A4

Muhammet Furkan Isik

Assignment #4.

2021-12-12

I pledge on my honor that I have not given or received any unauthorized assistance on this assignment/examination. I further pledge that I have not copied any material from a book, article, the Internet or any other source except where I have expressly cited the source.

By filling out the following fields, you are signing this pledge. No assignment will get credit without being pledged.

Name:Muhammet Furkan Isik

CWID:10472193

Date: 12/11/2021

Instructions

In this assignment, you should use R markdown to answer the questions below. Simply type your R code into embedded chunks as shown above. When you have completed the assignment, knit the document into a PDF file, and upload both the .pdf and .Rmd files to Canvas.

```
CWID = -1 #Place here your Campus wide ID number, this will personalize
#your results, but still maintain the reproduceable nature of using seeds.
#If you ever need to reset the seed in this assignment, use this as your seed
#Papers that use -1 as this CWID variable will earn 0's so make sure you chan
ge
#this value before you submit your work.
personal = CWID %% 10000
set.seed(personal)#You can reset the seed at any time in your code,
#but please always set it to this seed.
```

1 point for every item of every question. Total = 22. There is a final extra question (2 points).

Pilgrim Bank.

This exercise is based on the case Pilgrim Bank A (602104), Harvard Business School. In order to buy this case, you must register in the HBS website following this link: https://hbsp.harvard.edu/import/859412.

You must read the case to understand the main problem proposed that would help you to answer the questions in the proper way.

Using the dataset pilgrim.csv from the Pilgrim Bank case, please answer the following questions to evaluate the impact of the online channel and if its adoption requires pay a rebate or receive a fee from the customers. The dataset uses the following convention: variables xxx9 and xxx0 refer to 1999 and 2000 respectively. Observations of 2000 with missing observations are from customers that have already left the bank.

You can answer most of the questions until 5.c. using linear regression (OLS). The program should be written in R.

1. Calculate average customer profitability with 95% confidence level

```
df <- read.csv("/Users/metuhead/Desktop/FA590/HW4/pilgrim.csv")
df_1 <- na.omit(df)
data=df_1

library(tidyr)
p9 <- data$Profit9
p9 <- p9[!is.na(p9)]
p9.mean <- mean(p9)
p9.std <- sd(p9)
n1 <- length(p9)
p0 <- data$Profit0
p0 <- p0[!is.na(p0)]
p0.mean <- mean(p0)
p0.std <- sd(p0)
n2 <- length(p0)
t1 <- qt(p = 0.95,
df = n1 - 1,
lower.tail = T)
t2 <- qt(p = 0.95,
df = n2 - 1,
lower.tail = T)
conf_lvl <- function(n, mean, std, t) {
u <- mean + (t * std / sqrt(n))
l <- mean - (t * std / sqrt(n))
return(c(l, u))
}
ul9 <- conf_lvl(n1, p9.mean, p9.std, t1)
ul0 <- conf_lvl(n2, p0.mean, p0.std, t2)
conf.table <- data.frame(ul9, ul0)
rownames(conf.table) <- c('low', 'high')
```

```r
colnames(conf.table) <- c('1999', '2000')
conf.table
```

```
##          1999    2000
## low  126.6479 149.905
## high 133.1100 158.884
```

2.a. Evaluate if online channel has a significant impact on 1999 profitability (Profit9).

- According to t-test p value is 0.2254, then null hypothesis can not be rejected
- Null hypothesis : Mean of group 1= mean of group 2
- Hence, online channel usage is significant
- Moreover, regression p value for Variable Online is 0.21 not significant

```r
t.test(data$Profit9 ~ data$Online9, mu=0, alt="two.sided", conf=0.95, var.eq=
F, paired=F)
```

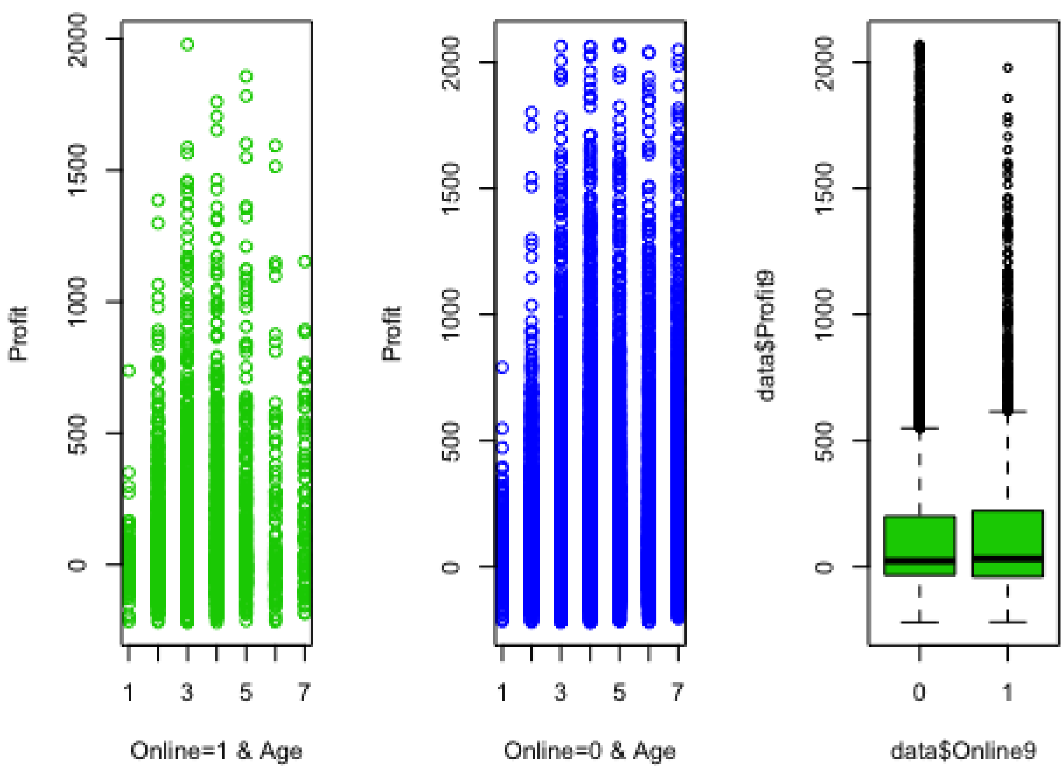
```
##
##  Welch Two Sample t-test
##
## data:  data$Profit9 by data$Online9
## t = -1.2909, df = 3494.6, p-value = 0.1968
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -19.61678   4.04068
## sample estimates:
## mean in group 0 mean in group 1
##        128.8771        136.6652
```

```r
summary(lm(Profit9~Online9, data=data))
```

```
##
## Call:
## lm(formula = Profit9 ~ Online9, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -356.67 -162.88 -105.88   73.12 1942.12
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  128.877      2.104  61.248   <2e-16 ***
## Online9        7.788      5.867   1.327    0.184
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 285.2 on 21081 degrees of freedom
## Multiple R-squared:  8.358e-05,  Adjusted R-squared:  3.615e-05
## F-statistic: 1.762 on 1 and 21081 DF,  p-value: 0.1844
```

