Data Visualisation for Business

ANL 201

Four Stages of Data Visualisation
Study Unit 3

January 2024

SUSS
SINGAPORE UNIVERSITY
OF SOCIAL SCIENCES

Recap- Science and Art of Data Visualisation

- Data Visualisation- definition, benefits and examples

- Four components of Data visualisation-
  - *Visual cues, Coordinate systems, Scales and Context*

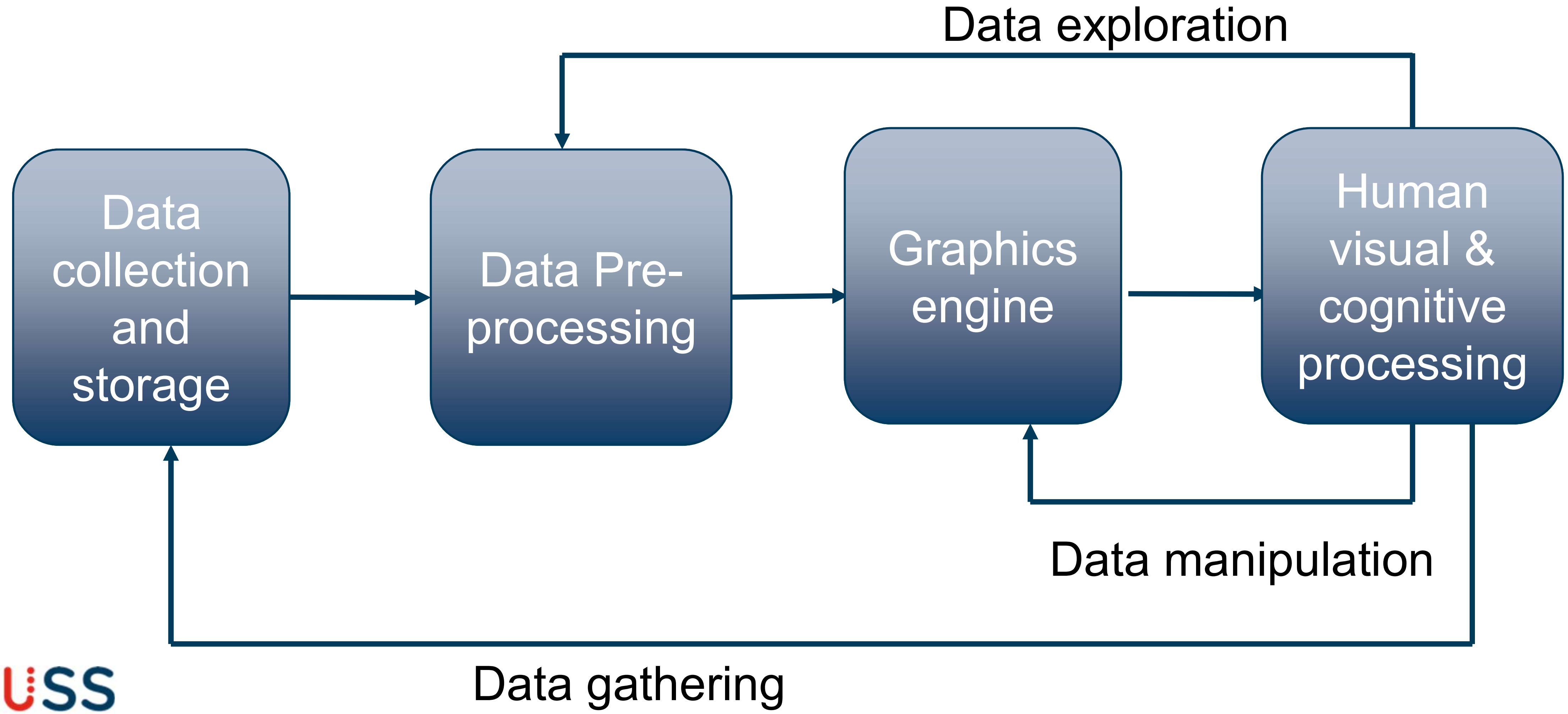
- Tableau activity
  - *Table join, Data blending, Pivot, Split, Calculated field, Quick table calculations*

Data Visualisation Stages

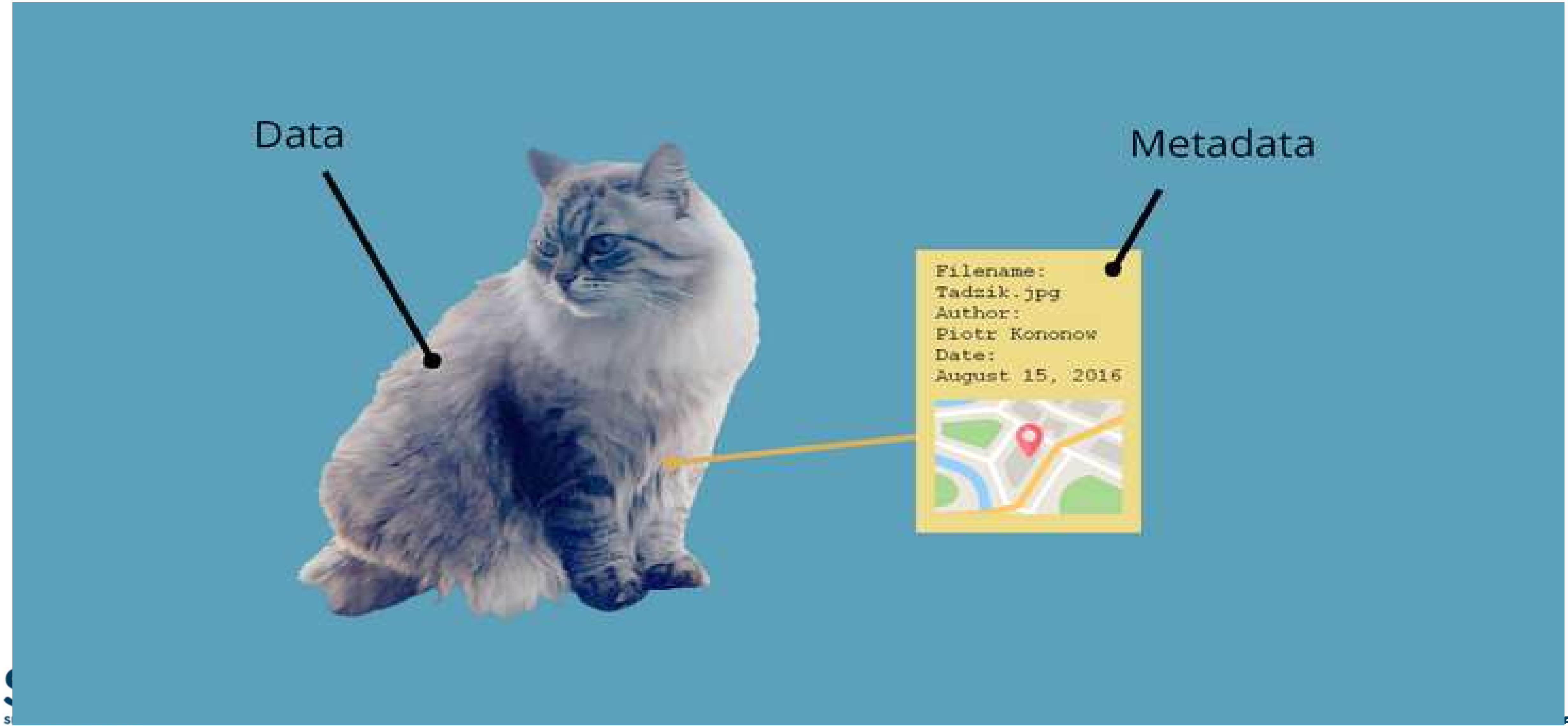
SUSS
SINGAPORE UNIVERSITY
OF SOCIAL SCIENCES

Data Visualisation Stages

Four stages of the data visualisation process



Metadata


Data
Metadata
Filename:
Tadzik.jpg
Author:
Piotr Kononow
Date:
August 15, 2016

S

Data Dictionary

No.	Name	Description	Measurement	Data Type	Frequency
1	USER_ID	ID of the customer	Nominal	ID	Monthly
2	AGE	Age of the customer	Integer	Numeric	Monthly
3	GENDER	Gender of the customer	Nominal	Character	Monthly
4	SECTOR	The sector that the customer is working in	Nominal	Character	Monthly
7	TOTAL_WEEKLY_SALES	Total purchases made by the customer in a week	Integer	Numeric	Weekly
8	SALES_RANK_MTH	Ranking of customer based on the total purchases made by the customer in a month	Ordinal	Numeric	Monthly

Types of Datasets

Record Data

- No explicit relationship among records or data fields, and every record (object) has the same set of attributes
- Transaction or market basket data: each record contains a set of items
- The data matrix: data objects can be thought of as points (vectors) in a multidimensional space, where each dimension represents a distinct attribute describing the object
- The sparse data matrix: is a special case in which the attributes are of the same type and are asymmetric; i.e., only non-zero values are important

Types of Datasets

Graph-based Data

- Data with relationships among objects
  - Example- web pages on the World Wide Web, which contain both text and links to other pages
- Data with objects that are graphs
  - Example, the structure of chemical compounds can be represented by a graph, where the nodes are atoms and the links between nodes are chemical bonds

Types of Datasets

Ordered Data

- For some types of data, the attributes have relationships that involve order in time or space.
- Sequential data: each record has a time associated with it. Example- retail transaction data with timestamp
- Sequence data: sequence of words or letters. Example- genetic information of plants and animals.
- Time series data: each record is a time series. Example- daily prices of stocks.
- Spatial data: data objects have spatial attributes, such as positions or areas. Example- weather data collected from different geographical locations.

