M5 Forecasting - Data Exploratory Analysis

Bryan Florence, Jordan Robles, Dustin Vasquez

05/04/2020

### Abstract

The M5 Forecasting data is a set of datasets provided by Walmart to Kaggle for a competition in forecasting the quantity of sales a month out. The link for the competition page, where the datasets can be found is here:

<https://www.kaggle.com/c/m5-forecasting-accuracy>

Walmart captured the sales of their retail goods from Jan 2011 to June 2016. The variables that have been tracked are: time, state, department, category, item, cost of item, quantity of items purchased, holiday, holiday type, acceptance of SNAP food stamps. This data is set up in a hierarchical fashion so that we can look at the information in many ways. We can look at it from the:

* Item Level: 3,490 cases
* Department Level: 7 cases
* Category Level: 3 cases
* Store Level: 10 cases
* State Level: 3 cases

The dataset comes from three separate files containing the sales information, the price information, and the calendar information. There are 1941 days, or cases, in this data set.

The two variables we will be interested in will be the quantity of items sold and the price of items. Some of the questions we are interested in answering are as follows:

* How does time effect the variables? (i.e. seasonality, weekly, holiday)
* Is there a difference between sales by the different location?
* How about the 3 different categories and their prices in each state?

### Import the Data

The first step in doing our analysis will be importing the data and trying to get an understanding of how it is set up.

#### Train Data

The train data is of size (30490, 1919) with the following names. This data set is set up where the variables are id, item\_id, dept\_id, cat\_id, store\_id, state\_id, d\_1, d\_2, …., d\_1913. The id variable is just a combination of the other 5 id’s.

Here is a quick look at the top 5 rows for some of the columns.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| item\_id | dept\_id | cat\_id | store\_id | state\_id | d\_1 | d\_1909 | d\_1910 | d\_1911 | d\_1912 | d\_1913 |
| HOBBIES\_1\_001 | HOBBIES\_1 | HOBBIES | CA\_1 | CA | 0 | 1 | 3 | 0 | 1 | 1 |
| HOBBIES\_1\_002 | HOBBIES\_1 | HOBBIES | CA\_1 | CA | 0 | 1 | 0 | 0 | 0 | 0 |
| HOBBIES\_1\_003 | HOBBIES\_1 | HOBBIES | CA\_1 | CA | 0 | 1 | 0 | 1 | 1 | 1 |
| HOBBIES\_1\_004 | HOBBIES\_1 | HOBBIES | CA\_1 | CA | 0 | 0 | 1 | 3 | 7 | 2 |
| HOBBIES\_1\_005 | HOBBIES\_1 | HOBBIES | CA\_1 | CA | 0 | 1 | 2 | 2 | 2 | 4 |

Looking at the unique values in the qualitative columns of the dataframe, shown in the table below, gives an idea of what this dataframe is telling us. There are 3049 items spread out through 10 stores in 3 different states. A lot of these items are repeated through the different stores. In each store there are 3 categories with 7 departments.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | item\_id | dept\_id | cat\_id | store\_id | state\_id |
| unique | 3049 | 7 | 3 | 10 | 3 |

The table below is the distribution of 3 random days for all the items through all of the stores and categories. With the Median value being 0, you can see that a lot of these values are zero. This is because they will not always sell all of the items every day, or some may have been discontinued. The amount of zeros doesn’t pose much of a threat since we are aggregating over different variables and time.

|  |  |  |  |
| --- | --- | --- | --- |
|  | d\_193 | d\_1171 | d\_158 |
| Mean | 0.8597245 | 1.511414 | 0.872286 |
| Std.Dev | 3.6145279 | 4.492225 | 3.591203 |
| Min | 0.0000000 | 0.000000 | 0.000000 |
| Median | 0.0000000 | 0.000000 | 0.000000 |
| Max | 106.0000000 | 136.000000 | 91.000000 |

#### Price Data

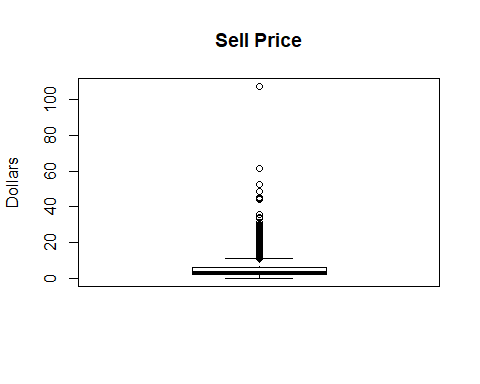
The prices data frame has a size of (6841121, 4) with the following variables; store\_id, item\_id, wm\_yr\_wk, sell\_price. There are a couple of rows that tie this data frame to the train data frame, which are store\_id, item\_id.

Here is a quick look at the head of the dataframe.

|  |  |  |  |
| --- | --- | --- | --- |
| store\_id | item\_id | wm\_yr\_wk | sell\_price |
| CA\_1 | HOBBIES\_1\_001 | 11325 | 9.58 |
| CA\_1 | HOBBIES\_1\_001 | 11326 | 9.58 |
| CA\_1 | HOBBIES\_1\_001 | 11327 | 8.26 |
| CA\_1 | HOBBIES\_1\_001 | 11328 | 8.26 |
| CA\_1 | HOBBIES\_1\_001 | 11329 | 8.26 |
| CA\_1 | HOBBIES\_1\_001 | 11330 | 8.26 |

The only new real interesting variable we get from this dataframe is the “sell\_price” variable, which we can use to calculate revenue later. The sell price is given as a weekly average of the price for that item at that store. The stats and box-plot of this value are shown below.

|  |  |
| --- | --- |
|  | sell\_price |
| Mean | 4.410952 |
| Std.Dev | 3.408814 |
| Min | 0.010000 |
| Median | 3.470000 |
| Max | 107.320000 |



Looking at the box plot, we can see that a lot of the item prices are pretty low with a select few of them reaching over 40 dollars.

#### Calendar Data

The calendar data frame has a size of (1969, 14) with the following variables; date, wm\_yr\_wk, weekday, wday, month, year, d, event\_name\_1, event\_type\_1, event\_name\_2, event\_type\_2, snap\_CA, snap\_TX, snap\_WI.

The column d is the column name that ties this dataset back to the train dataset and represents d\_0001, d0002, … , d1913.

By looking at the first and last values in the data frame, we can see that the time frame from 2011-01-29 to 2016-06-19. This data set also gives data on whether there is an event and what that event type is. It also gives information on whether that day is a SNAP day or not in either of the three states. The number of events and event types are displayed in the table below. The unique descriptor also includes regular days without any events.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | event\_name\_1 | event\_type\_1 | event\_name\_2 | event\_type\_2 | snap\_CA | snap\_TX | snap\_WI |
| count | 162 | 162 | 0 | 0 | 650 | 650 | 650 |
| unique | 31 | 5 | 1 | 1 | 2 | 2 | 2 |

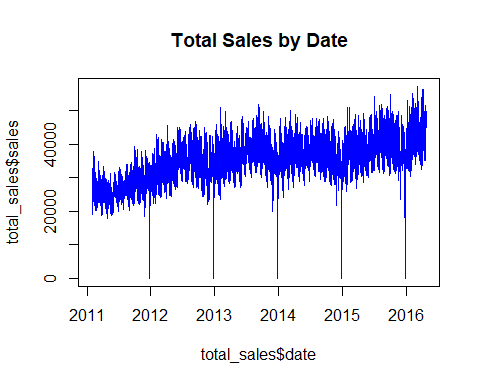
There are not a lot of events throughout the year that this keeps track of, just 30 unique events in all of the 5 plus years. Each state only has a total of 650 SNAP days because they are required to have so many in 1 year.

The “wm\_yr\_wk” column tracks the number of weeks in a year but it does it in a tricky way. It start the 1st week on the first day of data collection, and starts it count there. So week 52 of the first year will actually be around 2012-01-30. This can be seen with the following table, which looks at the 1st 5 values of the first year and the second year.

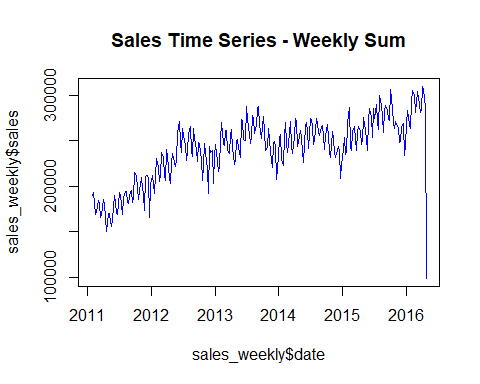
|  |  |
| --- | --- |
| date | wm\_yr\_wk |
| 2011-01-29 | 11101 |
| 2011-01-30 | 11101 |
| 2011-01-31 | 11101 |
| 2011-02-01 | 11101 |
| 2011-02-02 | 11101 |
| 2012-01-30 | 11201 |
| 2012-01-31 | 11201 |
| 2012-02-01 | 11201 |
| 2012-02-02 | 11201 |
| 2012-02-03 | 11201 |

## Exploratory Analysis

By aggregating all of the sales and pivoting the training table we can see the amount sales for each day in all 10 of the stores. The following time series chart shows how the total sales change per day.



Seasonality in quantity of sales over the years can be seen. The drops at the end of every year are Christmas where the stores are closed for part of the day. We can also look at this weekly instead of daily to reduce the noise by summing up the total sales in a week, which is displayed below.

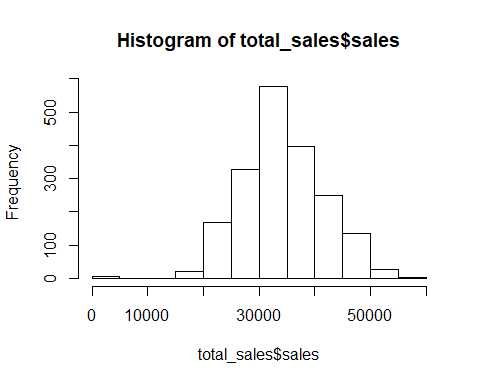


One thing we can dive deeper into is how the sales correlates to different aspects of times, such as year, month, week, week day. The following is the correlation between sales and those factors.

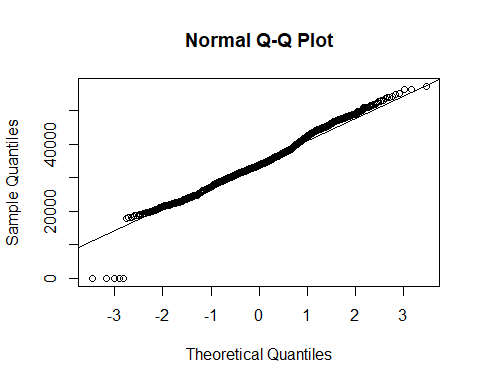
|  |  |
| --- | --- |
|  | r |
| sales | 1.0000000 |
| year | 0.5311059 |
| month | -0.0037429 |
| week | -0.0348703 |
| wday | -0.4544096 |

The two factors that correlate the most with sales are year and weekday, and even those are not great. Month and week number in the year have pretty much no correlation.

The following is a dive into the distribution of the sales data and determining whether to use parametric or non-parametric analysis. It can be seen in the histogram below that the data has good shape with some potential outliers.



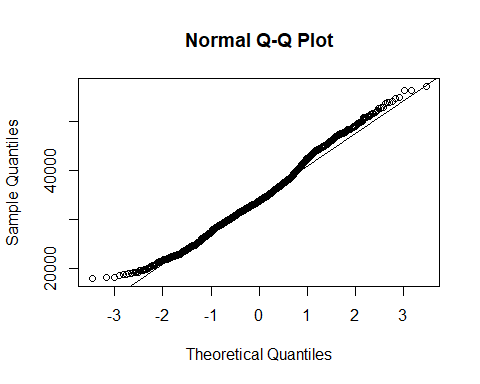
Looking at the Normal Q-Q Plot below, it can definitely be seen that there are some outliers in the data. Outside of that the data appears normal.



Doing a Shapiro-Wilks to do a quick check on the normality of the sales data gives us the results below.

## Shapiro-Wilk normality test  
##   
## data: total\_sales$sales  
## W = 0.98562, p-value = 6.226e-13

It can be seen from the test that the data is not normal at all. But lets try and remove the outliers and recheck the assumptions for parametric models. The outliers that need to be removed fall on Christmas days, where there are significantly less sales on this event than on any other day. The Q-Q Plot for sales without the outliers is below, followed by the Shapiro-Wilk normality test.

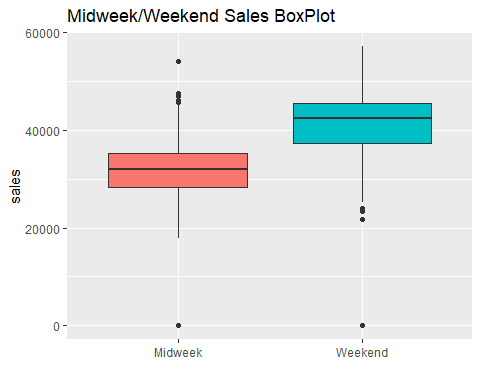


## Shapiro-Wilk normality test  
##   
## data: X$sales  
## W = 0.98945, p-value = 1.424e-10

It can be seen that the data, even with the removed outliers, is not normal by the p-value being so small. This means statistical analysis on the sales part of the data will have to use non-parametric methods.

#### Comparison of Midweek Sales vs. Weekend Sales

The following is a box plot for midweek and weekend sales. It appears that there are more sales on Saturday and Sunday compared to the 5 other days.



One of our questions of interest was if weekend sales differed from weekday sales. Leading us to the following hypothesis:  
H0: µmidweek = µweekend

Ha: µmidweek ≠ µweekend

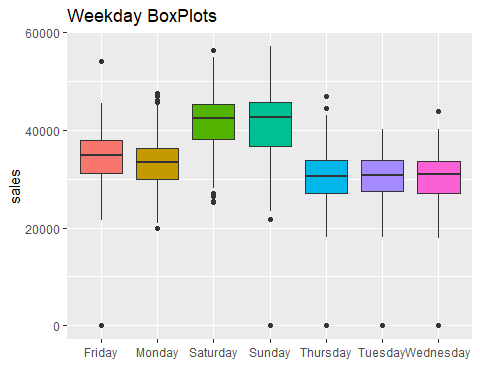
The p-value is <.05 so we reject the null hypothesis and claim that there is a difference in sales between midweek days and the weekend.

## Kruskal-Wallis rank sum test  
##   
## data: total\_sales$sales and total\_sales$weekend  
## Kruskal-Wallis chi-squared = 652.91, df = 1, p-value < 2.2e-16

The results from this test shows that we have enough evidence to support our claim that there is a difference between the sales on the weekend and midweek. This leads us to exploring which days have the least amount of sales and which days have the most.

### Comparison of Sales and Weekday

The following plot is a box plot of the days in the week and sales. By looking at this plot, it appears that there are less sales between Tuesday, Wednesday, and Thursday than in the other days of the week. It also appears that the most shopping is done on Saturday and Sunday.



This proposes the following hypothesis:

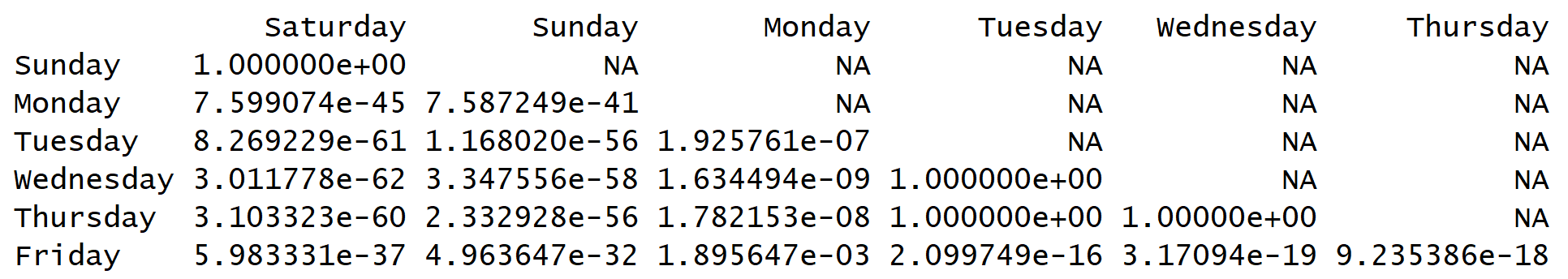
The below table is the descriptive stats for the days in the week. Just by looking at the descriptive stats, it appears that there is some difference between the day of the week means.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Mean** | **Std\_Dev** | **Count** | **Median** |
| Friday | 34225.99 | 5602.040 | 273 | 34775 |
| Monday | 32852.97 | 5223.532 | 273 | 33444 |
| Saturday | 41546.89 | 6120.411 | 274 | 42437 |
| Sunday | 41130.02 | 6997.728 | 274 | 42586 |
| Thursday | 30205.01 | 5342.799 | 273 | 30658 |
| Tuesday | 30368.78 | 5088.428 | 273 | 30710 |
| Wednesday | 30010.02 | 5164.106 | 273 | 30911 |

Looking at the Kruskal-Wallis test for the days of the week, we can see that there is enough evidence to reject the null hypothesis and know there is a difference for at least one of the days of the week.

## Kruskal-Wallis rank sum test  
##   
## data: X$sales and X$weekday  
## Kruskal-Wallis chi-squared = 756.21, df = 6, p-value < 2.2e-16

The Wilcox test was applied to look at the relationship between the individual weekdays and their respective significance levels. The table below has been created and shows the p-value for the Wilcox test between the row and column days. Since multiple tests were completed, the p-value has been adjusted vie the Bonferroni method.



From this table, you can see that Saturday and Sunday are significantly different from every other day in the week, but are practically equal against each other. It also looks like the sales on Tuesday, Wednesday, and Thursday are not different from each other, but those three days are different from all of the others. Using this table and the above plots, a conclusion can be made that if a crowd is desired, shop on Saturday and Sunday. If the goal is to avoid the crowd, as much as possible, then shop on Tuesday, Wednesday and Thursday.

#### Comparison of Sales and Event Type

Next, the different types of events are looked at and compared between each other to see if there is any specific event that stands out from a normal day. Below is a list of the descriptive statistics for the average sales per day for following types of events. It appears that the event type ‘National’ may have the largest difference in mean sales per day.

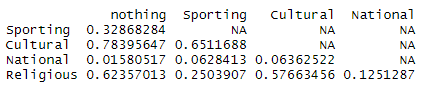
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Event Type** | **Mean** | **Std\_Dev** | **Count** | **Median** |
| Cultural | 34234.74 | 7380.082 | 35 | 35089 |
| National | 29458.51 | 12117.064 | 51 | 32656 |
| nothing | 34489.21 | 7133.702 | 1759 | 33714 |
| Religious | 33760.69 | 7069.625 | 52 | 33760 |
| Sporting | 35796.06 | 6651.927 | 16 | 34998 |

The following hypothesis can be stated:

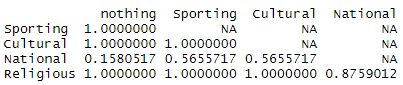
The following is the Krusdal-Wallis.

##   
## Kruskal-Wallis rank sum test  
##   
## data: sales.events$sales and sales.events$event\_type\_1  
## Kruskal-Wallis chi-squared = 7.1705, df = 4, p-value = 0.1271

According to this test, there is not enough evidence to reject the null hypothesis that the sales on events is different from the sales on a day with no event. Below is a table of p-values using the Wilcox test for the p-values. Looking at this table, it appears that National is significantly different from a day with no events. This table was created without an adjustment.



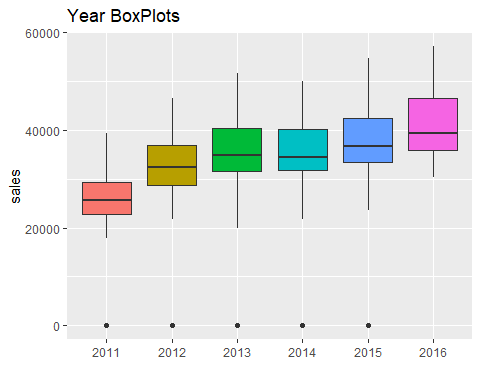
The following table was created with an adjustment and the significance has been lost between the National events and days with no events.



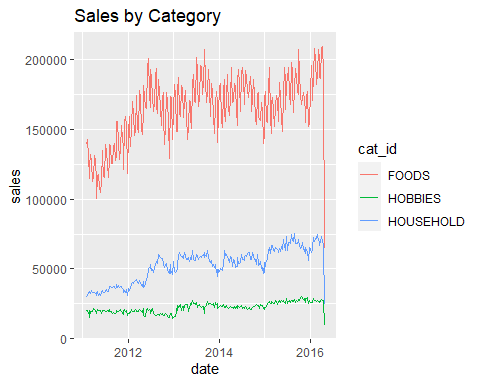
In conclusion, there is not enough evidence to support a claim that any events have a significantly different sales than a day with no events.

#### Look at Sales and Year

This next chart shows the box plots of the sales through each year.

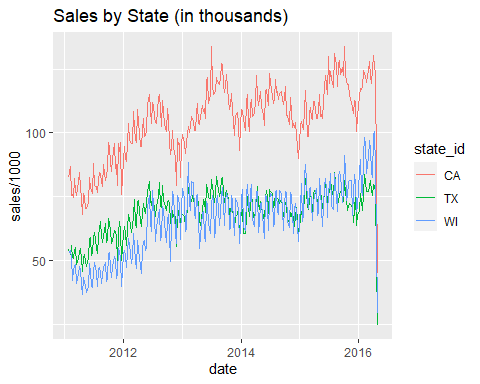


This appears to show that there is a steady increase in sales as the years progress. The next graph shows a time series of sales by each category: food, household, and hobbies.



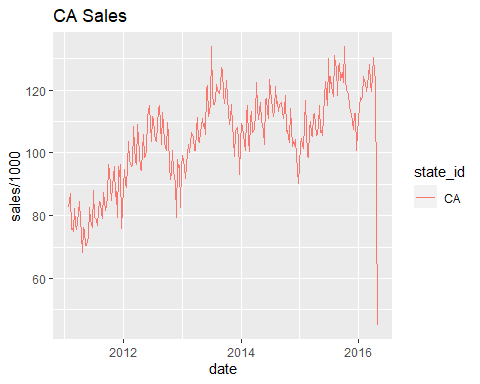
### Comparison of States

The next graph shows a time series of sales by each category: California, Texas, and Wisconsin.



Looking at the graph, it almost appears that every state has its own rate for increase in sales per time. It appears that CA and WI may be similar but TX seems to be less steep than the other two. We will also like to compare the rates of sales increase in each step by looking at the following.

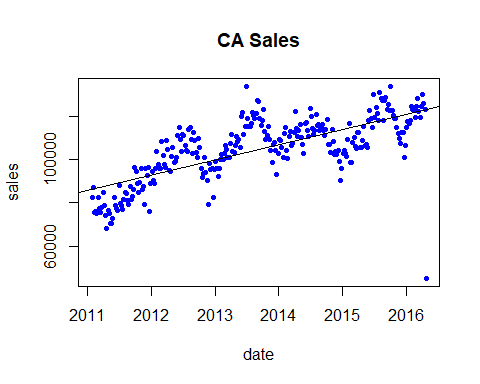
The following is a look at weekly sales by state. First is CA.



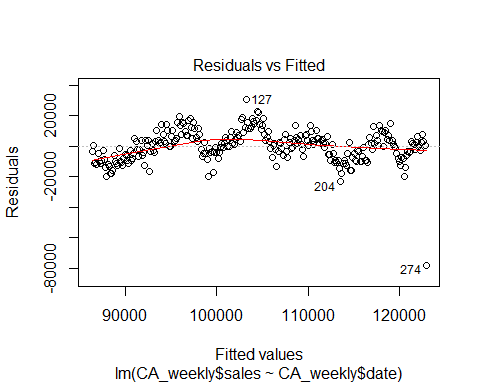
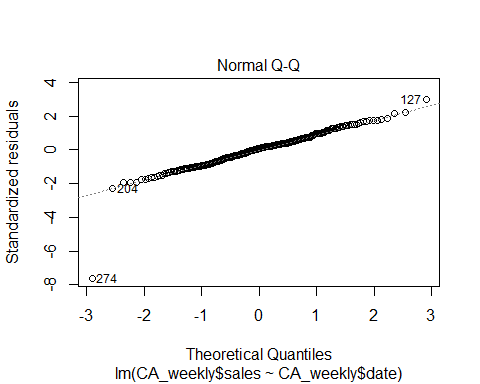
A linear regression gives us our and coefficients, as displayed in the following table.

##   
## Call:  
## lm(formula = CA\_weekly$sales ~ CA\_weekly$date, data = CA\_weekly)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -77919 -6363 876 5937 30469   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.007e+05 1.796e+04 -11.18 <2e-16 \*\*\*  
## CA\_weekly$date 1.914e+01 1.125e+00 17.01 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10310 on 272 degrees of freedom  
## Multiple R-squared: 0.5155, Adjusted R-squared: 0.5137   
## F-statistic: 289.4 on 1 and 272 DF, p-value: < 2.2e-16

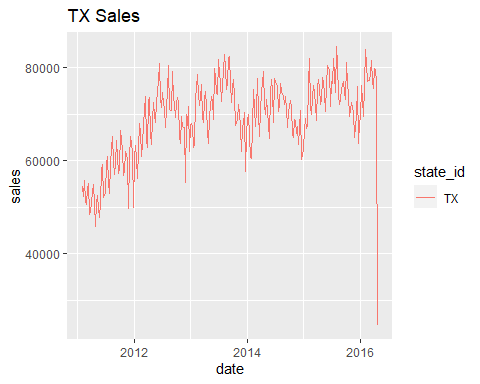
We can also see our fitted line with the following chart. It is obvious from this plot that Sales data follows a cyclical/seasonal pattern.



The following two plots shows that CA sales are normally distributed with a fairly equal variance. Index 274 seems to be a bit of an outlier, corresponds to the week of May 23, 2016, the final week in the analysis.



Next is a look at weekly sales in Texas.

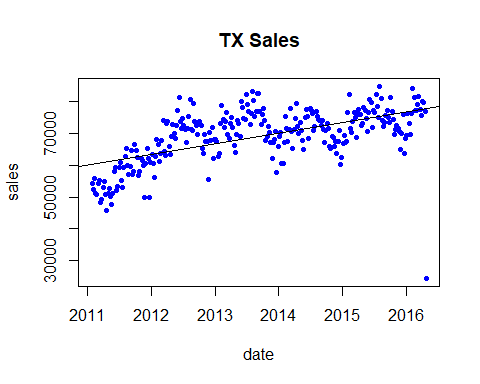


The following is the linear model table for Texas.

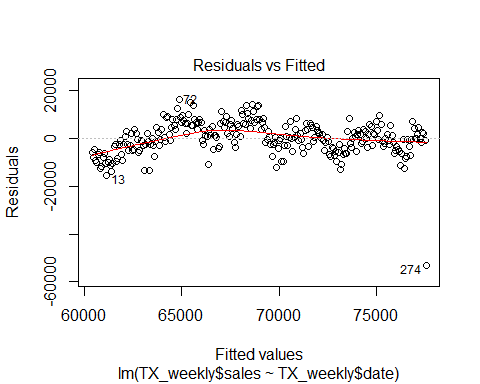
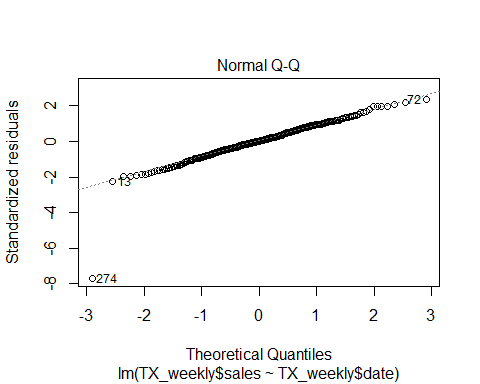
##   
## Call:  
## lm(formula = TX\_weekly$sales ~ TX\_weekly$date, data = TX\_weekly)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -53013 -3732 69 4466 16196   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -7.443e+04 1.210e+04 -6.153 2.71e-09 \*\*\*  
## TX\_weekly$date 8.986e+00 7.576e-01 11.862 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6943 on 272 degrees of freedom  
## Multiple R-squared: 0.3409, Adjusted R-squared: 0.3385   
## F-statistic: 140.7 on 1 and 272 DF, p-value: < 2.2e-16

Texas’ slope is 8.986286 and California’s slope is 19.1362012.

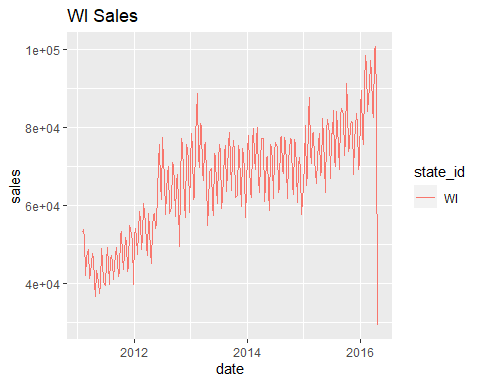
The following is the plot of the fitted line. Again, a seasonal sales pattern is obvious in the plot.



The following two plots shows that TX sales are normally distributed with a fairly equal variance. As with California, Index 274 corresponds to the week of May 23, 2016, the final week in the analysis.



The following is a look at Wisconsin’s weekly sales.

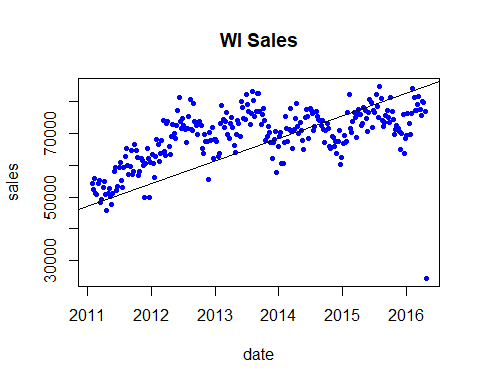


The following is the linear model table for Wisconsin.

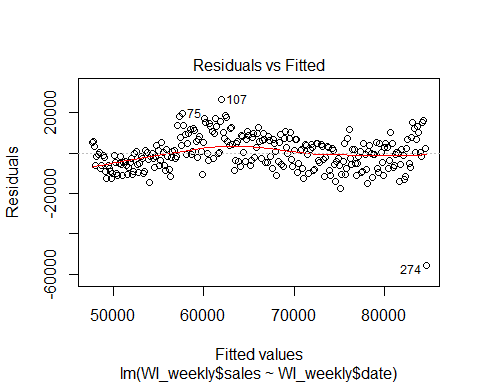
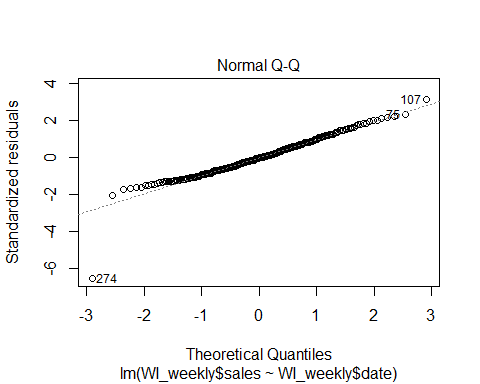
##   
## Call:  
## lm(formula = WI\_weekly$sales ~ WI\_weekly$date, data = WI\_weekly)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -55450 -5777 -285 5433 26518   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.429e+05 1.489e+04 -16.31 <2e-16 \*\*\*  
## WI\_weekly$date 1.936e+01 9.327e-01 20.76 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8548 on 272 degrees of freedom  
## Multiple R-squared: 0.6131, Adjusted R-squared: 0.6117   
## F-statistic: 431 on 1 and 272 DF, p-value: < 2.2e-16

Texas’ slope is 8.986286, California’s slope is 19.1362012, and Wisconsin’s slope is 19.3650254.

The following is plot of the fitted line.



The following two plots shows that WI sales are normally distributed with a fairly equal variance.



The next step will be doing the Hypothesis test on the difference in the coefficients for each of these states.

The null hypothesis is that the β1 coefficients between the 3 states are equal.

H0: β1 California = β1 Texas= β1 Wisconsin

Ha: At least one β1 is different.

From the linear regression, the 95% confidence interval for California’s β1 coefficient is (21.35, 16.92)

## 2.5 % 97.5 %  
## (Intercept) -236093.69411 -165365.87862  
## CA\_weekly$date 16.92154 21.35086

The 95% confidence interval for Texas’ β1 coefficient is (10.48, 7.49)

## 2.5 % 97.5 %  
## (Intercept) -98248.701163 -50617.63879  
## TX\_weekly$date 7.494842 10.47773

The 95% confidence interval for Wisconsin’s β1 coefficient is (21.20, 17.53)

## 2.5 % 97.5 %  
## (Intercept) -272224.43024 -213580.06296  
## WI\_weekly$date 17.52873 21.20132

Therefore, the decision is to reject the null hypothesis and conclude that at least one β1 coefficient is different than the others. In this particular case, Texas’ β1 coefficient differs from the other two states since its confidence interval does not overlap with the other two.

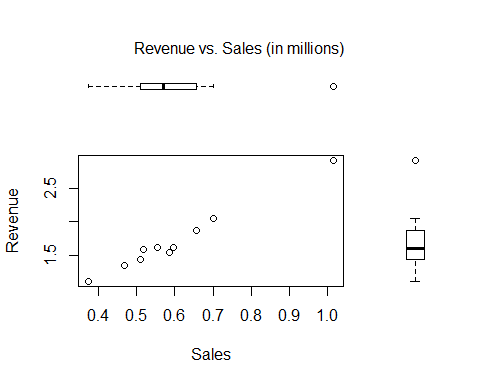
### Comparison of Category Price by State

The next thing we want to look at is the sales, revenue, and average price of the ten stores compares to each other.

The following table shows the average yearly sales and revenue for each store.

|  |  |  |
| --- | --- | --- |
| store\_id | sales | revenue |
| CA\_1 | 699837.8 | 2047296 |
| CA\_2 | 516861.4 | 1582677 |
| CA\_3 | 1017107.3 | 2918975 |
| CA\_4 | 373061.5 | 1109741 |
| TX\_1 | 508662.9 | 1429924 |
| TX\_2 | 655853.1 | 1865294 |
| TX\_3 | 553575.5 | 1617884 |
| WI\_1 | 468096.5 | 1342658 |
| WI\_2 | 594910.2 | 1608249 |
| WI\_3 | 584343.8 | 1538810 |

The following graph show the scatter plot with box plots of this data. Our next steps will be to do a multivariate analysis of the data and see if we find anything significant.



This next table shows the basic descriptive statics for sales and revenue for all of the stores. Each store sells on average about 597,231 items brining in about 1.7 million dollars over these 10 stores.

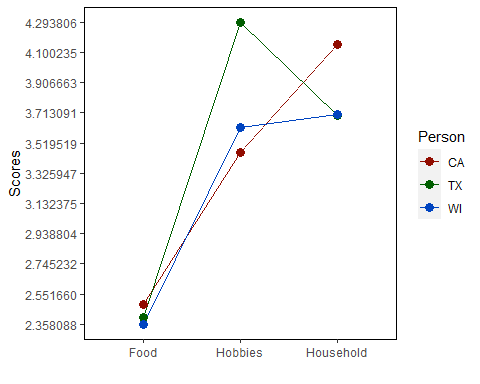
|  |  |  |
| --- | --- | --- |
|  | **sales** | **revenue** |
| Mean | 597231.0 | 1706150.6 |
| Std.Dev | 174346.2 | 498598.7 |
| Min | 373061.5 | 1109740.6 |
| Median | 568959.6 | 1595462.9 |
| Max | 1017107.3 | 2918974.5 |

That is interesting, but what about how much each state compares for each of three categories Food, Hobbies, and Household? The following table is the descriptive statistics for each category in each store.

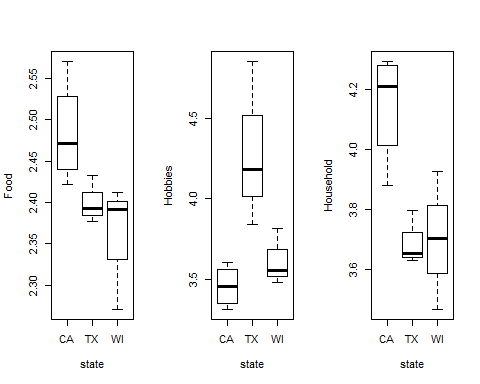
|  |  |  |  |
| --- | --- | --- | --- |
|  | sales | revenue | avg\_price |
| Mean | 199077.00 | 568716.9 | 3.3520975 |
| Std.Dev | 171107.05 | 380413.8 | 0.7372562 |
| Min | 33655.82 | 128554.6 | 2.2707652 |
| Median | 127481.73 | 461423.7 | 3.5398370 |
| Max | 683253.64 | 1654610.9 | 4.8520368 |

In each category, (Food, Hobbies, and Household), sells on average 199077 items and brings on average 57000 dollars for each store.

The following plot is the profile plot of the average sales price in each category per state. It appears according to this that Hobbies are priced higher in Texas than in the other states. It also appears that California prices higher its Household items than the other states. All states have a similar price for food.

