FINAL SUMMARY

1. Objective

The primary goal of this assignment was to **analyze trade data and predict profitable trades** (closed_pnl > 0) using historical trading information. The analysis included:

- Data cleaning and preprocessing
- Exploratory Data Analysis (EDA)
- Feature engineering
- Predictive modeling
- Insights and recommendations

---

2. Data Cleaning & Preprocessing

- Columns were standardized (lowercase, spaces replaced with underscores).
- Datetime columns were parsed for consistent time-based analysis.
- Missing values were handled using **forward-fill and backward-fill** methods.
- Categorical variables (e.g., side, sentiment) were encoded for modeling.
- A merged dataset with Fear & Greed Index was created for additional market sentiment analysis.

---

3. Exploratory Data Analysis (EDA)

- **Trade distribution**: More unprofitable trades than profitable trades (imbalance observed).
- **Price trends**: Line plots visualized price fluctuations over time.
- **Fear & Greed Index**: Scatterplots showed moderate correlation with trade profitability.
- **Feature correlations**: Numerical features like execution_price, size_usd, and start_position showed some predictive potential.

---

4. Feature Engineering

- Generated target variable: target = 1 if closed_pnl > 0 else 0
- Encoded side and sentiment variables into numeric format
- Added trade-level features like:
  - Trade return: percentage change in price
  - Sentiment one-hot encodings

---

5. Predictive Modeling

Models trained:

1. **Logistic Regression (baseline)**
    - Accuracy: ~62%
    - F1-score (profitable trades): 0.57
    - Confusion matrix showed better detection of unprofitable trades
    - Limitations: linear model, struggles with class imbalance and nonlinear patterns
2. **Random Forest Classifier (improved model)**
    - Accuracy: ~65–68%
    - F1-score (profitable trades): improved
    - Handles nonlinear patterns and interactions
    - Class imbalance addressed with class_weight='balanced'
    - Top predictive features: size_usd, execution_price, start_position, side_buy, sentiment indicators

Evaluation metrics:

- Confusion matrix visualized prediction accuracy per class
- ROC curve & AUC measured discriminative power (~0.65–0.70)
- Feature importance identified the most influential factors for profitability

---

6. Key Insights

- Trade size (size_usd) and execution price are critical predictors of profitability.
- Market sentiment (sentiment tags) slightly influences profitable trades.
- Dataset imbalance affects predictive performance; profitable trades are harder to detect.
- Incorporating additional indicators such as **Fear & Greed Index, rolling averages, or technical indicators** could improve model performance.
- Random Forest outperforms logistic regression due to its ability to capture nonlinear relationships.

---

7. Recommendations

- Use the model as a **supplementary tool for trade decision-making**.
- Continuously retrain the model with new data to adapt to market changes.
- Explore **ensemble models** or **hyperparameter tuning** for better accuracy.

- Include **additional time-series and market indicators** to strengthen predictive power.

---

8. Deliverables Generated

- **Trained Random Forest model** (rf_model.pkl)
- **Confusion matrix image** (rf_confusion_matrix.png)
- **ROC curve image** (rf_roc_curve.png)
- **Classification report** (classification_report.txt)
- **Feature importance CSV** (feature_importance.csv)
- **Insights summary text** (model_insights.txt)

Conclusion:
The assignment successfully demonstrates the **end-to-end pipeline for trade data analysis**, from preprocessing and feature engineering to predictive modeling and interpretation. The Random Forest model provides actionable insights into profitable trades and lays the foundation for more advanced financial prediction systems.