Data analysis on Weekly Sales Transaction of products over 52 weeks

Course: MSBA 305

Term paper Project

Instructor: Dr. Mayank Johri

Student: Sadegh Bamohabbat Chafjiri

November 9th 2018

Introduction 3

Data Collection and Description 3

Descriptive Statistics 4

Correlation Analysis 7

Statistical tests 8

Regression Analysis 10

Conclusion 19

Reference 20

Introduction

In this project, I perform the statistical analysis including descriptive, predictive, and prescriptive analysis on Weekly Sales Transaction. The dataset is a .csv file with a clustering Characteristics reporting the number of Sales Transaction during the 52 weeks based on the Product Codes. The data set has 52 columns, one column for each week, and 819 rows, one row for each Product Code. Moreover, Data set provided a normalized value in extra 52 columns. The Variables of datasets are:

- Product_Code
- 52 weeks: W0, W1, ..., W51
- Normalized values of weekly data: Normalized 0, Normalized 1, ..., Normalized 51

In this project, I do correlation analysis on all weeks to find the relationship between different weeks. Moreover, I perform market basket analysis in week 25 to measure the sales relations between this week and other weeks. It would be a good measure to model the quantity growth. In this project, I perform multiple regression for week 25 and then will show this method is not accurate for clustering data type. Then, I will perform clustering with 6 different classifiers to find the fit model to predict week 25.

Data Collection and Description:

According to Crawford (2017), dataset is related to a Weekly Sales Transaction. Crawford got this file from Tan, Chuan, Lau, Pei (2014) which uploaded on UCI Machine Learning Repository. The Format of data set is csv file. I remove the extra columns in different sections of project which are not related to that part of analysis.

Descriptive Statistics:

In this section, I remove the normalized values and reduced the table size to 52 columns for 52 weeks. I show a summary of statistical characteristics for 52 weeks including characteristics of each week such as *standard deviation, mean, minimum, maximum, 1st, 2nd (median) and 3rd quartile values* (See table 1, 2 and 3). Table 1 shows that week 51 has the biggest range of Sales from 0 to 73 for different barcodes whereas the week 29 varies in the smallest interval of sales quantities between 0 to 42. In addition, median varies between 3 to 5 between week 0 to 51. Interestingly, in most of weeks, median varies from 3 to 4. However, the maximum median is 5 which is belonged to 4 weeks; week 24, 25, 50, 51. Week 25 and 50 have the highest 3rd quartile value on 15 and 14 respectively which means 75% of quantities would be less than 15 in week 25 (see table 2 and fig. 1) and 75% of quantities would be less than 14 in week 51 (see table3 and fig. 1)

Table 1. Summary of attributes

	W0	W1	W2	W3	W4	W5	W6	W7	W8	W9	...	W42	W43	
count	811.000000	811.000000	811.000000	811.000000	811.000000	811.000000	811.000000	811.000000	811.000000	811.000000	...	811.000000	811.000000	81
mean	8.902589	9.129470	9.389642	9.717633	9.574599	9.466091	9.720099	9.585697	9.784217	9.681874	...	8.394575	8.318126	
std	12.067163	12.564766	13.045073	13.553294	13.095765	12.823195	13.347375	13.049138	13.550237	13.137916	...	11.348777	11.250455	1
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	
25%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	
50%	3.000000	3.000000	3.000000	4.000000	4.000000	3.000000	4.000000	4.000000	4.000000	4.000000	...	4.000000	4.000000	
75%	12.000000	12.000000	12.000000	13.000000	13.000000	12.500000	13.000000	12.500000	13.000000	13.000000	...	10.000000	11.000000	1
max	54.000000	53.000000	56.000000	59.000000	61.000000	52.000000	56.000000	62.000000	63.000000	52.000000	...	52.000000	50.000000	4

