

# **Podstawy uczenia maszynowego**

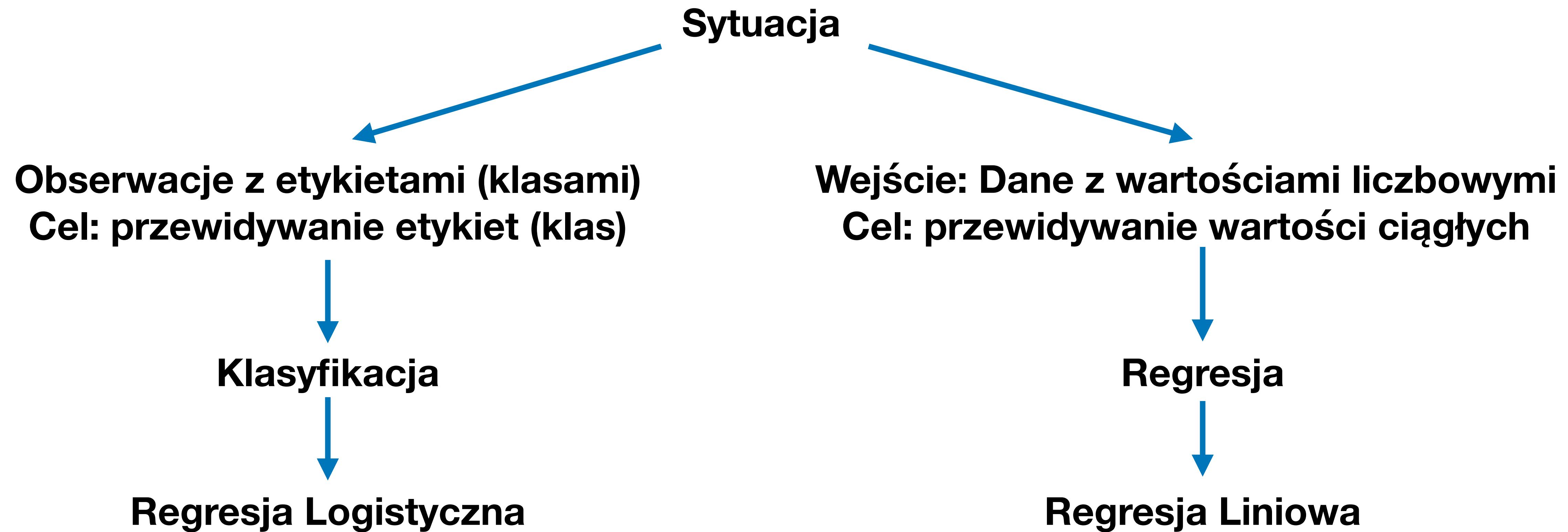
## **Część 3**

**Dr Mateusz Radzimski**

# **Regresja wielowymiarowa**

(Multivariate linear regression)

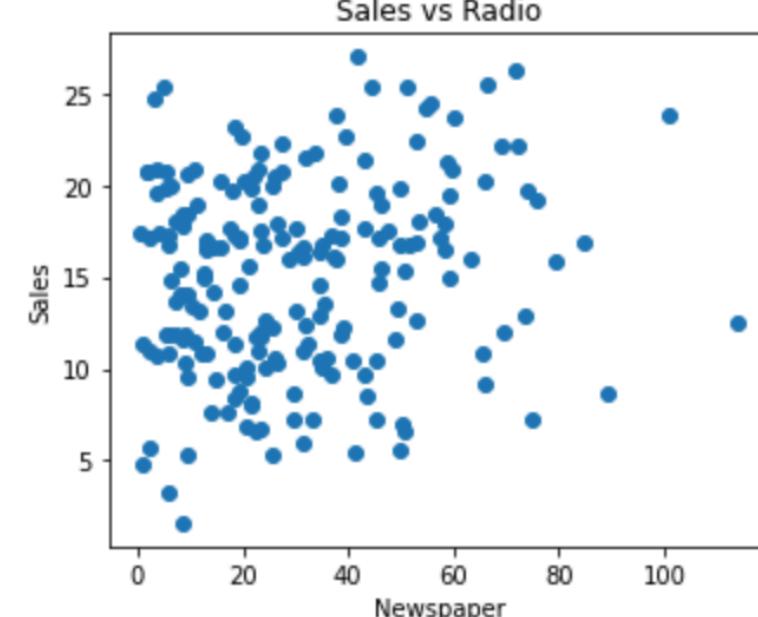
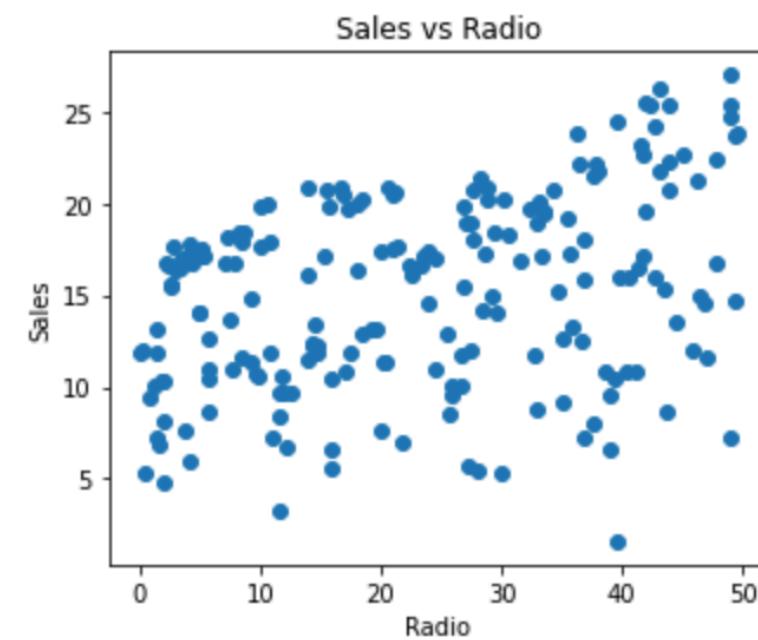
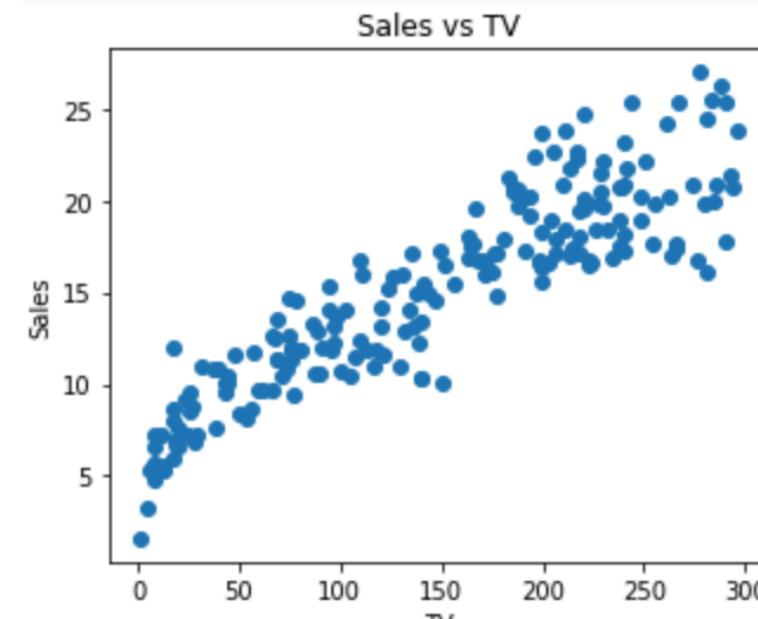
# Przypomnienie dotychczasowych metod



