Relatório Desafio Análise de Dados

Marina Rocha Guimarães

January 9, 2020

0.1 Introdução

O desafio proposto apresenta três arquivos para serem analisados: um sobre as vendas em lojas físicas, outro sobre as vendas feitas através do site e outro sobre as visitas na página da loja. Com isso, abriu-se cada um dos arquivos e algumas analises foram feitas. Vale ressaltar que analisou-se o tamanho de cada arquivo antes de abri-lo e percebeu-se que o arquivo referente às visitas na página da loja era muito grande, então algumas estratégias foram utilizadas - as quais serão explicadas posteriormente.

Este relatório irá apresentar primeiramente todas as análises realizadas e, após isso, serão respondidas as perguntas propostas 1,2 e 3 com base nas análises apresentadas. As questões 4 e 5 não foram inteiramente respondidas.

0.2 Análises

A fim de analisar de uma forma organizada os dados existentes, a análise foi dividida em três partes:

- Vendas em lojas físicas
- Encomendas online
- Visualizações da página online

Abaixo cada um desses itens citados será explicado. Vale ressaltar que, para fins de organização, as variáveis presentes no código da análise de cada um dos itens citados acima possuem a terminação *ofs*, *ono* e *onp*, respectivamente.

0.2.1 Vendas em Lojas Físicas

Primeiramente analisou-se o número de linhas presentes no arquivo em questão antes de abri-lo completamente, e descobriu-se que o arquivo apresenta 29372 linhas. Decidiu-se então abri-lo direto como um *data frame* utilizando a biblioteca *pandas* - já que o número de linhas não é muito grande.

A fim de facilitar algumas análises posteriores, criou-se uma coluna no *data frame* chamada *price_quantity*, a qual apresenta o resultado da multiplicação do valor de cada produto com a quantidade comprada. Pôde-se então calcular o faturamento total existente no tempo analisado: somou-se os valores presentes na nova coluna criada. Fazendo isso, obteve-se um valor de faturamento igual a R$ 13903005,32.

Após isso, decidiu-se analisar o faturamento por estado, a fim de descobrir qual deles mais contribuiu no faturamento total. O resultado da análise é mostrado na Figura 1.

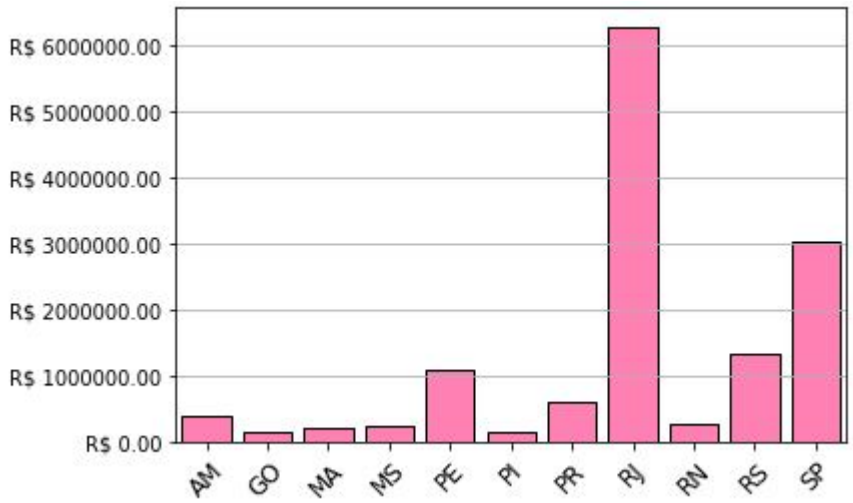

Figure 1: Faturamento de lojas físicas por estado.

Ao analisar a Figura 1, nota-se que o estado do Rio de Janeiro é o que mais contribuiu com o faturamento das lojas físicas - sendo sua contribuição muito maior do que a maioria dos outros estados. Além disso, nota-se que São Paulo é o segundo estado que mais contribuiu com o faturamento. Ao concluir isso, surge a curiosidade sobre quantas lojas cada estado possui e qual o faturamento médio das lojas por estado.

