MATES ED2MIT
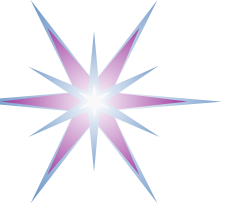Education and Training for Data Driven Maritime Industry

Tutorial D04

Data Analysis Principles and Techniques
Exploratory Data Analysis

Yuri Demchenko MATES Project

University of Amsterdam

Maritime Alliance for fostering the European Blue economy through a Marine Technology Skilling Strategy

Outline

- General aspects of Data Analysis
  - Concepts of Data Analysis
  - Principles of Data Analysis
- Data Analysis techniques
  - Some tips for data analysis
- General aspects of the Exploratory Data Analysis
- Example EDA: Procrastination

Data Analysis – "The Concept"

- Approach to de-synthesizing data, informational, and/or factual elements to answer research questions

- Method of putting together facts and figures
  to solve research problem

- Systematic process of utilizing data to address research questions

- Breaking down research issues through utilizing controlled data and factual information
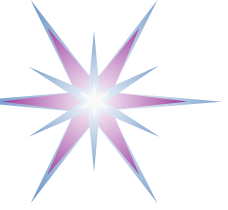
Which one to use?

- Nature of research
  - Descriptive in nature?
  - Attempts to "infer", "predict", find "cause-and-effect", "influence", "relationship"?
  - Is it both?
- Research design (incl. variables involved)
- Outputs/results expected
  - research issue
  - research questions
  - research hypotheses

At post-graduate level research, failure to choose the correct data analysis technique is an almost safe ingredient for thesis failure.

Principles of analysis – Goals of analysis

- Goal of an analysis:
  - To explain cause-and-effect phenomena
  - To relate research with real-world event
  - To predict/forecast the real-world phenomena based on research
  - Finding answers to a particular problem
  - Making conclusions about real-world event based on the problem
  - Learning a lesson from the problem

Principles of analysis (contd.)

- Data can't "talk"
- An analysis contains some aspects of scientific reasoning/argument:
  - Define
  - Interpret
  - Evaluate
  - Illustrate
  - Discuss
  - Explain
  - Clarify
  - Compare
  - Contrast

Principles of analysis (contd.)

- An analysis must have four elements:
  - Data/information (what)
  - Scientific reasoning/argument (what? who? where? how? what happens?)
  - Finding (what results?)
  - Lesson/conclusion (so what? so how? therefore,…)

Principles of data analysis

- Basic guide to data analysis:
  - "Analyse" NOT "narrate"
  - Go back to research flowchart
  - Break down into research objectives and research questions
  - Identify phenomena to be investigated
  - Visualise the "expected" answers
  - Validate the answers with data
  - Don't tell something not supported by data

Principles of data analysis (contd.)

Shoppers	Number
Male	
Old	6
Young	4
Female	
Old	10
Young	15

More female shoppers than male shoppers

More young female shoppers than young male shoppers

Conclusion: Young male shoppers are not interested to shop at the shopping complex

Data analysis (contd.)

- When analysing:
  - Be objective
  - Accurate
  - True
- Separate facts and opinion
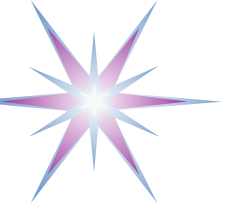- Avoid "wrong" reasoning/argument. E.g. mistakes in interpretation.

Quantitative Data Analysis

- Definitions
- Examples of a data set
- Creating a data set
- Displaying and presenting data – frequency distributions
- Grouping and recoding
- Visual presentations
- Summary statistics, central tendency, variability
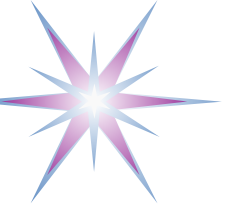
What do we analyze?

- **Variable** – characteristic that varies
- **Data** – information on variables (values)
- **Data set** – lists variables, cases, values
- **Qualitative variable** – discrete values, categories.
  - Frequencies, percentages, proportions
- **Quantitative variable-** range of numerical values
  - Mean, median, range, standard deviation, etc.

Creating a data set

- Entering/Retrieving data into statistical program
- May involve coding and data entry
- Coding = assigning numerical value to each value of a variable
  – Gender: 1= male, 2 = female
  – Year in school: 1= freshman, 2= sophomore, etc.
  – May need codes for missing data (no response, not applicable)
  – Large data sets come with codebooks

Displaying and Presenting Data

- Frequency distribution – list of all possible values of a variable and the # of times each occurs
  - May require grouping into categories
  - May include percentages, cumulative frequencies, cumulative percentages

Categories of Data Analysis

- Narrative (e.g. laws, arts)
- Descriptive (e.g. social sciences)
- Statistical/mathematical (pure/applied sciences)
- Audio-Optical (e.g. telecommunication)
- Others

Most research analyses, arguably, adopt the first three.

