News Data analysis final notebook

Data cleaning and profiling for U.S. equities news data

FINISHED

The dataset used for this profiling is data of financial news related to U.S. equities. The aim of this notebook is to do data profiling, cleaning, and ingestion. The news dataset is obtained from Kaggle platform at : Link to dataset on Kaggle (https://www.kaggle.com/datasets/gennadiyr/us-equities-news-data)

Acknowledgements

The origial datasource is from https://www.investing.com/ (https://www.investing.com/) . Investing.com is an online data and news website that provides financial information. Every row of this dataset includes attribution to the data provider and link on the source.

Took 5 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:36:44 PM.

Data loading

FINISHED

The first step in the step in the data cleansing and profiling is to load the data from the source, in apropriate format. Loading the data from csv file stored in HDFS system.

Took 0 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:36:44 PM.

```
val newsFilePath = "us_equities_news_dataset.csv"

val rawDF = spark.read
  .option("header", "true")
  .option("multiLine", "true")
  .option("inferSchema", "true")
  .option("escape", "\"")
  .csv(newsFilePath)


rawDF.cache()
z.show(rawDF, 5)
```

SPARK JOB FINISHED

settings

id	ticker	title	category	content	release_date	provider
221515	NIO	Why Shares of Chinese Electric Car Maker NIO Are Flying High Today	news	What s happening Shares of Chinese electric car maker NIO NYSE NIO were sharply higher on Wednesday morning after a Chinese business news outlet reported that the cash strapped company had secured new financing from a major automaker As of 12 p m EST NIO s American depositary shares ADS were up about 16 from Tuesday s closing price So what According to a	2020-01-15	The Motley Fool

Output is truncated to 102400 bytes. Learn more about **ZEPPELIN_INTERPRETER_OUTPUT_LIMIT**

Took 1 min 9 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:37:53 PM.

```
println(s"Total columns : ${rawDF.columns.length}")
println(s"Total rows : ${rawDF.count()}")

rawDF.printSchema
```

SPARK JOB FINISHED

```
Total columns : 9
Total rows : 221513
root
 |-- id: integer (nullable = true)
 |-- ticker: string (nullable = true)
 |-- title: string (nullable = true)
 |-- category: string (nullable = true)
 |-- content: string (nullable = true)
 |-- release_date: string (nullable = true)
 |-- provider: string (nullable = true)
 |-- url: string (nullable = true)
 |-- article_id: integer (nullable = true)
```

Took 3 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:37:56 PM.

News Data analysis final notebook

```
import org.apache.spark.sql.functions.{col,when, count}
import org.apache.spark.sql.{Column, SparkSession}

def countCols(columns:Array[String]):Array[Column]={
    columns.map(c=>{
      count(when(col(c).isNull,c)).alias(c)
    })
}
```

FINISHED

```
import org.apache.spark.sql.functions.{col, when, count}
import org.apache.spark.sql.{Column, SparkSession}
countCols: (columns: Array[String])Array[org.apache.spark.sql.Column]
```

Took 1 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:37:57 PM.

```
println("Count of null entries in the columns : ")
z.show(rawDF.select(countCols(rawDF.columns):_*))
```

SPARK JOB FINISHED

```
Count of null entries in the columns :
```

settings

id	ticker	title	category	content	release_date	provider	url
0	0	0	0	8	0	0	0

Took 1 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:37:58 PM.

```
val nullRemovedDF = rawDF.filter($"content".isNotNull)
                         .withColumnRenamed("article_id","articleId")
                         .withColumnRenamed("release_date","releaseDate")

println(s"Total columns : ${nullRemovedDF.columns.length}")
println(s"Total rows : ${nullRemovedDF.count()}")
println("Count of null entries in the columns : ")

z.show(nullRemovedDF.select(countCols(nullRemovedDF.columns):_*))
```

SPARK JOB FINISHED

```
Total columns : 9
Total rows : 221505
Count of null entries in the columns :
```

settings

id	ticker	title	category	content	releaseDate	provider	url
0	0	0	0	0	0	0	0

```
nullRemovedDF: org.apache.spark.sql.DataFrame = [id: int, ticker: string ... 7 more fields]
```

