Statistics and Machine Learning 1

Lecture 3B: Multivariate Visualisation

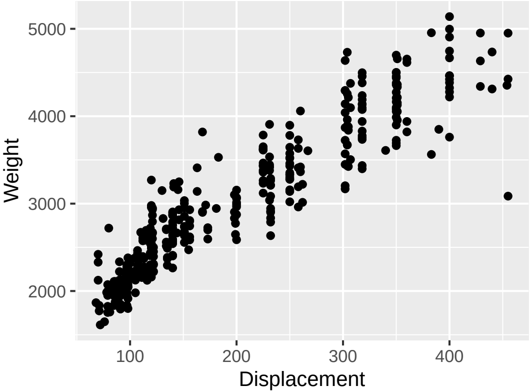
Mark Muldoon
Department of Mathematics, Alan Turing Building
University of Manchester

Week 3

Bivariate visualisation techniques: Scatter Plots

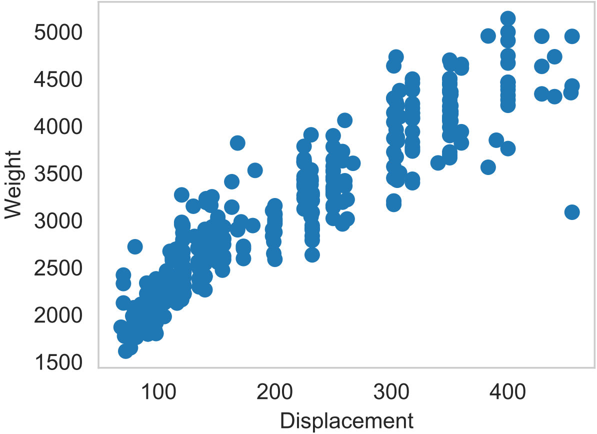
A scatter plot is a lossless visualisation that involves placing a marker at (x_{ia}, x_{ib}) for each i and some a, b.

```
ggplot(auto.data, aes(x=Displacement,y=Weight)) + geom_point()
```



Bivariate visualisation techniques: Scatter Plots

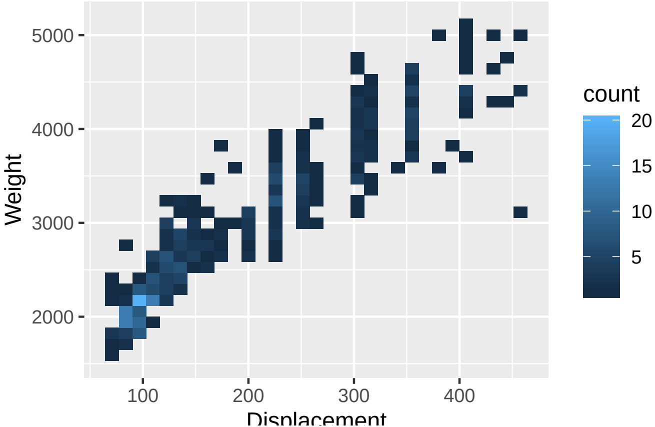
```
plt.scatter(dis,wgt)
```



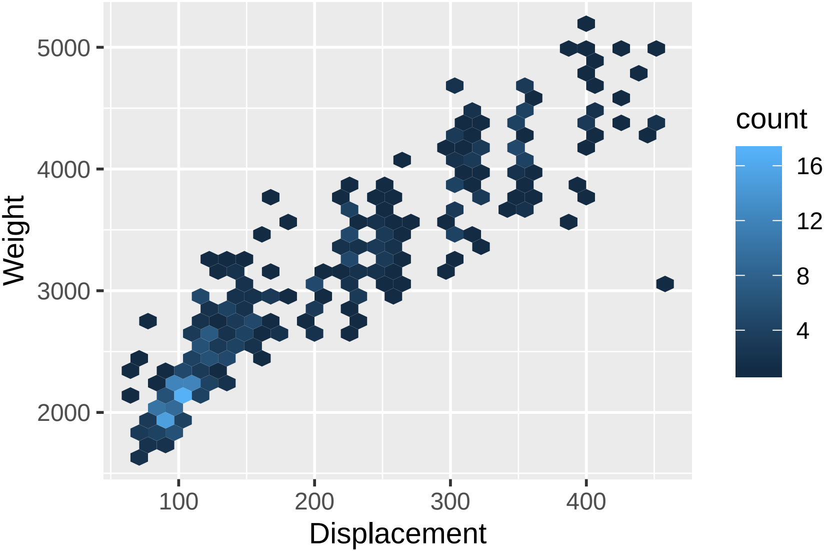
Bivariate visualisation techniques: 2d Histograms

A 2d histogram generalised the univariate in the natural way as the count of data points falling inside a given two-dimensional area.

```
ggplot(auto.data, aes(x=Displacement,y=Weight)) + geom_bin2d()
```

