

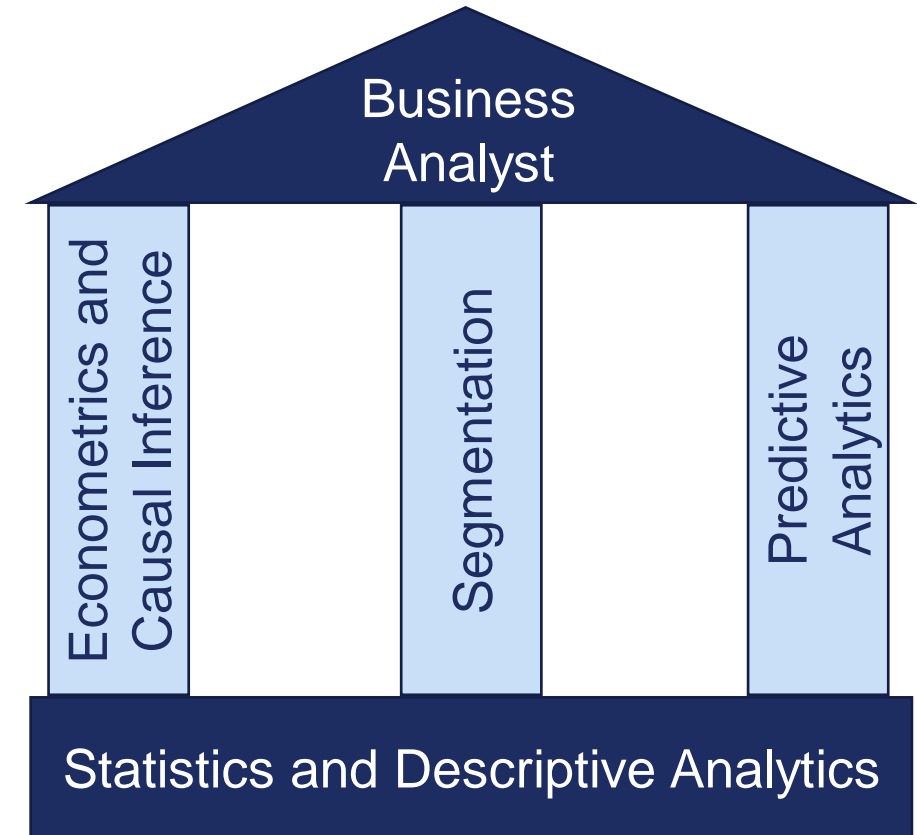
# **Business Analyst course**

# Introduction to the course

## Description

---

- 1 This video is dedicated to bureaucracies
- 2 The course has statistics as the base
- 3 A business analyst needs to know 3 Analytics types
- 4 The course is practice-focused
- 5 The course materials are in the next lecture



# The Modern-day Business Analyst

## Description

---

- 1 Being only good with numbers is a thing of past
- 2 Proficient with Statistics & Analytics methodologies
- 3 A bridge between technical and non-technical people

# The impact of weather on sales

## Description

---

- 1 Weather influences seasonal industries
- 2 External factors are uncontrollable by nature
- 3 How to prove weather influences sales?
- 4 If weather influences, then sales move when weather changes and are constant, all else being equal
- 5 The Technique I used was Google Causal Impact

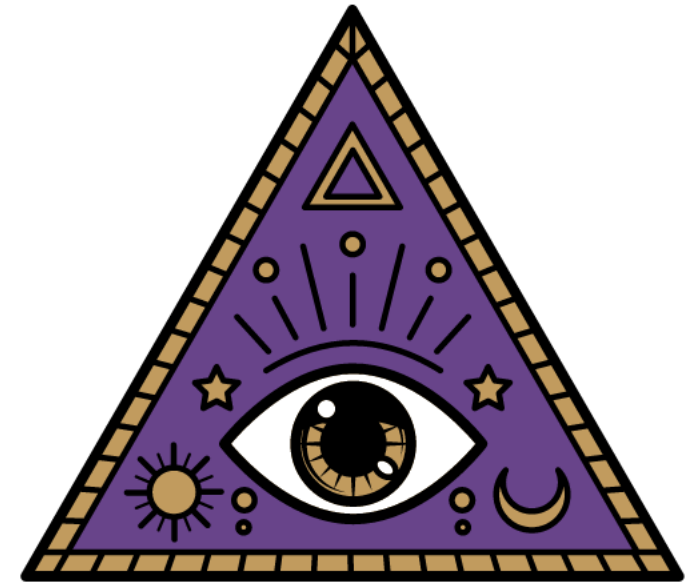


# Predicting the future

## Description

---

- 1 Commercial teams are belief-driven in nature
- 2 Advanced Analytics give numbers to beliefs
- 3 But making the change is difficult
- 4 Simple and interpretable usually has high errors
- 5 One of the techniques used was Facebook Prophet



# **BASIC STATISTICS**

# Game Plan

## Description

---

- 1 Backbone for the full course
- 2 Master the principles to make it easier in the future
- 3 You will do an exercise for each statistic learned
- 4 Moneyball case study at the end

# (Arithmetic) Mean

## Description

Same thing as average

When we say mean, we refer to the arithmetic mean

Represents the expected value

## Methodological Representation

$$\bar{x} = \frac{\sum x_i}{n}$$

## Visualization





# Case Study

## Briefing –

## Baseball

### Description

---

- 1 We have a dataset with baseball teams' data from 1962 to 2012
- 2 There are 12 KPIs for each Team, League and Year
- 3 We will practice statistical concepts on the dataset

# Mode and Median

## Mode

---

The most frequent number in a set

Fashion is a statistical term 😊

## Visualization

---

X	2	3	3	5
---	---	---	---	---

Mode is 3

## Median

---

The central number of an ordered set

If even numbers, then you average both middle points

**Used with skewed dataset**

## Visualization

---

X	2	3	5
---	---	---	---

Median is 3

X	2	3	5	10
---	---	---	---	----

Median is 4

# (Pearson) Correlation

## Description

Measures the relationship strength between 2 variables

Varies between -1 and 1

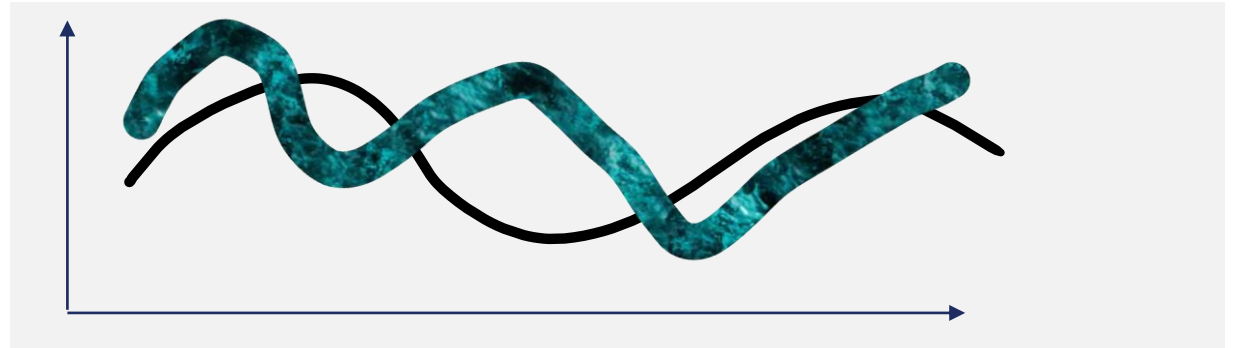
1 means strong positive relationship

-1 means strong negative relationship

0 indicates no relationship

Correlation does not imply causation

## Visualization



## Methodological Representation

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

# Standard Deviation

## Description

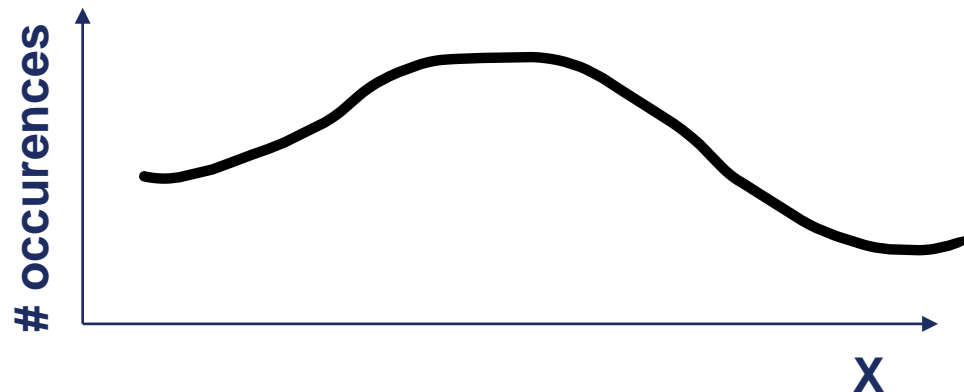
Measures the variation or dispersion of a set of values

High values mean higher variability

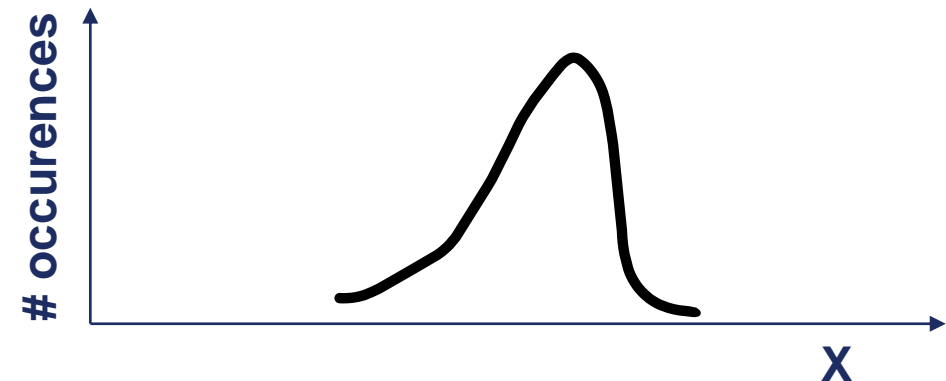
## Methodological Representation

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

## High variability



## Low variability



# Moneyball

## Case Study

### Description

---

- 1 Money ball is set on the world of baseball
- 2 The A's had success despite financial struggles
- 3 The team looked for undervalued players
- 4 Other teams did not look at statistics
- 5 The General Manager looked at specific statistics
- 6 With the right system, you can beat anyone

# **INTERMEDIARY STATISTICS**

# Game Plan

## Description

---

- 1 Level up our statistics game
- 2 Please use the Q&A
- 3 We have 2 datasets

# Normal Distribution aka Gaussian Distribution

## Description

Symmetric distribution with the mean in the middle

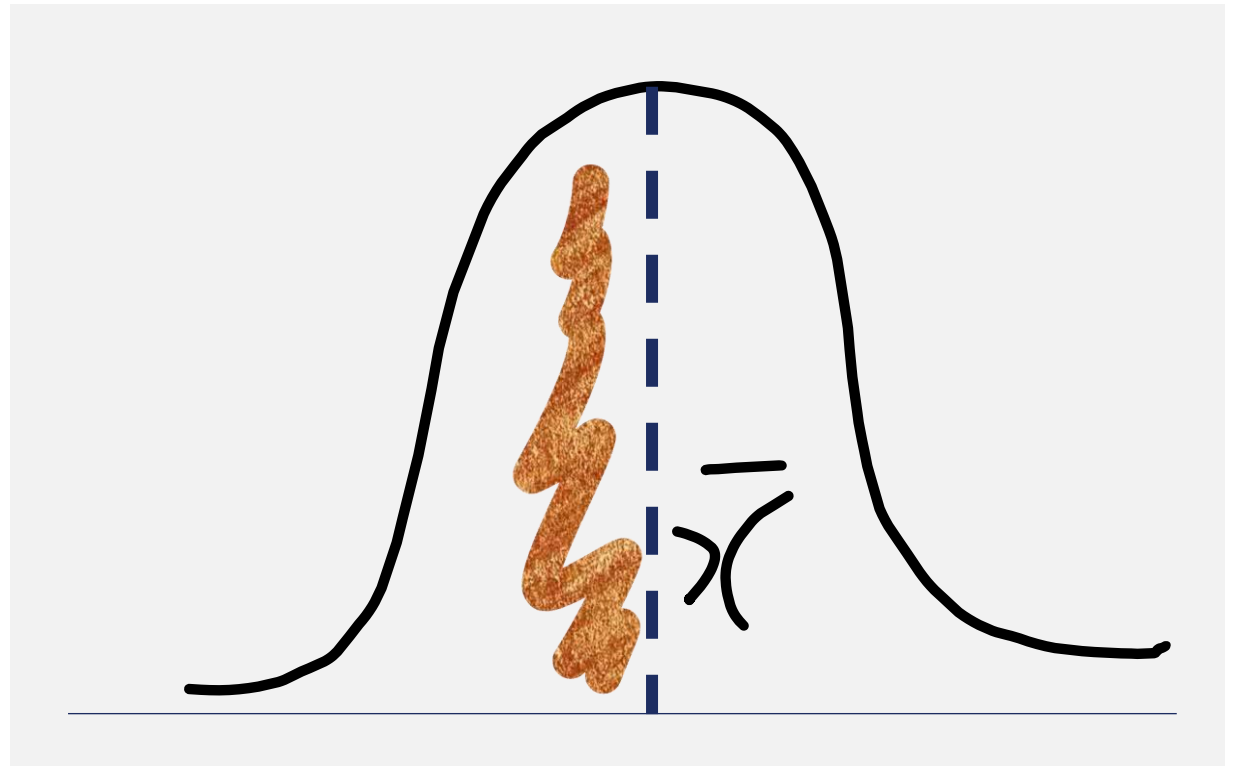
Data occurring near the mean is more frequent

Graph is similar to bell shaped curve

Statistical methods, i.e. regression assumes normalization of errors

In real-life, there will be some degree of similarity in most problems

## Visualization





# 68-95-99 Rule in Normal Distributions

## Description

Within  $\pm 1$  standard distributions, you can find 68% of observations

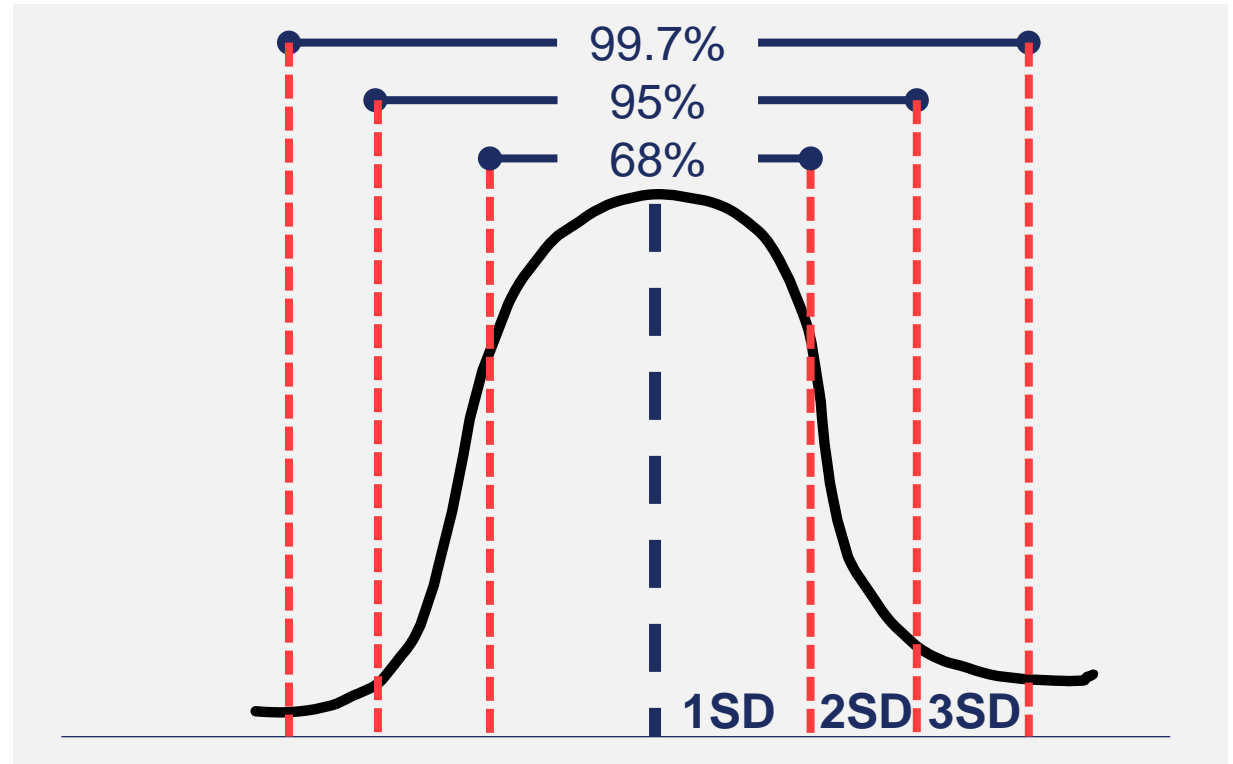
Within  $\pm 2$  SD, you should encounter 95% of deviations

Within  $\pm 3$  SD, it is 99.7%

## Key Idea

A normal distributions is a pattern, and patterns enables us to categorize data with more confidence

## Visualization



# Case Study – Wine Quality

## Description

---

**You are a wanna be Wine Statistics Connoisseur**

### Challenge<sup>1</sup>



- 1 Normal Distribution
- 2 Standard Errors
- 3 Confidence Intervals

# P-value is all about likelihood

## Description

The probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming the null hypothesis is correct.

