ADVANCED ANALYTICS

MOON KARMAKAR_40389123

2023-04-04

1. Introduction: -

The topic of the assignment is to inclination of healthcare professionals such as clinicians and nurses towards emerging technologies holds great significance for policy formulation and decision-makers. We have a personalized section of panel data from a larger dataset containing an evaluation study on the technology inclination of healthcare employees in the United States. This dataset includes all these following variables: -

From this dataset we must find some of the insights

```
library('ggplot2')
library('corrplot')

## corrplot 0.92 loaded

library('dplyr')

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library('ggplot2')
library('corrplot')
library('readxl')
library('dplyr')
library('caret')

## Loading required package: lattice

library('lattice')
library('rpart')
library('rpart.plot')
```

2. Methodology: -

Data Loading: -

Loading the Dataset for Preliminary Analysis Read the data and ignore the first column because it is of no use for our visualization.

Checking null values:

Checking for null values is important in data analysis because null values represent missing or undefined data. If null values are not handled properly, they can result in errors and inaccuracies in data analysis, and can potentially lead to incorrect conclusions being drawn.

As we are seeing we do not have any null values in any of the column.

Checking for Zeroes:

Missing data- Zero values in a dataset can indicate missing data. It is important to identify missing data because it can affect the accuracy and validity of any analyses performed on the dataset.

Data quality: Zero values can also indicate data quality issues such as errors in data collection or data entry. Identifying these issues early on can help prevent potential problems later on.

In column name “final_grad_year,” “rank”, “global.rank” we have 0 values.

```
# Loading the Dataset and Preliminary Analysis
# Read the data and ignore the first column
data <- read.csv("C:/Users/moon/dataset-four_states- WA - TX - IA - NH -
student 90  Moon.csv")
Init_ds_df <- as.data.frame(data)
attach(data)
head(data)

##   X.1  X         ID State max_tech min_tech median_tech mean_tech
## 1  40 40 1003002379    WA 2013.417 1998.853    2007.835  2007.114
## 2  41 41 1003002627    TX 2010.503 1997.553    2008.892  2005.314
## 3  43 43 1003002742    IA 2012.043 2001.116    2009.329  2007.939
## 4  53 53 1003003153    WA 2004.374 2000.531    2004.215  2003.040
## 5  54 54 1003003153    WA 2004.374 2000.531    2004.215  2003.040
## 6  56 56 1003003609    TX 2012.434 1998.853    2008.332  2007.431
##               final_primary_speciality final_grad_year final_gender
## 1                    INTERNAL MEDICINE            2003            F
## 2                OBSTETRICS/GYNECOLOGY            2001            F
## 3 PHYSICAL MEDICINE AND REHABILITATION            2003            M
## 4                          DERMATOLOGY            2006            F
## 5                          DERMATOLOGY            2006            F
## 6                           NEPHROLOGY            1999            M
##                                         final_medical_school Rank
Global.Rank
## 1                                                     OTHER    0
0
## 2                                                     OTHER    0
0
```

```
## 3 MICHIGAN STATE UNIVERSITY COLLEGE OF OSTEOPATHIC MEDICINE   70
343
## 4   UNIVERSITY OF ILLINOIS AT CHICAGO HEALTH SCIENCE CENTER   50
208
## 5   UNIVERSITY OF ILLINOIS AT CHICAGO HEALTH SCIENCE CENTER  154
673
## 6                                                     OTHER    0
0
##    accuracy
## 1 0.5500000
## 2 0.1428571
## 3 0.3571429
## 4 1.0000000
## 5 1.0000000
## 6 0.7857143
```

```
View(data)
```

```
class(data)
```

```
## [1] "data.frame"
```

```
str(data)
```

```
## 'data.frame':    109870 obs. of  15 variables:
##  $ X.1                     : int  40 41 43 53 54 56 57 61 63 66 ...
##  $ X                       : int  40 41 43 53 54 56 57 61 63 66 ...
##  $ ID                      : int  1003002379 1003002627 1003002742
1003003153 1003003153 1003003609 1003003633 1003003963 1003004185 1003004490
...
##  $ State                   : chr  "WA" "TX" "IA" "WA" ...
##  $ max_tech                : num  2013 2011 2012 2004 2004 ...
##  $ min_tech                : num  1999 1998 2001 2001 2001 ...
##  $ median_tech             : num  2008 2009 2009 2004 2004 ...
##  $ mean_tech               : num  2007 2005 2008 2003 2003 ...
##  $ final_primary_speciality: chr  "INTERNAL MEDICINE"
"OBSTETRICS/GYNECOLOGY" "PHYSICAL MEDICINE AND REHABILITATION" "DERMATOLOGY"
...
##  $ final_grad_year         : int  2003 2001 2003 2006 2006 1999 0 0 0 0
...
##  $ final_gender            : chr  "F" "F" "M" "F" ...
##  $ final_medical_school    : chr  "OTHER" "OTHER" "MICHIGAN STATE
UNIVERSITY COLLEGE OF OSTEOPATHIC MEDICINE" "UNIVERSITY OF ILLINOIS AT
CHICAGO HEALTH SCIENCE CENTER" ...
##  $ Rank                    : int  0 0 70 50 154 0 0 0 0 0 ...
##  $ Global.Rank             : int  0 0 343 208 673 0 0 0 0 0 ...
##  $ accuracy                : num  0.55 0.143 0.357 1 1 ...
```

```
summary(data)
```

```
##       X.1               X                 ID               State
##  Min.   :    40   Min.   :    40   Min.   :1.003e+09   Length:109870
##  1st Qu.:235988   1st Qu.:235988   1st Qu.:1.255e+09   Class :character
##  Median :471313   Median :471313   Median :1.509e+09   Mode  :character
##  Mean   :469727   Mean   :469727   Mean   :1.503e+09
##  3rd Qu.:703514   3rd Qu.:703514   3rd Qu.:1.750e+09
##  Max.   :934797   Max.   :934797   Max.   :1.993e+09
##     max_tech       min_tech      median_tech     mean_tech
##  Min.   :1978   Min.   :1960   Min.   :1978   Min.   :1978
##  1st Qu.:2010   1st Qu.:1994   1st Qu.:2006   1st Qu.:2006
##  Median :2012   Median :1999   Median :2008   Median :2007
##  Mean   :2011   Mean   :1999   Mean   :2007   Mean   :2006
##  3rd Qu.:2013   3rd Qu.:2004   3rd Qu.:2009   3rd Qu.:2008
##  Max.   :2015   Max.   :2014   Max.   :2014   Max.   :2014
##  final_primary_speciality final_grad_year final_gender
##  Length:109870            Min.   :   0    Length:109870
##  Class :character         1st Qu.:   0    Class :character
##  Mode  :character         Median :1985    Mode  :character
##                           Mean   :1259
##                           3rd Qu.:1998
##                           Max.   :2014
##  final_medical_school      Rank         Global.Rank       accuracy
##  Length:109870        Min.   :  0.00   Min.   :  0.0   Min.   :-9.0000
##  Class :character     1st Qu.:  0.00   1st Qu.:  0.0   1st Qu.: 0.3824
##  Mode  :character     Median :  0.00   Median :  0.0   Median : 0.5556
##                       Mean   : 53.08   Mean   :228.2   Mean   : 0.4915
##                       3rd Qu.:115.00   3rd Qu.:559.0   3rd Qu.: 0.7778
##                       Max.   :272.00   Max.   :895.0   Max.   : 1.0000

data_1 <- data
```

We are dropping null and 0 values rows because it may impact our analysis, visualisation, and modelling. Sometime the zero values represent missing or incomplete data, dropping those rows can help to ensure that the remaining data is complete and accurate. This can be especially important in cases where missing data could introduce bias or affect the validity of the analysis.

Here we are creating our 1st dataset as mentioned in question

```
# Count the number of unique values
unique_values <- unique(data_1$State)
num_unique_values <- length(unique_values)
print(num_unique_values)    # We have four datasets (IA, NH, TX, WA)

## [1] 4

# Count the number of occurrences of unique value
occurrences <- table(data_1$State)
print(occurrences)
```

```
##
##    IA    NH    TX    WA
##  8378  4727 76224 20541

# IA    NH    TX    WA
# 8378  4727 76224 20541

# We will create our first data with TX because it has the maximum number of
records/data
```

The number of occurrences is counted for unique value. In my dataset we have four datasets (IA, NH, TX, WA), we will create our first data with TX because it has maximum amount of records/data.

The number of occurrences are counted for each value and sorted the results for final_primary_speciality. Getting top 5 occurrence values as mentioned in assignment. top_5_values <- names(occurrences_1) [1:6] print(top_5_values).

```
# Count the number of unique values
unique_values <- unique(data_1$final_primary_speciality)
num_unique_values <- length(unique_values)
print(num_unique_values)    # We have 77 primary specialties

## [1] 77

# Count the number of occurrences of unique values
occurrences <- table(data_1$final_primary_speciality)
print(occurrences)

##
##
##                                                    40479
##                                       ADDICTION MEDICINE
##                                                       15
##                    ADULT CONGENITAL HEART DISEASE (ACHD)
##                                                        2
##         ADVANCED HEART FAILURE AND TRANSPLANT CARDIOLOGY
##                                                       80
##                                       ALLERGY/IMMUNOLOGY
##                                                      564
##                                           ANESTHESIOLOGY
##                                                      635
##                                CARDIAC ELECTROPHYSIOLOGY
##                                                      338
##                                          CARDIAC SURGERY
##                                                      174
##                      CARDIOVASCULAR DISEASE (CARDIOLOGY)
##                                                     2982
##                CERTIFIED CLINICAL NURSE SPECIALIST (CNS)
##                                                       71
##                                  CERTIFIED NURSE MIDWIFE
##                                                      10
##                               CERTIFIED NURSE MIDWIFE (CNM)
##                                                       2
##                          CERTIFIED REGISTERED NURSE ANESTHETIST
##                                                      14
##                   CERTIFIED REGISTERED NURSE ANESTHETIST (CRNA)
##                                                       1
##                                            CHIROPRACTIC
##                                                       6
##                               CLINICAL NURSE SPECIALIST
##                                                       3
##                                  CLINICAL SOCIAL WORKER
##                                                       1
##                         COLORECTAL SURGERY (PROCTOLOGY)
##                                                     208
##                            CRITICAL CARE (INTENSIVISTS)
##                                                     282
##                                             DERMATOLOGY
##                                                    1958
##                                    DIAGNOSTIC RADIOLOGY
##                                                     264
##                                      EMERGENCY MEDICINE
##                                                    1639
##                                           ENDOCRINOLOGY
##                                                     802
##                                         FAMILY MEDICINE
##                                                   14000
##                                         FAMILY PRACTICE
##                                                     328
##                                        GASTROENTEROLOGY
##                                                    1946
##                                        GENERAL PRACTICE
##                                                     339
##                                         GENERAL SURGERY
##                                                    1436
##                                      GERIATRIC MEDICINE
##                                                     186
##                                    GERIATRIC PSYCHIATRY
##                                                      27
##                                  GYNECOLOGICAL ONCOLOGY
##                                                     117
##                                            HAND SURGERY
##                                                     160
##                                              HEMATOLOGY
##                                                      85
##                                     HEMATOLOGY/ONCOLOGY
##                                                     977
## HEMATOPOIETIC CELL TRANSPLANTATION AND CELLULAR THERAPY
##                                                      10
##                                 HOSPICE/PALLIATIVE CARE
##                                                     154
##                                             HOSPITALIST
##                                                     654
##                                      INFECTIOUS DISEASE
##                                                     535
##                                       INTERNAL MEDICINE
##                                                    9559
##                               INTERVENTIONAL CARDIOLOGY
##                                                     177
##                          INTERVENTIONAL PAIN MANAGEMENT
##                                                     313
##                                INTERVENTIONAL RADIOLOGY
##                                                      48
##                                   MAXILLOFACIAL SURGERY
##                                                     110
##                                        MEDICAL ONCOLOGY
##                                                     560
##                                              NEPHROLOGY
##                                                    1314
##                                               NEUROLOGY
##                                                    1590
##                                         NEUROPSYCHIATRY
##                                                       1
##                                            NEUROSURGERY
##                                                     496
##                                        NUCLEAR MEDICINE
##                                                       1
##                                      NURSE PRACTITIONER
##                                                    3293
##                                   OBSTETRICS/GYNECOLOGY
##                                                    3395
##                                           OPHTHALMOLOGY
##                                                    2716
##                                               OPTOMETRY
##                                                    1583
##                                            ORAL SURGERY
##                                                      62
##                                      ORTHOPEDIC SURGERY
##                                                    3154
##                       OSTEOPATHIC MANIPULATIVE MEDICINE
##                                                      14
##                                          OTOLARYNGOLOGY
##                                                    1714
##                                         PAIN MANAGEMENT
##                                                     216
##                                               PATHOLOGY
##                                                      14
##                                      PEDIATRIC MEDICINE
##                                                     109
##                             PERIPHERAL VASCULAR DISEASE
##                                                       5
##                    PHYSICAL MEDICINE AND REHABILITATION
##                                                     834
##                                        PHYSICAL THERAPY
##                                                       7
##                                     PHYSICIAN ASSISTANT
##                                                      93
##                      PLASTIC AND RECONSTRUCTIVE SURGERY
##                                                     303
##                                                PODIATRY
##                                                     795
##                                   PREVENTATIVE MEDICINE
##                                                      10
##                                              PSYCHIATRY
##                                                    1560
##                                       PULMONARY DISEASE
##                                                    1082
##                                      RADIATION ONCOLOGY
##                                                     356
##                                            RHEUMATOLOGY
##                                                     601
##                               SLEEP LABORATORY/MEDICINE
##                                                      26
##                                         SPORTS MEDICINE
##                                                      22
##                                       SURGICAL ONCOLOGY
##                                                      39
##                                        THORACIC SURGERY
##                                                     138
##                                                 UROLOGY
##                                                    1665
##                                        VASCULAR SURGERY
##                                                     411
```

```
# We have 40k approx null values in this column
# So we remove null values from that column and find the top 5 occurrence

# Count the number of occurrences of each value and sort the results
occurrences_1 <- sort(table(data_1$final_primary_speciality), decreasing =
TRUE)

# Get the top 5 occurrence values
top_5_values <- names(occurrences_1)[1:6]
print(top_5_values)
```

```
## [1] ""                      "FAMILY MEDICINE"       "INTERNAL MEDICINE"
## [4] "OBSTETRICS/GYNECOLOGY" "NURSE PRACTITIONER"    "ORTHOPEDIC SURGERY"
```

```
#  "FAMILY MEDICINE"       "INTERNAL MEDICINE"     "OBSTETRICS/GYNECOLOGY"
#  "NURSE PRACTITIONER",      "ORTHOPEDIC SURGERY"
```

