Data quality report V2

Opdracht Data Exploratie ADS 2023-2024

School: HAN

Course: ADS  
Studenten:

Timoy van Balkom - 1592442  
Daan van Vugt – s636474  
Docent:  
Matthijs de Haan

Datum: 2024-03-08

Inhoudsopgave

[Inleiding 3](#_Toc160663246)

[Informatievragen 3](#_Toc160663247)

[Descriptieve statistische kentallen 4](#_Toc160663248)

[Continue features 4](#_Toc160663249)

[Categoriale features 4](#_Toc160663250)

[Histogrammen 5](#_Toc160663251)

[Temperatuur 5](#_Toc160663252)

[Dauwpunt 7](#_Toc160663253)

[Luchtvochtigheid 8](#_Toc160663254)

[Windrichting 10](#_Toc160663255)

[Windsnelheid 12](#_Toc160663256)

[Boxplots 14](#_Toc160663257)

[Vertrek vertraging 14](#_Toc160663258)

[Aankomst vertraging 16](#_Toc160663259)

[Temperatuur 18](#_Toc160663260)

[Dauwpunt 20](#_Toc160663261)

[Luchtvochtigheid 22](#_Toc160663262)

[Windrichting 24](#_Toc160663263)

[Windsnelheid 25](#_Toc160663264)

[Testen 26](#_Toc160663265)

[Proportietesten 26](#_Toc160663266)

[Difference of means testen 26](#_Toc160663267)

[Anova testen 27](#_Toc160663268)

[Scatter plots 29](#_Toc160663269)

[Pair plot 29](#_Toc160663270)

[Correlatiecoëfficiënten 30](#_Toc160663271)

[Vertrek vertraging en aankomst vertraging 31](#_Toc160663272)

[Wind richting en vertrek vertraging 32](#_Toc160663273)

[Wind snelheid en vertrek vertraging 33](#_Toc160663274)

[Temperatuur en dauwpunt 34](#_Toc160663275)

[Temperatuur en luchtvochtigheid 35](#_Toc160663276)

[Dauw punt en luchtvochtigheid 36](#_Toc160663277)

[Conclusie 37](#_Toc160663278)

[Wat is de verdeling van de vertragingstijden voor vluchten in de dataset? 37](#_Toc160663279)

[Hoe is de spreiding van temperatuur, dauw punt, luchtvochtigheid, windrichting en windsnelheid in de dataset? 37](#_Toc160663280)

[Hoe verhouden de verschillende weer data met elkaar? 38](#_Toc160663281)

[Figuren lijst 40](#_Toc160663282)

[Bibliografie 41](#_Toc160663283)

# Inleiding

In dit document wordt de datakwaliteit geverifieerd van de door de HAN aangeleverde nycflights13 dataset, samengesteld door Hadley Wickham. De dataset is te groot om volledig te controleren, daarom zal er gekeken worden naar 7 continue features en 3 categoriale features die specifiek gekozen zijn om te helpen bij het beantwoorden van onderstaande informatievragen. In de [conclusie](#_Conclusie) staan de antwoorden op de informatievragen, samengevat gebaseerd op de informatie die uit dit verslag is verkregen.

## Informatievragen

1. Wat is het effect van vertrek vertragingen op aankomst vertragingen?
2. Kunnen we op basis van temperatuur andere weersomstandigheden voorspellen?
3. Hebben wind snelheid en wind richting een effect op vertragingen?

# Descriptieve statistische kentallen

Hieronder staan voor de continue features en de categoriale features de descriptieve statische kentallen beschreven, met in totaal 7 continue en 3 categoriale features. Dit is het minimale aantal benodigde features voor het beantwoorden van de informatievragen, ook al zijn dit er meer dan het minimum van 5 continue features vanuit de opdracht.

De waardes in onderstaande tabellen zijn berekend met behulp van de matplotlib.pyplot library in Python.

## Continue features

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Feature | Count | % Miss. | Card. | Min. | 1ste Qrt. | Mean | Median | 3rd Qrt. | Max. | Std. Dev. |
| Dep\_delay | 336776 | 2.451 | 527 | -43.0 | -5.0 | 12.639 | -2.0 | 11.0 | 1301.0 | 40.210 |
| Arr\_delay | 336776 | 2.800 | 577 | -86.0 | -17.0 | 6.895 | -5.0 | 14.0 | 1272.0 | 44.633 |
| Temp | 336776 | 0.467 | 168 | 10.94 | 42.08 | 56.996 | 57.2 | 71.96 | 100.04 | 17.965 |
| Depw | 336776 | 0.467 | 147 | -9.94 | 26.06 | 41.631 | 42.8 | 57.92 | 78.08 | 19.353 |
| Humid | 336776 | 0.467 | 2442 | 12.74 | 43.99 | 59.557 | 57.73 | 75.33 | 100.0 | 19.656 |
| Wind\_dir | 336776 | 2.909 | 37 | 0.0 | 130.0 | 201.540 | 220.0 | 290.0 | 360.0 | 104.820 |
| Wind\_speed | 336776 | 0.485 | 34 | 0.0 | 6.905 | 11.114 | 10.357 | 14.960 | 42.579 | 5.574 |

## Categoriale features

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Feature | Count | % Miss. | Card. | Mode | Mode Freq. | Mode % | 2nd Mode | 2nd Mode Freq. | 2nd Mode % |
| Dest | 336776 | 0.0 | 105 | ORD | 17283 | 5.132 | ATL | 17215 | 5.112 |
| Origin | 336776 | 0.0 | 3 | EWR | 120835 | 35.880 | JFK | 111279 | 33.042 |
| Carrier | 336776 | 0.0 | 16 | UA | 58665 | 17.420 | B6 | 54635 | 16.223 |

# Histogrammen

De histogrammen zijn gemaakt met behulp van de matplotlib.pyplot library en Python en de density plots zijn gemaakt met behulp van de Seaborn library in Python.

## Temperatuur

Als we kijken naar het histogram van de temperatuur zien we een bimodale distributie, het histogram heeft 2 duidelijke pieken. Uit deze pieken kunnen we opmaken dat er twee temperatuurranges zijn die het meest voorkomen, met name rond de 40 en 70 graden Fahrenheit.