Tabular data					
Features & Label Samples	F1	F2	F3	F4	Label
Sample_1	v <sub>1,1</sub>	v <sub>1,2</sub>	v <sub>1,3</sub>	v <sub>1,4</sub>	L <sub>1</sub>
Sample_2	v <sub>2,1</sub>	v <sub>2,2</sub>	v <sub>2,3</sub>	v <sub>2,4</sub>	L <sub>2</sub>
Sample_3	v <sub>3,1</sub>	v <sub>3,2</sub>	v <sub>3,3</sub>	v <sub>3,4</sub>	L <sub>3</sub>
.	.	.	.	.	.
.	.	.	.	.	.
Sample_n	v <sub>n,1</sub>	v <sub>n,2</sub>	v <sub>n,3</sub>	v <sub>n,4</sub>	L <sub>n</sub>

Homework Grade (x)	Test Grade (y)
94	98
95	94
92	95
87	89
82	85
80	78
75	73
65	67
50	45
20	40

# Przypomnienie dotychczasowych metod



	TV	Radio	Newspaper	Sales
0	230.1	37.8	69.2	22.1
1	44.5	39.3	45.1	10.4
2	17.2	45.9	69.3	12.0
3	151.5	41.3	58.5	16.5
4	180.8	10.8	58.4	17.9
...	...	...	...	...
195	38.2	3.7	13.8	7.6
196	94.2	4.9	8.1	14.0
197	177.0	9.3	6.4	14.8
198	283.6	42.0	66.2	25.5
199	232.1	8.6	8.7	18.4

200 rows × 4 columns

W przykładzie ze zbiorem danych „Advertisement” tworzyliśmy model regresji liniowej na podstawie jednej zmiennej niezależnej: np.  $\text{Sales} \sim \text{TV}$ .

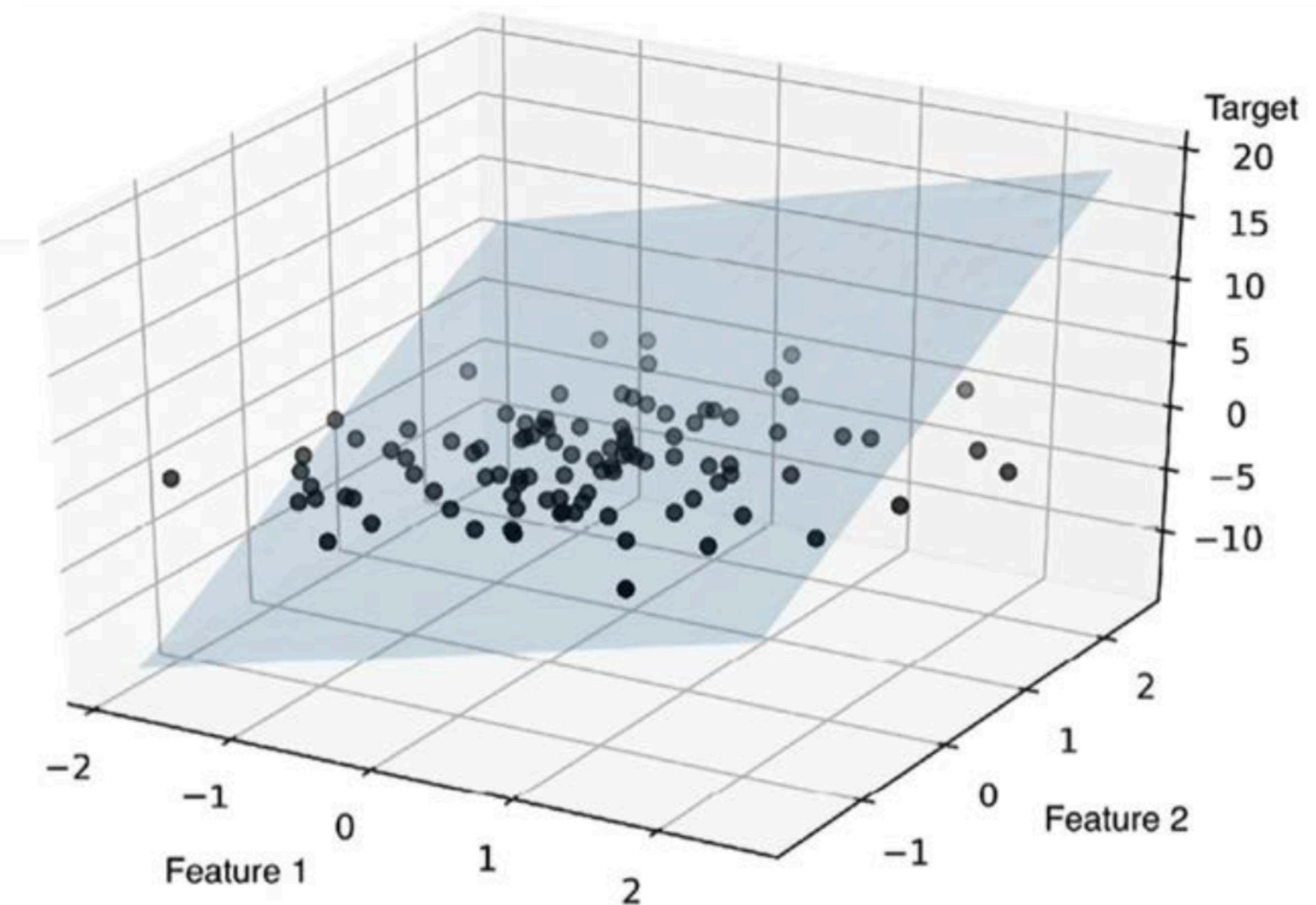
Jest to tzw. prosta regresja liniowa (simple linear regression).

Co zrobić w przypadku gdy mamy więcej zmiennych niezależnych?

