Guided Project Report

Exploratory Factor Analysis

Name: Shruti Verma
Course: AI and ML
(Batch 4)
Duration: 10 months

Problem Statement: Build a machine learning model using EFA for dimensionality reduction

Prerequisites

What things you need to install the software and how to install them:

Python 3.8 or higher versions This setup requires that your machine has latest version of python. The following url https://www.python.org/downloads/ can be referred to download python. Once you have python downloaded and installed, you will need to setup PATH variables (if you want to run python program directly, detail instructions are below in how to run software section). To do that check this: https://www.pythoncentral.io/add-python-to-path-python-is-not- recognized-as-an-internal-or-external- command/. Setting up PATH variable is optional as you can also run program without it and more instruction are given below on this topic.

Second and easier option is to download anaconda and use its anaconda prompt to run the commands. To install anaconda check this url https://www.anaconda.com/download/ You will also need to download and install below 3 packages after you install either python or anaconda from the steps above Sklearn (scikit-learn) numpy scipy if you have chosen to install python 3.8 then run below commands in command prompt/terminal to install these packages pip install -U scikit-learn pip install numpy pip install scipy if you have chosen to install anaconda then run below commands in anaconda prompt to install these packages conda install -c scikit-learn conda install -c anaconda numpy conda install -c anaconda scipy

Dataset used

The data source is airline satisfaction datasets(14 columns which contribute to passenger ratings) dataset provided in the kaggle https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction. The dataset is divided into 80 percent training data and 20 percent a test data. We focused mainly on 14 columns to understand what factors are highly

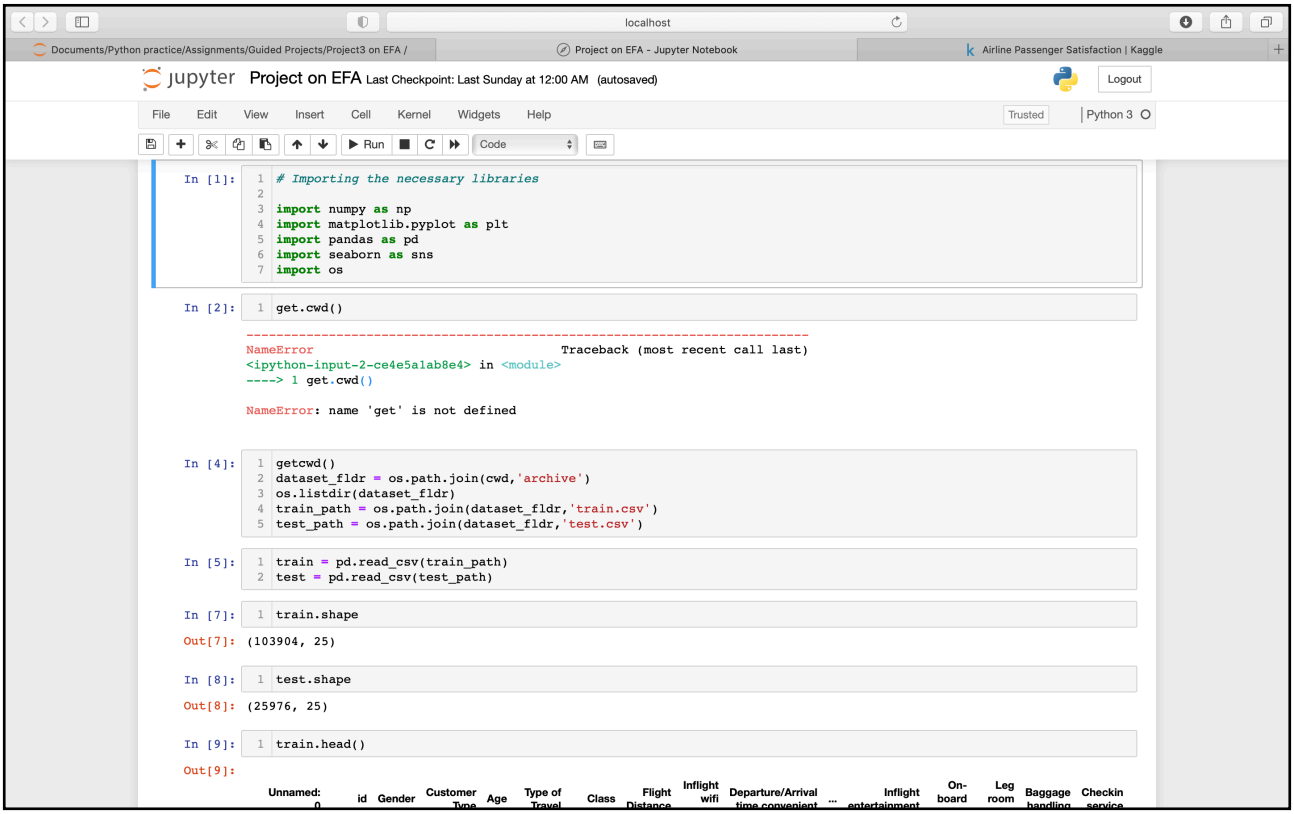
correlated to a satisfied or dissatisfied passenger. Train data is having more than one lakh records and test data is around 25 thousand records.

