COBIFY CASE STUDY

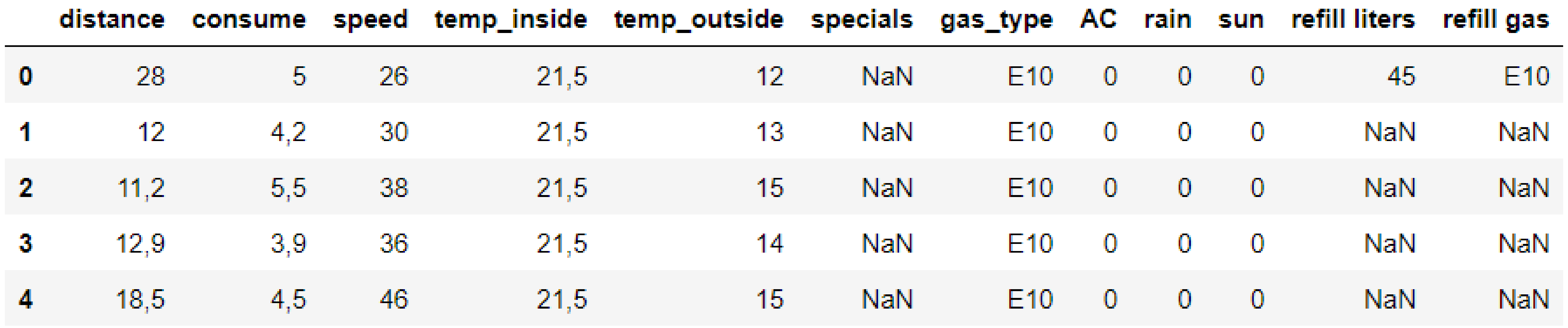
Ricardo Beato

A Tech challenge example

INTRODUCTION

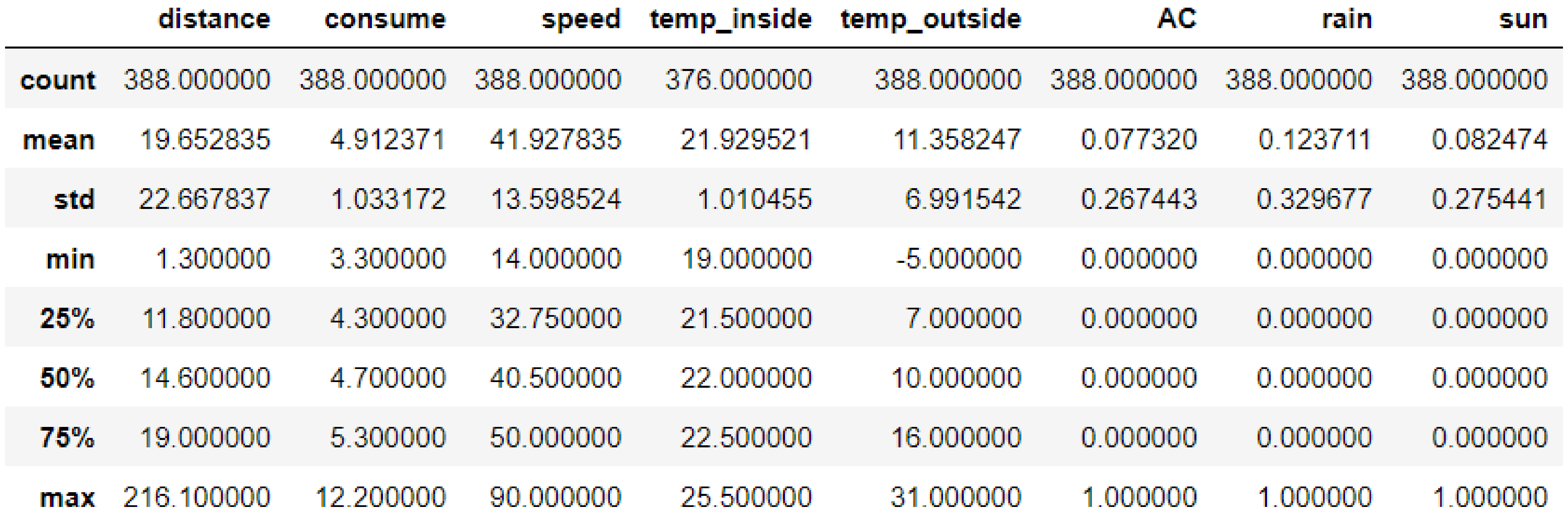
My approach to the challenge: used Python and libraries like pandas, numpy, seaborn and matplotlib to dissecate the data and visualize it. Dropped some columns that were not providing information at all due to the large amount of NaN and additionally a few rows with NaN in the prevailing ones.

