Technology Management for Organisations

Workshop 4

Data Mining & Data Dictionary

Aims of the workshop

In Lectures we have looked at Data Mining for Organisations, and introduced the concept of Data Warehousing as well as the importance of Data Dictionaries. This workshop will involve some coding tasks, as well as paper-based tasks to get to grips with these concepts as well as some independent reading.

The concept behind this workshop is about discovery, and experimentation surrounding topics covered so far.

Make sure to fill in your ePortfolio for Teaching Week 4 / Workshop 4 after the workshop.

For the paper-based exercises, feel free to scan them and e-mail them to yourselves or use a mobile phone to take a photo of any evidence you wish to include.

As a reminder, you should be writing reflectively, what worked well? What challenges did you face? How did you overcome them? It's about documenting process like a diary. **Do not submit just a single complete piece of code; it's about how you got to that, the steps, the struggles, etc.**

Feel free to discuss the work with peers, or with any member of the teaching staff.

Reminder

We encourage you to discuss the contents of the workshop with the delivery team, and any findings you gather from the session.

Workshops are not isolated, if you have questions from previous weeks, or lecture content, please come and talk to us.

Exercises herein represent an example of what to do; feel free to expand upon this.

Exercises

With the exercises for this workshop, please discuss answers with a peer as well as a member of the delivery team.
You should divide your time accordingly over each of the tasks (at least 15-20 minutes each) in today's workshop, do not rush these. Take your time.

Use your ePortfolio here to document your responses, and how you got to them.

Exercise 1: Research individually, or as part of a group, the trends in data of a company in your field or a company of interest for you.
In particular ensure you consider and answer the following:

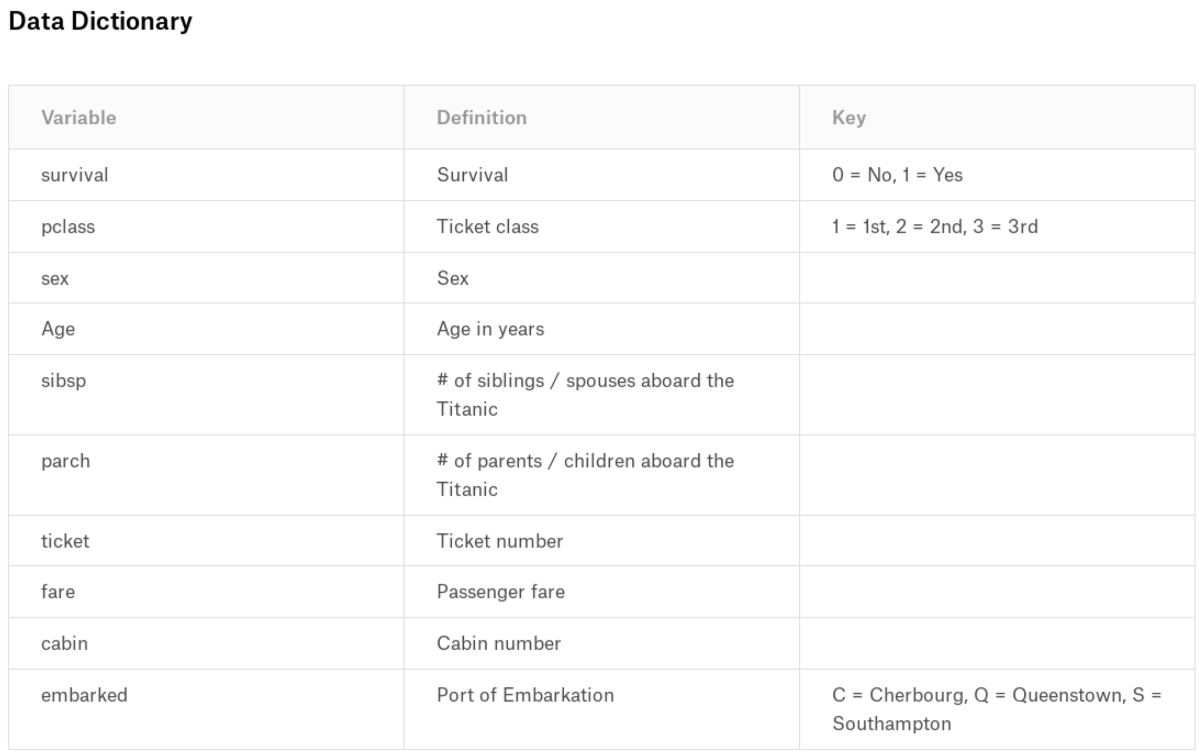
1. How much data is generated by this company (and the time frame)?
2. What data is used for day-to-day operations, is any generated data used for analysis?
3. How is this data used?
4. Where do they get the data from?
5. How do they get this data?
6. Who gave them permission for these data?

Exercise 2: Find an example of a company utilising data mining for each of the use-cases outlined in the lectures. How do they extract value from the data within these use-cases to drive their business? What technologies do they use for these, where applicable?
When answering these questions for the following use-cases, you should aim to find any relevant literature such as journals, conference proceedings, or technical whitepapers which may have been produced by, or about, the chosen company.

1. Log analytics
2. Commerce
3. Recommendation
4. Fault Detection / Prediction
5. Fraud Detection

Can you think of any other use-cases not listed above, for your respective programme of study?

Exercise 3: Read in the data from the titanic dataset CSV file found on Canvas using techniques covered in previous workshops (I.e Context managers, File I/O).
Using Python, do some basic printing of the data, looking for attributes and the raw data stored in the CSV.
For each attribute, using “pen-and-paper” (not programming) and knowledge from the lectures, provide the correct data type. E.g Is it boolean, numeric, categoric (ordinal/nominal).
You may need to consult the Data Dictionary below for the titanic dataset.



Data Dictionary

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Exercise 4: Using the Titanic Data and Python (Numpy may be helpful here):

1. Calculate the min and max fare in the dataset.
2. What Data Type is the **fare** attribute? (Both Python data type, and Data Science)
3. If we were writing a better Data Dictionary, how would we represent this data attribute as a range descriptor? Remember the difference between [10, 1000] and (10, 1000).
4. Given the values within the dataset and what fare represents, what would be a good degree of precision to use for this attribute? Defend your position.
5. Separate out the data into the three pclass (1, 2, and 3). How does the min and max fare price change for each of these groups?
6. Consider the **name** attribute, what is it composed of, and how you might express this better?

   Example: How do we break up names in our society? What about titles (Mr, Mrs, Ms, Dr, Prof, …).

Exercise 5: Look at publicly available datasets for data mining / data analysis which have data dictionaries. Find 2 different data dictionaries and compare these against each other and the titanic dataset above, and the games example from the lecture slides.
You should note the level of description, the accuracy, as well as useful information for multiple roles within the organisation: analyst, designers, developers.

Exercise 6: In the lecture we looked at the titanic dataset data dictionary and compared it against a more well-defined version for a games-selling business. Given access to the titanic dataset, improve the data dictionary provided with Exercise 3 to be more complete and useful. For each modification / addition, you should justify your reasoning.

END OF EXERCISES