Capstone Project

Rossmann Sales Prediction

ROSSMANN

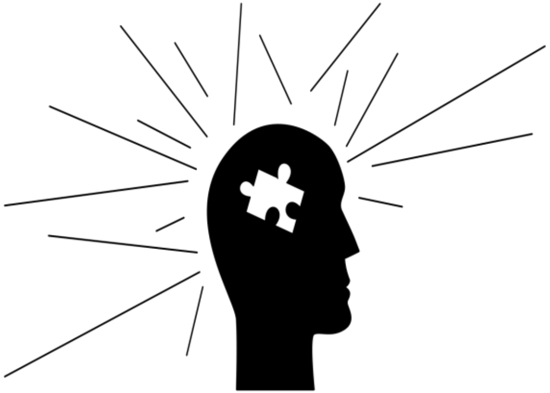
Team Consists of:

1. Vridhi Parmar
2. Pradip Solanki
3. Hrithik Chourasia
4. Ameen Attar

Problem Statement:

- Rossmann store managers are tasked with predicting their daily sales for up to six weeks in advance.

- We were provided with historical sales data for 1,115 Rossmann stores and were asked to predict the same



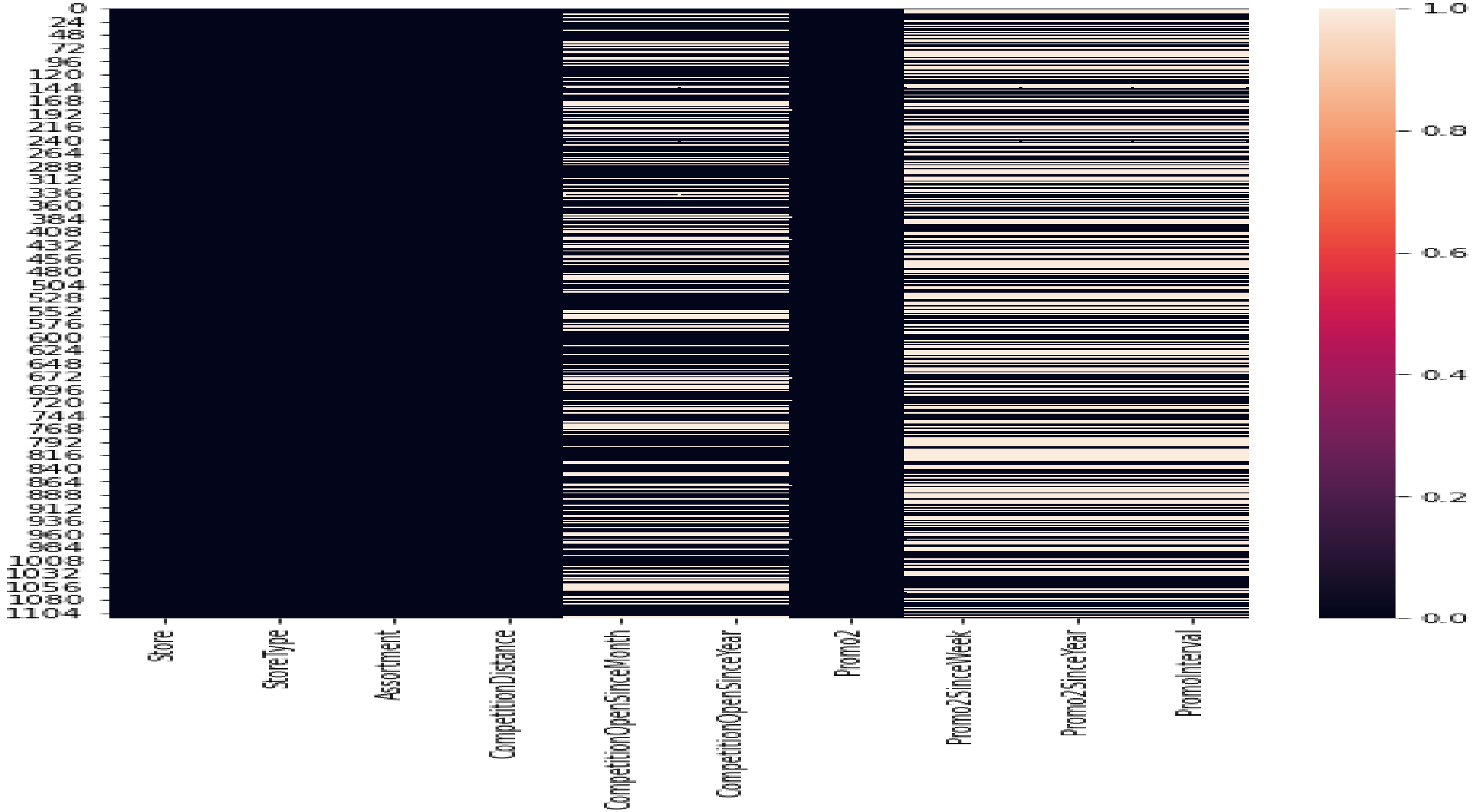
Description of data provided

We are provided with 2 data sets:

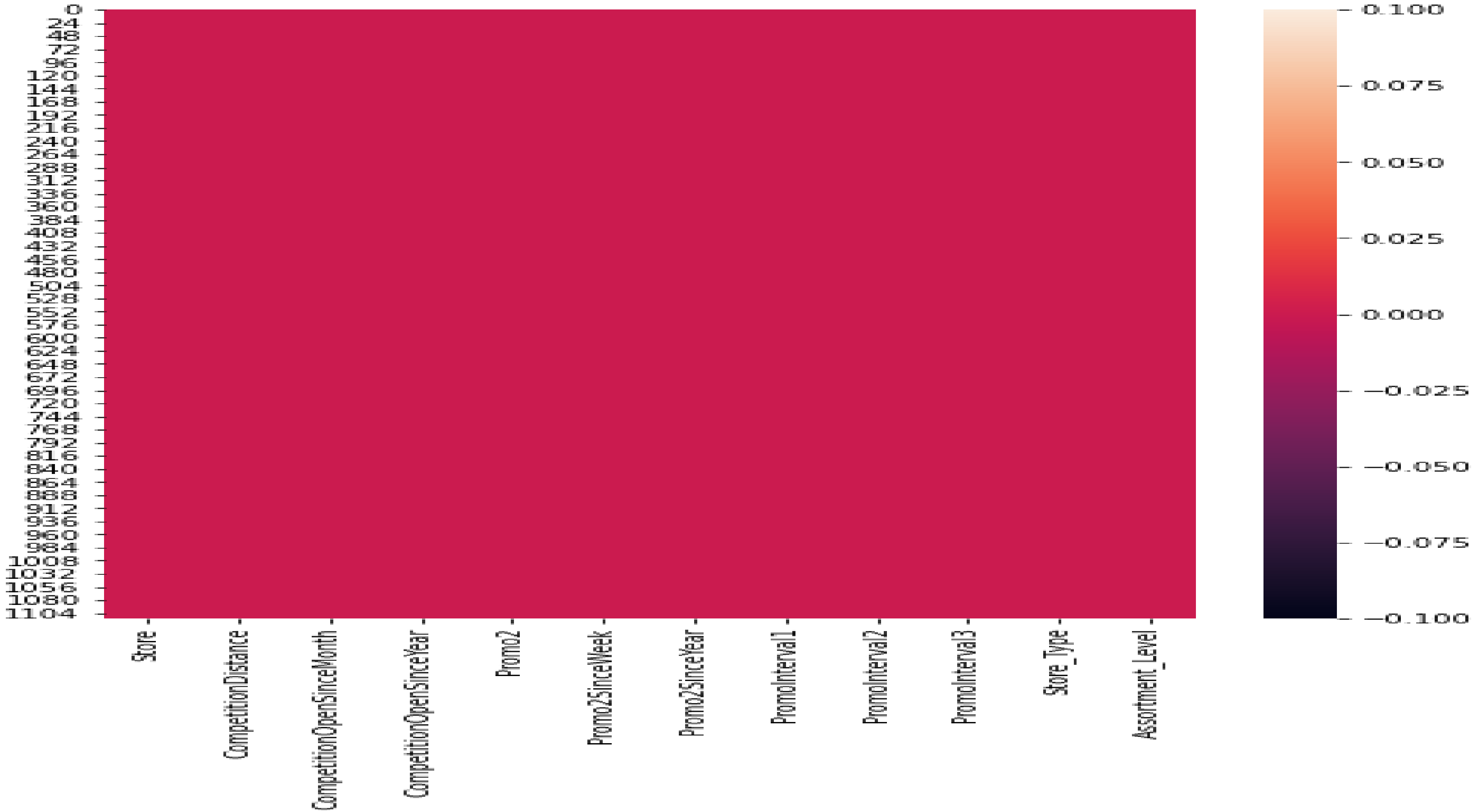
1. Rossmann Stores Data.csv – This dataset includes the historical data including Sales. This dataset contain features like Sales, Customers, Open, StateHoliday, SchoolHoliday.

2. store.csv – This includes supplemental information about the stores. This dataset contain features like Assortment, CompetitionDistance, CompetitionOpenSince[Month/Year], Promo, Promo2, Promo2Since[Year/Week]

