RTrack Assignment

CSORNAI, Gyula

February 25, 2017

Report summary

1. This assignment targets to analyze the flights dataset from the nycflights13 data package
2. First steps explore the data and get rid of unusable particles
3. After this feature engineering will happen, in which 3.1 Some variebles will be transformed to modeling format 3.2 Some variables will be created to help visualization 3.3 Certain dependencies will be visualized
4. Finally KNN modeling method will be used for prediction purposes

Exploratory data analysis

Summary statistics

Table 1: Table continues below

year	month	day	dep_time	sched_dep_time
Min. :2013	Min. : 1.000	Min. : 1.00	Min. : 1	Min. : 106
1st Qu.:2013	1st Qu.: 4.000	1st Qu.: 8.00	1st Qu.: 907	1st Qu.: 906
Median :2013	Median : 7.000	Median :16.00	Median :1401	Median :1359
Mean :2013	Mean : 6.549	Mean :15.71	Mean :1349	Mean :1344
3rd Qu.:2013	3rd Qu.:10.000	3rd Qu.:23.00	3rd Qu.:1744	3rd Qu.:1729
Max. :2013	Max. :12.000	Max. :31.00	Max. :2400	Max. :2359
NA	NA	NA	NA's :8255	NA

Table 2: Table continues below

dep_delay	arr_time	sched_arr_time	arr_delay	carrier
Min. : -43.00	Min. : 1	Min. : 1	Min. : -86.000	Length:336776
1st Qu.: -5.00	1st Qu.:1104	1st Qu.:1124	1st Qu.: -17.000	Class :character
Median : -2.00	Median :1535	Median :1556	Median : -5.000	Mode :character
Mean : 12.64	Mean :1502	Mean :1536	Mean : 6.895	NA
3rd Qu.: 11.00	3rd Qu.:1940	3rd Qu.:1945	3rd Qu.: 14.000	NA
Max. :1301.00	Max. :2400	Max. :2359	Max. :1272.000	NA
NA's :8255	NA's :8713	NA	NA's :9430	NA

Table 3: Table continues below

flight	tailnum	origin	dest	air_time
Min. : 1	Length:336776	Length:336776	Length:336776	Min. : 20.0
1st Qu.: 553	Class :character	Class :character	Class :character	1st Qu.: 82.0
Median :1496	Mode :character	Mode :character	Mode :character	Median :129.0
Mean :1972	NA	NA	NA	Mean :150.7
3rd Qu.:3465	NA	NA	NA	3rd Qu.:192.0

flight	tailnum	origin	dest	air_time
Max. :8500	NA	NA	NA	Max. :695.0
NA	NA	NA	NA	NA's :9430

distance	hour	minute	time_hour
Min. : 17	Min. : 1.00	Min. : 0.00	Min. :2013-01-01 05:00:00
1st Qu.: 502	1st Qu.: 9.00	1st Qu.: 8.00	1st Qu.:2013-04-04 13:00:00
Median : 872	Median :13.00	Median :29.00	Median :2013-07-03 10:00:00
Mean :1040	Mean :13.18	Mean :26.23	Mean :2013-07-03 05:02:36
3rd Qu.:1389	3rd Qu.:17.00	3rd Qu.:44.00	3rd Qu.:2013-10-01 07:00:00
Max. :4983	Max. :23.00	Max. :59.00	Max. :2013-12-31 23:00:00
NA	NA	NA	NA

Elininating NA values

After dropping rows with NA value in dep_time, arr_time, dep_delay and arr_delay features, we are left with 317916 clean and usable rows.

Feature engineering

The year is 2013 for all rows, so it can be dropped from the dataset. The weekday will be transformed from the time_hour. Hour and minute variables are fundamentally the same as dep_time, which is the original departure time (see above), so for modeling it also can be dropped.

From the other side dep_time feature's hour and minute parameters need to be extracted, which will be done by the metric functions of R.

The sched_arr_time and arr_time will be processed similarly.

