



IBA Challenge 2019 Case

AIESEC'S DIGITAL TRANSFORMATION Final Report



Team: 5th Quartile

Team Members

Caner Irfanoglu
Diven Kumar Sambhwani
Sunil Padikar M
Vinay Govindan



Table of Contents

1. Executive Summary	3
2. Data Preparation & Initial Findings.....	4
3. Labelling Opportunities.....	4
4. Country Filter.....	5
5. Skill & Background Filter.....	6
5.1 Analysis of Skills Across Clusters	6
5.2 Association Analysis	6
6. Analytic Hierarchy Process (AHP).....	7
7. Input & Output Example.....	8
8. Assumptions List	9
9. Discussion & Conclusion.....	9
10. Appendix.....	10
10.1 Data Cleaning Process	10
10.2 Exploratory Analysis Code	12
10.3 Preparing Cluster Analysis Data	13
10.4 The Elbow Plot.....	14
10.5 Association Analysis Code.....	14
10.6 Snapshot of Hierarchical Country Filter.....	17
10.7 Skill Preferred & Background Required.....	18
10.8 Background required.....	19
10.9 Background Preferred	20
10.10 Higher Resolution Ranking Opportunities.....	21
10.11 Resumes of Team Members.....	22
11. References	28

1. Executive Summary

The purpose of this report is to describe our analysis for addressing AIESEC's needs and requirements. Furthermore, the report explains how we alleviate the problem of low ratio of opportunity realizations by improving the similar opportunity recommendation system.

While performing exploratory analysis, we found that number of applicants rapidly increased from 7,200 to 157,490 during the last ten years. However, 55% of these opportunities had no applicants. Also, the locations where most opportunities are posted were extremely different than the locations where candidates applied. Considering these evidence, we defined our task as to recommend candidates, the most relevant opportunities while increasing the ratio of realized opportunities.

To do so, we decided to group internship opportunities using a clustering algorithm ('k-means') based on 5 suggestive characteristics which are 'Openings', 'Duration', 'Favorite Count', 'Applicant Count' and 'Recency'. This methodology provided us insights about factors which made an internship opportunity more popular. Based on the groups we found, we eliminated the expired opportunities and the ones already getting extreme number of applications.

We created a hierarchy between countries based on the ratio of applicants per opening ('app_by_opening'). Then, when a country is selected by the user, it became possible to identify similar countries he/she might be interested in. We performed [Association Analysis](#) for the 'Skill', 'Background' and 'Language' information (SBL). Like the country selection, this is done to identify related SBL information on the given input. [Analytic Hierarchy Process](#) (AHP) is a simple technique that is used for tuning the contribution of the parameters mentioned above.

As a result, we get a versatile and robust recommendation system with adjustable weights. To provide a match score, the weights are summed up, 100 being the perfect score, and then multiplied by a popularity factor derived from 'app_by_opening'. When other parameters are the same, the popularity factor favors the least applied opportunities.

The long-tailed distributions, in which a few products are in very high demand while most are in very low demand, describe endless-inventory e-businesses like Amazon and Netflix (Goldstein, 2014). AIESEC opportunities also follow the long-tailed distribution. The efficiency of recommendation system defines whether an applicant can discover the most suitable match or not. The current recommender seems to be limited to exact matching capability and heavily suggests opportunities from the searched company but, unrelated to input parameters. Therefore, it is observed that, it often suggests results not aligned with candidates search criteria(s).

The recommendation system we developed is flexible and can be integrated to AIESEC's site with ease. It can also be further developed by adding the information currently not available such as the organization name, logistics and compensation. With these extended capabilities, it can improve the current recommendation system as well as the search engine. As a result, thousands of undiscovered opportunities might be matched with the right applicants.

2. Data Preparation & Initial Findings

First, during the [data cleaning process](#), we considered unique opportunities (duplicate opportunities were present) in the dataset. Irrational values such as negative openings or 0-week durations are excluded. Extreme values for openings, duration, application and favorite counts are excluded with the help of box plots. Sub setting the program to global volunteer, date time conversions and creating new 'Recency' field (indicates how many weeks passed since the opportunity was created) are also done in this part.

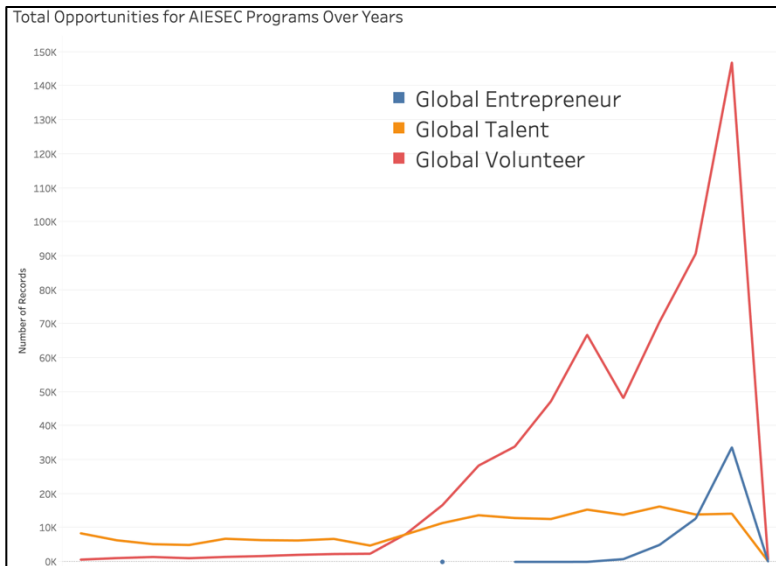


Figure 2.1: Number of opportunities for different programs since 2000

Observation:

- Global Volunteer program has the most opportunities in recent years
- Before 2009, Global Talent program had more applicants than Global Volunteer. However, starting at 2008, there is significant increase in volunteer opportunities, reaching around 10 times more opportunities than Global Talent in 2018
- Global Entrepreneur program was started by AIESEC in the year 2012 and had maximum number of opportunities in year 2018

3. Labelling Opportunities

'k-means Clustering' is performed on the cleaned data to categorize the similar opportunities into clusters. The optimal number of clusters 'k' is found as 5 based on the '[Elbow Method](#)'.

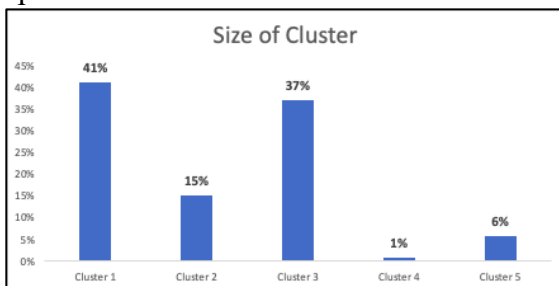


Figure 3.1: Distribution of opportunities within clusters

Value	Favorite Count	Application Count	Openings	Recency
Cluster 1	0.35	1.36	1.09	264
Cluster 2	0.01	0.50	1.02	486
Cluster 3	1.62	4.98	10.08	53
Cluster 4	62.17	191.30	14.71	118
Cluster 5	1.63	5.84	37.81	67

Table 3.1: Clusters and centers for each characteristic

Observation:

- Clusters 1 and 3 have the most opportunities while cluster 4 has the least (*Figure 3.1*)
- Cluster 3 and 5 have more recent opportunities and the latter has higher number of openings (*Table 3.1*)
- Clusters 1 and 2 are not useful for recommendations since the opportunities are already closed and application counts increased significantly after 2008 (*Figure 2.1*)
- Cluster 4 contains the most popular opportunities, but corresponds to less than 1% of opportunities

The cluster 3 and 5 show the most potential for betterment as interest shown is low but number of openings is high. These opportunities are our target for improving the ratio of realizations. Therefore, our recommendation system uses a subset of opportunities only coming from these clusters.

