Análise de evasão dos cursos de Psicologia no Brasil

Alisson Rosa Pereira, Caroline Cogo Carneosso, Vítor Bernardo Silveira Pereira

08/07/2021

Sumário

1 Introdução

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) disponibiliza o Censo [1] da educação superior, o qual fornece informações detalhadas sobre os cursos superiores no Brasil. O objetivo do trabalho é estudar como as variáveis selecionadas influenciam na váriavel de desfecho. Trata-se de um estudo observacional transversal, uma vez que os ingressantes são estudados em um momento específico do tempo, neste caso no ano de 2019. Neste relatório faremos uma breve análise sobre o curso de Psicologia das Instituições de Ensino Superior (IES), estudando a permanência dos estudantes que ingressaram no ano de 2019, utiliza-se como variáveis de estudo:

Como variável de interesse tem-se a **Situação** do estudante, após o primeiro ano matriculado no ensino superior, ela é classificada em Retido e Evadido.

Turno: Uma variável categórica nominal que indica o tipo de turno que o estudante está vinculado, sendo dividido em 4 categorias.

Sexo: Uma variável categórica binária que indica se o estudante é do sexo masculino ou feminino.

Nacionalidade: Uma variável categórica nominal que indica se o aluno é Brasileiro, Exterior/Naturalizado ou Estrangeiro.

Apoio Social Uma variável categórica dicotômica que informa se o estudante possui ou não apoio social.

Idade Uma variável que foi discretizada em intervalos, isto é tornando-se uma variável categórica ordinal

[1] Acesse o Censo aqui

E por último a variável **Mobilidade** que foi criada pela junção do banco de dados dos alunos com o banco das Instituições de Ensino Superior, sendo esta uma variável categórica dicotômica para verificar se os alunos são do mesmo estado da universidade ou não.

Adota-se como convenção a partir de agora que todas as observações referem-se ao ano 2019. Todas as tabelas e gráficos foram elaboradas pelos autores com base nas 99119 observações do banco de dados, foram selecionadas 7 variáveis .

2 Estudo Descritivo

A partir dos dados, foram desenvolvidas tabelas e gráficos de frequência, para melhor compreensão e interpretação das variáveis a serem estudadas:

Situação:

Tabela 1: Frequência absoluta e relativa para a variável Situação.

Situação	Frequência Absoluta	Frequência Relativa
Retido	81538	0,823
Evadido	17581	0,177
Total	99119	1,000

Fonte: Elaborado pelos autores

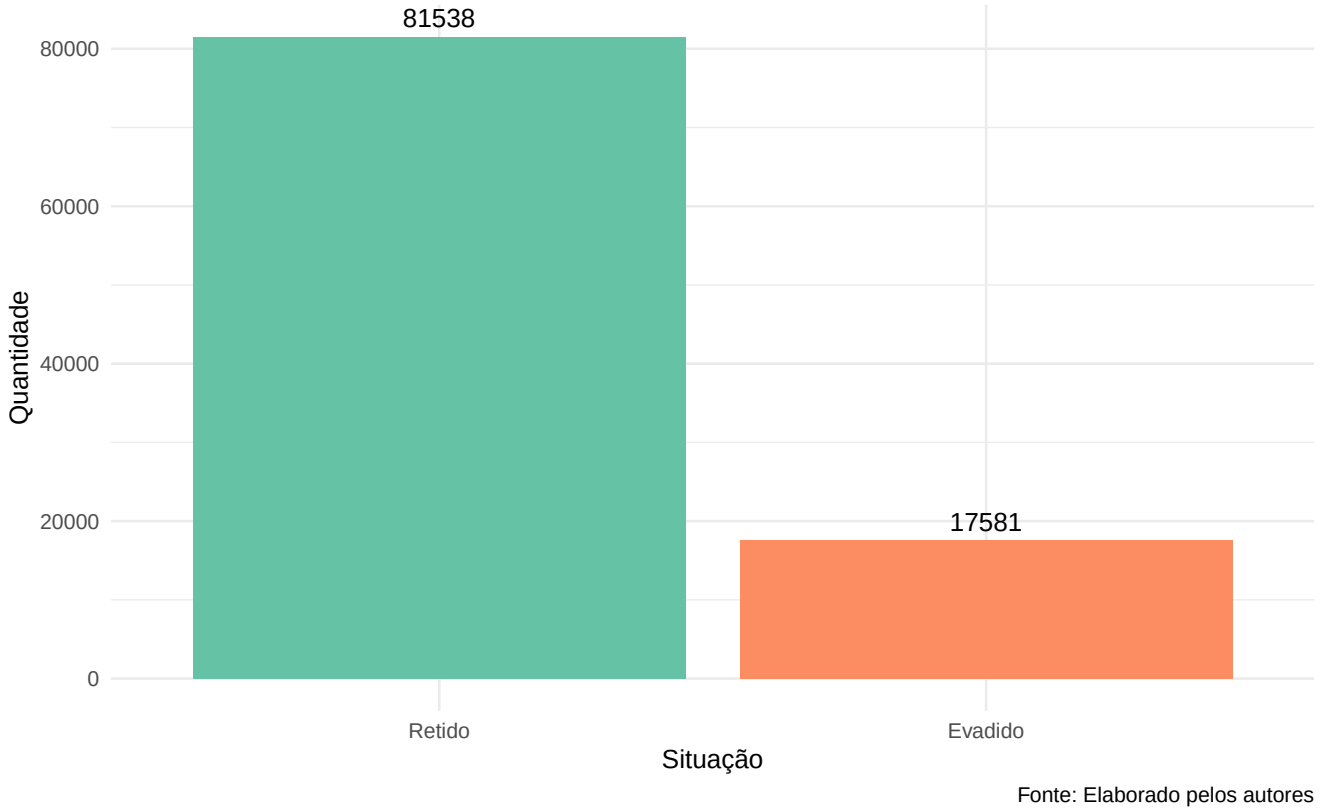


Figura 1: Gráfico de barras com os valores absolutos da variável Situação.

A partir da tabela 1 e do gráfico 1, verifica-se que cerca de $82,26\%$ foram retidos, ou seja continuaram no curso de Psicologia, já $17,74\%$ foram evadidos do curso.

Turno:

Tabela 2: Frequência absoluta e relativa para a variável Turno.

Turno	Frequência Absoluta	Frequência Relativa
Matutino	31491	0,318
Vespertino	2398	0,024
Noturno	59777	0,603
Integral	5453	0,055
Total	99119	1,000

