Data Quality Report – Case Study

Fundamentals of Machine Learning for Predictive Data Analytics,

John Kelleher and Brian Mac Namee and Aoife D'Arcy

Chapter 3

Case Study: Analytical Base Table

ID	Type	Inc.	Marital Status	Num Clmnts.	Injury Type	Hospital Stay	Claim Amnt.	Total Claimed	Num Claims	Num Soft Tiss.	% Soft Tiss.	Claim Amt Rcvd.	Fraud Flag
1	CI	0		2	Soft Tissue	No	1,625	3250	2	2	1.0	0	1
2	CI	0		2	Back	Yes	15,028	60,112	1		0	15,028	0
3	CI	54,613	Married	1	Broken Limb	No	-99,999	0	0	0	0	572	0
4	CI	0		4	Broken Limb	Yes	5,097	11,661	1	1	1.0	7,864	0
5	CI	0		4	Soft Tissue	No	8869	0	0	0	0	0	1
6	CI	0		1	Broken Limb	Yes	17,480	0	0	0	0	17,480	0
7	CI	52,567	Single	3	Broken Limb	No	3,017	18,102	2	1	0.5	0	1
8	CI	0		2	Back	Yes	7463	0	0	0	0	7,463	0
9	CI	0		1	Soft Tissue	No	2,067	0	0	0	0	2,067	0
10	CI	42,300	Married	4	Back	No	2,260	0	0	0	0	2,260	0
		:				:					:		
300	CI	0		2	Broken Limb	No	2,244	0	0	0	0	2,244	0
301	CI	0		1	Broken Limb	No	1,627	92,283	3	0	0	1,627	0
302	CI	0		3	Serious	Yes	270,200	0	0	0	0	270,200	0
303	CI	0		1	Soft Tissue	No	7,668	92,806	3	0	0	7,668	0
304	CI	46,365	Married	1	Back	No	3,217	0	0		0	1,653	0
		:				:					:		
458	CI	48,176	Married	3	Soft Tissue	Yes	4,653	8,203	1	0	0	4,653	0
459	CI	0		1	Soft Tissue	Yes	881	51,245	3	0	0	0	1
460	CI	0		3	Back	No	8,688	729,792	56	5	0.08	8,688	0
461	CI	47,371	Divorced	1	Broken Limb	Yes	3,194	11,668	1	0	0	3,194	0
462	CI	0		1	Soft Tissue	No	6,821	0	0	0	0	0	1
		:				:					:		
491	CI	40,204	Single	1	Back	No	75,748	11,116	1	0	0	0	1
492	CI	0		1	Broken Limb	No	6,172	6,041	1		0	6,172	0
493	CI	0		1	Soft Tissue	Yes	2,569	20,055	1	0	0	2,569	0
494	CI	31,951	Married	1	Broken Limb	No	5,227	22,095	1	0	0	5,227	0
495	CI	0		2	Back	No	3,813	9,882	3	0	0	0	1
496	CI	0		1	Soft Tissue	No	2,118	0	0	0	0	0	1
497	CI	29,280	Married	4	Broken Limb	Yes	3,199	0	0	0	0	0	1
498	CI	0		1	Broken Limb	Yes	32,469	0	0	0	0	16,763	0
499	CI	46,683	Married	1	Broken Limb	No	179,448	0	0		0	179,448	0
500	CI	0		1	Broken Limb	No	8,259	0	0	0	0	0	1

