UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE ESTATÍSTICA

Análise exploratória de dados no Local Search hagah

Autor: Diego Barbosa de Souza

Orientador: Professor Dr. Danilo Marcondes Filho

Porto Alegre, 18 de dezembro de 2013.

Resumo

Com o desenvolvimento contínuo na área da computação, o armazenamento de dados não é mais um problema. Antigamente, era impossível, por exemplo, que uma empresa controlasse perfeitamente o seu estoque e volume de vendas diário de cada produto. Hoje, isso pode ser medido com alto nível de detalhamento, extraindo estas informações por filial, faixa de horário e forma de pagamento. Mas apenas armazenar não traz soluções às empresas. Este trabalho se propõe a transformar essa massa de dados em inteligência para o local search hagah, empresa do pilar digital do Grupo RBS, através do uso de técnicas exploratórias estatísticas. O objetivo central é entender de maneira mais aprofundada os problemas que o negócio enfrenta, como a inadimplência de seus assinantes, para então serem tomadas ações corretivas que preservem o faturamento.

1. Introdução

O *hagah* é um *local search* criado em 2006 pelo Grupo RBS, visando o crescimento no pilar digital da empresa, que sempre teve foco na comunicação por TV e rádio. O site facilita a procura dos seus usuários, apontando de forma rápida e fácil os produtos e serviços na região; desta forma, potencializa as chances de um estabelecimento ser encontrado, aumentando sua demanda. O site atua como "meio de campo" entre consumidores e empresas nos três estados da região Sul e São Paulo.

Para melhorar esta atuação, a área de marketing precisa escolher as direções certas, e para isto, a empresa confia este trabalho para a célula de Planejamento e Inteligência de Mercado, onde procuramos respaldar as ações através de análise aprofundada de grandes bancos de dados. Em resumo, devemos seguir os seguintes passos (MALHOTRA, 2012): identificar, coletar, analisar e disseminar informações objetiva e sistematicamente, assessorando a gerência na tomada de decisões. Também é importante ressaltar a preocupação de grandes empresas em dominar suas bases de dados, tendo BIs (áreas de Business Inteligence) que explorem e identifiquem as variações minuciosas do negócio. Por exemplo, se uma safra de vendas se apresentou muito mais inadimplente que a média, deve-se analisar se ela está concentrada em alguma forma de pagamento, em algum canal de venda, em alguma categoria de estabelecimentos, cidade/bairro, etc.

Com o crescente desenvolvimento tecnológico, temos armazenamento diário de grande quantidade de dados. Com a posse dos dados, precisamos utilizar as técnicas corretas, para alcançarmos as respostas que precisamos para prosperar o negócio, analisando tanto o perfil do usuário (quem acessa, onde mora, como e para onde faz suas buscas, etc.), quanto o perfil dos estabelecimentos anunciantes (tipo de serviço anunciado, local dos estabelecimentos, tipo de plano de anúncio, etc.). Juntando as duas pontas, a atuação do hagah como local search certamente terá evolução.

O objetivo da análise é entender a base de dados do site, transformando os dados em inteligência, buscando informações que expliquem o comportamento dos anunciantes do *hagah*, através da utilização de técnicas estatísticas exploratórias para buscar identificar os perfis abaixo:

- Anunciantes do site - entender quem procura por divulgação no site, quais as categorias e em que regiões atuam;
- Buscas no site - como os visitantes do *hagah* interagem com a ferramenta, quais estabelecimentos são mais visitados, etc.;
- Através de análise exploratória dos dados, apontar as principais questões que o hagah deve ter atenção, principalmente no que se refere ao cancelamento ou inadimplência dos assinantes.

Com estes três pontos esclarecidos, teremos material para balizar as estratégias para uma futura mudança no produto, que avançará para o âmbito nacional, ampliando para o Rio de Janeiro, Minas Gerais já em 2014. E além de pensar na expansão, teremos dados consistentes para rever as estratégias tanto no âmbito comercial (para quem vender, em que região, em que época do ano, etc.) quanto no âmbito de rentabilização da carteira de anunciantes (entender quem pode cancelar o anúncio, quem tem poder econômico para trabalhar com um plano mais caro, etc.).

O restante da análise está organizado da seguinte forma: a seção 2 apresenta uma breve descrição sobre técnicas de data mining, focando nas técnicas estatísticas de sumarização de dados. A seção 3 apresenta a metodologia do trabalho. A seção 4 apresenta o resultado das análises dos bancos de dados da empresa *hagah*. A seção 5 apresenta as conclusões do estudo.

2. Técnicas exploratórias estatísticas

Esta seção apresenta as etapas e técnicas estatísticas exploratórias que serão utilizadas para cumprimento dos objetivos propostos neste trabalho.

Para o estudo que será realizado, usaremos técnicas para relacionar dados envolvendo estatística descritiva (análise de parâmetros estatísticos clássicos, tais como média, mediana, desvio padrão, etc.), técnicas de inferência (testes de hipóteses) e técnicas gráficas exploratórias (gráficos de barra, boxplot).

- Teste qui-quadrado de associação (AGRESTI, 2000): A estatística Qui Quadrado calcula o total de desvios entre o nº de ocorrências observadas e esperadas, pondera sua probabilidade de ocorrência segundo uma distribuição qui-quadrado com n graus de liberdade. É apropriado para testar a hipótese nula de não associação entre as categorias. As hipóteses são dadas por H_0: não existe alguma associação entre as categorias; ou H_1: existe associação entre as categorias. O teste calcula primeiramente uma estatística utilizando a fórmula abaixo:

$$\chi^2 = \sum_{i=1}^{l}\sum_{j=1}^{c}\frac{\left(A_{ij} - E_{ij}\right)^2}{E_{ij}},$$

em que:

Aij = frequência real na i-ésima linha, j-ésima coluna
Eij = frequência esperada na i-ésima linha, j-ésima coluna
l = número de linhas e c = número de colunas

Valores relativamente grandes da estatística levam à rejeição de H_0, isto é, de que existe dependência entre linhas e colunas da tabela de dados.

- Teste de Kruskal-Wallis (UPTON & COOK, 2008): É útil para decidir se k amostras independentes provêm de populações diferentes. Desta forma, se procura descobrir se a amostra representa variações casuais ou diferenças efetivas entre as populações. Cada observação é substituída pelo seu posto, ou seja, ficam dispostas em ordem crescente, onde o posto 1 fica com a observação de menor valor e o posto N com a de maior valor. Após, calcula-se a soma dos postos para cada amostra, e então, o teste determina se as somas são suficientemente diferentes que seja improvável que ambas se originem da mesma população. As hipóteses são dadas por H_0: Todas as populações possuem funções de distribuição iguais; ou H_1: Uma ou mais funções são diferentes. Quando o teste aponta resultado significativo, quer dizer que pelo menos uma das amostras é significativamente diferente das demais. A estatística é dada por:

$$H = \frac{12}{N(N+1)}\sum_{j=1}^{k}\frac{R_j^2}{n_j} - 3(N+1),$$

onde R_i é a soma dos postos da amostra i, n_i é o tamanho da amostra i, e N representa o total geral de observações. A estatística H tem distribuição aproximada χ^2 com gl = k-1 graus de liberdade (desde que o tamanho das k amostras não sejam muito pequenos). Valores relativamente grandes de k levam à rejeição de Ho.

- Boxplot (TUKEY, 1977): O gráfico boxplot indica a simetria dos dados através de um retângulo construído com os quartis. Utiliza cinco medidas estatísticas: máximo, mínimo, mediana, primeiro quartil e terceiro quartil. Com isto, consegue mostrar a dispersão e salientar os outliers, o que facilita uma análise entre dois ou mais conjuntos de dados que se queira comparar. O centro do boxplot traz a mediana, enquanto a caixa representa o intervalo interquartílico. O eixo vertical traz a variável analisada, e o horizontal traz um fator de interesse para esta variável.

3. Metodologia

Para cada objetivo do estudo, há um banco de dados diferente. Isto ocorre em função do *hagah* ser um *local search* que estabelece uma conexão entre os estabelecimentos e os usuários do site. Desta forma, há um banco somente para os estabelecimentos, outro somente para os visitantes, e outro para todos os estabelecimentos que um dia foram pagantes.

Os dados analisados são de fevereiro de 2011 até março de 2013 (início do estudo). Antes do período inicial, os dados não eram confiáveis, em função da extração utilizada na época. É interessante reforçar que, em agosto de 2012, o produto passou a ter 4 opções de planos pagos, e anterior à este mês, só existiam 2 planos. Esta mudança ocorreu em função de muitos clientes sentirem falta de uma opção mais barata que o plano inicial, e também por um plano intermediário entre eles, em função da grande distância financeira (R$ 99,90 para R$ 599,90). Os novos planos entraram na faixa inicial de R$ 49,90 e o intermediário em R$ 299,90.

Iremos identificar três perfis, e cada um terá um banco de dados diferente, como se segue:

- Para identificação do perfil dos anunciantes, temos os 1,1 milhão de estabelecimentos cadastrados no site. Neles, observam-se dados geográficos (estado e cidade), modelo de anúncio (gratuito ou pago – se pago, qual seu plano de destaque), tipo de negócio (se é uma loja, um professor de química, uma farmácia, etc.). Os dados foram retirados da base do site, que é controlada por uma empresa contratada, disponibilizada na internet pela ferramenta Effective Campaign.
- Para identificação do perfil de busca dos visitantes do site, temos todas as páginas visitadas no período, com informações geográficas (de onde o visitante faz sua procura e para onde ele procura – exemplo: morador de Porto Alegre procura restaurantes em Gramado), de modelo de anúncio, de tipo de negócio e de tempo (data). Esta última variável será muito útil para entendermos em que dia da semana o site é mais/menos acessado, e se ou não há uma variabilidade grande neste ponto. Os dados foram extraídos pelo Google Analytics.
- Para análise exploratória do comportamento financeiro do assinante, olhamos para variáveis do 1º banco de dados (perfil dos anunciantes), adicionando informações de forma de pagamento (boleto bancário, débito em conta ou cartão de crédito), desconto (se paga o preço normal do plano assinado ou não) e atrito com a base (se já passou por processo de intenção de cancelamento e foi ou não revertido). Os dados foram extraídos pelo CRM do *hagah.*

No entanto, para estes dados conversarem entre si, foi feito um trabalho de categorização dos anunciantes.

