Linear Relationship Between Material Type and Zara's Sales

C7: Ruiyu Zhao, Joanna Zhu, Jie Cai, Clarence Xie, Joshua Zhang, Richa Desai

1. Problem Definition & Research Question

Exploring how material choice influences Zara's sales performance

Focus:	How different materials (e.g., cotton, polyester, wool blend) affect Zara's sales volume.
Why it matters:	Material directly influence shoppers' comfort and quality, which strongly shape their willingness to buy. Industry reports (e.g.,McKinsey's 2023 State of Fashion Report) highlights material as a major driver of product cost and consumer value.
Research Question:	Do different material types and other factors significantly impact Zara's sales?
Goal:	Use data to identify the effect of material while accounting for other key factors such as price and promotion.
Hypothesis:	Products made with natural materials (like cotton) may have lower sales volumes than those made with synthetic materials.
Dataset:	Real-world Zara retail dataset from Kaggle, including over 20,000 product observations with attributes such as price, material, promotion (aggregated by individual products)

2. Ideal Experiment

01

Independent Variable

Material as a categorical attribute such as cotton, linen, wool, polyester

02

Dependent Variable

Sales Volume measured in units

03

Other Relevant Variables

Collection Season, In-Store Product Position, Promotion Status

2. Ideal Experiment

Potential Confounders

- **Product category** — polyester might dominate in dresses, while cotton might lead in basics like T-shirts
- **Regional preferences** — customers in warmer climates may prefer light natural fabrics
- **Season** — wool might sell better in winter, linen in summer
- **Price** — higher prices may lower sales volume; lower prices may raise it without true demand
- **Promotion** — promotions boost sales temporarily. Even an expensive cotton item could sell well if discounted

2. Ideal Experiment

The Limitations of Observational Data

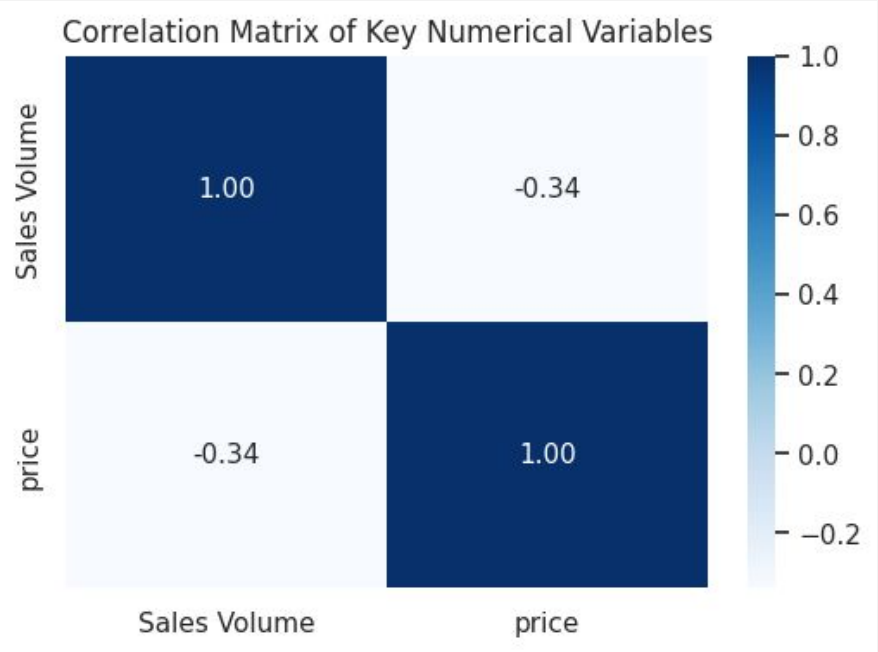
- The dataset lacks full numeric detail (most are categorical).
- Missing key variables that affect sales (inventory, production cost, marketing).
- Nothing is randomized, we can't isolate material effects from other factors.

Boundaries of the Setup

- We can identify linear relationship between material and sales volume.
- We cannot prove that material alone causes higher or lower sales.

