# Content

# 1. Introduction

This report aims to examine how fast-food meal prices relate to neighbourhood characteristics in New Jersey (NJ) and Pennsylvania (PA), using descriptive analysis, hypothesis testing and regression models, referring to the essay from Graddy (1997)

# 2. Data & Preprocessing

After the data cleaning, the dataset contains 369 fast-food outlets. The main outcome variable, meal_price_avg(average total meal price (entrée + fries + soda) across two survey waves). Define a high-price indicator (meal_price_avg ≥ 3.075) and construct a median income split. For the regression analysis, the original continuous measures of income and prpblck （proportion of Black residents） are retained to avoid loss of information.

Table 1. Descriptive statistics for meal_price_avg (N = 369)

| N | Mean | SD | Median |
|---|---|---|---|
| 369 | 3.32 | 0.62 | 3.08 |

*Note.* Mean and SD are in USD. The constructed price is the average of entrée, fries, and soda across two waves. No inferential claims are made.

Table 2. High-price (meal_price_avg ≥ 3.075) counts by state (NJ vs PA)

| | High-price (=1) | Total |
|---|---|---|
| NJ | 168 | 295 |
| PA | 17 | 74 |

*Note.* “High-price” is defined as meal_price_avg ≥ 3.075 (sample median). Counts are row totals by state; used for the two-proportion test.

# 3. Descriptive Analysis

This section summarises descriptive patterns in the data using charts matched to the variable types.

A bar chart is appropriate for category-wise means. The figure shows visible between-brand differences, motivating brand controls in regressions. KFC records the highest average meal prices, while Burger King is consistently the lowest. The wide gap across chains highlights the need to control for brand identity.


Figure 1. Average meal price by chain

A Pareto chart is appropriate for ranked frequencies and concentration. The sample is concentrated in a few brands, cautioning against unadjusted mean comparisons. A small number of brands make up most of the sample, with cumulative shares rising steeply. This imbalance indicates that raw price comparisons may be misleading unless brand mix is accounted for, reinforcing the use of chain dummies.

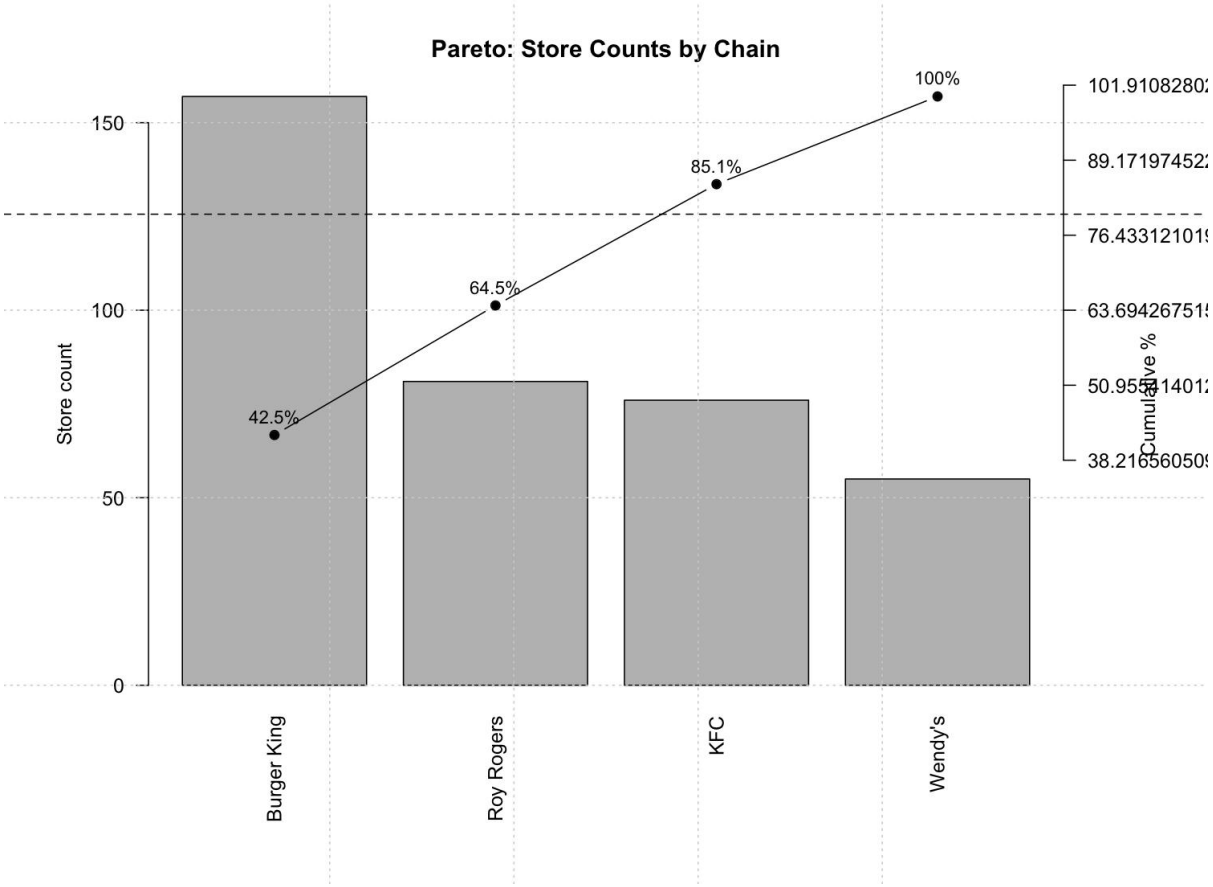


Figure 2. Pareto chart of store counts by chain

The pie chart is well-suited to show how each category contributes to the whole sample as proportions, which shows that NJ contributes more outlets than PA. Because state-level differences can influence prices, this imbalance motivates the NJ–PA proportion test in Section 4.2.

