Exercise 2

for Data Analysis

Valentino Lazarevic - 1223211

19.03.2024

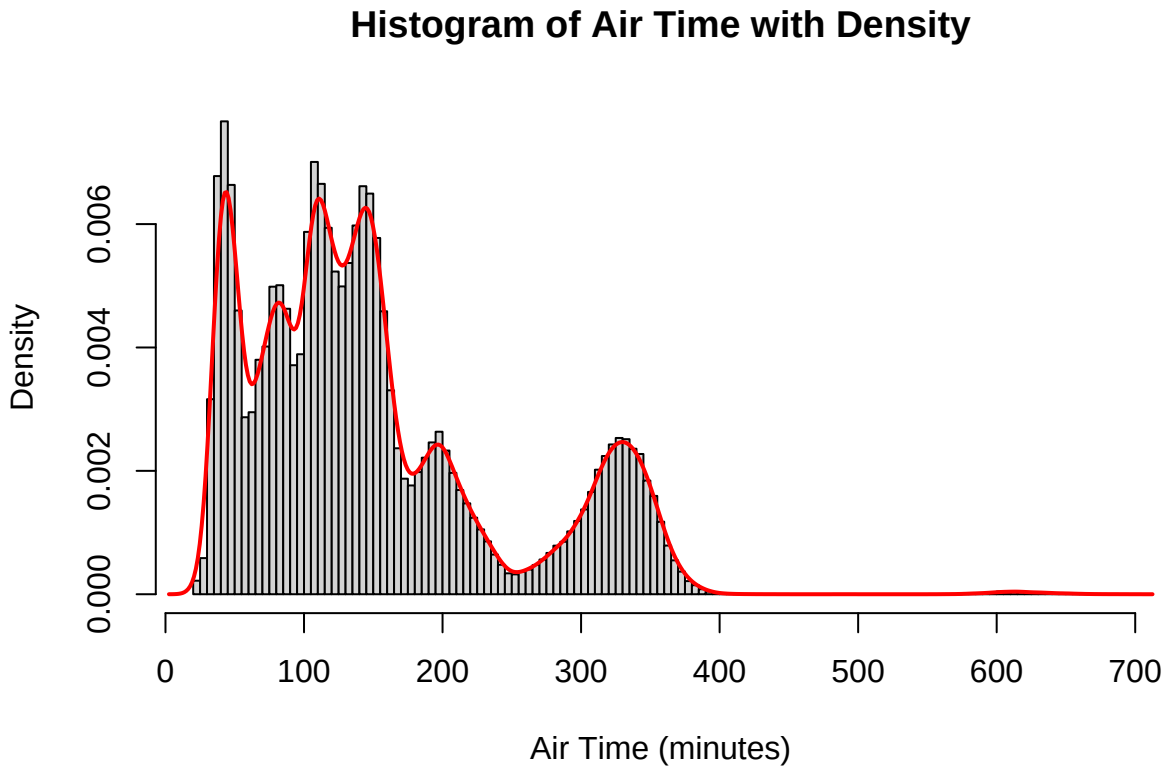
Setup

```
library(nycflights13)
library(car)
library(dplyr)
library(ggplot2)
data("flights")
attach(flights)
```

Exercise 1

Show air_time in a histogram with relative frequencies, and plot on top of the histogram the density estimate. What do you see?

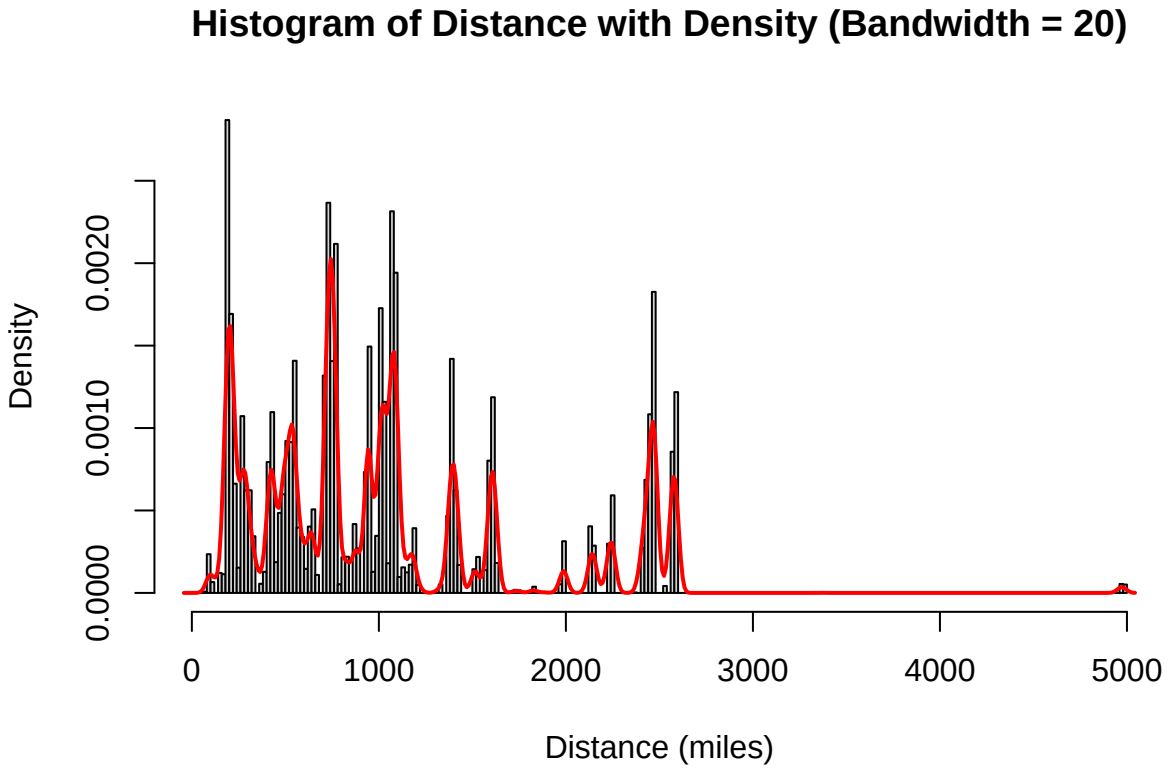
We can see that the density line is smoothly following the bars of the histogram, specifying that the density curve represents the flight durations distribution