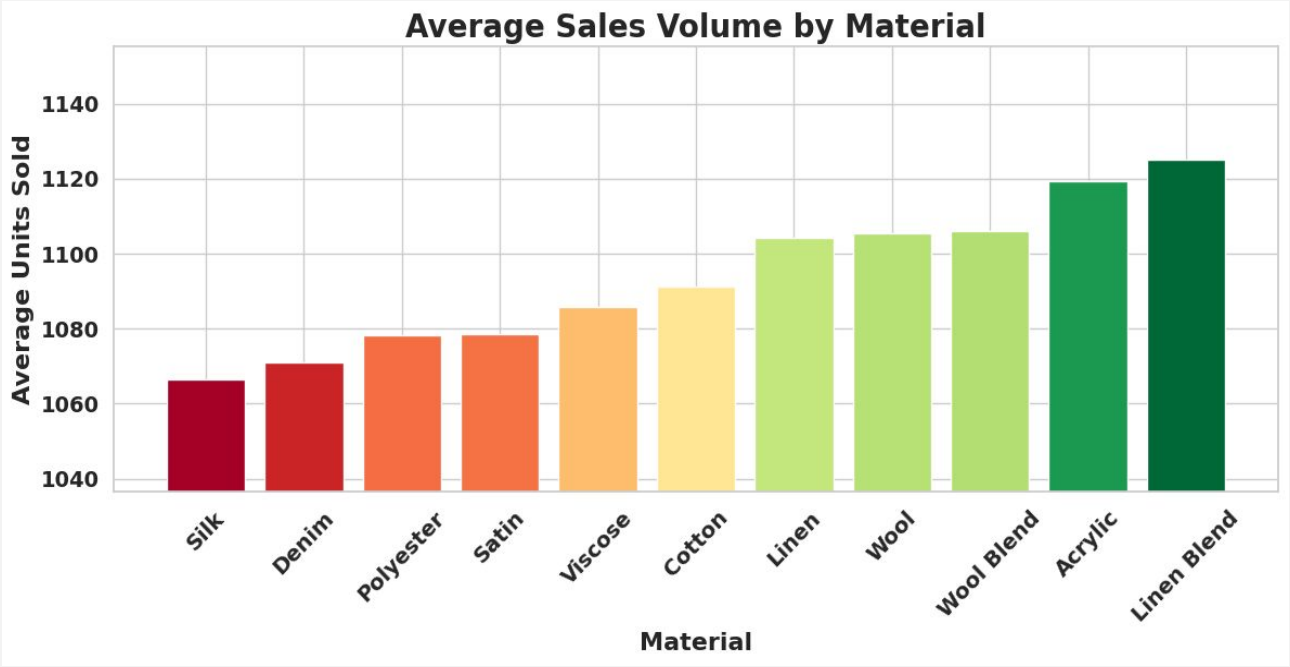
Descriptive Statistics & Correlation Matrix

	Sales Volume	price
count	20252.00	20252.00
mean	1097.40	41.95
std	298.23	23.38
min	518.00	12.00
25%	849.00	23.95
50%	990.00	35.95
75%	1364.25	53.95
max	1940.00	134.99
median	990.00	35.95
mode	848.00	19.99

