Final Project Step 4 PCA

```
Course:   DS 5001
Module:   Final
Date:     8 May 2022
Author:   Thomas McIntyre gem5cm@virginia.edu
Purpose:  This notebook will utlize the data created in step 2 to perform
PCA and visualize the components.
```

In [1]:

```
import pandas as pd
import numpy as np
# from sklearn.decomposition import PCA
from scipy.linalg import norm, eigh
import plotly_express as px
import seaborn as sns; sns.set()
import sys
%matplotlib inline
```

In [2]:

```
data_home = "data"
local_lib = "code"
OHCO = ['book_id', 'chap_num', 'para_num', 'sent_num', 'token_num']
SENTS = OHCO[:4]
PARAS = OHCO[:3]
CHAPS = OHCO[:2]
BOOKS = OHCO[:1]
```

In [3]:

```
LIB = pd.read_csv(f"{data_home}/LIB.csv").set_index(OHCO[:1])
CORPUS = pd.read_csv(f"{data_home}/CORPUS.csv").set_index(OHCO)
VOCAB = pd.read_csv(f"{data_home}/VOCAB.csv").set_index("term_str")
BOW = pd.read_csv(f"{data_home}/BOW.csv").rename(columns = {"Unnamed: 2": "term_str"}).
TFIDF = pd.read_csv(f"{data_home}/TFIDF.csv").set_index(CHAPS)
DOC = pd.read_csv(f"{data_home}/DOC.csv").set_index(CHAPS)
```

In [4]:

```
DOC = DOC.join(LIB, on='book_id')
DOC
```

Out[4]:

		doc_str	n_tokens	book_title	book_file	chap_
book_id	**chap_num**					
1	**1**	we should start back gared urged as the woods ...	3860	A Game of Thrones, by RR Martin	corpus/MARTIN_A_GAME_OF_THRONES-pg1.txt	[A-Z]+[A-Z
	2	the morning had dawned clear and cold with a c...	3037	A Game of Thrones, by RR Martin	corpus/MARTIN_A_GAME_OF_THRONES-pg1.txt	[A-Z]+[A-Z

		doc_str	n_tokens	book_title	book_file	chap_
book_id	chap_num					
	3	catelyn had never liked this godswood she had ...	2074	A Game of Thrones, by RR Martin	corpus/MARTIN_A_GAME_OF_THRONES-pg1.txt	[A-Z]+[A-Z
	4	her brother held the gown up for her inspectio...	4161	A Game of Thrones, by RR Martin	corpus/MARTIN_A_GAME_OF_THRONES-pg1.txt	[A-Z]+[A-Z
	5	the visitors poured through the castle gates i...	3801	A Game of Thrones, by RR Martin	corpus/MARTIN_A_GAME_OF_THRONES-pg1.txt	[A-Z]+[A-Z
...	...	...	...	...	...	
26654	13	do you believe in fairies the more quickly thi...	2873	Peter and Wendy, by James Matthew Barrie	corpus/BARRIE_PETER_AND_WENDY-pg26654.txt	((CHAPTER)\s
	14	the pirate ship one green light squinting over...	2337	Peter and Wendy, by James Matthew Barrie	corpus/BARRIE_PETER_AND_WENDY-pg26654.txt	((CHAPTER)\s
	15	hook or me this time odd things happen to all ...	3211	Peter and Wendy, by James Matthew Barrie	corpus/BARRIE_PETER_AND_WENDY-pg26654.txt	((CHAPTER)\s
	16	the return home by two bells that morning they...	2982	Peter and Wendy, by James Matthew Barrie	corpus/BARRIE_PETER_AND_WENDY-pg26654.txt	((CHAPTER)\s

		doc_str	n_tokens	book_title	book_file	chap_
book_id	**chap_num**					
	17	when wendy grew up i hope you want to know wha...	3801	Peter and Wendy, by James Matthew Barrie	corpus/BARRIE_PETER_AND_WENDY-pg26654.txt	((CHAPTER)\s

