FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Sistemas de Gestão de Dados

2020/2021 - 2º Semestre

SSB Perf Bench and Processing Design

Grupo 14
Francisco Miranda - 2015250592
Maria Paula Viegas - 2017125592

Index

Avaliação do Grupo

- O grupo foi capaz de avaliar os dataset fornecido a fim de:
    - descobrir padrões, tais como variações em vendas e outros
    - produzir estatísticas e charts
    - testar hipóteses e intervalos de confiança
- Autoavaliação do grupo
    - 16 Valores
- Contribuição individual
    - Francisco: Hipóteses, intervalos de confiança e relatório
    - Maria Paula: Produção de plots e análise
- Autoavaliação individual
    - Francisco: 16 Valores
    - Maria Paula: 16 Valores
- Tempo de esforço
    - Francisco: 8 hrs
    - Maria Paula: 10 hrs

1. Introdução

Foi nos proposto que exploremos e analisemos um dataset relativo a dados de vendas de produtos ao longo de 2 anos. Foram-nos fornecidos pelo docente múltiplos datasets correspondendo a segmentos do SSB, o nosso grupo utilizou o dataset número 7. Para cada encomenda sabemos qual o produto comprado, quantas unidades foram compradas, qual o custo de fornecimento e o custo para o consumidor final, bem como quaisquer descontos e taxas aplicadas. Além disso, é sabido o modo de transporte e a prioridade da entrega. Para cada produto é conhecido o seu nome, marca e categoria. Para cada cliente ou fornecedor sabe-se a cidade, nação e região da sua localização.

Além da informação à partida disponível calculamos o valor da receita com base nos preços de venda, na percentagem de desconto e nas taxas.

Uma primeira abordagem revelou que as métricas de média, quartis, mediana, valores mínimo e máximo bem como desvio padrão das diversas colunas do dataset era exatamente igual para os dois anos considerados. Outras observações gráficas vieram a confirmar a hipótese de que os dados para os 2 anos eram exatamente iguais e como tal optámos por não fazer uma exploração centrada na comparação entre os 2 anos.

A nossa abordagem começou por num primeiro momento criar gráficos para visualizar os dados em diversas condições. Depois identificámos padrões de consumo e calculamos o grau de confiança das nossas observações. Por fim concluímos intervalos de confiança para o valor de vendas mensal de produtos por categoria e por marca.

2. Análise de Dados e Formulação de Hipóteses

Num primeiro momento observámos a distribuição de vendas por diversas regiões e nações como visível nas figuras 1 e 2. Observámos que a distribuição das vendas é aproximadamente uniforme pelo mundo todo, tendo África as maiores receitas e a América as menores, correspondendo a 21.2% e 16.8% respectivamente. A distribuição de receitas pelas nações é também aproximadamente uniforme, sendo o único desnível observável na região Americana, onde o Brasil é responsável por 33.2% das receitas em comparação com os apenas 19.1% da Argentina.

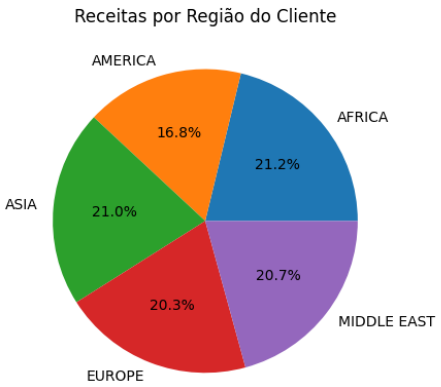

Figura 1: Distribuição das receitas pelas diversas regiões dos clientes.