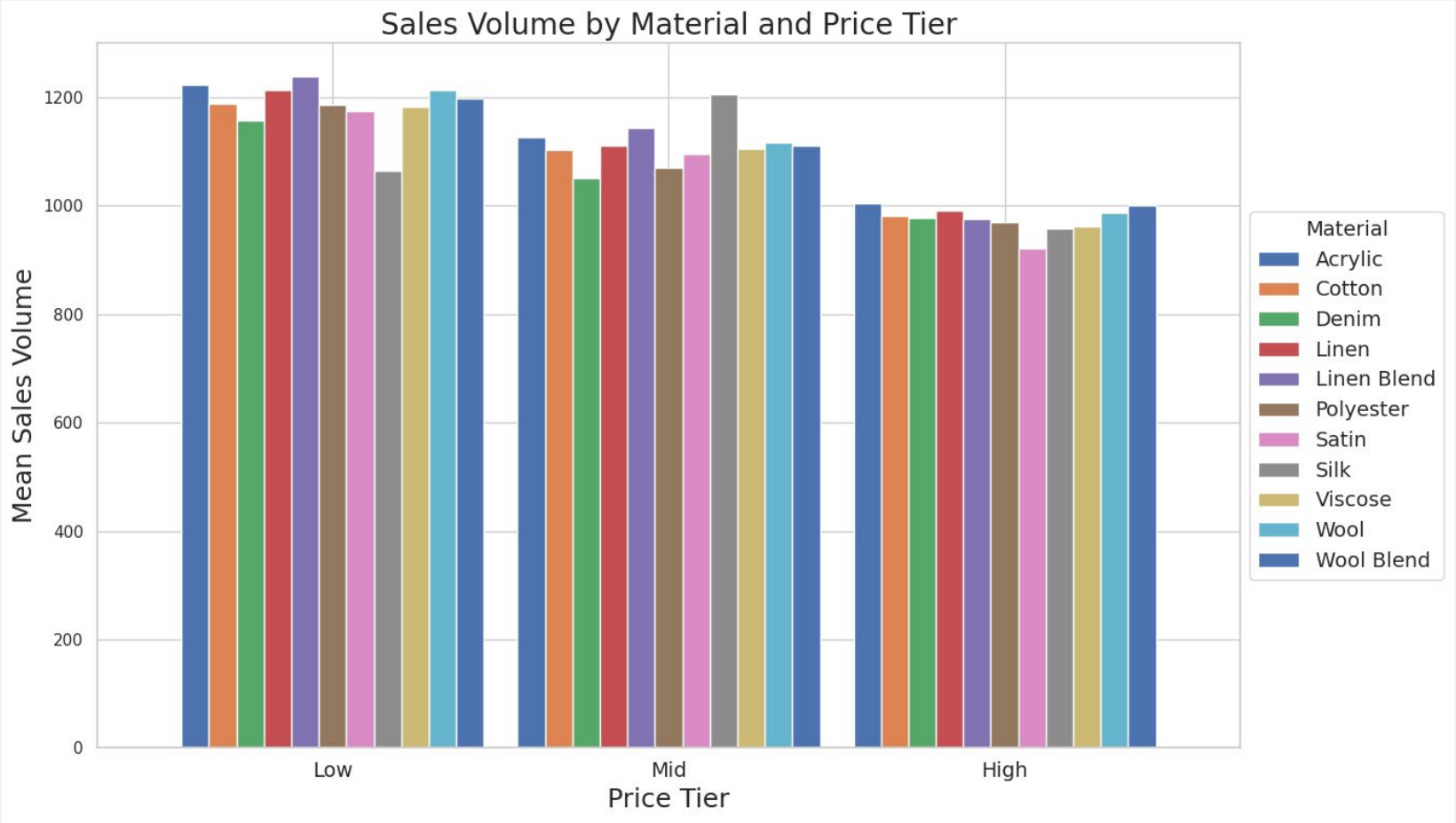


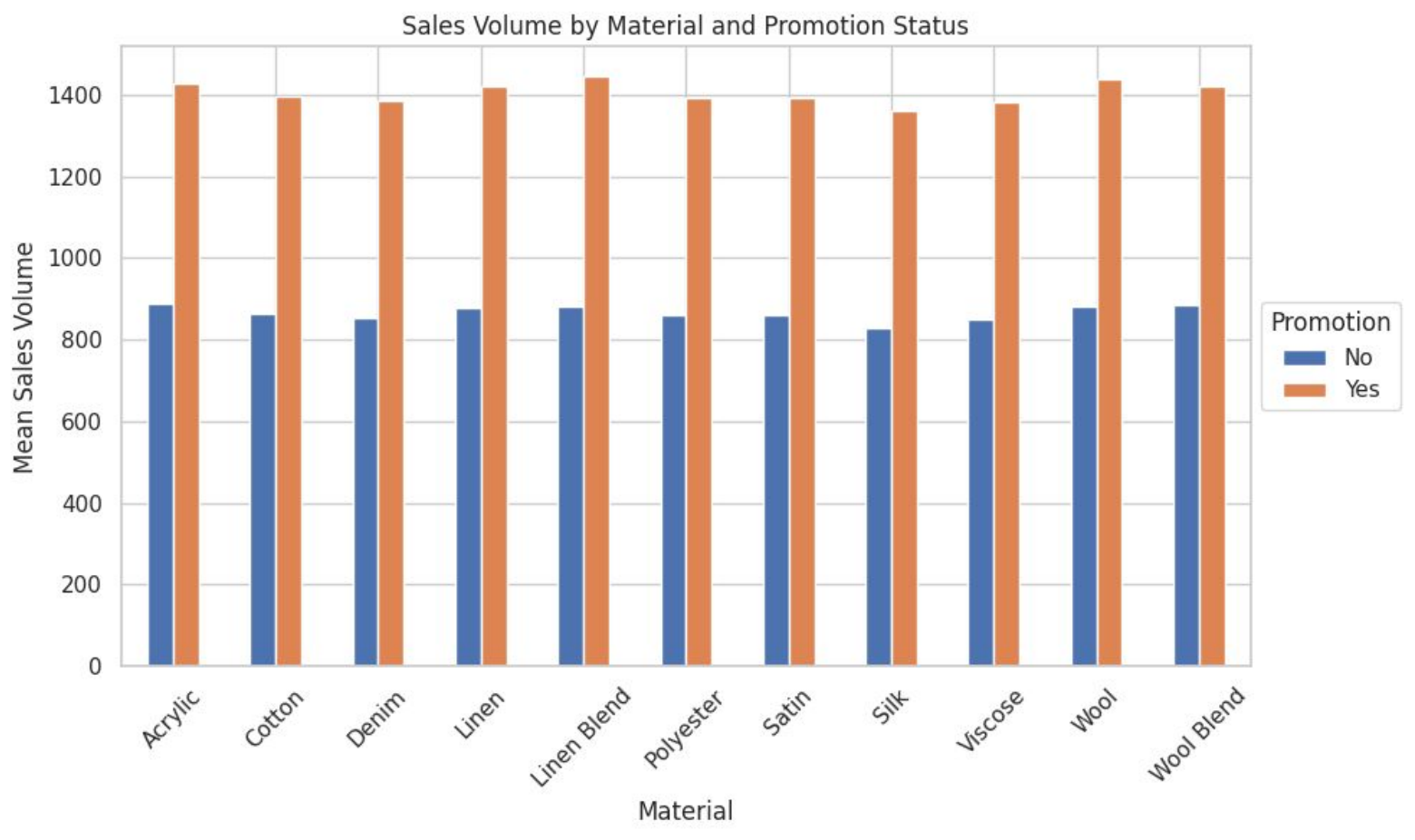
	Sales Volume	price
Sales Volume	1.00000	-0.33778