Hoje, ao cadastrar seu estabelecimento no site, o proprietário pode digitar qual a sua categoria para atendimento ao público (restaurante, pizzaria, fast-food, etc.). Esta liberdade de escolha termina poluindo o site com mais de 2 mil categorias diferentes. Para melhor interpretação dos dados, agrupamos todas em 20 categorias:

- **Comércio** (atacados, varejo, equipamentos, representantes comerciais, lojas e produtos diversos);
- **Serviços** (agências, administração, organizações, construções, manutenção);
- **Profissionais** (advogados, arquitetos, consultores, estatísticos, etc.);
- **Indústria** (extração, fabricação, produção, tecelagem, etc.);
- **Transportes e Veículos** (transportes aéreos, hidroviários, oficinas mecânicas, etc.);
- **Serviços Públicos** (instituições e fundações, ONGs, prefeituras, sindicatos, etc.);
- **Alimentação** (restaurantes, pizzarias, fast-food, etc.);
- **Saúde** (médicos, enfermeiros, laboratórios, etc.);
- **Casa e Jardim** (encanadores, jardineiros, floriculturas, etc.);
- **Comunicação** (edição de jornais, revistas, produção teatral/musical, etc.);
- **Educação** (educação infantil à educação superior, ensino de idiomas, de música, escola de DJ, etc.);
- **Autônomos** (acupuntura, esteticista, podólogo, etc.);
- **Condomínios** (Condomínios prediais e empresariais, lares para idosos, etc.);
- **Religião;**
- **Telefonia** (Oi, Terra, UOL, etc.);
- **Animais** (Criação de animais, zoológicos, pet shops, veterinários, etc.);
- **Hotéis** (Hospedagem, Hotéis, Pousadas, etc.);
- **Eventos** (Casas de eventos, filmagens, etc.);
- **Esporte** (Academias, Quadras de esportes, etc.);
- **Agricultura** (Cultivo e atividades pós-colheita).

Ou seja, se uma pessoa procura por nutricionistas em Porto Alegre, e a Maria da Rosa se cadastra no site na categoria "Nutricionistas", contaremos como uma busca por "Saúde" na capital gaúcha. Depois de feita esta categorização, algumas pequenas categorias, que representam menos de 0,1% dos estabelecimentos foram desconsideradas. Esse pequeno grupo era composto por cemitérios, parques, praças, etc.

Em função do volume de dados de visitas ao site ter mais de 20 milhões de registros, foi utilizado o software Qlikview 11 e SPSS para análise dos dados. Os testes de significância irão ser avaliados utilizando nível de significância $\alpha = 0,05$.

4. Estudo de Caso - *hagah*

Nesta seção, iremos apresentar o estudo de caso no *hagah*. Conforme a divisão citada na metodologia, teremos três partes: seção de estabelecimentos do site, falando sobre as características das empresas que anunciam no site; seção de páginas visitadas, mostrando o perfil dos usuários do site e o que eles procuram; seção de análise exploratória buscando identificar um perfil de comportamento como pagante do anunciante.

4.1 Estabelecimentos do site

Aqui, iremos descrever o perfil dos estabelecimentos que anunciam no site, olhando para o tipo de plano, quais as categorias que mais anunciam, qual a distribuição geográfica, quais as formas de pagamento que os anunciantes pagos utilizam e, dentre os estabelecimentos que hoje tem planos gratuitos, quantos já cancelaram anteriormente e voltaram por processo de reconquista.

O *hagah* diferencia os cerca de 1,1 milhão de estabelecimentos cadastrados no site de algumas formas, e a principal delas, para a visualização do usuário do site, é o plano. A predominância é do plano gratuito, com mais de 99%. Quando olhamos somente para a fatia de planos pagos – ao redor de ***7 mil*** - temos 4 planos, listados a seguir por ordem crescente de preço: Básico, Pleno, Avançado e Top. Os planos Pleno (80%) e Top (8%) existem desde o início do site (2006). Em agosto/2012, foram criados os planos Básico, que hoje já abrange 10%, e Avançado, que fica com a menor fatia (2%).

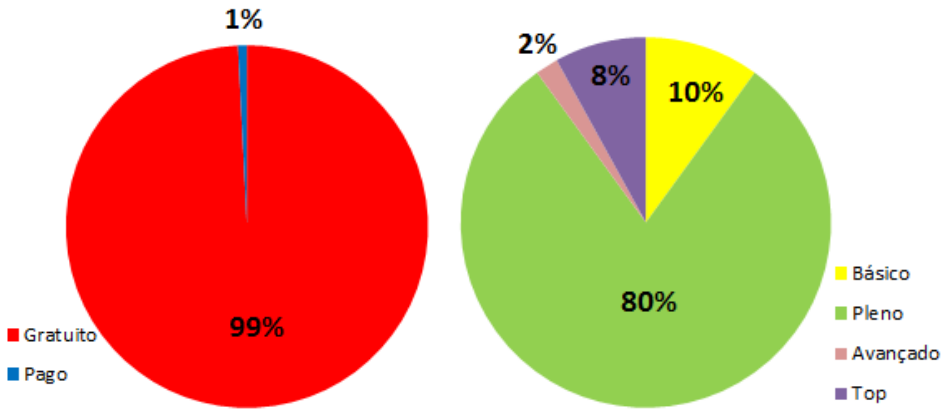

Fig. 1: percentual entre gratuitos e pagos, e dentre os pagos, seus planos de anúncio.

Cada plano tem um valor diferente. Os planos acima descritos serão chamados de 1 a 4, e seus valores, sem descontos ou promoções, seguem abaixo:

#	Plano	Valor mensal em R$
1	Básico	49,90
2	Pleno	99,90
3	Avançado	299,90
4	Top	599,90

Quadro 1: preços de cada **plano** pago no hagah.

Abaixo, na Fig. 2, observamos a distribuição do valor médio pago dos estabelecimentos em cada plano – todos variando abaixo do valor padrão (Quadro 1), em função de ações da área comercial, que usam descontos de até 50% em certos meses:

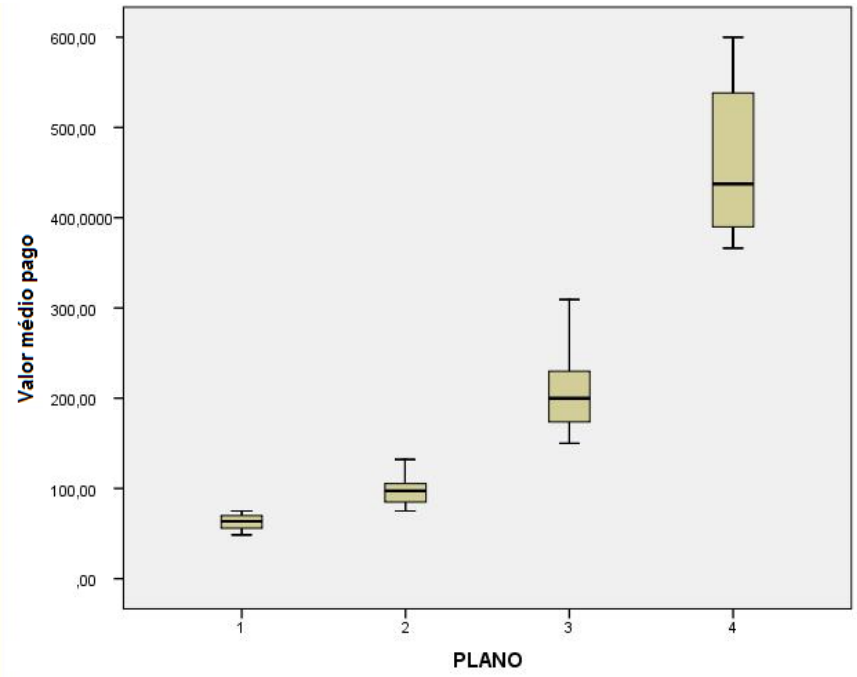

Fig. 2: variação do **valor médio pago** conforme o plano contratado pelo estabelecimento.

Todo estabelecimento escolhe uma categoria. Através dela, seus possíveis clientes vão encontrar seus serviços. Conforme explicado no capítulo de metodologia, agrupamos as 2 mil categorias existentes no site em 20, para melhor visualização dos dados.

A lógica da busca dos usuários do site funciona da seguinte forma: digita-se "Fisioterapia", por exemplo, e marca-se uma cidade na lista. Quem será encontrado nesta busca? Todos os profissionais do ramo que atuam naquela cidade e cadastraram-se no site. Abaixo (Fig. 3), o histograma das categorias, cujas 5 principais correspondem a 66% do total de estabelecimentos cadastrados:

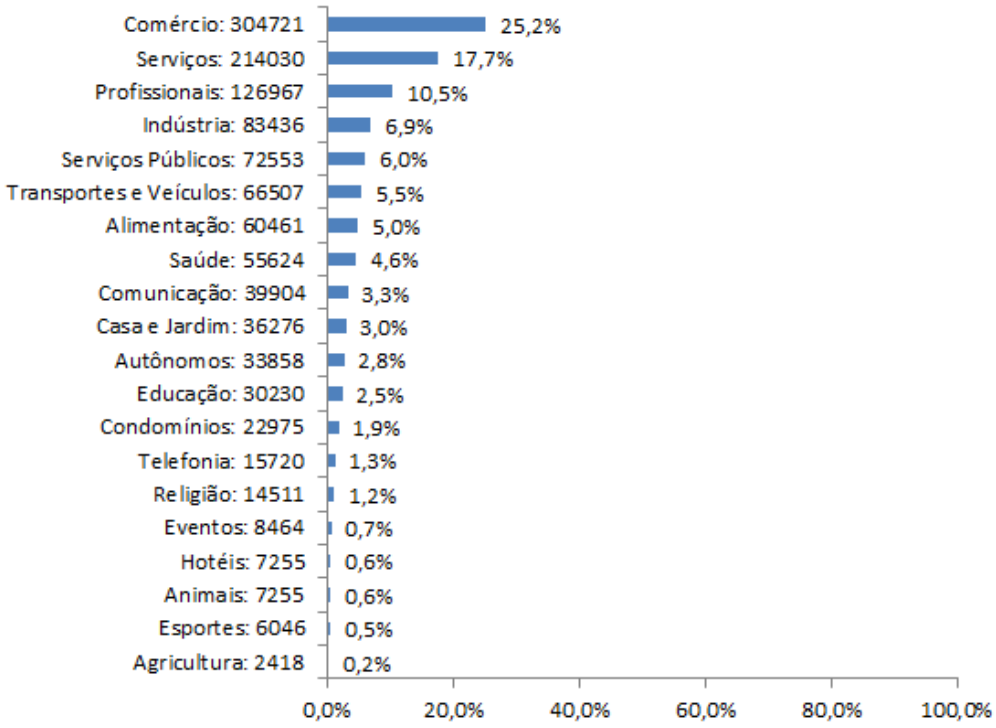

Fig. 3: distribuição percentual de todos os estabelecimentos cadastrados no site entre as 20 **categorias**, com seus respectivos volumes.

No entanto, quando olhamos para os estabelecimentos pagantes (Fig. 4), temos uma visão bem diferente. Casa e Jardim, por exemplo, é o 10º em volume no site, mas o 3º em volume de pagantes. Isto ocorre em função de uma percepção da operação comercial, que sentiu a necessidade desta categoria em ser encontrada na internet e focou as vendas neste tipo de cliente.

Na contramão, vemos a categoria Indústria, que é a 4ª em volume no site, mas apenas a 14ª em volume de pagantes. Este ponto é facilmente entendível porque o ramo não é o foco de atuação do *hagah*, que trabalha quase exclusivamente com micro e pequenas empresas.

Se abrirmos as categorias por planos (Fig. 5), vemos que a menor faixa azul (plano básico = 1, conforme descrito no Quadro 1) está na categoria alimentação. Por ser uma categoria com grande volume de estabelecimentos pagos, quem quer se diferenciar entre os concorrentes precisa de um plano mais completo, ou seja, o básico não é o suficiente.

