|  |
| --- |
|  |

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | |  |
| April 1/1/2015 | |  |
| Walmart Recruiting - Sales in stormy weather | |  |
| Notes | |  |
|  | |  |
| Isaac | |  |
| Authored by: Gino Tesei | |  |
|  | |  |

Exploratory analysis

# Weather

Exploratory analysis on weather data has been performed and related results are shown in the Excel document [weather\_elab.xlsx](https://github.com/gtesei/fast-furious/blob/master/competitions/walmart-recruiting-sales-in-stormy-weather/weather_elab.xlsx)

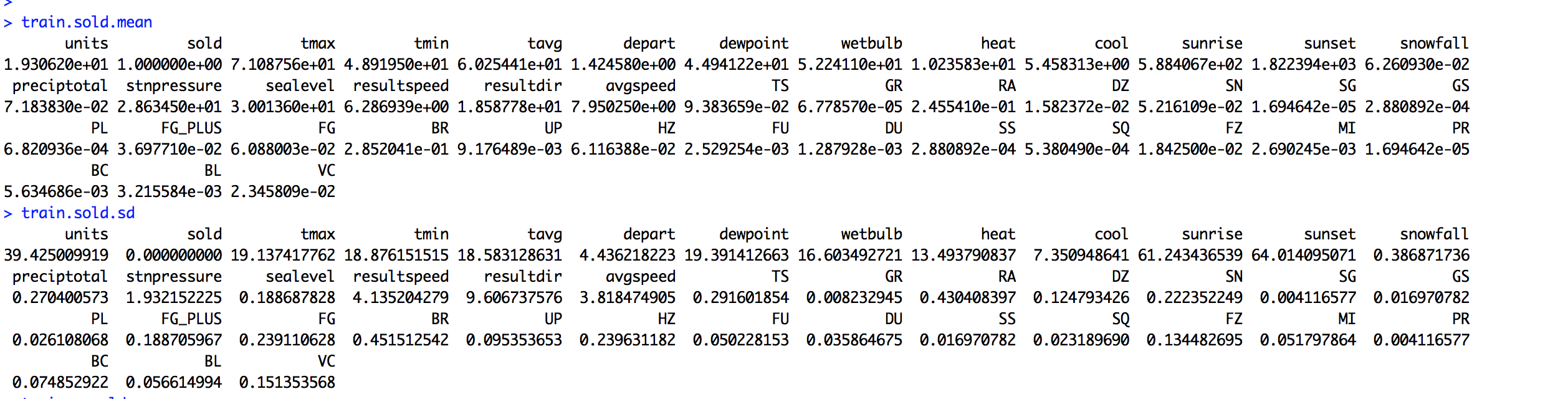
# Stores and products

In training data there’s only 2.5% of sold <units> with a value > 0, i.e. 97.5% of <date,store,item> in the training are 0s (=unsold).

