

SESSION 4: UNSUPERVISED LEARNING

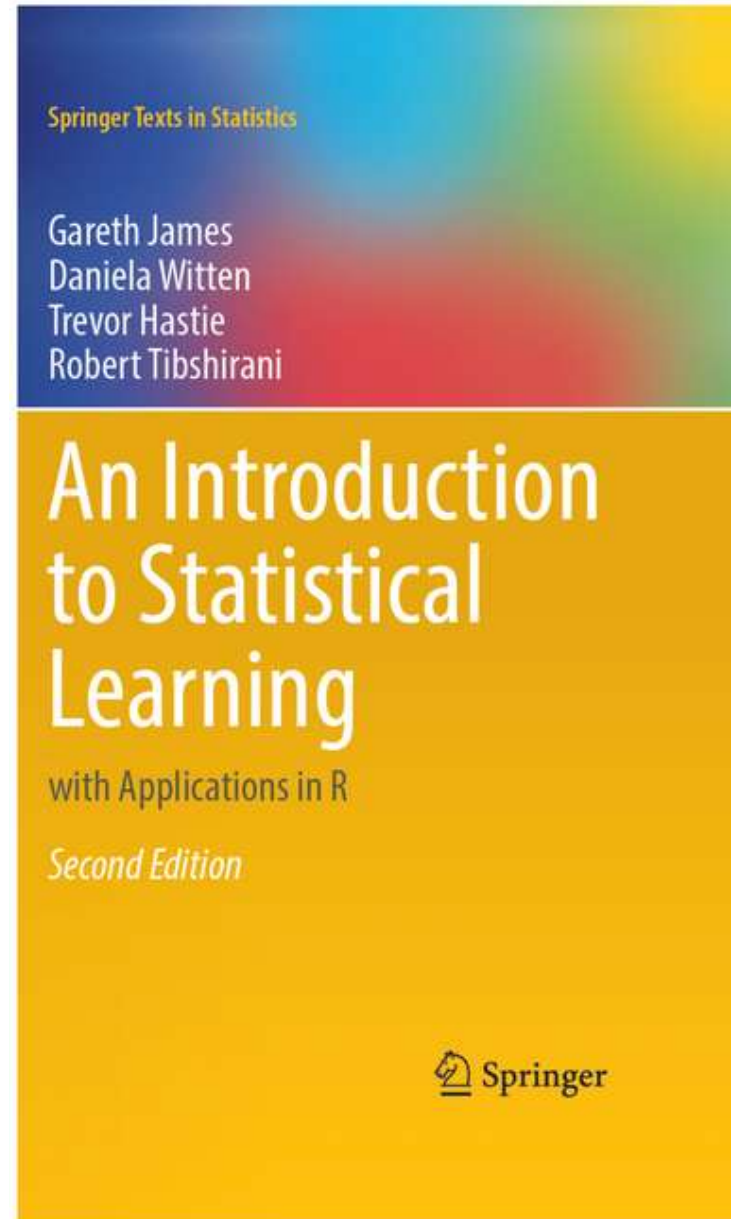
DR. SOFIA GIL-CLAVEL

- ❖ The basics of Statistical Learning
- ❖ Unsupervised Methods

WE WILL BE FOLLOWING:

You can find the book for free here:

<https://www.statlearning.com/>



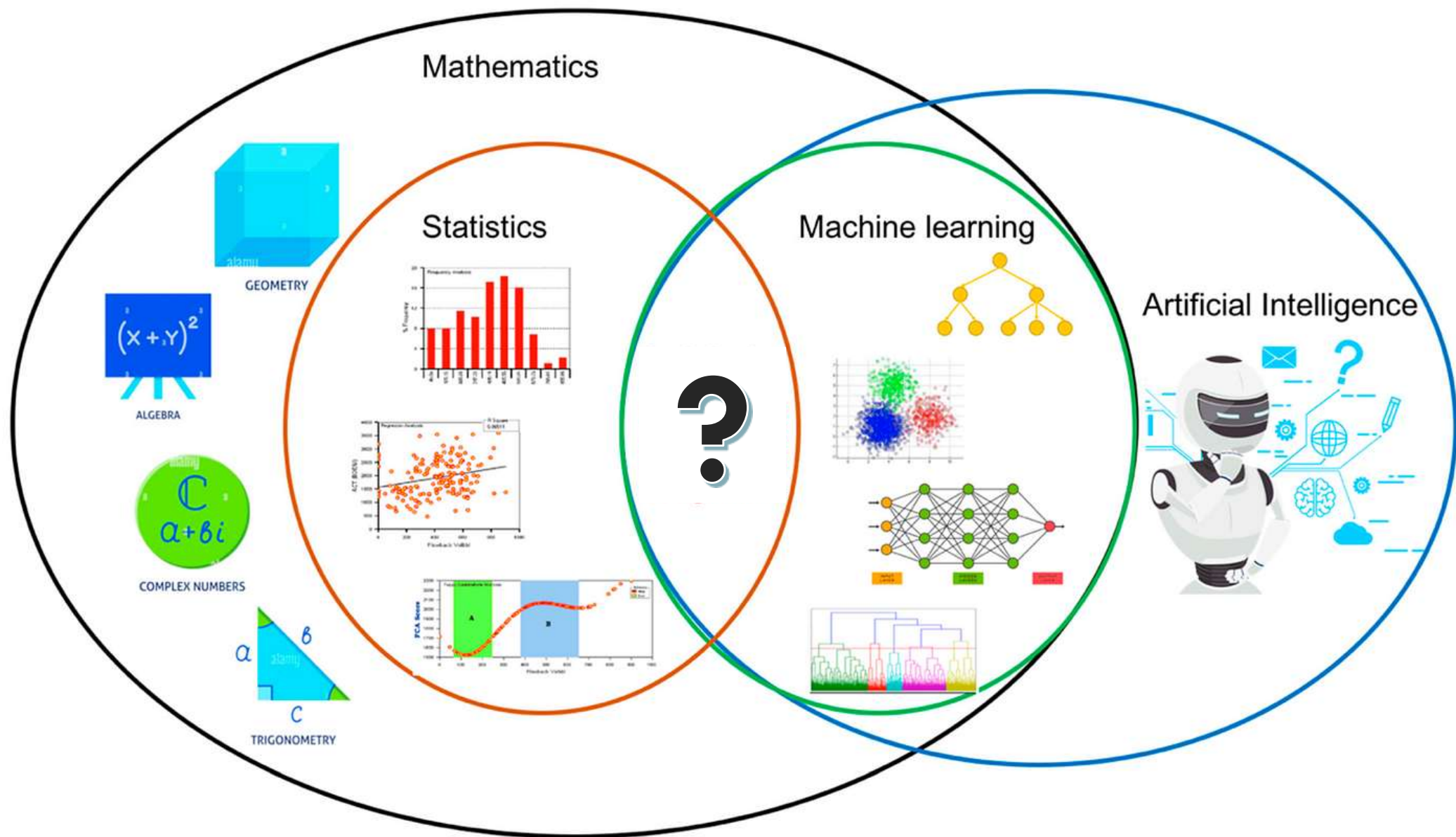
1. THE BASICS OF STATISTICAL LEARNING

1.1 Statistical learning

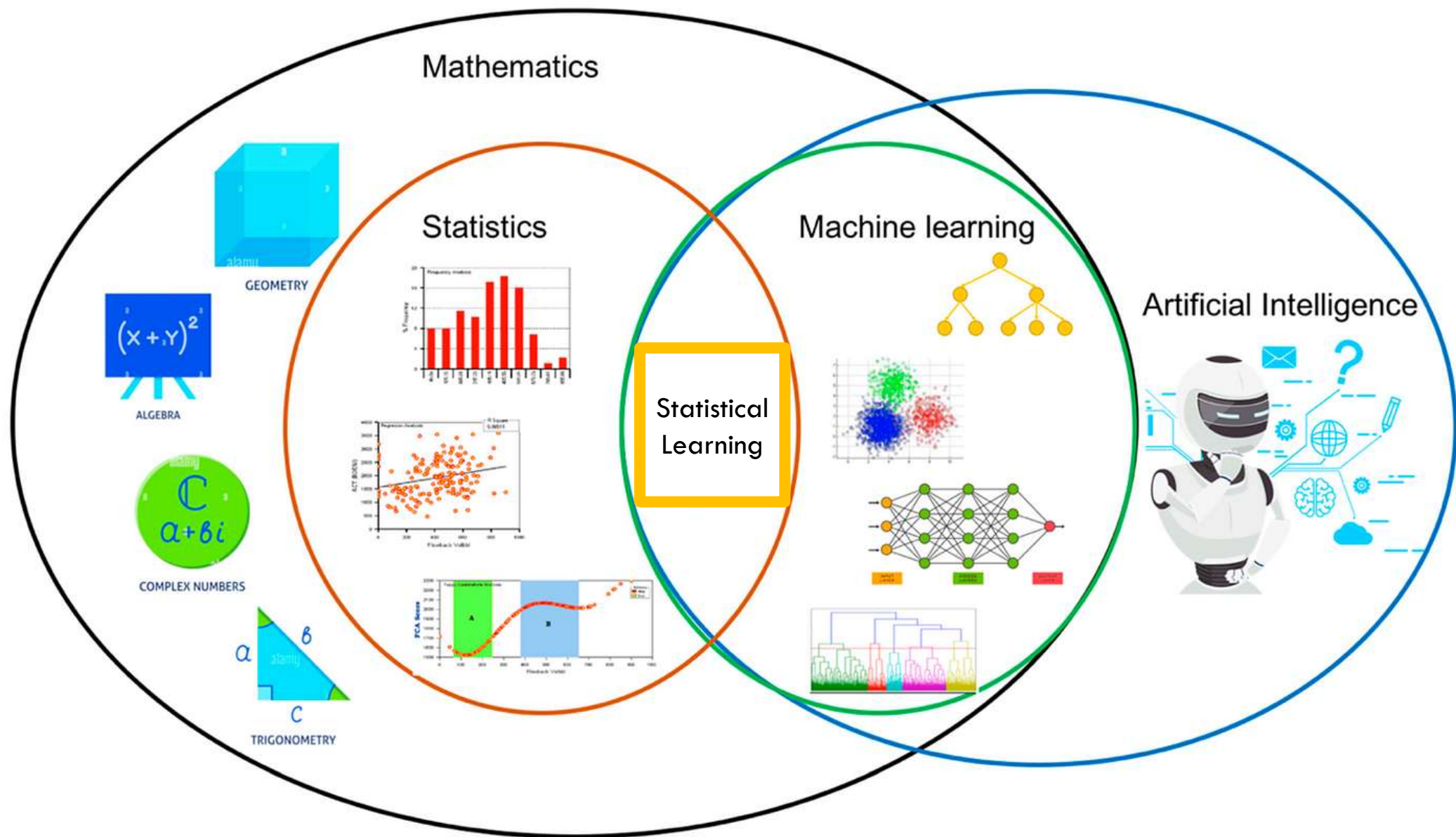
1.2 Supervised vs. Unsupervised

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.

