Instituto Superior de Contabilidade e Administração de Coimbra
Instituto Politécnico de Coimbra

Mestrado em Análise de Dados e Sistemas de Apoio à Decisão

Trabalho N. º1 de
Análise Estatística de Dados

BÁRBARA XAVIER SOBRAL (2024104488)

MARIA RIBEIRO MARGARIDO (2024104121)

MARIANA SOFIA MENDES PRATA DE ALMEIDA (2024102807)

COIMBRA

15-10-2024

Introdução

O presente relatório foi realizado no âmbito da unidade curricular de Análise Estatística de Dados e tem como objetivo analisar o comportamento das vendas de café, considerando vários fatores, numa cadeia de lojas.

A análise vai ter como base dados o ficheiro 'DadosTrabalho1.sav' de onde se gerou uma amostra aleatória de dimensão N=880 – 5 * 4 = 860, obtida através do programa SPSS e seguindo os seguintes passos: Data -> Select Cases -> Random sample of cases -> Sample -> Exactly -> 860 ... 880.

A amostra em estudo apresenta 6 variáveis que podem ser divididas em dois grupos:

Variáveis quantitativas:

Vendas: valor das vendas (em dólares) dos dois tipos de café mais vendidos;

Despmark: despesas relacionadas com a promoção do produto (em dólares);

Vendasesp: valor esperado de vendas (em dólares).

Variáveis qualitativas:

Tipo: tipo de café (tipo1/tipo2);

Regiao: região onde a loja está localizada (Norte, Centro, Sul);

Dim: dimensão da loja (Pequena/Grande).

Estamos agora na posse de todos os dados necessários para avançar com o nosso estudo de forma clara e objetiva.

Análise Descritiva e Comparativa

Vamos iniciar o nosso relatório estatístico com uma análise descritiva e comparativa das ***Vendas*** segundo a dimensão das lojas (***Dim)***. Para tal, através do seguinte caminho, Analyze -> Descriptive Statistics -> Explore, conseguimos obter outputs com informações suficientes para fazer a nossa análise. A variável ***Vendas*** é uma variável quantitativa, sendo por isso a variável dependente na "*Dependent list*", enquanto que a variável **Dim**, variável qualitativa, é colocada no "*Factor List*".

Vendas de café por dimensão da loja

		Casos					
		Válido		Omisso		Total	
	Dimensão da loja	N	Porcentagem	N	Porcentagem	N	Porcentagem
Valor de vendas	Grande	343	100,0%	0	0,0%	343	100,0%
	Pequena	517	100,0%	0	0,0%	517	100,0%

Tabela 1- Distribuição da amostra por dimensão da loja

A tabela 1 demonstra a distribuição da variável ***Vendas*** por dimensão da loja (***Dim)***. Após uma observação cuidada e assumindo a normalidade da amostra por questões de simplicidade, encontramos-nos em condições de analisar as estatísticas descritivas obtidas.

Descritivas

	Dimensão da loja			Estatística	Estatística do teste Padrão
Valor de vendas	Grande	Média		291,42	9,105
		95% de Intervalo de Confiança para Média	Limite inferior	273,51	
			Limite superior	309,33	
		5% da média aparada		282,97	
		Mediana		242,00	
		Variância		28432,700	
		Erro Padrão		168,620	
		Mínimo		63	
		Máximo		678	
		Amplitude		615	
		Amplitude interquartil		300	
		Assimetria		,718	,132
		Curtose		-,694	,263

Tabela 2- Estatísticas descritivas - Vendas da loja grande

Ao analisar os dados referentes ao volume de vendas na loja de maior dimensão, observa-se que a **média** foi de **291,42**, enquanto a **mediana** se estabeleceu em **242,00**.

No que concerne à dispersão, ou variabilidade dos dados, identificamos que o volume de vendas variou entre um **valor mínimo** de **63,00** e um **máximo** de **678,00**, ou seja, a **amplitude**, diferença entre o maior e o menor volume de vendas, foi de **615,00**. Por outro lado, a **amplitude interquartil**, que abrange os 50% dos dados centrais, é de **300,00**, mostrando a dispersão nesse conjunto de dados. O **desvio padrão** foi de **168,620**, o que revela o quanto o volume de vendas se afasta da sua média.

