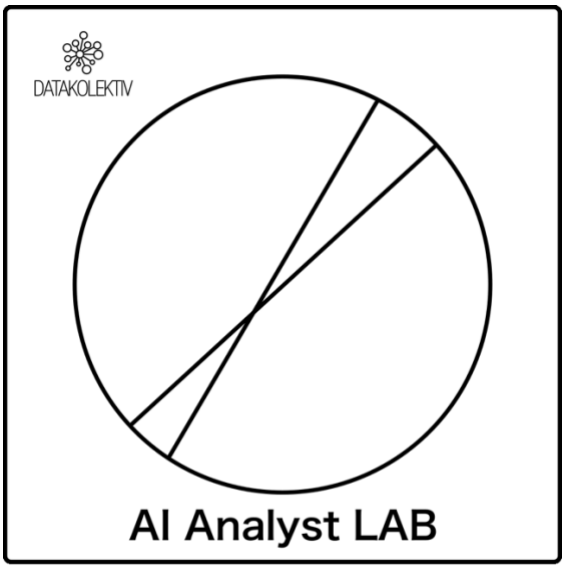
AI Analyst Lab Curriculum



For Analysts, Tech

AI Analyst Lab at a glance

AI Analyst Lab is an eight-week, eight-session, hands-on course designed to help you become fluent in **business data analysis in Python** - while using modern **LLMs** (ChatGPT-style assistants) as learning partners and productivity tools. You will learn the *statistical thinking* that business analysis depends on, the *Python workflow* that makes the analysis real, and the *prompt engineering + API* habits that let you keep learning independently after the course ends.

This course is built for **beginner Data Analysts**:

- You may have **no Python background** (we build fundamentals as we go).
- You do **not** need advanced math (we emphasize intuition, simulation, visualization, and plain-English interpretation).
- You will work in **Visual Studio Code + Python**, and you will learn to use LLMs in two ways:
  - through their **UI** (chat interface),
  - through **API calls** from Python (OpenAI Responses API and Anthropic Messages API patterns).
- You will practice using open datasets with clear licensing, including datasets from the UCI Machine Learning Repository that are explicitly released under **CC BY 4.0** (allowing reuse with proper attribution).

Outcomes and value for you

What you will be able to do after the full course

By the end of AI Analyst Lab, you will be able to:

Run a complete business analysis workflow in Python

- Load and clean messy real-world datasets.
- Perform exploratory data analysis (EDA) and produce clear, decision-ready visualizations.
- Write short, structured business summaries (what happened, why it matters, what to do next).

Use core statistics correctly and confidently

- Explain and compute descriptive statistics (sample vs population, center, spread).
- Recognize when Normal/Binomial/Poisson models are useful and when they are not.
- Use the sampling distribution of the mean (Central Limit Theorem intuition) to reason about uncertainty.

Apply probability in business contexts

- Use conditional probability and expected value to frame risk and decision tradeoffs.

Design and interpret A/B tests

- Use a hypothesis testing framework to choose the right test and interpret results responsibly.
- Understand practical choices like Welch vs classic t-test (when variances differ) and when to use chi-square for categorical outcomes.

Build and interpret baseline ML models used constantly in analytics

- Correlation, simple and multiple linear regression for forecasting and "drivers analysis."
- Binomial (logistic) regression for churn/propensity-style classification.
- K-means clustering for segmentation.
- t-SNE for visualizing cluster solutions in 2D (and communicating them responsibly).
- Decision trees for interpretable prediction and rule-based recommendations.

Use LLMs in a way that makes you more effective (not dependent)

- Build your own "session learning assistant" prompts that teach you on demand.
- Use APIs to generate structured analysis plans, reporting outlines, and stakeholder-ready summaries - *only after you supply computed results*.
- Learn reliable prompting patterns such as checklists, assumption testing, and schema-based structured output.

What you will have in your portfolio

You will finish the course with a practical portfolio you can reuse:

- **8 analysis notebooks** (one per session): cleaned dataset + EDA + model/test + interpretation.

- **8 short stakeholder memos** (Markdown) that summarize the business decision.
- A reusable **"AI-assisted reporting script"** pattern: read computed results → call an LLM API for narrative/report scaffolding → export a clean report draft.