- Descriptive Statistics

```
par(mfrow=c(1,3))

plot( data$Age9[data$Online9==1],data$Profit9[data$Online9==1],col=3, xlab= "
Online=1 & Age ", ylab= "Profit")
plot( data$Age9[data$Online9==0],data$Profit9[data$Online9==0],col=4 , xlab=
" Online=0 & Age ", ylab= "Profit")

boxplot(data$Profit9~data$Online9, col=3)
```



2.b. Does age help to explain if online channel has a significant impact on 1999 profitability?

- According to regression summary, yes Age hels to expalin that online channel has signifact impact since the p value is too small close to zero

```
online_lm <- lm(Profit9~Age9+Online9, data = df_1)
summary(online_lm)

##
## Call:
## lm(formula = Profit9 ~ Age9 + Online9, data = df_1)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -420.76 -163.15  -90.51   71.29 1964.85
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   21.235      5.507   3.856 0.000116 ***
## Age9          25.640      1.214  21.115  < 2e-16 ***
## Online9       28.688      5.890   4.871 1.12e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 282.2 on 21080 degrees of freedom
## Multiple R-squared:  0.02079,    Adjusted R-squared:  0.0207
## F-statistic: 223.8 on 2 and 21080 DF,  p-value: < 2.2e-16

#Yes , Age and Online9 are impactful for profitability
```

3. To adjust for missing observations in the case of the variables Age9 and Inc9 (income) and adjust other variables:

```
- Substitute missing observations with zeros: create variables Age0 and Inc0
- Substitute missing observations with averages: create variables AgeAvg and
IncAvg
- Include additional dummy variables where 1 if there is data and 0 otherwise
: create variables AgeExist and IncExist (define as factor variable).
- Retain takes a value of 0 when Profit0 has a missing observation and 1 othe
rwise: create variable retainD (define as factor variable).
- Create dummy variables D1100 and D1200 for districts 1100 and 1200 respecti
vely from the variable District9 (define as factor variables).
- Variables Online9, Billpay9, Online0, Billpay0 should be defined as factor
variables.
```

To test for bias of missing data, evaluate if missing data has an effect on profitability analysis: 3a. Evaluate the effect of online channel on 1999 profits when Age0 is included. 3b. Evaluate if adjusting missing data using Age0 or AgeAvg is relevant. In both cases, it is still necessary to include the additional variable AgeExist to control for the missing data 3c. Repeat above steps with income. Evaluate if adjusting missing data using Inc0 or IncAvg is relevant.

```
df$Age0 <- df$Age9
df$Age0[is.na(df$Age0)] <- 0


df$Inc0 <- df$Inc9
df$Inc0[is.na(df$Inc0)] <- 0
```

```
df$AgeAvg[is.na(df$Age9)]<-mean(df$Age9,na.rm=TRUE)


df$AgeAvg <- df$Age9
df$AgeAvg[is.na(df$Age9)]<-mean(df$Age9,na.rm=TRUE)


df$IncAvg <- df$Inc9
df$IncAvg[is.na(df$Inc9)]<-mean(df$Inc9,na.rm=TRUE)


AgeExist <- ifelse(is.na(df$Age9),0,1)
df$AgeExist <- as.factor(AgeExist)

IncExist <- ifelse(is.na(df$Inc9),0,1)
df$IncExist <- as.factor(IncExist)

retainD <- ifelse(is.na(df$Profit0),0,1)
df$retainD <- as.factor(retainD)

D1100 <- ifelse(df$District9 == 1100,1,0)
df$D1100 <- as.factor(D1100)

D1200 <- ifelse(df$District9 == 1200,1,0)
df$D1200 <- as.factor(D1200)

fact_cols <- c("Online9","Billpay9","Online0","Billpay0")

#df[fact_cols] <- as.factor(df[fact_cols])


df[,fact_cols] <- lapply(df[,fact_cols], as.factor)
```

3a. Evaluate the effect of online channel on 1999 profits when Age0 is included

- According to Regression summary,
- Model is Profit9= 57+13.8 (Online9) + 17.7 (Age0) was obtained.
- All the variables are signifacant since the p values quite close to 0
- Age0 helps explaining the effect of online channel on 1999 profits

```
online_lm <- lm(Profit9~Age0+Online9, data = df)
summary(online_lm)

##
## Call:
## lm(formula = Profit9 ~ Age0 + Online9, data = df)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max 
## -393.91 -147.07  -82.03   49.97 1976.97 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  57.0311     2.6014  21.923  < 2e-16 ***
## Age0         17.6803     0.6697  26.402  < 2e-16 ***
## Online91     13.7925     4.6487   2.967  0.00301 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 269.9 on 31631 degrees of freedom
## Multiple R-squared:  0.02161,    Adjusted R-squared:  0.02155 
## F-statistic: 349.3 on 2 and 31631 DF,  p-value: < 2.2e-16
```