Class Discussion 1

Types of Datasets

What are some examples for the three types of datasets?

- Record data
- Graph-based data
- Ordered data

Class Discussion 1

Types of Datasets

What are some examples for the three types of datasets?

- Record data – retail sales data, survey data
- Graph-based data - Facebook, Linkdln, Instagram and Twitter
- Ordered data- market indices datasets (Dow Jones, S&P 500, Nasdaq, Nikkei etc.), price of a product across time/across locations

Data Collection and Storage

SUSS
SINGAPORE UNIVERSITY
OF SOCIAL SCIENCES

Data Collection and Storage

Data Collection Methods

- Surveys gather both qualitative and quantitative data
- Transactional Tracking records Customer purchase history
- Interviews and focus groups consist of talking to subjects face-to-face
- Observing people interacting with your website or product
- Online tracking- using pixels and cookies to gather behavioural data
- Online forms are beneficial for gathering qualitative data about users
- Monitoring your company's social media channels

Data Collection and Storage

Data Collection Challenges and Improvements

Challenges	Improvements
Inconsistent data collection standards	Data items have pre-defined responses
Data collection is not core to business function	Using data-related key performance indicators (KPIs)
Lack of training in data collection	Staff training
Lack of quality assurance processes	Creating mandatory data fields
Economic and IT restrictions	Commitment from all levels of an organisation

Data Collection and Storage

Data Storage Challenges and Solutions

Challenges	Solutions
Infrastructure	Cloud hosting
Cost	Outsource the work
Security	Run a tight operation
Corruption	Use multiple backups
Scale	Explore options
UI and accessibility	Use a system with good UI
Compatibility	Open API

Data Preparation

SUSS
SINGAPORE UNIVERSITY
OF SOCIAL SCIENCES

Data Preparation

Benefits

- Ensure the data used in analytics applications produces reliable results
- Identify and fix data issues that otherwise might not be detected
- Enable more informed decision-making by business executives and operational workers
- Reduce data management and analytics costs
- Avoid duplication of effort in preparing data for use in multiple applications
- Get a higher ROI from BI and analytics initiatives

Data Preparation

Data Preparation Principles

- Understand the data consumer
- Understand the data
- Save the raw data
- Ensure that transforms are reproducible and deterministic
- Future proof your data pipeline

Data Preparation

Steps in the Data Preparation Process

The data preparation process comprises of the following steps:

- Understand the data: data types
- Understanding the data: dataset architecture
- Exploratory data analysis (EDA) & appropriate treatment methods

Understanding Data: Data Types

S.S. Stevens (1946) defined four data types

1. *Nominal* - classifies entities based on their labels or categories – color of a car

2. *Ordinal* - orders entities based on rank – quality rating of a car

★★★★★ > ★★★★ > ★★★

3. *Interval* - measures the degree of difference between entities- year of manufacture of a car

4. *Ratio* – measures the equality of ratios for continuous variables – engine horsepower

Class Discussion 2

Four types of attributes of an entity

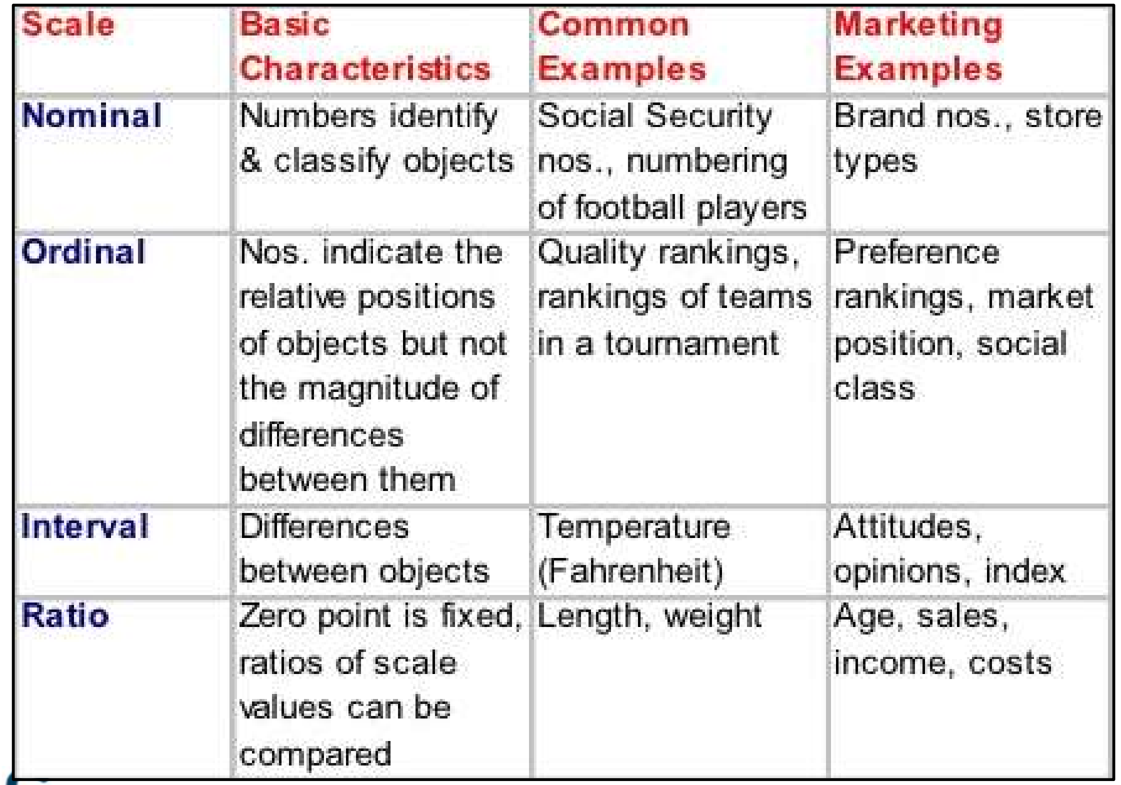
What are some examples for the four types of attributes that you can identify in your company, or any other organisations you are familiar with?

- Nominal
- Ordinal
- Interval
- Ratio

Class Discussion 2

Four types of attributes of an entity

What are some examples for the four types of attributes that you can identify in your company, or any other organisations you are familiar with?

