Data Exploration
Sections 3.5, 3.6, 3.7

John D. Kelleher and Brian Mac Namee and Aoife D'Arcy

Advanced Data Exploration

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

- 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.

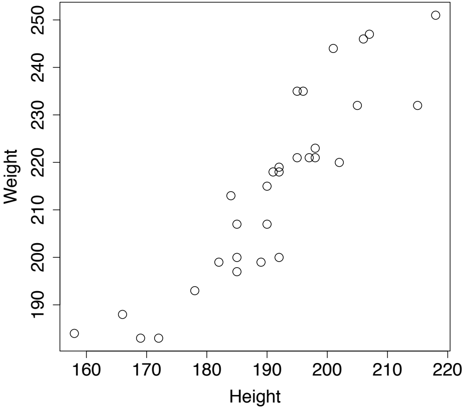


Figure: An example scatter plot showing the relationship between the HEIGHT and WEIGHT features from the professional basketball squad dataset in Table 4 [4].

Visualizing Relationships Between Features

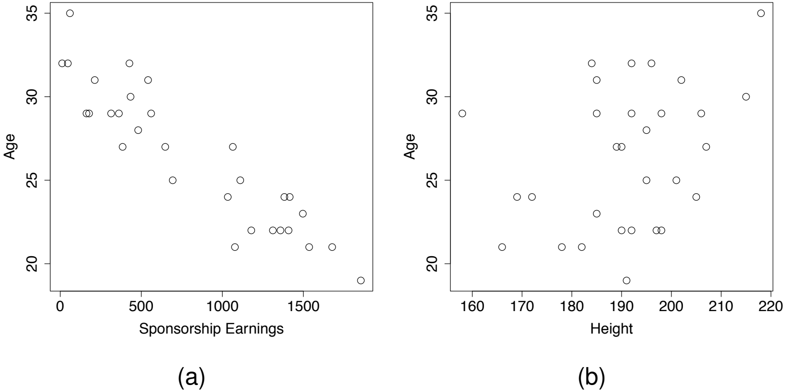


Figure: Example scatter plots showing (a) the strong negative covariance between the SPONSORSHIP EARNINGS and AGE features and (b) the HEIGHT and AGE features from the dataset in Table 4 [4].

- 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.

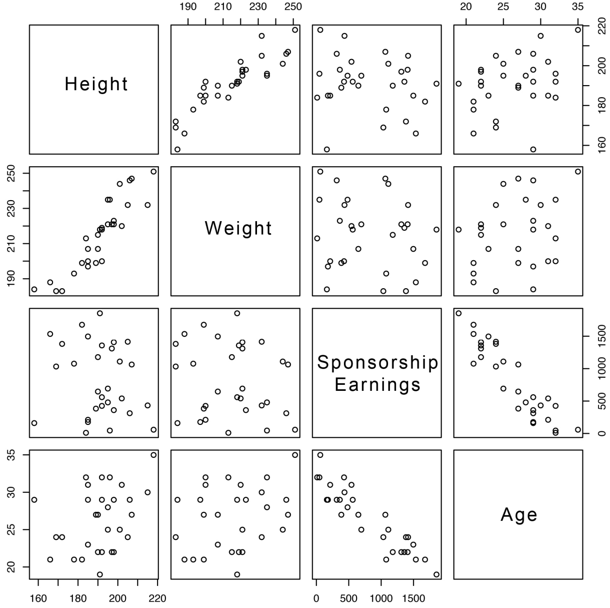


Figure: A scatter plot matrix showing scatter plots of the continuous features from the professional basketball squad dataset.

- The simplest way to visualize the relationship between two categorical variables is to use a collection of **small multiple** bar plots.

Visualizing Relationships Between Features

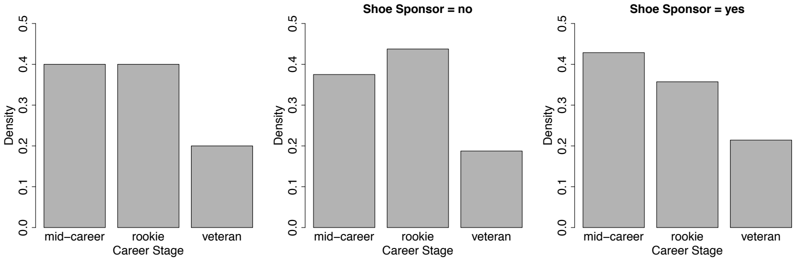


Figure: Using small multiple bar plot visualizations to illustrate the relationship between the CAREER STAGE and SHOE SPONSOR features.

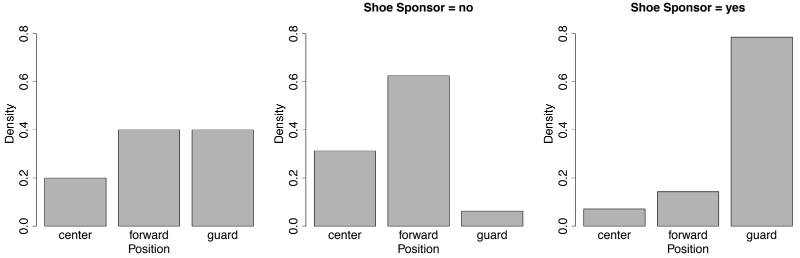
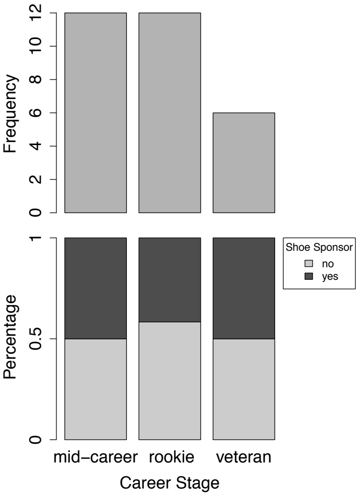
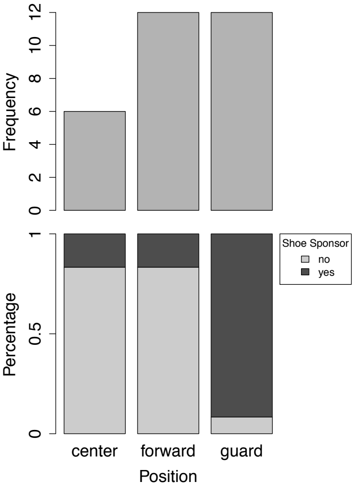


Figure: Using small multiple bar plot visualizations to illustrate the relationship between the POSITION and SHOE SPONSOR features.

- 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.



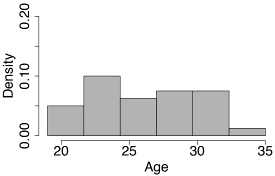
(a) Career Stage and Shoe Sponsor



(b) Position and Shoe Sponsor

Figure: Stacked bar plot visualizations.

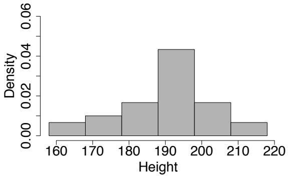
- To visualize the relationship between a continuous feature and a categorical feature a **small multiples** approach that draws a histogram of the values of the continuous feature for each level of the categorical feature is useful.



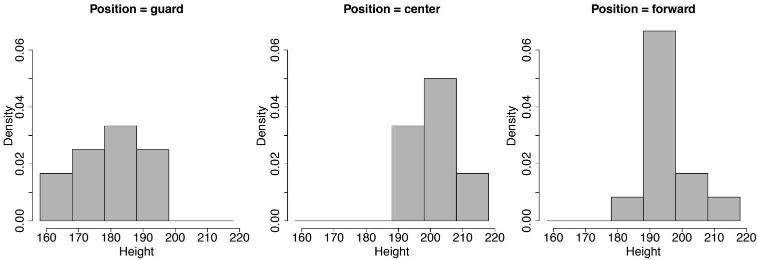
(a) Age

(b) Age and Position

Figure: Using small multiple histograms to visualize the relationship between the AGE feature and the POSITION FEATURE.



(a) Height



(b) Height and Position

Figure: Using small multiple histograms to visualize the relationship between the HEIGHT feature and the POSITION feature.

- A second approach to visualizing the relationship between a categorical feature and a continuous feature is to use a collection of box plots.
- For each level of the categorical feature a box plot of the corresponding values of the continuous feature is drawn.