3b. Evaluate if adjusting missing data using Age0 or AgeAvg is relevant. In both cases, it is still necessary to include the additional variable AgeExist to control for the missing data

- The coefficients are significantly not zero sincce p values are quite close to 0.
- In comparison to model in 3a, a higher R-squared was obtained. Hence, this model better explains the variation.
- Equation of Profit9= 70.9+19.6 (Online9) + 25.6 (Age0) – 51.85(AgeExist) was obtained.
- When we just use Age0 and AgeExist, both are relevant to predict Profit
- When we just use AgeAvg and AgeExist, both are relevant to predict Profit9

```
Age0_lm <- lm(Profit9~Age0+AgeExist, data = df)
summary(Age0_lm)

## 
## Call:
## lm(formula = Profit9 ~ Age0 + AgeExist, data = df)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -404.86 -144.10  -82.98   51.95 1963.90 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   72.962      2.961  24.640   <2e-16 ***
## Age0          24.939      1.074  23.212   <2e-16 ***
## AgeExist1    -48.682      5.548  -8.775   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 269.6 on 31631 degrees of freedom
## Multiple R-squared:  0.02372,    Adjusted R-squared:  0.02365 
## F-statistic: 384.2 on 2 and 31631 DF,  p-value: < 2.2e-16
```

```
AgeAvg_lm <- lm(Profit9~AgeAvg+AgeExist, data = df)
summary(AgeAvg_lm)

##
## Call:
## lm(formula = Profit9 ~ AgeAvg + AgeExist, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -404.86 -144.10  -82.98   51.95 1963.90
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -27.944      5.260  -5.313 1.09e-07 ***
## AgeAvg        24.939      1.074  23.212  < 2e-16 ***
## AgeExist1     52.224      3.447  15.151  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 269.6 on 31631 degrees of freedom
## Multiple R-squared:  0.02372,    Adjusted R-squared:  0.02365
## F-statistic: 384.2 on 2 and 31631 DF,  p-value: < 2.2e-16
```

3c. Repeat above steps with income. Evaluate if adjusting missing data using Inc0 or IncAvg is relevant. Include AgeExist and AgeAvg in the calculations.

- When we include Inc0 and Online to predict Profit9, Online9 is not relevant due to higher p value
- When we just use Inc0 and IncExist, both are relevant to predict Profit9
- When we just use IncAvg and IncExist, both are relevant to predict Profit9

```
onlineInc_lm <- lm(Profit9~Inc0+Online9, data = df)
summary(onlineInc_lm)

##
## Call:
## lm(formula = Profit9 ~ Inc0 + Online9, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -397.48 -148.50  -80.71   47.30 1975.28
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  58.7162     2.5071  23.420   <2e-16 ***
## Inc0         13.1962     0.4853  27.192   <2e-16 ***
## Online91     -3.5896     4.6491  -0.772     0.44
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 269.7 on 31631 degrees of freedom
## Multiple R-squared:  0.02289,    Adjusted R-squared:  0.02283
## F-statistic: 370.5 on 2 and 31631 DF,  p-value: < 2.2e-16
```

```r
Inc0_lm <- lm(Profit9~Inc0+IncExist, data = df)
summary(Inc0_lm)
```

```
##
## Call:
## lm(formula = Profit9 ~ Inc0 + IncExist, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -408.41 -145.99  -80.71   47.72 1992.29
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   71.365      2.965  24.073  < 2e-16 ***
## Inc0          17.713      0.751  23.586  < 2e-16 ***
## IncExist1    -42.365      5.357  -7.908  2.7e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 269.4 on 31631 degrees of freedom
## Multiple R-squared:  0.0248, Adjusted R-squared:  0.02474
## F-statistic: 402.2 on 2 and 31631 DF,  p-value: < 2.2e-16
```

```r
IncAvg_lm <- lm(Profit9~IncAvg+IncExist, data = df)
summary(IncAvg_lm)
```

```
##
## Call:
## lm(formula = Profit9 ~ IncAvg + IncExist, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -408.41 -145.99  -80.71   47.72 1992.29
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -25.325      5.059  -5.006 5.59e-07 ***
## IncAvg        17.713      0.751  23.586  < 2e-16 ***
## IncExist1     54.324      3.449  15.751  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 269.4 on 31631 degrees of freedom
## Multiple R-squared:  0.0248, Adjusted R-squared:  0.02474
## F-statistic: 402.2 on 2 and 31631 DF,  p-value: < 2.2e-16
```

4.a. Evaluate if online channel has a significant impact on 1999 profitability after controlling for demographic variables: age, income, tenure, and geographic district. You can evaluate the impact of geographic district using the dummy variables D1100 and D1200.

- The coefficients are significantly not zero.
- The R-squared increases even more than the model in 3c
- We find that Online9 is significant due to its p-value which is 0.00271

```
dem_lm <- lm(Profit9~Online9+Age0+AgeExist+Inc0+IncExist+Tenure9+D1100+D1200,
data = df)
summary(dem_lm)

##
## Call:
## lm(formula = Profit9 ~ Online9 + Age0 + AgeExist + Inc0 + IncExist +
##     Tenure9 + D1100 + D1200, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -487.17 -141.21  -65.88   48.87 1993.27
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  23.2098     5.0959   4.555 5.27e-06 ***
## Online91     13.8233     4.6091   2.999  0.00271 **
## Age0         16.6701     1.1482  14.519  < 2e-16 ***
## AgeExist1   -63.0567     9.1844  -6.866 6.74e-12 ***
## Inc0         16.8530     0.7554  22.310  < 2e-16 ***
## IncExist1   -57.1191     8.9956  -6.350 2.19e-10 ***
## Tenure9       4.7464     0.1918  24.742  < 2e-16 ***
## D11001       -7.9955     6.2582  -1.278  0.20140
## D12001       13.1986     4.4734   2.950  0.00318 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 264.2 on 31625 degrees of freedom
## Multiple R-squared:  0.06234,    Adjusted R-squared:  0.0621
## F-statistic: 262.8 on 8 and 31625 DF,  p-value: < 2.2e-16
```

5.a. Evaluate the drivers of customer profitability for the year 2000 (Hint: you can evaluate the variables explored for profitability of 1999).

