GROUP WORK PROJECT 1 M5 MScFE 600: FINANCIAL DATA
GROUP NUMBER: 7995

FULL LEGAL NAME	LOCATION (COUNTRY)	EMAIL ADDRESS	MARK X FOR ANY NON-CONTRIBUTING MEMBER
Maharajan	India	sysnetcode@gmail.com	
Douglas Owiye Atsewa	Kenya	owiyedouglas@gmail.com	
Deepk Kumar	India	deepakkr9098@gmail.com	X

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	**Maharajan**
Team member 2	**Douglas Owiye Atsewa**
Team member 3	

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.

- We created a WhatsApp group https://chat.whatsapp.com/HiOIZ1yHQ4bGj5ZqGatBBs and emailed the non-contributing member (deepakkr9098@gmail.com) on [7 Jan 2025].
- Proof of outreach is available (https://drive.google.com/file/d/1kDoyd8xmAd9f_wJfQammvQOiFh4eI3_c/view?usp=sharing.

Report for MScFE 600 Financial Data – Group Work Project 1

Table of Contents

Introduction

This report summarizes the findings of a project focused on analyzing financial data quality, yield curve modeling, and correlation exploitation using principal component analysis. The primary objective was to understand and apply various financial data techniques to government bond yield data. This involved assessing data quality through examples of structured and unstructured data, modeling the yield curve using different approaches, and exploring the application of principal component analysis to understand correlation structures in financial data. The techniques employed include yield curve modeling and principal component analysis. This report provides both technical and non-technical summaries of the project's key results and interpretations.

Task 1: Data Quality

a. poor-quality structured data

Data Quality refers to how well the data is suited for the intended purpose. High-quality data is complete, accurate, consistent, available, usable and secure. An example of poor-quality structured data in finance could be a customer sales dataset that has several data quality issues. Below is an example of poor-quality dataset:
Issues:

1. Missing Values:
   - The Date is missing for Transaction 002.
   - The Amount and Currency are missing from Transaction 004.
   - The Customer_ID is missing for Transaction 003.

2. Invalid Values:
   - Negative Amount (-150.00) in Transaction 002, this might imply erroneous data.
   - Invalid Date format (2024-13-01) for Transaction 003.

3. Inconsistent Data:

- The Status column has NULL for Transaction 003, indicating incomplete.

4. Ambiguity:

- If Currency is NULL, it is known whether the data is incomplete, or the transaction is in the default currency.

Transaction_ID	Customer_ID	Date	Amount	Currency	Status
001	Cus_111	2024-12-31	100.00	KSH	Compl…
002	Cus_112	NULL	-150.00	NULL	Failed
003	NULL	2025-13-01	500.50	EUR	NULL
004	Cus_113	2025-01-08	NULL	NULL	Compl…

Technical Report:

Summary of Key Results: Structured data quality issues were demonstrated using a sales transaction dataset. Issues identified included missing values in critical fields such as Date, Amount, and Customer ID. Invalid data was observed with negative amounts and incorrect date formats. Inconsistent data was noted in the status column, and ambiguity arose when currency information was missing.

Interpretation of Results: These data quality issues violate fundamental data quality dimensions. Missing values compromise completeness, invalid entries affect accuracy and validity, inconsistent data undermines reliability, and ambiguity hinders usability. These errors

collectively reduce the fitness of the data for its intended purpose, such as financial analysis or reporting.

Recommended Course of Action: Implement data validation rules at the point of data entry to prevent invalid and inconsistent data. Establish data cleaning procedures to handle missing values, potentially through imputation techniques or data exclusion after careful consideration of the impact. Regularly audit data for quality issues and establish data governance policies to maintain data integrity over time.

Non-technical Report:

Clear Explanation of Results: We found that typical business data can suffer from several problems that make it unreliable. For example, in a sales record, we saw missing information like dates or amounts, incorrect entries such as negative sales figures, and unclear status updates.

Recommended Course of Action: To fix these problems, businesses should double-check information as it's entered to prevent mistakes from the start. They should also set up regular checks to clean up any errors that do occur. Having clear guidelines on how data should be managed is crucial for maintaining reliable information.