### Histogram

Afbeelding met Perceel, diagram, schermopname, lijn

Automatisch gegenereerde beschrijving

Figuur 1 Histogram temperatuur

### Density plot

Afbeelding met tekst, diagram, Perceel, schermopname

Automatisch gegenereerde beschrijving

Figuur 2 Density plot temperatuur

## Dauwpunt

Voor zowel het dauwpunt als de temperatuur geldt dat het een bimodale verdeling is. Alhoewel het in het histogram minder gemakkelijk te zien is, laat het density plot duidelijk zien dat er 2 pieken zijn rond 30 en 60 graden Fahrenheit.

### Histogram

Afbeelding met diagram, Perceel, lijn, schermopname

Automatisch gegenereerde beschrijving

Figuur 3 Histogram dauwpunt

### Density plot

Afbeelding met tekst, diagram, schermopname, Perceel

Automatisch gegenereerde beschrijving

Figuur 4 Density plot dauwpunt

## Luchtvochtigheid

Als we kijken naar de luchtvochtigheid is het wat lastiger om de verdeling te bepalen. Het heeft het meeste weg van een rechts-scheve verdeling, wat vooral in het density plot zichtbaar is. Het lijkt alsof er een lange staart aan de rechterkant ontstaat, wat betekent dat er een paar zeer hoge waardes zijn die het gemiddelde omhoogtrekken. Dit is goed zichtbaar in het histogram.

### Histogram

Afbeelding met diagram, Perceel, schermopname, lijn

Automatisch gegenereerde beschrijving

Figuur 5 Histogram luchtvochtigheid

### Density plot

Afbeelding met tekst, diagram, Perceel, schermopname

Automatisch gegenereerde beschrijving

Figuur 6 Density plot luchtvochtigheid

## Windrichting

Bij de windrichting zien we wederom een histogram die lijkt op een bimodale verdeling. Zeker in het density plot is duidelijk te zien dat er twee pieken zijn, de linker piek weliswaar een stuk lager dan de rechter piek. Nu is het wel zo dat de windrichting gemeten is in graden, waarbij 0 en 360 in elkaar overlopen en het diagram dus zichzelf herhaalt. Wat lijkt op een linker piek is eigenlijk nog een daling van de rechter piek, we hebben dus te maken met een gaussverdeling.

### Histogram

Afbeelding met Perceel, lijn, diagram, tekst

Automatisch gegenereerde beschrijving

Figuur 7 Histogram wind richting

### Density plot

Afbeelding met tekst, diagram, Perceel, schermopname

Automatisch gegenereerde beschrijving

Figuur 8 Density plot wind richting

## Windsnelheid

Het histogram van de windsnelheid lijkt op een gaussverdeling, omdat de meeste waardes dicht bij het gemiddelde liggen. We zien een uitschieter aan de linker kant, maar alles tezamen lijkt het gros van de waardes dicht bij het gemiddelde te liggen zonder al te veel uitschieters. Als we terugkijken naar de waardes in de tabel met continue features zien we dat het gemiddelde en de mediaan erg dicht bij elkaar liggen. De afstand van het 1ste kwartiel tot het gemiddelde en het kwartaal tot het 3de kwartiel zijn ook vrijwel gelijk. Dit wijst op een gelijke verdeling.

### Histogram

Afbeelding met Perceel, lijn, diagram, tekst

Automatisch gegenereerde beschrijving

Figuur 9 Histogram wind snelheid

### Density plot

Afbeelding met tekst, diagram, Perceel, schermopname

Automatisch gegenereerde beschrijving

Figuur 10 Density plot wind snelheid

# Boxplots

De boxplots zijn gemaakt met behulp van de Seaborn library in Python.

## Vertrek vertraging

Als we kijken naar de boxplot van de vertragingen bij vertrek zien we dat de spreidingen tussen de minimale en maximale waarde en daarbij ook tussen het 1ste kwartiel het 3de kwartiel klein is. Wel zien we veel outliers, met enorme uitschieters vergeleken met de bovengrens.

Afbeelding met lijn, tekst, Perceel, diagram

Automatisch gegenereerde beschrijving

Figuur 11 Boxplot vertrek vertraging

Als we specifieker gaan kijken naar de verhoudingen van vertragingen bij vertrek per carrier lijkt het op het eerste ogenblik alsof de spreiding redelijk gelijk is. Toch zien we als we goed kijken dat met name EV een grote delta heeft tussen min en max met redelijk wat outliers vergeleken met de rest en AS slechts 1 outlier met een gemiddelde delta tussen min en max vergeleken met derest.

Afbeelding met tekst, diagram, Perceel, lijn

Automatisch gegenereerde beschrijving

Figuur 12 Boxplot vertrek vertraging gegroepeerd op luchtvaartmaatschappij

## Aankomst vertraging

Als we naar de boxplot van vertragingen bij aankomst kijken zien we een beeld dat we verwachten te zien. De spreiding is iets groter dan bij de vertragingen bij vertrek, dat is met name te danken aan waardes aan de linkerkant van de boxplot. Dit is logisch, omdat vluchten ook eerder kunnen arriveren, in tegenstelling tot eerder vertrekken. Verder is de spreiding van de outliers vergelijkbaar met die van de vertragingen bij vertrek, wat kan betekenen dat de vluchten die vertraagd waren dezelfde delta hadden van hun aankomst tijd als hun vertrek tijd. dit is ook logisch omdat de vliegtijd niet per se langer of korter wordt wanneer er vertraging plaats vind voor vertrek.

Afbeelding met tekst, lijn, Perceel, diagram

Automatisch gegenereerde beschrijving

Figuur 13 Boxplot aankomst vertraging

Als we kijken naar de verhouding tussen vertraging bij aankomst en de vliegtuigmaatschappij zien we net als bij de boxplot hierboven dat deze redelijk in lijn ligt met de boxplot van vertraging bij vertrek, met wederom meer spreiding naar de linkerkant van de boxplot (onder de 0 in dit geval) en daarmee ook een grotere delta tussen min en max.