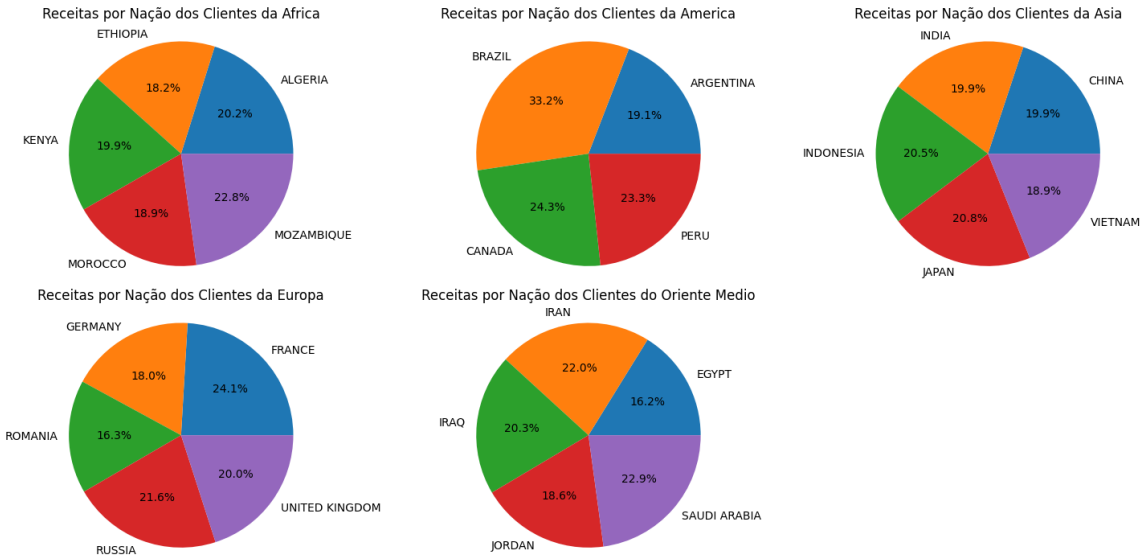

Figura 2: Distribuição das receitas pelas diversas nações dos clientes.

De seguida explorámos a distribuição das vendas ao longo dos diversos meses e dias da semana em cada região de clientes (figuras 3 e 4). Aqui identificamos que Agosto e Novembro são consistentemente os meses com mais vendas, sendo Agosto particularmente superior na Europa e Novembro particularmente superior no Médio Oriente. Os gráficos sugerem ainda que existe uma quebra nas vendas aos domingos.

