Data Mining and Machine Learning

Comp 3027J

Dr Catherine Mooney
Assistant Professor

catherine.mooney@ucd.ie

Lectures and Text

- **Core Text:**
  Fundamentals of Machine Learning for Predictive Data Analytics
  By John D. Kelleher, Brian Mac Namee and Aoife D'Arcy
- Did anyone have any problems getting the book out of the library?
- Last week we covered Chapter 2, sections 2.3 and 2.4 and Chapter 3, sections 3.1 – 3.4
- This week we will cover Chapter 3, sections 3.5 and 3.6
- Please read these sections of the book

Review of Lab

Review of Lab 1

- Most people seemed happy with the level of difficulty
- Only slight problems with Question 6
- If you you are taking this class and did not submit anything on Thursday please come and talk to me after class

Review of Lab 1

Question 6: Suppose you wanted to print out only the grade for John. Type out the command you would use.

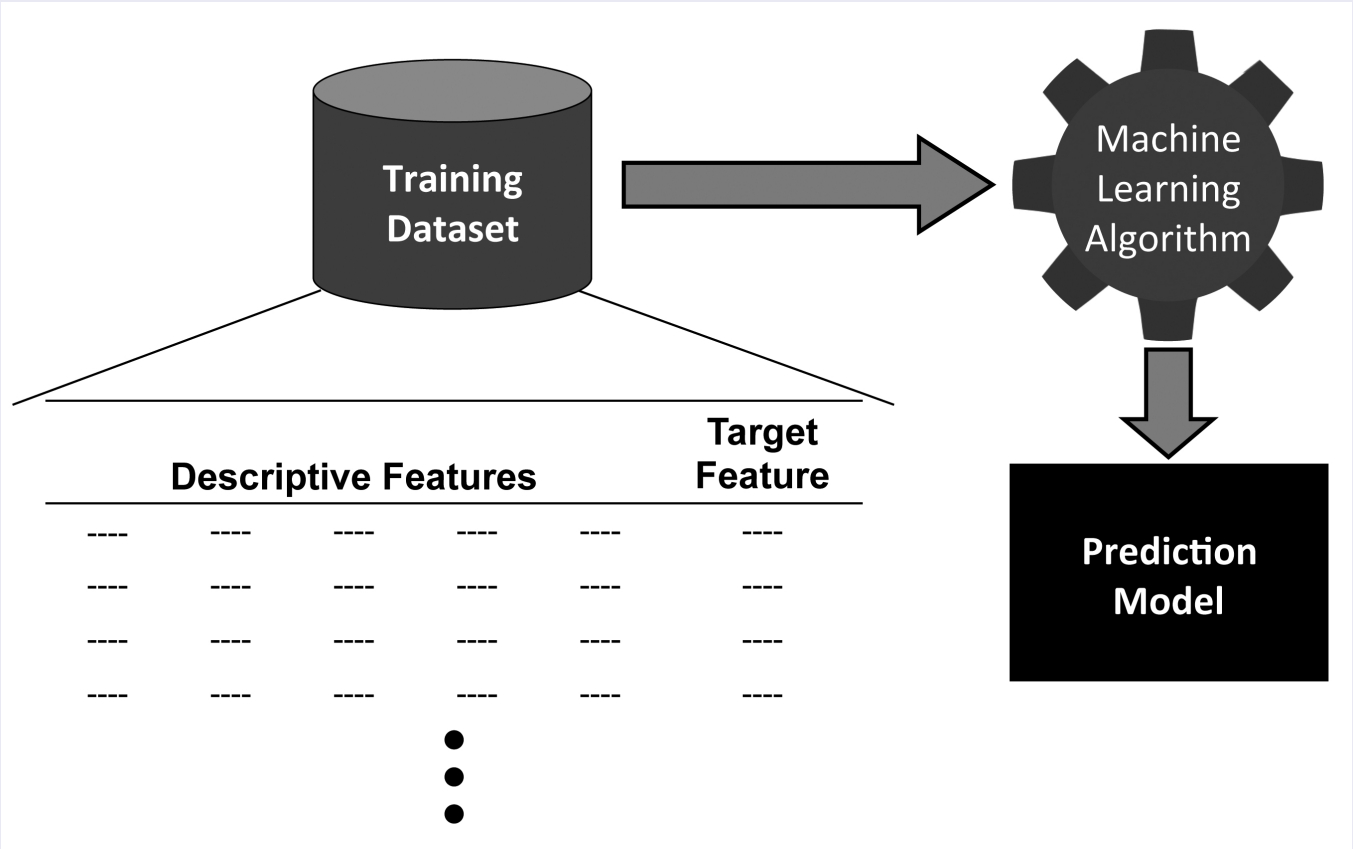
There are many ways to do this... here are some:

```
> results[1,5]
[1] A
Levels: D C B A
> results2[1,2]
[1] A
Levels: D C B A
> results$grade[1]
[1] A
Levels: D C B A
> results2$results.grade[1]
[1] A
Levels: D C B A
```

Review of last week

Step 1:

Supervised machine learning techniques automatically learn the relationship between a set of **descriptive features** and a **target feature** from a set of historical **instances** (referred to as a **training dataset**) to build a **prediction model**.



Step 2:

We can then use this **prediction model** to make predictions for new instances

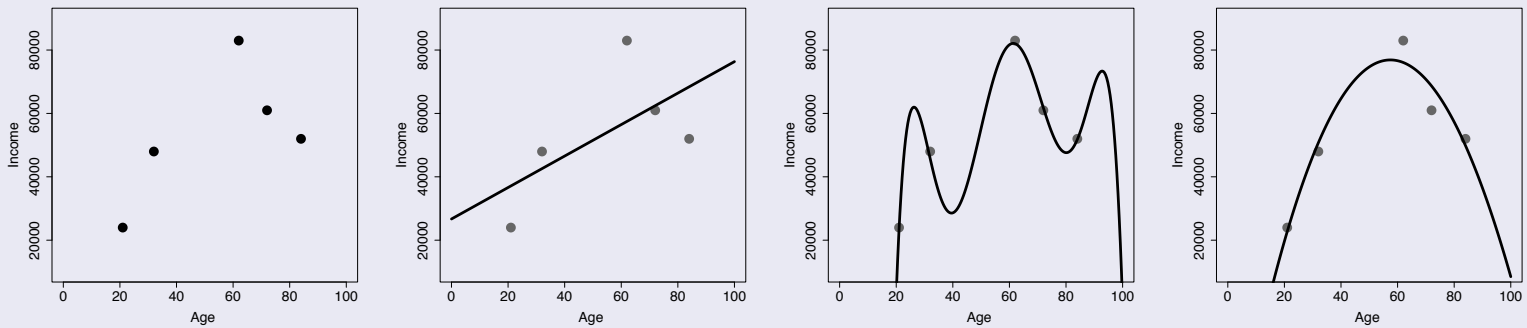
- A prediction model that makes the correct predictions for these queries is said to **generalise** well.
- The goal of machine learning is to find the predictive model that **generalises** best.
- To find the best prediction model, a machine learning algorithm must use some criteria for choosing among the candidate prediction models it considers during its search.

What criteria should we use?

- Lots of different machine learning algorithms.
- Each machine learning algorithm uses different model selection criteria to drive its search for the best **predictive model**.
- The set of assumptions that defines the model selection criteria of a machine learning algorithm is known as the **inductive bias** of the machine learning algorithm.
- It has been shown that there is no particular **inductive bias** that on average is the best one to use.
- **The ability to select the appropriate machine learning algorithm (and hence inductive bias) to use for a given predictive task is one of the core skills that a data analyst must develop!!**

What happens if we choose the wrong inductive bias:

- **Underfitting**
- **Overfitting**
- Striking the right balance between **model** simplicity and complexity (between underfitting and overfitting) is the hardest part of machine learning.



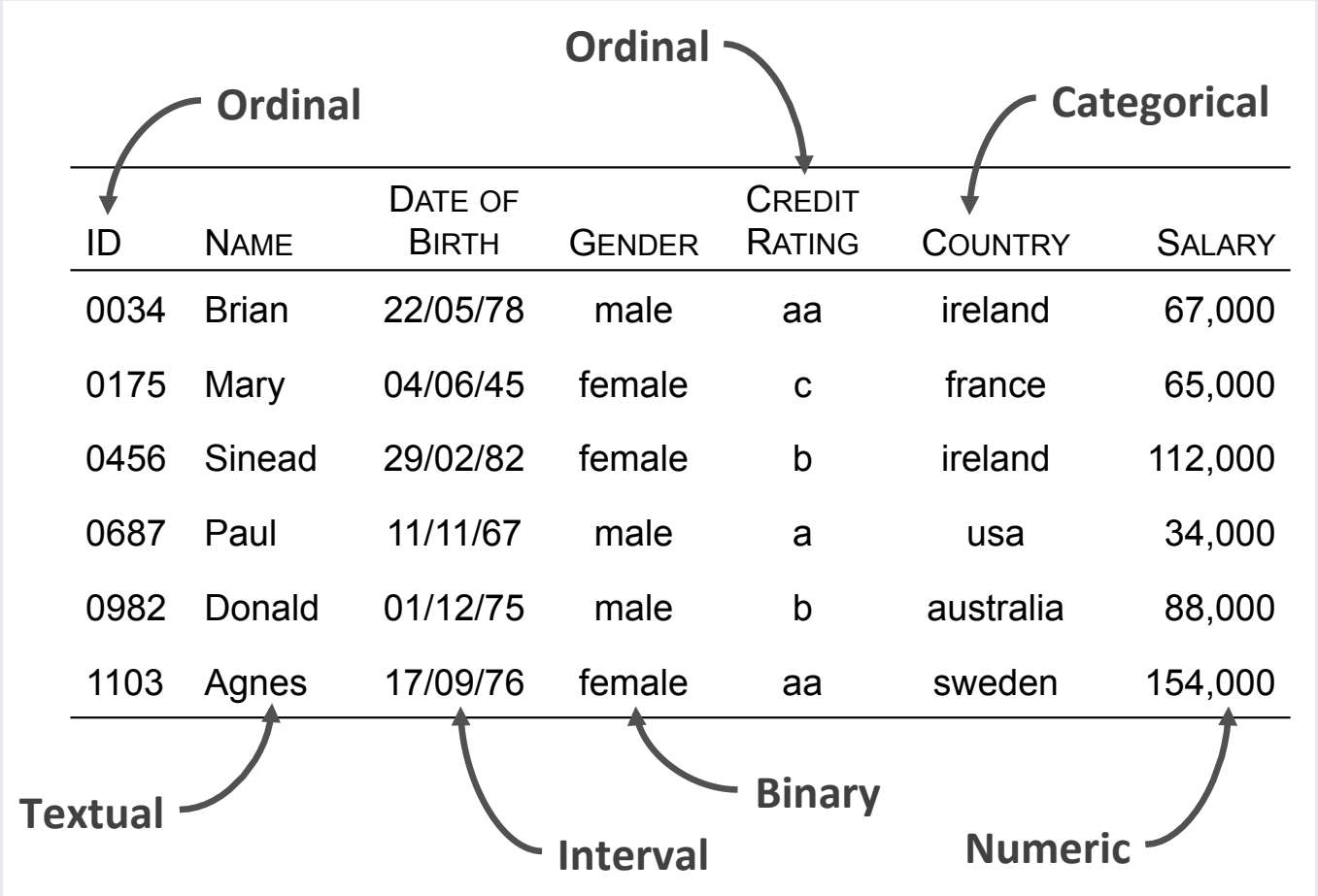
(a) Dataset (b) Underfitting (c) Overfitting (d) Just right

The Analytics Base Table

- A simple, flat, tabular data structure made up of rows and columns.
- The columns are divided into a set of descriptive features and a single target feature.
- Each row contains a value for each descriptive feature and the target feature.
- Each row represents an instance about which a prediction can be made.