How you will learn in the lab

The teaching style you can expect

Every session is taught as a **real business case**. You will repeatedly practice one core professional rhythm:

1) Clarify the business question (what decision are we making?)
2) Understand and validate the data (what do we trust?)
3) Visualize first (what does the data *look like*?)
4) Apply a statistical/ML method (what's the best tool here?)
5) Interpret in business language (what should we do?)

How LLMs are used in this course

You will use two complementary LLM modes:

UI mode (learning and coaching)

- You will set up a "session tutor" prompt so your assistant teaches with the right level, asks good questions, and keeps you on track.

API mode (repeatable workflows)

- You will learn to call LLMs from Python for tasks like:
- generating structured analysis checklists,
- drafting a report outline in JSON,
- translating computed outputs into a stakeholder narrative.

For OpenAI, we use the **Responses API** as the primary interface and learn schema-based reliability with **Structured Outputs** (JSON Schema adherence).
For Anthropic, we use the **Messages API** via the official Python SDK and learn API patterns (multi-turn, streaming, and reliable request structure).

Responsible-use habit you will practice every week

LLMs are powerful at drafting and explaining, but they are not calculators. In this course you will develop a repeatable "LLM + verification loop":

- The model proposes a plan → **Python computes** → you sanity-check → the model helps communicate the result.
- When you need machine-readable output, you use **schemas** or strict JSON constraints instead of "free-form text."

Weekly curriculum roadmap

Below is what will happen each week, written for you as a prospective participant: what you will do, why it matters, what you will be able to do immediately after the session, and what to study (StatQuest videos + Khan Academy resources) to make the most of the lab.

Session 01.

Session focus: Demand and operations analytics for a bike-sharing business

Statistics focus: sample vs population, descriptive statistics, Normal/Binomial/Poisson intuition, sampling distribution of the mean (Central Limit Theorem intuition)

Business case (why you care). You are the analyst for a city bike-sharing operator. Stations sometimes run empty, and leadership wants to understand **demand patterns** and **uncertainty** so they can staff operations and reposition bikes more intelligently.

Dataset (open). Bike Sharing Dataset (CC BY 4.0) from the UCI Machine Learning Repository.

What will happen in the session. You will load the dataset into Python, build your first "EDA notebook," and learn how to summarize a business process using descriptive statistics and visualizations. You will explore count-like outcomes and discuss why Poisson/Binomial/Normal are different "stories" about how the world generates data.

After this session, you will be able to:

- Create a clean Python notebook that loads a dataset, checks columns/types, and produces a first EDA.
- Produce the "analyst's starter toolkit" of plots (histograms, boxplots, time patterns) and explain them clearly.
- Explain the difference between a sample and a population, and why uncertainty exists in business reporting.
- Simulate sampling and see why averages become more stable (sampling distribution intuition / CLT).

Prompt engineering + API progression.

- UI: You will create your **Session Tutor** prompt that teaches gently, avoids heavy math, and asks you to confirm what you computed before interpreting.
- API: You will make your first successful API call (OpenAI or Anthropic) to generate:
    - a structured EDA checklist, and
    - a plain-English explanation template **after** you paste your computed stats.

StatQuest videos to study (StatQuest).

- Histograms, Clearly Explained.
- The Main Ideas behind Probability Distributions.
- The Normal Distribution, Clearly Explained.
- Population and Estimated Parameters.
- Calculating the Mean, Variance and Standard Deviation.

- Sampling from a Distribution, Clearly Explained.
- The Binomial Distribution and Test.
- The Central Limit Theorem, Clearly Explained.
- Standard Deviation vs Standard Error.
- The Standard Error.

Khan Academy resources to study.

- Statistics & Probability (entry point for descriptive stats and visuals).
- Histograms review.
- Box plot review.
- Measures of spread: variance & standard deviation.
- Normal distribution introduction.
- Sampling distributions library (and CLT practice).

Session 02.

Session focus: Risk assessment with probability

Statistics focus: probability rules, conditional probability, expected value, uncertainty via simulation/bootstrapping

Business case (why you care). You have to write a risk brief for leadership: "When is the risk of severe incidents higher?" You will learn how probability becomes a decision-support tool.

Dataset (open). Great Britain Road Safety open data (STATS19) published by the UK Department for Transport under the Open Government Licence; the official description notes it covers personal injury collisions reported to police and recorded via STATS19.

