Project Brief:
The Data Consultant Challenge

1. Project Overview & Business Case

In teams of 2–4, you are working as **ML Strategy Consultants**. Your goal is to identify a high-impact business problem within a raw dataset and build a machine learning pipeline that provides both **strategic classification** and **operational regression** solutions.

The "Why" Document & Presentation (Part 1)

Before you write a single line of code, you must understand and produce the **Business Rationale** that covers:

- **The Opportunity:** Identify the specific "AI Use Case." (e.g., "We are building a tool for record labels to filter talent and forecast revenue.")
- **The Problem Statement:** What pain point does this address? (e.g., "Manual scouting is slow and revenue projections are currently based on guesswork.")
- **Business Impact:** Why should a CEO care? Define the potential impact (e.g., "Reducing risk in signing new artists and increasing marketing ROI by 15%").

Dataset(s) Selection Process

You can select ***any*** dataset(s) (either one joint dataset for classification **and** regression, **or** dataset for each case - note, this may involve more work). Alternatively, use any of the following suggestions for inspiration:

1. **Supply Chain / Logistics:** Predict "Late Delivery" (Class) and "Arrival Delay Time" (Reg).
2. **Healthcare Operations:** Predict "Readmission Risk" (Class) and "Length of Stay" (Reg).
3. **Real Estate Investment:** Predict "Property Tier" (Class) and "Appraised Value" (Reg).
4. **E-commerce Customer Value:** Predict "Churn Risk" (Class) and "Total Lifetime Spend" (Reg).
5. **Hotel Booking Demand:** Predict "Cancellation" (Class) and "Lead Time/Revenue" (Reg).
   Challenge: High seasonality and sensitive to time-series trends.
6. **Telco Customer Churn:** Predict "Churn" (Class) and "Total Charges" (Reg).
   Challenge: Heavily imbalanced classes (fewer people churn than stay).
7. **NYC Taxi Trips:** Predict "High Tip" (Class) and "Trip Duration" (Reg).
   Challenge: Massive dataset that requires efficient processing.

---

2. Phase A: Technical & Strategic Foundations

This phase is worth the most points, as it sets the stage for the entire pipeline.

I displayed 5 to-the-point plots in total

Did it for outliers tracking

I display a jointplot using average monthly charges and tenure in relation to is_Churned. In our case, where the problem is simple, we dont need a pairplot. Jointplots reveal exactly the same information without making things more complex.

- **EDA Storytelling:** Use at least 3 types of plots (e.g., Heatmaps, Boxplots, Pairplots) to explain the data's "current state." **Requirement:** Every plot must have a caption explaining what it highlights to a business stakeholder.
- **Feature Engineering & Justification:**
  - **New Feature engineering / creation:** Engineer at least one new feature. *Justify why this new variable helps the business problem.*
  - ~~**Categorical Target:** Find or create a classification label. *Explain the business logic if you used a threshold* (e.g., why is a "Hit" defined as the top 20% and not the top 10%?).~~

I created the normalized by tenure average monthly charges column which may be better than total charges for predictions. This also helps us apply basic feature selection (filtering).

Already have a class target (Is_Churned)

- **Dimensionality Reduction & Cleanup:** Investigate the use of techniques like **PCA** or **Correlation Filtering** to reduce the number of features. *Explain why removing "noise" could make the model more robust for a real-world environment.*
- **Imputing**: Conduct one or more imputing strategies (and justify why) in the case of null values
- **Transformation:** Apply scalers, outlier removal and/or handle categorical encoding appropriately.

I applied a simple feature selection process only based on .corr. We also reduced features through pd.get_dummies(drop_first=True). These sounds ok to assume we get all points.

You can apply whatever scaler you like during your model implementation. I think it keeps team work issues simpler.

KIMON's FINAL COMMENT:
I strongly believe that we dont need PCA to get 100%. If someone wants to apply it just for an extra work just to test model performance, without prioritizing interpretability can do so. I believe that we are ok as we are right now for this stage.