The second and third are, arguably, most popular in pure, applied, and social sciences

Descriptive statistics

- Use sample information to explain/make abstraction of population "phenomena".
- Common "phenomena":
  - Association (e.g. $\sigma_{1,2.3} = 0.75$)
  - Tendency (left-skew, right-skew)
  - Causal relationship (e.g. if X, then, Y)
  - Trend, pattern, dispersion, range
  - Trends are similar but no logical relation between X and Y (possibly dependence on value z)
- Used in non-parametric analysis (e.g. chi-square, t-test, 2-way anova)
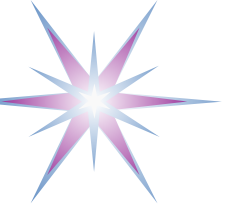
Statistical Methods

- Something to do with "statistics"
- Statistics: "meaningful" quantities about a *sample* of objects, things, persons, events, phenomena, etc.
- Widely used in social sciences.
- Simple to complex issues. E.g.
  - * correlation
  - * anova
  - * manova
  - * regression
  - * econometric modelling
- Two main categories:
  - * Descriptive statistics
  - * Inferential statistics

Exploratory Data Analysis (EDA)

A preliminary exploration of the data to better understand its characteristics.
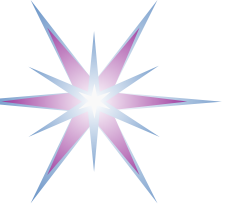
- Key motivations of data exploration include
    - Helping to select the right tool for preprocessing or analysis
    - Making use of humans' abilities to recognize patterns
        - People can recognize patterns not captured by data analysis tools

- Related to the area of Exploratory Data Analysis (EDA)
    - Created by statistician John Tukey
    - Seminal book is Exploratory Data Analysis by Tukey
    - A nice online introduction can be found in Chapter 1 of the NIST Engineering Statistics Handbook

    http://www.itl.nist.gov/div898/handbook/index.htm

Techniques Used In Data Exploration

- In EDA, as originally defined by Tukey
  - The focus was on visualization
  - Clustering and anomaly detection were viewed as exploratory techniques
  - In data mining, clustering and anomaly detection are major areas of interest, and not thought of as just exploratory

- In our discussion of data exploration, we focus on
  - Summary statistics
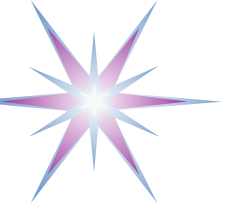 - Visualization

Analyzing the Data

~ Exploratory Methods ~

This method often involves a lot of calculating averages and percentages, and displaying the information on a graph. Although Exploratory methods may provide many pieces of information, it may not answer specific questions or make definite statements about a problem.

~ Confirmatory Methods ~

This method is used to conclude the results of the survey and the statistical information by answering specific questions. For example, using a confirmatory method, a statistician can say "Oil Prices leaving Saudi Arabia has been increasing, and will increase in prices."

Not one of these methods should be overlooked. Both methods should be used extensively to analyze the results of a statistical activity and will have to come to varieties of extremely specific conclusions with credibility and accuracy.

Reporting the Results

Inference is used to draw conclusion from a statistical activity; even from a small collection of observations or experimental results, careful and rational inference can create an accurate and reliable generalization that can be used to used to the social benefits.

There are many forms of presentations, and they include bar graphs, pie graphs, tables, or a set of percentages.

However, when drawing conclusions, one must take into consideration the fact that the survey was carried on a specifically selected sample population, not the entire population. Therefore, using *probability*, the conclusions must reflect and include the uncertainty possibly excluded or misrepresented in the statistics.

Iris Sample Data Set

- Many of the exploratory data techniques are illustrated with the Iris Plant data set.
  - Can be obtained from the UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html
  - From the statistician Douglas Fisher
  - Three flower types (classes):
    - Setosa
    - Virginica
    - Versicolour
  - Four (non-class) attributes
    - Sepal width and length
    - Petal width and length

Virginica. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute.

