data_exploration

June 30, 2021

Table of Contents

1 Data exploration

```
[NbConvertApp] Converting notebook dictionaries_rename.ipynb to script
[NbConvertApp] Writing 113060 bytes to dictionaries_rename.py
```

1.1 Preparing some functions for plotting

1.2 Does time matter?

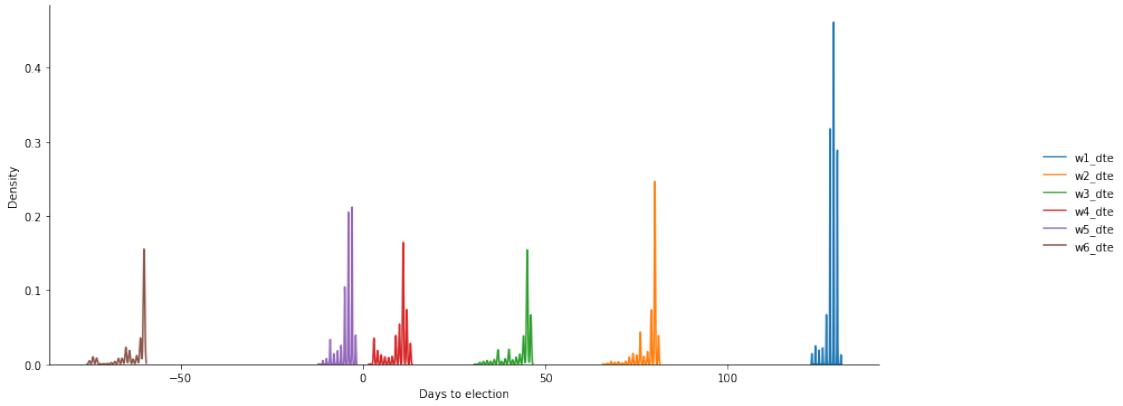
In this section we want to answer the question whether the point in time at which the survey is completed matters.

1.2.1 `Days to election`

We first noticed that in various decision tree (DT) models the so called `days_to_election` feature was not only important but even the most important (it made up the "root split").

Density Let us look at the distribution first to get a feel for how this feature looks.

[4]: `<seaborn.axisgrid.FacetGrid at 0x25e7ea9be50>`

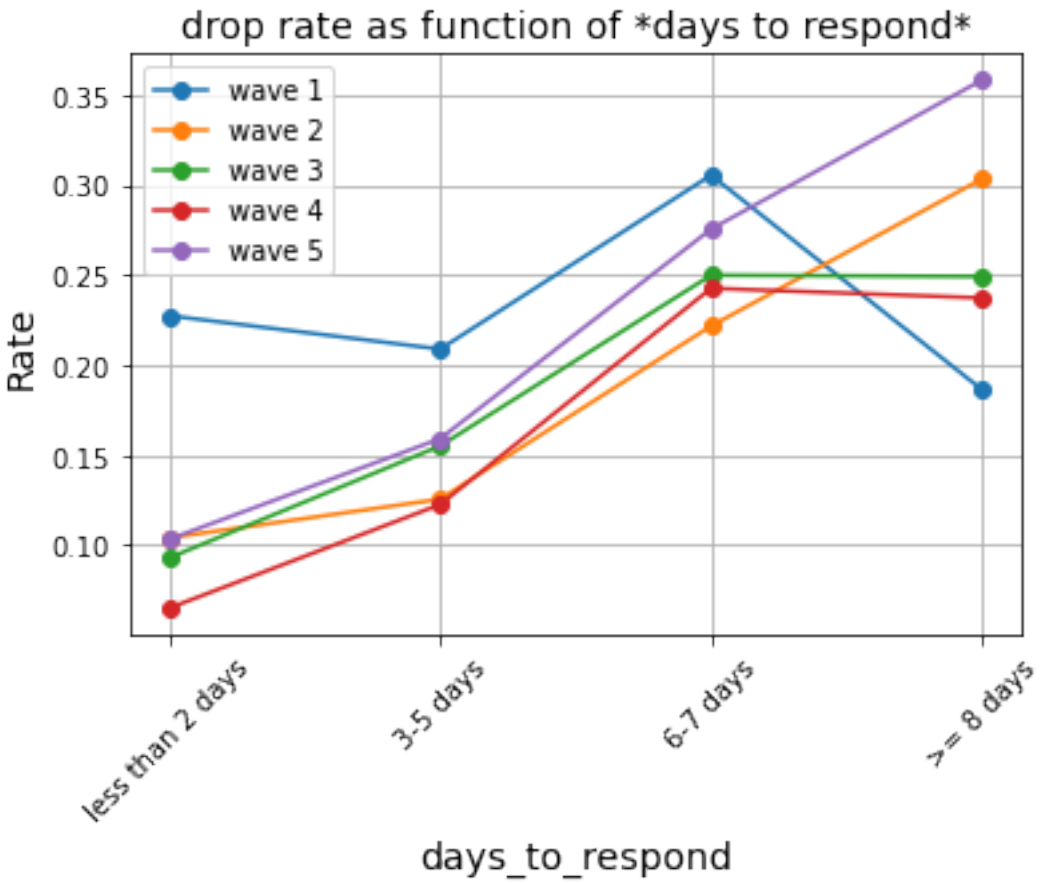


We immediately see the different waves; each spanning approximately a week. We also note that most people seem to respond early in the wave. This means that most people respond right away. The question remains, who are the people that respond late in the wave? Are they contacted later, or are they just deferring?

Also, the first day never seems to be the most popular (with the exception of wave 6). This is surprising as I would assume that most people do the survey right away.. The most of responds were made either on Wednesday or Thursday, so in general day of the week when survey has been sent does not matter

Influence on attrition First we normalize each wave individually, in order to make this feature comparable across waves. We do this by finding for each wave the earliest day somebody responded and take this as "day zero".

Next we do some binning to get a bit more stability at plot the drop out rate of participant in the next wave depending on which day the answered the previous wave.



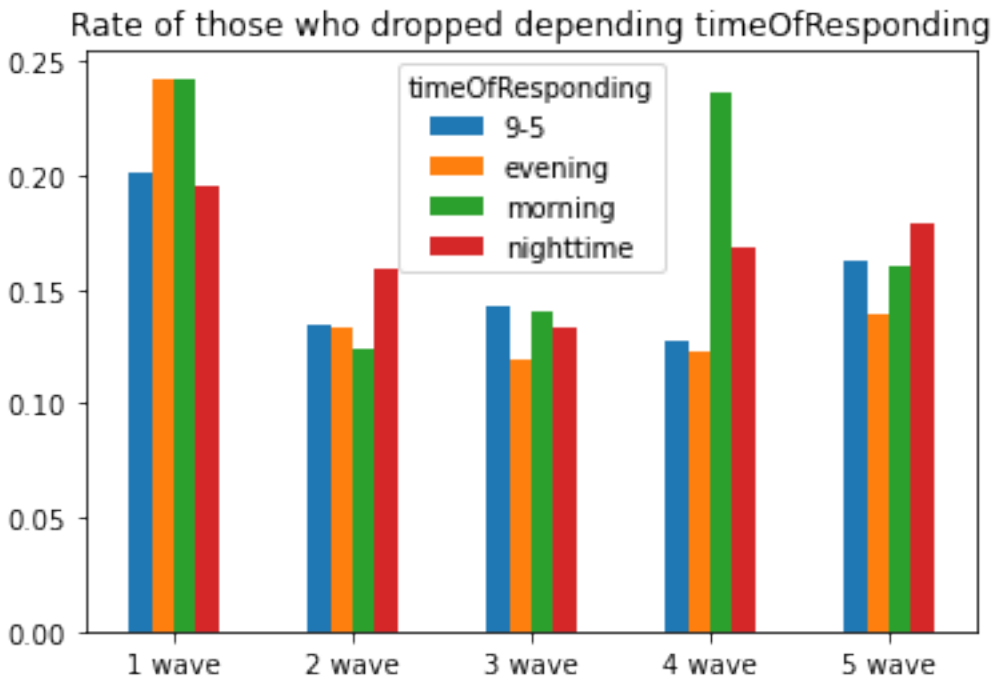
If people respond *sooner*, they are *less likely to drop* (except wave 1). The guess would be that either people who defer "work" do so consistently and are therefore more likely to forget about it in the future - or, people who do not enjoy answering the survey tend to defer and ultimately even drop it.

