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](#_Toc160825202)

[Informatievragen 3](#_Toc160825203)

[Descriptieve statistische kentallen 4](#_Toc160825204)

[Continue features 4](#_Toc160825205)

[Categoriale features 5](#_Toc160825206)

[Descriptieve statistieken beschrijving 5](#_Toc160825207)

[Histogrammen 6](#_Toc160825208)

[Temperatuur 6](#_Toc160825209)

[Dauwpunt 8](#_Toc160825210)

[Luchtvochtigheid 10](#_Toc160825211)

[Windrichting 12](#_Toc160825212)

[Windsnelheid 14](#_Toc160825213)

[Boxplots 16](#_Toc160825214)

[Vertrek vertraging 16](#_Toc160825215)

[Aankomst vertraging 18](#_Toc160825216)

[Temperatuur 20](#_Toc160825217)

[Dauwpunt 22](#_Toc160825218)

[Luchtvochtigheid 24](#_Toc160825219)

[Windrichting 26](#_Toc160825220)

[Windsnelheid 27](#_Toc160825221)

[Testen 28](#_Toc160825222)

[Proportietesten 28](#_Toc160825223)

[Difference of means testen 28](#_Toc160825224)

[Anova testen 29](#_Toc160825225)

[Scatter plots 31](#_Toc160825226)

[Pair plot 31](#_Toc160825227)

[Correlatiecoëfficiënten 32](#_Toc160825228)

[Vertrek vertraging en aankomst vertraging 33](#_Toc160825229)

[Windrichting en vertrek vertraging 34](#_Toc160825230)

[Windsnelheid en vertrek vertraging 35](#_Toc160825231)

[Windsnelheid en windrichting 36](#_Toc160825232)

[Temperatuur en dauwpunt 37](#_Toc160825233)

[Temperatuur en luchtvochtigheid 38](#_Toc160825234)

[Dauwpunt en luchtvochtigheid 39](#_Toc160825235)

[Temperatuur en windsnelheid 40](#_Toc160825236)

[Temperatuur en windrichting 41](#_Toc160825237)

[Conclusie 42](#_Toc160825238)

[Figurenlijst 44](#_Toc160825239)

[Bibliografie 45](#_Toc160825240)

# 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 windsnelheid en windrichting effect op vertragingen?

# Descriptieve statistische kentallen

Om te beginnen is het zaak om met gebruik van descriptive statistics een beeld te schetsen van de data die we voor ons hebben.

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 (Ratio) | 336776 | 2.451 | 527 | -43.0 | -5.0 | 12.639 | -2.0 | 11.0 | 1301.0 | 40.210 |
| Arr\_delay (Ratio) | 336776 | 2.800 | 577 | -86.0 | -17.0 | 6.895 | -5.0 | 14.0 | 1272.0 | 44.633 |
| Temp (Interval) | 336776 | 0.467 | 168 | 10.94 | 42.08 | 56.996 | 57.2 | 71.96 | 100.04 | 17.965 |
| Depw (Interval) | 336776 | 0.467 | 147 | -9.94 | 26.06 | 41.631 | 42.8 | 57.92 | 78.08 | 19.353 |
| Humid (Ratio) | 336776 | 0.467 | 2442 | 12.74 | 43.99 | 59.557 | 57.73 | 75.33 | 100.0 | 19.656 |
| Wind\_dir (Ratio) | 336776 | 2.909 | 37 | 0.0 | 130.0 | 201.540 | 220.0 | 290.0 | 360.0 | 104.820 |
| Wind\_speed (Ratio) | 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 (nominaal) | 336776 | 0.0 | 105 | ORD | 17283 | 5.132 | ATL | 17215 | 5.112 |
| Origin (nominaal) | 336776 | 0.0 | 3 | EWR | 120835 | 35.880 | JFK | 111279 | 33.042 |
| Carrier (nominaal) | 336776 | 0.0 | 16 | UA | 58665 | 17.420 | B6 | 54635 | 16.223 |

## Descriptieve statistieken beschrijving

Wanneer we naar de descriptive statistics kijken van de gekozen features zien we over het algemeen realistische data. Alhoewel het gevaarlijk is om al te grote aannames te maken zonder diep op de cijfers in te gaan zien we vooral bij de data rondom weersomstandigheden cijfers die overeenkomen met wat in werkelijkheid mogelijk is. Een ander opvallend punt is dat het minimum en het maximum van vluchtvertragingen bijzonder ver uit elkaar ligt. Wat ook opvalt is dat bij de origin van vluchten de mode frequentie van de eerste twee airports vele malen hoger ligt dan bij de destination airports. Verder is de windrichting een interessante feature om de kengetallen van in te zien. We zien bijvoorbeeld dat de gemiddelde richting naar het zuidwesten wijst.