# Regresja wielowymiarowa

- Regresję możemy stosować również w przypadku, gdy mamy do czynienia z wieloma zmiennymi niezależnymi

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	Price
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33	36.2



# Regresja wielowymiarowa - model

- Model wielowymiarowej regresji liniowej przyjmuje postać:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon$$

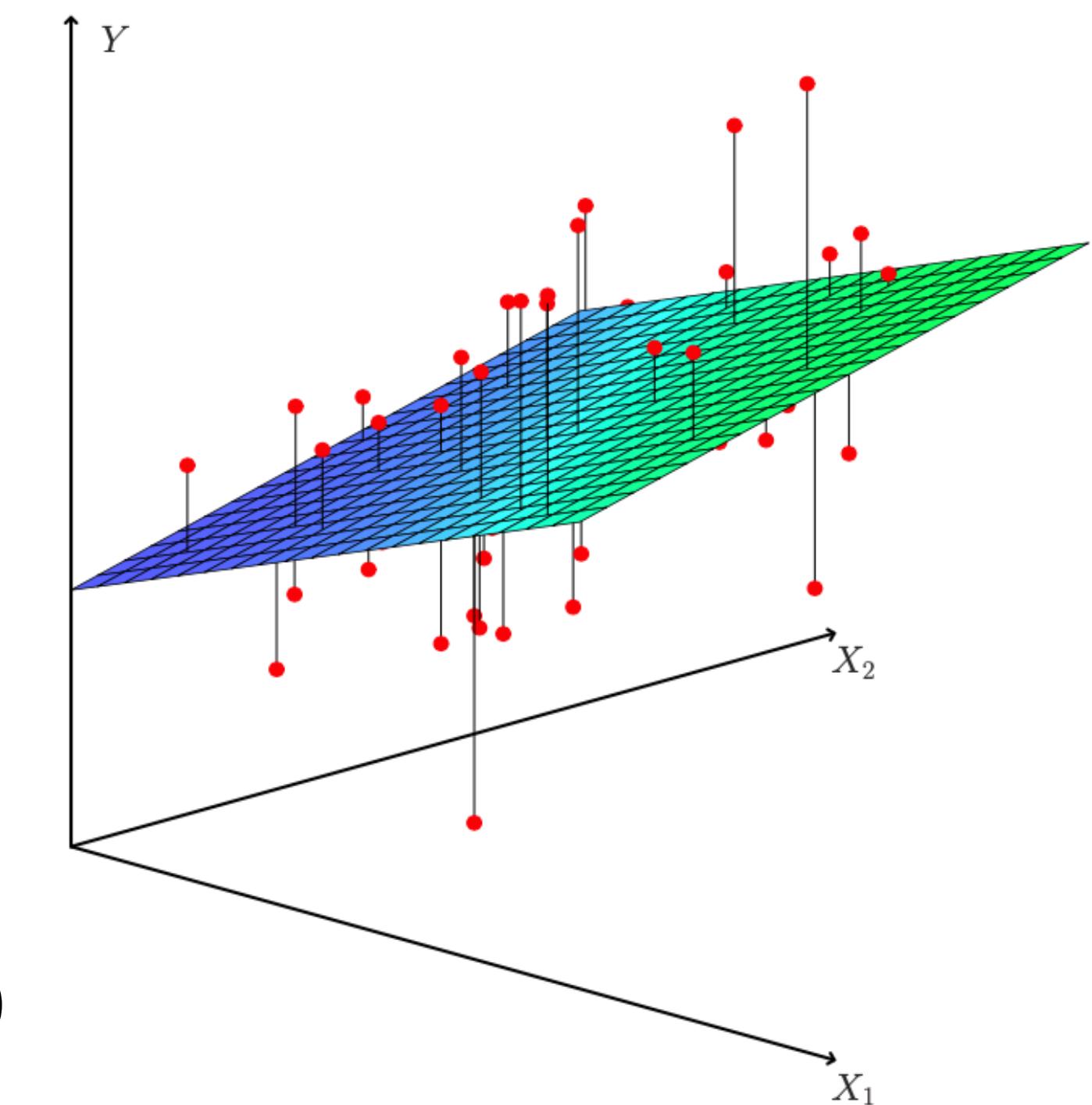
- Parametry  $\beta_1 \dots \beta_p$  określają w jaki sposób zmienne niezależne  $X_j$  wpływają na zmienną zależną  $Y$ .
- Wracając do przykładu ze zbioru danych „Advertising”:

$$\text{Sales} = \beta_0 + \beta_1 \cdot \text{TV} + \beta_2 \cdot \text{Radio} + \beta_3 \cdot \text{Newspaper} + \epsilon$$

# Regresja wielomianowa - estymacja parametrów

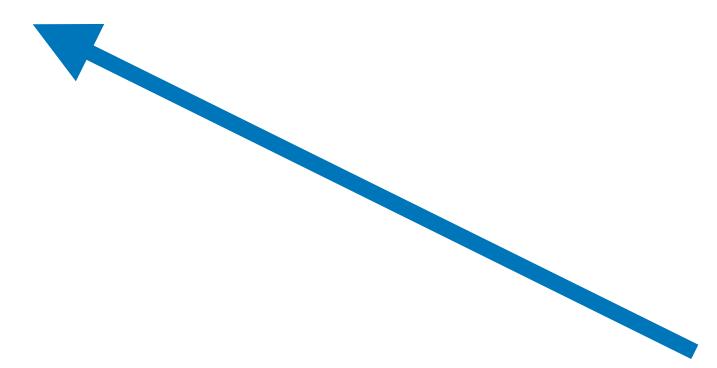
- Podobnie jak w przypadku regresji prostej, używamy metody najmniejszych kwadratów.
- Suma kwadratów odległości naszych obserwacji od **płaszczyzny** (hyperplane) definującą naszą regresję.

$$\text{RSS} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \dots - \hat{\beta}_p x_{ip})^2$$



# Ewaluacja parametrów

- Podobnie jak w przypadku regresji prostej, chcemy odpowiedzieć na pytanie: czy jest (istotna statystycznie) zależność między zmiennymi niezależnymi a zmienną zależną
  - Hipoteza zerowa:  $H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$
  - Hipoteza alternatywna:  $H_1$  : conajmniej jeden parametr  $\beta_j$  jest różny od zera
- Wciąż moglibyśmy użyć p-value pojedynczych parametrów, tak jak w przypadku regresji prostej, jednakże może być to problematyczne dla dużej ilości zmiennych
- Do odrzucenia hipotezy zerowej używamy więc testu F, a w szczególności statystyki F (F-statistics)
  - Test F bada łączną istotność parametrów dla całego modelu
  - Wartość  $F$ -statistics powinna być dużo większa od 1



Polecam zainteresowanym

# Evaluacja parametrów

Współczynniki  $\beta$

```
X_train = df[["TV", "Radio", "Newspaper"]]  
y_train = df["Sales"]  
  
X2 = sm.add_constant(X_train)  
est = sm.OLS(y_train, X2)  
regression_model = est.fit()  
print(regression_model.summary())
```

OLS Regression Results						
Dep. Variable:	Sales	R-squared:	0.903			
Model:	OLS	Adj. R-squared:	0.901			
Method:	Least Squares	F-statistic:	605.4			
Date:	Fri, 12 May 2023	Prob (F-statistic):	8.13e-99			
Time:	21:25:35	Log-Likelihood:	-383.34			
No. Observations:	200	AIC:	774.7			
Df Residuals:	196	BIC:	787.9			
Df Model:	3					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	4.6251	0.308	15.041	0.000	4.019	5.232
TV	0.0544	0.001	39.592	0.000	0.052	0.057
Radio	0.1070	0.008	12.604	0.000	0.090	0.124
Newspaper	0.0003	0.006	0.058	0.954	-0.011	0.012
Omnibus:	16.081	Durbin-Watson:	2.251			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	27.655			
Skew:	-0.431	Prob(JB):	9.88e-07			
Kurtosis:	4.605	Cond. No.	454.			

