Chapter 03 - Data Exploration

Sunday, December 6, 2020 1:19 AM

Before constructing the analytics base table, it is important to carefully examine the data. This process is called exploratory data analysis, or data exploration.

- What are the types of the data?
- What are the ranges of values?
- What is the distribution of values for each feature?
- What is the quality of the data? Are there missing or extreme values?

3.1 Data Quality Report

A "data quality report" can be generated to quickly describe the data. For each feature, calculate:

- Summary statistics for quantitative: count, mean, median, mode, min, max, standard deviation, percentiles, number of missing values, number of unique values
    - One feature per row to form an easy to read table
- Summary statistics for categorical values: count, count and % missing, how many unique values (cardinality), number of values in each category (along with the %)
    - One feature per row
- Basic distribution plots: histograms or bar plots, box plots
- Basic relationship plots: scatter plot matrix

3.2 Getting to Know the Data

For categorical variables:

- Examine the 1st and 2nd mode and percentage of representation to identify the most common values
- Bar plots of each variable give a visual representation of how common each value is

For continuous variables:

- The mean and standard deviation, along with histograms and box plots, describe the central tendency and variation of the distribution.

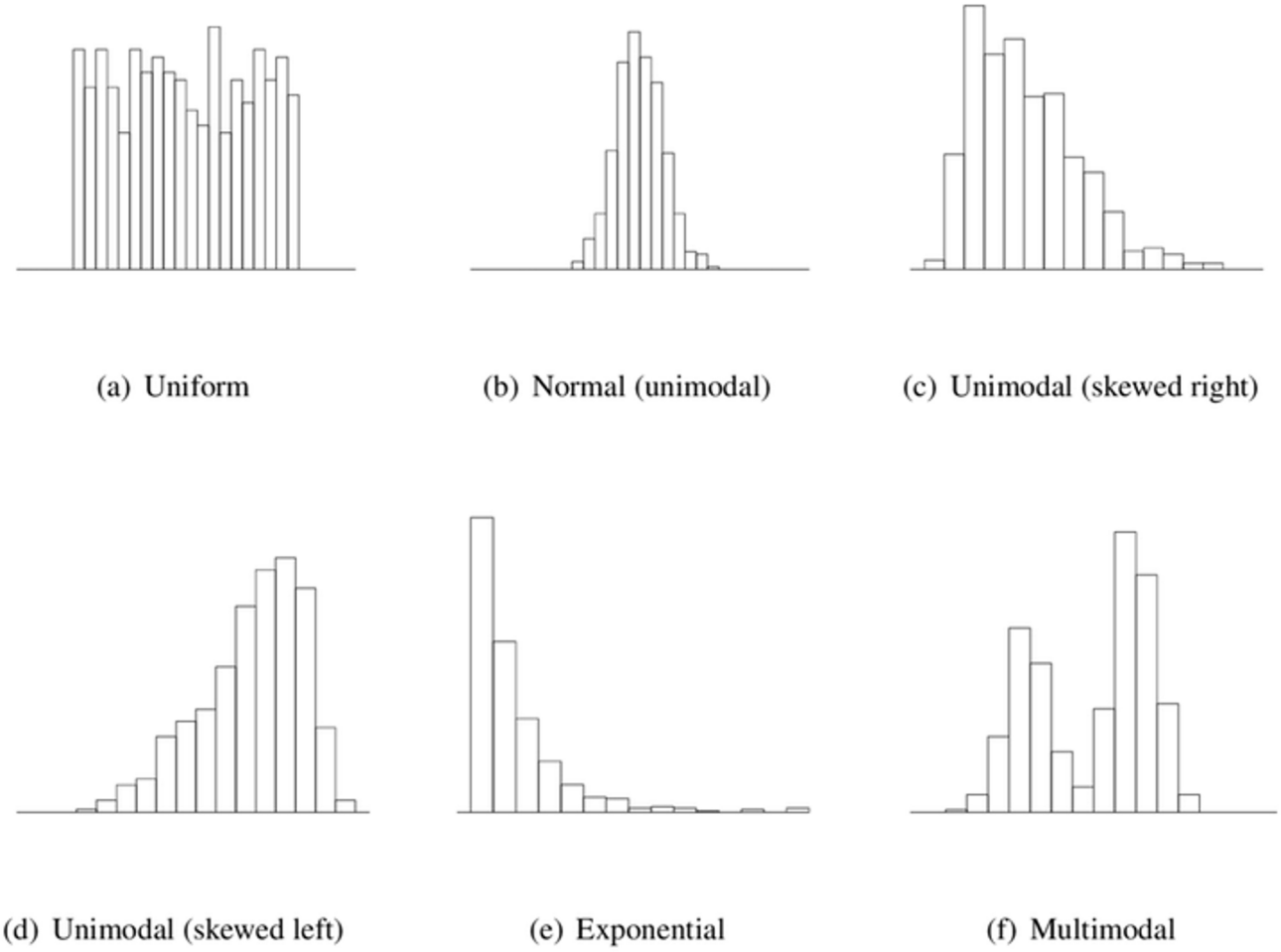
The distribution of a variable is important because it helps us determine the best type of model to use for our solution.

