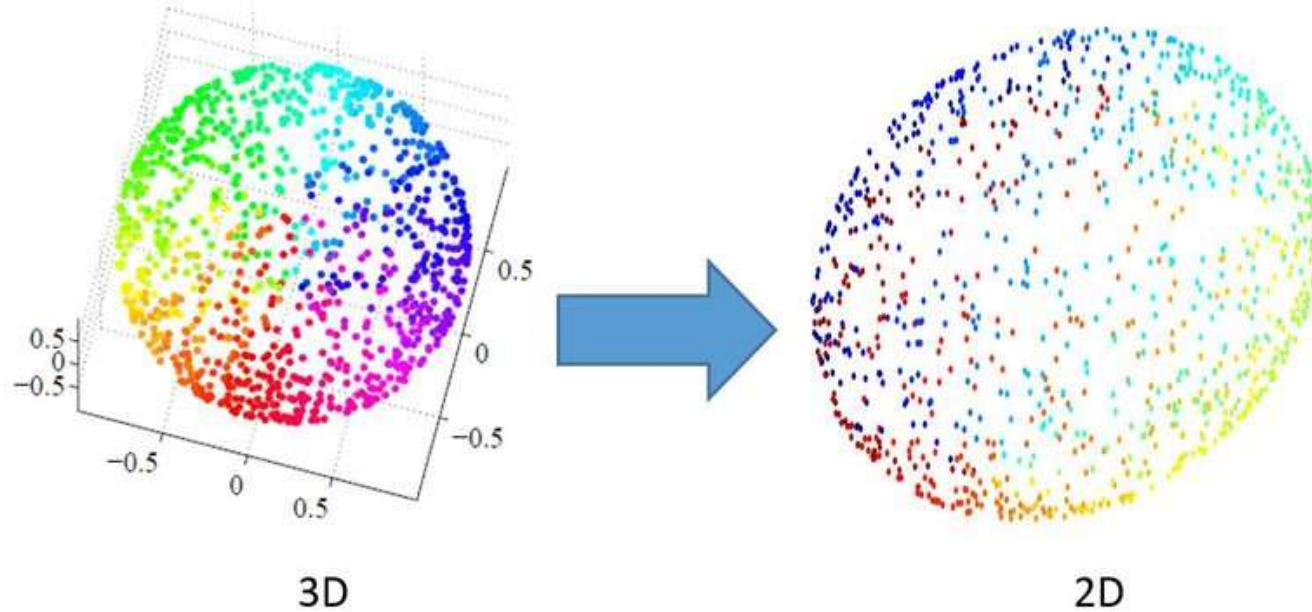


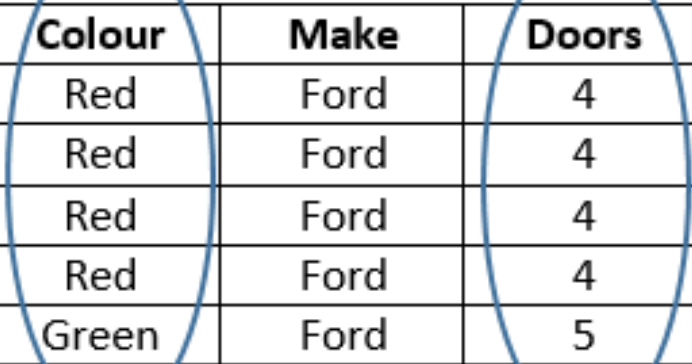
# Boyut Küçültme Nedir?

- Gerçek hayattaki veriler çok fazla boyuta (özniteliğe) sahip oluyor ve boyut büyüdükçe veri temizlemeden model kurmaya bütün süreçlerde harcamamız gereken **zaman ve kaynaklar** artıyor. Boyut azatma bu sorunun önüne geçmek için kullanılan yöntemlerden biridir. Hemen her veri setinde bazı öznitelikler arasında yüksek korelasyon oluyor ve bu bizim gereksiz bilgiye sahip olmamıza ve modelimizde **overfitting problemine** sebep olabiliyor.



# Nasıl Uygulanır ?

- Boyut küçültmenin en kolay yolu verimizi en iyi tanımlayan öznitelikleri bulup diğerlerini atmaktır (**öznitelik seçimi — feature selection**). Dikkat edilmesi gereken nokta en az bilgi kaybıyla bu işi yapmaktır ve aslında önemli olan öznitelikleri silmemektir. Bunu yapmak için verideki dağılımın maksimum varyansını-bilgisini tutan minimum sayıda değişken oluşturuyoruz. Eğer bir değişken her örnek için aynı değere sahip ise gereksiz bir değişkendir. Biz en yüksek varyansa sahip olan değişkenleri bulmalıyız.



Colour	Make	Doors	Exhaust
Red	Ford	4	1
Red	Ford	4	1
Red	Ford	4	1
Red	Ford	4	1
Green	Ford	5	1

# **HDI(İnsani Kalkınma Endeksi)**

- Buna en güzel örneklerden biri HDI(İnsani Kalkınma Endeksi) hesaplaması olabilir. Ülkelerin sahip olduğu değişkenlerden bazıları (wiki): resmi dil, yüz ölçümü (toplam), su oranı, nüfus, nüfus yoğunluğu, GSYH, para birimi, trafik akışı, hukuk, dış ilişkiler, din, eğitim, sağlık, sanayi, tarım, turizm... Bunlara ek olarak ülkeden ülkeye değişen yüzlerce değişken sıralayabiliriz.


