TOURISM IN SAUDI ARABIA: DATA-DRIVEN INSIGHTS

Course: IT 362: Principles of Data Science
Supervised by: Dr. Reem Alqifari

PROJECT OVERVIEW

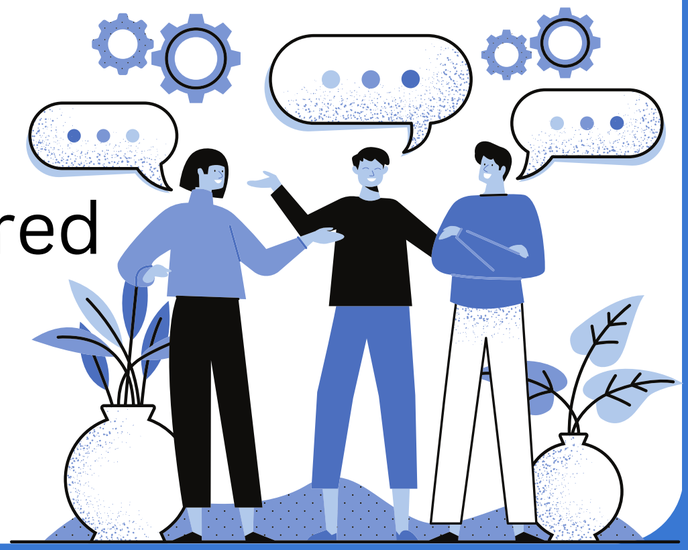
- Saudi Arabia is rapidly emerging as a top global tourist destination under Vision 2030.
- This project investigates key factors that drive tourism using a data-driven approach.
- Research Question: What are the key factors driving tourism in Saudi Arabia, and how can this data be used to enhance the country’s tourism strategy?

(A) STEPS FROM DATA COLLECTION TO MODEL EVALUATION

DATA COLLECTION

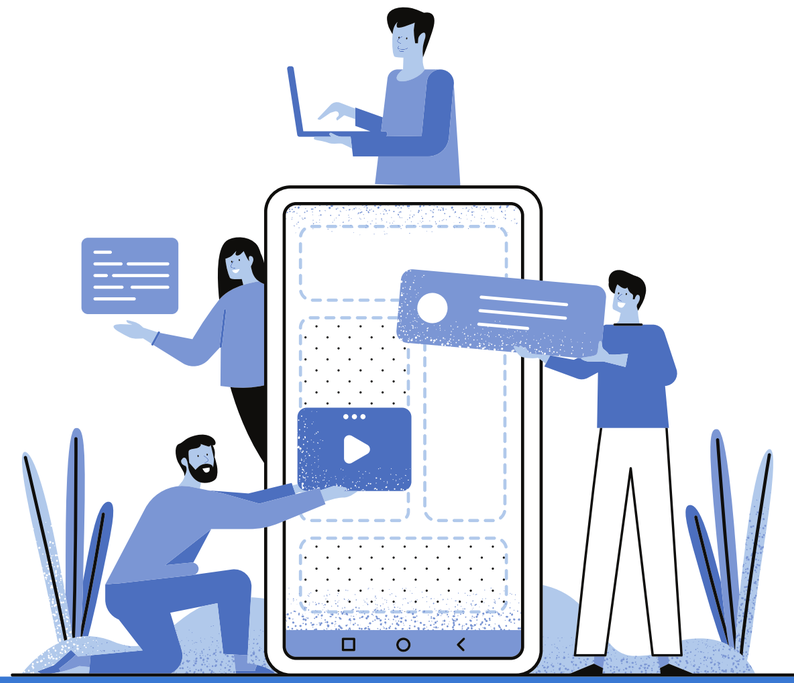
- Data was collected from multiple sources:
  - TripAdvisor & Booking.com: Web scraping using Apify
  - Survey: Google Forms (Feb 4–6, 2025)
  - Twitter/X: API via Bearer Token
  - Ministry of Tourism: Public datasets

- The combination of structured and unstructured data helped us gain a broad perspective.