What will happen in the session. You will compute conditional probabilities ("given nighttime, what is the severity mix?"), compare scenarios, and translate probability results into a risk narrative. You will also learn what data like STATS19 includes - and what it systematically misses - so you can state limitations responsibly.

After this session, you will be able to:

- Turn raw event data into risk rates and conditional risk tables (and visualize them clearly).
- Use conditional probability to explain "risk changes under conditions," not just overall averages.
- Use expected value as a simple decision tool (e.g., expected severe cases per period).
- Use simulation/bootstrapping to show uncertainty without heavy math.

Prompt engineering + API progression.

- UI: You will create a **Risk Analyst Assistant** prompt that forces clarity: outcome definition, exposure definition, and data limitations before recommendations.
- API: You will generate a structured "risk brief" outline (JSON headings + required tables/plots), validating the JSON in Python (schema checks are part of professional practice).

StatQuest videos to study (StatQuest).

- Conditional Probabilities, Clearly Explained.
- Bayes' Theorem, Clearly Explained.
- Expected Values, Main Ideas.
- Expected Values for Continuous Variables.
- In Statistics, Probability is not Likelihood.

Khan Academy resources to study.

- Probability library (conditional probability and independence).
- Tree diagrams and conditional probability.
- Conditional probability and independence.
- Expected value basics.
- Random variables unit (expected value practice).

Session 03.

Session focus: A/B testing in growth analytics

Statistics focus: hypothesis testing framework, t-tests (Welch awareness), non-param alternatives conceptually, chi-square for categorical outcomes, multiple testing awareness

Business case (why you care). You work with a growth team testing headlines and creatives. Your job: decide what to ship and explain the confidence level *without overclaiming*.

Dataset (open). The Upworthy Research Archive: large-scale A/B tests of headlines conducted by Upworthy (2013–2015). The archive is documented publicly, and the site notes a June 2024 update describing randomization issues affecting a subset of tests in a specific window, with guidance to avoid confirmatory use of that period.

What will happen in the session. You will learn the hypothesis testing workflow end-to-end: define a metric, define null/alternative, choose a test based on outcome type, check assumptions, compute results, and communicate them. You will also learn what "p-hacking" looks like in practice and why multiple comparisons matter when you look at many experiments.

After this session, you will be able to:

- Translate an A/B test into a clear statistical question and decision rule.
- Choose appropriate tests for common business outcomes (means vs proportions).
- Communicate p-values, confidence intervals, and practical significance clearly.
- Write "what we know / what we don't know / what to test next" with credibility.

Prompt engineering + API progression.

- UI: You will create an **Experiment Coach** prompt that forces you (and the model) to specify: metric type, unit of analysis, number of variants, assumptions, and multiple-testing risk before claiming a winner.

- API: You will generate a structured A/B test analysis plan as JSON, using schema-based reliability (OpenAI Structured Outputs) *or* strict JSON validation in Python, then auto-generate a stakeholder memo draft from your computed results.

StatQuest videos to study (StatQuest).

- Hypothesis Testing and the Null Hypothesis.
- Alternative Hypotheses: Main Ideas.
- p-values: What they are and how to interpret them.
- How to calculate p-values.
- StatQuickie: Which t test to use.
- Using Linear Models for t-tests and ANOVA.
- Confidence Intervals, Clearly Explained.
- p-hacking: What it is and how to avoid it.
- False Discovery Rates (FDR), clearly explained.
- Statistical Power, Clearly Explained + Power Analysis.
- Bootstrapping (Main Ideas + p-values).

Khan Academy resources to study.

- Significance tests (hypothesis testing) unit.
- Z-statistics vs T-statistics.
- Introduction to t statistics.
- Two-sample t test for difference of means.
- Inference for categorical data (chi-square tests).
- Chi-square statistic video.

Session 04.

Session focus: Drivers analysis with correlation and regression

Statistics/ML focus: correlation, simple linear regression, multiple linear regression, interpretation and diagnostics

Business case (why you care). You need to answer a classic business question: “What drivers are associated with quality, cost, or performance - and how confident are we?” This is the backbone of many analyst roles.

Dataset (open). Wine Quality dataset from the UCI Machine Learning Repository (CC BY 4.0).

What will happen in the session. You will learn how to interpret relationships without overclaiming causality: correlation as association, regression coefficients as “all else equal” relationships (carefully explained), and how to validate a model with visual diagnostics and held-out evaluation.