Enkel de descriptive statistics zijn niet genoeg om een definitieve conclusie te trekken over de bruikbaarheid van de dataset. Daarom zal in het volgende hoofdstuk dieper in worden gegaan op features met gebruik van diagrammen.

# 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

Het histogram van de temperatuur laat een bimodale distributie zien, wat betekent dat er twee duidelijke pieken zijn in de frequentie van temperatuurwaarden. De pieken liggen rond de 40 en 70 graden Fahrenheit. Dit suggereert dat er twee dominante temperatuur bereiken zijn die het meest voorkomen in de dataset en het density plot bevestigt dit.

Deze informatie kan relevant zijn voor de tweede vraag over het voorspellen van andere weersomstandigheden op basis van temperatuur. Je zou kunnen onderzoeken of bepaalde temperatuur bereiken correleren met specifieke weersomstandigheden. Dit kan helpen bij het identificeren van patronen in het weer en het voorspellen van weersomstandigheden op basis van temperatuur.

### Histogram

Afbeelding met diagram, Perceel, lijn, schermopname

Automatisch gegenereerde beschrijving

Figuur 1 Histogram temperatuur

### Density plot

Afbeelding met tekst, diagram, Perceel, schermopname

Automatisch gegenereerde beschrijving

Figuur 2 Density plot temperatuur

## Dauwpunt

Het density plot van het dauwpunt toont twee pieken rond 30 en 60 graden Fahrenheit, wat suggereert dat de verdeling van het dauwpunt bimodaal is. Het moeilijk om een eenduidige conclusie te trekken op basis van alleen de diagrammen, de bewijzen zijn in dit geval niet extreem duidelijk.

Het gebrek aan informatie uit de diagrammen kan te verklaren zijn door verschillende factoren, zoals de complexiteit van de relatie tussen het dauwpunt en andere variabelen, of mogelijke externe invloeden die niet in de dataset zijn opgenomen.

Om meer informatie te verkrijgen rondom deze feature kan het nuttig zijn om aanvullende analyses uit te voeren, zoals correlatieanalyses tussen het dauwpunt en andere weersomstandigheden, of het onderzoeken van seizoen patronen in het dauwpunt.

In het kader van de vraag over het effect van weersomstandigheden op vertragingen, zou je kunnen proberen te onderzoeken of bepaalde dauwpunt bereiken correleren met verhoogde vertragingen, en of deze correlaties consistent zijn bij verschillende seizoenen of bepaalde gebieden.

### Histogram

Afbeelding met diagram, Perceel, schermopname, lijn

Automatisch gegenereerde beschrijving8-3-2024

Figuur 3 Histogram dauwpunt

### Density plot

Afbeelding met tekst, diagram, Perceel, schermopname

Automatisch gegenereerde beschrijving

Figuur 4 Density plot dauwpunt

## Luchtvochtigheid

De luchtvochtigheidsverdeling lijkt rechts-scheef te zijn, met enkele zeer hoge waarden die het gemiddelde verhogen. Deze extreme waarden zijn zichtbaar in het histogram en tot op zekere hoogte in het density plot. Dit kan wijzen op uitschieters of extreme omstandigheden in de dataset, wat relevant kan zijn bij het analyseren van het effect van luchtvochtigheid op vluchtvertragingen.

### 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

De beschrijving van de windrichting suggereert een histogram met een bimodale verdeling, waarbij twee pieken zichtbaar zijn in het density plot. De windrichting wordt echter gemeten in graden, waarbij 0 en 360 in elkaar overlopen. Wat lijkt op een linker piek is eigenlijk een daling van de rechter piek, waardoor we te maken hebben met een gaussverdeling.

Het begrijpen van deze cijfers is belangrijk bij het begrijpen van de patronen van windrichting en kunnen relevant zijn voor het onderzoeken van het effect van windrichting op vluchtvertragingen. Het is belangrijk op te merken dat het effect van windrichting op vertragingen mogelijk complex zijn en afhankelijk kunnen zijn van andere factoren zoals windsnelheid.

### Histogram

Afbeelding met diagram, Perceel, lijn, schermopname

Automatisch gegenereerde beschrijving

Figuur 7 Histogram windrichting

### Density plot

Afbeelding met tekst, diagram, Perceel, schermopname

Automatisch gegenereerde beschrijving

Figuur 8 Density plot windrichting

## Windsnelheid

Het histogram van de windsnelheid lijkt op een gaussverdeling, waarbij de meeste waarden zich dicht bij het gemiddelde bevinden. Hoewel er een uitschieter aan de linker kant te zien is, liggen de meeste waarden toch dicht bij het gemiddelde zonder al te veel extreme waarden. De gemiddelde en mediaan van de windsnelheid zitten dicht bij elkaar (11.114 & 10.357), wat aangeeft dat de verdeling symmetrisch is.

Om deze beschrijving te verbinden met je vraag over het effect van windsnelheid en windrichting op vertragingen, zou je kunnen onderzoeken of er een correlatie bestaat tussen hoge windsnelheden en verhoogde vertragingen in aankomst- of vertrektijden. Je zou ook kunnen analyseren of bepaalde windrichtingen vaker leiden tot vertragingen, bijvoorbeeld als vliegtuigen moeite hebben om tegen de wind in te vliegen of als er zijwinden zijn die de landingsprocedures beïnvloeden.

### Histogram

Afbeelding met diagram, Perceel, lijn, schermopname

Automatisch gegenereerde beschrijving

Figuur 9 Histogram windsnelheid

### Density plot

Afbeelding met tekst, diagram, Perceel, schermopname

Automatisch gegenereerde beschrijving

Figuur 10 Density plot windsnelheid

# Boxplots

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

## Vertrek vertraging

De boxplot van vertrekvertragingen toont een relatief kleine spreiding tussen de minimale en maximale waarden, evenals tussen het eerste en derde kwartiel, wat suggereert dat de meerderheid van de vluchten binnen een redelijke tijd vertrekken. Echter, de aanwezigheid van veel outliers wijst op vertragingen die aanzienlijk hoger zijn dan de gebruikelijke vertrektijden. In dit geval zijn de outliers dus geen punten om te verwerpen, ze zijn juist interessant.

Het effect van vertrekvertragingen op aankomstvertragingen kan worden onderzocht door te analyseren of vluchten met hoge vertrekvertragingen ook meer kans hebben op aankomstvertragingen.

Afbeelding met lijn, tekst, Perceel, diagram

Automatisch gegenereerde beschrijving

Figuur 11 Boxplot vertrek vertraging

Het diagram met de boxplots van vertrekvertragingen per vliegtuigmaatschappij laat zien hoe verschillend ze presteren als het gaat om op tijd vertrekken. Je ziet dat sommige maatschappijen zoals EV en YV eruit springen met een opvallend groot derde kwartiel. De grote derde kwartielen bij EV en YV geven aan dat deze maatschappijen vaak te maken hebben met vertragingen bij het vertrekken. De spreiding tussen de verschillende maatschappijen qua outliers loopt ook ver uiteen.

Het zou interessant zijn om te onderzoeken of er een causaal verband bestaat tussen vertrekvertragingen en aankomstvertragingen, en zo ja, in welke mate dit verband verschilt tussen verschillende maatschappijen.

Afbeelding met tekst, diagram, Perceel, lijn

Automatisch gegenereerde beschrijving

Figuur 12 Boxplot vertrek vertraging gegroepeerd op luchtvaartmaatschappij

## Aankomst vertraging

Ten eerste valt op dat de spreiding van aankomstvertragingen iets groter is dan die van vertrekvertragingen. Dit verschil kan worden verklaard met de mogelijkheid dat vluchten zowel vroeger als later dan gepland kunnen aankomen, terwijl vertrektijden in principe niet eerder kunnen plaatsvinden dan gepland. Bovendien tonen de outliers in de boxplot een vergelijkbare spreiding als die van de vertrekvertragingen, wat zou kunnen suggereren dat vluchten met vertragingen bij vertrek vaak ook vergelijkbare vertragingen ervaren bij aankomst.

Afbeelding met tekst, lijn, Perceel, diagram

Automatisch gegenereerde beschrijving

Figuur 13 Boxplot aankomst vertraging

De boxplot van vertraging bij aankomst per maatschappij vertoont vergelijkbare patronen als die van vertrekvertragingen. De grotere spreiding van het minimum en het maximum van de boxplot duidt op vluchten die vroeger dan gepland aankomen en dat het met meer regelmaat voorkomt dat een vlucht wat later aankomt dan gepland. Er lijkt uit deze diagrammen een consistentie te zijn tussen vertrek- en aankomstvertragingen.

Afbeelding met tekst, diagram, Perceel, nummer

Automatisch gegenereerde beschrijving

Figuur 14 Boxplot aankomst vertraging gegroepeerd op luchtvaartmaatschappij

## Temperatuur

Het lijkt erop dat de temperatuur over het algemeen binnen een beperkt bereik blijft, zoals aangegeven door de kleine spreiding tussen het 1ste en 3de kwartiel. Het ontbreken van outliers suggereert dat er geen extreme temperatuurwaarden zijn buiten het gebruikelijke bereik.

Deze informatie kan relevant zijn bij het beantwoorden van de vraag over het voorspellen van andere weersomstandigheden op basis van temperatuur, aangezien het aangeeft hoe de temperatuur is verdeeld en welke variabiliteit er is.