Three of the most common distributions are the **uniform distribution**, **normal distribution, exponential distribution**.

- Uniform distributions occur when each value of a random variable is equally likely
- Normal distributions occur most often for naturally occuring phenomenon
- Exponential distributions occur most often when dealing with how long it takes for an event to occur

Distributions can be unimodal or multimodal, and can be skewed right (peak is on the left side) or skewed left (peak is on the right side).

- When the distribution is multimodal, the mean is a very misleading value; but the presence of a multimodal distribution may indicate multiple clearly distinct "groups" within the data



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

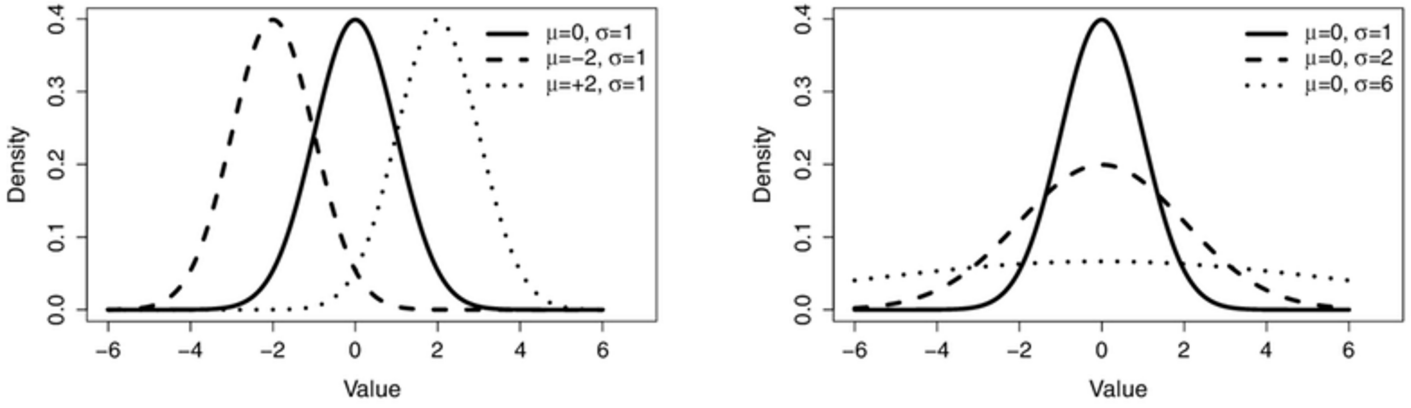
(d) Unimodal (skewed left) (e) Exponential (f) Multimodal

Figure 3.2
Histograms for six different sets of data, each of which exhibit well-known, common characteristics.

The Normal Distribution

2 parameter continuous distribution defined by

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



(a) Different means (b) Different standard deviations

Figure 3.3
(a) Three normal distributions with different means but identical standard deviations; and (b) three normal distributions with identical means but different standard deviations.

The **standard normal distribution** is defined as a normal distribution with a mean of 0 and sd of 1

68% of the values will be within 1 standard deviation, 95% within 2 sd, and 99.7% within 3

- If the distribution is normal, one method of detecting outliers is to look for values more than 3 or 4 standard deviations away from the mean

3.3 Identifying Data Quality Issues

Once the data has been summarized, it can be analyzed for quality issues.

- Missing values
- Irregular counts of unique values (cardinality)
- Outliers

If quality issues within the data are not resolved, then the data will not yield a good model.

