# Data Mining Project
# Investment Fund Analytics
# Social Media World News Impact on Stock Index Values

## Final Presentation

Bernardo Najlis

DS8004 – Winter 2017

March 27, 2017

Ryerson University

# Agenda

01

Case Description and Problem Presentation

02

Proposed Methodology as Data Mining Problem

03

Data Acquisition and Engineering

04

Descriptive Analytics on the Data

05

Predictive Models Description and Results

06

Conclusions, Lessons Learned and Next Steps

# Background on News Analytics (1)

- A large number of companies use news analysis to help them make business decisions, especially with regards to predicting
  - Stock Price movements
  - Volatility
  - Trade Volume
- Provided a set of values such as sentiment and relevance, as well as the frequency of news arrivals, it is possible to construct news sentiment scores for multiple asset classes

Source: Wikipedia (https://en.m.wikipedia.org/wiki/News_analytics)

# Background on News Analytics (2)

- Applications / Strategies
  - Absolute Return Strategy
  - Relative Return Strategy
  - Financial Risk Management
  - Algorithmic Order Execution
- Sources
  - Tetlock, Paul C. - Does Public Financial News Resolve Asymmetric Information? (http://ssrn.com/abstract=1303612)
  - Elizabeth A. Demers, Clara Vega – The Impact of Credibility on the Pricing of Managerial Textual Content (https://dx.doi.org/10.2139%2Fssrn.1153450)
  - Gsb.Columbia.edu - More Than Words – Qualifying Language to Measure Firms' Fundamentals (http://www1.gsb.columbia.edu/mygsb/faculty/research/pubfiles/3096/More_Than_Words_tetlock.pdf)
  - Northinfo.com - Equity Portfolio Risk (volatility) estimation using market information and sentiment (http://www.northinfo.com/documents/313.pdf)

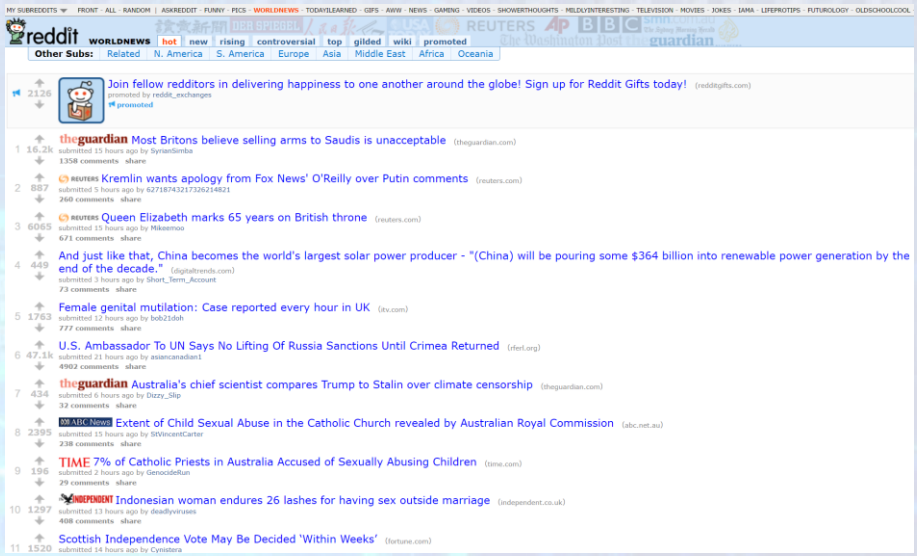
Source: Wikipedia (https://en.m.wikipedia.org/wiki/News_analytics)

# Objectives

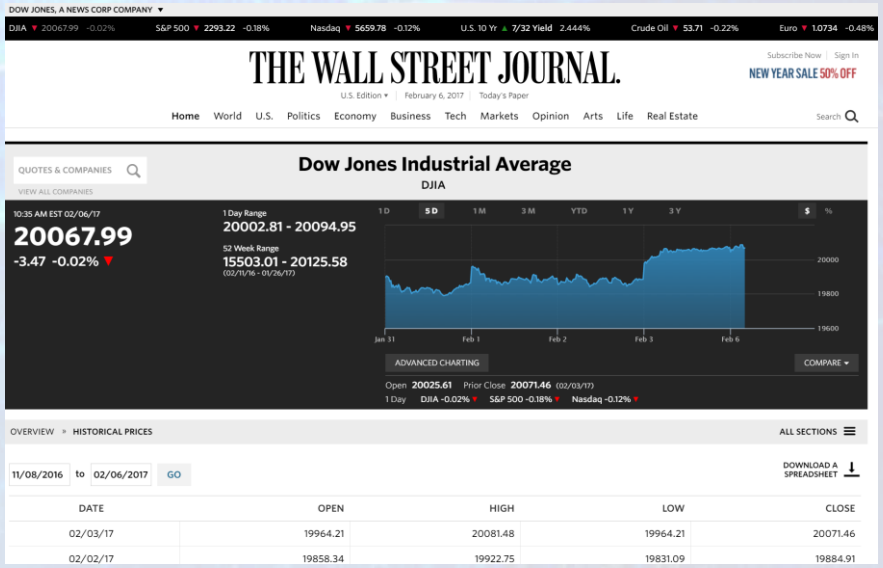
To create a model that predicts the DJIA stock index trends (up, down) by looking at the correlation between world news events and stock market index using text analytics

# Data sources

- News data:
  - Historical news headlines from Reddit **/r/worldnews**




- Stock data:
  - Dow Jones Industrial Average (DJIA) daily historical



# Proposed Methodology

1. Connect to Reddit API and download news headlines from **/r/worldnews**
    1. If API imposes restrictions, use available Reddit data dumps
2. Find the TOP 25 sorted by 'hot'
3. Perform the text analytics
    - tokenization
    - Stop word removal
    - Stemming
    - Sentiment detection / classification
4. Download DJIA daily historical (WSJ or other online sources available)
5. Label daily news headlines (0 or 1) based on the index value move comparing open for the same day

- Train the model with the output of 3 and 5
- Test and tune the model with test set

