Technische Universität Berlin

Quantitative Decision Making: Metrics and Performance Indication with Spreadsheet Software

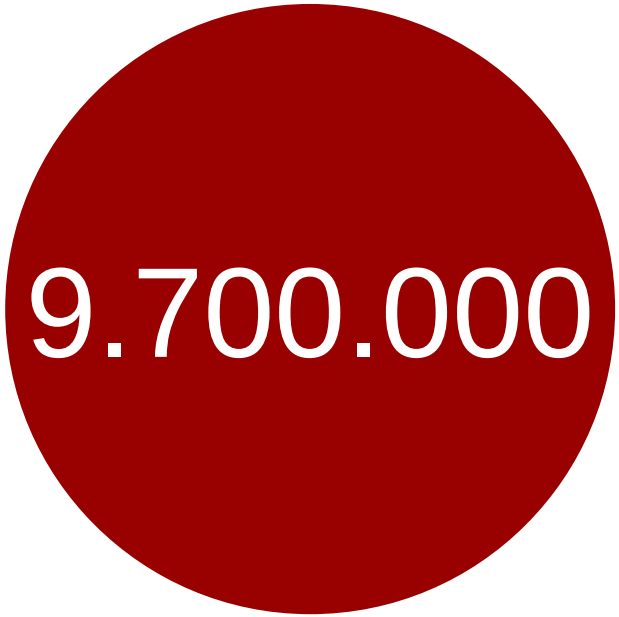
Fakultät Wirtschaft & Management, Institut für Technologie und Management

25. July 2018

Agenda

1. Data cleaning
2. Performance Evaluation
3. Finding Errors

Discussion



... $ are the average financial impact on companies due to poor data quality.

- It is assumed that these costs are underestimated

Few companies measure data quality

- 8% have formal measuring systems, 22% have informal measuring systems, 59% have no measuring systems

Recommendation against poor data quality is the use of a data quality tool

- These have a rising market
- The software focuses primarily on man-made data

Data Quality Assurance programs are also recommended, including

- Data Quality roles (e.g. Data Quality Steward)
- Business Use Cases
- Monitoring and measuring processes
- Data Quality specific guidelines

Data Preparation - Introduction

- Data may be of poor quality due to
  - Inconsistency (for example, different values represent the same)
  - Incompleteness / Missing values
  - duplicates
  - Impossible values
  - runaways
  - Error during data entry / manual entry (e.g. lower and upper case)
  - Problems of merging several data sources (e.g. different aggregation of data or units)
- A common principle of data analysis is "garbage in, garbage out".
  - Poor data quality will lead to poor models
- If the data quality is adjusted, the steps should be...
  - well-founded / reasonable
  - validated
  - documented
  - ...and also executed

Source: Baesens (2014), Cielen (2016)

Sample

- Some technologies already allow to analyze huge amounts of data - samples are still useful

➡ the sample should be representative of the test objective, e.g.:

  - Data on events closer in time Significance of events to be expected in the short term higher
  - Restriction to a specific target group
  - Distorted distribution equalize situation: e.g. ratio of 90% to 10% on-time delivery to non-on-time delivery interest: What differentiates on-time deliveries from non-on-time (~50% to 50% helpful)

Source: Baesens (2014)

Strategies of data cleaning

Erors	Strategies
▪ Data has wrong entries	▪ E.g. Using Frequency table to identifiy (e.g. spelling mistakes) and change manually
▪ Impossible values	▪ Define range of plausible values, identify values outside that range and change manually
▪ Missing values	▪ Delete observation or estimate value
▪ Outliers	▪ Use Scatter plot, Histogram or boxplot to identify and delete, substitute or use value if plausible

Source: Cielen (2016)

