# **CSE Financial Analysis - Documentation**

## **Project Overview**

This project analyzes financial data from quarterly reports of two companies on the Colombo Stock Exchange (CSE): Dipped Products PLC (DIPD.N0000) and Richard Pieris Exports PLC (REXP.N0000). The work involved extracting data from PDF reports, creating structured datasets, building a dashboard, and forecasting future profits.

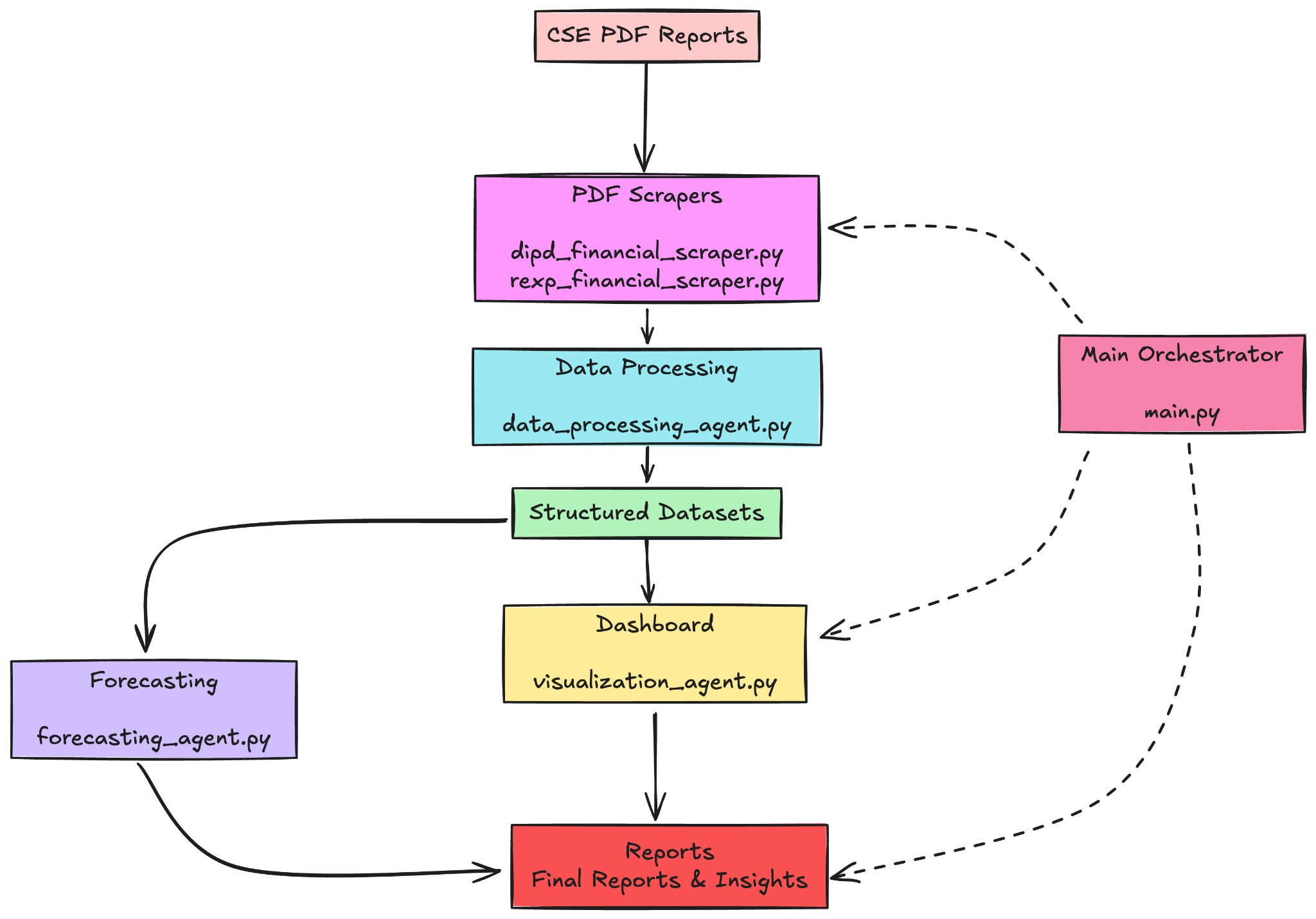
## **Project Structure**

The project uses a modular approach with the following components:

1. **Data Scraping**:  
   * dipd\_financial\_scraper.py - Gets DIPD financial data
   * rexp\_financial\_scraper.py - Gets REXP financial data
2. **Data Processing**:  
   * data\_processing\_agent.py - Cleans and structures the data
3. **Visualization**:  
   * visualization\_agent.py - Makes the dashboard with Dash and Plotly
4. **Forecasting**:  
   * forecasting\_agent.py - Runs the forecasting workflow
   * forecast\_profits.py - Actual forecasting code
5. **Main Program**:  
   * main.py - The main script that runs everything