Relativamente à forma da distribuição dos dados, a análise de simetria (**Skewness**) indica uma **distribuição assimétrica positiva** dos dados, com um valor de **0,718**, sendo por isso representados num gráfico com uma cauda superior longa. Isso significa que há mais vendas concentradas abaixo da mediana, com uma menor quantidade de valores altos, confirmando o padrão Moda < Mediana < Média. Tais conclusões podem também ser tiradas a partir do diagrama de extremos e quartis (Gráfico 1). O **coeficiente de curtose**, por sua vez, apresenta um valor de **-0,694**, caracterizando uma **distribuição platicúrtica**, ou seja, **mais achatada**.

Pequena	Média		127,69	2,830
	95% de Intervalo de Confiança para Média	Limite inferior	122,13	
		Limite superior	133,25	
	5% da média aparada		122,15	
	Mediana		117,00	
	Variância		4141,415	
	Erro Padrão		64,354	
	Mínimo		32	
	Máximo		387	
	Amplitude		355	
	Amplitude interquartil		72	
	Assimetria		1,304	,107
	Curtose		2,084	,214

Tabela 3- Estatísticas descritivas - Vendas da loja pequena

Por outro lado, ao realizar a análise das estatísticas descritivas da loja pequena, verifica-se que a **média** do volume de vendas foi de **127,69** e a **mediana** foi de **117,00**, valores expectavelmente mais baixos.

Relativamente às medidas de dispersão, verifica-se um valor **mínimo** de volume de vendas foi de **32,00** e um valor **máximo** foi de **387,00**, sendo por isso a **amplitude** de

355,00. A **amplitude interquartis**, que representa a dispersão dos 50% centrais do conjunto de dados, é de **72,00**. O **desvio padrão** medido é de **64,354**.

Em relação à simetria (**Skewness**), é também visível uma **distribuição assimétrica positiva** no volume das vendas da loja de menor dimensão, assumindo um valor de **1,304** e novamente é verificado que Moda < Mediana < Média, pelo que se identifica um desvio à esquerda da distribuição dos dados, associado a um maior volume de vendas inferiores à mediana (Gráfico 1). No que concerne ao **coeficiente de curtose**, identifica-se uma curtose de **2,084** pelo que a curva de distribuição é uma **curva leptocúrtica**, sendo por isso **mais alongada**.

Percentis

			Percentis						
		Dimensão da loja	5	10	25	50	75	90	95
Média Ponderada (Definição 1)	Valor de vendas	Grande	96,40	114,80	140,00	242,00	440,00	546,00	627,60
		Pequena	46,00	56,00	82,00	117,00	153,50	213,00	265,10
Teste de Tukey	Valor de vendas	Grande			140,50	242,00	431,50		
		Pequena			82,00	117,00	153,00		

Tabela 4- Percentis relativos às vendas por dimensão de loja

De referir ainda, a importância da análise dos **percentis** nas vendas de café por dimensão de loja. Deste modo, na loja de grande dimensão, observa-se que 25% do volume de vendas foram inferiores a **140,00** e outros 25% foram superiores a **440,00**.

Já na loja de pequena dimensão, 25% das vendas foram inferiores a **82,00** e outros 25% foram superiores a **153,50**.

