Nome Completo dos Integrantes do Grupo:
Daniele Lyra, Eduarda Rousseau e Gabriela Lea Alves.

Clusterização de compras de e-commerce: a relação dos dados geoestatísticos com a formação de clusters.

Contexto

Segundo o estudo Webshoppers, o comércio eletrônico brasileiro avançou significativamente cerca de 387% nos últimos 10 anos, e houve um salto no faturamento das Lojas B2C e Marketplace de bens de consumo (novos) de R$ 14,8 bilhões, em 2008, para R$ 53,2 bilhões em 2018.

A empresa Nielsen, que realiza esse estudo desde 2001, é uma empresa global que faz estudos de tendências e hábitos de consumo ao redor do mundo. E sua principal referência é o seu site E-bit que faz a avaliação da reputação do comércio eletrônico brasileiro.

O estudo aponta também a evolução do faturamento do Digital Commerce de R$ 94 bilhões em 2016, para R$ 133 bilhões em 2018. Digital Commerce se refere às vendas B2C e Marketplaces, incluindo lazer (viagens, passagens aéreas, ingressos) e Marketplaces de produtos novos e usados (Mercado Livre, Enjoei e Elo 7).

Durante o Fórum anual de Inovação para América Latina e o Caribe (ALC) em 2019, da empresa Mastercard anunciou um estudo que mostra que o Brasil é o país onde os consumidores fazem compra online com mais frequência, com 15% comprando a cada 03 dias ou menos, 21% comprando uma vez por semana e 12% comprando uma vez a cada duas semanas. Revelando que os dispositivos mais utilizados para as compras online foram: 53% o uso do smartphone, 45% utilizam o desktop e apenas 2% o tablet.

Em 2019, o faturamento do comércio eletrônico no Brasil foi de R$ 75 bilhões uma alta de 22,7% em relação ao ano anterior, de acordo com o estudo da NeoTrust que analisa o varejo digital por trimestre com base nos dados coletados pela empresa de inteligência de mercado Compre&Confie.

O e-commerce oferece oportunidades de negócio tanto para as empresas já consolidadas no mercado, quanto para aquelas pessoas que buscam empreender ou já são empreendedores. Para o microempreendedor o potencial do comércio eletrônico é ainda maior, pois reduz custos com o ponto comercial e expande suas vendas para clientes em todo o país.

O comércio eletrônico possui uma cadeia de atividades econômicas que fomentam o mercado brasileiro como um todo, que vai da indústria que produz bens de consumo, ao atacadista e/ou varejista, aos serviços de entregas e distribuição, aos profissionais da tecnologia, aos serviços de depósitos, seguros de mercadorias, centrais de atendimentos, prestação de serviços, entre outros.

Dado o crescimento exponencial desse setor e o interesse em analisar o mercado brasileiro de e-commerce, encontramos uma dataset de conjunto de dados públicos do e-commerce brasileiro da empresa Olist disponível no site do Kaggle.

A Olist se considera a maior loja de departamentos do mercado brasileiro, conectando através de seu site os comerciantes que vendem seus produtos aos clientes que usam a plataforma para consumo e por outro também oferece um produto que conecta diversos lojistas aos maiores marketplaces do país.

A plataforma funciona da seguinte forma, um cliente compra o produto na Olist Store, o vendedor é notificado sobre o pedido e deve providenciar o envio do produto. Depois que o cliente recebe o produto, ou a data estimada de entrega é vencida, ele recebe uma pesquisa de satisfação por e-mail, onde irá avaliar a experiência da compra e poderá fazer comentários. Já caso o lojista queira anunciar seus produtos em grandes marketplaces pode também utilizar a Olist como oportunidade de aumentar a sua relevância dentro desses marketplaces, aumentando o tráfego e a conversão de vendas desses lojistas.

Dado essas duas oportunidades que a Olist oferece no varejo online, a empresa está posicionada em uma visão completa do mercado transacionando milhões de pedidos através da sua plataforma. Nesse cenário, o estudo da base de dados desse player no varejo nacional pode trazer inputs gerenciais importantes para quem está inserido nesse mercado.

Relevância

Considerando o crescente cenário do mercado online no Brasil, temos a necessidade de identificar perfis cada vez mais distintos e estrategicamente entregar produtos e serviços que sejam adequados a sua necessidade.

Isso traz a notoriedade de tecnologias que tenham como objetivo identificar, reter e trabalhar dados para que as empresas sejam ainda mais centradas no cliente. (Dhandayudam et al., 2012). Pela diversidade que temos dos dados disponibilizados pela Olist, a alternativa que consideramos ser adequada a nossa proposta de identificar padrões de consumo e realizar a segmentação a partir dos mesmos é a clusterização. Esta técnica estatística multivalorada estabelece grupos que demonstrem ter similaridades e possibilita que a base de dados seja segmentada, a fim de identificarmos dependências e relações entre as variáveis associadas às semelhanças encontradas (Mingoti, 2005).

A ideia de utilizar técnica de clusterização (análise de agrupamentos) aplica-se a necessidade de segmentar os clientes e identificar o perfil dos mesmos com o objetivo de fazer marketing direcionado ao determinado público ou recomendações a novos clientes sugerindo produtos com base em perfis de consumo dos mesmos. (MARTINS, Pedro; HEIZEN, Renato , 2018)

Considerando os dados que serão apresentados, optamos pela abordagem comportamental trabalhando com variáveis, como quantidade de compras realizadas (Kotler & Armstrong, 2015),

que podem ou não ser sequenciais, e demonstrarão as similaridades entre comportamentos de cada cliente (YC Liu & YL Chen, 2017).

Procurando aumentar a assertividade da segmentação, vamos também agregar dados geográficos e demográficos ao modelo, uma vez que estes podem contribuir para novas e significativas similaridades e relações (Bailey & Gatrell, 1995). Dessa forma, clusters terão compras de itens similares, ainda que não seja exatamente os mesmos itens, refletindo padrões de cada grupo (C.Y. Tsai & C.C. Chiu, 2004).

Para operacionalização desta visão, serão utilizados os dados da Amostra do Censo IBGE 2010, realizada no período de 1 de Agosto de 2010 a 30 de Outubro de 2010, do qual consideramos dados de renda das regiões censitárias, utilizando variáveis de renda média e quantidade de moradores por residência nas regiões analisadas.

Objetivo

Com base nos dados disponíveis de compra, o presente estudo de caso tem como objetivo verificar a relevância de dados geográficos apresentam sobre um modelo de clusterização. Existem pontos que queremos observar, por exemplo qual o impacto de vizinhos sobre o cluster, se a influência deles agrega ao conjunto ou se a análise considerando apenas os aspectos da venda consegue atingir de forma mais assertiva cada cliente.

A segmentação através da técnica de clusterização adicionada a análise variáveis geoespaciais permite a identificação de segmentos de clientes apropriados diante da base analisada que permite a gestão dos clientes de maneira estratégica. Adicionada a isso identifica-se quais grupos possuem características de rentabilidade mais atrativas que permitem o direcionamento de campanhas de marketing direcionadas para esses clientes.

Dados

O dataset possui informações de mais de 100 mil pedidos feitos na Olist Store no período de 2016 a 2018 em várias regiões do Brasil. O conjunto de dados permite visualizar um pedido de várias dimensões: status do pedido, preço do produto e do frete, os meios de pagamentos e se utilizou mais de uma forma para pagamento, desempenho do frete até a localização do cliente, os atributos dos produtos, para qual vendedor foi o pedido e avaliação das compras escritas pelos clientes.

Também possuem um conjunto de dados de geolocalização que relaciona os códigos postais às coordenadas de latitude e longitude, possibilitando uma visão espacial da distância entre compradores e vendedores no território brasileiro. Esses dados não possuem uma relação direta com os pedidos efetuados na plataforma, ou seja, nem sempre o endereço de entrega do pedido será o mesmo do que consta no cadastro do cliente.

Entretanto, se considerarmos que os pedidos efetuados no site serão entregues no mesmo endereço de cadastro isso nos permitirá ter uma visão geoespacial da venda mensurando o raio de atuação de determinado vendedor.

No sistema da Olist cada pedido é atribuído a um único customer_id. Isso significa que o mesmo cliente receberá IDs diferentes para pedidos diferentes, e a estrutura dos registros que constam nas tabelas nos levariam a um entendimento errado que cada pedido teria um cliente diferente associado.

Com o objetivo de identificar os clientes que fizeram compras na loja com um único ID independentemente da quantidade de pedidos que esse tenha foi criado o campo **customer_unique_id** na tabela de customers.

Pontos que tivemos que levar em conta sobre os pedidos desse conjunto de dados, são que um único pedido pode ter vários itens, cada item pode ser atendido por um vendedor distinto por se tratar de uma única plataforma com diferentes vendedores. E também que todos os textos que fazem referências às empresas e parceiros no texto foram substituídos pelos nomes das grandes casas da série Game of Thrones para manter o anonimato.

Para fins desse estudo de caso, após a análise do dataset original identificamos que para aplicar a técnica de segmentação dos clientes seria melhor focar na cidade de São Paulo sobre a visão de distritos. Essa base de possui 18 mil pedidos realizados entre 2016 a 2018 e utilizando o Censo Geográfico do IBGE de São Paulo de 2010, a análise detalhada da base estará na sessão Análise e Modelagem.

Estrutura dos Dados

Como o dataset está dividido em diversos arquivos e as informações estão dispersas, dessa forma não temos uma estrutura "pronta" para que pudéssemos usar em nossas análises. Foi necessário pensar na estrutura ideal e construir uma única base a ser usada para realizar as análises de Clusterização e Geoespacial.

