1BM110: Data Analytics for Business Intelligence - Assignment 1

Group 29 (Kiran, T.: 1690906, Le Noheh, P.: 1815202, Ledemé, J.: 1815210, Zhang, X.: 1750348)

Supermarket chain Favorita stores sales prediction

Introduction

This report illustrates the use of data analytics to help supermarket chain Favorita stores to predict the daily sales of dairy products across 54 stores in Ecuador. Historical data about the sales and stores is provided, including the following: “data” (main dataset gives sales info for product families on a given date), “stores” (info about stores such as their city, state and cluster), “transaction” (daily transaction of each stores), “oil” (daily oil price) and “holidays” (national, regional, and local holidays information).

The business experts had some inputs about the possible influential factors on the dairy sales. First, the time may play a role, meaning the season, the day of the month and week, and the wage day may affect the sales. Secondly, holidays could also make a difference with sales. Thirdly, oil influences the national economy a lot, thus may influence the purchase behaviour of people as well. Based on those inputs, the preliminary explorations of the data is done as below.

Exploratory Data Analysis

The exploratory data analysis (EDA) for the Favorita sales data starts with the main data file “data" using Python v3.9 (packages: pandas, numpy, matplotlib, seaborn, scipy, and dython). The dataset provides information about the daily sales of different product families. An overview of the mean total sales of all Favorita stores is shown (Figure 1) to see the general trend across days of the week, days of the month, months of the year, and across years:

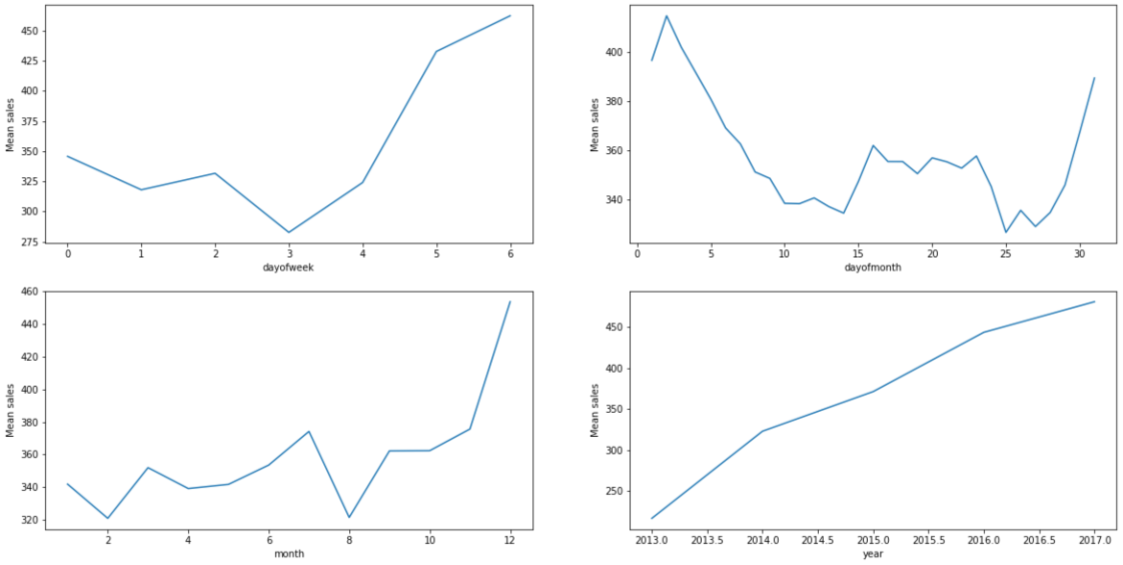


Figure 1. Mean total sales across days of week, days of month, months of year, and years

A trend of sales being higher towards the weekend is seen, and a rising trend towards the last few days of a month followed by a peak of sales during the first couple of days of the month. For a given year, sales are seen to be increased towards the end of the year with dips in

February and August. There is also an overall increasing trend of sales across years of the data (from 2013 to 2017).

We see a similar trend for the mean sales of product family DAIRY (Figure A.1, Appendix A).

To take a closer look at the different product families, the correlation between all product families (Figure B.1, Appendix B) and overall mean sales of all product families (Figure B.2, Appendix B) were checked. From Figure B.2, it is clear that some product families had quite small sales numbers both in general and compared to DAIRY sales. Those product families contribute very little to the overall sales, and hence are not likely to have an effect on the prediction of DAIRY sales and may even generate bias for the prediction model.

A quick glance at the data shows quite a few values of 0 for sales. However, we do not have any inventory information but only the sales and transactions data, so the reason for the 0 sales is not known (it could be because of lack of inventory as well). In this case, we decided to just keep the 0 sales to avoid losing any data and introducing biases.

Further EDA results in the following insights about other features and the supporting datasets as well:

- The feature *onpromotion* is positively correlated to the sales of the product families.
- The dataset “stores” gives an overview of the location and cluster of a particular store, which corresponds to the feature *store_nbr* from “data”.
- The dataset “holidays” identifies the days of national holidays in Ecuador and could result in different sale patterns (e.g., holiday shopping).
- The dataset “oil” corresponds to the daily price of oil, since Ecuador is an oil-dependent economy and vulnerable to shocks in oil prices. We see that the volume of sales is negatively correlated with the price of oil, with a correlation coefficient of -0,69 (Figure 2). This means that the higher the price of oil is, the less stores are able to sell. Since Ecuador is an oil exporting country, higher oil prices should positively affect the purchasing power of people. We think that this correlation could either be random, or a delay in the effect of drop in oil price on sales could explain this counter intuitive result.

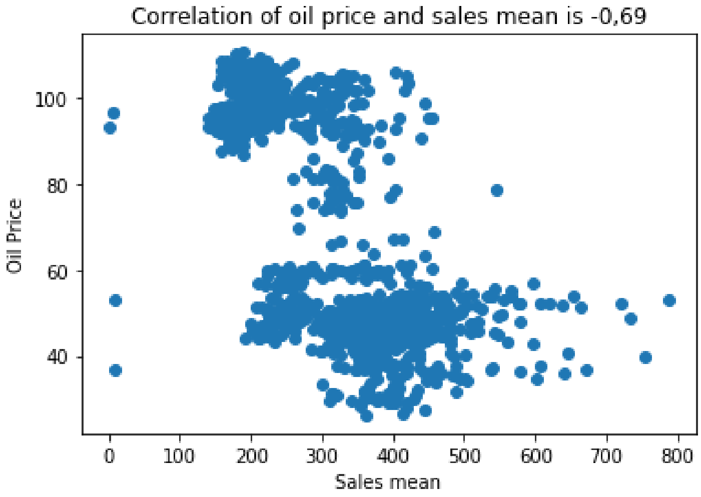


Figure 2. Correlation between oil price and mean sales

Data Preprocessing

- ***Data integration:*** Since all the data is from the same source, data integration here only involves finding the related information across all the data sets such that all required information can be accessed in a single dataset. The main dataset “data” contains the information for sales across all stores from 01/01/2013 to 31/07/2017. Additional sales information from 01/08/2017 to 15/08/2017 is given in the dataset “bonusQ”. Some datasets have information across different time periods but we use data only for the time period that corresponds to the main and bonus dataset, i.e., 2013-01-01 to 2017-08-15.