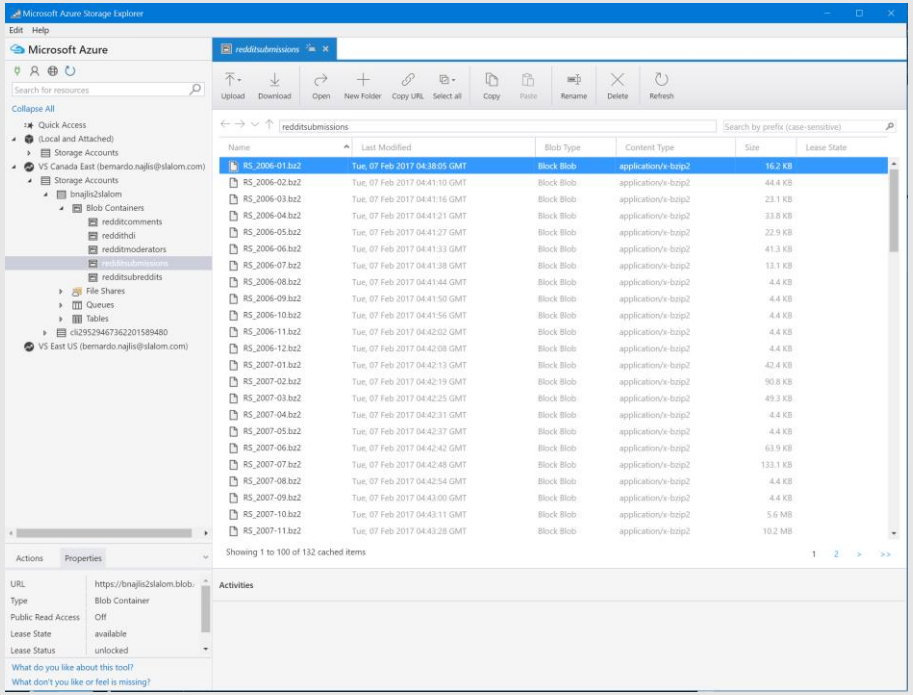
# Data Acquisition

Methodology and Reproducibility

Cleaning, Transformation and Feature Engineering

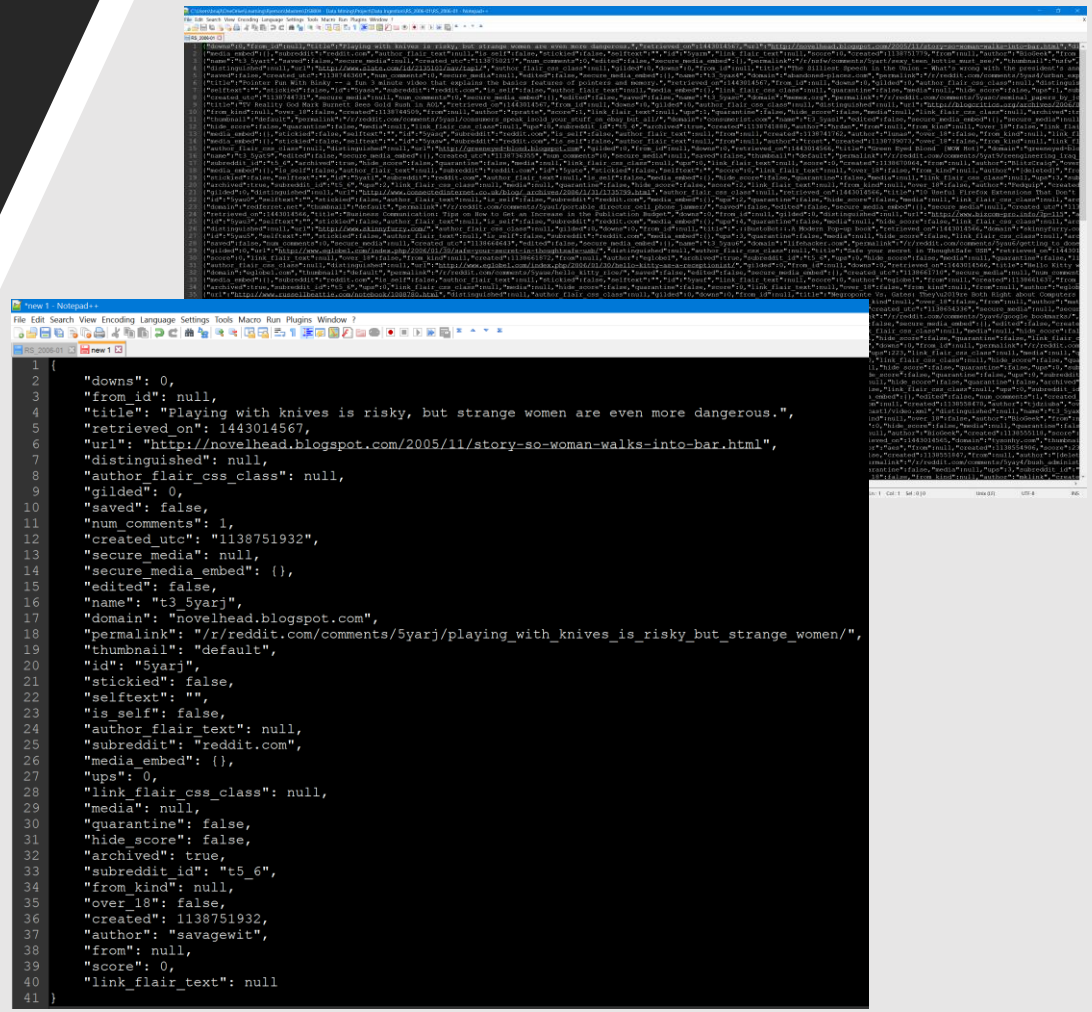
# Data Acqusition - Methodology

- Reddit API imposes restrictions on data download (60 requests per minute, 100 objects per call) =>  As complete dataset is 1.7 billion objects, would take about 231 days to download
- We got a complete data dump of all Reddit posts going back to 01-2006 (via [http://files.pushshift.io/](http://files.pushshift.io/))
- Complete dataset with all submissions for all 10 years from all subreddits is 74.1 GB !!! => Azure
- Download all files using a bash script in a Virtual Machine and upload them to Blob Storage



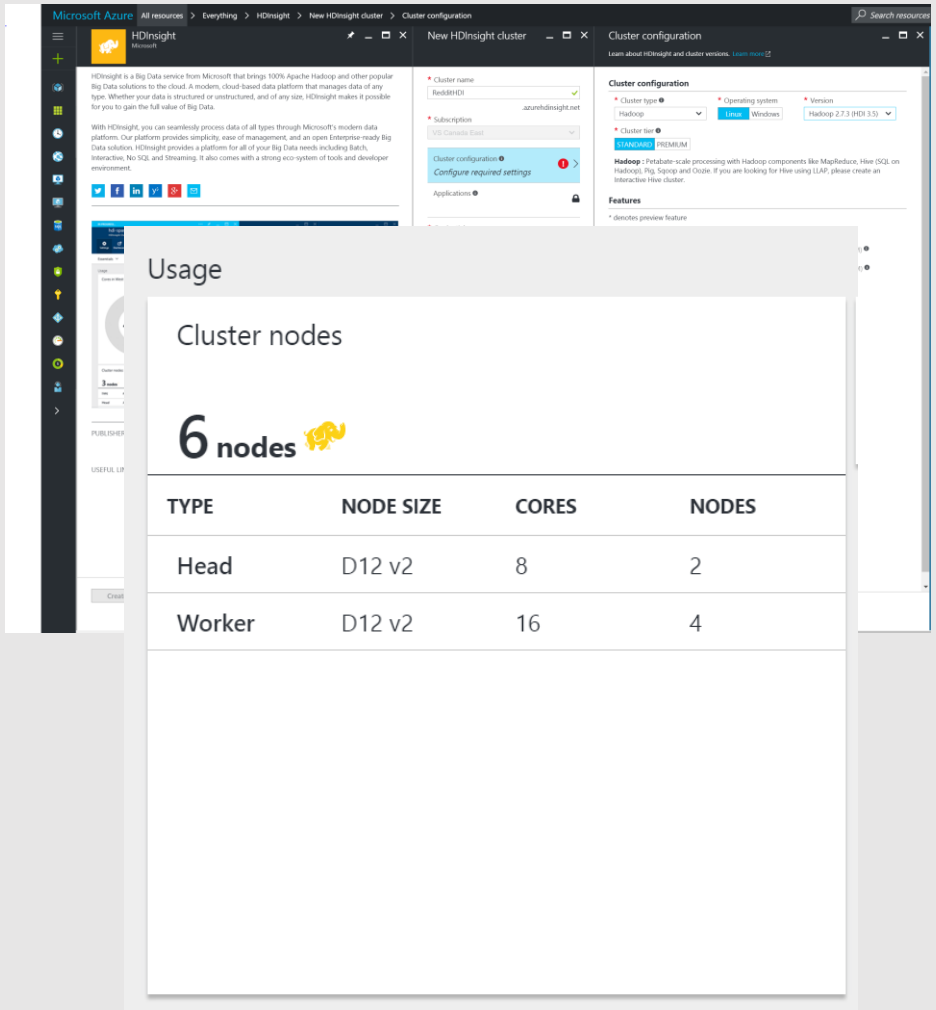
# Data Acquisition – Raw Format

- Data is in monthly .bz2 compressed files in JSON format (one JSON doc per line)
- From each JSON document we only need to extract just a couple of fields:
  - subreddit = “worldnews”
  - title (news headline)
  - created_utc (date-time) => year
  - score, ups, downs
- Put all this data in a Hadoop Cluster with Hive to query efficiently



# Data Acquisition –Processing in HDInsight Hadoop Cluster

- Azure HDInsight Cluster can be created ad-hoc, hourly cost depends on number of nodes and node size
- Cluster creation is done through Web UI
- Used smaller size cluster for modelling and development, larger cluster for actual query