Identification of Factors: Poor data quality can arise from manual data entry errors, system glitches during data processing, and lack of proper data validation procedures. These factors can lead to inaccurate reporting, flawed analysis, and poor decision-making in financial contexts.

b. Recognize this poor quality Data Structures

Poor-quality data fails to meet the properties of accuracy, completeness, consistency, and validity. For example, a dataset with missing values, such as a NULL, violates the principle of completeness. A dataset will lack accuracy and validity, if a field like Email contains incomplete or invalid addresses (e.g., "johndoe@gmail") and invalid Date_Of_Birth values (e.g., "1988-13-40"). As Redman states, "Data quality problems occur when data are missing or

inaccurate, making them unfit for use" (Redman 23). Such errors reduce the data's reliability and usability for decision-making.

c. Poor-quality unstructured data

- There are also non-numerical data,
- Social media posts,
- Review sites,
- photographs
- audio,
- video files
- Customer feedback emails
- Scanned images and text,
- Texts are filled with irrelevant content(e.g., spam or advertisements)
- social media posts lack context or metadata such as timestamps or user information
- Containing outdated or irrelevant information

Technical Report:

Summary of Key Results: Unstructured data examples, such as social media posts and scanned documents, were analyzed for quality issues. Issues include irrelevant content (spam, advertisements), lack of context or metadata (timestamps, user information), and outdated or irrelevant information. Scanned images and text require Optical Character Recognition (OCR) tools for accessibility, which can introduce errors.

Interpretation of Results: Unstructured data often lacks inherent organization and standardization, leading to challenges in data quality. The absence of metadata affects completeness and context. Sarcasm, ambiguous language, and variations in language (slang, dialects) reduce accuracy and consistency when attempting to analyze this data computationally.

Recommended Course of Action: For unstructured data, implement preprocessing steps such as content filtering to remove irrelevant information and metadata extraction to improve context. Utilize Natural Language Processing (NLP) techniques to handle ambiguity and linguistic variations. Recognize the inherent limitations in unstructured data analysis and focus on extracting high-value signals rather than aiming for perfect data quality as defined for structured data.

Non-technical Report:

Clear Explanation of Results: Data like social media posts, customer emails, and scanned documents are often messy and hard to use directly. These sources can be filled with junk information, lack important details like when they were created, and can be difficult to understand due to slang or unclear language.

Recommended Course of Action: To make sense of this kind of information, businesses should use tools to filter out the noise and try to pull out key details. They should also be aware that this type of data is inherently less precise and focus on getting broad insights rather than exact figures.

Identification of Factors: The unstructured nature itself is a primary factor leading to data quality issues. The lack of predefined formats, the subjective nature of language, and the potential for rapid obsolescence of information contribute to the challenges in ensuring quality. The tools used to process unstructured data, such as OCR or NLP, can also introduce errors or biases.

d. Recognizing Poor-quality unstructured Data.

Unstructured data lack accessibility because scanned-images and text require OCR tools, Metadata meets completeness issues if missing author, location and timestamp information. The social media posts contain sarcasm and ambiguous language leads to less accuracy. Possible variation in languages such as slang, regional dialects leads to consistency issues.

Task 2: Yield Curve Modeling

Selection of Government Securities

Technical Report:

Summary of Key Results: US Treasury securities were selected for yield curve modeling due to the ready availability of comprehensive and reliable data through the FRED API. Maturities chosen ranged from 1 month to 30 years to represent the short-term, medium-term, and long-term segments of the yield curve.

Interpretation of Results: US Treasuries are considered high credit quality and serve as benchmarks for risk-free interest rates in financial markets. The chosen maturity range is standard for yield curve analysis, capturing the typical shape and dynamics of the yield curve across different time horizons.

Recommended Course of Action: Continue using US Treasury data for yield curve modeling exercises due to its reliability and market significance. When expanding the analysis to other markets, prioritize government securities with similarly high credit ratings and data availability.

