GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Flight Delays Prediction

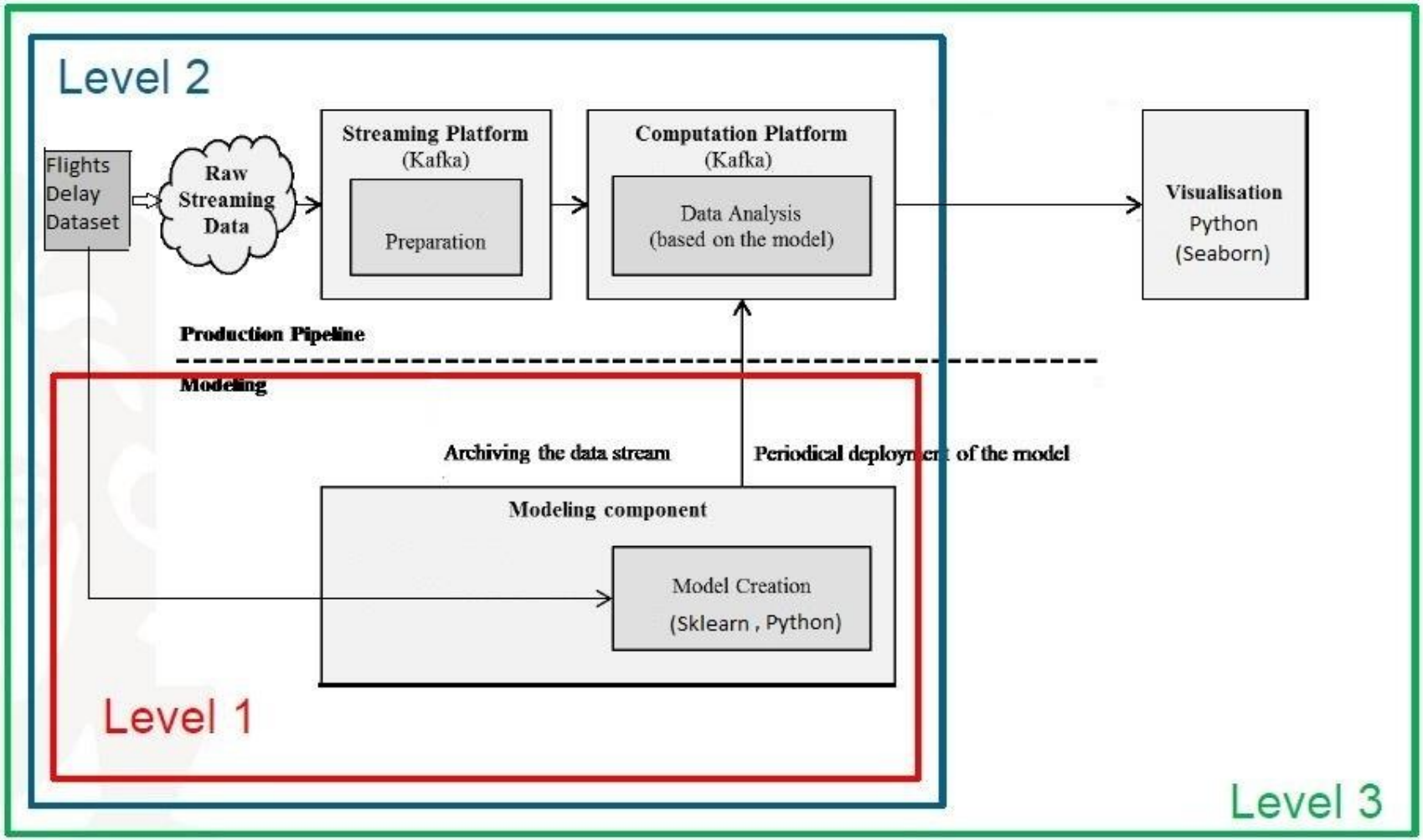
Big Data Technologies

Yavor Obreshkov
Vasil Radushev
WS 2017/2018

Motivation

