RESTAURANT RISK ANALYSIS: TECHNICAL REPORT

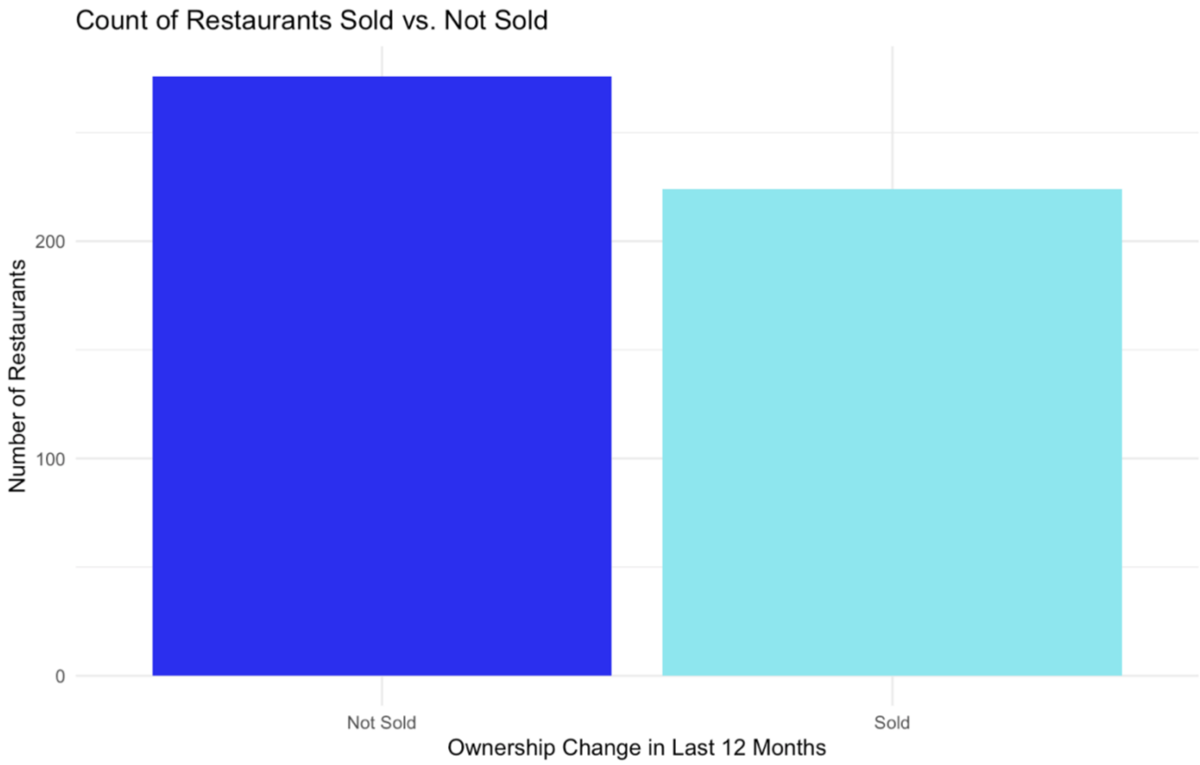
Coert, Miss Madeline

Table of Contents

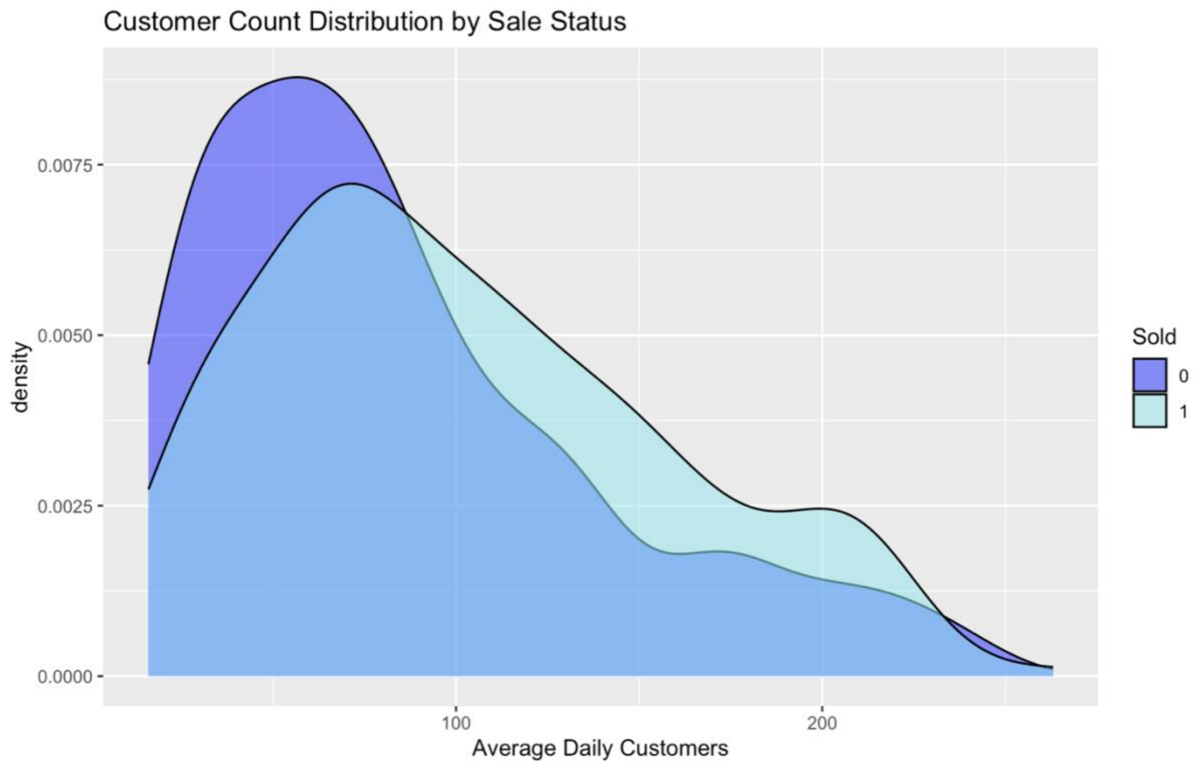
Introduction

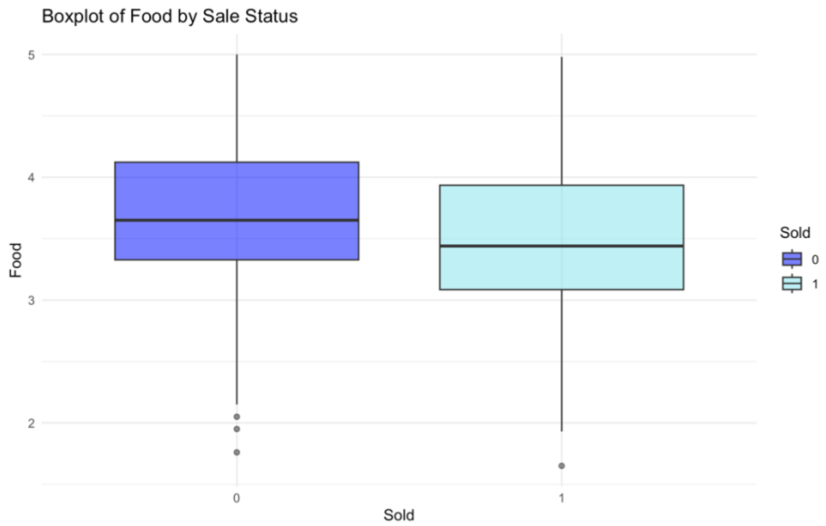
Franchise operators need timely insights backed by data to identify thriving restaurants and those at risk. This analysis addresses a key question: What are the warning signs of an underperforming restaurant at risk of closure*?