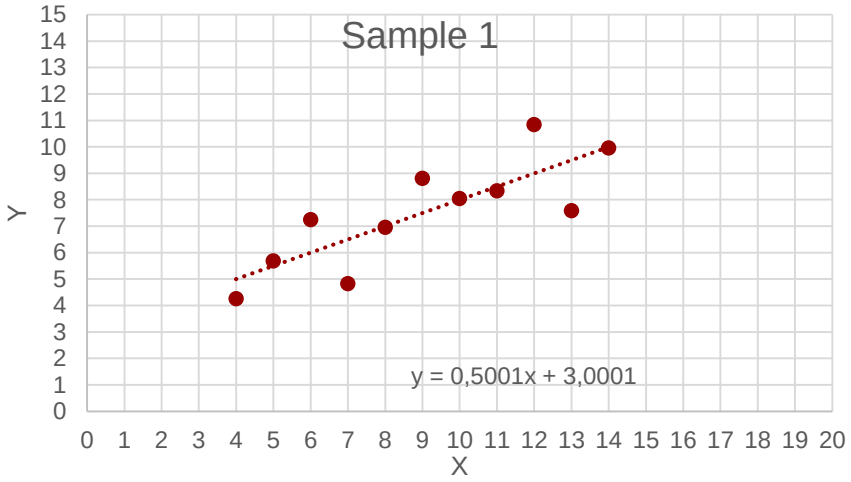
Visual Data Exploration
Motivation

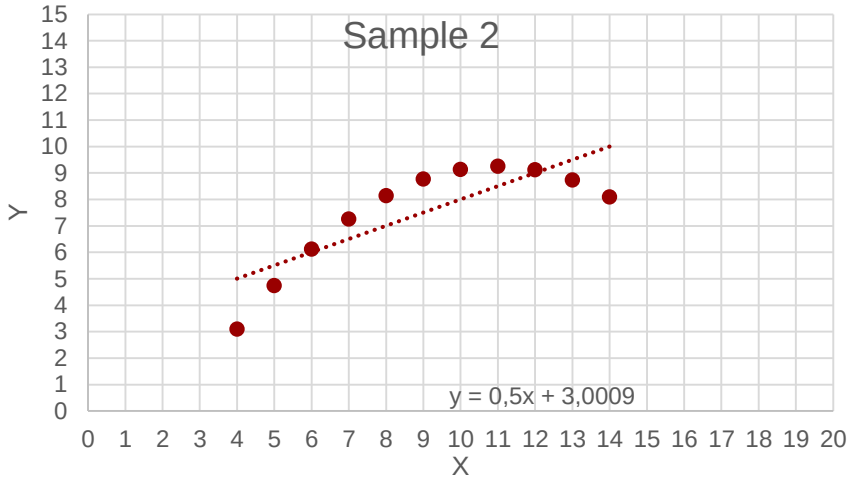
Anscombe Quartet:

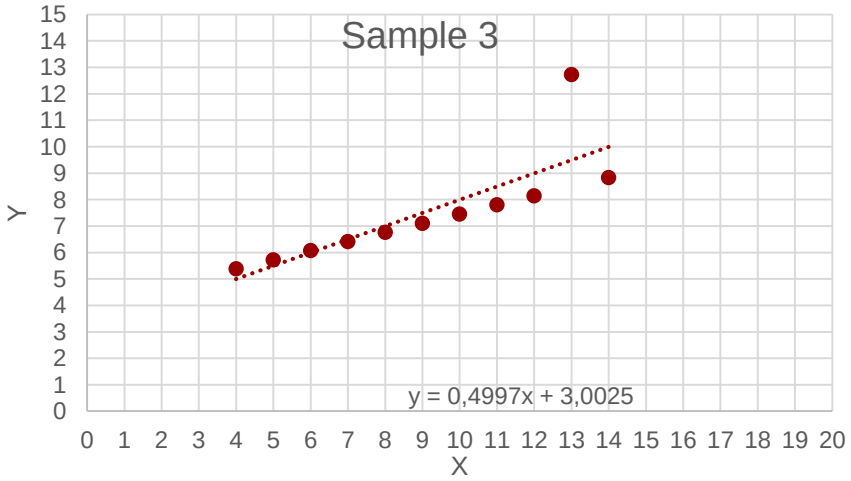
Sample 1		Sample 2		Sample 3		Sample 4	
x	y	x	y	x	y	x	y
10,0	8,04	10,0	9,14	10,0	7,46	8,0	6,58
8,0	6,95	8,0	8,14	8,0	6,77	8,0	5,76
13,0	7,58	13,0	8,74	13,0	12,74	8,0	7,71
9,0	8,81	9,0	8,77	9,0	7,11	8,0	8,84
11,0	8,33	11,0	9,26	11,0	7,81	8,0	8,47
14,0	9,96	14,0	8,10	14,0	8,84	8,0	7,04
6,0	7,24	6,0	6,13	6,0	6,08	8,0	5,25
4,0	4,26	4,0	3,10	4,0	5,39	19,0	12,50
12,0	10,84	12,0	9,13	12,0	8,15	8,0	5,56
7,0	4,82	7,0	7,26	7,0	6,42	8,0	7,91
5,0	5,68	5,0	4,74	5,0	5,73	8,0	6,89