We have these top 5 primary specialities “FAMILY MEDICINE”, “INTERNAL MEDICINE” “OBSTETRICS/GYNECOLOGY”, “NURSE PRACTITIONER”, “ORTHOPEDIC SURGERY”

The number of rows are selected where score is greater than 0.5.

Here we created our first dataset with only one state TX because it has the maximum amount of data in comparison to others, and after that we have selected top 5 primary specialities and after that we have selected only those data who have accuracy more than 0.5. we are referring this data from data_1.

```
# We are creating data with 1(TX) states
# Select rows where final_primary_speciality is Family Medicine or Pediatrics
selected_dataset_1 <- data_1[data_1$final_primary_speciality %in% c("Family
Medicine", "INTERNAL MEDICINE","OBSTETRICS/GYNECOLOGY","NURSE
PRACTITIONER","ORTHOPEDIC SURGERY"), ]
selected_dataset_1 <- selected_dataset_1[selected_dataset_1$State %in%
c("TX"), ]
# Select rows where score is greater than 0.5
selected_data <- selected_dataset_1[selected_dataset_1$accuracy > 0.5, ]
#removing null and 0 values from data
selected_data <- subset(selected_data, Rank != 0)

#reseting index
selected_data <- data.frame(selected_data, row.names = NULL)

class(selected_data)
```

```
## [1] "data.frame"
```

```
str(selected_data)
```

```
## 'data.frame':    4837 obs. of  15 variables:
##  $ X.1                     : int  264 265 266 267 268 269 270 3020 3049
3164 ...
##  $ X                       : int  264 265 266 267 268 269 270 3020 3049
3164 ...
##  $ ID                      : int  1003014366 1003014366 1003014366
1003014366 1003014366 1003014366 1003014366 1003804428 1003805292 1003807801
...
##  $ State                   : chr  "TX" "TX" "TX" "TX" ...
##  $ max_tech                : num  2013 2013 2013 2013 2013 ...
##  $ min_tech                : num  2003 2003 2003 2003 2003 ...
##  $ median_tech             : num  2007 2007 2007 2007 2007 ...
##  $ mean_tech               : num  2008 2008 2008 2008 2008 ...
##  $ final_primary_speciality: chr  "NURSE PRACTITIONER" "NURSE
PRACTITIONER" "NURSE PRACTITIONER" "NURSE PRACTITIONER" ...
##  $ final_grad_year         : int  2007 2007 2007 2007 2007 2007 2007 1992
1990 1997 ...
##  $ final_gender            : chr  "F" "F" "F" "F" ...
##  $ final_medical_school    : chr  "UNIVERSITY OF TEXAS MEDICAL BRANCH AT
GALVESTON" "UNIVERSITY OF TEXAS MEDICAL BRANCH AT GALVESTON" "UNIVERSITY OF
```

```
TEXAS MEDICAL BRANCH AT GALVESTON" "UNIVERSITY OF TEXAS MEDICAL BRANCH AT
GALVESTON" ...
##  $ Rank                    : int  61 66 81 115 157 169 188 29 97 61 ...
##  $ Global.Rank             : int  301 320 403 559 681 698 733 97 489 301
...
##  $ accuracy                : num  0.667 0.667 0.667 0.667 0.667 ...
```

```
summary(selected_data)
```

```
##       X.1               X                ID               State          
##  Min.   :   264   Min.   :   264   Min.   :1.003e+09   Length:4837       
##  1st Qu.:245726   1st Qu.:245726   1st Qu.:1.265e+09   Class :character  
##  Median :496245   Median :496245   Median :1.538e+09   Mode  :character  
##  Mean   :482680   Mean   :482680   Mean   :1.517e+09                     
##  3rd Qu.:713786   3rd Qu.:713786   3rd Qu.:1.760e+09                     
##  Max.   :934181   Max.   :934181   Max.   :1.993e+09                     
##     max_tech       min_tech     median_tech     mean_tech   
##  Min.   :1994   Min.   :1978   Min.   :1994   Min.   :1994  
##  1st Qu.:2010   1st Qu.:1995   1st Qu.:2007   1st Qu.:2006  
##  Median :2012   Median :1999   Median :2008   Median :2007  
##  Mean   :2011   Mean   :1999   Mean   :2008   Mean   :2007  
##  3rd Qu.:2013   3rd Qu.:2000   3rd Qu.:2009   3rd Qu.:2008  
##  Max.   :2014   Max.   :2012   Max.   :2012   Max.   :2012  
##  final_primary_speciality final_grad_year final_gender      
##  Length:4837              Min.   :1907    Length:4837       
##  Class :character         1st Qu.:1989    Class :character  
##  Mode  :character         Median :1996    Mode  :character  
##                           Mean   :1996                      
##                           3rd Qu.:2004                      
##                           Max.   :2013                      
##  final_medical_school      Rank        Global.Rank     accuracy     
##  Length:4837           Min.   :  2.0   Min.   :  2   Min.   :0.5040  
##  Class :character      1st Qu.: 66.0   1st Qu.:320   1st Qu.:0.6000  
##  Mode  :character      Median :115.0   Median :559   Median :0.7273  
##                        Mean   :121.5   Mean   :526   Mean   :0.7636  
##                        3rd Qu.:169.0   3rd Qu.:698   3rd Qu.:1.0000  
##                        Max.   :272.0   Max.   :895   Max.   :1.0000  
```

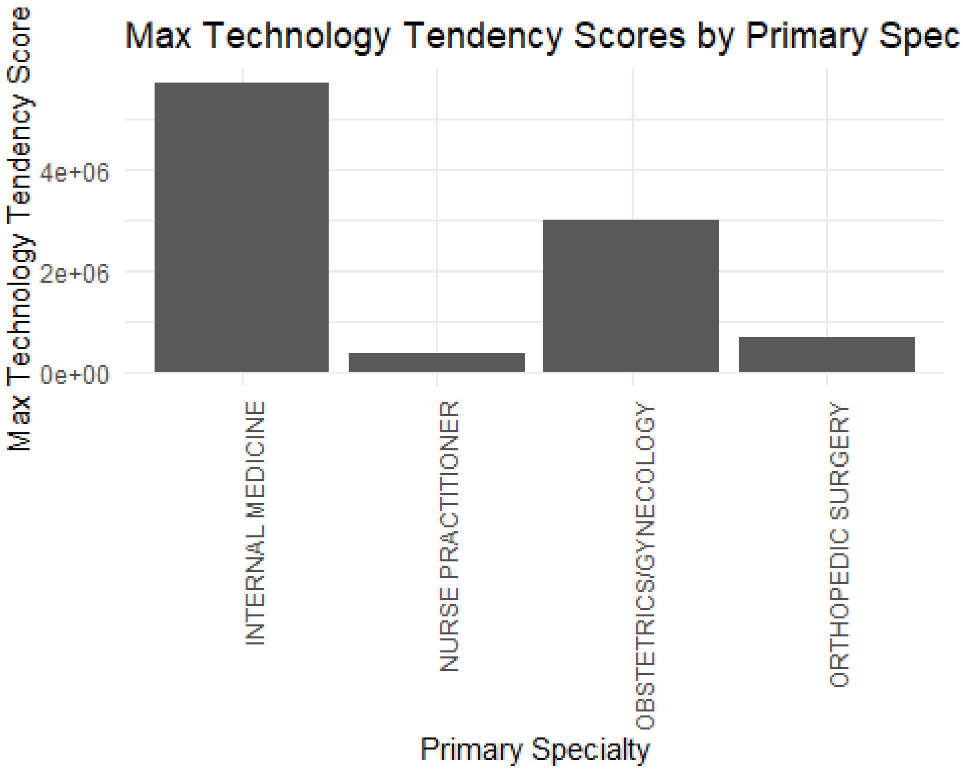
```
# Print the result
head(selected_data)
```

```
##   X.1   X         ID State max_tech min_tech median_tech mean_tech
## 1 264 264 1003014366    TX 2013.417 2003.033     2006.62   2007.69
## 2 265 265 1003014366    TX 2013.417 2003.033     2006.62   2007.69
## 3 266 266 1003014366    TX 2013.417 2003.033     2006.62   2007.69
## 4 267 267 1003014366    TX 2013.417 2003.033     2006.62   2007.69
## 5 268 268 1003014366    TX 2013.417 2003.033     2006.62   2007.69
## 6 269 269 1003014366    TX 2013.417 2003.033     2006.62   2007.69
##   final_primary_speciality final_grad_year final_gender
## 1       NURSE PRACTITIONER            2007            F
## 2       NURSE PRACTITIONER            2007            F
## 3       NURSE PRACTITIONER            2007            F
## 4       NURSE PRACTITIONER            2007            F
## 5       NURSE PRACTITIONER            2007            F
## 6       NURSE PRACTITIONER            2007            F
##                               final_medical_school Rank Global.Rank
accuracy
## 1 UNIVERSITY OF TEXAS MEDICAL BRANCH AT GALVESTON   61         301
0.6666667
## 2 UNIVERSITY OF TEXAS MEDICAL BRANCH AT GALVESTON   66         320
0.6666667
## 3 UNIVERSITY OF TEXAS MEDICAL BRANCH AT GALVESTON   81         403
0.6666667
## 4 UNIVERSITY OF TEXAS MEDICAL BRANCH AT GALVESTON  115         559
0.6666667
## 5 UNIVERSITY OF TEXAS MEDICAL BRANCH AT GALVESTON  157         681
0.6666667
## 6 UNIVERSITY OF TEXAS MEDICAL BRANCH AT GALVESTON  169         698
0.6666667
```