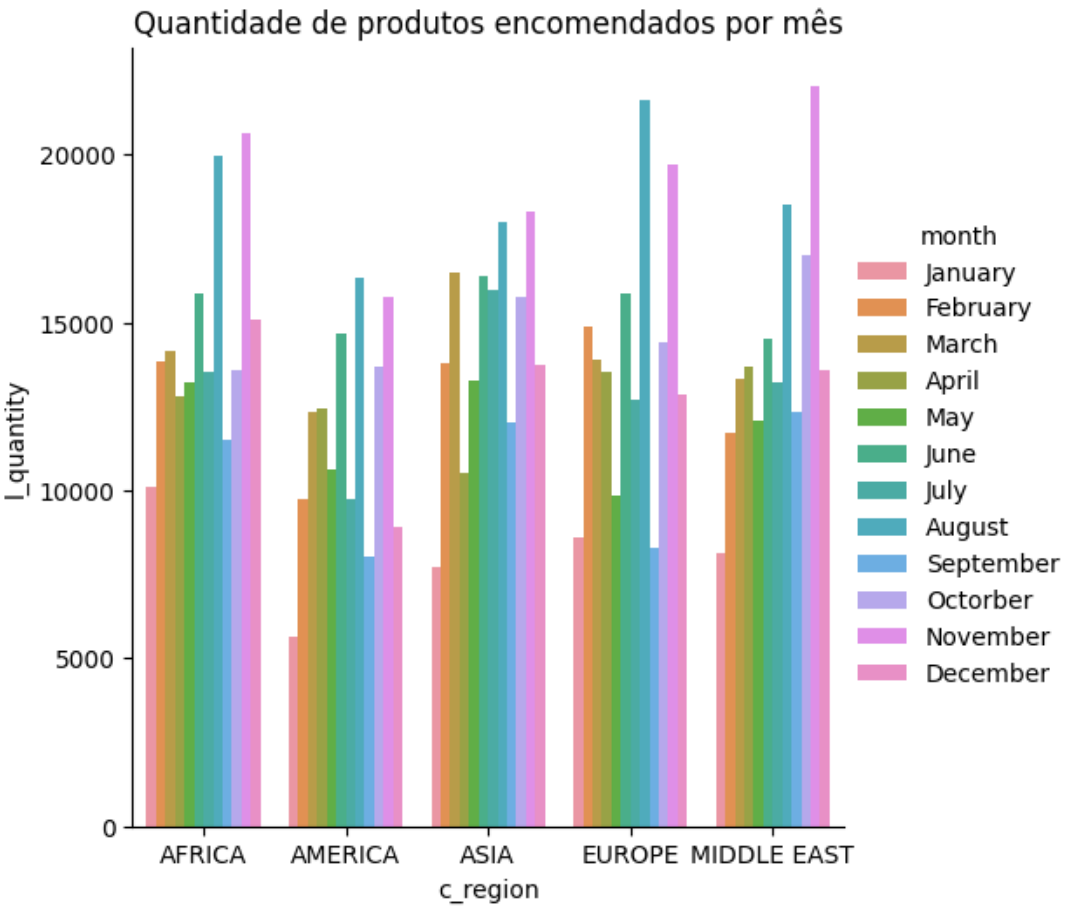

Figura 3: Número de produtos encomendados por mês nas diversas regiões.

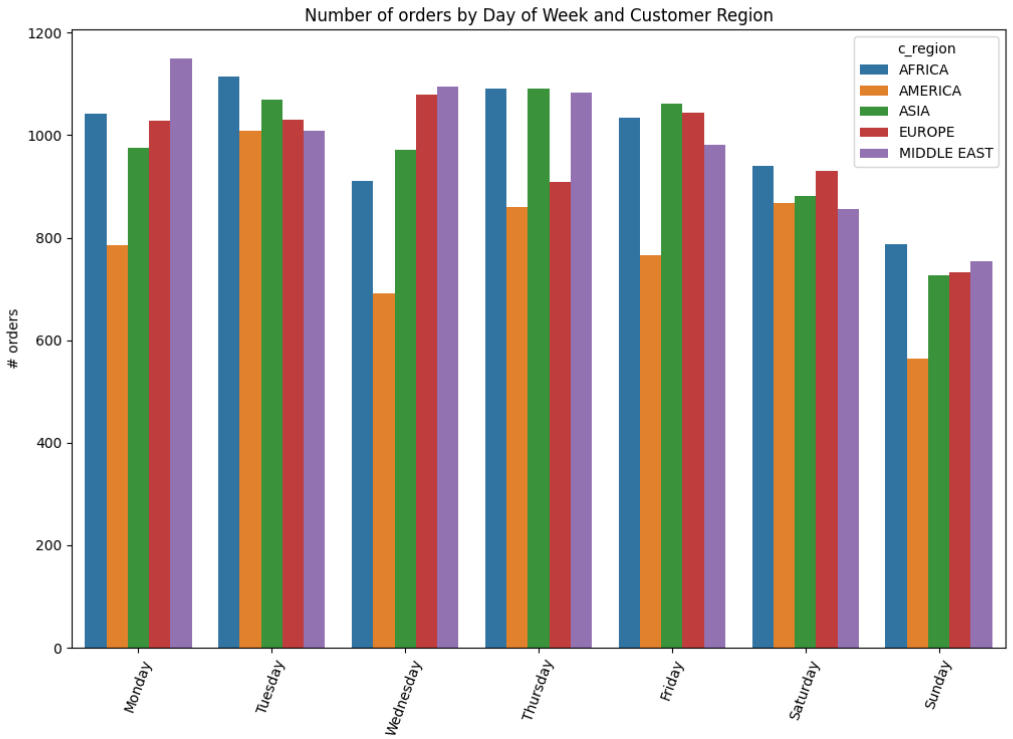

Figura 4: Total de receitas obtidas em cada dia da semana para cada região.

Explorando a relação entre a região dos clientes e o meio de transporte de encomendas (figura 5) concluímos que o meio de transporte mais comum é o transporte marítimo e o menos comum o transporte aéreo registado, sendo que os restantes 5 meios de transporte são igualmente populares.

