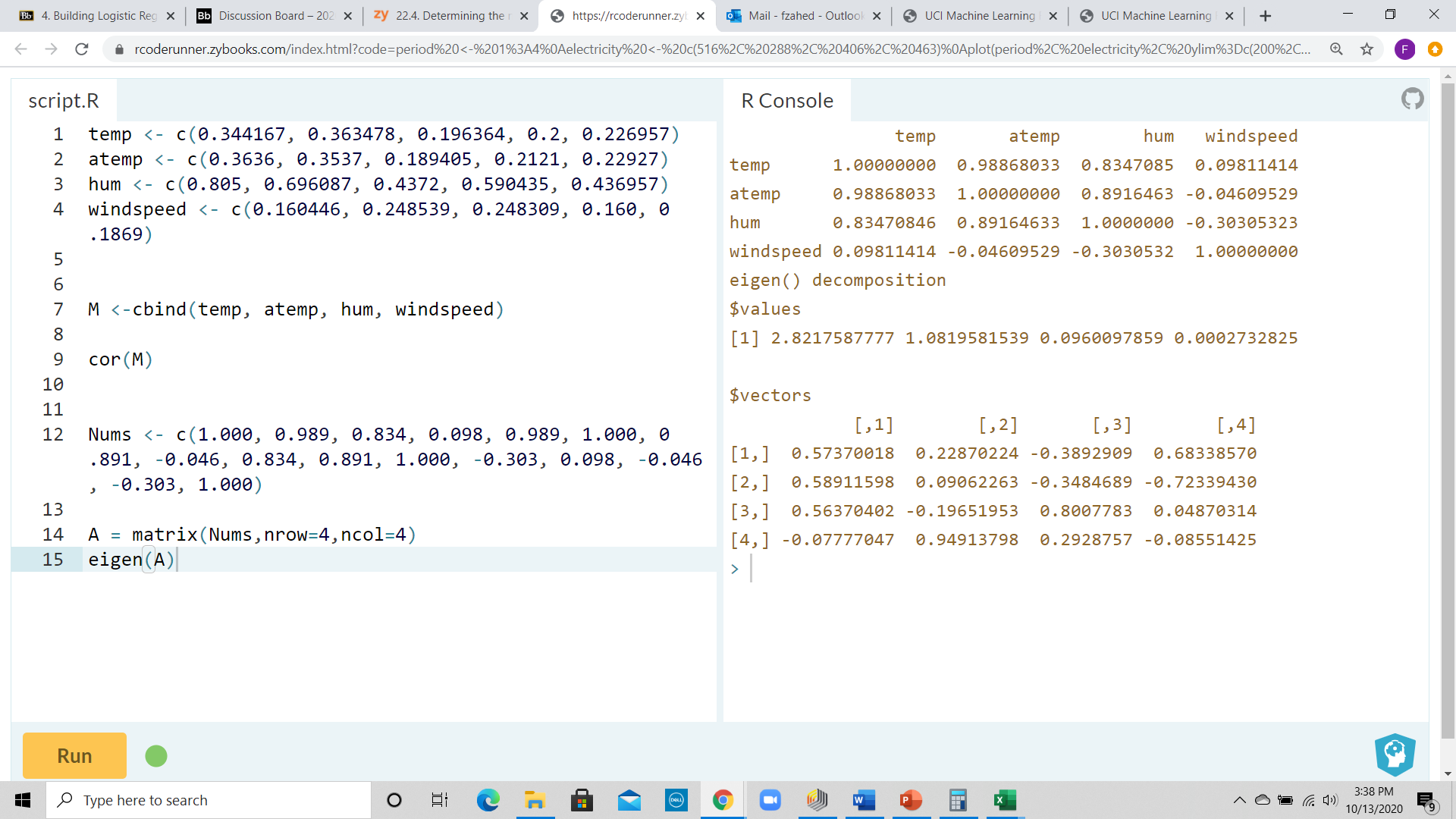
1. When a statistical model has a lot of predictor variables, those variables can be transformed into different components based on their interdependence. In other words, a component is formed by transforming several predictor variables that are weighted according to the correlations between them. For instance, there are many (predictor) variables affecting a country’s GDP such as the unemployment rate, government spending, inflation rate, private sector investments, interest rates, personal income levels, public consumption, to name a few. These variables can be transformed to components based on their interdependence with each other, with weights being assigned accordingly to each variable in the component.
2. The main premise of Principal Component Analysis (PCA) is to reduce the number of parameters of a data set without losing too much information, such that a few components can account for the maximum variability in the data. This is also known as dimensionality reduction in large data sets, done for ease of analysis.

The first step of performing a PCA is to standardize the variables in order to bring the variables to the same scale, if the data ranges vary considerably for these variables. Then the correlation coefficients of every variable pair need to be worked out and transformed into a matrix. Next, we calculate the eigenvalues and corresponding eigenvectors of the correlation matrix which gives us the components of each eigenvector. (In instances where the data ranges do not vary much, a covariance matrix can be used for this calculation). Each principal component is expressed as an equation containing a weighted value of all the original variables. The number of variables will determine how many eigenvalues we generate; so for *n* number of variables, we will have an *n* x *n* correlation/covariance matrix and *n* eigenvalues, with the first one being the largest. Hence, the first principal component with the largest eigenvalue accounts for the maximum variability in the data. Arranging the eigenvectors according to their eigenvalues in descending order will help us determine the principal components in order of significance, whereby we can eliminate the ones that only account for a small amount of variability. This can be done with the help of a scree plot which allows us to visually see where the bend of the curve happens in a scatterplot (plotting Principal Component vs eigenvalues), which indicates that the eigenvalues are leveling off and, based on the specific goal we pick, we can determine how many principal components to retain in our model.

1. I have the following dataset containing the number of bike rentals over a period of time (for ease of analysis, I only included data from 5 days and 4 standardized predictor variables – temperature in Celsius (temp), “Feels like” temperature in Celsius (atemp), humidity (hum), windspeed (windspeed); the last column, cnt, is the number of bike rentals for that day):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| temp | atemp | hum | windspeed | cnt |
| 0.344167 | 0.363625 | 0.805833 | 0.160446 | 985 |
| 0.363478 | 0.353739 | 0.696087 | 0.248539 | 801 |
| 0.196364 | 0.189405 | 0.437273 | 0.248309 | 1349 |
| 0.2 | 0.212122 | 0.590435 | 0.160296 | 1562 |
| 0.226957 | 0.22927 | 0.436957 | 0.1869 | 1600 |

After doing a PCA run in R, I obtained the following eigenvalues and eigenvectors after formulating the correlation matrix:



As we can see in the output above, the first eigenvalue 2.821 accounts for 70% of the variability in bike rentals for this data set. The first two eigenvalues together account for more than 95% of the variability in our data set, so it is safe to say that we can drop the third and the fourth principal components. The equations for these two Principal Components are:

PC1 = 0.5737 temp + 0.5891 atemp + 0.5637 hum – 0.0777 windspeed

PC2 = 0.2287 temp + 0.0906 atemp – 0.1965 hum + 0.9491 windspeed

Of course in a real world setting, a PCA run is more justifiable for a data set that contains several predictor variables as opposed to just the 4 that I used in my example.

\*Data set obtained from:

<https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset>