

Flight Delays Prediction

Big Data Technologies

Yavor Obreshkov

Vasil Radushev

WS 2017/2018

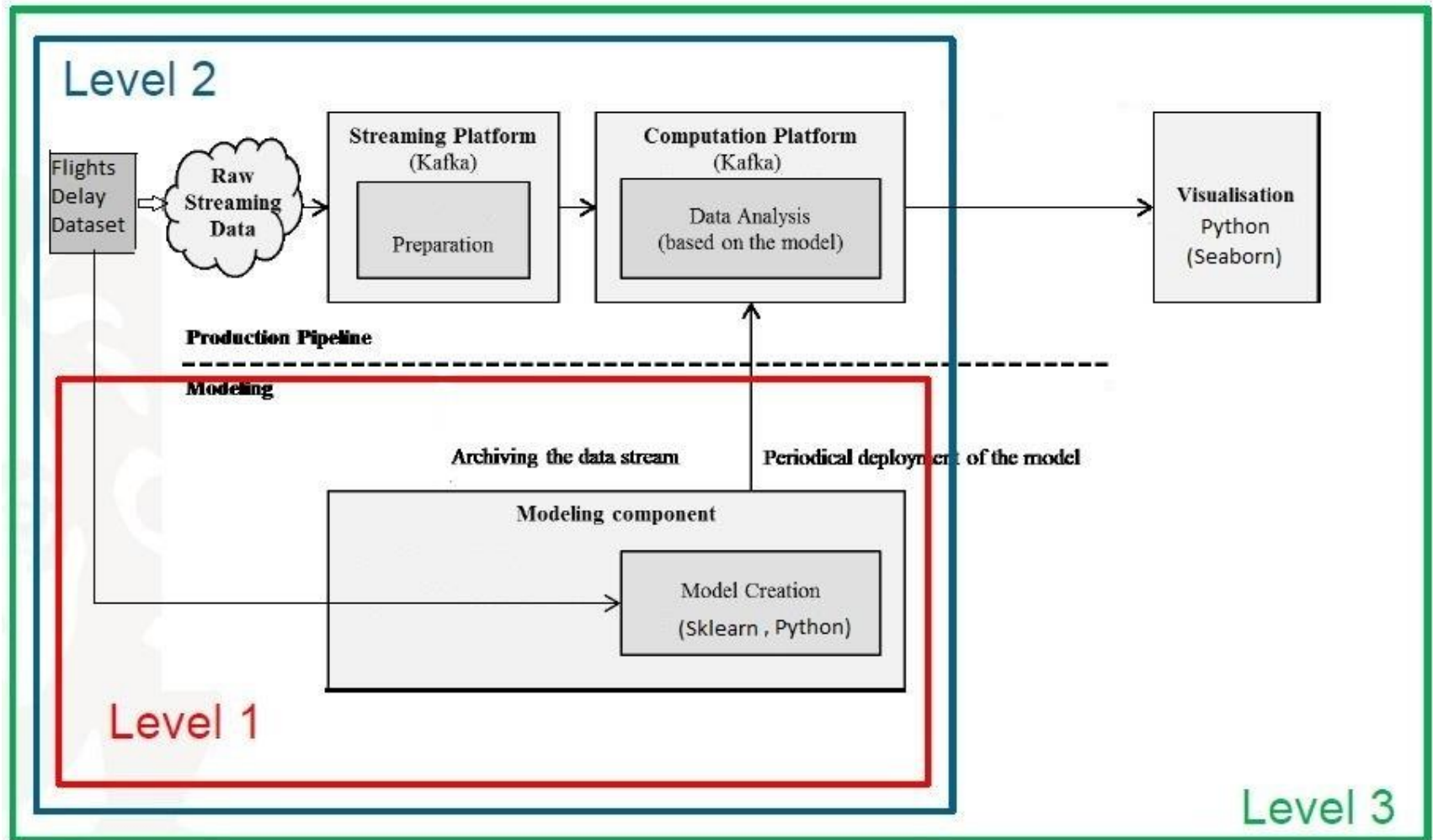
Motivation

- Predicting Flight Delays
- Finding Patterns and Dependencies about the Delays



DESTINATION	FLIGHT	GATE	REMARKS
BERLIN	LH543	09	DELANED
NEW YORK	AA978	28	CANCELLED
TORONTO	AC902	11	CANCELLED
MADRID	IB343	15	CANCELLED
BEIJING	HK205	11	CANCELLED