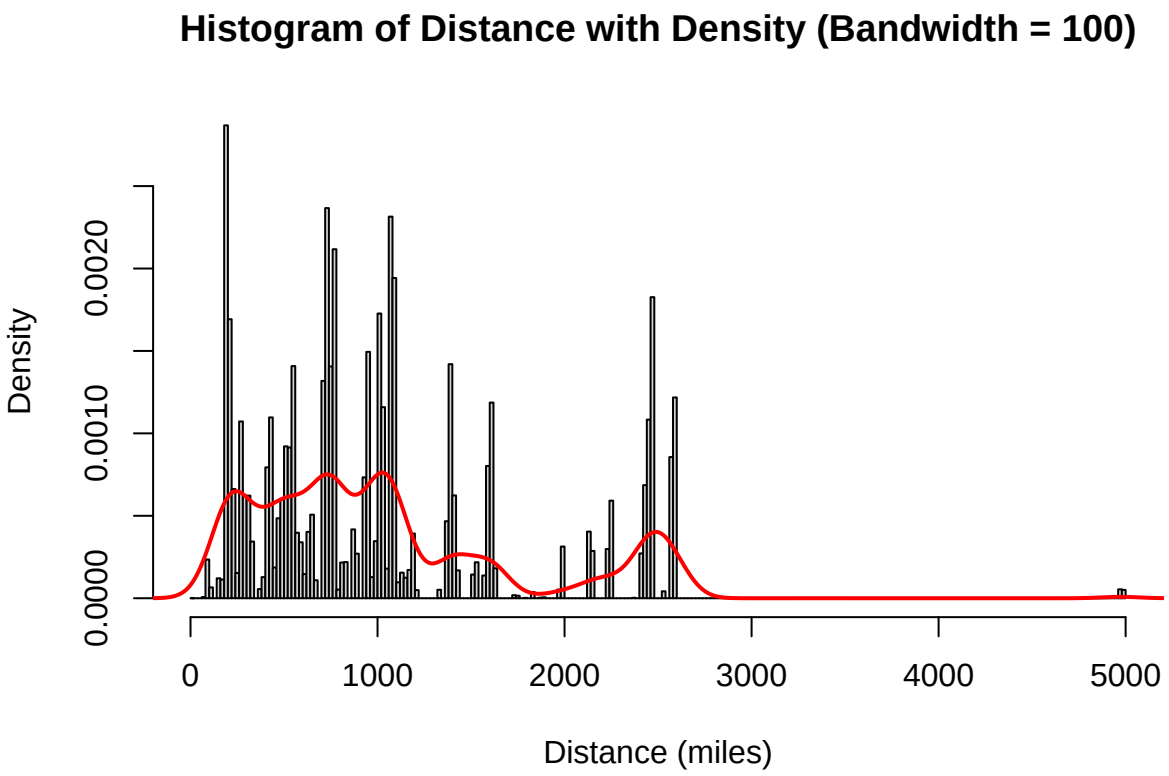


Exercise 2

Do the same as in 1. for the variable distance. What do you see?

We can see that the density line is not following the bars of the histogram, even if we change the bandwidth of the histogramm. It is leading to an over smoothing effect which not represent the underlying patterns.