- ***Data cleaning:*** We start with looking for missing values, which are only found in the dataset “oil” (43 missing values). No apparent reason is made clear for these missing values, and hence a possible way to overcome this is by replacing the missing values with the local mean (between the previous and the next value). (Deleting these rows is also possible, however, since oil prices could influence the economy, replacing the values might be a better choice.)

- ***Data reduction***
    - For the dataset “holidays”, we can remove the rows where the feature *transferred* is true, as these holidays were shifted to another day. The holidays with *type* Bridge were additional holidays, but *type* Work Day was used to compensate for the bridge days. So we can remove work days as well.
    - As discussed previously, the product families which had low overall sales would be removed. Here, 1% of the total sales was chosen as the threshold. The product families which contribute to 99% of the overall sales are retained while other sales from other product families were removed.

- ***Data transformation and discretization:*** Normalisation for the sales pertaining to each product family should be done separately, as some product families have significantly higher sales than others and normalising across the entire dataset could create biases. We use the min-max normalisation, with every feature normalised with its own minimum and maximum values, and this could also handle the possible outliers.

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

Experimental Setup

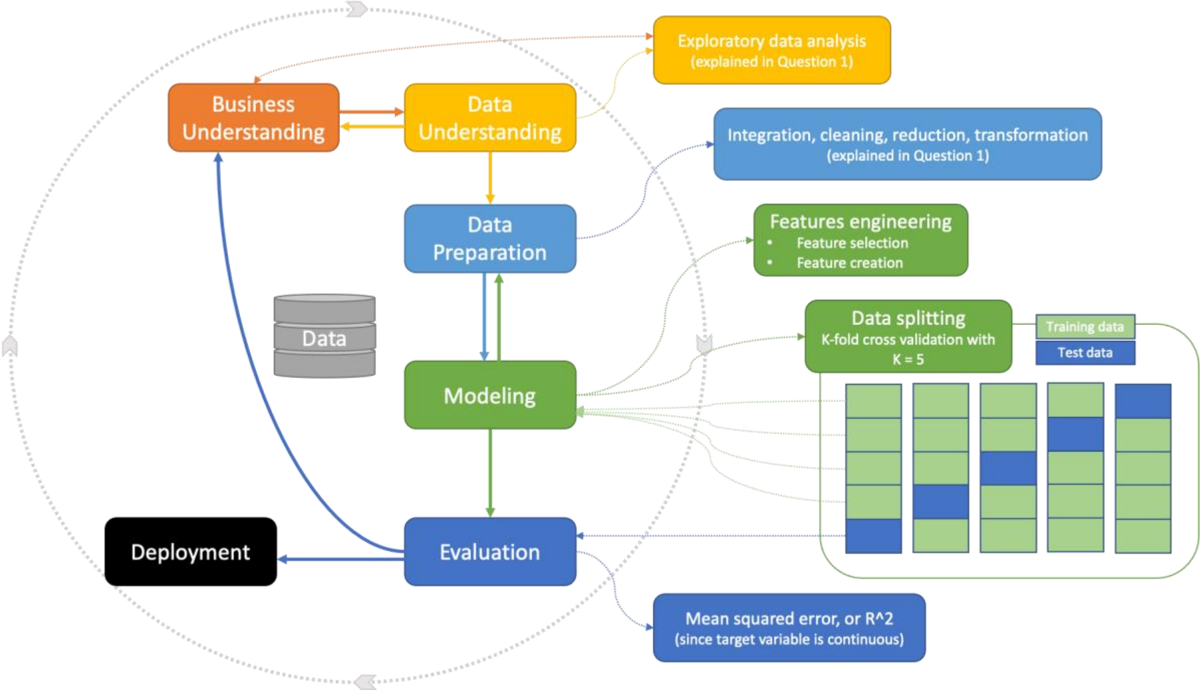

Exploratory data analysis
(explained in Question 1)
Business Understanding
Data Understanding
Integration, cleaning, reduction, transformation
(explained in Question 1)
Data Preparation
Features engineering
• Feature selection
• Feature creation
Data
Data splitting
K-fold cross validation with K = 5
Training data
Test data
Modeling
Deployment
Evaluation
Mean squared error, or R^2
(since target variable is continuous)

Figure 3. Environment setup

Data splitting: K-fold cross validation was selected because it is a commonly used validation method which results in a less biassed or less optimistic model.

Feature selection

- **Product family "DAIRY":** The aim of the model is to predict the sales of dairy products for the next day, hence past sales of this product family are important.
- **Other (correlated) product families:** Sales from other product families could also have an influence on the sales of dairy products. These were selected based on correlation with dairy sales and percentage of contribution to overall sales.
- **Promotion on products in store:** Promotions on products are likely to affect the sales as we see a positive correlation between the number of products being promoted and the sale numbers.
- **Holidays:** National holidays could affect sales patterns, so a binary variable indicating if the particular sale was on a holiday can be introduced. To better catch the possible influence, a feature of *holiday_tomorrow*, that is, the effect on sales if there is a holiday on the next day was created as well.
- **Oil price:** Since Ecuador is an oil-dependent country, the sales could be influenced by changes in the oil prices (though the change might not be visible immediately).
- **Store information:** The aim is to predict sales per store, identified by store number. Even though the information for *type* or *cluster* does not have a known meaning for us, it is a classification by Favorita and can be included. Geographical location of stores was included in the form of *city* of the store.
- **Transaction:** Overall transactions of the store will be used as well to provide information about sales in the store.

Features creation

In addition to the features described above, the following additional features were created and added/removed to the selected features based on iterative performance evaluation of the models.

- **Dates:** Breaking the date column into columns of *year*, *month*, *dayofmonth* and *dayofweek*. The experts pointed out that the sales were cyclical, relating to the season of the year, day of the month and day of the week. In the meantime, the wage day may affect the sales as well. This trend was also proved by the plots of trend for mean total sales and for the mean sales of product family DAIRY. Thus it makes sense to use those extra data info as features. Additionally, breaking the dates could also take away the time dependence of the features. Therefore randomization can be employed when splitting the data for training and testing.
- **Sales of other product families:** The sales of DAIRY products were related to the sales of a few product families and the relation can also be seen in the correlation map (Figure B.1, Appendix B). So a column for each product family with the sales for a certain day at a certain store were created.
- **DAIRY sales for next day:** Since the aim of the model is to predict dairy sales for the next day using information from past days, a column with the sales value for DAIRY from the next day *DAIRY_nextday* is added.
- **Payment days:** Experts speculated that the payment of wages (on the 15th and last day of the month) could have an impact on the sales. To this effect, a feature for *ispayment* is created to mark the days of payment of wages. An additional feature of *paymentdays* is also created to mark the next two days after the payment of wages as the trend of overall sales shows a cyclic effect of sales increasing during the first few days of a month (after the payment day on the last day of the previous month) and during the middle of the month (after the payment day on the 15th of the month).
- **One-hot encoding for categorical variables:** Variables such as *store_nbr*, *dayofweek*, *month* of the year, store *type*, and store *city* are one-hot encoded to enable the model to learn that these are categorical variables and should not be considered with their numerical value (like *store_nbr* or *dayofweek*).

