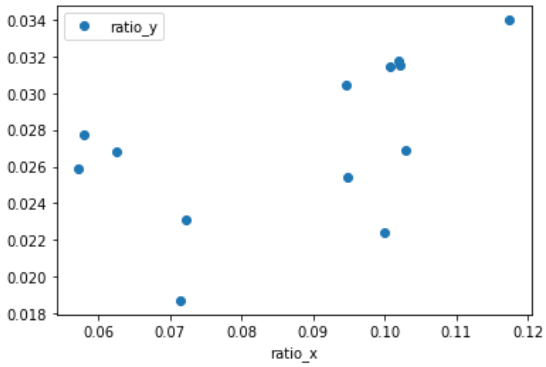
Exercise 1

4.
The optimal number of topics when using the lemmatizer is 4 whereas using the porter stemmer is 6.

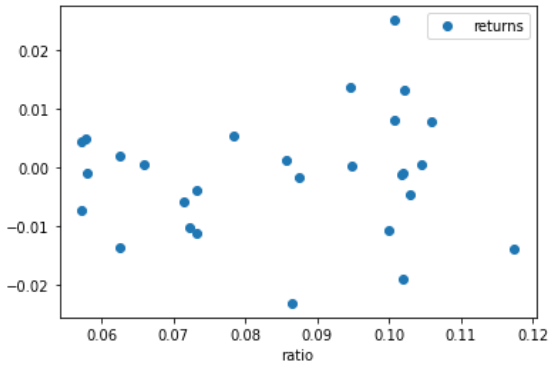
The topics generated from the lemmatizer are as follows: ['Pandemic', 'Politics', '2016 Election', 'Vaccination'].

The topics generated from the porter stemmer are as follows: ['Presidential Election', 'Republican Trump', 'Protest', 'Democratic Biden', 'Vaccination and Trump's impeachment'].

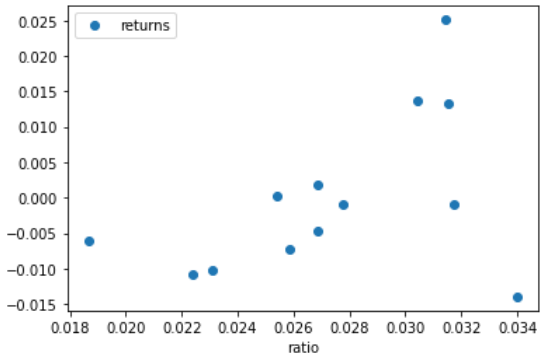
8.



On the x-axis, we observe the covid uncertainty index whereas, on the y-axis, we observe the coarse economic policy uncertainty index. We observe a positive linear relationship between these two indices thus, as the covid uncertainty index increases, we would expect the coarse economic policy uncertainty index to increase.



On the x-axis, we observe the covid uncertainty index whereas, on the y-axis, we observe the daily SNP500 returns. We do not observe any particular trend between the index and the returns since we observe a random scatter in the plot. Hence, there is no relationship between the covid uncertainty index and the daily SNP500 returns.



On the x-axis, we observe the coarse economic policy uncertainty index whereas, on the y-axis, we observe the daily SNP500 returns. We observe a positive relationship between these two indices thus, as the coarse economic policy uncertainty index, we would expect the daily SNP500 to increase.

10.

From the study, we determined that the topics of the headlines are mostly related to the pandemic as well as politics. Furthermore, we determined the relationship between the indices in question 8. Lastly, determining the sentiment index yielded the following results.

```
Aggregate Sentiment Index:
    polarity_neg  polarity_neu  polarity_pos  polarity_neg-pos
0          0.049         0.872         0.079             -0.03
```

Therefore, we can conclude that most headlines are neutral in terms of their sentiment since the neutral polarity yielded the highest value at 0.872.

Exercise 2

1.

$$1)\ E\left[(y_0-\hat{f}(x_0))^2\right]$$

$$= E\left[(f(x_0)+\epsilon-\hat{f}(x_0))^2\right]$$

$$= E\left[(f(x_0)+\epsilon)^2+\hat{f}(x_0)^2-2(f(x_0)+\epsilon)(\hat{f}(x_0))\right]$$

$$= E\left[f(x_0)^2+\epsilon^2+2f(x_0)\epsilon+\hat{f}(x_0)^2-2f(x_0)\hat{f}(x_0)-2\hat{f}(x_0)\epsilon\right]$$

$$= E\left[f(x_0)^2+\epsilon^2+0+\hat{f}(x_0)^2-2f(x_0)\hat{f}(x_0)-0\right]$$

$$= E\left[f(x_0)^2+\hat{f}(x_0)^2-2(f(x_0)\hat{f}(x_0)+\epsilon^2\right]$$

$$= E\left[(f(x_0)-\hat{f}(x_0))^2\right]+E(\epsilon^2)$$

$$= E\left[(f(x_0)-\hat{f}(x_0))^2\right]+Var(\epsilon)-(E(\epsilon))^2$$

$$= E\left[(f(x_0)-\hat{f}(x_0))^2\right]+Var(\epsilon)$$

2.