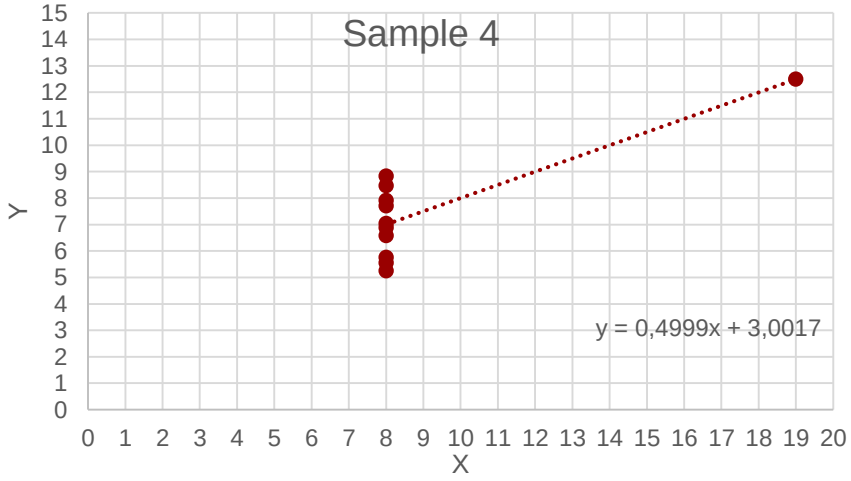
Mean X	9		9		9		9	
Mean Y		7,50		7,50		7,50		7,50
(Sample) Var X	11		11		11		11	
(Sample) Var Y		4,12726909		4,12726909		4,12726909		4,12726909
Correlation X and Y	0,81642052		0,81623651		0,81628674		0,81642052	
Linear Regression line	y = 3,00 + 0;500x		y = 3,00 + 0;500x		y = 3,00 + 0;500x		y = 3,00 + 0;500x	

Visual Data Exploration
Motivation – Anscombe Quartett visualized

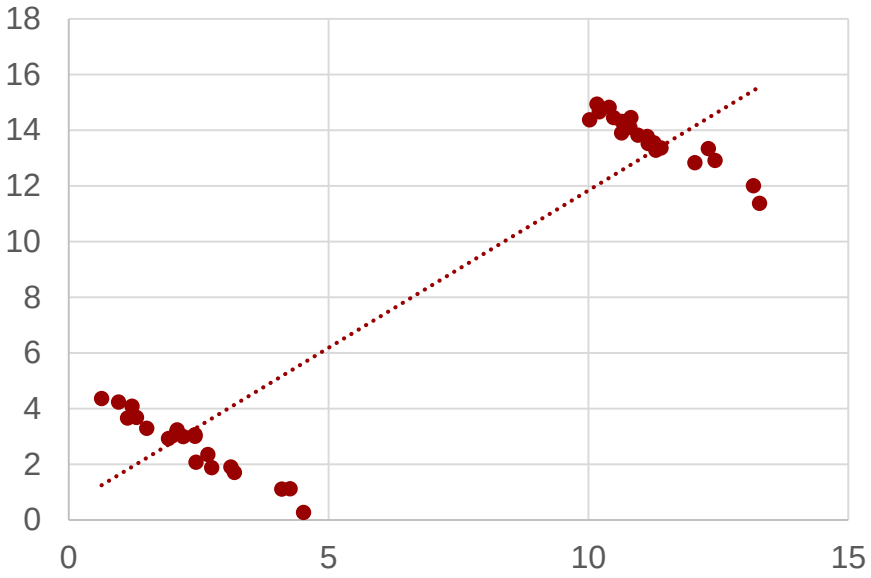








Visual Data Exploration
Motivation – Simpson‘ Paradox



- Within the data, undetected groups may exist
- In this example, the linear regression line (y = 0,53+1,13x) has a coefficient of determination (R^2) of 0,85