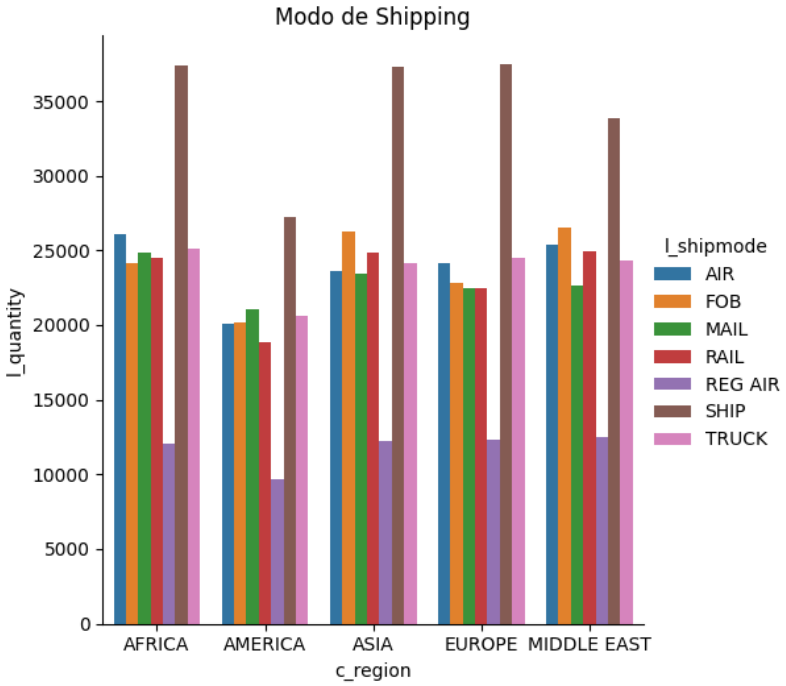

Figura 5: Número de produtos transportados por cada opção de envio por região.

Obtivemos gráficos que mostram a média de receitas obtidas pela venda de produtos ao longo do ano (figuras 6 e 7), mostrando o desvio nas receitas para diferentes marcas e categorias de produtos. Como os desvios - representados no gráfico como a sombra da linha principal - são muito pequenos, concluímos que a competição entres as diferentes marcas é muito renhida, sendo que todas têm picos de receita nas mesmas alturas. Observamos também uma alta instabilidade na receita obtida ao longo dos meses.

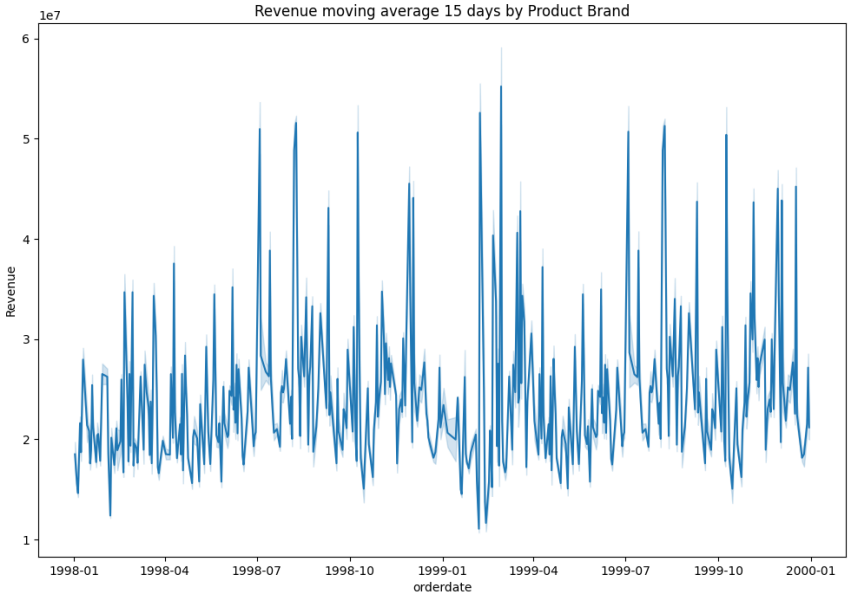

Figura 6: Média móvel de 15 dias de Receitas para cada marca ao longo dos 2 anos.