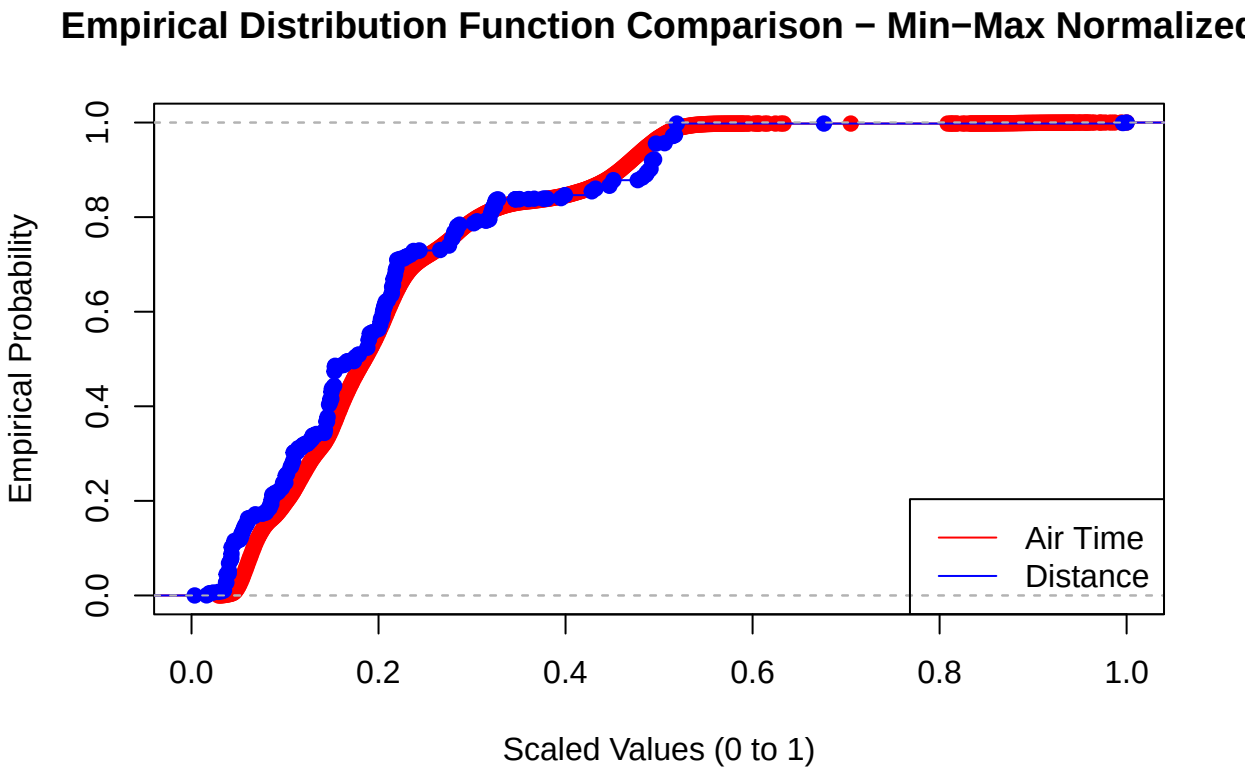



Histogram of Distance with Density (Bandwidth = 100)
Density
0.0000
0.0010
0.0020
0
1000
2000
3000
4000
5000
Distance (miles)

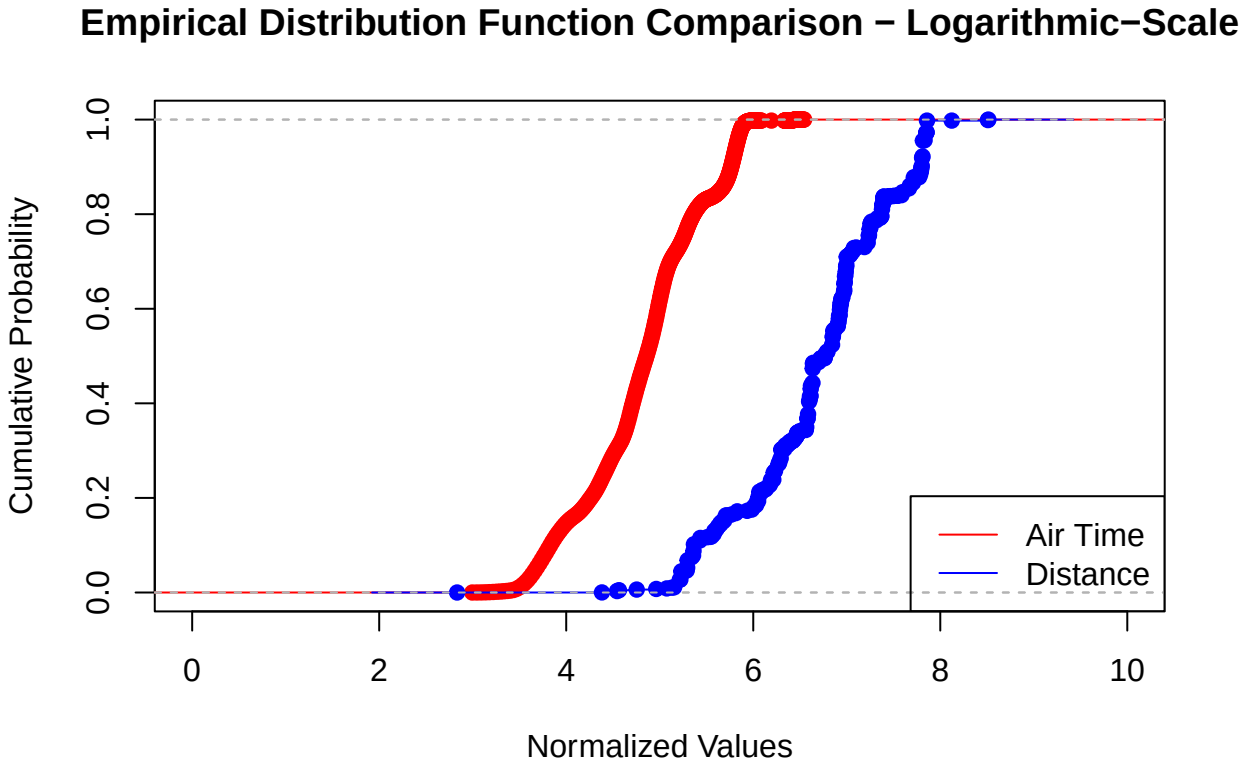
Exercise 3

Compare the empirical distribution function for air_time and distance with the function ecdf(). Try to plot the two functions on top of each other. Since the scale is very different, you should first try to normalize the variables appropriately. Which conclusions can you draw?

