Marketing-Campaign - Case Study

Timon-Laurin Krämer

11/11/2021

Problem Statement

—- INSERT ———

Main Objectives

—- INSERT ———-

Key Ressources

—- INSERT ———

High-Level Process

—- INSERT ———-

Part 1

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

# 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)

# 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

Part 1a - Regression - Predictive

Label: Web Purchases Features: All numerical

Motivation

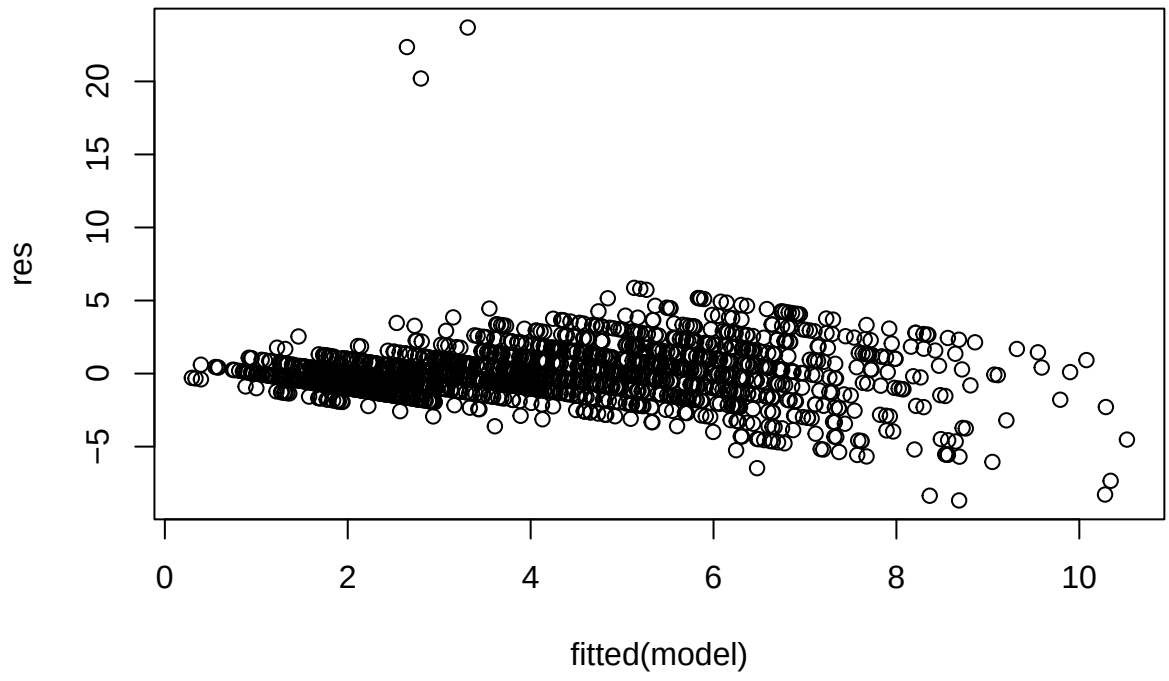
Use of linear regression to classify the important variables that help us predict web purchases.

