US Fast Food Chain Regression

Evan Shields, Shoaib Ansari, Yash Sihag
Department of Mathematics, University of Arizona
DATA 467: Introduction to Applied Regression and Generalized Linear Models
Taryn Laird
December 6, 2023

Introduction

With a market size of over 322 billion U.S. dollars in 2021, the fast food industry, formally known as the quick service restaurant (QSR) industry, is consistently growing at a rapid rate year over year (Statista, 2023). According to the CDC, in a study with data from between 2013 and 2016, on any given day, over a third of American adults consume fast food (Fryar et al., 2018). In any given municipality in our country, it is hard to be able to walk around without seeing at least one fast food restaurant, especially one of the top 50 fast food chains in the United States. With the fast food industry's success showing no sign of slowing down anytime soon, we wanted to find, on an industry-wide level, how fast food sales can be predicted by the number of fast food restaurants a chain has and the growth of the chain. This paper seeks to answer the following question: is the number of stores a fast food chain has along with the growth of the fast food chain and sales per store correlated with how a fast food chain performs in terms of sales?

Methods

Data for this correlation study was sourced from the Kaggle dataset "Top 50 Fast Food Chains in the United States" (Banerjee, 2019), which contains information on the top 50 fast-food chains in the United States for 2021. Though the dataset focuses on the top 50 chains, we still expect wide variation in store numbers and total systemwide sales due to the diversity in chain sizes and popularity. As this data was collected directly from the restaurant themselves (QSR Magazine, 2022), a potential source of bias in this study may arise from the accuracy of the data. Data quality may vary between different fast-food chains, and discrepancies in reporting or data collection methods can introduce bias.

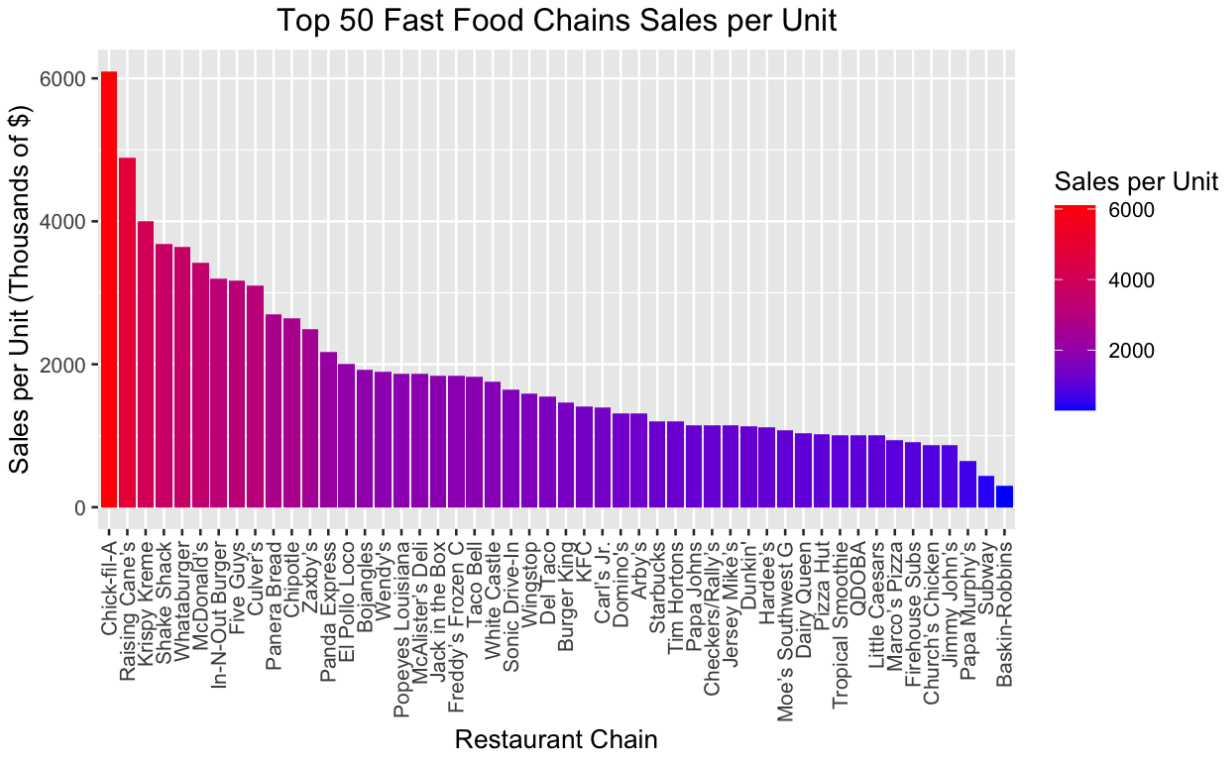
In addition, the data was collected with the help of a third party consumer behavior analysis company, Circana, in order to provide accurate estimates of sales, as it was not possible to acquire exact figures for total systemwide sales of each fast-food chain (QSR Magazine, 2022). With this being said, it is important to note that the number of stores nationwide and the change in this number from 2020 to 2021 are not estimates and are exact.