Fonte: Elaborado pelos autores

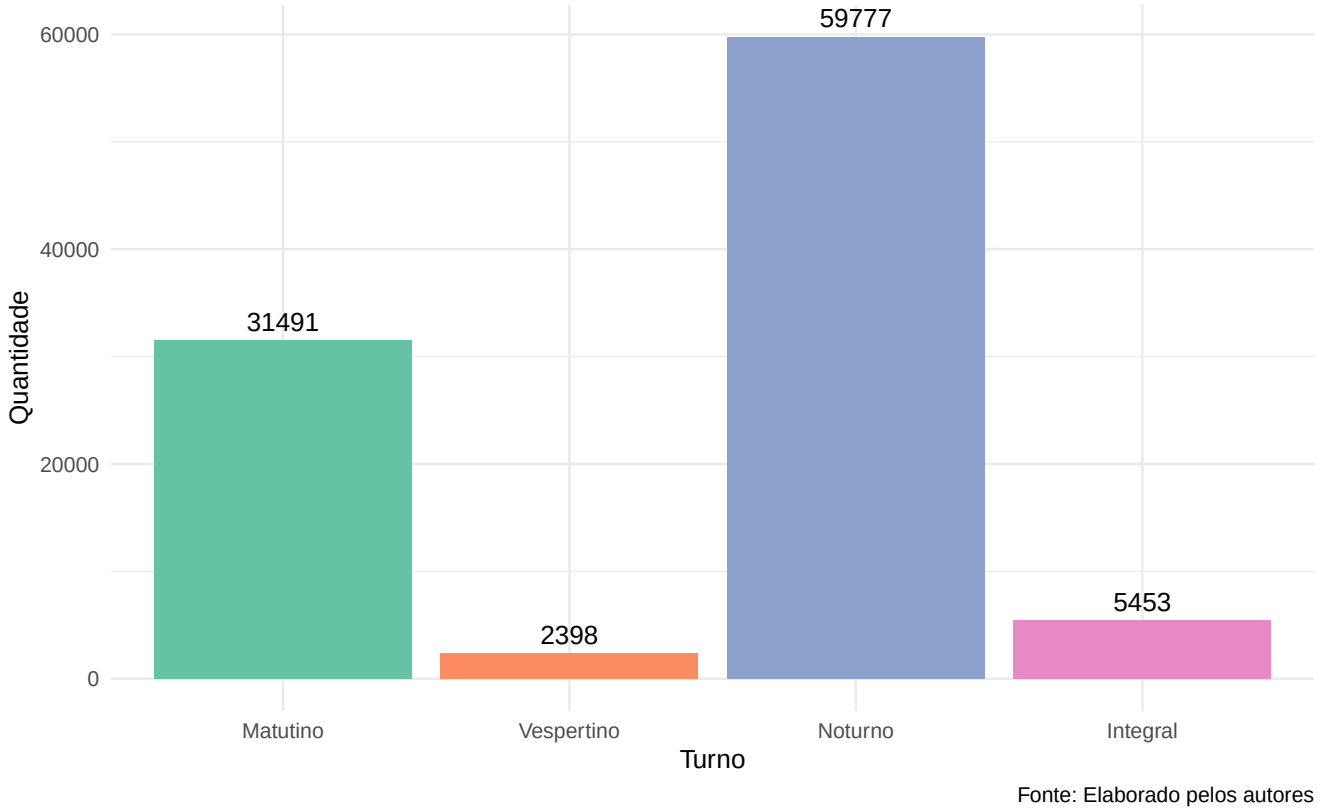


Figura 2: Gráfico de barras com os valores absolutos da variável Turno

Observa-se na tabela 2 e no gráfico 2 que aproximadamente 60% dos estudantes são do turno noturno, apenas 2% são do turno vespertino.

Sexo:

Tabela 3: Frequência absoluta e relativa para a variável Sexo.

Sexo	Frequência Absoluta	Frequência Relativa
Feminino	76304	0,77
Masculino	22815	0,23
Total	99119	1,00

Fonte: Elaborado pelos autores

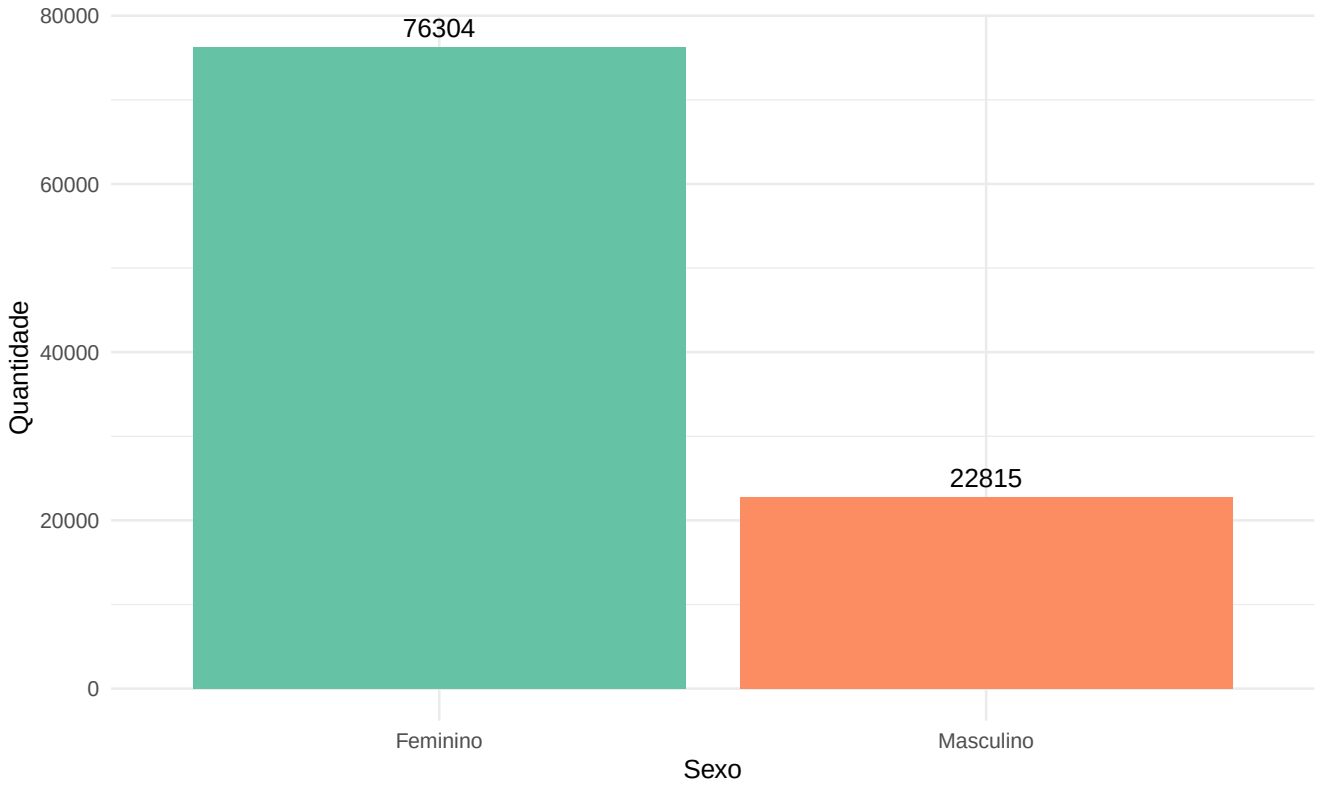


Figura 3: Gráfico de barras com os valores absolutos da variável Sexo.

Como é observado na tabela 3 e no gráfico 3, mais de 75% dos alunos ingressantes são do sexo feminino.

Nacionalidade:

Tabela 4: Frequência absoluta e relativa para a variável Nacionalidade.

Nacionalidade	Frequência Absoluta	Frequência Relativa
Brasileira	98824	0,997
Exterior/Naturalizado	139	0,001
Estrangeira	156	0,002
Total	99119	1,000

Fonte: Elaborado pelos autores

Ao analisar a tabela 4 é possivel constatar que a maioria dos ingressantes é brasileira, cerca de $99,7\%$.

Apoio Social

Tabela 5: Frequência absoluta e relativa para a variável Apoio Social.

Apoio social	Frequência Absoluta	Frequência Relativa
Não	96353	0,972
Sim	2766	0,028
Total	99119	1,000

Fonte: Elaborado pelos autores

Percebe-se pela tabela 5 que aproximadamente 97% dos estudantes não recebem apoio social.

A tabela 6 e o gráfico 4 representam a **idade** discretizada, ou seja a quantidade de observações nos respectivos intervalos para cada idade.

Tabela 6: Frequência absoluta e relativa para a variável Idade dividida em classes.

Idade	Frequência Absoluta	Frequência Relativa
[15,20]	41198	0,416
(20,25]	23038	0,232
(25,30]	10086	0,102
(30,65]	24614	0,248
(65,90]	183	0,002
Total	99119	1,000