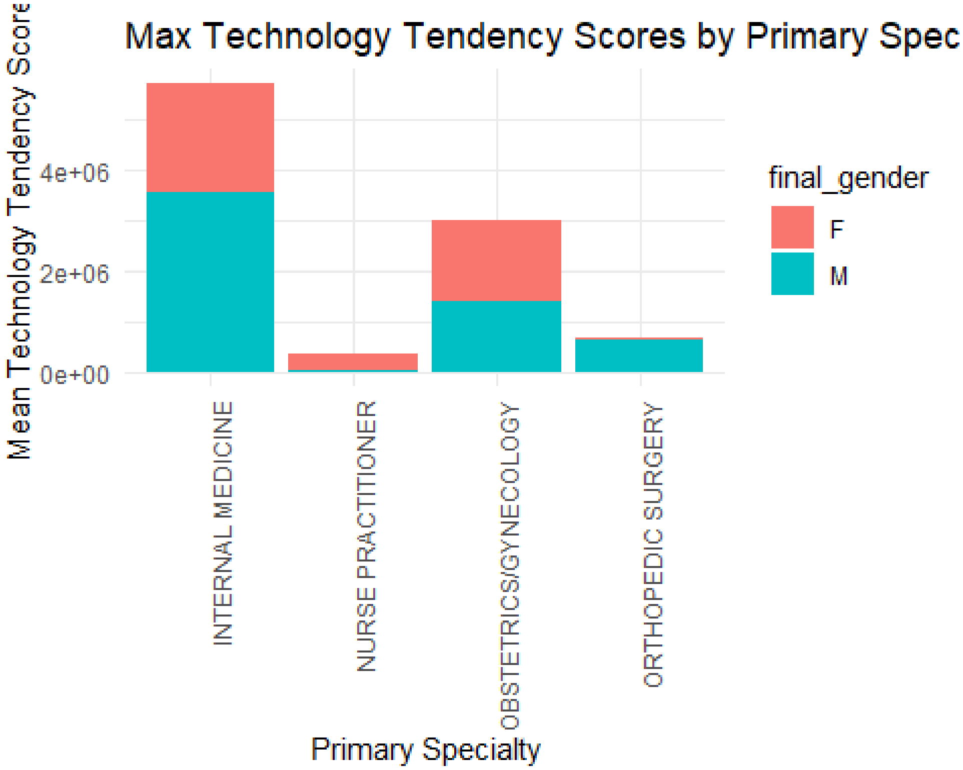
Q1. Do workers of a particular primary speciality (e.g., internal medicine) have more technology tendency than those workers with other primary speciality (e.g., general surgery)

Ans:Yes, workers of a particular primary speciality (internal medicine) have more technology tendency than those workers with other primary speciality (FAMILY MEDICINE”, “OBSTETRICS/GYNECOLOGY” ,“NURSE PRACTITIONER”, “ORTHOPEDIC SURGERY)

```
#Question 1



# group the data by primary specialty, gender, and calculate the mean
technology tendency score for each group
df_means <- selected_data %>%
  group_by(final_primary_speciality, final_gender) %>%
  summarise(sum_tech_tendency = sum(max_tech))

## `summarise()` has grouped output by 'final_primary_speciality'. You can
## override using the `.groups` argument.

# create a stacked bar chart of mean technology tendency scores for each
primary specialty and gender combination
ggplot(df_means, aes(x = final_primary_speciality, y = sum_tech_tendency)) +
  geom_col(position = "stack") +
  labs(title = "Max Technology Tendency Scores by Primary Specialty and
Gender",
       x = "Primary Specialty",
       y = "Max Technology Tendency Score") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



```
# create a stacked bar chart of mean technology tendency scores for each
primary specialty and gender combination
ggplot(df_means, aes(x = final_primary_speciality, y = sum_tech_tendency,
fill = final_gender)) +
  geom_col(position = "stack") +
  labs(title = "Max Technology Tendency Scores by Primary Specialty and
Gender",
       x = "Primary Specialty",
       y = "Mean Technology Tendency Score") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Q1.1 Is it dependent on the gender of the workers?

Ans: As shown above graph, yes it depends on the genders of the workers.

Q2.1

- New column is created based on the “rank” variable that specifies whether the employee graduated from a high-ranking or low-ranking institution.

- Data is grouped by rank category, and each group’s mean technology tendency score is calculated.

- Depending on the “global.rank” variable, a new column is added stating whether the employee graduated from a high or low rank school.

- Data is grouped by global rank category, and the maximum technology tendency score is calculated for each group.

- Using a t-test for both “rank” and “global. rank,” the mean technological inclination scores are compared for high and low rank schools.

```
# question-2
data<- selected_data
# add a new column that indicates whether the employee graduated from a high
or low rank school based on the "rank" variable
data$rank_category <- ifelse(data$Rank <= median(data$Rank), "Low Rank",
"High Rank")
```

```r
# group the data by rank category and calculate the mean technology tendency
score for each group
df_rank_means <- data %>%
  group_by(rank_category) %>%
  summarise(mean_tech_tendency = mean(max_tech))

# add a new column that indicates whether the employee graduated from a high
or low rank school based on the "global.rank" variable
data$global_rank_category <- ifelse(data$Global.Rank <=
median(data$Global.Rank), "Low Rank", "High Rank")

# group the data by global rank category and calculate the max technology
tendency score for each group
df_global_rank_means <- data %>%
  group_by(global_rank_category) %>%
  summarise(mean_tech_tendency = mean(max_tech))

# compare the mean technology tendency scores for high and low rank schools
using a t-test for both "rank" and "global.rank"
t.test(max_tech ~ rank_category, data = data)
```

```
##
##  Welch Two Sample t-test
##
## data:  max_tech by rank_category
## t = -0.72448, df = 4783.2, p-value = 0.4688
## alternative hypothesis: true difference in means between group High Rank
and group Low Rank is not equal to 0
## 95 percent confidence interval:
##  -0.21670954  0.09975924
## sample estimates:
## mean in group High Rank  mean in group Low Rank
##                2011.317                2011.375
```

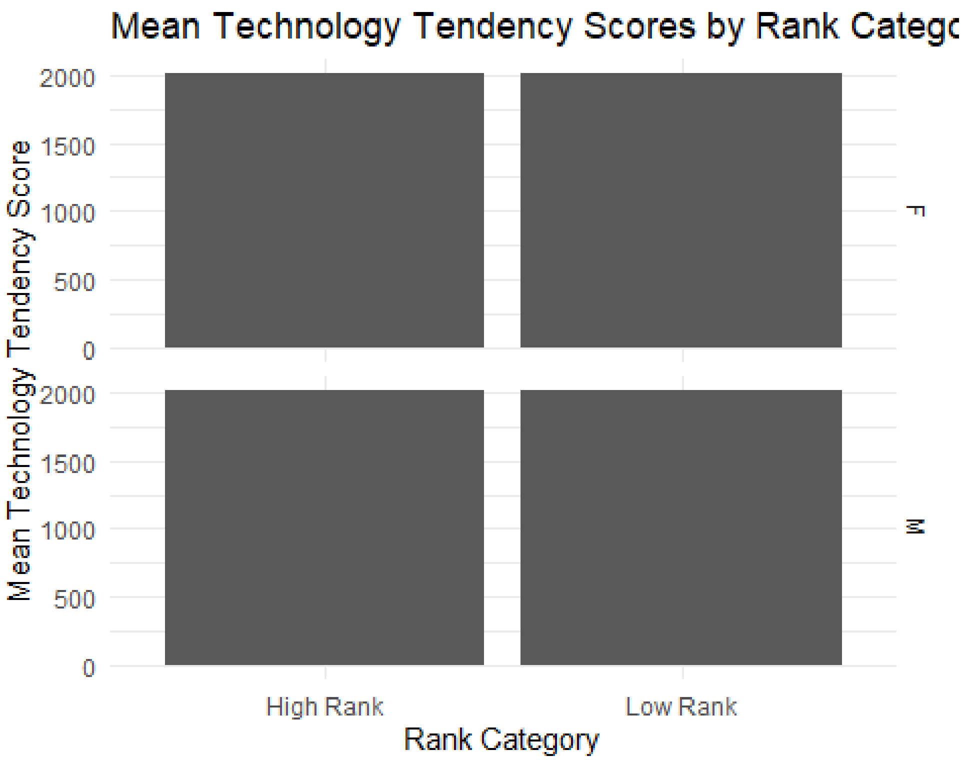
```r
t.test(max_tech ~ global_rank_category, data = data)
```

```
##
##  Welch Two Sample t-test
##
## data:  max_tech by global_rank_category
## t = -0.72448, df = 4783.2, p-value = 0.4688
## alternative hypothesis: true difference in means between group High Rank
and group Low Rank is not equal to 0
## 95 percent confidence interval:
##  -0.21670954  0.09975924
## sample estimates:
## mean in group High Rank  mean in group Low Rank
##                2011.317                2011.375
```

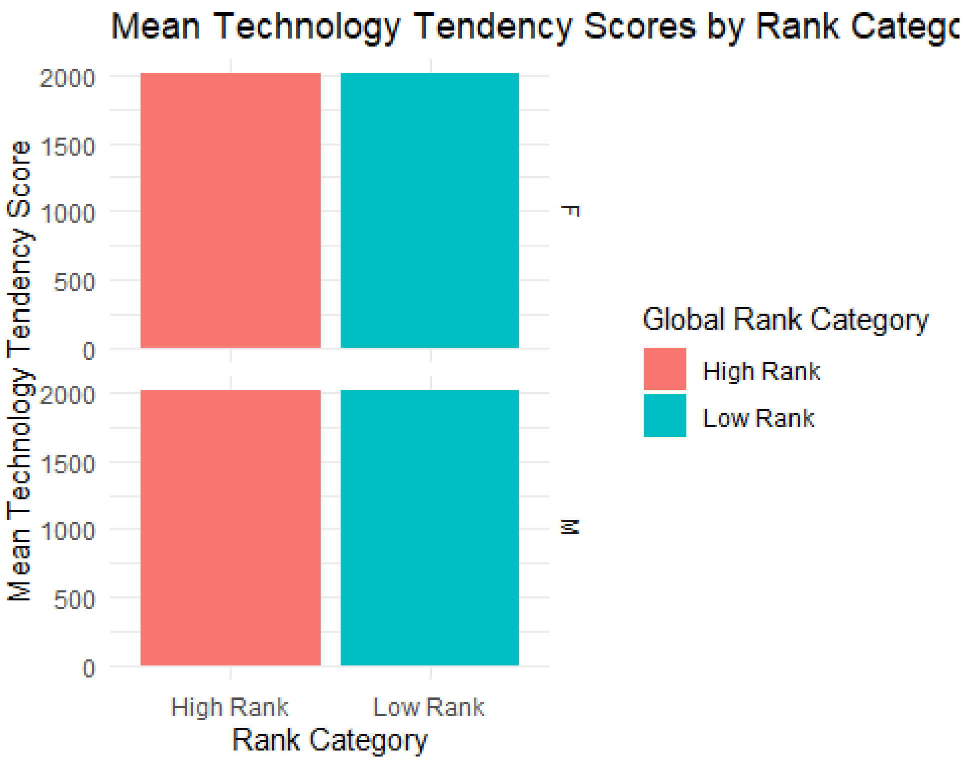
```r
# add new columns that indicate whether the employee graduated from a high or
low rank school based on the "rank" or "global.rank" variables
data$rank_category <- ifelse(data$Rank <= median(data$Rank), "Low Rank",
"High Rank")
data$global_rank_category <- ifelse(data$Global.Rank <=
median(data$Global.Rank), "Low Rank", "High Rank")



# group the data by rank category, global rank category, and gender, and
calculate the mean technology tendency score for each group
df_means <- data %>%
  group_by(rank_category, global_rank_category, final_gender) %>%
  summarise(mean_tech_tendency = mean(max_tech))
```

```
## `summarise()` has grouped output by 'rank_category',
'global_rank_category'.
## You can override using the `.groups` argument.
```

```r
# create a stacked bar chart of mean technology tendency scores for high and
low rank schools, by global rank category and gender
ggplot(df_means, aes(x = rank_category, y = mean_tech_tendency)) +
  geom_col(position = "stack") +
  facet_grid(rows = vars(final_gender)) +
  labs(title = "Mean Technology Tendency Scores by Rank Category, Global Rank
Category, and Gender",
       x = "Rank Category",
       y = "Mean Technology Tendency Score",
       fill = "Global Rank Category") +
  theme_minimal()
```



```
ggplot(df_means, aes(x = rank_category, y = mean_tech_tendency, fill =
global_rank_category)) +
  geom_col(position = "stack") +
  facet_grid(rows = vars(final_gender)) +
  labs(title = "Mean Technology Tendency Scores by Rank Category, Global Rank
Category, and Gender",
       x = "Rank Category",
       y = "Mean Technology Tendency Score",
       fill = "Global Rank Category") +
  theme_minimal()
```

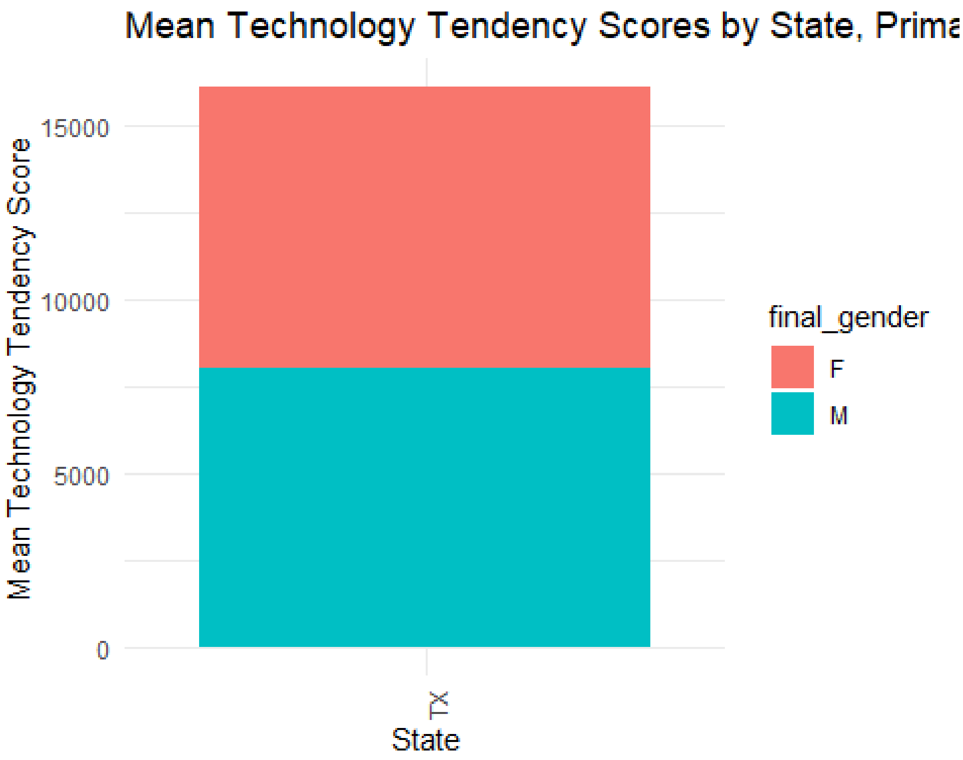


• Additional columns that, based on the “rank” or “global.rank” variables are added, show whether the employee graduated from a high-ranking or low-ranking institution.

