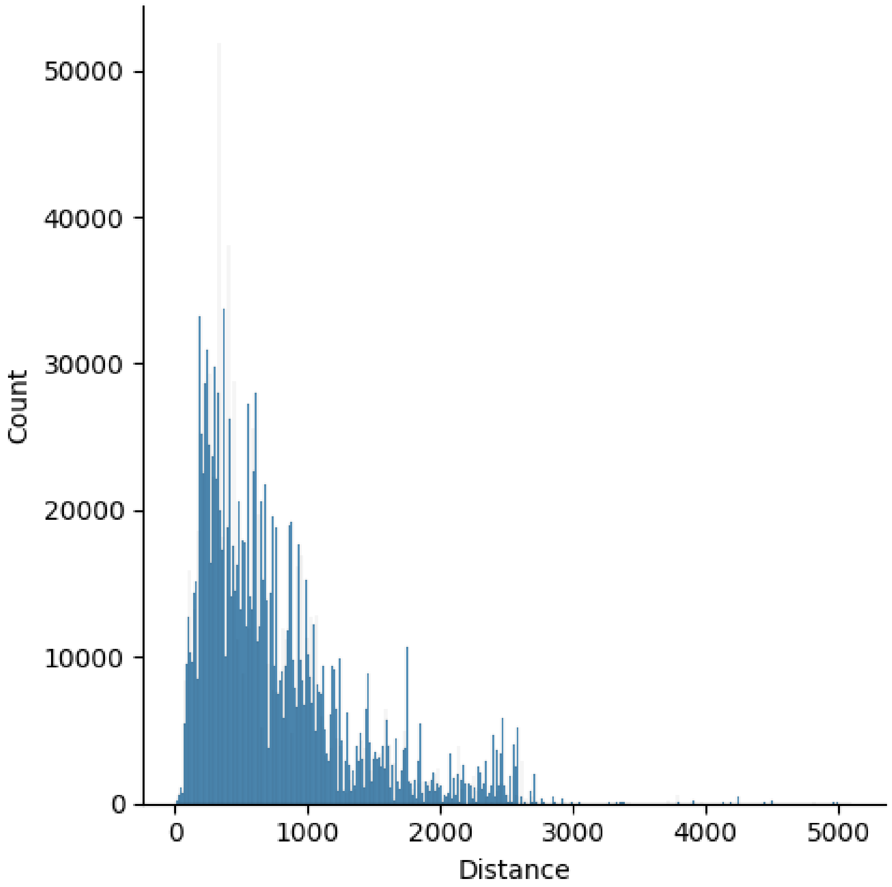
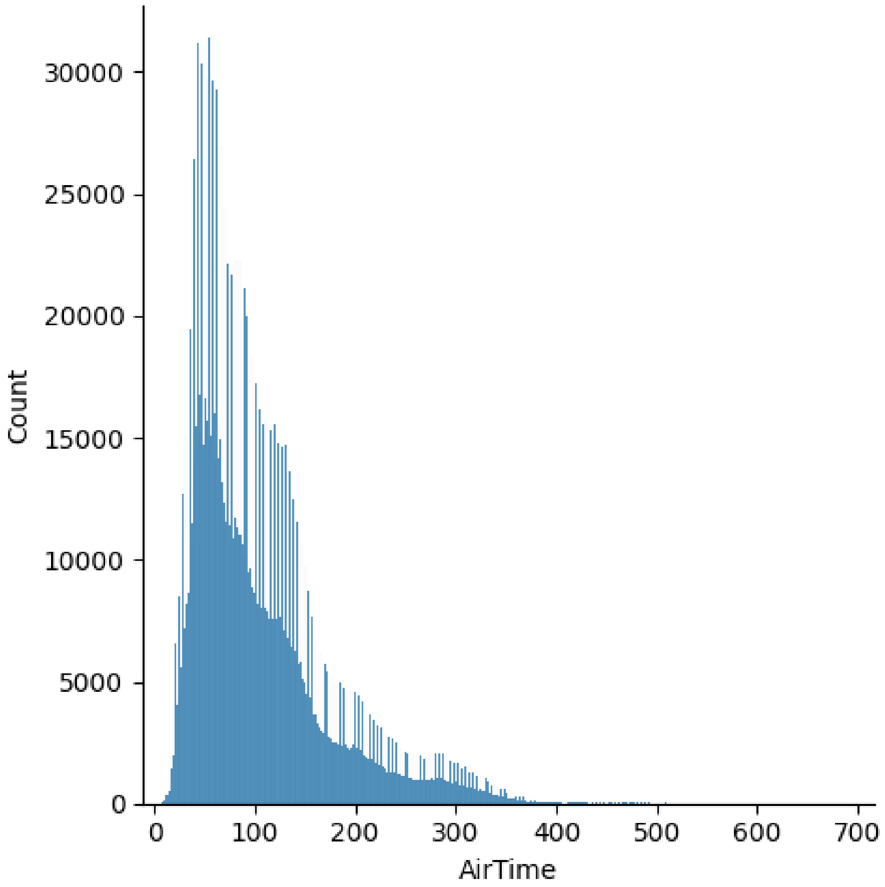
Step 3.2: Displaying histograms for some variables in the data set