We can conclude a strong relationship between air_time and disatnce when using the Min-Max Normalization.



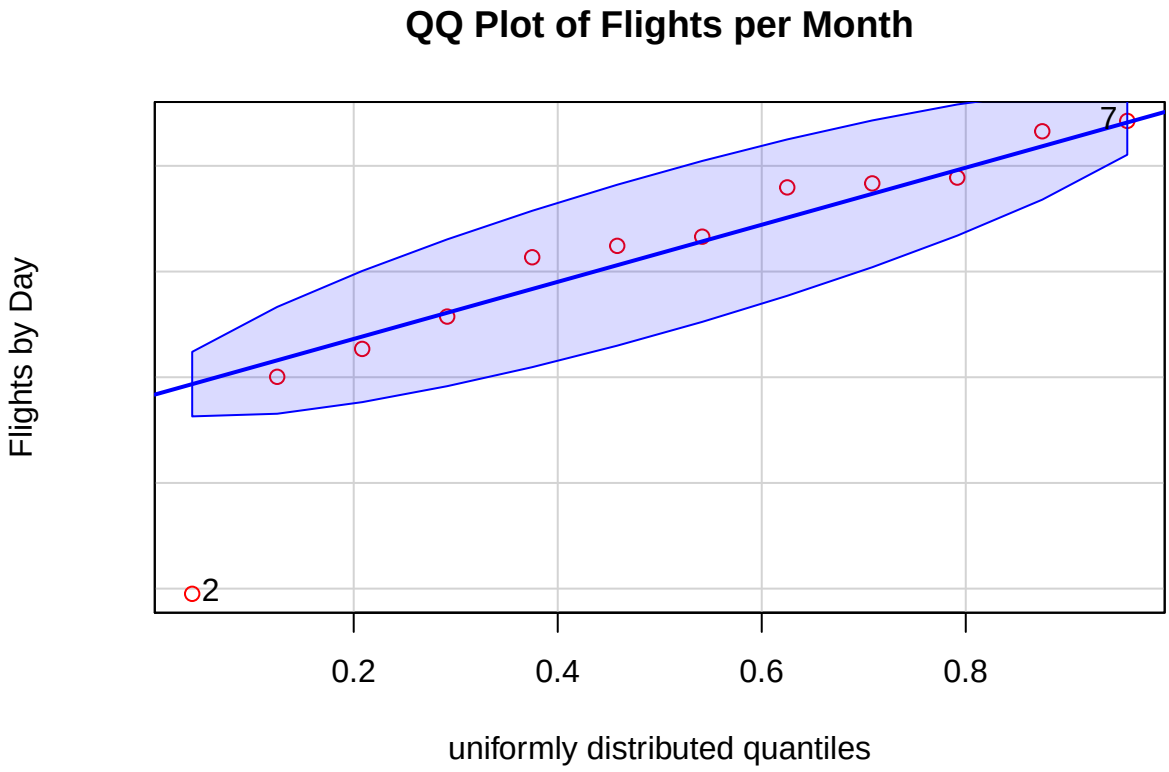
We can conclude that the lines are shaped very much alike but the distance curve is shifted to the right because the disatnce data has a far wider range than the data of air_time when using Logarithmic-Scale Normalization.