O esquema que montamos para esse dataset ficou da seguinte forma:

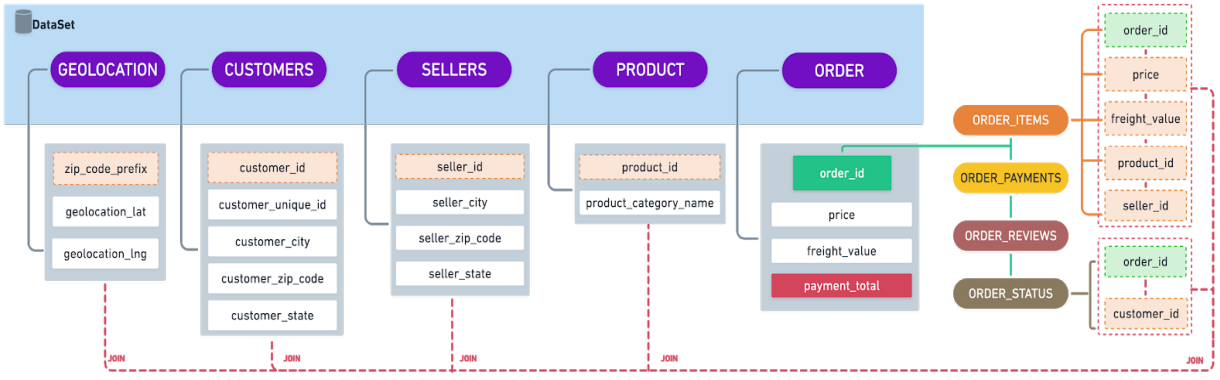

Figura 1: Nova estrutura de dados

Os dados de geolocalização fornecidos pela Olist são as coordenadas geoespaciais dos 5 primeiros números de CEP, ou seja, os prefixos de CEP e não possuem correlação com os pedidos registrados.

Devido à falta de informação se o produto foi entregue em local diferente do cadastro do cliente, assumimos que os produtos foram entregues no endereço de cadastro do cliente e usamos o prefixo do cadastro do cliente para o join com os dados de geolocalização fornecidos.

Tratamento dos Dados

Para o tratamento e análise dos dados utilizamos a ferramenta **JupyterLab** com a linguagem programação Python 3. Primeiro montamos a estrutura de dados acima definida e fizemos a análise exploratória dos dados, tais como observar o tamanho do dataset, identificar os tipos de dados existentes, validar se existem campos nulos e identificar a concentração dos pedidos por região do Brasil.

- O tamanho do dataset é de: **112.650 linhas** que são as ordens de pedido e **14 colunas** que são os atributos (customer_id, order_id, etc)

- Foram tratados apenas os **1.603 campos nulos** identificados que são relevantes para a criação do cluster, que são: **product_category_name** e **payment_total**.

Na distribuição dos dados por região verificamos que 69% do total de ordens de pedidos está concentrado na região sudeste (RJ, MG, SP e ES), sendo nesses que focaremos o estudo, com foco no estado com maior concentração de pedidos, o Estado de São Paulo.

Realizamos mais um recorte na amostra, trazendo dados de venda da capital, uma vez que pelas análises iniciais, ali tínhamos a concentração das amostras mais relevantes. Essa amostra com aproximadamente 18 mil observações foi utilizada ao longo do trabalho que será apresentado a seguir.

Análise e Modelagem

Para o presente estudo, queremos determinar a relação das informações com a variável de volume de vendas, sendo esta nossa referência principal e nosso y nas visões estatísticas desenvolvidas analisando essas variáveis sobre a ótica da técnica estatística de clusterização.

Para a análise geoespacial, consideramos a quebra de polígonos de distritos para a cidade de São Paulo. Gostaríamos de ter uma visão mais granular sobre a base de dados, mas a divisão de subdistrito não é mais praticada hoje em dia para a cidade em questão. Escolhemos essa divisão por distrito, pois acreditamos que era uma divisão que se adequa melhor ao volume de dados que temos.

Diante dos dados, a nossa análise considerou como faríamos a geolocalização dos dados com as informações disponíveis. No dataset utilizado tínhamos o CEP dos clientes com 5 dígitos o que não nos trazia a posição exata e nem permitia que conseguíssemos tal localização através de

ferramentas. Portanto, utilizamos a informação dos 5 primeiros dígitos para fazer a análise geoespacial.

Um outro ponto que avaliamos na base escolhida foi o foco do estudo. Inicialmente pensamos em focar na região sudeste por ter um grande volume e variedade nos dados. No entanto, ao analisar a base do ponto de vista geográfico observamos que o estado de São Paulo, representava 77 % de nossa amostra e permitiria uma análise interessante do ponto de vista de aplicação e em termos de explorar a granularidade dos dados.

Na Figura 1, criamos uma visão geográfica da dispersão das vendas da Olist de 2016 a 2018 no Estado de São Paulo utilizando o Software Tableau e ficou evidente a concentração dos dados em torno da capital do estado.

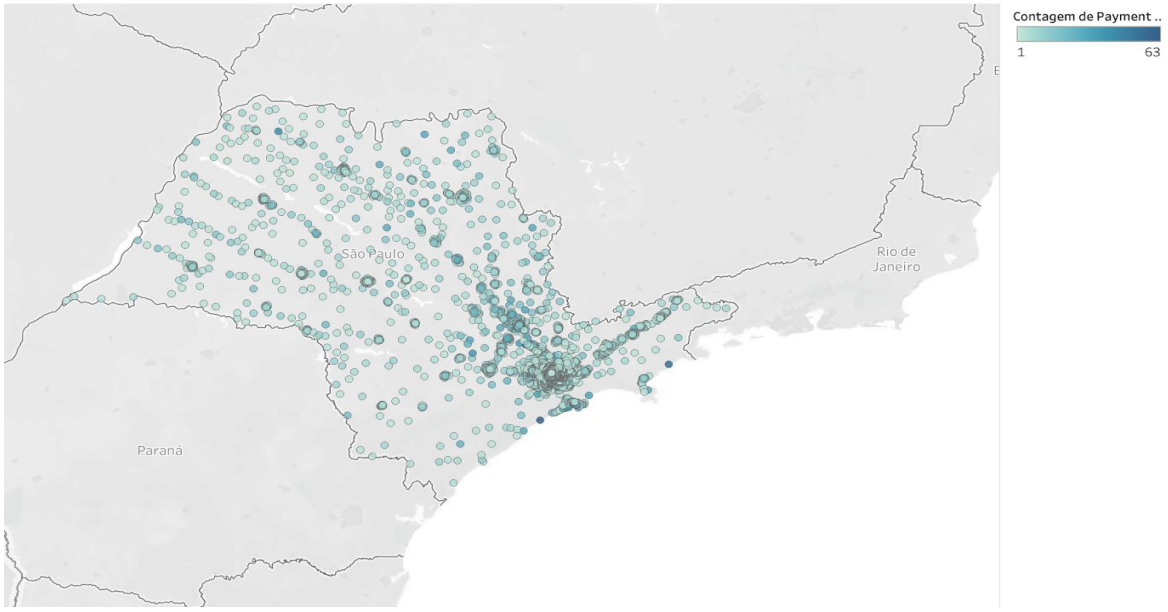

Figura 2 - Distribuição de vendas no estado de São Paulo

Como o propósito da análise é identificar padrões de consumo e os cinco primeiros dígitos do CEP são aqueles que concentram o maior nível de informações da localização, acreditamos ser adequado - apesar de não o ideal - para o propósito que possuímos e diante dos dados disponíveis. Diante disso, utilizamos o QGIS para projetar os dados de quantidade de vendas por cada localidade, através do método de quebras naturais, ilustrado na Figura 3: Volume de vendas no estado de SP pelo método de quebras naturais.

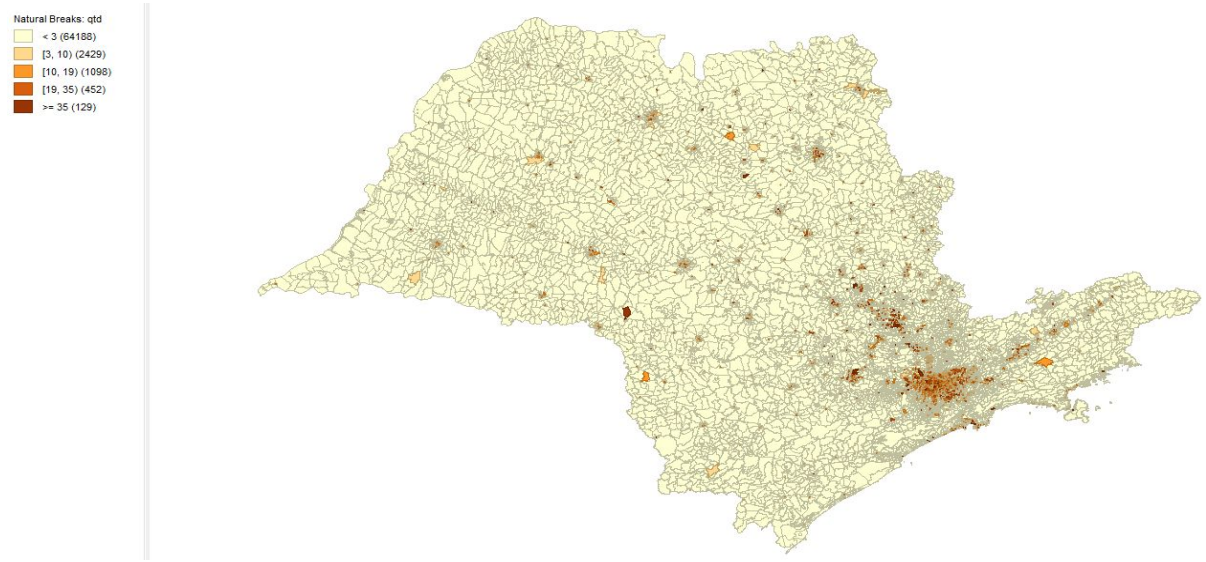

Figura 3: Volume de vendas no estado de SP pelo método de quebras naturais.