It helps us understand what is the likelihood of “accepting” aka “fail to reject” the hypothesis

A small p-value (small probability) would mean we favor the alternate hypothesis

P-value threshold usually used: 0.05

## Examples

H0: The average salary of business analysts is €60k

H1: business analysts' average salary **is not** €60k

P-value = 0.2 -> We fail to reject the null hypothesis

-----

H0: Blueberries prevent cancer

H1: Blueberries **do not** prevent cancer

P-value = 0.01 -> We reject the null hypothesis

# Shapiro-Wilk test

## Description

---

Quantifies how likely it is that the data was drawn from a Gaussian distribution

Created in 1965 and is one of many normality tests

## Interpretation

---

H0: The distribution is gaussian

If p-value > 0.05

The distribution appears to have a normal distribution

If p-value < 0.05

The distribution does not look Gaussian -> reject the null hypothesis

# Standard Error (of the sample mean)

## Description

---

The standard error of the sample mean is an estimate of how far the sample mean is likely to be from the population mean.

Standard deviation is the degree to which individuals within the sample differ from the sample mean.

## Methodological Representation

---

$$SE = \frac{\sigma}{\sqrt{n}}$$

# Z-Score

## Description

---

Gives you an idea of how far from the mean is a data point.

Z-scores are a way to compare results to a “normal” population

It is a way to standardize values

## Methodological Representation

---

$$Z = \frac{x - \mu}{\sigma}$$

## Example – Diogo College grades

---

$$\text{Uni: } Z = (16 - 13) / (2) = 1.5$$

$$\text{GMAT: } Z = (680 - 560) / (120) = 1$$

# Confidence Interval (when $n > 30$ )

## Description

A range that gives a sense of how precisely a statistic estimates a parameter.

The associated confidence level gives the probability with which an estimated interval will contain the true value of the parameter

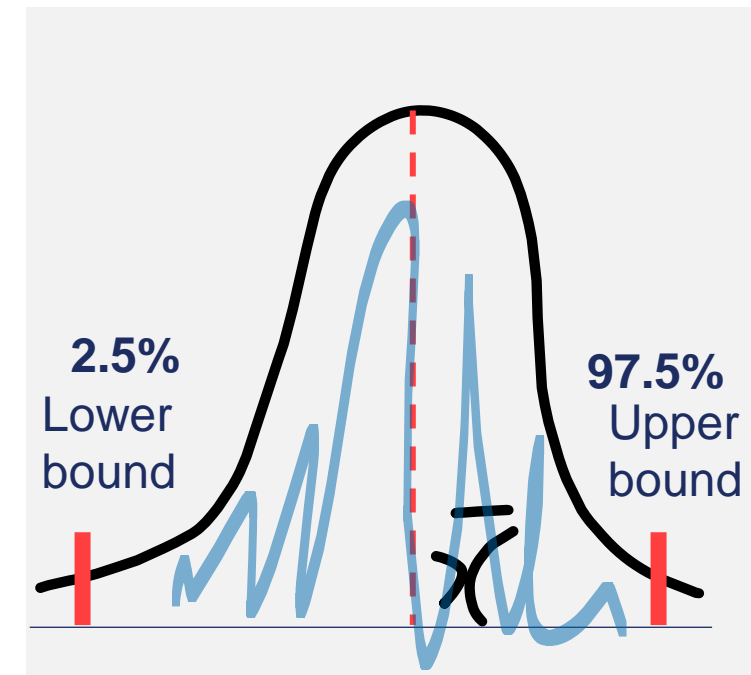
## Methodological Representation

$$CI = \bar{x} \pm z * \frac{\sigma}{\sqrt{n}}$$

## Z-values table

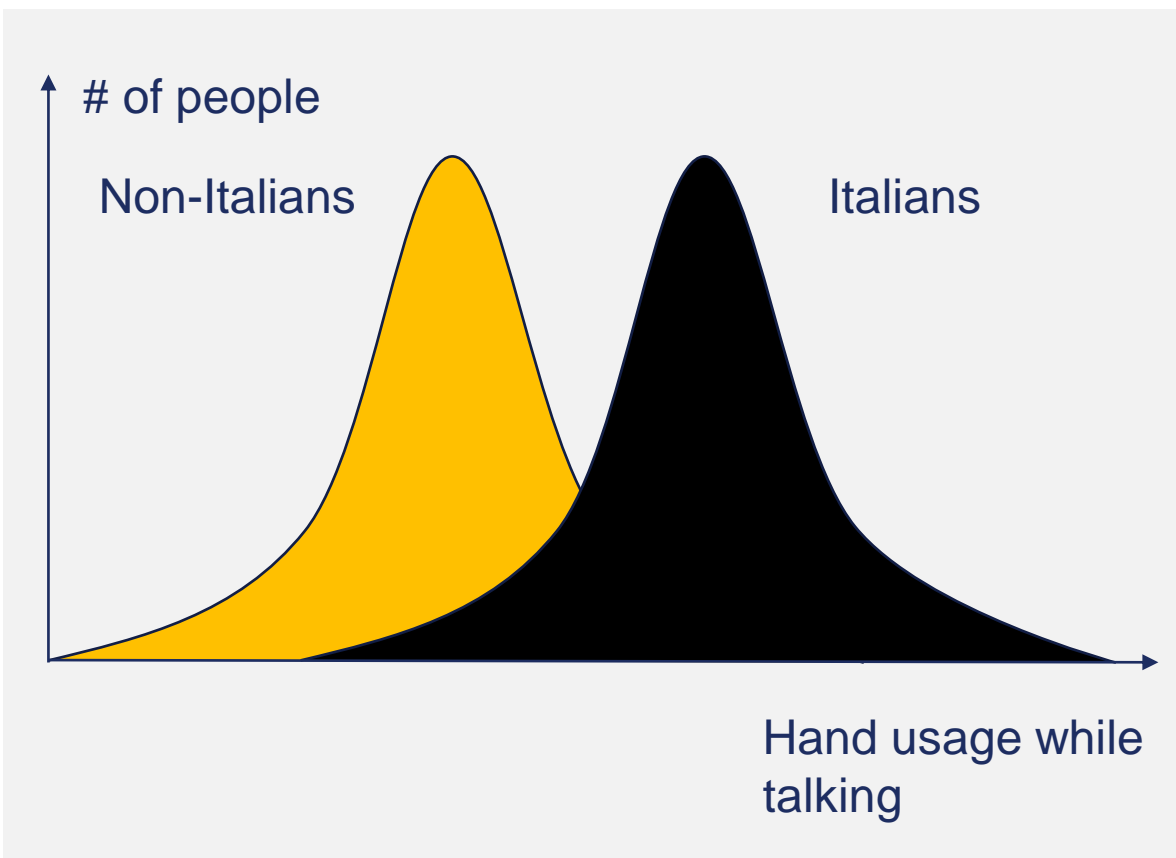
Confidence Interval	Z-Value
80%	1.28
85%	1.44
90%	1.65
95%	1.96
99%	2.58
99.9%	3.29

## Visualization – 95% CI



# T-Tests

## Visualization



## T Test formally

Test any statistical hypothesis in which the test statistic follows a Student's t-distribution under the null hypothesis.

## In practical terms

Helps us understand whether one group is (statistically) different than from the other

## How do we know?

If p-value less than 0.05, then the groups are statistically different



# Challenge – Understanding Remote Work predictions

## Challenge<sup>1</sup>







### Stack Overflow dataset

Worker's characteristics, and job related queries

- 1 T-tests
- 2 Chi-square tests

# (Person) Chi-square test

## Visualization

		Wears black	
		Yes	No
Lives in Berlin	Yes		
	No		

## Null hypothesis

There is no relationship between variables

## Chi-square test

Determine whether there is a statistically significant difference between the expected frequencies and the observed frequencies

## Difference from t-test

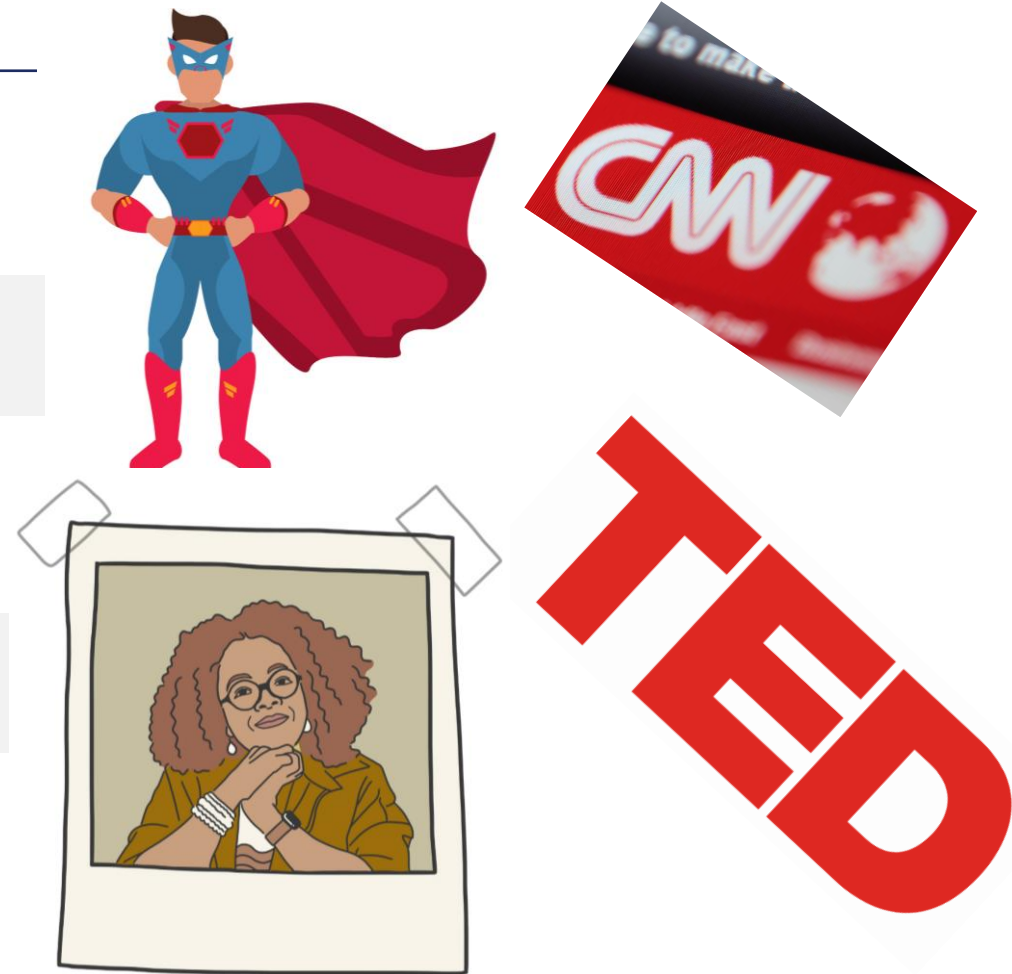
A t-test tests a null hypothesis about two means;

A chi-square test requires categorical variables, each having any number of levels.

# Powerposing and p-hacking

## Description

- 1 You put your body in a powerpose
- 2 You would perform better in high-pressure moments
- 3 Powerposing is not backed up by science
- 4 Powerposing results were not replicated by others
- 5 P-hacking is the removal of some individuals to achieve statistical significance



# **LINEAR REGRESSION**

# Game Plan

## Description

---

- 1 Building block in our learning capacity
- 2 I learned how to do it by hand. Yeah, really!
- 3 We will have a practice-focused approach

# Case Study

## Briefing –

## Pricing

## Diamonds

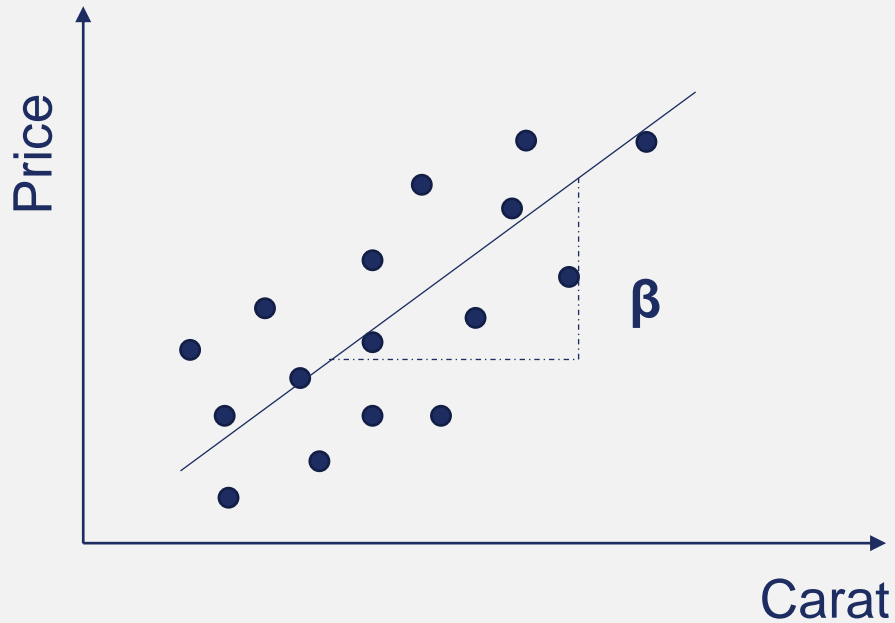
### Description

---

- 1 We have a dataset of roughly 300 diamonds
- 2 We have the price, carats and other KPIs
- 3 We want to understand how carats influence Diamond Prices

# (Linear) Regression crash course

## Visualization



## Definition

Study of a relationship between a dependent variable and at least one independent variable

## Intuition perspective

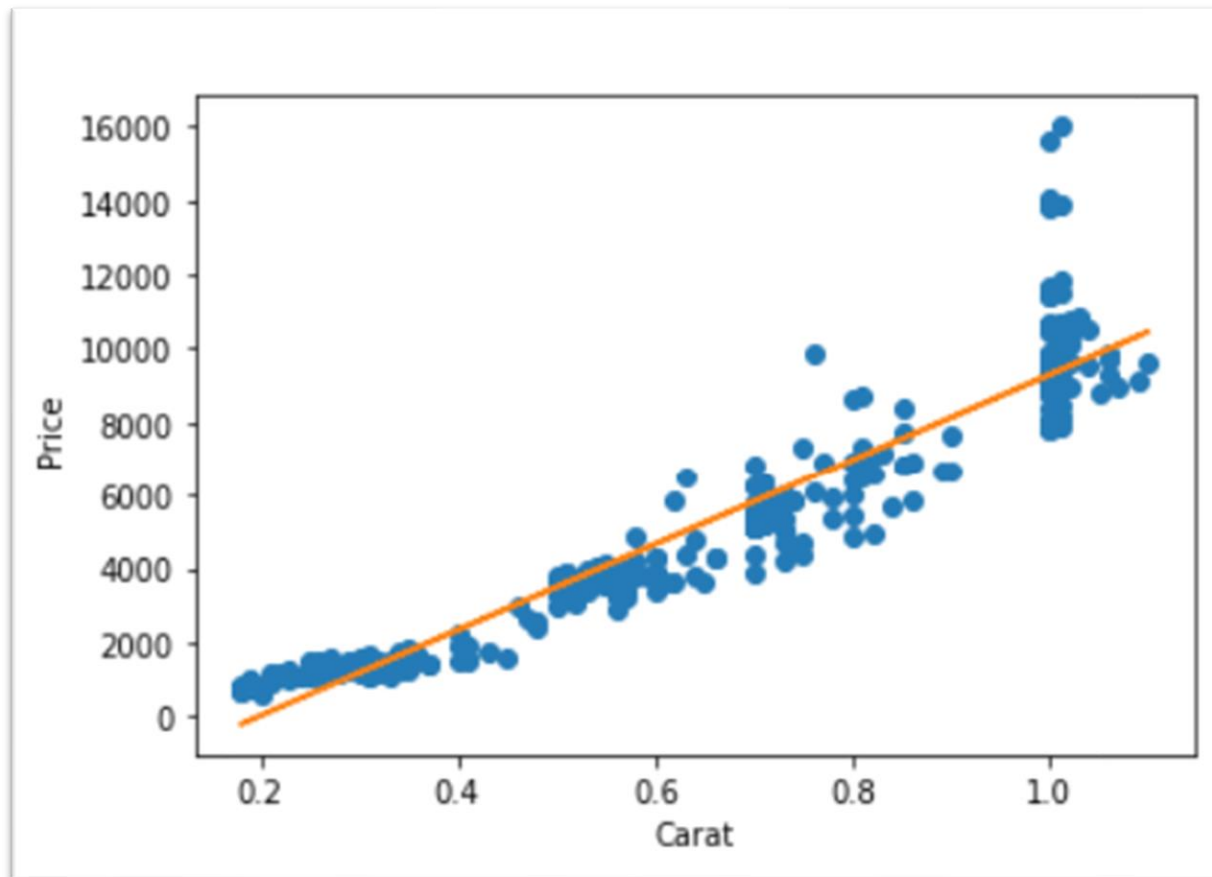
Method for “What is the impact of X on Y?”

