




# DATA EXPLORATION

CS576 MACHINE LEARNING



Dr. Jin Soung Yoo, Professor  
Department of Computer Science  
Purdue University Fort Wayne

# Reading

- Kelleher et al., Machine Learning for Predictive Data Analytics, Ch 3.1 – Ch 3.5

# Outline

- Data Exploration
  - Descriptive Statistics
  - Data Visualization
  - Common Histogram Shapes
- Data Quality Issues
  - Missing Values
  - Irregular Cardinality
  - Outliers
- Visualizing Relationships Between Features
- Measuring Covariance & Correlation
- Sampling
- Summary

# Getting To Know The Data

- Before attempting to build predictive models, it is important that we do some exploratory analysis of a data, or **data exploration**.
- The stage of data exploration is mostly for an information-gathering – just a better understanding of the content of the data analyzed

# Descriptive Statistics for Machine Learning

- Two important things to understand the characteristics of each feature in the data are **central tendency** and **variation**
  - The **central tendency** of a feature refers to the value that is *typical* of the feature and therefore can be used to summarize it
  - The **variation** of a feature indicates variation of the values within the data of the feature.
- We examine the central tendency and variation of each feature using **fundamental statistical measures**.

# Descriptive Statistics for Continuous Features

## ■ Central tendency

For a sample of  $n$  values for a feature  $x$ ,

- **arithmetic mean** (or **sample mean** or **mean**):  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ 
    - Mean is very sensitive to the presence of **outliers**
  - **median**: the middle value of a feature among ordered values from lowest to highest
    - Median is not as sensitive to outliers
  - **mode**: the most commonly occurring value in a feature
- Any measure of central tendency is just an **approximation**

# Descriptive Statistics for Continuous Features

For a sample of  $n$  values for a feature  $x$ ,

## ■ Variation

- **range**:  $range = \max(x) - \min(x)$

- **variance**:  $var(x) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$

- the average difference between each value in a sample and the mean of that sample.

- **Standard deviation**:  $sd(x) = \sqrt{var(x)} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$

- **Percentiles**: A proportion of  $\frac{i}{100}$  of the values in a sample take values equal to or lower than the  $i^{\text{th}}$  percentile of that sample

- **25<sup>th</sup> percentile** (**1<sup>st</sup> quartile**), **50<sup>th</sup> percentile** (**median**),  
**75<sup>th</sup> percentile** (**3<sup>rd</sup> quartile**)

- **Inter-quartile range (IQR)**: **3<sup>rd</sup> quartile** - **1<sup>st</sup> quartile**

# Descriptive Statistics for Categorical Features

- For categorical features we are interested primarily in **frequency counts** and **proportions**.
- **Frequency count**
  - The frequency count of each level of a categorical feature is the number of items that level appears in the sample
- **Proportion:**  $\frac{\text{frequency count of that level}}{\text{total sample size of the feature}}$

- **Example:**

Level	Count	Proportion
guard	8	40%
forward	7	35%
center	5	25%

Frequency table of Position feature

Training			Training		
ID	Position	Expenses	ID	Position	Expenses
1	center	56.75	11	center	550.00
2	guard	1,800.11	12	center	223.89
3	guard	1,341.03	13	center	103.23
4	forward	749.50	14	forward	758.22
5	guard	1,150.00	15	forward	430.79
6	forward	928.30	16	forward	675.11
7	center	250.90	17	guard	1,657.20
8	guard	806.15	18	guard	1,405.18
9	guard	1,209.02	19	guard	760.51
10	forward	405.72	20	forward	985.41



# Data Quality Report

## (a) Continuous Features

Feature	Count	% Miss.	Card.	Min.	1 <sup>st</sup> Qrt.	Mean	Median	3 <sup>rd</sup> Qrt.	Max.	Std. Dev.
_____	_____	_____	_____	_____	_____	_____	_____	_____	_____	_____
_____	_____	_____	_____	_____	_____	_____	_____	_____	_____	_____
_____	_____	_____	_____	_____	_____	_____	_____	_____	_____	_____

## (b) Categorical Features

Feature	Count	% Miss.	Card.	Mode	Mode Freq.	Mode %	2 <sup>nd</sup> Mode	2 <sup>nd</sup> Mode Freq.	2 <sup>nd</sup> Mode %
_____	_____	_____	_____	_____	_____	_____	_____	_____	_____
_____	_____	_____	_____	_____	_____	_____	_____	_____	_____
_____	_____	_____	_____	_____	_____	_____	_____	_____	_____

**Table:** The structures of the tables included in a data quality report to describe (a) continuous features and (b) categorical features.

The card. (**cardinality**) of each feature measures the number of distinct values present in the data for a feature. The **mode** and **2<sup>nd</sup> mode** are the two most frequent levels for each feature, and **mode freq.** and **2<sup>nd</sup> mode freq.** are the frequency with which these appear.

# Example: Data for Motor Insurance Fraud Detection

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	Married	2	Soft Tissue	No	1,625	3250	2	2	1.0	0	1
2	CI	0		2	Back	Yes	15,028	60,112	1	1	0	15,028	0
3	CI	54,613		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	Single	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		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	Married	1	Soft Tissue	No	2,067	0	0	0	0	2,067	0
10	CI	42,300		4	Back	No	2,260	0	0	0	0	2,260	0
300	CI	0	Married	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	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		1	Broken Limb	Yes	3,194	11,668	1	0	0	3,194	0
462	CI	0	Divorced	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	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		1	Broken Limb	No	5,227	22,095	1	0	0	5,227	0
495	CI	0	Married	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		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	0	179,448	0
500	CI	0		1	Broken Limb	No	8,259	0	0	0	0	0	1

**Table:** Portions of the ABT (Analytics Base Table) for the motor insurance claims fraud detection problem.

# Example: Data Quality Report

## (a) Continuous Features

Feature	Count	% Miss.	Card.	Min	1 <sup>st</sup> Qrt.	Mean	Median	3 <sup>rd</sup> 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 %	2 <sup>nd</sup> Mode	2 <sup>nd</sup> Mode Freq.	2 <sup>nd</sup> 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

**Table:** A data quality report for the motor insurance claims fraud detection data

# Outline

- Data Exploration
  - Descriptive Statistics
  - ☞ **Data Visualization**
    - Common Histogram Shapes
- Data Quality Issues
- Visualizing Relationships Between Features
- Measuring Covariance & Correlation
- Sampling
- Summary

# Data Visualization

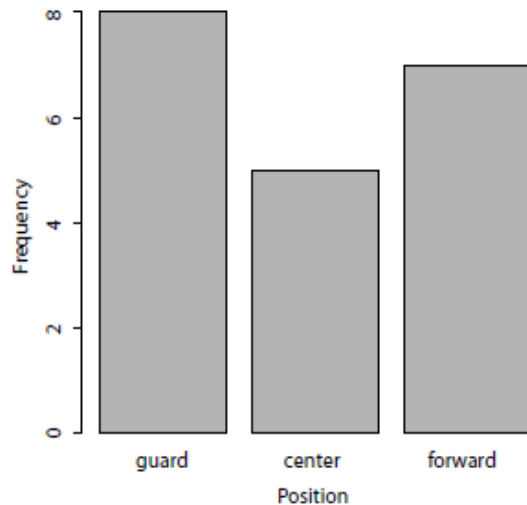
- When performing data exploration **data visualization** can help enormously.
- Three important data visualization techniques that can be used to visualize the values in a single feature:
  - **bar plot**
  - **histogram**
  - **box plot**

# Visualization for Data Distribution

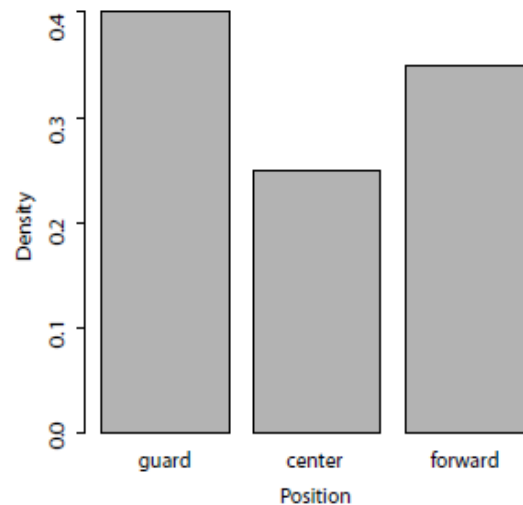
- The data quality reports are accompanied by **data visualizations** that illustrate the distribution of the values in each feature:
  - A **bar plot** for **categorical feature**.
  - A **histogram** for **continuous feature**.

# Example: Bar Plots

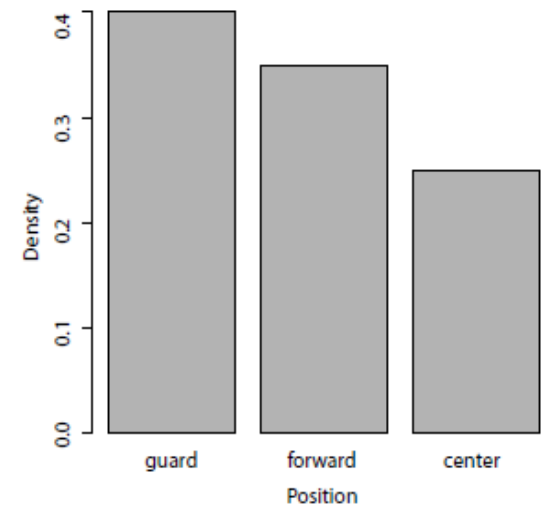
Training Expenses			Training Expenses		
ID	Position	Training Expenses	ID	Position	Training Expenses
1	center	56.75	11	center	550.00
2	guard	1,800.11	12	center	223.89
3	guard	1,341.03	13	center	103.23
4	forward	749.50	14	forward	758.22
5	guard	1,150.00	15	forward	430.79
6	forward	928.30	16	forward	675.11
7	center	250.90	17	guard	1,657.20
8	guard	806.15	18	guard	1,405.18
9	guard	1,209.02	19	guard	760.51
10	forward	405.72	20	forward	985.41



(a) Frequency



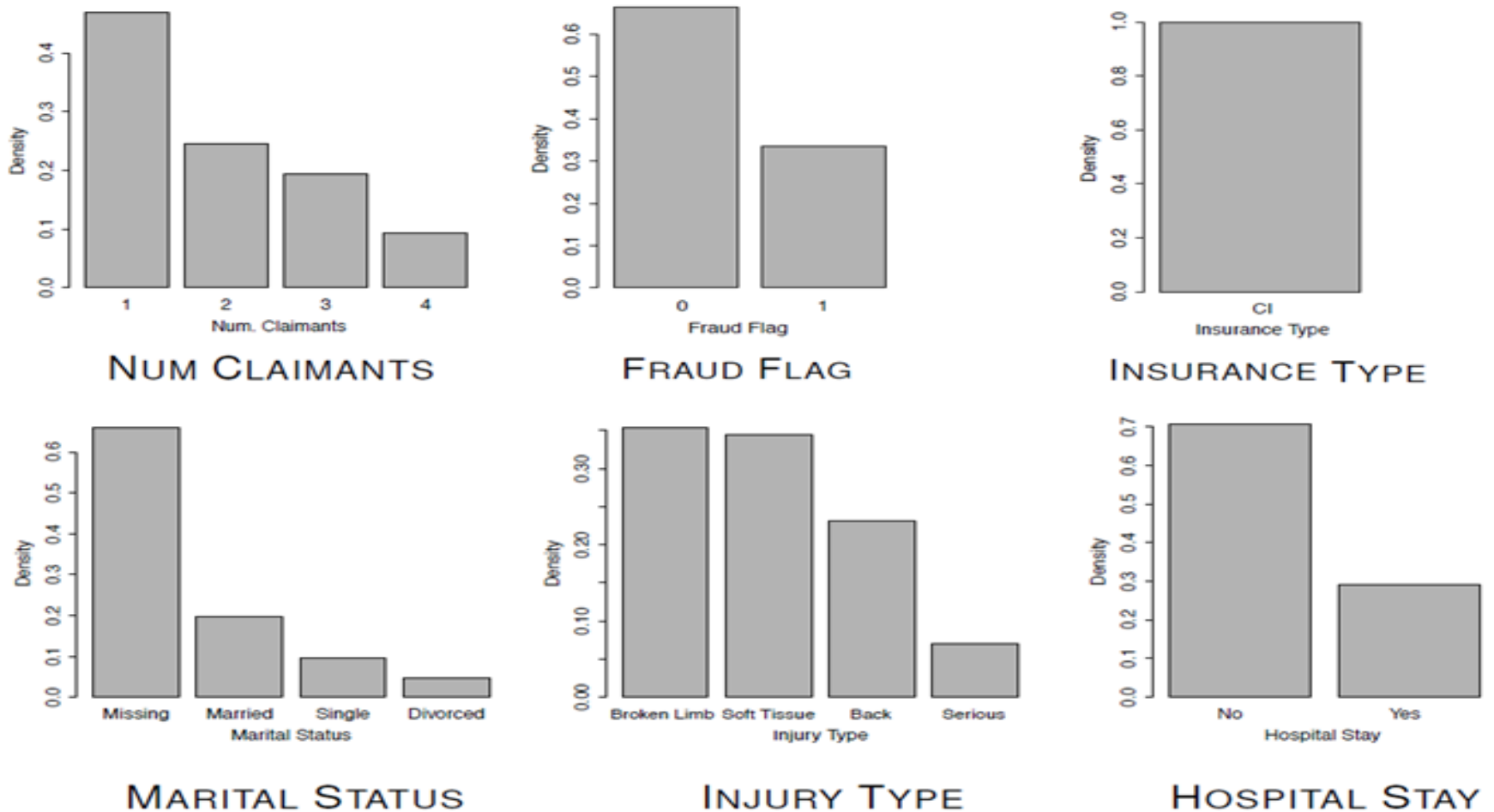
(b) Proportion



(c) Ordered

- From the bar plots, *Guard* is the most frequent level of POSITION feature in a basketball team dataset