* Sales data can reflect both distress and success therefore, misinterpreting it increases the risk of poor decision making. This report aimed to identify characteristics linked to customer volume and ownership change and whether time series patterns can signal decline early enough for franchisors to intervene and improve performance.

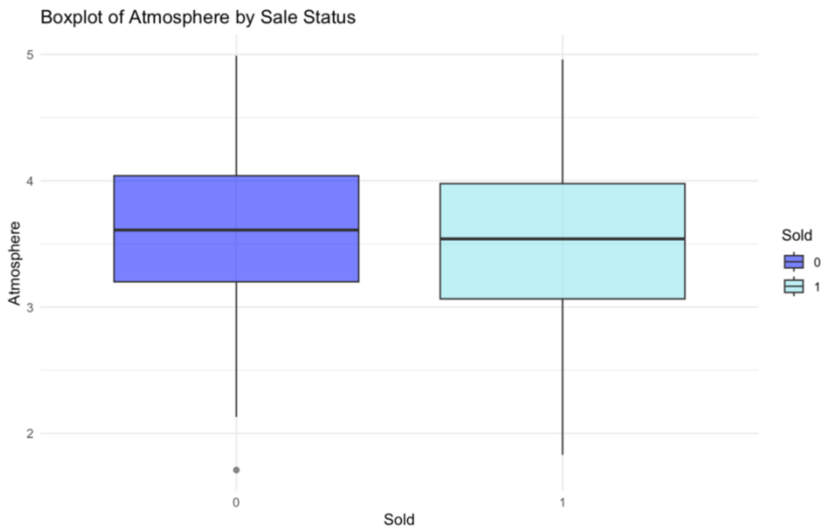
Exploratory Data Analysis

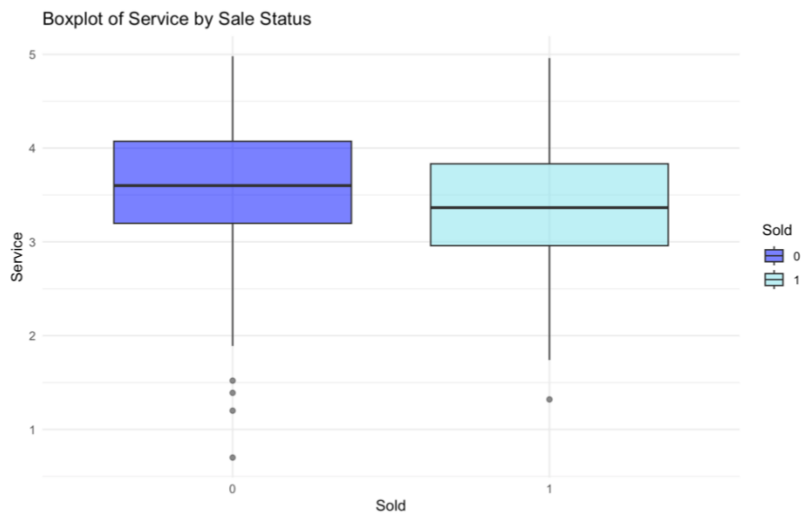


The sample is uneven with 52 more unsold restaurants, which may affect the reliability of insights.

