THE UNIVERSITY *of* ADELAIDE

Faculty of the Professions

Assignment Cover Sheet

To be completed by Individual Student for Individual Assignment, or Group Leader for Group Assignment.

First Name:	Chinmay Praveen	**Last Name:**	Shetty
Student ID:	a 1916150	**Assignment Due Date:**	25th August 2024
Course Name:	Predictive and Visual Analytics for Business	**Course Code:**	BUSANA 7001
Lecturer:	Professor Sigitas Karpavicius	**Tutor:**	Mr Md Jakir Hasan Talukder
Tutorial Day:		**and Time:**	

To be completed by all other Group Members for Group Assignments.

Student Name:	**Student ID:**	**Declaration Signature:**
	a	
	a	
	a	
	a	
	a	

CHECKLIST

1. All pages are securely stapled or bound	2. You have kept a copy of the assignment	3. The assignment is submitted without a plastic cover or folder

DECLARATION

This declaration must be signed before a mark can be awarded.

I/We declare the following to be my/our own work as understood by the University's Policy on Plagiarism (see *Statement and Definition of Plagiarism and Related Forms of Cheating*, www.adelaide.edu.au/policies/230).

I/We give permission for my/our assignment to be scanned for electronic checking of plagiarism.

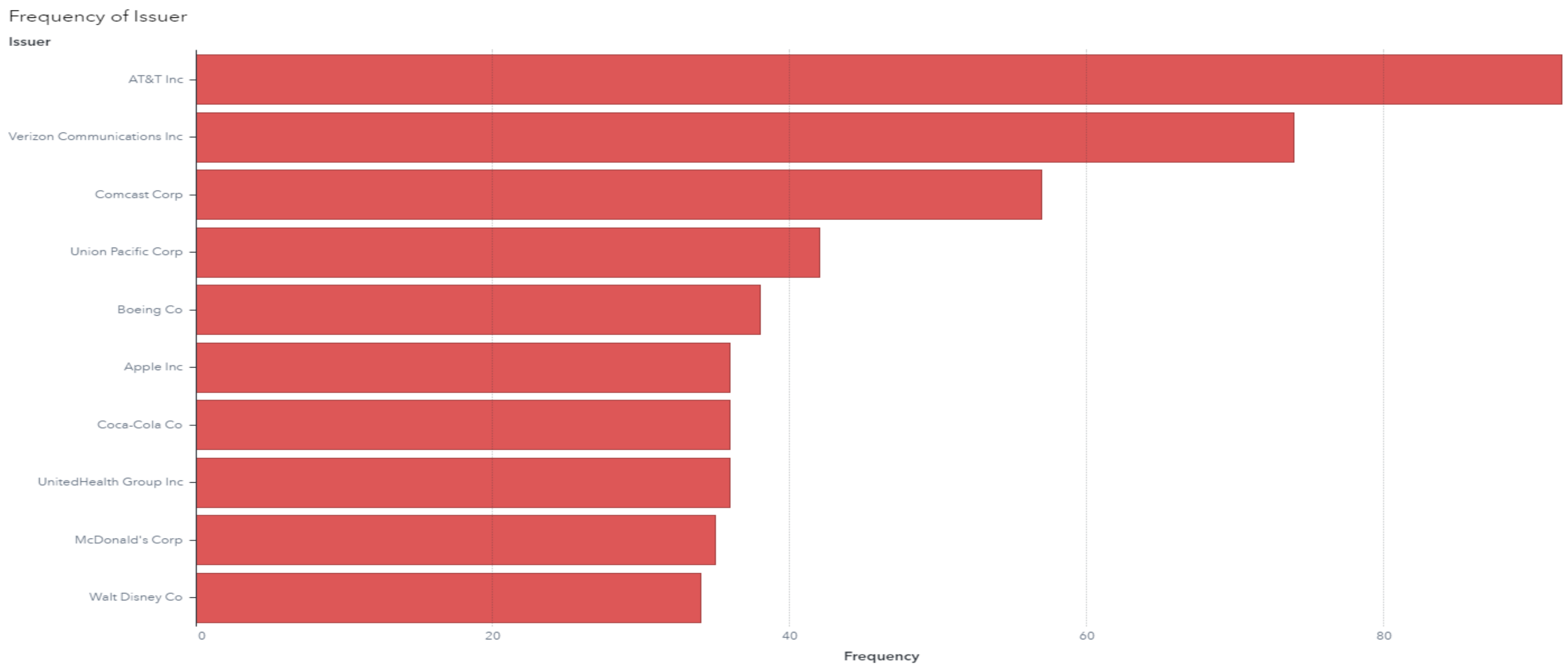
Signed:	Shetty	Date:	25th August 2024

OFFICE USE ONLY (for late assignments)
Assignment received by Professions Student Support Hub

Date:		Time:	
Staff Signature:		Name (print):	

Mark:

Chinmay Praveen Shetty (a1916150)
BUSANA 7001 Visual and Predictive Analytics
Assignment 1
Task 1 – Visual Slides



Frequency of Issuer

This horizontal bar-chart displays the frequency of bond issuances by various companies, with AT&T Inc. being the most frequent issuer, followed by Verizon Communications Inc. and Comcast Corp. This shows the strong presence of telecom companies in the bond market, which can be interpreted as their huge requirement of capital, issuers from different sectors, such as technology (Apple Inc.) and consumer goods (Coca-Cola Co.), displays the wide representation in the data. This kind of diversity is very important as it help analysis of the bond market and how various industries affect the bond yields. The frequency of issuance can influence the liquidity and investor familiarity, which are important factors to consider in the regression analysis of bond yields.

Coupon by Sector & Frequency

This heat map chart illustrates the frequency distribution of bond coupons across various sectors. In a heat map, the colour intensity represents the frequency of coupon values within each sector, with darker shades indicating higher frequencies.

This chart highlights how coupon rates are spread among sectors such as Cable/Media, Electronics, and Telecommunications. The frequency values range from 0 to just over 2, with sectors like "Service - Other" having the highest frequency of bonds issued with a specific coupon rate.

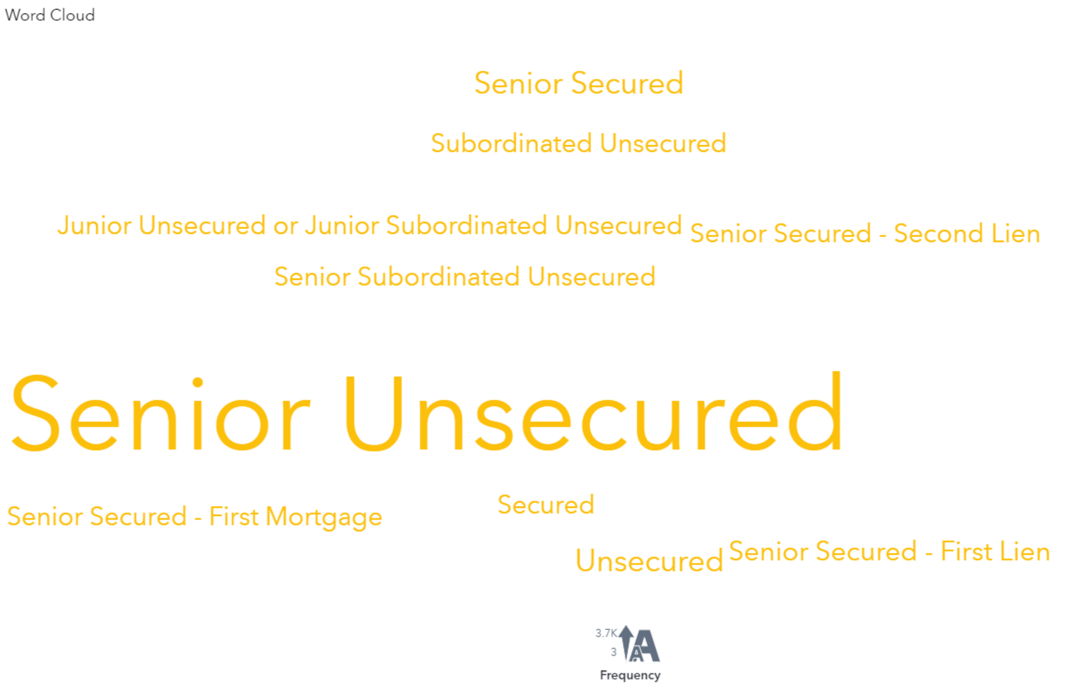
It is necessary to understand this distribution as we can get to know the sector specific risks that influence the bond yields, which is an essential insight for regression analysis on bond yields. For example, sectors like "Service – Others" show a higher concentration of bonds with a particular coupon rate, displaying this sectors characteristics.