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

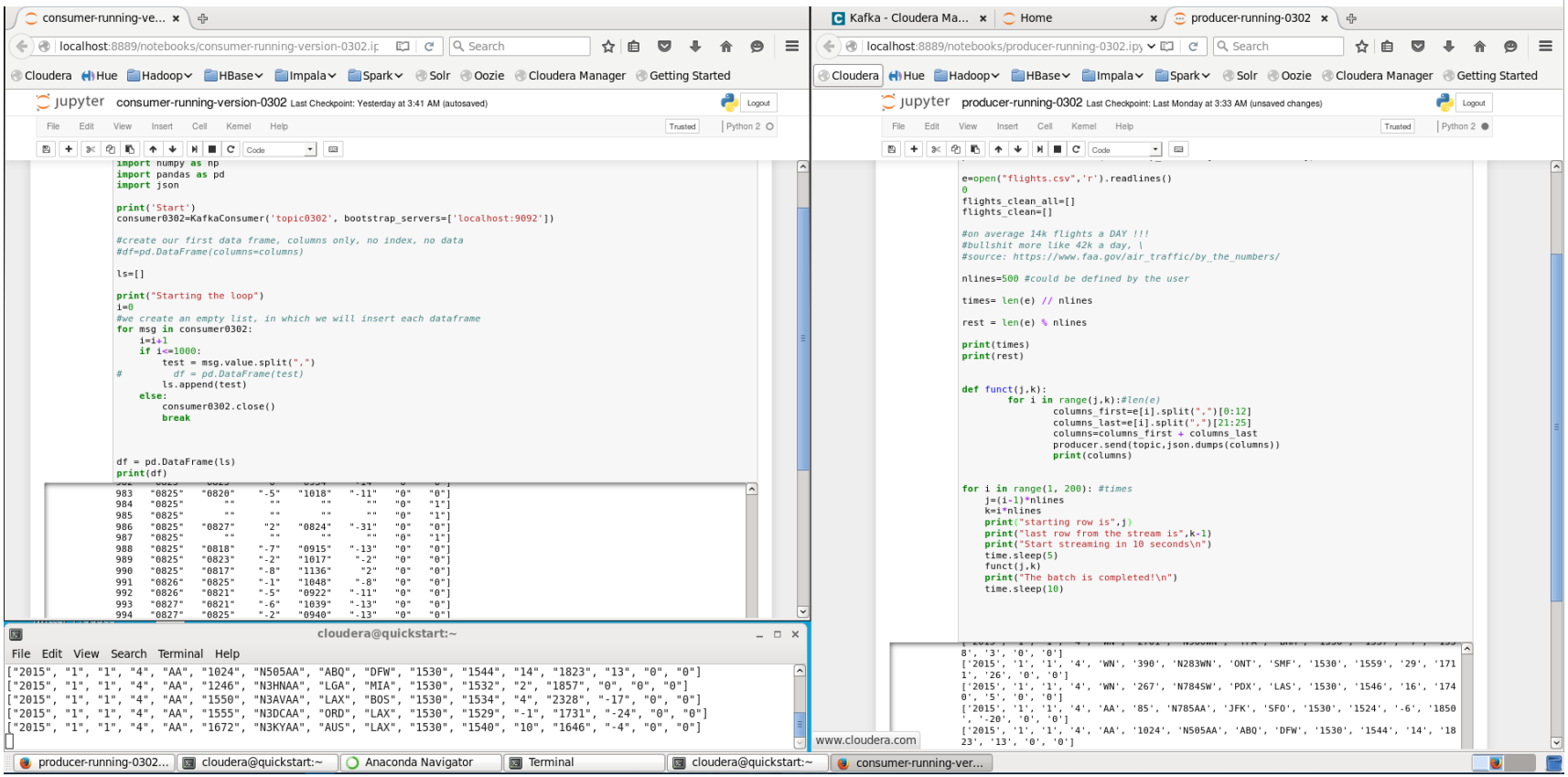
System Architecture



Data Streaming

Kafka Producer

Kafka Consumer

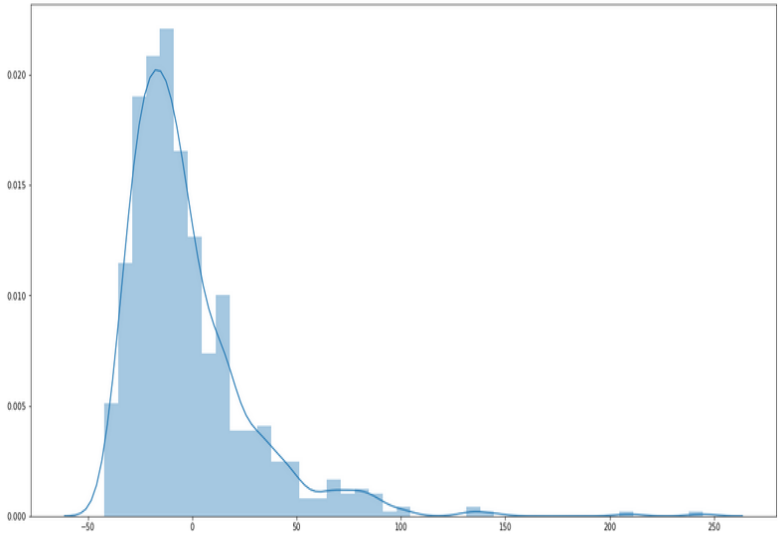


Use Describe in the Linux terminal to verify streaming process