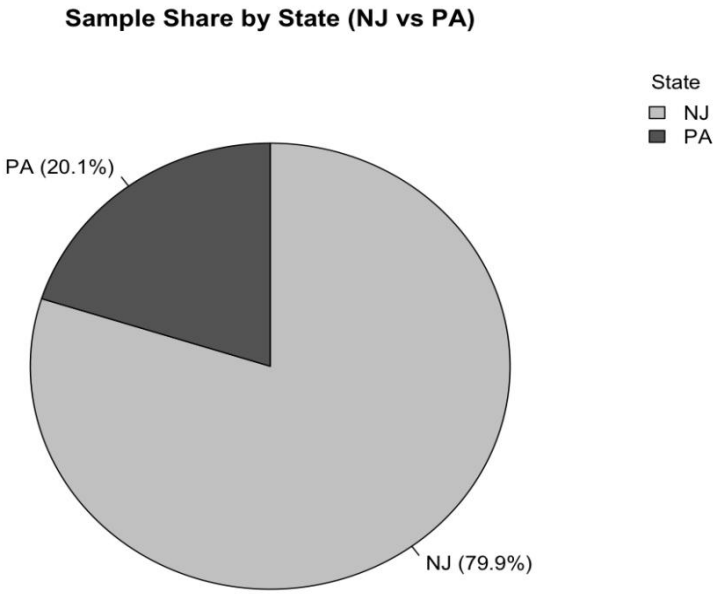


Figure 3. Sample share by state (NJ vs PA)

The histogram is appropriate for continuous distributions, which confirms a right-tailed price distribution, with most outlets clustered around the centre and a small group of higher-priced stores stretching the tail. Although meal_price_avg is right-skewed, the sample size makes the test robust to deviations from normality, so comparing means on the original price scale is still appropriate. At the same time, the right-skewed distribution motivates using log_meal_price in the regression models to reduce the influence of high-price outlets.

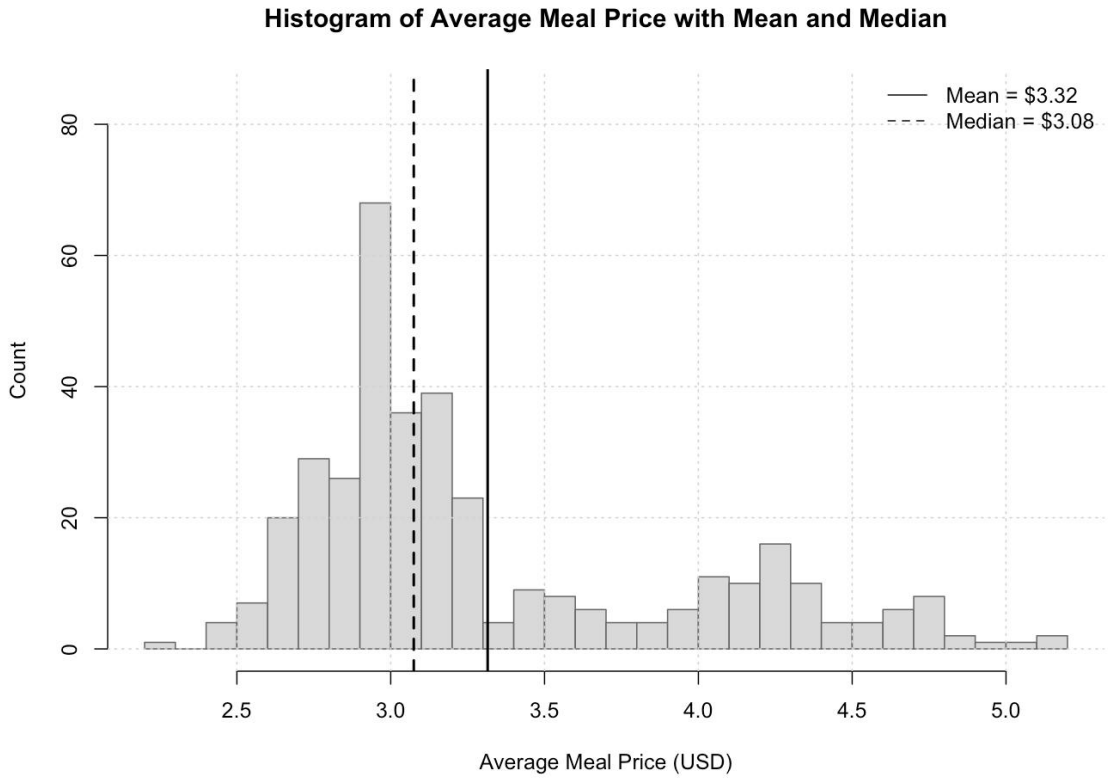


Figure 4. Histogram of Average Meal Price with Mean and Median

Boxplots summarise central tendency and dispersion for a continuous outcome across groups, making them a useful complement to the histogram, so we use Boxplots to compare meal_price_avg between high- and low-income areas using boxplots. The higher median and upper quartiles in the High-income group suggest higher meal prices.