- Except from district demographics and Age0, everything is useful to predict profit0

```r
#Due to NA values in Profit0 we impute the median values

summary(df$Profit0)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## -5643.0   -30.0    23.0   144.8   206.0 27086.0    5238

#Median is 23.0; we don't choose mean because it is right skewed


df$Profit0[is.na(df$Profit0)] <- 23.0


profit0_lm <- lm(Profit0~Online9+Tenure9+Age0+AgeAvg+IncAvg+Inc0+D1100+D1200,
data = df)
summary(profit0_lm)

##
## Call:
## lm(formula = Profit0 ~ Online9 + Tenure9 + Age0 + AgeAvg + IncAvg +
##     Inc0 + D1100 + D1200, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5919.8  -145.1   -62.8    26.5 26811.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -79.4275     9.7040  -8.185 2.82e-16 ***
## Online91     25.6657     6.1413   4.179 2.93e-05 ***
## Tenure9       4.5015     0.2556  17.611  < 2e-16 ***
## Age0          2.7286     2.7010   1.010 0.312393
## AgeAvg       10.2573     3.0246   3.391 0.000697 ***
## IncAvg       12.2929     2.1958   5.598 2.18e-08 ***
## Inc0          8.0082     2.0028   3.999 6.39e-05 ***
## D11001      -12.0921     8.3387  -1.450 0.147037
## D12001        9.5609     5.9606   1.604 0.108724
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 352.1 on 31625 degrees of freedom
## Multiple R-squared:  0.03904,    Adjusted R-squared:  0.0388
## F-statistic: 160.6 on 8 and 31625 DF,  p-value: < 2.2e-16
```

#5.b. Evaluate if the variable Profit9 should be included in the customer profitability analysis for 2000.

- Adding Profit 9, increases the Adjusted R-squared value dramatically from 0.03417 to 0.3613. Therefore, it should definitely be added to the model. However, some of the variables become not statistically significant.

- Hence, it's wise to create another model definitely including Profit9, and removing non-significant variables, which gives us Adjusted R-squared: 0.3614 value

```
profit0_9lm <- lm(Profit0~Profit9+Online9+AgeAvg+IncAvg+Tenure9+Age0+Inc0+D11
00+D1200, data = df)
summary(profit0_9lm)

##
## Call:
## lm(formula = Profit0 ~ Profit9 + Online9 + AgeAvg + IncAvg +
##     Tenure9 + Age0 + Inc0 + D1100 + D1200, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6806.4   -72.3   -23.8    32.0 26901.5
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.287442   8.172533  -1.136  0.25579
## Profit9      0.723347   0.006293 114.953  < 2e-16 ***
## Online91    15.666707   5.158427   3.037  0.00239 **
## AgeAvg      -1.015874   2.542053  -0.400  0.68943
## IncAvg       4.724011   1.845254   2.560  0.01047 *
## Tenure9      1.068148   0.216739   4.928 8.34e-07 ***
## Age0         1.943552   2.268388   0.857  0.39156
## Inc0         3.386565   1.682460   2.013  0.04414 *
## D11001      -6.308569   7.003305  -0.901  0.36770
## D12001       0.013718   5.006618   0.003  0.99781
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 295.7 on 31624 degrees of freedom
## Multiple R-squared:  0.3222, Adjusted R-squared:  0.3221
## F-statistic:  1671 on 9 and 31624 DF,  p-value: < 2.2e-16
```

5.c. Forecast customer profitability of the test sample for 2000 after adding electronic billpay and evaluate the most important variables using OLS.

- According to summary, these are the statistically significant variables: retainD1, Tenure9 ,Age0 ,AgeExist1,Online01 ,Inc0

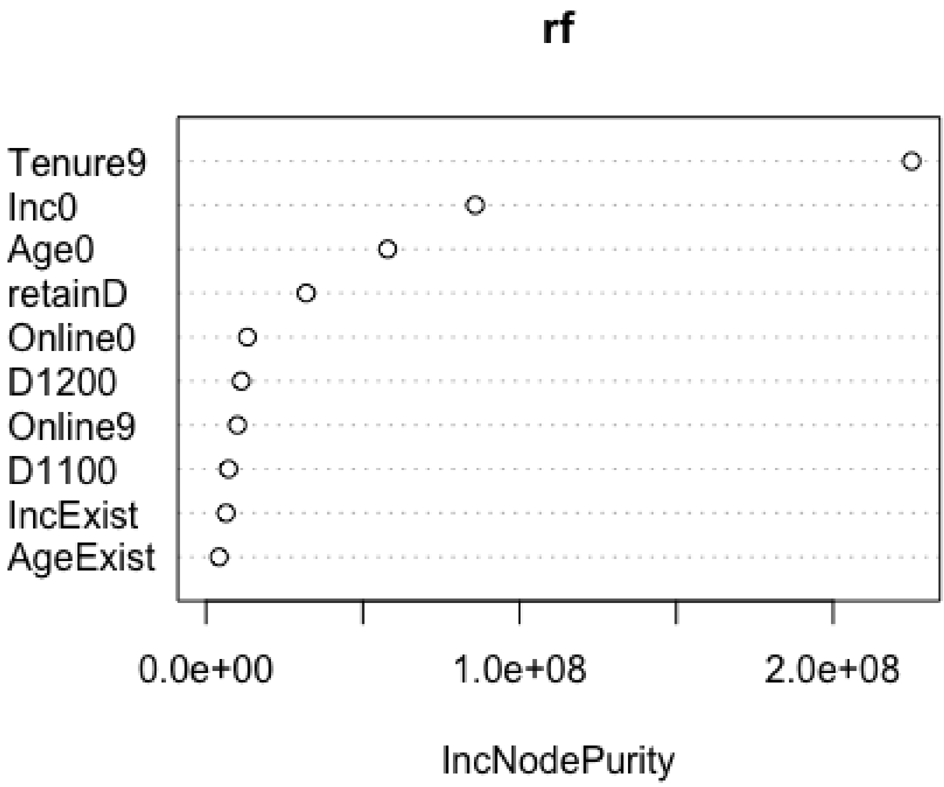
```
#Split the data in 2/3 training and 1/3 testing.

#Train Test Split

df$Online0[is.na(df$Online0)] <- 0
df$Billpay0[is.na(df$Billpay0)] <- 0
```

```r
train = df[1:21099,]
test = df[21100:31634,]

#Online0 has NA values, changing these to 0;

profit0_lm <- lm(Profit0~Online9+retainD+Tenure9+Age0+AgeExist+Online0+Inc0+IncExist+D1100+D1200, data = train)
summary(profit0_lm)

##
## Call:
## lm(formula = Profit0 ~ Online9 + retainD + Tenure9 + Age0 + AgeExist +
##     Online0 + Inc0 + IncExist + D1100 + D1200, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -914.0  -150.0   -58.3    36.3 14678.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -19.2534     8.1975  -2.349 0.018848 *
## Online91      5.3428     8.4091   0.635 0.525198
## retainD1     84.6107     6.6195  12.782  < 2e-16 ***
## Tenure9       4.2395     0.2805  15.115  < 2e-16 ***
## Age0         12.6044     1.6852   7.480 7.75e-14 ***
## AgeExist1   -52.8075    13.6624  -3.865 0.000111 ***
## Online01     25.7761     7.5651   3.407 0.000657 ***
## Inc0         19.7936     1.1058  17.900  < 2e-16 ***
## IncExist1   -82.8764    13.3988  -6.185 6.31e-10 ***
## D11001      -13.6859     9.0910  -1.505 0.132227
## D12001       10.3102     6.4724   1.593 0.111184
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 316.2 on 21088 degrees of freedom
## Multiple R-squared:  0.05616,    Adjusted R-squared:  0.05571
## F-statistic: 125.5 on 10 and 21088 DF,  p-value: < 2.2e-16

profit0_testpreds <- predict(profit0_lm,test)
lm_testmse <- mean((test$Profit0 - profit0_testpreds)^2)
cat("Test MSE of Linear Regression is:", lm_testmse,'\n')

## Test MSE of Linear Regression is: 169099.7
```

