Project Big Data, Business Intelligence and NoSQL   
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
Teamleden:

Thomas Billiet

Ishan Ameel

Sven Depickere

Jonas Anseel

Inhoud

[1. Preprocessing van de dataset 2](#_Toc513734454)

[1.1. Data loading 2](#_Toc513734455)

[1.2. Cleaning functions 3](#_Toc513734456)

[1.3. Convert military time to datetime 5](#_Toc513734457)

[1.4. Add time zones 6](#_Toc513734458)

[1.5. Change local time to UTC 7](#_Toc513734459)

[1.6. Calculate duration 8](#_Toc513734460)

[1.7. Calculate distance 9](#_Toc513734461)

[1.8. Calculate speed 10](#_Toc513734462)

[1.9. Remove too long and too short flights 11](#_Toc513734463)

[1.10. Execute cleaning 12](#_Toc513734464)

[2. Tableau 13](#_Toc513734465)

[3. Machine learning met Scikit learn 22](#_Toc513734466)

[4. NetworkX 23](#_Toc513734467)

# Preprocessing van de dataset

We hebben onze code in verschillende stappen opgesplitst namelijk:

* Data loading
* Cleaning functions
* Convert military time to datetime
* Add time zones
* Change local time to UTC
* Calculate duration
* Calculate distance
* Calculate speed
* Remove too long too short flights
* Execute cleaning

## Data loading

1. IMPORT\_PATH = '../Initial\_Data/Unzipped'
2. IMPORT\_ALL\_PATH = IMPORT\_PATH + '/\*'
3. import pandas as pd
4. import numpy as np
5. import glob as glob
6. def readAllFiles():

files = glob.glob(IMPORT\_ALL\_PATH)

frames = []

for file in files:

df = pd.read\_csv(file, index\_col = 0)

frames.append(df)

return pd.concat(frames)

def readOneFile(url):

return pd.read\_csv(url, index\_col = 0)

df = readAllFiles()

Hier in lezen we de data in als een panda frame. De csv file bevinden zich in een bepaalde direct die luisterd naar de naam */Initial\_Data/Unzipped*.

Het aantal records dat we in het begin hebben zijn er **11.401.196**

## Cleaning functions

1. import pandas as pd
2. import numpy as np
3. import glob as glob from math
4. import \*
5. usaStates = ["AL", "AK", "AZ", "AR", "CA", "CO", "CT", "DE", "FL", "GA", "HI", "ID", "IL", "IN", "IA", "KS", "KY", "LA", "ME", "MD", "MA", "MI", "MN", "MS", "MO", "MT", "NE", "NV",  "NH", "NJ", "NM", "NY", "NC", "ND", "OH", "OK", "OR", "PA", "RI", "SC", "SD", "TN", "TX",  "UT", "VT", "VA", "WA", "WV", "WI", "WY"]
6. def deleteWrongStates(df):

print("Aantal records:", len(df))

for el in df.departure\_state.unique():

if (el not in usaStates):

df = df.drop(df[df['departure\_state'] == el].index)

print("Aantal records na verwijderen foute vertrek staat:", len(df))

for el in df.arrival\_state.unique():

if (el not in usaStates):

df = df.drop(df[df['arrival\_state'] == el].index)

print("Aantal records na verwijderen foute aankomst staat:", len(df))

return df

def convertColumnTypes(df):

df.departure\_schedule = df.departure\_schedule.astype(int)

df.departure\_delay = df.departure\_delay.astype(float)

df.arrival\_schedule = df.arrival\_schedule.astype(int)

df.arrival\_delay = df.arrival\_delay.astype(float)

df.arrival\_actual = df.arrival\_actual.astype(int)

df.departure\_actual = df.departure\_actual.astype(int)

return df

def dropMoreAdvancedDuplicates(df):

df = df.groupby(['date', 'airline', 'airline\_code', 'departure\_airport',

 'departure\_state', 'departure\_lat', 'departure\_lon', 'departure\_schedule',

 'arrival\_airport', 'arrival\_state', 'arrival\_lat', 'arrival\_lon',

'arrival\_schedule']).mean().reset\_index()

df = convertColumnTypes(df)

if 'index' in df.columns:

df = df.drop(['index'], axis = 1) #remove old index

return df

The United States Of America telt 50 staten. In de data dat we gekregen hebben zijn er 3 staten die geen staat zijn van The United States Of America. Deze zullen we namelijk verwijderen.