## How is it different from a correlation?

- Correlation studies the direction
- Regression studies the impact

# Linear Regression

## Visualization



## Methodological View

$$Y = a + b * X + e$$

## Interpretation

If I increase X by 1, Y increases by b

If X happens, Y increases by b



# How to read a Regression result

## OLS Regression Results

```
=====
Dep. Variable:          price    R-squared:          0.893
Model:                  OLS      Adj. R-squared:       0.892
Method:                 Least Squares    F-statistic:       2541.
Date:                  Tue, 19 Oct 2021    Prob (F-statistic): 3.04e-150
Time:                  06:25:09    Log-Likelihood:    -2597.9
No. Observations:      308        AIC:                5200.
Df Residuals:          306        BIC:                5207.
Df Model:              1
Covariance Type:       nonrobust
=====
```

```
=====
               coef      std err          t      P>|t|      [0.025      0.975]
-----
const      -2298.3576    158.531    -14.498    0.000    -2610.306    -1986.410
carat       1.16e+04     230.111     50.406    0.000     1.11e+04     1.21e+04
=====
```

```
Omnibus:            170.301    Durbin-Watson:           1.216
Prob(Omnibus):       0.000    Jarque-Bera (JB):        1324.411
Skew:                2.168    Prob(JB):                2.56e-288
Kurtosis:            12.187    Cond. No.                 5.13
=====
```

## Coefficients

One carat increases the price by 11.6k

## R-Squared and adj R-squared

We can explain 89.3% of the variance

## Confidence interval (95%)

The Carat coefficient is between 11.1k-12.1k

## Statistical Significance

If  $P > |t|$  is less than 0.05, we have statistical significance

# Dummy variable trap

Observation	Coca cola	Pepsi
a	1	0
b	1	0
c	1	0
d	1	0
e	0	1
f	0	1
g	0	1
h	0	0
j	0	1

## Multicollinearity

Correlation between Coca-Cola and Pepsi is -1

Solution: remove one dummy variable

## Removing does not mean information is lost

A zero also represents information

The removed dummy variable becomes part of the intercept.

You can see it as being your baseline.

# Dummy variable trap

Observation	Coca cola	Pepsi	White Label
a	1	0	0
b	1	0	0
c	1	0	0
d	0	0	1
e	0	0	1
f	0	1	0
g	0	1	0
h	0	0	1
j	0	1	0

## Multicollinearity

Correlation between Coca-Cola and Pepsi is -1

Solution: remove one dummy variable

## Removing does not mean information is lost

A zero also represents information

The removed dummy variable becomes part of the intercept.

You can see it as being your baseline.

# **MULTILINEAR REGRESSION**

# Game Plan

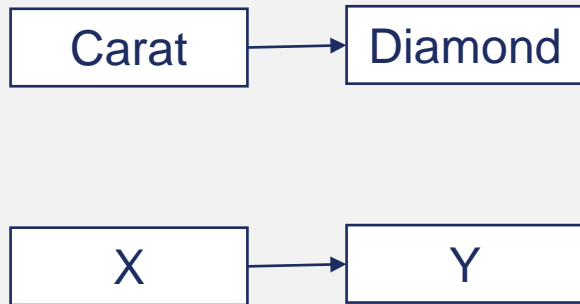
## Description

---

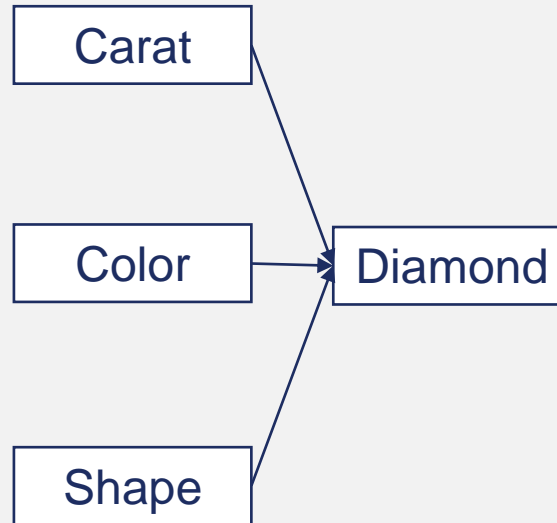
- 1 Topics: outliers, assessment ad overfitting
- 2 Practice tutorial: Teacher's salaries
- 3 Challenge: Retail Store drivers
- 4 Regression value adding is interpretability

# Multilinear Regression

## Linear Regression



## Multilinear Regression



## Description

It is very rare that one input explains the output

We often need more predictors to improve the models

Beware of multicollinearity or overfitting

# Case Study

## Briefing –

## Professors’

## salaries

### Description

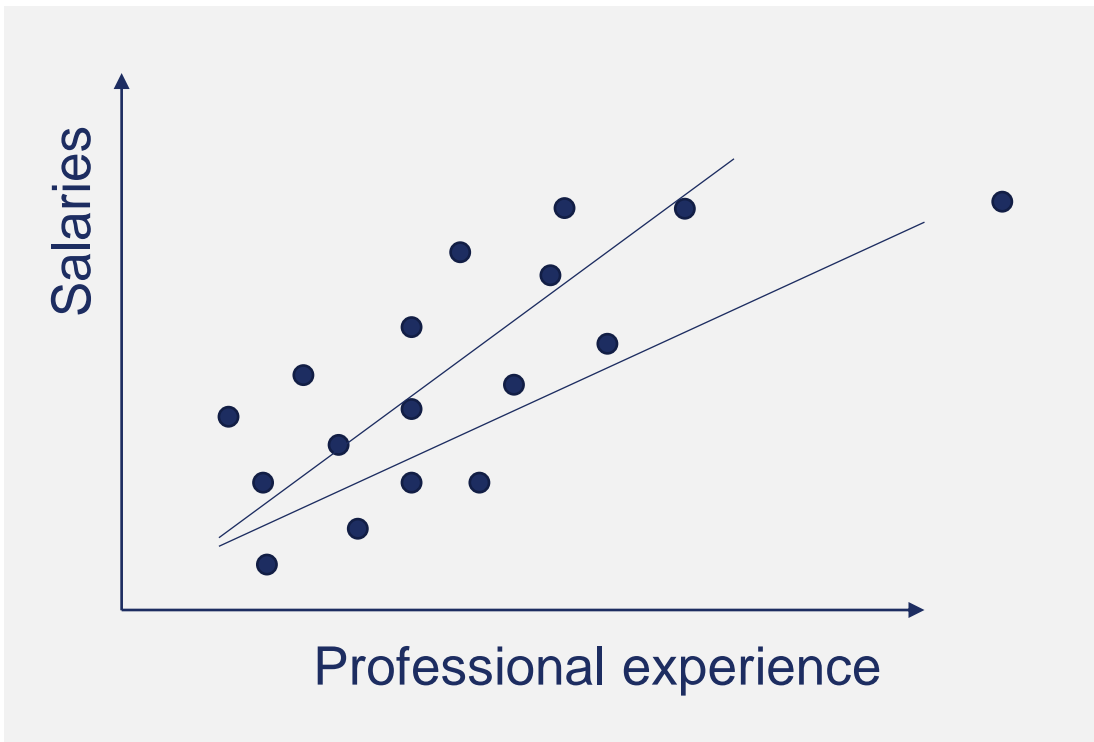
---

- 1 The 2008-09 academic salary for Professors in a college in the U.S.
- 2 The data was collected to monitor salary differences between male and female faculty members.
- 3 Use Multilinear Regression to study

# Outliers

## Visualization

---



## Interpretation

---

Outliers can damage your analysis

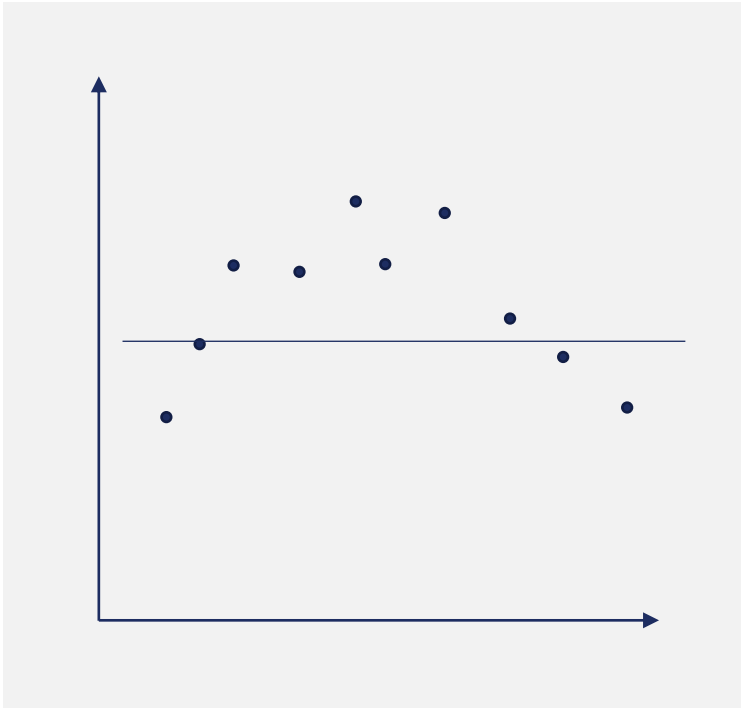
You must distinguish between noise and valuable information

Consider using models good with outliers or non-linearity (e.g., Random Forest)

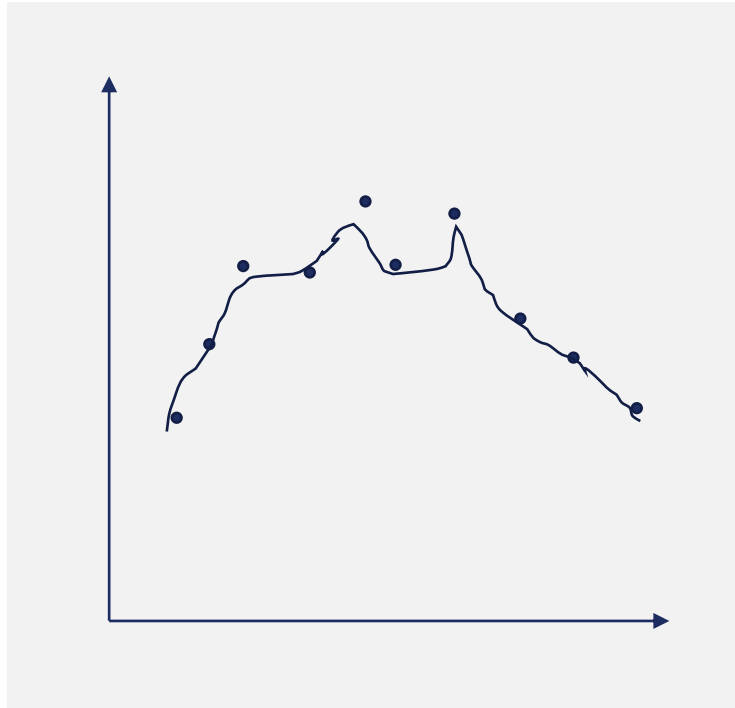


# Modelling is finding the balance between under and overfitting

## Underfitting



## Overfitting



## Insights

Having a too simple model will get you nowhere

Too complex will not yield results in other testing scenarios

You should iterate based on results

**Let's imagine this is our full data set**

**Description**

---



# Splitting between training and test enables an unbiased model assessment

Training Set



Model

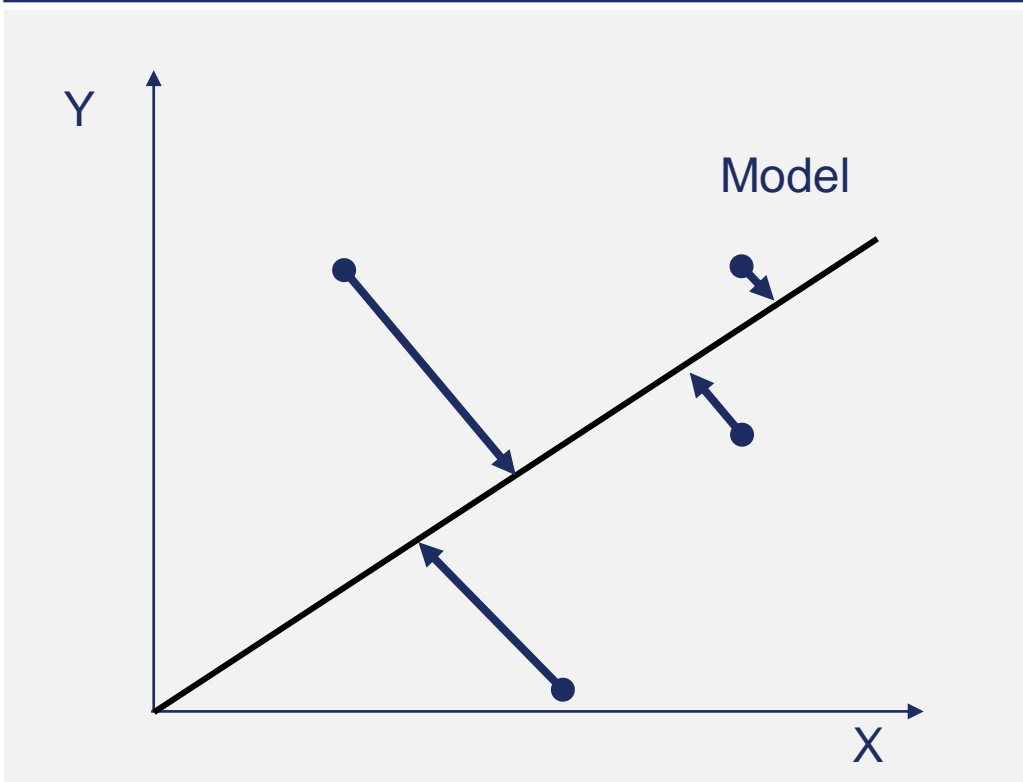
Test Set



Assessment

# Mean Absolute Error (MAE) vs Root Squared Mean Error (RSME)

## Visualization



## Key ideas

MAE and RSME are performance indicators for Regression models with continuous outputs

$$MAE = \frac{\sum |y - \hat{y}|}{n}$$

$$RSME = \sqrt{\frac{\sum (\hat{y} - y)^2}{n}}$$

RSME is useful for models with extremes / outliers

MAE is more interpretable.

