Análise dos datasets e resultados

Felipe B. Diniz

1 Parte 1

1.1 Primeiro contato com o dataset

Todos as figuras neste texto podem ser reproduzidas, basta usar os código disponíveis no meu Github.[1] Começamos analisando a *BASE A*. Antes de proceder para as análises de fato, verificamos que não há dados omissos ou duplicatas. Abaixo, temos uma amostra destes dados.

StudentId	RegisteredDate	SubscriptionDate
98723802	2017-11-01	2017-11-01
86905029	2017-11-01	2017-11-17
40935842	2017-11-01	2017-11-01
83184096	2017-11-01	2018-05-18
12771137	2017-11-01	2017-11-01

Figura 1

1.2 Séries temporais

Afim de simplificar as análises, agrupamos os dados por dia, somando o número de ocorrências (registros ou inscrições em plano premium) que houveram naquele dia. Abaixo nós podemos ver as séries temporais com a evolução do número de ocorrências diária. É visível que a maior parte dos planos premium foram criados em novembro, havendo um declínio após este período. A partir de março podemos ver leve um aumento na abertura destes planos, e este aumento aos poucos vai se intensificando.

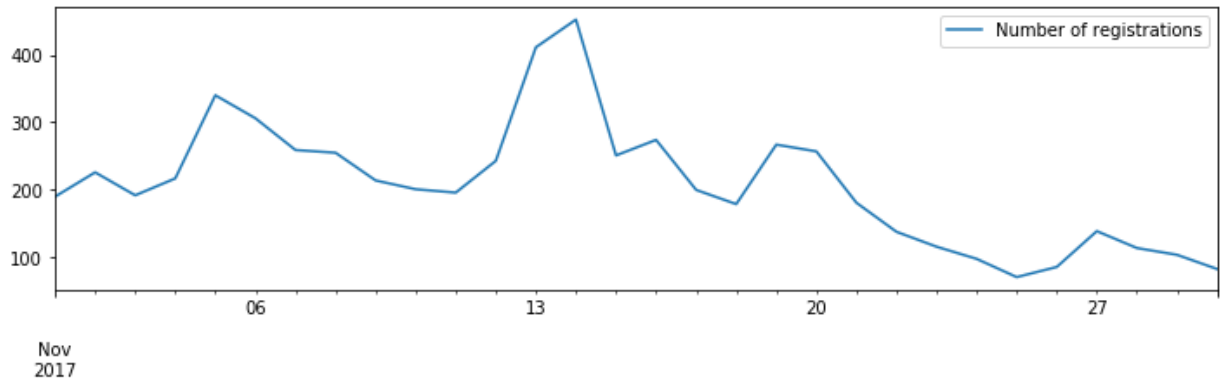

Figura 2

[1] https://github.com/felipebottega/Machine-Learning-Codes/tree/master/Desafio

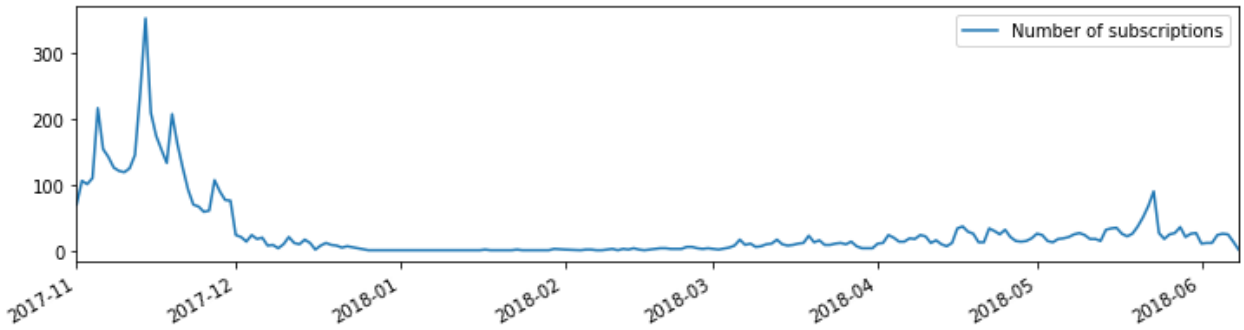

Figura 3

1.3 Probabilidades

Como estamos interessados no dias passados para a compra de um plano premium, começamos vendo a diferença, em dias, entre a data do registro e a data da compra do plano premium. Faremos isso para cada estudante.

Uma primeira abordagem ao problema seria simplesmente contar o número de vezes que ocorreu uma certa diferença e calcular a razão desta com o total. No código, temos a lista `probs` que satisfaz a equação abaixo:

$$\texttt{probs}[d] \approx \mathbb{P}(X = d),$$

onde X é o número de dias que um aluno aleatório levou para comprar o plano premium. Podemos ver no gráfico abaixo que mais ou menos 40% dos estudantes compra o plano premium nos primeiro dias após ter se registrado. Isso está de acordo com a observação anterior, de que a maior parte compra o plano em novembro, o mesmo mês que fizeram registro na plataforma.