Die staten zijn:

* TT (Trust Territory of the Pacific Islands):
  + bestaat niet meer sinds 1986
    - Bestaat nu uit:
      * Republic of the Marshall Islands
      * Federated States of Micronesia
      * Republic of Palau
      * Commenwealth of the Northern Mariana Islands
* PR (Puerto Rico):
  + is een territorium
* VI (U.S Virgin Islands) :
  + is een territorium

Er zijn **56 413** vluchten die vertrekken van uit staten die gen officiële staten zijn.

Er zijn **53 184** vluchten die landen in staten die geen officiële staten zijn.

We veranderen ook de kolommen naar het gepaste type voor dat verder bewerking gemakkelijk gaan. Daarna gaan we nog eens controleren op verborgen dubbels die we dan opnieuw verwijderen.

## Convert military time to datetime

1. import math from datetime
2. import datetime
3. def getTotalMinutes(time):

time\_str = str(time).rjust(4, "0")

hours = time\_str[: -2]

hours = int(hours)

if hours == 24:

hours = 0 minutes = int(time\_str[-2: ])

return datetime(2010, 1, 1, hours, minutes, 0)

def convertMilitaryTime(df):

df["departure\_schedule"] = df["departure\_schedule"].map(getTotalMinutes)

df["arrival\_schedule"] = df["arrival\_schedule"].map(getTotalMinutes)

return df

## Add time zones

1. #
2. UTC offset by state timezones = {
3. "AL": -6,
4. "AK": -9,
5. "AZ": -7,
6. "AR": -6,
7. "CA": -8,
8. "CO": -7,
9. "CT": -5,
10. "DE": -5,
11. "FL": -5,
12. "GA": -5,
13. "HI": -10,
14. "ID": -7,
15. "IL": -6,
16. "IN": -5,
17. "IA": -6,
18. "KS": -6,
19. "KY": -6,
20. "LA": -6,
21. "ME": -5,
22. "MD": -5,
23. "MA": -5,
24. "MI": -5,
25. "MN": -6,
26. "MS": -6,
27. "MO": -6,
28. "MT": -7,
29. "NE": -6,
30. "NV": -8,
31. "NH": -5,
32. "NJ": -5,
33. "NM": -7,
34. "NY": -5,
35. "NC": -5,
36. "ND": -6,
37. "OH": -5,
38. "OK": -6,
39. "OR": -8,
40. "PA": -5,
41. "RI": -5,
42. "SC": -5,
43. "SD": -6,
44. "TN": -6,
45. "UT": -7,
46. "TX": -6,
47. "VT": -5,
48. "VA": -5,
49. "WA": -8,
50. "DC": -5,
51. "WV": -5,
52. "WI": -6,
53. "WY": -5
54. }
55. def addTimezones(df):

df['arrival\_tz'] = df['arrival\_state'].map(lambda state: timezones[state])

df['departure\_tz'] = df['departure\_state'].map(lambda state: timezones[state])

return df

## Change local time to UTC

1. from datetime
2. import timedelta
3. def timeToUTC(schedule, tz):

tdelta = timedelta(hours = abs(tz))

if tz > 0:

schedule -= tdelta

else :

schedule += tdelta

return schedule

1. def convertLocalToUTC(df):

arrival\_schedules = []

departure\_schedules = []

for i, row in df.iterrows():

arrival\_schedule = row["arrival\_schedule"]

arrival\_tz = row["arrival\_tz"]

departure\_schedule = row["departure\_schedule"]

departure\_tz = row["departure\_tz"]

arrival\_schedules.append(timeToUTC(arrival\_schedule, arrival\_tz))

departure\_schedules.append(timeToUTC(departure\_schedule, departure\_tz))

df["arrival\_schedule"] = arrival\_schedules

df["departure\_schedule"] = departure\_schedules

return df

## Calculate duration

1. def addDurations(df):

durations = []

for i, row in df.iterrows():

arrival = row["arrival\_schedule"]

departure = row["departure\_schedule"]

