**DIRETORIA DE DESENVOLVIMENTO E ENSINO**

**DEPARTAMENTO DE ENSINO SUPERIOR**

**UNIDADE ACADÊMICA DE INFORMAÇÃO E COMUNICAÇÃO**

**CURSO SUPERIOR TECNOLÓGICO EM SISTEMAS PARA INTERNET**

**RELATÓRIO DESCRITIVO E ANALÍTICO DE UM BANCO DE DADOS UTILIZANDO O R COMO FERRAMENTA COMPUTACIONAL**

**DISCENTE: Natanael Guedes da Silva Neto**

**Nayanna de Oliveira Silva**

João Pessoa, 23 de fevereiro de 2018.

# Introdução

O presente relatório descreve um estudo que foi feito a partir de um banco de dados utilizando os conceitos que foram adquiridos ao longo da disciplina de estatística e que é apresentado na prática no software estatistico RStudio.

Esse trabalho consiste em fazer uma análise detalhada de um banco de dados utilizando o software RStudio, o mesmo possui funções pré-definidas e que são essenciais para realizar cálculos estatístico, além da criação de gráficos, o que contribuiu para uma análise mais descritiva e eficiente.

Metodologia

Para o desenvolvimento deste estudo da metodologia estatística, foi dividida em três partes: estatística descritiva, inferência, correlação e regressão.

Inicialmente foi abordado a estatística descritiva para fazer uma análise de um banco de dados com o auxílio do software RStudio com intuito de analisar e interpretar os dados de uma amostra ou população, bem como criar gráficos de acordo com cada tipo de variável e realizar cálculos estatístico utilizando as funções pré-definidas do RStudio.

Na segunda parte deste trabalho foi a inferência estatística, na qual utilizamos o RStudio para fazer uma análise dos dados, e em seguida, realizar testes de médias, de hipótese e de proporção com o propósito de estudar a população através de evidências fornecidas pela amostra e comparar uma com a outra.

Na terceira parte foi empregado os conceitos de correlação e regressão que estudamos na disciplina de estatística, onde utilizamos as funções pré definidas do software RStudio para calcular a correlacão das variáveis quantitativas, tanto da população quanto da amostra , comparando-as com o objetivo de testar para saber se ambas estão correlacionadas ou não , isto é, se elas possuem um relacionamento , além de construir o gráfico de pontos para cada uma. Em relação a regressão,foi feito uma análise para verificar se há uma relação dependente ou independente entre as variáveis de um banco de dados, comparando a amostra com a população, utilizando as funções pré-definidas do RStudio para realizar os cálculos de regressão, além de criar o modelo de regressão que é o gráfico de ponto com a reta.

# Resultados encontrados

# Por meio de entrevistas foi realizada uma análise de dados a partir de uma amostra de 60 pessoas de uma população de 533, no qual foram feitas frequências, correlação, regressão, média e proporções.

# Nas entrevistas foi questionado a cidade e bairro em que cada pessoa reside e onde nasceu, por meio dos resultados foi mostrado que Recife é a cidade mais frequente entre entrevistados, por segundo Florianópolis e por fim Vitória. Como também a maioria nasceu onde atualmente reside, há pessoas que nasceram em cidades diferentes mas não chegando a ser em região diferente, apenas 4 em outra região. E 95% das pessoas moram no bairro aplicado. As três tabelas a seguir representam os resultados adquiridos.

# 

| **NCidade** | **Frequência Absoluta** | **Frequência Relativa** | **Frequência Acumulada** | **F. Acumulada Relativa** |
| --- | --- | --- | --- | --- |
| Recife | 33 | 55% | 33 | 55% |
| Vitória | 11 | 18% | 44 | 73% |
| Florianópolis | 16 | 27% | 60 | 100% |

Tabela para variável NCidade

| **Cidade** | **Frequência Absoluta** | **Frequência Relativa** | **Frequência Acumulada** | **F. Acumulada Relativa** |
| --- | --- | --- | --- | --- |
| Onde foi aplicado o questionário | 42 | 70% | 42 | 8% |
| Outra no mesmo estado | 9 | 15% | 51 | 35% |
| Outra em outro estado, mas na mesma região | 5 | 8% | 56 | 67% |
| Outra região | 4 | 7% | 60 | 100% |

Tabela para variável Cidade

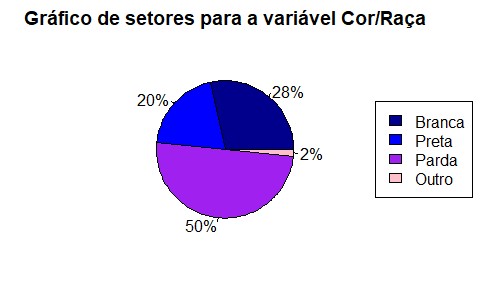
| **Bairro** | **Frequência Absoluta** | **Frequência Relativa** | **Frequência Acumulada** | **F. Acumulada Relativa** |
| --- | --- | --- | --- | --- |
| Onde foi aplicado o questionário | 57 | 95% | 57 | 95% |
| Outro na mesma região | 3 | 5% | 60 | 100% |