Method used for detection

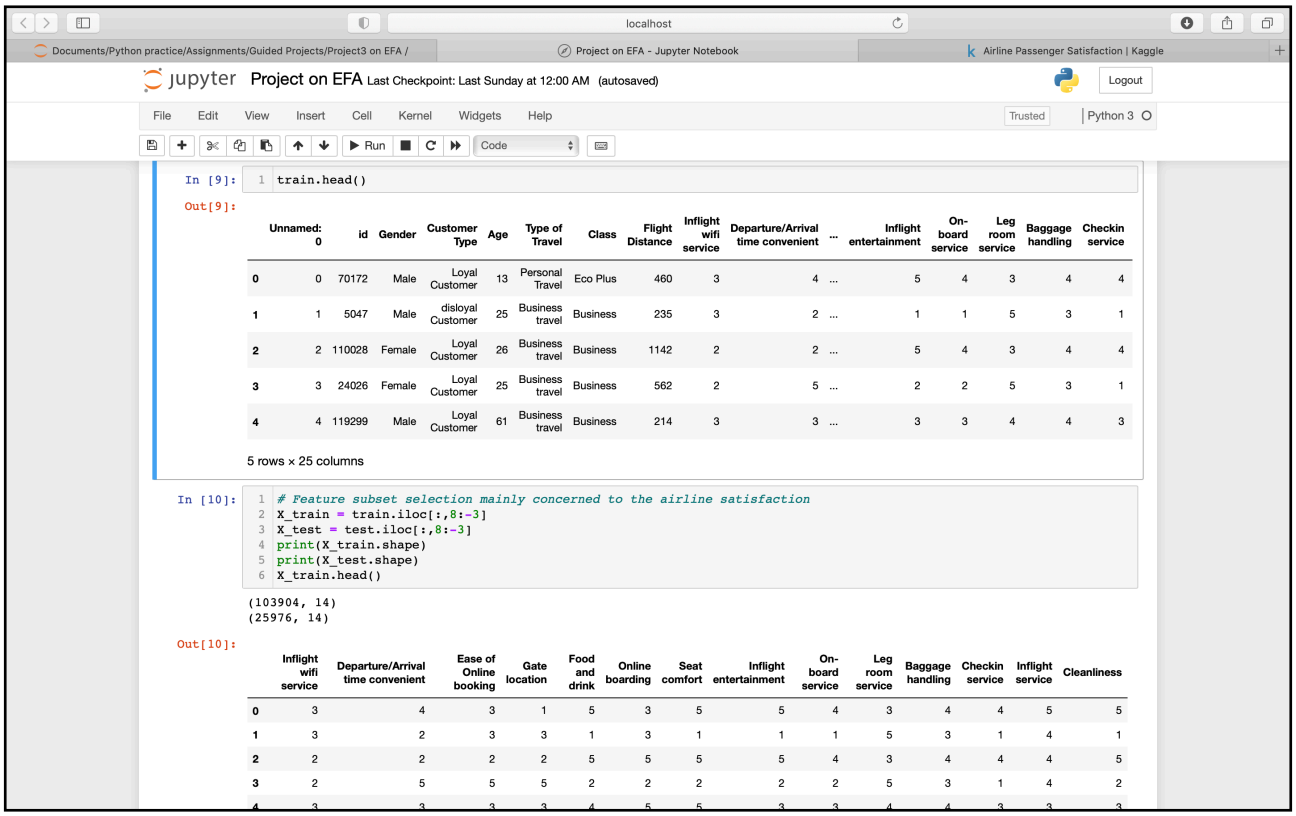
EFA

Feature subset dataset -> Zero centring dataset -> correlation among features -> based on variance and Eigen values -> dimensionality reduction transformation

Importing the libraries and capturing images:

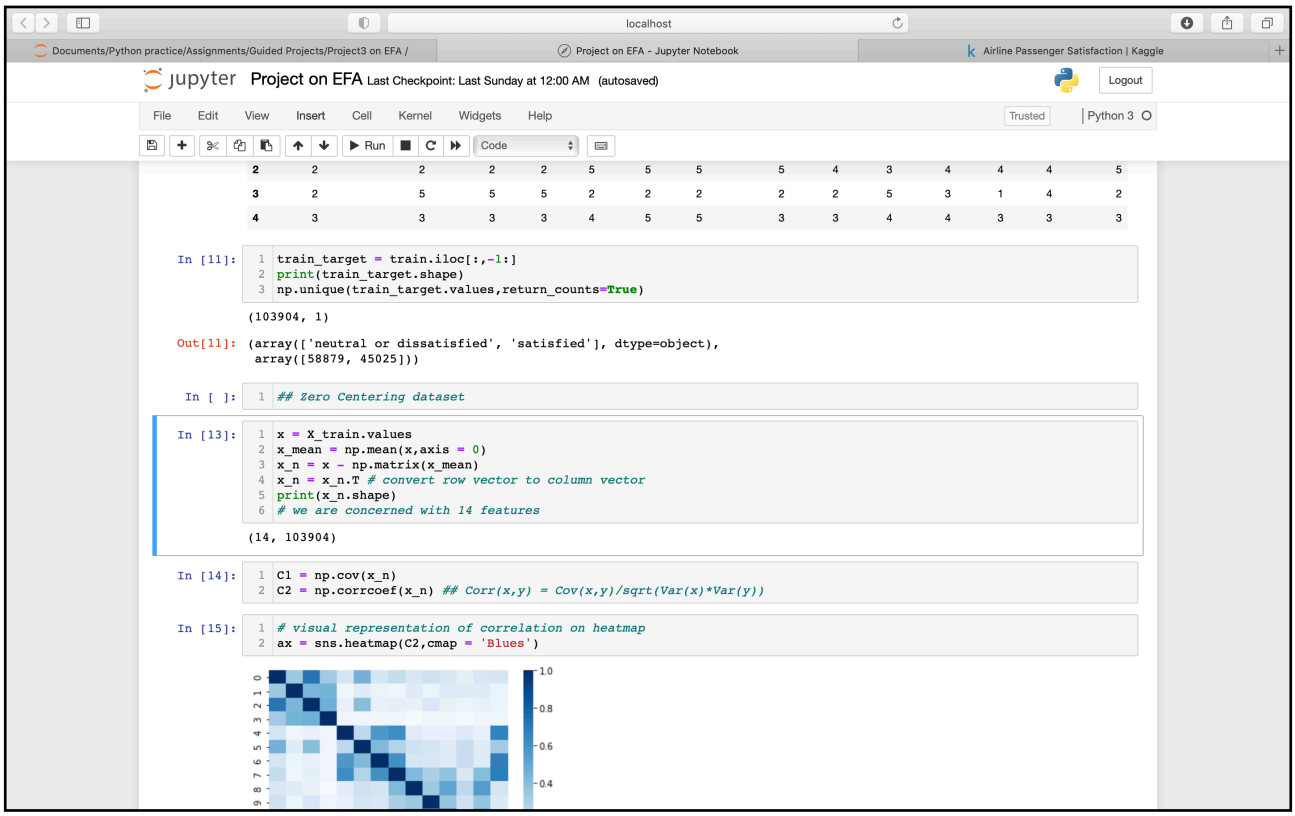


Importing the train and test data and feature subset