Enquanto na Figura 4 - Ticket médio no estado de SP pelo método de quebras naturais, observamos o comportamento da variável ticket médio na dispersão geográfica dos dados.

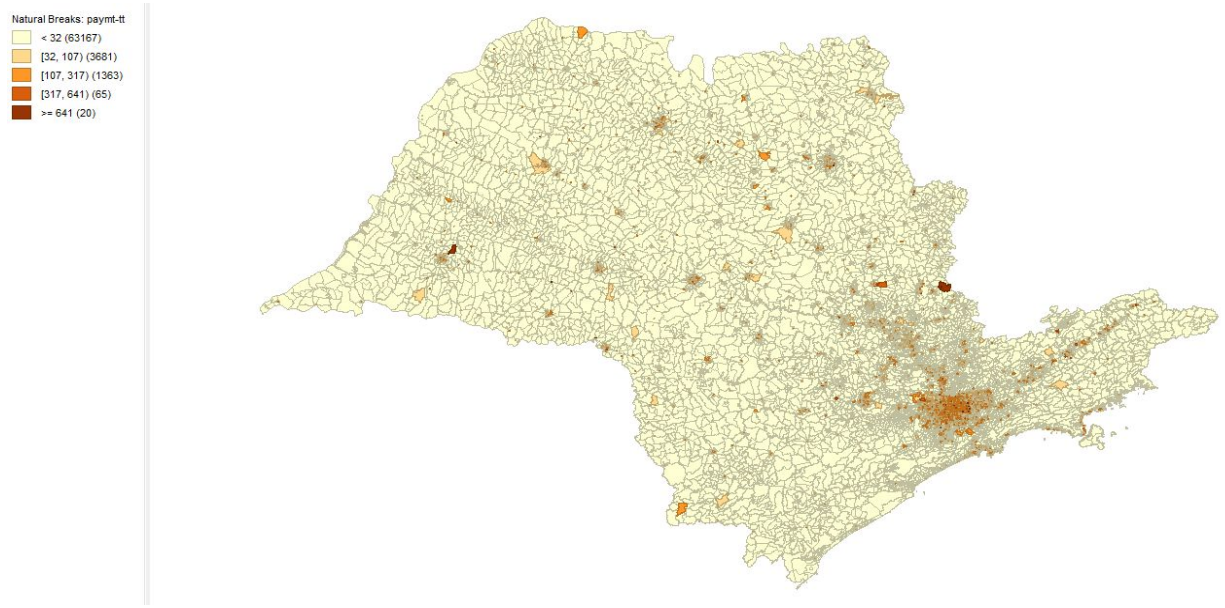

Figura 4: Ticket médio no estado de SP pelo método de quebras naturais.

Conseguimos observar que a concentração dos pontos efetivamente se dá na região da Grande São Paulo e que os pontos de maior ticket médio também se concentram nesta região. Existem, no entanto, algumas regiões de interesse que poderiam também ser analisadas como ao centro-leste, perto de São Carlos, onde temos alto ticket médio, ainda que baixa concentração de vendas; e centro-oeste, próximo a Bauru, que tem um alto volume, mas com ticket médio não tão atrativo.

Ainda que tivéssemos uma amostra menor, ela ainda requisitava muito do processamento disponível nas máquinas particulares que dispúnhamos, tornando inviável a continuidade de forma

fluida das análises mais complexas, optamos por seguir com o Estudo de Caso com foco na capital de São Paulo.

A análise das base de dados original também teve como direcionamento analisar variáveis qualitativas como a Dispersão de vendas na cidade de São Paulo por departamentos visando identificar se existia uma predominância de pedidos por departamento que posteriormente resultaria em uma segmentação de clientes diferenciadas.

Na Figura 7: Dispersão de vendas na cidade de São Paulo por departamentos, tendo cor da observação referente ao valor e tamanho pelo volume de vendas, temos a dispersão pelos departamentos mais recorrentes no dataset de vendas da cidade de São Paulo, bem como a informação de total que as vendas desse departamento representaram no total que estamos trabalhando.

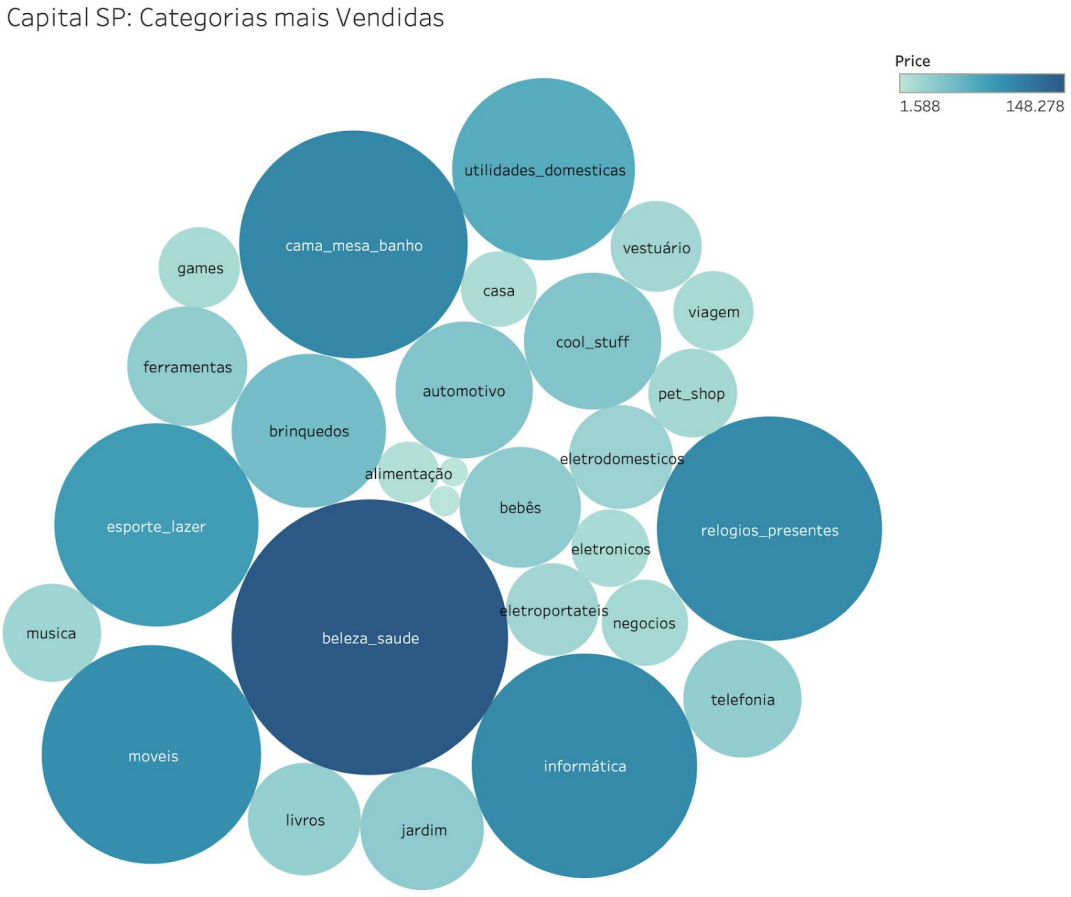

Figura 7: Dispersão de vendas na cidade de São Paulo por departamentos, tendo cor da observação referente ao valor e tamanho pelo volume de vendas.

Percebemos que grande parte das vendas se refere a produtos da categoria de Beleza & Saúde e Cama, Mesa & Banho. Existem alguns departamentos em faixa de volume e preços médios como Esporte & Lazer, Móveis, Informática, Relógios & Presentes e Utilidades Domésticas. Os demais departamentos não são relevantes para a análise atual.

A análise exploratória dos dados também considerou a dispersão dos valores de venda por valores pagos de frete como ilustrado na Figura 8: Dispersão dos valores de venda por valores

pagos de frete. Essa análise indicou que o ticket médio fica em torno de até R$400,00 reais e que o valor do frete não passa na média de R$50,00.

Dado os resultados da análise de valor de frete e ticket médio da Figura 8, estão em linha com a análise que fizemos do ticket médio das categorias mais vendidas e do frete gasto também estar nessa faixa de valores.

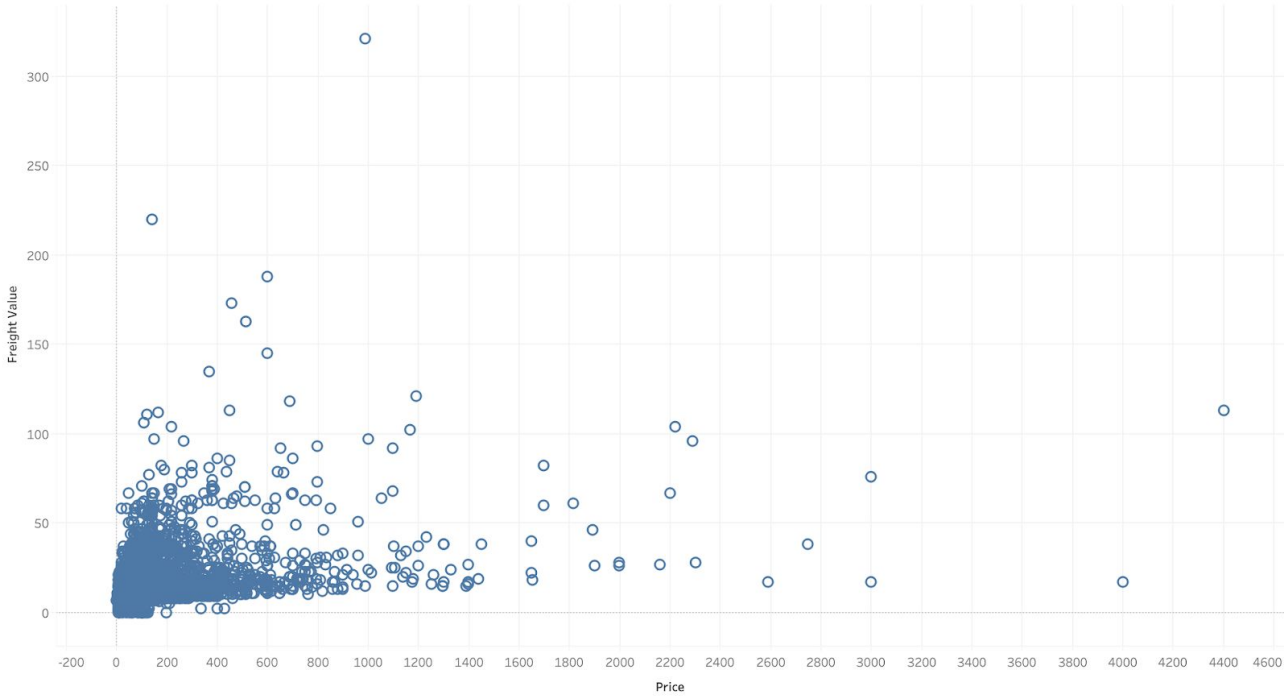

Figura 8: Dispersão dos valores de venda por valores pagos de frete.

