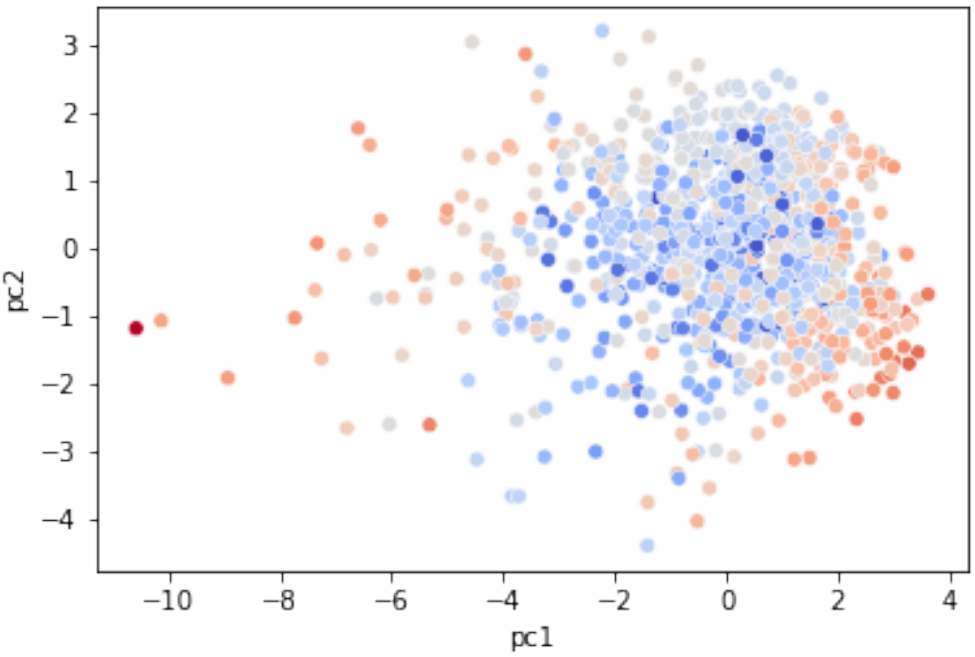
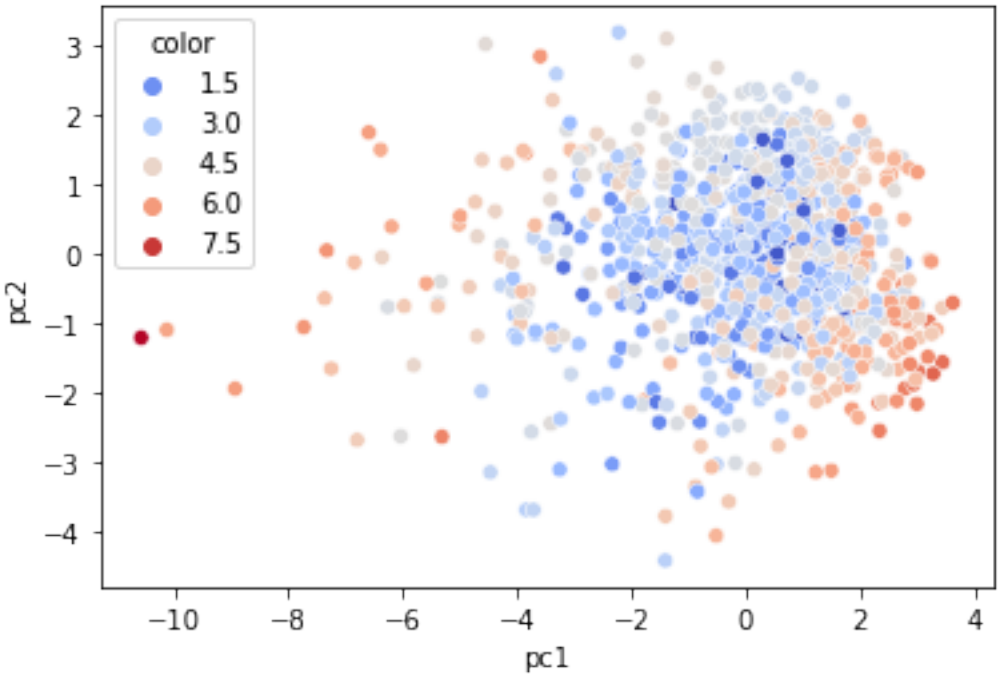
0.1 Question 2d