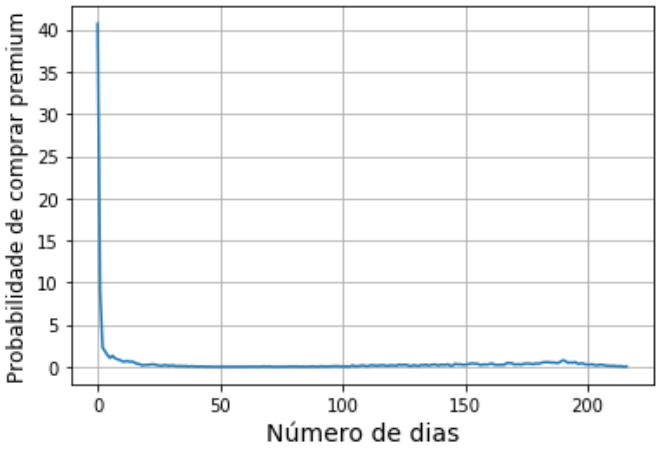

Figura 4

Uma limitação desta abordagem é o fato de que há alguns dias específicos com zero ocorrências, mesmo que os dias vizinhos não tenham zero ocorrências. Por exemplo `prob`[85] = 0,

mas isto é apenas um fator aleatório, não significa que ninguém nunca irá comprar um plano premium após 85 dias. Com mais dados certamente iríamos observar alguma ocorrência. De todo modo, afim de evitar este tipo de coisa, talvez seja mais interessante fazer um histograma e então obter a curva de densidade da função de probabilidade.

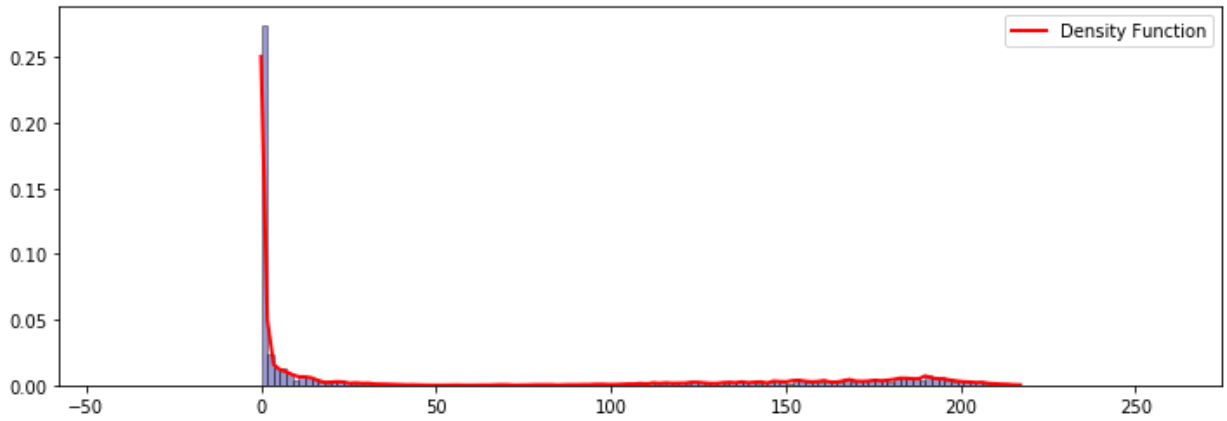

Figura 5

Vamos denotar por f a função de densidade de probabilidade e F a função de distribuição acumulada correspondente. Deste modo, a probabilidade de um estudante comprar um plano premium entre d_1 e d_2 dias é dada por

$$\mathbb{P}(d_1 \leq X \leq d_2) = \int_{d_1}^{d_2} f(x)\, dx = F(d_2) - F(d_1).$$

Usando a densidade obtida pela função *KernelDensity* (do scikit learn) acima, podemos começar a estimar probabilidades à vontade. Algumas probabilidades do usuário virar Premium após o cadastro em ranges de dias são mostradas abaixo.