if arrival <= departure:

arrival += timedelta(days = 1)

durations.append(arrival - departure)

df["duration"] = durations

return df

## Calculate distance

1. def addDistances(df):

distances = []

for index, row in df.iterrows():

distances.append(calcTheDistance(row))

df["distance"] = distances

return df

def calcTheDistance(el):

slat = radians(float(el["arrival\_lat"]))

slon = radians(float(el["arrival\_lon"]))

elat = radians(float(el["departure\_lat"]))

elon = radians(float(el["departure\_lon"]))

return 6371.01 \* acos(sin(slat) \* sin(elat) + cos(slat) \* cos(elat) \* cos(slon - elon))

We kunnen de afstand berekenen aan de hand van coördinaten van de luchthavens.

We zijn er bewust van dat die afstand niet 100% perfect is vanwege de gebruikte methode de *Great-circle distance* bereking. Die is namelijk niet 100% perfect want dat komt omdat de aarde namelijk geen perfecte bol is. En vliegtuigen vliegen namelijk niet altijd dezelfde routes vanwege de straalstroom, weeromstandigheden… Maar het is een goede benadering.

## Calculate speed

1. def addSpeed(df):

speeds = []

for i, row in df.iterrows():

duration = row["duration"]

speed = row["distance"] / ((duration.total\_seconds() / 60) / 60)

speeds.append(speed)

df["speed"] = speeds

return df

## Remove too long and too short flights

1. def deleteImpossibleFlights(df):

shortest\_distance\_in\_us = 100 #KM

longest\_distance\_in\_us = 9000#KM

shortest\_duration\_in\_us = timedelta(minutes = 16)

longest\_duration\_in\_us = timedelta(hours = 11, minutes = 40)

df = df[df.distance.between(shortest\_distance\_in\_us, longest\_distance\_in\_us)]

df = df[df.duration.between(shortest\_duration\_in\_us, longest\_duration\_in\_us)]

return df

## Execute cleaning

1. print("Aantal records om te beginnen:", len(df))
2. cleaning = df
3. cleaning = cleaning.drop\_duplicates()
4. print("Aantal records na verwijderen van dubbels:", len(cleaning))
5. cleaning = cleaning.reset\_index()
6. print("Aantal records na nieuwe index:", len(cleaning))
7. cleaning = cleaning.dropna()
8. print("Aantal records na verwijderen lege waarden:", len(cleaning))
9. cleaning = deleteWrongStates(cleaning)
10. cleaning = convertColumnTypes(cleaning)
11. cleaning = dropMoreAdvancedDuplicates(cleaning)
12. print("Aantal records na het verwijderen van de verborgen dubbels:", len(cleaning))
13. cleaning = convertMilitaryTime(cleaning)
14. cleaning = addTimezones(cleaning)
15. cleaning = convertLocalToUTC(cleaning)
16. cleaning = addDurations(cleaning)
17. cleaning = addDistances(cleaning)
18. cleaning = addSpeed(cleaning)
19. cleaning = deleteImpossibleFlights(cleaning)
20. print("Aantal records na het verwijderen van de onmogelijk lange of korte vluchten:", len(cleaning))
21. cleanData = cleaning

Als je dit uitvoert print python dit:

Aantal records om te beginnen: 11401196

Aantal records na verwijderen van dubbels: 10751921

Aantal records na nieuwe index: 10751921

Aantal records na verwijderen lege waarden: 10751919

Aantal records: 10751919

Aantal records na verwijderen foute vertrek staat: 10695506

Aantal records na verwijderen foute aankomst staat: 10642322

Aantal records na het verwijderen van de verborgen dubbels: 10642032

Aantal records na het verwijderen van de onmogelijk lange of korte vluchten: 10552133

Dus we waren begonnen met **11.401.196** records en hebben er nog **10.552.133** over. Dat wilt zeggen dat er **849.063** verkeerde de data tussen zit, dat is ongeveer **7.447%** van de data.