Para isso, foi realizada uma análise sobre a coluna nomeada como *store_id*. Percebeu-se que quando a venda era feita no mesmo estado, o *store_id* era igual em alguns casos e em outros não. Calculou-se então o número de estados existentes na tabela (igual a 11) e o número de *store_id* existentes (igual a 39), ou seja, cada estado apresenta um ou mais *store_id* associado a ele. Com o intuito de descobrir se um mesmo número de *sore_id* pertence a mais de um estado, analisou-se quantos *store_id* tinham por estado e somou-se esses valores, obtendo-se o número 39 - exatamente igual ao número de *store_id* existentes na tabela. Logo, concluiu-se que o *store_id* representa a identificação de cada loja, sendo que um mesmo *store_id* não é encontrado em diferentes estados.

A partir disso, analisou-se o número de lojas existentes em cada estado. O resultado pode ser visto na Figura 2 abaixo.

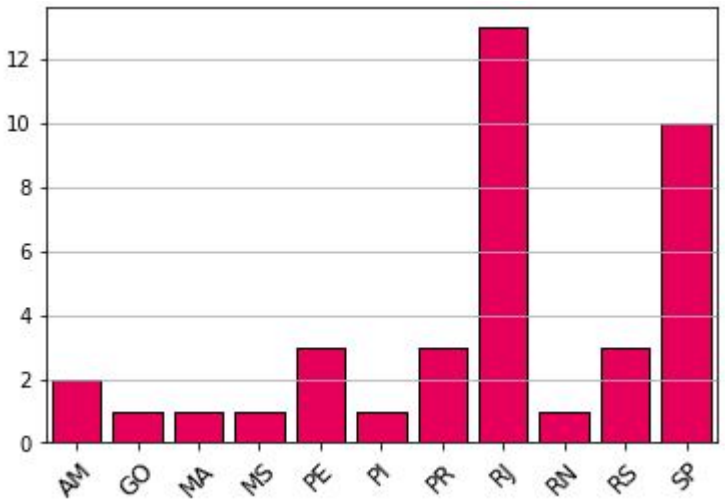

Figure 2: Número de lojas existentes por estado.

Percebe-se que o número de lojas no Rio de Janeiro e em São Paulo é muito maior do que nos outros estados. Logo, realizou-se outra análise referente ao faturamento por estado considerando agora o número de lojas, ou seja, dividiu-se os valores obtidos na Figura 1 pelo número de lojas existentes em cada estado a fim de se obter uma média do faturamento das lojas (por estado). Esse resultado é mostrado na Figura 3.

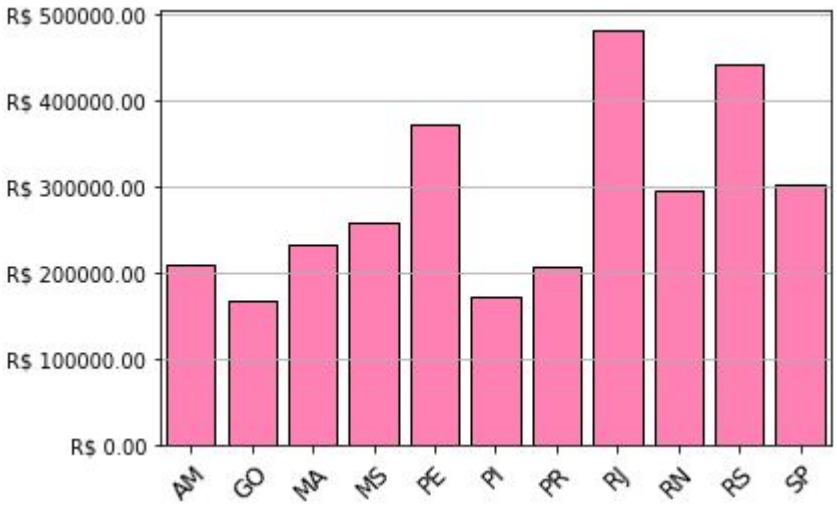

Figure 3: Média do faturamento das lojas por estado.

Ao observar a Figura 3 percebe-se que a média do faturamento das lojas

existentes no Rio de Janeiro é realmente maior do que a média do faturamento das lojas presentes nos outros estados. Um ponto importante a analisar é que no estado do Rio Grande do Sul a média do faturamento das lojas é maior do que em São Paulo, mas para a contribuição do faturamento total de todos os estados, São Paulo contribuiu mais do que o Rio Grande do Sul. Isso ocorre pois o número de lojas existentes em SP é maior que duas vezes o número de lojas existentes no RS.

Uma outra análise feita foi a relação da quantidade de compras realizadas por cariocas em relação aos dias da semana. O resultado dessa análise pode ser visto na Figura 4 abaixo.