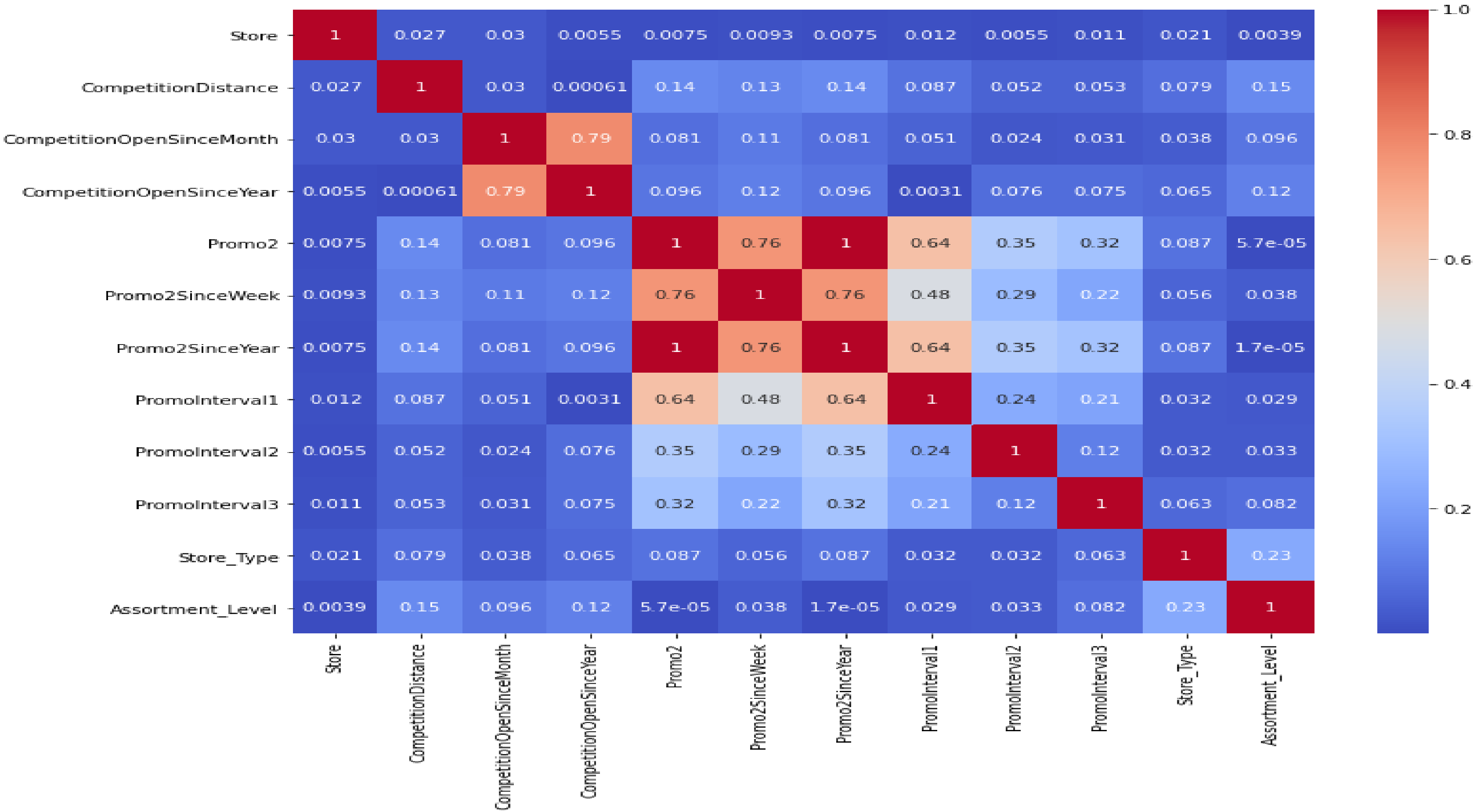
Missing Values in the Datasets



After filling NaN Values in the Datasets

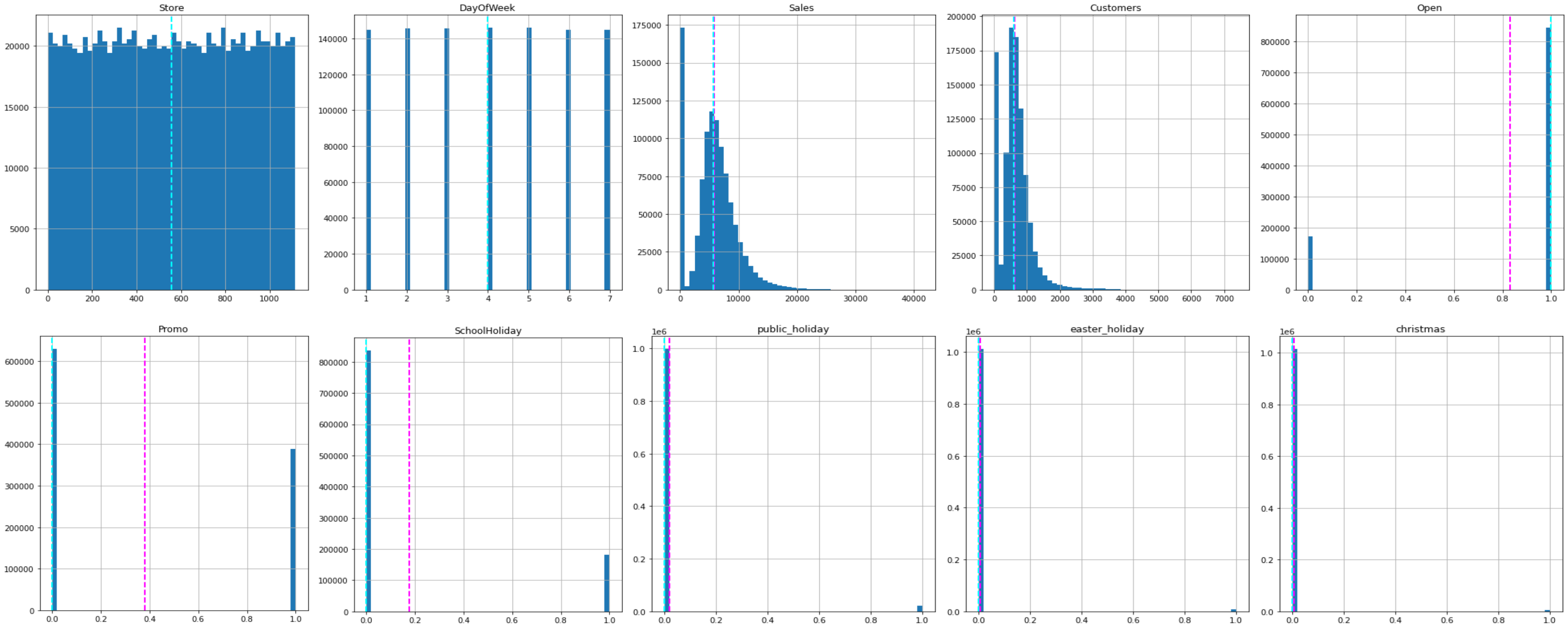


Correlation of variables with each others

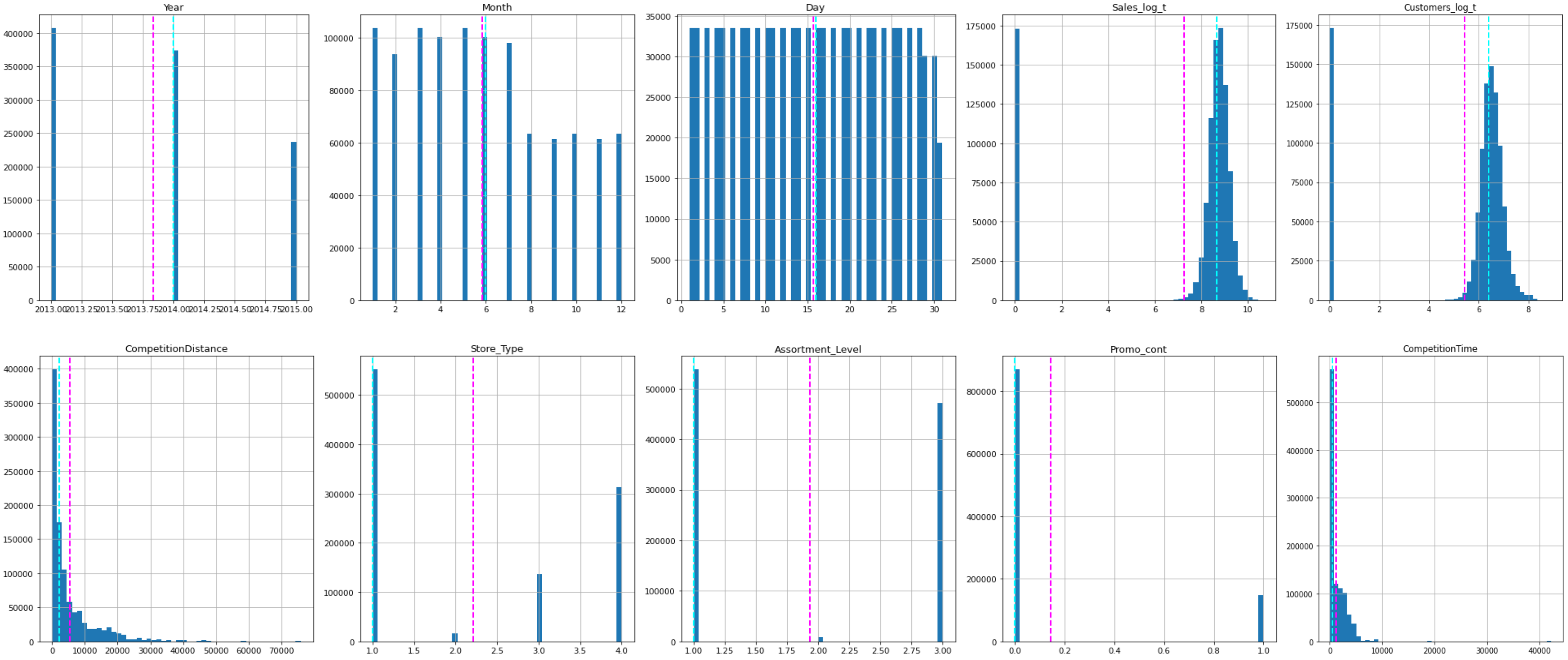


	Store	CompetitionDistance	CompetitionOpenSinceMonth	CompetitionOpenSinceYear	Promo2	Promo2SinceWeek	Promo2SinceYear	PromoInterval1	PromoInterval2	PromoInterval3	Store_Type	Assortment_Level
Store	1	0.027	0.03	0.0055	0.0075	0.0093	0.0075	0.012	0.0055	0.011	0.021	0.0039
CompetitionDistance	0.027	1	0.03	0.00061	0.14	0.13	0.14	0.087	0.052	0.053	0.079	0.15
CompetitionOpenSinceMonth	0.03	0.03	1	0.79	0.081	0.11	0.081	0.051	0.024	0.031	0.038	0.096
CompetitionOpenSinceYear	0.0055	0.00061	0.79	1	0.096	0.12	0.096	0.0031	0.076	0.075	0.065	0.12
Promo2	0.0075	0.14	0.081	0.096	1	0.76	1	0.64	0.35	0.32	0.087	5.7e-05
Promo2SinceWeek	0.0093	0.13	0.11	0.12	0.76	1	0.76	0.48	0.29	0.22	0.056	0.038
Promo2SinceYear	0.0075	0.14	0.081	0.096	1	0.76	1	0.64	0.35	0.32	0.087	1.7e-05
PromoInterval1	0.012	0.087	0.051	0.0031	0.64	0.48	0.64	1	0.24	0.21	0.032	0.029
PromoInterval2	0.0055	0.052	0.024	0.076	0.35	0.29	0.35	0.24	1	0.12	0.032	0.033
PromoInterval3	0.011	0.053	0.031	0.075	0.32	0.22	0.32	0.21	0.12	1	0.063	0.082
Store_Type	0.021	0.079	0.038	0.065	0.087	0.056	0.087	0.032	0.032	0.063	1	0.23
Assortment_Level	0.0039	0.15	0.096	0.12	5.7e-05	0.038	1.7e-05	0.029	0.033	0.082	0.23	1

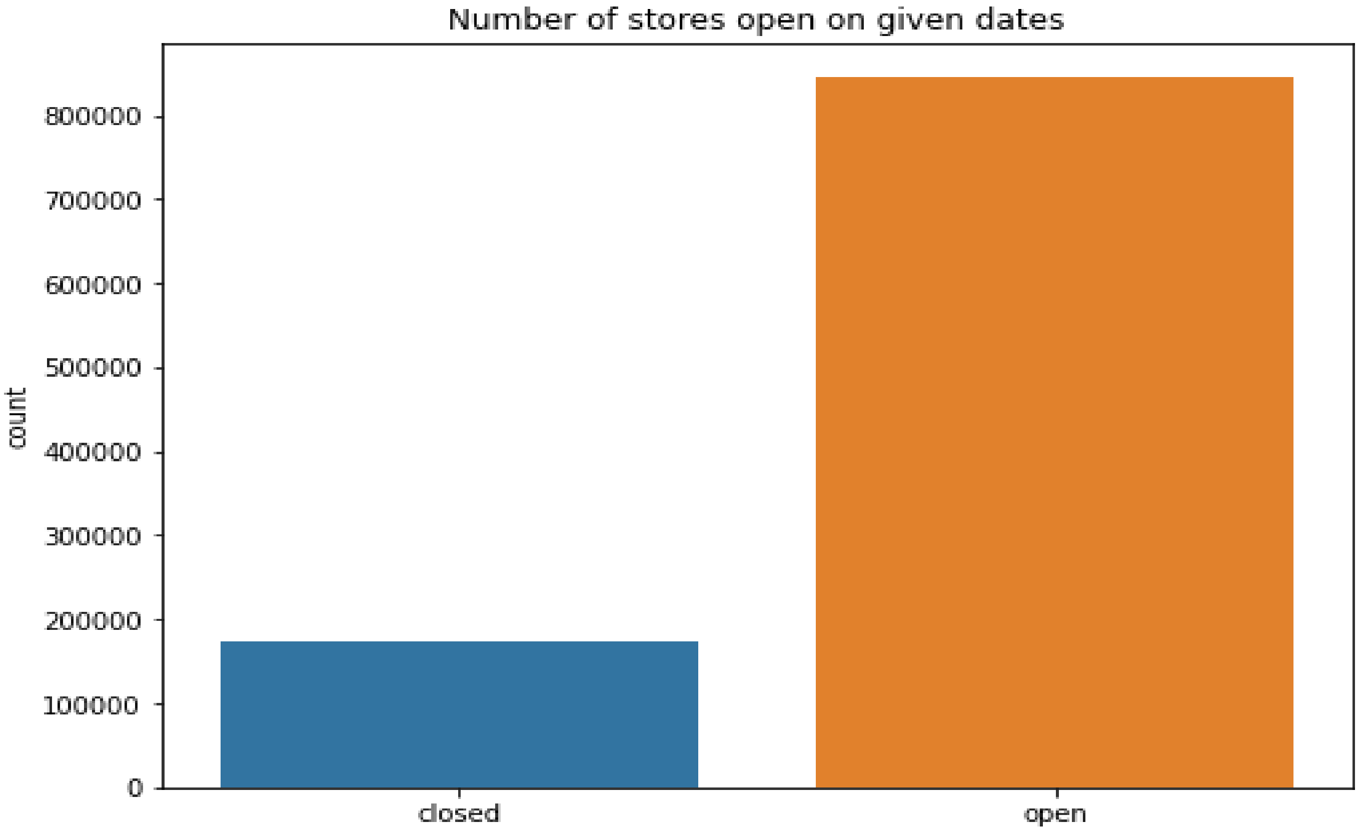
Plot of all the variables:



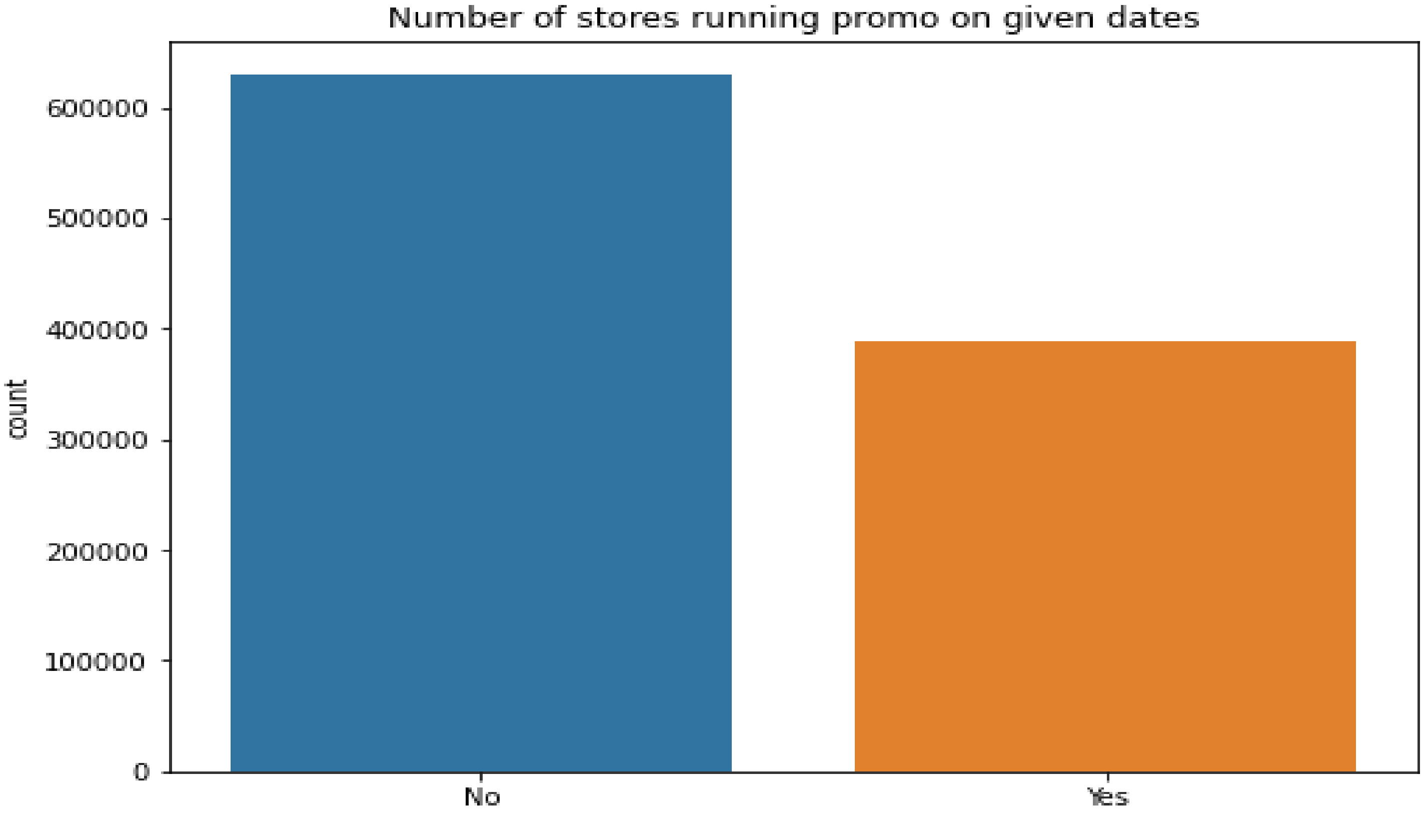
Plot of all the variables:



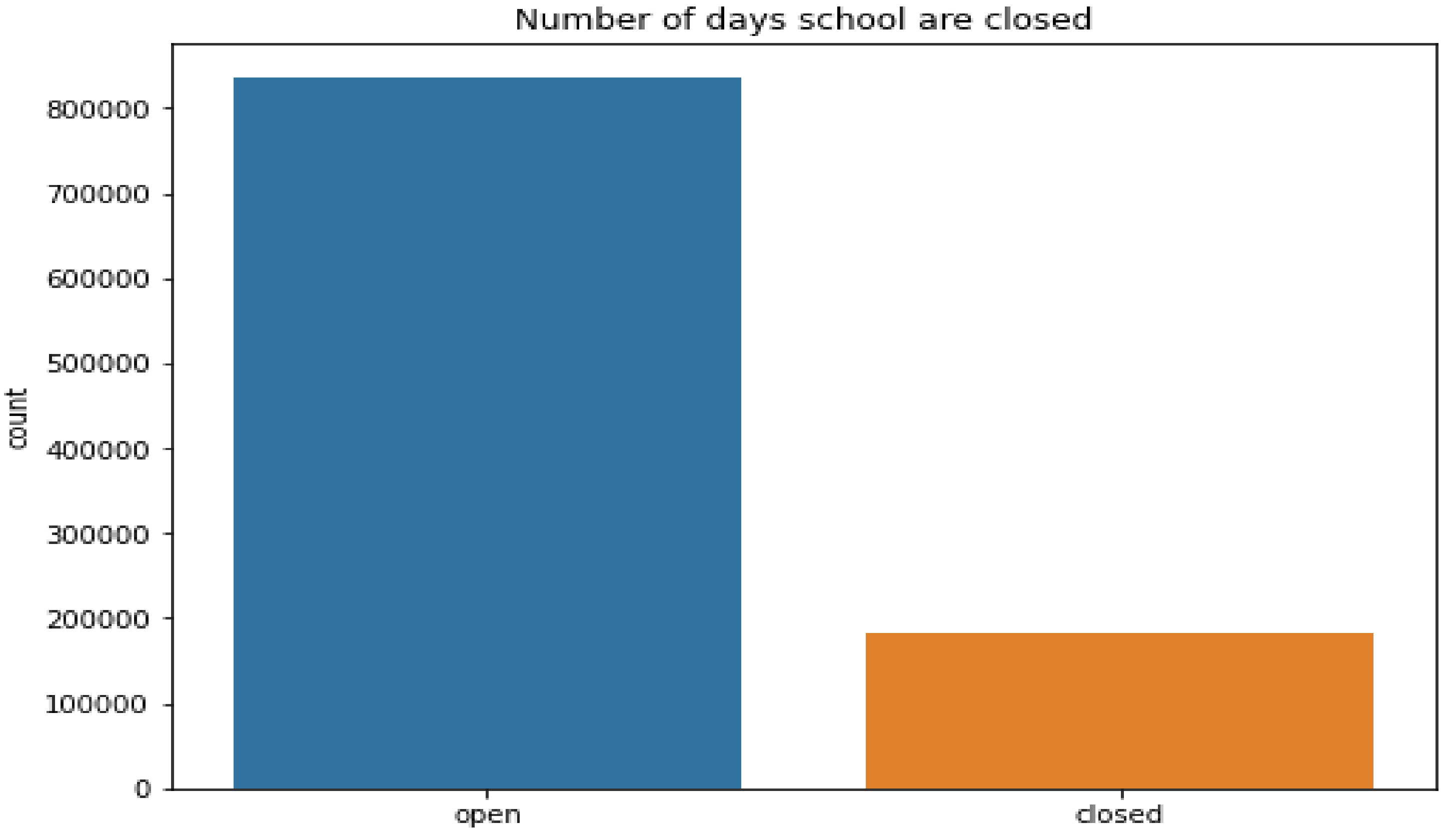
Numbers of stores open on a given Dates



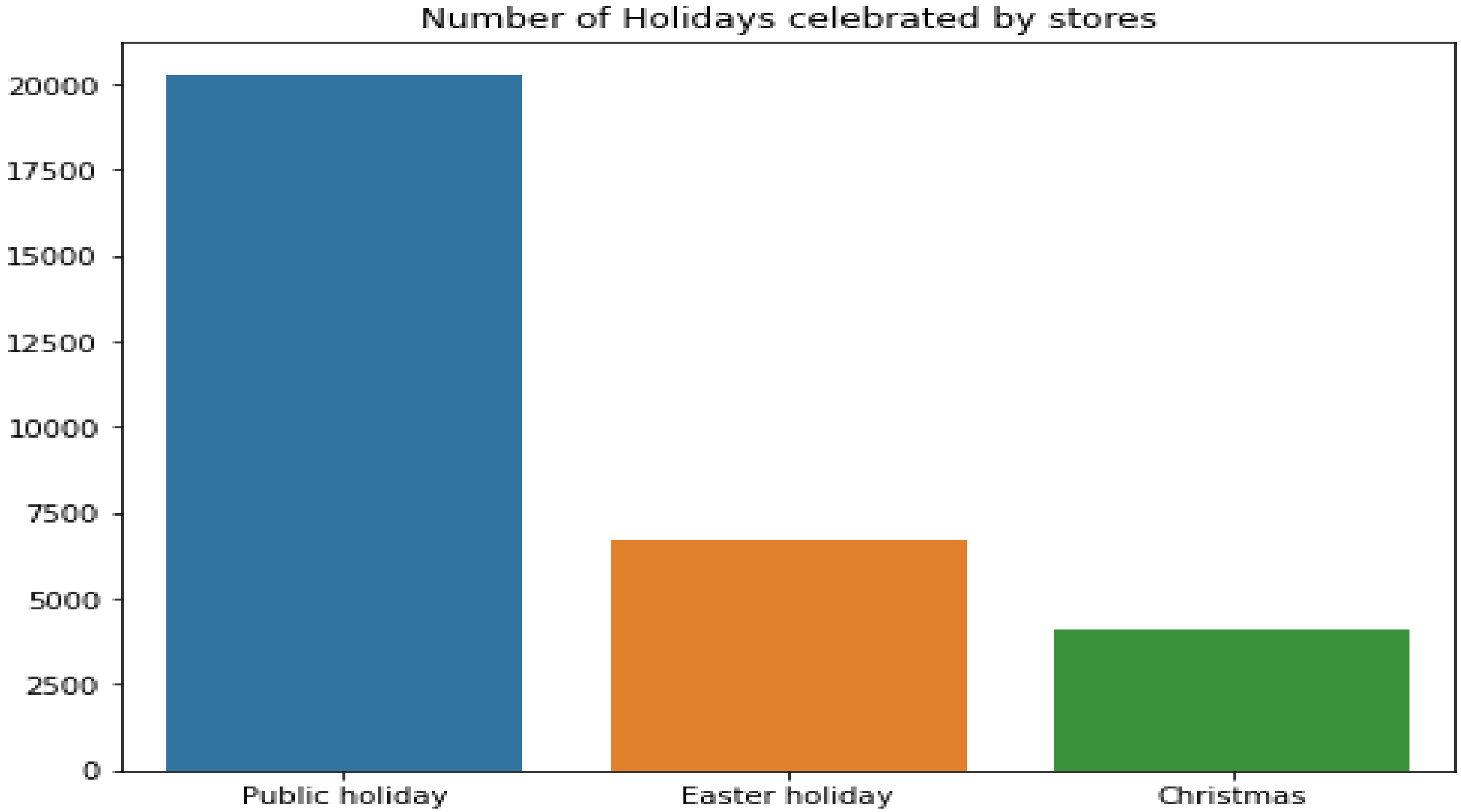
Number of stores running promo on given Dates