De 0 a 3 dias $\rightarrow$ 54.32%
De 4 a 8 dias $\rightarrow$ 5.00%
De 8 a 30 dias $\rightarrow$ 9.85%
De 30 a 200 dias $\rightarrow$ 30.25%

2 Parte 2

2.1 Primeiro contato com o dataset

Agora vamos analisar os datasets da *BASE B*. Abaixo, temos uma visão geral de como são os datasets com os quais iremos trabalhar.

students

	UniversityName	SignupSource	StudentId	StudentClient	CourseName	RegisteredDate	State	City
0	PUC-RIO	Email	12970655	NaN	Administração	2012-05-29	Rio de Janeiro	NaN
1	UFF	Facebook	59873654	NaN	Direito do Trabalho e Segurança Social	2012-09-03	Rio de Janeiro	Rio de Janeiro
2	UNB	Facebook	3664695	NaN	Enfermagem	2012-09-10	Distrito Federal	NaN
3	UERJ	Facebook	15207697	NaN	Engenharia de Produção Mecânica	2012-09-05	Rio de Janeiro	Resende
4	UFU	Facebook	36988693	NaN	Engenharia Elétrica	2012-10-15	NaN	NaN

sessions

	StudentClient	StudentId	SessionStartTime
0	Website	12970655	2017-02-20 14:51:37
1	Website	12970655	2017-02-22 14:04:34
2	Website	12970655	2017-02-23 13:46:14
3	Website	12970655	2017-02-23 14:52:24
4	Website	12970655	2017-03-03 20:47:21

subjects

	StudentId	FollowDate	SubjectName
0	12970655	2015-09-12	Disciplina Integradora II
1	12970655	2015-09-12	Pesquisa Operacional
2	12970655	2016-06-07	Cálculo I
3	12970655	2015-09-12	Introdução à Administração
4	12970655	2015-09-12	Contabilidade Aplicada à Administração

questions

	StudentId	QuestionDate	StudentClient	QuestionSnippet
0	12970655	2013-09-04	NaN	O que é mais importante para um projeto: escop...
1	12970655	2013-09-04	NaN	Você pode ter um negócio/produto bem sucedido ...
2	12970655	2013-10-30	NaN	Custos com consultoria para planejar um projet...
3	12970655	2013-10-30	NaN	CAPEX: investimento em consultoria entra como ...
4	12970655	2013-10-30	NaN	Custos com consultoria para planejar um projet...

answers

	StudentId	AnswerDate	StudentClient	AnswerSnippet
0	12970655	2013-08-30	NaN	<p>Vou tentar <span style="text-decoration: un...
1	12970655	2013-09-04	NaN	<p>O centil divide algo em 10...
2	12970655	2013-09-05	NaN	<p>Ou será que o problema que esses pro...
3	12970655	2013-09-04	NaN	<p>Excelente pergunta! Abaixo estão as ...
4	12970655	2013-11-12	NaN	<p>Vamos supor que você se compromete a ...

fileViews

	StudentId	ViewDate	FileName	Studentclient
0	12970655	2017-02-23	Exercicios Resolvidos do Halliday sobre Rotaçã...	Website
1	12970655	2017-02-23	Exercicios Resolvidos do Halliday sobre Rotaçã...	Website
2	12970655	2017-05-25	CALCULO I	Website
3	12970655	2017-05-25	CALCULO I	Website
4	12970655	2017-05-25	CALCULO I	Website

premium_payments

	StudentId	PlanType	PaymentDate
0	12970655	Anual	2017-05-26
1	12970655	Anual	2018-05-26
2	12970655	Mensal	2016-07-21
3	12970655	Mensal	2016-08-21
4	12970655	Mensal	2016-09-21

premium_cancellations

	CancellationDate	StudentId
0	2016-05-05	34129668
1	2016-05-05	34129668
2	2016-05-20	82394932
3	2016-05-31	82942835
4	2016-06-01	85382416

Figura 6

2.2 Lifetime Value

Queremos analisar como o LTV varia com o tempo. O primeiro passo é saber o tempo que cada cliente passou pagando o plano premium. Iremos considerar este tempo em dias, como sendo a diferença entre a abertura e o cancelamento da conta premium. Este número será chamado de *student time*. A partir destas considerações, criamos o dataframe abaixo.

	StudentId	StudentTime	CancellationDate
0	43621	5	2017-06-16
1	51498	20	2017-03-31
2	105540	7	2018-05-03
3	344828	15	2017-11-30
4	440239	295	2018-05-06

Figura 7

Primeiramente, iremos calcular a média de compras de planos premium pelos estudantes anualmente. Este valor é dado pela fórmula

$$\text{número médio de planos por ano} = \frac{\text{soma das compras no ano}}{\text{número de estudantes que fizeram compras no ano}}.$$

Obsevamos que este valor está sendo atualizado semanalmente na análise feita aqui. A ideia é basicamente usar a fórmula acima em uma certa data e os valores dos 365 dias anteriores. Além disso, como o plano mensal e anual possuem valores diferentes, iremos produzir dois dataframes de valores médios e depois iremos tirar a média entre eles para ser o valor médio final.