Apesar de já termos apresentado a distribuição do volume de vendas pelo estado, uma vez que este é nosso objeto de análise, achamos relevante realizar a abertura na cidade de São Paulo para observarmos.

A Figura 4: Volume de vendas na cidade de São Paulo pelo método de quebras naturais traz a visualização em questão, nela podemos perceber que existe grande concentração de pedidos na região central da cidade e pontos de interesse nas zonas sul e oeste que também fazem parte da quebra.

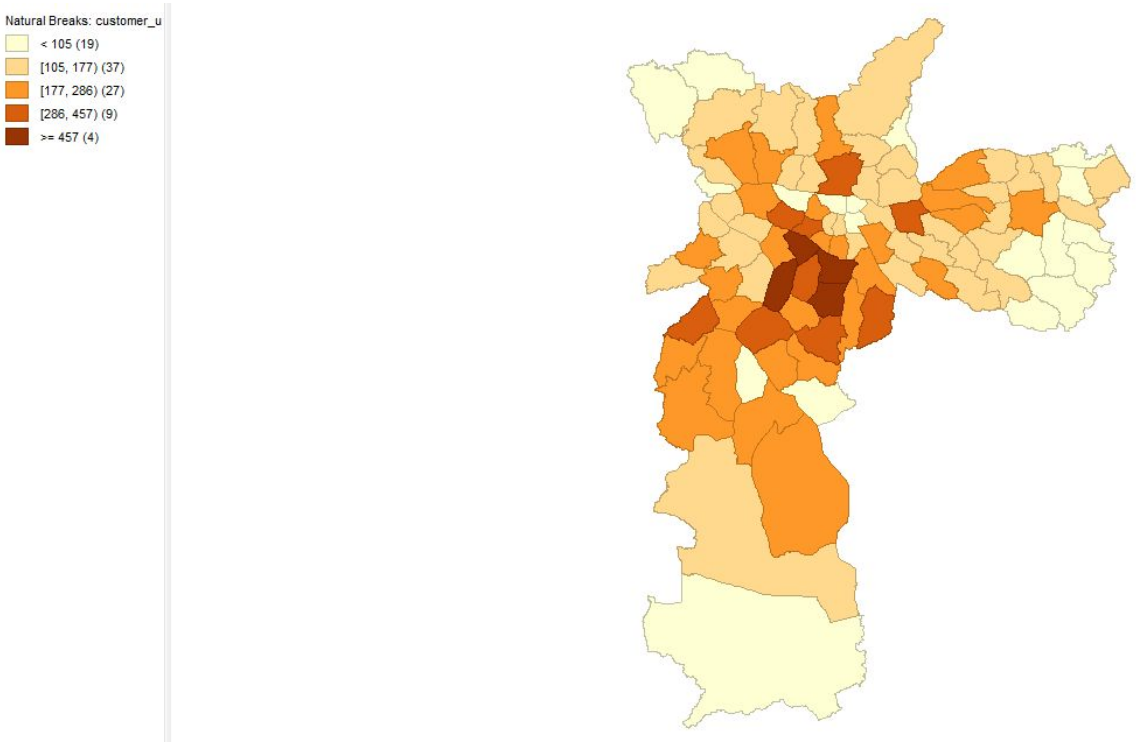

Figura 4: Volume de vendas na cidade de São Paulo pelo método de quebras naturais.

Analisando os demais dados do dataset em questão, ainda sem inclusão de variáveis externas, tivemos interesse em ver qual seria a distribuição do ticket médio dessas regiões e, principalmente, qual o desvio padrão desses valores.

Com essa perspectiva, pretendemos determinar se os valores médios são mais altos de forma estável dentro de cada região da cidade ou se existe grande variação em regiões com maior volume de vendas.

Tendo em vista as Figura 5: Desvio padrão dos valores de venda dos distritos da cidade de São Paulo pelo método de quebras naturais e Figura 6: Ticket médio de venda dos distritos da cidade de São Paulo pelo método de quebras naturais, identificou-se que regiões como o extremo da Zona Sul que tem ticket médio a alto possui um desvio padrão baixo.

Isso é interessante, pois indica um ponto de interesse para análises futuras, uma vez que demonstra uma venda de produtos de valores mais relevantes. Confrontando isso com os dados apresentados na Figura 4: Ticket médio no estado de SP pelo método de quebras naturais., percebemos que esta região possivelmente pode ser melhor explorada com ações de marketing ou incentivos subsidiados.

Por outro lado, temos pontos de alta variação de valores nas vendas, como pontos na zona norte e leste.

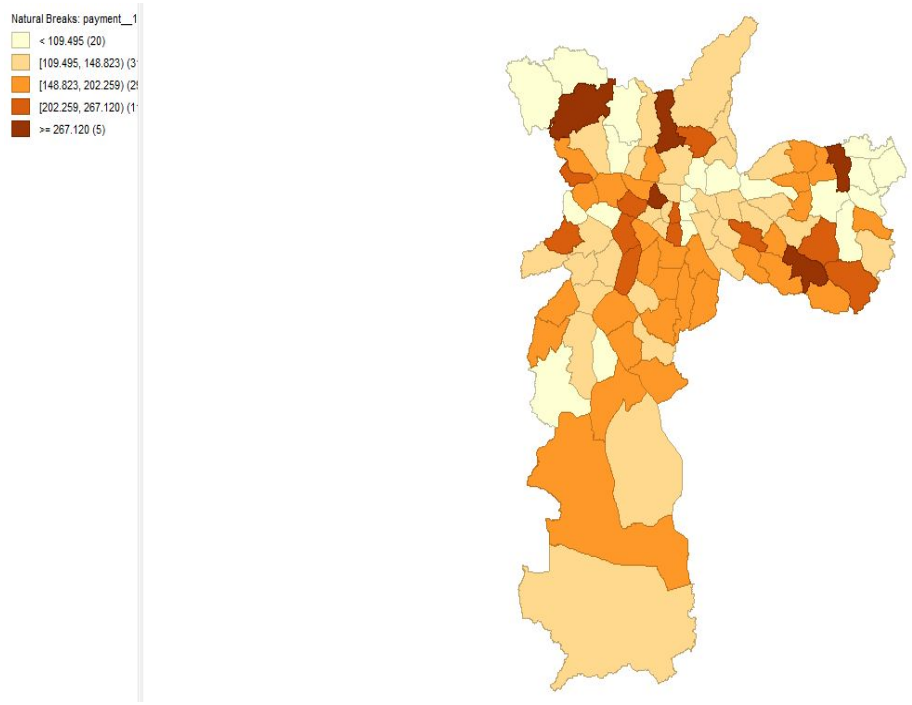

Figura 5: Desvio padrão dos valores de venda dos distritos da cidade de São Paulo pelo método de quebras naturais.

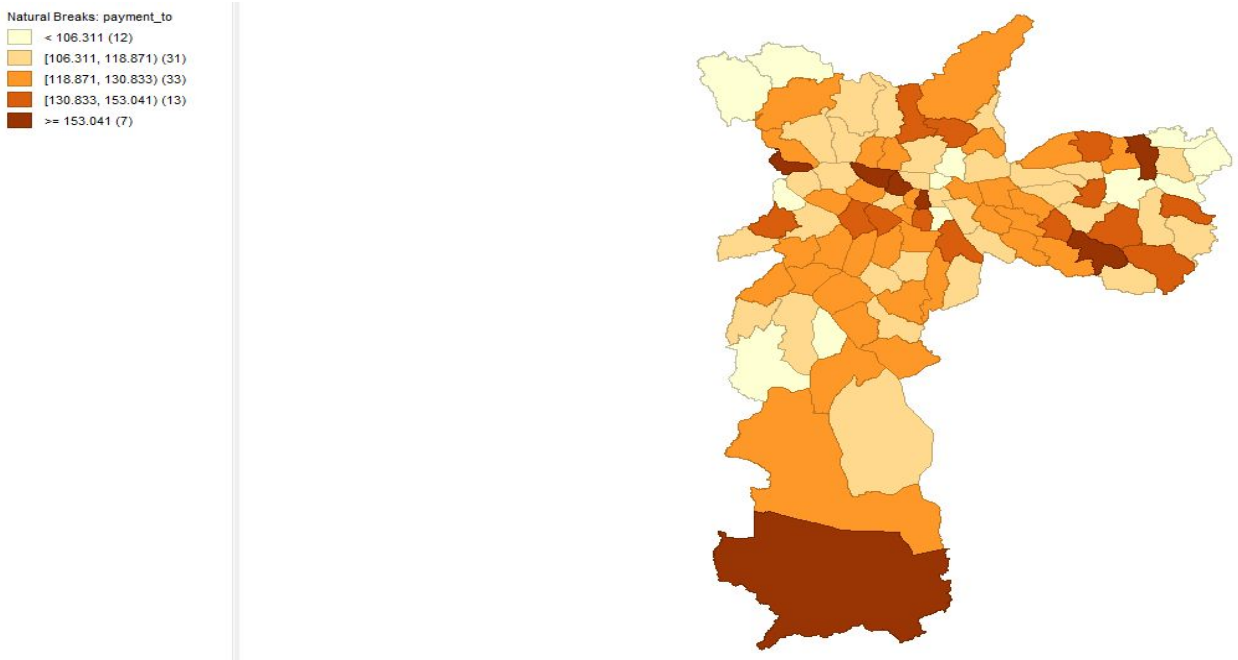

Figura 6: Ticket médio de venda dos distritos da cidade de São Paulo pelo método de quebras naturais.