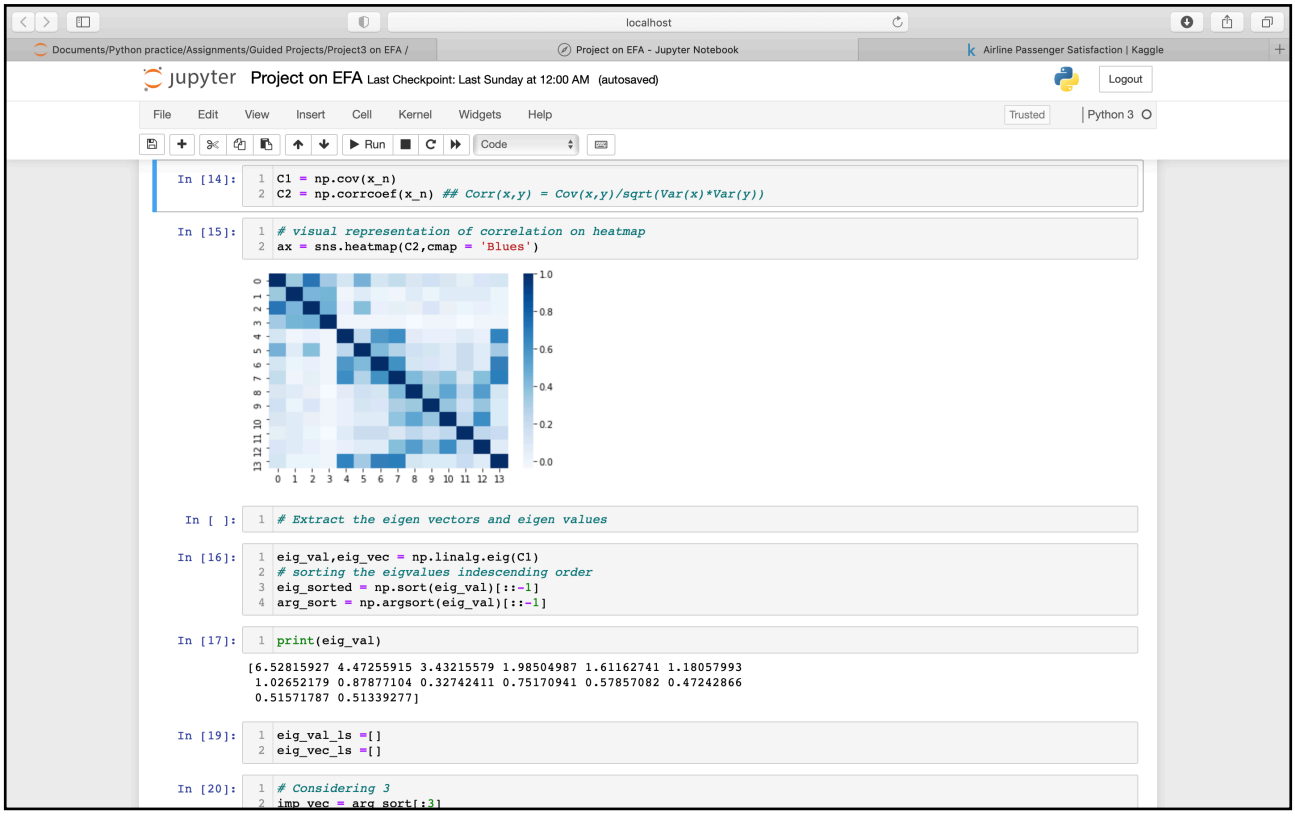


Splitting the training and testing data into 80 :20.