# Example: Bar Plots



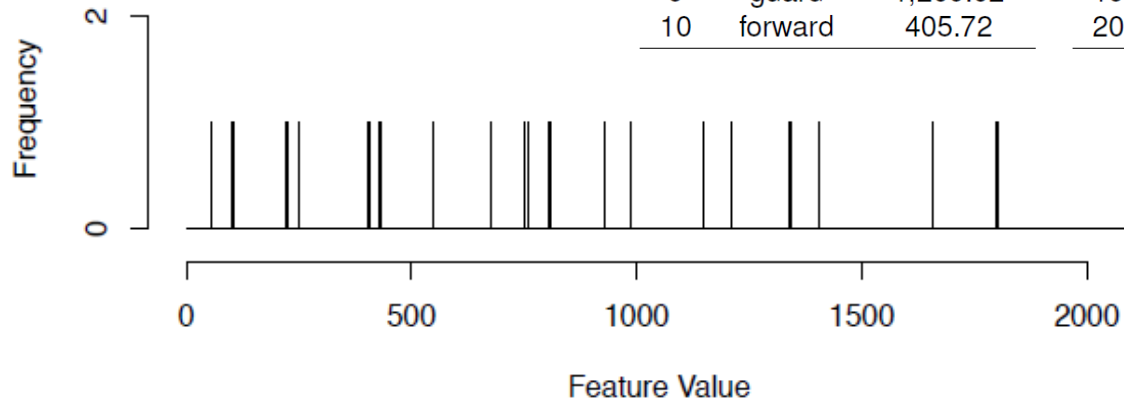
**Figure.** Visualizations of the categorical features in the motor instance claims fraud detection data using Bar Plot (Bar Chart)



# Histograms

- Bar plots don't work for continuous features

- E.g., A bar plot of the “training expenses” feature



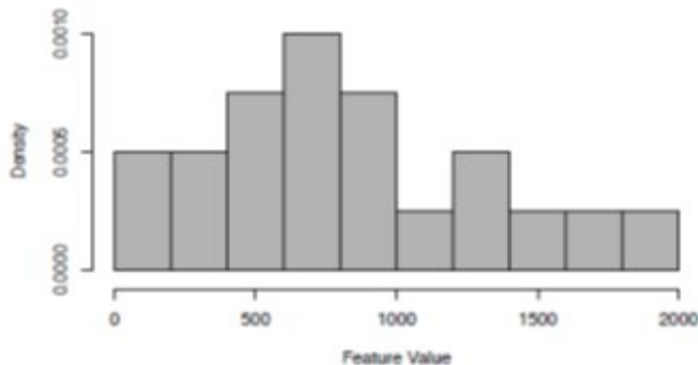
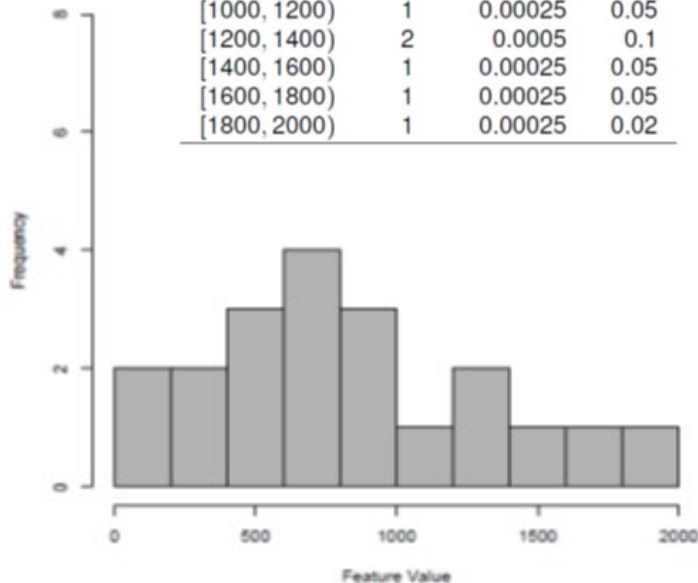
ID	Position	Training Expenses	ID	Position	Training Expenses
1	center	56.75	11	center	550.00
2	guard	1,800.11	12	center	223.89
3	guard	1,341.03	13	center	103.23
4	forward	749.50	14	forward	758.22
5	guard	1,150.00	15	forward	430.79
6	forward	928.30	16	forward	675.11
7	center	250.90	17	guard	1,657.20
8	guard	806.15	18	guard	1,405.18
9	guard	1,209.02	19	guard	760.51
10	forward	405.72	20	forward	985.41

- By dividing the range of a continuous feature into intervals, or bins, we can generate **histograms**

# Example: Histograms of Training Expenses Feature

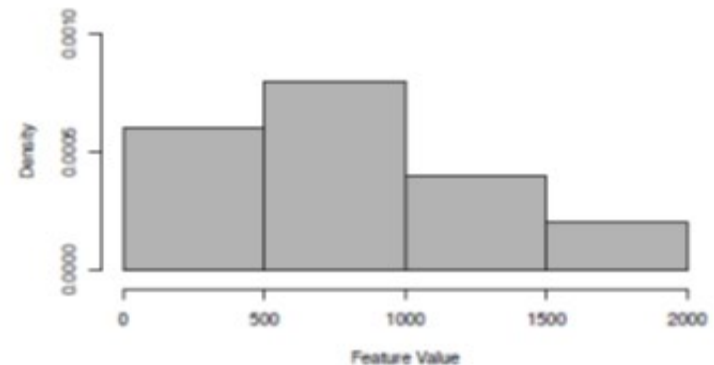
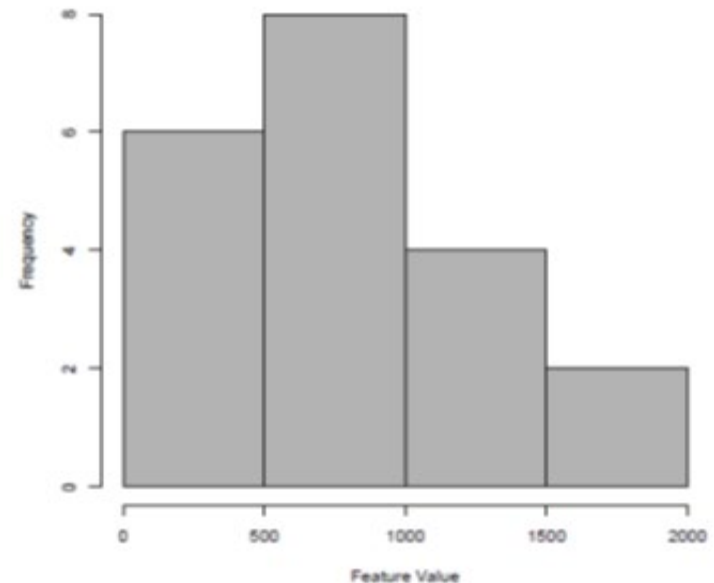
200 unit intervals

Interval	Count	Density	Prob
[0, 200)	2	0.0005	0.1
[200, 400)	2	0.0005	0.1
[400, 600)	3	0.00075	0.15
[600, 800)	4	0.001	0.2
[800, 1000)	3	0.00075	0.15
[1000, 1200)	1	0.00025	0.05
[1200, 1400)	2	0.0005	0.1
[1400, 1600)	1	0.00025	0.05
[1600, 1800)	1	0.00025	0.05
[1800, 2000)	1	0.00025	0.02

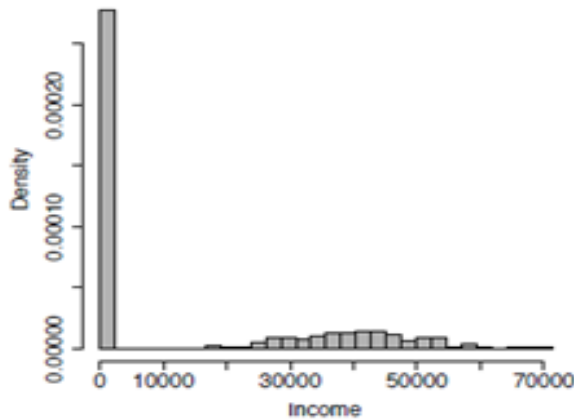


500 unit intervals

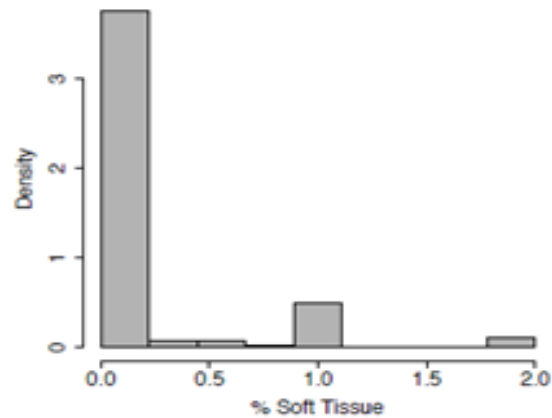
Interval	Count	Density	Prob
[0, 500)	6	0.0006	0.3
[500, 1000)	8	0.0008	0.4
[1000, 1500)	4	0.0004	0.2
[1500, 2000)	2	0.0002	0.1