price	-0.33778	1.00000

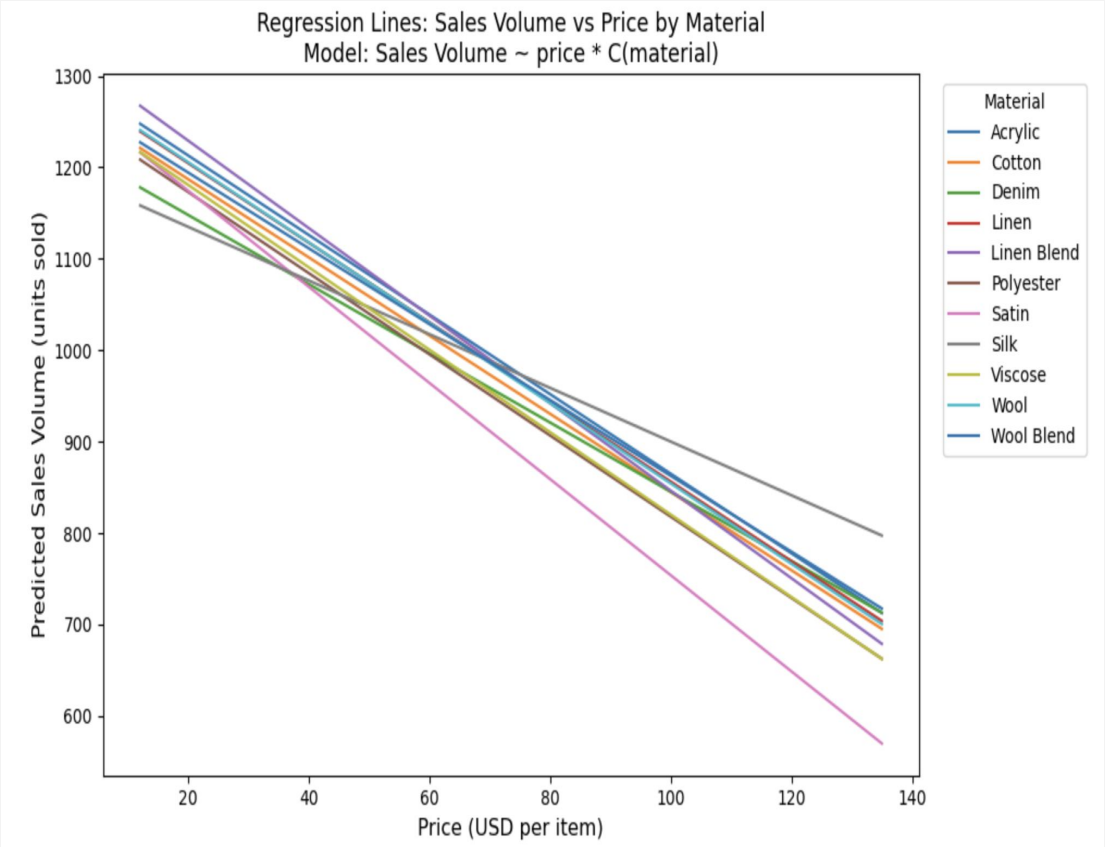
material	price
Silk	43.292895
Linen	43.060688
Wool	42.772962
Cotton	42.379917
Linen Blend	41.705118
Acrylic	41.518774
Polyester	41.335874
Wool Blend	41.248346
Viscose	41.035859
Denim	40.327945
Satin	38.195682