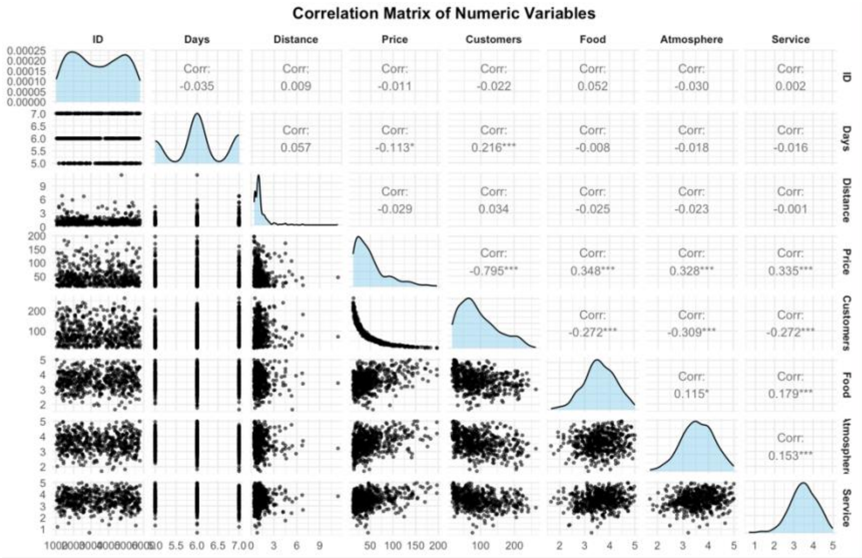




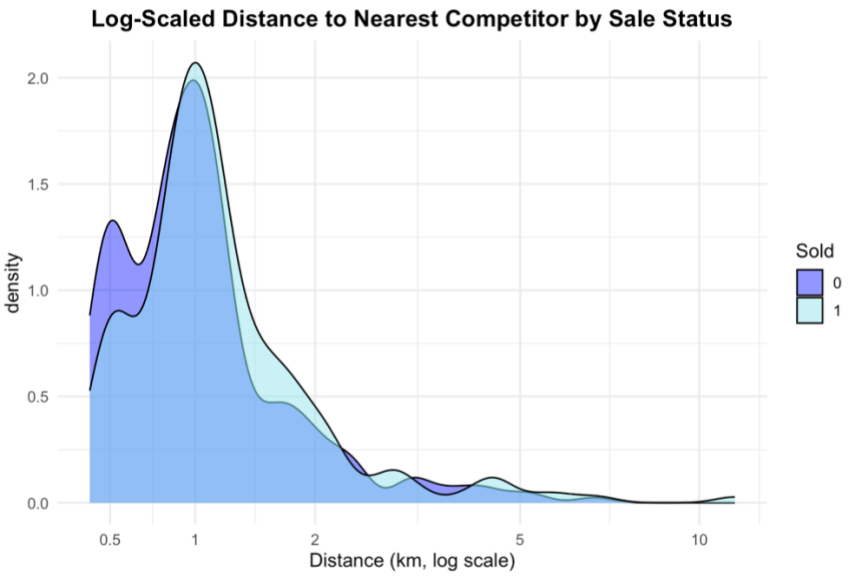




The density plot identified that sold restaurants have more daily customers. Box plots revealed slightly higher customer ratings among unsold venues, suggesting a weak link between experience and sale likelihood. However, greater variability implies inconsistent quality.



Proximity to competitors had no significant effect, indicating that customer volume is likely driven by quality or brand loyalty. Higher customer numbers were linked to lower food, service and atmosphere ratings which is possibly due to operational strain or increased scrutiny. These trends may highlight pressure points for layout or service redesign.



A log transformation addressed skew in competitor distance, revealing that most restaurants lie within 2km of a rival. However, distance showed little difference between sold and unsold restaurants thus, justifying its exclusion from the final model. Therefore, the refined business question became: Can we identify early signals that a restaurant is under strain and distinguish between distress-driven exits and strategic sales?

Multiple Linear Regression

Hypothesis:

It was hypothesised that restaurants with **higher food** and **service ratings**, **city locations**, and **lower prices** would attract **more customers**.	
Null Hypothesis (H_0)	None of the independent variables significantly predict customer count.
Alternative Hypothesis (H_1)	At least one independent variable significantly predicts customer count.

Model Specification:

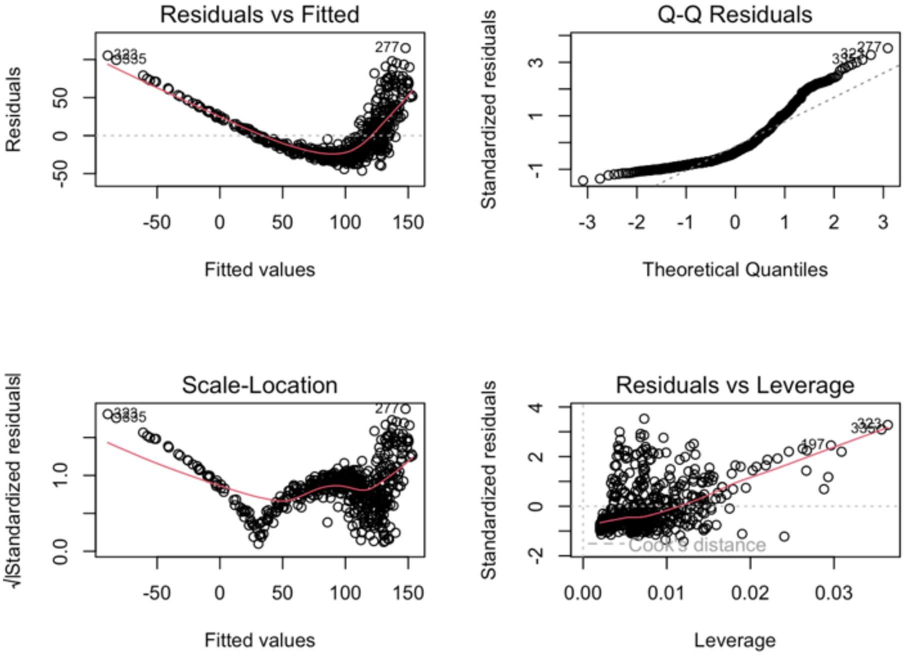
```
model_1 <- lm(Customers ~ Price + Food + Atmosphere + Service + Distance + Days +
Location + Sunday + Monday, data = restaurants)
```

Menu was excluded due to complexity and Sold was unsuitable for predicting customer count. The model explained 65% of variance ($R^2 = 0.653$, $p=0.001$) and Monday was removed due to collinearity. Price, Atmosphere and Days were significant where higher prices reduced customer numbers, more days increased them and better atmosphere slightly reduced volume. AIC comparison (4922.88 vs null model's 5435.76) confirmed an improved model fit.

