**GROUP WORK PROJECT 1** MScFE 600: FINANCIAL DATA

**GROUP NUMBER:** 8050

| FULL LEGAL NAME | LOCATION (COUNTRY) | EMAIL ADDRESS | MARK X FOR ANY NON-CONTRIBUTING MEMBER |
|---|---|---|---|
| **DANIEL SIMWABA** | ZAMBIA | sulayman.m.f@gmail.com | |
| **MBOUZOU FOMENA PALMAS** | CAMEROON | mbouzoufomena@gmail.com | |
| **MUHAMMAD SULAYMAN** | NIGERIA | danielsimwaba@gmail.com | |

| **Statement of integrity:** By typing the names of all group members in the text boxes below, you confirm that the assignment submitted is original work produced by the group (excluding any non-contributing members identified with an “X” above). | |
|---|---|
| **Team member 1** | **DANIEL SIMWABA** |
| **Team member 2** | **MBOUZOU FOMENA PALMAS** |
| **Team member 3** | **MUHAMMAD SULAYMAN** |

Use the box below to explain any attempts to reach out to a non-contributing member. Type (N/A) if all members contributed.
**Note:** You may be required to provide proof of your outreach to non-contributing members upon request.

N/A

# Question 1: Data Quality

## a. Examples of Poor Quality Structured Data

***Part 1 : Mobile Money transactions :***

| Transaction ID | Sender Name | Transaction Amount | Date |
|---|---|---|---|
| 1001 | Daniel SIMWABA | 350 | 2025-01-15 |
| 100& | Palmas MBOUZOU FOMENA | | 2025-01-15 |
| 1003 | Daniel SIMWABA | 350 | 2025-01-15 |
| 1004 | | Muhammad FOLAHANMI SULAYMAN | 15 January, 2025 |

***Part 2 : Customer Address Dataset for an eCommerce Business :***

| Customer ID | Name | Email | Address | Phone Number | Date of Birth | Gender |
|---|---|---|---|---|---|---|
| 001 | Jon Dolo | johndoe@gmail.com | 123 Main St | 555-123-4567 | 1985-06-15 | M |
| 002 | Janet Samuli | | | 555-789-1011 | 1992-03-28 | F |
| 003 | Alice Banda | aliceb@abc.com | 456 Elm St, Apt 5 | | Not provided | Female |

| | | | | | | |
|---|---|---|---|---|---|---|
| 004 | Bob Kata | | Unknown | 555-999-8888 | 1988/08/09 | M |
| 005 | Carlos Lumba | carloslum@yahoo | 789 Oak Ave Apt 2B | 555-555-5555 | 15/07/1990 | |

***<u>Part 3 : Financial Market Dataset of Stocks and cryptocurrency</u>***

| Date | Asset | Opening Price | Closing Price |
|---|---|---|---|
| ***Bitcoin Prices*** | | | |
| 2025-01-17<br>Friday | BTC/USD<br>(Crypto) | 100,000 | 106,489 |
| 2025-01-18<br>Saturday | BTC/USD<br>(Crypto) | 98,100 | 99,510 |
| 2025-01-19<br>Sunday | BTC/USD<br>(Crypto) | 99,100 | 99,860 |
| ***Apple Prices*** | | | |
| 2025-01-17<br>Friday | AAPL (Stock) | 225 | 229 |
| 2025-01-18 | AAPL (Stock) | **N/A** | **N/A** |

| | | | |
|---|---|---|---|
| Saturday | | | |
| 2025-01-19<br>Sunday | AAPL (Stock) | **N/A** | **N/A** |

**The dropna() function in Python (typically used with pandas DataFrames or Series) is used to remove missing values (NaN) from a dataset. It allows for cleaning the dataset by removing rows or columns with missing data.**

## b. Data Quality Issues

***Part 1 : Mobile Money transactions :***

Poor quality structured data can be recognized by its failure to meet key data quality dimensions:

1. **Accuracy:** Missing values in the "Transaction Amount" column indicate a lack of completeness, making the dataset unreliable for analysis.
2. **Uniqueness:** Duplicate entries, such as repeated Transaction IDs or identical records, compromise data integrity.
3. **Consistency:** If the data types or formats in a field, such as dates, vary (e.g., "January 15, 2025" vs. "2025-01-15"), it signals inconsistency.