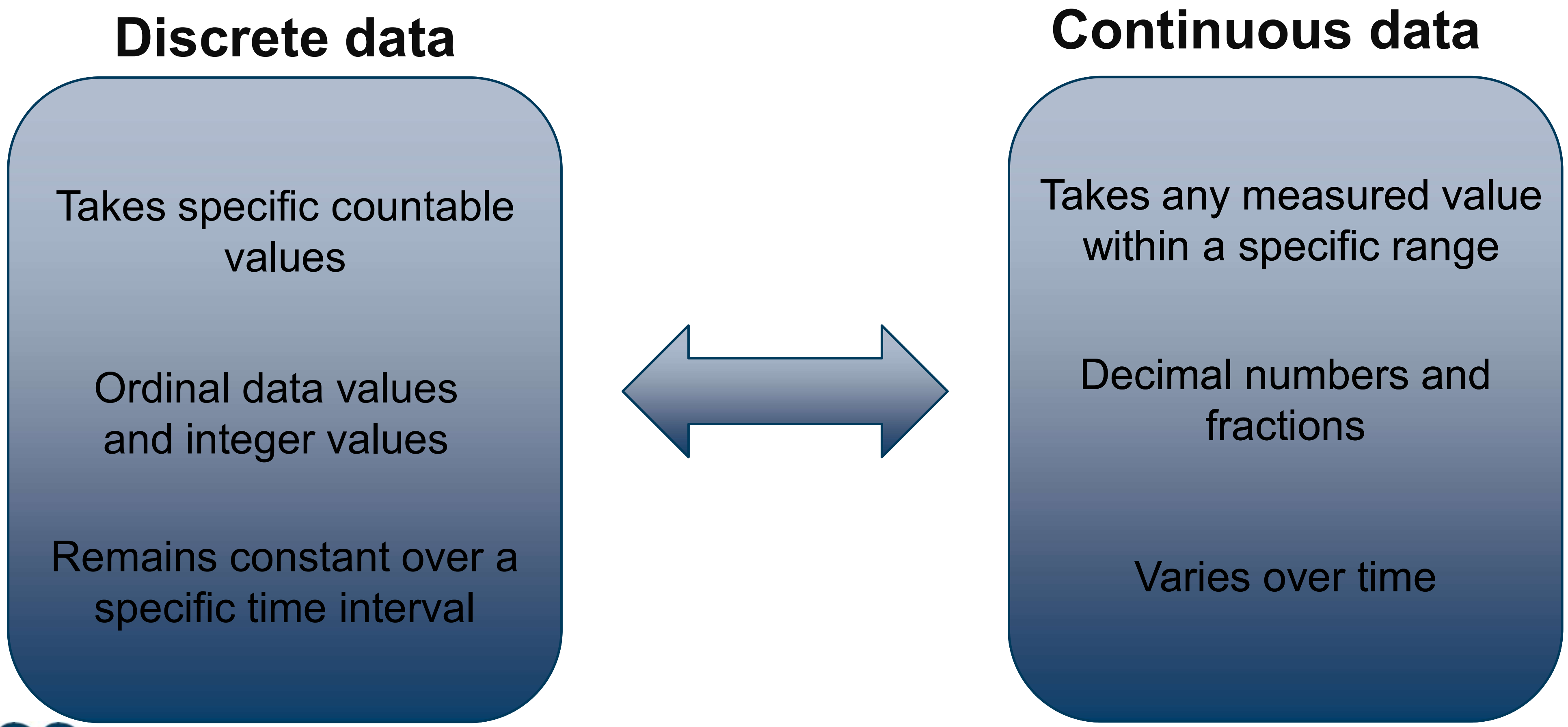


Scale	Basic Characteristics	Common Examples	Marketing Examples
Nominal	Numbers identify & classify objects	Social Security nos., numbering of football players	Brand nos., store types
Ordinal	Nos. indicate the relative positions of objects but not the magnitude of differences between them	Quality rankings, rankings of teams in a tournament	Preference rankings, market position, social class
Interval	Differences between objects	Temperature (Fahrenheit)	Attitudes, opinions, index
Ratio	Zero point is fixed, ratios of scale values can be compared	Length, weight	Age, sales, income, costs

Source: http://image.slidesharecdn.com/measurementandscales-120411075801-phpapp02/95/measurement-and-scales-6-728.jpg?cb=1334131308

Understanding Data: Data Types

Discrete and Continuous data



Discrete data

Takes specific countable values

Ordinal data values and integer values

Remains constant over a specific time interval

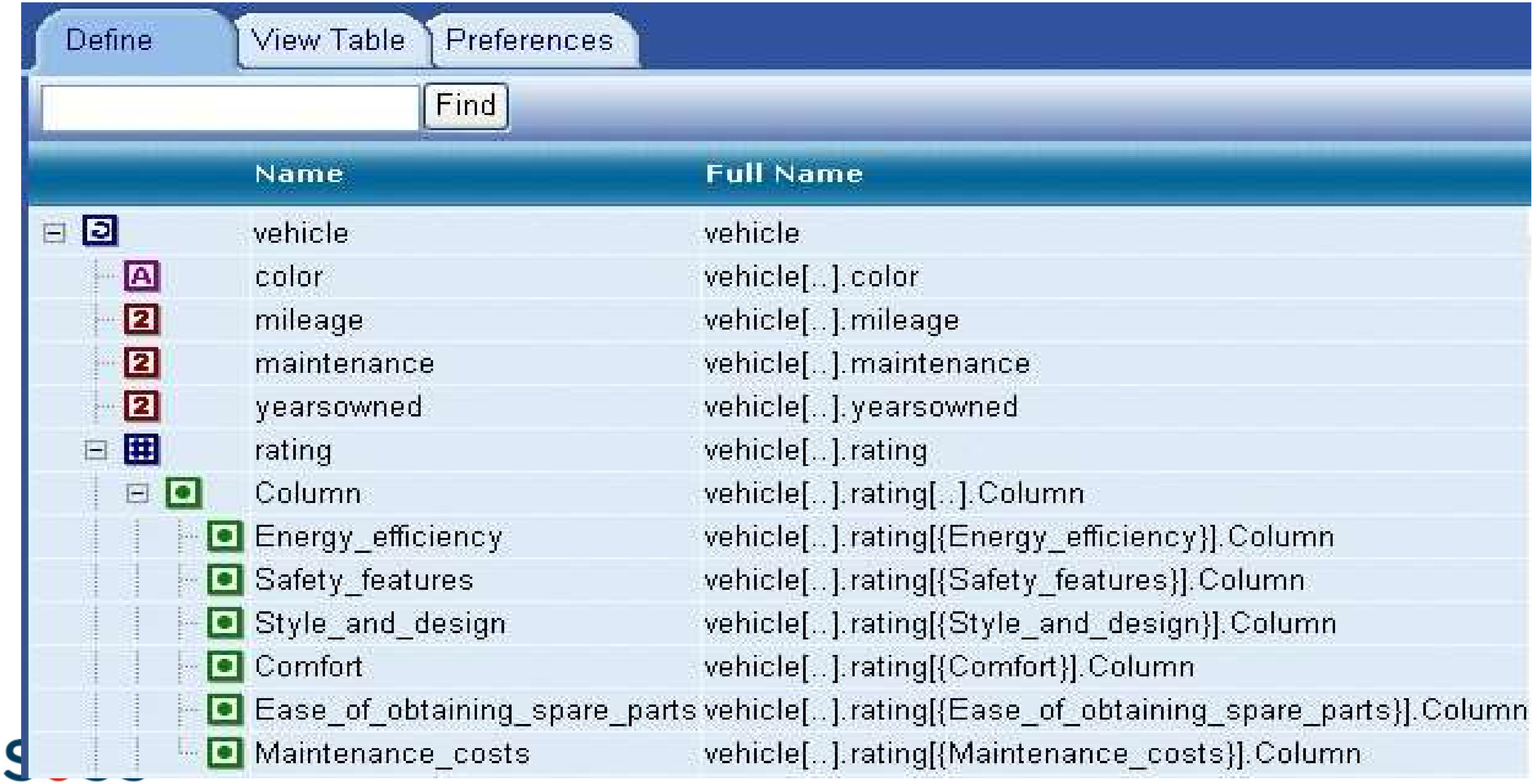
Continuous data

Takes any measured value within a specific range

Decimal numbers and fractions

Varies over time

Understanding Data: Dataset Architecture



Name	Full Name
vehicle	vehicle
color	vehicle[..].color
mileage	vehicle[..].mileage
maintenance	vehicle[..].maintenance
yearsowned	vehicle[..].yearsowned
rating	vehicle[..].rating
Column	vehicle[..].rating[..].Column
Energy_efficiency	vehicle[..].rating[{Energy_efficiency}].Column
Safety_features	vehicle[..].rating[{Safety_features}].Column
Style_and_design	vehicle[..].rating[{Style_and_design}].Column
Comfort	vehicle[..].rating[{Comfort}].Column
Ease_of_obtaining_spare_parts	vehicle[..].rating[{Ease_of_obtaining_spare_parts}].Column
Maintenance_costs	vehicle[..].rating[{Maintenance_costs}].Column

Source:*Variable hierarchy example (UNICOM Intelligence, n.d.)*

Understanding Data: Dataset Architecture

Data granularity

- Measure of the level of detail in a data structure
  - Examples- the granularity of measurement might be based on intervals of years, months, weeks, days, or hours
  - For ordering transactions, granularity might be at the purchase order level, or line item level, or detailed configuration level for customised parts.