# Data Acquisition –Processing in HIVE

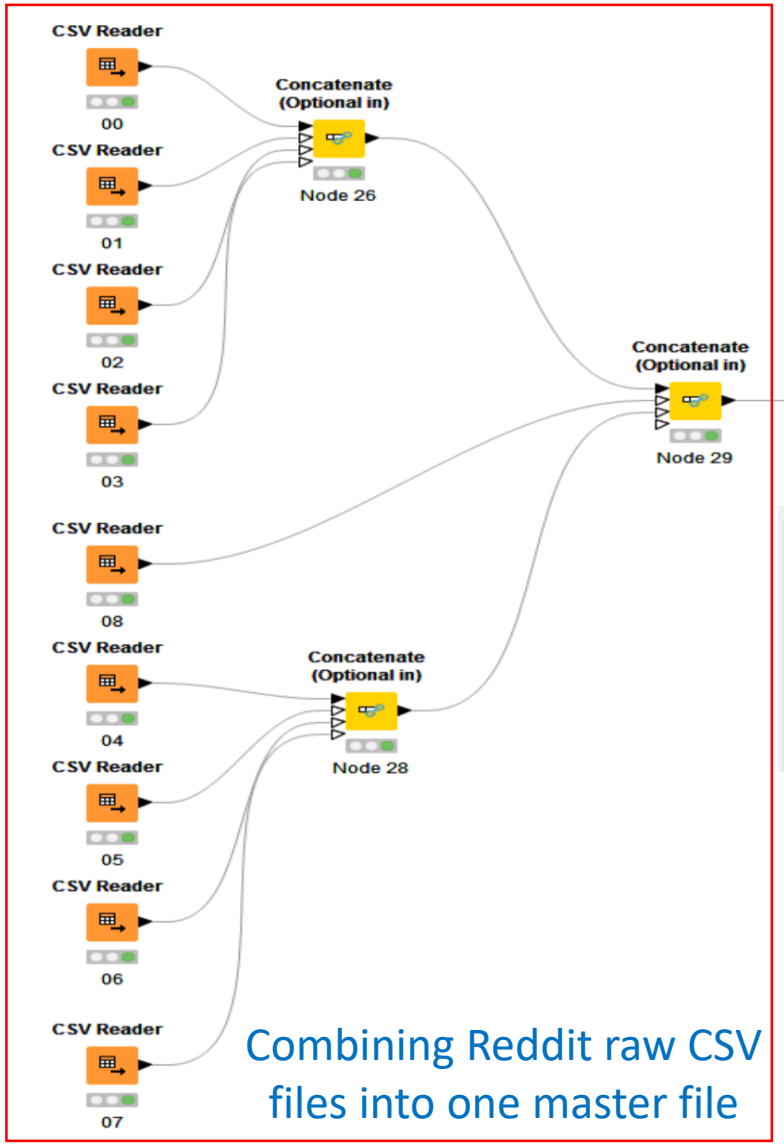
- 1) Create external tables to read compressed .bz2 JSON data files that read raw text line by line
- 2) Create ORC (columnar store) HIVE compressed tables
- 3) Parse JSON from raw external tables to move just required fields from “worldnews” submissions into ORC files for faster querying
  - Submission_year
  - Subreddit
  - Submission_Date
  - Title
  - Score
  - Ups
  - downs
- 4) Export into CSV format
  - Remove \t, \n
  - Add quotes to strings

```
CREATE DATABASE IF NOT EXISTS reddit;

USE reddit;

CREATE EXTERNAL TABLE IF NOT EXISTS reddit.submissions_raw
(
  value STRING
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' LINES TERMINATED BY '\n'
LOCATION 'wasb://subtest@bnajlis2slalom.blob.core.windows.net/';
```

```
INSERT INTO reddit.submissions_orc(submission_year, subreddit, submission_date, title, score, ups, downs)

SELECT
    CAST(from_unixtime(CAST(v.created_utc AS BIGINT), 'yyyy') AS SMALLINT) as submission_year,
    v.subreddit,
    from_unixtime(CAST(v.created_utc AS BIGINT), 'yyyy-MM-dd') as submission_date,
    v.title,
    v.score,
    v.ups,
    v.downs
FROM reddit.submissions_raw jt
     LATERAL VIEW json_tuple(jt.value, 'subreddit', 'created_utc', 'title', 'score', 'ups', 'downs') v
     AS subreddit, created_utc, title, score, ups, downs
WHERE v.subreddit = "worldnews";
```

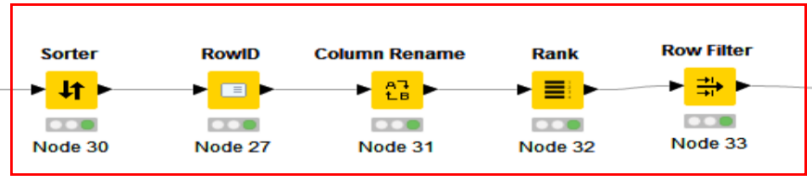
```
INSERT OVERWRITE DIRECTORY
'/user/admin/reddit_worldnews'
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
WITH SERDEPROPERTIES (  "separatorChar" = " ",
                        "quoteChar"  = "'")
SELECT
    submission_date,
    regexp_replace(title, "\n|\t", "") as title,
    score
FROM reddit.submissions_orc;
```

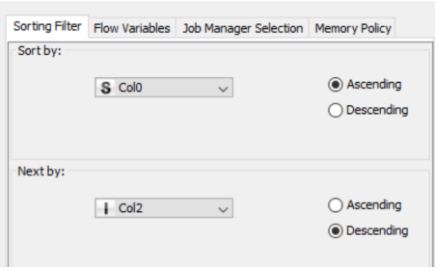
# Data Acquisition –Ingestion into KNIME