***Part 2 : Customer Address Dataset for an eCommerce Business :***

1. **Missing Data:** Jane Smith's email and address are missing, and Alice Banda's phone number is absent.
2. **Inconsistent Formatting:** Date of Birth has different formats (e.g., 1988/08/09 vs. 15/07/1990).
3. **Incomplete Entries:** The "Gender" field for Carlos Lumba is blank, and Alice Banda's "Date of Birth" is labeled "Not provided."
4. **Invalid Data:** Carlos Lumba's email address (carloslum@yahoo) is incomplete and lacks a domain.
5. **Ambiguous Data:** Bob Kata' address is "Unknown," which provides no actionable information.

***Part 3 : Financial Market Dataset of Stocks and cryptocurrency***

1. **Accuracy**: The missing data for stock prices on weekends creates gaps that render comparative analysis between crypto and stock markets unreliable.
2. **Consistency**: The inclusion of cryptocurrency weekend data but the absence of corresponding stock data violates consistency, making cross-market analyses challenging.

3. **Completeness**: Missing fields for critical data points, such as stock prices on specific dates, reduce the dataset's overall integrity and usability.

## C. Example of Poor Quality Unstructured Data

***Part 2 : Customer Address Dataset for an eCommerce Business :***

An example of poor quality unstructured data could be a collection of customer reviews via amazon where the text contain unstructured information, excessive abbreviations, emojis, or incomplete sentences. For instance:

- "Gr8 prdct bt svce wz 😡!!!"
- "Delivery took forever will not... "
- "Happy :)"

***Part 3 : Financial Market Dataset of Stocks and cryptocurrency***

In the context of financial markets, poor quality unstructured data could be seen in social media posts or cryptocurrency forum discussions where the information may contain misinformation, excessive jargon, or unverified claims about asset performance. For example:

- "BTC 🚀💰 gonna hit 500k by weekend! #HODL"
- "The stock market is collapsing. Everything down!"
- "XRP outperforming ETH because whales buying!!!"

## D. Recognizing Poor Quality in Unstructured Data

Unstructured data can be more difficult to assess because it lacks a predefined format or schema, but its poor quality can be recognized as follows:

1. Relevance: Data must be applicable and helpful for the intended purpose. According to Wang and Strong (1996), relevance refers to the extent to which data is applicable and helpful for the task at hand.

2. Completeness: This dimension assesses whether all required data is present. Incomplete data can lead to inaccurate analyses and decisions. Wang and Strong (1996) define completeness as the extent to which data is of sufficient breadth, depth, and scope for the task at hand.

3. Readability: Data should be easily understood by its intended audience. This includes proper grammar, spelling, and the avoidance of excessive jargon. While not always listed as a separate dimension, readability impacts the interpretability and usability of data.

4. Accuracy: Accuracy refers to the closeness of data values to the true values. High accuracy ensures that data correctly represents the real-world constructs it is intended to model. Wang and Strong (1996) describe accuracy as the extent to which data is correct, reliable, and certified free of error.

5. Consistency: Consistency ensures that data remains uniform across different datasets and systems. Inconsistencies can arise when data is recorded differently in various places, leading to confusion and errors. Wang and Strong (1996) define consistency as the extent to which data is presented in the same format and compatible with previous data.

# Question 2: NS MODEL AND CUBIC SPLINE MODEL FITTING FOR NIGERIAN GOVERNMENT BOND

## 2-a) Introduction

This report presents model fitting of yield curve for Nigerian government bond, we apply Nelson-Siegel (NS) model and cubic spline interpolation to fit the yield curve. In the analysis, we use data obtained from the Debt Management Office (DMO) of Nigeria.

## Analysis and Results

## 2-b) Data Source

The data for this analysis was sourced from the Debt Management Office (DMO) of Nigeria, which provide the yield for different maturities of 2017 Nigerian Government Bond. The data use are as follows:

| Maturity (Years) | Yield (%) |
|---|---|

| | |
|---|---|
| 0.21 | 14.05 |
| 0.46 | 18.20 |
| 0.55 | 19.12 |
| 1.30 | 19.79 |
| 2.38 | 16.13 |
| 2.70 | 16.35 |
| 3.01 | 16.56 |
| 4.42 | 16.56 |
| 4.96 | 16.23 |
| 7.09 | 16.20 |
| 8.95 | 16.76 |
| 11.80 | 16.50 |

| | |
|---|---|
| 12.28 | 16.46 |
| 12.78 | 16.42 |
| 13.45 | 16.36 |
| 17.43 | 16.00 |
| 19.10 | 16.84 |

