Fundamentals of Machine Learning for Predictive Data Analytics

Chapter 3: Data Exploration

Sections 3.1, 3.2, 3.3, 3.4

John Kelleher and Brian Mac Namee and Aoife D'Arcy

john.d.kelleher@dit.ie brian.macnamee@ucd.ie aoife@theanalyticsstore.com

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** and **variation**.
- 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.

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

(a) Continuous Features

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

(b) Categorical Features

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

Case Study: Motor Insurance Fraud

The following slides show a portion of the ABT that has been developed for the motor insurance claims fraud detection.

A portion of the ABT developed for this solution is shown first.

Table: Portions of the ABT for the motor insurance claims fraud detection problem.

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

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

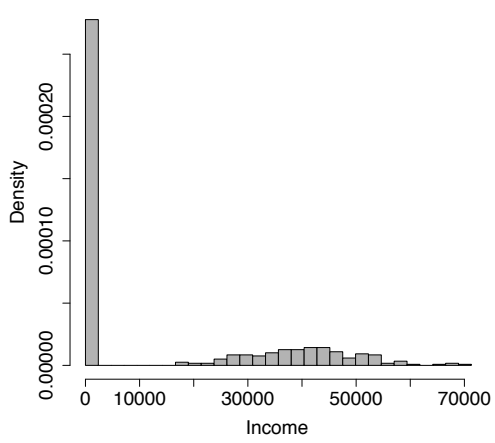
(a) Continuous Features

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

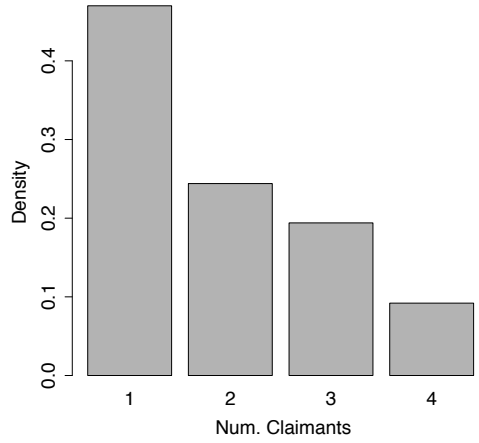
Table: A data quality report for the motor insurance claims fraud detection ABT.

(a) Categorical Features

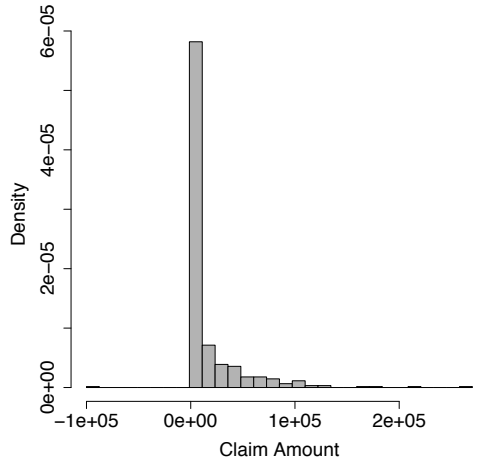
Feature	Count	% Miss.	Card.	Mode	Mode Freq.	Mode %	2^{nd} Mode	2^{nd} Mode Freq.	2^{nd} Mode %
INSURANCE TYPE	500	0.0	1	CI	500	1.0	–	–	–
MARITAL STATUS	500	61.2	4	Married	99	51.0	Single	48	24.7
INJURY TYPE	500	0.0	4	Broken Limb	177	35.4	Soft Tissue	172	34.4
HOSPITAL STAY	500	0.0	2	No	354	70.8	Yes	146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