Case study: Data Quality Reports

(a) Continuous Features

Feature	Count	% Miss.	Card.	Min	1st Qrt.	Mean	Median	3rd Qrt.	Max	Std. Dev.
INCOME	500	0.0	171	0.0	0.0	13,740.0	0.0	33,918.5	71,284.0	20,081.5
NUM CLAIMANTS	500	0.0	4	1.0	1.0	1.9	2	3.0	4.0	1.0
CLAIM AMOUNT	500	0.0	493	-99,999	3,322.3	16,373.2	5,663.0	12,245.5	270,200.0	29,426.3
TOTAL CLAIMED	500	0.0	235	0.0	0.0	9,597.2	0.0	11,282.8	729,792.0	35,655.7
NUM CLAIMS	500	0.0	7	0.0	0.0	0.8	0.0	1.0	56.0	2.7
NUM SOFT TISSUE	500	2.0	6	0.0	0.0	0.2	0.0	0.0	5.0	0.6
% SOFT TISSUE	500	0.0	9	0.0	0.0	0.2	0.0	0.0	2.0	0.4
AMOUNT RECEIVED	500	0.0	329	0.0	0.0	13,051.9	3,253.5	8,191.8	295,303.0	30,547.2
FRAUD FLAG	500	0.0	2	0.0	0.0	0.3	0.0	1.0	1.0	0.5

(a) Categorical Features

Feature	Count	% Miss.	Card.	Mode	Mode Freq.	Mode %	2nd Mode	2nd Mode Freq.	2nd Mode %
INSURANCE TYPE	500	0.0	1	CI	500	1.0	–	–	–
MARITAL STATUS	500	61.2	4	Married	99	51.0	Single	48	24.7
INJURY TYPE	500	0.0	4	Broken Limb	177	35.4	Soft Tissue	172	34.4
HOSPITAL STAY	500	0.0	2	No	354	70.8	Yes	146	29.2

Analysis – Missing values

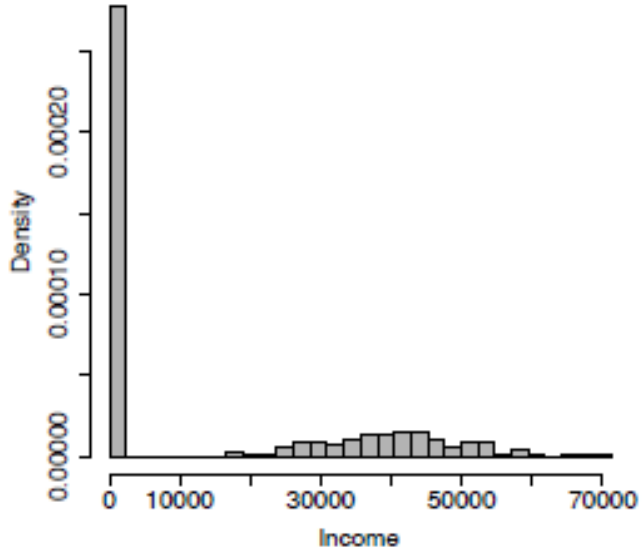
NUM SOFT TISSUE – 2%

Possible solution: leave it as it is, document it in the data quality plan

MARITAL STATUS – 61.2%

Possible solution – remove the feature

- Examining the raw data we notice that the fields where MARITAL STATUS is missing we have income value 0
- Unusual pattern for INCOME from the histogram large number of 0s
- Discussion with the business revealed that MARITAL STATUS and INCOME were collected together, and INCOME 0 represents a missing value



(a) INCOME

Decision → remove both MARITAL STATUS and INCOME from ABT

Analysis – Irregular cardinality

INSURANCE TYPE – cardinality 1

Nothing wrong with the data, refers to the type of claim (CI = Car Insurance). However, not useful for predictions

Decision → feature is removed from the ABT

Analysis – Irregular cardinality

Continuous features with low cardinality:

NUM CLAIMANTS, NUM CLAIMS, SOFT TISSUE, SOFT TISSUE %, FRAUD FLAG – all have cardinality <10 (dataset of 500 instances)

- Discussing with the business indicates that NUM CLAIMANTS, NUM CLAIMS, SOFT TISSUE, SOFT TISSUE% naturally take low number of values