1.2.2 Influence of `time of day` on attrition

Does the time of day a person answers the survey impact their attrition?

```
wave: 1
Counter({'9-5': 2271, 'evening': 1033, 'morning': 169, 'nighttime': 143})
wave: 2
Counter({'9-5': 1497, 'evening': 1035, 'morning': 162, 'nighttime': 132})
wave: 3
Counter({'9-5': 1407, 'evening': 1014, 'morning': 164, 'nighttime': 113})
wave: 4
Counter({'9-5': 1721, 'evening': 919, 'morning': 131, 'nighttime': 77})
wave: 5
Counter({'9-5': 1644, 'evening': 899, 'morning': 106, 'nighttime': 84})
```

```
<Figure size 720x360 with 0 Axes>
```

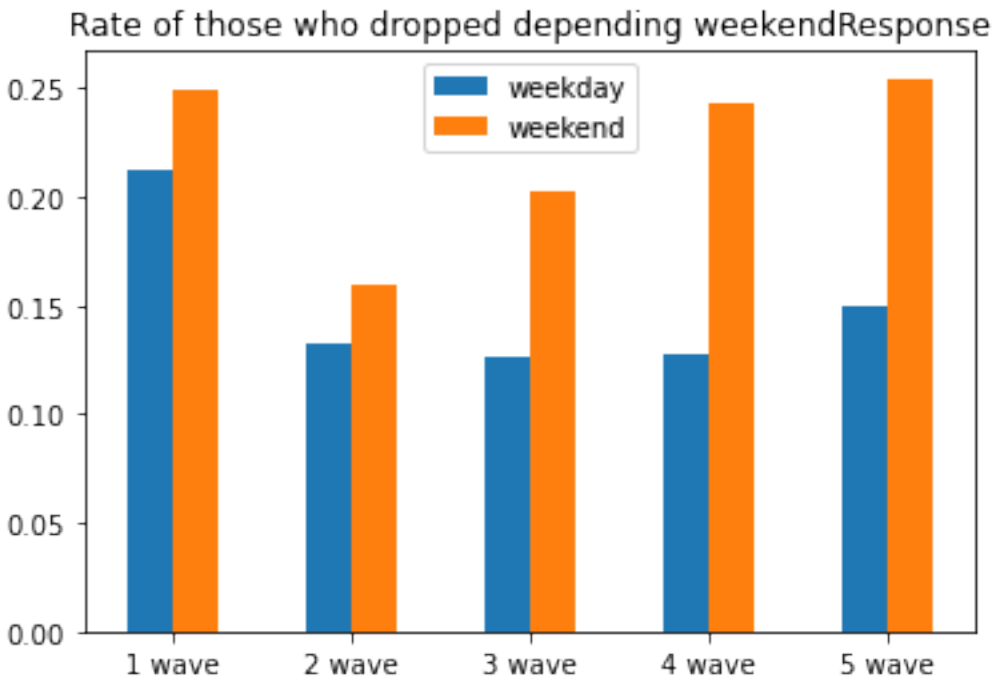


Evening seems to be a bit better than the other times but still not much. It would be interesting to know what happened in the morning of wave 4?

1.2.3 *Weekdays vs weekends* and their effect on attrition

We form to groups of peoples - those who answered the previous wave on a *weekday* and those who did so on the *weekend* - and check their difference in terms of attrition.

```
wave: 1
Counter({0: 3359, 1: 257})
wave: 2
Counter({0: 2650, 1: 176})
wave: 3
Counter({0: 2466, 1: 232})
wave: 4
Counter({0: 2745, 1: 103})
wave: 5
Counter({0: 2568, 1: 165})
```

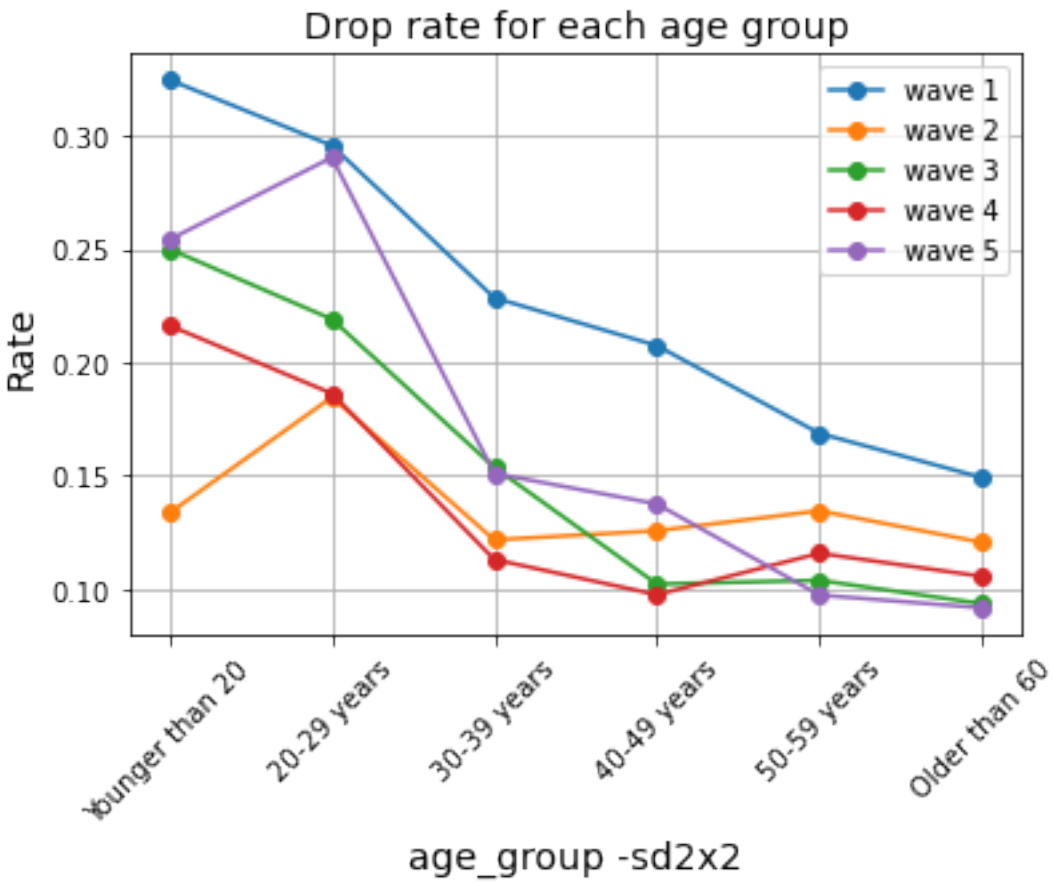


While it seems very clear that the people who answered on the *weekend* are *more likely* to drop out, the number of respondents on the weekend was **much** lower (typically **less than 10%**).