Afbeelding met schermopname, Rechthoek, tekst, plein

Automatisch gegenereerde beschrijving

Figuur 15 Boxplot temperatuur

Het lijkt erop dat de temperatuur een duidelijk patroon laat zien gedurende de dag, waarbij het oploopt tot een hoogtepunt rond de middaguren en vervolgens weer afneemt richting de avond. Dit patroon is goed te verklaren en is in overeenstemming met de verwachtingen van dagelijkse temperatuurschommelingen.

Het feit dat er alleen informatie beschikbaar is van 5u tot 23u suggereert dat de temperatuur alleen gedurende deze uren wordt gemeten.

Afbeelding met tekst, diagram, lijn, Perceel

Automatisch gegenereerde beschrijving

Figuur 16 Boxplot temperatuur gegroepeerd op uur

## Dauwpunt

Voor dit boxplot geldt eigenlijk vrijwel hetzelfde als voor die van de temperatuur. We zien een gemiddelde net boven de 40, we zien dat de spreiding onder het gemiddelde wat groter is dan erboven.

Afbeelding met schermopname, diagram, Rechthoek, plein

Automatisch gegenereerde beschrijving

Figuur 17 Boxplot dauwpunt

Het lijkt erop dat het dauwpunt gedurende verschillende uren van de dag binnen een relatief consistent bereik blijft, zoals aangegeven door de boxplots. Het feit dat de delta’s tussen het 1ste en 4de kwartiel redelijk gelijk zijn, suggereert dat er geen grote variaties zijn in het dauwpunt gedurende de dag.

Afbeelding met tekst, lijn, diagram, Rechthoek

Automatisch gegenereerde beschrijving

Figuur 18 Boxplot dauwpunt gegroepeerd op uur

## Luchtvochtigheid

Wederom zien we een boxplot vergelijkbaar met de temperatuur. Dit is goed te verklaren, aangezien hogere temperaturen vaak samengaan met hogere luchtvochtigheid. Deze observatie kan van belang zijn bij het voorspellen van andere weersomstandigheden op basis van temperatuur.

Afbeelding met diagram, schermopname, Rechthoek, lijn

Automatisch gegenereerde beschrijving

Figuur 19 Boxplot luchtvochtigheid

De boxplot van luchtvochtigheid, gegroepeerd per uur, toont een opvallend tegenovergesteld patroon ten opzichte van de boxplot van temperatuur. Wat is dan nu precies het verband tussen deze features? Aangezien luchtvochtigheid en temperatuur doorgaans hand in hand gaan. Het afnemende gemiddelde van luchtvochtigheid gedurende de dag, met een dieptepunt rond de middaguren en een latere stijging, vraagt om verdere analyse. Deze observatie benadrukt de complexiteit van de relatie tussen meteorologische variabelen en suggereert dat er mogelijk meer interacties zijn die van invloed zijn op lokale weerspatronen. Deze observatie kan als basis worden gebruikt voor verdere analyse om meer inzicht te krijgen in de relatie tussen temperatuur, luchtvochtigheid en andere weerkundige statistieken.

Afbeelding met tekst, diagram, lijn, Perceel

Automatisch gegenereerde beschrijving

Figuur 20 Boxplot luchtvochtigheid gegroepeerd op uur

## Windrichting

De boxplot van de windrichting toont geen uitschieters en laat zien dat de wind het vaakst in de richting van ongeveer 200-250 graden waait.

Afbeelding met schermopname, tekst, diagram, Rechthoek

Automatisch gegenereerde beschrijving

Figuur 21 Boxplot windrichting

## Windsnelheid

Bij de boxplot van de windsnelheid valt op dat het gemiddelde rond de 10 ligt, met een iets grotere spreiding tussen het 3de kwartiel en de maximumwaarde dan tussen het minimum en het 1ste kwartiel. Er zijn enkele uitschieters aan de rechterkant vanaf ongeveer 27 en hoger, wat aangeeft dat deze hoge windsnelheden gering voorkomen, mogelijk bij stormachtige omstandigheden. Het is essentieel om deze outliers mee te nemen in de analyse, aangezien ze met uitstek invloed zouden kunnen hebben op vertragingen bij vluchten.

Afbeelding met diagram, lijn, schermopname, Rechthoek

Automatisch gegenereerde beschrijving

Figuur 22 Boxplot windsnelheid

# Testen

## Proportietesten

Voor de proportietesten worden betrouwbaarheidsintervallen berekend voor de vertraging van vluchten. Eerst wordt gekeken naar alle vluchten, waarna specifieke luchtvaartmaatschappijen apart worden behandeld, waarvoor ook betrouwbaarheidsintervallen worden berekend. Bij elke berekening wordt gebruik gemaakt van de normale verdeling vanwege de omvang van de dataset, met een alpha van 0.05, tenzij anders aangegeven.

### Alle vluchten

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

Dit betekent dat we met 95% zekerheid kunnen zeggen dat er een kans op vertraging is tussen 37.97% en 38.30% voor vertrek, en tussen 39.33% en 39.66% voor aankomst. We kunnen dus met een hoge mate van zekerheid aangeven waar de werkelijke proportie waarschijnlijk ligt.