Descriptive Features						Target Feature
----	----	----	----	----	----	----
----	----	----	----	----	----	----
----	----	----	----	----	----	----
----	----	----	----	----	----	----
----	----	----	----	----	----	----

Different Types of Data



Different Types of Data

- **Numeric**: True numeric values that allow arithmetic operations (e.g., price, age)
- **Interval**: Values that allow ordering and subtraction, but do not allow other arithmetic operations (e.g., date, time)
- **Ordinal**: Values that allow ordering but do not permit arithmetic (e.g., size measured as small, medium, or large)
- **Categorical**: A finite set of values that cannot be ordered and allow no arithmetic (e.g., country, product type)
- **Binary**: A set of just two values (e.g., gender)
- **Textual**: Free-form, usually short, text data (e.g., name, address)

We often reduce this categorization to just two data types:

- **Continuous** (encompassing the numeric and interval types)
- **Categorical** (encompassing the categorical, ordinal, binary, and textual types)
  - We refer to the set of possible values that a categorical feature can take as the **levels** of the feature
  - For example, the levels of the CREDIT RATING feature are aa, a, b, c and the levels of the GENDER feature are male, female.

The presence of different types of descriptive and target features can have a big impact on how an algorithm works.

Different Types of Features

- The features in an ABT can be of two types:
  - **Raw features** – features that come directly from raw data sources (e.g. customer age, gender, loan amount)
  - **Derived features** – do not exist in any raw data source, so they must be constructed from data in one or more raw data sources (e.g. average customer purchases per month, loan-to-value ratios)

Legal Issues

- Data analytics practitioners can often be frustrated by legislation that stops them from including features that appear to be particularly well suited to an analytics solution in an ABT.
- There are significant differences in legislation in different jurisdictions, but a couple of key relevant principles almost always apply.
  1. **Anti-discrimination legislation** in most jurisdictions prohibits discrimination on the basis of some set of the following grounds: sex, age, race, ethnicity, nationality, sexual orientation, religion, disability, and political opinions.
  2. **Data protection legislation** – rules surrounding the use of personal data.

Data Quality Issue

- A **data quality issue** is loosely defined as anything *unusual* about the data in an ABT.
- The most common data quality issues are:
  - **missing values**
  - **irregular cardinality**
  - **outliers**

The data quality report

- A data quality report includes tabular reports that describe the characteristics of each feature in an ABT using standard statistical measures of **central tendency** (mean, mode, and median) and **variation** (standard deviation and percentiles).
- The tabular reports are accompanied by data visualizations:
  - A **histogram** for each continuous feature in an ABT.
  - A **bar plot** for each categorical feature in an ABT.

The data quality report – continuous features

Should include:

- the total number of instances in the ABT
- the percentage of instances in the ABT that are missing a value for each feature
- the cardinality of each feature (cardinality measures the number of distinct values present in the ABT for a feature)
- the minimum, 1st quartile, mean, median, 3rd quartile, maximum, and standard deviation statistics

Feature	Count	% Miss.	Card.	Min.	1^{st} Qrt.	Mean	Median	3^{rd} Qrt.	Max.	Std. Dev.
___	___	___	___	___	___	___	___	___	___	___
___	___	___	___	___	___	___	___	___	___	___
___	___	___	___	___	___	___	___	___	___	___

The data quality report – categorical features

Should include:

- the total number of instances in the ABT
- the percentage of instances in the ABT that are missing a value for each feature
- the cardinality of each feature
- for the two most frequent levels for the feature (the mode and 2nd mode) – the frequency with which these appear (both as raw frequencies and as a proportion of the total number of instances in the dataset)

Feature	Count	% Miss.	Card.	Mode	Mode Freq.	Mode %	2^{nd} Mode	2^{nd} Mode Freq.	2^{nd} Mode %
___	___	___	___	___	___	___	___	___	___
___	___	___	___	___	___	___	___	___	___
___	___	___	___	___	___	___	___	___	___

- The data quality issues we identify from a data quality report will be of two types:
  - Data quality issues due to **invalid data** – take immediate action to correct them, regenerate the ABT, and recreate the data quality report.
  - Data quality issues due to **valid data** – record any data quality issues due to valid data in a data quality plan so that we remain aware of them and can handle them later if required.

The structure of a data quality plan

Feature	Data Quality Issue	Potential Handling Strategies
____	____	____
____	____	____
____	____	____
____	____	____

For each of the data quality issues found:

- include the feature it was found in
- the details of the data quality issue
- information on potential handling strategies

Handling Data Quality Issues

- Handling Missing Values
- Handling Irregular Cardinality
- Handling Outliers

- The key outcomes of the **data exploration** process are that the practitioner should
  1. Have *gotten to know* the features within the ABT, especially their central tendencies, variations, and **distributions**.
  2. Have identified any **data quality issues** within the ABT, in particular **missing values**, **irregular cardinality**, and **outliers**.
  3. Have corrected any data quality issues due to **invalid data**.
  4. Have recorded any data quality issues due to **valid data** in a **data quality plan** along with potential handling strategies.
  5. Be confident that enough good quality data exists to continue with a project.

Any questions on what we covered last week?

Advanced Data Exploration

Visualizing Relationships Between Features

Relationships Between Features

- So far we have focused on the characteristics of individual features.
- Now we will look at techniques that enable us to examine relationships between pairs of features.
- This can help identify which features might be useful for predicting a target feature.
- Can help find pairs of descriptive features that are closely related.
- If the relationship between two descriptive features is strong enough, we may not need to include both.

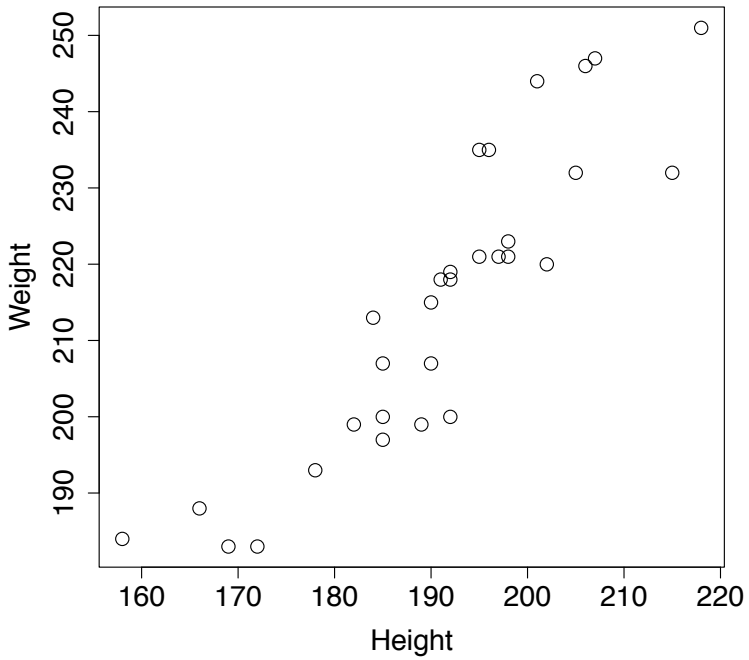
Professional Basketball Team Dataset

- HEIGHT, WEIGHT, and AGE of each player.
- The POSITION that the player normally plays (guard, center, or forward).
- The CAREER STAGE of the player (rookie, mid-career, or veteran).
- The average weekly SPONSORSHIP EARNINGS of each player.
- If the player has a SHOE SPONSOR (yes or no).

