

**PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS**  
**NÚCLEO DE EDUCAÇÃO A DISTÂNCIA**  
**Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data**

**Hermes Ribeiro da Mota Junior**

**APRENDIZADO DE MÁQUINA PARA CLASSIFICAR E PREVER A RELAÇÃO  
DOS EFEITO DA EDUCAÇÃO SUPERIOR SOBRE O INDICE DE  
DESENVOLVIMENTO HUMANO DOS MUNICIPIOS BRASILEIROS**

Belo Horizonte  
2023

**Hermes Ribeiro da Mota Junior**

**APRENDIZADO DE MÁQUINA PARA CLASSIFICAR A RELAÇÃO E PREVISÃO  
SOBRE O EFEITO DA EDUCAÇÃO SUPERIOR SOBRE O ÍNDICE DE  
DESENVOLVIMENTO HUMANO DOS MUNICÍPIOS BRASILEIROS**

Trabalho de Conclusão de Curso apresentado  
ao Curso de Especialização em Ciência de  
Dados e Big Data como requisito parcial à  
obtenção do título de especialista.

Belo Horizonte

2023

## SUMÁRIO

1	INTRODUÇÃO.....	7
1.1	Contextualização .....	8
1.2	O problema proposto .....	12
2	COLETA DE DADOS.....	14
3	PROCESSAMENTO DOS DADOS .....	16
4	ANALISE E EXPLORAÇÃO DOS DADOS .....	20
4.1	Dados Estatísticos das variáveis IDHM e IDHM educação dos municípios brasileiros da década dos anos 1990.....	21
4.2	Dados Estatísticos das variáveis IDHM e IDHM Educação dos municípios brasileiros da década dos anos 2000.....	28
4.3	Dados Estatísticos das variáveis IDHM e IDHM Educação dos municípios brasileiros da década dos anos 2010.....	35
4.4	Municípios Seleccionados e Amostras Aleatórias .....	41
4.5	Criação das amostras testes dos municípios escolhidos aleatoriamente .....	51
5	CRIAÇÃO DO MODELO DE MACHINE LEARNING .....	56
5.1	Modelo Regressão Linear para os Municípios Brasileiros .....	56
5.2	Modelo Regressão Linear para os Municípios Seleccionados.....	60
5.3	Modelo Regressão Linear para a Amostra 1.....	64
5.4	Modelo Regressão Linear para a Amostra 2.....	66
6	INTERPRETAÇÃO DOS RESULTADOS .....	69
6.1	Shapiro-Wilk-Test .....	69
6.2	Normalidade dos outliers residuais .....	70
6.3	Teste de homocedasticidade .....	70
6.4	Teste Durbin-Watson .....	71
7	APRESENTAÇÃO DOS RESULTADOS .....	74
7.1	QUESTIONAMENTO SOBRE O RESULTADO OBTIDO .....	74
	LINKS.....	76

## FIGURAS

Figura 1 – Percentual de populações com ensino superior.....	9
Figura 2 – Arquivo shapefile adquirido na página do IBGE .....	15
Figura 3 – Unificação dos dataset`s IDHM e Coordenadas .....	16
Figura 4 – Dataset inserido no R-Studio .....	16
Figura 5 – Dataset municípios selecionados e seus respectivos IES.....	17
Figura 6 – Planilha com todos os dados inserido no R-studio para análise estatística .....	18
Figura 7 – Head dos dados no R- Studio .....	19
Figura 8 – Pacotes instalados no R-Studio para análise estatística dos dados .....	20
Figura 9 e 10 – Histograma dos dados IDHM e IDHM Educação década anos 90 ...	22
Figura 11 e 12 – Boxplot dos dados IDHM e IDHM Educação década anos 90 .....	22
Figura 13 – Boxplot nuvem dos dados IDHM Educação década anos 90.....	23
Figura 14 - Mapa de cor do IDHM década dos anos 1990 dos municípios brasileiros com e legenda do IDH definido pela ONU .....	24
Figura 15 - Mapa de cor do IDHM Educação década dos anos 1990 dos municípios brasileiros com e legenda do IDH definido pela ONU .....	25
Figura 16 – Gráfico Correlação IDHM x IDHM Educação anos 90.....	26
Figura 16 e 17 – Histograma dados IDHM, IDHM Educação anos 2000 .....	29
Figura 18 e 19 – Boxplot dados IDHM, IDHM Educação anos 2000 .....	29
Figura 20 e 21 - Boxplot nuvem dos dados IDHM, IDHM Educação anos 2000 .....	30
Figura 22 – Correlação IDHM x IDHM EDUCAÇÃO anos 2000.....	31
Figura 23 - Mapa de cor do IDHM Educação década dos anos 2000 dos municípios brasileiros com e legenda do IDH definido pela ONU .....	32
Figura 24 - Mapa de cor do IDHM década dos anos 2000 dos municípios brasileiros com e legenda do IDH definido pela ONU .....	33
Figura 24 e 25 – Histograma dos dados municípios brasileiros década anos 2010..	35
Figura 26 e 27 – Boxplot dados IDHM e IDHM Educação década anos 2010 .....	36
Figura 28 e 29 – Boxplot nuvem dados IDHM e IDHM Educação década anos 2010 .....	36
Figura 30 - Mapa de cor do IDHM década dos anos 2010 dos municípios brasileiros com e legenda do IDH definido pela ONU .....	37

Figura 31 - Mapa de cor do IDHM Educação década dos anos 2010 dos municípios brasileiros com e legenda do IDH definido pela ONU .....	38
Figura 33 – Gráfico Correlação IDHM x IDHM Educação década 2010.....	39
Figura 34 – Dados dos municípios brasileiros selecionados pelo IES descentralizados no R-Studio.....	42
Figura 35 – Histograma dos dados dos municípios selecionados década anos 90 ..	43
Figura 36 – Correlação IDHM x IDHM Educação municípios selecionados década anos 90 .....	44
Figura 38 – Histograma IDHM Educação municípios selecionados década anos 2000 .....	45
Figura 39 – Gráfico Correlação IDHM x IDHM Educação municípios selecionados anos 2000.....	46
Figura 40 e 41 - Histograma IDHM e IDHM Educação municípios selecionados década anos 2010.....	47
Figura 42 – Correlação IDHM x IDHM Educação década anos 2010 .....	48
Figura 43 - Mapa de Cores Municípios Brasileiros que receberam os IES e os Municípios Adjacentes .....	50
Figura 44 – Gráfico correlação IDHM x IDHM Educação Amostra 1 dos municípios selecionados aleatoriamente.....	53
Figura 45 – Gráfico correlação IDHM x IDHM Educação Amostra 2 dos municípios selecionados aleatoriamente.....	54
Figura 46 – Gráficos estatísticos dos pressupostos validação do modelo de aprendizado de máquina municípios brasileiros.....	58
Figura 47 – Gráfico de dispersão dos municípios brasileiros .....	60
Figura 48 - Gráficos estatísticos dos pressupostos validação do modelo de aprendizado de máquina municípios selecionados .....	61
Figura 49 - Gráfico de dispersão dos municípios selecionados .....	63
Figura 50 - Gráficos estatísticos dos pressupostos validação do modelo de aprendizado de máquina Amostra 1.....	64
Figura 51 – Gráfico de dispersão Amostra 1 .....	66
Figura 52 - Gráficos estatísticos dos pressupostos validação do modelo de aprendizado de máquina Amostra 2.....	67
Figura 53 – Gráfico de dispersão Amostra 2.....	68
Figura 54 – Worflow deste trabalho de Ciência dos Dados.....	75

## TABELAS

Tabela 1 - Campo coluna Dataset localização das universidades .....	14
Tabela 2 - Campo coluna dataset IDHM municípios brasileiros .....	14
Tabela 3 - Dados estatísticos apurados de todos os municípios brasileiros. ....	40
Tabela 4 – Dados estatísticos apurados dos Municípios selecionados .....	49
Tabela 5 – Dados estatísticos apurados das amostra 1 e 2.....	55
Tabela 6 – Pacotes baixados no R-Studio para criação do modelo de machine learning .....	57
Tabela 7 – Dados apurados no Teste Shapiro-Wilk .....	69
Tabela 8 – Dados apurados Normalidade dos Outliers Residuais .....	70
Tabela 9 – Dados apurados no teste de homocedasticidade.....	71
Tabela 10 – Dados apurados no teste de Durbin-Watson.....	71
Tabela 12 – Teste R- Squared .....	73
Tabela 13 – Teste de hipótese F- Statistic .....	73

## 1 INTRODUÇÃO

O índice de desenvolvimento humano (IDH), um indicador mundialmente conhecido que tem como meta traçar um espectro da forma como a sociedade de um determinado país se encontra em relação a três indicadores sociais a saber, educação, saúde e renda. Dentre estes três indicadores sociais, a educação é vista como o verdadeiro pilar transformador social desempenhando um papel central que renova a vida em sociedade no que diz respeito aos rumos que uma determinada nação almeja atingir conduzindo suas ações no sentido de permitir que prolifere um ambiente de reflexão sobre si mesma norteando suas decisões a respeito do futuro.

A educação por ser um tema bastante amplo e complexo ainda mais se considerarmos a extensão do território brasileiro e toda as suas peculiaridades culturais, regionais e sociais, este trabalho foca em uma pequena fração deste universo que é o ensino superior público que a partir de um determinado momento da nossa história, por meio de políticas públicas amparado pelo MEC, INEP e outros órgãos da educação pública, decidiu expandir as ações das universidades públicas das capitais para regiões do interior do Brasil visando o desenvolvimento social dos municípios brasileiros em determinadas regiões do território nacional

## 1.1 Contextualização

Em 2020 a ONU divulgou através do Programa das Nações Unidas para o Desenvolvimento (PNUD) a relação dos países com maior IDH entre as nações do Mundo e a posição entre os 12 primeiro ficou da seguinte forma respectivamente : Noruega – 0,957, Suíça e Irlanda – 0,955 (empatados), Hong Kong (China) e Islândia 0,949 (empatados), Alemanha – 0,947, Suécia – 0,945, Austrália e Holanda – 0,944 (empatados), Dinamarca -0,940, Singapura e Finlândia – 0,938 (empatados), Finlândia – 0,938, Nova Zelândia e Bélgica – 0,931 (empatados), Canadá – 0,929, Estados Unidos – 0,926. O IDH do Brasil cresceu de 0,762 para 0,765, mas caiu cinco posições no ranking em relação ao ano anterior, ficando em 84º lugar entre 189 países avaliados.

O IDH brasileiro é medido todos os anos pelo IPEA e divulgado nacional e globalmente pelo Programa das Nações Unidas para o Desenvolvimento (PNUD), órgão esse que faz parte da Organização das Nações Unidas. O atlas do desenvolvimento humano, que é o relatório final da coleta de dados para o estudo o IDH nacional, traça um perfil da evolução ou regressão do IDH nas macrorregiões do Brasil.

[...] A versatilidade do Atlas nos permite pensar desde o micro – vendo a realidade a nível de Unidades de Desenvolvimento Humano (UDH) – até o macro – pensando o país, as unidades da federação e agora, as macrorregiões (Atlas, PNUD, 2016, p. 4).

De acordo com o PNUD, o relatório emitido pela entidade diz que:

[...] O Relatório de Desenvolvimento Humano (RDH) é reconhecido pelas Nações Unidas como um exercício intelectual independente e uma importante ferramenta para aumentar a conscientização sobre o desenvolvimento humano em todo o mundo. Com sua riqueza de dados e abordagem inovadora para medir o desenvolvimento, o RDH tem um grande impacto nas reflexões sobre o tema no mundo todo (Atlas, PNUD, 2016, p. 4).

O Censo da Educação Superior do ano de 2020 divulgado pelo INEP constatou que no Brasil existia 2457 instituições de ensino superior espalhadas pelo território brasileiro sendo que 87,6% deste montante são de instituições privadas e 12,4%



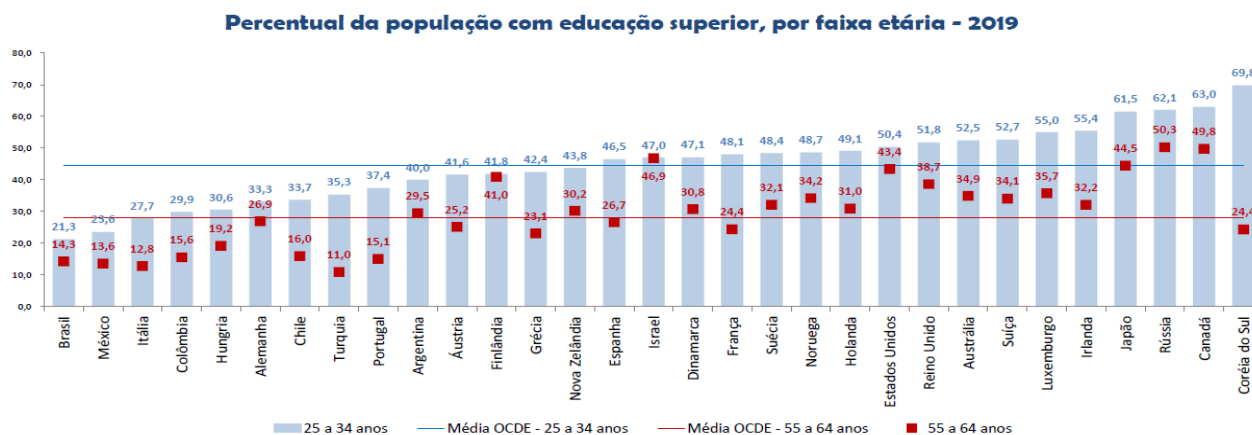
instituições de ensino superior da rede pública onde mais de 33 mil cursos de graduação são oferecidos desses 66% são dos cursos de bacharelado, 19,7% Licenciatura e 14,3% Tecnologia.

Pelo fato da grande maioria dos IES privados visar a lucratividade ao tempo que presta serviços de educação de nível superior, este trabalho optou por excluir esses IES de caráter privado, mesmo tendo grande participação percentual no contexto nacional brasileiro, sendo assim, as mesmas não formam consideradas para análise de suas contribuições nos IDH's dos municípios. Outro ponto, com relação a esse tema, é a tendência que esses IES privados tenderem a se estabelecer em grandes centros e cidades com maior número de pessoas onde na maioria desses municípios, já existe grandes centros universitários.

Muitos são os desafios na esfera educacional no país, o INEP por exemplo tem como metas:

Elevar a escolaridade média da população de 18 a 29 anos, de modo a alcançar, no mínimo, 12 anos de estudo no último ano de vigência deste Plano, para as populações do campo, da região de menor escolaridade no País e dos 25% mais pobres, e igualar a escolaridade média entre negros e não negros declarados à Fundação Instituto Brasileiro de Geografia e Estatística (IBGE.)  
Elevar a taxa bruta de matrícula na educação superior para 50% e a taxa líquida para 33% da população de 18 a 24 anos, assegurando a qualidade da oferta e expansão para, pelo menos, 40% (quarenta por cento) das novas matrículas, no segmento público. (Censo Educação, INEP, 202 p. 2).

A figura abaixo, extraída do censo da educação superior INEP, apresenta o percentual da população com educação superior em relação a alguns países relacionando os com o contexto internacional.



*Figura 1 – Percentual de populações com ensino superior*

Fonte: INEP – Censo da Educação Superior extraídos em 25 de maio de 2022

A educação é encarada por muitos países como um vetor de mudanças no que diz respeito a desenvolvimento humano. O presidente Lula em seu discurso quando foi descoberto o Pré sal no Brasil disse: “Segundo ele, o maior desafio do país hoje é a educação. O presidente quer inclusive que parte da receita obtida com o petróleo do Pré-sal seja destinada para investimento na área. “Não basta ensinar a ler e escrever é preciso muito mais. É um investimento no futuro dos nossos filhos e netos”, disse. [...] Então, a primeira definição que nós tivemos era que em Educação neste país não utilizaríamos no orçamento mais a palavra gasto, iríamos utilizar a palavra investimento. Estava proibido qualquer ministro utilizar a palavra gasto com Educação.

Um dos pioneiros nas reflexões sobre as universidades, Friedrich von Humboldt, descreveu em sua obra o texto *Sobre a Organização Interna e Externa das Instituições Científicas Superiores em Berlim*, a argumentação de que parte do pressuposto de que às Instituições Científicas cabe a responsabilidade pelo "enriquecimento da cultura moral da Nação." Afirma ainda que a organização interna destas instituições é caracterizada "pela combinação de ciência objetiva e formação subjetiva." A organização externa teria uma finalidade pragmática, ao preparar para a saída da escola e o ingresso na Universidade. Apresenta também uma concepção de ciência pura, que assim deve permanecer para não ser deturpada pelas demandas sociais

*Friedrich von Humboldt* (1767-1835) acreditava que a base e fim de qualquer sistema educacional era a formação de cidadãos, e defendia que somente a autonomia permitiria às universidades atingirem seus mais altos propósitos (Rohe, 2017). *Vannevar Bush* (1890-1974), no texto seminal ‘Science, *The Endless Frontier*’ (Bush, 1945) argumentou que o progresso científico é resultado da interação livre de intelectos livres: cientistas trabalhando em temas de sua escolha, definidos a partir de sua curiosidade para explorar o desconhecido e avançar as fronteiras da ciência.

Os países conhecidos como os Tigres Asiáticos são reconhecidos internacionalmente pelo rápido desenvolvimento econômico, marcado pela aplicação maciça de capital em áreas como a educação e a profissionalização. Eles conseguiram, por meio do investimento externo e da produção voltada para a exportação, consolidar a sua economia. Sendo assim, desenvolveram um consistente setor industrial, em conjunto com melhorias sociais, como o aumento da escolaridade

e da renda. Em 1990 quando o IDH passou a ser divulgado, Coreia do Sul, Singapura e Hong Kong apresentavam índices de 0,732, 0,721 e 0,784 ocupando as posições 36º, 42º e 19º respectivamente, em 2019 o PNUD divulgou os índices de IDH desses mesmos países: 0,916, 0,938 e 0,949 ocupando as respectivas posições no ranking 23º, 11º e 4º, este é um claro sinal do poder do investimento em educação que causam ganhos no desenvolvimento social consideráveis.

## 1.2 O problema proposto

O propósito deste trabalho consiste em lançar um olhar para o quesito educação delimitado aos municípios brasileiros que receberam os Institutos superiores públicos descentralizados e inferir se este fato contribui para mais, nada ou menos para o desenvolvimento do IDHM destes municípios em comparação com os demais municípios do Brasil e duas amostras de municípios colhida aleatoriamente entre os mais de 5500 municípios brasileiros.

Pretende-se mapear as unidades de ensino superior nas cinco regiões brasileiras (Norte, Sul, Sudeste, Nordeste e Centro-Oeste) e inferir sobre qual a relação entre o fato de termos unidades de ensino superior descentralizado e o desenvolvimento em frações regionais (municípios selecionados e seus adjacentes) dentre as cinco regiões.

Partindo do pressuposto que os países com os maiores IDH tem entre sua população altos percentuais de pessoas com nível de escolaridade superior, será avaliado os *Dataset* referentes a distribuição das unidades de ensino superior espalhadas pelo país e confrontado com o *Dataset* do IDHM educacional ao longo das décadas dos anos 90, 2000 e 2010 e deste modo efetuar análise dos dados estatísticos e correlaciona-los com o propósito investigativo e assim aplicar modelos de *Machine Learning*, que possibilite inferir o quanto a disseminação do ensino superior no Brasil contribui para o aumento no Índice de Desenvolvimento Humano (IDH) sobre as regiões do Brasil e destacar a ocorrência de fatos relevantes refletindo desta forma o quadro de desenvolvimento social brasileiro.

Outro aspecto importante é que em anos recentes, os (Institutos de Ensino Superior) IES estão sendo reavaliadas sob uma perspectiva territorial por profissionais acadêmicos e gestores públicos, suscitando estudos e projetos de pesquisas internacionais, com a finalidade de examinar mais detalhadamente a influência e o impacto do sistema de ensino superior para o desenvolvimento regional. O objetivo, de modo geral, tem sido o de estabelecer um quadro teórico e empírico mais compreensivo, visando subsidiar a formulação de políticas públicas endereçadas, principalmente, a mobilizar os IES (Institutos de Ensino Superior) a favor das áreas

geográficas nas quais estão sediadas e, assim, contribuir para fazer face às desigualdades econômicas regionais

Para melhor entendimento do problema proposto foi utilizada a técnica recomendada do 5W`s, técnica esta que consiste em perguntas direcionadas de modo que todas as informações necessárias para compreensão do problema proposto sejam definidas de modo claro visando a melhor compreensão por parte dos leitores. Por fim será respondido o teste de hipótese abaixo:

Ho: O ensino superior descentralizado influencia no aumento do IDHM Educação, variável independente, que por sua vez influencia no aumento do IDHM, variável dependente, mais rapidamente que os municípios que não possuem polos de IES descentralizados

H1: O ensino superior descentralizado tem pouca ou nenhuma influência sobre o desenvolvimento IDHM Educação que por sua vez não influencia no IDHM.

(Why) Por que esse problema é importante? Porque a educação superior é encarada com a fonte de mudanças nas sociedades pauta de políticas públicas que visam usa lá com a maior eficiência possível na busca por melhorias no incremento dos níveis de desenvolvimento social sendo, os demais indicadores de desenvolvimento do IDH, compreendidos como derivados das questões educacionais.

(Who?) De quem são os dados analisados? Os dados analisados provêm das instituições governamentais brasileiras como INEP, IBGE e IPEA e instituição internacional, Organização das Nações Unidas (ONU) como o PNUD.

(What?): Quais os objetivos com essa análise? Através dos dados coletados pretende se inferir como a educação superior pública, pode interferir nos níveis de desenvolvimento humano nas microrregiões beneficiadas pelo recebimento de polos descentralizados do IES públicos comparada com as demais macros e microrregião do Brasil que não receberam os mesmos benefícios.

(Where?): A pesquisa abrange todos os municípios do território nacional, com foco nas cidades e microrregiões brasileiras de cada região do país.

(When?): Qual o período está sendo analisado? O período a ser analisado corresponde as décadas dos anos 90, 2000 e 2010.

## 2 COLETA DE DADOS

Os *Dataset* sobre a localização das universidades no território brasileiro foram adquiridos nas páginas do governo federal nos dados abertos no endereço do INEP - <https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/microdados/censo-da-educacao-superior>, arquivo *microdados\_censo\_da\_educacao\_superior\_2020* zip *MICRODADOS\_CADASTRO\_CURSOS\_2020.xlsx*, e os *Dataset* do IDHM dos municípios brasileiros, dados montados a partir da tabela disponibilizada na página do PNUD <https://www.br.undp.org/content/brazil/pt/home/idh0/rankings/idhm-municipios-2010.htm>! devidamente configurado em uma planilha de Excel para melhor atender a proposta do trabalho sem perda na aquisição dos dados brutos conforme disponibilizado na referida página acima.