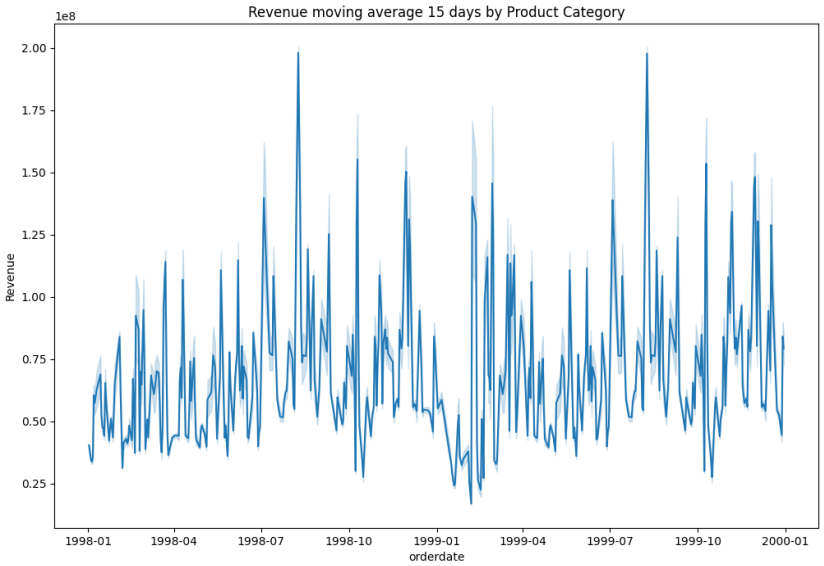

Figura 7: Média móvel de 15 dias de Receitas para cada categoria de produtos ao longo dos 2 anos.

3. Teste de Hipóteses

3.1 Distribuição de receitas uniforme

Na secção 2 observámos que a distribuição das receitas por região aparenta ser uniforme. Aqui decidimos testar a veracidade desta observação recorrendo a um teste de hipótese. Como demonstrado abaixo, esta observação foi confirmada pelo teste estatístico.

Hipótese nula: A região do cliente não afeta a capacidade de venda.

Método: Calculamos as receitas para os clientes de cada cidade. Considerámos que o conjunto das receitas de todas as cidades de uma região é uma amostra, ficando cada região com cerca de 40 observações. Verificamos a normalidade das distribuições com o teste de Shapiro, e a igualdade das variâncias com teste de Bartlett e por fim a igualdade das distribuições de cada amostra com o teste ANOVA independente. Todos os testes foram realizados ao nível de significância $\alpha = 0.01$.

Objetivo	Teste	Estatística	P-valor	Conclusão
Averiguar Normalidade Amostra de África	Shapiro	0.978	0.526	Normal
Averiguar Normalidade Amostra da América	Shapiro	0.964	0.277	Normal
Averiguar Normalidade Amostra da Ásia	Shapiro	0.983	0.728	Normal
Averiguar Normalidade Amostra da Europa	Shapiro	0.957	0.092	Normal
Averiguar Normalidade Amostra do Médio Oriente	Shapiro	0.980	0.613	Normal
Averiguar Igualdade de Variâncias das amostras	Bartlett	10.85	0.028	Variâncias iguais
Testar Hipótese Nula	Anova Independente	0.200	0.938	Aceitar H0

Tabela 1: Testes estatísticos realizados para decidir sobre a hipótese de que a região do cliente não afeta a capacidade de venda.

3.2 Quebra de vendas ao domingo

Os gráficos da distribuição das receitas por dia da semana e por região sugeriram também que haveria uma quebra nas vendas ao domingo. Aqui verificamos esta observação. Como demonstrado abaixo, esta observação foi rejeitada pelo teste estatístico que afirma não haver diferença nas amostras dos diversos dias da semana.

Hipótese nula: Não existe diferença entre as receitas de vendas a cada dia da semana.

Método: Calculamos as receitas obtidas para cada dia da semana em cada semana ao longo dos 2 anos. As amostras para cada dia da semana obtidas têm tamanhos que variam entre 44 e 69. Verificamos a não normalidade das distribuições com o teste de

Shapiro e a igualdade das distribuições de cada amostra com o teste de Kruskal-Wallis. Todos os testes foram realizados ao nível de significância α = 0.01.

Objetivo	Teste	Estatística	P-valor	Conclusão
Averiguar Normalidade Amostra de Sexta-Feira	Shapiro	0.786	9.11e-08	Não Normal
Averiguar Normalidade Amostra de Segunda-Feira	Shapiro	0.897	7.30e-05	Não Normal
Averiguar Normalidade Amostra de Sábado	Shapiro	0.743	4.23e-08	Não Normal
Averiguar Normalidade Amostra de Domingo	Shapiro	0.757	3.90e-07	Não Normal
Averiguar Normalidade Amostra de Quinta-Feira	Shapiro	0.829	5.39e-07	Não Normal
Averiguar Normalidade Amostra de Terça-Feira	Shapiro	0.925	5.84e-04	Não Normal
Averiguar Normalidade Amostra de Quarta-Feira	Shapiro	0.902	5.30e-05	Não Normal
Testar Hipótese Nula	Kruskal-Wallis	9.85	0.131	Aceitar H0

Tabela 2: Testes estatísticos realizados para decidir sobre a hipótese de que não existe diferença entre as receitas de vendas a cada dia da semana.