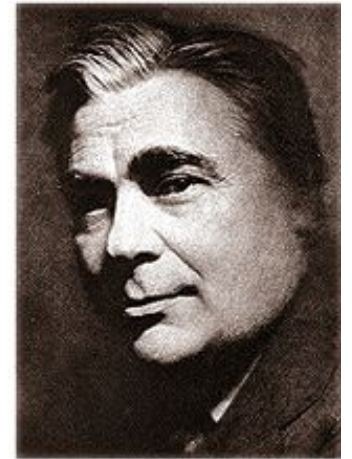# Guess the species for A and B
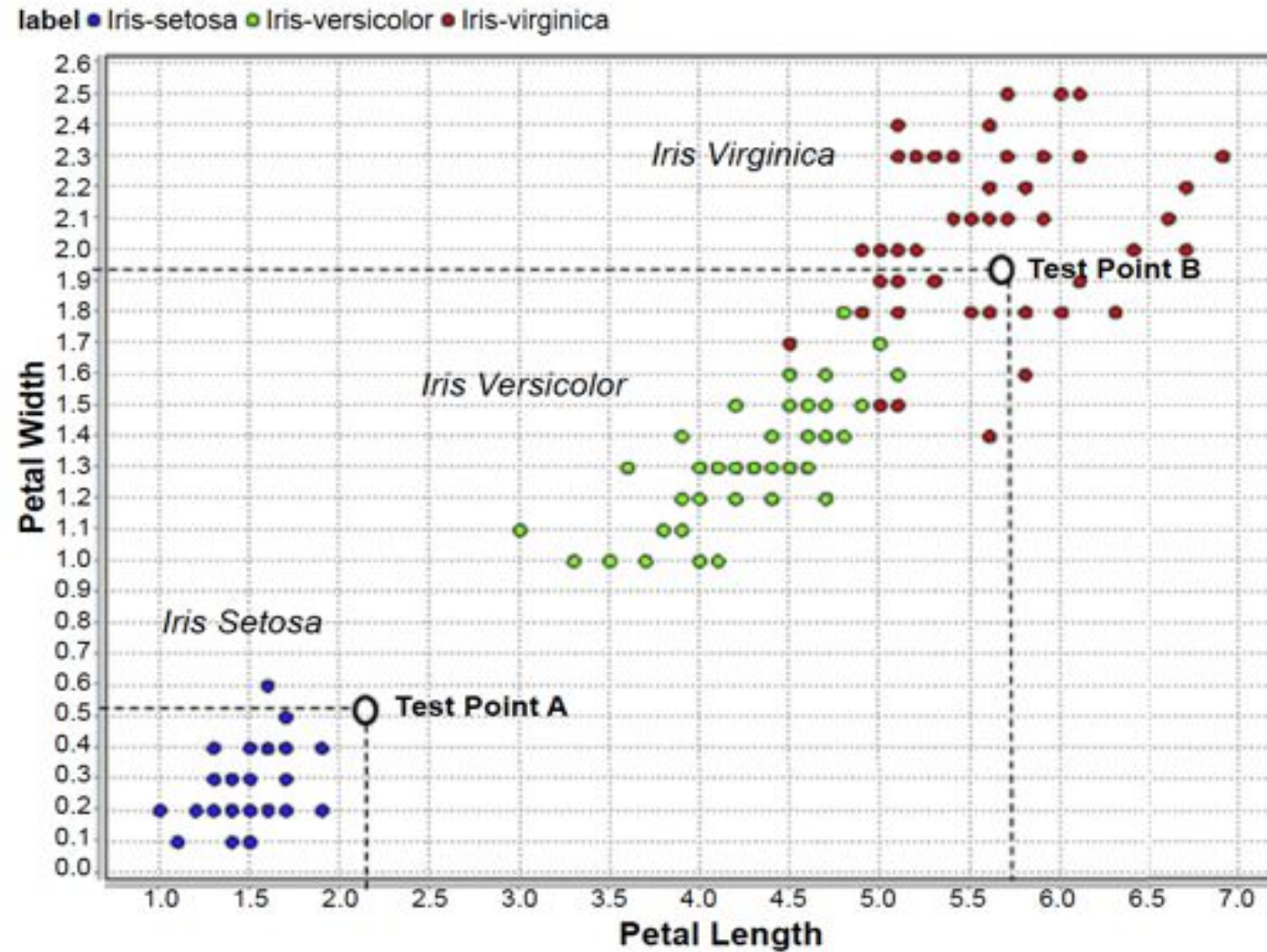
Case study: Procrastination

- Procrastination in Online Exams: What Data Analytics Can Tell Us?
  - By Yair Levy, Ph.D., Graduate School of Computer and Information Sciences