• The data is grouped by “rank” category, “global.rank” category, and “gender”. The mean technology tendency score is then calculated for each group.

• Based on global rank category, gender, and mean technological inclination scores for high and low rank schools, a stacked bar chart is generated.

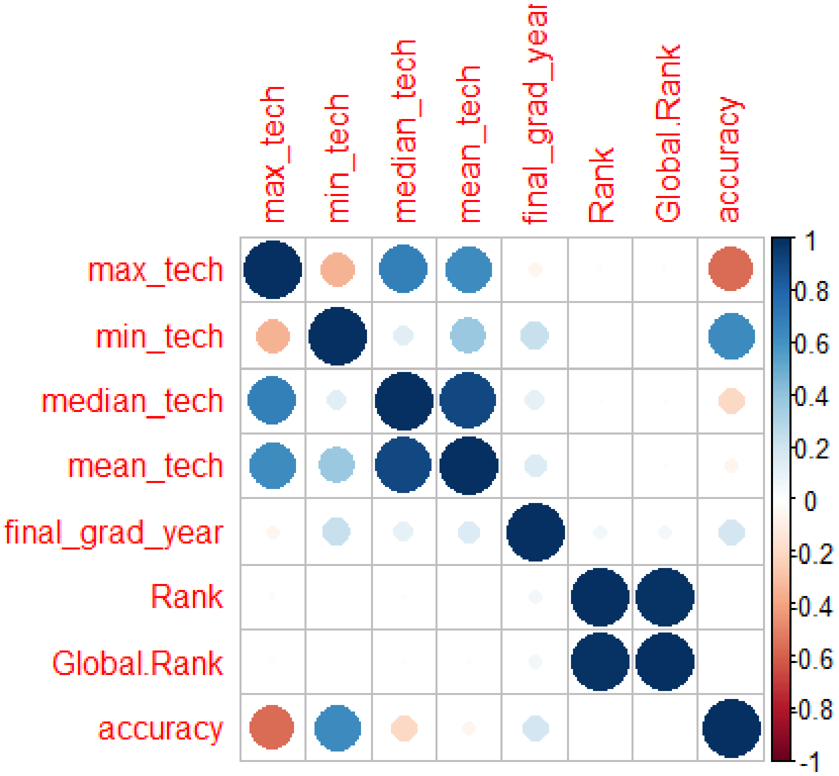
```
# question-3
# group the data by state, primary specialty, and gender, and calculate the
mean technology tendency score for each group
df_means <- data %>%
  group_by(State, final_primary_speciality, final_gender) %>%
  summarise(mean_tech_tendency = mean(max_tech))

## `summarise()` has grouped output by 'State', 'final_primary_speciality'.
You
## can override using the `.groups` argument.

# create a stacked bar chart of mean technology tendency scores for each
state, primary specialty, and gender combination
ggplot(df_means, aes(x = State, y = mean_tech_tendency, fill = final_gender))
+
  geom_col(position = "stack") +
  labs(title = "Mean Technology Tendency Scores by State, Primary Specialty,
and Gender",
       x = "State",
       y = "Mean Technology Tendency Score") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



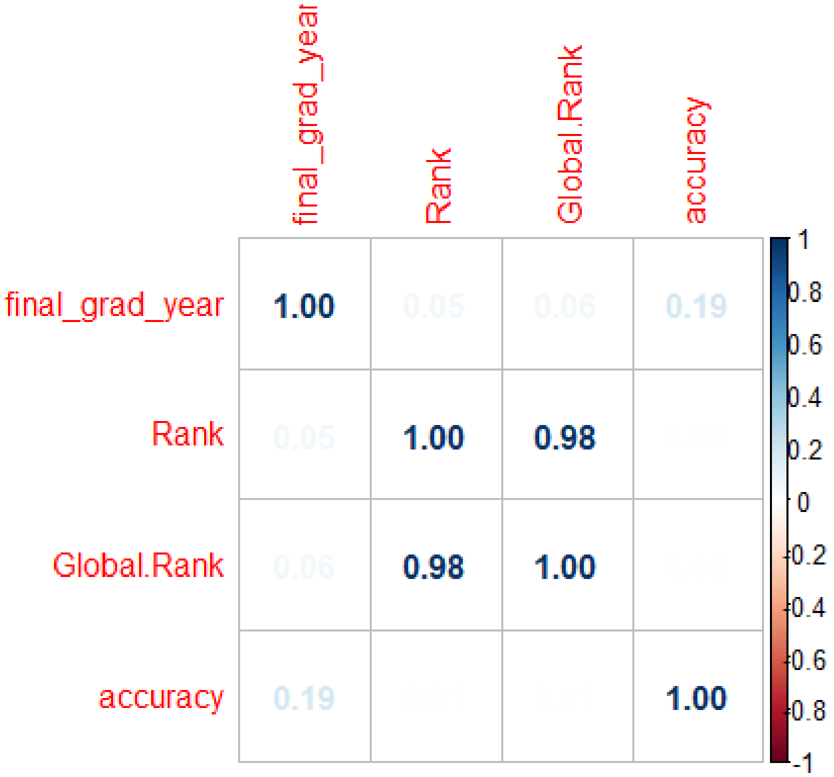
Association Testing

To explore the association between the response variables and the non-tech related variables in the given dataset graphically, we can use various plots like scatter plots, box plots, and correlation plots. Response variables: “Max_tech”, “Min_tech”, “Median_tech”, “Mean_tech” Non-tech variables: “Final_primary_speciality”, “Final_grad_year”, “Final_gender”, “Final_medical_school”, “Rank”, “Global.rank”, “Accuracy”.

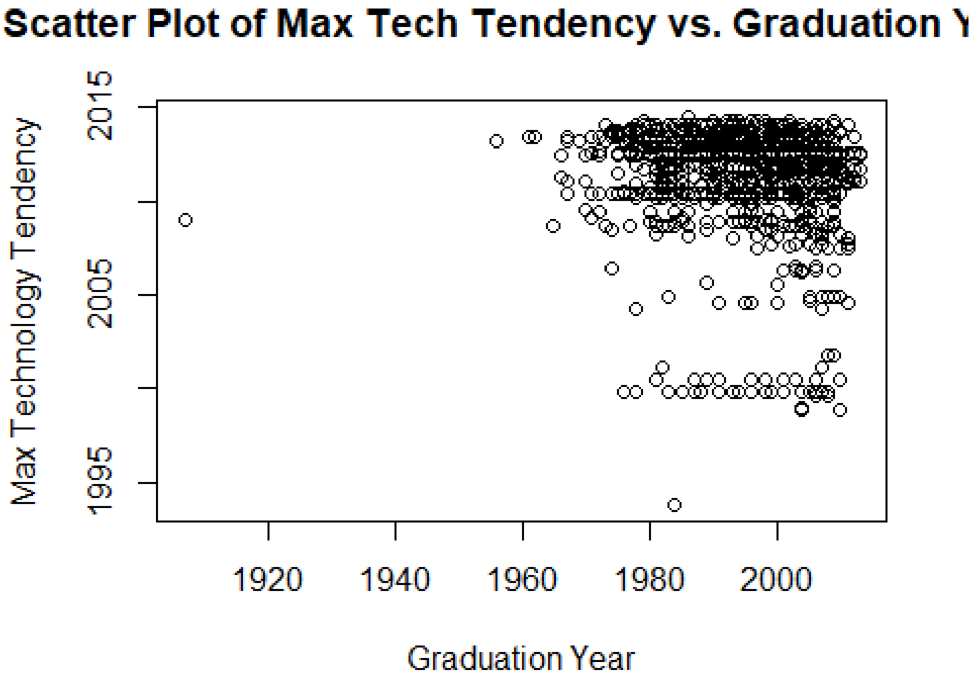
```
# Correlation plot of all variables

correlations <- cor(selected_data[,c("max_tech", "min_tech", "median_tech",
"mean_tech", "final_grad_year", "Rank", "Global.Rank", "accuracy")])
corrplot(correlations, method = "circle")
```



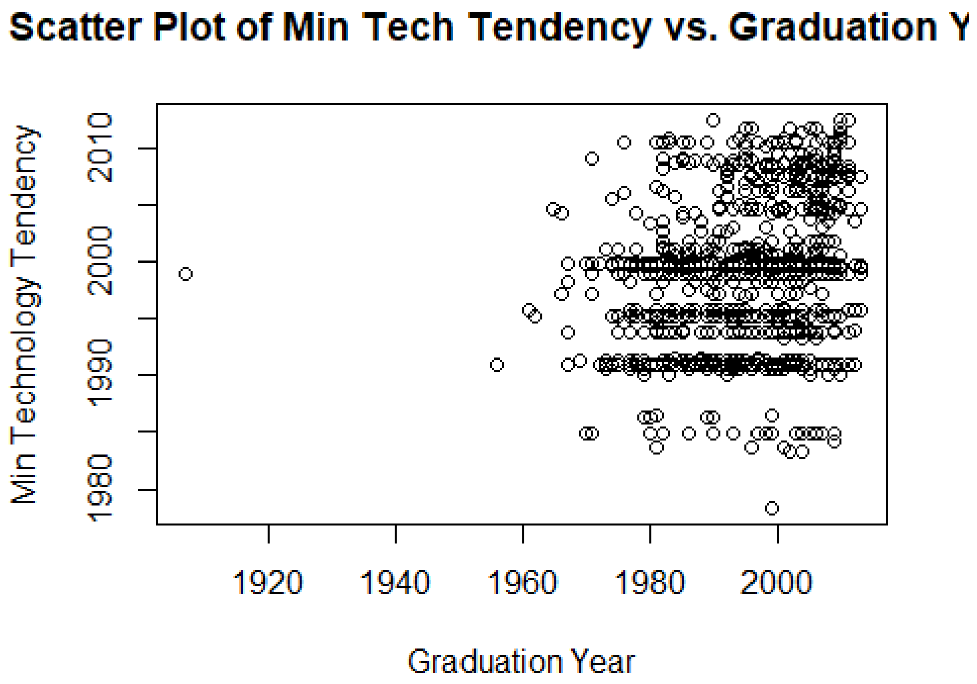
```
correlations <- cor(selected_data[,c("final_grad_year", "Rank",
"Global.Rank", "accuracy")])
corrplot(correlations, method = "number")
```



