AULA 4 - COMPARAÇÃO DE MÉDIAS

1. Moving Up

Essas pessoas estão em ascensão na vida, tanto em termos de carreira quanto financeiramente. Geralmente são profissionais jovens ou de meia-idade, que têm um forte foco em suas carreiras e buscam constantemente melhorias. Eles investem em educação e desenvolvimento pessoal e estão sempre em busca de novas oportunidades para subir na vida. Preferem produtos e serviços que reflitam seu status crescente e que possam facilitar ainda mais seu progresso. São consumidores exigentes, que valorizam qualidade e inovação.

2. Suburb Mix

Este grupo é composto por famílias que vivem em áreas suburbanas. Eles valorizam a segurança, a estabilidade e um bom ambiente para criar seus filhos. As atividades desse grupo incluem visitas regulares a parques, atividades extracurriculares para os filhos e participação em eventos comunitários. Preferem serviços e produtos que atendam às necessidades de toda a família, como supermercados, escolas de qualidade e entretenimento familiar. Embora sejam cuidadosos com seus gastos, estão dispostos a pagar mais por conveniência e produtos de qualidade.

3. Travelers

Este segmento adora explorar novos lugares e experimentar diferentes culturas. São aventureiros, abertos a novas experiências e estão sempre planejando sua próxima viagem, seja para destinos exóticos ou para descobrir cidades vizinhas. Eles preferem gastar dinheiro em experiências, como viagens, gastronomia e atividades culturais, ao invés de bens materiais. Valorizam produtos que possam facilitar suas viagens, como equipamentos de viagem de alta qualidade, seguros de viagem e programas de milhas aéreas.

4. Urban Hip

Esse grupo é composto principalmente por jovens adultos que vivem em grandes cidades e valorizam o estilo de vida urbano. Eles estão antenados nas últimas tendências, tanto na moda quanto na tecnologia, e têm um forte senso de individualidade. Frequentam restaurantes modernos, eventos culturais, e adoram descobrir novos lugares "da moda" na cidade. Estão sempre conectados e preferem produtos que sejam únicos, inovadores e que lhes permitam expressar sua personalidade. Eles gostam de marcas que apoiam causas sociais e ambientais e que se alinham com seus valores pessoais

Base de Dados

	age	gender	income	kids	ownHome	subscribe	Segment
1	47	Male	49482.8104	2	ownNo	subNo	Suburb mix
2	31	Male	35546.2883	1	ownYes	subNo	Suburb mix
3	43	Male	44169.1864	0	ownYes	subNo	Suburb mix
4	37	Female	81041.9864	1	ownNo	subNo	Suburb mix
5	41	Female	79353.0144	3	ownYes	subNo	Suburb mix
6	43	Male	58143.3633	4	ownYes	subNo	Suburb mix
7	38	Male	19282.2306	3	ownNo	subNo	Suburb mix
8	28	Male	47245.2385	0	ownNo	subNo	Suburb mix

library(lattice)

Puxando Dados

```
seg.df <- read.csv("trabalho.csv", sep=",", header = TRUE)
seg.df$age <- round(seg.df$age, 0)
View(seg.df)
```

Encontrando Descritivos por Grupo em R

Usando a Função `by()`para descobrir a média da renda por segmento

```
by(seg.df$income, seg.df$Segment, mean)
```

```
seg.df$Segment: Moving up
[1] 53090.97
------------------------------------------------------------------------
seg.df$Segment: Suburb mix
[1] 55033.82
------------------------------------------------------------------------
seg.df$Segment: Travelers
[1] 62213.94
------------------------------------------------------------------------
seg.df$Segment: Urban hip
[1] 21681.93
```

```
by(seg.df$income, list(seg.df$Segment, seg.df$subscribe), mean)
```

```
: Moving up
: subNo
[1] 53633.73
------------------------------------------------------------------------
: Suburb mix
: subNo
[1] 54942.69
```

```
------------------------------------------------------------------------
: Travelers
: subNo
[1] 62746.11
------------------------------------------------------------------------
: Urban hip
: subNo
[1] 22082.11
------------------------------------------------------------------------
: Moving up
: subYes
[1] 50919.89
------------------------------------------------------------------------
: Suburb mix
: subYes
[1] 56461.41
------------------------------------------------------------------------
: Travelers
: subYes
[1] 58488.77
------------------------------------------------------------------------
: Urban hip
: subYes
[1] 20081.19
```

Usando a Função `aggregate()` para o mesmo exercício feito anteriormente

aggregate(seg.df$income, list(seg.df$Segment), mean)

```
    Group.1        x
1  Moving up 53090.97
2 Suburb mix 55033.82
3  Travelers 62213.94
4  Urban hip 21681.93
```

#Variáveis respostas a esquerda e explicativas a direita
#y ~ x # Fórmula Simples

#Agregar renda por segmento
#aggregate(formula, data, FUN)
aggregate(income ~ Segment, data = seg.df, mean)

```
    Segment   income
1  Moving up 53090.97
2 Suburb mix 55033.82
3  Travelers 62213.94
4  Urban hip 21681.93
```

#Descritivo para grupo de duas variáveis
aggregate(income ~ Segment + ownHome, data = seg.df, mean)

```
    Segment ownHome   income
1  Moving up   ownNo 54497.68
2 Suburb mix   ownNo 54932.83
3  Travelers   ownNo 63188.42
4  Urban hip   ownNo 21337.59
5  Moving up  ownYes 50216.37
6 Suburb mix  ownYes 55143.21
7  Travelers  ownYes 61889.12
8  Urban hip  ownYes 23059.27
```