- Determines what analysis can be performed on the data, and whether results from that analysis lead to appropriate conclusions

Understanding Data: Dataset Architecture

Data aggregation

- Data can be aggregated over a given period to provide statistics such as sum, count, and average, minimum, maximum

- There are two types of data aggregation:

  - Time aggregation- data points for a single resource over a specified period

  - Spatial aggregation- data points for a group of entities (sku of a product, products in a basket, places in a region/country etc.) over a specified period

Understanding Data: Dataset Architecture

Slicing and Dicing

- Refers to a way of segmenting, viewing and comprehending data in a database

- Large blocks of data is cut into smaller segments and the process is repeated until the correct level of detail is achieved for proper analysis

- Presents the data in new and diverse perspectives and provides a closer view of it for analysis

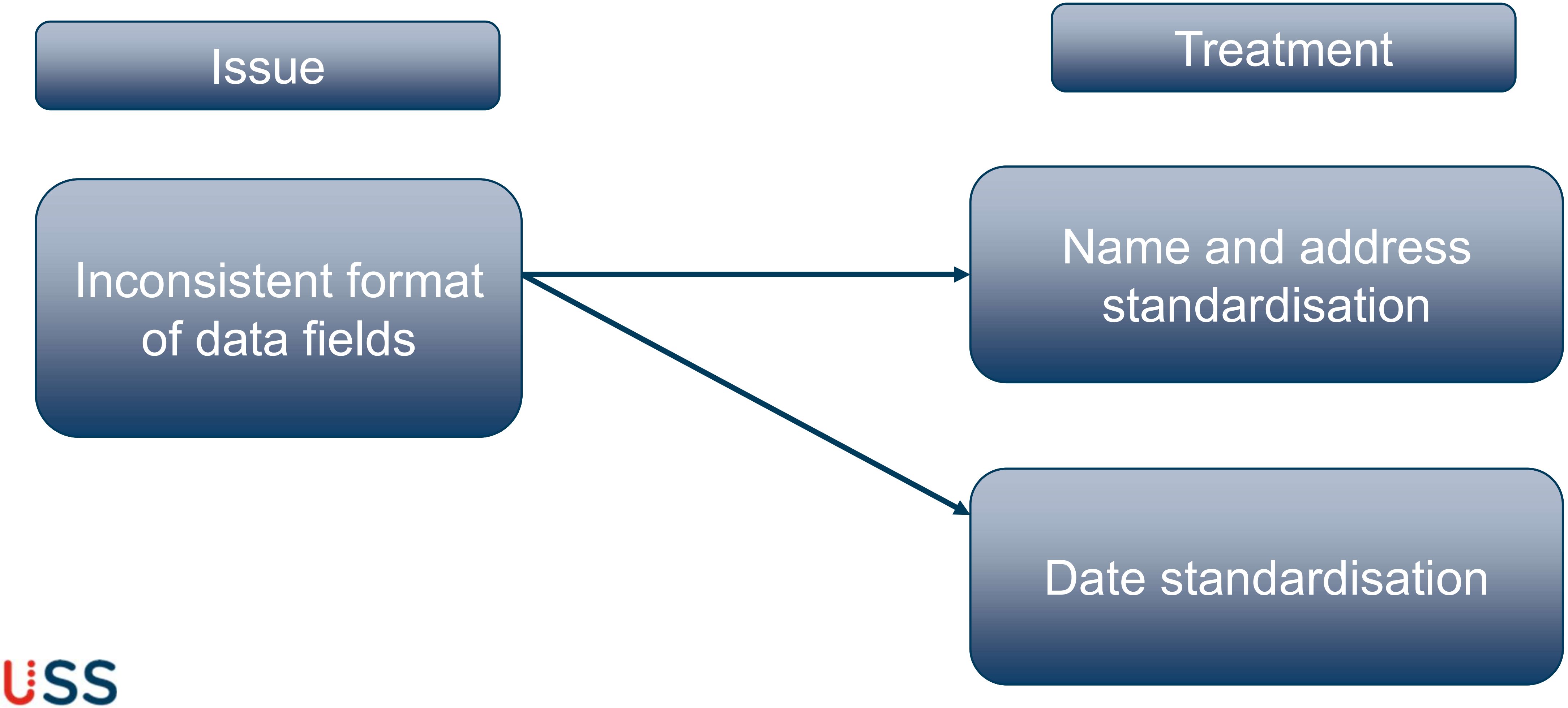
- It is mainly done using the filter actions in the software- *drill down the annual performance of a product to the quarterly level using filters*

Exploratory Analysis and Treatment Methods

SUSS
SINGAPORE UNIVERSITY
OF SOCIAL SCIENCES

Exploratory data analysis & Treatment Methods

1. Visual Examination


Issue
Treatment
Inconsistent format of data fields
Name and address standardisation
Date standardisation

Exploratory data analysis & Treatment Methods

1. Visual Examination- Excel demonstration (global_superstore_2016_raw)

- Select excel file “Ctrl+A” and click on filter tab
- Click on drop down arrow next to “Country” and scroll down the list
  - Two spelling of United States and United Kingdom
  - Filter out the cells with “US” and “UK” to standardize the spelling
- Scroll down the “Order date” column and visually examine the cells
  - Other date format exist
  - Select the entire column and right click. Select “Format Cells” and select the “dd/mm/yyyy” format

Exploratory data analysis & Treatment Methods

2. *Summary Statistics*

Measure	Definition	Issues
Number of observations	Number of non-null or missing values	Missing values
Mean	Average of all the observations present in the table	
Median	Middle value of the collection of data when arranged in ascending order and descending order	
Mode	Value which is repeated maximum number of times in the set	
Minimum & Maximum	Lowest & Highest value of the variable respectively	Inconsistent data values
Range	Maximum value – Minimum value	

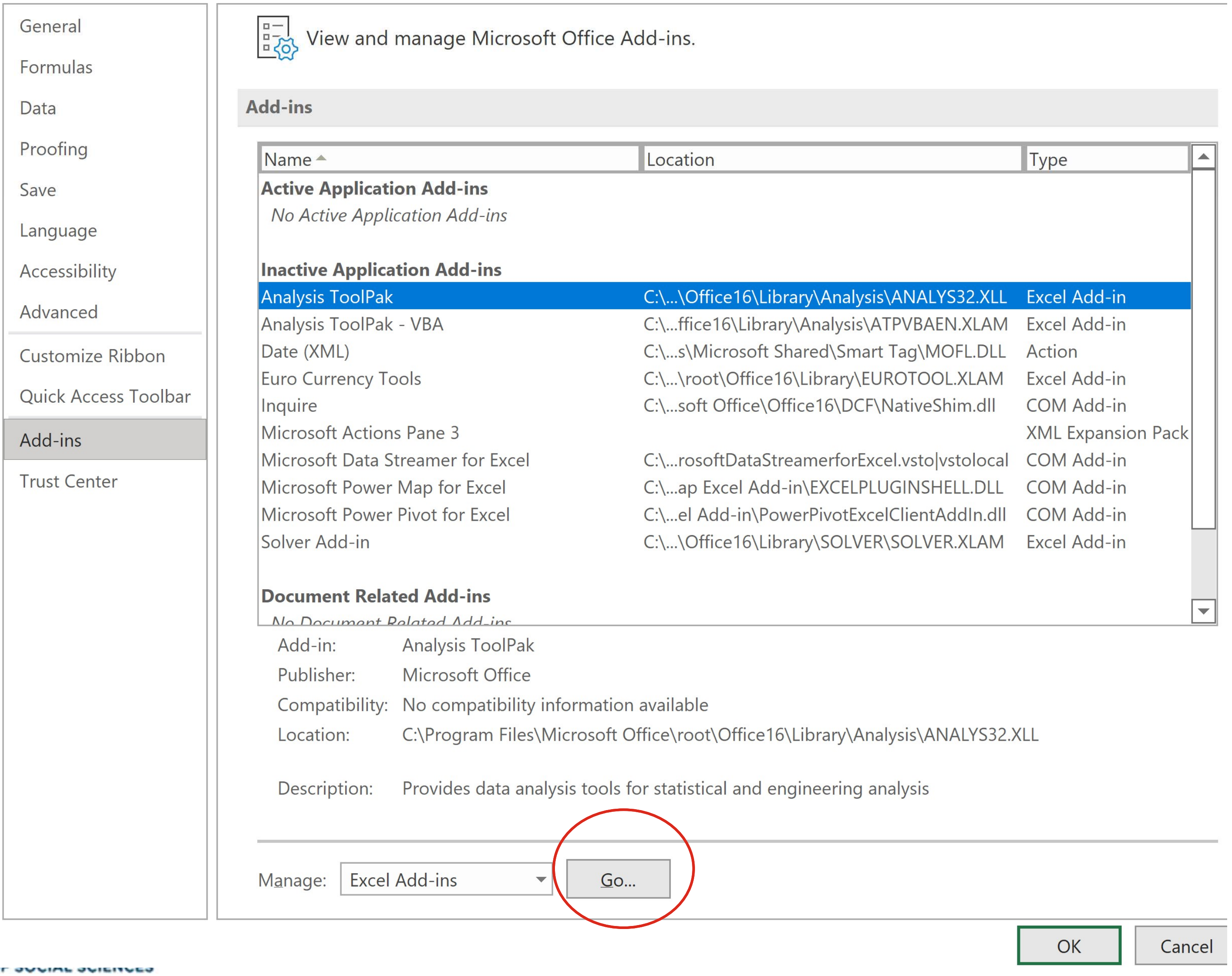
Exploratory data analysis & Treatment Methods

2. Summary Statistics- Installing data analysis toolkit in excel

- Click on “File” tab followed by “Options” followed by “Add-ins”
- Select “Analysis ToolPak” and click “Go” at the bottom of the window
- Select all the options and select ok

Exploratory data analysis & Treatment Methods

2. *Summary Statistics- Installing data analysis toolkit in excel*



- General
- Formulas
- Data
- Proofing
- Save
- Language
- Accessibility
- Advanced
- Customize Ribbon
- Quick Access Toolbar
- Add-ins
- Trust Center

View and manage Microsoft Office Add-ins.

