**Regork Grocery Chains Report**

**By: Moureen Mathenge and Shehryar Naveed**

Dear Regork Grocery Chains:

**Introduction**

We received compiled data from 127 Regork Stores to analyze and provide suggestions to help the company function better. Detailed below are some action steps backed up with an analysis that we think would benefit the company.

**How the economic status of customers affects sales**

Regork Grocery Stores attract both low-income and high-income communities. However, some stores have higher ratios of either high-income or low-income customers. We used a randomization test to test the hypothesis that there is a difference in sales prices between the store with the lowest number of high-income customers (Store ID 288) and the store with the highest number of high-income customers (Store ID 406).  We found that the p-value < 0.0001. This means that the difference observed is outside two standard deviations of the mean difference. We, therefore, concluded that there is significant statistical evidence that there is a difference in mean sales prices between the two stores.  For the store to increase its sales and profits, it can target each of the low and high-income districts with more products that are specifically demanded by each. This will lead to more of their products sold which in turn increases the profits.

We made no assumptions using the randomization test.

**Brand and Department Association.**

Regork sells both national and private branded products, on average, the company stocks more national brands than private brands. The distribution of products is as follows:

|  |  |
| --- | --- |
| Private | National |
| 0.3073248 | 0.6926752 |

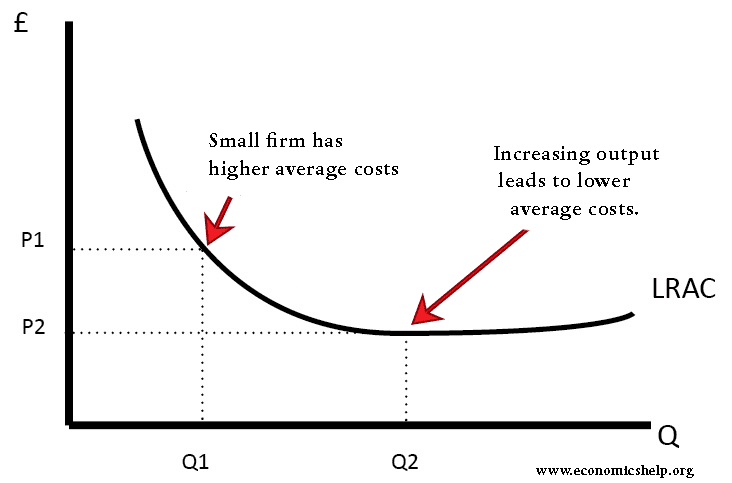
 However, we were interested in whether there is a difference in the expected and observed number of private and national brands within each of the departments. We conducted a chi-squared test statistic using a contingency table to test the hypothesis that there was an association between brand and department. We used categorical variables ( Department and Brand) to run the chi-squared test and the observations are independent of each other. The table below shows our findings.

|  |  |  |
| --- | --- | --- |
| **Chi-squared test statistic** | **Degrees of freedom** | **P-value** |
| 2482.31856 | 7 | <0.0001 |

Since the p-value< 0.0001, we concluded that there is significant statistical evidence that there is an association between brand and department. The distribution of private and national brands is not as expected across all the departments. The most variance in expected and observed was in grocery, pharmacy, meat, and produce departments. In grocery, there were more private brands than expected while in the other three departments there were fewer private brands than expected.