THE ANALYSIS



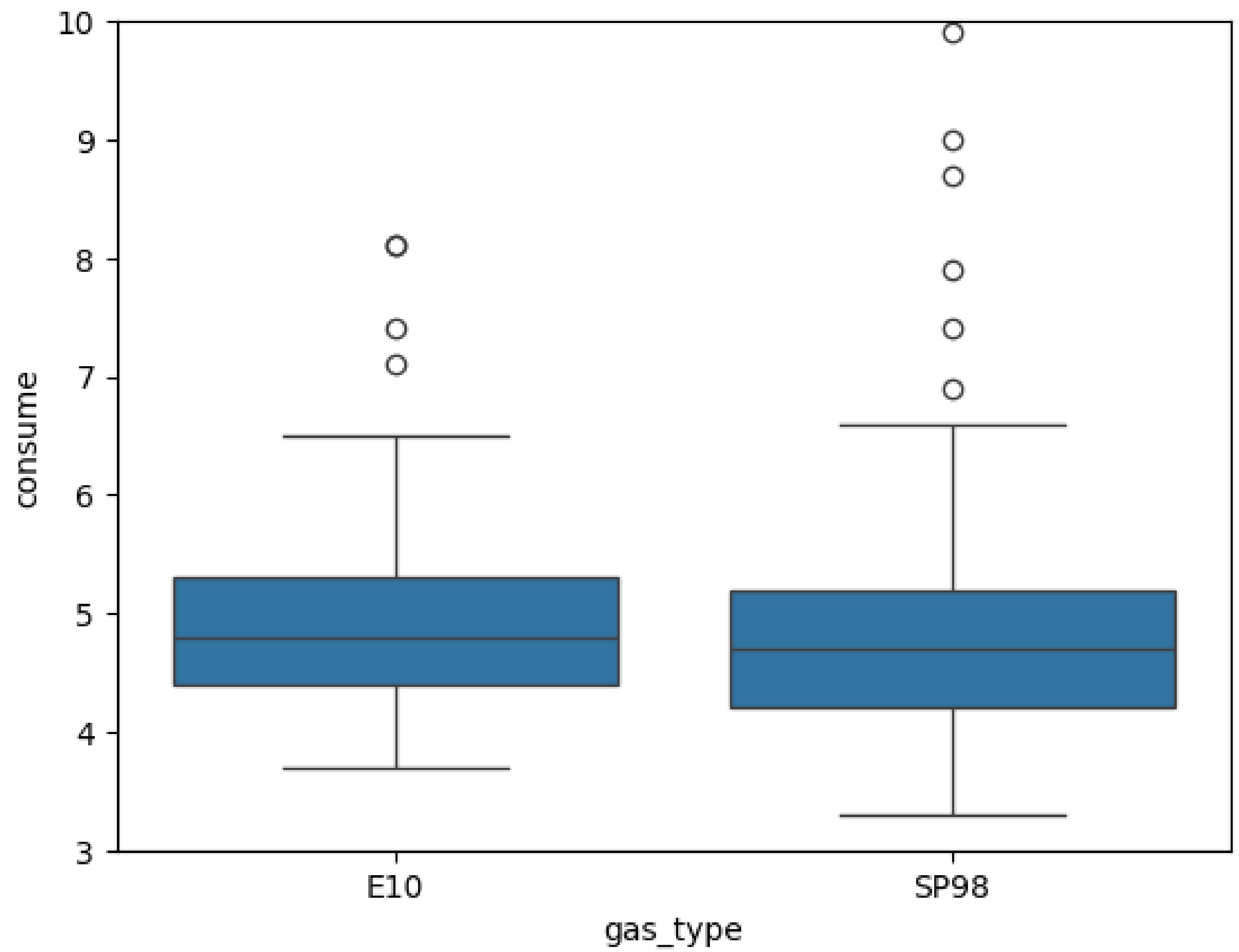
	distance	consume	speed	temp_inside	temp_outside	specials	gas_type	AC	rain	sun	refill liters	refill gas
0	28	5	26	21,5	12	NaN	E10	0	0	0	45	E10
1	12	4,2	30	21,5	13	NaN	E10	0	0	0	NaN	NaN
2	11,2	5,5	38	21,5	15	NaN	E10	0	0	0	NaN	NaN
3	12,9	3,9	36	21,5	14	NaN	E10	0	0	0	NaN	NaN
4	18,5	4,5	46	21,5	15	NaN	E10	0	0	0	NaN	NaN