data frame
name

columns

rows

variable

value/element

```
> Bikeshare
```

	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

10,055 more rows
3 more variables: windspeed <dbl>, casual <dbl>,
registered <dbl>
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

```
# 18,455 more rows
# 3 more variables: windspeed <dbl>, casual <dbl>
#   registered <dbl>
# 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.

$$Y \sim X1 + X2 + X3$$

 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 $i = 1, \dots, n$, we observe a vector of measurements X_i but no associated response Y_i .

$$\text{No } Y_i \sim X1_i + X2_i + X3_i$$

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 $X1, X2, \dots, Xp$. 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.

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

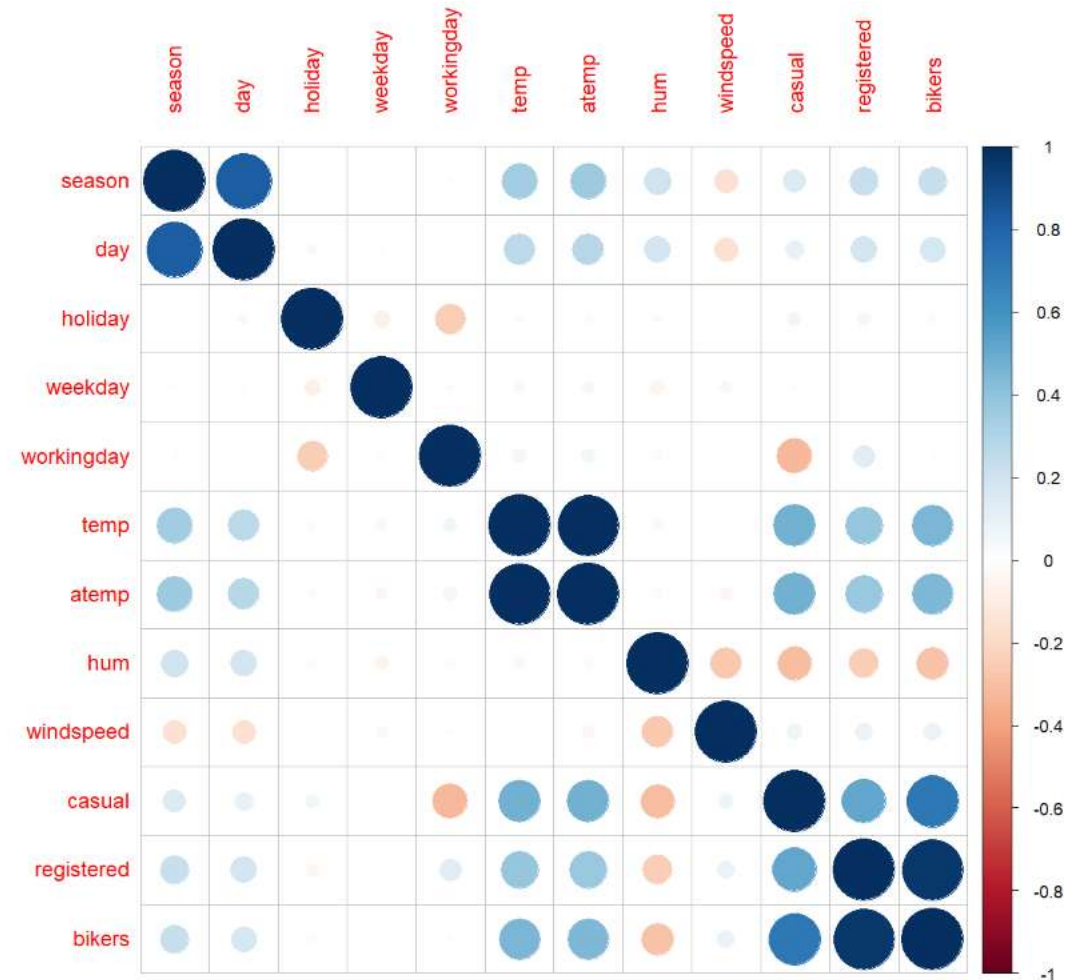
Take a minute to understand the variables. Which ones do you think are correlated?

- **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.

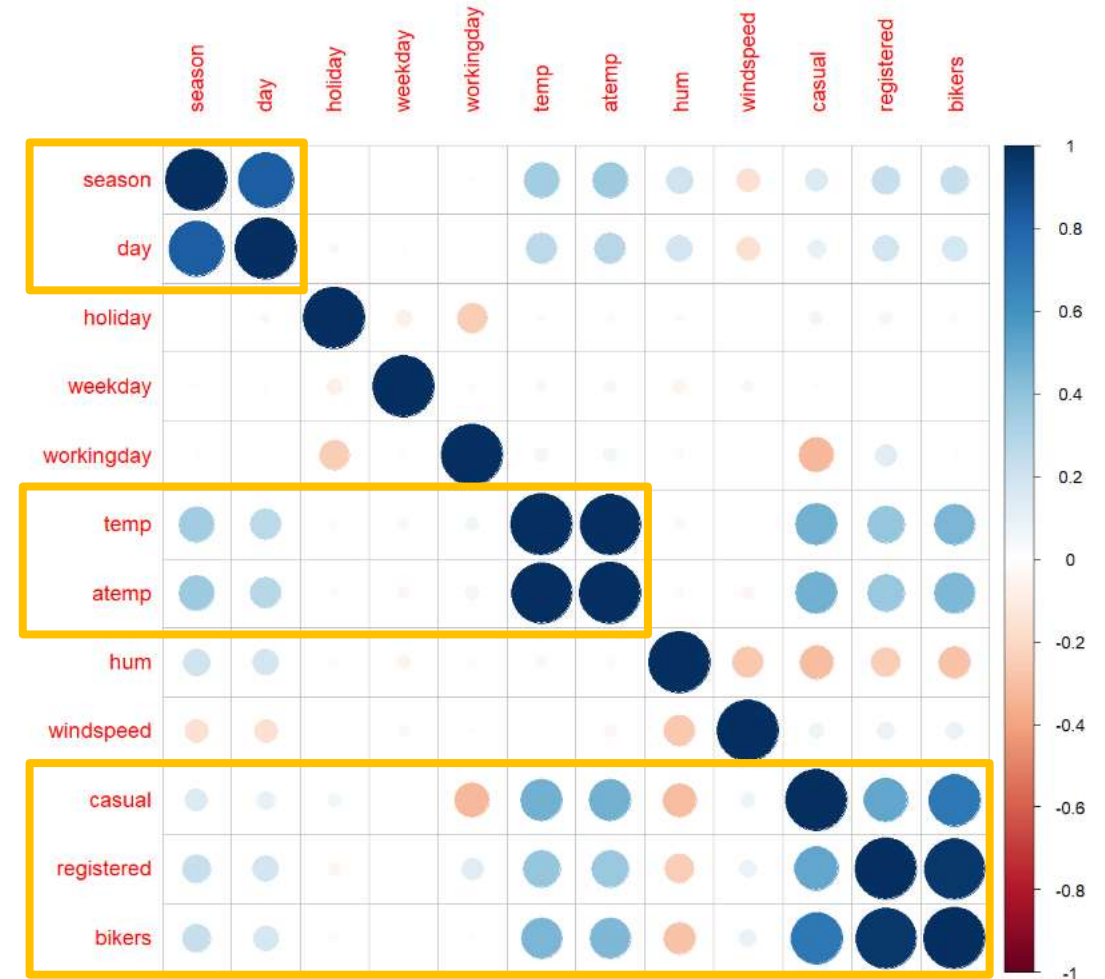
5 MINS BREAK



THE “BIKESHARE” DATA



THE “BIKESHARE” DATA



THE “BIKESHARE” DATA

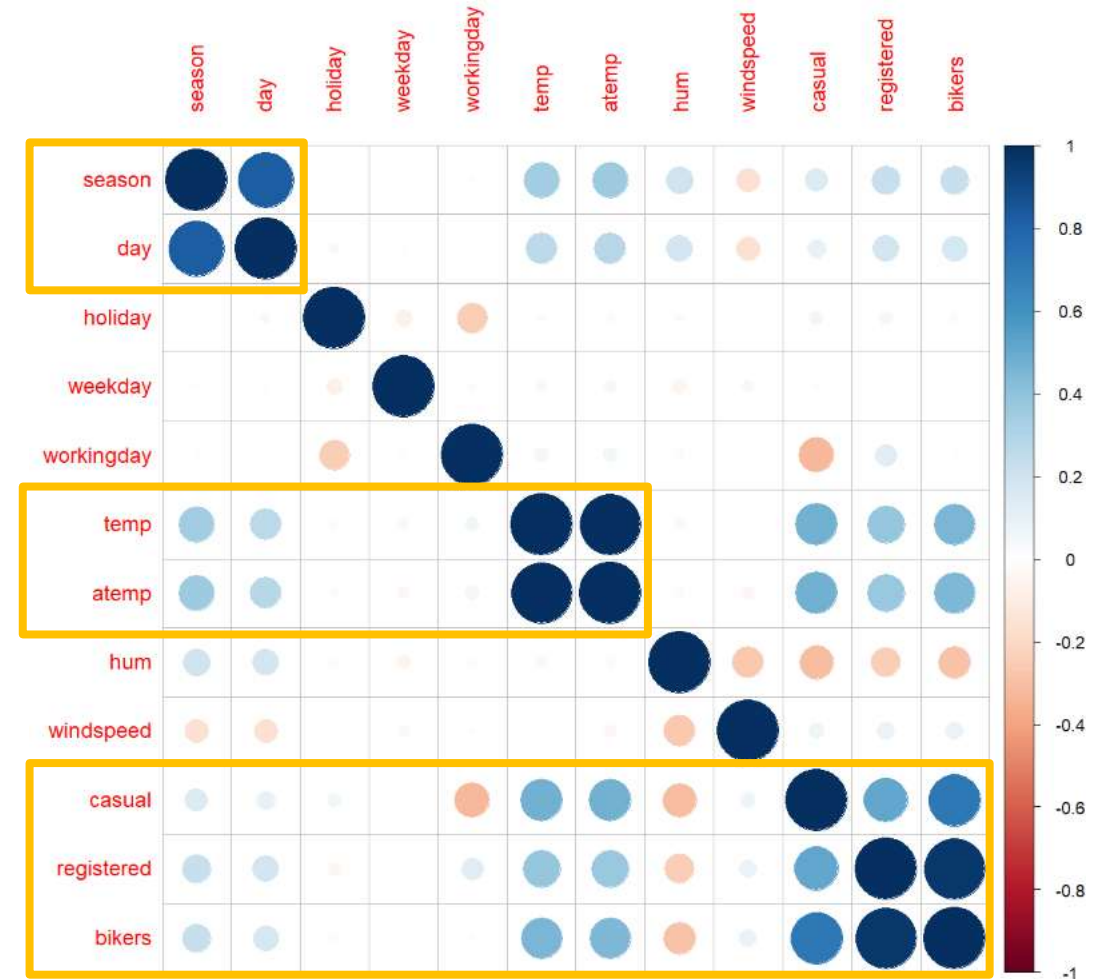