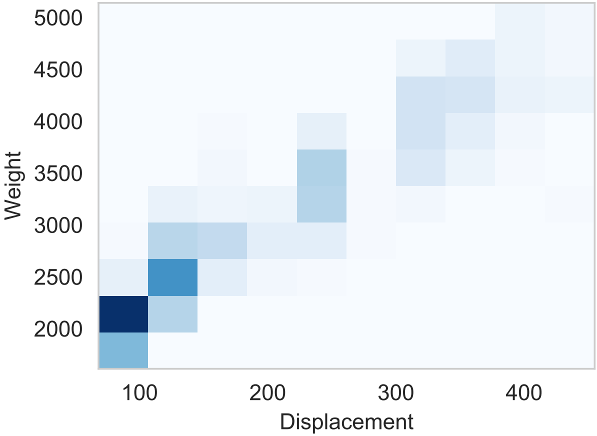


And the area need not be a rectangle!



Bivariate visualisation techniques: 2d Histograms

```
plt.hist2d(dis,wgt,cmap='Blues')
```



Multivariate DKE

- The kernel density estimate (KDE) approximates the population distribution function (as before) and is defined by

$$\hat{f}(\mathbf{x} \mid \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} K(\mathbf{x} \mid \mathbf{x}_i, \boldsymbol{\theta}). \tag{1}$$

  though here the bandwidth is replaced by a more general, potentially multivariate set of parameters, $\boldsymbol{\theta}$.

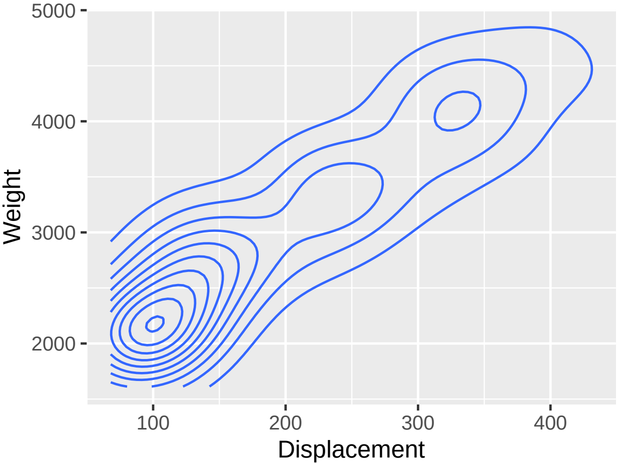
- Typically the *kernel function* K will be chosen to be the multivariate normal probability density function:

$$K(\mathbf{x} \mid \mathbf{x}_i, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x} \mid \mathbf{x}_i, \boldsymbol{\sigma}). \tag{2}$$

- A 2d kernel density plot shows estimated curves of constant $f(\mathbf{x})$.

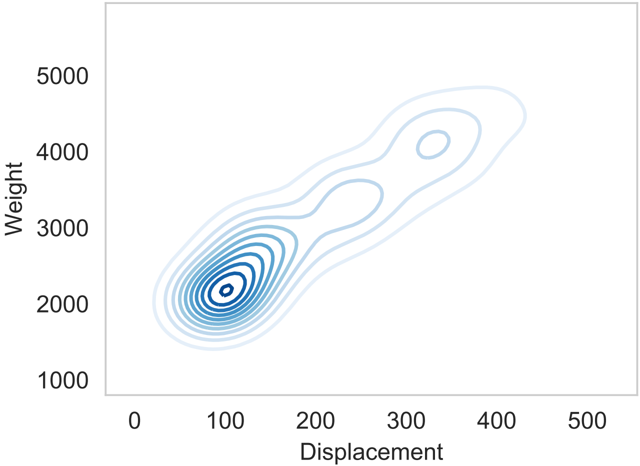
Bivariate visualisation techniques: 2d KDE

```
ggplot(auto.data, aes(x=Displacement,y=Weight)) + geom_density_2d()
```