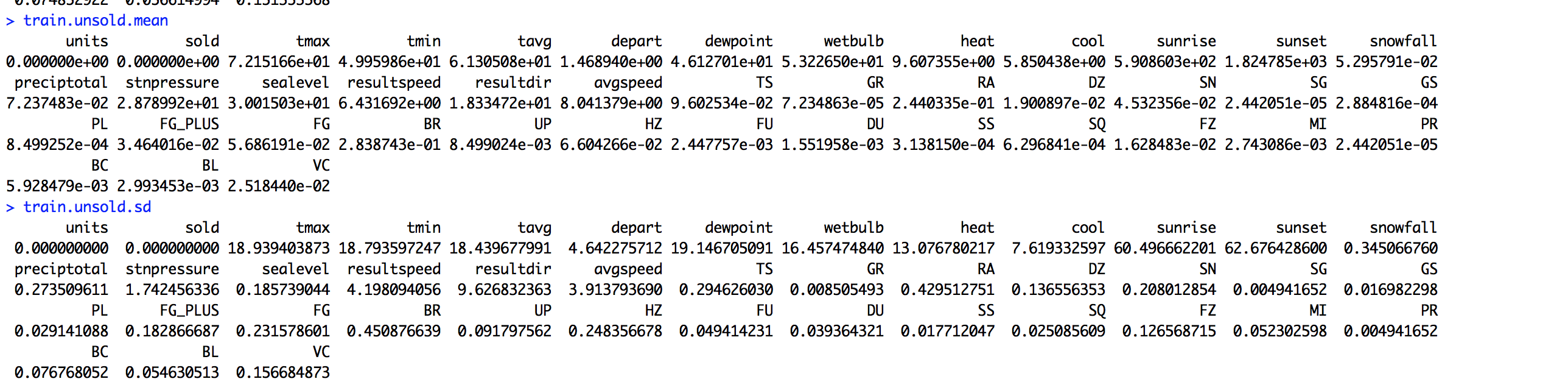
There are

* **20 weather stations** 
  + that are associated to **2.25 stores** on average (standard dev. = 1.51)
* **45 stores**
* **111 items**
* **4995 different combinations** of < store\_nbr, item\_nbr> (45 \* 111 = 4995), whose
  + **255 (5.1%)** combinations of < store\_nbr, item\_nbr> has at least one unit > 0 (=sold) one day in the training set
    - where the average units sold are 19.35 (standard dev.= 33.02)
    - corresponding to 236038 observations, i.e. 5.1% of total training observations
  + **4740 (94.9%)** combinations of < store\_nbr, item\_nbr> has units sold = 0 (=unsold) every day in the training set
    - corresponding to 4381562 observations, i.e. 94.9% of total training observations

# Predictors’ distribution in stations/products sold (5.1%)

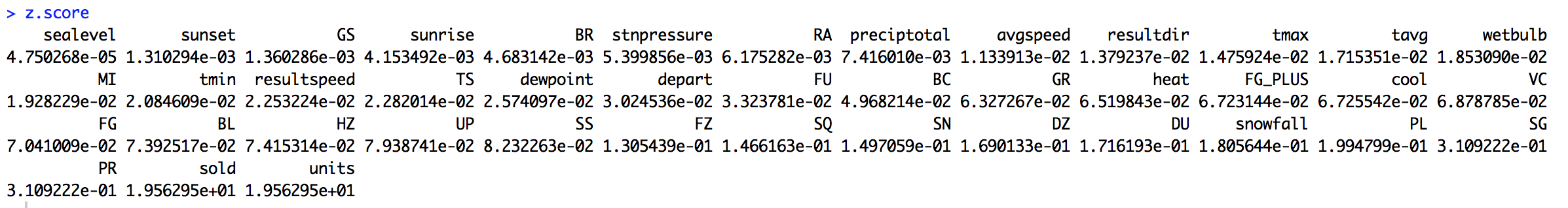


# Predictors’ distribution in stations/products unsold (94.9%)



# Z-score

Ordering by the difference of the predictors’ means in the two sets and dividing for the standard deviation, we have the following situation



So, the most promising predictors seem

* PR (z-score 0.31)
* SG (z-score 0.31)
* PL (z-score 0.19)
* Snowfall (z-score 0.18)
* DU (z-score 0.17)
* SN (z-score 0.15)

On the other hand, **z-score is** **always minor than 1**.