Fonte: Elaborado pelos autores

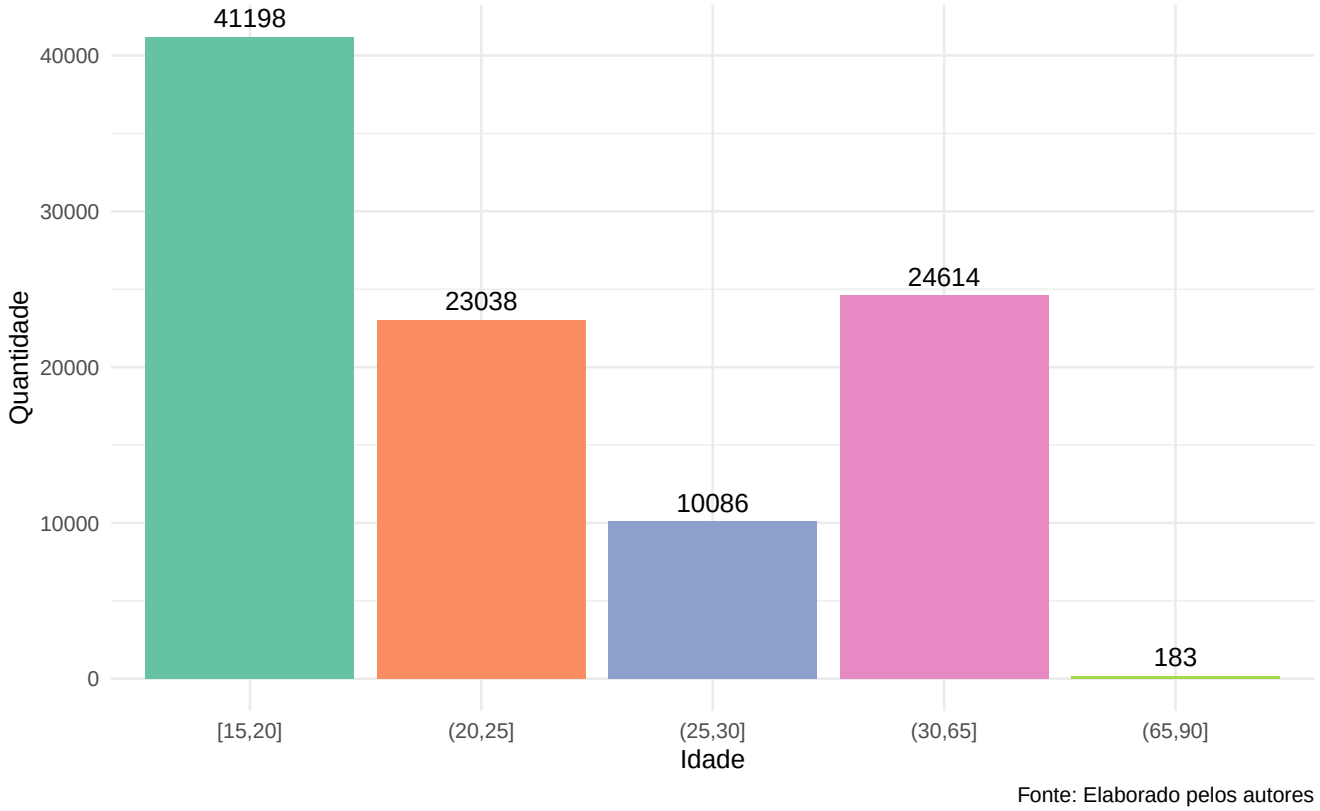


Figura 4: Gráfico de barras com os valores absolutos da variável Idade dividida em classes.

Como existe uma grande concentração de ingressantes na faixa etária entre 15 a 20 anos, cerca de 41%, é construído o gráfico abaixo, o qual se refere a quantidade de alunos com a respectiva idade.

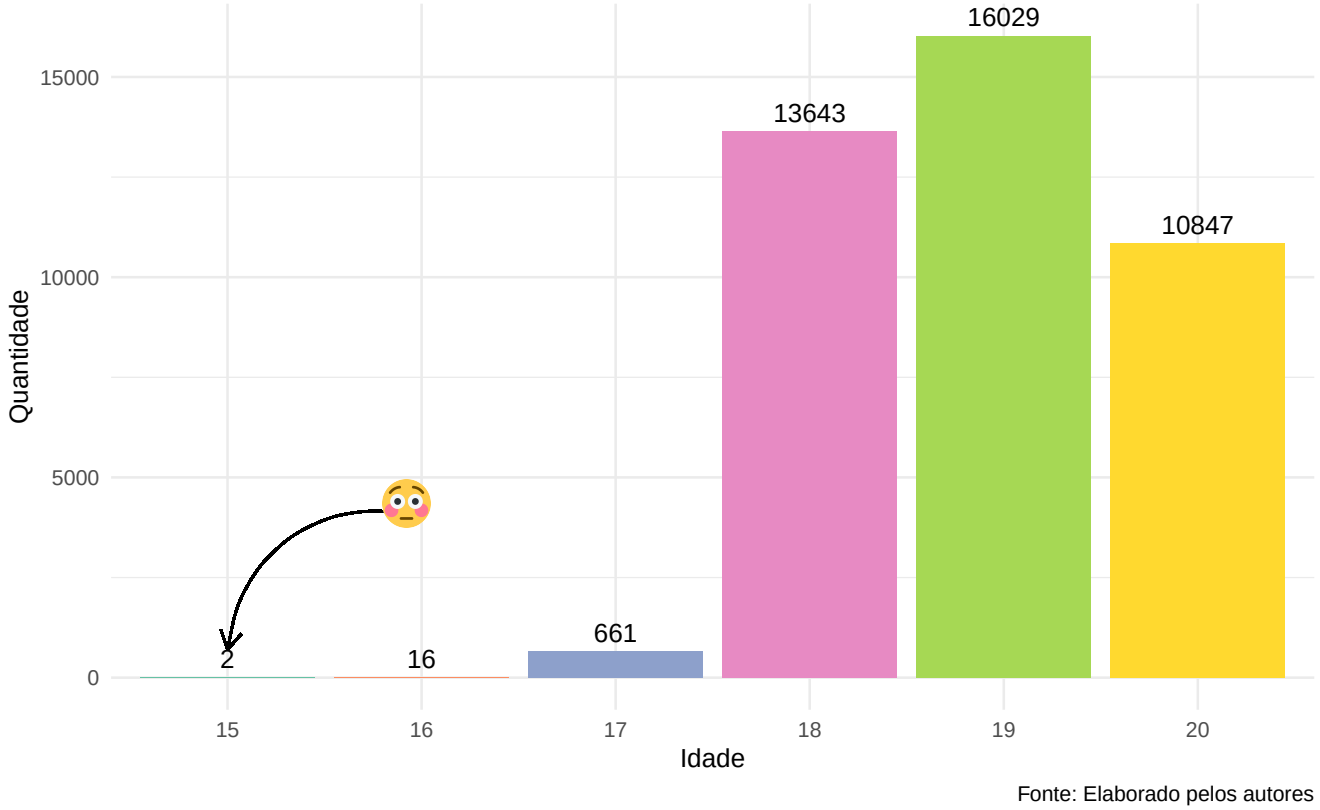


Figura 5: Valores Absolutos da variável Idade na faixa etaria 15 a 20 anos

Mobilidade

Tabela 7: Frequência absoluta e relativa para a variável Mobilidade.

Mobilidade	Frequência Absoluta	Frequência Relativa
Estado diferentes	12919	0,179
Mesmo estado	59131	0,821
Total	72050	1,000