4. Country Filter

We plotted total number of applicants and job openings for clusters 4 and 3 & 5 aggregated for each country. This allowed us to observe the demand (application counts) and supply (number of openings). Consequently, the inconsistencies can be seen below:

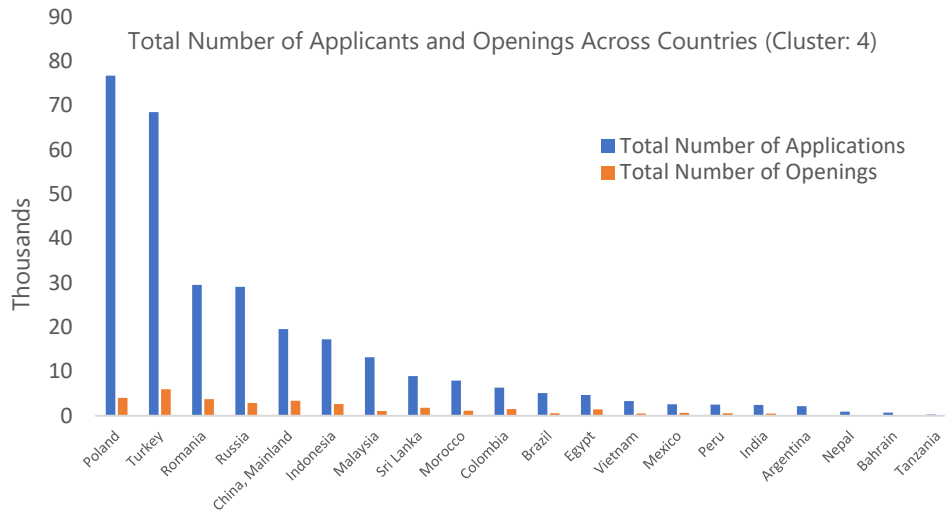


Figure 4.1: Most popular opportunities and their countries

Observation:

- As it can be seen in *Figure 4.1* number of applicants greatly exceeds the number of opportunities in these nations

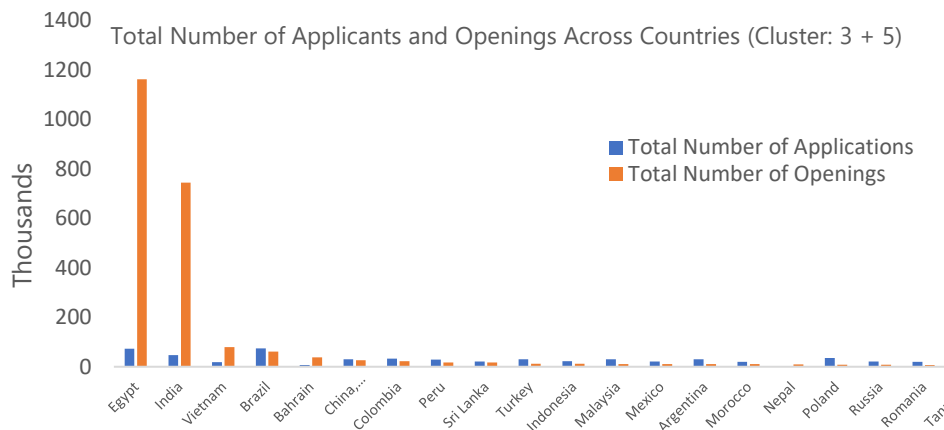


Figure 4.2: Least popular opportunities and their countries

Observation:

- As it can be seen in *Figure 4.2* number of opportunities greatly exceeds the number of applications significantly in Egypt and India

Compared to *Figure 4.1*, the bars in *Figure 4.2* is reversed. This inconsistency showcases the imbalance of application to opening ratio across the countries.

To address the unequal distribution, we created a new measure by dividing number of applications by number of openings for each opportunity. We labeled this measure 'app_by_opening'. We used 'app_by_opening' to rank countries as follows:

1. Group the data by country for opportunities
2. Get the average of 'app_by_opening' for each country
3. Sort the obtained list in descending order

In this 'country list', close proximities represented countries getting similar attention. We assumed that when candidates are interested in a certain opportunity based in a country, they will also be interested in those countries which are near in the list. The recommendation system suggests opportunities from similar countries using a real-time percentage range in the final model. This will not only ensure that the applicants will receive well-balanced recommendations, but also help promoting the less popular opportunities.

5. Skill & Background Filter

Skill, Background and Language information (SBL) are provided as ‘required’ and ‘preferred’ for each category totaling 6 separate features in data set. Also, there exist multiple values in cells indicating association for SBL information.

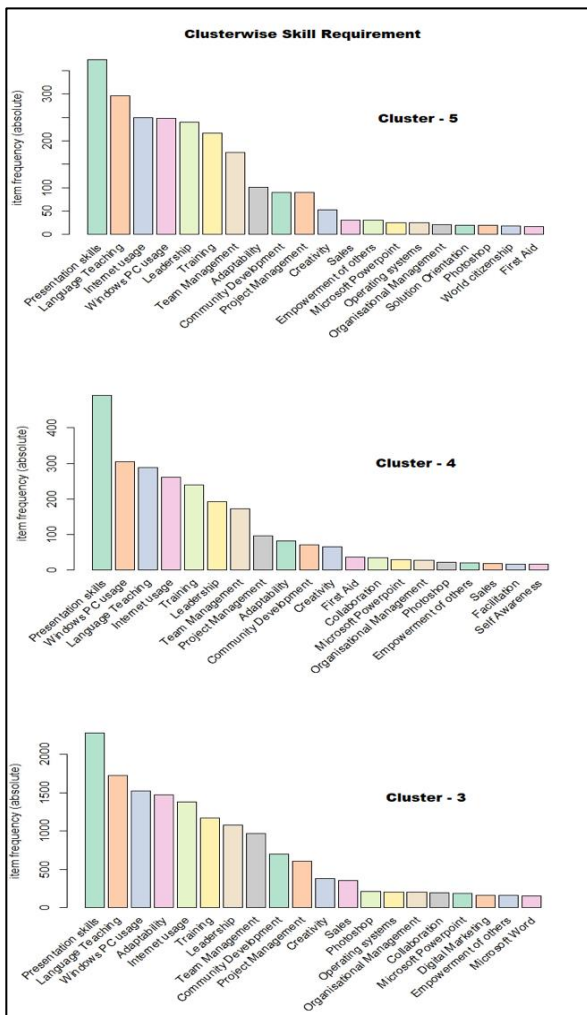


Figure 5.1: Most required skills among clusters of interest

5.1 Analysis of Skills Across Clusters

In Figure 5.1. the primary skills such as ‘Presentation Skills’, ‘Internet Usage’, ‘Windows PC Usage’, ‘Training’, ‘Language Teaching’ are generic for each category. Also, following less common skills have similar order and distribution. Analogous characteristics are noted across the [other SBL fields](#). Consequently, we concluded that the SBL information cannot be considered as a significant attribute that influences the chances of an opportunity getting applied or realized.

5.2 Association Analysis

To obtain the association of each item in SBL sections following steps are followed:

1. Get a list of unique items in the category of interest
2. Read each opportunity as transaction object
3. Set the parameters for APRIORI¹ properly (Setting support² parameter low for capturing all items, while setting confidence³ parameter high for capturing significant relations)
4. Sort the results of association by Lift⁴
5. Feed each unique item to the APRIORI function in a loop and append output to data frame

After performing above procedure for all SBL categories following observations are made:

- ‘Language Required’ ‘Language Preferred’ were generic across opportunities
- ‘Skills Required’ and ‘Background Required’ columns had many missing values (94% and 97% respectively) within the subset for recommendations (‘open’ status opportunities from clusters 3 & 5)

Due to the above reasons, out of 6 SBL categories only ‘Skills Preferred’ and Background ‘Preferred’ are used for recommendations. The resultant two lists obtained from APRIORI function output consisted of the skill or the background itself and associated skills. Next, these lists are used for further filtering the subset. Eventually, the quality of matches is rated for remaining opportunities.

¹APRIORI: This function does association analysis in a computationally effective way (Dhanabhakym & Punithavalli, 2011)

²Support: Fraction of transactions that contain the item A, that being searched

³Confidence: How often item B appears with item A

⁴Lift: The lift ratio indicates how efficient the rule is and it is found by dividing confidence to support

6. Analytic Hierarchy Process (AHP)

AHP is a simple yet powerful tool that was first developed within the management science field around 40 years ago (Saaty, 1980). It was developed to help managers make more effective decisions by structuring and evaluating the relative attractiveness of competing options or alternatives (Handfield et al., 2002). Within the scope of this project, AHP will be used for deciding relative influence of ‘Country’, ‘Skill’ and ‘Background’ information on overall rating (weights). Also, different weights will be assigned to sub-sections of the main categories based on different criteria. The weights set in the model can be adjusted in the future based on the performance of the model. Initial settings are as follows:

- Country, Skill and Background parameters define the overall rating out of 100
- Initial weights for ‘Country’, ‘Skill’ and ‘Background’ are set to 20, 40, 40 respectively
- If a parameter equals to itself, it gets 100% of its parent weight
- For country parameter, if suggestion is a physical neighbor, it gets 75% of total ‘Country’ weight
- Countries within proximity in the ‘country list’ are selected as a proportion of ‘app_by_opening’. The countries with higher proximity gets 60% of initial ‘Country’ weight
- For the ‘Skill’ and ‘Background’ total points are broken into ‘n’ & ‘m’ respectively. These numbers represent the total number of skills and backgrounds exist for given input
- If a ‘Skill’ or a ‘Background’ found in the ‘associate list’, they get 80% of parent value for first associate and declines 5% progressively
- After the weights summed up, opportunities having ‘app_by_opening’ between 1 & 2 multiplied by 0.85 (85%) and opportunities with ‘app_by_opening’ above 2 multiplied by 0.7 (70%) for ensuring the low interest opportunities will get higher ranking

Below chart visualizes the Rating Process for opportunities.

Please refer to Appendix for larger resolution.