5.d. Forecast customer profitability of the test sample for 2000 after adding electronic billpay and evaluate the most important variables using any nonlinear machine learning algorithm.

- According to variables importance plot, important variables are Tenure 9, Inc0, Age0, retain ID, Online 0

```
library(randomForest)

rf <- randomForest(Profit0~Online9+retainD+Tenure9+Age0+AgeExist+Online0+Inc0
+IncExist+D1100+D1200, data = train,mtry=3,ntree=100)
importance(rf)

##            IncNodePurity
## Online9         10029263
## retainD         31956025
## Tenure9        225033534
## Age0            57925902
## AgeExist         4181691
## Online0         13193504
## Inc0            85875110
## IncExist         6410485
## D1100            7170284
## D1200           11268597

varImpPlot(rf)
```



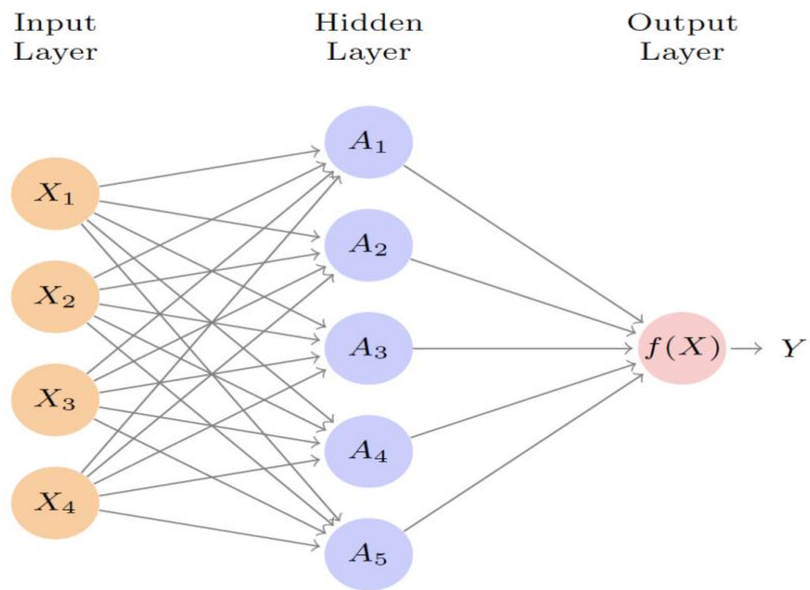

```
rf_pred <- predict(rf,test)

rf_mse <- mean((test$Profit0 - rf_pred)^2)
cat("Test MSE for Random Forest:",rf_mse)

## Test MSE for Random Forest: 170201
```

#5.e. Which one provides the best ranking for these variables? why?

- Comparing these two models OLS and Random Forest, since OLS gives lower MSE, it's wise to say OLS provides better ranking for these variables

5.f. Forecast customer profitability of the test sample for 2000 after adding electronic billpay using 1 layer neural network (NN).

10.1 Single Layer Neural Networks



Neural Networks

```
library(neuralnet)


cols_tonum <- c("retainD","Online9","Age0","AgeExist","Online0","Inc0","IncEx
ist","D1100","D1200","Billpay0","Billpay9")
train[,cols_tonum] <- lapply(train[,cols_tonum], as.numeric)
test[,cols_tonum] <- lapply(test[,cols_tonum], as.numeric)

nn = neuralnet(Profit0~Online9+retainD+Tenure9+Age0+AgeExist+Online0+Inc0+Inc
Exist+D1100+D1200+Billpay0,data = train,hidden = 1, linear.output = F)


nn1_preds <- compute(nn, test)
nn1_mse <- mean((test$Profit0 - nn1_preds$net.result)^2)
cat("Test MSE for NN with 1 layers:",nn1_mse)

## Test MSE for NN with 1 layers: 191258
```

5.g. Forecast customer profitability of the test sample for 2000 after adding electronic billpay using 2 layers neural network (NN).

```
nn2 = neuralnet(Profit0~Online9+retainD+Tenure9+Age0+AgeExist+Online0+Inc0+In
cExist+D1100+D1200+Billpay0,data = train,hidden =c(4,2), linear.output = F)


nn2_preds <- compute(nn2, test)
nn2_mse <- mean((test$Profit0 - nn2_preds$net.result)^2)
cat("Test MSE for NN with 2 layers:",nn2_mse)

## Test MSE for NN with 2 layers: 191258
```

5.h. Build a table with the mean squared error (MSE) of these 4 methods. Discuss your results.

