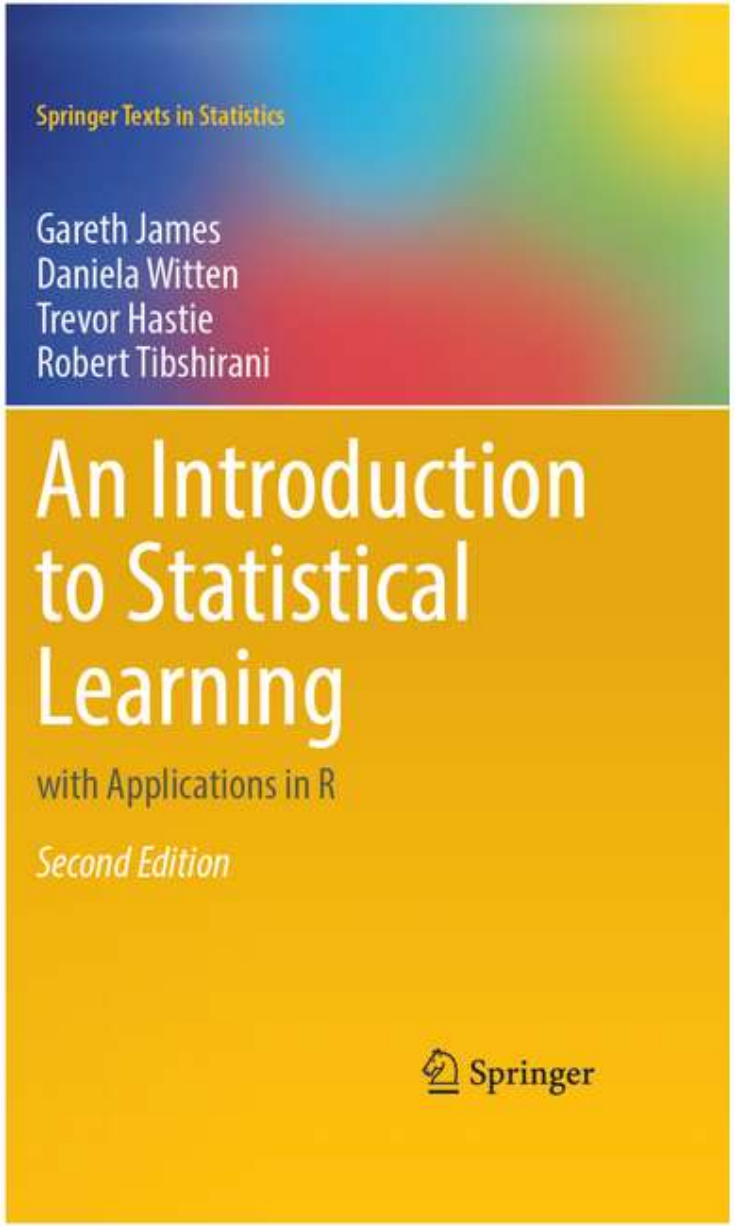
SESSION 5.1: UNSUPERVISED LEARNING

DR. SOFIA GIL-CLAVEL

- The basics of Statistical Learning
- Unsupervised Methods
- Unsupervised Methods for Text

WE WILL BE FOLLOWING:

You can find the book for free here:
https://www.statlearning.com/



Springer Texts in Statistics

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R

Second Edition

Springer

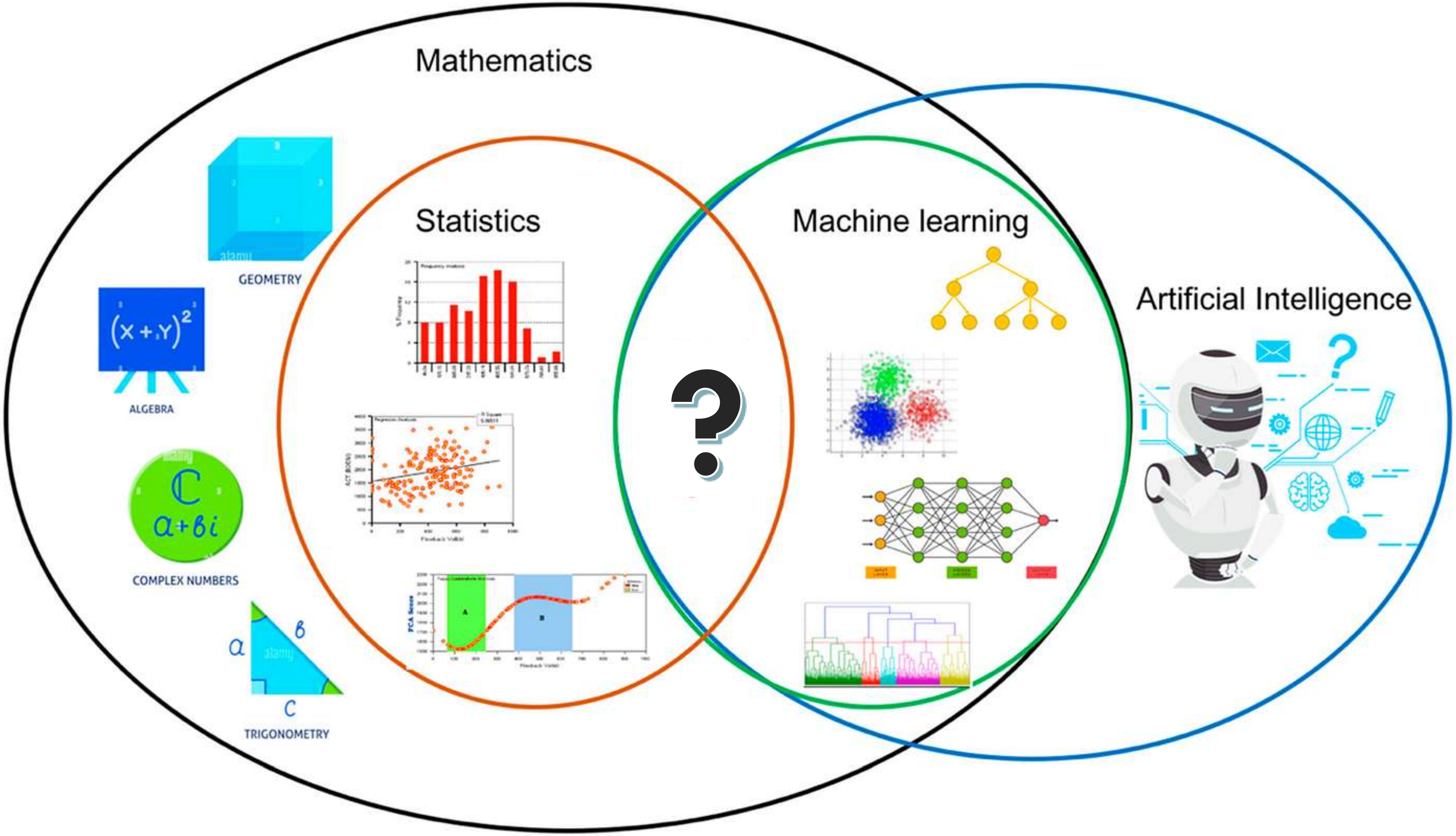
1. THE BASICS OF STATISTICAL LEARNING

1.1 Statistical learning

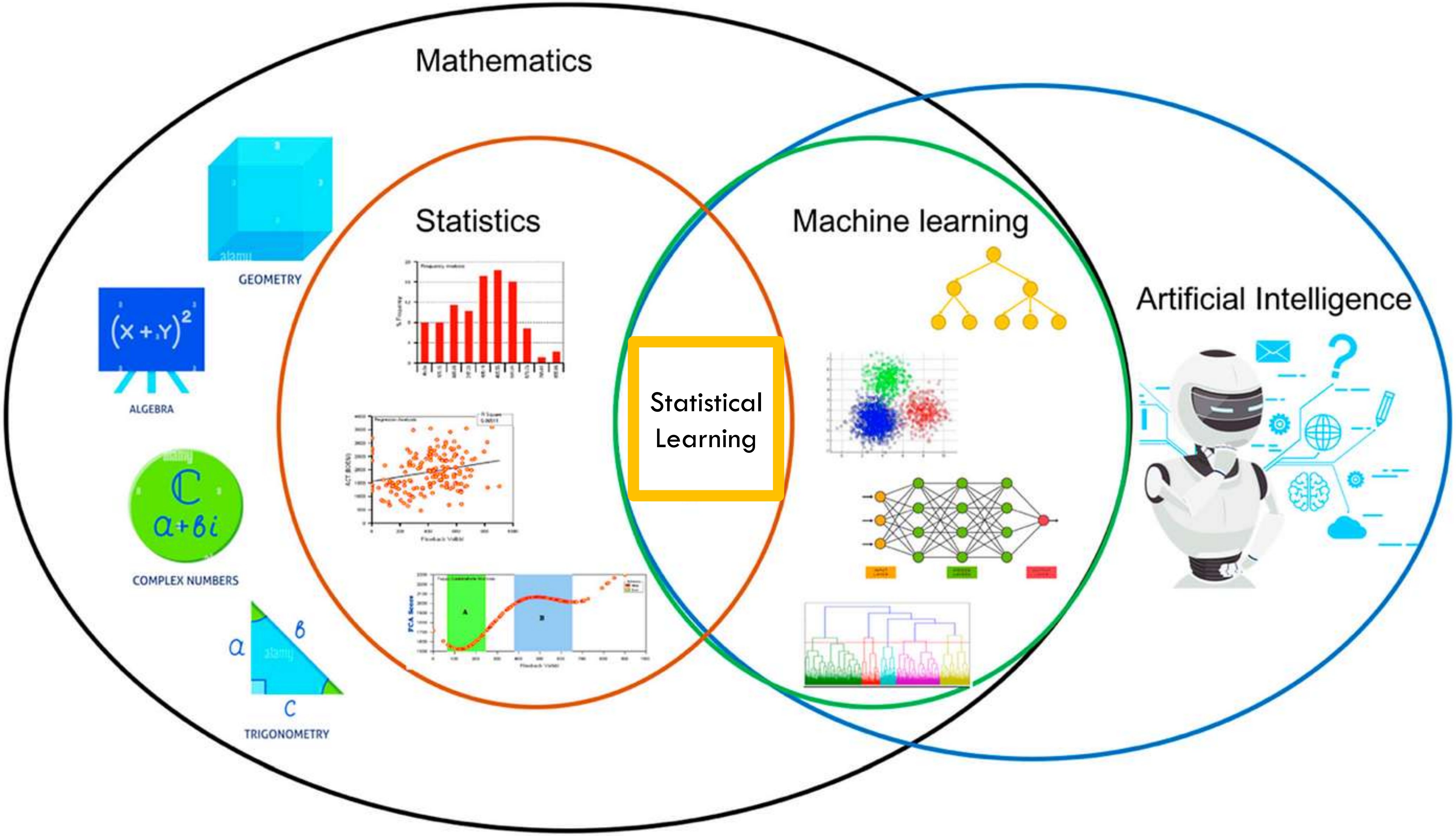
1.2 Supervised vs. Unsupervised

1.3 Overview of unsupervised methods

1.1 STATISTICAL LEARNING



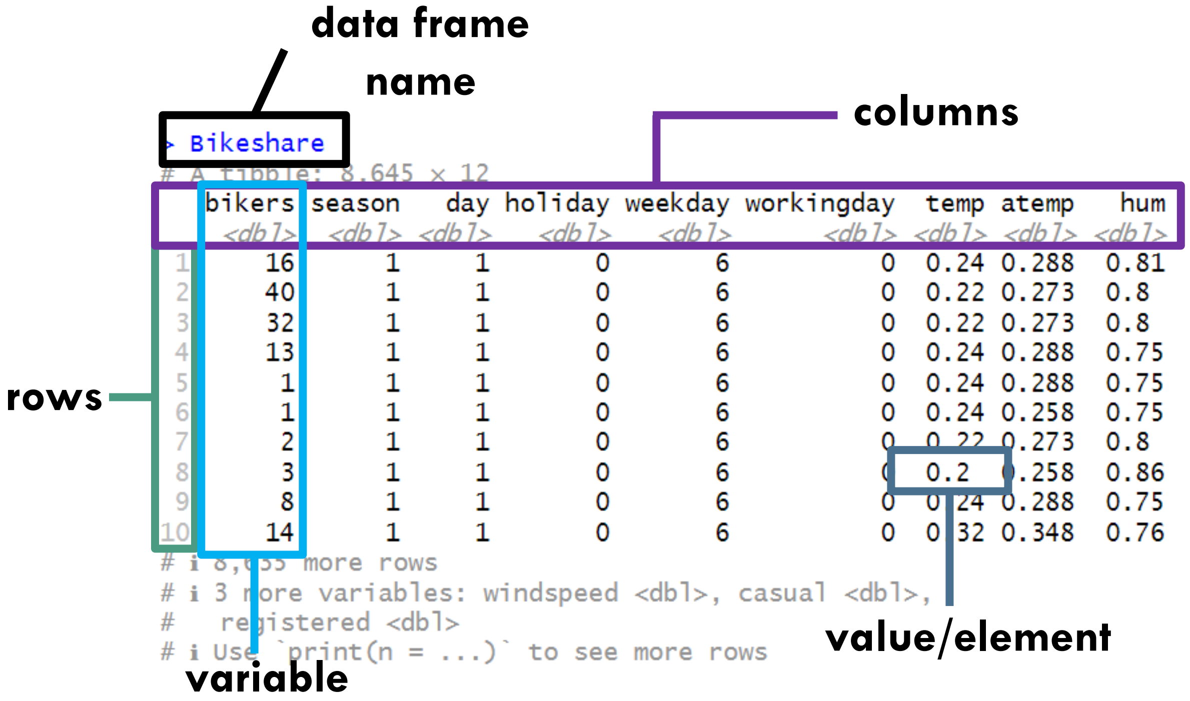
Original source: https://medium.com/stats-learning/differences-and-synergies-between-statistics-and-machine-learning-90cea85d4cf5



Original source: https://medium.com/stats-learning/differences-and-synergies-between-statistics-and-machine-learning-90cea85d4cf5

WHY STATISTICAL LEARNING?

So far in these workshops we have learned to handle data frames in R.