Stepwise Regression:

Stepwise regression retained Price, Atmosphere and Days as key predictors. All remained statistically significant and the simplified model explained 65% of the variance (Adjusted $R^2 = 0.65$). This reinforced price and availability as key drivers and outlined that better ambience may not increase customers.

Model Assumptions:



Diagnostic checks revealed mild violations of linearity and homoscedasticity as indicated by curved and U-shaped patterns. The Q-Q plot revealed small deviations from normality in the tails, which is expected in larger samples. Multicollinearity was not a concern, as all VIF values were low. Additionally, no influential outliers were detected based on Cook's Distance.

Log-transformed Model:

To address assumption violations, a log-transformed model and a polynomial model were tested. The log-transformed model gave the best fit (Adjusted $R^2 = 0.894$) and had a very low residual error. Price remained highly significant ($p < 0.001$), with each $1 increase linked to a 1.6% decrease in customer numbers. Days had a positive effect with a 9.8% increase per day and Atmosphere was slightly significant ($p = 0.079$). The polynomial model also performed well (Adjusted $R^2 = 0.870$), highlighting a curved effect of price, suggesting that mid-range pricing is optimal. However, the log-transformed model was retained for its stronger fit and clearer assumptions.

Logistic and Multiple Logistic Regression

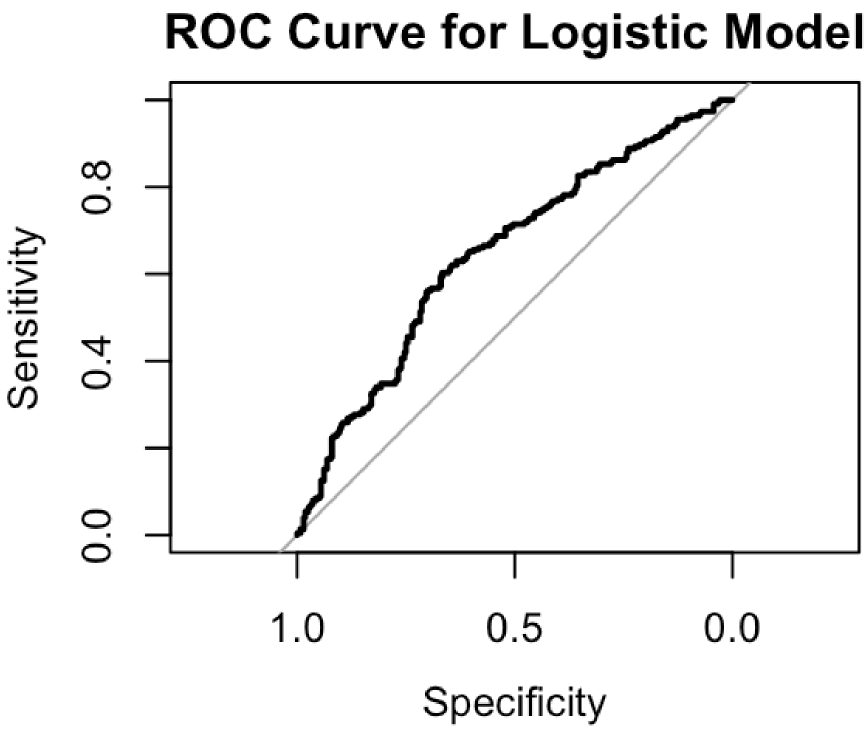
In the initial model, only food quality and suburban location were significant predictors, with lower ratings and suburban sites linked to higher sale odds. Most variables were not significant hence, the model was simplified. In the refined version, food and location remained significant and price also emerged as a key factor. The model was then evaluated for predictive performance:

Table 1: Logistic Regression Results:

Metric	Value	Interpretation
Accuracy	0.618	61.8% of predictions were correct overall
Sensitivity (Recall)	0.721	72.1% of the not sold restaurants were correctly predicted
Specificity	0.491	Only 49.1% of the sold restaurants were correctly predicted
Kappa	0.215	Weak to moderate agreement beyond chance
Balanced Accuracy	0.606	Average of sensitivity and specificity, shows some predictive value
P-value	0.0016	Model performs significantly better than guessing
Mcnemar's Test P-value	0.009	Asymmetry in errors