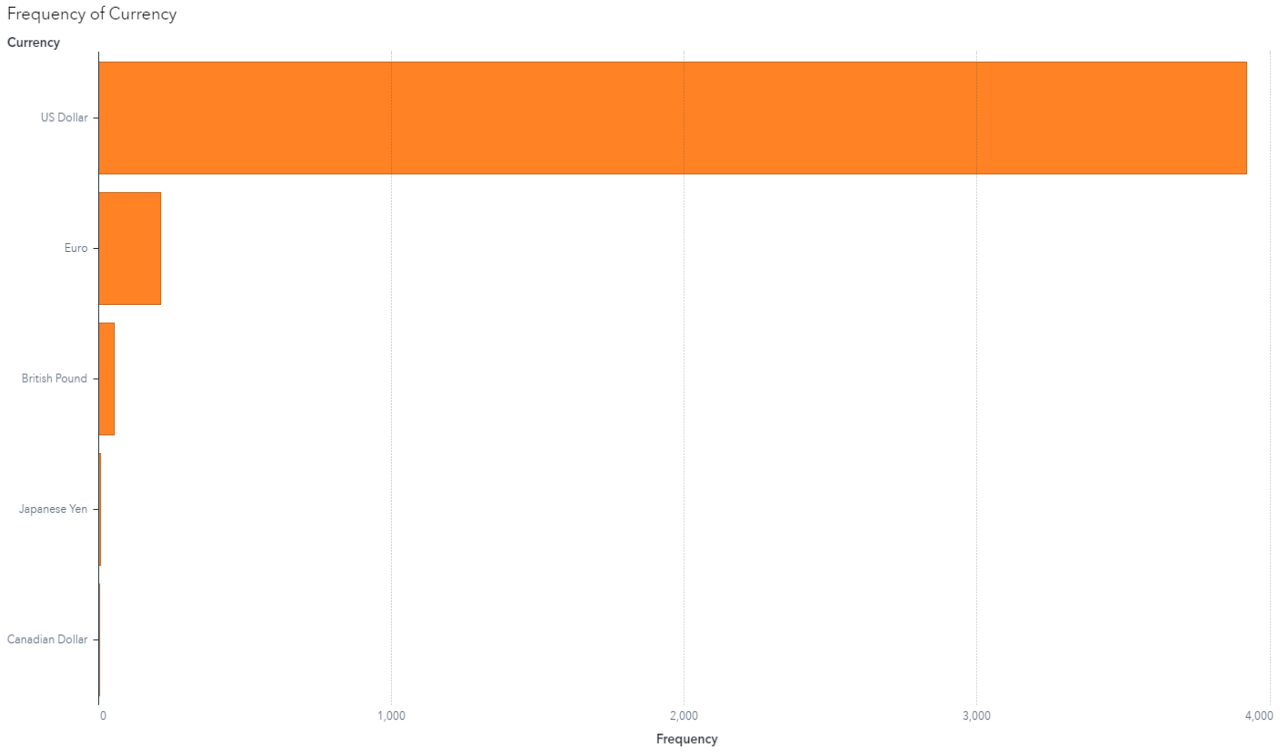
Word Cloud

This is a word cloud chart, which highlights the frequency of different bond seniority classifications in the dataset. In a word cloud, the size of each word depends on the frequency, which means larger words have higher frequency.

The biggest word here is "Senior Unsecured“, which indicates that this classification is the most common among the bonds in the dataset. This is significant because the seniority of a bond affects its risk profile and, consequently, its yield.

The “Senior Unsecured” bonds are dominating the dataset, compared to the relative sizes of other classifications, such as “Senior Secured” and “Subordinated Unsecured”, this gives us an estimate of the risk distribution across the bond sample. These variations in seniority could significantly impact the yield predictions and the overall risk assessment in the U.S. corporate bond market.

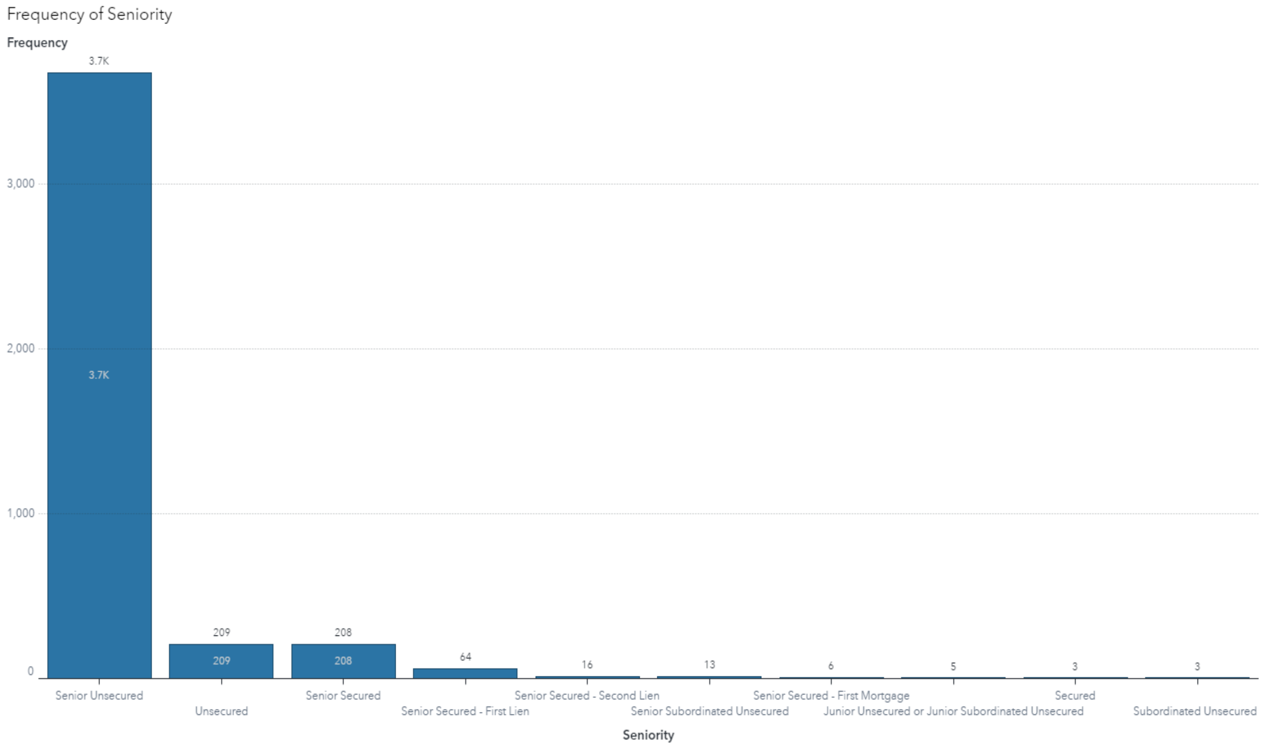




Frequency of Currency

This horizontal bar chart shows the frequency distribution of bonds based on their currency denomination. With a frequency near 4000, this graphic demonstrates that the majority of the bonds are in US Dollars. There are also some lower frequency currencies like Euro and British Pound, while currencies like the Japanese Yen and Canadian Dollar appear only marginally.

Of course since the dataset is based upon the U.S corporate bond market, this kind of US Dollar oriented chart was expected. This dominance is important for the analysis as it suggests that the findings, especially in the context of yield predictions, will be heavily influenced by U.S. market conditions and monetary policies. On the other hand the low represented currencies should be carefully interpreted during the analysis due to smaller sample size.



Frequency of Seniority

This vertical bar chart illustrates the frequency distribution of different bond seniority levels within the dataset. "Senior Unsecured" bonds has the tallest bar with a frequency of 3.7K, which is much higher than the other categories like "Unsecured" (209) and "Senior Secured" (208).

This overwhelming presence of "Senior Unsecured" bonds indicates that the dataset is heavily weighted towards bonds that are not backed by collateral, which tend to carry a higher risk. This factor is critical when analysing bond yields, as higher risk typically correlates with higher yields.

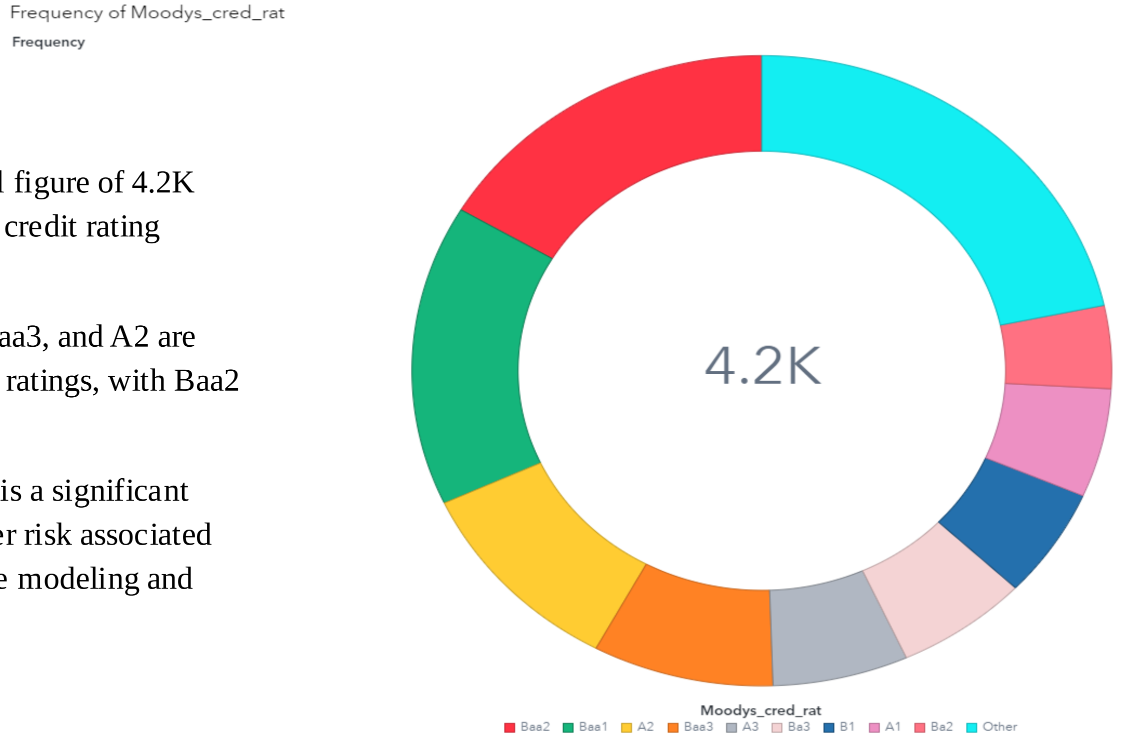
The dataset has very less representation of secured debt as it can be seen with the lower frequencies of other seniority levels, such as "Senior Secured" and "Subordinated Unsecured“. This skew should be considered when interpreting the results of regression models, as it may influence the robustness of the yield predictions.

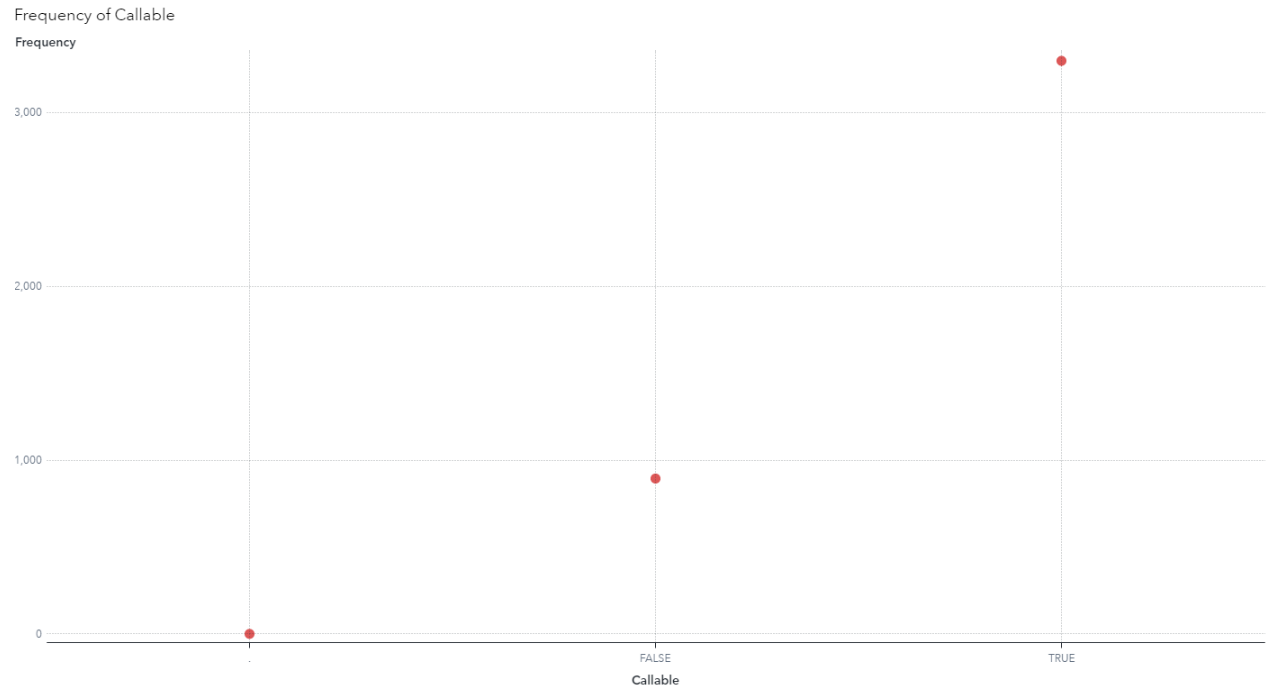
Frequency of Moodys_cred_rat

This donut chart displays the frequency distribution of Moody's credit ratings in the dataset. The central figure of 4.2K represents the total number of observations, while each segment of the donut corresponds to a different credit rating category.

The chart highlights the distribution of various credit ratings among the bonds. Categories like Baa2, Baa3, and A2 are prominently represented, indicating that the majority of the bonds in the dataset have investment-grade ratings, with Baa2 being a key rating category.

Understanding this distribution is crucial for the regression analysis on bond yields, as the credit rating is a significant determinant of bond yield. Bonds with lower credit ratings typically have higher yields due to the higher risk associated with them. This chart provides a clear visualization of how the bonds are rated, which will influence the modeling and prediction of bond yields in your analysis.





Frequency of Callable

This scatter plot chart shows the frequency distribution of callable and non-callable bonds in the dataset. Callable bonds, indicated as "TRUE," have a frequency of approximately 3,000, whereas non-callable bonds, indicated as "FALSE," have a much lower frequency, close to 1,000.

The high frequency of callable bonds, with around 3,000 occurrences, indicates that a substantial portion of the bonds in the dataset includes a call option. This is crucial for the analysis, as callable bonds generally offer higher yields to compensate investors for the risk of early redemption. On the other hand, the lower frequency of non-callable bonds (about 1,000) suggests that these types of bonds are less common in the dataset, which might influence the yield analysis, as non-callable bonds usually have a more predictable cash flow.

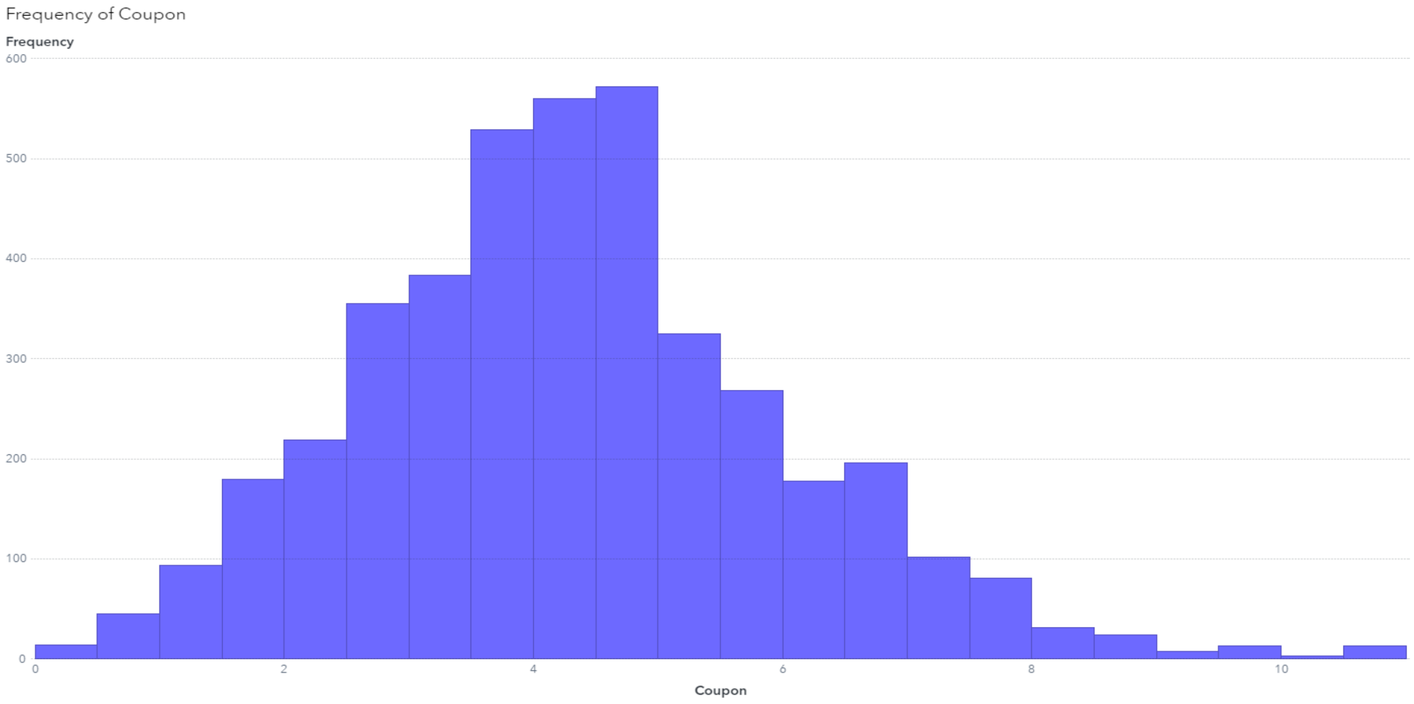
Understanding this distribution is important when analysing bond yields, as the callable feature can significantly impact the bond's return profile. This distribution will need to be accounted for in any regression models to ensure accurate yield predictions.