Method

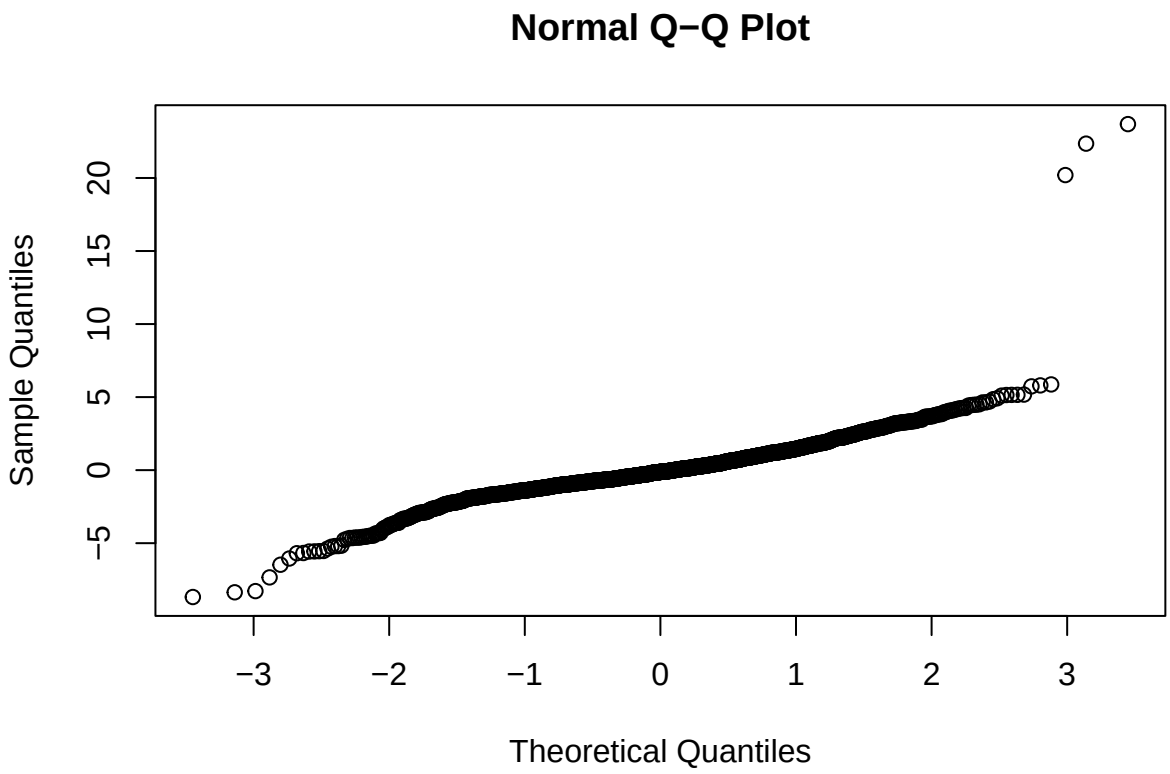
- Checklist for Regression
- Split into training and test data
- Scale data
- Train model using train data
- Use the fitted model to predict unseen values (test-data)
- Plot predicted vs actual values

Mechanics

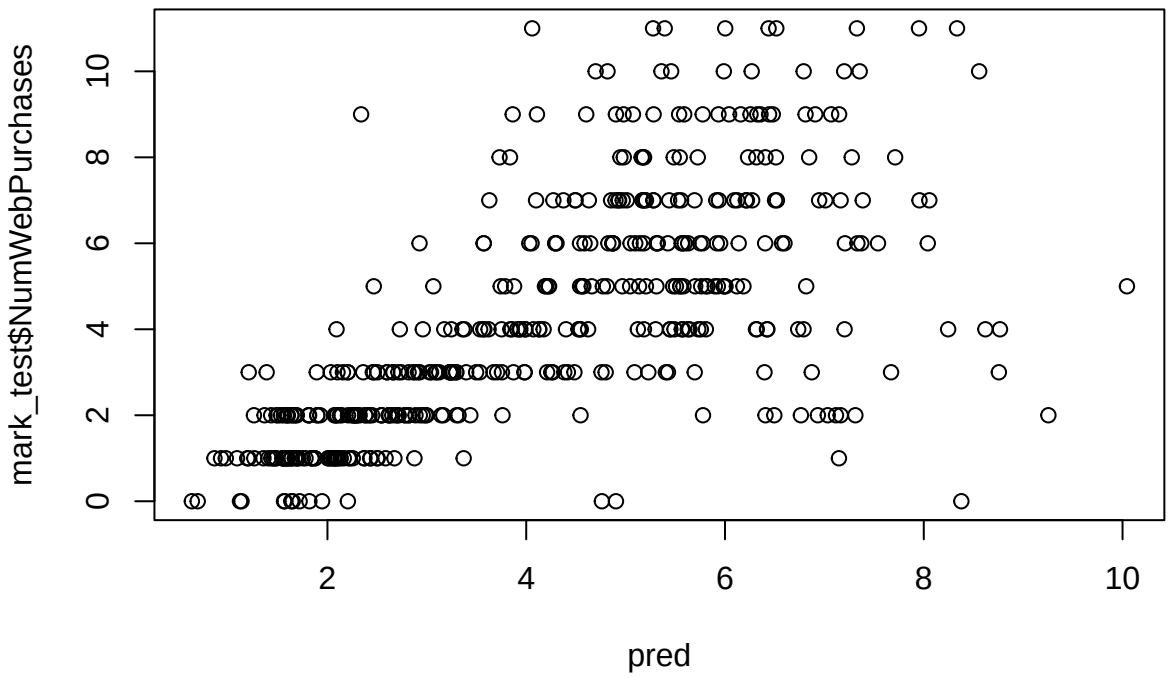
```
# Scale the data
scaled_mark = normalize(mark_cmp, method = "range", range = c(0, 1))

# 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)
```



```
qqnorm(res)
```




```
# Produce predicted vs actual values
plot(pred,mark_test$NumWebPurchases)
```



Message

(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.