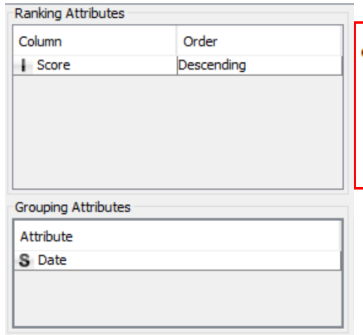



Combining Reddit raw CSV files into one master file

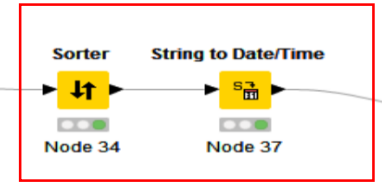
Selecting top rated posts

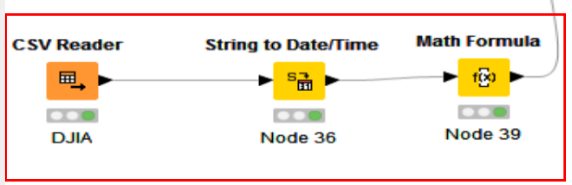


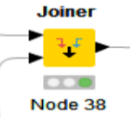



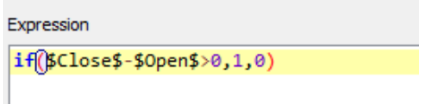




Preparing to join



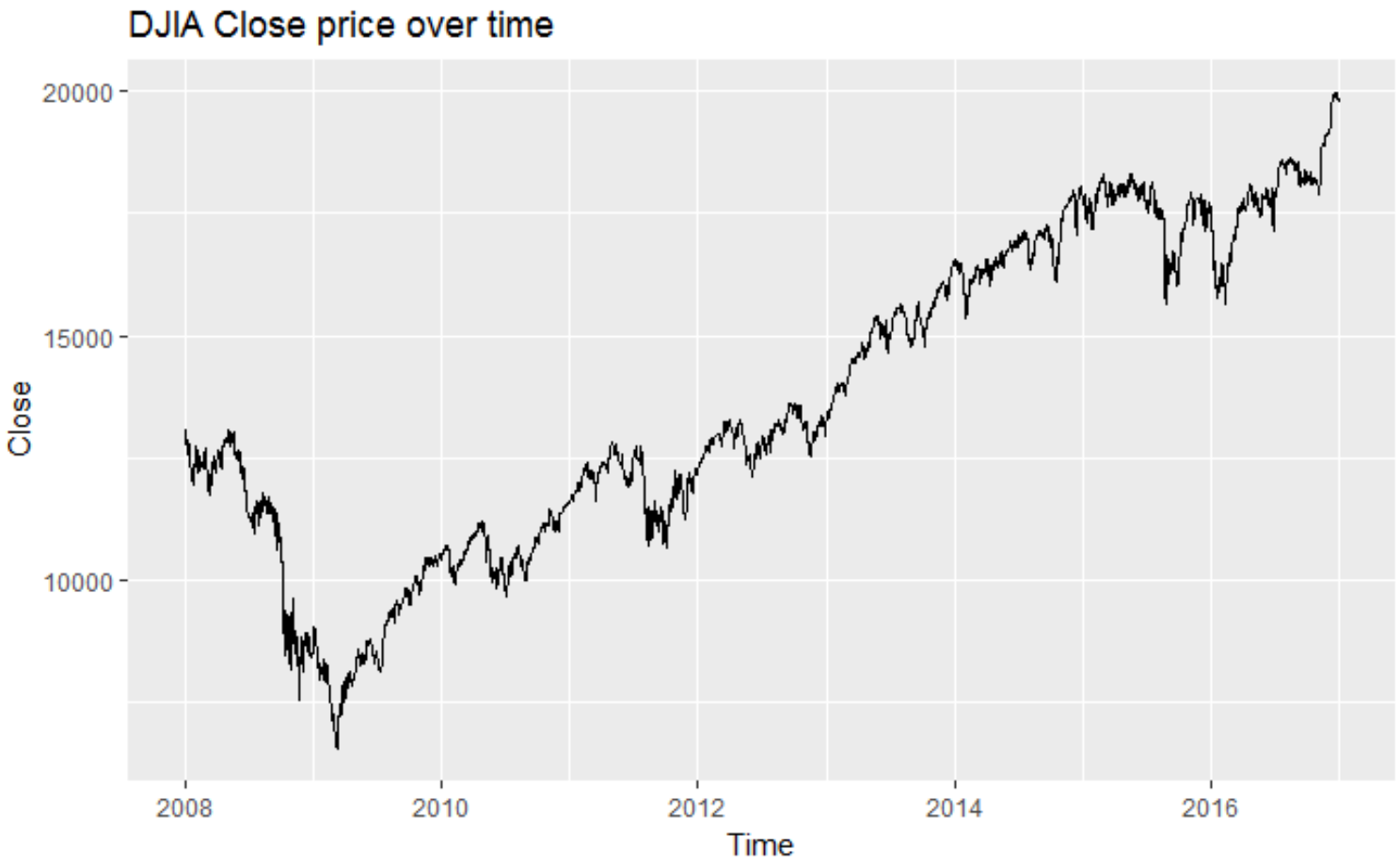

Ingesting DJIA index data
&
Assigning the labels

# Descriptive Analytics

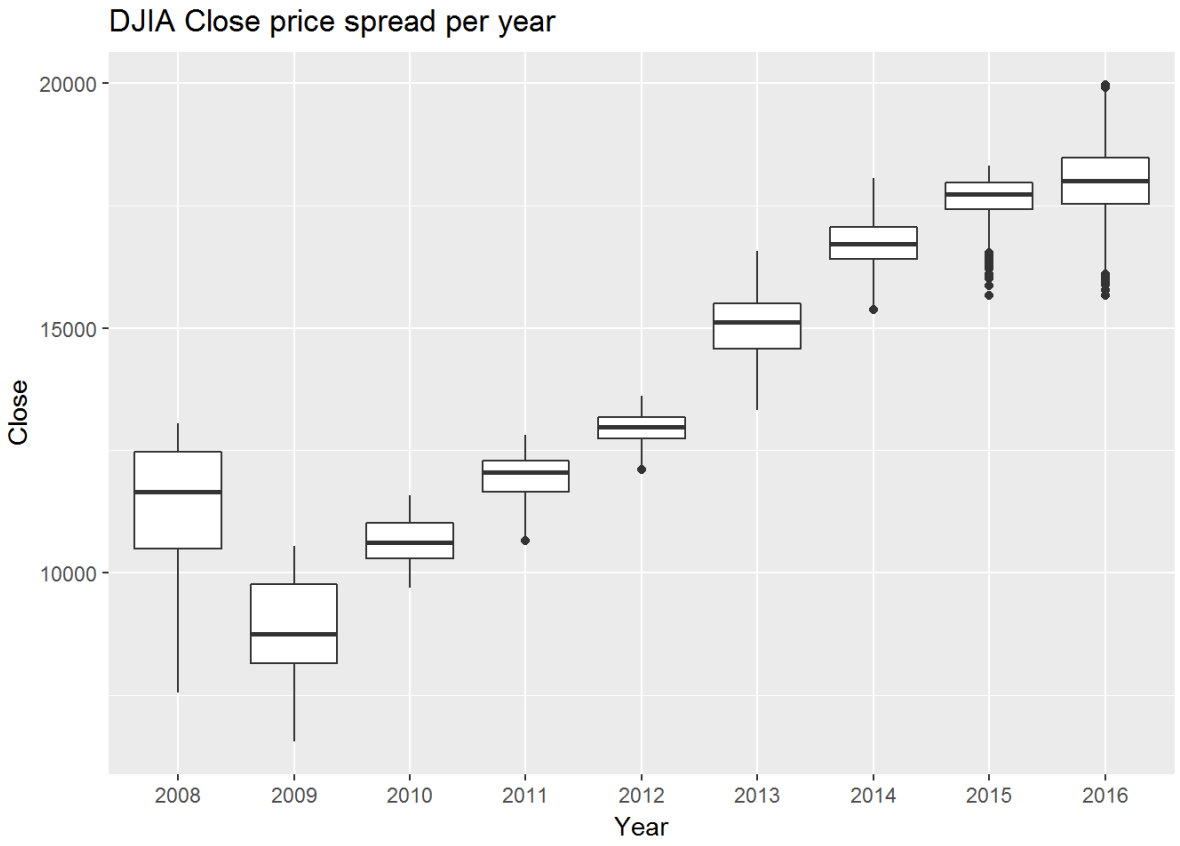
Exploratory Data Analysis

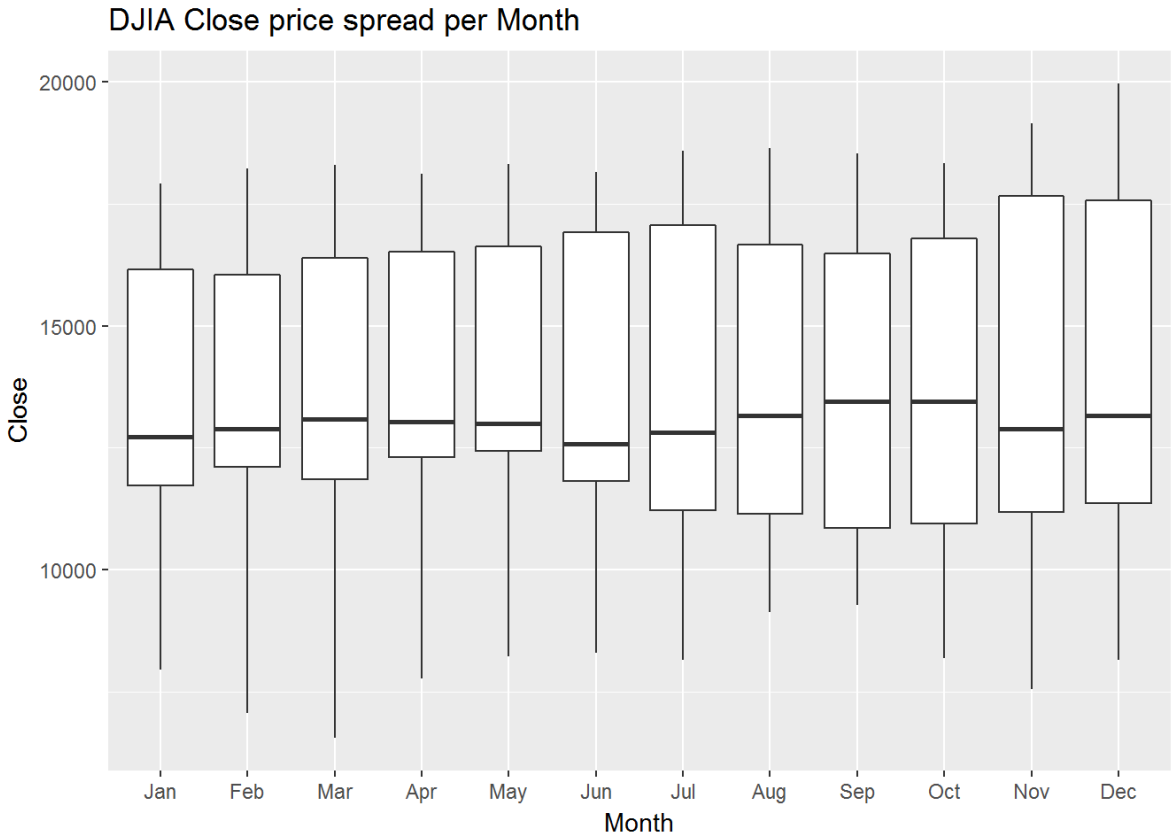
# Descriptive Analytics – DJIA (1)