Afbeelding met tekst, diagram, Perceel, nummer

Automatisch gegenereerde beschrijving

Figuur 14 Boxplot aankomst vertraging gegroepeerd op luchtvaartmaatschappij

## Temperatuur

De boxplot van de temperatuur heeft geen outliers zoals hieronder te zien, wat we wel zien is dat de delta tussen het minimum en het 1ste kwartiel en de delta tussen het 3de kwartiel en het maximum vrij groot is, waarbij de spreiding tussen het 1ste kwartiel en het 3de kwartiel minder groot is in verhouding. Dit laat duidelijk zien dat de temperatuur over het algemeen binnen een klein bereik schommelt met een aantal uitzonderingen.

Afbeelding met schermopname, Rechthoek, tekst, plein

Automatisch gegenereerde beschrijving

Figuur 15 Boxplot temperatuur

Als we kijken naar de temperatuur op verschillende momenten van de dag zien we een goed te verklaren beeld. De temperatuur stijgt tot een hoogtepunt rond de middaguren, waarna het weer langzaam afzakt rond de avond. Ook zien we een goede spreiding in uren, van 5 uur ’s nachts tot 11 is ’s avonds.

Afbeelding met tekst, diagram, lijn, Perceel

Automatisch gegenereerde beschrijving

Figuur 16 Boxplot temperatuur gegroepeerd op uur

## Dauwpunt

Als we kijken naar de boxplot van het dauwpunt zien we een grote delta tussen het minimum en het 1ste kwartiel. Ook zien we een relatief kleine delta tussen het 3de kwartiel en het maximum. Hier kunnen we uit concluderen dat het dauwpunt gemiddeld wat aan de hogere kant zit met uitschieters naar onder. Er zijn geen outliers.

Afbeelding met schermopname, diagram, Rechthoek, plein

Automatisch gegenereerde beschrijving

Figuur 17 Boxplot dauwpunt

Als we kijken naar het dauwpunt per uur van de dag zien we een redelijk lineaire lijn voor de data tussen het 1ste en het 3de kwartiel met wat fluctuatie tussen het minimum en maximum. Deze min en max fluctuatie is voor iedere uur van de dag redelijk in lijn met elkaar, er is geen moment waar de max significant verder uitschiet dan de min in vergelijking met de overige uren.

Afbeelding met tekst, lijn, diagram, Rechthoek

Automatisch gegenereerde beschrijving

Figuur 18 Boxplot dauwpunt gegroepeerd op uur

## Luchtvochtigheid

Als we kijken naar de boxplot van luchtvochtigheid zien we geen outliers. We zien hetzelfde patroon als bij de temperatuur: Een grote delta tussen het minimum en maximum, met een relatief kleine delta tussen het 1ste kwartiel en 3de kwartiel. Dit is te verwachten omdat luchtvochtigheid en temperatuur veelal in lijn met elkaar liggen.

Afbeelding met diagram, schermopname, Rechthoek, lijn

Automatisch gegenereerde beschrijving

Figuur 19 Boxplot luchtvochtigheid

Als we kijken naar de luchtvochtigheid op verschillende punten van de dag zien we het spiegelbeeld van de boxplot met de temperatuur per uur. Idealiter zouden we hier een zelfde soort trend terug zien omdat naarmate de lucht opwarmt er ook meer vocht in de lucht kan blijven hangen echter is dit ook afhankelijk van het klimaat waardoor het moeilijk te concluderen is of onderstaande boxplot een realistisch beeld schetst van de luchtvochtigheid.

Afbeelding met tekst, diagram, lijn, Perceel

Automatisch gegenereerde beschrijving

Figuur 20 Boxplot luchtvochtigheid gegroepeerd op uur

## Windrichting

Als we kijken naar de boxplot van de windrichting zien we geen outliers en zien we dat de windrichting het vaakst richting de 200 - 250 graden ligt. Waar wel rekening mee gehouden moet worden is dat na 360 de wind weer 0 is omdat het wordt gemeten in graden t/m 360 graden.

Afbeelding met schermopname, tekst, diagram, Rechthoek

Automatisch gegenereerde beschrijving

Figuur 21 Boxplot wind richting

## Windsnelheid

Bij de boxplot van de windsnelheid zien we dat het gros van de data in het midden zit, met een iets grotere delta tussen 3de kwartiel en max als tussen de min en 1ste kwartiel. In de boxplot zien we een paar outliers aan de rechter kant vanaf ongeveer 27 en hoger. Windsnelheden met deze snelheid kunnen bij uitzondering voorkomen. Hierom lijken de outliers logische data punten die hoogst waarschijnlijk zijn gemeten bij hevige wind/storm, het is wenselijk deze data mee te nemen omdat het zeker voor kan komen in de praktijk en een groot effect heeft op het vliegverkeer.

(het gros van het data zit in het midden – dat is altijd zo bij een boxplot, wat is het midden? Een betere beschrijving zou zijn dat het gemiddelde rond de 10 ligt (10 wat? Mph? Kph?))  
(ik denk ook dat een belangrijke conclusie uit dit diagram is dat je deze outliers niet moet verwerpen omdat die stormwinden juist interessant zijn voor vertragingen)

Afbeelding met diagram, lijn, schermopname, Rechthoek

Automatisch gegenereerde beschrijving

Figuur 22 Boxplot wind snelheid

# Testen

## Proportietesten

Voor de proportietesten wordt het betrouwbaarheidsinterval voor vertraging van vluchten berekend. Ten eerste wordt er gekeken naar alle vluchten, waarna 3 specifieke luchtvaartmaatschappijen los zijn beschreven waar ook een betrouwbaarheidsintervalberekening voor is gemaakt. Er wordt voor iedere berekening gebruik gemaakt van de normaal distributie in verband met de grootte van de dataset en een alpha van 0.05, tenzij anders gespecificeerd.

### Alle vluchten

* Vertrek: 0.3797 – 0.3830
* Aankomst: 0.3933 – 0.3966

Dit betekent dat we voor het vertrek 95% zeker zijn dat er een kans op vertraging is tussen de 37.97% en 38.30% en voor de aankomst tussen de 39.33% en 39.66%. Dit betekent dat we met veel zekerheid kunnen zeggen waar de werkelijke proportie waarschijnlijk ligt.

### DL

* Vertrek: 0.3126 – 0.3210
* Aankomst: 0.3369 – 0.3454

Dit betekent dat we voor het vertrek 95% zeker zijn dat er een kans op vertraging is tussen de 31.26% en 32.10% en voor de aankomst tussen de 33.69% en 34.54% voor deze luchtvaartmaatschappij. Dit betekent dat we met veel zekerheid kunnen zeggen waar de werkelijk proportie waarschijnlijk ligt.

### HA

* Vertrek: 0.1592 – 0.2443
* Aankomst: 0.2359 – 0.3314

Dit betekent dat we voor het vertrek 95% zeker zijn dat er een kans op vertraging is tussen de 15.92% en 24.43% en voor de aankomst tussen de 23.59% en 33.14% voor deze luchtvaartmaatschappij. Dit betekent dat we met minder zekerheid kunnen zeggen waar de proportie ligt maar zeker niet met weinig zekerheid.

### EV

* Vertrek: 0.4230 – 0.4313
* Aankomst: 0.4478 – 0.4562

Dit betekent dat we voor het vertrek 95% zeker zijn dat er een kans op vertraging is tussen de 42.30% en 43.13% en voor de aankomst tussen de 44.78% en 45.62% voor deze luchtvaartmaatschappij. Dit betekent dat we met veel zekerheid kunnen zeggen waar de werkelijk proportie waarschijnlijk ligt.

## Difference of means testen

Voor de difference of means testen wordt er bij een aantal vliegtuigmaatschappijen gekeken naar de verschillen in het gemiddelde van vertragingen bij zowel vertrek als aankomst. De nulhypothese mag alleen verworpen worden als de p-waarde kleiner is dan 0.05.

### Nulhypothese

* H0 = μ1 = μ2
* HA = μ1 ≠ μ2

### F9 en YV

In onderstaande tabel zien we dat voor zowel vertrek als aankomst de p-waarde hoger is dan 0.05, hierdoor kunnen we niet concluderen of er verschillen zitten in de gemiddelde vertragingstijden tussen F9 en YV zowel bij vertrek als aankomst, de gemeten verschillen kunnen door toeval zijn ontstaan.

|  |  |  |
| --- | --- | --- |
| **F9 en YV** | **T-statistiek** | **P-waarde** |
| **Vertrek** | 0.3896 | 0.6969 |
| **Aankomst** | 1.9101 | 0.0563 |

### UA en OO

In onderstaande tabel zien we dat voor zowel vertrek als aankomst de p-waarde hoger is dan 0.05, hierdoor kunnen we niet concluderen of er verschillen zitten in de gemiddelde vertragingstijden tussen UA en OO zowel bij vertrek als aankomst, de gemeten verschillen kunnen door toeval zijn ontstaan.

|  |  |  |
| --- | --- | --- |
| **AA en MQ** | **T-statistiek** | **P-waarde** |
| **Vertrek** | -0.0724 | 0.9423 |
| **Aankomst** | -1.0998 | 0.2714 |

### EV en F9

In onderstaande tabel zien we dat voor vertrek de p-waarde hoger is dan 0.05, hierdoor kunnen we niet concluderen of er verschillen zitten in de gemiddelde vertragingstijden tussen UA en OO bij vertrek, de gemeten verschillen kunnen door toeval zijn ontstaan.

Voor de aankomst is de p-waarde lager dan 0.05, hierdoor kunnen we de nulhypothese verwerpen en concluderen dat het verschil in gemiddelde vertraging significant is tussen EV en F9 bij aankomst, de waargenomen verschillen kunnen niet alleen door kans toeval verklaard worden.

|  |  |  |
| --- | --- | --- |
| **EV en F9** | **T-statistiek** | **P-waarde** |
| **Vertrek** | -0.1444 | 0.8815 |
| **Aankomst** | -3.1731 | 0.0015 |

## Anova testen

Onderstaande nulhypothese is van toepassing op de Anova testen.

### Nulhypothese

* H0 = μ1 = μ2
* HA = μ1 ≠ μ2

### Temperatuur per maand

Om de temperatuur per maand te testen wordt gebruik gemaakt van een one way Anova test met behulp van de Scipy library in Python. Hieronder staan de f-statistiek en p-waarde die met behulp van de test berekent zijn. We verwachten hier te zien dat er een duidelijk verschil is in temperatuur per maand.

* F-statistiek: 108415.6325
* P-waarde: 0.0

Met een p-waarde van 0.0 kunnen we de nulhypothese verwerpen en kunnen we overduidelijk concluderen dat de temperatuur per maand verschilt. Om dit verder te onderzoeken wordt er een post-hoc test uitgevoerd, meer specifiek de Tukey HSD test.

Hieronder zien we een boxplot met op de x-as de maanden en op de y-as de temperaturen. We zien hier heel duidelijk dat de temperatuur stijgt en piekt naarmate we richting de zomer gaan, waarna de gemiddelde waarde weer daalt wanneer we richting de winter gaan. Dit is exact wat we verwachten te zien, hiermee kunnen we concluderen dat er een wezenlijk verschil is in de temperatuur per maand in deze dataset.

Afbeelding met tekst, diagram, Plan, Technische tekening

Automatisch gegenereerde beschrijving

Figuur 23 Boxplot Anova test temperatuur per maand

# Scatter plots

Hieronder zien we allereerst een pair plot van alle numerieke data in de dataset. Daaronder wordt er verder ingegaan op 6 feature-to-feature relaties, namelijk dep\_delay en arr\_delay, wind\_dir en dep\_delay, wind\_speed en dep\_delay, temp en dewp, temp en humid, dewp en humid

## Pair plot

Afbeelding met tekst, kaart, patroon

Automatisch gegenereerde beschrijving

Figuur 24 Pairplot

## Correlatiecoëfficiënten

In onderstaande tabel staan de correlatiecoëfficiënten van het bovenstaande pair plot, alle rood gemarkeerde kolommen zijn kolommen waar een feature wordt vergeleken met zichzelf en daarbij altijd een correlatiecoëfficiënt van 1 zal hebben.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Dep\_delay | Arr\_delay | Temp | Dewp | Humid | Wind\_dir | Wind\_speed |
| Dep\_delay | 1 | 0.9148 | 0.0615 | 0.1024 | 0.1175 | -0.0176 | 0.0474 |
| Arr\_delay | 0.9148 | 1 | 0.0330 | 0.0869 | 0.1414 | -0.0254 | 0.0641 |
| Temp | 0.0615 | 0.0330 | 1 | 0.8823 | 0.0355 | -0.0999 | -0.1468 |
| Dewp | 0.1024 | 0.0869 | 0.8823 | 1 | 0.4922 | -0.2360 | -0.2212 |
| Humid | 0.1175 | 0.1414 | 0.0355 | 0.4922 | 1 | -0.3250 | -0.1922 |
| Wind\_dir | -0.0176 | -0.0254 | -0.0999 | -0.2360 | -0.3250 | 1 | 0.3416 |
| Wind\_speed | 0.0474 | 0.0641 | -0.1468 | -0.2212 | -0.1922 | 0.3416 | 1 |