Tabela para a variável Bairro

Os locais da aplicação de entrevistas variam entre a Unidade de saúde, casa do entrevistado tanto sozinho como acompanhado e outros locais. Na tabela de frequências abaixo é possível perceber que grande parte dos locais não são comuns estando representados por “outros” e por fim nas casas seja sozinho ou acompanhado.

| **Local** | **Frequência Absoluta** | **Frequência Relativa** | **Frequência Acumulada** | **F. Acumulada Relativa** |
| --- | --- | --- | --- | --- |
| Unidade de saúde | 5 | 8% | 5 | 8% |
| Casa da entrevistada (sozinha) | 16 | 27% | 21 | 35% |
| Casa da entrevistada (com presença de outras pessoas) | 19 | 32% | 40 | 67% |
| Outro | 20 | 33% | 60 | 100% |

Tabela para a variável Local



Nosso gráfico de setores para raça nos mostra que metade das pessoas da amostra se declaram pardas, e com umas diferença de 8% há mais brancos do que pretos, os outros não se declaram como nenhum desses.

| **Religião** | **Frequência Absoluta** | **Frequência Relativa** | **Frequência Acumulada** | **F. Acumulada Relativa** |
| --- | --- | --- | --- | --- |
| Sim | 60 | 100% | 60 | 100% |

Tabela para a variável Religião

Por unanimidade todos entrevistados possuem e praticam uma religião, predominantemente Católicos e Evangélicos. Espíritas e Umbanda chegam a ter abaixo de 5 praticantes.

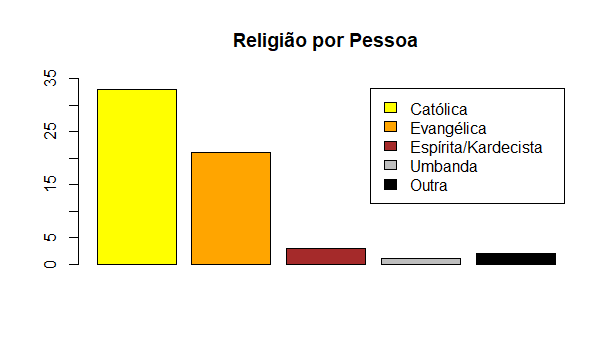
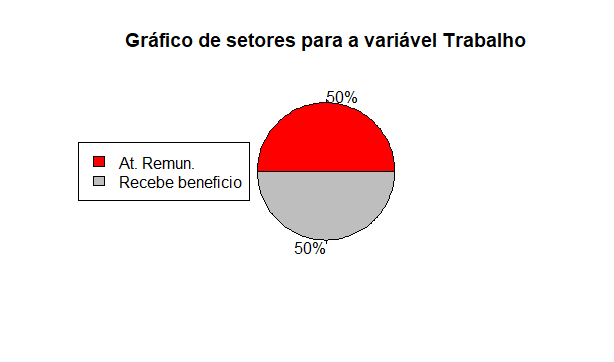


Gráfico de barras para a variável Religião-qual

Metade dos entrevistados exercem um trabalho remunerado e a outra metade recebe benefícios sejam pensão, aposentadoria ou bolsa de estudo.



Mais da metade da amostra não estuda e a maioria possui o ensino médio incompleto e por pouco menos completo, pouco mais de 10 pessoas possuem ensino fundamental incompleto e por menos incompleto.

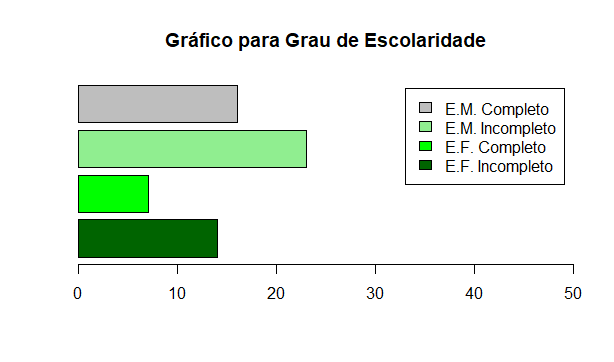
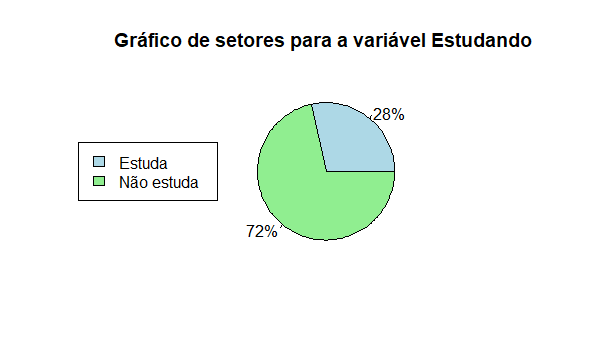