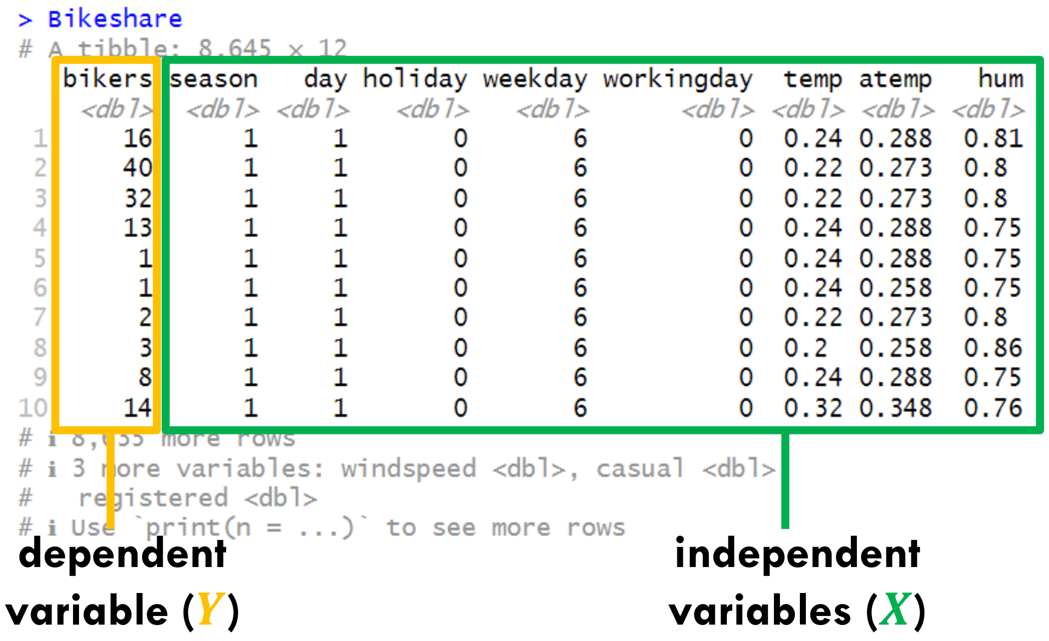


```
> Bikeshare
# A tibble: 8,645 × 12
   bikers season   day holiday weekday workingday  temp atemp   hum
    <dbl>  <dbl> <dbl>   <dbl>   <dbl>      <dbl> <dbl> <dbl> <dbl>
 1     16      1     1       0       6          0  0.24 0.288  0.81
 2     40      1     1       0       6          0  0.22 0.273  0.8
 3     32      1     1       0       6          0  0.22 0.273  0.8
 4     13      1     1       0       6          0  0.24 0.288  0.75
 5      1      1     1       0       6          0  0.24 0.288  0.75
 6      1      1     1       0       6          0  0.24 0.258  0.75
 7      2      1     1       0       6          0  0.22 0.273  0.8
 8      3      1     1       0       6          0  0.2  0.258  0.86
 9      8      1     1       0       6          0  0.24 0.288  0.75
10     14      1     1       0       6          0  0.32 0.348  0.76
# ℹ 8,635 more rows
# ℹ 3 more variables: windspeed <dbl>, casual <dbl>,
#   registered <dbl>
# ℹ Use `print(n = ...)` to see more rows
```

WHY STATISTICAL LEARNING?

So far in these workshops we have learned to handle data frames in R. This is very convenient, as we can analyze these data frames using statistical symbolic language!

- **Dependent** (*Y*): Also known as response variable.
- **Independent** (*X*): Also known as input, predictor, feature, or just variable.



```
> Bikeshare
# A tibble: 8,645 × 12
   bikers season   day holiday weekday workingday  temp atemp   hum
    <dbl>  <dbl> <dbl>   <dbl>   <dbl>      <dbl> <dbl> <dbl> <dbl>
 1     16      1     1       0       6          0  0.24 0.288  0.81
 2     40      1     1       0       6          0  0.22 0.273  0.8 
 3     32      1     1       0       6          0  0.22 0.273  0.8 
 4     13      1     1       0       6          0  0.24 0.288  0.75
 5      1      1     1       0       6          0  0.24 0.288  0.75
 6      1      1     1       0       6          0  0.24 0.258  0.75
 7      2      1     1       0       6          0  0.22 0.273  0.8 
 8      3      1     1       0       6          0  0.2  0.258  0.86
 9      8      1     1       0       6          0  0.24 0.288  0.75
10     14      1     1       0       6          0  0.32 0.348  0.76
# ℹ 8,635 more rows
# ℹ 3 more variables: windspeed <dbl>, casual <dbl>,
#   registered <dbl>
# ℹ Use `print(n = ...)` to see more rows
```

dependent variable (*Y*)

independent variables (*X*)

WHY STATISTICAL LEARNING?

So far in these workshops we have learned to handle data frames in R. This is very convenient, as we can analyze these data frames using statistical symbolic language!

Symbolic Language

- **Dependent** (Y): Also known as response variable.
- **Independent** (X): Also known as input, predictor, feature, or just variable.

$$\underbrace{Y \sim X1 + X2 + X3}_{\text{R was built to understand symbolic language!}}$$

1.2 UNSUPERVISED VS. SUPERVISED

SUPERVISED STATISTICAL LEARNING

In supervised learning, for each observation of the predictor measurement(s) X_{i} there is an associated response measurement Y_i.

$$Y_i \sim X1_i + X2_i + X3_i$$

i refers to the data frame row

We wish to fit a model that relates the response to the predictors, with the aim of:

- Prediction: accurately predicting the response for future observations.
- Inference: better understanding the relationship between the response and the predictors.

Many classical statistical learning methods such as linear regression and logistic regression, as well as more modern approaches such as GAM, boosting, and support vector machines, operate in the supervised learning domain.

UNSUPERVISED STATISTICAL LEARNING

Unsupervised learning describes the somewhat more challenging situation in which for every observation i = 1,…, n, we observe a vector of measurements $\boldsymbol{X}_{\mathrm{i}}$ but no associated response Y_i.

$$\cancel{Y_i} \sim \boldsymbol{X1}_i + \boldsymbol{X2}_i + \boldsymbol{X3}_i$$

We are not interested in prediction, because we do not have an associated response variable Y. The goal is to discover interesting things about the measurements on X1, X2,…, Xp. Is there an informative way to visualize the data? Can we discover subgroups among the variables or among the observations?

Unsupervised learning is often performed as part of an exploratory data analysis.

2. UNSUPERVISED METHODS

1. Overview
2. Principal Components Analysis (PCA)
3. Visualizing dimensionality reduction
4. Using K-means
5. Visualizing Clusters

2.1 OVERVIEW OF UNSUPERVISED METHODS

WE WILL FOCUS ON TWO PARTICULAR TYPES OF UNSUPERVISED LEARNING:

- **Dimensionality reduction:** a tool to reduce the dimensionality of your data. It is used for data visualization or data pre-processing (e.g., data imputation) before supervised techniques are applied.
  - Principal Components Analysis (PCA): looks to find a low-dimensional representation of the observations that explain a good fraction of the variance.

- **Clustering:** Clustering refers to a very broad set of techniques for finding subgroups, or clustering clusters, in a data set. When we cluster the observations of a data set, we seek to partition them into distinct groups so that the observations within each group are quite similar to each other, while observations in different groups are quite different from each other.
  - K-means: we seek to partition the observations into a pre-specified number of clusters.

5 MINS BREAK

2.2 PRINCIPAL COMPONENTS ANALYSIS (PCA)

WHAT IS PRINCIPAL COMPONENTS?

When faced with a large set of **correlated variables**, **principal components allow us to summarize this set with a smaller number of** representative **variables** that collectively explain most of the variability in the original set.

THE “BIKESHARE” DATA

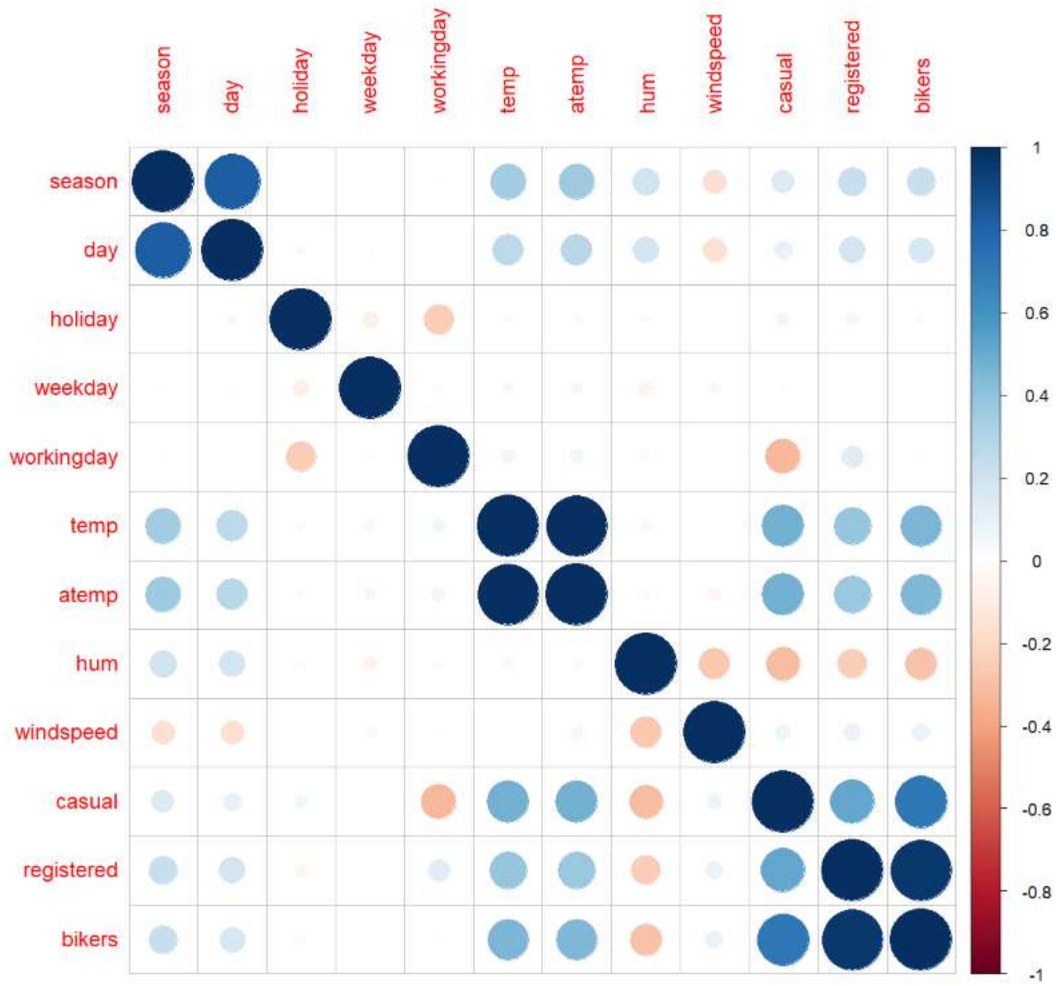
- **season:** Season of the year, coded as Winter=1, Spring=2, Summer=3, Fall=4.
- **mnth:** Month of the year, coded as a factor.
- **day:** Day of the year, from 1 to 365
- **hr:** Hour of the day, coded as a factor from 0 to 23.
- **holiday:** Is it a holiday? Yes=1, No=0.
- **weekday:** Day of the week, coded from 0 to 6, where Sunday=0, Monday=1, Tuesday=2, etc.
- **workingday:** Is it a work day? Yes=1, No=0.
- **weathersit:** Weather, coded as a factor.
- **temp:** Normalized temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=-8, t_max=+39.
- **atemp:** Normalized feeling temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=-16, t_max=+50.
- **hum:** Normalized humidity. The values are divided to 100 (max).
- **windspeed:** Normalized wind speed. The values are divided by 67 (max).
- **casual:** Number of casual bikers.
- **registered:** Number of registered bikers.
- **bikers:** Total number of bikers.

THE "BIKESHARE" DATA

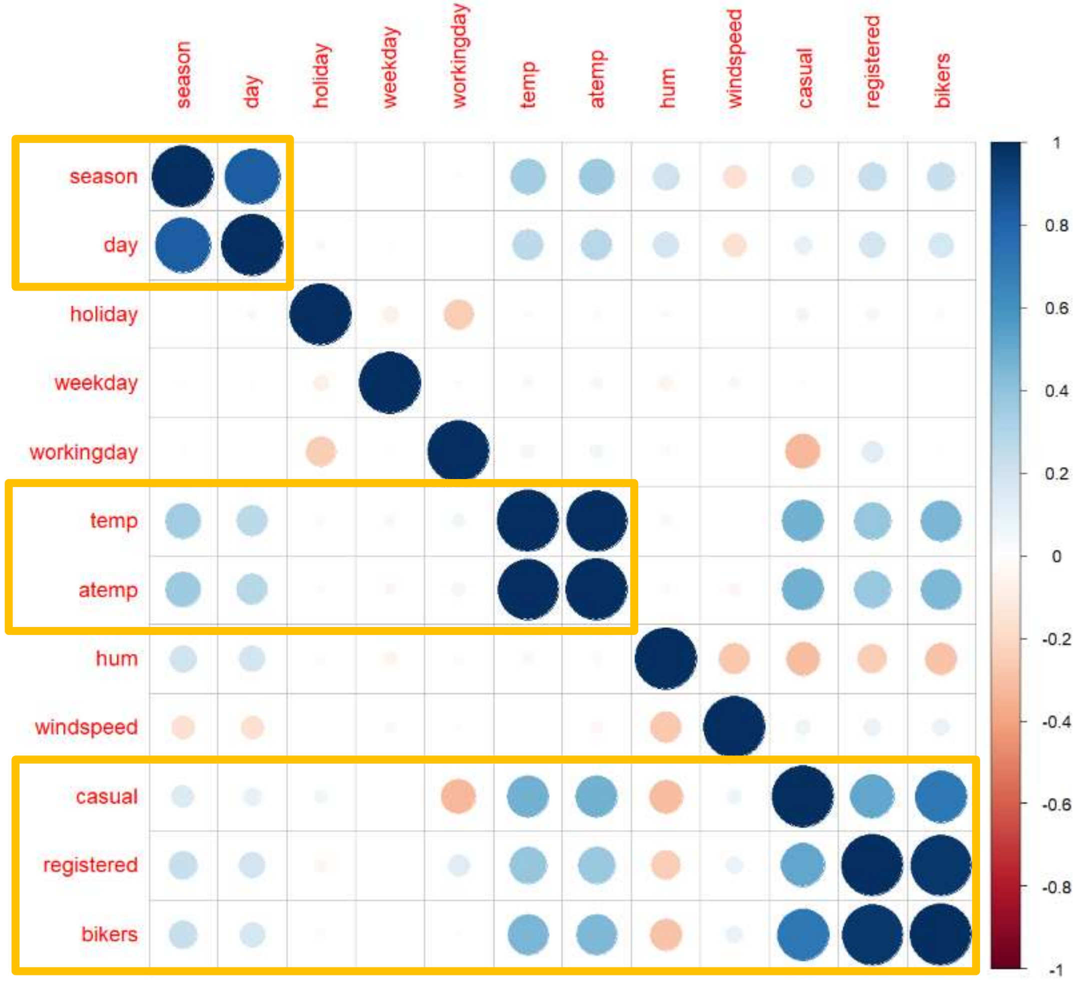
Take a minute to understand the variables. Which ones do you think are correlated?