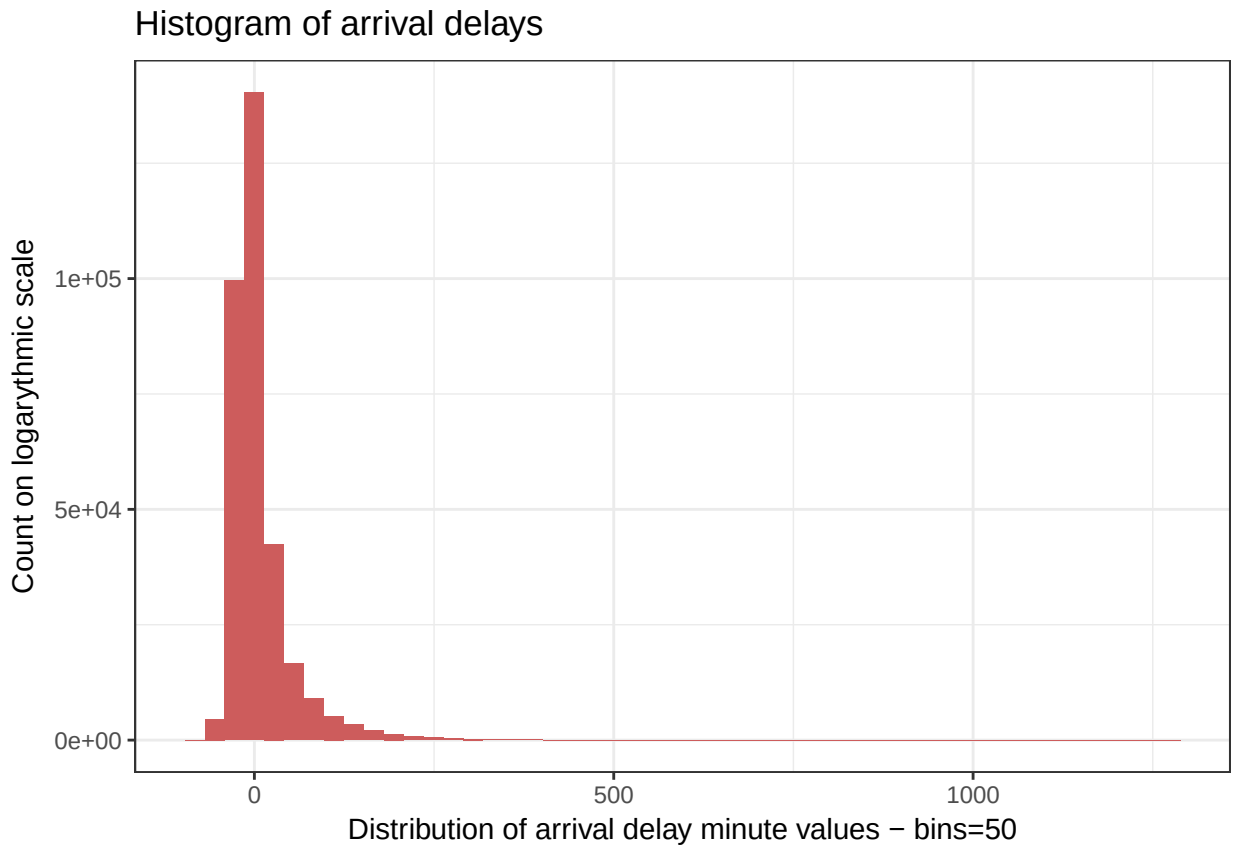
Data visualization

Distribution of delays on arrival side

The following chart shows the histogram of arrival delays with 50 bins. Later on this will be the predicted variable. The distribution is heavily skewed, most of the flights arrive with less than 15m delay, or even arrive in time.

