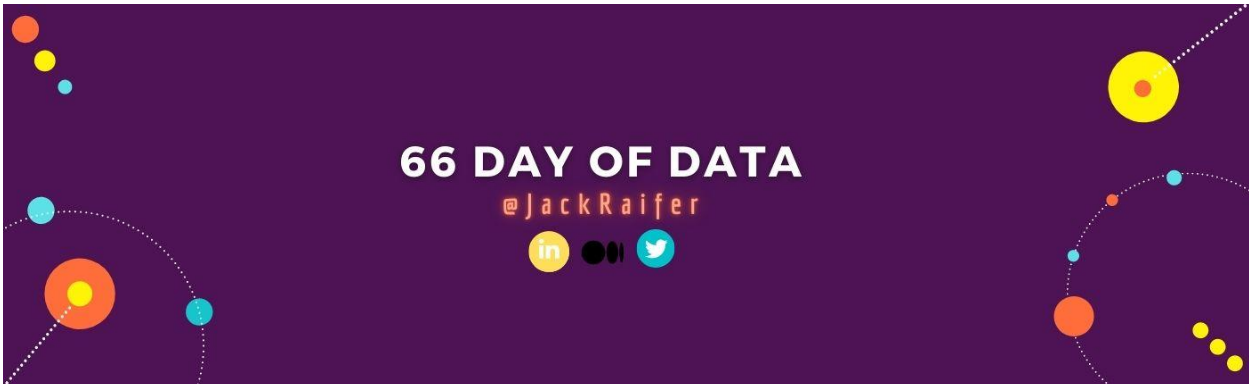
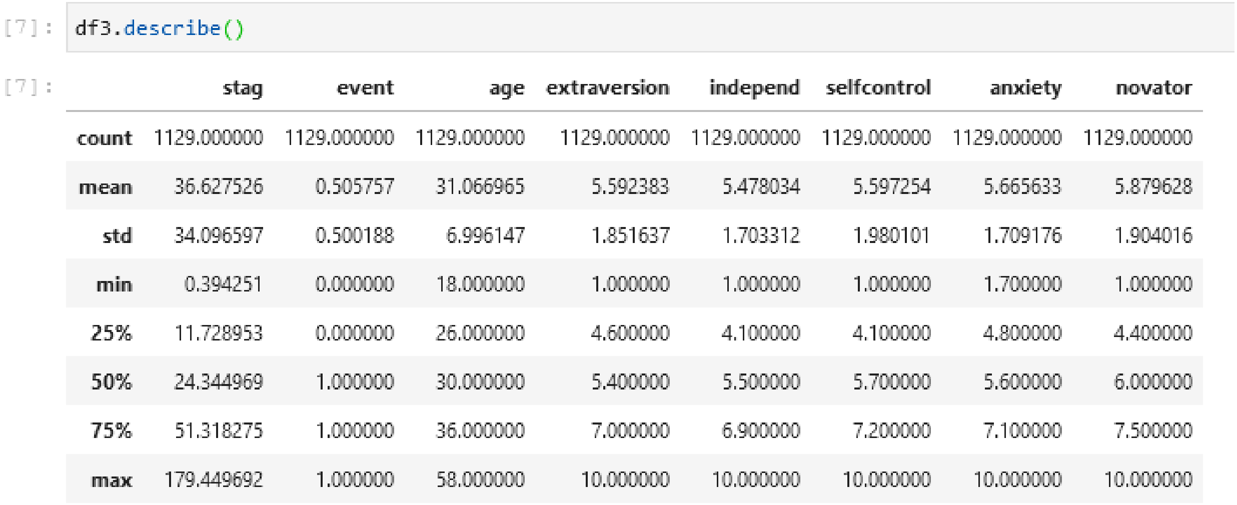


#66DaysOfData – Day 9: One more dataset and making decisions to move forward

Yesterday we cracked open the general HR and the IBM HR datasets and got some interesting insights. The bad news is: we have not yet found any set that can correlate attrition to the OCEAN personality model.

Today, we open one more dataset, the Turnover data set by Edward Babushkin. Here is the data after running the describe method on the dataframe:



