Mathias Caelenberghe, Nick Engels, Matthias Durivet, Matthias Bruynooghe

Big Data Project

Inhoudsopgave

[1 Data Cleaning 2](#_Toc513811778)

[2 Tableau 2](#_Toc513811779)

[2.1 Analyse aantal vluchten 2](#_Toc513811780)

[2.2 Analyse Vertragingen 4](#_Toc513811781)

# Data Cleaning

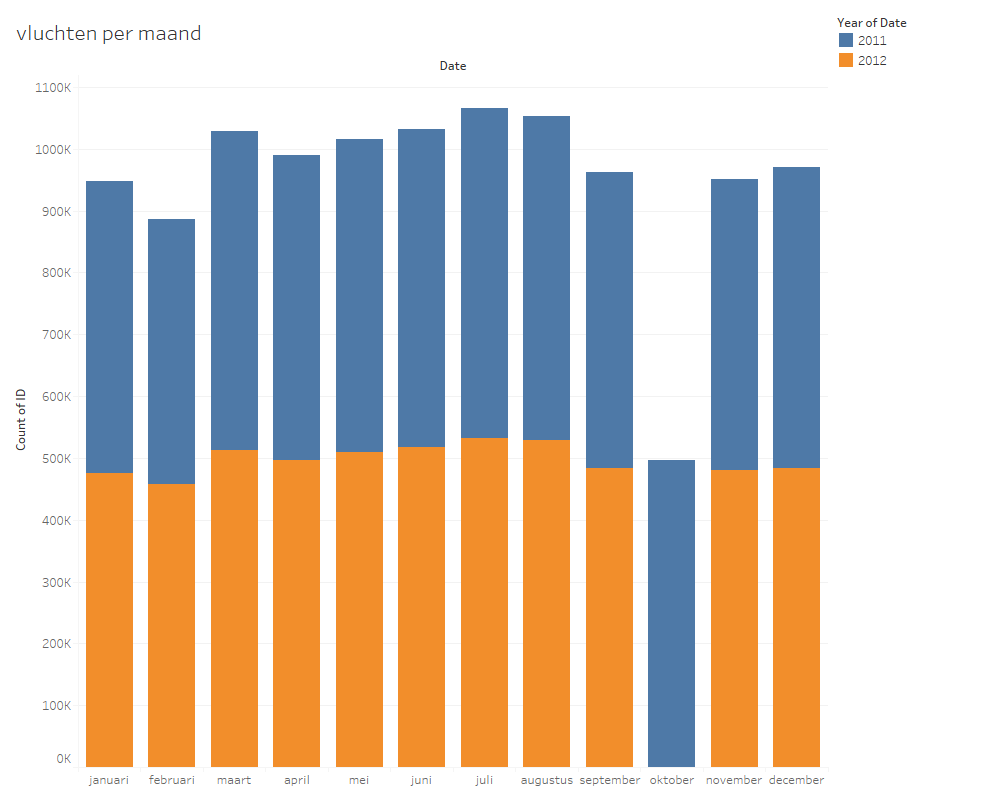
Om te beginnen bij de data cleaning hebben we alle 6 sets ingeladen en gecombineerd in 1 grote set. We hebben de dubbele entries eruit verwijderd uit de tabel.

We hebben gecontroleerd of er onvolledige kolommen waren deze bleken de kolommen ‘Scheduled\_departure’ en ‘scheduled\_arrival’ incompleet te zijn. Aangezien de data uit de kolommen ‘actual\_departure’, ‘actual\_arrival’ en de elk respectievelijke delay kolommen volledig in gevuld zijn kunnen we hieruit de scheduled kolommen aanvullen.

Vervolgens hebben we de kolommen ‘departure\_delay’ en ‘arrival\_delay’ omgezet naar integers.

# Tableau

Bij het invoeren van de data in Tableau hebben we ontdekt dat de vluchtdata van oktober 2012 volledig ontbreekt. Na overleg hebben we beslist deze data niet te genereren op basis van de data van 2011 omdat dit mogelijks aanzien zou kunnen worden als vervalsing.

