INTERNSHIP PROJECT PHASE (2 WEEKS)

PROJECT GUIDELINES

- **Choose any 1 or 2 projects from the given list.**
- **You are free to improvise — take the given project as a base and modify it as you like.**
- **You can use any tools, technologies, or steps you're comfortable with — there are no restrictions.**
- **Focus and work sincerely so that you have complete clarity and can explain the project confidently in interviews.**
- **Go through the Top 50 Interview Questions for your domain (attached at the end).**
- **Update your project status regularly when the Google Form is shared in group.**
- **while working on the project YOU CAN CHOOSE ANY DATASET RELAVENT TO THE PROJECT.**

.

After project completion, prepare a 1–2 page report in PDF format, containing:

- Introduction
- Abstract
- Tools Used
- Steps Involved in Building the Project
- Conclusion

◆ Note: Report must not exceed 2 pages.

DEAR INTERNS,

YOU HAVE TO UPDATE STATUS OF YOUR PROJECT EVERY 3 OR 4 DAYS ONCE WHEN THE UPDATION LINK IS SHARED IN THE GROUP.

Final submission links will be shared later.

READ ALL THE GUIDELINES CAREFULLY

LIST OF PROJECTS

1. Retail Business Performance & Profitability Analysis

Objective: Analyze transactional retail data to uncover profit-draining categories, optimize inventory turnover, and identify seasonal product behavior.

Tools: SQL, Python (Pandas, Seaborn), Tableau

Mini Guide:

Import data into SQL and clean missing/null records
Use SQL to calculate profit margins by category and sub-category
Use Python (Pandas) to run correlation between inventory days and profitability
Build Tableau dashboard with filters for region, product type, and season
Derive strategic suggestions for slow-moving and overstocked items

Deliverables:

Tableau Dashboard
SQL queries (.sql file)
PDF Report with key insights

2. Customer Lifetime Value Prediction Model

Objective: Predict the lifetime value (LTV) of customers based on their purchase behavior to aid in targeted marketing.

Tools: Python (Sklearn, XGBoost), Excel

Mini Guide:

Preprocess customer purchase history (merge transactions with customer IDs)
Feature engineering: Frequency, Recency, AOV (Avg Order Value)
Train regression model (XGBoost or Random Forest)
Validate using MAE, RMSE
Segment customers based on predicted LTV

Deliverables:

Python notebook
Trained model + visualizations
Final LTV prediction CSV

3. HR Analytics - Predict Employee Attrition

Objective: Use analytics to understand the main causes of employee resignation and predict future attrition.

Tools: Python (Pandas, Seaborn), Power BI, Sklearn

Mini Guide:

Perform EDA on HR data (department-wise attrition, salary bands, promotions)
Build a classification model (Logistic Regression or Decision Tree)
Visualize attrition factors using Power BI
Perform SHAP value analysis to explain model predictions

Deliverables:

Power BI dashboard
Model accuracy report + confusion matrix
PDF of attrition prevention suggestions

4. YouTube Trending Video Analytics

Objective: Uncover patterns in trending videos by analyzing YouTube datasets across regions.

Tools: Python (Matplotlib, Seaborn), SQL, Tableau

Mini Guide:

Clean and standardize YouTube trending datasets from different countries

Perform sentiment analysis on titles and tags

Use SQL to rank categories by avg views

Create time-series visualizations for trending duration

Deliverables:

Dashboard: Most popular genres, sentiments

Region-wise comparison visuals

Final report with data storytelling

5. E-commerce Return Rate Reduction Analysis

Objective: Identify why customers return products and how return rates vary by category, geography, and marketing channel.

Tools: Python, Power BI, SQL

Mini Guide:

Clean return and order dataset

Analyze return % per category and supplier

Use logistic regression to predict probability of return

Use Power BI to create return risk score dashboard

Deliverables:

Interactive dashboard with drill-through filters

Python codebase for prediction

CSV of high-risk products

6. Customer Churn Analysis for Telecom Industry

Objective: Predict churn and derive actionable strategies to retain users in a highly competitive telecom environment.

Tools: Python (Scikit-learn, ELI5), SQL

Mini Guide:

Use SQL for data aggregation (call duration, complaints, recharge frequency)

Build binary classification model for churn

Use ELI5 or SHAP for model explainability

Create customer segments: At Risk, Loyal, Dormant

Deliverables:

Python ML notebook

Customer churn report (PowerPoint)

Final recommendations

7. Financial KPI Analysis for a Startup

Objective: Analyze monthly revenue, burn rate, CAC, LTV, and run rate for an early-stage startup.

Tools: Excel, Tableau, Python (Pandas)

Mini Guide:

Collect financials: expenses, revenue, customer base

Compute LTV:CAC ratio

Build dashboard with trend indicators

Perform cohort analysis (monthly customer groups)

Deliverables:

Tableau dashboard

LTV:CAC report in PDF

Excel model template

8. Movie Success Prediction and Sentiment Study

Objective: Predict movie success using IMDB/Kaggle data, and analyze sentiment of viewer reviews.

Tools: Python (NLTK, VADER, Sklearn), Excel

Mini Guide:

Scrape or import IMDB movie + rating data

Use VADER for sentiment on user reviews

Create regression model to predict box office success

Analyze genre-wise sentiment trends

Deliverables:

Python notebooks

Sentiment visuals

Predictive model summary

9. Airbnb Dynamic Pricing Recommendation Engine

Objective: Analyze historical Airbnb data to suggest optimal pricing based on location, season, and listing quality.