533 rows × 11 columns

In [5]:
```
top2000 = VOCAB.loc[VOCAB["max_pos"].isin(["NN","NNS", "VB", "VBD", "VBG", "VBN", "VBP"
top2000
```

Out[5]:

	n	tfidf_mean	df	dfidf	p	i	max_pos	n_pos	stop
term_str									
easy	264	0.003811	196	282.883248	0.000305	11.676997	JJ	10.0	0
present	359	0.005527	197	282.880161	0.000415	11.233551	JJ	9.0	0
fool	388	0.005420	195	282.878975	0.000449	11.121478	NN	10.0	0
big	489	0.006594	195	282.878975	0.000566	10.787701	JJ	6.0	0
loved	381	0.005573	195	282.878975	0.000441	11.147744	VBD	9.0	0
...	...	...	...	...	...	...	...	...	...
footsteps	72	0.001811	53	176.493777	0.000083	13.551466	NNS	5.0	0
wretch	80	0.001917	53	176.493777	0.000093	13.399463	NN	7.0	0
yielded	67	0.001791	53	176.493777	0.000078	13.655302	VBN	7.0	0
bench	99	0.002380	53	176.493777	0.000115	13.092035	NN	4.0	0
mocking	59	0.001500	53	176.493777	0.000068	13.838748	NN	3.0	0

2000 rows × 9 columns

In [6]:
```
TFIDF = TFIDF[top2000.index]
TFIDF
```

Out[6]:

		easy	present	fool	big	loved	darkness	steps	shut	
book_id	**chap_num**									
1	**1**	0.000000	0.000000	0.011165	0.011165	0.011165	0.033494	0.000000	0.000000	0.00
	2	0.009425	0.000000	0.000000	0.009449	0.009449	0.000000	0.000000	0.000000	0.01
	3	0.016520	0.000000	0.000000	0.000000	0.016562	0.016562	0.000000	0.000000	0.03
	4	0.006601	0.000000	0.013235	0.033088	0.006618	0.006618	0.000000	0.000000	0.00
	5	0.008575	0.000000	0.000000	0.008597	0.034388	0.051581	0.008532	0.000000	0.00

book_id	chap_num	easy	present	fool	big	loved	darkness	steps	shut	
...	...	...	...	...	...	...	...	...	...	
26654	13	0.014295	0.028517	0.000000	0.000000	0.057326	0.028663	0.000000	0.000000	0.0
	14	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.014083	0.0
	15	0.007893	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0
	16	0.000000	0.000000	0.000000	0.021334	0.000000	0.000000	0.000000	0.010586	0.0
	17	0.006578	0.006561	0.000000	0.013189	0.006595	0.006595	0.000000	0.000000	0.0

533 rows × 2000 columns

In [7]:

```
exec(open("code/methods.py").read())
LOADINGS, DCM, COMPINF = get_pca(TFIDF, k=10, norm_docs=True, center_by_mean=False, cen
```

In [8]:

```
LOADINGS
```

Out[8]:

pc_id term_str	0	1	2	3	4	5	6	7	
easy	-0.002884	0.007308	0.006637	-0.002790	-0.010034	-0.005328	-0.000367	-0.014834	-0.0246
present	0.038323	-0.007629	-0.020272	0.030516	-0.041704	-0.016748	0.021737	-0.010968	-0.0007
fool	-0.037409	0.023864	-0.038274	0.001731	-0.018462	-0.009816	0.002549	0.016554	0.0076
big	-0.046661	0.035626	0.049218	-0.009898	0.017349	0.009410	0.047200	-0.001070	-0.0023
loved	-0.008453	-0.016437	-0.033531	0.002029	0.003691	-0.056289	0.004368	-0.003358	0.0202
...	...	...	...	...	...	...	...	...	
footsteps	-0.003974	0.003658	0.000333	-0.008707	-0.006167	0.001133	0.000736	0.017392	-0.0109
wretch	0.008919	-0.020050	-0.006966	-0.007600	-0.006786	0.010699	0.017039	0.017996	0.0014
yielded	0.000186	0.003630	0.000756	-0.000111	-0.006991	-0.006691	0.015336	0.002343	-0.0014
bench	-0.006755	0.016277	0.004147	-0.014952	-0.011422	0.010962	0.017917	0.003427	0.0048
mocking	-0.006068	0.004620	-0.003585	0.006803	0.009306	-0.007563	-0.000830	-0.003530	0.0008