2) $E\left[\left(f(x_0) - \hat{f}(x_0)\right)^2\right] + \mathrm{Var}(\epsilon)$

$= E\left[\left(f(x_0)^2 + \hat{f}(x_0)^2 - 2f(x_0)\hat{f}(x_0)\right)\right] + \mathrm{Var}(\epsilon)$

$= f(x_0)^2 + E\left[\hat{f}(x_0)^2\right) - 2f(x_0)E\left(\hat{f}(x_0)\right) + \mathrm{Var}(\epsilon)$ # by linearity of expectation

$= f(x_0)^2 + E\left(\hat{f}(x_0)\right)^2 + \mathrm{Var}\left[\hat{f}(x_0)\right] - 2f(x_0)E\left(\hat{f}(x_0)\right) + \mathrm{Var}(\epsilon)$

by the formula for variance

$= f(x_0)^2 + E\left(\hat{f}(x_0)\right)^2 - 2f(x_0)E\left(\hat{f}(x_0)\right) + \mathrm{Var}\left[\hat{f}(x_0)\right] + \mathrm{Var}(\epsilon)$

$= \left(f(x_0) - E\left[\hat{f}(x_0)\right)\right)^2 + \mathrm{Var}\left[\hat{f}(x_0)\right] + \mathrm{Var}(\epsilon)$

$= \mathrm{Var}\left[\hat{f}(x_0)\right] + \left(f(x_0) - E\left[\hat{f}(x_0)\right)\right)^2 + \mathrm{Var}(\epsilon)$

$= E\left[\left(E\hat{f}(x_0) - \hat{f}(x_0)\right)^2\right] + \left(f(x_0) - E\hat{f}(x_0)\right)^2 + \sigma^2$

□

3.
a, d, g

4.
(a) In machine learning, bagging is where we take a single dataset D with n data points and generate m new datasets, by sampling n training examples from D, with replacement. We then finally average the predictions of models trained on each of these datasets. Random forests are decision trees that employ the bagging algorithm. Therefore, since random forests average over independent samples, we will expect the variance to be reduced. Furthermore, we will expect the bias to remain unchanged since the averaged prediction has the same expectation. Therefore, Random Forests have lower variance and the same bias as Decision Trees.

(b) Bias represents how wrong the expected prediction is whereas variance represents the amount of variability in the predictions. Hence, an algorithm with low bias and high variance indicates that the algorithm learns all the relevant structures as well as fits the quirks of the data we happened to sample. Therefore, we can conclude that our friend is correct about bias but incorrect about variance. By claiming that the predictions “are quite different for different datapoints”, the friend does not specify that the variance “represents the amount of variability in the predictions”.

Exercise 3
3.1

Q1:

2. 1 variable is required to perform clustering since we have to put an observation into at least one of the clusters based on the variable.

Q2:

2. Clustering. Based on the clicking history of the user, we can form a user-item matrix where it determines the relationship between the user and the item. Thus, we can perform clustering where we cluster the user's tastes and preferences accordingly for the website to provide outfit recommendations.

Q3:

2. 6 is the best choice for the number of clusters since by using the elbow method, we observe a kink at 6 clusters.

Q4:
2. and 3. are wrong statements.

2. k-nearest neighbour is a supervised learning algorithm used for classification and regression whereas k-means is an unsupervised learning algorithm used for clustering

3. After each run of K-means clustering, we would expect the variation of the observations within a cluster to decrease hence, we would reassign observations to their corresponding cluster to achieve a lower objective function. Hence, after every run, the objective function converses on a local minimum which will have different clustering results.

3.2
In a linear regression model, we wish to obtain the value of a target variable from the data i.e. obtaining y from y = ax + b. If we were to incorporate clustering into a linear regression model, we can add indicator functions into our linear regression model of clusters with specific characteristics. For example, if the linear regression model is described as follows: GDP = aStock_Price + c, we can add an indicator variable where it indicates whether if the country has high/low poverty rates through clusters. Thus, our model would be GDP = aStock_Price + bI(poverty_rates) + c which would improve the fit and accuracy of our model.