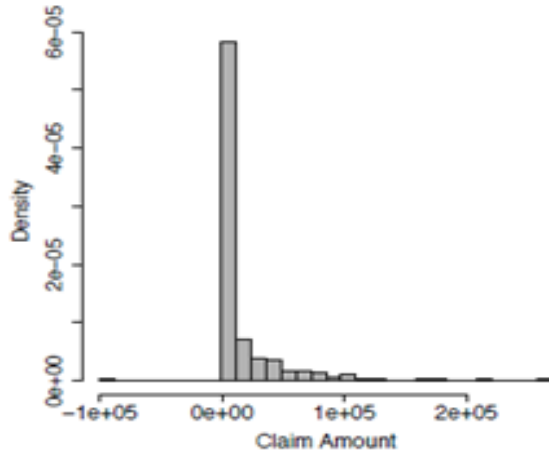
# Example: Histograms



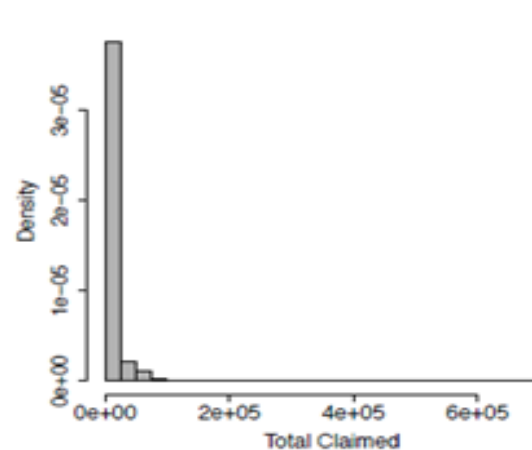
INCOME



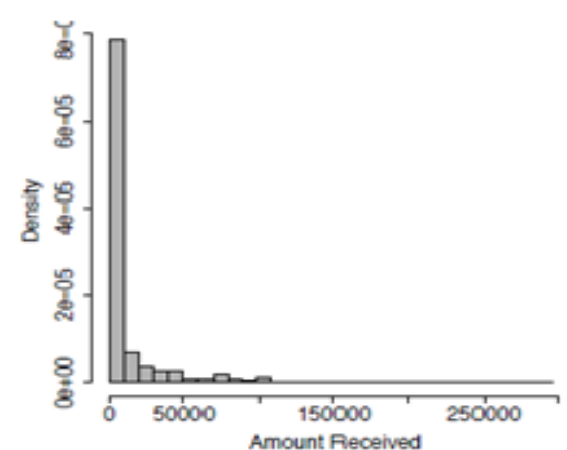
% SOFT TISSUE



CLAIM AMOUNT



TOTAL CLAIMED

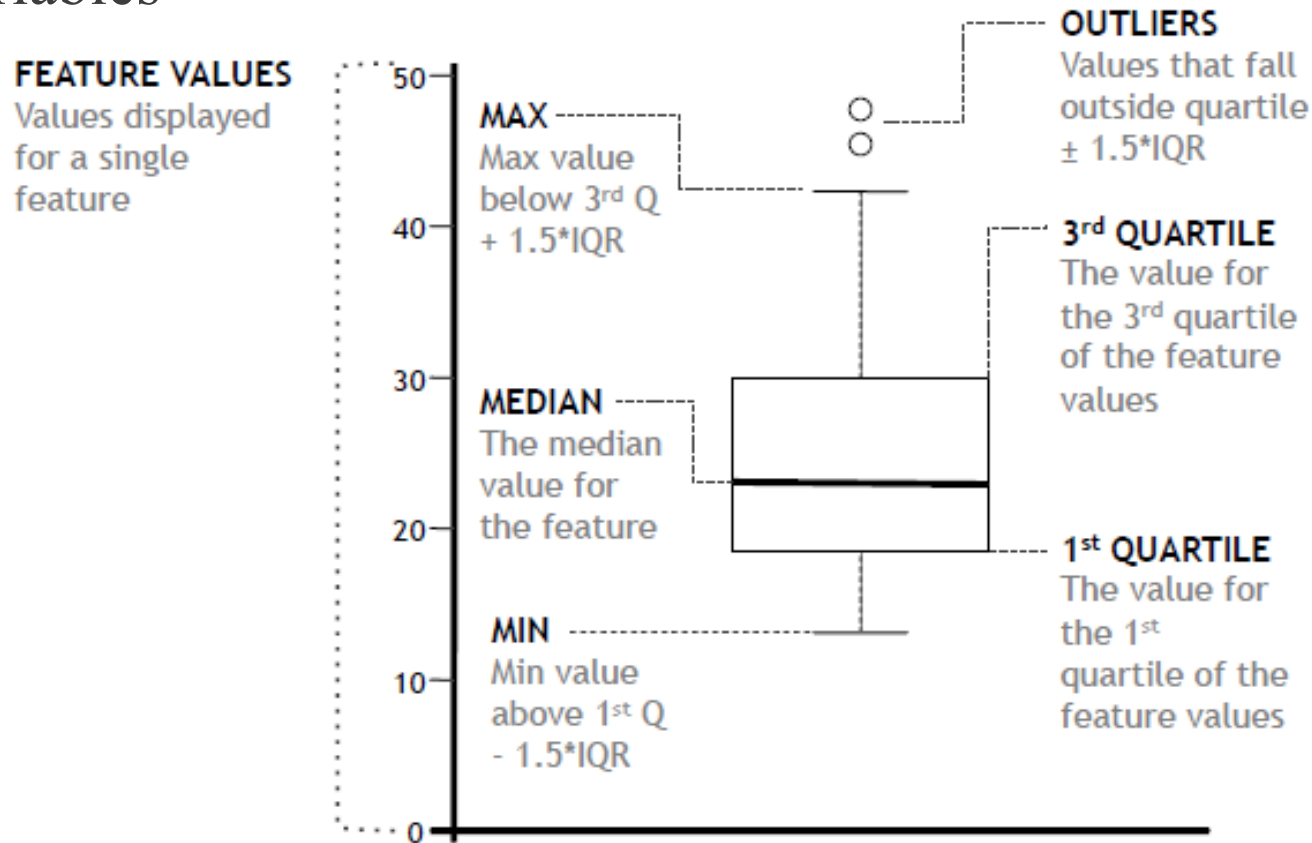


AMOUNT RECEIVED

**Figure.** Visualizations of the continuous features in the motor instance claims fraud detection data using Histogram

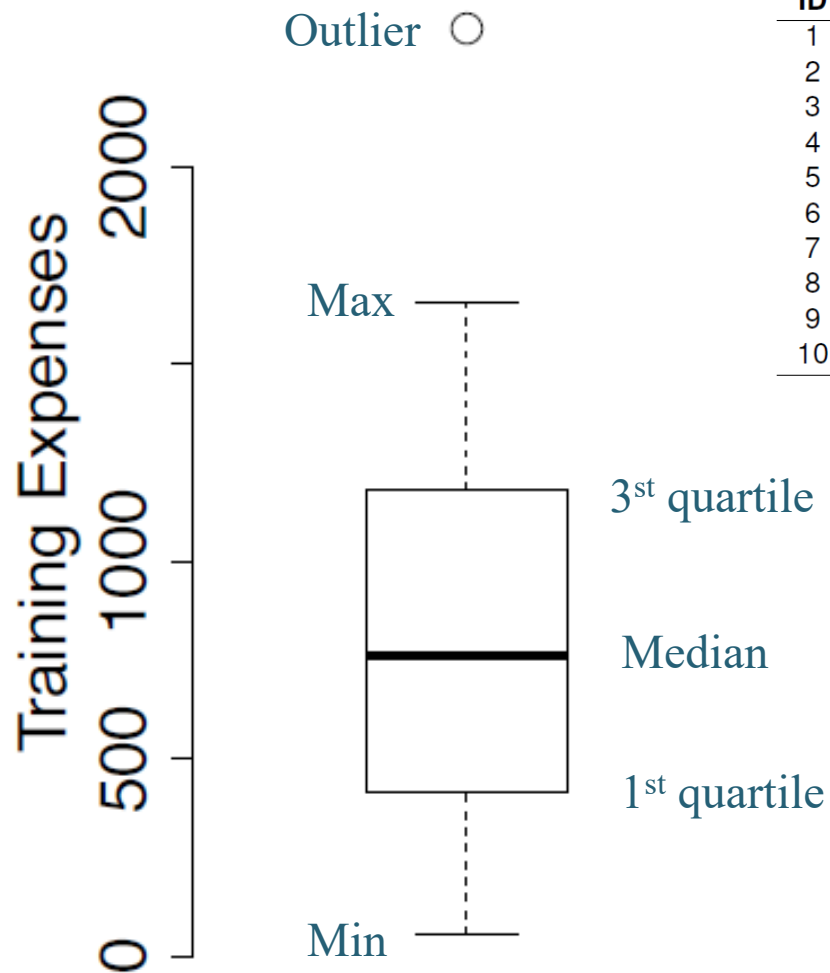
# Boxplot

- **Box plots** are another useful way of visualizing continuous variables



**Figure:** The structure of a box plot

# Example: Boxplot of Training Expenses Feature



Training Expenses			Training Expenses		
ID	Position	Training Expenses	ID	Position	Training Expenses
1	center	56.75	11	center	550.00
2	guard	1,800.11	12	center	223.89
3	guard	1,341.03	13	center	103.23
4	forward	749.50	14	forward	758.22
5	guard	1,150.00	15	forward	430.79
6	forward	928.30	16	forward	675.11
7	center	250.90	17	guard	1,657.20
8	guard	806.15	18	guard	1,405.18
9	guard	1,209.02	19	guard	760.51
10	forward	405.72	20	forward	985.41

**Table:** a basketball team dataset

# Outline

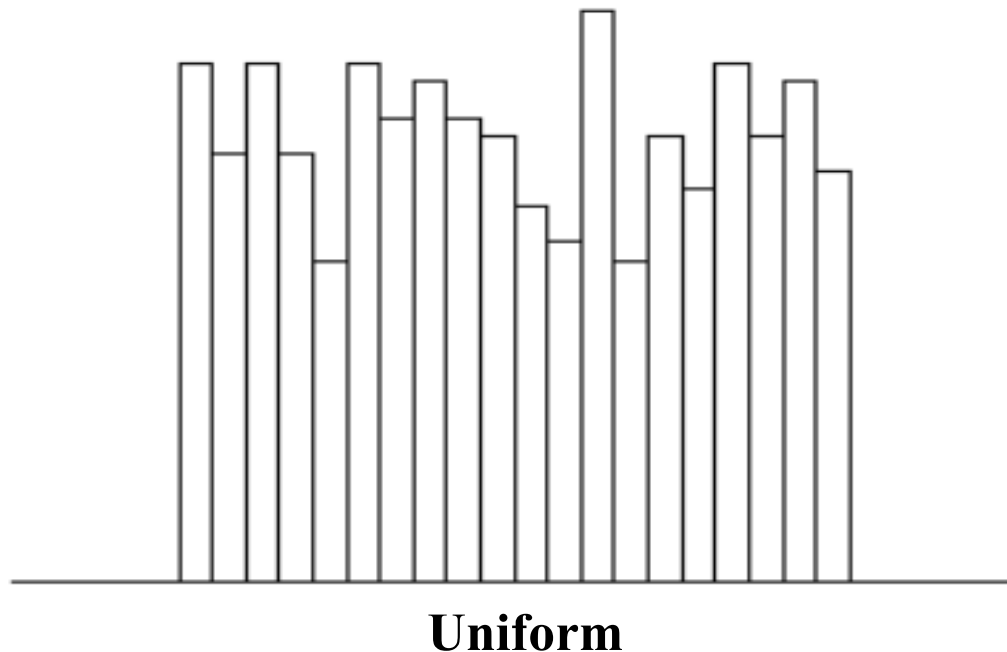
- Data Exploration
  - Descriptive Statistics
  - Data Visualization
  - ☞ **Common Histogram Shapes**
- Data Quality Issues
- Visualizing Relationships Between Features
- Measuring Covariance & Correlation
- Sampling
- Summary

# Common Shapes of Histograms

- Histograms can take on various shapes, each indicating different characteristics of the data distribution.
- Some common and well-known shapes of histograms are
  - Uniform Distribution
  - Normal Distribution (Bell Curve)
  - Skewed Right (Positive Skew)
  - Skewed Left (Negative Skew)
  - Bimodal Distribution
  - Multimodal Distribution
  - Exponential Distribution

# Uniform Distribution

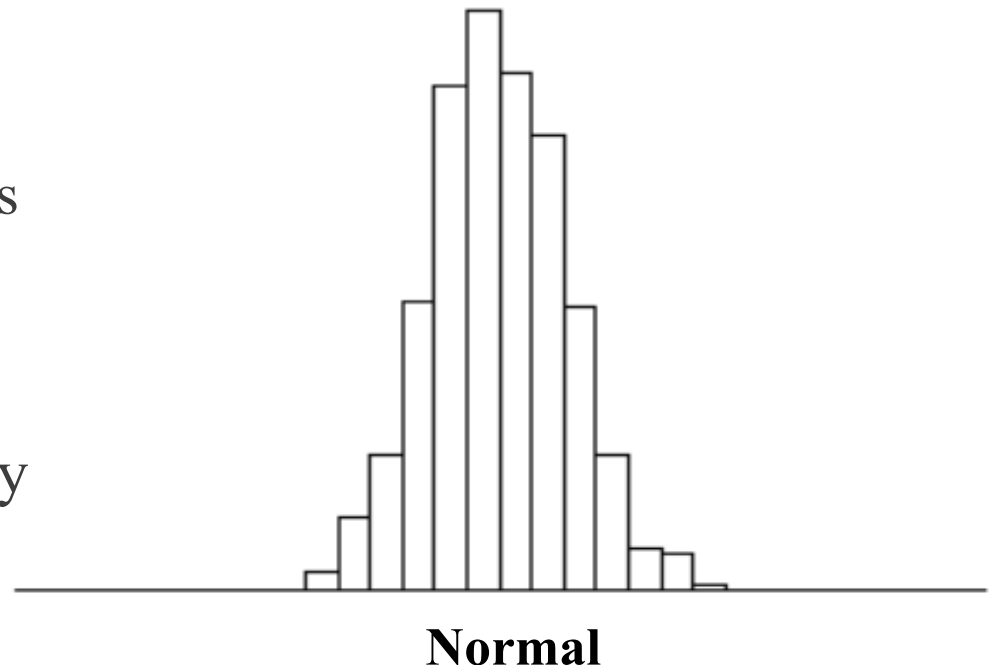
- A **uniform distribution** indicates that a feature is equally likely to take a value in any of the ranges present.
- All values have approximately the same frequency, resulting in a rectangular-shaped histogram.





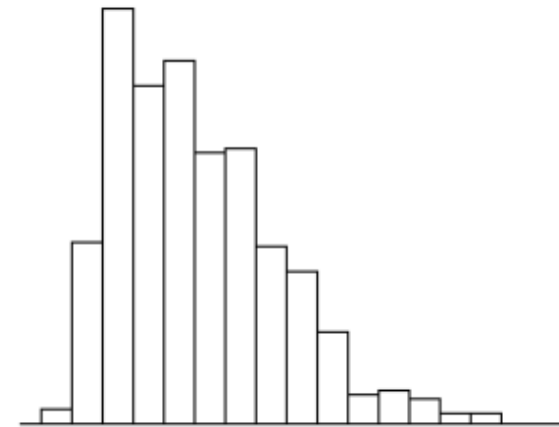
# Normal Distribution (Gaussian Distribution)

- A **normal distribution** (bell shape, bell curve) shows a strong tendency towards a central value and symmetrical variation to either side of this.
  - The mean, median, and mode of the data are all around the center of the distribution.
- A normal distribution shows naturally occurring phenomena.
  - **Example:** The height values of a randomly selected group of men or women.
- Many of modeling techniques work particularly well with normally distributed data

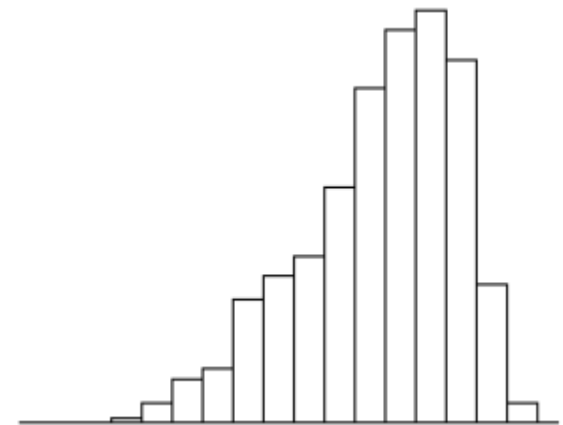


# Skewed Distribution

- **Skewed distributions have long tails.**
- **Skewed Right (Positive Skew):** A skewed right histogram has a tail that extends to the right side of the graph.
  - This indicates that there are some unusually high values in the dataset, and the mean is typically greater than the median.
  - **Example:** salary - central tendency, but there are usually a small number of people who are paid very large salaries.
- **Skewed Left (Negative Skew):** A skewed left histogram has a tail that extends to the left side of the graph.
  - This indicates that there are some unusually low values in the dataset, and the mean is typically smaller than the median.



