

Passei Direito

Teste – Vaga Data Scientist

Marcos Correa Martins Junior
Doutor em Física e Cientista de Dados

Rio de Janeiro, 25 de Novembro de 2019

Conteúdo

- Motivação e Objetivos
- Passei Direto
- Desafio
- Conclusão

Motivação

O que nos move?

Motivação

- O que nos move?
- Compartilhar conhecimento
- Empoderar o estudante
- Transformar o amanhã
- Conectar alunos e seus conhecimentos
- Proporcionar um aprendizado mais enriquecedor

Objetivo

A proposta da apresentação é mostrar o processamento e análise de dados no desafio para a vaga de Cientista de Dados.

Os dados para a análise correspondem a uma amostra dos dados gerados pela plataforma do Passei Direto.

O objetivo é explicar o processo de decisão diante das questões apresentadas.

Passei Direto

**Estude tudo.
Aprenda com todos.**

- Maior rede de estudos e compartilhamento do Brasil.
- Mais de 5 milhões de conteúdos compartilhados.
- Tecnologia e pluralidade.



Desafio

- O desafio está dividido em duas partes.
- A 1ª parte utiliza a Base A com apenas uma amostra de dados.
- A 2ª parte utiliza a Base B com diversas amostras de dados para uma análise mais completa do comportamento dos usuários.
- Os dados estão no formato .json (JavaScript Object Notation).
- O teste consiste de 6 questões: uma para a 1ª parte e cinco para a 2ª parte.



Desafio

- Para o processamento dos dados foi utilizado a linguagem de programação Python.
- Bibliotecas Pandas, Numpy e Matplotlib.
- Resumo dos dados no calc do linux para análise.



Questão 1

Dentre os usuários cadastrados em Nov/2017 que assinaram o Plano Premium, qual a probabilidade do usuário virar Premium após o cadastro em ranges de dias?

- A 1ª parte utiliza a Base A com amostra `premium_students.json`.
- Esta amostra de dados contém aproximadamente 6.2 mil registros no período de 2017-11-01 a 2017-11-30 contendo 3 atributos do usuário: Id, Data de Registro e Data da assinatura.
- Registros de usuários cadastrados no Passei Direto em Novembro de 2017 e que também assinaram o Plano Premium.

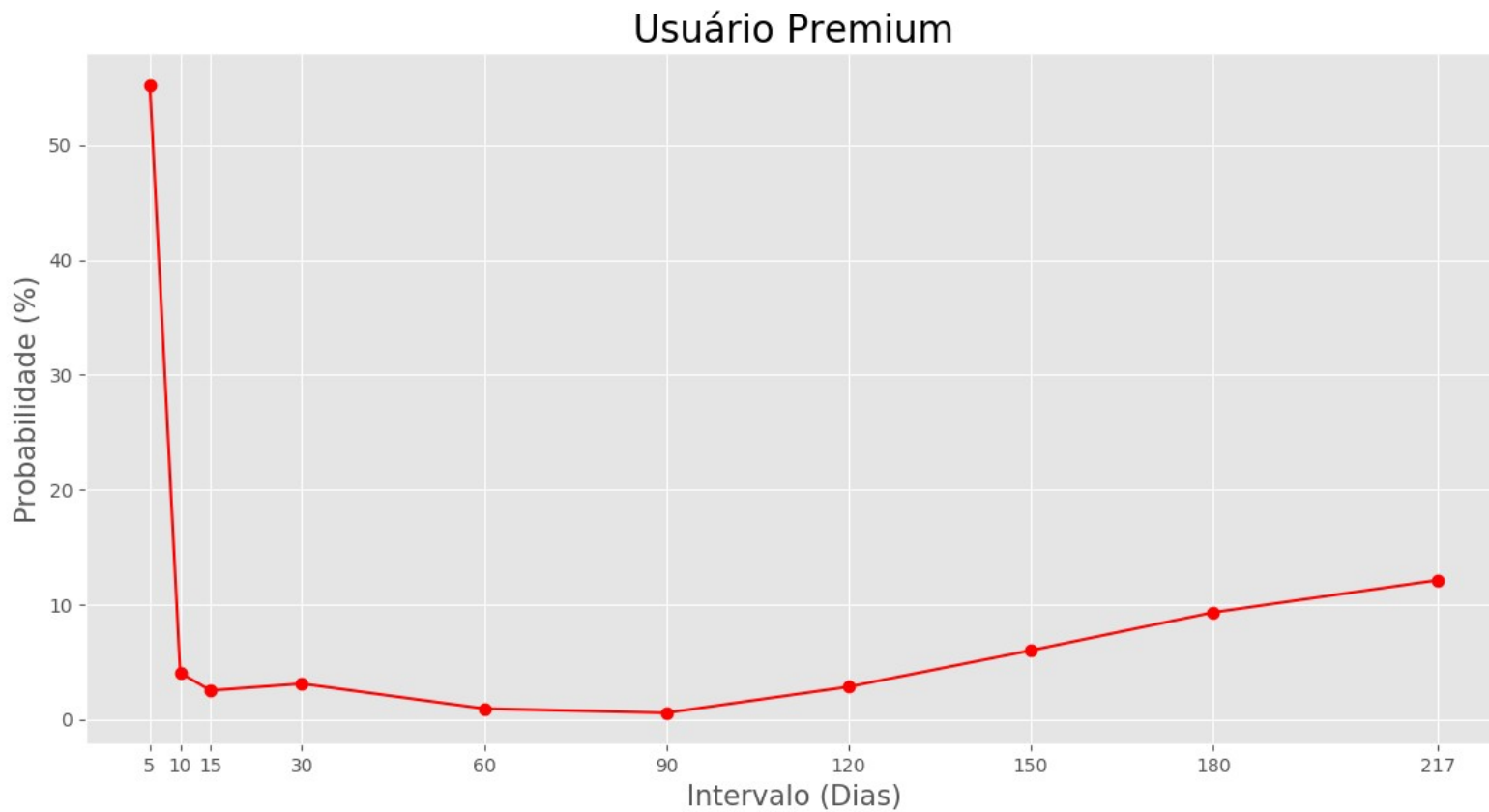
Desafio

Questão 1

- Foi calculado a diferença de dias entre a Data de Registro e Data da assinatura.
- Foram escolhidos alguns destes intervalos para realizar contagens de ocorrências.
- A probabilidade foi calculada como a razão do número de ocorrências para cada intervalo dividida pela soma total da amostra.