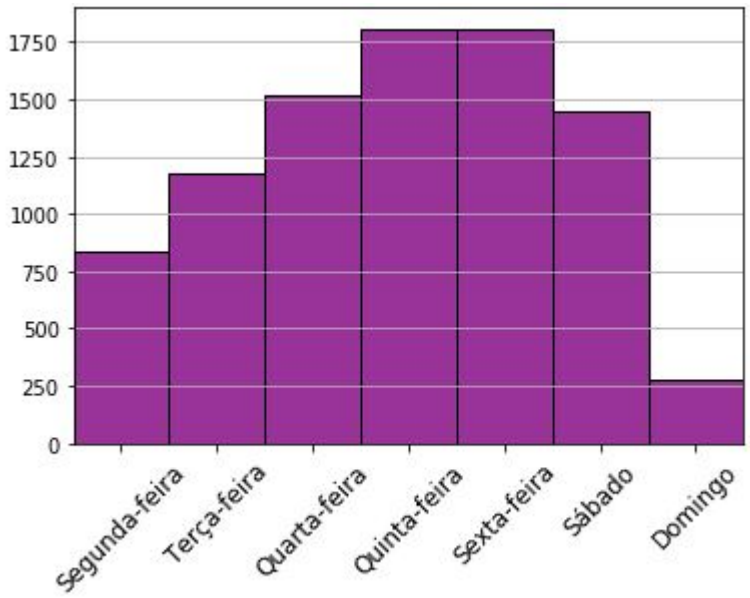

Figure 4: Número de compras realizadas por cariocas em lojas físicas por dias da semana.

Ao analisar a Figura 4 nota-se que a maior quantidade de compras ocorre na quinta-feira e na sexta-feira. Surge então a curiosidade se esse resultado é realmente válido, ou seja, se não houve algum dia do mês em que a quantidade de vendas foi absurda e fez com que o resultado apresentado acima na Figura 4 seja enganoso. Para verificar isso, analisou-se então cada dia da semana do período proposto (o mês de agosto inteiro). O resultado dessa analise é mostrado na Figura 5 abaixo.

