GROUP WORK PROJECT # 1 MScFE 600: FINANCIAL DATA

GROUP NUMBER: 11446

FULL LEGAL NAME	LOCATION (COUNTRY)	EMAIL ADDRESS	MARK X FOR ANY NON-CONTRIBUTING MEMBER
Harris Tekenah	Nigeria	harrisedwardtekenah@gmail.com	
Nyasha Katonha	Zimbabwe	katonhanyasha@gmail.com	
Ka Leung Godfrey Cheung	Hong Kong	Godfreycheungjob1234@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	Harris Tekenah
Team member 2	Nyasha Katonha
Team member 3	Ka Leung Godfrey Cheung

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.

GROUP WORK PROJECT # 1 MScFE 600: FINANCIAL DATA

Group Number: 11446

Task 1

a.

	person_id	name	email	phone	linkedin
0	1	Database Administrator	NaN	NaN	NaN
1	2	Database Administrator	NaN	NaN	NaN
2	3	Oracle Database Administrator	NaN	NaN	NaN
3	4	Amazon Redshift Administrator and ETL Develope...	NaN	NaN	NaN
4...	5	Scrum Master Scrum Master Scrum Master	NaN	NaN	NaN

(Source: Kaggle ,

https://www.kaggle.com/datasets/suriyaganesh/resume-dataset-structured?select=01_people.csv)

b. When we first examined the dataset, we observed that many of the key fields were blank: columns such as email (97%missing), phone (97%missing), and LinkedIn (85%missing), which immediately made the table feel incomplete. As we looked closer, we noticed the name column is inconsistent, often containing job titles instead of individuals' names. The name column also contain inaccurate entry like one row listed "Scrum Master" three times. These issues were discussed by ("Guide to Structured vs Unstructured Data") and (Freeman) when discussing poor data quality: missing values, inconsistent records, and information that can't really be trusted.

c.

	v1	v2	Unnamed: 2	Unnamed: 3	Unnamed: 4
0	ham	Go until jurong point, crazy.. Available only ...	NaN	NaN	NaN
1	ham	Ok lar... Joking wif u oni...	NaN	NaN	NaN
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	NaN	NaN	NaN
3	ham	U dun say so early hor... U c already then say...	NaN	NaN	NaN
4...	ham	Nah I don't think he goes to usf, he lives aro...	NaN	NaN	NaN

(Source: Kaggle ,

https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset)

d. .The first two columns (v1 and v2) contain the useful data with labels and messages, while the other columns (Unnamed: 2–4) are mostly empty (99% missing) and lack a consistent format, making them difficult to use. Multiple rows are duplicated, such as lines 95 and 899, which could lead to misleading and

inaccurate conclusions due to this noise. Upon closer inspection, some of the text messages appear unfinished, suggesting that the data may be potentially incomplete. Unstructured data quality is less mature and demands more research attention compared to structured data, which has relatively clearer processing frameworks (Mirzaie, Behkamal, and Paydar 10). Jonker and Gomstyn also highlighted that uneven data, like this dataset, can easily lead to errors, illustrating the another challenges of unstructured information (Jonker and Gomstyn)

Task 2

Yield Curve Modeling Task

a. We picked the government securities from Hong Kong. The data source is from investing.com.
b. We picked a range of government bond/note of different tenor. 8 tenor are selected, they are : 6month, 1Year, 3Year , 5Year, 7Year , 10Year, 15year and 20 year.
c. We fit the data using Nelson-Siegel(NS) model.
d. We have also fit the data using Cubic-Spline(CS) model.
e. Comparing the fit of NS and CS, we can observe that NS does not perfectly fit all observed data, only a single smooth curve that passes through most points. In contrast, CS demonstrates the ability to fit observed data closely, but its fitted line may exhibit a "kink."

In terms of interpretation:

a) NS offers better interpretability because its fitted line is smooth, potentially yielding more explainable results. Additionally, NS can handle extrapolation effectively. Since CS relies on interpolation, it may not be well-defined for out-of-sample calculations. However, as a parametric method, NS can provide results for out-of-sample inputs.
b) NS also provides intuitive meaning for its model parameters. According to M1L3, □ represents the decay rate, which ranges between 0 and 1. The parameter □0 describes the level of the yield curve, □1 describes its slope, and □2 describes its shape.

f. For Nelson Siegel Curve, the fitted values are :

Beta0 3.7992
Beta1 -0.5993
Beta2 -3.450
Tau 1.730

For Cubic spline the model parameter is:

	(0.5 to 1.0)	(1.0 to 3.0)	(3.0 to 5.0)	(5.0 to 7.0)	(7.0 to 10.0)	(10.0 to 15.0)	(15.0 to 20.0)
Cubic(x^3)	0.1938	-0.0574	0.0206	-0.0189	0.0045	0.0007	-0.0005

Quadratic(x^2)	0.0000	0.2907	-0.0540	0.0699	-0.0434	-0.0028	0.0074
Linear(x^1)	-0.5504	-0.4051	0.0683	0.1002	0.1532	0.0147	0.0378
Constant(x^0)	2.8810	2.6300	2.5230	2.6090	2.9380	3.1290	3.2180

g. Although Nelson-Siegel is smoothing the yield curve, this is not necessarily considered as unethical. It is because:

a) The smoothing is not intended to mislead the true. The smoothing is by construction of the model design. It is also intend to produce a interpretable representation of the yield curve for analysis.
b) In Econometrics, you will smooth time series to filter out noise. Nelson-Siegel can filter out the noise to prevent over fitting. So the intention is not unethical.

Group Number: 11446

Task 3

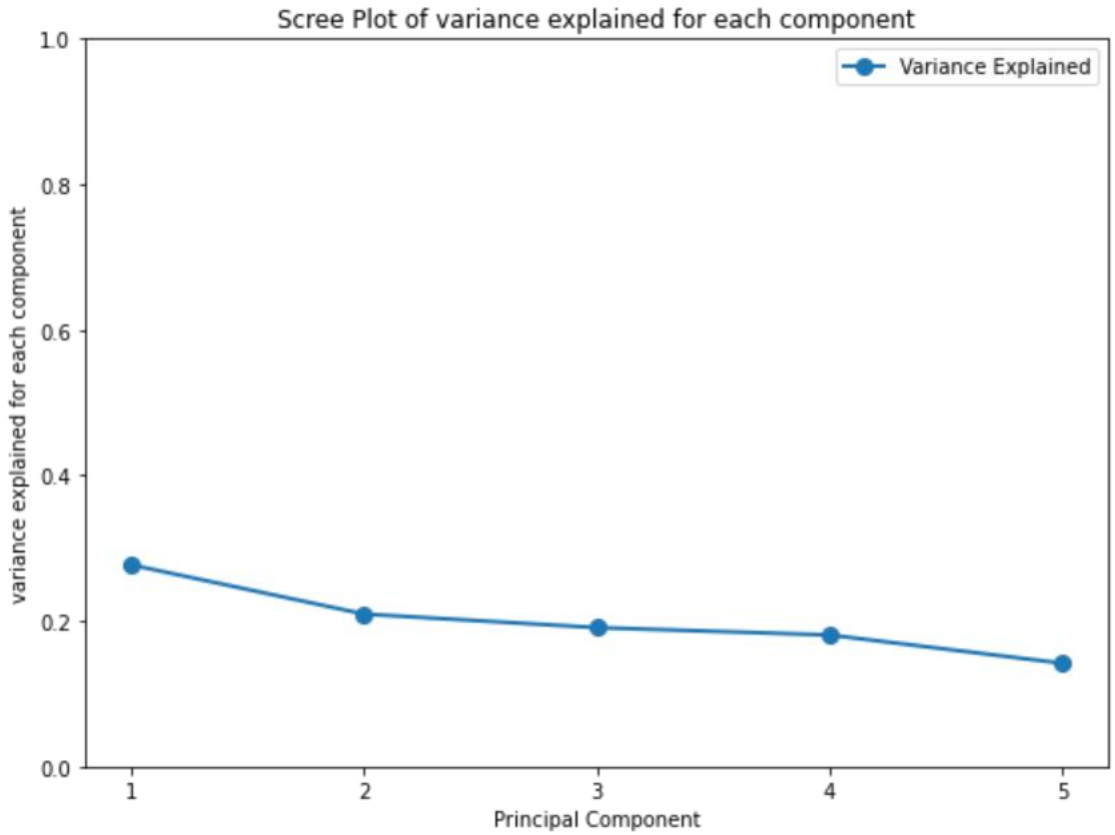
a. We have generated 5 uncorrelated Gaussian random variables, with size equal 100, assuming mean = 0, sd = 0.05 and saved the result into 1 dataFrame.

b. Result of the PCA on the uncorrelated Data:

	Eigenvalues	Explained proportion
1	1.384668	27.69%
2	1.047263	20.95%
3	0.954461	19.09%
4	0.903765	18.08%
5	0.709842	14.20%

c. The eigenvalues from PCA are in descending order. The leading principal components can explain higher portions of the variance of the dataset than the rest of the principal components. Component 1 can explain 27.69%, Component 2 can explain 20.95% and Component 3 can explain 19.09%

d. Scree plot:



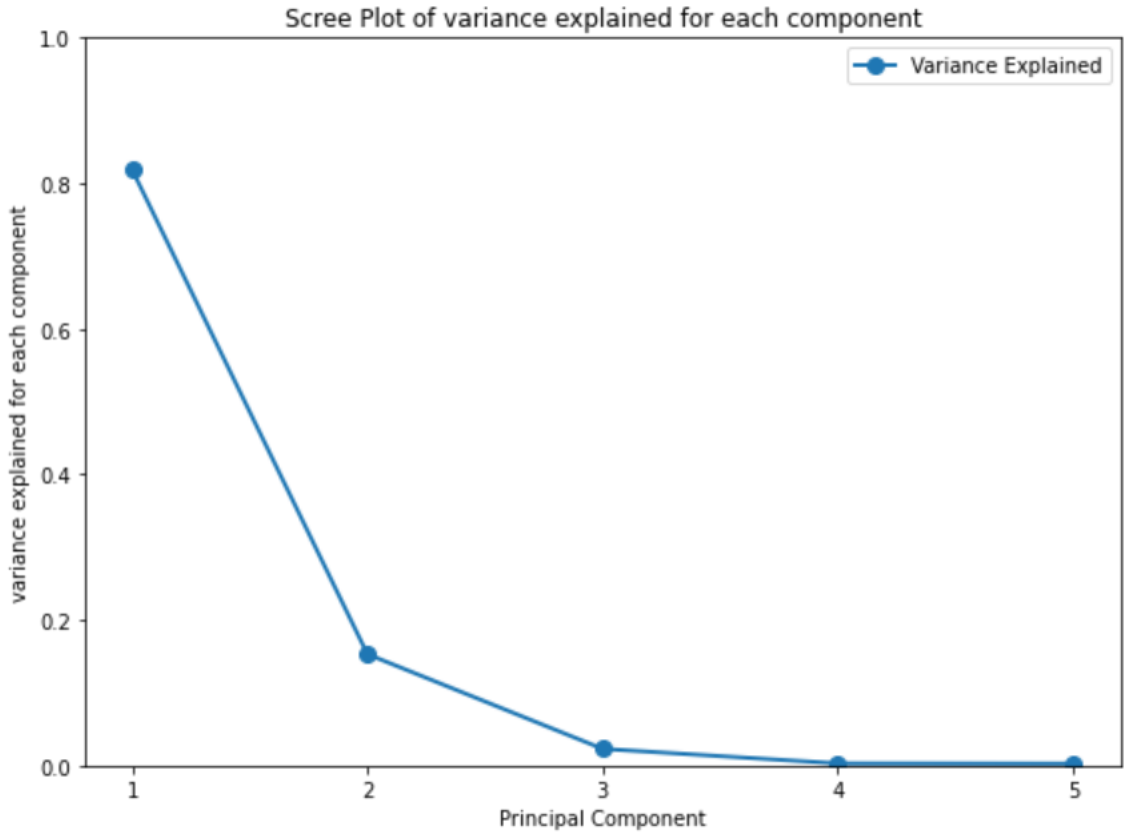
e. We have chosen the US Gov Note/Bond for the task. With different tenor: 1Year, 5Year, 10Year, 20 year and 30 year.

g. Result of the PCA on the government securities data:

	Eigenvalues	Explained proportion
1	1.384668	27.69%
2	1.047263	20.95%
3	0.954461	19.09%
4	0.903765	18.08%
5	0.709842	14.20%

h. Component 1 can explain 81.86%, Component 2 can explain 15.30% and Component 3 can explain 2.29%

i. Screeplot:



j. Screeplot from the uncorrelated data is relatively flat, while for government data, the screeplot show a steep decline in principal components.

Since the uncorrelated data are uncorrelated, meaning for the standardized matrix, the off-diagonal are closer to 0. So the standardized matrix is "similar" to an identity matrix. As a result Eigenvalues are close to 1 and as a result the variance explained are relatively similar .Thus the screeplot is quite flat.

Unlike uncorrelated data, government data are often correlated, so we can expect the off-diagonal is different from 0. As opposite to uncorrelated data, the Eigenvalues will be different. As a result some eigenvalues will be significantly higher than some, as a result we could have a steeper screeplot.