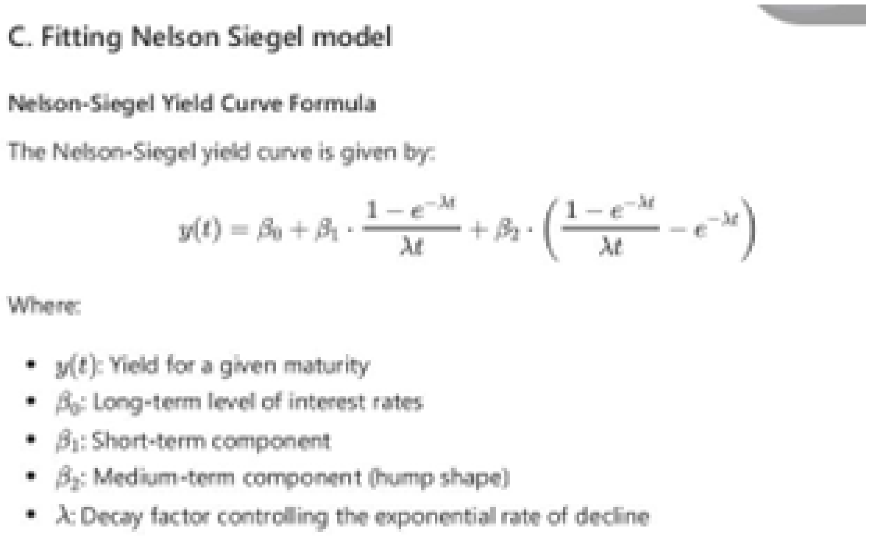
### Yield Curve Plot

The yield curve was plotted using the maturity and yield data. It shows the relationship between bond maturities and their corresponding yields.

## 2-c) Nelson-Siegel Model

**Formula:**

The Nelson-Siegel yield curve **is** defined as:



### C. Fitting Nelson Siegel model

#### Nelson-Siegel Yield Curve Formula

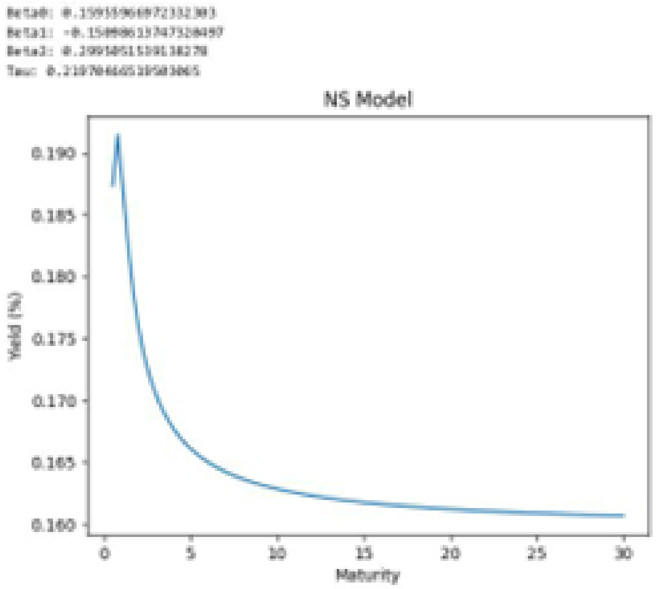
The Nelson-Siegel yield curve is given by:

$$y(t) = \beta_0 + \beta_1 \cdot \frac{1 - e^{-\lambda t}}{\lambda t} + \beta_2 \cdot \left( \frac{1 - e^{-\lambda t}}{\lambda t} - e^{-\lambda t} \right)$$

Where:

- $y(t)$: Yield for a given maturity
- $\beta_0$: Long-term level of interest rates
- $\beta_1$: Short-term component
- $\beta_2$: Medium-term component (hump shape)
- $\lambda$: Decay factor controlling the exponential rate of decline

A smooth yield curve was generated using the estimated parameters below:

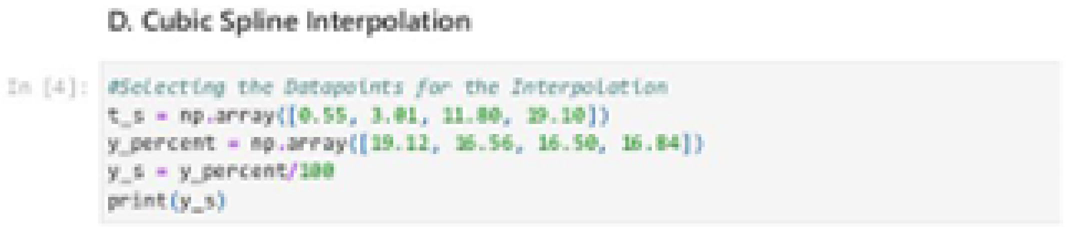


```
Beta0: 0.15935960972332303
Beta1: -0.15000613747328497
Beta2: 0.19950515391382278
Tau: 0.21070465319583065
```

## 2-d) Cubic Spline Interpolation

### Spline Construction

The cubic spline was constructed using four key points:

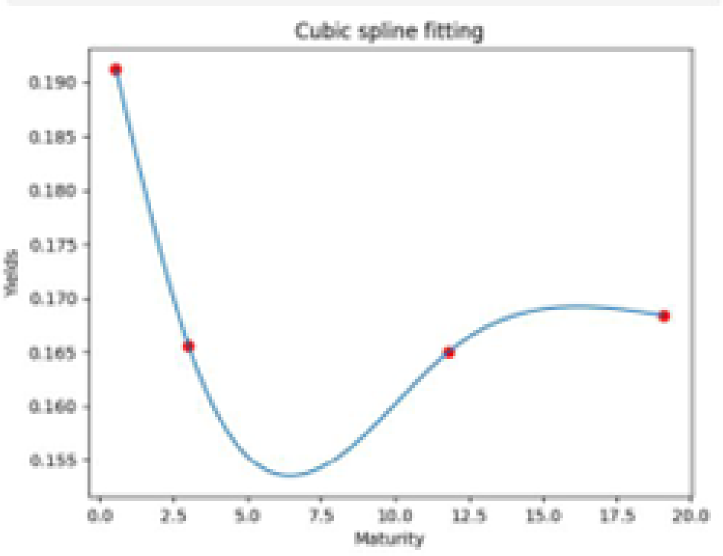


D. Cubic Spline Interpolation

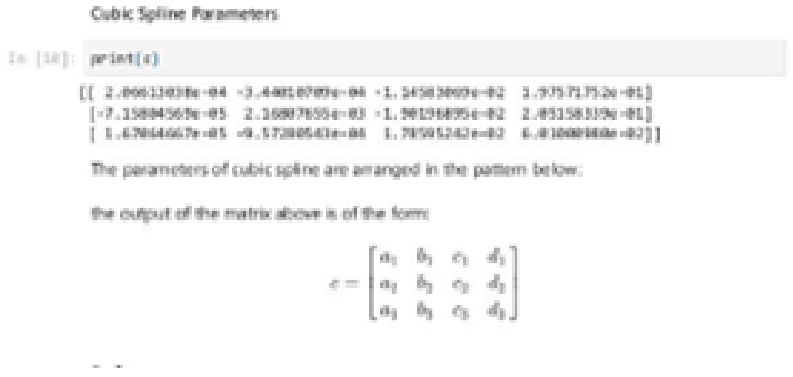
```
In [4]: #Selecting the Datapoints for the Interpolation
        t_s = np.array([0.55, 3.01, 11.80, 19.10])
        y_percent = np.array([19.12, 16.56, 16.50, 16.84])
        y_s = y_percent/100
        print(y_s)
```

The cubic spline equations were derived for each segment of the curve and solved using matrix methods

**Results**



**Parameters**



Cubic Spline Parameters

```
In [10]: print(c)

[[ 2.06612038e-04 -3.44018709e-04 -1.14583069e-02  1.97571752e-01]
 [-7.15804569e-05  2.16807655e-03 -1.90196895e-02  2.05158335e-01]
 [ 1.67044667e-05 -9.57300543e-04  1.78505242e-02  6.01000088e-02]]
```

The parameters of cubic spline are arranged in the pattern below:

the output of the matrix above is of the form:

$$c = \begin{bmatrix} a_1 & b_1 & c_1 & d_1 \\ a_2 & b_2 & c_2 & d_2 \\ a_3 & b_3 & c_3 & d_3 \end{bmatrix}$$

## 2-e) Comparison of Models

**In terms of fit:**

1. 1. NS model is very smooth and is more ideal for capturing general trends. (Nelson & Siegel 473 - 489)
2. 2. Cubic spline model offers high flexibility which gives it the advantage of fixing irregularies in data points and the disadvantage of overfitting ( McCulloch 811 - 830)