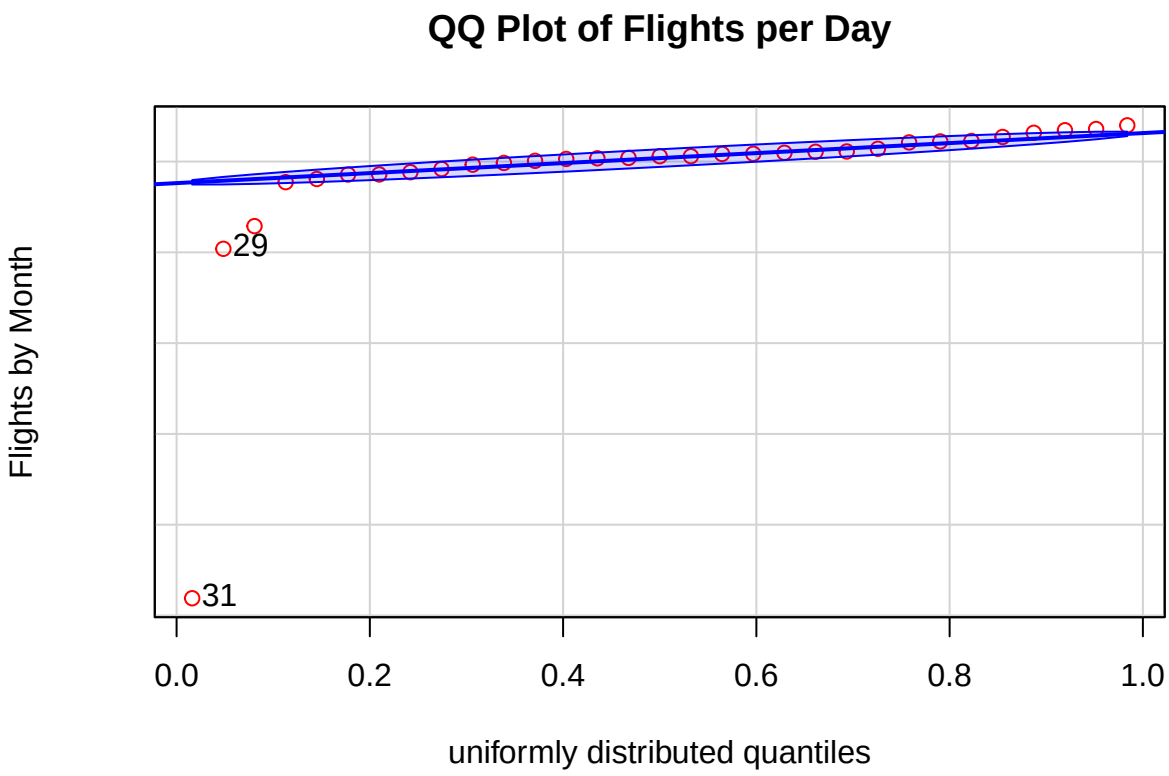
Exercise 4

Use the function qqPlot() of the package car, and check if the number of flights is uniformly distributed over the months (1–12). Do the same for the days (1–31).

The strong left tail of both plots indicates the presence of outliers or extreme values that are significantly lower than the majority of the data points, which would lead to the conclusion that our data deviate from a uniform distribution, but further testing is needed to determine whether to make a more accurate statement




```
## [1] 2 7
```

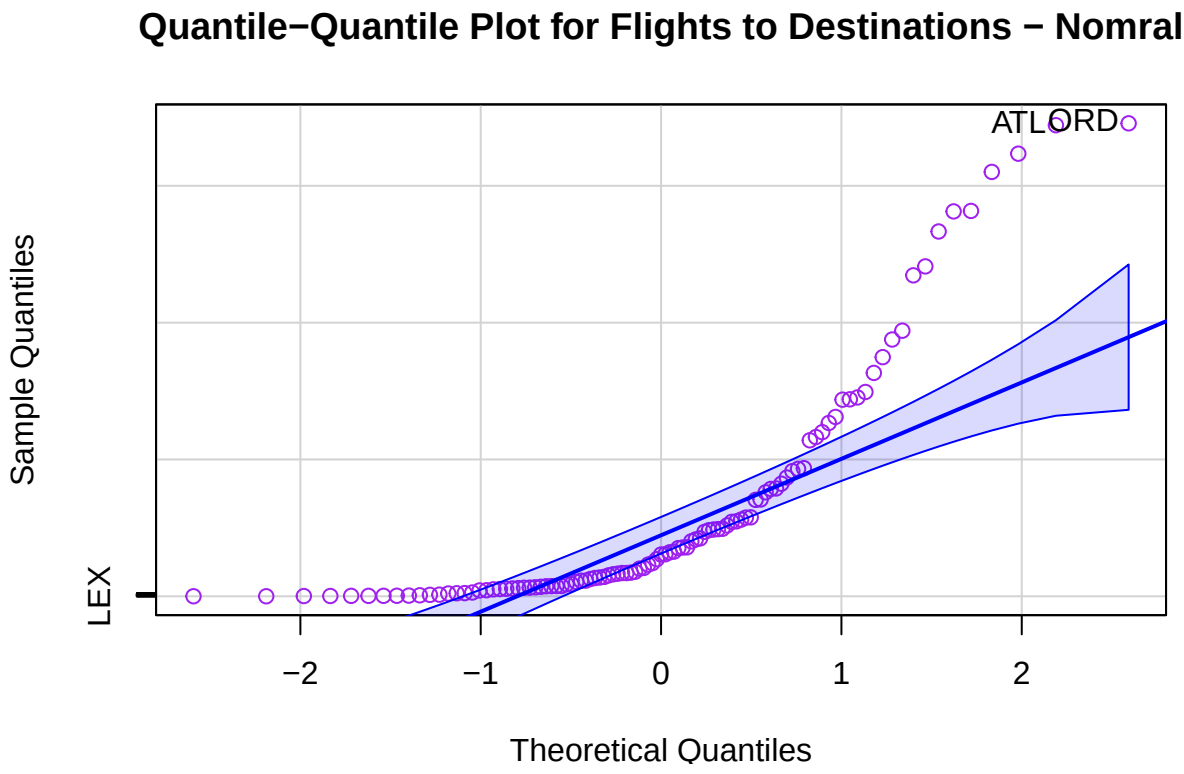

QQ Plot of Flights per Day
Flights by Month
29
31
0.0
0.2
0.4
0.6
0.8
1.0
uniformly distributed quantiles

```
## [1] 31 29
```