#Estrutura permite quantas quiser
aggregate(income ~ Segment + ownHome + subscribe, data = seg.df, mean)

```
     Segment ownHome subscribe   income
1   Moving up   ownNo     subNo 55402.89
2  Suburb mix   ownNo     subNo 54579.99
3   Travelers   ownNo     subNo 65852.54
4   Urban hip   ownNo     subNo 21604.16
5   Moving up  ownYes     subNo 49898.85
6  Suburb mix  ownYes     subNo 55354.86
7   Travelers  ownYes     subNo 61749.71
8   Urban hip  ownYes     subNo 23993.93
9   Moving up   ownNo    subYes 50675.70
10 Suburb mix   ownNo    subYes 63753.97
11  Travelers   ownNo    subYes 48091.75
12  Urban hip   ownNo    subYes 20271.33
13  Moving up  ownYes    subYes 51359.44
14 Suburb mix  ownYes    subYes 52815.13
15  Travelers  ownYes    subYes 62944.64
16  Urban hip  ownYes    subYes 19320.64
```

#Atribuir resultado a dataframe
agg.data <- aggregate(income ~ Segment + ownHome, data = seg.df, mean)

#Obter frequencias de diferentes Segmentos e Casa Própria
with(seg.df, table(Segment, ownHome))

```
           ownHome
Segment      ownNo ownYes
  Moving up     47     23
  Suburb mix    52     48
  Travelers     20     60
  Urban hip     40     10
```

#Obter frequencias de crianças por segmentos
with(seg.df, table(kids, Segment))

```
kids Moving up Suburb mix Travelers Urban hip
   0         13          11         80         17
   1         17          36          0         17
   2         18          22          0         11
   3         13          19          0          4
   4          5           7          0          1
   5          3           3          0          0
   6          0           2          0          0
   7          1           0          0          0
```

#Somatório de filhos por segmentos
xtabs(kids ~ Segment, data = seg.df)

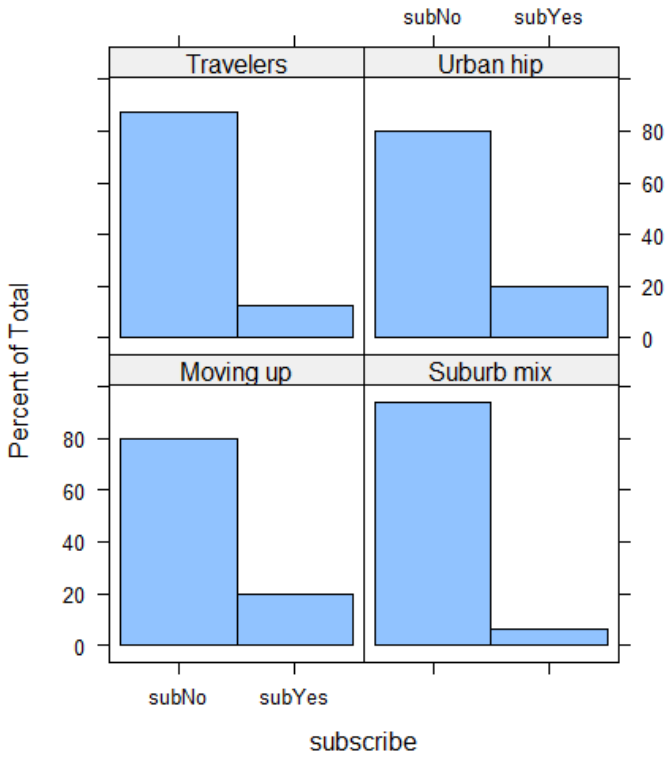
```
    Segment kids
1  Moving up  134
2 Suburb mix  192
3  Travelers    0
4  Urban hip   55
```

#ou

aggregate(kids ~ Segment, data = seg.df, sum)

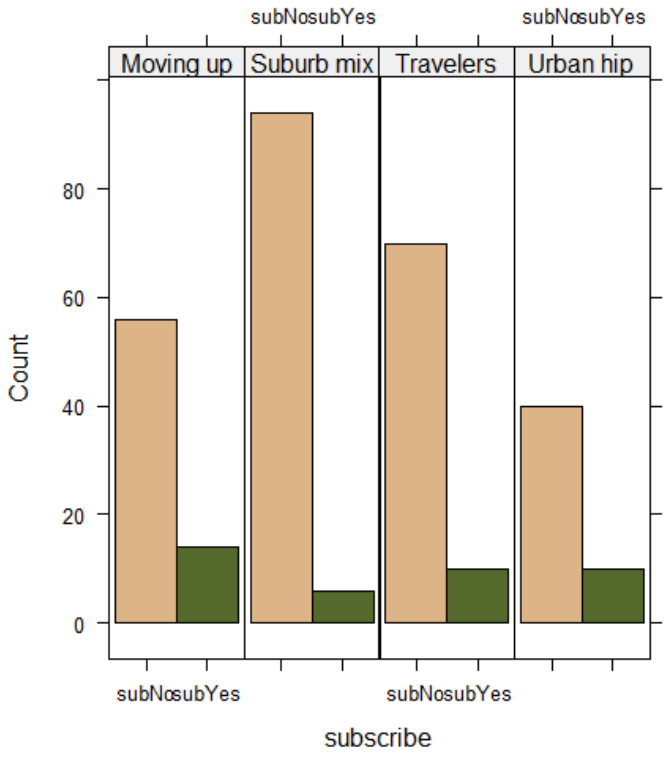
```
Segment
 Moving up Suburb mix  Travelers  Urban hip
       134         192           0         55
```

