**Dataset analysis**

**Relational vs. non-relational databases: Which one is right for you?**

Relational databases like MySQL, PostgreSQL and SQLite3 represent and store data in tables and rows. They're based on a branch of algebraic set theory known as relational algebra. Meanwhile, non-relational databases like MongoDB represent data in collections of JSON documents. The Mongo import utility can import JSON, CSV and TSV file formats. Mongo query targets of data are technically represented as BSON (binary JASON).

Relational databases use Structured Querying Language (SQL), making them a good choice for applications that involve the management of several transactions. The structure of a relational database allows you to link information from different tables through the use of foreign keys (or indexes), which are used to uniquely identify any atomic piece of data within that table.

In non-relational databases like Mongo, there are no joins like there would be in relational databases. This means you need to perform multiple queries and join the data manually within your code -- and that can get very ugly, very fast.

Since Mongo doesn’t automatically treat operations as transactions the way a relational database does, you must manually choose to create a transaction and then manually verify it, manually commit it or roll it back. Even the documentation on the MongoDB site warns

SQL Syntx:

Select \* from “table” (\* can be replaced my multiple columns in the table)

Where date>’01-01-2017’ and price > 20

**Statistics**

You should be able to explain phrases like null hypothesis, P-value, maximum likelihood estimators and confidence intervals.

**Confounding Variables:**

A confounding variable is an “extra” variable that you didn’t account for. They can ruin an experiment and give you useless results. Let’s say you test 200 volunteers (100 men and 100 women). You find that lack of exercise leads to weight gain. One problem with your experiment is that is lacks any [control variables](http://www.statisticshowto.com/control-variable/). For example, the use of placebos, or random assignment to groups. So you really can’t say for sure whether lack of exercise leads to weight gain. One confounding variable is how much people eat.

**P-Value:**

When you perform a hypothesis test in statistics, a p-value helps you determine the significance of your results. [Hypothesis tests](https://www.dummies.com/education/math/statistics/how-hypothesis-tests-are-used-in-statistics/)are used to test the validity of a claim that is made about a population. This claim that’s on trial, in essence, is called the [null hypothesis](https://www.dummies.com/education/math/statistics/how-to-determine-a-p-value-when-testing-a-null-hypothesis/).

The alternative hypothesis is the one you would believe if the null hypothesis is concluded to be untrue. The evidence in the trial is your data and the statistics that go along with it. All hypothesis tests ultimately use a p-value to weigh the strength of the evidence (what the data are telling you about the population). The p-value is a number between 0 and 1 and interpreted in the following way:

* A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis.
* A large p-value (> 0.05) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis.
* p-values very close to the cutoff (0.05) are considered to be marginal (could go either way). Always report the p-value so your readers can draw their own conclusions.

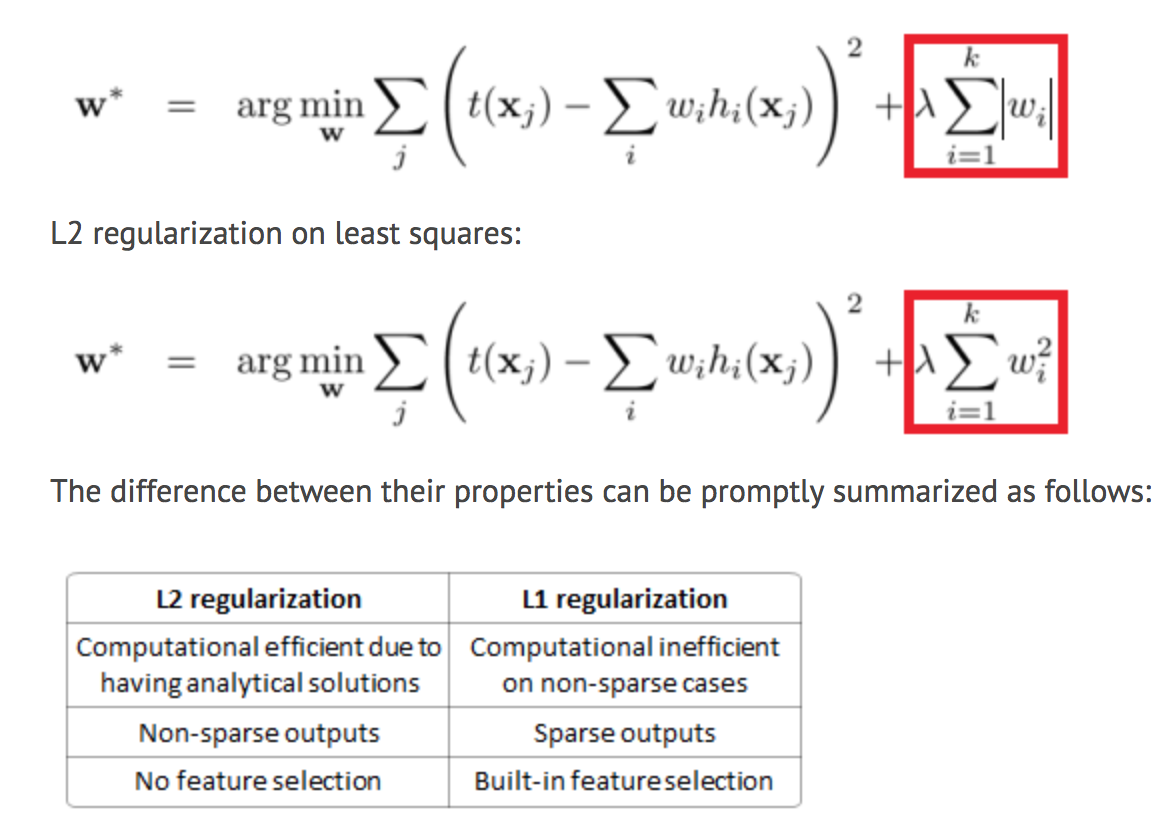
For example, suppose a pizza place claims their delivery times are 30 minutes or less on average but you think it’s more than that. You conduct a hypothesis test because you believe the null hypothesis, Ho, that the mean delivery time is 30 minutes max, is incorrect. Your alternative hypothesis (Ha) is that the mean time is greater than 30 minutes. You randomly sample some delivery times and run the data through the hypothesis test, and your p-value turns out to be 0.001, which is much less than 0.05. In real terms, there is a probability of 0.001 that you will mistakenly reject the pizza place’s claim that their delivery time is less than or equal to 30 minutes. Since typically we are willing to reject the null hypothesis when this probability is less than 0.05, you conclude that the pizza place is wrong; their delivery times are in fact more than 30 minutes on average.