System Architecture



Data Streaming

Kafka Producer

Kafka Consumer

The image shows two Jupyter notebooks running on a Cloudera environment. The left notebook, titled 'consumer-running-version-0302', contains Python code for connecting to a Kafka consumer. The right notebook, titled 'producer-running-0302', contains Python code for connecting to a Kafka producer. Both notebooks show the output of the code execution, including a list of flight data in the consumer notebook and a confirmation message in the producer notebook.

Consumer Notebook Code:

```
import numpy as np
import pandas as pd
import json

print('Start')
consumer0302=KafkaConsumer('topic0302', bootstrap_servers=['localhost:9092'])

#create our first data frame, columns only, no index, no data
#df=pd.DataFrame(columns=columns)

ls=[]

print('Starting the loop')
i=0
#we create an empty list, in which we will insert each dataframe
for msg in consumer0302:
    i=i+1
    if i<=1000:
        test = msg.value.split(",")
        df = pd.DataFrame(test)
        ls.append(test)
    else:
        consumer0302.close()
        break

df = pd.DataFrame(ls)
print(df)
```

Consumer Notebook Output:

```
983 "0825" "0820" "-.5" "1018" "-.11" "0" "0"
984 "0825" "0820" "-.5" "1018" "-.11" "0" "0"
985 "0825" "0820" "-.5" "1018" "-.11" "0" "0"
986 "0825" "0820" "-.5" "1018" "-.11" "0" "0"
987 "0825" "0820" "-.5" "1018" "-.11" "0" "0"
988 "0825" "0820" "-.5" "1018" "-.11" "0" "0"
989 "0825" "0820" "-.5" "1018" "-.11" "0" "0"
990 "0825" "0820" "-.5" "1018" "-.11" "0" "0"
991 "0825" "0820" "-.5" "1018" "-.11" "0" "0"
992 "0825" "0820" "-.5" "1018" "-.11" "0" "0"
993 "0825" "0820" "-.5" "1018" "-.11" "0" "0"
994 "0825" "0820" "-.5" "1018" "-.11" "0" "0"
```

Producer Notebook Code:

```
e=open("flights.csv", 'r').readlines()
0
flights_clean_all=[]
flights_clean=[]

#on average 14k flights a DAY !!!
#bullshit more like 42k a day, \
#source: https://www.faa.gov/air_traffic/by_the_numbers/

nlines=500 #could be defined by the user

times= len(e) // nlines
rest = len(e) % nlines

print(times)
print(rest)

def func(j,k):
    for i in range(j,k):
        columns_first=e[i].split(",")[0:12]
        columns_last=e[i].split(",")[21:25]
        columns=columns_first + columns_last
        producer.send(topic,json.dumps(columns))
        print(columns)

for i in range(1, 200): #times
    j=(i-1)*nlines
    k=i*nlines
    print("starting row is",j)
    print("last row from the stream is",k-1)
    print("Start streaming in 10 seconds\n")
    time.sleep(5)
    func(j,k)
    print("The batch is completed\n")
    time.sleep(10)
```

