Intro to Data Science

and SQL

Alicja Laszuk

okta

[uh-LEE-ts-ya]

Hi, I'm Alicja!

Work:

- Senior Site Reliability Engineer @Okta, with a focus on Datastores
- a few jobs ago, Quantitative Economist @Ministry of Health, Poland

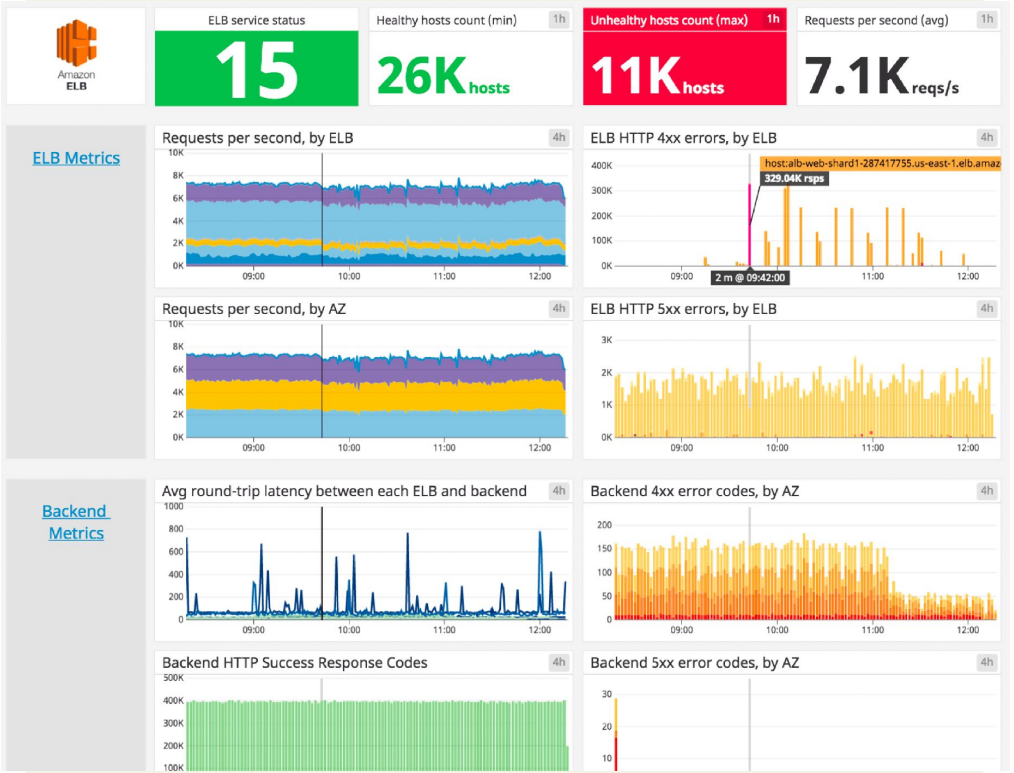
While I no longer work as a data scientist, I still use data in my job all the time:

- monitoring resources
- investigating issues
- digging into performance
- understanding how our customers are using our product

That's why I wanted to talk about data science.

I also knit and crochet, so the data we're going to use is on yarn and patterns :)

Note: none of the data or graphs in this workshop contain internal Okta information



okta

Workshop overview

go to the workshop GitHub repository

1. Brief intro to what data science is about (3mins)
2. Dataset overview (2mins)
3. Basic SQL (15mins):
   a. SELECT statements
   b. aggregate functions
4. Exploratory data analysis (15mins)
   a. distribution metrics
   b. visualizations

Feel free to ask questions as we go through the material! I'll make plenty of pauses.

Also, no need to take notes, everything (including these slides) is on GitHub.

okta

Data Science is creating insights from data

Understanding of the domain

- general terms in the problem space
- business needs
- user behaviours and pathways
- usual kinds of problems
- gaps

Example:

In order to create an item made of yarn, one usually requires a pattern, appropriate yarn, and tools that allow them to meet the gauge.

Programming

- finding the data
- getting the data
- organizing the data
- cleaning up the data
- writing code

Example:

Main sources of information for yarn crafts are yarn and pattern databases. They include details such us yarn size and fibre, as well pattern gauge and yardage.

Statistics

- exploring the data
- understanding distributions and patterns
- building models
- providing actionable insights

Example:

In order to build a recommendation system for patterns, we need to analyse what the user is making, their search queries, and the type of yarn they usually work with.

About the dataset

Name - string

Parent category - string, Hat|Hands|Sweater

Parent category id - integer

Craft name - string, Knitting|Crochet

Price - float

Currency - string, three-letter currency code

Yarn weight - string, standard names defined by Craft Yarn Council

Yardage - float, number of yards required

Gauge - string|float, stitches per 4x4inch square

Dataset size: 3000

Format: .csv

Source: ravelry.com

Let’s explore some data!

```
Name,Parent_category,Craft,Yarn_weight,Yardage,Gauge,Price,Currency,Parent_category_id
Hot Dish Hat,Hat,Knitting,Aran,300,18,4,USD,411
Musselburgh,Hat,Knitting,Any gauge,130,6,6,GBP,411
Explicate,Hat,Knitting,Any gauge,200,,10,USD,411
The Caliper Beanie,Hat,Knitting,Bulky,110,10,6.25,CAD,411
Totally Textured Beanie,Hat,Crochet,Aran,200,14,,,411
The Basketweaver Sweater,Sweater,Knitting,DK,900,20,6,EUR,319
smoking,Sweater,Knitting,DK,1170,21,6.7,EUR,319
Carlisle (Saddle),Sweater,Knitting,Fingering,1340,25,7,USD,319
Carlisle (Raglan),Sweater,Knitting,Fingering,1340,25,7,USD,319
Bohemian Scrapsody,Sweater,Knitting,Aran,1121,14,9,CAD,319
Hot Dish Mitts,Hands,Knitting,Aran,260,20,4,USD,390
3-Hour Mitts,Hands,Knitting,Bulky,50,10,,,390
The World's Simplest Mittens,Hands,Knitting,Any gauge,70,20,,,390
Lolina_DCmittens,Hands,Knitting,Sport,248,33,5,EUR,390
Goldie Mittens,Hands,Knitting,Fingering,240,28,10.5,USD,390
```

okta