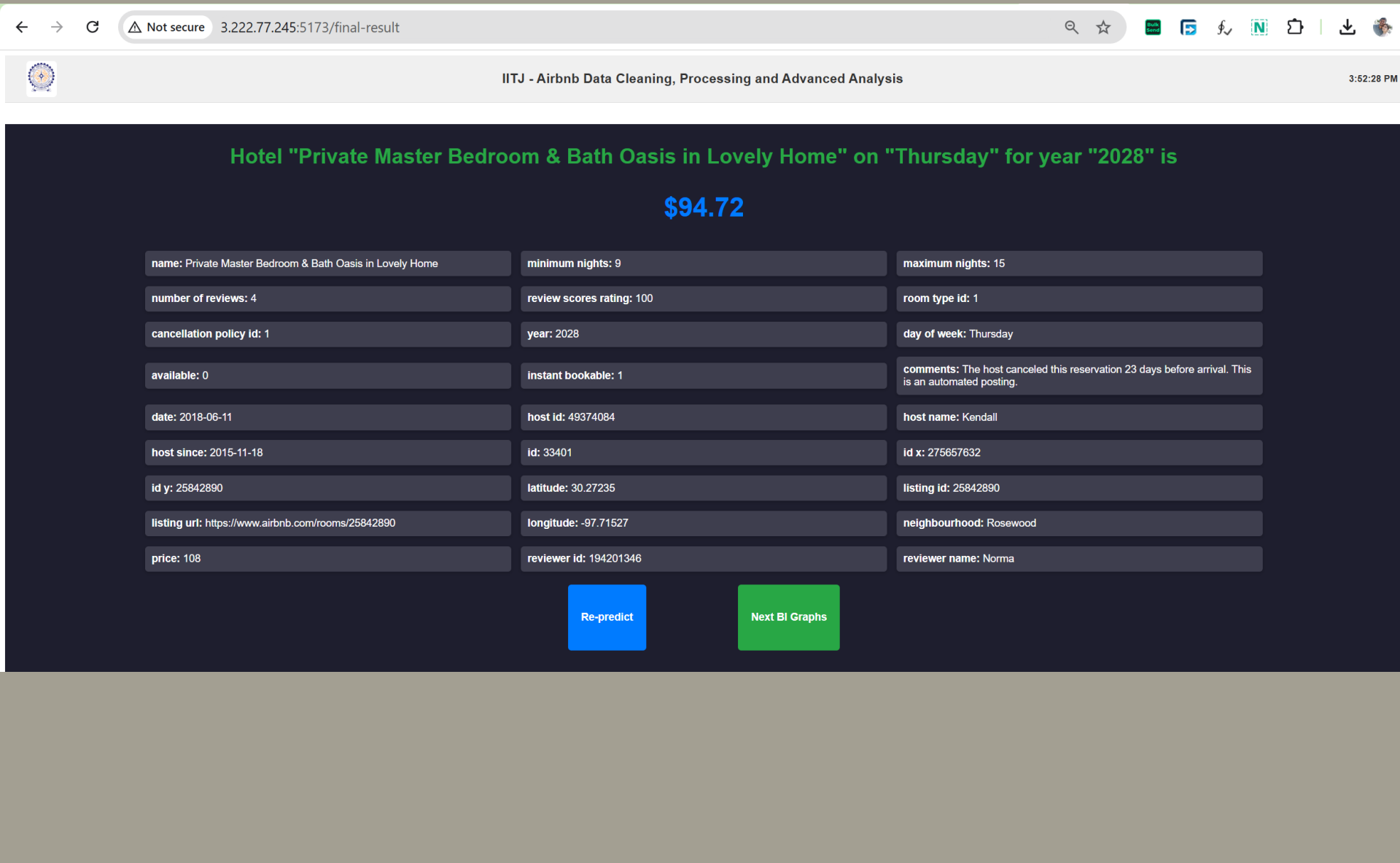
Training the data...
Building and refining the model

