Employee Turnover Analytics assessment is organized and approached as follows:

1. Data Inspection & Quality
   1. Built a function to conduct a complete inspection of the data, printing out basic checks for possible investigation
2. Exploratory Data Analysis
   1. HeatMap of Correlation Matrix
   2. Univariate Analysis
      1. Distribution of Selected Features
   3. Bivariate Analysis: # of Projects vs. Left
3. K-Means Clustering was performed by filtering the dataframe for those that left then proceeding to use K-Means clustering to find 3 clusters based on satisfaction and evaluation.
   1. From this we identified 3 clusters and plotted to see the branching from the centroid and built a dataframe describing the means of the clusters
4. Data Pre-Processing
   1. Performed by encoding the categorical values using get\_dummies and attaching it to our dataframe
   2. Checked imbalance of our target by plotting and getting ratio than applying SMOTE after splitting with designated stratification and random state
5. Model Training
   1. Built a loop to fit our 3 models using the Stratified K-Fold Cross Validation set at 5 splits
6. Evaluate Model
   1. Above loop also ploted the ROC curve to determine the AUC
   2. Further evaluation was performed by displaying the confusion matrix and classification report from each model that we trained
      1. **Recall is favored over precision** as we favor a high true positive rate, meaning that a false negative (predicting that they will not leave) when they actually are very likely to leave can be more problematic.
   3. Given this, we will proceed with the **\*\*Random Forest Classifier\*\***
7. Predicted using our random forest model for both probabilities and classes
   1. Created a function that would categorize our probabilities into “Risk Zones”
   2. Investigated Our Categorizations in Various Ways:
      1. Model feature importances (probabilities), SHAP summary plots (class) to understand feature importance
      2. Distribution of our risk zones across the probabilities but also across the Top 5 Feature Importance
      3. Quick check of correlation analysis to see if any further insight gained
      4. Understanding the characteristics of our risk zones through descriptive statistics and plotting the means
      5. Performing a comparison of risk zone 1 (risk zone 1) vs risk zone 4 (>90%) to understand opposite sides of the spectrum

Preliminary Recommendation for Retention Strategies:

* Based on the high training and testing metrics we used a Random Forest Classifier to predict feature importances and obtained the top 5 features which contribute to 94% of explaining the prediction probabilities.
* Evaluating the mapped risk zones (1=<20%, 2=20-60%, 3=60-80%, 4 =>90%) to the probabilities of leaving and analyzed them according to these 5 features.
* The 3 clusters we saw in our K-means clustering analysis with Employee Satisfaction and Employee Evaluation Scores were seen again in our plotted distribution, so it is important to begin to see when these clusters begin to emerge.

At year 2, preventative retention strategies can be enabled and continued especially through year 5 that revolve around:

1. Check satisfaction levels: particularly for scores 0, 0.3-0.5, and 0.7-0.9 this will make the most difference in retention as time with the company extends to through year 5
2. In regards to evaluation scores, while scores between 0.45-0.55 are of particular importance, even when evaluations were high (between 0.75-1.00) it is important addition to ensure their satisfaction is high.
3. It is important to monitor that monthly hours are between 170 to 210 as there are particular increases in the probability of leaving when below or above those marks.
4. The same can be said for the range of projects where 3 is the ideal, particularly below 2 and still above 4 there is an increase in the ratio of wanting to leave.