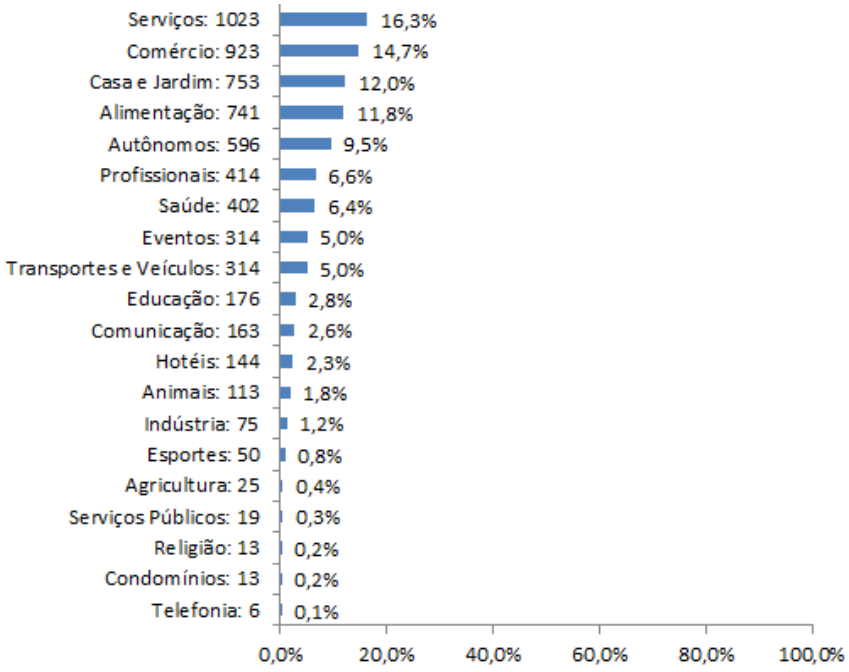

Fig. 4: distribuição dos estabelecimentos pagantes entre as **categorias**, com seus respectivos volumes.

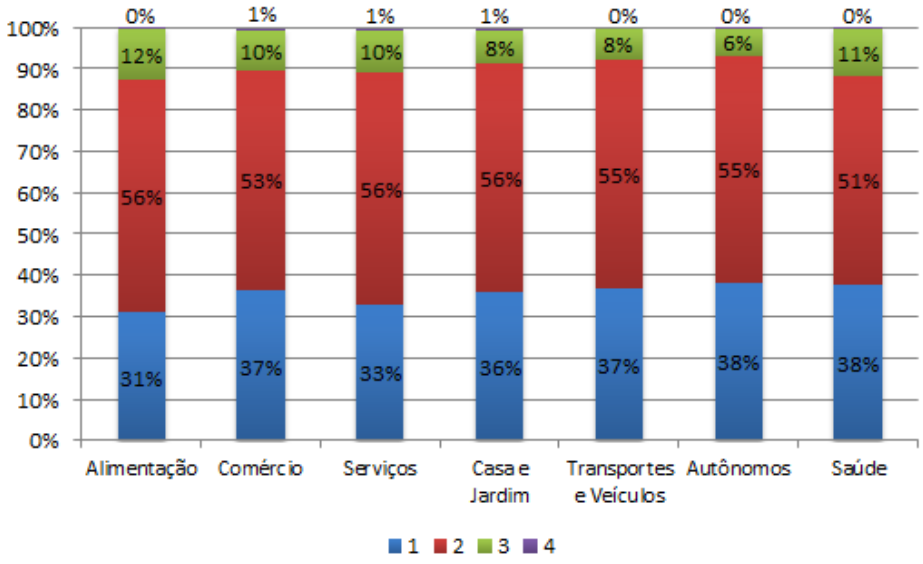

Fig. 5: distribuição percentual da participação dos **planos** entre as principais **categorias**.

Hoje, o site atende os 3 estados da Região Sul (RS, SC e PR) e São Paulo. Olhando para a base total, com gratuitos e pagos, temos predomínio de SP, como vemos na Fig. 6(a). No entanto, quando excluímos os gratuitos, vemos que a carteira de anunciantes do *hagah* é gaúcha (75%) – Fig. 6(b). Restam 16% para SC, 8% para o PR e 1% para SP. Isso se deve principalmente pelo fato de o *hagah* ser uma empresa do Grupo RBS, que tem nome forte no RS e em SC. No Paraná, poucos associam a empresa à marca, e em São Paulo, quase não há divulgação.

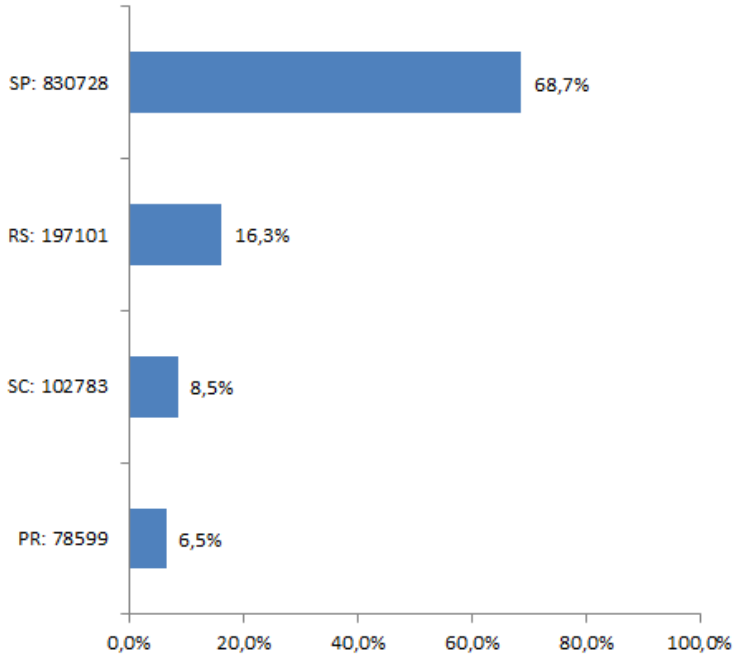

Fig. 6 (a): distribuição por **UF** do total de estabelecimentos, com seus respectivos volumes.

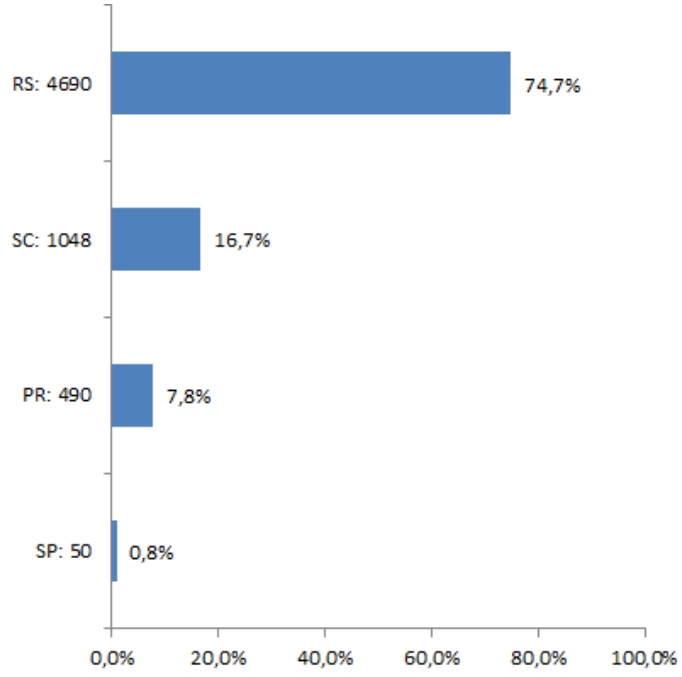

Fig. 6 (b): distribuição por **UF** dos estabelecimentos pagantes, com seus respectivos volumes.

Olhando por cidade, na Fig. 7(a), vemos que as 10 principais abrangem pouco mais de 50% do volume no site. Quando restringimos a visão de destacados, na Fig. 7(b), vemos maior concentração, com as 10 principais cidades representando 73% do total.

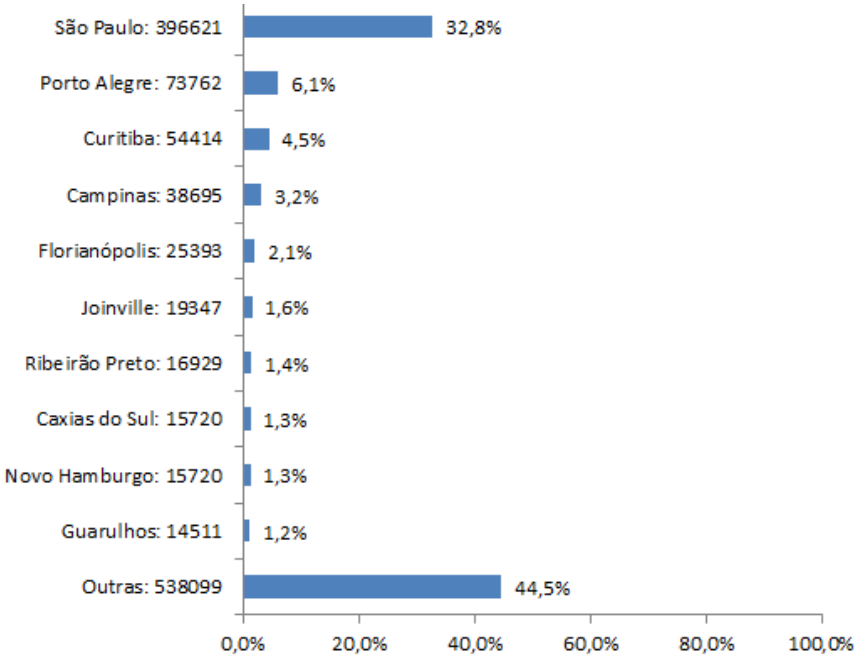

Fig. 7 (a): distribuição por **cidade** do total de estabelecimentos, com seus respectivos volumes.

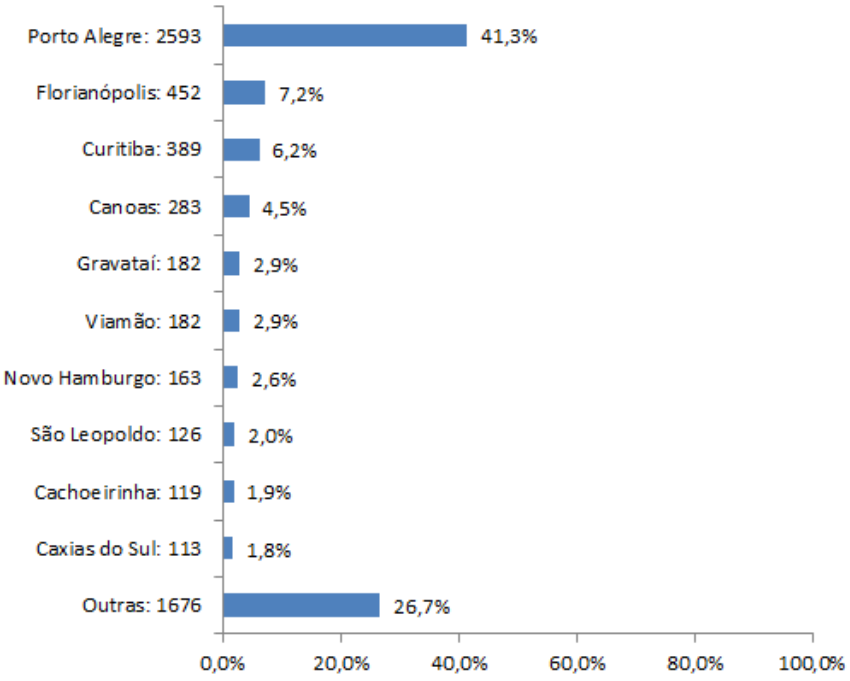

Fig. 7 (b): distribuição por **cidade** dos estabelecimentos pagantes, com seus respectivos volumes.