Based on the definition



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

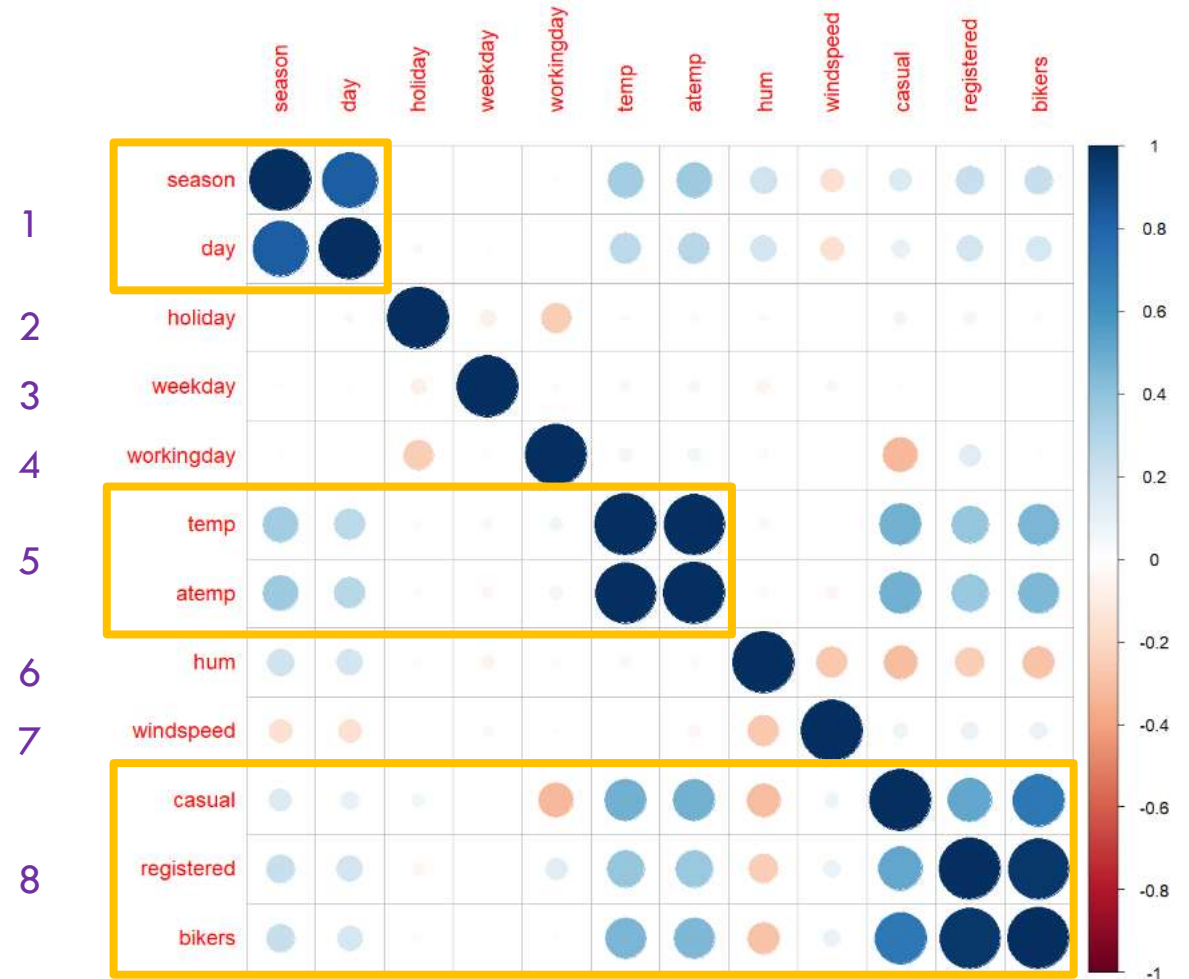
Based on the definition



When faced with a large set of **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**?

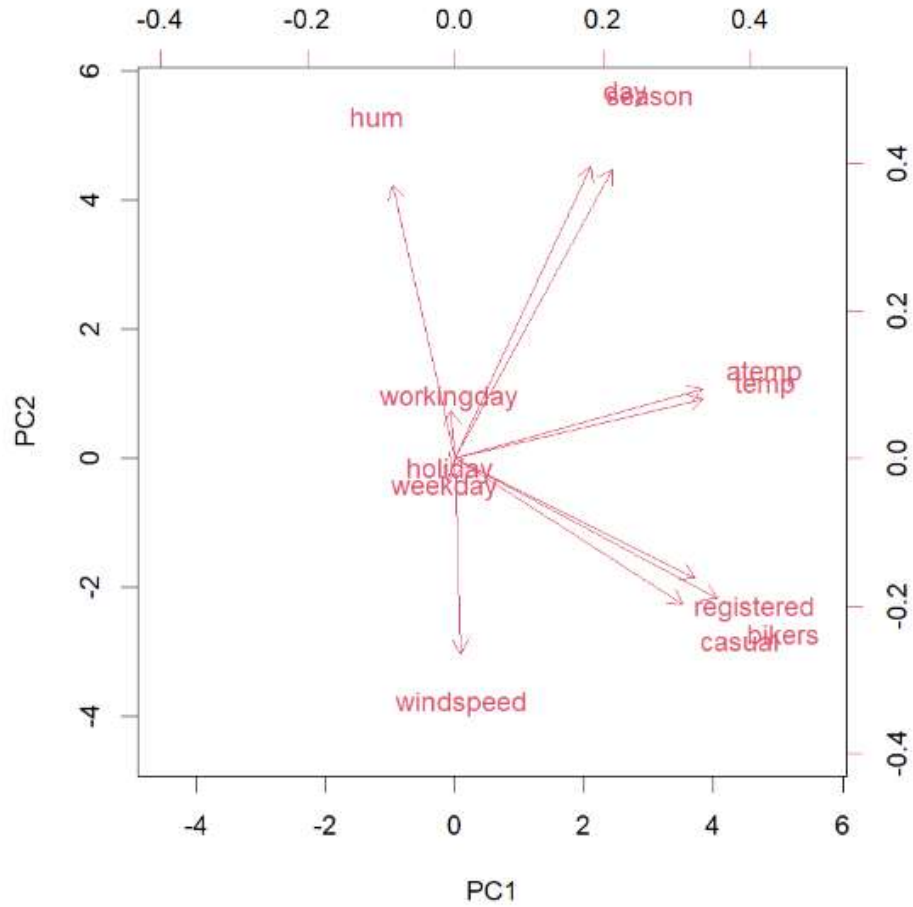


LET'S APPLY PCA TO THE DATA

2.3 VISUALIZING DIMENSIONALITY REDUCTION

VISUALIZING PRINCIPAL COMPONENTS

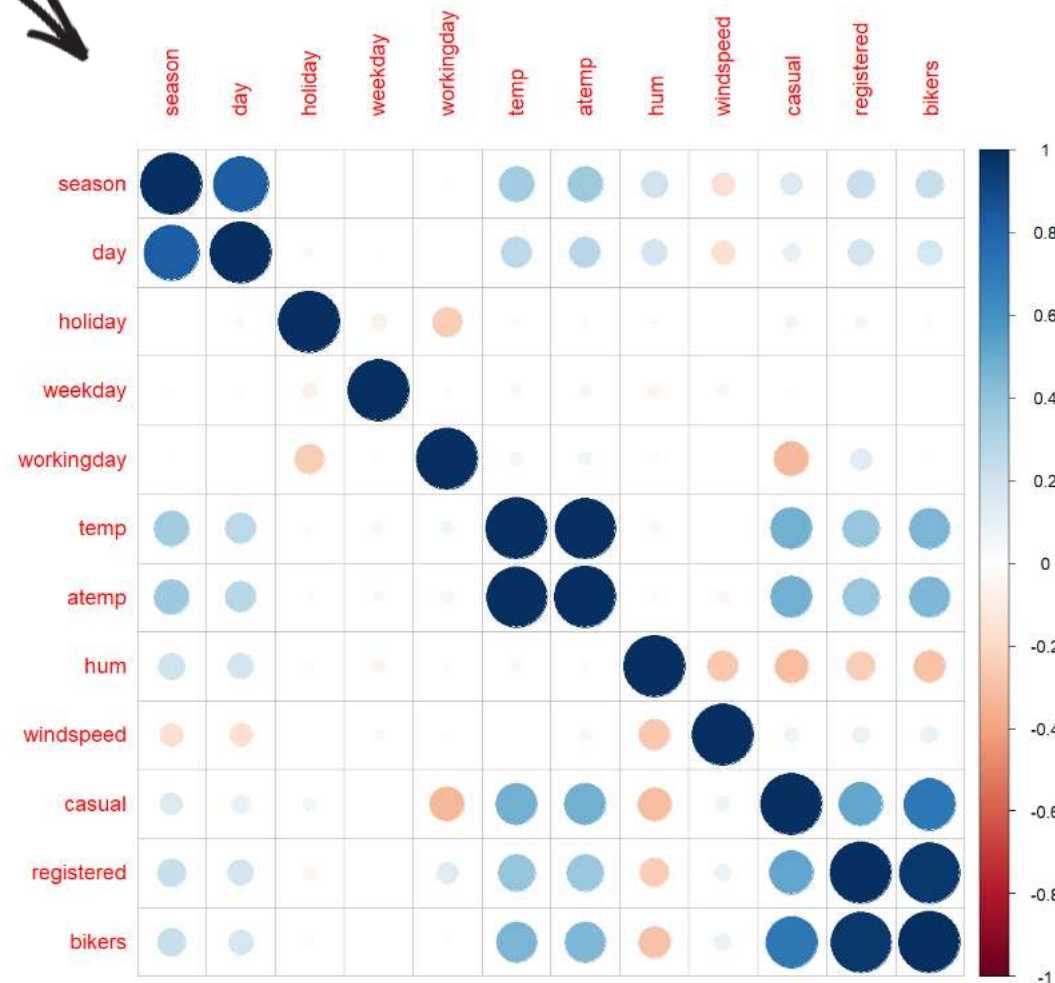
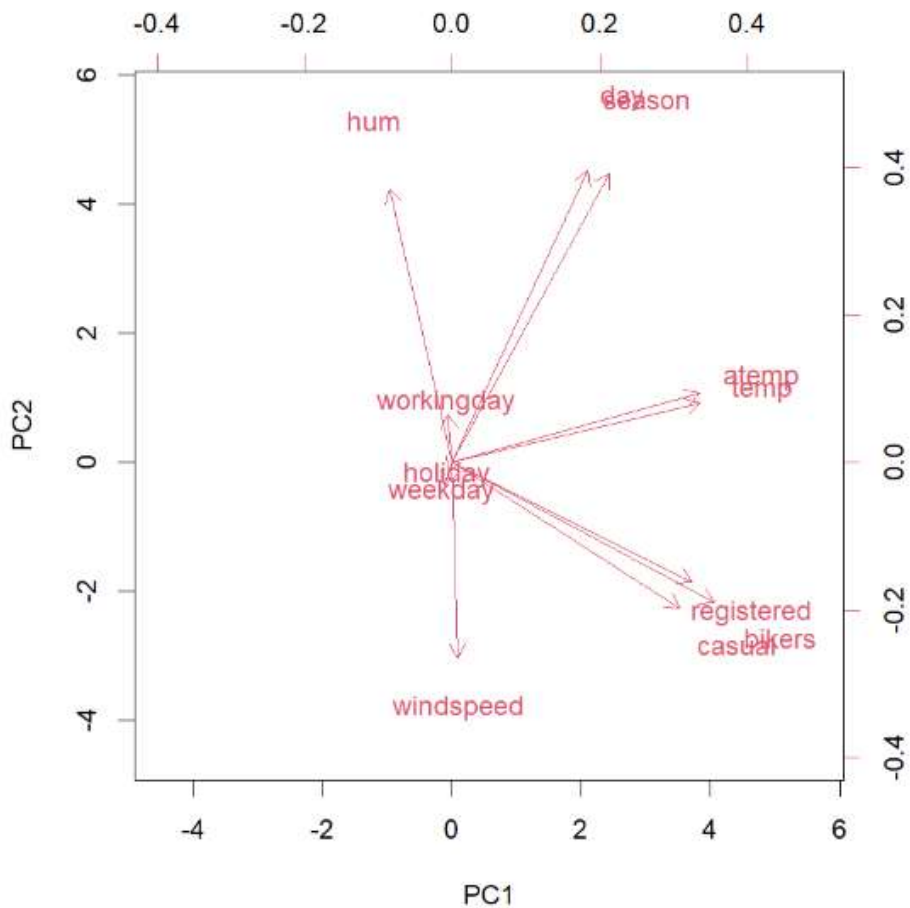
Take a minute to understand the graph. How would you interpret it?