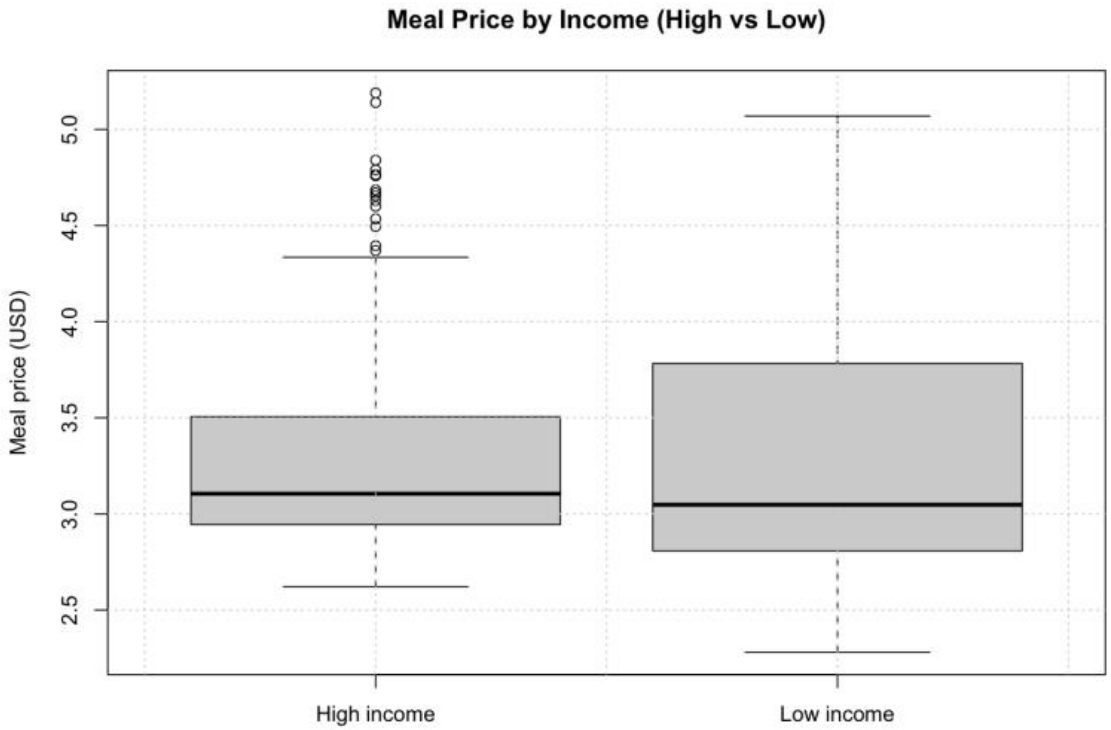


Figure 5. Meal price by income (median split)

# 4. Hypothesis Testing

We conduct the hypothesis tests using the raw average meal price. Although the distribution of meal_price_avg is mildly right-skewed, the Welch t-test is robust with our large sample ($n > 180$ per group), so using the untransformed variable is acceptable for mean comparisons.

## 4.1 Two-Sample t-Test

Income differences across neighbourhoods are widely used to test for potential price discrimination (Hastings, Hortacsu & Syverson, 2019). Using the sample median of neighbourhood income, we classify outlets into high- and low-income areas and apply a Welch t-test to compare mean meal prices, which is appropriate given the large sample size.

$$H_0: \mu_{\text{high}} = \mu_{\text{low}}$$

$$H_1: \mu_{\text{high}} \neq \mu_{\text{low}}$$

The mean meal price is 3.321 (n = 185) in high-income areas and 3.309 (n = 184) in low-income areas. The test yields $t = 0.194$ with $p = 0.846$. As $p > 0.05$, we fail to reject $H_0$ at the 5% significance level, so we find no statistically significant difference in average meal prices between high-income and low-income areas. Because the median split discards within-group income variation, this result should be interpreted as exploratory.

## 4.2 Proportion Test

We then assess whether the proportion of high-price outlets (meal_price_avg ≥ the sample median of 3.075) differs between NJ and PA using a two-sample proportion test.

$$H_0: p_{\text{NJ}} = p_{\text{PA}}$$

$$H_1: p_{\text{NJ}} \neq p_{\text{PA}}$$

In NJ, 56.95% outlets are high-priced, compared with 22.97% in PA. The proportion test yields $\chi^2(1) = 27.317$, $p < 0.001$, so we reject $H_0$ at the 1% significance level and conclude that high-price outlets are significantly more prevalent in NJ than in PA. This large difference suggests that state-level location is associated with larger price gaps than the modest income-based comparison and points to systematic cross-state differences in how fast-food chains set prices.

# 5. Regression Analysis

To inform the regression specification, we computed pairwise correlations among the key neighbourhood predictors. Our main demographic measure is prpblck. Income and other neighbourhood characteristics enter the models as control variables

Table 3. Correlation Matrix

| var | prpblck | lincome | prppov | lhseval | ldensity | crmrte | wagest_avg | emp_avg |
|---|---|---|---|---|---|---|---|---|
| prpblck | 1.000 | -0.511 | 0.697 | -0.336 | 0.404 | 0.598 | 0.020 | -0.073 |
| lincome | -0.511 | 1.000 | -0.843 | 0.805 | -0.318 | -0.470 | 0.353 | 0.072 |
| prppov | 0.697 | -0.843 | 1.000 | -0.565 | 0.389 | 0.665 | -0.205 | -0.118 |
| lhseval | -0.336 | 0.805 | -0.565 | 1.000 | -0.034 | -0.297 | 0.385 | 0.124 |
| ldensity | 0.404 | -0.318 | 0.389 | -0.034 | 1.000 | 0.307 | -0.041 | 0.041 |
| crmrte | 0.598 | -0.470 | 0.665 | -0.297 | 0.307 | 1.000 | -0.011 | -0.197 |
| wagest_avg | 0.020 | 0.353 | -0.205 | 0.385 | -0.041 | -0.011 | 1.000 | 0.025 |
| emp_avg | -0.073 | 0.072 | -0.118 | 0.124 | 0.041 | -0.197 | 0.025 | 1.000 |

*Note.* Pearson correlations among prpblck, lincome, prppov, lhseval, ldensity, crmrte, wagest_avg, emp_avg. Values are rounded to three decimals. Matrix is descriptive and used to screen collinearity.

## 5.1 Multicollinearity Test and Variable Selection (Model 1)

We first estimated a comprehensive model including racial, socio-economic, crime, labour-cost and brand/state variables. Variance Inflation Factors (VIFs) remained below the conventional threshold (≈10), indicating no serious multicollinearity. Variables with limited theoretical relevance and negligible statistical contribution—density, crmrte, wagest_avg and emp_avg—were therefore dropped to obtain a more parsimonious baseline specification

$$log\_meal\_price \sim prpblck + lincome + prppov + lhseval + BK + KFC + RR + NJ \quad \text{(Model 1)}$$

**Interpretation in Relation to the Research Question**

**Model fit**: Model 1 explains around 80% of the variation in log meal prices, indicating that neighbourhood and brand/state factors jointly matter for pricing.

**Racial composition**: Racial composition (prpblck) is positive and statistically significant, so outlets in areas with a higher share of Black residents charge higher prices, conditional on socio-economic, brand and state controls.

**Socio-economic indicators**: Income is negative, while poverty is negative and housing value is positive, suggesting distinct effects of different socio-economic indicators.

Brand and state effects: Brand and state dummies are also important: KFC is priced above, and Burger King below, Wendy's, and NJ outlets charge higher prices than those in PA.

**Overall implication**: Overall, racial composition, socio-economic conditions, brand identity and state location jointly shape fast-food pricing.