Decision → no issues

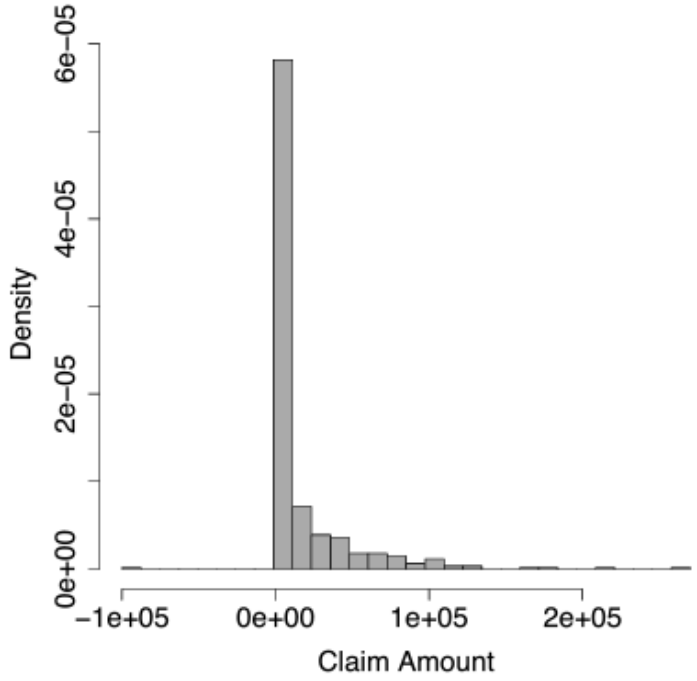
- FRAUD FLAG has cardinality 2, indicates categorical feature labelled as continuous with values 0 and 1.

Decision → change FRAUD FLAG to categorical feature (note – this is also our target feature, so the type will have large impact on the ML approach we take)

Analysis - Outliers

CLAIM AMOUNT – min value -99,999

- Examining the raw data we notice this value comes from the d_3 in our table.
- Examining the histogram for CLAIM AMOUNT we don't see a large bar at that value, so this is isolated instance
- Value looks like input error, which was confirmed with the business



Decision → invalid outlier, remove it and replace it with missing value

Analysis - Outliers

CLAIM AMOUT, TOTAL CLAIMED, NUM CLAIMS, AMOUNT RECIEVED – all have unusually high maximum values, especially compared to median and 3rd quartile

TOTAL CLAIMED, NUM CLAIMS – both high values are from d_{460} in ABT

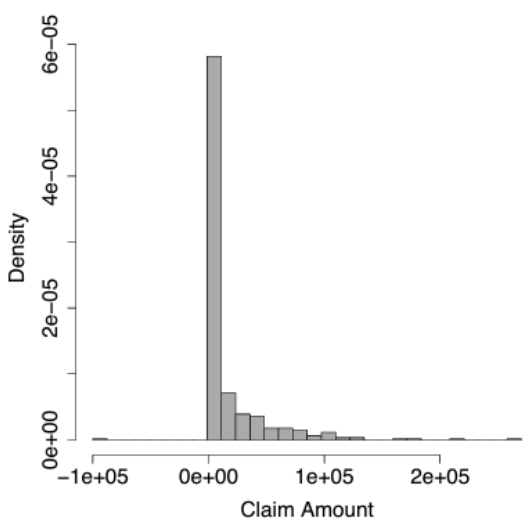
- This policy seems to have made many more claims compared to the others
- Investigation with the business shows this is a valid outlier, however this was a company policy and was included in the ABT by mistake.

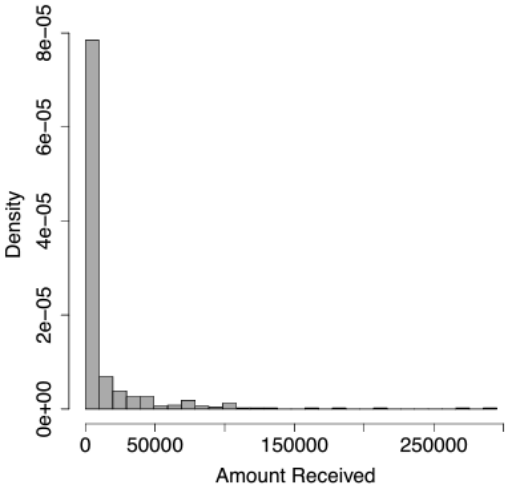
Decision → d_{460} was removed from the ABT