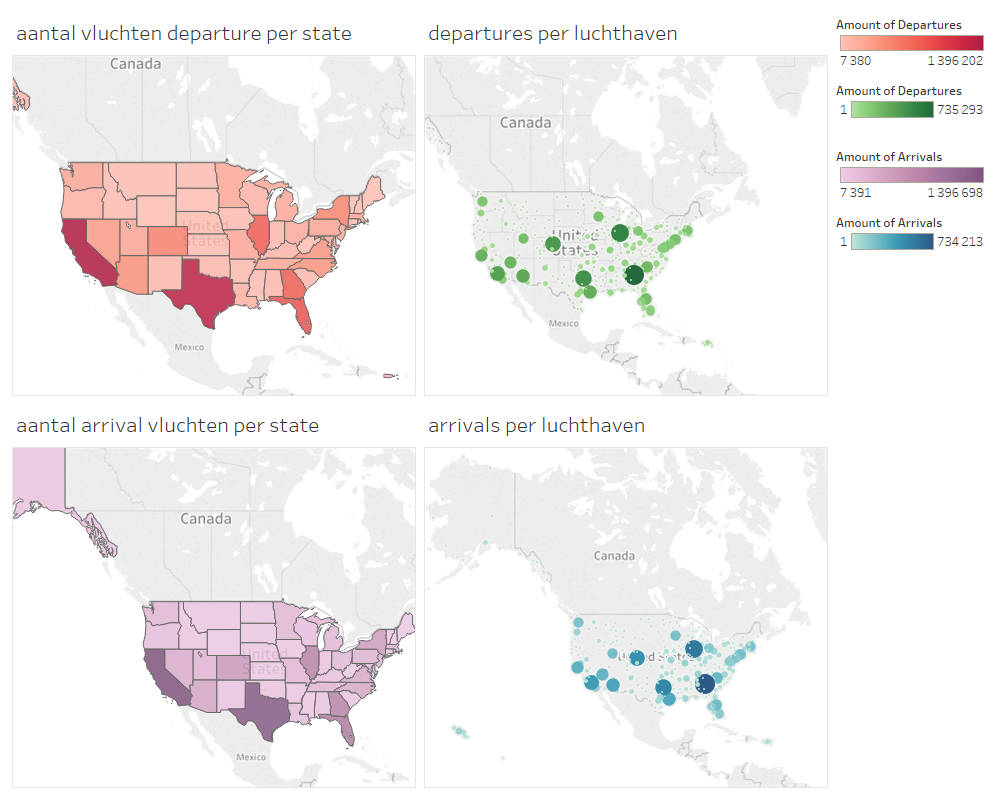


## Analyse aantal vluchten

Uit de analyse via Tableau hebben we ontdekt dat het grootste aantal vluchten aankomen en vertrekken uit Atlanta Airport, maar de twee drukste staten van zijn Californië gevolgd door Texas.

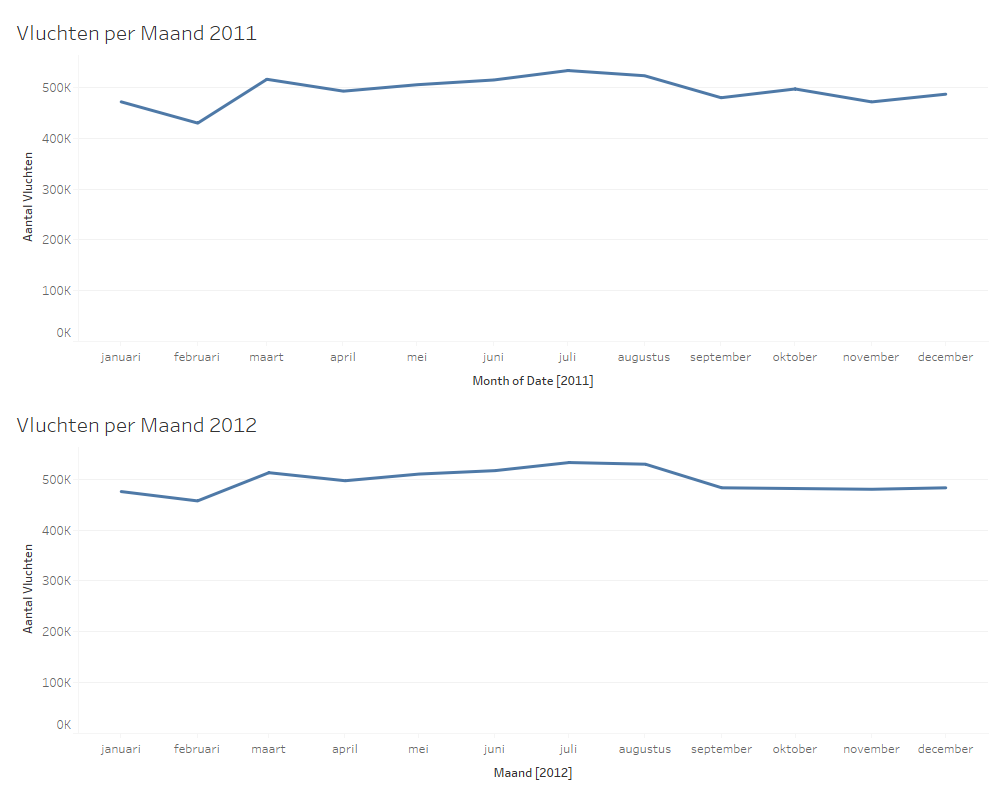
hoewel Atlanta de luchthaven is met de meeste vluchten heeft is dit de enige grote luchthaven in de staat Georgia Atlanta met 735 293 vertrekkende vluchten per jaar en de tweede grootste luchthaven is Savannah Hilton Head International Airport met slechts 17 550 vluchten per jaar.

