IN-CLASS ASSIGNMENT

From Linear Regression to Regularization

BrewRight Coffee Co. — Store Performance Analysis

Duration: 90 min	Tool: Python	Total:	Dataset: brewright_stores.csv

Name: ______________________________________ Date: _________________

Business Context

You have been hired as a data analyst at BrewRight Coffee Co., a growing coffee chain with 150 store locations across the United States. The VP of Strategy has asked you to analyze what drives store-level monthly revenue so the company can make smarter decisions about new store locations, marketing budgets, and operations.

You have a dataset (brewright_stores.csv) with 150 stores and the following columns:

Column	Description
monthly_revenue_K	Monthly revenue in $K (TARGET variable)
marketing_spend_K	Local monthly marketing spend in $K
store_sqft	Store size in square feet
avg_daily_foot_traffic	Average daily walk-in customers
num_employees	Number of employees at the store
neighborhood_median_income_K	Median household income of area ($K)
drive_through	1 = has drive-through, 0 = no
competitor_count	Number of competing coffee shops within 1 mile
yelp_rating	Store's Yelp rating (2.5–5.0)
avg_latte_price	Average latte price at this store ($)
parking_spots	Number of dedicated parking spots
num_menu_items	Total items on the menu
seating_capacity	Indoor seating capacity
wifi_speed_mbps	WiFi speed in Mbps
distance_to_nearest_atm_miles	Distance to nearest ATM (miles)
avg_barista_experience_months	Avg months of barista experience
loyalty_program	1 = store has active loyalty program, 0 = no

Your analysis will proceed in five phases — each building on the last.

Part A: Simple Linear Regression (20 points)

The marketing team believes that local marketing spend is the single biggest driver of revenue. Let's test this claim with a simple linear model.

Q1. Load the dataset and create a scatter plot of monthly_revenue_K (y-axis) vs marketing_spend_K (x-axis). Does the relationship appear linear? Describe what you see.

[5 points]

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression

df = pd.read_csv('brewright_stores.csv')
# Your code here...
```

Q2. Fit a Simple Linear Regression: monthly_revenue_K ~ marketing_spend_K.

[8 points]

(a) Report the intercept and slope coefficient.

(b) Interpret the slope in business language: 'For every additional $1,000 in marketing spend, revenue changes by...'

(c) Report the R^2 value. What percentage of revenue variation does marketing spend alone explain?

Q3. If the VP proposes spending $15K/month on marketing at a new store, what revenue does your simple model predict? Is this estimate reliable? Why or why not?

[7 points]

Part B: Multiple Linear Regression (30 points)

The VP now wants to know: 'Marketing can't be the only factor. What else matters?' Time to bring in more predictors.

Q4. Fit a Multiple Linear Regression using these 5 features: marketing_spend_K, store_sqft, avg_daily_foot_traffic, num_employees, competitor_count.

[10 points]

```
features = ['marketing_spend_K', 'store_sqft', 'avg_daily_foot_traffic',
           'num_employees', 'competitor_count']
# Your code here...
```

(a) Report the R^2 value. How much did it improve over the simple model?

(b) List each coefficient and its sign (+/−). Do the signs make business sense? Explain for at least two features.

Q5. Now fit an MLR with ALL 16 predictor columns (everything except store_id and monthly_revenue_K).

[10 points]

(a) What is the R^2 on the training data? Did adding more features improve it?

(b) Split the data 80/20 using train_test_split (random_state=42). Report the R^2 on the TEST set. Compare it to training R^2. What do you observe?

```
from sklearn.model_selection import train_test_split
# Your code here...
```

Q6. Look at the coefficients from Q5's full model. Do any look suspiciously large or have unexpected signs? What business concern does this raise?

[5 points]

Q7. Calculate the Variance Inflation Factor (VIF) for all 16 features. Which features have VIF > 5? What does this mean in plain English?

[5 points]

```
from statsmodels.stats.outliers_influence import variance_inflation_factor
# Your code here...
```

Part C: Regularization — Ridge & Lasso (35 points)

The full model has issues — possible overfitting and multicollinearity. Let's see if regularization can help.

Important: Standardize all features before applying regularization. Use StandardScaler.

```
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import Ridge, Lasso, RidgeCV, LassoCV

scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

Q8. Fit a Ridge Regression using RidgeCV with alphas = [0.01, 0.1, 1, 10, 100, 1000]. Report:

[10 points]

(a) The best alpha chosen by cross-validation.

(b) The test R^2 score. How does it compare to the plain OLS test R^2 from Q5b?

(c) List all 16 coefficients. Did Ridge shrink any to near-zero? Did it set any to exactly zero?

Q9. Fit a Lasso Regression using LassoCV. Report:

[10 points]

(a) The best alpha chosen by cross-validation.

(b) The test R^2 score.

(c) Which features did Lasso eliminate (set coefficient to exactly zero)? List them.

(d) Which features survived? Do these make intuitive business sense as revenue drivers?