DATA PREPROCESSING – GENERAL STEPS

- Removed duplicates
- Handled missing values:
  - Categorical → “Unknown”
  - Numerical → Median
- Standardized column names
- Removed irrelevant fields



DATASET-SPECIFIC PREPROCESSING

Source	Cleaning Actions
TripAdvisor	Removed reviews with nulls, dropped high-missing cols
Booking.com	Dropped metadata (photos, contact info)
Twitter	Removed emojis, URLs, duplicates; formatted datetime
Survey	Removed empty/duplicate rows, standardized columns
Ministry Data	Removed headers, dropped unnecessary cols, added type

EXPLORATORY DATA ANALYSIS (EDA)

- Statistical summaries (mean, median, std)
- Visualization tools: Seaborn, Matplotlib
- Examples:
  - Most visited destinations
  - Spending trends
  - Ratings distribution
  - Event influence score

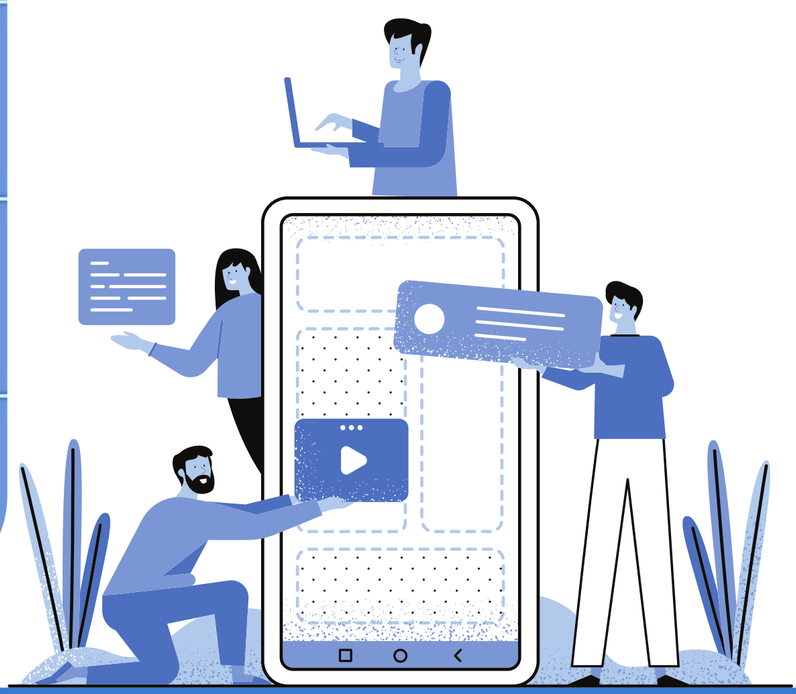
(B) ALGORITHMS AND RATIONALE

OVERVIEW OF ALGORITHMS USED

- **Clustering:** KMeans, Agglomerative
- **Regression:** Linear, Logistic
- **Classification:** SVM, Random Forest

ALGORITHM RATIONALE

Algorithm	Reason for Selection
KMeans	Segment cities by popularity & satisfaction
Agglomerative	Hierarchical grouping of tourist types
Logistic Regression	For binary outcomes (e.g., high vs. low spenders)
SVM	Effective in small datasets with nonlinear separation
Random Forest	For feature importance, although slightly lower performance



(C) EVALUATION METRICS & THEIR MEANING

MODEL EVALUATION METRICS

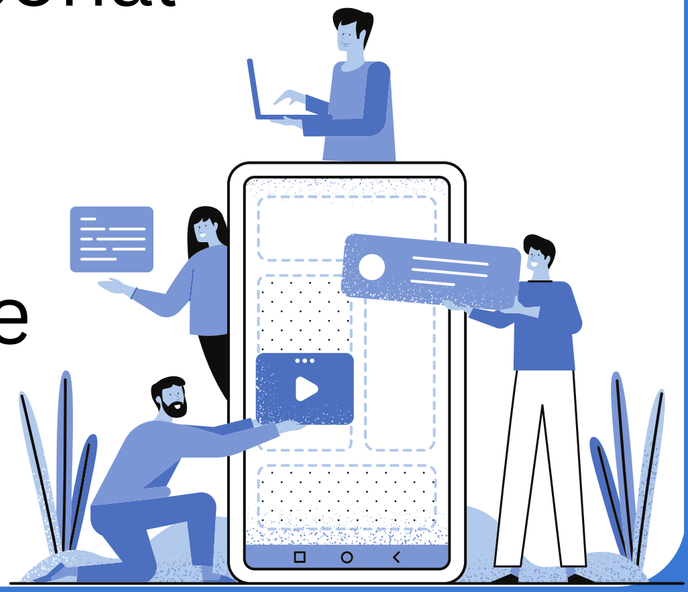
- **Classification Models:**
  - Accuracy, Precision, Recall, F1-score
- **Regression Models:**
  - R^2 Score
- **Clustering:**
  - Silhouette Score

WHY THESE METRICS?

