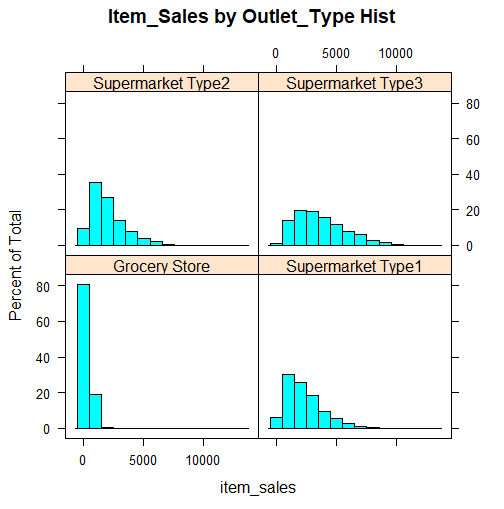
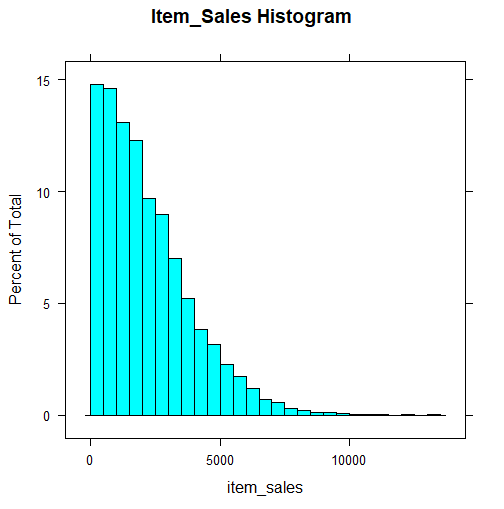
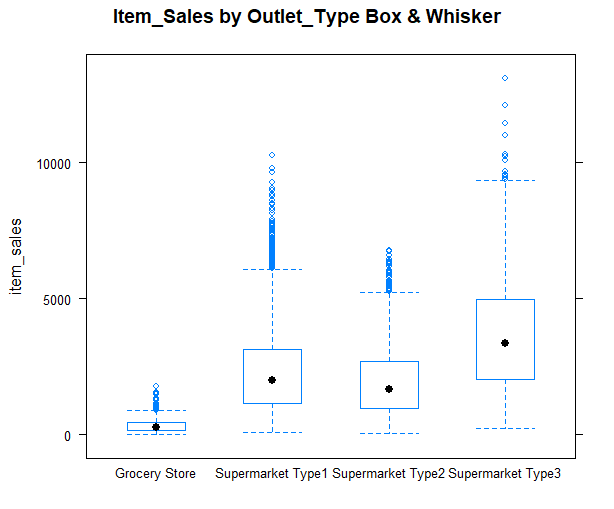
Grocery Sales

**1. We want to analyze how sales vary by outlet type: Grocery store versus Supermarket Type 1, 2, or 3. Ignore city information for this analysis. Conduct the necessary analysis and present the output in a nice, concise manner. Interpret the findings of this analysis, i.e., which type of store has the highest sales, by what amount, etc.**

**Dataset Analysis**





The histogram of item\_sales alone looks almost like a Poisson distribution. However, when splitting it by Outlet\_Type we see that it is better described as a distribution with a strong right-skew. The split histogram also shows that Grocery Stores have disproportionately low sales compared to the other types, while Type 3 supermarkets are more likely to have higher sales than the rest. A box and whisker plot allows us to see this more clearly.