- Index values for 2,265 days (from 2008-01-02 to 2016-12-30)
- Index values ranging from 6,547.05 to 19,974.62 points
- The lowest close price was recorded on 2009-03-09, and the highest on 2016-12-20

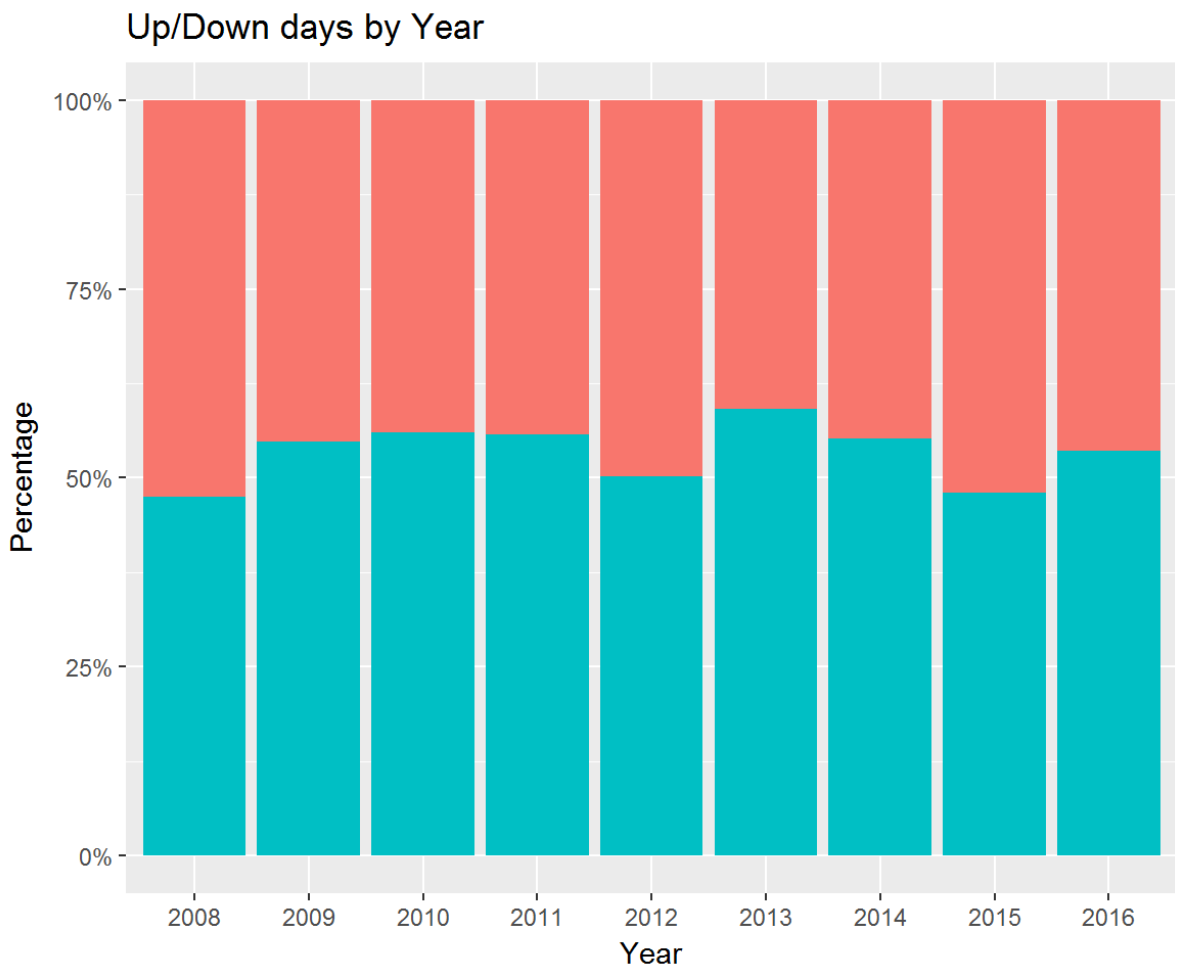


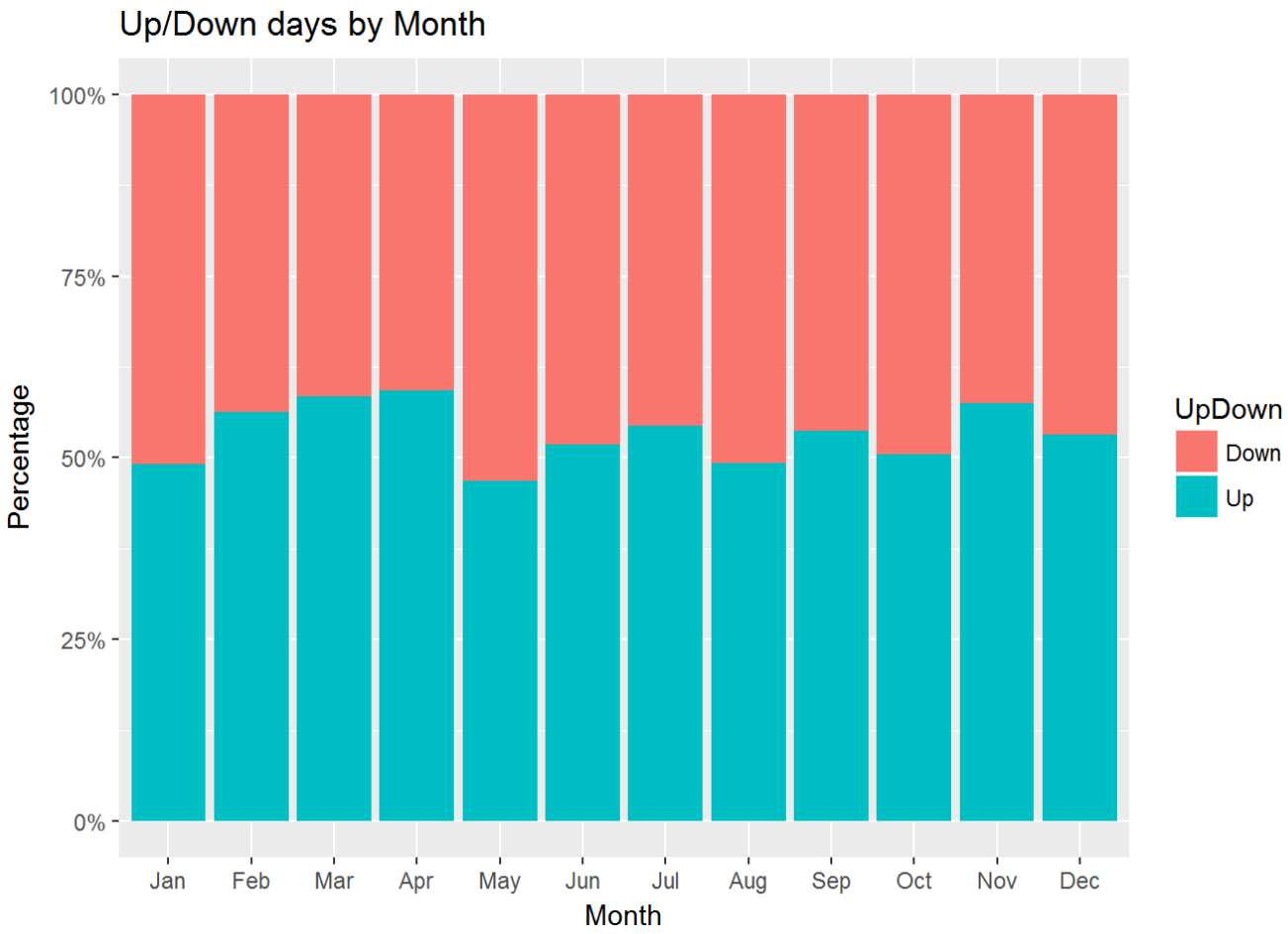
# Descriptive Analytics – DJIA (2)