# Tableau

Inhoud:

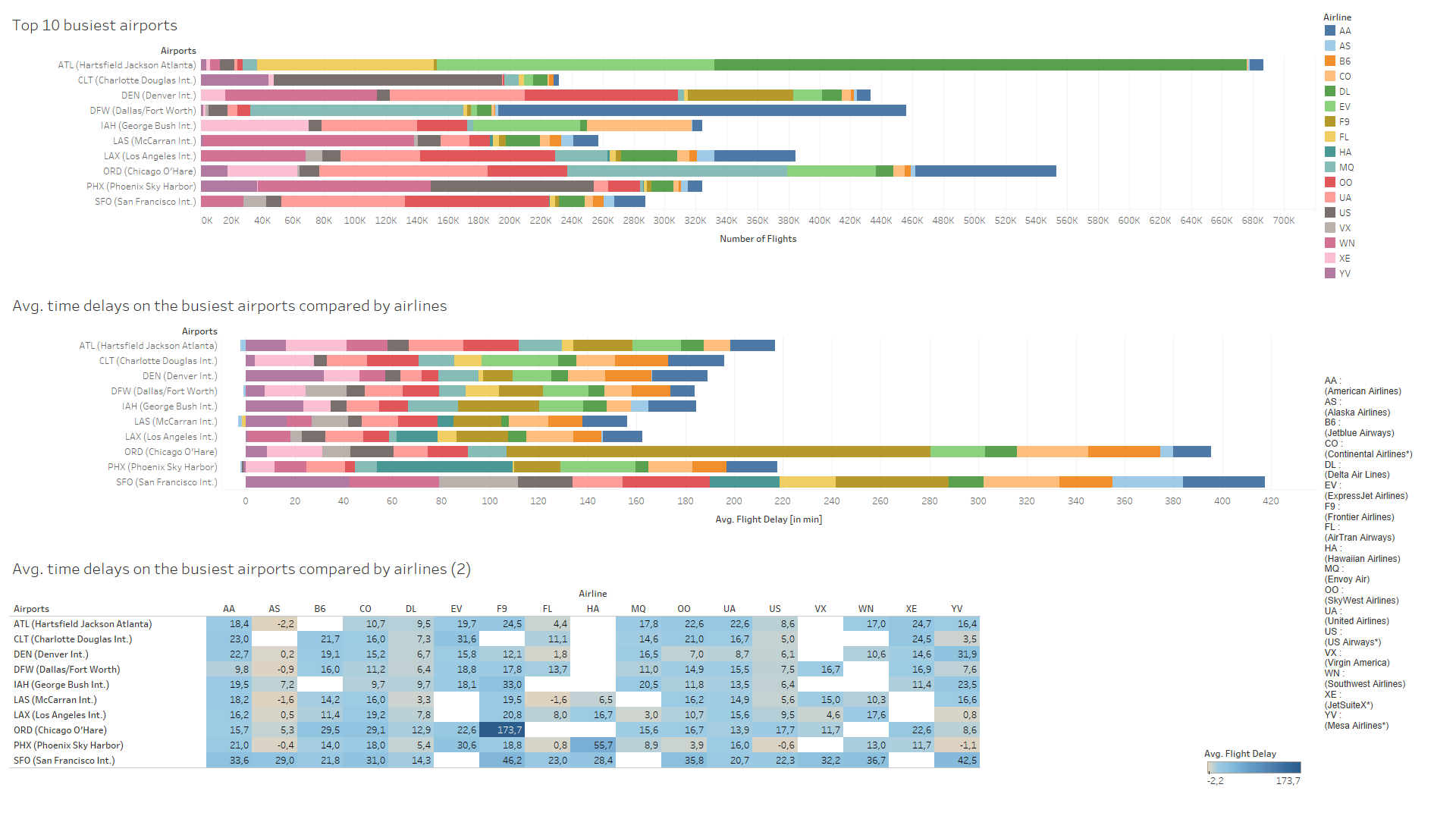