Californië en Texas daarentegen hebben meerdere grote luchthavens waar de top 3 meer dan 90 000 vertrekkende Vluchten heeft jaar.

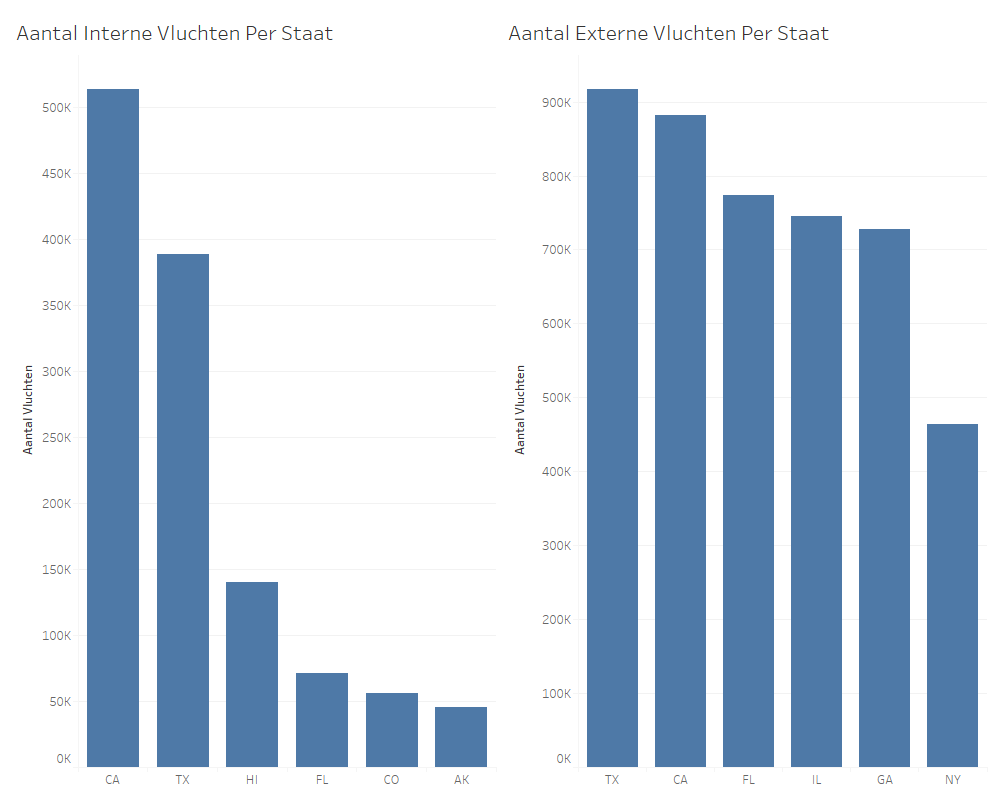


Als we naar het aantal vluchten per maand kijken, rekening houdend met de ontbrekende data van oktober 2012, kunnen we zien dat het aantal vluchten per maand overeen komen met elkaar over de twee jaren.

