|  |  |  |
| --- | --- | --- |
| **Code** | **Output** | **Comment** |
| #opening file  library(readxl)  df <- read\_excel("Documents/HULT Academic/DSR/datasets\_marketing\_campaign\_SF.xlsx")  #Removings rows with NA values  #install.packages("tidyr")  library(tidyr)  df\_clean <- df %>% drop\_na()  #Recoding chr variables into numeric variables for correlation matrix for Q1  df\_clean$education\_num <- revalue(df\_clean$Education, c("2n Cycle"=1,"Basic"=2,"Graduation"=3,"Master"=4,"PhD"=5))  df\_clean$education\_num <- mapvalues(df\_clean$Education, from = c("2n Cycle","Basic","Graduation","Master","PhD"), to = c("1", "2","3","4","5"))  df\_clean$marstatus\_num <- revalue(df\_clean$Marital\_Status, c("Alone"=1,"Divorced"=2,"Married"=3,"Single"=4,"Together"=5,"Widow"=6))  df\_clean$marstatus\_num <- mapvalues(df\_clean$Marital\_Status, from = c("Alone","Divorced","Married","Single","Together","Widow"), to = c("1", "2","3","4","5","6"))  df\_clean$marstatus\_num <- as.numeric(df\_clean$marstatus\_num)  df\_clean$education\_num <- as.numeric(df\_clean$education\_num) | > #opening file  > library(readxl)  > df <- read\_excel("Documents/HULT Academic/DSR/datasets\_marketing\_campaign\_SF.xlsx")  >  > #Removings rows with NA values  > #install.packages("tidyr")  > library(tidyr)  > df\_clean <- df %>% drop\_na()  >  > #Recoding chr variables into numeric variables for correlation matrix for Q1  > df\_clean$education\_num <- revalue(df\_clean$Education, c("2n Cycle"=1,"Basic"=2,"Graduation"=3,"Master"=4,"PhD"=5))  > df\_clean$education\_num <- mapvalues(df\_clean$Education, from = c("2n Cycle","Basic","Graduation","Master","PhD"), to = c("1", "2","3","4","5"))  >  > df\_clean$marstatus\_num <- revalue(df\_clean$Marital\_Status, c("Alone"=1,"Divorced"=2,"Married"=3,"Single"=4,"Together"=5,"Widow"=6))  > df\_clean$marstatus\_num <- mapvalues(df\_clean$Marital\_Status, from = c("Alone","Divorced","Married","Single","Together","Widow"), to = c("1", "2","3","4","5","6"))  >  > df\_clean$marstatus\_num <- as.numeric(df\_clean$marstatus\_num)  Warning message:  NAs introduced by coercion  > df\_clean$education\_num <- as.numeric(df\_clean$education\_num) |  |
| #####Question B #Creating an empty variable for the total sum of amounts  df\_clean$total\_amount <- c()  #Creating a function that will get the total amount  totalamount <- function(var1,var2,var3,var4,var5,var6){  total\_amount = var1+var2+var3+var4+var5+var6    return(total\_amount)    } #closing unique\_score UDF  #calling a function  df\_clean$total\_amount <- totalamount(var1=df\_clean$MntWines,var2=df\_clean$MntFishProducts,var3 = df\_clean$MntFruits, var4 = df\_clean$MntGoldProds,  var5 = df\_clean$MntMeatProducts, var6 = df\_clean$MntSweetProducts)  #Creating a binary variable for US 0 and Rest of World 1  df\_clean$USorROW <- recode(df\_clean$Country, "AUS" = 1, "CA" = 1, "GER" = 1, "IND" = 1, "ME"=1, "SA" =1, "SP"=1, "US"=0)  #Updating training and testing data sets for new variables  index <- sample(1:nrow(df\_clean), size = 0.7\*nrow(df\_clean))  dfclean\_train <- df\_clean[index,]  dfclean\_test <- df\_clean[-index,]  QB\_logit <- glm(USorROW ~ total\_amount, data = dfclean\_train, family = "binomial")  summary(QB\_logit) #####Question B  #Creating an empty variable for the total sum of amounts  df\_clean$total\_amount <- c()  #Creating a function that will get the total amountt  totalamount <- function(var1,var2,var3,var4,var5,var6){  total\_amount = var1+var2+var3+var4+var5+var6    return(total\_amount)    } #closing unique\_score UDF  #calling a function  df\_clean$total\_amount <- totalamount(var1=df\_clean$MntWines,var2=df\_clean$MntFishProducts,var3 = df\_clean$MntFruits, var4 = df\_clean$MntGoldProds,  var5 = df\_clean$MntMeatProducts, var6 = df\_clean$MntSweetProducts)  #Creating a binary variable for US 0 and Rest of World 1  df\_clean$USorROW <- recode(df\_clean$Country, "AUS" = 1, "CA" = 1, "GER" = 1, "IND" = 1, "ME"=1, "SA" =1, "SP"=1, "US"=0)  #Updating training and testing data sets for new variables  index <- sample(1:nrow(df\_clean), size = 0.7\*nrow(df\_clean))  dfclean\_train <- df\_clean[index,]  dfclean\_test <- df\_clean[-index,]  QB\_logit <- glm(USorROW ~ total\_amount, data = dfclean\_train, family = "binomial")  summary(QB\_logit) | Call:  glm(formula = USorROW ~ total\_amount, family = "binomial", data = dfclean\_train)  Deviance Residuals:  Min 1Q Median 3Q Max  -2.4690 0.3119 0.3123 0.3133 0.3156  Coefficients:  Estimate Std. Error z value Pr(>|z|)  (Intercept) 3.000e+00 1.691e-01 17.73 <2e-16 \*\*\*  total\_amount -9.865e-06 1.972e-04 -0.05 0.96  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  (Dispersion parameter for binomial family taken to be 1)  Null deviance: 594.72 on 1550 degrees of freedom  Residual deviance: 594.71 on 1549 degrees of freedom  AIC: 598.71  Number of Fisher Scoring iterations: 5 | In the first part of this code we created a new variable “total amount” to compute the sum of all the purchases made by every customer. To compute US and non-US customers we created a binary variable “USorROW”, where 0 was assigned to US customers and 1 to Rest of World.  With variables ready, we then proceeded to run a logistic regression model. As seen on the output, we observe that the P-value is much higher than 0.05 therefore “total amount” is not at all significant. We can say with 95% certainty that there is no significant difference within US and non-US customers and the amount purchased. |
| ######Question C ## Finding average amount spend on Gold ##  av\_amt\_gold <- mean(df\_clean$MntGoldProds)  ## Creating a vector for dummy variable (0 <- below av. & 1 <- above av.) ##  df\_clean$gold\_group <- c()  for (i in 1:nrow(df\_clean)) {  if(df\_clean$MntGoldProds[i]>=av\_amt\_gold){  df\_clean$gold\_group[i] <- 1  } else {df\_clean$gold\_group[i] <- 0}  }  #Updating training and testing data sets for new variables  index <- sample(1:nrow(df\_clean), size = 0.7\*nrow(df\_clean))  dfclean\_train <- df\_clean[index,]  dfclean\_test <- df\_clean[-index,]  #Statistical Test (Response: Gold\_group (0,1); Exp: NumStorePurchases) #####  QC\_logit <- glm(dfclean\_train$gold\_group~dfclean\_train$NumStorePurchases, data=dfclean\_train, family="binomial")  summary(QC\_logit) | Call:  glm(formula = dfclean\_train$gold\_group ~ dfclean\_train$NumStorePurchases,  family = "binomial", data = dfclean\_train)  Deviance Residuals:  Min 1Q Median 3Q Max  -1.6469 -0.7603 -0.5979 0.9659 2.2458  Coefficients:  Estimate Std. Error z value Pr(>|z|)  (Intercept) -2.43809 0.13403 -18.19 <2e-16 \*\*\*  dfclean\_train$NumStorePurchases 0.26895 0.01869 14.39 <2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  (Dispersion parameter for binomial family taken to be 1)  Null deviance: 1928.7 on 1550 degrees of freedom  Residual deviance: 1692.5 on 1549 degrees of freedom  AIC: 1696.5  Number of Fisher Scoring iterations: 3 | Ho: Customer who spend above average (44.02) in gold are conservative    From the result of the regression,  P value is less than 0.05*(<2e-16)*; Therefor we can reject the Null hypothesis (Ho).   Conclusion: Supervisors statement is incorrect. |
| ######Question D #Create dummy for married and PhD  df\_clean$Dummy\_Married <- ifelse(df\_clean$Marital\_Status == "Married", "1", "0")  df\_clean$Dummy\_PhD <- ifelse(df\_clean$Education == "PhD", "1", "0")  df\_clean$Age <- (2021 - df\_clean$Year\_Birth)  #Updating training and testing data sets for new variables  index <- sample(1:nrow(df\_clean), size = 0.7\*nrow(df\_clean))  dfclean\_train <- df\_clean[index,]  dfclean\_test <- df\_clean[-index,]  #Creating model  QD\_logit <- lm(MntFishProducts ~ Dummy\_Married+Dummy\_PhD+Dummy\_Married\*Dummy\_PhD+Income, data=dfclean\_train)  summary(QD\_logit) | Call:  lm(formula = MntFishProducts ~ Dummy\_Married + Dummy\_PhD + Dummy\_Married \*  Dummy\_PhD + Income, data = dfclean\_train)  Residuals:  Min 1Q Median 3Q Max  -583.08 -27.21 -13.48 7.28 209.37  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -3.123e+00 3.098e+00 -1.008 0.314  Dummy\_Married1 -3.928e+00 2.974e+00 -1.321 0.187  Dummy\_PhD1 -1.922e+01 3.907e+00 -4.919 9.62e-07 \*\*\*  Income 8.913e-04 4.798e-05 18.578 < 2e-16 \*\*\*  Dummy\_Married1:Dummy\_PhD1 -7.750e-01 6.275e+00 -0.124 0.902  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 