### DL

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

Dit betekent dat we met 95% zekerheid kunnen zeggen dat er een kans op vertraging is tussen 31.26% en 32.10% voor vertrek, en tussen 33.69% en 34.54% voor aankomst bij deze luchtvaartmaatschappij. Hieruit kunnen we afleiden waar de werkelijke proportie waarschijnlijk ligt met een hoge mate van zekerheid.

### HA

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

Dit betekent dat we met 95% zekerheid kunnen zeggen dat er een kans op vertraging is tussen 15.92% en 24.43% voor vertrek, en tussen 23.59% en 33.14% voor aankomst bij deze luchtvaartmaatschappij. Hoewel we minder zekerheid hebben over de exacte proportie, kunnen we nog steeds met een redelijke mate van zekerheid aangeven waar deze waarschijnlijk ligt.

### EV

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

Dit betekent dat we met 95% zekerheid kunnen zeggen dat er een kans op vertraging is tussen 42.30% en 43.13% voor vertrek, en tussen 44.78% en 45.62% voor aankomst bij deze luchtvaartmaatschappij. Hieruit kunnen we met een hoge mate van zekerheid afleiden waar de werkelijke proportie waarschijnlijk ligt.

## Difference of means testen

Voor de difference of means testen wordt bij verschillende luchtvaartmaatschappijen gekeken naar de verschillen in het gemiddelde van vertragingen zowel bij vertrek als aankomst. De nulhypothese mag alleen worden verworpen 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 EV en F9 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 zijn berekend. We verwachten hier een duidelijk verschil in temperatuur per maand te zien.

* 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 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 precies wat we verwachten te zien. Hiermee kunnen we concluderen dat er een significant 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 een aantal feature-to-feature scatter plots die van toepassing zijn op de [informatievragen](#_Informatievragen).

## Pair plot

Afbeelding met tekst, kaart, patroon

Automatisch gegenereerde beschrijving

Figuur 24 Pairplot

## Correlatiecoëfficiënten

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

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 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 de onderstaande scatter plot zien we een duidelijke correlatie tussen vertragingen bij vertrek en vertragingen bij aankomst. Naarmate de vertraging bij vertrek toeneemt, neemt ook de vertraging bij aankomst toe. Dit is ook goed te zien in de [correlatiecoëfficiënten tabel,](#_Correlatie_coëfficiënten) waar de correlatiecoëfficiënt tussen dep\_delay en arr\_delay 0.9148 is. Deze waarde ligt bijzonder dicht bij 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 aankomstvertragingen 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

## Windrichting en vertrek vertraging

Als we naar de onderstaande scatter plot kijken, zien we geen duidelijke correlatie tussen de windrichting en de vertrekvertraging. Als we ook naar de correlatiecoëfficiënt kijken, zien we een waarde van -0.0176. Dit wijst op een extreem lichte negatieve correlatie, maar de waarde ligt te dicht bij nul om te concluderen dat er een significante correlatie is tussen windrichting en vertrekvertraging.

Hierdoor kunnen we ook niet concluderen dat windrichting een effect heeft op de vertrekvertraging, althans niet op basis van deze scatter plot.

Afbeelding met tekst, schermopname, diagram, lijn

Automatisch gegenereerde beschrijving

Figuur 26 Scatter plot windrichting en vertrek vertraging

## Windsnelheid en vertrek vertraging

Als we naar de onderstaande scatter plot kijken, zien we geen duidelijke correlatie tussen windsnelheid en vertrekvertragingen. De correlatiecoëfficiënt is 0.0474, wat aangeeft dat er nauwelijks een correlatie is tussen de windsnelheid en vertrekvertragingen. Opmerkelijk is dat het niet lijkt uit te maken of de windsnelheid hoog is, aangezien we ook bij hoge windsnelheden geen significant langere vertragingen waarnemen dan bij lage windsnelheden. Het tegenovergestelde is zelfs waar, aangezien we een aantal aanzienlijke vertragingen zien bij vrij lage windsnelheden. We zouden kunnen veronderstellen dat hoge windsnelheden van korte duur zijn en daarom de vertraging relatief kort is, maar deze conclusie kan niet rechtstreeks uit de scatter plot worden afgeleid.

Hierdoor kunnen we niet concluderen of de windsnelheid een effect heeft op de vertrekvertragingen, althans niet op basis van deze scatter plot.

Afbeelding met tekst, schermopname, diagram, lijn

Automatisch gegenereerde beschrijving

Figuur 27 Scatter plot windsnelheid en vertrek vertraging

## Windsnelheid en windrichting

Als we naar onderstaande scatter plot kijken, zien we een zwakke correlatie tussen windrichting en windsnelheid, zit zou betekenen dat er wel een verband kan zijn tussen de windsnelheid en de windrichting maar dat dit zeker niet altijd het geval is. Een correlatiecoëfficiënt van 0.3416 laat ook zien dat er sprake is van een zwakke positieve correlatie tussen de beide features.

We kunnen wel de aanname maken dat de wind uit bepaalde richtingen een hogere snelheid kan hebben, dit zien we in de realiteit ook vaak zat gebeuren namelijk, echter kunnen we dit niet concluderen. We zien wel uitschieters rond de 270 en 50 graden maar ook in het omliggende gebied en rond de 180 graden vinden hoge windsnelheden plaats, dit is een heel groot bereik waardoor het moeilijk is om te zeggen of windsnelheid en windrichting elkaar echt sterk beïnvloeden.

Afbeelding met tekst, schermopname, diagram, lijn

Automatisch gegenereerde beschrijving

Figuur 28 Scatter plot windsnelheid en windrichting

## Temperatuur en dauwpunt

In de 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 waar te nemen. Er is meer variatie in de datapunten, wat te zien is aan de grotere spreiding en een zichtbaar lineair verband. Als we naar de [correlatiecoëfficiënten tabel](#_Correlatie_coëfficiënten) kijken, 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 de temperatuur het dauwpunt kunnen voorspellen. Door de hoge correlatie tussen beide waarden kunnen we redelijkerwijs een goede voorspelling doen met betrekking tot het dauwpunt aan de hand van de temperatuur.

Afbeelding met tekst, schermopname, diagram, lijn

Automatisch gegenereerde beschrijving

Figuur 29 Scatter plot temperatuur en dauw punt

## Temperatuur en luchtvochtigheid

In het onderstaande scatter plot zien we geen correlatie tussen temperatuur en luchtvochtigheid. De datapunten zijn erg verspreid en er is geen duidelijk verband zichtbaar in het diagram. We observeren geen positieve noch negatieve correlatie tussen beide variabelen. Wanneer we naar de [correlatiecoëfficiënten tabel](#_Correlatie_coëfficiënten) kijken, zien we een waarde van 0.0355. Dit getal ligt zeer dicht bij nul, wat bevestigt dat er geen significante correlatie bestaat tussen temperatuur en luchtvochtigheid.

Hieruit kunnen we concluderen dat we niet in staat zijn om de luchtvochtigheid te voorspellen aan de hand van de temperatuur. De correlatie is te laag, of eigenlijk nauwelijks aanwezig, waardoor het moeilijk is om de luchtvochtigheid te voorspellen op basis van de temperatuur.

Afbeelding met tekst, schermopname, diagram, patroon

Automatisch gegenereerde beschrijving

Figuur 30 Scatter plot temperatuur en luchtvochtigheid

## Dauwpunt en luchtvochtigheid

In het onderstaande scatter plot zien we een matige tot sterke correlatie tussen het dauwpunt en de luchtvochtigheid. Er lijkt enigszins een lineair verband duidelijk te worden uit het diagram, hoewel de spreiding groot is. Als we kijken naar de correlatiecoëfficiënt in de tabel, zien we een correlatiecoëfficiënt van 0.4922 tussen dauwpunt en luchtvochtigheid. Dit toont aan dat er een matige correlatie bestaat tussen de features, die zich vrijwel precies tussen 0 en 1 bevindt. Het is moeilijk te zeggen of beide features een directe invloed hebben op elkaar, maar er is wel een correlatie waar te nemen tussen dauwpunt en luchtvochtigheid.

Omdat we aan de hand van de temperatuur het dauwpunt kunnen voorspellen, en er toch een sterke positieve correlatie bestaat tussen dauwpunt en luchtvochtigheid, zouden we met behulp van de temperatuur via het dauwpunt nog steeds een redelijk betrouwbare voorspelling kunnen doen met betrekking tot de luchtvochtigheid. Wel moeten we oppassen, omdat we dan een voorspelling zouden doen op basis van een andere voorspelling, wat de betrouwbaarheid enigszins kan verminderen. Echter, dit is nog steeds betrouwbaarder dan het voorspellen van luchtvochtigheid op basis van alleen de temperatuur.

Afbeelding met tekst, schermopname, diagram, kaart

Automatisch gegenereerde beschrijving

Figuur 33 Scatter plot dauwpunt en luchtvochtigheid

## Temperatuur en windsnelheid

Als we naar onderstaande scatter plot kijken, dan zien we een lichte negatieve correlatie tussen temperatuur en windsnelheid, dit wijst erop dat ze elkaar niet beïnvloeden. Ook de correlatiecoëfficiënt van -0.1468 laat zien dat er een zeer matige negatieve correlatie is tussen de beide features wat er wederom op wijst dat ze elkaar niet beïnvloeden, alleen met zo’n lage correlatiecoëfficiënt kun je ook wel stellen dat er geen correlatie is.

