

# DATA ANALYSIS PROJECTS

## Objective

Each team will design, analyze, and interpret a real-world dataset from a chosen domain. Projects will demonstrate mastery of the **entire data analysis workflow**, including **data wrangling, EDA (Exploratory Data Analysis), feature engineering, feature selection, probability analysis, hypothesis testing, and dimensionality reduction (PCA)**.

Each team's final product should include:

- A comprehensive **Jupyter Notebook** documenting all steps.
  - **Visualizations** interpreting key findings.
  - **Statistical validation** of insights through hypothesis tests.
  - A concise **presentation/report** highlighting actionable insights.
- 

## Team Structure

Each team will consist of **5–6 members** with distinct, well-defined roles:

- Data Acquisition & Cleaning
  - Exploratory Data Analysis & Visualization
  - Feature Engineering & Selection
  - Statistical Modeling & Hypothesis Testing
  - PCA & Insight Visualization
  - Documentation & Presentation
- 

## Learning Outcomes

By the end of this project, students will:

- Collect, clean, and preprocess real-world data using Python Data Analysis libraries (Pandas, NumPy).
- Conduct univariate and multivariate analyses using visualization tools (Matplotlib, Seaborn).
- Engineer meaningful features, apply selection techniques, and reduce dimensionality.
- Apply probability distributions and statistical hypothesis tests for decision making.
- Communicate findings through visual storytelling and recommendations.
- Collaborate effectively within teams using versioned notebooks and structured documentation.

## General Project Requirements

Component	Description	Points
<b>Data Wrangling</b>	Clean dataset: handle missing values, duplicates, outliers, inconsistent formats.	3 pts
<b>Exploratory Data Analysis (EDA)</b>	Perform univariate & multivariate analyses. Use appropriate visualizations with insights.	4 pts
<b>Feature Engineering &amp; Selection</b>	Create new features, encode variables, apply selection methods (filter, Lasso, RFE).	3 pts
<b>Probability &amp; Hypothesis Testing</b>	Fit distributions, calculate probabilities, and conduct hypothesis tests relevant to dataset.	3 pts
<b>Dimensionality Reduction (PCA)</b>	Apply PCA, interpret components, visualize clusters/patterns.	2 pts
<b>Documentation &amp; Presentation</b>	Clear report (Jupyter Notebook + slides); concise interpretation of results.	3 pts
<b>Total</b>		<b>18 pts (Technical) + 2 pts (Team Participation) = 20 pts Total</b>

---

## Common Deliverables

All teams must submit:

1. **A Jupyter Notebook (.ipynb)** documenting the full analysis pipeline.
2. **A cleaned dataset** and code used for loading/cleaning.
3. **Visualizations** demonstrating insights (EDA, correlations, PCA).
4. **Statistical tests and outcomes** clearly explained.

5. **Final report slides or PDF summary** connecting insights to business or research context.
- 

## ◆ **Detailed Project Ideas**

### 1. **Customer Churn Analysis**

- Goal: Identify churn factors and recommend retention strategies.
- Tasks: Wrangling (clean customer data), EDA (usage patterns), Feature Engineering (tenure groups), Feature Selection (Lasso), Probability ( $P(\text{churn} | \text{contract})$ ), Hypothesis Testing (effects of payment type), PCA (cluster customers).

### 2. **Movie Ratings Prediction & Insights**

- Goal: Analyze variables influencing movie ratings and revenue.
- Tasks: Clean budgets and release dates, extract years, encode genres, correlation analysis, t-tests between genres, PCA to visualize genre similarity.

### 3. **E-commerce Sales Optimization**

- Goal: Analyze sales trends and customer behavior to maximize revenue.
- Tasks: Handle missing orders, analyze product/customer-level metrics, create new temporal and segment features, Poisson model of sales pattern, t-test for promotions.

### 4. **Health Risk Prediction**

- Goal: Analyze medical and lifestyle factors influencing disease risk.
- Tasks: Clean health data, correlation between metrics, create BMI bins, calculate  $P(\text{disease} | \text{smoker})$ , run t-tests, visualize healthy vs. at-risk clusters.

### 5. **Bank Marketing Campaign Response Analysis**

- Goal: Predict customer response likelihood and optimize targeting.
- Tasks: Wrangling demographic & campaign data, EDA response rates, encode categorical jobs, calculate  $P(\text{response} | \text{job})$ , Chi-square test, PCA for segment visualization.