De forma a averiguar o porquê de o teste estatístico contradizer a observação anterior fizemos um boxplot das diferentes amostras (figura 8) que confirmou uma distribuição idêntica das vendas para os diversos dias da semana. A explicação para a observação errada na secção anterior prende-se com o facto de a amostra para os domingos ser mais pequena que aquelas para todos os outros dias. Ou seja, aos domingos vende-se o mesmo que nos restantes dias, mas é menos provável que se venda algo de todo, provavelmente por motivos de fecho da loja.

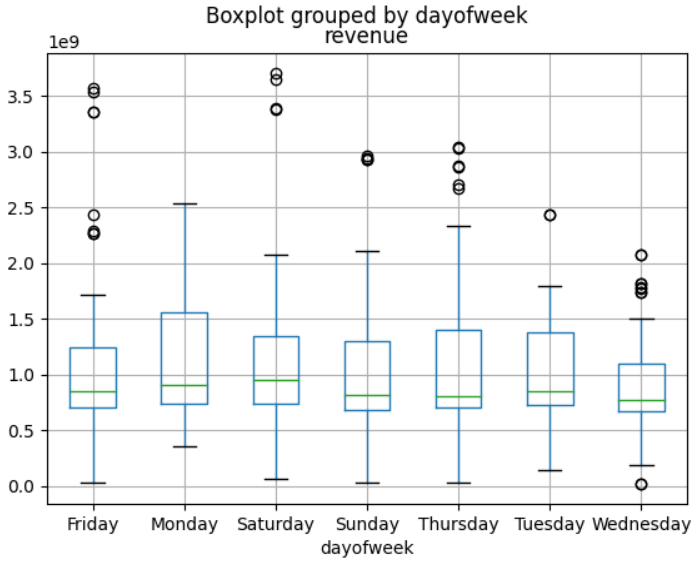

Figura 8: Distribuição das receitas por dia da semana com atividade comercial.

3.3 Caracterizar capacidade de geração de receitas

Dada a elevada instabilidade observada nas vendas ao longo dos diversos meses, achamos importante ser capaz de definir um intervalo de confiança para a capacidade de produção de receitas. Com este objetivo calculamos o valor das receitas por cidade e por mês para um total de 2264 amostras.

Estas amostras de receita são caracterizadas por uma média de *1.96e+08* e desvio padrão *1.64e+08*, têm um valor mínimo de *2.80e+05* e máximo de *1.13e+09* e tem os três quartis situados em *7.00e+07*, *1.59e+08* e *2.80e+08*. Para referência deixamos uma visualização destas amostras por mês na figura 9.

Com estes dados concluímos que as receitas de uma qualquer cidade em um qualquer mês situam-se entre *1.88e+08* e *2.05e+08* com um nível de confiança de *99%*.

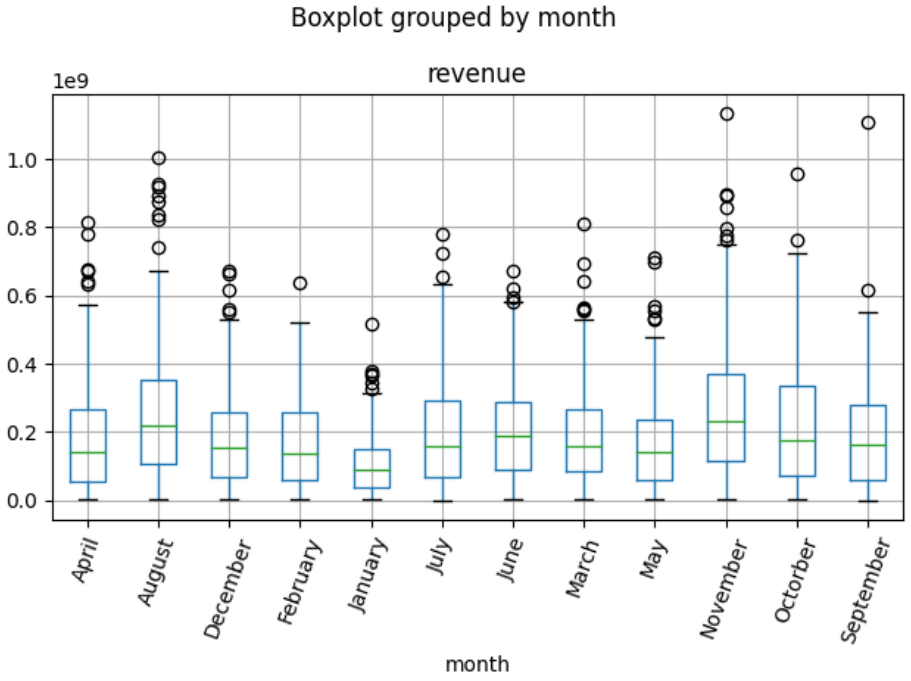

Figura 9: Distribuição das receitas por cidade para cada mês.