**Feature Selection**

|  |  |  |
| --- | --- | --- |
| **Predictor** | **Effect** | **Reasoning** |
| Outlet\_Type | +/- | As shown in the histogram and box & whisker plot, the type of outlet will likely have a measurable effect on sales |
| Item\_MRP | + | By definition, higher priced items will aggregate to higher total sales. |

**Models**

re1 <- lmer(item\_sales ~ item\_mrp + outlet\_type + (1 | outlet\_id), data = df, REML=FALSE)

re2 <- lmer(item\_sales ~ item\_mrp\*outlet\_type + (1 | outlet\_id), data = df, REML=FALSE)

fe2 <- lm(item\_sales ~ item\_mrp\*outlet\_type + outlet\_id, data=df)

===========================================================================================================

Dependent variable:

---------------------------------------------------------------------

item\_sales

linear OLS

mixed-effects

re1 re2 fe2

-----------------------------------------------------------------------------------------------------------

item\_mrp 15.561\*\*\* (0.196) 2.473\*\*\* (0.527) 2.473\*\*\* (0.527)

outlet\_typeSupermarket Type1 1,961.971\*\*\* (51.561) -20.864 (95.510) 28.248 (99.288)

outlet\_typeSupermarket Type2 1,634.043\*\*\* (73.280) -68.636 (130.434) -66.928 (122.785)

outlet\_typeSupermarket Type3 3,361.789\*\*\* (73.210) 157.208 (130.485) 158.916 (122.840)

outlet\_idOUT013 -68.088 (49.652)

outlet\_idOUT017 6.872 (49.732)

outlet\_idOUT018

outlet\_idOUT019 3.428 (65.126)

outlet\_idOUT027

outlet\_idOUT035 43.597 (49.682)

outlet\_idOUT045 -166.802\*\*\* (49.691)

outlet\_idOUT046 -99.720\*\* (49.679)

outlet\_idOUT049

item\_mrp:outlet\_typeSupermarket Type1 14.127\*\*\* (0.575) 14.127\*\*\* (0.575)

item\_mrp:outlet\_typeSupermarket Type2 12.146\*\*\* (0.768) 12.146\*\*\* (0.768)

item\_mrp:outlet\_typeSupermarket Type3 22.877\*\*\* (0.774) 22.877\*\*\* (0.774)

Constant -1,843.244\*\*\* (53.619) -7.072 (86.768) -8.780 (87.072)

-----------------------------------------------------------------------------------------------------------

Observations 8,523 8,523 8,523

R2 0.607

Adjusted R2 0.606

Log Likelihood -72,002.000 -71,557.270

Akaike Inf. Crit. 144,018.000 143,134.500

Bayesian Inf. Crit. 144,067.400 143,205.000

Residual Std. Error 1,071.245 (df = 8509)

F Statistic 1,009.000\*\*\* (df = 13; 8509)

===========================================================================================================

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)

re1 7 144018 144067 -72002 144004

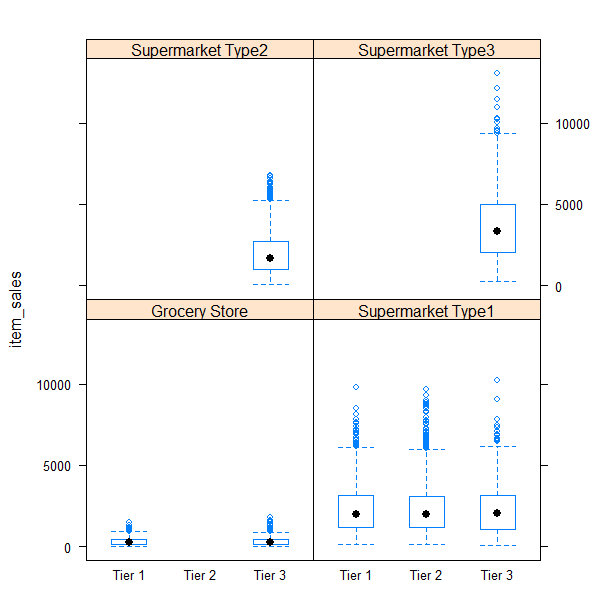
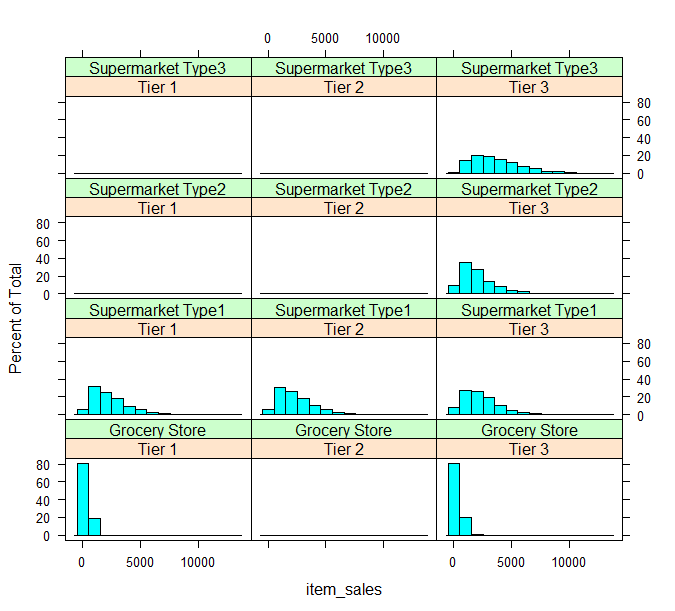
re2 10 143135 143205 -71557 143115 889.465 3 < 2.2e-16 \*\*\*