#Proporção subscribe por segmento. Proporção vem por default
seg.df$subscribe <- as.factor(seg.df$subscribe)
histogram(~subscribe | Segment, data=seg.df)

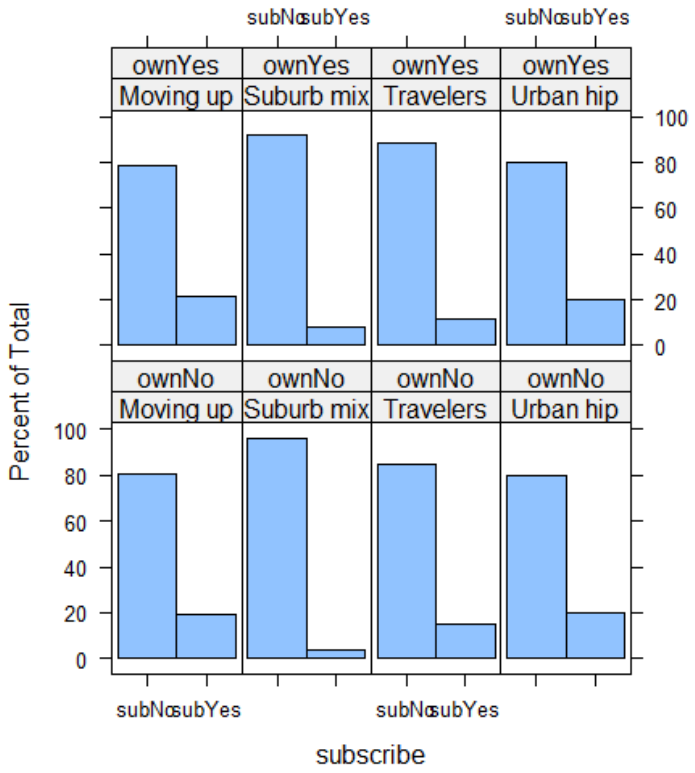


#se quisermos contar os subscribes

histogram(~subscribe | Segment, data=seg.df, type="count", layout=c(4,1),
col=c("burlywood", "darkolivegreen"))

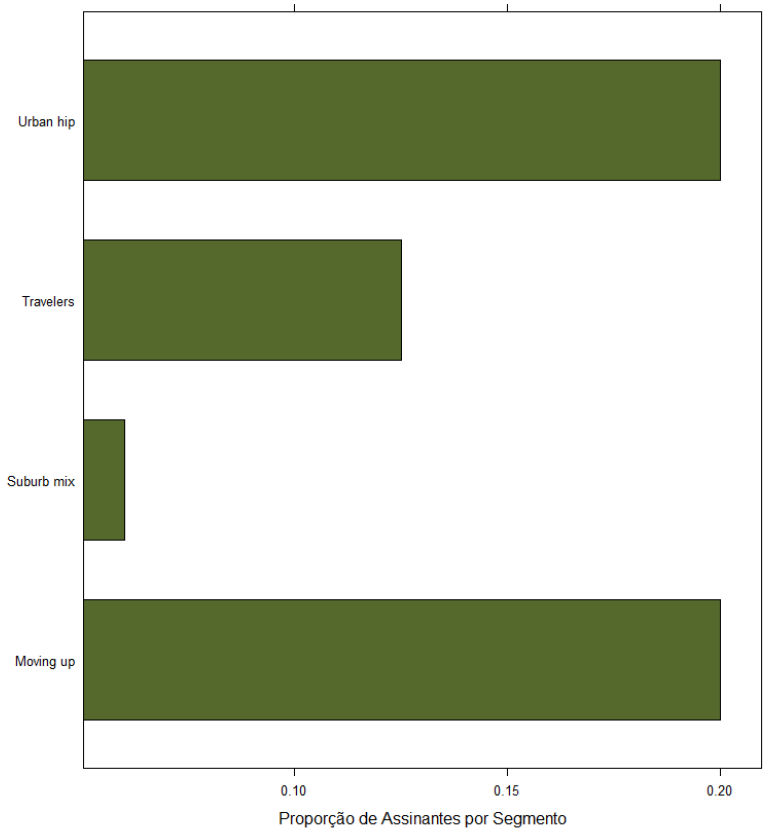


#Ver dentro de cada segmento, discriminado por casa própria ou não
histogram(~subscribe | Segment + ownHome, data=seg.df)



#Por fim, poderíamos plotar apenas as proporções de "sim" em vez de barras de "sim" e "não".

```
prop.table(table(seg.df$subscribe, seg.df$Segment), margin=2)
barchart(prop.table(table(seg.df$subscribe, seg.df$Segment), margin=2)[2,],
        xlab="Proporção de Assinantes por Segmento", col="darkolivegreen")
```