Add-ins

Name	Location	Type
Active Application Add-ins		
No Active Application Add-ins		
Inactive Application Add-ins		
Analysis ToolPak	C:\...\Office16\Library\Analysis\ANALYS32.XLL	Excel Add-in
Analysis ToolPak - VBA	C:\...ffice16\Library\Analysis\ATPVBAEN.XLAM	Excel Add-in
Date (XML)	C:\...s\Microsoft Shared\Smart Tag\MOFL.DLL	Action
Euro Currency Tools	C:\...\root\Office16\Library\EUROTOOL.XLAM	Excel Add-in
Inquire	C:\...soft Office\Office16\DCF\NativeShim.dll	COM Add-in
Microsoft Actions Pane 3		XML Expansion Pack
Microsoft Data Streamer for Excel	C:\...rosoftDataStreamerforExcel.vsto|vstolocal	COM Add-in
Microsoft Power Map for Excel	C:\...ap Excel Add-in\EXCELPLUGINSHELL.DLL	COM Add-in
Microsoft Power Pivot for Excel	C:\...el Add-in\PowerPivotExcelClientAddIn.dll	COM Add-in
Solver Add-in	C:\...\Office16\Library\SOLVER\SOLVER.XLAM	Excel Add-in
Document Related Add-ins		
No Document Related Add-ins		

Add-in: Analysis ToolPak

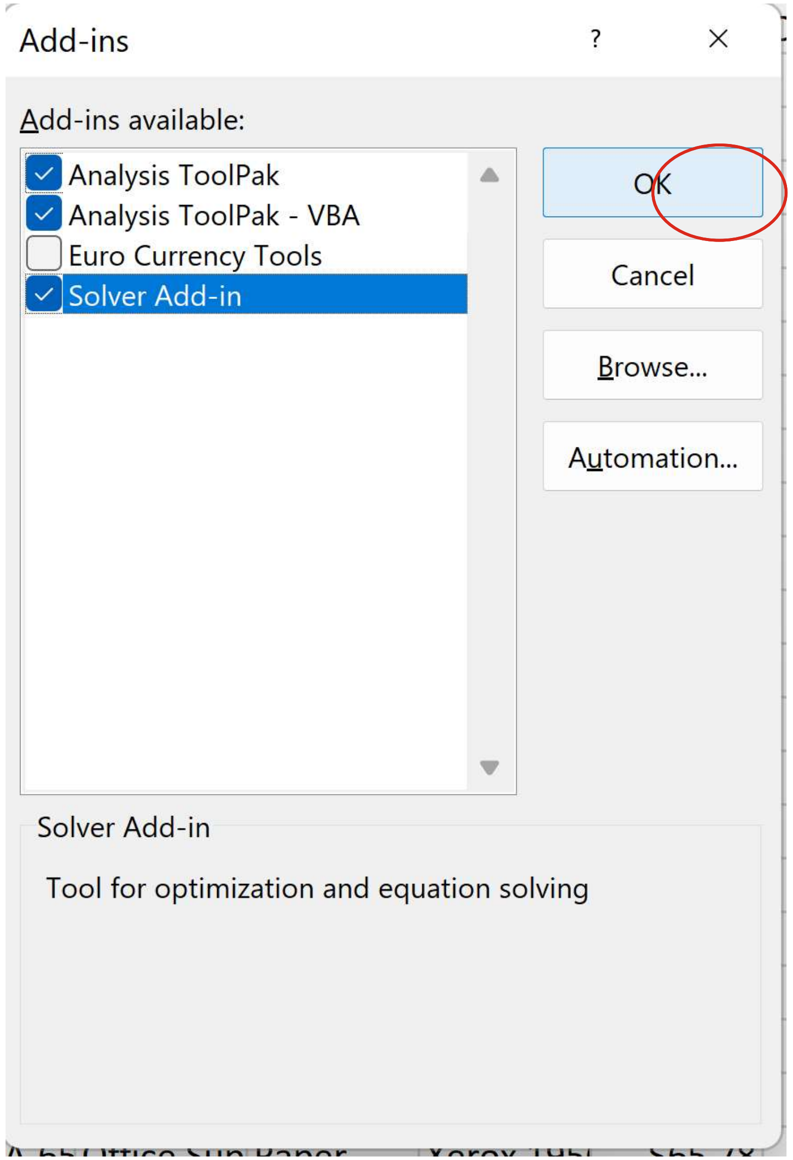
Publisher: Microsoft Office

Compatibility: No compatibility information available

Location: C:\Program Files\Microsoft Office\root\Office16\Library\Analysis\ANALYS32.XLL

Description: Provides data analysis tools for statistical and engineering analysis

Manage: Excel Add-ins



Add-ins

Add-ins available:

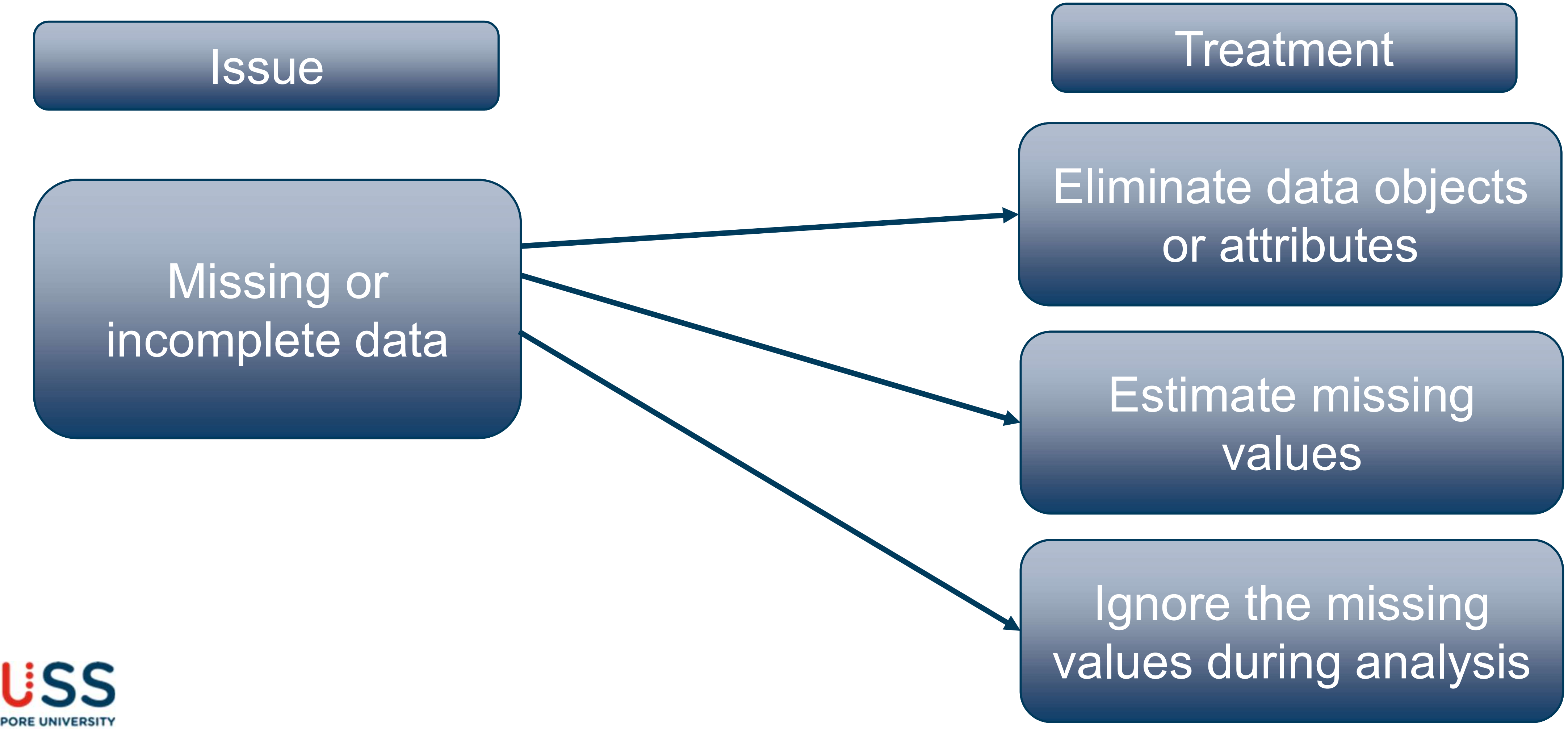
- ☑ Analysis ToolPak
- ☑ Analysis ToolPak - VBA
- □ Euro Currency Tools
- ☑ Solver Add-in