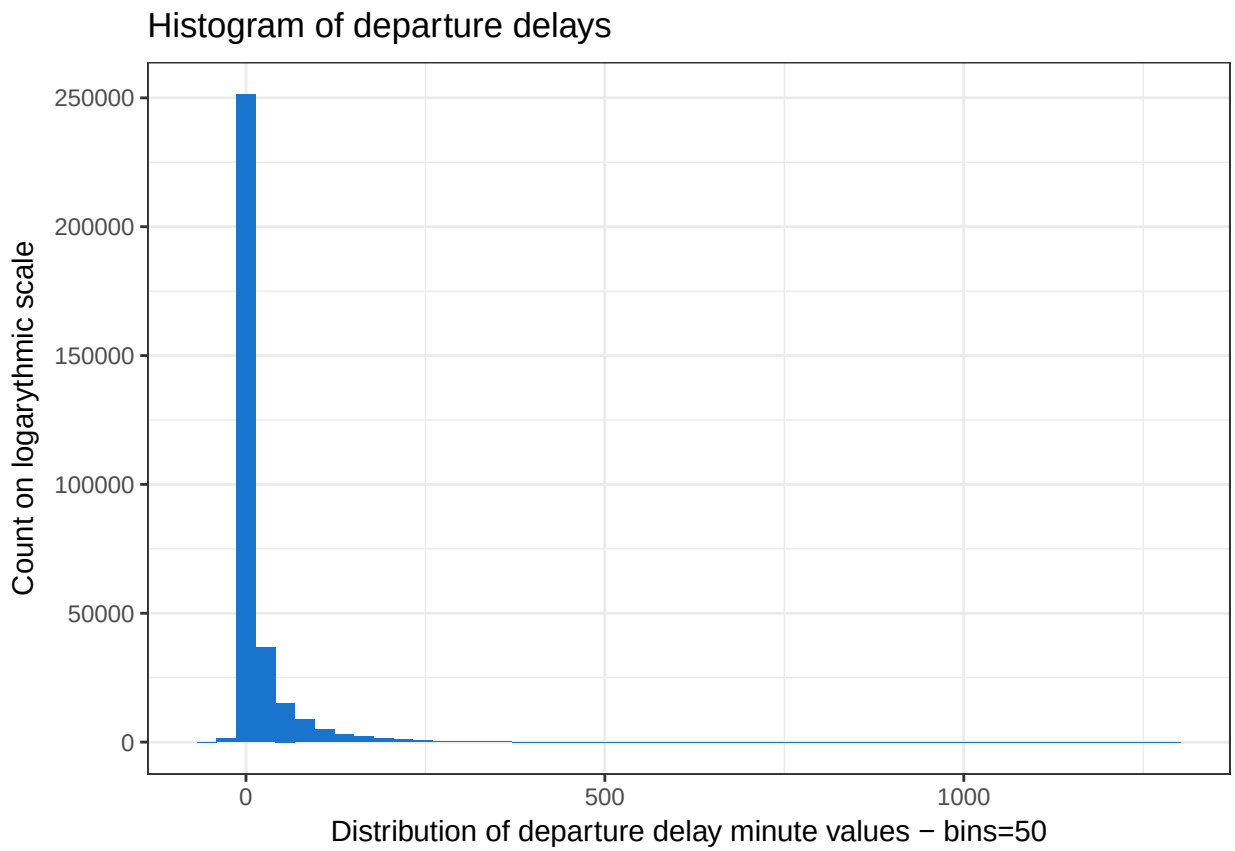
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-86	-17	-5	6.895	14	1272

The following chart shows the numbers on a logarythmic Y-axis.