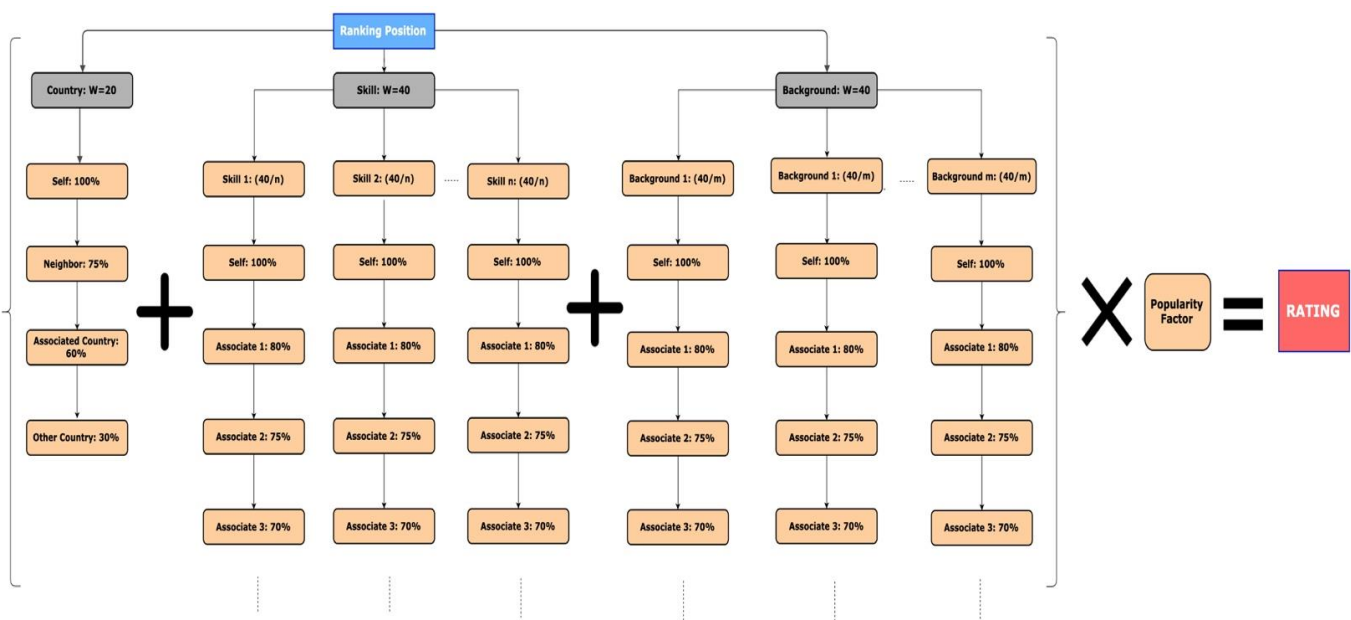



Figure 6.1. Ranking the opportunities

7. Input & Output Example

Below is a screenshot of the recommendation results generated by the prototype. When user specifies Country, Background and Skill Information, the recommender displays most relevant opportunities listed by match rating.

AIESEC




86%

723091

Country: Brazil

Skills Matched: Windows PC usage, Presentation skills, Leadership, Corel Draw, Adobe Illustrator, Team Management, Training, Mac usage, Photoshop, Internet usage, Organisational Management

Background Matched: Economics, Agriculture, Public relations, Biology, Business administration, Marketing, Environmental engineering




77%

795133

Country: India

Skills Matched: Windows PC usage, Presentation skills, Leadership, Adobe Illustrator, Team Management, Training, Mac usage, Photoshop, Internet usage, Organisational Management

Background Matched: Economics, Agriculture, Marketing




70%

798083

Country: India

Skills Matched: Windows PC usage, Presentation skills, Leadership, Team Management, Training, Mac usage, Internet usage, Organisational Management

Background Matched: Biology, Economics, Agriculture




70%

299544

Country: Senegal

Skills Matched: Windows PC usage, Presentation skills, Leadership, Training, Photoshop, Internet usage

Background Matched: Business administration, Environmental engineering




69%

259923

Country: Egypt

Skills Matched: Windows PC usage, Presentation skills, Leadership, Team Management, Training, Internet usage, Organisational Management

Background Matched: Agriculture, Marketing



69%

759917

Country: Egypt

Skills Matched: Windows PC usage, Presentation skills, Leadership, Team Management, Training, Internet usage, Organisational Management

Background Matched: Agriculture, Marketing

Figure 7.1. Sample Input & Output for Recommendations

8. Assumptions List

The following assumptions have been made during the process of analysis and building a recommendation system:

- Five separate groups (clusters) represent all unique opportunities
- Clusters 3 & 5 show the most potential for betterment
- When candidates are interested in a certain opportunity based in a country, they might also be interested in the nearby countries in the hierarchical '[country list](#)'
- Combination of 'Country', 'Skill Preferred' and 'Background Preferred' features provide the most relevant recommendations
- Initial weights assigned by assuming 'Skill Preferred' and 'Background Preferred' are more influential than 'Country' when rating a relevant opportunity

9. Discussion & Conclusion

We began the IBA challenge by analyzing datasets provided by AIESEC and drilling down further in our exploratory analysis. Our main findings were the following:

- Between 2008 and 2018 number of applicants rapidly increased from 7,200 to 157,490
- About 55% of opportunities have no applicants
- Global Volunteer program consists around 70% of the opportunities
- 81% of data has 'open' status

After this stage, we decided to concentrate on opportunities data. We applied 'k-means Clustering' for grouping opportunities and getting insight about how various characteristics affect the applicants' behavior. We are convinced that 5 clusters were appropriate for representing the data, and the country was the main factor leading into inconsistencies in applicants' behavior.

Next, we decided to run association analysis on 'Skills', 'Background' and 'Language' parameters. This provided us information about how different items in these categories related with each other.

Afterwards, by putting all our findings together, we developed a recommendation system focusing on similar opportunities where applicants can stand out. The purpose of this system is two-folds. Improving the accuracy of the suggestions while highlighting less popular but related opportunities.

Finally, Analytic Hierarchy Process (AHP) is selected for combining all the filters and dispensing weights for organizing their individual contribution. Currently, the working prototype of AHP based recommender is programmed with a user interface.

Since the information for applicants' profile is not accessible, it was not possible to develop tailored suggestions. Similarly, some information such as visa requirements, logistics and salary provided were not available in the dataset. The prototype can be further improved in the future by adding this information.

To conclude, for AIESEC's Digital Transformation, we first analyzed and explored the provided data sets and then applied unsupervised machine learning, data mining and segmentation techniques on opportunities. Ultimately, we used AHP to connect our findings and programmed a recommendation system. When the recommendation system is integrated to AIESEC's site with the above-mentioned parameters, it can innovate the current recommendation system as well as the search engine. As a result, thousands of undiscovered opportunities can be matched with the right applicants.

10. Appendix

10.1 Data Cleaning Process

- The csv files are read to R with read.csv command. The Null values are filtered out as ""(empty string), "(space) and NA. Those values marked as NA for catching all possible missing combinations.

```
opp_data <-  
  read.csv("/Users/Caner/Desktop/projects/iba/iBA_ass1/opportunity_iba_ch  
    allenge.csv",  
  na.strings=c("", " ", "NA"), stringsAsFactors=FALSE)
```

- In case required, code for reading chunks is provided. The code can be easily restructured for reading random lines to get a homogenous sample from the overall dataset.

```
con <- file("/Users/Caner/Desktop/projects/iba/opportunity_iba_challenge.csv", "r")  
opp_chunk <- list()  
for (i in 1:ceiling(813635/50000)){  
  opp_chunk(i) <- readLines(con, 50000)  
}  
close(con)
```

- Issue 1:** Column Name "X" is only indicating the index.

Action Taken: The column is dropped from the dataset.

```
opp_data <- opp_data[, -which(names(opp_data) %in% c("X"))]
```

- Issue 2:** Opportunity ids are not unique

Action Taken: Data frame is grouped by opportunity id. So that, similar data treated as duplicates. Hence, they are eliminated.

```
opp_data_unique <- sqldf("select * from opp_data group by opportunity_id")
```

- Issue 3:** Date Columns have erroneous data types

Action Taken: Date Columns are converted to type posix

```
columns_to_convert_dt <-  
  c("created_at", "applications_close_date", "earliest_start_date", "latest_  
    end_date", "matched_or_rejected_at", "experience_start_date",  
    "experience_end_date")  
for (col in columns_to_convert_dt){  
  opp_data[,col] <- as.POSIXct(opp_data[,c(col)], format="%Y-%m-%d %H:%M:%S")  
}
```

- Issue 4:** Recency Column is missing

Action Taken: Recency is created by subtracting latest created date from the corresponding row value.

```
recency = as.integer(round(difftime(max(created_at), created_at , units =  
  "weeks"), 0))
```

- Issue 5:** Data for Global Volunteer needs to be extracted

Action Taken: Required data is filtered out

```
opp_data_unique %>% filter(programme_id == 'Global Volunteer')
```

- Issue 6:** Openings Column has outliers

Action Taken: There were 967 observations having opening greater than 100. That count corresponds to 0.23% of the total data. Since the percentage is low enough and the upper values are extreme, these values are considered as outliers. Openings having values higher than 100 filtered out.

```
opp_data_kmeans <- opp_data_kmeans[opp_data_kmeans$openings <100,]
```

- **Issue 7:** Application Count Column has outliers

Action Taken: There were 12 observations having application count greater than 1000. That count corresponds to 0.003% of the total data. Since the percentage is low enough and the upper values are extreme, these values are considered as outliers. Openings having values higher than 1000 filtered out.

```
opp_data_kmeans <- opp_data_kmeans[opp_data_kmeans$opportunity_applications_count < 1000,]
```

