Machine learning academy (BOSCH)

Airline data anlysis

Nenad Knežević

Introduction

Hello, my name is Nenad Knežević and I love to explore and learn new things about machine learning.

Thank you very much for taking time to consider my application

Due to the limitations of my CPU, I analyzed data for 3 years (2014, 2015 and 2016) instead of 5.

For the same reason, I used data from year 2018 (20% of total data for the year) for linear regression model.

Mean and median delay for years

	Arrivals		Departures	
Year	Mean	Median	Mean	Median
2014	7.328	-2.000	10.624	-1.000
2015	4.407	-5.000	9.370	-1.000
2016	3.519	-6.000	8.938	-2.000

Table 1. Mean and median for arrivals and departures by years

Year	Mean	Median
2014	17.970	-3.000
2015	13.777	-6.000
2016	12.457	-8.000

Table 2. Mean and median summed for year

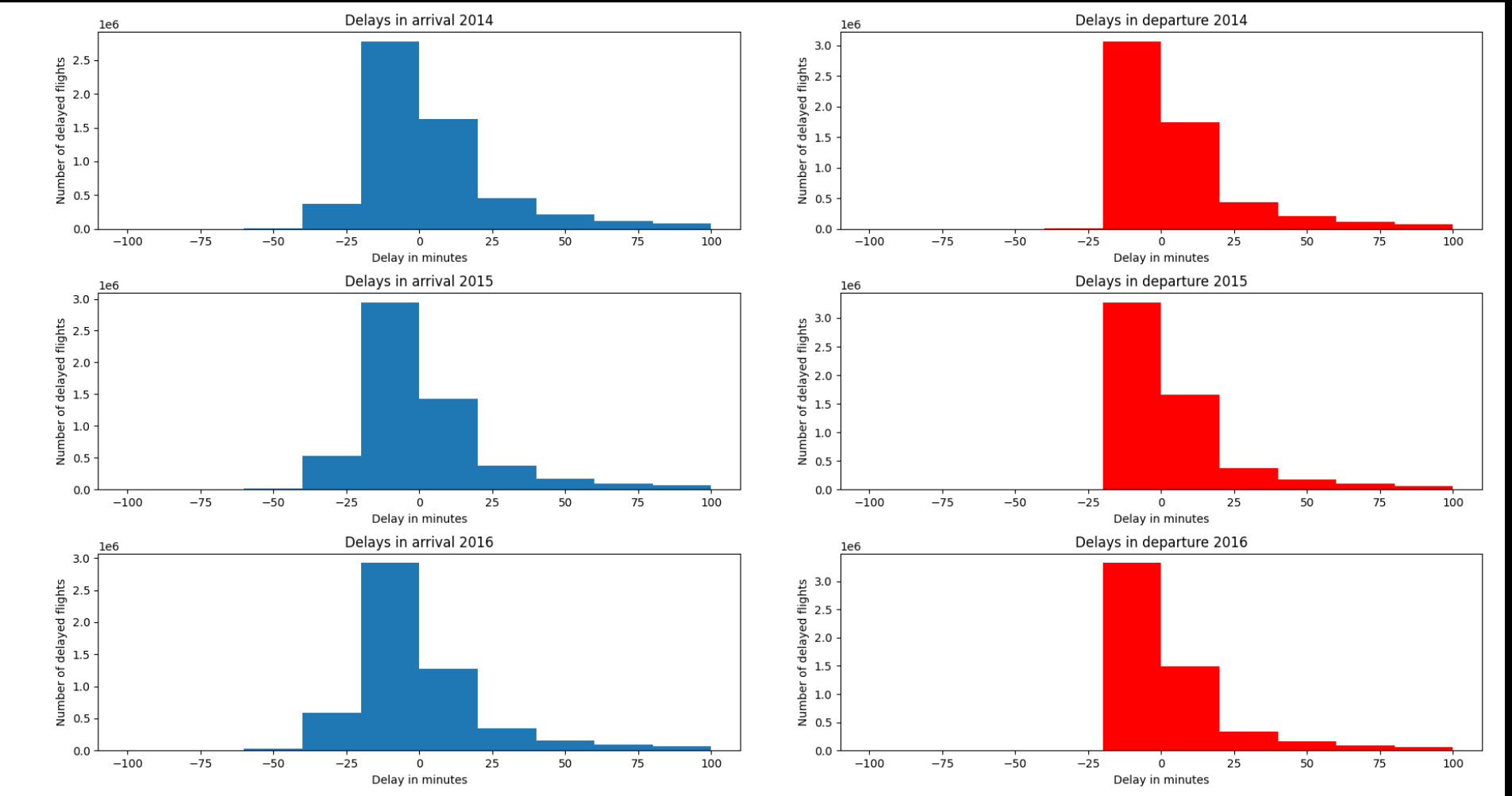
Skewness and kurtosis

	Arrivals		Departures	
Year	Skewness	Kurtosis	Skewness	Kurtosis
2014	6.173	96.983	7.183	122.403
2015	6.562	99.592	7.650	124.899
2016	7.800	129.816	8.961	157.686

Table 3. Skewness and kurtosis by year

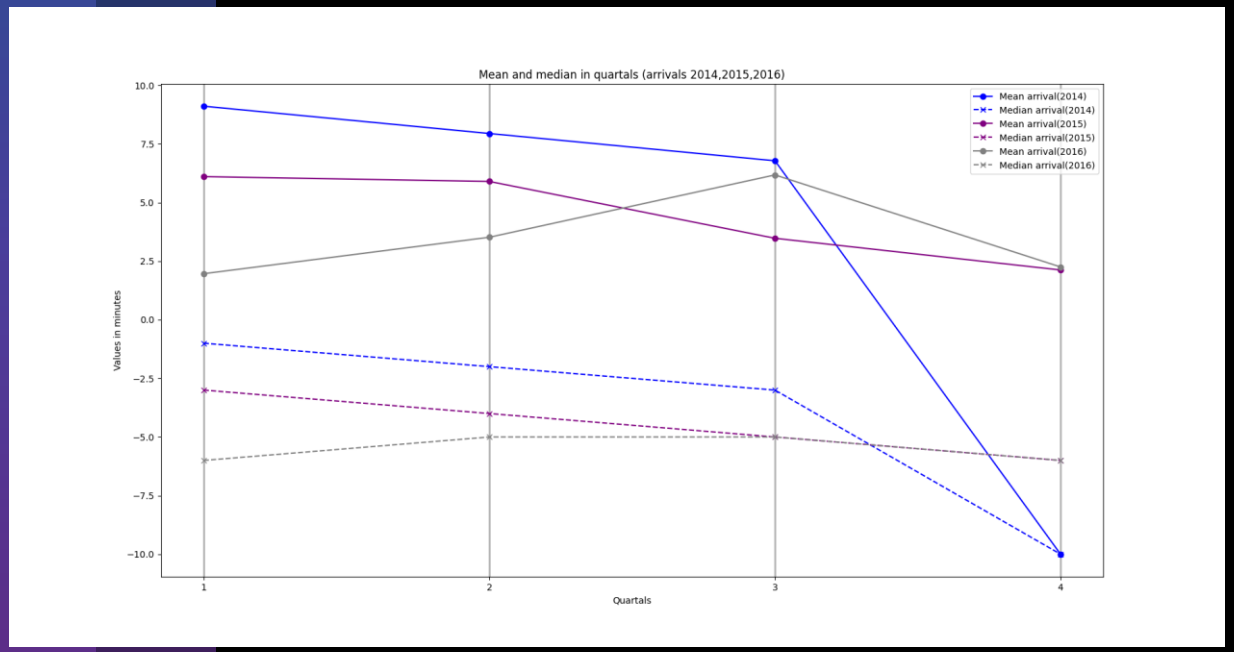
- Positive kurtosis tells us that values are concentrated around mean value.
- Skewness tells us that there are some flights with high delays.
- I would use a normal gaussian distribution because of central limit theorem, which says that if the number of samples is growing, it will lean towards normal gaussian distribution.