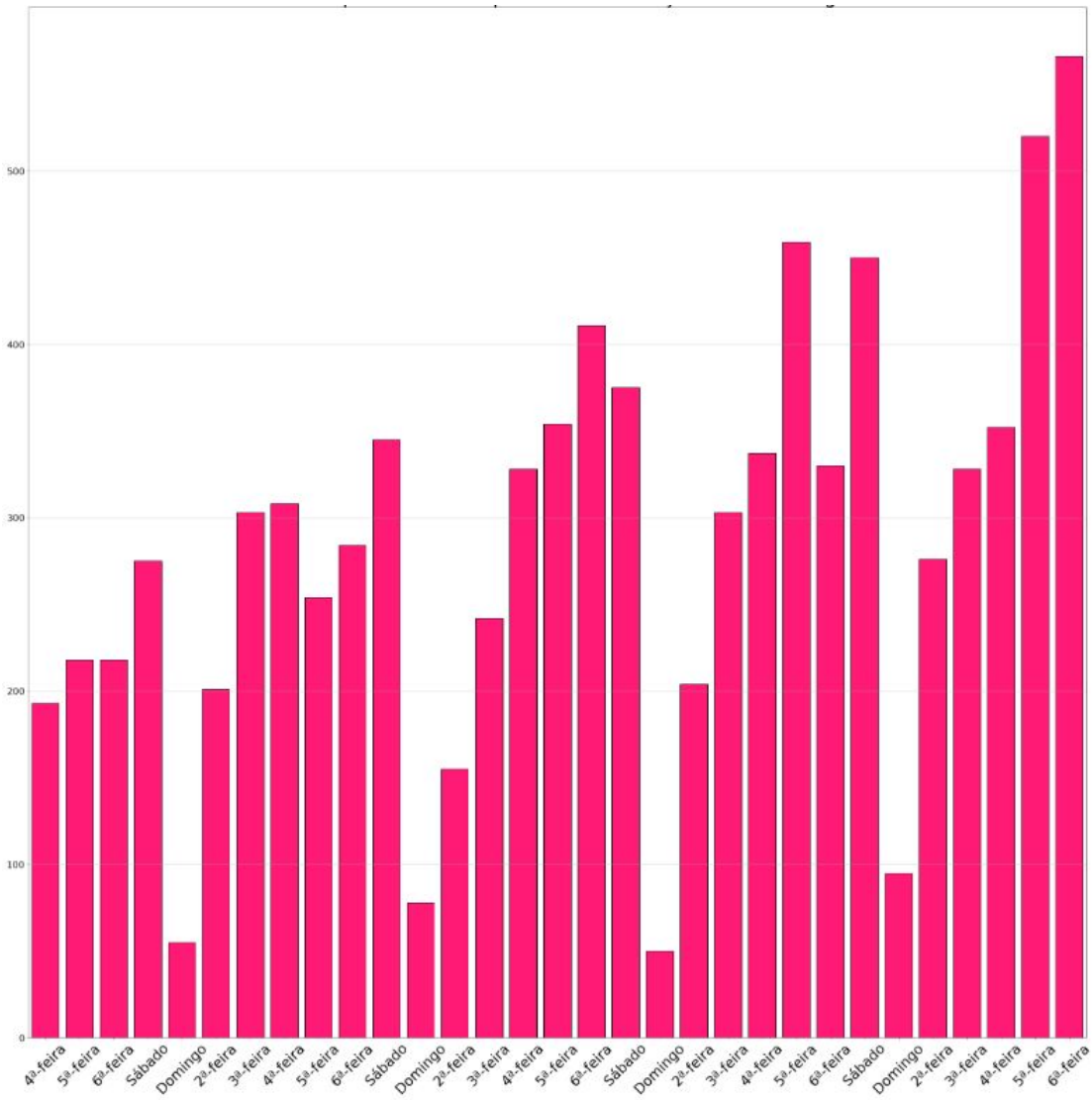

Figure 5: Número de compras realizadas por cariocas em lojas físicas por cada dia da semana do mês.

Percebe-se que as últimas quinta e sexta-feira do mês obtiveram mais vendas do que ao longo do mês, porém não foi um resultado absurdamente diferente - o que não torna inválido o resultado obtido anteriormente. Por outro lado, percebe-se que há cinco quintas, quartas e sextas-feiras mas apenas quatro dos outros dias da semana, o que influencia o resultado mostrado na Figura 4. Logo, o resultado apresentado na Figura 4 tem sentido porém é um pouco enganoso.

Ao perceber isso, dividiu-se o número de compras realizadas por dia de

semana pelo número de vezes que o dia da semana ocorreu no mês de agosto, a fim de se obter um resultado ponderado, o qual pode ser visto na Figura 6 abaixo.

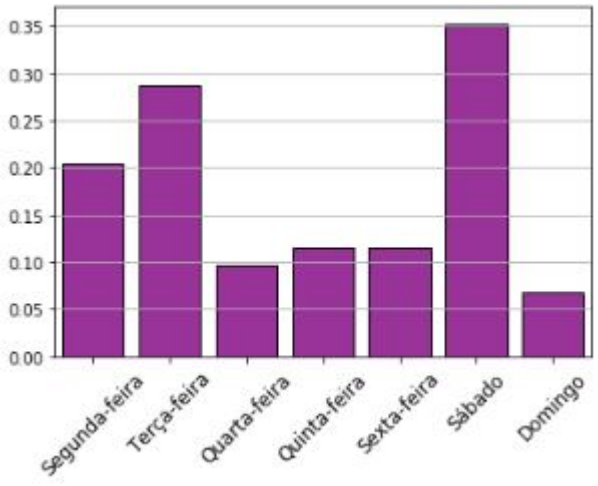

Figure 6: Número de compras realizadas por cariocas em lojas físicas por dias da semana, ponderando a ocorrência de cada dia da semana no mês.

Agora, ao analisar a Figura 6, percebe-se que o dia da semana que apresenta em média o maior número de compras é sábado, seguido de terça-feira e segunda-feira - resultado diferente do encontrado anteriormente quando a ponderação não havia sido considerada.

Ao analisar agora a relação de compras nos dias de semana e no final de semana, é necessário realizar mais uma ponderação: são cinco dias de semana e apenas dois dias no final de semana. Logo, para essa análise somou-se o valor encontrado na Figura 6 dos dias de semana e dividiu-se por cinco, assim como somou-se os valores encontrados no sábado e domingo e dividiu-se por dois. O resultado final é o seguinte: O número ponderado de compras realizadas por cariocas em lojas físicas nos finais de semana é igual a 0.2103271484375, ao passo que esse mesmo número, porém em dias de semana, é igual a 0.1639317875. Nota-se que os cariocas gostam mais de comprar nos finais de semana do que em dias de semana - principalmente no sábado.