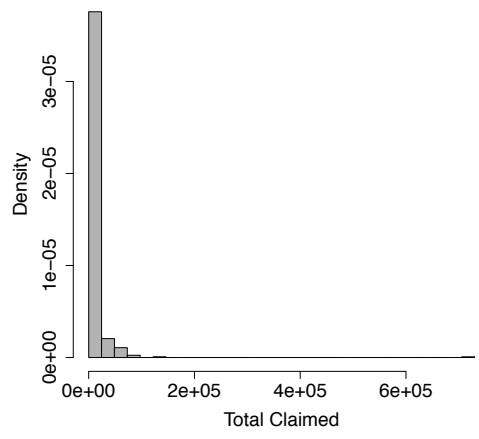
(a) INCOME



(b) NUM CLAIMANTS

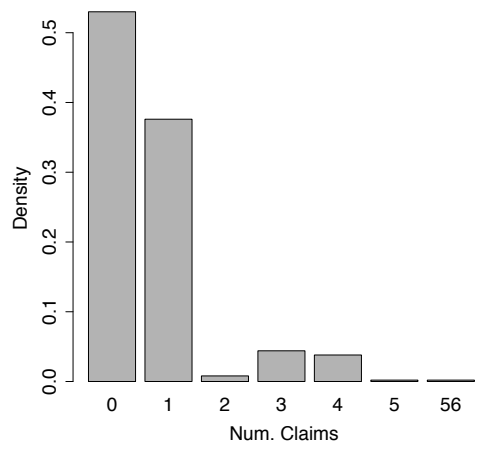


(c) CLAIM AMOUNT

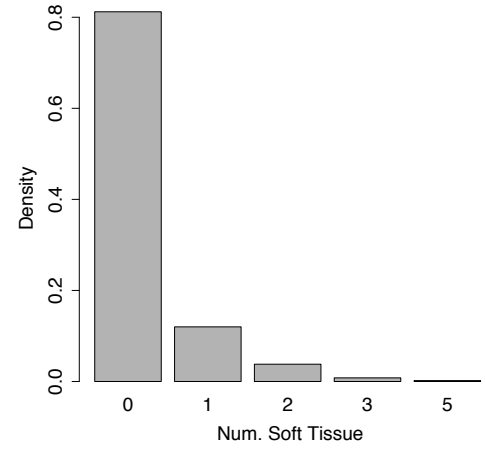


(d) TOTAL CLAIMED

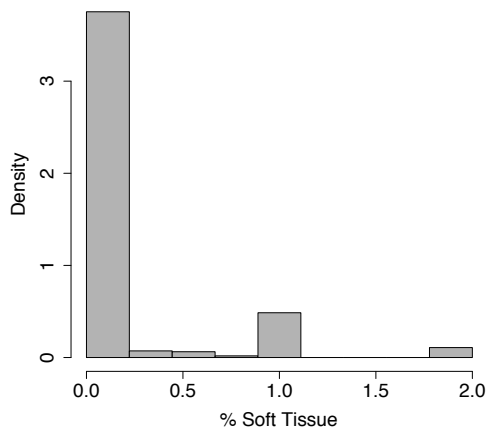
Figure: Visualizations of the continuous and categorical features in the motor insurance claims fraud detection ABT in Table 2 [7].



