Data Collection and Cleaning

Representation, Missing Values, Outliers

Topics

1. Measurement Scales
2. Data Collection
3. Missing Values
4. Outliers

Measurement Scales

- Non-metric / Categorical / Qualitative
  - Nominal (labels) - gender, CNIC, eye color, postal-code - maths operations: mode only
  - Ordinal (ranking - order only, no magnitude) - Grades, Very satisfied to Not at all satisfied - allowed: + median, percentile
- Metric
  - Interval (arbitrary zero) - temperature in Celsius & Fahrenheit (80F is twice as hot than 40F?) allowed: + all except ratio or percentage
  - Ratio (absolute zero) - height, weight, income - allowed: all operations

Data Collection - representativeness

- Aggregation or omission errors
  - Example: rating of 3 colas unknowingly involving 2 subgroups students, and stay-at-home mothers
  - Example (real): new drug testing initially failed but afterwards identified that it worked for one group
- Including irrelevant groups
  - Example: You want to study smokers. How do you define a smoker?
  - Example: Voters survey
  - Example: (motivation) Identify who is randomly responding by including questions to detect
- Nonresponse and generalizability
  - Example: Students from stigmatized groups usually choose not to answer certain questions

Data Collection - representativeness

- Consent procedures and sample bias
  - Example: Patients, parents of minors
  - Look for ways to get consent in a non-alarming way (e.g. at the time of registration)
- Generalizability of internet surveys
  - Who has access to the internet, who might be willing to participate, fake and repeated responses
- Restriction of range
  - Example: Study students for self-esteem (very high to very low). Students generally have high-self esteem and does not represent population
  - Ceiling and floor effects (may be caused by measuring instrument - too sensitive?)
- Extreme groups analysis
  - Example: Select students with very high and very low IQ rating for comparison and intervention. This kind of study will not be representative of population results.

Missing Data

- Legitimate missing data
  - Are you married? For how long have you been married?
  - Employee satisfaction survey when an employee has left.
  - Long term medical study when patient has been cured and not responding to questions
  - You can adjust the denominator (subsample) to handle this kind of situation
- Illegitimately missing data
  - Sensor malfunction
  - Research participants choose to skip questions
  - Caused by data cleaning

Missing Data - Missing Data Process

Any systematic event

external to the respondent (such as data entry errors or data collection problems)

or any action on the part of the respondent (such as refusal to answer) that leads to missing values.

Investigate:

Is it random or there is a pattern?

How prevalent is missing data?

Missing Data - Levels of Randomness

- Level of randomness depends on 2 conditions
  1. The randomness of missing values of a variable among its own values
  2. The degree of association between the missingness of one variable and other variables

- Missing Completely at Random (MCAR)
  - Conditions #1 & #2 are absent
- Missing Data at Random (MAR)
  - Condition #1 absent and #2 present
- Missing Not at Random (MNAR)
  - Condition #1 present

Figure 2.6

Missing Data Processes: MCAR, MAR and MNAR

Complete Data		Missing Data Process for *Y*		
X	*Y*	MCAR:	MAR:	MNAR:
3	9	9	Missing	9
3	5	5	Missing	5
4	1	Missing	Missing	Missing
4	3	3	Missing	Missing
5	2	Missing	2	Missing
6	6	Missing	6	6
7	7	7	7	7
7	4	4	4	Missing
8	5	5	5	5
9	9	Missing	9	9
		Characteristics of the Missing Data Process		
Pattern of missing values of *Y*		Random: Across all values of *Y*	Random: Across all values of *Y*	Nonrandom: Only lowest values of *Y*
Relationship of *X* to missingness of *Y*		No Across all values of *X*	Yes Lowest values of *X*	No Across all values of *X*

Adapted from [31].

Missing Data - Imputation

- Imputation for MCAR using only valid data
  - Complete Case Approach (list-wise)
    - Include only observations with complete data [reduced sample size]
  - Using All-Available Data (pair-wise)
    - Calculates distribution characteristics (e.g., means or standard deviations) or relationships (e.g., correlations) from every valid value ignoring the missing values [out of range correlation values can occur]
- Imputation for MCAR by using known replacement values
  - Hot-deck - replace value from a “similar” observation
  - Cold-deck - replace value from external source (prior studies, other samples)
  - Case Substitution - replace entire observation by a non-sampled (external) observation

Missing Data - Imputation

- Imputation for MCAR by calculating replacement values
  - Mean substitution [understate variance, distort distribution, group mean substitution?]
  - Regression imputation [reinforces relationships already in data, stochastic terms?, assumes sufficient correlation]
- Imputation for MAR
  - Multiple Imputation [generated multiple imputed datasets then combine, alternate realities]