Além de explorar a base original da Olist do ponto de vista de clusterização e da análise geoespacial, utilizamos dados censitários para complementar essa metodologia de segmentação de clientes.

Nesse sentido, utilizamos a base de dados do censo do IBGE de 2010 e realizamos o join das informações que gostaríamos de considerar nesta análise. Foram utilizados os campos V003 e V005 que correspondem a quantidade de habitantes por residência e valor médio de renda, respectivamente.

Foi necessário realizarmos um tratamento na base que estávamos utilizando, uma vez que o campo de distrito nela estava em formato de string e nos resultados censitários o campo estava como numérico. O tratamento foi feito através do próprio QGIS e, a partir daí, o join mencionado anteriormente foi feito.

Utilizando ambas bases de dados com o join realizado, fizemos análises espaciais considerando quatro pesos diferentes: Queen de ordem 1 e 2, Rook de ordem 1 e 2. Pretendemos observar qual modelo possui melhor aderência aos dados. Nas Figuras 9 à 12, temos o histograma de vizinhos de cada uma delas para referência. Diante dos dados, optamos por analisar a significância com os dados Queen de ordem 1.

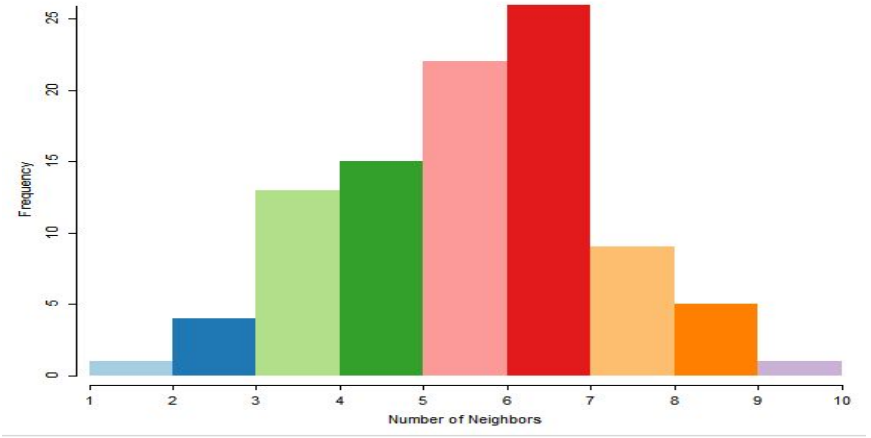

Figura 9: Histograma vizinhos Queen de ordem 1

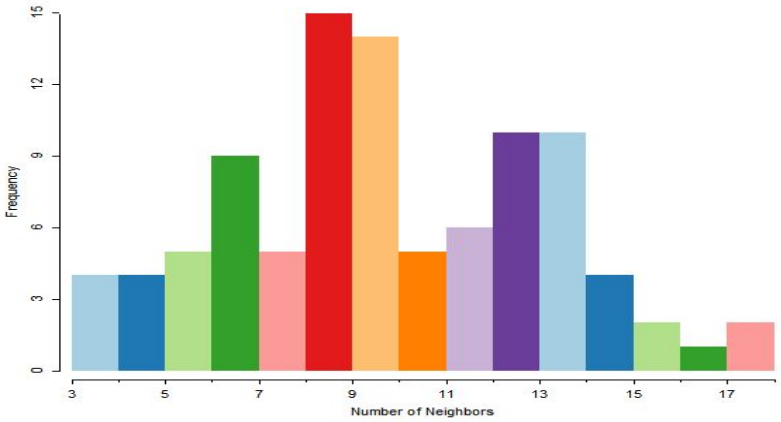

Figura 10: Histograma vizinhos Queen ordem 2

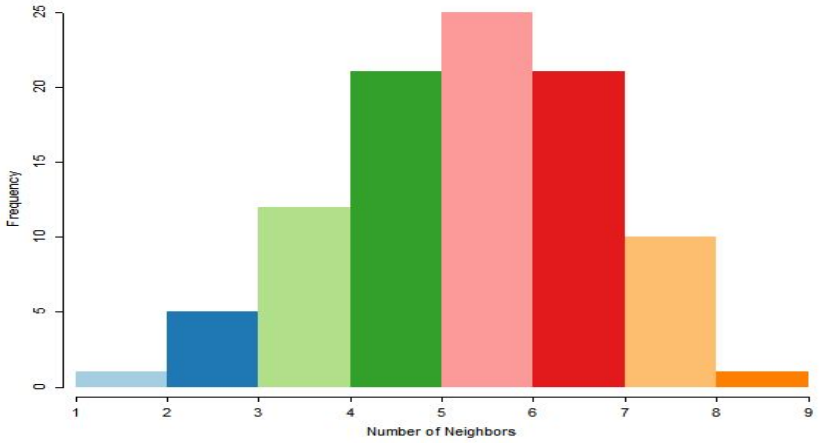

Figura 11: Histograma vizinhos Rook de ordem 1

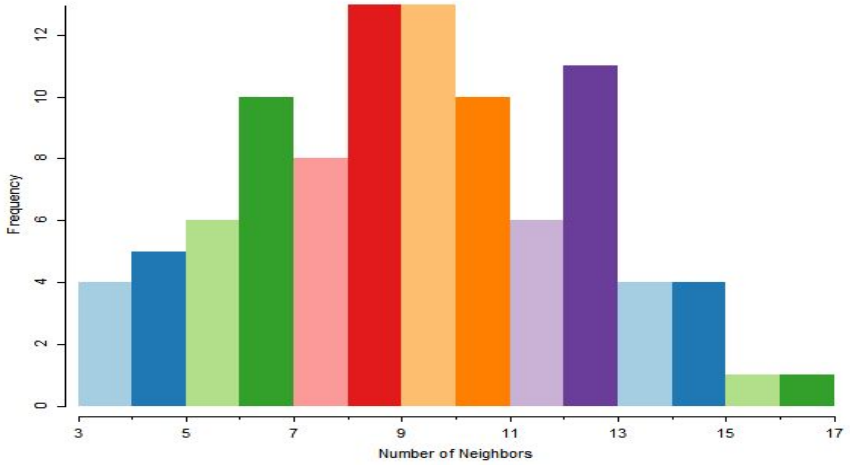

Figura 12: Histograma vizinhos Rook ordem 2

Também realizamos análises de regressões feitas com e sem peso espacial, seus resultados se encontram nos Anexos de 2 a 8 para consulta.

Iniciamos a análise de regressões buscando explicar o volume de vendas com uma regressão múltipla sem peso espacial com apenas as variáveis de ticket médio e desvio padrão do valor pago. Obtendo um R-quadrado de 7.83%.

O segundo passo foi incluir variáveis adquiridas através do censo IBGE 2010, quantidade de moradores por residência e renda média. Isso tem relevância para o modelo e nos traz um R-quadrado de 31.71%.

Notamos, portanto, o quão relevante são características socioeconômicas para nossa análise, sendo responsáveis pelo aumento significativo na capacidade do modelo explicar nossa variável objetivo, volume de vendas.

Passamos agora para a inclusão das referências espaciais no modelo. Foram feitas 4 regressões, usando peso de queen ordem 1, queen ordem 2, rook ordem 1 e rook ordem 2, respectivamente, eles resultaram nos seguintes R-quadrados: 45,19%, 42,40%, 46,78%, 39,46%.

Temos duas abordagens espaciais que mais se destacam, Queen de ordem 1 e Rook de ordem 1. Daqui em diante, faremos uso destes dois pesos para as demais visões espaciais.

Observamos agora os resultados que correspondem a autocorrelação espacial dos dados usando as matrizes de contiguidade elencadas anteriormente com a metodologia do I de Moran.

Usando a referência de contiguidade Queen de ordem 1, temos a autocorrelação espacial apresentada na Figura 13, de I = 0,403. Observamos que temos dois quadrantes com grande volume de observações, baixo-baixo e alto-alto.

