Marketing-Campaign - Case Study

Team 13

11/11/2021

Main Objectives

We are tasked to present to the Chief Marketing Officer our analysis of data set to increase total purchases and be more efficient with spending on the marketing department. If we are able to successfully correlate and come up with any significant regression models to show what variables are the most impactful in spending, we would consider our data research a success.

Key Ressources

We are given the data of 30 variables in our data to analyze. We are given data of customers such as: Year of birth, educational level, country, spending habits on certain items, online shopping, and other characteristics on a customer. From there, we will decide to further analyze and what to take out of our model in terms of importance.

High-Level Process

—— INSERT ———-

Part 1

```
# Import libraries
library(readxl)
library(BBmisc)
library(fastDummies)
library(dplyr)
library(ggplot2)

# Import data
mark_cmp <- read_excel("datasets_marketing_campaign_SF.xlsx")

# Inspect
colSums(is.na(mark_cmp))

# Convert
mark_cmp <- as.data.frame(mark_cmp)
```

Importing

```
library(tidyr)

# Remove NAs
mark_cmp <- drop_na(mark_cmp)

# Change date variable from character to date
mark_cmp$Dt_Customer <- as.Date(mark_cmp$Dt_Customer)

# Change Year Birth to Age
mark_cmp$actual_age <- 2021 - mark_cmp$Year_Birth

# Create sub dfs for types numerical, character and dates

bool <- sapply(mark_cmp,is.numeric)
num_cols <- mark_cmp[,bool]

#bool <- sapply(mark_cmp,is.Date)
#date_cols <- mark_cmp[,bool]

bool <- sapply(mark_cmp,is.character)
cat_cols <- mark_cmp[,bool]

# Create dummy variables for the categories
cat_dummy <- dummy_cols(cat_cols, select_columns = c('Education', 'Marital_Status','Country'))

# Remove the character entries
cat_dummy <- cat_dummy[,4:ncol(cat_dummy)]

# Combine data
cat_num_cols <- cbind(cat_dummy,num_cols)

# Write csv
write.csv(mark_cmp,"C:/Users/LK/Nextcloud7/Personal/Docs/case-studies/Marketing Campaign/data.csv", row
```

Massaging

Descriptive

```
# Plot Histogram of customer age
hist(mark_cmp$actual_age,
     main="Histogram of Actual Age",
     xlab="Age of customers",
     border="black",
     col="blue",
     xlim=c(10, 90),
     las=1,
     breaks=20)
```

Part 1a - Regression - Predictive

Label: Web Purchases Features: All numerical

Motivation

Use of linear regression to classify the important variables that help us predict web purchases.

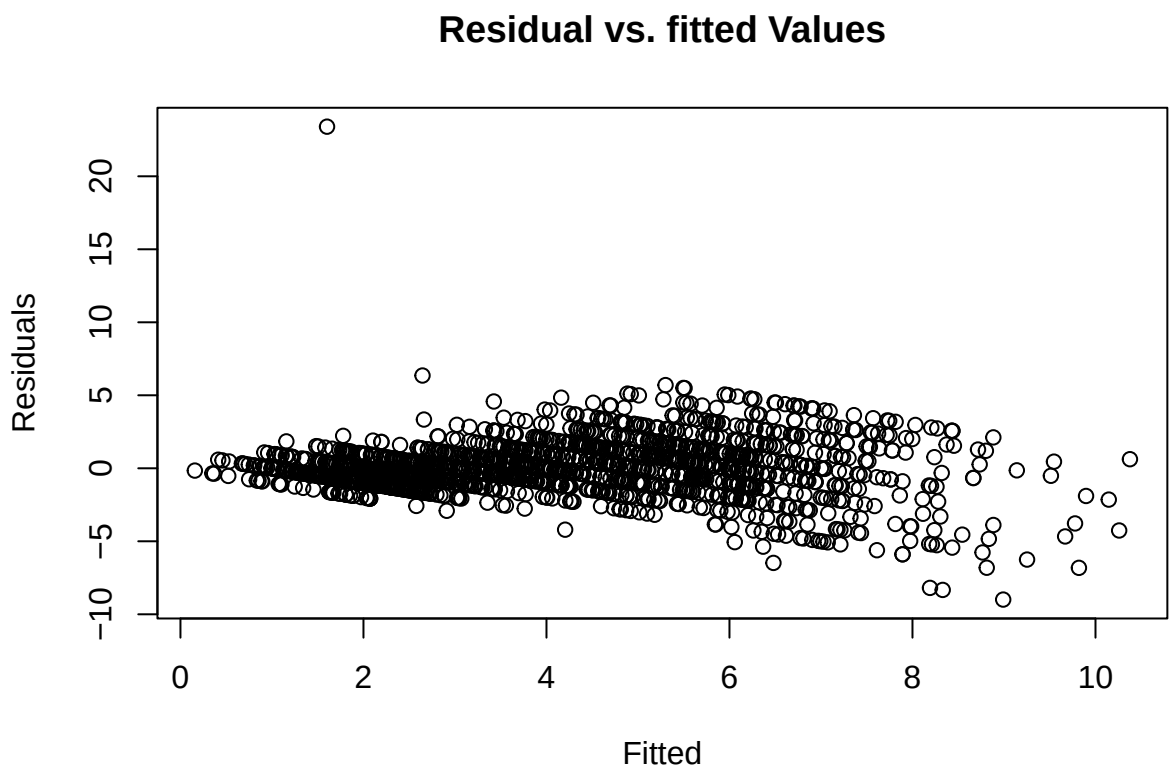
Method

- Checklist for Regression
- Split into training and test data
- Scale data
- Train model using train data
- Use the fitted model to predict unseen values (test-data)
- Plot predicted vs actual values