- Do expensive materials actually sell more or less?
- Does material help or hurt sales volume?
- What is the relationship between material, price, and sales?


Sales Volume by Material and Price Tier
Mean Sales Volume
0
200
400
600
800
1000
1200
Low
Mid
High
Price Tier
Material
Acrylic
Cotton
Denim
Linen
Linen Blend
Polyester
Satin
Silk
Viscose
Wool
Wool Blend


Sales Volume by Material and Promotion Status
Mean Sales Volume
0
200
400
600
800
1000
1200
1400
Acrylic
Cotton
Denim
Linen
Linen Blend
Polyester
Satin
Silk
Viscose
Wool
Wool Blend
Material
Promotion
No
Yes



- Price * material predicting Sales Volume
- Materials as categorical variable so using acrylic as dummy variables
- Each material's coefficient shows how its sales differ from acrylic at the same price
- Higher prices → lower sales

OLS Regression Results

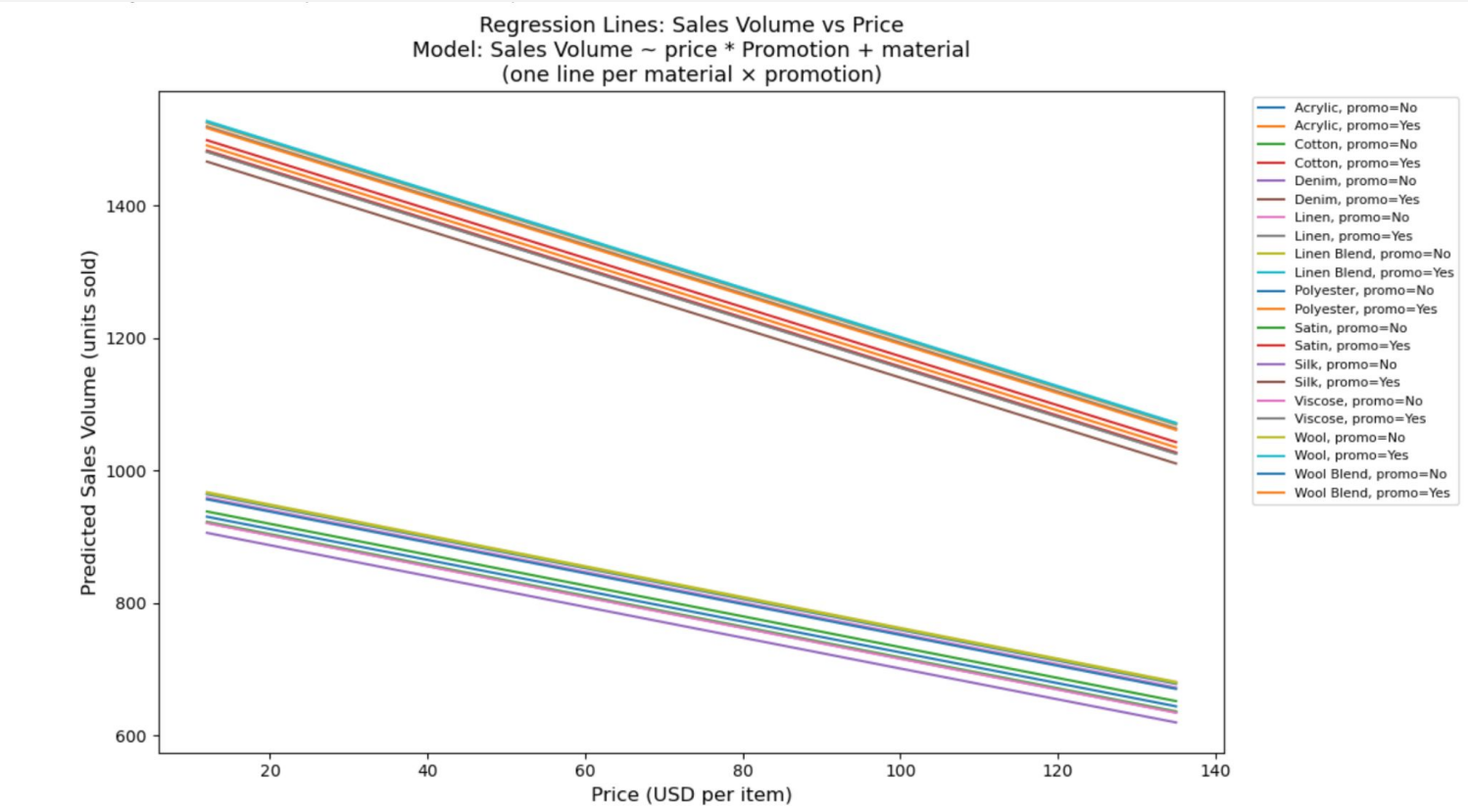
Dep. Variable:	Q('Sales Volume')	R-squared:	0.117
Model:	OLS	Adj. R-squared:	0.116
Method:	Least Squares	F-statistic:	127.7
Date:	Wed, 03 Dec 2025	Prob (F-statistic):	0.00
Time:	14:59:16	Log-Likelihood:	-1.4287e+05
No. Observations:	20252	AIC:	2.858e+05
Df Residuals:	20230	BIC:	2.860e+05
Df Model:	21		
Covariance Type:	nonrobust		