Gráfico de barras para variável Estudo

Das pessoas que se encaixam nas alternativas de habitações pela entrevista, quase todos moram em uma casa e só uma pessoas que não se incluiu em nenhum desses tipos, declarou residir em um barraco. Por volta da metade tem sua habitação própria e maior parte dos resto tem habitações alugadas.



Gráfico de setores para a variável Habitação

| **Outra habitação** | **Frequência Absoluta** | **Frequência Relativa** | **Frequência Acumulada** | **F. Acumulada Relativa** |
| --- | --- | --- | --- | --- |
| Barraco | 1 | 2% | 1 | 2% |
| Não se aplica | 59 | 98% | 60 | 100% |

Tabela para a variável Habit-outro

| **Habitação**  **(propriedade)** | **Frequência Absoluta** | **Frequência Relativa** | **Frequência Acumulada** | **F. Acumulada Relativa** |
| --- | --- | --- | --- | --- |
| Próprio (comprado) | 28 | 47% | 28 | 47% |
| Próprio (“pós-invasão”) | 8 | 13% | 36 | 60% |
| Alugado | 18 | 30% | 54 | 90% |
| Cedido/Emprestado | 5 | 8% | 59 | 98% |
| Outro | 1 | 2% | 60 | 100% |

Tabela para a variável Habit-alug

Das classes relatadas predominam classe média e baixa, 38% e 28% respectivamente.

| **Classe** | **Frequência Absoluta** | **Frequência Relativa** | **Frequência Acumulada** | **F. Acumulada Relativa** |
| --- | --- | --- | --- | --- |
| Não sabe | 4 | 7% | 4 | 7% |
| Abaixo da linha da pobreza | 1 | 2% | 5 | 9% |
| Humilde | 1 | 2% | 6 | 11% |
| Pobre | 8 | 13% | 14 | 24% |
| Baixa | 17 | 28% | 31 | 52% |
| Nobre | 1 | 2% | 32 | 54% |
| Média-baixa | 4 | 6% | 36 | 60% |
| Estável | 1 | 2% | 37 | 62% |
| Média | 23 | 38% | 60 | 100% |

Tabela para a variável Classe

O estado civil dos entrevistados na grande maioria foi declarado como solteiro, apenas 10 pessoas casadas e 3 pessoas se declaram como amigadas, união consensual e moram junto.

| **Civil** | **Frequência Absoluta** | **Frequência Relativa** | **Frequência Acumulada** | **F. Acumulada Relativa** |
| --- | --- | --- | --- | --- |
| Solteira | 47 | 78% | 47 | 78% |
| Amigada | 1 | 2% | 48 | 80% |
| Mora junto | 1 | 2% | 49 | 82% |
| União Consensual | 1 | 2% | 50 | 84% |
| Casada | 10 | 16% | 60 | 100% |

Tabela para a variável Civil

Ao realizamos teste para média da amostra Renda da família contra a média da população , conclui-se que, a média da renda da família na amostra não é igual a 265.5392, ou seja, devemos rejeitar a hipótese nula em favor da alternativa

O teste para a média da amostra da variável renda div pessoa com ela mesma na população, conclui-se que,devemos rejeitar a hipótese nula, em favor da alternativa, logo não há diferença entre as médias.

Ao realizarmos o teste para a média entre a amostra e população de Recife concluímos que, a média não é igual a 0. Dessa forma, devemos rejeitar a hipótese nula em favor da alternativa com 95% de confiança

Ao realizarmos o teste de hipótese da idade na amostra contra a idade da população, descobrimos que a idade da amostra é menor do que a da população, ou seja, devemos rejeitar a hipótese nula em favor da alternativa com 95% de confiança.

Ao realizamos o teste de hipótese, verificamos a variável recursos e concluímos com 95% de confiança que a média da amostra é maior que a da população, sendo assim, aceitamos hipótese alternativa e rejeitamos a hipótese nula

Ao realizarmos o teste de hipótese, descobrimos que média da renda familiar da amostra é maior que a média da renda familiar da população , ou seja, rejeitamos a hipótese alternativa em favor da hipótese nula

Ao realizar o teste de hipótese,concluímos que média de pessoas da amostra que moram na casa é menor que a da população com 95% de confiança, ou seja, rejeitamos a hipótese nula em favor da hipótese alternativa .