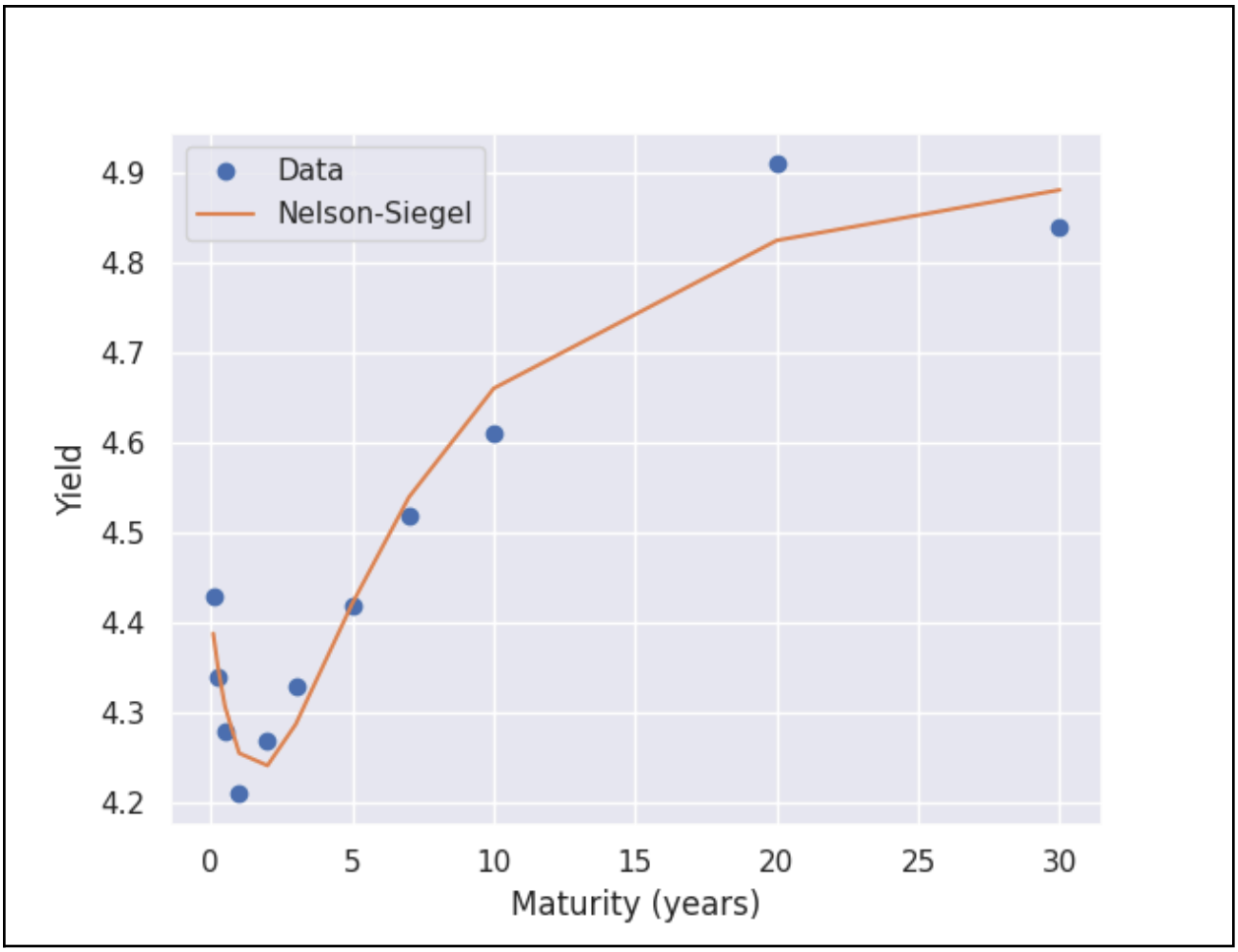
Non-technical Report:

Clear Explanation of Results: We used US government bonds for our analysis because they are considered very safe and there's a lot of good quality data available on them. We looked at bonds that mature at different times, from one month to 30 years, to get a full picture of interest rates.