1.3 Age and gender

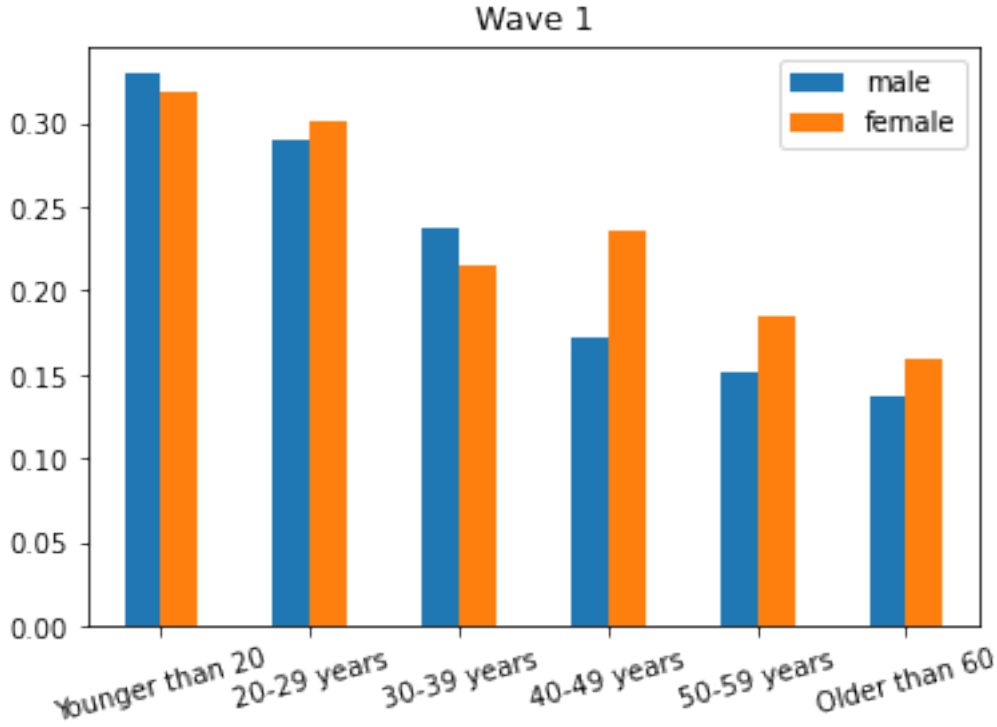
1.3.1 The effect of age on attrition

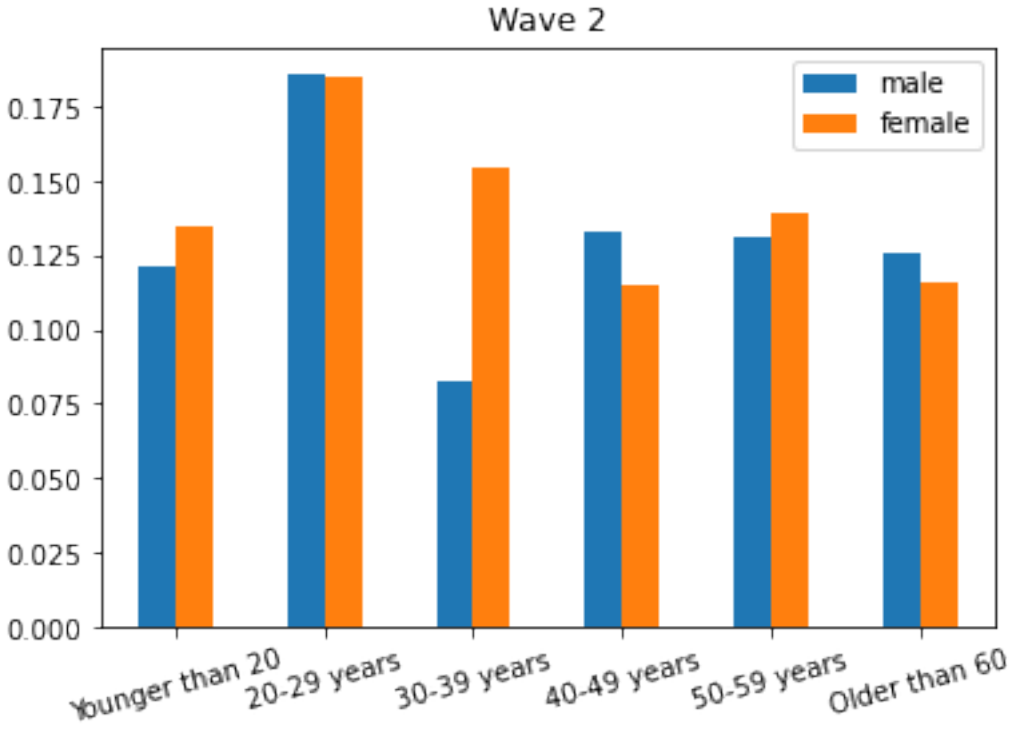
Age consistently showed up as one, if not the, most important feature across models (logistic regression, DT).

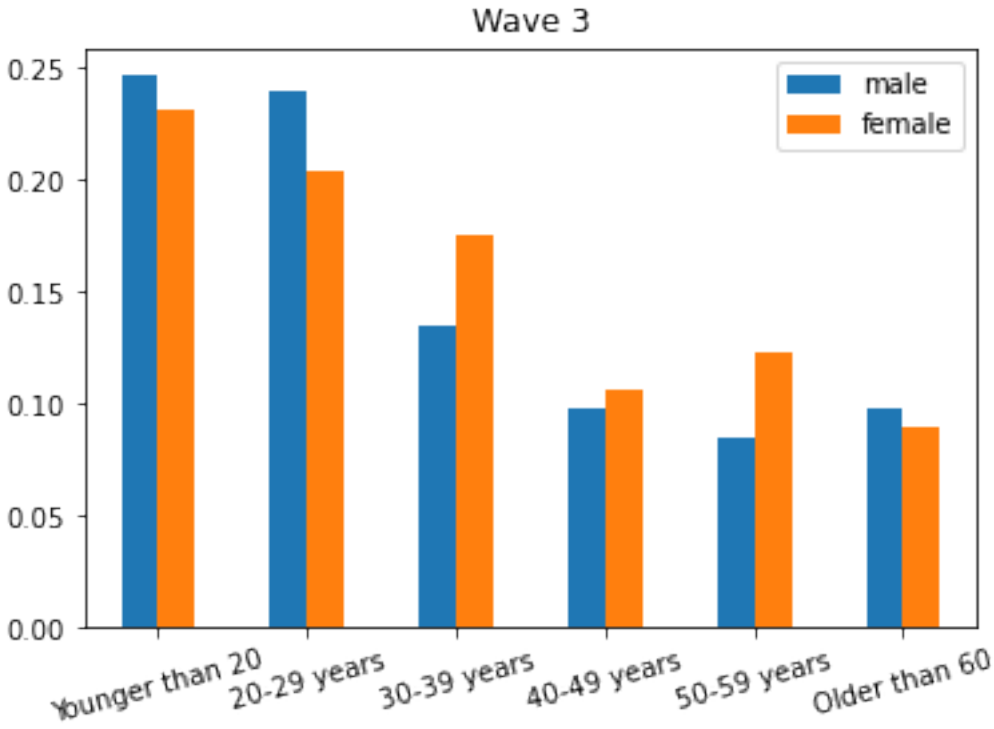


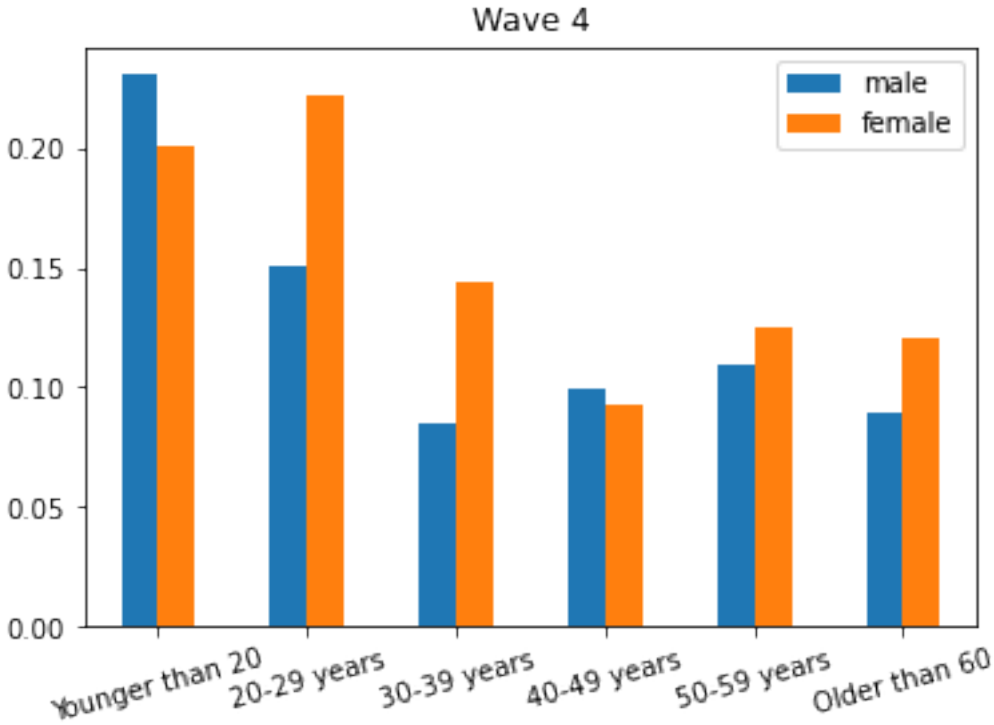
Clearly, the older respondents are less likely to drop the survey. However, if we exclude wave 1, there does not seem to be a significant change after the age of 30.

