****

**EDGAR DOCUMENTATION**

Edgar Securities Data Extraction Model

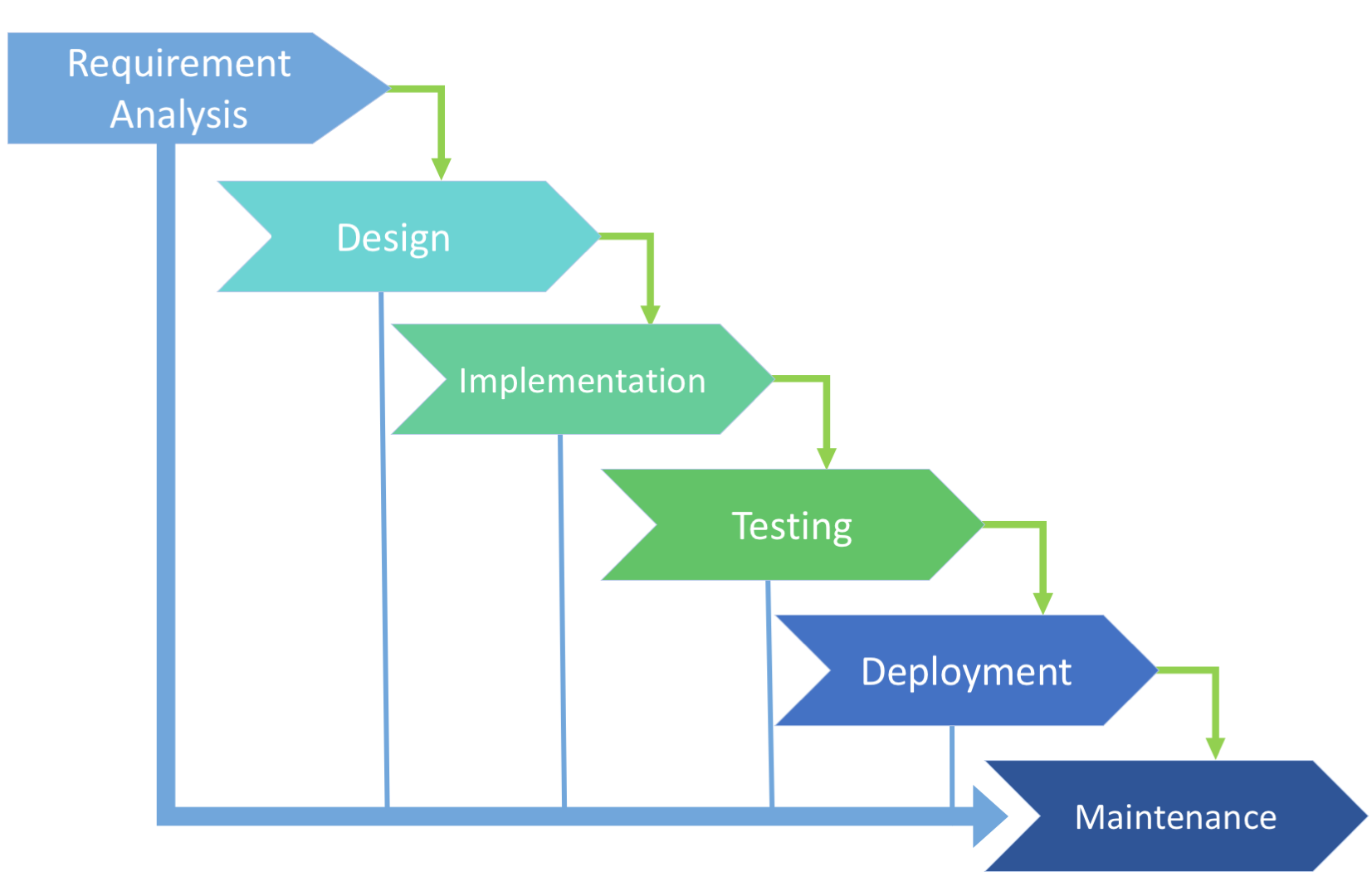
**CONTENT**

|  |  |  |
| --- | --- | --- |
|  |  | **Page** |
| 1. | Edgar Project Over View | 3 |
| 2. | Edgar Securities Over View | 5 |
| 3. | Taxonomy and Model Terms and Definition | 6 |
| 4. | Code Overview | 19 |
| 5. | Issuer Code Overview | 31 |
| 6. | Reconciliation Code Overview | 33 |
| 7. | Basic Commands | 36 |
| 8.  9.  10. | Data Dictionary  Model Installation Process and Requirements  Sanity Check | 42 |

1. **EDGAR NLP PROJECT OVERVIEW**

The purpose of the project was to have a framework which is capable enough to extract the key features of an unstructured document based on predefined taxonomy terms to reduce the cost of purchase from the vendor, and able to provide results with a minimum time lag. This exercise will not only try to reduce the cost but also help the research team to take decisions effectively by providing features in lesser time for the analysis.

For the first Phase of implementation a standardized generic framework was created which can be implemented to another prospectus apart from 424B type with minimum modifications. Project was segmented into multiple decision cycle to formulate the problem.



The Entire Edgar project based on the Natural Language Processing technique which is an automatic manipulation of a language by a machine and opens new ways for machines to analyze, understand, and derive meaning from language in an intelligent and useful way. By utilizing NLP, we can organize and structure the knowledge to perform multiple tasks such as named entity recognition, translation, relationship extraction, sentiment analysis, and subject classification.

Here we used **SpaCy’s** as NLP framework and trained this statistical NLP model to extract linguistic features such as part-of-speech tags, dependency labels, named entities and using other techniques of words or sentence tokenizing and rule-based matcher. Processing the raw text intelligently is difficult, as most words are rare, and it's common for words that look completely different to mean almost the same thing and here SpaCy helps to understand the right meaning of a word using its corpus and some of the Data preprocessing techniques has been used to clean the data for Spacy usability

The project is categorized into three major components which form a pipeline i.e. output of one will be the input of other

As of now, the model was trained on 28 different taxonomy terms with an average processing time of 2-3 minutes per document. Results of the model were tested against vendor’s data and able to extract information with an average coverage of 78.7% and accuracy of 89.7%, results based on the test performed on 100 different records into three different phases. Still, the model is getting trained to improve its coverage and accuracy by improving Named Entity Recognition (NER) and terms identification.

1. **EDGAR SECURITIES**

**EDGAR**, the **Electronic Data Gathering, Analysis, and Retrieval** system, performs automated collection, validation, indexing, acceptance, and forwarding of submissions by companies and others who are required by law to file forms with the [U.S. Securities and Exchange Commission](https://en.wikipedia.org/wiki/U.S._Securities_and_Exchange_Commission) (the "SEC"). The database contains a wealth of information about the Commission and the securities industry which is freely available to the public via the Internet

The objective of the SEC is to protect investors; maintain fair, orderly, and efficient markets; and facilitate capital formation. The SEC strives to promote a market environment that is worthy of the public's trust.

Model use this freely available platform to extract the required documents (424B) for the trading perspective. All links to the platforms are mentioned below:

**Main Website Page:** <https://www.sec.gov/about.shtml>

**Company Filling Search:** <https://www.sec.gov/edgar/searchedgar/companysearch.html>

**Daily Filling Search:** <https://www.sec.gov/edgar/searchedgar/currentevents.htm>

**CIK Lookup:** <https://www.sec.gov/edgar/searchedgar/cik.htm>

**Today’s Filling:** <https://www.sec.gov/cgi-bin/current?q1=0&q2=0&q3=424B>

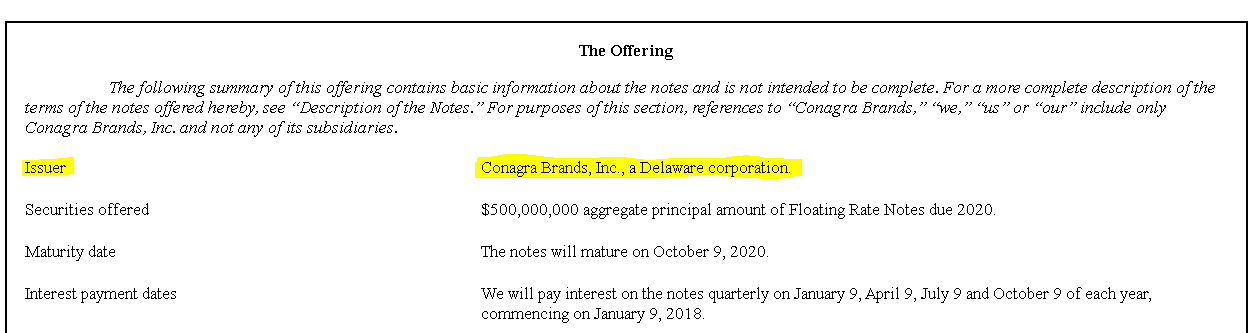
* 1. **TAXONOMY AND MODEL TERMS AND DEFINITIONS**

Taxonomy is an arrangement or division and *nomos* meaning law. Taxonomies act as dictionaries of business terms and their corresponding tags. Taxonomies can be thought as collective information about the structure of a document in terms of its content.

In this project, model has extracted information for about 25 taxonomies terms of 424B type prospectus document.

1. **Issuer Name**

An issuer is a **legal entity** that develops, registers and sells securities to finance its operations. Issuers may be corporations, investment trusts, or domestic or foreign governments. Sometime same issuer can also act as guarantor for the securities. One example related to the issuer is shared below

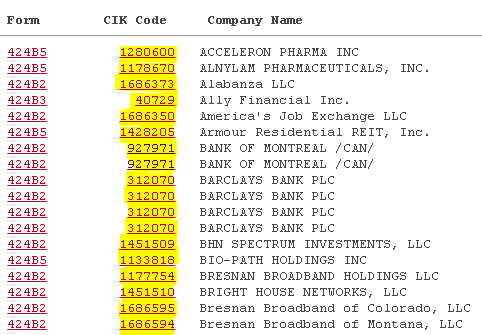


Source: <https://www.sec.gov/Archives/edgar/data/23217/000119312517307880/d466260d424b5.htm>

In this Security Conagra Brands, Inc. is the Issuer and it’s a sister corporation of the Delaware Corporation

1. **CIK NUMBER**

A **Central Index Key** or CIK number is a **number** given to an individual or company by the United States Securities and Exchange Commission. The **number** is used to identify the filings of a company, person, or entity in several online databases, including EDGAR.

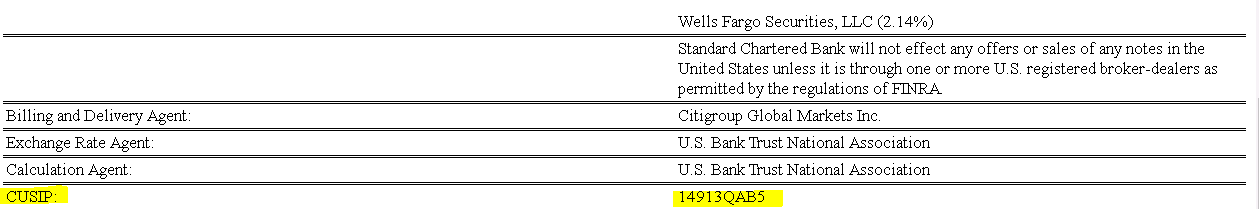


Source: <https://www.sec.gov/cgi-bin/current?q1=0&q2=0&q3=424B>

1. **CUSIP Number:**

A CUSIP is a nine-character [alphanumeric](https://en.wikipedia.org/wiki/Alphanumeric) code that identifies a [North American](https://en.wikipedia.org/wiki/North_America) financial [security](https://en.wikipedia.org/wiki/Security_(finance)) for the purposes of facilitating [clearing](https://en.wikipedia.org/wiki/Clearing_(finance)) and [settlement](https://en.wikipedia.org/wiki/Settlement_(finance)) of trades. The CUSIP was adopted as an American National Standard under Accredited Standards X9.6.

In the case of Apple, Inc. (trading symbol AAPL), for example, a CUSIP of 037833100 has been assigned to identify this stock.



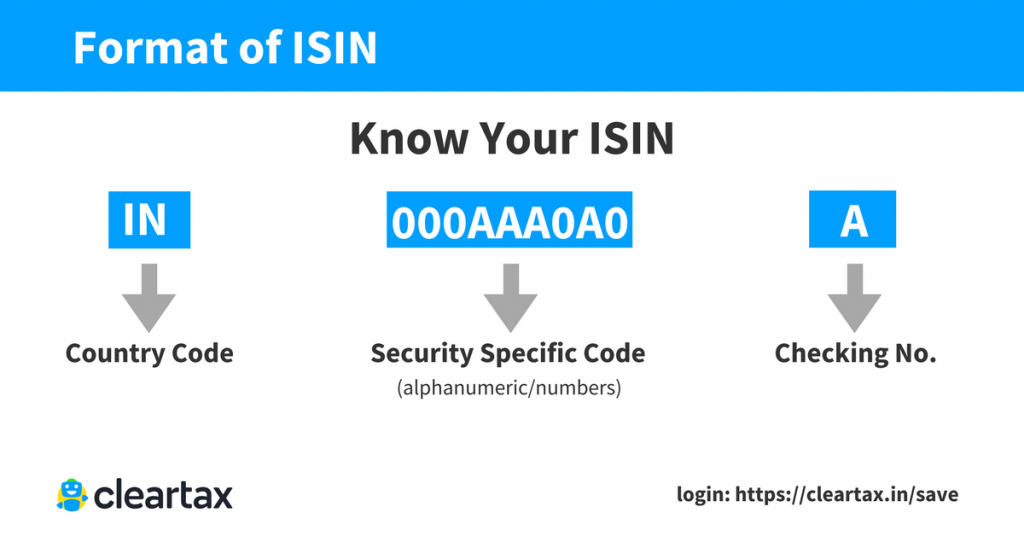
Source: <https://www.sec.gov/Archives/edgar/data/764764/000076476417000091/pricingsuppjune2017mtnfloa.htm>

1. **ISIN Number:**

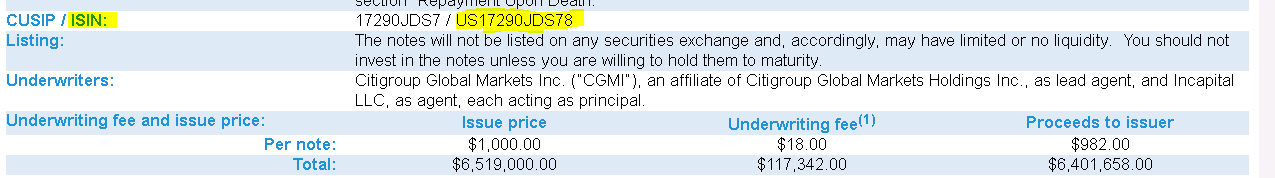
An International Securities Identification Number (ISIN) uniquely identifies a [security](https://en.wikipedia.org/wiki/Security_(finance)). Its structure is defined in [ISO 6166](https://en.wikipedia.org/wiki/ISO_6166).

All ISIN codes have 12 characters, and this code is made up of both letters and numbers. This is how the numbering code is broken down:

* First two digits refer to the country in which the issuing company is headquartered
* The middle nine digits link back to the specific security and act as a unique identifier or CUSIP Number
* The final character called the ‘check digit’ acts as a check and prevents forgery



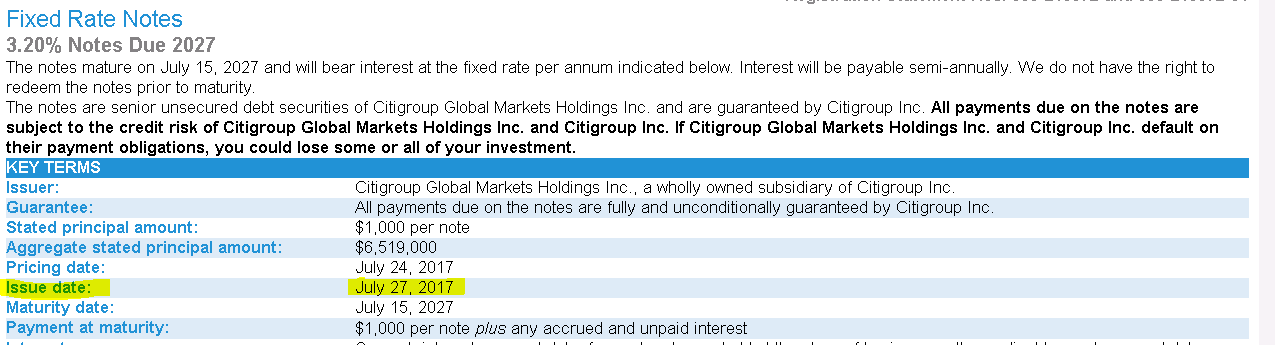
Source: https://cleartax.in/s/isin-code



Source: <https://www.sec.gov/Archives/edgar/data/200245/000095010317007122/dp78715_424b2-1139.htm>

1. **Issue Date:**

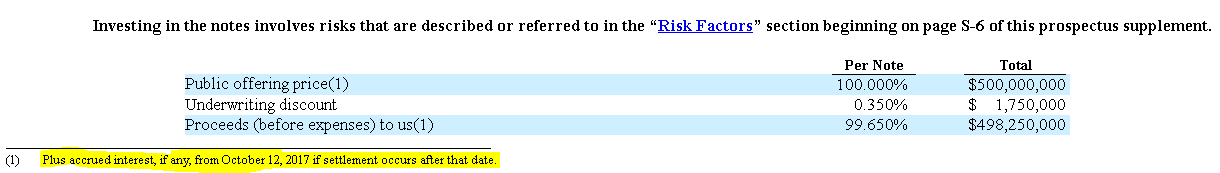
Date on which a security is issued. In case of a bond, the date from which a bond-holder is entitled to receive interest irrespective of the date the bond was purchased or delivered.



Source: <https://www.sec.gov/Archives/edgar/data/200245/000095010317007122/dp78715_424b2-1139.htm>

1. **Interest Accrual Date:**

For new issue fixed-income offerings (e.g., Notes and bonds), this is the date on which coupon interest will begin to accrue. For secondary market fixed-income offerings (e.g., bonds), this is generally the date the last coupon interest payment was paid but we are not using the definition related to secondary market fixed income offerings for our model.



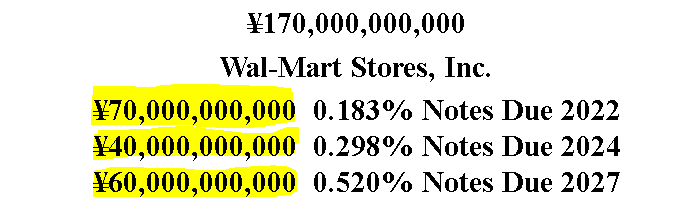
Source: https://www.sec.gov/Archives/edgar/data/23217/000119312517307880/d466260d424b5.htm

1. **Original Face Amount:**

Original Face Amount is also known as Face value/Principal Amount. Face Value is the [nominal value](https://www.investopedia.com/terms/n/nominalvalue.asp) or dollar value of a security stated by the issuer.

For stocks, it is the [original cost](https://www.investopedia.com/terms/o/originalcost.asp) of the stock shown on the certificate.