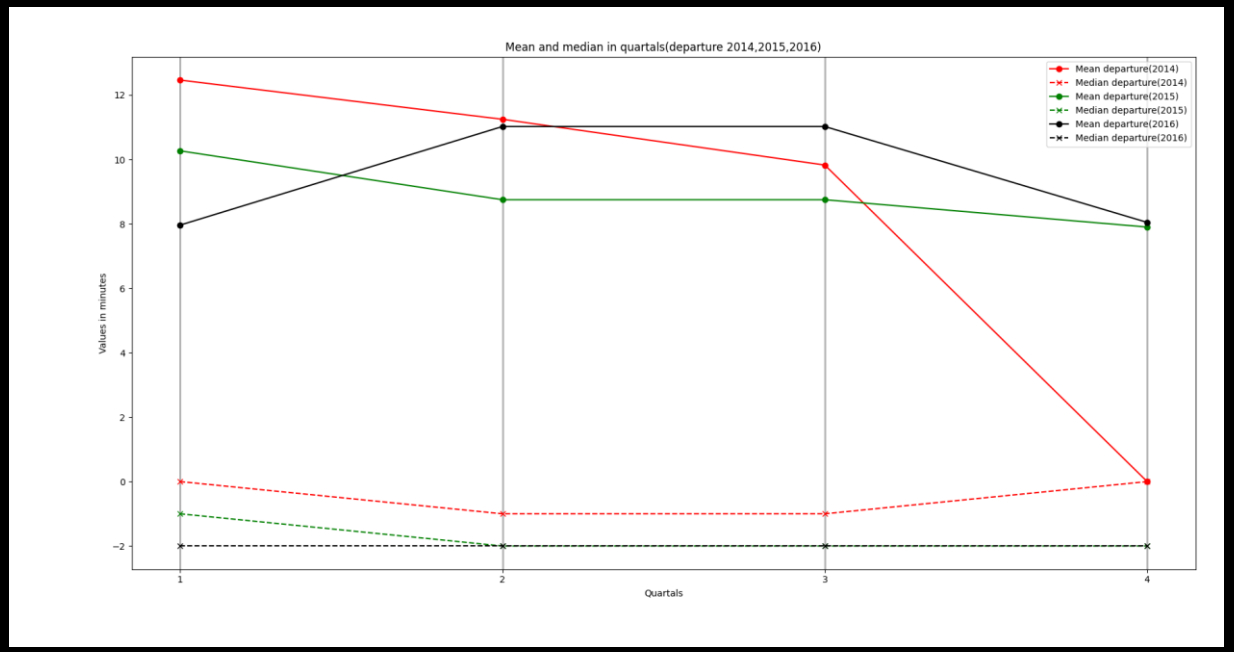
Skewness and kurtosis