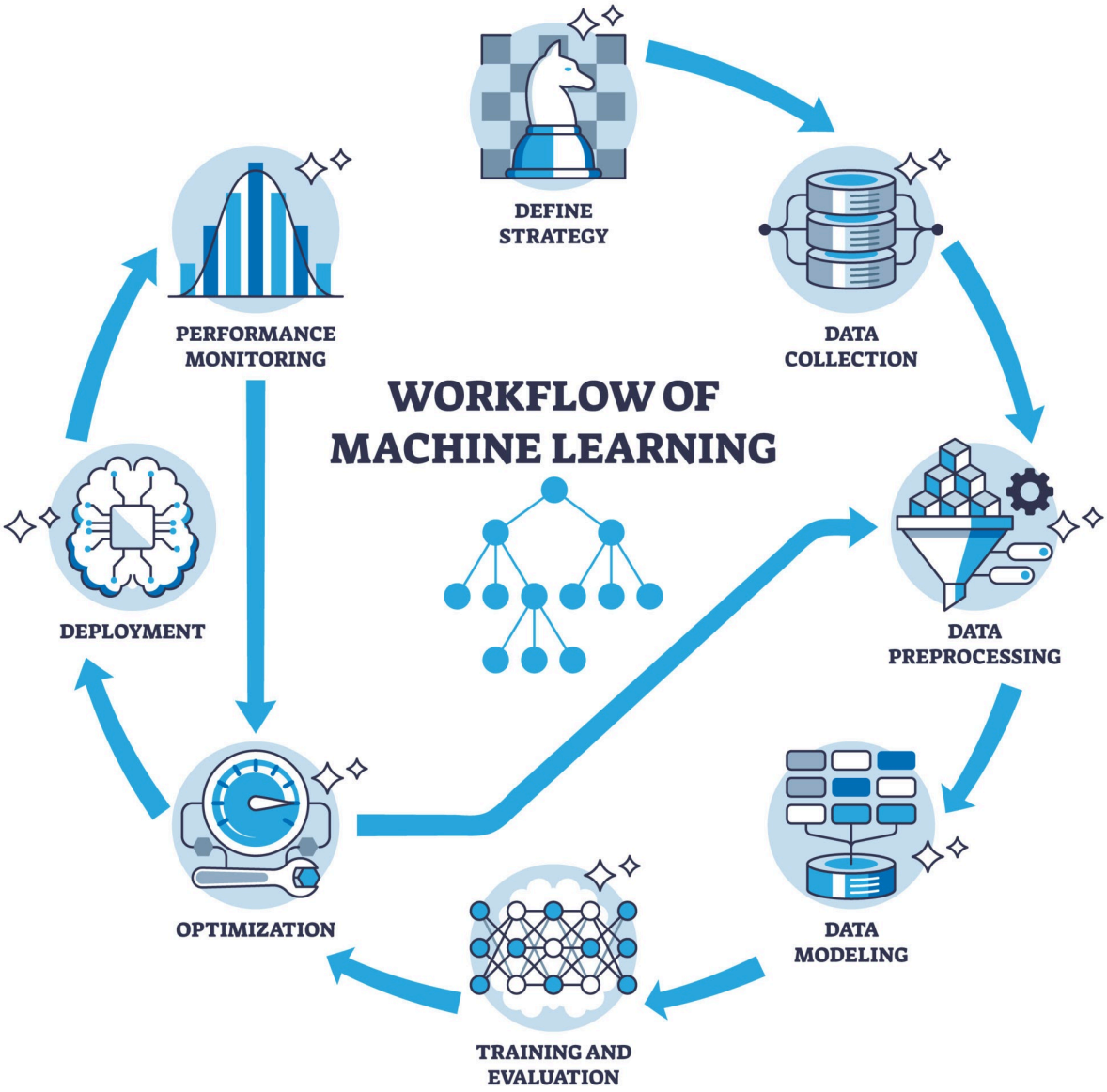
We reduced dimensionality a bit using correlation. I didn't apply PCA.

3. Phase B & C: The Predictive Engines

Requirement	Classification (Part B)	Regression (Part C)
Model Count	2 Different Models (optimised)	2 Different Models (optimised)
Logic	Predict a category (e.g., "High Risk")	Predict a number (e.g., "Loss Amount")
Metrics	Accuracy, F1-Score & Confusion Matrices	MAE and R^2 Score
Explainability (new domain)	**XAI (use SHAP or LIME):** Identify the top 3 drivers of a decision.	**XAI:** Show which feature most impacts the numeric output.

4. Grading (Total: 100 Points)

Criteria	Points	Requirements for Full Marks
Business Rationale & Dataset Selection	15	Clear link between the dataset, the ML solution and a real-world business impact.
EDA & Storytelling	15	Visuals that explain trends and "justify" the project's direction.
Feature Engineering	20	Engineering new features, imputing, scaling, PCA/ Feature Selection with logical justification.
Modeling Quality	30	Correct implementation of all 4 models (2 Clf, 2 Reg) and associated metrics.
Explainable AI (XAI)	10	Using SHAP/LIME to prove *why* the model should be trusted.
Technical + Document Clarity	10	Code is clean and reproducible. Everything is well-documented and justified in the report and presentation.



Student Self-Assessment Checklist

- [ ] Did we define a **Business Problem** that justifies the AI use case?
- [ ] Did we explain the **Potential Impact** (Revenue, Efficiency, Risk)?
- [ ] Do our EDA plots have "Business Insight"?
- [ ] Did we use **Dimensionality Reduction** and explain *why* we dropped specific columns?
- [ ] Did we create any new **Features** and explain the logic behind its creation?
- [ ] Do we have SHAP/LIME plots explaining a specific "Black Box" prediction?