Visual Data Exploration
Heatmap

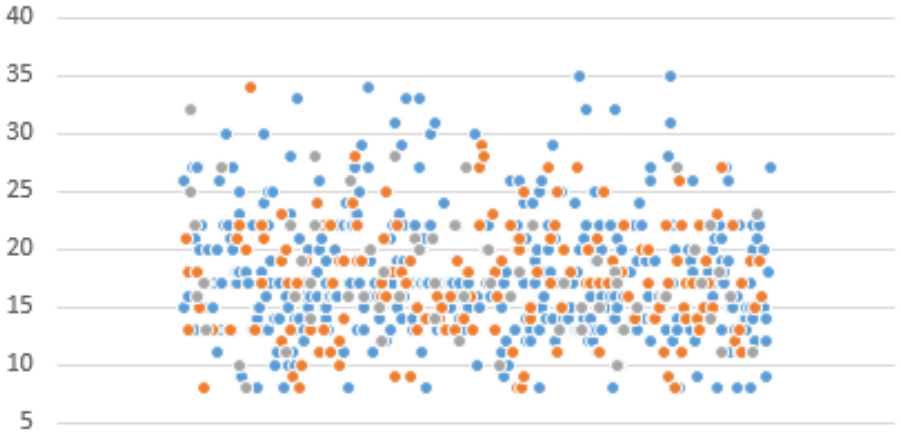
- Shading values according to desired criteria
  - e.g. particularly large or small values red, opposite values green respectively
  - Shows: "Density" of values (frequency of a value in certain ranges)
- Exploration of the data
  - patterns
  - Extreme values (e.g. outliers, minimum, maximum)

	1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10	11
2	3	4	5	6	7	8	9	10	11	12
3	4	5	6	7	8	9	10	11	12	13
4	5	6	7	8	9	10	11	12	13	14
5	6	7	8	9	10	11	12	13	14	15
6	7	8	9	10	11	12	13	14	15	16
7	8	9	10	11	12	13	14	15	16	17
8	9	10	11	12	13	14	15	16	17	18
9	10	11	12	13	14	15	16	17	18	19
10	11	12	13	14	15	16	17	18	19	20

Source: Kirk (2012)

Visual Data Exploration
Scatter Plot

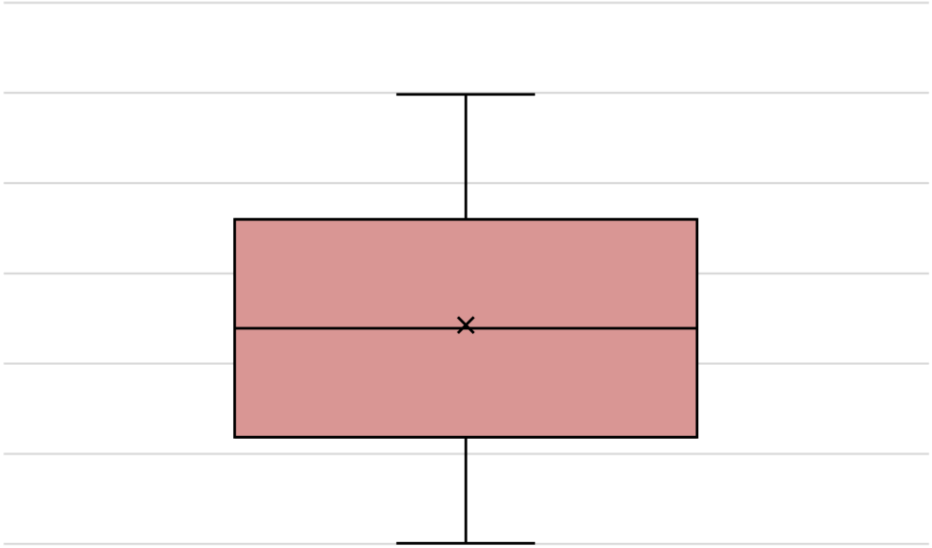
- Behavior of two variables to each other as points in two-dimensional space
  - (Unorganized) clear display of all values
  - Shows: Overview of difficult-to-understand values
- Exploration of the data
  - Anomalies (e.g. outliers)
  - Relationships of variables
  - Density of values