# Best selling combinations of stores / products

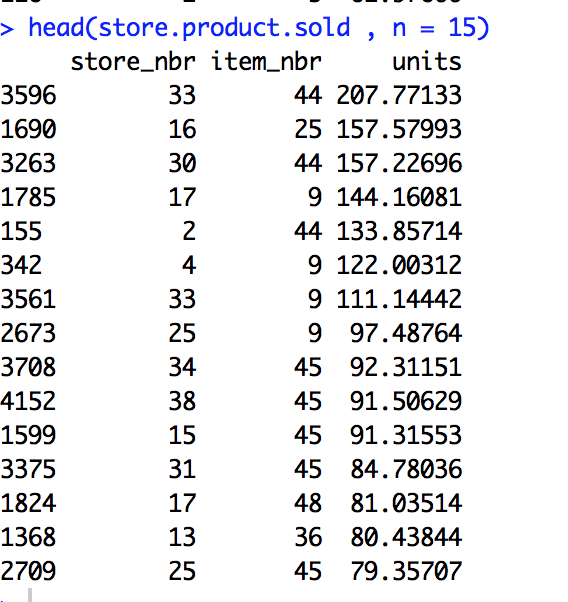


Figure units is the daily average of sold units

# Most sold products

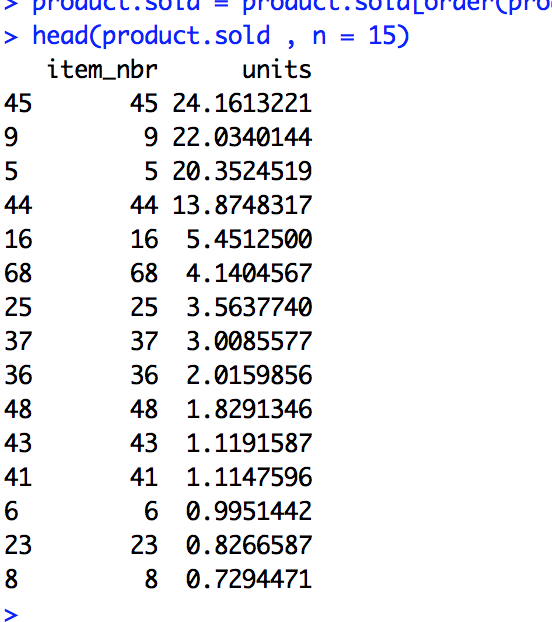


Figure units is the daily average of sold units

# Best selling stores

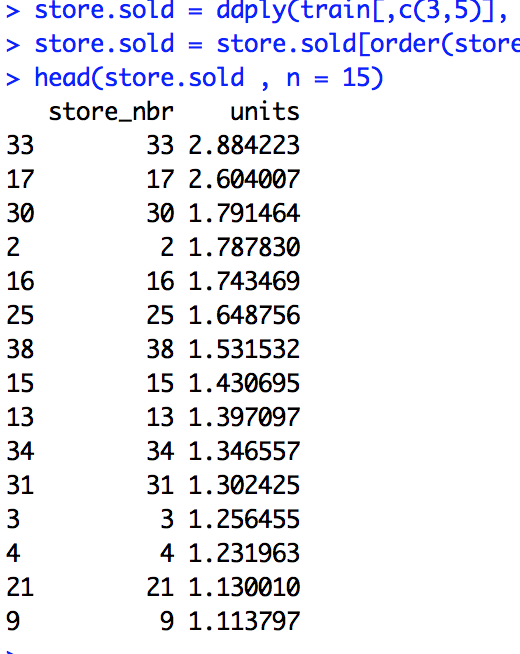


Figure units is the daily average of sold units

# Correlation among sold products (cross selling?)

This analysis can be done at three levels, i.e. for each observed day in the training set

* among the products sold in each store (45 correlation matrices)
* among the products sold in the stores associates to the same weather station (20 correlation matrices)
* among the products sold in all stores (1 correlation matrix)

# Correlation among products sold in each store

Let’s focus only in a given store and let’s consider the correlation among products sold in such a store in the same day.