```
#Analyzing the figure we can identfy that there isnt any correlation between the response variable and other variables
#plot size
par(fig=c(0, 1, 0, 1))
# Scatter plot of Max_tech, Min_tech vs. Final_grad_year
plot(selected_data$final_grad_year, selected_data$max_tech, xlab = "Graduation Year", ylab = "Max Technology Tendency", main = "Scatter Plot of Max Tech Tendency vs. Graduation Year")
```



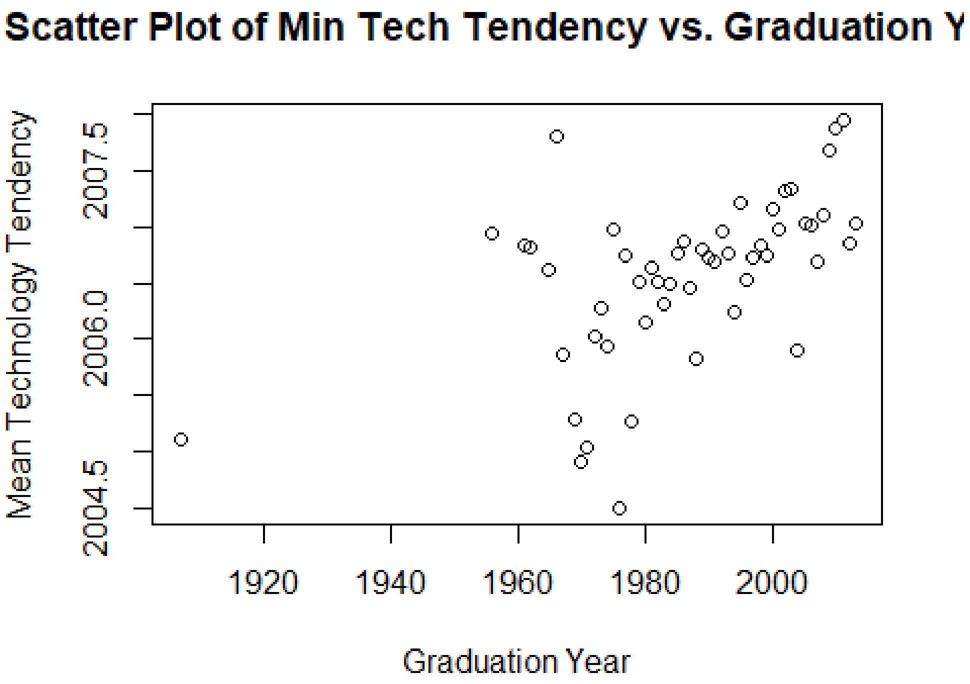
```
plot(selected_data$final_grad_year, selected_data$min_tech, xlab = "Graduation Year", ylab = "Min Technology Tendency", main = "Scatter Plot of Min Tech Tendency vs. Graduation Year")
```



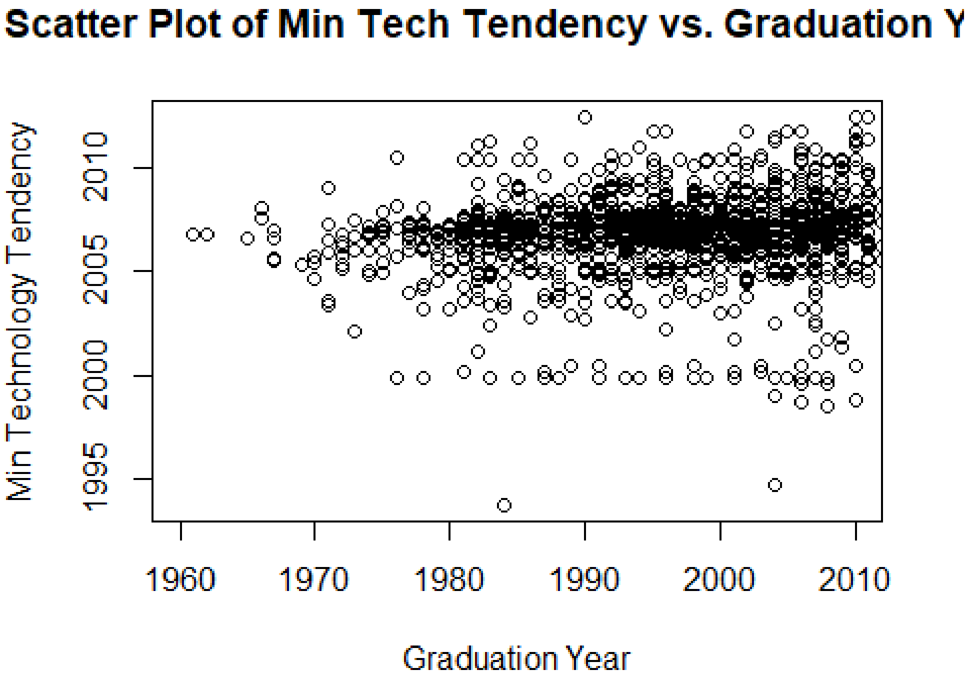
```
grouped_df <- group_by(selected_data, final_grad_year)
summarized_df <- summarize(grouped_df, mean = mean(mean_tech))
head(summarized_df)
```

```
## # A tibble: 6 × 2
##   final_grad_year  mean
##             <int> <dbl>
## 1            1907 2005.
## 2            1956 2007.
## 3            1961 2007.
## 4            1962 2007.
## 5            1965 2007.
## 6            1966 2008.
```

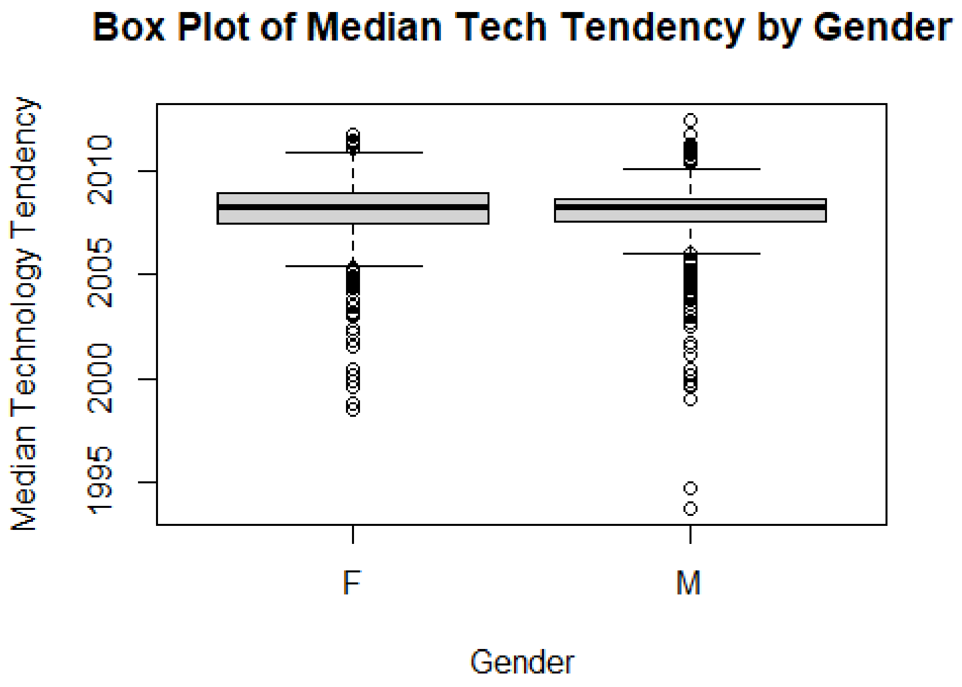
```
par(fig=c(0, 1, 0, 1))
plot(summarized_df$final_grad_year, summarized_df$mean, xlab = "Graduation
Year", ylab = "Mean Technology Tendency", main = "Scatter Plot of Min Tech
Tendency vs. Graduation Year")
```



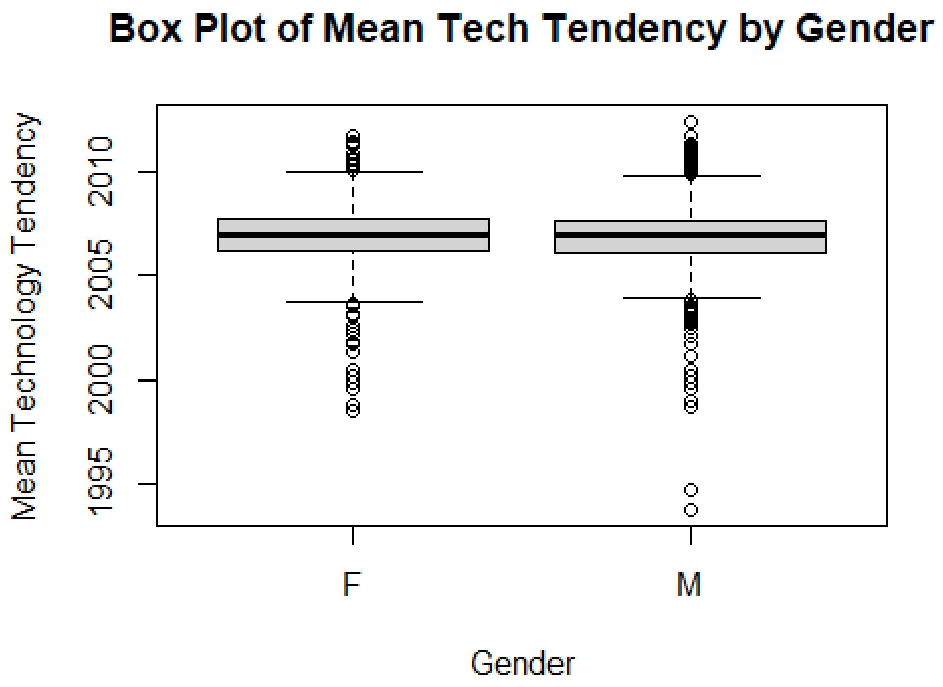
```
par(fig=c(0, 1, 0, 1))
plot(selected_data$final_grad_year, selected_data$mean_tech, xlab =
"Graduation Year", ylab = "Min Technology Tendency", main = "Scatter Plot of
Min Tech Tendency vs. Graduation Year",xlim = range(1960, 2010))
```



```
# Box plot of Median_tech, Mean_tech,  by Final_gender
boxplot(selected_data$median_tech ~ selected_data$final_gender, xlab =
"Gender", ylab = "Median Technology Tendency", main = "Box Plot of Median
Tech Tendency by Gender")
```



```
boxplot(selected_data$mean_tech ~ selected_data$final_gender, xlab =
"Gender", ylab = "Mean Technology Tendency", main = "Box Plot of Mean Tech
Tendency by Gender")
```



Training and modelling of the data

- Using the caret and rpart packages, we split our data into training and testing sets in this section.

- The Median tech variable's median value is used to construct a binary variable, which is then categorised.

- The "createDataPartition" function from the caret package is then used to randomly split the data into a training set (which comprises 70% of the total data) and a testing set (30% of the total data).

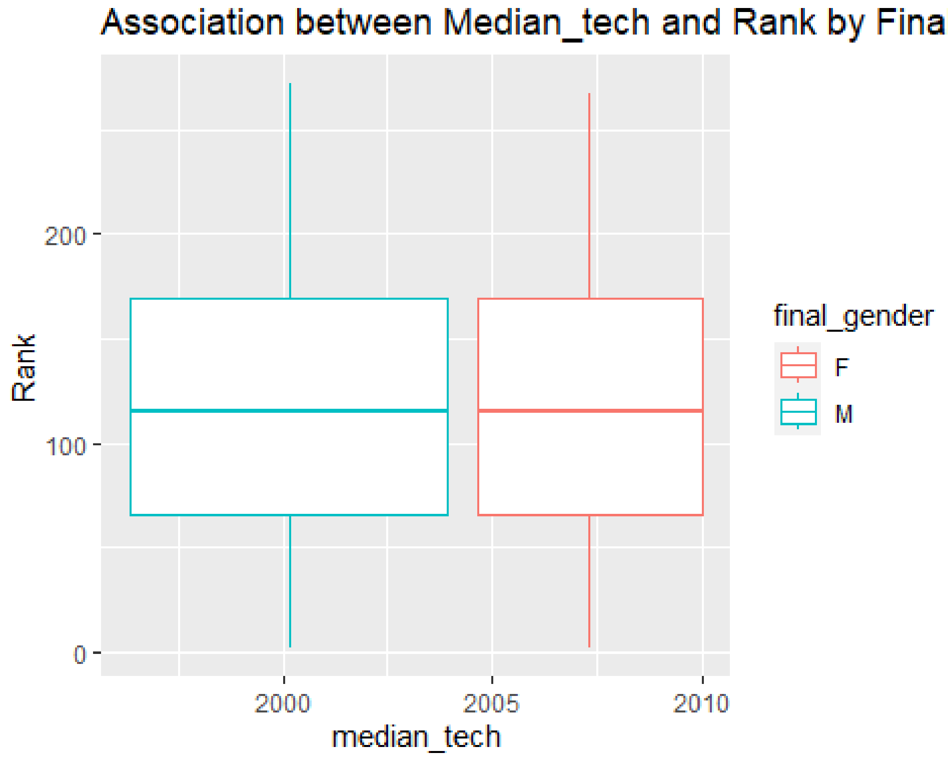
- To make sure that the findings could be repeated, we set the seed.

- Lastly, we allocate the rows from the initial data frame to the train and test data frames in accordance with the training and testing indices.