NU_ANO_C ENSO	NO_RE GIÃO	CO_RE GIÃO	NO_ UF	SG_ UF	CO_ UF	NO_MUNI CÍPIO	CO_MUNICÍ PIO
------------------	---------------	---------------	-----------	-----------	-----------	------------------	------------------

CO _IE S	NO_CINE_ ROTULO	CO_CINE_ _ROTULO	CO_CINE_A REA_GERAL	NO_CINE_A REA_GERA L	CO_CINE_ARE A_ESPECIFIC A
----------------	--------------------	---------------------	------------------------	----------------------------	---------------------------------

Tabela 1 - Campo coluna Dataset localização das universidades

IDHM 2010 Ranking	Municípios	IDHM 2010	IDHM Educação
-------------------	------------	-----------	---------------

Tabela 2 - Campo coluna dataset IDHM municípios brasileiros

OS *Dataset*'s relacionados a construção de mapas para agregar informações a este trabalho de conclusão de curso, digo, dados e Polígono das coordenadas dos municípios brasileiros foi buscado no endereço do IBGE abaixo:

<https://www.ibge.gov.br/geociencias/organizacao-do-territorio/malhas-territoriais/15774-malhas.html?=&t=acesso-ao-produto>.

Estes arquivos em shape file é possível construir o mapa do Brasil dividido por regiões, Estados e municípios, para este trabalho foi usado o mapa do Brasil dividido em municípios para melhor confrontar as informações vista de forma mais pontual em relação ao território brasileiro. A ideia e poder juntar as informações dos *Dataset*'s

ranking IDH dos municípios brasileiros e condensar as informações de forma visual tanto da disposição dos IES e dos seus respectivos municípios e adjacências.



*Figura 2 – Arquivo shapefile adquirido na página do IBGE*

### 3 PROCESSAMENTO DOS DADOS

De modo a se criar um Dataset contendo todas as informações dos IDH's municípios, código dos municípios, classificação e coordenadas e *datas* de todos municípios, foi criado uma tabela no Excel que posteriormente pudesse ser utilizado tanto para cálculos estatísticos no R e Machine Learning e também gerar mapas de cores de todos os municípios brasileiros em relação aos dados estatísticos apurados ao longo da elaboração deste trabalho. A figura abaixo representa uma amostra com os dados dos dois *Dataset*'s unificados e posteriormente foi acrescentado os IDHM's e IDHM's Educação na planilha abaixo totalizando desta forma a planilha principal para o desenvolvimento do trabalho.

Classificação	Municípios	Código Município	IDHM	IDHM Educação	Longitude	Latitude
1 º	SÃO CAETANO DO SUL (SP)	3548807	0.862	0.811	-465.715.146.086.306	-23.614.705
2 º	ÁGUAS DE SÃO PEDRO (SP)	3500600	0.854	0.825	-478.839.747.409.776	-225.973.395.538.539
3 º	FLORIANÓPOLIS (SC)	4205407	0.847	0.8	-485.476.373.781.933	-2.758.779.554.855
4 º	BALNEÁRIO CAMBORIÚ (SC)	4202008	0.845	0.789	-486.346.174.770.265	-269.918.186.052.546
4 º	VITÓRIA (ES)	3556958	0.845	0.805	-504.808.069.702.363	-201.987.385.744.561
6 º	SANTOS (SP)	3160702	0.84	0.807	-435.537.521.652.921	-214.575.319.947.014
7 º	NITERÓI (RJ)	3303302	0.837	0.773	-430.758.231.672.735	-228.964.523.851.273
8 º	JOAÇABA (SC)	4209003	0.827	0.771	-515.066.897.333.377	-271.739.445.292.328
9 º	BRASÍLIA (DF)	3108602	0.824	0.742	-444.275.330.930.915	-16.205.872.688.841
10 º	CURITIBA (PR)	4106902	0.823	0.768	-492.718.478.850.774	-25.432.956
11 º	JUNDIAÍ (SP)	4112900	0.822	0.768	-502.493.535.652.284	-23.437.909.451.052
12 º	VALINHOS (SP)	3556206	0.819	0.763	-469.966.300.275.552	-22.971.244
13 º	VINHEDO (SP)	3556701	0.817	0.739	-469.764.763.090.797	-230.305.383.241.408
14 º	ARARAQUARA (SP)	3503208	0.815	0.782	-481.744.399.375.437	-217.903.595
14 º	SANTO ANDRÉ (SP)	3547809	0.815	0.769	-465.308.742.576.295	-2.365.751
16 º	SANTANA DE PARNAÍBA (SP)	2312007	0.814	0.725	-402.094.628.598.127	-345.977.534.173.113
17 º	NOVA LIMA (MG)	3144805	0.813	0.704	-438.497.833.626.163	-199.838.929.938.666
18 º	ILHA SOLTEIRA (SP)	3520442	0.812	0.782	-51.344.890.657.635	-204.293.725
19 º	AMERICANA (SP)	3501608	0.811	0.76	-473.303.629.263.814	-227.408.835
20 º	BELO HORIZONTE (MG)	3106200	0.81	0.737	-439.264.531.735.305	-199.375.242.937.751

Figura 3 – Unificação dos dataset's IDHM e Coordenadas

Classificação	Municípios	Cod_Município	IDHM.2010	IDHM.2010.Educação	LONGITUDE	LATITUDE
1 1 º	SÃO CAETANO DO SUL (SP)	3548807	0.862	0.811	-46.57151	-23.614705
2 2 º	ÁGUAS DE SÃO PEDRO (SP)	3500600	0.854	0.825	-47.88397	-22.597340
3 3 º	FLORIANÓPOLIS (SC)	4205407	0.847	0.800	-48.54764	-27.587796
4 4 º	BALNEÁRIO CAMBORIÚ (SC)	4202008	0.845	0.789	-48.63462	-26.991819
5 4 º	VITÓRIA (ES)	3556958	0.845	0.805	-50.48081	-20.198739
6 6 º	SANTOS (SP)	3160702	0.840	0.807	-43.55375	-21.457532
7 7 º	NITERÓI (RJ)	3303302	0.837	0.773	-43.07563	-23.006453

Showing 1 to 7 of 5,565 entries, 7 total columns

Figura 4 – Dataset inserido no R-Studio



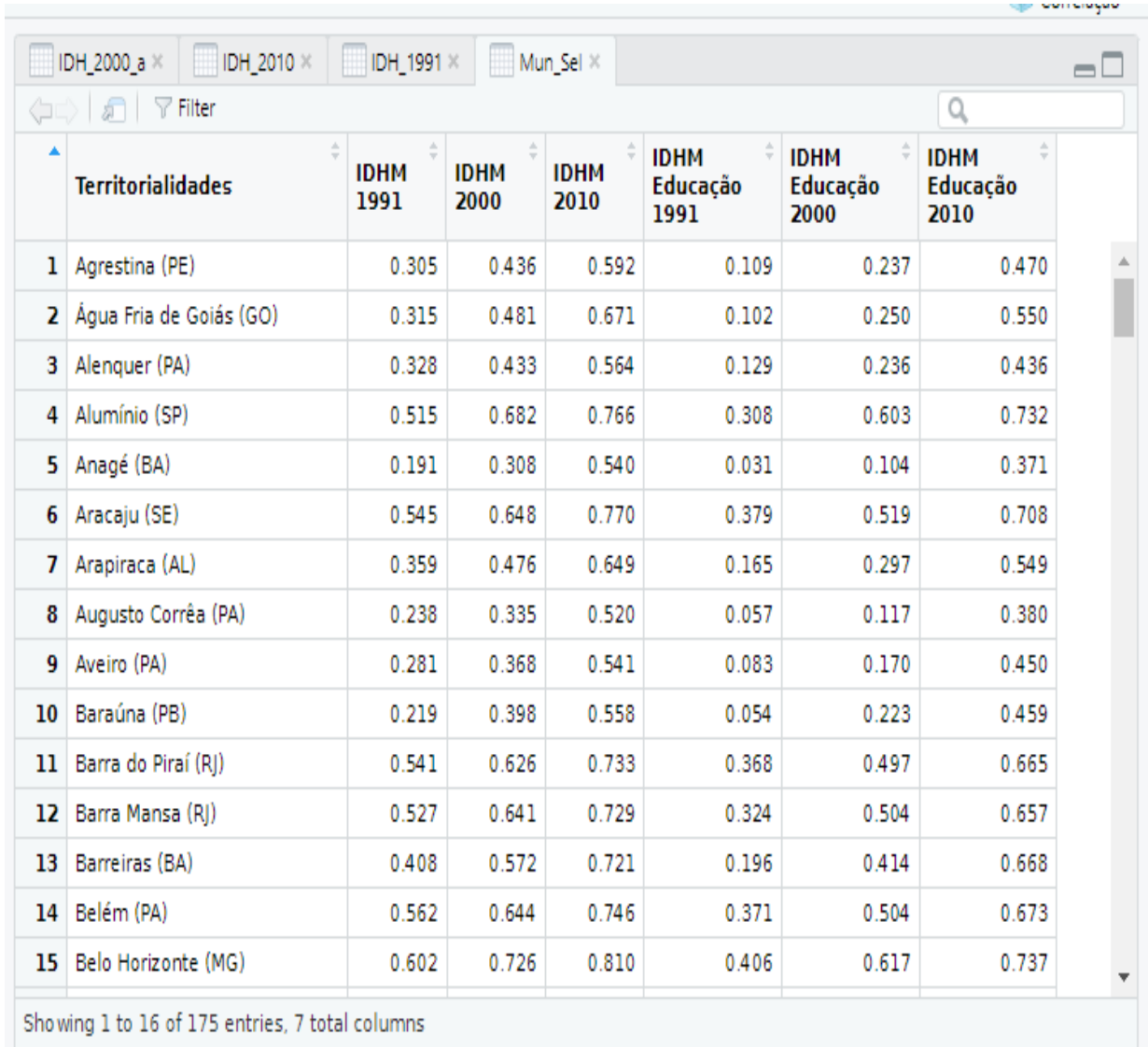
*Dataset* de atributos inserida no QGIS versão 3.32 Lima com a localização dos IES públicos dos municípios brasileiros identificado pelo código comum que cada município possui.

Universidades a partir de 1990 Planilha1 :: Feições de totais: 38, filtrado: 38, selecionado: 0

	Cidade	Estado	Cod Municipio	Universidade
1	Nova Iguaçu	RJ	3303500	Universidade Federal Rural do Rio de Janeiro - Campus de Nova Iguaçu
2	Serra Talhada	PE	2613909	Universidade Federal Rural de Pernambuco - Campus de Serra Talhada
3	Garanhuns	PE	2606002	Universidade Federal Rural de Pernambuco - Campus de Garanhuns)
4	Volta Redonda	RJ	3306305	Universidade Federal Fluminense - Campus de Volta Redonda
5	Teófilo Otoni	MG	3168606	Universidade Federal dos Vales do Jequitinhonha e Mucuri - Campus de Teóf...
6	Frederico Westp...	RS	4308508	Universidade Federal do Rio Grande do Sul - Campus de Frederico Westphal...
7	Picos	PI	2208007	Universidade Federal do Piauí - Campus de Picos
8	Parnaíba	PI	2207702	Universidade Federal do Piauí - Campus de Parnaíba
9	Bom Jesus	PI	2201903	Universidade Federal do Piauí - Campus de Bom Jesus do Gurguéia
10	Matinhos	PR	4115705	Universidade Federal do Paraná - Campus do Litoral
11	Santarém	PA	1506807	Universidade Federal do Para (Campus Santarém)
12	Marabá	PA	1504208	Universidade Federal do Para (Campus Marabá)
13	Castanhal	PA	1502400	Universidade Federal do Para (Campus Castanhal)
14	Bragança	PA	1501709	Universidade Federal do Para (Campus Bragança )
15	Imperatriz	MA	2105302	Universidade Federal do Maranhão - Campus de Imperatriz

*Figura 5 – Dataset municípios selecionados e seus respectivos IES*

*Dataset* inserido no programa R a partir da planilha de Excel elaborada previamente com todas as informações dos IDHM e IDHM educação de todos os municípios brasileiros, e deste modo facilitar a manipulação dos dados para as devidas análise estatísticas das variáveis em relação a todos os municípios para a criação do modelo de aprendizado de máquina.



The screenshot shows an RStudio spreadsheet with the following tabs: IDH\_2000\_a, IDH\_2010, IDH\_1991, and Mun\_Sel. The spreadsheet displays a table with 8 columns: Territorialidades, IDHM 1991, IDHM 2000, IDHM 2010, IDHM Educação 1991, IDHM Educação 2000, and IDHM Educação 2010. The first 15 rows of data are visible, showing municipalities from Agrestina (PE) to Belo Horizonte (MG). The status bar at the bottom indicates 'Showing 1 to 16 of 175 entries, 7 total columns'.

	Territorialidades	IDHM 1991	IDHM 2000	IDHM 2010	IDHM Educação 1991	IDHM Educação 2000	IDHM Educação 2010
1	Agrestina (PE)	0.305	0.436	0.592	0.109	0.237	0.470
2	Água Fria de Goiás (GO)	0.315	0.481	0.671	0.102	0.250	0.550
3	Alenquer (PA)	0.328	0.433	0.564	0.129	0.236	0.436
4	Alumínio (SP)	0.515	0.682	0.766	0.308	0.603	0.732
5	Anagé (BA)	0.191	0.308	0.540	0.031	0.104	0.371
6	Aracaju (SE)	0.545	0.648	0.770	0.379	0.519	0.708
7	Arapiraca (AL)	0.359	0.476	0.649	0.165	0.297	0.549
8	Augusto Corrêa (PA)	0.238	0.335	0.520	0.057	0.117	0.380
9	Aveiro (PA)	0.281	0.368	0.541	0.083	0.170	0.450
10	Baraúna (PB)	0.219	0.398	0.558	0.054	0.223	0.459
11	Barra do Piraí (RJ)	0.541	0.626	0.733	0.368	0.497	0.665
12	Barra Mansa (RJ)	0.527	0.641	0.729	0.324	0.504	0.657
13	Barreiras (BA)	0.408	0.572	0.721	0.196	0.414	0.668
14	Belém (PA)	0.562	0.644	0.746	0.371	0.504	0.673
15	Belo Horizonte (MG)	0.602	0.726	0.810	0.406	0.617	0.737

*Figura 6 – Planilha inserido no R-Studio para análise estatística todos os dados*

```
> head(ID_Mun_IDH)
  Classificação      Municípios Cod_Município IDHM.2010
1      1 ° SÃO CAETANO DO SUL (SP)      3548807      0.862
2      2 ° ÁGUAS DE SÃO PEDRO (SP)      3500600      0.854
3      3 ° FLORIANÓPOLIS (SC)      4205407      0.847
4      4 ° BALNEÁRIO CAMBORIÚ (SC)      4202008      0.845
5      4 ° VITÓRIA (ES)      3556958      0.845
6      6 ° SANTOS (SP)      3160702      0.840

  IDHM.2010.Educação LONGITUDE LATITUDE
1      0.811 -46.57151 -23.61471
2      0.825 -47.88397 -22.59734
3      0.800 -48.54764 -27.58780
4      0.789 -48.63462 -26.99182
5      0.805 -50.48081 -20.19874
6      0.807 -43.55375 -21.45753
>
```

*Figura 7 – Head dos dados no R- Studio*

Para que pudéssemos realizar a estatística dos dados nestes *Dataset's*, tivemos que considerar o *rank* e os municípios como variável numérica de modo que o programa R pudesse realizar os cálculos como devido. E assim foi desenvolvido a análise estatística e os mapas dos municípios brasileiros conforme cada desempenho no que diz respeito ao IDHM e IDHM Educação. Entre as décadas dos anos 1990, 2000 e 2010. No final do trabalho encontra se um script da forma como foi criado os mapas no software Qgis, e a planilha com os municípios que receberam os IES públicos descentralizados

#### 4 ANÁLISE E EXPLORAÇÃO DOS DADOS

Começamos as análises e exploração dos dados estatísticos dos *Dataset's* baixando os seguintes pacotes no R para auxiliar tanto na interpretação dos dados brutos em dados compilados, e também poder gerar gráficos de histogramas, *boxplot* e análise de correlação entre as variáveis. O intuito é respaldar a criação do modelo de machine learning que melhor se adequará as pretensões deste trabalho

Alguns pacotes instalados no R para análise estatística e na criação dos gráficos.

```
install.packages("rcompanion")
install.packages("funModeling")
install.packages("skimr")
install.packages("ggpubr")
install.packages("ggdist")
install.packages("ggthemes")
library(rcompanion)
library(funModeling)
library(skimr)
library(ggpubr)
library(ggdist)
library(ggthemes)
```

Figura 8 – Pacotes instalados no R-Studio para análise estatística dos dados

Como mencionado anteriormente devido a praticidade e a gama de recursos que o *software* QGIS oferece, em comparação a outros softwares, o R inclusive, para a elaboração de mapas passamos a utiliza-lo. De posse do arquivo *shapefile* buscado no IBGE e a planilha com a lista do IDH dos municípios brasileiros foi feito a anexação da coluna de interesse da planilha Excel com o arquivo shp pelo software QGIS, lembrando que as duas planilhas só tinham em comum o código do município, variável essa utilizada para fazer a correspondência entre a localização no município no mapa com seu respectivo IDHM e IDHM educação. A partir do novo *Dataset* do arquivo *shapefile* do IBGE e a planilha Excel com o IDH dos municípios brasileiros das décadas dos anos de 1990, 2000 e 2010 foi gerado o mapa de cores, de acordo o IDHM e IDHM educação possibilitando ao leitor visualizar de forma mais rápida a situação do desenvolvimento humano dos municípios brasileiros e a sua evolução ao longo do decorrer das décadas assim descritas acima. A escala da legenda dos mapas está de acordo com Programa das Nações Unidas para o Desenvolvimento (PNUD).

#### 4.1 Dados Estatísticos das variáveis IDHM e IDHM educação dos municípios brasileiros da década dos anos 1990.

```
> var(IDHM_1991$IDHM)
[1] 0.01062867
> sd(IDHM_1991$IDHM)
[1] 0.1030954
> var(IDHM_1991$`IDHM Educação`)
[1] 0.008459542
> sd(IDHM_1991$`IDHM Educação`)
[1] 0.09197577
> IDH_1991 <- read_excel("IDH_1991.xlsx")
> View(IDH_1991)
> summary(IDH_1991)
Código do Município Município IDHM IDHM Educação
Min. :110001 Length:5564 Min. :0.1200 Min. :0.0100
1st Qu.:251218 Class :character 1st Qu.:0.2990 1st Qu.:0.1060
Median :314623 Mode :character Median :0.3820 Median :0.1680
Mean :325324 Mean :0.3814 Mean :0.1787
3rd Qu.:411903 3rd Qu.:0.4630 3rd Qu.:0.2450
Max. :530010 Max. :0.6970 Max. :0.5570
> summary(IDH_1991$IDHM)
Min. 1st Qu. Median Mean 3rd Qu. Max.
0.1200 0.2990 0.3820 0.3814 0.4630 0.6970
> summary(IDH_1991$`IDHM Educação`)
Min. 1st Qu. Median Mean 3rd Qu. Max.
0.0100 0.1060 0.1680 0.1787 0.2450 0.5570
```

De posse desses dados estatísticos pode se gerar os histogramas, os *Boxplots* e a correlação entre os dados de modo inferir sobre o comportamento do IDHM e o IDHM Educação distribuídos por municípios brasileiros ao longo dos anos 1990, 2000 e 2010.