The response variable we are using is systemwide sales in 2021 (Millions of US$). Systemwide sales includes sales from both franchised stores and company-owned stores. The first predictor is

the total number of stores in 2021, which includes franchised as well as company-owned locations. The second predictor is the growth of the restaurant chain, which is calculated as the difference in total store count in 2021 compared to 2020. The third predictor is sales per unit (store). The fourth predictor is the total number of franchised stores. The fifth predictor is the total number of company-owned stores. Before performing any analysis, we predict that there will be a moderately strong positive correlation between our predictor variables and our response variable, with sales per unit, our third predictor being the strongest predictor.

Data Analysis

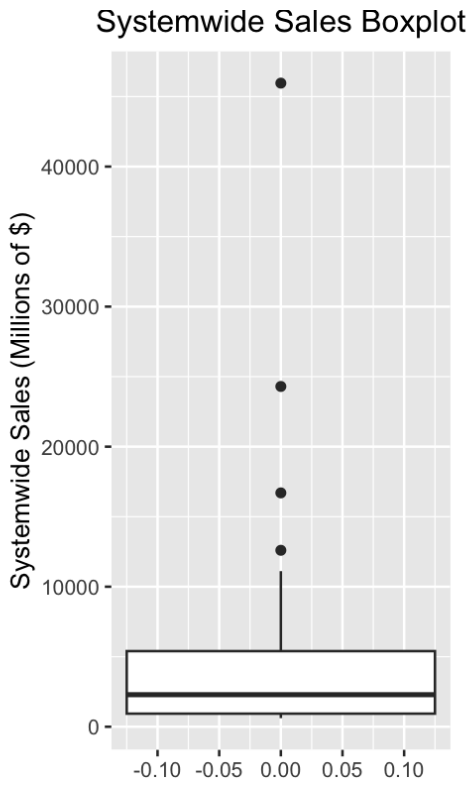
The dataset contains information on the top 50 fast food chains in the United States, including systemwide sales, average sales per unit, number of franchised and company stores, and total units in 2021.

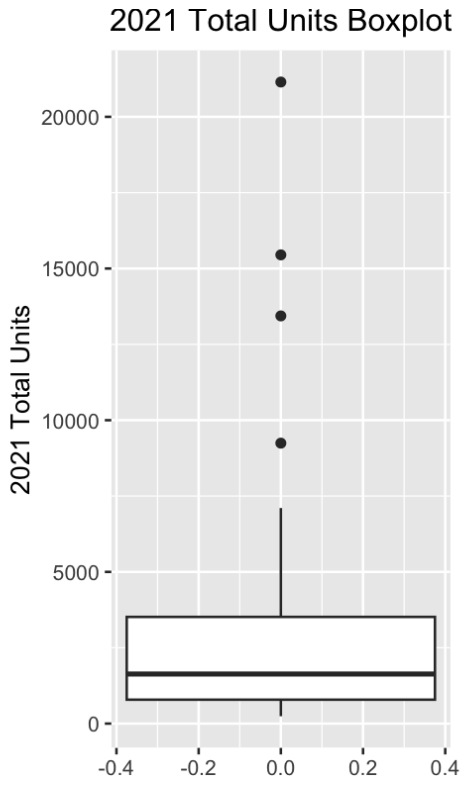


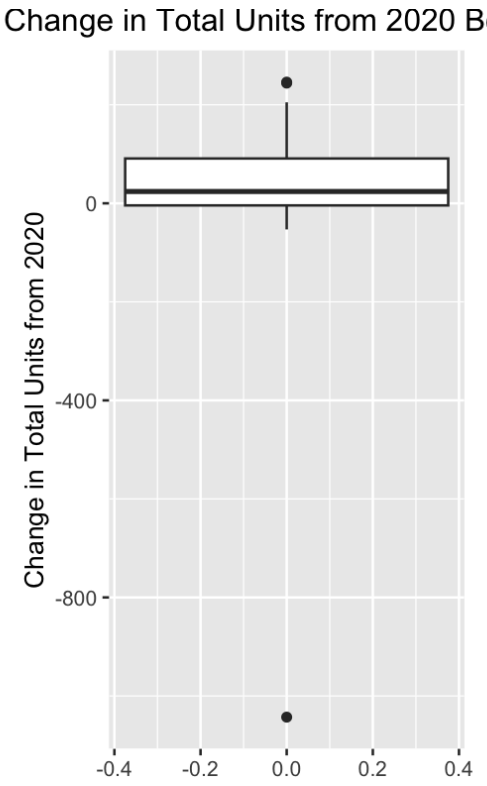
Initial data exploration shows that across the 50 chains, total systemwide sales were $248 billion, with a median of $2.3 billion per chain. The total number of units was 158,370, with a median of 1,634 units per chain. Between 2020 and 2021, the total change in units was an increase of 1,485, with a median increase of 24 units per chain.