- **Issue 8:** Openings column has values lower than 1

Action Taken: Since an opportunity needs to be available for at least 1 intern, those rows are considered as invalid. 186 observations having invalid values are removed from the dataset.

```
opp_data <- opp_data[opp_data$openings > 0,]
```

- Ggpairs

```
ggpairs_product <- ggpairs(opp_data_kmeans[,which(names(opp_data_kmeans) %nin%  
c("opportunity_id", "programme_id", "created_at"))],  
upper = list(continuous = ggally_points),  
lower = list(continuous = points),  
title = "Opportunity data pairs")
```

- Dropping Row Number Column

```
opp_data <- opp_data[, -which(names(opp_data) %in% c("X"))]
```

- Filtering out Duplicate Opportunities

```
opp_data_unique <- sqldf("select * from opp_data group by opportunity_id")
```

- Converting Date Columns

```
columns_to_convert_dt <-  
c("created_at", "applications_close_date", "earliest_start_date",  
"latest_end_date", "matched_or_rejected_at", "experience_start_date",  
"experience_end_date")  
for (col in columns_to_convert_dt){  
  opp_data[,col] <- as.POSIXct(opp_data[,c(col)], format="%Y-%m-%d %H:%M:%S")  
}
```

- Mutating Recency Column

```
recency = as.integer(round(difftime(max(created_at),  
created_at, units = "weeks"),0))
```

- Filtering out Global Volunteer

```
opp_data_unique %>% filter(programme_id == 'Global Volunteer')
```

- Remove outliers from Openings

```
opp_data_kmeans <- opp_data_kmeans[opp_data_kmeans$openings <100,]
```

- Remove outliers from Applications Count

```
opp_data_kmeans <- opp_data_kmeans[opp_data_kmeans$opportunity_applications_count < 1000,]
```

- Scaled Ranges

```
scaled_ranges <- sapply(as.data.frame((opp_data_kmeans[, which(names(opp_data_kmeans) %nin% c("opportunity_id", "programme_id", "created_at"))])), range)
```

- Scaling

```
opp_data_scaled <- scale(opp_data_kmeans[, which(names(opp_data_kmeans) %nin% c("opportunity_id", "programme_id", "created_at"))])
```

- Opening values lower than 1

```
opp_data <- opp_data[opp_data$openings > 0,]
```

- Unscaling

```
opp_centers_4 <- unscale(opp_4_clusters$centers, opp_data_scaled)  
# dataframe of centers for each cluster & column
```

- Appending the Cluster identifier to dataset

```
opt_w_cluster_numbers_4 <- cbind(opp_data_kmeans, opp_4_clusters$cluster)  
# appended corresponding cluster to each product  
names(opt_w_cluster_numbers_4)[8] <- "CLUSTER"
```

10.2 Exploratory Analysis Code

- **Issue 1:** Column names “X.1” and “X” are only indicating the index
Action Taken: The columns are dropped from the dataset

```
opp_data_unique <- opp_data_unique[, which(names(opp_data_unique) %nin% c("X.1", "X"))]  
  
# Dropped the row identifier excess columns
```

- **Issue 2:** Opportunity dataset application counts and application dataset row counts needs to be cross verified

Action Taken: Find the ratio of matching application counts for all opportunities

```
# CHECKING IF opportunity data opportunity_applications_count &
# counts of rows in applications dataset are matching
app_counts_df <- opp_app_data %>%
  group_by(opportunity_id) %>%
  summarise(n= n())
# Dataframe for row counts of opportunity id's in application df

opp_data_matchings <- opp_data_unique[opp_data_unique$opportunity_id %in%
  app_counts_df$opportunity_id,
  c("opportunity_id",
    "opportunity_applications_count")]

# Subset opp_data_unique. Only existing opportunities in application df
app_counts_common <- app_counts_df[app_counts_df$opportunity_id %in%
  opp_data_matchings$opportunity_id,]

# Subset app_counts_df. Only existing opportunities in opportunity df
binded_df <- cbind(app_counts_common,
  opp_data_matchings$opportunity_applications_count)

# bind application counts from 2 datasets
final_df <- binded_df %>% mutate(isEqual = n ==
  opp_data_matchings$opportunity_applications_count)

# Mutate bool column for checking equality
match_percentage <- sum(final_df$isEqual) / nrow(final_df)

# 98.37 % match. 279427 / 284047. Remaining ~5000 is very close to differs only a
few mostly
```

- **Issue 3:** Opportunity and Application datasets need to be merged

Action Taken: Datasets are merged on opportunity ids

```
opp_data_join <- sqldf("select * from opp_data_unique A inner join opp_app_data B
  ON A.opportunity_id = B.opportunity_id")

# Dataframes Joined
```

- **Issue 4:** Combined dataset has two status columns, leading into confusion

Action Taken: Renamed status columns appropriately

```
names(opp_data_join)[which(names(opp_data_join) == "status")] <-
  "status_opportunity"
names(opp_data_join)[which(names(opp_data_join) == "status..24")] <-
  "status_application"
```

10.3 Preparing Cluster Analysis Data

- **Issue 1:** Cluster dataset needs to be written to csv

Action Taken: Cluster dataset is written to csv with appropriate parameters

```
write.csv(opp_associate[,c("opportunity_id", "opp_background_pref")],
  "/Users/Caner/Desktop/projects/iba/opp_associate_full.csv",
  row.names = FALSE, quote = FALSE)
# writing data to csv. quote = FALSE is significant. Otherwise it is treated as
single string
```

- **Issue 2:** Cluster dataset needs to be read as transaction object

Action Taken: Cluster dataset csv is read from file to R with appropriate parameters

```
trans = read.transactions("/Users/Caner/Desktop/projects/iba/opp_cluster_full.csv"
                          , format = "basket", sep = ",")
# reading cluster data / Takes around 1.5min
```

10.4 The Elbow Plot

'k-means' implements the "elbow" method to help data scientists select the optimal number of clusters by fitting the model with a range of values. This is due the total error within do not vary significantly as cluster number increases after 5.

<https://www.scikit-yb.org/en/latest/api/cluster/elbow.html>

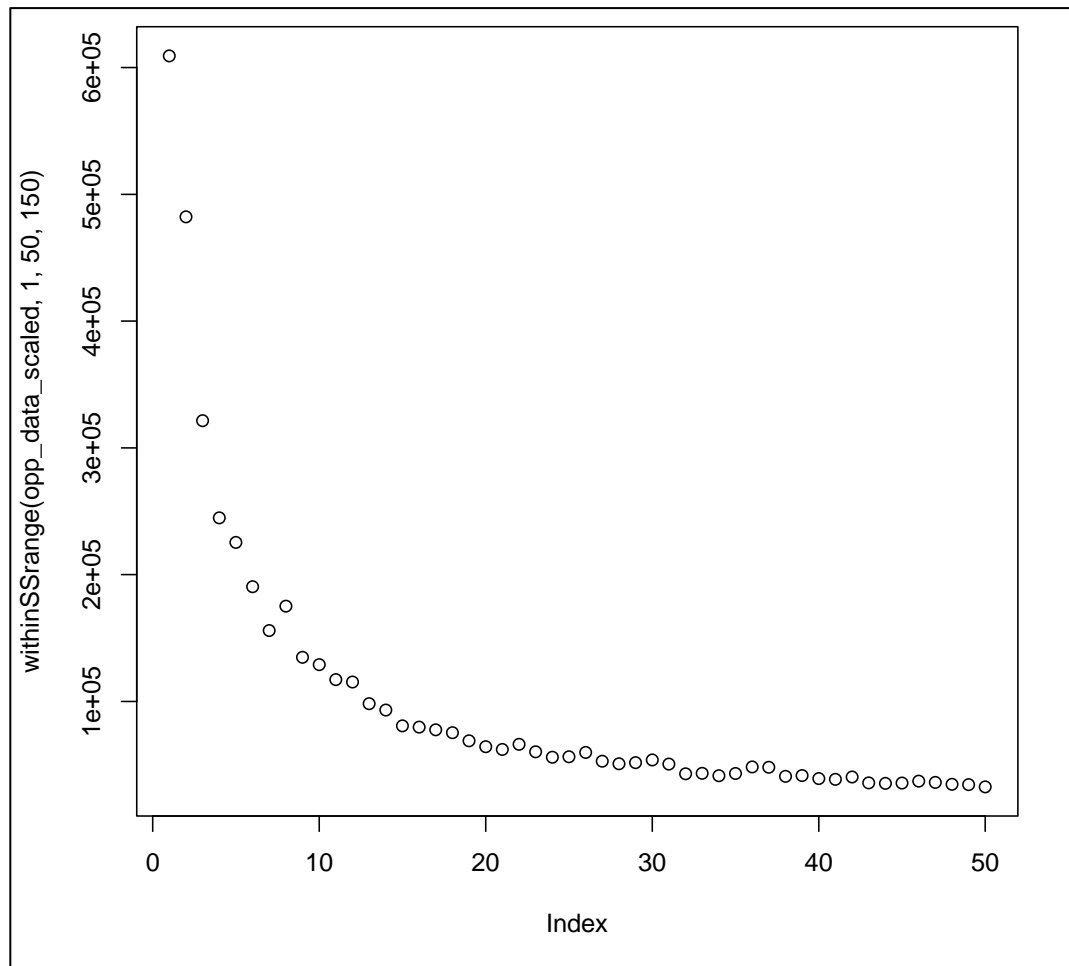


Figure 10.1: Elbow Plot for selecting number of clusters 'k'