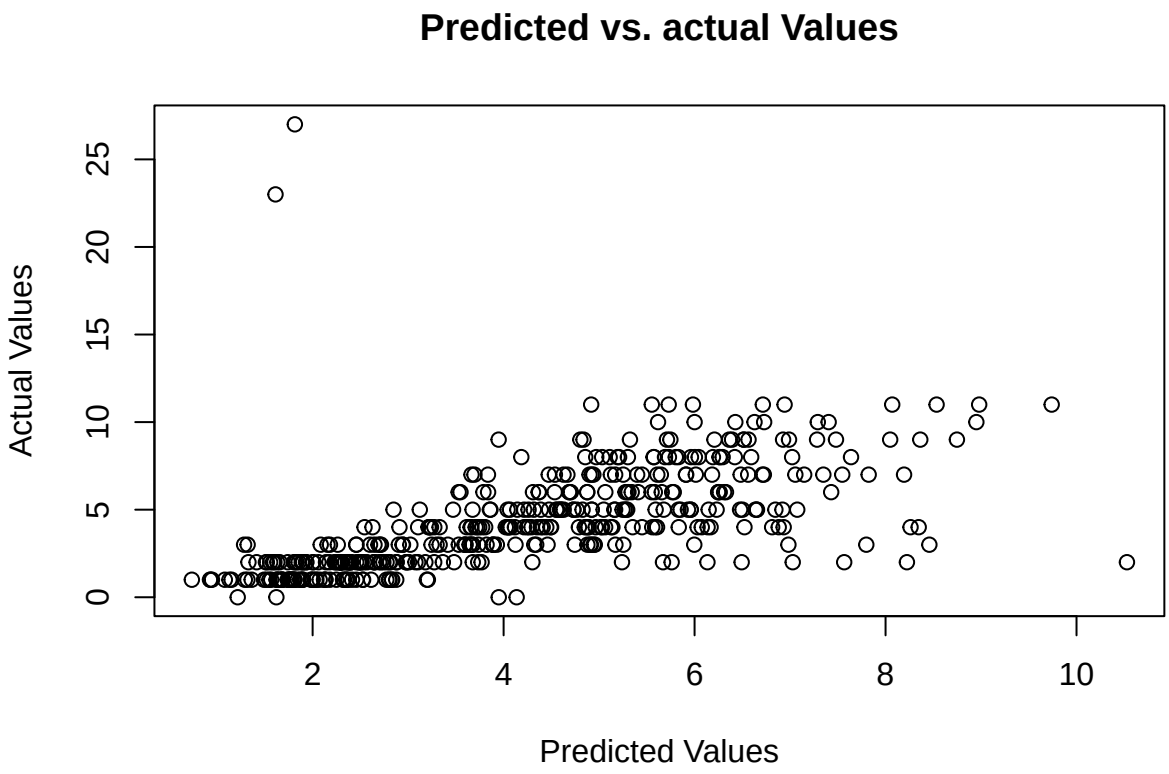
Mechanics

```
# Scale the data
scaled_mark = normalize(mark_cmp, method = "range", range = c(0, 1))
```

```r
# Split into training and test
train_index <- sample(1:nrow(mark_cmp),size=0.8*nrow(mark_cmp))
mark_train <- mark_cmp[train_index,]
mark_test <- mark_cmp[-train_index,]

# Train the model
model <- lm(NumWebPurchases ~ ., data=mark_train)
summary(model)

# Let the model predict values for the test data
pred <- predict(model,mark_test)

# Get residuals
res <- resid(model)

# Produce residual vs. fitted plot
plot(fitted(model), res, main="Residual vs. fitted Values", xlab='Fitted', ylab='Residuals')
```




```r
# Produce predicted vs actual values
plot(pred,mark_test$NumWebPurchases, main="Predicted vs. actual Values", xlab='Predicted Values', ylab=
```