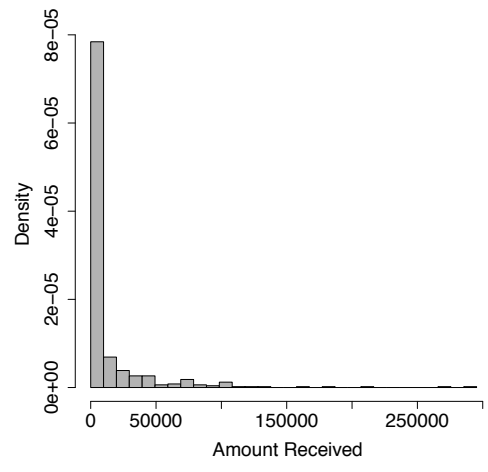
(a) NUM CLAIMS



(b) NUM SOFT TISSUE

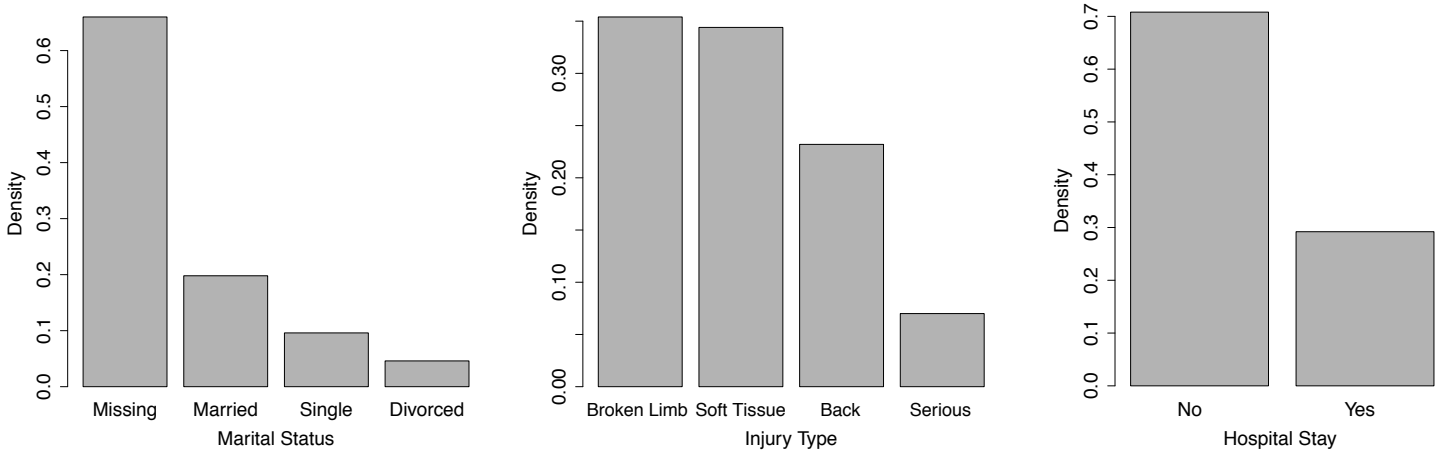


(c) % SOFT TISSUE



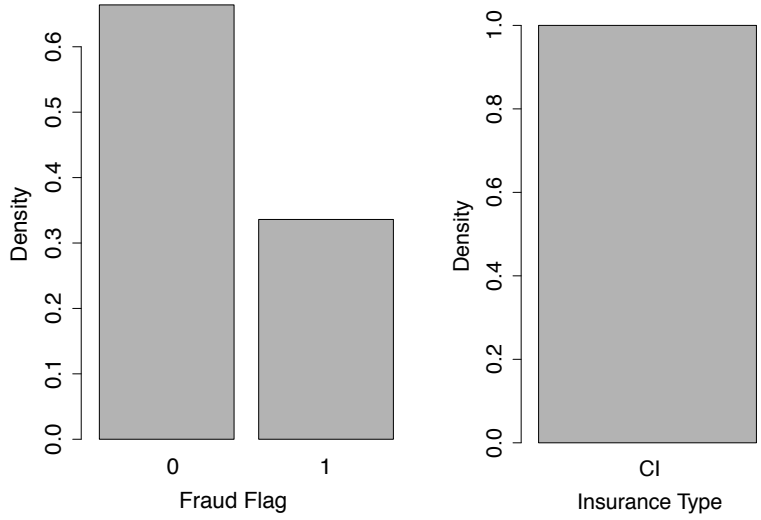
(d) AMOUNT RECEIVED

Figure: Visualizations of the continuous and categorical features in the motor insurance claims fraud detection ABT in Table 2 [7].



(a) MARITAL STATUS (b) INJURY TYPE (c) HOSPITAL STAY

Figure: Visualizations of the continuous and categorical features in the motor insurance claims fraud detection ABT in Table 2 [7].