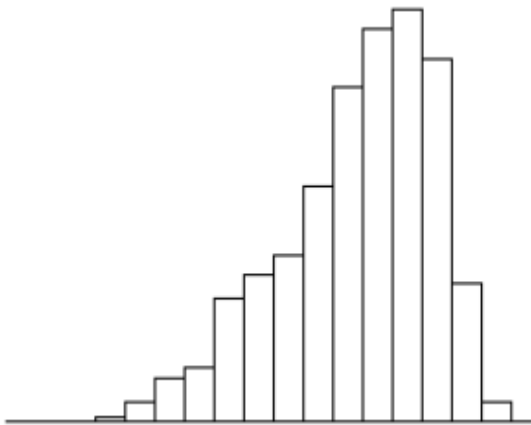
**Skewed right**



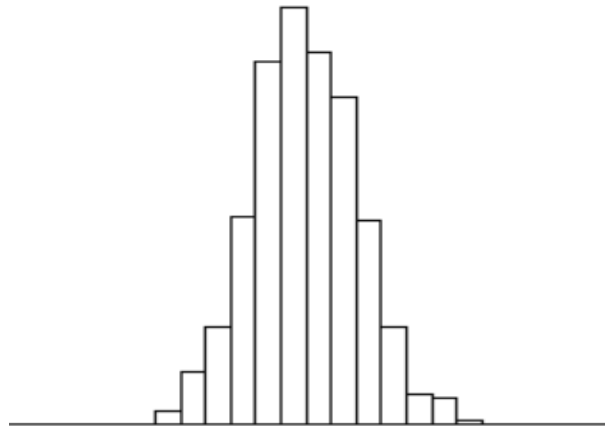
**Skewed left**

# Unimodal Distribution

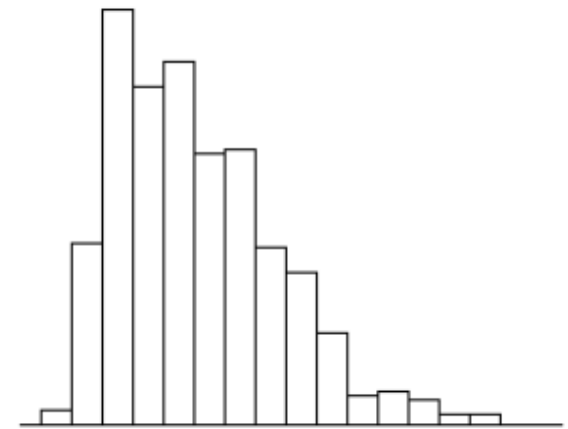
- **Unimodal distributions** have a single peak around the central tendency.



**Unimodal** (skewed right)



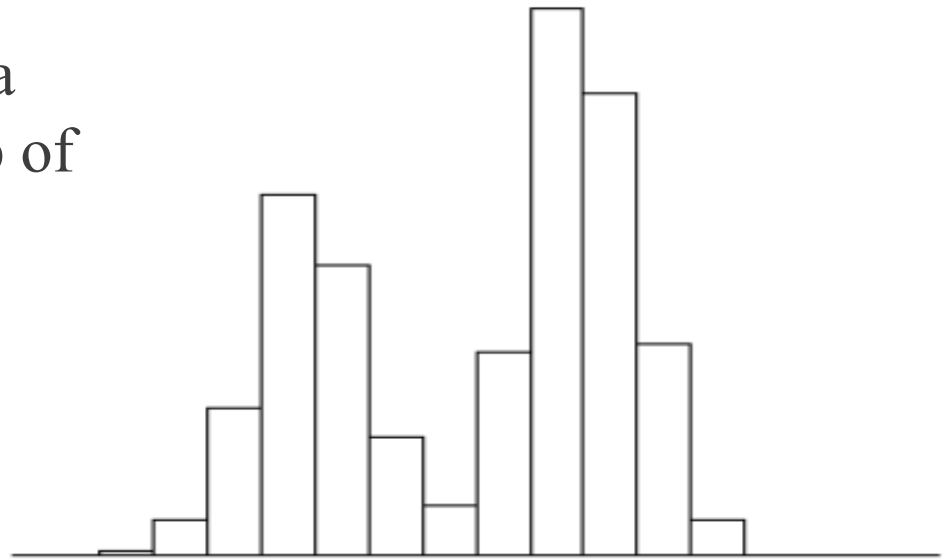
**Unimodal** (normal)



**Unimodal** (skewed left)

# Multimodal Distribution

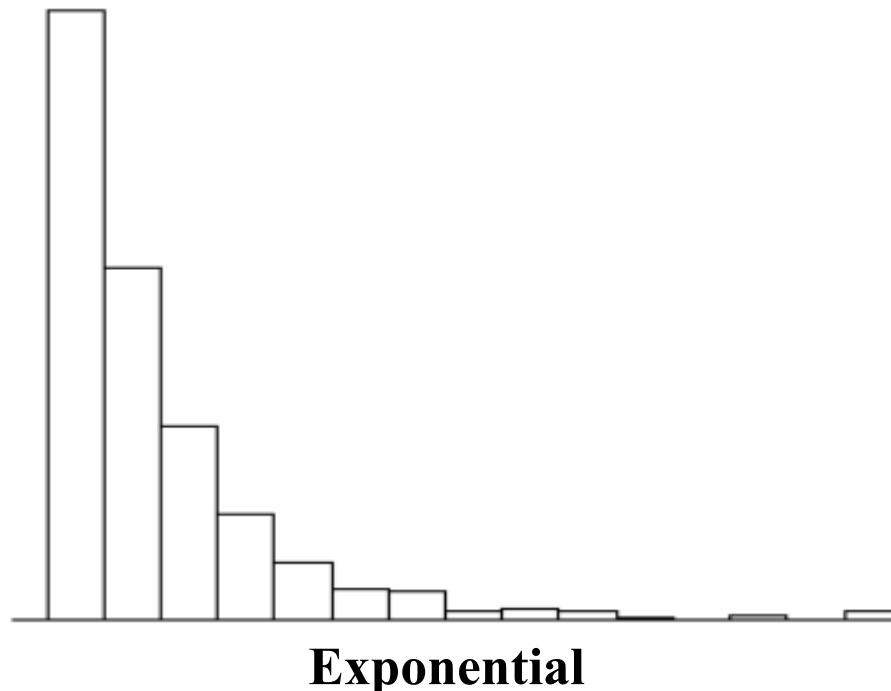
- A **multimodal distribution** results in a histogram with more than two distinct peaks, indicating multiple underlying distributions or processes in the data.
- A **bimodal distribution** shows two distinct peaks, suggesting that the data is derived from two separate underlying distributions or processes.
- **Example:** the height of a randomly selected group of men AND women.



**Multimodal (Bimodal)**

# Exponential Distribution

- In a feature following an **exponential distribution**, the likelihood of occurrence of a small number of low values is very high, but sharply diminishes as values increase, often resembling a decreasing curve.
- **Example:** the number of times a person has been married.



# Normal Distribution (Gaussian Distribution)

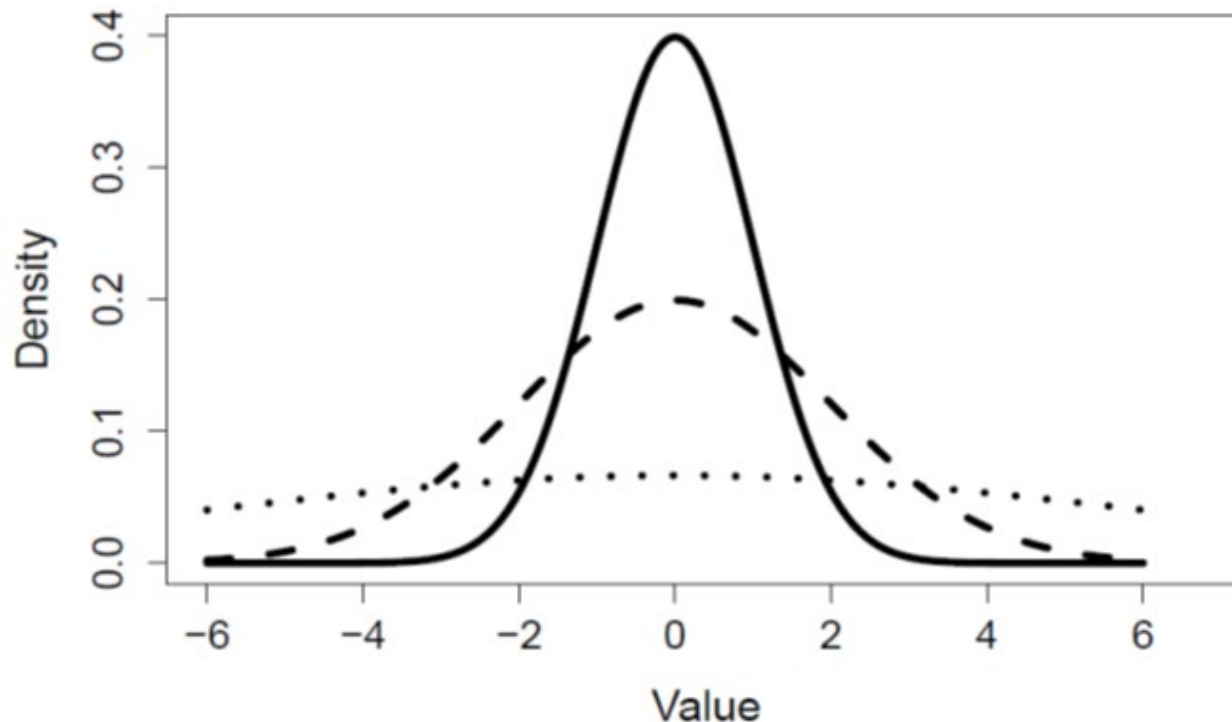
- A **normal distribution** (also known as **Gaussian distribution**) is important.
- The characteristics of a distribution is defined by a probability density function.
- The **probability density function for the normal distribution** is

$$N(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

, where  $x$  is any value,  $\mu$  and  $\sigma$  are parameters that define the shape of the distribution: the **population mean** and **population standard deviation**.

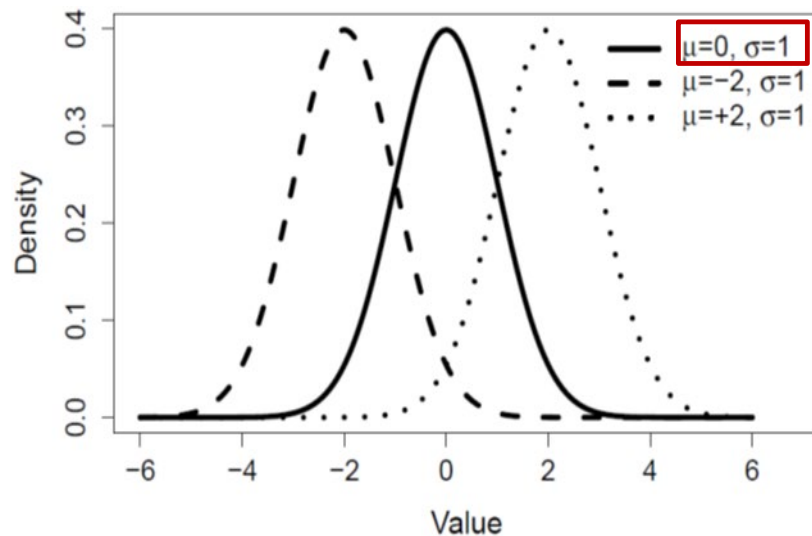
# Density Curve

- Given a probability density function, the **density curve** associated with a distribution can be plotted.
- The higher the curve for a particular value on the horizontal axis, the more likely that value is.

