Netflix_Analise

Bruno Florêncio

29/09/2020

Carregando as bibliotecas utilizadas neste projeto

```
library(readxl)
library(plyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

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

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

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

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

```
library(reshape2)
library(ggplot2)
library(magrittr)
library(stringr)
library(psych) # para usar a função describBY
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
library(RVAideMemoire) # função utilizada para fazer o teste de shapiro
```

```
## *** Package RVAideMemoire v 0.9-78 ***

##
## Attaching package: 'RVAideMemoire'

## The following object is masked from 'package:magrittr':
##
##     mod

library(car) # para fazer o teste de levane

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:psych':
##
##     logit

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

library(DescTools) # para os teste post-hoc Tukey HSD

##
## Attaching package: 'DescTools'

## The following object is masked from 'package:car':
##
##     Recode

## The following objects are masked from 'package:psych':
##
##     AUC, ICC, SD
```

Carregando e transformando o arquivo “Netflix_Data.xlsx”

```
Netflix_Dados<- as.data.frame (read_xlsx ("Netflix_Data.xlsx",
sheet="Dados", col_names = TRUE))
View(Netflix_Dados)
```

Trocando o nome das colunas (variáveis)

A opção deve-se primeiramente para melhorar as buscas pelo nome das variáveis e fiz a tradução, isso me ajudou entender melhor o dataset

```
Netflix_Dados <- rename (Netflix_Dados, "Data"                 = "Time",
                                        "Assinatura_Total"     = "Total
subscriptions at end of period",
                                        "Assinatura_Paga"      = "Paid
subscriptions at end of period",
```

```
                                        "Teste_gratuito"      = "Free
Trails",
                                        "Receita"             =
"Revenue",
                                        "Custo_Receita"       = "Cost of
revenues",
                                        "Custo_Marketing"     =
"Marketing",
                                        "Margem_Lucro"        =
"Contribution profit",
                                        "Margem_Contribuição" =
"Contribution Margin",
                                        "Custo_Cliente"       = "Cost per
Customer (excluding marketing)",
                                        "Receita_Cliente"     = "Revenue
per Customer",
                                        "Ganho_cliente"       = "Earnings
per Customer",
                                        "Segmento"            =
"Segment")
```

Transformando as datas para varíaveis categóricas

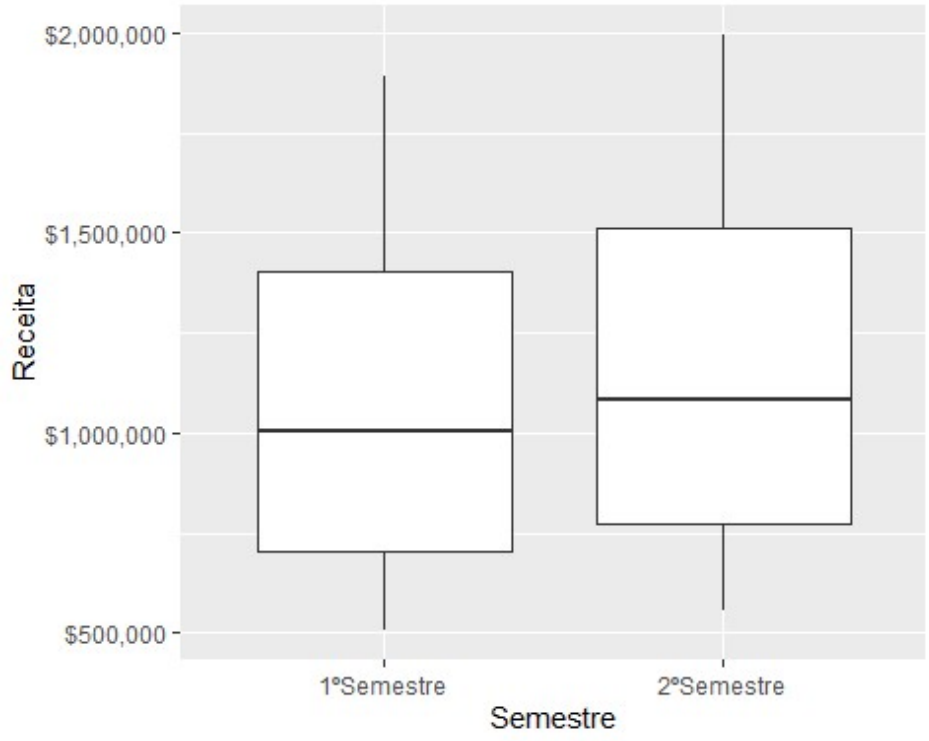
No contexto do problema de negócio, optei por transformar as datas em variáveis categóricas, pois assim, pude analisar as médias da RECEITA (Revenue) e CUSTO_MARKETING (Marketing) desenvolvendo alguns testes de hipóteses para cada uma, tendo em vista que a média populacional é conhecida, dado o dataset pequeno. Além disso podemos encontrar uma correlação positiva entre as variáveis demonstrado mais a frente. Define uma coluna para Semestre e Trimestre

```
Netflix_Dados$Data<- mdy(Netflix_Dados$Data)
Netflix_Dados$Mes<- month(Netflix_Dados$Data)
Netflix_Dados$Ano<- year(Netflix_Dados$Data)

Netflix_Dados$Mes_Ano<- paste(month(Netflix_Dados$Mes, label = TRUE),
Netflix_Dados$Ano, sep="/")

Netflix_Dados$Semestre <- as.factor (mapvalues(Netflix_Dados$Mes,
c(3,6,9,12),
                        c("1ºSemestre",
"1ºSemestre","2ºSemestre","2ºSemestre")))
Netflix_Dados$Trimestre <- as.factor (mapvalues(Netflix_Dados$Mes,
c(3,6,9,12),
                        c("1ºTrimestre",
"2ºTrimestre","3ºTrimestre","4ºTrimestre")))

View(Netflix_Dados)
```

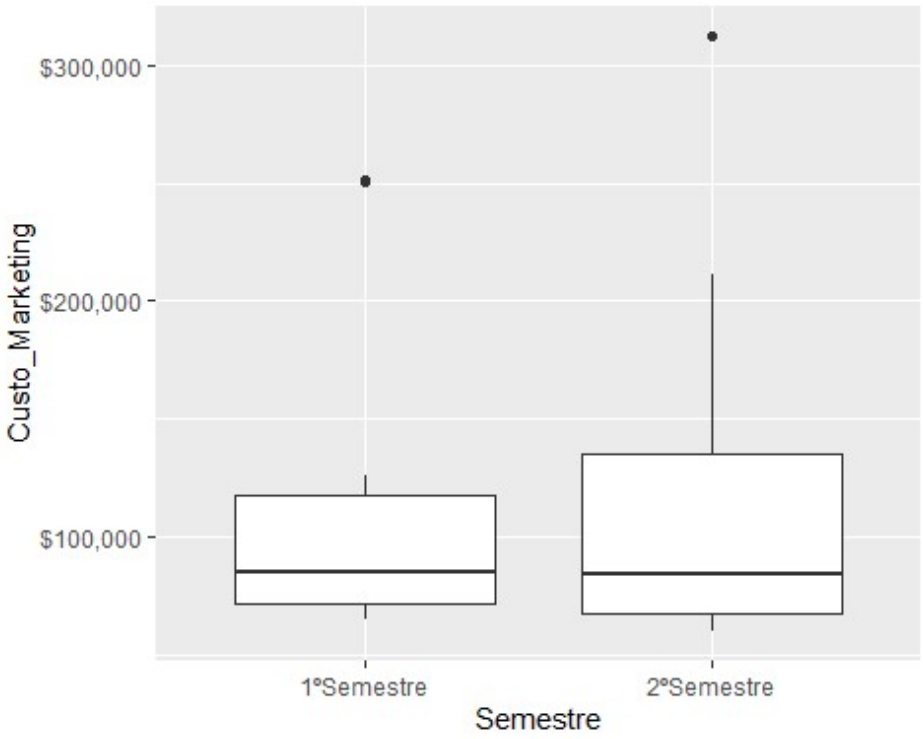
Iniciando a análise descritiva dos dados

Inicialmente trabalhei com a variável receita e criando amostras para análise do semestre. Embnora não seja umas das atividades afim, procurei observar possíveis correlações entre as variáveis em questão. Pode-se observar que há uma tendência positiva em que, na medida que aumento o custo de marketing, houve um aumento das receitas, porém essa relação ficou mais evidente a partir de 2014. Aproveitei tambéma a oportunidade para exmplorar a biblioteca ggplot2

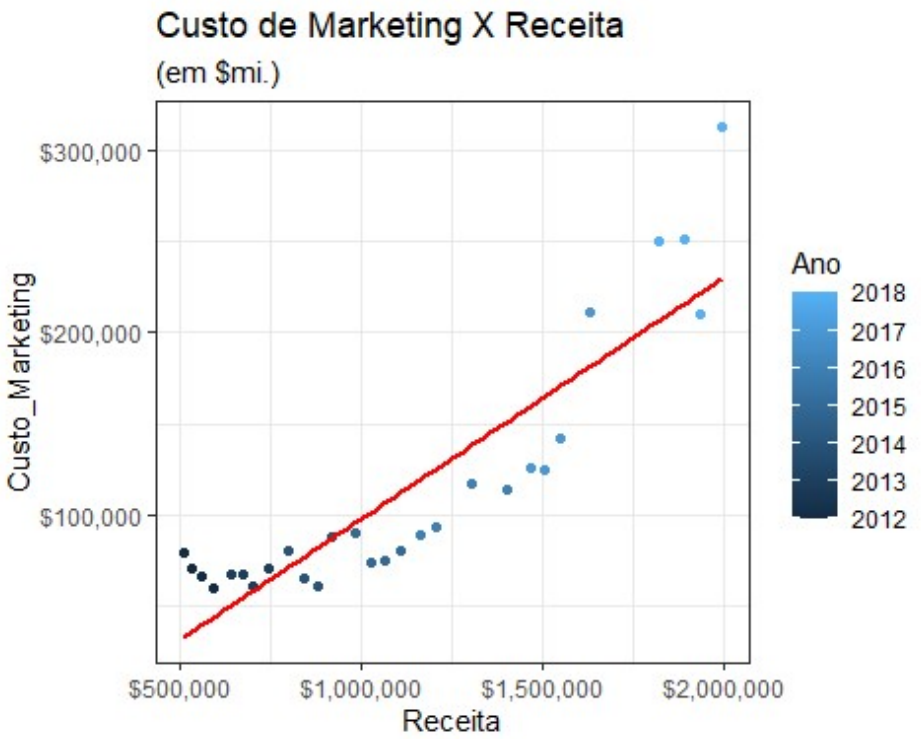
```
#Plotando os dados para averiguar as diferenças nas categorias do
semestre
Receita_plot<- ggplot (Netflix_Dados, aes(x = Semestre, y = Receita)) +
geom_boxplot()+
                scale_y_continuous(labels = scales::dollar)
Receita_plot
```