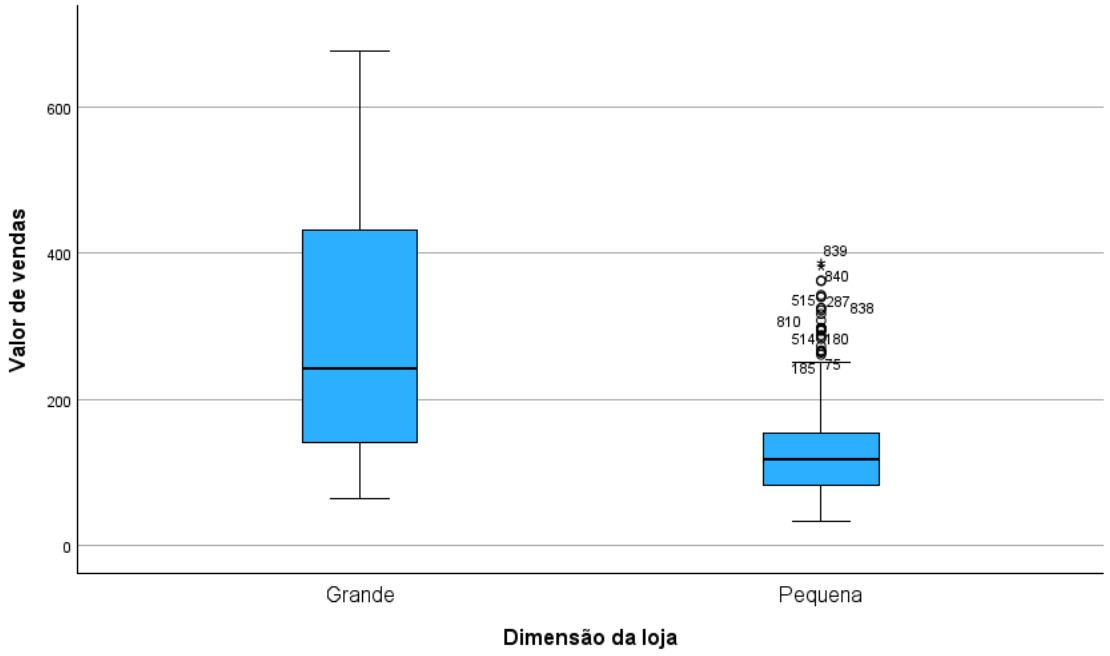

Gráfico 1 - Box Plot de variação dos dados observados de vendas em ambas as lojas

O gráfico 1 representa um Box Plot que interpreta a variação dos valores das vendas observadas nos dois tipos de lojas: grande e pequena. Após uma análise cuidada é possível concluir que as lojas de maior dimensão apresentam uma maior variabilidade nos valores das vendas, indicada pela maior extensão da caixa. Esta interpretação sugere que as vendas em lojas grandes são mais suscetíveis a flutuações, enquanto que, por outro lado, as lojas pequenas têm uma menor variabilidade, traduzindo-se em vendas mais consistentes e menos sujeitas a variações.

Conforme dito anteriormente, a mediana é visível neste tipo de gráfico, nomeadamente na linha dentro de cada caixa. Assim, comparando as duas categorias, podemos observar que a mediana das vendas é maior em lojas grandes.
Além disso, os outliers severos e moderados são indicados pelos valores fora do corpo principal do Box Plot, representando valores de vendas significativamente diferentes dos dados restantes. A presença de outliers pode indicar eventos excepcionais, como promoções ou períodos de procura elevada de café.

Correlação

Uma vez que vamos analisar simultaneamente a correlação entre duas variáveis (***Vendas*** e ***Despmark)*** numa amostra de dimensão N=860 e ambas as variáveis são quantitativas rácio temos de utilizar o coeficiente de Pearson para podermos tirar conclusões.

Através do caminho Analyze -> Correlate -> Bivariate -> escolha as variáveis ***Vendas*** e ***Despmark*** e selecione Pearson -> OK, obtemos o seguinte output:

Correlações

		Despesas de marketing	Valor de vendas
Despesas de marketing	Correlação de Pearson	1	**,654****
	Sig. (2 extremidades)		<,001
	N	860	860
Valor de vendas	Correlação de Pearson	**,654****	1
	Sig. (2 extremidades)	<,001	
	N	860	860

**. A correlação é significativa no nível 0,01 (2 extremidades).

Tabela 5 - Correlação de Pearson

O coeficiente de Pearson ou de correlação linear pode variar entre -1 e 1, e mede a intensidade e a direção da relação linear entre duas variáveis. Quanto maior for o valor do coeficiente, maior será o grau de associação linear entre as variáveis. Um valor do coeficiente negativo indica uma associação negativa entre as duas variáveis, isto é, quando os valores de uma das variáveis aumentam, existe tendência para que os valores da outra variável diminuam. Pelo contrário, um valor do coeficiente positivo indica uma associação linear positiva entre as duas variáveis, isto é, quando os valores de uma das variáveis aumentam, existe tendência para que os valores da outra variável também aumentem.