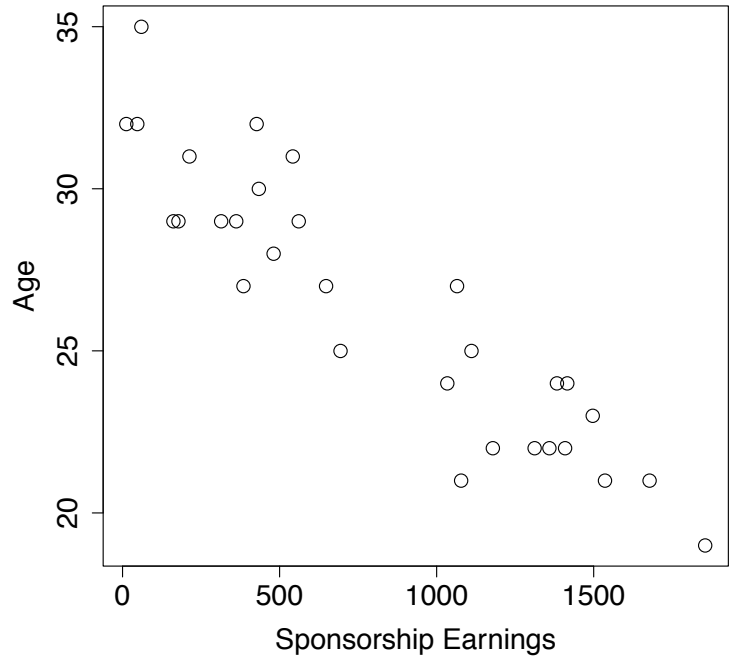
ID	Position	Height	Weight	Career Stage	Age	Sponsorship Earnings	Shoe Sponsor
1	forward	192	218	veteran	29	561	yes
2	center	218	251	mid-career	35	60	no
3	forward	197	221	rookie	22	1,312	no
4	forward	192	219	rookie	22	1,359	no
5	forward	198	223	veteran	29	362	yes
6	guard	166	188	rookie	21	1,536	yes
7	forward	195	221	veteran	25	694	no
8	guard	182	199	rookie	21	1,678	yes
9	guard	189	199	mid-career	27	385	yes
10	forward	205	232	rookie	24	1,416	no
11	center	206	246	mid-career	29	314	no
12	guard	185	207	rookie	23	1,497	yes
13	guard	172	183	rookie	24	1,383	yes
14	guard	169	183	rookie	24	1,034	yes
15	guard	185	197	mid-career	29	178	yes
16	forward	215	232	mid-career	30	434	no
17	guard	158	184	veteran	29	162	yes
18	guard	190	207	mid-career	27	648	yes
19	center	195	235	mid-career	28	481	no
20	guard	192	200	mid-career	32	427	yes
21	forward	202	220	mid-career	31	542	no
22	forward	184	213	mid-career	32	12	no
23	forward	190	215	rookie	22	1,179	no
24	guard	178	193	rookie	21	1,078	no
25	guard	185	200	mid-career	31	213	yes
26	forward	191	218	rookie	19	1,855	no
27	center	196	235	veteran	32	47	no
28	forward	198	221	rookie	22	1,409	no
29	center	207	247	veteran	27	1,065	no
30	center	201	244	mid-career	25	1,111	yes

Visualizing Pairs of Continuous Features

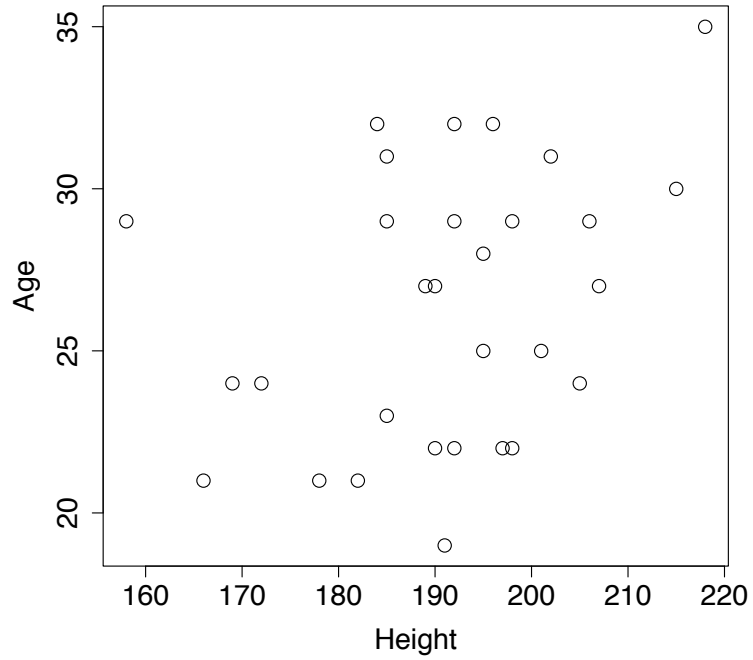
- A **scatter plot** is based on two axes: the horizontal axis represents one feature and the vertical axis represents a second.
- Each instance in a dataset is represented by a point on the plot determined by the values for that instance of the two features involved.



- Scatter plot showing the relationship between the HEIGHT and WEIGHT features.
- Broadly linear pattern diagonally across the scatter plot.
- This suggests that there is a strong, positive, linear relationship between the HEIGHT and WEIGHT features
- We say that features with this kind of relationship are **positively covariant**.

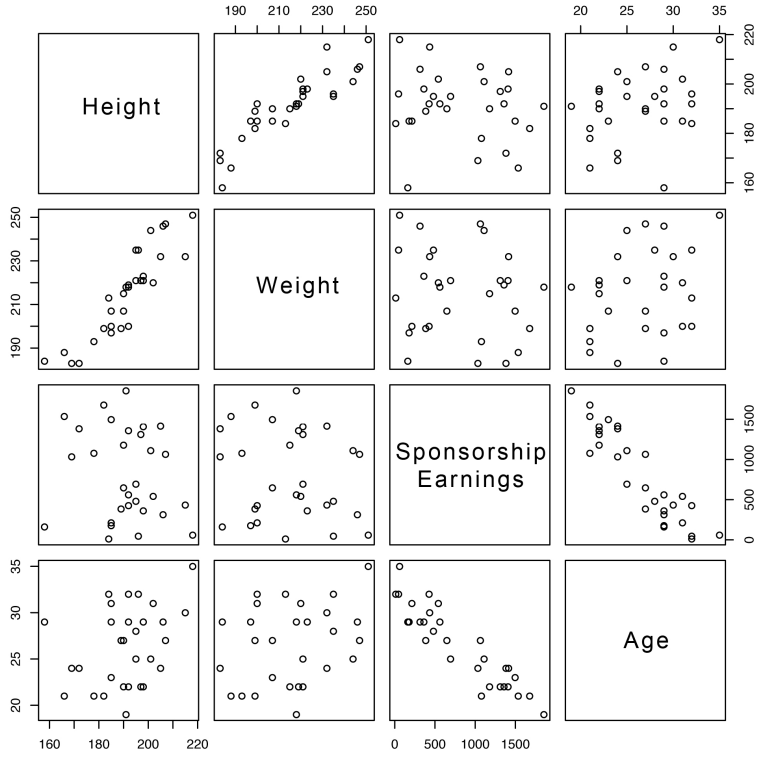


- Scatter plot showing the relationship between the SPONSORSHIP EARNINGS and AGE features.
- We say that features with this kind of relationship are **negatively covariant**.



- Scatter plot showing the relationship between the HEIGHT and AGE
- These features are not strongly covariant either positively or negatively.

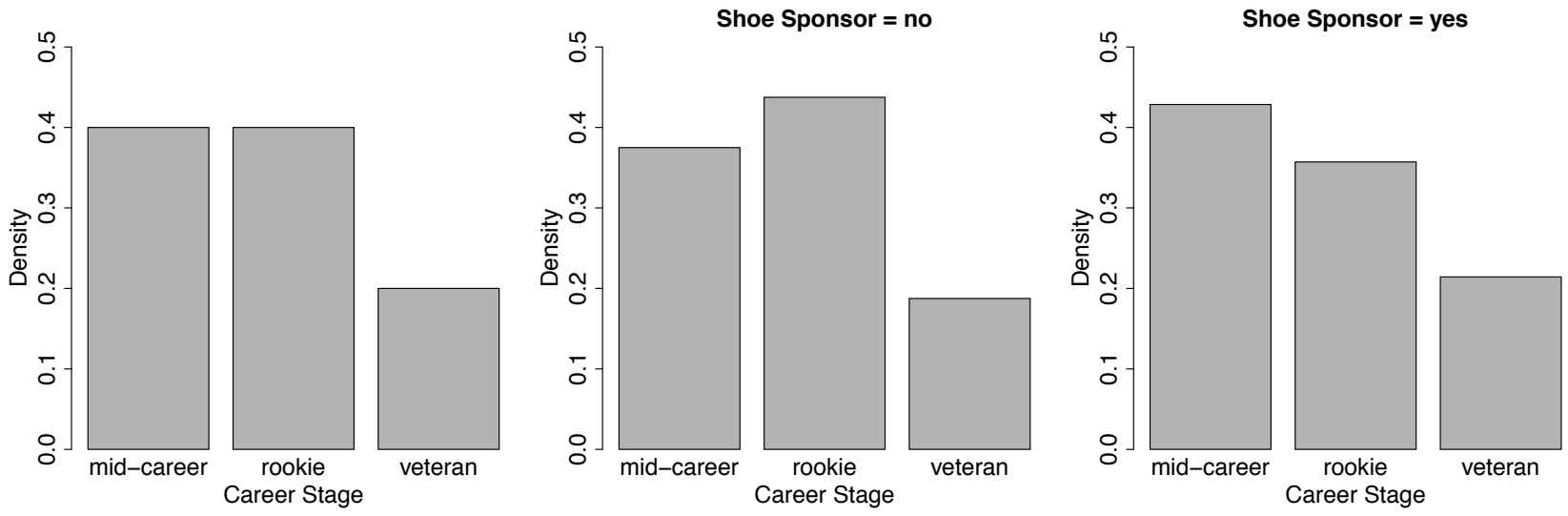
- A **scatter plot matrix** (**SPLOM**) shows scatter plots for a whole collection of features arranged into a matrix.
- This is useful for exploring the relationships between groups of features - for example all of the continuous features in an ABT.
- However, once the number of features in the set goes beyond 8 the graphs become too small.



- A **scatter plot matrix** (**SPLOM**).
- Each row and column represent the feature named in the cells along the diagonal.
- The cells above and below the diagonal show scatter plots of the features in the row and column that meet at that cell.