### **Data Flow**

1. **Acquisition**: PDF reports → Text extraction → Raw financial metrics
2. **Processing**: Raw metrics → Data cleaning → Structured datasets
3. **Analysis**: Structured data → Visualization → Interactive dashboard
4. **Forecasting**: Historical data → Model training → Future projections



## **Implementation Approach**

### **Data Acquisition**

The project involved writing Python scripts to:

* Download PDF reports directly from the CSE website
* Extract text using PyPDF2
* Find financial numbers using regex patterns
* Map each report to the right quarter and year
* Handle inconsistencies like different formats

The most challenging part was dealing with different PDF formats - this required multiple regex patterns for each metric and fallback methods when direct extraction failed.

### **Data Processing**

The data processing phase included:

* Standardizing dates and company names
* Fixing numeric data that had formatting issues
* Calculating financial ratios (margins, growth rates)
* Combining separate company datasets for comparison
* Addressing the scale difference between companies (DIPD is about 1000x larger than REXP)

### **Dashboard Development**

The interactive dashboard built with Dash and Plotly allows users to:

* Select which company to view
* Choose date ranges
* Pick different financial metrics
* Compare both companies

The main charts include:

* Financial performance over time
* Margin analysis
* Growth rates
* Company comparisons

### **Forecasting**

The forecasting component implemented two models:

* Holt-Winters Exponential Smoothing
* ARIMA (AutoRegressive Integrated Moving Average)

For each financial metric, the process involved:

1. Testing both models on the historical data
2. Using MAPE (Mean Absolute Percentage Error) to select the better model
3. Generating 4-quarter forecasts with confidence intervals

## 

## **Assumptions**

### **Data Scraping Assumptions**

1. **Report Structure**: Although the PDF reports have different formats, they all follow basic accounting principles and have similar P&L structures that can be parsed.
2. **Data Completeness**: The necessary financial metrics (Revenue, COGS, etc.) are somewhere in the reports, even if they're called different names or in different sections.
3. **Regular Reporting**: Both companies publish quarterly reports following a regular schedule, providing at least 3 years of continuous data.
4. **Accounting Relationships**: Basic accounting relationships (Revenue - COGS = Gross Profit) can be used to check data and fill in missing values.
5. **Consistent Units**: Financial values in a single report use the same scale (millions, billions), though different reports might use different scales.

### **Data Processing Assumptions**

1. **Quarter Mapping**: Both companies follow similar fiscal year patterns, making it possible to align quarters for comparison.
2. **Consistent Definitions**: The meaning of financial metrics (like "Operating Income") stays consistent over time for each company.
3. **Manageable Missing Data**: There aren't too many missing values, and any gaps can be filled using interpolation or calculated from other metrics.
4. **Meaningful Growth Rates**: Despite seasonal variations, calculating quarter-over-quarter and year-over-year growth rates still provides useful information.

### **Forecasting Assumptions**

1. **Pattern Continuation**: The patterns seen in historical data will generally continue in the future.
2. **Seasonal Factors**: The quarterly data has seasonal patterns that repeat and can be modeled.
3. **Stability**: No major company restructuring or significant market changes will happen during the forecast period.
4. **Economic Conditions**: The overall economic environment will remain relatively stable during the forecast period.
5. **Sufficient Historical Data**: Three years of quarterly data (12 data points) is enough to identify meaningful patterns for forecasting.

## **Handling Data Inconsistencies**

Several strategies were used to deal with data inconsistencies in the reports:

### **1. Multiple Pattern Matching**

Since companies use different terms and formats for the same financial metrics, multiple regex patterns were created for each metric:

revenue\_patterns = [

r'Revenue from contracts\s+with customers[\s\n]\*[\d,.]+[\s\n]\*([\d,.]+)',

r'Revenue\s+([\d,.]+)',

r'Revenue\s\*(?:Rs\.|Rs)?[\s\n]\*([\d,.]+)',

# More patterns for different formats...

]

This helps catch the same metric even when it appears in different formats across reports.

### **2. Step-by-Step Extraction**

The extraction follows these steps when looking for financial data:

1. Try to find exact matches using specific patterns
2. If that fails, look for values in table structures
3. If tables don't work, search the general text
4. As a last resort, calculate missing values using accounting formulas

### **3. Using Accounting Relationships**

When a value can't be found directly, accounting relationships help fill the gaps:

# If Revenue and Gross Profit exist but COGS doesn't

if 'Cost of Goods Sold' not in metrics and 'Revenue' in metrics and 'Gross Profit' in metrics:

metrics['Cost of Goods Sold'] = metrics['Revenue'] - metrics['Gross Profit']

Other examples:

* Gross Profit = Revenue - COGS
* Operating Income = Gross Profit - Operating Expenses
* Net Margin = (Net Income / Revenue) \* 100

