Entendimento dos Dados

24 de janeiro de 2017

Seleção dos Dados

Utilização dos dados de 2000 até 2009 como treino e de 2010 até 2015 para testes.

Atributos que serão levados em consideração foram explicados na fase de entendimento dos dados: sexo, idade, UF, cotista, tipo da escola, curso, forma de ingresso, IRA.

Construção dos Atributos Derivados

Atributos derivados que serão construídos já foram definidos na fase anterior.

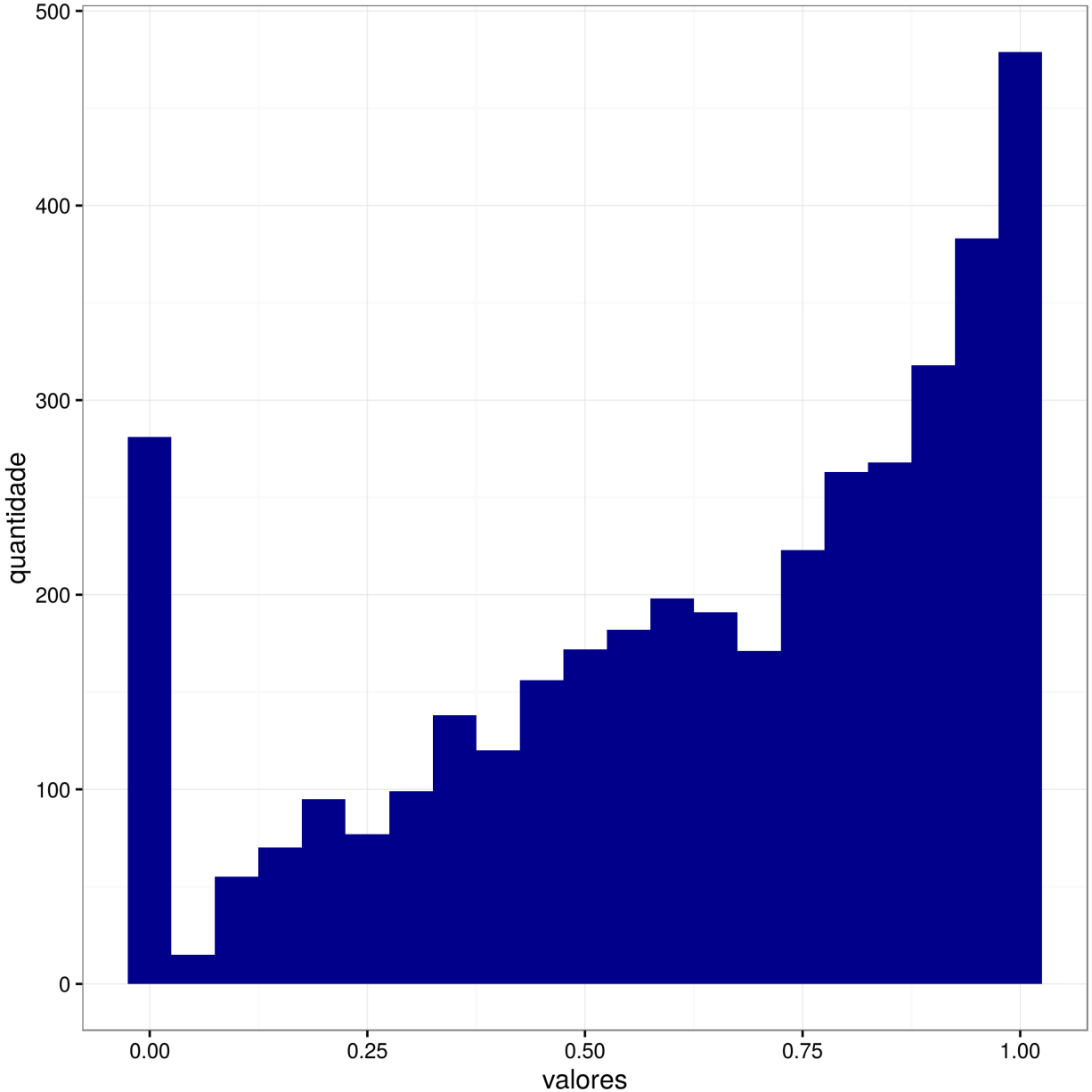
A seguir, histograma mostrando a distribuição de tais atributos

Histograma da Distribuição dos Atributos Derivados - Coeficiente de Melhora Acadêmica