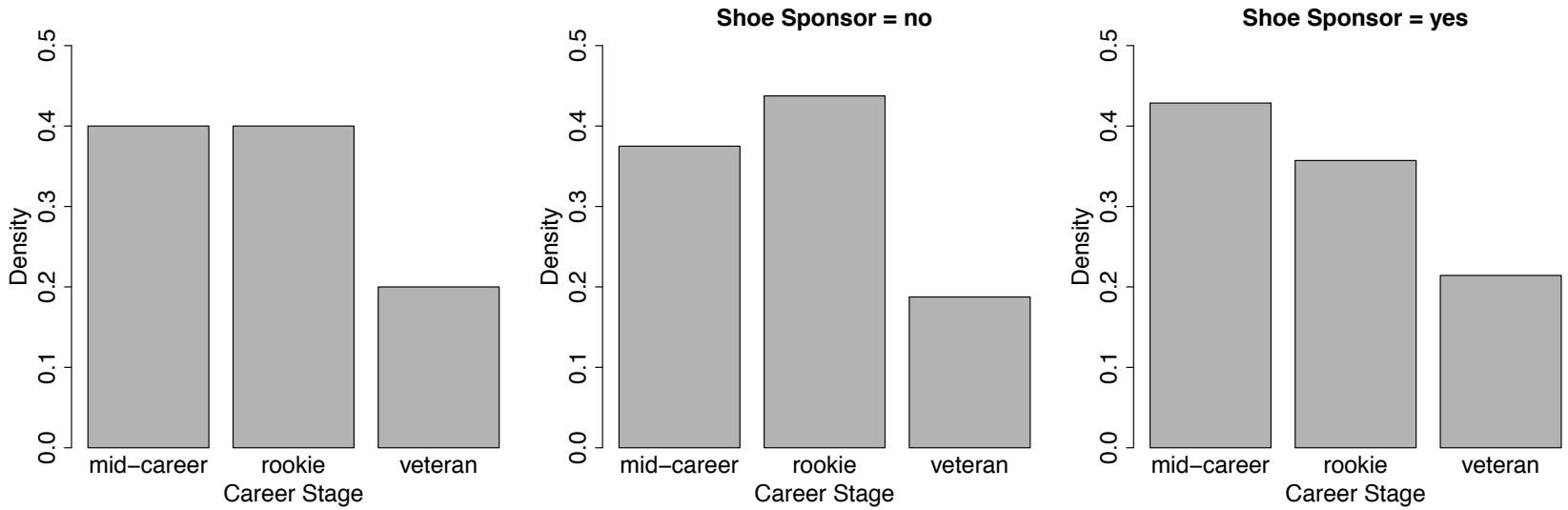
Visualizing Pairs of Categorical Features

- The simplest way to visualize the relationship between two categorical variables is to use a collection of **small multiple** bar plots.
- First, we draw a simple bar plot showing the densities of the different levels of the first feature.
- Then, for each level of the second feature, we draw a bar plot of the first feature using only the instances in the dataset for which the second feature has that level.

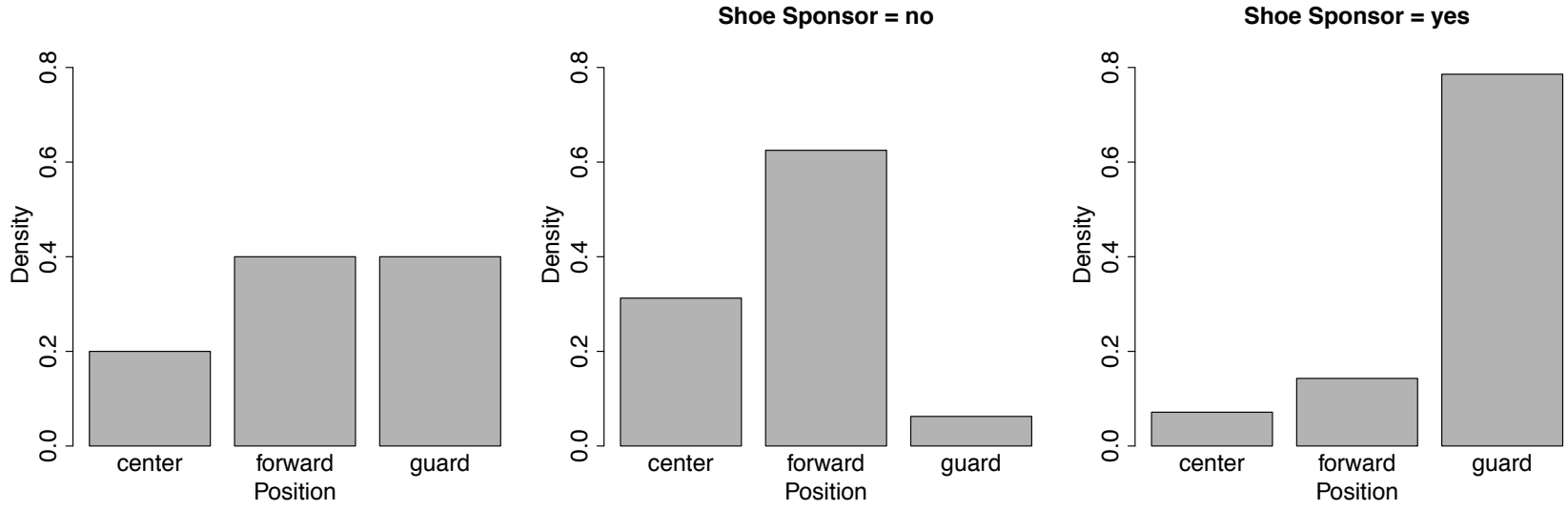


Using small multiple bar plot visualizations to illustrate the relationship between the CAREER STAGE and SHOE SPONSOR features.

- If the two features being visualized have a strong relationship, then the bar plots for each level of the second feature will look noticeably different to one another and to the overall bar plot for the first feature.
- If there is no relationship, then the levels of the first feature will be evenly distributed amongst the instances having the different levels of the second feature, so all bar plots will look much the same.



- The bar plot on the left shows the distribution of the different levels of the CAREER STAGE feature across the entire dataset.
- The two plots on the right show the distributions for those players with and without a SHOE SPONSOR.
- Since all three plots show similar distributions, we can conclude that no real relationship exists between these two features

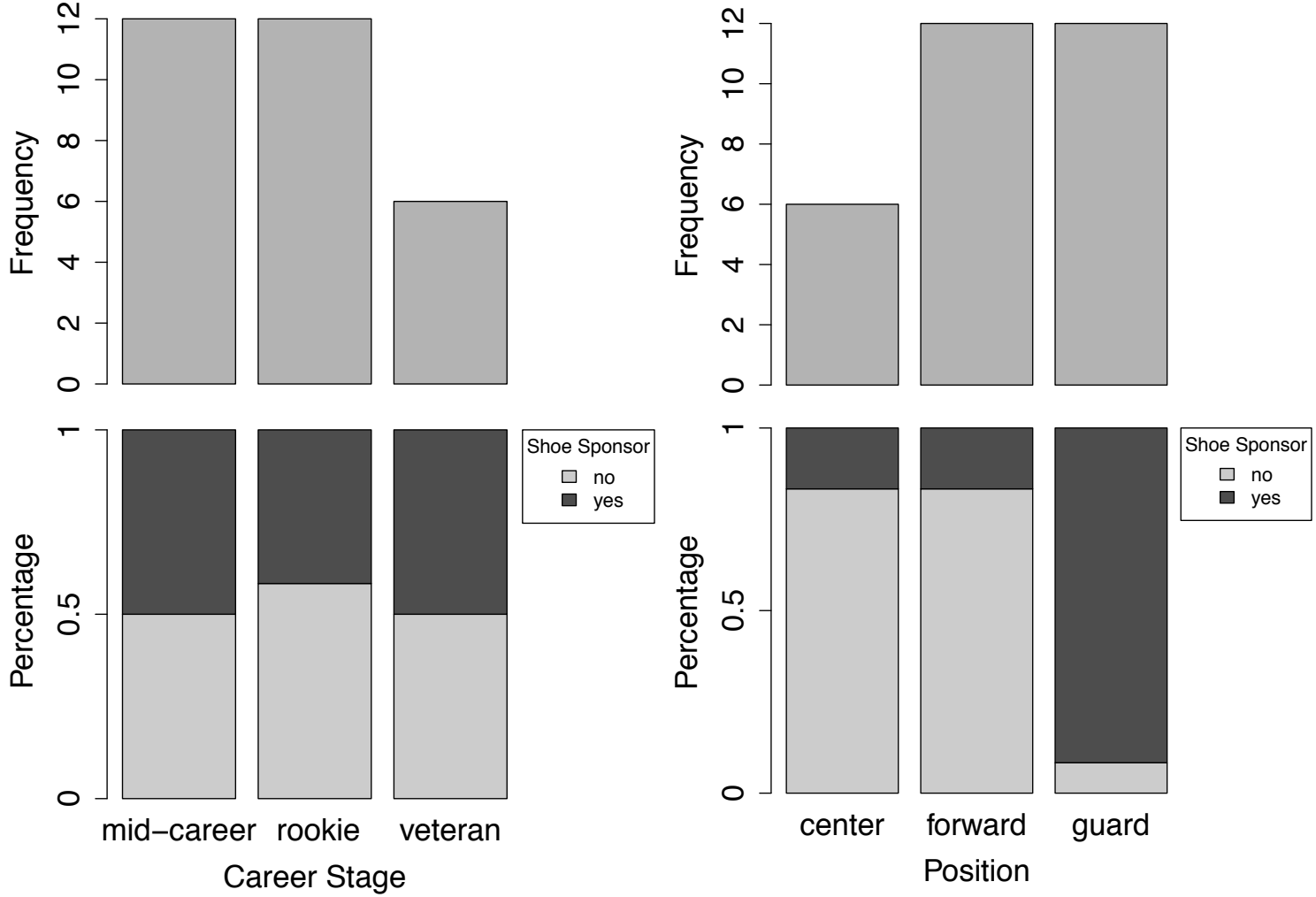


- Using small multiple bar plot visualizations to illustrate the relationship between the POSITION and SHOE SPONSOR features.
- In this case, the three plots are very different, so we can conclude that there is a relationship between these two features.
- Players who play in the guard position are much more likely to have a shoe sponsor than forwards or centers.

Important

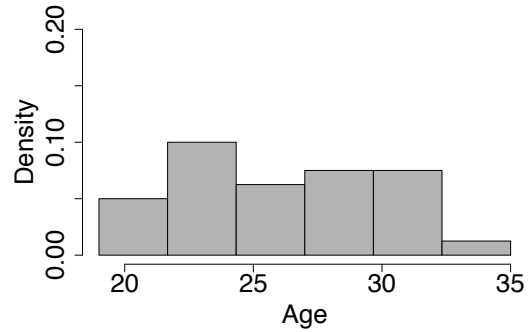
- It is important that all the charts are kept consistent as this ensures that only genuine differences within the data are highlighted, rather than differences that arise from formatting.
- The scales of the axes must always be kept consistent, as should the order of the bars in the individual bar plots.
- Densities are shown rather than frequencies as the overall bar plots on the left of each visualization cover much more of the dataset than the other two plots, so frequency-based plots would look very uneven.

- If the number of levels of one of the features being compared is no more than three we can use **stacked bar plots** as an alternative to the small multiples approach.