Analysis of output

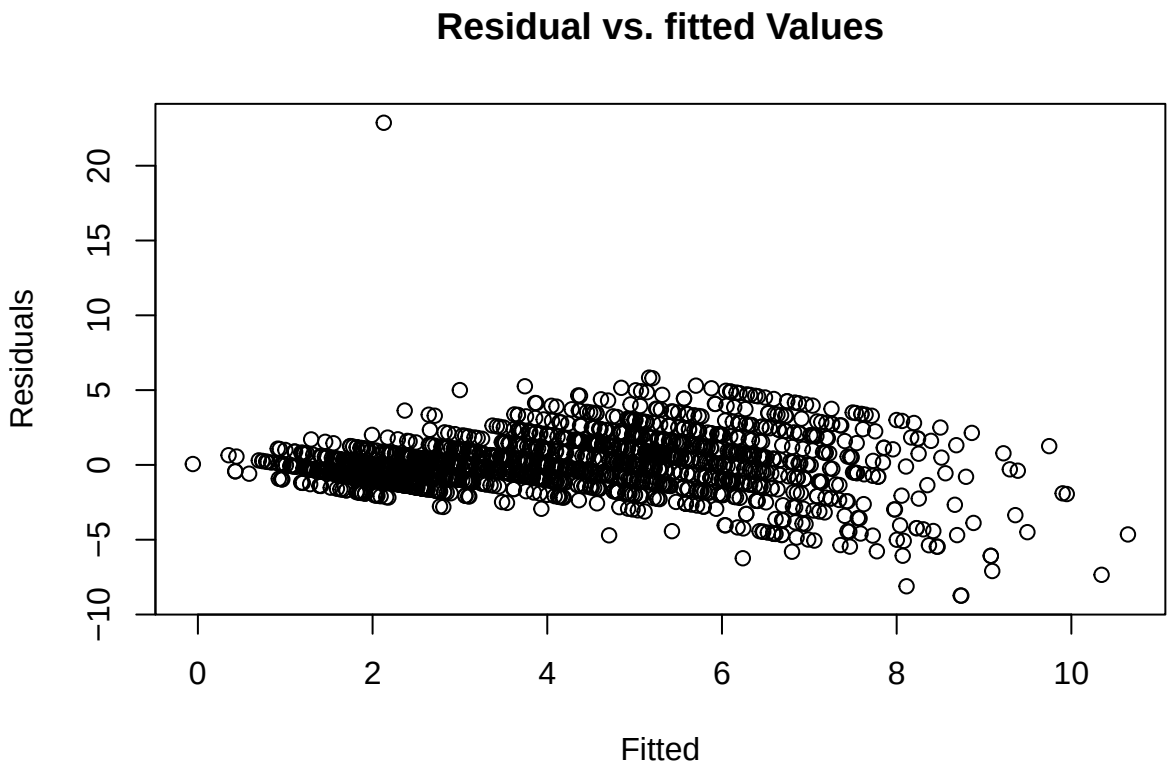
The first regression we run was built including all the variables we had from the case. This regression was meant to give us a first screening of the statistical significance every single variable brought into the model. Looking at the p-value from each row of the summary (which we decided to be $< .05$ to have a significant impact,) we could identify the variable with the highest potential. If we could have expected some variables to have an impact on web purchases, it would be number of web visits per month or number of store purchases. While on the other we see variables, initially we seemed not related to our goal, having a very high statistical significance in the model (i.e., amount sweet products, amount of gold products and number of kids residing in home. Now, in order to obtain a better regression, we excluded all the variables with a low impact and included just the ones with high potential, always basing our selection on p-values. We didn't get rid of the variable of if a customer accepted the campaign: although we had a few observations that didn't seem to be statistically important, we thought it wouldn't make sense to eliminate this bunch of variables.

Better Regression

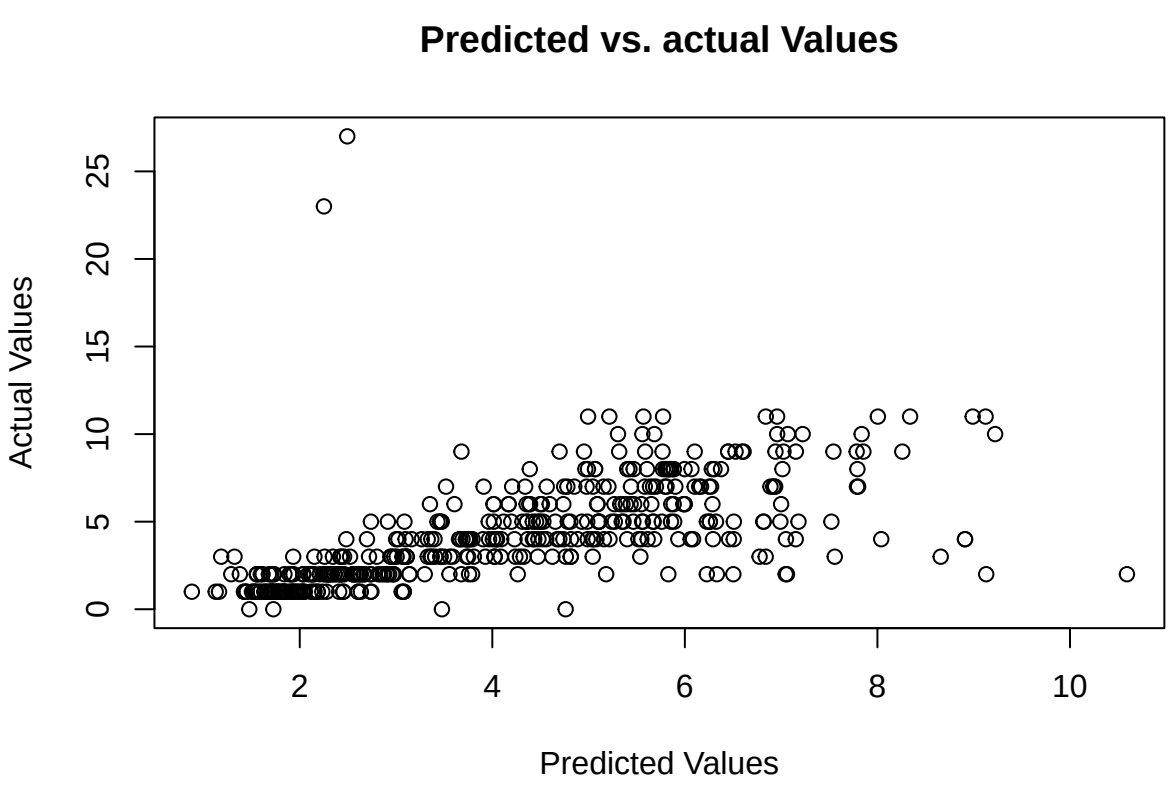
Label: Web Purchases, Features: Numerical Variables selected based on findings in Part a

```
# Train the model
better_model <- lm(NumWebPurchases ~ Income + Kidhome + Teenhome + MntWines + MntSweetProducts + MntGol
summary(model)

# Get residuals
```

```r
res <- resid(better_model)

# Produce residual vs. fitted plot
plot(fitted(better_model), res, main="Residual vs. fitted Values", xlab='Fitted', ylab='Residuals')
```




```r
#qqnorm(res)

# Let the model predict values for the test data
pred <- predict(better_model,mark_test)

# Visualize the models performance
plot(pred,mark_test$NumWebPurchases, main="Predicted vs. actual Values", xlab='Predicted Values', ylab=
```