After this session, you will be able to:

- Create correlation and regression visuals that directly support a business narrative.

- Fit and interpret a simple regression and a multiple regression in Python.
- Explain what R-squared does and does not mean.
- Use residual-style thinking (what the model keeps missing) to talk about limitations.

Prompt engineering + API progression.

- UI: You will create a **Regression Interpreter** prompt that translates coefficients into plain English and always asks for diagnostic plots before conclusions.
- API: You will send your computed coefficients / metrics to the API and generate a "drivers memo" section with:
- key findings,
- caveats (confounding, data limitations),
- and what data would strengthen the conclusion.

StatQuest videos to study (StatQuest).

- Covariance, Clearly Explained.
- Pearson's Correlation, Clearly Explained.
- R-squared, Clearly Explained.
- The Main Ideas of Fitting a Line to Data (Least Squares).
- Linear Regression, Clearly Explained.
- Multiple Regression, Clearly Explained.

Khan Academy resources to study.

- Exploring bivariate numerical data unit.
- Correlation coefficient review.
- Calculating correlation coefficient r.
- Linear regression review.
- Regression line example.
- Covariance and the regression line.

Session 05.

Session focus: Binomial regression for classification

ML focus: logistic regression, probability outputs, thresholds, evaluation (confusion matrix, sensitivity/specificity)

Business case (why you care). You're asked to reduce churn. You must predict who is at risk and decide how to intervene under limited capacity (e.g., "we can contact only 10% of customers").

Dataset (open). Iranian Churn dataset from the UCI Machine Learning Repository (CC BY 4.0).

What will happen in the session. You will build a logistic regression model in Python, interpret it as a probability model, and learn how model evaluation ties to business cost tradeoffs (false positives vs false negatives).

After this session, you will be able to:

- Build a baseline churn model that outputs churn probabilities.
- Choose and justify a decision threshold based on business constraints.
- Explain performance using confusion matrix thinking (not just "accuracy").
- Produce a simple churn triage policy (who gets outreach first).

Prompt engineering + API progression.

- UI: You will create a **Classification Decision Coach** prompt that forces a cost/benefit discussion and prevents the assistant from inventing model results.
- API: You will produce a structured "action policy" JSON (tiers, thresholds, messages) and generate a stakeholder-friendly explanation of how to interpret probabilities and limitations.

StatQuest videos to study (StatQuest).

- Odds and Log(Odds), Clearly Explained.
- Odds Ratios and Log(Odds Ratios).
- Logistic Regression.
- Machine Learning Fundamentals: The Confusion Matrix.
- Machine Learning Fundamentals: Sensitivity and Specificity.
- ROC and AUC (optional but strong for classification intuition).

Khan Academy resources to study.

- Confidence intervals unit (helps you talk about uncertainty in business recommendations even when modeling).
- Statistics & Probability hub (for reinforcement and practice).
- Inference for categorical data (chi-square tests) as background for categorical outcomes and contingency thinking.

(Note: Khan Academy's core statistics track is excellent for probability, inference, and regression foundations; for logistic regression specifically, your primary external video track in this course is StatQuest.)

Session 06.

Session focus: Customer segmentation with k-means clustering

ML focus: k-means, scaling intuition, choosing k, cluster profiling and business actions

Business case (why you care). You are asked to segment customers so your company can tailor sales and retention strategies. The output is not a "model score" - it's a **usable segmentation story**.

Dataset (open). Wholesale customers dataset from the UCI Machine Learning Repository (CC BY 4.0).

What will happen in the session. You will build a clustering pipeline in Python, learn why scaling matters for distance-based clustering, and produce cluster profiles that can be turned into actions.

After this session, you will be able to:

- Build a k-means pipeline that you can reuse for segmentation work.
- Choose a reasonable number of clusters using visual heuristics (and explain the tradeoff).
- Profile clusters using summary tables and visualization (who is in each segment).
- Translate cluster profiles into "what we should do differently for each segment."

Prompt engineering + API progression.

- UI: You will create a **Segmentation Strategist Assistant** prompt that requires evidence-based segment naming (it must reference the summary stats you provide).
- API: You will generate a structured "segment profile" JSON and a short sales playbook draft per segment.

StatQuest videos to study (StatQuest).

- A Gentle Introduction to Machine Learning.
- K-means clustering.

Khan Academy resources to study.