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.

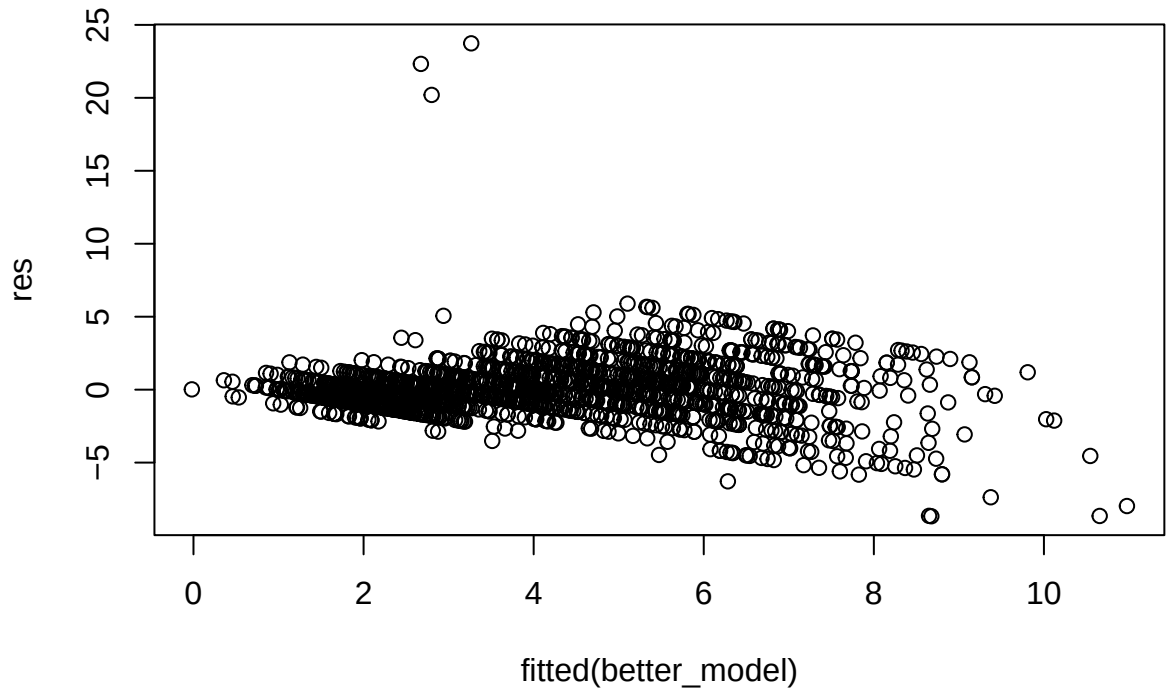
We should also use the "income" variable to see the different type of classes they have within their data set. We should break this category into three tiers: low-income, middle-income, and high-income brackets to acquire more specific information for targeting campaigns to the varying income classes 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.