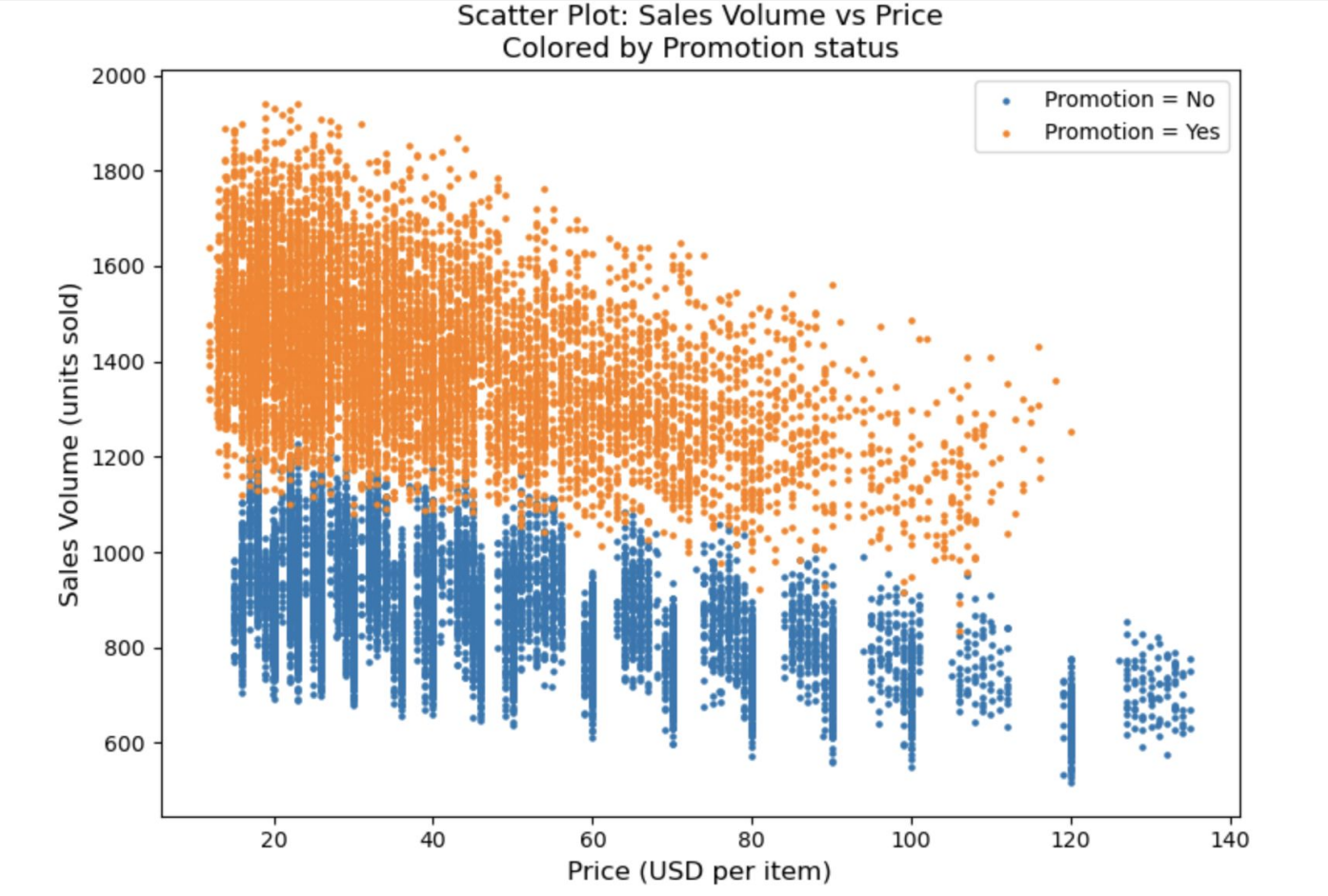
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1299.8990	19.652	66.147	0.000	1261.380	1338.418
C(material)[T.Cotton]	-27.4250	21.746	-1.261	0.207	-70.049	15.199
C(material)[T.Denim]	-76.4423	26.401	-2.895	0.004	-128.190	-24.695
C(material)[T.Linen]	-8.2863	22.661	-0.366	0.715	-52.703	36.130
C(material)[T.Linen Blend]	24.9149	27.923	0.892	0.372	-29.817	79.646
C(material)[T.Polyester]	-38.1971	22.576	-1.692	0.091	-82.448	6.054
C(material)[T.Satin]	-20.4443	56.621	-0.361	0.718	-131.425	90.537
C(material)[T.Silk]	-106.3235	98.535	-1.079	0.281	-299.460	86.813
C(material)[T.Viscose]	-29.1062	26.959	-1.080	0.280	-81.948	23.735
C(material)[T.Wool]	-6.5006	21.726	-0.299	0.765	-49.084	36.083
C(material)[T.Wool Blend]	-22.7634	22.037	-1.033	0.302	-65.957	20.431
price	-4.3508	0.415	-10.483	0.000	-5.164	-3.537
price:C(material)[T.Cotton]	0.0741	0.457	0.162	0.871	-0.822	0.971
price:C(material)[T.Denim]	0.5677	0.562	1.009	0.313	-0.535	1.670
price:C(material)[T.Linen]	-0.0031	0.474	-0.007	0.995	-0.932	0.925
price:C(material)[T.Linen Blend]	-0.4345	0.585	-0.742	0.458	-1.582	0.713
price:C(material)[T.Polyester]	-0.0880	0.477	-0.184	0.854	-1.024	0.848
price:C(material)[T.Satin]	-0.9060	1.303	-0.696	0.487	-3.459	1.647
price:C(material)[T.Silk]	1.4150	2.011	0.704	0.482	-2.526	5.356
price:C(material)[T.Viscose]	-0.1559	0.572	-0.272	0.785	-1.277	0.966
price:C(material)[T.Wool]	-0.0416	0.456	-0.091	0.927	-0.935	0.852
price:C(material)[T.Wool Blend]	0.2046	0.466	0.439	0.660	-0.708	1.118

Omnibus:	8045.236	Durbin-Watson:	2.012
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1627.271
Skew:	0.435	Prob(JB):	0.00
Kurtosis:	1.917	Cond. No.	2.61e+03