Fonte: Elaborado pelos autores

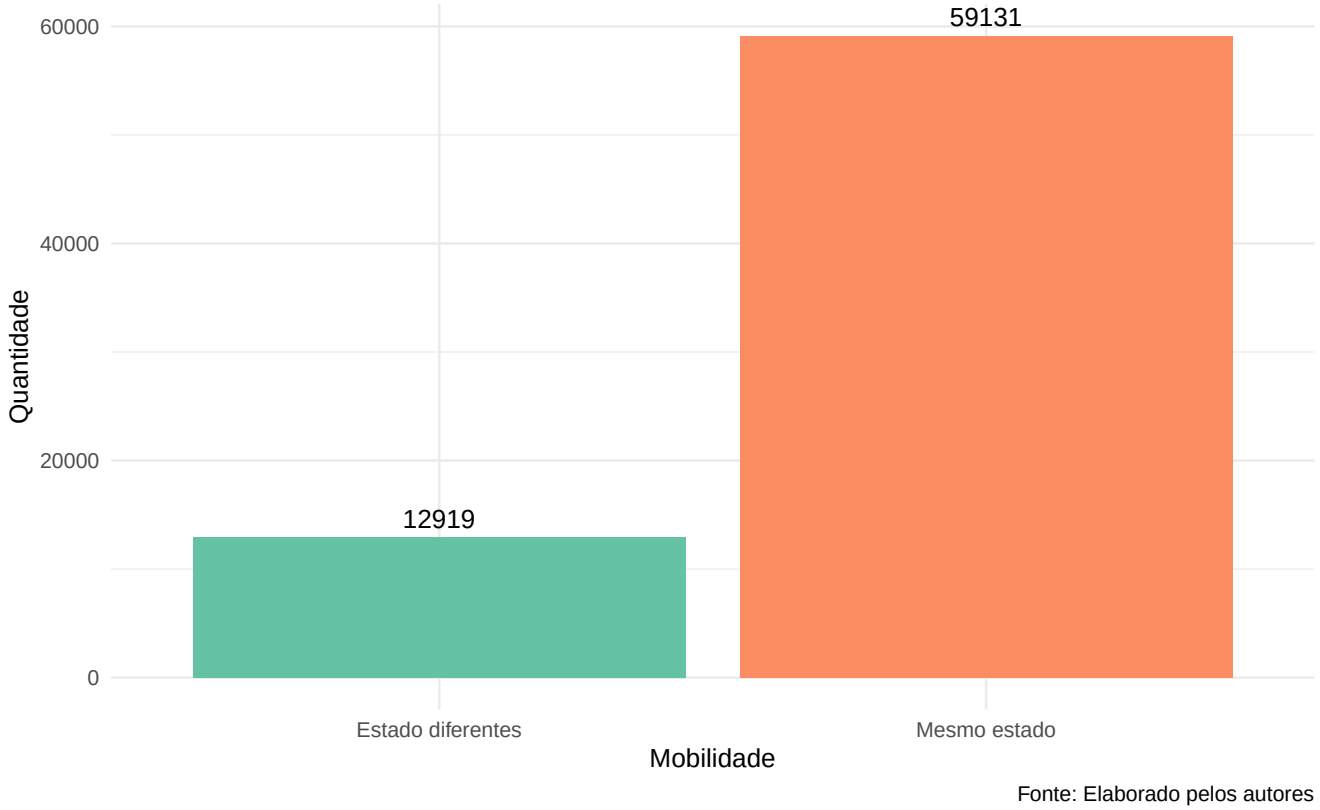


Figura 6: Gráfico de barras com os valores absolutos da variável Mobilidade.

Nota-se que a partir da tabela 7 e da figura 6, cerca de 82% dos alunos ingressaram na universidade do seu estado de origem.

3 Estudo de Associação

A partir do estudo de frequência realizado pelo Estudo Descritivo, iremos construir tabelas, gráficos e testes para verificar a associação entre a variável de desfecho e as variáveis de interesse, com as seguintes hipóteses:

$$H_0 : \text{Ausência de associação entre as variáveis}$$

versus a hipótese alternativa

$$H_1 : \text{Presença de associação}$$

.

3.1 Tabelas de Contigência e Testes

Nesta seção, iremos verificar a relação entre as variáveis com a visualização das tabelas de contingência, gráfico e a análise dos testes qui-quadrado.

Turno vs Situação

Tabela 8: Tabela de contingência com os valores absolutos da variável Turno, considerando a variável de desfecho.

Turno	Situação		
	Retido	Evadido	Total
Matutino	26304	5187	31491
Vespertino	2137	261	2398
Noturno	47978	11799	59777
Integral	5119	334	5453
Total	81538	17581	99119

Fonte: Elaborado pelos autores

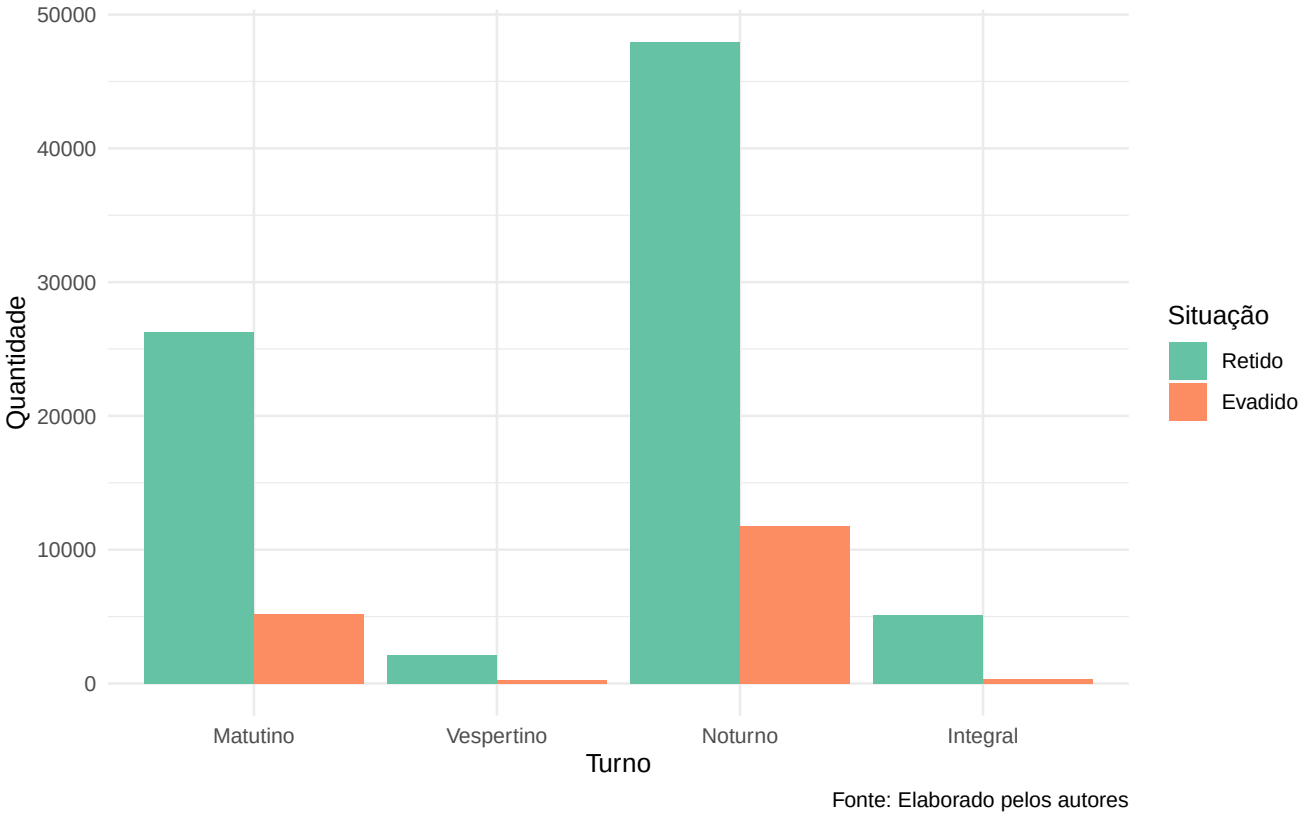


Figura 7: Gráfico de barras para o turno dos ingressantes conforme a variável de desfecho.

A tabela 8 e o gráfico 7 expõem que há prevalência dos alunos no curso de Psicologia em todos os turnos, sendo o noturno o turno com maior evasão em valores absolutos.

Para o teste de associação, temos que a estatística de teste é igual a 779,759, com p-valor menor que 0.01, ou seja, rejeita-se H0 com todos os níveis de significância, assim há evidência de que há associação entre as variáveis.

Sexo vs Situação

Tabela 9: Tabela de contingência com os valores absolutos da variável Sexo, considerando a variável de desfecho.

Sexo	Situação		
	Retido	Evadido	Total
Feminino	63107	13197	76304
Masculino	18431	4384	22815
Total	81538	17581	99119