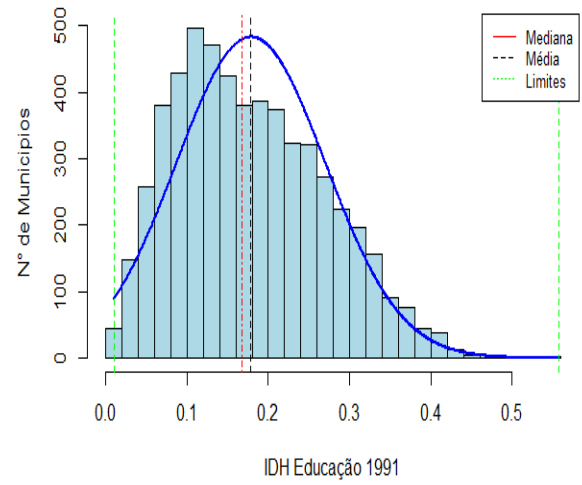
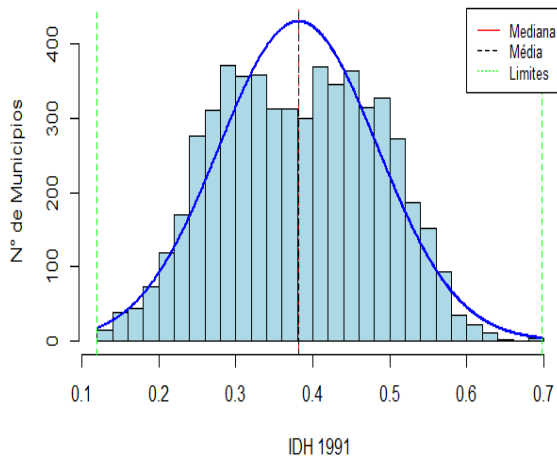


Figura 9 e 10 – Histograma dos dados IDHM e IDHM Educação década anos 90

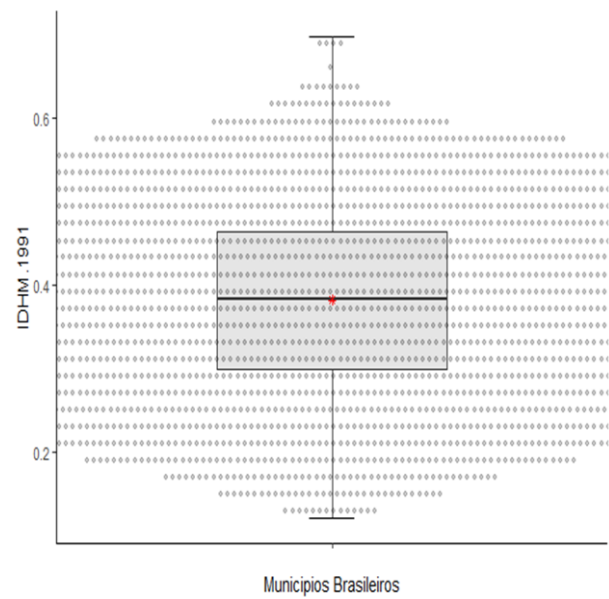
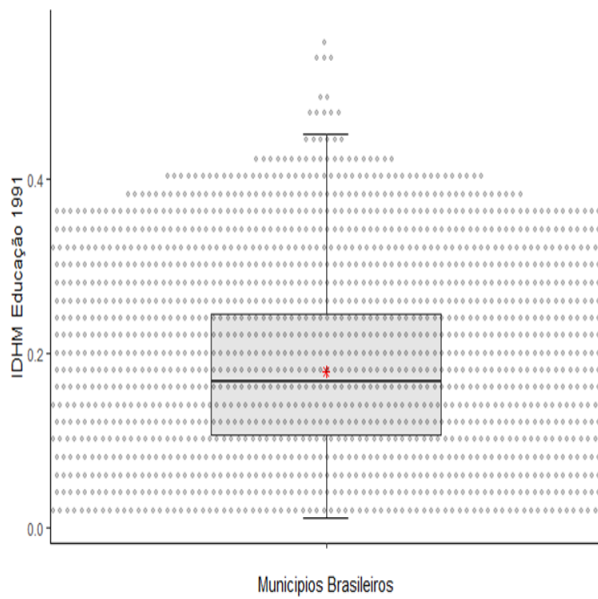
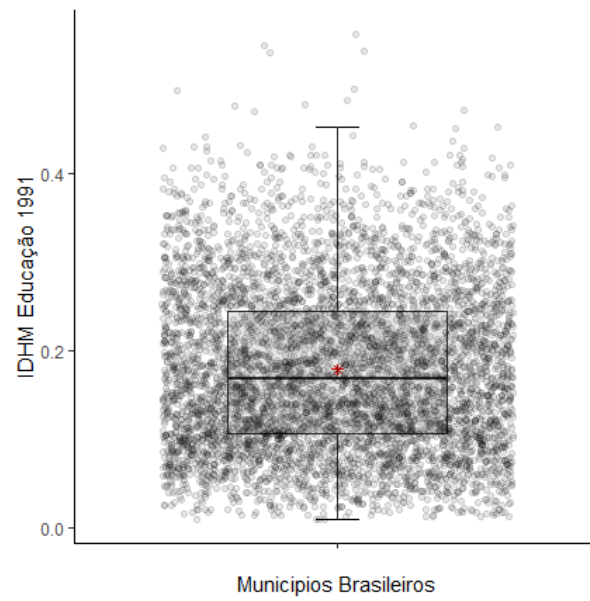
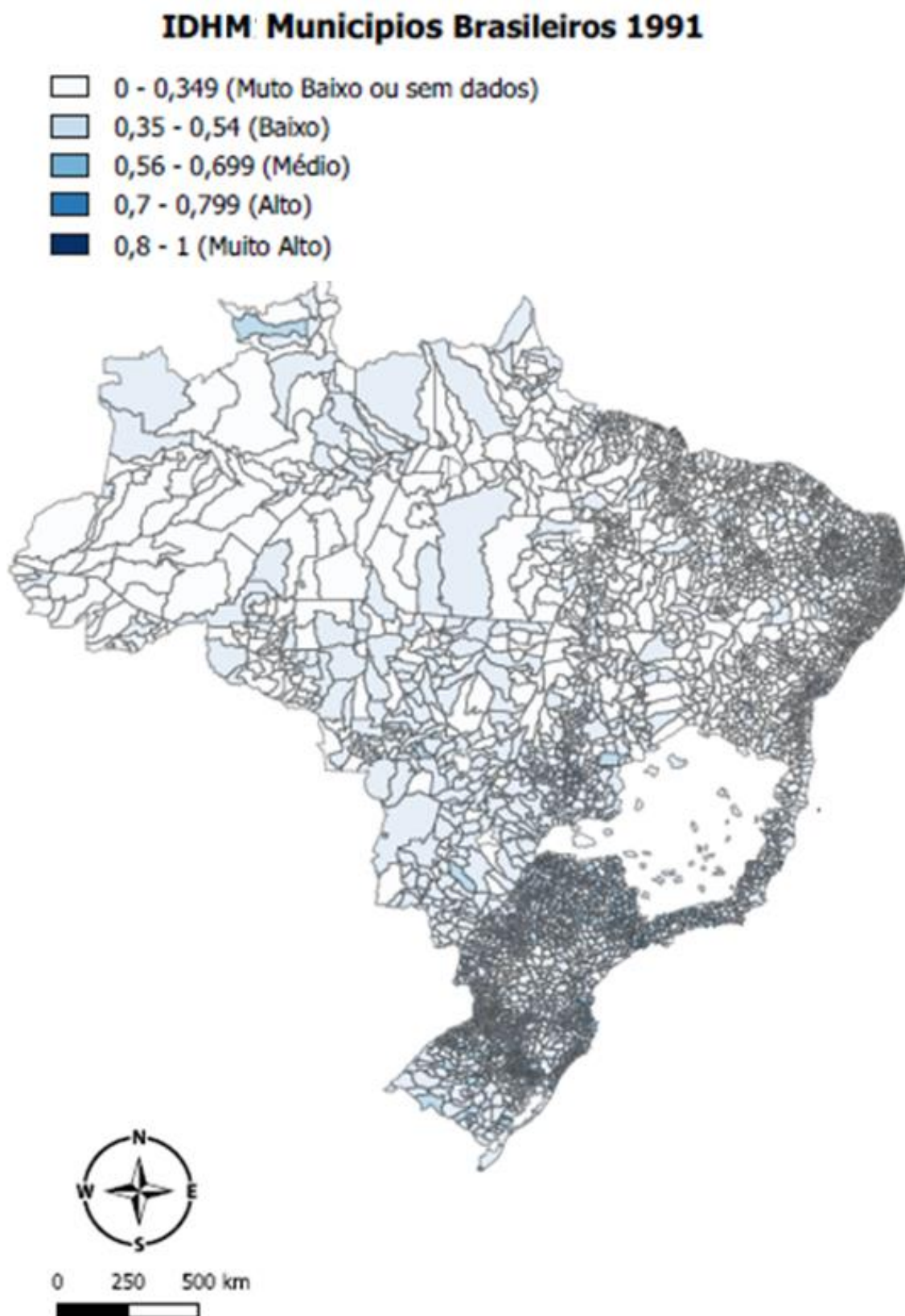


Figura 11 e 12 – Boxplot dos dados IDHM e IDHM Educação década anos 90

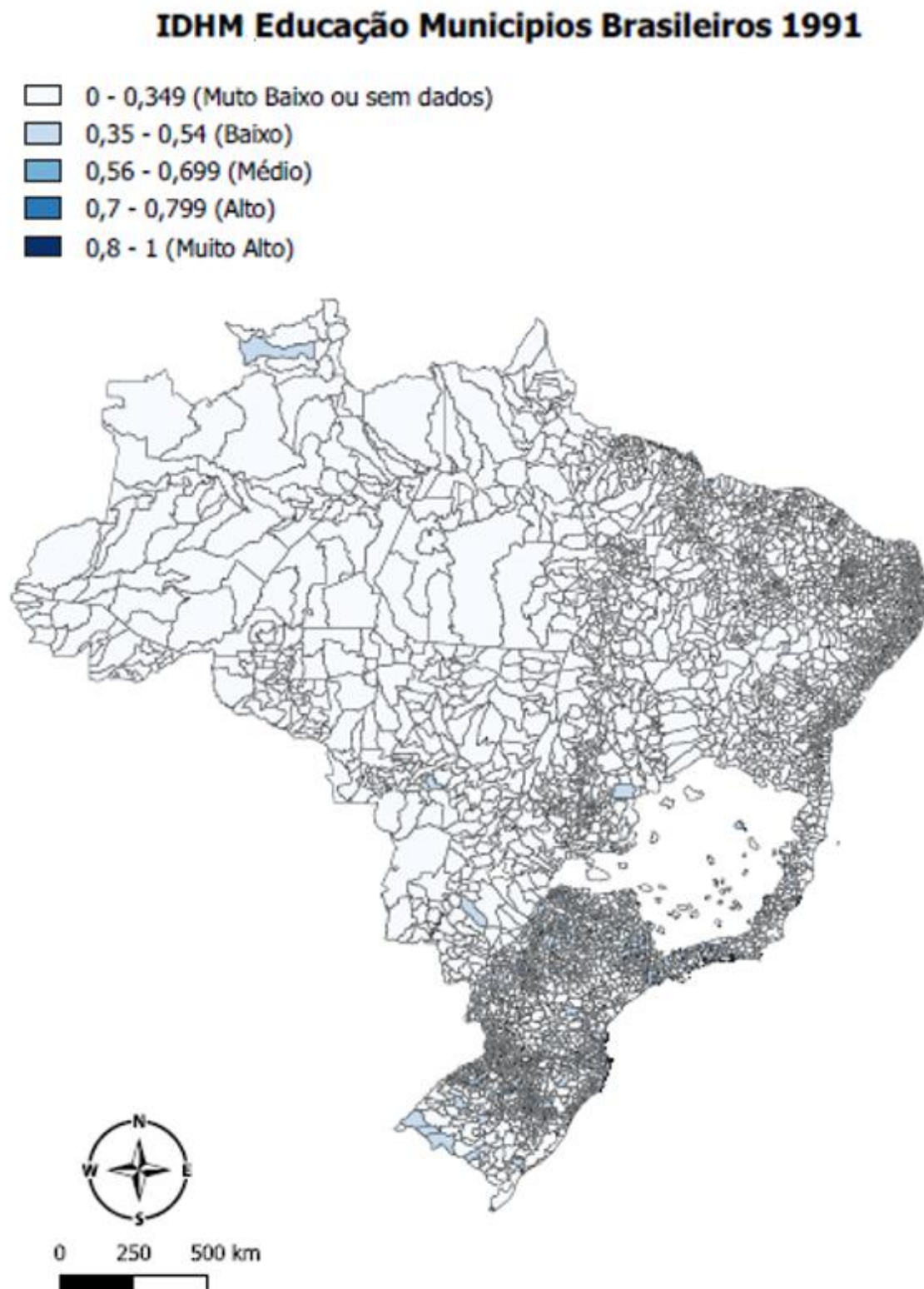


*Figura 13 – Boxplot nuvem dos dados IDHM Educação década anos 90*



*Figura 14 - Mapa de cor do IDHM década dos anos 1990 dos municípios brasileiros com e legenda do IDH definido pela ONU*





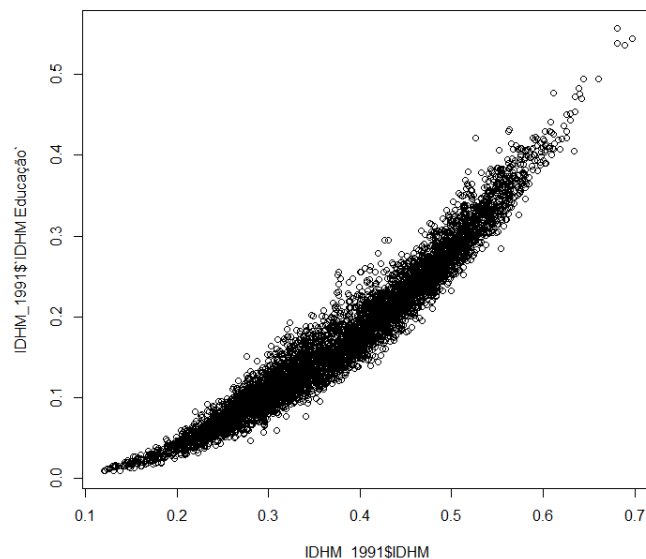
*Figura 15 - Mapa de cor do IDHM Educação década dos anos 1990 dos municípios brasileiros com e legenda do IDH definido pela ONU*

## Correlação entre o IDHM x IDHM Educação ano 1991

```
> plot(IDHM_1991$IDHM, IDHM_1991$`IDHM Educação`)
> cor.test(IDHM_1991$IDHM, IDHM_1991$`IDHM Educação`)

Pearson's product-moment correlation

data: IDHM_1991$IDHM and IDHM_1991$`IDHM Educação`
t = 290.81, df = 5562, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9669895 0.9702347
sample estimates:
      cor
0.9686534
```



*Figura 16 – Gráfico Correlação IDHM x IDHM Educação anos 90*

O coeficiente de correlação de Pearson pode ter um intervalo de valores de +1 a -1. Um valor de 0 indica que não há associação entre as duas variáveis. Um valor maior que 0 indica uma associação positiva. Isto é, à medida que o valor de uma variável aumenta, o mesmo acontece com o valor da outra variável. Um valor menor que 0 indica uma associação negativa. Isto é, à medida que o valor de uma variável aumenta, o valor da outra diminui.

Pelo teste de Spearman também podemos inferir sobre a forte relação de correlação entre as variáveis, uma vez que o valor  $\rho$  (0.9766) se aproxima de +1, então eles têm uma associação de classificação quase perfeita.

O método de verificação da correlação pelo método Kendall também obtivemos um valor de Tau de 0.8717 bem acima dos valores de classificação sugerido por Rumsey (2016):

- $\tau = 0,30$  -> correlação fraca
- $\tau = 0,50$  -> correlação moderada
- $\tau = 0,70$  -> correlação forte

```
cor.test(IDHM_1991$IDHM, IDHM_1991$`IDHM Educação`, method = "spearman")
Spearman's rank correlation rho

data: IDHM_1991$IDHM and IDHM_1991$`IDHM Educação`
S = 669541695, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.9766779
cor.test(IDHM_1991$IDHM, IDHM_1991$`IDHM Educação`, method = "kendall")
Kendall's rank correlation tau
data: IDHM_1991$IDHM and IDHM_1991$`IDHM Educação`
z = 97.228, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
tau 0.8717488
```

De posse desses dados iniciais, podemos inferir que em média o IDHM e IDHM Educação dos municípios nesta década 0,3814 e 0,1787 respectivamente demonstram que nesta época o Brasil, de um modo geral apresentava indicadores sociais bem baixo. Isso fica bem evidente quando analisamos os Histogramas e percebemos que as colunas que representam os municípios estão na sua maioria em torno da média, que é bem baixa, e pelo boxplot podemos observar também esse mesmo efeito além do mesmo gráfico mostrar que a densidade abaixo da média tem comportamento parecido enquanto acima da média observa-se uma pulverização da densidade ou seja cada vez menos e mais rápidos os municípios com maior IDHM vão diminuindo o que nos leva a crer que não há ou existe muito pouco para expansão do IDHM o que é ainda pior quando comparamos com o boxplot do IDHM Educação com índices bem baixos sendo seu limite inferior bem próximo a zero e ainda assim, alguns municípios encontram-se um pouco abaixo do extremo inferior do referido gráfico e pode-se contar nos dedos os municípios que estão acima do limite superior que é de apenas 0,5570.

E através dos mapas com escala de cores de acordo com a faixa de IDHM e IDHM Educação, fica bem claro como os municípios brasileiros nesta época não se encontravam em boa situação em relação a essas variáveis. O mapa do Brasil está em sua quase totalidade em azul clarinho quase branco uma das menores escalas no

IDHM/IDHM Educação de acordo com os limites definidos pela ONU. Os testes de correlação ficam evidente a clara relação entre o IDHM e o IDHM Educação pois todos os testes realizados demonstraram esse fator de correlação entre essas variáveis.

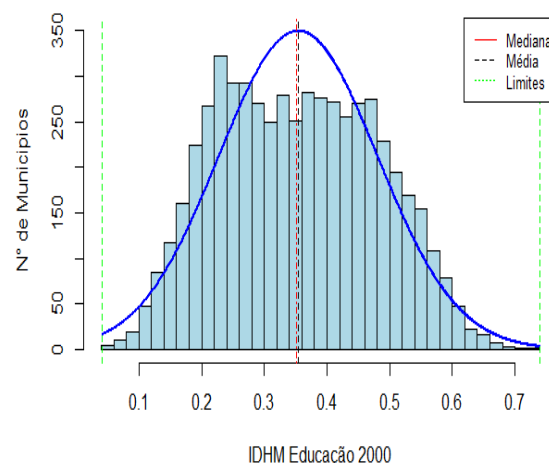
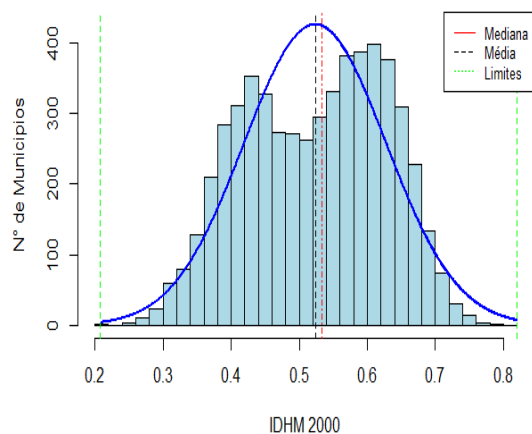
#### 4.2 Dados Estatísticos das variáveis IDHM e IDHM Educação dos municípios brasileiros da década dos anos 2000.

```
> var(IDH_2000_a$IDHM)

> library(readxl)
> IDH_2000_a <- read_excel("IDH_2000_a.xlsx")
> view(IDH_2000_a)
> summary(IDH_2000_a)
Código do Município Município          IDHM          IDHM Educação
Min.   :110001      Length:5564      Min.   :0.2080      Min.   :0.0410
1st Qu.:251218      Class :character      1st Qu.:0.4360      1st Qu.:0.2490
Median :314623      Mode  :character      Median :0.5330      Median :0.3520
Mean   :325324                                Mean   :0.5235      Mean   :0.3542
3rd Qu.:411903                                3rd Qu.:0.6090      3rd Qu.:0.4550
Max.   :530010                                Max.   :0.8200      Max.   :0.7400
> summary(IDH_2000_a$IDHM)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.2080 0.4360  0.5330  0.5235 0.6090  0.8200
> summary(IDH_2000_a$`IDHM Educação`)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.0410 0.2490  0.3520  0.3542 0.4550  0.7400

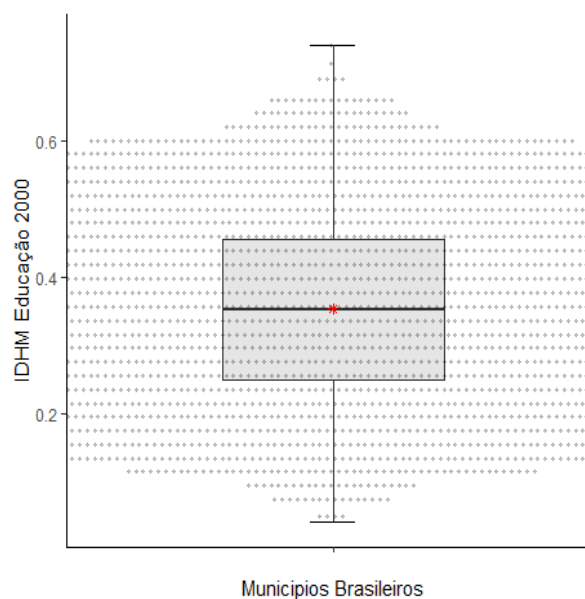
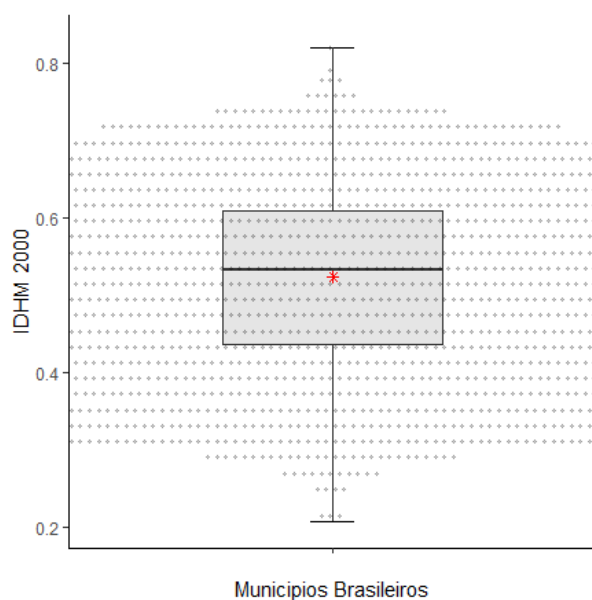
> rcompanion::plotNormalHistogram(ID_Mun_IDH$IDHM.2010, breaks = 30, col =
"lightblue", xlab = "IDH 2010", ylab = "Nº de Municípios")
> abline(v=median(ID_Mun_IDH$IDHM.2010), col="red", lty=2)
> abline(v=mean(ID_Mun_IDH$IDHM.2010), col="black", lty=2);
> abline(v=0.862, col="green", lty=2)
> abline(v=0.418, col="green", lty=2)
> legend(x = "topleft", legend = c("Mediana", "Média", "Limites"), col = c
("red", "black", "green"), lwd = 1, lty = c(1,2,3), cex = 0.8)

[1] 0.01089726
> sd(IDH_2000_a$IDHM)
[1] 0.1043899
> var(IDH_2000_a$`IDHM Educação`)
[1] 0.01611634
> sd(IDH_2000_a$`IDHM Educação`)
[1] 0.1269501
```



*Figura 16 e 17 – Histograma dados IDHM, IDHM Educação anos 2000*

Abaixo os comandos inseridos no R para a formação do Histogramas deste trabalho, os comandos basicamente são os mesmos bastando apenas substituir as respectivas colunas com os dados que se deseja criar os Histogramas ao mesmo tempo que representa as linhas de média, mediana, e limites superiores e inferiores mais a legenda.