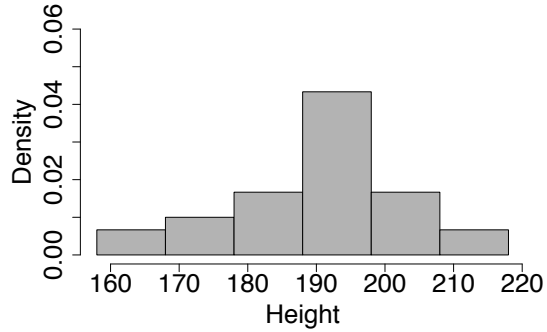


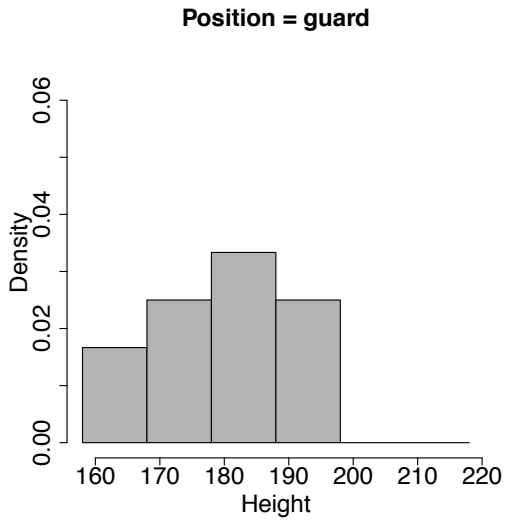
Stacked bar plot visualizations. If two features are unrelated, we expect to see the same proportion of each level of the second feature within the bars for each level of the first.

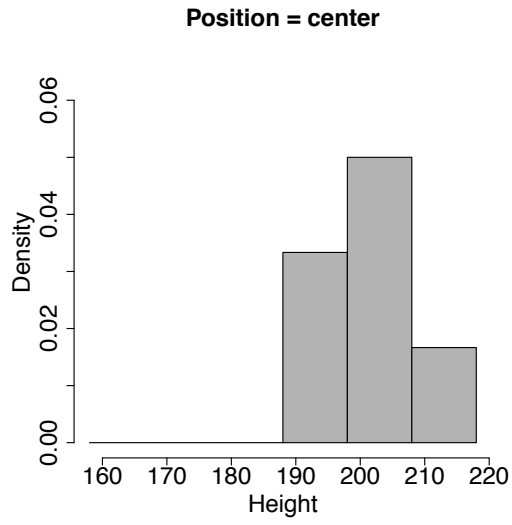
- To visualize the relationship between a continuous feature and a categorical feature a **small multiples** approach that draws a histogram of the values of the continuous feature for each level of the categorical feature is useful.

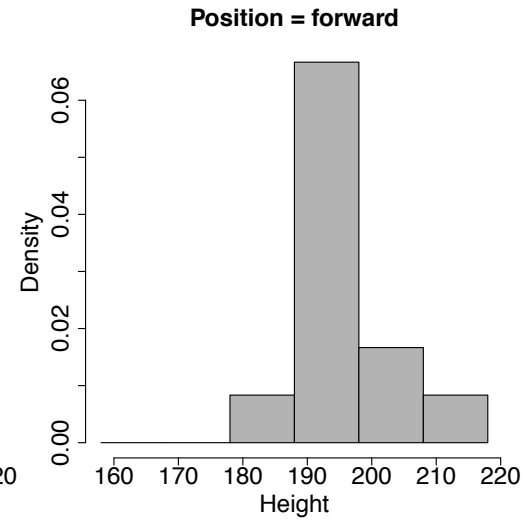


- Using small multiple histograms to visualize the relationship between the AGE feature and the POSITION FEATURE.
- A slight tendency for centers to be a little older than guards and forwards





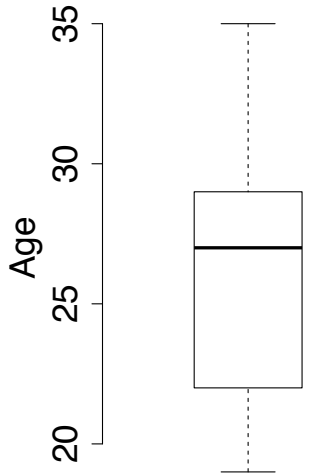




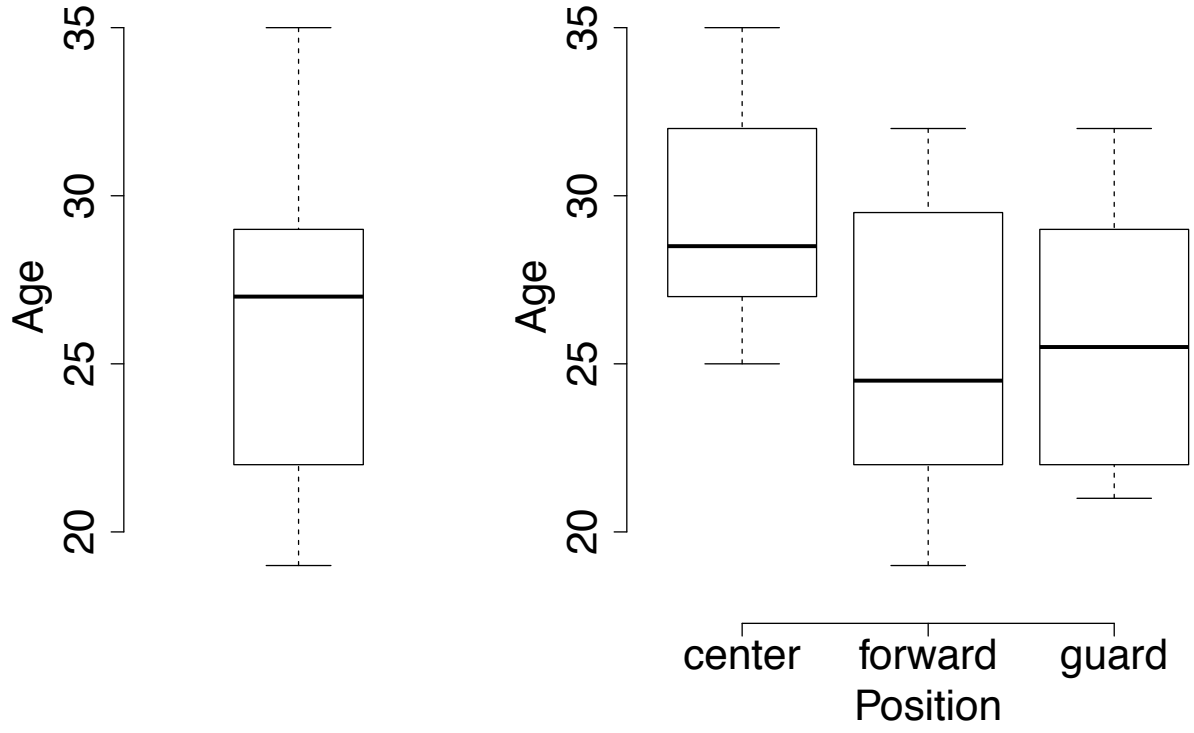
- Visualising the relationship between the HEIGHT feature and the POSITION feature.
- HEIGHT follows a normal distribution.
- The three smaller histograms depart from this distribution and suggest that centers tend to be taller than forwards, who in turn tend to be taller than guards.

Box Plots

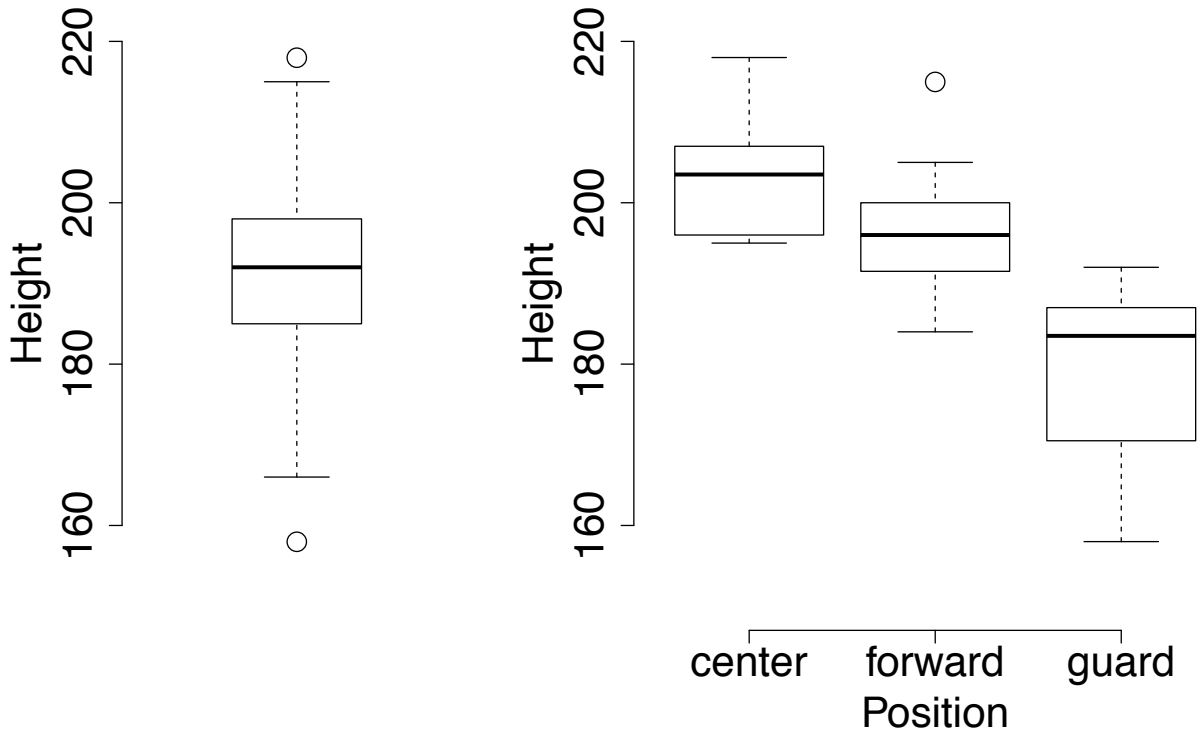
- A second approach to visualizing the relationship between a categorical feature and a continuous feature is to use a collection of box plots.
- A box plot is a visual representation of the five key descriptive statistics for a continuous feature: minimum, 1^{st} quartile, median, 3^{rd} quartile, and maximum.
- For each level of the categorical feature a box plot of the corresponding values of the continuous feature is drawn.