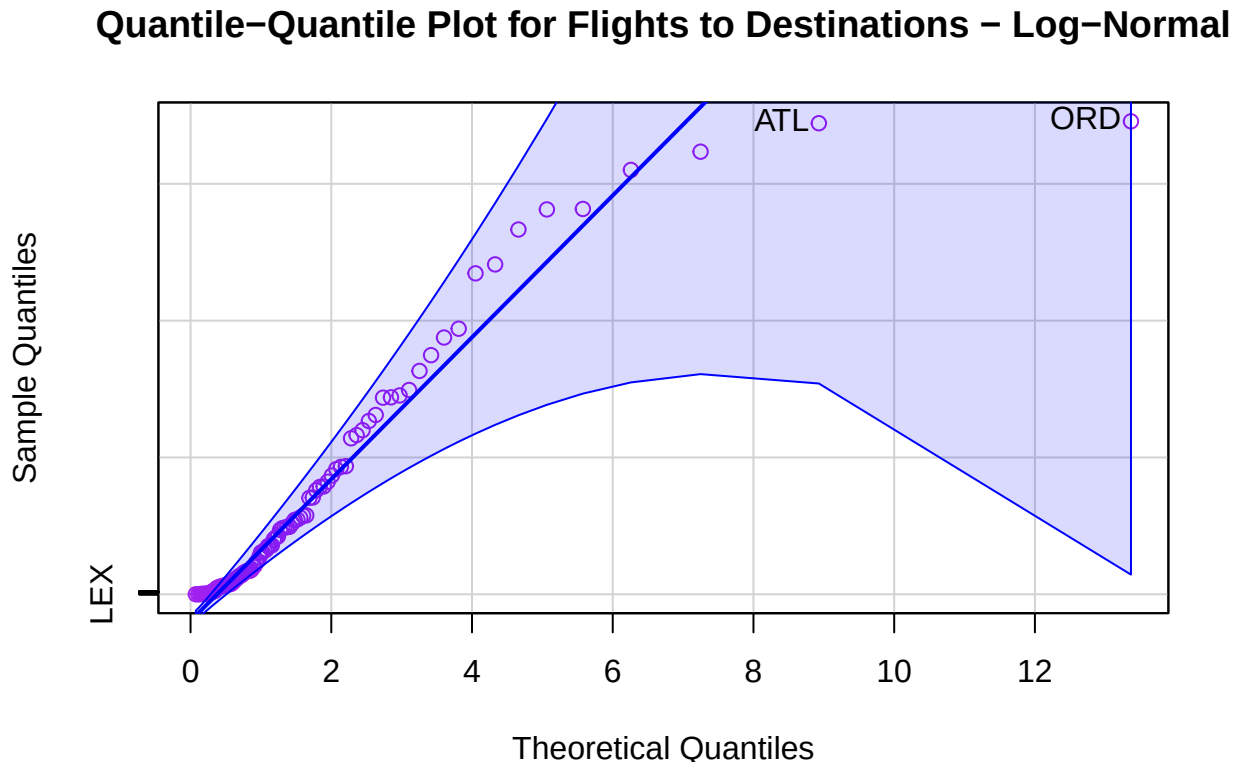
Exercise 5

Similar as before: Check if the number of flights to the different destinations follows a normal, log-normal, or exponential distribution.

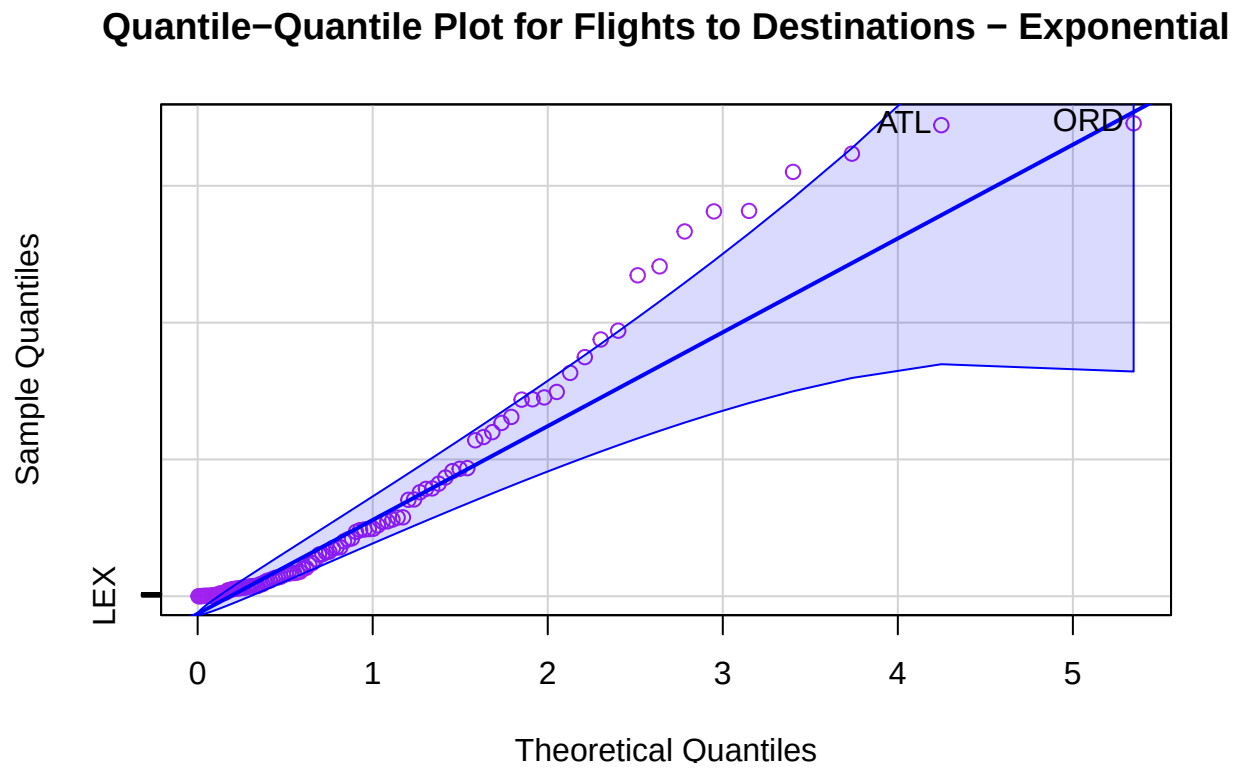
Given a distribution, the data is likely to be exponentially distributed, since both the normal and lognormal distributions deviate from base.




```
## ORD ATL
##  70   5
```




```
## ORD ATL
##  70   5
```