Took 2 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:38:01 PM.

```
val cleanDF = nullRemovedDF.withColumn("content", trim(col("content")))
                           .where(length($"content") >= length($"title") )
                           .where(length($"content") > 10)
                           .withColumn("content", regexp_replace($"content", "\t", " "))

z.show(cleanDF)
```

SPARK JOB (http://nyu-dataproc-w-0.c.hpc-dataproc-19b8.internal:35207/jobs/job?id=11) FINISHED

settings

id	ticker	title	category	content	releaseDate	provider	url
221515	NIO	Why Shares of Chinese Electric Car Maker NIO Are Flying High Today	news	What s happening Shares of Chinese electric car maker NIO NYSE NIO were sharply higher on	2020-01-15	The Motley Fool	https://invst

News Data analysis final notebook

Wednesday morning after a Chinese business news outlet reported that the cash strapped company had secured new

Output is truncated to 102400 bytes. Learn more about **ZEPPELIN_INTERPRETER_OUTPUT_LIMIT**

Took 1 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:38:02 PM.

Handling duplicate values

FINISHED

Next task is to remove duplicate news from the dataset, thus counting distinct news and removing duplicates if any.

Took 0 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:38:03 PM.

```
val dropDisDF = cleanDF.dropDuplicates("content")

println(s"Total columns : ${dropDisDF.columns.length}")
println(s"Total rows : ${dropDisDF.count()}")
```

SPARK JOB FINISHED

```
Total columns : 9
Total rows : 220984
dropDisDF: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [id: int, ticker: string ... 7 more fields]
```

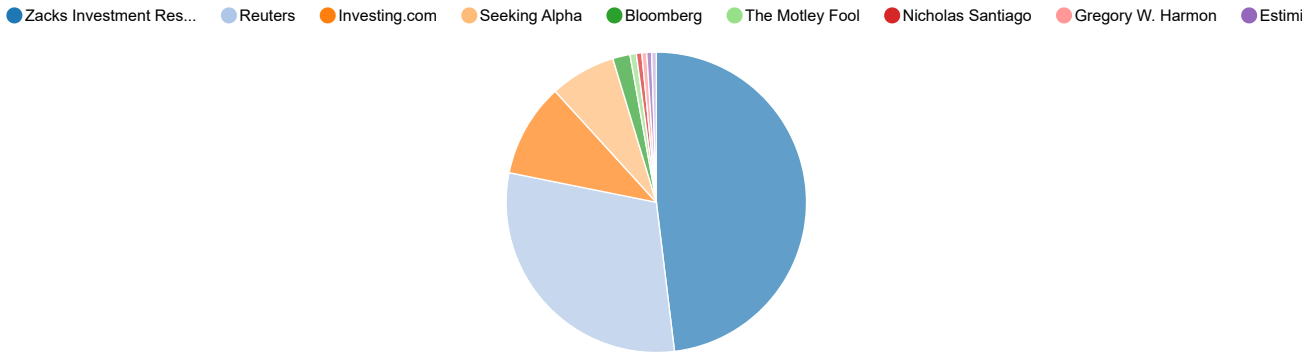
Took 37 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:38:40 PM.

```
val countByPublishers = dropDisDF.groupBy("provider").count().sort(col("count").desc)

println(s"The total number of distinct news providers covered by this dataset : ${countByPublishers.count()}")
println("Top 10 news providers in this dataset are : ")

countByPublishers.show(10, false)
z.show(countByPublishers.limit(10), 10)
```

SPARK JOB FINISHED

```
The total number of distinct news providers covered by this dataset : 979
Top 10 news providers in this dataset are :
+-------------------------+-----+
|provider                 |count|
+-------------------------+-----+
|Zacks Investment Research|87933|
|Reuters                  |55054|
|Investing.com            |18436|
|Seeking Alpha            |12957|
|Bloomberg                |3407 |
|The Motley Fool          |1250 |
|Nicholas Santiago        |1064 |
|Gregory W. Harmon        |972  |
|Estimize                 |961  |
|Ryan Mallory             |905  |
+-------------------------+-----+
only showing top 10 rows
```

settings




```
countByPublishers: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [provider: string, count: bigint]
```

Took 2 min 18 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:40:58 PM.

```
val correctDateDF = dropDisDF.withColumn("releaseDate", to_timestamp(col("releaseDate")))
z.show(correctDateDF)
```