Feature re-selection

The correlation of all features was calculated (Figure C.1, Appendix C). A few features show extremely low correlation to other features and DAIRY sales for the next day, so they are removed since they cannot contribute much to the prediction. The dropped features are *isholiday*, *ispayment*, *holiday_tomorrow*, *paymentdays* and *dayofmonth*. Other features that are removed are *year* since it is a non-cyclic feature of time and predictions will be done for future, *state* as it has a very high correlation of 1 with chosen feature *city*, and *type* as store information is available from *store_nbr* and *cluster*.

Model Choice

Based on the features and the target variable to be predicted, the prediction models have to be chosen.

- **Random Forest Regression:** It employs a number of techniques to maintain low variance and low bias, and hence it has good accuracy. A number of hyperparameters such as *max_features* and *min_samples_leaf* can be tuned to navigate the tradeoff between bias and variance so as to minimise errors. Besides that, in the task, being able to explain the result does not have a high priority, so Random Forest Regression is a reasonable choice for the prediction model.
- **Linear Regression:** It is a relatively easy yet powerful way of modelling. The exploratory graphs show certain trends which linear regression can pick up, which makes linear regression one valid option of modelling. The advantages of linear regression are its interpretability and lower computation requirements. With the interpretability, the dominant features of the sales can be found and used as input for business decisions. With lower computation requirements, insights can be acquired cheaper and faster.

Parameter configurations

Random Forest Regression**:** A few important parameters were chosen for achieving the best performance. We fed most of the parameters with a couple of choices, and used a randomised grid search (RandomizedSearchCV) to get the best combination.

- *n_estimators*: The number of trees in Random Forest. Random Forest uses the averaged result of all the trees as the final result. A too small number of trees may lead to underfitting, and more trees would not necessarily increase the performance anymore. We used 50 to 200 as feeding, and it turned out that **150 is the preferred number**.
- *max_features*: The features that the trees can use to split further. Here we **used the "sqrt"**, meaning the square root of the input features. This one is considered the best performance by previous studies.
- *Min_samples_leaf* : The minimum number of samples required to be at a leaf node. For example, if this is set as 1, then the split will only stop when there is 1 data point in one node, which may lead to overfitting especially if the dataset has a big amount of data. We chose numbers of 2, 5 and 10 for this parameter to avoid overfitting issues. **The optimised value turned out to be 2**.
- *min_samples_split*: the minimum number of samples required to split an internal node. This one is similar to min_samples_leaf, with our choice of 2, 5 and 10, **5 turned out to be the optimised value**.
- *max_depth*: the maximum depth/layer of the tree. This parameter is also used to control the tree from getting overfitting. We used 10 to 80, and the **optimised value for max_depth turned out to be 56**.
- *bootstrap*: True is used to make sure the bootstrap samples are used for building trees.

Linear Regression: There is no need to pre-define parameters for linear regression. While doing the modelling, it is seen that some features have more weight than others (~10^{10} for dates or geographical parameters vs. ~10^{-2} for the other product families). We tried to abandon the features with extremely low weight, but it lowered the R2 performance by 5%, so we decided to keep all features.

Performance of models

Both models showed good performance for the train and validation data. The performance metrics of R-squared (R2, the proportion of variance of dependent variables that can be predicted by the independent variables) and mean squared error (MSE, the mean squared error between the predictions and actual values) are shown below for both models.

Random forest regression

	R2	MSE
Training	0.982	0.00021
Validation	0.892	0.0018
Test	0.784	0.0024
Overall	0.940	0.00085

Table 1. R2 and MSE scores for Random Forest Regression modelling

For the validation phase, the results of the first 50 predictions are shown in Figure 4, most of the predicted values are quite close to the actual values, meaning a relatively accurate prediction. The R2 values for all stores can be seen in Figure 5, the performance is good in general for most stores. Only one store has a relatively low R2 value, the reason can be investigated further if necessary.

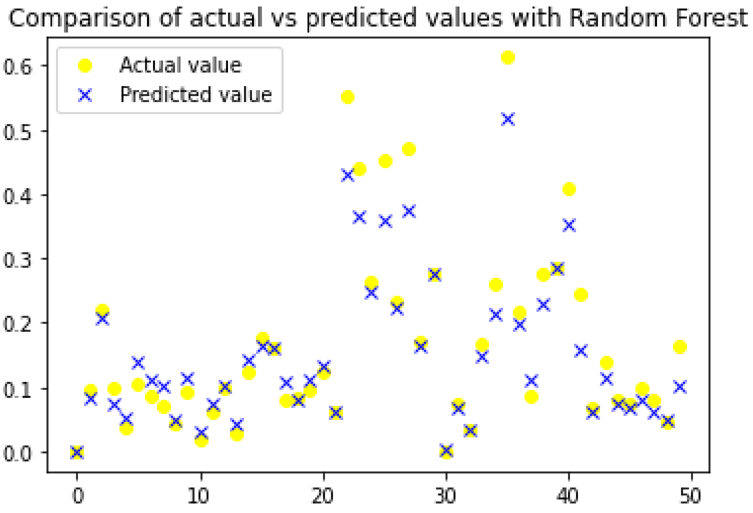


Figure 4. Actual and predicted values with Random Forest Regression

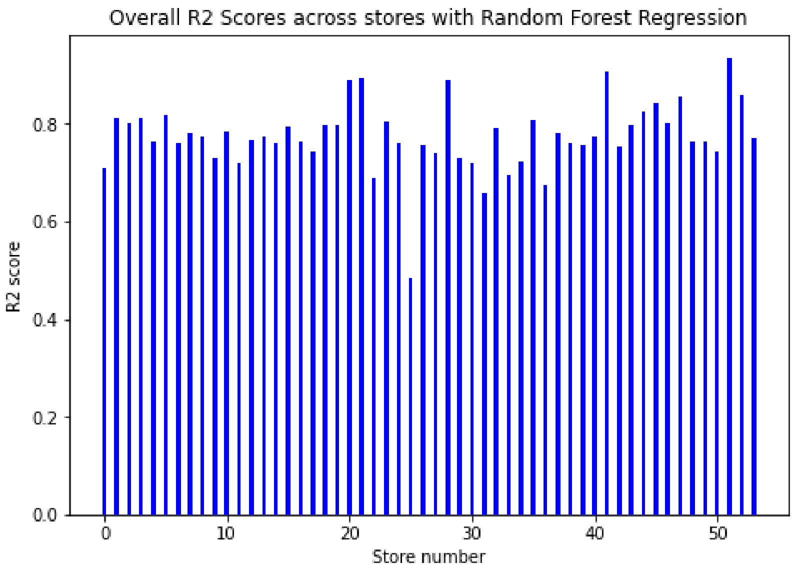


Figure 5. R2 scores for all stores

Linear regression

	R2	MSE
Training	0.880	0.0017
Validation	0.866	0.0019
Test	- 4.86e+23	5.42e+21
Overall	-3.40e+21	4.83e+19

