

Ulike verdier i variablene "type" og "Forsterket vegoppmerking"

Ved skriving av dokumentasjon stusset jeg på å finne egenskapen "Forsterket vegoppmerking". Den burde være overflørdig. La oss sjekke den nærmere...

```
In [1]: import pandas as pd

vegopp = pd.read_csv( 'trafikkulykke-vegoppmerking_v8.csv', sep=';')
```

Hva er i variabelen *"Forsterket vegoppmerking"*?

```
In [2]: vegopp['Forsterket vegoppmerking'].value_counts()

Out[2]: Midtoppmerking          19
        Kantoppmerking         15
        Både midt- og kantoppmerking    9
        Name: Forsterket vegoppmerking, dtype: int64
```

Altså kun 43 rader der vi har data på denne variabelen.

Min første tanke var at variabele *"Forsterket vegoppmerking"* kom fra fra prosesens **trinn 4: Overlapp kant og midtoppmerking**. Dette er feil: Den kommer fra **ulykkesdataene**

La oss sammenligne med variabelen *"type"*:

```
In [3]: vegopp.type.value_counts()

Out[3]: Forsterket midtoppmerking    2500
        Forsterket kantoppmerking   1010
        Både forst. kant- og midtoppm    875
        Name: type, dtype: int64
```

La oss grave litt dypere i de radene der det finnes data på varibelen *"Forsterket vegoppmerking"*:

```
In [4]: sjekk = vegopp[ vegopp['Forsterket vegoppmerking'].notnull() ]
sjekk[['id','type', 'Forsterket vegoppmerking', 'strekningsslengde']]
```

Out[4]:

	id	type	Forsterket vegoppmerking	strekningsslengde
208	760718791	Forsterket midtoppmerking	Midtoppmerking	1152
212	623612669	Forsterket midtoppmerking	Midtoppmerking	1151
350	760718835	Forsterket midtoppmerking	Midtoppmerking	112
2084	760718762	Forsterket midtoppmerking	Midtoppmerking	83
2331	579115623	Forsterket midtoppmerking	Midtoppmerking	5405
2431	628626561	Forsterket midtoppmerking	Midtoppmerking	45
2559	628626562	Forsterket midtoppmerking	Midtoppmerking	172
2627	760718766	Forsterket midtoppmerking	Midtoppmerking	240
2895	628626561	Forsterket midtoppmerking	Midtoppmerking	45
3265	375749404	Forsterket midtoppmerking	Midtoppmerking	195
3467	760718762	Forsterket midtoppmerking	Midtoppmerking	407
4104	760718765	Forsterket midtoppmerking	Midtoppmerking	959
4283	633729844	Forsterket kantoppmerking	Kantoppmerking	495
4287	642372563	Forsterket kantoppmerking	Kantoppmerking	46
4291	671926923	Forsterket midtoppmerking	Midtoppmerking	1130
4296	760718834	Forsterket midtoppmerking	Både midt- og kantoppmerking	168
4299	602968288	Forsterket kantoppmerking	Kantoppmerking	428
4301	555194363	Forsterket kantoppmerking	Kantoppmerking	1040
4302	747066224	Forsterket midtoppmerking	Midtoppmerking	11
4304	671955699	Forsterket midtoppmerking	Midtoppmerking	1095
4311	747066224	Forsterket midtoppmerking	Midtoppmerking	11
4313	555194362	Forsterket kantoppmerking	Kantoppmerking	1040
4318	845648888	Forsterket kantoppmerking	Kantoppmerking	930
4324	579006804	Forsterket midtoppmerking	Midtoppmerking	171
4328	745399377	Både forst. kant- og midtoppm	Både midt- og kantoppmerking	5332
4329	600611289	Forsterket kantoppmerking	Kantoppmerking	219
4330	481668619	Både forst. kant- og midtoppm	Både midt- og kantoppmerking	1062
4332	600326668	Forsterket kantoppmerking	Kantoppmerking	373
4337	481668616	Forsterket midtoppmerking	Både midt- og kantoppmerking	1062
4338	633729848	Forsterket kantoppmerking	Kantoppmerking	495
4341	646597916	Forsterket midtoppmerking	Midtoppmerking	206
4343	845648880	Forsterket kantoppmerking	Kantoppmerking	930
4344	481668616	Forsterket midtoppmerking	Både midt- og kantoppmerking	1062
4348	845648888	Forsterket kantoppmerking	Kantoppmerking	930
4349	602968286	Forsterket kantoppmerking	Kantoppmerking	505
4352	657379653	Forsterket midtoppmerking	Midtoppmerking	504
4361	846063910	Både forst. kant- og midtoppm	Både midt- og kantoppmerking	478
4365	481668618	Både forst. kant- og midtoppm	Både midt- og kantoppmerking	1062
4366	845648880	Forsterket kantoppmerking	Kantoppmerking	930
4367	845648897	Både forst. kant- og midtoppm	Kantoppmerking	183
4368	760669150	Forsterket midtoppmerking	Både midt- og kantoppmerking	91
4369	845648906	Både forst. kant- og midtoppm	Kantoppmerking	153
4380	745399375	Både forst. kant- og midtoppm	Både midt- og kantoppmerking	5332

De fleste radene ser ut til å ha samme svar i de to variablene - men ikke alle.