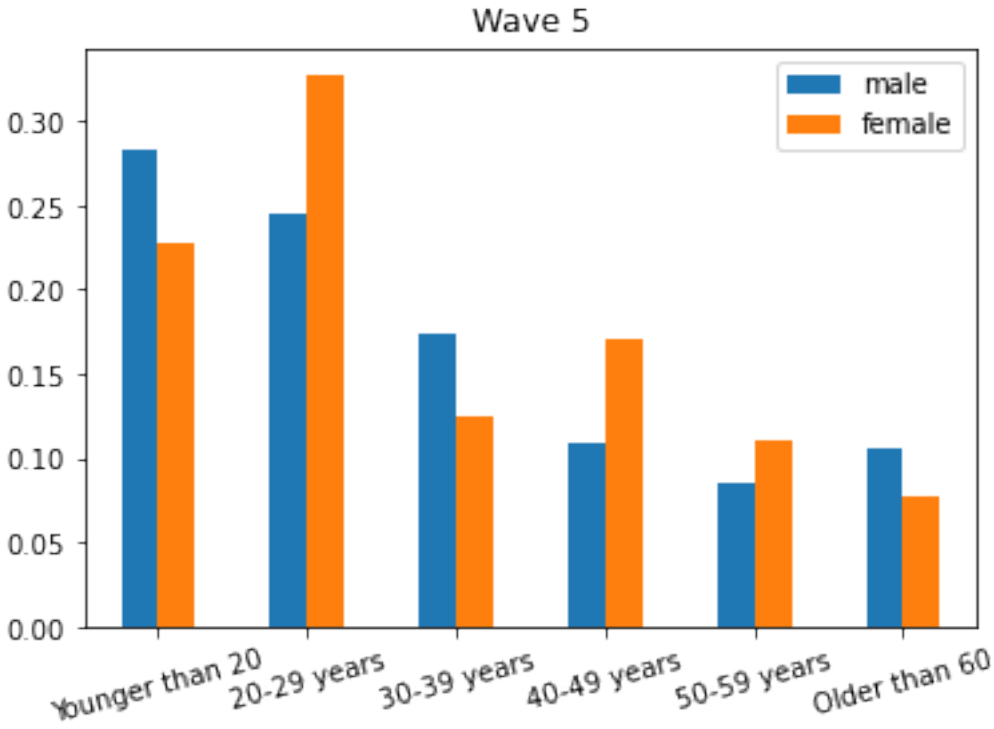
1.3.2 Attrition depending on gender and age






Wave 3
male
female
0.25
0.20
0.15
0.10
0.05
0.00
Younger than 20
20-29 years
30-39 years
40-49 years
50-59 years
Older than 60


Wave 4
male
female
0.20
0.15
0.10
0.05
0.00
Younger than 20
20-29 years
30-39 years
40-49 years
50-59 years
Older than 60



We can see the largest differences between genders at the age of 20-59, but no consistent trends among all the waves.

```
Counter({0: 3595, 1: 21})
Counter({0: 2816, 1: 10})
Counter({0: 2682, 1: 16})
Counter({0: 2834, 1: 14})
Counter({0: 2720, 1: 13})
```

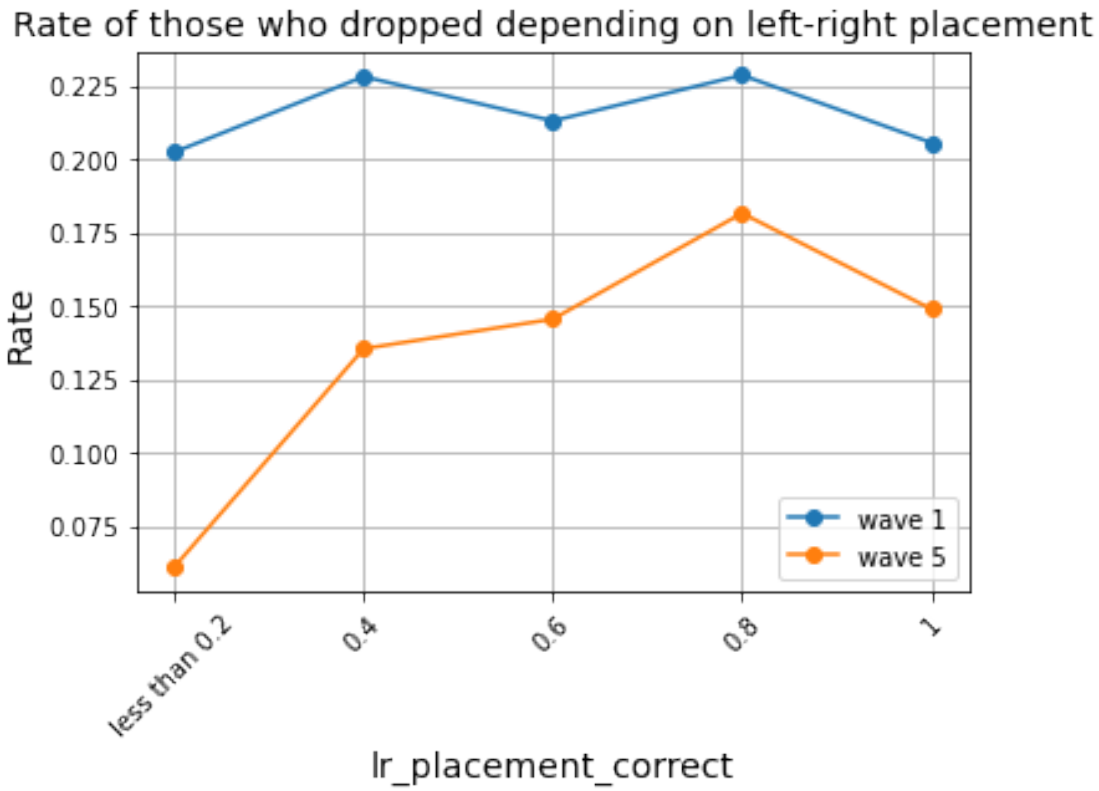
Drop rate for people who preferred not to say gender is much larger, up to 70-90% (but it is only 13-21 respondents for each wave, so due to scaling reasons we did not plot it with female and male).

1.4 Quality/behavioral issues

1.4.1 Correct left-right assessment

Here we investigate our engineered feature which measures the percentage of correctly assessed parties and politicians on a political left-right spectrum.

```
wave: 1
Counter({1.0: 1712, 0.8: 1085, 0.6: 507, 0.4: 228, 0.2: 84})
wave: 5
Counter({1.0: 1376, 0.8: 782, 0.6: 371, 0.4: 155, 0.2: 49})
```

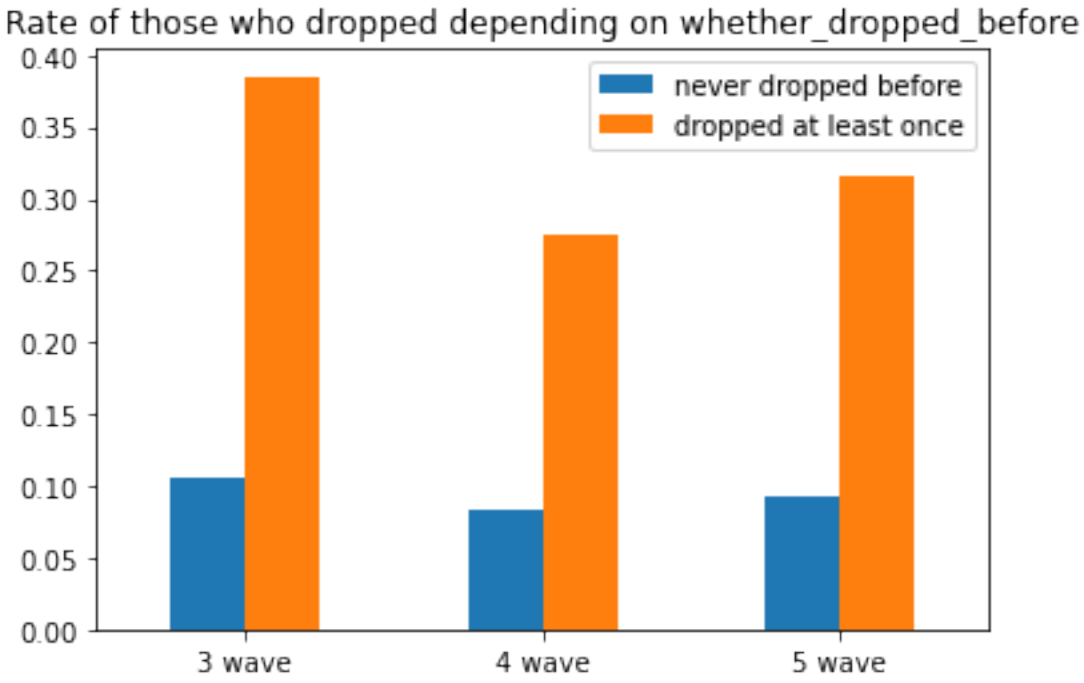


LR placement questions were only in waves 1, 4, 5. In wave 4 nearly 95% values are missing (we turned to 0 all features with >50% of values missing). After all cannot be sure about whether it makes sense (in wave 5 the more correct answers - more likely to drop).