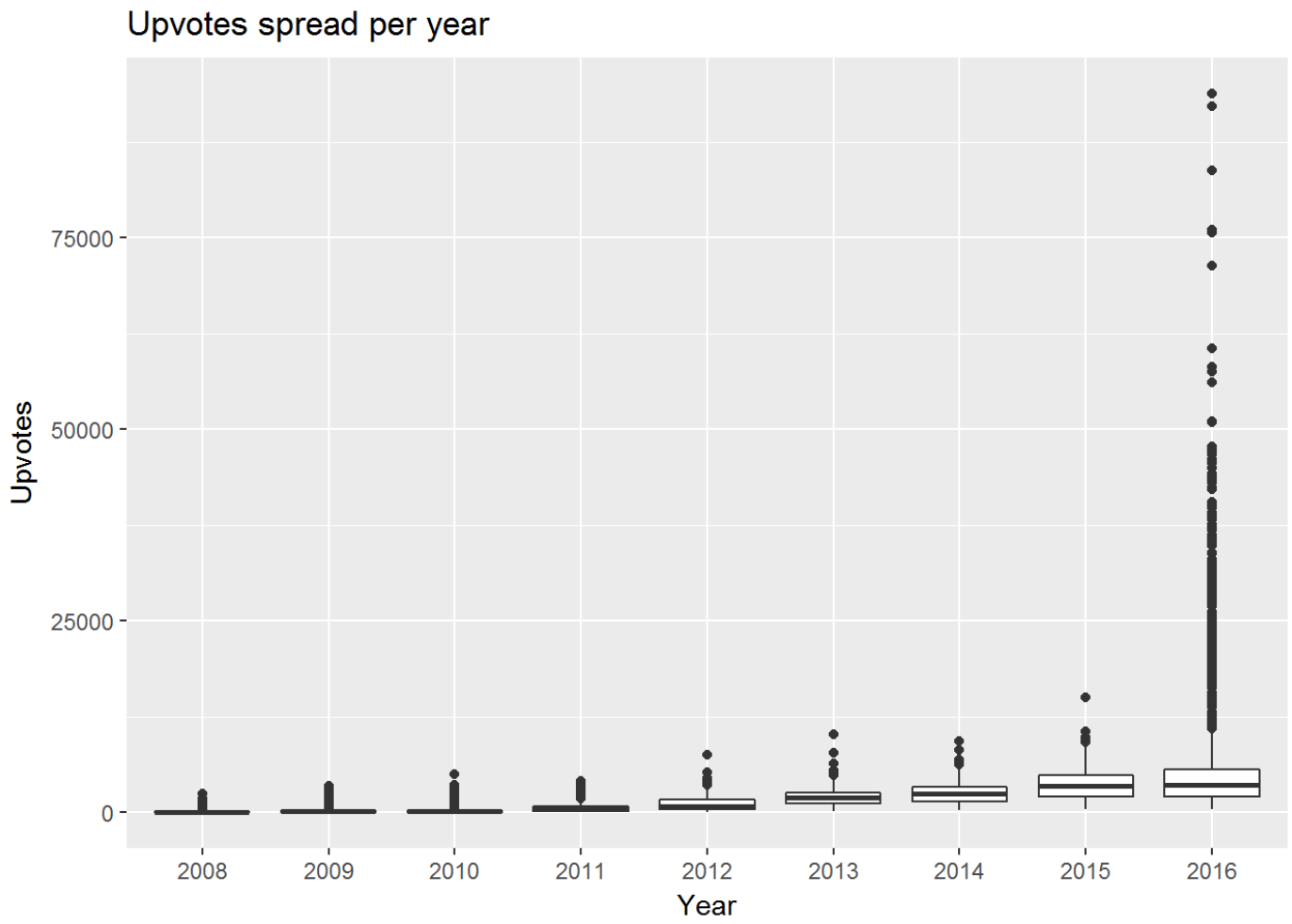
# Descriptive Analytics – DJIA (3)



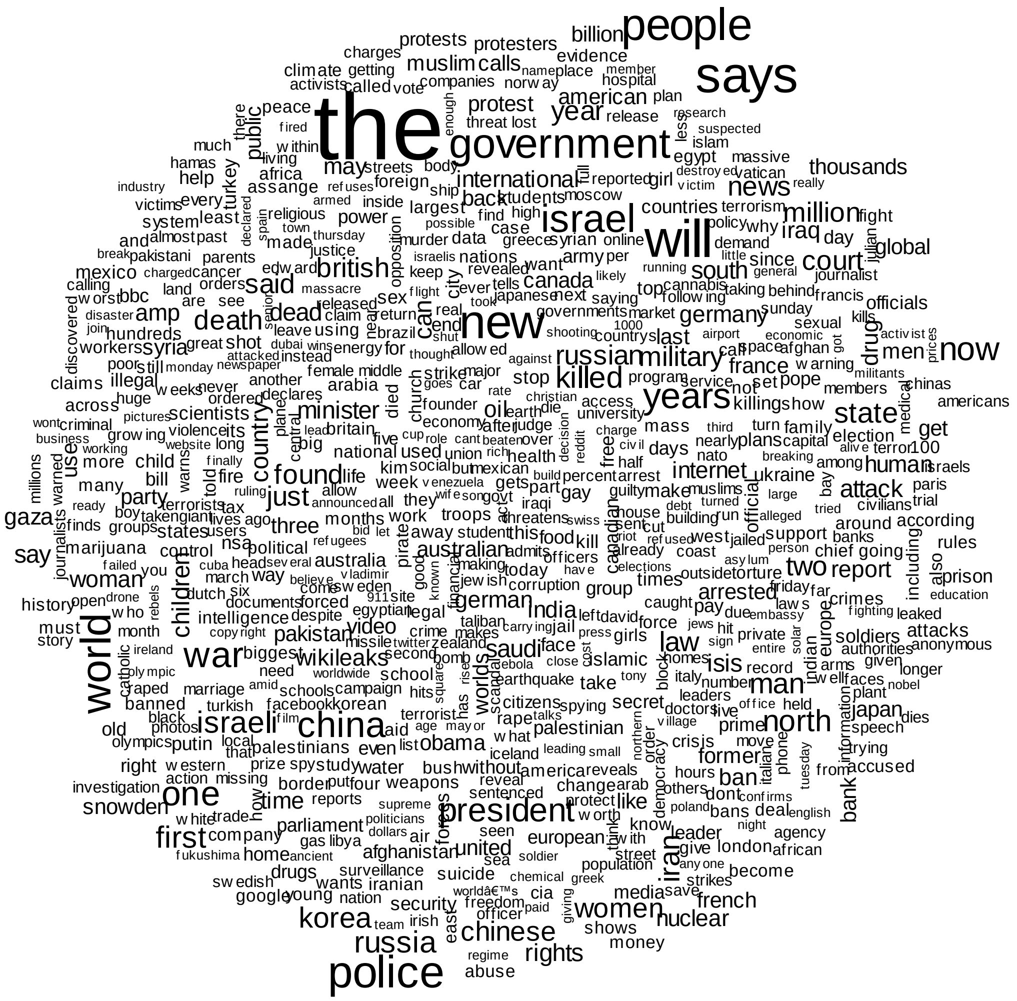



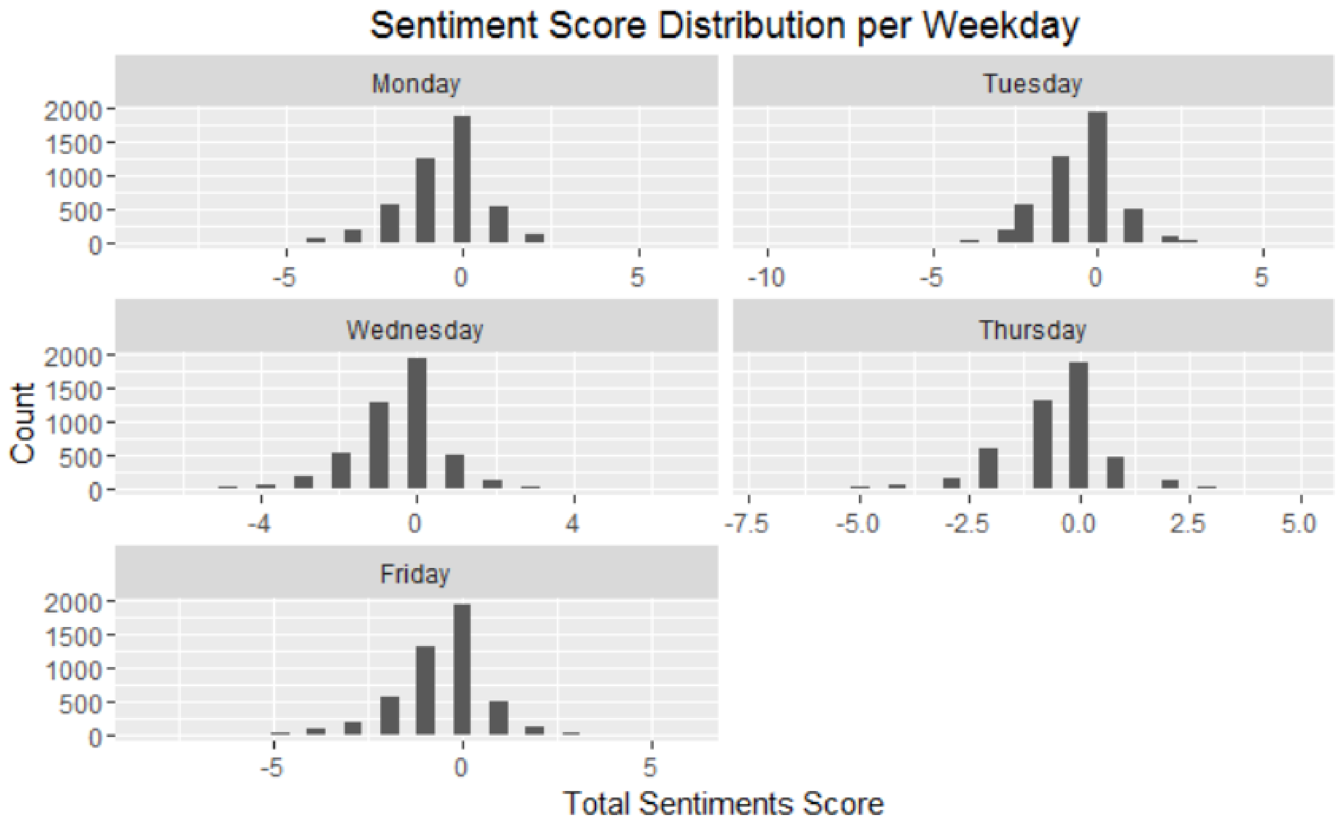
# Descriptive Analytics – Reddit /r/worldnews