Solver Add-in

Tool for optimization and equation solving

Exploratory data analysis & Treatment Methods

2. Summary Statistics



Issue

Treatment

Missing or incomplete data

Eliminate data objects or attributes

Estimate missing values

Ignore the missing values during analysis

SUSS SINGAPORE UNIVERSITY OF SOCIAL SCIENCES

Exploratory data analysis & Treatment Methods

2. Summary Statistics- Excel Demonstration

- Apply Counta function to all variables. This gives the number of non-blanks observations for each column

- There are 2 missing values in Order ID, 41296 in postal code and 2 missing values in Sales

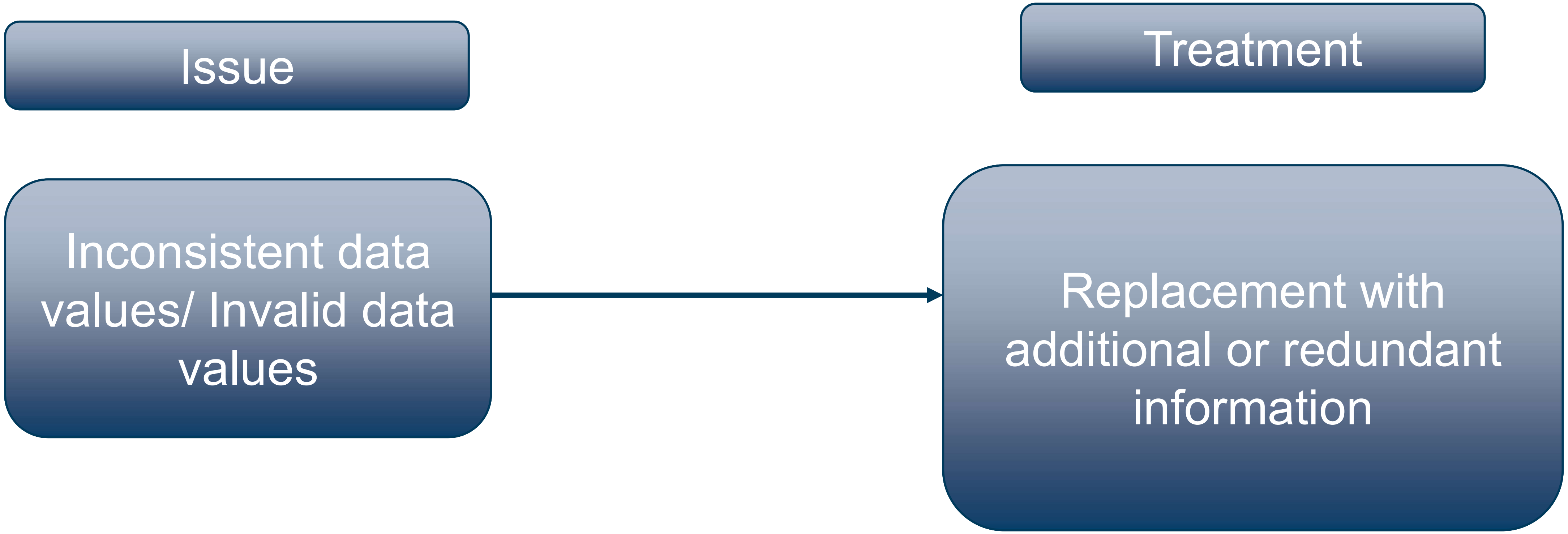
Exploratory data analysis & Treatment Methods

2. Summary Statistics- Excel Demonstration

Treatment Method	How to do it?
Eliminate data objects or attributes	For Mitch Webber, not possible to replace 41296 missing values- can drop postal code from analysis
Estimate missing values	For Greg Hansen, it is possible to replace the order ID For Greg Hansen, customer ID GH-4665138, we can leave the missing value as it is or replace with the mean sales value of “Office supplies” category =$27.1
Ignore the missing values during analysis	Similarity index between two persons based on four attributes, with two missing attributes can be based on balanced two non-missing attributes

Exploratory data analysis & Treatment Methods

2. Summary Statistics



Issue

Treatment

Inconsistent data values/ Invalid data values

Replacement with additional or redundant information

Exploratory data analysis & Treatment Methods

2. *Summary Statistics- Excel Demonstration*

- Use Mode, Median, Mean, Min and Max function in excel to calculate the balance summary statistics

- From minimum statistic for continuous variables, we observe the following:
  - Quantity variable has a value =-5 (quantity cannot be negative)
  - Discount variable has one inconsistent value =-1 (cannot be negative)

- Using drop down menu on the side of the Quantity variable, we also see another inconsistent value:
  - Quantity variable has a value =1.1 (quantity cannot be decimal, has to be integers)

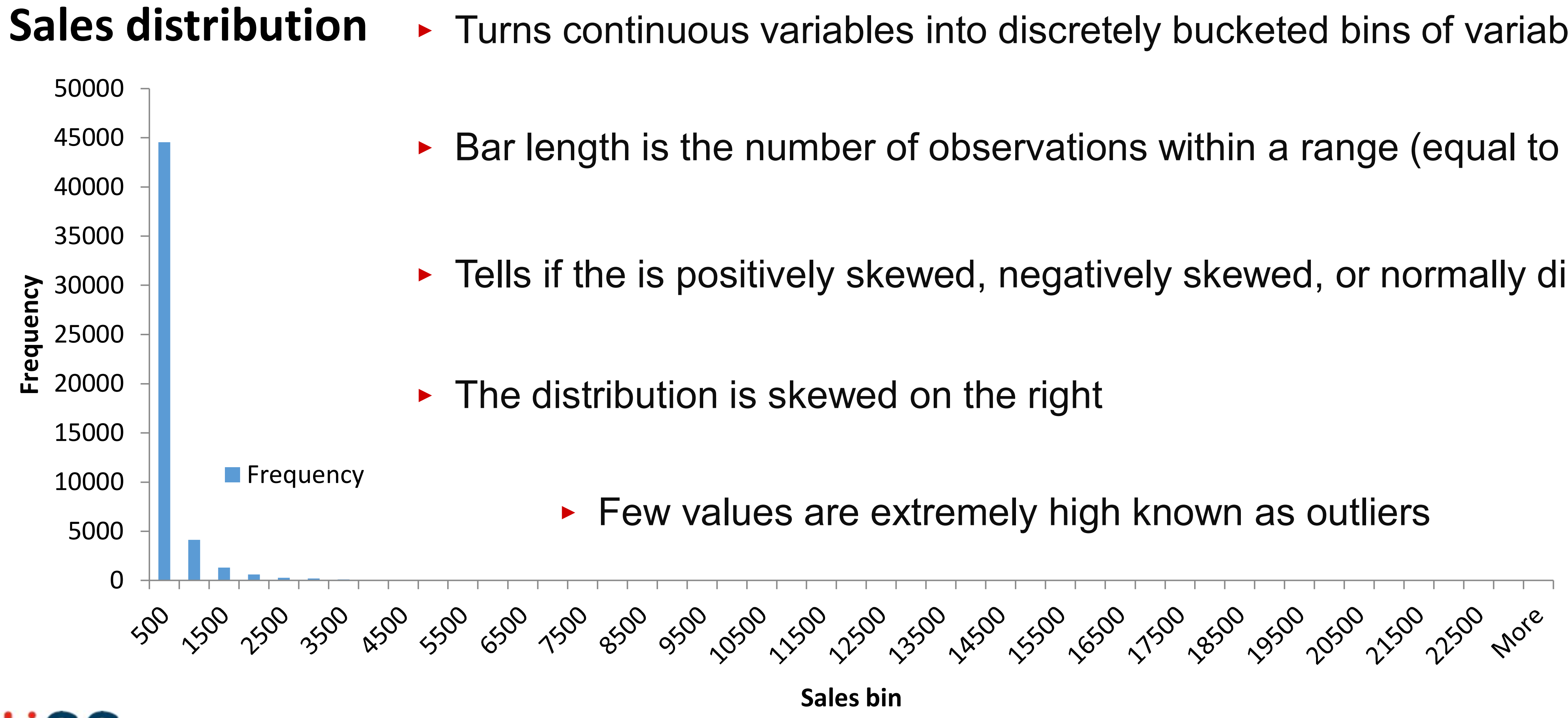
Exploratory data analysis & Treatment Methods

2. Summary Statistics- Excel Demonstration

Treatment Method	How to do it?
Replacement with additional or redundant information	For quantity=1.1, there is only one observation corresponding to this customer for that product name. Ideal is to replace it with null (blank) For Quantity =-5, the customer is Adam Hart with another bill of 5 chairs (different brand). We may replace -5 with 5 if it is really needed. For Discount =-1, customer Emily Grady has ordered other categories with same order ID with discount=0%. Though a high probability that this category also did not have discount, it still cannot be said with 100% surety, hence recommended approach is to replace -1 with null value (blank).