From the original DF, I dropped the last two columns as well as “specials”. There are two types of gas under scope (E10 and SP98)

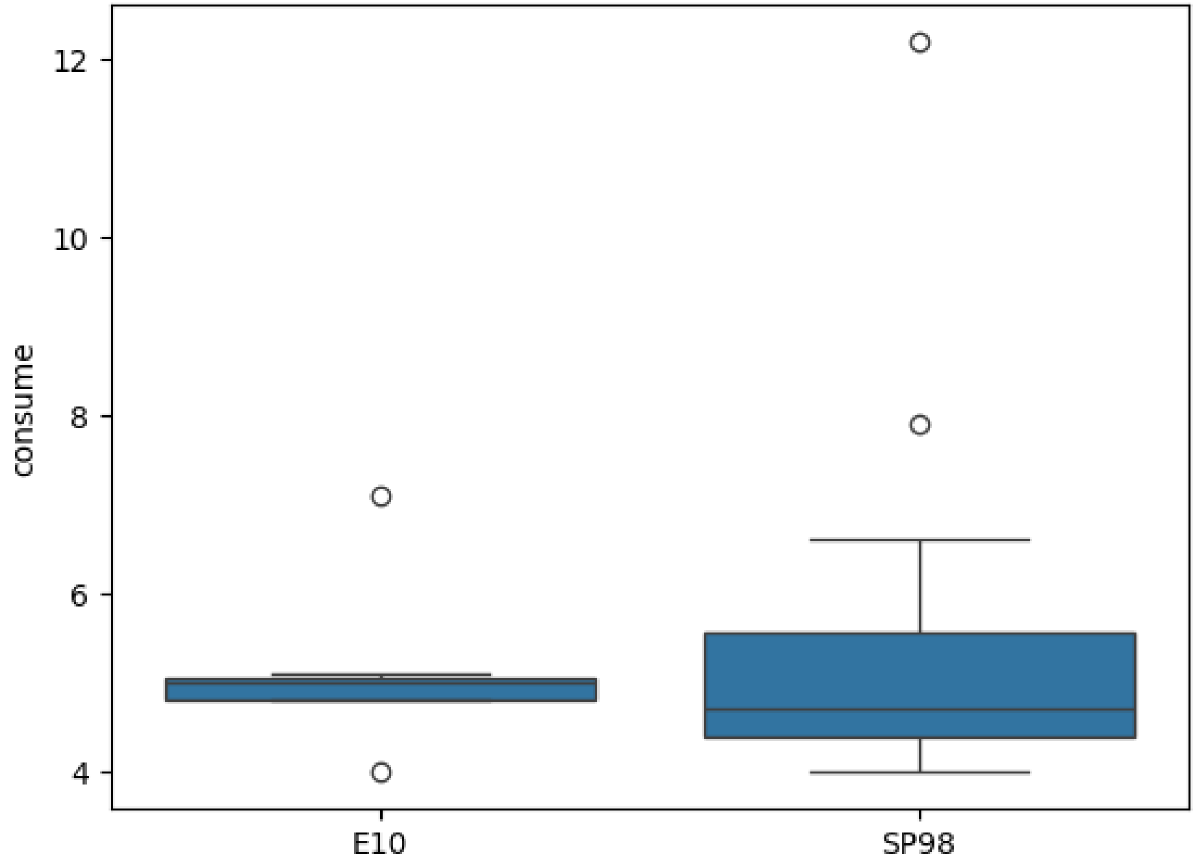


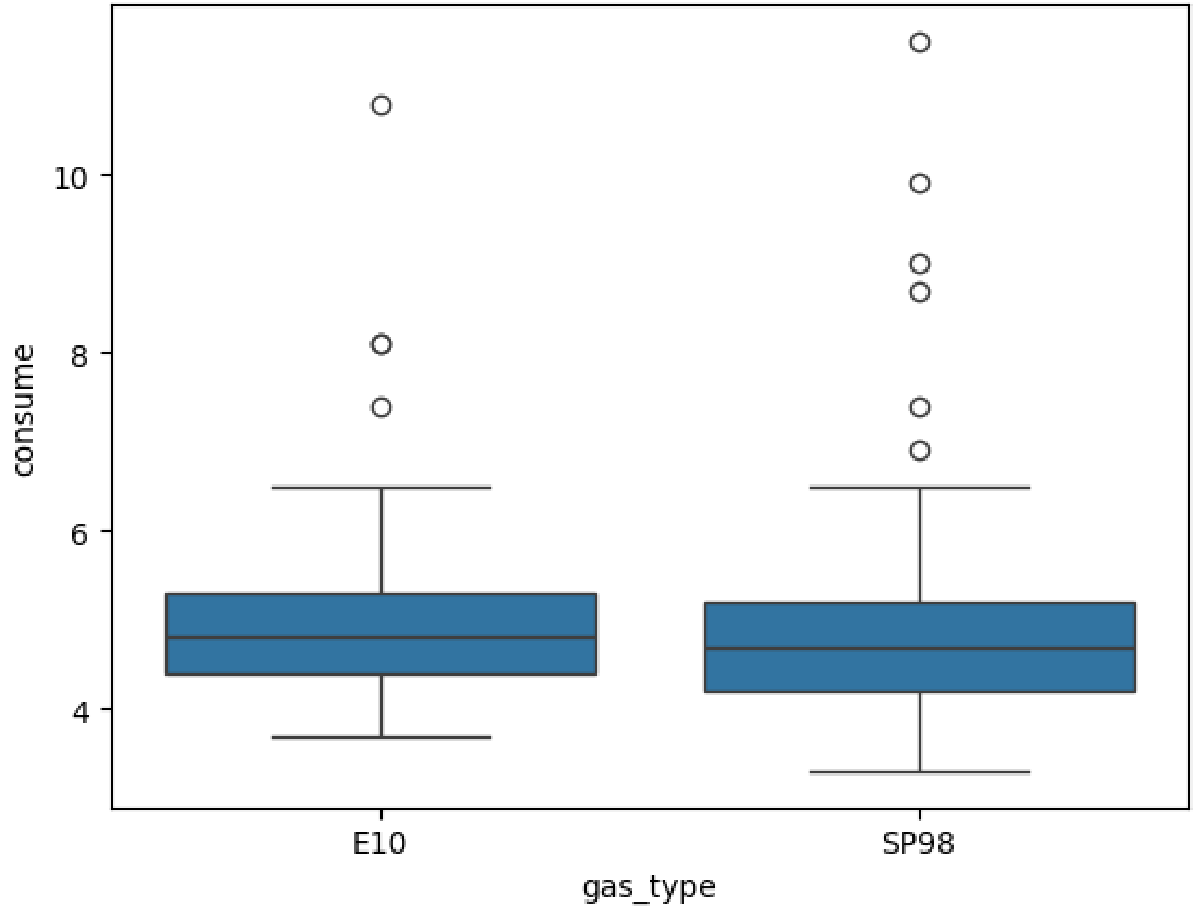
	distance	consume	speed	temp_inside	temp_outside	AC	rain	sun
count	388.000000	388.000000	388.000000	376.000000	388.000000	388.000000	388.000000	388.000000
mean	19.652835	4.912371	41.927835	21.929521	11.358247	0.077320	0.123711	0.082474
std	22.667837	1.033172	13.598524	1.010455	6.991542	0.267443	0.329677	0.275441
min	1.300000	3.300000	14.000000	19.000000	-5.000000	0.000000	0.000000	0.000000
25%	11.800000	4.300000	32.750000	21.500000	7.000000	0.000000	0.000000	0.000000
50%	14.600000	4.700000	40.500000	22.000000	10.000000	0.000000	0.000000	0.000000
75%	19.000000	5.300000	50.000000	22.500000	16.000000	0.000000	0.000000	0.000000
max	216.100000	12.200000	90.000000	25.500000	31.000000	1.000000	1.000000	1.000000

The descriptive statistics suggest some binary variables (last 3) and some other variables that explain the consumption for the possible fuels



At naked eye, SP98 seems to perform better (lower consumption) for the comparable quartiles and median. For the analysis I am disregarding some outliers (everything after Q4).