10.5 Association Analysis Code

- Reading data

```
opp_data_unique <-
  read.csv("/Users/Caner/Desktop/projects/iba/opp_data_unique.csv", na.strings=c("", " ", "NA"), stringsAsFactors=FALSE)
```


- Getting required columns for association mining

```
opp_associate <- opp_data_unique %>% select(opportunity_id, opp_background_req,
opp_language_req, opp_skill_req, opp_background_pref, opp_language_pref, opp_skill_pref)
```

- Writing extracted data to csv

```
write.csv(opp_associate[,c("opportunity_id", "opp_background_pref")],
"/Users/Caner/Desktop/projects/iba/opp_associate_full.csv", row.names = FALSE, quote
= FALSE)
```

- Reading the data as transaction or associate mining

```
trans =
read.transactions("/Users/Caner/Desktop/projects/iba/opp_associate_full.csv",
format = "basket", sep = ",")
```

- Summary of the transaction object

```
sum_tr <- summary(trans)
```

- Frequency plot of each skills

```
freq_plt <-
itemFrequencyPlot(trans, topN=20, type="relative", col=brewer.pal(8, 'Pastel2'),
main="BG Pref. Absolute Frequency Plot")
```

- Create a list of unique skills to iterate through each skill to find association

```
skills_single_string <- paste(opp_associate$opp_background_pref, collapse = ",")
# Combine all skills in a single string
unique_skills_list <- unique(unlist(strsplit(skills_single_string, ",")))
# Separate the string by commas, unlist for single list and get the unique ones
only 1 306 items
unique_skills_list <- unique_skills_list[unique_skills_list != c("Driver's
licence")]
unique_skills_list <- unique_skills_list[unique_skills_list != c("NA")]
CREATING EMPTY DATA FRAME TO STORE ASSOCIATION OUTPUT AND SETTING SEED VALUE
tic()
no_association <- vector()
df <- setNames(data.frame(matrix(ncol = 6, nrow = 0)), c("LHS", "RHS", "support",
"confidence", "lift", "count"))
# Created empty df for appending associations
set.seed(31)
```

- For each skills find the associated skills. Get top 3 associations and append it to data frame created above

```
for (u in unique_skills_list){
  # sample(1:length(unique_skills_list),10, replace = FALSE)
  # to work with a subset of unique_skills
  association.rules <- apriori(trans, parameter = list(supp=0.00001,
conf=0.03,maxlen=38, maxtime = 5, minlen =2),
                             appearance = list(lhs=u,default="rhs"), control =
list(verbose = FALSE))
  if(length(association.rules) > 0){
    df_new <- DATAFRAME(association.rules) %>% arrange(desc(lift))
    if (nrow(df_new) > 2){
      df <- rbind(df, df_new[1:3,])
      print(c("FINISHED NUMBER ", which(unique_skills_list == u)))
    }
    else if(nrow(df_new) > 0){
      df <- rbind(df, df_new)
    }
  }
  else {
    no_association <- append(no_association, u)
  }
}
toc()
```

After running association for columns skills_required, background_required, skills_preferred, background_preferred. below are the results.

- skills_required has 304 unique skills. 73 did not have any association, but remaining has 76% match rate
- background_required has 71 unique skills. 18 did not have any association, but remaining has 75% match rate
- skills_preferred has 317 unique skills. 39 did not have any association, but remaining has 88% match rate
- background_preferred has 73 unique skills. 2 did not have any association, but remaining has 97% match rate

LHS	RHS	support	confidence	lift	count
{Presentation skills}	{Training}	0.065108069	0.421637165	4.646400153	41495
{Presentation skills}	{Team Management}	0.043520967	0.281839982	4.293006061	27737
{Presentation skills}	{Leadership}	0.050796689	0.328957262	4.095098975	32374
{Java}	{Eclipse}	0.000811203	0.102477701	92.50934928	517
{Java}	{Jscript}	0.000663712	0.083845391	65.40607604	423
{Java}	{Perl}	0.000305966	0.038652131	59.50234125	195

Table 9.1: Sample Output of association mining

10.6 Snapshot of Hierarchical Country Filter

Row Labels	Count of name_entity	Average of app_by_opening
Group1	16414	11.39651169
Slovenia	5	45
France	23	41.74275362
Belgium	13	40.20512821
Iceland	78	33.77478632
Denmark	38	32.08377193
Switzerland	33	30.7020202
Germany	69	27.97168894
Australia	14	25.51785714
Moldova	39	23.87216117
Bosnia and Herzegovina	67	23.77304549
Greece	906	23.10509047
Hungary	841	22.08495129
Spain	49	22.06727891
Singapore	43	21.54728682
Italy	2090	21.20164236
Austria	148	20.38247533
Japan	51	20.35294118
United States	191	20.15538519
Latvia	45	18.50148148
Portugal	692	18.21938997
Croatia	237	18.16672192
Serbia	587	18.07088675
United Kingdom	10	17.63333333
Uruguay	2	17
Estonia	26	16.94234432
United Arab Emirates	27	16.7962963
Finland	125	16.60885796
Czech Republic	145	14.20413912
Lithuania	72	13.56358025
Thailand	126	13.35405882
Macedonia	18	12.64814815
Montenegro	96	12.171875
Canada	27	11.49382716
South Africa	72	11.37164021
South Korea	210	11.2098997
Bulgaria	77	10.77398689
Slovakia	161	10.43173761
Cabo Verde	71	10.2100939
Ukraine	445	9.286826829