Desafio

Questão 1

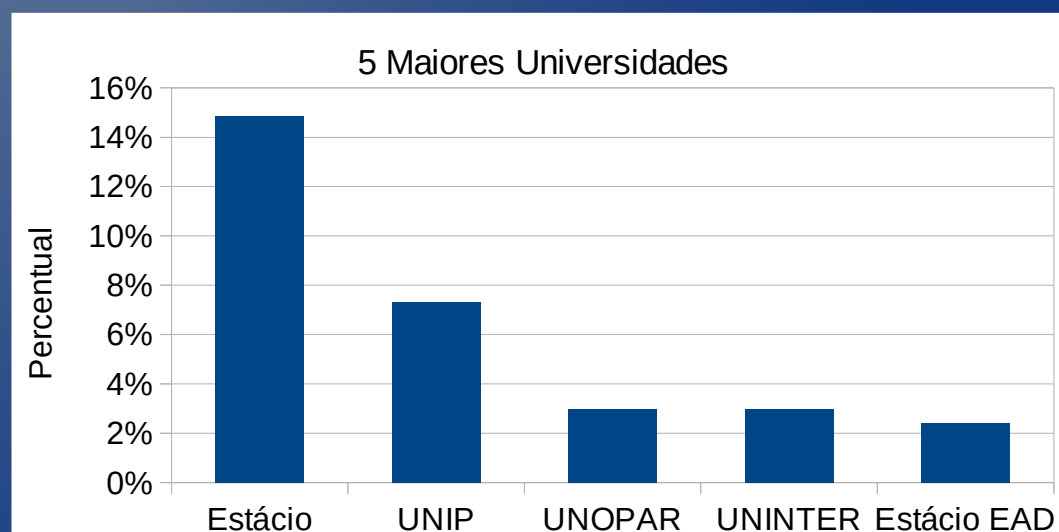
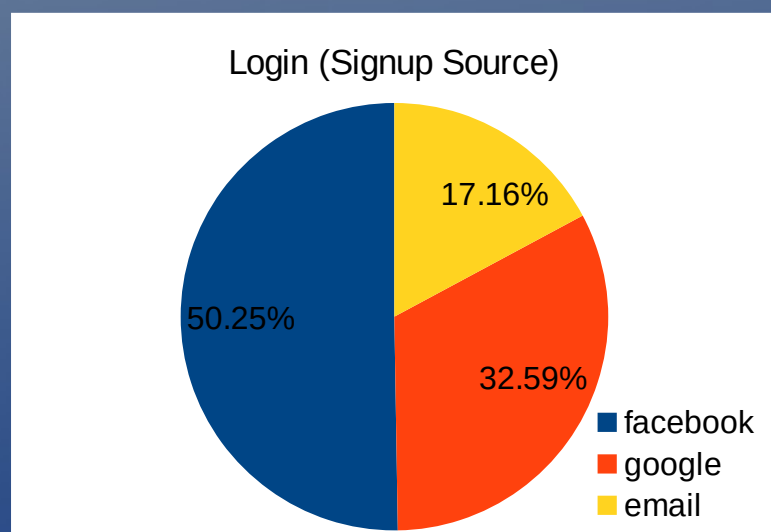


Questão 2

Análise geral dos dados e apresente as informações que julgar mais relevantes dessa base.

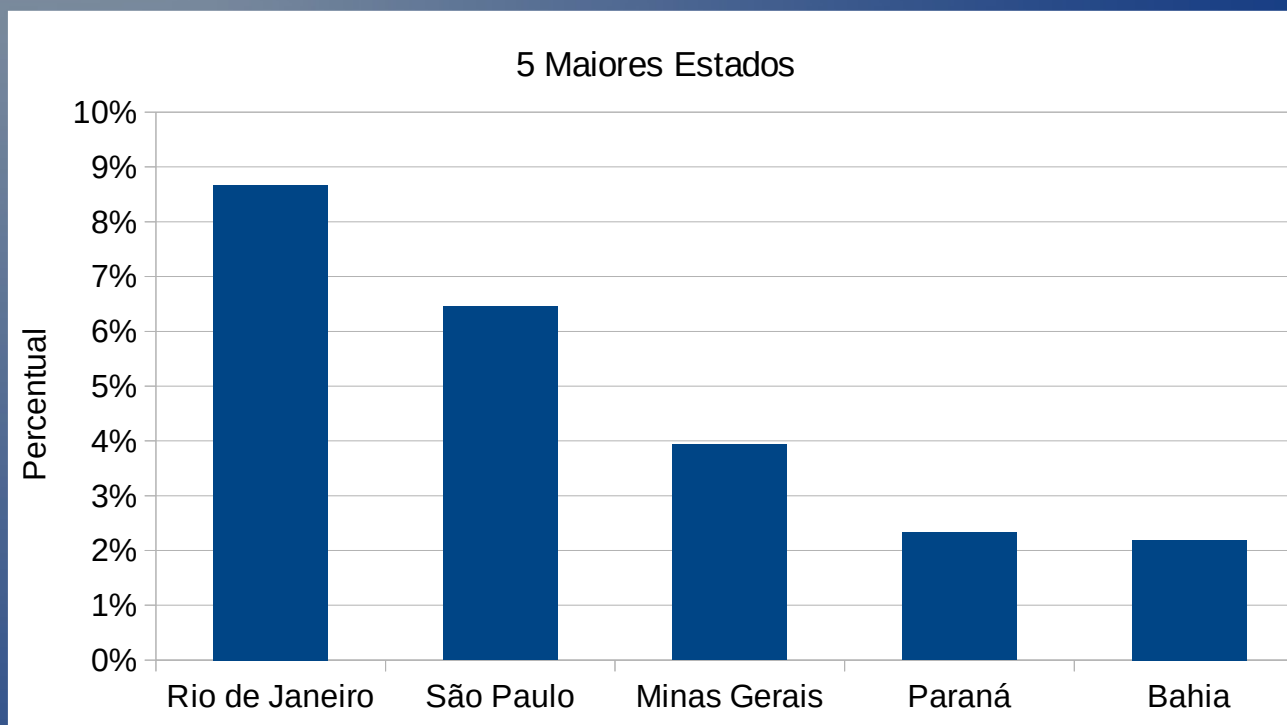
- A 2ª parte utiliza a Base B com algumas amostras de dados.
- Todas as amostras se relacionam através do Id do usuário.
- Amostras: students.json, sessions.json, subjects.json, questions.json, answers.json, fileViews.json, studyPlanViews.json, textBookSolutionViews.json, evaluations.json, premium_payments.json e premium_cancellations.json

- Amostra students.json contem 60 mil registros no período de 2012-05-29 a 2017-11-30 contendo 8 atributos do usuário: Id, Data de registro, Universidade, Curso, Estado, Cidade e Forma de login.