The distribution of systemwide sales is right-skewed, with most chains having sales below $10 billion but a few major chains like McDonald's and Starbucks over $20 billion. Sales per unit also varies widely from $296,000 to $6.1 million, reflecting differences in store size and sales volume across chains. The number of total units in 2021 appears symmetric and bell-shaped, centered around the median of 1,634 units.

In terms of outliers, Subway stands out with over 21,000 total units compared to the next highest chains around 7,000 units. Being such an extreme outlier, Subway could distort the model and may need to be removed depending on the research question. In addition, McDonald’s will likely have to be removed due to having a much higher total systemwide sales than the other franchises. Overall, this dataset provides a useful snapshot of recent fast food industry trends, laying the groundwork for the upcoming regression analysis.




2021 Total Units Boxplot
2021 Total Units
20000
15000
10000
5000
0
-0.4
-0.2
0.0
0.2
0.4


Change in Total Units from 2020 Boxplot
Change in Total Units from 2020
0
-400
-800
-0.4
-0.2
0.0
0.2
0.4

Linear Model

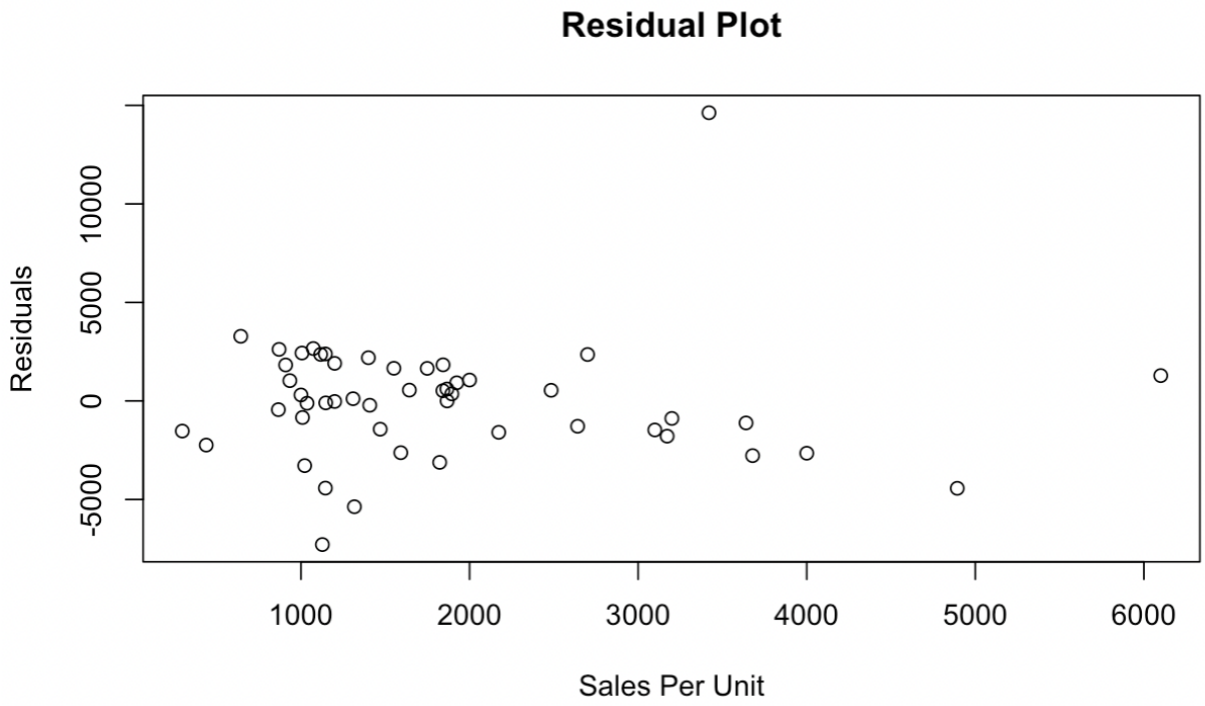
Multiple Linear Regression Model:

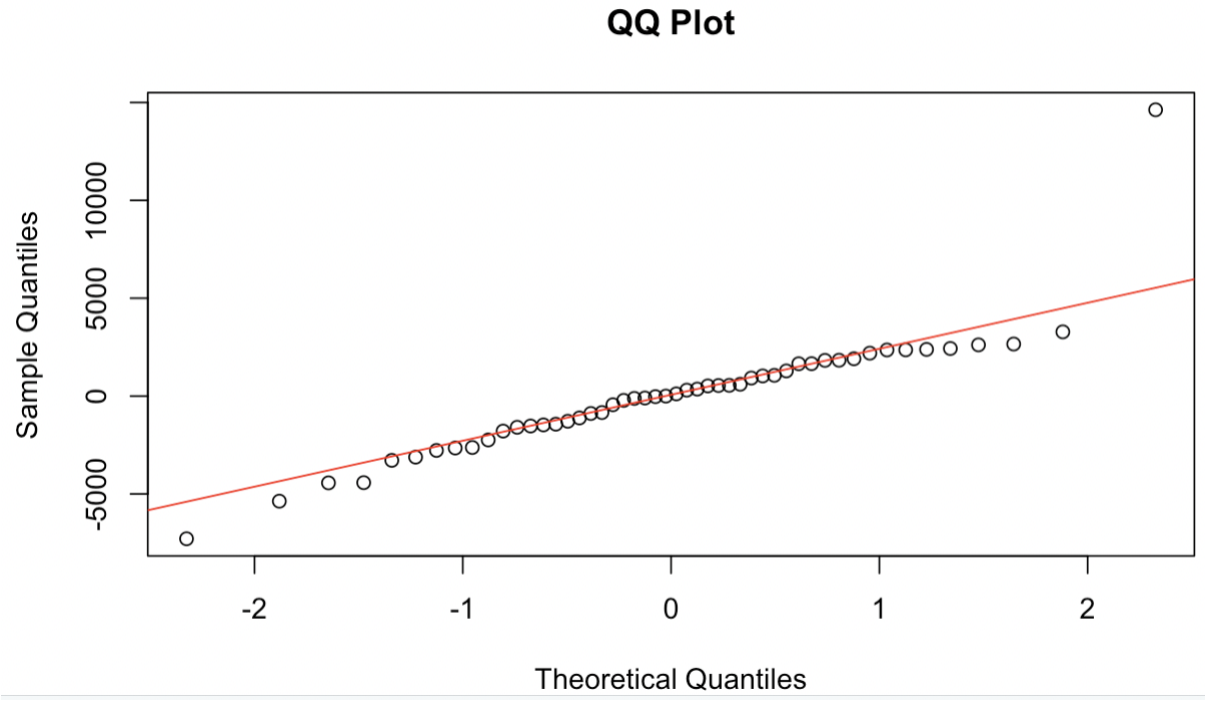
- Response: Systemwide sales
- Predictors: Sales per Unit, Franchise Stores, Company Stores, 2021 Total Units, Change in Total Units from 2020
- -4804.6938 + 1.9503(Sales per Unit) - 747.9679(Franchised Stores) - 748.3852(Company Stores) + 749.7836(2021 Total Units) + 21.9119(Change in TU from 2020)

The regression model output shows a quantitative and categorical predictor and their association with the dependent variable, Systemwide Sales. The coefficient for the quantitative predictor Sales per Unit is 1.9503, indicating a positive link with Systemwide Sales. Assuming all else is held constant, we expect Systemwide Sales to increase by 1.9503 million for each 1000 unit increase in Sales per Unit. A p-value of 2.81e-05, far below 0.05, shows that this link is statistically significant.

Moving on to the category predictor Company Stores, the coefficient is -748.3852. If we use Company Stores as a binary variable (1 for company shops, 0 otherwise), this coefficient shows that company stores lower Systemwide Sales by 748.3852 units on average. The enormous standard error of 1114.7567 compared to the coefficient and the t-value of -0.671 imply that this finding is not statistically significant, since the p-value is 0.506, much beyond the 0.05 threshold. Thus, corporate stores may not affect Systemwide Sales, and we would not reject the null hypothesis that sales are the same regardless of their presence. In this model, Sales per Unit predicts Systemwide Sales, while Company Stores does not. These findings must be considered alongside the model's other diagnostics to completely assess the predictors' effects.

In addition, the near-identical and extremely high p-values of around .5 indicate that there may be multicollinearity issues in the data. This makes sense, as the total number of units is calculated as the sum of franchised stores and company stores, all three of which are predictors in this model. Two out of three of these predictors will be removed in part 4 of this paper.