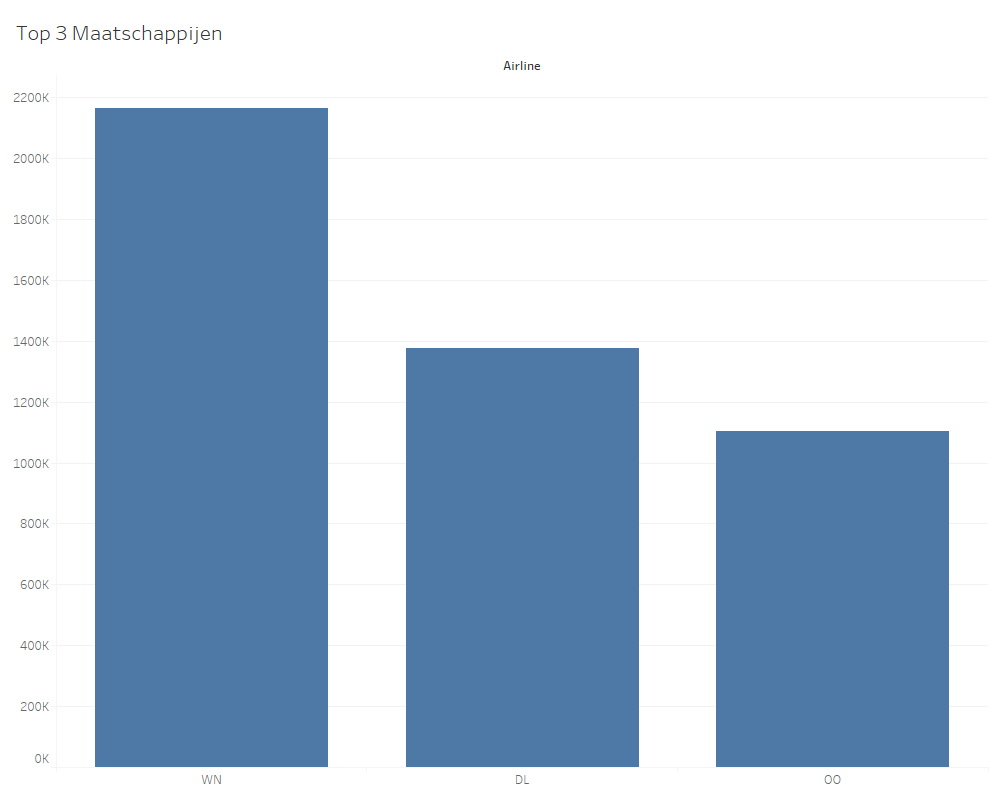
In beide jaren is februari de maand met het minst aantal vluchten dat mogelijks deels te verklaren is met de kortere duur van de maand. In maart is er een sterke stijging in het aantal vluchten om terug te zakken in april gevolgd door een geleidelijke stijging van het aantal vluchten over de maanden juni en juli gevolgd door een lichte daling augustus. In september zakt het aantal vluchten. In oktober 2011 stijgt deze terug in oktober, we kunnen aannemen dat dit mogelijk ook zo zou kunnen zijn in oktober 2012.



Bij Californië is het ons opgevallen dat iets meer dan een derde van de vluchten die vertrekken vanuit een luchthaven van Californië ook landen op een luchthaven van Californië. Texas heeft een gelijkaardig fenomeen maar hier gaat het slechts om iets meer dan een vijfde van de vluchten.



Het aantal vluchten per luchtvaartmaatschpij blijkt dat de top 3 maatschappijen de low-cost maatschappij Southwest Airlines voor 2 165 114 vand de vluchten verantwoordlijke zijn gevolgd door Delta Airlines voor 1 376 347 vluchten en SkyWest Airlines op de derde plaats met 1 103 192 vluchten.



## Analyse Vertragingen

Om over te gaan naar wat niemand graag heeft op zijn vlucht de vertragingen. Van de data die we hebben kunne we zien dat er tussen de maand april en juli in 2011 gemiddeld meer dan 10 minuten vertraging was per vlucht.

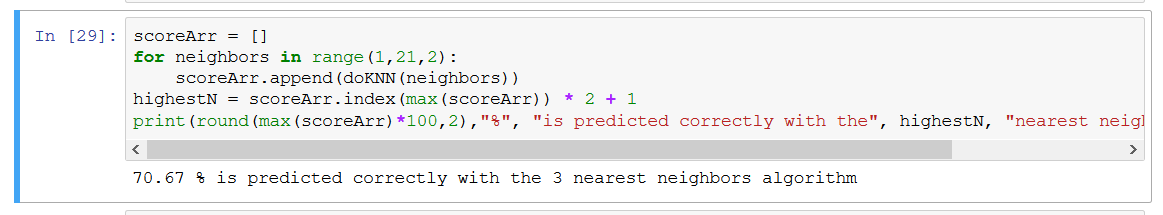
# Machine Learning

## Oorspronkelijk doel

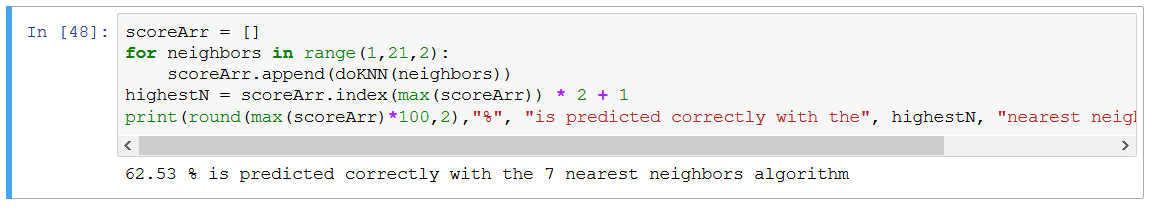
Toen we begonnen wilden we a.d.h.v. regressie de “departure\_delay”, of dus de vertraging, van een nieuwe vlucht voorspellen. We begonnen met gewone lineaire regressie met als input parameters de datum van vertrek, het uur van vertrek en het uur van aankomst. Al snel bleek dat lineaire regressie met deze parameters geen correcte voorspellingen maakte.