1.4.2 Will they do it again?

If people miss one wave but return at a later point - are they more likely to show up for a later wave as well?

```
wave: 1
Counter({0: 3616})
wave: 2
Counter({0: 2826})
wave: 3
Counter({0: 2428, 1: 270})
wave: 4
Counter({0: 2125, 1: 723})
wave: 5
Counter({0: 1965, 1: 768})
```



If a person dropped at least once in wave 1 or 2 then probability that they drop again is ~200% higher!

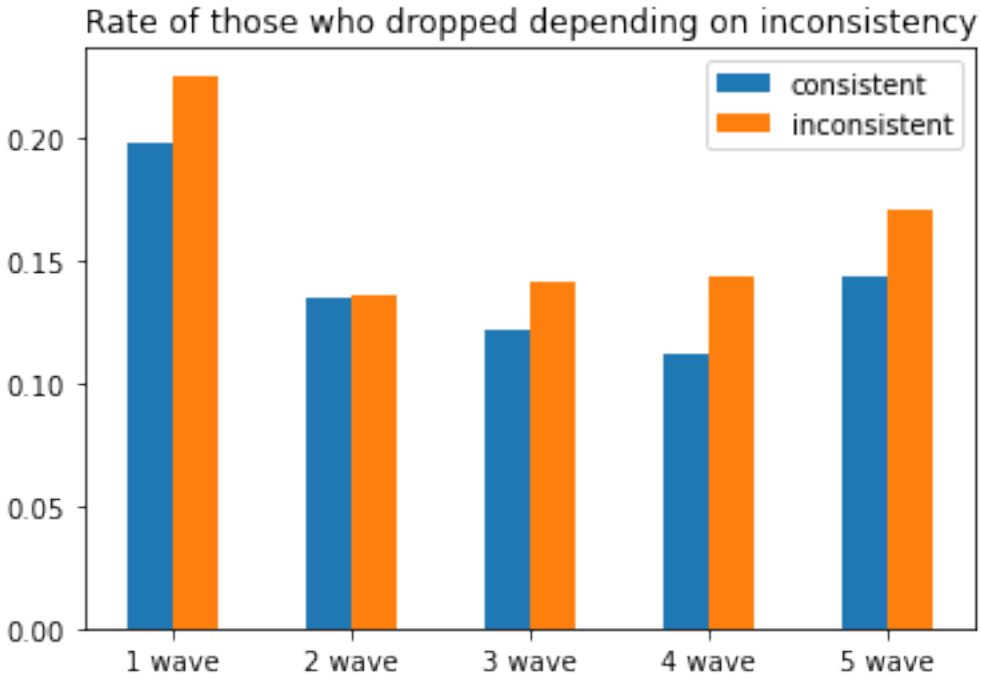
1.4.3 Is consistency in answers relevant for attrition?

Here we investigate the relevance of another engineered feature. For this we picked some questions which are very similar (or different) and measured if peoples answers are consistent (only higher degrees of sureness are taken into account, e.g. if person strongly agrees that they feel like a stranger due to many muslims but at the same wave strongly agrees that european and muslim lifestyles are easily comparable, then such a sample considered as inconsistent)

Manually, we selected the following questions: * PREFER INDEPENDENT CITIZEN INSTEAD OF A PARTY MEMBER -w1_q27x8
* THE PEOPLE SHOULD TAKE MOST IMPORTANT DECISIONS, NOT POLITICIANS -w1_q27x7 * FEELING LIKE A STRANGER DUE TO THE MANY MUSLIMS -w2_q21x4, * EUROPEAN AND MUSLIM LIFESTYLE ARE EASILY COMPATIBLE -w2_q21x5, * PEOPLE LIKE ME HAVE RECEIVED LESS THAN THEY DESERVE -w3_q35x1 * PEOPLE LIKE ME GET LESS ATTENTION THAN OTHERS -w3_q35x2 * SAME ACCESS TO SOCIAL BENEFITS: ASYLUM SEEKERS -w4_q65x2, * SAME ACCESS TO SOCIAL BENEFITS: NON-AUSTRIANS -w4_q65x1, * IMMIGRANTS GET MORE ATTENTION -w5_q30x2
* IMMIGRANTS HAVE RECEIVED MORE THAN THEY DESERVE -w5_q30x1

```
wave: 1
Counter({1: 2180, 0: 1436})
wave: 2
Counter({0: 2723, 1: 103})
```

```
wave: 3
Counter({1: 1552, 0: 1146})
wave: 4
Counter({1: 1803, 0: 1045})
wave: 5
Counter({0: 1540, 1: 1193})
```

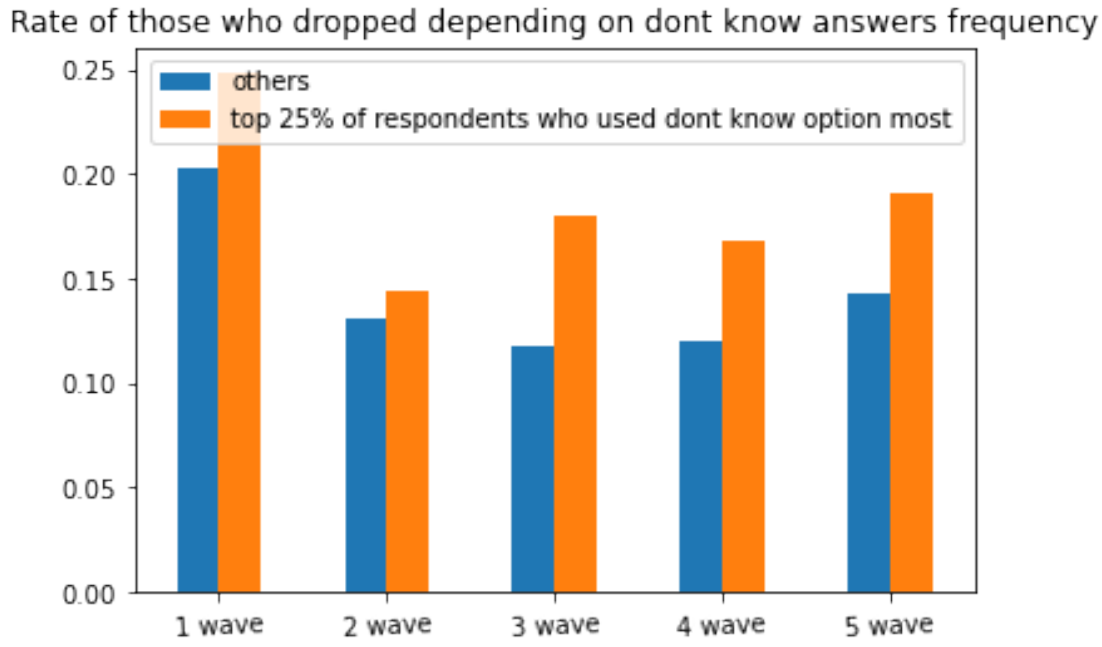


No significant differences, but "inconsistent" respondents tend to drop a bit more: 2-4 percentage points