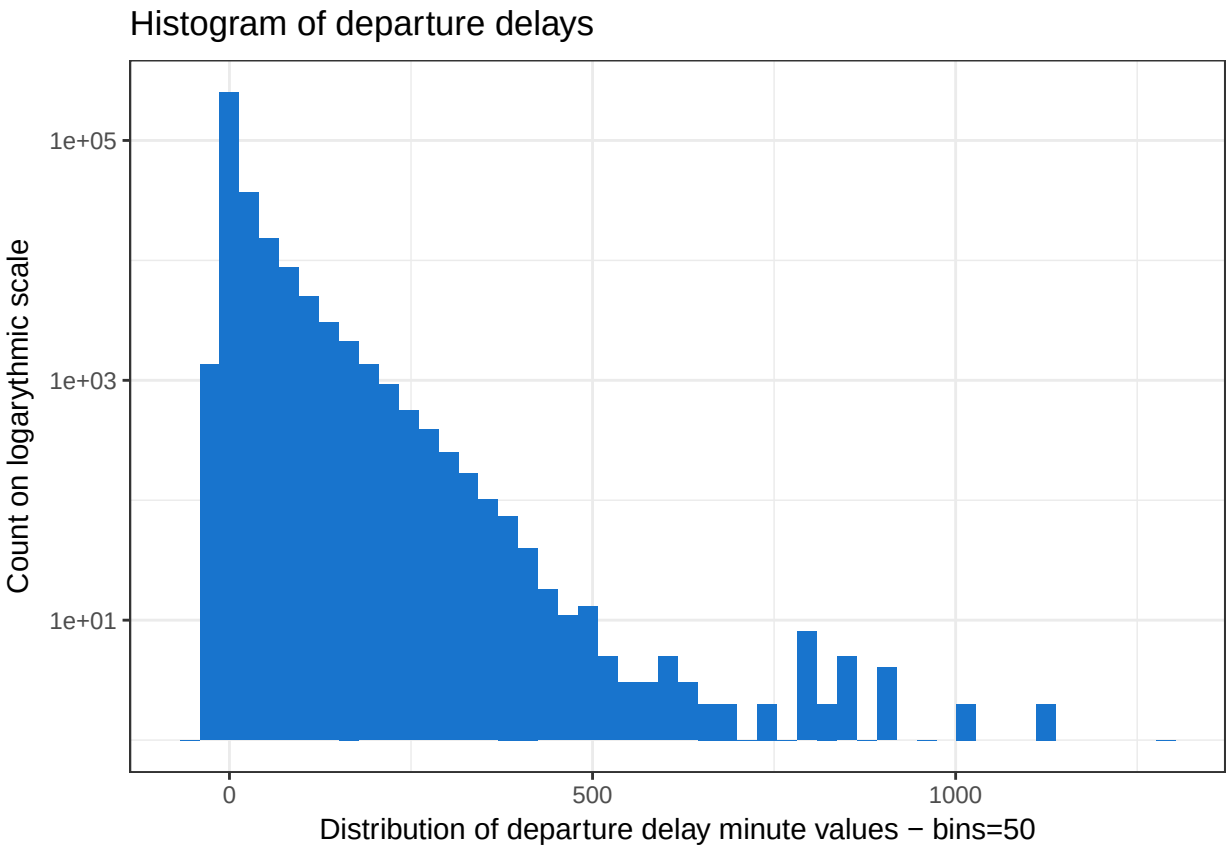
Distribution of delays on departure side

The following chart shows the histogram of arrival delays with 50 bins. Later on this will be one of the predictor variable.



The departure delays show a much more exotic distirbution, due to the fact, that there are a lot of flight departing on time, the mean is interestingly bigger than the 3rd quantile.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-43	-5	-2	12.56	11	1301