O testes de proporção para amostra da cidade de recife nos mostra que, a proporção de pessoas que é de Recife não é igual a 0.4, ou seja, devemos rejeitar a hipótese alternativa e aceitamos a hipótese nula

O teste de proporção nos mostra que não há diferença entre a amostra e a população em relação às pessoas que moram em Recife com 95% de confiança

O teste de proporção nos mostra com 99% de confiança, que não há diferença entre as pessoas que moram em Vitória na amostra com as que mora em Vitória na população

Para a cada variável quantitativa foi feito o gráfico de pontos, conforme ilustra as figuras abaixo:

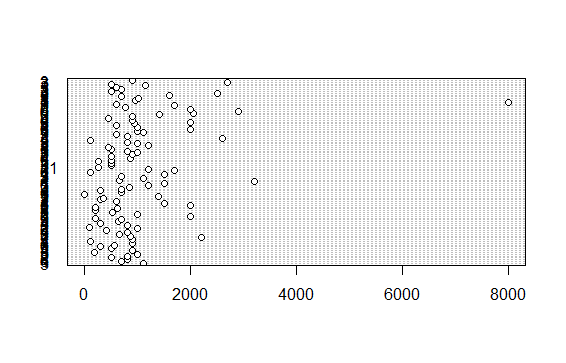


Gráfico de pontos entre Renda familiar e casa na população

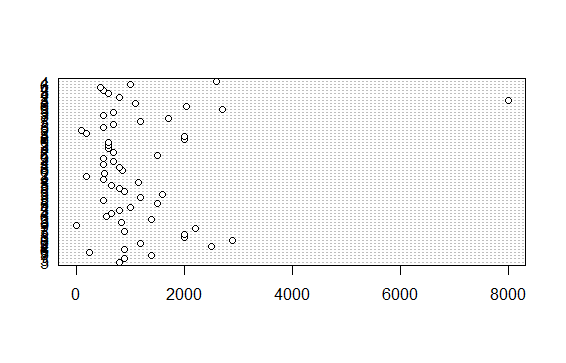
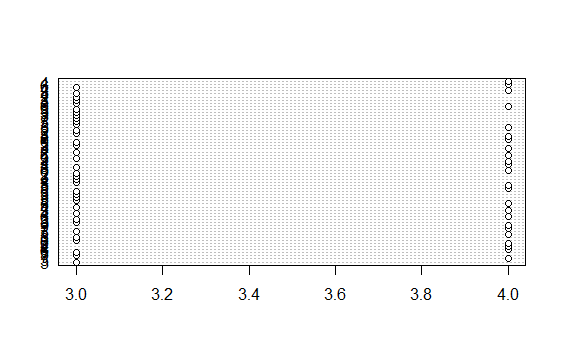


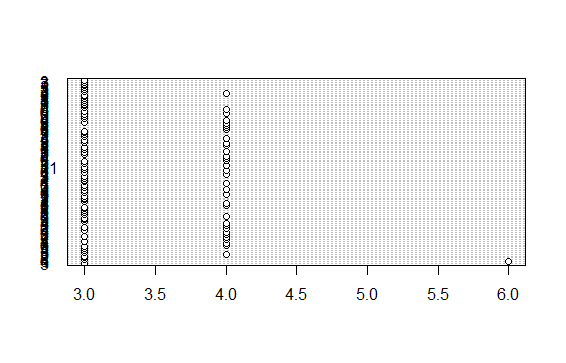
Gráfico de pontos entre Renda familiar e casa na amostra

Concluímos que a correlação é fraca entre renda familiar e casa tanto na amostra quanto na população

Concluímos que ao fazer o teste de correlação entre renda familiar e casa, notamos que a correlação não é igual a zero,ou seja, rejeitamos a hipótese nula em favor da alternativa com 95% de confiança



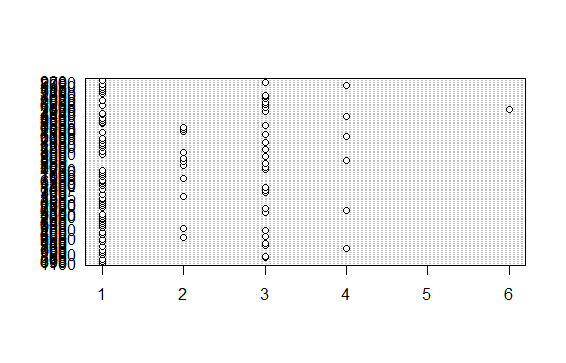
A figura acima, mostra o Gráfico de Ponto de cômodos e casa na amostra