fe2 15 143126 143232 -71548 143096 18.599 5 0.002282 \*\*

The three models whose results are displayed above include two random-effects models and one fixed-effects model. While each includes the outlet\_type and item\_mrp, fe2 and re2 multiply item\_mrp by outlet\_type to create a compound predictor. Looking at the ANOVA output, this appears to improve the model (judging by a reduced AIC).

In either case, it appears that the Supermarket Type 3 (ST3) has the highest sales of all of the outlet types. The coefficients of re1 indicate that ST3 outlets have $3361.79 higher sales than Grocery Stores, whereas ST1 and ST2 are only $1961.97 and $1634.04 higher respectively.

Re2 and fe2 must be interpreted differently due to the feature engineering. Each of these models indicates that ST3 outlets increase their sales by $22.88 per item.

**2. We also want to analyze if sales vary by city type: Tier 1, 2 and 3. For this analysis, you have to account for both city and outlet. Again, conduct the necessary analysis in a succinct way and present the results and interpretation of your analysis.** 

What’s immediately obvious from the histogram and box & whisker plot is that not every type of outlet is present in each type of city. Supermarket Type 1 is the only outlet type in all three types of city, and Tier 3 cities are the only ones that have all types of outlets. It’s less obvious, however, the effect that city type has on the sales of each of the outlet types.

**Feature Selection**

The features for this question are the same as the first, with the addition of city\_type.

|  |  |  |
| --- | --- | --- |
| **Predictor** | **Effect** | **Reasoning** |
| Outlet\_Type | +/- | As shown in the histogram and box & whisker plot, the type of outlet will likely have a measurable effect on sales |
| Item\_MRP | + | By definition, higher priced items will aggregate to higher total sales. |
| City\_Type | +/- | Outlets in Tier 1 cities should have higher sales due to their larger populations and possible higher-income residents. The effect is reversed for Tier 3 cities. |

**Models**

reC1 <- lmer(item\_sales ~ item\_mrp + outlet\_type + city\_type + (1 | outlet\_id), data=df, REML=FALSE)

feC1 <- lm(item\_sales ~ item\_mrp + outlet\_type + outlet\_id + city\_type, data=df)

feC2 <- lm(item\_sales ~ item\_mrp + outlet\_type\*city\_type + outlet\_id , data=df)

=================================================================================================================

Dependent variable:

--------------------------------------------------------------------

item\_sales

linear OLS

mixed-effects

(1) (2) (3)

