Analyzing the PIMA Indians dataset

The context

Pima Indians are a group of Native Americans living in an area consisting of what is now central and southern Arizona.
They have the highest prevalence of type 2 diabetes in the world.
This is determined by genetic and environmental factors. 34% of men and 47 % of woman have diabetes.

The dataset[1]

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old.

Variables

- preg: Number of times pregnant
- plas: Plasma glucose concentration at 2 hours in an oral glucose tolerance test
- pres: Diastolic blood pressure (mm Hg)
- skin: Triceps skin fold thickness (mm)
- insu: 2-Hour serum insulin (mu U/ml)
- mass: Body mass index (weight in kg/(height in m)^2)
- pedi: Diabetes pedigree function
- age: Age (years)
- class: Class variable (0 = no diabetes or 1 = diabetes)

Exploratory data analysis

	Data types	Is null?	zeros count
preg	int64	0	111
plas	int64	0	5
pres	int64	0	35
skin	int64	0	227
insu	int64	0	374
mass	float64	0	11
pedi	float64	0	0
age	int64	0	0
class	int64	0	500

	preg	plas	pres	skin	insu	mass	pedi	age	class
count	768.00	768.00	768.00	768.00	768.00	768.00	768.00	768.00	768.00
mean	3.85	120.89	69.11	20.54	79.80	31.99	0.47	33.24	0.35
std	3.37	31.97	19.36	15.95	115.24	7.88	0.33	11.76	0.48
min	0.00	0.00	0.00	0.00	0.00	0.00	0.08	21.00	0.00
25%	1.00	99.00	62.00	0.00	0.00	27.30	0.24	24.00	0.00
50%	3.00	117.00	72.00	23.00	30.50	32.00	0.37	29.00	0.00
75%	6.00	140.25	80.00	32.00	127.25	36.60	0.63	41.00	1.00
max	17.00	199.00	122.00	99.00	846.00	67.10	2.42	81.00	1.00

Observations

- Funny number: 17 times pregnant
- Value range between the observations is high
- Big jump between 75% and max for preg, skin and insu -> outlier? BMI of 67 realistic?
- No systolic blood pressure -> why?

Reflection

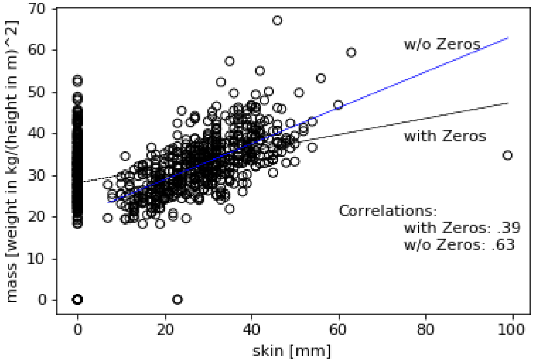
- If there is an article about the data read it?
- Look at the data and the metadata diligently with domain knowledge, like units of measures etc.
- Coming up with a clear interpretation from the descriptive statistics / viz is not always that straight forward

Correlations highlights:

- preg – age: 0.54
- class – plas: 0.47
- skin – insu: 0.44
- skin – mass: 0.39

But…

Bivariate regression and the impact of the zeros



Observations

- Corr between age and number of pregnancies makes sense
- Plasma glucose might be a good discriminator
- Zeros have a big impact on the extent of association between the variables

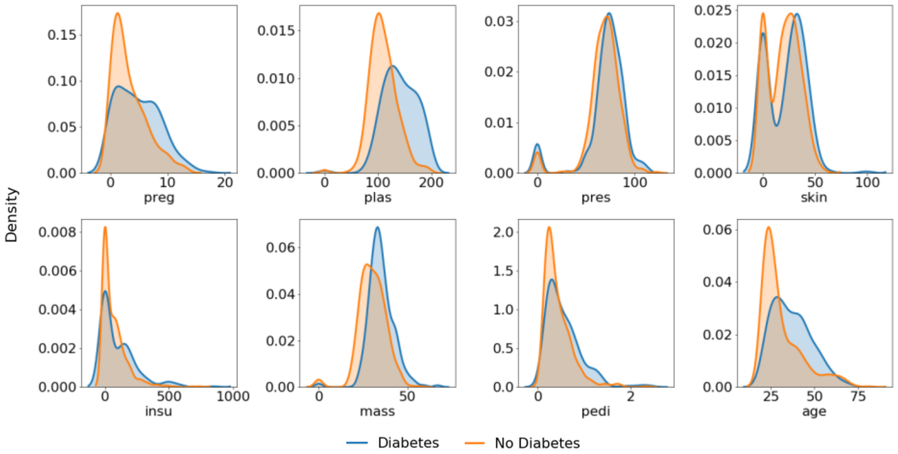
Reflection

- What is a weak, moderate, high correlation?
- Be aware of how you deal with missing values and outliers and how this might affect the conclusions you draw

Hypothesis testing

Class counts: Without diabetes / with diabetes: n = 500 / n= 268

Kernel density plots



Do central tendencies differ?

	Mann Whitney with Zeros	Mann Whitney w/o Zeros	T-Test with Zeros	T-Test w/o Zeros
preg	p < .05	p < .05	p < .05	p < .05
plas	p < .05	p < .05	p < .05	p < .05
pres	p < .05	p < .05	p > .05	p < .05
skin	p < .05	p < .05	p < .05	p < .05
insu	p < .05	p < .05	p < .05	p < .05
mass	p < .05	p < .05	p < .05	p < .05
pedi	p < .05	p < .05	p < .05	p < .05
age	p < .05	p < .05	p < .05	p < .05

Observations

- Blood pressure and BMI seem to be normally distributed, but they are not
- Diabetes cases are associated with greater levels of plasma glucose and BMI
- Visually all attributes seem not to discriminate a lot, but the differences are statistically significant

Reflection

- Defining which test to use on a given data set might be challenging – random sample?
- Even small numbers, e.g. 35 zeros for blood pressure can make a “big” difference
- Be diligent – check requirements for the test
- Manipulating sizes and layout in matplotlib is rather cumbersome

[1] Smith, J. W.. et. al (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261–265). IEEE Computer Society Press.