Task 4

Executive Summary

For Task 4, we have chosen The Real Estate Select Sector SPDR Fund (XLRE) for analysis. The 30 largest holdings of XLRE were identified using data provided by State Street Investment Management[1]. The data was downloaded in CSV format and is already sorted in descending order. We then record the tickers of the top 30 holdings. Historical data for these holdings, covering approximately six months from March 23, 2025, to September 23, 2025, was obtained from Yahoo Finance for further analysis.

To analyze the data, we first calculated and standardized the price return data. Then, we applied Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) to the standardized covariance matrix to understand data further.

The first key finding is, from PCA, we found 74% of total variance will be explained by first 3 PC. While the top ten principal components explain 90% of the total variance, with immaterial contribution for subsequent components. This indicates that the majority of the ETF's systematic risk is driven by a relatively small number of factors. Additionally, both PCA and SVD produce identical eigenvalues, and their eigenvectors are equivalent, differing only in sign, aligning with the theoretical concepts discussed in Module 3, Lecture 4.

Methodology

We first set up the essential libraries: OS for importing data, yfinance to pull XLRE data, pandas and NumPy for data manipulation, and Seaborn and Matplotlib for plotting. Our goal is to fetch ETF prices, compute returns, and analyze their risk structure.

Next we're pulling 6 months of adjusted closing prices for XLRE's 30 largest holdings using Yahoo Finance from period: March 23, 2025, to September 23, 2025. Forming the basis for calculating daily returns and analyzing the ETF's risk and return.

We've converted the raw price series into daily returns, which represent the percentage change in price. This transformation is important because daily returns provide a standardized format to compare the relative performance of different holdings for XLRE. Using raw prices, which vary across holdings due to differing share prices, would affect the results based on the magnitude of the share price.

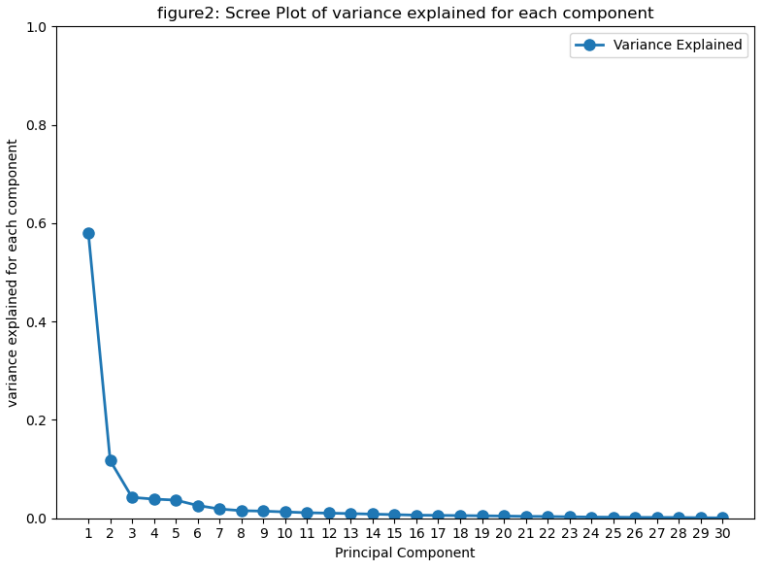
The second transformation is to apply a common statistical technique which transform the variable into a standard scale. It converts a variable so that it has a mean of 0 and a standard deviation of 1. It is important since PCA is a variance maximizing procedure, we want the variable having a same scale to uncover the maximum components and for fair comparison.

Group Number: 11446

PCA Results

Table1: Top 10 PCs

	Eigenvalues	Explained	Cumulative explained
1	17.426312	58.09%	58.09%
2	3.514055	11.71%	69.80%
3	1.275606	4.25%	74.05%
4	1.168617	3.90%	77.95%
5	1.109116	3.70%	81.65%
6	0.779281	2.60%	84.24%
7	0.565794	1.89%	86.13%
8	0.462744	1.54%	87.67%
9	0.434802	1.45%	89.12%
10	0.386216	1.29%	90.41%



74% of total variance will be explained by first 3 PC, if we want to explain 80% of the variance we could select the top 5 PCs. If we want more than 90% of explained variance, we can include top 10 PCs, after that the increment are extremely immaterial.