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

$R^2$

Wartość statystyki F

P-value dla hipotezy  $H_0$   
(Niska wartość → odrzucamy  $H_0$ )

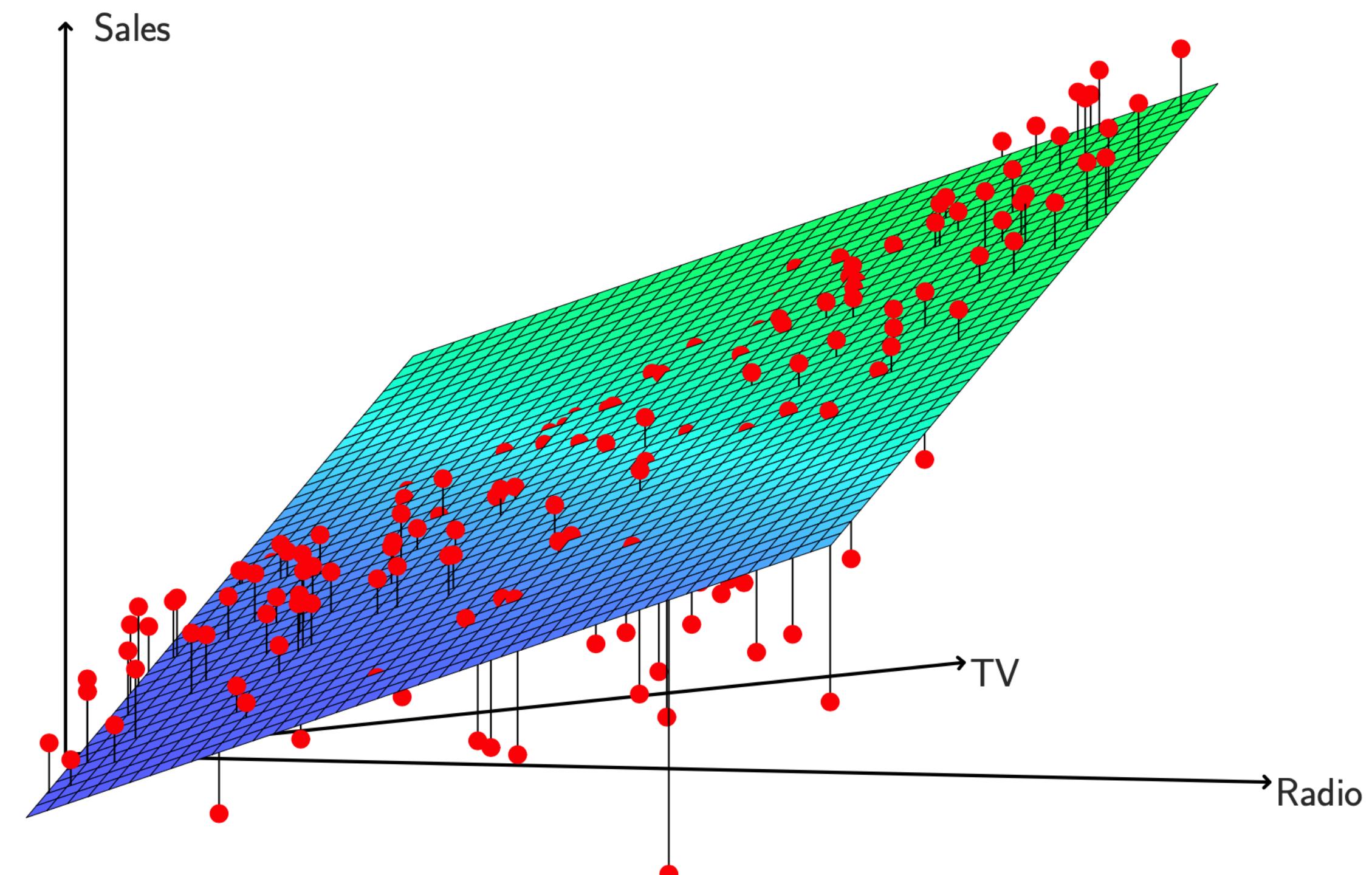
# 04 - Intro to multivariate regression

# Wybór zmiennych

- Na poprzednim przykładzie widzieliśmy, że nie zawsze warto „wrzucać” wszystkie zmienne do regresji liniowej
- Strategie wyboru zmiennych:
  - *Forward selection* - najpierw robimy  $p$  prostych regresji, patrzymy która zmienna ma największy współczynnik  $R^2$  (lub najmniejszy RSS) i wybieramy ją do modelu, kontynuujemy aż do ustalonego momentu
  - *Backward selection* - budujemy model regresji liniowej ze wszystkich zmiennych, a następnie usuwamy zmienną o największym parametrze p-value, kontynuujemy aż do ustalonego momentu (np. Wszystkie zmienne mają p-value poniżej ustalonego progu)
  - *Mixed selection* - kombinacja powyższych: zaczynamy tak jak przy strategii *forward selection*, ale gdy dla którejkolwiek zmiennej parametr p-value wzrośnie powyżej pewnej wartości, wtedy taką zmienną usuwamy.

# Efekt „synergii”

- Regresja liniowa zakłada, że zależność między zmiennymi niezależnymi a zmienną zależną jest liniowa i *addytywna*, tj. Zależność między  $Y \sim X_j$  nie zależy od innych zmiennych.
- W przypadku zbioru „advertising” nasza regresja niedoszacowuje sytuacji, w których wydajemy budżet na dwa kanały: radio i telewizję.
- Pojawia się nieliniowość sugerująca pewien rodzaj synergii lub interakcji między obiema kanałami reklamy, skutkującym większą sprzedażą niż gdyby użyć tych kanałów odziennie.