```
Custo_MKT_plot<-ggplot (Netflix_Dados, aes(x = Semestre, y =
Custo_Marketing)) + geom_boxplot() +
                scale_y_continuous(labels = scales::dollar)
Custo_MKT_plot
```




```
p1   <- ggplot(Netflix_Dados, aes (x = Receita, y= Custo_Marketing))+
        geom_point(aes(color = Ano))+
        geom_smooth(color = "red", se = FALSE, method = "lm" )+
        scale_y_continuous(labels = scales::dollar)+
        scale_x_continuous(labels = scales::dollar)+
        theme_bw()+
        labs (title = "Custo de Marketing X Receita", subtitle = "(em
$mi.)")
p1

## `geom_smooth()` using formula 'y ~ x'
```



Teste de hipoteses e ANOVA

Teste de Hipoteses para Receita

H0 : Não há diferença significativa entre a média das receitas no 1º e 2º semestre
H1 : Há média de receita no 1º semestre é < que a do 2º Semestre Unicaudal a esquerda

Teste de Hipoteses para Custo de Marketing

H0 : Não há diferença significativa entre a média do Custo de Marketing no 1º e 2º semestre
H1 : Há média do Custo de Marketing no 1º semestre é < que a do 2º Semestre
#Criando um novo objeto para não alterar o banco de dados original

```
Dados<- Netflix_Dados
attach(Dados)
View(Dados)
```

#Resumos estatísticos para varíavel Receita e Custo de Marketing

```
describeBy(Dados$Receita, group = Dados$Semestre)

##
##  Descriptive statistics by group
## group: 1ºSemestre
##    vars  n    mean     sd  median trimmed      mad    min     max
```

```
range skew
## X1    1 14 1075335 456138 1005223 1054567 519434.1 506665 1893222
1386557 0.43
##    kurtosis     se
## X1    -1.22 121908
## ------------------------------------------------------------
## group: 2ºSemestre
##    vars  n    mean       sd  median trimmed      mad    min     max
range
## X1    1 14 1169308 478865.7 1084947 1151516 539856.9 556027 1996092
1440065
##    skew kurtosis       se
## X1 0.35    -1.33 127982.2

describeBy(Dados$Custo_Marketing, group = Dados$Semestre)