### **4. Data Validation Checks**

Several validation checks were implemented:

1. **Scale Check**: Flag values that are significantly different from previous quarters
2. **Logical Relationships**: Make sure Revenue ≥ Gross Profit ≥ Operating Income ≥ Net Income
3. **Growth Rate Check**: Flag unusually large quarter-to-quarter changes (>50%)
4. **Cross-Reference**: Compare values across different parts of the same report

## **Challenges Faced**

### **Data Extraction Problems**

1. **Inconsistent PDFs**: Every report had slightly different formats - some had tables, others had inline text
2. **Missing Data**: Some reports didn't clearly state all metrics
3. **Scale Issues**: Some values were in millions, others in thousands, not always clearly marked
4. **Text Extraction Quality**: PyPDF2 sometimes messed up table formatting

To solve these, I used multiple regex patterns for each metric and derived missing values using accounting relationships (e.g., Revenue - COGS = Gross Profit).

### **Processing Challenges**

1. **Different Scales**: DIPD operates at ~1000x the scale of REXP
2. **Missing Quarters**: Had to handle cases where data was missing
3. **Inconsistent Naming**: Companies used different terms for the same metrics

### **Forecasting Limitations**

The biggest issue with forecasting was the limited historical data (only 12 quarters), which isn't really enough for reliable time series forecasting. This probably explains the high MAPE values, especially for REXP's Net Income (285.34%!).

## **Key Findings**

### **Financial Performance Analysis**

**DIPD.N0000:**

* Larger company (about 1000x bigger than REXP)
* More stable performance with better margins
* Average gross margin: ~25.2%
* Average net margin: ~8.3%

**REXP.N0000:**

* Smaller operation with more volatile results
* Lower but still decent margins
* Average gross margin: ~23.8%
* Average net margin: ~7.1%

### **Forecast Results**

**DIPD.N0000 Forecast:**

| **Metric** | **Model** | **Q1 2025** | **Q2 2025** | **Q3 2025** | **Q4 2025** | **Growth** |
| --- | --- | --- | --- | --- | --- | --- |
| Revenue | Holt-Winters | 75.2B | 39.2B | 50.2B | 64.1B | 11.2% |
| Net Income | Holt-Winters | 3.9B | -0.4B | 0.9B | 1.8B | -77.0% |
| Gross Profit | Holt-Winters | 17.4B | 3.8B | 8.4B | 12.8B | 35.3% |
| Operating Income | Holt-Winters | 5.7B | -0.3B | 1.6B | 2.7B | -169.1% |

**REXP.N0000 Forecast:**

| **Metric** | **Model** | **Q1 2025** | **Q2 2025** | **Q3 2025** | **Q4 2025** | **Growth** |
| --- | --- | --- | --- | --- | --- | --- |
| Revenue | Holt-Winters | 1.5B | 2.0B | 1.8B | 1.2B | -2.1% |
| Net Income | ARIMA | 67.9M | 70.9M | 101.9M | 72.3M | -5.6% |
| Gross Profit | ARIMA | 449.5M | 396.0M | 441.0M | 405.3M | 3.8% |
| Operating Income | ARIMA | 148.1M | 130.8M | 139.3M | 139.3M | -0.8% |

The forecast results are pretty wild, especially for DIPD showing negative income in Q2 2025. The high MAPE values (Holt-Winters had 20-50% for DIPD, and ARIMA had up to 285% for REXP) show that these forecasts aren't very reliable.

## **Implications**

### **For Investors**

* DIPD has been more stable historically but forecasts suggest potential volatility
* REXP shows limited growth potential but potentially less dramatic swings
* The forecasts have high uncertainty, so should be interpreted cautiously

### **For the Companies**

* DIPD should focus on maintaining its margin advantage
* REXP might benefit from stabilizing performance and improving margins
* Both show seasonal patterns that could be addressed through diversification

## **Analysis Limitations**

1. **Limited Data**: Only 3 years (12 quarters) of data - insufficient for reliable forecasting
2. **Data Quality Issues**: Some values needed to be derived when direct extraction wasn't possible
3. **Forecasting Reliability**: The high MAPE values indicate low forecast reliability
4. **Missing Context**: The analysis is purely quantitative and doesn't incorporate management changes, strategy shifts, or industry trends

## **Conclusion**

The analysis reveals that DIPD is the larger, more stable company with better margins, while REXP is smaller with more volatile performance. The forecasts suggest potential volatility ahead, especially for DIPD, but the high error rates indicate these projections should be viewed as possible scenarios rather than predictions.

Despite the limitations, this project demonstrates a systematic approach to extracting financial data from PDFs, processing it into structured datasets, visualizing through an interactive dashboard, and forecasting future performance. This methodology could be applied to other companies or extended with more sophisticated analysis techniques.