Morey 2U Interview Assignment: Junior Data Scientist

Section	Goals	Skills	Dataset	Short Description	Example Learning Objective	Relevant Tools and Technology
Detective Work: Data Retrieval & Decomposition	Decompose abstract business questions into specific data requirements. Separate "What do I do?" from "How do I do?" Develop a mental model of Relational Databases.	Translate natural language questions into executable SQL queries. Filter, sort, and aggregate data using WHERE, GROUP BY. Combine tables using JOINs. Structure complex logic using CTEs and Window Functions.	E-commerce Sales Logs (SQL Database)	With a SQL database at the core, students explore data fundamentals by answering broad business questions (e.g., "Why did sales drop in November?"). The focus is on translating abstract business questions into precise syntax. Syntax is taught explicitly, and help from the in house AI assistant is encouraged but focused on leading students to a conclusion rather than just giving the answer.	Given a vague email about "slumping sales," write a SQL query to isolate the specific regions/products responsible.·	SQL (PostgreSQL) PG Admin (or a hosted cloud solution) AI Instructional Companion
Wrangling: Cleaning & Organizing Data	Use DataFrames to build, modify, and analyze datasets. Quantify the business impact of "Dirty Data." Automate data ingestion from disparate sources.	Ingest tabular data from Excel files. Ingest JSON data from REST APIs. Clean nulls and fix types using Polars. Perform high-performance transformations on large datasets. Establish Python-to-SQL connections for ETL.	Weather API City Event Excel Data	Students move to Python to access Excel and API data along with SQL data. Using Pandas, students build ETL pipelines that ingest raw weather data, clean it, and load it into the SQL DB. A key focus is comparing "clean" vs "dirty" analysis, emphasizing how dirty data can still lead to conclusions, but faulty ones.	Write a Python script to collect real time weather data via API and merge it with static Excel events, storing the resulting data back in PostgreSQL. Use Polars to clean a 1M row dataset.·	Python 3.10+ Pandas (or Polars) Jupyter Lab Requests Library SQLAlchemy
Science: Tests & Metrics	Validate business hypotheses using statistical tests. Distinguish between "Statistical Significance" and "Business Relevance." Detect anomalies that skew analysis.	Use stats, statsmodels, and SciPy libraries Calculate descriptive stats (Mean, Variance). Identify outliers using Z-Scores/IQR. Perform Hypothesis Testing (T-tests, Chi-Squared). Interpret p-values in a business context.	Customer Churn Data	Before predicting the future, students must understand the present. This module focuses on validation: using math to confirm if a pattern is real or just noise. Students must statistically justify their answers to business questions.	Conduct a T-test to see if churn rates differ significantly between two subscription tiers.· Use Box Plots to identify and remove outliers.· Write a report explaining why a 2% conversion increase is (or isn't) significant.	Numpy SciPy Stats Statsmodels Seaborn Jupyter Lab
Storytelling: Ingestion to Visualization	Tailor technical findings to non-technical audiences. Apply design principles (e.g., Tufte) to minimize cognitive load. Differentiate between "Exploratory" and "Explanatory" visuals.	Connect BI tools directly to SQL for live reporting. Create Calculated Fields and LOD expressions. Design interactive dashboards with drill-downs. Present findings in video format.	Sales Data (Callback to Section 1)	Students shift from analysis to communication. Using Tableau/Power BI, they learn to visualize the Science and Detective findings. Lessons focus on design principles and tailoring the delivery format to specific stakeholders.	Design an executive dashboard that shows the health of the business at a glance.· Critique and refactor a "bad" visualization.· Record a 3-minute video summary for the VP of Sales.	Tableau Public/Power BI Google Slides Loom/Zoom
AI Agents: Foundations	Understand architecture/limitations of LLMs. Move beyond chat to programmatic AI interaction. Orchestrate complex logic flows. Describe possible business risks and ethical concerns that accompany AI Agents.	Construct prompts using Chain-of-Thought. Interact with LLMs via OpenAI/Gemini APIs. Build sequential logic chains using LangChain.	Customer Reviews (Text Corpus)	Introduction to the modern AI stack. Students move beyond "chatting" with bots to programming them. We cover how LLMs work, how to connect them to data, and how to use LangChain to orchestrate logic.	Write a Python script using OpenAI API to perform sentiment analysis on reviews.· Build a LangChain pipeline to summarize documents and extract entities.· Compare cost/latency of different prompts.	Chat GPT/Gemini LangChain Python
AI Agents: Applications	Operationalize analysis into self-service tools. Implement software deployment best practices (CI/CD). Evaluate "Buy vs Build" for AI tools.	Build interactive web apps using Streamlit. Integrate SQL/Python into a UI. Deploy apps to cloud via Git. Implement input validation.	Integrated Course Data	Students wrap their analysis or agent into a usable product using Streamlit. Emphasis is placed on CI/CD and deployment making the work reproducible and accessible via a public URL.	Build a "Chat with your Data" Streamlit app converting natural language to SQL.· Deploy the app to the cloud with auto-redeploy on Git push.· Add user filters to the app sidebar.	Streamlit Git and GitHub
Machine Learning: Under the Hood	Demystify ML as "Automated Statistics." Understand trade-offs (Complexity vs. Interpretability). Master the full ML lifecycle.	Preprocess features (Encoding, Scaling). Train Supervised models (LogReg, Random Forest). Evaluate models using Precision/Recall/ROC-AUC. Interpret Feature Importance.	Customer Churn (Callback to Section 3)	Students apply ML algorithms, treating Scikit-Learn as automated statistics. We reuse the Churn dataset to transition from inference (Section 3) to prediction, emphasizing evaluation metrics.	Train a Random Forest to predict churn with ROC-AUC > 0.8.· Use Grid Search to optimize hyperparameters.· Generate a Feature Importance chart to explain drivers of churn to stakeholders.	Scikit-Learn Imblearn
Conclusions: The Future Data Scientist	Cultivate a mindset of continuous learning. Develop a strategic career roadmap. Understand ethical responsibilities.	Synthesize a professional portfolio. Evaluate emerging tools (Vector DBs, AutoML). Articulate ethical considerations (bias/privacy).	N/A	A reflective capstone. We discuss how tools like AutoML change the junior role from "Code Writer" to "Code Reviewer" and "Problem Solver."	Compile all course projects into a unified portfolio.· Write a reflective essay on the ethical implications of the Churn model.· Create a 6-month self-directed learning plan.	GitHub Pages