Outliers

- Error outliers
- Interesting outliers
- Influential outliers

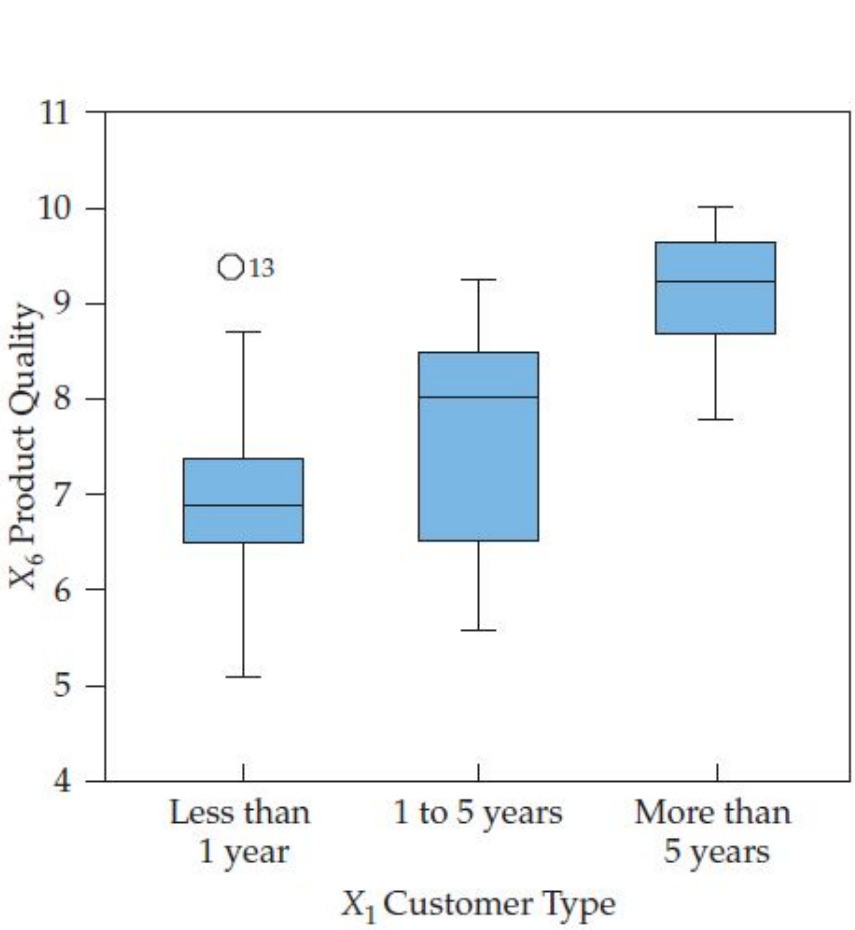
Detecting outliers

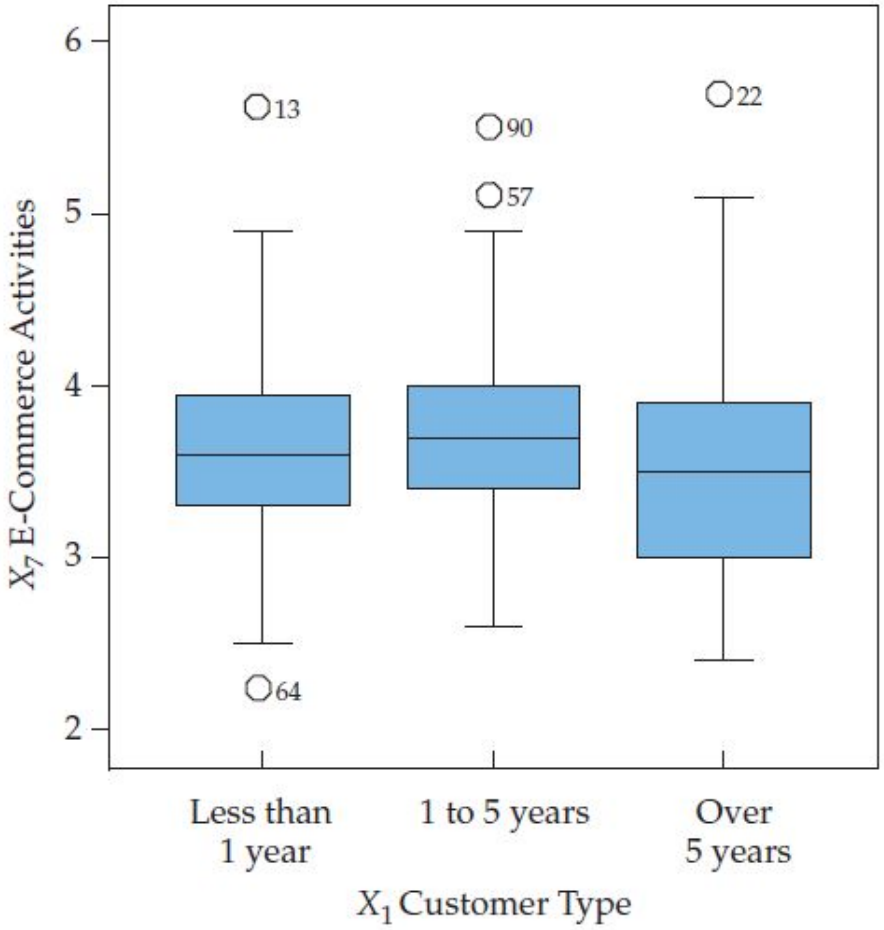
- Normal distribution z score (e.g. +/- 3 sigma or more)
- Dbscan (Density Based Spatial Clustering of Applications with Noise)
- Isolation Forests
- Plots (histogram, scatter, box plot)