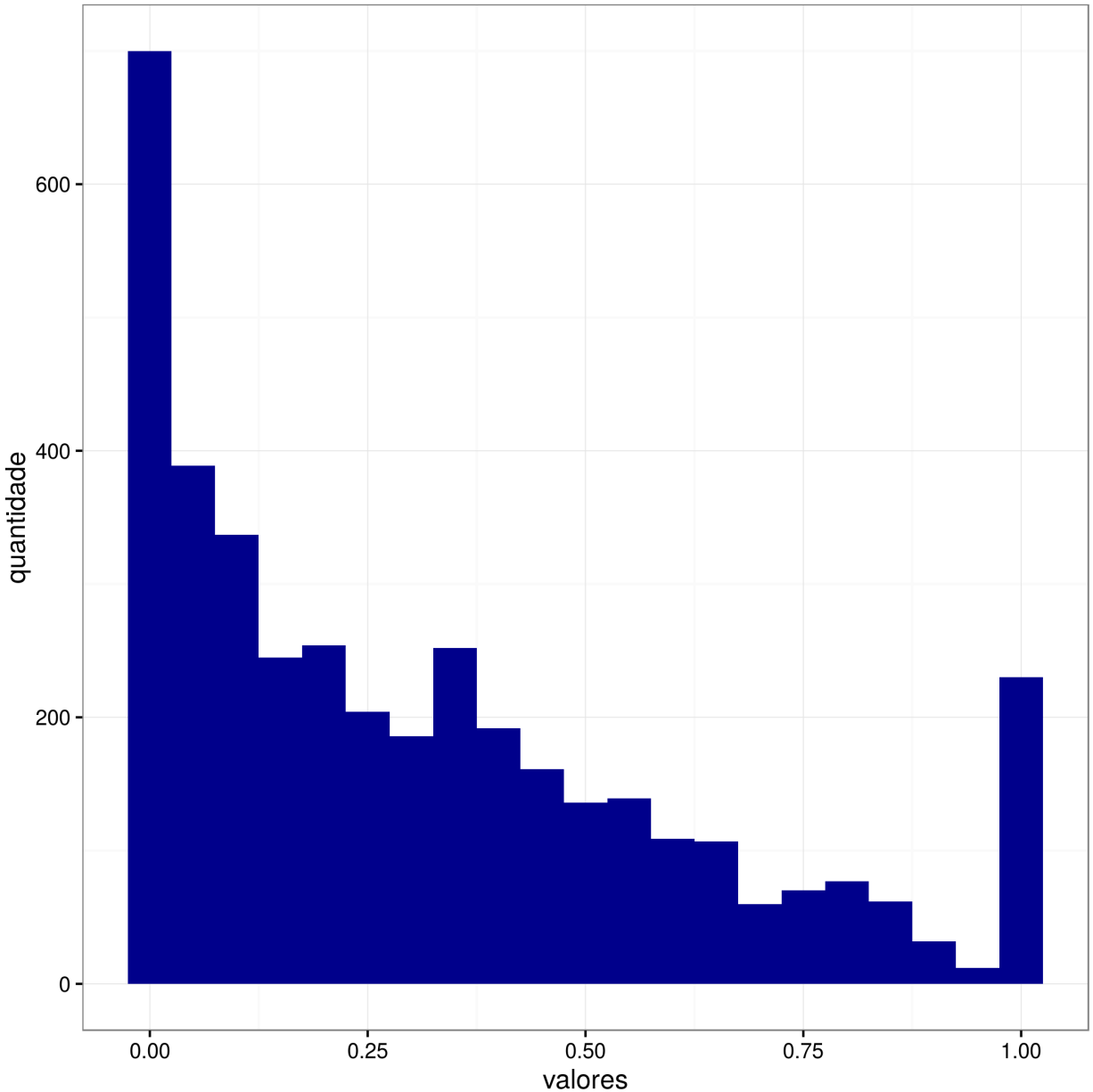
O coeficiente de melhora acadêmica é definido como sendo a razão entre as notas do semestre anterior e as notas do semestre anterior ao anterior.