Table 2. R2 and MSE scores for Linear Regression modelling

This model showed good performance for the training and validation, as we can see above, however, for the test data (which wasn't provided at the time of modelling) we see a significant drop in performance. The results of the first 54 predictions are shown in Figure 6. As for Random Forest, most of the predicted values are quite close to the actual values for the validation data points. So the prediction is also relatively precise as we can see on the correlation chart.

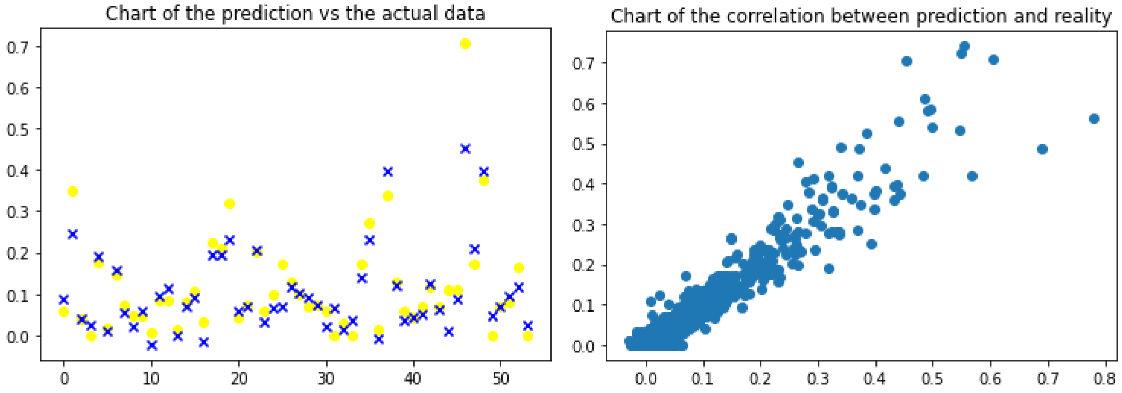


Figure 6. Performances of Linear Regression

Comparison between the two models

- In general, Random Forest has higher R2 and lower MSE values, which is desired as it indicates better predictions.
- Random Forest can make better use of the features compared to linear regression. Random Forest uses part of the features for each tree and gives the average values of all trees, this way can achieve low bias and low variance.
- The sales trend can be picked up by linear regression at the moment, however, future sales might not be so linear. Then Random Forest is a more safer choice than linear regression.
- With linear regression, predictions are becoming less accurate for high values of diary sales as shown in Figure 10. Random Forest is doing better for this section.

Recommendation to Favorita

We would recommend ***Random Forest Regression*** modelling to Favorita based on following considerations:

- Based on the comparison, it is clear that Random Forest has better general performance than linear regression.
- Even though Random Forest is not as explainable as linear regression, it is not making much sense to explain the models in this case. Because the features used in the models are mainly sales of other product families, and explaining this will not help to make a business decision anyway. It is more important to have good predictions in this business case.
- The two models have similar building complexity and Random Forest requires slightly more computation power than linear regression. Considering the small differences in this aspect, Random Forest wins over linear regression for being better with other aspects.

Bonus Predictions

Random Forest model was used to do the predictions of dairy sales for store 1 for each next day from 01/08/2017 to 15/08/2017 because of better performance between the two models (RMSE = 0.0293).