A figura acima, mostra o Gráfico de Ponto de cômodos e casa na população

Concluímos que há uma correlação moderada entre cômodo na amostra, já na população, temos um correlação fraca

Concluímos que a correlação entre cômodo e casa na população não é igual a zero,ou seja, rejeitamos a hipótese nula em favor da alternativa com 95% de confiança



habit alug e renda familiar população

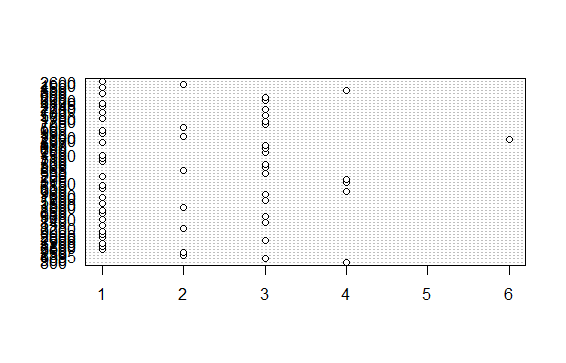


Gráfico de pontos habit alug e renda familiar na amostra

Concluímos que na amostra existe uma correlação forte entre habit alug e renda familiar, enquanto na população a correlação entre as duas variáveis é moderada

Ao realizarmos o teste de correlação entre habit alug e renda familiar, concluímos que a correlação não é igual a 0, ou seja, devemos aceitar a hipótese alternativa e rejeitar a hipótese nula

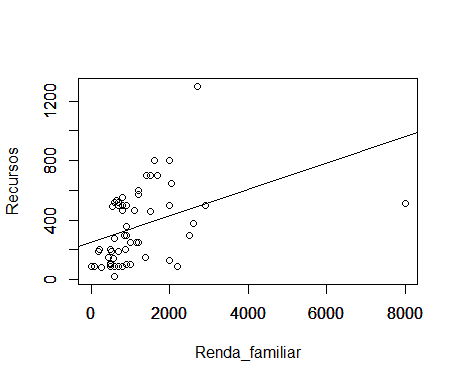


Gráfico de regressão entre recursos e renda familiar na amostra

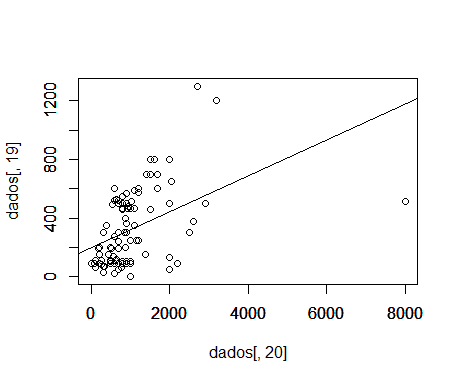
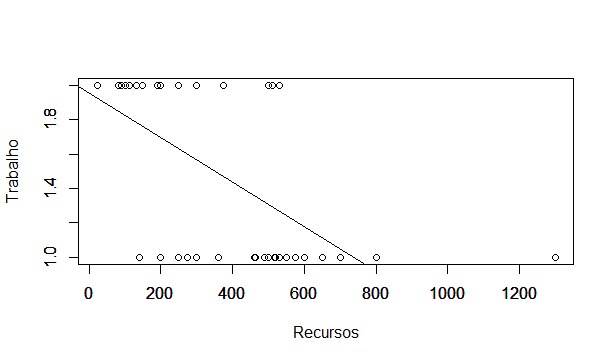


Gráfico de regressão entre recursos e renda familiar na população

Ao realizarmos o teste de regressão, descobrimos que a variável recurso influência na variável renda familiar tanto na amostra e quanto na população, ou seja, quanto mais uma pessoa ganha, maior será seu recurso ou seja, sua renda.

Os gráficos de regressão abaixo no mostras se as variáveis têm uma relação dependentes ou independentes entre elas.

 Gráfico de regressão entre trabalho e recursos na amostra

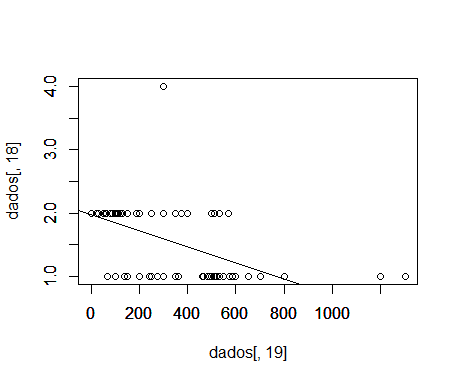


Gráfico de regressão entre trabalho e recursos na população

Pelo teste de regressão, concluímos que a variável trabalho tanto na amostra, quanto na população influencia no recurso ou seja, depende da atividade que a pessoa exerce, por exemplo, se ela exercer alguma atividade remunerada, ou seja quanto maior o salário que receber, maior o recurso que tem, isto é , quanto mais elevado o cargo de seu trabalho,maior o salário que a pessoa ganhará por mês.