Dit betekent dat het dus niet uit lijkt te maken wat de temperatuur of de windsnelheid is voor de windsnelheid of de temperatuur, respectievelijk. Dit betekent ook dat we aan de hand van de temperatuur geen voorspellingen kunnen maken over de windsnelheid.

Afbeelding met tekst, schermopname, Lettertype, diagram

Automatisch gegenereerde beschrijving

Figuur 31 Scatter plot temperatuur en windsnelheid

## Temperatuur en windrichting

Als we naar onderstaande scatter plot kijken, dan zien we dat er geen correlatie is tussen temperatuur en windrichting. Ook de correlatiecoëfficiënt van -0.0999 laat zien dat er geen correlatie is tussen beide features, hoewel de coëfficiënt wel onder de 0 zit en daarmee negatief is zit het zo dicht op de 0 dat er geen sprake is van een correlatie.

Dit is wel interessant omdat wij wel verwacht hadden dat met name wind die van het zuiden komt meer warme lucht mee zou nemen dan windt die van het noorden af komt maar als we naar de scatter plot kijken zien we tussen de 250 en 20 graden (0 en 360 graden liggen immers naast elkaar, ook al zijn ze ver van elkaar verwijderd op de scatter plot) een grote spreiding in temperaturen en tussen de 50 en 200 graden juist een veel kleinere spreiding van temperaturen.

De grote spreiding van temperatuur komt met wind vanuit het noorden tot en met het oosten en de kleine spreiding komt vanuit het noordwesten tot en met het zuiden. Dit is dan weer wel in lijn met onze verwachten in zekere zin omdat het betekend dat uit het zuiden meer warme lucht komt alleen zien we vanuit het noorden zowel warme als koude lucht komen, deze spreiding maakt het moeilijk om voorspellingen te doen aan de hand van de temperatuur op de windrichting of visa versa. Om hier een beter beeld van te krijgen zouden we ook kunnen kijken naar de temperatuur en windrichting per maand of seizoen, echter valt dit buiten wat zichtbaar is in deze scatter plot.

Afbeelding met tekst, schermopname, Lettertype, nummer

Automatisch gegenereerde beschrijving

Figuur 32 Scatter plot temperatuur en windrichting

# Conclusie

Op basis van de analyses lijkt de dataset bruikbaar om de gestelde informatievragen te beantwoorden. Er zijn duidelijke verbanden tussen verschillende features en de onderzochte aspecten van vluchtvertragingen en weersomstandigheden.

1. **Wat is het effect van vertrekvertragingen op aankomstvertragingen?**

Met de verzamelde informatie kunnen we concluderen dat er een significant effect is van vertrekvertragingen op aankomstvertragingen, zoals blijkt uit de correlatie tussen deze twee variabelen.

Uit het scatter plot blijkt een duidelijke correlatie tussen vertrekvertragingen en aankomstvertragingen. Naarmate de vertrekvertraging toeneemt, neemt ook de aankomstvertraging toe. Dit wordt ondersteund door de hoge correlatiecoëfficiënt van 0.9148 tussen de vertrek- en aankomstvertragingen. Deze waarde ligt dicht bij 1 wat duidt op een bijna perfecte positieve correlatie tussen de twee features.

Hieruit kunnen we concluderen dat het effect van vertrekvertragingen significant is op de aankomstvertragingen. Wanneer een vlucht vertraging heeft bij vertrek, is de kans groot dat deze ook vertraging zal hebben bij aankomst, en de mate van vertraging bij aankomst kan redelijkerwijs worden voorspeld aan de hand van de vertrekvertraging. Het is belangrijk om voorop te stellen dat dit enkel een correlatie is, een causaal verband kunnen we nog niet definitief aannemen.

1. **Kunnen we op basis van temperatuur andere weersomstandigheden voorspellen?**

Op basis van de informatie die is verzameld over de correlatie tussen temperatuur en andere weersomstandigheden, kunnen we concluderen dat temperatuur een goede voorspeller is voor bepaalde weersomstandigheden:

Er is een sterke positieve correlatie (correlatiecoëfficiënt van 0.8823) tussen temperatuur en dauwpunt. Dit suggereert dat we het dauwpunt kunnen voorspellen aan de hand van de temperatuur, gezien het sterke verband tussen de variabelen.

Hoewel er geen significante correlatie is tussen temperatuur en luchtvochtigheid (correlatiecoëfficiënt van 0.0355), kunnen we indirect via het dauwpunt redelijk betrouwbaar voorspellingen doen over de luchtvochtigheid. Dit komt door de sterke correlatie tussen dauwpunt en luchtvochtigheid en de correlatie tussen temperatuur en dauwpunt.