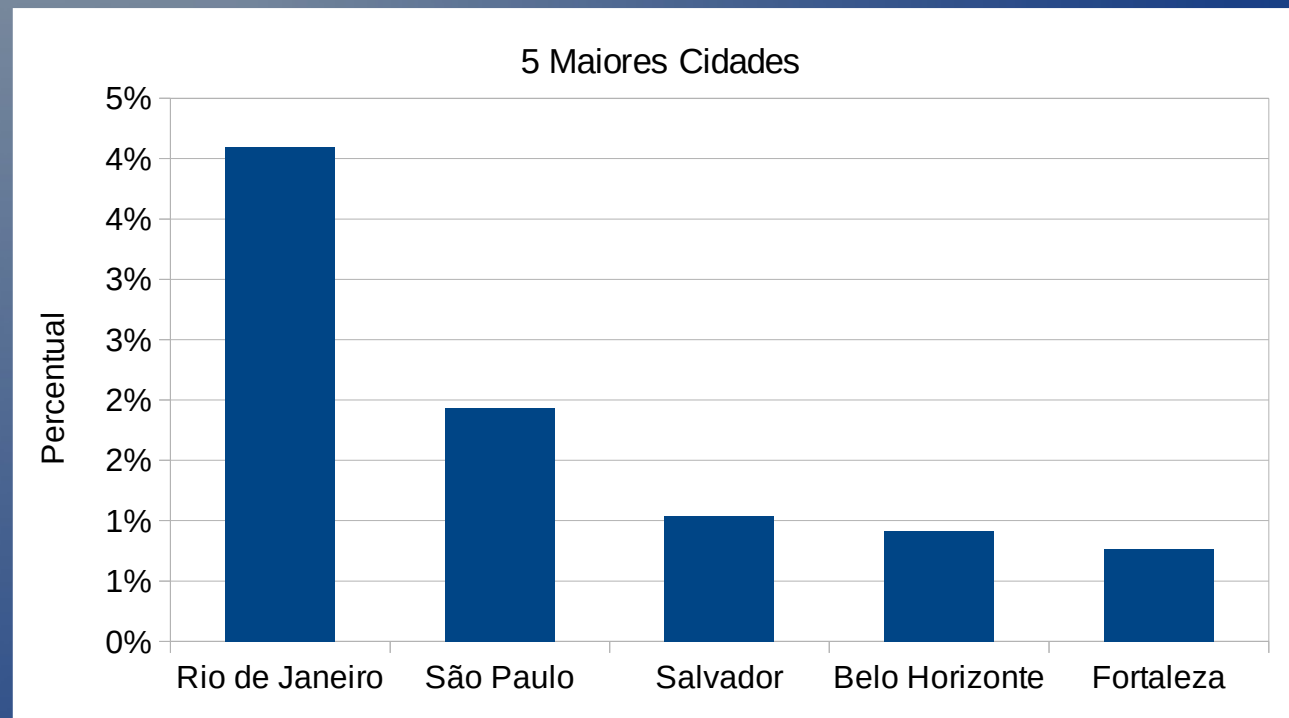
Desafio

Questão 2



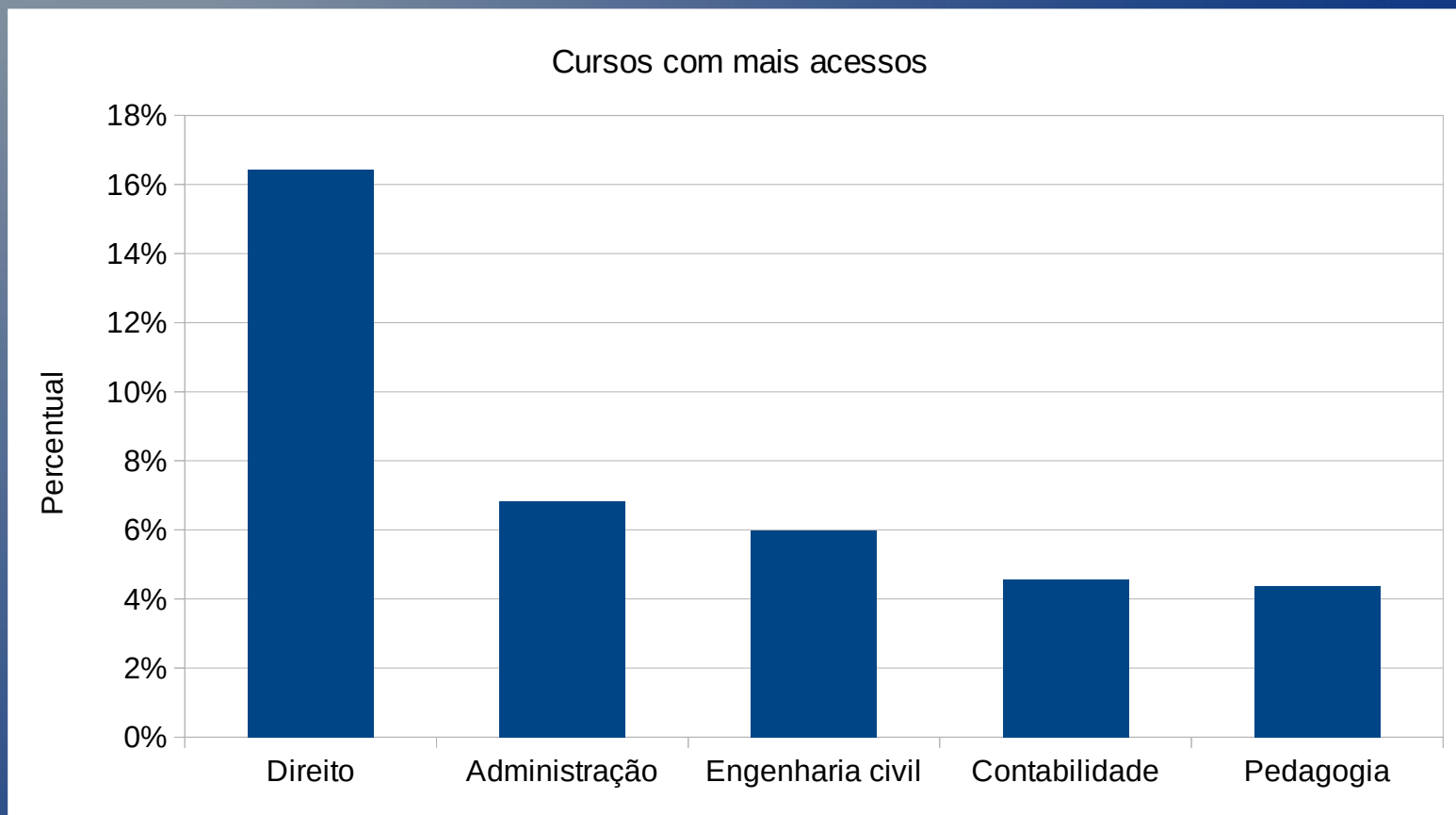
Desafio

Questão 2



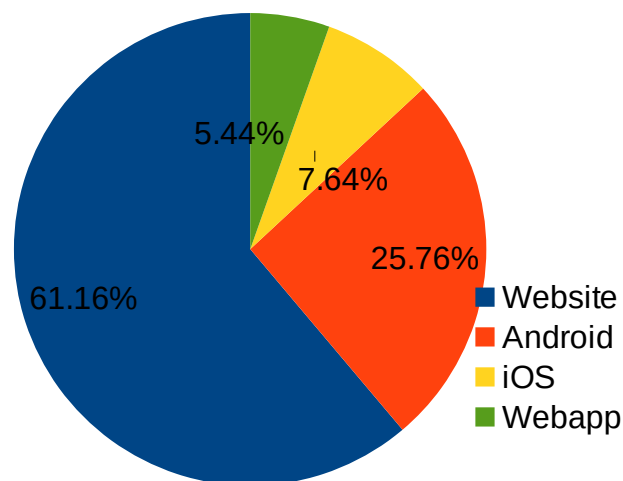
Desafio

Questão 2

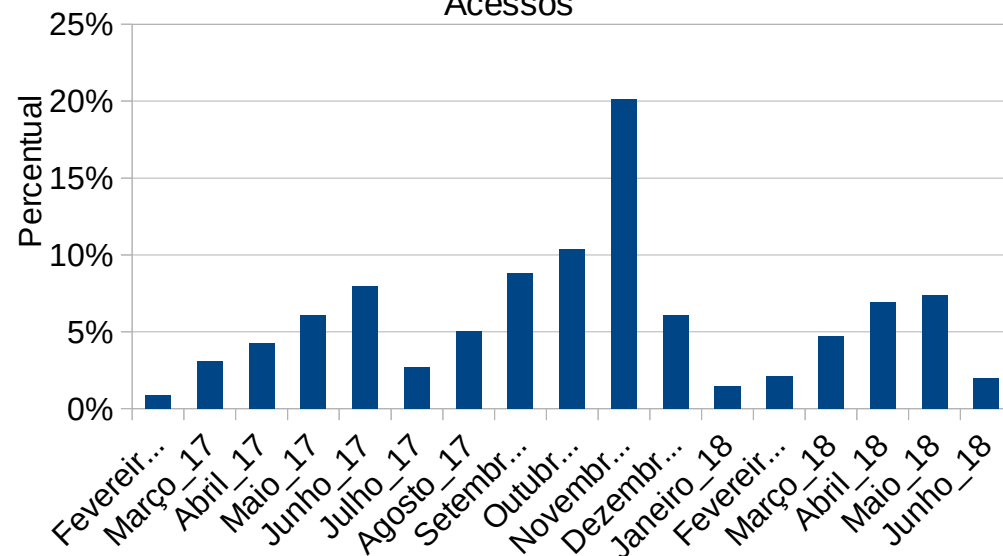


- Amostra sessions.json contem aproximadamente 1.4 milhão de registros no período de 2017-02-07 a 2017-11-30 contendo 3 atributos do usuário: Id, Data do início dos acessos e Meio de login.

Meio de Acesso (Student Client)