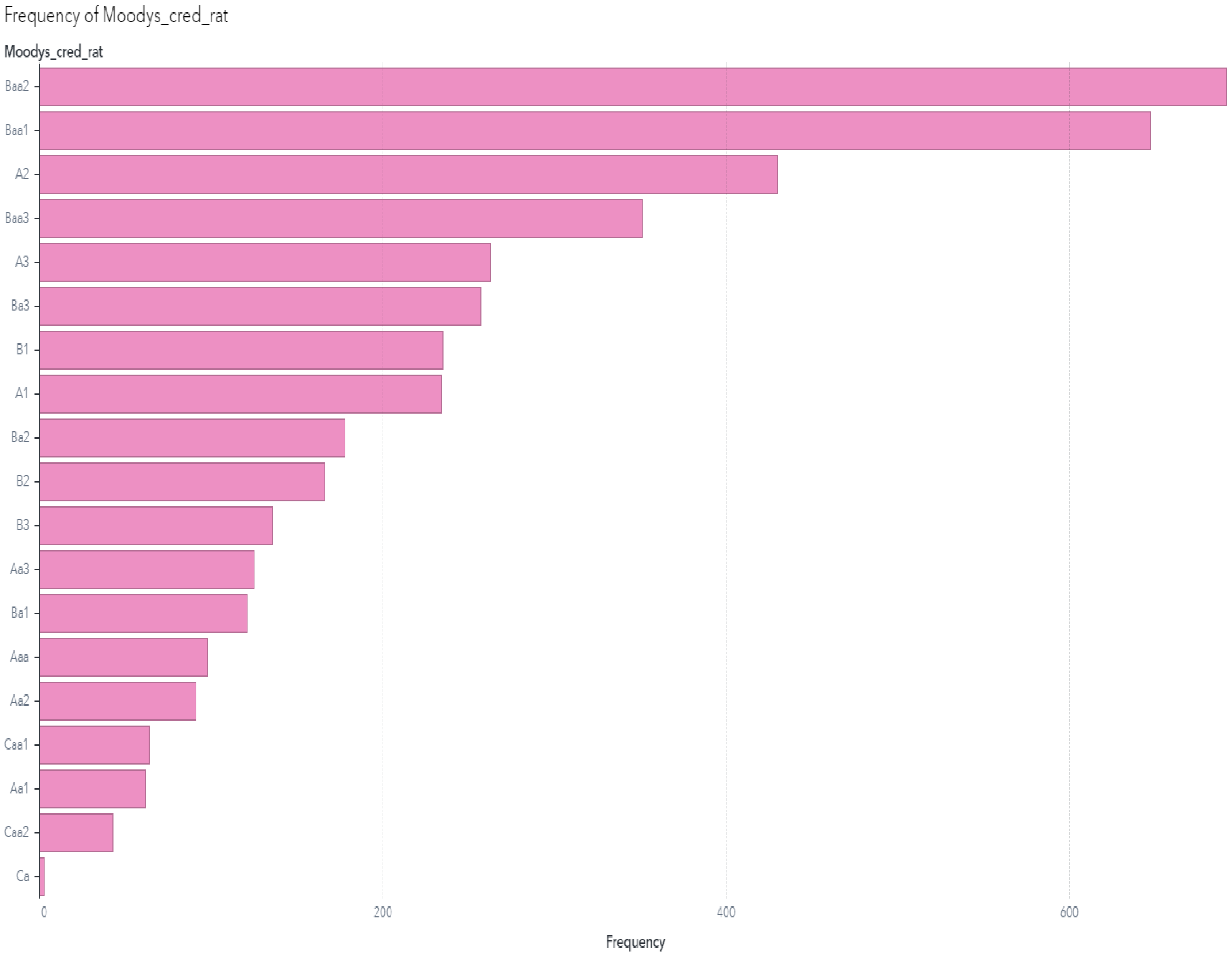
Frequency of Coupon

This histogram displays the frequency distribution of bond coupon rates in the dataset. The chart shows that most bonds have coupon rates clustered between 2% and 6%, with the highest frequency around the 4% mark. The distribution appears to be skewed slightly to the right, with fewer bonds offering higher coupon rates above 6%.

The distribution of coupon rates is critical for understanding the overall yield landscape of the bonds in your dataset. The high concentration of bonds with coupon rates around 4% suggests that this is a common rate offered in the market, possibly reflecting prevailing interest rates at the time these bonds were issued.

The lower frequency of bonds with very high coupon rates (above 6%) indicates that such bonds are less common, likely due to the higher cost of borrowing associated with these rates. This distribution should be taken into account when building regression models to predict bond yields, as the prevalence of mid-range coupon rates will have a significant impact on the model's predictions.





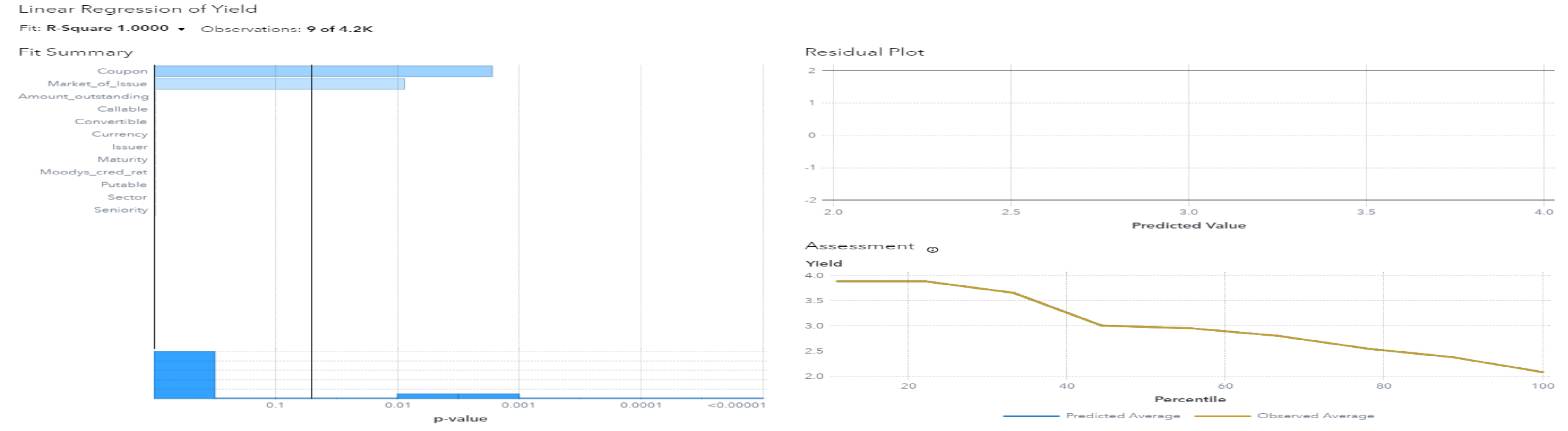
Frequency of Moodys_cred_rat

This horizontal bar chart shows the frequency distribution of Moody's credit ratings among the bonds in the dataset. The most frequent ratings are Baa2 and Baa1, each with a frequency close to 600, followed by A2 with a frequency around 500. Lower ratings like Ca and Caa2 have the least representation in the dataset.

The dataset contains a sizable percentage of bonds with investment-grade ratings that are towards the lower end of the investment-grade spectrum, as evidenced by the concentration of bonds with Baa2 and Baa1 ratings. These ratings typically represent bonds that are more susceptible to economic changes but are still considered a safe investment.

The presence of a smaller number of bonds with lower ratings, such as Caa2 and Ca, suggests that high-yield or "junk" bonds are less represented in this dataset. This distribution is essential for yield analysis, as the bonds with lower ratings typically give greater yields to offset the additional risk

The skew towards investment-grade bonds may influence the regression models used for predicting bond yields, as these bonds are typically less volatile than those with lower ratings. This distribution should be considered when interpreting the results and making predictions about bond yields.



Linear Regression of Yield

Fit Summary:

R-Square: The model's R-squared value is 1.0000, indicating that the regression model explains 100% of the variance in the yield. This is a perfect fit, which suggests that the model is highly accurate, though in practice, such a result might indicate overfitting.

Significant Variables:

Coupon: This variable shows a very low p-value (<0.00001), meaning it is highly significant in predicting bond yield.

Market of Issue and **Amount Outstanding** are also significant predictors, as indicated by their low p-values.

Residual Plot: The residual plot, which typically shows the difference between observed and predicted values, appears to be flat and centred around zero. This could indicate that the model has little to no error in predictions, aligning with the perfect R-squared value. However, this lack of residuals can also be a sign that the model may be overfitted.

Assessment: The assessment graph shows how the predicted yield compares with the observed average yield across percentiles. The close alignment between the predicted average (blue line) and the observed average (yellow line) suggests that the model's predictions are very accurate across the dataset.

This regression analysis is crucial in understanding which factors most significantly influence bond yield. The high significance of the coupon rate, market of issue, and amount outstanding emphasizes their importance in yield determination. The perfect fit of the model, while impressive, should be critically assessed to ensure the model isn't overfitted and remains generalizable to other datasets.

TASK 2 & 3 BUSINESS REPORT

Contents

Executive Summary

The primary goal of this project is to analyse the U.S corporate bond market based on the provided data. The bond dataset has been provided in 2 files - 'Sample_a.csv' and 'Sample_b.csv', which encompasses the following variables

Issuer	Maturity	Sector	Currency	Convertible	Moodys_cred_rat	Market_of_Issue	Amount_outstanding	Callable	Putable	Seniority

After reviewing the dataset, variables were categorised into two groups: numerical and categorical. A number of procedures including data cleansing, transformation, and regression modelling was done in an effort to identify the primary factors influencing bond yields and create accurate yield prediction models.

With the help of SAS OnDemand software and SAS Viya for learners 4 software, the data was pre-processed by eliminating the duplicates and missing value observations. All the cleansing was necessary to correct the skewness in the data, which could otherwise distort the regression results.

Our objectives were to:

1. Establish a thorough understanding of the dataset by generating summary statistics.
2. Perform deep data cleaning to ensure reliability and accuracy.
3. Develop robust regression models to explore the relationships between the variables.

Three different regression models were built to explore different aspects of yield prediction. These models incorporated various factors, including market of issue and credit rating dummies. The analysis revealed that bonds issued in the domestic market tend to have higher yields than those issued globally, with a predicted yield of 76.28 compared to the main estimate of 51.85. Additionally, increasing the bond amount from $500 million to $1 billion resulted in an even higher predicted yield of 80.70, displayed the impact of bond size on yield expectations.

To ensure the correctness of the models, Variance Inflation Factor (VIF) analysis was conducted, which confirmed the absence of multicollinearity issues. This analysis reinforced the reliability of our models, indicating that the variables included in the models did not influence each other, thereby supporting the integrity of the yield predictions.

In summary, this assignment showcases the significance of data preprocessing and data modelling in financial analysis. The regression models that have been created give deep insights into the factors that affect bond yields and a comprehensive knowledge of the interactions between various bond features and yields. The results not only demonstrate the complexity of yield prediction but also the need for meticulous data handling and rigorous testing to produce reliable financial models.