- **season:** Season of the year, coded as Winter=1, Spring=2, Summer=3, Fall=4.
- **mnth:** Month of the year, coded as a factor.
- **day:** Day of the year, from 1 to 365
- **hr:** Hour of the day, coded as a factor from 0 to 23.
- **holiday:** Is it a holiday? Yes=1, No=0.
- **weekday:** Day of the week, coded from 0 to 6, where Sunday=0, Monday=1, Tuesday=2, etc.
- **workingday:** Is it a work day? Yes=1, No=0.
- **weathersit:** Weather, coded as a factor.
- **temp:** Normalized temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=-8, t_max=+39.
- **atemp:** Normalized feeling temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=-16, t_max=+50.
- **hum:** Normalized humidity. The values are divided to 100 (max).
- **windspeed:** Normalized wind speed. The values are divided by 67 (max).
- **casual:** Number of casual bikers.
- **registered:** Number of registered bikers.
- **bikers:** Total number of bikers.

THE "BIKESHARE" DATA



THE “BIKESHARE” DATA

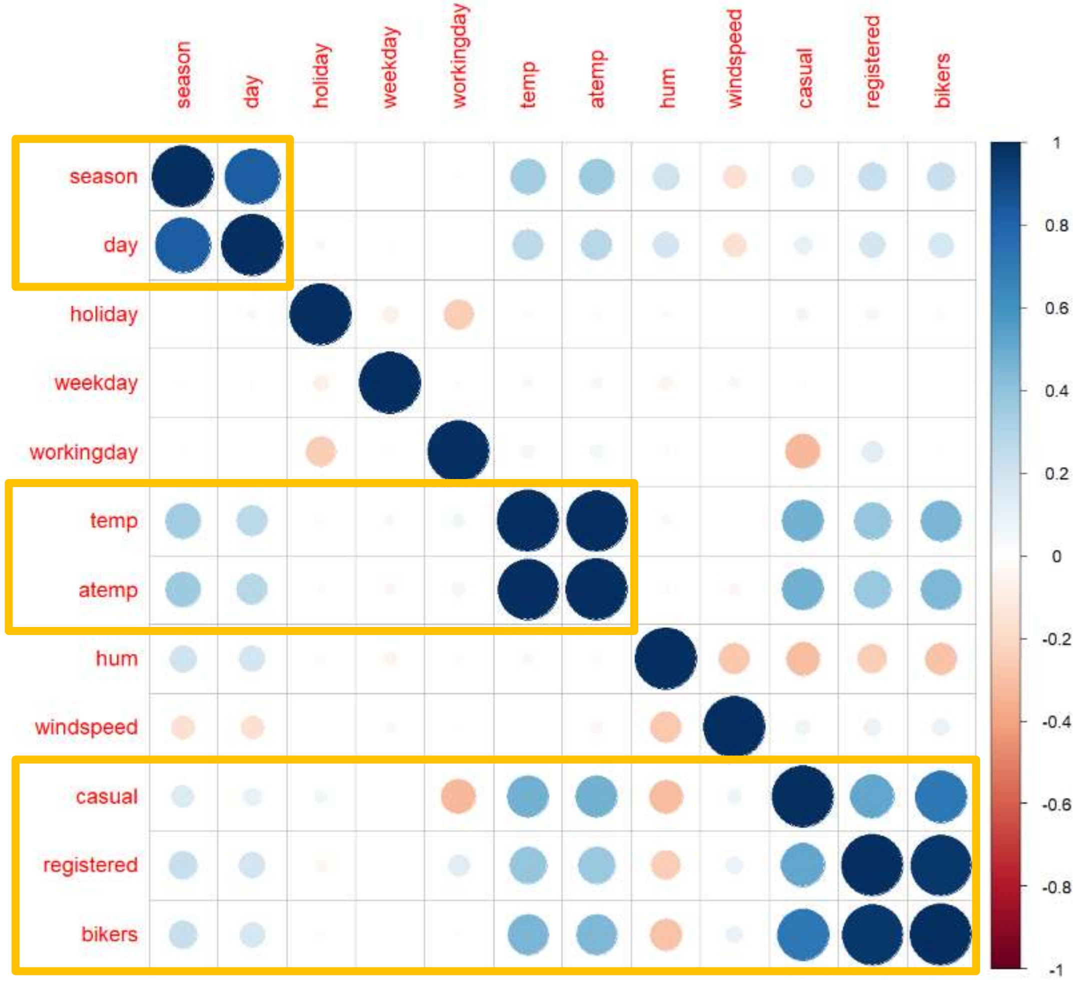


THE "BIKESHARE" DATA

Based on the definition

When faced with a large set of **correlated variables, principal components allow us to summarize this set with a smaller number of** representative **variables** that collectively explain most of the variability in the original set.

What do you expect to happen after using PCA?



THE "BIKESHARE" DATA

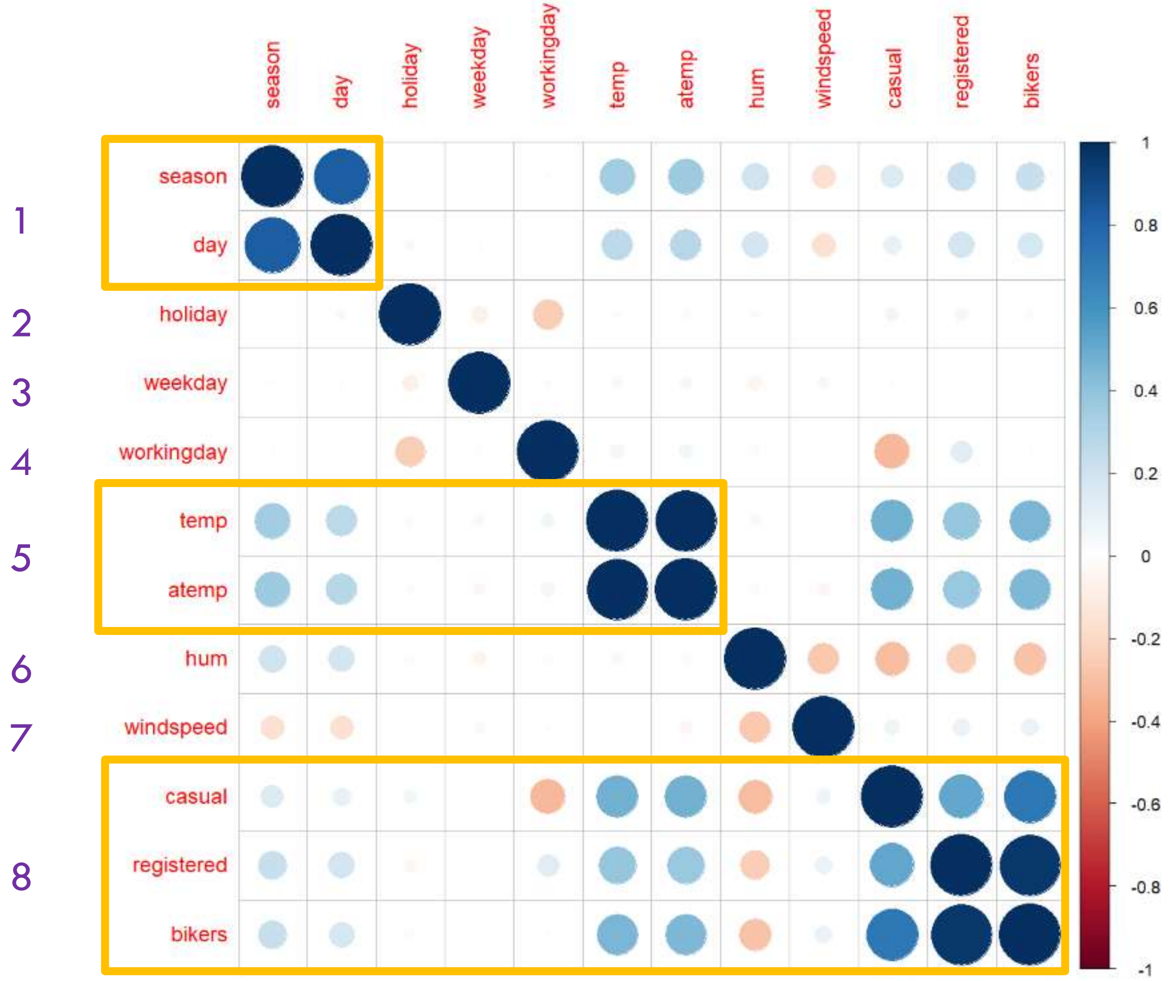
Based on the definition

↓

When faced with a large set of <u>**correlated variables**</u>**, principal components allow us to** <u>**summarize this set**</u> **with a smaller number of** representative **variables** that collectively explain most of the variability in the original set.

↓

What do you expect to happen after using PCA?

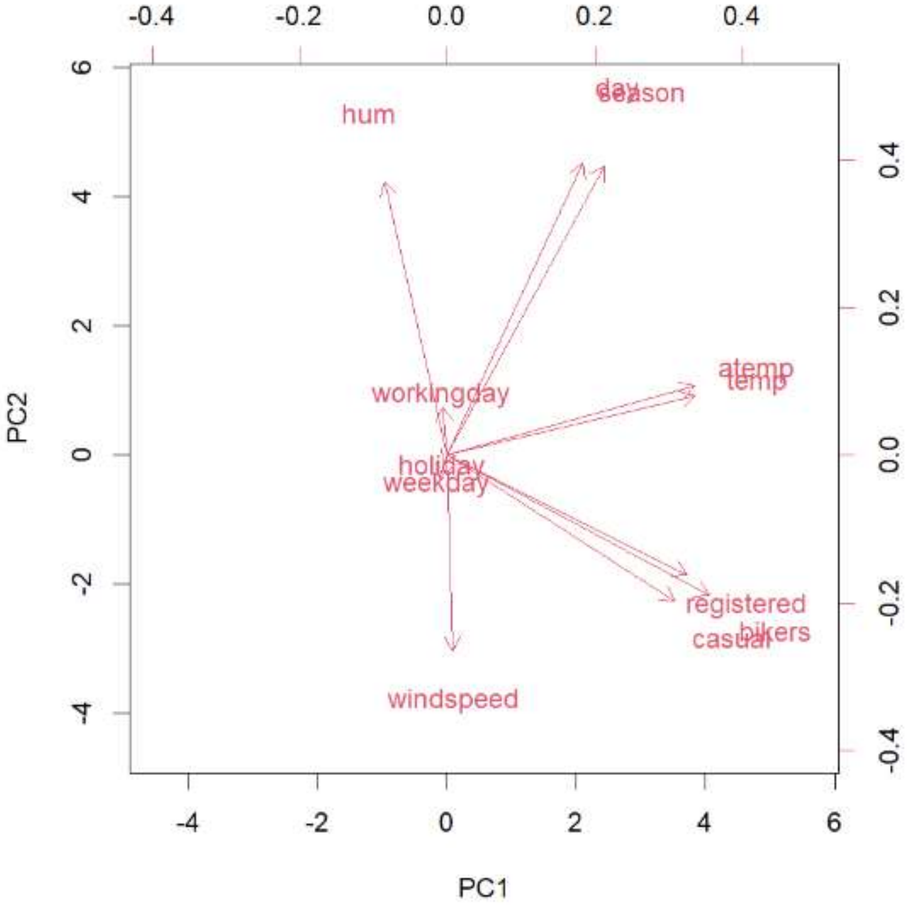


LET'S APPLY PCA TO THE DATA

2.3 VISUALIZING DIMENSIONALITY REDUCTION

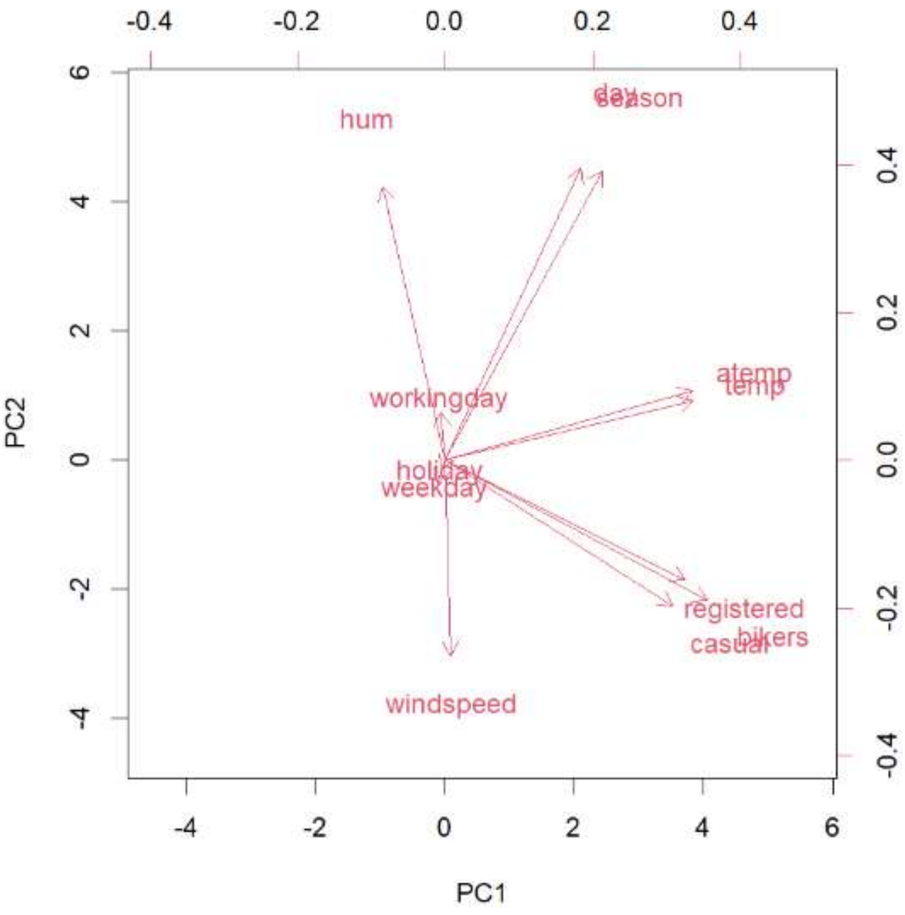
VISUALIZING PRINCIPAL COMPONENTS

Take a minute to understand the graph. How would you interpret it?