Missing values:

- Why are the values missing? Collection problems? Integration problems? Intentionally missing?
- If a large portion of the values for a feature are missing (~60% is a good rule of thumb), then it may be best to not use that feature

Irregular Cardinality

- Occurs when there are more or fewer unique values for a categorical variable than expected
- If all of the values in a feature are the same (Cardinality 1), then that feature should be removed if there are no errors - it will not be useful in the model
- If the cardinality is close to the number of instances in the dataset, then the feature may be continuous and not categorical (and vice versa)
- Cardinality values larger than expected may indicate invalid levels that need to be recoded

Outliers

- Outliers can be invalid (actual errors), or valid (correct values, but unusual for some reason)
- To detect outliers, you can
    - Examine the minimum and maximum values for sensibility
    - Look at a box plot to see where the whiskers are

3.4 Handling Data Quality Issues

How to handle Missing Values

- Drop the entire row; this might be acceptable if there is a large amount of data, but can lead to bias
- Drop the feature if a large number of its values are missing (generally more than 60%)
- Create a new feature that indicates if the value is present or missing - this can be used to train the model when it should ignore the feature
- Impute the missing value by replacing it with a mean, median, mode, or other aggregate value
    - This is normally not the best idea as it can lead to bias in the data
    - Only consider it if a small number of features are missing (generally less than 30%)
- Build a predictive [regression] model based on the dataset to try and predict the missing features

Handling Outliers

- Clamp the values
    - $a_i = \begin{cases} lower, & a_i < lower \\ upper, & a_i > upper \\ a_i, & otherwise \end{cases}$
    - Method 1: Determine upper and lower values is to use the whiskers of a box plot (1.5 * 1st quartile and 3rd quartile)
    - Method 2: Set the upper and lower values to the mean plus/minus a multiple of the standard deviation
    - Always inspect the data to see how much of an impact clamping will have. If too many values will be changed, you may need to do something else.
- Consider leaving them alone if the values are valid
- Consider removing the observations if the values are invalid and the feature is important

Regardless of how missing values are handled, you should try to do it in a way that does not change the underlying distribution of the data.

3.5 Advanced Data Exploration

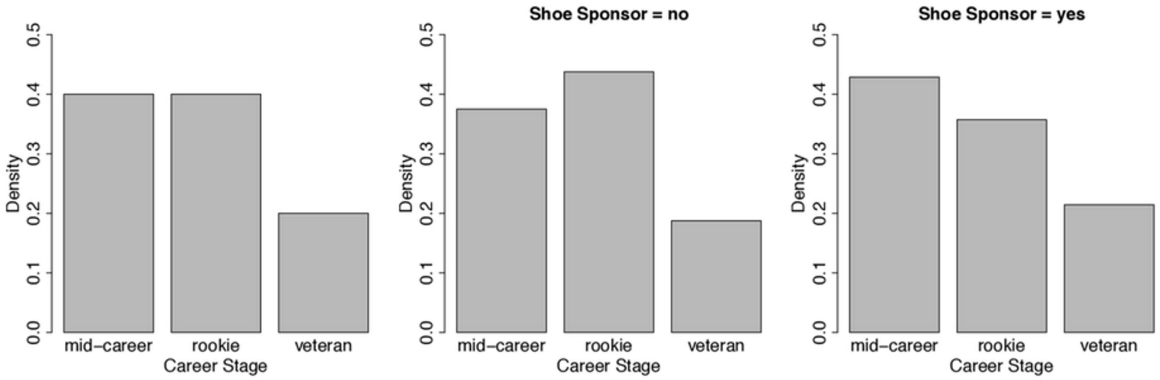
Scatter plots can be used to visualize the relationship between two variables

- Positive covariance (increase or decrease together)
- Negative covariance (as one goes up, the other goes down)
- No apparent relationship