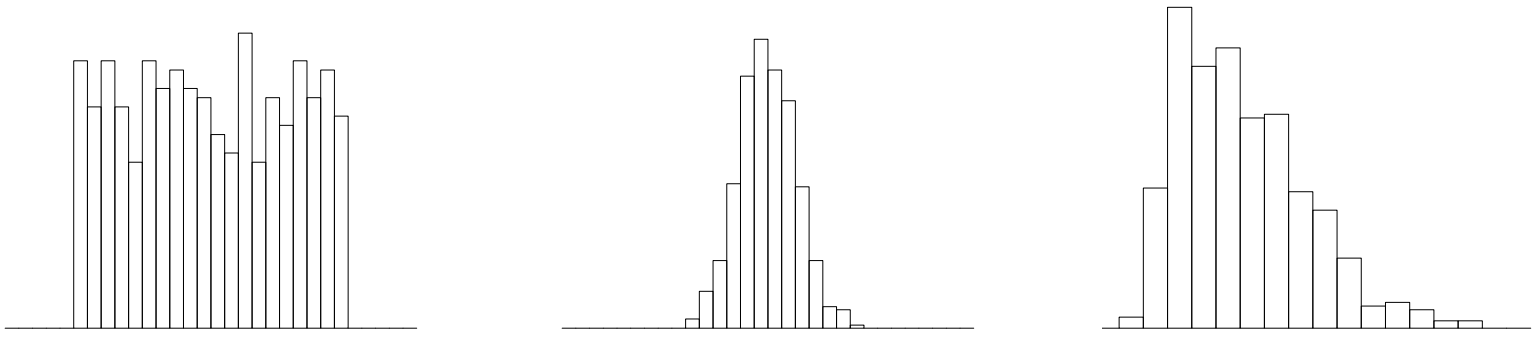
(a) FRAUD FLAG (b) INSURANCE TYPE

Figure: Visualizations of the continuous and categorical features in the motor insurance claims fraud detection ABT in Table 2 [7].

Getting To Know The Data

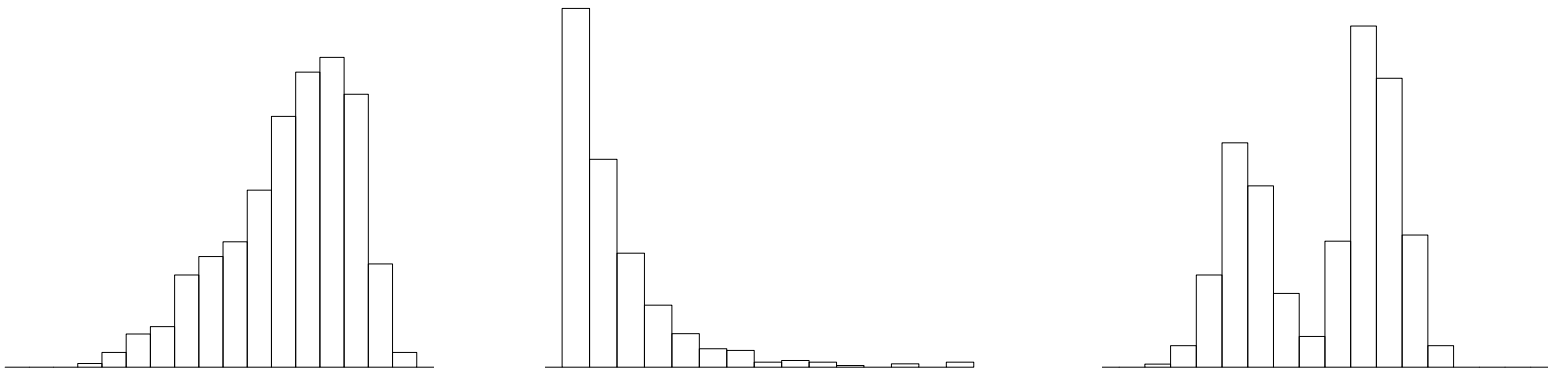
- For categorical features, we should:
  - Examine the mode, 2^{nd} mode, mode %, and 2^{nd} mode % as these tell us the most common levels within these features and will identify if any levels dominate the dataset.
- For continuous features we should:
  - Examine the mean and standard deviation of each feature to get a sense of the central tendency and variation of the values within the dataset for the feature.
  - Examine the minimum and maximum values to understand the range that is possible for each feature.