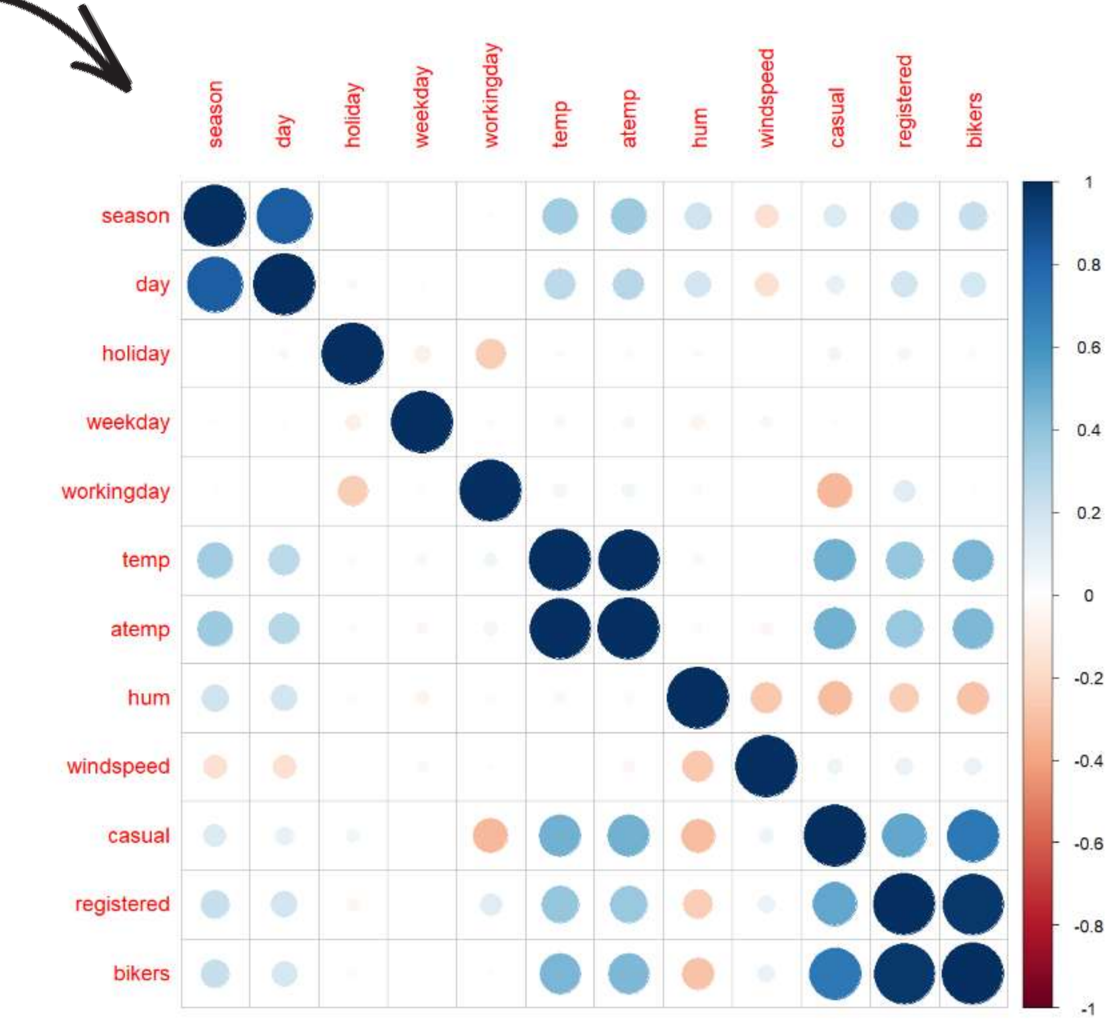


VISUALIZING PRINCIPAL COMPONENTS

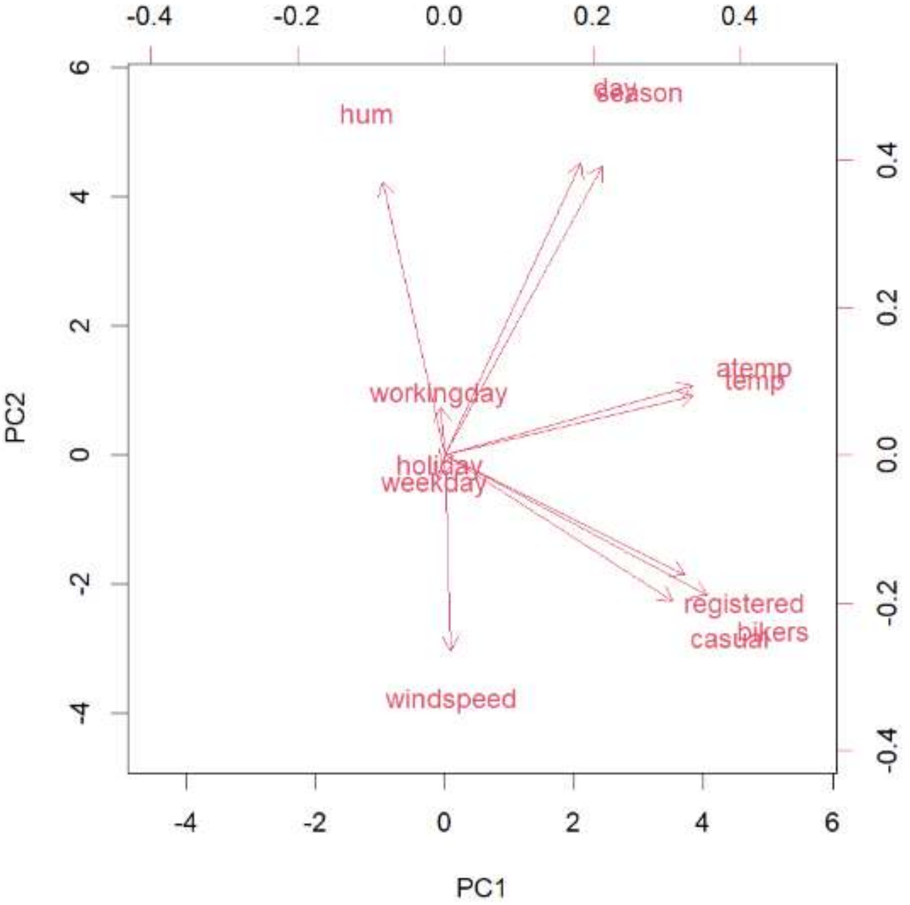
Take a minute to understand the graph. How would you interpret it?



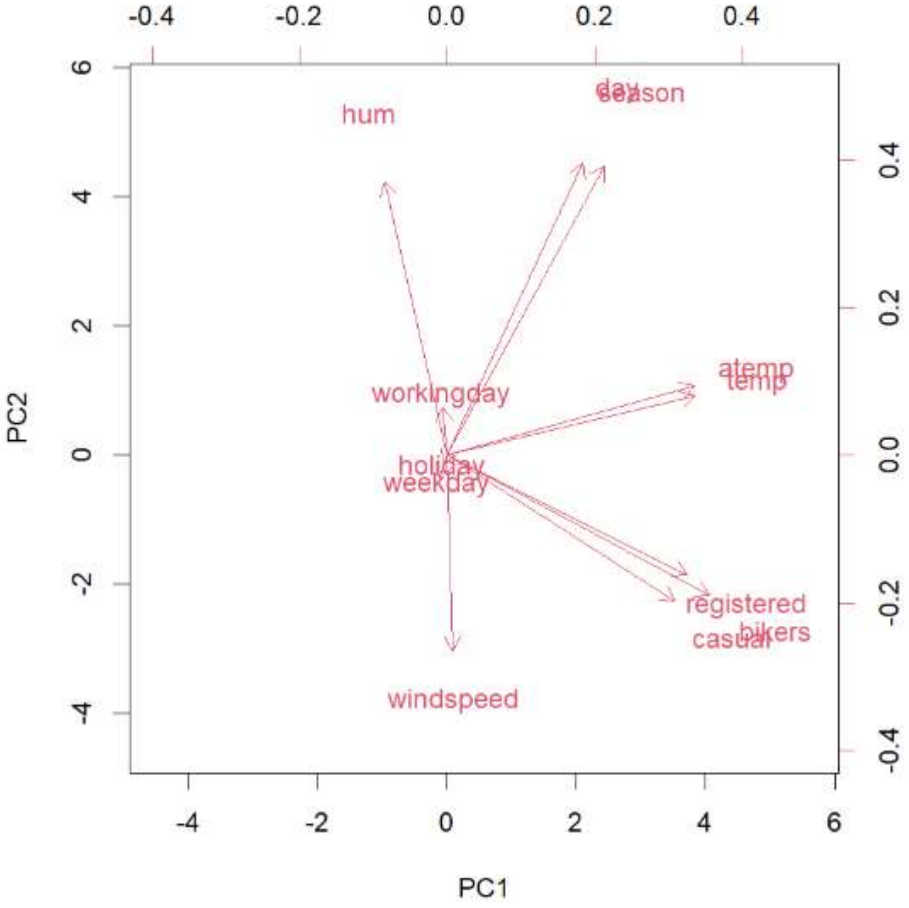
Hint



WHAT VARIABLES ARE REPRESENTED IN EACH PC?

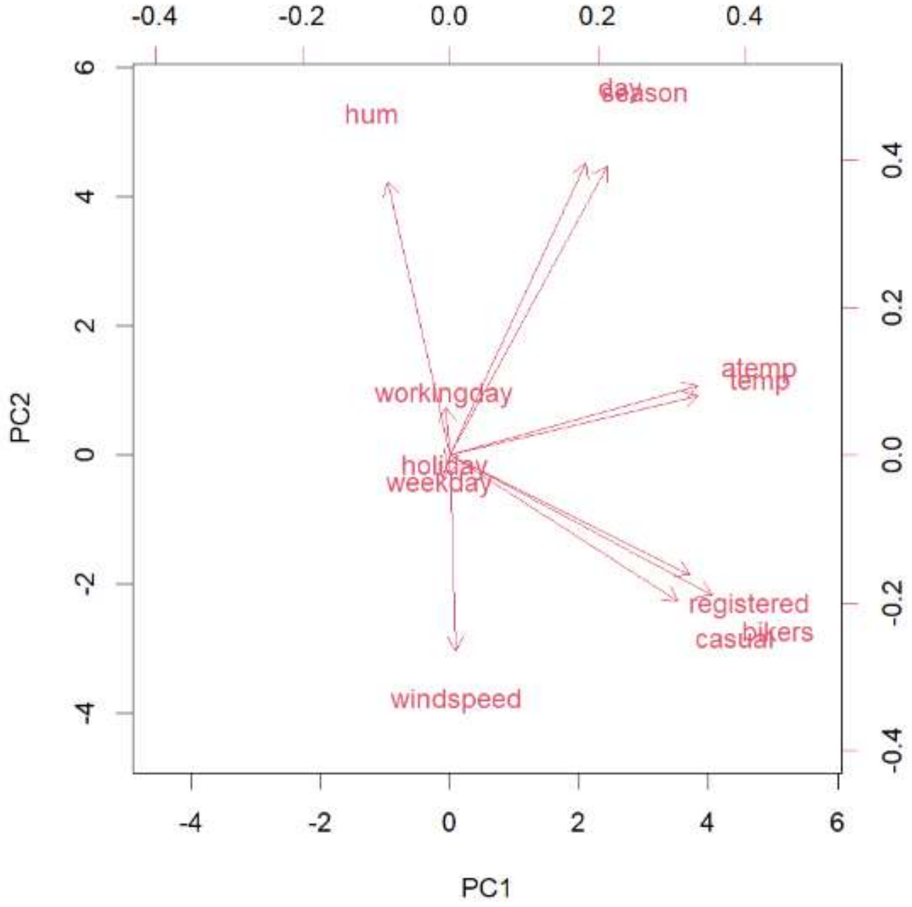


WHAT VARIABLES ARE REPRESENTED IN EACH PC?

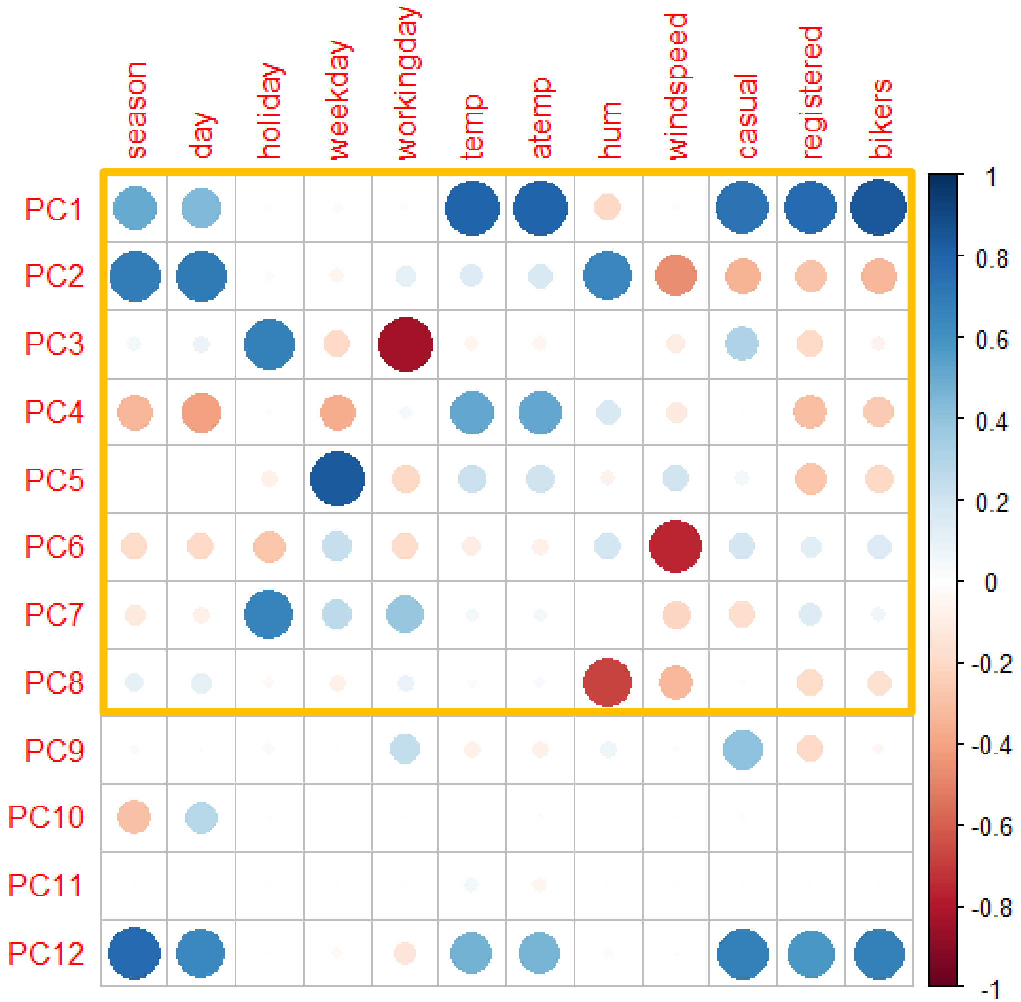


Let’s check the correlations between the principal components and the original data.