# Efekt „synergii”

- Można taką sytuację zapisać:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

- Albo:

$$\text{Sales} = \beta_0 + \beta_1 \cdot \text{TV} + \beta_2 \cdot \text{Radio} + \beta_3 \cdot \text{TV} \cdot \text{Radio} + \epsilon$$

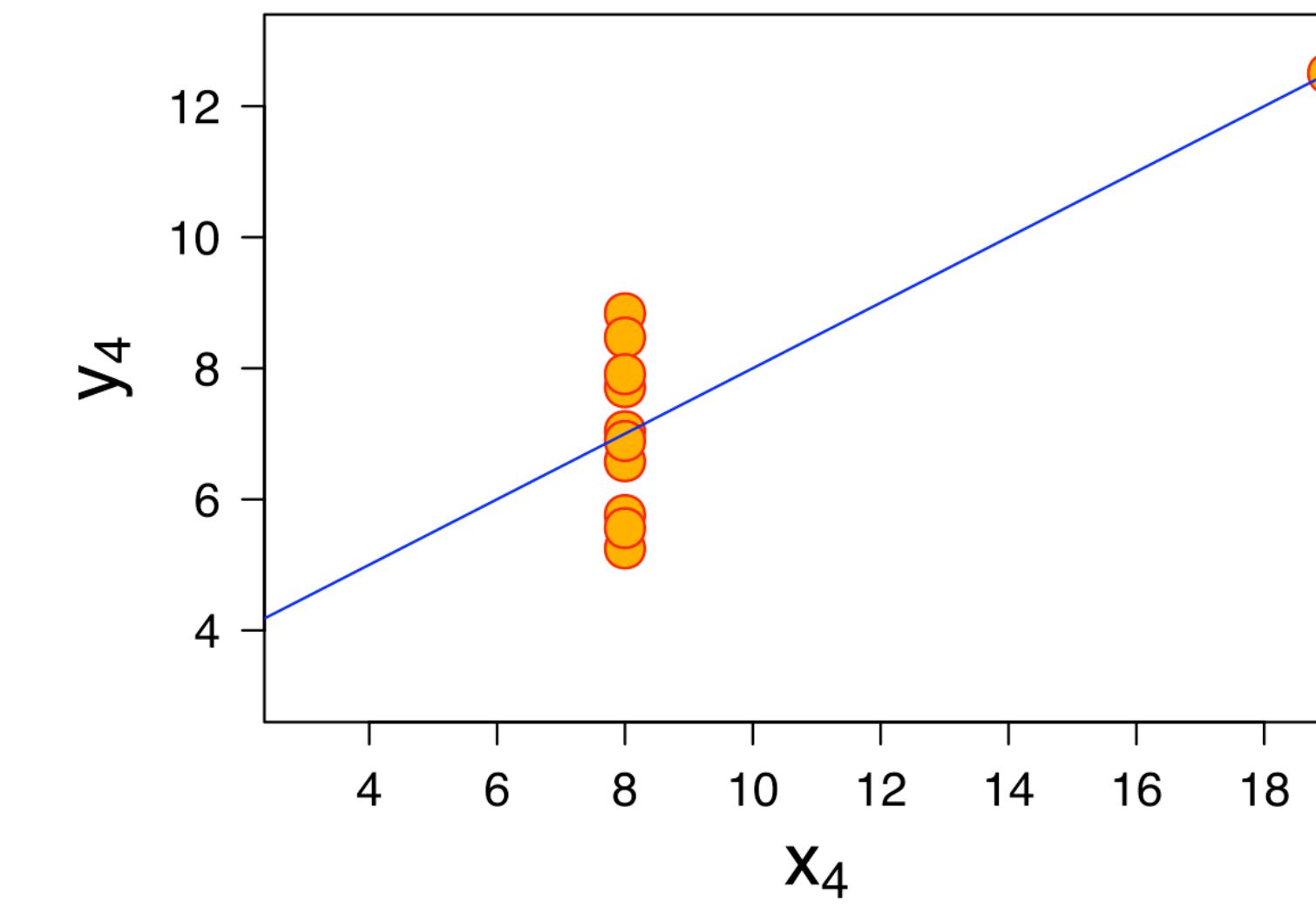
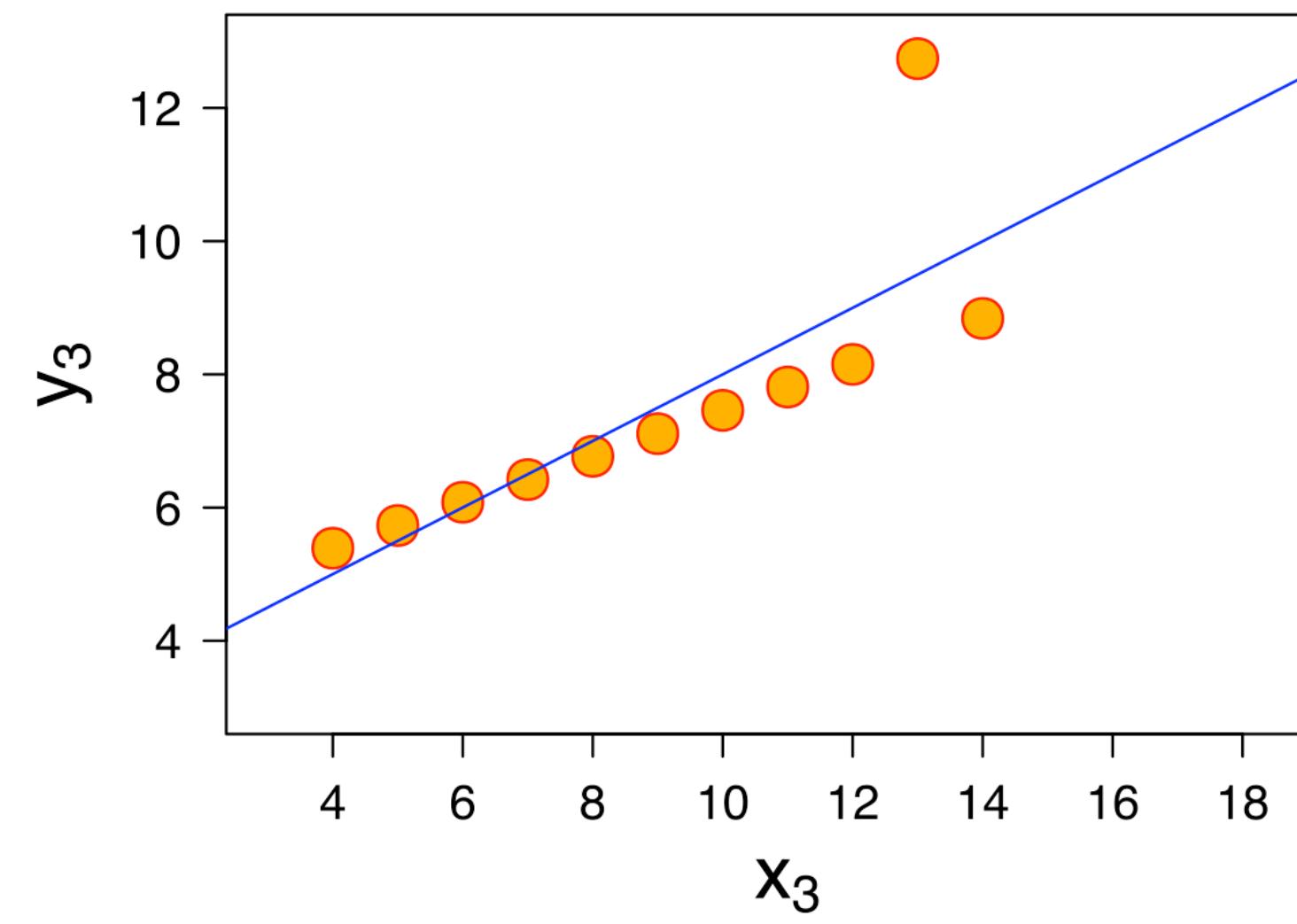
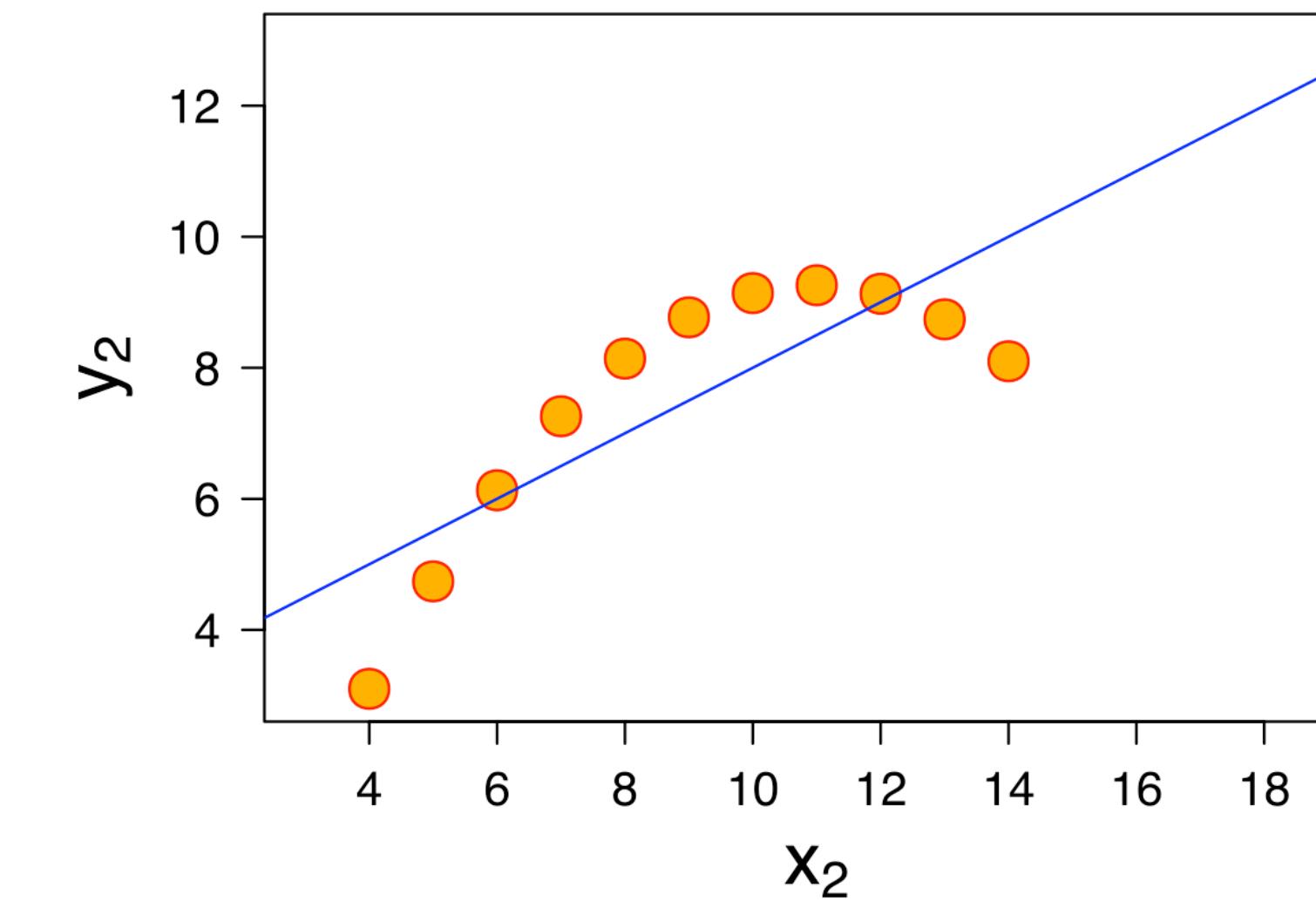
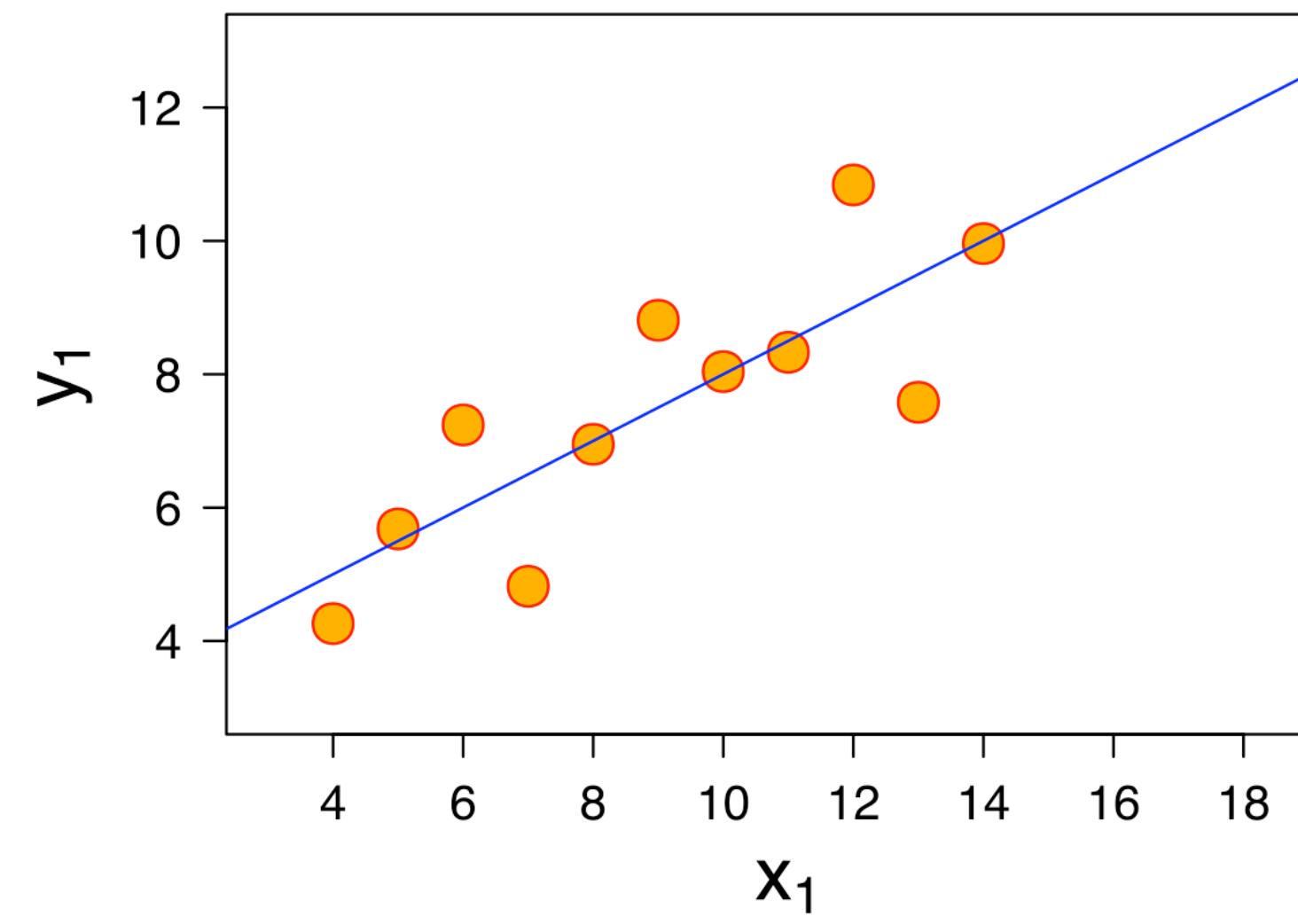
- Taki model znów staje się liniowy