SPARK JOB FINISHED

settings

id	ticker	title	category	content	releaseDate	provider	url
240268	MU	3 Ideas For Investing In Tech Stocks Amid Trade War Volatility	opinion	0 Approaching The Recent Selloff 3 Focusing On Earnings 5 Oversold Stocks 9 Avoid The Trade	2018-04-09 00:00:00.0	Zacks Investment Research	https://www com/analys for-investing stocks-amid

News Data analysis final notebook

				Concerns Small Cap StocksOn today s episode of the Tech Talk Tuesday podcast Ryan McQueeney			war-volatilit 200304397

Output is truncated to 102400 bytes. Learn more about **ZEPPELIN_INTERPRETER_OUTPUT_LIMIT**

Took 34 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:41:32 PM.

```
correctDateDF.printSchema
```

FINISHED

```
root
 |-- id: integer (nullable = true)
 |-- ticker: string (nullable = true)
 |-- title: string (nullable = true)
 |-- category: string (nullable = true)
 |-- content: string (nullable = true)
 |-- releaseDate: timestamp (nullable = true)
 |-- provider: string (nullable = true)
 |-- url: string (nullable = true)
 |-- articleId: integer (nullable = true)
```

Took 1 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:41:33 PM.

```
val financialQuarterDF = correctDateDF.withColumn("quarter",quarter(correctDateDF.col("releaseDate")))
                                      .withColumn("year",year(correctDateDF.col("releaseDate")))
                                      .select(concat($"year", lit("Q"), $"quarter") as "financialQuarter")
                                      .groupBy("financialQuarter").count()
                                      .sort("financialQuarter")
z.show(financialQuarterDF)
```

SPARK JOB FINISHED

settings

financialQuarter	count
2008Q4	479
2009Q1	632
2009Q2	919
2009Q3	2147
2009Q4	807
2010Q1	22
2010Q2	16
2010Q3	74

```
financialQuarterDF: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [financialQuarter: string, count: bigint]
```

Took 38 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:42:11 PM.

```
val maximumCountForTitleProviderArticle = correctDateDF.groupBy("title", "provider", "articleId").count().select(max("count"))
maximumCountForTitleProviderArticle.show()
```

SPARK JOB FINISHED

```
+----------+
|max(count)|
+----------+
|         1|
+----------+

maximumCountForTitleProviderArticle: org.apache.spark.sql.DataFrame = [max(count): bigint]
```

Took 46 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:42:57 PM.

```
z.show(correctDateDF)
```

SPARK JOB FINISHED

settings

id	ticker	title	category	content	releaseDate	provider	url
240268	MU	3 Ideas For Investing In Tech Stocks Amid Trade War Volatility	opinion	0 Approaching The Recent Selloff 3 Focusing On Earnings 5 Oversold Stocks 9 Avoid The Trade Concerns Small Cap StocksOn today s episode of the Tech Talk Tuesday podcast Ryan McQueeney discusses three tech investment ideas for	2018-04-09 00:00:00.0	Zacks Investment Research	https://www com/analys for-investing stocks-amid war-volatilit 200304397

Output is truncated to 102400 bytes. Learn more about **ZEPPELIN_INTERPRETER_OUTPUT_LIMIT**

News Data analysis final notebook

Notebook Last updated by sz3369_nyu_edu at December 17 2022, 3:43:30 PM.

```
import java.util.Properties
import edu.stanford.nlp.pipeline.StanfordCoreNLP
import edu.stanford.nlp.ling.CoreAnnotations
import edu.stanford.nlp.neural.rnn.RNNCoreAnnotations
import edu.stanford.nlp.sentiment.SentimentCoreAnnotations
import scala.collection.JavaConverters._
import org.apache.spark.SparkContext
import edu.stanford.nlp.util.CoreMap
```

FINISHED

```
import java.util.Properties
import edu.stanford.nlp.pipeline.StanfordCoreNLP
import edu.stanford.nlp.ling.CoreAnnotations
import edu.stanford.nlp.neural.rnn.RNNCoreAnnotations
import edu.stanford.nlp.sentiment.SentimentCoreAnnotations
import scala.collection.JavaConverters._
import org.apache.spark.SparkContext
import edu.stanford.nlp.util.CoreMap
```