Create a 2D scatterplot of the first two principal components of `mid1_grades_centered_scaled`. Use `colorize_midterm_data` to add a `color` column to `mid1_1st_2_pcs`. Your code will be very similar to the code from problems 2a and 2b. Your result should look like this




```
In [35]: um, sm, vtm = np.linalg.svd(mid1_grades_centered_scaled , full_matrices = False)
         mid1_1st_2_pcs = np.matmul(mid1_grades_centered_scaled, vtm[0:2, :].T)
         mid1_1st_2_pcs = mid1_1st_2_pcs.rename(columns = {0 : 'pc1', 1: 'pc2'})
         sns.scatterplot(data = colorize_midterm_data(mid1_1st_2_pcs), x = "pc1", y = "pc2", hue = "colo
```

```
Out[35]: <AxesSubplot:xlabel='pc1', ylabel='pc2'>
```


color
1.5
3.0
4.5
6.0
7.5
pc2
pc1
3
2
1
0
−1
−2
−3
−4
−10
−8
−6
−4
−2
0
2
4

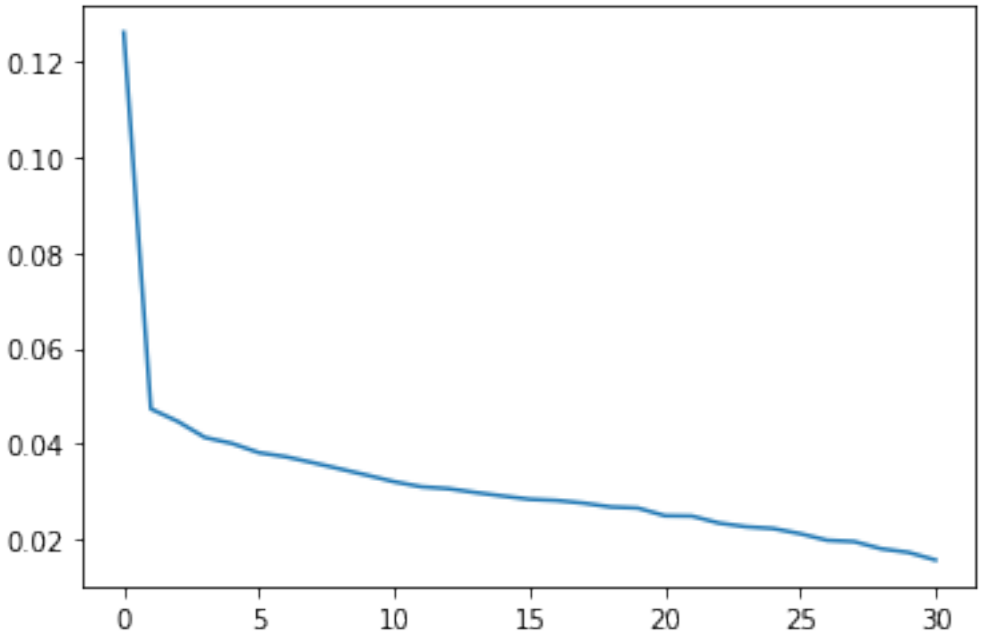
0.2 Question 2e

If you compute the fraction of the variance captured by this 2D scatter plot, you'll see it's only 17%, roughly 12% by the 1st PC, and roughly 5% by the 2nd PC. **In the cell below, create a scree plot showing the fraction of the variance explained by each principal componant using the data from 2d.**

Informally, we can say that our midterm scores matrix has a high rank. More formally, we can say that 2 principal components only capture a small fraction of the variance, and thus the data are not particularly amenable to 2D PCA scatterplotting.

```
In [36]: plt.plot(sm**2 / sum(sm**2))
```

```
Out[36]: [<matplotlib.lines.Line2D at 0x7f1d82a6c730>]
```



0.3 Question 3a

What does each row in `df_clean` represent?