Grafiek 1

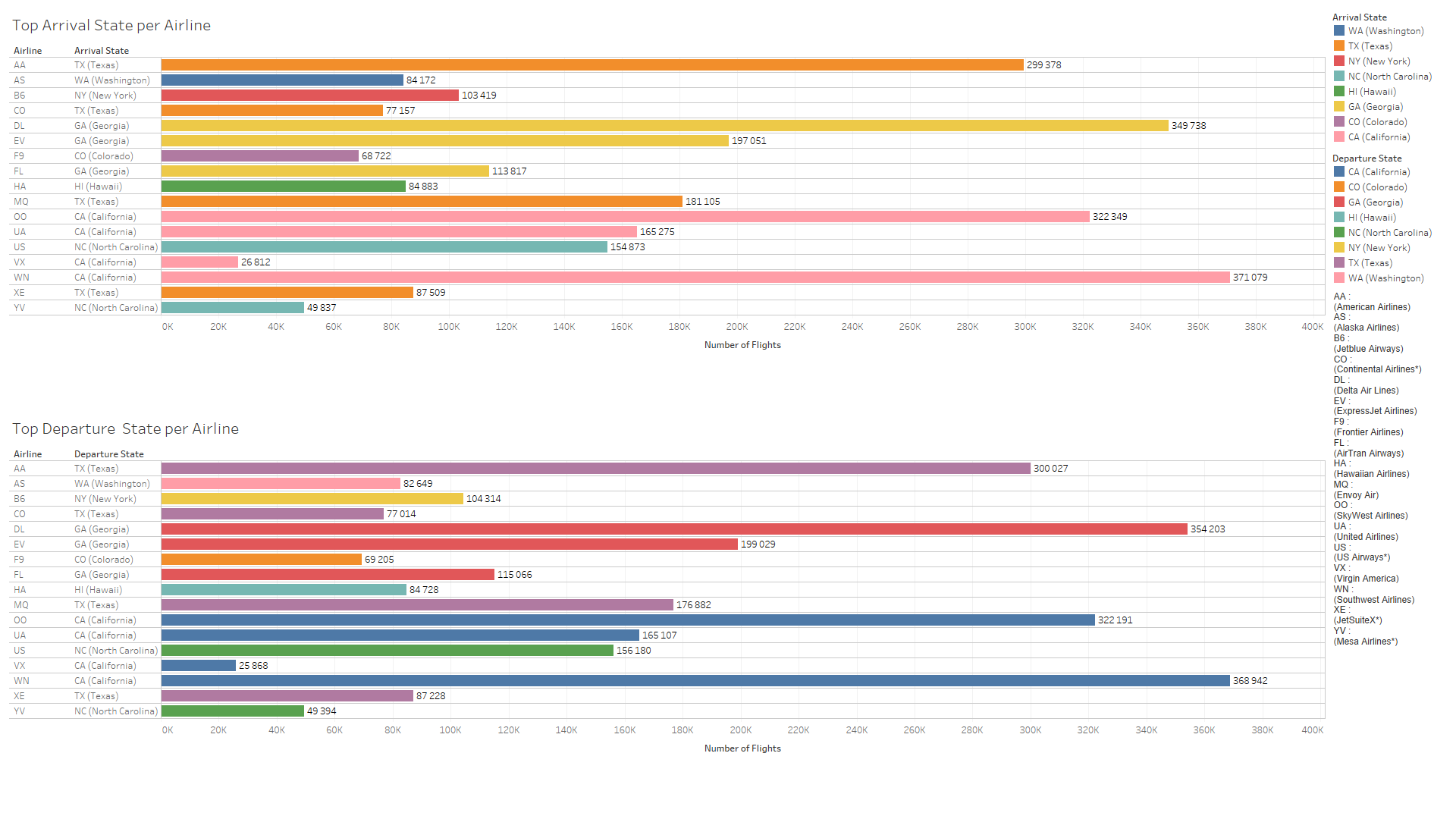