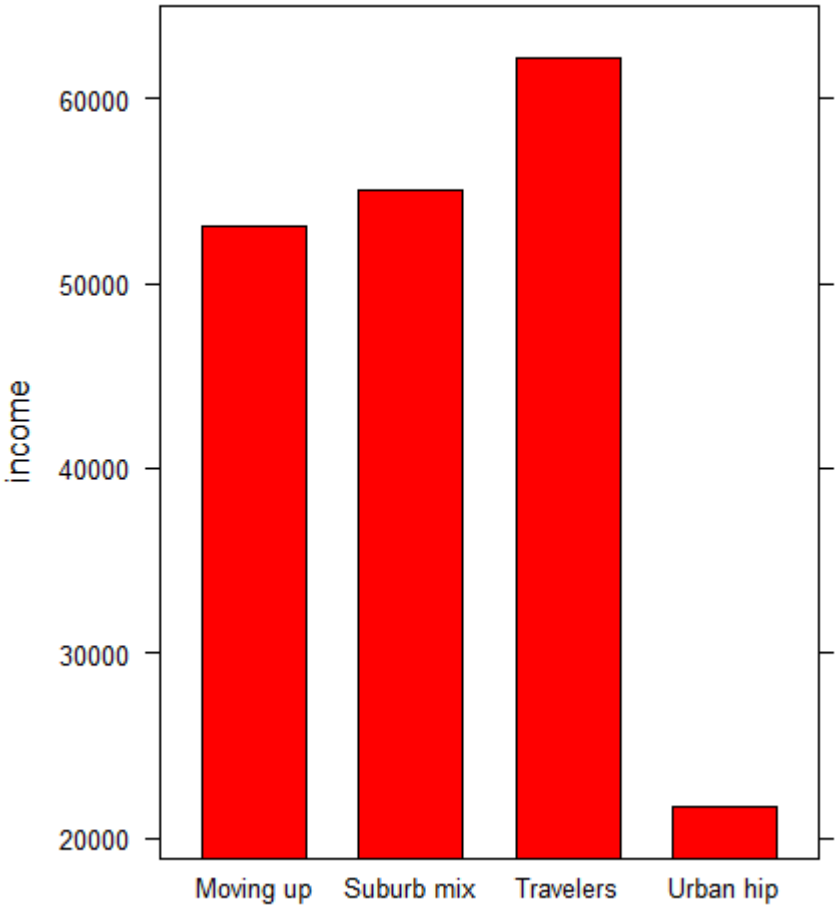


#O resultado comunica fortemente que o segmento "Suburb mix" tem uma taxa de #assinatura aparentemente baixa

#Visualização por Grupos - dados contínuos

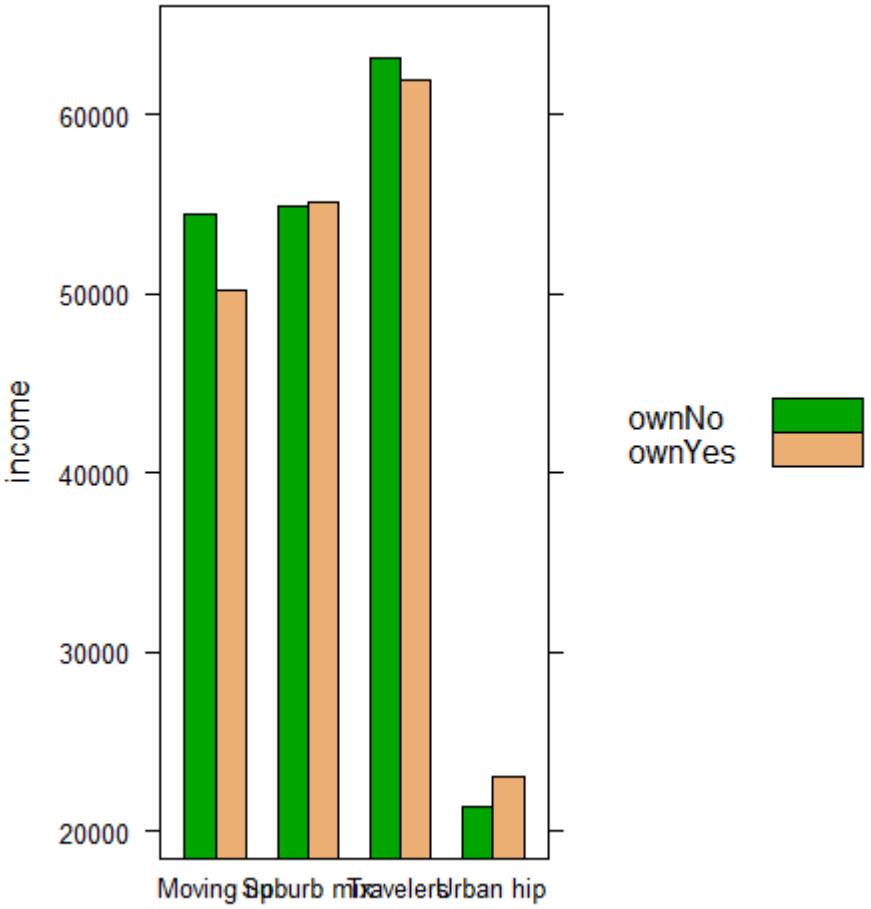
#Mas e quanto aos dados contínuos? Como plotar a renda por segmento em nossos dados?

```
seg.mean <- aggregate(income ~ Segment, data=seg.df, mean)
library(lattice)
barchart(income ~ Segment, data=seg.mean, col="red")
```



