# Ting jeg har gjort

## Dataforberedelse

* Ændret **længde- og breddegrader til dansk**, for at det virker på kortet
* *Year*, *month* og *day* er **lagt sammen i én *dato* kolonne**, og ændret til dansk format (DD-MM-ÅÅÅÅ)
* **Klokkeslæt** ændres til noget underligt. Forslag fra Amalie: Gå et "trin tilbage" i transformér data
* **Erstatte *null* med 0?** I de sidste 5 kolonner med delayes (air system delay, airline delay osv.) er der både *null*-værdier og 0-værdier. Jeg har vurderet, at det giver mening at erstatte *null* med 0 i disse kolonner, da der her er tale om antal minutter, et fly er forsinket. Om et fly er rapporteret at være rettidigt (0), eller der ikke er nogen indrapporteringer (null), vil jeg umiddelbart vurdere skulle give det samme resultat i denne sammenhæng, hvorfor jeg har valgt at konvertere *null* til 0 af hensyn til datarenhed og konsekvent måde at skrive 0/*null* på.

Til gengæld beholder jeg *null* i de øvrige kolonner, da det her kan være relevant at vide, at der er tale om enten manglende rapportering, eller et fly der er aflyst. Eksempelvis er der fly, hvor *air time* står til *null* (pga. aflysning), og vi kan derfor have interesse i at kunne filtrere disse data fra.

* **Fejlbehæftet data?** Der er fejlbehæftede rækker, fejl f.eks. manglende afgangs- eller ankomstdata. Disse skyldes dog enten aflyste fly, eller omdirigerede fly.
* Derudover er der **486.165 rækker**, hvor lufthavne figurerer med et nummer frem for initialer. Disse forekommer i oktober 2015. Jeg har valgt at filtrere disse rækker fra.
* Efter denne filtrering ender jeg med et datasæt på **5.332.914 rækker, jeg kan arbejde med**.
* **Slettede kolonner**:
  + Tail number
  + Taxi out
  + Wheels off
  + Air time
  + Wheels on
  + Taxi in
  + Cancelled reason (Da referencerne ikke findes andetsteds)
* Jeg har **pivoteret delay kolonnerne** i et nyt ark, så jeg kan arbejde med dem i et cirkeldiagram

## Kontrol

Jeg har udtrukket et udsnit af data, og eksporteret til et Excel-ark, så jeg nemt kan kontrollere, om de data jeg præsenterer, er rigtige. Jeg har brugt Lehigh Valley International Airport (ABE) som kontrol-afgangslufthavn.

# Forbehold

* Der findes ingen unikke nøgle-ID'er (både flight number og tail number går igen flere gange)

# Forståelse af funktionaliteter

**Semantisk model/stjerneskema**

Det er dén der findes under "Modelvisning"

**Tabeller**

* Fact table (f.eks. *flights*)
* Dimension table (f.eks. *airports*)

**Relationer**

* One to many
* Many to one
* Many to many

**Links**<https://youtu.be/4ePNrdxWtY0> – forklaring af hvordan Modelvisning virker