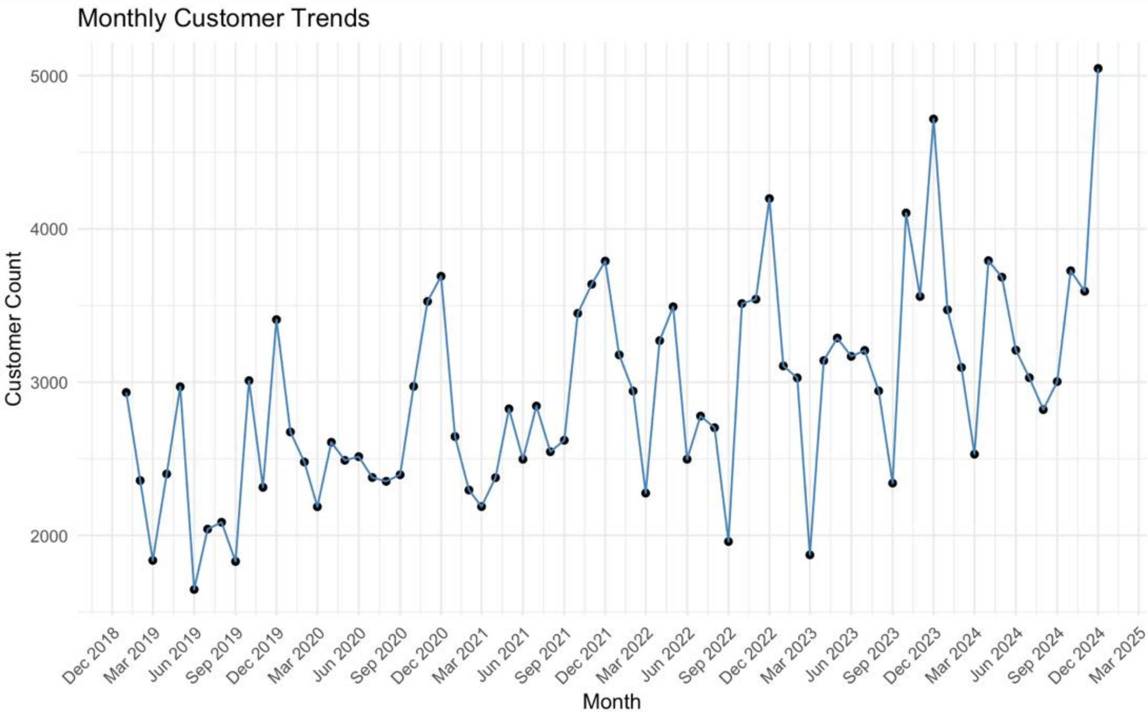
The final logistic regression model achieved 61.8% accuracy and outperformed random guessing ($p = 0.0016$). It displayed high sensitivity (72%) in identifying restaurants that were not sold but lower specificity (49%) for those that were. This imbalance suggests that sales may result from factors beyond operational data. Low food quality, lower prices and suburban location predicted sales which are useful warning signals.

ROC Curve and AUC:

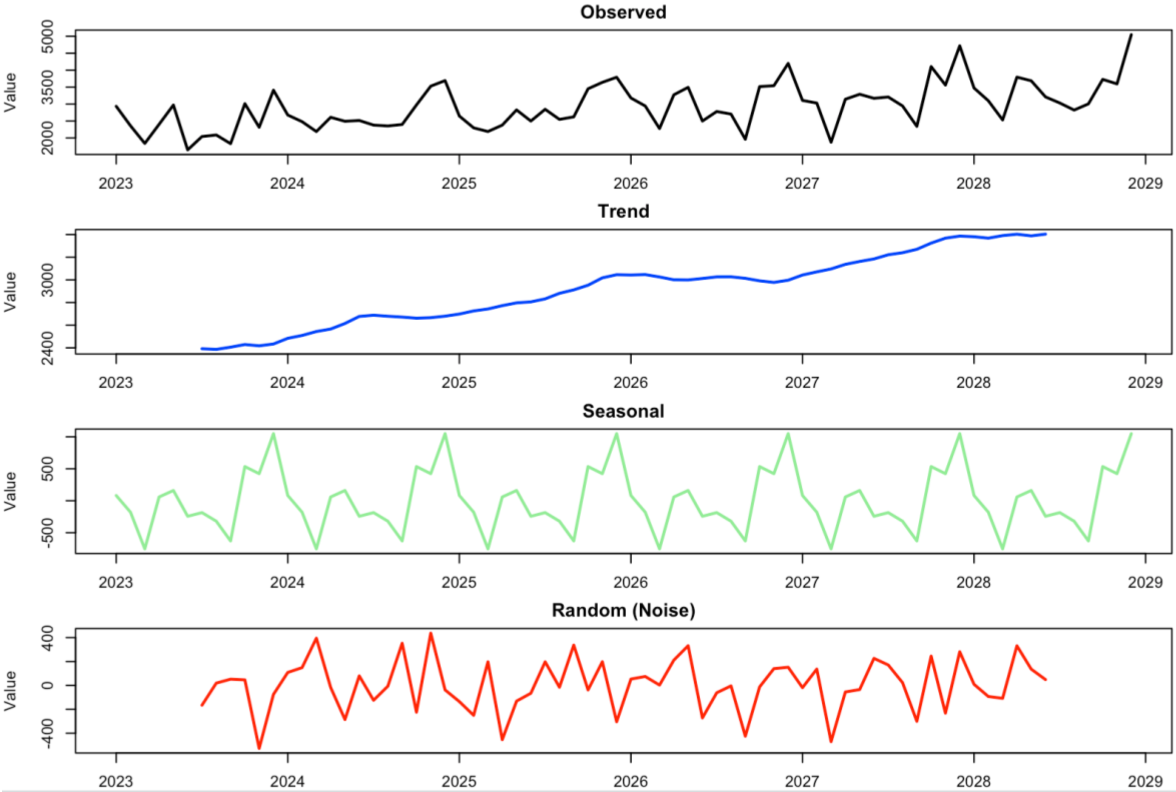


The final logistic model had modest predictive power (AUC = 0.646) but still outperformed random guessing. While not highly accurate the model performs meaningfully better than guessing and can help flag restaurants at an increased risk of sale based on operational characteristics. Starting with the single-variable model Price, adding Food and Location improved fit and reduced AIC. Interaction terms were tested but did not significantly enhance performance thus, the simpler model was retained.