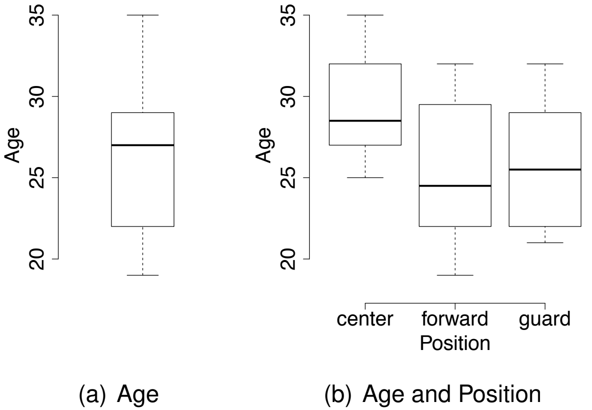
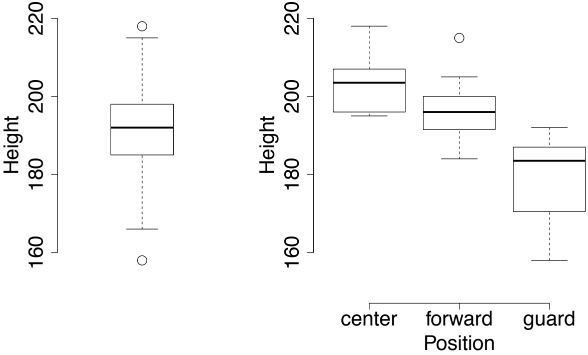


Figure: Using box plots to visualize the relationship between the AGE and the POSITION feature.



(a) Height (b) Height and Position

Figure: Using box plots to visualize the relationship between the HEIGHT feature and the POSITION feature.

- 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

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

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

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.

$$
\begin{aligned}
cov(\text{HEIGHT}, \text{WEIGHT}) &= \frac{7{,}009.9}{29} = 241.72 \\
cov(\text{HEIGHT}, \text{AGE}) &= \frac{570.8}{29} = 19.7
\end{aligned}
$$

- **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.
- 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.

$$
\begin{aligned}
corr(Height, Weight) &= \frac{241.72}{13.6 \times 19.8} = 0.898 \\
corr(Height, Age) &= \frac{19.7}{13.6 \times 4.2} = 0.345
\end{aligned}
$$

- 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:

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

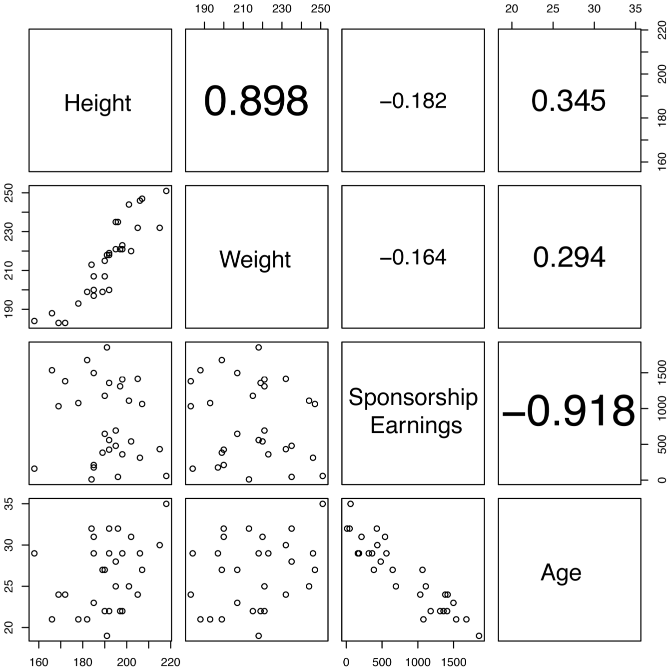
- Calculating covariances matrix for the HEIGHT feature and the WEIGHT and AGE features from the basketball players dataset.

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

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

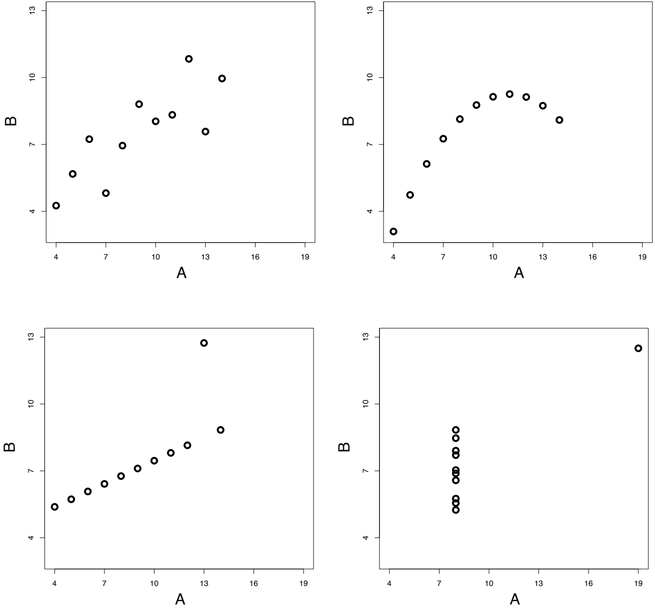
$$\underset{<\mathit{Height},\mathit{Weight},\mathit{Age}>}{\text{correlation matrix}} = \begin{bmatrix} 1.0 & 0.898 & 0.345 \\ 0.898 & 1.0 & 0.294 \\ 0.345 & 0.294 & 1.0 \end{bmatrix}$$

- 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.