3.4 Popularidade de meios de transporte

Como observámos anteriormente o transporte mais popular é o marítimo e o menos popular o correio aéreo registado. Aqui confirmamos essa hipótese com um teste estatístico.

Hipótese nula: Não existe diferença entre a popularidade de cada transporte.

Método: Calculamos o número de encomendas transportadas por cada meio em cada semana ao longo dos 2 anos. Verificamos a não normalidade das distribuições com o teste de Shapiro e a desigualdade das distribuições de cada amostra com o teste de Kruskal-Wallis. Por fim fazemos uma análise post-hoc com testes pareados de Dunn para encontrar as distribuições diferentes. Todos os testes foram realizados ao nível de significância $\alpha = 0.01$, tendo os testes pareados usado correção de Bonferroni.

Objetivo	Teste	Estatística	P-valor	Conclusão
Averiguar Normalidade Amostra de AIR	Shapiro	0.965	0.006	Não Normal
Averiguar Normalidade Amostra de FOB	Shapiro	0.970	0.016	Normal
Averiguar Normalidade Amostra de MAIL	Shapiro	0.963	0.005	Não Normal
Averiguar Normalidade Amostra de RAIL	Shapiro	0.958	0.002	Não Normal
Averiguar Normalidade Amostra de REG AIR	Shapiro	0.947	3.39e-04	Não Normal
Averiguar Normalidade Amostra de SHIP	Shapiro	0.967	9.85e-03	Não Normal
Averiguar Normalidade Amostra de TRUCK	Shapiro	0.978	0.079	Normal
Testar Hipótese Nula	Kruskal-Wallis	156.8	2.78ze-31	Rejeitar H0

Tabela 3: Testes estatísticos realizados para decidir sobre a hipótese de que não existe diferença entre a popularidade de cada transporte.

	AIR	FOB	MAIL	RAIL	REG AIR	SHIP	TRUCK
AIR	1.00	1.00	1.00	1.00	2.53e-13	1.44e-04	1.00
FOB	1.00	1.00	1.00	1.00	2.52e-14	5.37e-04	1.00
MAIL	1.00	1.00	1.00	1.00	8.740e-12	1.45e-05	1.00
RAIL	1.00	1.00	1.00	1.00	5.31e-12	2.05e-05	1.00
REG AIR	1.00	1.00	1.00	1.00	1.00	5.50e-33	1.00
SHIP	1.00	1.00	1.00	1.00	5.50e-33	1.00	1.00
TRUCK	1.00	1.00	1.00	1.00	5.36e-14	3.54e-04	1.00

Tabela 4: P-valores para testes pareados de Dunn na análise post-hoc à rejeição da hipótese de que não existe diferença entre a popularidade de cada transporte. A vermelho identificam-se os valores que indicam uma diferença significativa ao nível de significância $\alpha = 0.01$.

4. Conclusão

No decurso deste trabalho, a elevada quantidade de atributos a serem explorados levou a que utilizássemos múltiplas formas de agregação dos dados e subsequentemente diversas ferramentas de visualização de dados, como pie e bar charts, boxplots e line charts com representação dos desvios.

As diferentes vistas sobre os dados levaram a conclusões muito distintas, por vezes erróneas, reforçando a importância de escolher técnicas de agregação e visualização adequadas. Torna-se também evidente a necessidade de realizar testes estatísticos e calcular intervalos de confiança de forma a criar confiança nas conclusões obtidas.

Sendo o dataset utilizado gerado aleatoriamente para os fins deste trabalho foi difícil encontrar padrões significativos no consumo que seriam de esperar por diferenças culturais ou inerentes ao comportamento humano. Esta ausência de padrões tornou difícil a utilização significativa de outras técnicas de análise e exploração de dados como regressões e clustering.