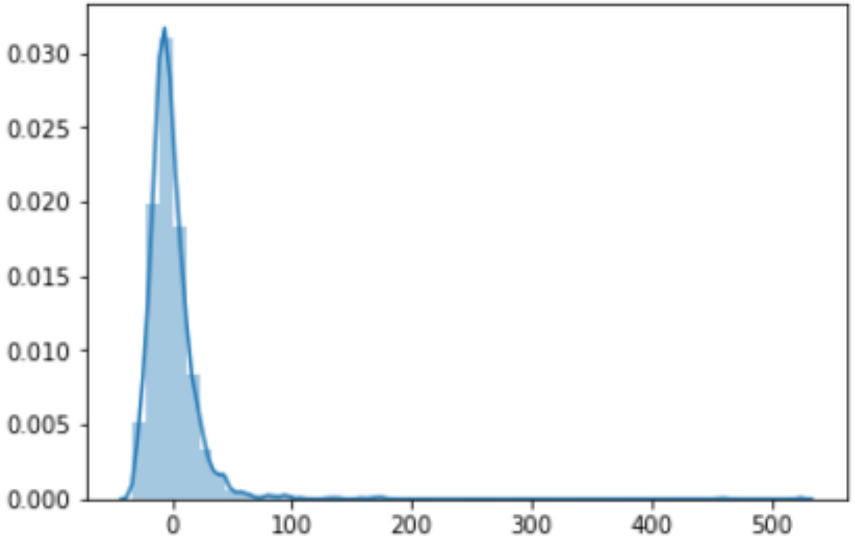
Model

Training the model with - 75 % to 25% split

Training the model on 100k data rows

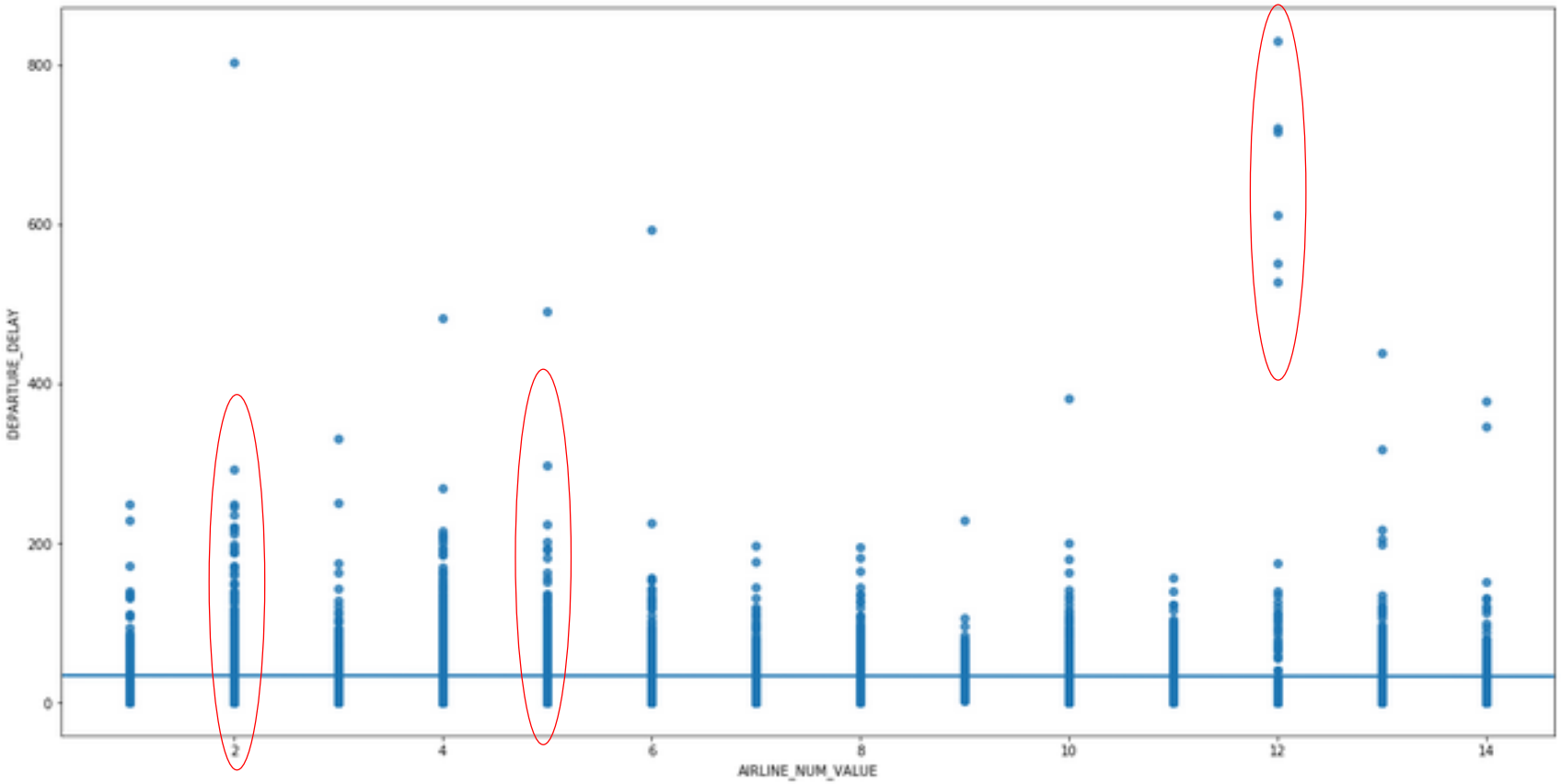