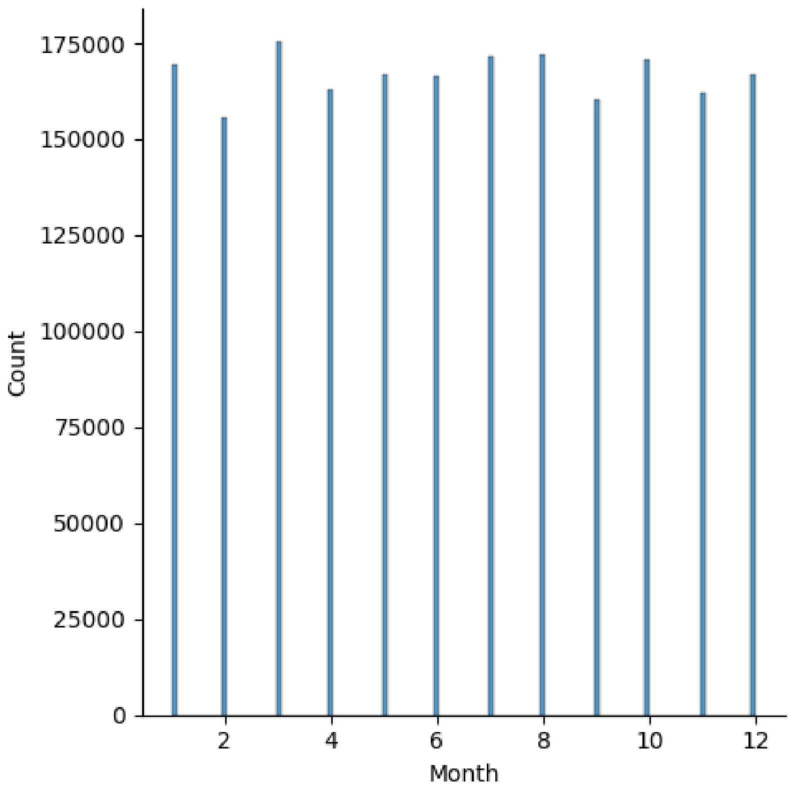
3_2-Histogram-Distance-Frequency-tiny



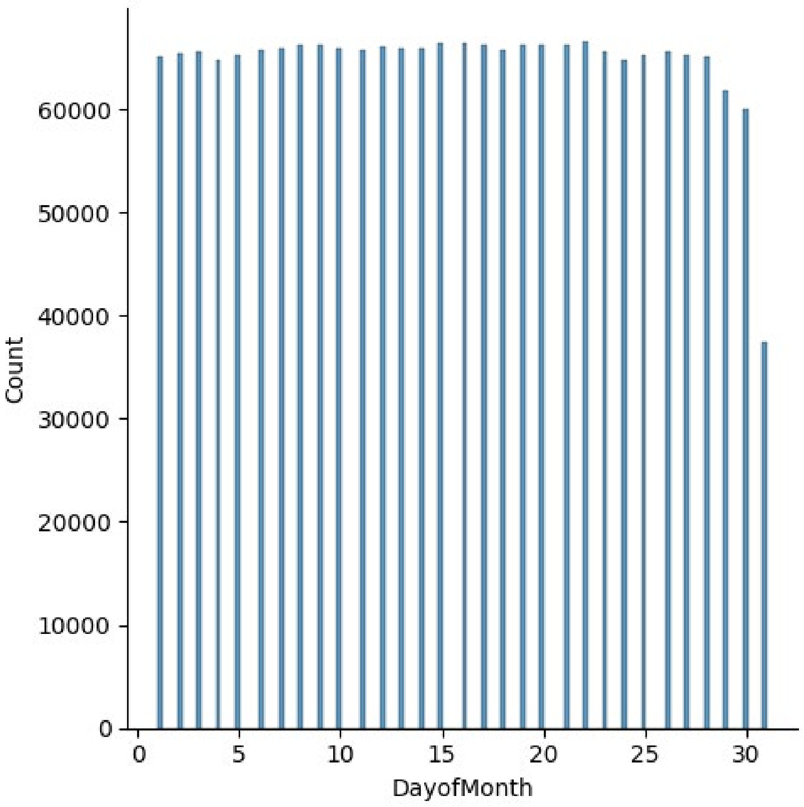
3_2-Histogram-AirTime-Frequency-tiny

Here we can see data displayed in histogram form regarding the number of flights related to the total Air Time and total Distance. This makes sense with the actual data frame, since we are analysing domestic flights within the United States, and most of those have a short distance value. I arrive to the same conclusion after plotting histogram for Air time: most flights are no longer than 200 minutes (3.3 hours).

Step 3.3 and 3.4: Identifying and removing outliers from the data frame



3_2-Histogram-Month-Frequency-tiny

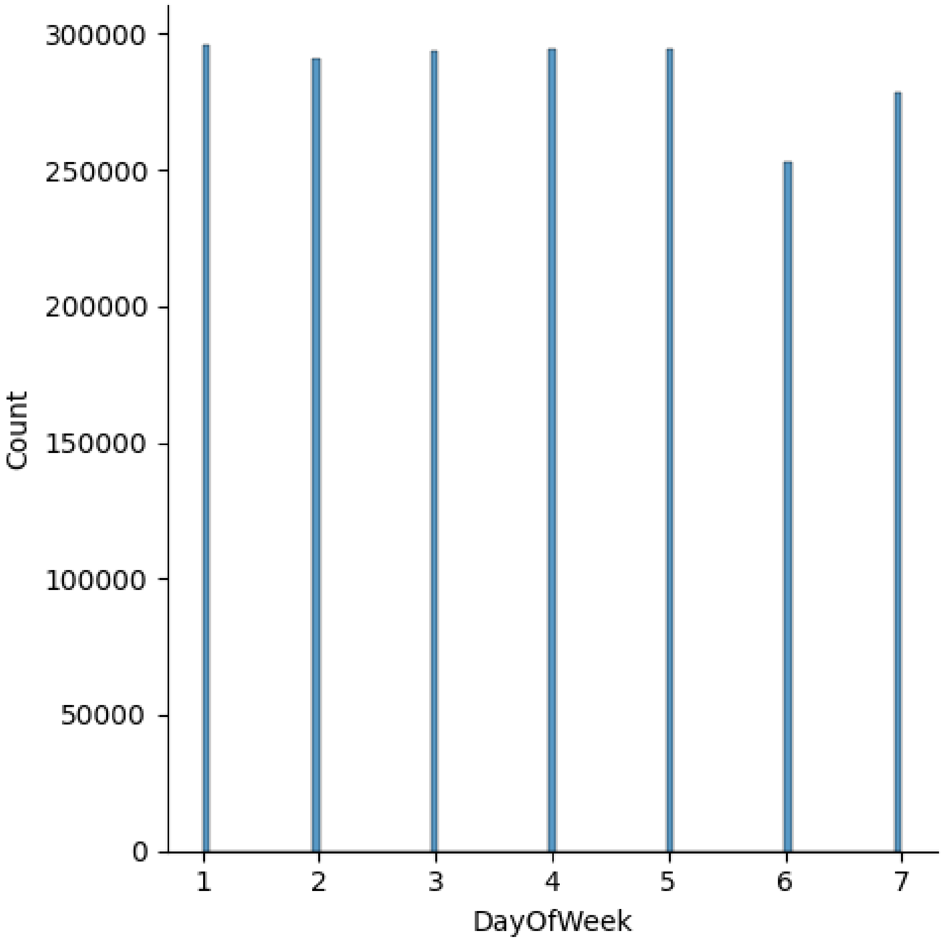


3_2-Histogram-DayOfMonth-Frequency-tiny

As a first approach to the dataset, some histograms have shown logical information, such as: there are less flights, in total, on day 31st, since not all months have 31 days.