# HDI(İnsani Kalkınma Endeksi)

- PCA metodu sanayi, tarım, turizm vs. ekonomiyle ilgili yüzlerce değişkeni ve nüfusu kullanıp kişi başına düşen gayri safi milli gelir diye tek bir değişken oluşturuyor. Yani yüzlerce boyutluk bilgi en az bilgi kaybıyla tek boyutta tutulmuş oluyor.
- HDI hesaplamalarında sadece 5 öz niteliğe bakılarak pek de itiraz edilmeyecek bir tablo karşımıza çıkıyor.
- Kişi başına düşen gayri safi milli gelir
- İnsan gelişmişlik endeksi — HDI sıralaması
- Beklenen ortalama yaşam süresi
- Beklenen eğitim yılı
- Ortalama eğitim yılı

		Human Development Index (HDI)	Life expectancy at birth	Expected years of schooling		Mean years of schooling		Gross national income (GNI) per capita		GNI per capita rank minus HDI rank	HDI rank
HDI rank	Country	Value	(years)	(years)		(years)		(2011 PPP \$)			
		2015	2015	2015	a	2015	a	2015		2015	2014
	VERY HIGH HUMAN DEVELOPMENT										
1	Norway	0.949	81.7	17.7		12.7		67,614		5	1
2	Australia	0.939	82.5	20.4	b	13.2		42,822		19	3
2	Switzerland	0.939	83.1	16.0		13.4		56,364		7	2
4	Germany	0.926	81.1	17.1		13.2	c	45,000		13	4
5	Denmark	0.925	80.4	19.2	b	12.7		44,519		13	6
5	Singapore	0.925	83.2	15.4	d	11.6		78,162	e	-3	4
7	Netherlands	0.924	81.7	18.1	b	11.9		46,326		8	6
8	Ireland	0.923	81.1	18.6	b	12.3		43,798		11	8
9	Iceland	0.921	82.7	19.0	b	12.2	c	37,065		20	9
10	Canada	0.920	82.2	16.3		13.1	f	42,582		12	9
10	United States	0.920	79.2	16.5		13.2		53,245		1	11
12	Hong Kong, China (SAR)	0.917	84.2	15.7		11.6		54,265		-2	12
13	New Zealand	0.915	82.0	19.2	b	12.5		32,870		20	13
14	Sweden	0.913	82.3	16.1		12.3		46,251		2	15
15	Liechtenstein	0.912	80.2	14.6	g	12.4	h	75,065	e,i	-11	14
16	United Kingdom	0.909	80.8	16.3		13.3		37,931		10	16
17	Japan	0.903	83.7	15.3		12.5	c	37,268		10	17
18	Korea (Republic of)	0.901	82.1	16.6		12.2		34,541		12	18
19	Israel	0.899	82.6	16.0		12.8		31,215		16	19
20	Luxembourg	0.898	81.9	13.9		12.0		62,471		-12	20

# PCA Metodu

- Boyut azaltma işlemlerinde sıklıkla kullanılan PCA verideki gerekli bilgileri ortaya çıkarmada oldukça etkili bir yöntemdir. PCA'nın arkasında yatan temel mantık çok boyutlu bir veriyi, verideki temel özellikleri yakalayarak daha az sayıda değişkenle göstermektir. En iyi sonucu elde etmek için genelde öncelikle standardizasyon yapılır.
- `x = StandardScaler().fit_transform(x)`

	sepal length	sepal width	petal length	petal width	Standardization 		sepal length	sepal width	petal length	petal width
0	5.1	3.5	1.4	0.2		0	-0.900681	1.032057	-1.341272	-1.312977
1	4.9	3.0	1.4	0.2		1	-1.143017	-0.124958	-1.341272	-1.312977
2	4.7	3.2	1.3	0.2		2	-1.385353	0.337848	-1.398138	-1.312977
3	4.6	3.1	1.5	0.2		3	-1.506521	0.106445	-1.284407	-1.312977
4	5.0	3.6	1.4	0.2		4	-1.021849	1.263460	-1.341272	-1.312977

# PCA Metodu

- Sonrasında PCA uygulanılır.
- `from sklearn.decomposition import PCA`
- `pca = PCA(n_components=2)`
- `principalComponents = pca.fit_transform(x)`
- `principalDf = pd.DataFrame(data = principalComponents, columns = ['principal component 1', 'principal component 2'])`

	sepal length	sepal width	petal length	petal width		principal component 1	principal component 2
0	-0.900681	1.032057	-1.341272	-1.312977	PCA (2 components) ➔	-2.264542	0.505704
1	-1.143017	-0.124958	-1.341272	-1.312977		-2.086426	-0.655405
2	-1.385353	0.337848	-1.398138	-1.312977		-2.367950	-0.318477
3	-1.506521	0.106445	-1.284407	-1.312977		-2.304197	-0.575368
4	-1.021849	1.263460	-1.341272	-1.312977		-2.388777	0.674767