Table 2. The smallest range of sale in week 29

	W25	W26	W27	W28	W29
count	811.000000	811.000000	811.000000	811.000000	811.000000
mean	8.892725	6.950678	7.193588	7.383477	7.339088
std	10.016879	9.449151	9.729464	10.187970	10.161884
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1.000000	0.000000	0.000000	0.000000	0.000000
50%	5.000000	3.000000	3.000000	3.000000	3.000000
75%	15.000000	9.000000	10.000000	9.000000	10.000000
max	64.000000	46.000000	47.000000	44.000000	42.000000

Table 3. The biggest range of sales in week 29

W47	W48	W49	W50	W51
811.000000	811.000000	811.000000	811.000000	811.000000
8.670777	8.674476	8.895191	8.861899	8.889026
11.435870	11.222996	10.941375	10.492710	9.558011
0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	1.000000	1.000000	1.000000	1.000000
4.000000	4.000000	4.000000	5.000000	5.000000
12.000000	12.000000	12.000000	13.000000	14.000000
49.000000	50.000000	52.000000	57.000000	73.000000

Fig 1 shows all week are left skewed. The box plots illustrate the characteristics of each week including mean, median, maximum and minimum and and outlier values of quantities for each week schematically.

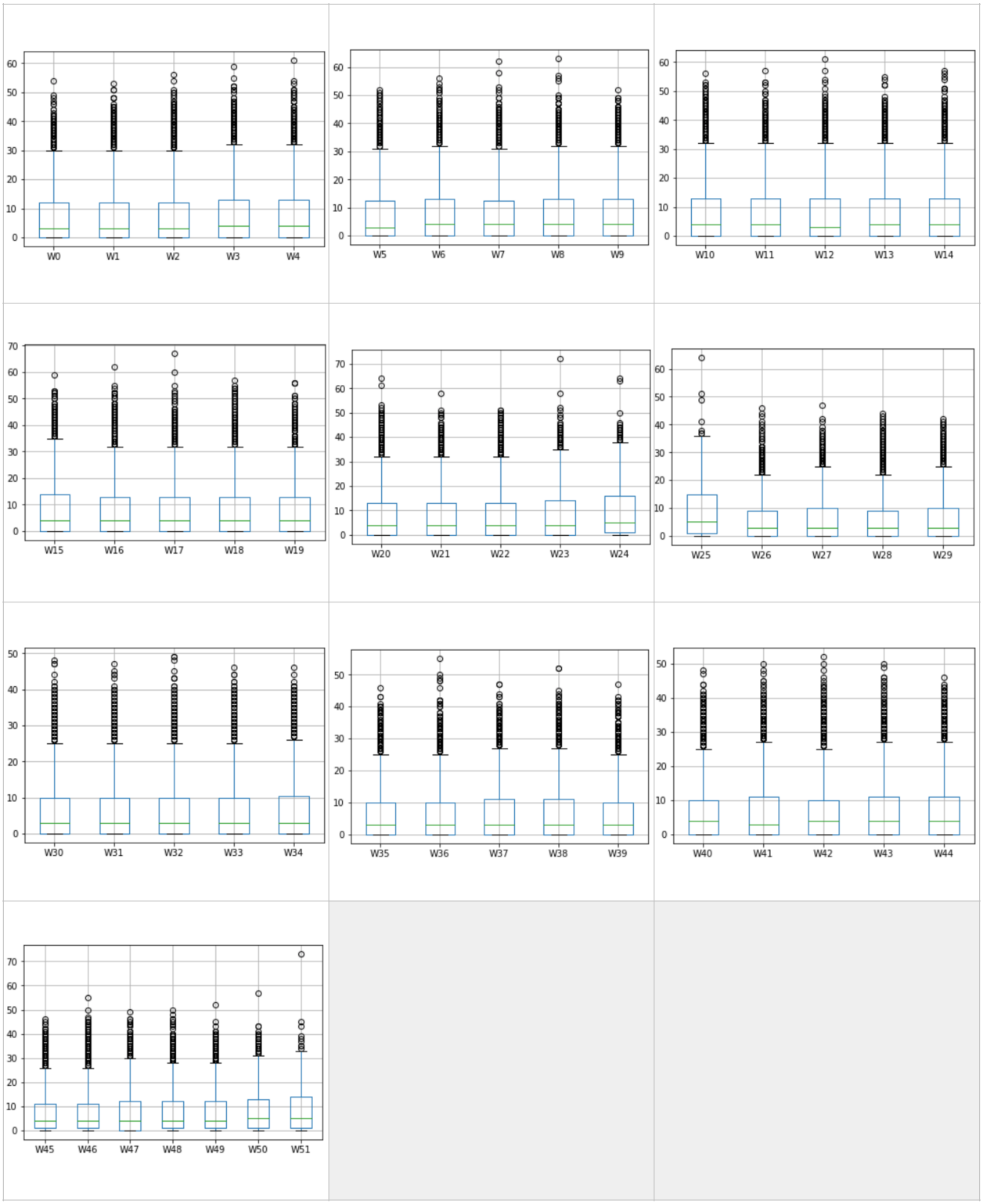

Fig.1 Box-plot of sales quantities in different weeks

Correlation Analysis

In Fig 2, I perform the correlation analysis on all weeks. It shows that roughly all weeks are highly correlated to each other. The correlation ratio is more than 0.8 for all weeks. However, week 24, 25, 50 and 51 have the minimum ratio of correlation around 0.85.

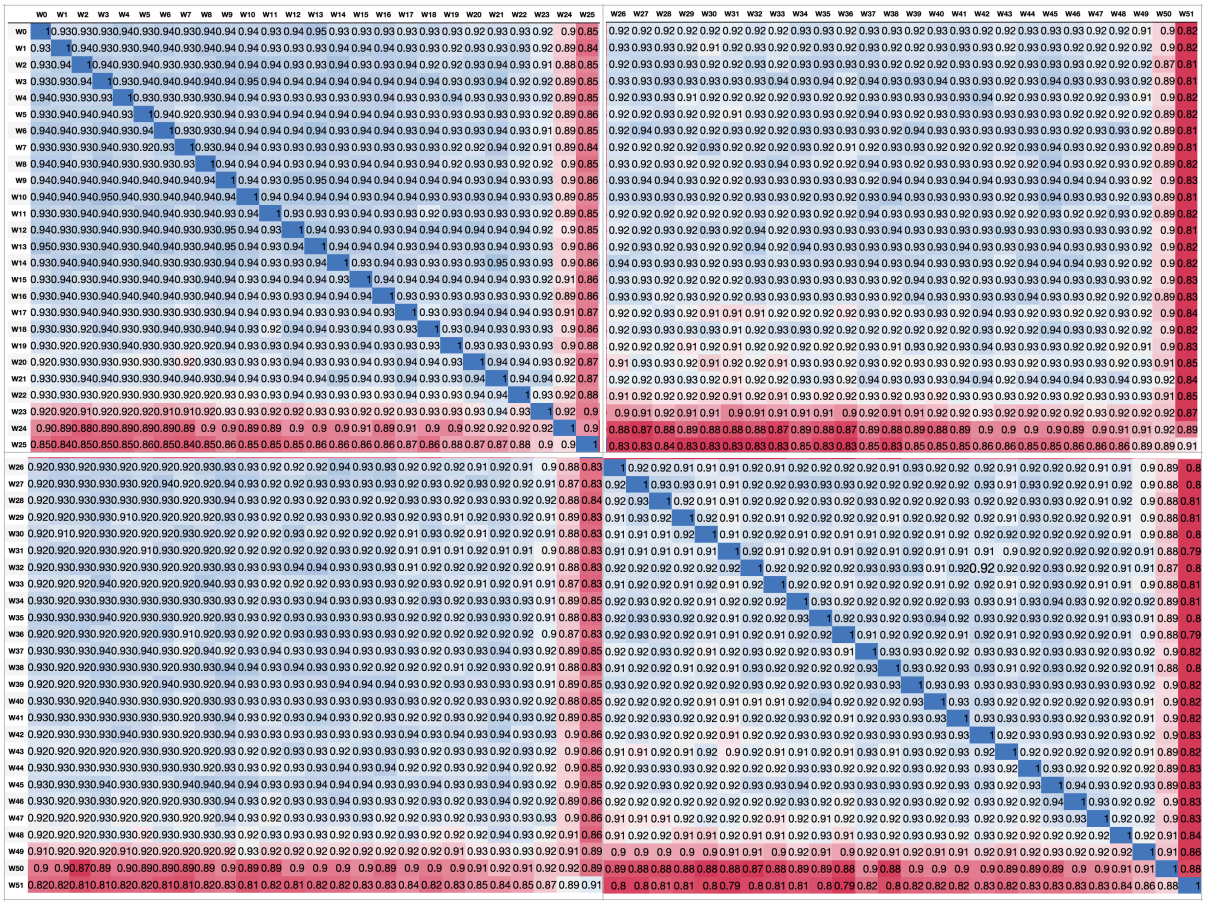

Fig 2. correlation analysis between all attributes

Statistical tests:

In the previous section, correlation analysis, I presented the relationship between weeks. However, it is evident that the strength of the correlation between attributes does not imply causation. In this section, I performed ANOVA (analysis of variance) to evaluate predictability (in the best situation causality) of weeks and find attributes which are statistical significant. My tool to achieve this purpose is p-value. I chose week 25 to evaluate how other weeks are statistical significant with this week . Ony-way ANOVA is a tool to evaluate if the groups of data could be statistical significant or not. In the regression section I will show that the best approach to find significant factors would be classification and performing clustering based on products.

Null Hypothesis: There was no difference between mean of W25 and attributes in below:

`W5,W9,W14,W19,W22,W23,W24,W29,W32,W33,W39,W49,W51`

Alternative Hypothesis: There was a difference between mean of W25 and attributes in below:

`W5,W9,W14,W19,W22,W23,W24,W29,W32,W33,W39,W49,W51`

The threshold for amount of statistical significant attributes in data analytics is 0.05. If the p-value is less than 0.05, null hypothesis is rejected and it means there is a statistical significance between assumed groups. In this case, p-values is less than 0.05 and null hypothesis is rejected. In other words, there is a statistical significance between Week 25 and Weeks 5,9,14,19,22,23,24,29,32,33,39,49 and 51 (see table 3).

table 4: analysis of variance between Week 25 and other weeks

	sum_sq	df	F	PR(>F)
W0	6.310168	1.0	0.596113	4.403052e-01
W1	51.968980	1.0	4.909440	2.700629e-02
W2	46.547358	1.0	4.397266	3.632693e-02
W3	3.719886	1.0	0.351413	5.534911e-01
W4	38.689844	1.0	3.654978	5.627817e-02
W5	53.303479	1.0	5.035508	2.512029e-02
W6	1.710073	1.0	0.161548	6.878481e-01
W7	10.368295	1.0	0.979479	3.226429e-01
W8	6.793122	1.0	0.641737	4.233332e-01
W9	54.335617	1.0	5.133013	2.375589e-02
W10	3.217965	1.0	0.303997	5.815498e-01
W11	33.657738	1.0	3.179601	7.496245e-02
W12	0.600620	1.0	0.056740	8.117895e-01
W13	8.693415	1.0	0.821255	3.651010e-01
W14	43.047918	1.0	4.066679	4.408891e-02
W15	8.066831	1.0	0.762063	3.829599e-01
W16	19.859386	1.0	1.876090	1.711843e-01
W17	1.050779	1.0	0.099266	7.527993e-01
W18	0.008272	1.0	0.000781	9.777064e-01
W19	283.901062	1.0	26.819753	2.863754e-07
W20	0.034059	1.0	0.003217	9.547809e-01
W21	2.881287	1.0	0.272191	6.020178e-01
W22	94.293174	1.0	8.907750	2.930797e-03
W23	85.396887	1.0	8.067329	4.627692e-03
W24	211.256548	1.0	19.957123	9.125264e-06
W26	36.228480	1.0	3.422456	6.470377e-02
W27	4.279678	1.0	0.404295	5.250709e-01
W28	21.303810	1.0	2.012542	1.564142e-01
W29	117.773592	1.0	11.125913	8.926835e-04
W30	26.863776	1.0	2.537785	1.115670e-01
W31	2.477310	1.0	0.234028	6.286925e-01
W32	89.497664	1.0	8.454724	3.747158e-03
W33	72.937287	1.0	6.890288	8.840906e-03
W34	0.477124	1.0	0.045073	8.319264e-01
W35	18.954798	1.0	1.790634	1.812496e-01
W36	10.459932	1.0	0.988136	3.205152e-01
W37	1.128848	1.0	0.106641	7.440915e-01
W38	36.154371	1.0	3.415455	6.497745e-02
W39	74.309770	1.0	7.019944	8.228381e-03
W40	3.492769	1.0	0.329957	5.658545e-01
W41	0.000254	1.0	0.000024	9.960945e-01
W42	2.651954	1.0	0.250527	6.168496e-01
W43	14.930653	1.0	1.410479	2.353485e-01
W44	34.302194	1.0	3.240482	7.223645e-02
W45	23.720487	1.0	2.240843	1.348234e-01
W46	0.891595	1.0	0.084228	7.717271e-01
W47	11.130165	1.0	1.051452	3.054996e-01
W48	14.341029	1.0	1.354778	2.448103e-01
W49	298.612073	1.0	28.209482	1.431565e-07
W50	63.673808	1.0	6.015179	1.440740e-02
W51	1477.774651	1.0	139.603390	1.085528e-29