Também foram observadas duas variáveis que serão importantes para análises subsequêntes. Primeiro, o tipo de faturamento dos clientes (Fig. 8), aonde temos a grande maioria (77%) utilizando o boleto bancário (faturado). Isto ocorre em função das restrições no momento da venda que o *hagah* aplica. Se o CPF/CNPJ do cliente constar com débito no mercado, rejeitamos seu cartão, sobrando assim, a forma de faturado.

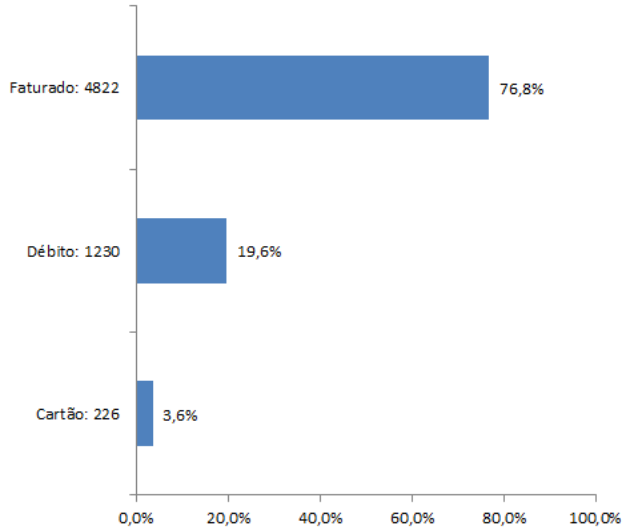

Fig. 8: distribuição dos estabelecimentos pagantes por **forma de pagamento**, com seus respectivos volumes.

Também analisamos a quantidade de vezes que cada um dos atuais clientes pagos passou por processo de recuperação (Fig. 9). O volume de clientes atuais que já cancelaram em algum momento é de 25%, sendo que 7% foram recuperados mais de uma vez. Sabendo disto, já temos uma ideia de que ¼ da carteira de anunciantes pagos tem uma propensão a voltarem a cancelar o produto, uma vez que já fizeram isto no passado.

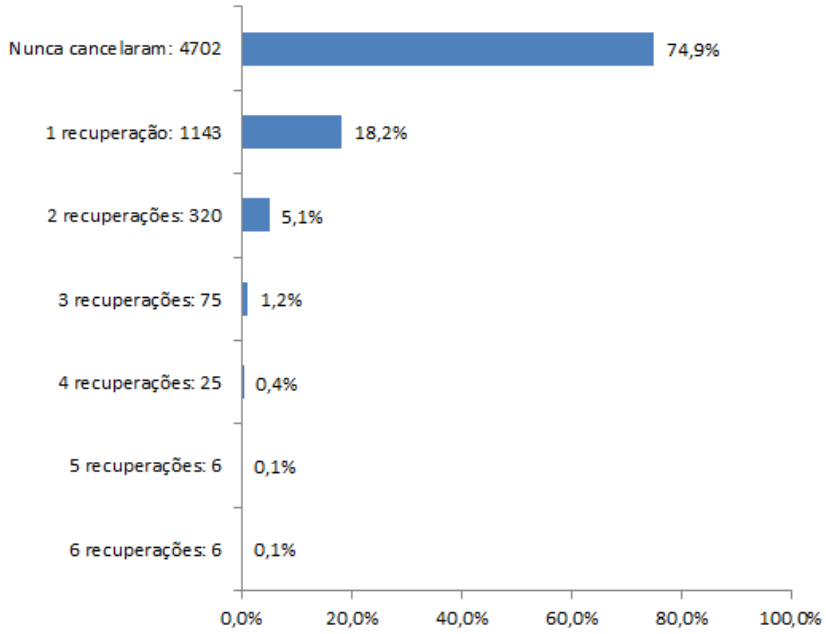

Fig. 9: percentual de anunciantes passaram por processo de **recuperação**, e seus respectivos volumes.

4.2 Páginas visitadas

Nesta seção, iremos descrever o perfil de busca dos visitantes do site. Serão analisados dados temporais, como o volume de acessos ao site, e também dados dos volumes totais por categorias de anúncio, com a distribuição geográfica, tipo de plano e suas performances, e também a relação entre a cidade do estabelecimento e a cidade do visitante do site. Este último ponto esclarecerá, por exemplo, se quem procura por um restaurante na Serra Gaúcha são moradores da região ou pessoas que residem em outros locais e estão à procura de estabelecimentos para visitar em uma viagem.

A Fig. 10 apresenta o volume diário de visualizações de página separados em anos. Os dados mostram que há um crescimento médio, pois a linha de 2012 está acima de 2011, assim como 2013 se posiciona acima de 2012.

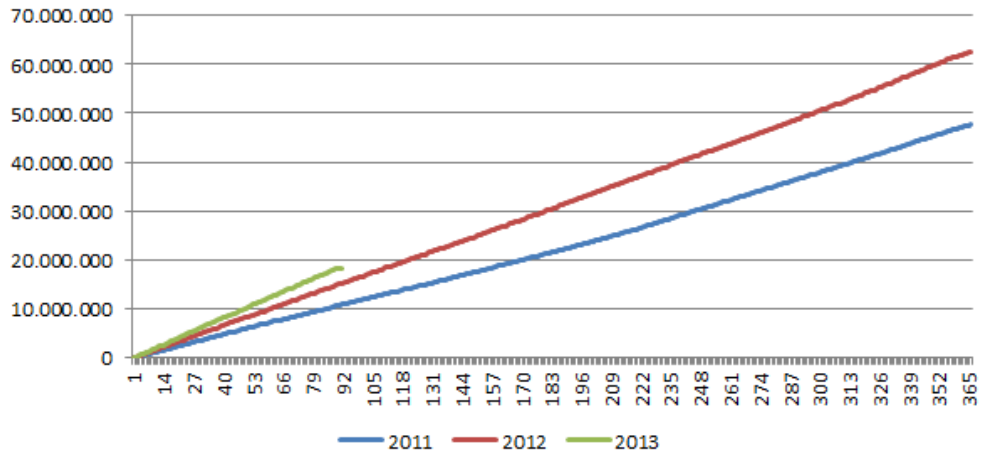

Fig. 10: volume de **visualizações de página** ao site hagah acumulados, dia a dia, ao longo de um ano (365 dias).

Na Fig. 11, analisamos a variação entre os dias da semana. Há uma diferença grande entre os dias úteis (2ª a 6ª feira) e os fins de semana. Temos o auge na 2ª feira e queda gradativa até 6ª feira. Este é um comportamento padrão do uso da internet no Brasil, em função do horário comercial, e se reflete também no uso do *hagah*.

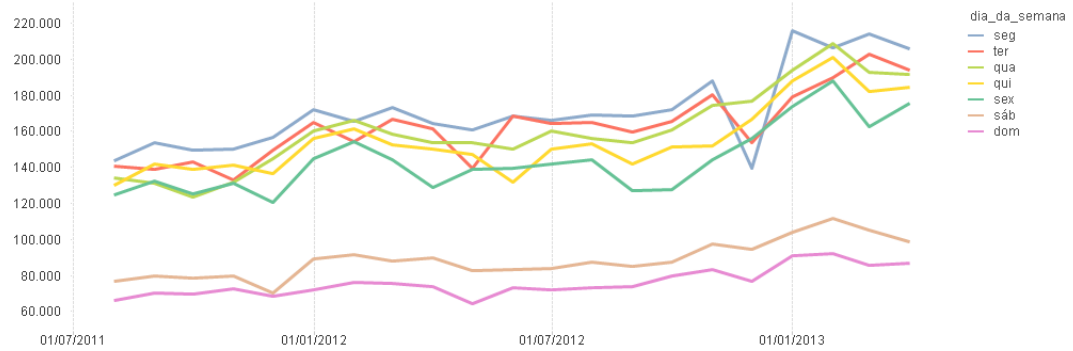

Fig. 11: evolução mensal do volume de **visualizações de página (média por dia)** no site por dia da semana.

Em relação às categorias mais acessadas, a Fig. 12 mostra que os estabelecimentos procurados pelos visitantes do *hagah* se mostram bem concentrados, com as 5 principais representando mais de 60% do total.

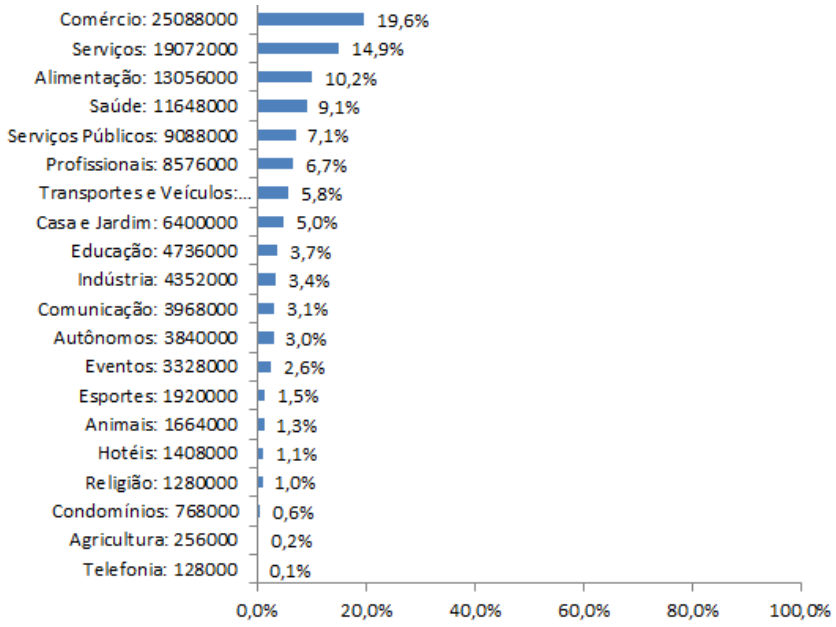

Fig. 12: **categorias** mais acessadas no site (volumes de **visualização** e percentuais).

Na Fig. 13, temos uma visão por cidade que mostra desencontros entre a procura dos visitantes e os estabelecimentos cadastrados no site. Porto Alegre e Curitiba tem uma base equilibrada, pois são procuradas na mesma proporção que oferecem serviços. No entanto, São Paulo tem um volume expressivo de visitantes que olham para estabelecimentos fora de sua cidade.