Figura 8

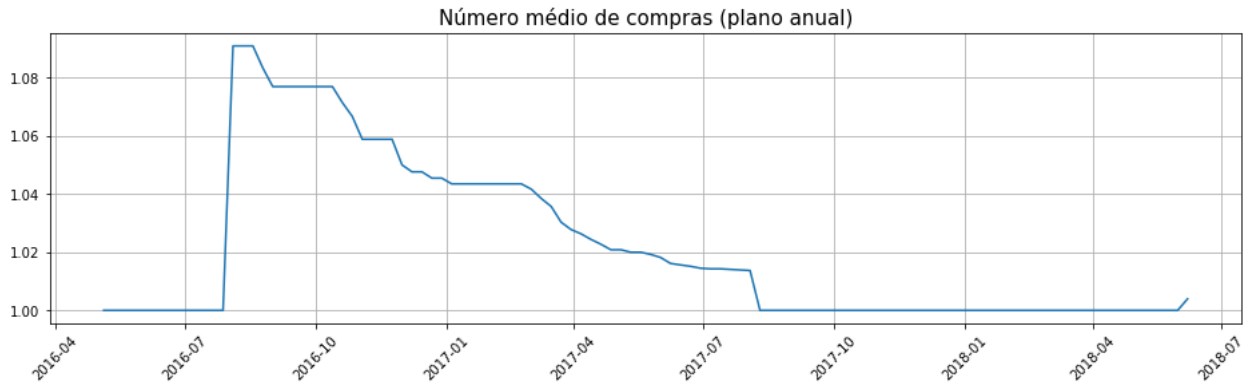

Figura 9

Agora podemos usar os valores de *churn rate* para deduzir o tempo médio que um cliente fica no plano premium. Note que o início do gráfico abaixo é atípico, mas é apenas um artefato, refletindo o fato de haver poucos estudantes para tirar médias após o primeiro ano. O LTV é então obtido a partir da seguinte fórmula

$$LTV = \text{valor médio} \times \text{número médio de compras} \times \text{churn rate de plano premium}.$$

Figura 10

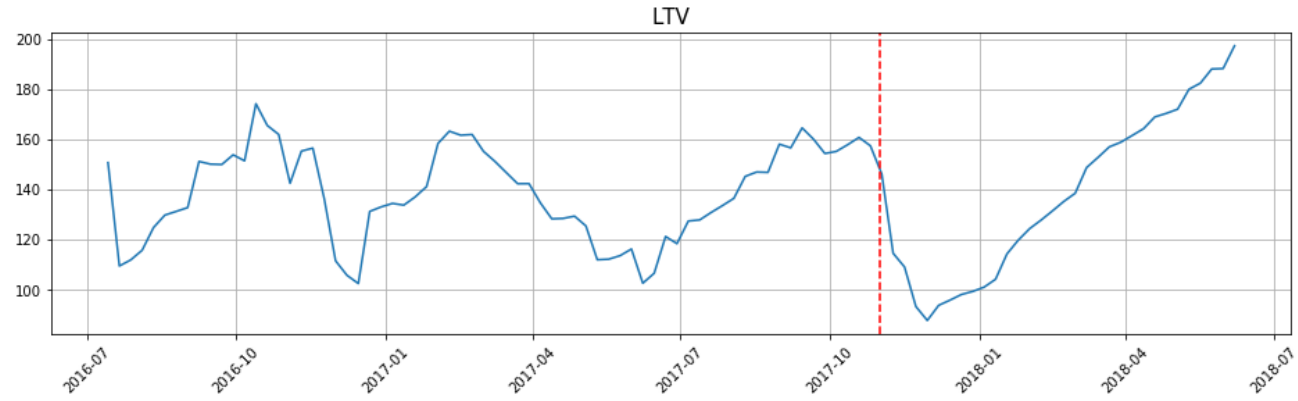

Figura 11

A respeito da questão 2: Assim que occoreu a mudança, o LTV sofreu uma forte queda seguida de um constante crescimento. Abaixo, podemos observar que a mudança de fato foi seguida de um aumento no número de cancelamentos e um aumento ainda mais evidente no número de compra de planos. Como a métrica usada aqui tem incrementos semanais, foi previso um pouco mais de tempo para observarmos o impacto da mudança, por isso a queda seguida do crescimento.