Time Series



This time series reflects a different restaurant that has been sold. The analysis highlighted a strong upward trend in monthly customer volume from late 2018 to early 2025, suggesting steady business growth. This could reflect improved offerings, expanded capacity or a growing customer base. Seasonal patterns are also clear, with predictable peaks and dips at similar times each year which are helpful for planning staffing, inventory and marketing. A few short-term declines interrupt the trend, possibly due to external disruptions (e.g., Covid lockdowns) and may serve as early warning signs if linked to operational changes. Overall, these trends do not reflect distress instead, they suggest that the restaurant is performing well. This reinforces the idea that not all ownership changes stem from underperformance as some may be strategic sales. While logistic regression identifies inactive risk factors like low food ratings and suburban location, the time series adds a dynamic view of performance over time.



The decomposition confirms that the restaurant has experienced sustained growth with predictable seasonal fluctuations**.** Importantly, there are no clear signs of customer decline, which supports the hypothesis that any potential ownership change would likely be strategic rather than due to distress**.**

Conclusion

In summary, the exploratory analysis outlined that sold restaurants often had more customers, lower food quality and greater variability in experience which pointed to possible performance issues. Atmosphere and service ratings declined as customer numbers rose, suggesting busier venues may sacrifice experience. Multiple regression confirmed higher prices were associated with fewer customers and highlighted availability as a strong driver. Logistic regression models outperformed chance identifying food quality, pricing and suburban location as predictors of sale likelihood. However, low specificity in identifying sold restaurants and a modest AUC, suggested sales may also stem from strategic decisions not captured in the data. The time series analysis supported this, revealing sustained customer growth with no signs of distress hence, reinforcing that not all sales indicate failure. Therefore, franchisors should monitor restaurants with low food ratings and high prices in suburban areas to help flag underperformance early. Venues with declining customer counts or deteriorating service must be prioritised for support. It is also recommended that future research is conducted as these models would benefit from richer data including profitability, staff turnover and customer complaints. Expanding time series across venues and incorporating forecasting could also improve proactive decision-making and performance management.

Appendix

AI Acknowledgement

AI Generative technologies were utilised in this analysis from the platform ChatGPT.

Prompts:
• Can you review my report and advise where I can cut down on words? • Should I pitch to franchise operators or to a marketing team? • How do I make my introduction and conclusion more succinct and persuasive? • How can I improve my refined business question to be more answerable with data? • Where in my code can I change the colours to blue and cadetblue2? • Where in my code can I change the colours of each aspect of the decomposition plot? • Can you see if I am missing any key points in the criteria within my report? • Should I add or change anything in my visualisations to be better suited with my report? • Are my visualisations suitable and cohesive for a technical report structure? • Can I make my ROC visualisation more interesting or is it better to keep it simple for the report? • How can I make my correlation matrix more spaced out and readable for my report? • Can you advise on how to make my multiple regression paragraphs more cohesive together? • Can you advise on what elements to add in my report to make it more of a narrative structure?