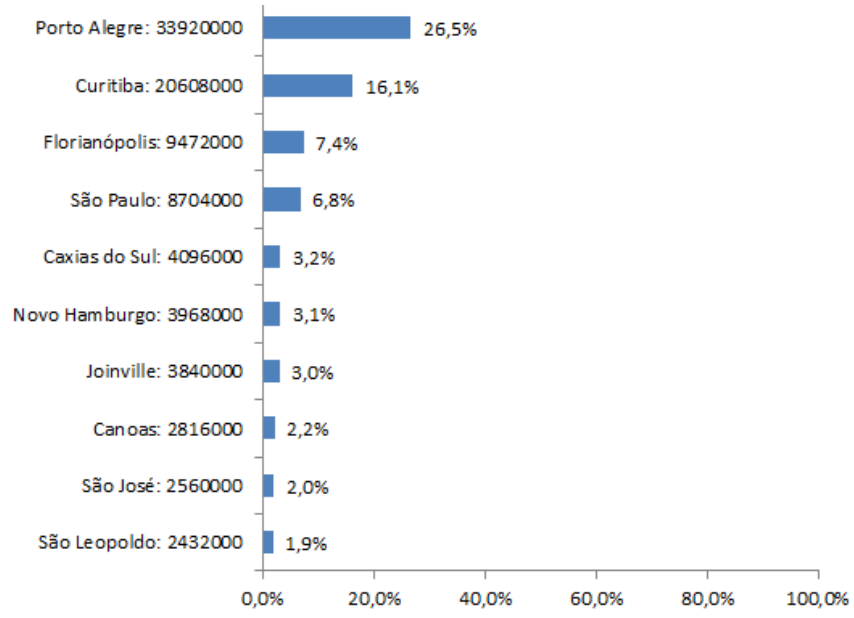

Fig. 13(a): **cidades** mais acessadas (volumes de **visualizações** e percentuais).

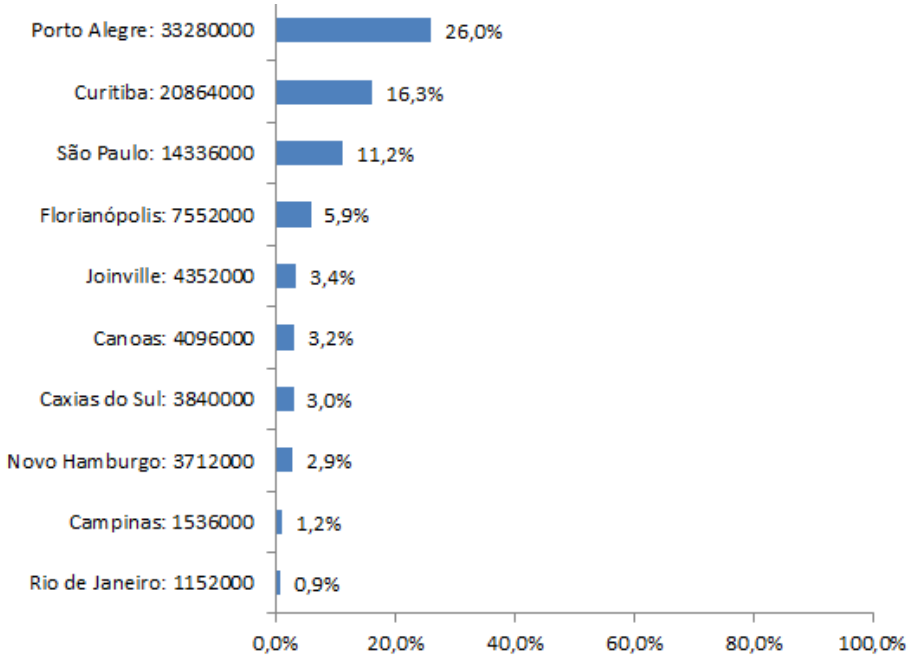

Fig. 13(b): **cidades** dos visitantes do site que mais acessam o portal (volume de **visualizações** e percentuais).

Como vimos nas Fig. 6 e 7 da seção de estabelecimentos, São Paulo quase não aparece em planos pagos, mesmo tendo alto volume de visualizações de página no site. A questão principal é o desconhecimento da marca do Grupo RBS neste estado, além da preferência em anunciar direto no Google (Maps, Trends, etc). São empresários que preferem a métrica de pagar por clique (método do Google) do que fazer um pagamento mensal, como o *hagah* trabalha.

Na Fig. 14(a), vamos olhar para uma métrica de performance: média de visualizações/mês. Assim como a escala de preços, o Top é o melhor plano neste quesito. A ordem segue com o Avançado, Pleno, Básico e Gratuito. No entanto, a distância entre os preços é muito maior do que a distância entre estas médias. Isto é explicado pelo fato de que o hagah não entrega somente pageviews ou cliques, e também agrega impressões de página aos planos mais caros. Como vemos na Fig. 14(b), os estabelecimentos pagos avançados ou tops aparecem na página dos concorrentes, o que aumenta o número de vezes em que ele aparece no site. Além disto, em suas páginas, assinantes que pagam por planos maiores têm direitos a usarem mais ferramentas do site, como o "Ligue Grátis", na Fig. 14(c), que conecta o cliente e o estabelecimento em ligação gratuita.

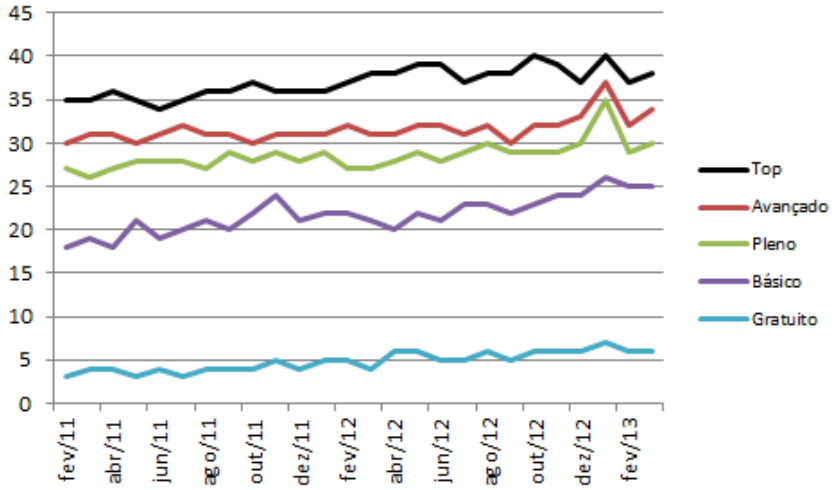

Fig. 14 (a): média de **visualizações de página** no portal *hagah* por **tipo de plano** de anúncio.

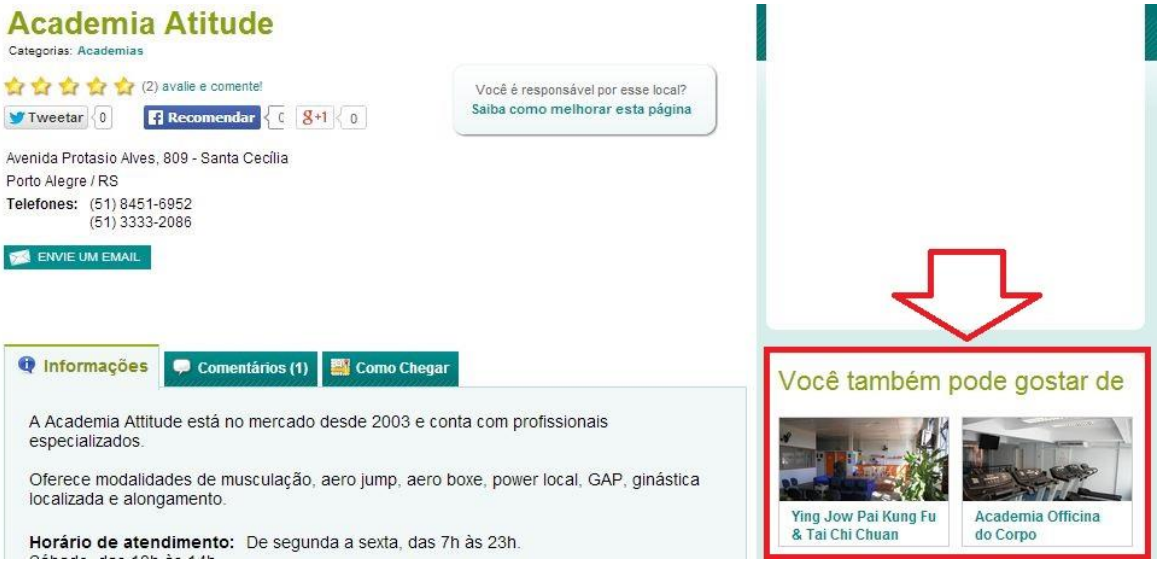

Fig. 14(b): como estabelecimentos pagos são vistos nas páginas dos concorrentes gratuitos.

Fig. 14(c): estabelecimento que faz uso do **Ligue Grátis** no site hagah.

Por último, vemos as cidades que mais tem estabelecimentos e visitantes em "sintonia", ou seja, se as buscas em uma cidade são feitas por seus próprios moradores ou por visitantes de fora da cidade. Na Fig. 15 (a), vemos que Curitiba e Porto Alegre, cidades que dominam o site em volume de acessos, são as cidades que mais apresentam % de busca de seus próprios moradores. No lado oposto, na Fig. 15 (b), vemos cidades onde quase nenhuma busca feita por estabelecimentos da cidade são feitas por seus moradores. Destacam-se a Região Metropolitana de Porto Alegre (Alvorada, Cachoeirinha, Ivoti, Esteio e Sapiranga) e a Serra Gaúcha (Gramado e Canela).

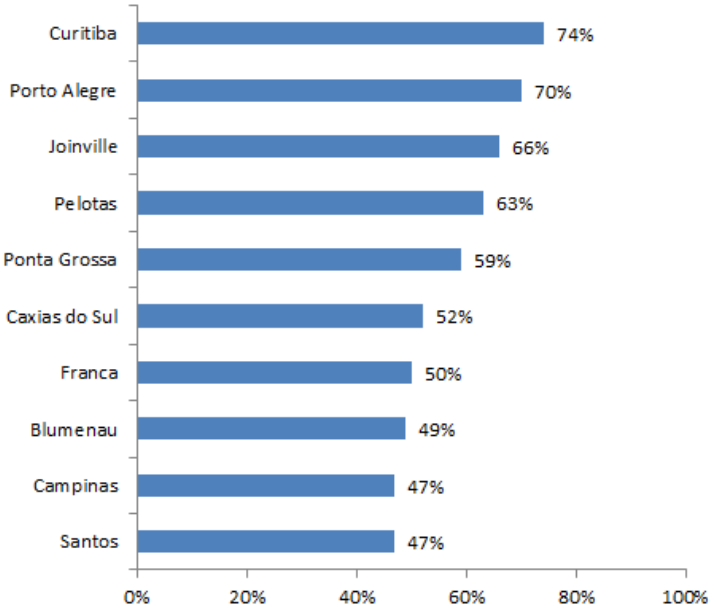

Fig. 15(a): as 10 **cidades** mais visualizadas por seus próprios moradores.