***Suggestions:***

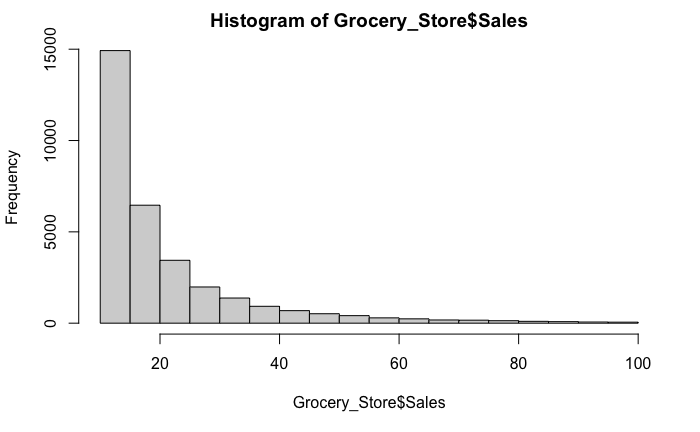
Depending on the objectives of Regork Grocery Chains, we recommend that Regork should work towards stocking more private products in its grocery stores. By owning more of the products they sell, they will have a higher profit margin due to the following reason: If they produce the products themselves, it is going to cost them much lower than buying the product from some other supplier. This all now relates to the concept of economies of scale- the bigger the production process gets, the lower your costs get. Hence, as the cost per unit decreases, the profit per unit increases.



As can be seen from the above graph, the larger the firm gets, the lower its cost gets. Hence, Regork investing in other lines of production will end in them benefiting in the long run with increased profits.

**Inventory and merchandise management**

Good inventory management can result in either losses or profits. We explored the nature of sales within Regork Stores to better understand the needs of the customers and the needs of the business. After modeling all sales observed in the data, we discovered that sales are ***exponentially distributed*** as shown in the diagram below.



As per the distribution, most of the sales of products appear to be below 20. We can use this model to predict the rate of decrease in sales frequency and calculate the risk that Regork is willing to take by stocking more or fewer products on the shelf. As this data is collected over a period of one week, it informs the company on how frequently or how much to stock items that may be perishable such as produce. Overstocking produce may lead to losses due to the products going bad.

This model can also be used to infer how to allocate space within the stores to maximize meeting the needs of customers while increasing sales and profit. After looking at the sales distribution for each department, we found that they all follow an exponential distribution model. Then if a department tends to have higher sales relative to another department, it would be profitable to expand the department to allow better flow and possibly more sales.

**Marketing Improvements**

***Introduction:***

We used Binomial distribution to check the probability of a certain number of products being advertised using the Mailer variable.

***Conducting the test:***

We found that 9.6% of products in the Regork Store Chains population are advertised. We randomly picked Store 286 and used a binomial distribution with its product number, 60, to find out how many products are being advertised and the results were very shocking.

***Findings:***

We ran a binomial distribution model with a sample size of store 286. Advertising 30% of their products is a bare minimum for each grocery store however, store 286 have only been advertising 13%. We calculated the probability of them advertising 30% or more products and the value we got was 0.000441 which is a very low value- this goes on to show that, with how it's functioning right now, store 286 cant compete with the rest of the competition in marketing.

***Suggestions:***

Our first solution is to advertise more. Upon advertising more, the people will know about most of the products available at the store beforehand which would likely drive the people to the store. In order to do that, they should be looking towards hiring a new marketing manager who has a more aggressive approach towards advertising. Another suggestion would be to diversify their advertising- they should look towards advertising more from social media- your sponsored post is likely to reach everyone around the area of the store which is better marketing than mailing advertisements.

**Maximizing Sales**

***Introduction:***

We decided on using the t-test to analyze if the average value of retail coupons is consistent with all the stores for the grocery store chain.

***Conducting the test:***

We tested the hypotheses below:

H0: True difference between the mean of retail coupons in two different stores is 0.

H1: True difference in the means is not 0.

***Findings:***

After conducting the test, we arrived at a p-value of 0.07428 which means that we strongly fail to reject the null hypothesis. This means that the average value of retail coupons in both the stores is the same however, we have a suggestion that might improve this number and lead to greater sales.