Uma última análise realizada sobre as lojas físicas foi a análise do valor médio gasto por compra em cada região, a qual pode ser visto na Figura 7

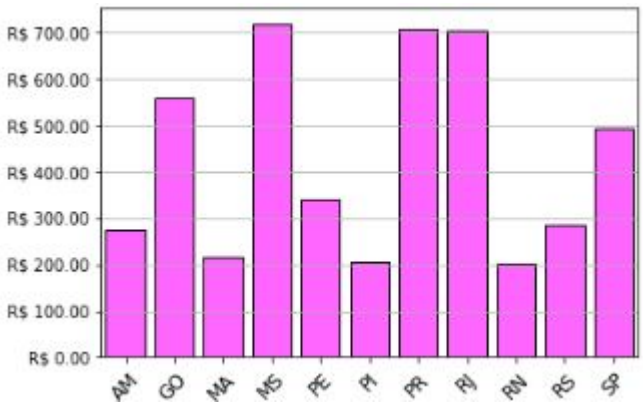

Figure 7: Valor médio gasto em uma compra por estado.

Nota-se que os estados Mato Grosso do Sul, Paraná e Rio de Janeiro apresentam os valores mais altos. Além disso, é interessante notar que o valor médio gasto por compra no estado do Rio Grande do Sul é menor do que em São Paulo, mesmo o Rio Grande do Sul apresentando um valor médio de faturamento das lojas maior que São Paulo (como foi visto na Figura 3). Isso pode significar que o número de compras por loja no RS é maior que o número de compras por loja em SP, porém as compras apresentam valores mais baixos (em média) no estado do Rio Grande do Sul.

0.2.2 Encomendas Online

Para as encomendas online, novamente foi analisado o número de linhas presentes no arquivo antes de abri-lo completamente, e descobriu-se que o arquivo apresenta 12237 linhas. Adicionou-se também a coluna *price_quantity* (já explicada anteriormente) e, a partir dela, calculou-se o faturamento total existente no tempo analisado: R$ 5649323.00.

Após isso, analisou-se a quantidade de produtos existentes nessa tabela referente às encomendas online, obtendo-se um valor igual a 2184 produtos. Analisou-se também qual foi o produto mais encomendado e obteve-se como resposta o produto que apresenta o *product_id* igual a 626664333563363 - foram vendidas 74 unidades dele. Analisou-se ainda os 5 produtos menos encomendados - a fim de propor algum possível desconto para que a venda deles aumente - que são os produtos com *product_id* iguais a: [**'656364326237623', '356164376139646', '656336386334386', '633835653265316', '356437616566363'**], sendo o último valor apresentado o *product_id* do produto menos encomendado.

A próxima análise sobre as encomendas online foi a relação da quantidade de encomendas realizadas por cariocas em relação aos dias da semana. O resultado dessa análise pode ser visto na Figura 8 abaixo.

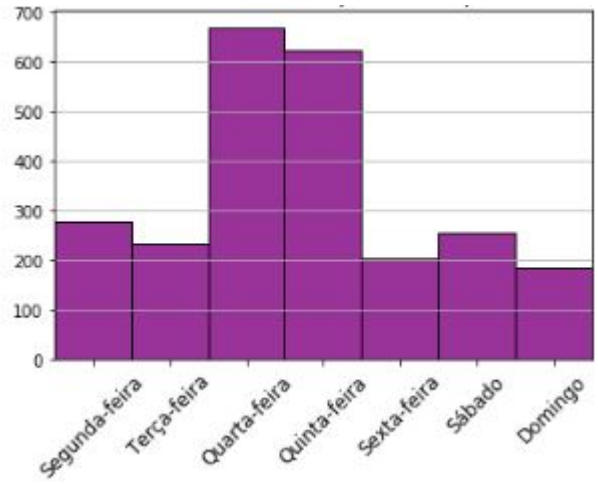

Figure 8: Número de encomendas realizadas por cariocas por dias da semana.

Observando a Figura 8 nota-se que a maior quantidade de compras ocorre na quarta-feira e na quinta-feira. Novamente questiona-se se esse resultado é válido. Mais uma vez então analisou-se cada dia da semana do período proposto. O resultado é mostrado na Figura 9 abaixo.