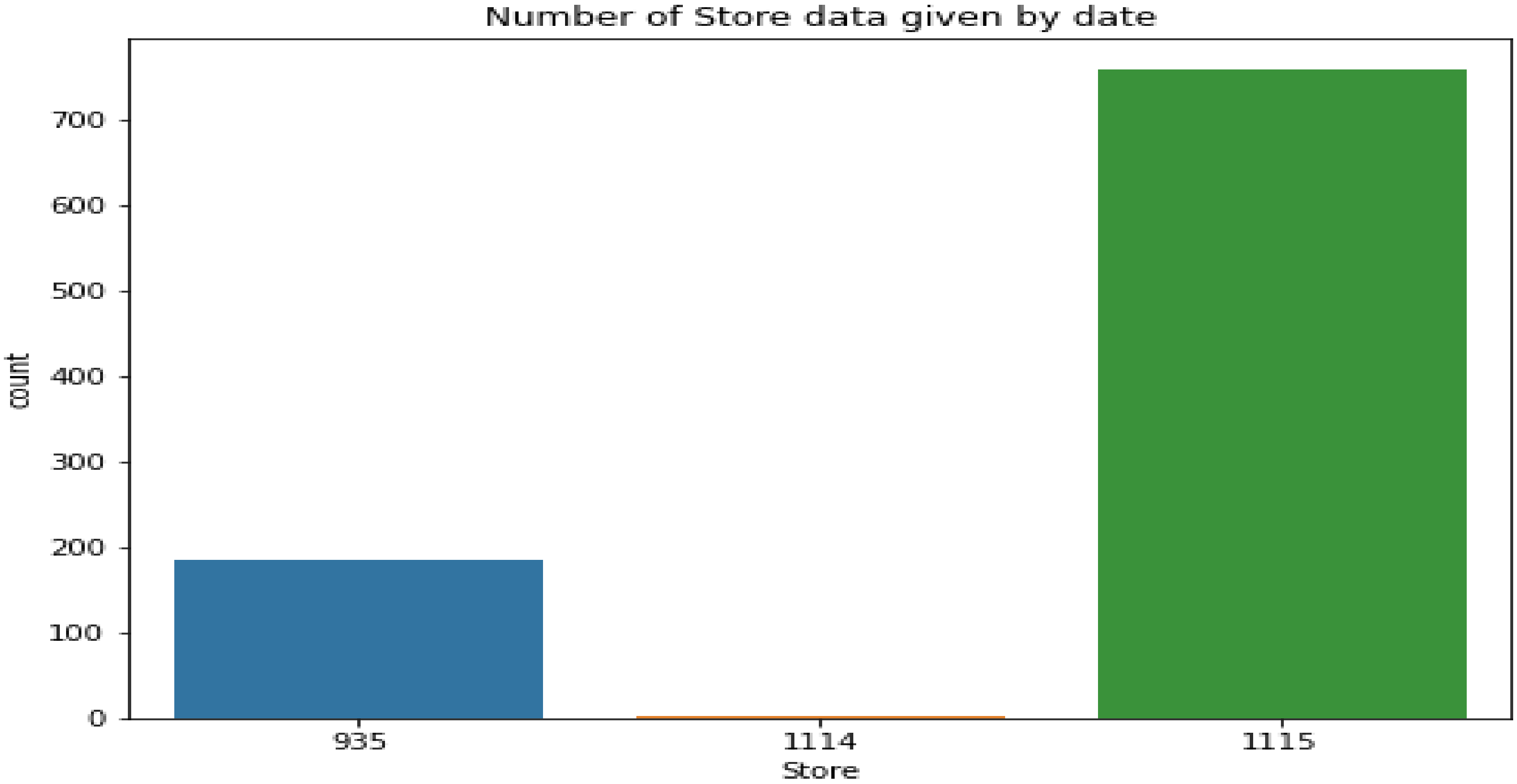
Number of school Holidays



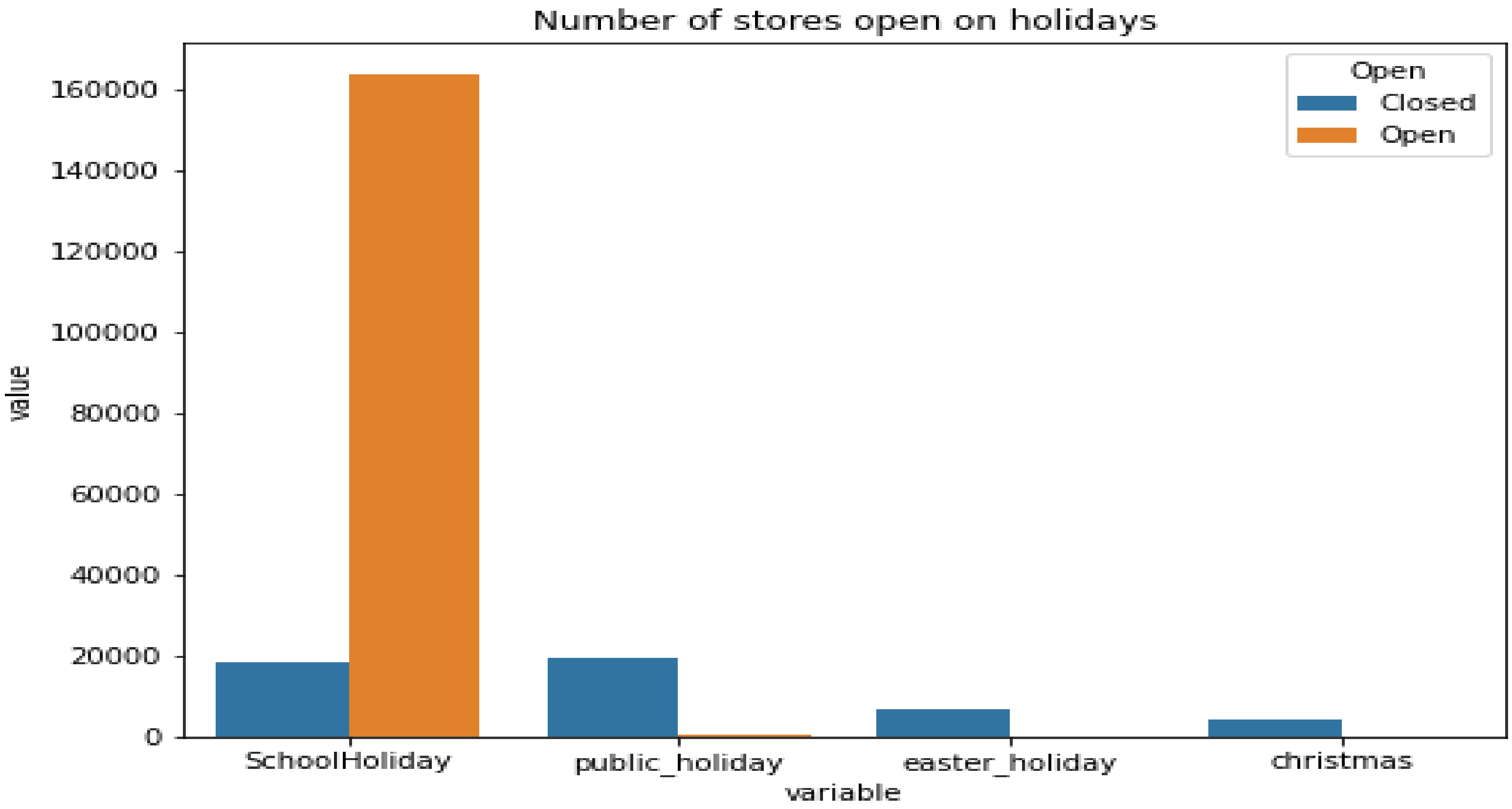
Number of Holidays celebrated by stores



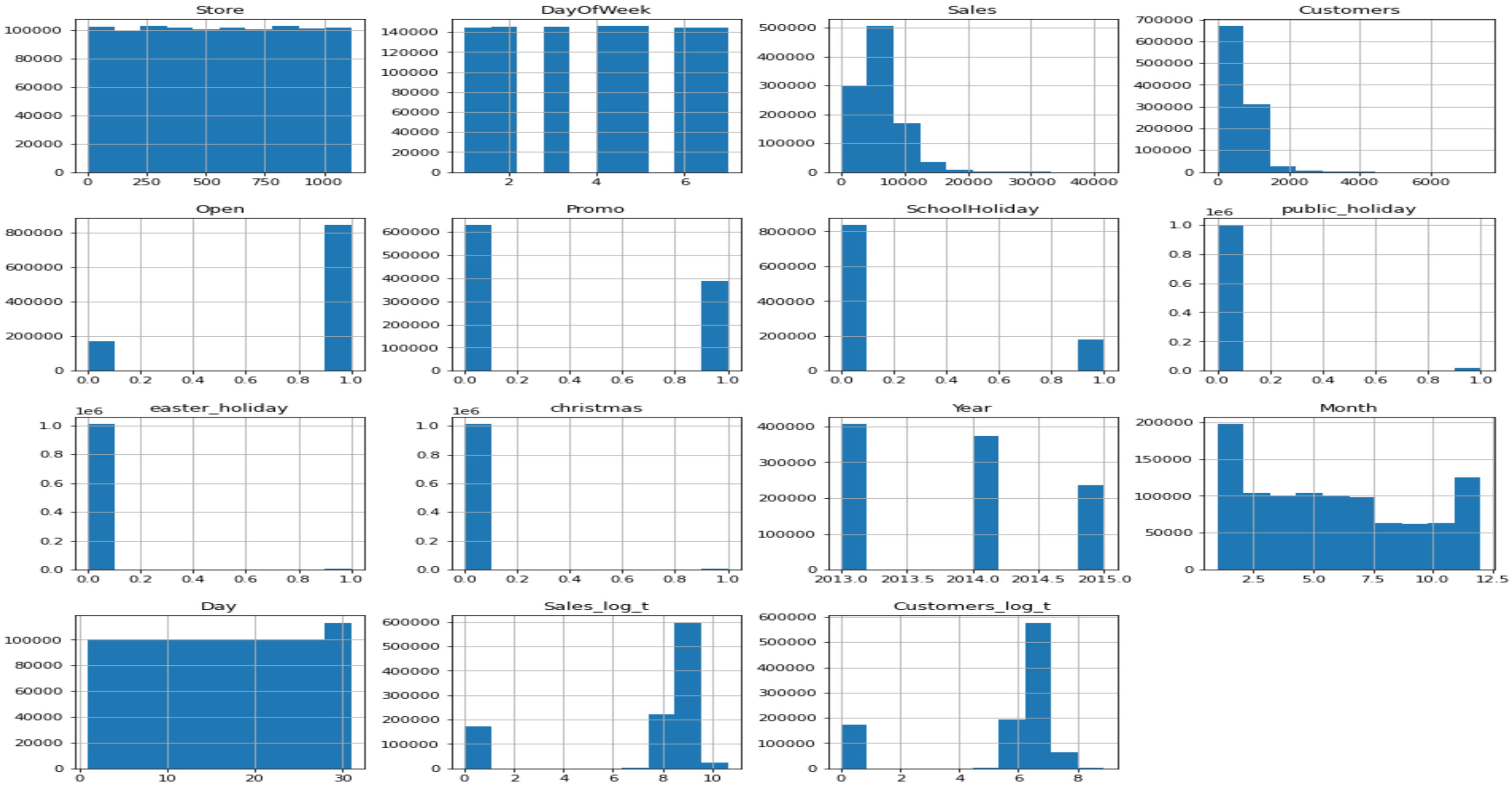
Number of store data given by Date



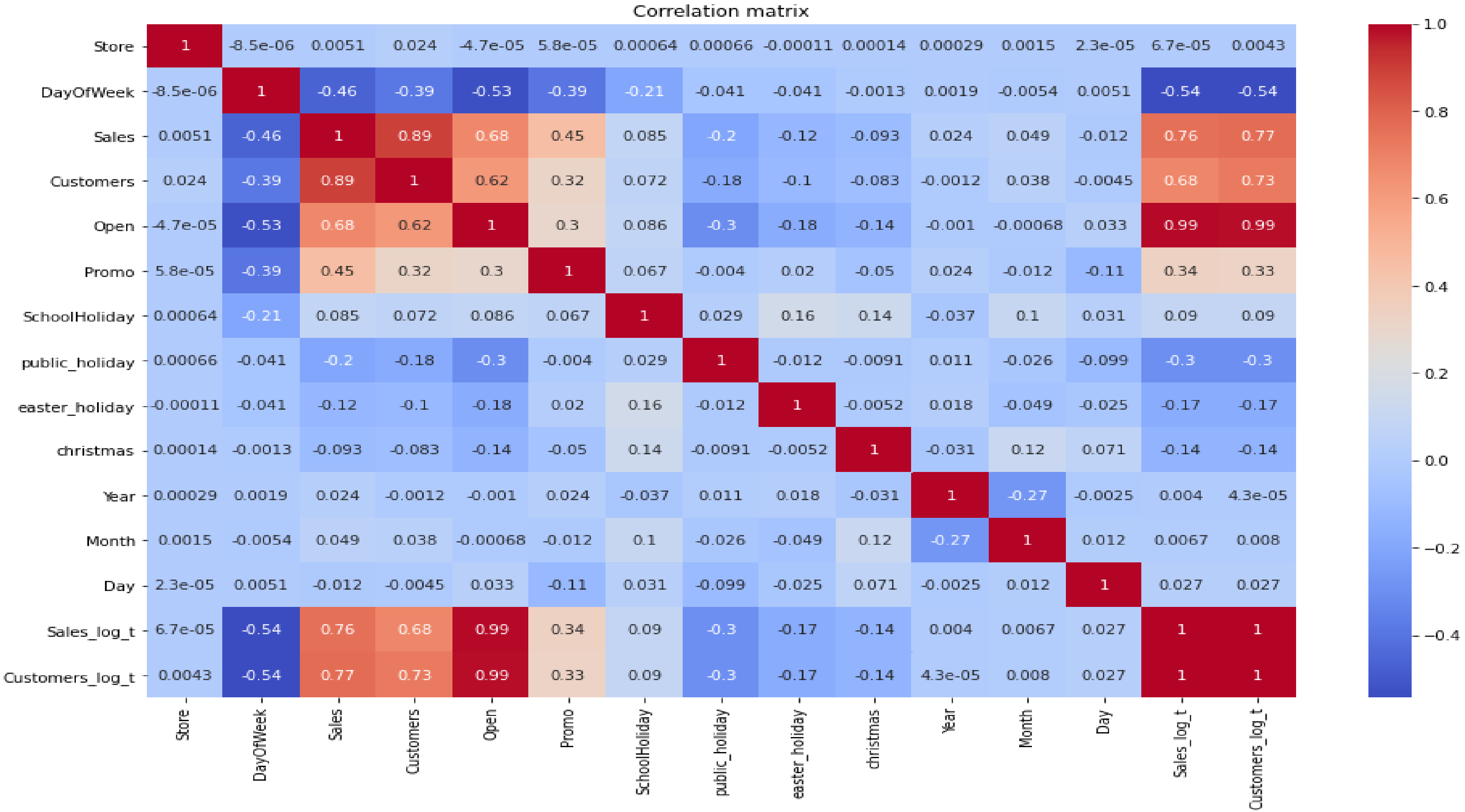
Number of stores open on Holidays



After removing the skewness

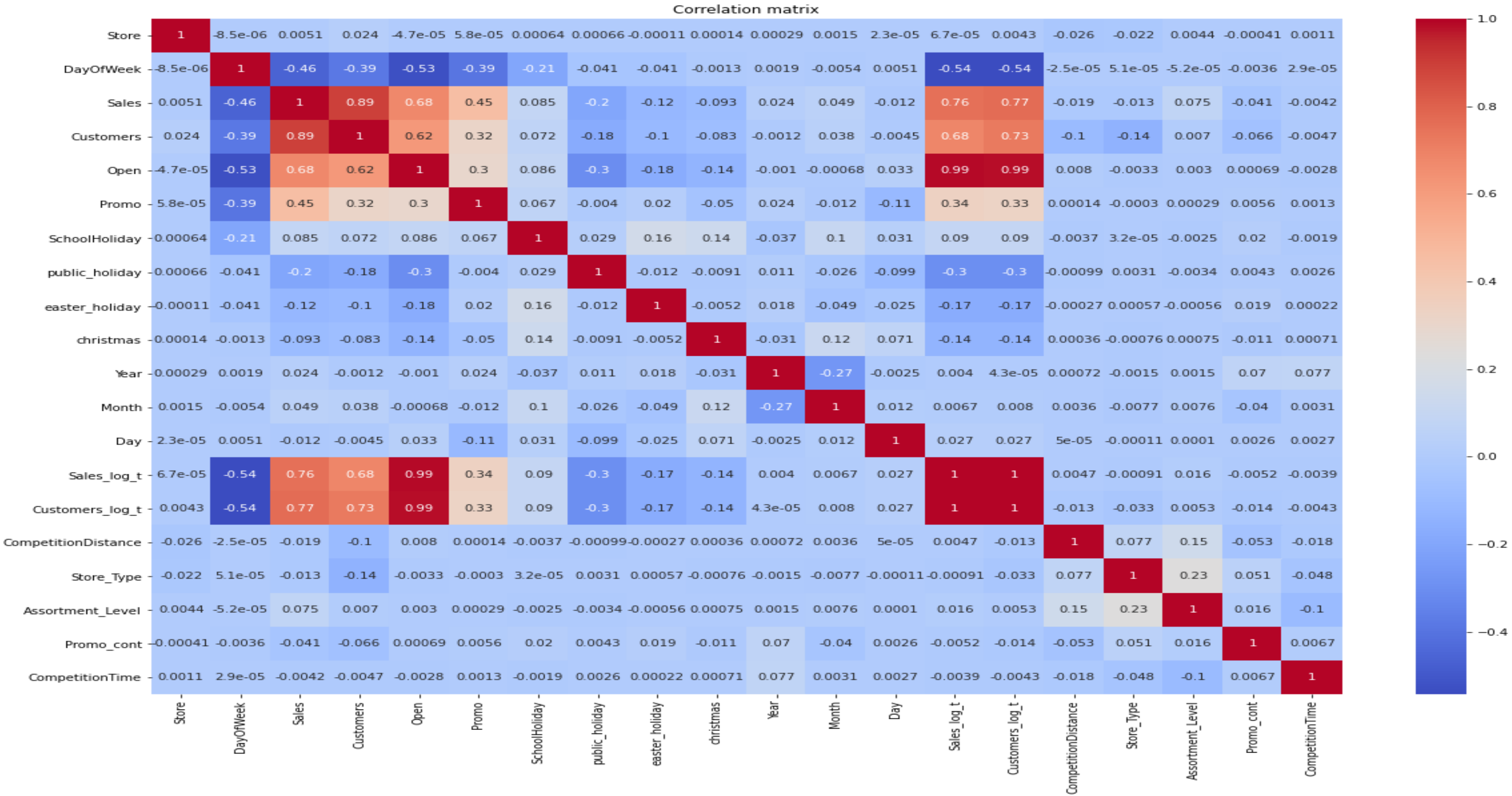


Correlation matrix of Rossmann's Dataset



	Store	DayOfWeek	Sales	Customers	Open	Promo	SchoolHoliday	public_holiday	easter_holiday	christmas	Year	Month	Day	Sales_log_t	Customers_log_t
Store	1	-8.5e-06	0.0051	0.024	-4.7e-05	5.8e-05	0.00064	0.00066	-0.00011	0.00014	0.00029	0.0015	2.3e-05	6.7e-05	0.0043
DayOfWeek	-8.5e-06	1	-0.46	-0.39	-0.53	-0.39	-0.21	-0.041	-0.041	-0.0013	0.0019	-0.0054	0.0051	-0.54	-0.54
Sales	0.0051	-0.46	1	0.89	0.68	0.45	0.085	-0.2	-0.12	-0.093	0.024	0.049	-0.012	0.76	0.77
Customers	0.024	-0.39	0.89	1	0.62	0.32	0.072	-0.18	-0.1	-0.083	-0.0012	0.038	-0.0045	0.68	0.73
Open	-4.7e-05	-0.53	0.68	0.62	1	0.3	0.086	-0.3	-0.18	-0.14	-0.001	-0.00068	0.033	0.99	0.99