***Interpretation:***

|  |  |  |
| --- | --- | --- |
| P-value | Degrees of freedom | Confidence interval |
| 0.07428 | 525.28 | (-0.22,  0.01) |

We conclude that there is no difference in the mean amounts of retail coupons offered in stores 288 and 406. We are 95% confident that the true difference in mean retail coupons is between -0.22 and 0.01. We assume normality in retail coupons for both stores, after testing this assumption we conclude that the data is sufficiently normally distributed and thus we can infer from the findings *(Figures below)*.  Therefore, the following are some of the recommendations we have for Regork based on this:

***Suggestions:***

Since the difference between means is zero, the average value of retail coupons is the same in both stores. As calculated above using a randomization test, we concluded that there is a difference between the sale price of products between these two stores. In order for the lower performing store to perform better, it needs to incentivize the customers with having more retail coupons- a discount would attract more customers and in turn increase sales and profits.

*Assumption of Normality for Sales Prices in Store ID 406*

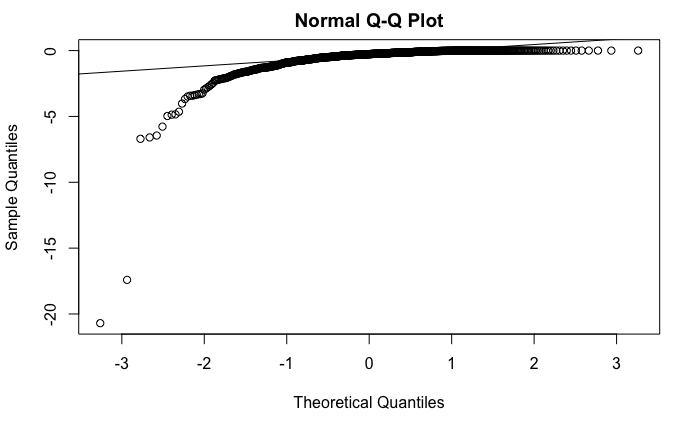


Figure 1:  The data points follow the QQline roughly with the tails trailing off.

*Assumption of Normality for Sales Prices in Store ID 288*

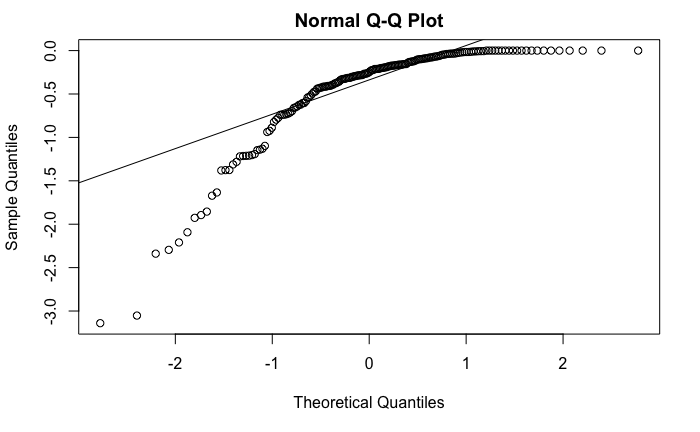


Figure 2:  The data points follow the QQline roughly with the tails trailing off.

**Targeting products/services to fit demographics**

We ran a simple regression model with the number of high-income individuals predicting the children coming into the store. We found an R-Squared value of 0.1625 which indicates the model is a good fit. Our interpretation of the data: the more high-income individuals there are, the greater the number of children who visit the store. Specifically, by increasing the number of high-income individuals by 1, the number of children increases by 0.394. This makes sense as the high-income individuals are more financially stable which allows them to have a greater number of kids as well. There is also a trend in the world with high-income individuals taking their children to the grocery stores with them. Through this, the store can predict the children coming in by knowing which district they are located in and how many high-income people come in. Then based on this, the stores can diversify their product selection and start selling more children-suited products- this will lead to an increase in sales and hence, an increase in profits. We checked assumptions of unbiased and homoskedascity as detailed by the plots below, we concluded that the assumptions held.

|  |  |  |
| --- | --- | --- |
| **R-Squared** | **Intercept** | **Slope** |
| 0.1625 | 7722.94 | 0.394 |