1.4.4 Plato, Sokrates and a *know-it-all* walk into a cave...

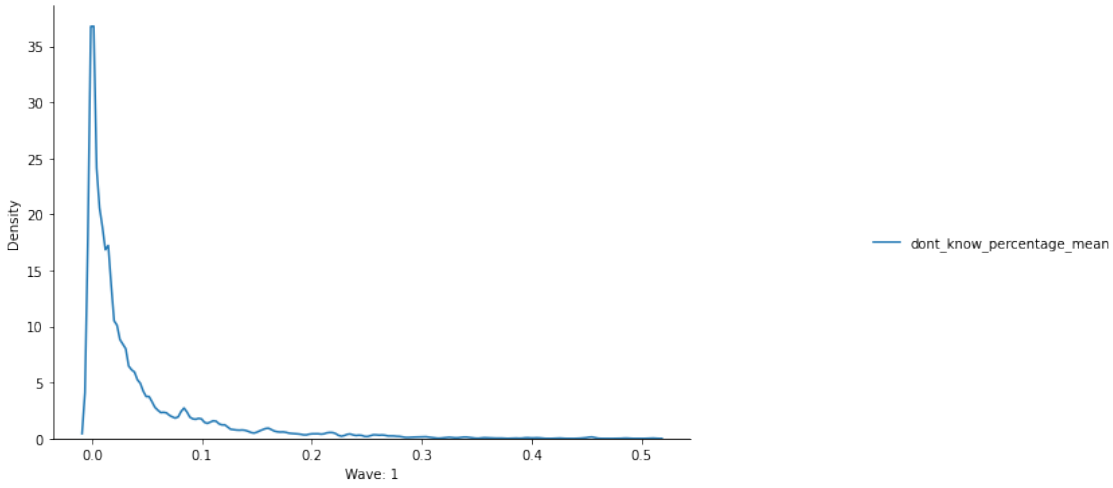
Lets look at the top quartil of respondents in terms of the number of `don't know` they chose.

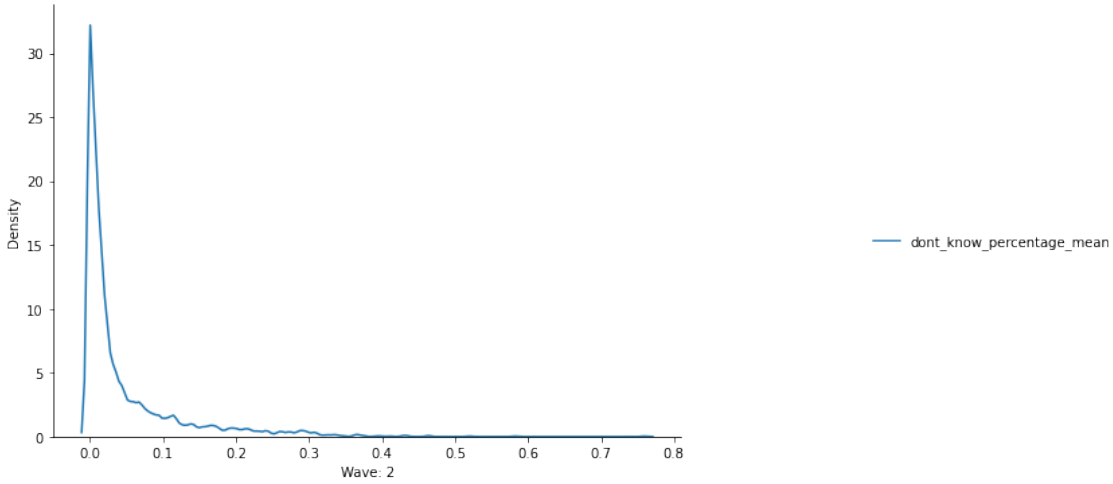
```
wave: 1
Counter({0.0: 2699, 1.0: 917})
wave: 2
Counter({0.0: 2119, 1.0: 707})
wave: 3
Counter({0.0: 2022, 1.0: 676})
wave: 4
Counter({0.0: 2134, 1.0: 714})
wave: 5
Counter({0.0: 1998, 1.0: 735})
```

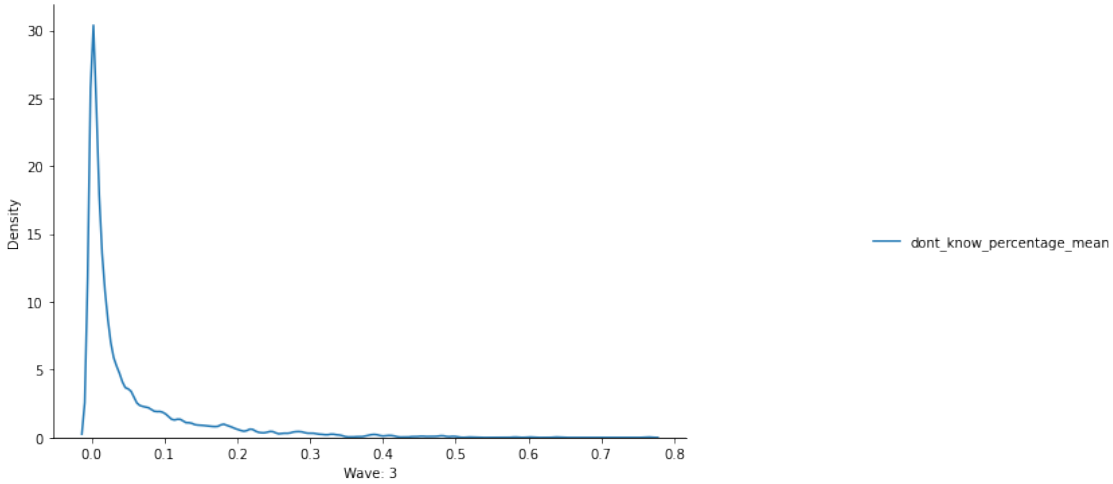


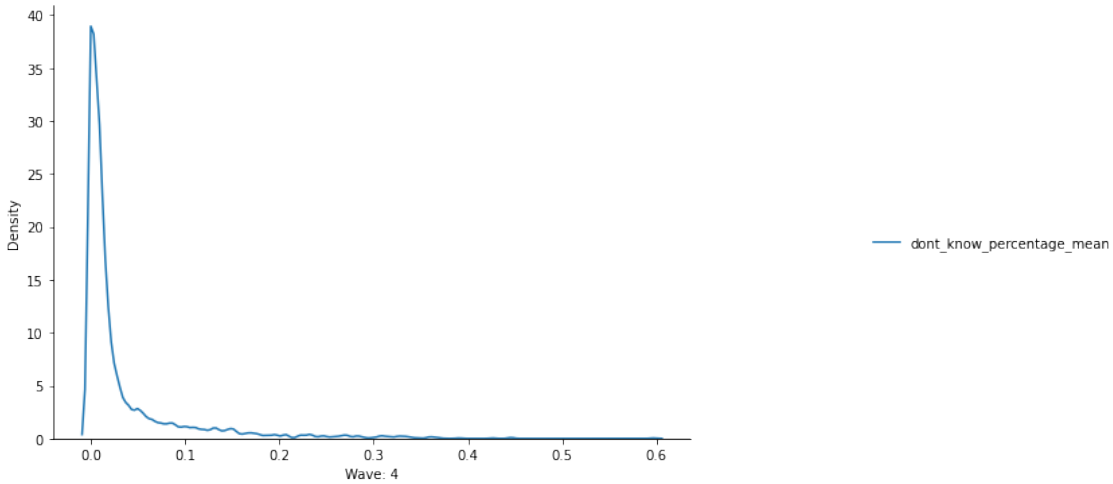
Drop rate slightly increases but only up to 5 percentage points (binary feature: top 25% with largest number of "don't know" responses are 1, tend to drop more often)

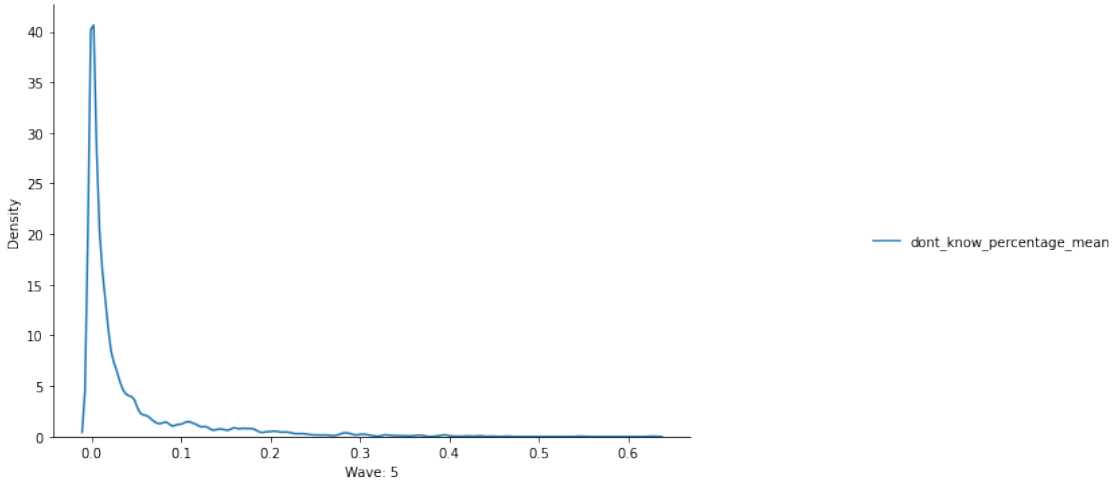
I might make sense to look at the distribution first. Maybe there is a linear trend? Maybe only the top 5% are different?




