* Examining the dependent variable, item\_sales.

|  |  |
| --- | --- |
|  |  |

* Item\_sales is right skewed for which OLS regression might not be suitable and log of it looks approximately normal which is more suited for OLS regression.
* Diagram

  Description automatically generated with medium confidenceBut the type of data is multi-level. Item, Outlet and City tier level. Hence going with a fixed effect/random effect or a mixed effect model would be the right choice.
* Positive correlation between item visibility and sales which is expected.

|  |  |  |
| --- | --- | --- |
| **Predictor** | **Effect** | **Rationale** |
| *DV: Item\_Sales* | | |
| Item Visibility | + | Products having higher visibility will have higher sales. |
| Item Type | +/- | Day-to-day items like diary, groceries might sell more compared to other items. |
| Item MRP | +/- | Low price items may sell more compared to costly items. |
| Outlet ID | +/- | There could be differences among the outlets (Higher sales/lower sales), and we need it to answer question 3. |
| Outlet Type | +/- | Will tell how the sales differ by outlet type and to answer question 1. |
| Outlet Age | +/- | Calculated. Established outlets might have more sales than newer ones. |
| City Type | +/- | Tier 1 cities might have large outlets and could have higher sales compared to Tier 2 and Tier 3. Also need to answer question 2. |
| Excluded |  |  |
| Item ID | +/- | There could be differences in specific items and their sales, but the Item Type would differentiate between the kinds of items which is more useful. |
| Item weight | n/a | Correlates with Item MRP |
| Item Fat Content | n/a | Fat content of an item might not effect the sales of that item. |
| Outlet Size | n/a | Outlet\_Size is a substitute for Outlet\_ID. |

Models-

fe\_sales <- lm(log(item\_sales) ~ item\_visibility + item\_type + item\_mrp + outlet\_type + outlet\_age + city\_type + outlet\_id, data=bms)

library(lme4)

re\_sales <- lmer(log(item\_sales) ~ item\_visibility + item\_type + item\_mrp + outlet\_type + city\_type + outlet\_age + (1 | outlet\_id), data=bms,REML=FALSE)

bms$item\_sales2 = round(bms$item\_sales, digits = 0)

re\_sales\_glmer<- glmer(item\_sales2 ~ item\_visibility + item\_type + item\_weight + item\_mrp + outlet\_type + city\_type + outlet\_age + (1 | outlet\_id), data=bms,family = poisson(link="log"))

Models Justification-

1. We start off by running a lm model with fixed effects on outlet ID. For comparison between outlets, it compares to a base level (OUT010) and provides the estimates. We have NA estimates for some of the outlets.
2. The 2nd model is a random effects model with outlet as the grouping level because Outlet sales vary from one outlet to another.
3. The 3rd model is a glmer which is for poison models. Since sales is a function of quantity and price in which quantity is a poison variable, sales must also be a poison variable. We round off the sales to make it an integer and run glmer on it with outlet ID as the group effect.

Stargazer output-

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

Dependent variable:

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

log(item\_sales) item\_sales2

OLS linear generalized linear

mixed-effects mixed-effects

(1) (2) (3)

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

item\_visibility -0.050 (0.118) -0.047 (0.118) -0.176\*\*\* (0.006)

item\_typeBreads 0.028 (0.040) 0.027 (0.040) 0.008\*\*\* (0.002)

item\_typeBreakfast -0.067 (0.056) -0.067 (0.055) -0.031\*\*\* (0.003)

item\_typeCanned 0.025 (0.030) 0.025 (0.030) -0.004\*\*\* (0.001)

item\_typeDairy -0.071\*\* (0.030) -0.070\*\* (0.030) -0.066\*\*\* (0.001)

item\_typeFrozen Foods -0.055\* (0.028) -0.054\* (0.028) -0.045\*\*\* (0.001)

item\_typeFruits and Vegetables -0.005 (0.026) -0.005 (0.026) -0.012\*\*\* (0.001)

item\_typeHard Drinks -0.028 (0.042) -0.029 (0.042) -0.015\*\*\* (0.002)

item\_typeHealth and Hygiene 0.004 (0.032) 0.004 (0.032) -0.016\*\*\* (0.001)

item\_typeHousehold -0.034 (0.028) -0.033 (0.028) -0.037\*\*\* (0.001)

item\_typeMeat 0.024 (0.034) 0.024 (0.034) -0.010\*\*\* (0.002)

item\_typeOthers -0.005 (0.046) -0.006 (0.046) -0.011\*\*\* (0.002)

item\_typeSeafood 0.005 (0.070) 0.004 (0.070) 0.157\*\*\* (0.003)

item\_typeSnack Foods -0.003 (0.026) -0.002 (0.026) -0.005\*\*\* (0.001)