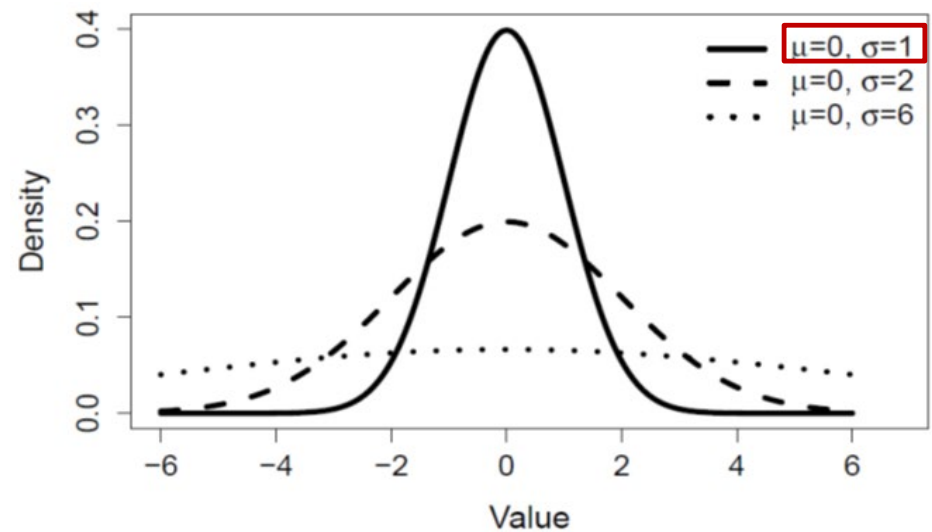


# Density Curves by Normal Distribution

- The density curve defined by a normal probability distribution,  $N(\mu, \sigma)$ , is symmetric around a single peak value.
  - The location of peak is defined by population mean  $\mu$
  - The height and slope of the curve is defined by population standard deviation  $\sigma$
  - Standard normal distribution:  $\mu=0, \sigma=1$



**Fig.** Three normal distribution with different means but identical standard deviations



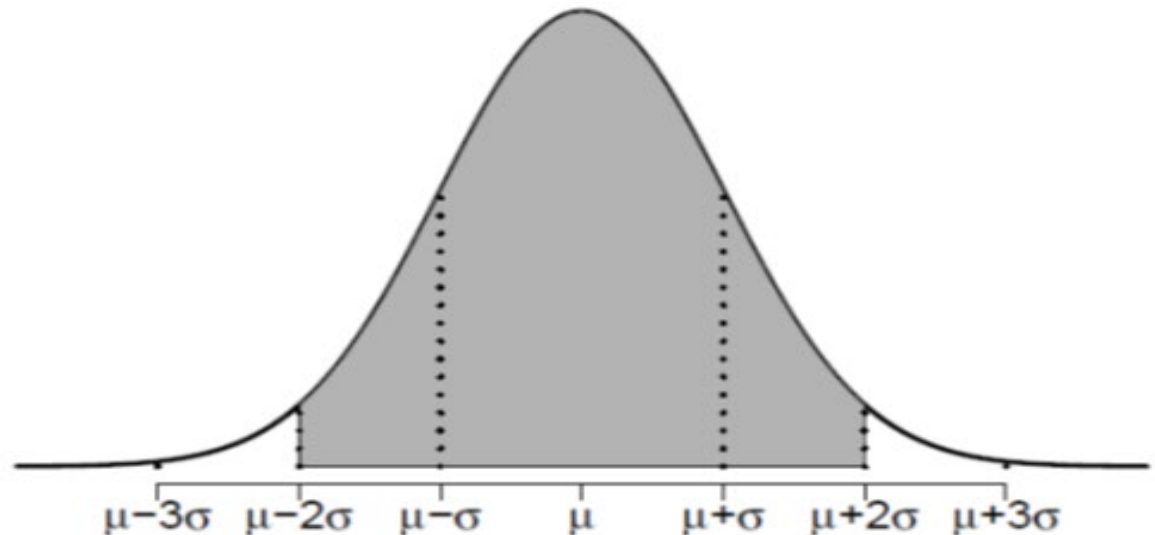
**Fig.** Three normal distributions with identical means but different standard deviations



# 68-95-99.7 Rule

- The **68 – 95 – 99.7 rule** is a useful characteristic of the normal distribution
- The rule states that approximately
  - 68% of the values in a sample that follows a normal distribution will be within one  $\sigma$  of  $\mu$
  - 95% of observations will be within two  $\sigma$  of  $\mu$
  - 99.7% of observations will be within three  $\sigma$  of  $\mu$

**Figure.** The grey region defines the area where 95% of observations are expected.



# Outline

- Data Exploration

- ☞ **Data Quality Issues**

- Missing Values
- Irregular Cardinality
- Outliers

- Visualizing Relationships Between Features

- Measuring Covariance & Correlation

- Sampling

- Summary

# Data Quality Issue

- A data quality issue is loosely defined as anything unusual about the data.
  - Data quality issues due to invalid data.
    - Errors in the process used to generate the data
    - May be easily corrected.
  - Data quality issues due to valid data
    - Error for a range of domain-specific reasons
    - No need to take any corrective action, but report the issue.
- The most common data quality issues are:
  - **missing values**
  - **irregular cardinality**
  - **outliers**

# Handling Missing Values

- **Approach 1: Drop any features** that have missing values
  - After examining the percentage of missing values for each feature, delete only features that are missing in excess of 60% of their values, in general.
- **Approach 2: Drop any instances** that are missing feature values
  - **Complete case analysis** - Delete any instances that are missing one or more feature values. → It may introduce data loss or a bias to data.
  - In general, only remove instances that are missing the value of target feature
- **Approach 3: Apply imputation**
  - **Imputation** replaces missing feature values with a plausible estimated value based on the feature values that are present.
  - The most common approach is to use a measure of the central tendency of that feature for the replacement
    - For continuous features, mean or median
    - For categorical features, mode

# Handling Irregular Cardinality

- Examine the **cardinality of each feature** - the number of distinct values presented for the feature
- When the cardinality for a feature does not match what we expect, a mismatch called an **irregular cardinality** is occurred.
- **Case 1: Features with a cardinality of 1.**
  - The features are not useful in building predictive models. Remove them.
- **Case 2: Continuous features with significantly lower cardinality than expected**
  - The feature is categorical but might be mistakenly identified as a continuous feature.
  - E.g., A “continuous” gender feature with a cardinality of 2, 1 for female and 0 for male. ➔ change the gender feature type to “categorical”

# Handling Irregular Cardinality (Cont.)

- **Case 3: Categorical features with a much higher cardinality than expected.**
  - **Example:** The cardinality of a “categorical” gender feature is 6
    - ➔ It might have multiple levels to represent the same thing, e.g., *male, female, m, f, M, and F*
- **Case 4: Categorical features with high cardinality**
  - It is worth to investigate features over 50 cardinality
- **Case 5: Features with a cardinality which is the same with the number of values.**
  - The feature with all different values (e.g., social security feature, student id) is not worth for building a model. Remove the feature.

# Handling Outliers

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

$$a_i = \begin{cases} \textit{lower} & \text{if } a_i < \textit{lower} \\ \textit{upper} & \text{if } a_i > \textit{upper} \\ a_i & \textit{otherwise} \end{cases}$$

- where  $a_i$  is a specific value of feature  $a$ ,
- $\textit{lower}$  and  $\textit{upper}$  are the lower and upper thresholds which are
  - set manually based on domain knowledge or
  - calculated from data, e.g.,
    - Method 1:  $\textit{upper} = 1\text{st quartile} - 1.5 \times \text{inter-quartile range}$   
 $\textit{lower} = 3\text{rd quartile} + 1.5 \times \text{inter-quartile range}$
    - Method 2:  $\textit{upper} = \text{mean} - 2 \times \text{standard deviation}$   
 $\textit{lower} = \text{mean} + 2 \times \text{standard deviation}$

# Outline

- Data Exploration
- Data Quality Issues
- ☞ **Visualizing Relationships Between Features**
- Measuring Covariance & Correlation
- Sampling
- Summary



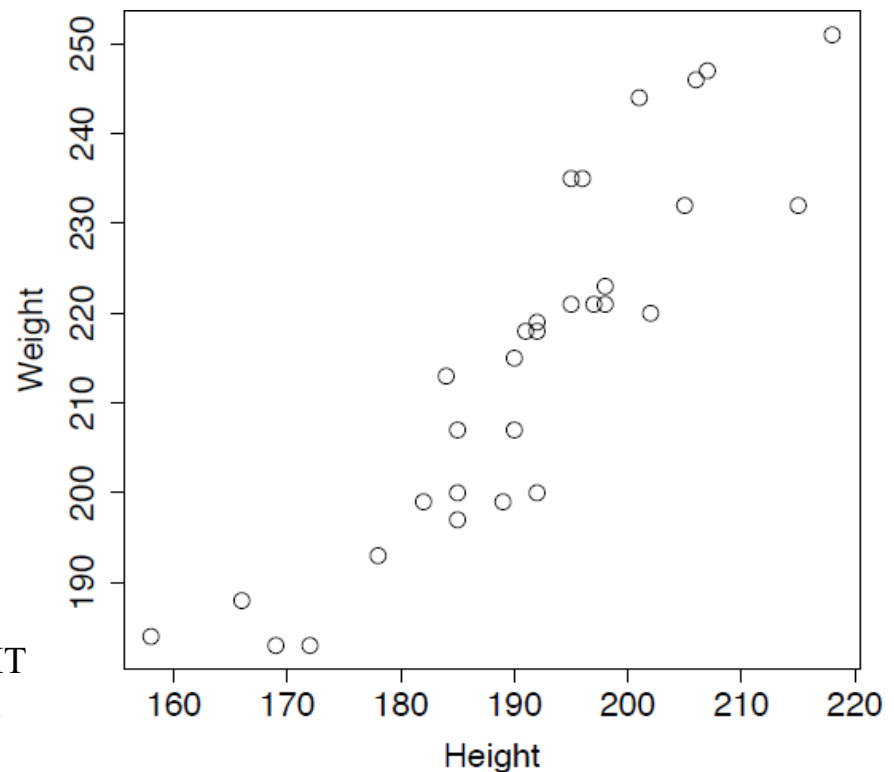
# Visualizing Relationships Between Features

- It is good to investigate the **relationships** between pairs of features to identify
  - which descriptive features might be useful for predicting a target feature, and
  - the pairs of descriptive features that are closely related.
- Cases of visualizing the relationships between
  - pairs of continuous features
  - pairs of categorical features
  - pairs including one categorical feature and one contiguous feature.
- Example: Professional basketball squad dataset

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

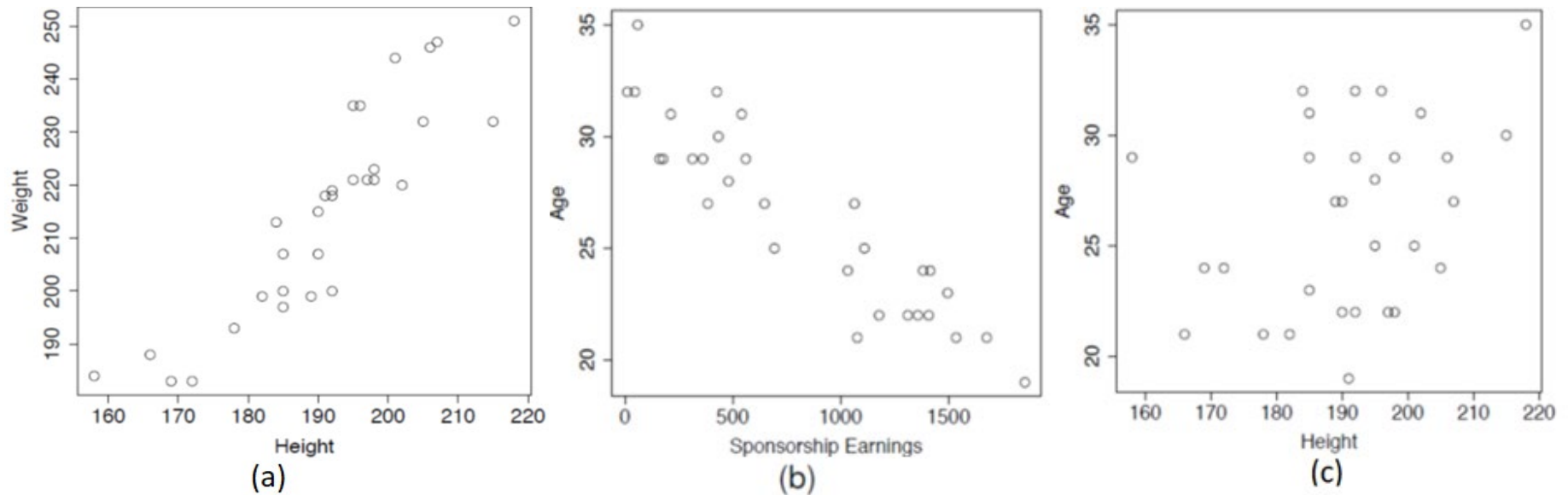
# Visualizing Relationships Between Contiguous Features

- A **scatter plot** is one of the most important tools in data visualization.
- 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 of the two features involved in the instance.