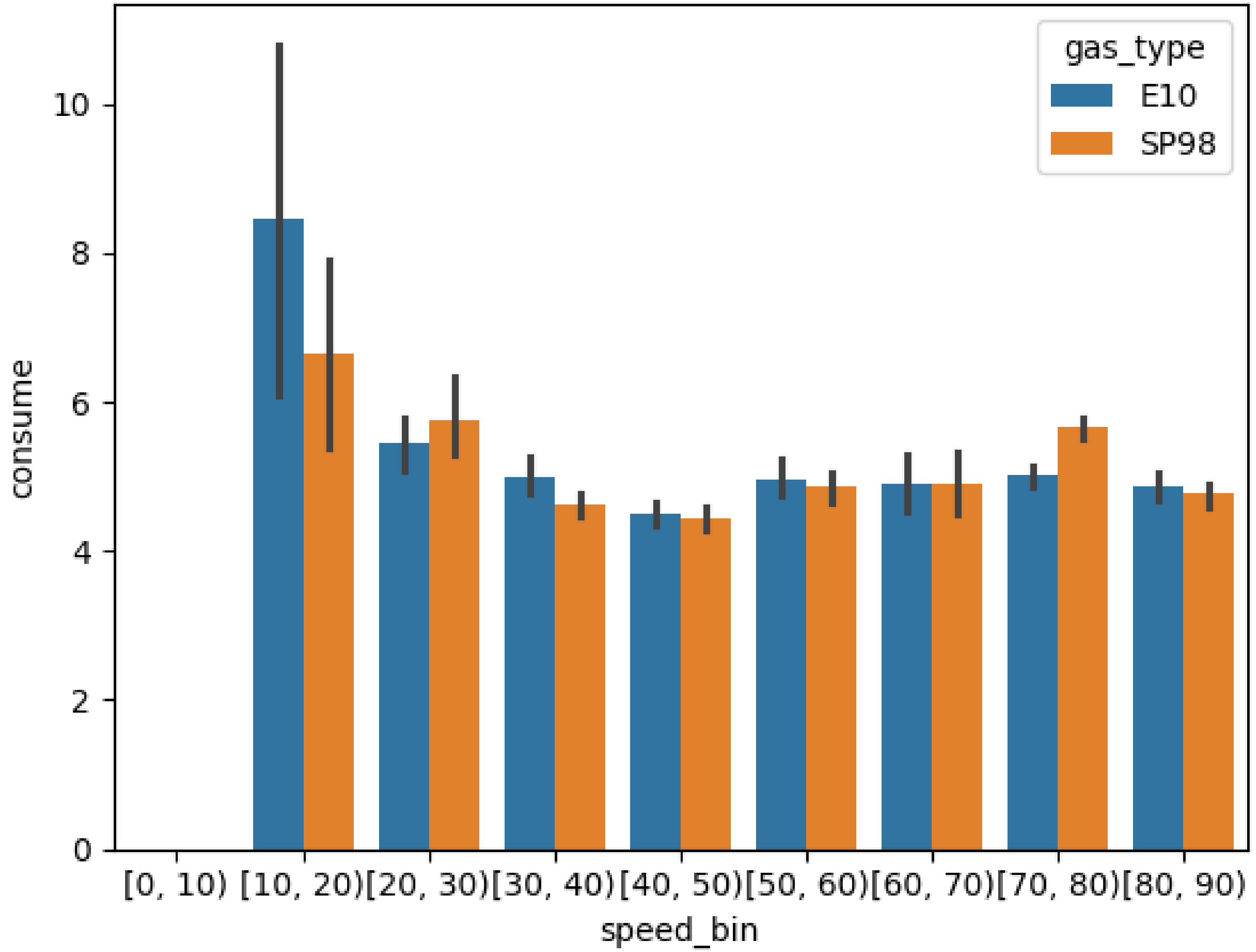


But to be fair, AC was penalizing SP98 more as more observations for this fuel has the AC on during the trip:

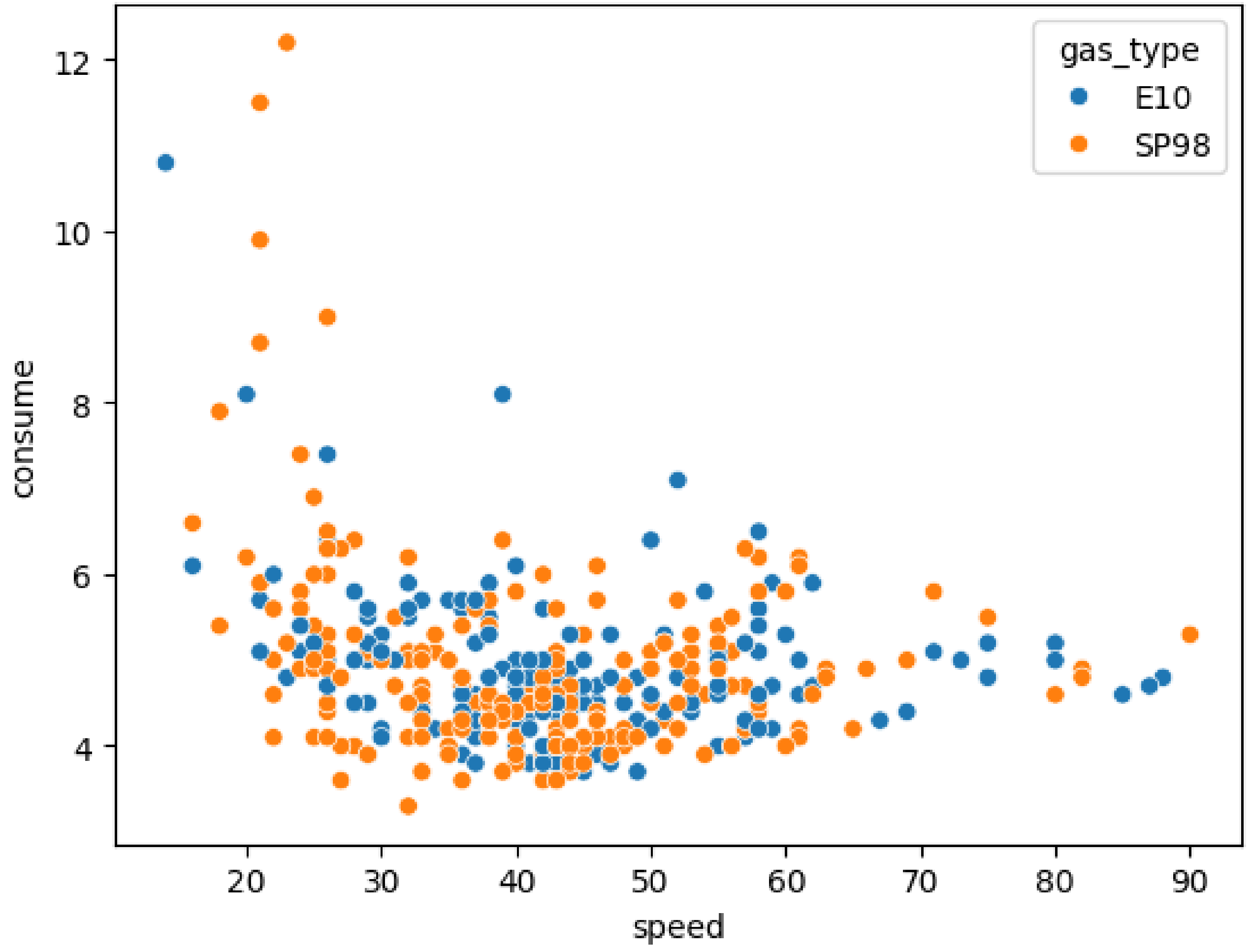
```
gas_type
E10     0.044586
SP98    0.105023
Name: AC, dtype: float64
```

Here we can see what the comparison would look like only for AC = ON observations

And here, only for AC = OFF. In both scenarios SP98 perofrmed better.