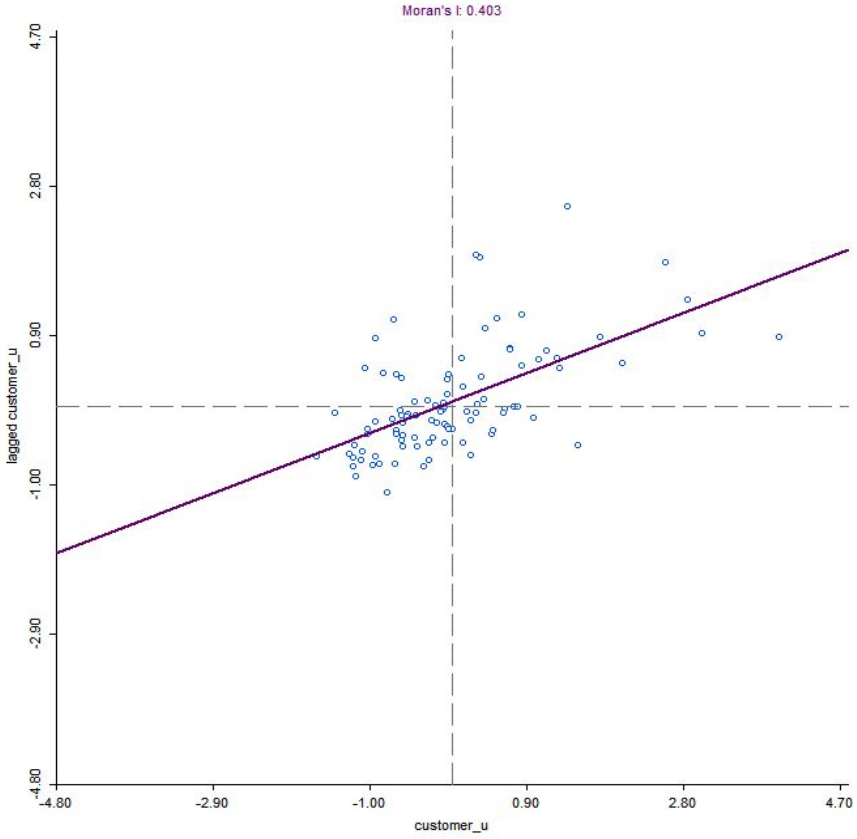

Figura 13: Índice Global de Moran com contiguidade Queen de ordem 1

Para validar se o valor resultante do I de Moran é significante fizemos uso do Teste de Pseudo Significância sobre 999 permutações. Obtivemos um P-valor < 0,001, que pode ser apreciado na figura 14, o que corrobora a significância e rejeita a hipótese nula de não haver autocorrelação espacial.

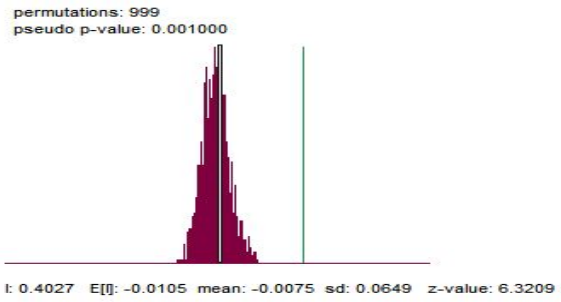

Figura 14: Pseudo-Significância por Queen de ordem 1

Como esperado, as visões de clusterização e significância também tem extremos bem destacados, de alta-alta e baixa-baixa correlação espacial.

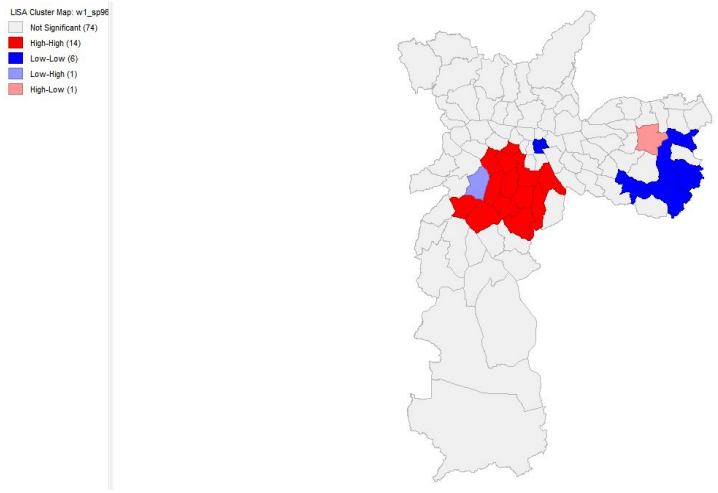

Figura 15: Clusters - Queen de ordem 1

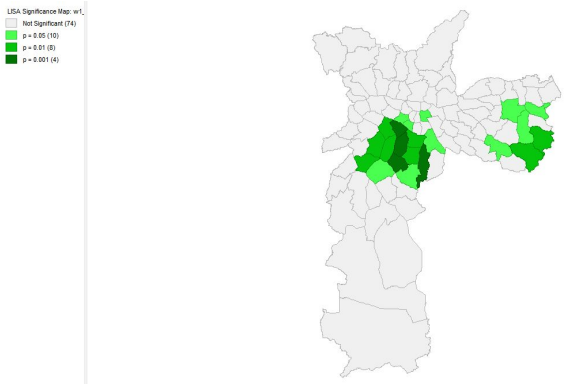

Figura 16: Significância por Queen de ordem 1

Realizando a mesma visão sobre os dados com a contiguidade Rook de ordem 1, temos os seguintes resultados.

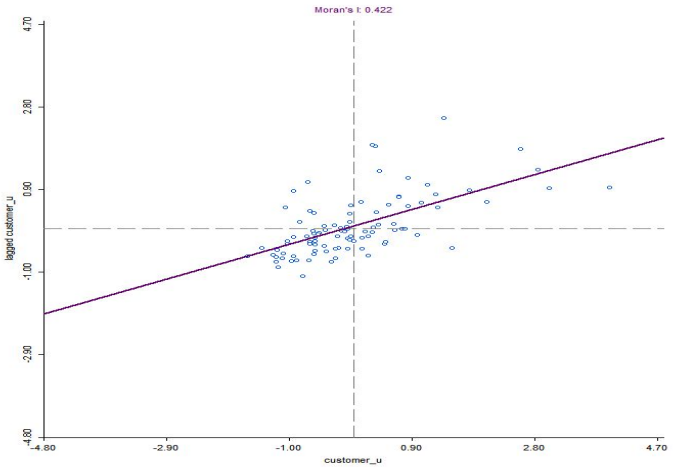

Figura 17: Índice Global de Moran com contiguidade Rook de ordem 1

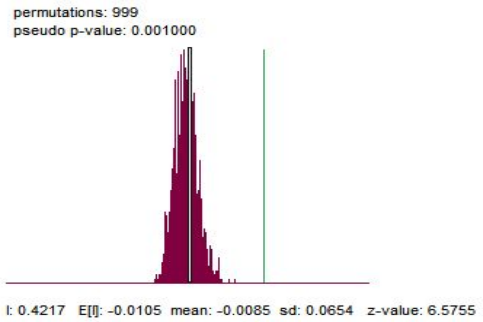

Figura 18: Pseudo-Significância por Rook de ordem 1

O Índice de Moran tem valor mais representativo, chegando a I = 0,422. Ainda assim, temos um p-valor < 0,001 quando realizamos o teste de pseudo-significância que rejeita a hipótese nula e permite que tenhamos uma abordagem ainda mais relevante.

Como a diferença entre resultados é baseada apenas em como a matriz de vizinhança é formada, considerando apenas vizinhos laterais e não incluindo diagonais, os resultados da clusterização e da significância visualmente são bastante similares aos apresentados anteriormente.

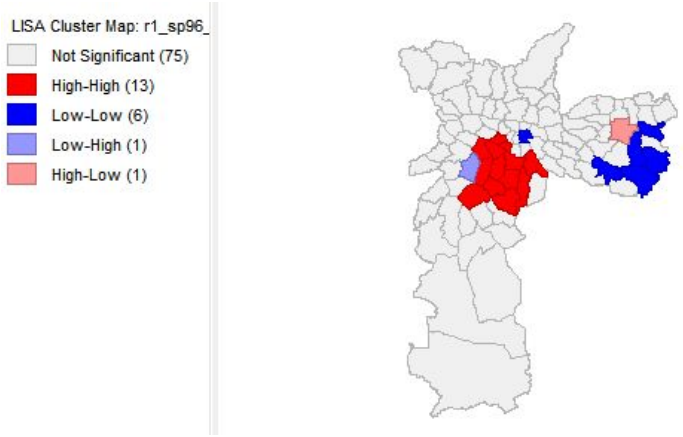

Figura 19: Clusters - Rook de ordem 1

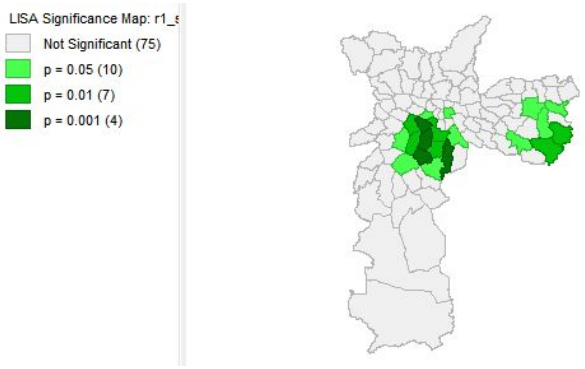

Figura 20: Significância por Rook de ordem 1

Enquanto para a presente análise, focamos em extrair o máximo possível das ferramentas de geoestatística, portanto faremos ambos os clusters através do GeoDa. Abordaremos clusters pela perspectiva hierárquica em ambos os casos, com e sem pesos espaciais.

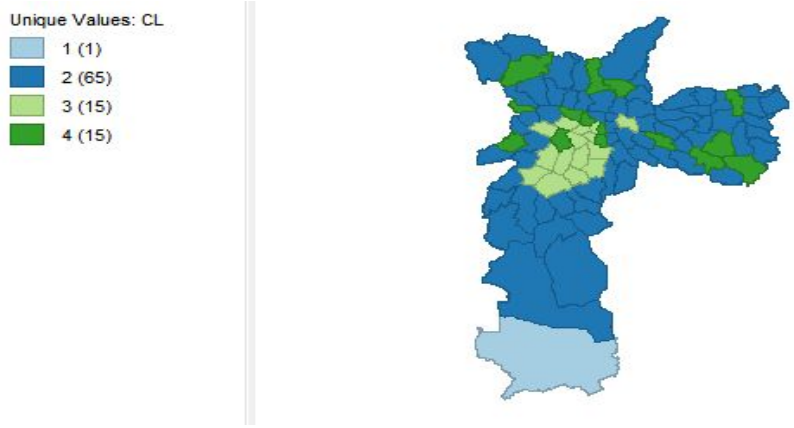

Figura 21: Cluster sem peso espacial

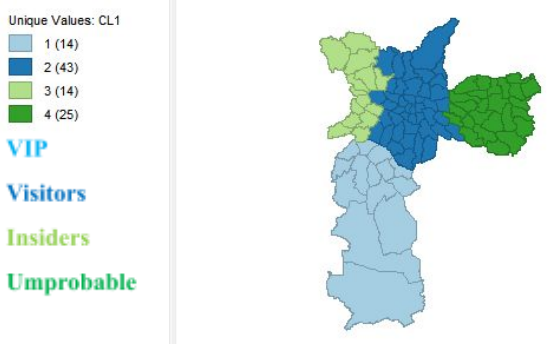

Figura 22: Cluster com peso espacial

Utilizamos o método Ward com distâncias euclidianas para determinar os 4 clusters solicitados hierarquicamente.