Grafiek 1

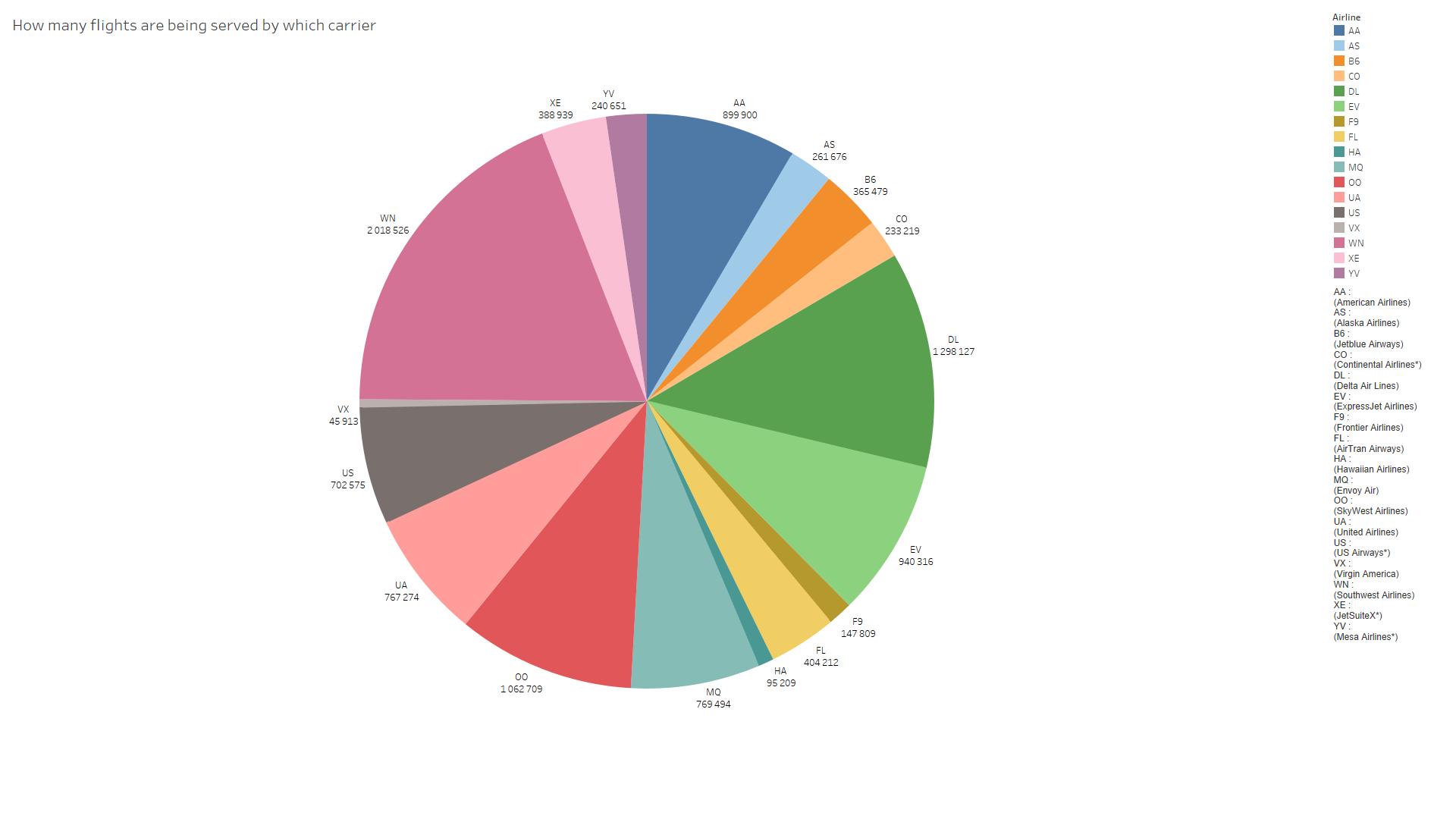