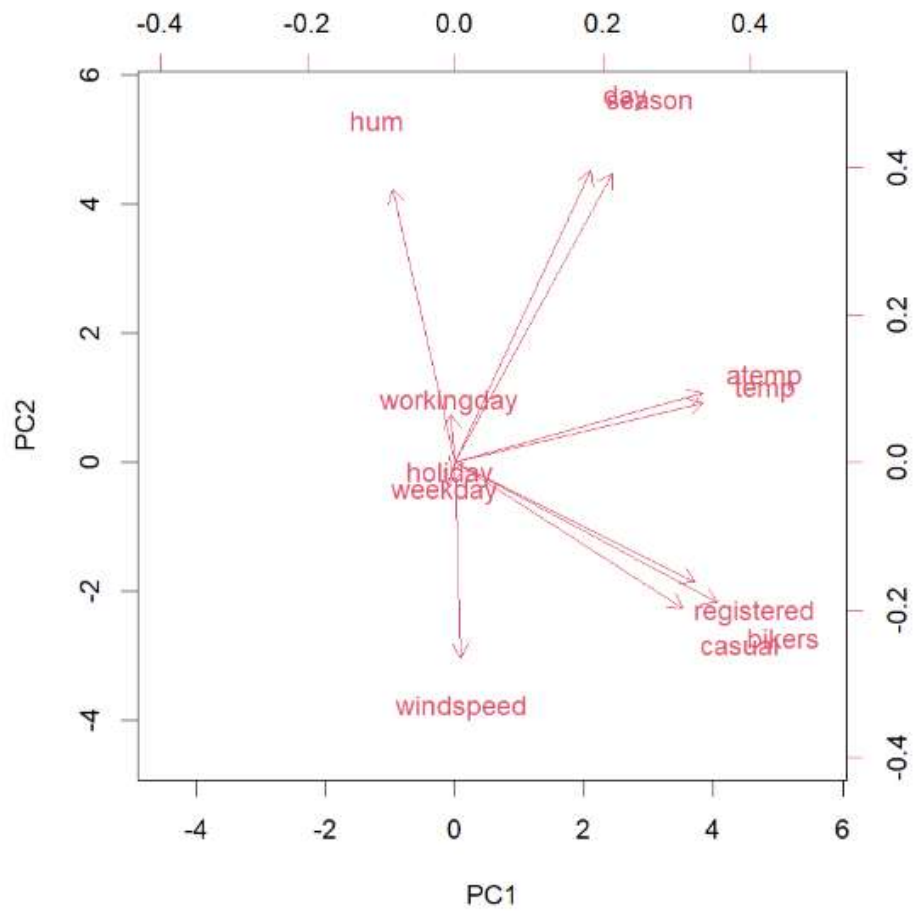
VISUALIZING PRINCIPAL COMPONENTS

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

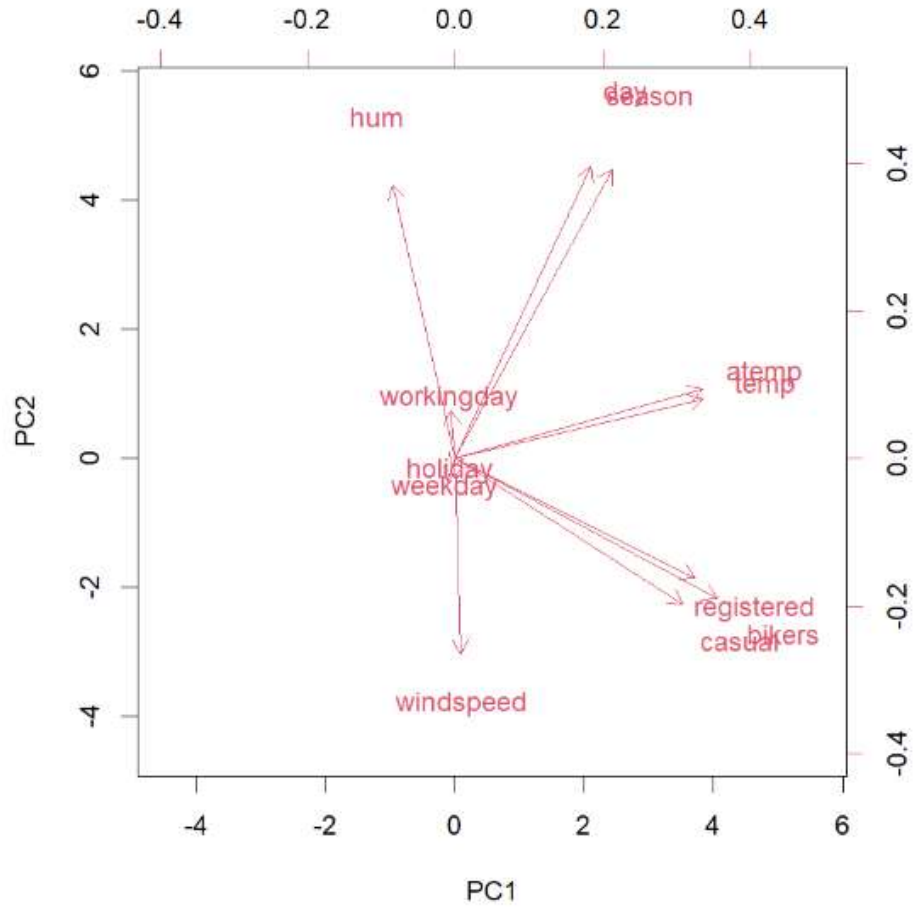
Hint



WHAT VARIABLES ARE REPRESENTED IN EACH PC?

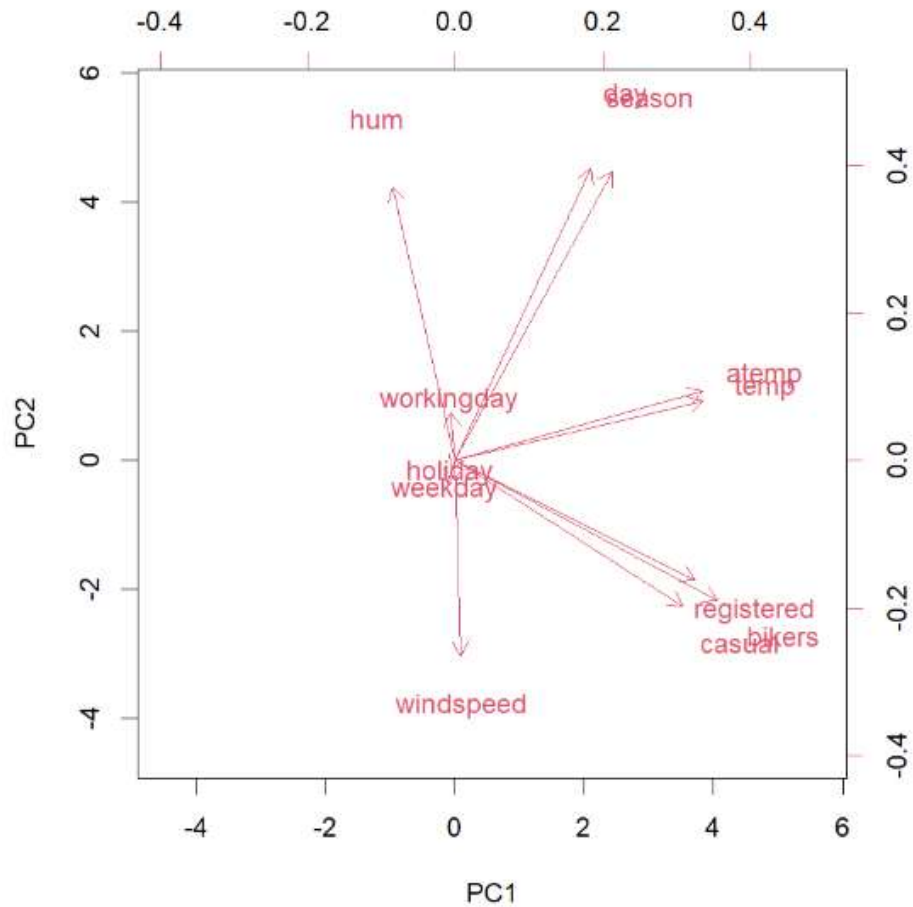


WHAT VARIABLES ARE REPRESENTED IN EACH PC?

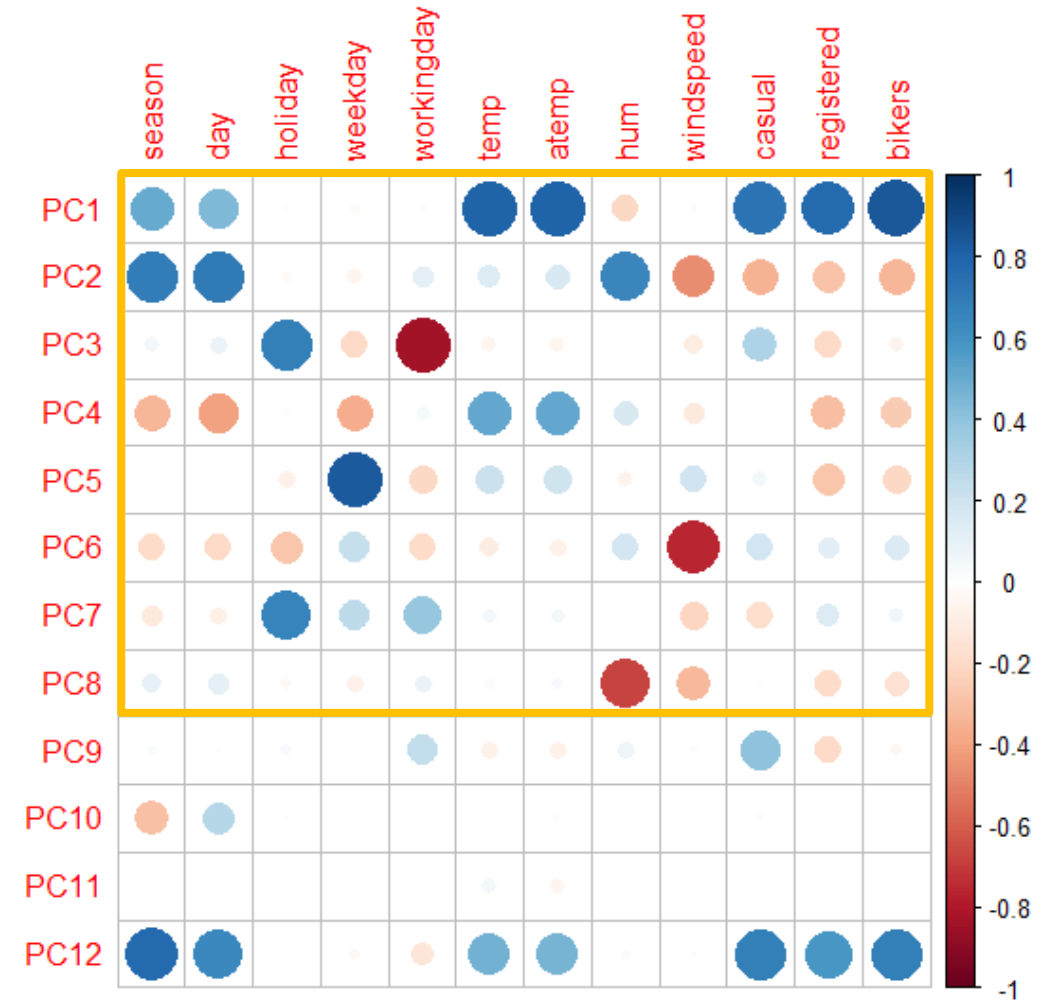


Let's check the correlations between the principal components and the original data.

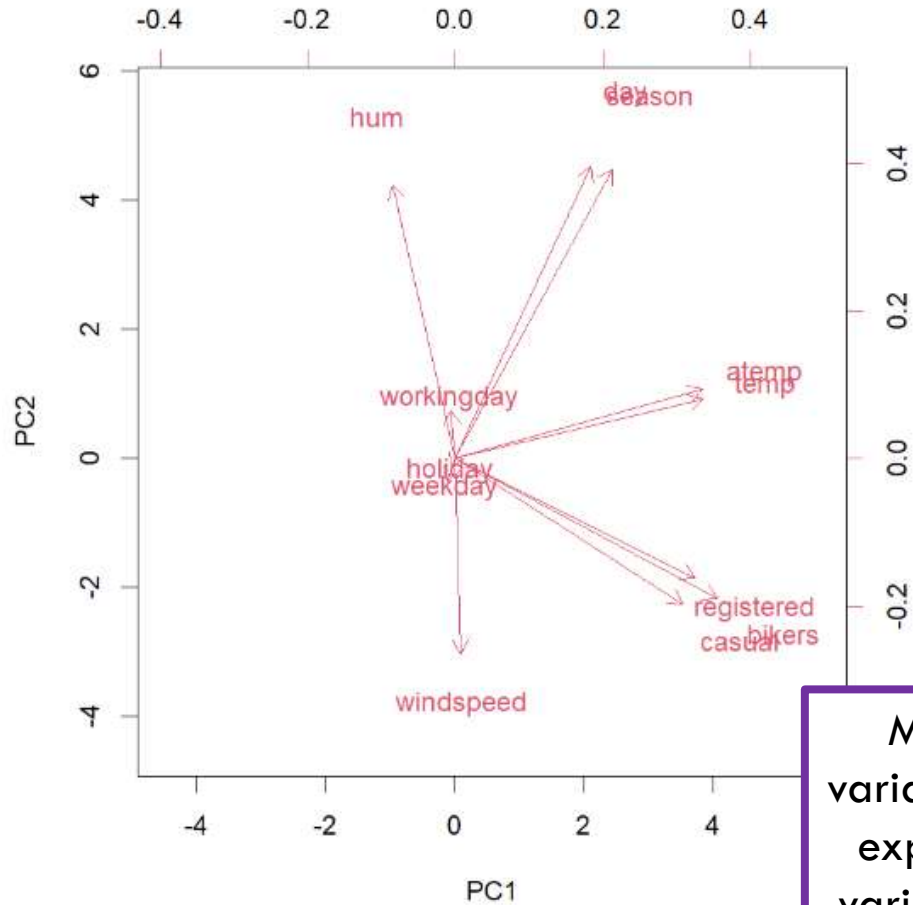
WHAT VARIABLES ARE REPRESENTED IN EACH PC?



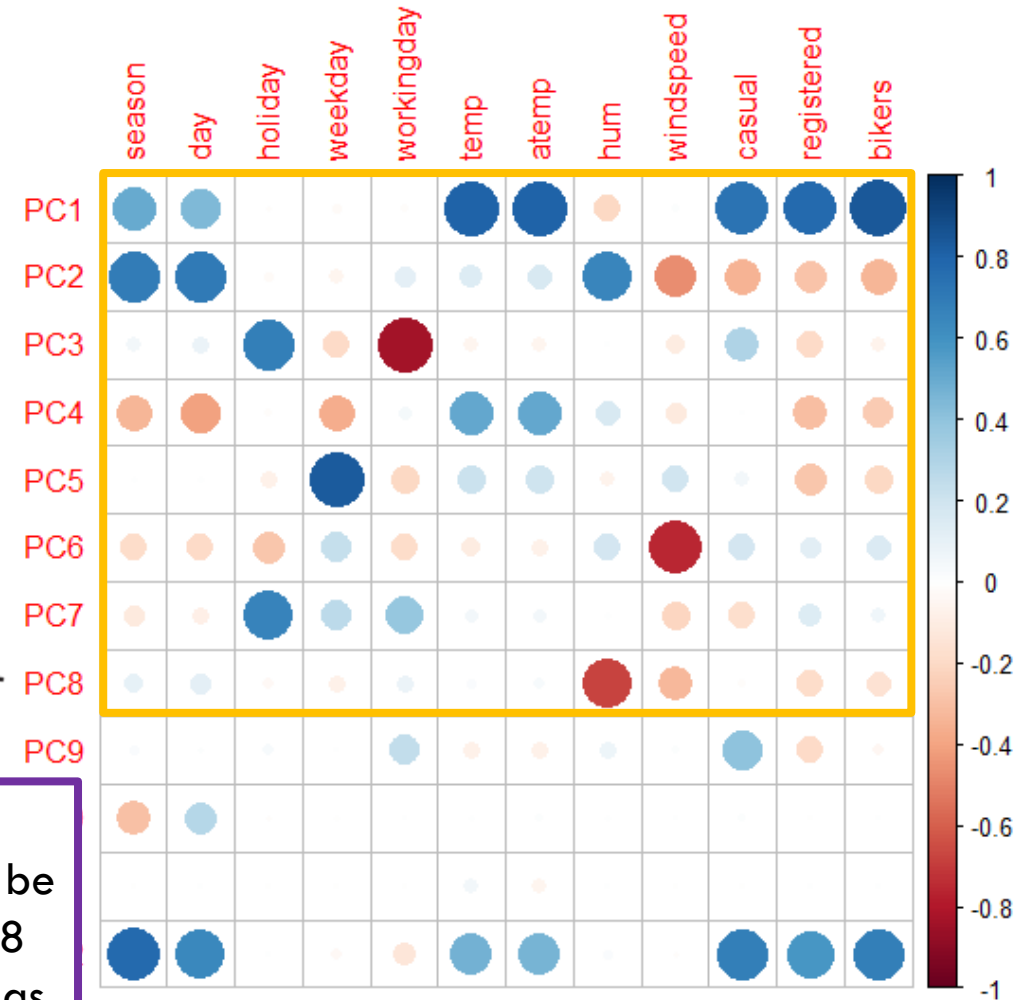