## Tweede poging

Om ook gebruik te maken van gegevens zoals vertrek luchthaven besloten we over te schakelen naar k-nearest neigbors classificatie. Als output klassen kozen we vluchten zonder vertraging, en vluchten met vertraging. Al snel behaalden we een redelijk resultaat van 70% met het ideaal aantal buren.



Daarnaast probeerden we ook classificatie met als klassen: geen vertraging, 0 tot 5 minuten vertraging, 5-10 minuten vertraging en 10+ minuten vertraging. Dit model behaalde een 62% accuraatheid. De verlaagde accuraatheid is uiteraard te verklaren door het toenemen van mogelijke klassen.



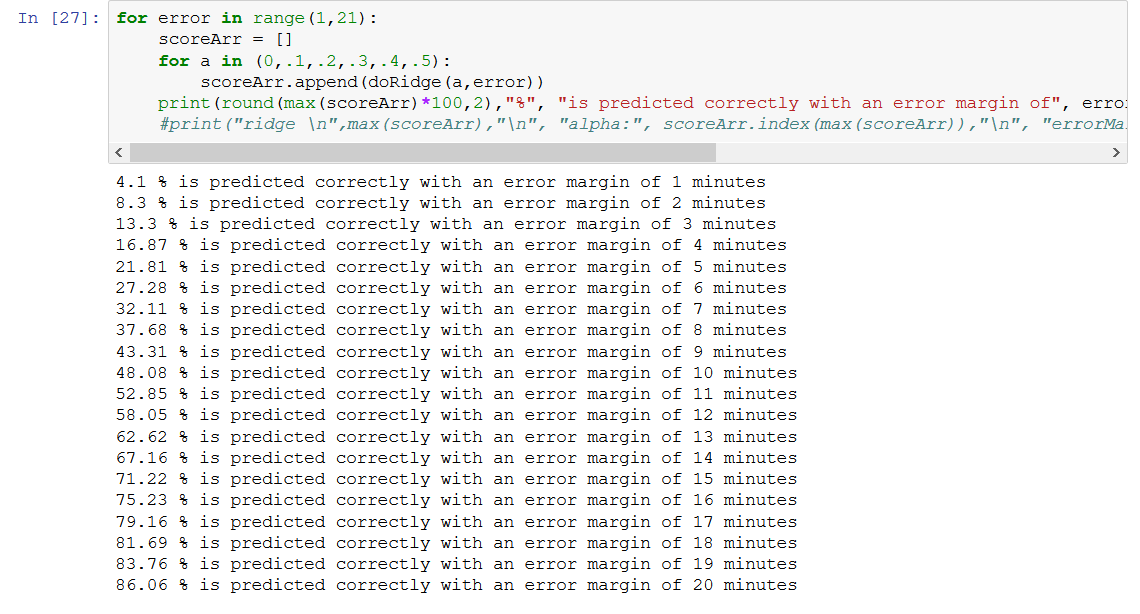
## Terugkeer naar het oorspronkelijke doel

Na ons wat meer te verdiepen in scikit-learn besloten we regressie toch nog een kans te geven, deze keer met succes. Door meer input parameters toe te voegen die een sterk verband hebben met de te voorspellen vertraging slaagden we erin om onze accuraatheid te verhogen.

Om dit te bereiken vroegen we ons af wat het dichtstbijzijnde getal is dat we kunnen berekenen. Dit waren de gemiddelde vertragingen per airline en airport.

We schreven ook onze eigen score functie die een voorspelling goedkeurt als ze binnen de 10 minuten van het echte antwoord zat.

## Laatste stappen

Onze eigen score functie is wellicht niet streng genoeg dus besloten we om niet enkel de accuraatheid met foutmarge 10 minuten te berekenen maar ook voor 1 tot en met 20 minuten. 

## Besluit

Omdat we met de ingebouwde score methode nooit meer dan 8% haalden nemen we aan dat deze methode zeer streng is op voorspellingen en ons eenvoudig model zeer zwaar afstraft. Maar uit onze eigen score methode kunnen we afleiden dat het model in de juiste richting gaat.

# NetworkX