**Interpretation**

1. The parameters of NS fitting has economic meaning capturing long term, short term and medium term yield curve component (Nelson and Siegel 473 - 489)
2. The parameters of cubic spline model are purely mathematical and don't have direct economic interpretation(Adams and van Deventer)

## Conclusion

The analysis of Nigerian government securities revealed the following:

1. **Nelson-Siegel Model:** Best suited for capturing overall trends with interpretable parameters.
2. **Cubic Spline Model:** Provides a highly accurate fit but lacks interpretability and can overfit.

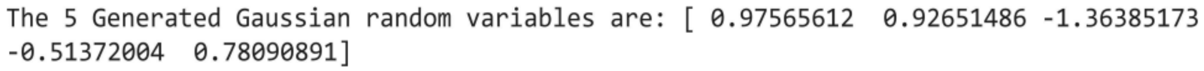
## Question 3. Exploiting Correlation

Financial Data is meant not only to process data but to understand how meaningful factors can be used to summarize or represent the data. Let’s understand the role that correlation and principal components play.

### PART 1: WORKING WITH RANDOM DATA

### a. Generate 5 uncorrelated Gaussian random variables that simulate yield:

Consult code for this question, below is the screenshot of results obtained.



```
The 5 Generated Gaussian random variables are: [ 0.97565612  0.92651486 -1.36385173
-0.51372004  0.78090891]
```