"Procrastination is the art of keeping up with yesterday."
~Don Marquis 1878-1937 (journalist/author – NYC)

- Procrastination – Voluntary postponing an activity to the last possible minute (Gafni & Geri , 2010)
- Ancient societies viewed procrastination in positive terms:
  - Avoid unnecessary work
  - Reduce impulsive behaviors
- E-learning Systems produce massive data sets

Data Analytics vs. Statistical Analysis

Data Analytics

Accumulation of raw data captured from various sources (i.e. discussion boards, emails, exam logs, chat logs in e-learning systems) can be used to identify fruitful patterns and relationships (Bose, 2009)

- Exploratory visualization – uses exploratory data analytics by capturing relationships that are perhaps unknown or at least less formally formulated
- Confirmatory visualization - theory-driven

Data Analytics

- Utilizes data mining techniques
- Identifies inexplicable or novel relationships/trends
- Seeks to visualize the data to allow the observation of relationships/trends

Statistical Analysis

- Utilizes statistical and/or mathematical techniques
- Used based on theoretical foundation
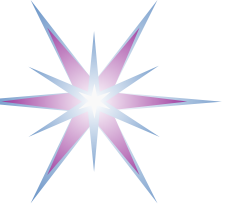- Seeks to identify a significant level to address hypotheses or RQs

Research Goals

- To uncover trends using data analytics about procrastination in online exams
- To examine a data set related to procrastination in online exams for trends in terms of:
  - Task completion time
  - Task completion scores
  - Gender
  - Academic level

  as time progress during the submission window

- To understand how to improve online exams performance, time-learning strategies, and overall learning experience.

Methodology

- The unit of analysis for this study is the task completed (i.e. an online exam)
- A data set of 1,629 online exam records
- Compiled from 10 courses distributed over five terms
- ~35 students/course
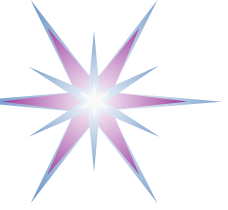- Six online exams/term

Data Extracted

- Two main time related measures were extracted:
  - task completion window (Monday 12am to Sunday 12pm)
  - task completion time
- Procrastination was measured based on the proximity to due time (in hr:min:sec)
- The task completion time is the time that it took to complete the online exam (in min:sec)
- Used SPSS 19, Excel 2011, and Google Visualization (i.e., Motion Chart Gadget)