Zero Centring data

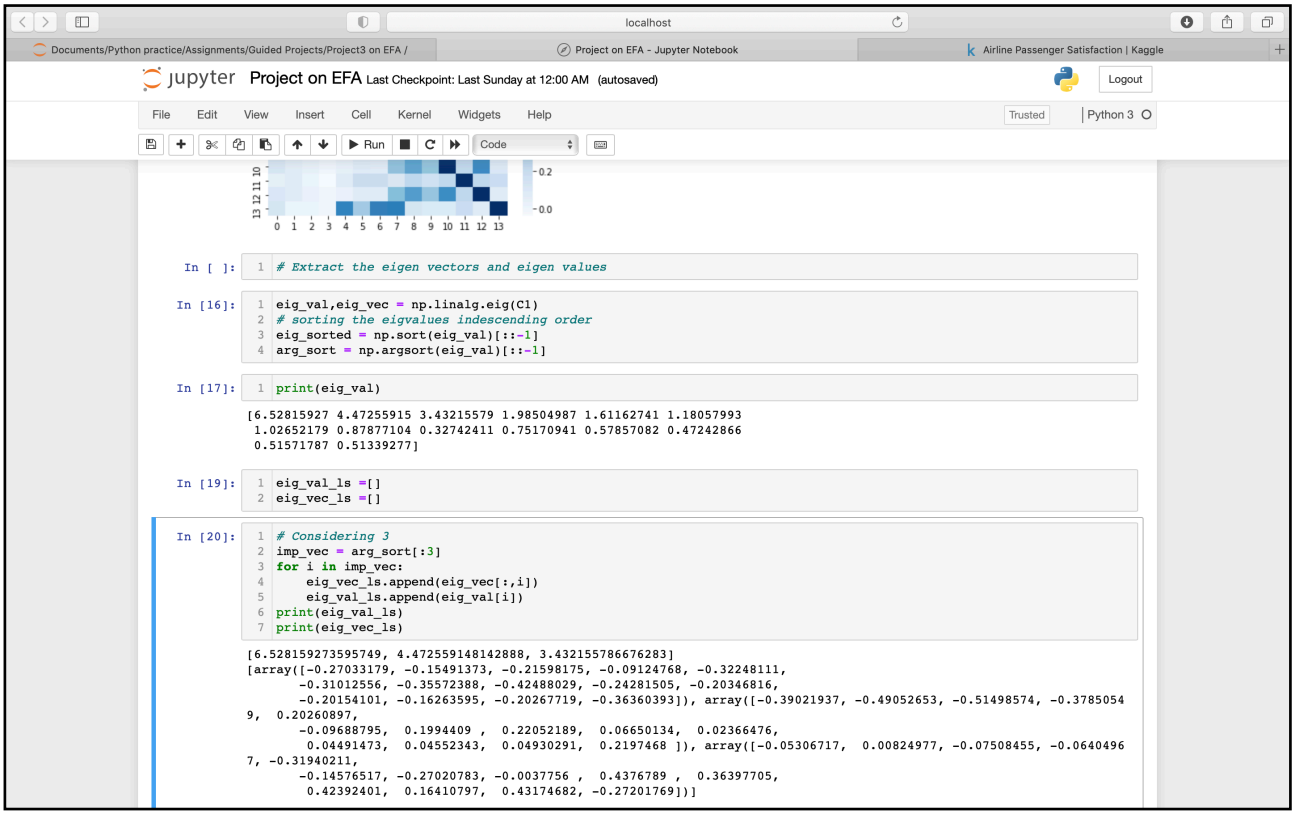


Visual Representation of correlation on heat map

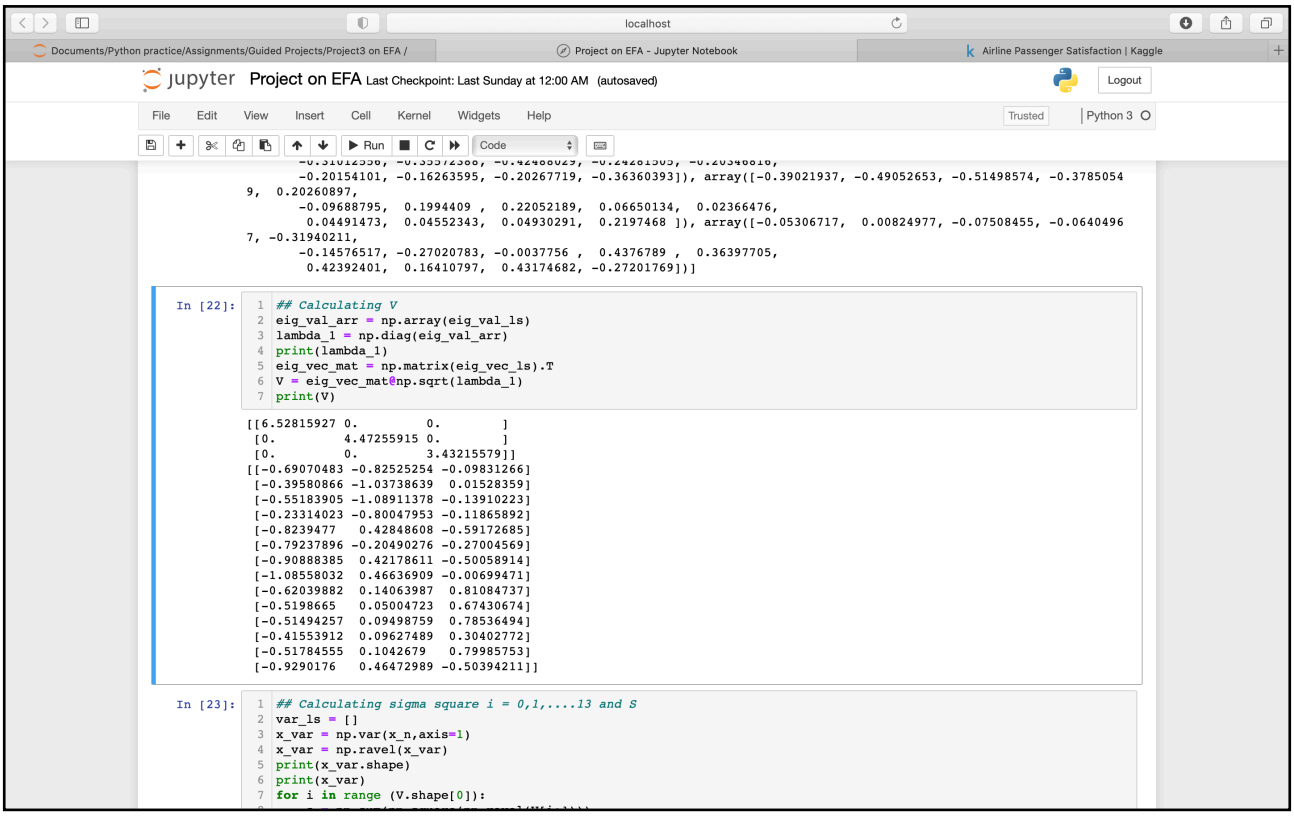


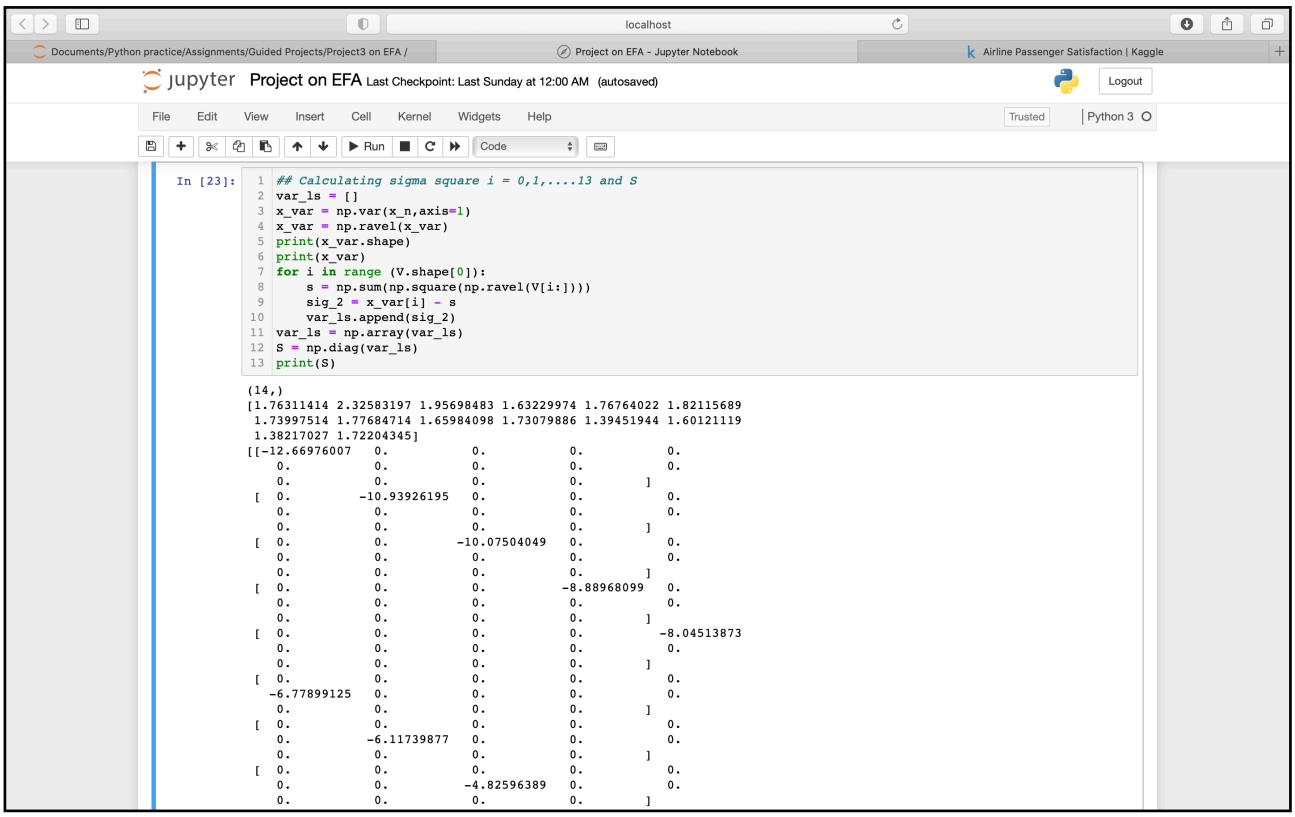
Considering 3 correlation arbitrarily

Extracting the Eigen vectors and values



Calculating Variance


localhost
Documents/Python practice/Assignments/Guided Projects/Project3 on EFA /
Project on EFA - Jupyter Notebook
Airline Passenger Satisfaction | Kaggle
jupyter
Project on EFA Last Checkpoint: Last Sunday at 12:00 AM (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
-0.20154101, -0.16263595, -0.20267719, -0.36360393]), array([-0.39021937, -0.49052653, -0.51498574, -0.3785054
9, 0.20260897,
-0.09688795, 0.1994409 , 0.22052189, 0.06650134, 0.02366476,