Analysis - Outliers

CLAIM AMOUT and AMOUNT RECEIVED – both from d_{302} in ABT

- Examining the raw data we note this is a large claim for a serious injury
- Analysis of the histograms show that the large values are no unique (several small bars in the right hand side of the histograms)





Decision → document the outliers and possibly handle them later

Handling missing values

Approach 1: Drop any features that have missing value

Rule of thumb – only use if 60% or more of values missing, otherwise look for different approach to handle

Approach 2: Apply **complete case analysis**.

Delete any instances where one or more features are missing values

Can result in significant data loss, can introduce bias (if distribution of missing values is not random)

Recommendation: Remove only instances with missing value for target feature

Approach 3: Derive a **missing indicator feature** from features with missing value.

Binary feature that flags whether the value was present or missing in the original data. May be useful if the reason the features are missing is related to the target feature

Handling missing values

Imputation replaces missing feature values with a plausible estimated value based on the feature values that are present.

The most common approach to imputation is to replace missing values for a feature with a measure of the central tendency of that feature.

- Continuous features – usually mean and median
- Categorical features – usually the mode

We would be reluctant to use imputation on features missing in excess of 30% of their values and would strongly recommend against the use of imputation on features missing in excess of 50% of their values.

Handling outliers – clamp transformation

The easiest way to handle outliers is to use a **clamp transformation** that clamps all values above an upper threshold and below a lower threshold to these threshold values, thus removing the offending outliers

$$a_i = \begin{cases} lower & \text{if } a_i < lower \\ upper & \text{if } a_i > upper \\ a_i & otherwise \end{cases} \qquad (2)$$

where a_i is a specific value of feature a, and *lower* and *upper* are the lower and upper thresholds.

Handling outliers - clamp transformation

Upper and lower limits can be set manually based on domain knowledge, or can be calculated from the data

Method 1:

lower = 1st quartile value – 1.5*inter-quartile range

higher = 3rd quartile value + 1.5*inter-quartile range

Method 2:

Use the mean value of a feature ± 2 times standard deviation

Handling outliers – clamp transformation

Clamp transformation – for and against

Performing clamp transformation may remove the most interesting (and predictive) data from the dataset

However, some machine learning techniques perform poorly in presence of outliers

Clamp transformations are only recommended when you suspect the outliers will affect the performance of the model

Case study: Data Quality Plan

Table: The data quality plan for the motor insurance fraud prediction ABT.

Feature	Data Quality Issue	Potential Handling Strategies
NUM SOFT TISSUE	Missing values (2%)	
CLAIM AMOUNT	Outliers (high)	
AMOUNT RECEIVED	Outliers (high)	

Case study: Data Quality Plan

NUM SOFT TISSUE

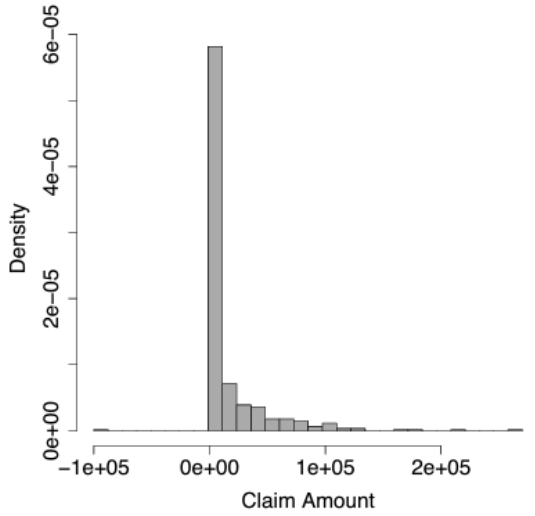
Missing only 2% of the values, we'll use imputation.