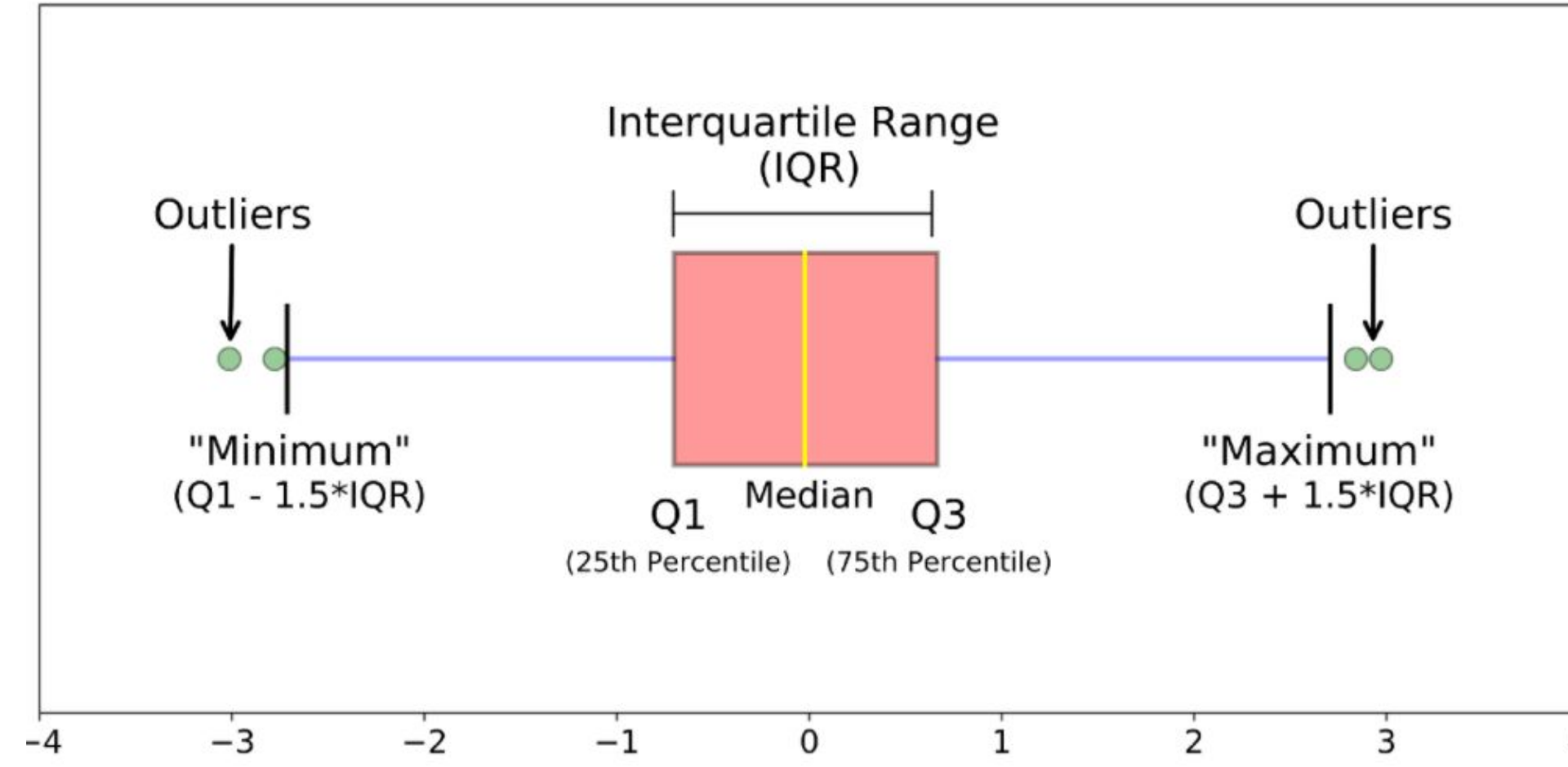
The Box Plot

- Demonstrate the locality, spread and skewness of data
- Can be used to identify the outliers

Figure 2.3 Bivariate Profiling of Group Differences: Boxplots of X_6 (Product Quality) and X_7 (E-Commerce Activities) with X_1 (Customer Type)







Different parts of a boxplot

Exercise: Draw Box Plot with outliers

Data:

65,199, 201, 236, 269,271,278,283,291, 301, 303, 341, 402

References

1. Multivariate Data Analysis - Chapter 1 (pp 11)
2. Best practices in Data Cleaning - Chapter 3
3. Multivariate Data Analysis - Chapter 2