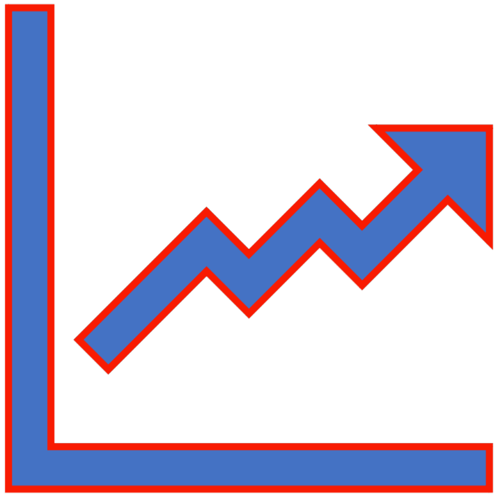


Sales Analyzing and Weekly Sales Forecasting

Rossman

Yang Liu Kunz

TOC

Overview

The aim of this project:

- Explore and analyze historical sales related data for Rossmann.
- Identify the key factors that influence store sales.
- Provide overall sales performance report to the management team with feasible strategies to increase future sales.
- Provide individual store performance report to each store manager.
- Develop time series model that predict the future sales for each store with certain level of accuracy.

Dataset① train.csv

- 1,017,209 data samples
- 9 variables (7 num, 2 cat)
- Date: 2013/01/01 ~ 2015/07/31

Column Name	**counts**	**unique_value_pct**	**nan_pct**	**data_type**
Sales	21734	2.14	0.0	int64
Customers	4086	0.40	0.0	int64
Store	1115	0.11	0.0	int64
Date	942	0.09	0.0	object
DayOfWeek	7	0.0007	0.0	int64
StateHoliday	5	0.0005	0.0	object
Open	2	0.0002	0.0	int64
Promo	2	0.0002	0.0	int64
SchoolHoliday	2	0.0002	0.0	int64

StateHoliday: a = public holiday, b = Easter holiday, c = Christmas, 0 = None

	Store	DayOfWeek	Date	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday
0	1	5	2015-07-31	5263	555	1	1	0	1
1	2	5	2015-07-31	6064	625	1	1	0	1
2	3	5	2015-07-31	8314	821	1	1	0	1
3	4	5	2015-07-31	13995	1498	1	1	0	1
4	5	5	2015-07-31	4822	559	1	1	0	1

- train.csv (head5)

Dataset② store.csv

- 1115 data samples
- 10 variables (7 num, 3 cat)

Colum Name	counts	unique_value_pct	nan_pct	data_type
Store	1115	100.00	0.0000	int64
CompetitionDistance	654	58.65	0.27	float64
Promo2SinceWeek	24	2.15	48.80	float64
CompetitionOpenSince Year	23	2.06	31.75	float64
CompetitionOpenSince Month	12	1.08	31.75	float64
Promo2SinceYear	7	0.63	48.80	float64
StoreType	4	0.36	0.0000	object
Assortment	3	0.27	0.0000	object
PromoInterval	3	0.27	48.80	object
Promo2	2	0.18	0.0000	int64

	Store	StoreType	Assortment	CompetitionDistance	CompetitionOpenSinceMonth	CompetitionOpenSinceYear	Promo2	Promo2SinceWeek	Promo2SinceYear	PromoInterval
0	1	c	a	1270.0	9.0	2008.0	0	NaN	NaN	NaN
1	2	a	a	570.0	11.0	2007.0	1	13.0	2010.0	Jan,Apr,Jul,Oct
2	3	a	a	14130.0	12.0	2006.0	1	14.0	2011.0	Jan,Apr,Jul,Oct
3	4	c	c	620.0	9.0	2009.0	0	NaN	NaN	NaN
4	5	a	a	29910.0	4.0	2015.0	0	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...
1110	1111	a	a	1900.0	6.0	2014.0	1	31.0	2013.0	Jan,Apr,Jul,Oct
1111	1112	c	c	1880.0	4.0	2006.0	0	NaN	NaN	NaN
1112	1113	a	c	9260.0	NaN	NaN	0	NaN	NaN	NaN
1113	1114	a	c	870.0	NaN	NaN	0	NaN	NaN	NaN
1114	1115	d	c	5350.0	NaN	NaN	1	22.0	2012.0	Mar,Jun,Sept,Dec

- store.csv

Data Processing

Steps	Action	Variable Names	Detail explanation	Dataset name
Step 1	Data Type Correction	Date	Convert object to datetime	train.csv
		StateHoliday	Convert object to int	tain.csv
Step 2	New Variable Creation	Year	Add new variable 'Year' by extracting year value from variable 'Date'	train.csv
		Month	Add new variable 'Month by extracting month value from variable 'Date'	train.csv
Step 3	Aggregation	All variables	Sum up all the values of all the variables for each store in train.csv	store_sum.csv (name after aggregation)
Step 4	New Variable Creation	AverageDailySales	AverageDailySales = Sum of 'Sales' ÷ Sum of 'Open'	store_sum.csv
		AverageDailyCustomer	AverageDailyCustomer = Sum of 'Customers' ÷ Sum of 'Open'	store_sum.csv
		SalesPerCustomer	SalesPerCustomer = Sum of 'Sales' ÷ Sum of 'Customers'	store_sum.csv
Step 5	Datasets Combination	-	Combine store_sum.csv with store.csv	combine_data.csv (name after combination)

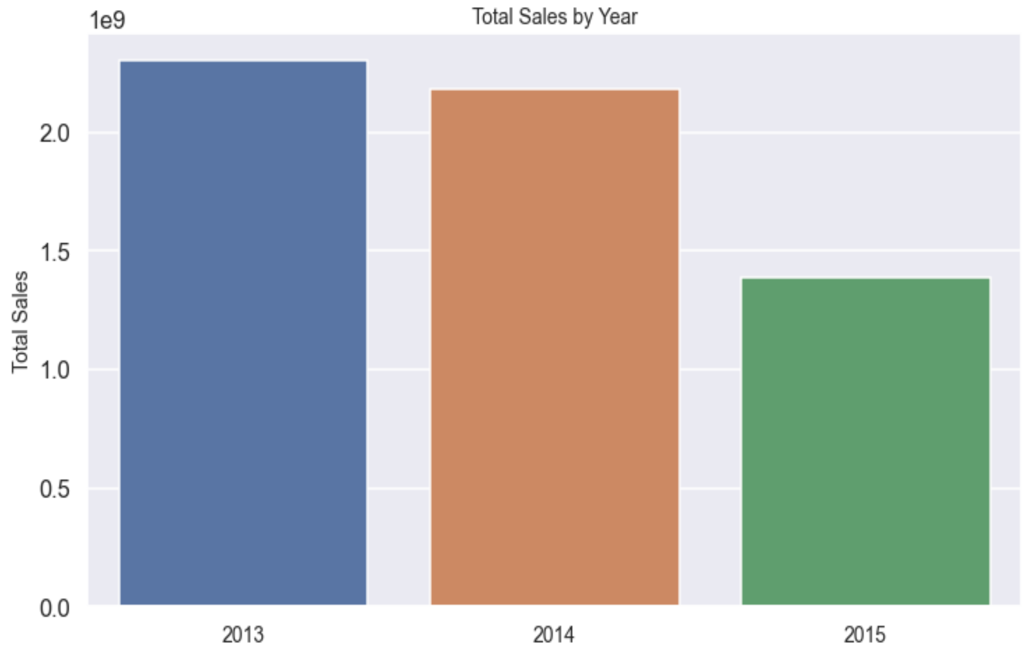
	Store	Customers	Open	Promo	Sales	SchoolHoliday	StateHoliday	AverageDailySales	AverageDailyCustomer	SalesPerCustomer	StoreType	Assortment	CompetitionDistance	Promo2
0	1	440523	781	360	3,716,854	193	27	4,759	564	8.44	c	a	1270	0
1	2	457855	784	360	3,883,858	167	25	4,954	584	8.48	a	a	570	1
2	3	584310	779	360	5,408,261	170	29	6,943	750	9.26	a	a	14130	1
3	4	1036254	784	360	7,556,507	173	24	9,638	1,322	7.29	c	c	620	0
4	5	418588	779	360	3,642,818	172	31	4,676	537	8.70	a	a	29910	0

- combined_data.csv

Key Findings

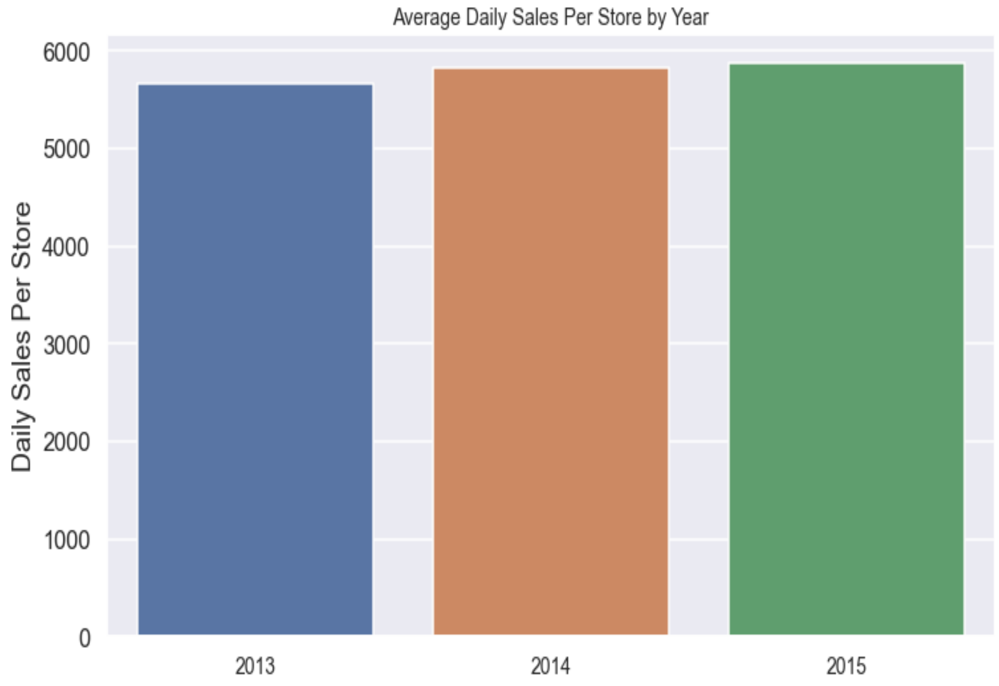
Year	Total Sales	Percentage Change
2013	2.303E+09	NaN
2014	2.181E+09	-5.3%
2015	1.389E+09	-36.3%