Exploratory data analysis & Treatment Methods

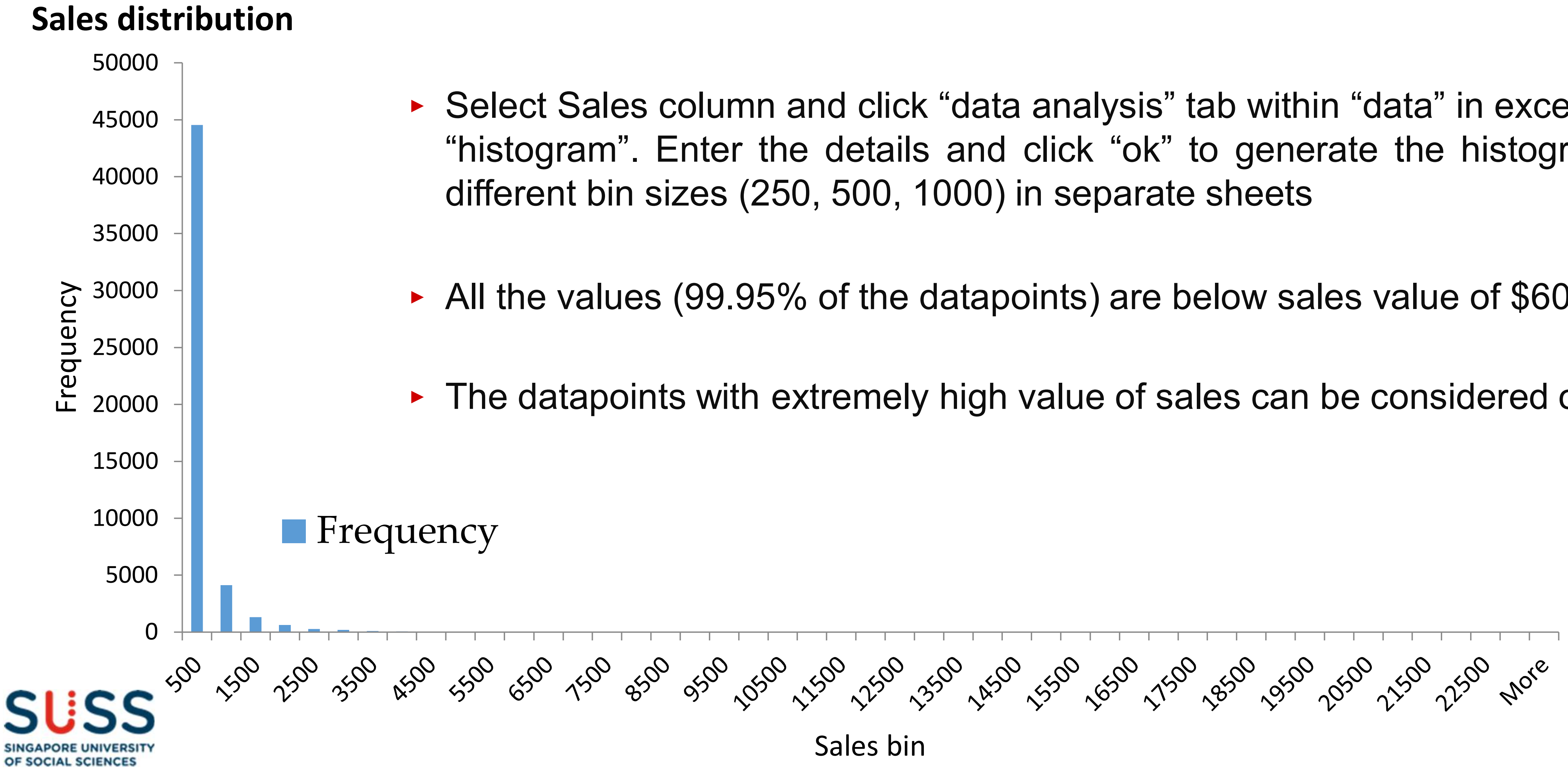
3. Data Distribution- Histogram



- Turns continuous variables into discretely bucketed bins of variables
- Bar length is the number of observations within a range (equal to bin size)
- Tells if the is positively skewed, negatively skewed, or normally distributed
- The distribution is skewed on the right
  - Few values are extremely high known as outliers

Exploratory data analysis & Treatment Methods

3. Data Distribution- Histogram (Excel demonstration)



- Select Sales column and click “data analysis” tab within “data” in excel. Select “histogram”. Enter the details and click “ok” to generate the histogram with different bin sizes (250, 500, 1000) in separate sheets
- All the values (99.95% of the datapoints) are below sales value of $6000
- The datapoints with extremely high value of sales can be considered outliers

Exploratory data analysis & Treatment Methods

3. Data Distribution- Box and Whisker plot

Upper Whisker (Q3+1.5 IQR)

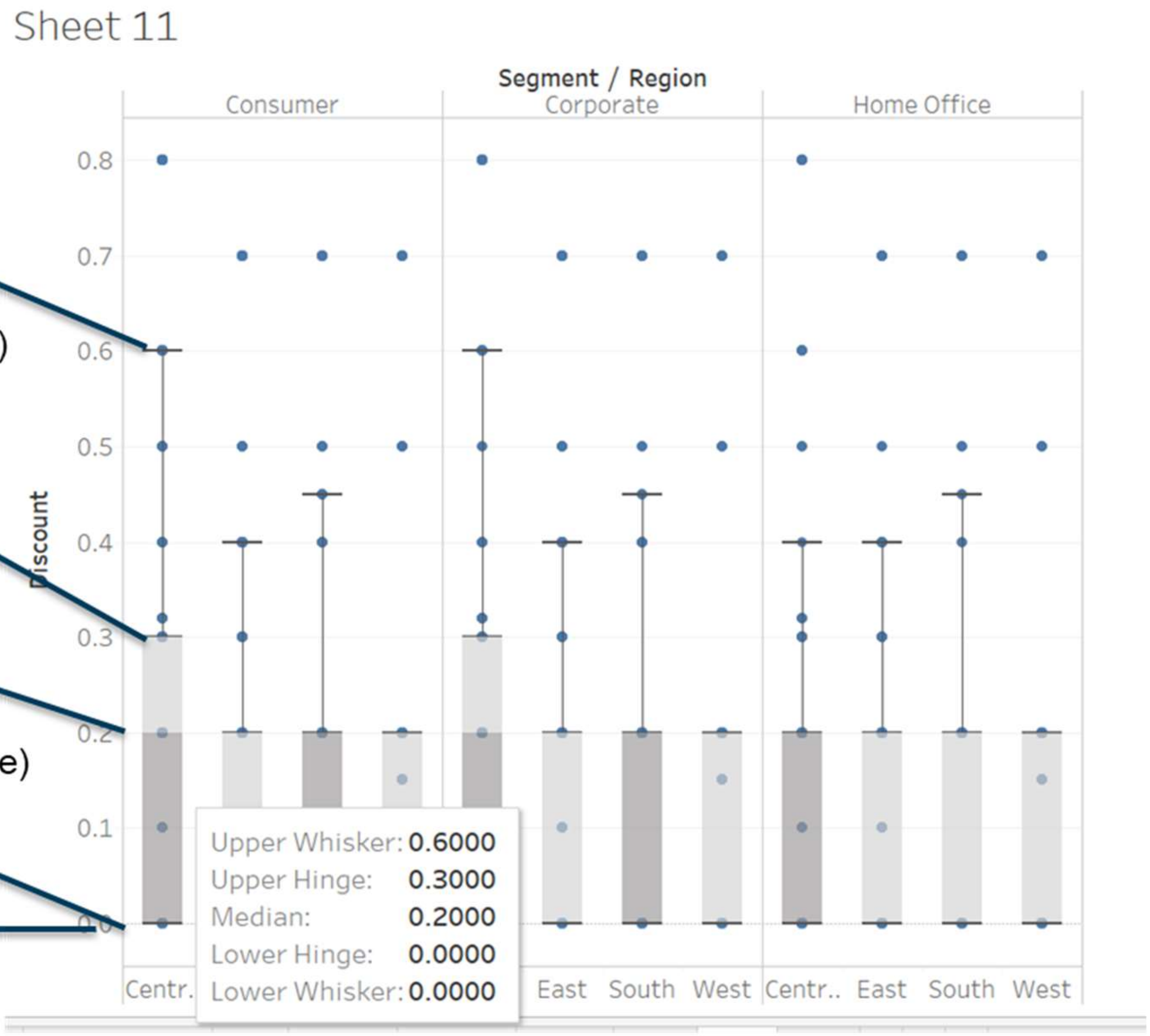
IQR= (Q3-Q1)

Upper Hinge (Q3, 75% percentile)

Median (Q2, 50% percentile)

Lower Hinge (Q1, 25% percentile)