##
##  Descriptive statistics by group
## group: 1ºSemestre
##    vars  n     mean       sd median  trimmed      mad   min    max
range skew
## X1    1 14 109138.8 63125.92  84416 100993.2 22754.94 64727 251298
186571 1.51
##    kurtosis       se
## X1     0.72 16871.11
## ------------------------------------------------------------
## group: 2ºSemestre
##    vars  n     mean       sd median  trimmed      mad   min    max
range skew
## X1    1 14 119034.6 75885.81  83628 107830.8 34723.97 59777 312739
252962 1.27
##    kurtosis       se
## X1     0.46 20281.34

ddply(Dados, ~ Semestre, summarize,
      Media = mean(Receita),
      DP  = sd(Receita),
      Erro_Margem = DP  / sqrt(length(Receita))
)

##     Semestre   Media       DP Erro_Margem
## 1 1ºSemestre 1075335 456138.0    121908.0
## 2 2ºSemestre 1169308 478865.7    127982.2

ddply(Dados, ~ Semestre, summarize,
      Media = mean(Custo_Marketing),
      DP  = sd(Custo_Marketing),
      Erro_Margem = DP  / sqrt(length(Custo_Marketing))
)
```

```
##     Semestre    Media       DP Erro_Margem
## 1 1ºSemestre 109138.8 63125.92    16871.11
## 2 2ºSemestre 119034.6 75885.81    20281.34
```

```
summary(Dados$Receita)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  506665  730686 1044937 1122321 1478906 1996092
```

```
summary(Dados$Custo_Marketing)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   59777   69634   83841  114087  125241  312739
```

##Criando amostra para o teste

```
amostra1<- Dados %>% filter (Semestre == "1ºSemestre" ) %>% sample_n(10)
amostra2<- Dados %>% filter (Semestre == "2ºSemestre" ) %>% sample_n(10)
```

```
summary(amostra1$Receita)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  506665  702971 1093577 1112324 1431192 1893222
```

```
summary(amostra2$Receita)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  556027  887223 1084947 1177933 1511273 1937314
```

```
summary(amostra1$Custo_Marketing)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   67177   74916   84416  118045  117115  251298
```

```
summary(amostra2$Custo_Marketing)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   61045   71549   83628  111665  134629  211057
```

Realizando o teste de hipótese

```
teste_t_dados <- t.test (amostra1$Receita, amostra2$Receita, alternative
= "less" )
teste_t_dados
```

```
##
##  Welch Two Sample t-test
##
## data:  amostra1$Receita and amostra2$Receita
## t = -0.31108, df = 17.658, p-value = 0.3797
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##    -Inf 300496
```

```
## sample estimates:
## mean of x mean of y
##   1112324   1177933
```

```
teste_t_dados <- t.test (amostra1$Custo_Marketing,
amostra2$Custo_Marketing, alternative = "less" )
teste_t_dados
```

```
##
##  Welch Two Sample t-test
##
## data:  amostra1$Custo_Marketing and amostra2$Custo_Marketing
## t = 0.21891, df = 17.179, p-value = 0.5853
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 57052.16
## sample estimates:
## mean of x mean of y
##  118044.7  111664.5
```

Nosso estudo constata que tanto a media da Receita e do custo de marketing, não diferem significativamente em relação aos semestre, portanto, falhamos em rejeitar a hipótese nula, considerando que o valor p ficou acima do nível significância.

##Teste de Anova

H0 : Não há diferença significativa entre a média das receitas do 1º, 2º, 3º e 4º trimestre
H1 : Há diferença significativa entre os trimestres

##Analisando as variáveis com o glimpse (similar ao str)

```
glimpse(Dados)
```

```
## Rows: 28
## Columns: 18
## $ Data                <date> 2012-03-31, 2012-06-30, 2012-09-30, 2012-
12-31...
## $ Assinatura_Total    <dbl> 23410, 23938, 25101, 27146, 29174, 29807,
31092...
## $ Assinatura_Paga     <dbl> 22022, 22686, 23801, 25471, 27913, 28624,
29925...
## $ Teste_gratuito      <dbl> 1388, 1252, 1300, 1675, 1261, 1183, 1167,
1708,...
## $ Receita             <dbl> 506665, 532705, 556027, 589471, 638649,
671089,...
## $ Custo_Receita       <dbl> 360776, 378574, 399124, 420390, 440334,
452598,...
## $ Custo_Marketing     <dbl> 79381, 70959, 65955, 59777, 66965, 67177,
60637...
## $ Margem_Lucro        <dbl> 66508, 83172, 90948, 109304, 131350,
```