## Vertrek vertraging en aankomst vertraging

In onderstaande scatter plot zien we een duidelijke correlatie tussen vertragingen bij vertrek en vertraging bij aankomst. Naarmate de vertraging bij vertrek toeneemt neemt ook de vertraging bij aankomst toe. Dit is ook goed te zien in de [correlatie coëfficiënten tabel](#_Correlatie_coëfficiënten) waar de correlatie coëfficiënt tussen dep\_delay en arr\_delay 0.9148 is. Deze waarde ligt bijzonder dicht tegen de 1 wat betekent dat er een bijna perfecte positieve correlatie is tussen de twee features.

Hieruit kunnen we concluderen dat het effect van vertragingen bij vertrek erg groot is op de vertragingen bij aankomst. Door de hoge correlatie tussen de vertrek en aankomst vertraging kunnen we een goede voorspelling doen dat wanneer een vlucht vertraging heeft bij vertrek, deze ook vertraging zal hebben bij aankomst en ook hoeveel die vertraging bij aankomst ongeveer zal zijn.

Afbeelding met tekst, schermopname, Perceel, lijn

Automatisch gegenereerde beschrijving

Figuur 25 Scatter plot vertrek vertraging en aankomst vertraging

## Wind richting en vertrek vertraging

Als we naar onderstaande scatter plot kijken zien we geen duidelijke correlatie tussen de wind richting en de vertrek vertraging. Als we ook kijken naar de correlatie coefficient zien we een waarde van -0.0176, dit wijst zelfs op een hele lichte negatieve correlatie maar de waarde ligt te dicht bij de nul waardoor we kunnen concluderen dat er geen correlatie is tussen wind richting en vertrek vertraging.

Hierdoor kunnen we ook niet concluderen dat wind richting een effect heeft op de vertrek vertraging, althans niet aan de hand van deze scatter plot.

Afbeelding met tekst, schermopname, diagram, lijn

Automatisch gegenereerde beschrijving

Figuur 26 Scatter plot wind richting en vertrek vertraging

## Wind snelheid en vertrek vertraging

Als we naar onderstaande scatter plot kijken zien we geen duidelijke correlatie tussen wind snelheid en vertrek vertragingen. Als we naar de correlatie coëfficiënt kijken zien we dat deze 0.0474 is, dit duidt dat er nauwelijks een correlatie is tussen de wind snelheid en vertrek vertragingen. Wat opvalt is dat het niet echt uit lijkt te maken of de windsnelheid hoog is, want ook bij hoge windsnelheden is de vertraging niet per se langer dan bij lage windsnelheiden, het tegendeel is zelfs waar want we zien een aantal flinke vertragingen bij vrij lage windsnelheden. We zouden de aanname kunnen maken dat hoge windsnelheden van korte duur zijn en daarom de vertraging relatief kort is maar dit is niet uit te lezen uit de scatter plot.

We kunnen hierom niet concluderen of de wind snelheid een effect heeft op de vertrek vertragingen, althans niet aan de hand van deze scatter plot.

Afbeelding met tekst, schermopname, diagram, lijn

Automatisch gegenereerde beschrijving

Figuur 27 Scatter plot wind snelheid en vertrek vertraging

## Temperatuur en dauwpunt

In onderstaande scatter plot zien we net als bij de vertragingen een positieve correlatie tussen temperatuur en dauwpunt. De overeenkomst is niet zo sterk als bij de vertragingen, maar er is wel degelijk een positieve correlatie. Er is meer variatie in de datapunten, wat je kan zien aan de grotere spreiding op een zichtbaar lineair verband. Als we kijken naar de [correlatie coëfficiënten tabel](#_Correlatie_coëfficiënten) zien we een correlatiecoëfficiënt van 0.8823. Dit laat zien dat er een sterk positieve correlatie is tussen de temperatuur en het dauwpunt, 0.8823 ligt ook dicht bij 1.

We kunnen daarom concluderen dat we op basis van temperatuur het dauwpunt kunnen voorspellen, door de hoge correlatie tussen beide waardes zouden we een goede voorspellen kunnen doen met betrekking tot het dauwpunt aan de hand van de temperatuur.

Afbeelding met tekst, schermopname, diagram, lijn

Automatisch gegenereerde beschrijving

Figuur 28 Scatter plot temperatuur and dauw punt

## Temperatuur en luchtvochtigheid

In het onderstaande scatter plot zien we geen correlatie tussen temperatuur en luchtvochtigheid. De data punten zijn erg ver verspreid en er is geen verband zichtbaar in het diagram. We zien geen positieve maar ook geen negatieve correlatie. Als we kijken naar de [correlatie coëfficiënten tabel](#_Correlatie_coëfficiënten) dan zien we een correlatie coëfficiënt van 0.0355, dit getal ligt heel dicht bij de 0. Daarmee wordt bevestigd dat er geen correlatie is tussen temperatuur en luchtvochtigheid.

Hieruit kunnen we ook concluderen dat we niet aan de hand van de temperatuur de luchtvochtigheid kunnen voorspellen, de correlatie is te laag, of eigenlijk bestaat deze nauwelijks, wat er voor zorgt dat er heel moeilijk is om aan de hand van de temperatuur de luchtvochtigheid te voorspellen.