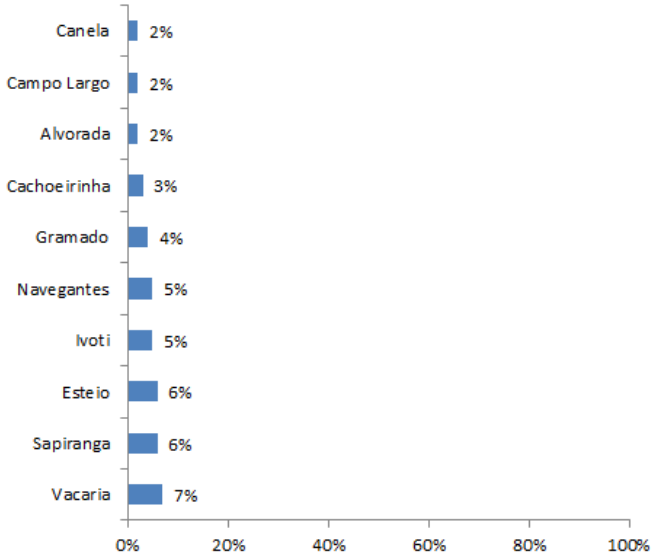

Fig. 15(b): as 10 **cidades** menos visualizadas por seus próprios moradores.

4.3 Análise exploratória do comportamento financeiro do assinante

Após entendermos como os estabelecimentos se distribuem pelo *local search* e analisarmos as origens de tráfego no site, iremos explorar variáveis de comportamento financeiro do assinante, verificando possíveis associações entre elas. Uma importante variável que analisaremos será sobre a adimplência do cliente. A partir desta variável, poderemos, por exemplo, diferenciar as categorias pelo percentual médio de adimplência.

Nos testes utilizados para verificar associação entre as variáveis, foi utilizado o teste de Kruskal Wallis, em função de não termos homogeneidade de variâncias nos dados. Se este pressuposto fosse atendido, poderíamos utilizar o teste ANOVA.

Para melhor visualização, aqui trataremos somente das sete maiores categorias, que juntas correspondem a 78% do total dos estabelecimentos pagos (Quadro 2).

Categorias	**Volume**	**%**
Alimentação	1.196	19%
Comércio	1.038	17%
Serviços	923	15%
Casa e Jardim	600	10%
Transportes e Veículos	399	6%
Autônomos	364	6%
Saúde	308	5%
Subtotal (7 maiores)	4.828	78%
Outras categorias	1.450	22%
Total	6.278	100%

Quadro 2: volumes de estabelecimentos de cada uma das sete maiores **categorias**.

Uma das variáveis mais analisadas no negócio é o tempo de permanência na carteira (TPC), medido em meses. Portanto, vamos compará-lo em algumas dimensões:

- Categorias (Fig. 16): a categoria Saúde apresenta TPC médio maior que as demais, enquanto Autônomos tem o menor resultado. O teste Kruskal-Wallis ($p < 0,05$) aponta diferenças no perfil de TPC nas categorias.

- Cidades (Fig. 17): olhando para as principais cidades entre os pagantes, vemos Curitiba e Porto Alegre com TPC acima das demais. Isto pode ocorrer em função das campanhas publicitárias serem muito voltadas a estas duas capitais. Negativamente, vemos que Gravataí está consideravelmente abaixo das demais cidades, com boa parte dos assinantes ficando abaixo dos 6 meses (tempo de fidelidade do produto). O teste Kruskal-Wallis ($p < 0,05$) apontou diferenças no perfil de TPC entre as cidades.

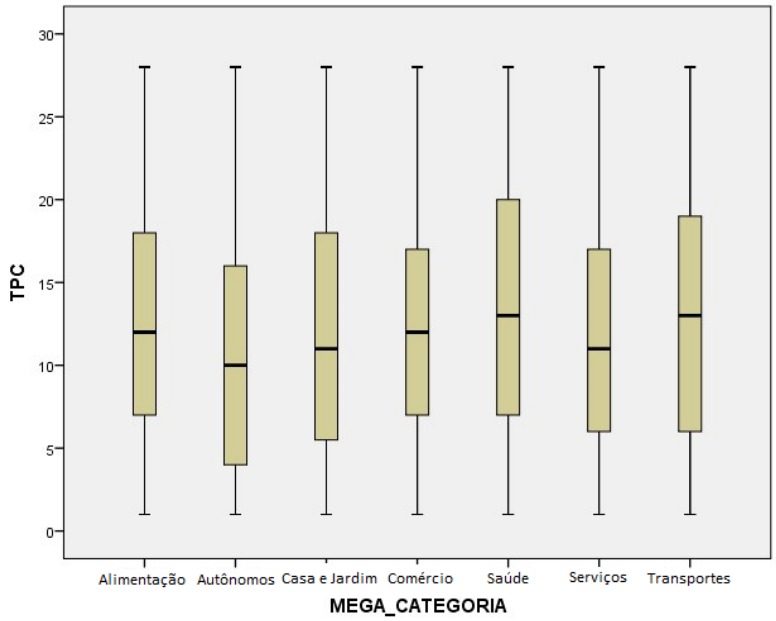

Fig. 16: **tempo de permanência na carteira** (em meses) das principais **categorias** do site.

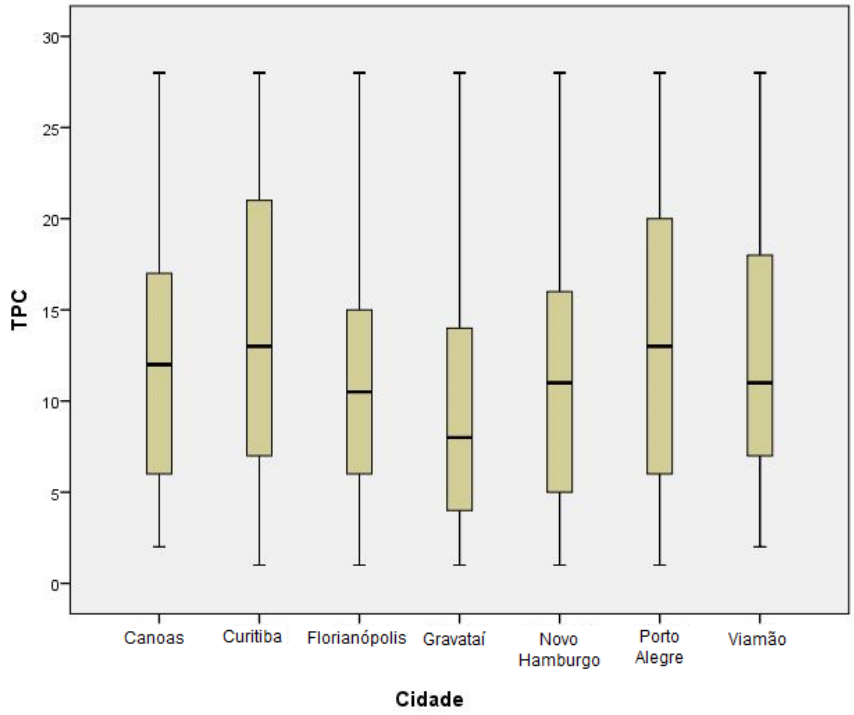

Fig. 17: **tempo de permanência em meses** na carteira das principais **cidades** do site.

- Plano (Fig. 18): temos leve diferença nas medianas de TPC entre os planos – quanto mais alto o plano, maior sua permanência como estabelecimento pagante. Aqui, excluímos da análise o plano 4 (top), por ser um plano com poucos casos.

Após realizar o teste de Kruskal Wallis, concluímos ($p < 0,05$) que temos diferença significativa entre os planos para o TPC.

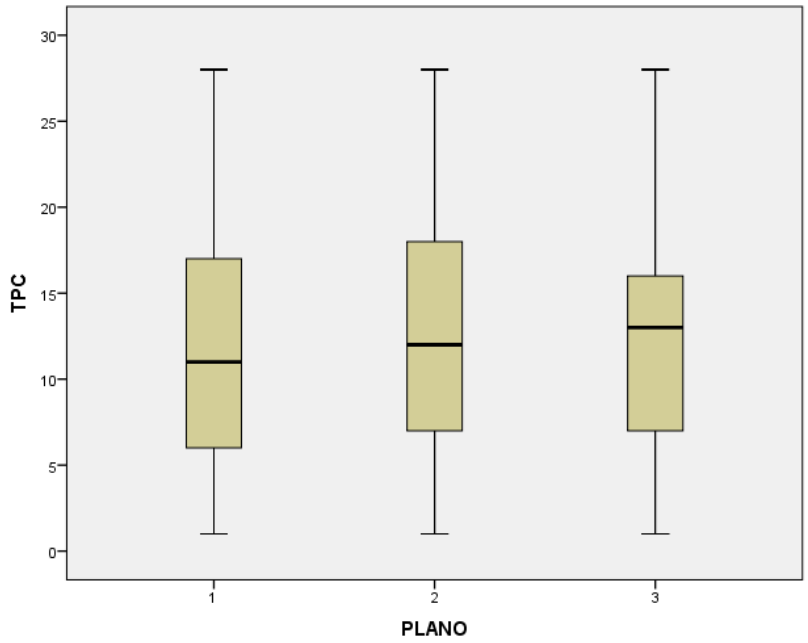

Fig. 18: **tempo de permanência na carteira**, em meses, de cada **plano** pago do hagah.

Abaixo, vamos trazer o foco para a variável categoria, e analisar as variações quando comparadas a outras dimensões:

- Tipo de faturamento (Fig. 19): chamamos de "Cartão" a soma de débito e crédito. Vemos grandes diferenças entre as Categorias: Comércio, Transportes e Saúde estão acima de 30%, enquanto Autônomos fica bem abaixo deste percentual, em 24%. O teste qui-quadrado apontou diferença significativa nas proporções entre as categorias ($p<0,05$). Analisando os resíduos do teste, vemos que as categorias Alimentação e Autônomos contribuem significativamente para o Boleto, enquanto a categoria Comércio contribui significativamente para a forma de pagamento com Cartão.

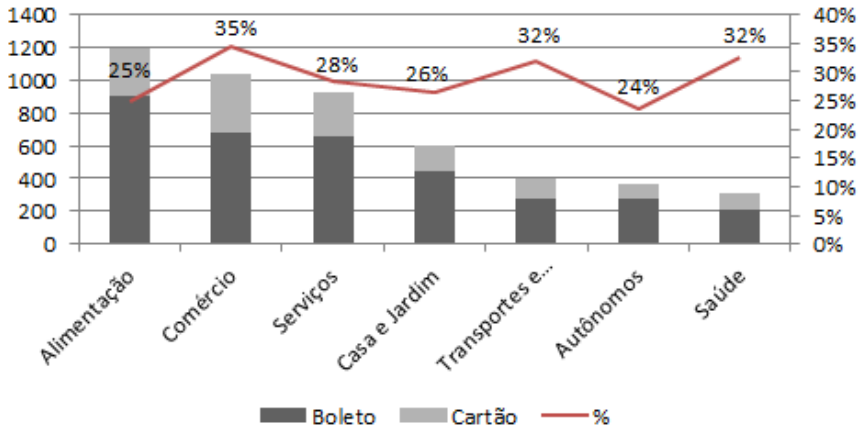

Fig. 19: comparação entre as principais **categorias** com a **forma de pagamento**.