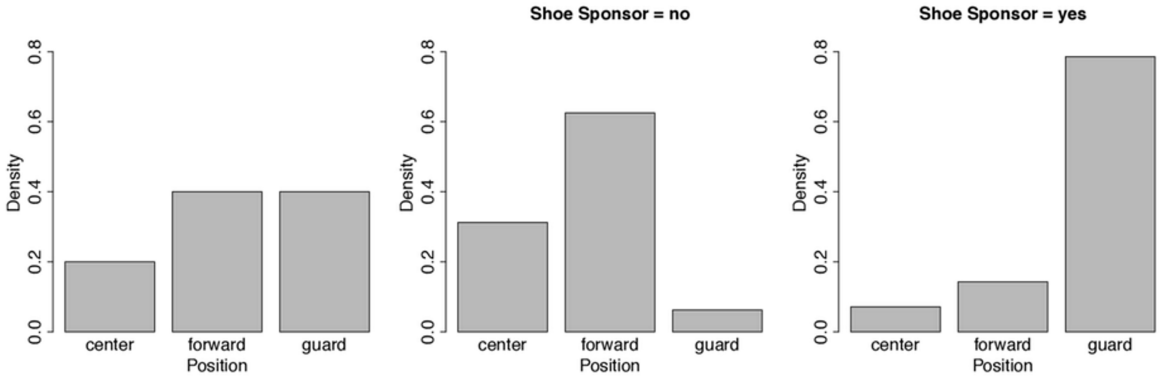
A scatter plot matrix displays scatter plots across all features in one visualization

For categorical data, you can draw a bar plot, factored across the levels of a feature.

- If there is a relationship present, then the distribution should differ across the levels of the second variable.
    - Why? Because if there is no relationship, then the value of the first variable should have no impact on the second



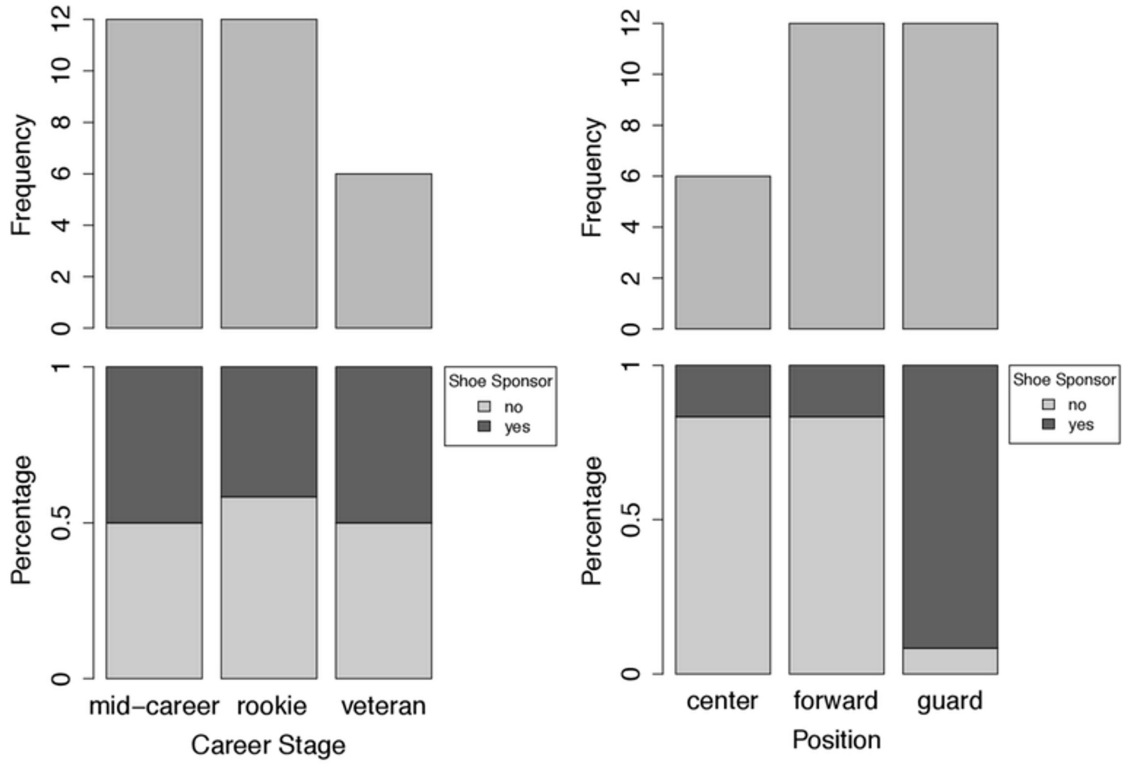
(a) Career Stage and Shoe Sponsor



(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[73].

Stacked bar charts can also be used to compare categorical variables. If there is a relationship, then the proportions of each level should differ by a large margin.