La oss finne de radene der kun en av kolonnene *type* eller *Forsterket vegoppmerking* har teksten "Både" i seg:

```
In [5]: feilbd = sjekk[ ( (~sjekk['type'].str.contains('Både')) & sjekk['Forsterket vegoppmerking'].str.contains('Både')) |
                        (( sjekk['type'].str.contains('Både')) & ~sjekk['Forsterket vegoppmerking'].str.contains('Både')) ) ]
feilbd[['id', 'type', 'Forsterket vegoppmerking', 'strekningsslengde']]
```

Out[5]:

	id	type	Forsterket vegoppmerking	strekningsslengde
4296	760718834	Forsterket midtoppmerking	Både midt- og kantoppmerking	168
4337	481668616	Forsterket midtoppmerking	Både midt- og kantoppmerking	1062
4344	481668616	Forsterket midtoppmerking	Både midt- og kantoppmerking	1062
4367	845648897	Både forst. kant- og midtoppm	Kantoppmerking	183
4368	760669150	Forsterket midtoppmerking	Både midt- og kantoppmerking	91
4369	845648906	Både forst. kant- og midtoppm	Kantoppmerking	153

```
In [6]: print( "Strekningsslengde", feilbd.strekningsslengde.sum())

Strekningsslengde 2719
```

Altså har vi 6 rader der egenskapene *type* og *forsterket midtoppmerking* gir ulike svar på om strekningen overlapper, totalt 2.7 km.

Vi ser litt nærmere på om det er andre avvik mellom to kolonnene. Hvis vi tar vekk tilfellene med overlapp (teksten "Både" i en av kolonnene)

```
In [7]: sjekk2 = sjekk[ ((~sjekk['type'].str.contains('Både'))
                    & (~sjekk['Forsterket vegoppmerking'].str.contains('Både')) ) ]
sjekk2[['type', 'Forsterket vegoppmerking']]
```

Out[7]:

	type	Forsterket vegoppmerking
208	Forsterket midtoppmerking	Midtoppmerking
212	Forsterket midtoppmerking	Midtoppmerking
350	Forsterket midtoppmerking	Midtoppmerking
2084	Forsterket midtoppmerking	Midtoppmerking
2331	Forsterket midtoppmerking	Midtoppmerking
2431	Forsterket midtoppmerking	Midtoppmerking
2559	Forsterket midtoppmerking	Midtoppmerking
2627	Forsterket midtoppmerking	Midtoppmerking
2895	Forsterket midtoppmerking	Midtoppmerking
3265	Forsterket midtoppmerking	Midtoppmerking
3467	Forsterket midtoppmerking	Midtoppmerking
4104	Forsterket midtoppmerking	Midtoppmerking
4283	Forsterket kantoppmerking	Kantoppmerking
4287	Forsterket kantoppmerking	Kantoppmerking
4291	Forsterket midtoppmerking	Midtoppmerking
4299	Forsterket kantoppmerking	Kantoppmerking
4301	Forsterket kantoppmerking	Kantoppmerking
4302	Forsterket midtoppmerking	Midtoppmerking
4304	Forsterket midtoppmerking	Midtoppmerking
4311	Forsterket midtoppmerking	Midtoppmerking
4313	Forsterket kantoppmerking	Kantoppmerking
4318	Forsterket kantoppmerking	Kantoppmerking
4324	Forsterket midtoppmerking	Midtoppmerking
4329	Forsterket kantoppmerking	Kantoppmerking
4332	Forsterket kantoppmerking	Kantoppmerking
4338	Forsterket kantoppmerking	Kantoppmerking
4341	Forsterket midtoppmerking	Midtoppmerking
4343	Forsterket kantoppmerking	Kantoppmerking
4348	Forsterket kantoppmerking	Kantoppmerking
4349	Forsterket kantoppmerking	Kantoppmerking
4352	Forsterket midtoppmerking	Midtoppmerking
4366	Forsterket kantoppmerking	Kantoppmerking

Ser riktig ut, men la oss telle: Har vi noen rader der *type* og *Forsterket vegoppmerking* gir ulike svar på om det er kant- eller midtoppmerking?

```
In [8]: sjekk2[ ( ( sjekk2['type'].str.contains('kant')) & (sjekk2['Forsterket vegoppmerking'].str.contains('Midt')) ) |
              ( (sjekk2['type'].str.contains('midt')) & (sjekk2['Forsterket vegoppmerking'].str.contains('Kant')) ) )]
```

Out[8]:

	from_measure	to_measure	route_id	from_date	to_date	fart_fra_aar	fart_til_aar	fagr_2021	type	fresemetode	...	post_tiltak_utl
0 rows × 96 columns												

Niks. Avvikene er altså avgrenset til de 6 radene vi i stedet, hvor den ene kolonnen antyder "Både kant og midt", men ikke den andre).

Årsakene til avvikene kan være mange og varierte, men en av de mest sannsynlige er manuelle feil ved registrering av trafikkulykker.