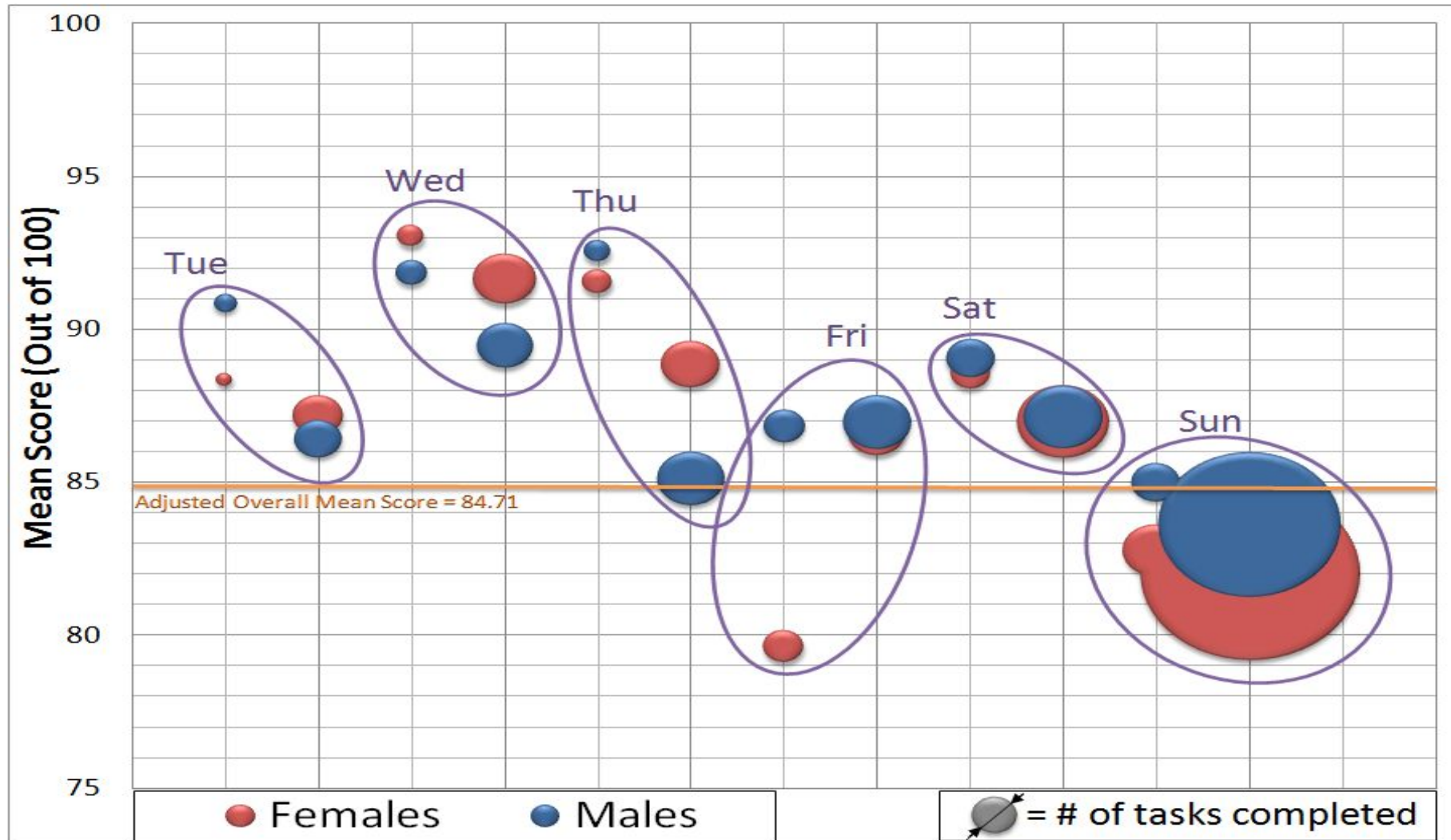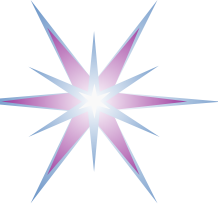# Procrastination (Day) based on Morningness-Eveningness
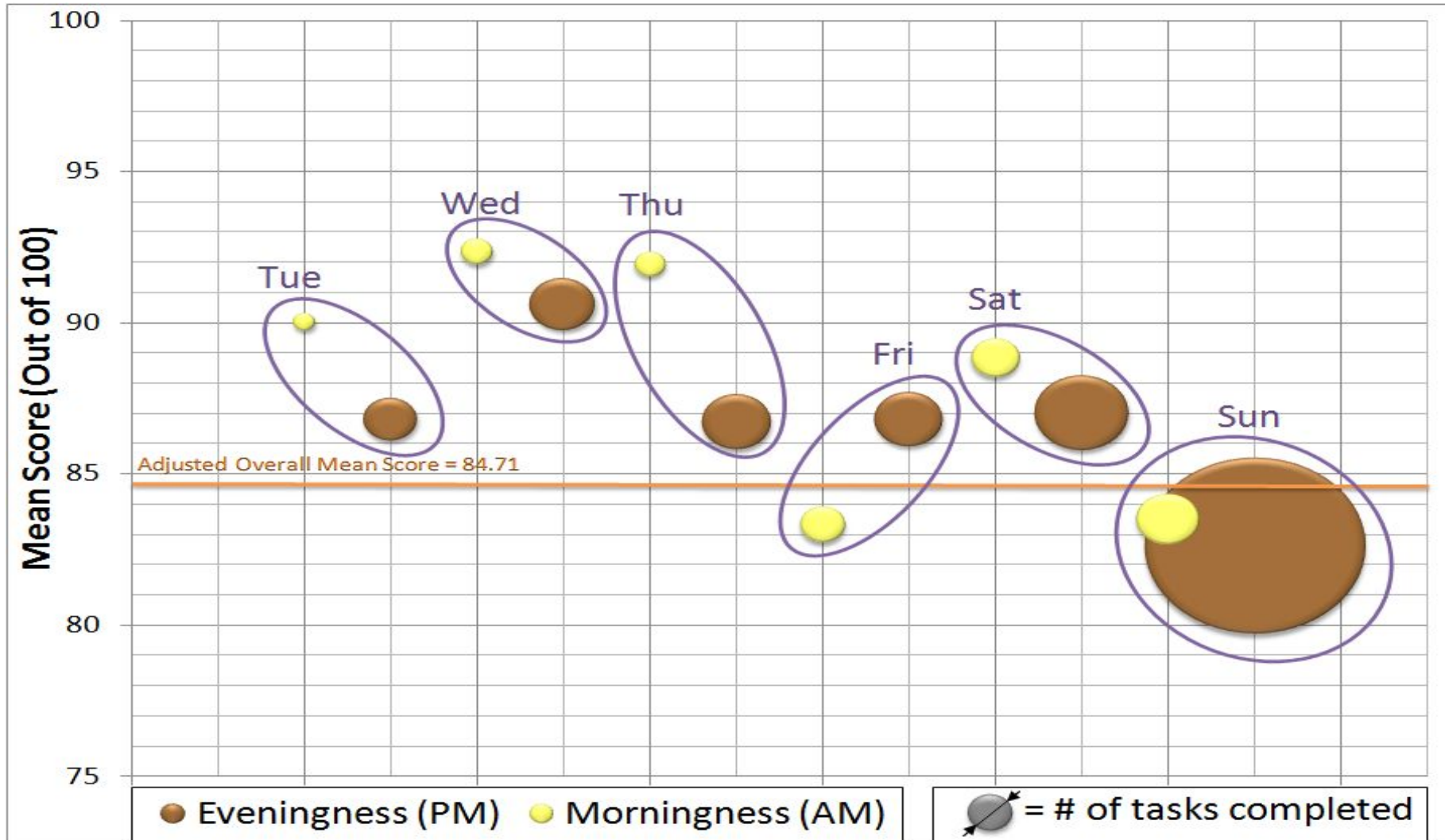
Procrastination (Day) based on Morningness-Eveningness and Gender