Table 4. Regression Results

| Predictor | Estimate | Std.Error | t.value | p.value |
|---|---|---|---|---|
| (Intercept) | 1.23862 | 0.32230 | 3.843 | 0.000144 |
| prpblck | 0.11057 | 0.03246 | 3.406 | 0.000734 |
| lincome | -0.14884 | 0.04032 | -3.692 | 0.000257 |
| prppov | -0.43581 | 0.14352 | -3.037 | 0.002566 |
| lhseval | 0.12400 | 0.01975 | 6.280 | 9.76e-10 |
| BK | -0.07132 | 0.01233 | -5.787 | 1.56e-08 |
| KFC | 0.33194 | 0.01392 | 23.849 | <2e-16 |
| RR | 0.01968 | 0.01382 | 1.423 | 0.155493 |
| NJ | 0.06110 | 0.01166 | 5.239 | 2.75e-07 |

*Note. OLS semi-log model (Model 1): outcome = log(meal_price). Predictors: prpblck, lincome, prppov, lhseval, brand dummies (BK, KFC, RR; Wendy's ref.), NJ (PA ref.). N = 369. Estimates with conventional SEs; t- and two-sided p-values shown.*

*Significance codes*: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

## 5.2 Stepwise Selection and Model Comparison (Model 2)

We used bidirectional stepwise regression with AIC as the selection criterion to assess whether any predictors in Model 1 could be removed to obtain a more concise model. The lowest AIC (−1873.5) was obtained by retaining all predictors; dropping RR, prppov, prpblck or lincome increased AIC and worsened model fit.

The stepwise procedure therefore returned exactly the same specification as Model 1: coefficients, standard errors and goodness-of-fit statistics (Adjusted $R^2 = 0.8025$) were unchanged. This indicates that the baseline model is already the most parsimonious AIC-optimal specification within the considered variable set and that each retained predictor contributes meaningfully when evaluated jointly.

## 5.3 Assumption Checks and Diagnostics

Residual diagnostics indicate:

**Linearity**: In the Residuals vs Fitted plot, the smoothing line is nearly horizontal with no strong systematic pattern, suggesting that a linear functional form is appropriate.

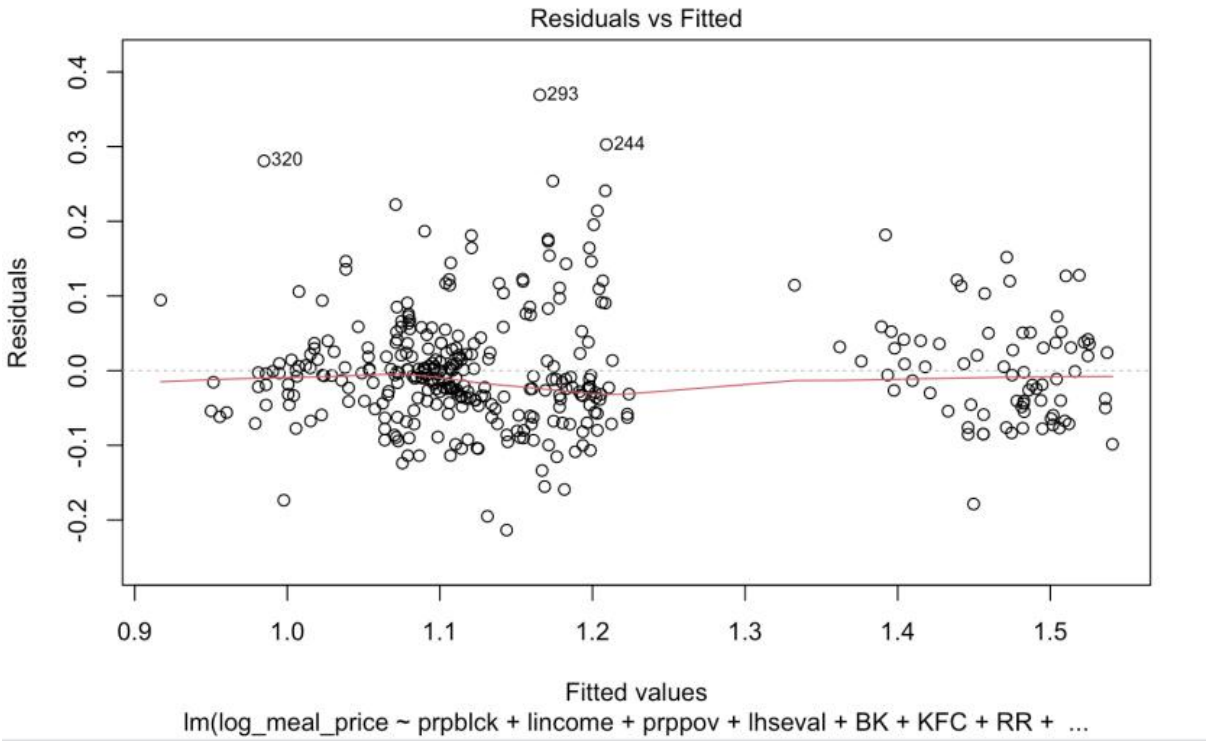


Figure 6. Residuals vs fitted values (Model 2 OLS, dependent variable log(meal_price))

**Normality**: The Q–Q plot shows some deviations in the tails, but residuals broadly follow the reference line, so departures from normality are unlikely to pose a substantial threat to inference given the sample size.