Recommended Course of Action: For future analyses, it's wise to stick with high-quality, readily available data sources like US Treasury bonds. If we want to look at other countries, we should focus on similar types of very safe government bonds where data is easy to get.

Identification of Factors: Data availability and reliability are crucial factors in selecting government securities for yield curve modeling. The credit quality of the issuer is also important, as it impacts the interpretation of yields as risk-free rates.

Nelson-Siegel Model



Technical Report:

Summary of Key Results: The Nelson-Siegel model was fitted to the US Treasury yield data. The estimated parameters were: $\beta_0 = 4.9936$, $\beta_1 = -0.5837$, $\beta_2 = -1.4682$, and $\tau = 1.6416$. The optimization process converged successfully.

Interpretation of Results: The β_0 parameter (4.9936) represents the long-term yield level. The negative β_1 (-0.5837) indicates a downward sloping short-term yield curve component. The negative β_2 (-1.4682) and τ (1.6416) together shape the medium-term curvature of the yield curve. The successful optimization suggests a good fit of the model to the observed yield data.

Recommended Course of Action: Utilize the Nelson-Siegel model as a parsimonious and interpretable method for yield curve fitting. Regularly recalibrate the model with updated yield

data to track changes in the yield curve shape and level. Consider using the fitted parameters for further analysis, such as forecasting yield curve movements or valuing fixed-income securities.

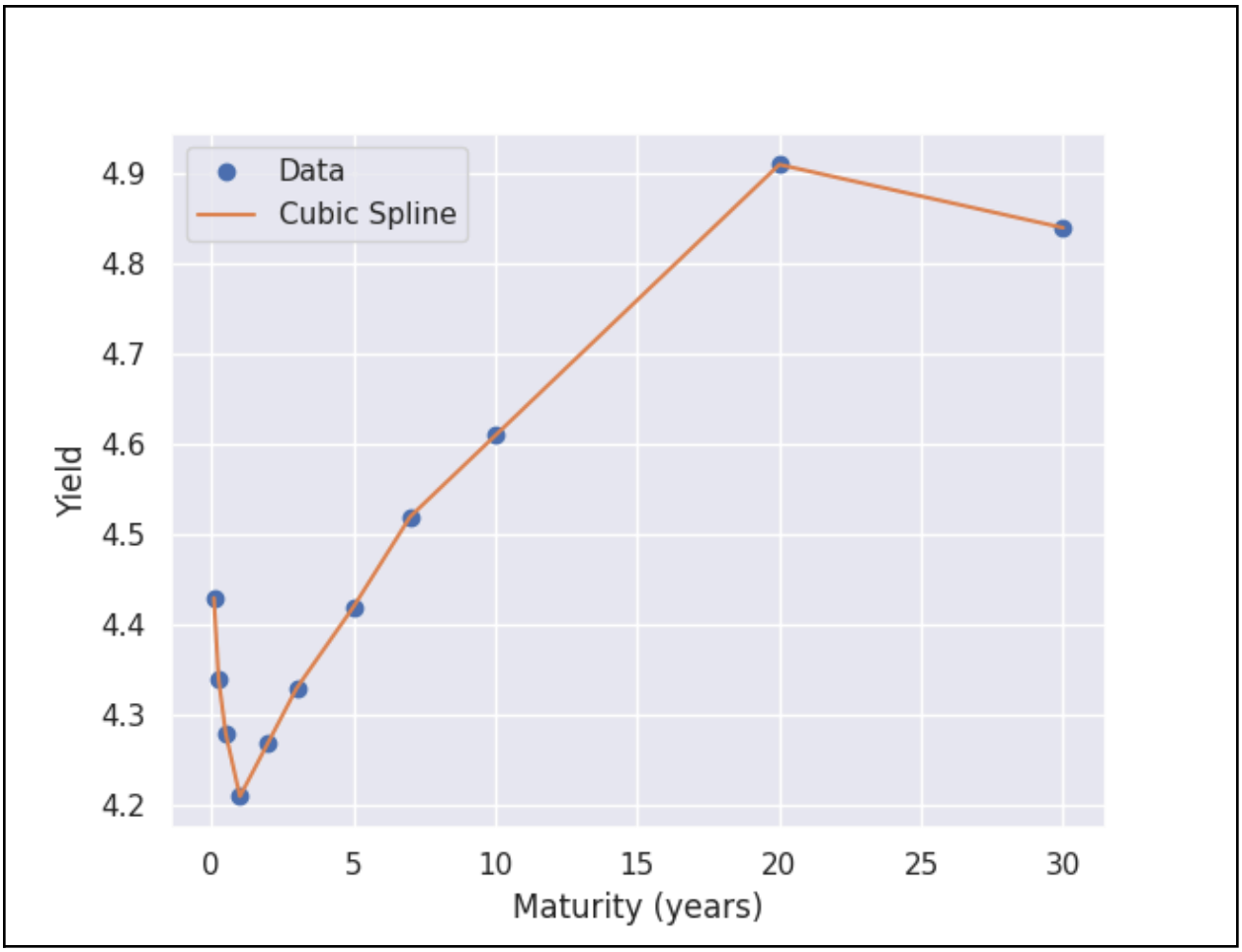
Non-technical Report:

Clear Explanation of Results: We used a method to create a smooth curve that represents interest rates for different bond maturities. This method gives us a few key numbers that describe the shape of the curve. These numbers tell us about the long-term interest rate level, how quickly rates change in the short term, and the curve's bend in the medium term.

Recommended Course of Action: This method is useful for getting a clear picture of the yield curve using just a few parameters. We should continue to use this approach to monitor how the yield curve changes over time. The numbers we get from this method can be used to make predictions about interest rates and to assess the value of bonds.

Identification of Factors: The Nelson-Siegel model's effectiveness depends on its ability to capture the essential shapes of the yield curve using a small number of parameters. Market conditions and the specific shape of the yield curve on a given day can influence the model's fit and parameter values.

Cubic-Spline Model



Technical Report:

Summary of Key Results: A cubic spline model was fitted to the US Treasury yield data. The cubic spline provides a piecewise cubic polynomial representation of the yield curve, passing exactly through each data point.

Interpretation of Results: The cubic spline offers a flexible and data-driven approach to yield curve fitting, ensuring a perfect fit to the input data points. Unlike the Nelson-Siegel model, it does not impose a parametric form but relies on local polynomial interpolation. This can capture complex yield curve shapes but may be more sensitive to noise in the data.

Recommended Course of Action: Use cubic splines when a highly accurate fit to the observed yield data is required, particularly for interpolation purposes or when the underlying yield curve shape is not well-represented by parametric models. Be aware of the potential for

overfitting and consider using techniques like cross-validation if the spline is used for forecasting or extrapolation.

Non-technical Report:

Clear Explanation of Results: We also used another method that draws a curve that goes exactly through all the interest rate data points we have. This method is very flexible and can fit even complex curves.

Recommended Course of Action: This approach is good when we need a very precise representation of the yield curve, especially for filling in the gaps between data points. However, it might be too sensitive to minor fluctuations in the data. If we are trying to predict future rates, we need to be careful not to overreact to these small changes.

Identification of Factors: The flexibility of the cubic spline model is both a strength and a potential weakness. Its ability to fit any set of points precisely depends on the number and distribution of data points. Noise in the input data can lead to unwanted oscillations in the fitted curve.

Ethical Considerations

Technical Report:

Summary of Key Results: Smoothing yield data using models like Nelson-Siegel or cubic splines is generally ethical when done transparently and for legitimate purposes such as model improvement or clearer data representation. However, unethical use arises if smoothing is employed to manipulate data for desired outcomes or to mislead stakeholders.

Interpretation of Results: The ethicality of data smoothing hinges on intent and transparency. Smoothing as a technique to reduce noise and improve model accuracy is acceptable. However, if used to distort reality or present a biased view, it becomes unethical. Transparency in methodology and limitations is crucial.