```
151314, 16...
## $ Margem_Contribuição <dbl> 0.131, 0.156, 0.164, 0.185, 0.206, 0.225,
0.237...
## $ Custo_Cliente       <dbl> 15.41119, 15.81477, 15.90072, 15.48626,
15.0933...
## $ Receita_Cliente     <dbl> 21.64310, 22.25353, 22.15159, 21.71484,
21.8910...
## $ Ganho_cliente       <dbl> 6.231909, 6.438758, 6.250866, 6.228579,
6.79766...
## $ Segmento            <chr> "Streaming", "Streaming", "Streaming",
"Streami...
## $ Mes                 <dbl> 3, 6, 9, 12, 3, 6, 9, 12, 3, 6, 9, 12, 3,
6, 9,...
## $ Ano                 <dbl> 2012, 2012, 2012, 2012, 2013, 2013, 2013,
2013,...
## $ Mes_Ano             <chr> "mar/2012", "jun/2012", "set/2012",
"dez/2012",...
## $ Semestre            <fct> 1ºSemestre, 1ºSemestre, 2ºSemestre,
2ºSemestre,...
## $ Trimestre           <fct> 1ºTrimestre, 2ºTrimestre, 3ºTrimestre,
4ºTrimes...
```

Análise descritiva

Resolvi verificar a distribuição dos dados, a homogeniedade da variância e alguns outliers

```
describeBy(Dados$Receita, group = Dados$Trimestre)

##
##  Descriptive statistics by group
## group: 1ºTrimestre
##    vars n    mean       sd median trimmed      mad    min     max
range skew
## X1    1 7 1054252 467887.9 984532 1054252 512806.1 506665 1820019
1313354 0.38
##    kurtosis     se
## X1    -1.52 176845
## ------------------------------------------------------------
## group: 2ºTrimestre
##    vars n    mean       sd  median trimmed      mad    min     max
range skew
## X1    1 7 1096418 480462.7 1025913 1096418 526062.1 532705 1893222
1360517 0.39
##    kurtosis       se
## X1    -1.47 181597.9
## ------------------------------------------------------------
## group: 3ºTrimestre
##    vars n    mean       sd  median trimmed      mad    min     max
range skew
```

```
## X1    1 7 1141011 489399.7 1063961 1141011 538002.9 556027 1937314
1381287 0.33
##    kurtosis       se
## X1    -1.53 184975.7
## ------------------------------------------------------------
## group: 4ºTrimestre
##    vars n    mean       sd  median trimmed      mad    min     max
range skew
## X1    1 7 1197604 505433.6 1105933 1197604 541710.9 589471 1996092
1406621 0.28
##    kurtosis       se
## X1    -1.62 191035.9
```

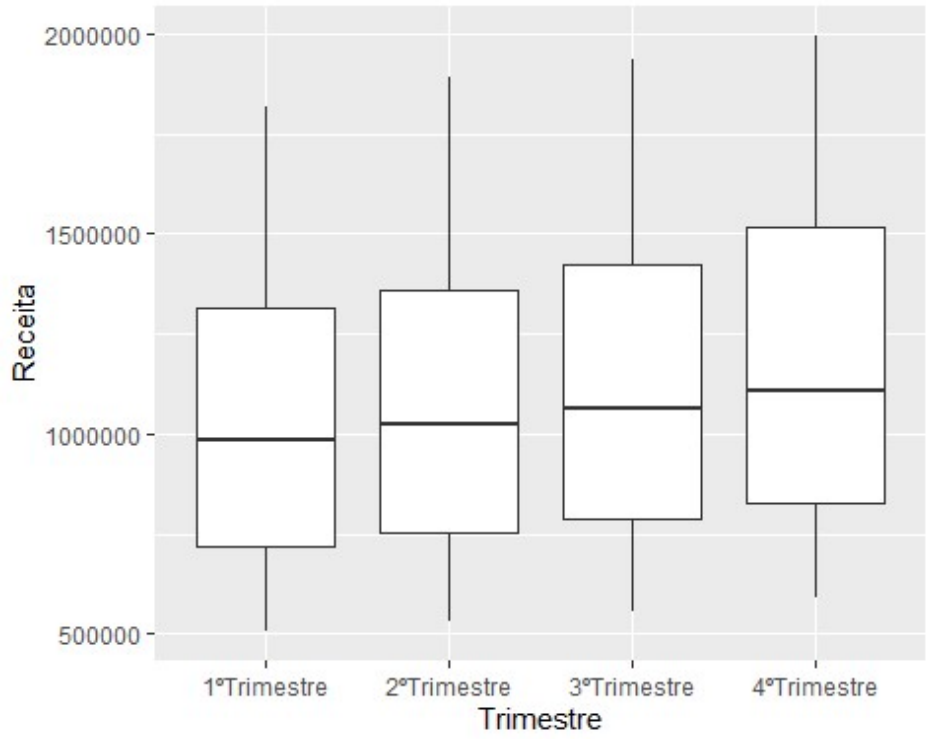
```
byf.shapiro(Receita ~ Trimestre, Dados) # verifica a distribuição normal
no grupo,se p maior que 0.05, é normalmente distribuida
```