Concluímos assim que **existe uma correlação positiva e moderada** entre as duas variáveis dado que o **coeficiente de Pearson** entre as ***Vendas*** e as ***Despmark*** é de ***0,654***. Ou seja, há uma tendência para que, à medida que as despesas de marketing aumentam, o valor das vendas também aumente. Por outras palavras, se existir um maior

investimento na promoção do produto, o número de vendas tende a aumentar, uma vez que é conhecido por mais pessoas.

Diferenças significativas

Médias das Vendas e Tipos de Café

Vamos agora analisar se existem diferenças significativas entre as médias das vendas (***Vendas***) dos dois tipos de café (***Tipo***). Para tal, foi utilizado o teste ***t-Student***, uma vez que se pretende comparar a média de uma variável quantitativa (***Vendas***) em dois grupos independentes (***Tipo***: café tipo 1 e café tipo 2). Para a realização deste teste teve de ser criada uma variável nova numérica, ***'Café'***, para se poder agrupar os tipos de café em apenas duas categorias (1 -Tipo 1 e 2 -Tipo 2) e assim poderem ser calculadas as estatísticas necessárias.

Os seguintes passos, Analyze -> Compare Means -> Independent-Samples T Test -> selecionar ***Vendas*** (test variable) e ***Café*** (grouping variable) -> em Define Groups colocar os códigos das duas categorias: 1 e 2 -> Continue -> OK, levam-nos aos outputs que precisamos de analisar.

Estatísticas de grupo

	Tipo de produto	N	Média	Desvio Padrão	Erro de média padrão
Valor de vendas	Tipo 1	401	201,93	142,371	7,110
	Tipo 2	459	185,18	141,886	6,623

Tabela 6 - Distribuição das vendas por tipo de café

Sendo as amostras independentes e assumindo que ***Vendas*** (variável dependente quantitativa) segue uma distribuição normal em cada grupo, as nossas hipóteses são as seguintes:

H0: Não há diferenças significativas entre as médias das vendas de café dos dois grupos.

vs

H1: Há diferenças significativas entre as médias das vendas de café dos dois grupos.

Teste de amostras independentes

		Teste de Levene para igualdade de variâncias		teste-t para Igualdade de Médias							
						Significância		Diferença média	Erro de diferença padrão	95% Intervalo de Confiança da Diferença	
		Z	Valor-p	t	df	Unilateral p	Bilateral p			Inferior	Superior
Valor de vendas	Variâncias iguais assumidas	,032	**,857**	**1,724**	858	,042	**,085**	16,752	9,714	-2,314	35,818
	Variâncias iguais não assumidas			1,724	841,793	,043	,085	16,752	9,716	-2,319	35,823

Tabela 7 - teste t-student amostras independentes

Como podemos verificar com o teste de Levene, a variável dependente ***Vendas*** possui homogeneidade de variâncias entre os grupos, uma vez que ***valor-p = 0,857 > α = 0,05***, assim, os nossos resultados vão ser lidos na primeira linha da tabela 7.

Dado que as hipóteses do teste são as médias serem ou não serem significativamente diferentes, consideramos que este teste é bilateral identificando na tabela um ***valor-p = 0,085***.

Podemos então, através de duas maneiras, constatar que **mantemos a hipótese nula (*H0*)**:

- utilizando o intervalo de confiança de 95%: 0 pertence ao IC;
- ***valor-p > α = 0,05***.

Neste sentido, podemos concluir que **não há uma diferença significativa** nas vendas médias entre o café tipo 1 e o café tipo 2.

Tamanhos de efeitos de amostras independentes

		Padronizador[a]	Estimativa de ponto	Intervalo de Confiança 95%	
				Inferior	Superior
Valor de vendas	d de Cohen	142,113	**,118**	-,016	,252
	Correção de Hedges	142,237	,118	-,016	,252
	Delta do vidro	141,886	,118	-,016	,252

a. O denominador usado na estimativa dos tamanhos dos efeitos.
O d de Cohen usa o desvio padrao agrupado.
A correçao de Hedges usa o desvio padrao agrupado, além de um fator de correçao.
O delta de Glass usa o desvio padrao de amostra do grupo de controle (ou seja, o segundo) grupo.