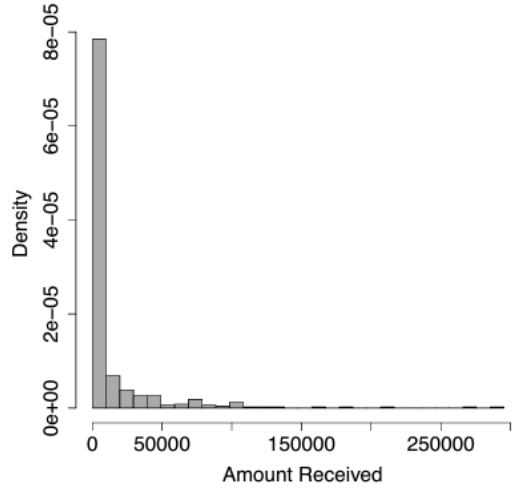
We'll use the mean or the median (value 0.2 doesn't naturally occur in the data set, so we'll use 0)

Case study: Data Quality Plan

CLAIM AMOUNT, AMOUNT RECEIVED

- we'll use clamp transformation
- exponential distribution, so methods we discussed won't work too well
- following a discussion with the business we were advised that lower limit of 0 and upper limit of 80,000 make sense for the clamp transformation





Case study: Data Quality Plan

Case Study: Motor Insurance Fraud

Table: The data quality plan for the motor insurance fraud prediction ABT.

Feature	Data Quality Issue	Potential Handling Strategies
NUM SOFT TISSUE	Missing values (2%)	Imputation (median: 0.0)
CLAIM AMOUNT	Outliers (high)	Clamp transformation (manual: 0, 80 000)
AMOUNT RECEIVED	Outliers (high)	Clamp transformation (manual: 0, 80 000)

Visualising relationship between features

In preparation to create predictive models it's always a good idea to investigate the relationship between the variables

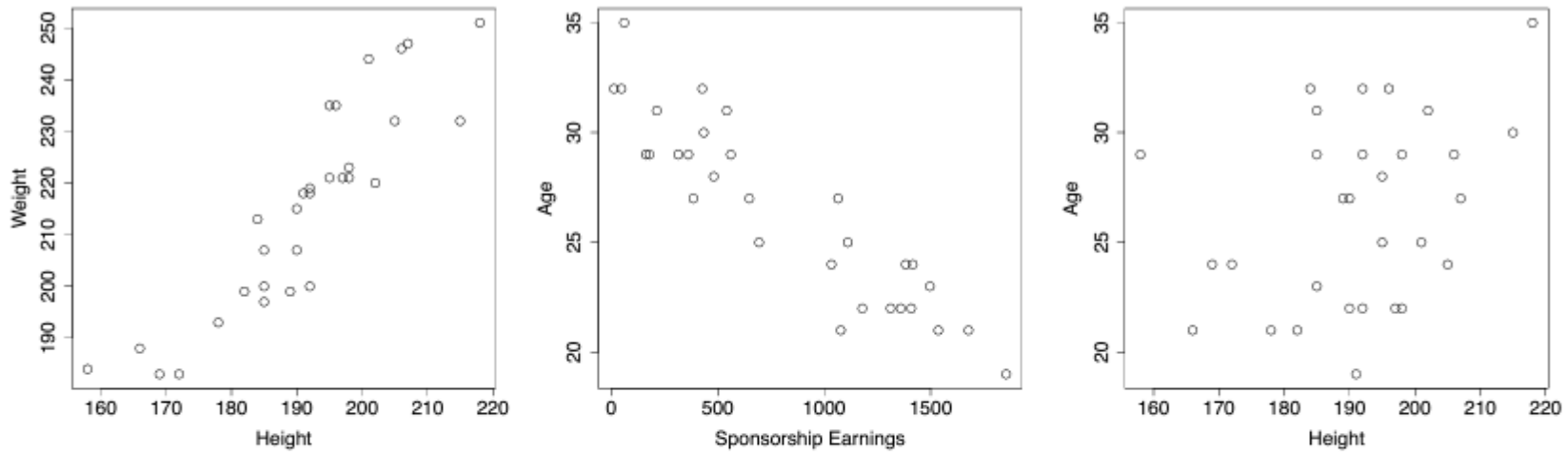
This can identify pairs of closely-related features (which in turn can be used to reduce the size of the ABT)

Table 3.7

The details of a professional basketball team.