### 6. **Taxi Fare & Trip Duration Analytics**

- Goal: Explore trip time, fare structure, and patterns by time or route.
- Tasks: Remove outliers, calculate trip distance, analyze fare patterns, fit trip duration distributions, compare peak vs off-peak fares (t-test).

## **7. Retail Store Inventory & Demand Forecasting**

- Goal: Predict product demand and improve inventory management.
- Tasks: Handle missing sales data, extract seasonal features, Poisson modeling for daily demand, PCA to visualize product clusters, create forecast dashboards.

## **8. Hotel Booking Cancellation Analysis**

- Goal: Understand cancellation drivers and seasonal patterns.
- Tasks: Handle missing stays, analyze cancellation distribution, calculate  $P(\text{cancel} \mid \text{room type})$ , Chi-square test, PCA for booking behavior clusters.

## **9. Environmental Air Quality Analytics**

- Goal: Model pollution behavior and correlation with weather.
- Tasks: Fill missing sensor data, analyze pollutant trends, fit distributions for PM2.5, test differences between workdays/weekends, PCA for city pattern clusters.

## **10. Student Performance Analysis**

- Goal: Discover factors affecting academic success.
- Tasks: Handle missing scores, analyze grade trends, extract attendance features, compute  $P(\text{pass} \mid \text{study time})$ , run t-tests, PCA for learning styles.

## **11. Credit Card Fraud Detection Analysis**

- Goal: Explore transaction data to identify suspicious patterns.
- Tasks: Clean transactions, EDA by amount/time, feature engineering (time difference), Probability  $P(\text{fraud} \mid \text{amount range})$ , Hypothesis test (mean comparisons), PCA for transaction grouping.

## **12. Social Media Sentiment & Engagement Analysis**

- Goal: Identify factors affecting engagement in posts/tweets.
- Tasks: Clean posts (remove emojis/links), EDA (likes, shares, sentiment), feature extraction (word length, hashtags), correlation analysis, hypothesis testing between post types.

### **13. COVID-19 Data Analysis**

- Goal: Analyze case trends, recovery rates, and vaccination impact.
- Tasks: Clean global case data, aggregate by region, create new features (case fatality rate), fit distributions (case daily counts), compare vaccination vs infection rates.

### **14. Real Estate Price Determinants**

- Goal: Analyze variables influencing housing prices.
- Tasks: Clean property listings, EDA on area, price, location; feature engineering (price per sq. ft.), regression feature selection, t-test between urban and rural prices, PCA visualization.

### **15. Energy Consumption Analytics**

- Goal: Understand how energy usage varies with time and conditions.
- Tasks: Clean energy meter data, analyze consumption patterns, create weekday/weekend flags, calculate  $P(\text{usage} > \text{threshold} | \text{hour})$ , test seasonal differences.

### **16. Employee Turnover Analysis**

- Goal: Identify HR patterns leading to employee attrition.
- Tasks: Clean records, explore turnover by department/experience, engineer satisfaction metrics, test difference in salary means, visualize clusters via PCA.

### **17. Traffic Accident Severity Analysis**

- Goal: Analyze the impact of weather/time on accident severity.
- Tasks: Clean accident data, calculate correlations between speed and severity, create categorical bins (weather type), perform Chi-square test, visualize via PCA.

### **18. Financial Market Volatility Insights**

- Goal: Study stock return distributions and volatility relationships.
- Tasks: Clean daily stock data, plot return distributions, fit normal/lnormal, test hypothesis (mean returns), PCA of sector correlations.

### **19. Wildlife Population & Conservation Data Analysis**

- Goal: Examine patterns in animal populations across regions.
- Tasks: Clean survey data, analyze population trends, compute  $P(\text{decline}|\text{region})$ , run t-tests across habitats, use PCA to visualize region-species similarity.

## 20. Online Learning Engagement Analytics

- Goal: Explore student engagement in e-learning environments.
  - Tasks: Clean LMS logs, create features (logins, time spent, quiz attempts), identify patterns, apply hypothesis testing for engagement by age/group, perform PCA for activity segmentation.
- 

### Final Deliverables

Each team must submit:

1. Complete Jupyter Notebook analysis (cleaning → visualization → PCA).
2. Final presentation/report summarizing insights.
3. Dataset (raw + cleaned version).
4. Clearly defined roles & contributions.
5. Visualizations, charts, and key takeaways supporting findings.