```
##
##  Shapiro-Wilk normality tests
##
## data:  Receita by Trimestre
##
##                  W p-value
## 1ºTrimestre 0.9575  0.7970
## 2ºTrimestre 0.9595  0.8145
## 3ºTrimestre 0.9646  0.8571
## 4ºTrimestre 0.9626  0.8405
```

```
leveneTest(Receita ~ Trimestre, center = mean) # verifica a
homogeneiedade da variância
```

```
## Levene's Test for Homogeneity of Variance (center = mean)
##       Df F value Pr(>F)
## group  3  0.0389 0.9895
##       24
```

```
# verificando se tem outliers no pelo boxplot
ggplot (Dados, aes(x = Trimestre, y = Receita)) + geom_boxplot()
```



##Analise de variância H0 : Não há diferença significativa entre a média das receitas do 1º, 2º, 3º e 4º trimestre
H1 : Há diferença significativa entre os trimestres

```
tab_anova <- aov(Receita~Trimestre, Dados) # função nativa do R para fazermos a tabela anova
summary(tab_anova)
```

```
##             Df    Sum Sq   Mean Sq F value Pr(>F)
## Trimestre    3 7.925e+10 2.642e+10   0.112  0.952
## Residuals   24 5.668e+12 2.362e+11
```

O resultado da Anova demonstrou que realmente não há diferença significativa entre os trimestres Portanto, falhamos em rejeitar a hipótese nula com [F(3,24) = 0.112; p= 0.95]. Inclusive cabe destacar que o F muito pequena indica que o valor p não detém tamanha significância para este teste. Apesar disso, pode-se dizer que, apesar de não deter uma diferença relevante entre as média dos grupos, pode haver, algumas médias significativas entre eles, por isso, faremos o teste post-hoc Tukey HSD, por se mostrar o mais equilibrado.

Teste Tukey HSD

```
PostHocTest(tab_anova, method = "hsd")
```

Novamente o teste de Tukey HSD demonstrou que há diferenças entre as medias entre os trimestres, porém sem um nível de significância

Relatório

Foi, para mim, desafiador, pois trabalhei com conceitos e fórmulas que até então nunca havia utilizado, porém acrescentou um novo conjunto de ferramentas em minhas análise agora por diante. Particularmente a definição dos testes de hipóteses foi a parte mais difícil, entender o problema de negócio, conforme evidenciado ao longo do treinamento, é o mais complexo. Face ao exposto, pude procurar na internet técnicas de extração de dados a análise dos dados. Não obstante, o fato de não ter tido uma variável do tipo fator no DATASET de forma explicíta, complicou inicialmente minha análise, até que optei transformara a coluna de datas como fator para semestre e trimestre.

A partir daí, procurei testar diversas váriaveis do DATASET, achei algumas análises com outliers, cujo tratei (e trago em anexo o código fonte ao final deste relatório), e vi como se comportavam, porém fique com a Receita e se ela mantinha ao longo do tempo o seu comportamento.

Os teste de hipóteses feitos levaram a falha da rejeição da hipótese nula, ou seja, as médias dos grupos de tempo, não se alterava, o que pode inferir-se que a receita da Netflix a príncpio, vem mantendo uma consistência ao longo do período delimitado no dados.

Por fim, foi interessante fazer o plot de alguns gráficos para entender a relação de algumas variáveis, normalmente, após o lançamento do serviço de streaming da Netflix (apesar de estar no mercado desde 1997), houve uma necessidade de investir em conteúdos próprios (a partir da produção da série “House of Cards” em 2013), porém houve um crescimento exponencial no número de assinaturas.