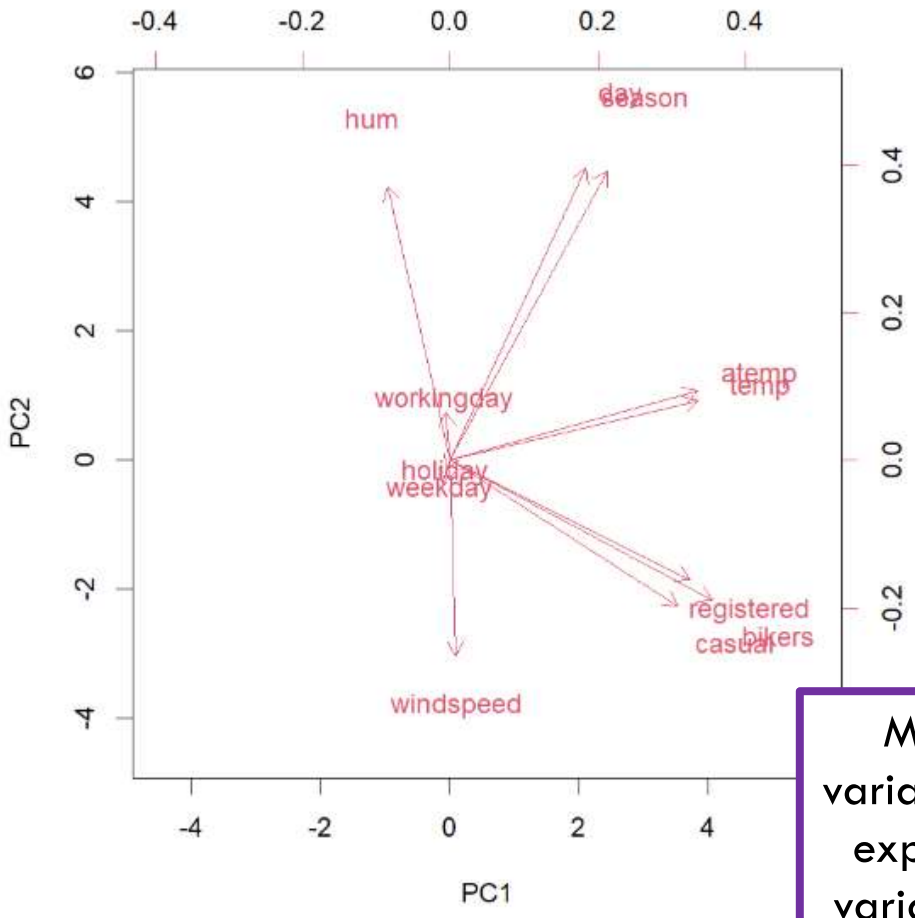
WHAT VARIABLES ARE REPRESENTED IN EACH PC?



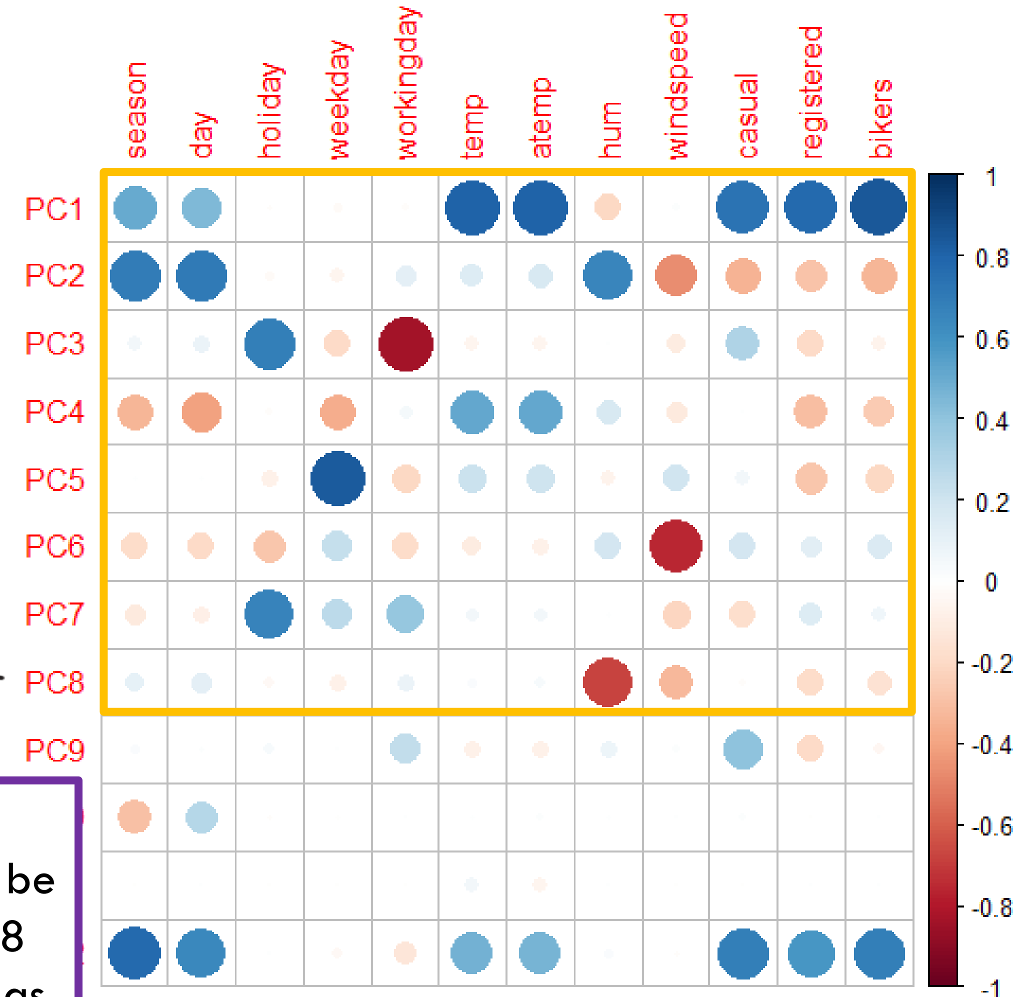
Let’s check the correlations between the principal components and the original data.



WHAT VARIABLES ARE REPRESENTED IN EACH PC?

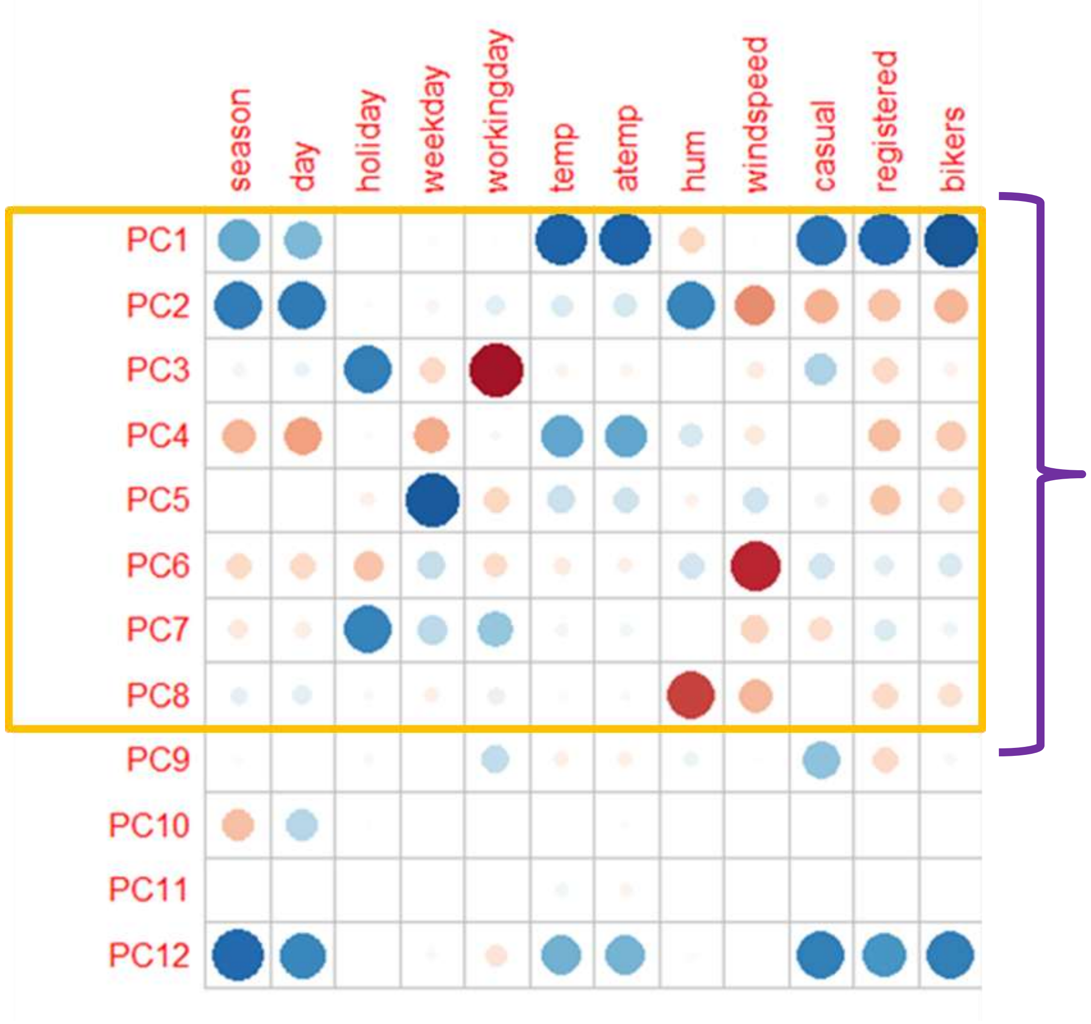


Let’s check the correlations between the principal components and the original data.



Most of the variability can be explained in 8 variables, just as we expected!

WHEN COULD YOU APPLY IT?



You can analyze the most meaningful principal components instead of the original data. Though, it may be more difficult to interpret!

2.4 K-MEANS

WHAT IS K-MEANS?

K-means clustering is a simple and elegant approach for partitioning a data set into K distinct, non-overlapping clusters. To perform K-means clustering, we must first specify the desired number of clusters K; then the K-means algorithm will assign each observation to exactly one of the K clusters.

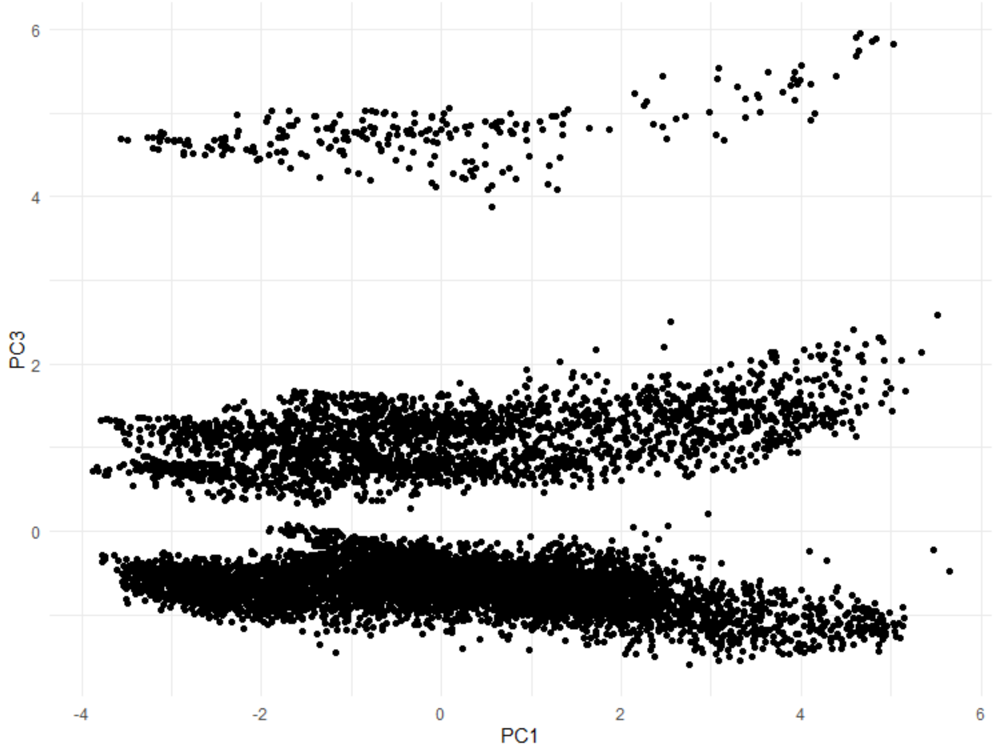
The idea behind K-means clustering is that a good clustering is one for which the within-cluster variation is as small as possible.

LET'S USE THE FIRST 8-PC FROM BEFORE

Visualize different pairs of PC. Can you find something interesting?

LET’S USE THE FIRST 8-PC FROM BEFORE

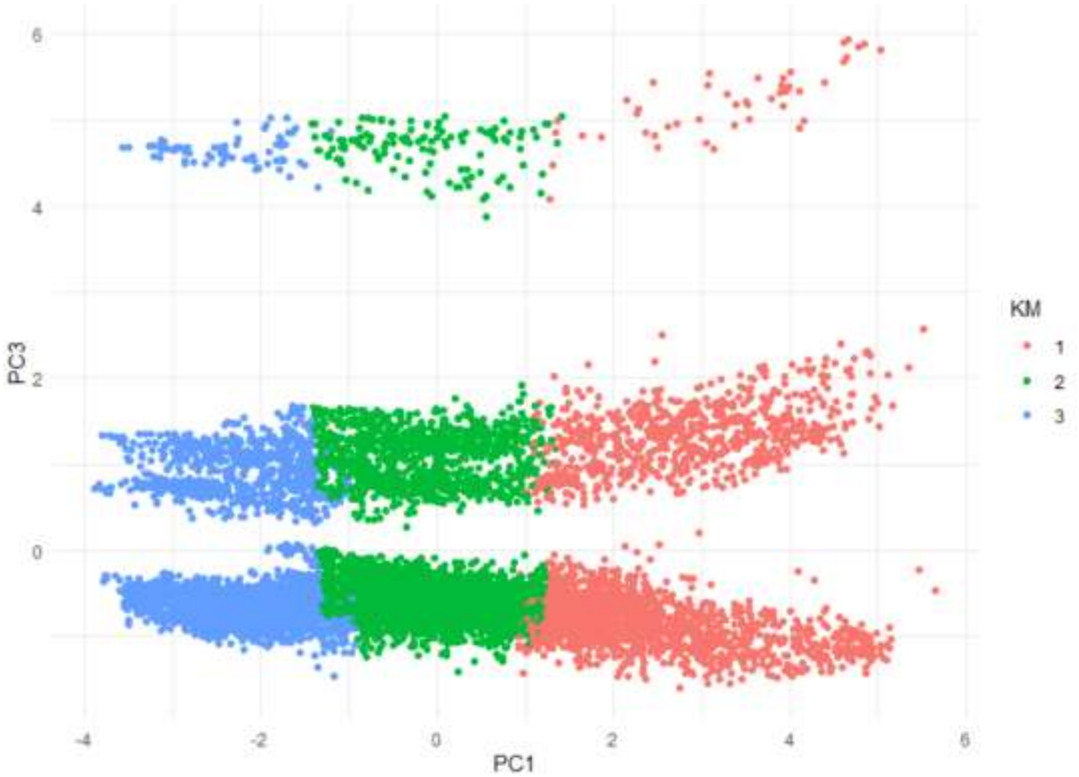
Visualize different pairs of PC. Can you find something interesting?