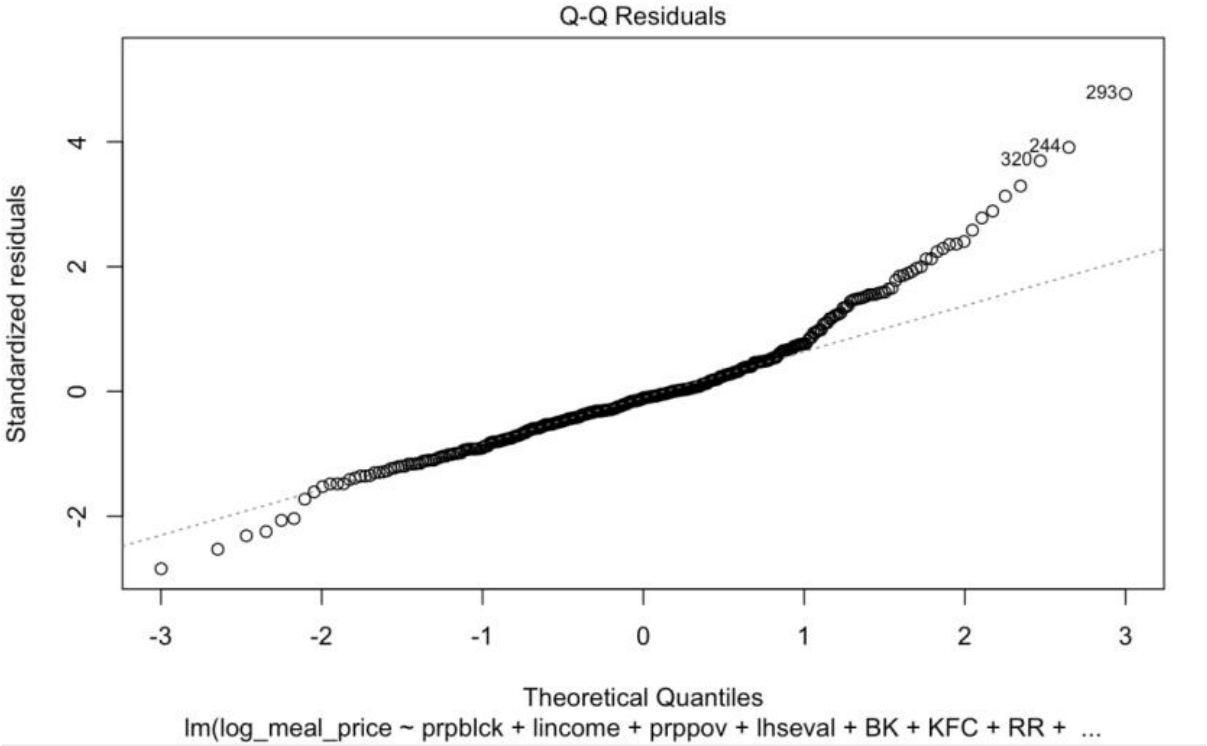


Figure 7. Normal Q–Q plot of standardised residuals (Model 2)

**Influential cases**: Cook's distance values are close to zero for all observations and remain below the common cut-off of 4/n, indicating that no single outlet unduly influences the regression estimates.

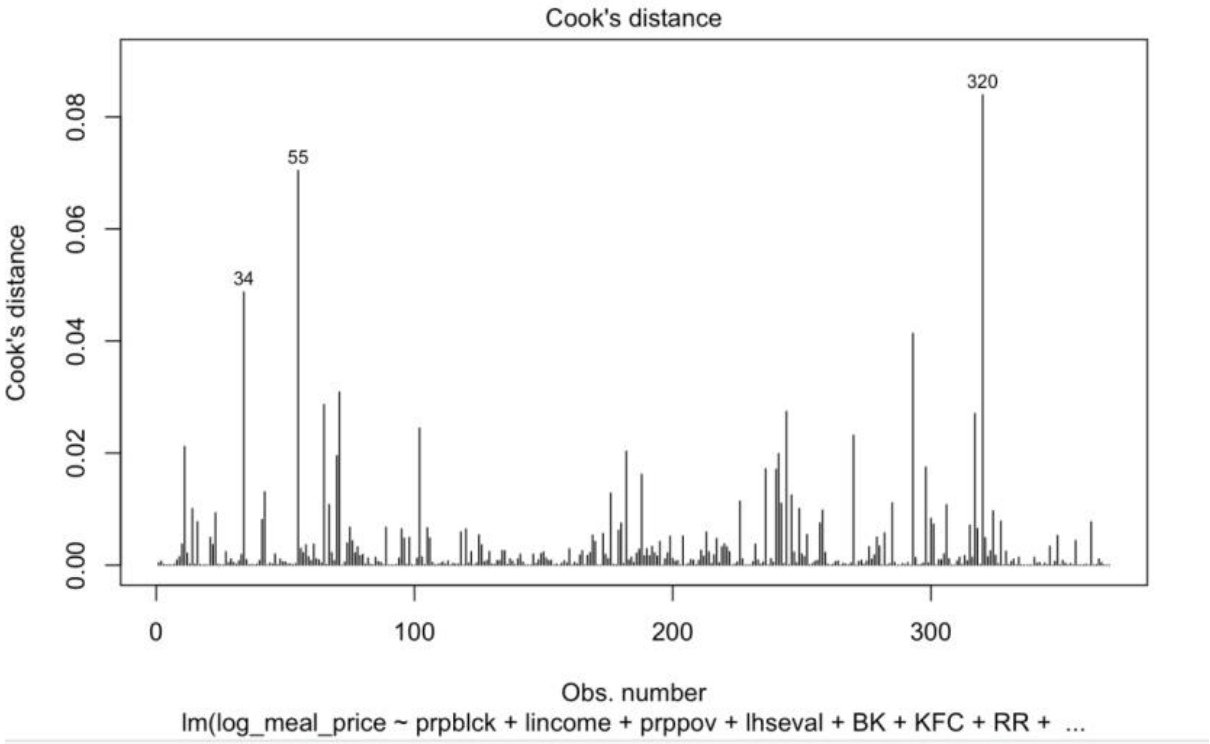


Figure 8. Cook's distance (Model 2)

**Homoscedasticity:** Residual spread is relatively stable across fitted values, with only a slight widening at higher fitted prices. However, the Breusch–Pagan test indicates statistically significant heteroscedasticity (BP = 34.10, $p < 0.001$), implying a violation of the constant-variance assumption, although the deviation appears modest.

## 5.4 Lasso and Ridge Regression

To assess whether regularisation improves on the OLS baseline, we estimated Ridge ($\alpha = 0$) and Lasso ($\alpha = 1$) models with the same predictors as Model 1, selecting $\lambda$ by 10-fold cross-validation.

### Ridge regression ($\alpha = 0$)

Ridge chose a small penalty ($\lambda = 0.0146$), shrinking all coefficients towards zero but not removing any predictors. Coefficients are smaller in magnitude than in OLS, yet RMSE is slightly higher, so the variance reduction does not translate into better predictive accuracy.