2000 rows × 10 columns

In [9]:

```
DCM
```

Out[9]:

	pc_id	0	1	2	3	4	5	6	7
book_id	chap_num								
1	1	-0.274565	0.128409	0.205919	0.031068	-0.204457	-0.063534	0.064545	0.081993
	2	-0.299369	0.092334	-0.045226	0.092463	-0.096451	-0.043168	0.211052	0.034234

	pc_id	0	1	2	3	4	5	6	7
book_id	**chap_num**								
	3	-0.229763	0.069198	-0.058106	0.083305	-0.133646	-0.153285	0.024623	-0.018159
	4	-0.305262	0.097910	-0.135328	0.131115	-0.146126	-0.036442	-0.130285	0.157397
	5	-0.345383	0.066444	-0.255441	0.065961	-0.186128	-0.064179	-0.093645	0.043644
...	**...**	...	...	...	...	...	...	...	..
26654	**13**	0.068367	0.026752	0.147885	0.151276	0.026228	-0.171708	0.026067	0.188403
	14	0.076021	-0.026692	0.030143	0.112838	0.152753	0.026854	-0.045133	0.137186
	15	0.032666	0.010361	0.128657	0.108654	0.155867	-0.000479	0.007253	0.281563
	16	0.137982	0.077220	-0.065199	0.200368	0.441117	-0.240444	0.038112	0.048022
	17	0.125263	0.102080	-0.050461	0.121180	0.317724	-0.270200	0.022683	0.093753

533 rows × 10 columns

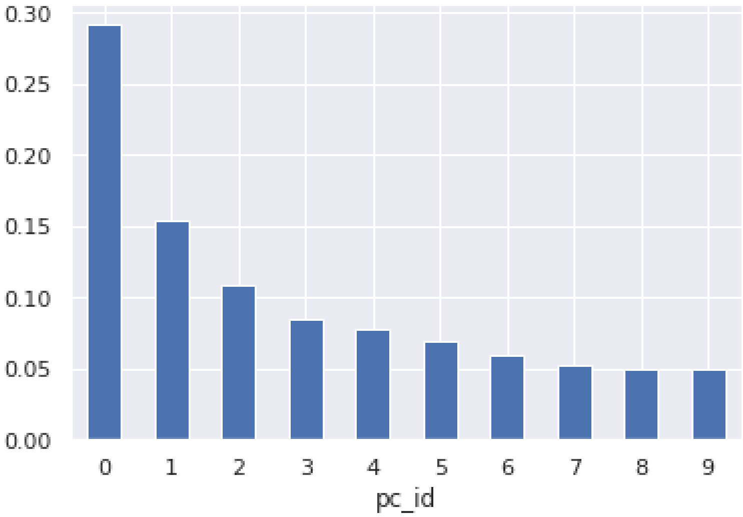
In [10]:
```
COMPINF
```

Out[10]:

	pos	neg	eig_val	exp_var
pc_id				
0	sir gentleman dear replied room	king men sword brother knight	0.050579	0.291869
1	lady boy girl door castle	heaven chief god war vain	0.026610	0.153558
2	trees wall tree water road	king lady father queen son	0.018965	0.109442
3	ship house sea home answered	gentleman boy replied dear sir	0.014813	0.085481
4	darling boys cried children mother	sir road master dark door	0.013624	0.078617
5	gentleman men ship dear boy	darling love lady oh children	0.011982	0.069142
6	father wall hes master hed	dragons king dragon sir gentleman	0.010402	0.060023
7	lady girl replied cried door	sir wall darling ice sister	0.009026	0.052084
8	dragons dragon sir blood city	king boy gentleman north son	0.008709	0.050257
9	sir gentleman replied lady mother	sister pocket kitchen queen forge	0.008582	0.049525