Tabela 8 - Medidas de Magnitude do efeito

Por fim, para analisar a magnitude da diferença entre as vendas médias, podemos utilizar o coeficiente de Cohen. Este coeficiente, ***d,*** assume um valor de ***0,118***, o qual revela, em termos práticos, uma diferença insignificante entre as vendas médias dos dois tipos de café (***d < 0,2***).

Vendas e Vendas Esperadas

Assim como no exercício anterior, também aqui, para analisar se existiam diferenças significativas entre as ***Vendas*** e as ***Vendasesp*** foi realizado um teste ***t-Student***, mas com uma diferença: neste caso, tratou-se de um teste para amostras emparelhadas em vez de amostras independentes. Este tipo de teste foi escolhido porque o objetivo era comparar duas médias — as vendas reais (***Vendas***) e as vendas esperadas (***Vendasesp***) — dentro do mesmo grupo, ***Vendas*** (variável quantitativa). Ou seja, a mesma amostra de ***Vendas*** foi avaliada em duas situações distintas.

Estatísticas de amostras emparelhadas

		Média	N	Desvio Padrão	Erro de média padrão
Par 1	Valor de vendas	192,99	860	142,276	4,852
	Valor esperado de vendas	194,17	860	153,186	5,224

Tabela 9 - Distribuição das vendas em duas situações

Correlações de amostras emparelhadas

		N	Correlação	Significância Unilateral p	Significância Bilateral p
Par 1	Valor de vendas & Valor esperado de vendas	860	**,968**	<,001	<,001

Tabela 10 - Correlação entre as variáveis

Na tabela 10, verifica-se que há uma correlação muito forte entre as variáveis uma vez que o coeficiente de Pearson é ***0,968***. Neste sentido, podemos concluir que temos efetivamente amostras emparelhadas.

Assim, sendo as amostras emparelhadas e assumindo que ***Vendas*** (variável dependente quantitativa) segue uma distribuição normal em cada grupo, as nossas hipóteses são as seguintes:

H0: Não há diferenças significativas entre as vendas reais e as vendas esperadas de café.

vs

H1: Há diferenças significativas entre as vendas reais e as vendas esperadas de café.

Teste de amostras emparelhadas

		Diferenças emparelhadas					t	df	Significância	
					95% Intervalo de Confiança da Diferença					
		Média	Desvio Padrão	Erro de média padrão	Inferior	Superior			Unilateral p	Bilateral p
Par 1	Valor de vendas - Valor esperado de vendas	-1,181	38,924	1,327	-3,787	1,424	**-,890**	859	,187	**,374**

Tabela 11 - teste t-student amostras independentes

Como verificamos pelas hipóteses, este teste também é bilateral, o qual nos permite identificar um ***valor-p = 0,374***. Utilizando um nível de significância de 5%, ***valor-p > α = 0,05***, logo **mantemos a hipótese nula (*H0*)**. Neste sentido, podemos concluir que **não há uma diferença significativa** entre as vendas reais e as vendas esperadas de café.

Tamanhos de efeitos de amostras em pares

			Padronizador[a]	Estimativa de ponto	Intervalo de Confiança 95%	
					Inferior	Superior
Par 1	Valor de vendas - Valor esperado de vendas	d de Cohen	38,924	**-,030**	-,097	,037
		Correção de Hedges	38,958	-,030	-,097	,036

a. O denominador usado na estimativa dos tamanhos dos efeitos.
O d de Cohen usa o desvio padrão de amostra da diferença média.
A correção de Hedges usa o desvio padrão de amostra da diferença média, além de um fator de correção.

Tabela 12 - Medidas de Magnitude do efeito

Em termos práticos, recorrendo ao coeficiente de Cohen, temos uma diferença de efeito bastante reduzido (***d = -0,030***).

Os outputs dos testes ***t-student*** anteriores foram obtidos a partir do seguinte caminho: Analyze -> Compare Means -> Paired-Samples T Test -> selecionar ***Vendas*** para Variable 1 e ***Vendasesp*** para Variable 2 -> OK.