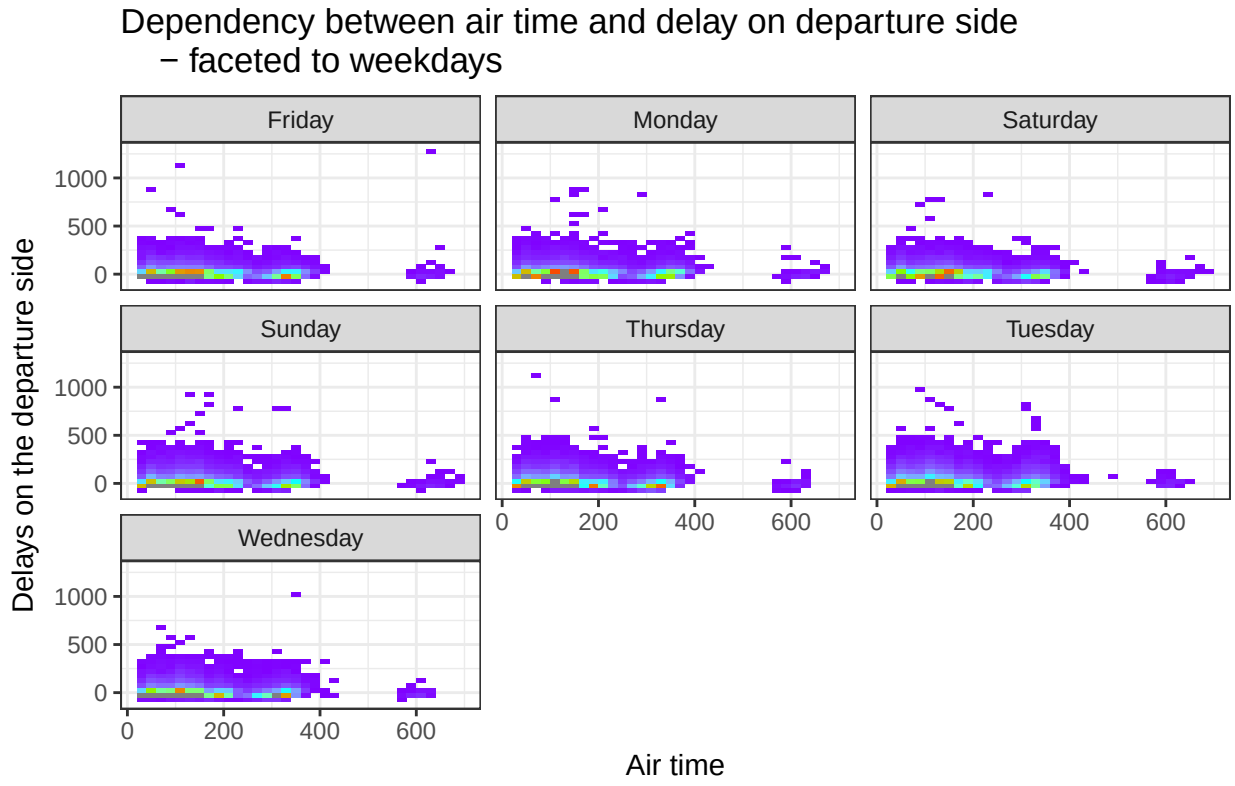


Dependency between air time and arrival side delay on different day of the week

The following chart shows the dependency between the air times and the arrival side delays of different flights. The color code of the point reflects the day when the flight has travelled. As the chart show, there is no dependency between air time and arrival side delay, from the other side Monday flights tend to arrive earlier.

It gives the following insights:

- The delay on the arrival side is less dependent on the air time, than on the day of week - on some days, we need to expect more delays in general (i.e. Tuesdays and Sundays)
- Long flights have a much better expected duration than medium duration flights
- The flights under approximately 150 minutes air time will typically not be delayed (grey area shows a huge mass of data)



o be interpreted as a heatmap with high numbers in the red range and lower numbers in the violet range. Grey areas need to be interpreted as higher numbers than 2000.

Heatmap of the arrival delays

The following chart shows the heatmap of the arrival delays broken down to hour of the day and day of the week.

The following code included shows that I have created 50 categories initially, to show equal categories in the arrival delay range. However due to the structure of the data, it has thrown error, that it cannot create 50 equal bins and defaulted back to 11.

As I personally cannot use blue heatmap well, I have switched to spectral palette.

```
ncolor <- 11
cuts<- split(dfm$arr_delay, cut_number(dfm$arr_delay, ncolor))
for (i in 1:ncolor) {
  dfm$adcat[sapply(cuts[i], is.element, el = dfm$arr_delay) ]<-i
}

ggplot(dfm) + geom_tile(aes(x=hour, y=weekday, fill=adcat)) +
  labs(x="Hour of the day", y="Day of the week",
       title="Flight delay heatmap")+
  scale_fill_distiller(palette = "Spectral")+
  theme_bw()
```

Modeling

In the next section I am going to use knn modeling with 2 different parameters: k=2 and k=10

Some more feature engineering

I get rid of the categorization variable, which I created only for visualization, it does not enrich the data with information.