### b. Run a Principal Components using covariance matrix:

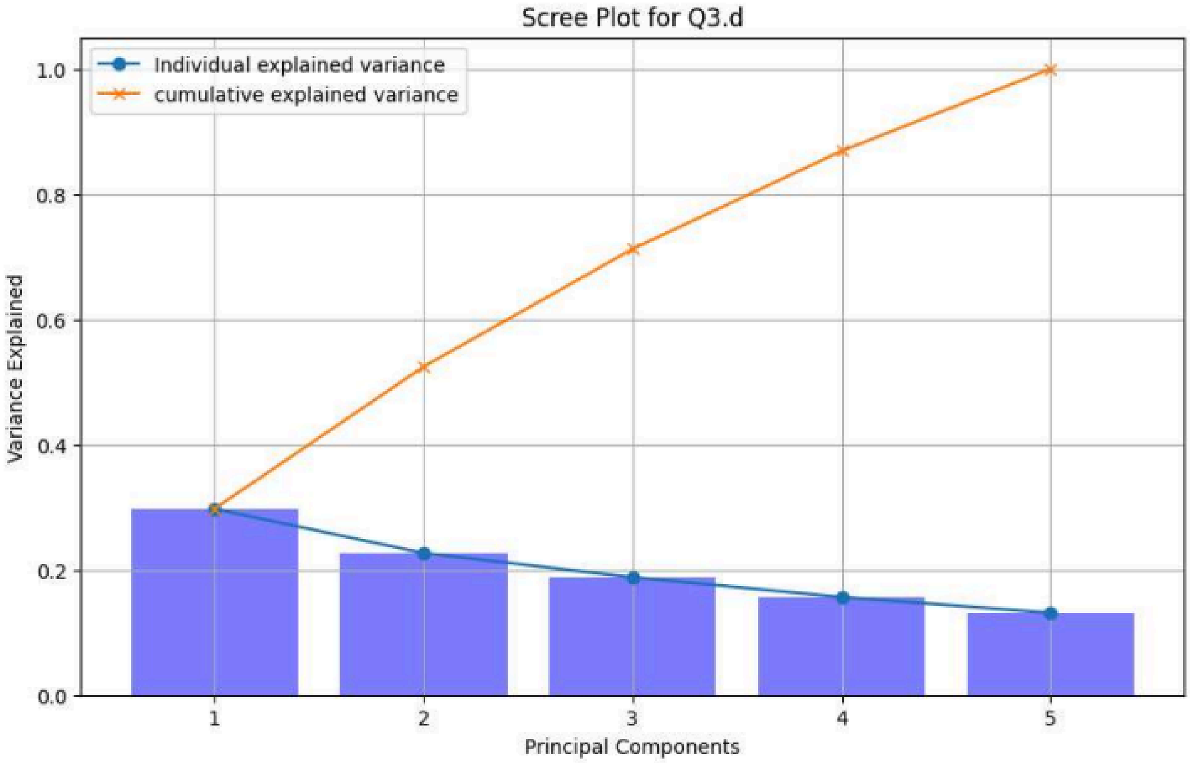
Consult code for this question

## c. Explanation of how the variances of each component compare with each other.

Following the results obtained in question 3b) above, the three principal components explained the variance in the data as follows: Component 1: 29.74% Component 2: 22.67% Component 3: 15.65%. Component 1 has the highest contribution to the variance, explaining 29.74% of the overall variance, Component 2 follows with a contribution of 22.67% to the variance, while Component 3 accounts for approximately 15.65% of the total variance. The remaining components contribute smaller portions, indicating that they capture less significant variations in the data. Overall, the first three components collectively explain almost 68% of the total variance.

## d. Production of a screeplot of the variance explained for each component:

Consult code for this question, results are shown below



## PART 2: WORKING WITH REAL DATA:

**e. Collect the daily closing yields for 5 government securities, say over 6 months:** Consult code for this question.

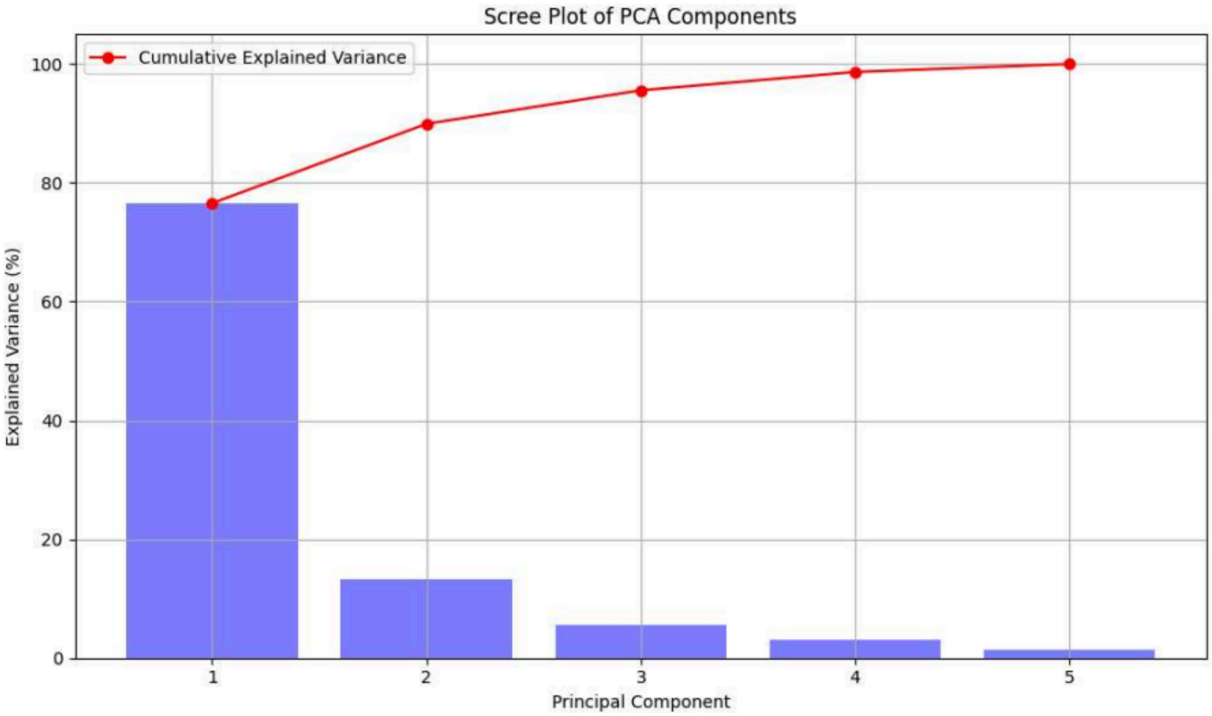
**f. Computing daily yield changes:** Consult code for this question.