Regression Analysis:

This section has two parts; A and B. In part A, I show that multiple regression model can not model clustering datatype accurately and in the next section, I perform 6 methods of classifications for clustering datatype.

A- Multiple Regression Analysis:

In this section, I perform the regression analysis to model week 25. Regression model can help decision makers to estimate their sales in different weeks based on the sales they had before. I perform multiple regression analysis in the first step to show that why the cluster analysis is more effective than multiple regression analysis for highly correlated attributes. I tell why the result of multiple regression is not acceptable for clustering despite giving a OLS model and why I should do clustering dataset to get prediction model.

Backward selection method:

In the first step, I built a multiple regression model by backward selection method. Our attributes are discrete, which this method works for Ordinary least squares (OLS). In this model by removing attributes one-by-one (which have p-value more than 0.05) I could find a linear model.

R Squared Equation for week 25 is equal to 0.0699*W9+ 0.1717*W19+ 0.0822*W22+ 0.1004*W23+ 0.1347*W24 -0.1401*W29 -0.1300*W32 -0.1213*W33+ 0.2097 *W49+ 0.0805*W50+ 0.3711*W51

Table5. Regression analysis for week 25

OLS Regression Results

```
==============================================================================
Dep. Variable:                    W25   R-squared:                       0.893
Model:                            OLS   Adj. R-squared:                  0.891
Method:                 Least Squares   F-statistic:                     605.2
Date:                Mon, 19 Nov 2018   Prob (F-statistic):               0.00
Time:                        19:57:25   Log-Likelihood:                -2113.4
No. Observations:                 811   AIC:                             4251.
Df Residuals:                     799   BIC:                             4307.
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.5041      0.160      3.151      0.002       0.190       0.818
W9             0.0699      0.034      2.051      0.041       0.003       0.137
W19            0.1717      0.031      5.615      0.000       0.112       0.232
W22            0.0822      0.033      2.518      0.012       0.018       0.146
W23            0.1004      0.034      2.993      0.003       0.035       0.166
W24            0.1347      0.031      4.333      0.000       0.074       0.196
W29           -0.1401      0.037     -3.833      0.000      -0.212      -0.068
W32           -0.1300      0.036     -3.649      0.000      -0.200      -0.060
W33           -0.1213      0.034     -3.582      0.000      -0.188      -0.055
W49            0.2097      0.036      5.845      0.000       0.139       0.280
W50            0.0805      0.035      2.296      0.022       0.012       0.149
W51            0.3711      0.030     12.559      0.000       0.313       0.429
==============================================================================
Omnibus:                       71.079   Durbin-Watson:                   1.927
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              249.859
Skew:                           0.356   Prob(JB):                     5.54e-55
Kurtosis:                       5.624   Cond. No.                         65.5
==============================================================================
```

As I have mentioned before, multiple regression approach specially backward selection method has no chance to give an accurate model with accurate coefficients for a clustering datatype for 2 reasons. Firstly, I can interpret this approach as an attempt to build a Causation model because in backward selection method I omit attributes one-by-

one. According to Prediction vs. Causation in Regression Analysis (n.d.), Causation has a simple methodology against dependent variables correlated with assumed variable. It does not care the complex dynamic of causes and effects among the variables and only track a part of correlation. For instance, I removed the weeks which had the p-value more than 0.05 one by one from top to bottom of table (week 0 to week 51). If I chose the reversed path from week 51 to week 0, the remained weeks after filtering process would be different. It means the path of attribute selections can impact the result. It happens because this method only see a part of data and does not care about integrity of correlation in filtering process (Prediction vs. Causation in Regression Analysis (n.d.)). Therefore, multiple regression on clustering datatype has no effective strategy for the clustering in which a pair of dependent variable are correlated with assumed variable. If one of those is omitted, the regression will conclude the invalid result. Secondly, multicollinearity is another problem. A regression should care about all possible combination of correlated variables to result the "true" coefficients. But it takes a huge process. Therefore, a close approximation of coefficient would be a good strategy to tolerate small differences. In the next section, I will perform clustering analysis by applying classifiers to depict the prediction model schematically and present coefficients and percentage of error for each classifier.

B- Clustering:

I perform clustering to depict a prediction model of sales in week 25. The basis of clustering would be based on classifications. I cluster on whole data set including normalized values to get the coefficients in integer. According to Crawford (2017), I did 7 well-known methods of classifiers; PCA, FastICA, Gauss, KMeans, SparsePCA, SparseRP, Birch, NMF on data set. I perform Clustering on product codes to model week 25 based on other weeks and then I make a prediction model. The output is a graphical scheme of model, predicted Model including coefficients of normalized values in different weeks (week 0 to 51, week 25 excepted), percentage of model error and Root Mean Squared Log Error (RMSLE). It is valuable to add that RMSLE is logarithmic version of RMSE.

$$RMSLE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\Big(log(p_i + 1) - log(a_i + 1)\Big)^2}$$

Where p_i is predicted and a_i is actual quantity. RMSLE is applicable in those model having a huge number of predicted and actual quantities.

The result of Clustering in week 25 in different methods would be in Fig 1 to Fig 7 :