Tools: Python, Tableau, Excel

Mini Guide:

Analyze pricing by city, property type, reviews

Run regression model to find pricing predictors

Create dashboard with price suggestion slider

Deliverables:

Tableau dashboard with filters

Python pricing engine script

Final PDF with suggestions

10. Real-Time Public Sentiment Dashboard (Twitter/X)
Objective: Track public opinion about a brand or product in real-time using Twitter data.
Tools: Python (Tweepy, NLTK), Tableau
Mini Guide:
Stream tweets using Tweepy
Apply NLTK to clean and score sentiment
Batch-update Tableau dashboard with summary charts
Deliverables:
Live sentiment dashboard
Python streaming + NLP script
Daily sentiment logs

11. Healthcare Appointment No-Show Prediction
Objective: Predict whether patients will miss their appointments and optimize scheduling.
Tools: Python (Sklearn, Pandas), Power BI
Mini Guide:
Import and clean appointment data
Train decision tree model to predict no-shows
Analyze trends like SMS reminders, age, weekday
Deliverables:
Prediction model
Power BI insight dashboard
Optimization recommendations

12. Electric Vehicle Charging Demand Forecasting
Objective: Forecast the demand at EV charging stations based on weather, time, and traffic.
Tools: Python, Excel, Tableau
Mini Guide: Merge EV usage + weather datasets
Create time-series models (ARIMA/Prophet)
Build Tableau dashboard to visualize demand curves
Deliverables: Forecasting model Tableau heatmaps Charging optimization strategy

13. Global CO2 Emissions Tracker by Sector

Objective: Build a dashboard to track carbon emissions from energy, transport, and industry sectors across countries.

Tools: Tableau, Excel, Python (for data prep)

Mini Guide:

Import multi-year emissions dataset

Prepare per capita and per GDP metrics

Use Tableau maps and bar graphs by sector

Deliverables:

Global emissions dashboard

PDF policy brief on top polluters

14. LinkedIn Job Trend Analysis (Web Scraping)

Objective: Scrape LinkedIn job postings to analyze skill demand trends across cities and roles.

Tools: Python (BeautifulSoup, Pandas), Excel

Mini Guide:

Scrape job titles, skills, locations using BeautifulSoup

Clean and parse skill tags

Generate heatmaps of top 10 skills by city

Deliverables:

Trend analysis visuals

Skill vs Role matrix

Job demand recommendation

15. Startup Investment Analysis (Shark Tank Data)

Objective: Analyze startup investment trends using Shark Tank India/US datasets.

Tools: Excel, Python, Tableau

Mini Guide:

Clean and organize data by domain, funding amount

Analyze founder profiles, funding stage success

Create Tableau visuals for industry trends

Deliverables:

Visual dashboard

PDF with industry-wise investor trends

Founder success pattern summary

! TOP 50 INTERVIEW QUESTIONS FOR DATA ANALYST !

1. What are the key differences between inner join and outer join in SQL?
2. How do you handle missing data in a dataset?
3. What is the difference between variance and standard deviation?
4. Explain the concept of normalization in databases.
5. What is the role of a primary key in a relational database?
6. How would you detect outliers in a dataset?
7. What is data wrangling and why is it important?
8. Describe a situation where you used data to solve a business problem.
9. What is the difference between a clustered and non-clustered index?
10. Explain the difference between supervised and unsupervised learning.
11. What is the purpose of the GROUP BY clause in SQL?
12. How do you handle duplicate data entries in a dataset?
13. What is a pivot table and how have you used it?
14. Explain the differences between a bar chart and a histogram.
15. How do you optimize a slow SQL query?
16. What are the common KPIs used in business analysis?
17. What is A/B testing and how is it used in data analysis?
18. How do you ensure data accuracy and integrity in a project?
19. What is a correlation matrix and how do you interpret it?
20. What is the difference between correlation and causation?
21. Describe a data project where you used Python.
22. What libraries do you use for data analysis in Python?
23. Explain the use of Pandas groupby() function.
24. How do you deal with imbalanced datasets?
25. What are the steps of a typical data analysis pipeline?
26. What is the purpose of data visualization?
27. Explain the difference between ETL and ELT.
28. What is the difference between OLAP and OLTP systems?
29. How do you decide which chart to use for a dataset?
30. What is time series analysis and where have you used it?
31. Describe your experience with Tableau or Power BI.
32. What are dimensions and measures in Tableau?
33. How do you track data quality over time?
34. What is multicollinearity and why is it a problem?
35. How would you analyze user behavior on a website?
36. What are your favorite Python functions for data analysis?
37. What is data cleaning and how do you perform it?
38. What does the term 'data storytelling' mean to you?
39. How do you handle large datasets efficiently?
40. What are lag and lead functions in SQL?
41. What is a hypothesis test and when would you use it?
42. How do you explain complex data insights to non-technical stakeholders?
43. What is the difference between a heatmap and a scatter plot?
44. How do you validate a machine learning model?
45. Describe a challenging dataset you worked on.
46. What is the role of feature engineering in data analysis?
47. What is the difference between a data analyst and a data scientist?
48. How do you prioritize tasks when working on multiple data projects?
49. What steps do you take before starting a data analysis project?
50. Describe a situation where your analysis had a measurable business impact.