item\_typeSoft Drinks -0.027 (0.033) -0.027 (0.033) -0.016\*\*\* (0.002)

item\_typeStarchy Foods -0.049 (0.049) -0.048 (0.049) -0.010\*\*\* (0.002)

item\_weight -0.0004\*\*\* (0.0001)

item\_mrp 0.008\*\*\* (0.0001) 0.008\*\*\* (0.0001) 0.007\*\*\* (0.00000)

outlet\_typeSupermarket Type1 1.655\*\*\* (0.164) 1.935\*\*\* (0.026) 2.066\*\*\* (0.061)

outlet\_typeSupermarket Type2 1.540\*\*\* (0.140) 1.755\*\*\* (0.051) 1.587\*\*\* (0.061)

outlet\_typeSupermarket Type3 2.784\*\*\* (0.165) 2.508\*\*\* (0.036)

outlet\_age -0.022\* (0.012) -0.002 (0.002) -0.015\*\*\* (0.005)

city\_typeTier 2 -0.146\*\*\* (0.054) -0.015 (0.028) -0.092\*\* (0.036)

city\_typeTier 3 -0.320\* (0.165) -0.033 (0.025) 0.158\*\*\* (0.059)

outlet\_idOUT013 0.524\* (0.302)

outlet\_idOUT017 -0.037 (0.067)

outlet\_idOUT018

outlet\_idOUT019

outlet\_idOUT027

outlet\_idOUT035 0.053 (0.035)

outlet\_idOUT045

outlet\_idOUT046

outlet\_idOUT049

Constant 5.040\*\*\* (0.351) 4.456\*\*\* (0.054) 4.820\*\*\* (0.036)

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

Observations 8,523 8,523 7,060

R2 0.721

Adjusted R2 0.720

Log Likelihood -6,800.317 -1,634,165.000

Akaike Inf. Crit. 13,652.630 3,268,381.000

Bayesian Inf. Crit. 13,835.950 3,268,553.000

Residual Std. Error 0.538 (df = 8496)

F Statistic 845.880\*\*\* (df = 26; 8496)

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

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

The AIC and BIC score indicate that lmer model fits the data well compared to glmer model. Hence, we will be interpreting based on re\_sales model.

ranef(re\_sales)

$outlet\_id

(Intercept)

OUT010 -3.697109e-03

OUT013 3.697109e-03

OUT017 6.010053e-03

OUT018 8.407592e-14

OUT019 3.697109e-03

OUT027 1.200758e-13

OUT035 1.949737e-02

OUT045 -2.550743e-02

OUT046 -1.169135e-02

OUT049 7.994245e-03

Testing Assumptions-

vif(re\_sales)

GVIF Df GVIF^(1/(2\*Df))

item\_visibility 1.066466 1 1.032698

item\_type 1.024725 15 1.000814

item\_mrp 1.012675 1 1.006318

outlet\_type 6.851140 3 1.378141

city\_type 4.259918 2 1.436648

outlet\_age 4.433776 1 2.105653

Multicollinearity Passed.

testDispersion(re\_sales)

DHARMa nonparametric dispersion test via sd of residuals fitted vs. simulated

data: simulationOutput

dispersion = 1.0008, p-value = 0.904

alternative hypothesis: two.sided

Overdispersion- Passed.

dwtest(resid ~ 1)

Durbin-Watson test

data: resid ~ 1

DW = 2.0085, p-value = 0.6518

alternative hypothesis: true autocorrelation is greater than 0

Independence- Passed.

What type of outlet will return him the best sales: Grocery store or Supermarket Type 1, 2, or 3.

According to the fixed effects estimates in the linear mixed model, the outlet type that returns the best sales is "Supermarket Type3" with an estimate of 2.508. This is followed by "Supermarket Type1" with an estimate of 1.935 and "Supermarket Type2" with an estimate of 1.755. So, among the given outlet types, "Supermarket Type3" has the highest potential for sales.

What type of city will return him the best sales: Tier 1, 2 or 3.

According to the fixed effects estimates in the linear mixed model, the city type does not have a significant effect on sales as the estimates for "Tier 2" and "Tier 3" are both negative and close to zero (1.5 and 3.3 % less respectively compared to Tier 1) . This means that the city type of Tier 1 have a higher impact on sales but not by much.

What are the top 3 highest performing and lowest performing stores in the sample.

Based on the output, the top 3 highest performing stores in the sample are OUT035 with random effects estimate of 0.0195, OUT017 with an estimate of 0.0060, and OUT049 with an estimate of 0.008. The lowest performing stores are OUT045 with an estimate of -0.0255, OUT046 with an estimate of -0.0117, and OUT010 with an estimate of -0.0037.