Source: Dietrich et al, (2015)

Visual Data Exploration
Boxplot

- Compact representation of the descriptive information of a distribution
  - One-dimensional representation!
  - Shows: Overview of the distribution of values
- Exploration of the data
  - Centre of distribution
  - diversification
  - symmetry
  - skewness

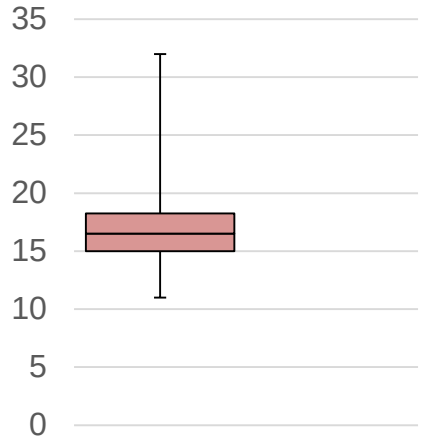


Source: Dietrich et al, (2015); Cielen (2016)

Visual Data Exploration
Boxplot II

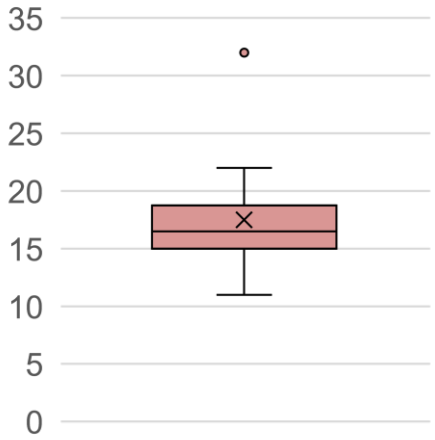
Skeletal Boxplot (5 Values -> 4 Ranges)

- Upper Whisker = Maximum **to** upper Quartile (Median of upper half)
- Lower Box = upper Quartile **to** median (Odd number of observations : Value exactly in the middle / even number of observations: Average of the two mean values)
- lower Box = Median **to** lower quartile (Median of lower half)
- lower Whisker = lower quartile **to** minimum



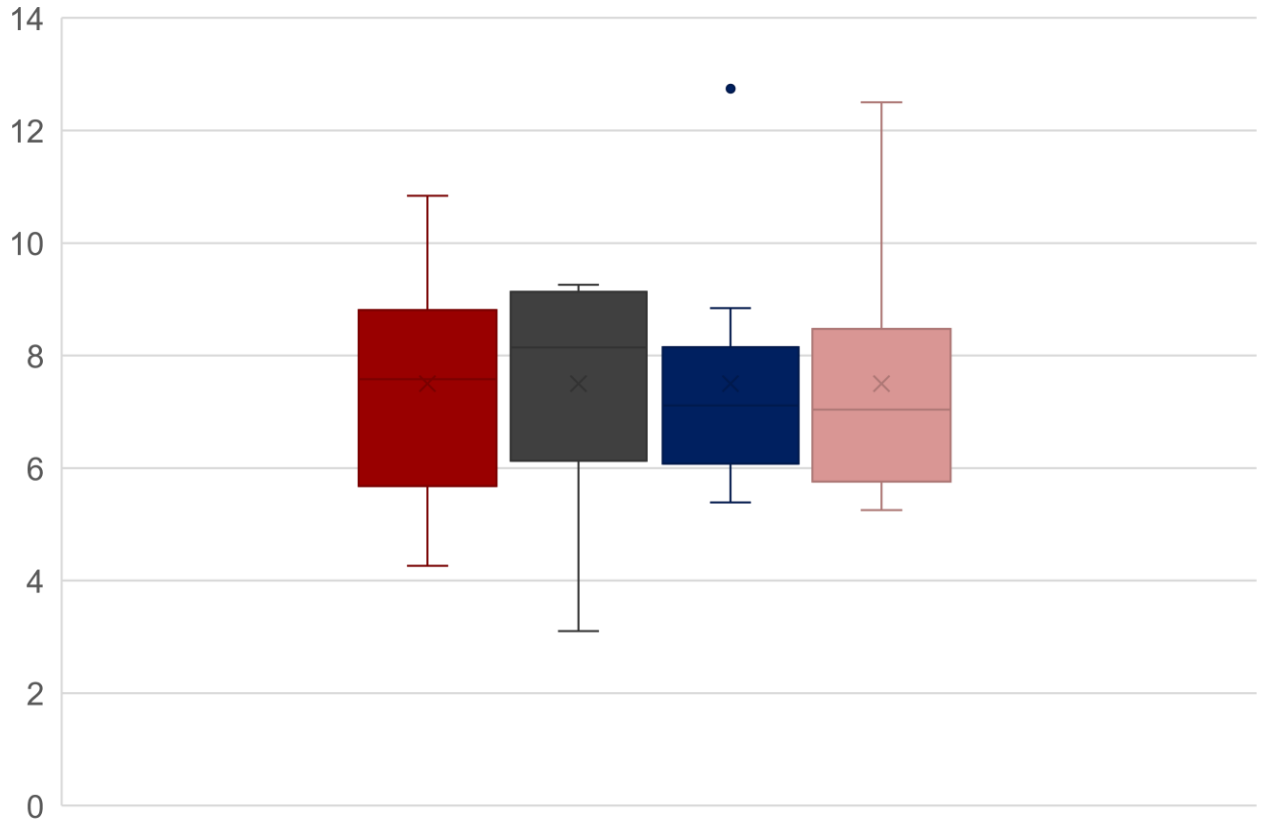
Modified Boxplot (excel 2016 Standard):