### Lasso regression ($\alpha = 1$)

Lasso selected an extremely small penalty ($\lambda = 0.00015$), yielding estimates that are almost identical to OLS. No coefficients were shrunk to zero, indicating little scope for variable selection, which is consistent with the relatively low multicollinearity and the theoretical relevance of all predictors.

Table 5. Model performance (10-fold CV RMSE)

| Model | RMSE_10fold_CV |
|---|---|
| OLS | 0.07731 |
| Ridge (alpha=0) | 0.08009 |
| Lasso (alpha=1) | 0.07966 |

*Note.* Same predictors as Model 1. RMSE is averaged over 10 folds. Lower values indicate better out-of-sample fit. Ridge ($\alpha = 0$) and Lasso ($\alpha = 1$) use CV-selected $\lambda$.

## 5.5 Bias–Variance Trade-off

Regularisation introduces bias by shrinking coefficients but reduces variance by penalising model complexity.

In this dataset, the bias introduced by Ridge and Lasso does not lead to improved predictive performance because the OLS model already exhibits low variance and does not overfit.

This explains why OLS achieves the best RMSE, while Ridge and Lasso offer no improvement.

# 6. Conclusion

This report used descriptive analysis, hypothesis testing and regression modelling to examine how fast-food prices vary across neighbourhoods in NJ and PA. Taken together, the results suggest that neighbourhood income on its own does not generate large or consistent price gaps, whereas operating in NJ is associated with a substantially higher prevalence of high-price outlets than operating in PA. In the regression models, prpblck remains a positive and robust predictor of log_meal_price after controlling for income, brand and other neighbourhood factors, indicating that both state location and neighbourhood racial composition are systematically associated with higher prices. These findings are based on observational data from two states and should therefore be interpreted as evidence of associations rather than causal effects.

# References

Graddy, K. (1997) 'Do Fast-Food Chains Price Discriminate on the Race and Income Characteristics of an Area?', *Journal of Business & Economic Statistics*, 15(4), pp. 391–401. doi: 10.1080/07350015.1997.10524717.

Hastings, J.S., Hortacsu, A. and Syverson, C. (2019) 'Price dispersion, household demand, and the gains from shopping online', *Quarterly Journal of Economics*, 134(2), pp. 713–768.

# APPENDIX B – R code

## Data Cleaning

```
# 1. Load dataset
df <- read.csv("Prices.csv")

# 2. Compute meal price for each wave (sum of entrée, fries, and soda)
df$meal_price_wave1 <- df$psoda + df$pfries + df$pentree
df$meal_price_wave2 <- df$psoda2 + df$pfries2 + df$pentree2

# 3. Compute averaged meal price across waves
df$meal_price_avg <- rowMeans(df[, c("meal_price_wave1", "meal_price_wave2")], na.rm = TRUE)

# 4. Average all store-level variables that appear in both waves
df$wagest_avg    <- rowMeans(df[, c("wagest",    "wagest2")],    na.rm = TRUE)
df$nmgrs_avg     <- rowMeans(df[, c("nmgrs",     "nmgrs2")],     na.rm = TRUE)
df$nregs_avg     <- rowMeans(df[, c("nregs",     "nregs2")],     na.rm = TRUE)
df$hrsopen_avg <- rowMeans(df[, c("hrsopen", "hrsopen2")], na.rm = TRUE)
df$emp_avg       <- rowMeans(df[, c("emp",       "emp2")],       na.rm = TRUE)

# 5. Remove all wave-specific and item-level price variables
df_clean <- subset(df, select = -c(
   psoda, pfries, pentree, psoda2, pfries2, pentree2,
   meal_price_wave1, meal_price_wave2,
   wagest, wagest2,
   nmgrs, nmgrs2,
   nregs, nregs2,
   hrsopen, hrsopen2,
   emp, emp2
))

# 6. Drop any remaining missing observations (final full-case sample)
df_clean <- na.omit(df_clean)

# 7. Create log-transformed dependent variable for regressions
df_clean$log_meal_price <- log(df_clean$meal_price_avg)
```

```
# 8. Check final sample size and descriptive statistics
cat("Final sample size:", nrow(df_clean), "observations\n")
summary(df_clean$meal_price_avg)
summary(df_clean$log_meal_price)

# 9. Export cleaned dataset
write.csv(df_clean, "cleaned_dataset.csv", row.names = FALSE)

# Notes:
# - Averaging is performed for all variables with both wave1 & wave2 values.
# - Missing values are only removed after all averages are constructed.
# - The resulting dataset (cleaned_dataset.csv) has ~369 observations,
#     slightly more than Graddy (1997)'s 322 due to fewer missing entries
#     in the public replication file.
```

