|  |
| --- |
|  |

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

Exploratory analysis

# Weather

Exploratory analysis 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)

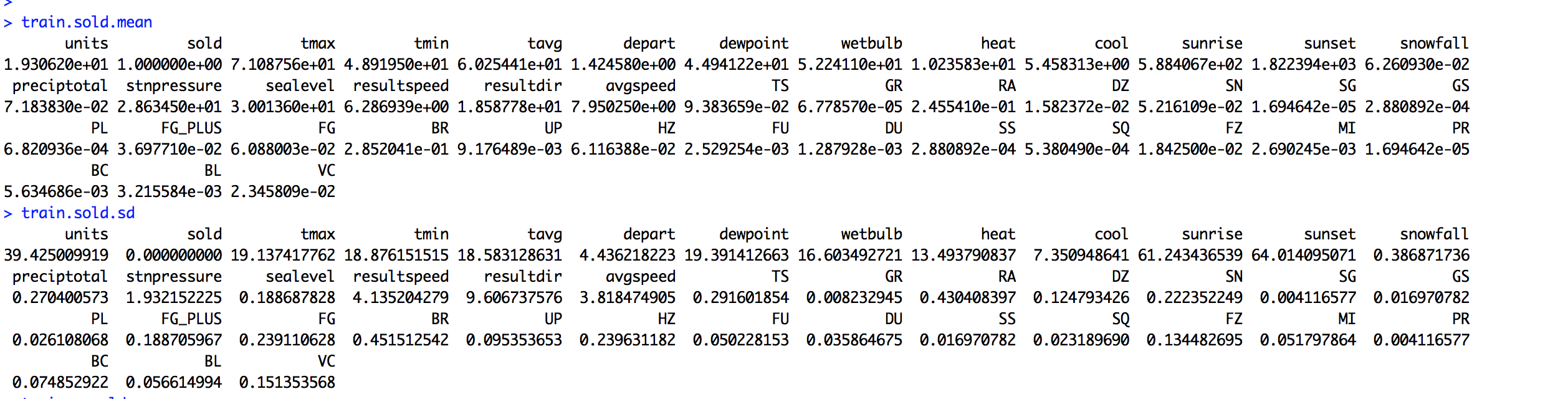
# Stores and products

There’s only 2.5% of <units> field in training data with a value > 0, i.e. 97.5% of <date,store,item> in the training are 0s.

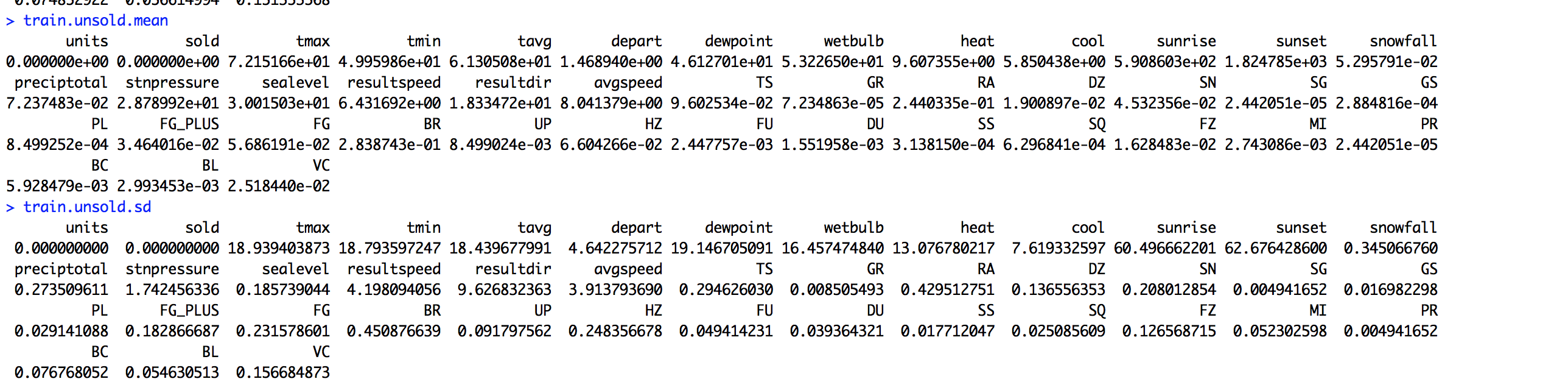
There are

* **20 weather stations** 
  + that are associated to **2.25 stores** on average (sd = 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 units > 0 one day in the training set
    - where the average units sold are 19.35 (sd = 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 = 0 each day in the training set
    - corresponding to 4381562 observations, i.e. 94.9% of total training observations

