Using data science to produce the V&A environmental report

Bhavesh Shah, Scientist (Environment), Victoria and Albert museum, b.shah@vam.ac.uk

Introduction

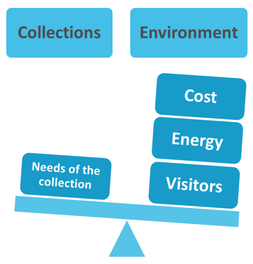
The V&As environmental monitoring system, OCEAN (Object Centred Environmental Analysis Network), continually gathers climate data from over 650 sensors measuring primarily temperature, humidity and light. Data is recorded every 15 minutes, generating over 40 million time-stamped data points a year. Once the data is collected decisions are made on how the climate could be a risk to the collections and what actions can be undertaken by the Estate team to manage the conditions. The challenge is finding a way to process and concisely present the data to key stakeholders.

Environmental report

The V&A aims to summarise the climate data in a monthly report that is quick and easy to interpret. The key objectives of summarising the data are:

- To identify the areas that are out of specification and any simple risks this may create to the collections, such as exceptionally high/low conditions.
- To report these areas to key stakeholders: curators, conservators, collections managers and engineers.
- To prioritise and record any actions undertaken for auditing.
- To consider the needs of the visitor and to reduce the museum's carbon footprint.

"The V&A environmental report summarises climate and environmental issues relevant to the collections and actions are taken to address specific concerns, whilst at the same time recognising the museum's sustainability ambitions, without adversely affecting visitor experience" – V&A Environmental report.



Data Science

High-level programming and statistical software are freely available under the growing umbrella of 'data science' to gather, process, analyse and communicate data in one complete package.
The advantages of using data science software to produce the environmental report are:

- The resources to learn data science are more available than before.
- A vast array of toolkits are available for graphing, summarising, modelling and machine learning data.
- The infrastructure for online collaboration, sharing code and development is established.
- There exists a large community for support both online and in-person.
- Outputs can be static such as PDF, Word or PowerPoint, as well as interactive such as websites.
- Once a report has been coded, the same code can be rerun every month using new data.
- This poster shows a data science process in R, but similar functionality is available in other languages such as Python, Julia, etc.

Data science software: python, R, julia

Help and learning: stackoverflow, Meetup, MANGO SOLUTIONS, R Ladies, edX

Collaboration and development: devtools, roxygen2, GitHub

Data Process in R

Different sources of data can be imported

- Text (CSV, txt)
- Excel
- SPSS, Stata, SAS
- Database
- json
- XML
- Web APIs
- HTML

'Messy' data

Space->	Space 1		Space 2		Space 3	
<-Time	T°C	RH	T°C	RH	T°C	RH

```
read_csv("data.csv")
read_excel("data.xls", sheet = 1)
```

Data is often imported in 'messy' formats. For example observations and variables are stored side-by-side for compactness.

Tools to tidy and transform data are available including tools to handle strings, dates and time.

```
%>% pipe operator is used to chain code so that it can be read from top to bottom.
```

```
tidy_data <-  messy_data %>%
  gather, spread, separate, unite, complete, arrange
```

'Tidy' data

<- Space <-Time	T°C	RH	New variable
Space 1			
Space 2			
Space 3			

'Tidy' data has each variable in each column, each observation in every row and so every cell holds a single value.

Once the data is 'tidy' it is easier to understand, to calculate new variables and combine with other tidy data sets. It is also easier to group, filter and subset tidy data.