**Figure:** An example scatter plot showing the relationship between the HEIGHT and WEIGHT features from the professional basketball squad dataset

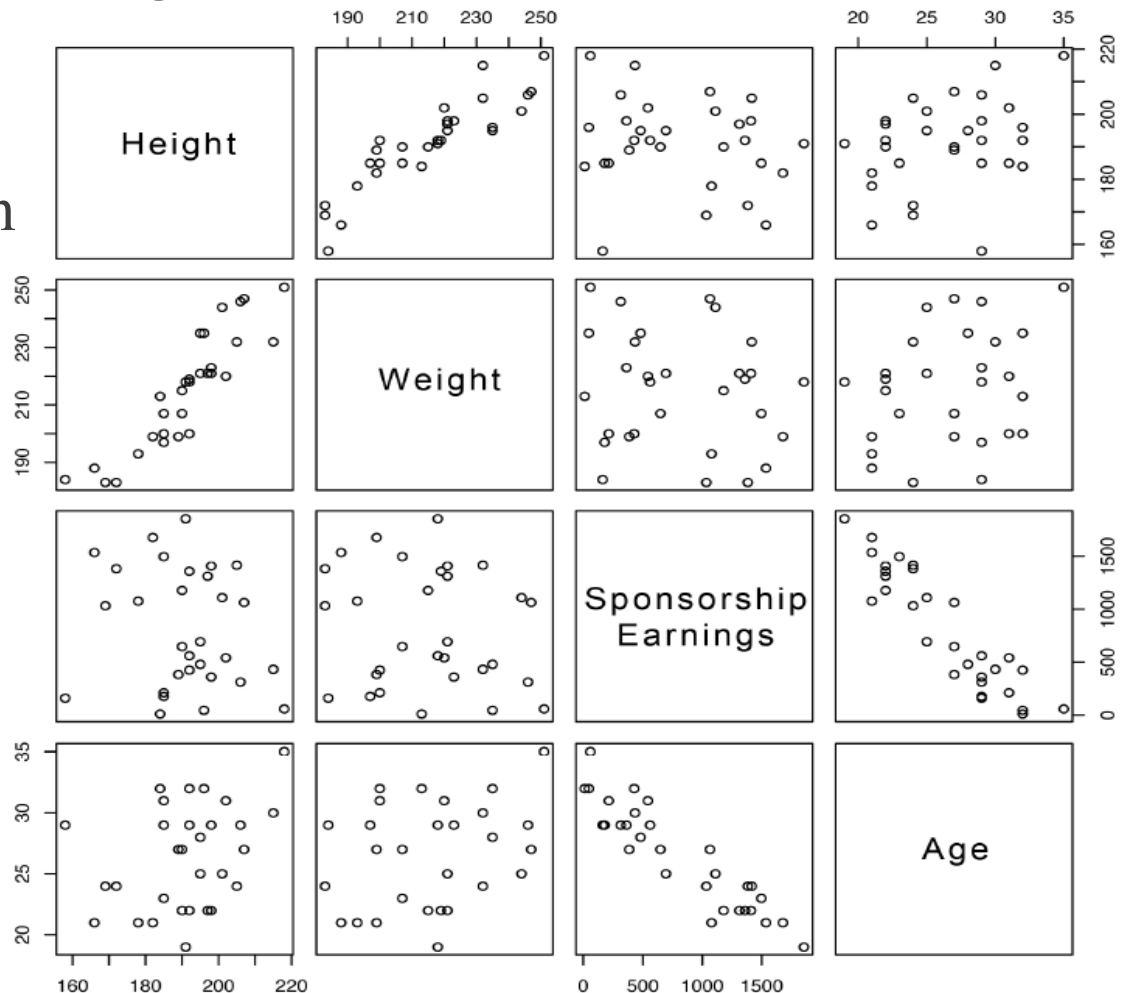
# Example Scatter Plots



**Figure:** Showing (a) the strong **positive covariance** (a strong, positive, linear relationship between the two features), (b) the strong **negative covariance**, and (c) the lack of strong covariance.

# Visualizing Relationships Between Contiguous Features

- A **scatter plot matrix** shows scatter plots for a whole collection of features arranged into a matrix.
- This is useful for exploring the relationships between groups of features

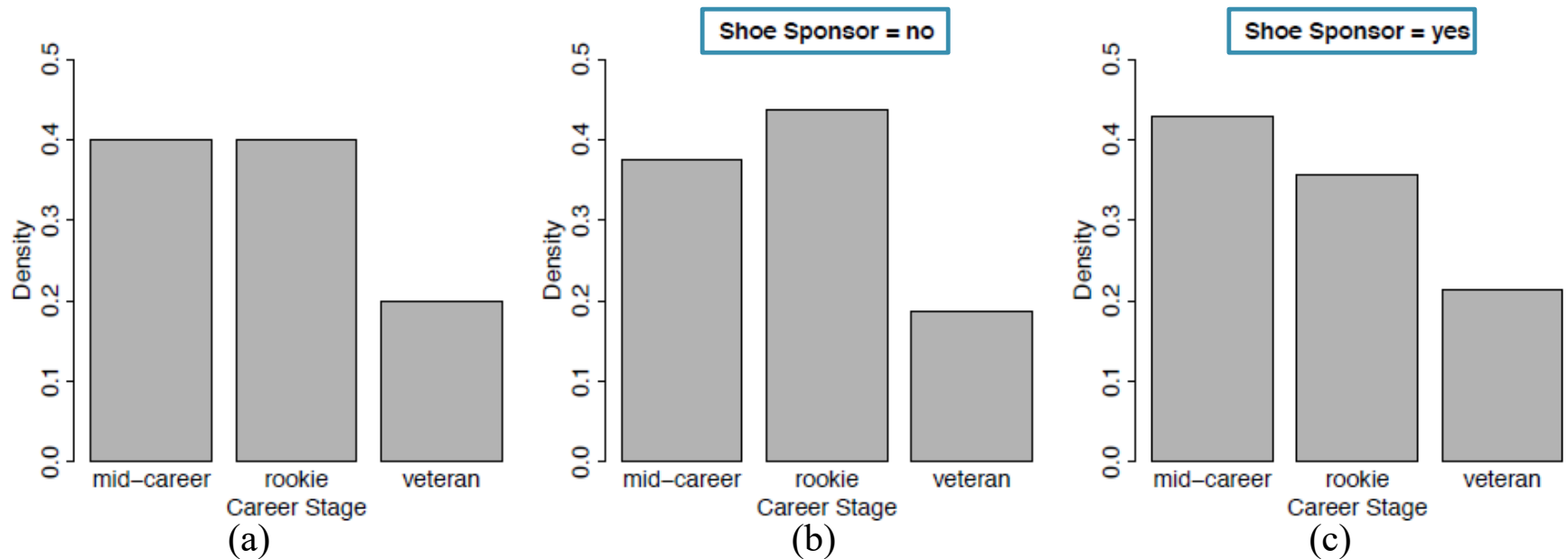


# Visualizing Relationships Between Categorical Features

- The simplest way to **visualize the relationship between two categorical features** is to use **a set of small multiple bar plots**.
- **Example:** for relationship between CAREER STAGE and SHOE SPONSOR, draw a bar plot of the first feature (CAREER STAGE) using only the instances in the dataset for which the second feature (SHOE SPONSOR) has that level (value).

# Example I

- **Bar plots** for relationship between CAREER STAGE and SHOE SPONSOR categorical features

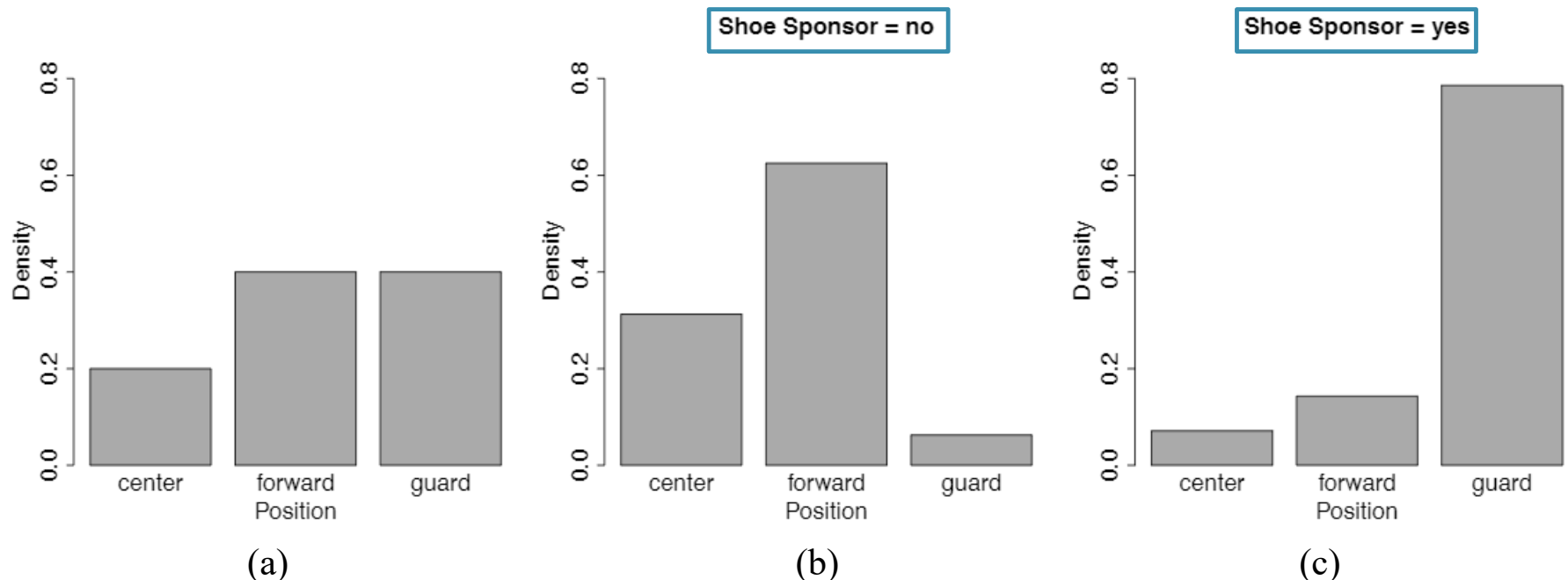


**Figure:** The bar plots show (a) the distribution of different levels of the career stage, and (b) (c) the distributions for those players **without and with a shoe sponsor**.

- All three plots show very similar distributions.
- No relationship exists between CAREER STAGE and SHOE SPONSOR – The players of any career stage are equally likely to have a shoe sponsor or not

## Example II

- Bar plots for **relationship between POSITION and SHOE SPONSOR**
  - All three plots show very different distributions.
  - There is a relationship between POSITION and SHOE SPONSOR – Players who play in the guard positions are much more likely to have a shoe sponsor than forwards or center.

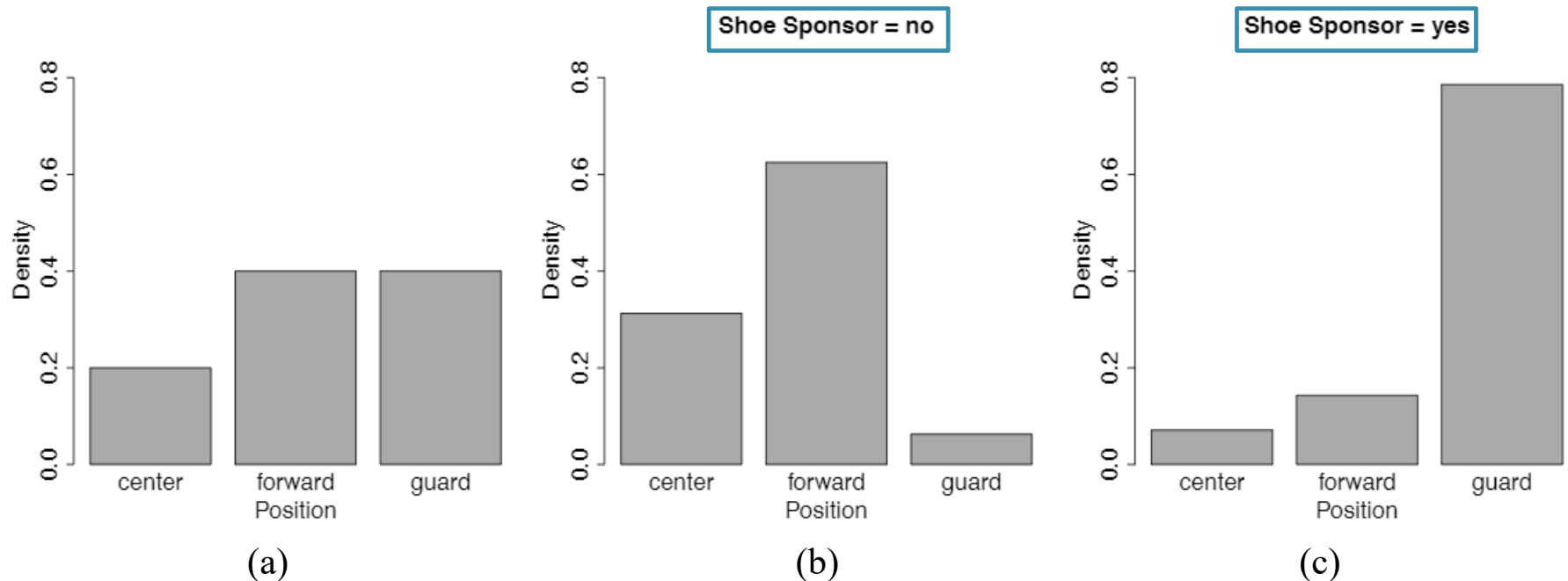


**Figure:** The bar plots show (a) the distribution of different levels of position, and (b) (c) the distributions for those players without and with a shoe sponsor.