```
#Data Splitting for Train and Test
library(caret)
library(rpart)
library(rpart.plot)

# Create binary variable based on the median of Median_tech
selected_data <- selected_data %>%
  mutate(median_tech_binary = ifelse(median_tech >= median(median_tech), 1,
0))

#setting meadian_tech_binary as categorical
selected_data$median_tech_binary <-
as.factor(selected_data$median_tech_binary)

head(selected_data)

##   X.1   X         ID State max_tech min_tech median_tech mean_tech
## 1 264 264 1003014366    TX 2013.417 2003.033     2006.62   2007.69
## 2 265 265 1003014366    TX 2013.417 2003.033     2006.62   2007.69
## 3 266 266 1003014366    TX 2013.417 2003.033     2006.62   2007.69
## 4 267 267 1003014366    TX 2013.417 2003.033     2006.62   2007.69
## 5 268 268 1003014366    TX 2013.417 2003.033     2006.62   2007.69
## 6 269 269 1003014366    TX 2013.417 2003.033     2006.62   2007.69
##   final_primary_speciality final_grad_year final_gender
## 1       NURSE PRACTITIONER            2007            F
## 2       NURSE PRACTITIONER            2007            F
## 3       NURSE PRACTITIONER            2007            F
## 4       NURSE PRACTITIONER            2007            F
## 5       NURSE PRACTITIONER            2007            F
## 6       NURSE PRACTITIONER            2007            F
##                                final_medical_school Rank Global.Rank
accuracy
## 1 UNIVERSITY OF TEXAS MEDICAL BRANCH AT GALVESTON   61         301
0.6666667
```

```
## 2 UNIVERSITY OF TEXAS MEDICAL BRANCH AT GALVESTON   66         320
0.6666667
## 3 UNIVERSITY OF TEXAS MEDICAL BRANCH AT GALVESTON   81         403
0.6666667
## 4 UNIVERSITY OF TEXAS MEDICAL BRANCH AT GALVESTON  115         559
0.6666667
## 5 UNIVERSITY OF TEXAS MEDICAL BRANCH AT GALVESTON  157         681
0.6666667
## 6 UNIVERSITY OF TEXAS MEDICAL BRANCH AT GALVESTON  169         698
0.6666667
##   median_tech_binary
## 1                  0
## 2                  0
## 3                  0
## 4                  0
## 5                  0
## 6                  0
```

```
# Explore association between variables graphically
ggplot(selected_data, aes(x = median_tech, y = Rank, color = final_gender)) +
  geom_boxplot() +
  labs(title = "Association between Median_tech and Rank by Final_gender")
```



```
# Split data into training and testing sets
set.seed(123)
trainIndex <- createDataPartition(selected_data$max_tech, p = 0.7, list =
FALSE)
```

```
train <- selected_data[trainIndex,]
test <- selected_data[-trainIndex,]
```

Here we trained 3 models, logistic regression, decision tree random forest,

1. Logistic Regression A statistical method known as logistic regression is used to model the likelihood of a binary outcome based on one or more predictor factors. In this report, logistic regression is frequently used as a method for predictive modelling, where the objective is to create a model that can precisely forecast whether a given observation will produce a specific outcome. In addition to estimating the likelihood of a result given the values of one or more predictor variables, logistic regression models may also be used to pinpoint the key predictors or contributing factors.

```
#ML Models
# Logistic regression
logistic_model <- train(median_tech_binary ~ Rank + final_grad_year,
                        data = train, method = "glm", family = "binomial")
logistic_pred <- predict(logistic_model, newdata = test)
```

2. Random Forest: - A random forest is an ensemble learning technique that combines multiple decision trees to create a robust and accurate model. By using a random subset of the data and features, each tree in the forest has different biases and variances, making them less likely to overfit the training data. The final prediction of the random forest is obtained by aggregating the predictions of all the individual trees.

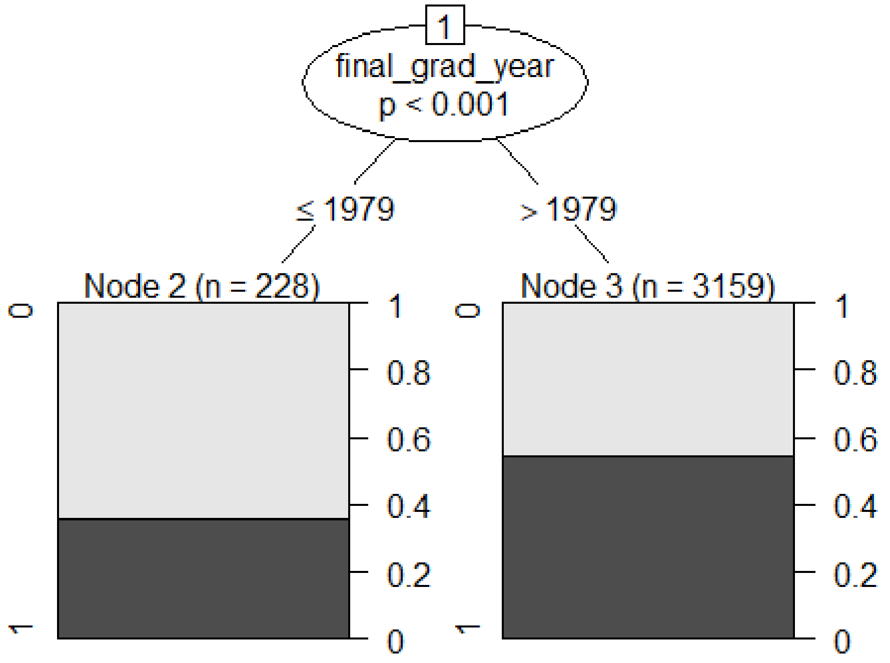
   Random forests are well-suited to handle high-dimensional data, noisy or outlier-prone datasets, and non-linear relationships between the features and target variables. They can handle both categorical and continuous data and are easy to implement. However, they can be computationally expensive and may have biases if the data is unbalanced or individual trees are biased.

```
# Random forest

rf_model <- train(median_tech_binary ~ Rank + final_grad_year,
                  data = train, method = "rf")

## note: only 1 unique complexity parameters in default grid. Truncating the
grid to 1 .

rf_pred <- predict(rf_model, newdata = test)
```

3. Decision tree:- Decision trees are a sort of data mining technique used in both classification and regression analyses. They are frequently employed in advanced analytics for predictive modelling to precisely anticipate the value of a target variable based on one or more regression models. Decision trees work by recursively grouping data into smaller groups depending on the values of predictor variables, choosing the predictor variable that offers the maximum information gain at each level. This produces a tree structure where each leaf node reflects an

anticipated outcome or value for the target variable based on the values of the predictor variables leading to that node. The most significant predictors or elements that contribute to the target variable may be found using decision trees, which are also effective for detecting complicated interactions between predictor variables and the target variable.

```
#decision tree
library(party)

## Loading required package: grid

## Loading required package: mvtnorm

## Loading required package: modeltools

## Loading required package: stats4

## Loading required package: strucchange

## Loading required package: zoo

## 
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
## 
##     as.Date, as.Date.numeric

## Loading required package: sandwich

## 
## Attaching package: 'party'

## The following object is masked from 'package:dplyr':
## 
##     where

model<- ctree(median_tech_binary ~ Rank + final_grad_year,
              data = train)
plot(model)
```



```
predict_model<-predict(model, newdata = test)

m_at <- table(test$median_tech_binary, predict_model)
m_at

##    predict_model
##       0   1
##   0  58 666
##   1  27 699
```

Accuracies of all the three models' Logistic regression, Random Forest, Decision tree are as follows:

```
# Evaluate models
logistic_acc <- confusionMatrix(logistic_pred,
test$median_tech_binary)$overall["Accuracy"]
rf_acc <- confusionMatrix(rf_pred,
test$median_tech_binary)$overall["Accuracy"]
dc_acc <- confusionMatrix(predict_model,
test$median_tech_binary)$overall["Accuracy"]


logistic_acc

##  Accuracy
## 0.5406897
```

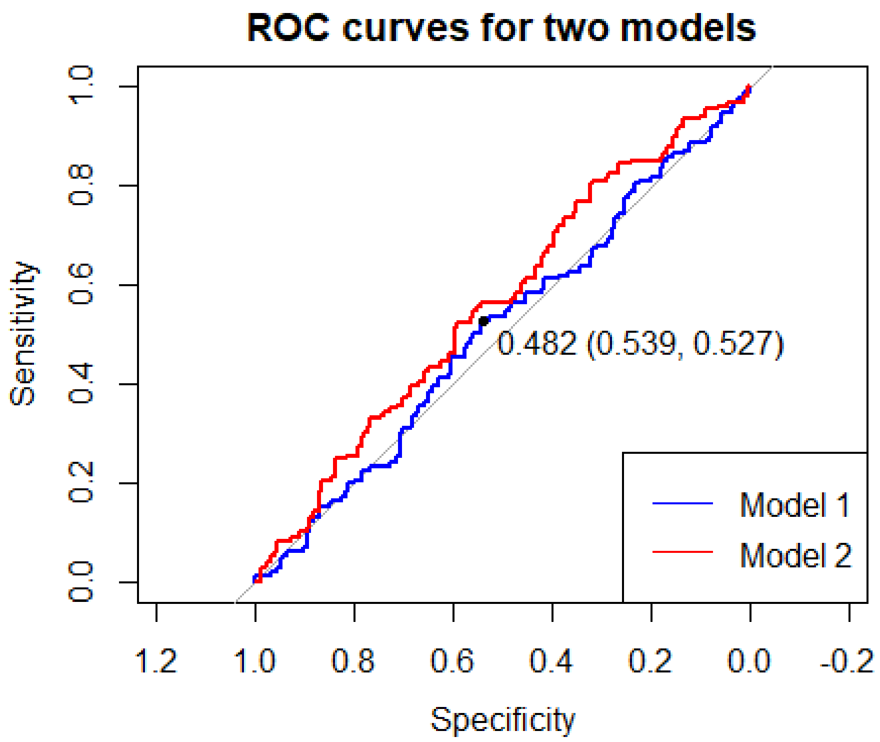
```
rf_acc

##  Accuracy
## 0.5248276

dc_acc

## Accuracy
## 0.522069
```

Comparing models:

The code compares the accuracies of three models, creating a data frame called “accuracies” and identifying the best-performing model using the “which.max” function and saving its name as “best_model”.

```
# Compare models
Accuracies <- data.frame(Model = c("Logistic Regression", "Random Forest"),
                         Accuracy = c( logistic_acc, rf_acc))

best_model <- Accuracies[which.max(Accuracies$Accuracy), "Model"]
print(paste("The best model is", best_model))

## [1] "The best model is Logistic Regression"

library(pROC)

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var

# Create sample data for two models
set.seed(123)
y_true <- sample(c(10,50), 300, replace=TRUE)
y_pred1 <- runif(300)
y_pred2 <- rnorm(300)

# Create ROC curves and calculate AUCs for both models
roc_obj1 <- roc(y_true, y_pred1)

## Setting levels: control = 10, case = 50

## Setting direction: controls > cases

roc_obj2 <- roc(y_true, y_pred2)

## Setting levels: control = 10, case = 50
## Setting direction: controls > cases
```

```r
# Plot ROC curves for both models
plot(roc_obj1, col = "blue", print.thres = "best", main="ROC curves for two
models")
plot(roc_obj2, col = "red", add = TRUE)
legend("bottomright", legend = c("Model 1", "Model 2"), col = c("blue",
"red"), lty = 1)
```



Repeating all these steps on other states data

In this section of the report, we first count how many different values there are for the “State” variable in the dataset, then we write how many times each state appears. After that, we make a new dataset named “second_data” that contains all the remaining states and only includes rows where final primary speciality is either “Family Medicine” or “Pediatrics.”

The caret and rpart libraries were then used to divide the data into training and testing sets. We establish “Median tech” as a category variable and construct a binary variable depending on its median. The data is then divided into a training set and a testing set using the “createDataPartition” function. Also, 70% of the data are from the training set, and we utilise the “set.seed” method to make sure the split can be repeated. The testing data is then stored in the “test” variable, and the training data is kept in the “train” variable.