Figure 10.2: Snapshot of Hierarchical Country Filter

10.7 Skill Preferred & Background Required

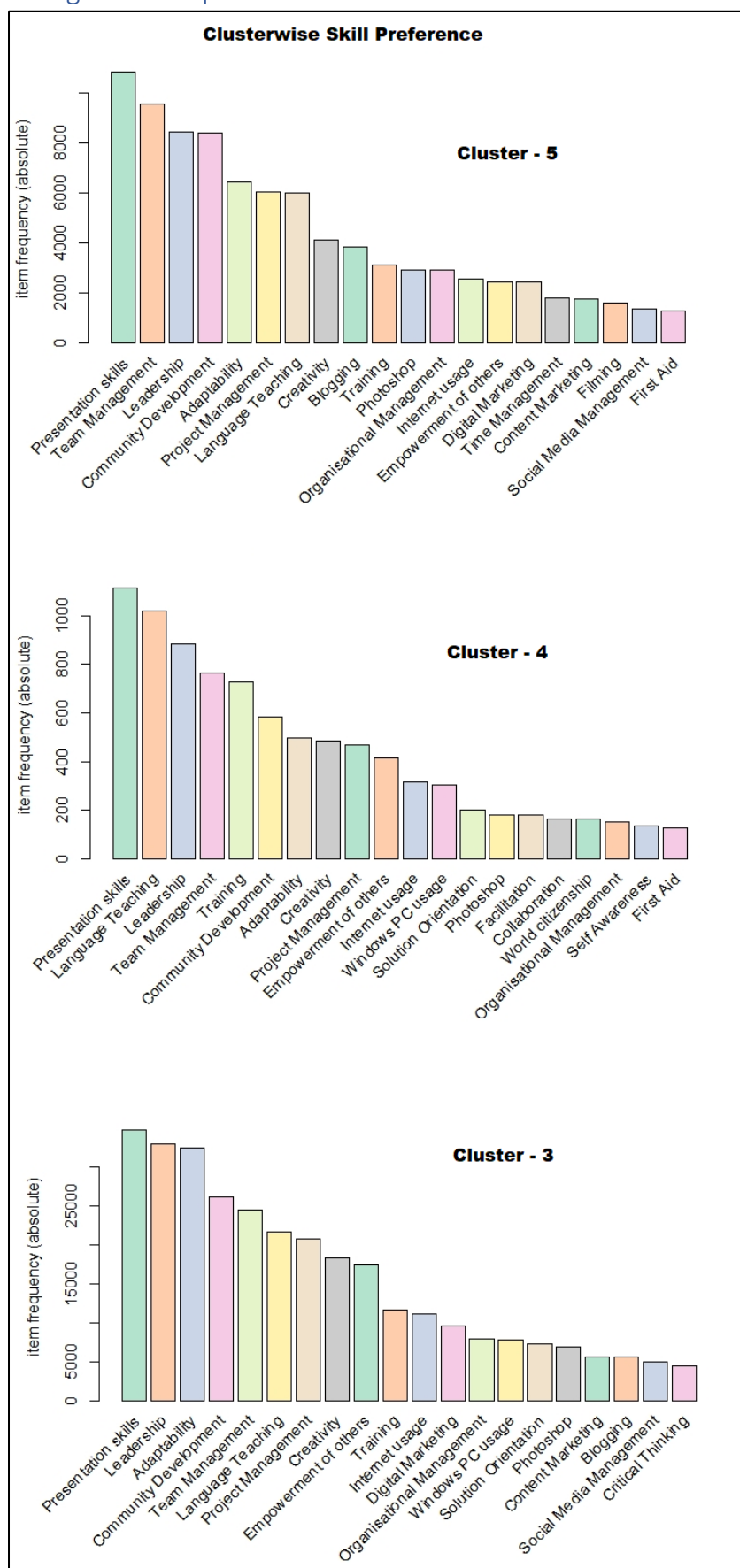


Figure 10.3: Most preferred skills among clusters of interest

10.8 Background required

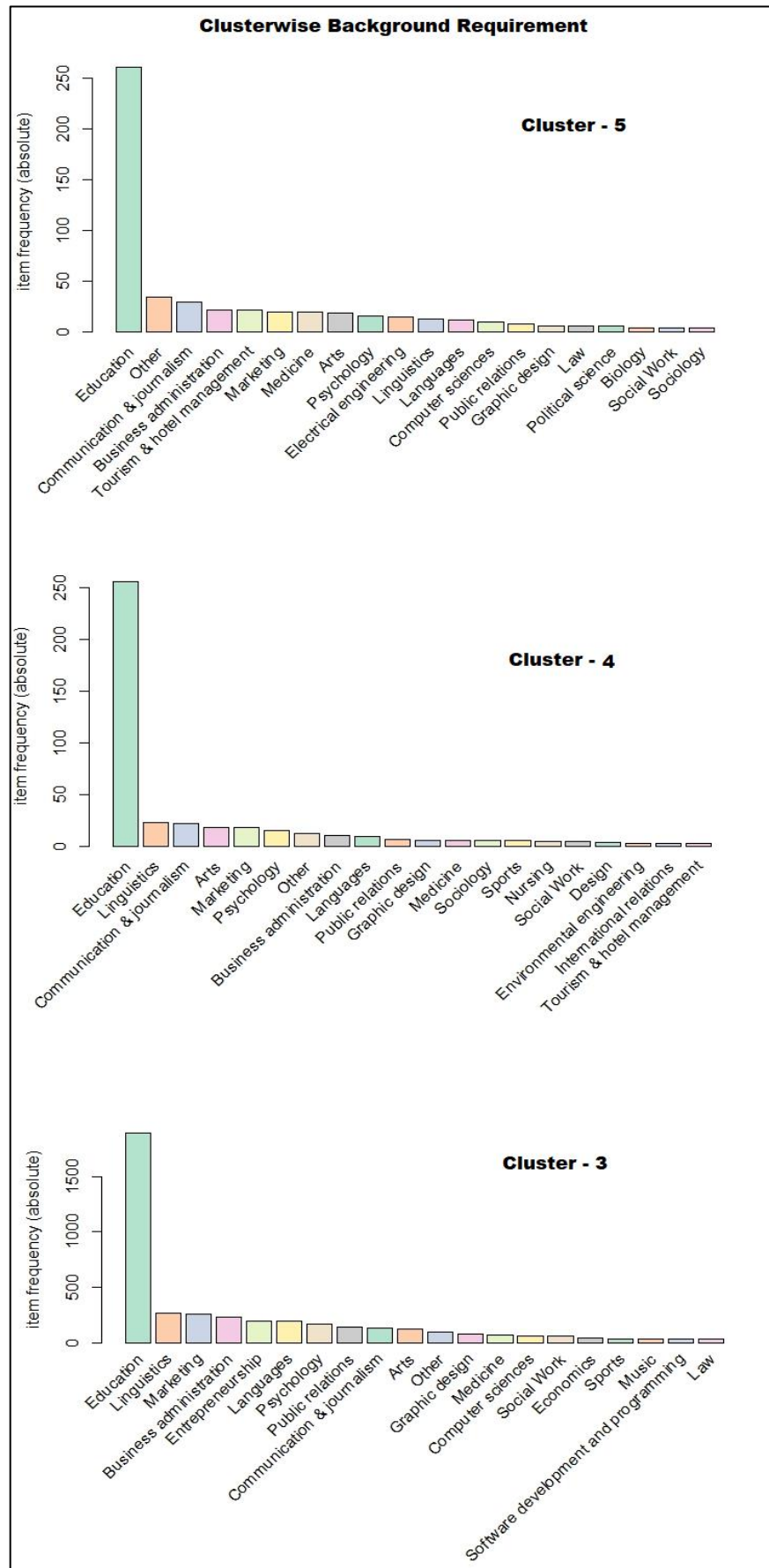


Figure 10.4: Most required backgrounds among clusters of interest

10.9 Background Preferred

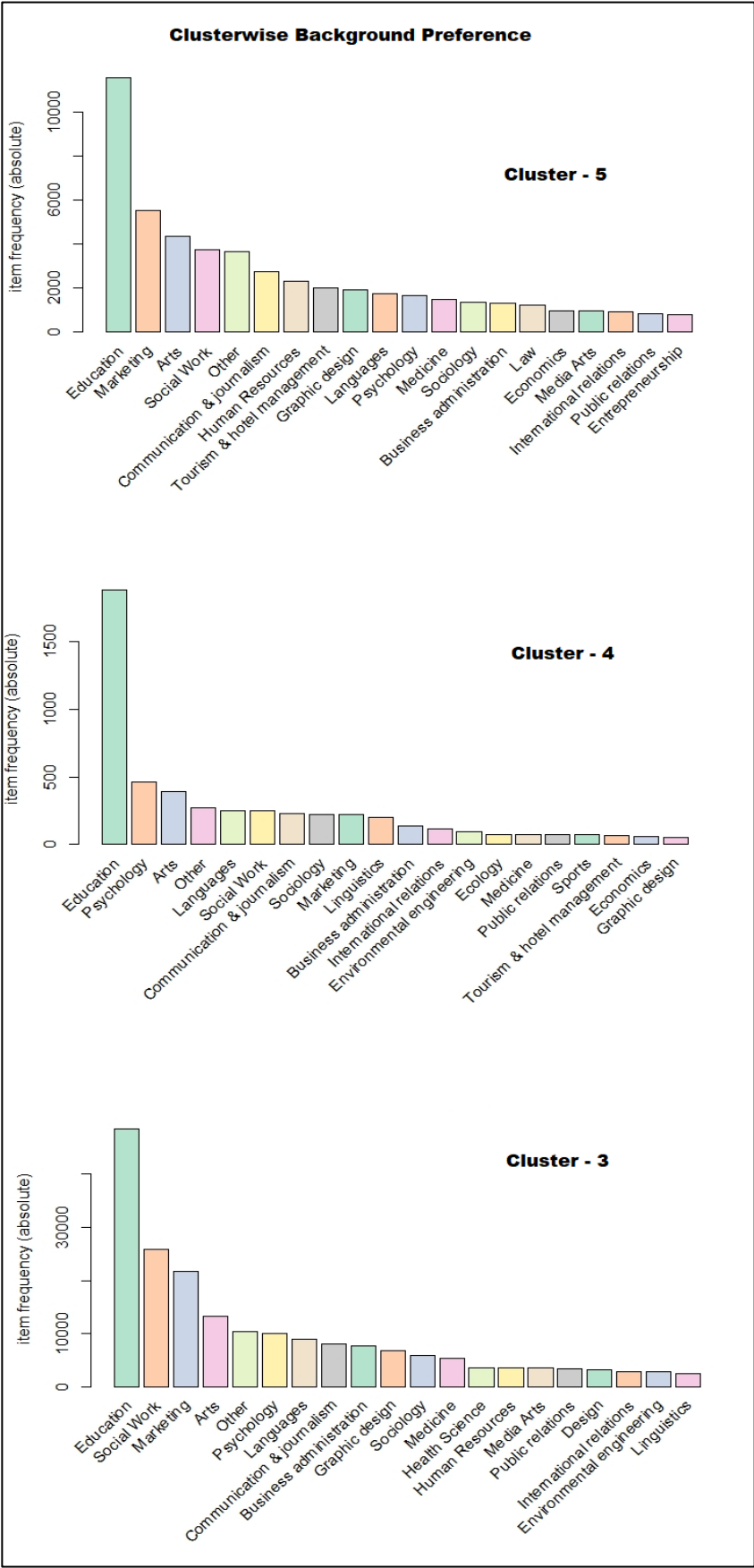


Figure 10.5: Most preferred backgrounds among clusters of interest

10.10 Higher Resolution Ranking Opportunities

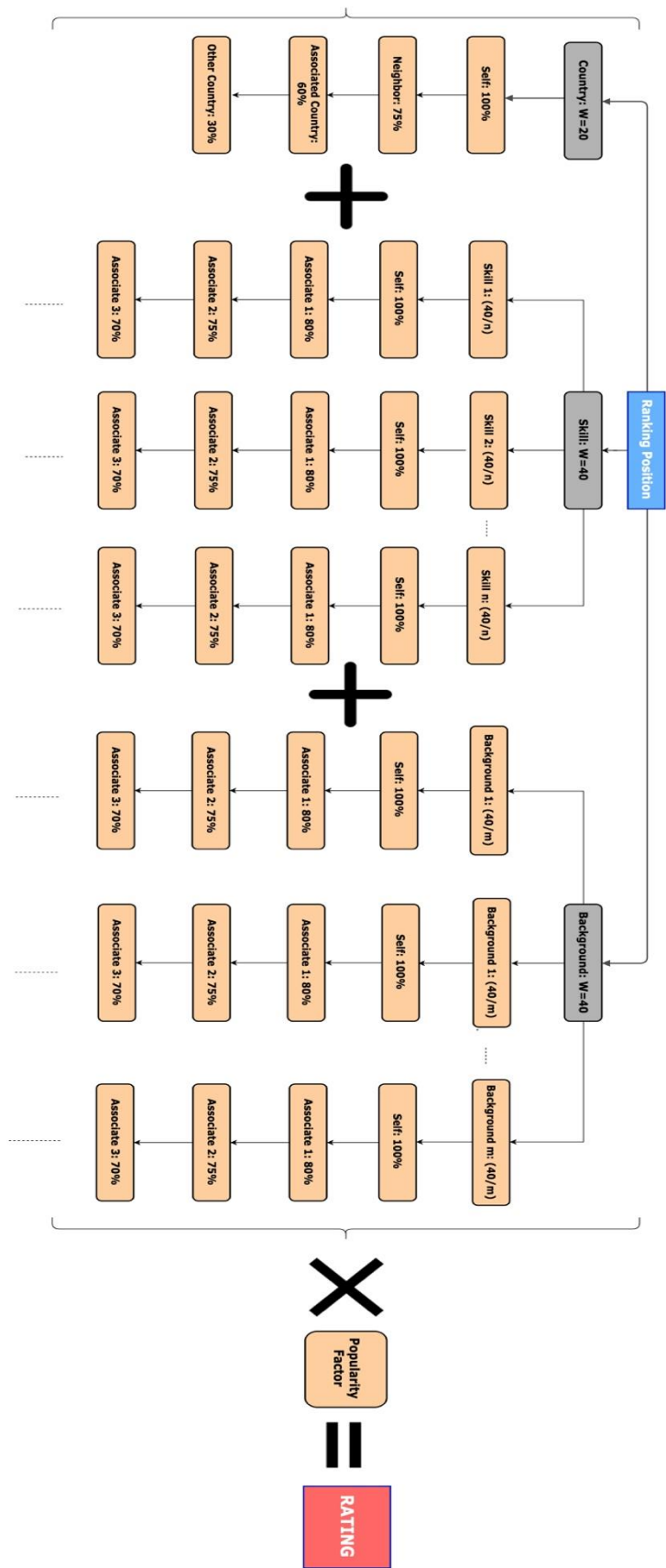


Figure 10.6: Weights Distribution for AHP Recommender

CANER IRFANOGLU

C: 1 (782) 234 80 85 | canerirfanoglu@gmail.com | [Linkedin](#)

CORE COMPETENCIES

Professional Interests	Machine learning, data analysis, algorithmic trading, decentralized apps
Personal Qualities	Energetic, driven, leadership, analytical thinker, well-rounded, team player
Technologies	Python, R, Tableau, Excel, VBA, SQL, Java, Git, Windows, Unix
Hobbies	Swimming, poker, ping-pong, yoga, meditation

RELEVANT PROJECTS



- Data Scientist** – AIESEC’s Digital Transformation Halifax, Nova Scotia | Jan 2019 – Ongoing
- Led a team of 5 for creating a recommendation system for International Business Analytics Challenge
 - Used R, Excel and Tableau for leveraging Unsupervised Machine Learning, Association Mining and Analytical Hierarchy Process techniques
- Data Analyst** – UNB Data Visualization Contest Halifax, Nova Scotia | Sep 2018 – Nov 2018
- Presented data driven analysis deriving information from New Brunswick’s Open Data Portal
 - Participated with a team of 4 and used Tableau to prepare an infographic
- Software Developer** – Jeffrey Trading Bot Istanbul, Turkey | Jan 2018 – Sep 2018
- Fully automated crypto-currency trading bot that can make long and short trades in multiple markets
 - Individual project coded in Python and R. Integrated with 2 major crypto-currency exchange API’s
- Data Science Student** – Data Science Specialization Irvine, California | Aug 2016 – Apr 2017
- Completed Coursera’s Data Science Specialization and DataCamp’s Data Scientist with Python
 - Shiny Country GDP App – Developing data products project - [CountryGDP](#)
 - Rpubs Quantitative Analysis Projects - [Rpubs](#)

EDUCATION

Saint Mary’s University
M.Sc. Computing and Data Analytics | Halifax NS, Canada (2018-2019)
Southern States University
M.B.A. | Newport Beach, CA, USA (2015-2017)
METU
B.S. Chemical Engineering | Ankara, Turkey (2007-2013)

CERTIFICATION

Algorithms - Stanford University
Machine Learning - Stanford University
Quantitative Analyst with R - Datacamp
Ultimate Hands-on Hadoop - Udemy
Complete Guide to Tensorflow -Udemy
Complete Java Certification - Udemy

CAREER OVERVIEW

- Data Analyst** – Real Mex Restaurants Orange County, California (Oct 2017 - Jan 2018)
- Automated routine data entry tasks by using Python and Excel VBA in collaboration with director of financial integration which resulted in improving the weekly work efficiency by 30%
 - Generated weekly, monthly and quarterly financial reports with Excel and Tableau, by gathered data from Microsoft SQL Server and Oracle Database
- Financial Analyst** – PDC Capital Group, LLC Costa Mesa, California (Feb 2016 - Dec 2016)
- Mapped financial data for multiple company projects using Quickbooks and Excel and reported the results to the CFO for assisting allocation of funds exceeding \$60m
 - Collaborated with legal, accounting and IT departments on a weekly basis

Sunil Padikar M