- When we generate histograms of features there are a number of common, well understood shapes that we should look out for.



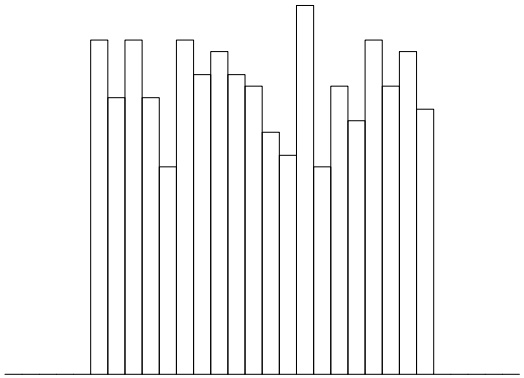
(a) Uniform (b) Normal (Unimodal) (c) Unimodal (skewed right)

Figure: Histograms for different sets of data each of which exhibit well-known, common characteristics.



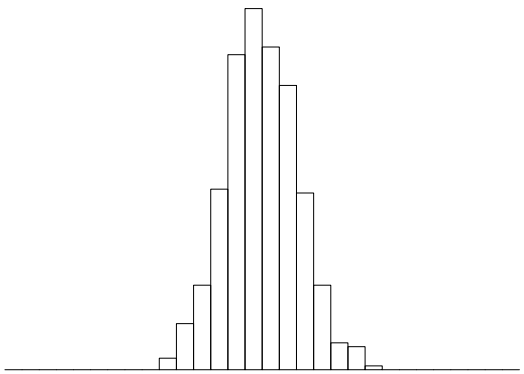
(a) Unimodal (skewed left) (b) Exponential (c) Multimodal

Figure: Histograms for different sets of data each of which exhibit well-known, common characteristics.



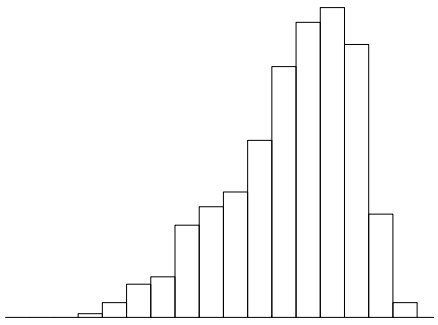
Uniform

- A uniform distribution indicates that a feature is equally likely to take a value in any of the ranges present.

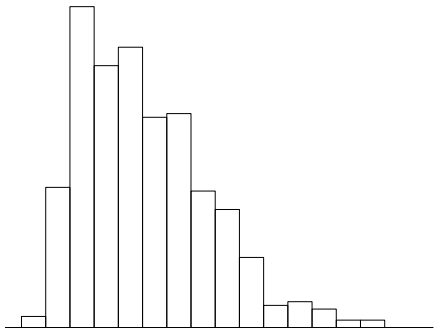


Normal (Unimodal)

- Features following a normal distribution are characterized by a strong tendency towards a central value and symmetrical variation to either side of this.

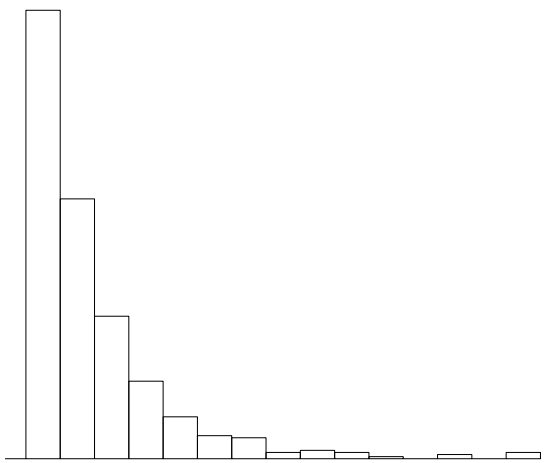


Unimodal (skewed left)