- **Accuracy:** General performance
- **F1-score:** Handles imbalanced data
- **R^2 Score:** Goodness of fit for regression
- **Silhouette Score:** Quality of clustering
- Example: SVM scored ~80% accuracy for satisfaction prediction.

WHAT THE METRICS REVEAL?

- **SVM Accuracy (~80%)**
  - → The model makes reliable predictions for tourist satisfaction.
- **R^2 Score (Regression)**
  - → Shows that the model explains seasonal trends in tourist inflow well.
- **Silhouette Score (Clustering)**
  - → Confirms that the tourist groups are well-defined and meaningful.



(D) TECHNICAL HURDLES & SOLUTIONS

KEY TECHNICAL CHALLENGES

Challenge	Solution
Missing data	Imputation (median/"Unknown")
Unstructured text (Twitter)	Manual preprocessing (cleaning, sentiment)
API limitations (Twitter: 100 max)	Combined multiple queries & sessions
Dataset inconsistency	Standardized names & column formatting
Survey privacy compliance	Removed identifying data, filled NA safely

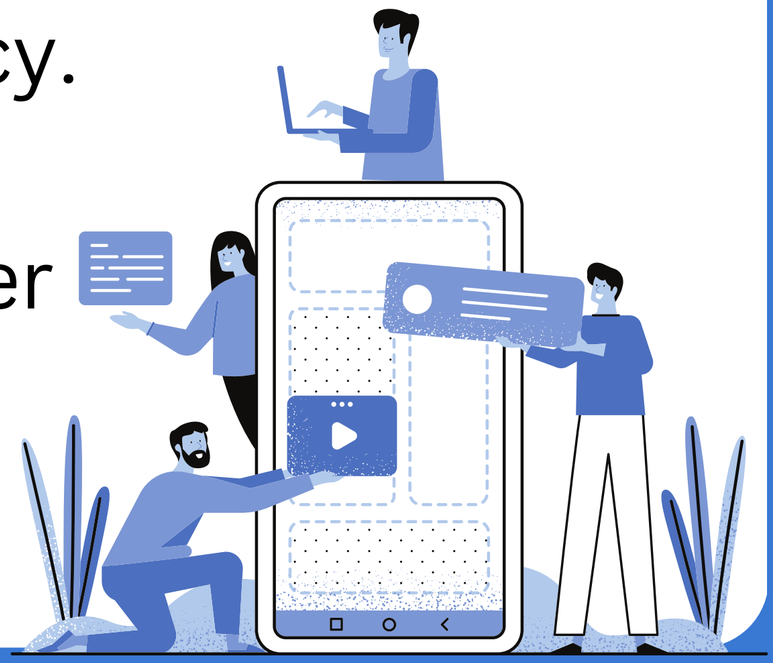
(E) IMPROVEMENTS & FUTURE WORK

FUTURE WORK

- Analyze demographics and sentiment in more depth
  - → Helps identify how age, nationality, and trip purpose affect satisfaction.
- Compare ratings across different review platforms
  - → Reveals possible biases and differences in user behavior between sources.

FUTURE WORK

- Use additional clustering algorithms like DBSCAN
  - →Helps discover hidden tourist groups and improves segmentation accuracy.
- Include data from different years
  - → Allows tracking tourism trends over time and measuring event impacts.



(F) CLOSING FOR TECHNICAL AUDIENCE

TECHNICAL TAKEAWAYS

- Multiple ML techniques tailored for tourism domain
- Challenges tackled with practical preprocessing strategies
- Insights drawn from integrated, diverse data sources

THANK YOU FOR YOUR ATTENTION
WE WELCOME YOUR QUESTIONS

Najla Almazyad-444200948
Jood Alkhrashi-444203007
Ghala Musallam-444200807
Reuof Alanazi-444200528