LET'S APPLY K-MEANS

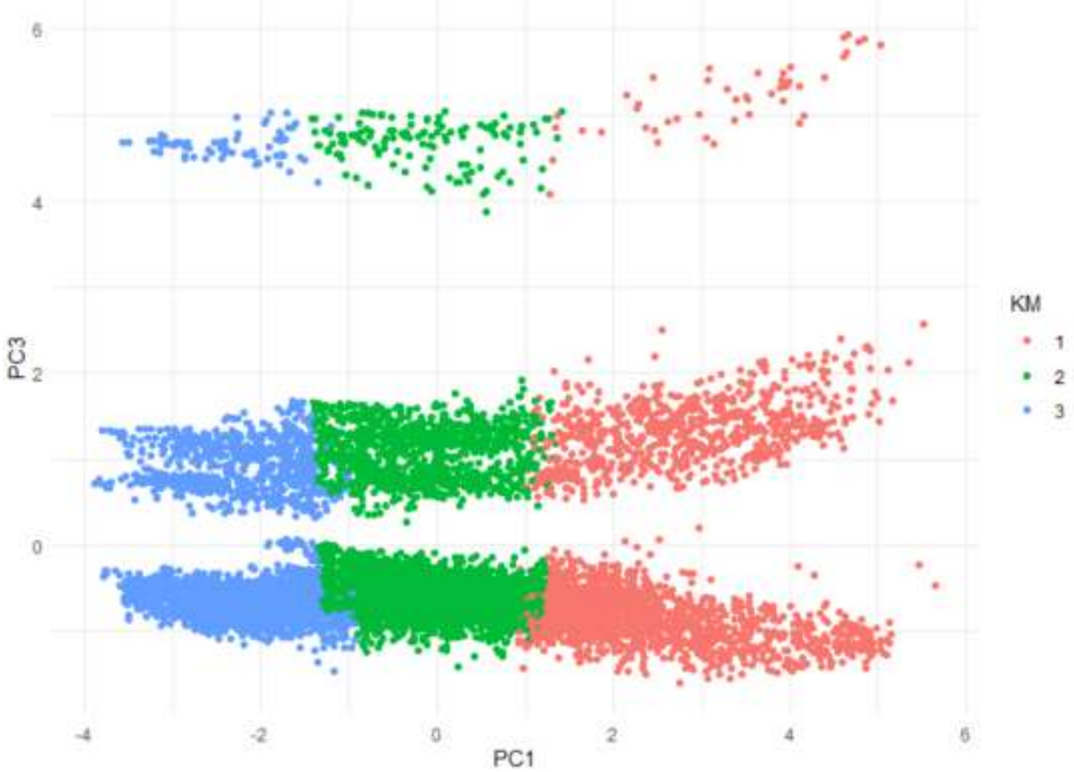
2.5 VISUALIZING CLUSTERS

LET'S VISUALIZE THE RESULTS

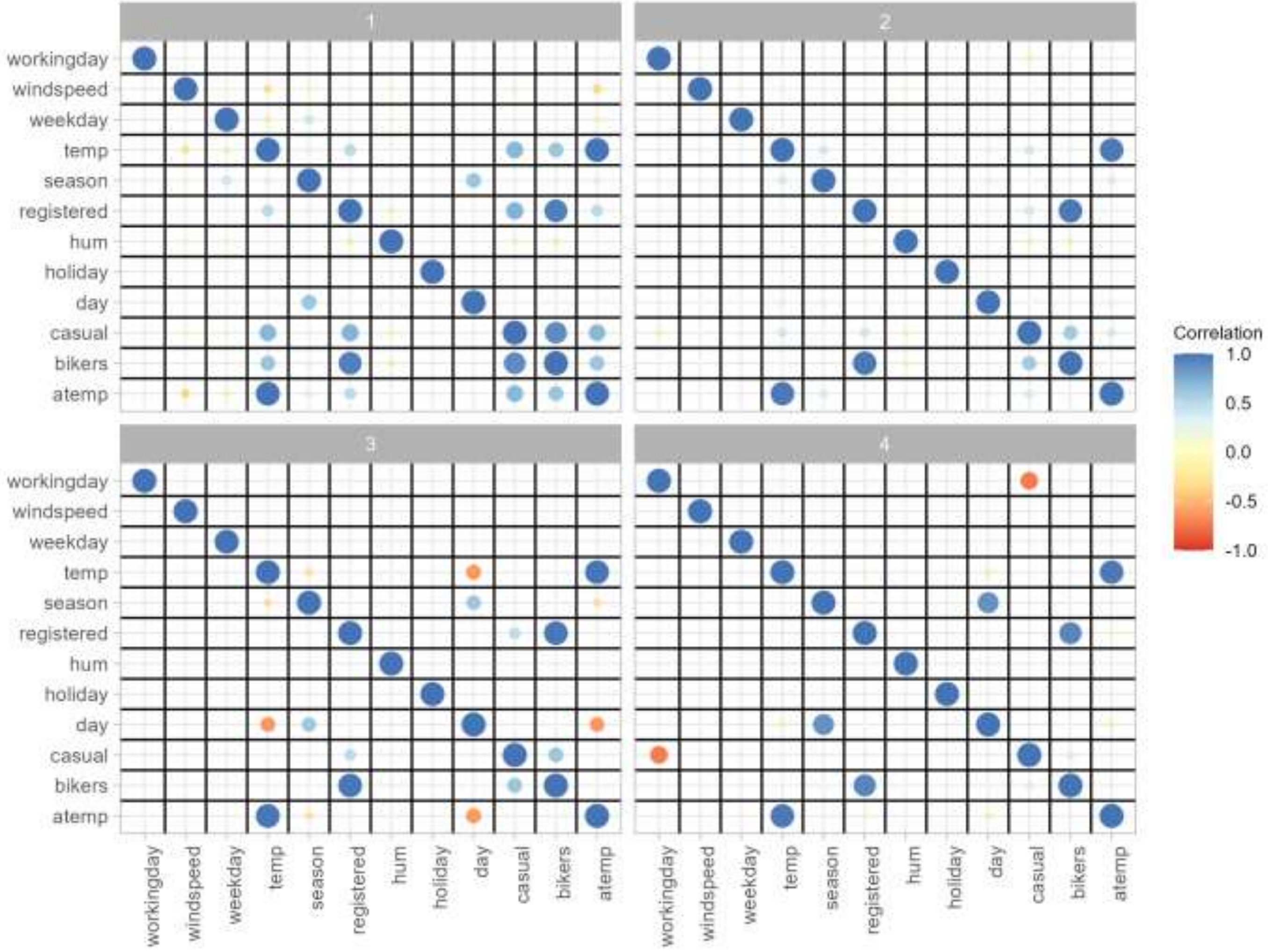


LET'S VISUALIZE THE RESULTS

Is this wrong?



LET'S VISUALIZE THE RESULTS

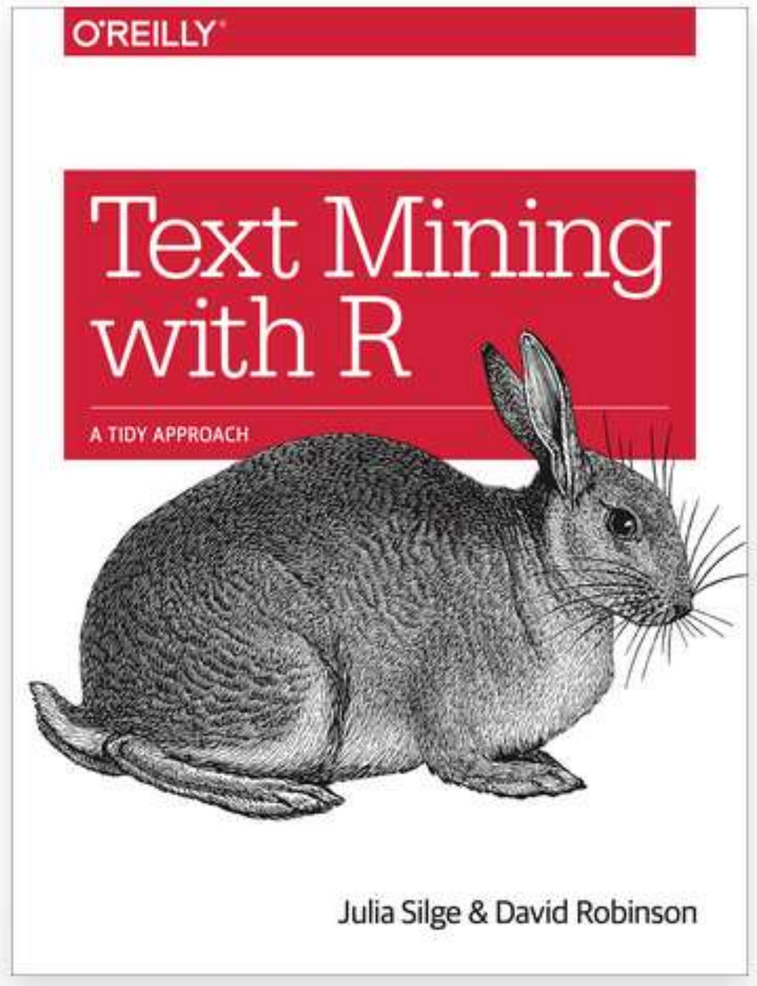


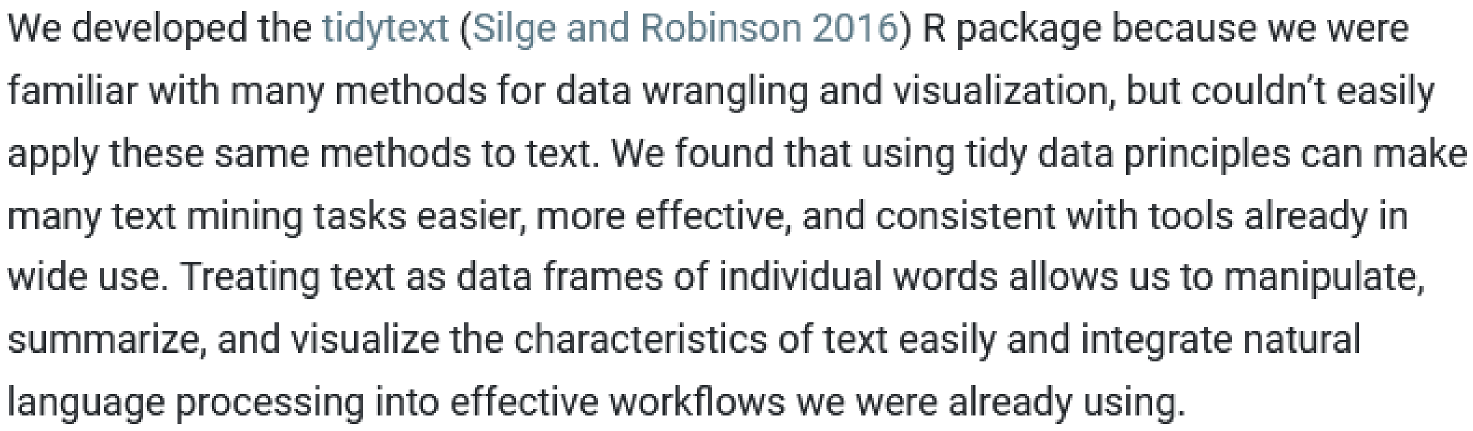
5 MINS BREAK

3. UNSUPERVISED METHODS FOR TEXT

1. Topic Modelling
2. From text to numbers
3. Latent Dirichlet allocation (LDA)

❖ TIDYTEXT: [HTTPS://WWW.TIDYTEXTMINING.COM/](https://www.tidytextmining.com/)





We developed the tidytext (Silge and Robinson 2016) R package because we were familiar with many methods for data wrangling and visualization, but couldn't easily apply these same methods to text. We found that using tidy data principles can make many text mining tasks easier, more effective, and consistent with tools already in wide use. Treating text as data frames of individual words allows us to manipulate, summarize, and visualize the characteristics of text easily and integrate natural language processing into effective workflows we were already using.

3.1 TOPIC MODELING

WHAT IS TOPIC MODELING?

Topic modeling is a collection of text-mining techniques that uses statistical and machine learning models to **automatically discover hidden abstract topics** in a collection of documents.

Topic modeling is also an amalgamation of a set of **unsupervised techniques** that’s capable of detecting word and phrase patterns within documents and **automatically cluster word groups** and similar expressions helping in best representing a set of documents.

Source: https://www.scaler.com/topics/nlp/topic-modelling-in-natural-language-processing/

ARE PCA, TOPIC MODELING, AND CLUSTERING THE SAME?

All three algorithms, PCA, topic modeling, and clustering are unsupervised and used to **simplify the data set with a small number of summaries,** the major difference lies in how they are used:

- *PCA* looks to find a low-dimensional representation of the observations that explain a good fraction of the variance. Since output representations are unrelated, they can also be used as **input to clustering.**
- *Topic* modeling and *PCA* both can be used for dimensionality reduction, but *LDA* provides **better accuracies and explainability in terms of text.**
- Clustering looks to find homogeneous subgroups among the observations by **maximizing the distance** between clusters, clusters are not known in advance.

Source: https://www.scaler.com/topics/nlp/topic-modelling-in-natural-language-processing/

METHODS FOR DIMENSIONALITY REDUCTION AND CLUSTERING

Here we can also perform Dimensionality Reduction! But in this case, there are other well-known methods:

- Non-Negative Matrix Factorization
- LDA – Latent Dirichlet Allocation
- LSA – Latent Semantic Allocation
- PLSA – Probabilistic Latent Semantic Analysis
- lda2vec – Deep Learning Model
- tBERT – Topic BERT

Source: https://www.scaler.com/topics/nlp/topic-modelling-in-natural-language-processing/

METHODS FOR DIMENSIONALITY REDUCTION AND CLUSTERING

Here we can also perform Dimensionality Reduction! But in this case, there are other well-known methods:

- Non-Negative Matrix Factorization
- LDA – Latent Dirichlet Allocation

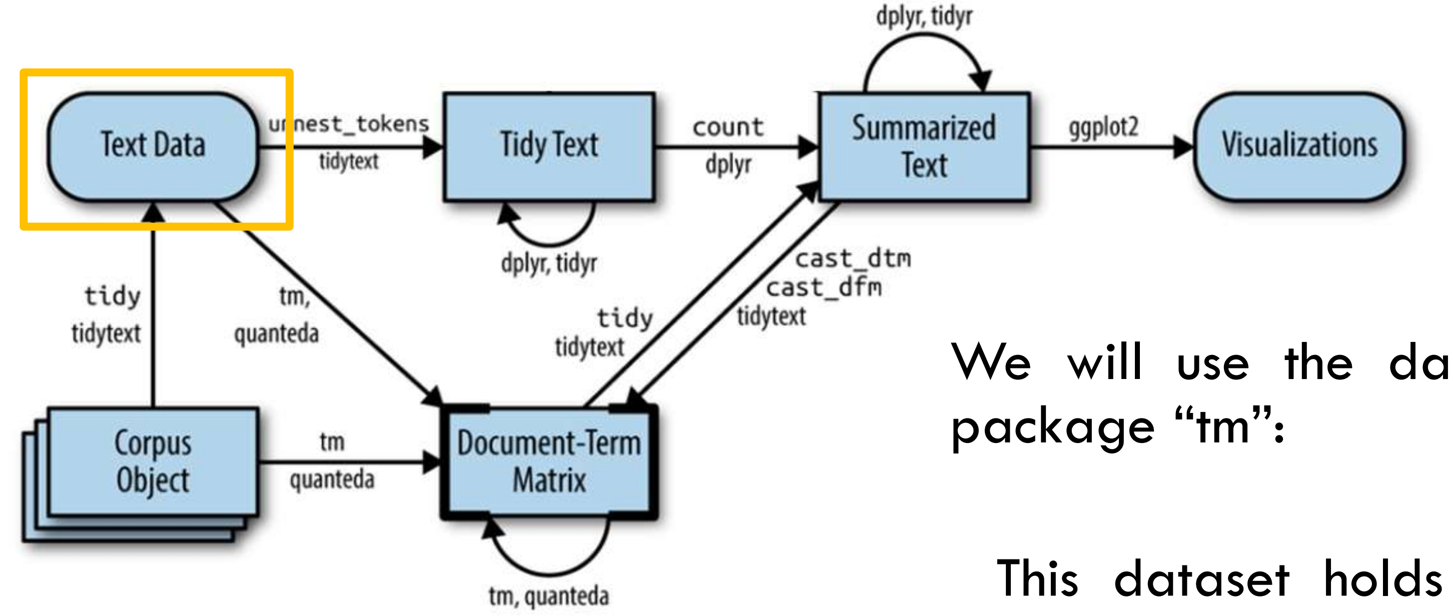
We will focus on this one!