Looking at the month frequency for the whole 2M sample dataset, there is not much information we can extract from that histogram: it is quite evenly divided, the total of flights among the 12 months in the whole period.

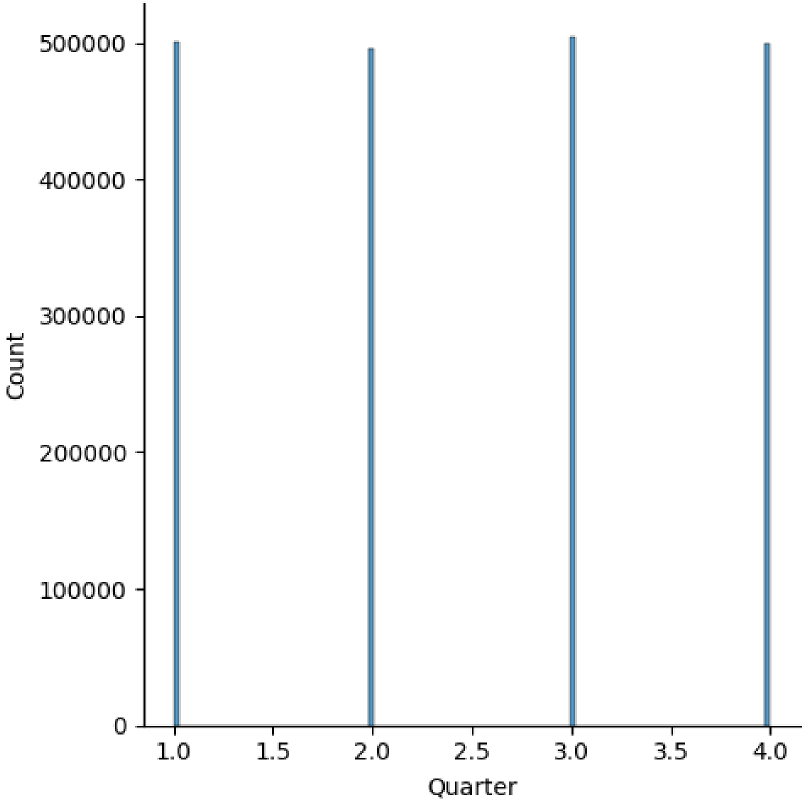
Step 3.3 and 3.4: Identifying and removing outliers from the data frame



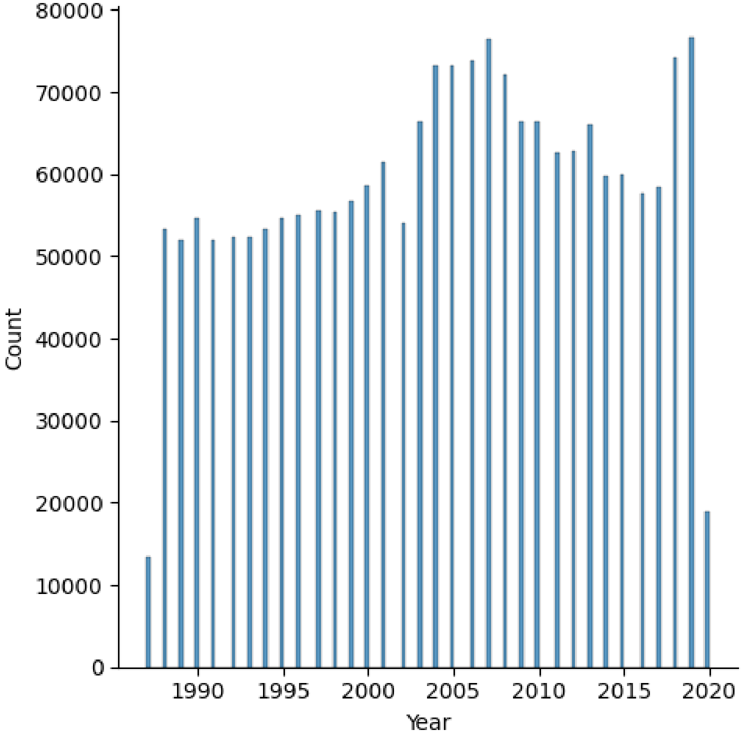
3_2-Histogram-DayOfWeek-Frequency-tiny

Saturdays seem to be a day when there are fewer domestic flights for the dataset.