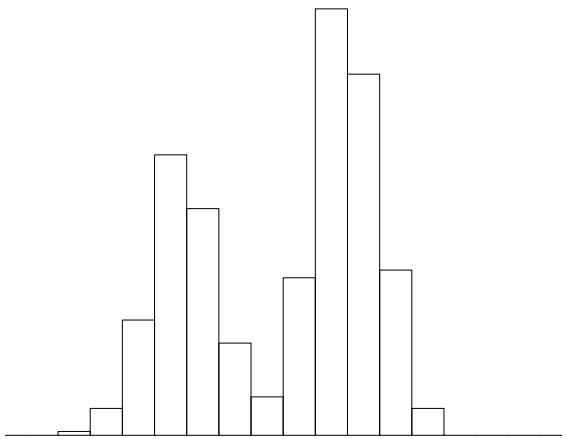
Unimodal (skewed right)

- Skew is simply a tendency towards very high (**right skew**) or very low (keywordleft skew) values.



Exponential

- 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.



Multimodal

- A feature characterized by a **multimodal distribution** has two or more very commonly occurring ranges of values that are clearly separated.

- The probability density function for the **normal** distribution (or **Gaussian distribution**) is

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

where x is any value, and μ and σ are parameters that define the shape of the distribution: the **population mean** and **population standard deviation**.