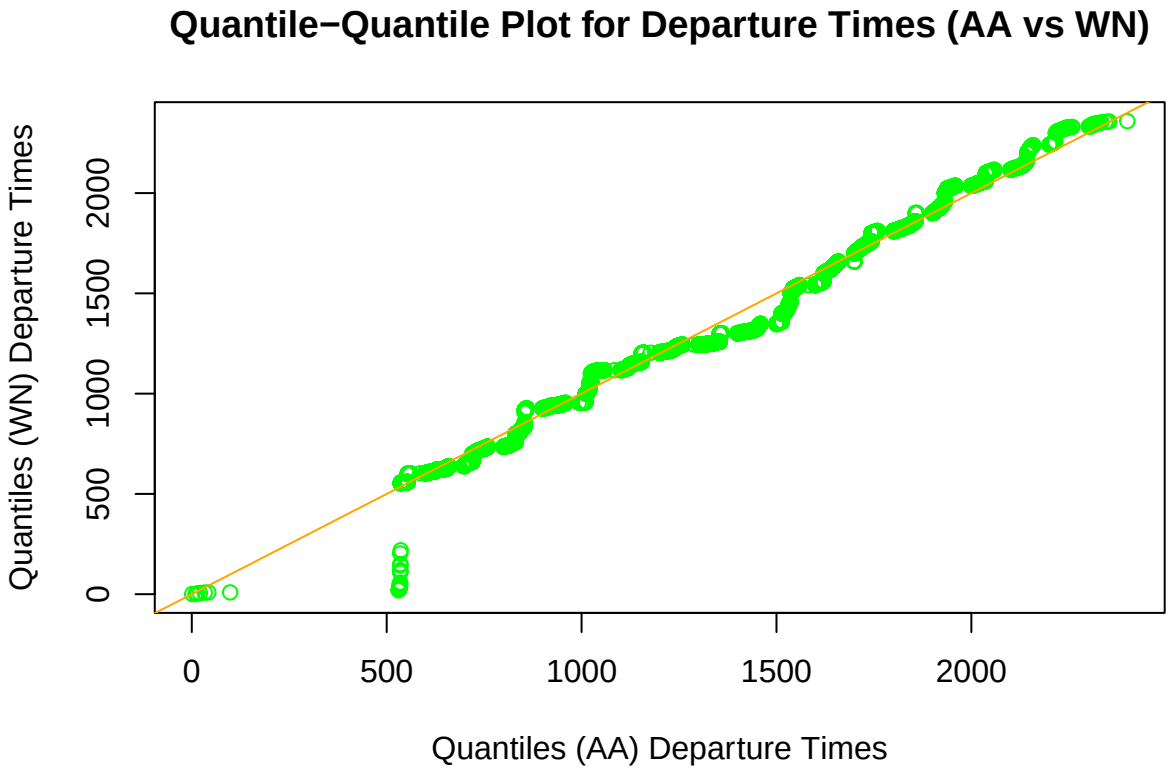



```
## ORD ATL
##  70   5
```

Exercise 6

Compare the departure times of the carriers AA and WN in a QQ-plot by usimg the function qqplot(). What can you conclude?