Step 3.3 and 3.4: Identifying and removing outliers from the data frame



3_2-Histogram-Quarter-Frequency-tiny

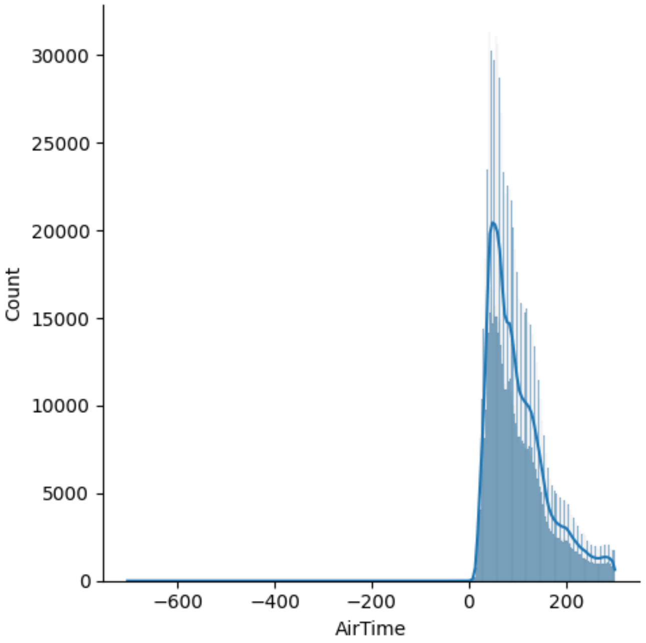


3_2-Histogram-Year-Frequency-tiny

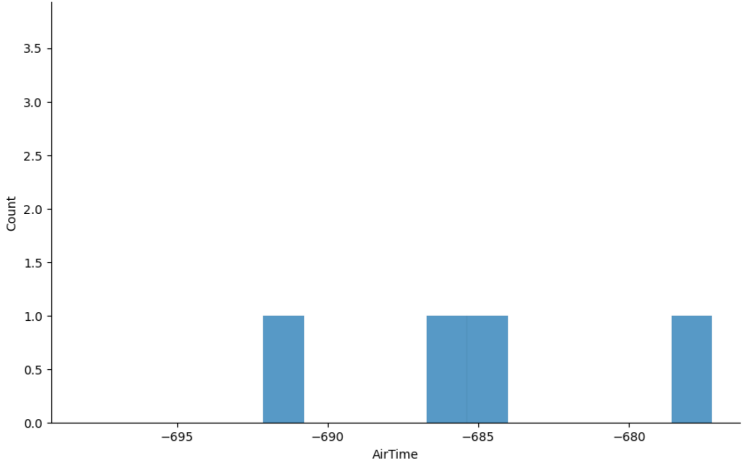
All 2 million flights are evenly distributed among the quarters of a year, for the whole period. I will analyse this in future steps, confirming there is no difference in season for what number of flights matter.

If we look at the year histogram, 1987 and 2020 shouldn't be analysed in full, since the sample starts from October 1987 and finishes in March 2020. The binary csvtool has been of great help to grab this information from the dataset.

Step 3.3 and 3.4: Identifying and removing outliers from the data frame



3_2-Histogram-AirTime-Frequency-NEG_VALUES-tiny



3_2-Histogram-AirTime-Frequency-NEGATIVE-zoom

Analysing again the AirTime histogram we can see that there are several values on the negative side of the X axis. This makes no sense to me, and I have zoomed in to confirm. Also, just in case I might have plotted the histogram wrong, I have confirmed with *csvtool* that there are, indeed, such negative values.

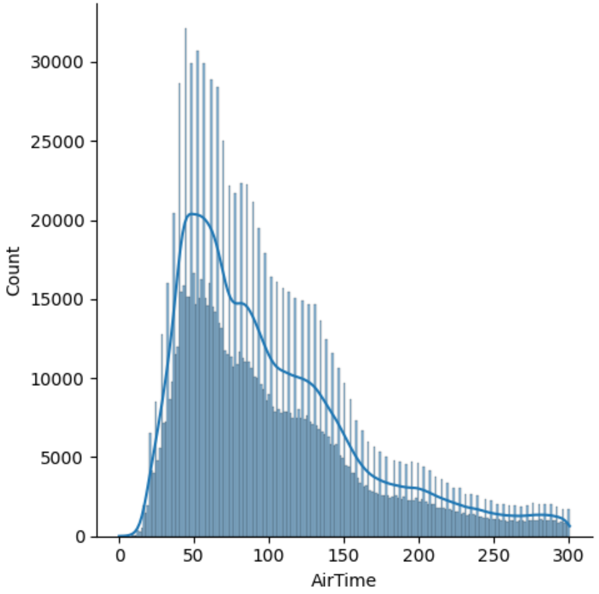
Just for the sake of curiosity, I have analysed how many rows, were affected by this and it's just 29 out of 2M. It's the exact definition for invalid data input, hence I will treat them as outliers.

If one wants to dive deeper, this is the output from *csvtool*:

```
2003-03-07,OO,SLC,SUN,-703.00
2004-01-03,OH,STL,CVG,-153.00
2004-01-14,OH,CVG,SYR,-39.00
2004-01-19,OH,CVG,CAK,-23.00
2004-01-29,OH,CVG,CAK,-19.00
2004-02-02,OH,BTV,CVG,-692.00
2004-02-12,OH,DCA,JAX,-19.00
2004-02-20,OH,MEM,CVG,-123.00
2004-02-23,OH,TYS,CVG,-626.00
2004-03-02,OH,SHV,CVG,-485.00
2004-05-03,OH,MCO,FLL,-12.00
2004-05-09,OO,CLL,IAH,-1.00
2004-05-16,OH,SLC,HLN,-60.00
2004-05-22,OH,ATL,DAB,-685.00
2004-06-01,OH,CVG,PIT,-609.00
2004-06-01,OO,LAX,SAN,-2.00
2004-06-04,OH,ATL,GRR,-36.00
2004-06-13,OH,CVG,GRR,-678.00
2004-06-23,OH,CVG,CLT,-669.00
2004-07-09,OH,CVG,MSP,-46.00
2004-08-15,OH,CVG,CAK,-15.00
2004-09-26,OH,ATL,SDF,-11.00
2004-10-19,OH,STL,CVG,-7.00
2004-10-26,OH,ATL,JFK,-686.00
2004-11-09,OH,MEM,CVG,-427.00
2004-11-10,OH,HPN,CVG,-63.00
2004-12-19,OH,CVG,TUL,-594.00
2004-12-22,OH,DCA,BOS,-206.00
2004-12-22,OO,SLC,RAP,-178.00
```

All affected flights are on 2004 except for one, and carriers having these values are just OO - SkyWest Airlines and most values for OH - PSA Airlines, Inc.