Fonte: Elaborado pelos autores

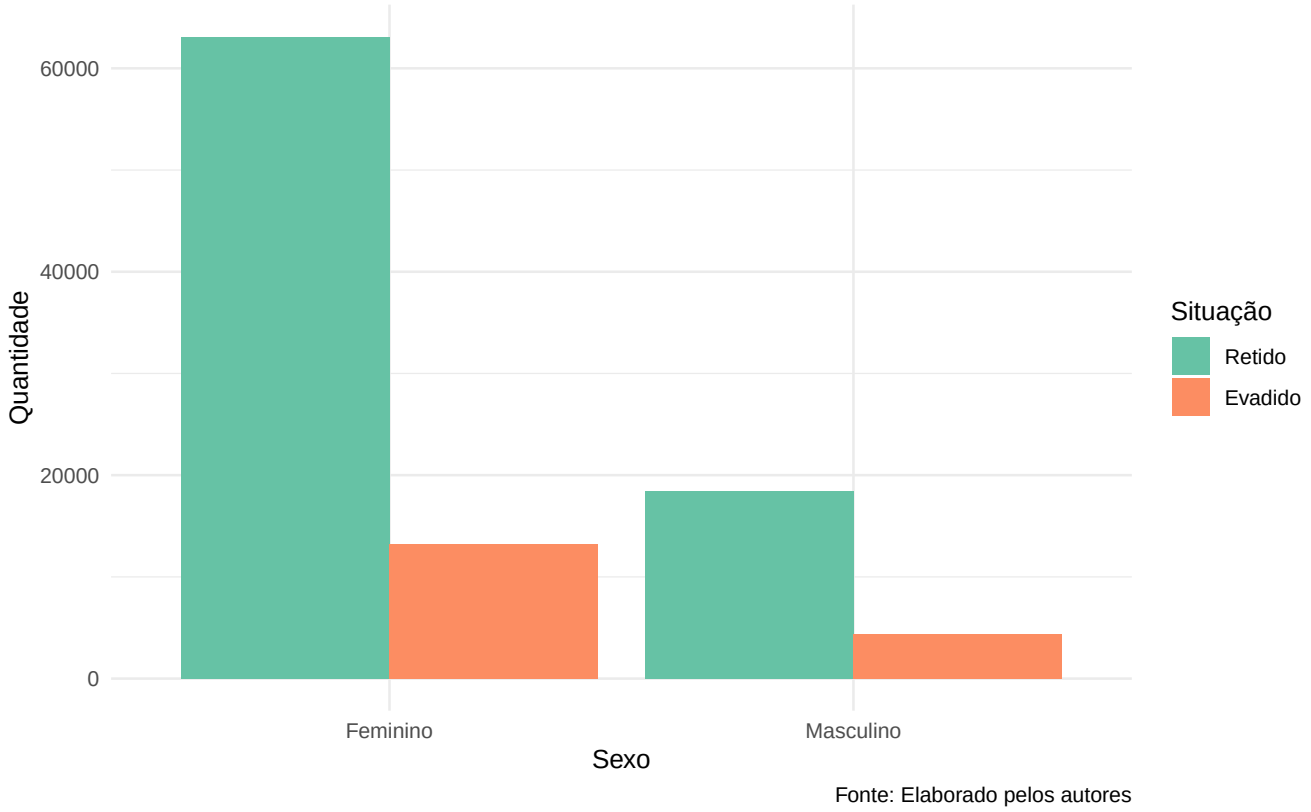


Figura 8: Gráfico de barras para o Sexo dos ingressantes de acordo com a variável de desfecho.

A tabela 9 e o gráfico 8 evidenciam em valores absolutos que a maior parte dos ingressantes permanece no curso em ambos os sexos.

Para o teste de associação, temos que a estatística de teste é igual a 44,38, com p-valor menor que 0.01 , sendo assim, rejeitamos H0 a todos os níveis de significância, logo há evidência de que há associação entre as variáveis. Sendo uma tabela 2x2 podemos calcular a razão de prevalências que é: 1,024, podemos confirmar com o no intervalo de confiança [1,017;1,031], então podemos ver que 1 não está contido no intervalo, reforçando esta associação.

Nacionalidade vs Situação

Tabela 10: Tabela de contingência com os valores absolutos da variável Nacionalidade, considerando a variável de desfecho.

Nacionalidade	Situação		
	Retido	Evadido	Total
Brasileira	81263	17561	98824
Exterior/Naturalizado	132	7	139
Estrangeira	143	13	156
Total	81538	17581	99119

Fonte: Elaborado pelos autores

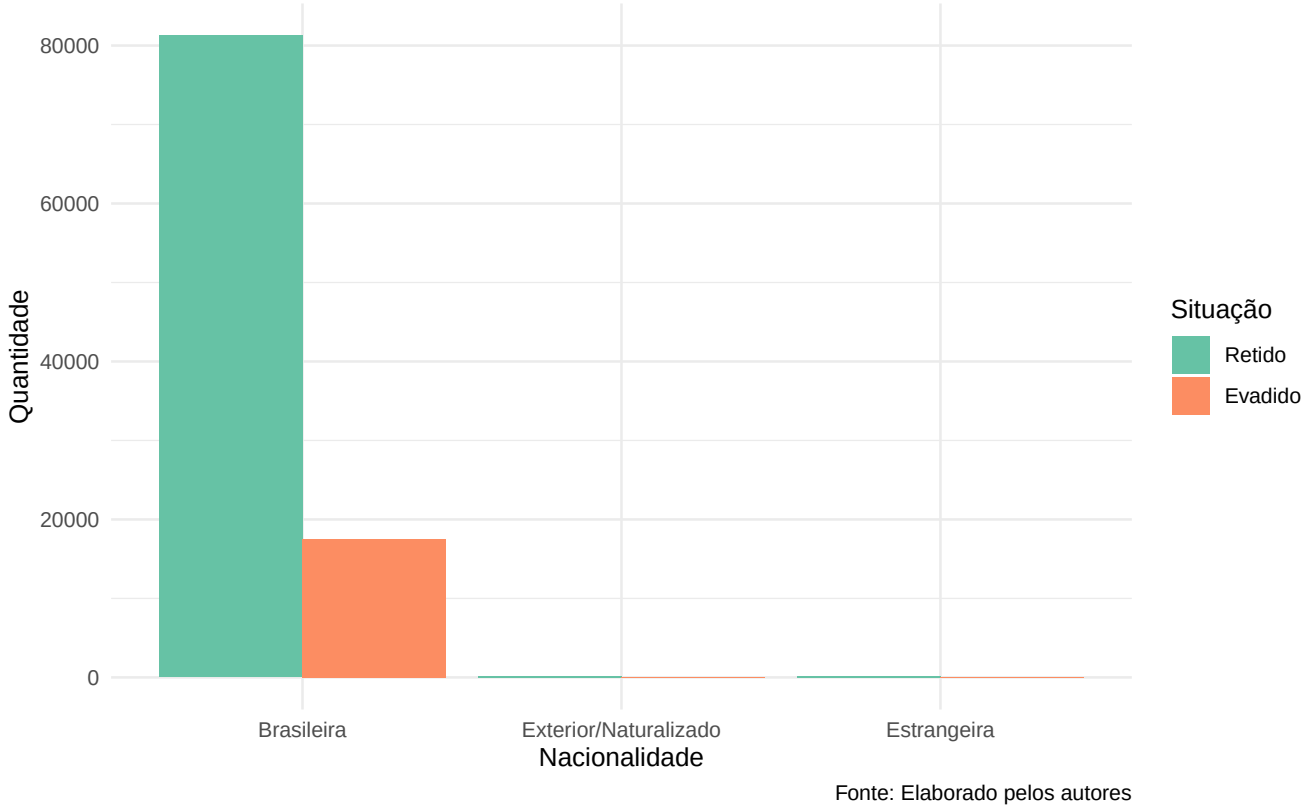


Figura 9: Gráfico de barras para o Nacionalidade dos ingressantes de acordo com a variável de desfecho.

A tabela 10 e o gráfico 9 apresentam em valores absolutos que a minoria dos alunos não-brasileiros evadem.

Para o teste de associação, temos que a estatística de teste é igual a 24,895, com p-valor menor que 0.01, ou seja, rejeita-se H0 com todos os níveis de significância, assim há evidência de que há associação entre as variáveis.

Apoio Social vs Situação

Tabela 11: Tabela de contingência com os valores absolutos da variável Apoio social, considerando a variável de desfecho.

