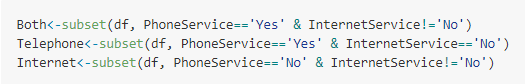
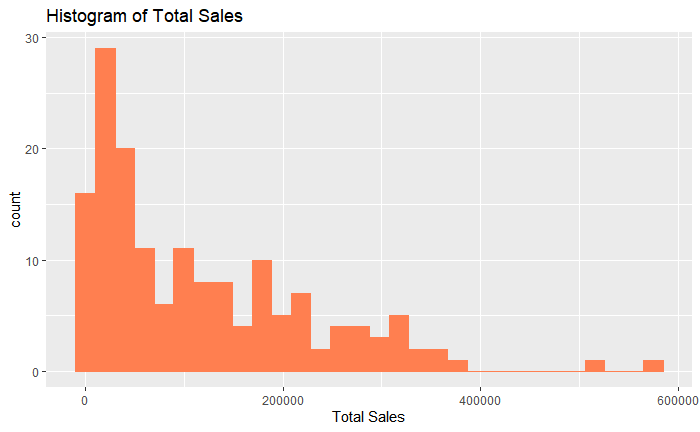
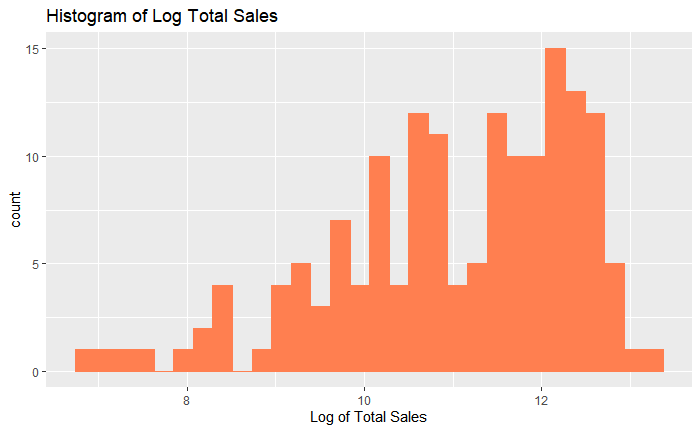
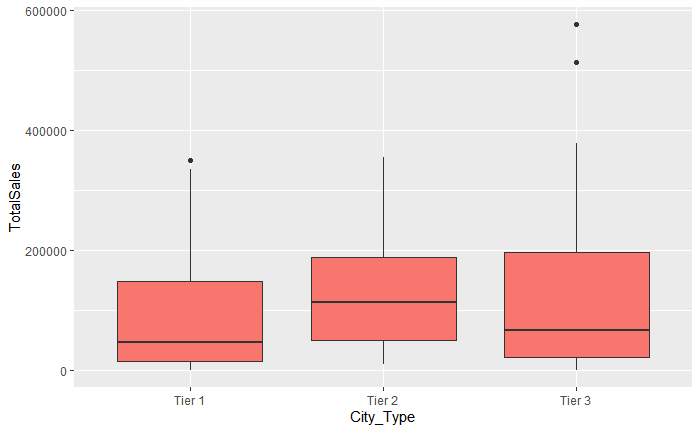
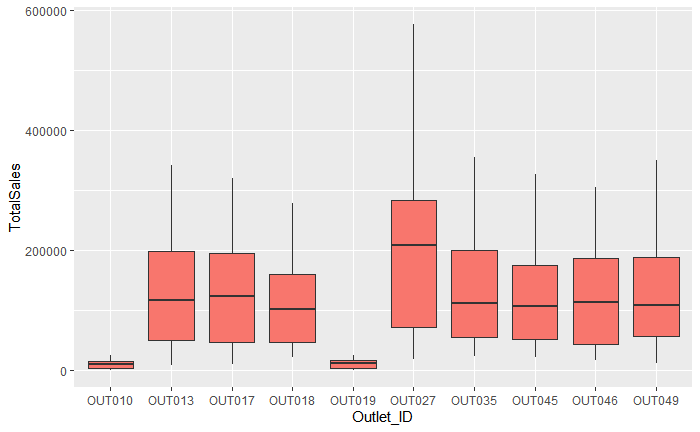
Assignment 7- Telco Churn

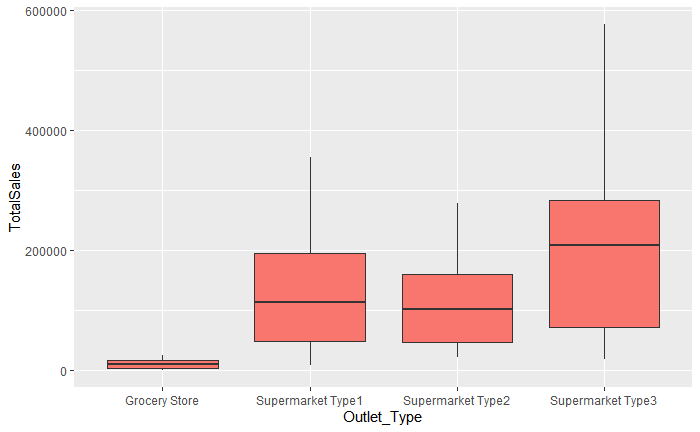
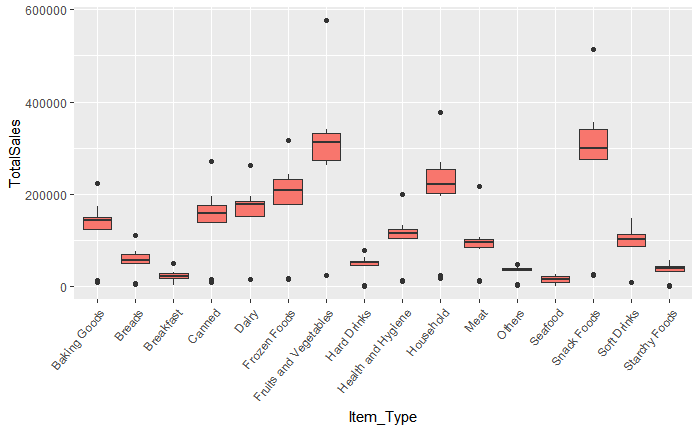
Submitted by: Hammad Muniem