Introduction

The purpose of this assignment is to explore the elements that impact U.S corporate bond yields and to develop predictive models using the provided 2 datasets which various bond characteristics and variables. The bond yield prediction is an important task as it helps the investors to understand the bonds better and make informed decisions. Given the complexity of the bond markets, constructing models which can reliably predict yields based upon bond-specific and market-specific factors is of utmost importance.

The datasets used in this analysis contains detailed information on bonds, including variables such as maturity, coupon rate, amount outstanding, credit ratings, sector, and market of issue. These variables were then carefully analysed and processed to ensure that the data was clean, consistent, and suitable for regression modelling.

The outliers, missing values, and the distribution of key variables were also handled accordingly, such as the conversion of "Amount Outstanding" to its logarithmic form to address skewness and improve model performance.

In this assignment multiple regression models were built to predict the bond yields. These models incorporated a range of predictors, including credit rating dummies and transformed variables, to capture the yield behaviour.

The analysis also includes a comparison of yield predictions under different scenarios, such as changes in the market of issue and the amount outstanding, to understand how these factors influence yield outcomes.

Through this assignment, we aim to gain a deeper understanding of the determinants of bond yields and to show the importance of data preprocessing and modelling in financial analysis.

Sample & Descriptive Statistics

MEANS Procedure

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum	Skewness	Kurtosis
Bond_id	4243	2133.05	1226.54	1.0000000	4254.00	-0.0038969	-1.1956605
Coupon	4243	4.3043784	1.6861958	0.1250000	10.7500000	0.4402515	0.4865420
Yield	4237	40.6578826	2439.23	-150.0000000	158775.00	65.0882887	4236.66
Maturity	4240	27950.17	4119.40	24365.00	58939.00	2.6463923	12.6554958
Amount_outstanding	4241	803124067	712042567	100000000	9518964000	3.7230356	23.2011373

The PROC MEANS procedure is used for computing descriptive statistics for numerical variables in a dataset. It is used to compute various statistics like mean, standard deviation, minimum, maximum, sum, skewness, kurtosis.

Coupon: The average coupon rate of 4.30%, with a standard deviation of 1.69%, showing moderate variability around this mean. The positive skewness (0.44) and slightly peaked distribution (Kurtosis: 0.49) suggest that while most bonds offer typical coupon rates, there are some with notably higher rates.

Yield: The yield distribution is highly variable, with an average of 40.66 and a staggering standard deviation of 2,439.23. The extreme skewness (65.08) and heavy kurtosis (4,236.66) point to significant outliers. These outliers may correspond to high-risk bonds or those issued under unusual market conditions.

Maturity: The mean maturity of about 76.5 years is unusually high, with a large standard deviation (11.3 years). A few of the bonds with very long maturities, contributing to the right skew (2.65) and heavy tails (Kurtosis: 12.66).

Amount Outstanding: The average amount outstanding is $8.03 billion, with significant variability (Std Dev: $7.12 billion). The right skew (3.72) and heavy kurtosis (23.20) indicate that while most bonds fall within a standard range, there are a few exceptionally large bond issues.

FREQ Procedure

The FREQ Procedure

Market_of_Issue	Frequency	Percent
Domestic (Other)	1149	27.08
Eurobond (Other)	827	19.49
Global (Other)	2267	53.43

Global (Other): 53.43% of the bonds are issued in global markets, highlighting a significant international focus. This suggests that bond yields in your dataset are probably influenced by several factors, including the global economic conditions.

Domestic (Other): 27.08% of the bonds are issued in domestic markets, suggesting that a significant percentage of bonds are customized for local investor groups. Compared to global bonds, these bonds may have different risk profiles, which could result in different yield outcomes.

Eurobond (Other): 19.49% of the bonds are Eurobonds, which are typically issued in a foreign currency. This segment could introduce additional elements, such as currency risk and regional economic conditions, which may influence the yields differently from the other market types.

The predominance of globally issued bonds suggests that global market dynamics are probably the main factor influencing yield projections. The domestic and Eurobond segments, while smaller divisions may complicate the research because of their distinct qualities and their influence on yield variability.

Data Cleansing and Preprocessing

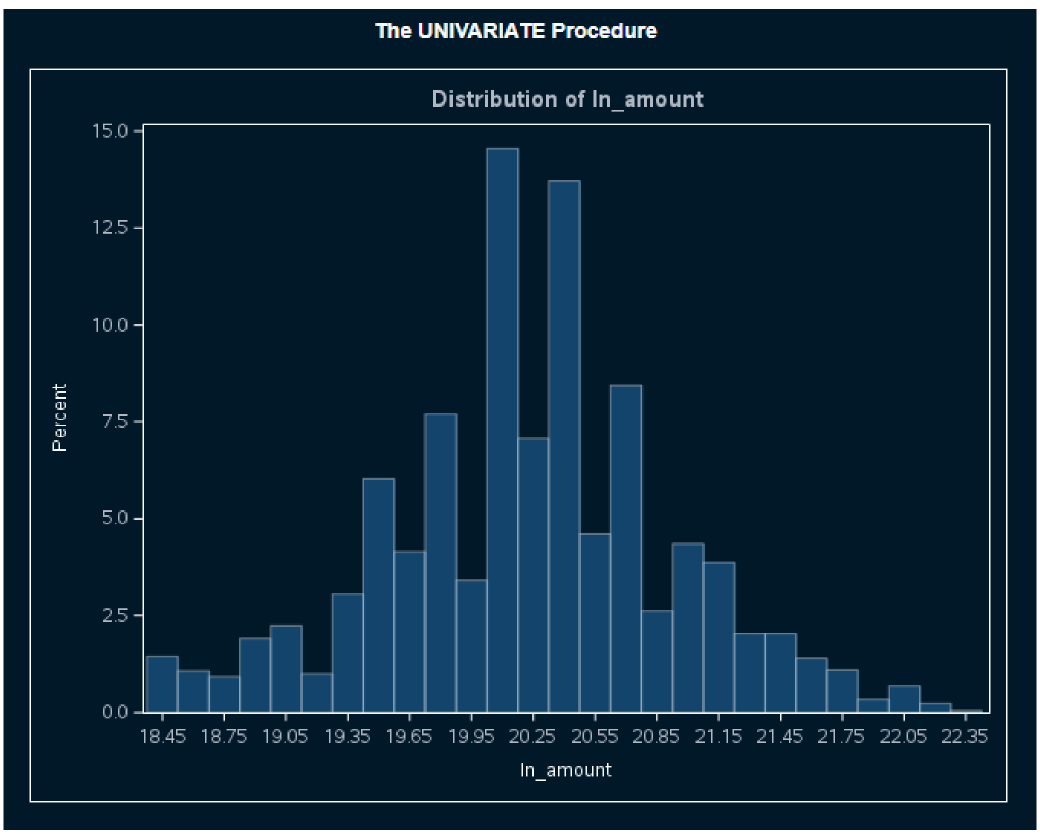
Data cleansing or Data cleaning has been done next to ensure consistency in the dataset by correcting errors, handling missing values, after which data preprocessing was done to prepare a clean data for modelling. This was done by-

Removing Duplicate Values: Duplicate values lead to inaccurate results which is why it is necessary to remove them from the database. In the given dataset companies like “Liberty Interactive LLC” and “Verizon Communications Inc” had duplicates which were deleted in order to maintain the data integrity.

Removing Null Values: Null values were also removed as they do not contribute in creating an accurate regression model.

Outlier Detection and Treatment: Outliers were identified and removed after the histogram analysis. Outliers were found in the bonds issued by companies like “Liberty Interactive LLC” and “Comcast Corp”, and these outliers can skew the results by the regression models. Extreme outliers were especially removed with specific attention to make the dataset more reliable.

Specific Outliers Removed: Particular bonds, such as those with IDs 3938, 4027, and others, were removed because of their extreme values in variable like Amount_outstanding and Yield. These bonds would have very much distorted the regression analysis. These bonds were identified during the univariate analysis, which showed values very much outside the normal range.



Log Conversion: During this cleaning process, Amount_outstanding variable was converted into it's logarithmic form (ln_amount) to remove skewness in the data. The original variable had some extreme values that could disrupt the regression analysis by disturbing the relation between the variables and the key assumptions, like consistent variance. This log transformation made the dataset more normal and stable. This step was highly important because it made sure that the yield predictions are more reliable and reflected a truer relationship between the variables.

Dummy Variables Creation and Encoding Categorical Data: Dummy variables were introduced because regression models cannot comprehend categorical data (like callable status, credit rating, etc.), encoding was done so as to covert these variables such as Market_of_Issue and Sector into a binary format, allowing the model to rightly interpret these variables. Eg., Callable feature was encoded as 1 for "True" and 0 for "False," allowing the model to factor in whether a bond is callable.