Residual Plot
Residuals
-5000
0
5000
10000
1000
2000
3000
4000
5000
6000
Sales Per Unit


QQ Plot
Sample Quantiles
-5000
0
5000
10000
-2
-1
0
1
2
Theoretical Quantiles

Diagnostics

According to the Residual Plot, the residuals do not create a recognizable pattern, which shows no non-linearity in the predictor-outcome connection. However, the “fan” shape (widening variance as Sales Per Unit grows) may imply heteroscedasticity, when error variance is not constant across all independent variable levels. Some aspects stand out, especially with larger Sales Per Unit values. These outliers may affect the regression model and represent leverage points.

QQ Plots assess if a dataset has a normal distribution. It compares sample data quantiles to theoretical distribution quantiles. The assumption that residuals are normally distributed is commonly tested in regression diagnostics. If points are close to the red line in the QQ Plot, residuals are normally distributed. The figure indicates that the points follow the line but diverge significantly in the tails, especially near the top. This suggests that the residuals may have a heavy-tailed distribution, which deviates from normality but not significantly considering this is real-world data.

Hypothesis Testing for Sales per Unit Coefficient

Null Hypothesis:
The null hypothesis states that the Sales per Unit coefficient (beta_1) is equal to zero, which means that Sales per Unit has no effect on Systemwide Sales.
H0: beta_1 = 0

Alternative Hypothesis:
The alternative hypothesis states that the Sales per Unit coefficient (beta_1) is not equal to zero, which means that Sales per Unit does have an effect on Systemwide Sales.
H1: beta_1 != 0

Test statistic: (t) = Estimate / Std. Error = 1.9503 / 0.4173 = 4.674
Degrees of Freedom (df):

The degrees of freedom for the t-test in a regression model is the number of observations minus the number of estimated parameters. In this case, there are 50 observations (44 degrees of freedom plus 5 estimated parameters plus 1 for the intercept).
df = n - (k + 1) = 50 - (5 + 1) = 44

P-value:
The p-value is a measure of the probability of observing a test statistic as extreme as, or more extreme than, the one observed if the null hypothesis is true. In this output, the p-value for the Sales per Unit coefficient is given as 2.81e-05.
p-value = 2.81e-05

Conclusion:
Given that the p-value is much smaller than the significance level of 0.05, we reject the null hypothesis. This means that there is statistically significant evidence at the 0.05 level to suggest that the Sales per Unit does have an effect on Systemwide Sales. The Sales per Unit has a positive relationship with Systemwide Sales, as indicated by the positive coefficient (1.9503), and this relationship is statistically significant. Therefore, it can be concluded that as Sales per Unit increases, Systemwide Sales are also expected to increase, holding all other variables constant.

Full Versus Reduced Model

Null hypothesis (H0): The reduced model fits the data as well as the full model.
Alternative hypothesis (H1): The full model fits the data better than the reduced model.

F-statistic = 65.663
Degrees of freedom:
Numerator (full model df): 3
Denominator (reduced model df): 44
P-value = 2.762e-16

Since the p-value is < 0.05, we reject the null hypothesis and conclude that the full model provides a significantly better fit than the reduced model. Adding the Franchised Stores and Company Stores variables improves model fit despite the individual variables not being significant. This suggests that together these variables explain additional variation in Systemwide Sales.

The reduced model's performance has deteriorated significantly, as seen by lower Multiple R-squared and Adjusted R-squared values compared to the entire model. Given this result, it's clear that even though some predictors were not individually significant, they contribute to the model when included with other variables. This is likely due to the aforementioned multicollinearity issues between three of the predictors.

Confidence Intervals

95% confidence interval for new observations using full model:
(1.75e+03, 1.12e+04)

95% confidence interval for new observations using reduced model:
(-1.20e+04, 1.20e+04)

The confidence interval for the full model ranges from 1,750 to 11,200 (in millions $). This indicates that for a new observation, we can be 95% confident the true 'Systemwide Sales' value lies within this range. The reduced model's confidence interval ranges from -12,000 to 12,000 (in millions $). This is much wider than the full model's interval. In addition, this interval includes negative values. However, in context, sales can never be negative, indicating that this is an unreliable interval to rely on. The reduced model cannot precisely estimate Systemwide Sales for new data points after excluding the Franchised Stores and Company Stores variables.

