1.Perform Exploratory Data Analysis and produce 2-3 plots that show either important features or interesting patterns in the data. It is up to you what you want to highlight.

Solution:

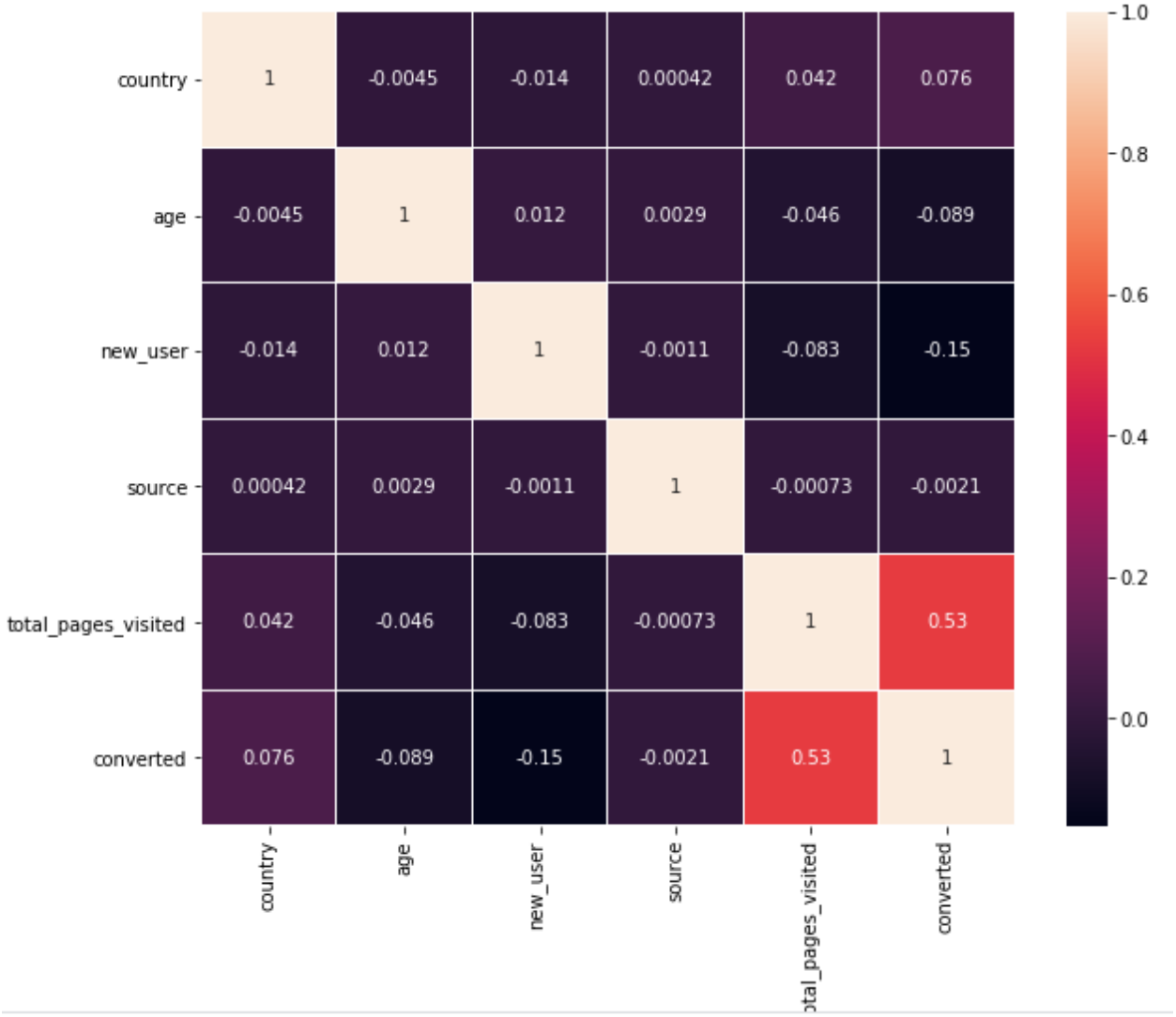

Fig. 1 Pearson's correlation among different variables

Fig. 1 shows that there is no correlation existing among input variables. So, there is no chance of multicollinearity. Moreover, regarding the relationship between the input variables and response variable 'converted', only 'total page visited' variable shows higher correlation with output variable. To confirm this, we again plotted the input variable importance in its impact on output variable as shown in Fig. 2.