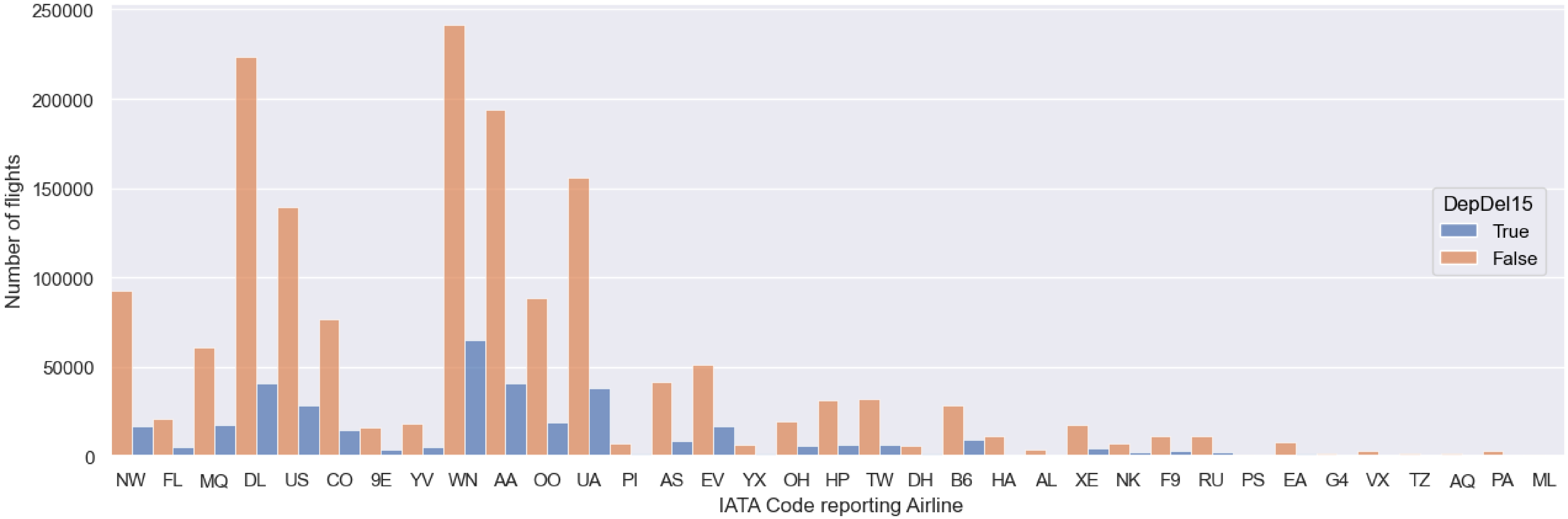
After seeing this, I have performed some outliers cleaning, working on the data below 0.98 percentile only and, for values like this, removing all rows having a negative value in AirTime, also in Distance value.



3_4-Histogram-AirTime-Frequency-NO_OUTLIERS-tiny

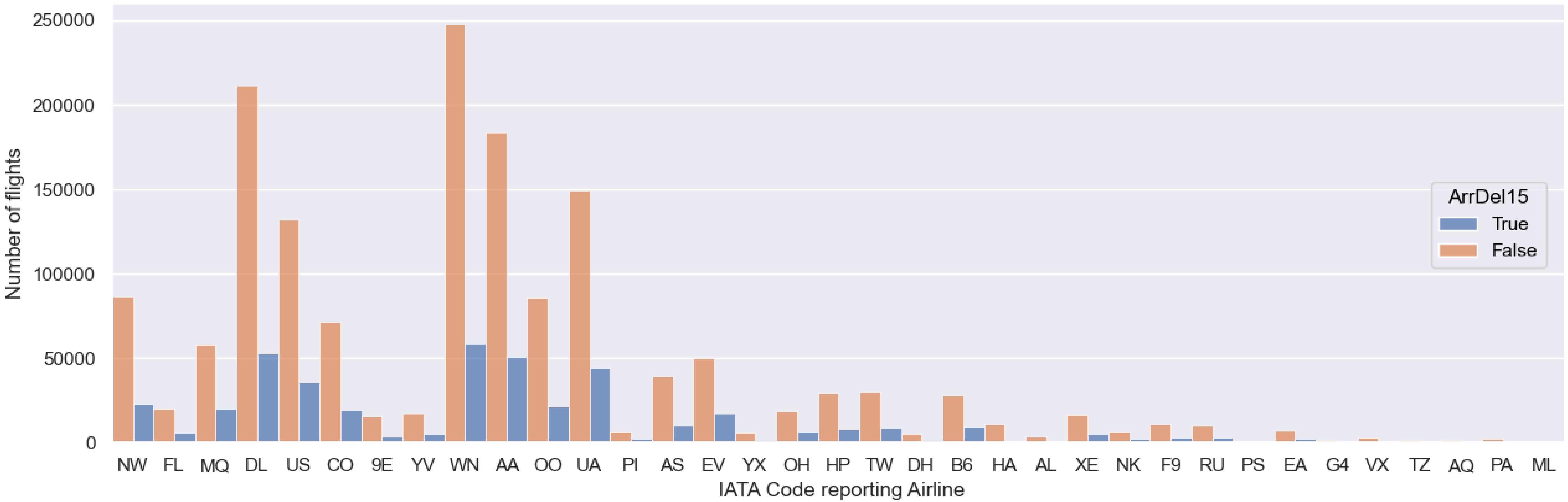
This is the final histogram for cleaned data on AirTime value.

Step 4.1: Analyzing the dataset and visualizing the trends



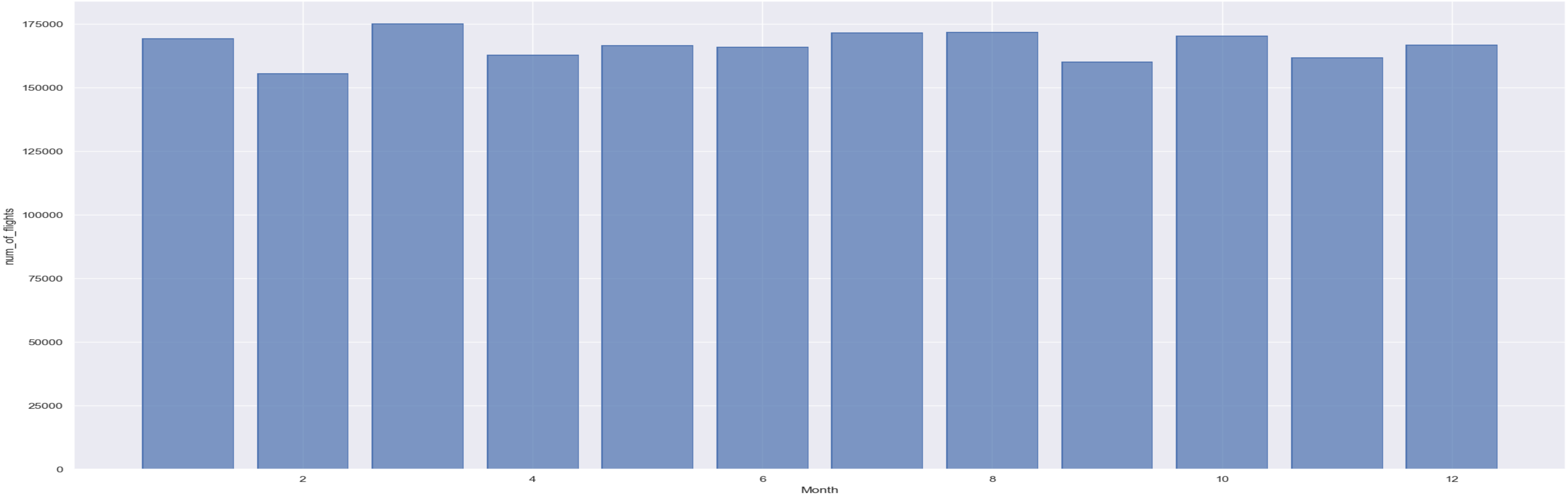
Most of the flights do not have departure delay > 15 minutes. These late departures, might have a big negative perception on customers and a big impact on the carrier operation.

Step 4.1: Analyzing the dataset and visualizing the trends



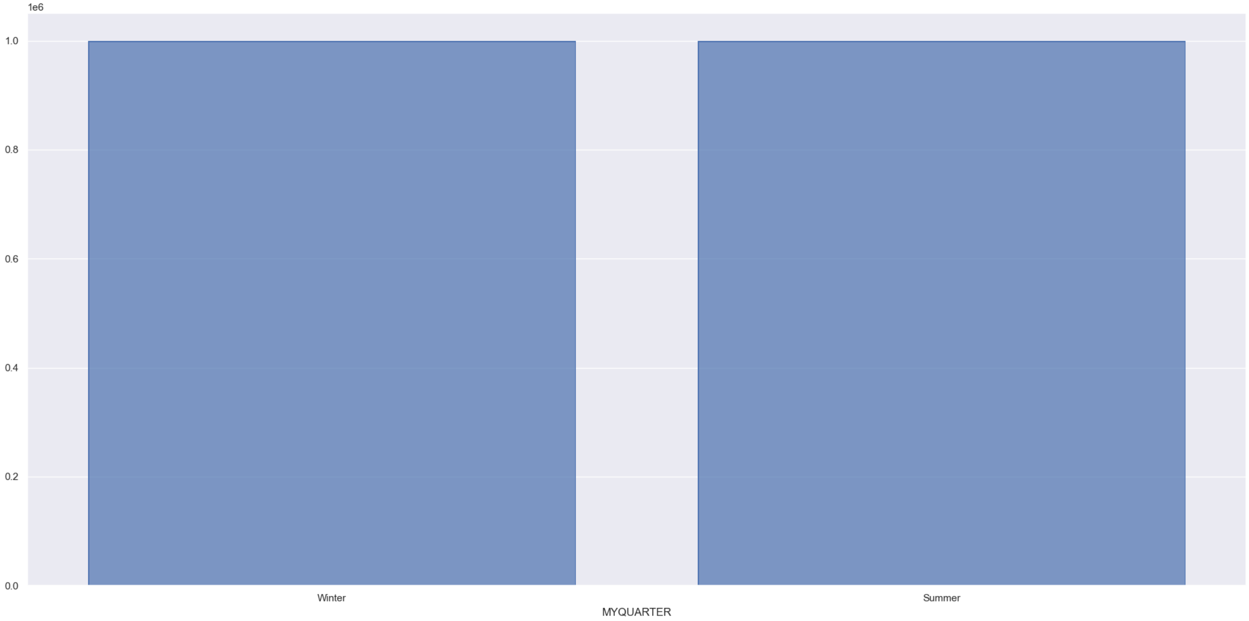
In this case, I am plotting and analysing the value of arrival delay > 15 minutes.
This is affecting customers on possible loss of connection flight forcing the company on compensating them.

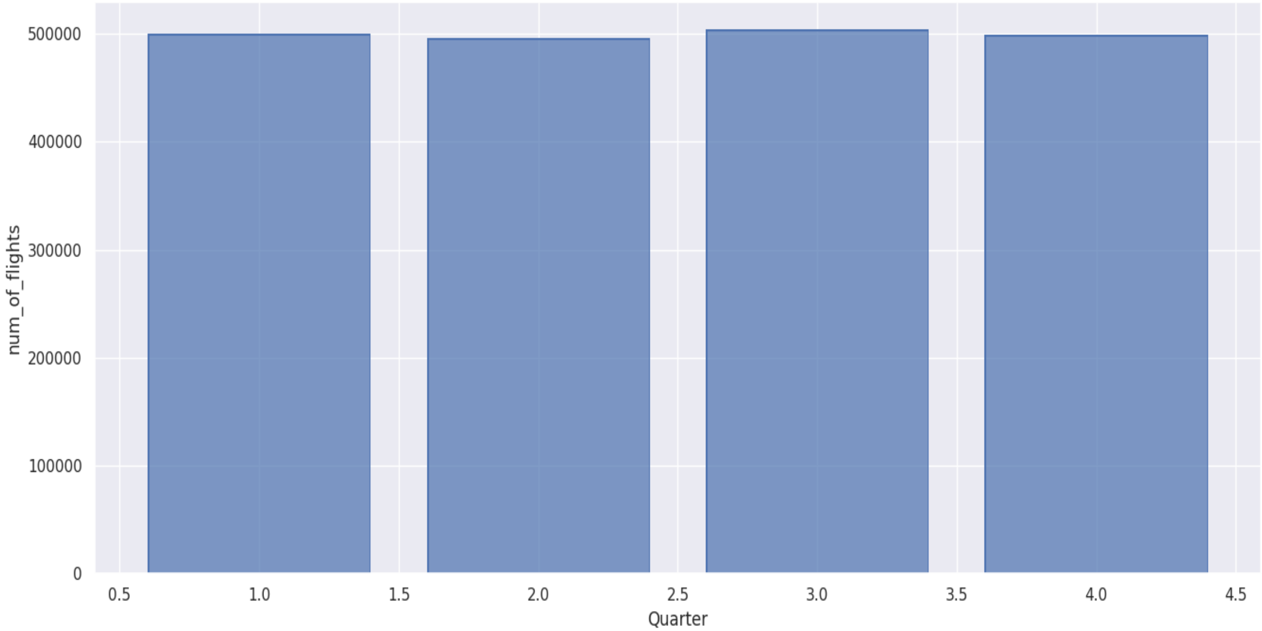
Step 4.2: Analyzing the dataset and visualizing the trends



This histogram shows the total number of flights per month. It includes all the values from the dataset, as removing outliers in this specific case makes no sense. This plot is the same as the one shown in the second slide, step 3.2.