# 04 - Intro to multivariate regression

# Kiedy stosować regresję liniową

- Zależność między zmiennymi niezależnymi a zmienną zależną jest liniowa
  - Eksploracja danych: wykresy punktowe (scatterplot)
- Regresja liniowa jest mało odporna na dane odstające (outliers)
  - Rozważyć odrzucenie outlierów
- Zmienne niezależne nie są skorelowane ze sobą
  - Odrzucić „nadmiarowe” zmienne
- Jednorodność wariancji, homoskedastyczność (homoscedasticity) - błędy wartości przewidywanej są w miarę stałe i np. nie rosną wraz ze wzrostem wartości przewidywanej

# Kiedy stosować regresję liniową



**“No one ever made a decision  
because of a number.  
They need a story.”**

*Daniel Kahneman*

# Follow up

- Nie damy rady omówić wszystkich matematycznych i technicznych aspektów regresji
- Dla zainteresowanych polecam zgłębić problemy:
  - Regularyzacji (regularisation)
  - Metoda gradientu prostego (gradient descent) w znajdowaniu minimum funkcji celu
  - Jak w praktyce wygląda trenowanie regresji liniowej
- Materiały:
  - Andrew Ng videos on YouTube: „Lecture 4.2 – Linear Regression With Multiple Variables”
  - Książka: „The Elements of Statistical Learning” Jerome H. Friedman, Robert Tibshirani i Trevor Hastie

# Rodzaje danych wejściowych

# Rzut oka na przykłady

- Titanic dataset

PassengerId	Survived	Pclass		Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th... Heikkinen, Miss. Laina	female	38.0	1	0	PC 17599 STON/O2. 3101282	71.2833	C85	C
2	3	1	3	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	26.0	0	0	113803	7.9250	NaN	S
3	4	1	1	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

# Rodzaje danych wejściowych

- Wartości liczbowe
- Wartości skategoryzowane
- Wartości tekstowe (strings)
- Daty
- Dźwięk
- Obrazy
- ...

# Wartości skategoryzowane

- Zamiana na tzw. „Dummy variables” albo „1 z n” (one hot encoding)
- Przykład

Person	Education	Salary(annum)
0 amit	Under-Graduate	\$90k
1 vishal	Diploma	\$80k
2 john	Under-Graduate	\$90k
3 marry	Diploma	\$60k
4 sherin	Under-Graduate	\$90k
5 komal	Under-Graduate	\$100k
6 jay	Under-Graduate	\$60k
7 shree	Under-Graduate	\$100k
8 kishore	Diploma	\$90k
9 geetha	Diploma	\$70k
10 savitha	Under-Graduate	\$50k
11 vinith	Under-Graduate	\$90k

	Diploma	Master's	Under-Graduate
0	0	0	1
1	1	0	0
2	0	1	0
3	0	1	0
4	0	0	1
5	1	0	0
6	0	0	1
7	0	0	1
8	0	0	1
9	0	0	1
10	1	0	0
11	1	0	0
12	0	0	1
13	0	0	1

# Brakujące dane

# Przykłady

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	Nan	S
1	2	1	Cumings, Mrs. John Bradley (Florence Briggs Th... Heikkinen, Miss. Laina	female	38.0	1	0	PC 17599 STON/O2. 3101282	71.2833 7.9250	C85 NaN	C S
2	3	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	Nan	S

**Nan - Not a Number**

**NA - Not Available**

**," - puste pole**

# Reprezentacja brakujących danych

- Strategie traktowania brakujących danych:
  - Usunięcie (wiersza - danej obserwacji, kolumny - odrzucenie zmiennej niezależnej)
  - Uzupełnienie danych
    - Wartości liczbowe: interpolacją wartości, średnia, taka sama jak poprzednia, inne bardziej zaawansowane metody
    - Wartości skategoryzowane: dodanie nowej klasy

# 05 - Handling data

# Zasady zaliczenia

- Zaliczeniem będzie projekt robiony w grupie 3-4 osobowej (max 5).
- Projekt polega na zastosowaniu regresji liniowej lub klasyfikacji (regresji logistycznej) na wybranym zbiorze danych
  - Zaproponuje 2 zbiory danych do wyboru: jeden na klasyfikacje, drugi na regresje
  - Każdy może też przyjść ze swoim zbiorem danych (**wtedy zalecam konsultacje ze mną, żeby określić zakres**)
- Rezultatem projektu powinien być notebook (collab), który może zawierać elementy omawiane na wykładzie:
  - Wczytanie danych, może jakaś wizualizacja zbioru
  - Zastosowanie algorytmu uczenia maszynowego
  - Ewaluacja: stworzenie macierzy błędów/metryk modelu/współczynnik R2 regresji itd.
- Zachęcam do wzorowania się na notebookach prezentowanych na wykładzie
- Projekt (notebook) najlepiej wysłać mi mejlem
- Termin do 18 czerwca

# Zbiory danych

- Titanic dataset (klasyfikacja)
- Penguins (klasyfikacja)
- Boston housing (regresja)
- Auto (regresja)
- Credit (klasyfikacja lub regresja)