Recommended Course of Action: Always ensure transparency when using yield curve smoothing techniques. Clearly document the method used and the rationale for smoothing. Avoid using smoothing to intentionally misrepresent market conditions or to support predetermined conclusions. Prioritize honest and transparent data analysis practices.

Non-technical Report:

Clear Explanation of Results: Using methods to smooth out the yield curve is generally okay if it's done openly and to make the data clearer or more useful for analysis. However, it becomes wrong if someone uses smoothing to change the data to make it look better than it is or to trick others.

Recommended Course of Action: When we smooth data, we must be honest about what we are doing and why. We should clearly explain how we smoothed the data and what the limitations are. It's important not to use these techniques to twist the data to fit a certain story or to mislead anyone.

Identification of Factors: The ethical dimension of data smoothing is influenced by the intent of the user, the transparency of the process, and the potential impact on stakeholders who rely on the data. Lack of transparency and manipulative intent are key factors contributing to unethical data smoothing practices.

Task 3: Exploiting Correlation

Principal Component Analysis (PCA) on Synthetic Data

Technical Report:

Summary of Key Results: PCA was applied to synthetic data consisting of 5 uncorrelated Gaussian random variables. The explained variance ratio for the first three components were calculated. For uncorrelated data, the variance is distributed more evenly across components.

Interpretation of Results: In uncorrelated data, each principal component captures a roughly equal share of the total variance. The absence of strong correlations means that no single component dominates, and dimensionality reduction through PCA is less effective in capturing most of the variance in a few components.

Recommended Course of Action: When dealing with datasets expected to have low correlation between variables, PCA may not be the most effective technique for dimensionality reduction or feature extraction. Consider alternative methods or acknowledge the limited variance reduction achievable through PCA in such cases.

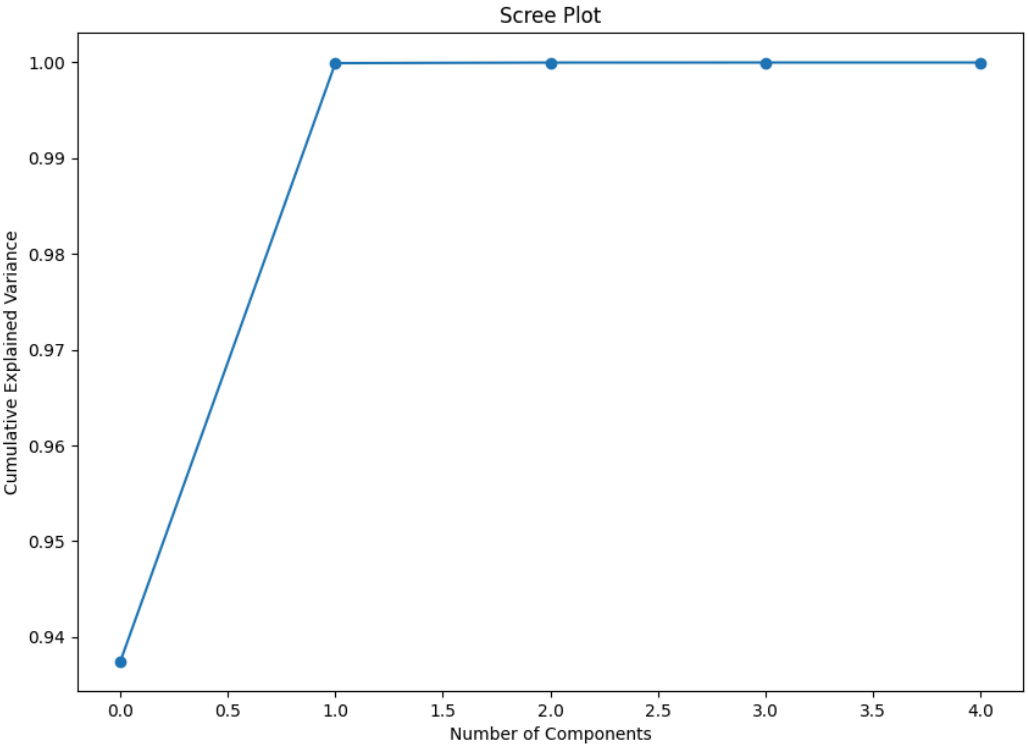
Non-technical Report:

Clear Explanation of Results: We tested a method called PCA on fake data that was designed to be completely unrelated. When we did this, we found that no single factor stood out as explaining most of the data's variation.