Highly skilled and accomplished software engineer with expertise in Software Development. Extremely passionate about programming with a strong desire to make benchmark contributions to the software industry and become an asset to the organization.

Work History

Clinical Logistics - Project Intern (12/2018 – Current) - Dartmouth, NS

Collaborated with all team members to build shipment schedule tracker using Asp.Net (MVC Architecture).
Boosted efficiency of sample tracking system using optimised algorithms.
Detected & diagnosed bugs. Rolled out patch sets for existing web applications.
Formulated requirements & provided continuous enhancement/support for all production work orders.
Involved in Technical design review and Code Review.

Skills used: ASP.NET, JAVA, jQuery, GIT, Visual Studio (VB.net)

Autoliv India Pvt. Ltd - Software Engineer (01/2014 - 08/2018) - Bengaluru, KA

Project 1: Heijunka Pattern for Production levelling – Autoliv being the largest automotive safety supplier required a technique for development of production efficiency.

- Architected & Implemented Heijunka single-handedly.
- Enabled production to efficiently meet customer demands while avoiding batching.
- Achieved results such as reduced wastage in production, interpersonal processes and minimum inventories.
- Significantly decreased the production lead time by 25%.
- Highly appreciated & Awarded for minimizing the capital costs by 6% through the implementation of Heijunka.

Skills used: VBA, Macros, Excel 2016.

Project 2: Design verification and Management (DVM) is a module for Enovia (PLM software) which enabled users to easily modify Bill of Materials for testing of various automobile parts.

- Enhanced the user experience of software testing.
- Involved in Software design, prototyping, investigating, coding, unit testing and Systems integration of DVM.
- Extensively handled performance issues & bug fixes before the given deadline.
- Created templates for effort-less storing of test results.
- Received 'Employee of the Quarter' award for Improving the testing process & Reducing the testing time by 10%.

✉ sunilpadikar77@gmail.com
☎ 902 818 9520
📍 South Park Street, Halifax, NS
[LinkedIn](#)

Skills

Technical Skills:

Java, JavaScript, JSP,
C, C++, VBA, VBS, Asp.Net
Python and R
HTML, CSS
Machine Learning Algorithms
MySQL, MSSQL, Mongo DB(Basic)
Troubleshooting and debugging

Soft Skills:

Problem-Solving
Decision making
Self-motivated
High Adaptability
Teamwork
Quick Learner

Competitions & Prizes

UNB Data visualization competition:

- **Won 3rd place.**
 - Assessed and analysed the accident data of Fredericton.
Designed a poster & presented to an audience to communicate the importance of public safety.
Involved in effective decision making.
Offered potential solutions with Clarity of Message.
Provided causality & explanation with visual emphasis.
- Skills used:** Tableau, Python, Excel 2016.

Skills used: Core Java, JavaScript, MQL, jQuery, SQL, HTML, CSS, Jsp, Git, VBA, Excel, XML.

Project 3: ESR Reporting (Engineering Service Request) is a tool which allows Monthly/Quarterly employee performance Report generation.

- Designed Data extractions and Reports generation modules.
- Involved in Data entry & data validation.
- Import and Export databases between multiple applications.
- Generated custom queries & extended capabilities of existing database used to track customer records.
- Troubleshoot and Maintain Data Inventory System.
- Extensive Documentation & Pivot Tables, Charts for Reports

Skills used: Visual Basic 6.0, MS Access, MS Excel, VBA, SQL Server, SharePoint.