Analysis of output

The second regression we run has similar R^2 and R^2 adjusted compared to the first one but this doesn't mean the model hasn't improved. In fact, we are now dealing with a lower number of variables. The F-statistic has raised from 41 to 183, meaning we gathered strong evidence that these variables are all statistically influential and that there's a very low chance they have a value of 0. We imagined that when we would have used fewer models in our regression that it would get stronger but that was not the case for us. This could be due to significance the isolated variables can have throughout our entire analysis.

Suggestions to CMO

Observing the estimates values from our multiple (better) regression we can select the variables that impact the most, positively, or negatively, on the case. For example, the number of kids at home is the voice that has the most negative impact we know that the odds of business success will decrease by 45% (exp (-6.06 e-01 -1) %) for each kid at home in our data seems to impact negatively on the odds of web purchases. On the other hand, the number of teens at home and number of web visits per home have the highest positive influence in the regression. These variables bring respectively 43 % (exp (3.562 e-01)-1) and 42% (exp (3.530 e-01) -1) increase in the odds of business success for each unit increased of both these voices. Last, the variable with the least impact in the model ends up being Income, as it's the closest to zero. It just brings 0.0032% (exp (3.175 e-05) -1) positive influence. Our recommendation would be of course to focus on the voices with the best positive impact and try to avoid as much as possible the ones lowering the odds of business success. For example, it's important to keep in mind that targeting people with kids at home seems to be negative, on the other hand, it could be convenient to rely on people with teenagers at home, probably because teens are more willing than kids to use online sources. Moreover, the CMO should focus more on

the number of web visits which has a 42% (exp (3.530 e-01) -1) positive impact. This could be done by allocating some budget on online ads or on everything that could bring visibility to the website. Speaking about budget allocation, it would be crucial to understand how to allocate and thus to select the variables with a low influence in the model. For example, our team thinks Income wouldn't be such a critical aspect to focus on, the company should save resources by not targeting this market segment and maybe invest in others with better potential. In fact, besides Income, there are a few more variables that seem to have higher importance but still not as significant to justify an effort. Amount spent on wines, gold, and sweet products in the past two years estimates, for instance, are close to zero, and despite their relevance, could also not be taken into consideration in the context of budget allocation. In conclusion, as we have seen, even though the better regression includes just the most significant variables, it seems that from a statistical point of view just the variation of a few manage to change the odds of business success and so the output of a business reasoning.

We should also use the "income" variable to see the different type of classes they have within their data set. We should break this category into three tiers: low-income, middle-income, and high-income brackets to acquire more specific information for targeting campaigns to the varying income classes that are useful in this data set. As the two campaigns did very well and were closely related to in-store purchases we need to see the specifics of why these campaigns were successful. On the other hand, we need to investigate the failure of purchases from the other campaigns and lean away from repeating these tactics again to not be wasteful in budgeting and marketing resources.

We also highly advise personalizing the communications, deals, and preferences with more specific characteristics for personnel. I would say that more specific marketing should be based on a customer demographic: education level, marital status, income level, and age, for example, to have a more pinpointed approach with advertising. This could be evident with doing more in-depth analysis of why one campaign was more successful to another, based on who saw it and made a purchase. We need to investigate performing A/B testing to increase deals. Using resources more effectively with data analysis leads to efficient marketing tactics for increased Return on investment for marketing budgets.

Part 1b - Descriptive - US vs rest of world

Motivation

We intend to find if there is any significance in demographics with purchasing behavior between the U.S. and the rest of the world. This is extremely vital for the marketing department due to attaining an overview of which region in the world purchases more and how to more effectively allocate budget and effort with different strategies.

Method

Our approach was to aggregate the columns which included: purchases made in store, catalog purchases, purchases through deals, and web purchases. From there, we created dummy variables with 1 and 0. If the country was US, we would place a 1, and if not, we would input a 0. We compared the two different categories with our explanatory variable of using the dummy variable for the country and our response will be the total purchases. To find out statistical significance, we will calculate the 95% interval for the difference in means.