Training the model on the full data set



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

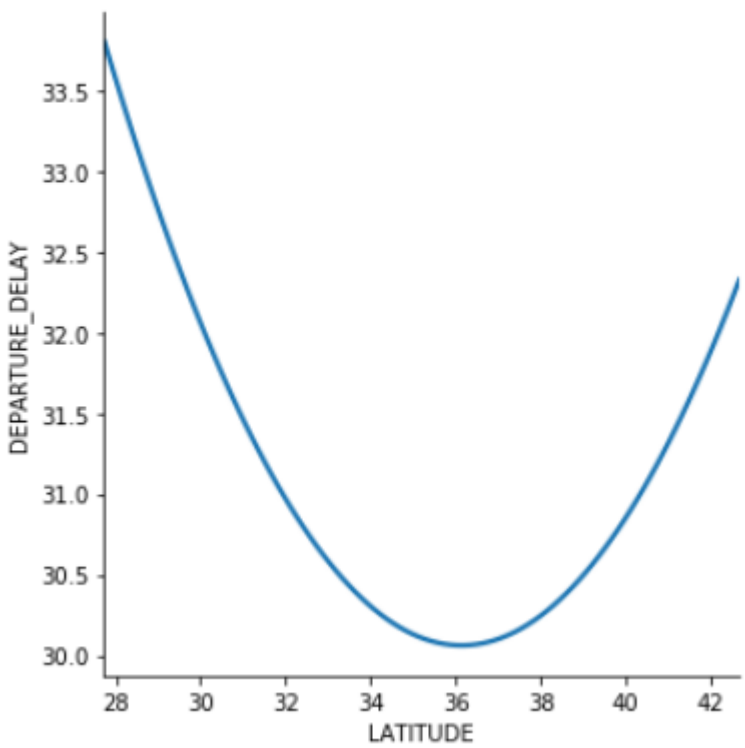
Plots

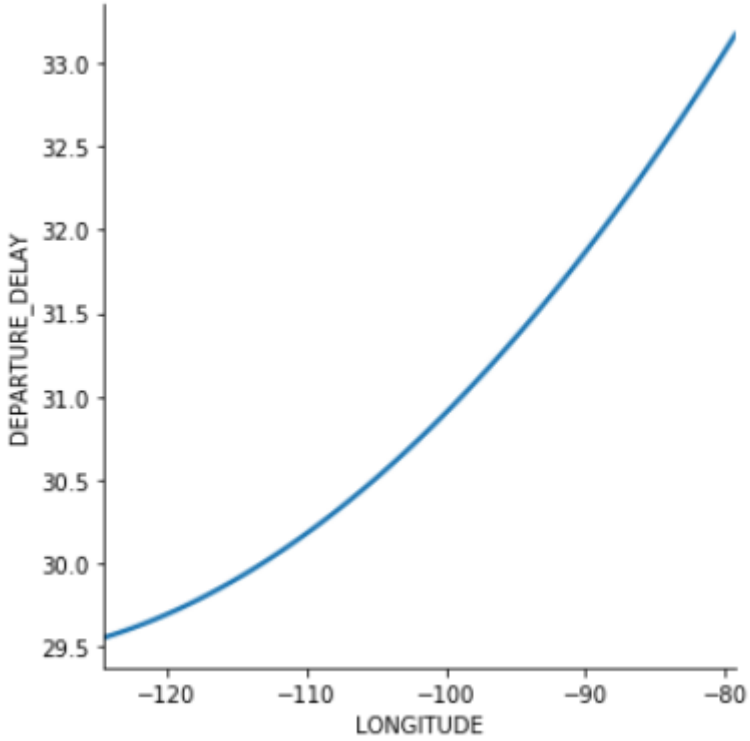


Using Airline_Num_Values, we can see the distribution