- All the model gives quite high MSE results
- According to the table, Linear Regression gives the lowest MSE value followed by Random Forest, and Neurelnet.
- Since, Linear regression less computationally expensive and more interperatable it’s wise to choose Linear regression among all those models.
- Among non linear model, it’s wise to choose random forrest since also it’s computationally less expensive and more explanory

```
x = c(lm_testmse,rf_mse,nn1_mse,nn2_mse)
y = c("Linear Regression","Random Forest","NN1","NN2")

table_MSE <- data.frame(x,y)
names(table_MSE) <- c("MSE","Model")
table_MSE

##        MSE             Model
## 1 169099.7 Linear Regression
## 2 170201.0     Random Forest
## 3 191258.0               NN1
## 4 191258.0               NN2
```

Forecast customer retention for the year 2000 using the variables Online9, Billpay9, Online0, Billpay0 and the following algorithms:

6.a. Naive Bayes. Hint: use library(e1071)

- NB is slightly better than NN and we prefer NB because it is much less complex thaa a NN
- Accuracy : 32.8% 1760+1696 / 10535

```
library(e1071)

nb <- naiveBayes(retainD~Online9+Billpay9+Online0+Billpay0, data = train)
```

```
nb.results <- predict(nb,test)
nbresultsdf <- data.frame(actual = test$retainD, prediction = nb.results)
table(nbresultsdf$actual,nbresultsdf$prediction)

##
##         1    2
##    1 1760    7
##    2 7072 1696
```

6.b Neural networks

- Accuracy = 32.017% 1579+1794 / 10535

```
nn_retainD <- neuralnet(retainD~Online9+Billpay9+Online0+Billpay0, data = tra
in, hidden=c(1,3),linear.output=F,threshold=0.3)

#Test the resulting output
nn.results <- compute(nn_retainD, test)
results <- data.frame(actual = test$retainD, prediction = ifelse(nn.results$n
et.result > 0.999999516907582,2,1))
attach(results)
table(actual,prediction)

##       prediction
## actual    1
##      1 1767
##      2 8768
```

6.c. Compare their accuracy and explain why one of these methods is more appropriate for this problem.

- NB is slightly better than NN and we prefer NB because it is much less complex than a NN

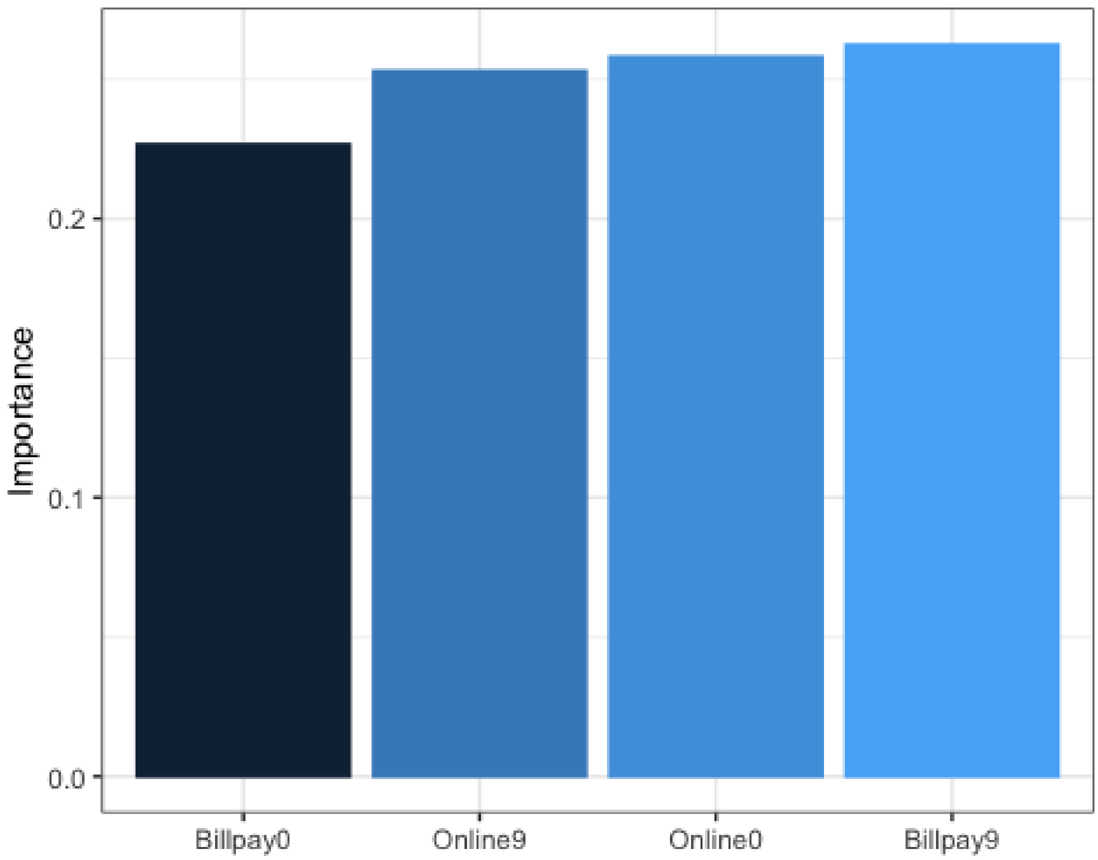
```
library(e1071)

nn_retainD <- neuralnet(retainD~Online9+Billpay9+Online0+Billpay0, data = tra
in, hidden=c(1,3),linear.output=F,threshold=0.3)

#Test the resulting output
nn.results <- compute(nn_retainD, test)
results <- data.frame(actual = test$retainD, prediction = ifelse(nn.results$n
et.result > 0.999999516907582,2,1))
attach(results)
table(actual,prediction)

##       prediction
## actual    1
```

```
##      1 1767
##      2 8768

#ACcuracy = 1579+1794 / 10535: 32.017%


nb <- naiveBayes(retainD~Online9+Billpay9+Online0+Billpay0, data = train)

nb.results <- predict(nb,test)
nbresultsdf <- data.frame(actual = test$retainD, prediction = nb.results)
table(nbresultsdf$actual,nbresultsdf$prediction)

##
##         1    2
##   1 1760    7
##   2 7072 1696

#Accuracy : 1760+1696 / 10535: 32.8%
```