# Challenge

## Description

---

### Use Multilinear Regression to study a Store sales' drivers

- 1 Pick variables for your model
- 2 Analyze the data i.e. summary statistics
- 3 Correlation Matrix
- 4 Create a Training a Test Set
- 5 Use Multilinear Regression
- 6 Assess Accuracy

Dataset: Ecdat package

# **LOGISTIC REGRESSION**

# Game Plan

## Description

---

- 1 We now face a classification problem
- 2 The question influences the analytical technique
- 3 How do we measure accuracy?
- 4 Case study: Which emails are spam?
- 5 Challenge: the sex of penguins

# Case Study

## Briefing – Is it spam?

### Description

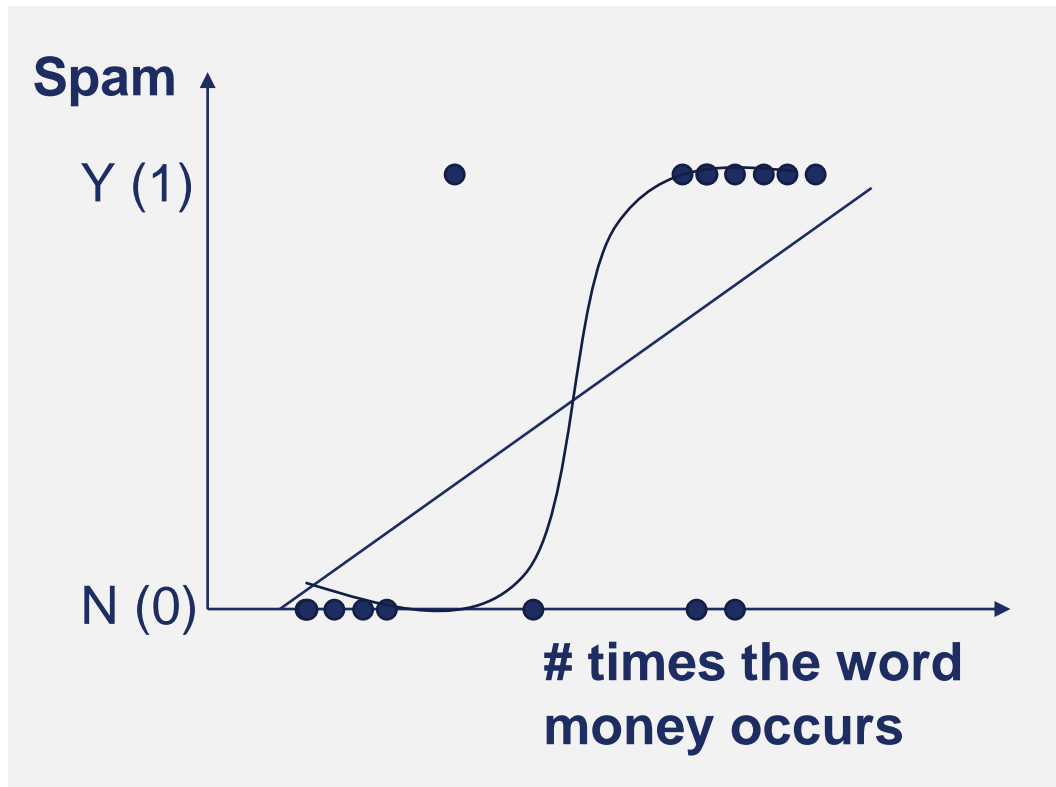
---

- 1 Dataset with ~5k emails
- 2 What makes an email spammy?
- 3 Can we predict which emails are spam?



# Logistic Regression crash course

## Visualization



## What is a Logistic Regression?

Relationship study between a discrete dependent variable and at least one independent variable

## From an Intuition Perspective?

What is the impact of X on Y happening?

## How is it different from a Linear Regression

Linear is for continuous, logistic is discrete

Linear we fit a straight line, logistic a curve

# How to read a Logistic Regression coefficients1

## Linear Regression

$$Y = a + b * X + e$$

## Interpretation

For each X unit increase, Y increases by ***b***

**B = 0.5:** For each X unit increase, Y increases by ***0.5***

## Logistic Regression

$$Y = \frac{e^{a+bX}}{1 + e^{a+bX}}$$

For each X unit increase, the probability of Y happening increases by ***exp(b) – 1 \* 100 %***

**B = 0.5:** For each X unit increase, the probability of Y happening increases by ***64%***

# The Confusion Matrix allows to access the results of a classifier

## Confusion Matrix

		Predicted	
		False	True
Actuals	False	True Negative	False Positive
	True	False Negative	True Positive

## Accuracy

Accuracy = (True positive + True negative ) / All

Balanced dataset

## F1-Score

F1-score =  $2 * TP / (2 * TP + FP + FN)$

Unbalanced dataset

# The Confusion Matrix allows to access the results of a classifier

## Confusion Matrix

Actuals	Predicted	
	False	True
	False	True
False	True Negative	False Positive
True	False Negative	True Positive

## Specificity or True Negative Rate

$\text{True Negative} / (\text{True Negative} + \text{False Positive})$

When we focus in False values accuracy

## Sensitivity, Recall or True Positive Rate

$\text{True Positive} / (\text{True Positive} + \text{False Negative})$

Focus is on True values

# Challenge

## Description

---

**Use Logistic Regression to predict the sex of penguins**

- 1 Pick variables for your model
- 2 Plot Histograms of the character variables
- 3 Transform the character variables into binary
- 4 Create a Training a Test Set
- 5 Use Logistic Regression
- 6 Assess Accuracy through the classification report

# **GOOGLE CAUSAL IMPACT**

# Why Econometrics and Causal Inference

## Description

---

- 1 Decision-Making
- 2 Understand and tackling biases

# According to BNP Paribas, Sustainability-focused companies perform better

Exhibit 1: Sustainability practices and business results





# BNP Paribas

## Description

---

- 1 Are there other differences between sustainability-focused companies and others?
- 2 People define politics and decision
- 3 Not including all factors is falling into the omitted variable bias


# Does Smoking prevent Parkinson's?

Continue Your  
Medical Education  
With Ease

Earn CME/CE Credits Conveniently on

**NeurologyAdvisor**

Get Started Now

CME/CE Powered by: 

Topics » [Movement Disorders](#)

December 10, 2015

## The Troubling Link Between Parkinson's and Smoking: Can We Deny the Benefits?

Tori Rodriguez, MA, LPC

# Smoking and Parkinson's

## Description

---

- 1 The incidence of Parkinson's in people between 55 and 75 is twice as significant in non-smokers
- 2 Is there a causal relationship between smoking and Parkinson's?
- 3 People are more likely to get Parkinson's the older they get
- 4 Smokers' life expectancy is lower than non-smokers'
- 5 **Non-smokers are more likely to have Parkinson's because they live longer, not because they don't smoke**

# Game Plan

## Description

---

- 1 Causal Impact was developed by Google
- 2 Practice Case Study: Paypal and Bitcoin
- 3 Challenge: Volkswagen CO2 scandal
- 4 Causal Impact is my most-used technique

# Case study:

## Paypal x

## Bitcoin

### Description

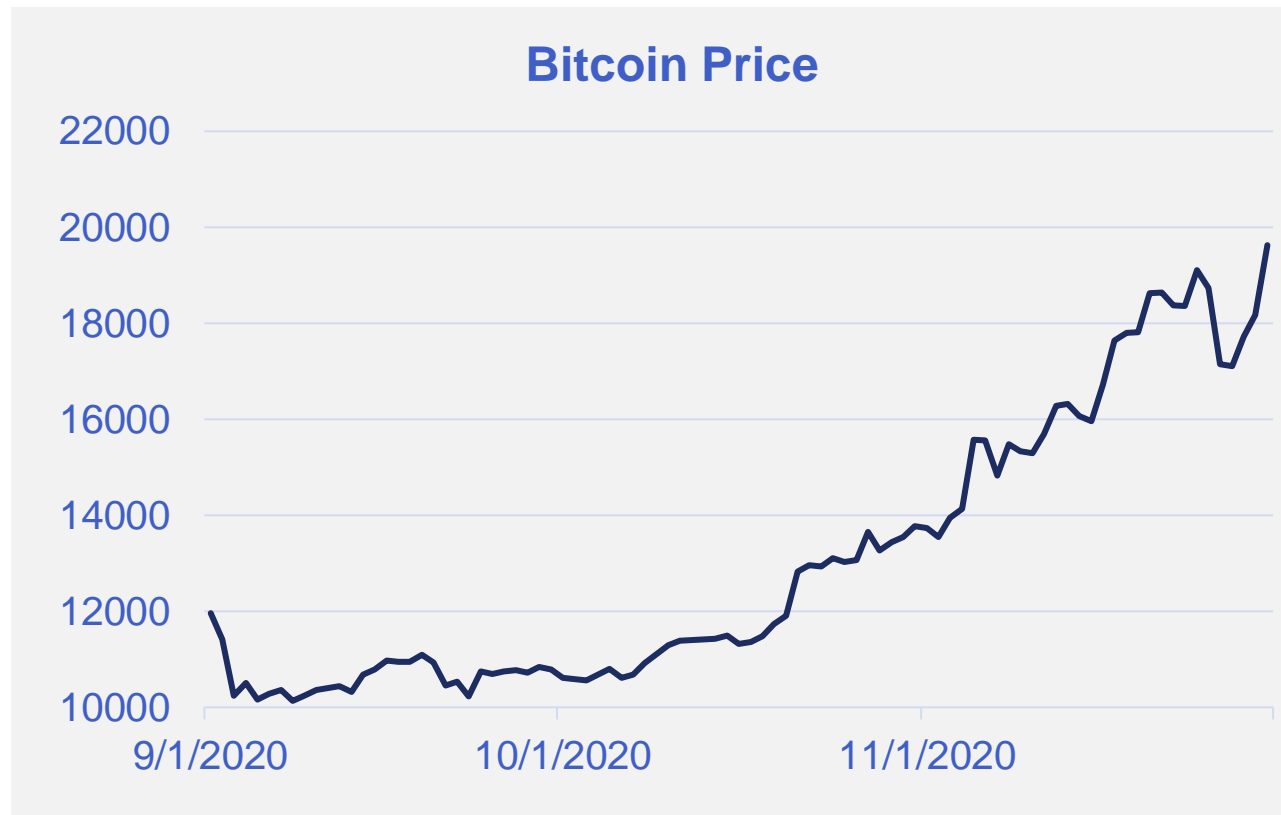
---

**Use Google Causal Impact to estimate the impact of Paypal allowing crypto payments on Bitcoin price**

- 1** In October 21st, 2020, Paypal announced entered the Crypto industry.
- 2** Given the bull market and other volatilities, we cannot compare the price before and after
- 3** We need to find comparable control groups

# What is Time Series Data?

## Visualization



## Key ideas

Sequence of data points in time order (oldest to newest)

Most commonly, it is data recorded in equally distanced time periods

Type of Panel Data (multidimensional dataset)

# Comparing before and after impact leads to omitted variable bias

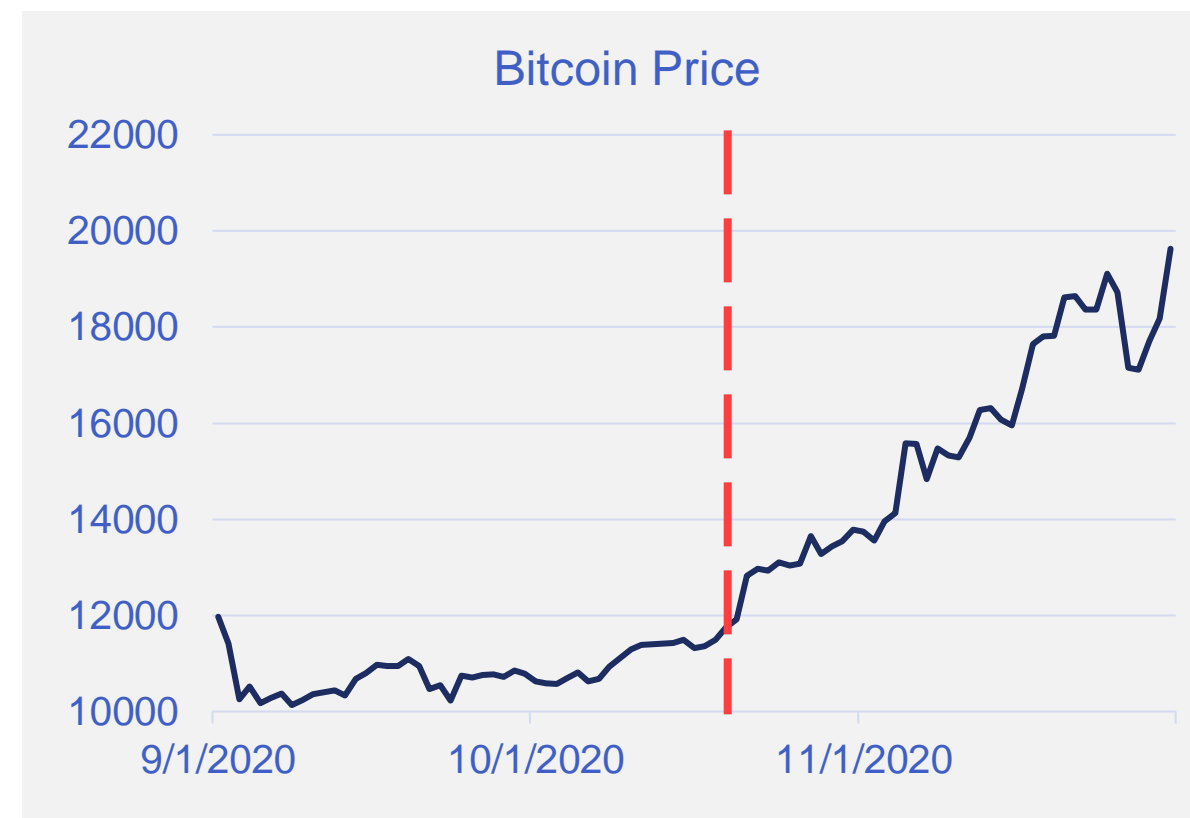
## Context

How to measure the impact of Paypal on Bitcoin price?

This graph shows the sales in the market. The event started where the red line is

Comparing before and after would subject you to omitted bias.

## Visualization



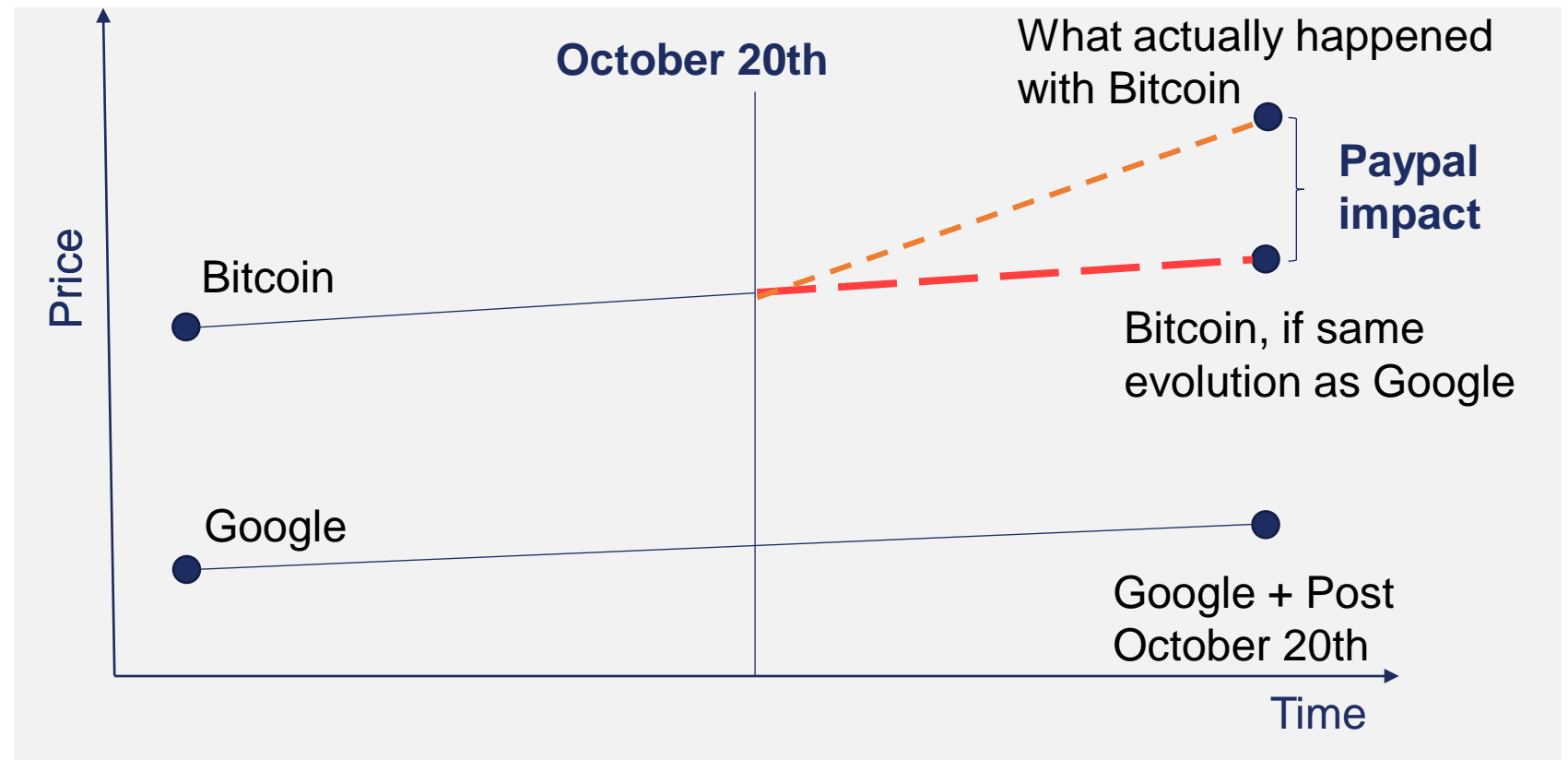
# Difference-in-differences framework

## Key ideas

We use Google to create an artificial control group

The delta between what actually happened and the what-if scenario is the **treatment impact**

## Visualization





# Causal Impact Step by Step

**Define pre and post period**



**Retrieve the data we need**



**Check whether the variables are correlated in the pre period**



**Remove non-correlated data**



**Use Causal Impact**

# Assumptions

## Parallel Trends Assumption

The Treatment and Control Groups are assumed to have the same evolution for the KPI

### Visualization



## Confounding Policy Change

There must be only one policy or initiative that differentiates the treatment from control groups.