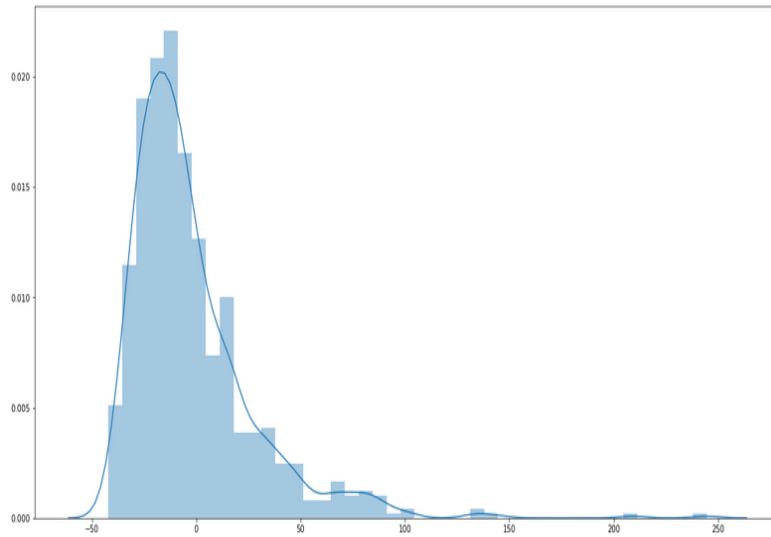
Producer Notebook Output:

```
8, "13", "0", "0"
["2015", "1", "1", "4", "AA", "1024", "N505AA", "ABQ", "DFW", "1530", "1544", "14", "1823", "13", "0", "0"]
["2015", "1", "1", "4", "AA", "1246", "N3HNA", "LGA", "MIA", "1530", "1532", "2", "1857", "0", "0", "0"]
["2015", "1", "1", "4", "AA", "1550", "N3AVAA", "LAX", "BOS", "1530", "1534", "4", "2328", "-17", "0", "0"]
["2015", "1", "1", "4", "AA", "1555", "N3DCAA", "ORD", "LAX", "1530", "1529", "-1", "1731", "-24", "0", "0"]
["2015", "1", "1", "4", "AA", "1672", "N3KYAA", "AUS", "LAX", "1530", "1540", "10", "1646", "-4", "0", "0"]
```

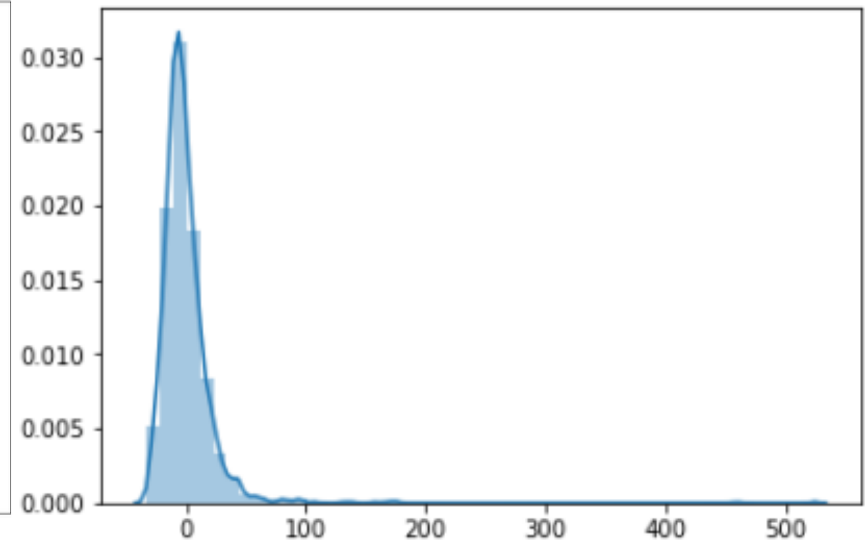
Use Describe in the Linux terminal to verify streaming process