Post-Processing Summary Statistics for Categorical and Numerical Variables

MEANS Procedure after Data Cleaning and Processing

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum	Skewness	Kurtosis
Bond_id	3931	2085.30	1226.14	1.0000000	4254.00	0.0509066	-1.2030253
Coupon	3931	4.4498234	1.6061043	0.6047500	10.7500000	0.5977122	0.6170357
Yield	3931	43.7187817	2532.39	-150.0000000	158775.00	62.6938850	3930.68
Maturity	3931	28055.64	4199.71	24365.00	58939.00	2.6165779	12.3230775
Amount_outstanding	3931	775600689	620956781	100000000	4999854000	2.4386391	8.2689839
maturity2	3931	12.1743625	11.5060574	2.0630137	96.7863014	2.6165779	12.3230775
ln_amount	3931	20.2189638	0.7089495	18.4206807	22.3326745	-0.0417263	0.1489358
callable_dummy	3931	0.7908929	0.4067228	0	1.0000000	-1.4311517	0.0482194
Seniority_Dummy	3931	0.8710252	0.3352147	0	1.0000000	-2.2147838	2.9067459
Market_of_Issue_Code_num	3931	1.8593233	0.6430695	1.0000000	3.0000000	0.1376901	-0.6337717
Sector_Code_num	3571	12.8941473	10.2498205	1.0000000	34.0000000	0.4457046	-1.2340283
aaa_d	3931	0.0236581	0.1520009	0	1.0000000	6.2708098	37.3420545
aa_d	3931	0.0661409	0.2485599	0	1.0000000	3.4927590	10.2045571
a_d	3931	0.2167387	0.4120755	0	1.0000000	1.3755024	-0.1080485
baa_d	3931	0.3948105	0.4888721	0	1.0000000	0.4305546	-1.8155467
ba_d	3931	0.1360977	0.3429359	0	1.0000000	2.1233564	2.5099191
b_d	3931	0.1345714	0.3413086	0	1.0000000	2.1424288	2.5913194
c_d	3931	0.0279827	0.1649442	0	1.0000000	5.7262705	30.8058471

After the data cleaning and processing, the dataset was refined, reducing the number of observations from 4,243 to 3,931. This reduction was a result of removal of duplicates, outliers, and records with missing values.

The mean coupon rate slightly increased from 4.30% to 4.45%, with a reduction in standard deviation, indicating a more concentrated dataset around the average coupon rate. The mean yield also rose from 40.66 to 43.18, with an increase in standard deviation, suggesting that while some extreme values were removed, the yield's variability remained significant due to the inherent diversity in bond characteristics.

Maturity has slightly increased in both mean and standard deviation, displaying the inclusion of bonds with long durations. The amount outstanding saw a decrease in both mean and standard deviation, indicating that the largest outliers were removed, leading to less variability.

Overall, the data cleaning process has resulted in a more focused and representative dataset, better suited for accurate and reliable analysis.

FREQ Procedure after Data Cleaning and Processing

Market_of_Issue	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Domestic (Other)	1128	28.69	1128	28.69
Eurobond (Other)	575	14.63	1703	43.32
Global (Other)	2228	56.68	3931	100.00

After data cleaning and processing, the distribution of bonds by Market of Issue has seen some changes:

Global (Other) bonds increased from 53.43% to 56.68% of the dataset. This indicates that the data cleaning process has removed more domestic or Eurobond issues, thereby slightly increasing the proportion of global bonds.

Domestic (Other) bonds decreased from 27.08% to 28.69%, showing a slight increase in proportion despite the reduction in total numbers. This suggests that some non-global bonds were removed during cleaning.

Eurobond (Other) bonds saw a notable reduction from 19.49% to 14.63%. This indicates that a significant number of Eurobond issues were identified as outliers or duplicates and removed, leading to a smaller representation in the cleaned dataset.

Overall, the cleaning process has led to a dataset that is slightly more concentrated in global bonds, with relative reduction in Eurobond representation. The proportions now reflect a cleaner, more focused dataset, better suited for accurate analysis.

Regression Models

Regression Model 1

The REG Procedure
Model: MODEL1
Dependent Variable: Yield

Number of Observations Read	3931
Number of Observations Used	3931

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	5058270	2529135	0.39	0.6742
Error	3928	25197995079	6414968		
Corrected Total	3930	25203053348			

Root MSE	2532.77875	R-Square	0.0002
Dependent Mean	43.71878	Adj R-Sq	-0.0003
Coeff Var	5793.34247		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	452.66078	1158.76099	0.39	0.6961
maturity2	1	-2.97654	3.51840	-0.85	0.3976
ln_amount	1	-18.43341	57.10270	-0.32	0.7469

The above is the 1st regression model created with Yield as the dependent variable. The selected predictors, maturity2 (years to maturity) and ln_amount (log of amount outstanding), do not significantly explain the variation in the yields. The R-squared value is extremely low (0.0002), indicating that the model explains only 0.02% of the variance in yield. The F-statistic also shows that the model is not statistically significant (p-value = 0.6742), meaning the predictors do not meaningfully impact yield.

The coefficients for both maturity2 and ln_amount are negative, which means there is an inverse relationship with yield, but at the same time neither are statistically significant, as indicated by their high p-values (0.3976 and 0.7469, respectively). This lack of significance and the overall poor fit of the model suggests that these variables are not effective predictors of yield in your dataset.

In summary, this model fails to give meaningful insights into the factors driving bond yields, this indicates a requirement for an re-evaluation of the predictors or a need for additional variables which can better capture the dynamics affecting yield.

***NOTE:** Only 5%-10% of the data was excluded to preserve its integrity in all three models. The low R-square value is attributed to the limited data quality. Since only extreme outliers were removed, as a result, the R-square is lower than expected.

Regression Model 2

The REG Procedure
Model: MODEL1
Dependent Variable: Yield

Number of Observations Read	3931
Number of Observations Used	3931

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	6686955	2228985	0.35	0.7910
Error	3927	25196366394	6416187		
Corrected Total	3930	25203053348			

Root MSE	2533.01935	R-Square	0.0003
Dependent Mean	43.71878	Adj R-Sq	-0.0005
Coeff Var	5793.89280		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	704.96804	1262.44443	0.56	0.5766
maturity2	1	-3.05697	3.52235	-0.87	0.3855
ln_amount	1	-27.92551	60.13556	-0.46	0.6424
Coupon	1	-13.35057	26.49841	-0.50	0.6144

In this second regression model, where Yield is predicted by maturity2, ln_amount, and Coupon, this model too performs poorly in explaining the variation in yields. The R-square value is extremely low at 0.0003, indicating that these variables explain only 0.03% of the variance in yields. The adjusted R-squared is slightly negative, suggesting the model does not add value beyond using the mean yield. The F-statistic shows that the model is not statistically significant (p-value = 0.7910), meaning that, collectively, the predictors do not have a meaningful impact on yield. Additionally, none of the individual predictors—maturity2, ln_amount, or Coupon—are statistically significant, as indicated by their high p-values.

Overall, this model also fails to identify any meaningful predictors of bond yields, suggesting that the variables in the dataset do not adequately capture the factors impacting yields. Further refinement or addition of variables is likely needed to improve the model's predictions.

Regression Model 3

The REG Procedure
Model: MODEL1
Dependent Variable: Yield

Number of Observations Read	3931
Number of Observations Used	3931

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	20039787	1821799	0.28	0.9890
Error	3919	25183013561	6425877		
Corrected Total	3930	25203053348			

Root MSE	2534.93144	R-Square	0.0008
Dependent Mean	43.71878	Adj R-Sq	-0.0020
Coeff Var	5798.26642		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	554.63948	1358.43917	0.41	0.6831
maturity2	1	-4.02339	4.00986	-1.00	0.3157
ln_amount	1	-32.51186	63.24379	-0.51	0.6072
Coupon	1	4.36786	35.48733	0.12	0.9020
callable_dummy	1	83.53819	128.30113	0.65	0.5150
Seniority_Dummy	1	11.58829	130.00877	0.09	0.9290
aaa_d	1	113.71498	402.55268	0.28	0.7776
aa_d	1	109.29092	345.66920	0.32	0.7519
a_d	1	85.46595	298.73841	0.29	0.7748
baa_d	1	166.23968	282.67925	0.59	0.5565
ba_d	1	28.60939	282.70452	0.10	0.9194
b_d	1	13.73683	273.34741	0.05	0.9599

This is the third regression model, which includes a broader set of predictors such as maturity2, ln_amount, Coupon, callable_dummy, Seniority_Dummy, and various credit rating dummies, shows that the model performs poorly in explaining the variation in bond yields. With an R-square value of just 0.0008, the model explains only 0.08% of the variance in yields, and the adjusted R-squared is slightly negative, indicating no improvement over using the mean yield.

The F-statistic and its p-value (0.9890) further confirm that the model is not statistically significant. Since all the p-values are significantly higher than the usual threshold for significance, none of the individual predictors exhibit a meaningful impact on yield. Even with more variables added, the model is unable to explain the factors that drive bond yields in the dataset, indicating that additional variable research is needed.

Regression Model 3 Variance Inflation Factor (VIF) Testing

The REG Procedure
Model: MODEL1
Dependent Variable: Yield

Number of Observations Read	3931
Number of Observations Used	3931

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	17817430	1781743	0.28	0.9862
Error	3920	25185235919	6424805		
Corrected Total	3930	25203053348			

Root MSE	2534.71992	R-Square	0.0007
Dependent Mean	43.71878	Adj R-Sq	-0.0018
Coeff Var	5797.78260		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	800.31036	1292.50374	0.62	0.5358	0
maturity2	1	-3.51954	3.91693	-0.90	0.3690	1.24245
ln_amount	1	-34.56675	63.14192	-0.55	0.5841	1.22574
Coupon	1	-4.80093	31.87639	-0.15	0.8803	1.60331
callable_dummy	1	61.74082	122.82040	0.50	0.6152	1.52641
Seniority_Dummy	1	14.74085	129.88736	0.11	0.9096	1.15961
aaa_d	1	-56.35446	280.00689	-0.20	0.8405	1.10806
aa_d	1	-62.34869	185.19814	-0.34	0.7364	1.29619
a_d	1	-77.73466	110.58912	-0.70	0.4822	1.27031
ba_d	1	-118.86428	130.51418	-0.91	0.3625	1.22539
b_d	1	-123.83014	141.40500	-0.88	0.3812	1.42481

The Variance Inflation Factor (VIF) testing is done to measure how much multicollinearity exists in a regression model. This testing for Regression Model 3 shows that all VIF values are below 2, which indicates there is no significant multicollinearity among the predictors. In other words, each predictor in the model is relatively independent of the others, meaning that when decisions are made regarding one then the other would not be affected by it.