Project 4: Time sheet management is a web-based harmonized logging tool which allows the employees to log & track their work, extract report & give detailed information to the respective manager.

- Involved in Design & Implementation of Database Schema.
- Coding and Testing of Application module by module
- Improved User experience by Creating Dashboards for viewing reports.
- Responsible for Project Execution and Deployment, customer-team coordination, User acceptance review & Documentation.
- Requirement Gathering and Implementation in agile way.

Skills used: MsSql, Asp.Net, jQuery, HTML, CSS, JavaScript, Ajax Call.

Academic Projects

- **Halifax Science Library:**

A vast library capable of storing the details of the Magazines, Articles, Authors and Sales of Magazines with a normalized Database design and a Rich UI to modify/update the details.

Skills used: PHP, HTML, CSS, SQL

Git Repo Link:

[Halifax Library](#)

- **Hotel Reservation System:**

A Hotel Reservation System through which any user can register and book a room. The user can pay the booking amount by adding a new card (new customer) or by selecting any available cards which interacts with a payment webservice API.

Skills used: ASP.NET, SOAP WS, GIT, jQuery

Git Repo Link:

[Hotel Reservation](#)

Education

Saint Mary's University Halifax, NS

09/2018 To 12/2019

M.Sc. in Computing & Data Analytics

Score: **4.15/4.3 GPA**

University Visvesvaraya College of

Engineering Bengaluru, KA

09/2009 To 07/2013

Bachelor of Engineering in Information Science and Engineering

Score: **72.5 %**

Certifications

Software for Embedded System C/C++

Udemy Machine Learning

Udemy Statistics: [Certification Link](#)

Hobbies

- Volleyball
- Swimming
- Badminton
- Cooking

VINAY GOVINDAN

Mobile: +1-9029893990 | Email ID: vinaygovindan94@gmail.com | LinkedIn: [Vinay Govindan](#)
Address: 22-8 Loyola Residence, 923 Robie Street, Halifax B3H 3C3

EXPERIENCE IN SUMMARY:

- Having 2 years of experience in the IT Industry broadly covering software development using PERL and Application operations in BSS side of a Telecommunications Network.
- Interested in learning new technologies and implement the best in projects.

SKILLS:

TECHNICAL SKILLS:

- **Programming Languages:**
Java(trained), Python(trained),
PERL(w.e), Shell(b.k)
- **Web Technology:** HTML5, CSS, JS
- **Web services:** SOAP, REST, XML
- **Development Tools:** Eclipse(trained),
PyCharm(h.o), Vim editor(w.e)
- **Database:** Oracle SQL(w.e),
MySQL(b.k), Mongo(b.k)
- **OS :** Windows (w.e), UNIX(w.e),
Linux(w.e)
- **Versioning tools:** Git (w.e)
- **Issue & Project tracking software:**
JIRA, BMC Remedy, HP QC, HP
ALM

SOFT SKILLS:

- Motivated self-learner.
- Can communicate a clear understanding of the subject at hand.
- Innovative & integrative thinker.
- Able to work independently, as a part of team, able to vaporize and grasp new things quickly
- Exhibit Leadership qualities and capable of handling multiple roles at the same time.

(*w.e- work experience, h.o - hands on, b.k - book knowledge)

PROJECT/EXPERIENCE PROFILE:

Company: **Infosys Limited**

Duration: **2.5 Years (Feb 2016 – August 2018)**

Projects: Multiple Developments & Application operations(AO) in parallel

Role: **Software Developer, Systems Engineer**

I've developed individual batch scripts, products, automation as a developer and worked across BSS layer of a major Telecom Service provide as a part of application operations.

Software Developer:

- Develop and provide support for the change requirements and enhancements for the Client.
- To provide New plans and products to the customers (Voice & Data) to adapt to the DOCSIS framework and enabling provisioning and billing of the Hybrid Fiber Connection.
- Design and Develop high quality software as deliverables on a PERL based stack with Oracle SQL as database and a web based CRM frontend.
- To enable Intercommunication between various stacks using REST/SOAP APIs thus reducing redundant components/processes from running individually in each stack.
- To provide solutions & hot fixes for production level exceptions and bugs.
- To create batch scripts and one off scripts using SHELL and PERL for mass migration of customer.
- To participate in triages and manage the builds through various Issue and project management software.
- To write and schedule CRON Jobs on a UNIX system using SHELL and PERL scripts.
- Use tools like Jenkins, GIT JIRA, HP QC, ALM as part of development cycle.
- Write Functions, Procedures and Triggers using ORACLE SQL

Systems Engineer:

- To monitor & provide solution to production issues and resolve the tickets across the BSS.
- To identify problems and provide workarounds and fixes for Mediation and Payment Gateways.
- To help automate long existing manual work using BASH scripts thus reducing latency.
- To monitor, perform health checks on systems and be proactive in identifying potential threats.

Intern:

- Trained in various Java Specifications including Hibernate, JSF
- Trained in Python and Oracle SQL
- Completed two projects as a part of the Internship:
 1. The Search Engine for blogs with JSoup, hibernate & Oracle SQL as the backend
 2. Service Assurance project for cellular networks with Python CGI Programming
- Completed Training as a High Performer at Infosys Limited

EDUCATIONAL QUALIFICATION:

- **Degree/Course : M.Sc. Computing and Data Science**
 - Saint Mary's University (In Progress)
 - GPA : 4.15
- **Degree/Course : B.E. Computer Science and Engineering**
 - Anna University (2016-2019)
 - CGPA : 7.16

ACCOLADES :

- Award for best automation initiatives
- Ad hoc awards and appreciation for in-depth analysis and development
- Award for best support in billing applications operations

TECHNOLOGY EXPERTISE

Programming Language(s): R, Python, SQL, Java, C#

DIVEN KUMAR SAMBHWANI

+1 902 430 7880 | [LinkedIn](#) | sambhwanidiven@gmail.com

Technologies and Skills: Statistical Analysis, Machine Learning, R Shiny, Google Analytics, Deep Learning, Financial Modelling, HTML, JS, Agile Methodology, Tableau, Linux, SPSS, A/B Testing, Business Intelligence

Familiarity: Hadoop and OLAP & OLTP datastores

WORK EXPERIENCE

Data Research Analyst, (Part Time) — Saint Mary's University, Halifax, Canada Oct 2018 - Present

- Collected and interpreted data using Excel and R
- Data wrangling and scripting which automates Fuzzy Mapping using R
- Reported the results back to the relevant members of the research

Data Analyst, (Co-op Term) — Hansa Cequity, Mumbai, India Jan - Jun 2018

- Developed understanding of architecture of health insurance, retail and automotive client (Business Logic Layer and database)
- Focused on project for automotive client for deep-dive analysis to understand behavior, uncover patterns and finding valuable insights that supported company to increase sales.
 - Analysis focused on: Customer feedback, Prospect, Service retention and Basket Analysis
- Created Ad hoc reports for the same and presented to clients.

Data Analyst, (Co-op Term) — GoSigmaway, Bangalore, India May - Jul 2017

- Implemented Black-Scholes formula and Margrabe formula for currency options using R for analyzing financial data that supported key business decisions which improved client engagement by 20%
- Used formulas on price derivatives including stock options, currency options, and exchange options, then used two-dimensional wiener processes on the patterns to see the currency fluctuation.

EDUCATION

M.Sc. - Computing and Data Analytics

Saint Mary's University, Halifax, NS — 2018 - Jan 2020 (Expected)

M.Sc. - Information Technology

DAIIT, India — 2016 – 2018

B.C.A - Computer Application

Ahmedabad University, India — 2013 – 2016

R Programming for Data Science

Python for Data Science

Machine Learning

Database Modelling and Design

Business Data Analysis

Intro. To language of SAS

Udemy

Udemy

Udemy

Udemy

Udemy

Udemy

CERTIFICATIONS

PROJECTS

Recommendation System

Developed an algorithm for recommendation system using Machine Learning Techniques for AIESEC in R

2019

[Restaurant Reviews](#)

Performed Sentimental Analysis of restaurant reviews using Machine Learning Techniques in python

2019

[Resume](#)

Developed an interactive resume using Tableau

2018

[Academic Projects](#)

2018

AWARDS AND ACHIEVEMENTS

President of Cultural Committee

2017

1st prize in System development project (Final Project) in bachelor's

2016

Represented state (Gujarat) in dance (V-fest)

2015

11. References

- Goldstein, D. G. (2014, August 21). *Profiting from the Long Tail*. Retrieved from <https://hbr.org/2006/06/profitting-from-the-long-tail>
- M. Dhanabhakym, Dr. M. Punithavalli, D. (2011). *A Survey on Data Mining Algorithm for Market Basket Analysis*. *Global Journal Of Computer Science And Technology*, . Retrieved from <https://computerresearch.org/index.php/computer/article/view/788>
- R. Handfield et al. *European Journal of Operational Research* 141 (2002) 70–87
- Saaty, T.L. (1980) *The Analytic Hierarchy Process*. McGraw-Hill, New York