In [11]:
```
COMPINF.exp_var.plot.bar(rot=0)
```

Out[11]:
```
<AxesSubplot:xlabel='pc_id'>
```

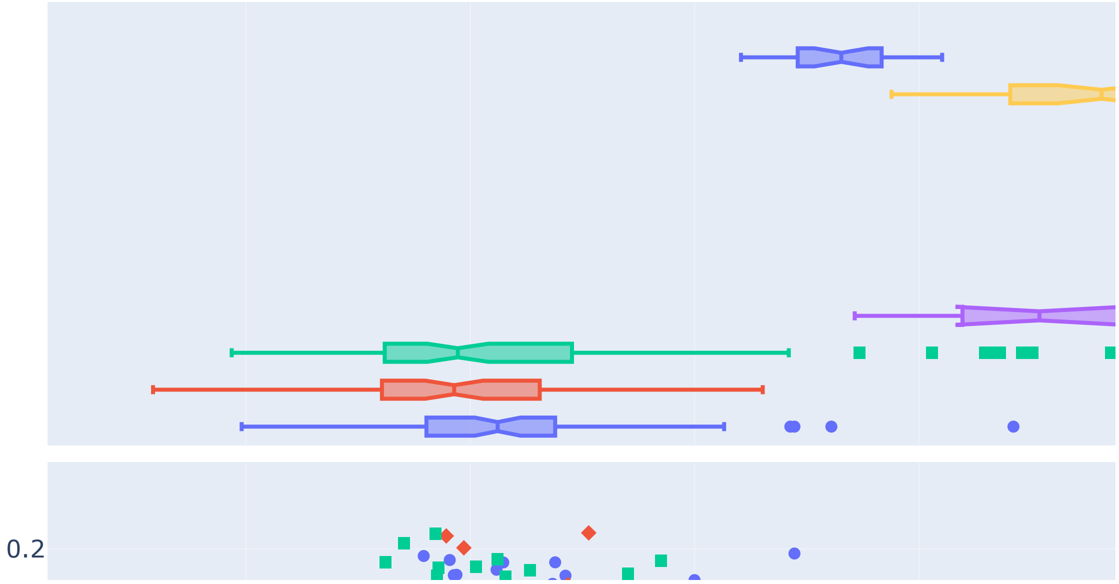


In [12]:

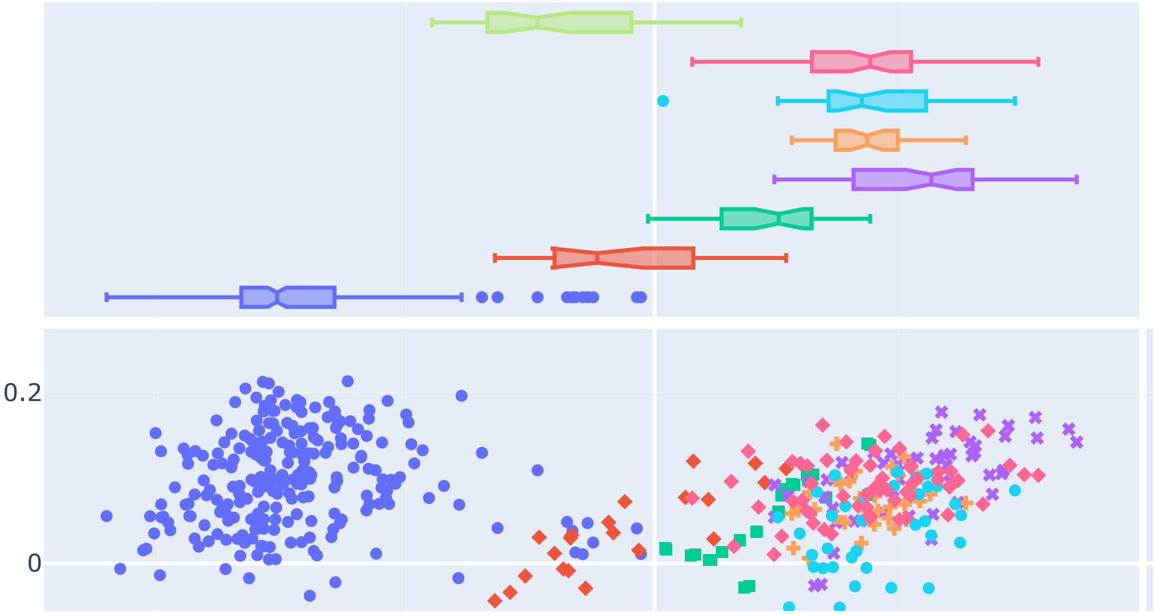
```
try:
    DOC = DOC.join(DCM)
except:
    pass
```