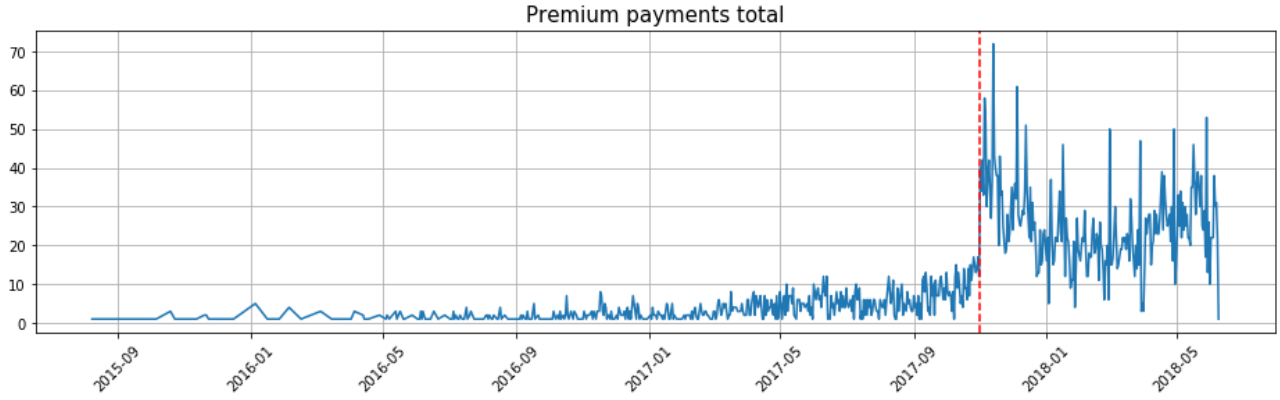

Figura 12

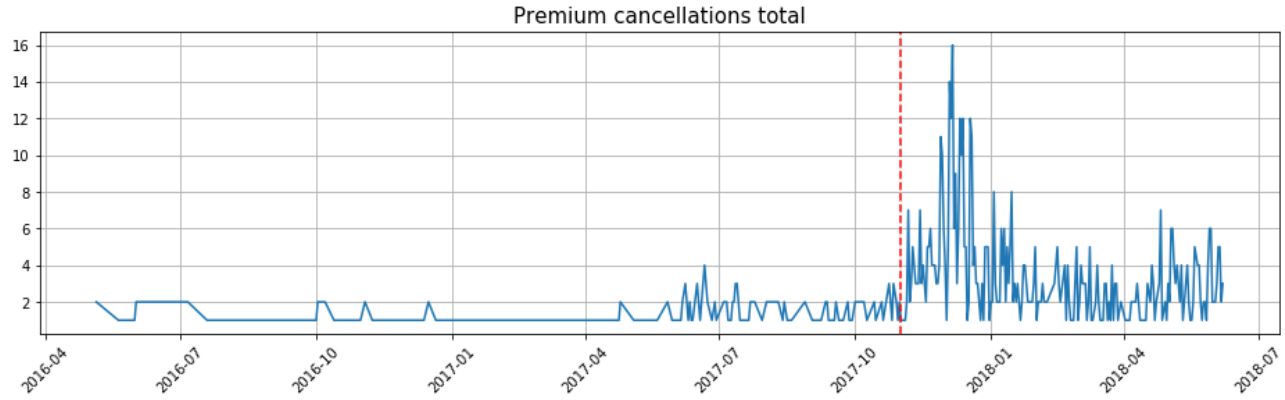

Figura 13

2.3 Experiência na plataforma

Sobre a questão 3: Depois de um tempo experimentando a plataforma Passei Direto, está claro que ela não se trata de apenas um repositório de materias didáticos, da mesma maneira que o Youtube não é apenas um repositório de vídeos. Ali há um espaço para as pessoas divulgarem seus trabalhos, de maneira personalizada. Para mim a analogia com o Youtube faz sentido, mas não precisa parar aí. Como acadêmico, sou sedento por artigos, e como a plataforma tem espaço para pós-graduandos, acho que seria natural também incorporar algo semelhante ao Google Scholar.[2] Mais precisamente, um ambiente com a possibilidade de divulgação de artigos, saindo um pouco do aspecto didático da plataforma e indo para o lado da pesquisa. Atualmente

[2] `https://scholar.google.com/`

eu (e mais um pequeno grupo) modero um grupo de Física e Matemática no Facebook.[3] Todos na moderação são atuantes tanto na questão didática quanto na de pesquisa, e nós ja tentamos criar um fórum brasileiro para uma troca mais séria de ideias acadêmicas, e não conseguimos. Também tentamos algo na plataforma Stack Exchange, e também acabou não vingando. Como erámos um bando de matemáticos e físicos amadores na internet, resolvemos deixar isso pra lá. Porém, a coisa poderia ser diferente se a iniciativa não fosse dada por meros acadêmicos amadores. Sem contar que o espaço na Passei Direto não seria para todas as disciplinas. Se tudo correr bem, é possível termos um Google Scholar brasileiro dentro da Passei Direto.

Um outro ponto que gostaria de fazer é o seguinte: em muitos lugares, tipo academias de dança, tem aquela coisa de fazer uma "aula experimental". Isto é, se você acaba de conhecer um local onde a ativadade x é ensinada, normalmente te deixam ter uma aula de graça para você sentir o gostinho de quero mais. Depois de me registrar e usar a Passei Direto por 5 minutos eu já atingi o limite do mês, não tive a chance de desfrutar da minha "aula experimental". Acho que seria interessante se no início houvesse algum tipo de regalia que não acabasse em 5 minutos, de modo que o usuário sentisse o que é ter um plano premium por um tempo razoável.