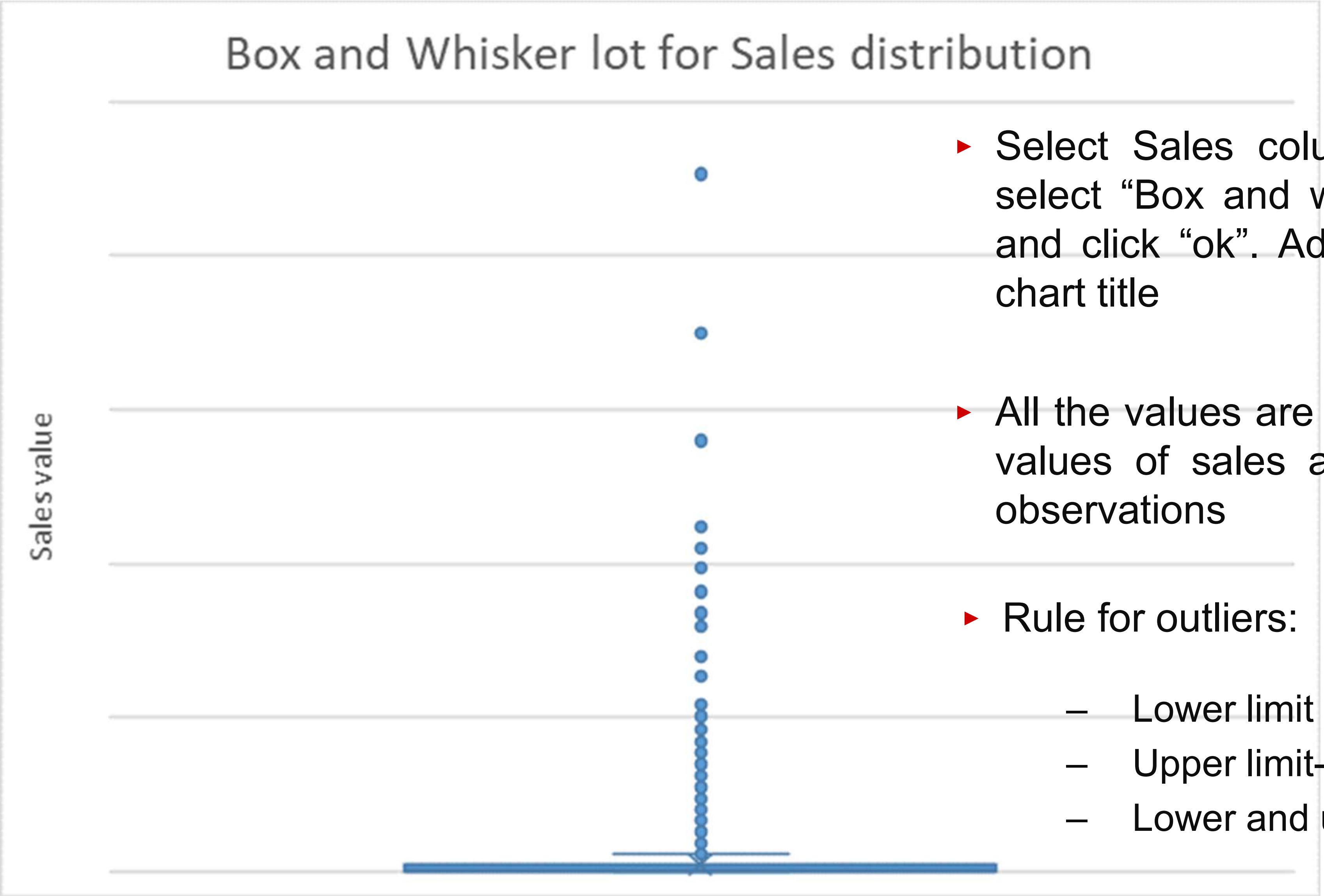
Lower Whisker (Q1-1.5 IQR)



Upper Whisker: 0.6000
Upper Hinge: 0.3000
Median: 0.2000
Lower Hinge: 0.0000
Lower Whisker: 0.0000

Exploratory data analysis & Treatment Methods

3. *Data Distribution- Box and Whisker plot (Excel demonstration)*



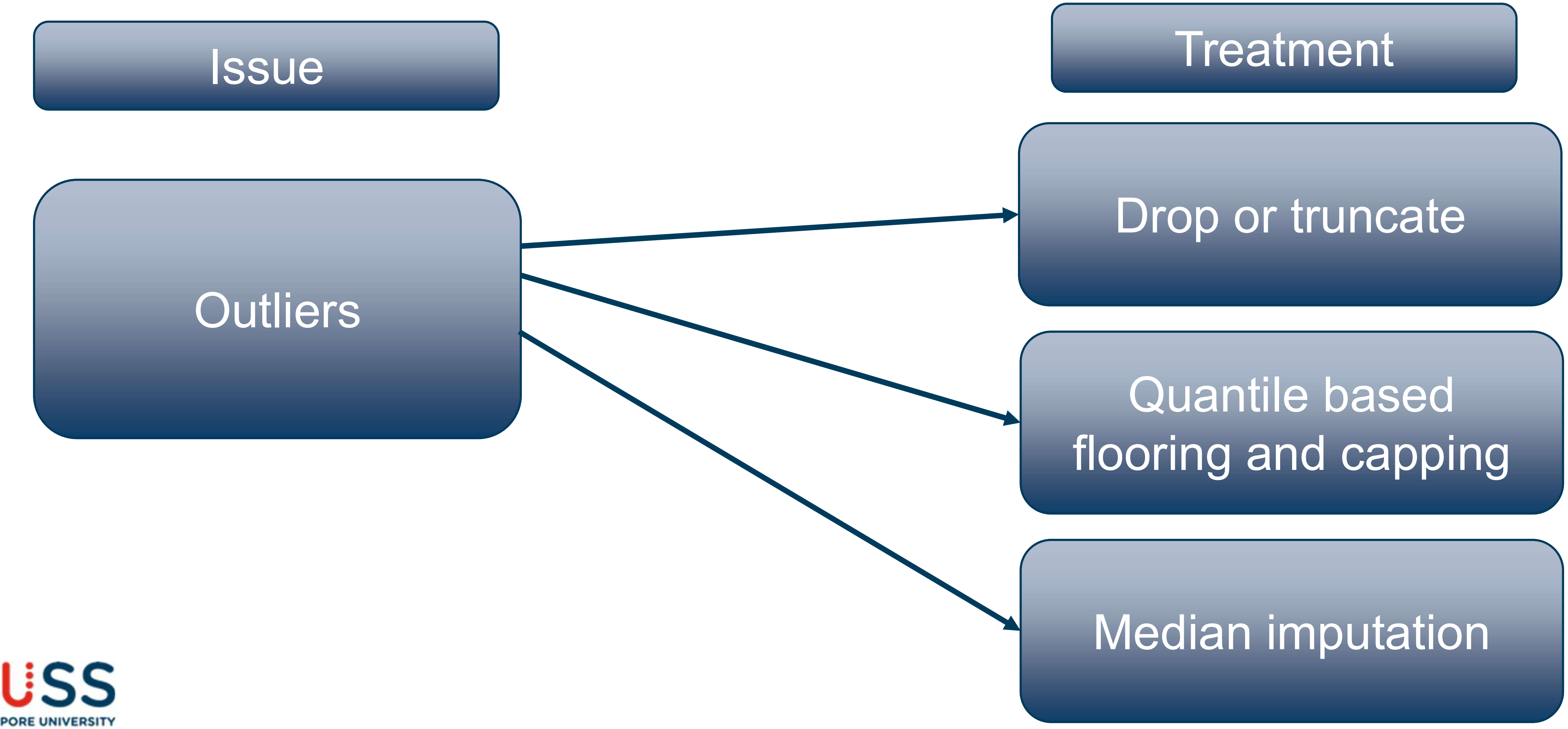
- Select Sales column and click “Insert chart”. Then select “Box and whisker plots” from within all charts and click “ok”. Adjust the axes titles and type in the chart title

- All the values are concentrated at the bottom or lower values of sales and the circles represent individual observations

- Rule for outliers:
  - Lower limit -5th to 10th percentile
  - Upper limit- 90th to 95th percentile
  - Lower and upper whisker

Exploratory data analysis & Treatment Methods

3. Data Distribution



Exploratory data analysis & Treatment Methods

3. Data Distribution- Excel demonstration

Outlier criteria	Limits	Number of observations	Treatment method
10th and 90th percentile	Lower- $14 Upper- $632	10427	Drop observations Replace with the capped value (90th or 95th percentile or upper whisker) Replace with the median value= $85
5th & 95th percentile	Lower- $9 Upper- $1016	5259	
Box and whisker plots	Lower whisker- (-$315) Upper whisker-$591	5544	

Summary

- There are four basic stages in the data visualisation process- data collection and storage, data pre-processing, graphics engine and human visual and cognitive processing
- Data collection is the methodological process of gathering information about a specific subject
- Data preparation is the process of combining, structuring and organising data so it can be used in business intelligence (BI), analytics and data visualisation applications
- Exploratory data analysis methods can be used to identify the data anomalies and appropriate treatment method should be used to prepare data