For bonds, it is the amount paid to the holder at maturity, generally $1,000. It is also known as "par value" or simply "par."

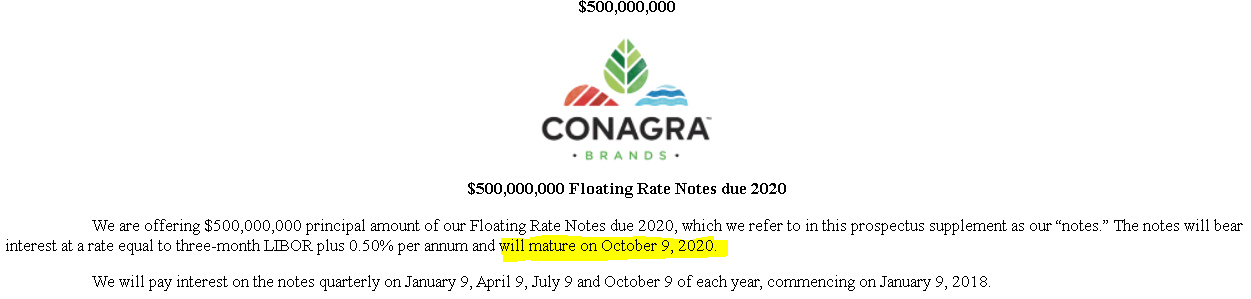


Source: https://www.sec.gov/Archives/edgar/data/104169/000119312517224293/d418506d424b5.htm

Here 70 Billion, 40 Billion and 60 Billion Japanese Yen are the face value for the securities issued by Walmart Stores, Inc. and due in 2022, 2024 and 2027 respectively.

1. **Maturity Date:**

The maturity date is the date on which the principal amount/Face Value of a note, draft, acceptance bond or another [debt instrument](https://www.investopedia.com/terms/d/debtinstrument.asp) becomes due and is repaid to the investor and interest payments stop. It is also the termination or due date on which an installment loan must be paid in full to buyers.

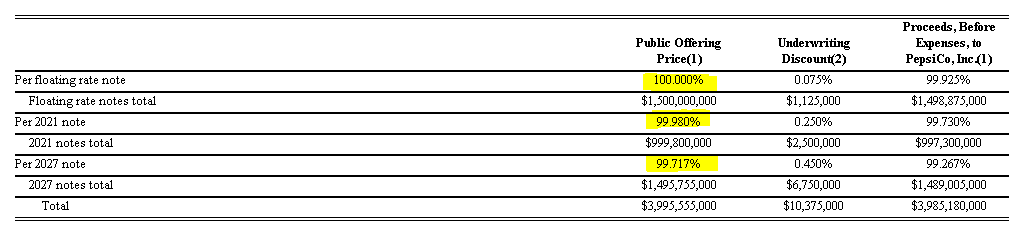


Source: <https://www.sec.gov/Archives/edgar/data/23217/000119312517307880/d466260d424b5.htm>

Here Conagra Floating Rate Notes due 2020 will mature on October 9,2020

1. **Issue Price:**

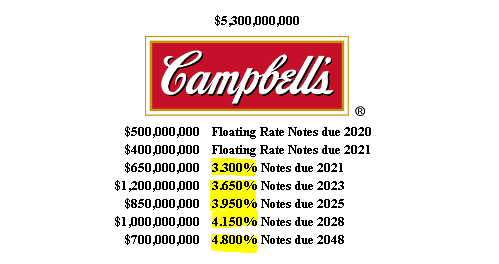
The issue price of shares is the price at which they are offered for sale when they first become available to the public. The issue price of a [bond](https://www.accountingtools.com/articles/2017/5/10/bond) is based on the relationship between the [interest rate](https://www.accountingtools.com/articles/2017/5/10/interest-rate) that the bond pays and the [market interest rate](https://www.accountingtools.com/articles/2017/5/8/market-interest-rate) being paid on the same date.



Source: <https://www.sec.gov/Archives/edgar/data/77476/000104746917006268/a2233478z424b2.htm>

1. **Current Coupon Rate:**

Coupon rate is the rate of interest paid by bond issuers on the bond’s face value. It is the periodic rate of interest paid by bond issuers to its purchasers. The coupon rate is calculated on the bond’s face value (or par value), not on the issue price or market value. For example, if you have a 10-year- $2,000 bond with a coupon rate of 10 per cent, you will get $200 every year for 10 years, no matter what happens to the bond price in the market.

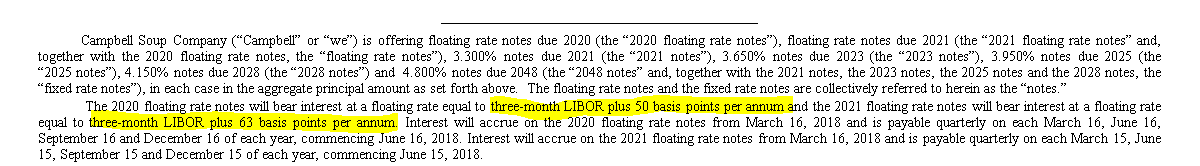


Source: <https://www.sec.gov/Archives/edgar/data/16732/000138713118001057/cpb-424b2_031318.htm>

1. **Floating Coupon Spread to Benchmark**

A floating rate fund is a fund that invests in financial instruments paying a variable or floating interest rate. A floating rate fund invests in bonds and debt instruments whose interest payments fluctuate with an underlying interest rate level, as opposed to paying fixed-rate income.

These Floating rate notes (FRNs) are [bonds](https://en.wikipedia.org/wiki/Bond_(finance)) that have a variable [coupon](https://en.wikipedia.org/wiki/Coupon_(bond)), equal to a [money market](https://en.wikipedia.org/wiki/Money_market) [reference rate](https://en.wikipedia.org/wiki/Reference_rate), like [LIBOR](https://en.wikipedia.org/wiki/LIBOR) or [federal funds rate](https://en.wikipedia.org/wiki/Federal_funds_rate), plus a quoted spread (also known as quoted margin). The spread is a rate that remains constant.



Source: <https://www.sec.gov/Archives/edgar/data/16732/000138713118001057/cpb-424b2_031318.htm>

Here in this Spread to floating rate notes due 2020 is **50 bps** and **63 bps** for the floating rate notes due 2021

1. **Coupon Type Code:**

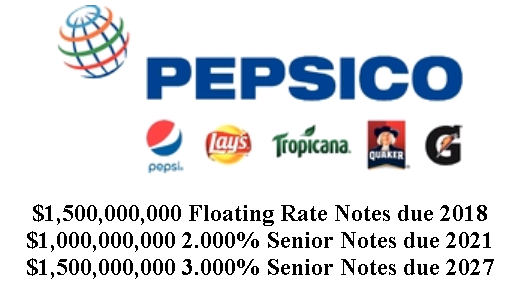
Categorizing the Note or Bond into “**FIX** **Type** **Coupon**” and “**FLOAT** **Type** **Coupon**” based on the Security.