Aangezien temperatuur een sterke correlatie heeft met het dauwpunt en een matige tot sterke correlatie met luchtvochtigheid, kunnen we stellen dat temperatuur een belangrijke factor is bij het voorspellen van weersomstandigheden die afhankelijk zijn van dauwpunt en luchtvochtigheid. Dit blijft een aanname, pas na tests met modelering of voorspellingsanalyses kunnen we dit met zekerheid zeggen.

Als we dan kijken naar de temperatuur en windrichting en windsnelheid zien we eigenlijk geen correlatie, dit maakt het heel lastig om aan de hand van temperatuur de windrichting of windsnelheid te voorspellen.

1. **Hebben windsnelheid en windrichting effect op vertragingen?**

Met de informatie die we hebben kunnen we concluderen dat er geen duidelijk verband is tussen windrichting, windsnelheid en vertrekvertragingen, althans niet op basis van de gevonden scatter plots en correlatiecoëfficiënten.

**Effect van windrichting op vertrekvertragingen:**

Uit dit scatter plot blijkt dat er geen significante correlatie is tussen windrichting en vertrekvertragingen, zoals aangegeven door de lichte negatieve correlatiecoëfficiënt (0.0176). Hieruit kunnen we niet concluderen dat windrichting een significant effect heeft op vertrekvertragingen.

**Effect van windsnelheid op vertrekvertragingen:**

Dit scatter plot toont eveneens geen duidelijk verband tussen windsnelheid en vertrekvertragingen, met een correlatiecoëfficiënt van 0.0474. Er lijkt geen consistent patroon te zijn waarbij hogere windsnelheden leiden tot langere vertrekvertragingen. De aanwezigheid van aanzienlijke vertragingen bij zowel lage als hoge windsnelheden suggereert dat andere factoren een grotere rol spelen bij vertrekvertragingen dan alleen windsnelheid.

# Figurenlijst

[Figuur 1 Histogram temperatuur 6](#_Toc160825241)

[Figuur 2 Density plot temperatuur 7](#_Toc160825242)

[Figuur 3 Histogram dauwpunt 8](#_Toc160825243)

[Figuur 4 Density plot dauwpunt 9](#_Toc160825244)

[Figuur 5 Histogram luchtvochtigheid 10](#_Toc160825245)

[Figuur 6 Density plot luchtvochtigheid 11](#_Toc160825246)

[Figuur 7 Histogram windrichting 12](#_Toc160825247)

[Figuur 8 Density plot windrichting 13](#_Toc160825248)

[Figuur 9 Histogram windsnelheid 14](#_Toc160825249)

[Figuur 10 Density plot windsnelheid 15](#_Toc160825250)

[Figuur 11 Boxplot vertrek vertraging 16](#_Toc160825251)

[Figuur 12 Boxplot vertrek vertraging gegroepeerd op luchtvaartmaatschappij 17](#_Toc160825252)

[Figuur 13 Boxplot aankomst vertraging 18](#_Toc160825253)

[Figuur 14 Boxplot aankomst vertraging gegroepeerd op luchtvaartmaatschappij 19](#_Toc160825254)

[Figuur 15 Boxplot temperatuur 20](#_Toc160825255)

[Figuur 16 Boxplot temperatuur gegroepeerd op uur 21](#_Toc160825256)

[Figuur 17 Boxplot dauwpunt 22](#_Toc160825257)

[Figuur 18 Boxplot dauwpunt gegroepeerd op uur 23](#_Toc160825258)

[Figuur 19 Boxplot luchtvochtigheid 24](#_Toc160825259)

[Figuur 20 Boxplot luchtvochtigheid gegroepeerd op uur 25](#_Toc160825260)

[Figuur 21 Boxplot windrichting 26](#_Toc160825261)

[Figuur 22 Boxplot windsnelheid 27](#_Toc160825262)

[Figuur 23 Boxplot Anova test temperatuur per maand 30](#_Toc160825263)

[Figuur 24 Pairplot 31](#_Toc160825264)

[Figuur 25 Scatter plot vertrek vertraging en aankomst vertraging 33](#_Toc160825265)

[Figuur 26 Scatter plot windrichting en vertrek vertraging 34](#_Toc160825266)

[Figuur 27 Scatter plot windsnelheid en vertrek vertraging 35](#_Toc160825267)

[Figuur 28 Scatter plot windsnelheid en windrichting 36](#_Toc160825268)

[Figuur 29 Scatter plot temperatuur en dauw punt 37](#_Toc160825269)

[Figuur 30 Scatter plot temperatuur en luchtvochtigheid 38](#_Toc160825270)

[Figuur 33 Scatter plot dauwpunt en luchtvochtigheid 39](#_Toc160825271)

[Figuur 31 Scatter plot temperatuur en windsnelheid 40](#_Toc160825272)

[Figuur 32 Scatter plot temperatuur en windrichting 41](#_Toc160825273)

# 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/