2.4 Dados demográficos

Para proceder com a **questão 4**, iremos levar em conta a questão regional. Primeiramente, note que toda a análise feita até agora levou em conta o Brasil como uma região única, mas poderíamos fazer as análises individualmente, uma para cada estado, por exemplo. Em vez disso, iremos ver alguns fatos gerais sobre os dados em mãos.

A primeira coisa a se observar é o número de estudantes registrados por estado, isto nos dá uma ideia geral do quanto a Passei Direto atua em cada lugar.

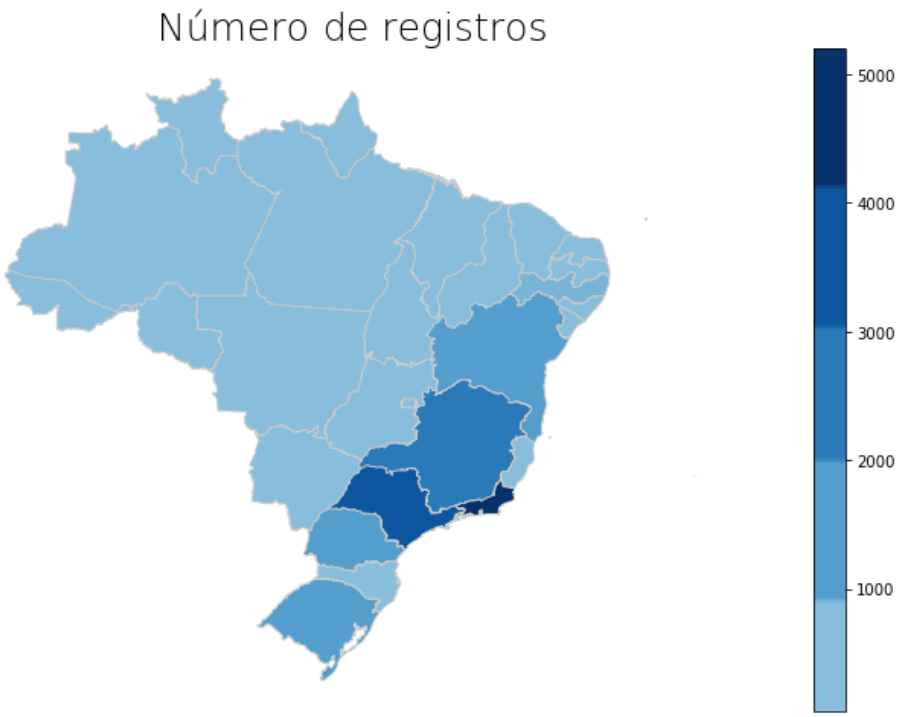

Figura 14

[3] https://www.facebook.com/groups/261270077250430/?ref=bookmarks

Não é surpresa que o Rio de Janeiro é o estado com o maior número de registros, seguido de São Paulo e depois Minas Gerais. Com exceção da Bahia, as regiões com mais registros são o sul e sudeste do Brasil. Certamente existe um grande número de estudantes no restante do país que podem ser potenciais clientes da Passei Direto. Há bastante espaço para crescimento ainda, basta saber como atuar nessas áreas de maneira mais efetiva. Mais interessante que o número total de registros é o número total de alunos com plano premium. Isto nos dá uma ideia de quais lugares estão dando mais retorno até agora.