Através de uma parceria com um fornecedor, o hagah abasteceu um vigésimo de sua base com um atributo de escore de crédito para adimplentes. Ele varia de 0 a 1000, e quanto mais alto, maior a chance do estabelecimento ser adimplente. De acordo com a estratégia de captação de anunciantes com bom comportamento como pagador, o hagah obteve uma base com empresas com escore superior a 600.

Na Fig. 20, temos uma visão parecida em relação à comparação das categorias de anunciantes e o tempo de permanência (Fig. 16): para este cruzamento, foi feito um teste de Kruskal-Wallis ($p < 0,05$), pelo qual confirmamos que há diferenças significativas entre as categorias em relação ao score de adimplência.

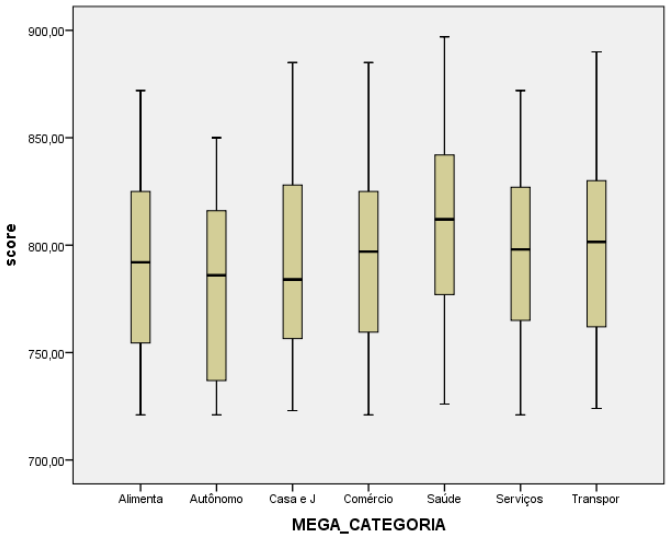

Fig. 20: comparação entre o **score de adimplência** nas **categorias** dos estabelecimentos pagos.

As próximas análises tratam da *inadimplência*, variável categórica binária de análise que terá resposta 'ad' (adimplente) se o estabelecimento nunca foi recuperado e 'inad' (inadimplente) se ele já cancelou e passou por pelo menos um processo de recuperação. Pensando no comportamento financeiro do estabelecimento, vemos que esta é uma variável importante.

Comparando a *inadimplência* com as categorias, vemos que há pouca variação (Fig. 21). O teste qui-quadrado ($p > 0,05$) confirma que não há diferença significativa entre as categorias.

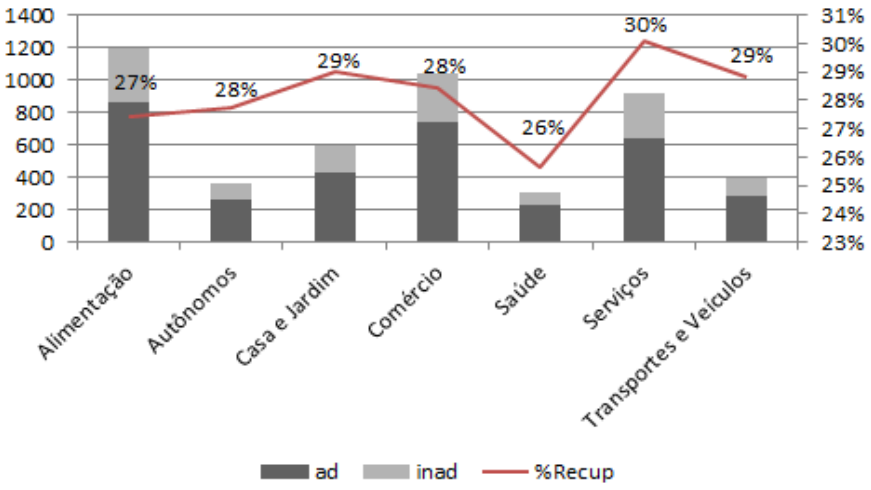

Fig. 21: comparação entre as **categorias** principais com a **inadimplência** (volumes de estabelecimentos e percentual da relação entre adimplentes e inadimplentes).

Comparando a *inadimplência* com as principais cidades, vemos que temos algumas diferenças (Fig. 22). Gravataí aparece abaixo das demais, enquanto Viamão e Curitiba aparecem na liderança. No entanto, o teste qui-quadrado ($p > 0,05$) mostra que não há diferença significativa entre as categorias.

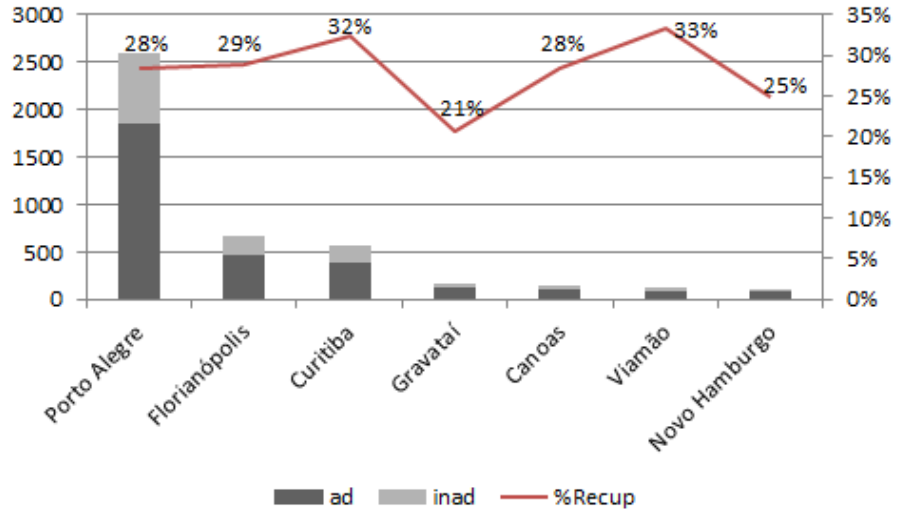

Fig. 22: comparação entre as principais **cidades** com a inadimplência (volumes de estabelecimentos e percentual da relação entre adimplentes e inadimplentes).

Outra análise feita é o cruzamento da *inadimplência* com uma variável de tempo. Separamos os clientes com entrada no hagah antes ou depois de 2012. A partir deste ano, o *hagah* passou por modificações no site, que trouxeram muitos benefícios para os assinantes. Desta forma, acredita-se que os clientes passaram a ver mais valor no produto. E o que o conceito e o gráfico (Fig. 23) nos sugerem é confirmado pelo teste qui-quadrado ($p < 0,05$), apontando que há diferença significativa entre os anos. Quem entrou após 2012 tem uma taxa de inadimplência consideravelmente mais baixa. Esta diferença era esperada, uma vez que a área comercial, desde 2012, passou a se preocupar mais com a qualidade da venda, e não só com a quantidade.

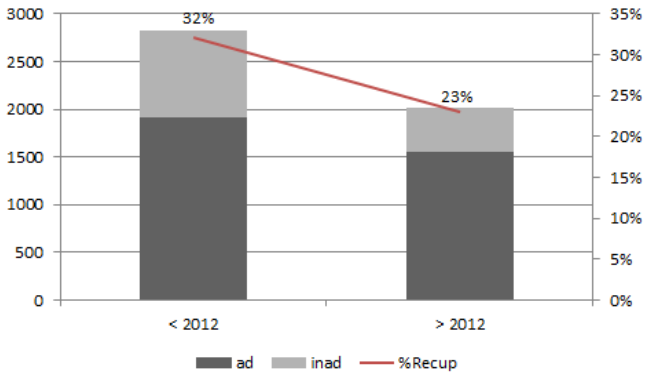

Fig. 23: comparação entre os clientes por ano de entrada no hagah com a inadimplência (volumes de estabelecimentos e percentual da relação entre adimplentes e inadimplentes).

Analisando a *inadimplência* com o plano do cliente (Fig. 24), vemos que o percentual varia pouco entre os planos 1, 2 e 3 (aqui, mais uma vez, excluímos o plano 4 por ter baixo número de casos). O teste qui-quadrado ($p > 0,05$) nos sugere que não há diferença significativa entre os planos e a inadimplência.

Junto ao plano do cliente, vamos analisar também sua forma de pagamento contra a variável *inadimplência* (Fig. 25). Mais uma vez, pelo teste qui-quadrado ($p > 0,05$), concluímos que não há diferença significativa em nenhum dos três planos entre suas formas de pagamento e a inadimplência.

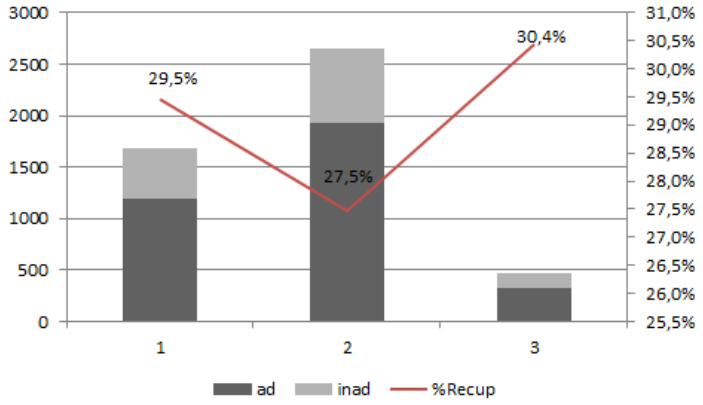

Fig. 24: comparação entre o plano dos clientes com a inadimplência (volumes de estabelecimentos e percentual da relação entre adimplentes e inadimplentes).

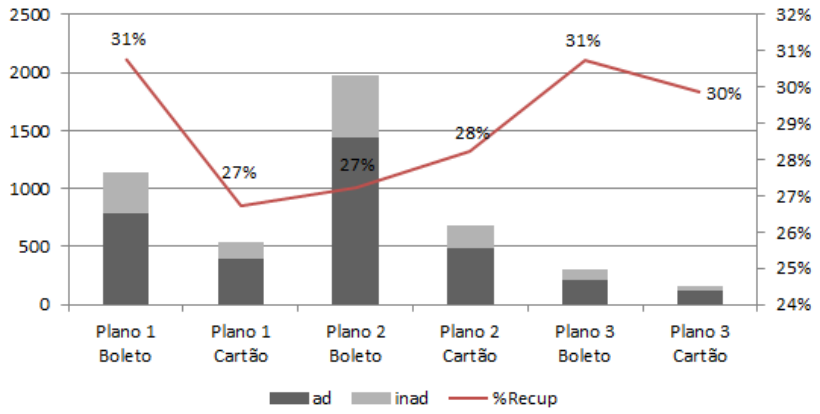

Fig. 25: comparação entre plano dos clientes e suas formas de pagamento com a inadimplência (volumes de estabelecimentos e percentual da relação entre adimplentes e inadimplentes).

Junto ao plano do cliente, também foi analisada a categoria do estabelecimento contra a variável *inadimplência* (Fig. 26). Pelo teste qui-quadrado ($p > 0,05$), só há diferença significativa entre os planos e a inadimplência para uma categoria específica: Saúde. Dentro desta categoria, há diferença entre os planos, o que é visível na Fig. 26, pois o plano 1 tem 35% de inadimplência, enquanto os planos superiores ficaram em valores bem mais baixos (19% e 24%). As demais categorias não se apresentaram distintas pelo mesmo teste.