You can only measure the impact of one treatment.

## How to Strengthen the assumptions

Use More control groups

Use a longer training period

Keep post-period to the bare minimum

# Correlation in Time Series

## Description

---

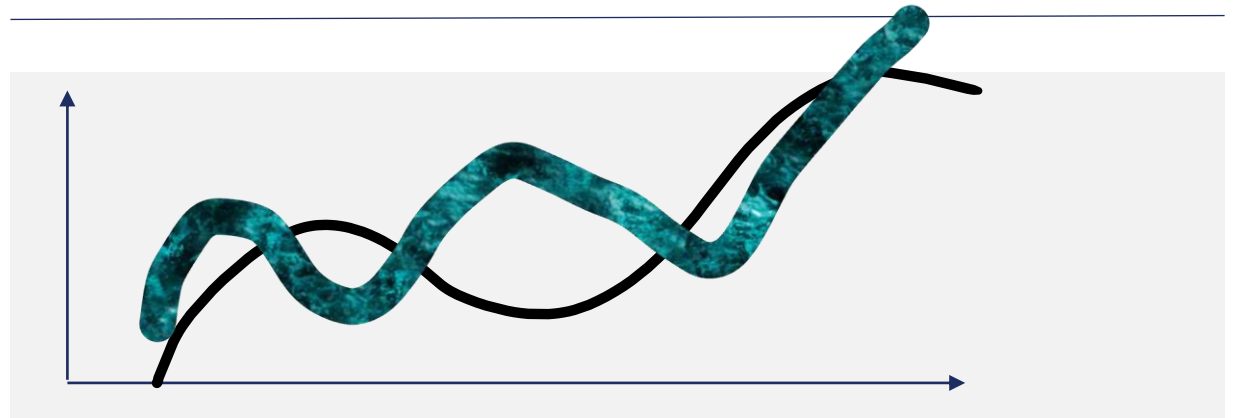
Measures the relationship strength between 2 variables

If the Time-Series grows over time, then the correlation might be random

The data must be stationary

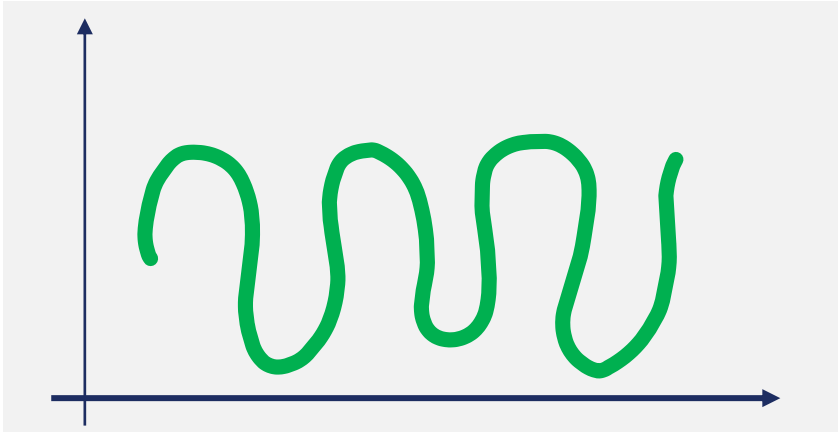
## Visualization

---

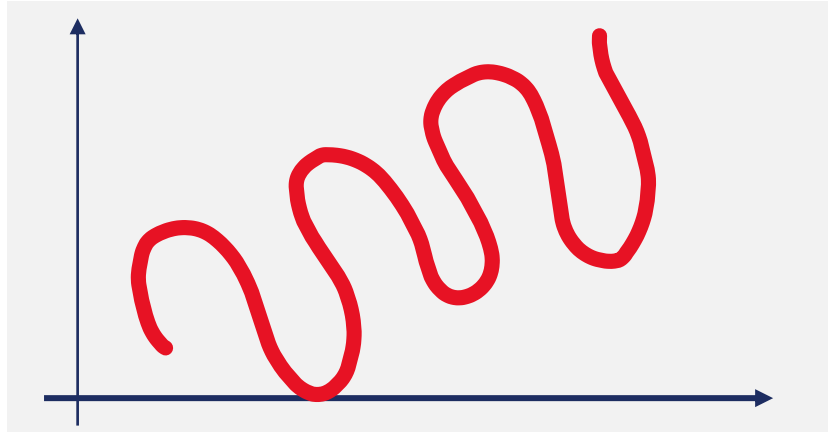


# Stationarity

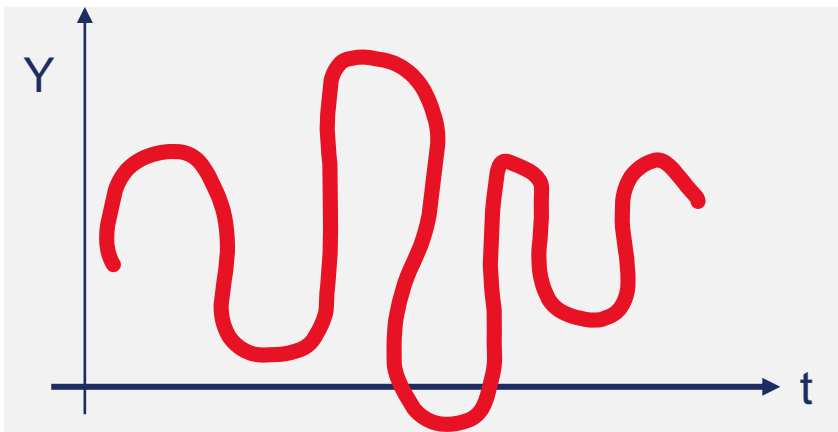
**Stationary Time Series**



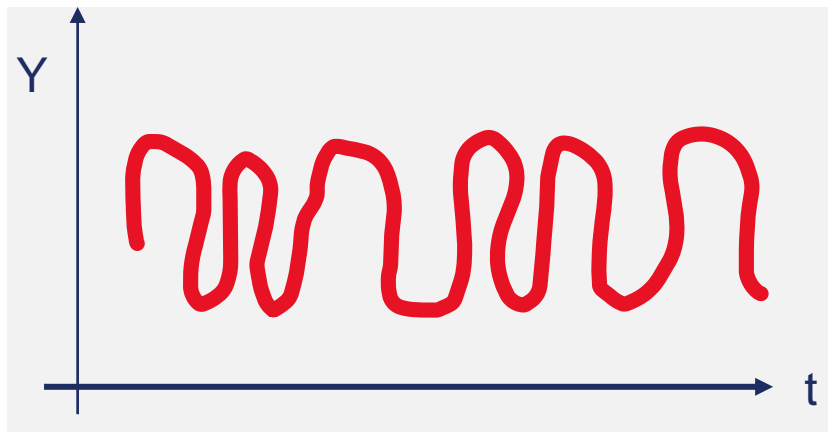
**Time dependent mean**



**Time dependent variance**



**Time dependent covariance**



## Key idea

Mean, variance and covariance are not time dependent

Stationary Time Series have a defined pattern

## Statistical test:

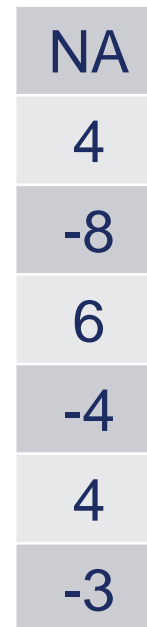
Dickey-Fuller test. If p-value is less than 0.05, time series is considered stationary

# Making Data Stationary

Time Series



Differencing



# Impact evolution

## Context

Let's discuss what should be the impact of Paypal adopting Bitcoin:

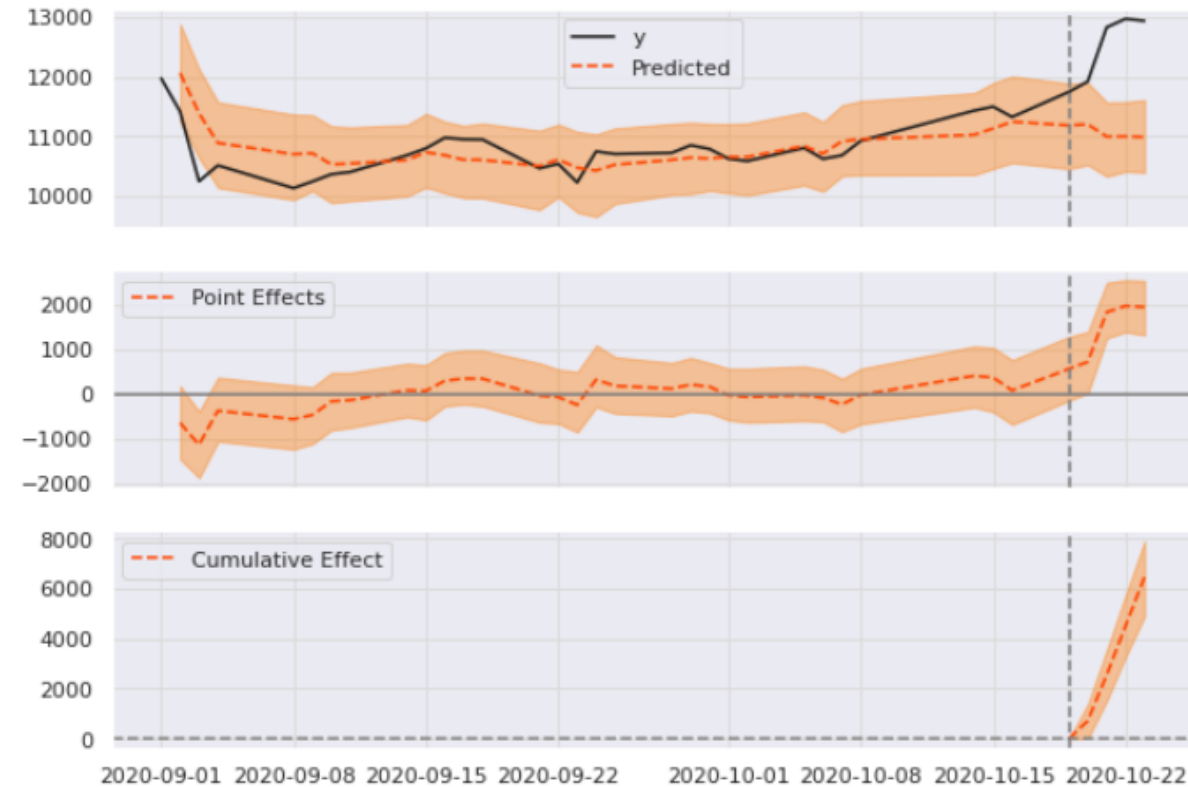
Greater in the beginning

Impact gradually increases

You can also point out that the impact should continue days after the announcement

**Causal Impact allows the impact variations over time**

## Visualization



- Brodersen, Kay H.; Gallusser, Fabian; Koehler, Jim; Remy, Nicolas; Scott, Steven L. Inferring causal impact using Bayesian structural time-series models. Ann. Appl. Stat. 9 (2015), no. 1, 247--274. doi:10.1214/14-AOAS788. <https://projecteuclid.org/euclid.aoas/1430226092>

# Challenge

## Description

---

**Use Causal Impact to measure the impact of the CO2 scandal in Volkswagen stock Price**

- 1 Pick Stocks for the control groups
- 2 Perform a correlation matrix
- 3 Measure the impact

**MATCHING**



# Game Plan

## Description

---

- 1 There is no comparable control group
- 2 Helps us with (self)-selection bias
- 3 How to measure referral programs?
- 4 What is the incremental value of Mobile Shopping?
- 5 Practice case study: Catholic Schools and scores
- 6 Challenge: Remote work and career satisfaction

# How do you figure out the value of Amazon Prime?

## Context

---

Amazon Prime is a loyalty program that provides free shipping, discounts and other services

The goal of program is fourfold:

- Increase customer loyalty
- Increase revenue per customer
- Decrease marketing spendings in customer re-activation
- Decrease paid advertising in conversion

The subscription lasts 1 year

If you were to asked to provide the impact of Amazon Prime on its financials, how would you do it?

# You cannot just simply compare the average Prime and non-prime subscriber

## Context

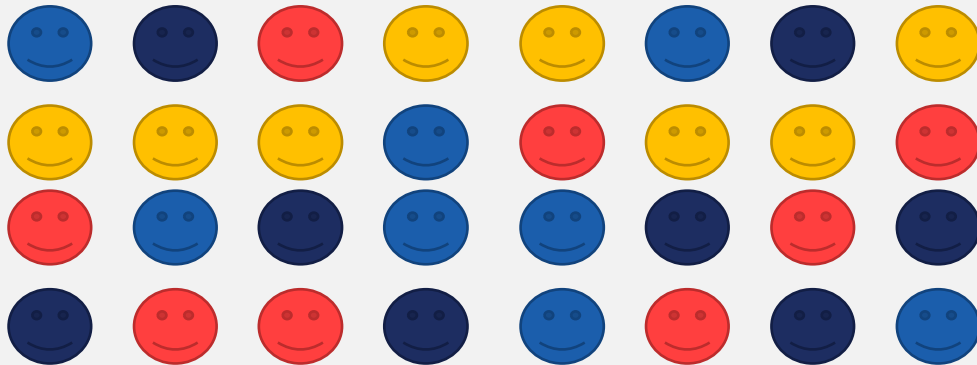
Both groups may be inherently different from the start. Hence, they are not comparable.

Beware of (self-)selection bias

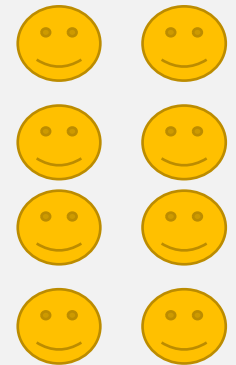
A possible solution is Matching.

In a nutshell, you create a counterfactual group with similar characteristics to your treatment group

Control



Treatment



# You cannot just simply compare the average Prime and non-prime subscriber

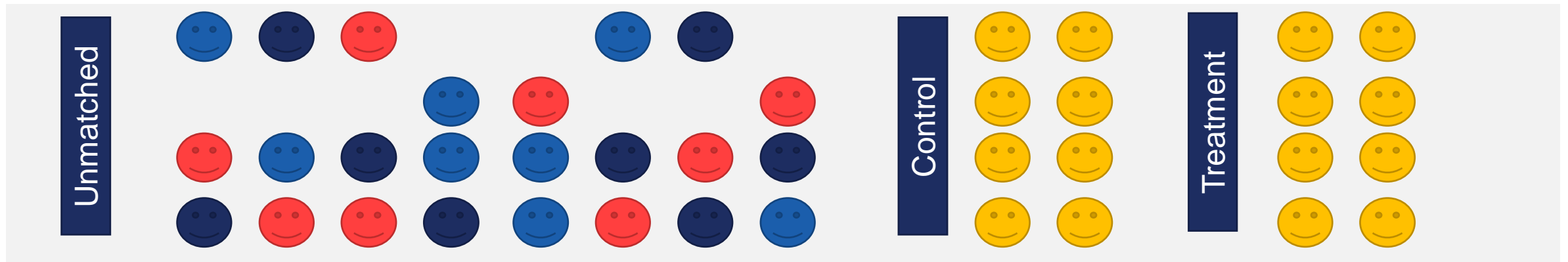
## Context

Both groups may be inherently different from the start. Hence, they are not comparable.

Beware of (self-)selection bias

A possible solution is Matching.