- The vertical axis shows the range of values.
- The rectangular box is determined by the 3^{rd} quartile at the top and the 1^{st} quartile at the bottom.
- The height of the rectangle shows the inter-quartile range.
- The strong black line across the middle of the rectangle shows the median.



- Using box plots to visualize the relationship between the AGE and the POSITION feature.
- Easy comparison of the central tendency and variation



- The relationship between the HEIGHT feature and the POSITION feature.
- When a relationship exists between the two features, the box plots should show differing central tendencies and variations.

Any questions so far?
5 min break...

Measuring Covariance & Correlation

- As well as visually inspecting scatter plots, we can calculate formal measures of the relationship between two continuous features using **covariance** and **correlation**.
- For two features, a and b, in a dataset of n instances, the **sample covariance** between a and b is

$$cov(a, b) = \frac{1}{n-1} \sum_{i=1}^{n} \left((a_i - \overline{a}) \times \left(b_i - \overline{b} \right) \right) \quad (1)$$

where a_i and b_i are values of features a and b for the i^{th} instance in a dataset, and $\overline{a}$ and $\overline{b}$ are the sample means of features a and b.

- Covariance values fall into the range $[-\infty, \infty]$ where negative values indicate a negative relationship, positive values indicate a positive relationship, and values near zero indicate that there is little or no relationship between the features.

Calculating covariance between the HEIGHT feature and the WEIGHT and AGE features from the basketball players dataset. The table shows how the $\left((a_i - \overline{a}) \times \left(b_i - \overline{b}\right)\right)$ portion of the equation is calculated for each instance in a dataset to arrive at the sum required in the calculation.

ID	HEIGHT (h)	$h - \overline{h}$	WEIGHT (w)	$w - \overline{w}$	$(h - \overline{h}) \times (w - \overline{w})$	AGE (a)	$a - \overline{a}$	$(h - \overline{h}) \times (a - \overline{a})$
1	192	0.9	218	3.0	2.7	29	2.6	2.3
2	218	26.9	251	36.0	967.5	35	8.6	231.3
3	197	5.9	221	6.0	35.2	22	-4.4	-26.0
4	192	0.9	219	4.0	3.6	22	-4.4	-4.0
5	198	6.9	223	8.0	55.0	29	2.6	17.9
				...				
26	191	-0.1	218	3.0	-0.3	19	-7.4	0.7
27	196	4.9	235	20.0	97.8	32	5.6	27.4
28	198	6.9	221	6.0	41.2	22	-4.4	-30.4
29	207	15.9	247	32.0	508.3	27	0.6	9.5
30	201	9.9	244	29.0	286.8	25	-1.4	-13.9
Mean	191.1		215.0			26.4		
Std Dev	13.6		19.8			4.2		
Sum					7,009.9			570.8

Calculating covariance between the HEIGHT feature and the WEIGHT and AGE features from the basketball players dataset.

$$cov(\text{HEIGHT}, \text{WEIGHT}) = \frac{7{,}009.9}{29} = 241.72$$

$$cov(\text{HEIGHT}, \text{AGE}) = \frac{570.8}{29} = 19.7$$

- These figures indicate that there is a strong positive relationship between the height and weight of a player, and a much smaller positive relationship between height and age.
- This example also illustrates a problem with using covariance.
- Covariance is measured in the same units as the features that it measures.
- As a result, comparing the covariance between pairs of features only makes sense if each pair of features is composed of the same mixture of units.

- **Correlation** is a normalized form of covariance that ranges between -1 and $+1$.
- The correlation between two features, a and b, can be calculated as

$$corr(a, b) = \frac{cov(a, b)}{sd(a) \times sd(b)} \tag{2}$$

where $cov(a, b)$ is the covariance between features a and b and $sd(a)$ and $sd(b)$ are the standard deviations of a and b respectively.

- Correlation values fall into the range $[-1, 1]$, where values close to -1 indicate a very strong negative correlation (or covariance), values close to 1 indicate a very strong positive correlation, and values around 0 indicate no correlation.
- Because correlation is normalized, it is dimensionless and, consequently, does not suffer from the interpretability difficulties associated with covariance.
- Features that have no correlation are said to be **independent**.

Calculating correlation between the HEIGHT feature and the WEIGHT and AGE features from the basketball players dataset.

$$corr(Height, Weight) = \frac{241.72}{13.6 \times 19.8} = 0.898$$

$$corr(Height, Age) = \frac{19.7}{13.6 \times 4.2} = 0.345$$

- These correlation values are much more useful than the covariances calculated previously
- They are on a normalized scale, which allows us compare the strength of the relationships to each other.
- There is a strong positive correlation between HEIGHT and WEIGHT features, but very little correlation between HEIGHT and AGE.

- In the majority of ABTs there are multiple continuous features between which we would like to explore relationships.
- Two tools that can be useful for this are the covariance matrix and the correlation matrix.

- The covariance matrix, usually denoted as $\sum$, between a set of continuous features, $\{a, b, \ldots, z\}$, is given as

$$\sum_{\{a,b,\ldots,z\}} = \begin{bmatrix} var(a) & cov(a,b) & \cdots & cov(a,z) \\ cov(b,a) & var(b) & \cdots & cov(b,z) \\ \vdots & \vdots & \ddots & \vdots \\ cov(z,a) & cov(z,b) & \cdots & var(z) \end{bmatrix} \tag{3}$$

- Similarly, the **correlation matrix** is just a normalized version of the covariance matrix and shows the correlation between each pair of features:

$$\underset{\{a,b,\ldots,z\}}{\textit{correlation matrix}} = \begin{bmatrix} corr(a,a) & corr(a,b) & \cdots & corr(a,z) \\ corr(b,a) & corr(b,b) & \cdots & corr(b,z) \\ \vdots & \vdots & \ddots & \vdots \\ corr(z,a) & corr(z,b) & \cdots & corr(z,z) \end{bmatrix} \quad (4)$$

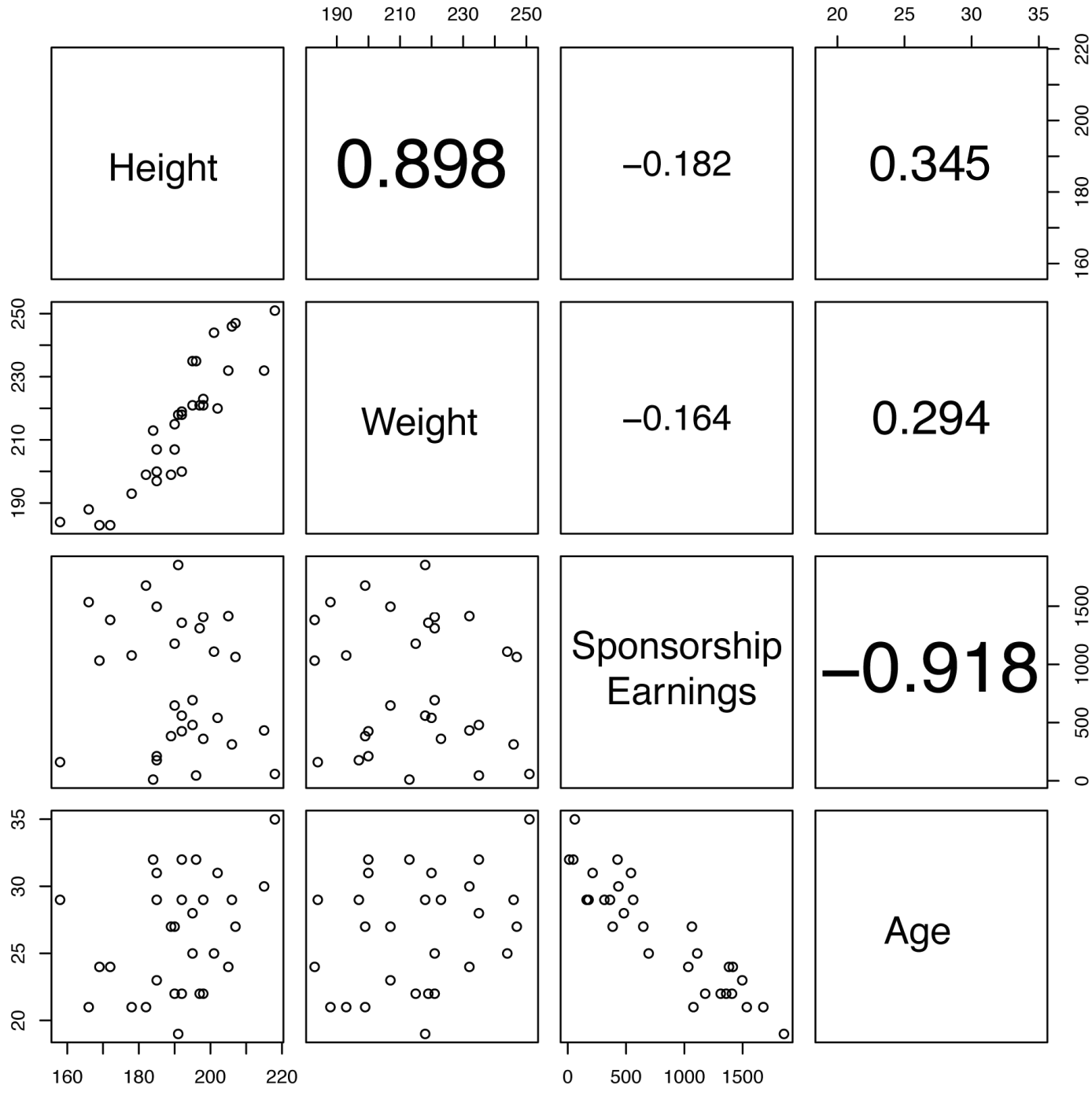
- Calculating covariances matrix for the HEIGHT feature and the WEIGHT and AGE features from the basketball players dataset.

$$\sum_{<Height,Weight,Age>} = \begin{bmatrix} 185.128 & 241.72 & 19.7 \\ 241.72 & 392.102 & 24.469 \\ 19.7 & 24.469 & 17.697 \end{bmatrix}$$

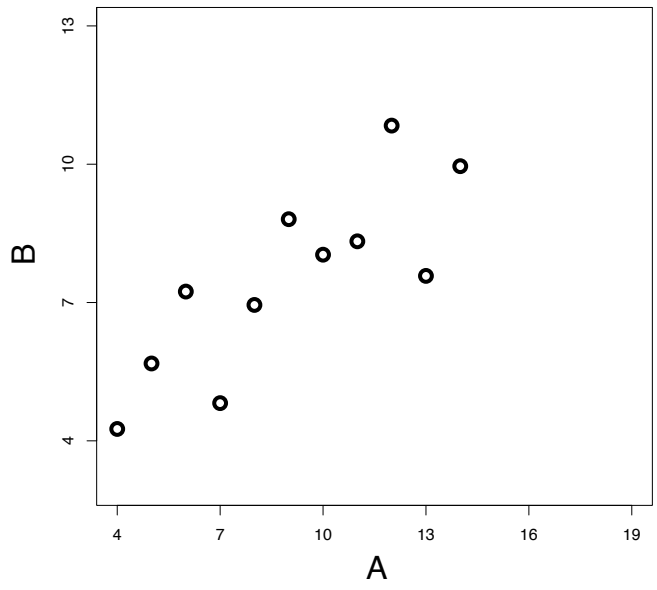
- Calculating correlation matrix for the HEIGHT feature and the WEIGHT and AGE features from the basketball players dataset.