**g. Rerunning the Principal Components using covariance matrix:** Consult code for this question.

### h. Explaining how the variances of each component compare?

Explained variance ratio by component are as follows: PC1: explains 76.57% of the total variance, PC2: explains 13.35% of the total variance, PC3: explains 5.65% of the total variance, PC4: explains 3.09% of the total variance and PC5: explains 1.34% of the total variance. PC1 captures the majority of the variance i.e. 76.57%, followed by PC2 and PC3. PC1 and PC2 alone captures almost 90% of the total variance

**i. Producing a screeplot of the variance explained for each component:** Consult code for this question, results are shown below



## j. Comparing the screeplot from the uncorrelated data with the screeplot from the government data?

Comparing the screeplot from the uncorrelated data with the screeplot from the government data reveals notable differences in how the variance is distributed across the principal components. In the uncorrelated data, the variance explained by the first three principal components (29.74%, 22.67%, and 15.65%, respectively) collectively accounts for around 68% of the total variance. This indicates a more gradual decline in the amount of variance captured by each subsequent component. Conversely, in the government data, the first three components (76.57%, 13.35%, and 5.65%, respectively) explain approximately 95.57% of the total variance, with a sharp drop-off after the first principal component. This steep decline in the government data suggests that the majority of the variance is concentrated in the first component, while the remaining components capture relatively small portions of the data's variability.

The practical and useful application of PCA in real-world data analysis is evident in the comparison of these two screeplots. With the government data, PCA effectively reduces the dimensionality of the dataset by identifying that the first component alone captures the majority of the variance. This allows for a more simplified and efficient representation of the data without losing significant information, which is particularly valuable in the field of finance where datasets are large and complex.

# Question 4: Empirical Analysis of ETFs: Real Estate Select Sector SPDR Fund (XLRE)

## Introduction

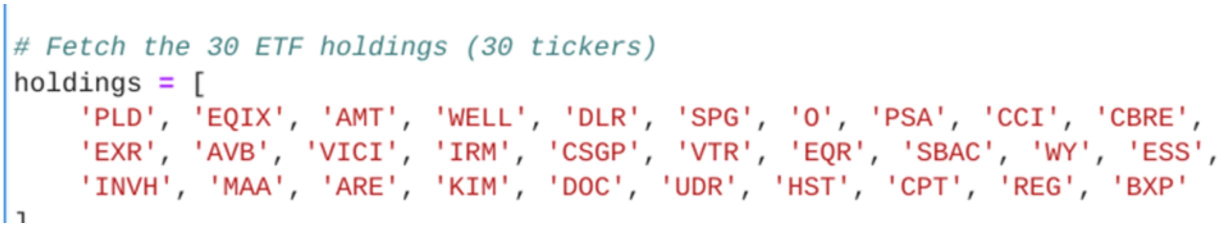
The portion of this report presents an empirical analysis of the Real Estate Select Sector SPDR Fund (XLRE), focusing on the top 30 holdings within the ETF. The analysis spans six months (July 1, 2024, to December 31, 2024), with approximately 120 daily data points. Key techniques applied include the computation of daily returns, covariance matrix, Principal Component Analysis (PCA), and Singular Value Decomposition (SVD). The findings highlight critical insights into the ETF's structure, diversification, and underlying drivers of performance.

## Analysis and Results

## a) Identifying the Top 30 Holdings

XLRE’s top holdings include leading real estate companies across various sectors:

- **Prologis, Inc. (PLD):** Specializes in logistics real estate (SPDR ETFs, 2024).
- **Equinix, Inc. (EQIX):** Operates global data centers (SPDR ETFs, 2024).
- **American Tower Corporation (AMT):** Focuses on wireless communication infrastructure (SPDR ETFs, 2024).
- **Simon Property Group, Inc. (SPG):** A major player in retail real estate (SPDR ETFs, 2024).
- **Digital Realty Trust, Inc. (DLR):** Invests in digital infrastructure (SPDR ETFs, 2024).



```
# Fetch the 30 ETF holdings (30 tickers)
holdings = [
    'PLD', 'EQIX', 'AMT', 'WELL', 'DLR', 'SPG', 'O', 'PSA', 'CCI', 'CBRE',
    'EXR', 'AVB', 'VICI', 'IRM', 'CSGP', 'VTR', 'EQR', 'SBAC', 'WY', 'ESS',
    'INVH', 'MAA', 'ARE', 'KIM', 'DOC', 'UDR', 'HST', 'CPT', 'REG', 'BXP'
]
```

These companies collectively account for a significant portion of XLRE’s performance, representing trends in logistics, retail, and digital transformation within the real estate sector.

## b) Get at least 6 months of data (~ 120 data points): Consult Code for this question

## c) Daily Returns

Daily returns were computed using the formula: Where is the return on day , is the closing price on day , and is the closing price on the previous day.