*Figura 18 e 19 – Boxplot dados IDHM, IDHM Educação anos 2000*

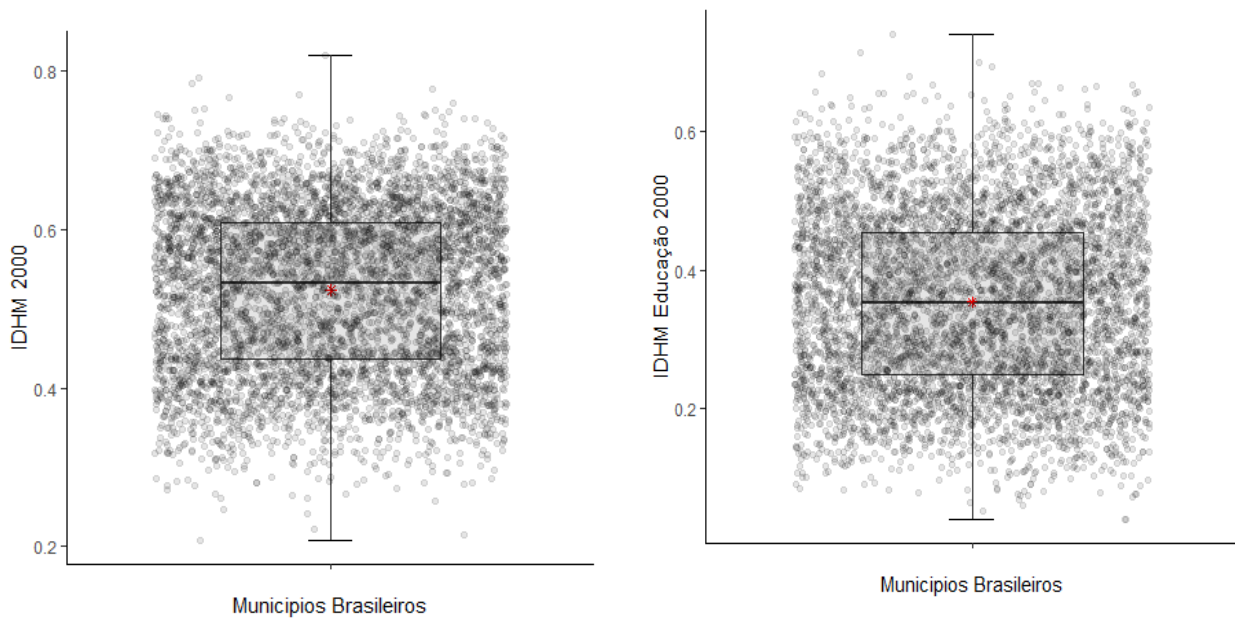
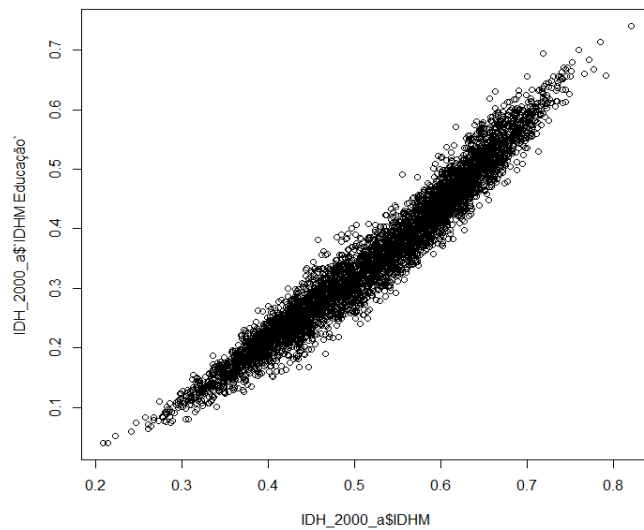


Figura 20 e 21 - Boxplot nuvem dos dados IDHM, IDHM Educação anos 2000

```
> plot(IDH_2000_a$IDHM, IDH_2000_a$`IDHM Educação`)
> cor.test(IDH_2000_a, IDH_2000_a$`IDHM Educação`)
Error in cor.test.default(IDH_2000_a, IDH_2000_a$`IDHM Educação`) :
  'x' deve ser um vetor numérico
> cor.test(IDH_2000_a$IDHM, IDH_2000_a$`IDHM Educação`)
```

Pearson's product-moment correlation

```
data: IDH_2000_a$IDHM and IDH_2000_a$`IDHM Educação`
t = 333.09, df = 5562, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9745521 0.9770625
sample estimates:
      cor
0.9758395
```



*Figura 22 – Correlação IDHM x IDHM EDUCAÇÃO anos 2000*

```
cor.test(IDH_2000_a$IDHM, IDH_2000_a$`IDHM Educação`, method = "spearman")
```

Spearman's rank correlation rho

data: IDH\_2000\_a\$IDHM and IDH\_2000\_a\$`IDHM Educação`

S = 585373904, p-value < 2.2e-16

alternative hypothesis: true rho is not equal to 0

sample estimates:

rho  
0.9796097

Kendall's rank correlation tau

data: IDH\_2000\_a\$IDHM and IDH\_2000\_a\$`IDHM Educação`

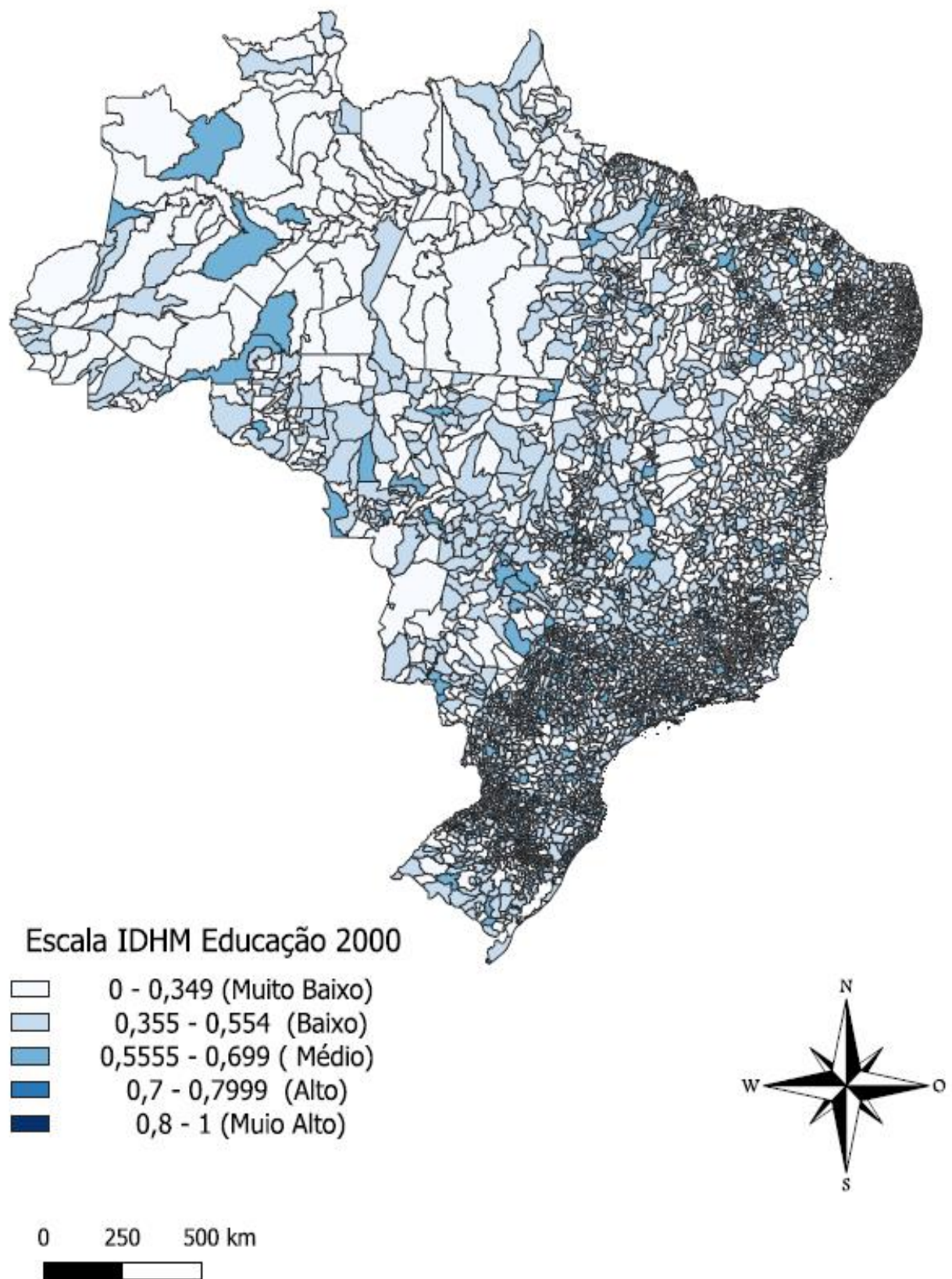
z = 97.808, p-value < 2.2e-16

alternative hypothesis: true tau is not equal to 0

sample estimates:

tau  
0.8765286





*Figura 23 - Mapa de cor do IDHM Educação década dos anos 2000 dos municípios brasileiros com e legenda do IDH definido pela ONU*



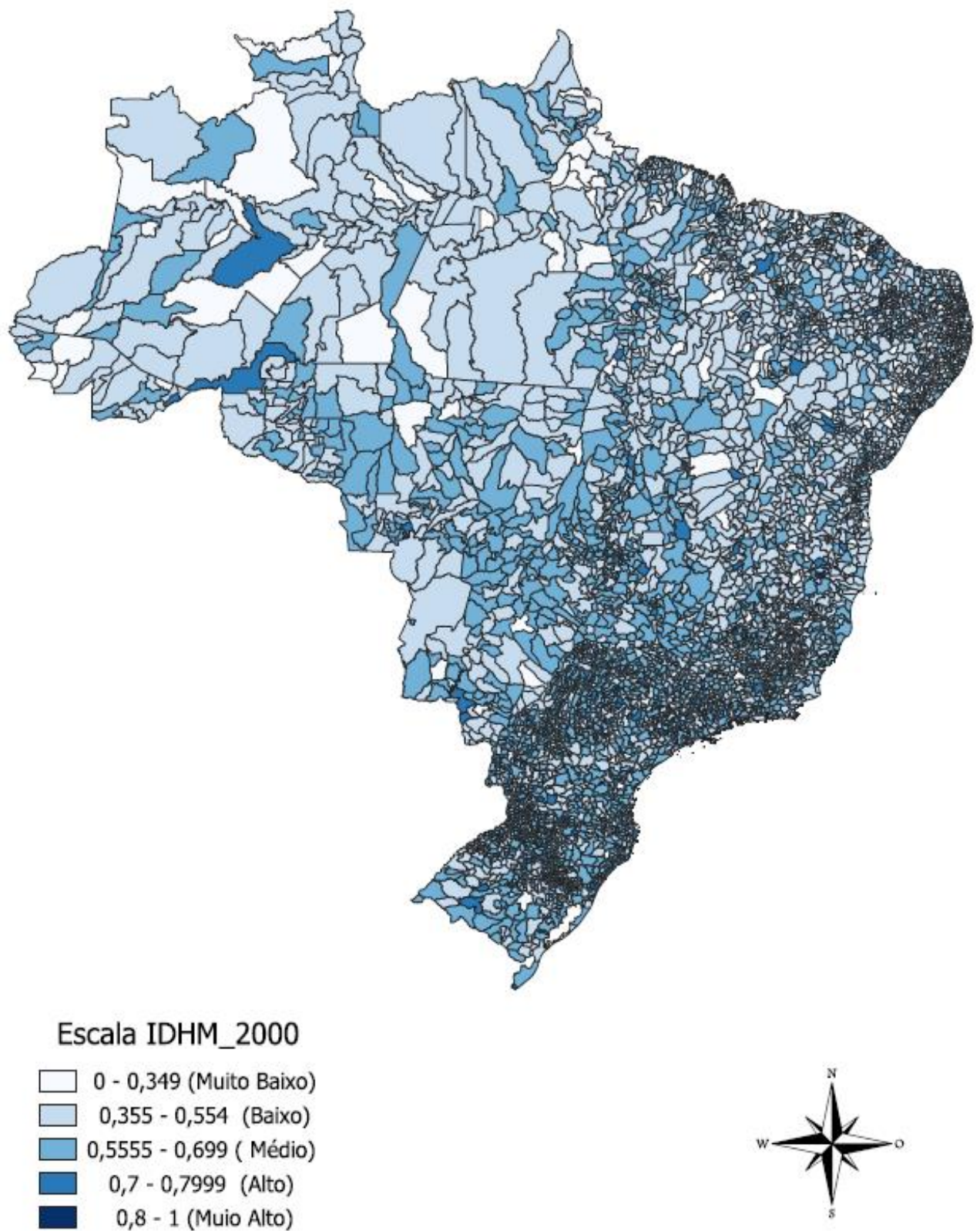


Figura 24 - Mapa de cor do IDHM década dos anos 2000 dos municípios brasileiros com e legenda do IDH definido pela ONU

De posse desses dados iniciais, podemos inferir que as medidas de posição dos IDHM e IDHM Educação dos municípios, ao longo desta década, estão se deslocando para cima, que facilmente pode ser observado tanto pelos índices apurados e também pelos gráficos e mapa de cores, em relação a década anterior, porém ainda bem abaixo das medidas de posição de outras nações. Quando analisamos os *Boxplots* podemos observar pouca dispersão dos dados e uma leve assimetria favorável ao terceiro quartil indicando uma certa influência dos valores acima da mediana demonstrando assim que realmente a uma ligeira ascensão positiva dos IDH's. Também pela análise do boxplot percebe se a existência de poucos outliers reforçando a ideia de que poucos ou nenhum município está ficando muito abaixo ou acima dos menores e maiores comprimento das caudas da distribuição ou seja há poucos municípios se destacando significativamente tanto muito acima ou muito abaixo dos dados estatísticos apurados .

### 4.3 Dados Estatísticos das variáveis IDHM e IDHM Educação dos municípios brasileiros da década dos anos 2010.

```
> var(IDHM_2010$IDHM)
[1] 0.005183609
> sd(IDHM_2010$IDHM)
[1] 0.07199728
> var(IDHM_2010$`IDHM Educação 2010`)
[1] 0.008710068
> sd(IDHM_2010$`IDHM Educação 2010`)
[1] 0.09332775
```

IDHM 2010		Município	IDHM	IDHM Educação 2010	Cod_Município
Length:5590	Length:5590	Min. :0.4180	Min. :0.2070	Min. :1100015	
Class :character	Class :character	1st Qu.:0.5990	1st Qu.:0.4900	1st Qu.:2512101	
Mode :character	Mode :character	Median :0.6650	Median :0.5600	Median :3146206	
		Mean :0.6592	Mean :0.5591	Mean :3253053	
		3rd Qu.:0.7180	3rd Qu.:0.6310	3rd Qu.:4119004	
		Max. :0.8620	Max. :0.8250	Max. :5300108	
		NA's :25	NA's :25	NA's :25	

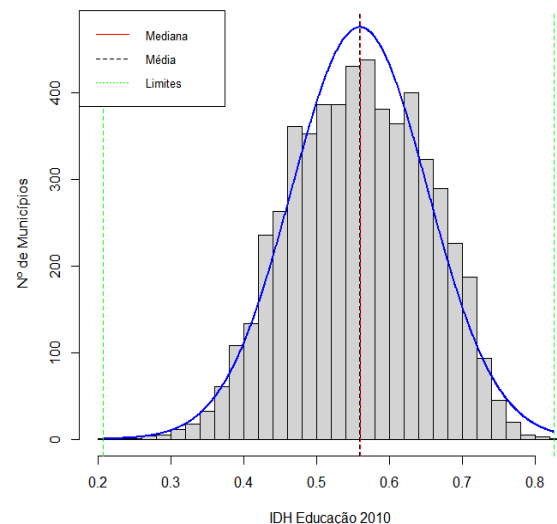
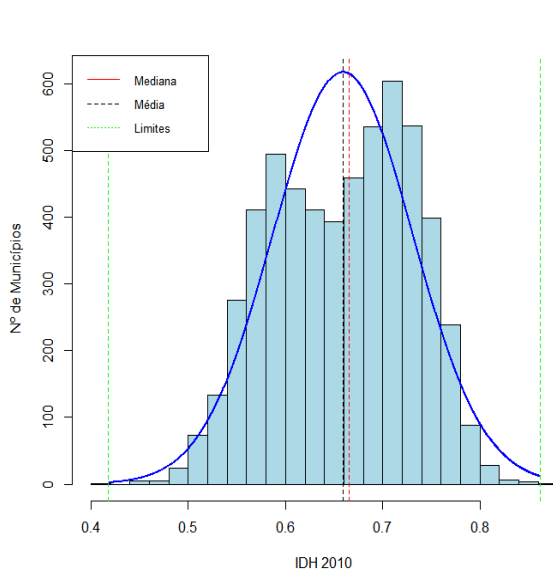


Figura 24 e 25 – Histograma dos dados municípios brasileiros década anos 2010

Para a criação dos Boxplot`s foi inserido os seguintes comandos no R e baixado os pacotes: `pacman::(dplyr e ggplot)`. Esse procedimento de realização dos boxplot`s foi feito para todos deste trabalho bastando apenas selecionar a coluna do *Dataframe* de acordo com a análise estatística em questão.

```
ggplot(data = ID_Mun_IDH, aes(y = IDHM.2010, x = "" )) + geom_errorbar(stat =
"boxplot", width = 0.1) + geom_boxplot(width = 0.5, fill = "grey90", outlier.shape =
1, outlier.size = 2, outlier.color = NA) + theme_classic() + labs(y = "IDHM.2010", x
= "Municipios Brasileiros") + geom_point(stat = "summary", fun = "mean", shape= 8,
color = "red") + geom_dotplot(binaxis = "y", stackdir = "center", alpha = 0.2,
dotsize = 0.3, stackratio = 2.0, binwidth = 0.02) + legend()
```

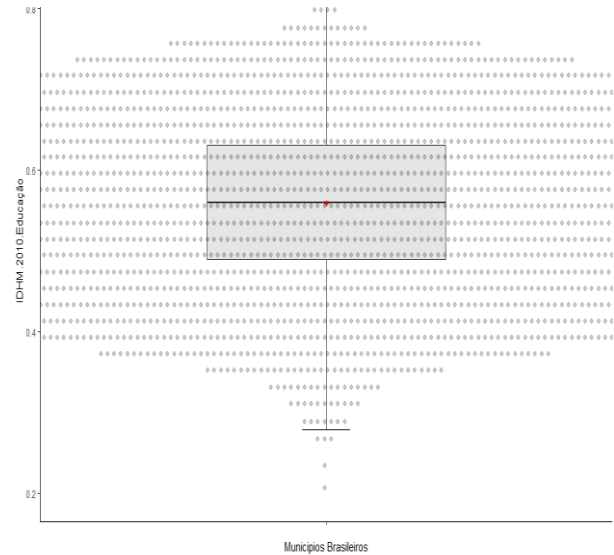
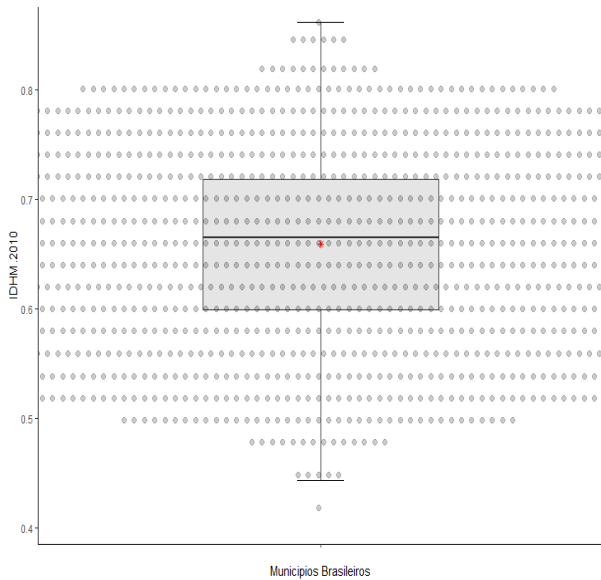


Figura 26 e 27 – Boxplot dados IDHM e IDHM Educação década anos 2010

```
> ggplot(data = ID_Mun_IDH, aes(y = IDHM.2010, x = "" )) + geom_errorbar(stat = "b
oxplot", width = 0.1) + geom_boxplot(width = 0.5, fill = "grey90", outlier.shape =
1, outlier.size = 2, outlier.color = NA) + theme_classic() + labs(y = "IDHM.201
0", x = "Municipios Brasileiros") + geom_point(stat = "summary", fun = "mean", sha
pe= 8, color = "red") + geom_jitter(alpha = 0.1, height = 0, width = 0.25)
```