Predicting the data...
Generating the final output





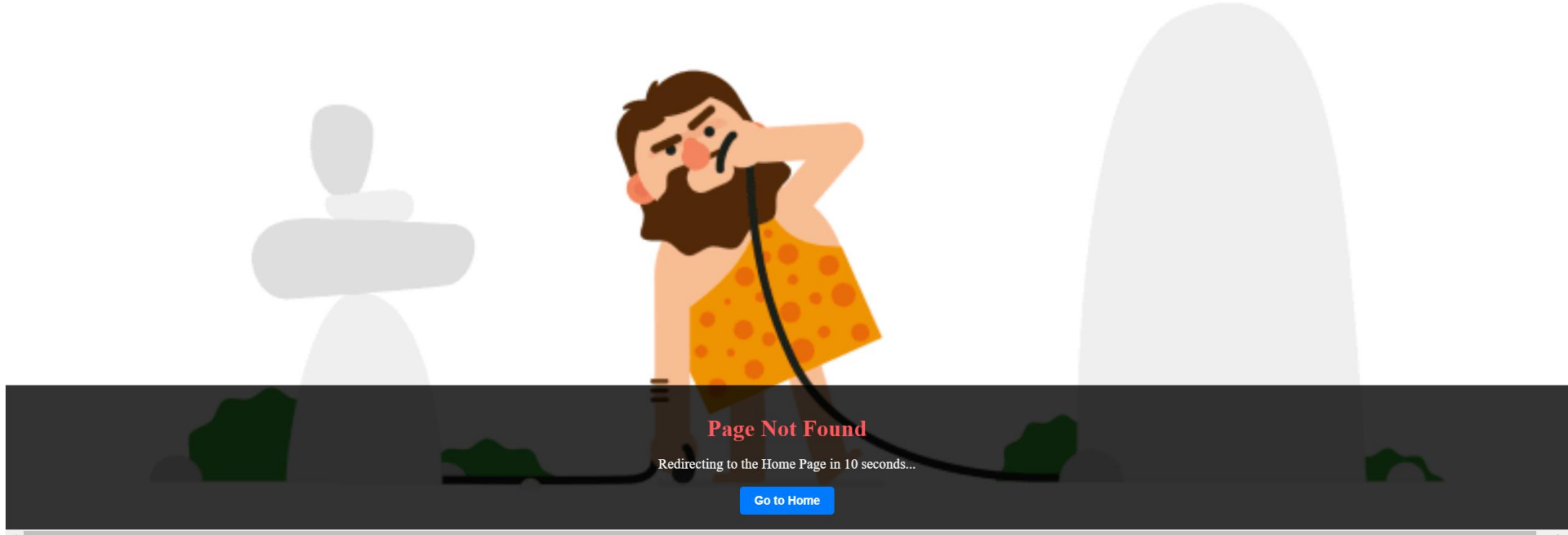

**Contributors:**
Shubham Raj Bhagchandani Niraj Bhavesh Arora Jai Singh Kushwah JATIN SHRIVAS PARAS PANDA


Not secure
3.222.77.245:5173/final-result
IITJ - Airbnb Data Cleaning, Processing and Advanced Analysis
3:52:28 PM
Hotel "Private Master Bedroom & Bath Oasis in Lovely Home" on "Thursday" for year "2028" is
$94.72
name: Private Master Bedroom & Bath Oasis in Lovely Home
minimum nights: 9
maximum nights: 15
number of reviews: 4
review scores rating: 100
room type id: 1
cancellation policy id: 1
year: 2028
day of week: Thursday
available: 0
instant bookable: 1
comments: The host canceled this reservation 23 days before arrival. This is an automated posting.
date: 2018-06-11
host id: 49374084
host name: Kendall
host since: 2015-11-18
id: 33401
id x: 275657632
id y: 25842890
latitude: 30.27235
listing id: 25842890
listing url: https://www.airbnb.com/rooms/25842890
longitude: -97.71527
neighbourhood: Rosewood
price: 108
reviewer id: 194201346
reviewer name: Norma
Re-predict
Next BI Graphs

Not secure 3.222.77.245:5173/abc

IITJ - Airbnb Data Cleaning, Processing and Advanced Analysis

3:53:48 PM



# Page Not Found

Redirecting to the Home Page in 10 seconds...