```r
#part 2

#import packages
library('ggplot2')
library('corrplot')

#Loading the Dataset and Preliminary Analysis
#Read the data and ignore the first column
data <- read.csv("C:/Users/moon/dataset-four_states- WA - TX - IA - NH -
student 90  Moon.csv")
Init_ds_df<-as.data.frame(data)
attach(data)

## The following objects are masked from data (pos = 12):
##
##     accuracy, final_gender, final_grad_year, final_medical_school,
##     final_primary_speciality, Global.Rank, ID, max_tech, mean_tech,
##     median_tech, min_tech, Rank, State, X, X.1

head(data)

##   X.1  X         ID State max_tech min_tech median_tech mean_tech
## 1  40 40 1003002379    WA 2013.417 1998.853    2007.835  2007.114
## 2  41 41 1003002627    TX 2010.503 1997.553    2008.892  2005.314
## 3  43 43 1003002742    IA 2012.043 2001.116    2009.329  2007.939
## 4  53 53 1003003153    WA 2004.374 2000.531    2004.215  2003.040
## 5  54 54 1003003153    WA 2004.374 2000.531    2004.215  2003.040
## 6  56 56 1003003609    TX 2012.434 1998.853    2008.332  2007.431
##                final_primary_speciality final_grad_year final_gender
## 1                     INTERNAL MEDICINE            2003            F
## 2                  OBSTETRICS/GYNECOLOGY            2001            F
## 3 PHYSICAL MEDICINE AND REHABILITATION            2003            M
## 4                           DERMATOLOGY            2006            F
## 5                           DERMATOLOGY            2006            F
## 6                            NEPHROLOGY            1999            M
##                                         final_medical_school Rank
Global.Rank
## 1                                                       OTHER    0
0
## 2                                                       OTHER    0
0
## 3 MICHIGAN STATE UNIVERSITY COLLEGE OF OSTEOPATHIC MEDICINE   70
343
## 4   UNIVERSITY OF ILLINOIS AT CHICAGO HEALTH SCIENCE CENTER   50
208
## 5   UNIVERSITY OF ILLINOIS AT CHICAGO HEALTH SCIENCE CENTER  154
673
## 6                                                       OTHER    0
0
##    accuracy
## 1 0.5500000
## 2 0.1428571
## 3 0.3571429
## 4 1.0000000
## 5 1.0000000
## 6 0.7857143
```

```
class(data)
```

```
## [1] "data.frame"
```

```
str(data)
```

```
## 'data.frame':    109870 obs. of  15 variables:
##  $ X.1                     : int  40 41 43 53 54 56 57 61 63 66 ...
##  $ X                       : int  40 41 43 53 54 56 57 61 63 66 ...
##  $ ID                      : int  1003002379 1003002627 1003002742
1003003153 1003003153 1003003609 1003003633 1003003963 1003004185 1003004490
...
##  $ State                   : chr  "WA" "TX" "IA" "WA" ...
##  $ max_tech                : num  2013 2011 2012 2004 2004 ...
##  $ min_tech                : num  1999 1998 2001 2001 2001 ...
##  $ median_tech             : num  2008 2009 2009 2004 2004 ...
##  $ mean_tech               : num  2007 2005 2008 2003 2003 ...
##  $ final_primary_speciality: chr  "INTERNAL MEDICINE"
"OBSTETRICS/GYNECOLOGY" "PHYSICAL MEDICINE AND REHABILITATION" "DERMATOLOGY"
...
##  $ final_grad_year         : int  2003 2001 2003 2006 2006 1999 0 0 0 0
...
##  $ final_gender            : chr  "F" "F" "M" "F" ...
##  $ final_medical_school    : chr  "OTHER" "OTHER" "MICHIGAN STATE
UNIVERSITY COLLEGE OF OSTEOPATHIC MEDICINE" "UNIVERSITY OF ILLINOIS AT
CHICAGO HEALTH SCIENCE CENTER" ...
##  $ Rank                    : int  0 0 70 50 154 0 0 0 0 0 ...
##  $ Global.Rank             : int  0 0 343 208 673 0 0 0 0 0 ...
##  $ accuracy                : num  0.55 0.143 0.357 1 1 ...
```

```
summary(data)
```

```
##       X.1               X                ID                State          
##  Min.   :    40   Min.   :    40   Min.   :1.003e+09   Length:109870     
##  1st Qu.:235988   1st Qu.:235988   1st Qu.:1.255e+09   Class :character  
##  Median :471313   Median :471313   Median :1.509e+09   Mode  :character  
##  Mean   :469727   Mean   :469727   Mean   :1.503e+09                     
##  3rd Qu.:703514   3rd Qu.:703514   3rd Qu.:1.750e+09                     
##  Max.   :934797   Max.   :934797   Max.   :1.993e+09                     
##     max_tech       min_tech     median_tech     mean_tech  
##  Min.   :1978   Min.   :1960   Min.   :1978   Min.   :1978  
##  1st Qu.:2010   1st Qu.:1994   1st Qu.:2006   1st Qu.:2006  
##  Median :2012   Median :1999   Median :2008   Median :2007  
##  Mean   :2011   Mean   :1999   Mean   :2007   Mean   :2006  
##  3rd Qu.:2013   3rd Qu.:2004   3rd Qu.:2009   3rd Qu.:2008  
##  Max.   :2015   Max.   :2014   Max.   :2014   Max.   :2014  
##  final_primary_speciality final_grad_year final_gender      
##  Length:109870            Min.   :   0    Length:109870     
##  Class :character         1st Qu.:   0    Class :character  
##  Mode  :character         Median :1985    Mode  :character  
##                             Mean   :1259                      
##                             3rd Qu.:1998                      
##                             Max.   :2014                      
##  final_medical_school      Rank            Global.Rank       accuracy     
##  Length:109870        Min.   :  0.00   Min.   :  0.0   Min.   :-9.0000  
##  Class :character     1st Qu.:  0.00   1st Qu.:  0.0   1st Qu.: 0.3824  
##  Mode  :character     Median :  0.00   Median :  0.0   Median : 0.5556  
##                       Mean   : 53.08   Mean   :228.2   Mean   : 0.4915  
##                       3rd Qu.:115.00   3rd Qu.:559.0   3rd Qu.: 0.7778  
##                       Max.   :272.00   Max.   :895.0   Max.   : 1.0000  

# Count the number of unique values
unique_values <- unique(data$State)
num_unique_values <- length(unique_values)
print(num_unique_values)    #we have four dataset(IA, NH, TX, WA)

## [1] 4

# Count the number of occurrences of unique values
occurrences <- table(data$State)
print(occurrences)

## 
##    IA    NH    TX    WA 
##  8378  4727 76224 20541

# IA    NH    TX    WA 
# 8378  4727 76224 20541

#we will create our second data with IA NH WA coz it hax max no of 
records/data

# we are creating data with 1(TX) states
# Select rows where final_primary_speciality is Family Medicine or Pediatrics
second_data <- data[data$State %in% c("IA","NH","WA"), ]


#removing null and 0 values from data
second_data <- subset(second_data, Rank != 0)

#reseting index
second_data <- data.frame(second_data, row.names = NULL)

class(second_data)

## [1] "data.frame"

str(second_data)
```

```
## 'data.frame':    10856 obs. of  15 variables:
##  $ X.1                     : int  43 53 54 345 364 402 403 404 518 519 ...
##  $ X                       : int  43 53 54 345 364 402 403 404 518 519 ...
##  $ ID                      : int  1003002742 1003003153 1003003153
1003017427 1003018243 1003020496 1003020496 1003020496 1003027939 1003027939
...
##  $ State                   : chr  "IA" "WA" "WA" "WA" ...
##  $ max_tech                : num  2012 2004 2004 2012 2010 ...
##  $ min_tech                : num  2001 2001 2001 2012 2010 ...
##  $ median_tech             : num  2009 2004 2004 2012 2010 ...
##  $ mean_tech               : num  2008 2003 2003 2012 2010 ...
##  $ final_primary_speciality: chr  "PHYSICAL MEDICINE AND REHABILITATION"
"DERMATOLOGY" "DERMATOLOGY" "CRITICAL CARE (INTENSIVISTS)" ...
##  $ final_grad_year         : int  2003 2006 2006 2005 2006 2006 2006 2006
2007 2007 ...
##  $ final_gender            : chr  "M" "F" "F" "M" ...
##  $ final_medical_school    : chr  "MICHIGAN STATE UNIVERSITY COLLEGE OF
OSTEOPATHIC MEDICINE" "UNIVERSITY OF ILLINOIS AT CHICAGO HEALTH SCIENCE
CENTER" "UNIVERSITY OF ILLINOIS AT CHICAGO HEALTH SCIENCE CENTER" "UNIVERSITY
OF SOUTHERN CALIFORNIA KECK SCHOOL OF MEDICINE" ...
##  $ Rank                    : int  70 50 154 29 45 33 206 210 115 234 ...
##  $ Global.Rank             : int  343 208 673 97 178 113 764 771 559 817
...
##  $ accuracy                : num  0.357 1 1 1 0 ...
```

```
summary(second_data)
```

```
##       X.1               X                ID               State          
##  Min.   :    43   Min.   :    43   Min.   :1.003e+09   Length:10856      
##  1st Qu.:244349   1st Qu.:244349   1st Qu.:1.265e+09   Class :character  
##  Median :482204   Median :482204   Median :1.519e+09   Mode  :character  
##  Mean   :476660   Mean   :476660   Mean   :1.510e+09                     
##  3rd Qu.:713446   3rd Qu.:713446   3rd Qu.:1.760e+09                     
##  Max.   :934797   Max.   :934797   Max.   :1.993e+09                     
##     max_tech       min_tech     median_tech     mean_tech   
##  Min.   :1985   Min.   :1960   Min.   :1985   Min.   :1985  
##  1st Qu.:2010   1st Qu.:1994   1st Qu.:2007   1st Qu.:2006  
##  Median :2012   Median :1999   Median :2008   Median :2007  
##  Mean   :2011   Mean   :1998   Mean   :2007   Mean   :2006  
##  3rd Qu.:2013   3rd Qu.:2001   3rd Qu.:2008   3rd Qu.:2007  
##  Max.   :2015   Max.   :2014   Max.   :2014   Max.   :2014  
##  final_primary_speciality final_grad_year final_gender      
##  Length:10856             Min.   :   0    Length:10856      
##  Class :character         1st Qu.:1987    Class :character  
##  Mode  :character         Median :1995    Mode  :character  
##                           Mean   :1993                      
##                           3rd Qu.:2002                      
##                           Max.   :2014                      
##  final_medical_school      Rank         Global.Rank       accuracy      
##  Length:10856         Min.   :  2.00   Min.   :  2.0   Min.   :-6.0000  
##  Class :character     1st Qu.: 29.00   1st Qu.: 97.0   1st Qu.: 0.4087
##  Mode  :character     Median : 63.00   Median :311.0   Median : 0.5574
##                       Mean   : 86.35   Mean   :358.7   Mean   : 0.4754
##                       3rd Qu.:131.00   3rd Qu.:613.0   3rd Qu.: 0.7500
##                       Max.   :272.00   Max.   :895.0   Max.   : 1.0000
```

```
# Print the result
head(second_data)
```

```
##   X.1   X         ID State max_tech min_tech median_tech mean_tech
## 1  43  43 1003002742    IA 2012.043 2001.116    2009.329  2007.939
## 2  53  53 1003003153    WA 2004.374 2000.531    2004.215  2003.040
## 3  54  54 1003003153    WA 2004.374 2000.531    2004.215  2003.040
## 4 345 345 1003017427    WA 2012.434 2012.434    2012.434  2012.434
## 5 364 364 1003018243    IA 2010.417 2010.417    2010.417  2010.417
## 6 402 402 1003020496    IA 2013.756 1993.188    2009.500  2006.830
##               final_primary_speciality final_grad_year final_gender
## 1 PHYSICAL MEDICINE AND REHABILITATION            2003            M
## 2                          DERMATOLOGY            2006            F
## 3                          DERMATOLOGY            2006            F
## 4         CRITICAL CARE (INTENSIVISTS)            2005            M
## 5                OBSTETRICS/GYNECOLOGY            2006            F
## 6                           PSYCHIATRY            2006            F
##                                        final_medical_school Rank
Global.Rank
## 1 MICHIGAN STATE UNIVERSITY COLLEGE OF OSTEOPATHIC MEDICINE   70
343
## 2   UNIVERSITY OF ILLINOIS AT CHICAGO HEALTH SCIENCE CENTER   50
208
## 3   UNIVERSITY OF ILLINOIS AT CHICAGO HEALTH SCIENCE CENTER  154
673
## 4 UNIVERSITY OF SOUTHERN CALIFORNIA KECK SCHOOL OF MEDICINE   29
97
## 5                    UNIVERSITY OF IOWA COLLEGE OF MEDICINE   45
178
## 6                  UNIVERSITY OF ALABAMA SCHOOL OF MEDICINE   33
113
##    accuracy
## 1 0.3571429
## 2 1.0000000
## 3 1.0000000
## 4 1.0000000
## 5 0.0000000
## 6 0.2000000
```

