internStudio

DATA SCIENCE INTERNSHIP

TASK GUIDE

Ask sharper questions of data. Test hypotheses, quantify uncertainty, find what's real.

SIX ADVANCED PROJECTS — CHOOSE ANY TWO

01 **A/B Testing & Experiment Analysis Platform**

02 **Customer Lifetime Value & Cohort Modeling**

03 **Survey Data Analysis & Statistical Inference**

04 **Statistical Health Risk Scoring**

05 **Market Basket & Association Rule Mining**

06 **Time Series Decomposition & Anomaly Detection**

INTERNSHIP
Data Science

REQUIRED TASKS
Complete any 2 projects

FOCUS
Statistics · Hypothesis Testing · Modeling · Experimentation

INTERNSHIP INSTRUCTIONS

Your Internship Tasks

Select and complete **any 2 projects** from the task list. The goal is rigorous statistical reasoning, not just a chart — quantify your confidence, not just your conclusion.

Collect	Clean	Explore	Model	Validate	Interpret	Report

EACH PROJECT SHOULD DEMONSTRATE

- [ ] Rigorous data cleaning and preparation
- [ ] Exploratory analysis with real statistical grounding
- [ ] An explicit hypothesis or research question
- [ ] Appropriate statistical tests or models
- [ ] Quantified uncertainty (confidence intervals, p-values, error bars)
- [ ] Validation against a held-out set where applicable
- [ ] Honest discussion of limitations
- [ ] Clear, defensible conclusions

EVERY SUBMISSION INCLUDES

- [ ] GitHub repository
- [ ] Jupyter Notebook or scripts
- [ ] Dataset / source
- [ ] README explaining the methodology
- [ ] Key findings and their statistical basis
- [ ] Visualizations that support the conclusions
- [ ] Requirements / dependency file

> **Suggested technologies:** Python · Pandas · NumPy · SciPy / statsmodels · Matplotlib / Seaborn · scikit-learn. Optional: Jupyter, PyMC for Bayesian work.

Submit your completed work here

Once both projects are analyzed, documented and pushed to GitHub, submit your project details through the official submission form.

https://forms.gle/vxL5JE8cvRXXfDBv8

> **A note on quality.** A p-value or a model score alone proves nothing. You should be able to explain what your method assumes, whether those assumptions hold, and how confident your conclusion really is.

01 A/B Testing & Experiment Analysis Platform

Design and analyze an A/B test end to end, from experiment design through statistical decision.

REQUIREMENTS

- Define a clear hypothesis and success metric
- Clean and validate the experiment data
- Run the appropriate significance test
- Discuss statistical vs. practical significance
- Determine required sample size / test duration
- Check for sample ratio mismatch and other validity issues
- Report effect size with confidence intervals
- Provide a clear ship/no-ship recommendation

Focus: Experimental design, hypothesis testing, and translating statistics into a business decision.

02 Customer Lifetime Value & Cohort Modeling

Build a statistical model estimating customer lifetime value and analyze how it varies across acquisition cohorts.

REQUIREMENTS

- Clean and structure transactional / subscription data
- Model retention curves per cohort
- Compare CLV across cohorts and segments
- Visualize cohort retention and CLV trends
- Build cohorts by acquisition period
- Estimate CLV with a defensible statistical approach
- Quantify uncertainty in the CLV estimates
- Recommend where to focus acquisition spend

Focus: Cohort analysis, retention modeling and translating CLV into acquisition strategy.

03 Survey Data Analysis & Statistical Inference

Analyze a real survey dataset to draw statistically defensible conclusions about the surveyed population.

REQUIREMENTS

- Clean and recode survey responses
- Compute confidence intervals for key estimates
- Correct for multiple comparisons where relevant
- Distinguish correlation from causation explicitly
- Check for response bias and non-response patterns
- Run appropriate hypothesis tests across subgroups
- Visualize findings with uncertainty shown
- Summarize findings for a non-technical audience

Focus: Survey methodology, inferential statistics and communicating uncertainty clearly.

04 Statistical Health Risk Scoring

Build a statistically grounded risk-scoring model (not a black-box ML model) for a health or insurance-style dataset.

REQUIREMENTS

- Clean and prepare the dataset
- Build an interpretable model (logistic regression or similar)
- Validate model calibration, not just accuracy
- Segment risk into meaningful bands
- Explore relationships between risk factors and outcomes
- Report odds ratios or coefficients with confidence intervals
- Check for confounding and multicollinearity
- Discuss ethical implications of the scoring

Focus: Interpretable statistical modeling, calibration and responsible scoring design.

05 Market Basket & Association Rule Mining

Analyze transactional data to discover statistically meaningful product associations.

REQUIREMENTS

- Clean and structure transaction-level data
- Compute support, confidence and lift for rules
- Visualize the strongest associations
- Validate findings on a held-out period
- Apply an association-rule mining algorithm (Apriori/FP-Growth)
- Filter to statistically and practically meaningful rules
- Segment findings by customer or store type
- Recommend concrete merchandising actions

Focus: Association-rule mining, statistical filtering and merchandising recommendations.

06 Time Series Decomposition & Anomaly Detection

Decompose a real time series into trend, seasonality and residual, and statistically flag anomalies.

REQUIREMENTS

- Clean and resample the time series appropriately
- Test for stationarity
- Flag anomalies against that band, not eyeballing
- Backtest the detection approach on known events
- Decompose into trend, seasonal and residual components
- Build a statistical control band for expected values
- Distinguish a true anomaly from a regime shift
- Visualize the decomposition and flagged anomalies

Focus: Time series decomposition, statistical anomaly detection and rigorous backtesting.

FINAL STEP

Ready to Submit?

Verify everything below before you send your work in.

SUBMISSION

- [ ] GitHub repository
- [ ] Notebook / source code
- [ ] README
- [ ] Requirements / dependency file
- [ ] Key findings documented
- [ ] Submission form completed
- [ ] LinkedIn post where required

VALIDATION

- [ ] Uncertainty quantified
- [ ] Results validated where applicable
- [ ] Limitations honestly discussed

DATA

- [ ] Dataset / source included
- [ ] Cleaning and preparation documented
- [ ] Missing values and outliers handled

METHOD

- [ ] Hypothesis or question stated explicitly
- [ ] Appropriate statistical method chosen
- [ ] Assumptions checked, not assumed

Important. Complete any 2 projects. The objective is statistical rigor, not a high-looking number — you should be able to defend your method choice and your confidence in the result.

Submission Form

Submit your repository, live link and screenshots here.

https://forms.gle/vxL5JE8cvRXXfDBv8

internStudio

LEARN · BUILD · GROW

www.internstudio.in

team.internstudio@gmail.com