H0: U.S. Total Purchases <= Rest of the world Ha: U.S. Total Purchases > Rest of the world

Mechanics

```
# Aggregate total purchases
cat_num_cols$Total_Purch <- cat_num_cols$NumDealsPurchases + cat_num_cols$NumWebPurchases + cat_num_cols
```

```r
# Create sub df for analysis for US customers and Rest of world
us_customer <- cat_num_cols[which(cat_num_cols$Country_US==1),]
world_customer <- cat_num_cols[-which(cat_num_cols$Country_US==1),]

# Observations & Mean of purchases & Standard deviation of US-Customer
n_us <- nrow(us_customer)
mean_us <- mean(us_customer$Total_Purch)
sd_us <- sd(us_customer$Total_Purch)

# Observations & Mean of purchases & Standard deviation of World
n_world <- nrow(world_customer)
mean_world <- mean(world_customer$Total_Purch)
sd_world <- sd(world_customer$Total_Purch)

# Calculate pooled variance
sp = ((n_us-1)*sd_us^2 + (n_world-1)*sd_world^2) / (n_us+n_world-2)

# Calculate margin of error
margin <- qt(0.975,df=n_us+n_world-1)*sqrt(sp/n_us + sp/n_world)

#calculate lower and upper bounds of confidence interval
low <- (mean_world-mean_us) - margin
high <- (mean_world-mean_us) + margin
```

Message

The 95% confidence interval for the true difference in population means is between -2.9700892 and 0.0094266

We saw that the U.S. and the rest of the world are not significantly different in terms of total purchases made. This was made through calculating the confidence intervals of the difference in means for the two different categories. Because our 95% confident interval included 0, we were able to see that there was no significant difference between total purchases made in the U.S. and the rest of the world. With the included box plot visual, you can see that they are very similar to each other with its boxplot graphing.

```r
library(plotly)

fig <- plot_ly(y = us_customer$Total_Purch, type = "box", name="US") %>%
        layout(title = 'Average Purchases per Country', plot_bgcolor = "#e5ecf6",xaxis = list(title = '
fig <- fig %>% add_trace(y = world_customer$Total_Purch, name='Rest of world')
fig
```

Average Purchases per country

Part 1c - Descriptive - Gold vs non-Gold

Motivation c)

Our motivation with this analysis is to see if people who bought gold above the average purchase in store more often those who did not. This information could prove useful to categorize customers into different labels for seeing the projection of customers buying habits.

Mechanics

```
#calculating the mean for gold purchasers and store purchasers
MntGoldProds.mean <- mean (cat_num_cols$MntGoldProds)
MntGoldProds.std <- sd (cat_num_cols$MntGoldProds)

NumStorePurchases.mean <- mean (cat_num_cols$NumStorePurchases)
NumStorePurchases.std <- sd (cat_num_cols$NumStorePurchases)

print (MntGoldProds.mean)
print (NumStorePurchases.mean)

for (i in 1:nrow(cat_num_cols)){
  if (cat_num_cols$MntGoldProds[i] > MntGoldProds.mean){
    if (cat_num_cols$NumStorePurchases[i] > NumStorePurchases.mean){
      cat_num_cols$test[i] <- c("CONSERVATORY")} else     {cat_num_cols$test[i] <- c("TECH")
      }
         }   else {cat_num_cols$test[i] <- c("LOWER")}
  }


which (cat_num_cols$test == 'LOWER')
cat_num_cols[which (cat_num_cols$test == 'LOWER') , ]

which (cat_num_cols$test == 'CONSERVATORY')
cat_num_cols[which (cat_num_cols$test == 'CONSERVATORY') , ]

which (cat_num_cols$test == 'TECH')
cat_num_cols[which (cat_num_cols$test == 'TECH') , ]

## confidence interval for CONSERVATORY
n_1 <- nrow(cat_num_cols[which (cat_num_cols$test == 'CONSERVATORY'),])
mean.conservatory <- mean (cat_num_cols$MntGoldProds[cat_num_cols$test == 'CONSERVATORY'])
std.conservarory <- sd (cat_num_cols$MntGoldProds[cat_num_cols$test == 'CONSERVATORY'])
error_1 <- qnorm(0.975)*std.conservarory/sqrt(n_1)

mean.conservatory-error_1
mean.conservatory+error_1

## confidence interval for TECH
n_2 <- nrow(cat_num_cols[which (cat_num_cols$test == 'TECH') , ])
mean.tech <- mean (cat_num_cols$MntGoldProds[cat_num_cols$test == 'TECH'])
std.tech <- sd (cat_num_cols$MntGoldProds[cat_num_cols$test == 'TECH'])
error_2 <- qnorm(0.975)*std.tech/sqrt(n_2)

mean.tech-error_2
mean.tech+error_2
```

Message

Through our analysis, we were able to see that we are going to disagree with our supervisor in that there is no correlation between the amount of gold purchased within the past two years would have more in-store

purchases. This is useful information to have for our marketing department because that would let us know the projection in which the customer buying habit is going towards. If we see that less customers are purchasing in-store and opting for buying online, we can spend more of our budget towards increasing the attraction of our online store.

Part 1d - Regression - Predictive

Label: mntFishProducts

Motivation

Use of linear regression to classify the important variables that help us predict the amount spent on Fish Products for married phd-candidates.