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

Scatter plot (continuous features)



(a) (b) (c)

Figure 3.5

Example scatter plots for pairs of features from the dataset in Table 3.7[78], showing (a) the strong positive covariance between HEIGHT and WEIGHT; (b) the strong negative covariance between SPONSORSHIP EARNINGS and AGE; and (c) the lack of strong covariance between HEIGHT and AGE.

SPLOM (Scatter plot matrix)

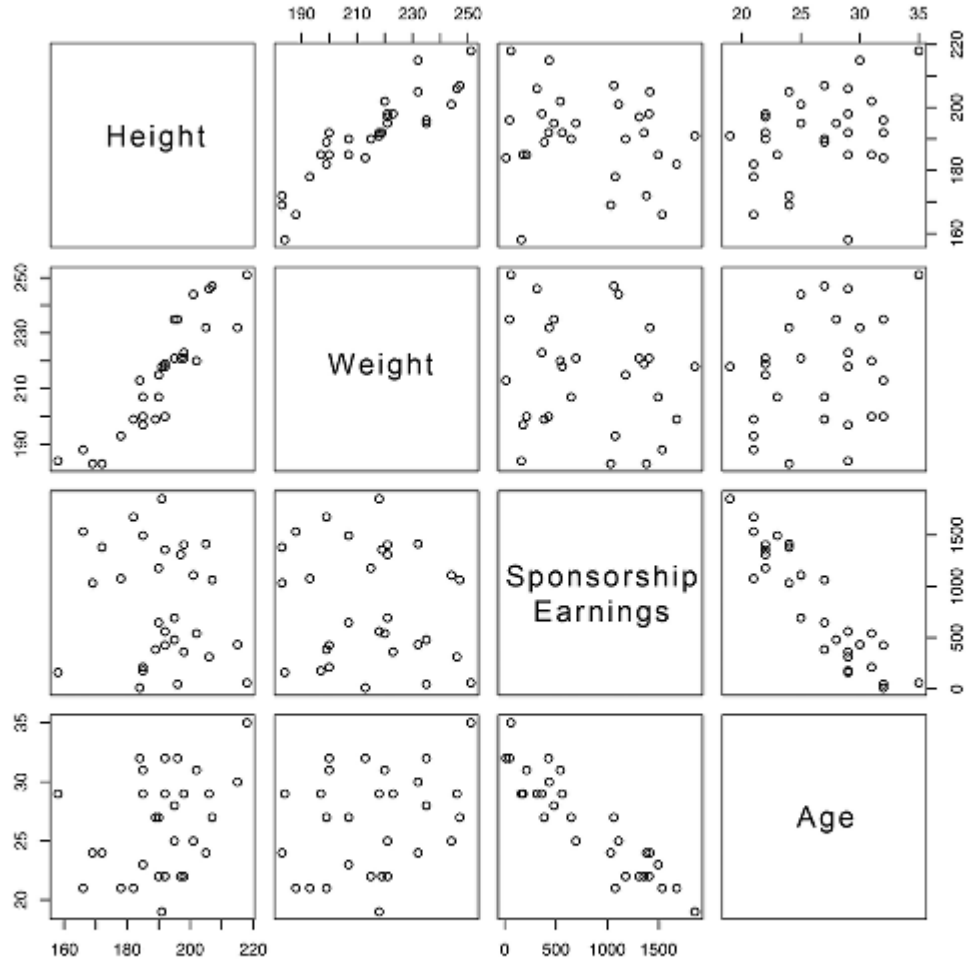
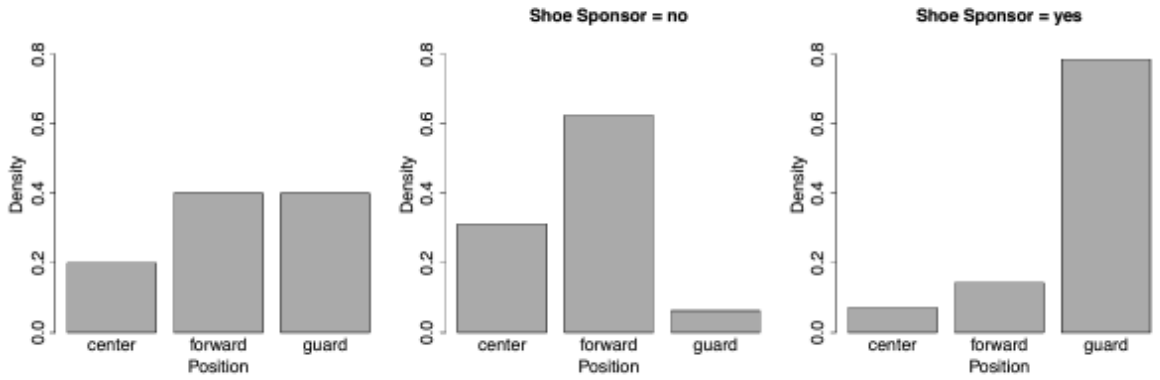


