DATA70002 – Understanding Data and their Environment

Forecasting Sales from a big US Retailer – An Exploratory Analysis

Student Number: 8477752
Word Count: 2,998

Introduction

Forecasting is a key component of sales analysis. Traditionally sales are forecasted using data from previous sales (e.g. Choi et al., 2014; Loureiro et al., 2018; Boone et al., 2019), however, previous research also links sales to unemployment rates, Consumer Price Index (CPI), promotional activities (Gür et al., 2009; Trapero et al., 2015), store types (Loureiro et al., 2018), and temperature, particularly in relation to clothing and online retail (Steele, 1951; Bahng and Kincade, 2012; Steinker et al., 2017; Bertrans and Parnaudeau, 2019). Using data from Walmart US, the current analysis explores links between these variables and weekly sales in order to assess which variables can most reliably predict future sales.

Method

Data Collection

The current analysis employs publicly available secondary sales data (see Walmart, 2014) from the years 2010-2012 from Walmart, a large multi-national retailer. The current data includes weekly sales data from several US-based stores and their departments, promotion information at store level, holiday information, as well as information on the stores. Variables at store level include CPI, unemployment rates, temperature, fuel price, store size, and store type. Overall, the datasets include information on 45 different stores, and 81 departments. Of all departments 27 are not present in every store, and 6 of these occur in fewer than 50% of all stores (see Figure 1).

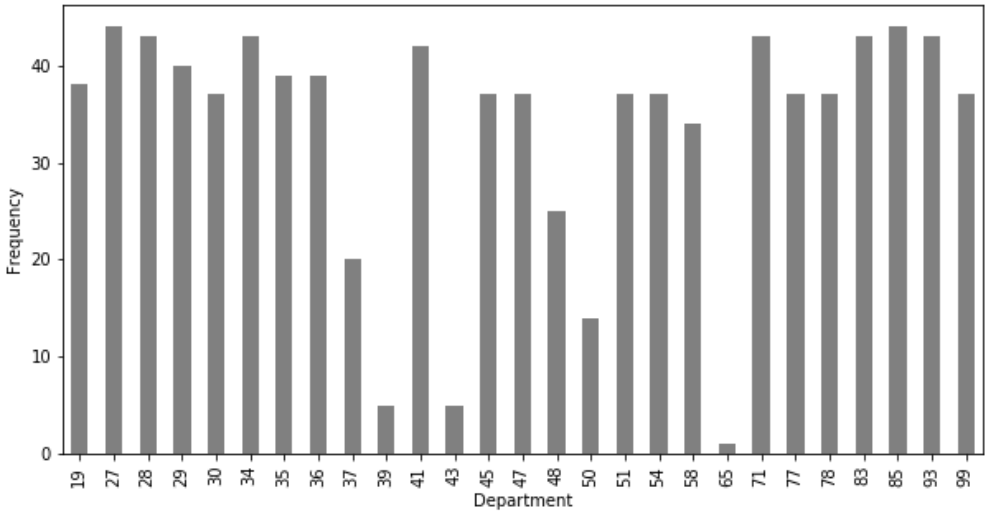


Figure 1. Frequency of department occurrence by store.

Data Linkage and Pre-processing

As the current data is split into four files, it must be linked prior to analysis. This is done – depending on the dataset – on store, or store and date. To ensure that dates within weeks do not differ, dates are converted into years and weeks first. Put differently, though all sales data is recorded on Fridays, the same may not be the case for promotion and other store data.

Conversion into week of year and year will ensure that dates match across data files regardless of the day of the week on which a recording was made.

For data for which weekly sales data are available (421,570 observations), 69.2% of promotion data are missing. However, considering only the period in which promotional data are available – after November 11th, 2011 – only 10.9% of all promotion data are missing (see Table 1). No other data are missing between February 5th, 2010 and October 26th, 2012. Similarly, for the period with no available weekly sales data (November 2nd, 2012 - July 26th, 2013) 10.4% of promotion data are missing. In addition, 33.3% of CPI and unemployment data are missing for this time period (see Table 2).

Table 1. Amount of missingness in promotion data (period: February 5th, 2010 – October 26th, 2012).

Variable		Missing Observations	
		Count	Percentage (%)
All Data	Promotion 1	4,155	64.6
	Promotion 2	4,798	74.6
	Promotion 3	4,389	68.2
	Promotion 4	4,470	69.6
	Promotion 5	4,140	64.3
After November 11th, 2011	Promotion 1	15	0.7
	Promotion 2	658	28.7
	Promotion 3	249	10.8
	Promotion 4	330	14.4
	Promotion 5	0	0.0

Table 2. Amount of missingness in data (period: November 2nd, 2012 - July 26th, 2013).

Variable	Missing Observations	
	Count	Percentage (%)
CPI	585	33.3
Fuel Price	0	0.0
Holiday Information	0	0.0
Promotion 1	3	0.2
Promotion 2	471	26.8
Promotion 3	188	10.7
Promotion 4	256	14.6
Promotion 5	0	0.0
Store Size	0	0.0
Store Type	0	0.0
Temperature	0	0.0
Unemployment	585	33.3

Missing promotion values are handled in four ways and evaluated later on in a sensitivity analysis. First, missing values are ignored. Second, missing values are treated as zeros. Third, missing values are imputed to be the median value. Fourth, as promotions appear to be seasonal (see Figure 2), missing values are copied from consecutive years, where available. If data are also missing in the following years for a given week, they are copied from the previous year if available, or else the following week. Thereafter, only two values from promotion 2 are still missing. These are replaced with the median.

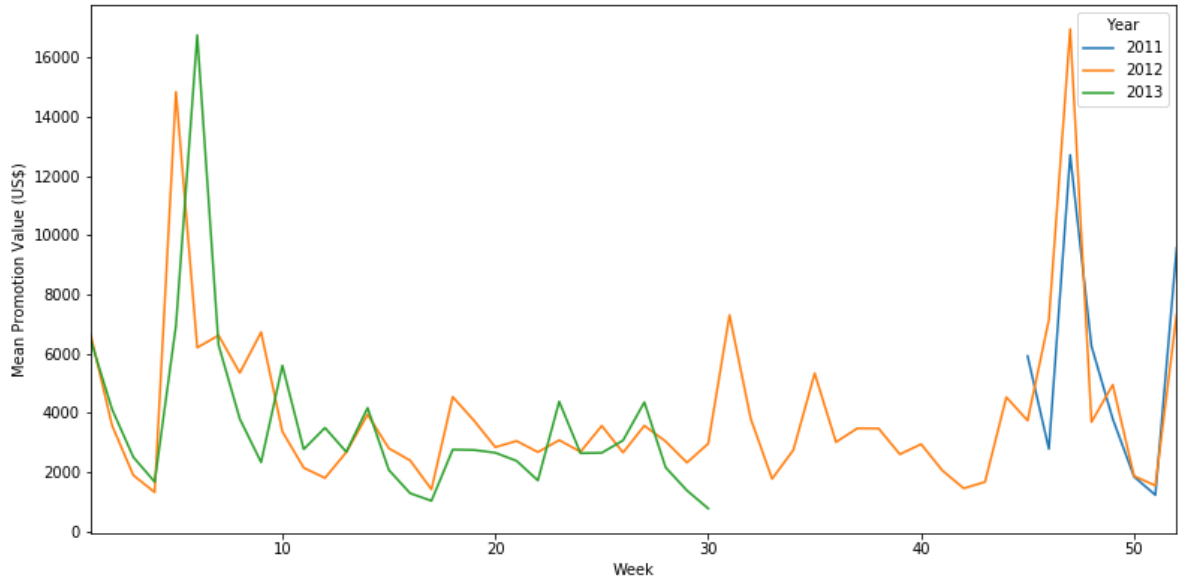


Figure 2. Mean promotion values by year.

Moreover, weekly sales data are missing for some dates. Including all combinations of Stores and Departments which exist in the dataset and all dates between February 5th, 2010 and October 26th, 2012, 11.5% of weekly sales data are missing. However, these are mostly data from departments with low sales which stop at certain points within the dataset. Thus, there is reason to assume that this is data from departments which were removed from stores due to not meeting performance targets. As a result, missing values for weekly sales are not imputed. Instead, the current analysis includes only observations of weekly sales which are reported in the dataset.

Negative values for weekly sales and promotion data are kept, as these likely represent returns and price reductions not linked directly to the promotions. This concerns 1,285 weekly sales data points as well as 23 promotion data points.

Finally, holiday and CPI data are pre-processed. CPI, as reported in the dataset, is bimodal. This is because it is not reported in percentage change – a typical measure of CPI. Thus, it is converted into increments in the current dataset (US: Bureau of Labor Statistics, n.d.). The holiday variable, on the other hand, is altered so that the week of the holiday, as well as the two weeks prior to the holiday are considered holiday periods. This time period is chosen as peaks in the data start appearing 2 weeks prior to holidays, in particular Christmas and

Thanksgiving. In this way the holiday variable captures busy retail periods leading up to holidays, rather than only the holiday week itself.

Data Exploration and Findings

Across all stores, departments and dates, weekly sales have a mean of US$ 15,981.26 (s.d. = US$ 22,711.18). Moreover, weekly sales range from US$ -4,988.94 to US$ 693,099.36, indicating that sales may differ largely across dates, departments, and stores. As visualised in Figure 3, weekly sales appear to indeed have seasonal trends, with peaks in November and December, as well as variance between store types. While the larger store types A and B show seasonality, aggregated weekly sales from C type stores appear more static.

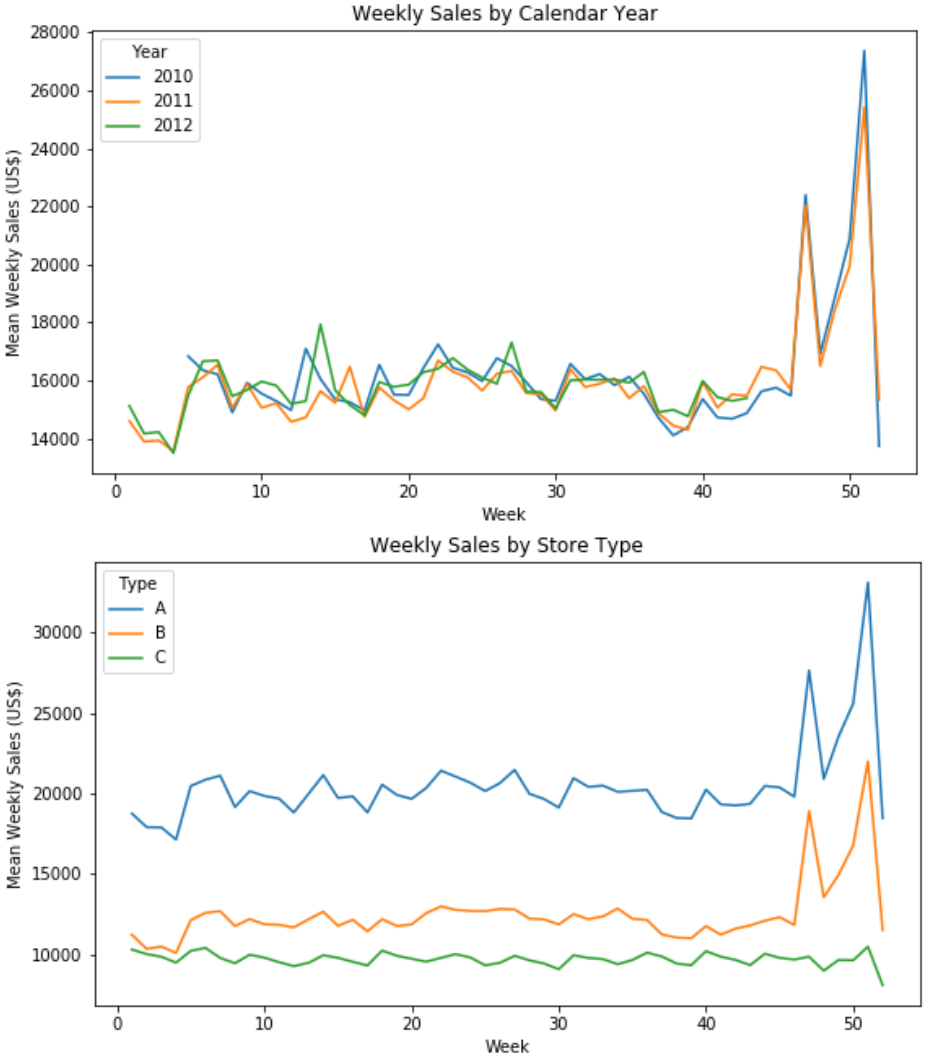


Figure 3. Mean weekly sales by year and store type.

Traditional Variables in Sales Forecasting

As previous research indicated links between sales data and temperature (Bahng and Kincade, 2012; Steinker et al., 2017; Bertrans and Parnaudeau, 2019), promotions (Gür et al., 2009; Trapero et al., 2015), unemployment rates, CPI (Cranage and Andrew, 1992; Boone et

al., 2019), store types (Loureiro et al., 2018), as well as previous sales (Choi et al., 2014; Loureiro et al., 2018; Boone et al., 2019), these variables are explored first.

Pearson's *r* correlation reveals a strong positive correlation between current and previous sales ($r = 0.98$). Moreover, dummy variables created for store types A ($r = 0.19$) and B ($r = 0.13$) appear to be weakly, positively correlated to weekly sales. Correlation coefficients for fuel price, unemployment, CPI, and temperature are very low between 0.05 and -0.03. Correlations between promotions and weekly sales are weak, but positive. Promotions with the highest correlation coefficients are promotion 1, imputed using other promotion data ($r = 0.09$) and with missing values ignored ($r = 0.08$), as well as promotion 5, imputed using other promotion data ($r = 0.09$) and with missing values ignored ($r = 0.09$). All other promotions, including the different imputation methods have correlation coefficients between 0.07 and 0.02.

Exploratory Data-Splitting

As indicated by Figure 3, trends within the data are not uniform. Thus, splitting the data appropriately may reveal patterns which are not evident at a high level of aggregation. In addition to further investigating store types, this section explores two approaches to splitting the data: analysis at department level and geographical grouping.

Geography

Geography can have a significant impact on sales, as consumer behaviours differ regionally (Jank and Kannan, 2005). As CPI is regional in the US (US: Bureau of Labor Statistics, n.d.), CPI values are matched across stores. This reveals 15 regions; six regions contain 1 store each, three contain 2 stores each, two contain 3 stores, two contain 4 stores each, one region contains 8 stores, and the last region 11 stores. As visualised in Figure 4, weekly sales appear to vary slightly between regions. However, this may reflect differences between individual stores and not geography, as each region only contains a small number of stores and, indeed, over one third of the regions contain only 1 store each. Similarly, the promotional data suggests some regional variance (see Appendix A), but again this cannot be attributed to regions with certainty, due to the small sample size per region. As a result of this geographical regions is not used further in the current analysis.

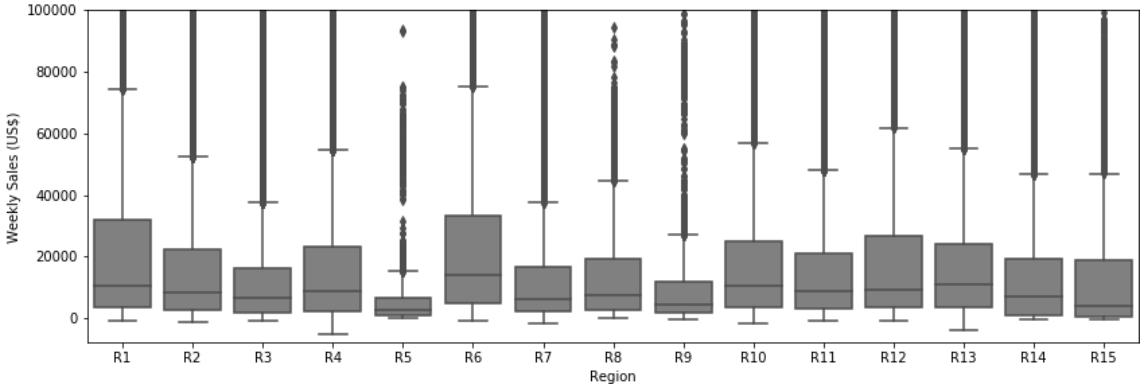


Figure 4. Weekly sales across regions.

Department and Store Type

Second, variables are investigated by store type. As shown in Table 3, means of weekly sales and promotion variables vary drastically between store types. Means appear to be slightly lower for type B than type A stores but are nonetheless comparable. Type C stores, however, have much lower mean promotion values. While mean sales of C type stores are approximately half of type A store mean sales, type C store promotion value means range between 1-23% of type A store means. The means of promotion values by store type over time are also visualised in Figure 5.

Table 3. Means and standard deviations of weekly sales and promotion variables by store type.

Variable	Statistic	Store Type		
		A	B	C
Weekly Sales	mean	20099.60	12237.10	9519.53
	s.d.	26423.50	17203.70	15985.40
Promotion 1	mean	8686.89	7108.81	394.65
	s.d.	8709.45	7886.56	484.01
Promotion 2	mean	3763.75	3063.89	447.07
	s.d.	10409.50	8524.78	686.51
Promotion 3	mean	1647.31	1483.70	18.00
	s.d.	10312.30	9731.39	30.85
Promotion 4	mean	3916.01	2925.56	65.04
	s.d.	6882.55	5577.13	98.03
Promotion 5	mean	5999.61	3688.56	1384.30
	s.d.	7273.57	3979.09	904.12

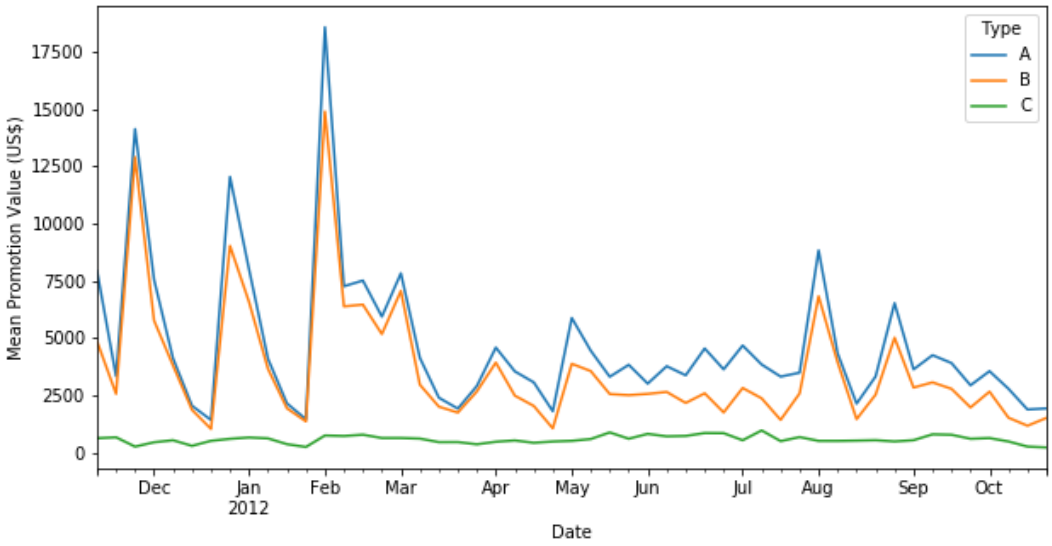


Figure 5. Mean promotion values by store type.

In addition to store type, and as shown in Figure 6, departments also appear to have a major impact on sales. This is due to product types differing between departments. Previous research indicates, for instance, that clothing sales are affected by temperature (Bertrand and Parnaudeau, 2019). Thus, departments also have varying levels of seasonality. As some departments appear to have comparable seasonality, the data are split by a combination of department and store type. As type C stores appear to have different departmental and seasonal trends to the other two types, they are grouped together (group 1). Store types A and B, on the other hand, are split by departments. Notably, some departments have a peak in weekly sales at Thanksgiving and Christmas time while others seem to be more static. Thus, departments with a range in sales between weeks 47 and 52, where the maximum is twice as large or larger than the minimum are grouped together (group 3), while departments with less evident seasonality form another group (group 2). Group 2 contains 35 departments, while group 3 contains 46 departments. A list of these can be found in Appendix B.

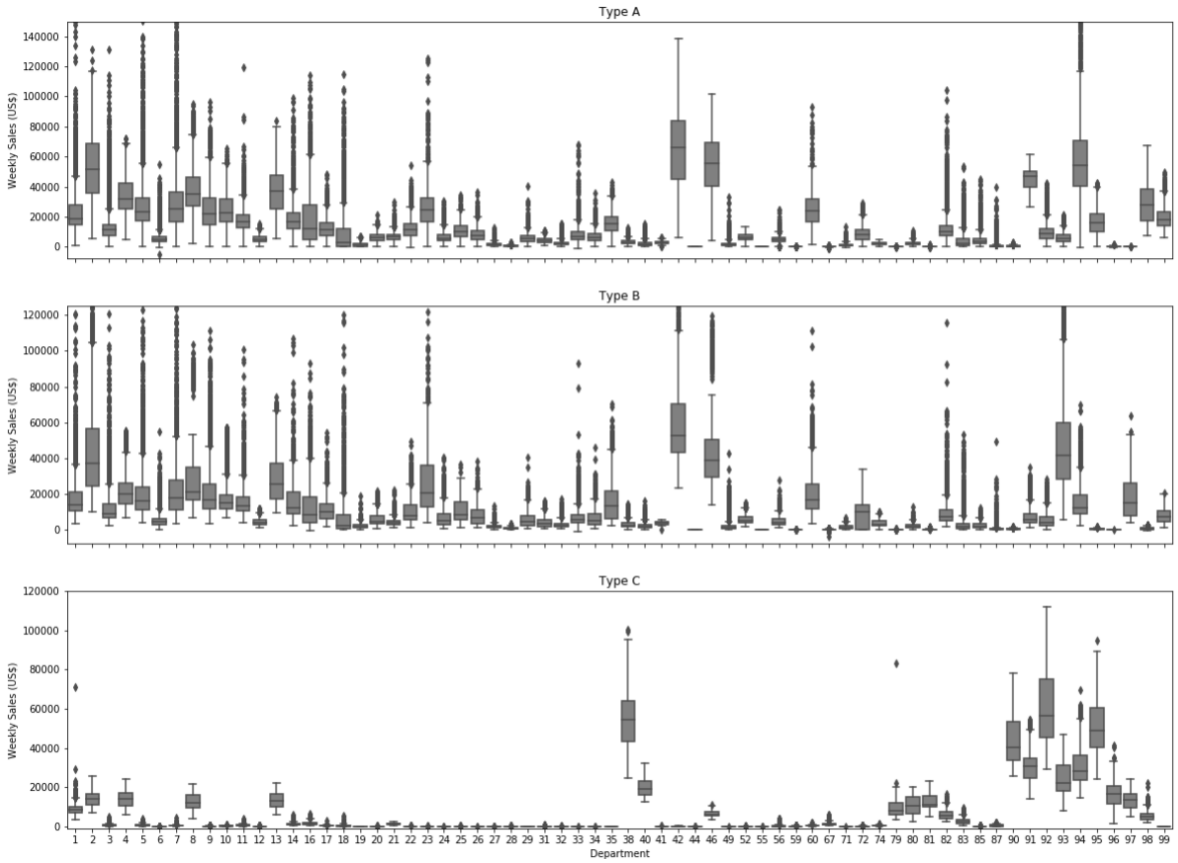


Figure 6. Weekly sales by departments and store types.

Total weekly sales for these groups are visualised in Figure 7. Interestingly, despite group 2 containing fewer departments, total weekly sales are higher throughout the year for group 2. This is also reflected in mean weekly sales, which are higher with lower levels of dispersion for group 2 (mean = US$ 10,503.68; s.d. = US$ 29,033.27) than for group 3 (mean = US$ 10,503.68; s.d. = US$ 15,410.57).

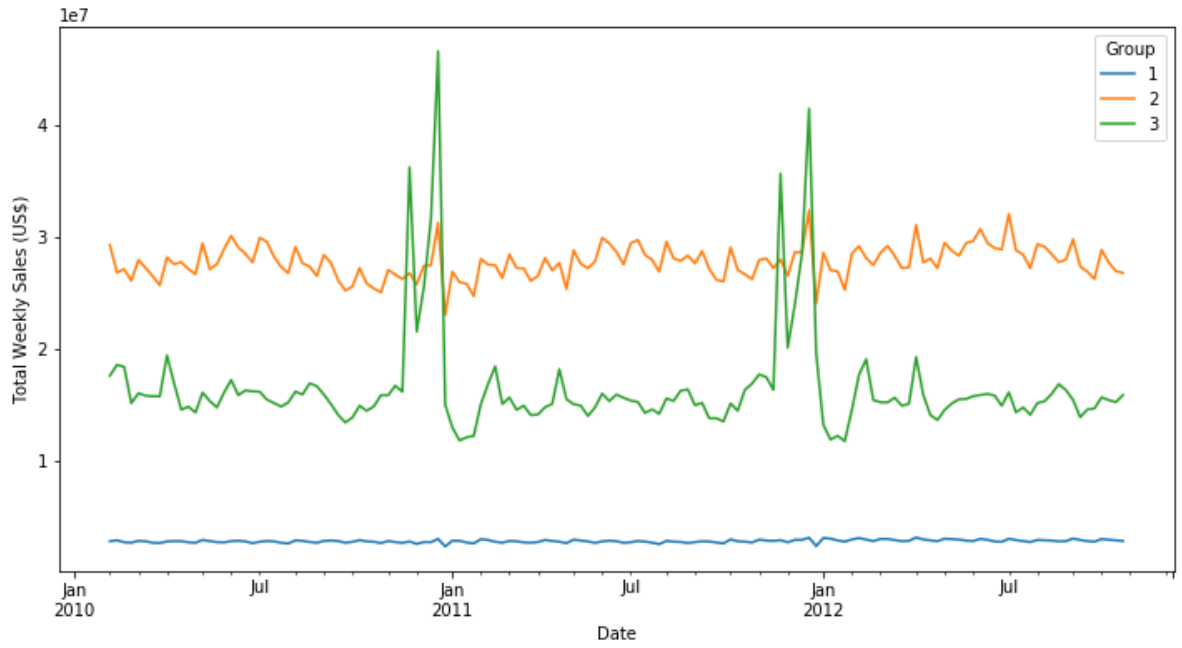


Figure 7. Weekly sales by groups; group 1 contains type C stores, group 2 contains departments with no peaks in November/December from store types A and B, group 3 contains departments with peaks at November/December from store types A and B.

Pearson's *r* correlation coefficients indicate stronger correlations between these groups and weekly sales than between store types and weekly sales. While group 1 is the same as store type C and therefore shows no change, group 2 and 3 have correlation coefficients of $r = 0.33$ and $r = -0.26$ respectively, compared to types A and B, with coefficients of $r = 0.19$ and $r = -0.13$ respectively.

Moreover, correlations between weekly sales and promotions vary by group. Group 1 has Pearson's *r* coefficients below +/- 0.05 for all promotions (see Table 4). The strongest correlations for group 2 are between promotion 1 and weekly sales and promotion 5 and weekly sales (both $r = 0.10$). The promotion most strongly correlated with weekly sales for group 3, on the other hand, is promotion 3 ($r = 0.13$). This is not surprising, given that promotion 3 also has a seasonal peak around November and December (see Appendix C). Findings are similar for promotions imputed using other promotions data (see Table 4).

Table 4. Correlation coefficients between promotions and weekly sales by groups.

	Promotion 1 (Imputed)	Promotion 2 (Imputed)	Promotion 3 (Imputed)	Promotion 4 (Imputed)	Promotion 5 (Imputed)
Group 1	-0.02	0.02	0.00	0.04	0.00
	(-0.02)	(0.07)	(0.00)	(0.02)	(0.00)
Group 2	0.10	0.02	0.01	0.06	0.10
	(0.11)	(0.03)	(0.01)	(0.08)	(0.10)
Group 3	0.04	0.03	0.13	0.02	0.06
	(0.05)	(0.02)	(0.12)	(0.04)	(0.06)

** *Note: Imputed refers to promotions imputed using other promotion data.*

Modelling

Linear regressions are run in consideration of variables for which Person's *r* suggests the presence of a linear relationship with weekly sales. This includes groups (which capture store type and department to some extent), previous sales, and some promotion variables. First, a baseline model is run on the entire dataset, using only previous sales as a predictor variable. This linear model reveals that previous sales are able to explain almost all of the variance in weekly sales, $R^2 = 0.967$, with previous sales being a significant predictor of weekly sales ($p < 0.001$). While the R^2 is high when running the model on grouped data, groups 1 and 2 see an improvement compared to the baseline model, whereas group 3 performs worse than the baseline model. These results are shown in Table 5 and Figure 8.

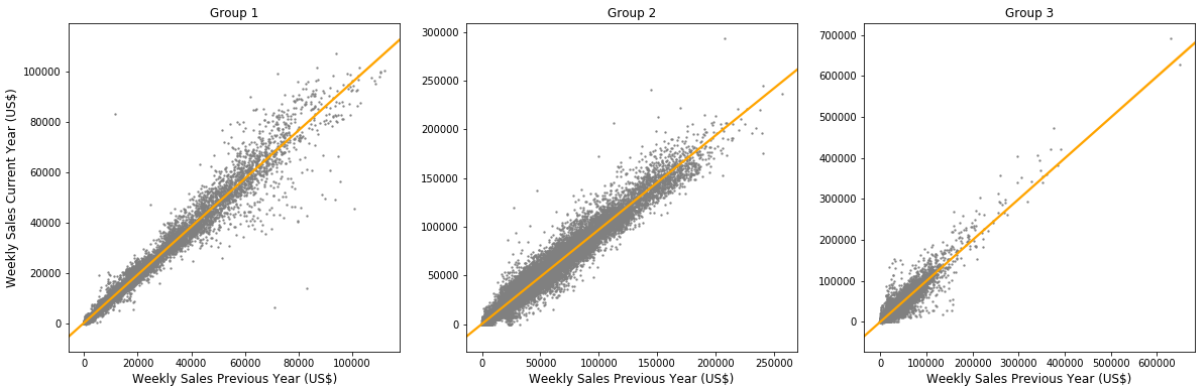


Figure 8. Weekly sales from current and previous years by group.

Table 5. Variance explained and model error for simple linear regression models run on the full dataset and different groups.

Data	Full Dataset		Cross-Validated RMSE
	R^2	RMSE	
All Data	0.967	4132.53	3562.04
Group 1	0.977	2499.50	2052.17
Group 2	0.975	4617.01	4197.84
Group 3	0.931	3860.08	3350.36

** *Note: 1) All the models in this table use previous sales as the only predictor variable of weekly sales. 2) For cross-validation, the models are trained on data before August 27th, 2012 and tested on data between and including August 27th, 2012 and October 26th, 2012.*

As relationships between weekly sales and previous sales vary slightly by group, and since promotions affect groups differently, further models are run by group. To measure how well a model is able to predict sales, Rooted Mean Squared Error (RMSE) is measured. This is a type of error often reported in both linear regression and neural networks (Kolehmainen et al., 2001; Loureiro et al., 2018). Linear regressions are run, however, here these are also compared to a multilayer perceptron (MLP), a type of neural network (see **Error! Reference**

source not found.). Groups 1 and 3 have the lowest RSME when promotion data are excluded. Group 2 has a slightly lower RSME when the imputed promotion 5 is included as a predictor variable than when no promotions are included.

Table 6. Model errors for multiple linear regression models and neural networks run on full dataset and different groups.

Data	Promotion Variables	RSME	
		Linear Regression	MLP
Group 1	None	**2052.17**	**2054.79**
	Promotion 2, imputed	2072.52	2141.61
Group 2	None	4197.84	4201.95
	Promotion 1	4276.78	4333.26
	Promotion 1, imputed	4201.37	4466.65
	Promotion 5	4266.55	4239.57
	Promotion 5, imputed	**4190.72**	**4137.26**
	Promotions 1 & 5, both imputed	4193.73	4171.79
Group 3	None	**3350.36**	**3336.72**
	Promotion 3	3445.12	3653.83
	Promotion 3, imputed	3352.26	3464.40

** *Note: All the models in this table are trained on data before August 27th, 2012 and tested on data between and including August 27th, 2012 and October 26th, 2012.*

The neural networks with the lowest RSME's are chosen to forecast sales from November 2nd, 2012 until July 26th, 2013. This means that for groups 1 and 3 the models without promotional data are chosen, while the group 2 model includes previous sales as well as promotion 5 (imputed using other promotion data) as predictors. Visualisations of mean weekly sales can be found in Figure 9, while visualisations of further disaggregation into departments and stores is attached in Appendix D.

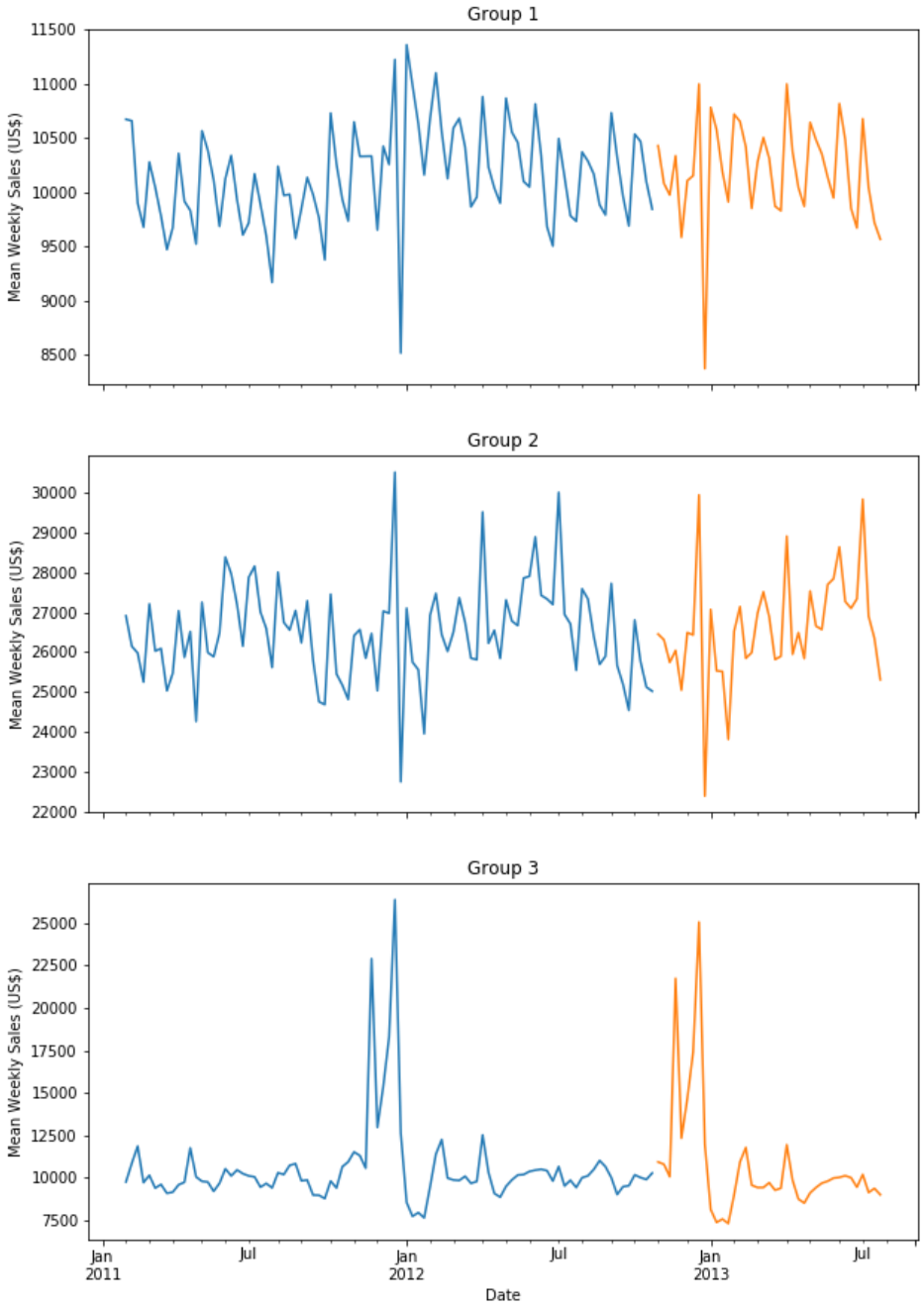


Figure 9. Mean actual and forecasted sales for all groups.

*** Note: 1) The models for groups 1 and 3 include previous sales as the only predictor, while group 2 also includes the imputed promotion 5. 2) Blue are actual sales, orange are forecasted sales.*

Discussion and Conclusion

In line with previous literature, past sales are found to be the strongest predictor of sales (Choi et al., 2014; Loureiro et al., 2018; Boone et al., 2019). Moreover, for the group made up of departments with no extreme peak in November and December from store types A and

B promotion 5 appears to reduce model error slightly. Nonetheless, the impact of promotions is not found to be strong in the current study and dataset. Perhaps having information on promotions at department level would lead to be a better predictor of sales. Moreover, due to the use of sales data from the same week but the year before, data for previous sales is only available from February 5th, 2011, a year after the first date recorded in the dataset. This may be one reason why using original or imputed promotions made only a small difference.

Other variables, including temperature, unemployment, CPI, and fuel price are not found to have strong correlations with sales, in contrast to some previous findings (e.g. Cranage and Andrew, 1992; Boone et al., 2019). Moreover, geographical patterns are not established in the current analysis, due to the small sample sizes per region. Nonetheless, in combination with geographical region, temperature may be used to establish patterns both seasonal and geographic consumer behaviour in future analyses. This may be able to capture some sales increases over the summer months, on which the current analysis does not focus. Furthermore, impacts of temperature on sales may vary by product type (see Bahng and Kincade, 2012) and thus may be better analysed at a departmental level.

The current analysis is limited by the use of linear regression for initial exploration and model testing. Linear regression assumes independence of observations, which this dataset does not have due to the hierarchical structure between stores and departments. To minimise the effect of this, the final models used in the current research are established using neural networks. Unlike linear regression, no assumptions are made about the data. Moreover, neural networks are frequently used in sales forecasting, and are found to perform well on sales forecasting tasks in previous studies (Thiesing and Vornberger, 1997; Loureiro et al., 2018). Despite this, a time series analysis at department level may be able to further improve predictions, as this takes sales from previous years, as well as from previous weeks into account (Nikolopoulos and Thomakos, n.d.). Perhaps using sales from previous weeks as well as from previous years may reduce the error from the neural network models. However, for linear regression this cannot be used, as sales from previous years and weeks would likely have high levels of covariance and thereby violate another assumption of linear regression. Thus, using a neural network in addition to linear regression helps validate the current analysis; nonetheless, a time series analysis may be most appropriate for this type of data.

Finally, the current analysis does not ensure that holiday weeks map onto one another between years. In other words, there is no control for mismatches between weeks in which holidays happen across years in the current analysis. Mapping holidays onto one another or incorporating specific dates for holidays may help make the model more accurate.

In conclusion therefore, the current analysis is able to establish sales from previous years as the strongest predictor of current sales. Although there are methodological limitations in applying linear regression to the current data, the current use of neural networks for the final forecasting models reduces this limitation. Notably, the current analysis does not find links between other variables and weekly sales, with the exception of one promotion for one sub-

set of the data. Other data-splits and an analysis of promotion impacts at store level may be able to make other inferences about the current data.

References

Bahng, Y. and Kincade, D.H. 2012. The relationship between temperature and sales: Sales data analysis of a retailer of branded women's business wear. *International Journal of Retail & Distribution Management*. **40**(6), pp.410–426.

Bertrand, J. and Parnaudeau, M. 2019. Understanding the economic effects of abnormal weather to mitigate the risk of business failures. *Journal of Business Research*. **98**(April 2017), pp.391–402.

Bertrans, J.-L. and Parnaudeau, M. 2019. Understanding the economic effects of abnormal weather to mitigate the risk of business failures. *Journal of Business Research*. **98**, pp.391–402.

Boone, T., Ganeshan, R., Jain, A. and Sanders, N.R. 2019. Forecasting sales in the supply chain: Consumer analytics in the big data era. *International Journal of Forecasting*. **35**(1), pp.170–180.

Choi, T.M., Hui, C.L., Liu, N., Ng, S.F. and Yu, Y. 2014. Fast fashion sales forecasting with limited data and time. *Decision Support Systems*. **59**(1), pp.84–92.

Cranage, D.A. and Andrew, W.P. 1992. A comparison of time series and econometric models for forecasting restaurant sales. . **11**(2), pp.129–142.

Gür, Ö., Sayın, S., Woensel, T. Van and Fransoo, J. 2009. Expert Systems with Applications SKU demand forecasting in the presence of promotions. *Expert Systems With Applications*. **36**(10), pp.12340–12348.

Jank, W. and Kannan, P.K. 2005. Understanding geographical markets of online firms using spatial models of customer choice. *Marketing Science*. **24**(4), pp.623–634.

Kolehmainen, M., Martikainen, H. and Ruuskanen, J. 2001. Neural networks and periodic components used in air quality forecasting. . **35**.

Loureiro, A.L.D., Miguéis, V.L. and Lucas, F.M. 2018. Exploring the use of deep neural networks for sales forecasting in fashion retail. *Decision Support Systems*. **114**(January), pp.81–93.

Nikolopoulos, K.I. and Thomakos, D.D. n.d. Forecasting Analytics. *In Press*., pp.1–46.

Steele, A.T. 1951. Weather's effect on sales of a department store. *Journal of Marketing*. **15**, pp.436–443.

Steinker, S., Hoberg, K. and Thonemann, U.W. 2017. The Value of Weather Information for E-Commerce Operations. *Production and Operations Management*. **26**(10), pp.1854–1874.

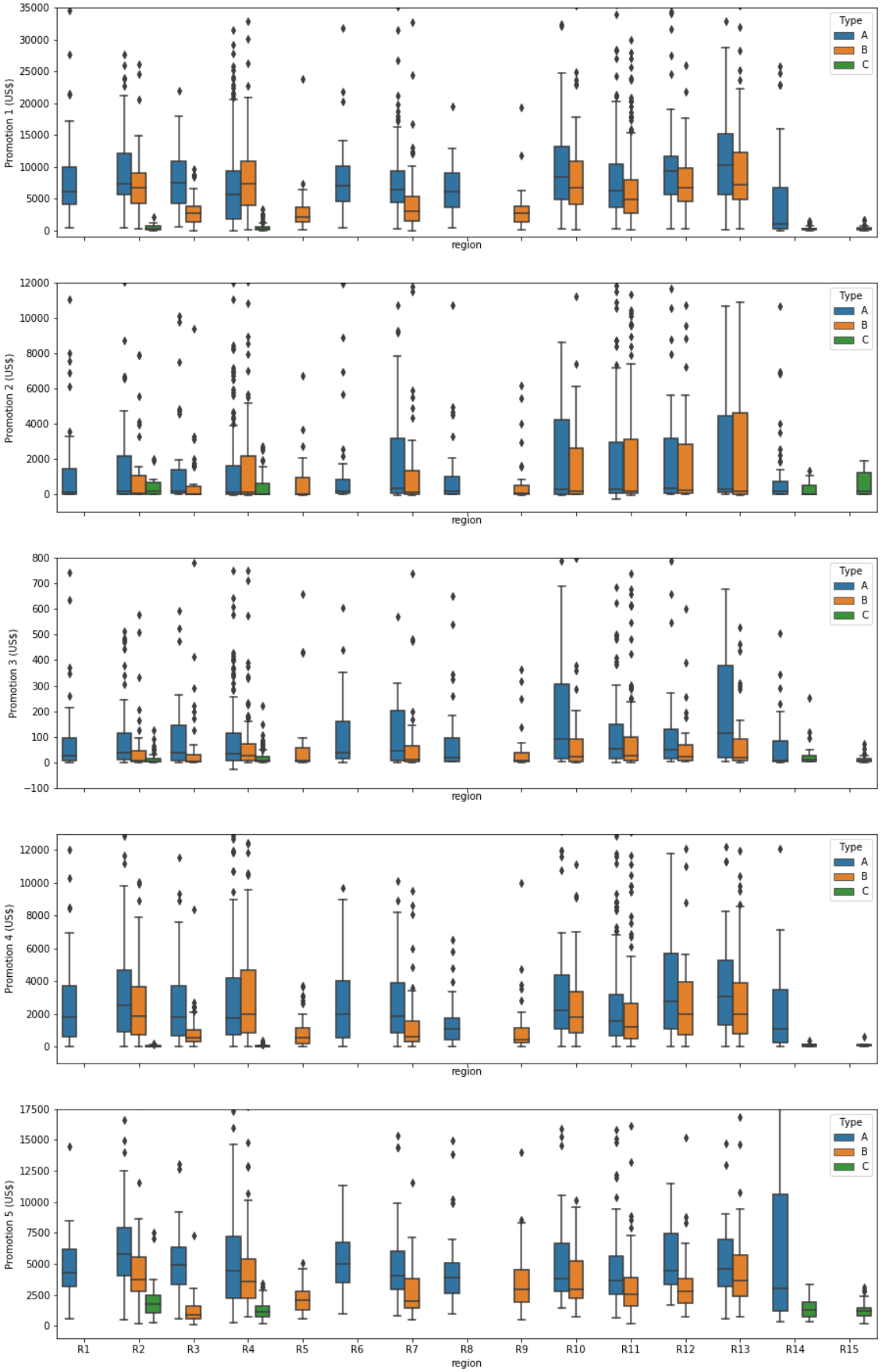
Thiesing, F.M. and Vornberger, O. 1997. Sales Forecasting Using Neural Networks *In*: *Proceedings of International Conference on Neural Networks (ICNN'97)*.

Trapero, J.R., Kourentzes, N. and Fildes, R. 2015. On the identification of sales forecasting models in the presence of promotions. *Journal of the Operational Research Society*. **66**(2), pp.299–307.

US: Bureau of Labor Statistics n.d. Consumer Price Index. *United States Department of Labor*.

Walmart 2014. Walmart Recruiting - Store Sales Forecasting. *kaggle*. [Online]. [Accessed 10 June 2019]. Available from: https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting.

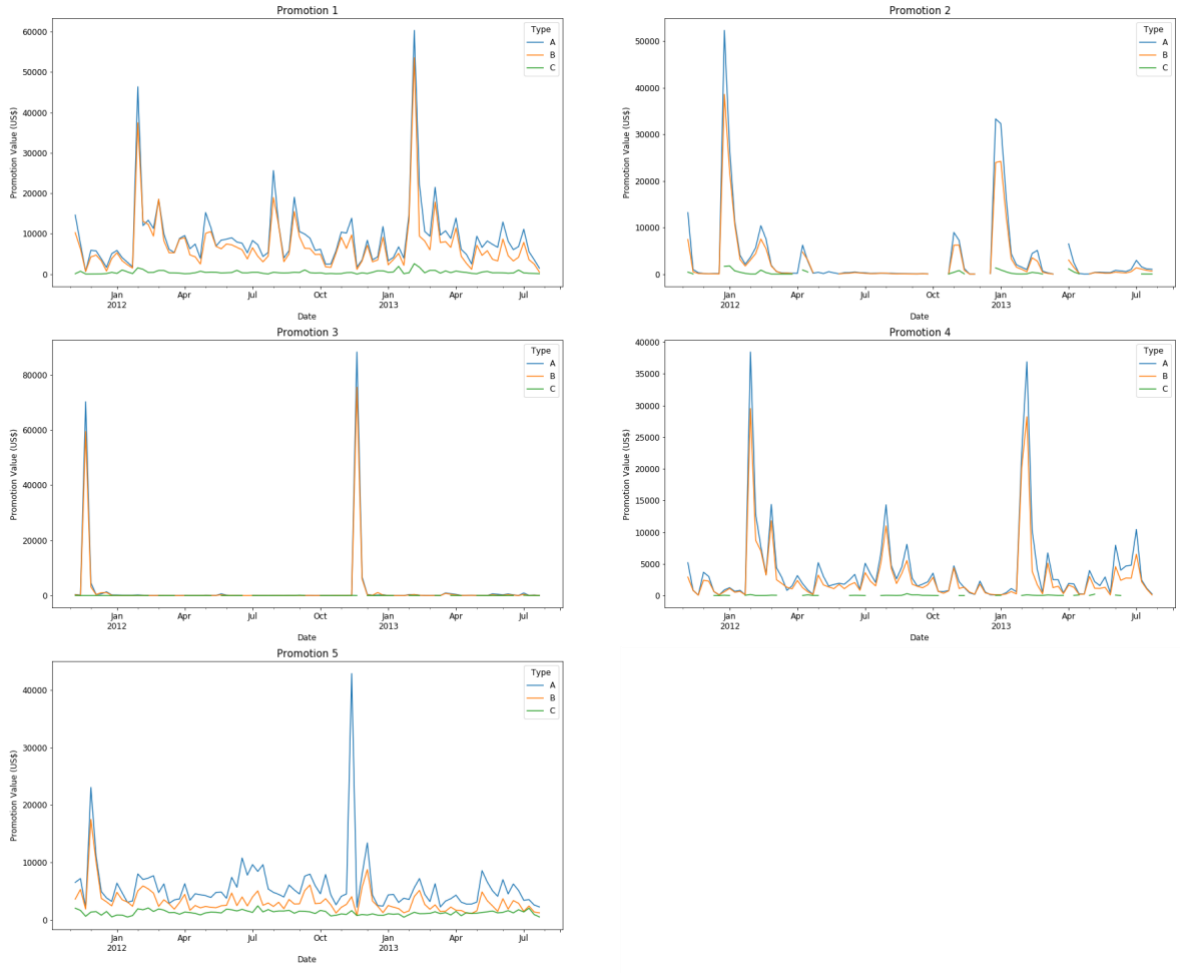
Appendix A: Figure A. Promotion values by region and store type



Appendix B: Table B. Group membership of departments

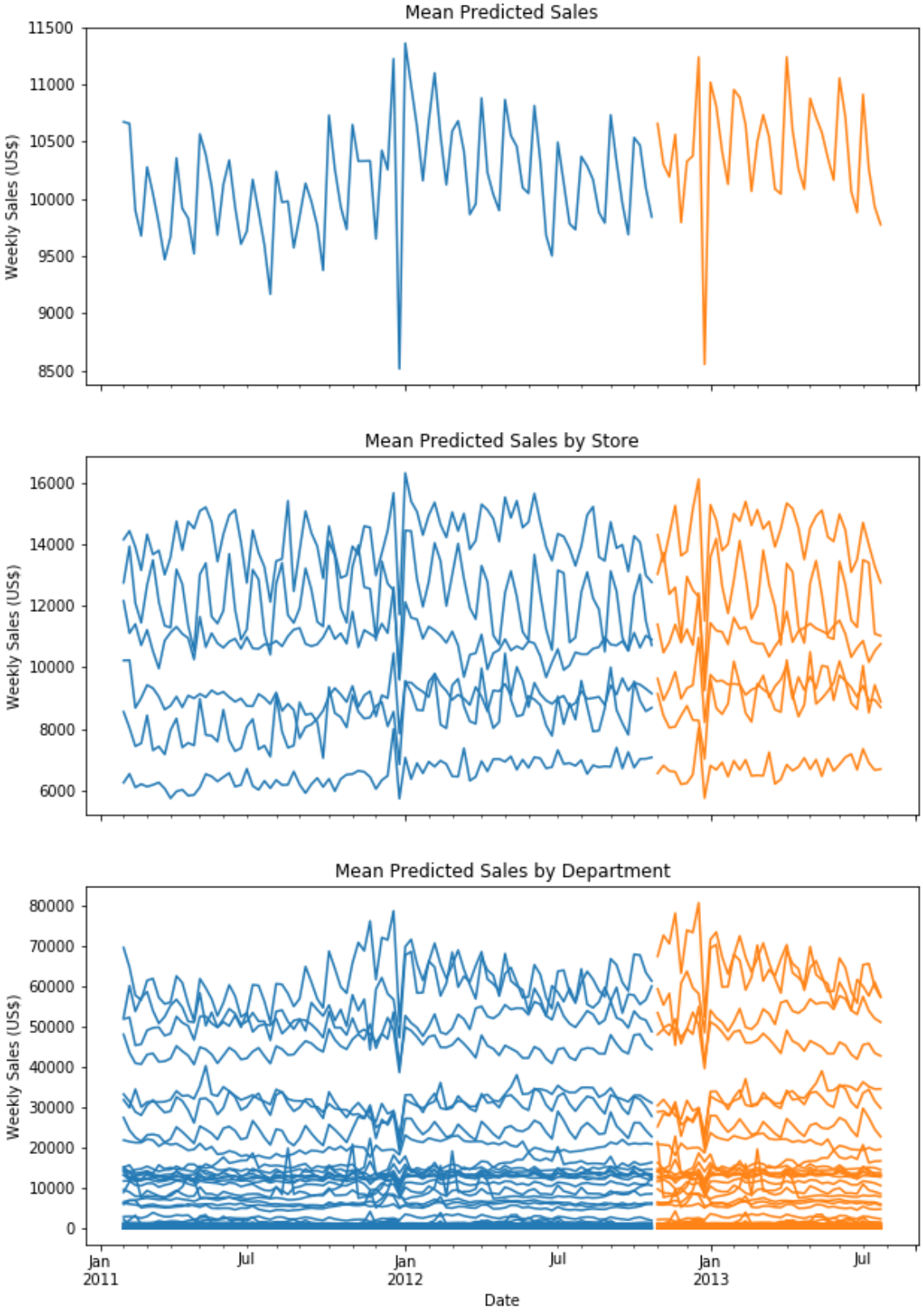
	Store Type	Departments
Group 1	C	All
Group 2	A and B	2, 4, 8, 10, 11, 12, 13, 16, 19, 28, 30, 37, 38, 39, 40, 42, 43, 49, 50, 65, 78, 79, 80, 81, 83, 87, 90, 91, 92, 93, 94, 95, 96, 97, 99
Group 3	A and B	1, 3, 5, 6, 7, 9, 14, 17, 18, 20, 21, 22, 23, 24, 25, 26, 27, 29, 31, 32, 33, 34, 35, 36, 41, 44, 45, 46, 47, 48, 51, 52, 54, 55, 56, 58, 59, 60, 67, 71, 72, 74, 77, 82, 85, 98

Appendix C: Figure C. Promotion trends by store type

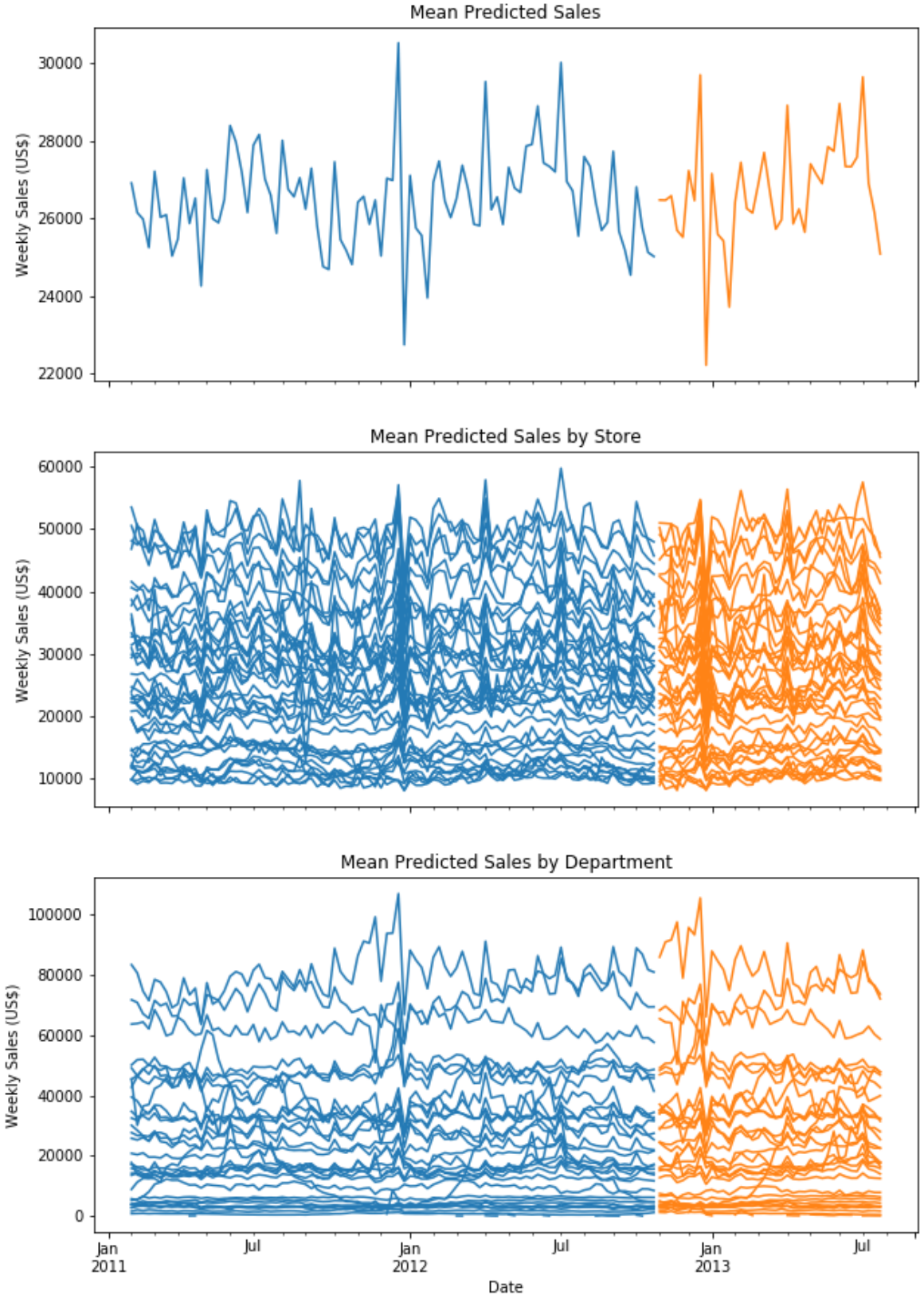


Appendix D: Figure D. Mean actual and predicted sales, aggregated means, by store, and by department for a) group 1, b) group 2, and c) group 3

a)



b)


Mean Predicted Sales
Weekly Sales (US$)
30000
28000
26000
24000
22000
Mean Predicted Sales by Store
Weekly Sales (US$)
60000
50000
40000
30000
20000
10000
Mean Predicted Sales by Department
Weekly Sales (US$)
100000
80000
60000
40000
20000
0
Jan
2011
Jul
Jan
2012
Jul
Jan
2013
Jul
Date

c)


Mean Predicted Sales
Weekly Sales (US$)
25000
22500
20000
17500
15000
12500
10000
7500
Mean Predicted Sales by Store
Weekly Sales (US$)
60000
50000
40000
30000
20000
10000
0
Mean Predicted Sales by Department
Weekly Sales (US$)
250000
200000
150000
100000
50000
0
Jan
2011
Jul
Jan
2012
Jul
Jan
2013
Jul
Date