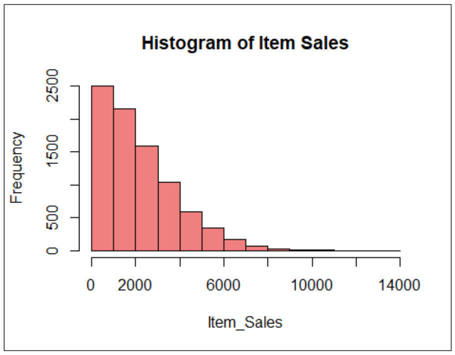
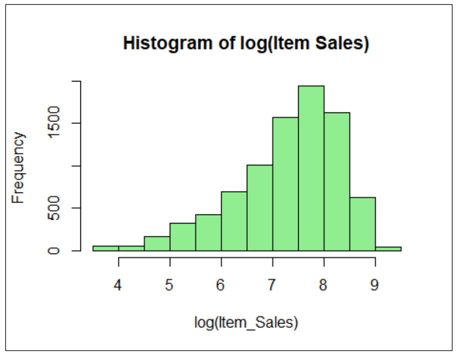
Retail Item Sales Prediction:

I. Create three reasonable models to do the analysis and present your analysis in a compact and succinct manner. Try to ensure that each of your three models answers all three questions and fits the multi-level nature of this model.

Item_Sales: My client is a business entrepreneur considering franchising one or more stores. Item_Sales is the sales data of the products from the retail chain based on multiple factors such as store type, city type, item type etc.





- Dependent variable (item_sales) seems to follow a non-normal distribution. Data is hierarchical so we should use Hierarchical Linear Models. Applying a log transform to the DV, it seems to follow a normal distribution.
- Item_weight and item_size have missing values so these columns can be omitted from our analysis. This is a two-level data.

II. In your document, explain clearly any assumptions that you make, and choice of predictors, models, and/or assumptions.

Factor Effects:
Positive Effect: If there is a direct proportionality (+X then +y / -X then -y) between the predictor variable (X) and the response variable (y), we can say there is a positive effect.

Negative Effect: If there is an inverse proportionality (+X then -y / -X then +y) between the predictor variable (X) and the response variable (y), we can say there is a negative effect.

Relevant Factors for Item_Sales		
Predictor	**Effect**	**Rationale**
Level_2		
Outlet_Year	+	This variable can be converted to age where it makes more sense as the older the store, the more customers it gets due to customer retention.
City_Type	+/-	This depends on the population. If there is a significant difference in the population between the cities, then it could have a massive impact on sales but if they are similar then the marketing efforts on either of these could yield the same results.
Outlet_Type	+/-	This variable talks about the size of the outlets. Supermarkets are larger than a grocery store and has better product visibility as most of the items are well organized on shelfs.
Level_1		
Item_Fat_Content	-	Fat content is measured as low fat and regular which could influence sales as people tend to be calorie conscious.
Item_Visibility	+	As mentioned in the store type variable, better product visibility could lead to better sales.
Item_Type	+	Products are categorized into several types based on their content. In this case we have 16 types of products. This could have a positive effect as more products means more choice for the customers.
Item_MRP	-	Higher MRP could have a negative effect as most consumers are cost conscious and would usually look for cheaper alternatives.

Irrelevant Factors		
Predictor	**Effect**	**Rationale**
Item_ID	No Effect	This variable is an ID. Can be omitted.
Item_Weight	Missing Entries	Since we have missing data on about 1400 entries, including this variable could skew our results.
Outlet_Size	Missing Entries	Store size correlates with store type. But there are about 2000 missing entries in the data which does not give a clear picture as supermarket 1 has all three sizes mixed in. So omitted this variable.