R Code:

```
#Read Data

Restaurants <- read.csv("Restaurants.csv")

Customers <- read.csv("Customers.csv")

#Installing packages - ggplot2 and dplyr

install.packages("ggplot2")

install.packages("dplyr")

#Loading packages

library(ggplot2)

library(dplyr)

#Data Structure

str(Restaurants)

str(Customers)

#Formatting Variables - converting chr to factors

Restaurants[c("Location", "Menu", "Sunday", "Monday")] <- lapply(Restaurants [c("Location", "Menu",
"Sunday", "Monday")], as.factor)

Customers[c("Month")] <- lapply(Customers [c("Month")], as.factor)

#Formatting Variables - converting int to num

Restaurants[c("ID", "Days", "Customers", "Sold")] <- lapply(Restaurants [c("ID", "Days", "Customers",
"Sold")], as.numeric)

Customers[c("Customers")] <- lapply(Customers [c("Customers")], as.numeric)

#Check for Missing Values

colSums(is.na(Restaurants))

colSums(is.na(Customers))

#EDA

summary(Restaurants)

summary(Customers)

#Distribution of Sold

table(Restaurants$Sold)

#Are sold restaurants getting more or less customers?

ggplot(Restaurants, aes(x = Customers, fill = as.factor(Sold))) + geom_density(alpha = 0.5) +
scale_fill_manual(values = c("0" = "blue", "1" = "cadetblue2")) + labs(title = "Customer Count Distribution by
Sale Status", x = "Average Daily Customers", fill = "Sold")

#Visualise Ratings Compare food, atmosphere, and service ratings across sold vs not sold

Restaurants$Sold <- as.factor(Restaurants$Sold)

str(Restaurants$Sold)

#Boxplot for Food
```

```r
ggplot(Restaurants, aes(x = Sold, y = Food, fill = Sold)) + geom_boxplot(alpha = 0.6) + labs(title = "Boxplot of Food by Sale Status", x = "Sold", y = "Food") +  scale_fill_manual(values = c("0" = "blue", "1" = "cadetblue2")) + theme_minimal()

#Boxplot for Atmosphere

ggplot(Restaurants, aes(x = Sold, y = Atmosphere, fill = Sold)) + geom_boxplot(alpha = 0.6) + labs(title = "Boxplot of Atmosphere by Sale Status", x = "Sold", y = "Atmosphere") + scale_fill_manual(values = c("0" = "blue", "1" = "cadetblue2")) + theme_minimal()

#Boxplot for Service

ggplot(Restaurants, aes(x = Sold, y = Service, fill = Sold)) + geom_boxplot(alpha = 0.6) + labs(title = "Boxplot of Service by Sale Status", x = "Sold", y = "Service") + scale_fill_manual(values = c("0" = "blue", "1" = "cadetblue2")) + theme_minimal()

#Correlation Plot

library(GGally)

Restaurants_numeric <- Restaurants %>% select_if(is.numeric)

ggpairs(Restaurants_numeric)

#Does being near a similar restaurant correlate with sale?

ggplot(Restaurants, aes(x = Distance, fill = as.factor(Sold))) + geom_density(alpha = 0.5) + labs(title = "Distance to Nearest Competitor by Sale Status", x = "Distance (km)", fill = "Sold")

#Any patterns by location or menu type?

ggplot(Restaurants, aes(x = Menu, fill = as.factor(Sold))) + geom_bar(position = "fill") + labs(title = "Menu Type Proportion Sold", y = "Proportion", fill = "Sold") + theme(axis.text.x = element_text(angle = 45, hjust = 1))

ggplot(Restaurants, aes(x = Location, fill = as.factor(Sold))) + geom_bar(position = "fill") + labs(title = "Location Proportion Sold", y = "Proportion", fill = "Sold")

#Time Series Customer Trends

ggplot(Customers, aes(x = Month, y = Customers, group = 1)) + geom_line(color = "darkred", size = 1.2) + geom_point(color = "black") + labs(title = "Monthly Customer Trends", x = "Month", y = "Customers") + theme_minimal()

                              #MULTIPLE LINEAR REGRESSION

#Initial Model

model_1 <- lm(Customers ~ Price + Food + Atmosphere + Service + Distance + Days + Location + Sunday + Monday, data = Restaurants)

summary(model_1)

#Fit the Null Model

model_null <- lm(Customers ~ 1, data = Restaurants)

AIC(model_null)

AIC(model_1)

#Stepwise Regression

install.packages("MASS")

library(MASS)

model_step <- step(model_1, direction = "both")
```

```
summary(model_step)

#Check Assumptions

par(mfrow = c(2, 2))

plot(model_step)

#Check Multicollinearity

install.packages("car")

library(car)

vif(model_step)

#Log Transformation

model_log <- lm(log(Customers) ~ Price + Atmosphere + Days, data = Restaurants)

summary(model_log)

#Polynomial Term

model_poly <- lm(Customers ~ Price + I(Price^2) + Atmosphere + Days, data = Restaurants)

summary(model_poly)

plot(model_poly)

#Check model fit

plot(model_log)

#Logistic Regression

model_2 <- glm(Sold ~ Price + Customers + Food + Atmosphere + Service + Distance + Location + Days +
Sunday + Monday, data = Restaurants, family = binomial)

summary(model_2)

#Refining the model

library(MASS)

step_model2 <- stepAIC(model_2, direction = "both")

summary(step_model2)

#Confusion Matrix

install.packages("caret")

library(caret)

pred_probs <- predict(step_model2, type = "response")

pred_class <- ifelse(pred_probs > 0.5, 1, 0)

confusionMatrix(as.factor(pred_class), as.factor(Restaurants$Sold))

#ROC Curve and AUC

install.packages("pROC")

library(pROC)

roc_obj <- roc(Restaurants$Sold, pred_probs)

plot(roc_obj, main = "ROC Curve for Logistic Model")
```

```
auc(roc_obj)

#Strengthening the model

model_interact <- glm(Sold ~ Price * Location + Food * Service + Food + Price + Location, data = Restaurants, family = binomial)

summary(model_interact)

#Comparing

AIC(step_model2, model_interact)

anova(step_model2, model_interact, test = "Chisq")

                                        #Time Series

#Visualise Data

install.packages("scales")

library(scales)

library(ggplot2)

Customers$Month <- as.Date(Customers$Month)

ggplot(Customers, aes(x = Month, y = Customers)) + geom_point(color = "black") +  geom_line(color = "steelblue") + labs(title = "Monthly Customer Trends", x = "Month", y = "Customer Count") + scale_x_date(date_breaks = "3 months", date_labels = "%b %Y") + theme_minimal() + theme(axis.text.x = element_text(angle = 45, hjust = 1))

#Convert to Time Series

ts_customers <- ts(Customers$Customers, start = c(2023, 1), frequency = 12)

#Decomposition

par(mfrow = c(4, 1), mar = c(2, 4, 2, 2))

plot(decomposed$x, col = "black", lwd = 2, main = "Observed", ylab = "Value")

plot(decomposed$trend, col = "blue", lwd = 2, main = "Trend", ylab = "Value")

plot(decomposed$seasonal, col = "lightgreen", lwd = 2, main = "Seasonal", ylab = "Value")

plot(decomposed$random, col = "red", lwd = 2, main = "Random (Noise)", ylab = "Value")
```

Analysis Outputs:

Multiple Regression:

```
model_1 <- lm(Customers ~ Price + Food + Atmosphere + Service + Distance + Days + Location + Sunday +
Monday, data = restaurants)

Call:

lm(formula = Customers ~ Price + Food + Atmosphere + Service +

    Distance + Days + Location + Sunday + Monday, data = Restaurants)

Residuals:

   Min     1Q Median     3Q    Max

-46.84 -24.89 -12.24  15.44 116.46

Coefficients: (1 not defined because of singularities)

               Estimate Std. Error t value Pr(>|t|)

(Intercept)    118.27142   21.58406   5.480 6.82e-08 ***

Price           -1.13833    0.04801 -23.708  < 2e-16 ***

Food            -0.08827    2.52796  -0.035  0.97216

Atmosphere      -5.03432    2.48137  -2.029  0.04301 *

Service         -0.81375    2.32210  -0.350  0.72616

Distance        -0.04380    1.58911  -0.028  0.97802

Days             8.15334    3.01515   2.704  0.00709 **

LocationSuburbs   2.03759    3.26250   0.625  0.53256

SundayYes        4.55028    4.49098   1.013  0.31146

MondayYes             NA         NA      NA       NA

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32.89 on 491 degrees of freedom

Multiple R-squared:  0.6528,         Adjusted R-squared:  0.6471

F-statistic: 115.4 on 8 and 491 DF,  p-value: < 2.2e-16
```