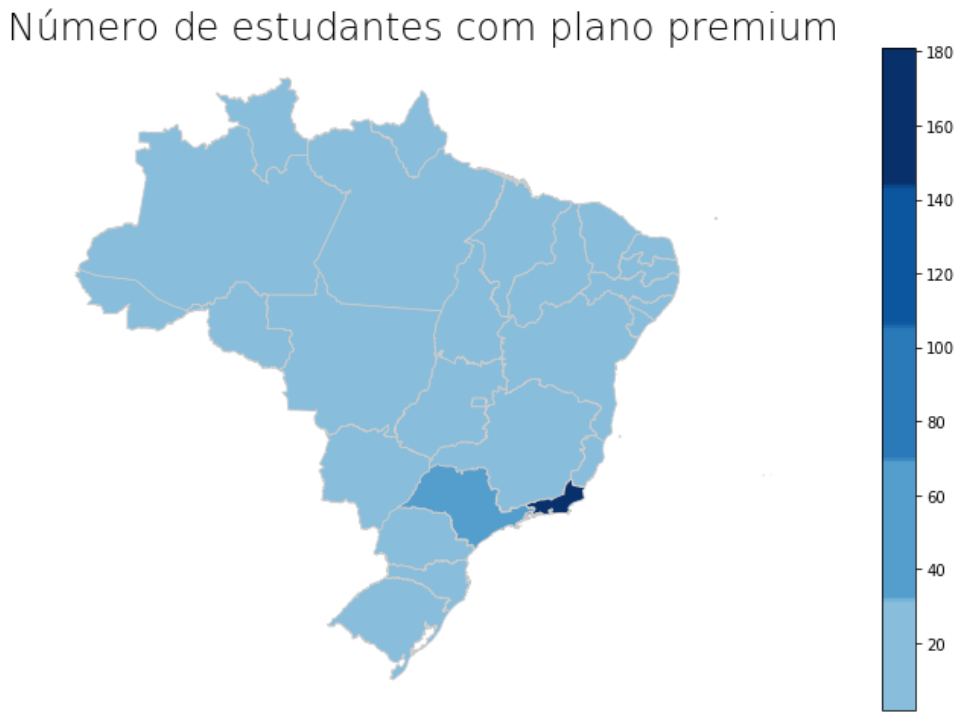

Figura 15

Os ganhos maiores com o Rio de Janeiro se devem bastante ao número de alunos registrados na plataforma, mas no quesito "fidelidade" (tempo usando o plano premium) as coisas mudam. Se tomarmos a média dos dias de plano premium por aluno, diversos estados começam a aparecer. Infelizmente isto pode ser apenas uma coincidência, pois na maior parte dos casos há poucos alunos para de fato termos uma média representativa. Por exemplo, Mato Grosso, que é o primeiro colocado, possui apenas 8 alunos. Como 6 destes alunos ficaram bastante tempo no plano premium, a média acabou ficando muito alta. Por um lado isto pode ser apenas consequência do baixo número de alunos para termos uma estatística representativa, por outro lado isto pode ser uma pequena amostra de um estado que tem mais tendência a permanecer mais tempo no plano premium.

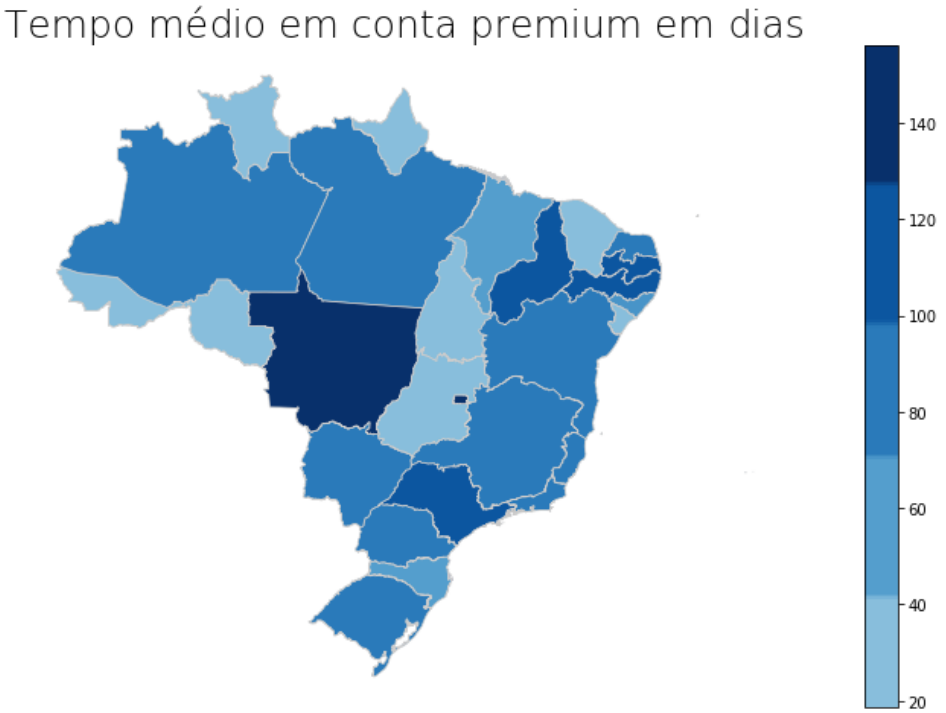

Figura 16

2.5 Cursos mais populares

Vamos ver apenas mais um tipo de segmentação, que será por popularidade de curso. Na figura abaixo nós podemos ver o número de estudantes com plano permium por curso. Os cursos mais relevantes (com mais de 20 estudantes pagantes) são: Engenharia de produção, Contabilidade / Ciências contábeis, Engenharia Civil, Administração e Direito. Com esse conhecimento ao nosso dispor, sabemos qual o público alvo que mais dá retorno. Com base nisso podemos dar um maior incentivo aos alunos destes cursos a adquirirem um plano premium.