- Statistics & Probability hub (for reinforcing data visualization and interpretation skills that make segmentation credible).
- Clusters in scatter plots (conceptual grounding for what "clusters" look like visually).

Session 07.

Session focus: Visualizing and communicating clusters with t-SNE

ML focus: t-SNE as a visualization method, parameter sensitivity (explained without heavy math), stakeholder-ready storytelling

Business case (why you care). You already have clusters. Now you must *sell the segmentation* to non-technical stakeholders and make it understandable without jargon.

Dataset (open). Same wholesale customers dataset (continuity helps you deepen interpretation rather than starting over).

What will happen in the session. You will run t-SNE to create a 2D embedding, overlay cluster labels, and learn how to communicate t-SNE responsibly (it is excellent for visualization, but you must avoid overstating what distance means).

After this session, you will be able to:

- Produce a 2D visualization that makes your clusters explainable in a meeting.
- Write stakeholder narrative: what the segments are, and how they differ operationally.
- Explain the limitation: why different settings can produce different pictures (without panic).
- Build a "cluster story pack" you can reuse as a template for future segmentation projects.

Prompt engineering + API progression.

- UI: You will create a **Stakeholder Translator Assistant** prompt that turns segment stats into a concise narrative with "so what" actions.
- API: You will auto-generate a segment story outline (headings + charts to include), suitable for a slide deck draft.

StatQuest videos to study (StatQuest).

- t-SNE, Clearly Explained.
- PCA, Step-by-Step (recommended as a comparison point for "dimension reduction" intuition).

Khan Academy resources to study.

- Statistics & Probability hub (practice on reading/communicating plots clearly).
- Box plot review (useful for comparing segments with robust summaries).

Session 08.

Session focus: Decision trees for regression and classification

ML focus: interpretable models, tree visualization, overfitting control, basic evaluation, turning models into decision rules

Business case (why you care). Many business teams prefer "if-then" logic they can understand. Trees are a gateway to interpretable decision-making, and they are also the foundation of many modern methods. This week is about **usable rules**, not black-box mystique.

Datasets (open).

- Wine quality (regression tree).
- Iranian churn (classification tree).

What will happen in the session. You will build both a regression tree and a classification tree, visualize shallow trees, learn how pruning/depth limits reduce overfitting, and convert a trained tree into a simple decision policy that leadership can understand.

After this session, you will be able to:

- Train a decision tree for either regression or classification and visualize it cleanly.
- Explain overfitting in plain language and apply practical controls (depth/pruning).
- Convert a model into a set of decision rules and recommended actions.
- Produce a final "AI-assisted analytics report" combining Python outputs + API-generated narrative based on your computed results (capstone-style workflow).

Prompt engineering + API progression.

- UI: You will create a **Model Governance Assistant** prompt that drafts a short model summary (intended use, limitations, monitoring ideas).
- API: You will generate a structured report skeleton and fill it with your computed results, using schema-based outputs for consistency and reliability.

StatQuest videos to study (StatQuest).

- Entropy (for decision tree intuition).
- Decision and Classification Trees, Clearly Explained.
- Decision Trees Part 2: Feature Selection and Missing Data.
- Regression Trees, Clearly Explained.
- How to Prune Regression Trees.
- Classification Trees in Python, from Start to Finish.
- Machine Learning Fundamentals: Cross Validation (recommended for evaluation maturity).

Khan Academy resources to study.

- Decision tree exploration (conceptual/interactive grounding).
- Statistics & Probability hub (for inference and interpretation practice).

Datasets and licensing transparency

A key part of this course is that you can legally reuse the same datasets and notebooks later (in your portfolio, in internal demos, and in learning projects).

- Bike Sharing Dataset (UCI): CC BY 4.0.
- Wine Quality dataset (UCI): CC BY 4.0.
- Iranian Churn dataset (UCI): CC BY 4.0.
- Wholesale Customers dataset (UCI): CC BY 4.0.
- Great Britain Road Safety open data (STATS19): published openly; official description emphasizes police-reported personal injury collisions recorded via STATS19, and the dataset is listed with an Open Government Licence on data.gov.uk.
- Upworthy Research Archive: the archive site discusses the dataset and explicitly states it is published under CC BY 4.0 by Cornell University, and it includes guidance about a randomization issue window (June 25, 2013 to January 10, 2014) that you will learn to handle responsibly in analyses.