(a) Career Stage and Shoe Sponsor

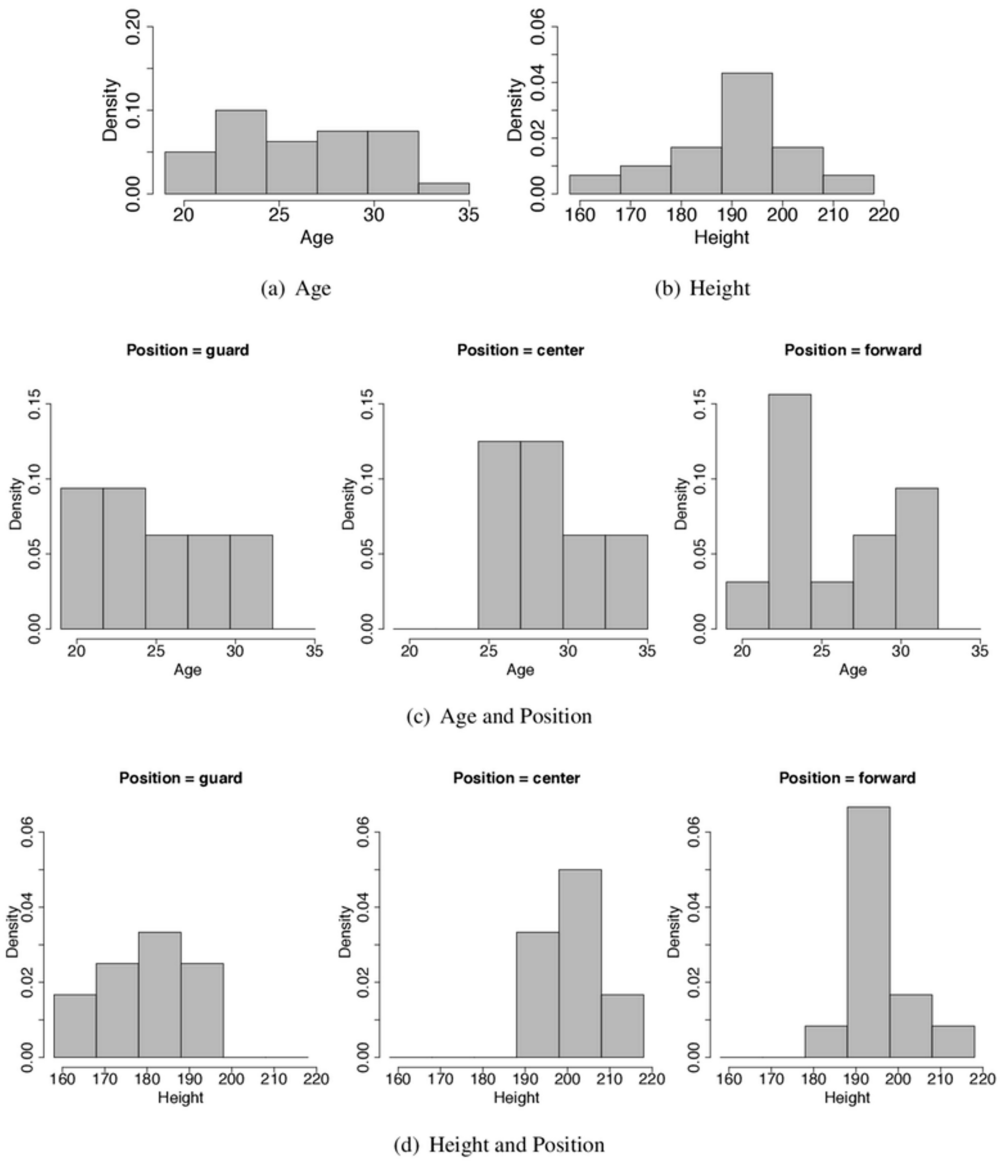
(b) Position and Shoe Sponsor

Figure 3.8

Examples of using stacked bar plot visualizations to illustrate the relationship between two categorical features: (a) CAREER STAGE and SHOE SPONSOR features; and (b) POSITION and SHOE SPONSOR features, all from Table 3.7[73].

To look for relationships between categorical and continuous variables, you can compare histograms when holding the level of the categorical variable steady.

Alternatively, you can also use box plots.