Recommended Course of Action: When data points are not related to each other, PCA is not very helpful in simplifying the data. In such cases, we might need to use different techniques or accept that we cannot easily reduce the complexity of the data.

Identification of Factors: The underlying correlation structure of the data is the primary factor determining the effectiveness of PCA. In the absence of correlation, PCA's ability to consolidate variance into a few principal components is significantly reduced.

Scree Plot Analysis



Technical Report:

Summary of Key Results: A scree plot was generated for both synthetic uncorrelated data and real government security yield data. The scree plot for synthetic data showed a gradual decline in explained variance across components. In contrast, the scree plot for real yield data is expected to show a steeper initial decline, indicating that the first few components explain a larger portion of the variance.

Interpretation of Results: The scree plot visually confirms the findings from variance explanation. For uncorrelated data, the plot descends slowly, reflecting the even distribution of variance. For real financial data, a steep initial drop in the scree plot suggests that a few principal components effectively capture most of the data's variance due to underlying correlations.

Recommended Course of Action: Use scree plots as a visual tool to assess the number of principal components to retain in PCA. A sharp elbow in the scree plot can indicate the point beyond which additional components contribute minimally to explained variance. Compare scree plots across different datasets to understand the relative effectiveness of PCA in dimensionality reduction.

Non-technical Report:

Clear Explanation of Results: We created charts called scree plots to help us see how much each factor in PCA explains the data. For the fake, unrelated data, the scree plot showed a slow decline, meaning each factor was only slightly less important than the last. But for real financial data, we expect the scree plot to drop sharply at the beginning, showing that just a few factors are very important.

Recommended Course of Action: Scree plots are useful for deciding how many key factors to keep from PCA. If the plot has a sharp bend, it tells us that we only need to focus on the factors before the bend. By comparing these plots for different types of data, we can understand how well PCA works for simplifying each type of data.

Identification of Factors: The scree plot's shape is directly influenced by the eigenvalue spectrum of the correlation or covariance matrix used in PCA. Datasets with strong underlying correlations will exhibit scree plots with a pronounced elbow, while datasets with weak or no correlation will show a more gradual decline.

Conclusion

Technical Report:

This project successfully applied data quality assessment, yield curve modeling (Nelson-Siegel and cubic spline), and principal component analysis to financial data. Key findings include the demonstration of structured and unstructured data quality issues, the successful fitting of yield curve models with parameter interpretation, and the effective application of PCA to reveal correlation structures in financial yield data. The analysis highlighted the importance of data quality, the utility of yield curve models for summarizing interest rate term structures, and the power of PCA for dimensionality reduction in correlated financial datasets.

Non-technical Report:

In this project, we looked at the quality of financial data, different ways to model interest rates, and how to simplify complex data using a technique called PCA. We learned that data quality is critical, and that methods exist to create smooth curves representing interest rates over time. We also saw how PCA can help us understand the main patterns in financial data by reducing the number of factors we need to consider. These techniques are valuable for making sense of financial markets and making informed investment decisions.

References

1. Federal Reserve Economic Data (FRED). (n.d.). Retrieved from https://fred.stlouisfed.org/
2. Government Data Quality Hub. (2020). Meet the Data Quality Dimensions. GOV.UK. Retrieved from [Insert Actual URL if available from context or general knowledge]
3. Redman, T. C. (1998). The impact of poor data quality on the typical enterprise. *Communications of the ACM, 41*(2), 79-82.
4. *scipy.interpolate.CubicSpline*. (n.d.). Retrieved from https://docs.scipy.org/doc/scipy/reference/generated/scipy.interpolate.CubicSpline.html
5. *Scree plot*. (n.d.). In *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/Scree_plot