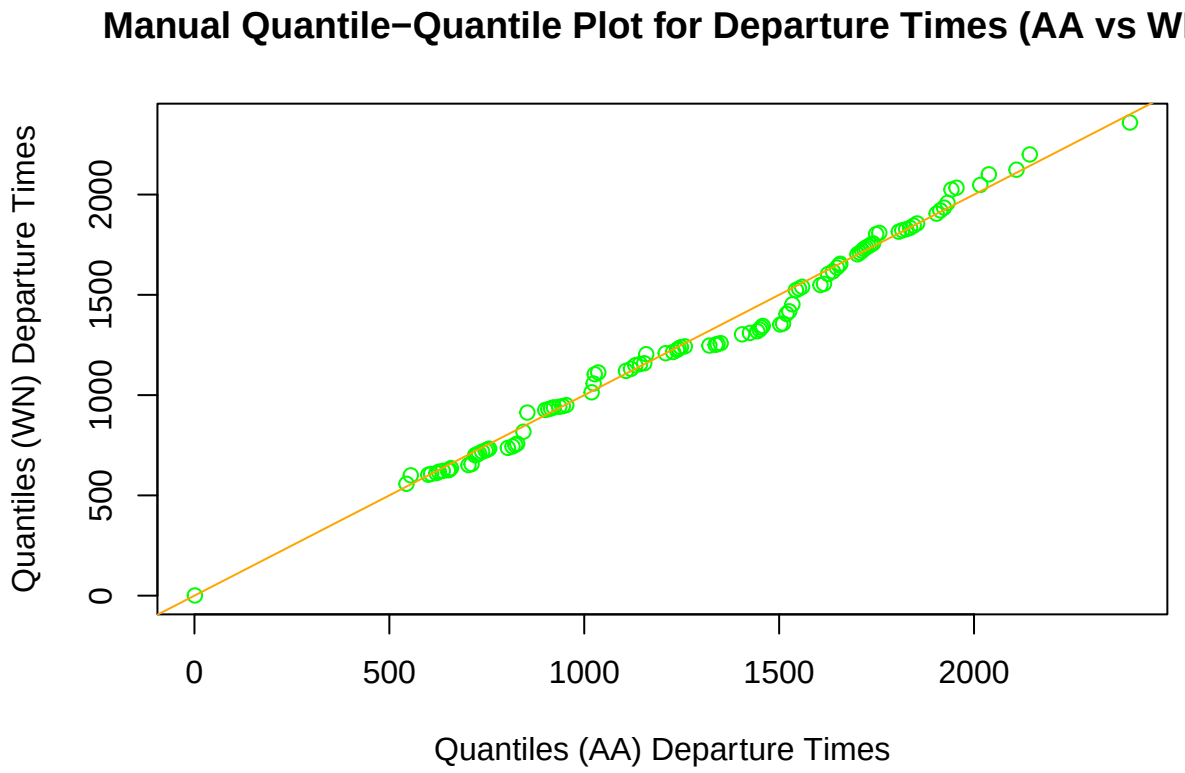
The clear similarity between the data points and the reference line indicates that the two airlines have similar distributions of departure times. Nonetheless, the number of 5:00 o'clock flights is still small, resulting in the behavior observed on the left in the Q-Q plot, as there is not enough data to draw insightful conclusions.



Exercise 7

Do the same as in 6. "by hand". I.e., plot the quantiles of the departure time of carrier AA versus those of carrier WN.

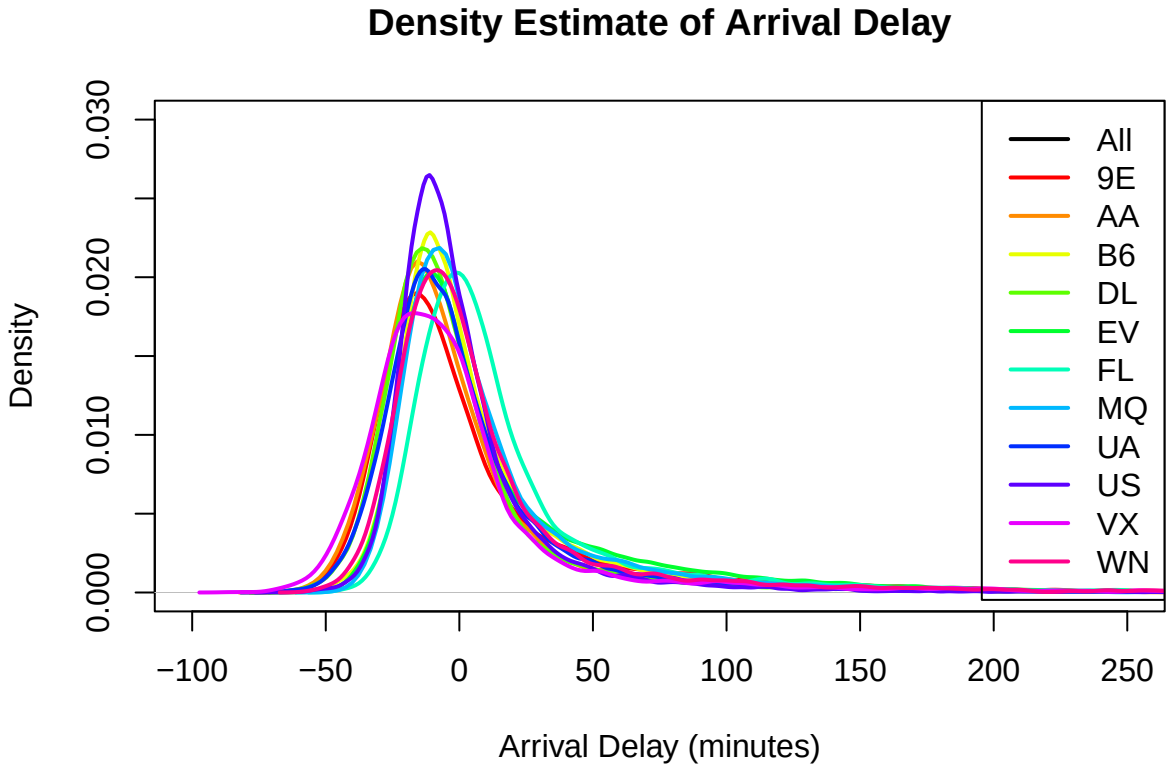
Hint: Use the function quantile()



Exercise 8

Show the density estimate of the arrival delay. Cut the x-range in the plot to just focus on the main part of the distribution. Show on top of that by differently colored lines the density estimates for the single main carriers (with at least 1000 flights). What can you conclude?

Unusually, most flights are not only on time, but even faster than expected. This could be for a variety of reasons, such as client experience or an overall conservative approach to estimating flight times.



Exercise 9

Do the same as in 8., but with QQ-plots (comparing with quantiles of the normal distribution).

We can clearly see, that all shown carriers have almost the same right-skew