Figure 3.6

A scatter plot matrix showing scatter plots of the continuous features from the professional basketball team dataset in Table 3.7[78].

Bar plot (categorical features)

(a) Career Stage and Shoe Sponsor



(b) Position and Shoe Sponsor

Figure 3.7

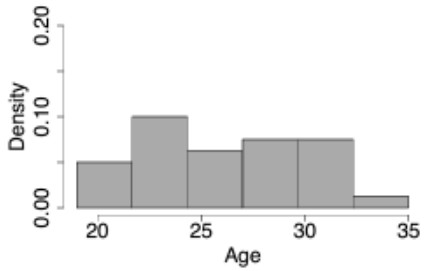
Examples of using small multiple bar plot visualizations to illustrate the relationship between two categorical features: (a) the CAREER STAGE and SHOE SPONSOR features; and (b) the POSITION and SHOE SPONSOR features. All data comes from Table 3.7[78].

"Small multiples approach"

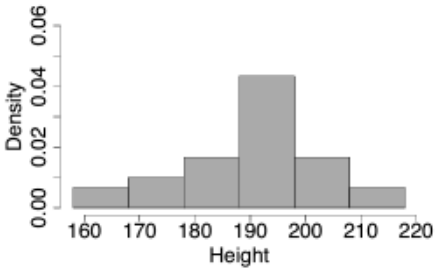
If the two features being visualised have a strong relationship, then the bar plots for each level of the second feature will look noticeably different to one another, and to the overall bar plot.

Small multiple histograms

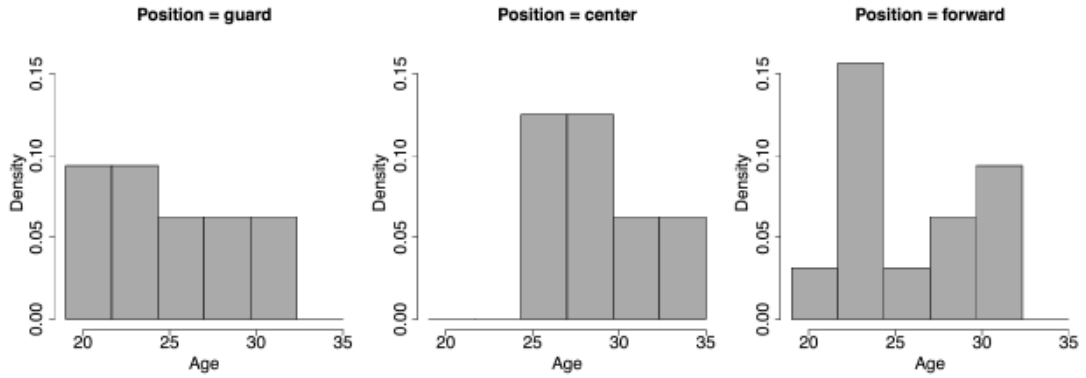
"Small multiple" histograms can be used to compare a categorical with a continuous feature