Took 0 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:43:30 PM.

```
 def sentiment(tweets: String): String = {
    var mainSentiment = 0
    var longest = 0;
    val sentimentText = Array("Very Negative", "Negative", "Neutral", "Positive", "Very Positive")
    val props = new Properties();
    props.setProperty("annotators", "tokenize, ssplit, parse, sentiment");
    new StanfordCoreNLP(props).process(tweets).get(classOf[CoreAnnotations.SentencesAnnotation]).asScala.foreach((sentence: CoreMap) => {
      val sentiment = RNNCoreAnnotations.getPredictedClass(sentence.get(classOf[SentimentCoreAnnotations.SentimentAnnotatedTree]));
      val partText = sentence.toString();
      if (partText.length() > longest) {
        mainSentiment = sentiment;
        longest = partText.length();
      }
    })
    sentimentText(mainSentiment)
  }

import org.apache.spark.sql.functions._
import org.apache.spark.sql.types._

def mysentiment = udf((x: String) =>
    {
        sentiment(x)
    }
    )
```

FINISHED

```
sentiment: (tweets: String)String
import org.apache.spark.sql.functions._
import org.apache.spark.sql.types._
mysentiment: org.apache.spark.sql.expressions.UserDefinedFunction
```

Took 1 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:43:31 PM.

```
sentiment("Why Shares of Chinese Electric Car Maker NIO Are Flying High Today,news, What s happenin Shares of Chinese electric car maker NIO  NYSE NIO  were sharply higher after a Chinese business news outlet reported that the cash strapped company had secured new financing from a major automaker  As of 12 p m  EST  NIO s American depos up about 16  from Tuesday s closing price  So what If this report is accurate and if the deal closes  then it s extremely bullish for NIO" )
```

FINISHED

```
res13: String = Negative
```

Took 7 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:43:38 PM.

```
val df3 = correctDateDF.groupBy("ticker","releaseDate").count()
```

FINISHED

```
df3: org.apache.spark.sql.DataFrame = [ticker: string, releaseDate: timestamp ... 1 more field]
```

Took 1 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:43:39 PM.

```
z.show(df3)
```

SPARK JOB FINISHED

settings

ticker	releaseDate	count
NUE	2016-05-04 00:00:00.0	1
CIEN	2017-04-10 00:00:00.0	2
NEE	2019-11-20 00:00:00.0	2
RF	2019-11-20 00:00:00.0	1
NWSA	2012-10-29 00:00:00.0	18
MSFT	2018-06-10 00:00:00.0	3
ABT	2016-06-17 00:00:00.0	1
K	2017-07-19 00:00:00.0	1

Output is truncated to 1000 rows. Learn more about **zeppelin.spark.maxResult**

Took 45 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:44:24 PM.

News Data analysis final notebook

```
df3.printSchema
```

FINISHED

```
root
 |-- ticker: string (nullable = true)
 |-- releaseDate: timestamp (nullable = true)
 |-- count: long (nullable = false)
```

Took 0 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:44:24 PM.

```
val dfWithWeekNumber = df3.withColumn("dayOfWeek", date_format(col("releaseDate"), "E"))

z.show(dfWithWeekNumber)
```

SPARK JOB FINISHED

settings

ticker	releaseDate	count	dayOfWeek
NUE	2016-05-04 00:00:00.0	1	Wed
CIEN	2017-04-10 00:00:00.0	2	Mon
NEE	2019-11-20 00:00:00.0	2	Wed
RF	2019-11-20 00:00:00.0	1	Wed
NWSA	2012-10-29 00:00:00.0	18	Mon
MSFT	2018-06-10 00:00:00.0	3	Sun
ABT	2016-06-17 00:00:00.0	1	Fri
K	2017-07-19 00:00:00.0	1	Wed

Output is truncated to 1000 rows. Learn more about **zeppelin.spark.maxResult**

```
dfWithWeekNumber: org.apache.spark.sql.DataFrame = [ticker: string, releaseDate: timestamp ... 2 more fields]
```

Took 45 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:45:09 PM. (outdated)

```
val noNulls = dfWithWeekNumber.na.drop("any")

z.show(noNulls)
noNulls.printSchema
```