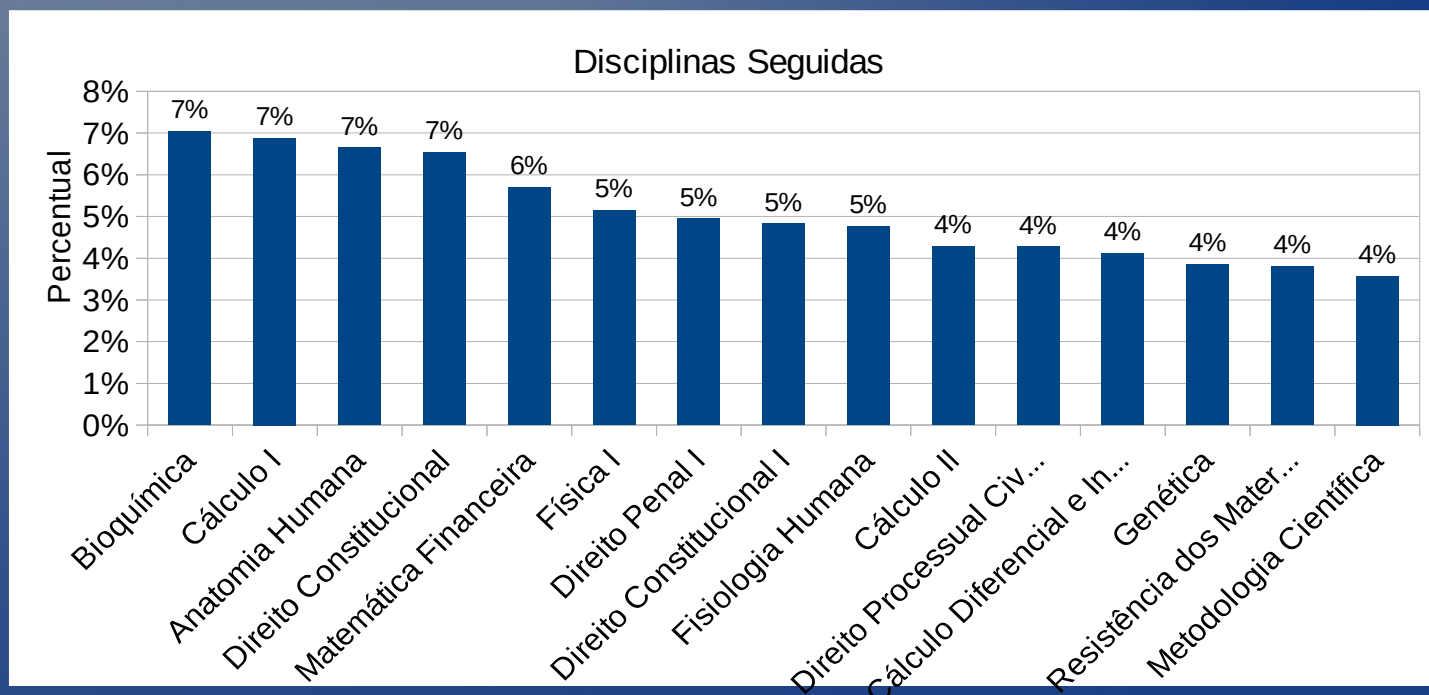
Acessos



Desafio

Questão 2

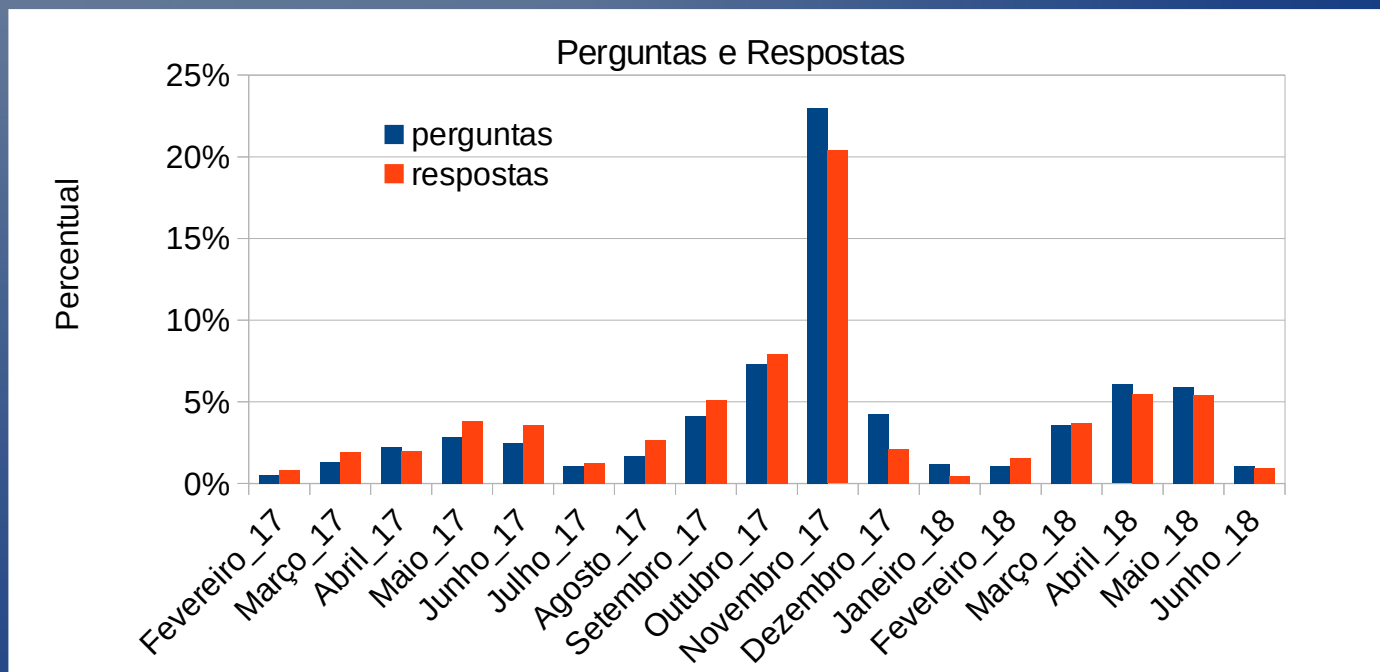
- Amostra subject.json contem aproximadamente 316 mil registros no período de 2017-08-31 a 2018-06-08 contendo 3 atributos do usuário: Id, Data de Seguidas e Disciplinas.



Desafio

Questão 2

- Amostra de Perguntas contem 3.9 mil registros no período de 2013-09-04 a 2018-06-07, amostra respostas, 7.5 mil registros de 2013-08-30 a 2018-06-07, 4 atributos do usuário: Id, Data das Perguntas/Respostas, Perguntas/Respostas e Meio de Login.



Desafio

Questão 2

- Amostra fileViews.json contem aproximadamente 3 milhões registros no período de 2016-11-24 a 2018-06-08 contendo 4 atributos do usuário: Id, Data de Acesso, Arquivo e Meio de Acesso.

Arquivos Visualizados	Nº de visualizações	Percentual
AV1	2303	0.08%
AV2	2169	0.07%
Avaliando o aprendizado	1804	0.06%
BDQ Prova	1541	0.05%
PLANEJAMENTO DE CARREIRA E SUCESSO PROFISSIONAL	1441	0.05%

Período	Nº de visualizações	Percentual
Set_17	291051	9.61%
Out_17	373772	12.34%
Nov_17	470164	15.52%
Dez_17	103643	3.42%
Jan_18	18098	0.60%

Desafio

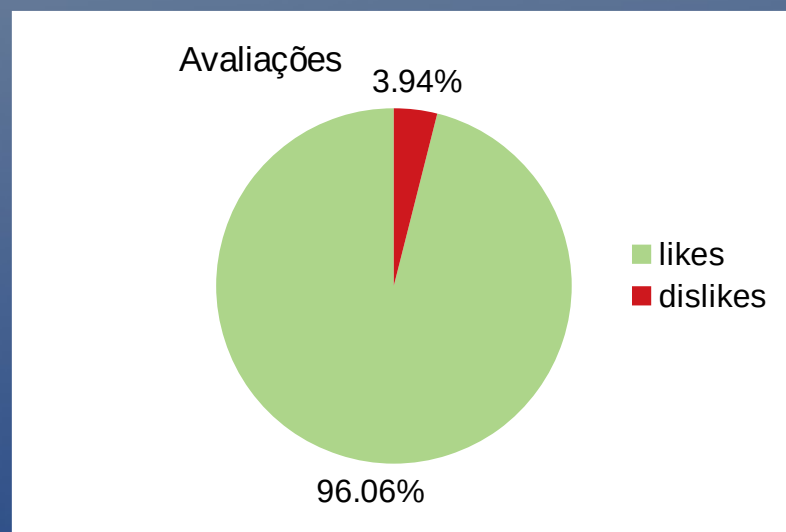
Questão 2

- Amostra studyPlanViews.json contem 7.7 mil registros no período de 2017-08-01 a 2018-06-07 contendo 4 atributos do usuário: Id, Data de Acesso, Tópico e Disciplina.