- The complete Reddit **/r/worldnews** dataset has 2,018,344 headlines
- We limited our dataset to a maximum of 10 headlines per day based on the upvoting score
- Headlines range from 2008-01-25 to 2016-12-31, with scores (upvotes) varying from 0 to 93,832.
- The highest scored headline on 2016-11-26 is :"**Fidel Castro is dead at 90**.".
- Sample 0-scored news:
    - "Avalanche Kills TV Star Christopher Allport"
    - "Immunizations"
    - "WHO to recommend ways to reduce harm of alcohol
    - "Nicolas Sarkozy and Carla Bruni marry "



# Descriptive Analytics – Reddit (2)





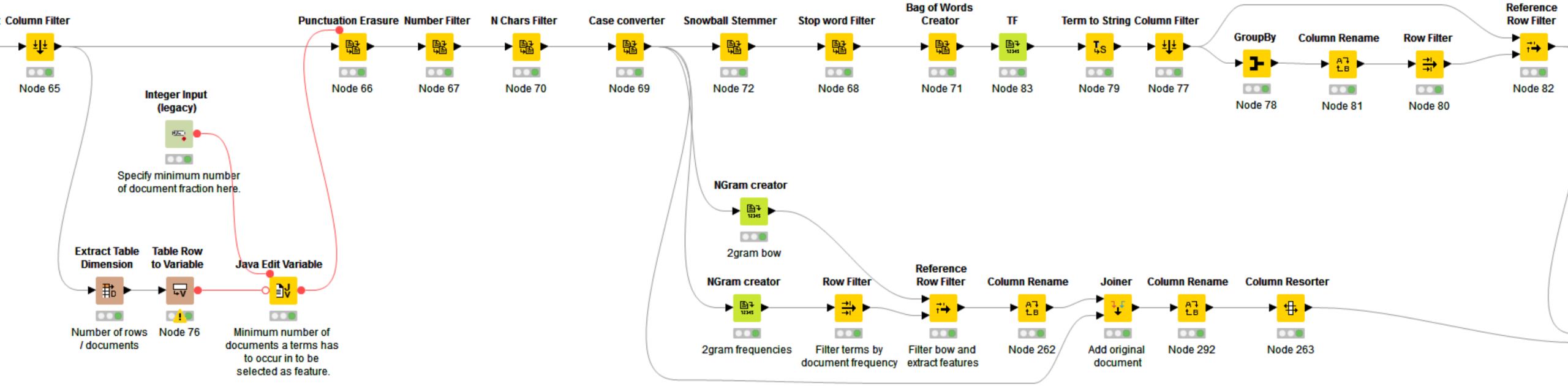
# Predictive Modeling

Models Description and Results

# Feature Engineering

- DJIA Dataset
  - Added Up/Down labels
    - if (Close – Open) > 0 => UP else => DOWN
- Reddit Dataset
  - Bag of words
  - Uni-gram / Bi-gram / N-gram
  - Pseudo TF-IDF
  - Headlines sentiment analysis scoring based on word dictionary

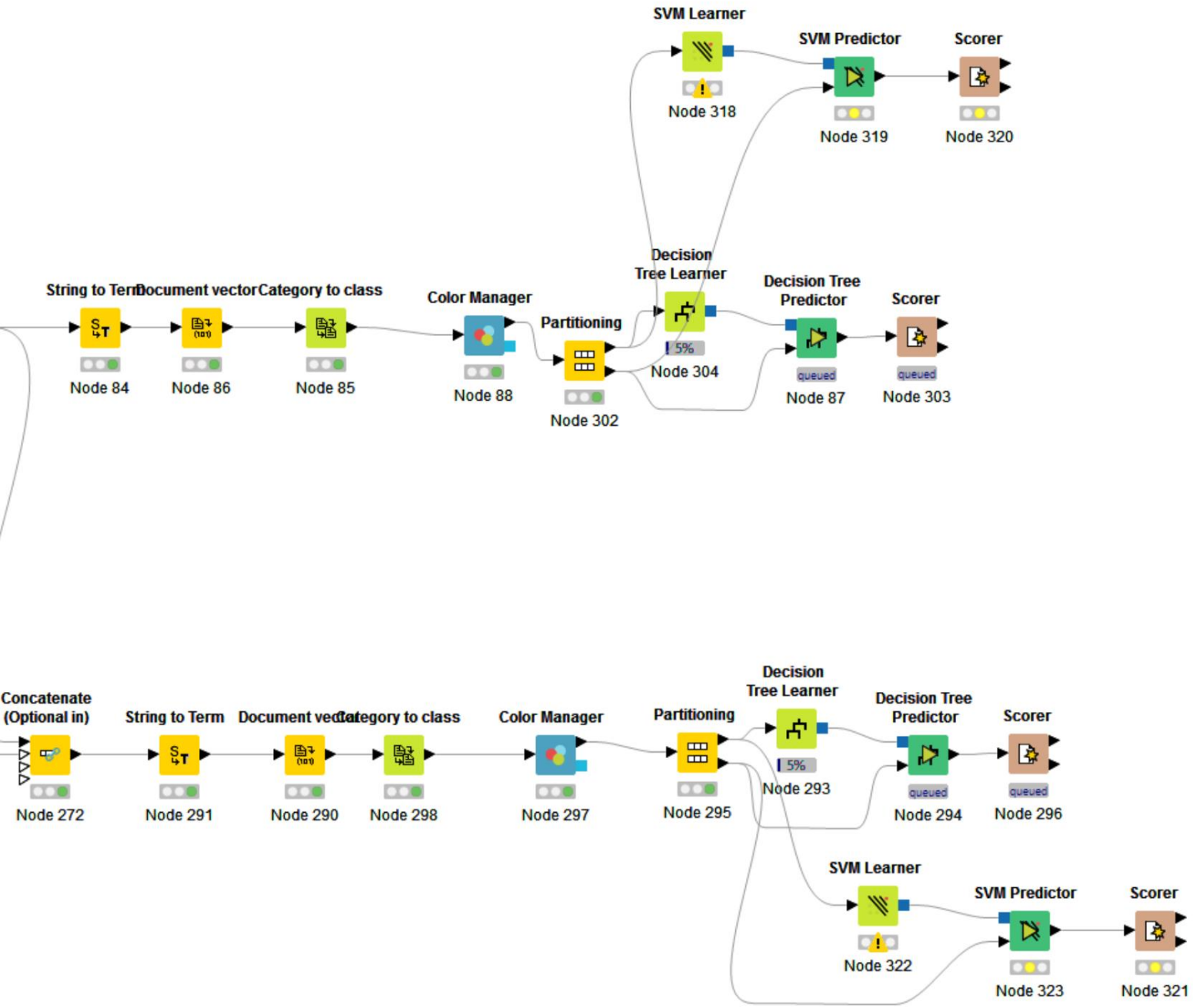
# Data Pre processing / Text Analytics in KNIME