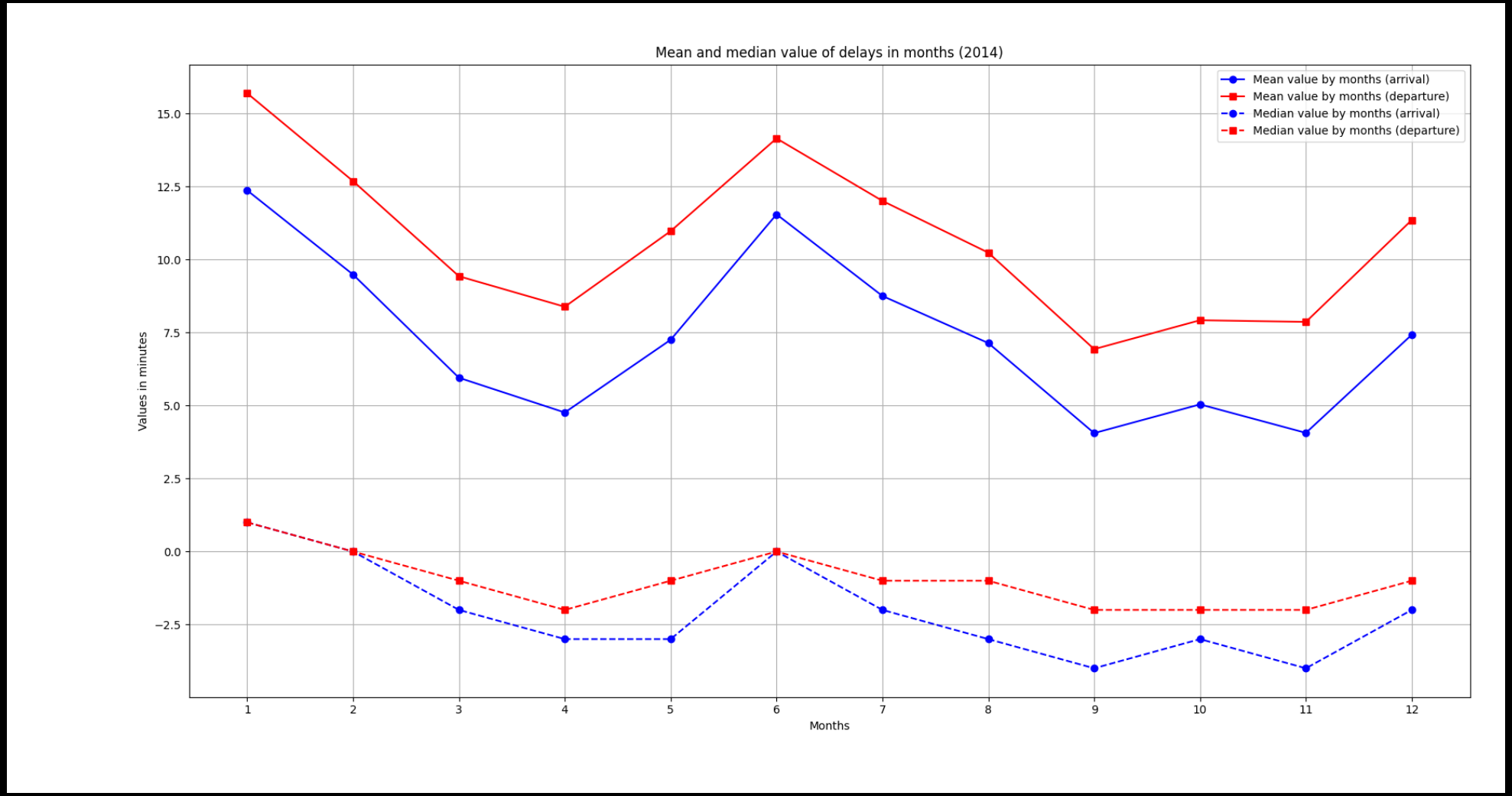
Picture 1. Delay densities

Mean and median (quarter of year)