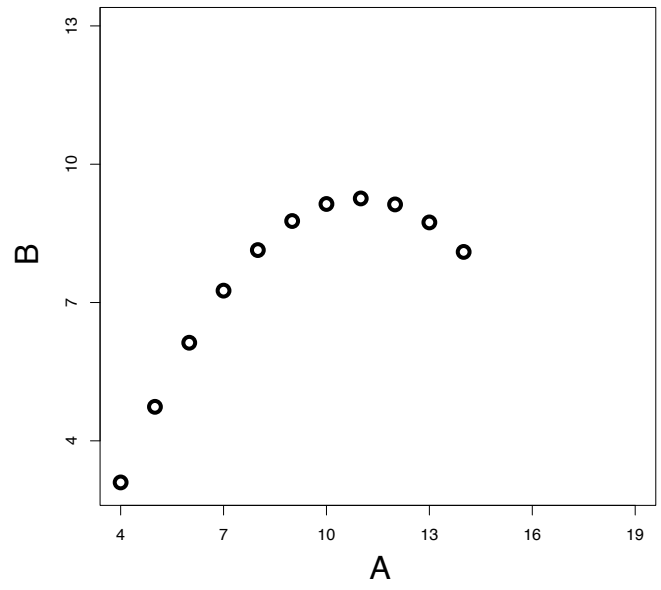
$$\underset{<Height, Weight, Age>}{\text{correlation matrix}} = \begin{bmatrix} 1.0 & 0.898 & 0.345 \\ 0.898 & 1.0 & 0.294 \\ 0.345 & 0.294 & 1.0 \end{bmatrix}$$

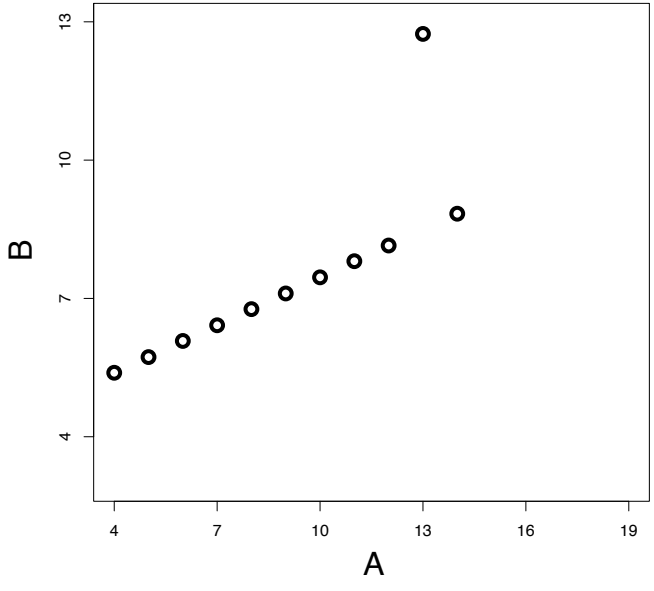
- The **scatter plot matrix** (SPLOM) is really a visualization of the correlation matrix.
- This can be made more obvious by including the correlation coefficients in SPLOMs in the cells above the diagonal.
- The font sizes of the correlation coefficients are scaled according to the absolute value of the strength of the correlation to draw attention to those pairs of features with the strongest relationships.

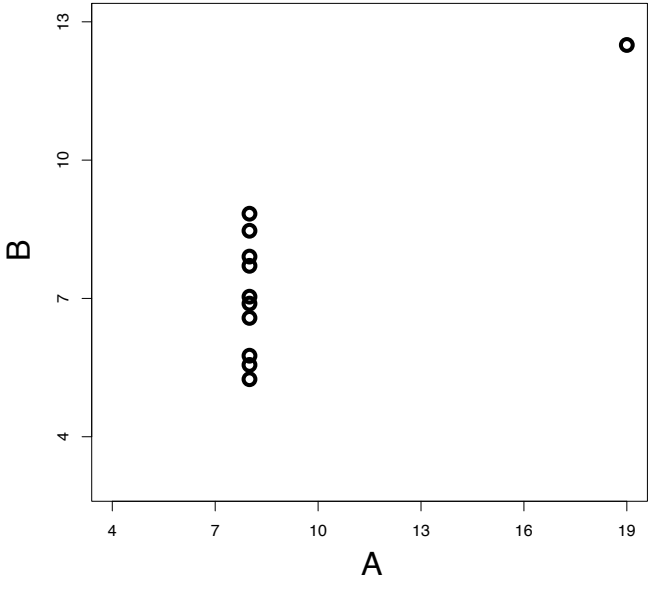

Height
0.898
−0.182
0.345
Weight
−0.164
0.294
Sponsorship Earnings
−0.918
Age
190 210 230 250
20 25 30 35
160 180 200 220
190 210 230 250
20 25 30 35
0 500 1000 1500
160 180 200 220
0 500 1000 1500

- Correlation is a good measure of the relationship between two continuous features, but it is not by any means perfect.
- Some of the limitations of measuring correlation are illustrated very clearly in the famous example of **Anscombe's quartet** by **Francis Anscombe**.
- This is a series of four pairs of features that all have the same correlation value of 0.816, even though they exhibit very different relationships.


B
A
4
7
10
13
16
19


B
A
4
7
10
13
16
19


B
A
4
7
10
13
16
19


B
A
4
7
10
13
16
19

- Perhaps the most important thing to remember in relation to correlation is that correlation does not necessarily imply causation.

- There are two main ways in which causation can be mistakenly assumed.
- The first is by mistaking the order of a causal relationship.
- For example, based on correlations tests alone, we might conclude that the presence of swallows cause hot weather, that spinning windmills cause wind, and that playing basketball causes people to be tall.
- In fact, swallows migrate to warmer countries, windmills are made to spin by wind, and tall people often choose to play basketball because of the advantage their height gives them in that game.

- The second kind of mistake that makes people incorrectly infer causation between two features is ignoring a third important, but hidden, feature.
- Read the story in the book about a causal relationship between young children sleeping with a night-light turned on and these children developing short-sightedness in later life.

Data Preparation

- Some data preparation techniques change the way data is represented just to make it more compatible with certain machine learning algorithms.
  - Normalization
  - Binning
  - Sampling

Normalization

- Having continuous features that cover very different ranges can cause difficulty for some machine learning algorithms.
- For example, a feature representing customer ages might cover the range [16, 96], whereas a feature representing customer salaries might cover the range [10,000, 100,000].
- **Normalization** techniques can be used to change a continuous feature to fall within a specified range while maintaining the relative differences between the values for the feature.

- We use **range normalization** to convert a feature value into the range [*low*, *high*] as follows:

$$a_i' = \frac{a_i - min(a)}{max(a) - min(a)} \times (high - low) + low \tag{5}$$

- Where a_i' is the normalized feature value, a_i is the original value, *min*(a) is the minimum value of feature a, *max*(a) is the maximum value of feature a, and low and high are the minimum and maximum values of the desired range.
- Typical ranges used for normalizing feature values are [0,1] and [-1,1].
- Drawback – quite sensitive to the presence of outliers.

- Another way to normalize data is to **standardize** it into **standard scores**.
- A standard score measures how many standard deviations a feature value is from the mean for that feature.
- We calculate a standard score as follows:

$$a_i' = \frac{a_i - \bar{a}}{sd(a)} \tag{6}$$

- Squashes the values of the feature so that the feature values have a mean of 0 and a standard deviation of 1.
- This results in the majority of feature values being in a range of [-1,1].
- Assumes that data is normally distributed. If this assumption does not hold, then it may introduce some distortions.

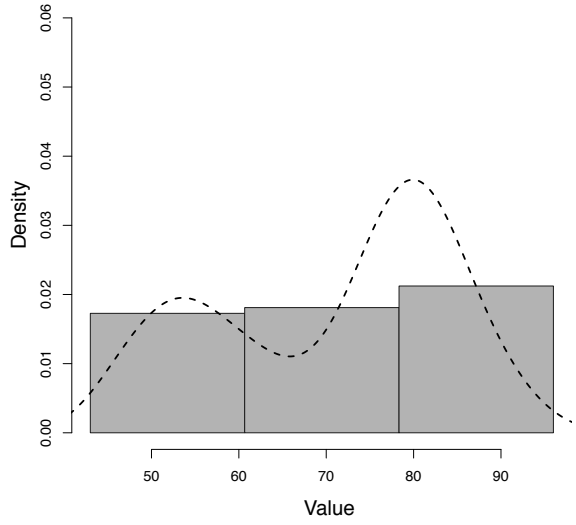
The result of normalising a small sample of the HEIGHT and SPONSORSHIP EARNINGS features from the professional basketball squad dataset.