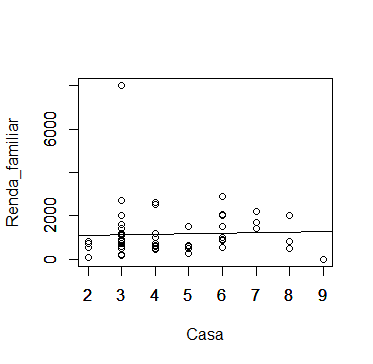
****

Gráfico de regressão entre Renda familiar e Casa na amostra

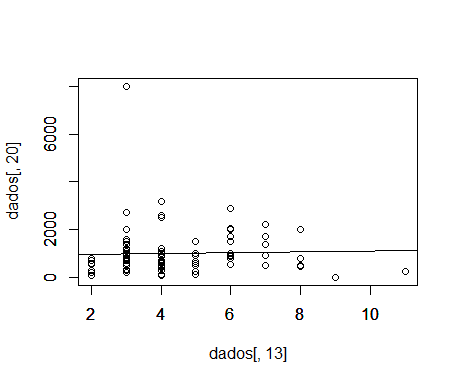


Gráfico de regressão entre Renda familiar e Casa na população

Concluímos tanto a renda familiar da amostra quanto a da população não influencia na quantidade de pessoas que moram na casa .

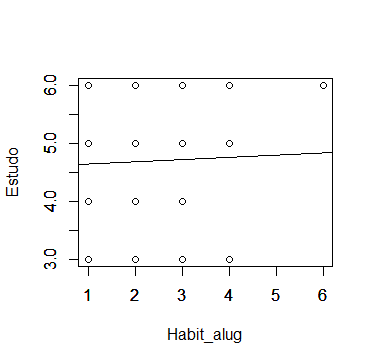


Gráfico de regressão entre estudo e habit\_alug na amostra

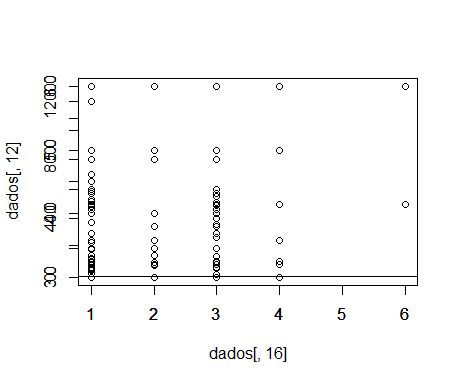


Gráfico de regressão entre estudo e habit\_alug na população

Concluímos tanto a variável Estudo da amostra quanto a da população não influencia na variável Habit\_alug

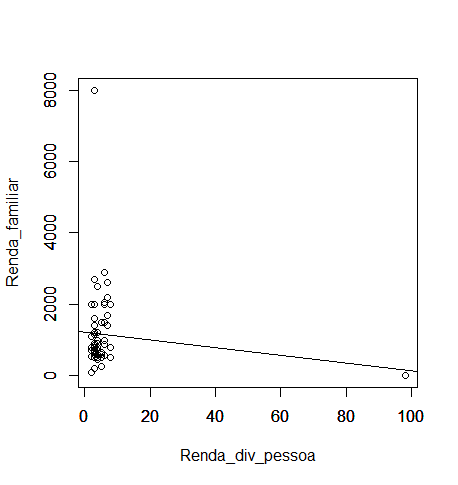


Gráfico de regressão entre Renda\_familiar e Renda\_div\_pessoa na amostra

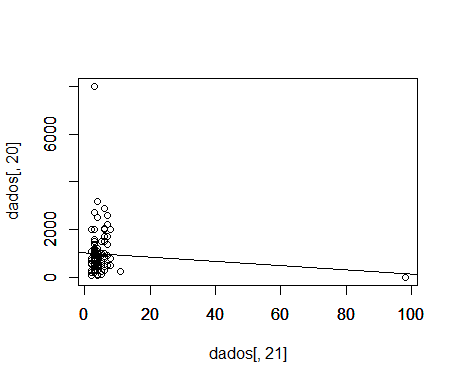


Gráfico de regressão entre Renda\_familiar e Renda\_div\_pessoa na população

Ao realizarmos o teste de regressão, descobrimos que a variável renda familiar não influencia na variável renda div pessoa na amostra e nem na população

# Linhas de código

Para os gráficos de barra foi utilizado o comando “barplot”, que diretamente retira a frequência absoluta da variável e cria o gráfico. Para as tabelas foi recolhida as frequências absolutas e relativas com os comandos “table(variável)” e “prop.table(table(variável))” respectivamente. Nos gráficos de setores é extraído a frequência relativa e com o comando “pie” cria se o gráfico referenciando a frequência.