1. I did the following steps:
   1. I checked for nulls in the main data set. There were 11 nulls so I removed those rows completely.
   2. Encoded Churn column
   3. I made separate data sets for Telephone only, Internet only and for customers using both services using the following block of code. 
   4. Then I converted all categorical variables to factors in the 3 data sets.
   5. Dropped irrelevant columns from each data set. For example, all internet related columns were dropped in the telephone only dataframe.

The data consists of item level information but the analysis that we are asked to do is Multi-Level. Therefore this requires some data transformation to reduce the granularity to atleast an item type.

After transformation, we have a column called TotalSales which would be our dependent variable. We then studied the distribution of our dependent variable. It doesn’t follow a normal distribution so we did a Log transform. The resultant distribution resembles a normal distribution more.

Following that, I did a few visualizations to get an idea of how total sales vary by different factors.



|  |  |  |
| --- | --- | --- |
| Predictor Variable | Expected sign of effect | Rationale |
| Item Type | Both | We can see from the boxplots that there are a few items like fruits and vegetables that are high sellers. Although the client has not asked for these, including this variable would improve the model fit and will result in unbiased estimates for the important variables. |
| Outlet Type | Positive | We expect the supermarkets to have higher sales than grocery stores. |
| City Type | Positive | We expect city tier 2 to give us higher sales than the other 2 city tiers. |
| Outlet\_ID | Both | There are certain stores like Outlet27 that have higher sales than the rest of the stores. We expect this variable to cause some variability in Total Sales at the store level as well due to factors not covered in this data. |

All Item level variables have not been included because the client wants an upper level analysis without any information on Items.  
Store variables such as Outlet\_size, Age of store and Outlet\_Year have not been included because the client wants store level information i.e. we have to include Outlet\_ID which includes size and age of the outlet.

Among the models posted, model 3 was the best according to anova tests and AIC.

**Type of outlet:**The type of outlet with the highest sales keeping everything else constant is Supermarket Type 3. Supermarket Type 3 will get 290% more sales than grocery store, 54% more sales than Supermarket Type 2 and 44% more sales than Supermarket Type 1 keeping all other factors constant.

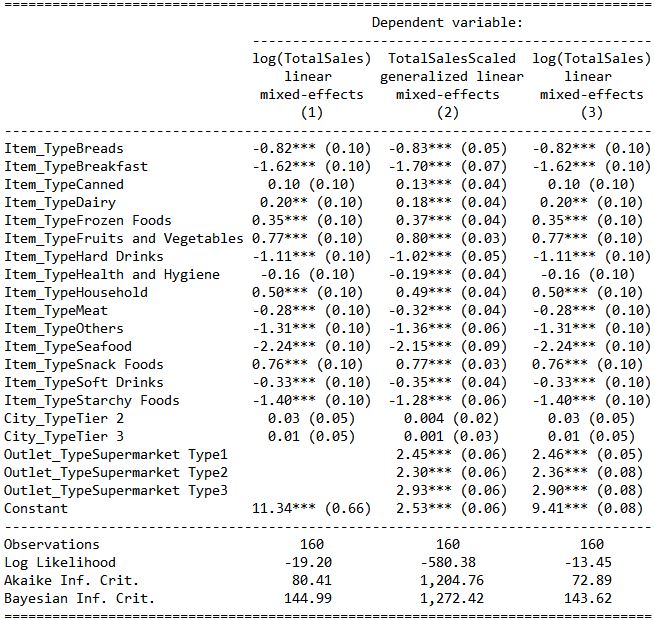
**City Tier:**City Tier 2 had the highest sales keeping all other factors constant. City Tier 2 had 3% higher sales than City Tier 1 and 2% higher sales than City Tier 3 keeping all other factors constant.

**Highest and lowest Performing Outlets:**At the store level, the economic significance is really low meaning that most of the variation has been accounted for by the other variables. Therefore, the variation caused by the lower level outlet itself is really low. However, I have still written down the best performing and least performing stores.  
The best performing outlets keeping all other factors constant were: (in order of best performing)

1. Outlet 35
2. Outlet 10
3. Outlet 49

The least performing outlets keeping all other factors constant were: (in order of least performing)

1. Outlet 13
2. Outlet 19
3. Outlet 17



**Recommendations:**

It is recommended that the entrepreneur invest in Supermarket Type 3 in city that is in City Tier 2. A bonus bit of advice would be stock up more on Snack foods and Fruits and Vegetables since they have the most sales as opposed to seafood which is expected to have the lowest sales.