III. Interpret your findings based on the BEST of your three models, with a set of actionable recommendations for the business entrepreneur.

```
#Regression models
m4 <- lmer(item_sales ~ item_type + item_visibility + item_mrp + city_type +
             outlet_type + outlet_age + (1 | city_type/outlet_id), data = bms)

m5 <- lmer(log(item_sales) ~ item_type + item_fat_content + item_visibility + item_mrp + city_type +
             outlet_type + outlet_age + (1 | city_type/outlet_id), data = bms)

m6 <- lmer(item_sales ~ item_type + item_fat_content + item_visibility + item_mrp + city_type +
             outlet_type + outlet_age + (1 | city_type/outlet_id), data = bms)
```

```
#Stargazer
stargazer(m4, m5, m6, type='text', single.row = TRUE)
```

```
##
## ===============================================================================================
##                                                        Dependent variable:
##                                  --------------------------------------------------------------
##                                       item_sales         log(item_sales)         item_sales
##                                          (1)                   (2)                   (3)
## -----------------------------------------------------------------------------------------------
## item_typeBreads                    3.160 (84.002)         0.028 (0.040)       5.177 (84.008)
## item_typeBreakfast                13.056 (116.537)       -0.069 (0.056)       7.718 (116.589)
## item_typeCanned                    24.891 (62.764)        0.025 (0.030)       25.624 (62.762)
## item_typeDairy                    -45.323 (62.097)      -0.069** (0.030)     -40.982 (62.166)
## item_typeFrozen Foods             -28.748 (58.848)      -0.054* (0.028)      -28.013 (58.847)
## item_typeFruits and Vegetables     29.184 (54.912)       -0.005 (0.026)       29.357 (54.908)
## item_typeHard Drinks              -20.212 (89.071)       -0.022 (0.043)       -0.181 (90.142)
## item_typeHealth and Hygiene       -30.890 (66.568)        0.011 (0.032)      -10.976 (67.982)
## item_typeHousehold                -59.690 (58.246)       -0.027 (0.029)      -39.694 (59.871)
## item_typeMeat                      4.078 (70.568)         0.022 (0.034)       -0.362 (70.631)
## item_typeOthers                   -42.495 (97.581)        0.001 (0.047)      -22.573 (98.549)
## item_typeSeafood                  181.635 (147.989)       0.005 (0.070)      184.527 (147.993)
## item_typeSnack Foods              -14.260 (55.189)       -0.002 (0.026)      -11.448 (55.220)
## item_typeSoft Drinks              -40.886 (69.529)       -0.022 (0.033)      -27.384 (70.153)
## item_typeStarchy Foods            19.317 (102.970)       -0.048 (0.049)      21.259 (102.972)
## item_fat_contentregular                                   0.014 (0.013)       40.685 (28.224)
## item_visibility                  -293.149 (248.615)      -0.052 (0.118)     -302.488 (248.683)
## item_mrp                           15.566*** (0.198)    0.008*** (0.0001)    15.565*** (0.198)
## city_typeTier 2                   -17.264 (275.420)      -0.014 (0.234)     -16.680 (194.947)
## city_typeTier 3                   -13.932 (268.250)      -0.035 (0.233)     -14.272 (184.668)
## outlet_typeSupermarket Type1    1,930.103*** (96.531)  1.935*** (0.043)   1,929.481*** (96.571)
## outlet_typeSupermarket Type2   1,576.256*** (189.514)  1.758*** (0.085)  1,576.709*** (189.592)
## outlet_typeSupermarket Type3   3,372.959*** (138.821)  2.507*** (0.062)  3,372.357*** (138.880)
## outlet_age                         -2.914 (7.546)        -0.002 (0.003)       -2.864 (7.549)
## Constant                       -1,731.272*** (258.723) 4.447*** (0.183)  -1,750.857*** (219.584)
## -----------------------------------------------------------------------------------------------
## Observations                          8,523                 8,523                 8,523
## Log Likelihood                     -71,883.450           -6,872.745           -71,878.150
## Akaike Inf. Crit.                  143,820.900           13,801.490           143,812.300
## Bayesian Inf. Crit.                144,011.300           13,998.910           144,009.700
## ===============================================================================================
## Note:                                                              *p<0.1; **p<0.05; ***p<0.01
```

The assumptions test for our models are multicollinearity and autocorrelation.