- Feature Selection
  - Keep terms that occur in x docs
- Text analytics
  - Punctuation Erasure
  - Remove words less than N chars
  - Convert all to lower case
  - Stemming (Snowball package)
  - Stop word removal
  - Bag of words

- Ngram creator for 2gram / Frequency calculation per document / corpus
- Filter 2grams with low frequency

# Decision Trees and SVM Models in KNIME



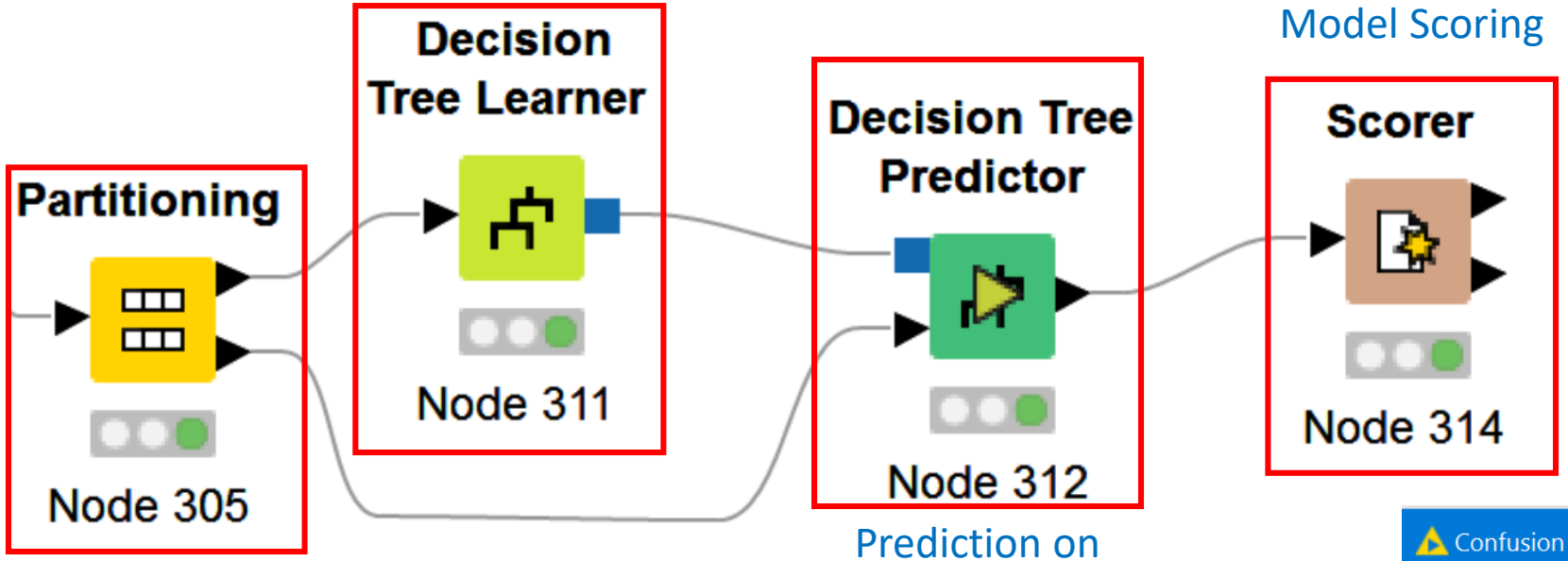

- 15 (14.5%) out of all 103 nodes in the workflow are for ML models
- 85.5% of the project is data processing, cleaning, feature creating

# Machine Learning Modelling in KNIME

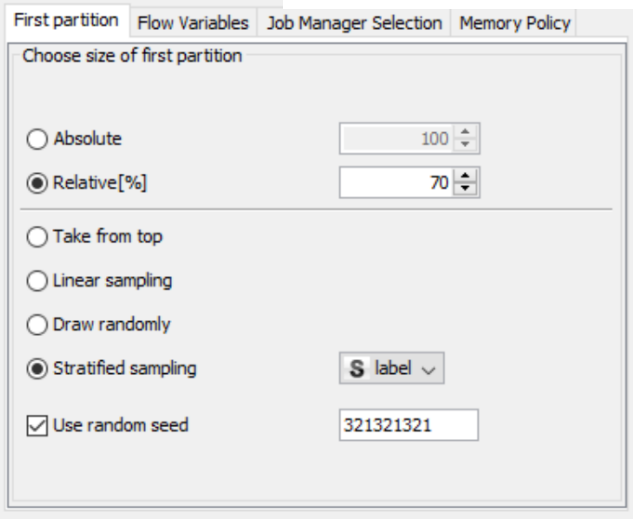
Model Training

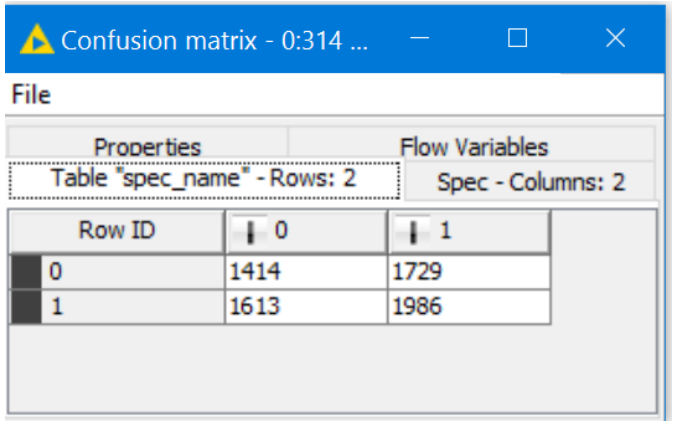
Model Scoring

Partition into Training and Test set




Prediction on Test Data




| Row ID | 0 | 1 |
|---|---|---|
| 0 | 1414 | 1729 |
| 1 | 1613 | 1986 |

| Row ID | TruePositives | FalsePositives | TrueNegatives | FalseNegatives | Recall | Precision | Sensitivity | Specifity | F-measure | Accuracy | Cohen's kappa |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1414 | 1613 | 1986 | 1729 | 0.45 | 0.467 | 0.45 | 0.552 | 0.458 | ? | ? |
| 1 | 1986 | 1729 | 1414 | 1613 | 0.552 | 0.535 | 0.552 | 0.45 | 0.543 | ? | ? |
| Overall | ? | ? | ? | ? | ? | ? | ? | ? | ? | 0.504 | 0.002 |

# Models Comparison - Accuracy

| Decision Trees | Uni-gram | Bi-gram | Sentiment Analysis |
|---|---|---|---|
| Top 1 | 0.503 | 0.492 | 0.487 |
| Top 3 | 0.493 | 0.5 | 0.506 |
| Top 5 | 0.503 | 0.509 | 0.503 |
| Top 10 | 0.502 | 0.501 | 0.504 |

| SVM | Uni-gram | Bi-gram |
|---|---|---|
| Top 1 | 0.517 | 0.495 |
| Top 3 | 0.511 | 0.507 |
| Top 5 | 0.509 | 0.501 |
| Top 10 | 0.507 | 0.497 |

# Conclusions

Lessons learned, conclusions and next steps

# Lessons Learned - Technical

- Azure HDInsight, Hive
- Knime and Knime/R integration
- Ggplot2 for descriptive analytics
- R limitations with “large” datasets
- Everything takes time …. A lot of time ….

# Lessons Learned - Analytics

- Data Science problem framing
- Bag of Words vs N-grams
- Pseudo TF-IDF
- Sentiment Analysis “as dimensionality reduction”
- Decision Trees
- SVM

# Conclusions

- Market efficiency
  - Market data at the wrong level (Daily vs streaming)
  - News data at the wrong level (Global News vs Industry or Company)

- More data doesn’t equal to better results
  - Cannot apply market data from 2008 to predict 2016
  - If data is not correlated, more data will not fix the problem

- Have realistic expectations about data analytics outcomes
  - Data availability is critical
  - Data processing takes a long time

# Next Steps

- Change data granularity
  - Streaming / hourly / by minute
  - Financial / Industry / Company focused news
- Use Deep Neural Network
- Use Hidden Markov Models
- Create models on a per-year basis