Procrastination: Summary of Findings and Conclusion

Simple data analytics showed that:

- Over 58% procrastinated to the last day
- About 40% procrastinated to the last 12 hours
- Significantly* more younger students procrastinate
- Percentage-wise, more females procrastinated
  - Enormous demand is placed on females (i.e. working mothers, balance work during the weekdays and family obligations during the weekends)

- Defining variables and basic coding are basic steps in data analysis
- Simple univariate analysis may be used with continuous and categorical variables
- Further analysis may require statistical tests such as chi-squares and other more extensive data analysis

Common mistakes in data analysis

- Wrong techniques
- Infeasible techniques

    How to design <u>ex-ante</u> effects of KLIA? Development occurs "before" and "after"! What is the control treatment?

- Abuse of statistics
- Simply exclude a technique

Issue	Data analysis techniques	
	Wrong technique	Correct technique
To study factors that "influence" visitors to come to a recreation site	Likert scaling based on interviews	Data tabulation based on open-ended questionnaire survey
"Effects" of KLIA on the development of Sepang	Likert scaling based on interviews	Descriptive analysis based on ex-ante post-ante experimental investigation

Note: No way can Likert scaling show "cause-and-effect" phenomena!

Common mistakes (contd.) – "Abuse of statistics"

Issue	Data analysis techniques	
	Example of abuse	Correct technique
Measure the "influence" of a variable on another	Using partial *correlation* (e.g. Spearman coeff.)	Using a regression parameter
Finding the "relationship" between one variable with another	Multi-dimensional scaling, Likert scaling	Simple regression coefficient
To evaluate whether a model fits data better than the other	Using R^2	Many – a.o.t. Box-Cox χ^2 test for model equivalence
To evaluate accuracy of "prediction"	Using R^2 and/or F-value of a model	Hold-out sample's MAPE
"Compare" whether a group is different from another	Multi-dimensional scaling, Likert scaling	Many – a.o.t. two-way anova, χ^2, Z test
To determine whether a group of factors "significantly influence" the observed phenomenon	Multi-dimensional scaling, Likert scaling	Many – a.o.t. manova, regression

How to avoid mistakes - Useful tips

- Crystalize the research problem → operability of it!
- Read literature on data analysis techniques.
- Evaluate various techniques that can do similar things w.r.t. to research problem
- Know what a technique does and what it doesn't
- Consult people, esp. supervisor
- Pilot-run the data and evaluate results
- Don't do research? Doing mistakes is big part of research

Summary and Takeaway

- Exploratory Data Analysis is important stage in each project and allows to understand the dataset and observed event or process

- Good understanding Data Analysis principles is a basis for efficient data projects