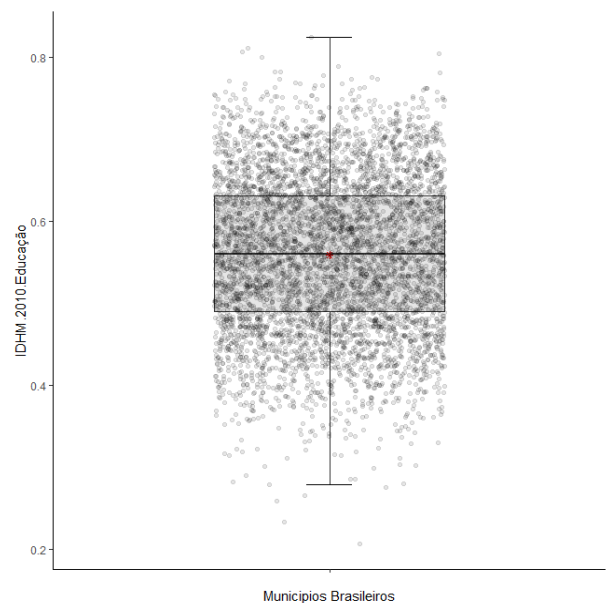
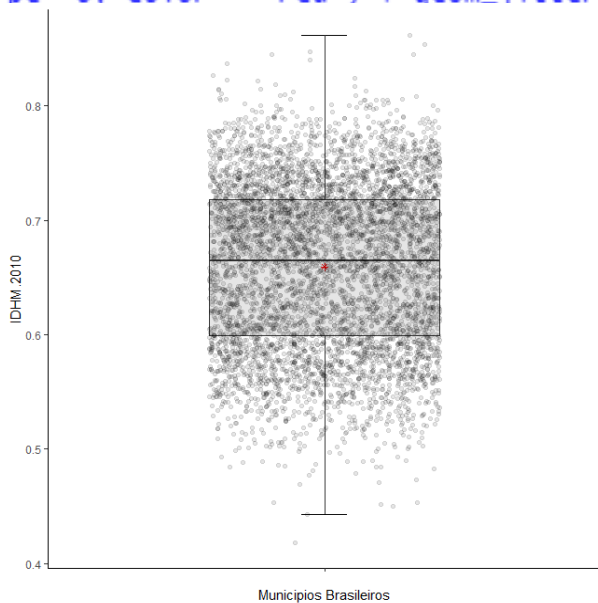
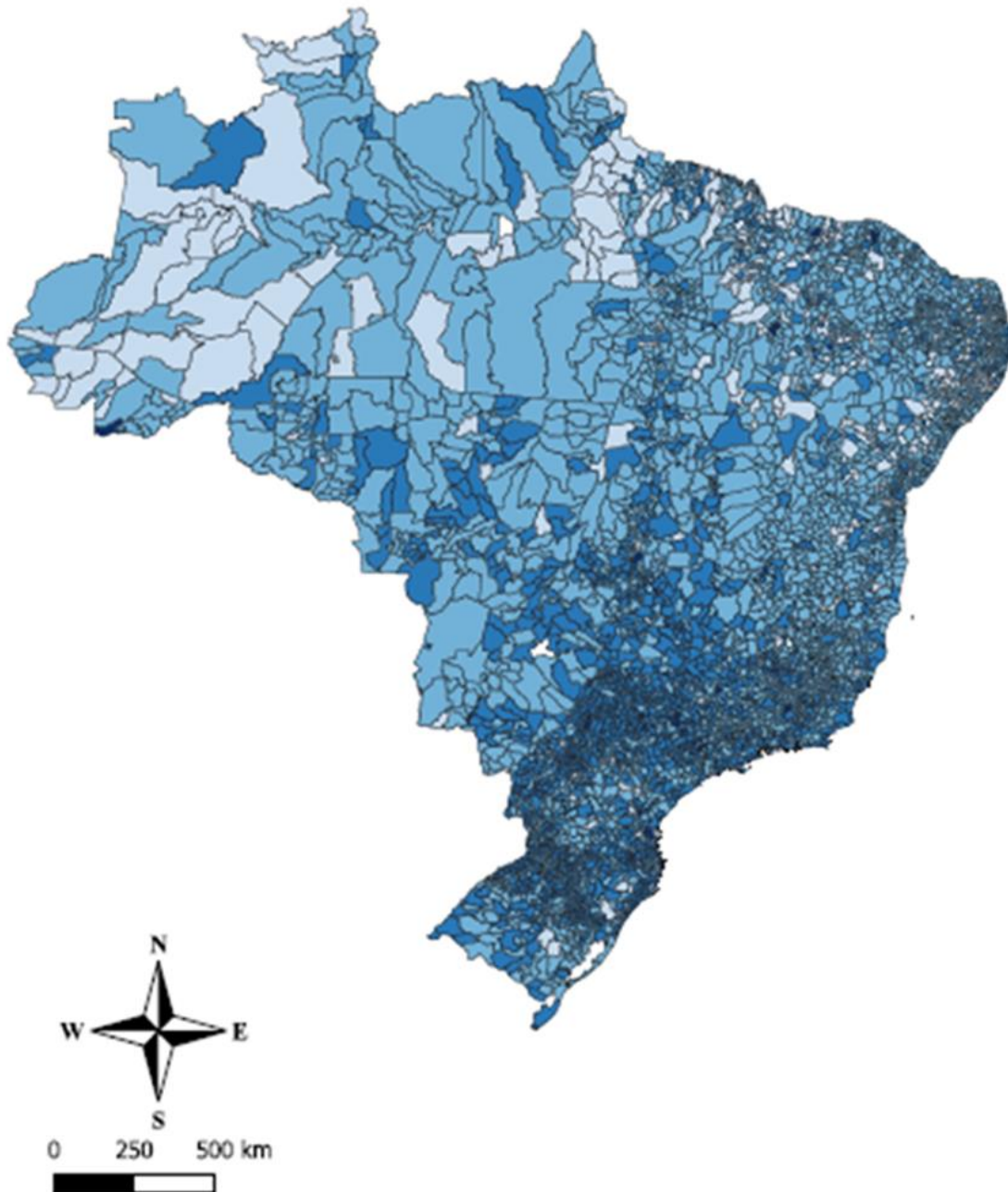


Figura 28 e 29 – Boxplot nuvem dados IDHM e IDHM Educação década anos 2010



**IDHM dos Municípios Brasileiros 2010**

- 0 - 0,349 (Muito Baixo)
- 0,349 - 0,554 (Baixo)
- 0,555 - 0,699 (Médio)
- 0,700 - 0,8000 (Alto)
- 0,8 - 1 (Muito Alto)



*Figura 30 - Mapa de cor do IDHM década dos anos 2010 dos municípios brasileiros com e legenda do IDH definido pela ONU*

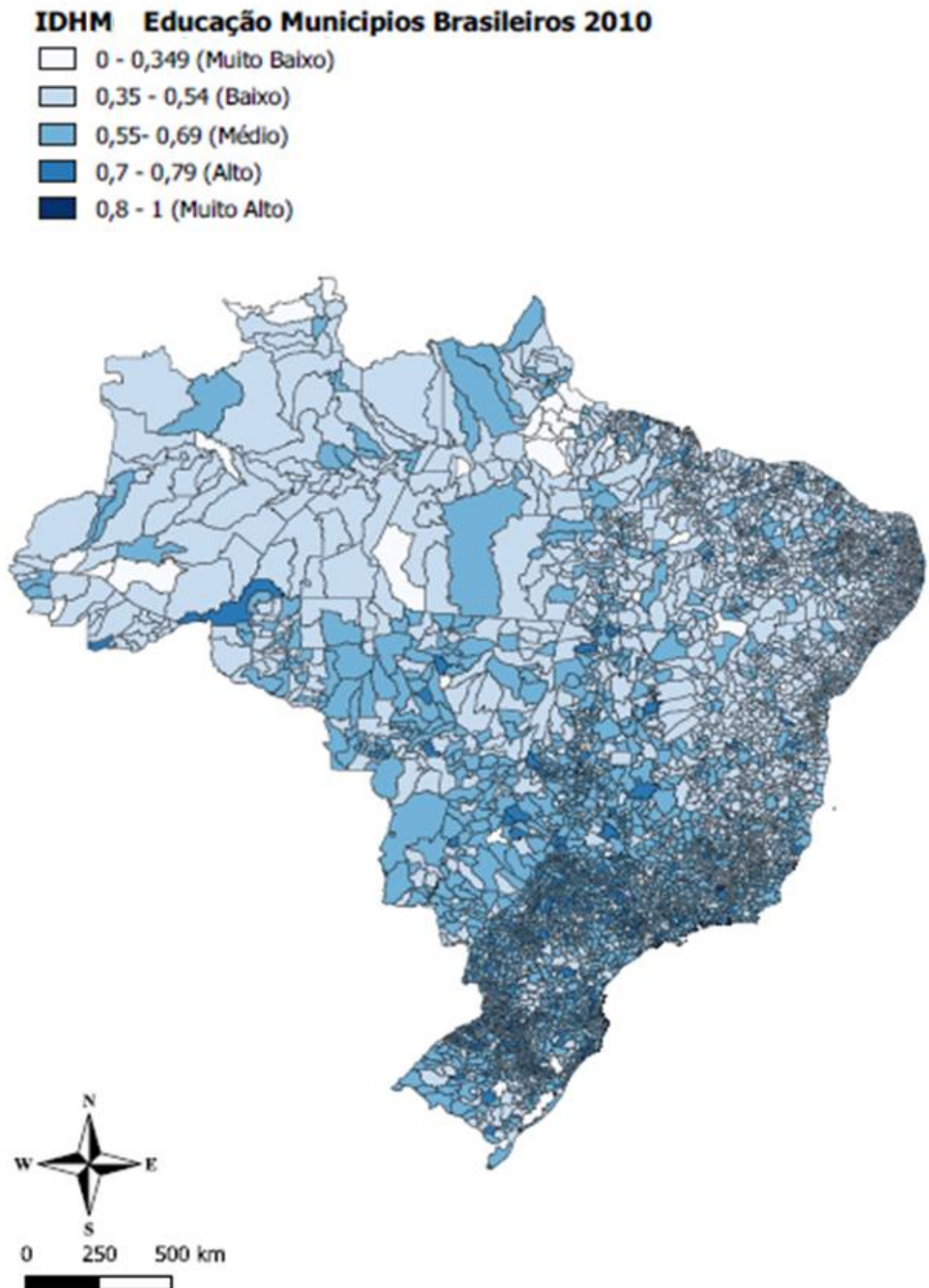
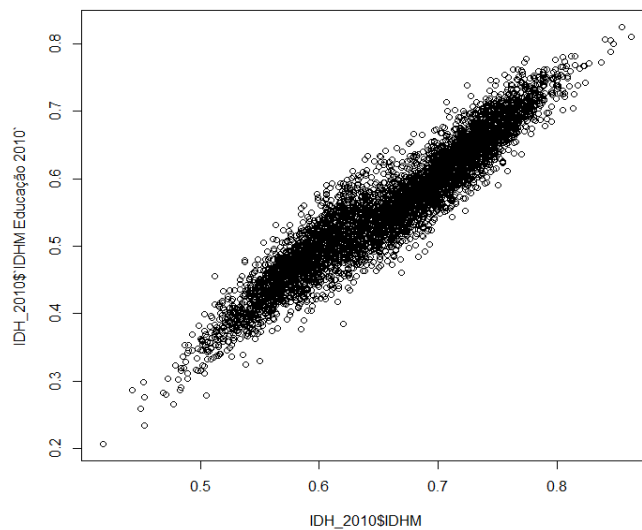


Figura 31 - Mapa de cor do IDHM Educação década dos anos 2010 dos municípios brasileiros com e legenda do IDH definido pela ONU

```
> plot(IDH_2010$IDHM, IDH_2010$`IDHM Educação 2010`)
> cor.test(IDH_2010$IDHM, IDH_2010$`IDHM Educação 2010`)
```

Pearson's product-moment correlation

```
data: IDH_2010$IDHM and IDH_2010$`IDHM Educação 2010`
t = 229.09, df = 5563, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9482927 0.9533319
sample estimates:
      cor
0.9508753
```



*Figura 33 – Gráfico Correlação IDHM x IDHM Educação década 2010*

```
cor.test(IDH_2010$IDHM, IDH_2010$`IDHM Educação 2010`, method = "spearman")
```

Spearman's rank correlation rho

```
data: IDH_2010$IDHM and IDH_2010$`IDHM Educação 2010`
S = 1360803121, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.9526248
```

```
> cor.test(IDH_2010$IDHM, IDH_2010$`IDHM Educação 2010`, method = "kendall")
)
```

Kendall's rank correlation tau

```
data: IDH_2010$IDHM and IDH_2010$`IDHM Educação 2010`
z = 90.768, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.8141625
```

O retrato do IDHM e IDHM dos municípios brasileiros divulgado pelo IBGE referente a década dos anos 2010 também podemos inferir que os dados estatísticos apurados referentes as medidas de posição seguem em alta em relação a década anterior. Apesar do aumento expressivo dos índices de IDH's em relação a década anterior, os municípios brasileiros ainda permanecem na escala média do ranking de IDH disponibilizado pela ONU.

Os boxplot's apontam pouca dispersão em torno da mediana e um número maior de outliers de municípios brasileiros que estão ficando bem abaixo do comprimento da cauda de distribuição inferior, ou seja, estão abaixo dos limites inferiores de IDH's baixo ou muito baixo em relação a escala de IDH da ONU.

Dados Estatísticos apurados						
Medidas de posição	Década anos 1990		Década anos 2000		Década anos 2010	
	IDHM	IDHM Ed.	IDHM	IDHM Ed.	IDHM	IDHM Ed.
Média	0,3814	0,1787	0,5235	0,3542	0,6592	0,5591
Mediana	0,3820	0,1680	0,5330	0,3520	0,6650	0,5600
Limite Inferior	0,1200	0,0100	0,2080	0,0410	0,4180	0,2070
Limite Superior	0,6970	0,5570	0,8200	0,7400	0,8620	0,8250
Medidas de dispersão	Década anos 1990		Década anos 2000		Década anos 2010	
	IDHM	IDHM Ed.	IDHM	IDHM Ed.	IDHM	IDHM Ed.
Desvio padrão	0,1030	0,0919	0,1044	0,1270	0,0720	0,0933
Variância	0,0106	0,0084	0,0109	0,0161	0,0052	0,0087
Dados Correlação						
Correlação	Década anos 1990		Década anos 2000		Década anos 2010	
	IDHM x IDHM Ed.	IDHM x IDHM Ed.	IDHM x IDHM Ed.	IDHM x IDHM Ed.	IDHM x IDHM Ed.	IDHM x IDHM Ed.
Pearson	0,9686		0,9758		0,9508	
Spearman	0,8717		0,9796		0,9526	
Kendall	0,9766		0,8765		0,8141	

Tabela 3 - Dados estatísticos apurados de todos os municípios brasileiros.



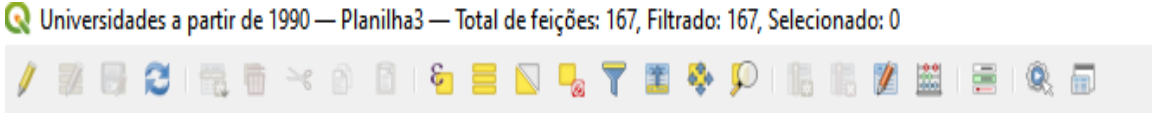
#### 4.4 Municípios Selecionados e Amostras Aleatórias

Aqui inicia se a busca por valores estatísticos referente a uma amostra de municípios previamente selecionados. A base da seleção desses municípios partiu da premissa que, os Municípios brasileiros que foram selecionados para obter Polo do IES público descentralizado. A base da amostra é formada por esses municípios e os municípios que ficam em seu entorno escolhido de forma aleatória para compor o restante da amostra, que subentende se que estes municípios poderiam de alguma forma ser beneficiado pela proximidade aos IES. Nesta etapa os dados são organizados em um novo Dataset contendo uma amostra de 172 municípios selecionados de acordo com os critérios previamente definidos. Uma vez criado o *Dataset* com as respectivas variáveis desta amostra a ideia é obter dados estatísticos que possam ser confrontados com os dados estatísticos a nível nacional feito anteriormente e de duas amostras de municípios escolhido aleatoriamente no próprio dataset. De posse de todos os dados estatísticos devidamente armazenado em uma tabela, esses dados serão a base para aprendizado de máquina que irá responder as Hipóteses a seguir:

Ho: O ensino superior descentralizado influencia no aumento do IDHM Educação, variável independente que por sua vez influencia no aumento do IDHM mais rapidamente que os municípios que não possuem polos de IES descentralizados

H1: O ensino superior descentralizado tem pouca ou nenhuma influência sobre o desenvolvimento IDHM Educação que por sua vez não influencia no IDHM no desenvolvimento humano.

Universidades a partir de 1990 — Planilha3 — Total de feições: 167, Filtrado: 167, Selecionado: 0



	Municípios	Estados	Cod_Mun
1	Baraúna	PB	2501534
2	Caetés	PE	2603207
3	Coité do Noia	AL	2702009
4	Currais	PI	2203230
5	Itaara	RS	4310538
6	Santa Luz	PI	2209302
7	Agrestina	PE	2600302
8	Água Fria de Goiás	DF	5200175
9	Alenquer	PA	1500404
10	Alumínio	SP	3501152
11	Anagé	BA	2901205
12	Aracaju	SE	2800308
13	Arapiraca	AL	2700300
14	Augusto Correa	PA	1500909
15	Aveiro	PA	1501006

*Figura 34 – Dados dos municípios brasileiros selecionados pelo IES descentralizados no R-Studio*

Dataset da amostra dos municípios selecionados para comparação com os dados estatísticos dos IDHM e IDHM Educação dos municípios brasileiros.

```
rcompanion::plotNormalHistogram(tabela_completa$`IDHM Educação 1991`, break
s = 10, col = "lightblue", xlab = "IDHM Educação 1991", ylab = "Municípios
selecionados")
> abline(v=mean(tabela_completa$`IDHM Educação 1991`), col = "red", lty=2)
> abline(v=median(tabela_completa$`IDHM Educação 1991`), col = "black", lty
=2)
> abline(v=0.0170, col="green", lty =2)
> abline(v=0.5380, col="green", lty =2)
```

```
> legend(x = "topright", legend = c("Mediana", "Média", "Limites"), col = c
      ("red", "black", "green"), lwd =1, lty = c(1,2,3), cex = 0.8)
> var(tabela_completa$IDHM_1991)
[1] 0.01506261
> sd(tabela_completa$IDHM_1991)
[1] 0.1227298
```

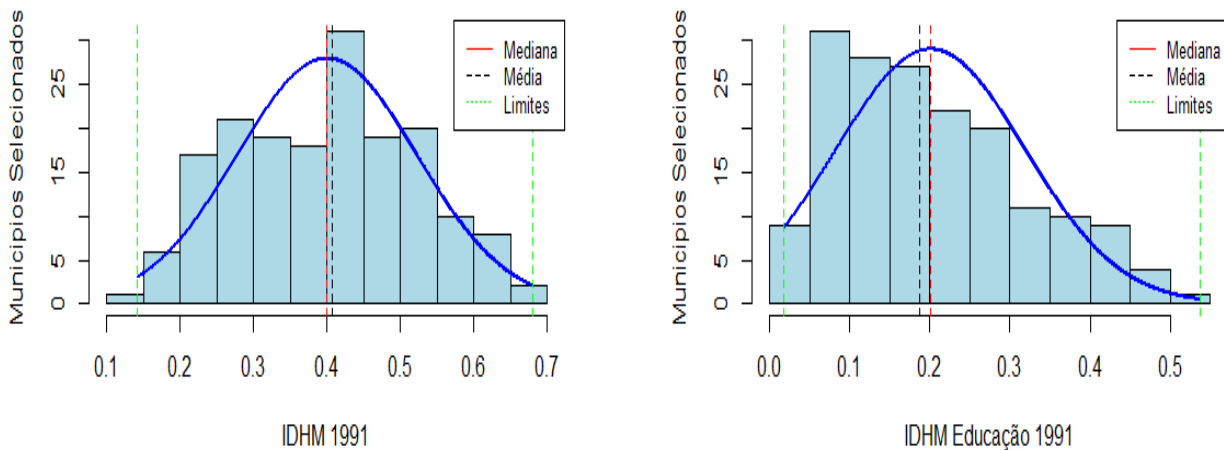


Figura 35 – Histograma dos dados dos municípios selecionados década anos 90

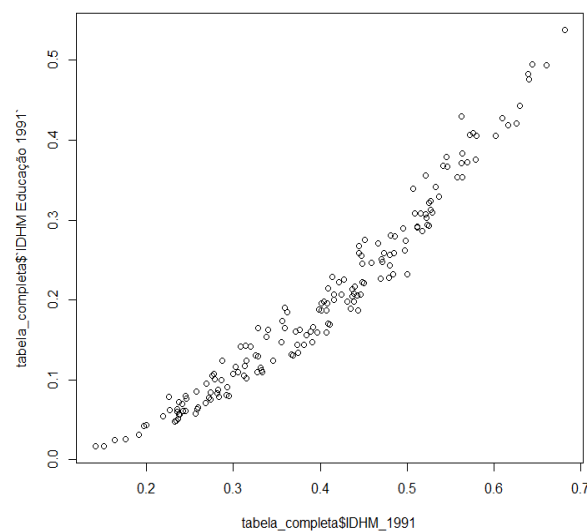
```
> var(tabela_completa$`IDHM Educação 1991`)
[1] 0.01400028
> sd(tabela_completa$`IDHM Educação 1991`)
[1] 0.1183228
> rcompanion::plotNormalHistogram(tabela_completa$`IDHM 2000`, breaks = 10,
col = "lightblue", xlab = "IDHM 2000", ylab = "Municípios Selecionados")
> abline(v=median(tabela_completa$`IDHM 2000`), col = "black", lty=2)
> abline(v=mean(tabela_completa$`IDHM 2000`), col = "red", lty=2)
> abline(v=0.7660, col="green", lty =2)
> abline(v=0.214, col="green", lty =2)
> legend(x = "topright", legend = c("Mediana", "Média", "Limites"), col = c
      ("red", "black", "green"), lwd =1, lty = c(1,2,3), cex = 0.8)

> summary(tabela_completa$`IDHM 2000`)
Min. 1st Qu. Median Mean 3rd Qu. Max.
0.2140 0.4240 0.5455 0.5258 0.6252 0.7660
> var(tabela_completa$IDHM_2000)
[1] 0.01466281
> sd(tabela_completa$IDHM_2000)
[1] 0.1210901
> var(tabela_completa$`IDHM Educação 2000`)
[1] 0.02225834
> sd(tabela_completa$`IDHM Educação 2000`)
[1] 0.1491923
```

```
plot(tabela_completa$IDHM_1991, tabela_completa$`IDHM Educação 1991`
)
cor.test(tabela_completa$IDHM_1991, tabela_completa$`IDHM Educação 1991`)
```

Pearson's product-moment correlation

```
data: tabela_completa$IDHM_1991 and tabela_completa$`IDHM Educação 1991`
t = 55.79, df = 170, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9646892 0.9805246
sample estimates:
cor 0.9737606
```



*Figura 36 – Correlação IDHM x IDHM Educação municípios selecionados década anos 90*

```
> cor.test(tabela_completa$IDHM_1991, tabela_completa$`IDHM Educação 1991`,
method = "spearman")
```

Spearman's rank correlation rho

```
data: tabela_completa$IDHM_1991 and tabela_completa$`IDHM Educação 1991`
S = 10948, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.9870904
```

```
> cor.test(tabela_completa$IDHM_1991, tabela_completa$`IDHM Educação 1991`,
method = "kendall")
```

Kendall's rank correlation tau

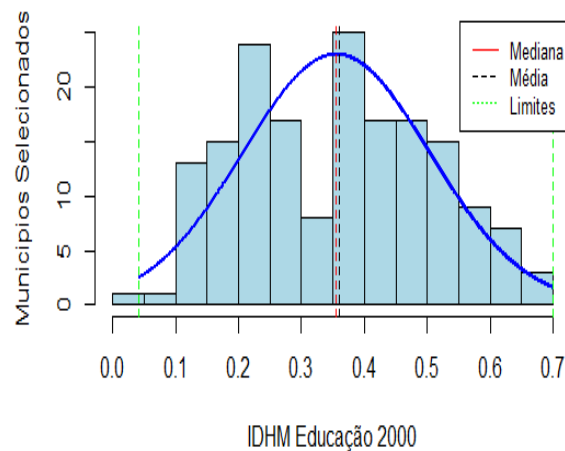
```
data: tabela_completa$IDHM_1991 and tabela_completa$`IDHM Educação 1991`
z = 17.605, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
tau
0.9059562
```

```
> summary(tabela_completa$`IDHM Educação 2000`)
```

```

      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.0410  0.2370  0.3600  0.3553  0.4730  0.7000
> var(tabela_completa$IDHM_2000)
[1] 0.01466281
> sd(tabela_completa$IDHM_2000)
[1] 0.1210901
> var(tabela_completa$`IDHM Educação 2000`)
[1] 0.02225834
> sd(tabela_completa$`IDHM Educação 2000`)
[1] 0.1491923