49.88 on 1546 degrees of freedom  Multiple R-squared: 0.1948, Adjusted R-squared: 0.1927  F-statistic: 93.52 on 4 and 1546 DF, p-value: < 2.2e-16 | From the result of the regression, the coefficient of married candidate is statistically insignificant, suggesting that there is no relation between married candidate and amount spend on fish. Meanwhile, PhD candidate have high negative coefficient, disproving the hypothesis that highly educate tend people buy more fish while in fact highly educate people statistically speaking tend to buy less fish. In terms of interaction between married and PhD candidate, the regression shows that the coefficient of the interaction term is statistically insignificant, suggesting that there is no interaction relationship between the two predictor variables (married and PhD). Other factor worth mentioning is income which the regression shows is statistically insignificant, suggesting that there is high relation between income and amount spend on fish |
| ######Question E campaign <- data.frame(campain = c("AcceptedCmp1", "AcceptedCmp2", "AcceptedCmp3", "AcceptedCmp4", "AcceptedCmp5", "Response"),  TotalAccepted = c(sum(df\_clean$AcceptedCmp1),sum(df\_clean$AcceptedCmp2),  sum(df\_clean$AcceptedCmp3),sum(df\_clean$AcceptedCmp4),  sum(df\_clean$AcceptedCmp5),sum(df\_clean$Response)))  #Understanding how the campaigns behaved  campaign  #Creating a model to better understand the customers of the worst performing campaign  #Logistic regression  QE\_logit <- lm(AcceptedCmp2 ~ Income+Age+Kidhome+Teenhome,dfclean\_train)  summary(QE\_logit)  QE2\_logit <- lm(Response ~ Income+Age+Kidhome+Teenhome, dfclean\_train)  summary(QE2\_logit) | Call:  lm(formula = AcceptedCmp2 ~ Income + Age + Kidhome + Teenhome,  data = dfclean\_train)  Residuals:  Min 1Q Median 3Q Max  -0.11851 -0.02102 -0.01525 -0.00176 0.98391  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 2.100e-02 1.501e-02 1.399 0.16188  Income 1.846e-07 1.153e-07 1.602 0.10947  Age -2.186e-04 2.555e-04 -0.856 0.39241  Kidhome -1.593e-02 5.772e-03 -2.760 0.00584 \*\*  Teenhome -1.542e-04 5.447e-03 -0.028 0.97742  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.1096 on 1546 degrees of freedom  Multiple R-squared: 0.01019, Adjusted R-squared: 0.00763  F-statistic: 3.979 on 4 and 1546 DF, p-value: 0.003232  > QE2\_logit <- lm(Response ~ Income+Age+Kidhome+Teenhome, dfclean\_train)  > summary(QE2\_logit)  Call:  lm(formula = Response ~ Income + Age + Kidhome + Teenhome, data = dfclean\_train)  Residuals:  Min 1Q Median 3Q Max  -1.08478 -0.16441 -0.13028 -0.06704 0.99717  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 1.421e-01 4.817e-02 2.951 0.00322 \*\*  Income 1.468e-06 3.700e-07 3.969 7.56e-05 \*\*\*  Age -9.277e-05 8.203e-04 -0.113 0.90996  Kidhome -3.215e-02 1.853e-02 -1.735 0.08293 .  Teenhome -9.779e-02 1.748e-02 -5.593 2.64e-08 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.3519 on 1546 degrees of freedom  Multiple R-squared: 0.04009, Adjusted R-squared: 0.0376  F-statistic: 16.14 on 4 and 1546 DF, p-value: 5.901e-13 | We want to investigate how well does the campaigns perform. We create a data frame showing the total amount of people who accepted each campaign. We can see from the data that most campaign performs roughly the same **except from Campaign 2** and Response where Campaign 2 perform a lot worse than average while Response perform much better than average. We want to investigate more on that so we perform logistic regression on both campaign by defining business success as candidates who accepted the campaign (1) and defining business failure as candidates who did not accept the campaign.  We can see from the regression output that Income and Teenhome have moderate relation with campaign success while Response has strong relation with Income and Teenhome. Since we already know Response is the more successful campaign, we can focus on the factors that have statistically significant on the campaign success. In this case, on the next campaign we can target candidates with high Income and high number of teens at home to increase the campaign success. |