- Knowing principles and best practices helps avoiding mistakes in data analysis and abuse of statistics
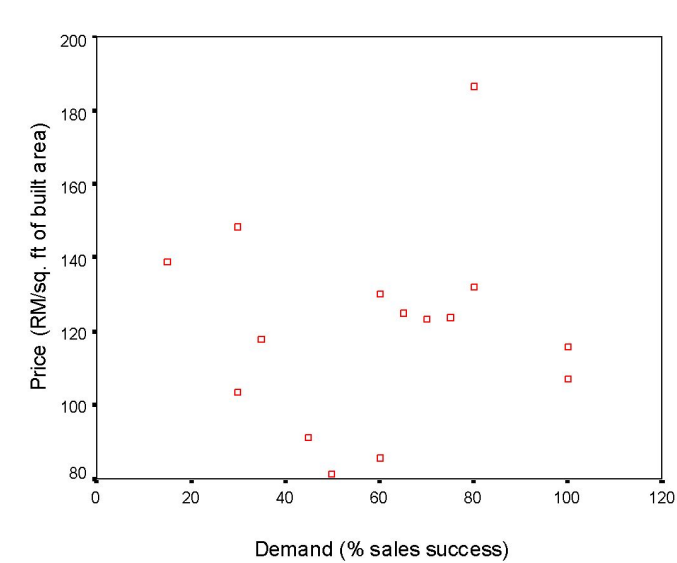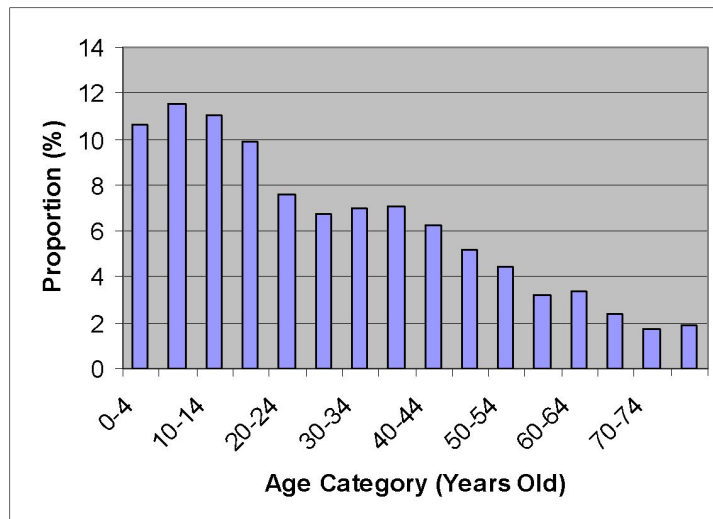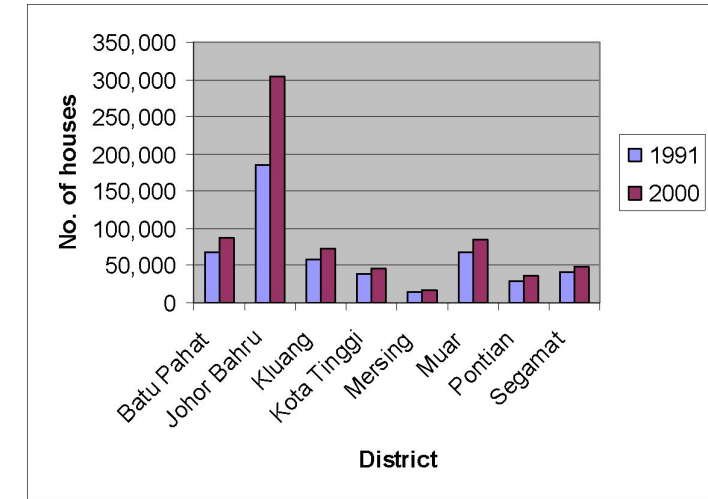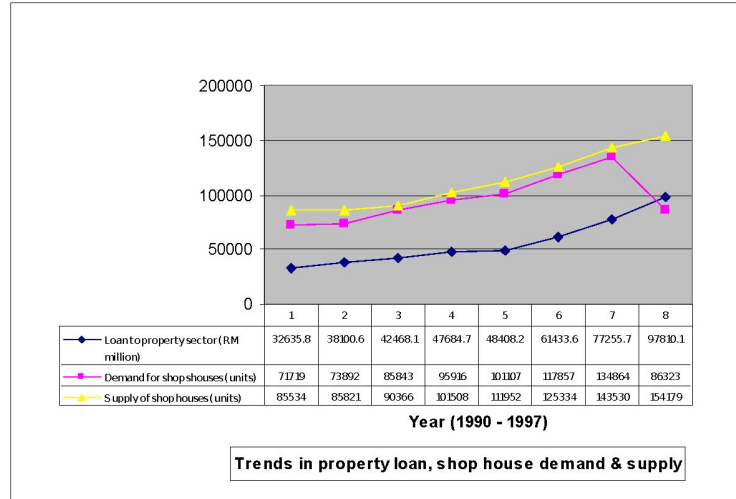
Practice Part – Exploratory Data Analysis