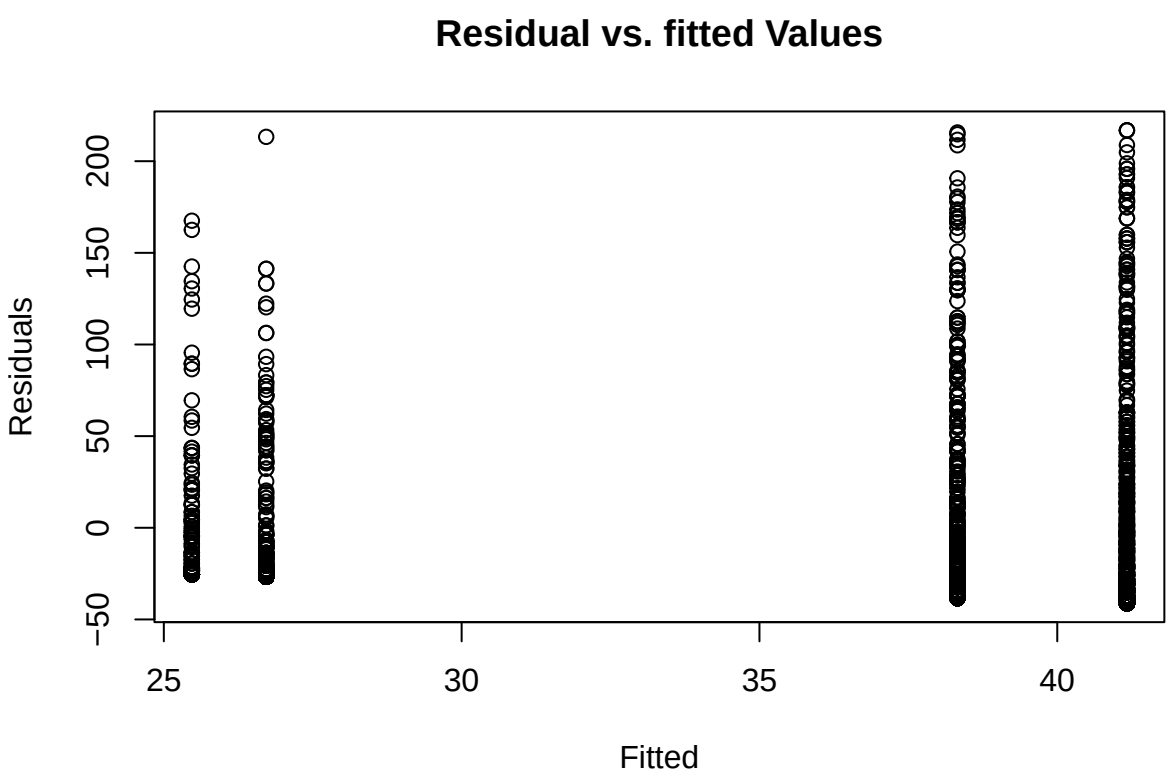
Mechanics

```
mark_train$Binary1 <- c() #build the variable to store the values of the loop

for (i in 1:nrow(mark_train)) {

  if(mark_train$Education[i] == "PhD"){
    mark_train$Binary1[i] <- "1"
  }else{
    mark_train$Binary1[i] <- "0"
  } #closing the if

} #closing the for loop

mark_train$Binary2 <- c() #build the variable to store the values of the loop

for (i in 1:nrow(mark_train)) {

  if(mark_train$Marital_Status[i] == "Married"){
    mark_train$Binary2[i] <- "1"
  }else{
    mark_train$Binary2[i] <- "0"
  } #closing the if

}

mark_train$Binary1<-as.numeric(mark_train$Binary1)
mark_train$Binary2<-as.numeric(mark_train$Binary2)

EDUXMAR<-function(var1,var2){
  PHDXMAR<-va1*var2
  return(PHDXMAR)

}
mark_train$Binary3 <- c(mark_train$Binary1*mark_train$Binary2)
```