Afbeelding met tekst, schermopname, diagram, patroon

Automatisch gegenereerde beschrijving

Figuur 29 Scatter plot temperatuur and luchtvochtigheid

## Dauw punt en luchtvochtigheid

In het onderstaande scatter plot zien we een matige tot sterke correlatie tussen het dauwpunt en de luchtvochtigheid. Er lijkt wel enigszins een lineair verband duidelijk te worden uit het diagram, de spreiding is echter groot. Als we kijken naar de [correlatie coëfficiënten tabel](#_Correlatie_coëfficiënten) zien we een correlatie coëfficiënt van 0.4922 tussen dauwpunt en luchtvochtigheid. Dit toont aan dat er een matige correlatie bestaat tussen de features, die vrijwel precies tussen de 0 en 1 in zit. Het is moeilijk te zeggen of beide features een directe invloed hebben op elkaar maar er is wel te zien dat er een correlatie bestaat tussen dauw punt en luchtvochtigheid.

Omdat we aan de hand van temperatuur wel het dauwpunt kunnen voorspellen en er een toch sterke positieve correlatie is tussen dauwpunt en luchtvochtigheid zouden we met de temperatuur via het dauwpunt nog wel een redelijk betrouwbare voorspelling kunnen doen met betrekking tot de luchtvochtigheid. Wel opgelet dat we dan een voorspelling zouden doen aan de hand van een voorspelling, dit maakt het wel minder betrouwbaar maar wel betrouwbaarder dan aan de hand van de temperatuur de luchtvochtigheid voorspellen.

Afbeelding met tekst, schermopname, diagram, kaart

Automatisch gegenereerde beschrijving

Figuur 30 Scatter plot dauwpunt en luchtvochtigheid

# Conclusie

Aan de hand van dit data-analyse verslag kunnen we antwoord geven op de 3 informatievragen, gespecificeerd in [de inleiding](#_Informatievragen). We kunnen ook concluderen dat de onderzochte data een goede spreiding heeft met logische data punten, dit betekent dat de data zeer waarschijnlijk realistisch is en compleet genoeg om te gebruiken voor trainingsdoeleinde.

## Wat is de verdeling van de vertragingstijden voor vluchten in de dataset?

Als we kijken naar het box plot met betrekking tot de vertragingen bij zowel aankomst als vertrek zien we dat het gros van de data punten rond de 0 ligt, met toch relatief veel uitschieters boven de maximum. Als we kijken naar de vliegtuigmaatschappijen apart dan zien we dat hier een goede verdeling is van de data punten, we zien ook een duidelijk onderscheid tussen vliegtuigmaatschappijen waar de een meer last heeft van vertragingen dan de ander, dit verwachten we ook maar we verwachten geen vliegtuigmaatschappijen met geen vertragingen en die zien we ook niet.

Als we naar de proportie testen en de difference of means testen kijken dan zien we bij de proportietesten dat er genoeg data is om met grote zekerheid te zeggen waar de proportie ligt en ook als we naar individuele maatschappijen kijken weten we bij 2 van de 3 maatschappijen die getest zijn ook met grote zekerheid waar de proportie ligt en bij 1 van de 3 gekozen maatschappijen is de zekerheid minder groot maar ook niet klein. Bij de difference of means testen zien we uit de gekozen test data op 1 plek een p-waarde onder de 0.05, dit betekent dat we slechts op 1 plek de nulhypothese kunnen verwerpen, op alle andere plekken kunnen we niet genoeg zekerheid concluderen dat er een verschil is in gemiddelde tussen de gekozen vliegtuigmaatschappijen.

Als laatste wordt er nog bij de [scatter plots](#_Vertrek_vertraging_en) gekeken naar de correlatie tussen vertraging bij vertrek en aankomst en hier zien we een hele duidelijke correlatie, dit willen we ook zien want als een vliegtuig vertraging heeft bij vertrek heeft deze een grote kans om ook vertraging op te lopen bij aankomst. De hoge correlatiecoëfficiënt toont aan dat de data in lijn met elkaar ligt.

Alle analyses samen genomen kunnen we concluderen dat de verdeling van vertragingstijden voor vluchten goed is, zeker goed genoeg om als dataset te gebruiken voor eventuele trainingsdoeleinden.

## Hoe is de spreiding van temperatuur, dauw punt, luchtvochtigheid, windrichting en windsnelheid in de dataset?

Als we kijken naar de temperatuur zien we een goede spreiding in de histogram en boxplot van ongeveer 10 tot 100 Fahrenheit, dit correspondeert met -12 tot 38 graden Celsius, met het gros van de dat in het midden, precies zoals we verwachten. Ook wanneer we kijken naar de boxplot waar naar de temperatuur per uur wordt gekeken zien we dat de temperatuur richting het warmste punt van de dag stijgt en daarna weer daalt, wederom precies zoals we verwachten.

Als we kijken naar het dauw punt zien we dat deze spreid van ongeveer -10 tot 80 graden Fahrenheit, dauw punt ligt veelal in directe correlatie met de temperatuur dus dat we hier een vergelijkbare spreiding zien als bij de temperatuur maar dan iets lager in waardes is exact wat we verwachten te zien.

Als we bij de luchtvochtigheid kijken zien we wederom een goede spreiding van data, de luchtvochtigheid ligt tussen de ongeveer 10 en 100 procent, dit zijn realistische waardes. We zien in de boxplot waar wordt gekeken naar de luchtvochtigheid per uur een spiegelbeeld boxplot van de temperatuur, dit is niet per se goed want over het algemeen geldt dat hoe warmer de lucht hoe meer vocht deze bevat maar dit is ook afhankelijk van het klimaat en het is lastig hier conclusies uit te trekken.

Als laatste hebben we nog de wind richting en wind snelheid, de windrichting loopt van 0 tot 360 graden zoals het hoort, het gemiddelde ligt hier op 220, dit zou betekenen dat de meeste wind uit het zuid westen komt. Het is moeilijk te beoordelen of deze data waarheidsgetrouw is omdat de gemiddelde windrichting verschilt per locatie. Voor de windsnelheid zien we dat deze ligt ergens tussen de 0 en 27 met een aantal uitschieters daarboven, dit is wel wat we verwachten te zien waar de uitschieters vermoedelijk zware windstoten/stormen waren die nauwelijks voorkomen.

We kunnen aan de hand van de analyses wel concluderen dat er een goede spreiding is van de verschillende weerdata, de waardes die we terug vinden zijn ook realistisch en niet waardes die eigenlijk nooit voor zouden kunnen komen. Ook de gemiddeldes hebben realistische waardes waardoor we kunnen zeggen dat de weer data spreiding goed is en geschikt om te gebruiken voor eventuele trainingsdoeleinde.