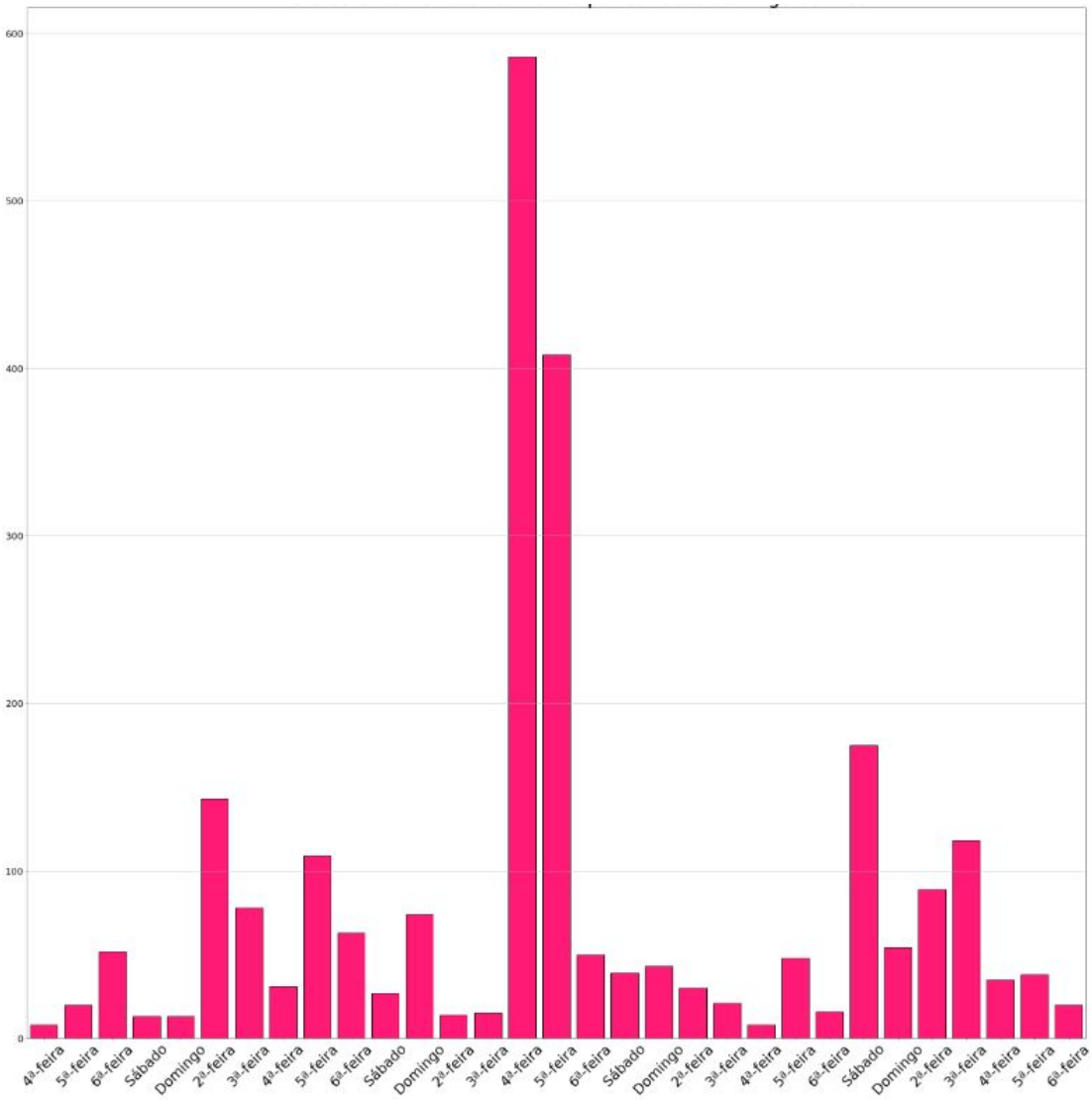

Figure 9: Número de encomendas realizadas por cariocas por cada dia da semana do mês.

Novamente nota-se que o número de ocorrências no mês dos dias da semana é diferente. Vale ressaltar que há uma diferença comparada com a analise feita nas vendas em lojas físicas: nas encomendas online percebe-se que na quarta-feira e na quinta-feira do mês de agosto correspondentes aos dias 15 e 16 o número de encomendas foi imensamente maior que em todos os outros dias, o que pode significar uma promoção oferecida pelo site - já que ao olhar o calendário do Rio de Janeiro de 2018 não há nenhum evento especial (aparentemente) nesses dias.