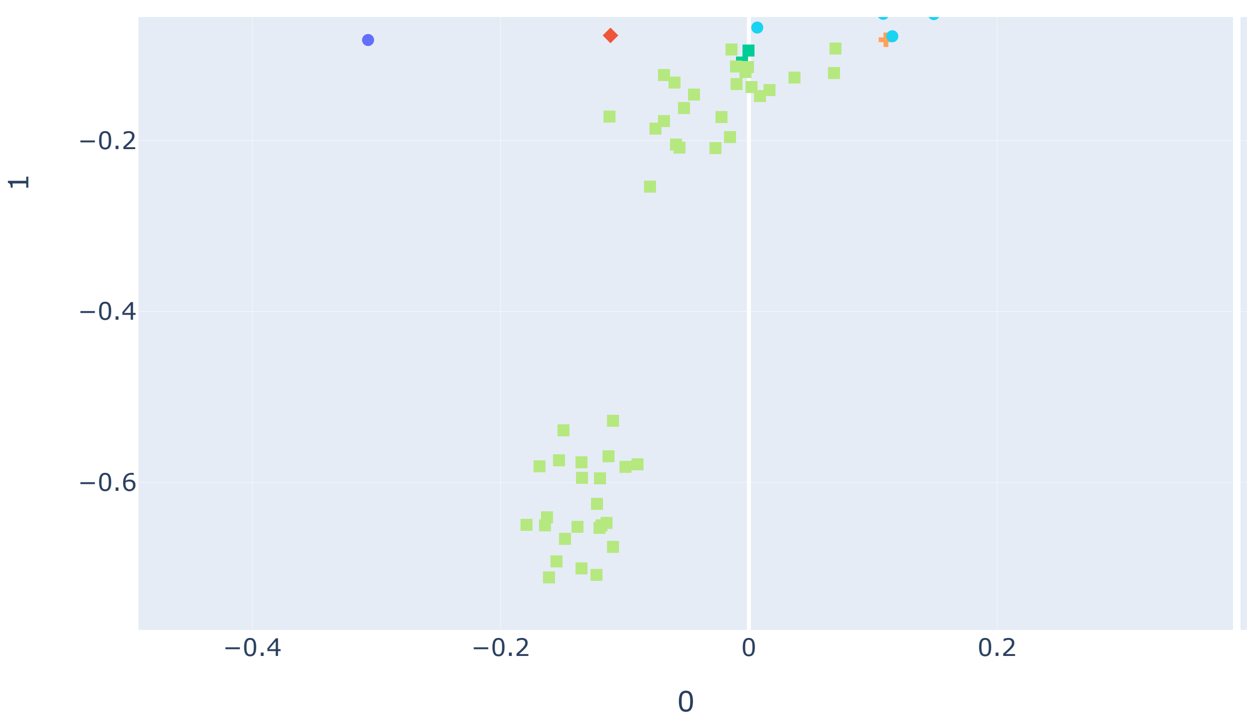
In [13]:

```
px.scatter(DOC, 0, 1,
           color='book_title', symbol='book_title',
           hover_name='book_title',
           marginal_x='box', marginal_y='box',
           height=1000)
```



```
In [18]: px.scatter(DOC, 0, 1,
                    color='author', symbol='author',
                    hover_name='author',
                    marginal_x='box', marginal_y='box',
                    height=750, width = 1000)
```