Assumption	DV Model: m6 (Item_Sales)
Multicollinearity: Passed • **Variance Inflation Factor (VIF – GVIF^(1/(2*Df)))** 1. **VIF = 1/T (Where T = 1 – R^2 , T < 0.1 is indicative of multicollinearity).** 2. **VIF > 5 indicates multicollinearity. VIF > 10 is strong evidence of multicollinearity.**	(R output below) **This model passed the VIF test.**

```
vif(m6)

##                     GVIF Df GVIF^(1/(2*Df))
## item_type        1.240235 15        1.007203
## item_fat_content 1.216101  1        1.102770
## item_visibility  1.027003  1        1.013411
## item_mrp         1.012666  1        1.006313
## city_type        1.455465  2        1.098375
## outlet_type      3.440377  3        1.228667
## outlet_age       3.107040  1        1.762680
```

Assumption	DV Model: m6 (Item_Sales)
Independence: Passed • **Durbin-Watson's Test (DW)** 1. **Ho: Residuals are not linearly auto-correlated.** 2. **DW ~ [0, 4]; values around 2 (i.e., 1.5 to 2.5) suggests no autocorrelation.**	**There is an error when I run the Durbin-Watson test for auto-correlation, but since the data does not follow a pattern we can safely assume this condition is met as well.**

[Selected Model = m6] – [Chose lmer model without log as there isn't much difference in the sign of the Beta-Coeff]

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

- From our analysis, Supermarket Type3 returns the best sales.

	Grocery Store in Sales
When considered for outlet_id and city_type level difference, Supermarket Type1 makes	$1929.5 (more than)
When considered for outlet_id and city_type level difference, Supermarket Type2 makes	$1576.7 (more than)
When considered for outlet_id and city_type level difference, Supermarket Type3 makes	$3372.3 (more than)

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

- From our analysis, Tier1 city returns the best sales.

	Tier 1 in Sales
When considered for outlet_id and city_type level difference, Tier 2 city makes	$16.6 (less than)
When considered for outlet_id and city_type level difference, Tier 3 city makes	$14.2 (less than)

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

- From our analysis,

Stores in city type		Sales (Negatives)
Top 3 Highest Performing Stores		
1	Store OUT035 in Tier 2 city makes	($1920.67) less in sales
2	Store OUT017 in Tier 2 city makes	($1961.77) less in sales
3	Store OUT049 in Tier 1 city makes	($1962.67) less in sales
Top 3 Lowest Performing Stores		
1	Store OUT045 in Tier 2 city makes	($2098.67) less in sales
2	Store OUT046 in Tier 1 city makes	($2040.67) less in sales
3	Store OUT010 in Tier 3 city makes	($2009.67) less in sales

```
ranef(m6)

## $`outlet_id:city_type`
##                  (Intercept)
## OUT010:Tier 3 -1.600991e+01
## OUT013:Tier 3  1.600991e+01
## OUT017:Tier 2  3.190131e+01
## OUT018:Tier 3 -3.472036e-11
## OUT019:Tier 1  1.600991e+01
## OUT027:Tier 3  1.474984e-10
## OUT035:Tier 2  7.332675e+01
## OUT045:Tier 2 -1.052281e+02
## OUT046:Tier 1 -4.704390e+01
## OUT049:Tier 1  3.103398e+01
##
## $city_type
##          (Intercept)
## Tier 1  1.063993e-09
## Tier 2 -2.319427e-10
## Tier 3  3.505110e-09
##
## with conditional variances for "outlet_id:city_type" "city_type"
```

Recommendations:

1. Invest in Type3 Supermarkets with coverage in Tier1 cities (Tier1 cities could be densely populated).
2. Regular products seem to contribute positively to sales than low fat products. So, recommended to invest in regular products.
3. Products like Canned, Sea Foods, Fruits and Veggies seem to contribute to the sales quite a bit, so recommended to invest in these products.