Key observations:

- **Volatility:** Smaller holdings exhibited higher daily fluctuations compared to larger, more stable firms like PLD and EQIX.
- **Performance trends:** Overall, daily returns revealed the ETF’s sensitivity to macroeconomic factors, such as interest rate changes. Overall, daily returns revealed the ETF’s sensitivity to macroeconomic factors, such as interest rate changes (Yahoo Finance, 2024).

## d) Covariance Matrix

The covariance matrix revealed relationships between the ETF’s holdings: The covariance matrix revealed relationships between the ETF’s holdings (WQU , 2024):

- **High correlations:** Most assets showed strong positive correlations, driven by sector-wide trends such as economic cycles and REIT demand.
- **Diversification:** Lower correlations between specific holdings (e.g., SPG and WELL) suggested opportunities for diversification within the ETF.

## e) Principal Component Analysis (PCA)

PCA reduced the dimensionality of the dataset, highlighting key drivers of variance: PCA reduced the dimensionality of the dataset, highlighting key drivers of variance (Scikit-learn, 2024):

- **Eigenvalues:** The first component explained 48.6% of the variance, with the top three components accounting for over 72%. This demonstrates that a few dominant factors shape the ETF’s overall performance.

- **Eigenvectors:** Holdings like PLD, AMT, and EQIX had high weights in the first component, reflecting their critical role in sector performance.

- **Explained Variance Ratio:** Most variance is concentrated in the top components, reinforcing the idea that XLRE’s behavior is governed by overarching economic and sector-specific trends.

## f) Singular Value Decomposition (SVD)

SVD provided additional insights into the ETF’s structure: SVD provided additional insights into the ETF’s structure (NumPy, 2024):

- **Singular Values:** A sharp decline in singular values after the first few confirmed the dominance of a small number of patterns, consistent with PCA results.
- **Matrix Decomposition:** The , , and matrices highlighted asset-specific effects, the importance of principal components, and time-series behaviors.

## Comparative Insights: PCA vs. SVD

- **PCA:** Focuses on variance and provides interpretable results through eigenvalues and eigenvectors. Ideal for identifying dominant factors in performance.
- **SVD:** Offers a detailed matrix decomposition, suitable for analyzing structural patterns in the data. Both methods are aligned in emphasizing the role of a few key components in XLRE's behavior.

## Conclusion

The empirical analysis of XLRE revealed:

1. **Sector Concentration:** XLRE's performance is driven by logistics, retail, and data center real estate.
2. **Key Drivers:** A few dominant factors, as identified by PCA and SVD, account for most of the ETF's variance.
3. **Diversification:** While highly correlated, opportunities for diversification exist among specific holdings.

This study underscores the importance of quantitative techniques in ETF analysis, offering valuable insights for portfolio construction, risk management, and investment strategies. The combination of PCA and SVD provided a robust understanding of XLRE's structure and performance dynamics.

# References

1. Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems, 12(4), 5–33. https://doi.org/10.1080/07421222.1996.11518099
2. Yahoo Finance: Historical data for XLRE and its holdings. Available at: https://finance.yahoo.com
3. SPDR ETFs Official Website: Information on XLRE holdings and sector breakdown. Available at: https://www.ssga.com
4. Scikit-learn Documentation: Principal Component Analysis (PCA). Available at: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
5. NumPy Documentation: Singular Value Decomposition (SVD). Available at: https://numpy.org/doc/stable/reference/generated/numpy.linalg.svd.html
6. WQU Masters Module: FD-MScFE600-M5-L1 Course Materials and Python implementation for ETF analysis.
7. Debt Management Office, "FGN bonds update", 13 february, 2017, https://www.dmo.gov.ng/fgn-bonds/fgn-bond-updates/1935-fgn-bonds-weekly-trading-highlights-report-as-at-february-06-to-february-10-2017/file.
8. Nelson, Charles R., and Andrew F. Siegel. "Parsimonious Modeling of Yield Curves." The Journal of Business, vol. 60, no. 4, 1987, pp. 473-489.
9. McCulloch, John H. "The Tax-Adjusted Yield Curve." The Journal of Finance, vol. 30, no. 3, 1975, pp. 811-830.
10. Adams, J., and R. van Deventer. "Maximum Smoothness Forward Rates and Related Yields." Basic Building Blocks of Yield Curve Smoothing, SAS Risk Data and Analytics, 2012, pp. 1-15.