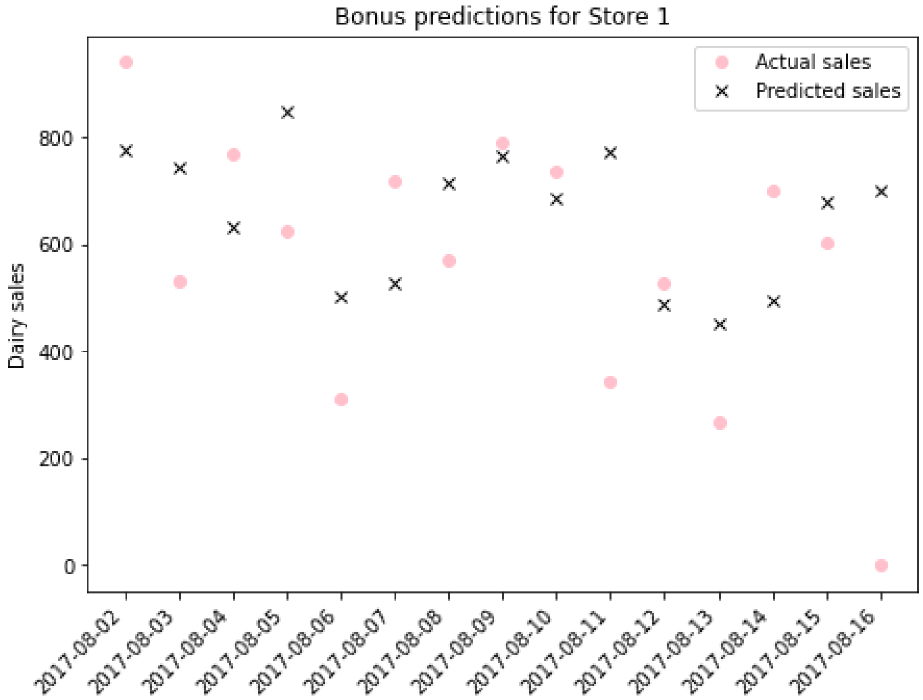


Figure 7. Bonus predictions for store 1 using random forest regression model

Appendix A

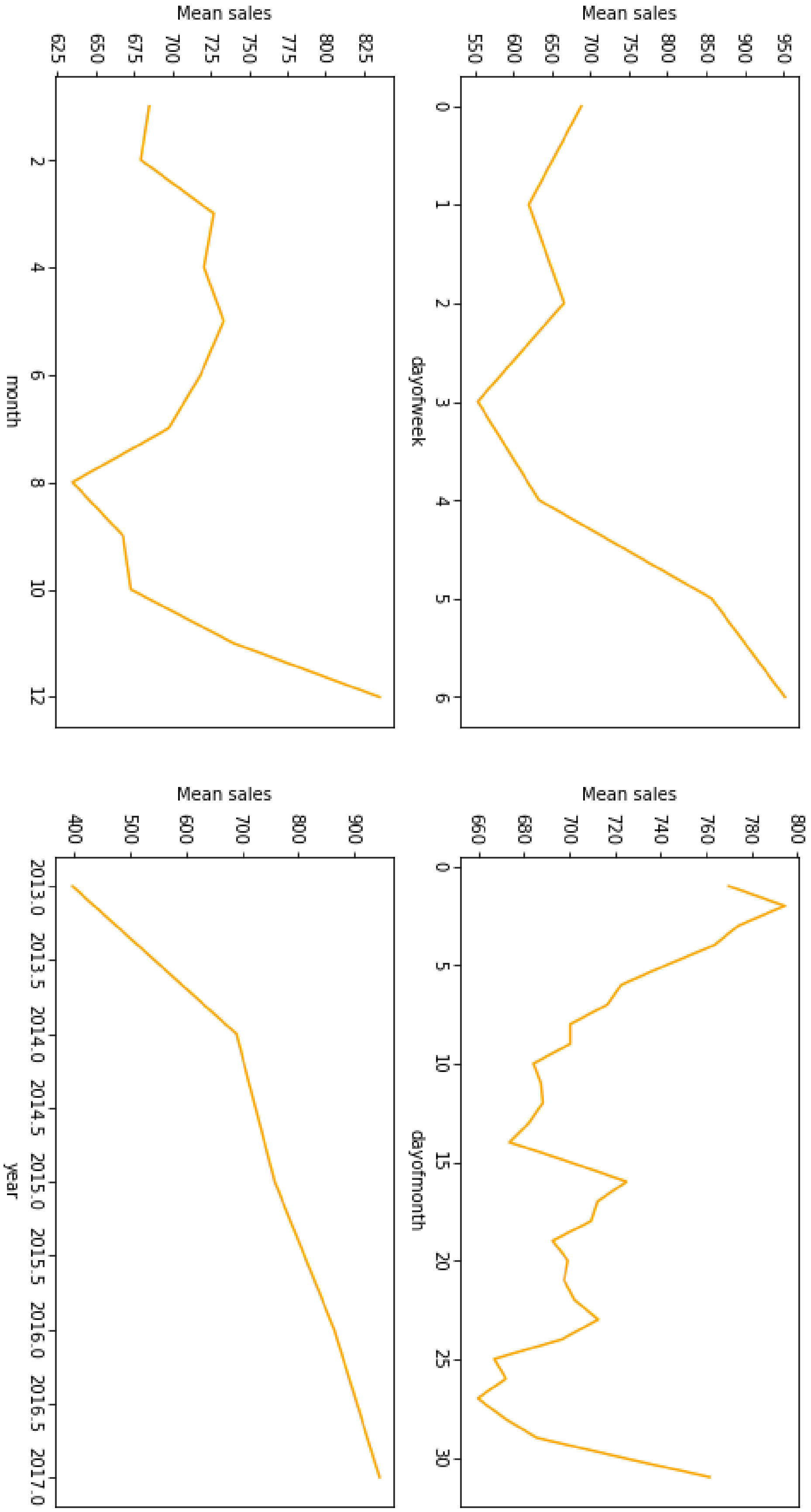
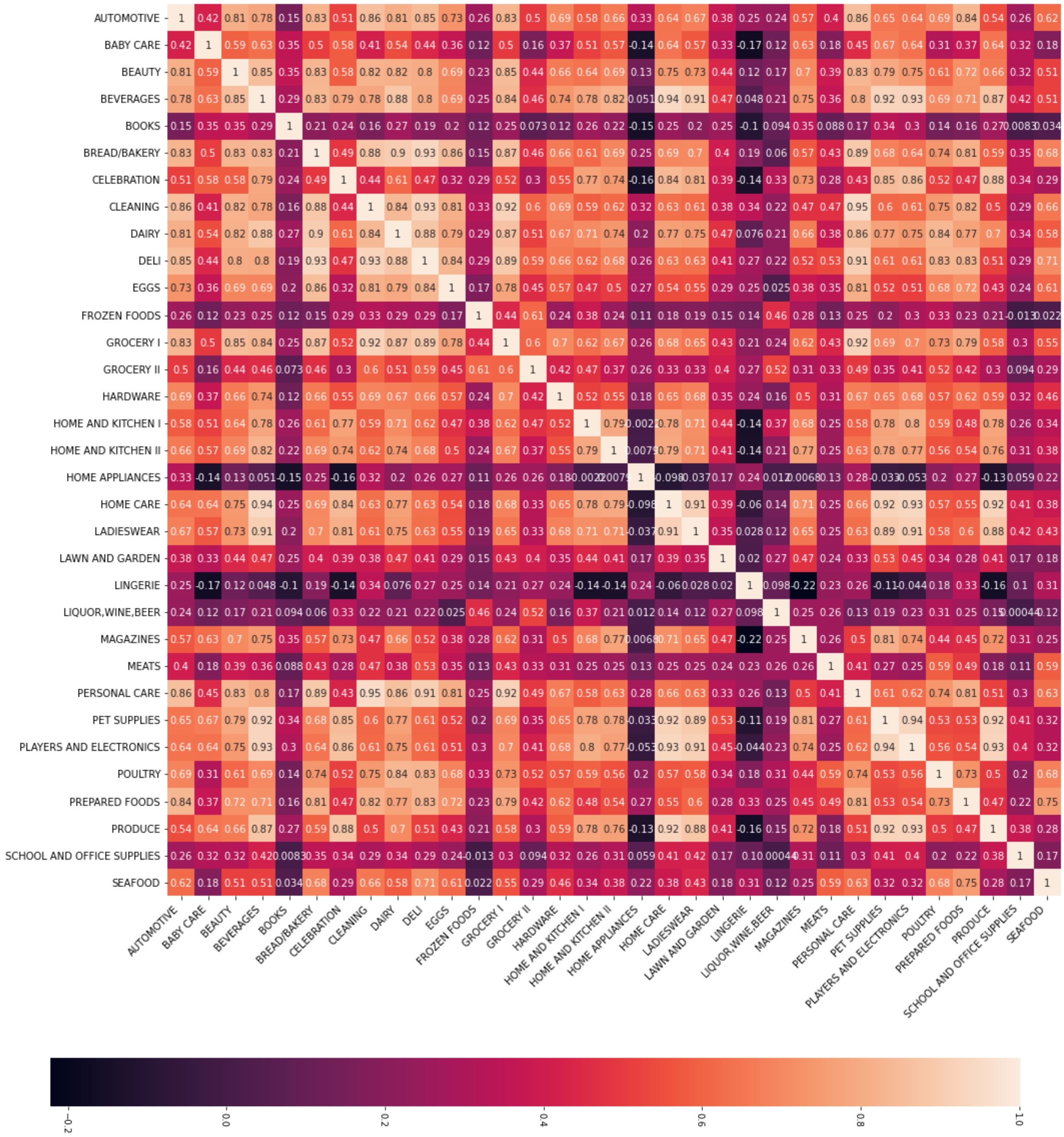


Figure A.1. Mean DAIRY sales across day of week, day of month, month of year, and year

Appendix B



	AUTOMOTIVE	BABY CARE	BEAUTY	BEVERAGES	BOOKS	BREAD/BAKERY	CELEBRATION	CLEANING	DAIRY	DELI	EGGS	FROZEN FOODS	GROCERY I	GROCERY II	HARDWARE	HOME AND KITCHEN I	HOME AND KITCHEN II	HOME APPLIANCES	HOME CARE	LADIESWEAR	LAWN AND GARDEN	LINGERIE	LIQUOR,WINE,BEER	MAGAZINES	MEATS	PERSONAL CARE	PET SUPPLIES	PLAYERS AND ELECTRONICS	POULTRY	PREPARED FOODS	PRODUCE	SCHOOL AND OFFICE SUPPLIES	SEAFOOD
AUTOMOTIVE	1	0.42	0.81	0.78	0.15	0.83	0.51	0.86	0.81	0.85	0.73	0.26	0.83	0.5	0.69	0.58	0.66	0.33	0.64	0.67	0.38	0.25	0.24	0.57	0.4	0.86	0.65	0.64	0.69	0.84	0.54	0.26	0.62
BABY CARE	0.42	1	0.59	0.63	0.35	0.5	0.58	0.41	0.54	0.44	0.36	0.12	0.5	0.16	0.37	0.51	0.57	-0.14	0.64	0.57	0.33	-0.17	0.12	0.63	0.18	0.45	0.67	0.64	0.31	0.37	0.64	0.32	0.18
BEAUTY	0.81	0.59	1	0.85	0.35	0.83	0.58	0.82	0.82	0.8	0.69	0.23	0.85	0.44	0.66	0.64	0.69	0.13	0.75	0.73	0.44	0.12	0.17	0.7	0.39	0.83	0.79	0.75	0.61	0.72	0.66	0.32	0.51
BEVERAGES	0.78	0.63	0.85	1	0.29	0.83	0.79	0.78	0.88	0.8	0.69	0.25	0.84	0.46	0.74	0.78	0.82	0.051	0.94	0.91	0.47	0.048	0.21	0.75	0.36	0.8	0.92	0.93	0.69	0.71	0.87	0.42	0.51
BOOKS	0.15	0.35	0.35	0.29	1	0.21	0.24	0.16	0.27	0.19	0.2	0.12	0.25	0.073	0.12	0.26	0.22	-0.15	0.25	0.2	0.25	-0.1	0.094	0.35	0.088	0.17	0.34	0.3	0.14	0.16	0.27	0.0083	0.034
BREAD/BAKERY	0.83	0.5	0.83	0.83	0.21	1	0.49	0.88	0.9	0.93	0.86	0.15	0.87	0.46	0.66	0.61	0.69	0.25	0.69	0.7	0.4	0.19	0.06	0.57	0.43	0.89	0.68	0.64	0.74	0.81	0.59	0.35	0.68
CELEBRATION	0.51	0.58	0.58	0.79	0.24	0.49	1	0.44	0.61	0.47	0.32	0.29	0.52	0.3	0.55	0.77	0.74	-0.16	0.84	0.81	0.39	-0.14	0.33	0.73	0.28	0.43	0.85	0.86	0.52	0.47	0.88	0.34	0.29
CLEANING	0.86	0.41	0.82	0.78	0.16	0.88	0.44	1	0.84	0.93	0.81	0.33	0.92	0.6	0.69	0.59	0.62	0.32	0.63	0.61	0.38	0.34	0.22	0.47	0.47	0.95	0.6	0.61	0.75	0.82	0.5	0.29	0.66
DAIRY	0.81	0.54	0.82	0.88	0.27	0.9	0.61	0.84	1	0.88	0.79	0.29	0.87	0.51	0.67	0.71	0.74	0.2	0.77	0.75	0.47	0.076	0.21	0.66	0.38	0.86	0.77	0.75	0.84	0.77	0.7	0.34	0.58
DELI	0.85	0.44	0.8	0.8	0.19	0.93	0.47	0.93	0.88	1	0.84	0.29	0.89	0.59	0.66	0.62	0.68	0.26	0.63	0.63	0.41	0.27	0.22	0.52	0.53	0.91	0.61	0.61	0.83	0.83	0.51	0.29	0.71
EGGS	0.73	0.36	0.69	0.69	0.2	0.86	0.32	0.81	0.79	0.84	1	0.17	0.78	0.45	0.57	0.47	0.5	0.27	0.54	0.55	0.29	0.25	0.025	0.38	0.35	0.81	0.52	0.51	0.68	0.72	0.43	0.24	0.61
FROZEN FOODS	0.26	0.12	0.23	0.25	0.12	0.15	0.29	0.33	0.29	0.29	0.17	1	0.44	0.61	0.24	0.38	0.24	0.11	0.18	0.19	0.15	0.14	0.46	0.28	0.13	0.25	0.2	0.3	0.33	0.23	0.21	-0.013	0.022
GROCERY I	0.83	0.5	0.85	0.84	0.25	0.87	0.52	0.92	0.87	0.89	0.78	0.44	1	0.6	0.7	0.62	0.67	0.26	0.68	0.65	0.43	0.21	0.24	0.62	0.43	0.92	0.69	0.7	0.73	0.79	0.58	0.3	0.55
GROCERY II	0.5	0.16	0.44	0.46	0.073	0.46	0.3	0.6	0.51	0.59	0.45	0.61	0.6	1	0.42	0.47	0.37	0.26	0.33	0.33	0.4	0.27	0.52	0.31	0.33	0.49	0.35	0.41	0.52	0.42	0.3	0.094	0.29
HARDWARE	0.69	0.37	0.66	0.74	0.12	0.66	0.55	0.69	0.67	0.66	0.57	0.24	0.7	0.42	1	0.52	0.55	0.18	0.65	0.68	0.35	0.24	0.16	0.5	0.31	0.67	0.65	0.68	0.57	0.62	0.59	0.32	0.46
HOME AND KITCHEN I	0.58	0.51	0.64	0.78	0.26	0.61	0.77	0.59	0.71	0.62	0.47	0.38	0.62	0.47	0.52	1	0.79	-0.0022	0.78	0.71	0.44	-0.14	0.37	0.68	0.25	0.58	0.78	0.8	0.59	0.48	0.78	0.26	0.34
HOME AND KITCHEN II	0.66	0.57	0.69	0.82	0.22	0.69	0.74	0.62	0.74	0.68	0.5	0.24	0.67	0.37	0.55	0.79	1	-0.0079	0.79	0.71	0.41	-0.14	0.21	0.77	0.25	0.63	0.78	0.77	0.56	0.54	0.76	0.31	0.38
HOME APPLIANCES	0.33	-0.14	0.13	0.051	-0.15	0.25	-0.16	0.32	0.2	0.26	0.27	0.11	0.26	0.26	0.18	-0.0022	-0.0079	1	-0.098	-0.037	0.17	0.24	0.012	0.0068	0.13	0.28	-0.033	-0.053	0.2	0.27	-0.13	0.059	0.22
HOME CARE	0.64	0.64	0.75	0.94	0.25	0.69	0.84	0.63	0.77	0.63	0.54	0.18	0.68	0.33	0.65	0.78	0.79	-0.098	1	0.91	0.39	-0.06	0.14	0.71	0.25	0.66	0.92	0.93	0.57	0.55	0.92	0.41	0.38
LADIESWEAR	0.67	0.57	0.73	0.91	0.2	0.7	0.81	0.61	0.75	0.63	0.55	0.19	0.65	0.33	0.68	0.71	0.71	-0.037	0.91	1	0.35	0.028	0.12	0.65	0.25	0.63	0.89	0.91	0.58	0.6	0.88	0.42	0.43
LAWN AND GARDEN	0.38	0.33	0.44	0.47	0.25	0.4	0.39	0.38	0.47	0.41	0.29	0.15	0.43	0.4	0.35	0.44	0.41	0.17	0.39	0.35	1	0.02	0.27	0.47	0.24	0.33	0.53	0.45	0.34	0.28	0.41	0.17	0.18
LINGERIE	0.25	-0.17	0.12	0.048	-0.1	0.19	-0.14	0.34	0.076	0.27	0.25	0.14	0.21	0.27	0.24	-0.14	-0.14	0.24	-0.06	0.028	0.02	1	0.098	-0.22	0.23	0.26	-0.11	-0.044	0.18	0.33	-0.16	0.1	0.31
LIQUOR,WINE,BEER	0.24	0.12	0.17	0.21	0.094	0.06	0.33	0.22	0.21	0.22	0.025	0.46	0.24	0.52	0.16	0.37	0.21	0.012	0.14	0.12	0.27	0.098	1	0.25	0.26	0.13	0.19	0.23	0.31	0.25	0.15	0.00044	0.12
MAGAZINES	0.57	0.63	0.7	0.75	0.35	0.57	0.73	0.47	0.66	0.52	0.38	0.28	0.62	0.31	0.5	0.68	0.77	0.0068	0.71	0.65	0.47	-0.22	0.25	1	0.26	0.5	0.81	0.74	0.44	0.45	0.72	0.31	0.25
MEATS	0.4	0.18	0.39	0.36	0.088	0.43	0.28	0.47	0.38	0.53	0.35	0.13	0.43	0.33	0.31	0.25	0.25	0.13	0.25	0.25	0.24	0.23	0.26	0.26	1	0.41	0.27	0.25	0.59	0.49	0.18	0.11	0.59
PERSONAL CARE	0.86	0.45	0.83	0.8	0.17	0.89	0.43	0.95	0.86	0.91	0.81	0.25	0.92	0.49	0.67	0.58	0.63	0.28	0.66	0.63	0.33	0.26	0.13	0.5	0.41	1	0.61	0.62	0.74	0.81	0.51	0.3	0.63
PET SUPPLIES	0.65	0.67	0.79	0.92	0.34	0.68	0.85	0.6	0.77	0.61	0.52	0.2	0.69	0.35	0.65	0.78	0.78	-0.033	0.92	0.89	0.53	-0.11	0.19	0.81	0.27	0.61	1	0.94	0.53	0.53	0.92	0.41	0.32
PLAYERS AND ELECTRONICS	0.64	0.64	0.75	0.93	0.3	0.64	0.86	0.61	0.75	0.61	0.51	0.3	0.7	0.41	0.68	0.8	0.77	-0.053	0.93	0.91	0.45	-0.044	0.23	0.74	0.25	0.62	0.94	1	0.56	0.54	0.93	0.4	0.32
POULTRY	0.69	0.31	0.61	0.69	0.14	0.74	0.52	0.75	0.84	0.83	0.68	0.33	0.73	0.52	0.57	0.59	0.56	0.2	0.57	0.58	0.34	0.18	0.31	0.44	0.59	0.74	0.53	0.56	1	0.73	0.5	0.2	0.68
PREPARED FOODS	0.84	0.37	0.72	0.71	0.16	0.81	0.47	0.82	0.77	0.83	0.72	0.23	0.79	0.42	0.62	0.48	0.54	0.27	0.55	0.6	0.28	0.33	0.25	0.45	0.49	0.81	0.53	0.54	0.73	1	0.47	0.22	0.75
PRODUCE	0.54	0.64	0.66	0.87	0.27	0.59	0.88	0.5	0.7	0.51	0.43	0.21	0.58	0.3	0.59	0.78	0.76	-0.13	0.92	0.88	0.41	-0.16	0.15	0.72	0.18	0.51	0.92	0.93	0.5	0.47	1	0.38	0.28
SCHOOL AND OFFICE SUPPLIES	0.26	0.32	0.32	0.42	0.0083	0.35	0.34	0.29	0.34	0.29	0.24	-0.013	0.3	0.094	0.32	0.26	0.31	0.059	0.41	0.42	0.17	0.1	0.00044	0.31	0.11	0.3	0.41	0.4	0.2	0.22	0.38	1	0.17
SEAFOOD	0.62	0.18	0.51	0.51	0.034	0.68	0.29	0.66	0.58	0.71	0.61	0.022	0.55	0.29	0.46	0.34	0.38	0.22	0.38	0.43	0.18	0.31	0.12	0.25	0.59	0.63	0.32	0.32	0.68	0.75	0.28	0.17	1