Let's check the correlations between the principal components and the original data.



WHAT VARIABLES ARE REPRESENTED IN EACH PC?

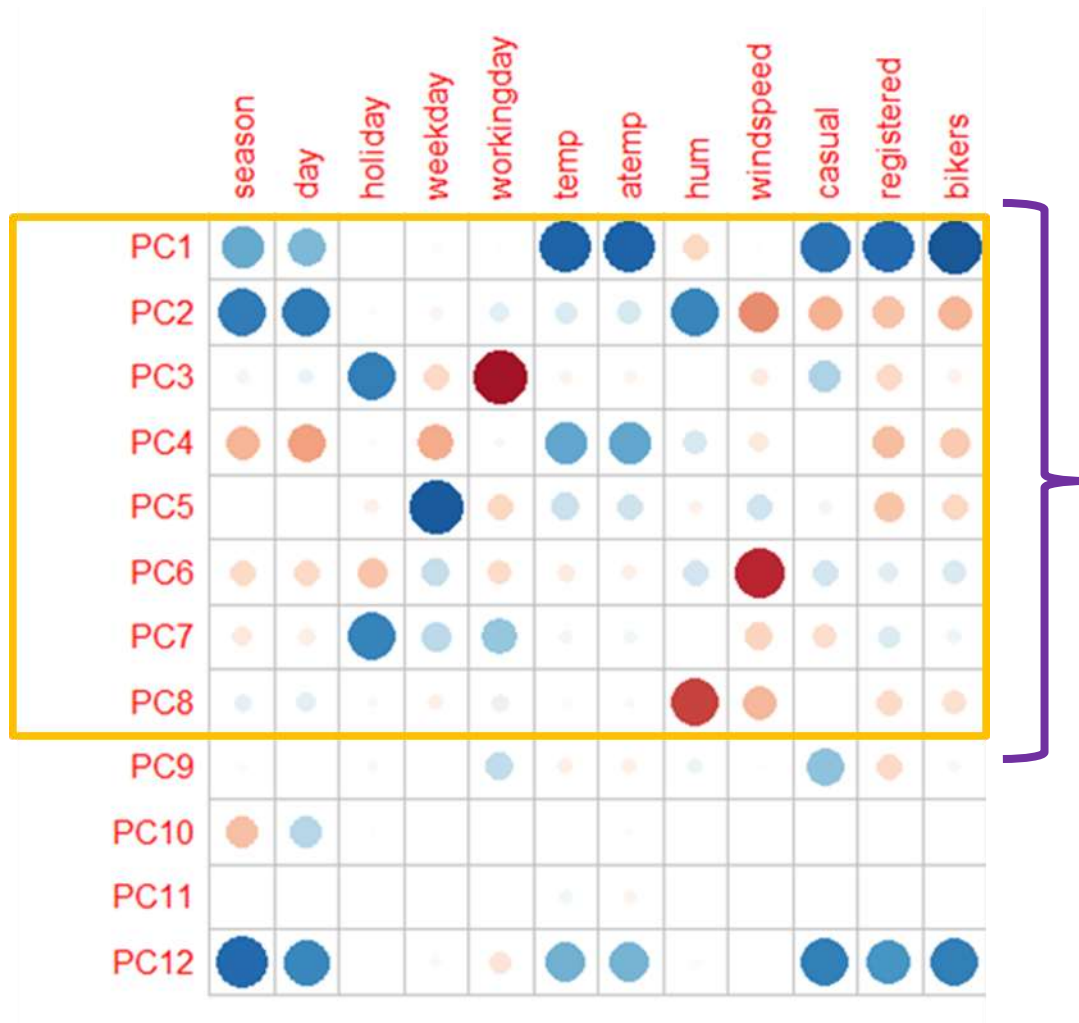


Let's check the correlations between the principal components and the original data.



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.

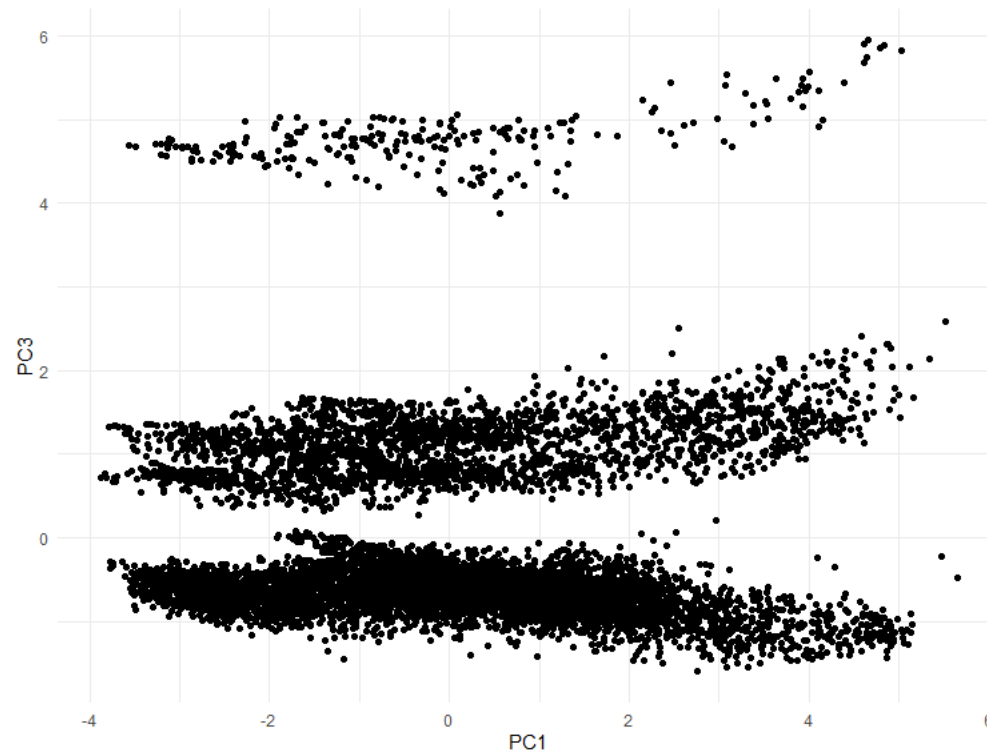
The idea behind K-means clustering is that a good clustering is one for which the within-cluster variation is as small as possible.

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

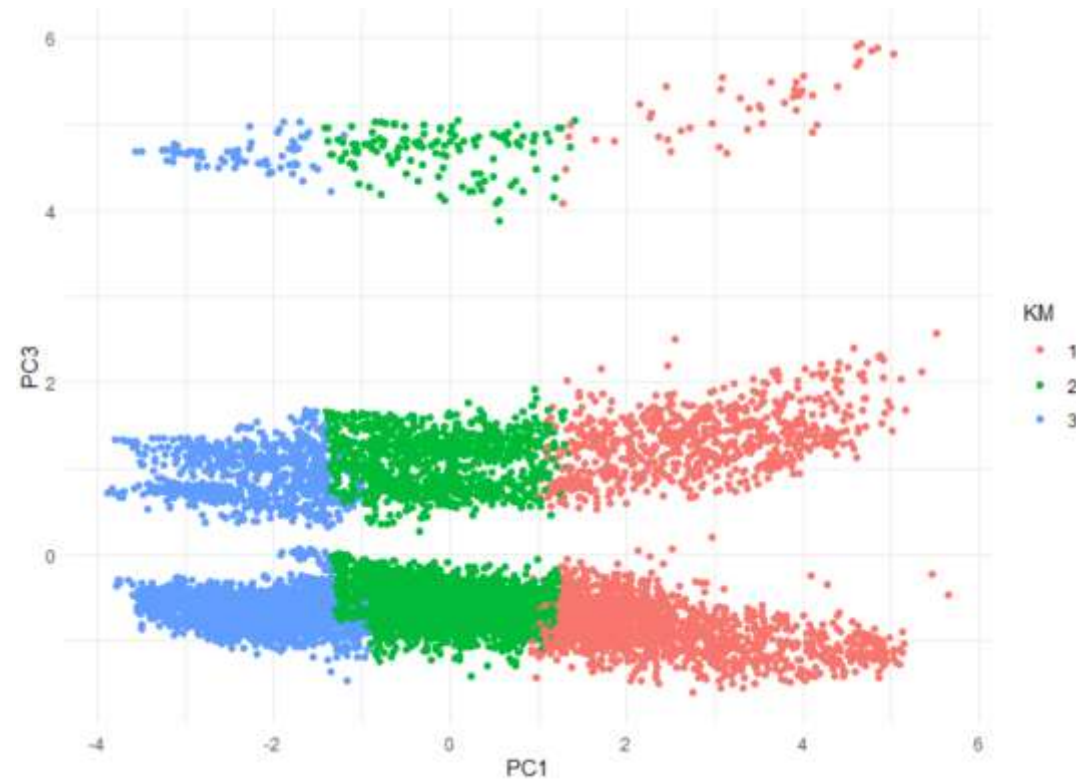