Promo	5.8e-05	-0.39	0.45	0.32	0.3	1	0.067	-0.004	0.02	-0.05	0.024	-0.012	-0.11	0.34	0.33
SchoolHoliday	0.00064	-0.21	0.085	0.072	0.086	0.067	1	0.029	0.16	0.14	-0.037	0.1	0.031	0.09	0.09
public_holiday	0.00066	-0.041	-0.2	-0.18	-0.3	-0.004	0.029	1	-0.012	-0.0091	0.011	-0.026	-0.099	-0.3	-0.3
easter_holiday	-0.00011	-0.041	-0.12	-0.1	-0.18	0.02	0.16	-0.012	1	-0.0052	0.018	-0.049	-0.025	-0.17	-0.17
christmas	0.00014	-0.0013	-0.093	-0.083	-0.14	-0.05	0.14	-0.0091	-0.0052	1	-0.031	0.12	0.071	-0.14	-0.14
Year	0.00029	0.0019	0.024	-0.0012	-0.001	0.024	-0.037	0.011	0.018	-0.031	1	-0.27	-0.0025	0.004	4.3e-05
Month	0.0015	-0.0054	0.049	0.038	-0.00068	-0.012	0.1	-0.026	-0.049	0.12	-0.27	1	0.012	0.0067	0.008
Day	2.3e-05	0.0051	-0.012	-0.0045	0.033	-0.11	0.031	-0.099	-0.025	0.071	-0.0025	0.012	1	0.027	0.027
Sales_log_t	6.7e-05	-0.54	0.76	0.68	0.99	0.34	0.09	-0.3	-0.17	-0.14	0.004	0.0067	0.027	1	1
Customers_log_t	0.0043	-0.54	0.77	0.73	0.99	0.33	0.09	-0.3	-0.17	-0.14	4.3e-05	0.008	0.027	1	1

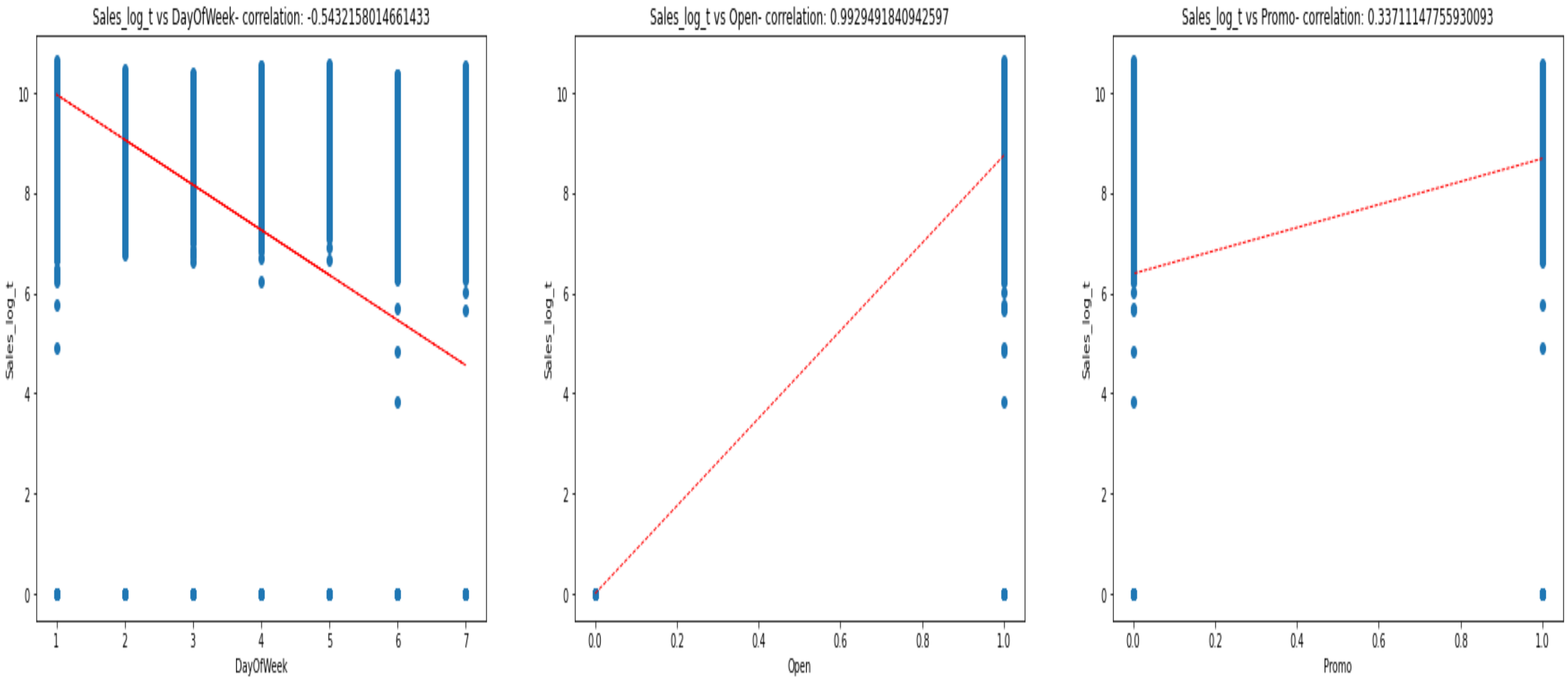
Correlation after merging the two dataset



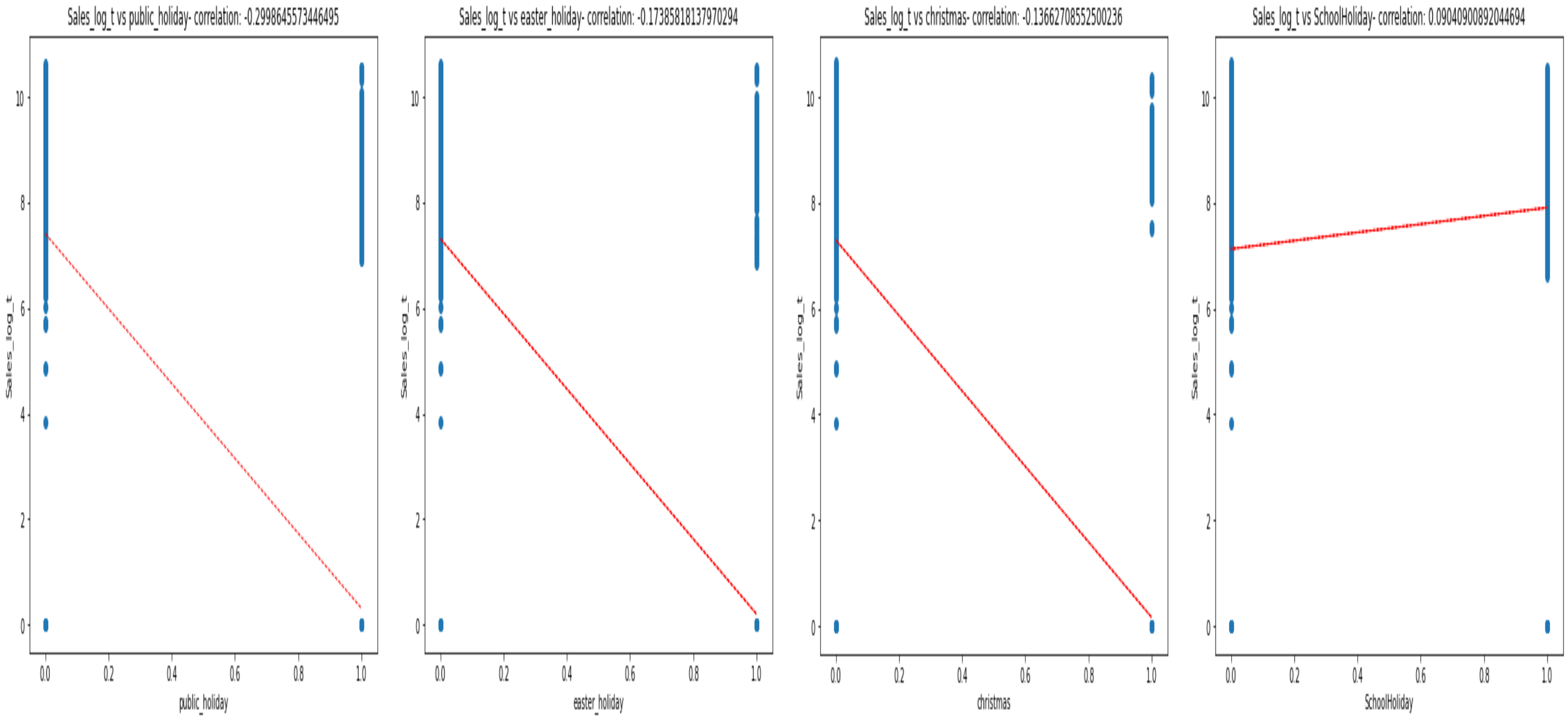
Correlation matrix

	Store	DayOfWeek	Sales	Customers	Open	Promo	SchoolHoliday	public_holiday	easter_holiday	christmas	Year	Month	Day	Sales_log_t	Customers_log_t	CompetitionDistance	Store_Type	Assortment_Level	Promo_cont	CompetitionTime
Store	1	-8.5e-06	0.0051	0.024	-4.7e-05	5.8e-05	0.00064	0.00066	-0.00011	0.00014	0.00029	0.0015	2.3e-05	6.7e-05	0.0043	-0.026	-0.022	0.0044	-0.00041	0.0011
DayOfWeek	-8.5e-06	1	-0.46	-0.39	-0.53	-0.39	-0.21	-0.041	-0.041	-0.0013	0.0019	-0.0054	0.0051	-0.54	-0.54	-2.5e-05	5.1e-05	-5.2e-05	-0.0036	2.9e-05
Sales	0.0051	-0.46	1	0.89	0.68	0.45	0.085	-0.2	-0.12	-0.093	0.024	0.049	-0.012	0.76	0.77	-0.019	-0.013	0.075	-0.041	-0.0042
Customers	0.024	-0.39	0.89	1	0.62	0.32	0.072	-0.18	-0.1	-0.083	-0.0012	0.038	-0.0045	0.68	0.73	-0.1	-0.14	0.007	-0.066	-0.0047
Open	-4.7e-05	-0.53	0.68	0.62	1	0.3	0.086	-0.3	-0.18	-0.14	-0.001	-0.00068	0.033	0.99	0.99	0.008	-0.0033	0.003	0.00069	-0.0028
Promo	5.8e-05	-0.39	0.45	0.32	0.3	1	0.067	-0.004	0.02	-0.05	0.024	-0.012	-0.11	0.34	0.33	0.00014	-0.0003	0.00029	0.0056	0.0013
SchoolHoliday	0.00064	-0.21	0.085	0.072	0.086	0.067	1	0.029	0.16	0.14	-0.037	0.1	0.031	0.09	0.09	-0.0037	3.2e-05	-0.0025	0.02	-0.0019
public_holiday	0.00066	-0.041	-0.2	-0.18	-0.3	-0.004	0.029	1	-0.012	-0.0091	0.011	-0.026	-0.099	-0.3	-0.3	-0.00099	0.0031	-0.0034	0.0043	0.0026
easter_holiday	-0.00011	-0.041	-0.12	-0.1	-0.18	0.02	0.16	-0.012	1	-0.0052	0.018	-0.049	-0.025	-0.17	-0.17	-0.00027	0.00057	-0.00056	0.019	0.00022
christmas	0.00014	-0.0013	-0.093	-0.083	-0.14	-0.05	0.14	-0.0091	-0.0052	1	-0.031	0.12	0.071	-0.14	-0.14	0.00036	-0.00076	0.00075	-0.011	0.00071
Year	0.00029	0.0019	0.024	-0.0012	-0.001	0.024	-0.037	0.011	0.018	-0.031	1	-0.27	-0.0025	0.004	4.3e-05	0.00072	-0.0015	0.0015	0.07	0.077
Month	0.0015	-0.0054	0.049	0.038	-0.00068	-0.012	0.1	-0.026	-0.049	0.12	-0.27	1	0.012	0.0067	0.008	0.0036	-0.0077	0.0076	-0.04	0.0031
Day	2.3e-05	0.0051	-0.012	-0.0045	0.033	-0.11	0.031	-0.099	-0.025	0.071	-0.0025	0.012	1	0.027	0.027	5e-05	-0.00011	0.0001	0.0026	0.0027
Sales_log_t	6.7e-05	-0.54	0.76	0.68	0.99	0.34	0.09	-0.3	-0.17	-0.14	0.004	0.0067	0.027	1	1	0.0047	-0.00091	0.016	-0.0052	-0.0039
Customers_log_t	0.0043	-0.54	0.77	0.73	0.99	0.33	0.09	-0.3	-0.17	-0.14	4.3e-05	0.008	0.027	1	1	-0.013	-0.033	0.0053	-0.014	-0.0043
CompetitionDistance	-0.026	-2.5e-05	-0.019	-0.1	0.008	0.00014	-0.0037	-0.00099	-0.00027	0.00036	0.00072	0.0036	5e-05	0.0047	-0.013	1	0.077	0.15	-0.053	-0.018
Store_Type	-0.022	5.1e-05	-0.013	-0.14	-0.0033	-0.0003	3.2e-05	0.0031	0.00057	-0.00076	-0.0015	-0.0077	-0.00011	-0.00091	-0.033	0.077	1	0.23	0.051	-0.048
Assortment_Level	0.0044	-5.2e-05	0.075	0.007	0.003	0.00029	-0.0025	-0.0034	-0.00056	0.00075	0.0015	0.0076	0.0001	0.016	0.0053	0.15	0.23	1	0.016	-0.1
Promo_cont	-0.00041	-0.0036	-0.041	-0.066	0.00069	0.0056	0.02	0.0043	0.019	-0.011	0.07	-0.04	0.0026	-0.0052	-0.014	-0.053	0.051	0.016	1	0.0067
CompetitionTime	0.0011	2.9e-05	-0.0042	-0.0047	-0.0028	0.0013	-0.0019	0.0026	0.00022	0.00071	0.077	0.0031	0.0027	-0.0039	-0.0043	-0.018	-0.048	-0.1	0.0067	1

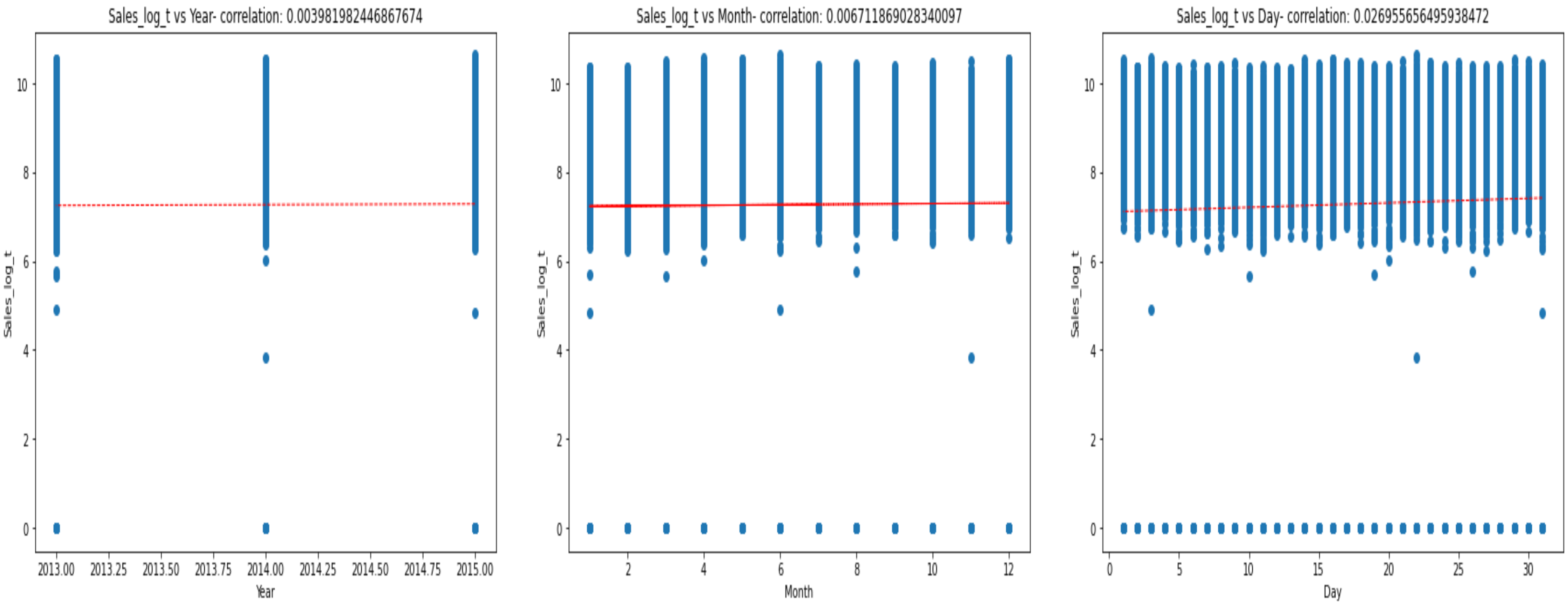
The scatter plot for sales_log_t with other Variables:



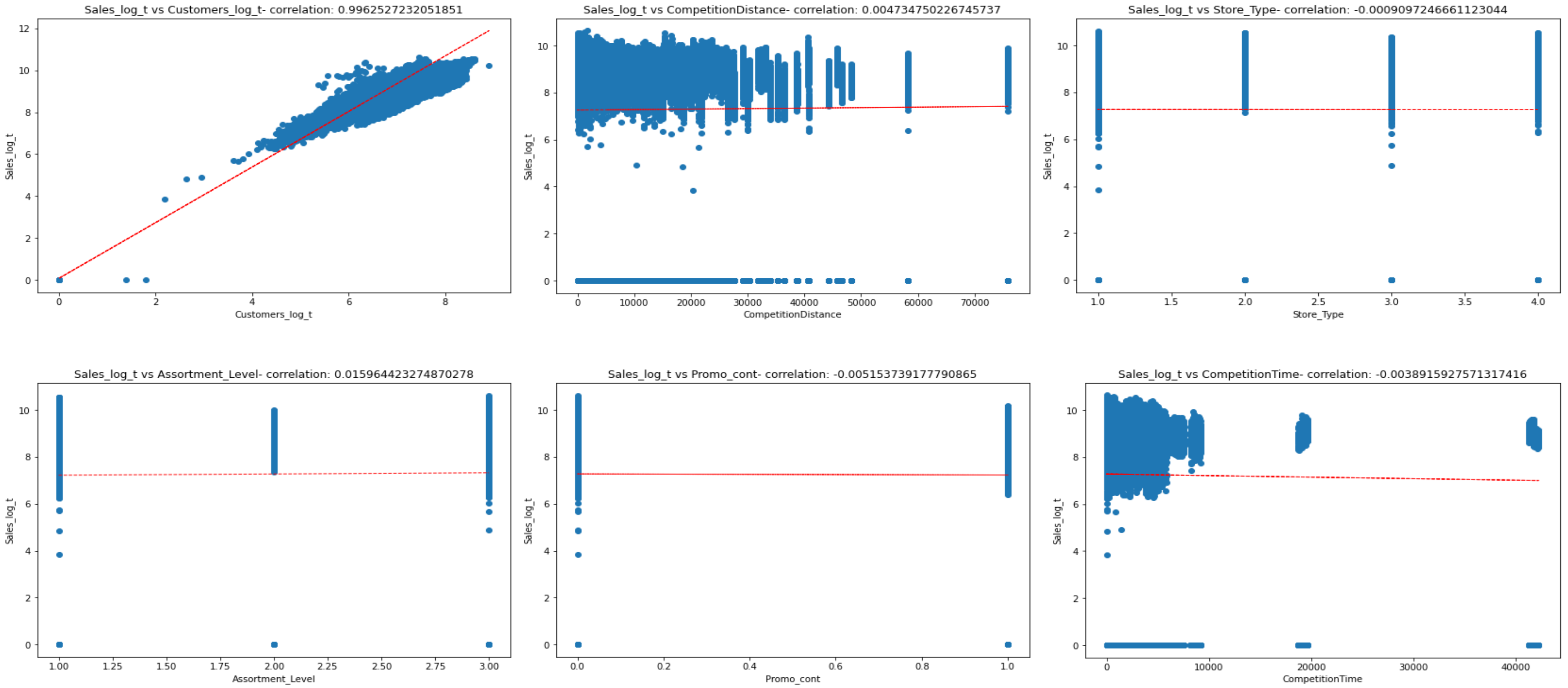
The scatter plot for sales_log_t with other Variables:



The scatter plot for sales_log_t with other Variables:



The scatter plot for sales_log_t with other Variables:



Sales_log_t vs Customers_log_t- correlation: 0.9962527232051851

Sales_log_t vs CompetitionDistance- correlation: 0.004734750226745737

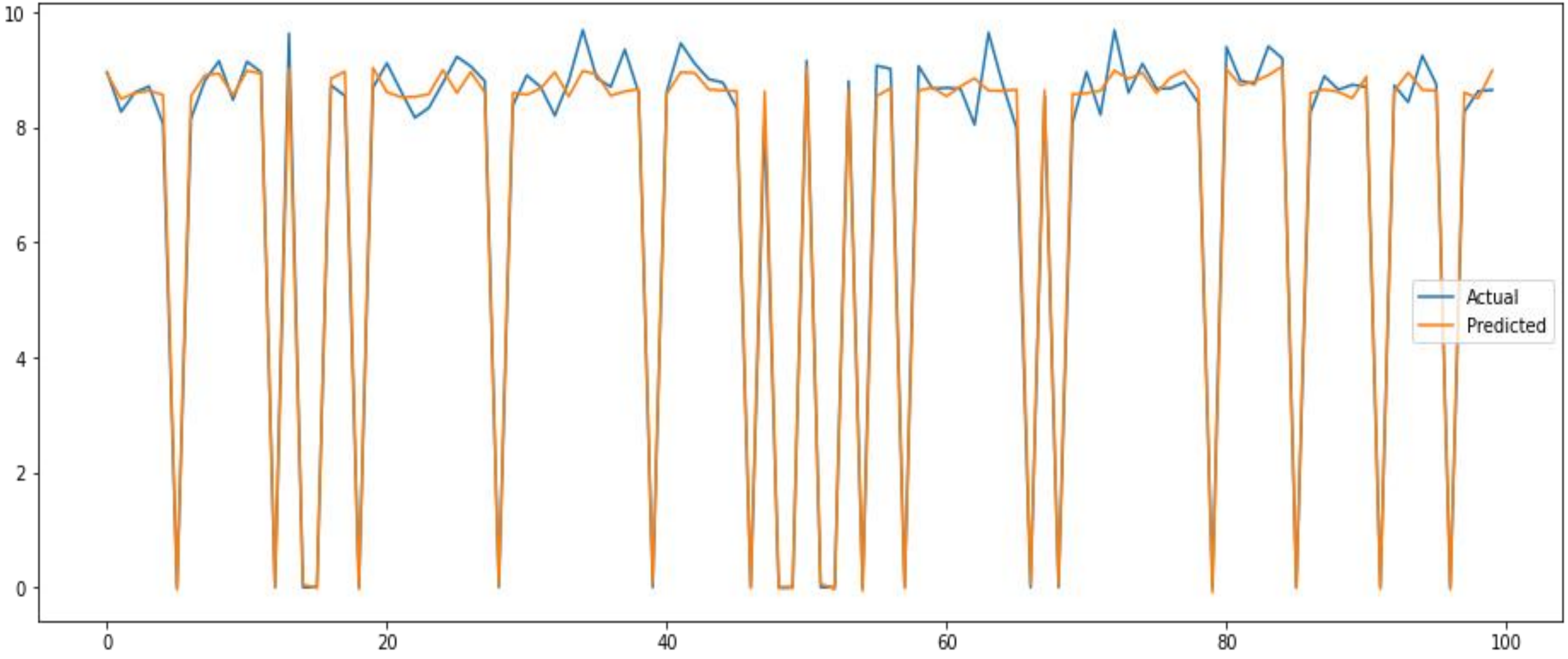
Sales_log_t vs Store_Type- correlation: -0.0009097246661123044

Sales_log_t vs Assortment_Level- correlation: 0.015964423274870278

Sales_log_t vs Promo_cont- correlation: -0.005153739177790865

Sales_log_t vs CompetitionTime- correlation: -0.0038915927571317416

Plot for first 100 observations between actual and predicted after building the linear regression model:

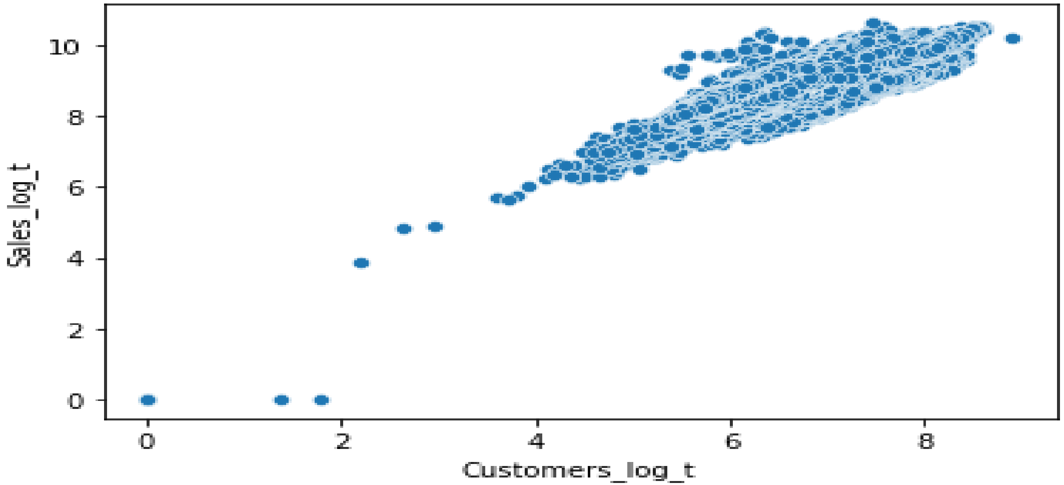


We are using linear regression model so, we need to check 4 basic assumptions of linear regression.

1. There should to be linear relationship between independent and dependent variables.
2. The sum of residuals/error should be near to 0.
3. There should not be multicollinearity.
4. There should not be heteroscedasticity.

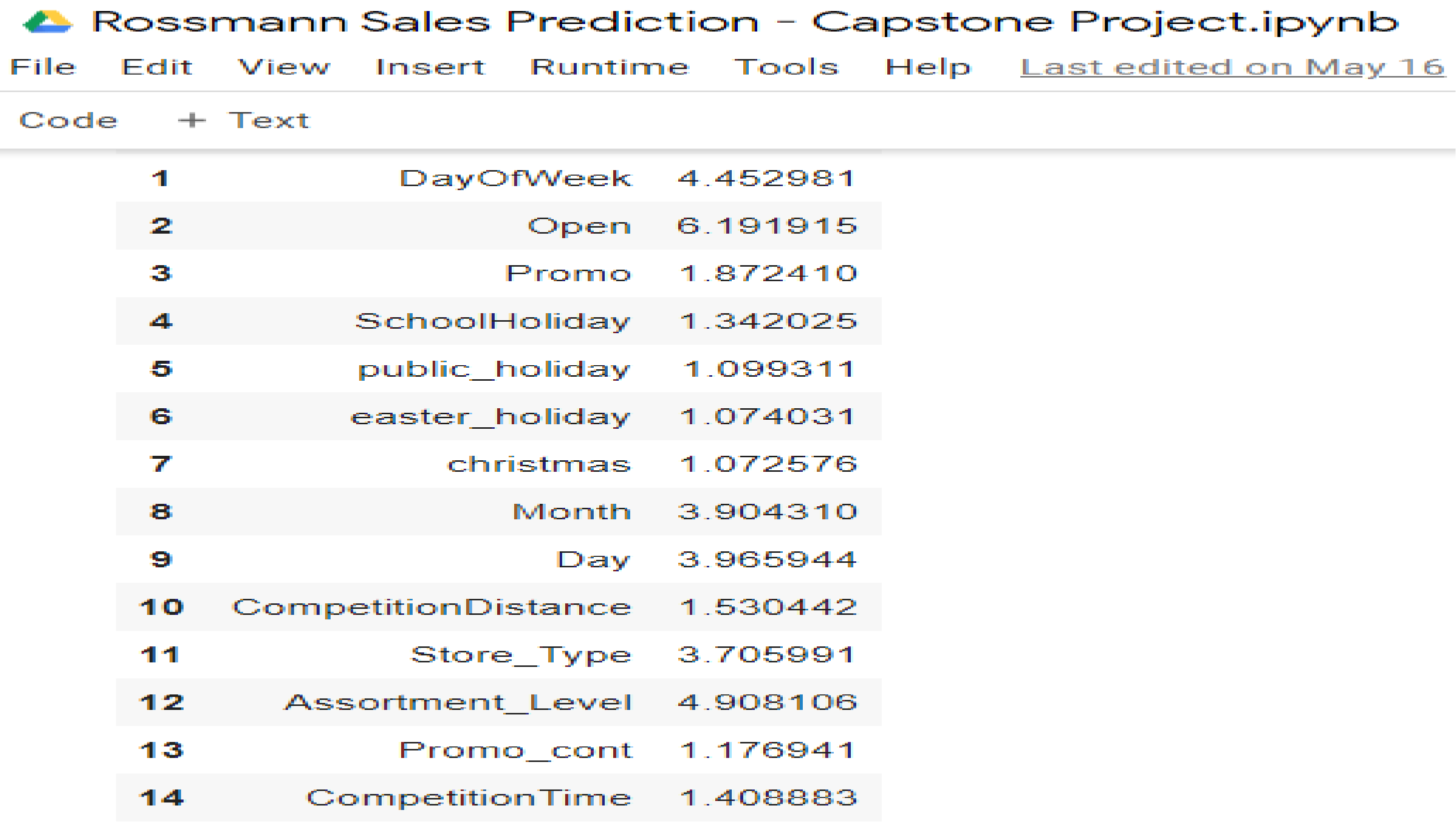
Assumptions:

1. The linear relationship between independent and dependent variables.



2. The sum of residuals/error should be near to 0, and the sum of residuals we have got is -0.0007.

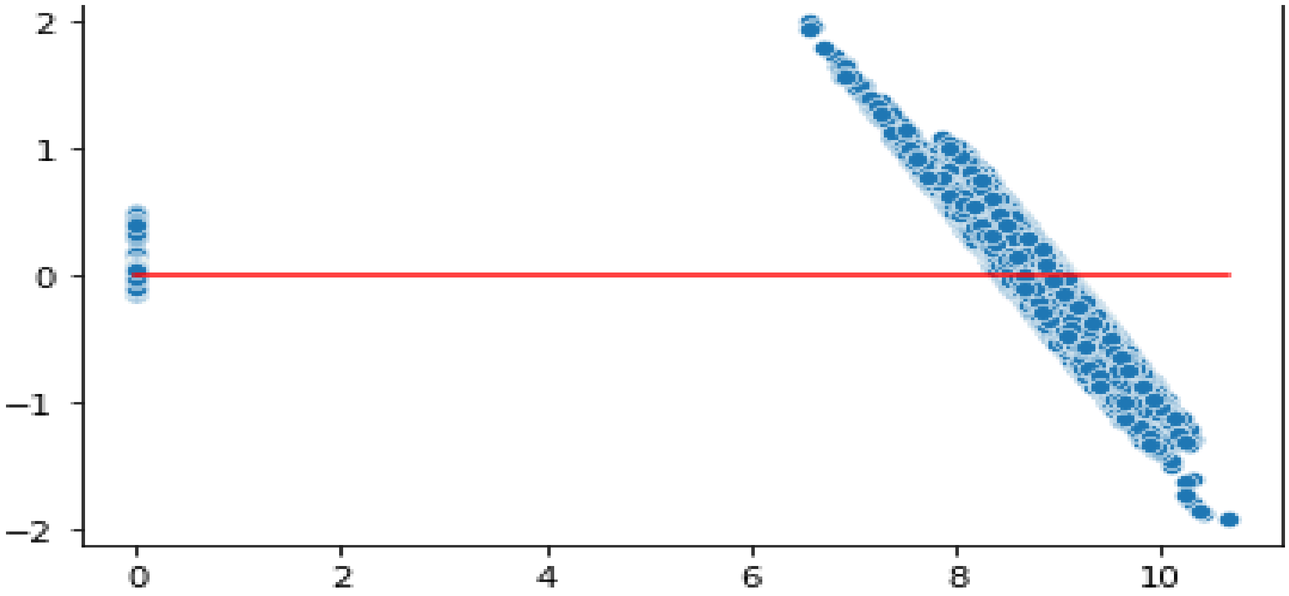
3. There should not be multicollinearity.



1	DayOfWeek	4.452981
2	Open	6.191915
3	Promo	1.872410
4	SchoolHoliday	1.342025
5	public_holiday	1.099311
6	easter_holiday	1.074031
7	christmas	1.072576
8	Month	3.904310
9	Day	3.965944
10	CompetitionDistance	1.530442
11	Store_Type	3.705991
12	Assortment_Level	4.908106
13	Promo_cont	1.176941
14	CompetitionTime	1.408883

Assumptions:

4. Checking the Heteroscedasticity.



Evaluation metrics for our multiple regression models used while model building:

Linear Regression model:

Mean Squared Error : 0.11
Root Mean Squared Error : 0
Mean Absolute Error : 0.25
Mean Absolute Percentage Error : 3.41 %
R-Square : 0.99
Adjusted R-Square : 0.99

Lasso Regression with cross validation:

Mean Squared Error : 0.11
Root Mean Squared Error : 0
Mean Absolute Error : 0.25
Mean Absolute Percentage Error : 3.41 %
R-Square : 0.99
Adjusted R-Square : 0.99

ElasticNet Regression:

Mean Squared Error : 0.11
Root Mean Squared Error : 0
Mean Absolute Error : 0.25
Mean Absolute Percentage Error : 3.41 %
R-Square : 0.99
Adjusted R-Square : 0.99

Ridge Regression with cross validation:

Mean Squared Error : 0.11
Root Mean Squared Error : 0
Mean Absolute Error : 0.25
Mean Absolute Percentage Error : 3.41 %
R-Square : 0.99
Adjusted R-Square : 0.99

Stack Model:

Stacking is a good way to combine all the predictions from different models into one. We can adjust weights for each model in stacking.

Evaluation metrics of our Stack model:

Mean Squared Error : 0.11

Root Mean Squared Error : 0

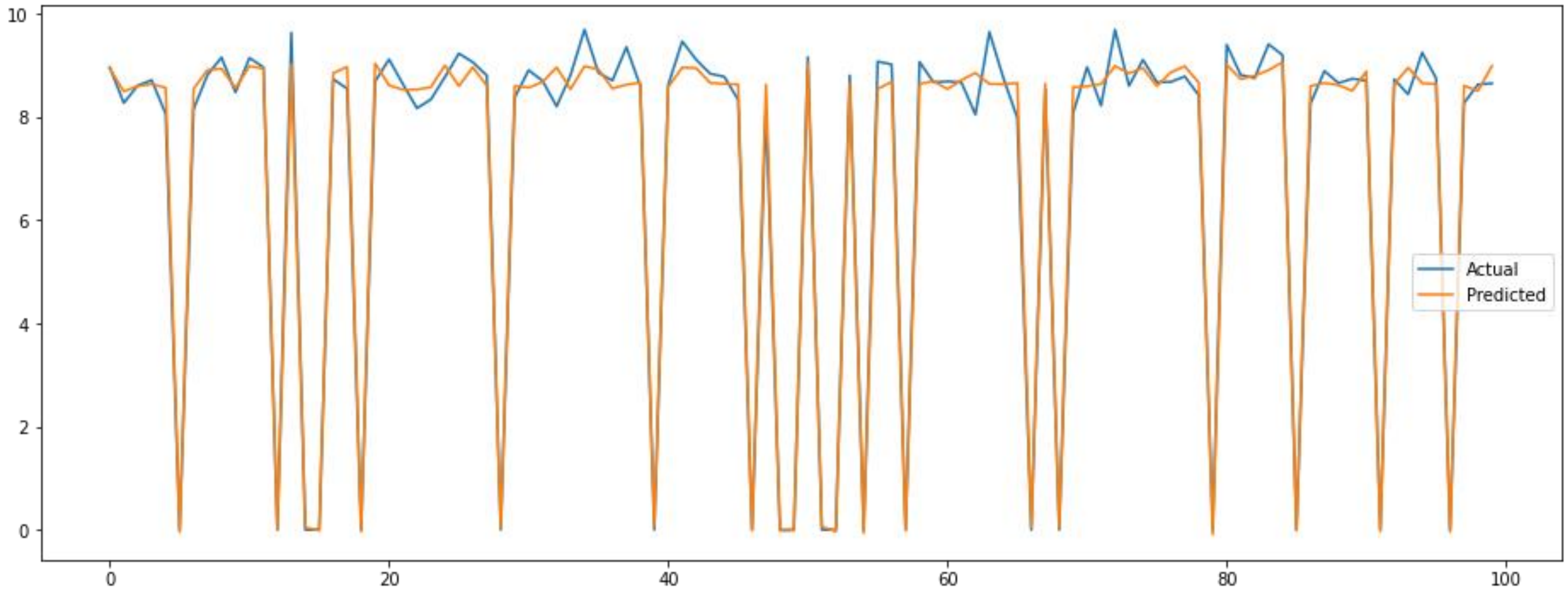
Mean Absolute Error : 0.25

Mean Absolute Percentage Error : 3.41 %

R-Square : 0.99

Adjusted R-Square : 0.99

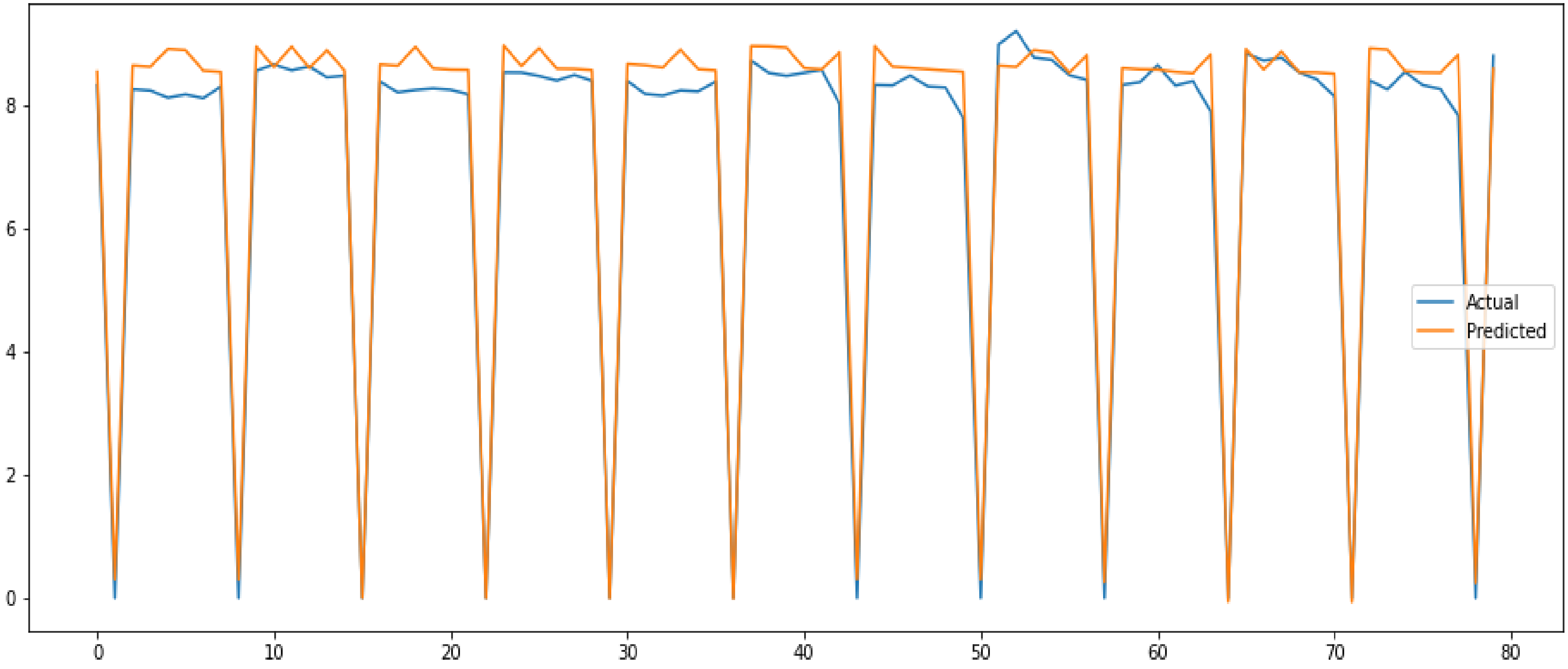
Graph for Stack Model:



Predicting Sales for next six weeks:

Now that we have linear regression models performing well.
we need to predict Sales for next 6 weeks in Advance.
This is what our problem statement states and hence we have considered the last 42 days as our test data.

Predicting Sales for next six weeks:



Challenges:

1. Here we had to deal with the larger datasets. We had to merge two datasets into one to get the meaningful insights from data.

2. We did a lot of feature engineering in order to get useful information from multiple columns.

3. We also applied log transform to a lot of columns to avoid skewness which helped us to predict our values well.

4. We generated the data for next six weeks using the test data which was the complex part in our project.

Conclusion:

The Rossmann store sales prediction is very engrossing data science problem to solve. We noticed that the problem is more concentrated towards the feature engineering and the feature selection part than on model selection. We had to spend around 60-70% of our time on analyzing data for trends in order to make our feature selection easier.

As we are building a liner regression model we emphasized on the basic 4 assumptions of a liner regression model.

Conclusion:

- We also have applied regularization techniques like Lasso, Ridge and Elastic Net to Avoid Overfitting.
- We also used Stacking to make predictions that have better performance than any single model.
- Most important feature came out to be customers, where sales is directly related to number of customers.
- The graph above is what our final model looks like at the end. As per our model prediction we can conclude that the Total revenue of Predicted Sales in next 6 weeks will be: 250,776,406 euros By 1115 Rossmann stores.