SPARK JOB FINISHED

settings

ticker	releaseDate	count	dayOfWeek
BA	2011-04-01 00:00:00.0	2	Fri
BA	2017-06-07 00:00:00.0	9	Wed
CIM	2019-07-24 00:00:00.0	1	Wed
SO	2020-01-23 00:00:00.0	3	Thu
FCX	2017-03-02 00:00:00.0	2	Thu
AMD	2019-11-19 00:00:00.0	1	Tue
AEGN	2019-07-09 00:00:00.0	1	Tue
CVX	2018-05-08 00:00:00.0	1	Tue

Output is truncated to 1000 rows. Learn more about **zeppelin.spark.maxResult**

```
root
 |-- ticker: string (nullable = true)
 |-- releaseDate: timestamp (nullable = true)
 |-- count: long (nullable = false)
 |-- dayOfWeek: string (nullable = true)

noNulls: org.apache.spark.sql.DataFrame = [ticker: string, releaseDate: timestamp ... 2 more fields]
```

Took 46 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:45:55 PM.

Shifting the news published on Saturday and Sunday to the next buisness day as per stock market working hours.

FINISHED

Took 0 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:52:08 PM.

```
val df4 = dfWithWeekNumber.withColumn("shiftedDate", when( col("dayOfWeek") === "Sat", date_add(col("releaseDate"),2))
.when(col("dayOfWeek") === "Sun", date_add(col("releaseDate"),1))
.otherwise(col("releaseDate")))
```

FINISHED

```
df4: org.apache.spark.sql.DataFrame = [ticker: string, releaseDate: timestamp ... 3 more fields]
```

Took 0 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:45:55 PM.

News Data analysis final notebook

```
df4.show()
```

SPARK JOB FINISHED

```
|    BA|2017-06-07 00:00:00|    9|      Wed|2017-06-07 00:00:00|
|   CIM|2019-07-24 00:00:00|    1|      Wed|2019-07-24 00:00:00|
|    SO|2020-01-23 00:00:00|    3|      Thu|2020-01-23 00:00:00|
|   FCX|2017-03-02 00:00:00|    2|      Thu|2017-03-02 00:00:00|
|   AMD|2019-11-19 00:00:00|    1|      Tue|2019-11-19 00:00:00|
|  AEGN|2019-07-09 00:00:00|    1|      Tue|2019-07-09 00:00:00|
|   CVX|2018-05-08 00:00:00|    1|      Tue|2018-05-08 00:00:00|
|   DDS|2018-02-11 00:00:00|    2|      Sun|2018-02-12 00:00:00|
|   NEM|2019-05-28 00:00:00|    1|      Tue|2019-05-28 00:00:00|
|    MS|2019-03-12 00:00:00|    3|      Tue|2019-03-12 00:00:00|
|    GM|2019-09-16 00:00:00|    7|      Mon|2019-09-16 00:00:00|
|   NUE|2018-03-05 00:00:00|    1|      Mon|2018-03-05 00:00:00|
|   NSC|2014-07-25 00:00:00|    1|      Fri|2014-07-25 00:00:00|
|    SO|2019-02-13 00:00:00|    2|      Wed|2019-02-13 00:00:00|
|    SO|2019-04-21 00:00:00|    1|      Sun|2019-04-22 00:00:00|
|    BX|2019-02-15 00:00:00|    3|      Fri|2019-02-15 00:00:00|
|   JNJ|2016-05-29 00:00:00|    1|      Sun|2016-05-30 00:00:00|
```

Took 46 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:46:41 PM.

```
val checkiIfWork = df4.filter(col("releaseDate") =!= col("shiftedDate"))
```

FINISHED

```
checkiIfWork: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [ticker: string, releaseDate: timestamp ... 3 more fields]
```

Took 0 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:46:41 PM.

Checking if the date shifting to next buisness day is working

FINISHED

Took 0 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:51:03 PM.

```
z.show(checkiIfWork)
```

SPARK JOB FINISHED

settings

ticker	releaseDate	count	dayOfWeek	shiftedDate
XLE	2012-05-13 00:00:00.0	1	Sun	2012-05-14 00:00
LMT	2019-05-05 00:00:00.0	1	Sun	2019-05-06 00:00
MSFT	2017-05-14 00:00:00.0	7	Sun	2017-05-15 00:00
XOM	2016-10-16 00:00:00.0	1	Sun	2016-10-17 00:00
J	2019-10-27 00:00:00.0	2	Sun	2019-10-28 00:00
AZO	2020-01-26 00:00:00.0	1	Sun	2020-01-27 00:00
BRKb	2012-04-01 00:00:00.0	1	Sun	2012-04-02 00:00
LRCX	2017-12-10 00:00:00.0	1	Sun	2017-12-11 00:00
CHK	2012-03-30 00:00:00.0	1	Sun	2012-04-02 00:00