Tópicos	Nº de visualizações	Percentual
N/A	1814	23.41%
Resumos-de-quimica	241	3.11%
Bioquimica-estrutural-parte-i	233	3.01%
Conceitos-gerais-de-contabilidade	213	2.75%
Introducao-ao-estudo-do-direito-civil	166	2.14%
Dimensão-do-direito	165	2.13%

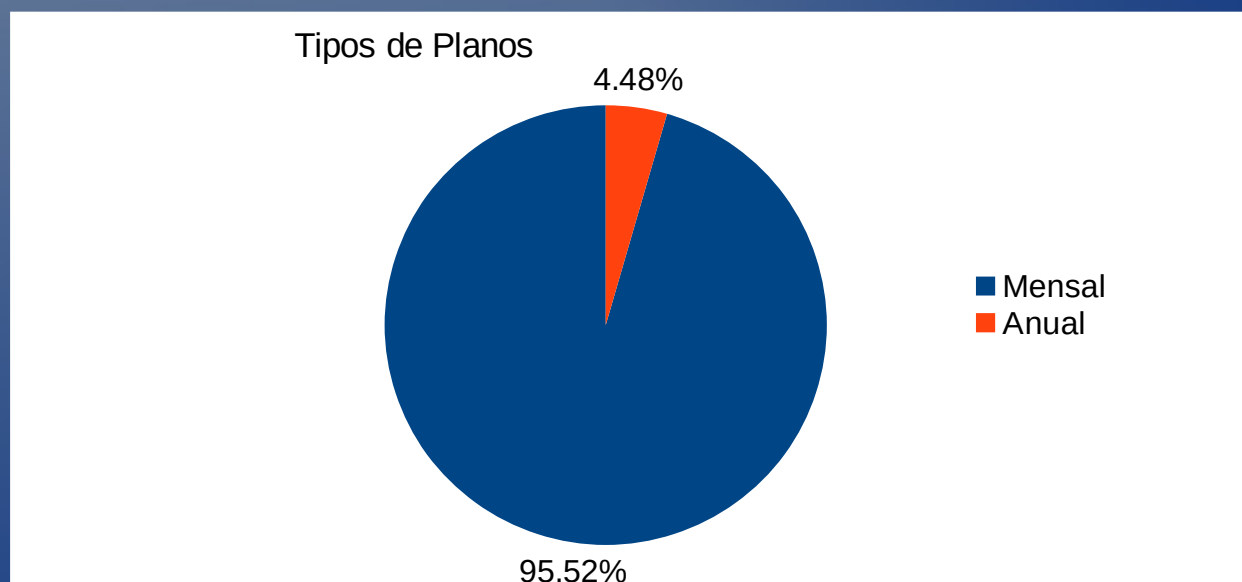
Disciplinas	Nº de visualizações	Percentual
Direito Penal I	622	8.03%
Bioquímica	554	7.15%
Cálculo I	551	7.11%
Direito Civil I	542	6.99%
Contabilidade I	394	5.08%
N/A	288	3.72%

- Amostra evaluations.json contem 106.8 mil registros aproximadamente, o período não é especificado, e ela contem 4 atributos do usuário: Id, Avaliação, Objeto da avaliação e Meio de acesso.



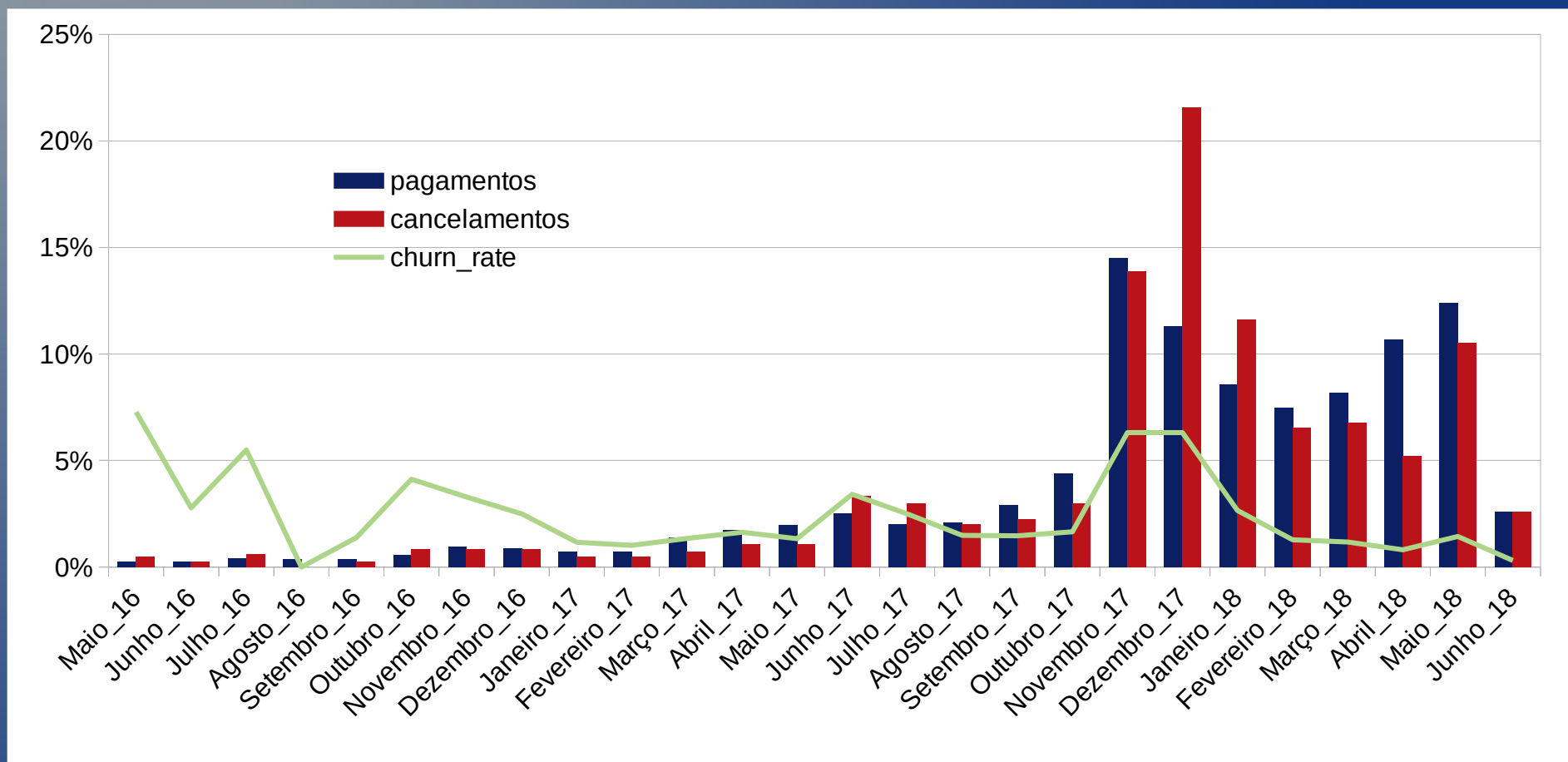
Item avaliado	Nº de avaliações	Percentual
File	3362	79.82
Question	470	11.16
Answer	293	6.96
Premium Content	34	0.81
Video	25	0.59
Comment	25	0.59
BookSolution	3	0.07