Visualize different pairs of PC. Can you find something interesting?



LET'S APPLY K-MEANS

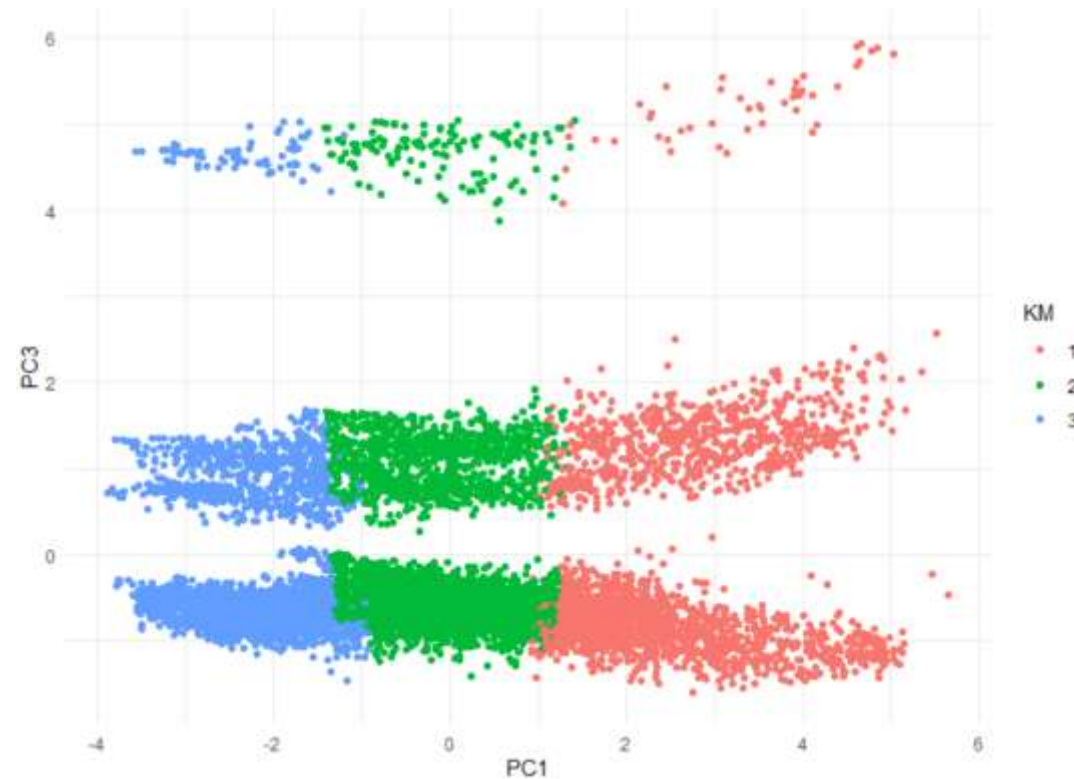
2.5 VISUALIZING CLUSTERS

LET'S VISUALIZE THE RESULTS

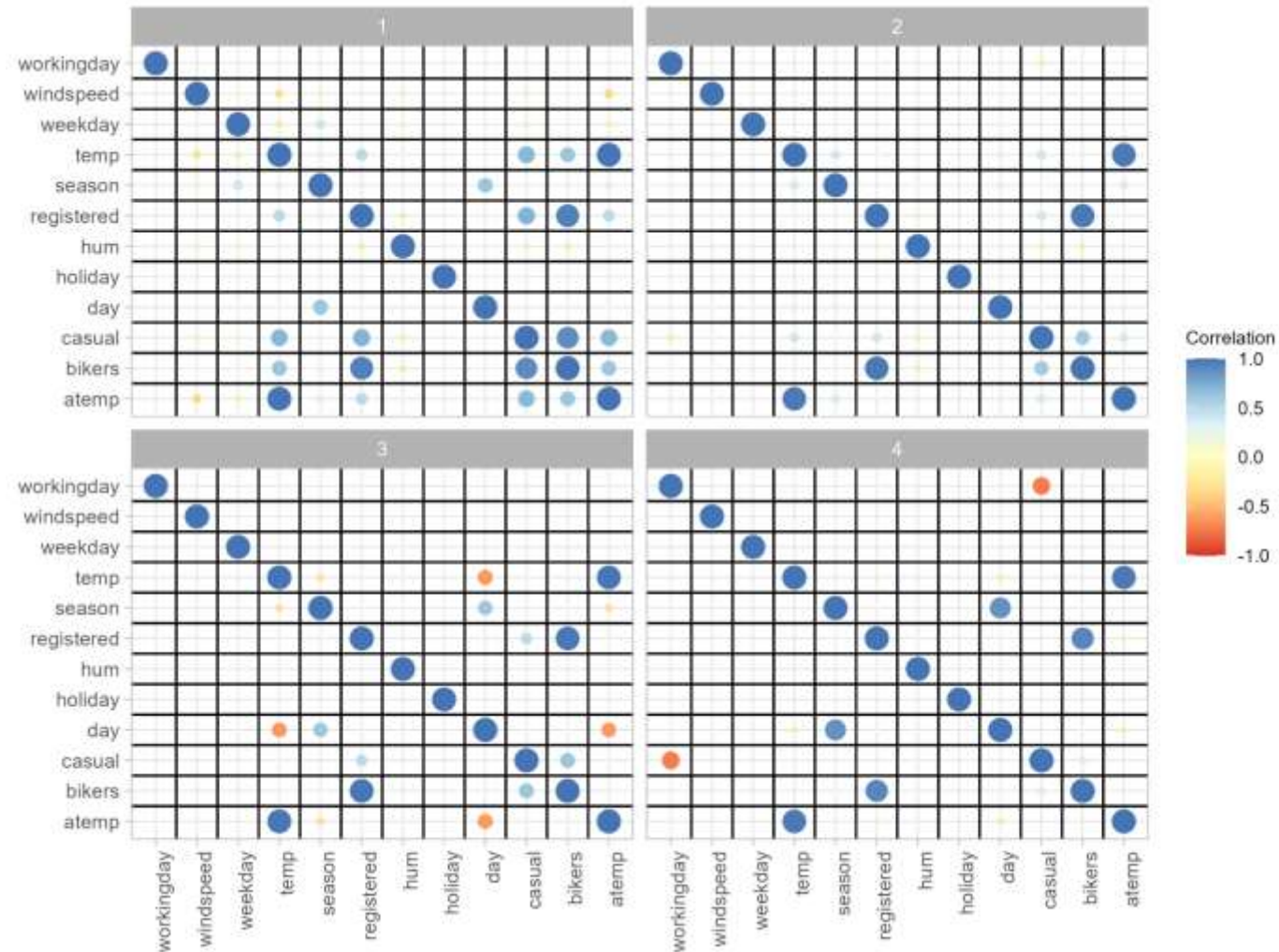


LET'S VISUALIZE THE RESULTS

Is this wrong?



LET'S VISUALIZE THE RESULTS





Join us!



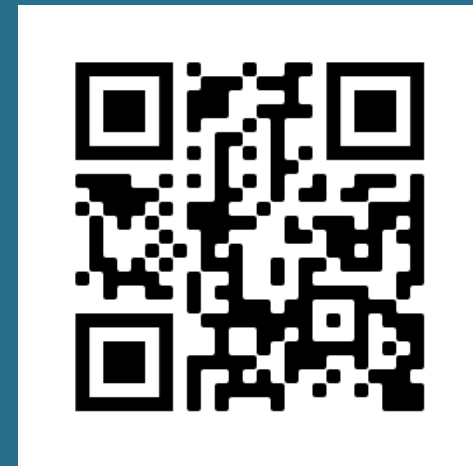
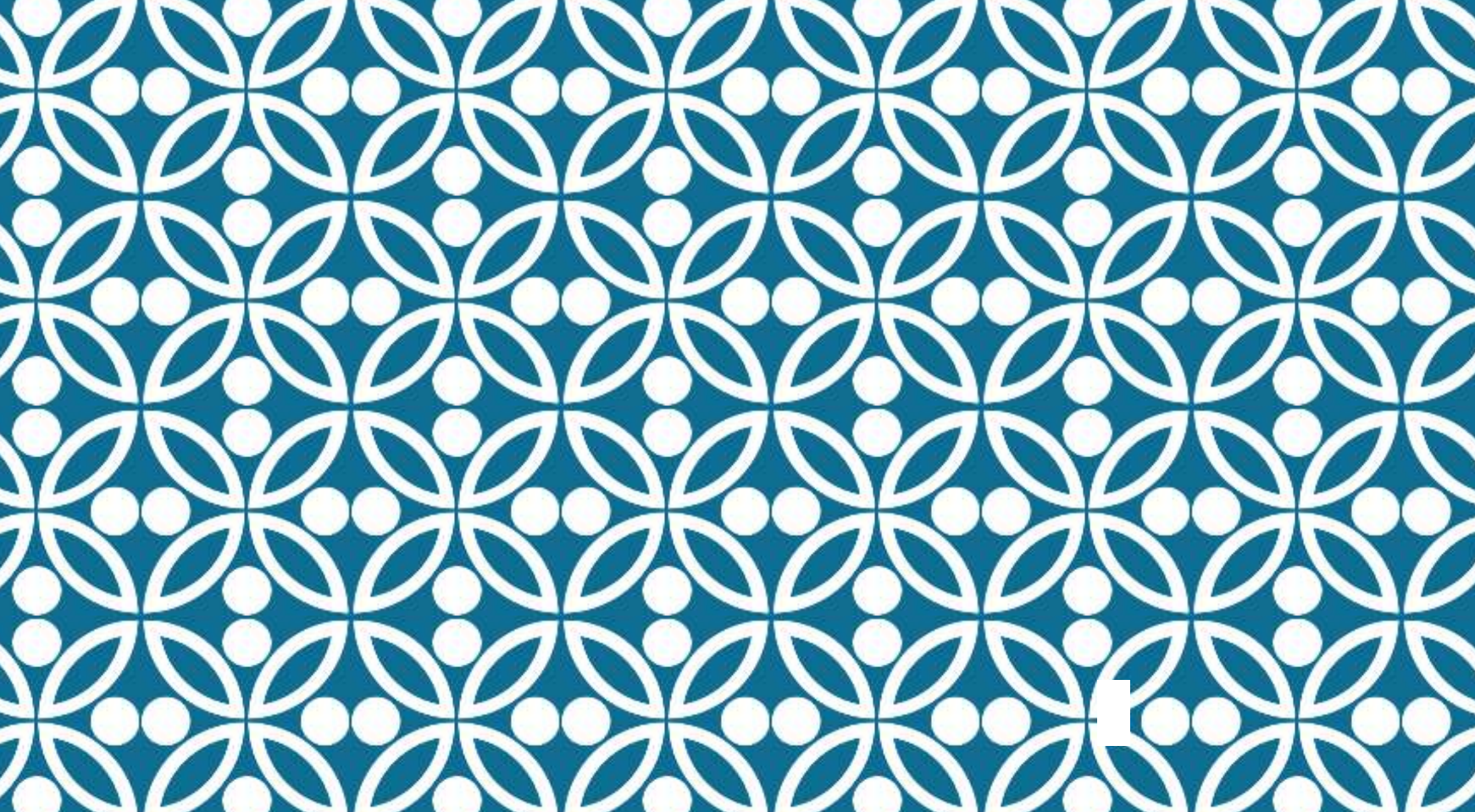
<https://forms.office.com/e/8Bgd2YsasJ>

All about the lab:

<https://societal-analytics.nl/>

Contact us at:

analytics-lab.fsw@vu.nl



<https://sofiagil.github.io/>

THANKS!

Dr. Sofia Gil-Clavel