Output is truncated to 1000 rows. Learn more about **zeppelin.spark.maxResult**

Took 46 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:47:27 PM.

```
df4.write.option("header",true).csv("/tmp/news_data_cleaned/news_features_final")
```

SPARK JOB FINISHED

Took 48 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:48:15 PM.

```
z.show(df4)
```

SPARK JOB FINISHED

settings

ticker	releaseDate	count	dayOfWeek	shiftedDate
BA	2011-04-01 00:00:00.0	2	Fri	2011-04-01 00:00
BA	2017-06-07 00:00:00.0	9	Wed	2017-06-07 00:00
CIM	2019-07-24 00:00:00.0	1	Wed	2019-07-24 00:00
SO	2020-01-23 00:00:00.0	3	Thu	2020-01-23 00:00
FCX	2017-03-02 00:00:00.0	2	Thu	2017-03-02 00:00
AMD	2019-11-19 00:00:00.0	1	Tue	2019-11-19 00:00
AEGN	2019-07-09 00:00:00.0	1	Tue	2019-07-09 00:00
CVX	2018-05-08 00:00:00.0	1	Tue	2018-05-08 00:00

Output is truncated to 1000 rows. Learn more about **zeppelin.spark.maxResult**

Took 45 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:49:00 PM.

News Data analysis final notebook

```
val tsla = df4.filter(col("ticker") === "TSLA")
```

FINISHED

```
tsla: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [ticker: string, releaseDate: timestamp ... 3 more fields]
```

Took 0 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:49:00 PM.

```
val plottslaNews = tsla.select(col("shiftedDate"), col("count"))
```

FINISHED

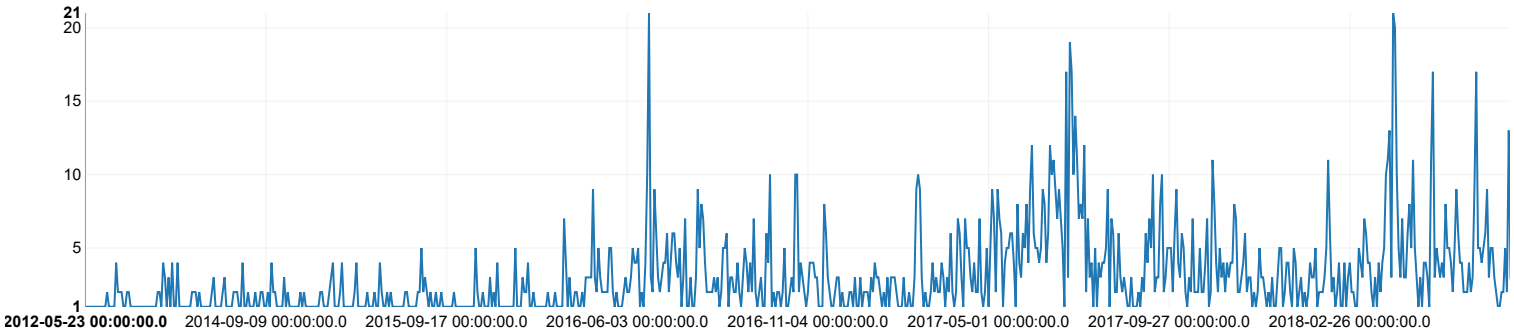
```
plottslaNews: org.apache.spark.sql.DataFrame = [shiftedDate: timestamp, count: bigint]
```

Took 1 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:49:01 PM.

```
val plottslaNewsSorted = plottslaNews.sort(col("shiftedDate"))
z.show(plottslaNewsSorted)
```

SPARK JOB FINISHED

settings



Output is truncated to 1000 rows. Learn more about **zeppelin.spark.maxResult**

```
plottslaNewsSorted: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [shiftedDate: timestamp, count: bigint]
```

Took 43 sec. Last updated by sz3369_nyu_edu at December 17 2022, 3:49:44 PM.

FINISHED