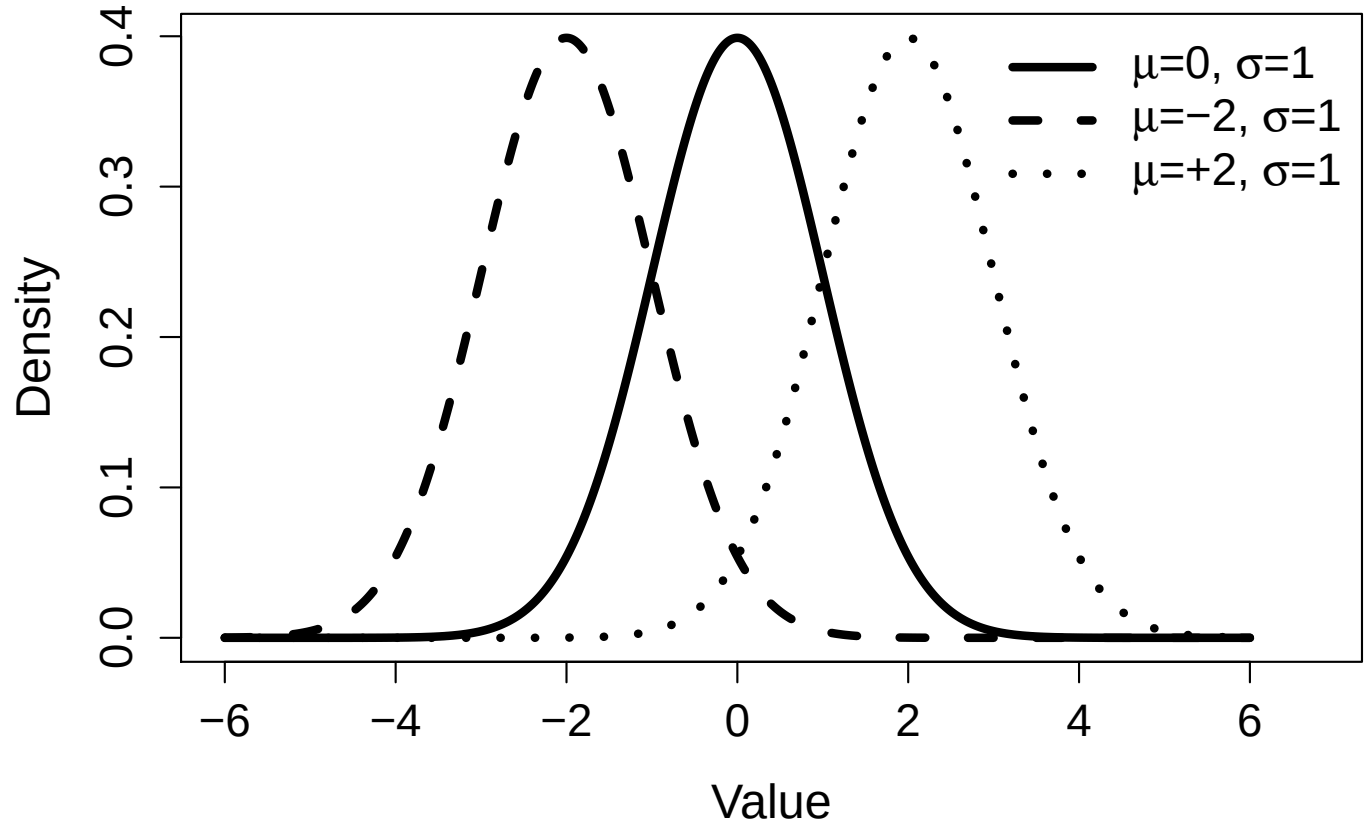


Figure: Three normal distributions with different means but identical standard deviations.