Each row represents a different U.S. state and how it voted during the presidential election year specified by the column. R represents Republican, while D represents Democratic.

Unfortunately, we have two problems:

1. There is a lot of overplotting, with only 28 distinct dots. This means that at least some states voted exactly alike in these elections.
2. We don't know which state is which because the points are unlabeled.

Let's start by addressing problem 1.

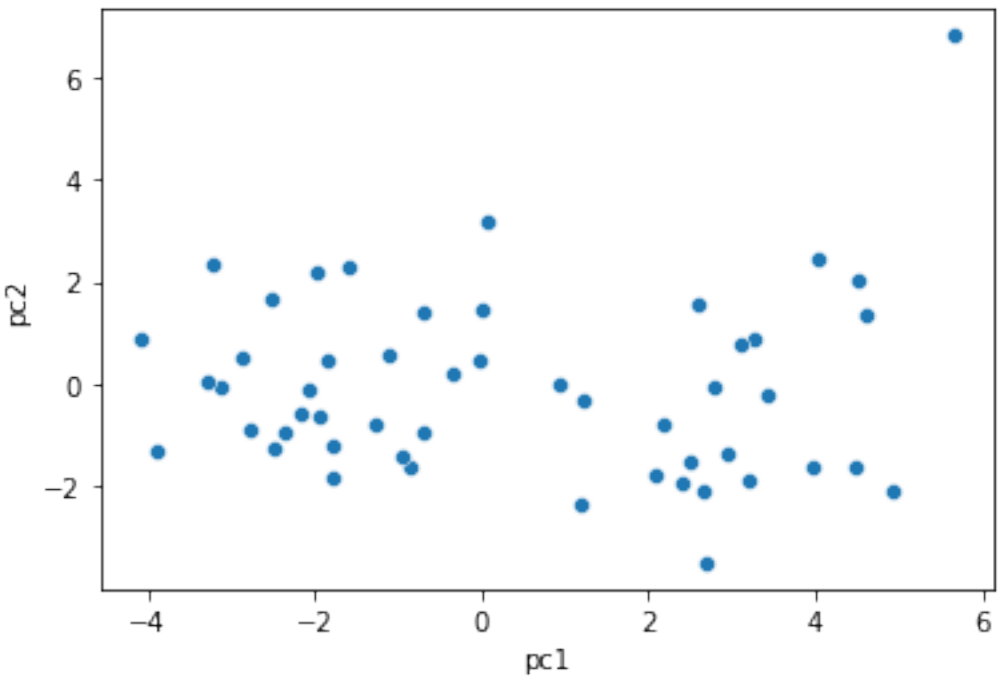
In the cell below, create a new dataframe `first_2_pcs_jittered` with a small amount of random noise added to each principal component. In this same cell, create a scatterplot.

The amount of noise you add should not significantly affect the appearance of the plot, it should simply serve to separate overlapping observations. Don't get caught up on the exact details of your noise generation, it's fine as long as your plot looks roughly the same as the original scatterplot.

Hint: See the pairplot from the intro to question 2 for an example of how to introduce noise.

```
In [46]: first_2_pcs_jittered=first_2_pcs.copy()
         noise = np.random.normal(0, 1, first_2_pcs_jittered.loc[:,'pc1':'pc2'].shape)
         first_2_pcs_jittered.loc[:,'pc1':'pc2'] += noise
         sns.scatterplot(data = first_2_pcs_jittered, x = "pc1", y = "pc2")
```

```
Out[46]: <AxesSubplot:xlabel='pc1', ylabel='pc2'>
```



Give an example of a cluster of states that vote a similar way. Does the composition of this cluster surprise you? If you're not familiar with U.S. politics, it's fine to just say 'No, I'm not surprised because I don't know anything about U.S. politics.'.

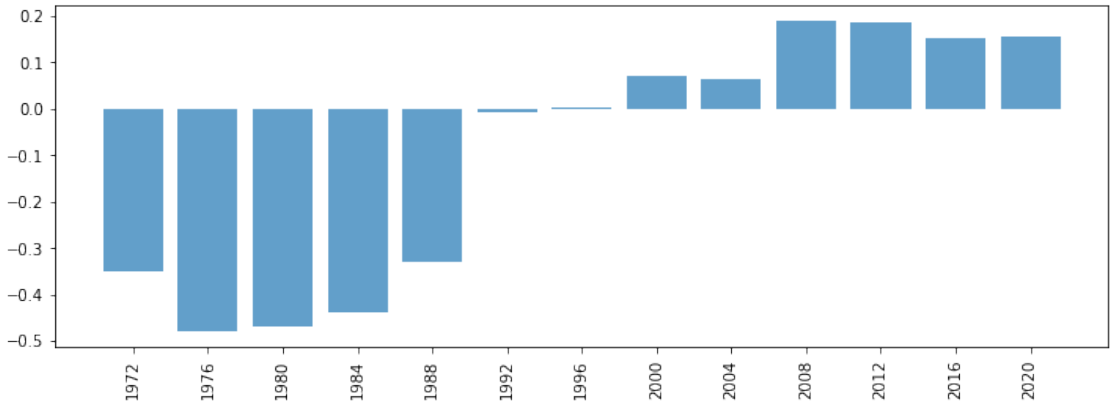
New York, Conneticut, and Pennsylvania voted in similar ways which was not that surprising since these are all large states that are in similar areas and tend to have similar demographics.

In the cell below, write down anything interesting that you observe by looking at this plot. You will get credit for this as long as you write something reasonable that you can take away from the plot.

It's interesting how Maryland and Florida voted similarly, since they are in such different regions and in my mind, they would vote for different parties–but then again, I don't know too much about U.S. politics. D.C. had the highest pc1 and pc2, which made it unique from the other states. This could be due to D.C. being the capital.

In the cell below, plot the the 2nd row of V^T.

```
In [49]: plt.figure(figsize=(12, 4))
         plot_pc(list(df_standardized.columns), vt_q3, 1);
```



0.4 Question 3h

Using your plots from question 3g as well as the original table, give a description of what it means to have a relatively large positive value for `pc1` (right side of the 2D scatter plot), and what it means to have a relatively large positive value for `pc2` (top side of the 2D scatter plot).

In other words, what is generally true about a state with relatively large positive value for `pc1`? For a large positive value for `pc2`?

Note: `pc2` is pretty hard to interpret, and the staff doesn't really have a consensus on what it means either. We'll be nice when grading.

Note: Principal components beyond the first are often hard to interpret (but not always; see question 1 earlier in this homework).

For pc1, large values indicate more democratic voting while smaller values indicate more republican voting. For pc2, large values could indicate the spread of the votes; maybe as more technology was introduced there was more variation due to differing opinions.

0.5 Question 3i

To get a better sense of whether our 2D scatterplot captures the whole story, create a scree plot for this data. On the y-axis plot the fraction of the total variance captured by the ith principal component. You should see that the first two principal components capture much more of the variance than we were able to capture when using the Data 100 Midterm 1 data. It is partially for this reason that the 2D scatter plot was so much more useful for this dataset.

Hint: Your code will be very similar to the scree plot from problem 1d. Be sure to label your axes appropriately!

```
In [51]: plt.plot(s_q3**2 / sum(s_q3**2))
```

```
Out[51]: [<matplotlib.lines.Line2D at 0x7f1d94acd880>]
```