(a) Age (b) Height

(c) Age and Position

(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[73]: (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.

Python and R both have graphics libraries that make this easy - Seaborn and ggplot.

- In R, you can specify the facet to have plots automatically generated across levels of a categorical variable

It is normally easier to conduct an exploratory data analysis in R than it is Python.

Measuring Covariance and Correlation

Plots alone are not sufficient to understanding relationships between variables. Covariance and correlation provide numerical metrics of the strength of these relationships.

Covariance is defined as

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

where \bar{a} and \bar{b} are the sample means of the features a and b.

Covariance measures the *linear* relationship between the variables. Values near 0 indicate that there is little to no relationship between the variables.

Covariance maintains the units of each variable, which may not make sense when compared to one another.

Correlation is the normalized covariance, removing units and limiting the range to [-1, 1].

$$corr(a,b) = \frac{cov(a,b)}{\bar{\sigma}_a \bar{\sigma}_b}$$

Where \bar{\sigma} is the sample standard deviation.

The covariance and correlation functions can be used to generate a covariance or correlation matrix, showing how every variable is linearly related to every other variable.

Correlation is not causation.

- Many times, the relationship between two variables exists because of confounded features that may or may not be directly observed
- Only careful experimentation can tease out causation

To measure the similarity between categorical variables, statistical techniques such as Chi-Squared tests and ANOVA can be used.

3.6 Data Preparation

Once the features have been identified, and the quantity assessed and corrected, the final step to conduct any transformations necessary to facilitate learning and model building.

Normalization

- Modify the range of a feature to be within [low, high]
  - $a_i = \frac{a_i - \min(a)}{\max(a) - \min(a)}(high - low) + low$
- Modify the a feature with a z-transform to be normally distributed with a mean of 0 and standard deviation of 1 (this really only works as expected if the feature is normally distributed to begin with)
  - $a_i = \frac{a_i - \bar{a}}{sd(a)}$
  - If the feature is not normally distributed to begin with, this may distort the data bit; but it might still be useful, you'll need to check.

Binning

- Converts a continuous feature into a categorical feature
  - Helps some algorithms handle continuous features "better"
  - Helps handle outliers
  - Discards information in the process
- If the number of bins is too low, then information is lost with respect to the distribution of the original values
- If the number of bins is too high, some of those bins may be empty
- Ideally, you want a number of bins that produces a representation close to the original distribution.

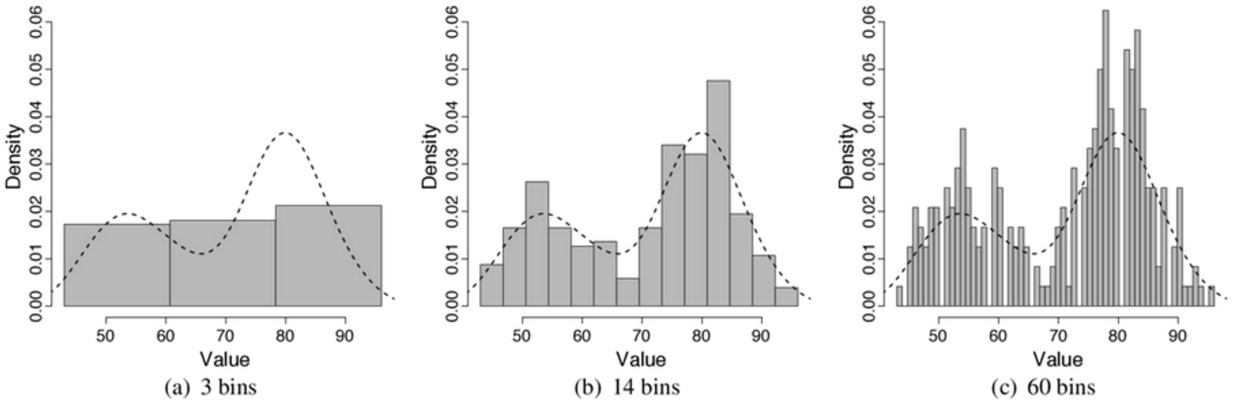


Figure 3.13
The effect of using different numbers of bins when using binning to convert a continuous feature into a categorical feature.