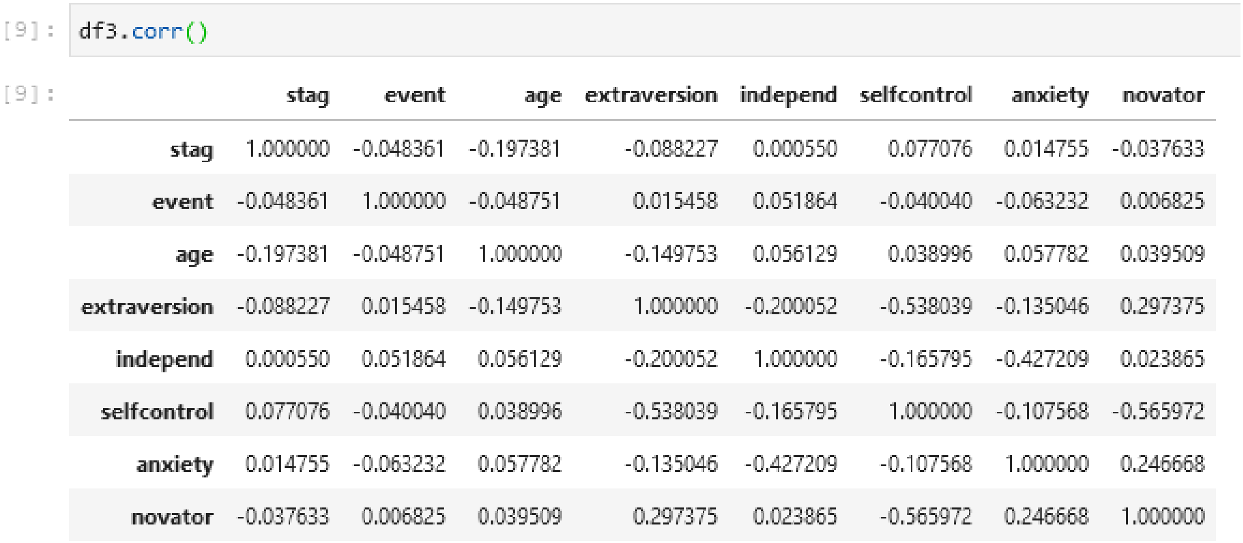
```
[7]: df3.describe()
```

[7]:

	stag	event	age	extraversion	independ	selfcontrol	anxiety	novator
count	1129.000000	1129.000000	1129.000000	1129.000000	1129.000000	1129.000000	1129.000000	1129.000000
mean	36.627526	0.505757	31.066965	5.592383	5.478034	5.597254	5.665633	5.879628
std	34.096597	0.500188	6.996147	1.851637	1.703312	1.980101	1.709176	1.904016
min	0.394251	0.000000	18.000000	1.000000	1.000000	1.000000	1.700000	1.000000
25%	11.728953	0.000000	26.000000	4.600000	4.100000	4.100000	4.800000	4.400000
50%	24.344969	1.000000	30.000000	5.400000	5.500000	5.700000	5.600000	6.000000
75%	51.318275	1.000000	36.000000	7.000000	6.900000	7.200000	7.100000	7.500000
max	179.449692	1.000000	58.000000	10.000000	10.000000	10.000000	10.000000	10.000000

Looking at the columns we find some names like extraversion, independ, selfcontrol, anxiety and novator. These seem to be personality factors from some model, but I am not sure what model it is (already reached out to the author and waiting for a response). It does seem like a 5 factor model, similar to OCEAN and we could venture a guess as to which of these is equivalent to the big five, but let us hold on before doing that.

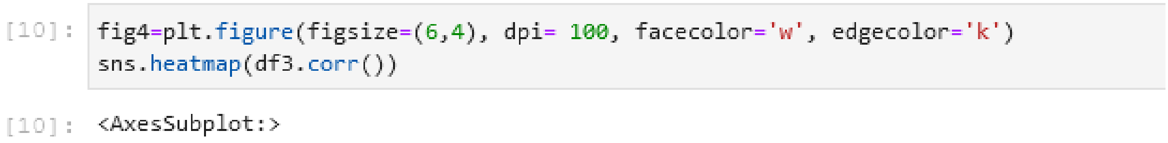
Putting that confusion aside, we do have another interesting column here, the event column, which, according to the author, relates to turnover, did the person leave the company or was fired (this second one is a bit bothersome, but nothing is perfect in the world of data). Let us see if we find any correlation between the personality factors and the event (turnover) column:



```
[9]: df3.corr()
```