Figure B.1. Correlation between sales of all product families

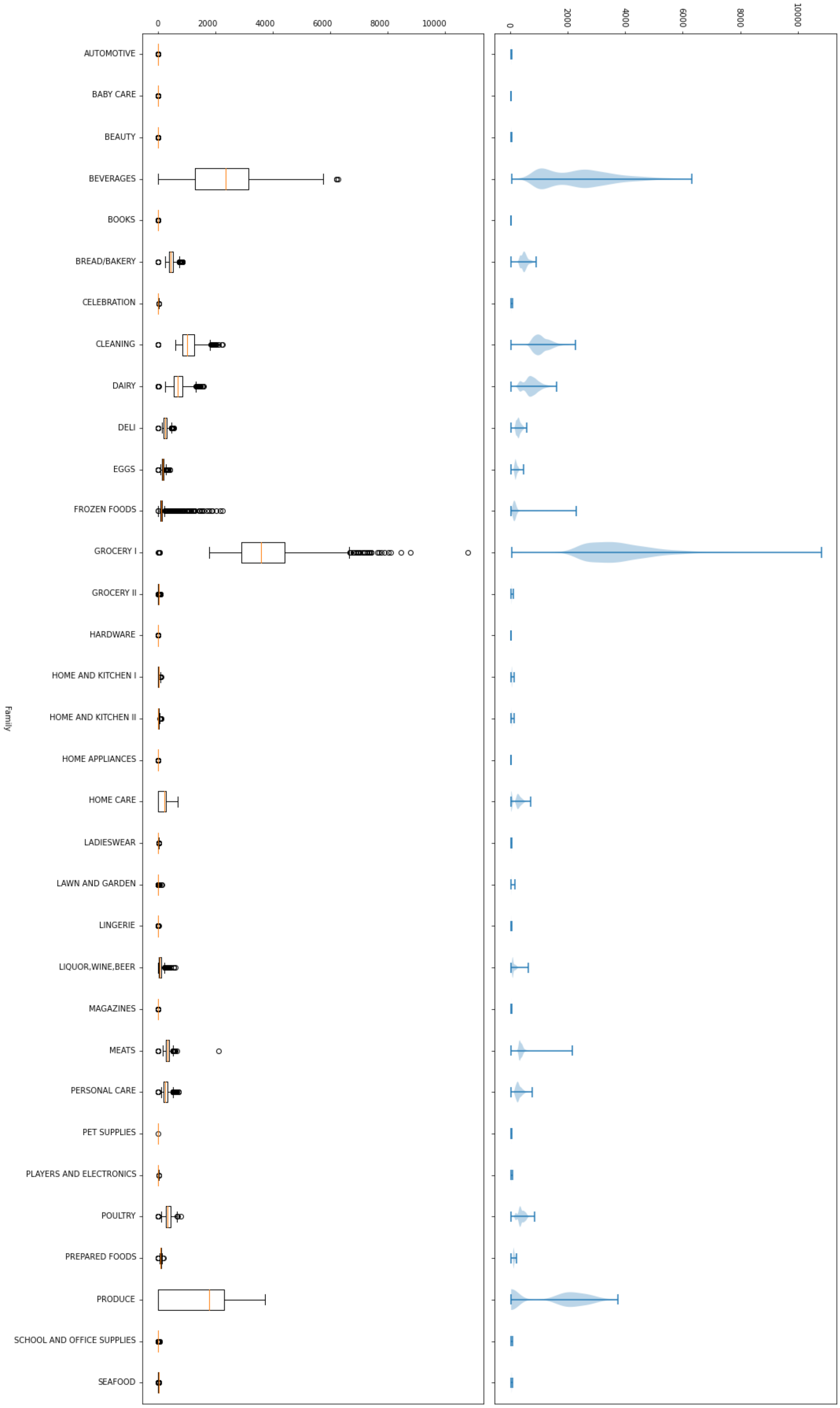


Figure B.2. Mean overall sales across all product families

Appendix C

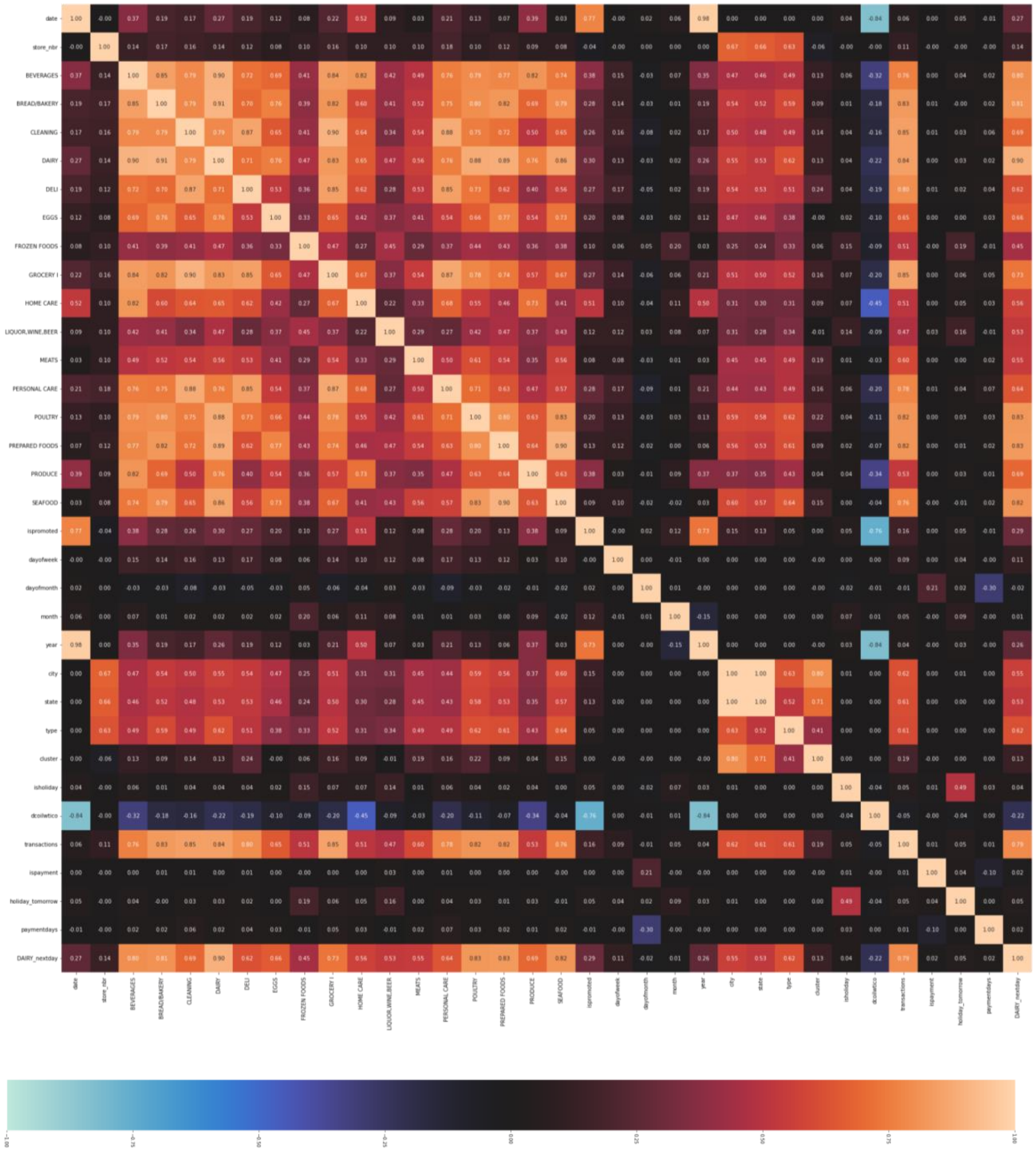


Figure C.1. Correlation between all features

Note: Submitted files.

- ***Jupyter notebooks for Python code:***
  - FavoritaEDA_29.ipynb (complete preprocessing and model training)
  - FavoritaModeling_29.ipynb (model training on preprocessed data)
- ***Data files:***
  - final_data_29.csv (preprocessed data to be used for model training)
  - results29.csv (bonus predictions for store 1)