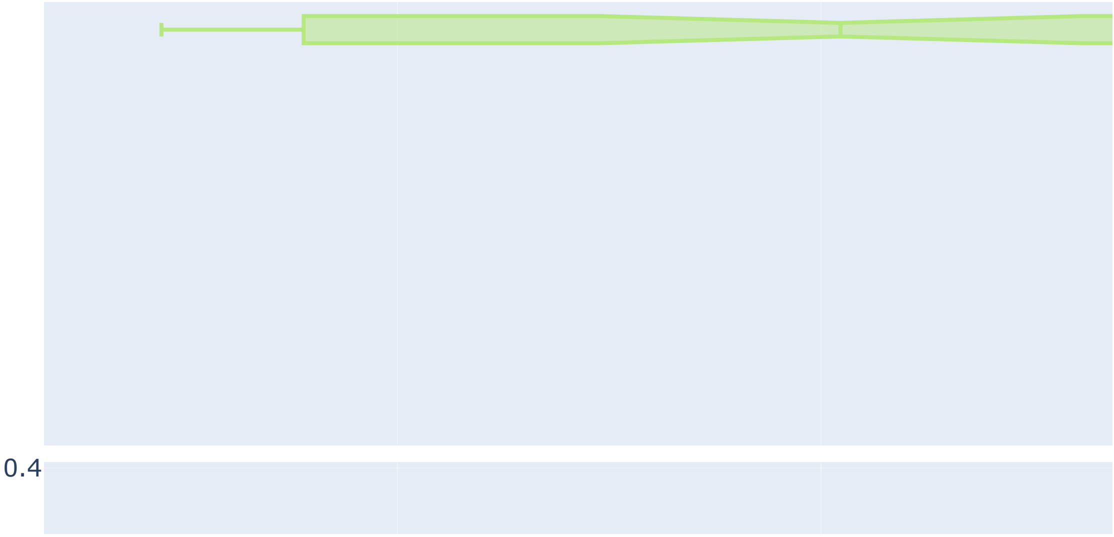


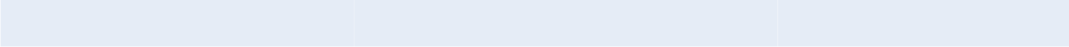


In [34]:

```
px.scatter(DOC, 1, 2,
           color='book_title', symbol='book_title',
           hover_name='book_title',
           marginal_x='box', marginal_y='box',
           height=1000)
```

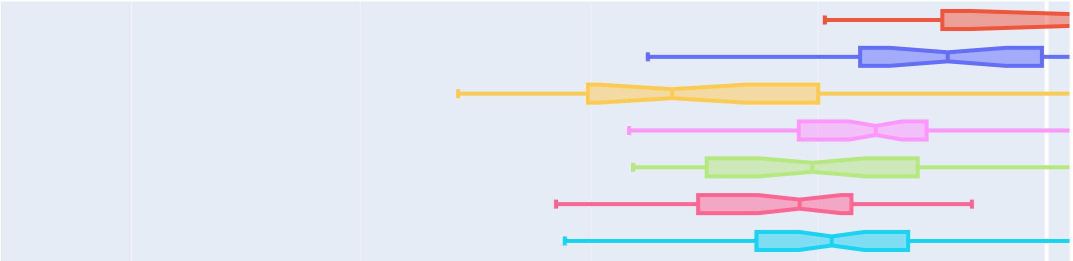
```
In [35]: px.scatter(DOC, 1, 2,
                    color='author', symbol='author',
                    hover_name='author',
                    marginal_x='box', marginal_y='box',
                    height=1000)
```