Go to Home



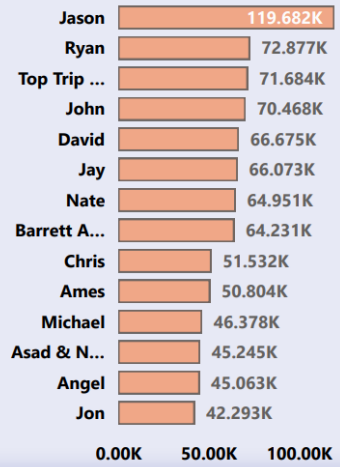
**Contributors:**

Shubham Raj Bhagchandani Niraj Bhavesh Arora Jai Singh Kushwah JATIN SHRIVAS PARAS PANDA

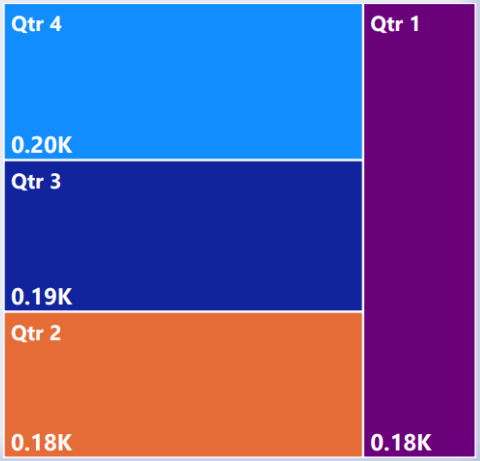
## City

- [ ] Select all
- [ ] Austin
- [ ] Austin Tx
- [ ] Austin, TX - Lake Travis
- [ ] Bee Cave
- [ ] Dripping Springs
- [ ] East Austin
- [ ] Lakeway
- [ ] Manchaca
- [ ] Pflugerville
- [ ] Round Rock

## Top Host by total reviews

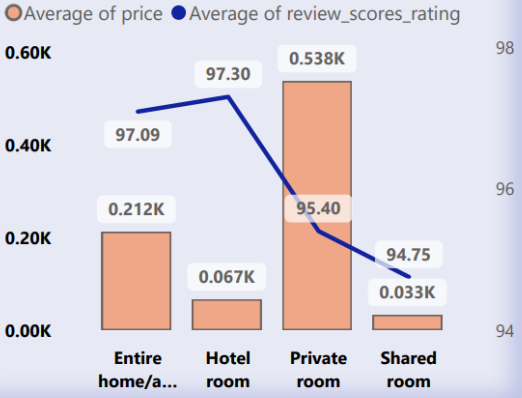



## Summarized Reviews ( Average ratings)

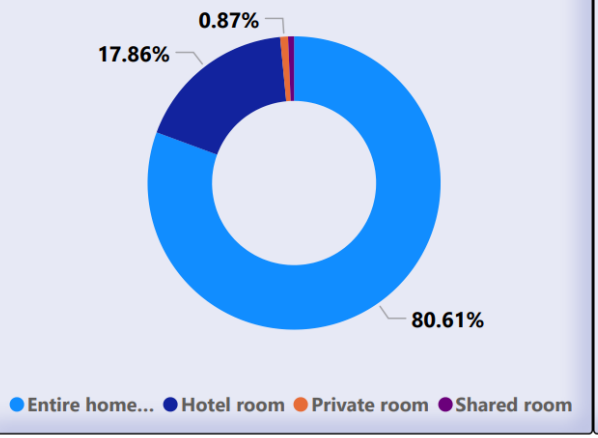



| Host | Count of Property |
|---|---|
| 1540 | 4320 |
| **Value for money** | **Check In** |
| 9.76 | 9.96 |
| **Communincation** | **Location** |
| 9.95 | 9.88 |
| **Cleanliness** | **Overall Rating** |
| 9.77 | 97.09 |

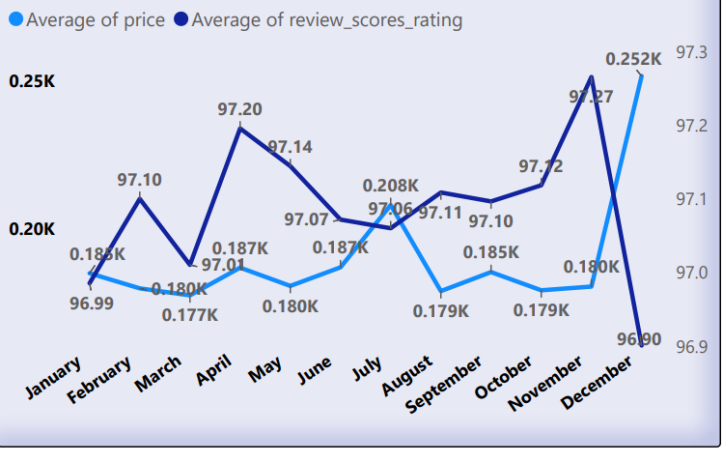
## Average of price and Average of review_scores_rating by room_type




## Room Type Distribution




## Pricing Trends




airbnb
THANK YOU !!!