Região e Vendas

Neste ponto queremos perceber se a região onde se localizam as lojas (***Regiao)*** tem um impacto significativo nas vendas de café (***Vendas***).

Para este efeito, foi utilizada a técnica ANOVA pois permite-nos analisar o efeito de um fator (com 3 ou mais níveis), que neste caso se trata da ***Regiao*** (Norte, Centro e Sul), numa variável dependente que será as ***Vendas***, testando se as médias da variável dependente em cada categoria do fator diferem ou não entre si. Tivemos novamente que criar uma nova variável numérica, ***Regiao_Cod*** (1 - Norte, 2 - Centro e 3 - Sul), para podermos realizar as estatísticas necessárias à análise.

Analyze -> Compare Means -> One-Way ANOVA -> selecionar ***Vendas*** (dependente list), e ***Regiao_Cod*** (factor) -> em Options selecionar Descriptives, Homogeneity of variance test e Means Plot -> OK, é o caminho que nos leva aos seguintes outputs.

Descritivas

Valor de vendas

	N	Média	Desvio padrão	Erro Padrão	95% de Intervalo de Confiança para Média		Mínimo	Máximo
					Limite inferior	Limite superior		
Norte	295	189,05	131,779	7,672	173,95	204,15	32	643
Centro	299	194,11	145,732	8,428	177,52	210,69	39	678
Sul	266	196,11	149,797	9,185	178,02	214,19	41	678
Total	860	192,99	142,276	4,852	183,47	202,51	32	678

Tabela 13 - Distribuição das vendas nas três regiões

Testes de homogeneidade de variâncias

		Estatística de Levene	df1	df2	Sig.
Valor de vendas	Com base em média	**1,087**	2	857	**,338**
	Com base em mediana	,010	2	857	,990
	Com base em mediana e com gl ajustado	,010	2	811,408	,990
	Com base em média aparada	,419	2	857	,658

Tabela 14 - Homogeneidade de variâncias Região e Vendas

Considerando os seguintes pressupostos:

- 3 grupos independentes (Norte, Centro e Sul),
- assumindo que variável dependente quantitativa (***Vendas***) segue distribuição normal em cada grupo,
- e homogeneidade de variâncias, uma vez que no teste de Levene, ***valor-p = 0.338 > α = 0,05***, não sendo por isso necessário ver o número de observações em cada grupo, podemos, deste modo, testar as próximas hipóteses:

H0: Não há diferenças significativas entre as médias das Vendas nas 3 regiões.

vs

H1: Existe pelo menos um par de médias que é significativamente diferente das restantes.

ANOVA

Valor de vendas

	Soma dos Quadrados	df	Quadrado Médio	Z	Sig.
Entre Grupos	7524,183	2	3762,091	,185	**,831**
Nos grupos	17380732,760	857	20280,902		
Total	17388256,943	859			

Tabela 15 - Teste ANOVA

Utilizando um nível de significância de 5%, ***valor-p = 0,831 > α = 0,05***, logo **mantemos a hipótese nula (*H0*)**. Neste sentido, podemos concluir que **não há um impacto significativo** entre as vendas de café (***Vendas***) e a região onde se localizam as lojas (***Regiao)***.

Tamanhos do efeito do ANOVA[a,b]

		Estimativa de ponto	Intervalo de Confiança 95% Inferior	Superior
Valor de vendas	Eta quadrado	**,000**	,000	,005
	Epsilon quadrado	-,002	-,002	,003
	Efeito fixo do Omega quadrado	-,002	-,002	,003
	Efeito aleatório do Omega quadrado	-,001	-,001	,001

a. Eta quadrado e Epsilon quadrado são estimados com base no modelo de efeito fixo.

b. As estimativas negativas, mas menos tendenciosas, são mantidas, não arredondadas para zero.

Tabela 16 - Medidas de magnitude do efeito (ANOVA)

Por fim, para analisar a magnitude da diferença entre as ***Vendas*** e a ***Regiao***, utilizamos o coeficiente $\eta 2$, que nos diz que, em termos práticos ($\eta 2 = 0{,}0 < 0{,}01$), existe uma diferença insignificante entre as médias das vendas de café e a região onde se localizam as lojas.

