# Inconsistencies in the data

Firstly, from my findings the data itself was not perfect. I believe this to be due to the web scraper not lowercasing all the data it gathered likely since it is a piece of software and in programming if anything Is spelt wrong it is treated separately. Also, even using Excel’s built in data analysis tools it does not pick this up and treats “Wisecars” and “wisecars”(There are a lot of these inconsistencies in the data) as a separate data entry as shown below:

A graph showing a number of points

AI-generated content may be incorrect.

In this graph Wisecars only shows up as 1000 from the count picked up from my data consolidation software I built. On the python frontend I had consolidated these false entries and Wisecars have more than 3219 total sales recorded as shown below.

A screenshot of a graph

AI-generated content may be incorrect.

During this I had to google possible solutions to problems I had encountered such as some data could also be combined due to the scraper gathering stressfreecarrental.com and stress free car rental and I had to google how I could actually figure this out as it’s a problem I’ve never had to deal with prior to today and I found out its called \*Fuzzy string matching\* which is a common issue to deal with in data analytics. This is a problem I never thought I had to deal with as in truth I had never heard of it. I have decided to implement fuzzy string matching using the python package TheFuzz to determine if strings are close enough to be combined into one piece of data.

A screenshot of a computer

AI-generated content may be incorrect.

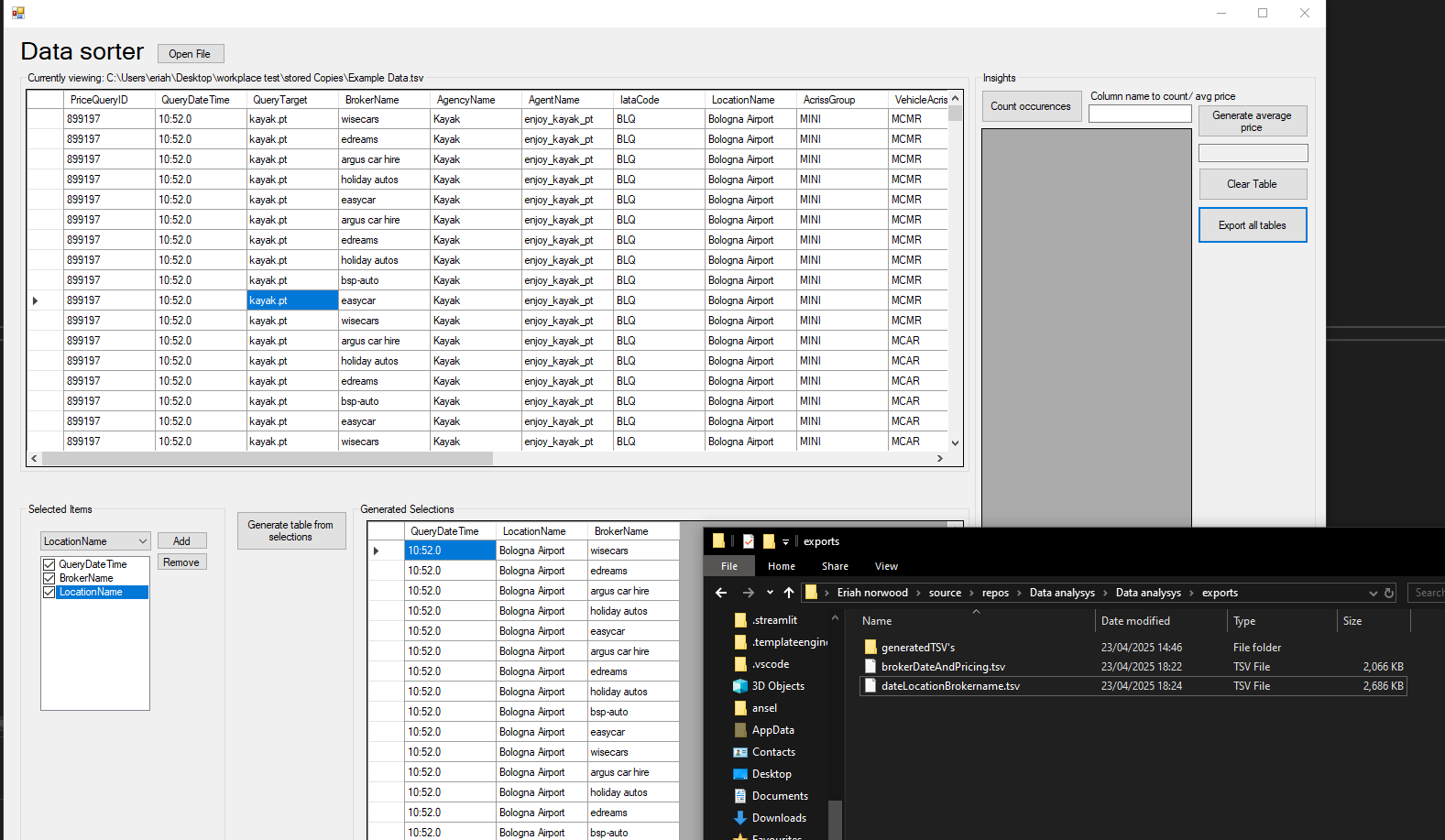
This data heatmap shows that there are no car hires from 22:52pm until 00:32am.

Also interesting how midday (11:21 and 1:32am) have over 5000 queries during that time. Perhaps that aligns with most common flight times. And there aren’t any specific dates in the dataset only just times.

A screenshot of a computer

AI-generated content may be incorrect.

Averages are skewed due to original erroneous data showing up in the original dataset. I am very sure that auto Europe does not rent out cars for 79,000gbp? So that’s some anomalistic data.



Personal Notes:

I also decided to use c# to process the 70,000 individual entries of data and convert the currency as it would complete the operation quicker as it is a compiled language. This was a personal choice. It could have been done in python, but it would be extremely slow rather than developing my own solutions to export separated data entries in the tsv file and organize it.

So the issues I ran into:

-- The c# software did not account for names with Expedia and expedia due to the lower and uppercase showing issues with the data. I had found these issues and corrected

data that was incorrect via combining the two (Obviously in a real case scenario I would ask a supervisor as it could have been the data showing up incorrectly but, in this case, I did not know so I handled it by combining the data.)

^ this means the base data had inconsistencies probably due to search queries from the web scraper not lowercasing any inputted data and programming languages

taking the two pieces of data as separate values but they’re, in fact, the same values.

Additionally, stress free car rentals was a value but also stressfreecarrentals.com was also a value so using fuzzy string matching helped kind of sort that although the .com was confusing the matching. I could have added a manual string sorter in hindsight.