-----------------------------------------------------------------------------------------------------------------

item\_mrp 15.561\*\*\* (0.196) 15.561\*\*\* (0.196) 15.561\*\*\* (0.196)

outlet\_typeSupermarket Type1 1,950.432\*\*\* (56.828) 2,016.471\*\*\* (60.530) 2,000.079\*\*\* (61.492)

outlet\_typeSupermarket Type2 1,642.621\*\*\* (77.428) 1,642.126\*\*\* (60.555) 1,642.126\*\*\* (60.555)

outlet\_typeSupermarket Type3 3,370.367\*\*\* (77.362) 3,369.871\*\*\* (60.470) 3,369.871\*\*\* (60.470)

outlet\_idOUT013 -66.917 (52.305) -50.525 (86.269)

outlet\_idOUT017 5.961 (52.389) 172.085\*\*\* (52.404)

outlet\_idOUT018

outlet\_idOUT019 16.391 (68.604)

outlet\_idOUT027

outlet\_idOUT035 46.531 (52.335) 212.655\*\*\* (52.348)

outlet\_idOUT045 -166.124\*\*\* (52.347)

outlet\_idOUT046 -97.893\* (52.334) -97.893\* (52.334)

outlet\_idOUT049

outlet\_typeSupermarket Type1:city\_typeTier 2

outlet\_typeSupermarket Type2:city\_typeTier 2

outlet\_typeSupermarket Type3:city\_typeTier 2

outlet\_typeSupermarket Type1:city\_typeTier 3

outlet\_typeSupermarket Type2:city\_typeTier 3

outlet\_typeSupermarket Type3:city\_typeTier 3

city\_typeTier 2 11.253 (49.549) -166.124\*\*\* (52.347)

city\_typeTier 3 -17.396 (55.032) -16.391 (68.604)

Constant -1,834.447\*\*\* (60.116) -1,851.295\*\*\* (55.304) -1,834.904\*\*\* (56.259)

-----------------------------------------------------------------------------------------------------------------

Observations 8,523 8,523 8,523

R2 0.563 0.563

Adjusted R2 0.563 0.563

Log Likelihood -72,001.890

Akaike Inf. Crit. 144,021.800

Bayesian Inf. Crit. 144,085.200

Residual Std. Error (df = 8512) 1,128.491 1,128.491

F Statistic (df = 10; 8512) 1,097.556\*\*\* 1,097.556\*\*\*

=================================================================================================================

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)

reC1 9 144022 144085 -72002 144004

feC1 12 144010 144095 -71993 143986 17.409 3 0.0005821 \*\*\*

feC2 12 144010 144095 -71993 143986 0.000 0 1.0000000

The three models above include one random-effects model and two fixed-effects models. reC1 and feC1 include city\_type as a standard predictor, while feC2 uses a compound feature of city\_type combined with outlet\_type.

The effects for city\_type are either missing or labeled non-significant with the exception of “city\_TypeTier 2” in the feC2 model. This coefficient says that Tier 2 cities can expect to have $166.12 lower sales than Tier 1 cities.

The issues with these models probably stem from the limited nature of the data as explained earlier. The lack of observations for every combination of outlet\_type and city\_type means that it is difficult to to analyze the effects of outlet\_type and city\_type together in the same model.

## 

## Raw Code

#ISM6317 Assignment 5

#Austin Johnson

library(tidyverse)

library(readxl)

library(ggplot2)

library(stargazer)

library(AER)

library(MASS)

library(car)

library(lmtest)

library(lattice)

library(corrplot)

library(PerformanceAnalytics)

library(lme4)

setwd("C:/Users/Ajohnson/Downloads")

df <- read\_excel("BigMartSales.xlsx", sheet = "Data")

names(df) <- tolower(names(df))

str(df)

summary(df)

#Replace inconsistently formatted values

df$item\_fat\_content[df$item\_fat\_content == "low fat"] <- "Low Fat"

colSums(is.na(df)) # Check for missing data, item\_weight and outlet\_size have missing values

# The weight of the item should be irrelavent to sales so throw it out

df$item\_weight <- NULL

#too many missing values in outlet\_size so we'll throw it out

#it also seems to be closely correlated with outlet\_type anyway

df$outlet\_size <- NULL

#change appropriate vars to factors

factorcols <- c("item\_fat\_content","item\_type","city\_type","outlet\_type")

df[factorcols] <- lapply(df[factorcols], factor)