Remove Outlier & Best Model Identification

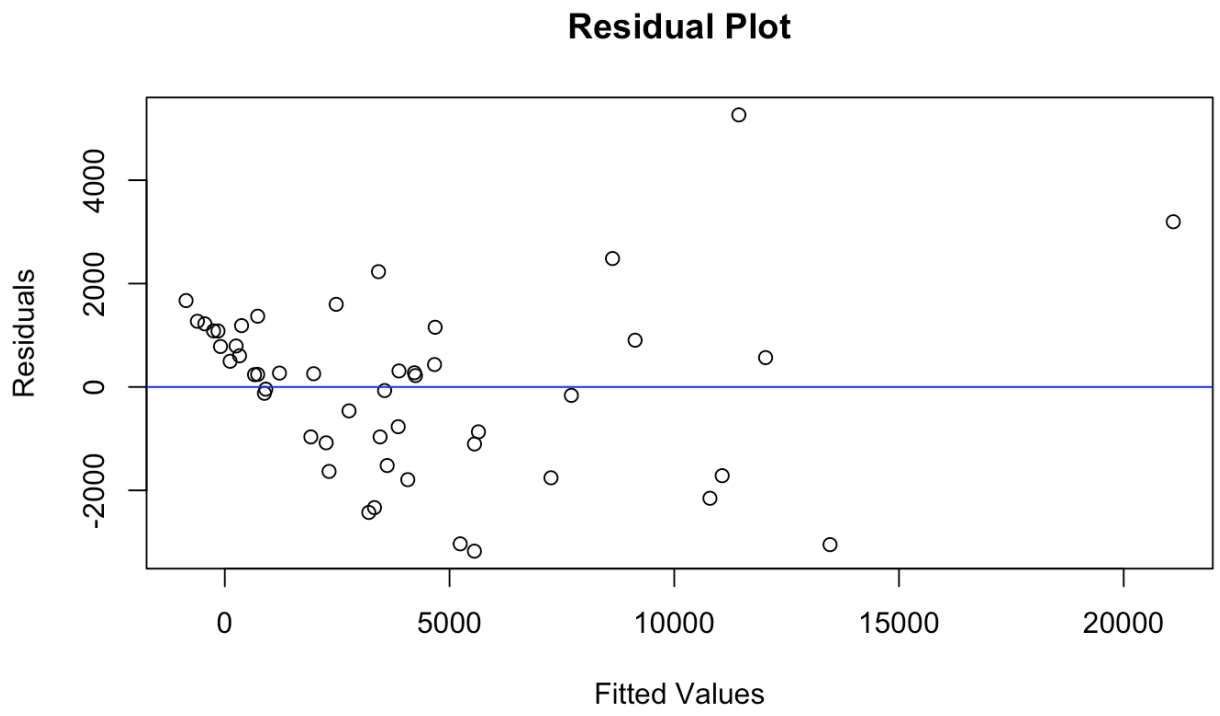
Part 4 - We wanted to do ridge regression, but we couldn't get the genridge library to install properly. We used AIC and BIC forward model selection instead to take care of the multicollinearity issues.

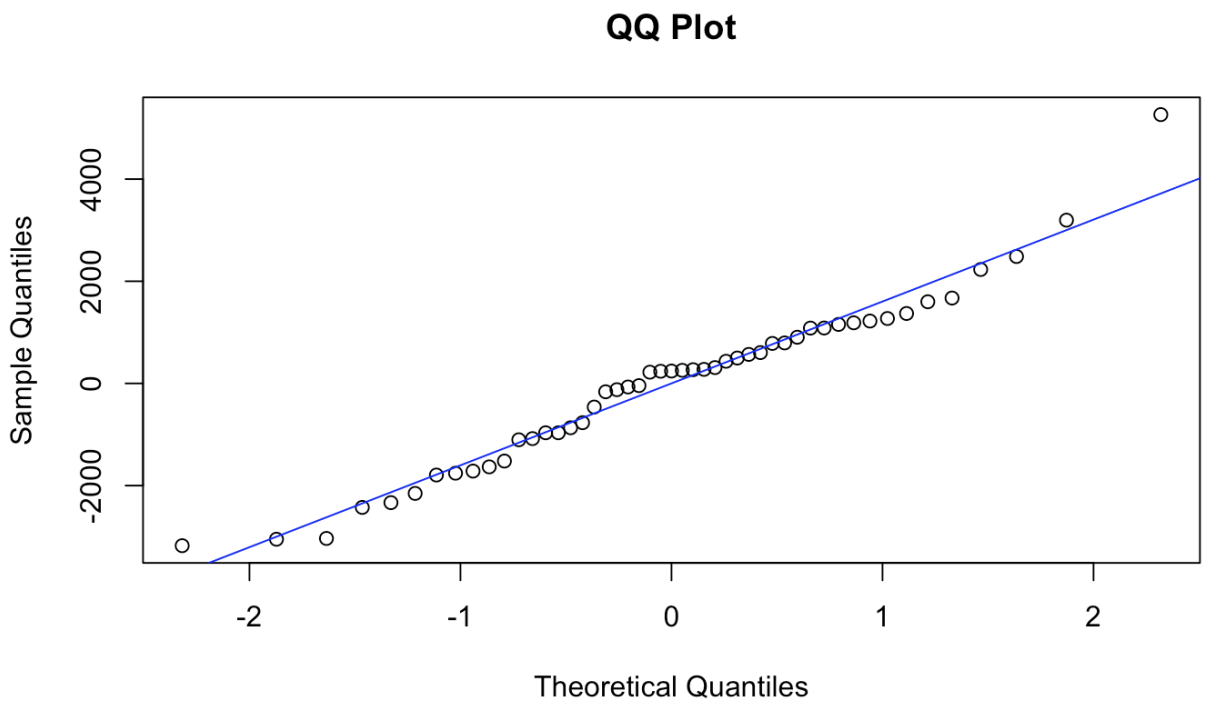
Both selection processes found the following model to be the best model:

- -2612 + 1.345(Sales per Unit) + 1.326(2021 Total Units) + 1.433(Change in TU from 2020)

In this model, all of the predictors are significant, and the extremely large standard errors for certain predictors are gone, indicating that the issues with multicollinearity are taken care of. This better model also has an R-squared value of .8754 compared to the original model at .8293, indicating that this latest model is a better fit.

Diagnostics For Newest Model





After removing the outlier and the predictors company stores and franchise stores, the qqplot has a lot less significant short-tail error, which is good. In addition, with the removal of the outlier, it is easier to see what is going on in the residual plot. While the scatter doesn't appear to be completely random, there also does not appear to be a distinct shape to the scatter either.

Conclusion

This paper sheds light on the intricate relationship between key predictors and systemwide sales in the fast-food industry. Through meticulous analysis of the top 50 U.S. chains in 2021 by systemwide sales, we identified Sales per Unit, Total Stores, and Change in Total Stores from the previous year as significant drivers of sales. Addressing challenges like outliers and multicollinearity, our optimized model can be used to predict systemwide sales with enhanced accuracy.

While our findings provide valuable insights into the dynamics of fast-food sales, the evolving nature of the industry suggests that there may be additional variables of importance that may aid in predicting systemwide fast food chain sales, especially as advancements in technology come about along with innovative ways to order and deliver food. Future research into this should

consider additional predictors and delve deeper into different nuances that shape the performance of each top-performing fast-food chain, ensuring a comprehensive understanding of the quick service restaurant industry.

References

Banerjee, S. (2022, December 9). *Top 50 Fast-Food Chains in USA*. Kaggle. Retrieved October 13, 2023, from https://www.kaggle.com/datasets/iamsouravbanerjee/top-50-fastfood-chains-in-usa/data

Fryar, C. D., Hughes, J. P., Herrick, K. A., & Ahluwalia, N. (2018, October). *Products - Data Briefs - Number 320 - September 2018*. CDC. Retrieved October 12, 2023, from https://www.cdc.gov/nchs/products/databriefs/db322.htm

QSR Magazine. (2022, August). *QSR 50*. QSR Magazine. https://www.qsrmagazine.com/downloads/the-2023-qsr-50/

Statista. (2023, August 31). *U.S. fast food restaurants statistics & facts*. Statista. Retrieved October 12, 2023, from https://www.statista.com/topics/863/fast-food/#topicOverview

Appendix

```r
fastFood <- read.csv("FastFood.csv")

colnames(fastFood) <- c("Chain Name", "Systemwide Sales", "Sales per Unit", "Franchised Stores",
                        "Company Stores", "2021 Total Units", "Change in TU from 2020")

# Increase of 1485 units of these top 50 chains from 2020 to 2021
# Median increase of 24 units per chain
# 5-num-sum: -1043, -6, 24, 102, 246
sum(fastFood$`Change in TU from 2020`)
median(fastFood$`Change in TU from 2020`)
fivenum(fastFood$`Change in TU from 2020`)

# 158370 total units across all top 50 restaurants
# Median amount of units is 1634
# 5-num-sum: 243, 773, 1634, 3552, 21147
sum(fastFood$`2021 Total Units`)
median(fastFood$`2021 Total Units`)
fivenum(fastFood$`2021 Total Units`)

# $248,253,000,000 total system wide sales across all top 50 restaurants
# Median amount of total system wide sales is $2,289,500,000
# 5-num-sum (in millions $): 615, 931, 2289.5, 5500, 45960
sum(fastFood$`Systemwide Sales`)
median(fastFood$`Systemwide Sales`)
fivenum(fastFood$`Systemwide Sales`)
```