In this FIX type of coupon have fixed interest rate associated to the note whereas with float type coupon spread is associated.

In the following example first Securities: Floating Rate Notes due 2018 is a **FLOAT** Coupon Type and

The rest of the securities of the document: Senior Notes due 2021 and Senior Notes due 2027 are of **FIX** Type

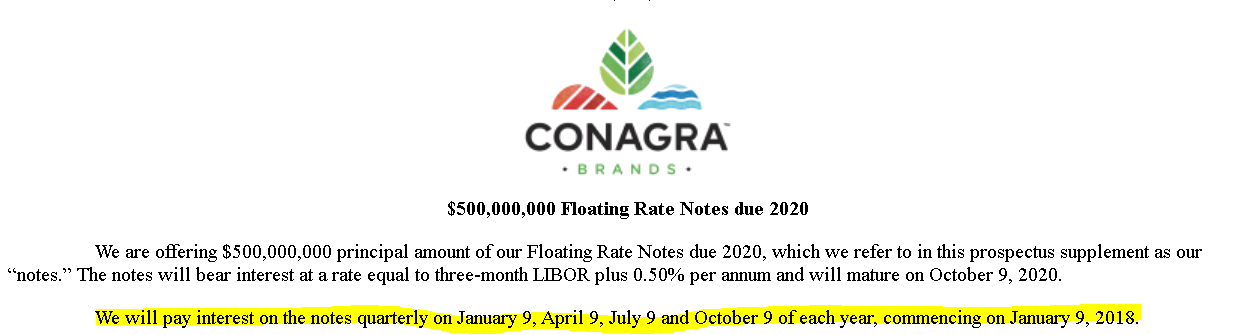
In this Model is identifying the value of Securities based on the Security names if it contains Floating value in the tag model call it as FLOAT, Second Not floating value in the Tag it’s FIX and if the tag is none Model not able to predict the value of Coupon type. In this case purposed method is checking for the Face Value and percentage related to the securities and then conforming it to FIX and FLOAT.



Source: https://www.sec.gov/Archives/edgar/data/77476/000104746917006268/a2233478z424b2.htm

1. **Coupon Pay Frequency Code:**

It’s the frequency of paying Interest to buyer in a year such as *Annually, Semi-Annually, Quarterly, Bi-Monthly, Monthly, Daily etc.*



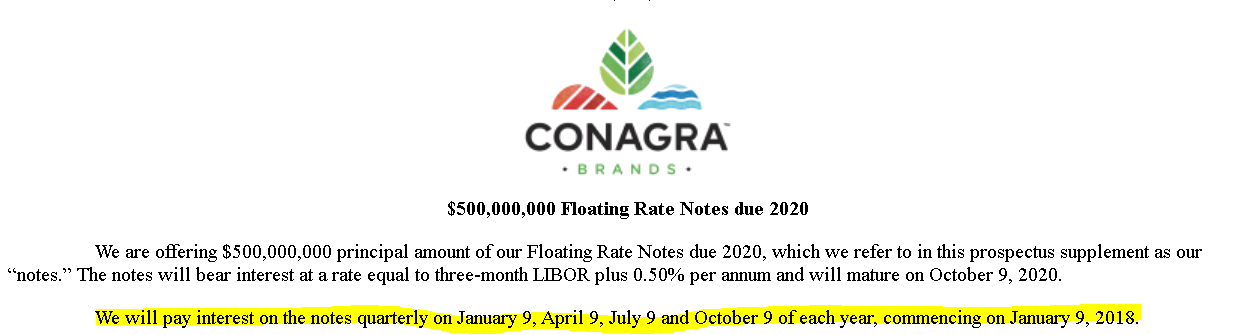
Source: <https://www.sec.gov/Archives/edgar/data/23217/000119312517307880/d466260d424b5.htm>

1. **First Coupon Date:**

It’s the date after issuing the notes or bond, when the interest paid to the customer for the very first time.

This signifies whether the first coupon payment is larger than normal to compensate for a long first coupon period, is smaller than normal to compensate for a short first coupon period, or is zero (and thereby creating a (normal) first period).

Note, all bonds have this feature, and it is implemented by setting the *First Coupon Date* parameter to a date that creates an odd first period

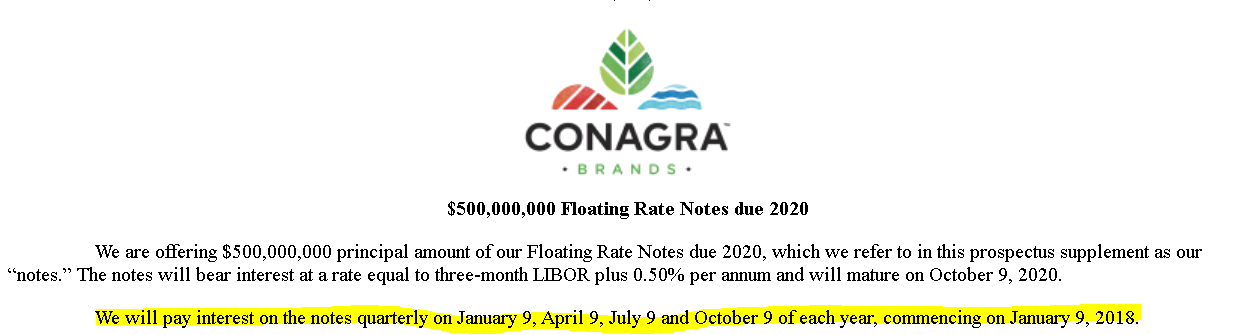


Source: <https://www.sec.gov/Archives/edgar/data/23217/000119312517307880/d466260d424b5.htm>

Here First coupon date is 9th January 2018

1. **Coupon Pay Date:**

It’s a list of interest payment dates associated to the coupon, paid **Annually, Semi-Annually, Quarterly, Bi-Monthly, Monthly, Daily** etc.