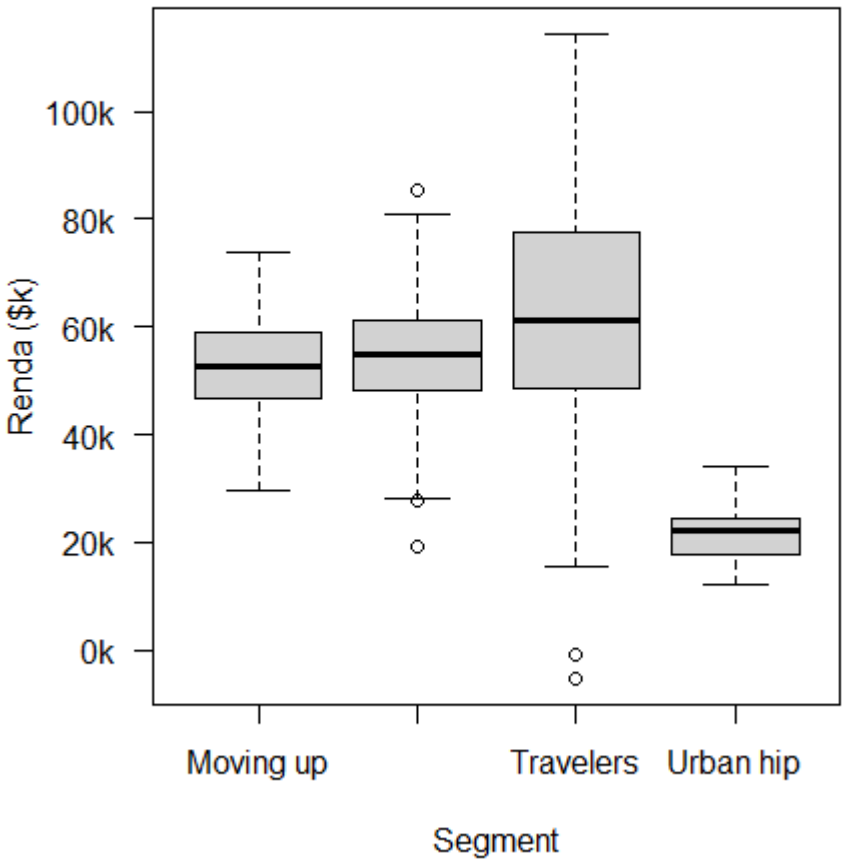
#Dividindo ainda mais os dados de posse por casa

```
seg.income.agg <- aggregate(income ~ Segment + ownHome, data=seg.df, mean)
barchart(income ~ Segment, data=seg.income.agg,
        groups=ownHome, auto.key=TRUE,
        par.settings = simpleTheme(col=terrain.colors(3)))
```



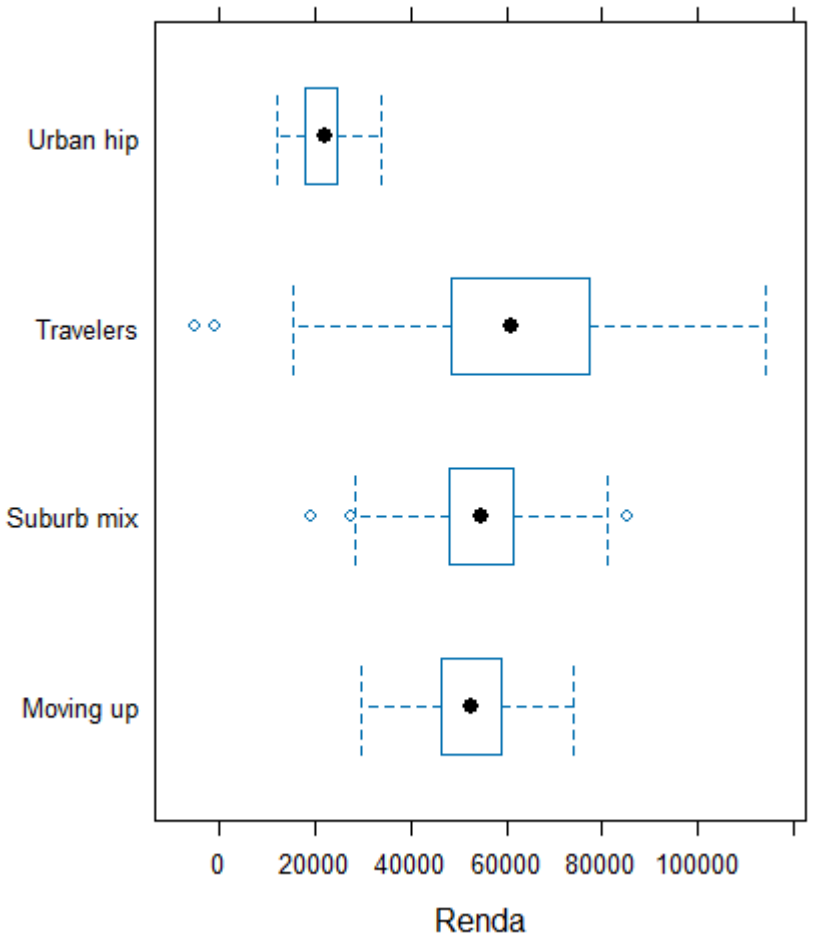
#Usando o `boxplot()` para plotar um box-and-whiskers plot por fator para comparar valores de dados contínuos, como a renda para diferentes grupos.

```
boxplot(income ~ Segment, data=seg.df, yaxt="n", ylab="Renda ($k)")
ax.seq <- seq(from=0, to=120000, by=20000)
axis(side=2, at=ax.seq, labels=paste(ax.seq/1000, "k", sep=""), las=1)
```