Group Number: 11446

SVD results

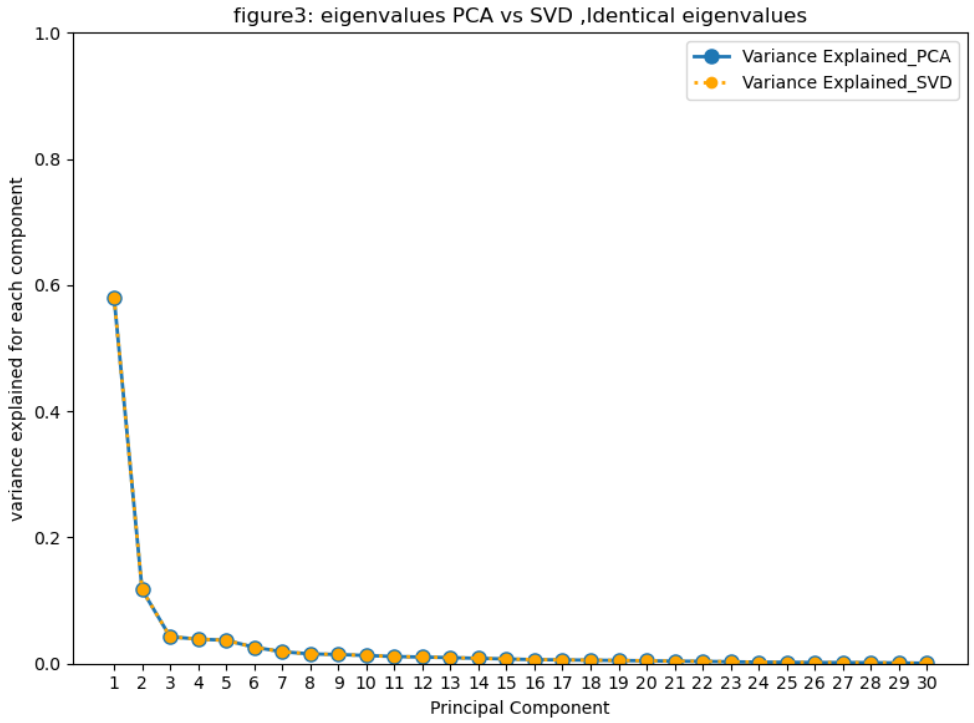
We perform SVD on standardized data, decomposing it into three components: U, a 127 x 127 orthonormal matrix; S, a 127 x 30 diagonal matrix with positive real numbers on the diagonal; and V a 30 x 30 orthonormal matrix.

The Singular values are:

[4.17448325	1.8745815	1.12942724	1.08102544	1.05314552	0.88276912
0.75219259	0.68025267	0.65939491	0.6214627	0.58067326	0.55895637
0.53315542	0.50283076	0.47109492	0.43491749	0.42084228	0.40438441
0.38847708	0.36812626	0.34160575	0.32937165	0.29648478	0.2689663
0.25305384	0.22984405	0.22607186	0.21473141	0.1801843	0.14362366]

To show the connection between SVD and principal component analysis (PCA), we follow what discussed in module 3 lesson 4, the squared singular values in matrix are eigen values of covariance matrix. We plot the results in figure 3, which show they are indeed the same.

As discussed, the transpose of the V matrix in SVD represents the eigenvectors. As expected, when compared with the eigenvectors from PCA, they differ only by sign. The results are plotted in Figure 4, the plot the ratio for each components of PCs, the magnitude of ratio is always 1.



REFERENCE

- Alexandra Jonker, and Alice Gomstyn. “Structured vs. Unstructured Data: What’s the Difference?” IBM Think, 7 Feb. 2025, [Link](https://www.ibm.com/think/topics/structured-vs-unstructured-data#:~:text=DynamoDB%2C%20Hadoop%20%20and%20,43.) Accessed 1 Oct. 2025.
- Freeman Mark. “Financial Data Quality: Modern Problems and Possibilities.” Gable, 7 Nov. 2024, [Link](https://www.gable.ai/blog/financial-data-quality-management#:~:text=Financial%20institutions%20often%20deal%20with,integration%20issues%20between%20different%20systems.) Accessed 28 Sept. 2025.
- Guide to Structured vs Unstructured Data: with Real-World Examples.” Domo, 17 June 2025, [Link](https://www.domo.com/learn/article/unstructured-data-vs-strucuctured-data?utm_source=%20https://www.domo.com/learn/article/unstructured-data-vs-strucuctured-data.) Accessed 1 Oct. 2025.
- Mirzaie, Mostafa, Behshid Behkamal, and Samad Paydar. “Data Quality: A Systematic Literature Review and Future Research Directions.” Ferdowsi University of Mashhad, 2019. [PDF](https://arxiv.org/pdf/1904.05353)