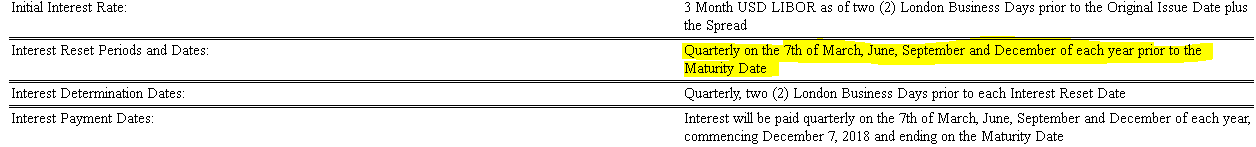


Source: <https://www.sec.gov/Archives/edgar/data/23217/000119312517307880/d466260d424b5.htm>

Here **January 9, April 9, July 9 and October 9** are the Coupon Pay Date and interest paid Quarterly for the Notes

1. **Rate Reset Frequency Code:**

Sometimes, the interest rate changes once a year, but some securities change interest rates as often as once a month or as seldom as every five years. The higher the reset frequency, the higher the financial [risk](https://financial-dictionary.thefreedictionary.com/risk) for the owner. For example, if the reset frequency is once a month, an owner could find the mortgage payment increasing every month for five monthsbefore it goes down again. This ties up more of the homeowner's [income](https://financial-dictionary.thefreedictionary.com/income) and increases the likelihood of [default](https://financial-dictionary.thefreedictionary.com/Default)



Source: //nasdevdump1/dsfiles/data\_extract/edgar/764764/20180905/0000764764-18-000164/424B3\_september2018\_mtn-pricingx.htm

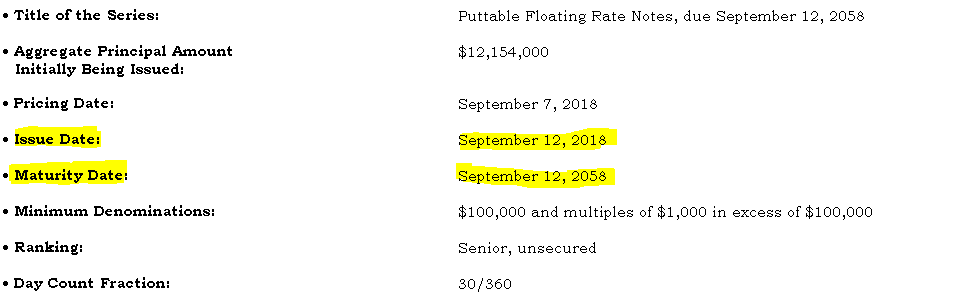
Here Coupon Rate reset frequency is quarterly and done on 7th of March, June, September and December of each year prior to the maturity date.

1. **Issue Type Code:**

It’s an indicator help to understand the kind of security **NOTE** or **BOND** as if the securities maturity within 10 years after issuing the securities it’s majorly a Note but if security mature after 10 years of the issuing date than it’s a **BOND.**

**NOTEs mature in 10 Years**

**BONDs mature after 10 Years**

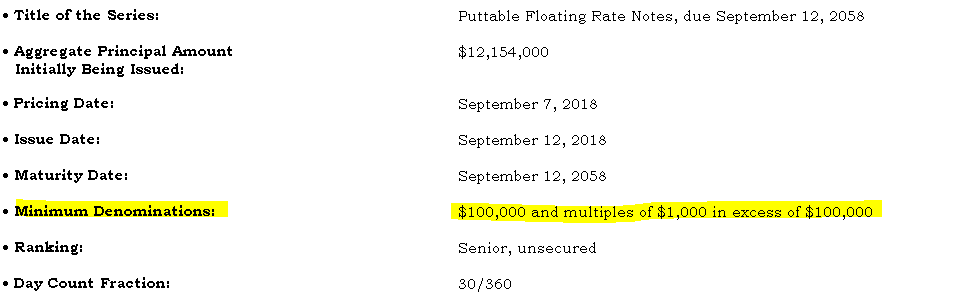


Source: //nasdevdump1/dsfiles/data\_extract/edgar/70858/20180911/0000891092-18-006721/424B2\_e2172\_424b2.htm

In this illustration year of issuing the Series is 2018 and year of maturity is 2058 and Series will mature **after 40 Years** of the issue date. So, Issue type Code for such securities will be BOND

1. **Minimum Denomination Value:**

A denomination is a unit of value most commonly assigned to physical currency, such as coins and notes, and other financial instruments that maintain a set value, such as government-issued [bonds](https://www.investopedia.com/terms/b/bond.asp). The value is often referred to the minimum nominal amount in which you can trade the bond.



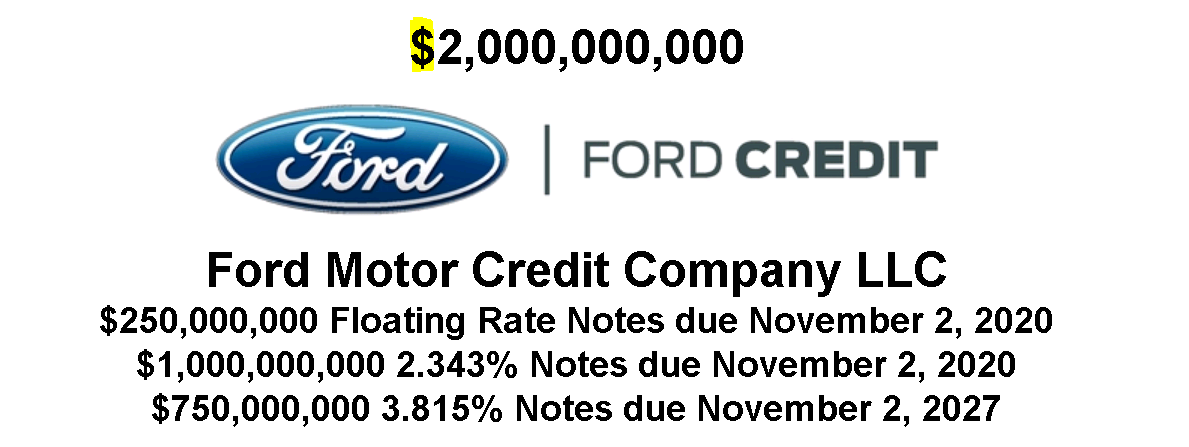
Source: //nasdevdump1/dsfiles/data\_extract/edgar/70858/20180911/0000891092-18-006721/424B2\_e2172\_424b2.htm

Here **$100,000** is the Minimum Denomination in which you can trade

1. **Currency of Issue:**

Currency of issue is the currency wherein the security is denominated.