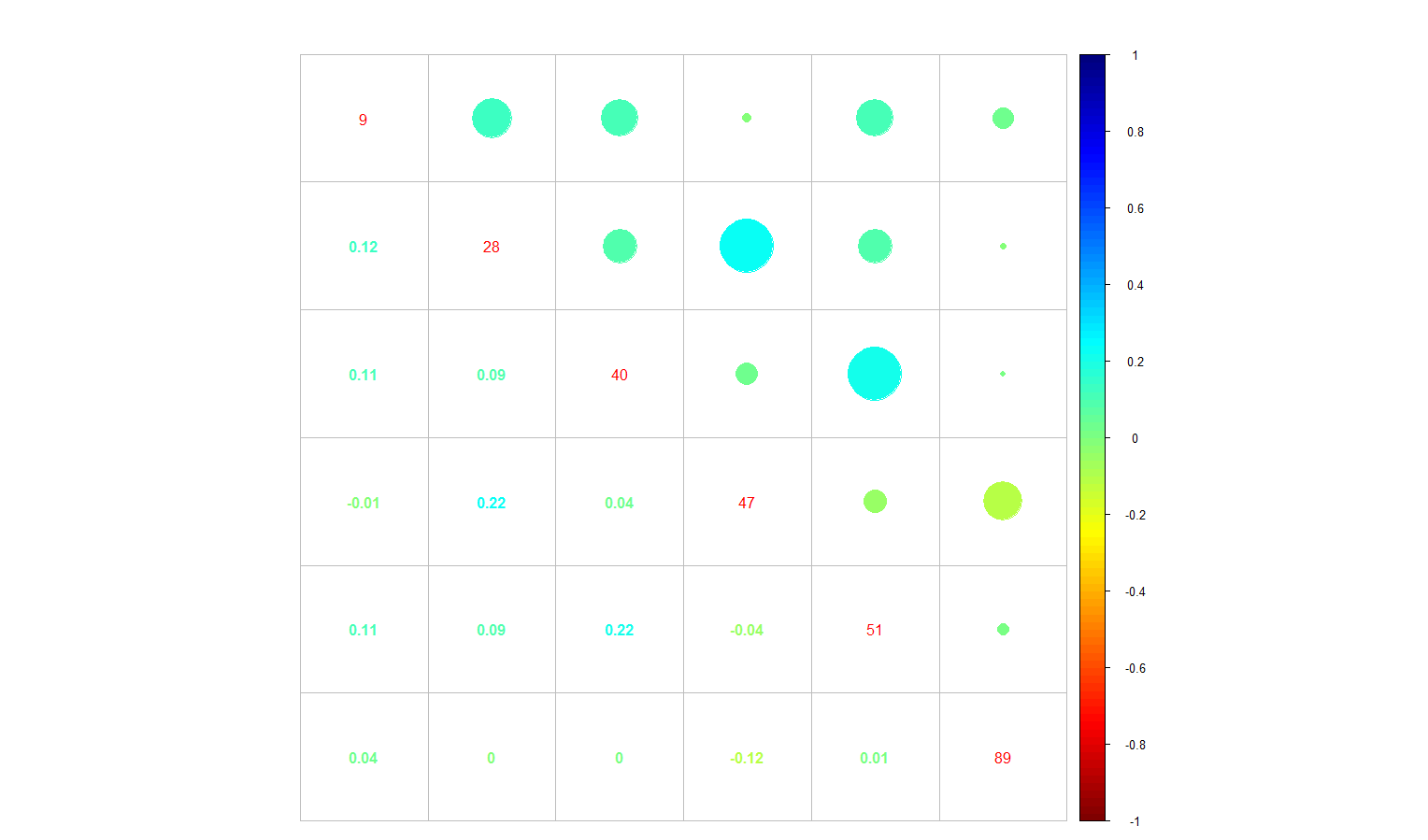


Figure Correlation plot of sold products sold in store N.1 (unsold products are discarded)

For example, as shown by the previous correlation plot, in the store N.1 we can observe a direct correlation (0.22) among products (40,51) and (47,28) while an inverse correlation (-0.12) between products (47,89).

Let’s focus now on the stores associated to the same weather station, i.e. assuming that selling happened with the same weather. For example, let’s consider the **stores 14 and 45 associated to the same weather station N. 16**.

