Project Task: Week 1
Data Science/Data Analysis

Data Collation and Cleaning:

Collate all files to consolidate the dataset in a single location for ease of analysis.

Inspect the dataset for missing values and address them appropriately.

Identify rows with trivial values (e.g., “?”) and calculate their percentage in the dataset. Remove such rows if they lack significant information.

Data Transformation:

Apply suitable transformation techniques for nominal and ordinal categorical variables.

Address the imbalance in state representation, focusing only on states R1011, R1012, and R1013 for dummy variable creation.

Clean the variable "NumberOfMajorSurgeries" by converting string values to a numeric format.

Derived Variables:

Calculate patients' ages using their dates of birth.

Determine patient gender based on the salutation in their name and create a new gender field.

Data Visualization:

Visualize cost distribution using histograms, box-and-whisker plots, and swarm plots.

Compare cost distribution across gender and hospital tiers.

Create a radar chart to show the median hospitalization cost for each hospital tier.

Generate a frequency table and a stacked bar chart to display the count of people across different city and hospital tiers.

Hypothesis Testing:

Test the following null hypotheses:
a. No significant difference in average hospitalization costs across hospital types.
b. No significant difference in average hospitalization costs across city tiers.
c. No significant difference in average hospitalization costs between smokers and non-smokers.
d. Smoking and heart issues are independent variables.

Use both Excel and Python for these analyses.

Project Task: Week 2
Machine Learning

Correlation Analysis:

Identify highly correlated predictors and visualize the relationships using a heatmap.

Model Development:

Build a regression model using a stochastic gradient descent optimizer.

Apply stratified 5-fold cross-validation.

Standardize data and perform hyperparameter tuning.

Use sklearn pipelines and implement regularization techniques to balance bias and variance.

For each fold, train on 80% of the data and validate on 20%, generating five distinct models and root mean squared error (RMSE) values.

Determine variable importance scores and identify redundant variables.

Advanced Model Building:

Build and evaluate Random Forest and Extreme Gradient Boosting models for cost prediction.

Share cross-validation results and calculate variable importance scores.

Case Scenario Prediction:

Estimate hospitalization costs for Ms. Jayna based on her provided demographic and health information.

Predict costs using the five models and calculate the mean of their predicted values.

Project Task: Week 2
SQL

Data Integration:

Merge tables by identifying columns suitable for creating a Primary Key constraint.

Remove duplicates and null values, then use the ALTER TABLE command to add the constraint.

Data Retrieval and Analysis:

Retrieve information about diabetic patients with heart problems, including their average age, number of dependent children, BMI, and hospitalization costs.

Calculate average hospitalization costs per hospital tier and city level.

Count individuals with major surgery and a history of cancer.

Determine the number of tier-1 hospitals per state.

Project Task: Week 2
Tableau

Dashboard Creation:

Develop a Tableau dashboard using appropriate chart types and business metrics.

Emphasize data storytelling for clear and impactful visual representation.