## Example II

- Bar plots for relationship between POSITION and SHOE SPONSOR

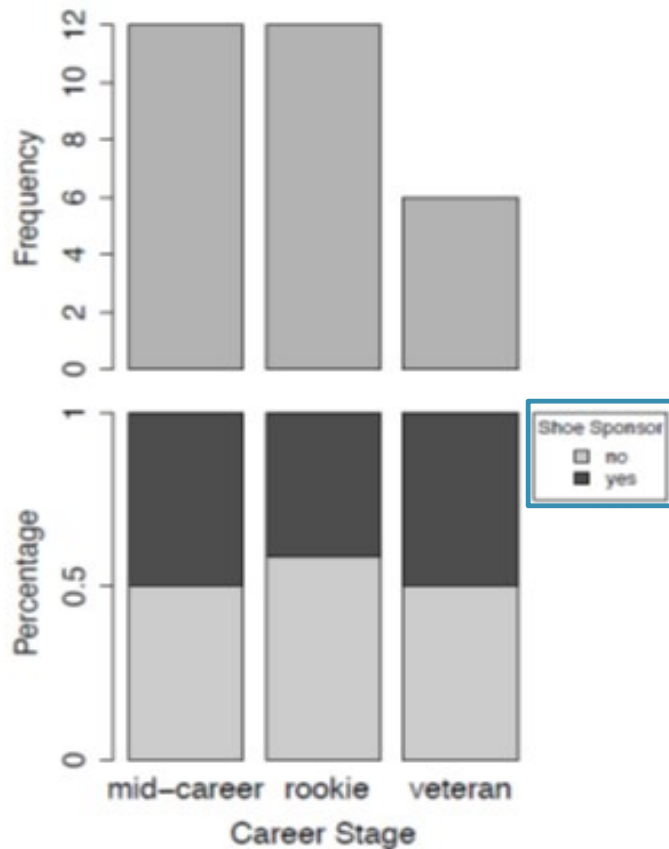


**Figure:** The bar plots show (a) the distribution of different levels of position, and (b) (c) the distributions for those players without and with a shoe sponsor.

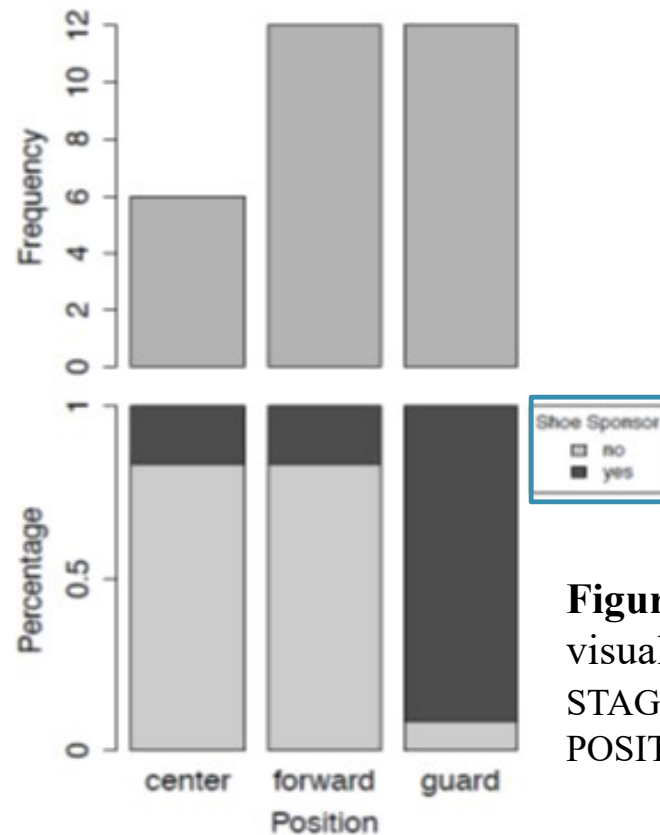
- All three plots show very different distributions.
- There is a relationship between POSITION and SHOE SPONSOR – Players who play in the guard positions are much more likely to have a shoe sponsor than forwards or center.

# Visualizing Relationships Between Categorical Features

- If the number of levels (distinct values) of one of the features is small (no more than three), **stacked bar plots** can be used as an alternative to the approach with a set of small multiple bar plots.



(a)



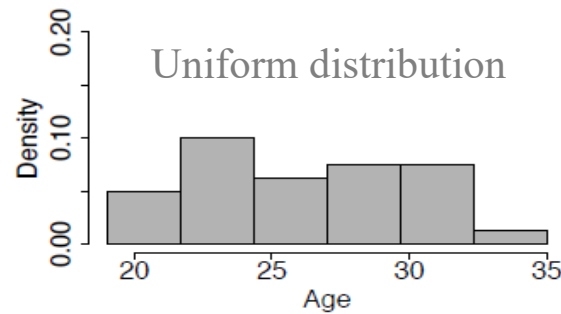
(b)

**Figure:** Stacked bar plot visualizations. (a) CAREER STAGE and SHOE SPONSOR (b) POSITION and SHOE SPONSOR

## Visualizing Relationship of Categorical and Contiguous Features

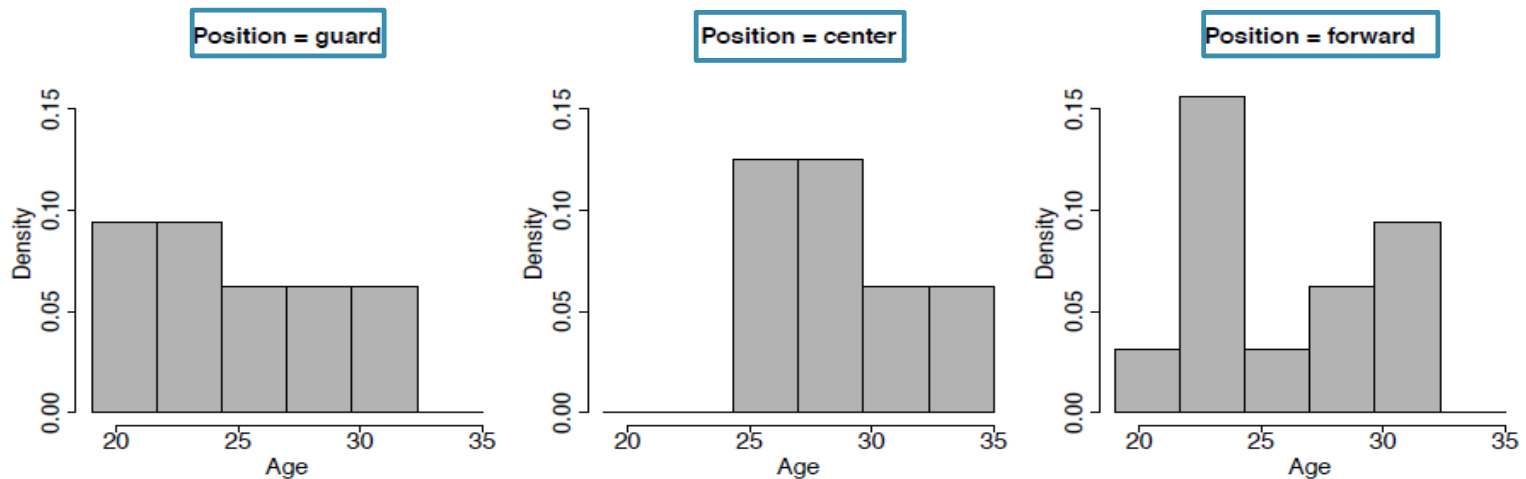
- To visualize the relationship between one continuous feature and one categorical feature, **a set of small multiple histograms** is used.
  - A histogram of the values of the continuous feature for each level (value) of the categorical feature

# Example I



**Figure:** Relationship between AGE and POSITION

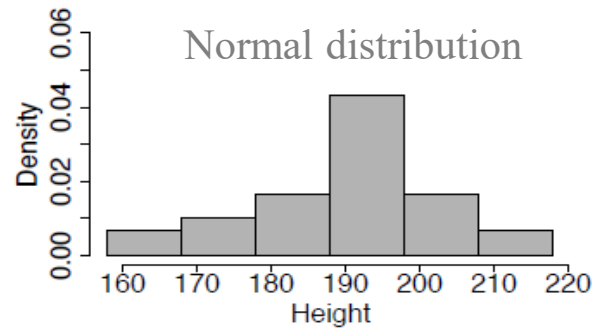
(a) Age



(b) Age and Position

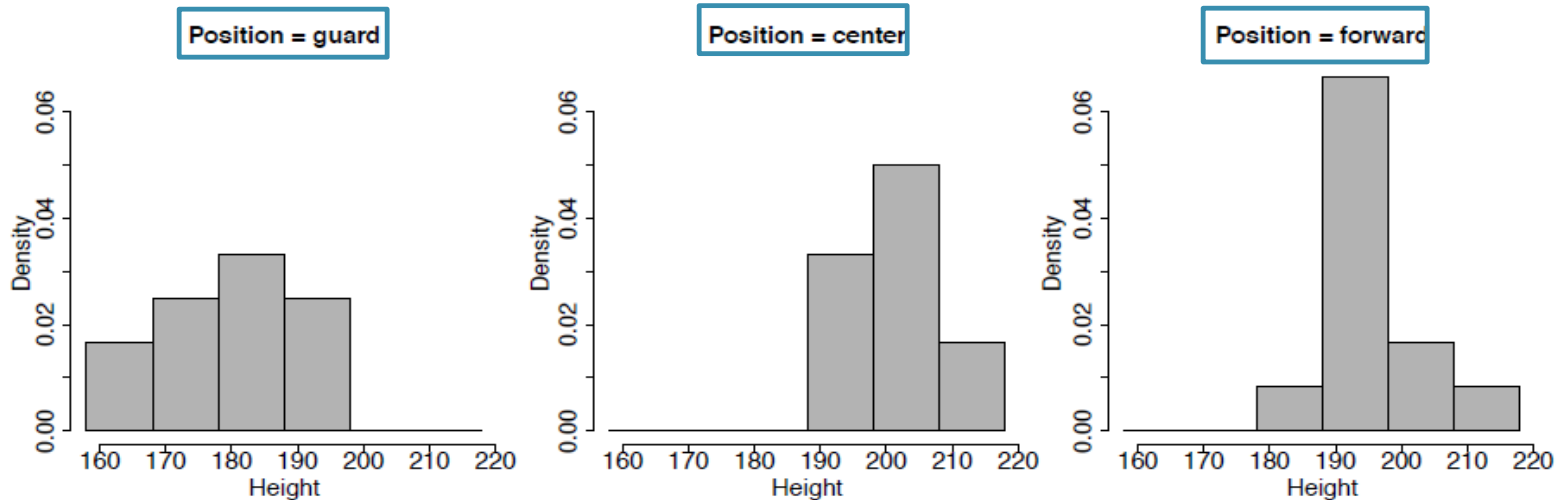
- A slight tendency for centers to be a little older than guards and forwards, but the relationship does not appear very strong, as each of the smaller histograms are similar to the overall uniform distribution of the AGE feature

## Example 2



**Figure:** Relationship between HEIGHT and POSITION

(a) Height



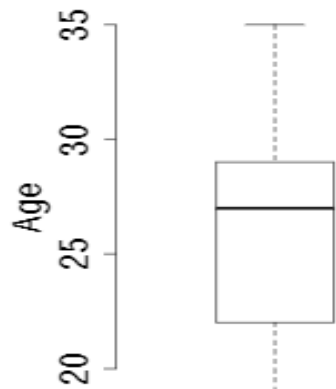
(b) Height and Position

- Centers tend to be taller than forwards. Forwards tend to be taller than guards.

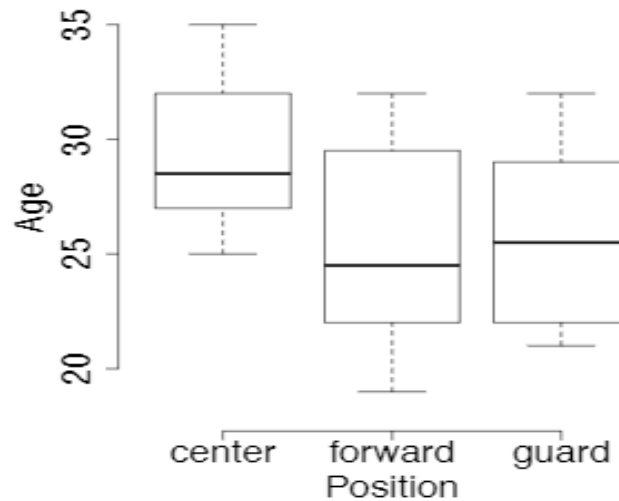
## Visualizing Relationship of Categorical and Contiguous Features

- 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.

# Example: A Set of Box Plots

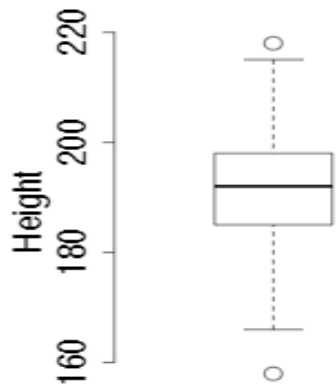


(a) AGE

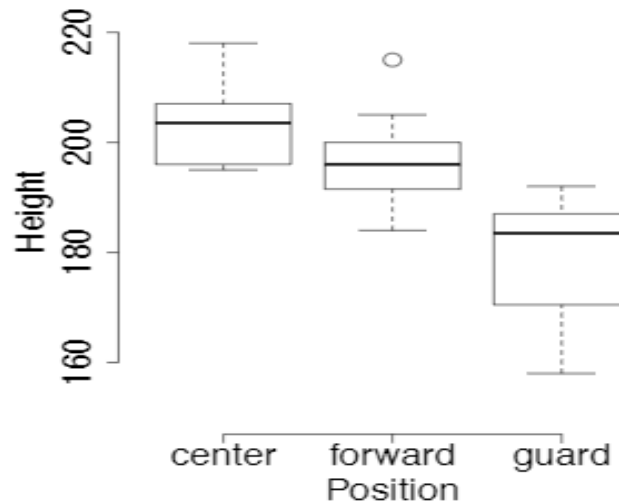


(b) AGE and POSITION

**Figure:** Relationship between AGE and POSITION



(c) HEIGHT



(d) HEIGHT and POSITION

**Figure:** Relationship between HEIGHT and POSITION

# Outline

- Data Exploration
- Data Quality Issues
- Visualizing Relationships Between Features
- ☞ **Measuring Covariance and Correlation**
- Sampling
- Summary



# 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**.

# Covariance

- For two features,  $a$  and  $b$ , in a dataset of  $n$  instances, the **sample covariance** between  $a$  and  $b$  is

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

- 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  $b$
- **Range of covariance values** is  $[-\infty, \infty]$ 
  - **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.