**Box-Cox transformation:**

In [statistics](https://en.wikipedia.org/wiki/Statistics), a power transform is a family of functions that are applied to create a [monotonic transformation](https://en.wikipedia.org/wiki/Monotonic_function) of data using [power functions](https://en.wikipedia.org/wiki/Power_function). This is a useful [data transformation](https://en.wikipedia.org/wiki/Data_transformation_(statistics)) technique used to stabilize variance, make the data more [normal distribution](https://en.wikipedia.org/wiki/Normal_distribution)-like.

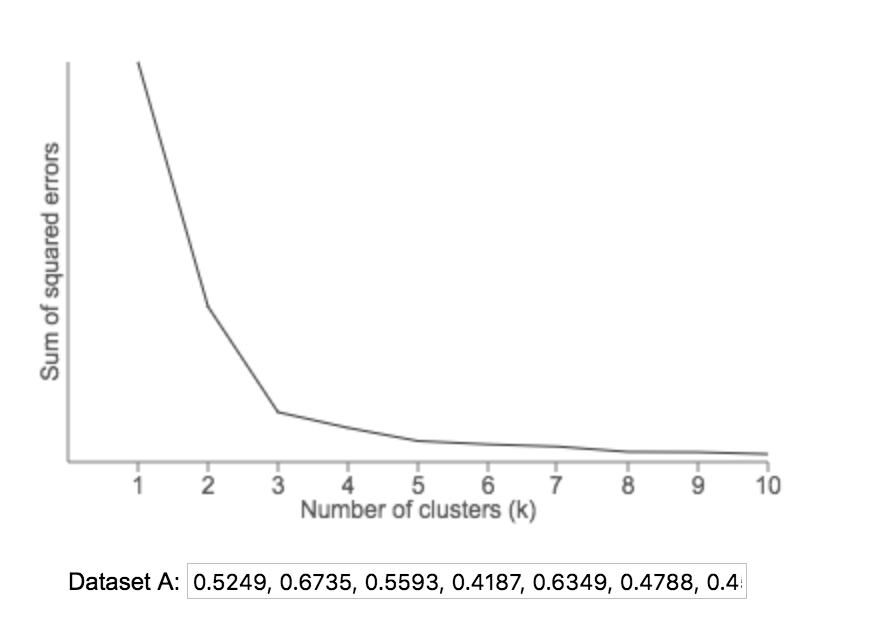
**Machine Learning:**

* What is Regularization? What is the difference in the outcome (coefficients) between the L1 and L2 norms? (usually I ask them to draw the geometric shape of the functions, just to make sure)



* How do you find the optimal k (k\*) in K-Mean? (I expect candidates to know at least 2 methods)

One method to validate the number of clusters is the ***elbow*** *method*. The idea of the elbow method is to run k-means clustering on the dataset for a range of values of *k* (say, *k* from 1 to 10 in the examples above), and for each value of *k* calculate the sum of squared errors (SSE). The "elbow" on the arm is the value of *k* that is the best. If you increase K, it will be like, each data becomes a cluster.



**What is A/B testing:**

<https://www.optimizely.com/optimization-glossary/ab-testing/>

**K-means Vs.** [**K-nearest neighbors**](http://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm)

[K-means](http://en.wikipedia.org/wiki/K_means) is a clustering algorithm that tries to partition a set of points into K sets (clusters) such that the points in each cluster tend to be near each other. It is unsupervised because the points have no external classification.

[K-nearest neighbors](http://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm) is a classification (or regression) algorithm that in order to determine the classification of a point, combines the classification of the K nearest points. It is supervised because you are trying to classify a point based on the known classification of other points.

**Review**

<https://github.com/ShuaiW/ml-interview/blob/master/README.md>

**Coding**

1. **Classes:**

<https://jeffknupp.com/blog/2014/06/18/improve-your-python-python-classes-and-object-oriented-programming/>

1. **\*args and \*\*kwargs:**

<http://book.pythontips.com/en/latest/args_and_kwargs.html>

1. **Itertools (not urgent, you can read after interview)**:

<https://realpython.com/python-itertools/>