- LSA – Latent Semantic Allocation
- PLSA – Probabilistic Latent Semantic Analysis
- lda2vec – Deep Learning Model
- tBERT – Topic BERT

Source: https://www.scaler.com/topics/nlp/topic-modelling-in-natural-language-processing/

3.2 FROM TEXT TO NUMBERS

FROM TEXT TO NUMBERS



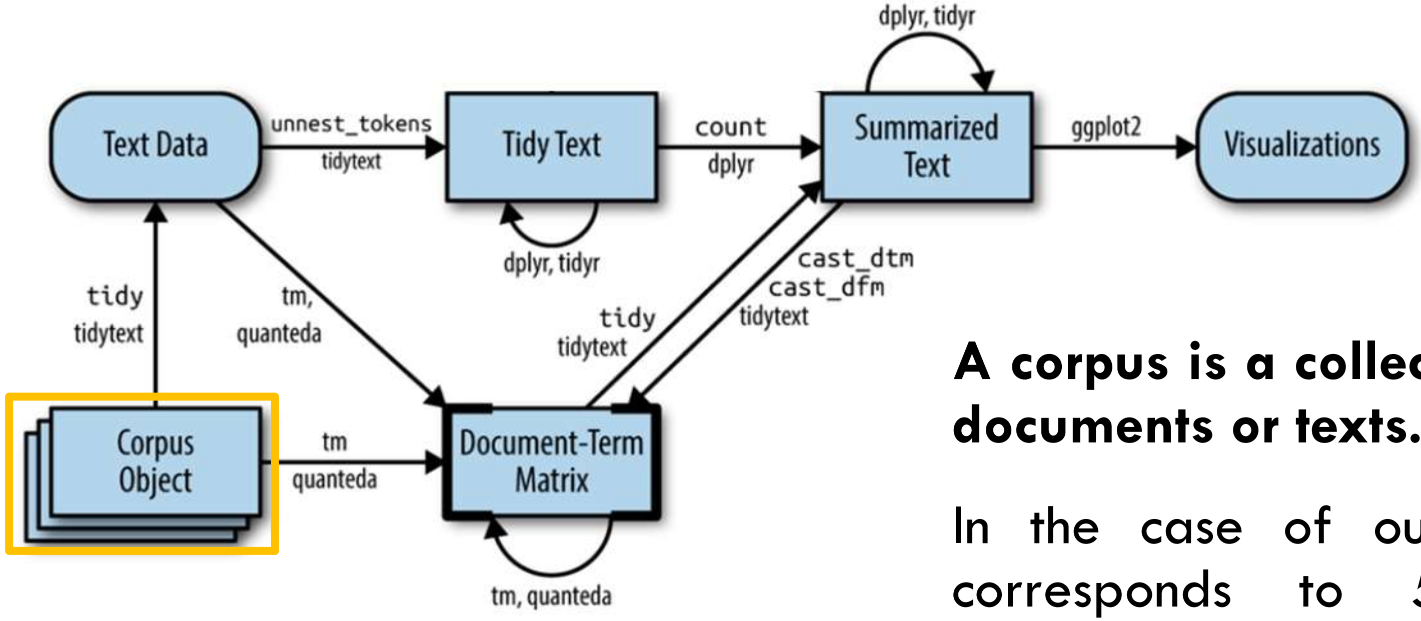
We will use the data "acq" from the package "tm":

This dataset holds **50 news articles** with additional meta information from the Reuters-21578 data set. **All documents belong to the topic acq dealing with corporate acquisitions.**

```r
library(tm)
data("acq")
```

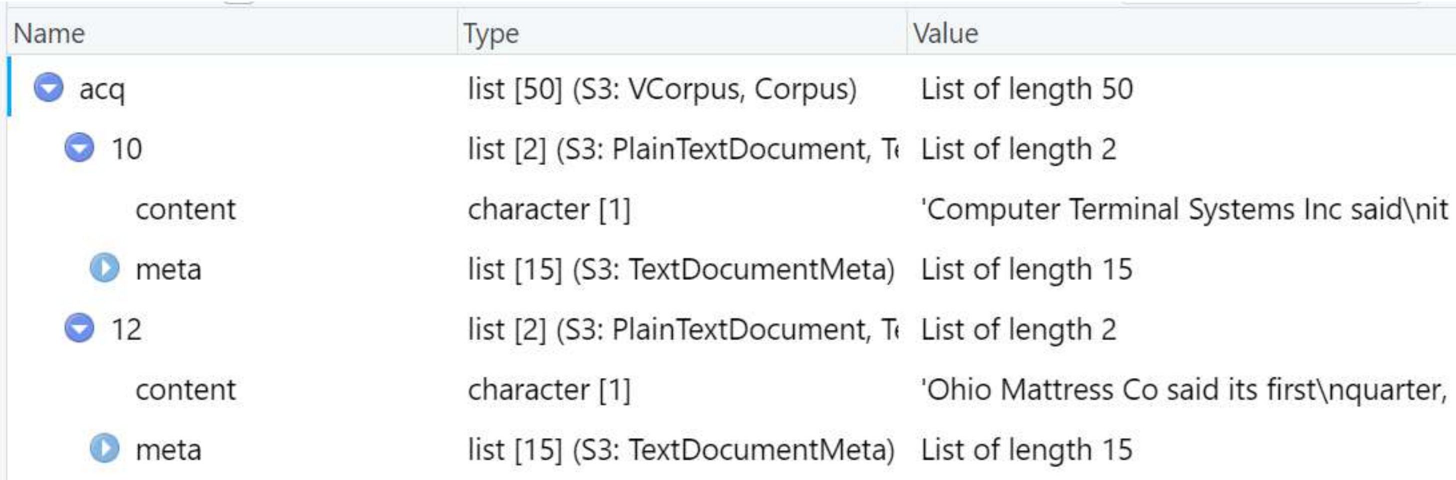
Source: https://www.tidytextmining.com/topicmodeling#topicmodeling
Also: https://rpubs.com/Argaadya/topic_lda

FROM TEXT TO NUMBERS



A corpus is a collection of unstructured documents or texts.

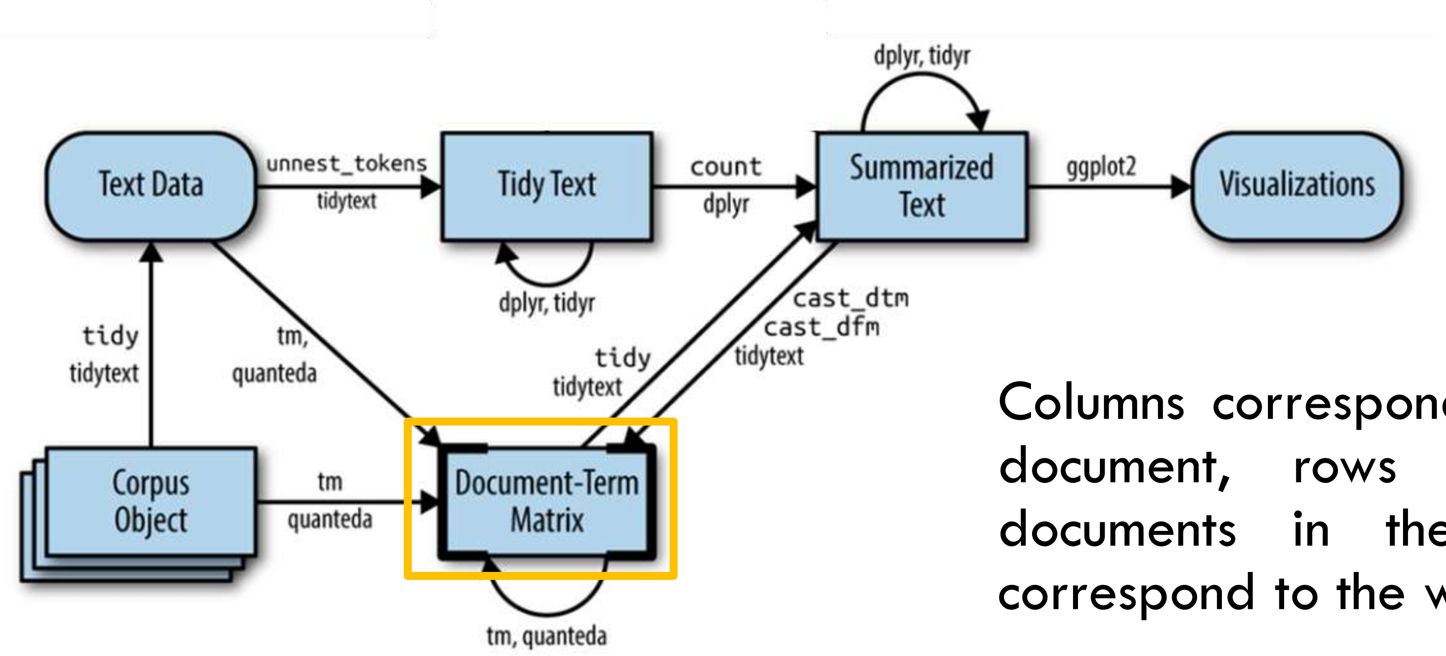
In the case of our data, the corpus corresponds to 50 articles about corporate acquisition.

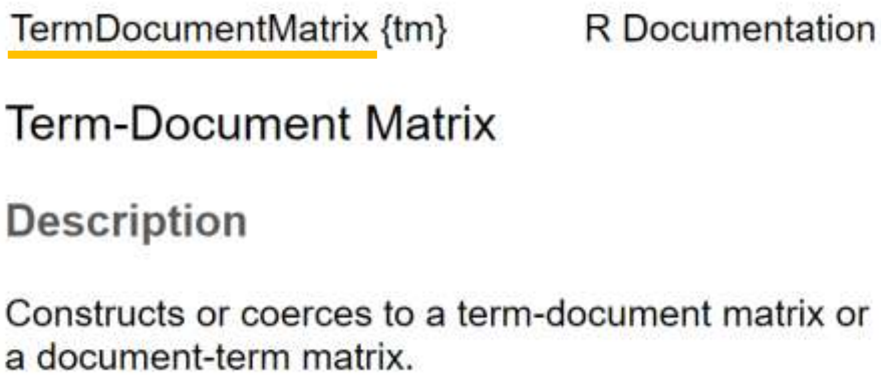


Name	Type	Value
acq	list [50] (S3: VCorpus, Corpus)	List of length 50
10	list [2] (S3: PlainTextDocument, T	List of length 2
content	character [1]	'Computer Terminal Systems Inc said\nit
meta	list [15] (S3: TextDocumentMeta)	List of length 15
12	list [2] (S3: PlainTextDocument, T	List of length 2
content	character [1]	'Ohio Mattress Co said its first\nquarter,
meta	list [15] (S3: TextDocumentMeta)	List of length 15

Source: https://www.tidytextmining.com/topicmodeling#topicmodeling
Also: https://rpubs.com/Argaadya/topic_lda

FROM TEXT TO NUMBERS





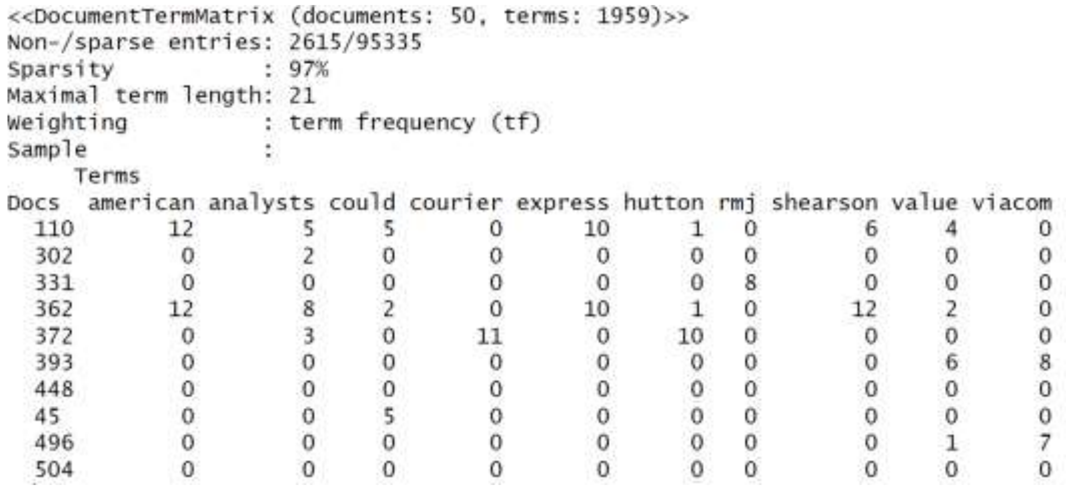
TermDocumentMatrix {tm}   R Documentation

Term-Document Matrix

Description

Constructs or coerces to a term-document matrix or a document-term matrix.