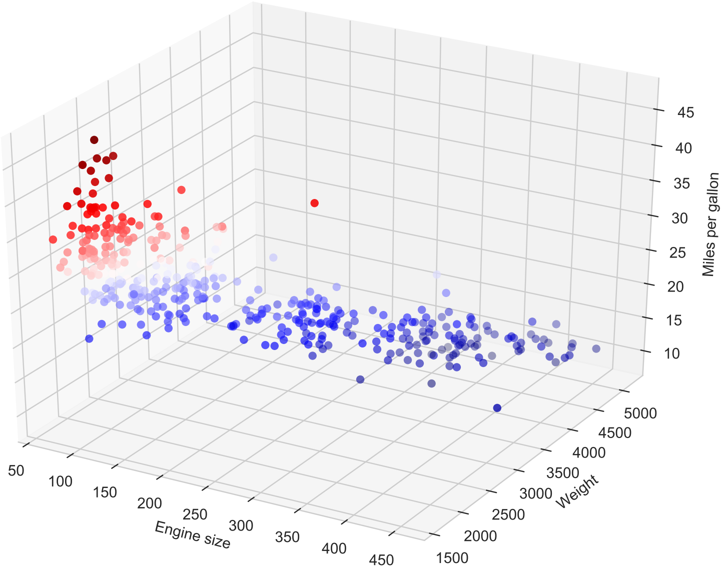


Bivariate visualisation techniques: 2d KDE

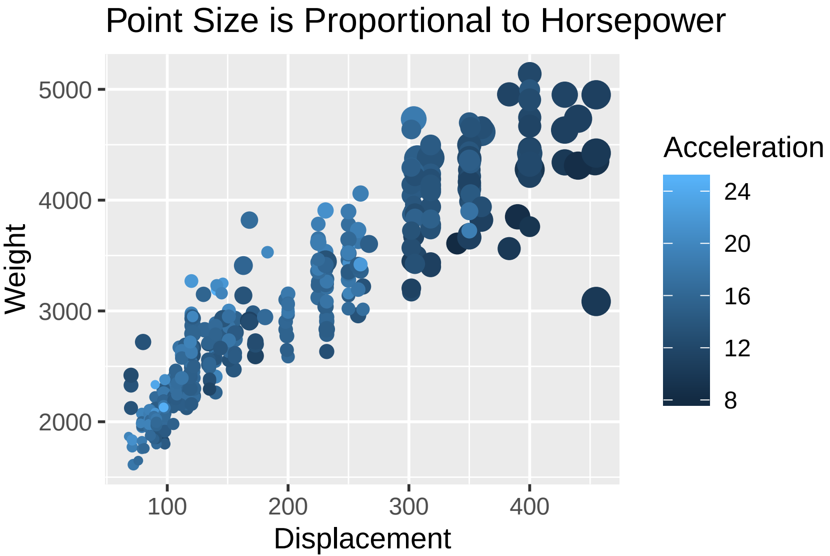
```
sns.kdeplot(dis,wgt,cmap="Blues")
```



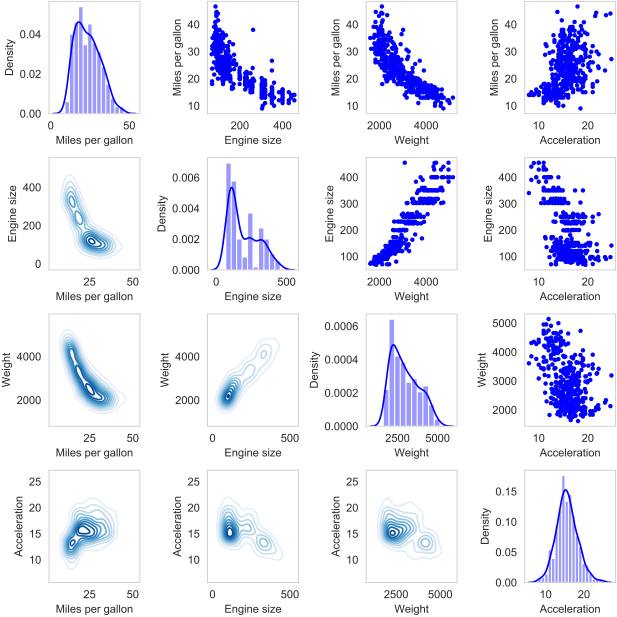
Higher dimensions: 3d Scatter



Higher dimensions: Scaled Scatter



Higher dimensions: Plot Matrices



Pairs of categorical variables

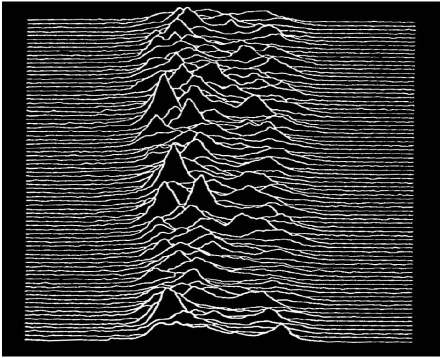
- Contingency tables:

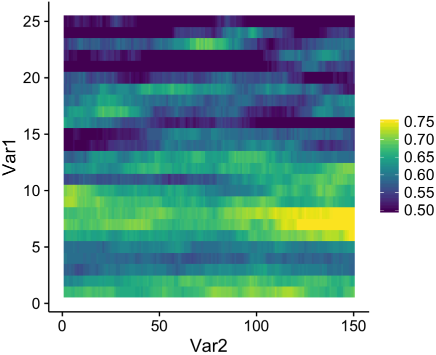
		Accident Occurred?		
		No	*Yes*	**Total**
Location:	*Offsite*	414	153	567
	Onsite	390	43	433
	Total	804	196	1000

- Can be used for categorical, ordinal and discrete variables, with more than two levels
- Can write values as proportions of each row or each column
- Can write them as proportions of the total
- Can compare them with what the values would be if the two variables were independent (values would be the products of respective marginals)

Categorical and continuous variables

For example, one density plot, or a heatmap strip, per category:




Var1
Var2