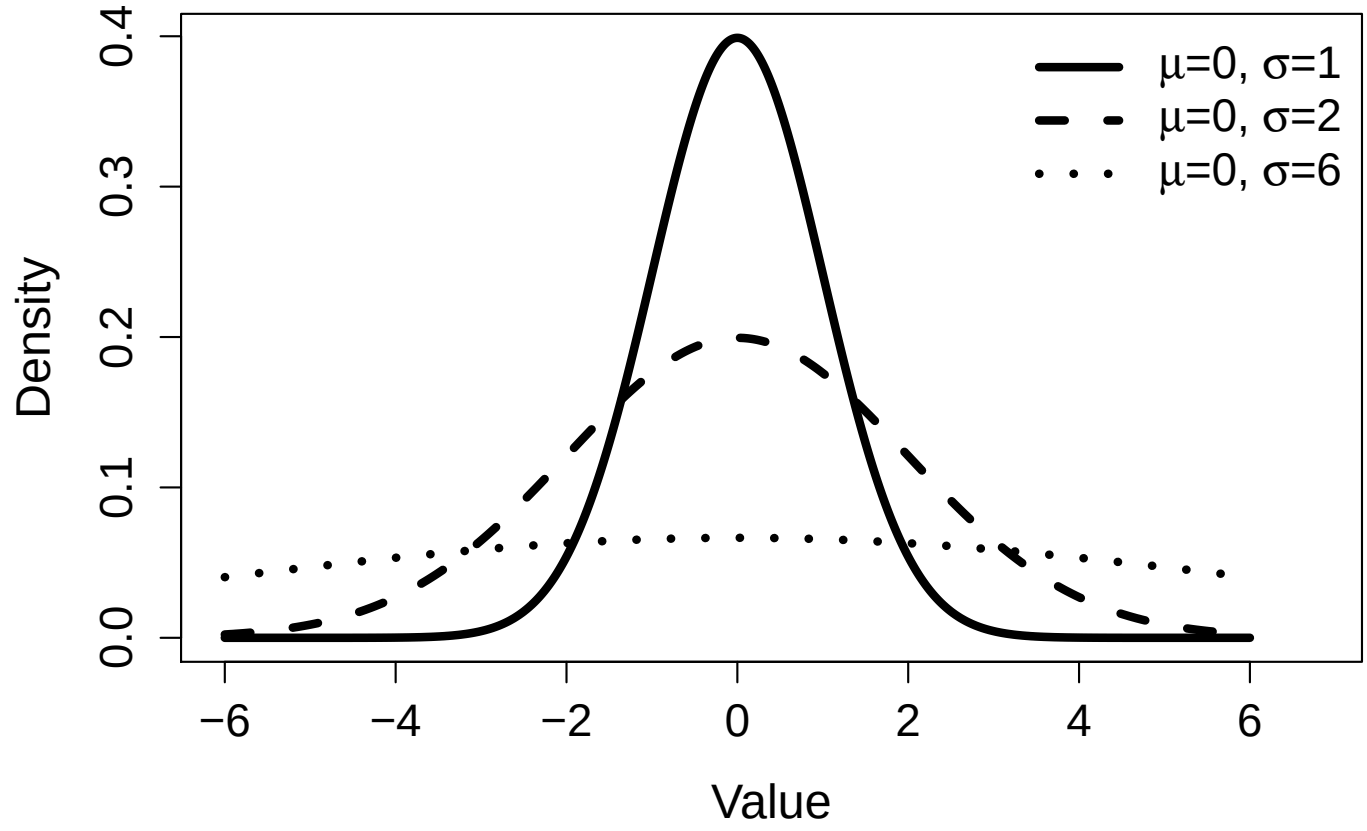


Figure: Three normal distributions with identical means but different standard deviations.

- The $68 - 95 - 99.7$ rule is a useful characteristic of the normal distribution.
- The rule states that approximately:
  - 68% of the observations will be within one σ of μ
  - 95% of observations will be within two σ of μ
  - 99.7% of observations will be within three σ of μ.

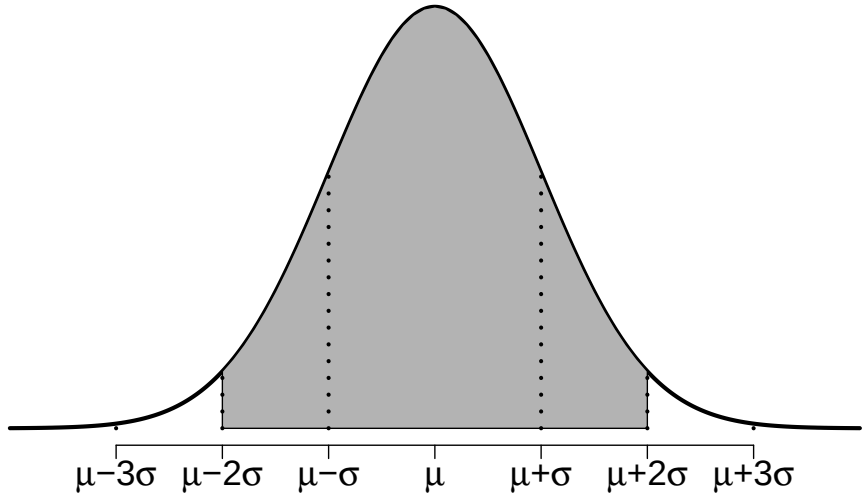


Figure: An illustration of the $68 - 95 - 99.7$ percentage rule that a normal distribution defines as the expected distribution of observations. The grey region defines the area where 95% of observations are expected.

Case Study: Motor Insurance Fraud

Examine the data quality report for the motor insurance fraud prediction scenario and comment on the central tendency and variation of each feature.

Identifying Data Quality Issues

- 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 issues we identify from a data quality report will be of two types:
  - Data quality issues due to **invalid data**.
  - Data quality issues due to **valid data**.

Table: The structure of a data quality plan.

Feature	Data Quality Issue	Potential Handling Strategies
———	———	———
———	———	———
———	———	———
———	———	———

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)	

Handling Data Quality Issues

- Approach 1: Drop any features that have missing value.
- Approach 2: Apply **complete case analysis**.
- Approach 3: Derive a **missing indicator feature** from features with missing value.

- **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.
- 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.

- 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

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

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

Case Study: Motor Insurance Fraud

What handling strategies would you recommend for the data quality issues found in the motor Insurance fraud ABT?

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)

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**.
  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.