Training the model with - 75 % to 25% split

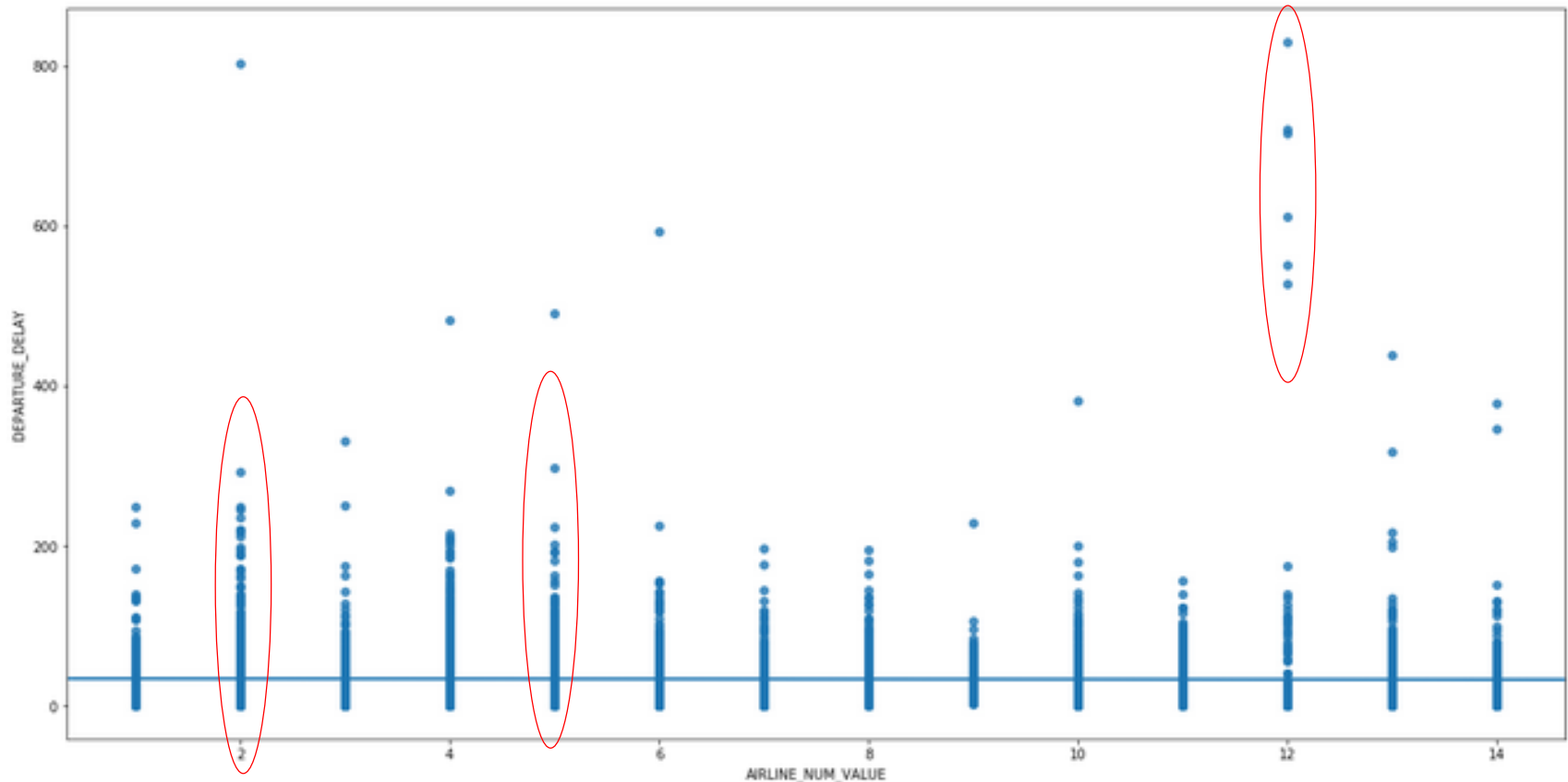
Training the model on 100k data rows



Training the model on the full data set



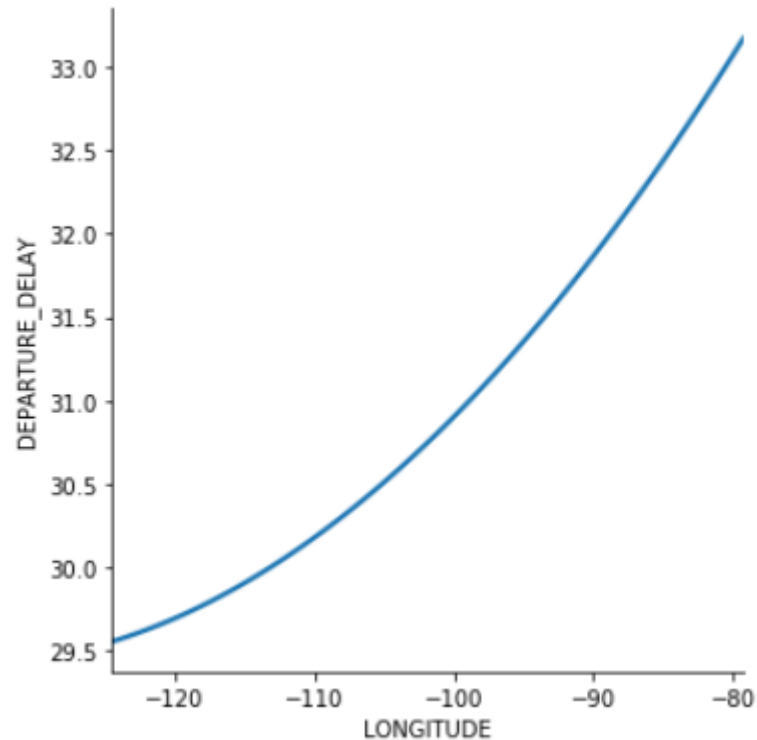
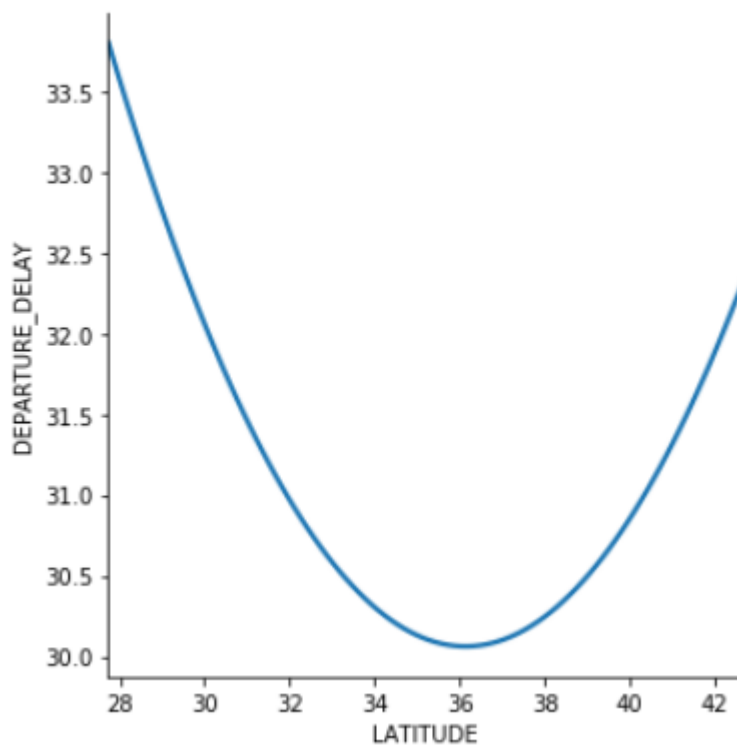
We can see that the difference between estimated and realized values declines



Using Airline_Num_Values, we can see the distribution and the extreme departure delays based on the airlines.

- American Airlines Inc.(2) and JetBlue Airways (5) are the two lines with the most severe delays
- Hawaiian Airlines Inc (12) had some major issues in 2015, as there were several departure delays of above 5 hours
- Southwest Airlines Co. (9) hat the mildest delays

