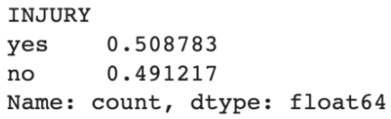
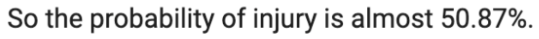
Q1.

a)



```
INJURY
yes     0.508783
no      0.491217
Name: count, dtype: float64
```



So the probability of injury is almost 50.87%.

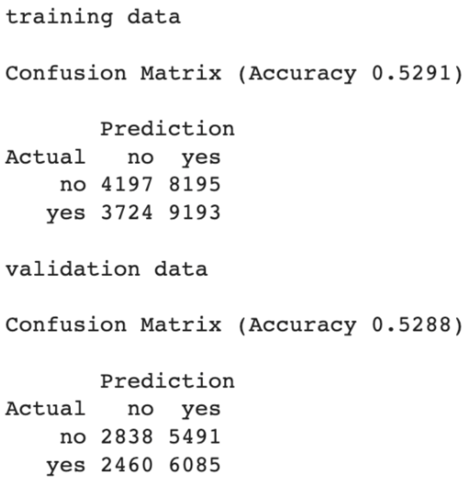
c)

i. Predictors which can be included in the analysis stated no intial reports on accident is availab le:

1. ALIGN_I -> Can be known based on location characteristics
2. WRK_ZONE→ Assuming this can be known using Google maps, as it shows divergence whenever there is construction on the road.
3. WKDY_I_R→ This would be known , based on day on which accident reported
4. INT_HWY→ Assuming this would be known based on location characteristic of incident
5. LGTCON_I_R→ Based on the time of accident
6. SPD_LIM→ Based on location characteristic of the area in which accident reported
7. TRAF_CON_R→ Based on location characteristics
8. HOUR_I_R→ Based on time on which accident reported
9. TRAF_WAY→ Based on location characteristics
10. WEATHER_R→ Weather report of the day

If an accident is reported in a particular area and we are not aware of the details involved in accident, the above variables can act as predictors based on location characteristics, weather characteristics.

ii.
Confusion Matrix and Statistics



```
training data

Confusion Matrix (Accuracy 0.5291)

       Prediction
Actual   no  yes
    no 4197 8195
   yes 3724 9193

validation data

Confusion Matrix (Accuracy 0.5288)

       Prediction
Actual   no  yes
    no 2838 5491
   yes 2460 6085
```

iii. Overall error for the validation set is 1-0.5288 = 47.12%.

iv. Overall error using validation set 0.4712

Naïve rule's error 0.4913

Improvement 3.95%

Q2.

```
Importance of components:
                       PC1      PC2      PC3     PC4     PC5     PC6
Standard deviation     632.048 196.88568 34.47943 3.61890 1.94474 1.49593
Proportion of Variance   0.909   0.08821  0.00271 0.00003 0.00001 0.00001
Cumulative Proportion    0.909   0.99724  0.99995 0.99998 0.99999 0.99999
```

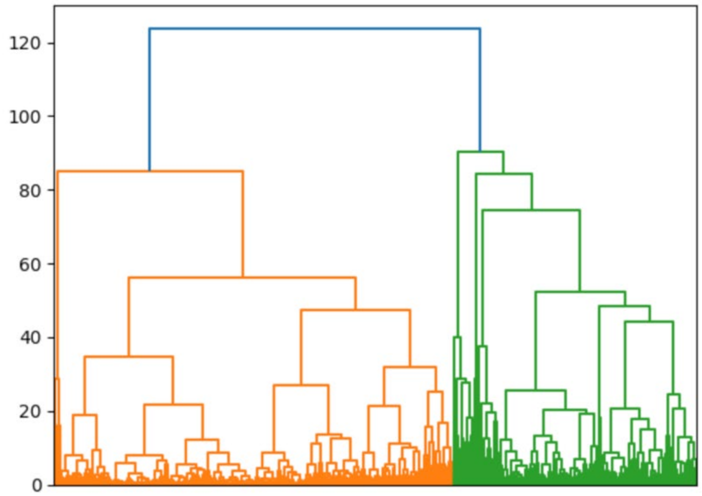
Based on the results from the code as shown above, we just need 2 PCs, as they explain >0.99 of the total variance.

Logistic Regression	Naïve Bayes	Random Forest
Reference Prediction X0 X1 X0 609 53 X1 229 213 Accuracy : 0.7446	Reference Prediction X0 X1 X0 489 173 X1 172 270 Accuracy : 0.6875	Reference Prediction X0 X1 X0 508 154 X1 135 307 Accuracy : 0.7382

Logistic regression has the highest accuracy and performs the best.

Q3.

a)



The dataset can be split into two clusters.

b) Normalizing the data is important to ensure that the "distance measured" accords equal weight to each variable - without normalization, the variable with the largest scale will dominate the measure. Here the variables that have the largest scale are Balance, Bonus_trans, Fligt_miles_12mo, and Days_since_enroll. These variables will dominate the measure if we don't standardize the data.

e) The cluster membership from the two methods is comparable.

f) For the "minimal, non-frequent" flyers, two types of offers might be used.

1. Offers to liquidate the mileage, to remove it as a liability (e.g., offers to purchase magazine subscriptions)
2. Offers for special mileage bonuses if a number of segments or miles is flown in a limited period of time, in case some of these flyers are regular customers of other airlines, or new flyers, in hopes that some of them will become more "invested" in East-West.

For the frequent, loyal flyer we might

1. Offer luxury goods in conjunction with partners (high end vacations, exclusive real estate, etc.) -- frequent flyers are likely to be relatively prosperous (compared to non-frequent flyers).