I also transform non-numeric features to numeric.

Finally I create a binary outcome variable of arrival side delay, which shows whether or not the flight was delayed by more than 15 minutes on the arrival side.

Table 7: Table continues below

month	day	dep_delay	flight	air_time
Min. : 1.000	Min. : 1.00	Min. : -43.00	Min. : 1	Min. : 20.0
1st Qu.: 4.000	1st Qu.: 8.00	1st Qu.: -5.00	1st Qu.: 544	1st Qu.: 82.0
Median : 7.000	Median :16.00	Median : -2.00	Median :1467	Median :129.0
Mean : 6.565	Mean :15.74	Mean : 12.56	Mean :1943	Mean :150.7
3rd Qu.:10.000	3rd Qu.:23.00	3rd Qu.: 11.00	3rd Qu.:3412	3rd Qu.:192.0
Max. :12.000	Max. :31.00	Max. :1301.00	Max. :8500	Max. :695.0

Table 8: Table continues below

distance	hour	minute	deph	depm
Min. : 80	Min. : 5.00	Min. : 0.00	Min. : 0.00	Min. : 0.00
1st Qu.: 509	1st Qu.: 9.00	1st Qu.: 8.00	1st Qu.: 9.00	1st Qu.:16.00
Median : 888	Median :13.00	Median :29.00	Median :14.00	Median :31.00
Mean :1048	Mean :13.14	Mean :26.23	Mean :13.17	Mean :31.75
3rd Qu.:1389	3rd Qu.:17.00	3rd Qu.:44.00	3rd Qu.:17.00	3rd Qu.:49.00
Max. :4983	Max. :23.00	Max. :59.00	Max. :24.00	Max. :59.00

Table 9: Table continues below

sch_ah	sch_am	arrh	arrm	isMT15
Min. : 0.00	Min. : 0.00	Min. : 0.00	Min. : 0.00	Min. :0.0000
1st Qu.:11.00	1st Qu.:14.00	1st Qu.:11.00	1st Qu.:14.00	1st Qu.:0.0000
Median :15.00	Median :30.00	Median :15.00	Median :29.00	Median :0.0000
Mean :15.04	Mean :29.04	Mean :14.72	Mean :29.47	Mean :0.2371
3rd Qu.:19.00	3rd Qu.:45.00	3rd Qu.:19.00	3rd Qu.:45.00	3rd Qu.:0.0000
Max. :23.00	Max. :59.00	Max. :24.00	Max. :59.00	Max. :1.0000

carrierfN	originfN	destfN	wdN
Min. : 1.00	Min. :1.000	Min. : 1.00	Min. :1.000
1st Qu.: 4.00	1st Qu.:1.000	1st Qu.: 28.00	1st Qu.:2.000
Median : 6.00	Median :2.000	Median : 50.00	Median :4.000
Mean : 7.15	Mean :1.951	Mean : 49.66	Mean :4.032
3rd Qu.:12.00	3rd Qu.:3.000	3rd Qu.: 71.00	3rd Qu.:6.000
Max. :16.00	Max. :3.000	Max. :104.00	Max. :7.000

2-Nearest neighbors model results

The following confusion table shows the the true negative, false negative, false positive, true positive predictions of the model using the 2 nearest neighbor based classification model.

	0	1
0	70689	3380
1	5787	18348

10-Nearest neighbor model results

The following confusion table shows the the true negative, false negative, false positive, true positive predictions of the model using the 10 nearest neighbor based classification model.

	0	1
0	72918	1151
1	7977	16158

Summarizing results

1. 2-NN model has found the good result in **90.6653497%** of the cases in total
2. 10-NN model has found the good result in **90.7050629%** of the cases in total
3. 2-NN has identified

i) less true negatives
ii) more true positives
iii) more false negative
iv) less false positive

compared to 10-NN.