Simple Linear Regression Model 1 Assumptions

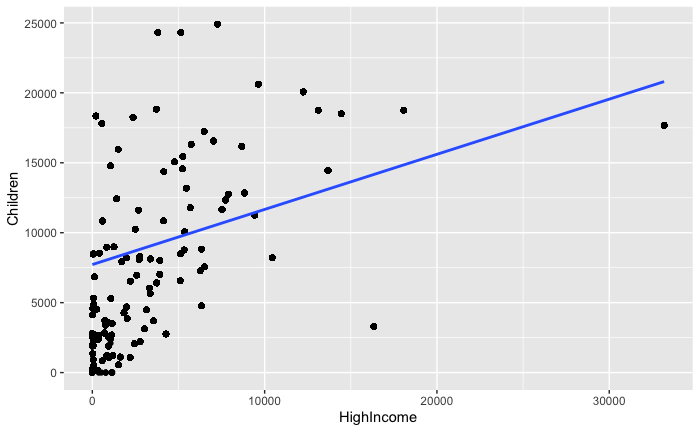


Figure 3: Linear Regression Model predicting number of children using high income.

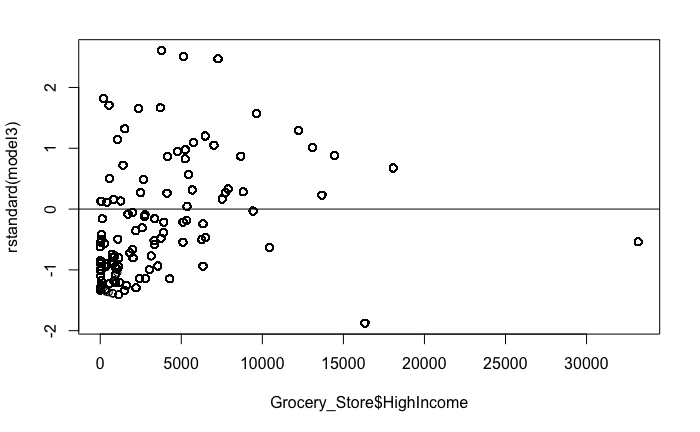


Figure 4: The residuals are randomly distributed, there is no clear pattern in the residuals.

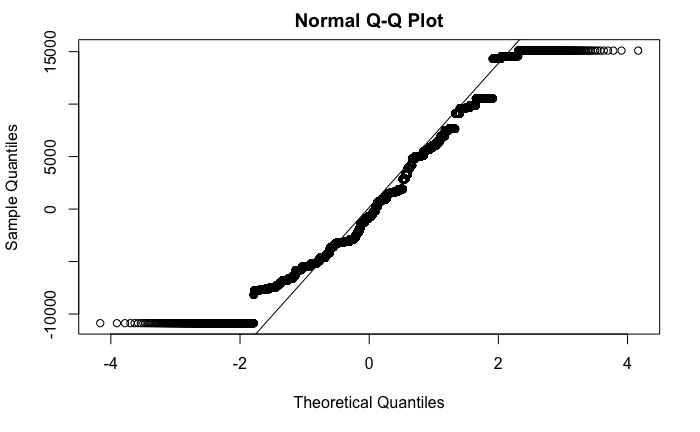


Figure 5:  The data points follow the QQline roughly with the tails trailing off.

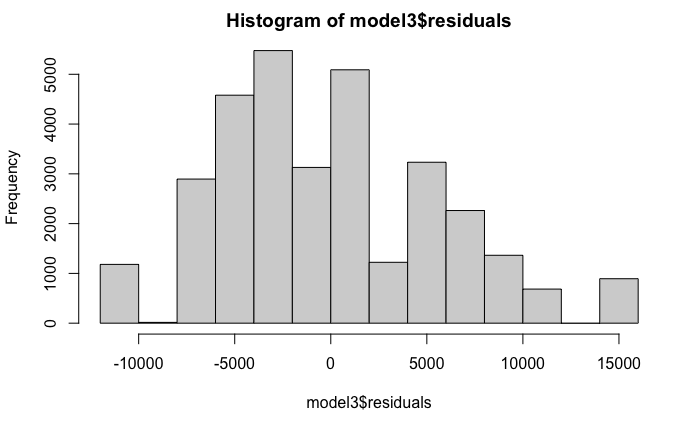


Figure 6: The histogram has roughly a bell curve shape with a bit of skew and outliers.