Equal width binning
- Splits the values into b bins, each of size range / b
  - E.g. [0, 10), [10, 20), …, [90, 100]
- Good for uniform distributions, but may produce many empty bins for non-uniform distributions

Equal frequency binning
- Sorts the values from smallest to largest and then puts an equal number of values in each bin
- The total number of instances in each bin is [count of instances] / [number of bins]

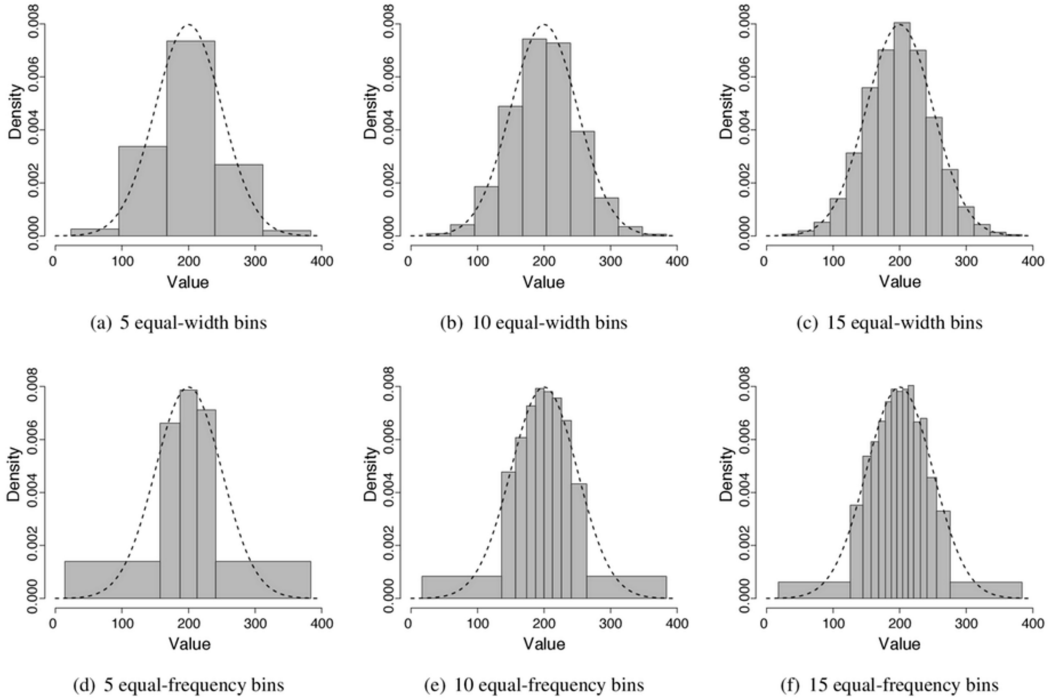


Figure 3.14
(a)–(c) Equal-frequency binning of normally distributed data with different numbers of bins; and (d)–(f) the same data binned into the same number of bins using equal-width binning. The dashed lines illustrate the distribution of the original continuous feature values, and the gray boxes represent the bins.

Sampling
- Selects a subset of data from the analytics base table
- Must be done carefully or the sample will be biased and not accurately represent the

population

Top Sampling
- Selects a flat % from the top of the dataset
- Is almost always biased and impacted by the ordering of the data

Random Sampling
- Randomly selects a flat % from the dataset
- Does not preserve relationships in the data

Stratified Sampling
- The dataset is grouped by one or more particular variables, and then s% of each group (called a strata) is selected for the sample.
- Maintains the relative frequency of each group within the dataset

Under-Sampling
- Creates a sample where all groups are equally represented
- Group the dataset by one or more variables; from each group, randomly sample (without replacement) N instances, where N is the number of instances in the smallest group

Over-Sampling
- Creates a sample where all groups are equally represented
- Group the dataset by one or more variables; from each group, randomly sample (with replacement) N instances, where N is the number of instances in the largest group

Under-sampling and over-sampling can be used to train predictive models that try to ignore sampling bias in the original dataset
- For example, when data is gathered unequally from different subpopulations

3.7 Summary

The key outcomes of the data exploration process should include:
1. Have gotten to know the features, especially their central tendencies, variations, and distributions
2. Have identified any data quality issues, in particular missing values, irregular cardinality, and outliers
3. Have corrected any data quality issues related 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 enough that good-quality data exists to continue with a project

Any steps taken to transform the data must be recorded so that they can also be applied as new data is made available.