# o código da média de Renda da família da amostra com a da população:

t.test(Renda\_da\_familia, alternative = "two.sided", mu = mean(dados[,22]), sigma.x = sd(dados[,22]))

# o código da média de Renda div pessoa da amostra com a da população:

amostraDeRendaDivPessoa = Rendadivpessoa

PopulacaoDeRendaDivPessoa = dados[,21]

t.test(amostraDeRendaDivPessoa,PopulacaoDeRendaDivPessoa)

#calculo para media de duas amostra cidade de recife da amostra com a populacao

testeMediaRecife <- t.test(x = c(59,36), n = c(102, 60))) #essa linha testeMediaRecife

# o código de hipótese da idade na amostra com a da população:

z.test(Idade, alternative = "less", mu = mean(dados[,4]), sigma.x = sd(dados[,4]))

# o código de teste de hipótese de Recursos da amostra com a da população:

z.test(Recursos, alternative = "greater", mu = mean(dados[,19]), sigma.x = sd(dados[,19]))

# o código de hipótese de Renda familiar da amostra com a da população:

z.test(Renda\_familiar, alternative = "less", mu = mean(dados[,20]), sigma.x = sd(dados[,20]))

# o código de teste de hipótese de Casa da amostra com a da população:

> z.test(Casa, alternative = "two.sided", mu = mean(dados[,13]), sigma.x = sd(dados[,13]))

# o código da proporção da amostra de Recife:

testePropCidadeRecifeAmostra = prop.test(NCidade[1],60, p = 0.4, conf.level = 0.99, correct = TRUE)

testePropCidadeRecifeAmostra

# o código de proporção entre duas amostras para população que mora em Recife:

testePropCidadeRecife <- prop.test(x = c(59,36), n = c(102, 60))

testePropCidadeRecife

# o código de proporção entre duas amostras para população que mora em Vitória:

testePropCidadeVitoria <- prop.test(x = c(21, 13), n = c(102,60), conf.level = 0.99)

testePropCidadeVitoria

//correlação

cor.test(Renda\_familiar, Casa) #faz o teste de correlação entre renda familiar e casa

cor(dados[,20], dados[,13]) #verifica a correlação entre renda familiar e casa na populacao

dotchart(Renda\_familiar,Casa) #gráfico de pontos de cômodos e casa na amostra

dotchart((dados[,20], dados[,13]) #gráfico de pontos de cômodos e casa na população

cor(dados[,20], dados[,13]) #verifica a correlação entre renda familiar e casa na população

cor.test(Renda\_familiar, Casa) #faz o teste de correlação entre renda familiar e casa

> cor.test(dados[,20], dados[,13]) # teste de correlação entre Renda familiar e casa da população

cor(Comodos,Casa) #correlação entre cômodos e casa na amostra:

cor(dados[,17] , dados[,13]) #correlação da população entre cômodos e casa

> cor.test(Comodos,Casa) #teste de correlação entre cômodos e casa na amostra

dotchart(dados[,17], dados[,13]) #gráfico de pontos de cômodos e casa na população

dotchart(Comodos,Casa) #gráfico de pontos de cômodos e casa amostra

cor(dados[,16] , dados[,20]) #correlação da população entre habit alug e renda familiar

> cor(Habit\_alug,Renda\_familiar) #verifica se a correlação entre habit alug e renda familiar

> cor.test(Habit\_alug, Renda\_familiar) #testar a correlação entre habit alug e renda familiar

dotchart(Habit\_alug,Renda\_familiar) #gráfico de pontos entre habit alug e renda familiar

#regressão

reg\_amos = lm( Recursos ~ Renda\_familiar)

summary(reg\_amos) #regressão entre recursos e renda familiar na amostra

reg\_pop2 = lm(dados[,19] ~ dados[,20]) #regressão entre recursos e renda familiar na população

summary(reg\_pop2)

reg\_pop2 = lm(dados[,18] ~ dados[,19]) #regressão trabalho e recurso na população

summary(reg\_pop2)

reg\_amos = lm( Trabalho ~ Recursos) #regressão de trabalho e recurso na amostra

summary(reg\_amos)

reg\_amos = lm( Renda\_familiar ~ Casa ) #regressão de renda familiar e casa

summary(reg\_amos)



reg\_amos = lm(dados[,20] ~ dados[,13]) //renda familiar e casa na população

summary(reg\_amos)

reg\_amos = lm( Estudo ~ Habit\_alug) #regressão de estudo e habit alug na amostra

summary(reg\_amos)