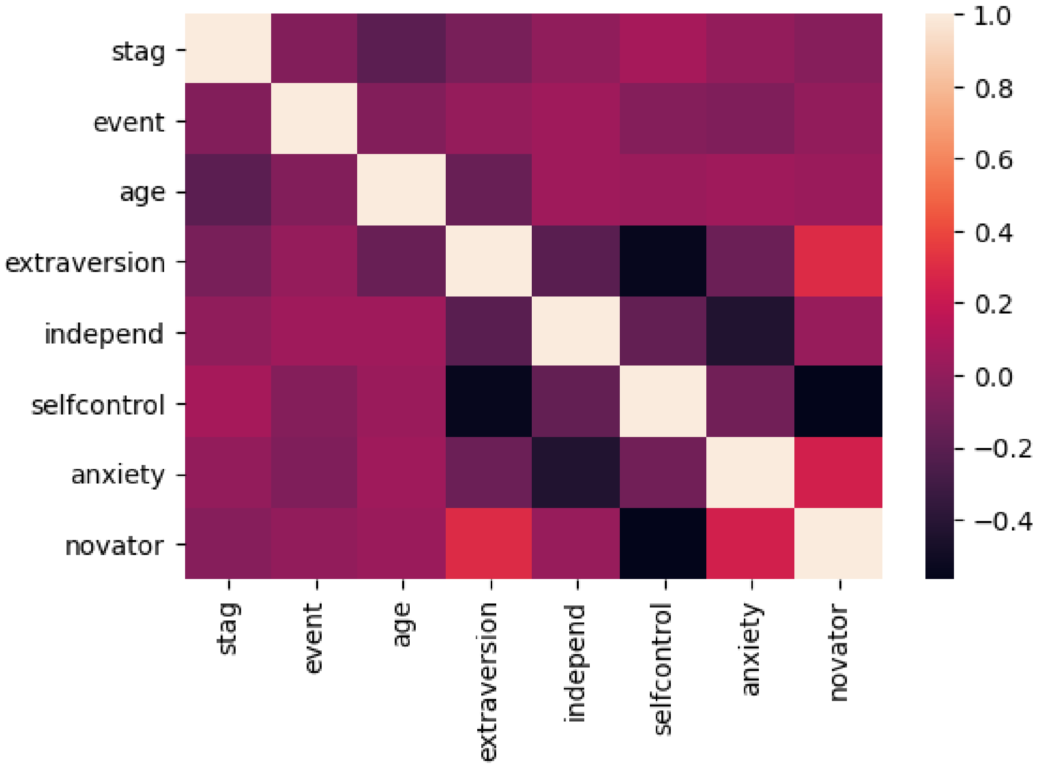
[9]:

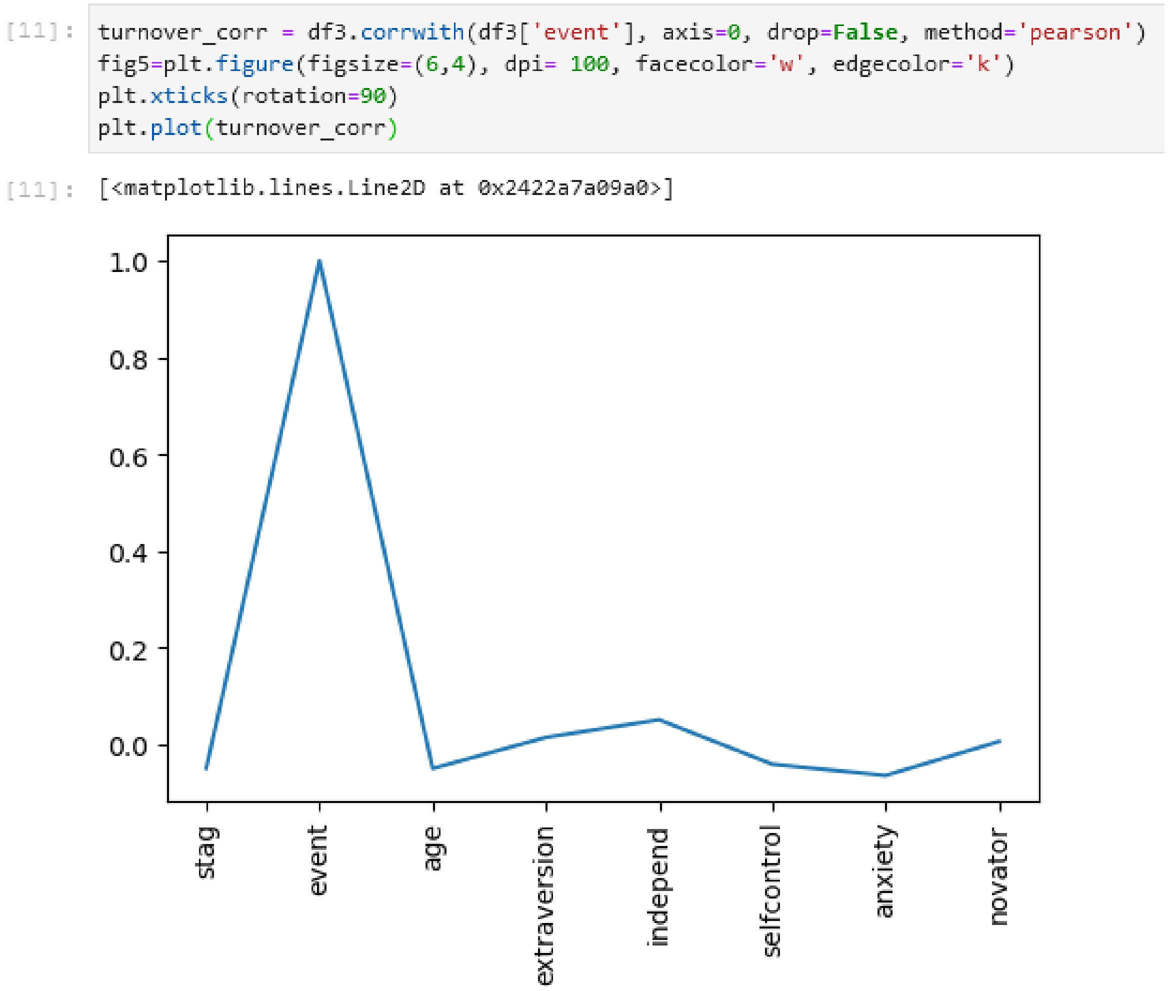
	stag	event	age	extraversion	independ	selfcontrol	anxiety	novator
stag	1.000000	-0.048361	-0.197381	-0.088227	0.000550	0.077076	0.014755	-0.037633
event	-0.048361	1.000000	-0.048751	0.015458	0.051864	-0.040040	-0.063232	0.006825
age	-0.197381	-0.048751	1.000000	-0.149753	0.056129	0.038996	0.057782	0.039509
extraversion	-0.088227	0.015458	-0.149753	1.000000	-0.200052	-0.538039	-0.135046	0.297375
independ	0.000550	0.051864	0.056129	-0.200052	1.000000	-0.165795	-0.427209	0.023865
selfcontrol	0.077076	-0.040040	0.038996	-0.538039	-0.165795	1.000000	-0.107568	-0.565972
anxiety	0.014755	-0.063232	0.057782	-0.135046	-0.427209	-0.107568	1.000000	0.246668
novator	-0.037633	0.006825	0.039509	0.297375	0.023865	-0.565972	0.246668	1.000000



```
[10]: fig4=plt.figure(figsize=(6,4), dpi= 100, facecolor='w', edgecolor='k')
      sns.heatmap(df3.corr())

[10]: <AxesSubplot:>
```





```
[11]: turnover_corr = df3.corrwith(df3['event'], axis=0, drop=False, method='pearson')
      fig5=plt.figure(figsize=(6,4), dpi= 100, facecolor='w', edgecolor='k')
      plt.xticks(rotation=90)
      plt.plot(turnover_corr)

[11]: [<matplotlib.lines.Line2D at 0x2422a7a09a0>]
```