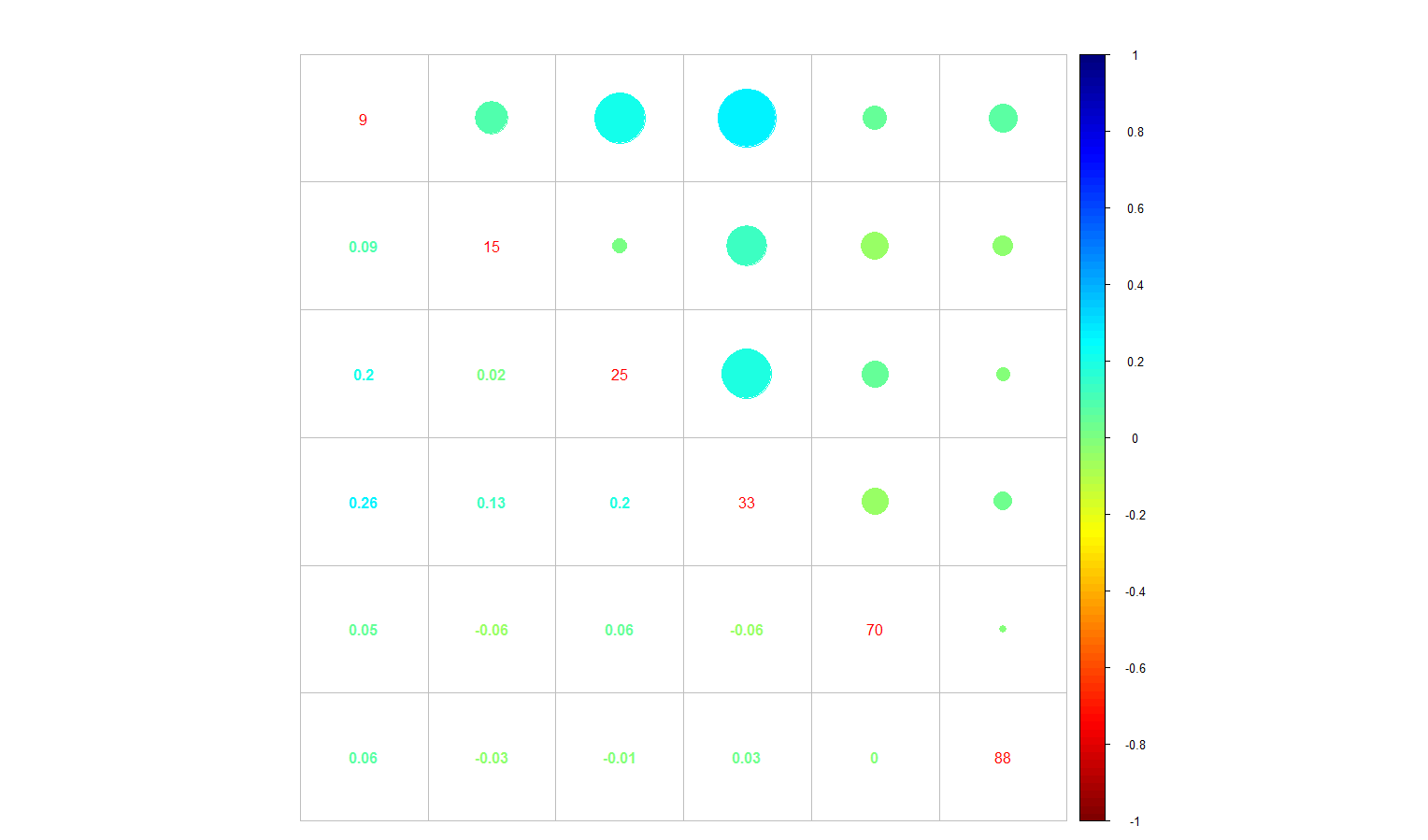


Figure Correlation plot of sold products (unsold products are discarded) sold in store N.14 (station N. 16)

In the **store N. 14**, we can observe

* a **direct correlation (0.26)** between **products (33,9)**,
  + p-value: 4.441e-16
  + 95 percent confidence interval: 0.1998593 0.3183700
  + 99 percent confidence interval: 0.1805916 0.3362195
* a **direct correlation (0.2)** between **products (25,9)**,
  + p-value: 2.75e-10
  + 95 percent confidence interval: 0.1409988 0.2628716
  + 99 percent confidence interval: 0.1213516 0.2813823
* an inverse poor correlation (-0.06) between products (9,88)
  + **p-value: 0.05624**
  + **95 percent confidence interval: -0.001646013 0.124943908**
  + **99 percent confidence interval: -0.02163447 0.14457201**