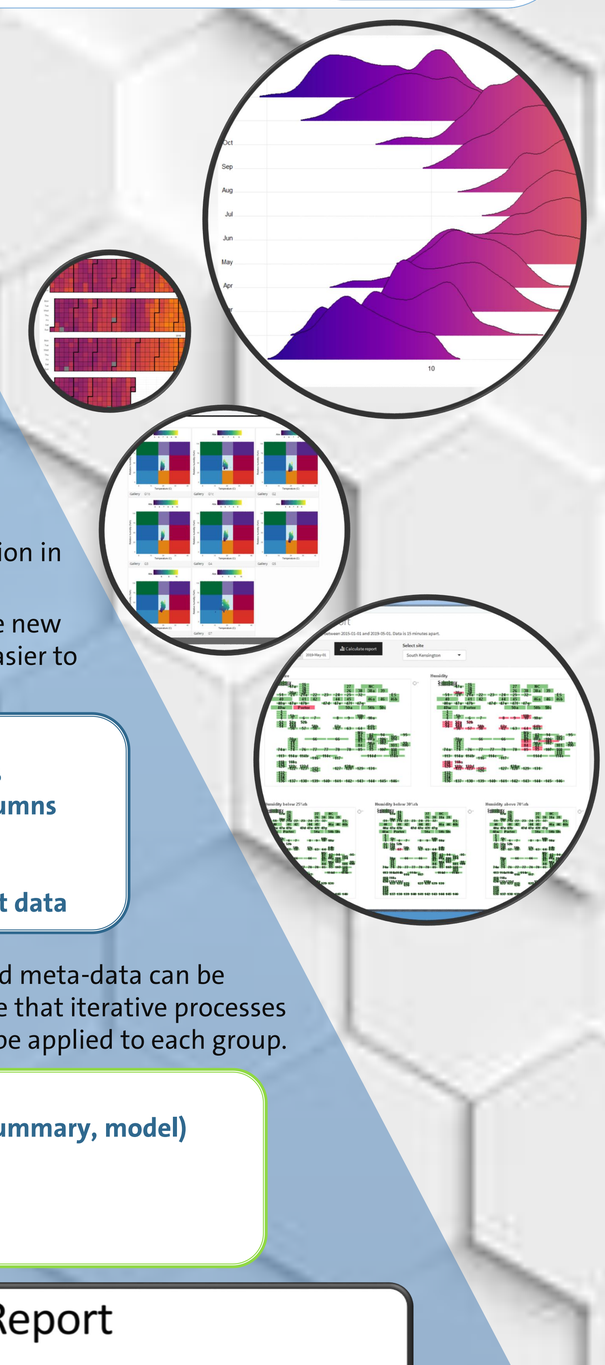
```
tidy_data %>%
  filter, distinct, sample_n, slice, top_n %>% # subset rows
  select, starts_with, contains, end_with %>% # select columns
  join, union, bind # combine, filter, join with other data
  mutate, transmutate, %>% # make new variables
  group_by, summarise, nest # group, summarise and nest data
```

'Nested' data

	Data	Meta data	Outputs
1			
2			
3			

Data can be 'nested' so that each row is a group of data and meta-data can be added to define the groups. Nested data has the advantage that iterative processes (loops) such as summarising, modelling and graphing can be applied to each group.

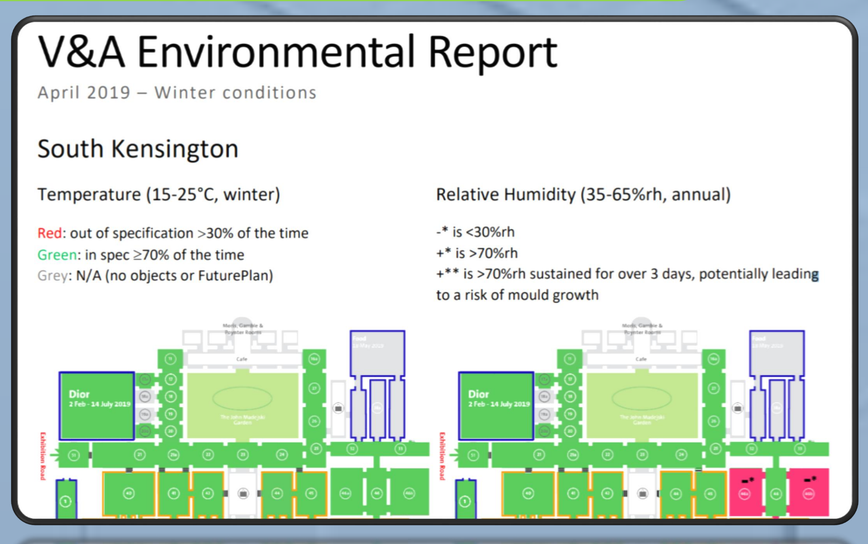
```
  %>% map(.data, function) %>%
# map can be used to create the desired outputs (graph, summary, model) whilst keeping a pipe flowing

  compose(import, tidy, nest, graph, summarise)
#  compose can put all the code together in one line!
```



Report

Outputs include:

- Interactive notebooks
- Website applications
- HTML
- Word/PDF
- Blogs
- Presentations
- Dashboards



Conclusion

Analysing and presenting environmental data to inform practical solutions is a challenge for any operational preventive conservator or scientist. Although the learning curve for coding can be steep, the benefits of investing time to automate recurring processes and creating insights into data that have previously been out of reach of non-coders is evident. As an increased volume of data is collected and as risk models grow in complexity, skills in data science will become more important in the future; in addition to the desire for organisations to make decisions driven by data.