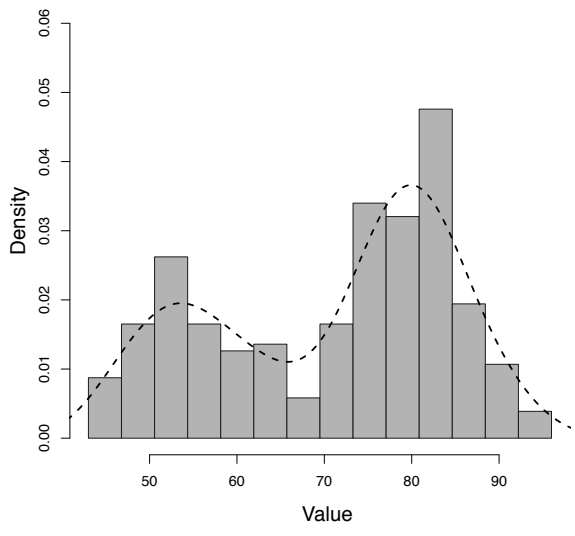
	HEIGHT			SPONSORSHIP EARNINGS		
	Values	Range	Standard	Values	Range	Standard
	192	0.500	-0.073	561	0.315	-0.649
	197	0.679	0.533	1,312	0.776	0.762
	192	0.500	-0.073	1,359	0.804	0.850
	182	0.143	-1.283	1,678	1.000	1.449
	206	1.000	1.622	314	0.164	-1.114
	192	0.500	-0.073	427	0.233	-0.901
	190	0.429	-0.315	1,179	0.694	0.512
	178	0.000	-1.767	1,078	0.632	0.322
	196	0.643	0.412	47	0.000	-1.615
	201	0.821	1.017	1111	0.652	0.384
Max	206			1,678		
Min	178			47		
Mean	193			907		
Std Dev	8.26			532.18		

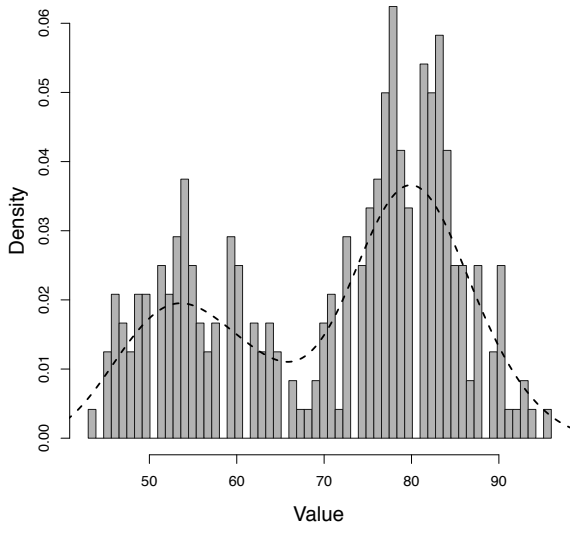
Binning

- **Binning** involves converting a continuous feature into a categorical feature.
- To perform binning, we define a series of ranges (called **bins**) for the continuous feature that correspond to the levels of the new categorical feature we are creating.
- We will introduce two of the more popular ways of defining bins:
  - **equal-width binning**
  - **equal-frequency binning**

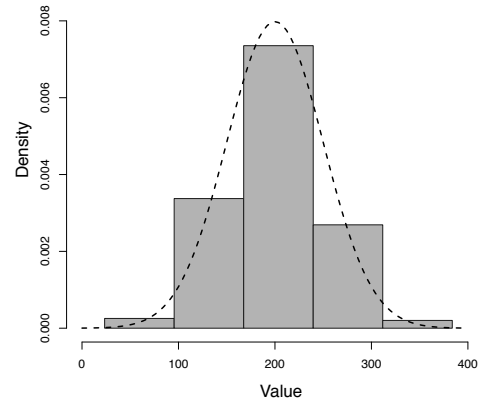
- Deciding on the number of bins can be difficult. The general trade-off is this:
  - If we set the number of bins to a very low number we may lose a lot of information
  - If we set the number of bins to a very high number then we might have very few instances in each bin or even end up with empty bins.

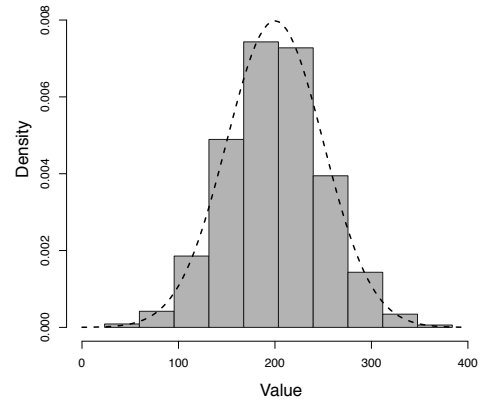

Density
0.00
0.01
0.02
0.03
0.04
0.05
0.06
50
60
70
80
90
Value

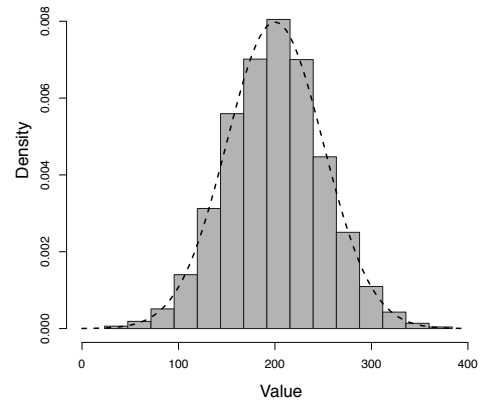

Density
0.00
0.01
0.02
0.03
0.04
0.05
0.06
50
60
70
80
90
Value


Density
0.00
0.01
0.02
0.03
0.04
0.05
0.06
50
60
70
80
90
Value

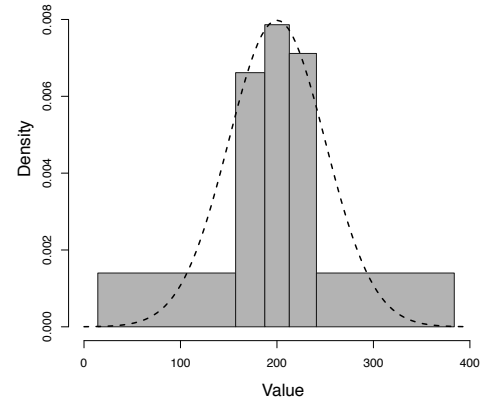
- The equal-width binning algorithm splits the range of the feature values into b bins each of size $\frac{range}{b}$.

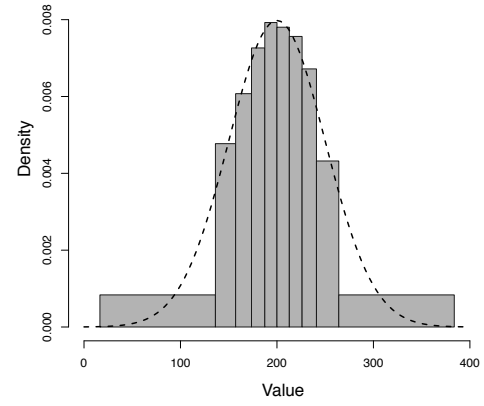

Density
0.000
0.002
0.004
0.006
0.008
0
100
200
300
400
Value

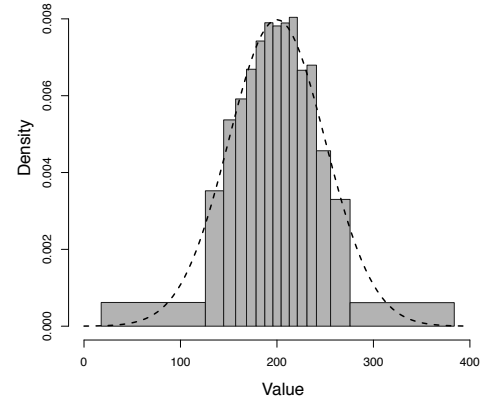

Density
0.000
0.002
0.004
0.006
0.008
0
100
200
300
400
Value


Density
0.000
0.002
0.004
0.006
0.008
0
100
200
300
400
Value

- **Equal-frequency binning** first sorts the continuous feature values into ascending order and then places an equal number of instances into each bin, starting with bin 1.
- The number of instances placed in each bin is simply the total number of instances divided by the number of bins, b.


Density
0.000
0.002
0.004
0.006
0.008
0
100
200
300
400
Value


Density
0.000
0.002
0.004
0.006
0.008
0
100
200
300
400
Value


Density
0.000
0.002
0.004
0.006
0.008
0
100
200
300
400
Value

Sampling

- Sometimes the dataset we have is so large that we do not use all the data available to us in an ABT and instead **sample** a smaller percentage from the larger dataset.
- We need to be careful when sampling, however, to ensure that the resulting datasets are still representative of the original data and that no unintended **bias** is introduced during this process.
- Common forms of sampling include:
  - **top sampling**
  - **random sampling**
  - **stratified sampling**
  - **under-sampling**
  - **over-sampling**

- **Top sampling** simply selects the top $s\%$ of instances from a dataset to create a sample.
- Top sampling runs a serious risk of introducing bias, however, as the sample will be affected by any ordering of the original dataset.
- **Top sampling should be avoided.**

- **Random sampling** randomly selects a proportion of $s\%$ of the instances from a large dataset to create a smaller set.
- Random sampling is a good choice in most cases as the random nature of the selection of instances should avoid introducing bias.

- **Stratified sampling** is a sampling method that ensures that the relative frequencies of the levels of a specific **stratification feature** are maintained in the sampled dataset.
- To perform stratified sampling:
  - the instances in a dataset are divided into groups (or strata), where each group contains only instances that have a particular level for the stratification feature
  - s% of the instances in each stratum are randomly selected
  - these selections are combined to give an overall sample of s% of the original dataset.

- In contrast to stratified sampling, sometimes we would like a sample to contain different relative frequencies of the levels of a particular feature to the distribution in the original dataset.
- To do this, we can use **under-sampling** or **over-sampling**.

- **Under-sampling** begins by dividing a dataset into groups, where each group contains only instances that have a particular level for the feature to be under-sampled.
- The number of instances in the *smallest* group is the under-sampling target size.
- Each group containing more instances than the smallest one is then randomly sampled by the appropriate percentage to create a subset that is the under-sampling target size.
- These under-sampled groups are then combined to create the overall under-sampled dataset.

- **Over-sampling** addresses the same issue as under-sampling but in the opposite way around.
- After dividing the dataset into groups, the number of instances in the *largest* group becomes the over-sampling target size.
- From each smaller group, we then create a sample containing that number of instances using **random sampling with replacement**.
- These larger samples are combined to form the overall over-sampled dataset.

Recommended Reading

- **Core Text:**
  Fundamentals of Machine Learning for Predictive Data Analytics
  By John D. Kelleher, Brian Mac Namee and Aoife D'Arcy
- This week we covered Chapter 3, sections 3.5 and 3.6
- I would suggest that you would read over these sections again
- Email me if you have any questions and I will cover them at the beginning of class next week