Regression Plots



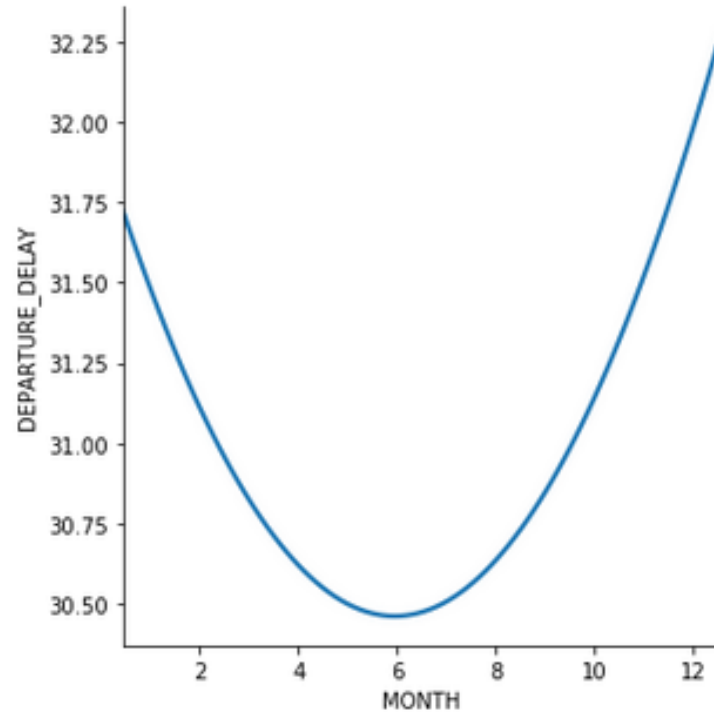
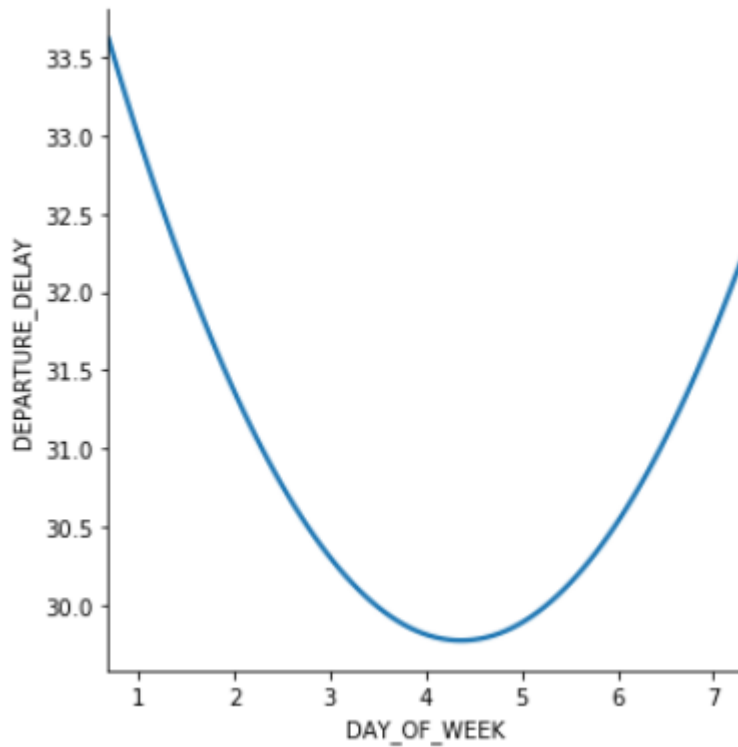
Explaining the departure delays via geographical coordinates:

- We conducted the regression based on airports' geographical position
- As the longitude increases, (going north), the departure delays increase
- Going further inland and off the coast, the departure delays decrease

Predictions

Further Factors we included in the model are:

- Days of the week
- Months



Predictions

In the consumer, we read each message and the model “predicts” the possible delay based on the

- origin airport (LAT and LON position),
- month
- day of the week,

```

Start
Starting the loop
The flight EV6099 scheduled for departure from DEN could be possibly delayed by 36 minutes.
The flight NK657 scheduled for departure from MSY could be possibly delayed by 37 minutes.
The flight 006413 scheduled for departure from BIL could be possibly delayed by 38 minutes.
The flight B6251 scheduled for departure from BOS could be possibly delayed by 35 minutes.
The flight 004797 scheduled for departure from SNA could be possibly delayed by 38 minutes.
The flight 004586 scheduled for departure from SJC could be possibly delayed by 38 minutes.
The flight UA202 scheduled for departure from ORD could be possibly delayed by 40 minutes.
The flight US462 scheduled for departure from PHL could be possibly delayed by 39 minutes.
The flight US450 scheduled for departure from PHL could be possibly delayed by 39 minutes.
The flight US887 scheduled for departure from JFK could be possibly delayed by 39 minutes.
The flight US1896 scheduled for departure from PHL could be possibly delayed by 39 minutes.
The flight US1969 scheduled for departure from RSW could be possibly delayed by 39 minutes.
The flight US2017 scheduled for departure from FLL could be possibly delayed by 39 minutes.
The flight US2063 scheduled for departure from DFW could be possibly delayed by 39 minutes.
The flight WN1457 scheduled for departure from ATL could be possibly delayed by 37 minutes.
The flight WN1361 scheduled for departure from ATL could be possibly delayed by 37 minutes.
The flight WN962 scheduled for departure from BNA could be possibly delayed by 37 minutes.
The flight WN736 scheduled for departure from FLD could be possibly delayed by 37 minutes.

In [102]: 1 string = str(prediction)
          2 print(string[2:5])
          3 int(string)
file:///home/cloudera/Document/cloudera-manager.html

umer+predictive... [cloudera@quickstart... [Anaconda Navigator] [Terminal] [Datasets]

```

Thank you for your attention 😊

Questions?