Logo, para realizar uma análise mais confiável que a análise feita na Figura 8, ponderou-se cada valor pela aparição do dia da semana em questão no mês (como já feito anteriormente) e excluiu-se o valor dos dias 15 e 16. Vale notar que agora o único dia da semana que aparece cinco vezes no mês de agosto é a sexta-feira. O resultado dessa análise é mostrado na Figura 10.

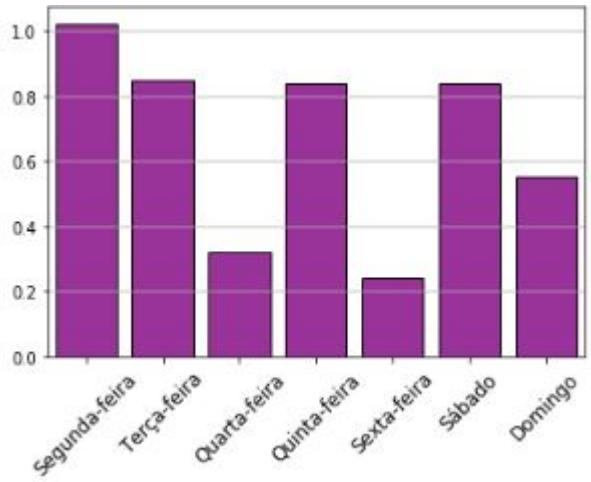

Figure 10: Número de encomendas realizadas por cariocas por dias da semana, ponderando a ocorrência de cada dia da semana no mês e excluindo os valores dos dias 15 e 16 de agosto.

Agora, ao analisar a Figura 10, percebe-se que o dia da semana que apresenta em média o maior número de compras é segunda-feira - resultado diferente do encontrado anteriormente quando a ponderação não havia sido considerada.

Analisando a relação de compras nos dias de semana e no final de semana (considerando novamente a ponderação sobre ser dia de semana ou final de semana feita anteriormente nas vendas em lojas físicas) obteve-se o seguinte: o número ponderado de encomendas realizadas por cariocas nos finais de semana é igual a 0.6953125 e o número ponderado de encomendas realizadas por cariocas nos dias de semana é igual a: 0.65457. Nota-se que os cariocas gostam mais de comprar online nos finais de semana do que em dias de semana.

Com uma outra análise realizada, nota-se que o gasto médio do consumidor em cada encomenda é de aproximadamente R$ 461.66. Ao analisar o gasto médio por dia de semana, tem-se o seguinte resultado mostrado na Figura 11.

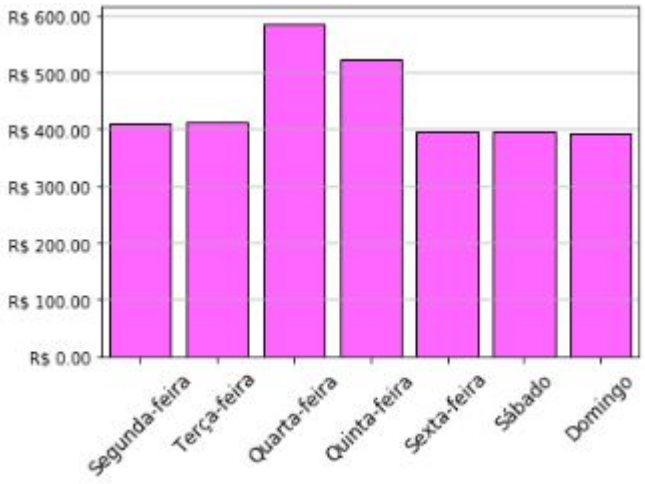

Figure 11: Valor médio gasto por encomenda por dia da semana.

Mais uma vez observa-se que esse resultado pode ser enganoso por motivos já vistos anteriormente. Então, para solucionar isso, excluiu-se novamente os dados dos dias 15 e 16 de agosto - dias com valores fora do comum. O resultado é mostrado na Figura 12.