As we can see, there does not seem to be any significant correlation between the event (turnover) and personality. Also, we have no idea how much of that turnover data has to do with attrition (quitting) or the person being fired, which makes it even less appealing.

What are my options now?

1. I could keep searching for more data, but after a few days of searching I do not believe it will be easy to find. Not having the perfect dataset is common, it is just a matter of how to move forward.

2. I could also completely scrap the project and think of something else. Well, although this crossed my mind a few times, after reading the research I found on day 2 (click here to check it out), I am certain that there is merit to the hypothesis of personality factors being able to predict attrition, and although I have yet to be able to access the datasets from these studies, I do have the numbers they found. So that leads me to a third option.

3. Since I do have one exceptionally clean, very useful and big enough (over 300 thousand observations) set of OCEAN data, I could use the research to build an algorithm that could, within

certain constraints, create the data for attrition. Although this could be considered like a bit of cheating, this is where domain knowledge becomes truly relevant, since it could allow us to build a theoretical predictive model. I agree, not good as the real thing, but with something on hand it could be a lot easier to convince an HR department to let us experiment with it, and if this happens, we could find the model works well if we are lucky, and worst case we are collecting real data that we can either retrain the model with or train a new model based on real data.

So on to creating a basic algorithm, based on research, that can help us create the data we are missing. Forward towards creating an OCEAN based attrition prediction model. And if we crash and burn with the project, we will have learned plenty of lessons.

Next Time – Creating an algorithm.

Jack Raifer Baruch

Follow me on Twitter: @JackRaifer

Follow me on LinkedIN: jackraifer

About the Road to Data Science - #66DaysOfData Series

Road to Data Science series began after I experienced the first round of Ken Jee´s #66DaysOfData challenge back in 2020. Since we are starting the second round of the challenge, I thought it would be a good idea to add small articles every day where I can comment my progress.

I will be sharing all the notebooks, articles and data I can on GitHub: https://github.com/jackraifer/66DaysOfData-Road-to-Data-Science

Please do understand I might have to withhold some information, including code, data, visualizations and/or models, because of confidentiality regards. But I will try to share as much as possible.

Want to follow the #66DaysOfDataChallenge?

Just follow Ken Jee on twitter **@KenJee_DS** and join the #66DaysOfData challenge.

You can also reach out to me at any time through **LinkedIN** or **Twitter.**