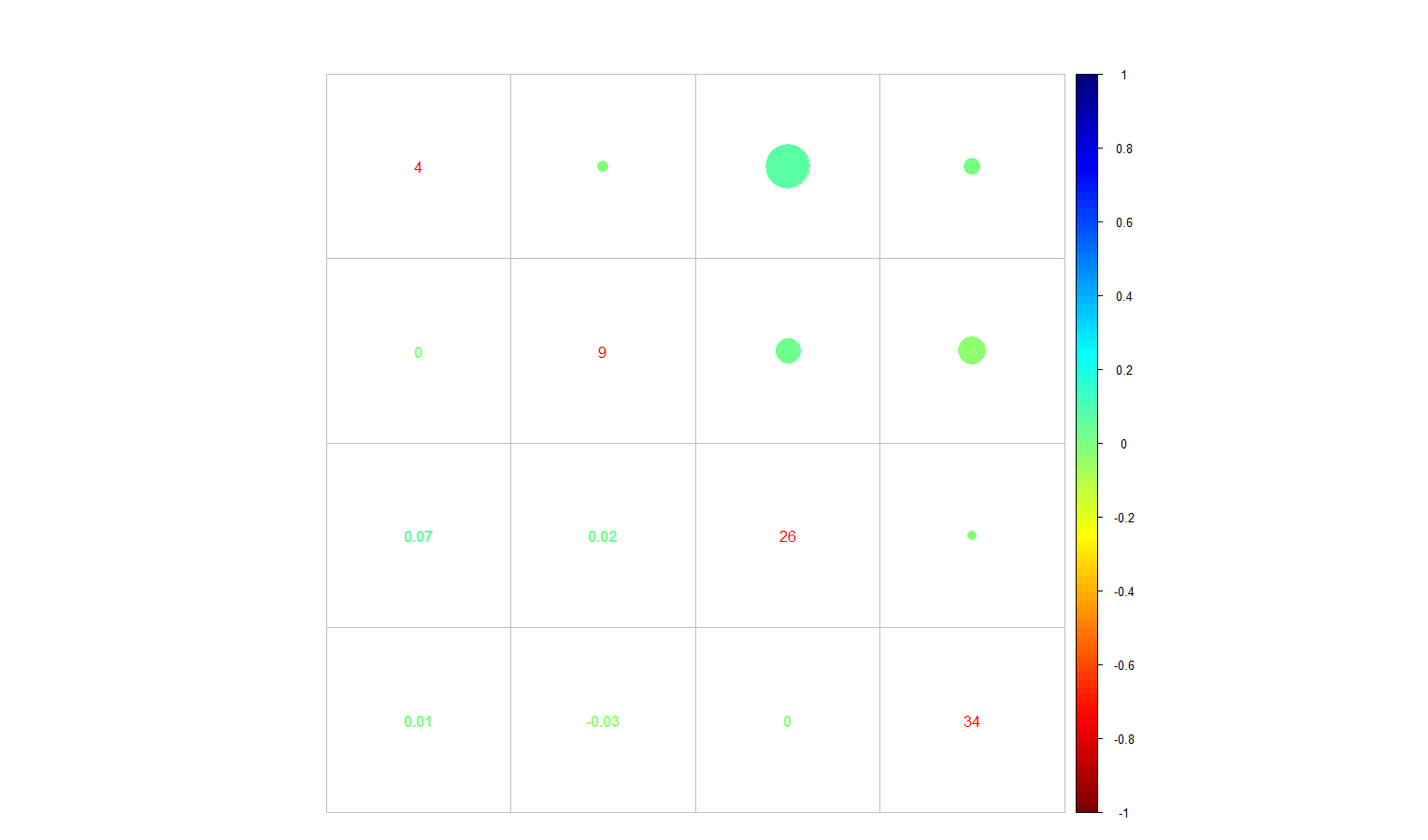


Figure Correlation plot of sold products (unsold products are discarded) sold in store N.45 (station N. 16)