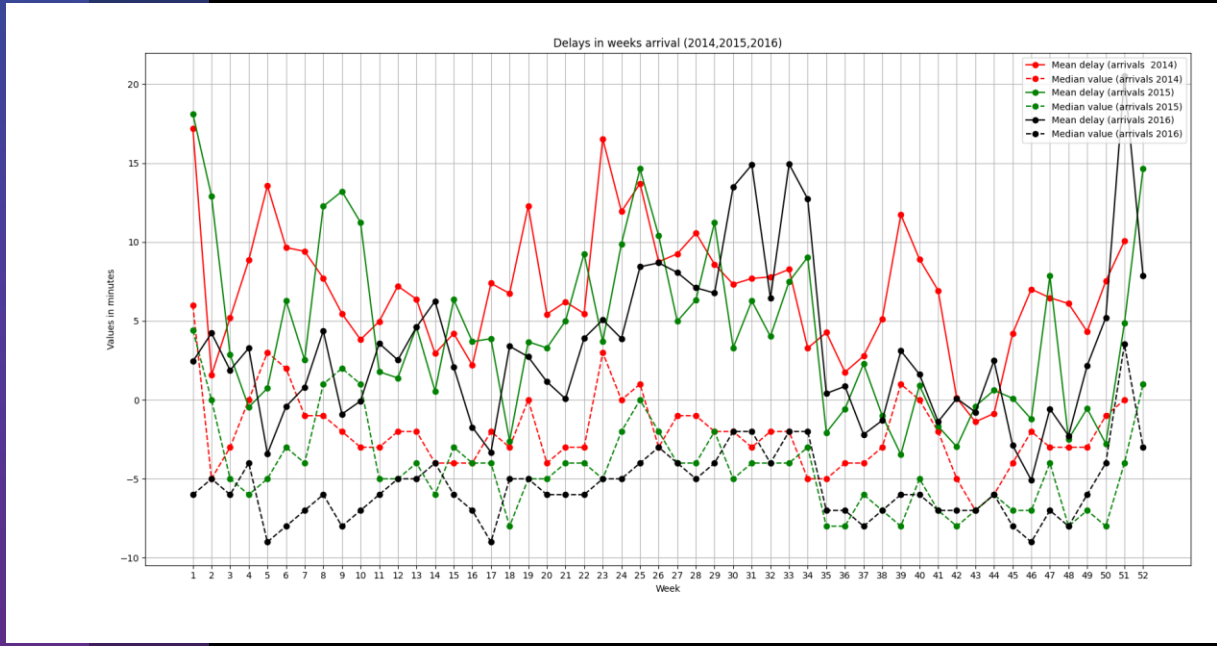


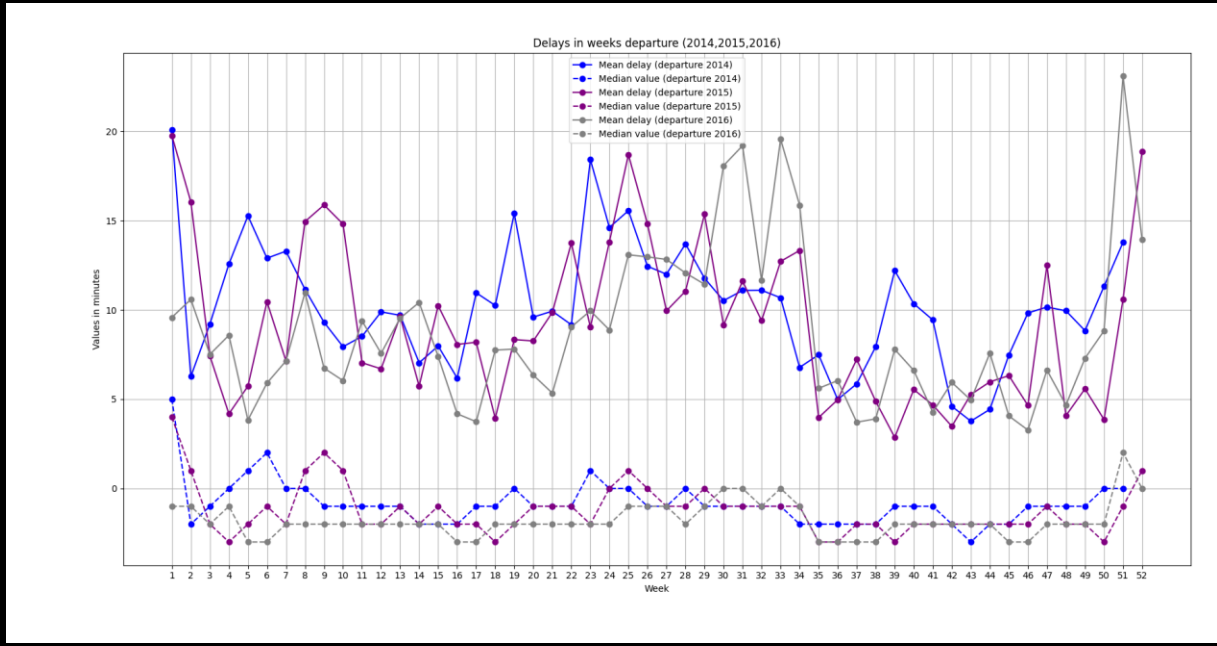


Mean and median (monthly)