In a nutshell, you create a counterfactual group with similar characteristics to your treatment group



# **Case Study**

## **Briefing – Are**

### **catholic**

### **schools**

### **better?**

#### **Description**

---

**Use Matching to understand whether catholic schools are better than others (from a standardized test score view)**

- 1** We have a dataset with kids' background, their parents upbringing among others
- 2** The key metric of success is the standardized test scores
- 3** We need to re-create a comparable control group

# Unconfoundedness

## Context

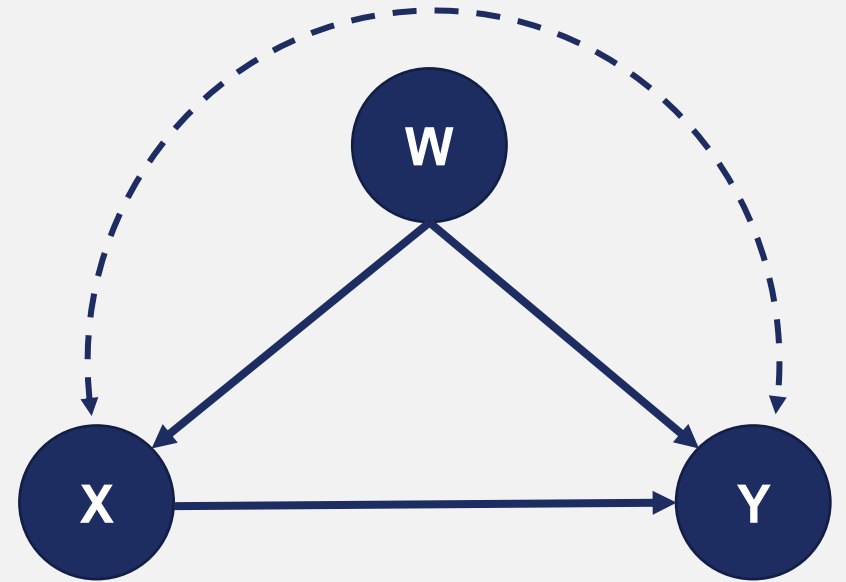
The variables (confounders) used are enough to describe the people or entities (W)

The characteristics affect the likelihood of someone being part of the treatment (X)

The combination of the confounders and the treatment leads to the outcome (Y)

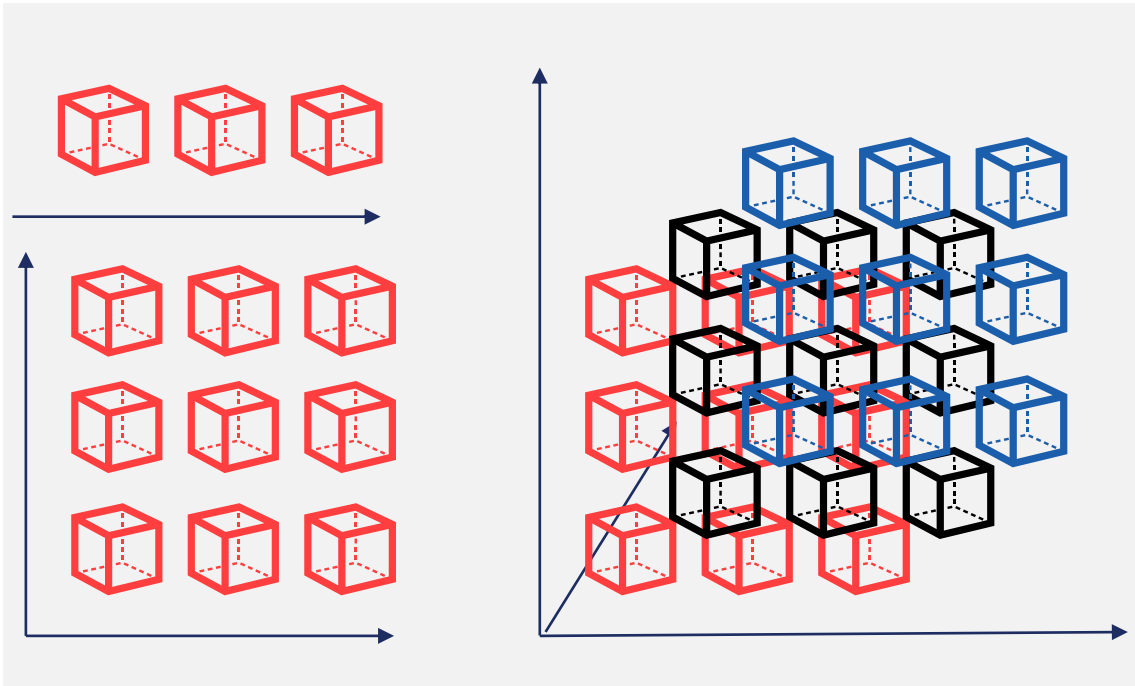
Meeting the Unconfoundedness assumption is a tall order

## Visualization



# Curse of Dimensionality

## Visualization



## Context

Imagine you have a variable with 3 options

Then you had a second with 3 more

Finally, a third

The observations needed to fill each bucket grows exponentially

The Matching outcome can be spurious, when few elements belong to a “dimension”

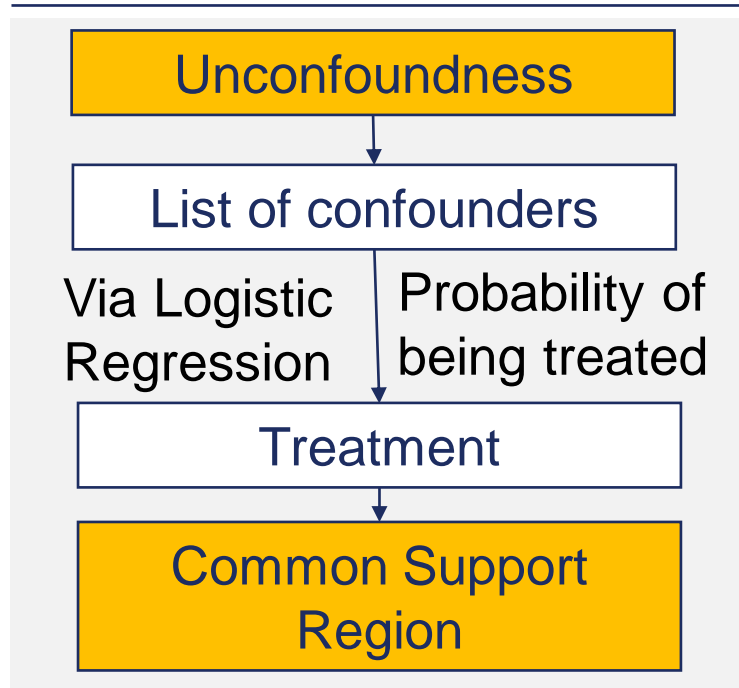


### Key Idea

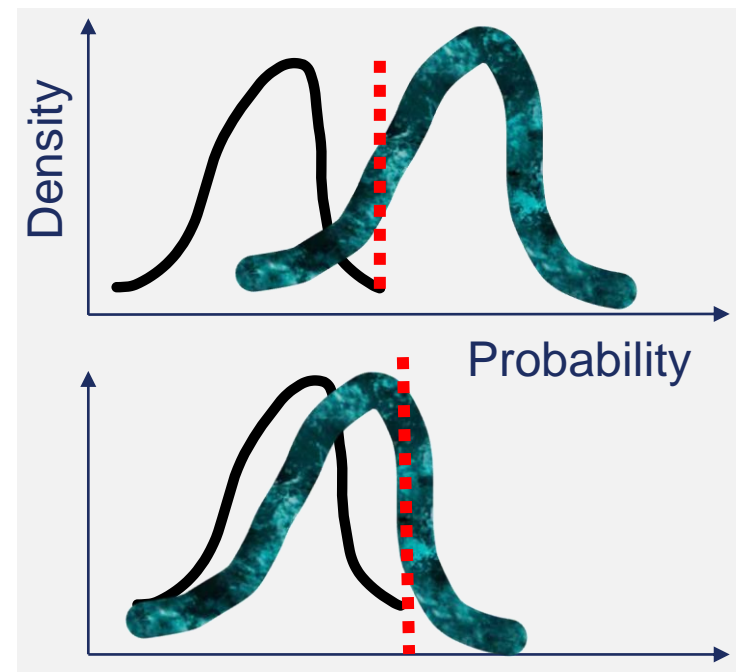
Make sure when you create a model as simple as possible

# How to determine the Common Support Region

## Visualization



## Examples



## Key ideas

We predict whether someone is part of the treatment group

There will be people with high likelihood of participating.

You are not likely to find a control group for them.

The greater the overlap, the higher the matching quality



Probability of the treated group being treated

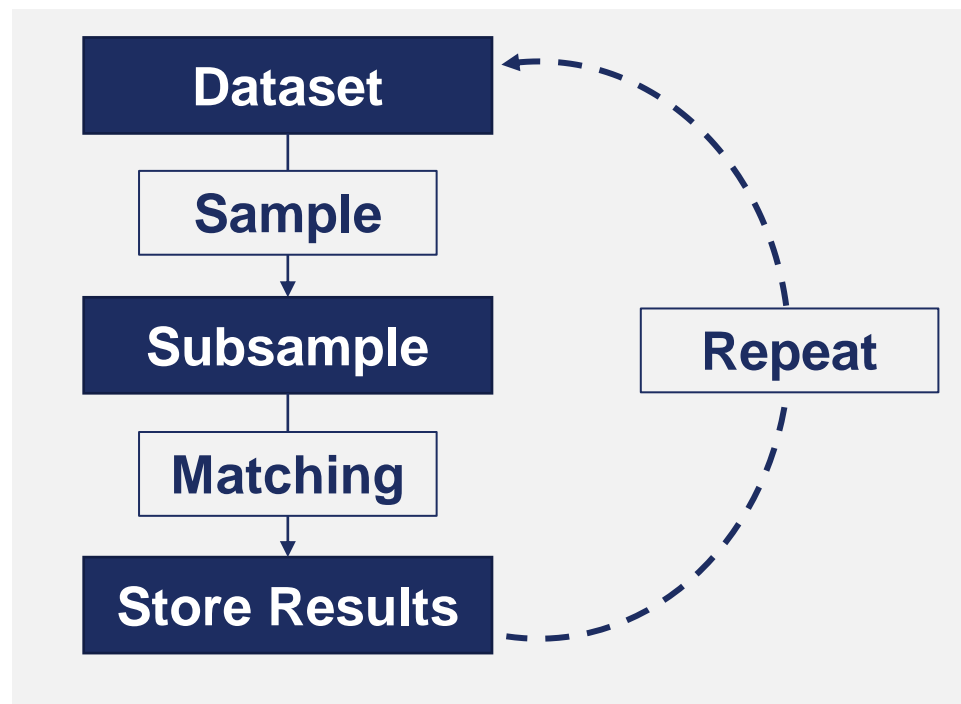


Probability of the non-treated group being treated

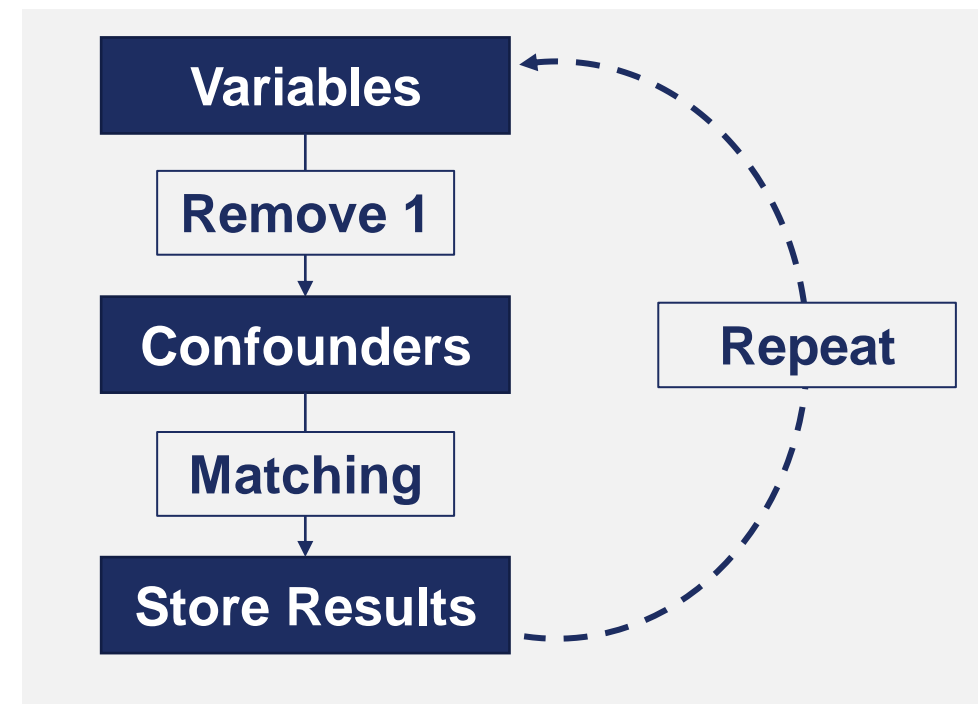


# Robustness checks

## Repeated experiment



## Removing 1 confounders



# Challenge

## Description

---

**Use Matching to figure out whether Remote workers have higher Career Satisfaction**

- 1 Pick variables for your model
- 2 T-test Loop
- 3 Transform the character variables into binary
- 4 Perform Matching
- 5 Perform a robustness check

# My experience with Matching

## Description

---

- 1 Introducing English in the Zalando.de website
- 2 What is the incrementality?
- 3 Difference audiences means non-comparability
- 4 Tiny treatment group = good common support region
- 5 Practice case study: Catholic Schools and scores
- 6 I used the repeated experiments for robustness

**RECENCY  
FREQUENCY  
MONETARY**

# Game Plan

## Description

---

- 1 Introducing value-based segmentation
- 2 Case study: online shoppers segmentation
- 3 Challenge: purchasing behavior
- 4 Simple yet powerful concepts in this section

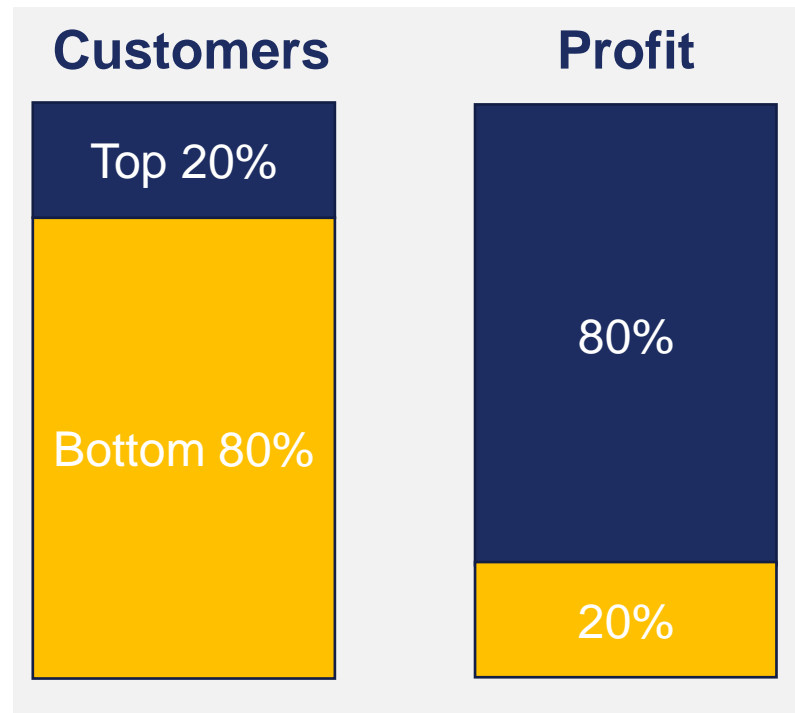
# Value-based segmentation

## Pareto rule

20% of the causes  
result in 80% of the  
consequences



## Visualization



## Description

Companies rank customers

Who to prioritize

Understand where to focus

Who is more loyal

# What is RFM?

Description	Typical RFM	Our Model	Meaning
Recency	Days gone	Days gone	How long since they last purchased
Frequency	Frequency (Q)	Frequency (Q)	How often have they purchased
Moneretary	Sales ( $P \cdot Q$ )	Basket (P)	Average Purchase Value

## What else?

Include Churn or Customer Retention

Include Time Horizon

Change Average Purchase by Average Profit

# How does it work?

## Frequency

Max	↑	4
		3
Median		2
		1
Min		

## Recency

Max	↓	1
		2
Median		3
		4
Min		

## Monetary

Max	↑	4
		3
Median		2
		1
Min		

## Final Values

11-12	Superstar
8-10	Future Champion
6-7	High Potential
3-5	Low Relevance



# Case Study

## Briefing

### Description

---

#### A Dataset with Online Shoppers data

- 1 We have a dataset with purchases of customers, detailed by items
- 2 Create a customer dataset with the Recency, Frequency and Monetary variables
- 3 Create an RFM model and apply to the dataset

# Challenge

## Customer

## Value

## segmentation

### Description

---

**A dataset with customer data**

- 1 Prepare basket variable
- 2 Rename variables
- 3 Create a RFM model with 3 levels
- 4 Define 3 segments
- 5 Prepare final table overview