- Pagamentos contem 7276 registros de 2015-08-08 a 2018-06-08 contendo 3 atributos de usuário (Id, Data de Pagamento e Tipo de Plano), cancelamentos, 844 registros de 2016-05-05 a 2018-06-07, contendo 2 atributos do usuário: Id e Data de Cancelamento.



Desafio

Questão 2



Questão 3

Análise comparativa do comportamento dos usuários não Premium e dos usuários Premium. Que tipos de ações podemos direcionar para usuários não Premium fazerem com o objetivo de termos um maior número de assinantes?

Fazendo uma análise comparativa das amostras de dados `students.json`, `premium_payments.json` e `evaluations.json`, verifica-se que os comportamentos de usuários premium e não premium são semelhantes em relação a origem, Estados e cidades, a universidade e a avaliações.

Desafio

Questão 3

Algumas ações podem direcionar usuários para o plano premium:

- 1) Aplicação de novas restrições, levando em conta que houve um aumento do número de usuários premium devido as restrições de conteúdo feitas em novembro de 2017;
- 2) Investir em conteúdos;
- 3) Criar relacionamento e atendimento personalizados, ações de engajamento e interações;
- 4) Buscando novas estratégias de marketing, como o marketing digital, diversidade de planos e novas tecnologias.

Questão 4


Em Novembro de 2017 fizemos uma grande mudança no PD: o Content Restriction. Os usuários não Premium passaram a poder consumir no máximo 3 arquivos diferentes por mês. Diante dessa mudança, qual passou a ser o Lifetime Value (LTV) dos usuários Premium a partir de Novembro de 2017?

O Lifetime Value (LTV) é definido como o Tempo de Retenção (TR) do usuário vezes o Ticket médio que neste caso seria de 29.90. O TR é calculado como o número de cancelamentos vezes a quantidade de meses dividido pelo total de cancelamentos. Desta forma pode-se separar a amostra em dois períodos: Antes de Novembro de 2017 (AN) e Depois Novembro de 2017 (DN).

Desafio

Questão 4

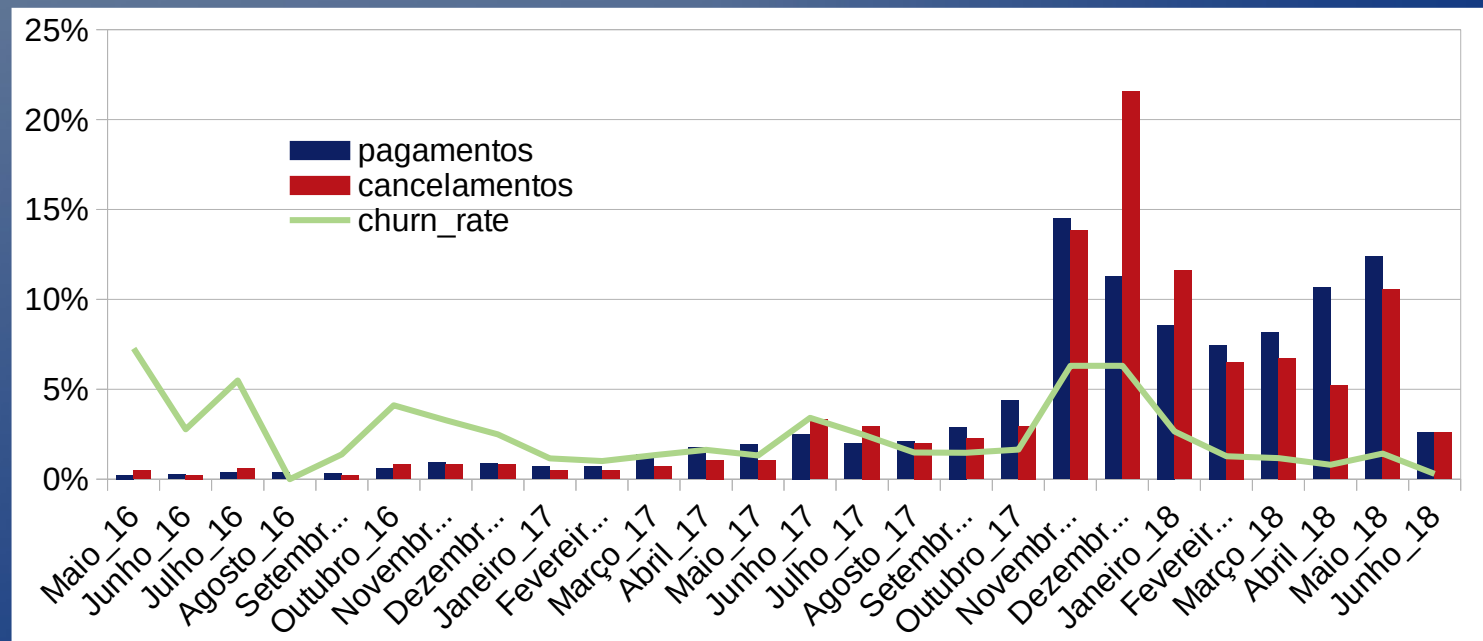
Para o período depois de Novembro de 2017, o TR e o LTV obtidos são iguais a 21.53 e 657.80, respectivamente. A mudança representou um percentual de quase 70% no LTV.