7. Evaluate the effect of the online channel and billpay on customer's retention with the variables Online9, Billpay9, Online0, Billpay0 using neural networks. Hint: use the function garson from the package NeuralNetTools.

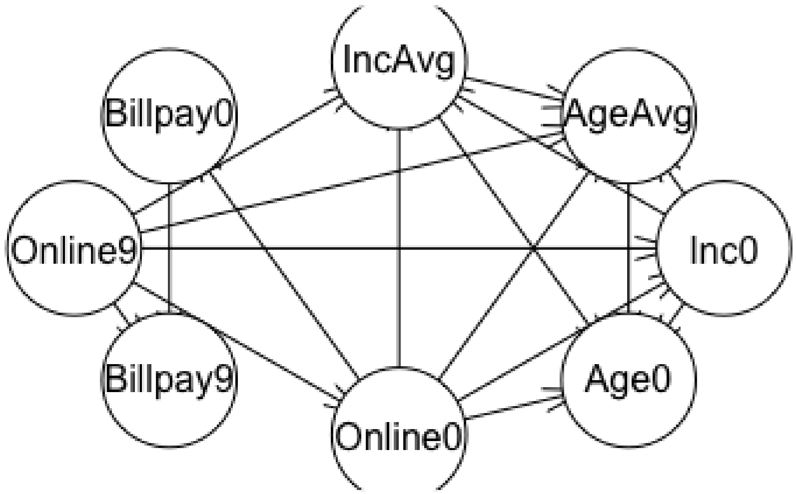
- According the barplot, we can see that Online0 is the most important and followed by Billpa9, Online9, Billpay 0

```
library("NeuralNetTools")



#nn_online <- neuralnet(retainD~Online9+Billpay9+Online0+Billpay0, data = tra
in, hidden = c(1,3), linear.output = T)
nn_online <- neuralnet(retainD~Online9+Billpay9+Online0+Billpay0, data = trai
n,hidden=50,threshold=0.01, linear.output=F)

garson(nn_online, bar_plot=T)
```



Nonprogramming question: You neither have to write a program nor make any direct calculations, only interpret the results of your previous calculations. # 8.a. Draw a Bayesian network that represents the main drivers of profitability for 2000 and customers' retention using the previous information.

```
library(bnlearn)

Pilgrim3= df[, c("Billpay0", "Online9", "Billpay9","Online0","Age0","Inc0","A
geAvg","IncAvg")]


bn_df = data.frame(Pilgrim3)

res= hc(bn_df)

plot(res)
```

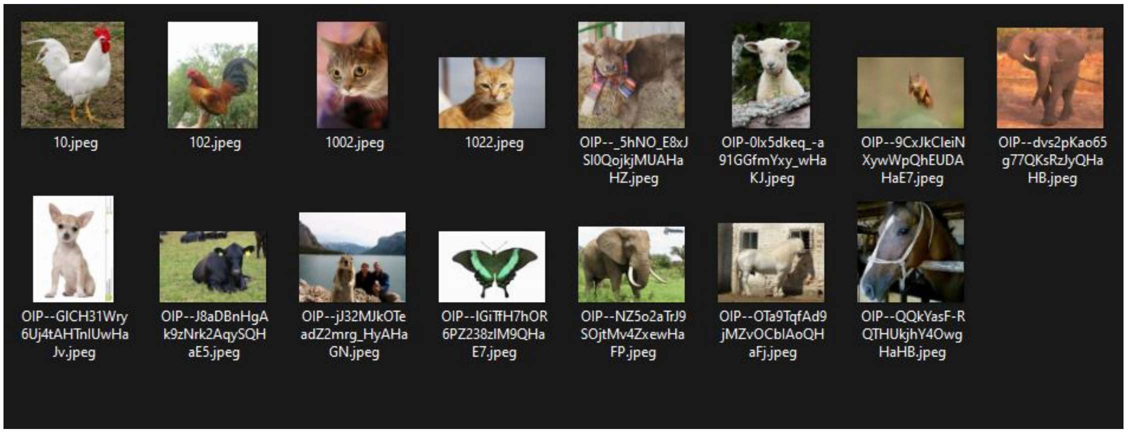


8.b Justify your Bayesian network and evaluate if the adoption of the online channel requires pay a rebate or receive a fee from the customers.

- We see that Online and Billpay plays importan role, so definitely adoption of the online channel requires pay a rebate.