```



*Figura 38 – Histograma IDHM Educação municípios selecionados década anos 2000*

```

> rcompanion::plotNormalHistogram(tabela_completa$`IDHM Educação 2010`, bre
aks = 10, col = "lightblue", xlab = "IDHM Educação 2010", ylab = "Município
s Seleccionados")
> abline(v=mean(tabela_completa$`IDHM Educação 2010`), col = "red", lty=2)
> abline(v=median(tabela_completa$`IDHM Educação 2010`), col = "black", lty
=2)
> abline(v=0.3020, col="green", lty =2)
> abline(v=0.8050, col="green", lty =2)
> legend(x = "topright", legend = c("Mediana", "Média", "Limites"), col = c
("red", "black", "green"), lwd =1, lty = c(1,2,3), cex = 0.8)
plot(tabela_completa$IDHM_2000, tabela_completa$`IDHM Educação 2000`)
> cor.test(tabela_completa$IDHM_2000, tabela_completa$`IDHM Educação 2000`)

```

```

Pearson's product-moment correlation
data: tabela_completa$IDHM_2000 and tabela_completa$`IDHM Educação 2000`
t = 70.954, df = 170, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9778013 0.9877930
sample estimates:
 cor 0.9835324

```

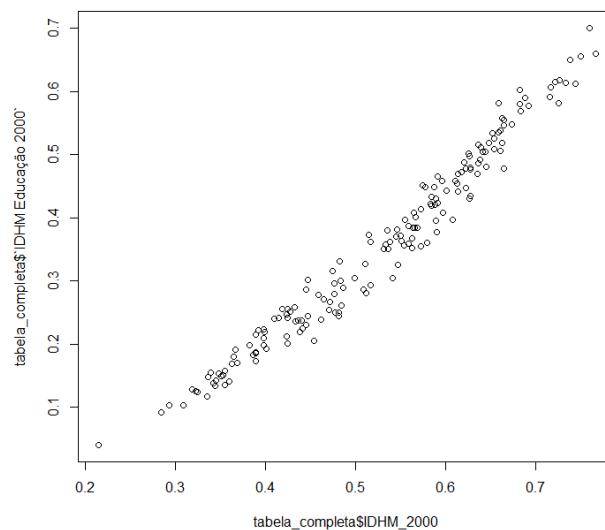


Figura 39 – Gráfico Correlação IDHM x IDHM Educação municípios selecionados anos 2000

```

> cor.test(tabela_completa$IDHM_2000, tabela_completa$`IDHM Educação 2000`,
method = "spearman")

```

Spearman's rank correlation rho

```

data: tabela_completa$IDHM_2000 and tabela_completa$`IDHM Educação 2000`
S = 11707, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
 rho
0.9861953

```

```

> cor.test(tabela_completa$IDHM_2000, tabela_completa$`IDHM Educação 2000`,
method = "kendall")

```

Kendall's rank correlation tau

```

data: tabela_completa$IDHM_2000 and tabela_completa$`IDHM Educação 2000`
z = 17.576, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
 tau
0.9044571

```

```

> summary(tabela_completa$`IDHM 2010`)

```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.4810	0.5917	0.6765	0.6668	0.7390	0.8470

```

> summary(tabela_completa$`IDHM Educação 2010`)

```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.3020	0.4785	0.5720	0.5682	0.6613	0.8050

```
> var(IDHM_2010$IDHM)
[1] 0.005183609
> sd(IDHM_2010$IDHM)
[1] 0.07199728
> var(IDHM_2010$`IDHM Educação 2010`)
[1] 0.008710068
> sd(IDHM_2010$`IDHM Educação 2010`)
[1] 0.09332775

rcompanion::plotNormalHistogram(tabela_completa$`IDHM Educação 2000`, break
s = 10, col = "lightblue", xlab = "IDHM Educação 2000", ylab = "Municípios
Selecionados")
> abline(v=mean(tabela_completa$`IDHM Educação 2000`), col = "red", lty=2)
> abline(v=median(tabela_completa$`IDHM Educação 2000`), col = "black", lty
=2)
> abline(v=0.0410, col="green", lty =2)
> abline(v=0.70, col="green", lty =2)
> legend(x = "topright", legend = c("Mediana", "Média", "Limites"), col = c
("red", "black", "green"), lwd =1, lty = c(1,2,3), cex = 0.8)
```

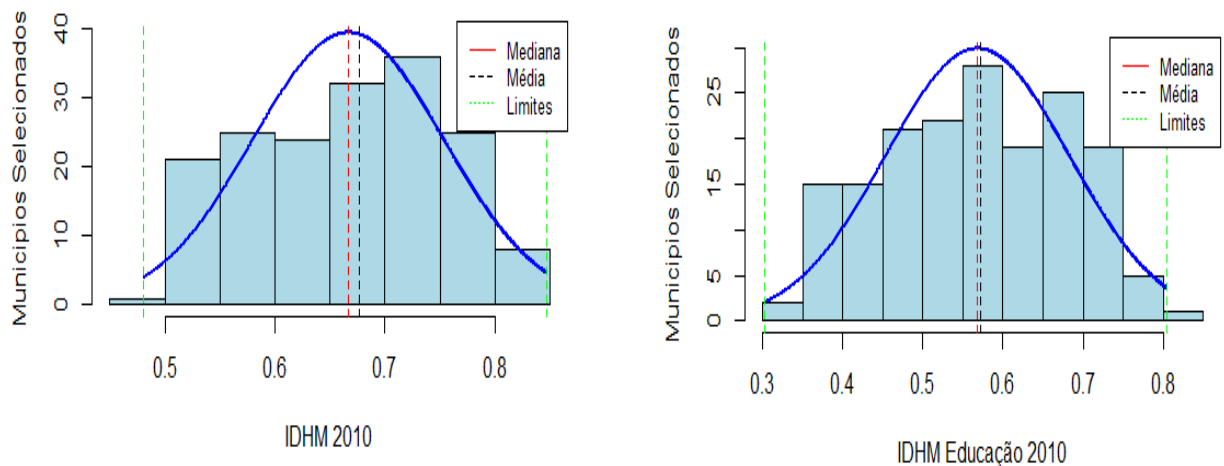
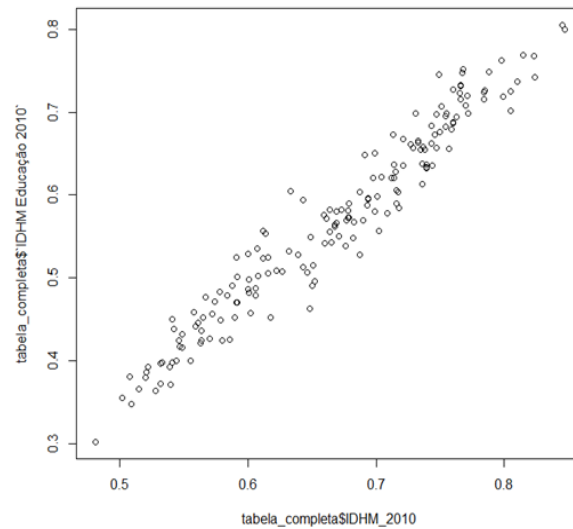


Figura 40 e 41 - Histograma IDHM e IDHM Educação municípios selecionados década anos 2010

```
plot(tabela_completa$IDHM_2010, tabela_completa$`IDHM Educação 2010`)
> cor.test(tabela_completa$IDHM_2010, tabela_completa$`IDHM Educação 2010`)
```

Pearson's product-moment correlation  
data: tabela\_completa\$IDHM\_2010 and tabela\_completa\$`IDHM Educação 2010`  
t = 53.594, df = 170, p-value < 2.2e-16  
alternative hypothesis: true correlation is not equal to 0  
95 percent confidence interval:  
0.961875 0.978959

sample estimates: cor 0.971659



*Figura 42 – Correlação IDHm x IDHM Educação década anos 2010*

```
> cor.test(tabela_completa$IDHM_2010, tabela_completa$`IDHM Educação 2010`,
method = "spearman")
```

Spearman's rank correlation rho

```
data: tabela_completa$IDHM_2010 and tabela_completa$`IDHM Educação 2010`
S = 22164, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.9738646
```

```
> cor.test(tabela_completa$IDHM_2010, tabela_completa$`IDHM Educação 2010`,
method = "kendall")
```

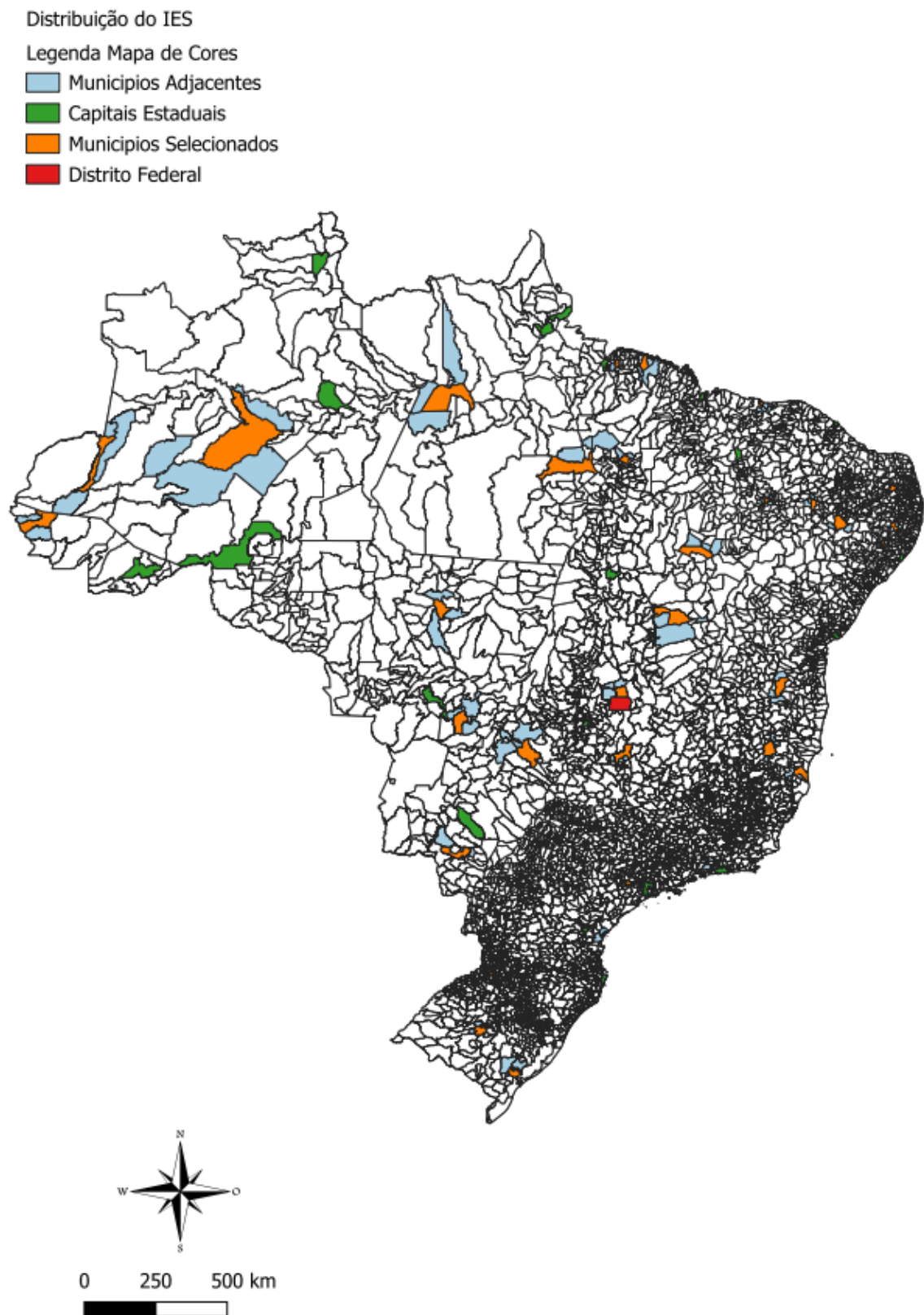
Kendall's rank correlation tau

```
data: tabela_completa$IDHM_2010 and tabela_completa$`IDHM Educação 2010`
z = 16.783, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.8638424
```



Dados Estatísticos apurados Municípios Selecionados						
Medidas de posição	Década anos 1990		Década anos 2000		Década anos 2010	
	IDHM	IDHM Ed.	IDHM	IDHM Ed.	IDHM	IDHM Ed.
Média	0,3553	0,2010	0,5258	0,3553	0,6668	0,5682
Mediana	0,3600	0,1875	0,5455	0,3600	0,6765	0,5720
Limite Inferior	0,0410	0,0170	0,2140	0,0410	0,4810	0,3020
Limite Superior	0,7000	0,5380	0,7660	0,7000	0,8470	0,8050
Medidas de dispersão	Década anos 1990		Década anos 2000		Década anos 2010	
	IDHM	IDHM Ed.	IDHM	IDHM Ed.	IDHM	IDHM Ed.
Desvio padrão	0,1227	0,1183	0,1211	0,1492	0,0870	0,1149
Variância	0,0151	0,0140	0,0147	0,0223	0,0076	0,0132
Dados Correlação						
Correlação	Década anos 1990		Década anos 2000		Década anos 2010	
	IDHM x IDHM Ed.		IDHM x IDHM Ed.		IDHM x IDHM Ed.	
Pearson	0,9738		0,9835		0,9716	
Spearman	0,9871		0,9862		0,9738	
Kendall	0,9059		0,9044		0,8638	

Tabela 4 – Dados estatísticos apurados Municípios selecionados



*Figura 43 - Mapa de Cores Municípios Brasileiros que receberam os IES e os Municípios Adjacentes*

## 4.5 Criação das amostras testes dos municípios escolhidos aleatoriamente

```
merge(dataframe_amostra, IDHM_1991, by = "municipios")
Error in fix.by(by.x, x) : 'by' must specify a uniquely valid column
> dataframe_amostra1 <- data.frame(Municipios = c(amostra))
> dataframe_amostra1
data.frame(id = c(amostra))
```

### Amostra 1

```
df <- sample(IDHM_1991a$id, size = 172, replace = FALSE)
> df
[1] "CAMPO MOURÃO" "SÍTIO NOVO DO TOCANTINS" "INÚBIA PAULI
STA" "COLNIZA"
[5] "ITAMARI" "Lagoa da Prata" "SANTA CRUZ D
O XINGU" "RIO DAS OSTRAS"
[9] "NOVA BRÉSCIA" "SEVERIANO MELO" "RODOLFO FERN
ANDES" "CAJARI"
[13] "Luminárias" "QUEDAS DO IGUAÇU" "Tiros"
"BOREBOREMA"
[17] "Andrelandia" "BELTERRA" "NAVEGANTES"
"ITAPEVA"
[21] "Cristais" "AFUÁ" "Campanha"
"SAO CARLOS"
[25] "COLOMBO" "JANIÓPOLIS" "São Lourenço"
"Peçanha"
[29] "PONTAL" "BARRA MANSA" "Galiléia"
"DESTERRO"
[33] "JOAQUIM TÁVORA" "ENTRE RIOS" "MAIRIPOTABA"
"Guidoval"
[37] "BATAGUASSU" "INDIAROA" "PEDRO AVELIN
O" "NOVA PRATA DO IGUAÇU"
[41] "IOMERÊ" "BARCELOS" "Itaúna"
"Iguatama"
```

Transformação da amostra em um data.table

```
df <- as.data.table(df)
> df
      df
1: CAMPO MOURÃO
2: SÍTIO NOVO DO TOCANTINS
3: INÚBIA PAULISTA
4: COLNIZA
5: ITAMARI
---
168: SOORETAMA
169: QUISSAMÃ
170: SOLIDÃO
171: APARECIDA DO RIO DOCE
172: SALDANHA MARINHO
```

Unificando os data.tables em uma única tabela associados pela coluna em com um "df" (Nome dos Municípios), resgatando para o dataframe da amostra os IDH M e IDHM's dos respectivos municípios.

```
k <- merge(df, IDHM_1991b, by = "df")
k
```

	df	Código do Município	IDHM	IDHM Educação
1:	ADOLFO	350020	0.513	0.296

2:	AFUÁ	150030	0.250	0.054
3:	ALDEIAS ALTAS	210030	0.257	0.078
4:	ALTANEIRA	230060	0.288	0.122
5:	AMORINÓPOLIS	520090	0.435	0.219
---				
168:	SÍTIO NOVO DO TOCANTINS	172080	0.302	0.134
169:	São Lourenço	316370	0.549	0.354
170:	São Sebastião do Oeste	316460	0.369	0.140
171:	TAPEROÁ	251650	0.285	0.123
172:	TAPEROÁ	293120	0.265	0.083

```
> summary(k)
```

df	Código do Município	IDHM	IDHM Educação
Length:172	Min. :110145	Min. :0.1310	Min. :0.0140
Class :character	1st Qu.:251249	1st Qu.:0.3078	1st Qu.:0.1140
Mode :character	Median :314703	Median :0.3845	Median :0.1690
	Mean :334579	Mean :0.3872	Mean :0.1837
	3rd Qu.:412120	3rd Qu.:0.4695	3rd Qu.:0.2425
	Max. :522020	Max. :0.6400	Max. :0.4760

```
> var(k$IDHM)
```

```
[1] 0.01039907
```

```
> sd(k$IDHM)
```

```
[1] 0.1019758
```

```
> var(k$`IDHM Educação`)
```

```
[1] 0.008136106
```

```
> sd(k$`IDHM Educação`)
```

```
[1] 0.09020037
```

```
### Gerando gráfico correlação IDHM e IDHM Educação Amostra 1 ###
```

```
plot(k$IDHM, k$`IDHM Educação`)
```

```
> > cor.test(k$IDHM, k$`IDHM Educação`)
```

```
Pearson's product-moment correlation
```

```
data: k$IDHM and k$`IDHM Educação`
```

```
t = 48.509, df = 170, p-value < 2.2e-16
```

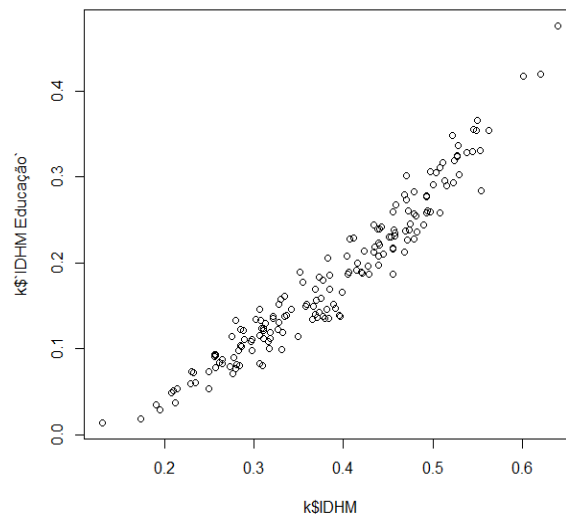
```
alternative hypothesis: true correlation is not equal to 0
```

```
95 percent confidence interval:
```

```
0.9539390 0.9745329
```

```
sample estimates:
```

```
cor 0.9657239
```



*Figura 44 – Gráfico correlação IDHM x IDHM Educação Amostra 1 dos municípios selecionados aleatoriamente*

```
cor.test(k$IDHM, k$`IDHM Educação`, method = "spearman")
```

Spearman's rank correlation rho

```
data: k$IDHM and k$`IDHM Educação`
S = 23388, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.9724208
```

O mesmo procedimento foi utilizado no dataframe dos IDHM de 2010 para selecionar amostra de municípios e aplicação de algoritmos no R para gerar a tabela com seus respectivos IDHM e IDHM Educação dos municípios amostrais.

```
cor.test(k$IDHM, k$`IDHM Educação`, method = "kendall")
```

Kendall's rank correlation tau

```
data: k$IDHM and k$`IDHM Educação`
z = 16.704, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.8602217
```

## Amostra 2

```
> summary(w)
```

df1	IDHM 2010	IDHM	IDHM Educação 2010
Length:172	Length:172	Min. :0.4500	Min. :0.2590
Class :character	Class :character	1st Qu.:0.5965	1st Qu.:0.4780
Mode :character	Mode :character	Median :0.6535	Median :0.5500
		Mean :0.6540	Mean :0.5510
		3rd Qu.:0.7105	3rd Qu.:0.6282
		Max. :0.8090	Max. :0.7490

```
> var(w$IDHM)
```

```
[1] 0.005322484
```

```
> var(w$`IDHM Educação 2010`)
```

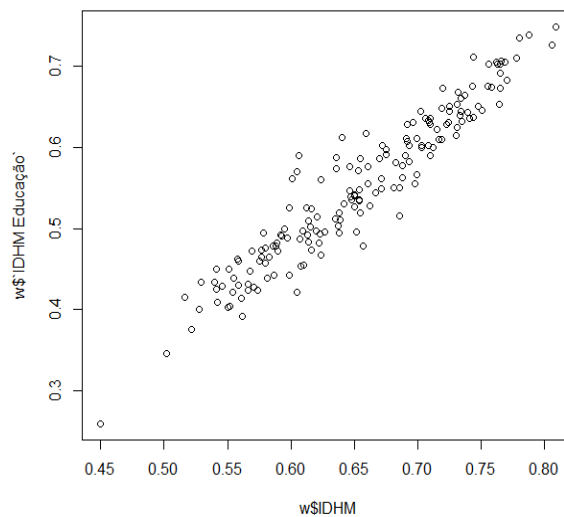
```
[1] 0.008721132
```

```
> sd(w$IDHM)
```

```
[1] 0.07295535
```

```
> sd(w$`IDHM Educação 2010`)
```

```
[1] 0.093387
```



*Figura 45 – Gráfico correlação IDHM x IDHM Educação Amostra 2 dos municípios selecionados aleatoriamente*

```
> cor.test(w$IDHM, w$`IDHM Educação`)
```

Pearson's product-moment correlation

data: w\$IDHM and w\$`IDHM Educação`

t = 41.513, df = 170, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.9383761 0.9658065

sample estimates:

cor  
0.9540492

```
cor.test(w$IDHM, w$`IDHM Educação`, method = "spearman")
```

Spearman's rank correlation rho

data: w\$IDHM and w\$`IDHM Educação`  
S = 40191, p-value < 2.2e-16  
alternative hypothesis: true rho is not equal to 0  
sample estimates:  
rho  
0.9526079

```
cor.test(w$IDHM, w$`IDHM Educação`, method = "kendall")
```

Kendall's rank correlation tau

data: w\$IDHM and w\$`IDHM Educação`  
z = 15.959, p-value < 2.2e-16  
alternative hypothesis: true tau is not equal to 0  
sample estimates: tau 0.8217805

Medidas de posição	Amostra 1		Amostra 2	
	IDHM	IDHM Ed.	IDHM	IDHM Ed.
Média	0.3872	0.1837	0.6540	0.5510
Mediana	0.3845	0.1690	0.6535	0.5500
Limite Inferior	0.1310	0.0140	0.4500	0.2590
Limite Superior	0.6400	0.4760	0.8090	0.7490
Medidas de dispersão	Década anos 1990		Década anos 2010	
Desvio padrão	0.1019	0.0902	0.0729	0.0933
Variância	0.0104	0.0081	0.0053	0.0087
Correlação	Década anos 1990		Década anos 2010	
	IDHM x IDHM Ed.		IDHM x IDHM Ed.	
Pearson	0.9657		0.9540	
Spearman	0.9724		0.9526	
Kendall	0.8602		0.8217	

Tabela 5 – Dados estatísticos apurados das amostra 1 e 2

## 5 CRIAÇÃO DO MODELO DE MACHINE LEARNING

De acordo com os dados estatísticos apurados previamente somado ao proposito deste trabalho o modelo de *Machine learning* compreendido como mais adequado para prosseguir com este trabalho foi o de regressão linear simples supervisionado que reconhecem os dados históricos e replicam esses valores conhecidos para uma aproximação futura não sendo descartado outras alternativas de aprendizado de máquina.

