Databases for data analytics

https://github.com/evidencebp/databases-course/

[MySql Functions](https://github.com/evidencebp/databases-course/)

MySql Functions

- MySql has over 400 functions, listed in its documentation
- Their vast majority is not relevant to the course and most uses.
- Function are mostly the same in databases, and even in programming languages, usually differing in names only
- No need to know the functions by heart, knowing the functionality is usually enough for successful search
- One can write new functions and extend the abilities

Useful functions families

- Operators
- Flow control
- Data type
  - numeric,
  - string
  - datetime (not for our course)
- Aggregate functions
- Window functions

Operators

- [CAST](#) - type conversion
- Is null - null is unknown, comparison is misleading. Is null identifies null
- Exist/in - existence in a subqueries. Handy but not elegant and very slow. Use joins instead.
- Like - weak pattern recognition on strings
- Regexp - strong yet more complex pattern recognition on strings
- All the usuals too…

Flow control

- If - conditional
- Case - multi value conditional
- Ifnull - replace null value

Numeric

- Rand - random function, useful for sampling
- Round - reduces precision, useful in grouping
- All the usual suspects (sin, cos, etc, handy if you work in a very unique domain)

String

- Concat - turns some strings into one
- Instr - search in string
- Length - string size
- Lower/upper - case enforcement
- Substr/left/right - get a part of a string

Aggregate

- Count - number of rows
- Count distinct <EXP> - number of different value of the expression (which is usually a column)
- Sum <EXP> - sum of <EXP> values
- Avg - average of <EXP> values
- Std - standard deviation of <EXP> values
- min/max - extreme values of <EXP> values
- Group_concat <EXP> - a list of the <EXP> values, can be used with distinct too

Window

- Not in the current scope
- Close in spirit to aggregation functions, has a window per record on which computations can be done
- Row_number - index selected rows
- lag/lead - rows before and after the current (handy when ordered)