(a) Age

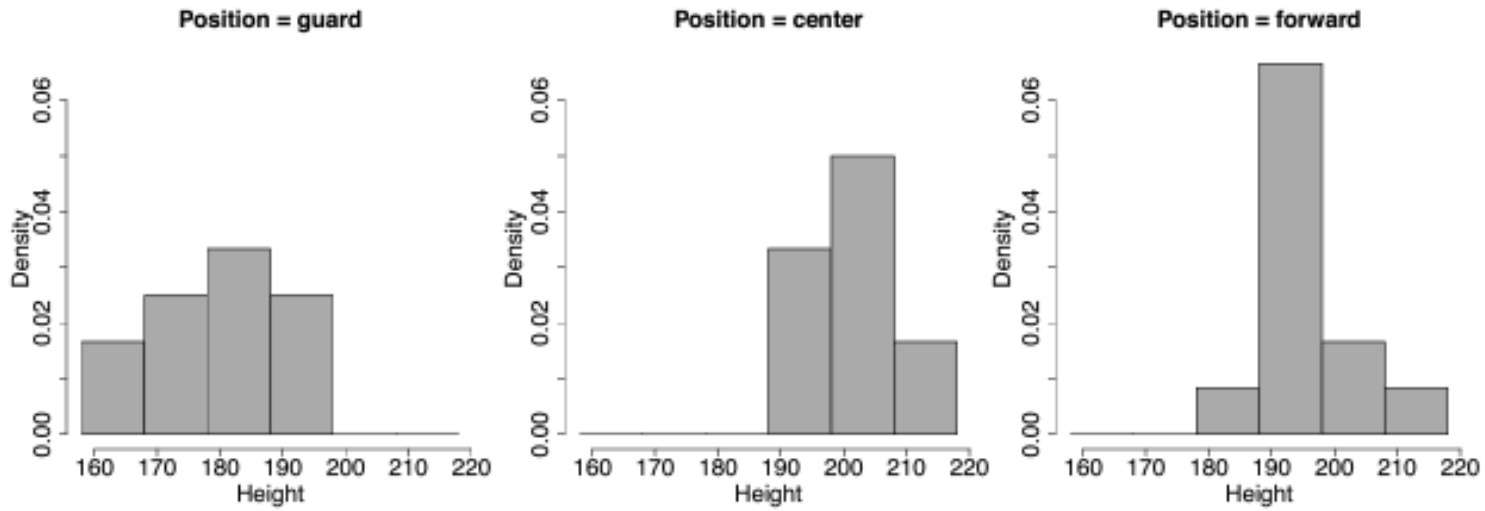


(b) Height



(c) Age and Position

Small multiple histograms (cont.)



(d) Height and Position

Figure 3.9

Example of using small multiple histograms to visualize the relationship between a categorical feature and a continuous feature. All examples use data from the professional basketball team dataset in Table 3.7[78]: (a) a histogram of the AGE feature; (b) a histogram of the HEIGHT feature; (c) histograms of the AGE feature for instances displaying each level of the POSITION feature; and (d) histograms of the HEIGHT feature for instances displaying each level of the POSITION feature.

Covariance and correlation

Covariance and **correlation** provide us with a formal measures of the relationship of two continuous variables

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

Covariance has a possible range of [-∞, ∞] where negative values indicate negative relationship, positive values indicate positive relationship, and values around 0 indicate little or no relationship

Correlation

Correlation is a normalized form of covariance

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

Correlation has a range of [-1, 1]

Correlation matrix

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

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

Summary (Data Quality Reports)

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.