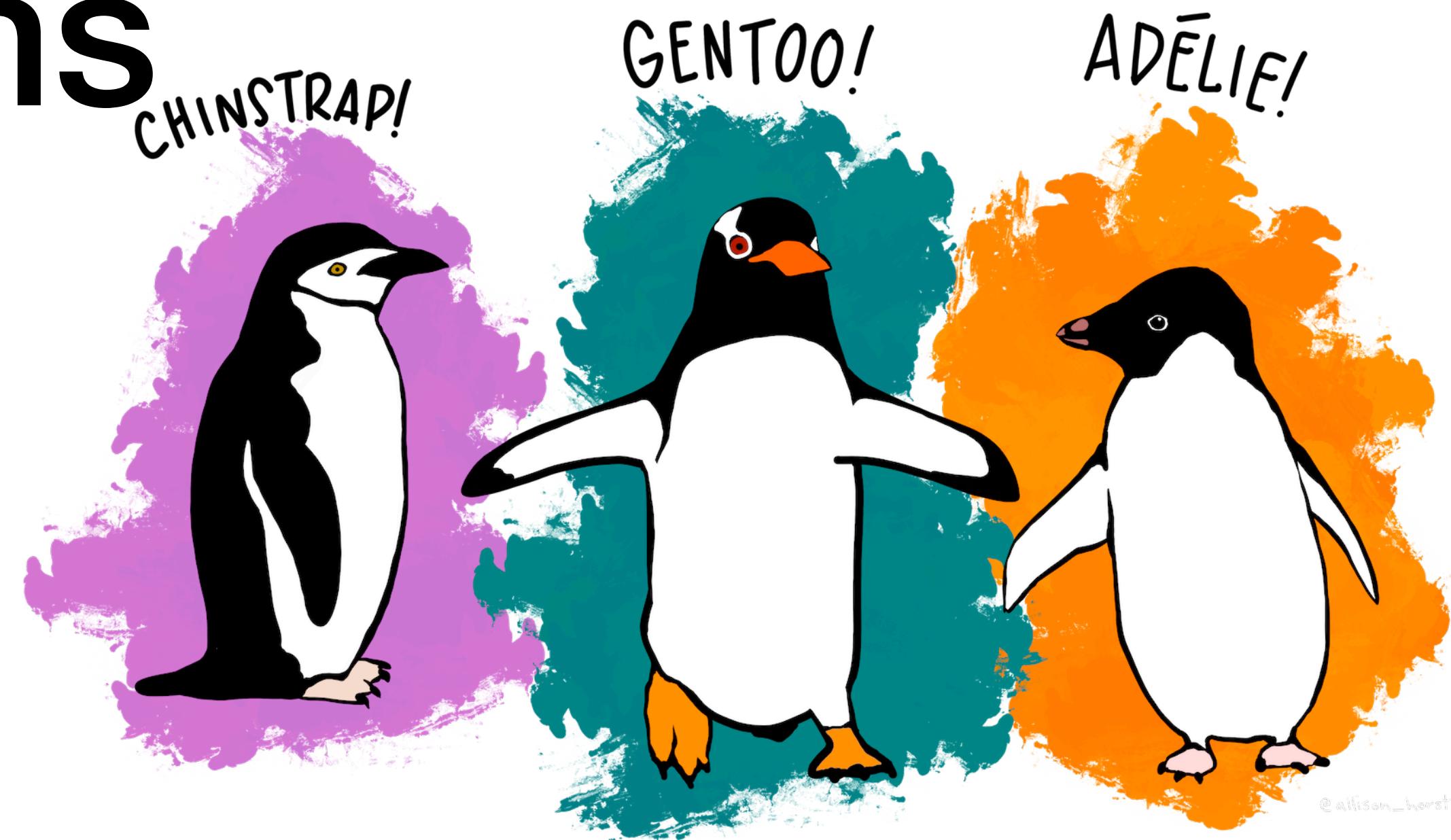
# Titanic

- Zbiór danych zwiera 12 kolumn i 891 wierszy. Znaczenie kolumn:
  - **survival: przeżycie 0 = Nie, 1 = Tak**
  - pclass: rodzaj biletu 1 = 1 Klasa, 2 = 2 Klasa, 3 = 3 Klasa
  - sex: płeć
  - age: wiek w latach
  - sibsp: ilość rodzeństwa (+ żona) na pokładzie
  - parch: ilość rodziców / dzieci na pokładzie
  - ticket: numer biletu
  - fare: opłata
  - cabin: numer kabiny
  - embarked: port zaokrętowania (wejścia na pokład) C = Cherbourg, Q = Queenstown, S = Southampton
- Przykład zadania: predykcja przeżycia



# Penguins

- Zbiór danych zawiera 7 kolumn
  - **species: gatunek (Chinstrap, Adélie, or Gentoo)**
  - culmen\_length\_mm: długość dzioba
  - culmen\_depth\_mm: wysokość dzioba
  - flipper\_length\_mm: długość skrzydła
  - body\_mass\_g: masa ciała
  - island: nazwa wyspy z Archipelagu Palmera (Antarktyda)
  - sex: płeć
- Przykład zadania: klasyfikacja gatunku na podstawie cech anatomicznych



CULMEN: RIDGE ALONG THE  
TOP PART OF A BIRD'S BILL



# MTCARS

- Zbiór danych z 392 obserwacjami i 9 zmiennymi:
  - mpg - miles per gallon - zużycie paliwa (w milach na galon paliwa)
  - cylinders - liczba cylindrów (od 4 do 8)
  - displacement - objętość skokowa (cale sześcienne)
  - horsepower - moc (konie mechaniczne)
  - weight - waga (w funtach)
  - acceleration - przyspieszenie w sekundach do 60 mil na godzinę
  - year - rok produkcji (z ostatnie cyfry roku, XX wiek)
  - origin - kraj produkcji (1. American, 2. European, 3. Japanese)
  - name - nazwa samochodu
- Przykład zadania: od czego zależy bardziej ekonomiczne zużycie paliwa?

# Credit

- Zbiór danych 400 fikcyjnych klientów banku:
  - Income - dochody (w tys \$)
  - Limit - limit kredytowy
  - Rating - rating
  - Cards - liczba kart kredytowych
  - Age - wiek (w latach)
  - Education - edukacja (w latach)
  - Own - czy jest właścicielem mieszkania/domu
  - Student - czy jest studentem
  - Married - czy jest żonaty/ma męża
  - Region - region (East, South, and West)
  - **Balance - dług w \$**
- Przykładowe zadanie: jak przewidzieć wysokość długu? Od czego on zależy?

# Boston housing

- Zbiór danych zawierający wartość nieruchomości z 506 obszarów Bostonu. Posiada 13 cech (zmiennych niezależnych):
  - crim - liczba przestępstw w przeliczeniu na mieszkańca
  - zn - procent działek/nieruchomości o wielkości pow. 25000 stopy kwadratowej
  - indus - procent obszaru przeznaczonego pod funkcje przemysłowe
  - chas - obszar bezpośrednio graniczy z rzeką
  - nox - współczynnik zanieczyszczenia tlenkami azotu NO (ppm)
  - rm - średnia ilość pokoi w mieszkaniu
  - age - procent budynków wybudowanych przez 1940 rokiem
  - dis - odległość (średnia ważona) od 5 głównych stref biurowo/przemysłowych
  - rad - index dostępu do obwodnic/dróg szybkiego ruchu
  - tax - stawka podatku gruntowego (od każdych 10.000\$)
  - ptratio - współczynnik ilości dzieci na jednego nauczyciela
  - lstat - procent mieszkańców o niższym statusie ekonomicznym
  - **medv - mediana wartości nieruchomości (w 1000\$)**
- Przykład zadania: od czego uzależniona jest cena nieruchomości? (Regresja liniowa)