Two csv files are attached: trends1.csv and predictions.csv

These have information for 3 types of tops.
In trends1.csv, each top has a score for popularity of those types of items, for the date.

predictions.csv has a fit that has then been extrapolated to the next 365 days.

Please complete the following tasks:

1) Assess the three trends in trends1.csv, which is performing best. Elaborate on how you defined "best performance".
   The Top 1, 2 and 3 data are plotted as shown in Q1.ipynb, as well as their trendlines.

   In my opinion, a trend is said to have the best performance if the spread of data around its line of best fit is the smallest compared to others. The performance of the model can be measured by calculating R^2 which is also known as coefficient of determination. R^2 is a statistical measure of how close the data are to the fitted regression line. In general, the higher the value of R^2, the better the model fits the given data.

   Calculating R^2 for all three trends using SciPy, we can see that Top 3 has the greatest value of R^2 compared to others. This implies implying that Top 3 trend performs the best.

2) Quantify the performance over the last year - to establish what proportion of demand has changed.
   In order to investigate the change in proportion of demand, root mean squared error(RMSE) between every year with 2017 are calculated. RMS tells how concentrated the data is around its line of best fit. In this case, we used RMS to tells how far the data for the corresponding year spread from data for 2017. The smaller the RMS, the smaller the deviation. The deviation can be seen as the change in proportion of demand between the year with 2017. This analysis is shown in Q2.ipynb.

   Over the last year, the change in demand for Top 2 data has been the greatest compared to that of Top 1 and 3. This can be confirmed as the RMS for Top 2 data is the greatest compared to others at every year. This is probably due to the a better branding and packaging. A good brand name and logo can both impact the popularity score. Same thing applies for the item packaging as a good item packaging indicates a good representation of selling item towards the customer. Besides that, a higher popularity score of items for Top 2 compared to others can also because of a better product placement. For example, a clear product placement such as on an end cap in a highly trafficked area of the store, or on the front page or in the side bar of an ecommerce business like Amazon, eBay can boost the popularity score. The fluctuation of the popularity score of the items over time for all datasets given are probably because of factors like pricing, reputation and availability of the items.

   For the case of pricing, there would probably be sales at certain times where the item price would be reduced, resulting in a rising in popularity score. The popularity score would then be dropped after the sale ends. Besides that, the reputation factor can also be considered in this case as a better advertisement of the item can bring a greater awareness to the customers regarding the item and thus raising the popularity score. A lesser advertisement induces a drop in the popularity score. Same thing applies for the availability of the item. If the item is out of stock, the popularity score would definitely be decreases until it is available again.

3) Look at the predictions.csv file - there are three fits for each trend, comment on which you find to be best.
In predictions.csv, there are three given prediction fit for every top data. The metrics between every three given prediction fits with trend are calculated for every three tops data and then displayed as a dataframe in Q3.ipynb. The analysis is done with a number of datasets used of 261.

Note that the fit is said to have the best performance if it has the smallest value of metrics compared to others. Therefore, the best fits are found to be Fit 1 for all three tops data. This is confirmed by comparing the coefficient of determination, R^2 of all three given prediction fits for every top data.

4) Quantify these predictions into an assessment of how you would expect the demand to change over the next 3, 6, 12 months.
The predictions are evaluated using the analysis of metrics in last question. By studying the distribution of all three tops data shown in Q1.ipynb, we found that both Top 2 and 3 show a clear seasonal trend where the popularity score gradient increases in the first 6 months of every year(Jan-Jun) and then decreases for the remaining months of the year.

For Top 1, there is no clear trend found but only an increase in popularity score gradient in 2015 for 12 months, after 2 years of having constant values of popularity score. If there is any trend exist for Top 1, we could probably say that this trend is not obvious and the demand in 2018 would probably increases again for 12 months, since there is no increase in popularity score for 2 years after 2015.

However, do keep in mind that if we take a consideration of external factor, it is not impossible for the demand to have a sudden change either gradually or rapidly, depending on the factor. The example of external factor would be such as the change in style or fashion trend, the strength of currency, and many more. The fluctuation of the popularity score would always occurs due to the mentioned factor such as pricing, reputation and availability.

5) Please create your own predictions based of the three types of tops.
Refer to Q5.ipynb.

6) Briefly explain your choice of model, and how it performs better or worse than the provided predictions.
The preferred model in this case is linear regression. This is because the variables are all numerical, instead of categorical and the data provided are al free of missing values and outliers. It is also because of the fact that the residuals of all three tops data shown in Q1.ipynb are all normally distributed. This model can be made easily using sci-kit learn.

In Q5.ipynb, the prediction is made using linear regression functions provided by scikit-learn package. The datasets was first divided into two parts; training and test datasets, with the training and test datasets being 4/5 and 1/5 of the number of actual dataset, respectively. For Top 1, the prediction performs better as the number of test dataset increases. The opposite thing occurred for Top 2 and 3.

It is noticed that all the Root Mean Squared Error calculated are all greater than ~10% of the Mean Absolute Error. This tells us that our algorithm was not very accurate but can still make reasonably good predictions.

We can also forecast the sales by creating a seasonal ARIMA model using Statsmodels package. However, forecasting sales using this way are much more complicated than using sci-kit learn package. I'm currently learning this method.

End.