Q10. Create a comparison table or bar chart showing the coefficients from OLS, Ridge, and Lasso side by side. Briefly describe the key pattern you see.

[8 points]

```
# Your code here (pd.DataFrame or plt.barh)...
```

Q11. Complete this summary table:

[7 points]

Metric	OLS (all features)	Ridge	Lasso
Train R^2			
Test R^2			
# of non-zero coefficients			
Best alpha (λ)			

Part D: Business Recommendation (15 points)

Now put on your MBA hat. The VP of Strategy is in the room.

Q12. Based on Lasso's results, what are the TOP 3 actionable recommendations you would give the VP for improving store revenue? Be specific — tie each recommendation to a coefficient.

[8 points]

Q13. The VP asks: 'Should we invest in faster WiFi and add a loyalty program at all stores?' Using your Lasso results, what is your data-driven answer?

[4 points]

Q14. If you had to pick ONE model (OLS, Ridge, or Lasso) to deploy in production for predicting revenue at potential new store locations, which would you choose and why?

[3 points]

Part E: What-If Scenarios (30 points)

The VP loved your analysis. Now the leadership team has follow-up questions. Each scenario below describes a real business situation — use your models and your judgment to reason through them.

SCENARIO: The Penalty Dial

Your colleague says: 'Why bother choosing lambda with cross-validation? Just set it to something really large like 1,000,000 to fully regularize, or really small like 0.00001 to barely regularize.'

Q15. What happens to your model's coefficients and predictions in each of these extreme cases? Test it with code.

[6 points]

(a) Fit a Lasso with alpha = 0.00001. How many features survive? What does the test R^2 look like? What model is this essentially equivalent to?

(b) Fit a Lasso with alpha = 1000. How many features survive? What does the test R^2 look like? What has this model effectively become?

(c) In one sentence, explain why cross-validation finds the 'sweet spot' between these two extremes.

SCENARIO: Expansion to College Towns

BrewRight is planning to expand into 20 college towns. These locations are very different from the current store mix: foot traffic is extremely high (800–1200/day), median income is low ($25K–$35K), and nearly all stores would be small (900–1100 sqft) with no drive-through. The VP asks: 'Can we just use our model to predict revenue for these new stores?'

Q16. Identify at least two specific reasons why your current model might give unreliable predictions for these college-town stores. Hint: look at the ranges in your training data.

[5 points]

Q17. If you were forced to give a prediction anyway, would you have more confidence in a confidence interval or a prediction interval for these stores? Which one is more appropriate here and why?

[3 points]

SCENARIO: The Interaction Effect

A regional manager argues: 'Marketing spend works differently depending on whether a store has a drive-through. $10K in marketing at a drive-through store has way more impact than at a sit-down-only store.' In other words, she believes there is an interaction effect between marketing_spend_K and drive_through.

Q18. Create a new feature: interaction_mkt_dt = marketing_spend_K × drive_through. Add it to your feature set and refit the Lasso model.

[5 points]

```
# Your code here...
```

(a) Does Lasso keep or eliminate this interaction term?

(b) Did the test R^2 improve? What does this tell you about the regional manager's theory?

SCENARIO: The Unstable Selection

Your colleague re-runs the exact same Lasso model but with a different random_state in train_test_split (try random_state=99 instead of 42). She notices that some features that were zeroed out before are now non-zero, and vice versa.

Q19. Re-run your Lasso pipeline with random_state=99. Compare which features are kept vs. dropped to your original run (random_state=42).

[5 points]

```
# Your code here...
```

(a) Which features 'flip' (kept in one run, dropped in the other)?

(b) Look at the VIF results from Q7. Is there a connection between the features that flip and their VIF values? Explain.

(c) If feature selection stability is critical for a business decision, would you recommend Lasso, Ridge, or Elastic Net? Why?

SCENARIO: The Budget Cut

The CFO announces a cost-cutting initiative: BrewRight can only afford to collect and maintain data on 5 features going forward (instead of 16). You need to choose which 5 to keep.

Q20. Using insights from your Lasso model, which 5 features would you recommend keeping? Justify each choice with both its coefficient magnitude and its business relevance. Then refit an OLS model with only those 5 features and report the test R^2.

[6 points]

```
# Your code here...
```

(a) Your 5 chosen features and justification:

(b) Test R^2 with only these 5 features: ___________

(c) How much test R^2 did you lose compared to the full Lasso model? Is the tradeoff worth it from a business perspective?

SCENARIO: What If We Had More Data?

The VP of Analytics says: 'We're about to onboard 300 more franchise stores into our data system. Once we have 450 stores instead of 150, how will that change things?'

Q21. Without running any code, reason through the following:

[5 points]

(a) Would you expect the gap between OLS training R^2 and test R^2 to increase, decrease, or stay the same? Why?

(b) Would Lasso likely eliminate more features, fewer features, or roughly the same number? Why?

(c) Would the confidence intervals for the mean response get wider or narrower? What about prediction intervals?

— End of Assignment —