Antecipadamente os *Dataset's* foram manipulados de modo que fosse obtido a média entre as variáveis dependentes (IDHM 1990, 2000 e 2010) e a variável independente IDHM Educação 1990, 2000 e 2010).

### 5.1 Modelo Regressão Linear para os Municípios Brasileiros

**# Passo 1 – Baixar pacotes necessário no R para a criação do modelo. Estes pacotes serão utilizados em todos os dataset`s a serem modelados**

```
library(car); library(dplyr); library(ggpubr); library(ggplot2);  
library(lmtest); library(rstatix), library(ggpmisc)
```

Uma breve explicação sobre cada pacote utilizado no R para criação do modelo de regressão linear.

dplyr	O pacote dplyr busca oferecer um conjunto de “verbos” (i.e., funções) voltados para as operações mais comumente aplicadas em tabelas. Ou seja, as funções desse pacote em geral aceitam um dataframe como input, e retornam um novo dataframe como output. Dito de forma menos técnica, você fornece uma tabela para essas funções, e elas lhe retornam como resultado uma nova tabela.
Car (Companion to Applied Regression)	Este pacote não é usado para realizar técnicas de Regressão Aplicada, ele complementa essas técnicas fornecendo inúmeras funções que realizam testes, criam visualizações e transformam dados. Para verificar a validade de inúmeras técnicas de regressão, precisamos realizar vários testes em nossos resultados. Este pacote fornece as ferramentas necessárias para isso.
Lmtest (Testing Linear Regression Models)	Uma coleção de testes, conjuntos de dados e exemplos para verificação de diagnóstico em modelos de regressão linear. Além disso, são fornecidas algumas ferramentas genéricas para inferência em modelos paramétricos.



ggplot 2	O ggplot2 é um pacote R para produção de gráficos que diferentemente da maioria dos outros pacotes, apresenta uma profunda gramática baseada no livro The grammar of graphics (Wilkinson 2005). Os gráficos originados em ggplot2 são baseados em camadas, e cada gráfico tem três componentes chave: data, os dados de onde o gráfico será criado; aes() (aesthetic mappings), que controla o mapeamento estético e as propriedades visuais do gráfico; e ao menos uma camada que irá descrever como cada observação será renderizada. Camadas são usualmente criadas utilizando uma função geom_(). A referência principal ao pacote é o livro Ggplot2 : elegant graphics for data analysis (Wickham 2009).
Rstatix (Pipe-Friendly Framework for Basic Statistical Tests)	Fornece uma estrutura simples e intuitiva, coerente com a filosofia de design 'tidyverse', para realizar testes estatísticos básicos, incluindo teste t, teste de Wilcoxon, ANOVA, Kruskal-Wallis e análises de correlação. A saída de cada teste é transformada automaticamente em um quadro de dados organizado para facilitar a visualização. Funções adicionais estão disponíveis para remodelar, reordenar, manipular e visualizar a matriz de correlação. Funções também estão incluídas para facilitar a análise de experimentos fatoriais, incluindo designs puramente 'dentro-Ss' (medidas repetidas), designs puramente 'entre-Ss' e designs mistos 'dentro e entre-Ss'. Também é possível calcular várias métricas de tamanho de efeito, incluindo "eta quadrado" para ANOVA, "d de Cohen" para teste t e 'Cramer V' para a associação entre variáveis categóricas. O pacote contém funções auxiliares para identificar outliers univariados e multivariados, avaliando a normalidade e a homogeneidade das variâncias.
Ggpubr (Based Publication Ready Plots)	O pacote fornece algumas funções fáceis de usar para criar e personalizar gráficos prontos para publicação baseados em 'ggplot2'.

Tabela 6 – Pacotes baixados no R-Studio para criação do modelo de machine learning

**# Passo 2 – Carregar arquivos do Banco de dados**

```

> library(readxl)
> IDH_br_1 <- read_excel("IDH_br_1.xlsx")
> View(IDH_br_1)

```

### # Passo 3 – Verificação dos Pressupostos para Regressão Linear

## Regressão Linear entre a Variável Dependente (VD) e a Variável Independente (VI)

### VD = Média IDHM

### VI = Média IDHM Educação

### ## Construção do Modelo

```
mod1 <- lm (IDH_br_1$`Média IDHM` ~ IDH_br_1$`Média IDHM Educação`)
par(mfrow=c(2,2))>
> plot(mod1)
```

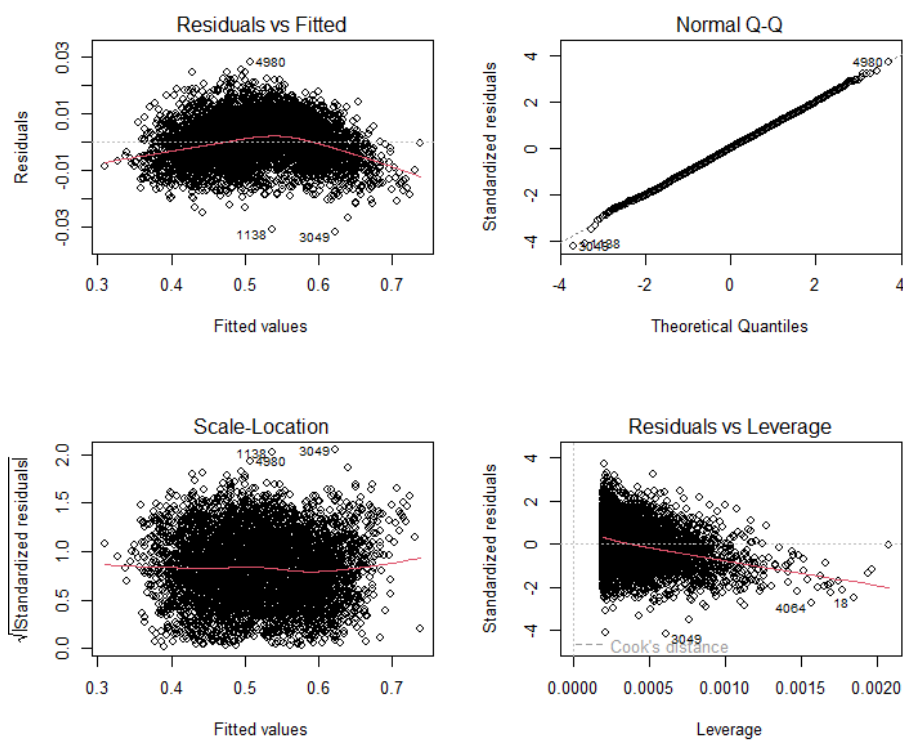


Figura 46 – Gráficos estatísticos dos pressupostos validação do modelo de aprendizado de máquina municípios brasileiros

### ## Normalidade dos Resíduos

```
shapiro.test(mod1$residuals)
```

shapiro-wilk normality test

data: mod1\$residuals

w = 0.99968, p-value = 0.6419

### ## Outliers nos Resíduos

```
> summary(rstandard(mod1))
```

Min.	1st Qu.	Median	Mean	3rd Qu.

```
-4.183626 -0.680815 0.005218 -0.000034 0.683159
      Max.
      3.720780
```

### ## Independência dos Resíduos

```
> durbinwatsonTest(mod1)
lag Autocorrelation D-W Statistic p-value
1 0.08041921 1.838028 0
Alternative hypothesis: rho != 0
```

### ## Homocedasticidade

```
bptest(mod1)
```

studentized Breusch-Pagan test

```
data: mod1
BP = 0.95916, df = 1, p-value = 0.3274
```

### # Passo 4 – Análise do Modelo

```
> summary(mod1)
```

```
Call:
lm(formula = IDH_br_1$`Média IDHM` ~ IDH_br_1$`Média IDHM Educação`)
```

```
Residuals:
      Min       1Q   Median       3Q      Max
-0.031956 -0.005201  0.000040  0.005219  0.028426
```

```
Coefficients:
              Estimate Std. Error
(Intercept)  0.0802453  0.0006769
IDH_br_1$`Média IDHM Educação` 0.9718531  0.0014748
              t value Pr(>|t|)
(Intercept)    118.5   <2e-16 ***
IDH_br_1$`Média IDHM Educação`  659.0   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.007641 on 4986 degrees of freedom
Multiple R-squared: 0.9886, Adjusted R-squared: 0.9886
F-statistic: 4.343e+05 on 1 and 4986 DF, p-value: < 2.2e-16
```

### # Passo 5 – Gráfico de Dispersão

```
ggplot(data = IDH_br_1, mapping = aes(x = IDH_br_1$`Média IDHM Educação`, y
= IDH_br_1$`Média IDHM`)) + geom_point() + geom_smooth(method = "lm", col =
"blue") + stat_poly_eq(aes(label = paste(..eq.label.., ..rr.label.., sep =
"*plain(\"\", \"\")~~\"")) + theme_classic()
```

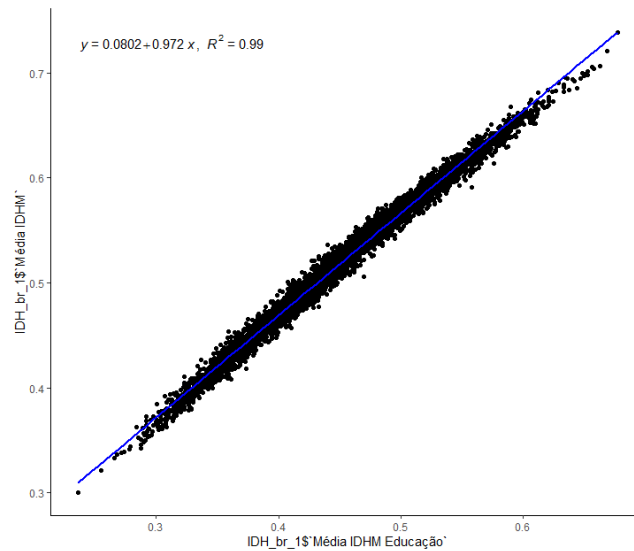


Figura 47 – Gráfico de dispersão dos municípios brasileiros

## 5.2 Modelo Regressão Linear para os Municípios Selecionados

### # Passo 2 – Carregar arquivos do Banco de dados

```
> library(readxl)
> Municipios_Selecionados <- read_excel("Municipios selecionados.xlsx")
> view(Municipios_Selecionados)
```

### # Passo 3 – Verificação dos Pressupostos para Regressão Linear

## Regressão Linear entre a Variável Dependente (VD) e a Variável Independente (VI)

### VD = Média IDHM

### VI = Média IDHM Educação

### ## - Construção do Modelo

```
mod2 <- lm(Municipios_Selecionados$`Média IDHM` ~ Municipios_Selecionados$`
Média IDHM Educação`)
par(mfrow=c(2,2))>
> plot(mod2)
```

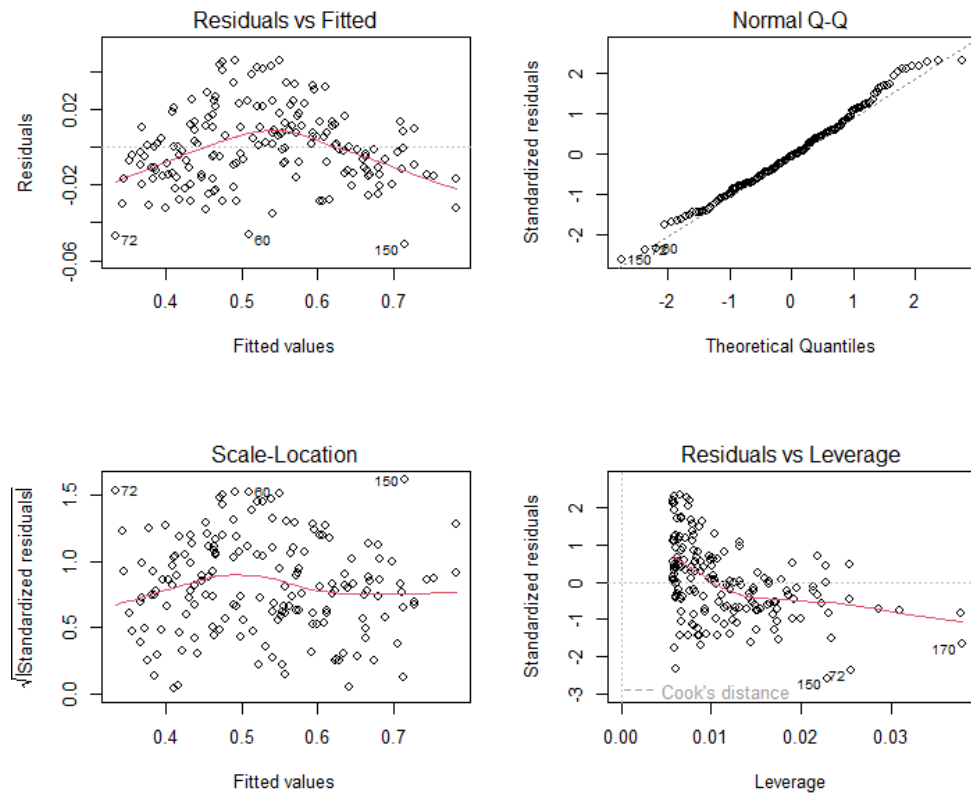


Figura 48 - Gráficos estatísticos dos pressupostos validação do modelo de aprendizado de máquina municípios selecionados

### ## Normalidade dos Resíduos

```
shapiro.test(mod2$residuals)
```

Shapiro-wilk normality test

```
data: mod2$residuals
W = 0.98911, p-value = 0.2091
```

### ## Outliers nos Resíduos

```
summary(rstandard(mod2))
```

	Min.	1st Qu.	Median	Mean	3rd Qu.
	-2.599423	-0.731473	-0.054807	-0.001338	0.586420
Max.	2.321612				

### ## Independência dos Resíduos

```
> durbinWatsonTest(mod2)
```

lag	Autocorrelation	D-W Statistic	p-value
1	-0.05844743	2.110497	0.454

Alternative hypothesis: rho != 0

## ## Homocedasticidade

```
bptest(mod2)
```

```
studentized Breusch-Pagan test
```

```
data: mod2  
BP = 0.60166, df = 1, p-value = 0.4379
```

## # Passo 4 – Análise do Modelo

```
> summary(mod2)
```

```
Call:  
lm(formula = Municipios_Selecionados$`Média IDHM` ~ Municipios_Selecionados  
$`Média IDHM Educação`)
```

```
Residuals:  
      Min       1Q   Median       3Q      Max  
-0.051484 -0.014547 -0.001093  0.011703  0.046366
```

```
Coefficients:
```

```
(Intercept)                Estimate  
Municipios_Selecionados$`Média IDHM Educação` 0.208646  
                                0.859701  
                                Std. Error  
(Intercept)                0.004858  
Municipios_Selecionados$`Média IDHM Educação` 0.012304  
                                t value  
(Intercept)                42.95  
Municipios_Selecionados$`Média IDHM Educação` 69.87  
                                Pr(>|t|)  
(Intercept)                <2e-16 ***  
Municipios_Selecionados$`Média IDHM Educação` <2e-16 ***  
---
```

```
Signif. codes:  
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.02004 on 170 degrees of freedom  
Multiple R-squared: 0.9663, Adjusted R-squared: 0.9662  
F-statistic: 4882 on 1 and 170 DF, p-value: < 2.2e-16
```

## # Passo 5 – Gráfico de Dispersão

```
ggplot(data = Municipios_Selecionados, mapping = aes(x = Municipios_Selecionados$`Média IDHM Educação`, y = Municipios_Selecionados$`Média IDHM`)) +  
geom_point() + geom_smooth(method = "lm", col = "red") + stat_poly_eq(aes(label = paste(..eq.label.., ..rr.label.., sep = "*plain(\\",\\")~~")))) + theme_
```

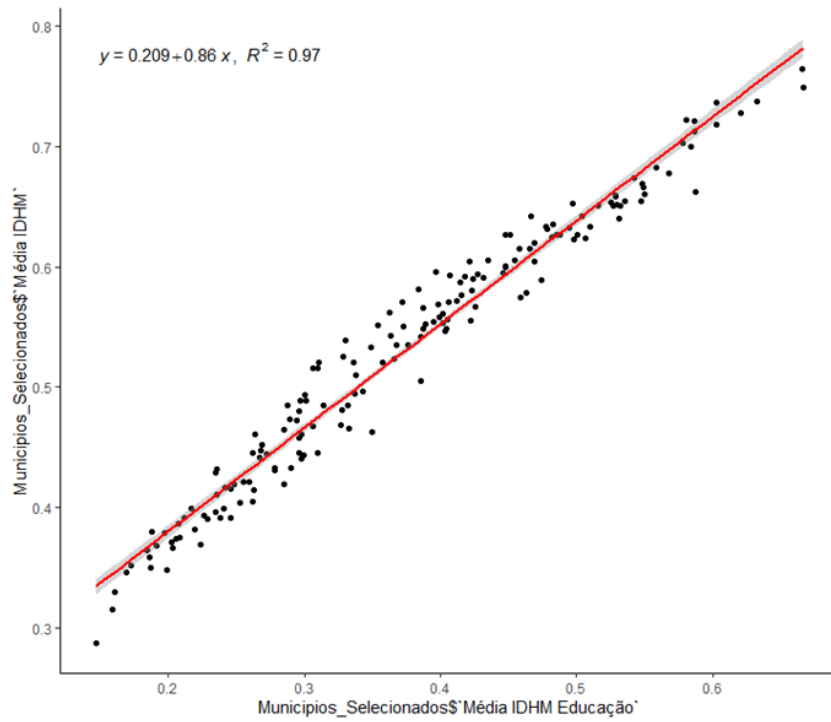


Figura 49 - Gráfico de dispersão dos municípios selecionados

### 5.3 Modelo Regressão Linear para a Amostra 1

#### # Passo 2 – Carregar arquivos do Banco de dados

```
> library(readxl)
> Amostra_1 <- read_excel("Amostra_1.xlsx")
> View(Amostra_1)
```

#### # Passo 3 – Verificação dos Pressupostos para Regressão Linear Amostra 1

## Regressão Linear entre a Variável Dependente (VD) e a Variável Independente (VI)

### VD = Média IDHM Amostra 1

### VI = Média IDHM Educação Amostra 1

#### ## - Construção do Modelo Amostra 1

```
> mod3 <- lm(Amostra_1$`Média IDHM Amostra 1` ~ Amostra_1$`Média IDHM Educ A
mostra 1`)
> par(mfrow=c(2,2))
> plot(mod3)
```

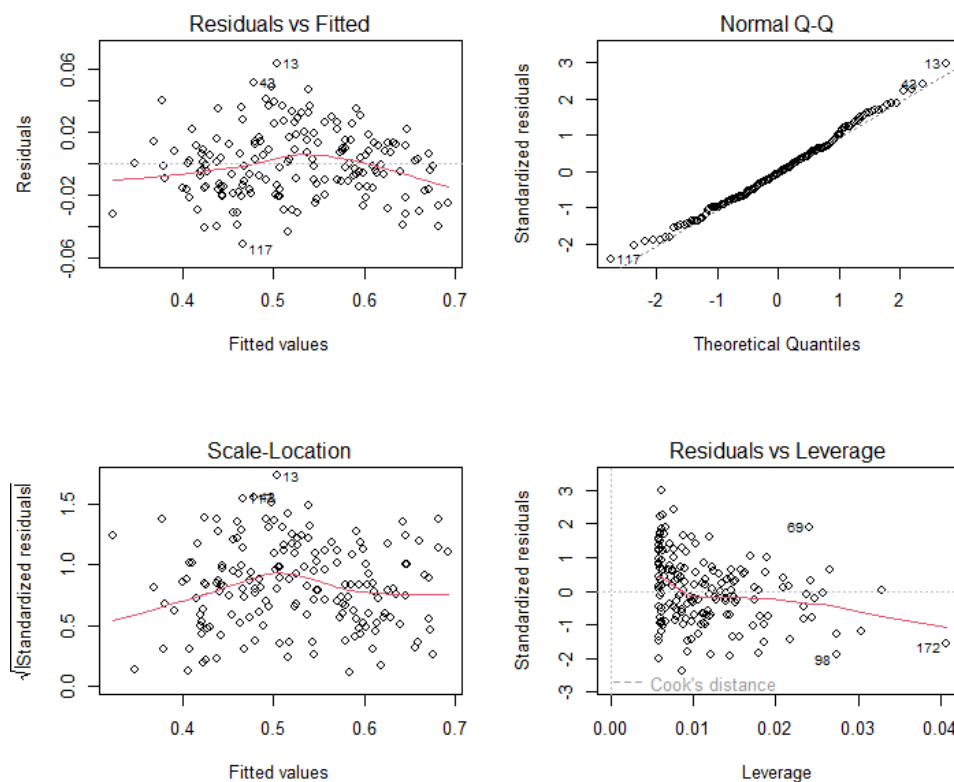


Figura 50 - Gráficos estatísticos dos pressupostos validação do modelo de aprendizado de máquina Amostra 1



### ## Normalidade dos Resíduos Amostra 1

```
shapiro.test(mod3$residuals)
```

Shapiro-wilk normality test

```
data: mod3$residuals
W = 0.99288, p-value = 0.5642
```

,

### ## Outliers nos Resíduos Amostra 1

```
> summary(rstandard(mod3))
```

Min.	1st Qu.	Median	Mean	3rd Qu.
-2.3998704	-0.7142281	-0.0565669	-0.0007884	0.6275097
Max.				
2.9922226				

### ## Independência dos Resíduos

```
> durbinwatsonTest(mod3)
```

lag	Autocorrelation	D-W Statistic	p-value
1	-0.01421589	2.012817	0.956

Alternative hypothesis: rho != 0

### ## Homocedasticidade Amostra 1

```
> bptest(mod3)
```

studentized Breusch-Pagan test

```
data: mod3
BP = 0.71676, df = 1, p-value = 0.3972
```