Período	LTV	
Antes de Novembro/17	388.70	
Depois de Novembro/17	657.80	 69.23%

Questão 5

Entre os usuários que "Churnaram" solicitando ativamente o cancelamento do Plano Premium, o que está fortemente correlacionado com o cancelamento?

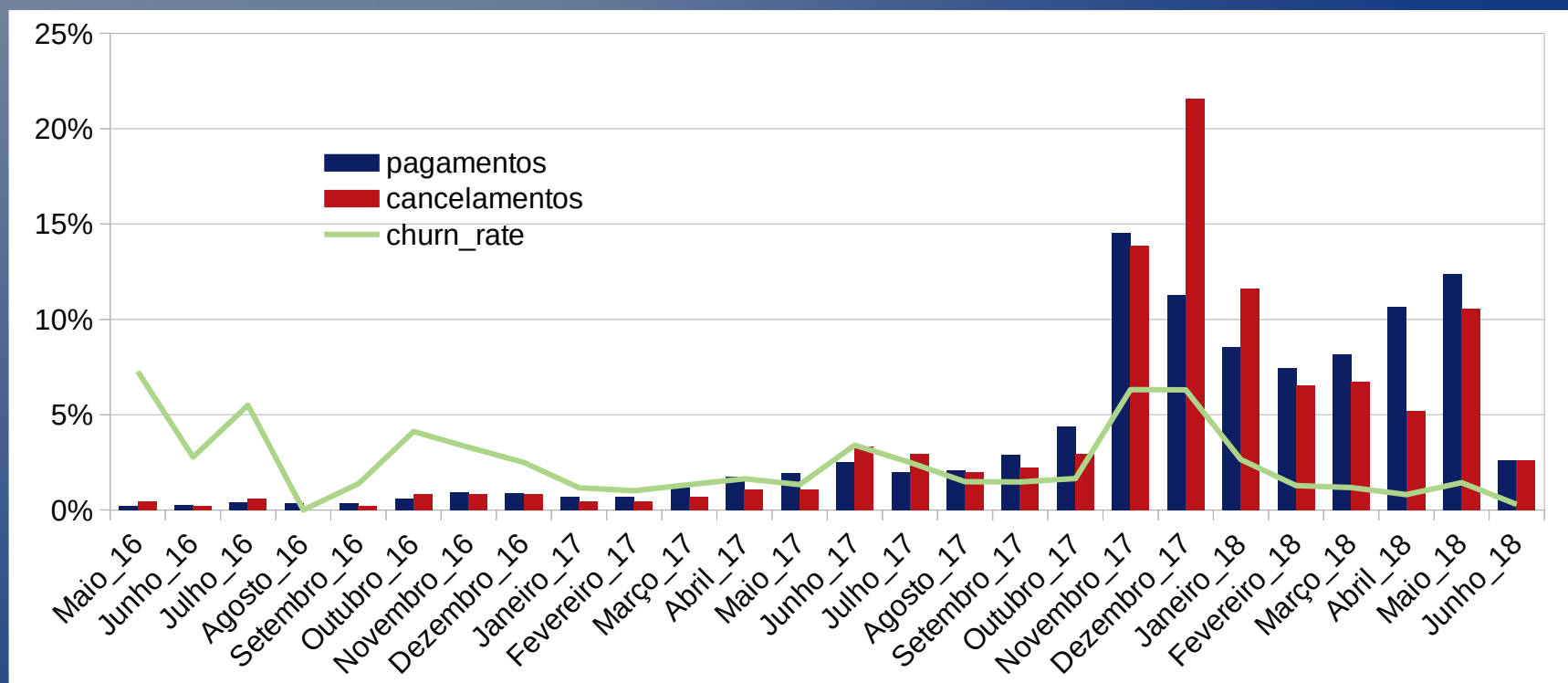
- Aumento significativo no percentual de cancelamentos dos usuários a partir de Novembro de 2017 que está fortemente correlacionado com o aumento no percentual de pagamentos;



Desafio

Questão 5

- A taxa de “churn” também possui um comportamento recíproco ao percentual de cancelamentos. Existe forte correlação com períodos de inatividade, períodos de férias em Dez, Jan, Fev e Jul.



Questão 6

Quais são as 5 maiores universidades no Passei Direto? E quais são os principais tipos de Arquivos consumidos em cada uma delas?

Universidade	Quantidade	Percentual
Estácio	8919	14.87%
UNIP	4392	7.32%
UNOPAR	1784	2.97%
UNINTER	1780	2.97%
Estácio EAD	1446	2.41%

Universidades	Arquivos mais consumidos
1 Estácio	Provas, resumo, planejamentos e exercícios
2 UNIP	Apanhados, resumos, disciplinas online, provas, atividades práticas supervisionada
3 UNOPAR	Provas presenciais, relatórios de estágio curricular obrigatório, produção textual inte
4 UNINTER	Atividade pedagógica on-line (APOL), provas e metodologias

Questão7

Considerando o seu conhecimento da plataforma do Passei Direto e a análise dos dados disponibilizados nesse teste, que oportunidades você vê para melhorarmos a experiência dos nossos usuários? Ofereça insights explorando o como e o porquê.

- A partir dos resultados da questão 1 e sobre o conhecimento da plataforma, considerando do intervalo de 10 dias para o intervalo de 120 dias, onde a probabilidade do usuário virar Premium se mantém entre 4% e 2%, uma oportunidade para melhorar a experiência dos usuários seria a recomendação de conteúdos individuais por um preço específico, direcionado para este perfil de usuário, baseando a recomendação no seu histórico de buscas e no histórico de buscas de perfis semelhantes;

Desafio

Questão 7

- A partir dos resultados e análise da questão 2, vale destacar o segundo lugar respectivamente do Estado de São Paulo e sua capital São Paulo no ranking de origem dos usuários. São Paulo é o Estado mais rico, tem as principais universidades e tem uma população maior (aproximadamente o dobro da população do RJ), apresentando um potencial muito grande a ser explorado;
- Investir na retenção dos clientes, fazendo com que ele veja a plataforma como a escolha certa, investindo em conteúdos, criando relacionamento e atendimento personalizados, ações de engajamento e interações. Desta forma, diminuir as taxas de “churn” ou desistência de planos, buscando novas estratégias de marketing, como o marketing digital, diversidade de planos e novas tecnologias.

Obrigado!