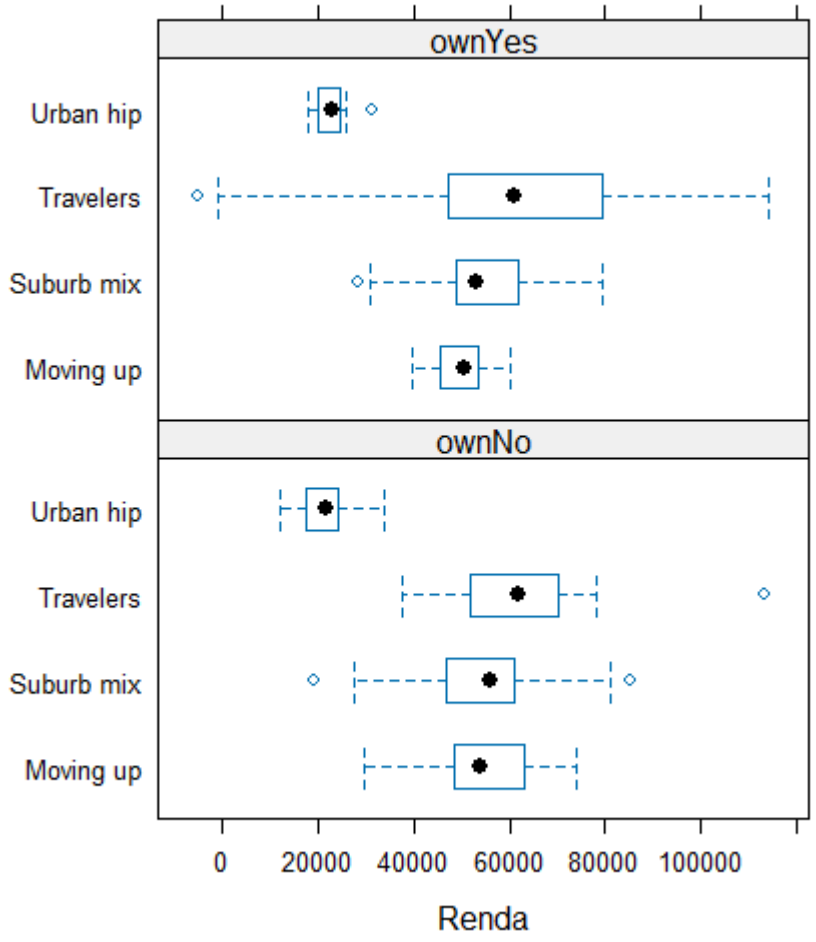
#Para plotar um gráfico mais bonito

bwplot(Segment ~ income, data=seg.df, horizontal=TRUE, xlab="Renda")



#detalhar a posse de casa como uma variável de condicionamento usando `| ownHome` na fórmula

bwplot(Segment ~ income | ownHome, data=seg.df, horizontal=TRUE, xlab="Renda")



#Neste gráfico, descobrimos—entre outras coisas—que, em nossos dados simulados, o segmento "Travelers" tem uma distribuição muito mais ampla de renda entre aqueles que possuem suas casas do que entre aqueles que não possuem.

#Dados para comparação de Grupos
#Verifica um resumo dos dados
summary(seg.df)

```
      age            gender             income            kids          ownHome
 Min.   :19.00   Length:300         Min.   : -5183   Min.   :0.00   Length:300
 1st Qu.:33.00   Class :character   1st Qu.: 39656   1st Qu.:0.00   Class :character
 Median :39.50   Mode  :character   Median : 52014   Median :1.00   Mode  :character
 Mean   :41.17                      Mean   : 50937   Mean   :1.27
 3rd Qu.:48.00                      3rd Qu.: 61403   3rd Qu.:2.00
 Max.   :80.00                      Max.   :114278   Max.   :7.00
  subscribe     Segment
 subNo :260   Length:300
 subYes: 40   Class :character
              Mode  :character
```

#Teste de Chi-Quadrado

#Tamanho da amostra por segmento
chisq.test(table(seg.df$Segment))

```
      Chi-squared test for given probabilities

data:  table(seg.df$Segment)
X-squared = 17.333, df = 3, p-value = 0.0006035
```

#O valor de p é 0.0006, o que indica que os tamanhos
#dos segmentos são significativamente diferentes.

#Independência entre fatores

#Para verificar se o status de assinatura é independente da posse de casa, construímos uma tabela cruzada e aplicamos o teste qui-quadrado:

table(seg.df$subscribe, seg.df$ownHome)
chisq.test(table(seg.df$subscribe, seg.df$ownHome))

```
      Pearson's Chi-squared test with Yates' continuity correction

data:  table(seg.df$subscribe, seg.df$ownHome)
X-squared = 0.010422, df = 1, p-value = 0.9187
```

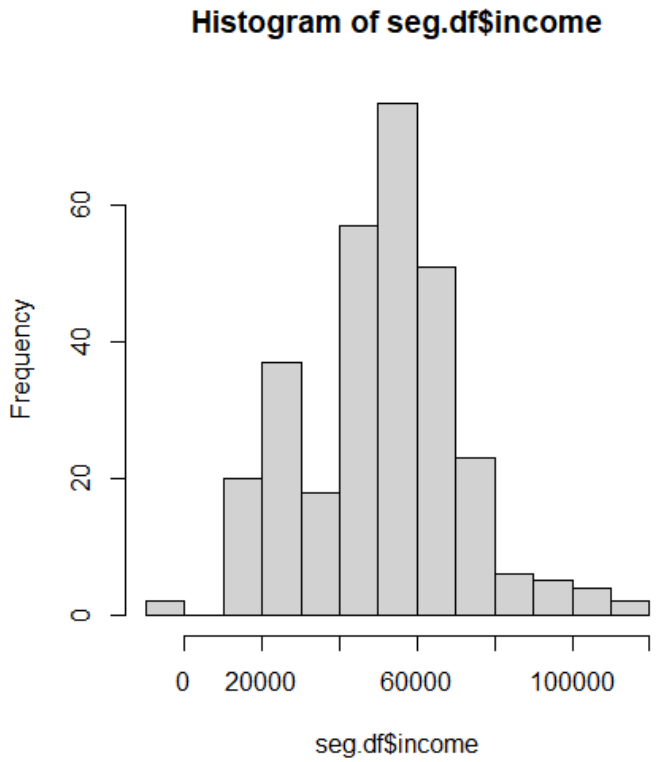
#O valor de p é 0.919, indicando que não há evidências suficientes para sugerir uma relação entre status de assinatura e posse de casa.