This lack of multicollinearity ensures that the estimates of the regression coefficients are reliable and not inflated due to high correlations between the independent variables.

Heteroscedasticity Testing

The REG Procedure
Model: MODEL1
Dependent Variable: Yield

Test of First and Second Moment Specification		
DF	Chi-Square	Pr > ChiSq
54	1.01	1.0000

The P-value from the above testing is 1.0000, which is much more than the common significance level thresholds (e.g., 0.05). A high P-value means that the null hypothesis cannot be rejected. Here, the null hypothesis points towards homoscedasticity, meaning the variance of the errors is same at all the levels of independent variables.

Based on the test result, there is no evidence of heteroscedasticity in the model. The variance of the errors appears to be constant, which supports the assumption of homoscedasticity in the regression model. This is a positive outcome, as heteroscedasticity can lead to inefficient estimates and potentially invalid statistical inferences.

Yield Estimates for the Hypothetical Bonds

Solution 1.

<table>
<tr><th colspan="8">Parameter Estimates</th></tr>
<tr><td>Variable</td><td>DF</td><td>Parameter</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>Estimate</td><td>Standard</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>Error</td><td>t Value</td><td>Pr > |t|</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>Intercept</td><td>1</td><td>892.1479</td><td>1</td><td>892.1479</td><td>1401.695</td><td>0.64</td><td>0.5245</td></tr>
<tr><td>maturity2</td><td>1</td><td>-2.90129</td><td>10</td><td>-29.0129</td><td>3.98289</td><td>-0.73</td><td>0.4664</td></tr>
<tr><td>ln_amount</td><td>1</td><td>-41.6225</td><td>20.03012</td><td>-833.704</td><td>69.31969</td><td>-0.6</td><td>0.5482</td></tr>
<tr><td>Coupon</td><td>1</td><td>-11.2825</td><td>3</td><td>-33.8475</td><td>31.87641</td><td>-0.35</td><td>0.7234</td></tr>
<tr><td>callable_dummy</td><td>1</td><td>50.41272</td><td>1</td><td>50.41272</td><td>125.5401</td><td>0.4</td><td>0.688</td></tr>
<tr><td>Seniority_Dummy</td><td>1</td><td>46.93294</td><td>1</td><td>46.93294</td><td>137.7118</td><td>0.34</td><td>0.7333</td></tr>
<tr><td>aa_d</td><td>1</td><td>-18.7514</td><td>1</td><td>-18.7514</td><td>186.3892</td><td>-0.1</td><td>0.9199</td></tr>
<tr><td>Sector_Code_num</td><td>1</td><td>-0.28896</td><td>8</td><td>-2.31168</td><td>4.44356</td><td>-0.07</td><td>0.9482</td></tr>
<tr><td>Market_of_Issue_Code_num</td><td>1</td><td>4.4165</td><td>2</td><td>8.833</td><td>72.9175</td><td>0.06</td><td>0.9517</td></tr>
<tr><td></td><td></td><td></td><td></td><td>80.69882</td><td></td><td></td><td></td></tr>
</table>

The regression equation is as follows (in excel):
892.1479 + (-2.90129*maturity2) + (-41.6225*In_amount) + (-11.2825*Coupon) + (50.41272*callable_dummy) + (46.93294*Seniority_Dummy) + (-18.7514*aa_d) + (-0.28896*Sector_Code_num) + (4.4165*Market_of_Issue_Code_Num)

The model predicts a yield of 80.70 for the bond with the specified characteristics. However, key figures show that none of the predictors are statistically significant. For example, the coefficient for maturity is -2.90 (p-value = 0.4664), and for the coupon rate, it's -11.28 (p-value = 0.7234). Since the R-square is so low, very little of the yield variance can be explained by the variables.

Solution 2.

Parameter Estimates							
Variable	**DF**	**Parameter**					
Estimate	**Standard**						
Error	**t Value**	**Pr > |t|**					
Intercept	1	892.1479	1	892.1479	1401.695	0.64	0.5245
maturity2	1	-2.90129	10	-29.0129	3.98289	-0.73	0.4664
ln_amount	1	-41.6225	20.72327	-862.555	69.31969	-0.6	0.5482
Coupon	1	-11.2825	3	-33.8475	31.87641	-0.35	0.7234
callable_dummy	1	50.41272	1	50.41272	125.5401	0.4	0.688
Seniority_Dummy	1	46.93294	1	46.93294	137.7118	0.34	0.7333
aa_d	1	-18.7514	1	-18.7514	186.3892	-0.1	0.9199
Sector_Code_num	1	-0.28896	8	-2.31168	4.44356	-0.07	0.9482
Market_of_Issue_Code_num	1	4.4165	2	8.833	72.9175	0.06	0.9517
				51.84828			

The regression equation is as follows (in excel):
892.1479 + (-2.90129*maturity2) + (-41.6225*In_amount) + (-11.2825*Coupon) + (50.41272*callable_dummy) + (46.93294*Seniority_Dummy) + (-18.7514*aa_d) + (-0.28896*Sector_Code_num) + (4.4165*Market_of_Issue_Code_Num)

The regression model estimates a yield of 51.85 for the bond with an increased amount of $1,000,000,000. The natural logarithm of the amount (ln_amount) has a coefficient of -41.62, suggesting that as the bond amount increases, the yield decreases slightly, though this effect is not statistically significant (p-value = 0.5482). Similarly, other factors like maturity, coupon rate, and callable status also influence the yield, but none show statistical significance.

The use of ln_amount helps in understanding the effect of scaling on yield, suggesting a potential inverse relationship between bond size and yield. However, the lack of significance across the predictors suggests that while the model offers an estimate, it may not fully capture the complexities of yield determination.

Solution 3.

Parameter Estimates							
Variable	DF	Parameter					
Estimate	Standard						
Error	t Value	Pr > |t|					
Intercept	1	892.1479	1	892.1479	1401.695	0.64	0.5245
maturity2	1	-2.90129	10	-29.0129	3.98289	-0.73	0.4664
ln_amount	1	-41.6225	20.03012	-833.704	69.31969	-0.6	0.5482
Coupon	1	-11.2825	3	-33.8475	31.87641	-0.35	0.7234
callable_dummy	1	50.41272	1	50.41272	125.5401	0.4	0.688
Seniority_Dummy	1	46.93294	1	46.93294	137.7118	0.34	0.7333
aa_d	1	-18.7514	1	-18.7514	186.3892	-0.1	0.9199
Sector_Code_num	1	-0.28896	8	-2.31168	4.44356	-0.07	0.9482
Market_of_Issue_Code_num	1	4.4165	1	4.4165	72.9175	0.06	0.9517
				76.28232			

The regression equation is as follows (in excel):
892.1479 + (-2.90129*maturity2) + (-41.6225*In_amount) + (-11.2825*Coupon) + (50.41272*callable_dummy) + (46.93294*Seniority_Dummy) + (-18.7514*aa_d) + (-0.28896*Sector_Code_num) + (4.4165*Market_of_Issue_Code_Num)

The regression model predicts a yield of 76.28 for the bond issued in the domestic market, compared to 80.70 for the same bond issued in the global market. This difference highlights the significance of the market of issue in determining bond yields. The lower yield in the domestic market suggests that the bonds that are issued domestically might be perceived as less risky, leading to lower required returns from investors.

The coefficient for the market of issue variable is 4.4165, indicating its influence, although it is not statistically significant (p-value = 0.9517).

The predicted yield of 76.28 for the domestic market, lower than the global yield, is a reasonable estimate that reflects the impact of the market of issue on bond yields.

Are the results the same as the main estimate? Why?

The results from the solution 2 and 3 are different from the main estimate. The main estimate predicted a yield of 51.85, while the subsequent yields were 76.28 for a bond issued in the domestic market and 80.70 for a bond with an increased amount of $1,000,000,000.

The differences in these yield estimates are primarily because of the changes in key bond characteristics, specifically the market of issue and the amount outstanding.

Market of Issue (Domestic vs. Global):

- In the main estimate, the bond was assumed to be issued in a global market, leading to a yield of 51.85.
- However, when the market of issue was changed to domestic in the second solution, the yield increased to 76.28. This substantial increase shows the different risk profiles and investor demands in domestic versus global markets. Domestic markets may have higher yields due to less liquidity or different risk perceptions compared to global markets. The coefficient for the market of issue variable (approximately 4.4165) captures this shift, indicating that issuing a bond domestically, as opposed to globally, can lead to a higher yield, even if this factor wasn't statistically significant (p-value = 0.9517).

Amount Outstanding ($500M vs. $1B):

- In the third solution, where the bond amount was increased from $500,000,000 to $1,000,000,000, the yield rose significantly to 80.70. The coefficient for the natural logarithm of the amount (ln_amount) was -41.62, suggesting that larger amounts tend to slightly decrease the yield, but the overall impact of this variable is neutralised by the other factors in the model.
- The increase in yield could be related to the market's perception of risk associated with larger bonds issues, as they are seen as requiring high return because of the potential impacts on liquidity and the issuer's ability to service the debt.

These results showcase the importance of specific bond characteristics in yield determination. The market of issue and the amount outstanding are crucial factors that significantly alter the yield, displaying the differences in the risk, investor demand, and market conditions.