**Optimal Pricing of products**

For the multiple regression model, we predicted the sales price from department, display, retail coupon, and brand. The R-Squared value we got is 0.24678 which indicates that the model is a good fit. First, we chose the department in this model because the price of a product is highly dependent on what department it belongs to the price of a product in electronics will be higher compared to the price of some products in soft drinks. Secondly, we used retail coupons for the following reason- the higher the retail coupon, the lesser the selling price so it directly influences the price the product is sold at. Thirdly, we used the predictor variables display and brand because, in our opinion, both of these variables can heavily influence the sales price- if a product is at the display, it is likely to have a more attractive price and if a product is privately sold, it is also likely to influence the price as there will be no third party commission costs in obtaining the product. The grocery store can decide the sales price of any product by running it through this regression model. This model is going to help them decide the best possible price which will earn them the highest possible profit. We checked assumptions of unbiased and homoskedascity as detailed by the plots below, we concluded that the assumptions held.

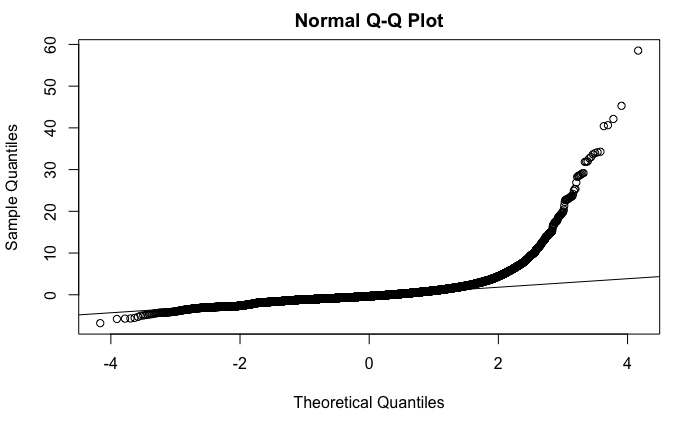


Figure 7:  The data points follow the QQline roughly with the tails trailing off.

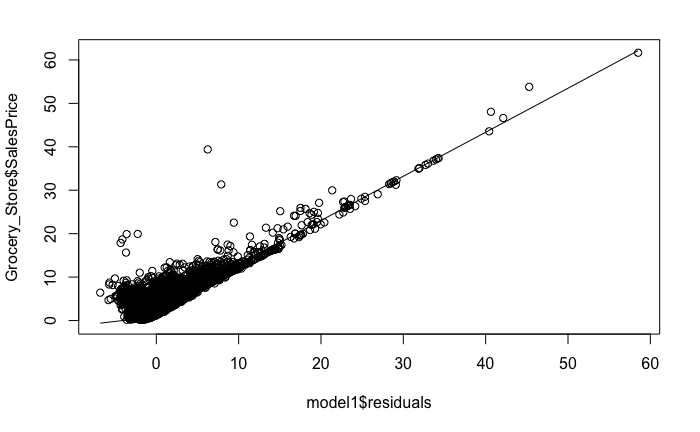


Figure 8: Scatter plot of residuals and response value.

**Conclusion**

We encountered some limitations with the data, largely to do with the low R2 values that we obtained while trying to run the linear regression models. We eventually used the highest R2  values that we could find, this translated to the data not being the best fit but it still provided an interpretation that was valuable.

Since that data set contained real-world data we expected to encounter such limitations and also expected results that were not the most ideal.