- $R^2 \approx 0.12$: price and material explain about 12% of sales variation
- At $\alpha = 0.05$, Denim have significantly lower sales than acrylic at the same price
- $1 increase in price → 4 fewer units sold on average
- Material has some effect, but most sales variation comes from other factors

Taking Promotion Status into Account




Scatter Plot: Sales Volume vs Price
Colored by Promotion status
Promotion = No
Promotion = Yes
Sales Volume (units sold)
Price (USD per item)
2000
1800
1600
1400
1200
1000
800
600
20
40
60
80
100
120
140

OLS Regression Results

Dep. Variable:	Q('Sales Volume')	R-squared:	0.849
Model:	OLS	Adj. R-squared:	0.849
Method:	Least Squares	F-statistic:	8747.
Date:	Wed, 03 Dec 2025	Prob (F-statistic):	0.00
Time:	15:07:37	Log-Likelihood:	-1.2499e+05
No. Observations:	20252	AIC:	2.500e+05
Df Residuals:	20238	BIC:	2.501e+05
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	992.3758	4.411	224.985	0.000	983.730	1001.021
C(Promotion)[T.Yes]	577.4662	3.435	168.117	0.000	570.734	584.199
C(material)[T.Cotton]	-26.2664	4.331	-6.065	0.000	-34.755	-17.778
C(material)[T.Denim]	-43.3458	5.325	-8.139	0.000	-53.784	-32.908
C(material)[T.Linen]	-5.6199	4.527	-1.241	0.214	-14.493	3.253
C(material)[T.Linen Blend]	2.8300	5.650	0.501	0.616	-8.245	13.905
C(material)[T.Polyester]	-34.1600	4.484	-7.618	0.000	-42.950	-25.370
C(material)[T.Satin]	-41.9240	10.824	-3.873	0.000	-63.139	-20.709
C(material)[T.Silk]	-58.5511	19.213	-3.048	0.002	-96.210	-20.893
C(material)[T.Viscose]	-43.8028	5.371	-8.156	0.000	-54.330	-33.275
C(material)[T.Wool]	1.4898	4.336	0.344	0.731	-7.009	9.988
C(material)[T.Wool Blend]	-7.7918	4.388	-1.776	0.076	-16.392	0.808
price	-2.3293	0.043	-53.831	0.000	-2.414	-2.244
price:C(Promotion)[T.Yes]	-1.3793	0.074	-18.533	0.000	-1.525	-1.233

Omnibus:	431.518	Durbin-Watson:	1.992
Prob(Omnibus):	0.000	Jarque-Bera (JB):	476.500
Skew:	0.334	Prob(JB):	3.38e-104
Kurtosis:	3.343	Cond. No.	1.23e+03

- Promotion status dramatically affects sales and price
- Low p-value
- R^2 = 0.849 shows strong model fit
- Material type still shows low significance

Additional Confounding Variable: Product Position

t-Test: Two-Sample Assuming Unequal Variances

	Aisle	*End-cap*
Mean	1090.4274	1100.45973
Variance	87639.7289	89998.0817
Observations	7810	6791
Hypothesized Mean D	0	
df	14264	
t Stat	-2.0279109	
P(T<=t) one-tail	0.02129394	
t Critical one-tail	1.64486046	
P(T<=t) two-tail	0.04258789	
t Critical two-tail	1.96013031	

t-Test: Two-Sample Assuming Unequal Variances

	Aisle	*Front of Store*
Mean	1090.4274	1103.361175
Variance	87639.7289	89396.96387
Observations	7810	5651
Hypothesized Mean	0	
df	12102	
t Stat	-2.4872115	
P(T<=t) one-tail	0.00644412	
t Critical one-tail	1.64497955	
P(T<=t) two-tail	0.01288825	
t Critical two-tail	1.96016003	

t-Test: Two-Sample Assuming Unequal Variances

	End-cap	*Front of Store*
Mean	1100.45973	1103.36118
Variance	89998.0817	89396.9639
Observations	6791	5651
Hypothesized Mean	0	
df	12047	
t Stat	-0.5381159	
P(T<=t) one-tail	0.2952535	
t Critical one-tail	1.64498912	
P(T<=t) two-tail	0.590507	
t Critical two-tail	1.96016092	

THANKS

Do you have any questions?

Team C7
Dream Team