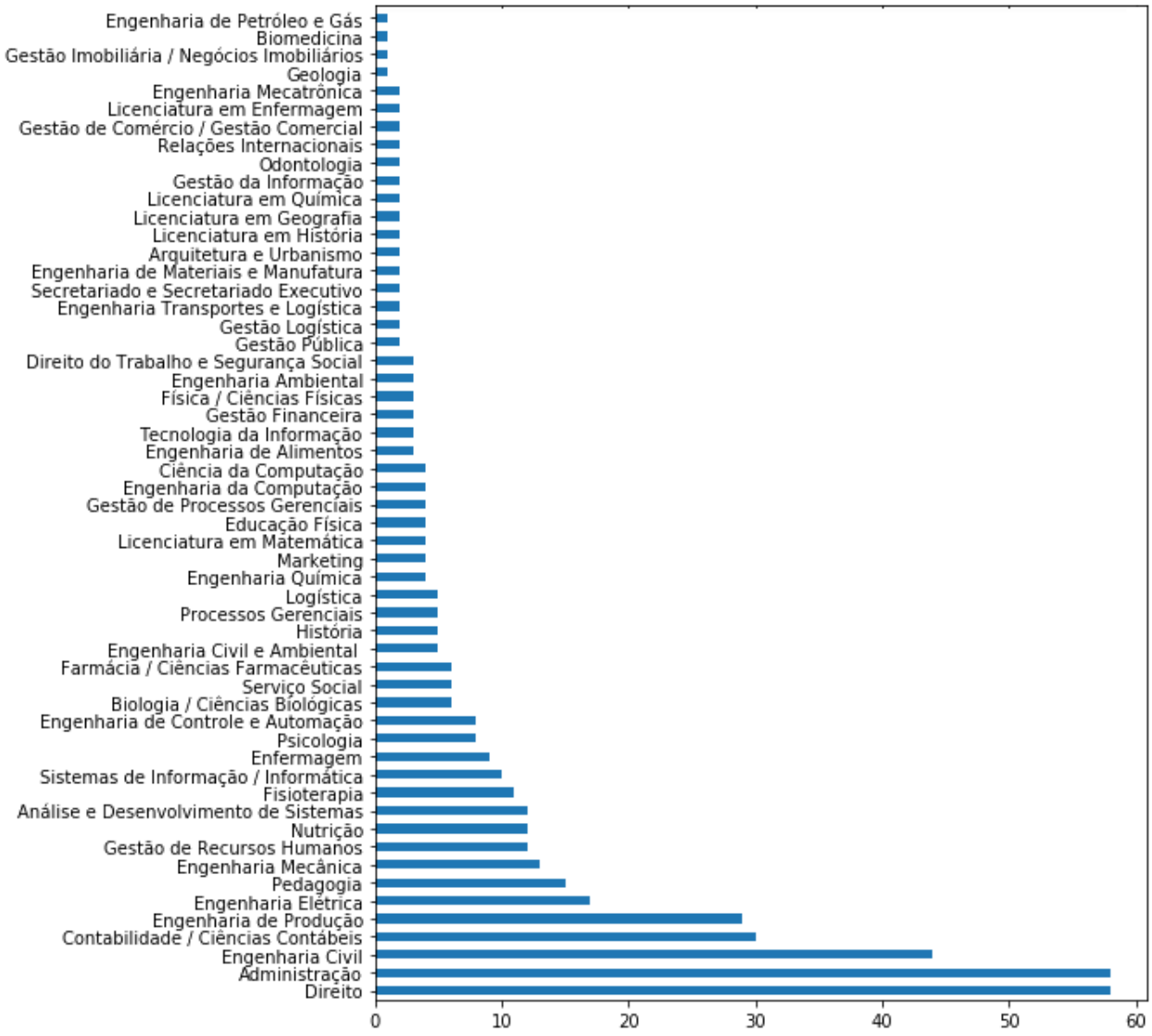
Engenharia de Petróleo e Gás
Biomedicina
Gestão Imobiliária / Negócios Imobiliários
Geologia
Engenharia Mecatrônica
Licenciatura em Enfermagem
Gestão de Comércio / Gestão Comercial
Relações Internacionais
Odontologia
Gestão da Informação
Licenciatura em Química
Licenciatura em Geografia
Licenciatura em História
Arquitetura e Urbanismo
Engenharia de Materiais e Manufatura
Secretariado e Secretariado Executivo
Engenharia Transportes e Logística
Gestão Logística
Gestão Pública
Direito do Trabalho e Segurança Social
Engenharia Ambiental
Física / Ciências Físicas
Gestão Financeira
Tecnologia da Informação
Engenharia de Alimentos
Ciência da Computação
Engenharia da Computação
Gestão de Processos Gerenciais
Educação Física
Licenciatura em Matemática
Marketing
Engenharia Química
Logística
Processos Gerenciais
História
Engenharia Civil e Ambiental
Farmácia / Ciências Farmacêuticas
Serviço Social
Biologia / Ciências Biológicas
Engenharia de Controle e Automação
Psicologia
Enfermagem
Sistemas de Informação / Informática
Fisioterapia
Análise e Desenvolvimento de Sistemas
Nutrição
Gestão de Recursos Humanos
Engenharia Mecânica
Pedagogia
Engenharia Elétrica
Engenharia de Produção
Contabilidade / Ciências Contábeis
Engenharia Civil
Administração
Direito
0
10
20
30
40
50
60

Figura 17