In the **store N. 45**, we can observe

* a direct **correlation (0.07)** between products **(26,4)**
  + p-value: 0.02467
  + 95 percent confidence interval: 0.00931338 0.13571730
  + **99 percent confidence interval: -0.01067778 0.15528520**
* an inverse poor correlation (-0.03) between products (34,4)
  + **p-value: 0.7548**
  + **95 percent confidence interval: -0.05343611 0.07362572**
  + **99 percent confidence interval: -0.07334693 0.09347732**

The question is why in store probably in the same area (i.e. associated to the same weather station) we have profiles so different? First of all, except for the product N.9, such stores has different sold products, i.e.

* Store N. 45: 9, 4, 26, 34
* Store N. 16: 9, 15, 25, 33, 70, 88

Moreover, the product correlation is different. We can make several hypothesis, e.g.

* different customers and/or different customer’s needs
* products out of stock in a store but not in the other
* etc.

In any case, we can conclude that **product correlation among the products sold in the stores associates to the same weather station is not very meaningful, in general**. For instance, referencing to the previous case, calculating the correlation between item N. 4 (sold in the store N. 45) and item N. 15 (sold in the store N. 16) is not very meaningful, as the previous item hasn’t been sold in the second store, and vice-versa.

# Correlation among products in the stores associates to the same weather station

Referencing to the previous example (i.e. stores 14 and 45 associated the weather station 16), different considerations can be done for the **item N.9 sold in both stores**. On the other hand, we observe a poor **direct correlation of 0.05579844**

* **p-value: 0.0853**
* **95 percent confidence interval: -0.007766513 0.118914267**
* **99 percent confidence interval: -0.02775138 0.13857405**

## Repeating same analysis for all store clusters

TODO

# Correlation among products sold in all stores

Referencing to the previous example (i.e. stores 14 and 45 associated the weather station 16), as the correlation observed for the only item sold in both stores is not statistic significant, it seems not very meaningful to consider sold units of item N. 9 in stores associated to other weather stations.

**Do test dates occur in train set?**

Weather imputation models

Discarded input variables

* station\_nbr
* date

All predictors are assumed as numeric (no factors)

**Models & Performance**

* Performed **basic** imputation with **BlackGuido[[1]](#footnote-1)** on Mode/Average/LineraReg and observed mean imputing performance (RMSE) **17.9**
* Performed **full** imputation with **BlackGuido** on Mode/Average/LineraReg/KNN\_Reg/PLS\_Reg/Ridge\_Reg/SVM\_Reg/Cubist\_Reg and observed mean imputing performance (RMSE) …

Predictive model #1 – basic

* For each date <d> and for each item <i> sold/predicted to be sold units in the store <s>, the related train/test set has been built with (imputed) weather data of the station <st> associated to the store <s> in the key.csv file, and the related output variable is the sold units for <d,s,i>
* Feature selection
  + Removing predictors that make ill-conditioned square matrix
  + Removing near zero variation predictors
  + Removing high correlated predictors
* Feature scaling
* Resampling: bootstrap + k-folds
* Models:
  + Average
  + Mode
  + LinearReg
  + RobustLinearReg
  + PLS\_Reg
  + Ridge\_Reg
  + Enet\_Reg
  + KNN\_Reg
  + BaggedTree\_Reg
* Leaderboard performance on dataset filled with basic imputation: **xxxx**

# Example (store n.1 / product n.9 / weather station n.1)

The output variable (units sold) has 929 observations with mean 29.48 and standard deviation 22.07. There are only 10 observations out of 929 (1%) with 0 units sold.