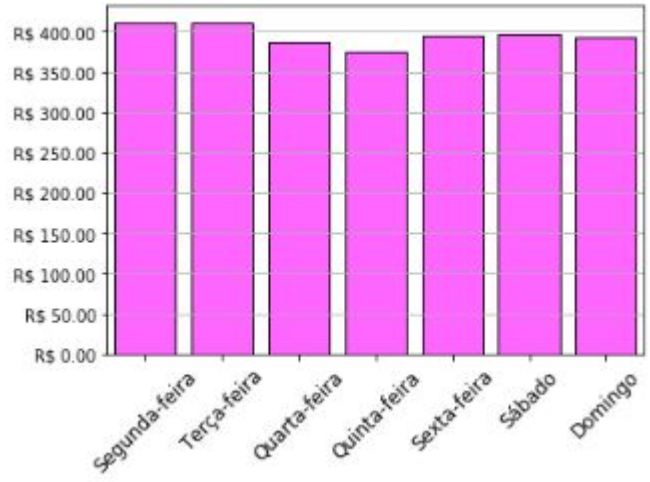

Figure 12: Valor médio gasto por encomenda por dia da semana.

Ao analisar a Figura 12 nora-se que o valor médio gasto por encomenda por dia de semana é parecido, sendo maior nas segundas e terças-feiras e menor na quinta-feira.

0.2.3 Visualizações da Página Online

Mais uma vez foi analisado o número de linhas presentes no arquivo, e obteve-se o número 3452540. Como esse número é muito grande, algumas tentativas para resolver esse problema foram feitas. Primeiramente, utilizou-se o site presente **nesse link** a fim de analisar quais informações estavam presentes no arquivo.

Tentou-se utilizar diversas ferramentas para abrir o arquivo por partes e guardar ele em diferentes *data frames* mas mesmo assim o problema de memória persistia. Então, a solução encontrada foi filtrar as informações necessárias no momento em que cada linha do arquivo era lida. Ou seja, para responder a questão 4, apenas as linhas em que o *pageType* é igual a *product* interessam e para responder a questão 5 apenas as linhas em que *pageType* são iguais a *cart* interessam. Fazendo isso, conseguiu-se salvar apenas as informações que seriam utilizadas.

No entanto, mesmo com essa solução as informações referentes ao *pageType* igual a *product* ainda foram divididas em 12 diferentes *data frames*, a fim de não ocasionar novamente um erro de memória.

A partir disso, as análises poderiam ser realizadas.

0.3 Questões propostas

0.3.1 Qual foi o faturamento total no período?

Somando-se os valores já obtidos anteriormente do faturamento das lojas físicas e do faturamento das encomendas online, tem-se o valor de faturamento total igual a R$ 19552328.32 no período analisado.

0.3.2 Qual o produto mais comprado online?

A partir do resultado obtido anteriormente, sabe-se que o produto mais comprado online foi o que possui o *product_id* igual a 62666433356336. Vale notar que 74 unidades desse produto foram vendidas.

0.3.3 Cariocas gostam de comprar no fim de semana?

A partir das análises realizadas anteriormente, percebe-se que os cariocas no geral realizam um maior número de compras nos finais de semana (ponderando esse resultado como explicado anteriormente) do que em dias de semana. Nota-se que, com relação às lojas físicas, o número de compras é muito maior no sábado do que no domingo - resultado esperado, já que muitas lojas não abrem nos domingos.

0.3.4 É comum escolher online e terminar a compra na loja física?

Não foi realizada a análise dessa parte do problema, porém a solução poderia ser feita analisando os *data frames* salvos a partir de *pageType* igual a *product* e comparando a coluna de *customer_id* com essa mesma coluna da tabela das compras em lojas físicas. O problema nessa análise é que talvez uma pessoa com um certo *customer_id* tenha olhado um produto no site, ido até uma loja física e comprado outro produto - o que não pode ser verificado pois o valor de *on_product_id* não é igual ao valor de *off_product_id*.

0.3.5 Estime o resultado da campanha proposta.

Essa analise também não foi realizada, mas para fazer isso teria que analisar a coluna de *customer_id* presente no *data frame* criado a partir de *pageType* igual a *cart* e comparar quais desses não estão presentes na tabela de vendas em lojas físicas. Porém, novamente, como não pode ser verificado o produto que o consumidor comprou a análise fica comprometida.

Para formular uma proposta a essa questão (após ser feita a análise), poderia ser observado qual estado possui uma menor média de faturamento por loja (mostrado na Figura 3) para aplicar o cupom. Além disso, poderia ser analisado pelo dia da semana em que o cupom poderia ser utilizado - o dia da semana com menor número de compras (essa análise de dia da semana também foi realizada).