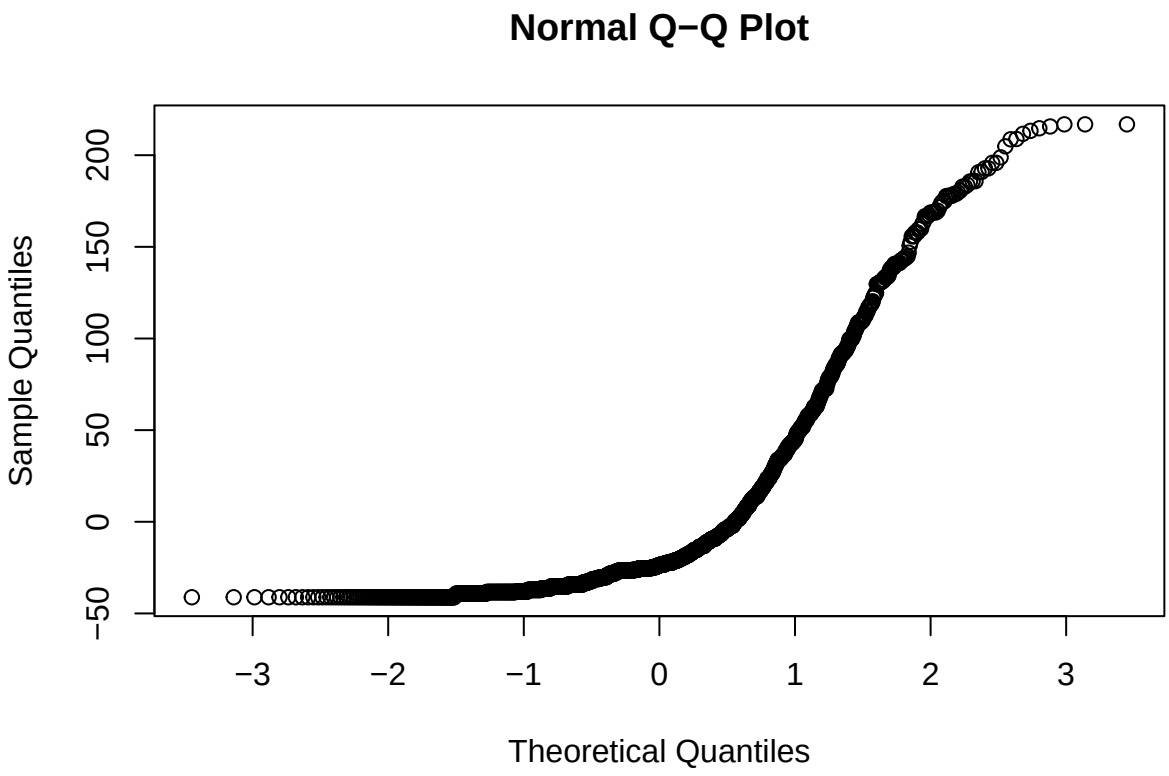
```r
my_logit <- glm(MntFishProducts~Binary1+Binary2+Binary3, data=mark_train)
summary(my_logit)
```

```
##
## Call:
## glm(formula = MntFishProducts ~ Binary1 + Binary2 + Binary3,
##     data = mark_train)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -41.17  -34.54  -23.72   10.94  216.83
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   41.170      1.862  22.112  < 2e-16 ***
## Binary1      -14.451      4.011  -3.603 0.000323 ***
## Binary2       -2.846      2.963  -0.961 0.336890
## Binary3        1.597      6.380   0.250 0.802326
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 2915.345)
##
##     Null deviance: 5214400  on 1771  degrees of freedom
## Residual deviance: 5154329  on 1768  degrees of freedom
## AIC: 19171
##
## Number of Fisher Scoring iterations: 2
```

```r
res <- resid(my_logit)
# Produce residual vs. fitted plot
plot(fitted(my_logit), res,  main="Residual vs. fitted Values", xlab='Fitted', ylab='Residuals')
```


Residual vs. fitted Values
Residuals
−50
0
50
100
150
200
25
30
35
40
Fitted

```
qqnorm(res)
```

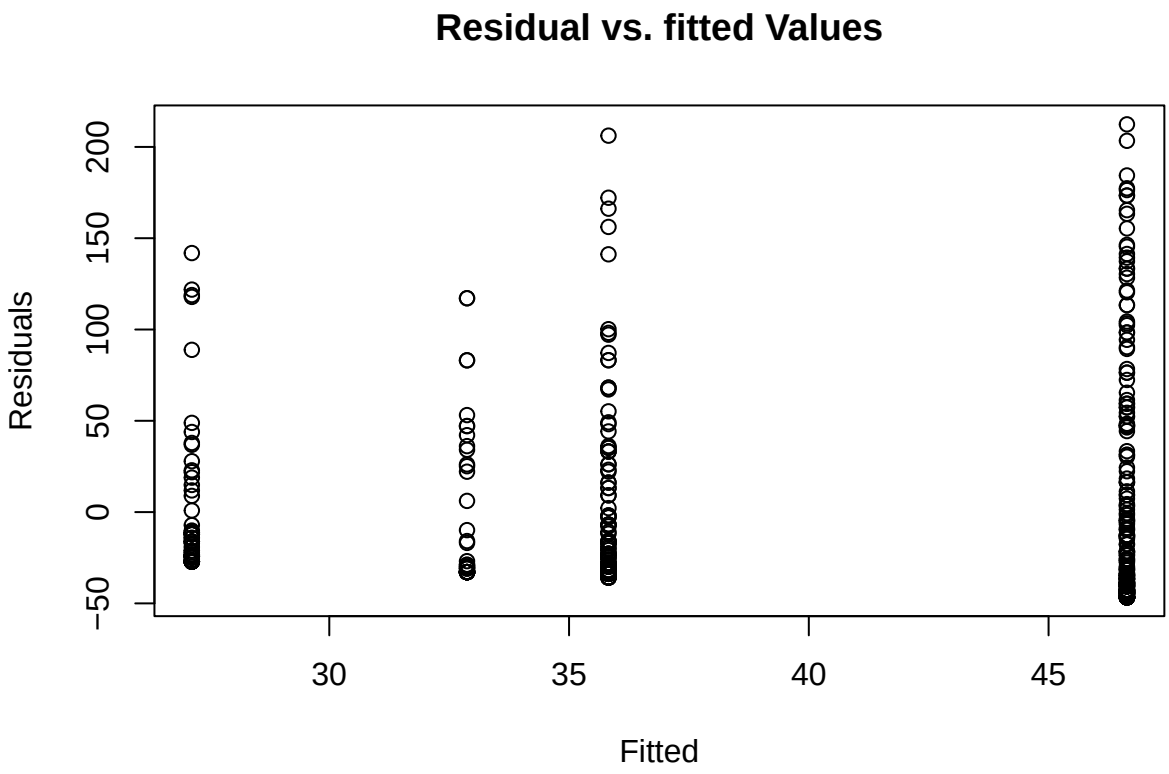



```
mark_test$Binary1 <- c() #build the variable to store the values of the loop

for (i in 1:nrow(mark_test)) {

  if(mark_test$Education[i] == "PhD"){
    mark_test$Binary1[i] <- "1"
  }else{
    mark_test$Binary1[i] <- "0"
  } #closing the if

} #closing the for loop

mark_test$Binary2 <- c() #build the variable to store the values of the loop

for (i in 1:nrow(mark_test)) {

  if(mark_test$Marital_Status[i] == "Married"){
    mark_test$Binary2[i] <- "1"
  }else{
    mark_test$Binary2[i] <- "0"
  } #closing the if

}

mark_test$Binary1<-as.numeric(mark_test$Binary1)
mark_test$Binary2<-as.numeric(mark_test$Binary2)
```