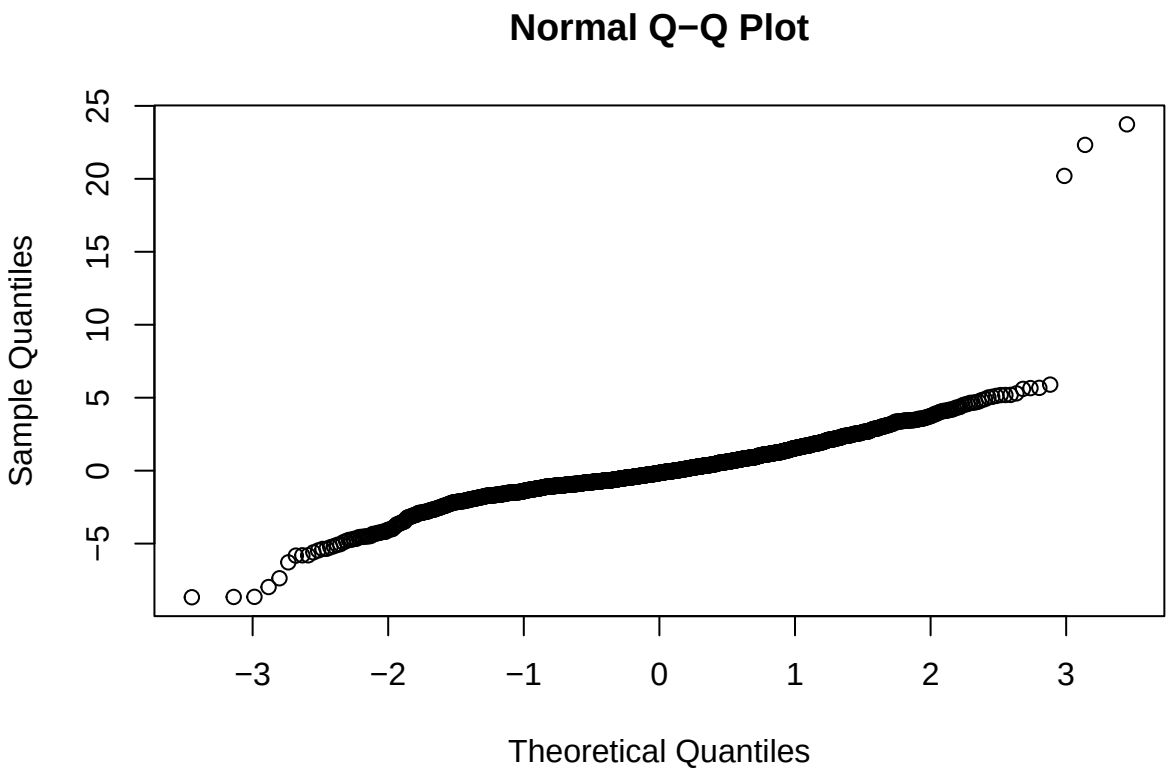
Better Regression

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

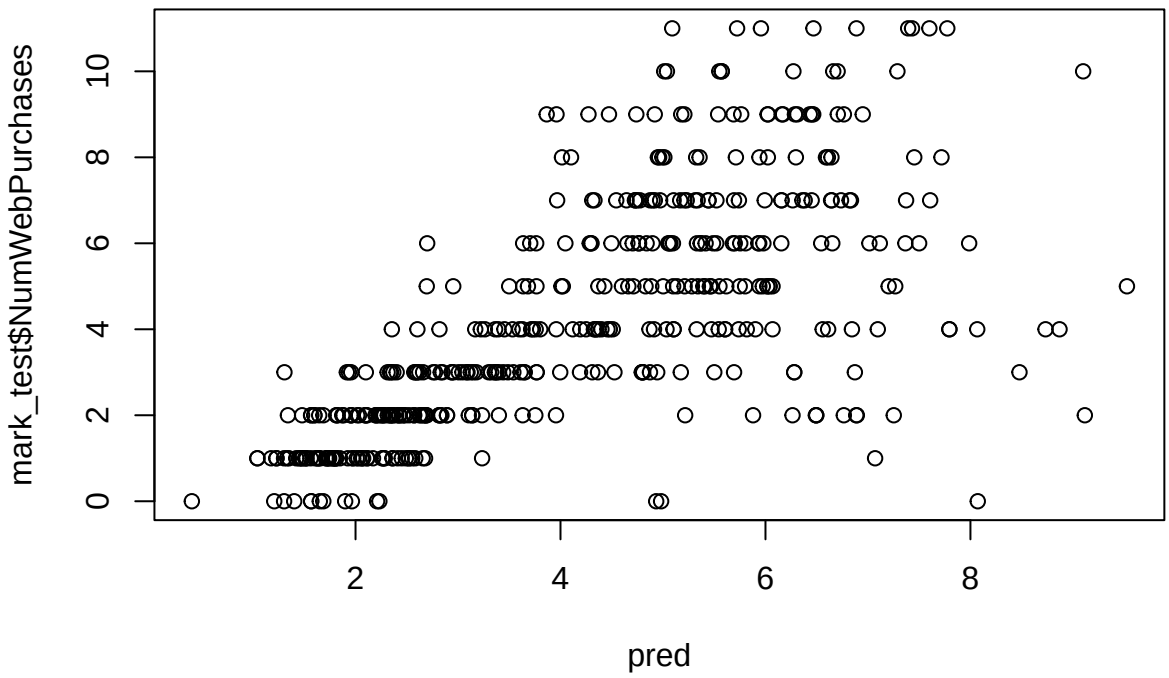
```
# Train the model
better_model <- lm(NumWebPurchases ~ Income + Kidhome + Teenhome + MntWines + MntSweetProducts + MntGol
summary(model)

# Get residuals
res <- resid(better_model)

# Produce residual vs. fitted plot
plot(fitted(better_model), res)
```




```
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)
```



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. <= Rest of the world Ha: U.S. > 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
```

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

```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
```

‘‘‘