## Hoe verhouden de verschillende weer data met elkaar?

Naast de spreiding is het ook goed om te weten hoe de verschillende weerdata met elkaar in verhouding staan. Hiervoor zijn verschillende scatter plots gemaakt die de correlatie tussen een aantal weerdata aantonen.

Als we allereerst kijken naar de temperatuur en het dauw punt zien we een zeer sterke positieve correlatie, dit willen we ook zien omdat de temperatuur en het dauw punt in de realiteit nauw met elkaar zijn verbonden en met een correlatie coëfficiënt van 0.8823 is het veilig om te stellen dat dit ook het geval is in de aangeleverde dataset.

Als we kijken naar de temperatuur en de luchtvochtigheid zien we een slecht positieve correlatie, en met een correlatie coëfficiënt van 0.0355 is het duidelijk dat in deze dataset geen meetbare verbinding is tussen temperatuur en luchtvochtigheid. We zien dit ook al terug in de boxplots waar we naar de temperatuur en luchtvochtigheid per uur kijken, hier zien we dat de luchtvochtigheid daalt naar mate de temperatuur stijgt en visa versa terwijl veelal warmer lucht een hogere luchtvochtigheid betekent, dit is echter ook sterk afhankelijk van het klimaat en de relatie tussen temperatuur en luchtvochtigheid is niet rechtlijnig. Desalniettemin hadden we een hogere correlatie verwacht tussen de beide features.

Als we kijken naar het dauw punt en de luchtvochtigheid zien we, met een correlatie coëfficiënt van 0.4922, een matige tot sterke positieve correlatie tussen het dauw punt en de luchtvochtigheid, over het algemeen zijn dauw punt en luchtvochtigheid wel met elkaar verbonden en het is daarom goed om te zien dat er een sterke positieve correlatie is tussen de 2 in deze dataset.

Als laatste hebben we de correlatie tussen de wind snelheid en wind richting, we zien hier een matige positieve correlatie, dit wel een beetje wat we verwachten omdat er wel zeker een relatie is tussen wind richting en wind snelheid is dit ook zeker niet altijd het geval, wind kan veranderen net als de snelheid, nu zijn er wel punten waar meestal hogere wind snelheden vandaan komen en dit correspondeert ook met de matige positieve correlatie.

Aan de hand van de analyses kunnen we, op de correlatie tussen temperatuur en luchtvochtigheid na, concluderen dat de verhoudingen tussen de verschillende weer data logisch en realistisch is en daarmee geschikt is om te gebruiken voor eventuele trainingsdoeleinden. Wat betreft de correlatie tussen temperatuur en luchtvochtigheid is dit wat lastig omdat deze veelal in verband liggen met elkaar maar het is niet rechtlijnig, wat betekent dat het ook niet hoeft, aan de hand van de verrichte analyse is het niet mogelijk om te concluderen of de correlatie tussen temperatuur en luchtvochtigheid logisch is of niet.

# Figuren lijst

[Figuur 1 Histogram temperatuur 5](#_Toc160663216)

[Figuur 2 Density plot temperatuur 6](#_Toc160663217)

[Figuur 3 Histogram dauwpunt 7](#_Toc160663218)

[Figuur 4 Density plot dauwpunt 7](#_Toc160663219)

[Figuur 5 Histogram luchtvochtigheid 8](#_Toc160663220)

[Figuur 6 Density plot luchtvochtigheid 9](#_Toc160663221)

[Figuur 7 Histogram wind richting 10](#_Toc160663222)

[Figuur 8 Density plot wind richting 11](#_Toc160663223)

[Figuur 9 Histogram wind snelheid 12](#_Toc160663224)

[Figuur 10 Density plot wind snelheid 13](#_Toc160663225)

[Figuur 11 Boxplot vertrek vertraging 14](#_Toc160663226)

[Figuur 12 Boxplot vertrek vertraging gegroepeerd op luchtvaartmaatschappij 15](#_Toc160663227)

[Figuur 13 Boxplot aankomst vertraging 16](#_Toc160663228)

[Figuur 14 Boxplot aankomst vertraging gegroepeerd op luchtvaartmaatschappij 17](#_Toc160663229)

[Figuur 15 Boxplot temperatuur 18](#_Toc160663230)

[Figuur 16 Boxplot temperatuur gegroepeerd op uur 19](#_Toc160663231)

[Figuur 17 Boxplot dauwpunt 20](#_Toc160663232)

[Figuur 18 Boxplot dauwpunt gegroepeerd op uur 21](#_Toc160663233)

[Figuur 19 Boxplot luchtvochtigheid 22](#_Toc160663234)

[Figuur 20 Boxplot luchtvochtigheid gegroepeerd op uur 23](#_Toc160663235)

[Figuur 21 Boxplot wind richting 24](#_Toc160663236)

[Figuur 22 Boxplot wind snelheid 25](#_Toc160663237)

[Figuur 23 Boxplot Anova test temperatuur per maand 28](#_Toc160663238)

[Figuur 24 Pairplot 29](#_Toc160663239)

[Figuur 25 Scatter plot vertrek vertraging en aankomst vertraging 31](#_Toc160663240)

[Figuur 26 Scatter plot wind richting en vertrek vertraging 32](#_Toc160663241)

[Figuur 27 Scatter plot wind snelheid en vertrek vertraging 33](#_Toc160663242)

[Figuur 28 Scatter plot temperatuur and dauw punt 34](#_Toc160663243)

[Figuur 29 Scatter plot temperatuur and luchtvochtigheid 35](#_Toc160663244)

[Figuur 30 Scatter plot dauwpunt en luchtvochtigheid 36](#_Toc160663245)

# Bibliografie

Gareth James, D. W. (2009). *An introduction to statistical learning with application in R.* Springer.