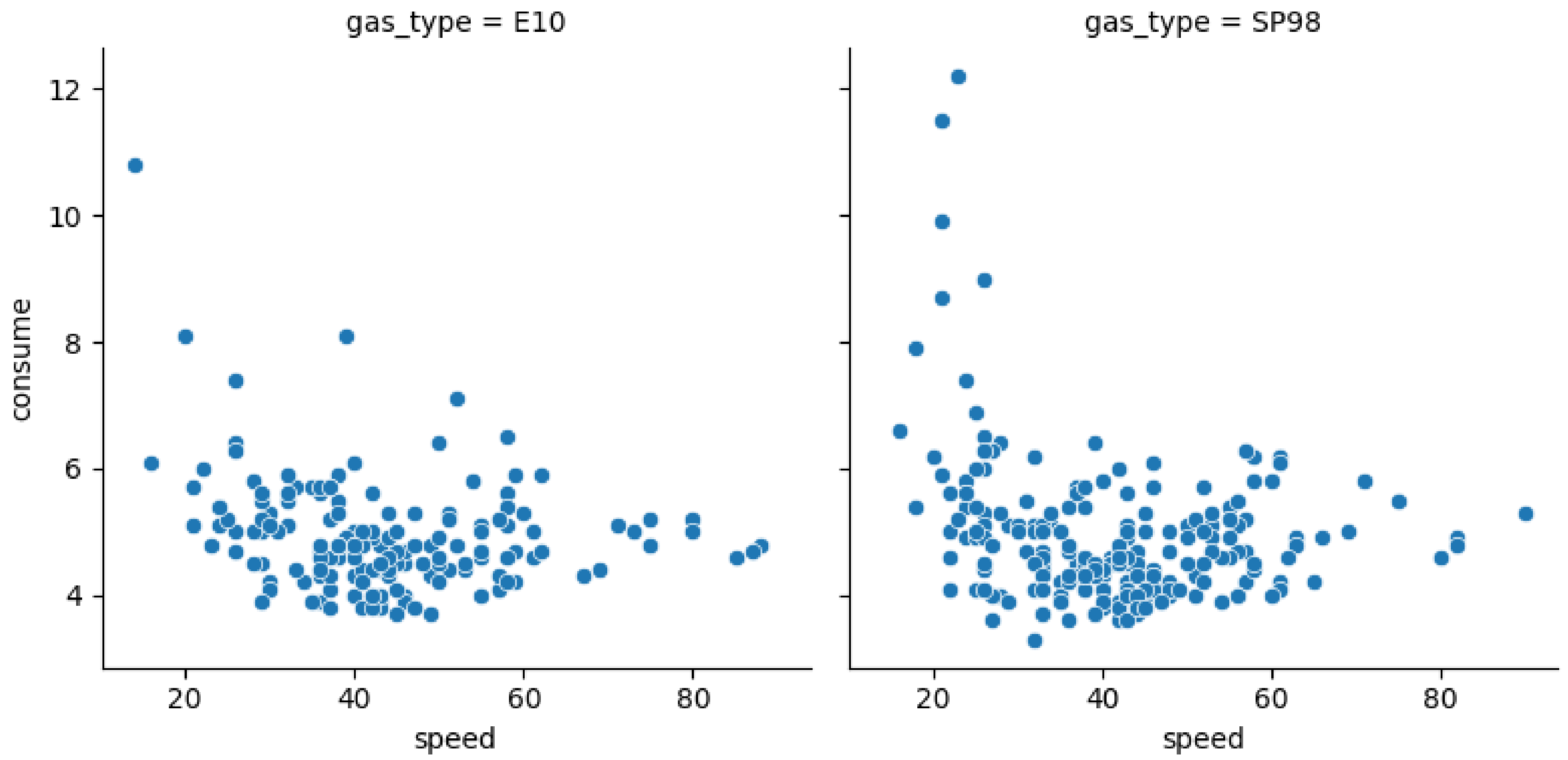


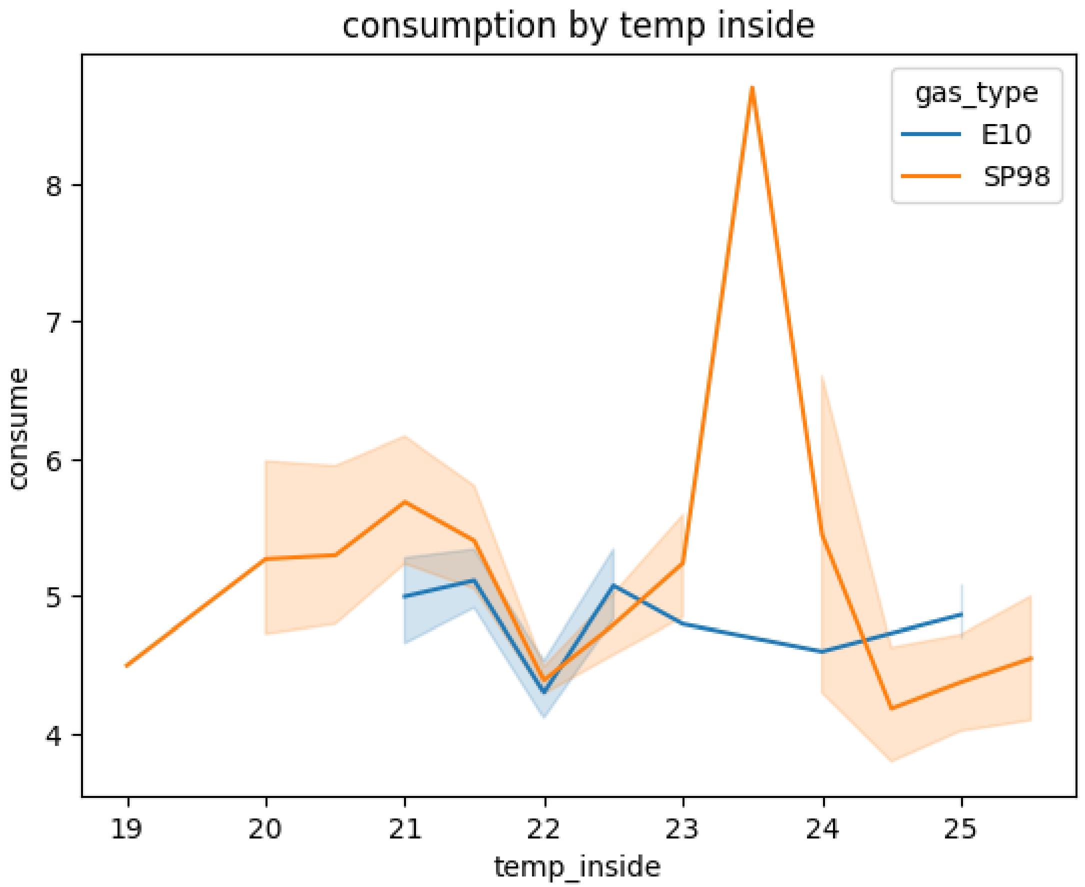
I also wenr on to analyse consumption at different speeds. To better visualize it, I've created bins for the mph/kmph. At specially lower speeds, SP98 is performing better, something that's not constant throughout the speed range. The speed factor seems not to be a clear factor in consumption as confirmed by the following scatter plot.