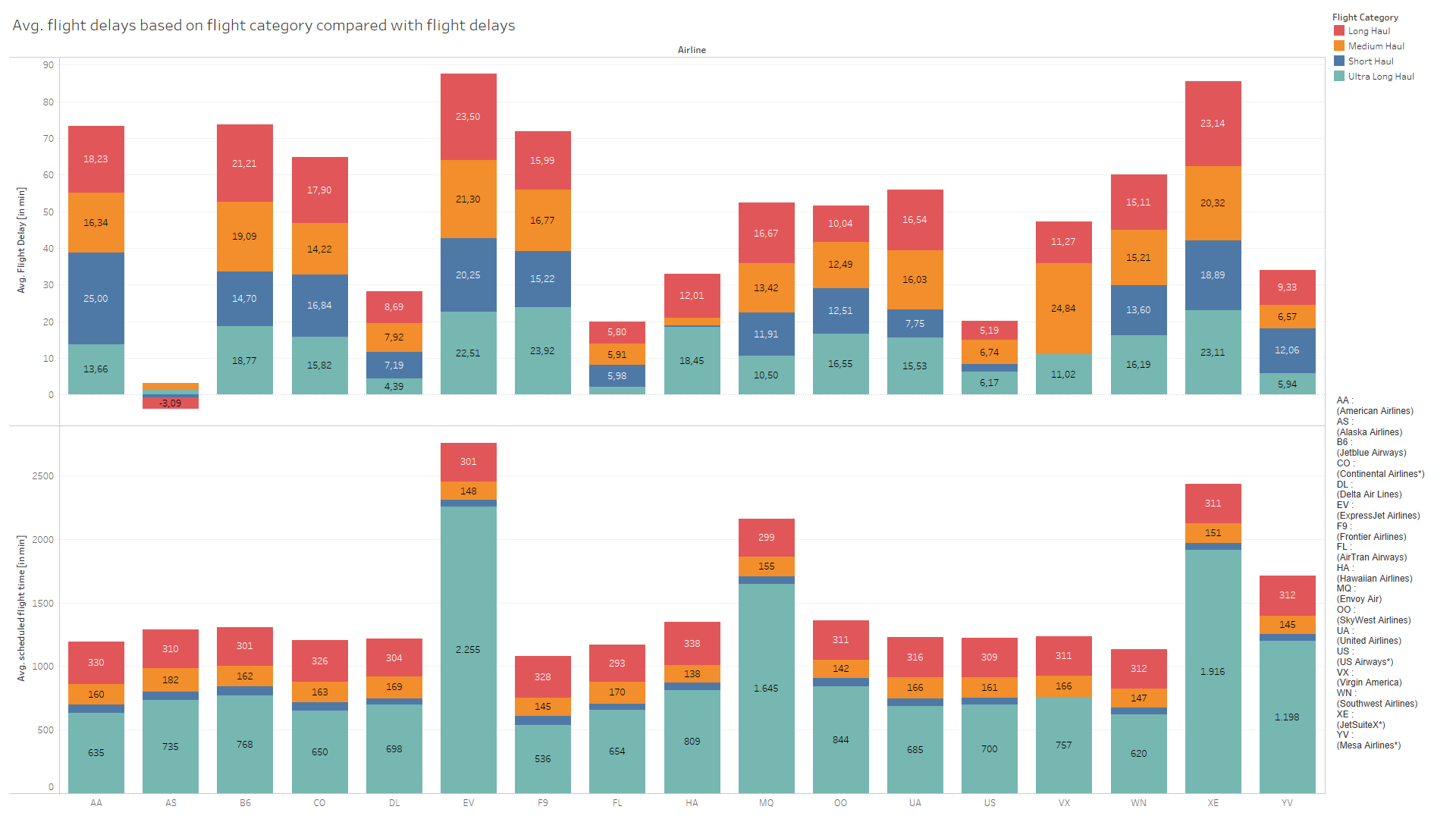
Grafiek 1