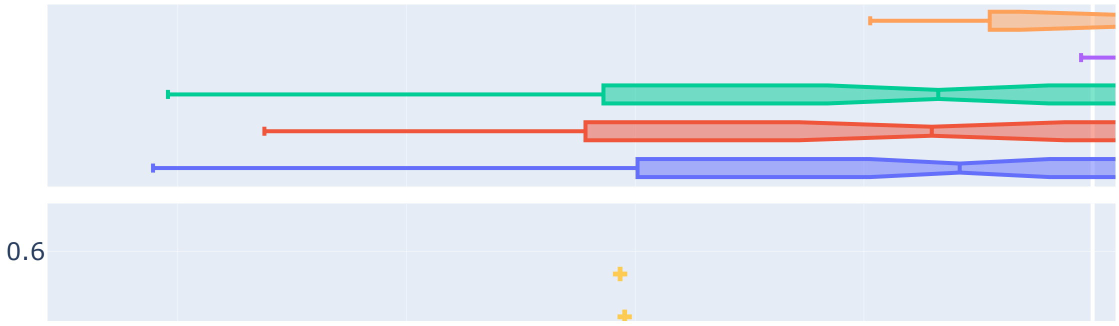




In [36]:

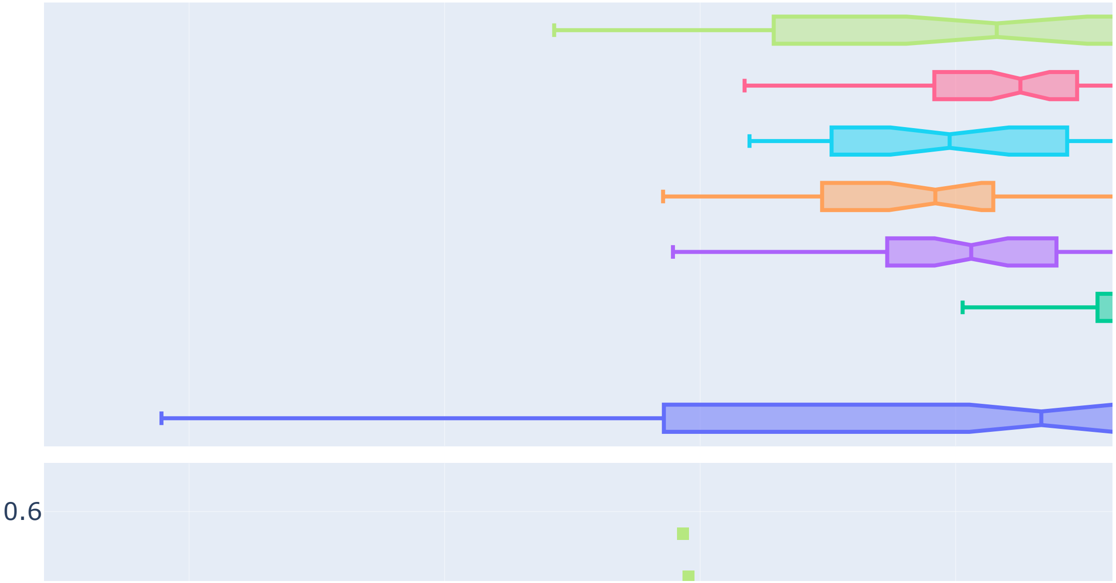
```
px.scatter(DOC, 2, 3,
           color='book_title', symbol='book_title',
           hover_name='book_title',
           marginal_x='box', marginal_y='box',
           height=1000)
```





In [37]:

```
px.scatter(DOC, 2, 3,
           color='author', symbol='author',
           hover_name='author',
           marginal_x='box', marginal_y='box',
           height=1000)
```


0.6

In [38]:

```
LOADINGS.to_csv("data/LOADINGS.csv")
DCM.to_csv("data/DCM.csv")
COMPINF.to_csv("data/COMPINF.csv")
```

In [ ]:

```
```