Height
0.898
−0.182
0.345
Weight
−0.164
0.294
Sponsorship Earnings
−0.918
Age

- 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**.



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.

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** 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:

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

- 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:

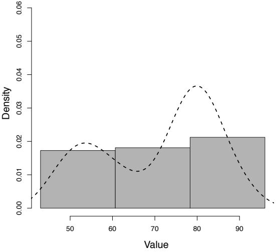
$$a_i^{'} = \frac{a_i - \overline{a}}{sd(a)} \tag{6}$$

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

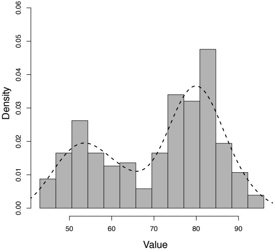
	HEIGHT			SPONSORSHIP EARNINGS		
	Values	Range	Standard	Values	Range	Standard
	192	0.500	-0.073	561	0.315	-0.649
	197	0.679	0.533	1,312	0.776	0.762
	192	0.500	-0.073	1,359	0.804	0.850
	182	0.143	-1.283	1,678	1.000	1.449
	206	1.000	1.622	314	0.164	-1.114
	192	0.500	-0.073	427	0.233	-0.901
	190	0.429	-0.315	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** 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**

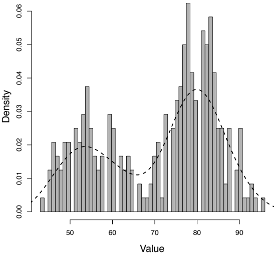
- Deciding on the number of bins can be difficult. The general trade-off is this:
    - If we set the number of bins to a very low number we may lose a lot of information
    - If we set the number of bins to a very high number then we might have very few instances in each bin or even end up with empty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



(e) 3 bins

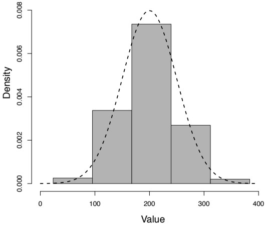


(f) 14 bins

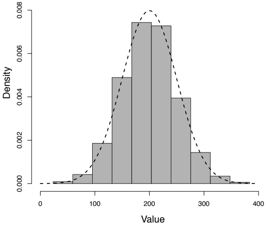


(g) 60 bins

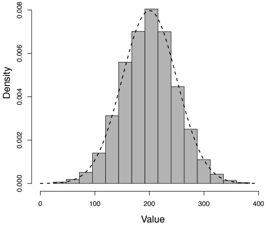
- The equal-width binning algorithm splits the range of the feature values into b bins each of size $\frac{range}{b}$.



(h) 5 Equal-width bins

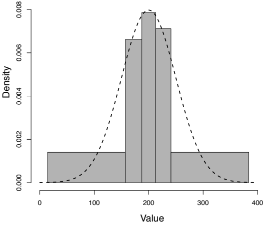


(i) 10 Equal-width bins

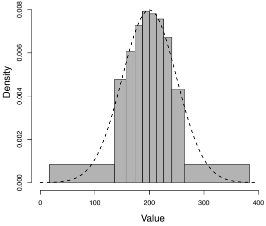


(j) 15 Equal-width bins

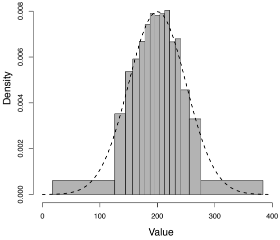
- **Equal-frequency binning** first sorts the continuous feature values into ascending order and then places an equal number of instances into each bin, starting with bin 1.
- The number of instances placed in each bin is simply the total number of instances divided by the number of bins, b.



(k) 5 Equal-frequency bins



(l) 10 Equal-frequency bins



(m) 15 Equal-frequency bins

- 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.
- We recommend that top sampling be avoided.

- Our recommended default, **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.

Summary

- 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**probability distribution.
  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.

1. **Advanced Data Exploration**
   - Visualizing Relationships Between Features
   - Measuring Covariance & Correlation

2. **Data Preparation**
   - Normalization
   - Binning
   - Sampling

3. **Summary**