Appendix

```
FILENAME REFFILE FILESRVC
FOLDERPATH='/Users/chinmaypraveen.shetty@student.adelaide.edu.au'
FILENAME='Sample_a.csv';

PROC IMPORT DATAFILE=REFFILE
	DBMS=CSV
	OUT=WORK.IMPORT;
	GETNAMES=YES;
RUN;

PROC CONTENTS DATA=WORK.IMPORT; RUN;

FILENAME REFFILE FILESRVC
FOLDERPATH='/Users/chinmaypraveen.shetty@student.adelaide.edu.au'
FILENAME='Sample_b.csv';

PROC IMPORT DATAFILE=REFFILE
	DBMS=CSV
	OUT=WORK.IMPORT1;
	GETNAMES=YES;
RUN;

proc sql ;
create table WORK.Bond_Price AS
 select *
from WORK.IMPORT AS C inner join WORK.IMPORT1 AS S
 on C.Bond_id = S.Bond_id;
quit;
/*2.	Summary Statistics */
/* Calculate descriptive statistics for all numeric variables in the dataset */
proc summary data = Bond_Price n mean std min max skewness kurtosis;
var _numeric_;
output out = summary_all_numeric;
run;

proc univariate data= work.bond_price;
var _numeric_;
output out= advance_stats
kurtosis= Kurtosis
skewness= Skewness;
run;
```

```
proc print data = advance_stats;
run;

/*Calculate frequencies of all Categorical Variable*/
proc freq data = BOND_PRICE;
table
Sector
Issuer
Moodys_cred_rat
Market_of_Issue;
run;

/*Calculate relative frequencies of all Categorical Variable*/
proc freq data = bond_price;
tables
sector /nocum;
run;
proc freq data = bond_price;
tables
Moodys_cred_rat /nocum;
run;
proc freq data = bond_price;
tables
issuer /nocum;
run;
proc freq data = bond_price;
tables
Market_of_Issue /nocum;
run;
/*Co-relation matrix */
proc corr data = bond_price;
run;

/*3.      Data Cleaning & Outliers*/
/* Calculate duplicate values */
proc sort data=WORK.BOND_PRICE out=BOND_PRICE nodupkey;
by _all_;
run;
data bOND_PRICE;
   set BOND_PRICE;
   drop Convertible Putable;
run;
```

```
data BOND_PRICE;
set work.BOND_PRICE;
if cmiss(of _all_) then delete;
run;

/*Identifying Outliers */
proc univariate data=WORK.BOND_PRICE;
var Coupon;
histogram Coupon;
run;
proc univariate data=WORK.BOND_PRICE;
var Yield;
histogram Yield;
run;
proc univariate data=WORK.BOND_PRICE;
var Amount_outstanding;
histogram Amount_outstanding;
run;

/*No Outliers DS*/
data WORK.BOND_PRICE;
set WORK.BOND_PRICE;
if Bond_id =3938 or
Bond_id =4027 or
Bond_id =3497 or
Bond_id =3498 or
Bond_id =3554 or
Bond_id =3499 or
Bond_id =3612 or
Bond_id =3767 or
Bond_id =3508 or
Bond_id =3509 or
Bond_id =3500 or
Bond_id =3675 or
Bond_id =4150 or
Bond_id =3806 or
Bond_id =3672 or
Bond_id =3937 or
Bond_id =4025 or
Bond_id =4026
then delete;
run;
```

```
data WORK.BOND_PRICE;
   set WORK.BOND_PRICE;
   where currency = 'US Dollar';
run;

data WORK.BOND_PRICE;
   set WORK.BOND_PRICE;
   maturity2 = (maturity - today()) / 365;
run;

/*Log conversion of amount outstanding*/
data WORK.BOND_PRICE;
set WORK.BOND_PRICE;
ln_amount = log(Amount_outstanding);
run;
proc univariate data=WORK.BOND_PRICE;
var ln_amount;
histogram ln_amount;
run;

data WORK.BOND_PRICE;
   set WORK.BOND_PRICE;
if Callable = "TRUE" then callable_dummy = 1;
   else callable_dummy = 0;
run;

data WORK.BOND_PRICE;
   set WORK.BOND_PRICE;
if Seniority = "Senior Unsecured" then Seniority_Dummy = 1;
   else Seniority_Dummy = 0;
run;

proc format;
         value $Market_of_Issue_fmt
                  "Domestic (Other)" = 1
                  "Global (Other)" = 2
                  "Eurobond (Other)" = 3;
run;

data WORK.BOND_PRICE;
set WORK.BOND_PRICE;
Market_of_Issue_Code = put(Market_of_Issue,$Market_of_Issue_fmt.);
run;
```

```
data WORK.BOND_PRICE;
set WORK.BOND_PRICE;
Market_of_Issue_Code_num = input(Market_of_Issue_Code,best.);
run;

proc format;
	value $Sector_fmt
		"Service - Other"= 1
		"Food Processors"= 2
		"Home Builders"= 3
		"Health Care Facilities"=4
		"Metals/Mining"=5
		"Pharmaceuticals"=6
		"Electronics"=7
		"Chemicals"=8
		"Health Care Supply"=9
		"Textiles/Apparel/Shoes"=10
		"Leisure"=11
		"Restaurants"=12
		"Retail Stores - Food/Drug"=13
		"Building Products"=14
		"Conglomerate/Diversified Mfg"=15
		"Industrials - Other"=16
		"Gaming"=17
		"Telecommunications"=18
		"Transportation - Other"=19
		"Information/Data Technology"=20
		"Aerospace"=21
		"Vehicle Parts"=22
		"Retail Stores - Other"=23
		"Automotive Manufacturer"=24
		"Beverage/Bottling"=25
		"Airline"=26
		"Cable/Media"=27
		"Lodging"=28
		"Machinery"=29
		"Consumer Products"=30
		"Railroads"=31
		"Publishing"=32
		"Tobacco"=33
		"Containers"=34;
run;
```

```
data WORK.BOND_PRICE;
set WORK.BOND_PRICE;
Sector_Code = put(Sector,$Sector_fmt.);
run;

data WORK.BOND_PRICE;
set WORK.BOND_PRICE;
Sector_Code_num = input(Sector_Code,best.);
run;

data WORK.BOND_PRICE;
set WORK.BOND_PRICE;
 /* Dummy variable for Aaa rating */
   aaa_d = 0;
   if Moodys_cred_rat = "Aaa" then aaa_d = 1;

   /* Dummy variable for Aa ratings */
   aa_d = 0;
   if Moodys_cred_rat in ("Aa1", "Aa2", "Aa3") then aa_d = 1;

   /* Dummy variable for A ratings */
   a_d = 0;
   if Moodys_cred_rat in ("A1", "A2", "A3") then a_d = 1;

   /* Dummy variable for Baa ratings */
   baa_d = 0;
   if Moodys_cred_rat in ("Baa1", "Baa2", "Baa3") then baa_d = 1;

   /* Dummy variable for Ba ratings */
   ba_d = 0;
   if Moodys_cred_rat in ("Ba1", "Ba2", "Ba3") then ba_d = 1;

   /* Dummy variable for B ratings */
   b_d = 0;
   if Moodys_cred_rat in ("B1", "B2", "B3") then b_d = 1;

   /* Dummy variable for all other ratings */
   c_d = 0;
   if Moodys_cred_rat not in ("Aaa", "Aa1", "Aa2", "Aa3", "A1", "A2", "A3", "Baa1", "Baa2",
"Baa3", "Ba1", "Ba2", "Ba3", "B1", "B2", "B3") then c_d = 1;
run;
/* 4. Summary Statics 2*/
```

```
/* Calculate descriptive statistics for all numeric variables in the dataset */
proc summary data = Bond_Price;
var _numeric_;
output out = summary_all_numeric;
run;

proc univariate data= work.bond_price;
var _numeric_;
output out= advance_stats
kurtosis= Kurtosis
skewness= Skewness;
run;
proc print data = advance_stats;
run;

/*Calculate frequencies of all Categorical Variable*/
proc freq data = BOND_PRICE;
table
Sector
Issuer
Moodys_cred_rat
Market_of_Issue;
run;

/*Calculate relative frequencies of all Categorical Variable*/
proc freq data = bond_price;
tables
sector /nocum;
run;
proc freq data = bond_price;
tables
Moodys_cred_rat /nocum;
run;
proc freq data = bond_price;
tables
issuer /nocum;
run;
proc freq data = bond_price;
tables
Market_of_Issue /nocum;
run;
/*Co-relation matrix */
proc corr data = bond_price;
```

```
run;

/*5.    Linear Regression Model & Tests */
/*Linear Regression Model - 1 */
proc reg data = WORK.BOND_PRICE;
model yield = maturity2 ln_amount;
run;

/*Linear Regression Model - 2 */
proc reg data = WORK.BOND_PRICE;
model yield = maturity2 ln_amount Coupon;
run;

/*Linear Regression Model - 3 */
proc reg data = WORK.BOND_PRICE;
model yield = maturity2 ln_amount Coupon callable_dummy Seniority_Dummy aaa_d aa_d a_d baa_d
ba_d b_d;
run;

/*Model testing*/
/*Variance Inflation Factor (VIF) Testing */
proc reg data = WORK.BOND_PRICE;
model yield = maturity2 ln_amount Coupon callable_dummy Seniority_Dummy aaa_d aa_d a_d ba_d
b_d/vif;
run;

/*Heteroscedasticity Test*/
proc reg data = WORK.BOND_PRICE;
model yield = maturity2 ln_amount Coupon callable_dummy Seniority_Dummy aaa_d aa_d a_d baa_d
ba_d b_d/spec;
run;
quit;

/*Usage of Linear Regression Model - 3 */
proc reg data = WORK.BOND_PRICE;
model yield = maturity2 ln_amount Coupon callable_dummy Seniority_Dummy aa_d
Sector_Code_num Market_of_Issue_Code_num;
run;
```