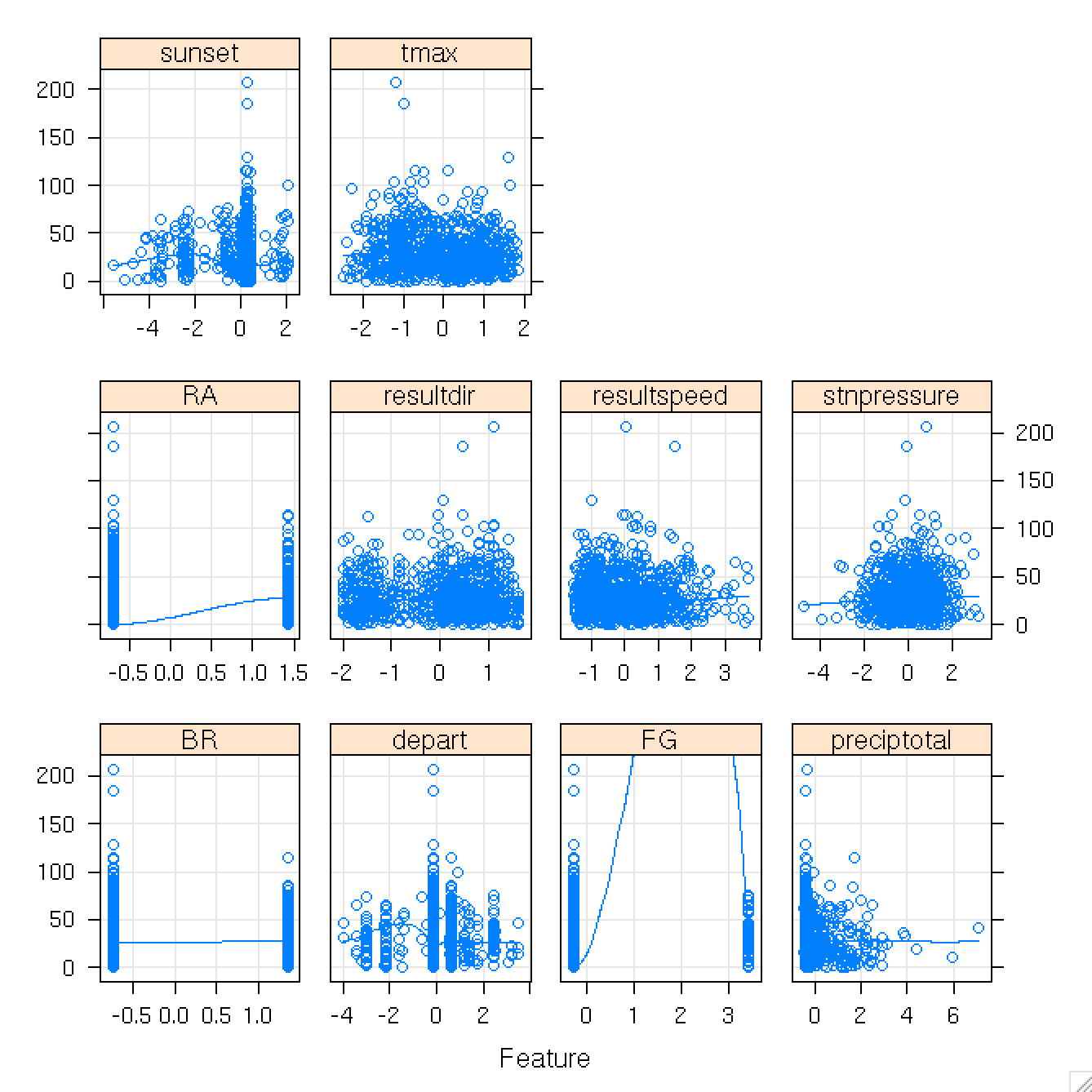


Figure Scatter plots of predictors for store n.1 / product n.9 / weather station n.1 versus units sold

In the scatter plots, features are discarded according to the described feature selection process and scaled.

**Are *a-priori* best store/product combinations still the best ones in prediction?**

**Are *a-priori* best-selling stores still the best ones in prediction?**

**Are *a-priori* most sold products still the most sold in prediction?**

**Do *a-priori* product correlations hold in predicted units sold?**

Predictive model #2 – product correlation

Predictive model #3 – using libsvm

1. BlackGuido is part of the proprietary machine learning framework *fast-furious* [↑](#footnote-ref-1)