#regressão de estudo e habit alug na população:

reg\_amos = lm(dados[,12] ~ dados[,16])

summary(reg\_amos)

#regressão de Renda\_da\_familia e Renda\_div\_pessoa na amostra

reg\_amos = lm(Renda\_da\_familia ~ Renda\_div\_pessoa)

summary(reg\_amos)

#regressão de Renda\_da\_familia e Renda\_div\_pessoa na populacao

reg\_pop = lm(dados[,22] ~ dados[,21])

summary(reg\_pop)

#grafico de recursos e renda familiar na amostra

plot(Recursos ~ Renda\_familiar)

par(new=TRUE)

plot(Recursos ~ Renda\_familiar, xlab="",ylab="")

par(new=TRUE)

plot(Recursos ~ Renda\_familiar, xlab="",ylab="")

abline(lm(Recursos ~ Renda\_familiar))

#grafico de recursos e renda familiar na população

plot(dados[,19] ~ dados[,20] )

par(new=TRUE)

plot(dados[,19] ~ dados[,20] , xlab="",ylab="")

par(new=TRUE)

plot(dados[,19] ~ dados[,20], xlab="",ylab="")

abline(lm(dados[,19] ~ dados[,20]))

#grafico de trabalho e recursos na amostra

plot(Trabalho ~ Recursos)

par(new=TRUE)

plot(Trabalho ~ Recursos, xlab="",ylab="")

par(new=TRUE)

plot(Trabalho ~ Recursos, xlab="",ylab="")

abline(lm(Trabalho ~ Recursos))

#grafico de trabalho e recursos na população

plot(dados[,18] ~ dados[,19])

par(new=TRUE)

plot(dados[,18] ~ dados[,19], xlab="",ylab="")

par(new=TRUE)

plot(dados[,18] ~ dados[,19], xlab="",ylab="")

abline(lm(dados[,18] ~ dados[,19]))

#grafico de renda familiar e casa na amostra

plot(Renda\_familiar ~ Casa)

par(new=TRUE)

plot(Renda\_familiar ~ Casa, xlab="",ylab="")

par(new=TRUE)

plot(Renda\_familiar ~ Casa, xlab="",ylab="")

abline(lm(Renda\_familiar ~ Casa))

#grafico de renda familiar e casa na população

plot( dados[,20] ~ dados[,13] )

par(new=TRUE)

plot( dados[,20] ~ dados[,13] , xlab="",ylab="")

par(new=TRUE)

plot( dados[,20] ~ dados[,13] , xlab="",ylab="")

abline(lm( dados[,20] ~ dados[,13] ))

#grafico de estudo e habit alug na amostra

plot(Estudo ~ Habit\_alug)

par(new=TRUE)

plot(Estudo ~ Habit\_alug, xlab="",ylab="")

par(new=TRUE)

plot(Estudo ~ Habit\_alug, xlab="",ylab="")

abline(lm(Estudo ~ Habit\_alug))

#grafico de estudo e habit alug na população

plot(dados[,12] ~ dados[,16] )

par(new=TRUE)

plot(dados[,12] ~ dados[,16] , xlab="",ylab="")

par(new=TRUE)

plot(dados[,19] ~ dados[,16], xlab="",ylab="")

abline(lm(dados[,12] ~ dados[,16]))

#grafico de renda familiar e renda div pessoa na amostra

plot(Renda\_familiar ~ Renda\_div\_pessoa)

par(new=TRUE)

plot(Renda\_familiar ~ Renda\_div\_pessoa, xlab="",ylab="")

par(new=TRUE)

plot(Renda\_familiar ~ Renda\_div\_pessoa, xlab="",ylab="")

abline(lm(Renda\_familiar ~ Renda\_div\_pessoa))

#grafico de renda familiar e renda div pessoa na população

plot( dados[,20] ~ dados[,21] )

par(new=TRUE)

plot( dados[,20] ~ dados[,21] , xlab="",ylab="")

par(new=TRUE)

plot( dados[,20] ~ dados[,21] , xlab="",ylab="")

abline(lm( dados[,20] ~ dados[,21] ))

# Conclusão

O RStudio é essencial para se trabalhar com dados estatísticos, visto que o mesmo possui uma série de recursos interessantes que possibilitam realizar uma análise mais completa e detalhada por meio de gráficos, o que nos ajuda a tomar decisões mais eficiente.

Pelos resultados adquiridos e analisados concluímos que na nossa amostra predominam pessoas que nasceram e residem no Recife, moram no mesmo bairro onde foram aplicadas as entrevistas, possuem suas casas próprias, são de classe média, solteiras, pardas, possuem o catolicismo como religião praticante, metade delas exercem trabalho remunerado ou recebem benefícios, não estudam e possuem ensino médio incompleto.