Density
Wave: 2
dont_know_percentage_mean

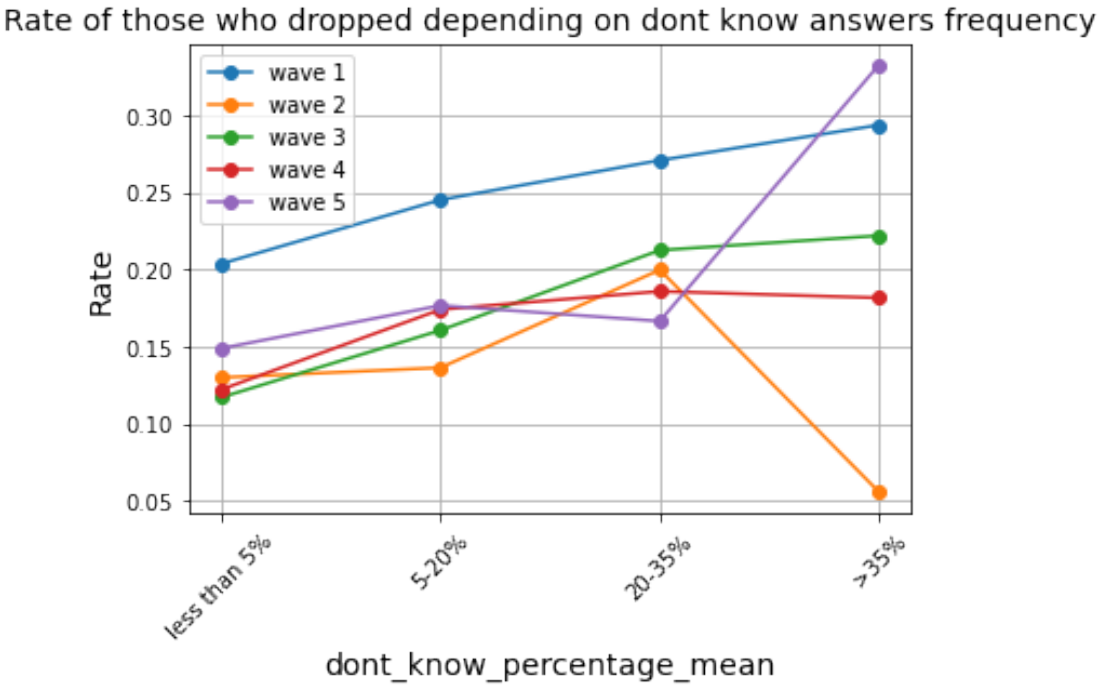

Density
Wave: 3
dont_know_percentage_mean


Density
Wave: 4
dont_know_percentage_mean



Most of the respondents choose "don't know" option ("don't know", 77, 88 (for numeric questions)) in less than 5% cases among all the waves. Hence is makes sense to check if there is linear trend for different rates of "don't know" number.

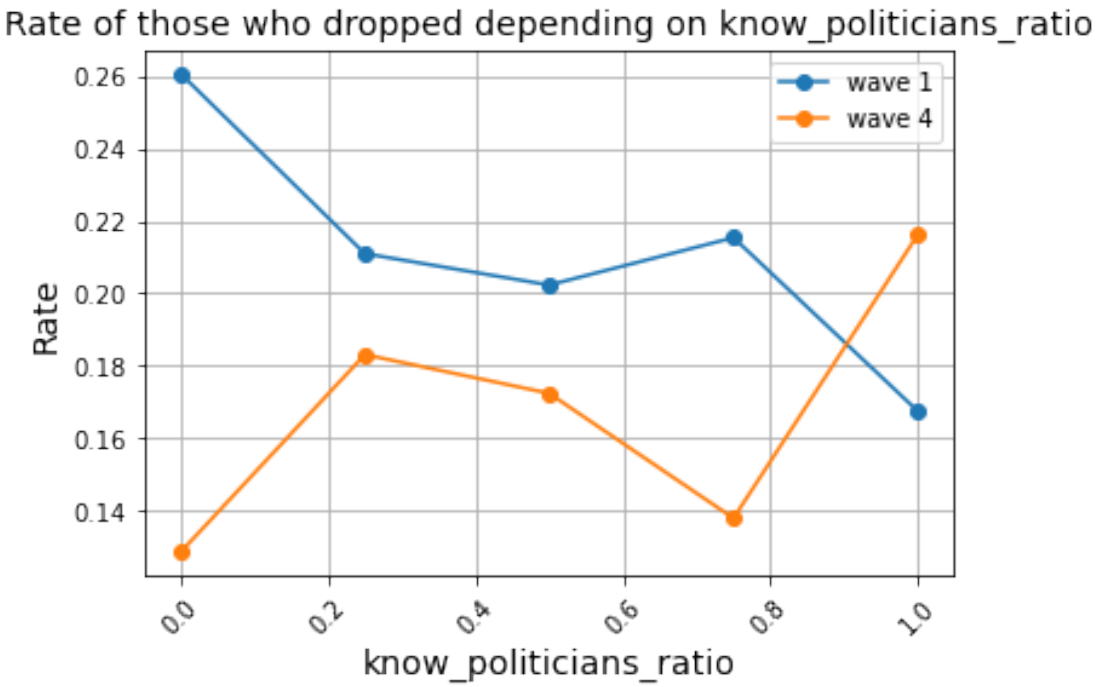
```
Counter({0.05: 2760, 0.15: 721, 0.3: 118, 0.5: 17})
Counter({0.05: 2084, 0.15: 579, 0.3: 145, 0.5: 18})
Counter({0.05: 1921, 0.15: 591, 0.3: 141, 0.5: 45})
Counter({0.05: 2332, 0.15: 419, 0.3: 86, 0.5: 11})
Counter({0.05: 2175, 0.15: 441, 0.3: 96, 0.5: 21})
```



We can see that probability to drop increases for 5-15 p.p. if one chooses "don't know" in more than 35% cases, but the trend is quite flat.

1.4.5 Correct left-right assessment of politicians only

```
wave: 1
Counter({0.0: 1055, 1.0: 806, 0.25: 692, 0.5: 608, 0.75: 455})
wave: 4
Counter({0.0: 2653, 0.25: 71, 0.5: 58, 1.0: 37, 0.75: 29})
```