# Relevel fat content so Regular is base

df$item\_fat\_content <- relevel(df$item\_fat\_content, "Regular")

str(df)

summary(df)

#---------------------------------------

# Question 1 - Sales by Outlet Type

#---------------------------------------

#Histogram and Density

histogram(~item\_sales, breaks = 20, main = "Item\_Sales Histogram", data=df) #right skewed distribution shows most items low sale amounts

densityplot(~item\_sales, data=df)

histogram(~item\_sales | outlet\_type, data=df, main = "Item\_Sales by Outlet\_Type Hist")

densityplot(~item\_sales | outlet\_type, data=df)

#Box and Whisker

bwplot(item\_sales ~ outlet\_type, data=df, main = "Item\_Sales by Outlet\_Type Box & Whisker")

#XY Plot

xyplot(item\_sales ~ outlet\_type, data=df)

#OLS Model as Control

ols1 <- lm(item\_sales ~ item\_mrp + outlet\_type, data = df)

ols2 <- lm(item\_sales ~ item\_mrp\*outlet\_type, data = df)

stargazer(ols1,ols2, type="text", single.row=TRUE)

#Fixed effects model (controls for between item and between outlet differences)

fe1 <- lm(item\_sales ~ item\_mrp + outlet\_type + outlet\_id, data=df)

fe2 <- lm(item\_sales ~ item\_mrp\*outlet\_type + outlet\_id, data=df)

stargazer(fe1,fe2, type="text", single.row=TRUE)

#Random effects model (controls for within item and within outlet differences)

re1 <- lmer(item\_sales ~ item\_mrp + outlet\_type + (1 | outlet\_id), data = df, REML=FALSE)

# summary(re1)

# fixef(re1)

# ranef(re1)

# coef(re1)

re2 <- lmer(item\_sales ~ item\_mrp\*outlet\_type + (1 | outlet\_id), data = df, REML=FALSE)

re3 <- lmer(item\_sales ~ item\_visibility + item\_mrp + outlet\_type + (1 | outlet\_id), data = df, REML=FALSE)

stargazer(re1, re2, fe2, type="text", single.row=TRUE)

anova(re1, re2, fe2)

#---------------------------------------

# Question 2 - Sales by City Type

#---------------------------------------

#Histogram and Density

histogram(~item\_sales | city\_type + outlet\_type, data=df)

densityplot(~item\_sales | city\_type, data=df)

#Box and Whisker

bwplot(item\_sales ~ city\_type | outlet\_type, data=df)

#XY Plot

xyplot(item\_sales ~ outlet\_type | city\_type, data=df)

#OLS Model as Control

olsC <- lm(item\_sales ~ item\_mrp + outlet\_type + city\_type, data = df)

summary(olsC)

#Fixed effects model (controls for individual-level differences)

feC1 <- lm(item\_sales ~ item\_mrp + outlet\_type + outlet\_id + city\_type, data=df)

feC2 <- lm(item\_sales ~ item\_mrp + outlet\_type\*city\_type + outlet\_id , data=df)

#summary(feC1)

#Random effects model (controls for within-block differences)

#reC1 <- lmer(item\_sales ~ item\_mrp + outlet\_type\*city\_type + (1 | outlet\_id) + (1 | city\_type), data=df, REML=FALSE)

reC1 <- lmer(item\_sales ~ item\_mrp + outlet\_type + city\_type + (1 | outlet\_id), data=df, REML=FALSE)

#summary(reC1)

#ranef(reC1)

#coef(reC1)

reC2 <- lmer(item\_sales ~ item\_mrp + outlet\_type\*city\_type + (1 | outlet\_id) , data=df, REML=FALSE)

reC3 <- lmer(item\_sales ~ item\_mrp + outlet\_type\*city\_type + (1 | outlet\_id/city\_type) , data=df, REML=FALSE)

stargazer(reC1, feC1, feC2, type="text", single.row=TRUE)

anova(reC1, feC1, feC2)