HAN. (sd). *PP1\_Desc\_Stat\_2024.pptx.* Opgehaald van Onderwijsonline: https://han.onderwijsonline.nl/elearning/lessonfile/VNzA5R1q/eyJpdiI6InoyaHlCbWRMZk15ZDRhTDVPeTJtbnc9PSIsInZhbHVlIjoiSXhOMTNLeDN1TGdBWXprUjBQakF0U3c3ZEg3eDMvT1NJbURpS0dubGY1UT0iLCJtYWMiOiI0ZmE3MDdlOTc0NjMwMzk3NjIzYjA0M2JmY2ZjMGU4MDlhOWI1NWNkZjFkMGMzZTQyND

HAN. (sd). *PP2\_Inf\_Stat\_CI\_2024.pptx.* Opgehaald van Onderwijsonline: https://han.onderwijsonline.nl/elearning/lessonfile/VNzA5R1q/eyJpdiI6IkltRmhVeEdSVW9kMXQzTGxyOWY1N1E9PSIsInZhbHVlIjoidXFmN1hlUldqb1MyMlErYTVRei9wVk5nUzBlNXcvMnloVk5zdU1HaXNyZ0kwZjZnRmNiRFlWbjhyQks5ZWQ5dCIsIm1hYyI6Ijg5ZTRmMmI4MTUxZDlhNTAyOGY2NTc3MTQ5OTdkND

HAN. (sd). *PP3\_Inf\_Stat\_Pval\_2024.pptx.* Opgehaald van Onderwijsonline: https://han.onderwijsonline.nl/elearning/lessonfile/VNzA5R1q/eyJpdiI6IjBXNWFWeFQ5R1NlcEpmUHI4RHZ4T0E9PSIsInZhbHVlIjoiU0VJR1FycjZrd3Y2dy9RY0RHNi9IUTBLWGk1NnhVd2lDS1pkb28vSGliRm4yNGdkTHFtRkdsLzJUOEFGK1dZSyIsIm1hYyI6IjIzZmIzNzhhYTdhZDYzNzcyNmE1ZDM1NTllZTFiYT

HAN. (sd). *PP4\_Inf\_Stat\_CorrRegr\_2024.pptx.* Opgehaald van Onderwijsonline: https://han.onderwijsonline.nl/elearning/lessonfile/VNzA5R1q/eyJpdiI6IkorR3o2UDdqUnptbE54TzVJN2poRFE9PSIsInZhbHVlIjoiRGRqeDUwTm56K2ZwaURVL0JtQmphZTlMYUQ1SGx4cVB5V2RmbkhtbVdzeGVIb2RIUGdXVHB0aXJNb0ZEUUN4ayIsIm1hYyI6Ijc2YThmYWIyNDQ2ZTQyZDVjMDk2ODVlZWEyYmE0MD

HAN. (sd). *PP5\_Inf\_Stat\_MultComp\_2024.pptx.* Opgehaald van Onderwijsonline: https://han.onderwijsonline.nl/elearning/lessonfile/VNzA5R1q/eyJpdiI6Ik5Xc2NSNnozdmp4enlFdzIraGdJV1E9PSIsInZhbHVlIjoiUS9LQ1U0Z3pNejBpNXkyMm9NemxRN05PVXlYV254NTJKTEJNVU92aTN0MmR3a1BWQkx2bzJjN0Y5WDVqSktTTSIsIm1hYyI6ImVhMzY5ZjVkOTMzODQwNmUyZmIxZDkzMWU3M2E0MD

HAN. (sd). *PPT2\_slide\_12.ipynb.* Opgehaald van Onderwijsonline: https://han.onderwijsonline.nl/elearning/lessonfile/VNzA5R1q/eyJpdiI6IlZVL2FaenBKSjNpb1czKzFSTTNrSFE9PSIsInZhbHVlIjoiR1lRZmJleitTV1AwYTd5OEsvdHhEWUZrVXliWmxJV1pRTE1qWkd5bDVYRT0iLCJtYWMiOiJjOGQ1YTA5ODFiZTE4MWZiYjg4ZTk1OWZkMTY1ZTdhZjdmNWEyZTFjNGE4ZDY0ZjVjNm

HAN. (sd). *PPT2\_slide\_16.ipynb.* Opgehaald van Onderwijsonline: https://han.onderwijsonline.nl/elearning/lessonfile/VNzA5R1q/eyJpdiI6Imtwc1BTWHZsRExRRCtBLzBJSDBwbGc9PSIsInZhbHVlIjoiWmlZdU1qWnZtc2gzS0cwWmFIa2FHUHE4K2FTQ1hQeHowT24vK1hHVHlSST0iLCJtYWMiOiIwZTY0YTE3YzFiNmFhNzZkYTE2YjcxMThkNjZmZmIxMmIwZGE2MmQ2YzUzYjNlMmUyYW

HAN. (sd). *PPT2\_slide\_25\_26\_27.ipynb.* Opgehaald van Onderwijsonline: https://han.onderwijsonline.nl/elearning/lessonfile/VNzA5R1q/eyJpdiI6IlJwWXlicFAyK21tVmNmMmJaamtnSFE9PSIsInZhbHVlIjoiUzFnZHRjNnJ2Z0dzSUVOMHlyY29iTXE4N29BVmJkNmZ5S2JULytWRUZOK2txbk5ObFNNN3JIVHUvNW55T0NueCIsIm1hYyI6IjE0NzZmMDU0NzZhNTY1NGI3ZTllOWZkZjZhZjNhMz

John D. Kelleher, B. M. (2015). *Fundamentals of machine learning for predictive data analytics.* The MIT Press.

Merkus, J. (2023, 02 09). *Boxplot maken en aflezen | Stappenplan & Voorbeelden*. Opgehaald van Scribbr: https://www.scribbr.nl/statistiek/boxplot/

Motulsky, H. (2018). *Intuitive biostatistics A nonmathemetical guide to statistical thinking.* New York: Oxford university press.

n.a. (sd). *QuickCalcs*. Opgehaald van Graphpad: https://www.graphpad.com/quickcalcs/

Scipy. (sd). *Docs*. Opgehaald van Scipy: https://docs.scipy.org/doc/scipy/

Scipy. (sd). *scipy.stats.f\_oneway*. Opgehaald van Scipy docs: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.f\_oneway.html

Scipy. (sd). *scipy.stats.tukey\_hsd*. Opgehaald van Scipy docs: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.tukey\_hsd.html

team, C. (sd). *Histogram*. Opgehaald van CFI: https://corporatefinanceinstitute.com/resources/excel/histogram/