Por histograma aqui

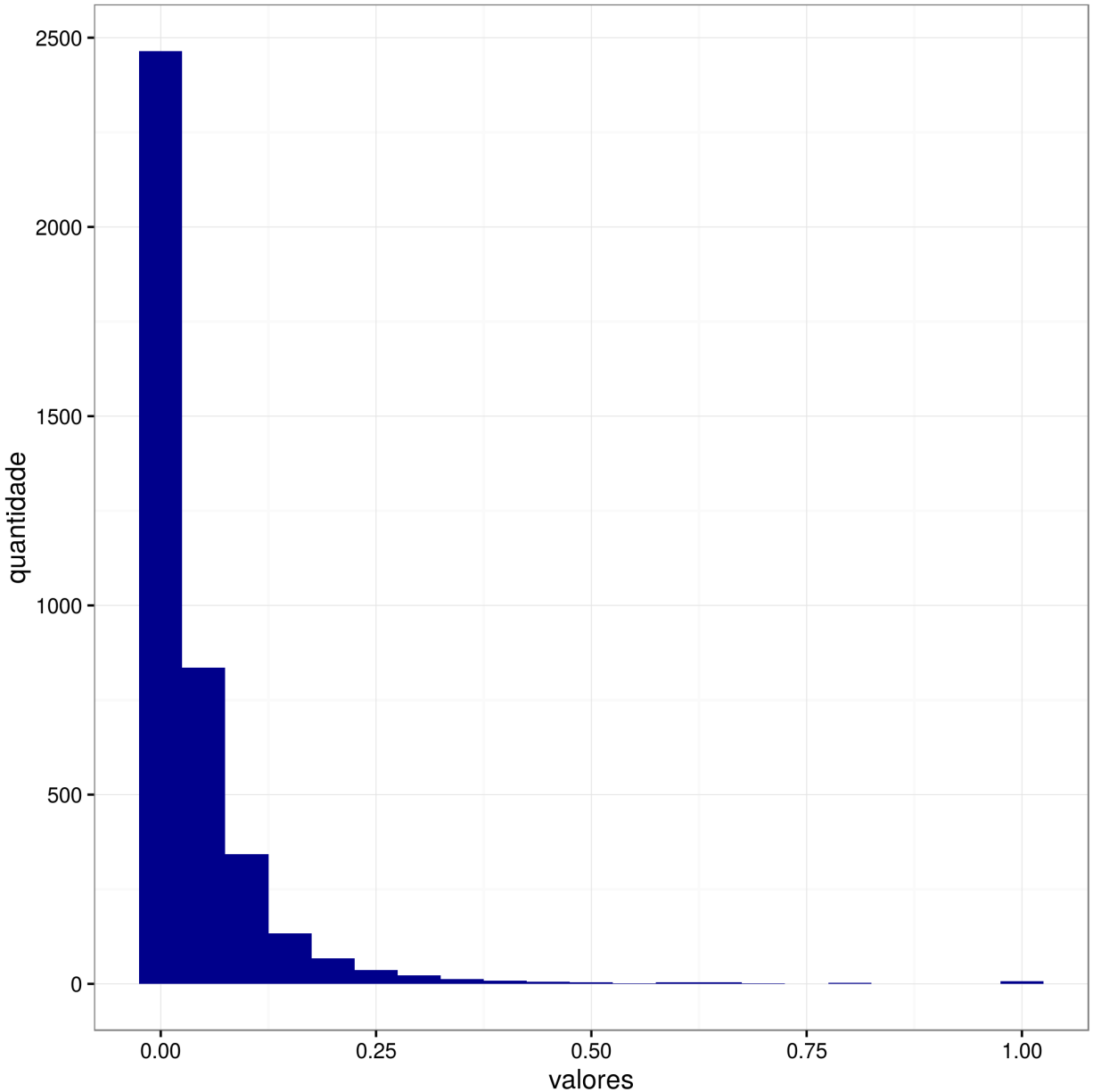
Histograma - Taxa de Aprovação



Histograma - Taxa de Reprovação



Histograma - Taxa de Trancamentos



Histograma - Razão entre disciplinas cursadas por semestre e disciplinas do curso

Histograma - Razão entre aprovações em disciplinas obrigatórias cursadas por semestre e disciplinas do curso

Taxa de Aprovação nas Disciplinas mais difíceis do Semestre

Demais Atributos Derivados

Demais atributos derivados incluem: booleano que indica se um aluno está ou não em condição e a posição em relação ao semestre (0 caso seja o pior aluno, 1 caso seja o melhor).

PCA

Utilização da técnica PCA (Principal Component Analysis) para verificar se features estão demasiadamente relacionados.

Limpeza dos Dados

Optou-se por não considerar features nos quais mais de 60% das entradas fossem missing values.

Descartou-se assim o feature raça.

Limpeza dos Dados

Para o caso de atributos que podem apresentar missing values, foi feita imputação (exemplo: coeficiente de melhora acadêmica).

Assim, caso haja missing value coloca-se a média do feature.

Integração dos Dados

Dados originais da SIGRA foram integrados com a informação dos currículos antigos e novos dos cursos, de modo a saber quais matérias são obrigatórias.

Transformação dos dados

Utilização de dummy variables para representar dados categóricos.

Referências