*Please be notified that the total sales of 2015 are not a full year data
*Total sales decreased 5.3% in 2014 as 180 stores were closed for renovation for a few months



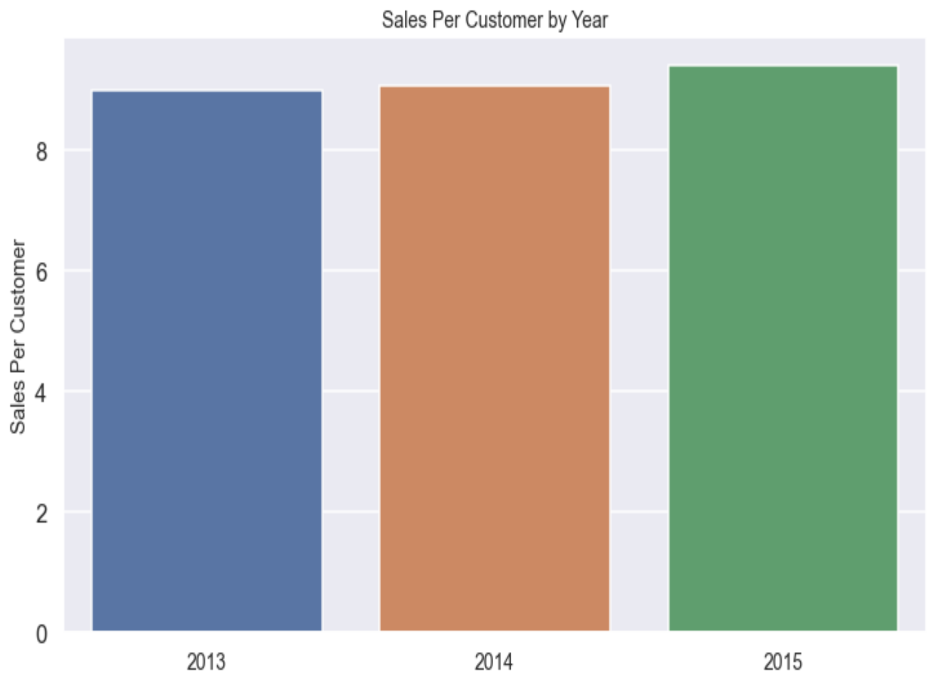
Total Sales by Year

Year	Daily Sales Per Store	Percentage Change
2013	5,659	NaN
2014	5,833	3.1%
2015	5,878	0.7%

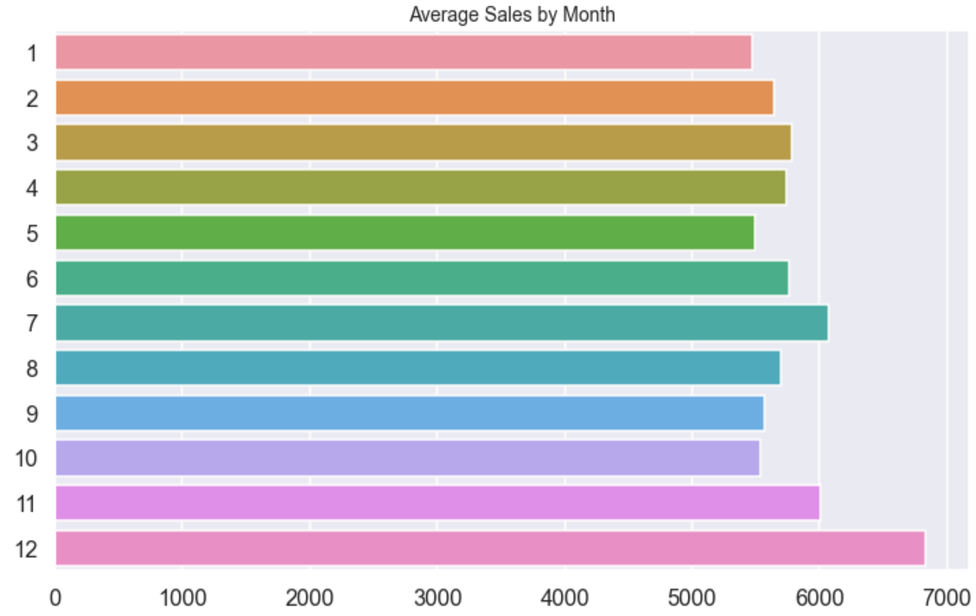


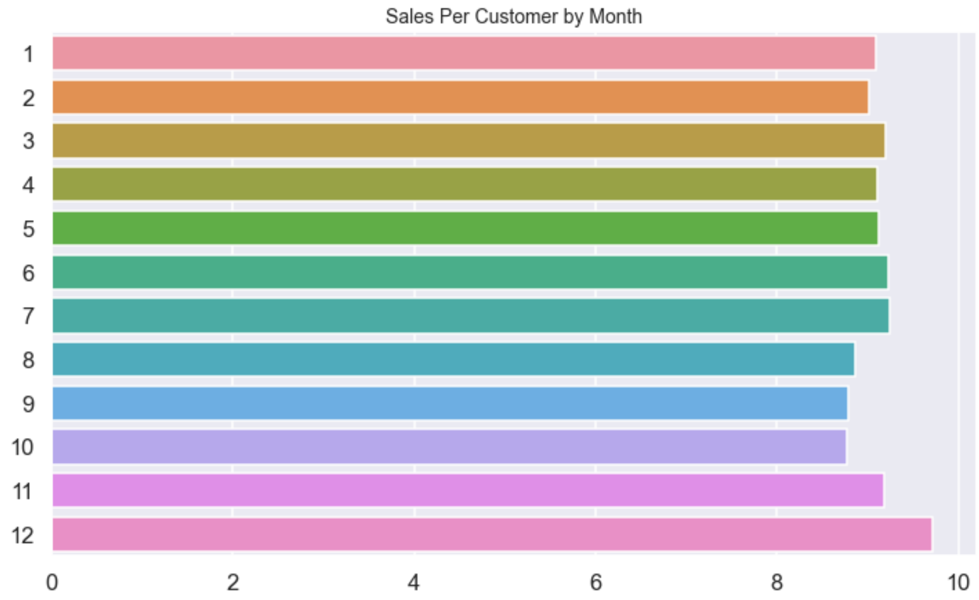
Daily Sales Per Store by Year

Year	Sales Per Customer	Percentage Change
2013	8.995	NaN
2014	9.068	0.8%
2015	9.417	3.8%



Sales Per Customer by Year

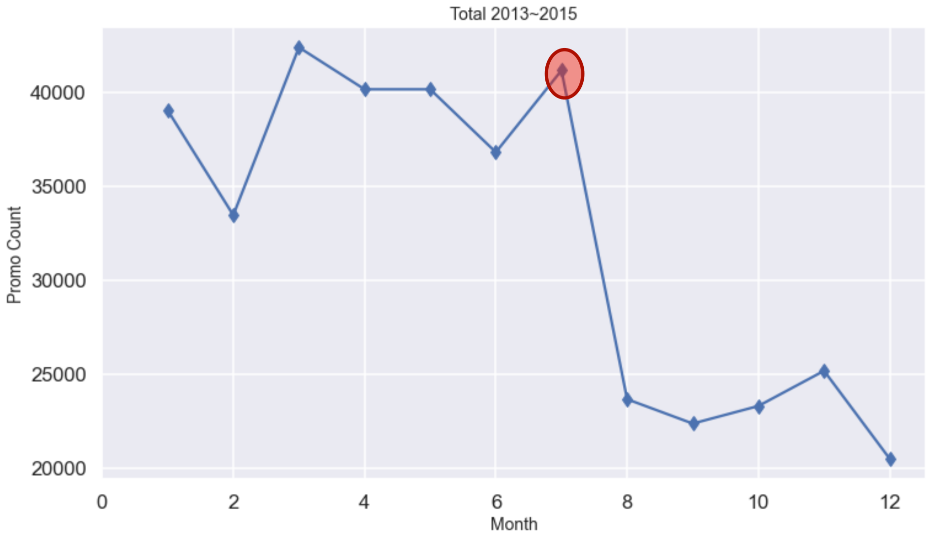


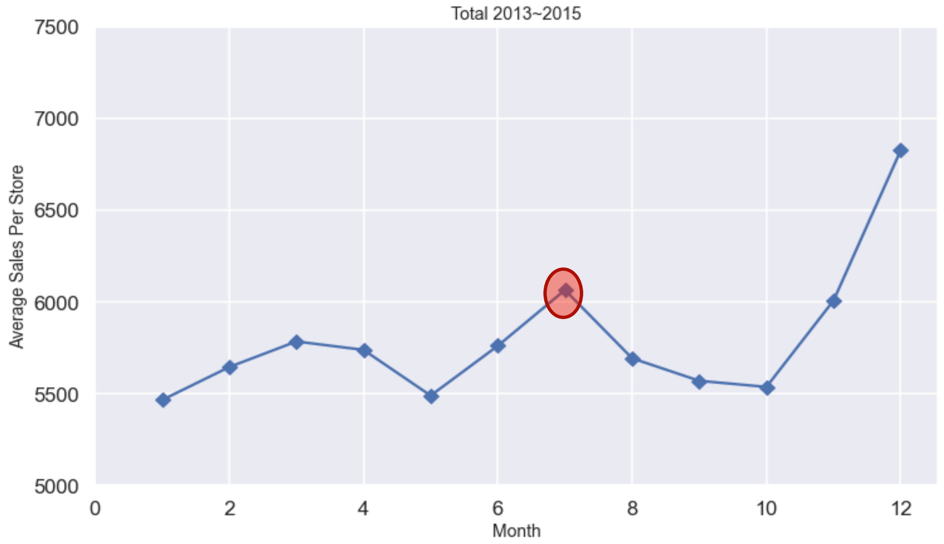


- December and July have the highest and second highest **daily sales per store**