# Distributions of predictors in stations/products sold (5.1%)

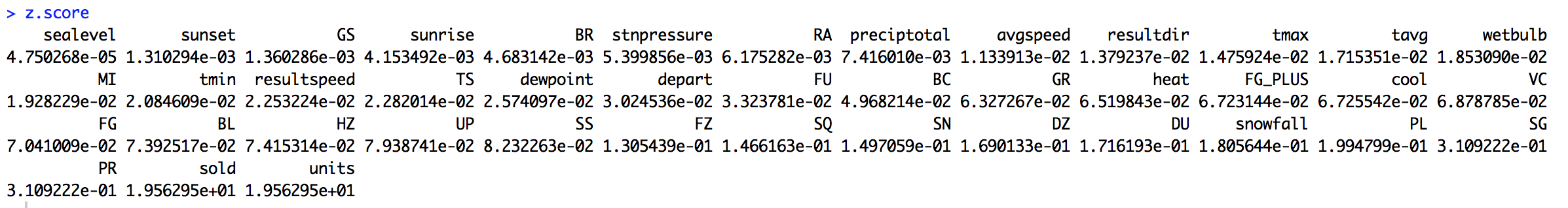


# Distributions of predictors in stations/products unsold (94.9%)



# Z-score

Sorting the difference of the mean of each predictor in the two sets dividing for the standard deviation of the predictor, we have the following situation

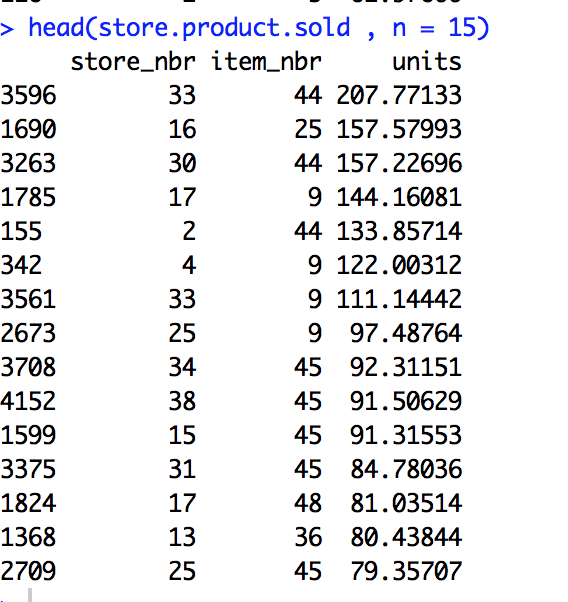


So, the most promising predictors seem

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

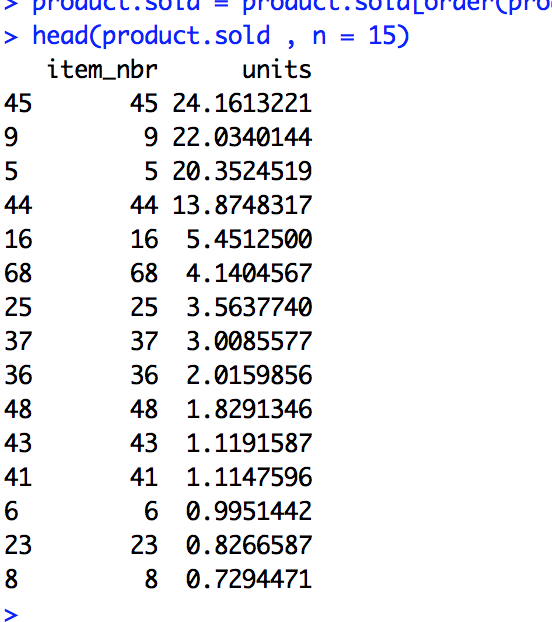
On the other hand, **z-score is** **always minor than 1**.

# Best selling combinations of stores / products



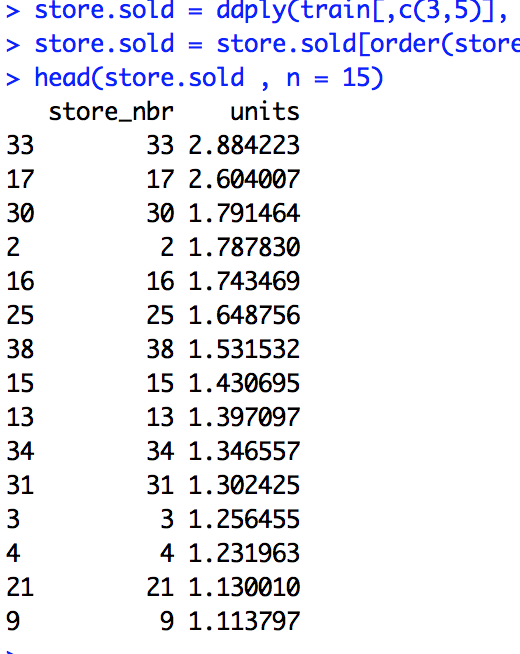
(where *units* is the average of sold units)