#TESTE T: Média entre grupos

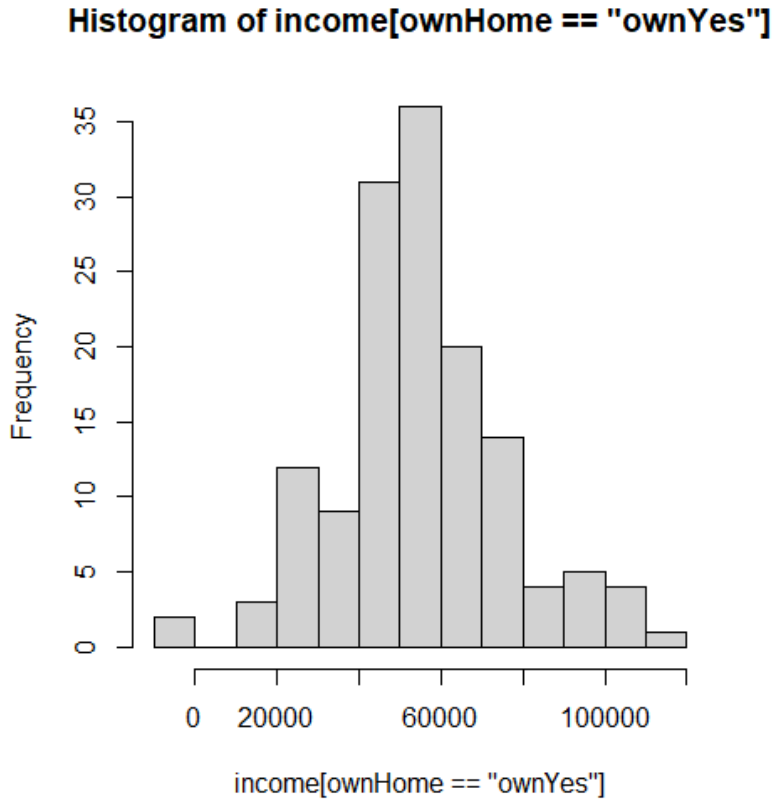
#Um teste t compara a média de uma amostra com a média de outra amostra (ou com um #valor específico, como 0). O ponto importante é que ele compara a média de exatamente #dois conjuntos de dados. Por exemplo, nos dados segmentados, poderíamos querer saber #se a renda domiciliar é diferente entre aqueles que possuem uma casa e aqueles que não #possuem.

#Verificando distribuição

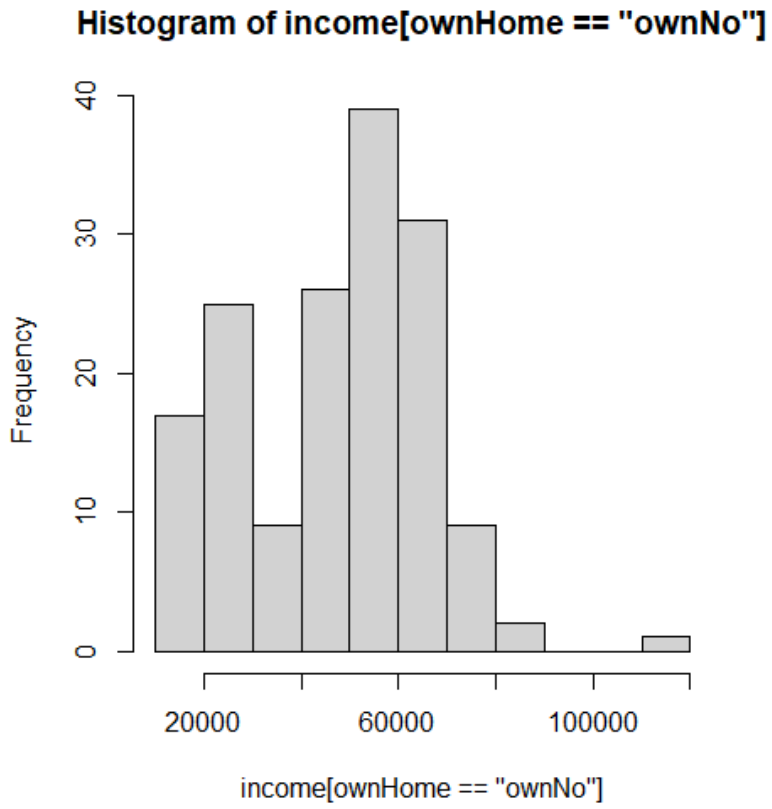
hist(seg.df$income)




```
with(seg.df, hist(income[ownHome == "ownYes"]))
```



with(seg.df, hist(income[ownHome == "ownNo"]))



#Teste T de renda por status de casa

t.test(income ~ ownHome, data=seg.df)

```
        Welch Two Sample t-test

data:  income by ownHome
t = -3.2731, df = 285.25, p-value = 0.001195
alternative hypothesis: true difference in means between group ownNo and
group ownYes is not equal to 0
95 percent confidence interval:
 -12080.155  -3007.193
sample estimates:
 mean in group ownNo mean in group ownYes
            47391.01             54934.68
```

#Há várias informações importantes na saída do `t.test()`.
#Primeiro, vemos que a estatística #t é -3.2, com um p-valor de 0.0012.
#Isso significa que a hipótese nula de nenhuma
#diferença na renda por posse de casa é rejeitada.
#Os dados sugerem que pessoas que #possuem suas casas têm uma renda mais alta.