Step 4.2: Analyzing the dataset and visualizing the trends

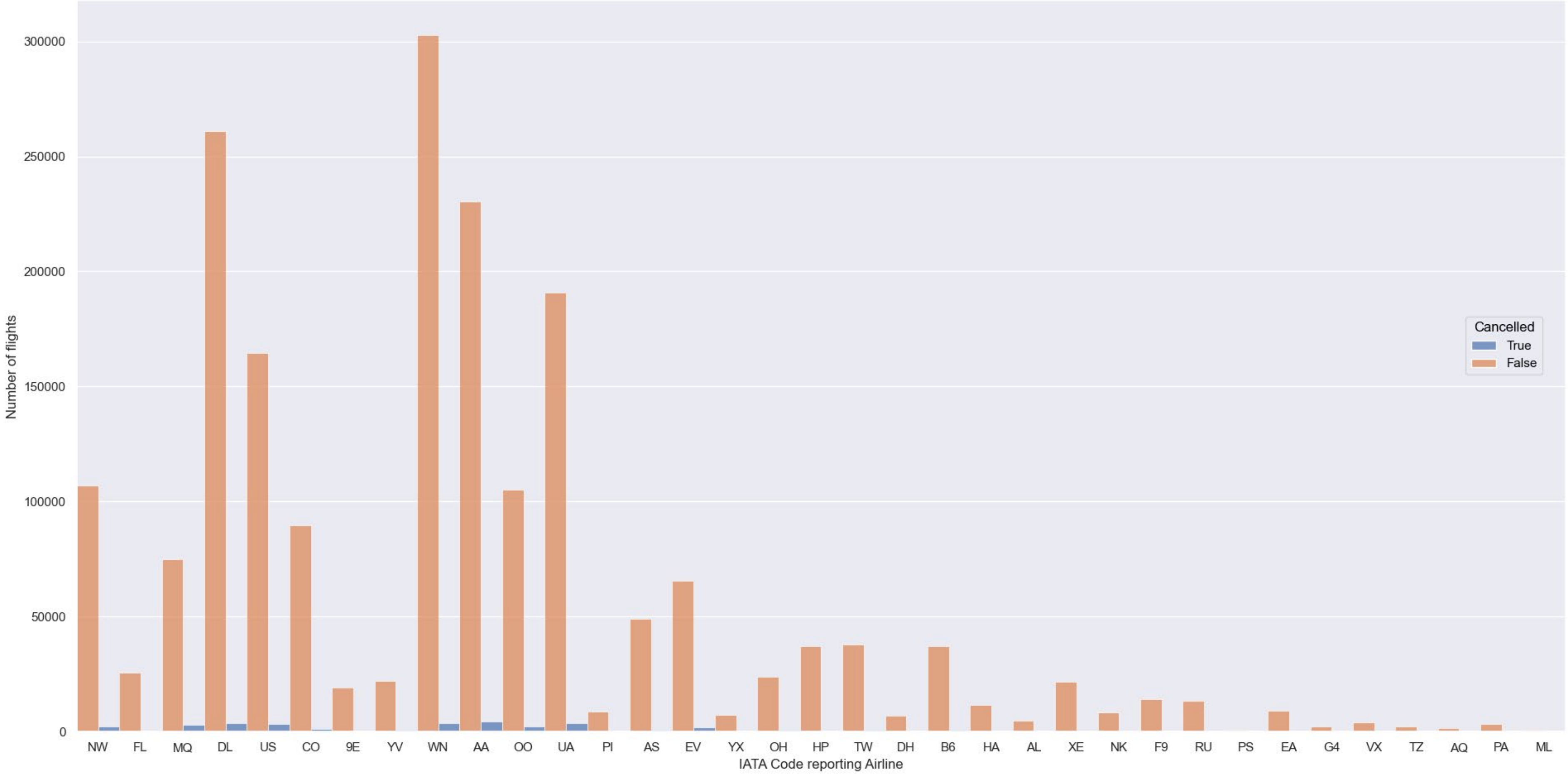




Similarly to the previous slide, these histograms show the total number of flights per quarter or season. The intention here was to uncover possible customer preferences between summer and winter season. Even though the number of flights per each group look almost the same, the final numbers are not.

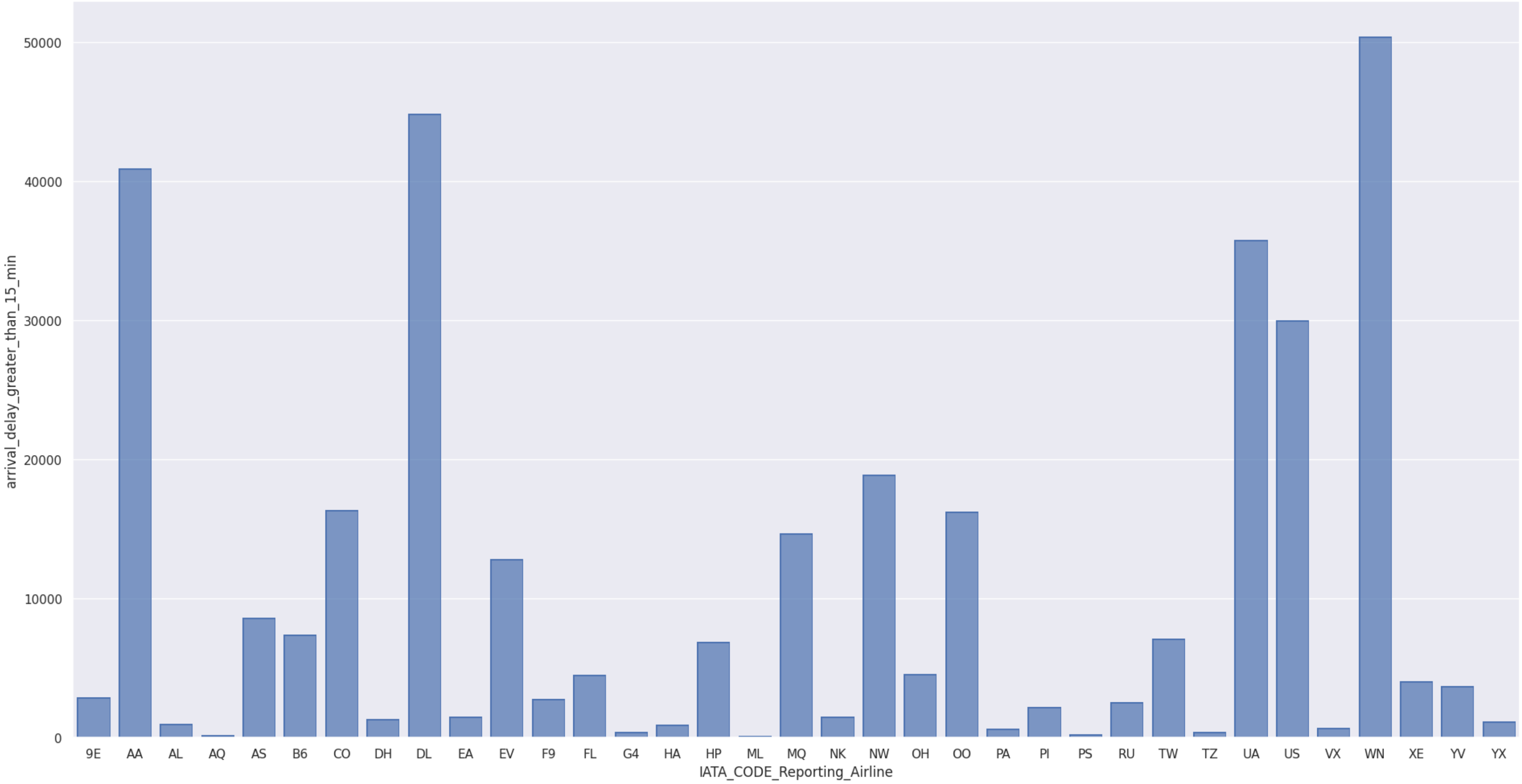
This might be due to the dataset containing only domestic flights.