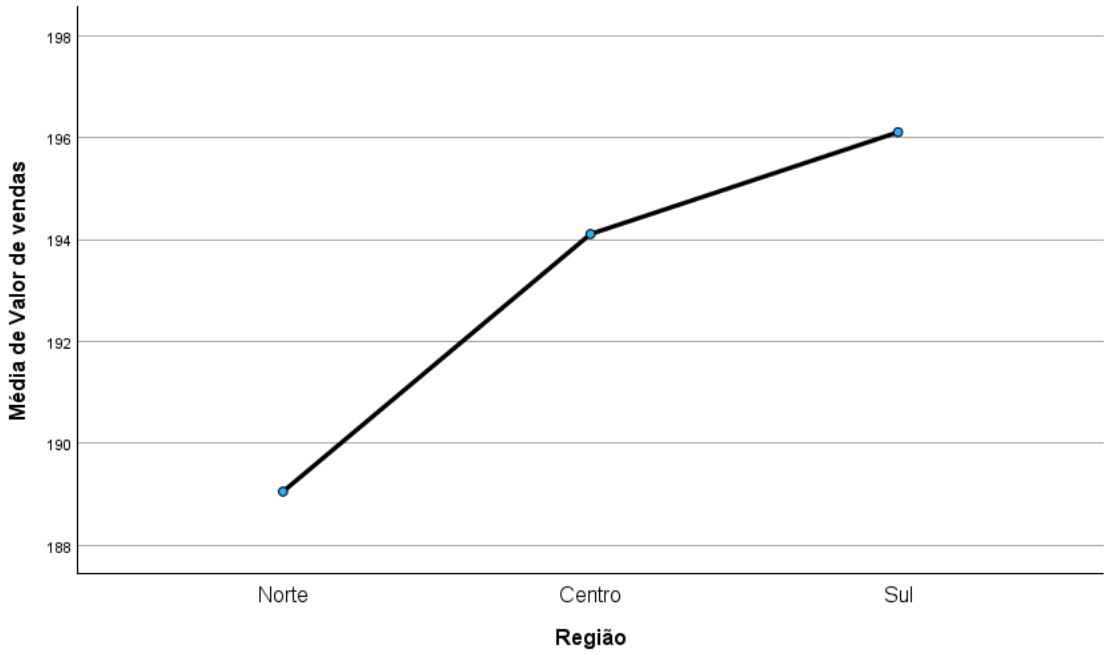

Gráfico 2 - Média das Vendas por Região

Conclusão

A análise estatística apresentada neste relatório permitiu identificar fatores que influenciam o comportamento das vendas de café numa cadeia de lojas. Primeiramente é fulcral referir que não foram encontradas diferenças significativas entre as médias de vendas dos dois tipos de café, assim como entre as vendas reais e as vendas esperadas. Esses resultados indicam que a escolha entre os tipos de café não afeta diretamente o volume de vendas.

A correlação entre as despesas de marketing e as vendas foi positiva e moderada, sugerindo que um maior investimento em publicidade tende a elevar o volume de vendas. As lojas de maior dimensão apresentaram uma média de vendas superior, mas com maior variabilidade nos resultados, indicando um comportamento mais volátil. Por outro lado, as lojas menores demonstraram uma estabilidade maior nas vendas.

A análise de variância (ANOVA) revelou que a localização das lojas por região (Norte, Centro, Sul) não tem um impacto significativo sobre as vendas, sugerindo que as vendas são consistentes entre as diferentes regiões.

Em suma, conclui-se que a gestão de marketing desempenha um papel crucial na maximização das vendas, evidenciado pela correlação positiva entre despesas de marketing e volume de vendas. As análises realizadas no software SPSS permitiram a aplicação de diversos testes estatísticos, como o teste t-Student e a ANOVA, que fundamentaram as conclusões sobre a comparação entre tipos de café e a avaliação do impacto da localização das lojas. No entanto, as variações entre os tipos de café, as diferenças entre vendas esperadas e reais, e a localização das lojas exercem um impacto limitado no desempenho comercial.