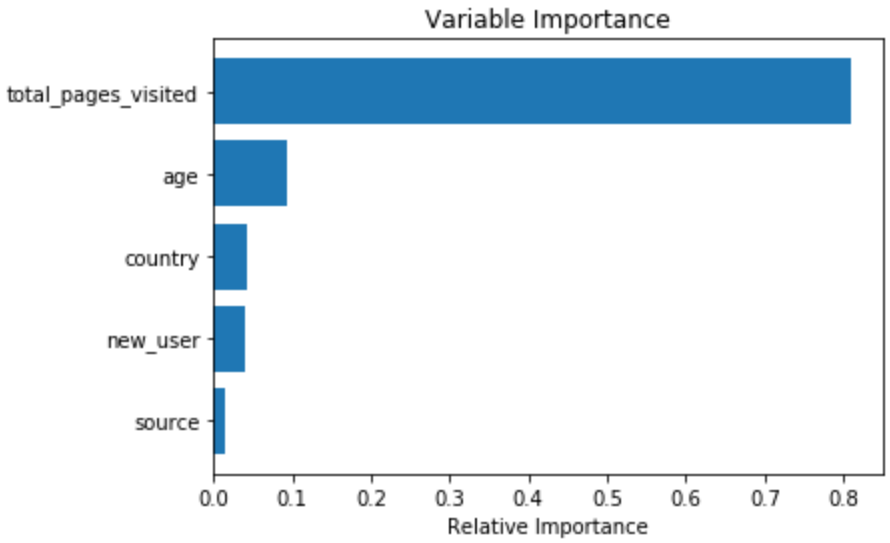

Fig. 2 Relative importance of input variable in its contribution to output variable

Fig. 2 clearly demonstrates that 'total pages visited' is the most significant factor contributing to the output response, whereas 'source' variable is the least significant.

2. Build a model to predict conversion rate and critically evaluate it, explaining your choice of model and performance metric.

Solution: Since 'source' variable is the least significant variable affecting output response, we would drop or delete this input variable and continue to develop a prediction model with the rest of the four significant factors. We use machine learning models including logistic regression, decision trees, random forest and neural network to train 80% of the data, and test rest of the 20% data using the 'accuracy' metric which is used as the classification performance metric. The 'accuracy' metric is same as the 'R-squared' used as regression performance metric.
Following are the accuracy scores for different machine learning models:
Logistic Regression: 0.986
Decision trees: 0.985
Random forest: 0.985
Neural network: 0.99
Clearly, logistic regression and neural network performs the best, however, other models are also good enough. Neural neutral is more like black box model providing no physical insights on the impact of input variables on response variable. It is used when there is a non-linear effect of the factors on response variable. Whereas, logistic regression model is used when the input-output relationship can be explained using linear response. If the relationship can be summarized using the simplified model such as logistic regression, why to choose complex model. So, I would select logistic regression whose model summary is shown below, which clearly demonstrates that rest of the four input variables are significant enough.

```
                          Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.702
Dependent Variable: y                AIC:              21603.4930
Date:               2020-05-14 18:15 BIC:              21655.6979
No. Observations:   252960           Log-Likelihood:   -10797.
Df Model:           4                LL-Null:          -36177.
Df Residuals:       252955           LLR p-value:      0.0000
Converged:          1.0000           Scale:            1.0000
No. Iterations:     10.0000
-----------------------------------------------------------------
          Coef.    Std.Err.      z       P>|z|     [0.025    0.975]
-----------------------------------------------------------------
const    -8.4748     0.1126   -75.2817   0.0000   -8.6954   -8.2541
x1        0.4912     0.0201    24.3884   0.0000    0.4517    0.5306
x2       -0.0742     0.0026   -28.4754   0.0000   -0.0794   -0.0691
x3       -1.7419     0.0390   -44.6959   0.0000   -1.8183   -1.6655
x4        0.7635     0.0068   112.0272   0.0000    0.7501    0.7769
=================================================================
```

3. Come up with recommendations for the product team and the marketing team to improve conversion rate.

Solution: Here, we perform the cross plot of the rest of the four input variables with the output response, to determine what range of input variables values significantly result in conversion or value '1', so we can target only those customers using the same configured advertisements, and design new advertisements or product for the other segmented customers who do not convert.