# Most sold products



(where *units* is the average of sold units)

# Best selling stores



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

# Correlation among products sold in all stores

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

# Correlation among products sold in each store

Weather imputation models

Discarded input variables

* station\_nbr
* date

All predictors are assumed as numeric (no factors)

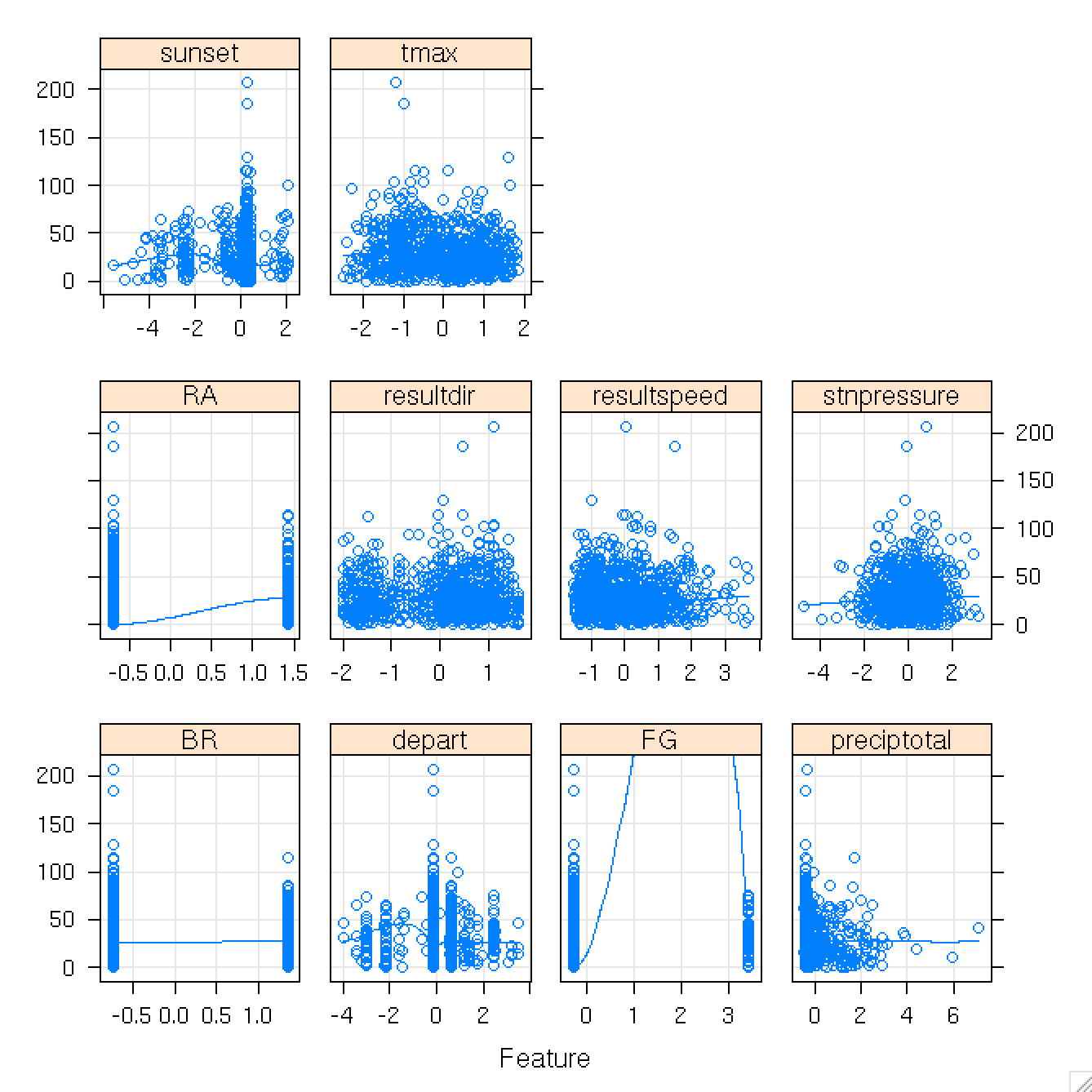
**Models & Performance**

* 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 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 training variable are the units sold in the train.csv file for <d,s,i>
* Feature selection
  + Removing predictors that make ill-conditioned square matrix
  + Removing near zero var 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)



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

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

Predictive model #2 – product correlation

If products

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