0.04491473, 0.04552343, 0.04930291, 0.2197468]), array([-0.05306717, 0.00824977, -0.07508455, -0.0640496
7, -0.31940211,
-0.14576517, -0.27020783, -0.0037756 , 0.4376789 , 0.36397705,
0.42392401, 0.16410797, 0.43174682, -0.27201769])]
In [22]:
Calculating V
eig_val_arr = np.array(eig_val_ls)
lambda_1 = np.diag(eig_val_arr)
print(lambda_1)
eig_vec_mat = np.matrix(eig_vec_ls).T
V = eig_vec_mat@np.sqrt(lambda_1)
print(V)
[[6.52815927 0. 0.]
[0. 4.47255915 0.]
[0. 0. 3.43215579]]
[[-0.69070483 -0.82525254 -0.09831266]
[-0.39580866 -1.03738639 0.01528359]
[-0.55183905 -1.08911378 -0.13910223]
[-0.23314023 -0.80047953 -0.11865892]
[-0.8239477 0.42848608 -0.59172685]
[-0.79237896 -0.20490276 -0.27004569]
[-0.90888385 0.42178611 -0.50058914]
[-1.08558032 0.46636909 -0.00699471]
[-0.62039882 0.14063987 0.81084737]
[-0.5198665 0.05004723 0.67430674]
[-0.51494257 0.09498759 0.78536494]
[-0.41553912 0.09627489 0.30402772]
[-0.51784555 0.1042679 0.79985753]
[-0.9290176 0.46472989 -0.50394211]]
In [23]:
Calculating sigma square i = 0,1,....13 and S
var_ls = []
x_var = np.var(x_n,axis=1)
x_var = np.ravel(x_var)
print(x_var.shape)
print(x_var)
for i in range (V.shape[0]):


localhost
Documents/Python practice/Assignments/Guided Projects/Project3 on EFA /
Project on EFA - Jupyter Notebook
Airline Passenger Satisfaction | Kaggle
jupyter
Project on EFA Last Checkpoint: Last Sunday at 12:00 AM (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
In [23]:
Calculating sigma square i = 0,1,....13 and S
var_ls = []
x_var = np.var(x_n,axis=1)
x_var = np.ravel(x_var)
print(x_var.shape)
print(x_var)
for i in range (V.shape[0]):
s = np.sum(np.square(np.ravel(V[i:])))
sig_2 = x_var[i] - s
var_ls.append(sig_2)
var_ls = np.array(var_ls)
S = np.diag(var_ls)
print(S)
(14,)
[1.76311414 2.32583197 1.95698483 1.63229974 1.76764022 1.82115689
1.73997514 1.77684714 1.65984098 1.73079886 1.39451944 1.60121119
1.38217027 1.72204345]
[[-12.66976007 0. 0. 0. 0.
0. 0. 0. 0. 0.
0. 0. 0. 0.]
[0. -10.93926195 0. 0. 0.
0. 0. 0. 0. 0.
0. 0. 0. 0.]
[0. 0. -10.07504049 0. 0.
0. 0. 0. 0. 0.
0. 0. 0. 0.]
[0. 0. 0. -8.88968099 0.
0. 0. 0. 0. 0.
0. 0. 0. 0.]
[0. 0. 0. 0. -8.04513873
0. 0. 0. 0. 0.
0. 0. 0. 0.]
[0. 0. 0. 0. 0.
-6.77899125 0. 0. 0. 0.
0. 0. 0. 0.]
[0. 0. 0. 0. 0.
0. -6.11739877 0. 0. 0.
0. 0. 0. 0.]
[0. 0. 0. 0. 0.
0. 0. -4.82596389 0. 0.
0. 0. 0. 0.]

Dimensionality Reduction Transformation