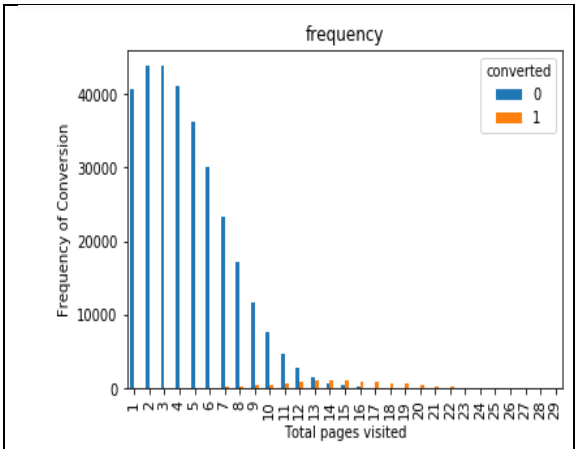

Fig. 3. Frequency of conversion by total pages visited

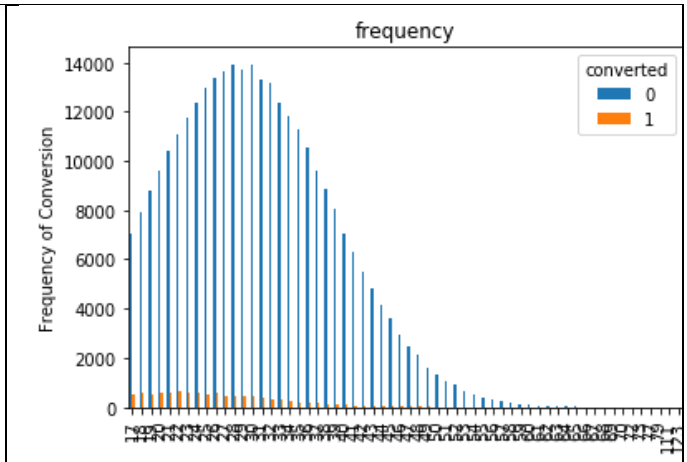

Fig. 4. Frequency of conversion by age

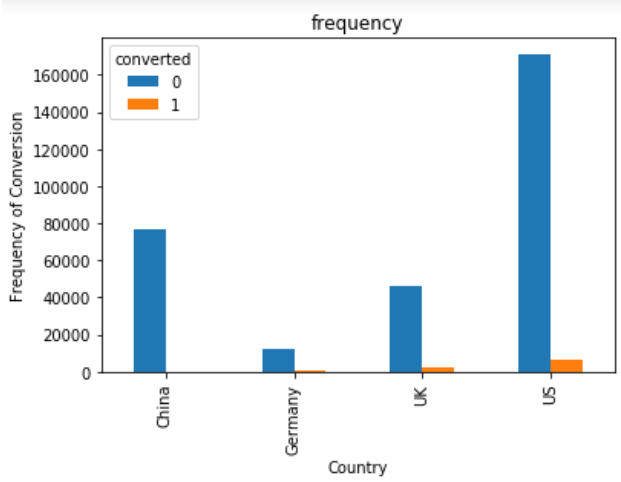

Fig. 5. Frequency of conversion by country

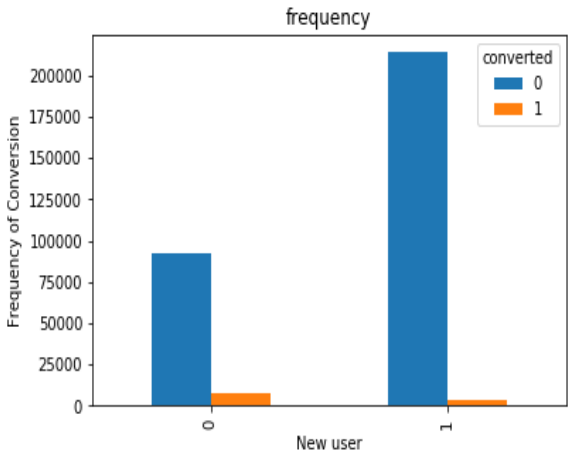

Fig. 6 Frequency of conversion by new user

Fig. 3: Total pages visited >12-13 has the highest conversion.

Fig. 4: Age <38-40 has the highest conversion.

Fig. 5: Users from US and UK typically converts

Fig. 6: Old users has the highest conversion

Users having above characteristics should be targeted for the current purchase, so the conversion rate would dramatically increase as all these users will tend to purchase the product.