- December and July have the highest and second highest **sales per customer**

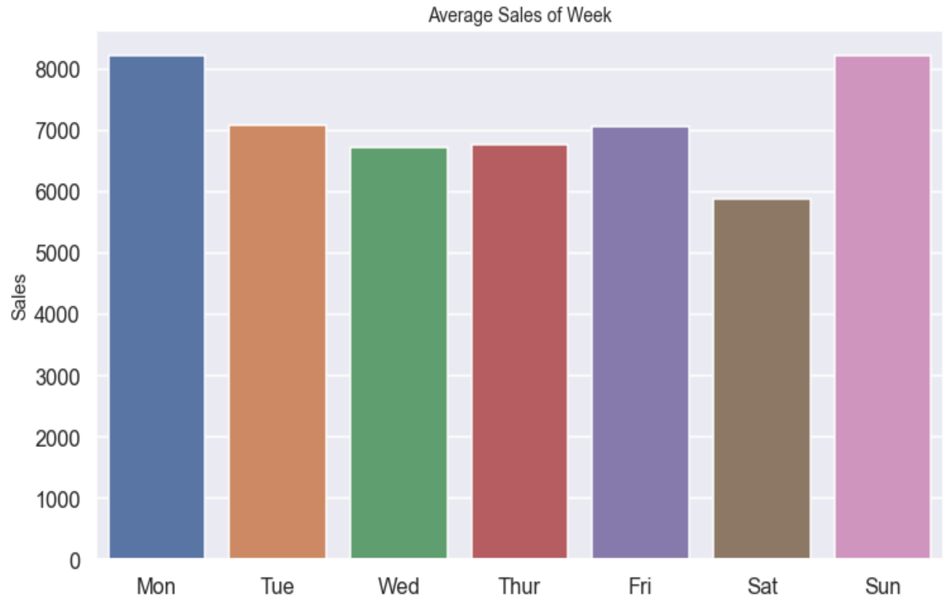
Average Sales by Month

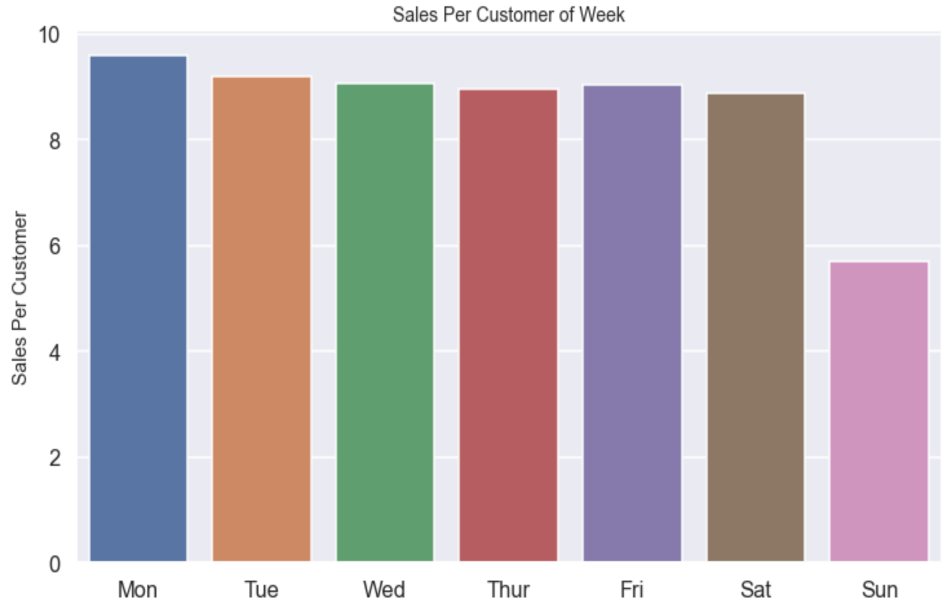




- July has the highest promotion counts in summer season. This strategy positively affected average daily sales per store in July. Also, we can see winter and autumn seasons are the off seasons of our business.

Promotion by Month





- Sunday has the second highest daily sales (per store) in a week, but its sales per customer is the lowest.

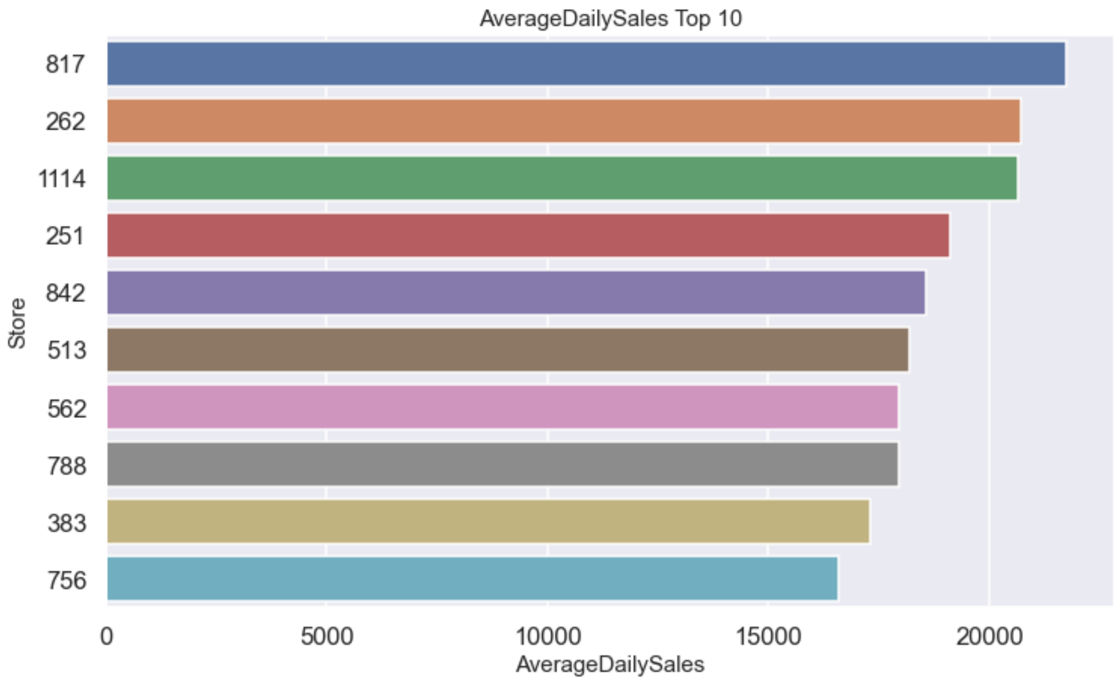
(possible reason: no promotion on the weekend, too crowded in the store on Sunday)

Average Sales by Day of Week

Ranking	Store id	Sales	Percentage %
1	262	19,516,842	0.332304
2	817	17,057,867	0.290437
3	562	16,927,322	0.288214
4	1114	16,202,585	0.275874
5	251	14,896,870	0.253642
6	513	14,252,406	0.242669
7	788	14,082,141	0.23977
8	733	14,067,158	0.239515
9	383	13,489,879	0.229686
10	756	12,911,782	0.219843
		Total Percentage	2.60

- Top 400 stores and the rest 715 stores' sales shares are roughly about 1:1.
- Half of the total sales (5.873 billion over 942 days) came from the top 400 stores and the rest 751 stores generated the other half.

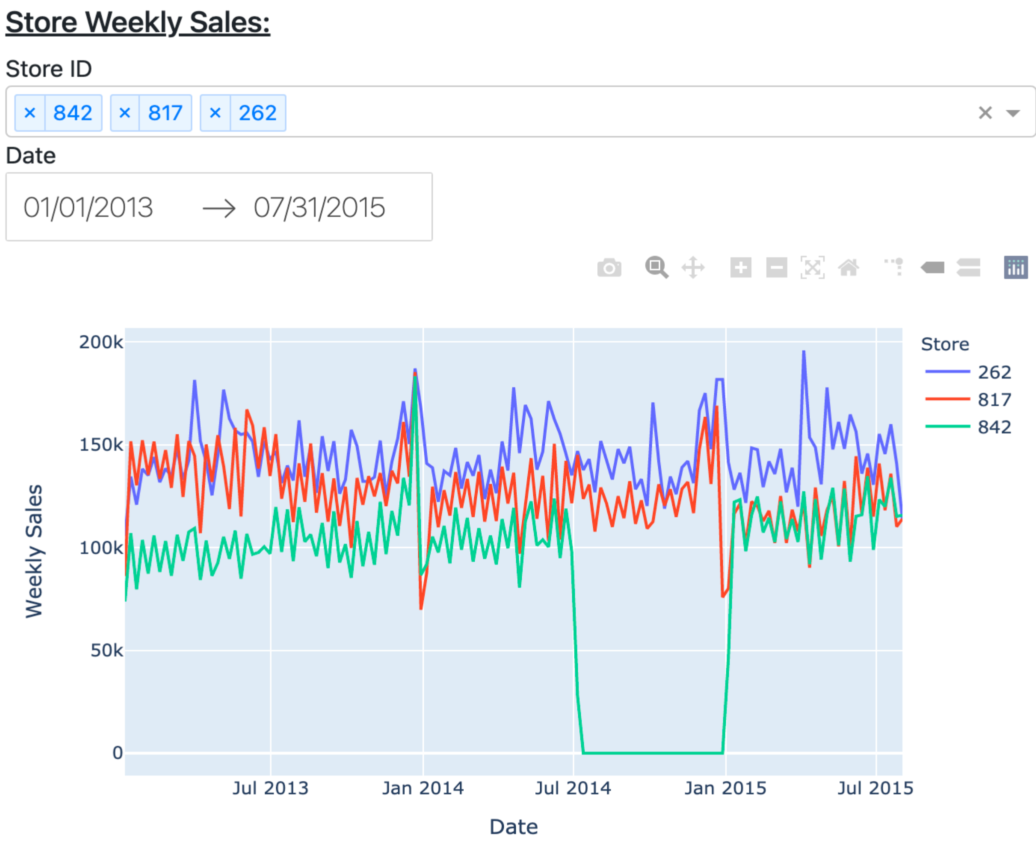
Total Sales Share Break Down by Stores



Store id	Sales Per Customer Ranking
817	1049
262	1099
1114	1081
251	919
842	1
513	717
562	1103
788	367
383	909
756	1038

- Stores which has top 10 highest average daily sales, tend to have very low value of sales per customer

Average Daily Sales Top 10

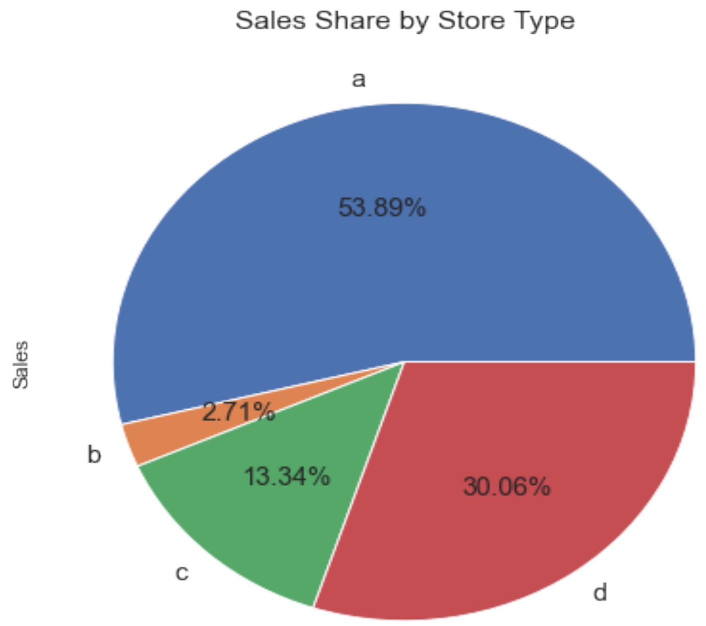


- Store842 has the highest value of sales per customer but it didn't rank in the top 10 total sales.
- As we can see from the left graph, store 842 stayed closed for half year due to renovation.
- Store262, Store817 has the highest and second highest value of total sales.)

Store842 Weekly Sales

Store Type	Count	Total Sales
a	602	3.16E+09
d	348	1.77E+09
c	148	7.83E+08
b	17	1.59E+08

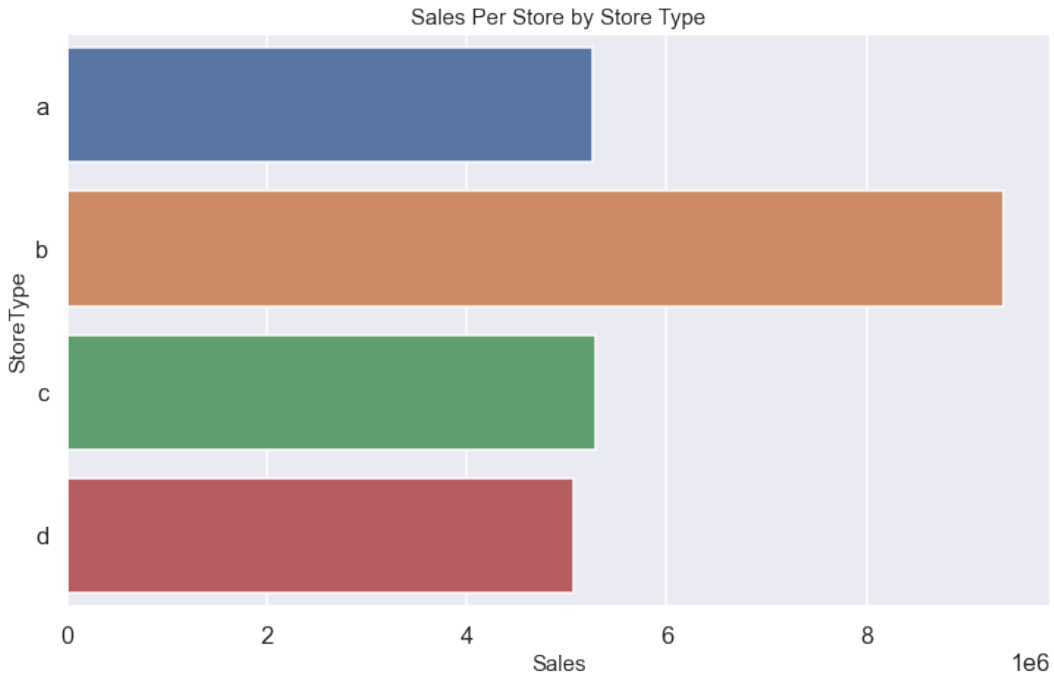
- A type has the highest value of store count, and total sales share



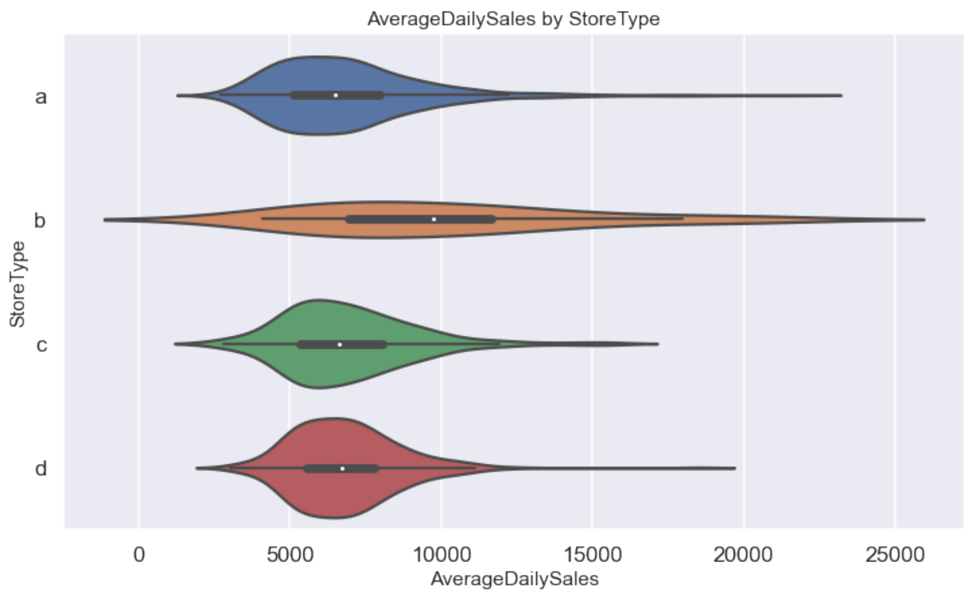
Sales by Store Type

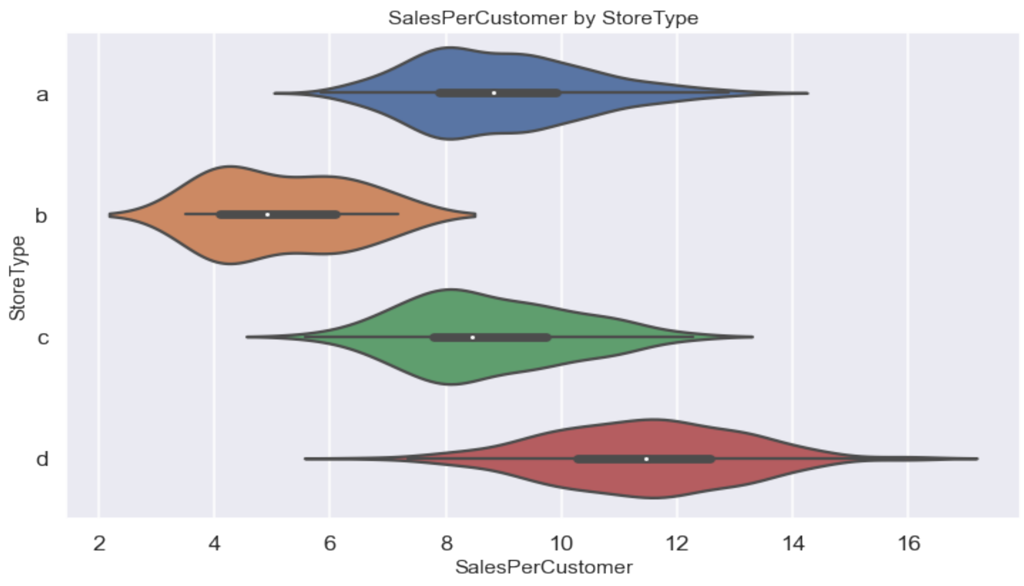
Store Type	Total Sales Per Store
a	5.26E+06
d	5.07E+06
c	5.29E+06
b	9.37E+06

- B type has the highest value of total sales per store



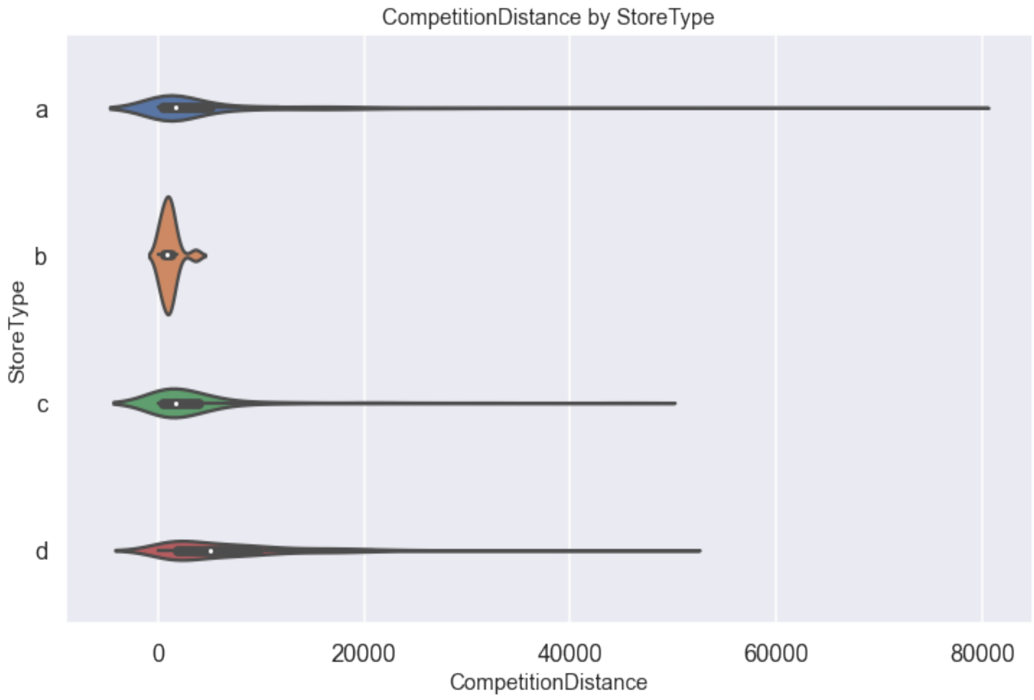
Sales by Store Type

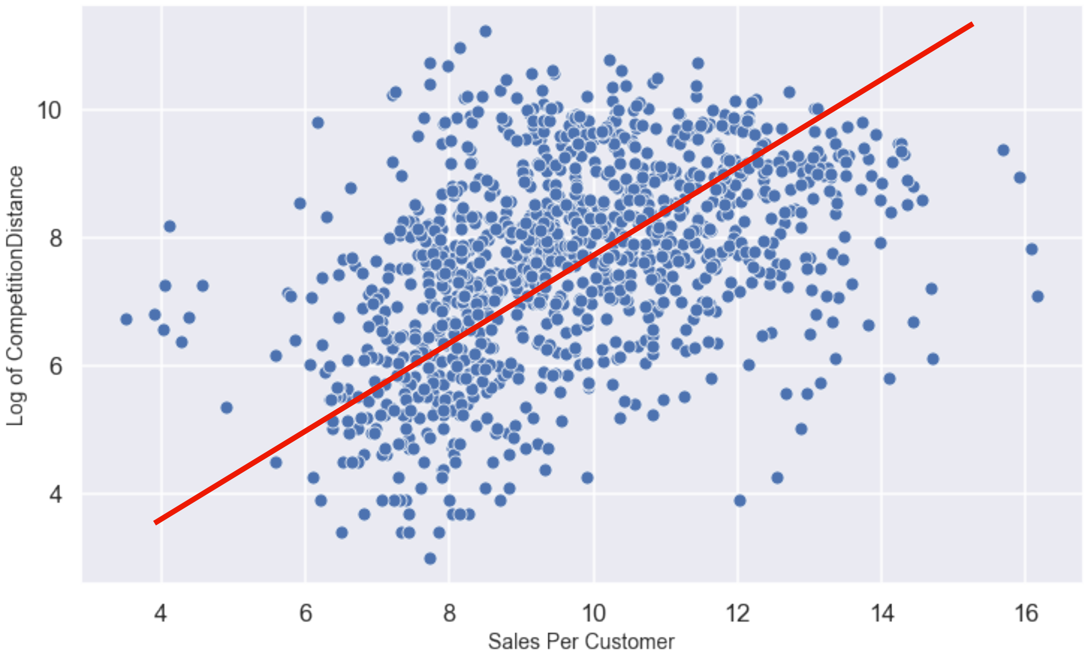




Store Type	Average Daily Sales	Sales Per Customer
a	6913	8.96
d	6824	11.43
c	6917	8.74
b	10111	5.17

Sales by Store Type





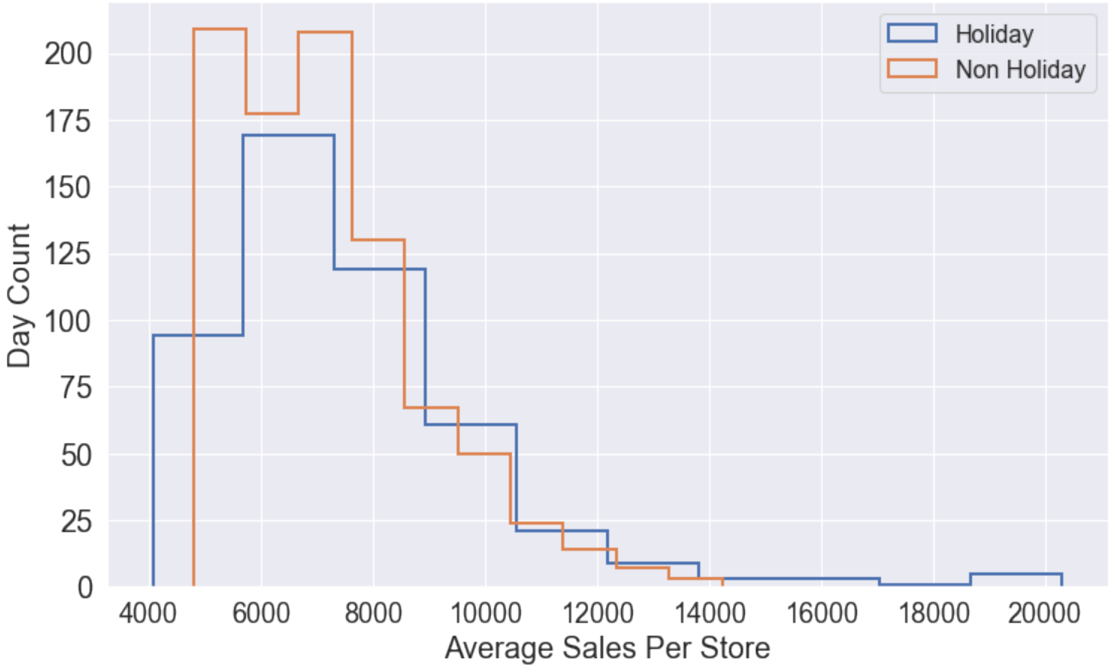
- B type stores have competitors more near by comparing to other 3 types.

- We observed a liner relationship between the log of competition distance and sales per customer
- The log of competition decrease, the value of sales per customer decrease correspondingly.

Competition Distance by Store Type

Holiday	Average Daily Sales	Sales Per Customer
Yes	7614	8.59
No	7208	8.53

- Holidays (state or school holidays) are affecting the amounts of sales.
- When it's a holiday, average sales per store is 5% higher than when it's not a holiday
- There were more daily sales higher than 14,000 when it's a holiday.



Holiday and Sales

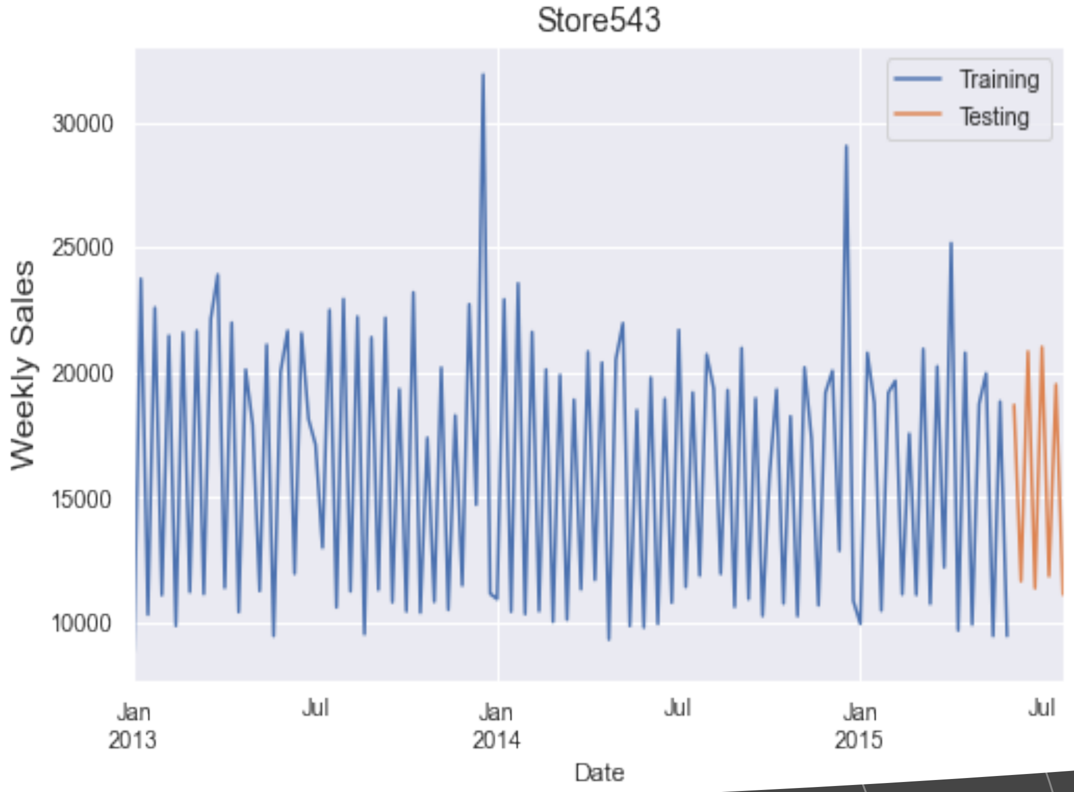
Modeling & Analysis

Data Processing and Testing Strategy

- When there's no zero weekly sales over the 135 weeks, 942 days, we split the dataset following **strategy 1**

Strategy 1:

Train (!27 weeks)	Test (8 weeks)	Forecast (8 weeks)
1/6/2013 ~ 5/31/2015	6/7/2015 ~ 7/26/2015	8/2/2015 ~ 9/20/2015



- When there’s continuing zero weekly sales over the 135 weeks (stores closed for renovation), we will do the train/test split following **strategy 2**

Strategy 2:

Train (22 weeks)	Test (8 weeks)	Forecast (8 weeks)
1/4/2015 ~ 5/31/2015	6/7/2015 ~ 7/26/2015	8/2/2015 ~ 9/20/2015

Strategy 3: If we don’t have enough data points after the store reopen (less than 16 weeks), we will then drop the zero values in the time series and do train/test split.

Data Processing and Testing Strategy

Baseline Model, Model and Metrics selection

Baseline model
- **Mean Method**: predict and forecast by using the average of the time series

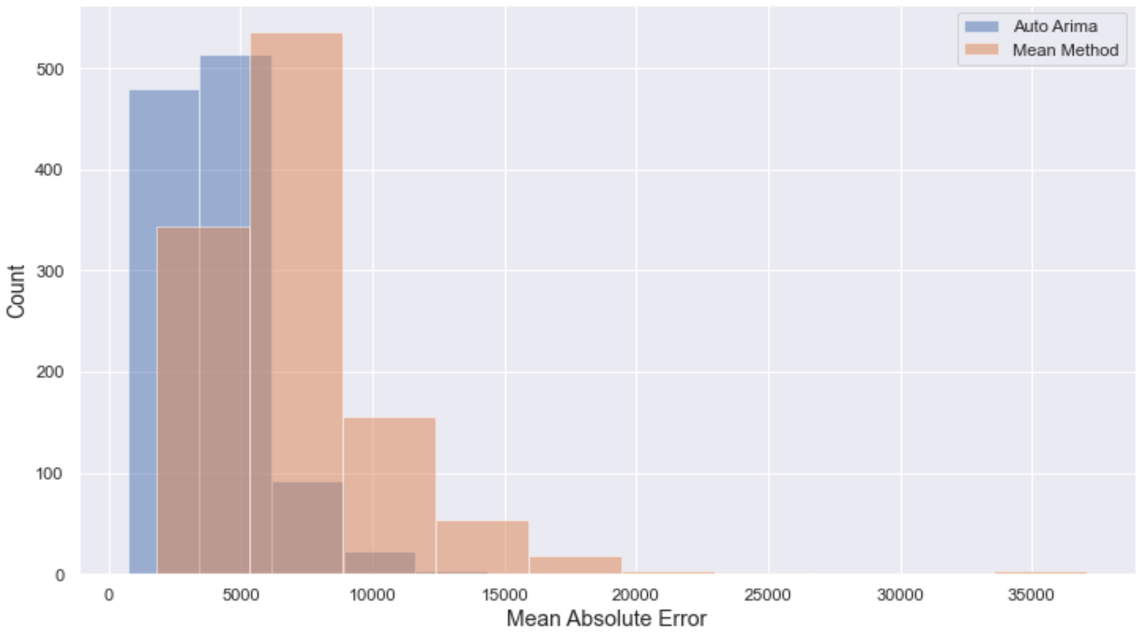
Model
- **Auto Arima:** generate optimal p,d,q for seasonal and non seasonal time series

Model evaluation metrics

MAE
- we need a robust model for each store that won't be influenced by outliers in the time series

WAPE
- Check model robustness over 8 weeks of prediction

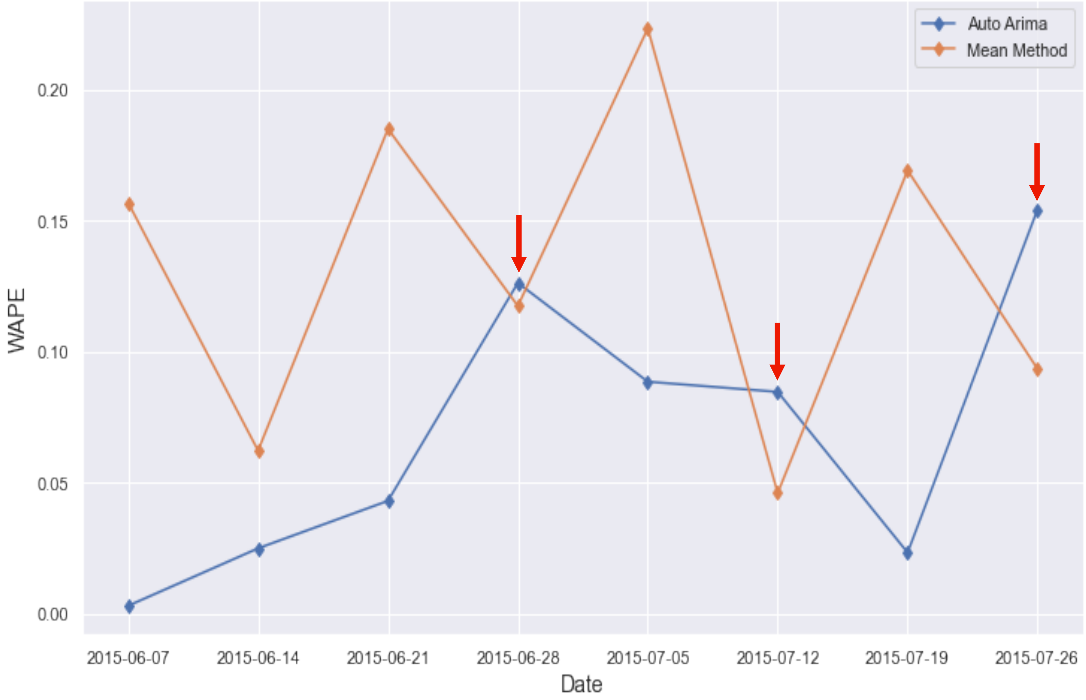
	MAE_AutoArima	MAE_MeanMethod	Performance Improvement
Average	4084	7143	1.75



- By comparing the average value of MAE we obtained from 1115 stores, we can see Auto Arima outperformed mean method by 1.75 times.

Model Evaluation - MAE

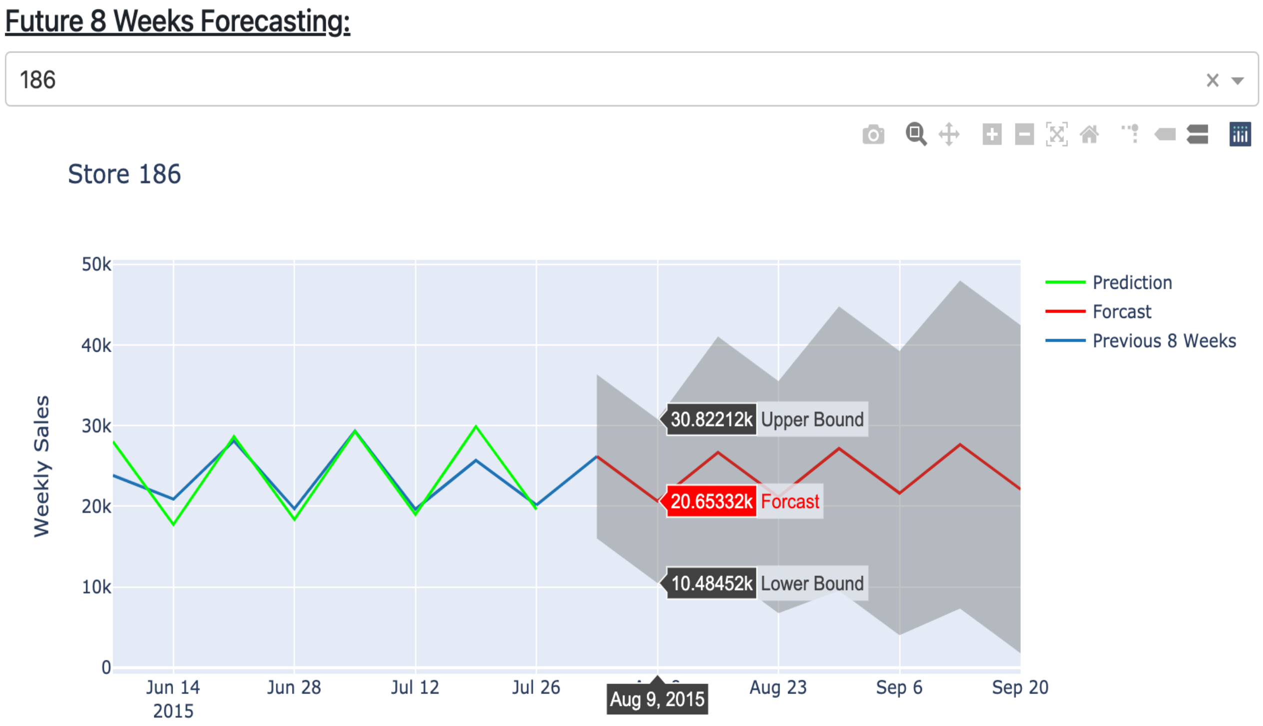
Date	WAPE_AutoArima	WAPE_MeanMethod
2015-06-07	0.0030	0.1568
2015-06-14	0.0249	0.0622
2015-06-21	0.0430	0.1850
2015-06-28	0.1262	0.1174
2015-07-05	0.0885	0.2233
2015-07-12	0.0846	0.0460
2015-07-19	0.0233	0.1691
2015-07-26	0.1542	0.0934



- We can see from the line plot that within 8 WAPE results, there were 5 of them are showing that Auto Arima model performed better than mean method, except week 6/28, week 7/12 and week 7/26.

Conclustion: Auto Arima model is our final model

Model Evaluation - WAPE



- After identifying the best model we used python dash package and deployed the model on Heroku cloud service.
- Enabling the store manager to confirm the forecast together with 95% confidence interval and with past 8 weeks sales and prediction values.

Model Deployment

Summary and Recommendation

Summary 1

Comparing to 2013, sales in 2014 decreased by 5.3% due to 180 stores closed for renovation. 5 of them ranked in top 10 stores that generated high values of sales per customer.

Recommendation 1

Try not to close these stores at the same time for renovation in the future if possible. (Store 842, Store 612, Store 52, Store 540 and Store 903).

Summary and Recommendation 1

Summary 2

Winter and autumn time are the off seasons of our business. They are also the seasons that we didn't run much daily promotions comparing to summer and winter seasons.

Recommendation 2

Increase daily promotions in Jan, Feb, Sept and Oct.

Summary and Recommendation 2

Summary 3

Stores that have high average daily sales tend to have low value of sales per customer.

Recommendation 3

Increase the value of sales per customer for Top 10 stores that have high average daily sales.
Increase the value of sales per customer for B type stores.
Encourage customers to spend more for per visit at these stores.

Summary and Recommendation 3

Summary 4

Customers tend to spend less at our stores when competitors are close by

Recommendation 4

Differentiate our products and customer service from our competitors

Summary and Recommendation 4

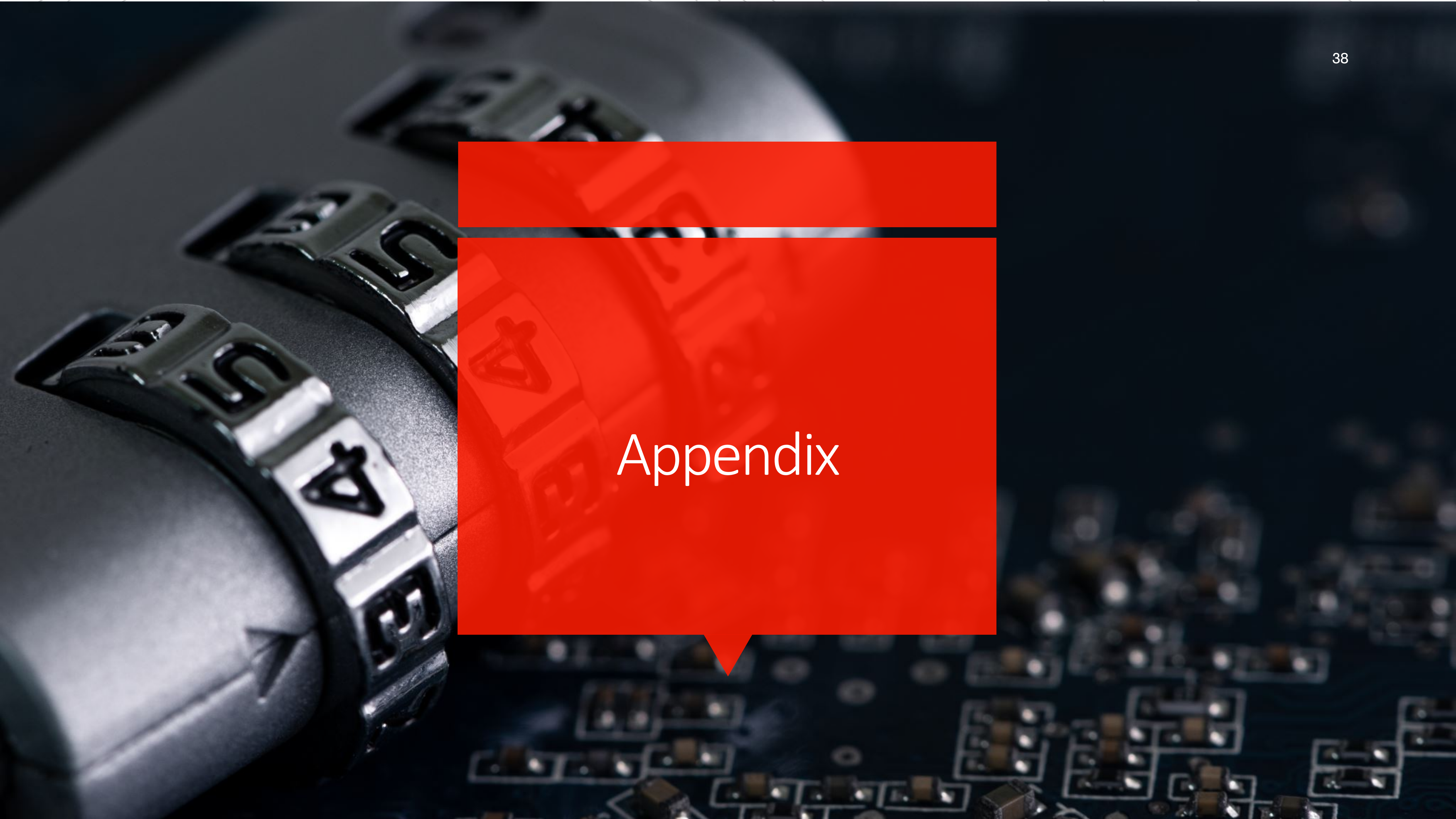
- ❖ Seasonality: Make sure our products are right for the season—whether it's the cold season or allergy season in order to maximize on customer needs and boost their spend. Stock the product before the season hits and have it there when consumers need it

- ❖ Large Display: Use big displays for seasonal items and redesign our display more frequently to make it more organized and interesting and dynamic with each season or holiday so customers will want to come into the store more often.

- ❖ Know Our Customer: Conduct customer surveys to find out:

1) reasons of why customers would choose competitors over us;
2) what are our customers' destination purchases vs. impulse purchases, then invest in displays of the former, make sure it's easy to find and easily called out.

Strategies for Recommendation 3 & 4



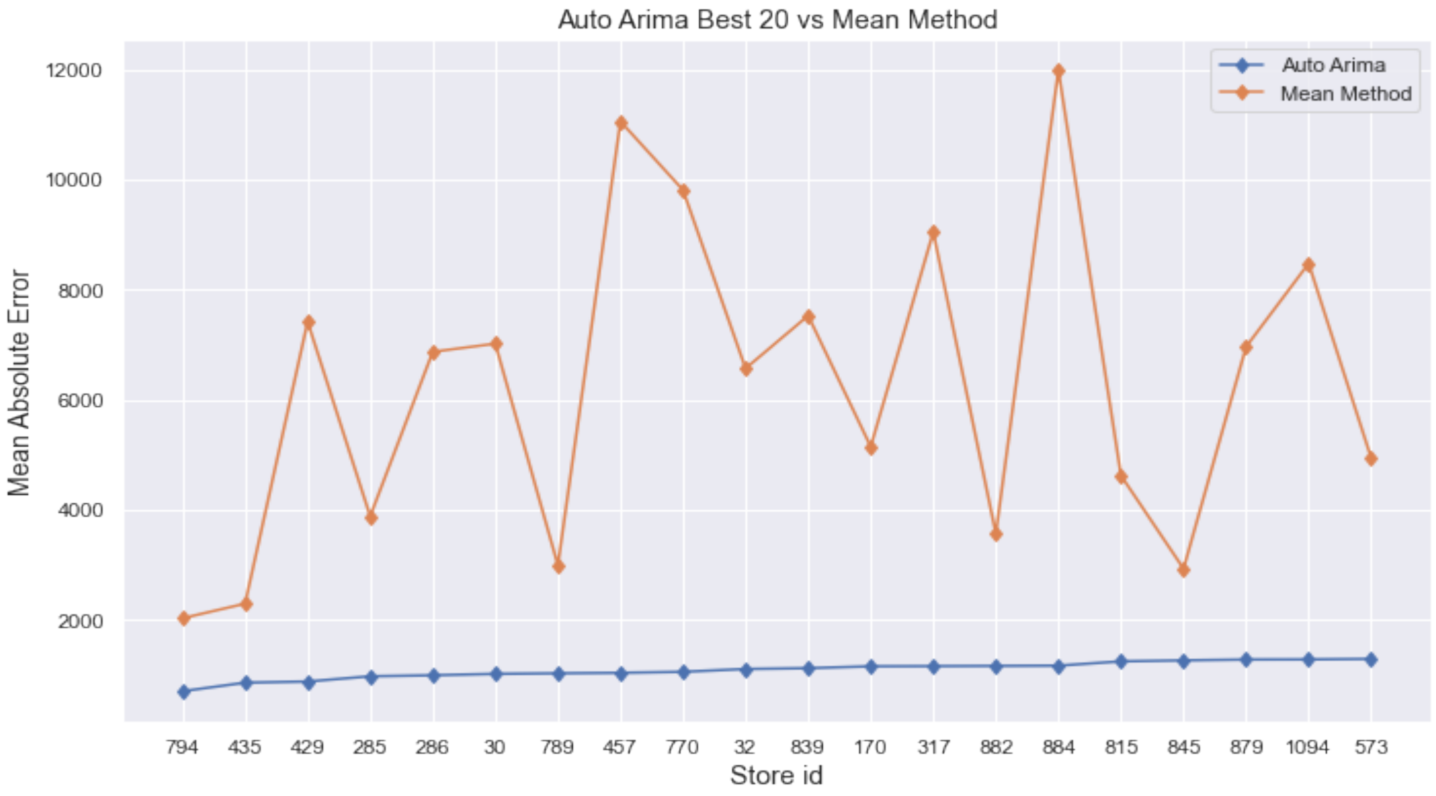
Appendix

- Mean Method: $\hat{y}_{T+h|T} = \bar{y} = (y_1 + \cdots + y_T)/T$.
- Autoregressive models of order p: $y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$
- Moving average model of order q: $y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} \cdots + \theta_q \varepsilon_{t-q}$,
- Non-seasonal ARIMA model: $y_t{}' = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t$
- Mean absolute error: $MAE = \frac{1}{n} \sum_{t=1}^{n} |A_t - F_t|$
- Weighted average percentage error: $WAPE = \frac{\sum_{t=1}^{n} |A_t - F_t|}{\sum_{t=1}^{n} |A_t|}$