## Descriptive Analysis

```
dat <- read.csv("cleaned_dataset.csv", stringsAsFactors = FALSE)
#1.Table 1. Descriptive Statistics for the Constructed Meal Price (USD).
N  <- sum(!is.na(dat$meal_price_avg))
M  <- mean(dat$meal_price_avg, na.rm = TRUE)
S  <- sd(dat$meal_price_avg, na.rm = TRUE)
Md <- median(dat$meal_price_avg, na.rm = TRUE)
tab1 <- data.frame(
  N = N, Mean = round(M, 2), SD = round(S, 2), Median = round(Md, 2)
)
write.csv(tab1, "table_summary_meal_cleaned.csv", row.names = FALSE, quote = FALSE)

tab1
file.exists("table_summary_meal_cleaned.csv")

#2.Figure 1. Mean Constructed Meal Price by Chain (USD).
mean_by_chain <- aggregate(meal_price_avg ~ chain_lab, data = dat, FUN = mean)
mean_by_chain <- mean_by_chain[order(mean_by_chain$meal_price_avg, decreasing = TRUE), ]
png("fig_bar_meal_by_chain_cleaned.png", width = 1400, height = 1000, res = 150)
par(mar = c(10, 6, 4, 2))
bp <- barplot(mean_by_chain$meal_price_avg,
```

```
                names.arg = mean_by_chain$chain_lab,
                las = 2, col = "grey85", border = "grey40",
                ylab = "Average meal price (USD)",
                main = "Average meal price by chain")
grid()
text(x = bp, y = mean_by_chain$meal_price_avg,
     labels = sprintf("$%.2f", mean_by_chain$meal_price_avg),
     pos = 3, cex = 0.9)
dev.off()
file.exists("fig_bar_meal_by_chain_cleaned.png")   # TRUE

#3.Figure 2. Pareto Chart of Store Counts by Chain.
freq <- sort(table(dat$chain_lab), decreasing = TRUE)
cum_pct <- cumsum(as.numeric(freq)) / sum(freq) * 100

png("fig_pareto_chain_cleaned.png", width = 1400, height = 1000, res = 150)
par(mar = c(10, 6, 4, 6))
bar_x <- barplot(freq, col = "grey85", border = "grey40",
                 las = 2, ylab = "Store count",
                 main = "Pareto chart of store counts by chain")
grid()
lines(bar_x, cum_pct/100 * max(freq), type = "b", lwd = 2)
ticks <- pretty(cum_pct)
ticks <- ticks[ticks >= 0 & ticks <= 100]
axis(4, at = (ticks/100) * max(freq), labels = round(ticks, 0))
mtext("Cumulative %", side = 4, line = 3)
text(x = bar_x, y = cum_pct/100 * max(freq),
     labels = paste0(round(cum_pct,1), "%"), pos = 3, cex = 0.9)
dev.off()

#4.Figure 3. Sample Composition by State (NJ vs PA).
state_counts <- table(dat$state_lab)
state_pct <- round(100 * state_counts / sum(state_counts), 1)
lbl <- paste0(names(state_counts), " (", state_pct, "%)")
png("fig_pie_state_cleaned.png", width = 1200, height = 900, res = 150)
par(mar = c(2, 2, 4, 2))
pie(state_counts, labels = lbl, col = gray.colors(length(state_counts)),
```

```
    main = "Sample share by state (NJ vs PA)", border = "grey40")
dev.off()
file.exists("fig_pie_state_cleaned.png")   # TRUE

#5.Figure 4a. Histogram of Constructed Meal Price with Mean and Median Reference Lines.
brks <- pretty(range(dat$meal_price_avg, na.rm = TRUE), n = 30)
ylim_top <- max(hist(dat$meal_price_avg, breaks = brks, plot = FALSE)$counts) * 1.25
png("fig_histogram_meal_price_avg_cleaned.png", width = 1400, height = 1000, res = 150)
par(mar = c(6, 6, 4, 2))
hist(dat$meal_price_avg, breaks = brks,
     col = "grey85", border = "grey40",
     xlab = "meal_price_avg (USD)", ylab = "Count",
     main = "Histogram of meal_price_avg",
     ylim = c(0, ylim_top))
grid()
abline(v = M,   lty = 1, lwd = 2)
abline(v = Md, lty = 2, lwd = 2)
legend("topright",
       legend = c(sprintf("Mean = $%.2f", M),
                  sprintf("Median = $%.2f", Md)),
       lty = c(1, 2), bty = "n")
dev.off()
file.exists("fig_histogram_meal_price_avg_cleaned.png")   # TRUE

#6.Figure 4b. Boxplots of Constructed Meal Price by Income Group (Median Split).
inc_med <- median(dat$income, na.rm = TRUE)
dat$income_bin <- ifelse(dat$income >= inc_med, "High income", "Low income")
png("fig_box_meal_by_income_bin.png", width = 1400, height = 1000, res = 150)
par(mar = c(6, 6, 4, 2))
boxplot(meal_price_avg ~ income_bin, data = dat,
        col = "grey85", border = "grey40",
        xlab = "", ylab = "Meal price (USD)",
        main = "Meal price by income (High vs Low)")
grid()
dev.off()
file.exists("fig_box_meal_by_income_bin.png")   # TRUE
```

```
#7.Table 2. Counts of High-Price Outlets (meal_price_avg ≥ 3.075) by State (NJ vs PA).
hi_cut <- 3.075
dat$high_price <- ifelse(dat$meal_price_avg >= hi_cut, 1, 0)
tab_hp <- as.data.frame.matrix(table(dat$state_lab, dat$high_price))
names(tab_hp) <- c("high_price_0","high_price_1")
tab_hp$Total <- rowSums(tab_hp)
tab_hp$state <- rownames(tab_hp)
tab_hp <- tab_hp[, c("state","high_price_1","Total")]
write.csv(tab_hp, "table_highprice_by_state_cleaned.csv", row.names = FALSE, quote = FALSE)
tab_hp
file.exists("table_highprice_by_state_cleaned.csv")   # TRUE
```

## Two-Sample t-Test

```
df <- read.csv("cleaned_dataset.csv", header = TRUE)

median_income <- median(df$lincome, na.rm = TRUE)
print(median_income)

income_group <- character(nrow(df))

for (i in 1:nrow(df)) {
  if (is.na(df$lincome[i])) {
    income_group[i] <- NA
  } else if (df$lincome[i] >= median_income) {
    income_group[i] <- "High"
  } else {
    income_group[i] <- "Low"
  }
}
df$income_group <- income_group

# delete null
high_income_prices <- df$meal_price_avg[df$income_group == "High" & !is.na(df$meal_price_avg)]
low_income_prices   <- df$meal_price_avg[df$income_group == "Low"
& !is.na(df$meal_price_avg)]
```

```
#high income
n_high_income <- length(high_income_prices)
mean_high_income <- mean(high_income_prices,na.rm = TRUE)
sd_high_income <- sd(high_income_prices,na.rm = TRUE)

#low    income
n_low_income <- length(low_income_prices)
mean_low_income <- mean(low_income_prices,na.rm = TRUE)
sd_low_income <- sd(low_income_prices,na.rm = TRUE)

#lincome t test
t_income <- t.test(meal_price_avg ~ income_group, data = df, var.equal = FALSE)
print(t_income)
```