# **GAUSSIAN MIXTURE MODEL**

# Game Plan

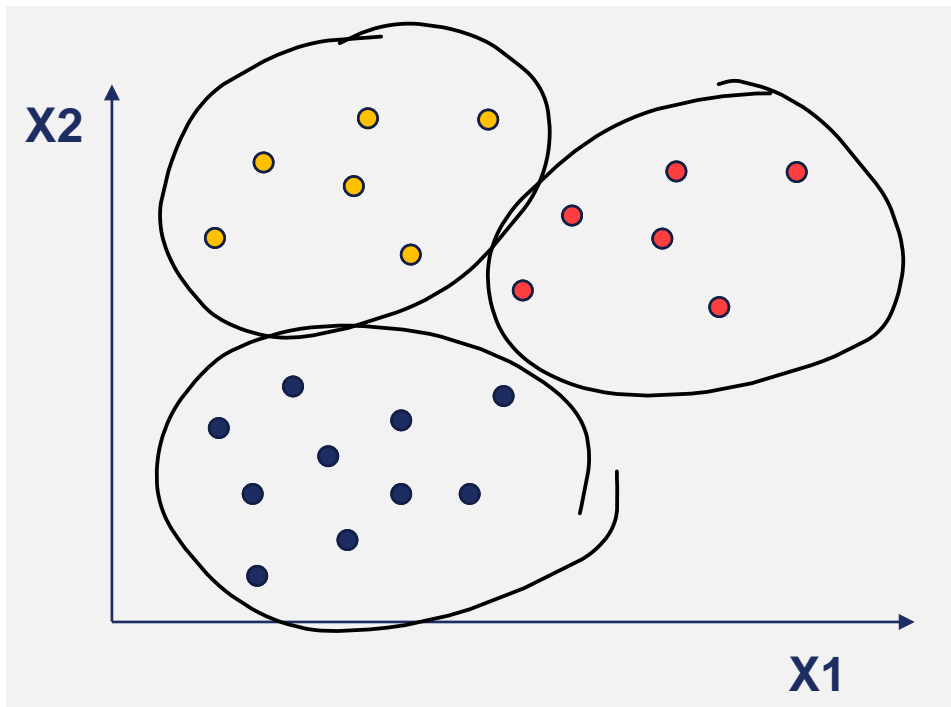
## Description

---

- 1 Clustering is a lazy person's favorite
- 2 Case study: Credit Card applicants
- 3 Challenge: Credit card users
- 4 New concepts: AIC and BIC

# What are clustering techniques?

## Visualization



## Key ideas

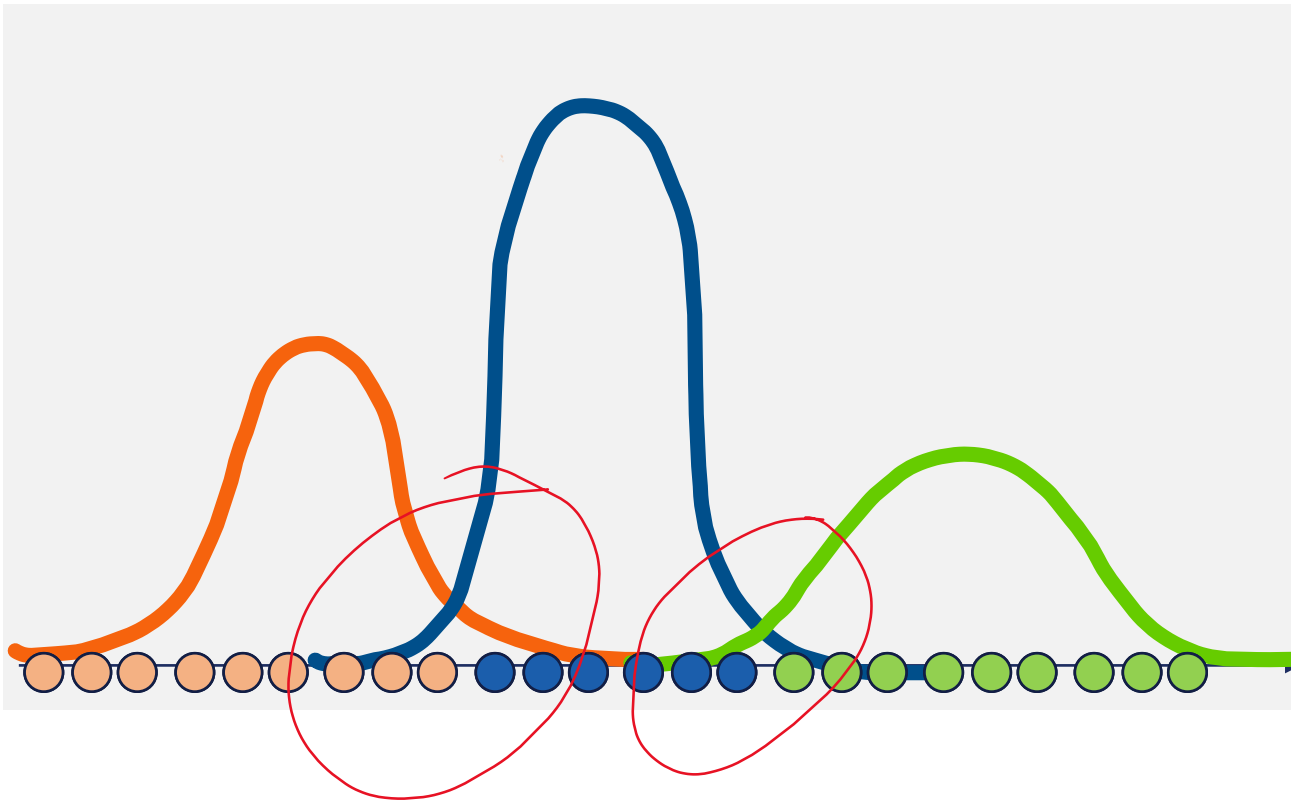
Groups observations in terms of their characteristics

Main task of exploratory data mining

Clustering is an art rather than Science

# Gaussian Mixture Model

## Visualization



## Key ideas

Gaussian Mixture Model is a probabilistic method for clustering

Better to use than traditional clustering algorithms, like Kmeans

The probabilities allow to better evaluate edge cases

# Case Study

## Briefing

### Description

---

#### A Dataset with Credit Card applicants

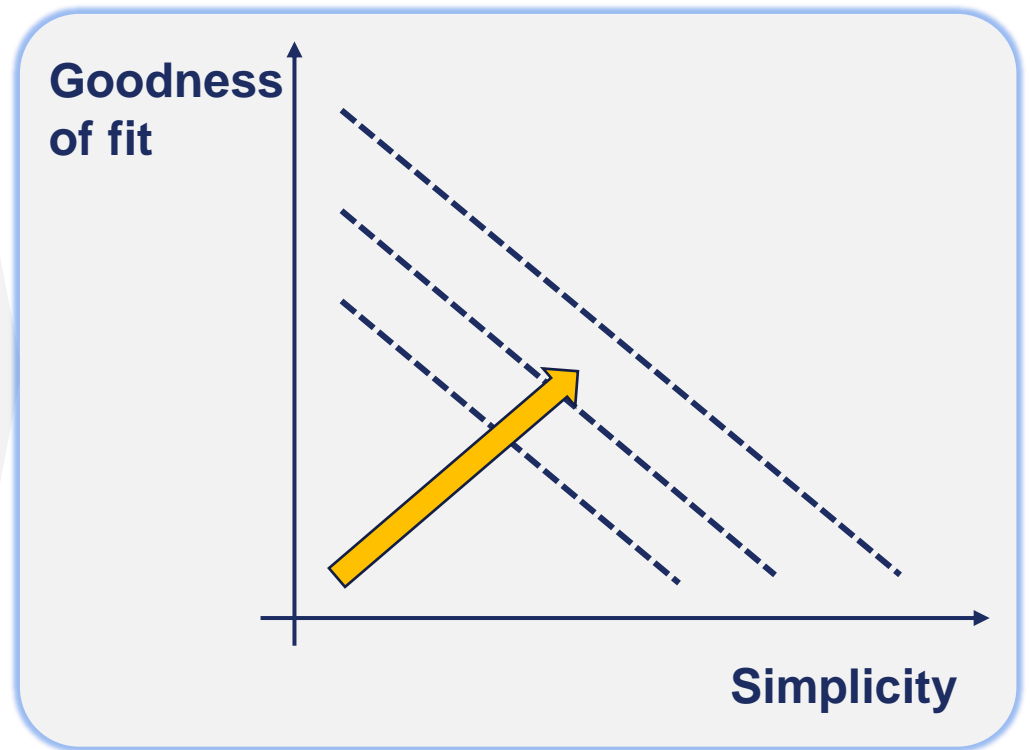
- 1 Determine the optimal number of segments for the dataset
- 2 Use Gaussian Mixture Model
- 3 Interpret the segments and name them according to their characteristics

# Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC)

## Key Ideas

- AIC and BIC helps us determining the optimal number of clusters
- AIC and BIC provide a means to select a model
- Trade-off between simplicity and goodness of fit
- Deal with overfitting
- BIC penalizes overfitting more than the AIC

## Pseudo-visualization





# Challenge – Gaussian Mixture Model

## Description

---

A Dataset with customer data

- 1 Prepare data set
- 2 Determine optimal number of clusters
- 3 Create GMM model
- 4 Interpret segments

# My experience with Segmentation

## Description

---

- 1 A closed contest for a big conglomerate
- 2 Their status-quo was a value-based segmentation
- 3 They wanted a behavioral segmentation
- 4 The first difficulty was how massive the data was
- 5 We tried to be hypothesis-driven
- 6 We had 7 interpretable segments in the end
- 7 We were complex and did not consider scalability

# **RANDOM FOREST**

# Game Plan

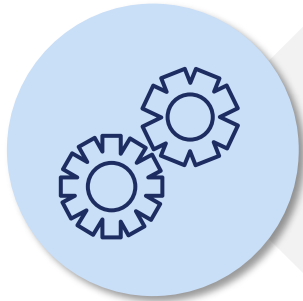
## Description

---

- 1 Random Forest is an advanced analytics technique
- 2 Learn about Decision Trees and Ensemble Learning
- 3 Practice case study: Credit card applicants
- 4 Challenge: Customer's income prediction

# Random Forest is an Ensemble Learning Algorithm

## What is it?

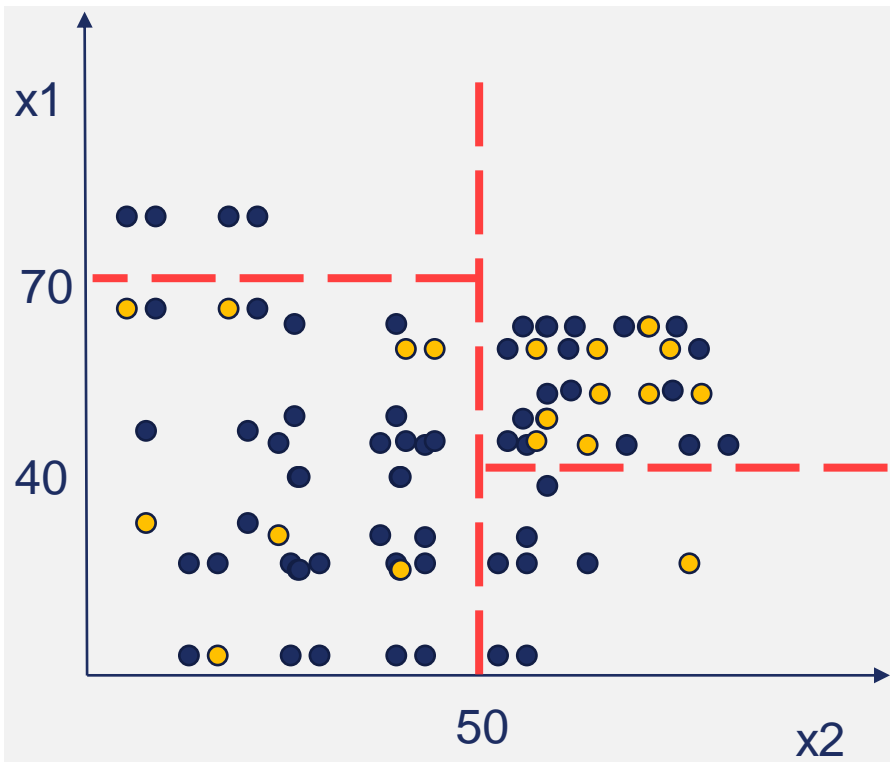


## Description

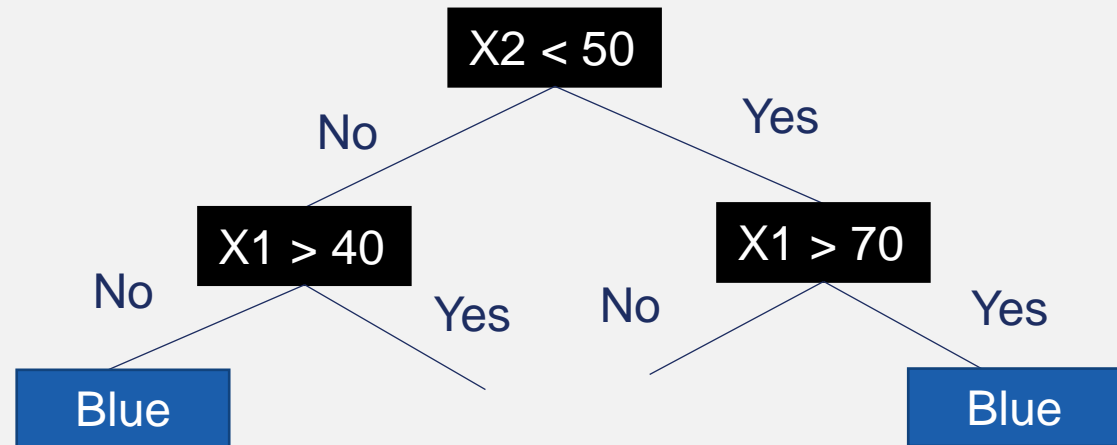
- 1 Ensemble Learning is when you have a plurality of models predicting your output
- 2 Ensemble is an average of Models
- 3 A Random Forest is a combination of decision trees
- 4 Can be used for Regression and Classification problems

# How do Decision trees work?

## Visualization



## Decision tree



### Key Ideas:

- A split or leaf is done taken a maximum entropy logic
  - Where would it yield more information
- The prediction would be done based on the relative frequency

# Case Study

## Briefing

### Description

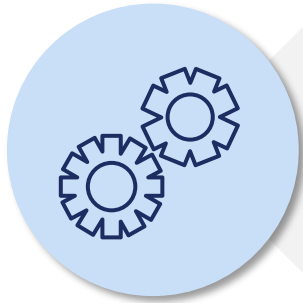
---

#### A Dataset with credit card applicants

- 1 The key metric of success is whether someone was accepted or not
- 2 We want to predict the acceptance
- 3 We also want to generate insights

# Random Forest quirks

## What is it?



## Description

- 1 Tendency to overfit
- 2 Good with multicollinearity
- 3 Works well with non-linearity
- 4 Robust to Outliers

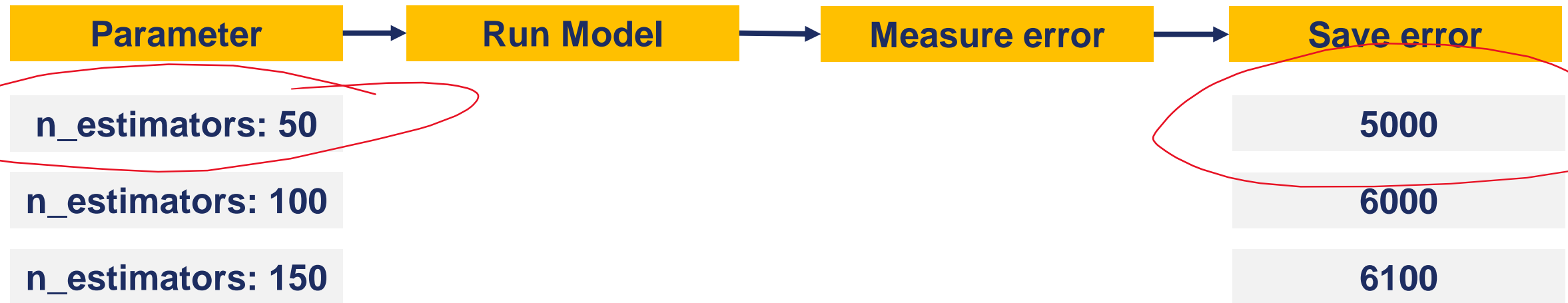


# Parameter Tuning

## Context

Advanced models have parameters to tune to optimize accuracy

## Description



# Challenge – Random Forest

## Description

---

What is the income of your customers?