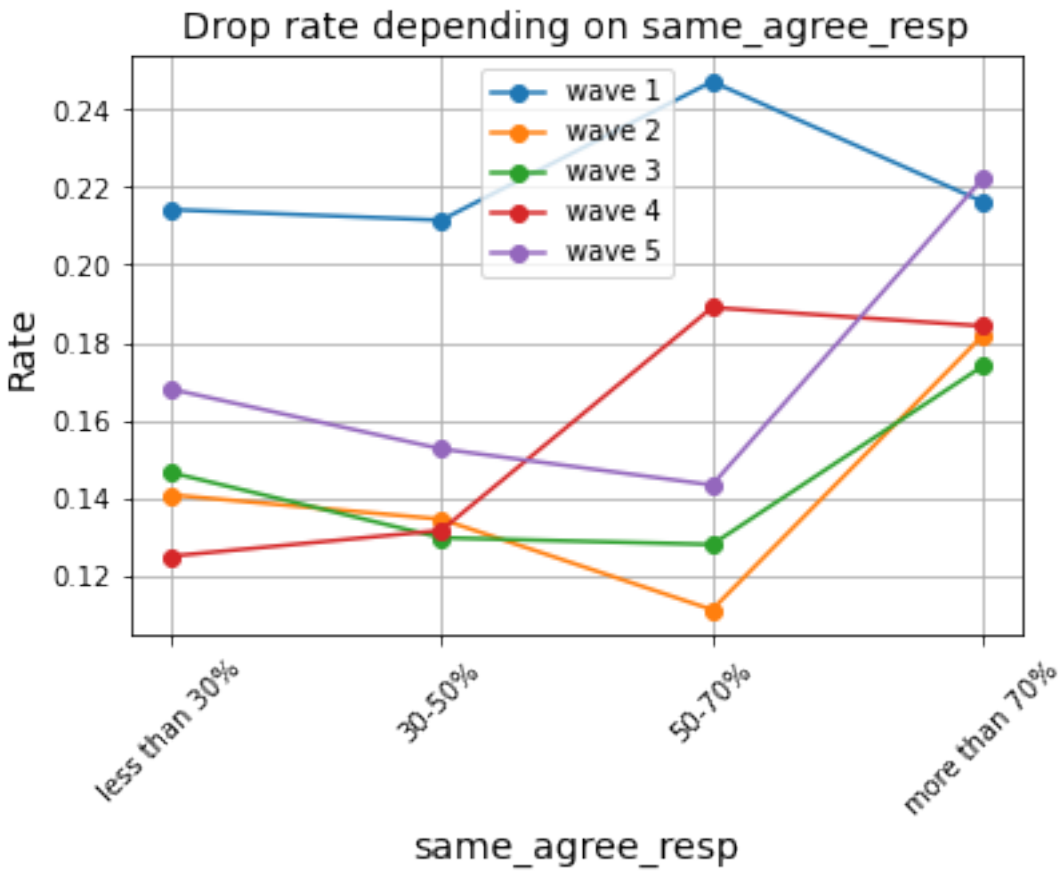
Question was asked only in wave 1, 4, 6

Features coded identically, but results are opposite! (these questions were asked to only 10% of respondents in wave 4 while in wave 1 everyone was asked, NaN coded as 0)

1.4.6 *Straightlining*

We try to capture **straightlining** (at least in the context of opinion questions. We do so by measuring how often the same degree of agreement was chosen.

```
wave: 1
Counter({0.5: 2369, 0.3: 943, 0.7: 267, 1.0: 37})
wave: 2
Counter({0.5: 2135, 0.3: 469, 0.7: 189, 1.0: 33})
wave: 3
Counter({0.5: 1882, 0.3: 512, 0.7: 258, 1.0: 46})
wave: 4
Counter({0.5: 1490, 0.3: 1193, 0.7: 127, 1.0: 38})
wave: 5
Counter({0.5: 1901, 0.3: 381, 0.7: 370, 1.0: 81})
```



People who picked the same degree of agreement on more than 70% of questions are more likely to drop out in the next wave.

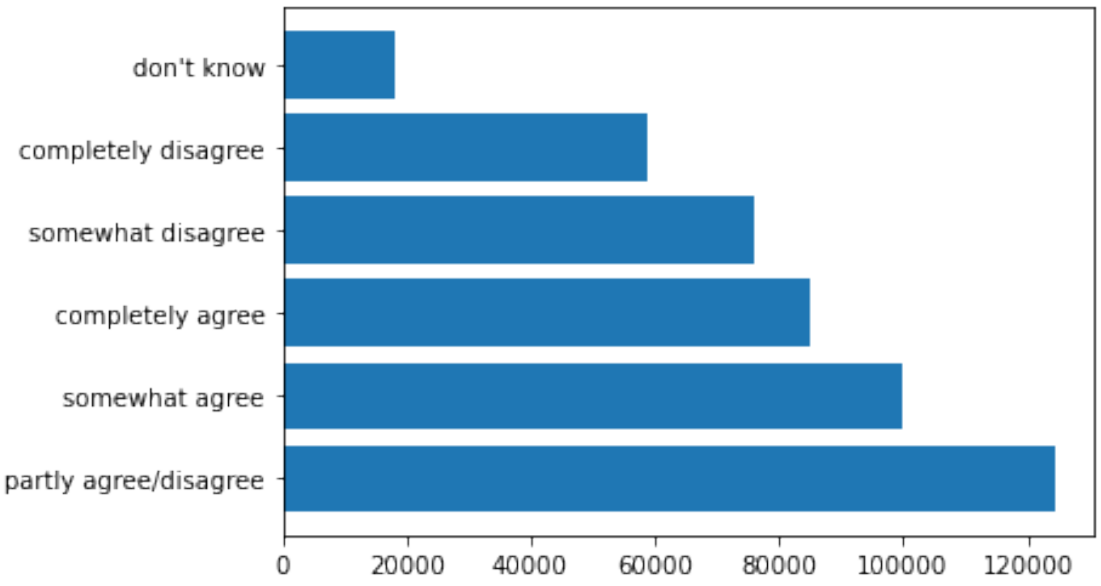
1.5 Distribution of answers for opinion responses

(e.g. immigration to Austria only in exceptional cases)

straightliners (chosen exactly the same answer in >= 60% of cases) are excluded

```
Initial number of respondents:  4453
Number of respondents after dropping straightliners:  4421
```

[65]: `<BarContainer object of 6 artists>`



"partly argee/disagree seems" to be versatile answer => most frequent one ("strongly agree/disagree" and "completely agree/disagree" are merged here)

2 The strongest correlations

[67]:
```
feature 1  \
15961  SAME ACCESS TO SOCIAL BENEFITS: EASTERN EUROPEANS (NEW EU MEMBER STATES)
-w4_q65x3
15807                            SAME ACCESS TO SOCIAL BENEFITS: ASYLUM SEEKERS
-w4_q65x2
10793                           IMMIGRANTS HAVE RECEIVED MORE THAN THEY DESERVE
-w3_q33x1
17633                           IMMIGRANTS HAVE RECEIVED MORE THAN THEY DESERVE
-w5_q30x1
10173                                                ASYLUM SEEKERS TAKE AWAY JOBS
-w3_q24x1
10791                           IMMIGRANTS HAVE RECEIVED MORE THAN THEY DESERVE
-w3_q33x1
15355                      CRIME RATES INCREASE IN AUSTRIA BECAUSE OF IMMIGRANTS
-w4_q52x5
10789                           IMMIGRANTS HAVE RECEIVED MORE THAN THEY DESERVE
-w3_q33x1
19595                        IMMIGRANTS PAY MORE INTO THE SOCIAL SECURITY SYSTEM
-w5_q43x6
8661                       CRIME RATES INCREASE IN AUSTRIA BECAUSE OF IMMIGRANTS
-w3_q18x5
```

```
feature 2  \
15961  SAME ACCESS TO SOCIAL BENEFITS: WESTERN EUROPEANS (OLD EU MEMBER STATES)
-w4_q65x4
15807                              SAME ACCESS TO SOCIAL BENEFITS: NON-AUSTRIANS
-w4_q65x1
10793                                          IMMIGRANTS GET MORE ATTENTION
-w3_q33x2
17633                                          IMMIGRANTS GET MORE ATTENTION
-w5_q30x2
10173                                      IMMIGRANTS TAKE AWAY AUSTRIAN JOBS
-w3_q18x4
10791                                          DUTY TO ACCEPT ASYLUM SEEKERS
-w3_q24x4
15355                             SAME ACCESS TO SOCIAL BENEFITS: ASYLUM SEEKERS
-w4_q65x2
10789                               EQUAL MEANS-TESTED INCOME FOR ASYLUM SEEKERS
-w3_q24x2
19595                            IMMIGRANTS HAVE RECEIVED MORE THAN THEY DESERVE
-w5_q30x1
8661                                       IMMIGRANTS ENRICH AUSTRIAN CULTURE
-w3_q18x2

       correlation
15961     0.735640
15807     0.741283
10793     0.766760
17633     0.792642
10173     0.817971
10791    -0.644819
15355    -0.639604
10789    -0.635237
19595    -0.609062
8661     -0.608252
```

These questions were candidates for inconsistency depending on correlation and logical relation between possible responses

No clusters among the most correlated features (only with opinion questions related to immigrants and asylum seekers). In the next step of clustering these questions were the most important to split respondents into groups

```
[NbConvertApp] Converting notebook data_exploration.ipynb to pdfviahtml
[NbConvertApp] Writing 1300355 bytes to output_pdf\data_exploration.pdf
```