and the extreme departure delays based on the airlines.

- 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





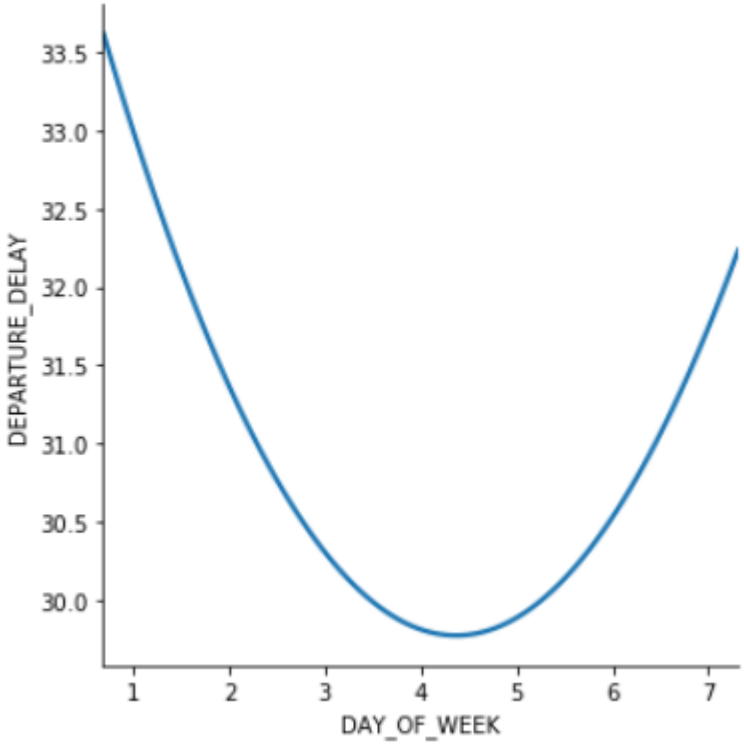
Explaining the departure delays via geographical coordinates:

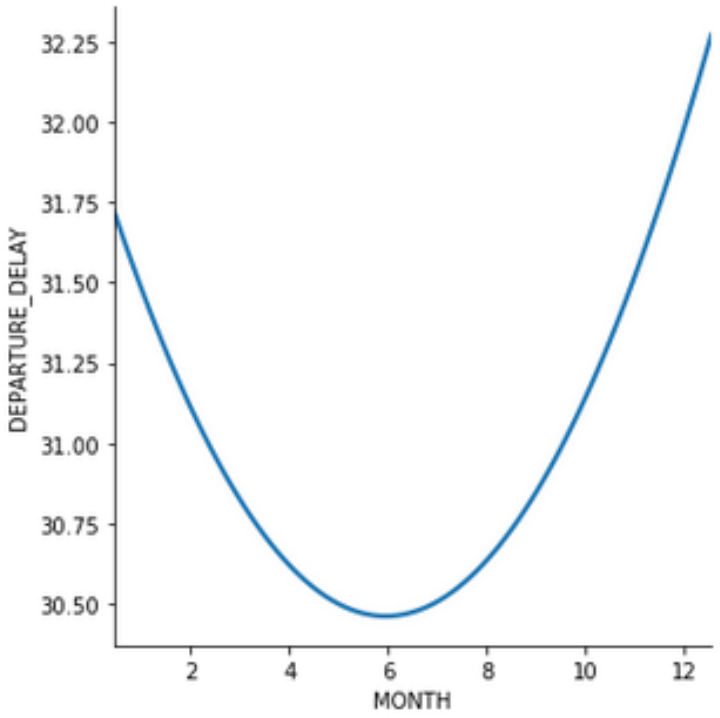
- We conducted the regression based on airports' geographical position
- As the longitude increases, (going north), the departure delays increase
- Going further inland and off the coast, the departure delays decrease

Predictions

Further Factors we included in the model are:

- Days of the week
- Months

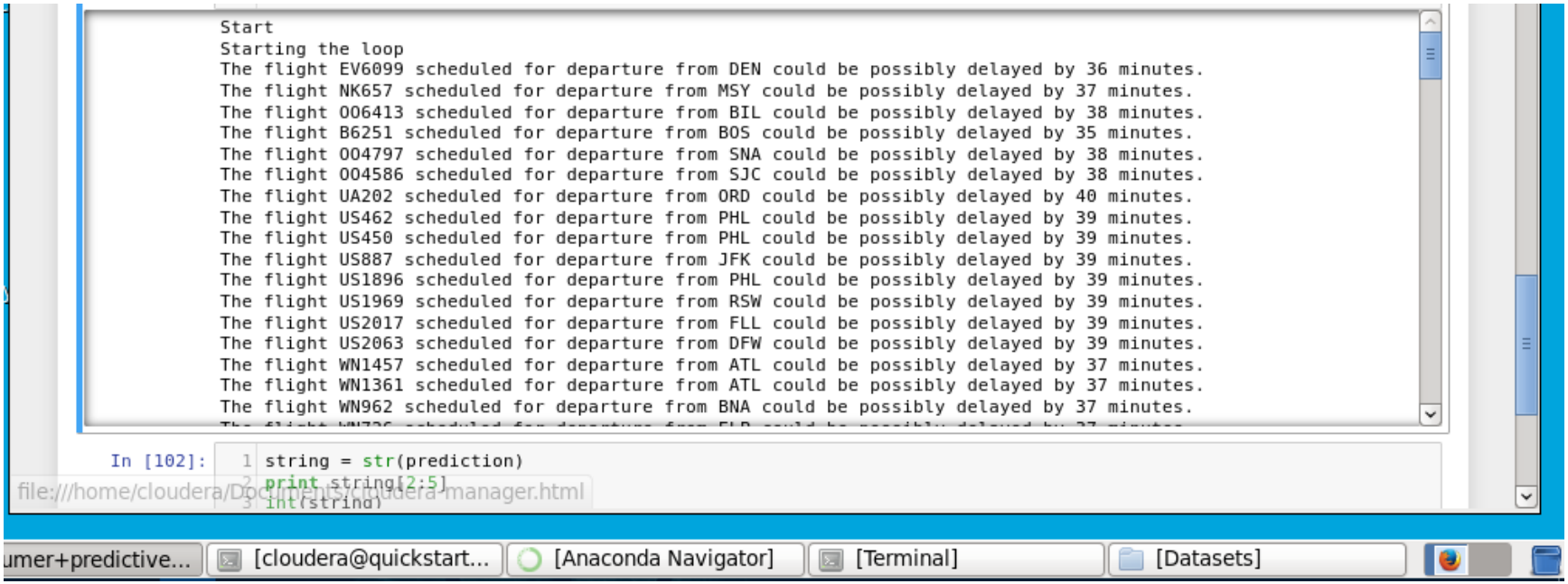




Predictions

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,**



GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Thank you for your attention ☺

Questions?