Step 4.3: Analyzing the dataset and visualizing the trends



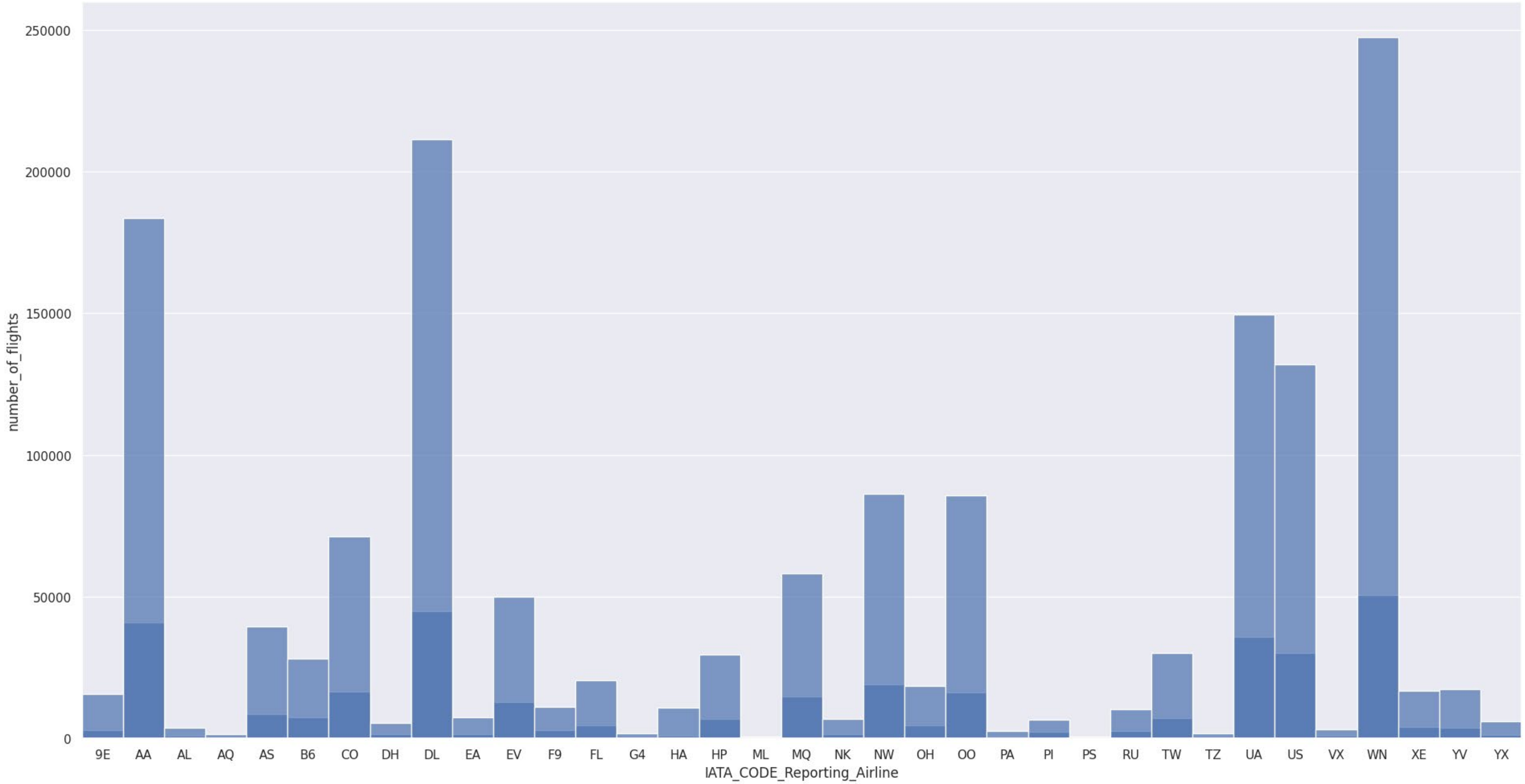
This histogram shows the total number of cancelled flights per company. Only the 0,018% of all the flights in the dataset have been cancelled due to a number of reason, which makes them not appealing to further analyse.

Step 4.4: Analyzing the dataset and visualizing the trends



This histogram shows the total number of delayed (above 15 minutes) flights per carrier on arrival. Some companies do have a lot of delayed flights, which will require a dedicate analysis.

Step 4.4: Analyzing the dataset and visualizing the trends



A big improvement over the previous one, as it provides context as well a comparison between carriers: this histogram shows the total number of delayed (above 15 minutes) flights per carrier on arrival, compared with the number of total flight for each carrier. We can see that some carries do have a big amount of number of flights with a sensible arrival delay.