# Example

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

	HEIGHT		WEIGHT		$(h - \bar{h}) \times$	AGE		$(h - \bar{h}) \times$
ID	(h)	$h - \bar{h}$	(w)	$w - \bar{w}$	$(w - \bar{w})$	(a)	$a - \bar{a}$	$(a - \bar{a})$
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
<b>Mean</b>	191.1		215.0			26.4		
<b>Std Dev</b>	13.6		19.8			4.2		
<b>Sum</b>					7,009.9			570.8

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

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

# Correlation

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

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

- where  $\text{cov}(a, b)$  is the covariance between features  $a$  and  $b$  and  $sd(a)$  and  $sd(b)$  are the standard deviations of  $a$  and  $b$  respectively.
- **Range of correlation values is  $[-1, 1]$** 
  - **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**.

# Example

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

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

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

# Covariance/Correlation with Multiple Features

- Tools useful for exploring the relationship of each pair of multiple continuous features are the **covariance matrix** and the **correlation matrix**.

# Covariance Matrix

- A **covariance matrix** (also known as auto-covariance matrix, dispersion matrix, variance matrix, or variance–covariance matrix) , usually denoted as  $\Sigma$ , between a set of continuous features,  $\{a, b, \dots, z\}$  is given as

$$\sum_{\{a,b,\dots,z\}} = \begin{bmatrix} \text{var}(a) & \text{cov}(a, b) & \dots & \text{cov}(a, z) \\ \text{cov}(b, a) & \text{var}(b) & \dots & \text{cov}(b, z) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(z, a) & \text{cov}(z, b) & \dots & \text{var}(z) \end{bmatrix}$$

- A covariance matrix is a square matrix giving the covariance between each pair of elements of a given feature set.
- Any covariance matrix is symmetric and its main diagonal contains variances (i.e., the covariance of each element with itself).

# Correlation Matrix

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

$$\text{correlation matrix}_{\{a,b,\dots,z\}} = \begin{bmatrix} \text{corr}(a, a) & \text{corr}(a, b) & \dots & \text{corr}(a, z) \\ \text{corr}(b, a) & \text{corr}(b, b) & \dots & \text{corr}(b, z) \\ \vdots & \vdots & \ddots & \vdots \\ \text{corr}(z, a) & \text{corr}(z, b) & \dots & \text{corr}(z, z) \end{bmatrix}$$



# Example

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

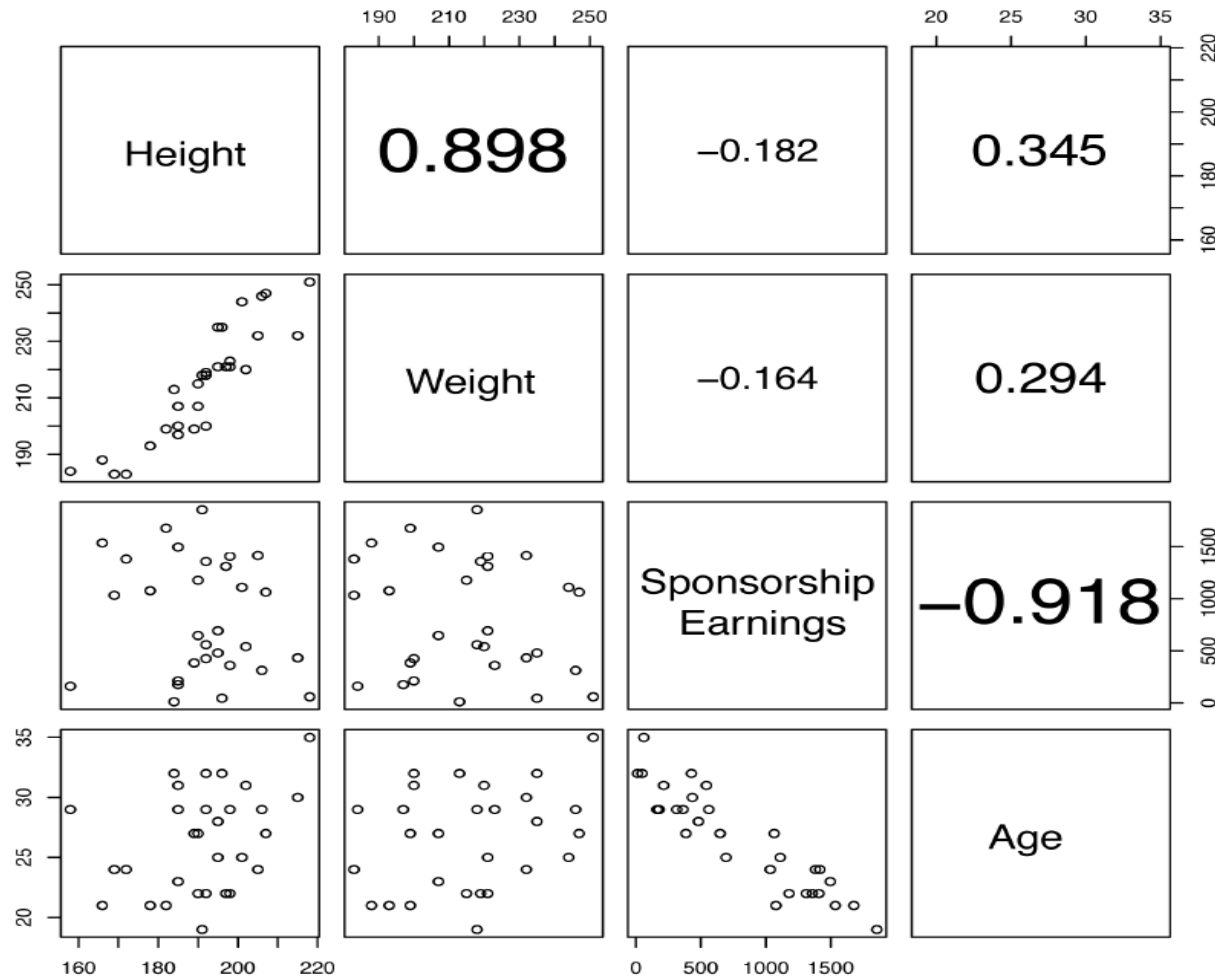
$$\sum_{\langle \text{Height}, \text{Weight}, \text{Age} \rangle} = \begin{bmatrix} 185.128 & 241.72 & 19.7 \\ 241.72 & 392.102 & 24.469 \\ 19.7 & 24.469 & 17.697 \end{bmatrix}$$

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

$$\text{correlation matrix}_{\langle \text{Height}, \text{Weight}, \text{Age} \rangle} = \begin{bmatrix} 1.0 & 0.898 & 0.345 \\ 0.898 & 1.0 & 0.294 \\ 0.345 & 0.294 & 1.0 \end{bmatrix}$$

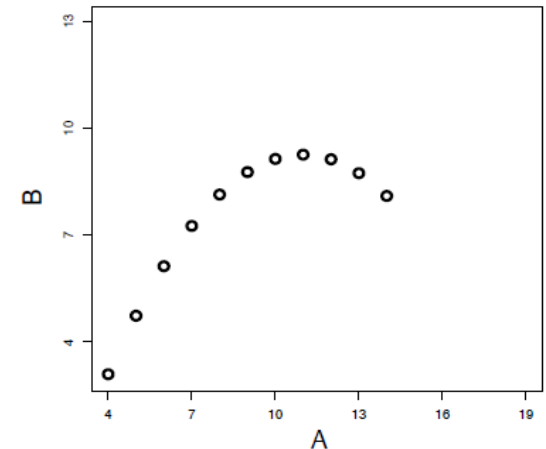
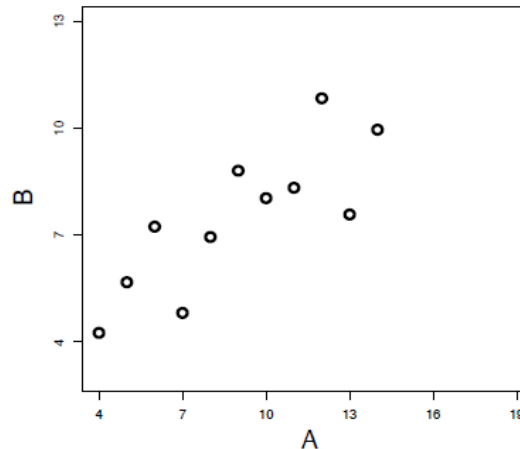
# Scatter Plot Matrix with Correlation Coefficients

- The **scatter plot matrix** (SPLOM) is really a visualization of the correlation matrix.

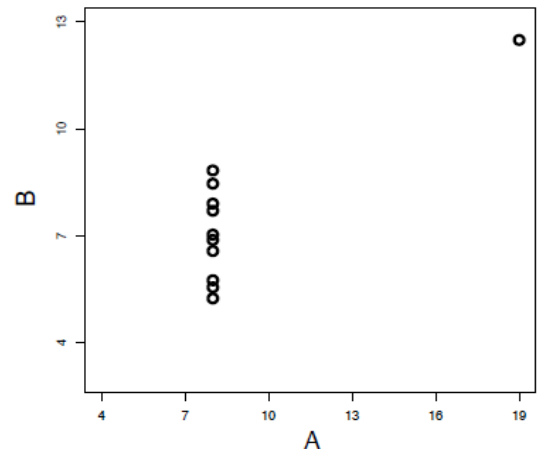
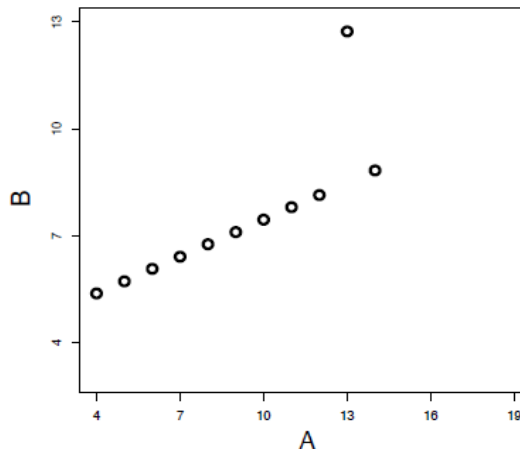


# Comment for Correlation

- Correlation is a good measure of the relationship between two continuous features, but it is not by any means perfect.
- Also notice that *correlation does not necessarily imply causation.*



The same correlation value, 0.816, for all four cases



# Outline

- Data Exploration
- Data Quality Issues
- Visualizing Relationships Between Features
- Measuring Covariance & Correlation
- ☞ **Sampling**
- Summary

# Sampling

- Sometimes the dataset we have is so large that we do not use all the data available to us 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

- **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, as the sample will be affected by any ordering of the original dataset.
- We recommend that top sampling be avoided.

# Random Sampling

- A 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

- **Stratified sampling** is a sampling method that ensures that the relative frequencies of the levels (distinct values) of a specific *stratification feature* are maintained in the sampled dataset.
- **To perform stratified sampling:**
  - the instances in a dataset are divided into homogeneous groups (called *strata*) based on specific characteristics (e.g., race, gender, location) where each group contains only instances that have a particular value (level) for the stratification feature
  - $s\%$  of the instances in each strata are randomly selected
  - these selections are combined to give an overall sample of  $s\%$  of the original dataset.
- Notice that each strata will contain a different number of instances. So by sampling on a percentage basis from each strata, the number of instances taken from each strata will be proportional to the number of instances in each strata.



# Under (or Over)-Sampling

- 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.
  - E.g., Create a sample in which the values of a particular categorical feature are represented equally, rather than with whatever distribution they have in the original dataset.
- To do this, we can use **under-sampling** or **over-sampling**

# Under-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

- **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.

# Outline

- Data Exploration
- Data Quality Issues
- Visualizing Relationships Between Features
- Measuring Covariance & Correlation
- Sampling
- ☞ **Summary**

# Summary

- The key outcomes of the **data exploration** process are that the practitioner should
  1. Have gotten to know the features of the data, especially their **central tendencies, variations, and distributions**.
  2. Have gotten to know simple data visualization
  3. Have identified any **data quality issues**, in particular **missing values, irregular cardinality, and outliers**.
  4. Have corrected any data quality issues due to **invalid data**.
- **Be aware of the relationships between features** in the data using **visualization** and **covariance & correlation measures**.
- Know various **sampling methods**.