Grafiek 1

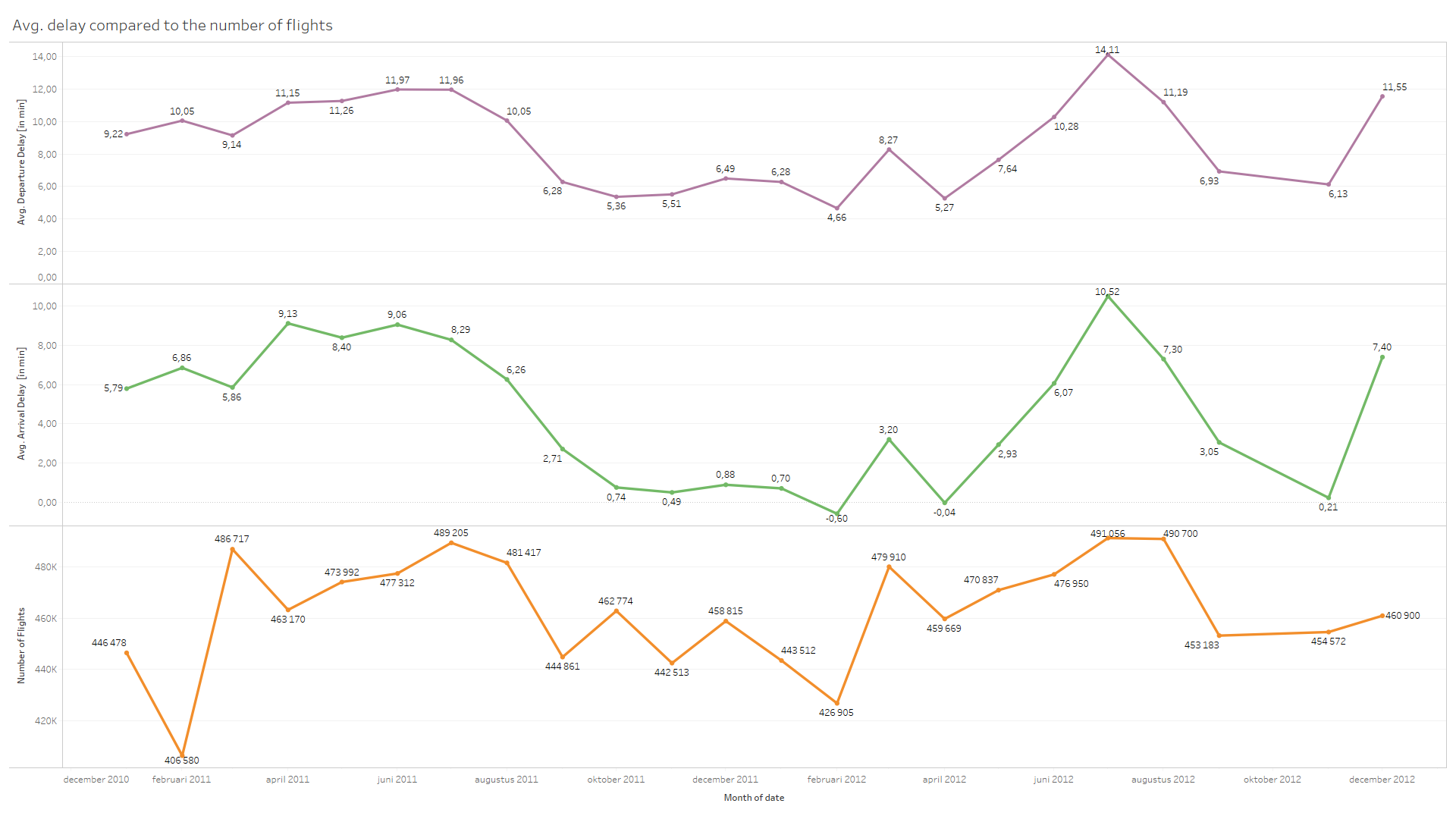
Grafiek 1

Grafiek 1,2 en 3:

Grafiek 4 en 5:

Grafiek 6: Grafiek 7: 

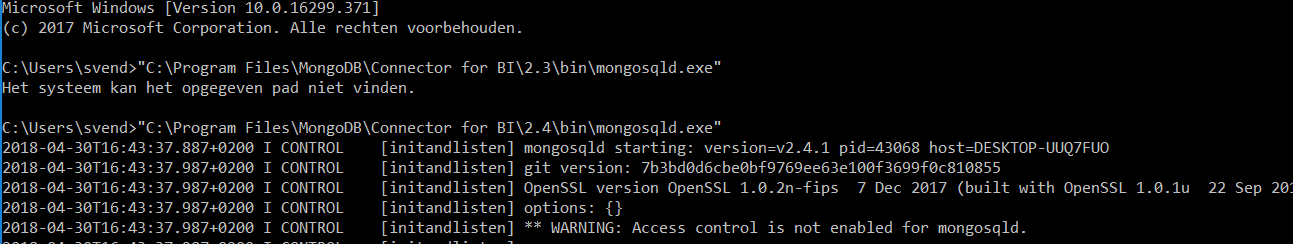
Grafiek 8:



"C:\Program Files\MongoDB\Connector for BI\2.4\bin\mongosqld.exe"

cd C:\Program Files\MongoDB\Server\3.6\bin

https://www.transtats.bts.gov/DL\_SelectFields.asp



# Machine learning met Scikit learn

# NetworkX