### # Passo 4 – Análise do Modelo Amostra 1

```
> summary(mod3)
```

Call:

```
lm(formula = Amostra_1$`Média IDHM Amostra 1` ~ Amostra_1$`Média IDHM Educ Amostra 1`)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.051499	-0.015295	-0.001213	0.013465	0.064292

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.199499	0.006685	29.84	<2e-16 ***
Amostra_1\$`Média IDHM Educ Amostra 1`	0.885937	0.017691	50.08	<2e-16 ***

---

Signif. codes:

0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02155 on 170 degrees of freedom  
Multiple R-squared: 0.9365, Adjusted R-squared: 0.9361  
F-statistic: 2508 on 1 and 170 DF, p-value: < 2.2e-16

### # Passo 5 – Gráfico de Dispersão Amostra 1

```
ggplot(data = Amostra_1, mapping = aes(x = Amostra_1$`Média IDHM Educ Amostra 1`, y = Amostra_1$`Média IDHM Amostra 1`)) + geom_point() + geom_smooth(
method = "lm", col = "red") + stat_poly_eq(aes(label = paste(..eq.label..,
..rr.label.., sep = "*plain(\",\"")~~")))) + theme_classic()
```

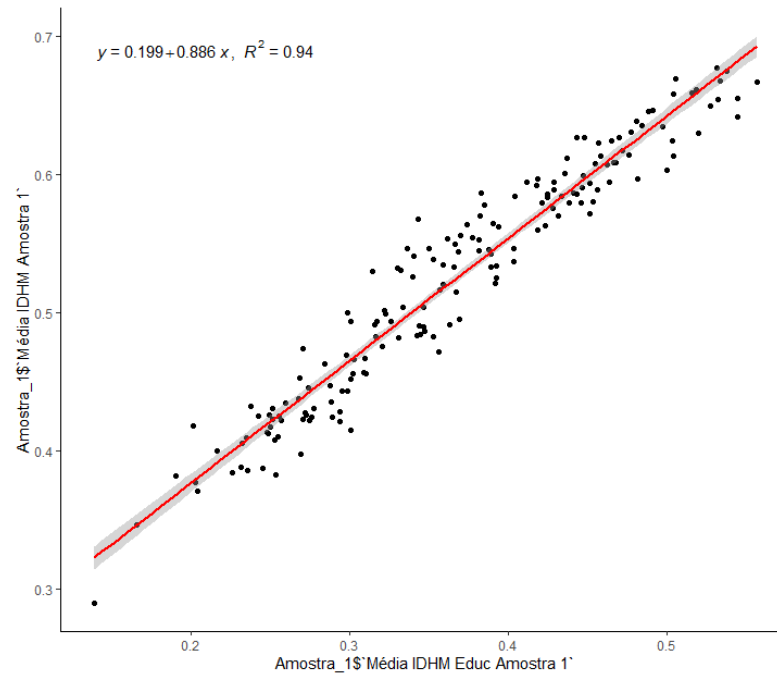


Figura 51 – Gráfico de dispersão Amostra 1

## 5.4 Modelo Regressão Linear para a Amostra 2

### # Passo 2 – Carregar arquivos do Banco de dados

```
> library(readxl)
> Amostra_2 <- read_excel("Amostra_2.xlsx")
> view(Amostra_2)
```

### # Passo 3 – Verificação dos Pressupostos para Regressão Linear Amostra 2

```
## Regressão Linear entre a Variável Dependente (VD) e a Variável Independente (VI)
### VD = Média IDHM Amostra 2
### VI = Média IDHM Educação Amostra 2
```

### ## - Construção do Modelo amostra 2

```
> mod4 <- lm(Amostra_2$`Média IDHM Amostra 2` ~ Amostra_2$`Média IDHM Educ A
mostra 2`)
> par(mfrow=c(2,2))
> plot(mod4)
```

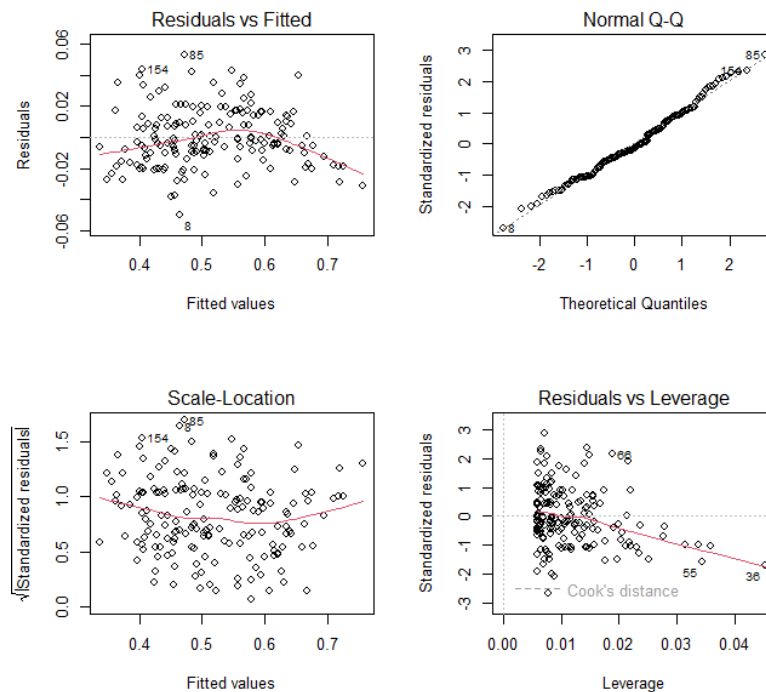


Figura 52 - Gráficos estatísticos dos pressupostos validação do modelo de aprendizado de máquina Amostra 2

### ## Normalidade dos Resíduos amostra 2

```
> shapiro.test(mod4$residuals)
```

shapiro-wilk normality test

```
data: mod4$residuals
W = 0.98894, p-value = 0.1991
```

### ## Outliers nos Resíduos amostra 2

```
> summary(rstandard(mod4))
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-2.686313	-0.633988	-0.140909	-0.000885	0.709884	2.866875

### ## Independência dos Resíduos amostra 2

```
> durbinwatsonTest(mod4)
```

lag	Autocorrelation	D-W	Statistic	p-value
1	0.01046727	1.978053	0.968	

Alternative hypothesis: rho != 0

### ## Homocedasticidade amostra 2

```
> bptest(mod4)
```

studentized Breusch-Pagan test

```
data: mod4
BP = 1.6955, df = 1, p-value = 0.1929
```

#### # Passo 4 – Análise do Modelo amostra 2

```
> summary(mod4)
```

```
Call:
lm(formula = Amostra_2$`Média IDHM Amostra 2` ~ Amostra_2$`Média IDHM Educ
Amostra 2`)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-0.049976 -0.011771 -0.002622  0.013156  0.053350
```

```
Coefficients:
```

	Estimate	Std. Error	t value
(Intercept)	0.195639	0.005179	37.77
Amostra_2\$`Média IDHM Educ Amostra 2`	0.893699	0.013865	64.46
	Pr(> t )		
(Intercept)	<2e-16	***	
Amostra_2\$`Média IDHM Educ Amostra 2`	<2e-16	***	

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.01868 on 170 degrees of freedom
Multiple R-squared:  0.9607, Adjusted R-squared:  0.9605
F-statistic: 4155 on 1 and 170 DF, p-value: < 2.2e-16
```

#### # Passo 5 – Gráfico de Dispersão Amostra 2

```
ggplot(data = Amostra_2, mapping = aes(x = Amostra_2$`Média IDHM Educ Amost
ra 2`, y = Amostra_2$`Média IDHM Amostra 2`)) + geom_point() + geom_smooth(
method = "lm", col = "red") + stat_poly_eq(aes(label = paste(..eq.label..,
..rr.label.., sep = "*plain(\",\")~~")))) + theme_classic()
```

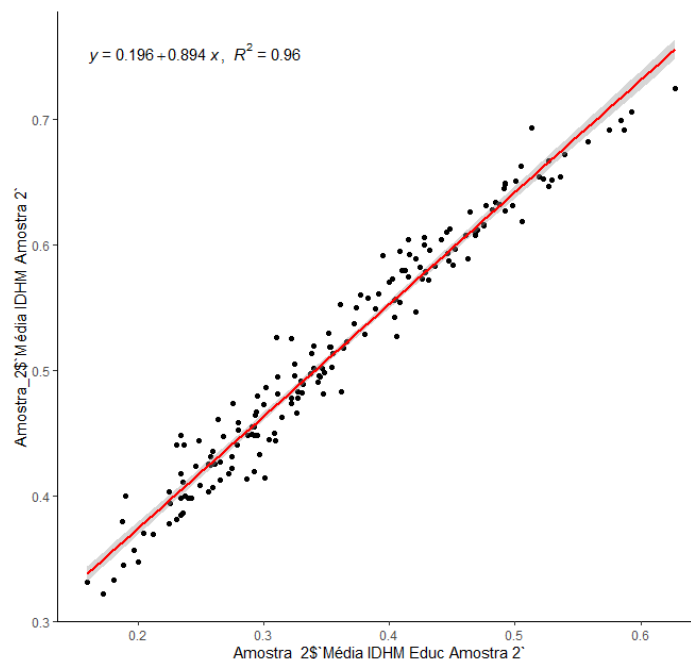


Figura 53 – Gráfico de dispersão Amostra 2

## 6 INTERPRETAÇÃO DOS RESULTADOS

Validação dos Pressupostos para considerar o modelo de aprendizado de máquina para regressão linear

Os resultados obtidos no modelo de *Machine learning* de regressão linear simples pode se inferir sobre os mesmos para concluir se o problema proposto pode ser aceito ou rejeitado.

Ho: O ensino superior descentralizado influencia no aumento do IDHM Educação, variável independente que por sua vez influencia no aumento do IDHM mais rapidamente que os municípios que não possuem polos de IES descentralizados

H1: O ensino superior descentralizado tem pouca ou nenhuma influência sobre o desenvolvimento IDHM Educação que por sua vez não influencia no IDHM.

A partir dos passos criados no desenvolvimento deste modelo, primeiramente verificou se os pressupostos para aplicação deste modelo baseado na regressão linear simples estão de acordo com o perfil dos dados analisados.

### 6.1 Shapiro-Wilk-Test

Teste Shapiro-wilk que consiste em avaliar se uma distribuição é semelhante a uma distribuição normal.

Como resultado, o teste retornará a estatística W, que terá um valor de significância associada, o valor-p. Para dizer que uma distribuição é normal, o valor p precisa ser maior do que 0,05.

Resultados Obtidos para teste de distribuição normal dos dados das amostras

Municípios Brasileiros	W = 0.99968	p-value = 0.6419
Municípios Selecionados	W = 0.98911	p-value = 0.2091
Amostra 1	W = 0.99288	p-value = 0.5642
Amostra 2	W = 0.98894	p-value = 0.1991

Tabela 7 – Dados apurados no Teste Shapiro-Wilk

De acordo com a tabela nesta etapa do pressuposto da normalidade de todos os quatro dataset obtiveram valor de  $p \geq 0,05$  que permite concluir que as distribuições dos dados são normais fato que já havia sido constatado na análise e exploração dos dados.

## 6.2 Normalidade dos outliers residuais

Um segundo teste para verificação da normalidade de outliers residuais, `summary(rstandard)`. Uma das suposições para o teste de hipóteses é que os erros seguem uma distribuição gaussiana. Como consequência, os resíduos também deveriam. As estatísticas de resumo residual fornecem informações sobre a simetria da distribuição residual. A mediana deve estar próxima de 0, pois a média dos resíduos é 0 e as distribuições simétricas têm mediana=média. Além disso, o 3Q e o 1Q devem estar próximos um do outro em magnitude. Eles seriam iguais sob uma distribuição de média 0 simétrica. O máximo e o mínimo também devem ter magnitude semelhante. No entanto, neste caso, não manter pode indicar um outlier em vez de uma violação de simetria

	Min	1st Qu.	Median	Mean	3rd Qu	Max.
Municípios Brasileiros	-2.599423	-0.731473	-0.054807	-0.001338	0.586420	2.321612
Municípios Seleccionados	-2.599423	-0.731473	-0.054807	-0.001338	0.586420	2.321612
Amostra 1	-2.399870	-0.714228	-0.056566	-0.000788	0.627509	2.992222
Amostra 2	-2.686313	-0.633988	0.140909	0.000885	0.709884	2.866875

Tabela 8 – Dados apurados Normalidade dos Outliers Residuais

Os dados apurados nesta etapa também indicam um comportamento padrão nos outliers das amostras neste caso podemos considerar retirar ou não os outliers do dataset. Optou-se por mantê-los já que sua presença não impactará no resultado do modelo em questão.

## 6.3 Teste de homocedasticidade

Em análise de variância (ANOVA), há um pressuposto que deve ser atendido que é de os erros terem variância comum, ou seja, homocedasticidade. Isso implica que cada tratamento que se está sendo comparado pelo teste F, deve ter aproximadamente a mesma variância para que a ANOVA tenha validade. Quando este pressuposto não é atendido dizemos que as variâncias não são homogêneas, ou ainda, que existe heterocedasticidade.

H<sub>0</sub> (hipótese nula): Há homocedasticidade.  $p > 0,05$

H<sub>A</sub> (hipótese alternativa): Não há homocedasticidade.  $P \leq 0,05$

	BP	p-value
Municípios Brasileiros	0.95916	0.3274
Municípios Selecionados	0.60166	0.4379
Amostra 1	0.71676	0.3972
Amostra 2	1.6955	0.1929

Tabela 9 – Dados apurados no teste de homocedasticidade

A estatística de teste dos dataset's o valor p em todos os casos supera 0,05 neste caso não rejeita se a hipótese nula. Não temos evidências suficientes para dizer que a heterocedasticidade está presente no modelo de regressão.

#### 6.4 Teste Durbin-Watson

Uma das principais suposições na regressão linear é que não há correlação entre os resíduos, por ex. os resíduos são independentes. Uma maneira de determinar se essa suposição é atendida é realizar um teste de Durbin-Watson, que é usado para detectar a presença de autocorrelação nos resíduos de uma regressão. Este teste usa as seguintes hipóteses:

H0 (hipótese nula): Não há correlação entre os resíduos.  $p > 0,05$

HA (hipótese alternativa): Os resíduos são auto correlacionados.

	Autocorrelation	D-W Statistic	p-value
Municípios Brasileiros	0.08041921	1.838028	0
Municípios Selecionados	-0.05844743	2.110497	0.454
Amostra 1	-0.01421589	2.012817	0.956
Amostra 2	0.01046727	1.978053	0.968

Tabela 10 – Dados apurados no teste de Durbin-Watson

O que fazer se a autocorrelação for detectada:

Se você rejeitar a hipótese nula e concluir que a autocorrelação está presente nos resíduos, então você tem algumas opções diferentes para corrigir esse problema, se considerar que é sério o suficiente:

Para correlação serial positiva, considere adicionar defasagens da variável dependente e/ou independente ao modelo.

Para correlação serial negativa, verifique se nenhuma de suas variáveis está super diferenciada.

Para correlação sazonal, considere adicionar variáveis fictícias sazonais ao modelo.

Para o nosso modelo somente não rejeitamos a hipótese nula no caso dos municípios brasileiros os demais o valor de p foi bem superior a ao valor  $p = 0,05$  indicando que devemos rejeitar a hipótese nula e assumir que os resíduos desse modelo são auto correlacionados. Neste caso optou se por prosseguir com o modelo de aprendizado de máquina sem interferir na base de dados e para uma futura ampliação desta pesquisa considere adicionar defasagens da variável dependente e/ou independente ao modelo., uma vez que os outros testes dos pressupostos atenderam o objetivo da pesquisa.

H0 (hipótese nula): A variável independente não tem impacto sobre a variável dependente Coeficiente = 0  $p > 0,05$

HA (hipótese alternativa): A variável independente tem impacto sobre a variável dependente Coeficiente  $\neq 0$   $p \leq 0,05$

A cada 1 valor aumentado na variável independente aumenta o valor Estimado apurado no valor da variável dependente.

	Estimate	Std. Error	t value	Pr(> t )
Municípios Brasileiros	0.9718531	0.0014748	659.0	<2e-16 ***
Municípios Selecionados	0.859701	0.012304	69.87	<2e-16 ***
Amostra 1	0.885937	0.017691	29.84	<2e-16 ***
Amostra 2	0.893699	0.013865	64.46	<2e-16 ***

**Tabela 11 – Teste de Hipótese para interpretação dos resultados**

De acordo com os dados apurados no modelo pode se inferir que os dados do dataset por apresentar valor  $p < 0.05$  devemos rejeitar a  $H_0$  e assumir que a variável independente tem impacto sim sobre a variável dependente portanto o modelo proposto atende as requisições de que os resultados da análise do modelo estão coerentes com o fato de quanto mais o IDHM Educação cresce mais o IDHM também cresce proporcionalmente ao valor estimado apurado no modelo.

Quanto da variável independente explica o aumento percentual da variável dependente



	R-squared
Municípios Brasileiros	0.9886
Municípios Seleccionados	0.9663
Amostra 1	0.9365
Amostra 2	0.9607

Tabela 12 – Teste R- Squared

Pelos valores apurados do R-squared percebe se que a variável independente tem grande influência sobre o comportamento da variável dependente como já era esperado.

Comparação entre um suposto modelo nulo que se baseia pela média da evolução do IDHM sem leva em conta o fator da variável independente comparado com o modelo aqui proposto.

H0 (hipótese nula): o modelo criado prevê tão bem quanto o modelo nulo  $p > 0,05$

HA (hipótese alternativa): Existe diferença entre o modelo criado e o modelo nulo (sendo o modelo proposto melhor que o nulo)  $p \leq 0,05$

	F-statistic	Graus de Liberdade	Graus de Liberdade	Pr(> t )
Municípios Brasileiros	4.343e+05	1	4986	<2.2e-16 ***
Municípios Seleccionados	4882	1	170	<2.2e-16 ***
Amostra 1	4155	1	170	<2.2e-16 ***
Amostra 2	4155	1	170	<2.2e-16 ***

Tabela 13 – Teste de hipótese F- Statistic

Todos os valores de p foram bem menores que 0.05 para o F-statistic portanto devemos rejeitar a hipótese nula e considerar que o modelo proposto é melhor para prever os resultados que um suposto modelo nulo que se baseia apenas na média temporal.

## 7 APRESENTAÇÃO DOS RESULTADOS

Através dos resultados apurados no modelo consegue se inferir se o problema proposto:

Ho: O ensino superior descentralizado influencia no aumento do IDHM Educação, variável independente que por sua vez influencia no aumento do IDHM mais rapidamente que os municípios que não possuem polos de IES descentralizados

H1: O ensino superior descentralizado tem pouca ou nenhuma influência sobre o desenvolvimento IDHM Educação que por sua vez não influencia no IDHM.

Os resultados apurados em todo seu contexto nos remetem a rejeitar a hipótese nula e aceitar a hipótese alternativa ou seja: O ensino superior descentralizado tem pouca ou nenhuma influência sobre o desenvolvimento IDHM Educação que por sua vez têm pouca influência sobre o IDHM no desenvolvimento humano dos municípios selecionados. A previsão do modelo apontou para uma tendência de avanços dos IDHM bem igualitária, sem discrepâncias significativas entre as amostras os municípios selecionados e a totalidade dos municípios brasileiros.

### 7.1 QUESTIONAMENTO SOBRE O RESULTADO OBTIDO

Deve se destacar que este tema é muito complexo e dinâmico que merece uma análise mais aprofundada para de fato entender o que pode ter gerado o resultado obtido. Alguns questionamentos foram levantados, porém sem nenhum embasamento de estudo científico que poderia justificar o resultado alcançado apenas em caráter especulativo foi levantado as seguintes possibilidades para o resultado do problema proposto a saber:

- MODELO DE ANÁLISE SIMPLES DEMAIS PARA O REFERIDO ESTUDO
- POLÍTICAS DE DESCENTRALIZAÇÃO DOS IES INEFICIENTES
- OS OUTROS DOIS PILARES DO IDH (RENDIMENTO E SAÚDE) FIZERAM A DIFERENÇA NO RESULTADO
- FALTA DE INFRAESTRUTURA NOS MUNICÍPIOS SELECIONADOS PARA RETER OS AS PESSOAS QUE GRADUAM NOS IES E PERMITE QUE EM SUA MAIORIA BUSQUEM OUTRAS REGIÕES PARA UTILIZAR SEUS CONHECIMENTOS ADQUIRIDOS

### Data Science Workflow Canvas\*

Start here. The sections below are ordered intentionally to make you state your goals first, followed by steps to achieve those goals. You're allowed to switch orders of these steps!

Title:		
<p><b>1 Problem Statement</b> What problem are you trying to solve? What larger issues do the problem address?</p> <p>O ensino superior descentralizado influencia no aumento do IDHM Educação, variável independente que por sua vez influencia no aumento do IDHM mais rapidamente que os municípios que não possuem polos de IES descentralizados</p>	<p><b>2 Outcomes/Predictions</b> What prediction(s) are you trying to make? Identify applicable predictor (X) and/or target (y) variables.</p> <p>Considerado a variável independente o IDHM Educação e variável Dependente o IDHM</p>	<p><b>3 Data Acquisition</b> Where are you sourcing your data from? Is there enough data? Can you work with it?</p> <p>Os dados analisados provêm das instituições governamentais brasileiras como INEP, IBGE e IPEA e instituição internacional, Organização das Nações Unidas (ONU) como o PNUD</p>
<p><b>4 Modeling</b> What models are appropriate to use given your outcomes?</p> <p>O modelo utilizado no aprendizado de máquina foi o de Regressão Linear simples</p>	<p><b>5 Model Evaluation</b> How can you evaluate your model's performance?</p> <p>O modelo foi capaz de atender as exigências do propósito do trabalho porém não limita a utilização de outros modelos que possam ser mais abrangentes na busca por mais respostas</p>	<p><b>6 Data Preparation</b> What do you need to do to your data in order to run your model and achieve your outcomes?</p> <p>Foi necessário utilizar mais de um software para preparação dos dados de modo a compilá-los de forma mais eficiente de acordo com o propósito deste trabalho</p>

### Activation

When you finish filling out the canvas above, now you can begin implementing your data science workflow in roughly this order.

1 Problem Statement → 2 Data Acquisition → 3 Data Prep → 4 Modeling → 5 Outcomes/Preds → 6 Model Eval

\* **Note:** This canvas is intended to be used as a starting point for your data science projects. Data science workflows are typically nonlinear.

*Figura 54 – Workflow deste trabalho de Ciência dos Dados*

## LINKS

[https://github.com/hermesmir/TCC\\_PUC/tree/main](https://github.com/hermesmir/TCC_PUC/tree/main)

<https://youtu.be/JpYxSouEXa8>