Columns correspond to the terms in the document, rows correspond to the documents in the corpus and cells correspond to the weights of the terms.



```
<<DocumentTermMatrix (documents: 50, terms: 1959)>>
Non-/sparse entries: 2615/95335
Sparsity           : 97%
Maximal term length: 21
Weighting          : term frequency (tf)
Sample             :
     Terms
Docs  american analysts could courier express hutton rmj shearson value viacom
  110       12        5     5       0      10      1   0        6     4      0
  302        0        2     0       0       0      0   0        0     0      0
  331        0        0     0       0       0      0   8        0     0      0
  362       12        8     2       0      10      1   0       12     2      0
  372        0        3     0      11       0     10   0        0     0      0
  393        0        0     0       0       0      0   0        0     6      8
  448        0        0     0       0       0      0   0        0     0      0
   45        0        0     5       0       0      0   0        0     0      0
  496        0        0     0       0       0      0   0        0     1      7
  504        0        0     0       0       0      0   0        0     0      0
```

Source: https://www.tidytextmining.com/topicmodeling#topicmodeling
Also: https://rpubs.com/Argaadya/topic_lda

3.3 LATENT DIRICHLET ALLOCATION

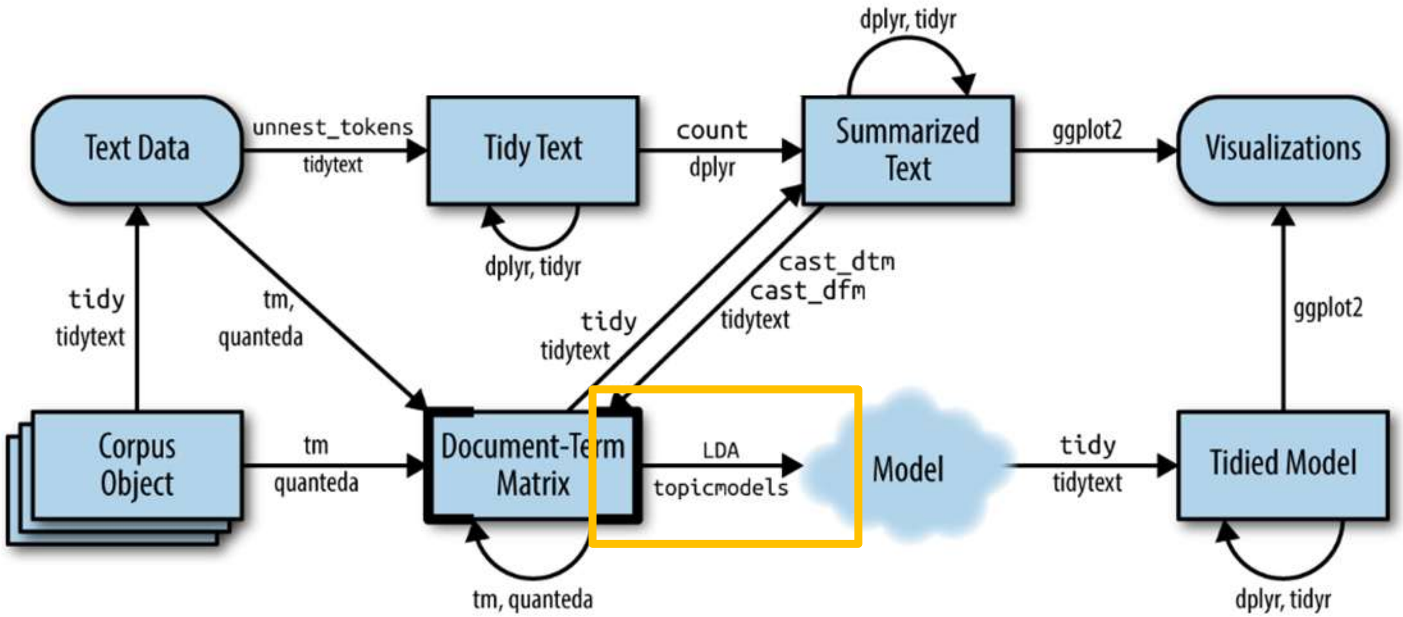
LATENT DIRICHLET ALLOCATION

Latent Dirichlet allocation is one of the most common algorithms for topic modeling. Without diving into the math behind the model, we can understand it as being guided by two principles.

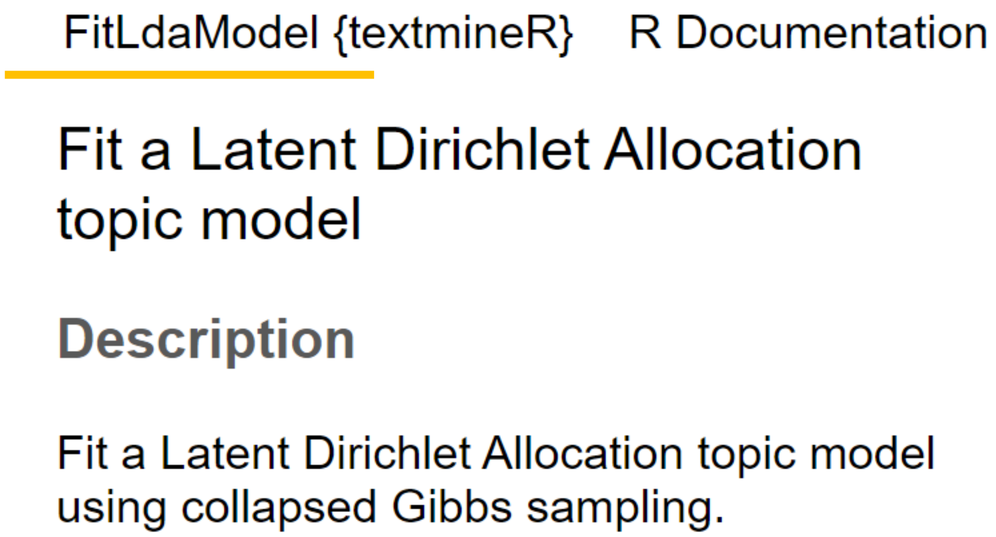
- **Every document is a mixture of topics.** We imagine that each document may contain words from several topics in particular proportions. For example, in a two-topic model we could say "Document 1 is 90% topic A and 10% topic B, while Document 2 is 30% topic A and 70% topic B."
- **Every topic is a mixture of words.** For example, we could imagine a two-topic model of American news, with one topic for "politics" and one for "entertainment." The most common words in the politics topic might be "President", "Congress", and "government", while the entertainment topic may be made up of words such as "movies", "television", and "actor". Importantly, words can be shared between topics; a word like "budget" might appear in both equally.

Source: https://www.tidytextmining.com/topicmodeling#topicmodeling

LATENT DIRICHLET ALLOCATION



LDA is a mathematical method for estimating both at the same time: finding the mixture of words that is associated with each topic, while also determining the mixture of topics that describes each document.



FitLdaModel {textmineR}   R Documentation

Fit a Latent Dirichlet Allocation topic model

Description

Fit a Latent Dirichlet Allocation topic model using collapsed Gibbs sampling.

Source: https://www.tidytextmining.com/topicmodeling#topicmodeling
Also: https://rpubs.com/Argaadya/topic_lda

MAKING SENSE OF THE RESULTS

Some important attributes acquired from the LDA Model:

- **phi** : Posterior per-topic-per-word probabilities
- **theta** : Posterior per-document-per-topic probabilities
- **alpha** : Prior per-document-per-topic probabilities
- **beta** : Prior per-document-per-topic probabilities
- **coherence** : The probabilistic coherence of each topic

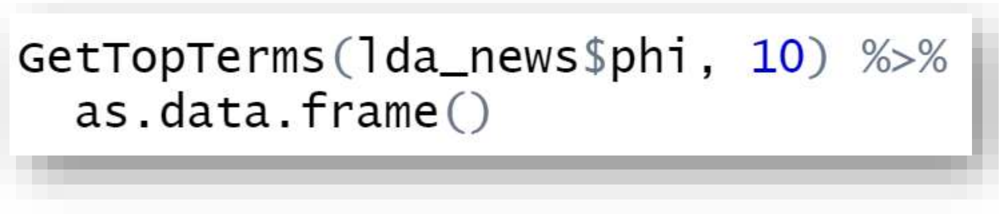
Source: https://www.tidytextmining.com/topicmodeling#topicmodeling
Also: https://rpubs.com/Argaadya/topic_lda

MAKING SENSE OF THE RESULTS

Some important attributes acquired from the LDA Model:

- **phi** : Posterior per-topic-per-word probabilities
- **theta** : Posterior per-document-per-topic probabilities
- **alpha** : Prior per-document-per-topic probabilities
- **beta** : Prior per-document-per-topic probabilities
- **coherence** : The probabilistic coherence of each topic

Let’s check the top ten words per cluster!

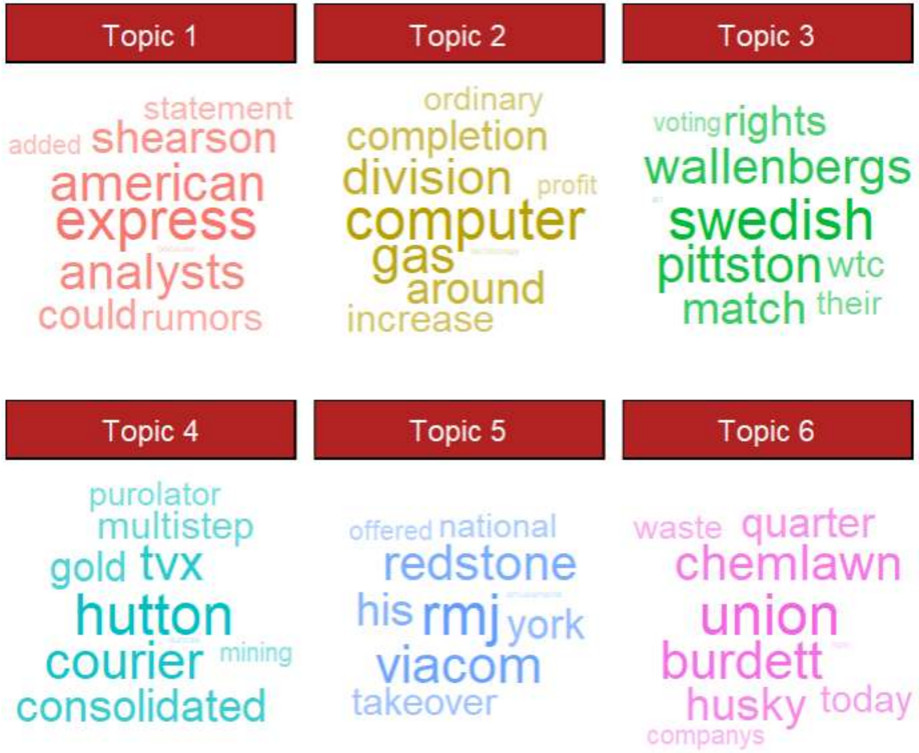


```
GetTopTerms(lda_news$phi, 10) %>%
  as.data.frame()
```

Source: https://www.tidytextmining.com/topicmodeling#topicmodeling
Also: https://rpubs.com/Argaadya/topic_lda

MAKING SENSE OF THE RESULTS

LDA doesn’t specifically inform us about what each topic is about. By looking at the representative words of each topic, we as the human will give meaning to each topic.



Source: https://www.tidytextmining.com/topicmodeling#topicmodeling
Also: https://rpubs.com/Argaadya/topic_lda

Join us!

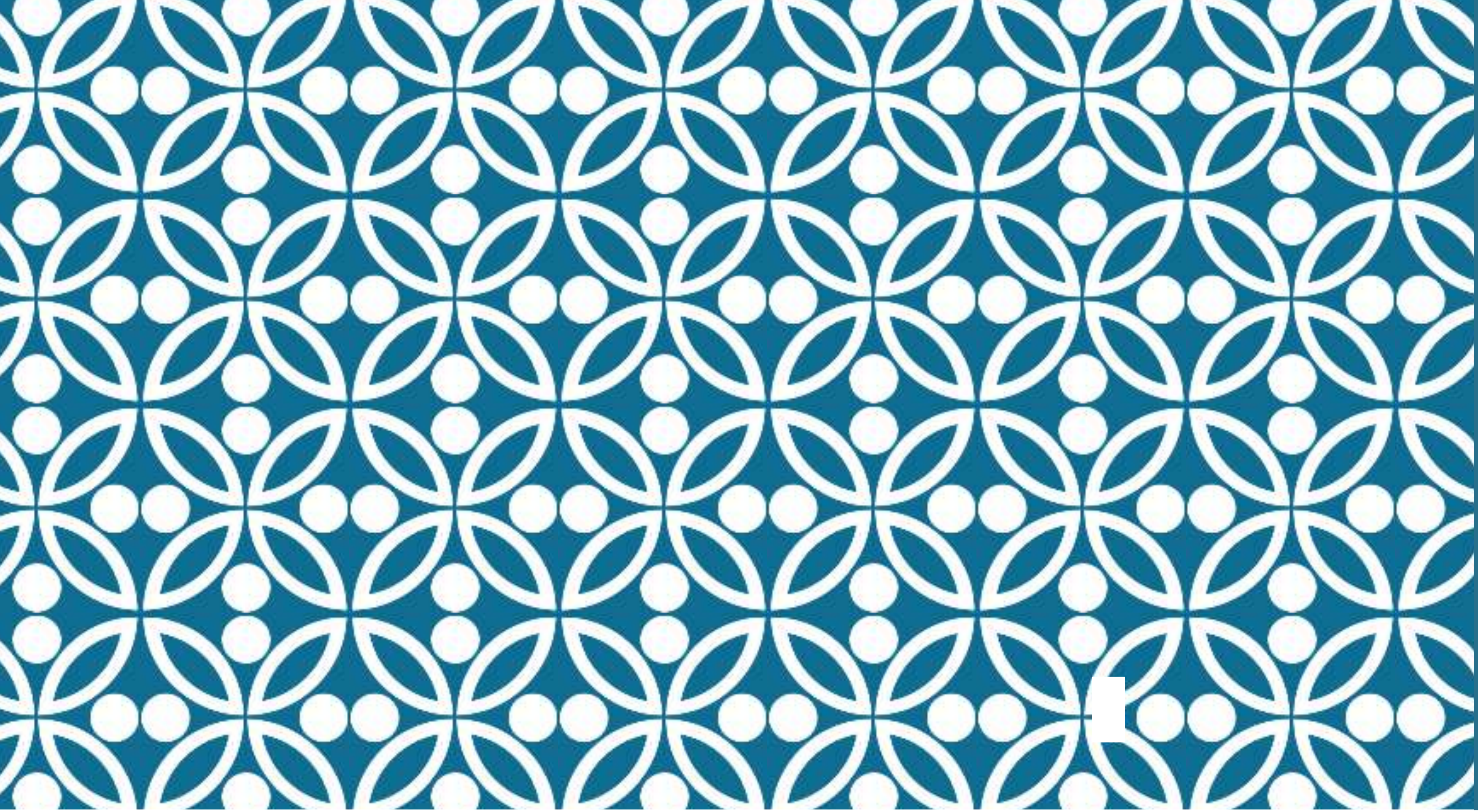
https://forms.office.com/e/8Bgd2YsasJ

All about the lab:

https://societal-analytics.nl/

Contact us at:

analytics-lab.fsw@vu.nl.



https://sofiag1l.github.io/

THANKS!

Dr. Sofia Gil-Clavel