#Mesma diferença mas dentro do grupo de viajantes

```
t.test(income ~ ownHome, data=subset(seg.df, Segment == "Travelers"))
```

```
      Welch Two Sample t-test

data:  income by ownHome
t = 0.26561, df = 53.833, p-value = 0.7916
alternative hypothesis: true difference in means between group ownNo and
group ownYes is not equal to 0
95 percent confidence interval:
 -8508.993 11107.604
sample estimates:
 mean in group ownNo mean in group ownYes
            63188.42             61889.12
```

#O intervalo de confiança de -8508 a 11107 inclui 0, e,
#portanto, concluímos—como evidenciado pelo p-valor de 0.79—que não há uma #diferença significativa na renda média
#entre os "Travelers" em nossos dados que possuem casas e os que não possuem.

#Localizando onde está a diferença de salario: ANOVA

#Uma análise de variância (ANOVA) compara as médias de múltiplos grupos. #Tecnicamente, isso é feito comparando o grau em que os grupos diferem, medido pela #variância em suas médias (entre os grupos), em relação à variância das observações em #torno de cada média (dentro de cada grupo).

#Renda por Status de Casa

```
seg.aov.own <- aov(income ~ ownHome , data=seg.df)
anova(seg.aov.own)
```

```
Analysis of Variance Table

Response: income
           Df     Sum Sq    Mean Sq F value   Pr(>F)
ownHome     1 4.2527e+09 4252661211  10.832 0.001118 **
Residuals 298 1.1700e+11  392611030
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

#Renda por Segmento

```
seg.aov.seg <- aov(income ~ Segment , data=seg.df)
anova(seg.aov.seg)
```

```
Analysis of Variance Table

Response: income
```

```
           Df     Sum Sq    Mean Sq F value    Pr(>F)
Segment     3 5.4970e+10 1.8323e+10  81.828 < 2.2e-16 ***
Residuals 296 6.6281e+10 2.2392e+08
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

#E para ambos?

anova(aov(income ~ Segment + ownHome , data=seg.df))

```
Analysis of Variance Table

Response: income
           Df     Sum Sq    Mean Sq F value Pr(>F)
Segment     3 5.4970e+10 1.8323e+10 81.6381 <2e-16 ***
ownHome     1 6.9918e+07 6.9918e+07  0.3115 0.5772
Residuals 295 6.6211e+10 2.2444e+08
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

#Rodar uma anova com todos as variáveis

seg.aov.step <- step(aov(income ~ ., data=seg.df))

```
income ~ age + gender + kids + ownHome + subscribe + Segment

            Df  Sum of Sq        RSS    AIC
- age        1 4.0618e+06 6.5661e+10 5777.2
- ownHome    1 1.0320e+08 6.5760e+10 5777.6
- kids       1 1.3383e+08 6.5790e+10 5777.8
- subscribe  1 1.5958e+08 6.5816e+10 5777.9
- gender     1 2.6883e+08 6.5925e+10 5778.4
<none>                    6.5657e+10 5779.2
- Segment    3 1.9353e+10 8.5010e+10 5850.7

Step:  AIC=5777.19
income ~ gender + kids + ownHome + subscribe + Segment

            Df  Sum of Sq        RSS    AIC
- ownHome    1 1.0159e+08 6.5762e+10 5775.7
- kids       1 1.3205e+08 6.5793e+10 5775.8
- subscribe  1 1.5794e+08 6.5819e+10 5775.9
- gender     1 2.7009e+08 6.5931e+10 5776.4
<none>                    6.5661e+10 5777.2
- Segment    3 4.9044e+10 1.1470e+11 5938.6

Step:  AIC=5775.66
income ~ gender + kids + subscribe + Segment

            Df  Sum of Sq        RSS    AIC
- kids       1 1.0707e+08 6.5869e+10 5774.1
```

```
- subscribe  1 1.6370e+08 6.5926e+10 5774.4
- gender     1 2.5520e+08 6.6017e+10 5774.8
<none>                    6.5762e+10 5775.7
- Segment    3 5.2897e+10 1.1866e+11 5946.7

Step:  AIC=5774.15
income ~ gender + subscribe + Segment

            Df  Sum of Sq        RSS    AIC
- subscribe  1 1.6226e+08 6.6032e+10 5772.9
- gender     1 2.4390e+08 6.6113e+10 5773.3
<none>                    6.5869e+10 5774.1
- Segment    3 5.3005e+10 1.1887e+11 5945.3

Step:  AIC=5772.88
income ~ gender + Segment

          Df  Sum of Sq        RSS    AIC
- gender   1 2.4949e+08 6.6281e+10 5772.0
<none>                  6.6032e+10 5772.9
- Segment  3 5.4001e+10 1.2003e+11 5946.2

Step:  AIC=5772.02
income ~ Segment

          Df Sum of Sq        RSS    AIC
<none>                  6.6281e+10 5772.0
- Segment  3 5.497e+10 1.2125e+11 5947.2
```