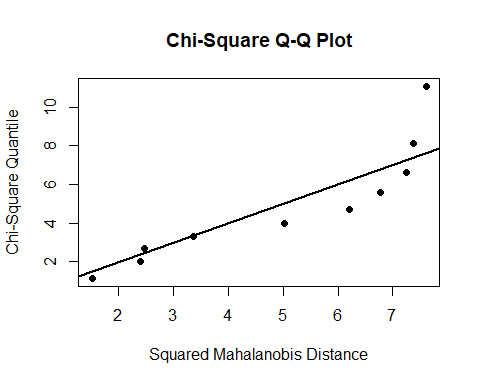


Below are 3 box plots, one for each category, showing the average price for item in each state and how they compare.



According to these plots, there may be a significant difference between the category costs for each state.

The multivariate normal test was applied to determine whether the data had a normal distribution. The below plot is a Q-Q plot of the and the distances. The data appears to follow the distribution quite well, until the end.



Looking at the results from the MVN test, it is shown that the assumption of normality is true.

## $multivariateNormality  
## Test Statistic p value Result  
## 1 Mardia Skewness 42.8050821440668 0.171113756339515 YES  
## 2 Mardia Kurtosis -0.940084372658701 0.347174284063481 YES  
## 3 MVN <NA> <NA> YES  
##   
## $univariateNormality  
## Test Variable Statistic p value Normality  
## 1 Shapiro-Wilk sales 0.8709 0.1023 YES   
## 2 Shapiro-Wilk revenue 0.8436 0.0487 NO   
## 3 Shapiro-Wilk Food 0.9520 0.6925 YES   
## 4 Shapiro-Wilk Hobbies 0.8270 0.0308 NO   
## 5 Shapiro-Wilk Household 0.9367 0.5166 YES   
##   
## $Descriptives  
## n Mean Std.Dev Median Min Max  
## sales 10 5.972310e+05 1.743462e+05 5.689596e+05 3.730615e+05 1.017107e+06  
## revenue 10 1.706151e+06 4.985987e+05 1.595463e+06 1.109741e+06 2.918975e+06  
## Food 10 2.421212e+00 7.771637e-02 2.417037e+00 2.270765e+00 2.570264e+00  
## Hobbies 10 3.757124e+00 4.616413e-01 3.579253e+00 3.313599e+00 4.852037e+00  
## Household 10 3.877956e+00 2.826975e-01 3.839093e+00 3.468583e+00 4.293026e+00  
## 25th 75th Skew Kurtosis  
## sales 5.107125e+05 6.406174e+05 1.14324583 0.6920951  
## revenue 1.457145e+06 1.803442e+06 1.24103067 0.7681932  
## Food 2.391436e+00 2.451135e+00 0.02047398 -0.1726584  
## Hobbies 3.493317e+00 3.838218e+00 1.23035986 0.3869497  
## Household 3.666367e+00 4.095470e+00 0.23155700 -1.5211614

The following Bartlett’s test are to test the assumption of equal variance. Since some of the states had only 3 stores and the number of variables are equal to 3, average price and 3 categories, the Bartlett test were implied instead of the BoxM Test. Looking at the test values, it appears that there is not enough evidence to support the rejection of equal variance for these categories.

## Bartlett test of homogeneity of variances  
##   
## data: Food by state  
## Bartlett's K-squared = 1.408, df = 2, p-value = 0.4946

## Bartlett test of homogeneity of variances  
##   
## data: Hobbies by state  
## Bartlett's K-squared = 4.2441, df = 2, p-value = 0.1198

## Bartlett test of homogeneity of variances  
##   
## data: Household by state  
## Bartlett's K-squared = 1.2964, df = 2, p-value = 0.523

## The following object is masked \_by\_ .GlobalEnv:  
##   
## state

The following shows the results from a MANOVA test, using Pillai’s Trace is displayed below. According to this method, there is enough evidence to support a statistical difference between the state's prices in each category with a p < 0.05.

## Df Pillai approx F num Df den Df Pr(>F)   
## state 2 1.2559 3.3757 6 12 0.0346 \*  
## Residuals 7   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Using the Wilks method, there is still some evidence to support significance between the state's prices in each category, but it not as much significance as Pillai’s Trace. This p value greater than .05 but less than .10.

## Df Wilks approx F num Df den Df Pr(>F)   
## state 2 0.12001 3.1445 6 10 0.05326 .  
## Residuals 7   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

To dig a little deeper, lets look at the individual ANOVA test and their respective p-values which are in the table below. If a significance level of .10 is used, then it can be seen that there is a difference in the states cost for the two categories Hobbies and Household. Food has no evidence to support any difference among the states.

|  |  |
| --- | --- |
| **Category** | **Adjust P-Value** |
| Food | 0.209 |
| Hobbies | 0.064 |
| Household | 0.055 |

### Conclusion

After analyzing the Walmart sales data, we were able to make several conclusions about the questions we set out to answer. Overall, there is enough evidence to conclude that the day of the week affects the amount of sales. The average sales on weekends are higher than on weekdays, with Tuesday, Wednesday, and Thursday having the lowest weekly sales. From the coefficient analysis, it is shown that although all three state’s sales increased in time, Texas had the slowest increase in sales over the years of analysis. It was found that categorical cost of goods differed between states, with California selling Food and Household items at higher prices, and Texas selling Hobby items at higher prices.

Surprisingly, it was found that Sporting events, National Events, Religious Holidays, and other Cultural events have no effect on weekly sales.

#### References

[1] <https://www.kaggle.com/headsortails/back-to-predict-the-future-interactive-m5-eda>

[2] *Applied Multivariate Statistical Analysis, Sixth Edition*

[3] *Dr. Poliak lecture notes, Spring 2020*