Podemos observar que sem o peso espacial, conseguimos identificar regiões estratégicas que fogem do padrão do entorno, que podem ser potenciais de venda inexplorados.

Para os clusters foram consideradas as variáveis volume de vendas, ticket médio, desvio padrão do valor pago, quantidade de moradores por residência e renda do distrito.

Parte importante de um cluster está em identificar padrões dentro deles e diferenças entre. Ao fim das análises quatro clusters buscando criar uma estratégia de marketing atendendo a seguinte segmentação de clientes por Fidelização denominados VIP, Insiders, Visitors e Umprobable. Os clientes foram segmentados pelas variáveis de renda, ticket médio, compras por departamento e localização.

A segmentação pode ser visualizada na Figura 22: Cluster com peso espacial. No qual os clientes VIP estão localizados principalmente ao sul da cidade de São Paulo, possuem alta renda, um ticket médio moderado a alto e preferência por Categorias de Beleza e Saúde.

Já os clientes Insiders são clientes mais concentrados ao centro e nordeste da cidade que também se mostraram um grupo relevante em volume, com renda atrativa, alto ticket médio gasto e consumo de categoria de Cama Mesa e Banho. Por último, os clientes Insiders e Umprobables fizeram poucas compras quando comparados aos grupos anteriores, com um ticket médio menor, uma renda média menor e foco em produtos de categorias como utilidades domésticas e outros, localizados ao oeste e leste da cidade respectivamente.

Do ponto de vista prático, após as diversas análises desse Estudo de Caso a aplicação da clusterização dos clientes utilizando a Geoanálise indicaria para a Diretoria de Marketing da empresa que se utiliza esses dados que a mesma direcione suas ações de marketing em clientes classificados como VIP e Insiders.

Além disso, desenvolve-se sua estratégia de marketing para engajar tráfego e conversão no site com foco nas categorias de Beleza e Saúde e Cama Mesa e o lançamento de uma Promoção com Frete Grátis para o sul da cidade de São Paulo com base no alto potencial que clientes tipo VIP demonstraram durante a análise.

Conclusão

Com a modelagem desenvolvida foi possível para a suposta Diretoria de Marketing tomar as seguintes decisões: Clusterizar com foco nos clientes VIP e Insiders segmentando a atuação, também vislumbram uma campanha de marketing direcionada as categorias de Beleza e Saúde e Cama Mesa e Banho e a criação de promoções de frete grátis na zona sul da capital de São Paulo.

Dessa maneira, a segmentação com a metodologia de clusterização e utilização de dados geoespaciais permitiu um direcionamento mais assertivo dos recursos destinado ao departamento de marketing e as campanhas que atendam às diferentes necessidades dos clientes da base, mostrando a importância no uso de modelos estatísticos e geoespaciais para a definição de perfis de consumo nas decisões de marketing das empresas.

Fontes

https://www.kaggle.com/olistbr/brazilian-ecommerce

https://www.hardware.com.br/noticias/2019-12/brasil-e-o-pais-da-america-latina-em-que-os-consumidores-mais-compram-online.html

https://revistapegn.globo.com/Banco-de-ideias/Mundo-digital/noticia/2014/11/50-oportunidades-no-e-commerce-para-voce-investir.html

https://www.nielsen.com/br/pt/insights/article/2018/webshoppers-38-e-commerce-fatura-vinte-e-tres-bilhoes-no-primeiro-semestre-de-2018-alta-de-doze-porcento/

https://www.ebit.com.br/webshoppers

https://g1.globo.com/economia/noticia/2020/02/13/comercio-eletronico-fatura-r-75-bilhoes-no-brasil-em-2019.ghtml

Bailey, T. C. & Gatrell, A. C. Interactive Spatial Data Analysis. London: Longman (1995).

C.-Y. Tsai, C.-C. Chiu. Expert Systems with Applications 27 (2004) 265–276

Dhandayudam, Prabha & Krishnamurthi, Ilango. International Journal of Engineering Science and Technology (IJEST) ISSN : 0975-5462, Vol. 4 No.02, February 2012, p. 695-702

KOTLER, P.; ARMSTRONG, G. Princípios de Marketing. 15ª ed. São Paulo: Pearson Prentice Hall, 2015.

MARTINS, Pedro; HEIZEN, Renato. Sistema para identificação de perfil de consumidores utilizando análise de agrupamento (clusterização). Tubarão: Editora UNISUL, 2018.

MINGOTI, Sueli Aparecida. Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada. Belo Horizonte: Editora UFMG, 2005.

Yen-Chung Liu; Yen-Liang Chen. Int. Journal of Engineering Research and Application ISSN : 2248-9622, Vol. 7, Issue 1, (Part -1) January 2017, pp.49-58

Apêndice 1: Estrutura das Tabelas do Dataset

O conjunto de dados é composto por 09 arquivos no formato csv divididos em:

- **Customers** - identificação do cliente.

customer_id	customer_unique_id	customer_zip_code_prefix	customer_city	customer_state
06b8999e2fba1a1fbc88172c00ba8bc7	861eff4711a542e4b93843c6dd7febb0	14409	franca	SP
18955e83d337fd6b2def6b18a428ac77	290c77bc529b7ac935b93aa66c333dc3	9790	sao bernardo do campo	SP
4e7b3e00288586ebd08712fdd0374a03	060e732b5b29e8181a18229c7b0b2b5e	1151	sao paulo	SP

- **Geolocation** – usado para plotar mapas e encontrar distâncias entre vendedores e clientes.

geolocation_zip_code_prefix	geolocation_lat	geolocation_lng	geolocation_city	geolocation_state
1037	-23.545621	-46.639292	sao paulo	SP
1046	-23.546081	-46.644820	sao paulo	SP
1046	-23.546129	-46.642951	sao paulo	SP

- **Order_items –** informações sobre os itens comprados em cada pedido

order_id	order_item_id	product_id	seller_id	shipping_limit_date	price	freight_value
00010242fe8c5a6d1ba2dd792cb16214	1	4244733e06e7ecb4970a6e2683c13e61	48436dade18ac8b2bce089ec2a041202	2017-09-19 09:45:35	58.90	13.29
00018f77f2f0320c557190d7a144bdd3	1	e5f2d52b802189ee658865ca93d83a8f	dd7ddc04e1b6c2c614352b383efe2d36	2017-05-03 11:05:13	239.90	19.93
000229ec398224ef6ca0657da4fc703e	1	c777355d18b72b67abbeef9df44fd0fd	5b51032eddd242adc84c38acab88f23d	2018-01-18 14:48:30	199.00	17.87

- **Payments –** informações sobre as opções de pagamentos dos pedidos, que renomeamos para **order_payment.**

order_id	payment_sequential	payment_type	payment_installments	payment_value
b81ef226f3fe1789b1e8b2acac839d17	1	credit_card	8	99.33
a9810da82917af2d9aefd1278f1dcfa0	1	credit_card	1	24.39
25e8ea4e93396b6fa0d3dd708e76c1bd	1	credit_card	1	65.71

- **Order_reviews –** informações sobre as avaliações dos pedidos feito pelo cliente.

review_id	order_id	review_score	review_comment_title	review_comment_message	review_creation_date	review_answer_timestamp
7bc2406110b926393aa56f80a40eba40	73fc7af87114b39712e6da79b0a377eb	4	NaN	NaN	2018-01-18 00:00:00	2018-01-18 21:46:59
80e641a11e56f04c1ad469d5645fdfde	a548910a1c6147796b98fdf73dbeba33	5	NaN	NaN	2018-03-10 00:00:00	2018-03-11 03:05:13
228ce5500dc1d8e020d8d1322874b6f0	f9e4b658b201a9f2ecdecbb34bed034b	5	NaN	NaN	2018-02-17 00:00:00	2018-02-18 14:36:24

- **Orders** – informações de status do pedido, que renomeamos para **orders_status.**

order_id	customer_id	order_status	order_purchase_timestamp	order_approved_at	order_delivered_carrier_date	order_delivered_customer_date	order_estimated_delivery_date
e481f51cbdc54678b7cc49136f2d6af7	9ef432eb6251297304e76186b10a928d	delivered	2017-10-02 10:56:33	2017-10-02 11:07:15	2017-10-04 19:55:00	2017-10-10 21:25:13	2017-10-18 00:00:00
53cdb2fc8bc7dce0b6741e2150273451	b0830fb4747a6c6d20dea0b8c802d7ef	delivered	2018-07-24 20:41:37	2018-07-26 03:24:27	2018-07-26 14:31:00	2018-08-07 15:27:45	2018-08-13 00:00:00
47770eb9100c2d0c44946d9cf07ec65d	41ce2a54c0b03bf3443c3d931a367089	delivered	2018-08-08 08:38:49	2018-08-08 08:55:23	2018-08-08 13:50:00	2018-08-17 18:06:29	2018-09-04 00:00:00

- **Products** - dados sobre os produtos vendidos pela Olist.

product_id	product_category_name	product_name_lenght	product_description_lenght	product_photos_qty	product_weight_g	product_length_cm	product_height_cm	product_width_cm
1e9e8ef04dbcff4541ed26657ea517e5	perfumaria	40.0	287.0	1.0	225.0	16.0	10.0	14.0
3aa071139cb16b67ca9e5dea641aaa2f	artes	44.0	276.0	1.0	1000.0	30.0	18.0	20.0
96bd76ec8810374ed1b65e291975717f	esporte_lazer	46.0	250.0	1.0	154.0	18.0	9.0	15.0

- **Sellers –** dados sobre os vendedores que atenderam aos pedidos feitos na Olist.

seller_id	seller_zip_code_prefix	seller_city	seller_state
3442f8959a84dea7ee197c632cb2df15	13023	campinas	SP
d1b65fc7debc3361ea86b5f14c68d2e2	13844	mogi guacu	SP
ce3ad9de960102d0677a81f5d0bb7b2d	20031	rio de janeiro	RJ

- **Product_category_name_translate –** converte o nome da categoria do produto de inglês para portugês.

product_category_name	product_category_name_english
beleza_saude	health_beauty
informatica_acessorios	computers_accessories
automotivo	auto

Apêndice 2: Regressão Múltipla com variáveis do dataset analisado.