## Proportion Test

```
df <- read.csv("cleaned_dataset.csv", header = TRUE)

median_meal <- median(df$meal_price_avg, na.rm = TRUE)

df$high_price <- ifelse(df$meal_price_avg >= median_meal, 1, 0)

# NJ GROUP（NJ == 1）
nj_data <- df[df$NJ == 1, ]
n_nj <- nrow(nj_data)
x_nj <- sum(nj_data$high_price, na.rm = TRUE)
p_nj <- x_nj / n_nj

# PA GROUP （NJ == 0）
pa_data <- df[df$NJ == 0, ]
n_pa <- nrow(pa_data)
x_pa <- sum(pa_data$high_price, na.rm = TRUE)
p_pa <- x_pa / n_pa

test_result <- prop.test(
  x = c(x_nj, x_pa),
```

```
  n = c(n_nj, n_pa),
  correct = FALSE
)
print(test_result)
```

## Regression

```
# install.packages(c("tidyverse", "broom", "lmtest", "sandwich", "car", "MASS", "glmnet", "caret"))

library(tidyverse)
library(broom)
library(lmtest)
library(sandwich)
library(car)
library(MASS)
library(glmnet)
library(caret)

set.seed(2025)

> df <- read.csv("cleaned_dataset.csv")

if(!"log_meal_price" %in% names(df) && "meal_price_avg" %in% names(df)){
  df <- df %>% mutate(log_meal_price = log(meal_price_avg))
}
```

### #1.Correlation Matrix

```
vars_cor <- c("prpblck", "lincome", "prppov", "lhseval",
              "ldensity", "crmrte", "wagest_avg", "emp_avg")

cor_mat <- cor(df[, vars_cor], use = "complete.obs")
round(cor_mat, 3)

#2.VIF
# Full model
fml_full <- log_meal_price ~
  prpblck + lincome + prppov + lhseval +
  ldensity + crmrte +
```

```
    wagest_avg + emp_avg +
    BK + KFC + RR + NJ

mod_full <- lm(fml_full, data = df)
summary(mod_full)

# VIF：official multicollinearity test
vif_full <- car::vif(mod_full)
vif_full

#3.Model1
#Model 1
fml_mod1 <- log_meal_price ~
    prpblck + lincome + prppov + lhseval +
    BK + KFC + RR + NJ

mod1 <- lm(fml_mod1, data = df)
summary(mod1)

vif_mod1 <- car::vif(mod1)
vif_mod1

#4.Stepwise
fml_lower <- log_meal_price ~ 1
scope_vars <- attr(terms(mod1), "term.labels")
fml_upper    <- as.formula(paste("log_meal_price ~", paste(scope_vars, collapse = " + ")))

#stepwise（both）
mod2 <- MASS::stepAIC(
    mod1,
    scope = list(lower = fml_lower, upper = fml_upper),
    direction = "both",
    trace = TRUE
)

summary(mod2)
```

```
Checks
plot(mod2, which = 1)
plot(mod2, which = 2)
lmtest::bptest(mod2)
plot(mod2, which = 4)

x_mod1 <- model.matrix(fml_mod1, data = df)[ , -1]
y_mod1 <- df$log_meal_price

#5.OLS
ctrl <- trainControl(method = "cv", number = 10)

ols_cv <- train(
   x = x_mod1,
   y = y_mod1,
   method = "lm",
   trControl = ctrl
)

ols_rmse <- ols_cv$results$RMSE
ols_rmse      # ≈ 0.0773

#6.Ridge
set.seed(2025)
ridge_cv <- cv.glmnet(
   x_mod1, y_mod1,
   alpha = 0,
   nfolds = 10,
   standardize = TRUE
)

ridge_lambda <- ridge_cv$lambda.min
ridge_rmse      <- sqrt(min(ridge_cv$cvm))

ridge_lambda
ridge_rmse

ridge_coef <- coef(ridge_cv, s = "lambda.min")
```

```
ridge_coef

#7.Lasso
set.seed(2025)
lasso_cv <- cv.glmnet(
   x_mod1, y_mod1,
   alpha = 1,
   nfolds = 10,
   standardize = TRUE
)

lasso_lambda <- lasso_cv$lambda.min
lasso_rmse      <- sqrt(min(lasso_cv$cvm))

lasso_lambda
lasso_rmse

lasso_coef <- coef(lasso_cv, s = "lambda.min")
lasso_coef
```

# APPENDIX C—CLEANED DATASET DESCRIPTION

## Data (“cleaned_dataset”)

A data frame with 369 observations on store-level prices, characteristics, and demographic variables used in the analysis of price differentials across New Jersey and Pennsylvania fast-food restaurants (Graddy, 1997 replication).

## Price and Dependent Variables

meal_price_avg – average total meal price (entrée + fries + soda) across two survey waves

log_meal_price – natural logarithm of the average meal price

lpsoda – log(price of soda)

lpfries – log(price of fries)

## Averaged Store-Level Characteristics

wagest_avg – average starting wage (across waves)

nmgrs_avg – average number of managers

nregs_avg – average number of registers

hrsopen_avg – average weekly hours open

emp_avg – average number of employees

## Regional and Demographic Characteristics (Census Data)

prpblck – proportion of Black residents in the ZIP code

prppov – proportion of residents below the poverty line

prpncar – proportion of residents without a car

income – median family income (ZIP code)

hseval – median housing value (ZIP code)

density – population density (town)

crmrte – crime rate (town)

county – county label

## Log-Transformed Census Variables

lincome – log(median family income)

lhseval – log(median housing value)

ldensity – log(population density)

## Ownership, Chain, and Location Indicators

compown – 1 if company-owned store, 0 if franchised

chain – numeric identifier for fast-food chain

BK, KFC, RR – dummy variables for Burger King, KFC, and Roy Rogers (Wendy's as base)

state – 1 = New Jersey, 2 = Pennsylvania

NJ – dummy for New Jersey stores (1 = NJ, 0 = PA)

nstores – number of stores operated by the chain in that county

## Notes

All two-wave variables (price and store-level characteristics) have been averaged before case deletion.

lpsoda and lpfries remain as single-item log prices for item-specific regressions (as in Tables 3–4 of Graddy, 1997).

The dataset contains 369 complete observations, slightly more than the original 322 due to fewer missing entries in the replication file.

# APPENDIX D—AI DECLARATION

We confirm that this submission is entirely our own work. No part of it has been generated or paraphrased using AI tools or similar technologies, except for basic spelling or grammar checking.