- 1 Prepare data set
- 2 Create Random Forest Regressor model
- 3 Measure accuracy
- 4 Tune the model
- 5 Generate insights

# **FACEBOOK PROPHET**

# Game Plan

## Description

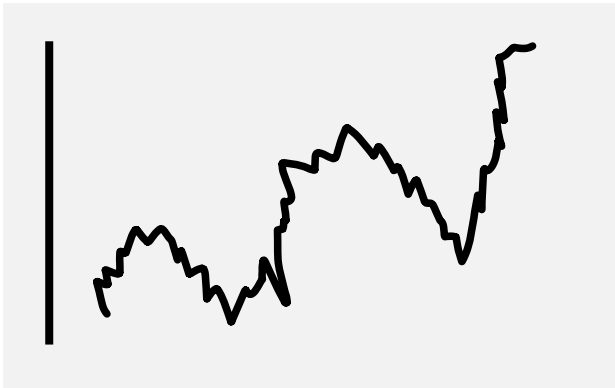
---

- 1 Technique to predict the future
- 2 Forecasting is a common task for Business analysts
- 3 Practice case study: Udemy Wikipedia page visits
- 4 Challenge: Shelter Demand in New York City

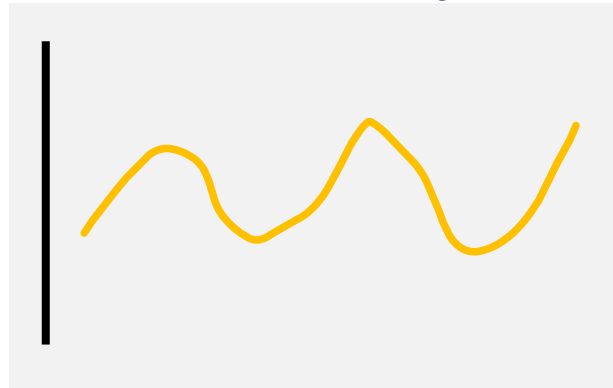
# Structural Time Series

## Visualization

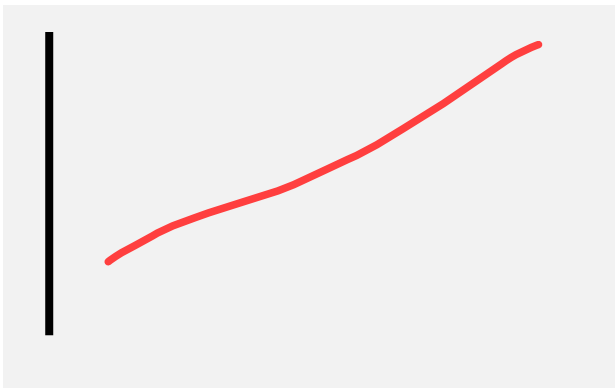
Data



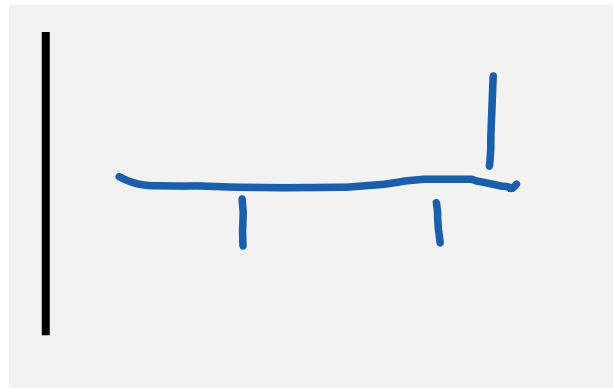
Seasonality



Trend



Exogenous impacts



## Description

Structural Time Series is the decomposition of the data in at least:

Trend

Seasonality

Exogenous impacts

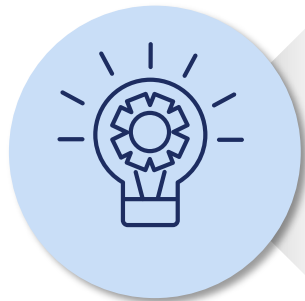
Error Term

## Methodological framework

$$y(t) = c(t) + s(t) + x(t) + \epsilon$$

# Facebook Prophet quick facts

Which?



## Description

- 1 Built by facebook
- 2 Stan background - probabilistic programming language for statistical inference
- 3 Dynamic Holidays
- 4 Prophet is customizable in ways that are intuitive to non-experts
- 5 Built-in Cross Validation

# Prophet Mechanics

## Methodological framework

---

$$y(t) = c(t) + s(t) + h(t) + x(t) + \epsilon$$

Where:

$c(t)$	Trend +
$s(t)$	Seasonality +
$h(t)$	Holiday effects +
$x(t)$	External regressors +
$e$	error

# Case Study

## Briefing

### Description

---

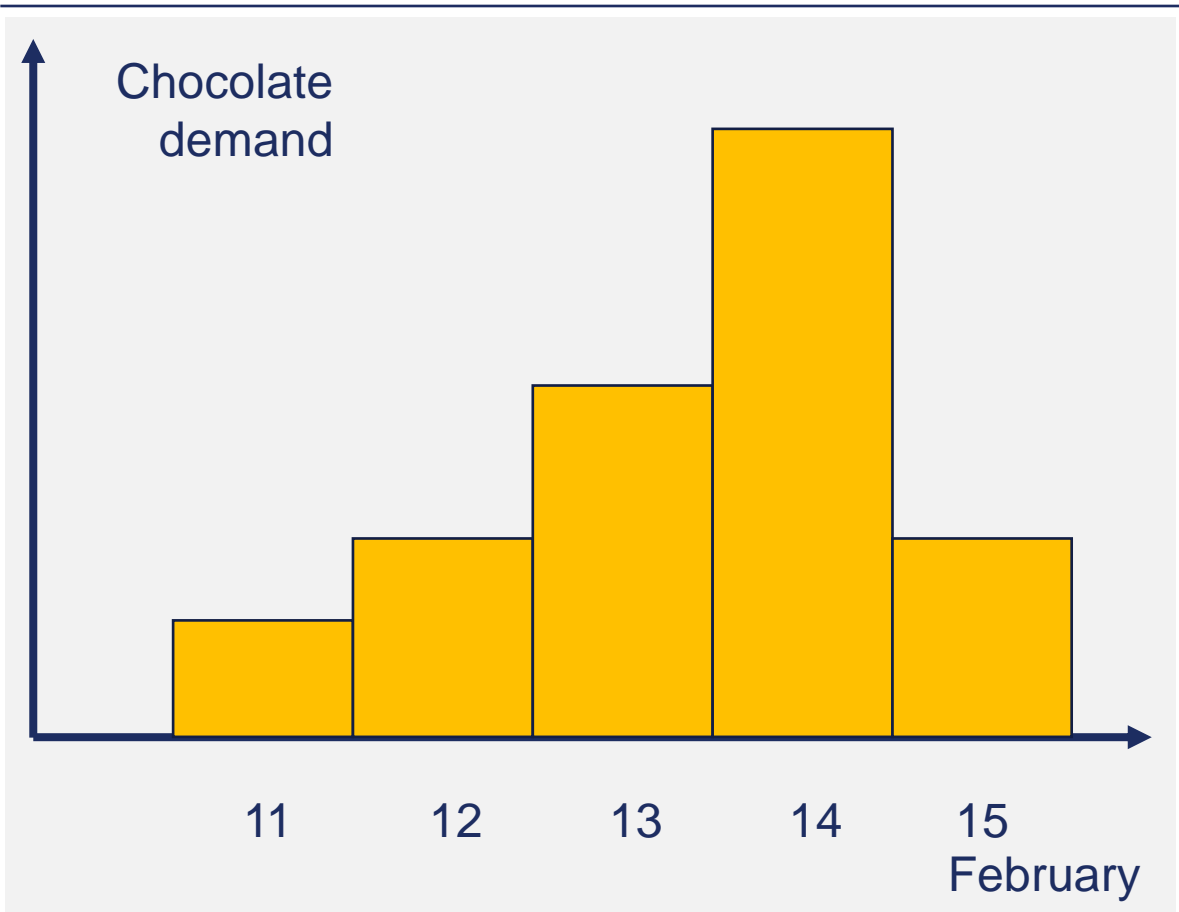
#### A Dataset with Daily Udemy Wikipedia Visits

- 1 Predict the number of visits to the Wikipedia page of Udemy
- 2 Learn cross-validation
- 3 Combine with Parameter Tuning



# Dynamic Holidays – Valentine's example

## Visualization



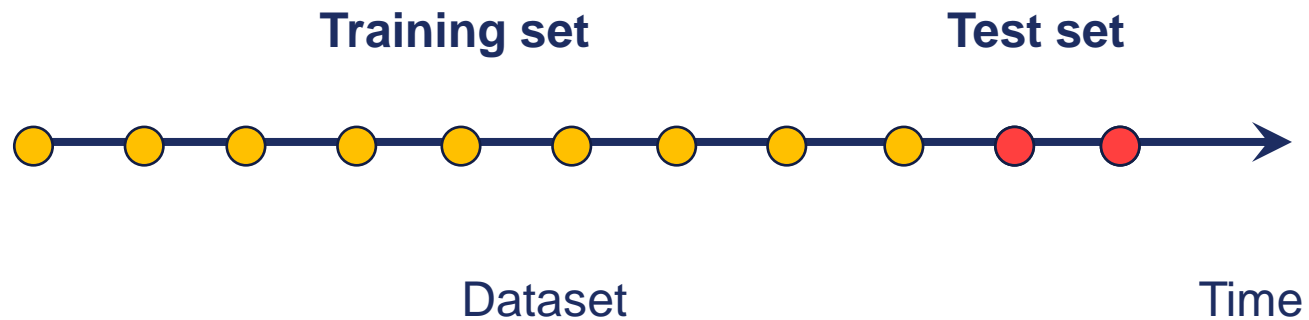
## Facebook Prophet

You state Valentine's as a key event and specify how many days before/after

## Other models:

You must create dummy variables for each day, if you believe they have different impacts

# Training and Test Set in Time Series



## Key Ideas

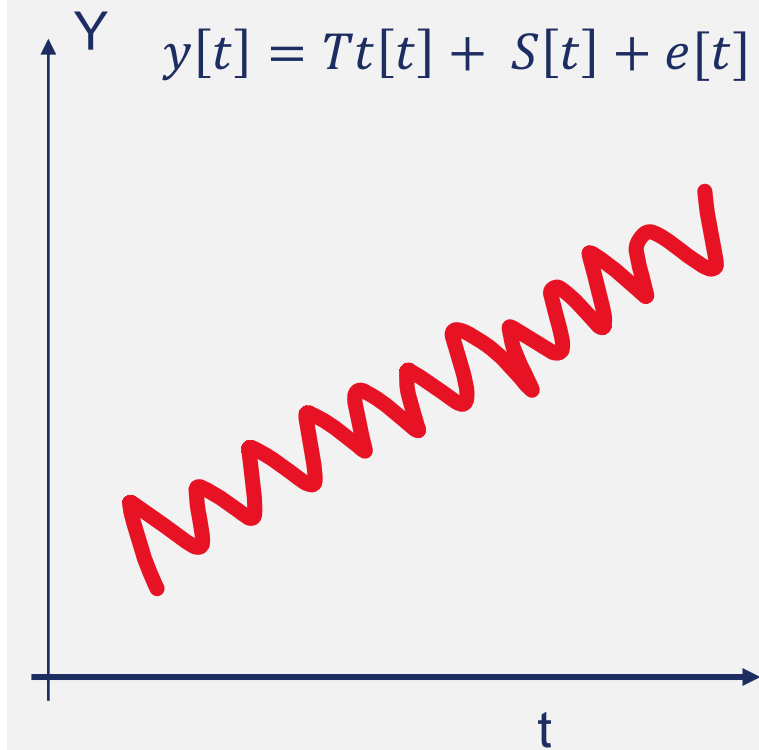
Forecasting Models are usually split into a pre and post period from a time perspective  
The Test Set should be of the size of a real-world forecast

# Facebook Prophet Model

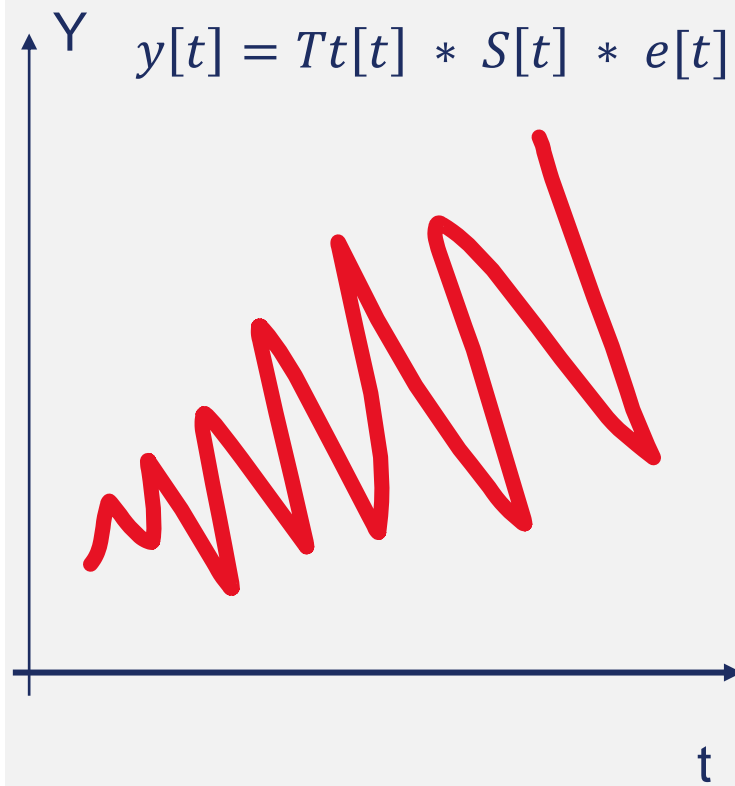
Component	Description
Growth	Linear or Logistic
Holidays	Dataframe that we prepared
Seasonality	Yearly, weekly or daily. True or False
Seasonality_mode	Multiplicative or additive
Seasonality_prior_scale	Strength of the seasonality
Holiday_prior_scale	Smaller values allow the model to fit larger seasonal fluctuations
Changepoint_prior_scale	Does the Trend change easily?

# Additive vs. Multiplicative

## Additive



## Multiplicative

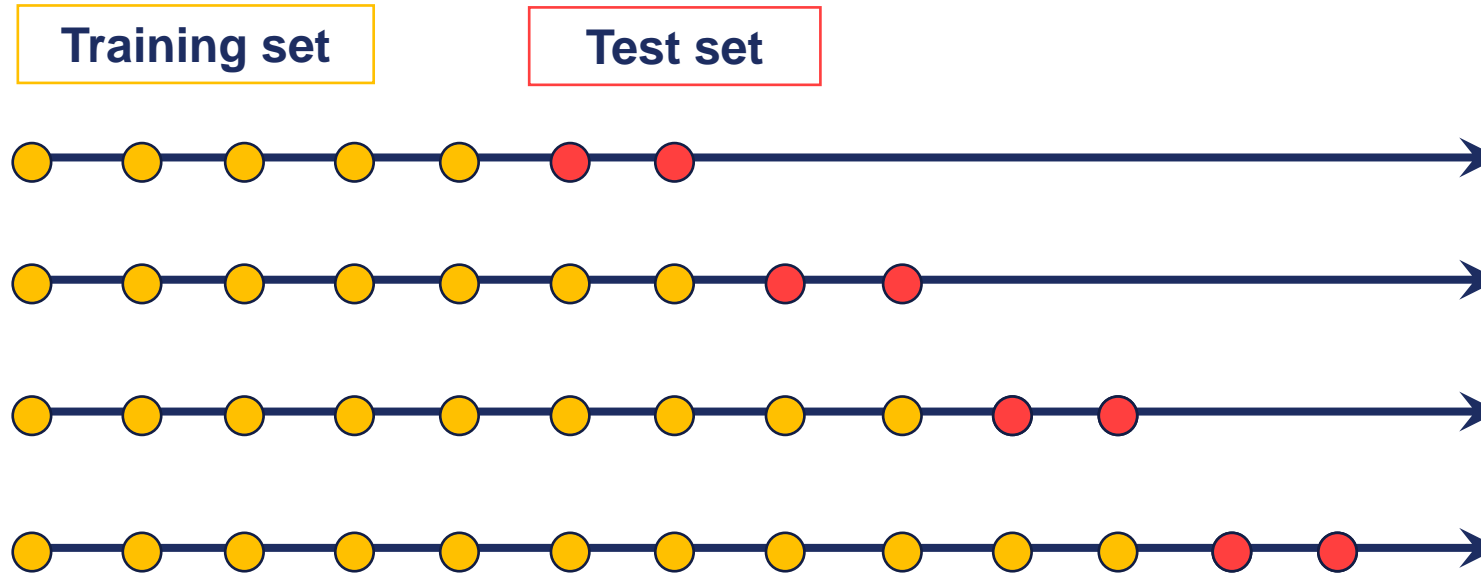


## Key ideas

If we talk about seasonality in terms of percentage, the seasonality is multiplicative

If it is in adding absolute values, then it is additive.

# Cross Validation



## Key Idea

Repeating the assessment of our model reinforces its evaluation

# Parameters to tune

Component	Description
Seasonality_prior_scale	Strength of the seasonality
Holiday_prior_scale	Smaller values allow the model to fit larger seasonal fluctuations
Changepoint_prior_scale	flexibility of the automatic changepoint selection

# Challenge - Demand Forecasting

## Shelter Demand

---

How many people need a shelter?

- 1 Prepare dataframe
- 2 Training and test set
- 3 Create model and assess accuracy
- 4 Visualize the output
- 5 Perform parameter tuning

# Forecasting at Uber

## Description

---

- 1 They need strategic and tactical forecasts
- 2 They need marketplace forecasts to allocate cars
- 3 But they also need it to measure investments
- 4 They need to forecast at scale
- 5 They use simple statistical models
- 6 Machine Learning when exogenous regressors are available
- 7 They try multiple approaches to find the best result