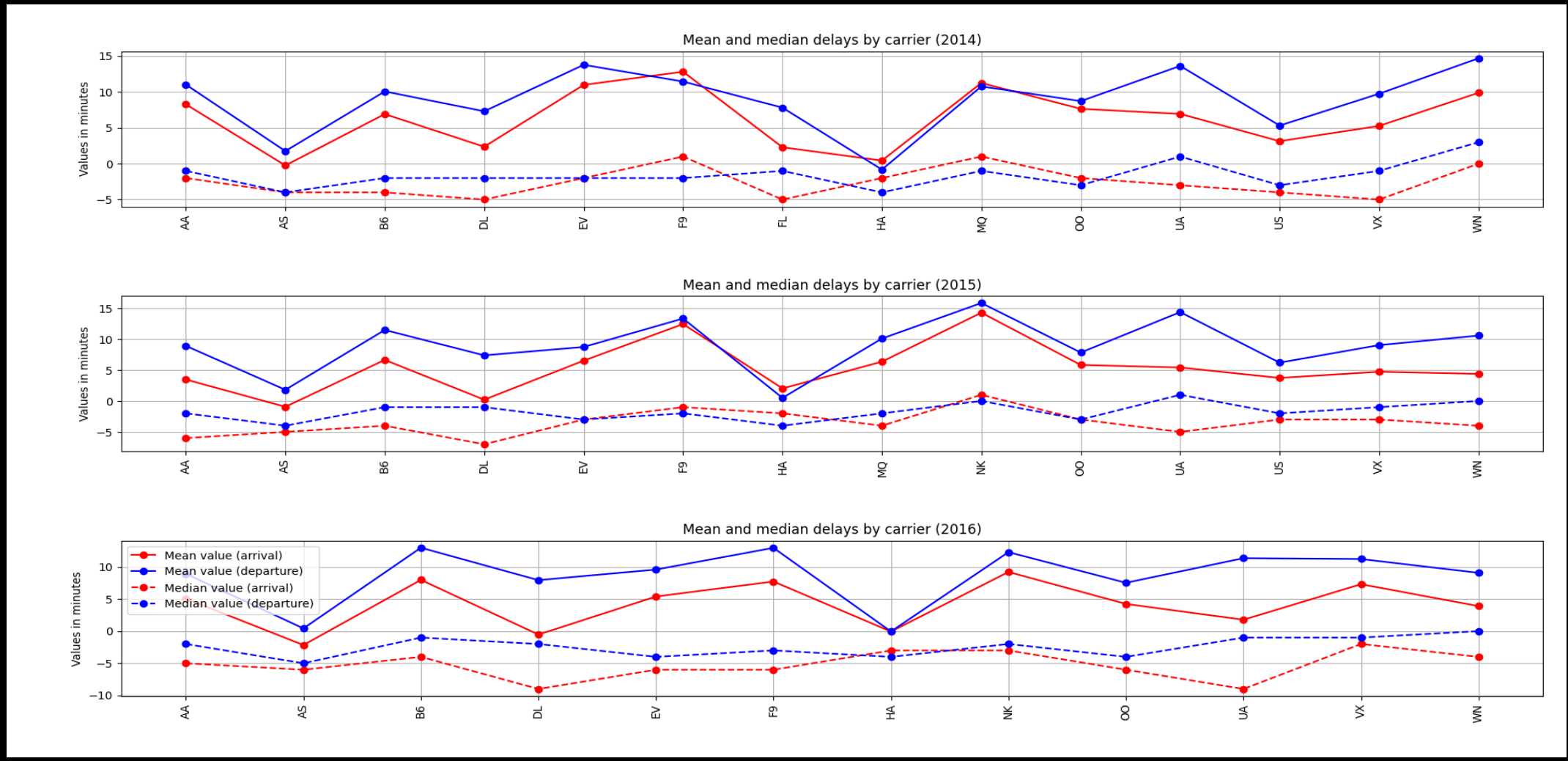


Mean and median (Week of year)