- Use self-study exercises provided for this course both in Python and RapidMiner
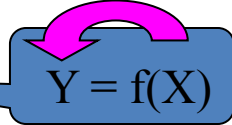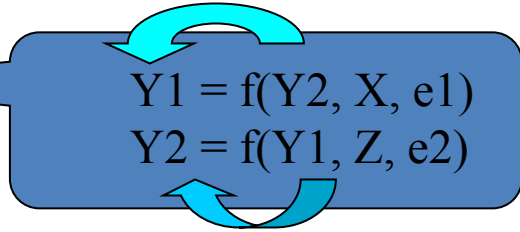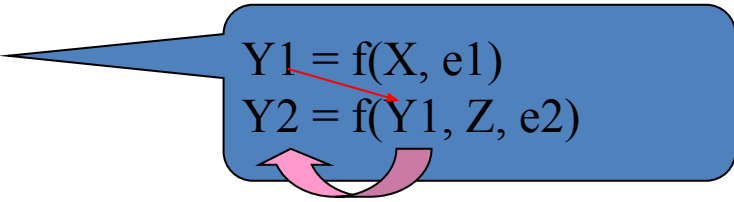- Investigate and visualize dataset characteristics

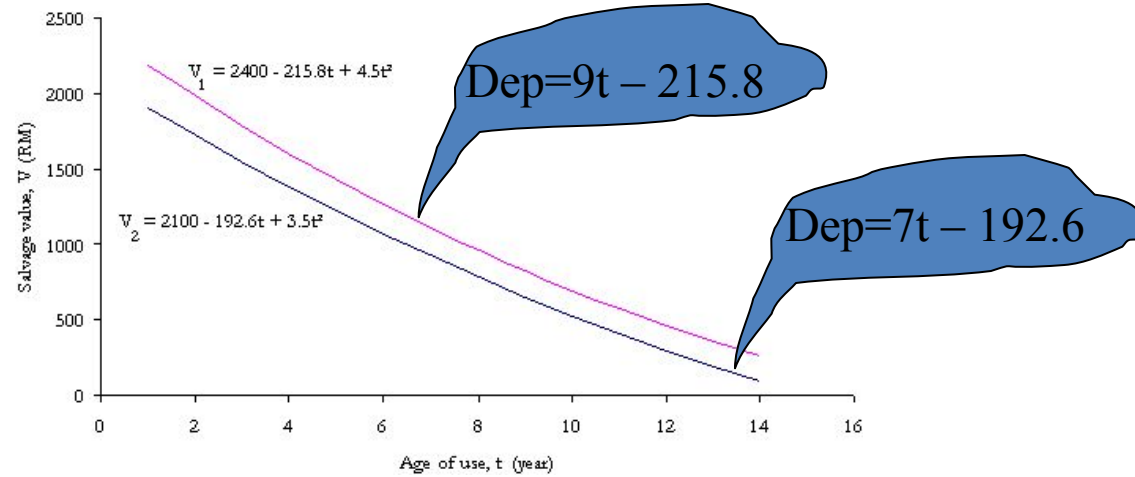Trends in property loan, shop house demand & supply

Inferential statistics

- Using sample statistics to infer some "phenomena" of population parameters
- Common "phenomena": cause-and-effect

- One-way relationship $Y = f(X)$

- Multi-directional relationship

$$Y1 = f(Y2, X, e1)$$
$$Y2 = f(Y1, Z, e2)$$

- Recursive

$$Y1 = f(X, e1)$$
$$Y2 = f(Y1, Z, e2)$$

- Use parametric analysis

Examples of relationship

$V_1 = 2400 - 215.8t + 4.5t^2$

$V_2 = 2100 - 192.6t + 3.5t^2$

Dep=9t – 215.8

Dep=7t – 192.6

Dependent Variable: Nilaism

Model		B	Std. Error	Beta	t	Sig.
			Unstandardized Coefficients	Standardized Coefficients		
a.						
	Flo_go	34.895	89.440	.020	.390	.697
	Umur	-27.893	6.108	-.241	-4.567	.000
	Ansilari	4.393	1.807	.139	2.431	.017
	Bangunan	6.938	.619	.705	11.209	.000
	Tanah	-4.472	1.199	-.190	-3.728	.000
1	(Constant)	1993.108	239.632		8.317	.000

Coefficients a