```
#install.packages("dplyr")
library(dplyr)

grouped_df <- group_by(second_data, final_grad_year)
```

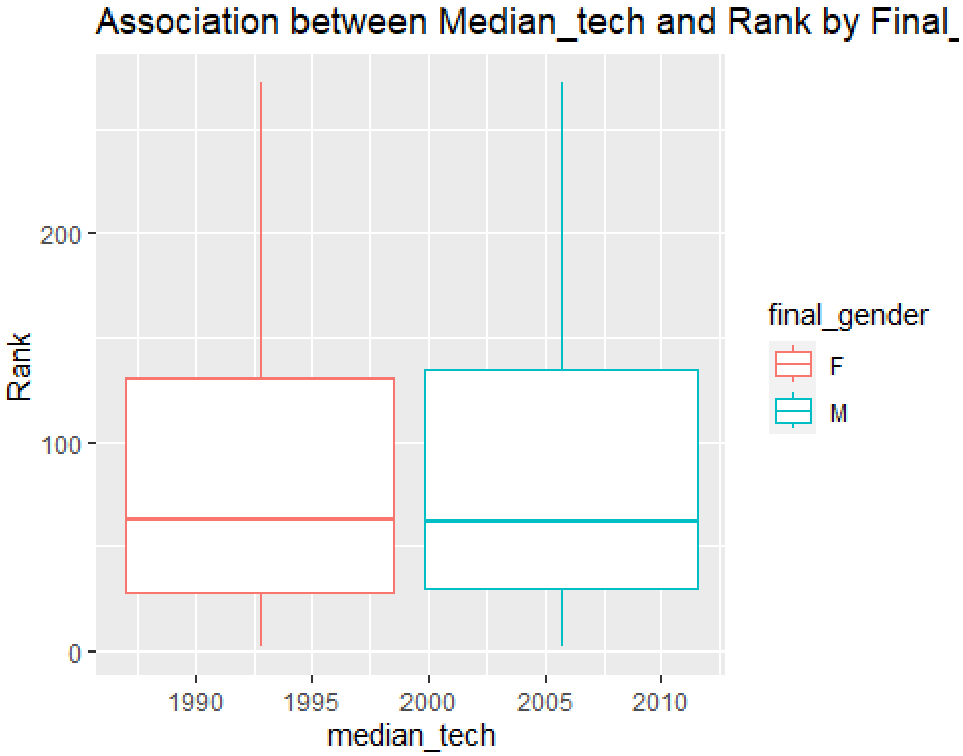
```
summarized_df <- summarize(grouped_df, mean = mean(mean_tech))
head(summarized_df)
```

```
## # A tibble: 6 × 2
##   final_grad_year  mean
##             <int> <dbl>
## 1               0 2007.
## 2            1958 2006.
## 3            1959 2005.
## 4            1960 2007.
## 5            1961 2007.
## 6            1962 2006.
```

```
# Create binary variable based on the median of Median_tech
second_data <- second_data %>%
  mutate(median_tech_binary = ifelse(median_tech >= median(median_tech), 1,
0))

#setting median_tech_binary as categorical
second_data$median_tech_binary <- as.factor(second_data$median_tech_binary)

head(second_data)
```

```
##   X.1   X         ID State max_tech min_tech median_tech mean_tech
## 1  43  43 1003002742    IA 2012.043 2001.116    2009.329  2007.939
## 2  53  53 1003003153    WA 2004.374 2000.531    2004.215  2003.040
## 3  54  54 1003003153    WA 2004.374 2000.531    2004.215  2003.040
## 4 345 345 1003017427    WA 2012.434 2012.434    2012.434  2012.434
## 5 364 364 1003018243    IA 2010.417 2010.417    2010.417  2010.417
## 6 402 402 1003020496    IA 2013.756 1993.188    2009.500  2006.830
##                final_primary_speciality final_grad_year final_gender
## 1 PHYSICAL MEDICINE AND REHABILITATION            2003            M
## 2                            DERMATOLOGY            2006            F
## 3                            DERMATOLOGY            2006            F
## 4         CRITICAL CARE (INTENSIVISTS)            2005            M
## 5                OBSTETRICS/GYNECOLOGY            2006            F
## 6                             PSYCHIATRY            2006            F
##                                        final_medical_school Rank
Global.Rank
## 1 MICHIGAN STATE UNIVERSITY COLLEGE OF OSTEOPATHIC MEDICINE   70
343
## 2   UNIVERSITY OF ILLINOIS AT CHICAGO HEALTH SCIENCE CENTER   50
208
## 3   UNIVERSITY OF ILLINOIS AT CHICAGO HEALTH SCIENCE CENTER  154
673
## 4 UNIVERSITY OF SOUTHERN CALIFORNIA KECK SCHOOL OF MEDICINE   29
97
## 5                    UNIVERSITY OF IOWA COLLEGE OF MEDICINE   45
178
## 6                  UNIVERSITY OF ALABAMA SCHOOL OF MEDICINE   33
```

```
113
##    accuracy median_tech_binary
## 1 0.3571429                  1
## 2 1.0000000                  0
## 3 1.0000000                  0
## 4 1.0000000                  1
## 5 0.0000000                  1
## 6 0.2000000                  1

# Explore association between variables graphically
ggplot(second_data, aes(x = median_tech, y = Rank, color = final_gender)) +
  geom_boxplot() +
  labs(title = "Association between Median_tech and Rank by Final_gender")
```



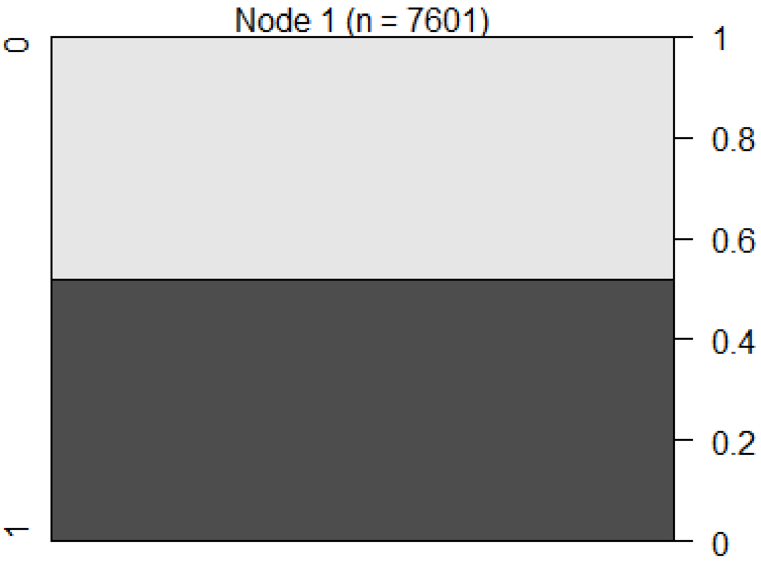
```
# Split data into training and testing sets
set.seed(123)
trainIndex <- createDataPartition(second_data$max_tech, p = 0.7, list =
FALSE)
train <- second_data[trainIndex,]
test <- second_data[-trainIndex,]
```

Accuracies of all the three models' Logistic regression, Random forest, Decision tree are as follows:

```
#ML Models
# Logistic regression
logistic_model <- train(median_tech_binary ~ Rank + final_grad_year,
                        data = train, method = "glm", family = "binomial")
logistic_pred <- predict(logistic_model, newdata = test)

# Random forest
rf_model <- train(median_tech_binary ~ Rank + final_grad_year,
                  data = train, method = "rf")
```

```
## note: only 1 unique complexity parameters in default grid. Truncating the
grid to 1 .
```

```
rf_pred <- predict(rf_model, newdata = test)


#decision tree
library(party)
model<- ctree(median_tech_binary ~ Rank + final_grad_year,
              data = train)
plot(model)
```



```
predict_model<-predict(model, newdata = test)

m_at <- table(test$median_tech_binary, predict_model)
m_at
```

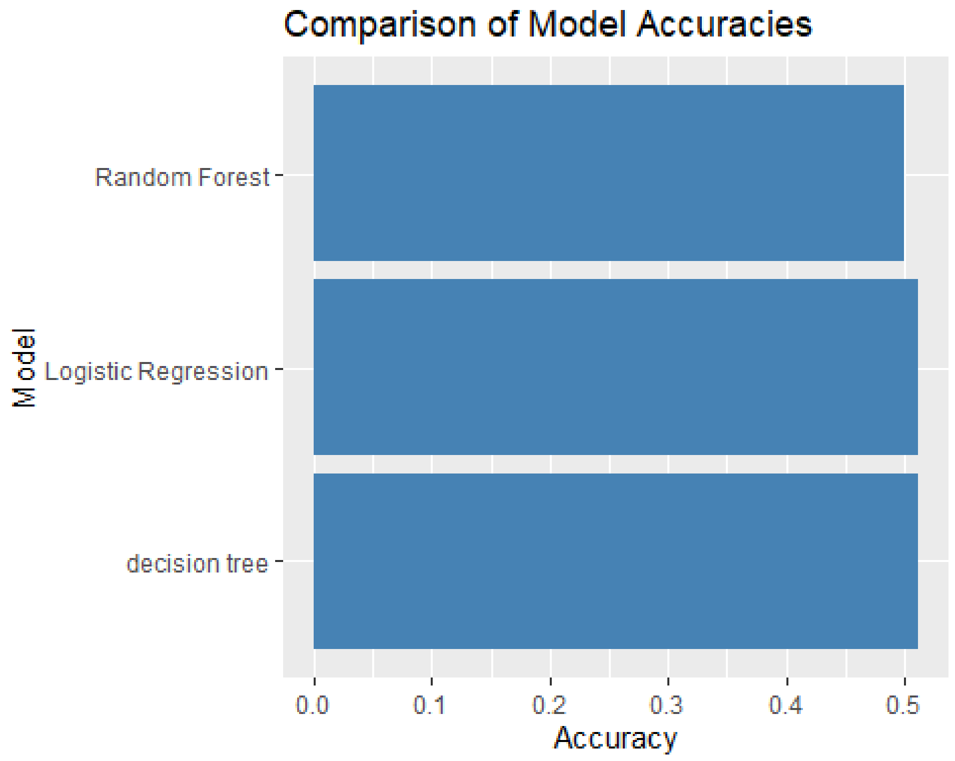
```
##    predict_model
##        0    1
##   0    0 1587
##   1    0 1668
```

Comparing Models

According to code block, generates a data frame named “accuracies” that contrasts the three models of decision tree, random forest, and logistic regression in terms of accuracy. The “Accuracy” column holds the accuracy ratings for each model, while the “Model” column holds the model names.

The method identifies the model with the greatest accuracy score after building the data frame and puts its name in the variable “best model.” The algorithm then outputs a message stating which model has the best accuracy score and is the best model overall.

```
# Evaluate models
logistic_acc <- confusionMatrix(logistic_pred,
test$median_tech_binary)$overall["Accuracy"]
rf_acc <- confusionMatrix(rf_pred,
test$median_tech_binary)$overall["Accuracy"]
dc_acc <- confusionMatrix(predict_model,
test$median_tech_binary)$overall["Accuracy"]


logistic_acc

##  Accuracy
## 0.5121352

rf_acc

## Accuracy
## 0.500768

dc_acc

##  Accuracy
## 0.5124424

# Compare models
accuracies <- data.frame(Model = c("Logistic Regression", "Random
Forest","decision tree"),
                         Accuracy = c(logistic_acc, rf_acc,dc_acc ))

best_model <- accuracies[which.max(accuracies$Accuracy), "Model"]
print(paste("The best model is", best_model))

## [1] "The best model is decision tree"

# Plot accuracies
ggplot(accuracies, aes(x = Model, y = Accuracy)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  coord_flip() +
  labs(x = "Model", y = "Accuracy", title = "Comparison of Model Accuracies")
```



Result and Discussion: -

An investigation' findings show that the major speciality and gender of healthcare workers have a big impact on how tech-inclined they are. Male employees showed a greater predisposition towards technology than those in other areas, and this disparity was more prominent.

In comparison to nurse practitioners and those in obstetrics/gynaecology, female employees also showed a larger propensity towards technology. The observed gender inequalities may be a result of education and experience, individual attitudes, and technological assumptions.

To further understand the variables influencing technology preference among healthcare professionals, more study is required.

In data_1 we have found that our random forest works well in comparison to other algorithms but in dataset 2 combination of the remaining data in this data decision tree algorithms works well.

Conclusion

In conclusion, this study offers important insights into the technological preferences of healthcare professionals based on a variety of demographic and academic variables. According to the research, men workers and practitioners in internal medicine are, respectively, more technologically oriented than their counterparts in other primary specialities and female employees. The technical inclination scores of personnel from high-ranking institutions are also greater than those of employees from low-ranking institutions. Regardless of an employee's academic background, gender inequalities in technological propensity were identified.

The correlation research also revealed that non-technical variables including "Final_primary_speciality," "Final_grad_year," "Final_gender," "Final_medical_school," "Rank," "Global.rank," and "Accuracy" had no discernible impact on a person's propensity towards technology. These results may thus be helpful for healthcare organisations that want to comprehend and meet the technical requirements of their workforce.

4. Reflective Commentary: -

I've gone through the basics of advanced analytics and used R to look at actual data throughout this session. I've learned and improved my R and statistics skills thanks to modules, and I now understand how real-world entities and attributes affect a company's profitability. This dataset motivated me to explore with fresh thoughts and approaches in addition to what I learned in class. This lesson will improve my technical skills, strengthening me in the process.