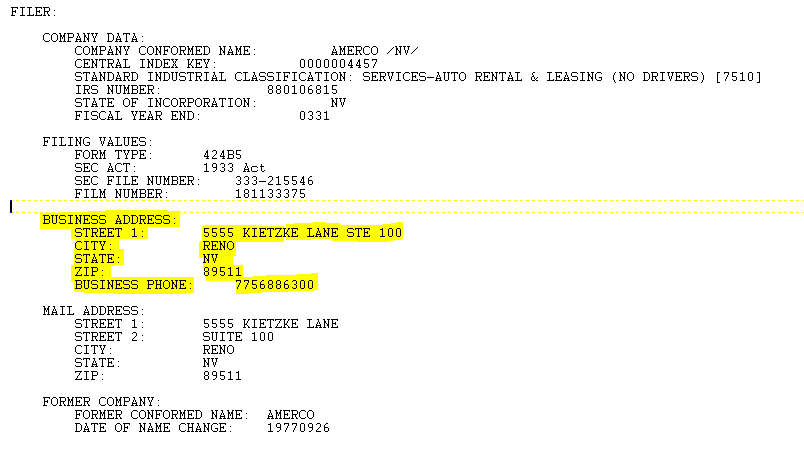
A bond that pays interest in one currency but pays the principal in a different currency is a concept of **Dual Currency Issue**. The amount of the principal repayment is set at initiation and paid at maturity. This principal amount usually allows for some appreciation in the exchange rate of the stronger currency. These issues are common in the Eurobond market and are a useful source of capital for multinational companies.



Source: <https://www.sec.gov/Archives/edgar/data/38009/000104746917006676/a2233666z424b2.htm>

1. **Country of Domicile:**

Country or state of incorporation or registration of a firm where it has its legal address or registered office, or which is considered in law as the center of its corporate affairs.  
According to the Banks It’s a place where a bill of exchange is made payable.



This Business address will help to find the Country of Domicile using Sates to Country Mapping



1. **Day Count Basis:**

In finance, a day count basis/convention determines how [interest](https://en.wikipedia.org/wiki/Interest) accrues over time for a variety of [investments](https://en.wikipedia.org/wiki/Investment), including [bonds](https://en.wikipedia.org/wiki/Bond_(finance)), notes, [loans](https://en.wikipedia.org/wiki/Loan), [mortgages](https://en.wikipedia.org/wiki/Mortgage_loan), medium-term notes, [swaps](https://en.wikipedia.org/wiki/Swap_(finance)), and [forward rate agreements](https://en.wikipedia.org/wiki/Forward_rate_agreement) (FRAs).

This determines the number of days between two [coupon](https://en.wikipedia.org/wiki/Coupon_(bond)) payments, thus calculating the amount transferred on payment dates and also the [accrued interest](https://en.wikipedia.org/wiki/Accrued_interest) for dates between payments.

The day count is also used to quantify periods of time when discounting a cash-flow to its [present value](https://en.wikipedia.org/wiki/Present_value). When a security such as a bond is sold between interest payment dates, the seller is eligible to some fraction of the coupon amount.

Day Count Convection looks like:

30/360

30/365

Actual/360

Actual/Actual etc.

### Fixed Rate

The fixed-rate leg of an interest rate swap and most fixed-rate bonds use either the **30/360**-day convention or **30/365**. This convention stipulates the month will always be treated as having 30 days in it, and the year will consistently be treated as having either 360 or 365 days.

Swap markets using the **30/360** convention for the fixed rate of a swap include the U.S. dollar, the euro and the Swiss franc.

Swaps in the British pound and the Japanese yen usually use the **30/365** convention; Australia, New Zealand and Hong Kong again follow the United Kingdom.

### Floating Rate

The floating-rate leg of most interest rate swaps uses some variation of an actual day count versus either a 360 or 365-day year.

The markets that use **30/360** for the fixed-rate leg, which include the U.S. dollar markets, use **actual/360** for the floating-rate leg. Those that use 30/365 on the fixed-rate leg use **actual/365** on the floating-rate leg.

### Treasury Bonds and Notes

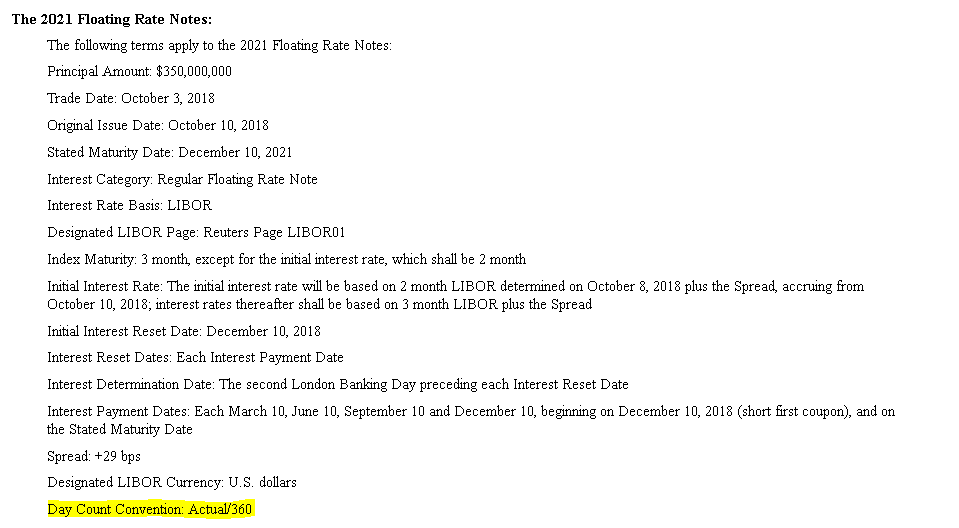
Bonds and notes issued by the U.S. Treasury earn interest calculated on an **actual/actual basis**. This means all days in a period carry equal value

it also means the length of coupon periods and the resultant payments vary.

### LIBOR Day Count

The [London Interbank Offered Rate (LIBOR)](https://www.investopedia.com/terms/l/libor.asp) is the most commonly used benchmark interest rate, and is posted daily at 11:45 a.m. London time.

For most currencies, interest at LIBOR is calculated on the **actual/360**-day basis and the major exception is again the British pound, which is calculated on the **actual/365**-day basis.



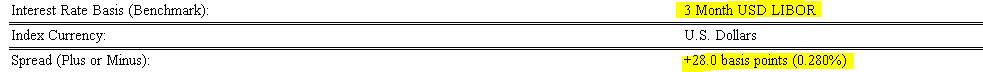
Source: //nasdevdump1/dsfiles/data\_extract/edgar/864270/20181005/0001193125-18-293772/424B2\_d617133d424b2.htm

1. **Rate Benchmark ID:**

An [interest rate](https://financial-dictionary.thefreedictionary.com/interest+rate) against which other interest rates are calculated.