Extra exercise (2 extra points). This is a completely optional exercise that requires the installation of the keras library following these instructions:
https://web.stanford.edu/~hastie/ISLR2/keras-instructions.html

9. Select 15 images of animals (such as dogs, cats, birds, farm animals, etc.). If the subject does not occupy a reasonable part of the image, then crop the image. Use a pretrained image classification CNN as in Lab 10.9.4 to predict the class of each of your images, and report the probabilities for the top five predicted classes for each image.




```r
library(tensorflow)
library(keras)
img_dir <- "./imgs"
image_names <- list.files(img_dir)
num_images <- length(image_names)
x <- array(dim = c(num_images, 224, 224, 3))
for (i in 1:num_images) {
  img_path <- paste(img_dir, image_names[i], sep = "/")
  img <- image_load(img_path, target_size = c(224, 224))
  x[i,,, ] <- image_to_array(img)
}
x <- imagenet_preprocess_input(x)
###
model <- application_resnet50(weights = "imagenet")
# summary(model)
###
pred6 <- model %>% predict(x) %>%  imagenet_decode_predictions(top = 5)
names(pred6) <- image_names
print(pred6)

## $`10.jpeg`
##   class_name class_description        score
## 1  n01514668              cock 0.5868831873
## 2  n01514859               hen 0.3733178973
## 3  n01855672             goose 0.0258939303
## 4  n01847000             drake 0.0087186862
## 5  n01807496         partridge 0.0008939427
##
## $`1002.jpeg`
##   class_name class_description       score
## 1  n02123045             tabby 0.560591280
## 2  n02123159         tiger_cat 0.219697207
## 3  n02124075      Egyptian_cat 0.168887615
```

```
## 4  n02127052              lynx 0.019455157
## 5  n02123394       Persian_cat 0.005819479
##
## $`102.jpeg`
##   class_name class_description        score
## 1  n01514668              cock 0.8487553596
## 2  n01514859               hen 0.1478639394
## 3  n01807496         partridge 0.0017673468
## 4  n01824575            coucal 0.0003589727
## 5  n01818515             macaw 0.0002182447
##
## $`1022.jpeg`
##   class_name class_description       score
## 1  n02124075      Egyptian_cat 0.810525239
## 2  n02123045             tabby 0.085191980
## 3  n02123159         tiger_cat 0.068304740
## 4  n02127052              lynx 0.020959303
## 5  n02123597       Siamese_cat 0.003721192
##
## $`OIP--_5hNO_E8xJSI0QojkjMUAHaHZ.jpeg`
##   class_name class_description      score
## 1  n02112137              chow 0.31188142
## 2  n02099601  golden_retriever 0.14063577
## 3  n02102480    Sussex_spaniel 0.12096537
## 4  n02111277      Newfoundland 0.05717488
## 5  n02094258   Norwich_terrier 0.02394269
##
## $`OIP--9CxJkCleiNXywWpQhEUDAHaE7.jpeg`
##   class_name class_description      score
## 1  n02422106        hartebeest 0.31105024
## 2  n02115913             dhole 0.30901545
## 3  n02454379         armadillo 0.06296137
## 4  n02356798      fox_squirrel 0.04911088
## 5  n01798484   prairie_chicken 0.03535118
##
## $`OIP--dvs2pKao65g77QKsRzJyQHaHB.jpeg`
##   class_name class_description        score
## 1  n02504458  African_elephant 5.189098e-01
## 2  n01871265            tusker 3.858417e-01
## 3  n02504013   Indian_elephant 9.522477e-02
## 4  n02397096           warthog 1.319695e-05
## 5  n02408429     water_buffalo 2.981413e-06
##
## $`OIP--GlCH31Wry6Uj4tAHTnIUwHaJv.jpeg`
##   class_name class_description        score
## 1  n02085620         Chihuahua 9.938471e-01
## 2  n02108915    French_bulldog 3.013794e-03
## 3  n02113978  Mexican_hairless 2.182680e-03
## 4  n02087046       toy_terrier 5.984782e-04
## 5  n02123597       Siamese_cat 8.282105e-05
```

```
##
## $`OIP--J8aDBnHgAk9zNrk2AqySQHaE5.jpeg`
##   class_name     class_description      score
## 1  n02088094          Afghan_hound 0.21020076
## 2  n02099267 flat-coated_retriever 0.16559061
## 3  n04604644            worm_fence 0.11866640
## 4  n02403003                    ox 0.07231194
## 5  n02105056           groenendael 0.06030094
##
## $`OIP--jJ32MJkOTeadZ2mrg_HyAHaGN.jpeg`
##   class_name class_description       score
## 1  n02361337            marmot 0.408747375
## 2  n09332890          lakeside 0.396046877
## 3  n02437616             llama 0.132158026
## 4  n02437312     Arabian_camel 0.009655411
## 5  n09246464             cliff 0.009020077
##
## $`OIP--lGiTfH7hOR6PZ238zlM9QHaE7.jpeg`
##   class_name class_description      score
## 1  n03532672              hook 0.12868321
## 2  n04599235              wool 0.08186767
## 3  n01795545      black_grouse 0.06289475
## 4  n02281787          lycaenid 0.04305092
## 5  n03888257         parachute 0.04219106
##
## $`OIP--NZ5o2aTrJ9SOjtMv4ZxewHaFP.jpeg`
##   class_name class_description        score
## 1  n01871265            tusker 6.628298e-01
## 2  n02504013   Indian_elephant 1.718147e-01
## 3  n02504458  African_elephant 1.653512e-01
## 4  n02397096           warthog 1.261633e-06
## 5  n02408429     water_buffalo 5.870044e-07
##
## $`OIP--OTa9TqfAd9jMZvOCbIAoQHaFj.jpeg`
##   class_name class_description      score
## 1  n02437616             llama 0.31493527
## 2  n01514668              cock 0.23414181
## 3  n02105505          komondor 0.11728832
## 4  n02437312     Arabian_camel 0.06234841
## 5  n02097474   Tibetan_terrier 0.04351031
##
## $`OIP--QQkYasF-RQTHUkjhY4OwgHaHB.jpeg`
##   class_name class_description       score
## 1  n02389026            sorrel 0.676176608
## 2  n03538406        horse_cart 0.300504297
## 3  n03868242            oxcart 0.006643793
## 4  n02795169            barrel 0.005548853
## 5  n03803284            muzzle 0.003036232
##
## $`OIP-0Ix5dkeq_-a91GGfmYxy_wHaKJ.jpeg`
```

```
##   class_name  class_description      score
## 1  n02134084           ice_bear 0.22257978
## 2  n02093647 Bedlington_terrier 0.19829135
## 3  n02113799    standard_poodle 0.11947756
## 4  n13044778          earthstar 0.07073054
## 5  n02114548         white_wolf 0.03674563
```