	Situação		
Apoio social	Retido	Evadido	Total
Não	78886	17467	96353
Sim	2652	114	2766
Total	81538	17581	99119

Fonte: Elaborado pelos autores

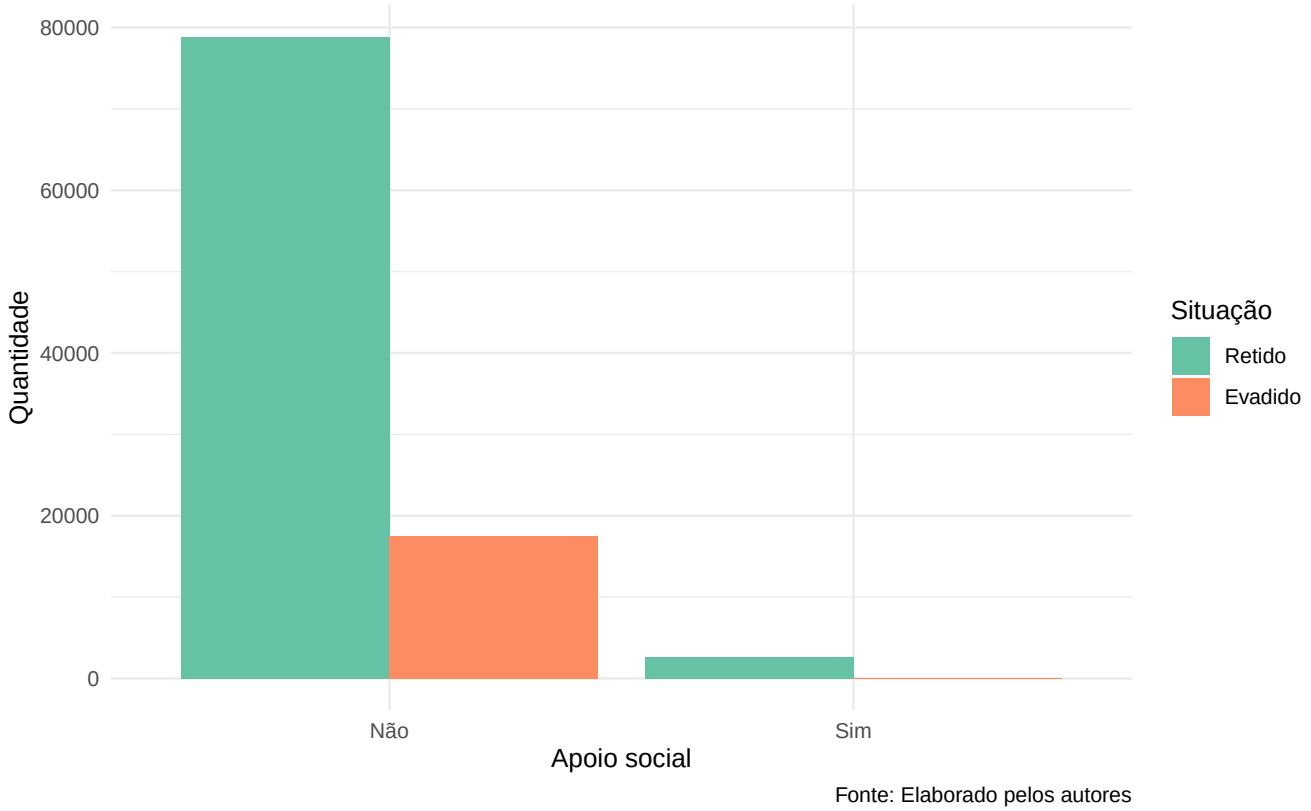


Figura 10: Gráfico de barras para o Apoio Social dos ingressantes de acordo com a variável de desfecho.

A tabela 11 e o gráfico 10 nos informa que evidentemente os estudantes que não possuem apoio social possuem maior retenção em valor absoluto (são 97% dos ingressantes), mas veremos na próxima seção em termos relativos quanto é essa retenção.

Para o teste de associação, temos que a estatística de teste é igual a 361,526, com p-valor menor que 0.01 , sendo assim, rejeitamos H0 a todos os níveis de significância, logo há evidência de que há associação entre as variáveis. Sendo uma tabela 2x2 podemos calcular a razão de prevalências que é: 0,854, podemos confirmar com o no intervalo de confiança [0,847;0,861], então podemos ver que 1 não está contido no intervalo, reforçando esta associação.

Idade vs Situação

Tabela 12: Tabela de contingência com os valores absolutos da variável Idade, considerando a variável de desfecho.

	Situação		
Idade	Retido	Evadido	Total
[15,20]	34567	6631	41198
(20,25]	18988	4050	23038
(25,30]	8147	1939	10086
(30,65]	19703	4911	24614
(65,90]	133	50	183
Total	81538	17581	99119

Fonte: Elaborado pelos autores

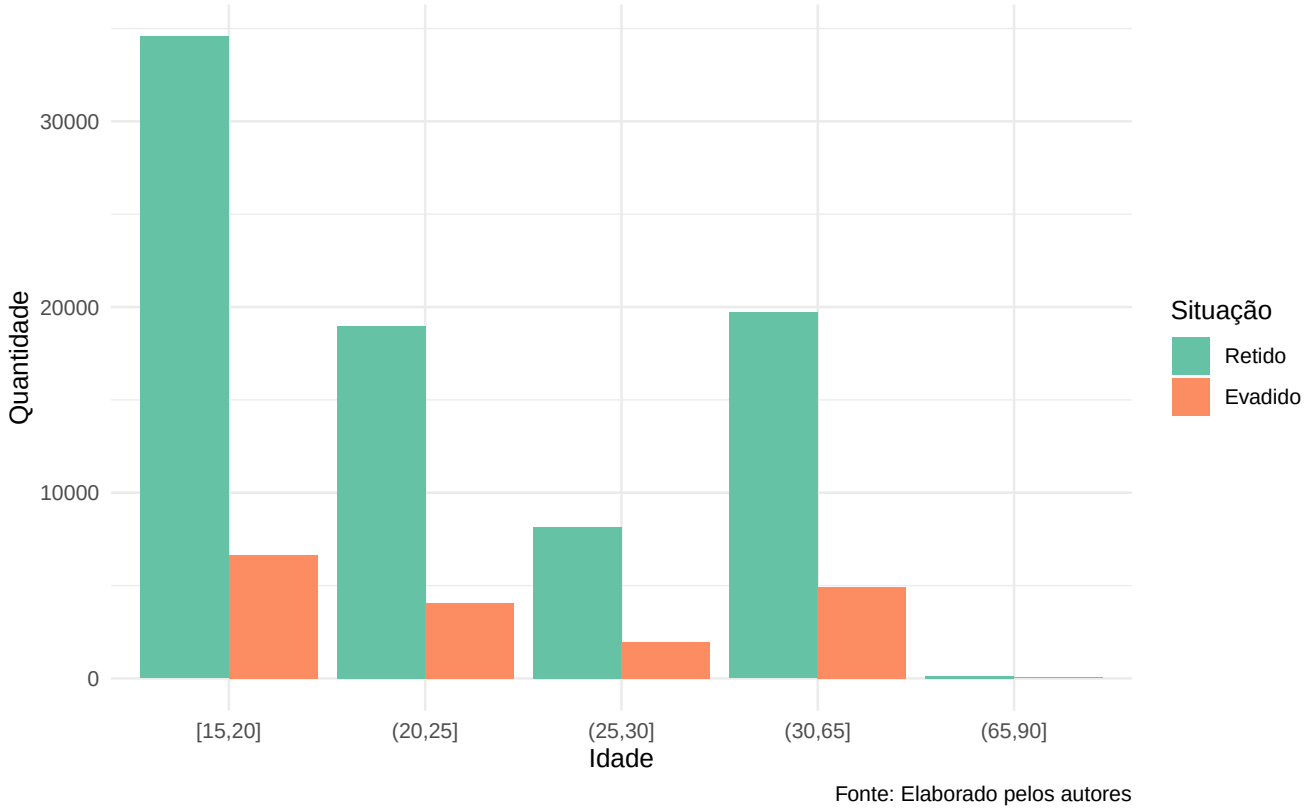


Figura 11: Gráfico de barras para o Idade dos ingressantes de acordo com a variável de desfecho.

Vemos que a faixa etária 15 a 20 anos possue maior retenção em valor absoluto, entretanto possui a maior quantidade de ingressantes, na proxima seção constataremos essa informação de forma mais detalhada.

Para o teste de associação, temos que a estatística de teste é igual a 0 com o p valor igual a 1, ou seja, não há nenhuma evidência de associação entre Idade e Situação. A correlação é igual a 0,043, então a relação linear entre as duas variáveis é muito baixa.

Mobilidade vs Situação

Tabela 13: Tabela de contingência com os valores absolutos da variável Mobilidade, considerando a variável de desfecho.