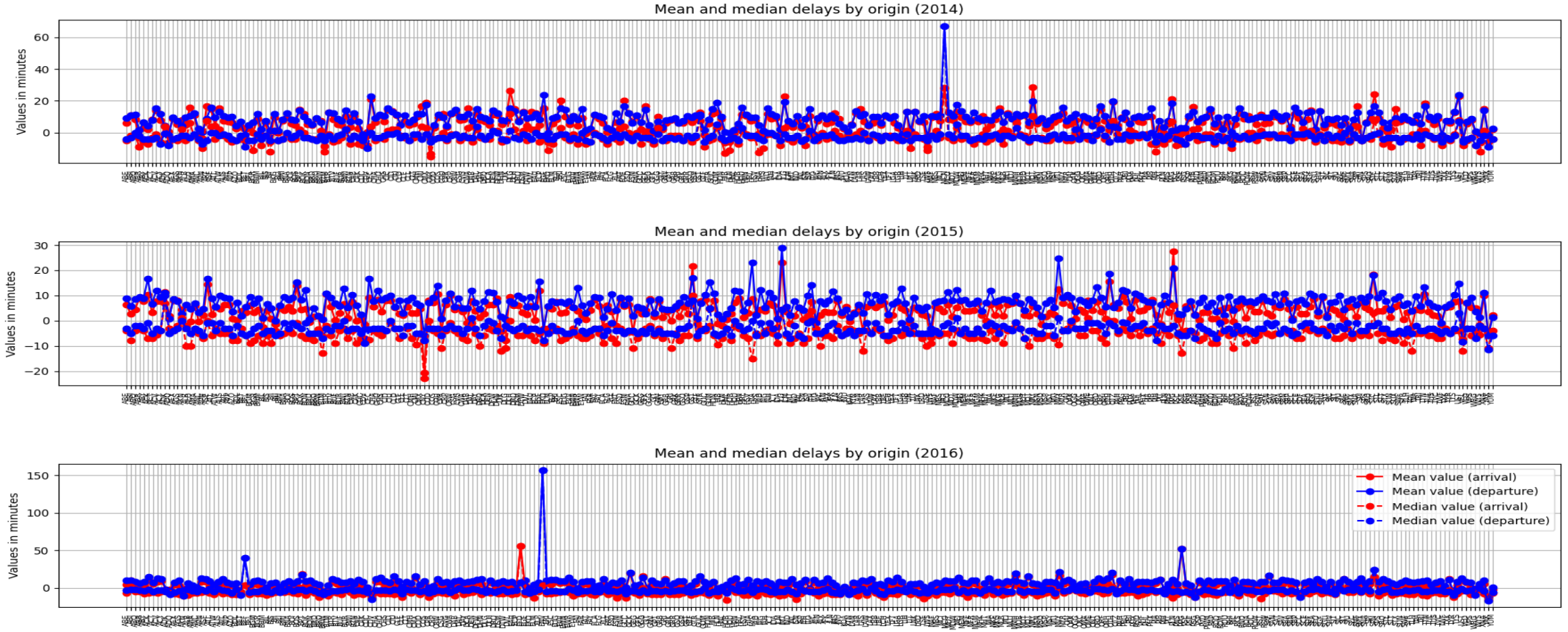




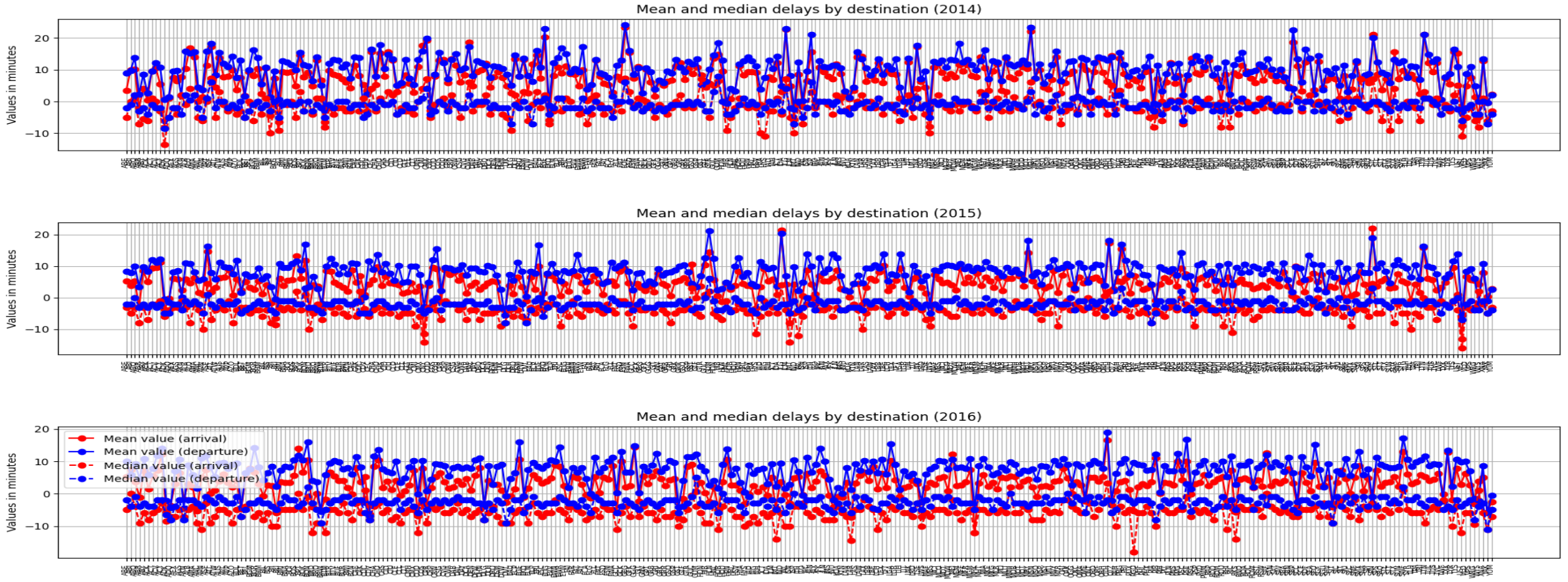
Mean and median (Carrier)



Mean and median (Origin)



Mean and median (Destination)

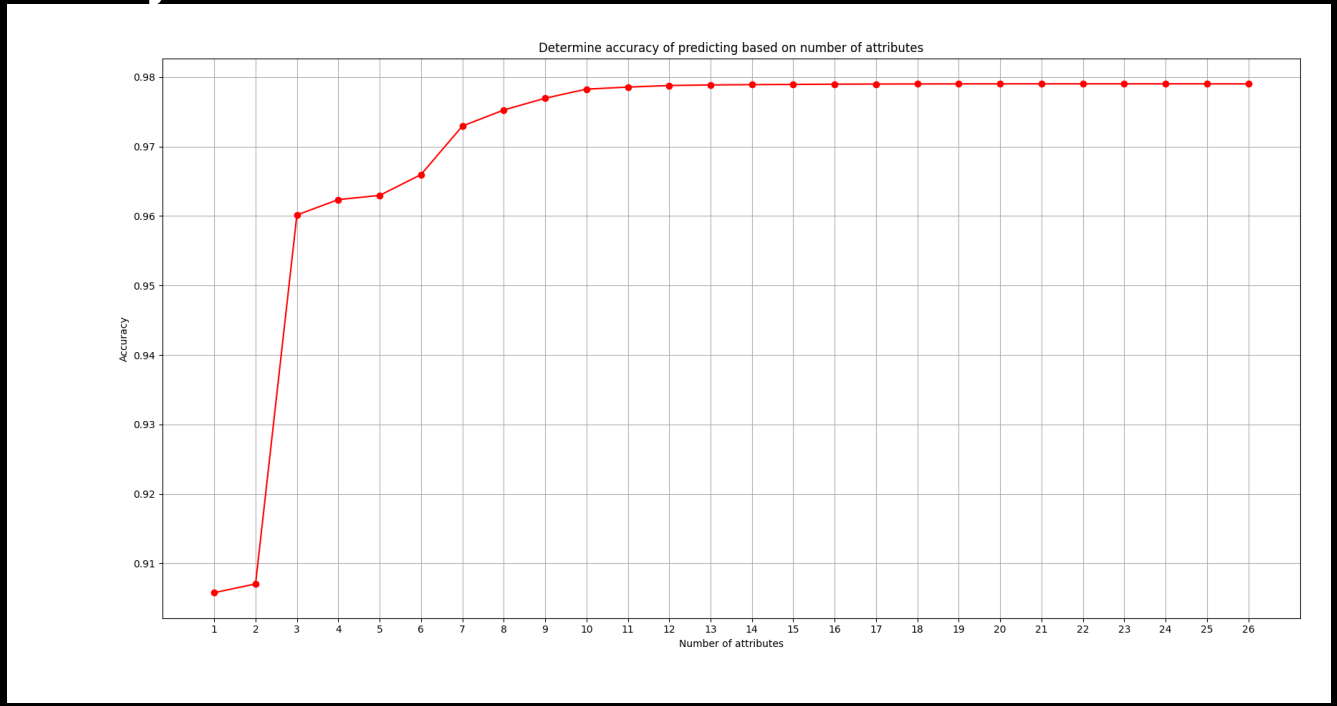


Conclusions

- Looking at the data aLooking just at the mean value, we could conclude that every flight will have some small delays.
- But, after looking at median value we conclude that most flights are actually landing slightly earlier.
- When looking at the results by year, we can conclude that average delay time is reducing, and median value is showing that more flights arrive before scheduled time (Table 2).

Using model to predict delays in arrival

- Linear regression model from sklearn library
- SBS (sequential backwards selection) is used for reducing dimensionality of our data.



Model conclusions

- After evaluating results, we can see from the graph above that after 13 dimensions accuracy of prediction stay the same.
- 13 attributes that are most important for prediction are (OP_CARRIER_FL_NUM, TAXI_OUT, CANCELATION_CODE, CRS_ELAPSED_TIME, ACTUAL_ELAPSED_TIME, AIR_TIME, DISTANCE, CARRIER_DELAY, WEATHER_DELAY, NAS_DELAY, SECURITY_DELAY, LATE_AIRCRAFT_DELAY, UNNAMED:27)
- Training accuracy of model (R2 score): 0.980
- Test accuracy of model (R2 score): 0.979