- Interquartile Range (IQR):
  0,25 Quantile **to** 0,75 Quantile
- Extreme *outlier*: > 3x IQR from 0,25 or 0,75 quantile
- moderate *outlier*: > 1,5x IQR from 0,25 or 0,75 quantile
- Whiskers: from 0,25 / 0,75 quantil 1,5x IQR OR Minimum / Maximum



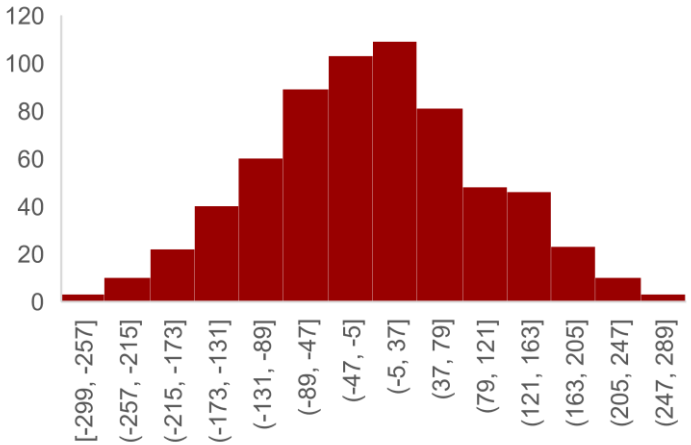
Source: Peck et al. (2015)

Visual Data Exploration
Boxplot for Ascombe Quartets (Y Values)



Visual Data Exploration
Histogramm

- Displaying the distribution of a variable in intervals
  - Intervals are mapped with quantity or frequency
  - One-dimensional representation!
  - Shows: Overview of distribution of values
  - Thumb rule for the number of intervals for continuous data: $= \sqrt{number\ of\ observations}$
- Exploration of the data
  - Center of distribution or typical values
  - diversification
  - General form of distribution
  - Position and number of distribution peaks
  - Position of gaps and outliers



Source: Peck et al. (2015)

Problem 3-1 – Data Import and finding errors

- BEAR Electronics produces a variety of electronic products in Europe and uses the logistics service providers AHL Express, Bange & Hammer, and the Combined Parcel Service (CPS) for delivery. Your task is to evaluate the performance of the processes in the past months. The recording of customer orders and provision of the goods is recorded as warehouse processing time and the transport as transit time.

a) Import the data from the.csv file

b) Calculate the average order cycle time of existing orders

c) Calculate the item fill rate of the existing purchase order

d) Calculate the on-time delivery rate of existing orders

Variables in the data set

OrderNo - Order number

ItemsOrdered - Number of ordered products in order

PriceTotal - Order volume (financial)

Customer - customer of the order

CustomerName - Name of the customer

Location - Location of the customer

ServiceLevel - Agreement on time between order and receipt of goods for order

LSP - The service provider executing the order

TimeWarehouse - Warehouse processing time

TimeTransit - Time of pure transport

ItemsReceived- Number of products received in order

Problem 3-2

- In order to evaluate the performance of the individual business units (sales, warehousing and transport), the key figures are to be displayed separately.

a) Determine the sales volume per country to valuate the sales units

b) Determine the item fill rate per customer in order to draw initial conclusions about the quality of warehousing.

c) Determine the average order transit time per logistics service provider per service level to evaluate the logistics service providers

Variables in the data set

OrderNo - Order number

ItemsOrdered - Number of ordered products in order

PriceTotal - Order volume (financial)

Customer - customer of the order

CustomerName - Name of the customer

Location - Location of the customer

ServiceLevel - Agreement on time between order and receipt of goods for order

LSP - The service provider executing the order

TimeWarehouse - Warehouse processing time

TimeTransit - Time of pure transport

ItemsReceived- Number of products received in order

Problem 3-3

- It is noticeable that some values cannot be correct. The record obviously contains errors that need to be corrected.

a) Create a frequency table of orders per service provider to find incorrect entries and replace them.

b) Search for missing values in the customer's location and replace them.

Variables in the data set

OrderNo - Order number

ItemsOrdered - Number of ordered products in order

PriceTotal - Order volume (financial)

Customer - customer of the order

CustomerName - Name of the customer

Location - Location of the customer

ServiceLevel - Agreement on time between order and receipt of goods for order

LSP - The service provider executing the order

TimeWarehouse - Warehouse processing time

TimeTransit - Time of pure transport

ItemsReceived- Number of products received in order

Problem 3-3

- It is noticeable that some values cannot be correct. The record obviously contains errors that need to be corrected.

c) Use a green-yellow-red scale to identify errors in the item fill rate and correct them by using the order quantity as the delivery quantity.

d) Create a boxplot of Transit Time to determine if one service provider behaves in a conspicuously different way than the others.

e) Create a histogram of the transit for the conspicuous service provider from 4-3D to better identify gaps and outliers in the distribution.

Variables in the data set

OrderNo - Order number

ItemsOrdered - Number of ordered products in order

PriceTotal - Order volume (financial)

Customer - customer of the order

CustomerName - Name of the customer

Location - Location of the customer

ServiceLevel - Agreement on time between order and receipt of goods for order

LSP - The service provider executing the order

TimeWarehouse - Warehouse processing time

TimeTransit - Time of pure transport

ItemsReceived- Number of products received in order

Problem 3-3

- It is noticeable that some values cannot be correct. The record obviously contains errors that need to be corrected.

f) Create a point cloud of the transit time (Y-axis) to the OrderNo to check whether a time pattern exists. If you find a temporal anomaly, delete it from the data.

g) Calculate the general item fill rate, the on-time delivery rate per service provider per service level and graphically display the distribution of orders among the service providers.

Variables in the data set

OrderNo - Order number

ItemsOrdered - Number of ordered products in order

PriceTotal - Order volume (financial)

Customer - customer of the order

CustomerName - Name of the customer

Location - Location of the customer

ServiceLevel - Agreement on time between order and receipt of goods for order

LSP - The service provider executing the order

TimeWarehouse - Warehouse processing time

TimeTransit - Time of pure transport

ItemsReceived- Number of products received in order

Summary: Metric and Indicators

- The quality of the data is decisive for the quality of the key figures and later the models ("garbage in, garbage out")
  - Therefore, the data must be checked for impossible values, missing values, outliers, entry errors and others.
  - The errors must be improved by correcting, deleting or estimating the value
- Visual methods for explorative data analysis are helpful for troubleshooting and for understanding the data (see Lecture01).

References

- Peck, R.; Olsen, C.; Devore, J.L. (2015): Introduction to Statistics and Data Analysis.