Stepwise Regression:

```
Call:
lm(formula = Customers ~ Price + Atmosphere + Days, data = Restaurants)


Residuals:
   Min     1Q Median     3Q    Max
-46.23 -24.36 -12.05  15.65 115.12


Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 107.70607   15.27996   7.049 6.08e-12 ***
Price        -1.15073    0.04262 -26.998  < 2e-16 ***
Atmosphere   -4.98064    2.46746  -2.019   0.0441 *
Days         10.12918    2.08516   4.858 1.59e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 32.77 on 496 degrees of freedom
Multiple R-squared:  0.6518,	Adjusted R-squared:  0.6497
F-statistic: 309.5 on 3 and 496 DF,  p-value: < 2.2e-16
```

Log transformation:

```
Call:
lm(formula = log(Customers) ~ Price + Atmosphere + Days, data = Restaurants)


Residuals:
     Min       1Q   Median       3Q      Max
-0.34257 -0.15853 -0.05831  0.12213  0.93571


Coefficients:
              Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  4.6677865  0.0983720  47.450  < 2e-16 ***
Price       -0.0162640  0.0002744 -59.270  < 2e-16 ***
Atmosphere  -0.0279224  0.0158855  -1.758   0.0794 .
Days         0.0983445  0.0134242   7.326 9.68e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2109 on 496 degrees of freedom
Multiple R-squared:  0.895,	Adjusted R-squared:  0.8944
F-statistic:  1410 on 3 and 496 DF,  p-value: < 2.2e-16
```

Polynomial Term:

```
Call:
lm(formula = Customers ~ Price + I(Price^2) + Atmosphere + Days,
    data = Restaurants)

Residuals:
   Min     1Q Median     3Q    Max
-76.71 -14.22  -4.45  12.51  86.17

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.686e+02  9.527e+00  17.696  < 2e-16 ***
Price       -3.436e+00  8.275e-02 -41.523  < 2e-16 ***
I(Price^2)   1.454e-02  5.001e-04  29.081  < 2e-16 ***
Atmosphere  -2.247e+00  1.504e+00  -1.494    0.136
Days         8.189e+00  1.270e+00   6.448  2.7e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.93 on 495 degrees of freedom
Multiple R-squared:  0.8714,	Adjusted R-squared:  0.8704
F-statistic: 838.8 on 4 and 495 DF,  p-value: < 2.2e-16
```

Logistic and Multiple logistic Regression:

```
Call:
glm(formula = Sold ~ Price + Customers + Food + Atmosphere +
    Service + Distance + Location + Days + Sunday + Monday, family = binomial,
    data = Restaurants)

Coefficients: (1 not defined because of singularities)
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     2.241810   1.414381   1.585   0.1130
Price          -0.004733   0.004816  -0.983   0.3258
Customers       0.002107   0.002910   0.724   0.4689
Food           -0.374270   0.160516  -2.332   0.0197 *
Atmosphere      0.036360   0.157095   0.231   0.8170
Service        -0.197134   0.147073  -1.340   0.1801
Distance        0.125861   0.104060   1.210   0.2265
LocationSuburbs  0.407664   0.207037   1.969   0.0489 *
Days           -0.146670   0.192938  -0.760   0.4471
SundayYes      -0.051536   0.284865  -0.181   0.8564
MondayYes             NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 687.73  on 499  degrees of freedom
Residual deviance: 652.41  on 490  degrees of freedom
AIC: 672.41

Number of Fisher Scoring iterations: 4
```

Refine the model:

```
Call:
glm(formula = Sold ~ Price + Food + Location, family = binomial,
    data = Restaurants)
```

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     1.290094   0.562661   2.293  0.02186 *
Price          -0.007764   0.002933  -2.647  0.00811 **
Food           -0.399527   0.159321  -2.508  0.01215 *
LocationSuburbs  0.512607   0.192583   2.662  0.00777 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 687.73  on 499  degrees of freedom
Residual deviance: 657.41  on 496  degrees of freedom
AIC: 665.41

Number of Fisher Scoring iterations: 4
```

Predict probabilities:

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 199 114
         1  77 110

               Accuracy : 0.618
                 95% CI : (0.5738, 0.6608)
    No Information Rate : 0.552
    P-Value [Acc > NIR] : 0.001649

                  Kappa : 0.2154

 Mcnemar's Test P-Value : 0.009191
```

```
          Sensitivity : 0.7210
          Specificity : 0.4911
       Pos Pred Value : 0.6358
       Neg Pred Value : 0.5882
           Prevalence : 0.5520
       Detection Rate : 0.3980
 Detection Prevalence : 0.6260
    Balanced Accuracy : 0.6060

     'Positive' Class : 0
```

Interaction terms:

```
Call:
glm(formula = Sold ~ Price * Location + Food * Service + Food +
    Price + Location, family = binomial, data = Restaurants)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           0.587949   2.864736   0.205   0.8374
Price                -0.005758   0.004363  -1.320   0.1869
LocationSuburbs       0.582505   0.336179   1.733   0.0831 .
Food                 -0.028455   0.806398  -0.035   0.9719
Service               0.160422   0.815098   0.197   0.8440
Price:LocationSuburbs -0.001250   0.005639  -0.222   0.8246
Food:Service         -0.103028   0.228453  -0.451   0.6520
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 687.73  on 499  degrees of freedom
Residual deviance: 655.23  on 493  degrees of freedom
AIC: 669.23
```

```
Number of Fisher Scoring iterations: 4
```

Comparing models:

```
              df      AIC
step_model2    4 665.4143
model_interact 7 669.2318


Analysis of Deviance Table


Model 1: Sold ~ Price + Food + Location
Model 2: Sold ~ Price * Location + Food * Service + Food + Price + Location
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       496     657.41
2       493     655.23  3   2.1825   0.5354
```