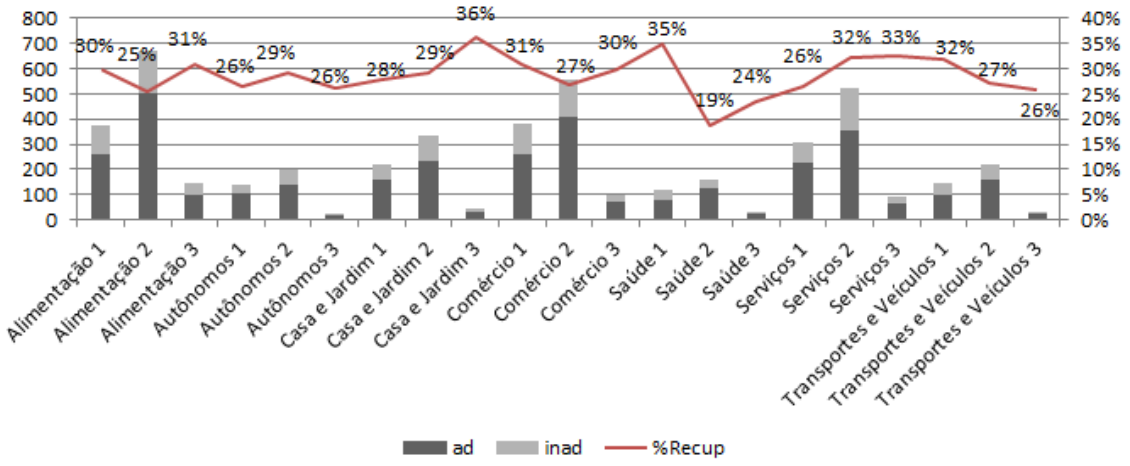

Fig. 26: comparação entre a categoria e o plano dos clientes com a inadimplência (volumes de estabelecimentos e percentual da relação entre adimplentes e inadimplentes).

Por último, analisamos o score do estabelecimento contra a variável *inadimplência* (Fig. 27). Separando os acima e abaixo de 600, vemos que há uma diferença significativa (teste qui-quadrado, $p < 0,05$), ou seja, podemos afirmar que quem tem score maior que 600 pontos têm menos risco de não pagar que o grupo abaixo de 600, o que reforça a consistência do escore e, assim, sua utilização para a prospecção de inadimplentes.

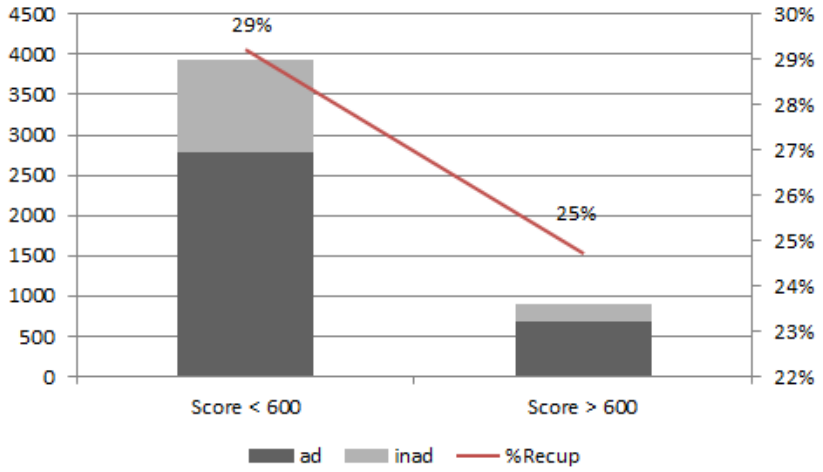

Fig. 27: comparação entre o score de adimplência com a inadimplência (volumes de estabelecimentos e percentual da relação entre adimplentes e inadimplentes).

Complementando as análises das variáveis vistas, vamos olhar para a curva de retenção dos planos dos assinantes (Fig. 28). Mês a mês, a curva aponta a saída dos estabelecimentos daquele plano, e assim pode-se comparar qual o plano com melhor liquidez para o negócio. No 1º mês, a curva de retenção mostra 100% dos clientes que compraram o produto, na safra x. Nos meses seguintes, a queda da curva em 5% indica que, naquele mês, 5% dos clientes deixaram de pagar.

Vemos que o plano 4 tem a melhor curva durante todo o período. No início da curva, o plano 2 chega a superar o 3 em dois meses, mas no geral, o plano 3 tem melhor performance. Em todo o período, o plano 1 tem a menor retenção.

É interessante notar também que as maiores quedas ocorrem entre o 6º e o 7º mês, justamente no momento em que acaba a fidelidade do produto. Portanto, é uma queda esperada.

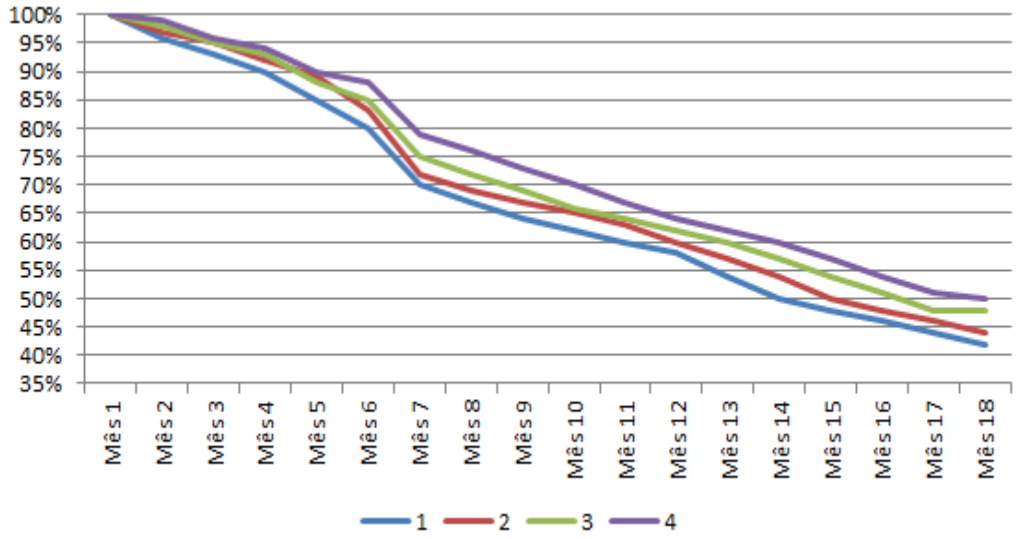

Fig. 28: curvas de retenção por plano de anúncio.

Para investigar a variável inadimplência em função de algumas das variáveis de comportamento econômico do anunciante apresentadas acima, foram testados modelos de regressão logística (HOSMER, D. W & LEMESHOW, 1989) com diferentes configurações de covariáveis preditoras. Essas variáveis envolvem: Plano (incluindo os quatro níveis de planos descritos), Categoria (utilizando as 7 maiores categorias de anunciantes, conforme Tab.2) que o cliente pertence, TPC (tempo de permanência em meses em que o cliente ficou ou está destacado), TPCi, para i=1,2,..,15 (variável binária, onde 0 é inadimplente e 1 é adimplente após o *i-ésimo* mês), e Faturamento (variável binária: boleto ou cartão). No entanto, nenhum modelo teve desempenho estatisticamente satisfatório, pois embora tenham alguma associação com a inadimplência, explicaram pouco dessa variável. A partir desta aparente falta de informação, estão sendo feitas reformulações no contrato do *hagah*, para que no momento do venda, sejam colhidas informações importantes sobre a empresa que contrata o anúncio, como por exemplo, tempo de mercado, porte (número de funcionários), faixa de faturamento anual, etc. Com variáveis desta natureza, conseguiremos avaliar melhor se uma empresa tende ou não a ser inadimplente com sua assinatura no *hagah.*

5. Conclusão

Este artigo desenvolveu extração, tratamento, e analisou um grande volume de dados. Foi possível alcançar os 3 objetivos principais: identificamos o perfil dos estabelecimentos que divulgam seu negócio no site, identificamos o perfil das buscas que os visitantes do site realizam, e por último foram analisadas os possíveis motivos da inadimplência dos assinantes do site.

Inicialmente, vimos que o ramo do *Comércio*, prestadores de *Serviços* e *Profissionais* em geral são aqueles que mais se preocupam em terem suas empresas divulgadas na internet, ao menos de forma gratuita. No entanto, vimos que estes *Profissionais* tem mais resistência em transformar seu anúncio gratuito em pago. Os ramos da *Alimentação* e *Casa e Jardim* apresentam maior preocupação em não só divulgar como ter um destaque maior no site, pagando por melhores indexações. Também foi entendido que, de todos os atuais assinantes, ¼ já passou por processo de intenção de cancelamento e foi revertido, voltando a pagar com algum nível de desconto. Isto preocupou a diretoria do negócio, que já redesenha a régua de relacionamento com os clientes, pois este grande volume gera impacto no valor médio pago da carteira de assinantes.

Na sequência, analisando as páginas do site, percebemos que o número de acessos vem crescendo ano após ano, e que estes acessos se concentram em horário comercial. Fica claro também que os visitantes que mais procuram são das mesmas cidades que os estabelecimentos que mais anunciam, o que aponta a importância da divulgação na internet. Ainda, detalhamos esta relevância para que esta divulgação seja paga, e o quanto fará diferença dependendo do plano escolhido para assinatura.

Por fim, a análise do comportamento do anunciante nos aponta alguns resultados esclarecedores. Analisando o TPC (tempo médio de permanência na carteira), encontramos as melhores cidades (Porto Alegre e Curitiba) e categorias (Saúde), assim como as piores (Gravataí e Autônomos). Então, quando a área comercial planejar suas ações, sabe em quem e aonde deve focar ou evitar suas vendas.

Por outro lado, a variável *adimplência* foi igualmente detalhada e não houve diferença significativa entre categorias, cidades, ou tipo de plano assinado. Isto levou a diretoria do negócio a pensar que o problema da inadimplência é gerado por uma venda sem qualidade, o que irá gerar uma reestruturação na área de treinamento dos vendedores.

6. *Referências*

- AGRESTI, A. *An Introduction to Categorical data Analysis.* New York: John Wiley & Sons, 2000.
- HOSMER, D. W.; LEMESHOW, S. *Applied logistic regression. New York:* John Wiley & Sons, 1989.
- IBM Corp. Released 2011. *IBM SPSS Statistics for Windows, Version 20.0.* Armonk, NY: IBM Corp.
- KIRK, R. E. *Experimental Design: procedures for the behavioral sciences.* 3.ed. Brooks/Cole, 1995.
- MALHOTRA, N. *Pesquisa de Marketing – uma orientação aplicada.* 6.ed. Bookman, 2012.
- QlikTech International AB. Released 2012. *Qlikview for Windows, Version 11.0.* Radnor, PA: Qlik Technologies Inc.
- TUKEY, J. W. *Exploratory Data Analysis.* Addison-Wesley, 1977.
- UPTON, G.; COOK, I. *A Dictionary of Statistics (2 rev).* Oxford University Press, 2008.