	Situação		
Mobilidade	Retido	Evadido	Total
Estado diferentes	9759	3160	12919
Mesmo estado	48616	10515	59131
Total	58375	13675	72050

Fonte: Elaborado pelos autores

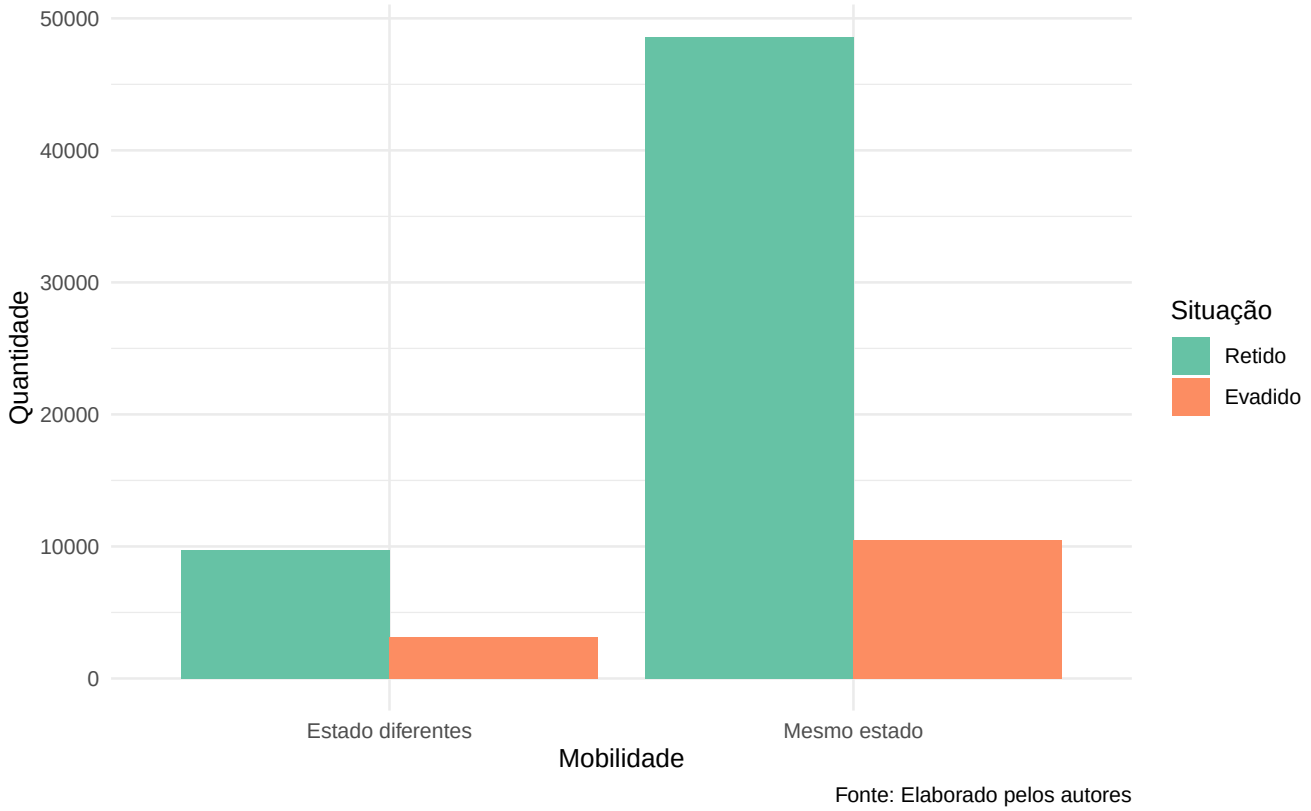


Figura 12: Gráfico de barras para o Mobilidade dos ingressantes de acordo com a variável de desfecho.

Percebe-se na tabela 13 e no gráfico 12 uma enorme retenção em valores absolutos dos ingressantes que são do mesmo estado que a universidade.

Para o teste de associação, temos que a estatística de teste é igual a 307,438, com p-valor menor que 0.01 , sendo assim, rejeitamos H0 a todos os níveis de significância, logo há evidência de que há associação entre as variáveis. Sendo uma tabela 2x2 podemos calcular a razão de prevalências que é: 0,919, podemos confirmar com o no intervalo de confiança [0,909;0,928], então podemos ver que 1 não está contido no intervalo, reforçando esta associação.

3.2 Tabelas Marginais

Nesta seção, iremos visualizar através de tabelas , o efeito marginal das variáveis explicativas na variável de desfecho

Tabela 14: Proporção marginal do Turno na Situação do ingressante

	Situação	
Turno	Retido	Evadido
Matutino	0,835	0,165
Vespertino	0,891	0,109
Noturno	0,803	0,197
Integral	0,939	0,061

Fonte: Elaborado pelos autores

Observando a tabela 14, constata-se que o turno que mais evade é o noturno, com proximidade do matutino, já o que menos evade é o turno integral.

Tabela 15: Proporção marginal do Sexo na Situação do ingressante

	Situação	
Sexo	Retido	Evadido
Feminino	0,827	0,173
Masculino	0,808	0,192

Fonte: Elaborado pelos autores

Apesar de termos aproximadamente 70% de mulheres ingressantes e como vimos no gráfico 8 existe maior retenção de estudantes do sexo feminino em valores absolutos, entretanto em termos relativos podemos notar que não há diferença significativa na permanência no curso em relação ao sexo.

Tabela 16: Proporção marginal da Nacionalidade na Situação do ingressante

	Situação	
Nacionalidade	Retido	Evadido
Brasileira	0,822	0,178
Exterior/Naturalizado	0,950	0,050
Estrangeira	0,917	0,083

Fonte: Elaborado pelos autores

O gráfico 9 nos informou que em valores absolutos existe mais retenção entre os ingressantes de naturalidade Brasileira e também sabemos que temos poucas pessoas que vieram do exterior ou foram naturalizadas, mas com tabela 16 notamos dentre os ingressantes os do esterior ou nacionalizados são os que possuem maior taxa de retenção no curso de psicologia.

Tabela 17: Proporção marginal de possuir Apoio Social na Situação do ingressante

	Situação	
Apoio social	Retido	Evadido
Não	0,819	0,181
Sim	0,959	0,041

Fonte: Elaborado pelos autores

É observado que os ingressantes que possuem apoio social, mesmo sendo poucos, são os que possuem maior retenção.

Tabela 18: Proporção marginal da Idade na Situação do ingressante

	Situação	
Idade	Retido	Evadido
[15,20]	0,839	0,161
(20,25]	0,824	0,176
(25,30]	0,808	0,192
(30,65]	0,800	0,200
(65,90]	0,727	0,273

Fonte: Elaborado pelos autores

Em relação a taxa de permanência para a variável idade, é perceptivel na faixa etária de 65 a 90 anos uma proporção maior de evasão do que nas outras faixas etárias.

Tabela 19: Proporção marginal da "Mobilidade"na Situação do ingressante

	Situação	
Mobilidade	Retido	Evadido
Estado diferentes	0,755	0,245
Mesmo estado	0,822	0,178

Fonte: Elaborado pelos autores

Para estudantes de origem do mesmo estado onde está localizada a universidade, há uma taxa de evasão relativamente menor em comparação com estudantes de origem de estados diferentes.