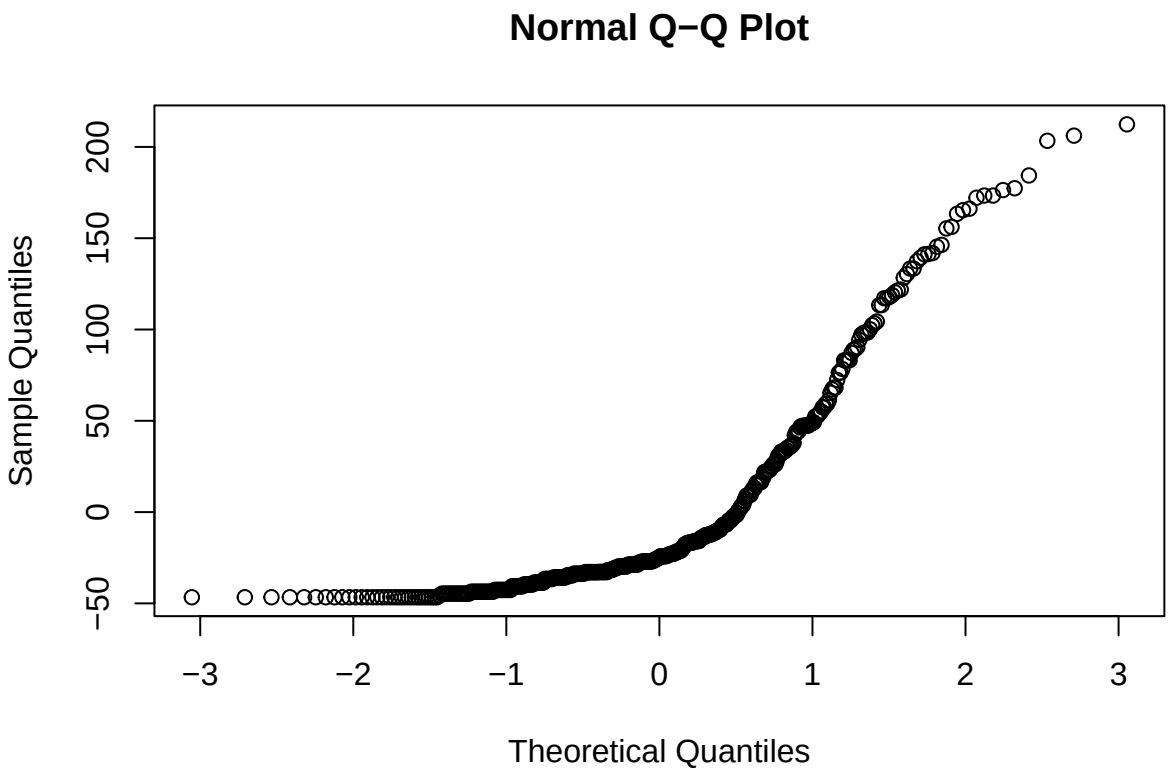
```
EDUXMAR<-function(var1,var2){
  PHDXMAR<-va1*var2
  return(PHDXMAR)

}
mark_test$Binary3 <- c(mark_test$Binary1*mark_test$Binary2)

my_logit1 <- glm(MntFishProducts~Binary1+Binary2+Binary3, data=mark_test)
summary(my_logit1)
```

```
##
## Call:
## glm(formula = MntFishProducts ~ Binary1 + Binary2 + Binary3,
##     data = mark_test)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -46.63  -35.82  -24.82   18.49  212.37
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   46.634      3.742  12.461   <2e-16 ***
## Binary1      -19.501      8.185  -2.383   0.0176 *
## Binary2      -10.812      6.399  -1.690   0.0918 .
## Binary3       16.551     13.246   1.250   0.2121
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 3179.225)
##
##     Null deviance: 1422822  on 443  degrees of freedom
## Residual deviance: 1398859  on 440  degrees of freedom
## AIC: 4846.6
##
## Number of Fisher Scoring iterations: 2
```

```
res <- resid(my_logit1)
# Produce residual vs. fitted plot
plot(fitted(my_logit1), res,  main="Residual vs. fitted Values", xlab='Fitted', ylab='Residuals')
```


Residual vs. fitted Values
Residuals
−50
0
50
100
150
200
30
35
40
45
Fitted

```
qqnorm(res)
```




```
model2 <- lm(MntFishProducts ~ ., data=mark_cmp)
```

Part 1e - Regression - Predictive

Motivation

Mechanics

Part 1: Key Findings