For example, [LIBOR](https://financial-dictionary.thefreedictionary.com/LIBOR) is considered a benchmark rate because floating-rate instruments are related to it (for example, one may be calculated as LIBOR + 1%).

[Central banks](https://financial-dictionary.thefreedictionary.com/central+bank) make [loans](https://financial-dictionary.thefreedictionary.com/loan) to [banks](https://financial-dictionary.thefreedictionary.com/bank) under their jurisdiction at certain interest rates, which are then used as benchmarks for the loans those banks make.



Source: //nasdevdump1/dsfiles/data\_extract/edgar/764764/20180905/0000764764-18-000164/424B3\_september2018\_mtn-pricingx.htm

1. **Penultimate Coupon Date:**

Indicates bonds that have a non-standard last coupon period length. Note, only the generic functions support this feature, and it is implemented by setting the Penultimate *Coupon Date* parameter to a date that creates an odd last period

Penultimate coupon date= Maturity date - Rate reset frequency code

{***Subtract the number of reset days from the maturity date***}

1. **Industry Classification:**

The **Standard Industrial Classification**(SIC) is a system for classifying [industries](https://en.wikipedia.org/wiki/Industry) by a four-digit code.

It is used by government agencies to classify industry areas. The SIC system is also used by agencies in other countries, e.g., by the [United Kingdom](https://en.wikipedia.org/wiki/United_Kingdom)'s [Companies House](https://en.wikipedia.org/wiki/Companies_House)

In the United States the SIC code is being supplanted by the six-digit [North American Industry Classification System](https://en.wikipedia.org/wiki/North_American_Industry_Classification_System) (NAICS code), which was released in 1997; however certain government departments and agencies, such as the [U.S. Securities and Exchange Commission](https://en.wikipedia.org/wiki/U.S._Securities_and_Exchange_Commission) (SEC), still use the SIC codes.

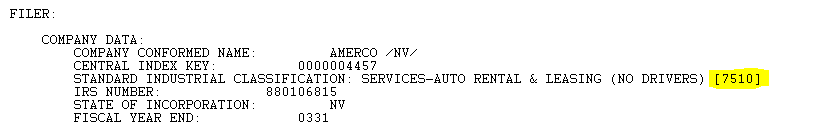
The SIC codes can be grouped into progressively broader industry classifications: industry group, major group, and division.

The **first 3 digits** of the **SIC code** indicate the **industry group**, and

the **first two digits** indicate the **major group**.

Each division encompasses a range of SIC codes

|  |  |
| --- | --- |
| **Range of SIC Codes** | **Division** |
| 0100-0999 | Agriculture, Forestry and Fishing |
| 1000-1499 | Mining |
| 1500-1799 | Construction |
| 1800-1999 | not used |
| 2000-3999 | Manufacturing |
| 4000-4999 | Transportation, Communications, Electric, Gas and Sanitary service |
| 5000-5199 | Wholesale Trade |
| 5200-5999 | Retail Trade |
| 6000-6799 | Finance, Insurance and Real Estate |
| 7000-8999 | Services |
| 9100-9729 | Public Administration |
| 9900-9999 | Non- classifiable |

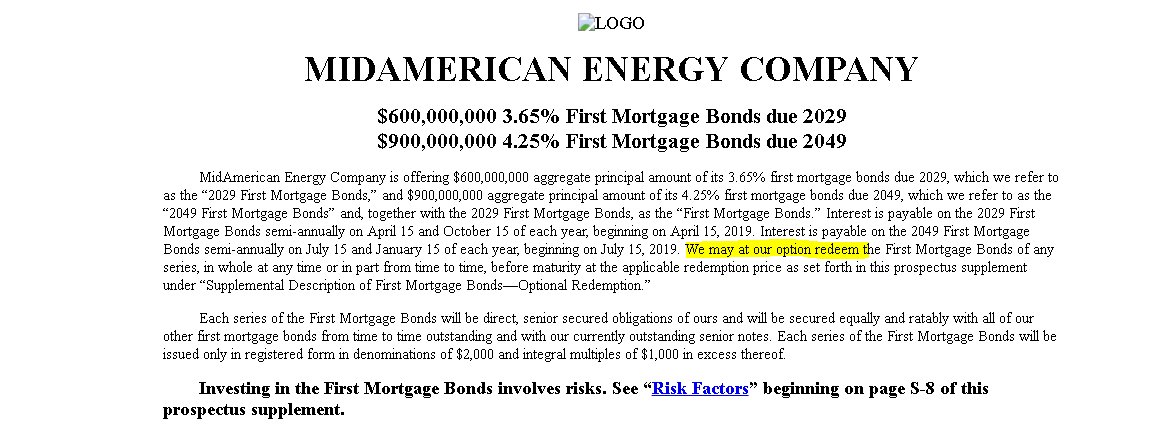


SIC or Industrial Classification of AMERCO/NV is 7510 and belongs to Service industry

1. **Make Whole Call Indicator:**

A make whole call indicator is a type of [call provision](https://www.investopedia.com/terms/c/callprovision.asp) on a bond allowing the issuer to pay off remaining debt early. The issuer typically should make a lump sum payment to the investor derived from a formula based on the [net present value](https://www.investopedia.com/terms/n/npv.asp) (NPV) of future coupon payments that will not be paid incrementally because of the call combined with the principal payment the investor would have received at maturity.

Many bonds issued today are “callable,” which means they can be redeemed by the issuer at set points before its listed maturity date. That means the issuer pays investors the call price and any accrued interest, and doesn’t make any future interest payments.



* 1. **Other Model Terms used in the project**

1. **Document to Skip:**

It’s a flag indicating to document is of 424B type or not.

Document to Skip is “Yes” when we can skip the document for the analysis and not belongs to 424B type of Security

Document to Skip is “NO” means it’s of 424B Type Security

1. **Correct Issuer**

Correct Issuer is a flag represent the Issuer picked by the model is correct or not.

If Correct Issuer Flag is “Yes” it means Issuer picked by the model is available in the Metadata of the document and related terms of the issuers are also correct.

Terms related to the Correct Issuer are

1. Issuer Name
2. CIK
3. Industry Classification or SIC Code and
4. Country of Domicile

Accuracy of the Correct issuer flag is above 80% and model can fetch the details almost correctly for all the other terms associated to correct issuer.