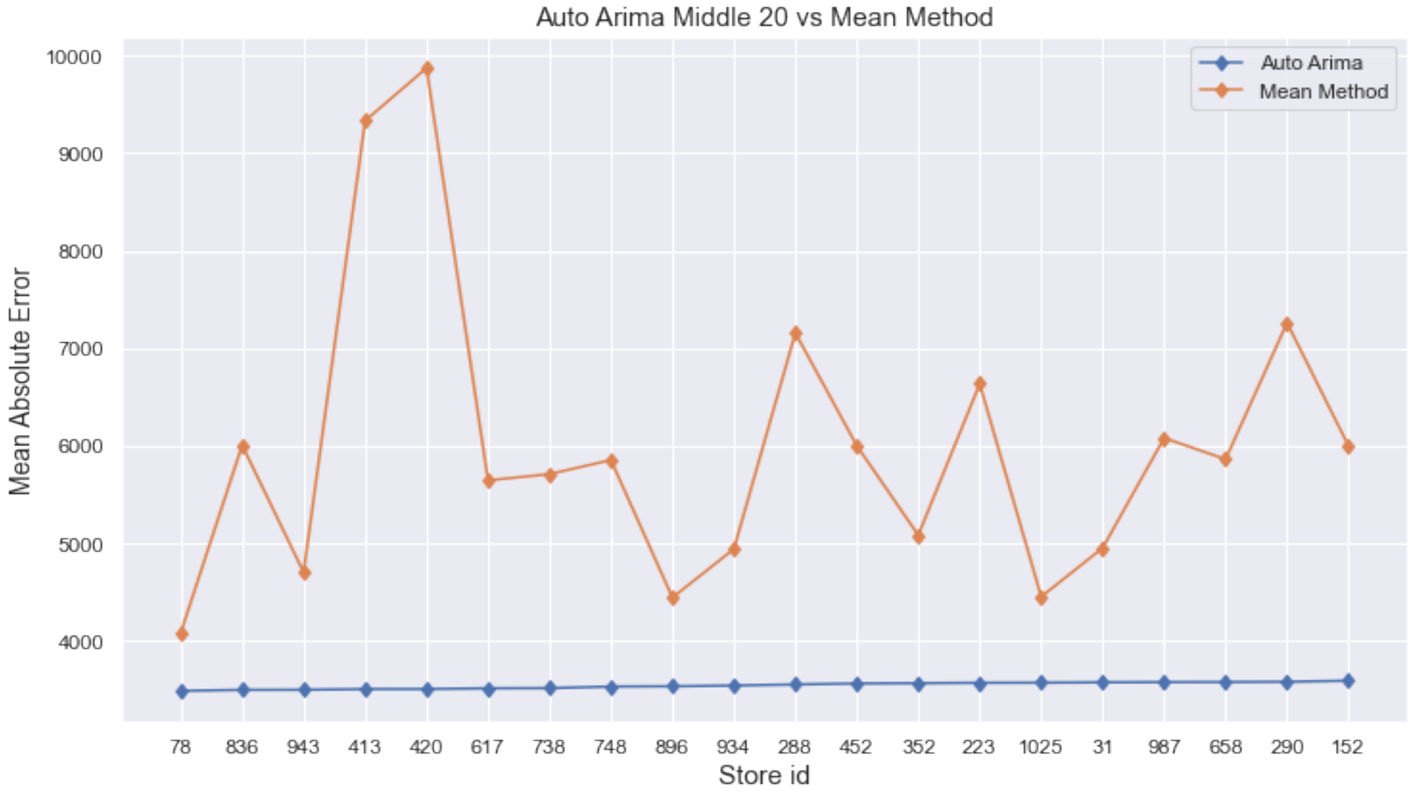
Equations

```
arima_model =  auto_arima(train,start_p=0, d=0, start_q=0,
                          max_p=5, max_d=5, max_q=5, start_P=0,
                          D=1, start_Q=0, max_P=5, max_D=5,
                          max_Q=5, m=2, seasonal=True,
                          error_action='warn',trace = False,
                          supress_warnings=True,stepwise = False,
                          random=True,
                          random_state=20, n_fits=50)
```

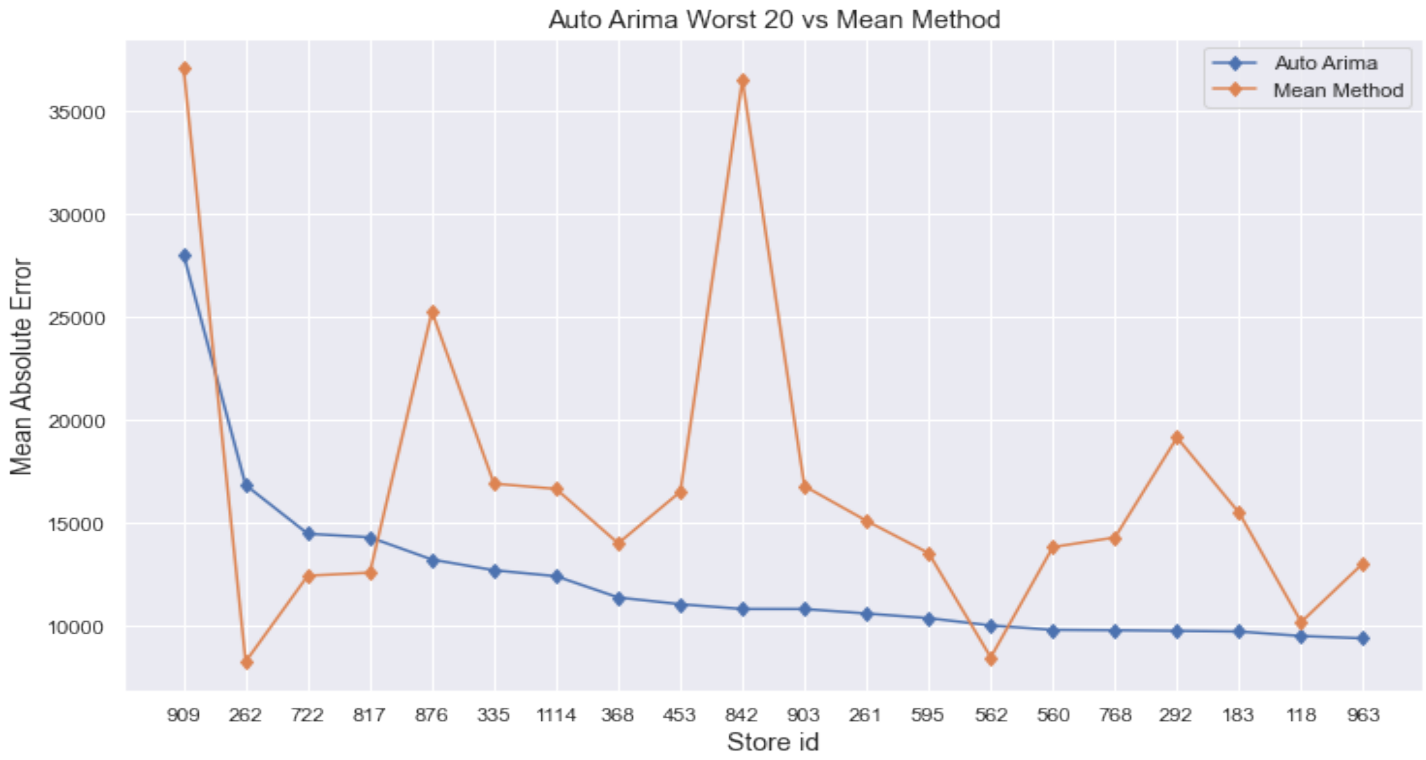
- Parameters and value set up for Auto Arima Model



Model Evaluation - Auto Arima Best 20 vs. Mean Method



Model Evaluation - Auto Arima Middle 20 vs. Mean Method



- Model Evaluation - Auto Arima Worst 20 vs. Mean Method

References

- https://www.kaggle.com/c/rossmann-store-sales/data?select=train.csv (assessed Dec 14, 2020)
- https://www.kaggle.com/c/rossmann-store-sales/data?select=store.csv (assessed Dec 14, 2020)
- https://towardsdatascience.com/time-series-forecasting-using-auto-arima-in-python-bb83e49210cd , *Sushmitha Pulagam* (assessed Dec 20, 2020)
- *Rob J Hyndman, George Athanasopoulos (2018) Forecasting: Principles and Practice,* 2nd edition, OTexts: Melbourne, Australia.
- https://www.linkedin.com/pulse/reading-acf-pacf-plots-missing-manual-cheatsheet-saqib-ali/ , *Saqib Ali* (assessed Dec 28, 2020)
- https://towardsdatascience.com/introduction-to-aic-akaike-information-criterion-9c9ba1c96ced , *Alexandre Zajic* (assessed Jan 2, 2021)
- https://uhurunetwork.com/business-dashboard/ (assessed Jan 14, 2021)
- https://mashmetrics.com/marketing-reports-dashboards-performance-alerts-pros-cons-of-each/ (assessed Jan 14, 2021)
- https://machinelearningmastery.com/how-to-develop-baseline-forecasts-for-multi-site-multivariate-air-pollution-time-series-forecasting/ ,*Jason Brownlee* (assessed Jan 20, 2021)
- https://machinelearningmastery.com/time-series-forecasting-methods-in-python-cheat-sheet/ , *Jason Brownlee* (assessed Jan 20, 2021)
- https://machinelearningmastery.com/remove-trends-seasonality-difference-transform-python/, *Jason Brownlee* (assessed Jan 22, 2021)
- https://www.baeldung.com/cs/mape-vs-wape-vs-wmape (assessed Jan 22, 2021)
- https://www.thebalancesmb.com/pharmacy-merchandising-how-to-boost-sales-2663853, *Amanda Baltazar* (assessed Jan 22, 2021)
- https://en.wikipedia.org/wiki/Kaggle (assessed Jan 1, 2021)
- Springboard – DSC Capstone Project II Detecting Potential Candidates Who are Looking for a New Job ((Nov 2020) *Yang Liu Kunz*