Restaurant vs. SPU bar chart

```r
fastFood$`Chain Name` <- substr(fastFood$`Chain Name`, 1, 17)

ggplot(fastFood, aes(x = reorder(`Chain Name`, -`Sales per Unit`), y = `Sales per Unit`,
                     fill = `Sales per Unit`)) +
  geom_bar(stat = "identity") +
  scale_fill_gradient(low = "blue", high = "red") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1), plot.title = element_text(hjust = 0.5)) +
  labs(x = "Restaurant Chain", y = "Sales per Unit (Thousands of $)") +
  ggtitle("Top 50 Fast Food Chains Sales per Unit")
```

Boxplots of predictors and response

```{r}
# Boxplot for Systemwide Sales
ggplot(fastFood, aes(y = `Systemwide Sales`)) +
  geom_boxplot(width = 0.25) +
  labs(title = "Systemwide Sales Boxplot", y = "Systemwide Sales (Millions of $)") +
  theme(plot.margin = margin(0, 6, 0, 6, "cm"), plot.title = element_text(hjust = 0.5))

# Boxplot for Sales per Unit
ggplot(fastFood, aes(y = `2021 Total Units`)) +
  geom_boxplot() +
  labs(title = "2021 Total Units Boxplot") +
  theme(plot.margin = margin(0, 6, 0, 6, "cm"), plot.title = element_text(hjust = 0.5))

# Boxplot for Change in TU from 2020
ggplot(fastFood, aes(y = (`Change in TU from 2020`))) +
  geom_boxplot() +
  labs(title = "Change in Total Units from 2020 Boxplot", y = "Change in Total Units from 2020") +
  theme(plot.margin = margin(0, 6, 0, 6, "cm"), plot.title = element_text(hjust = 0.5))
```

```{r}
# Define the full model
full_model <- lm(fastFood$`Systemwide Sales` ~ fastFood$`Sales per Unit` +
                   fastFood$`Franchised Stores`+ fastFood$`Company Stores` + fastFood$`2021 Total Units` +
                   fastFood$`Change in TU from 2020`)

# Summary of the full model
summary(full_model)

```

```
Call:
lm(formula = fastFood$`Systemwide Sales` ~ fastFood$`Sales per Unit` +
    fastFood$`Franchised Stores` + fastFood$`Company Stores` +
    fastFood$`2021 Total Units` + fastFood$`Change in TU from 2020`)

Residuals:
    Min      1Q  Median      3Q     Max
-7289.2 -1514.6    56.7  1655.3 14626.0

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -4804.5938  1027.2224  -4.677 2.78e-05 ***
fastFood$`Sales per Unit`         1.9503     0.4173   4.674 2.81e-05 ***
fastFood$`Franchised Stores`   -747.9679  1114.7198  -0.671    0.506
fastFood$`Company Stores`      -748.3852  1114.7567  -0.671    0.506
fastFood$`2021 Total Units`     749.7836  1114.7292   0.673    0.505
fastFood$`Change in TU from 2020`  21.9119     3.2078   6.831 2.02e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3280 on 44 degrees of freedom
Multiple R-squared:  0.8297,    Adjusted R-squared:  0.8103
F-statistic: 42.87 on 5 and 44 DF,  p-value: 7.913e-16
```

```{r}

plot(fastFood$`Sales per Unit`, residuals(full_model), xlab = "Sales Per Unit", ylab = "Residuals", main = "Residual Plot")

residuals <- residuals(full_model)
qqnorm(residuals, main = "QQ Plot")
qqline(residuals, col = "red")

```

```r
reduced_model <- lm(fastFood$`Systemwide Sales`~ fastFood$`Sales per Unit` + fastFood$`Change in TU from 2020`)
summary(reduced_model)
anova(reduced_model, full_model)
AIC(full_model, reduced_model)
BIC(full_model, reduced_model)
```

```
Call:
lm(formula = fastFood$`Systemwide Sales` ~ fastFood$`Sales per Unit` +
    fastFood$`2021 Total Units` + fastFood$`Change in TU from 2020`)

Residuals:
    Min      1Q  Median      3Q     Max
-3176.5 -1080.2   242.3  1083.2  5262.6

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -2.612e+03  5.697e+02  -4.584 3.62e-05 ***
fastFood$`Sales per Unit`        1.345e+00  2.248e-01   5.984 3.30e-07 ***
fastFood$`2021 Total Units`      1.326e+00  7.536e-02  17.591  < 2e-16 ***
fastFood$`Change in TU from 2020` 1.433e+01  1.664e+00   8.612 4.51e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1718 on 45 degrees of freedom
Multiple R-squared:  0.8753,    Adjusted R-squared:  0.8669
F-statistic: 105.2 on 3 and 45 DF,  p-value: < 2.2e-16
```