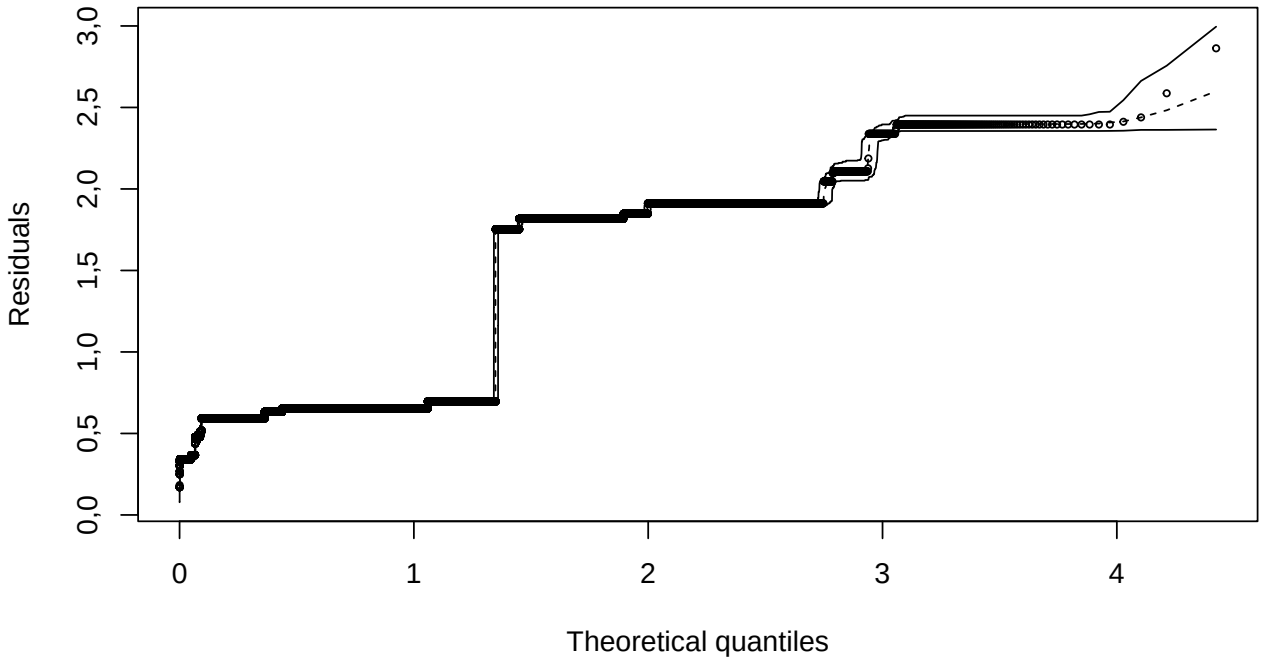
4 Modelagem

Nesta seçao, iremos ajustar uma regressão logistica e floresta aleatória, temos como variável a ser predita se o estudante vai ser estar retido no curso, ou vai evadir, utiliza-se como covariáveis o sexo, nacionalidade e turno do estudante utiliza-se. Como observado na tabela 1 tem-se que as observações nas categorias da variável resposta são muito desequilibradas, portanto a acurácia (accuracy) não será uma boa métrica para a escolha dos hiperparâmetros, assim escolhe-se a área da curva roc como métrica.

4.1 Regressão Logística

A regressão logistica foi necessário tomar como 0.808 o ponto corte para a previsão ser da categória Retido, pois como já dito existem muitas observações para essa categoria. Ajustando no conjunto de treino obtem-se como única variável não significativa a dummie Nacionalidade Estrangeira que tem como p-valor associado 0.054, o seguinte gráfico ilustra o ajuste sobre o modelo logístico, indicando que está bem ajustado inferencialmente dado que todos resíduos estão dentro da banda simulada.

```
## Binomial model
```



Testando no conjunto de teste que possui 34693 observações e deixando o ponto de corte padrão (0.5), a acurácia nos dados de teste é 0.823, poderia-se cair em tentanção e dizer que é um bom modelo, mas vejamos a matriz de confusão:

Tabela 20: Matriz de Confusão com ponte de corte 0.5

	Observado		
Predito	Retido	Evadido	Total
Retido	28539	6154	34693

Fonte: Elaborado pelos autores

Todas as previsões foram Retido!!, ou seja o modelo está completamente viesado, e possui uma alta acurácia pois a maior parte dos estudantes permaneceu no curso. Então é necessário ajustarmos esse ponto de corte, por otimização da área da curva roc ficamos com 0.808 , assim obtem-se area sobre a curva roc 0.555 e matriz de confusão

Tabela 21: Matriz de Confusão com ponte de corte 0.808

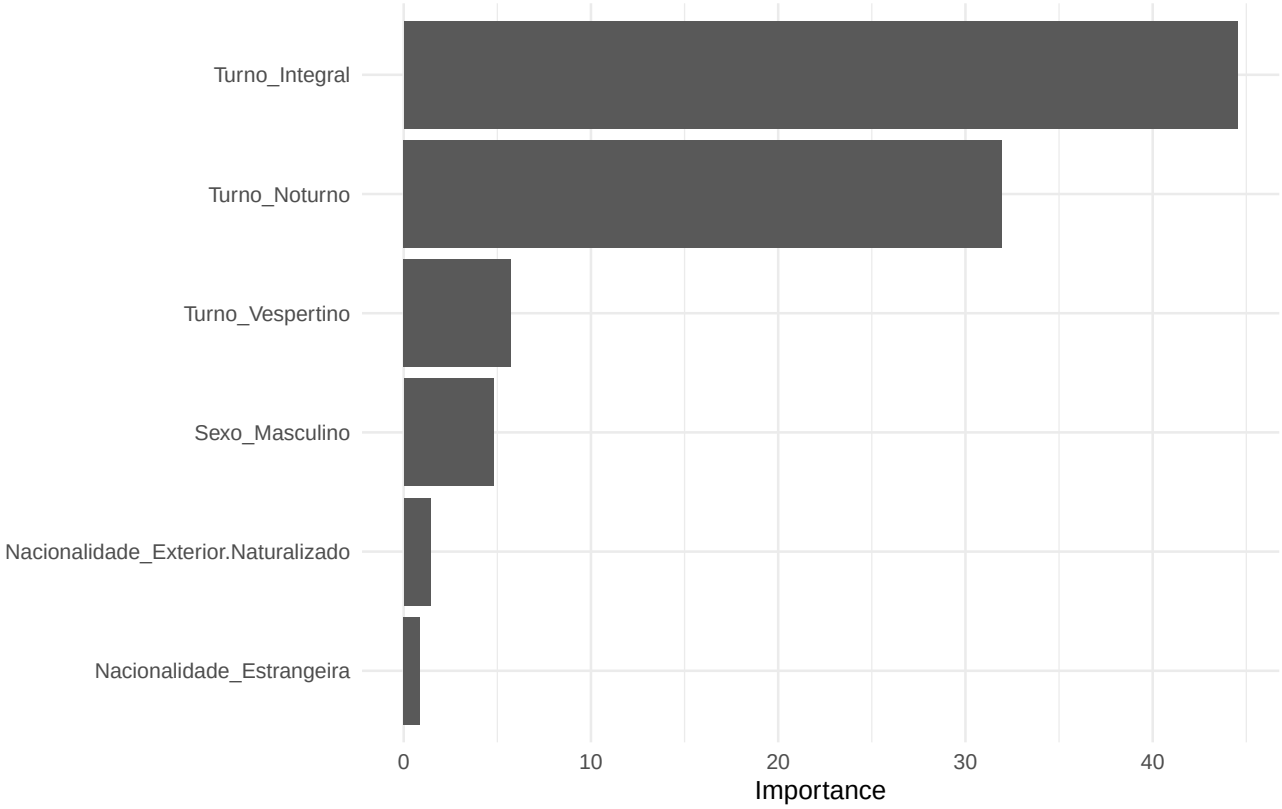
	Observado		
Predito	Retido	Evadido	Total
Retido	24615	5107	29722
Evadido	3924	1047	4971

Fonte: Elaborado pelos autores

4.2 Random Forest

O fenômeno de predizer todos como retido, também acontece na random forest, portanto o ponto de corte foi ajustado para 0.81, utilizou-se também a quantidade de arvores de decisão (ntree) igual a 500, e por otimização

mtry igual a 1 e min_n igual a 36. O seguinte gráfico fornece as variáveis mais importantes no conjunto de treino.



O ajuste final considerando o conjunto de teste teve area sobre a curva roc igual a 0.555 e acurácia 0.74. A matriz de confusão da random forest é dada por :

Tabela 22: Matriz de Confusão com ponte de corte 0.808

	Observado		
Predito	Retido	Evadido	Total
Retido	24615	5107	29722
Evadido	3924	1047	4971

Fonte: Elaborado pelos autores

5 Conclusão

De modo geral, para os ingressantes do curso de Psicologia do ano de 2019, podemos notar que a maioria dos ingressantes foi do sexo feminino, mas em termos da variável de desfecho não existe uma diferença significativa na taxa de evasão em relação ao sexo. A taxa de evasão para as idades em classes só não se manteve em tornos de 80% para a faixa etária de 65 a 90 anos que inclusive é a maior taxa de evasão de todos os níveis das variáveis estudadas. Para o turno do ingressante percebe-se que grande parte foi do noturno e em relação a variável de desfecho a maior taxa de permanência (retido) foi para os estudantes do turno integral. Apesar da variável resposta ser desbalanceada em um primeiro momento ambos os modelos tiveram um desempenho semelhante : mediano.