A messy scatter plot but overlapping both fuels as us seeing that overall the blue dots (E10) seems to be positioned higher than SP98 (bigger consumption)

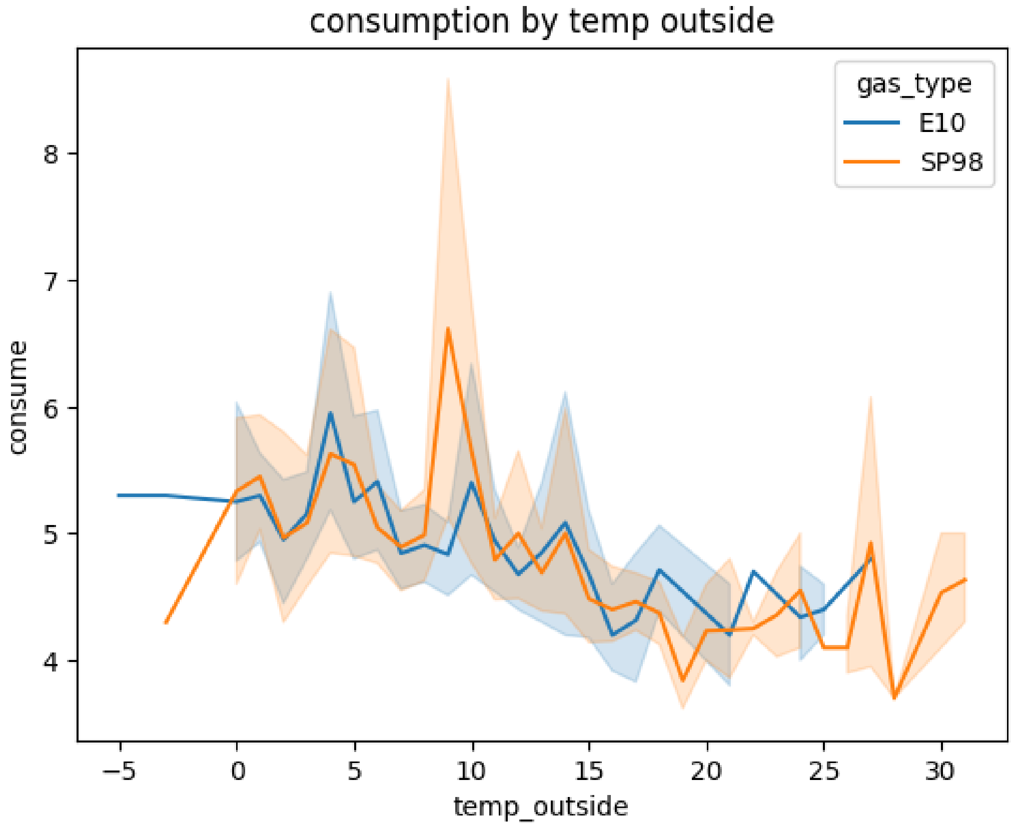
On a split visualization:





Temperature can also be a factor influencing consumption, both inside as seen here hinting towards a worse SP98 performance <--

As well as temperature outside:



CONCLUSIONS

I know this is an informal test, provided I had more time I would have made a ML model splitting the dataframe into 80/20 to train the model into learning the consumption for the different variables and test it afterwards. Additionally I'd also enrich this df with scrapped data from the internet or from an API. The conclusions still point out that SP98 performs generally better after the several shown evidences.