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            :  sp96_ibge
Dependent Variable  :  customer_u  Number of Observations:   96
Mean dependent var  :     182.396  Number of Variables   :    3
S.D. dependent var  :     105.649  Degrees of Freedom    :   93

R-squared           :    0.078327  F-statistic           :    3.95172
Adjusted R-squared  :    0.058506  Prob(F-statistic)     :  0.0225335
Sum squared residual:      987596  Log likelihood        :   -579.675
Sigma-square        :     10619.3  Akaike info criterion :    1165.35
S.E. of regression  :      103.05  Schwarz criterion     :    1173.04
Sigma-square ML     :     10287.5
S.E of regression ML:     101.427

------------------------------------------------------------------------------
       Variable      Coefficient      Std.Error    t-Statistic   Probability
------------------------------------------------------------------------------
          CONSTANT       331.616        96.3309        3.44247     0.00087
        payment__1      0.802049       0.285385        2.81041     0.00603
        payment_to      -2.23199         1.0237       -2.18032     0.03176
------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   25.987111
TEST ON NORMALITY OF ERRORS
TEST                   DF           VALUE             PROB
Jarque-Bera             2            35.1451          0.00000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                   DF           VALUE             PROB
Breusch-Pagan test      2            11.1255          0.00384
Koenker-Bassett test    2             5.8448          0.05380
================================ END OF REPORT =================================
```

Apêndice 3: Regressão Múltipla com variáveis acrescentadas do censo IBGE 2010.

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            :  sp96_ibge
Dependent Variable  :  customer_u  Number of Observations:   96
Mean dependent var  :     182.396  Number of Variables   :    5
S.D. dependent var  :     105.649  Degrees of Freedom    :   91

R-squared           :    0.317105  F-statistic           :     10.5641
Adjusted R-squared  :    0.287088  Prob(F-statistic)     :4.48145e-007
Sum squared residual:      731739  Log likelihood        :    -565.282
Sigma-square        :     8041.09  Akaike info criterion :     1140.56
S.E. of regression  :     89.6721  Schwarz criterion     :     1153.39
Sigma-square ML     :     7622.28
S.E of regression ML:     87.3057

------------------------------------------------------------------------------
       Variable      Coefficient      Std.Error    t-Statistic   Probability
------------------------------------------------------------------------------
          CONSTANT       410.813        88.8319        4.62461     0.00001
        payment_to      -1.60183       0.941626       -1.70114     0.09233
        payment__1      0.591997       0.267267          2.215     0.02926
              V005      0.013789     0.00421682           3.27     0.00152
              V003      -50.8925        15.3385       -3.31796     0.00130
------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   32.228844
TEST ON NORMALITY OF ERRORS
TEST                     DF           VALUE             PROB
Jarque-Bera               2            21.3441          0.00002

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                     DF           VALUE             PROB
Breusch-Pagan test        4             7.1748          0.12693
Koenker-Bassett test      4             4.4455          0.34906
================================ END OF REPORT =================================
```

Apêndice 4: Regressão Múltipla com Lag Espacial e Peso Queen de Ordem 1

```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : sp96_ibge
Spatial Weight      : w1_sp96_ibge
Dependent Variable  :  customer_u  Number of Observations:   96
Mean dependent var  :     182.396  Number of Variables   :    6
S.D. dependent var  :     105.649  Degrees of Freedom    :   90
Lag coeff.   (Rho)  :    0.487174

R-squared           :    0.451976  Log likelihood        :   -557.467
Sq. Correlation     : -            Akaike info criterion :    1126.93
Sigma-square        :     6116.89  Schwarz criterion     :    1142.32
S.E of regression   :     78.2106

------------------------------------------------------------------------------
       Variable       Coefficient     Std.Error       z-value    Probability
------------------------------------------------------------------------------
      W_customer_u      0.487174      0.107414         4.5355     0.00001
          CONSTANT       317.257       81.2229        3.90601     0.00009
        payment__1      0.545716      0.233139        2.34073     0.01925
        payment_to      -1.56337      0.821276       -1.90358     0.05696
              V003      -42.8997       13.6019       -3.15395     0.00161
              V005     0.0051727     0.0037691         1.3724     0.16994
------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                    DF      VALUE         PROB
Breusch-Pagan test                      4         3.8086      0.43254

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : w1_sp96_ibge
TEST                                    DF      VALUE         PROB
Likelihood Ratio Test                   1        15.6295      0.00008
================================ END OF REPORT =================================
```

Apêndice 5: Regressão Múltipla com Lag Espacial e Peso Queen de Ordem 2

```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : sp96_ibge
Spatial Weight      : w2_sp96_ibge
Dependent Variable  :  customer_u  Number of Observations:   96
Mean dependent var  :     182.396  Number of Variables   :    6
S.D. dependent var  :     105.649  Degrees of Freedom    :   90
Lag coeff.   (Rho)  :    0.528701

R-squared           :    0.424017  Log likelihood        :   -558.784
Sq. Correlation     : -            Akaike info criterion :    1129.57
Sigma-square        :     6428.96  Schwarz criterion     :    1144.95
S.E of regression   :     80.1808

------------------------------------------------------------------------------
       Variable       Coefficient     Std.Error       z-value    Probability
------------------------------------------------------------------------------
      W_customer_u      0.528701       0.13757        3.84313     0.00012
          CONSTANT       314.279         84.38        3.72457     0.00020
        payment__1      0.677501      0.239072        2.83388     0.00460
        payment_to      -1.86112      0.843007       -2.20771     0.02726
              V003      -44.0636       14.0457       -3.13715     0.00171
              V005     0.0102893     0.0038393           2.68     0.00736
------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                    DF      VALUE         PROB
Breusch-Pagan test                      4         6.3043      0.17754

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : w2_sp96_ibge
TEST                                    DF      VALUE         PROB
Likelihood Ratio Test                   1        12.9951      0.00031
================================ END OF REPORT =================================
```

Apêndice 6: Regressão Múltipla com Lag Espacial e Peso Rook de Ordem 1

```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : sp96_ibge
Spatial Weight      : rl_sp96_ibge
Dependent Variable  :  customer_u  Number of Observations:   96
Mean dependent var  :     182.396  Number of Variables   :    6
S.D. dependent var  :     105.649  Degrees of Freedom    :   90
Lag coeff.   (Rho)  :    0.509466

R-squared           :    0.467898  Log likelihood        :   -556.469
Sq. Correlation     : -            Akaike info criterion :    1124.94
Sigma-square        :     5939.18  Schwarz criterion     :    1140.32
S.E of regression   :     77.0661

------------------------------------------------------------------------------
       Variable       Coefficient     Std.Error       z-value    Probability
------------------------------------------------------------------------------
     W_customer_u      0.509466       0.103296        4.93209     0.00000
         CONSTANT        298.33        80.0382        3.72735     0.00019
        payment__1      0.521348       0.229747        2.26923     0.02325
        payment_to      -1.39098       0.809262       -1.71882     0.08565
             V003      -43.1405        13.3947       -3.22072     0.00128
             V005     0.0044882     0.00371413        1.20841     0.22689
------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                    DF      VALUE        PROB
Breusch-Pagan test                      4         3.4438     0.48648

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : rl_sp96_ibge
TEST                                    DF      VALUE        PROB
Likelihood Ratio Test                   1        17.6257     0.00003
================================ END OF REPORT =================================
```

Apêndice 7: Regressão Múltipla com Lag Espacial e Peso Rook de Ordem 2

```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : sp96_ibge
Spatial Weight      : r2_sp96_ibge
Dependent Variable  :  customer_u  Number of Observations:   96
Mean dependent var  :     182.396  Number of Variables   :    6
S.D. dependent var  :     105.649  Degrees of Freedom    :   90
Lag coeff.   (Rho)  :    0.440386

R-squared           :    0.394628  Log likelihood        :   -560.689
Sq. Correlation     : -            Akaike info criterion :    1133.38
Sigma-square        :        6757  Schwarz criterion     :    1148.76
S.E of regression   :      82.201

------------------------------------------------------------------------------
       Variable       Coefficient     Std.Error       z-value    Probability
------------------------------------------------------------------------------
      W_customer_u      0.440386         0.147        2.99583     0.00274
          CONSTANT       331.395        86.547        3.82908     0.00013
         payment__1      0.664629       0.24522        2.71034     0.00672
         payment_to      -1.81554      0.864884       -2.09917     0.03580
               V003      -45.4934       14.3848       -3.1626      0.00156
               V005     0.0105719    0.00394437        2.68025     0.00736
------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                    DF      VALUE        PROB
Breusch-Pagan test                      4         6.0800     0.19326

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : r2_sp96_ibge
TEST                                    DF      VALUE        PROB
Likelihood Ratio Test                   1         9.1858     0.00244
================================ END OF REPORT =====================================
```