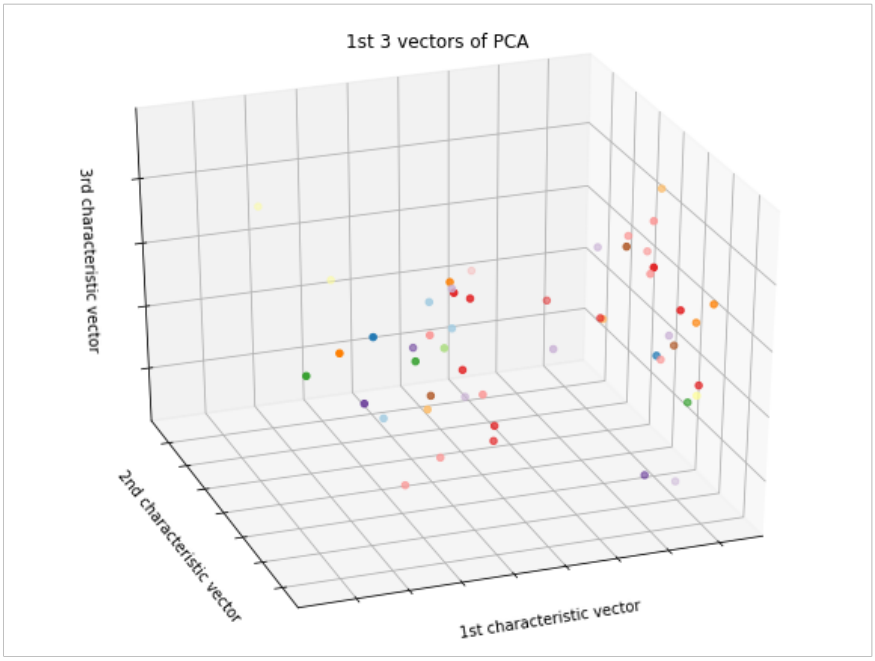

Fig1: PCA model

```
predicted Model [11.62745098 10.62745098 10.62745098 11.62745098 11.62745098 11.62745098
                 11.62745098 12.62745098 12.62745098 11.62745098 11.62745098 10.62745098
                 12.62745098 12.62745098 11.62745098 12.62745098 12.62745098 11.62745098
                 13.62745098 12.62745098 12.62745098 12.62745098 12.62745098 13.62745098
                 15.62745098 14.62745098  9.62745098  9.62745098  9.62745098 10.62745098
                 11.62745098 10.62745098 10.62745098 10.62745098 10.62745098  9.62745098
                 9.62745098 10.62745098 10.62745098 10.62745098 10.62745098 10.62745098
                 11.62745098 11.62745098 11.62745098 11.62745098 11.62745098 12.62745098
                 11.62745098 12.62745098 13.62745098]
PCA Percentage of error: 22.7 Root Mean Squared Logarithmic Error (RMSLE): 0.29032783632164133
```

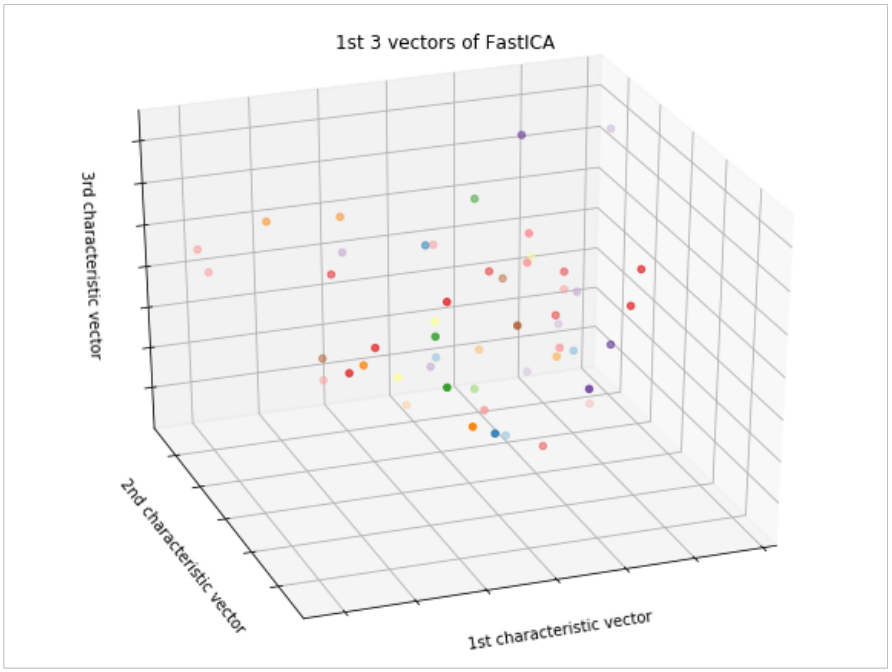

Fig2: FastICA model

```
predicted Model [11.62745098 10.62745098 10.62745098 11.62745098 11.62745098 10.62745098
                 11.62745098 12.62745098 12.62745098 11.62745098 11.62745098 10.62745098
                 12.62745098 11.62745098 11.62745098 12.62745098 12.62745098 11.62745098
                 13.62745098 12.62745098 12.62745098 12.62745098 12.62745098 13.62745098
                 15.62745098 14.62745098  9.62745098  9.62745098  9.62745098 10.62745098
                 10.62745098 10.62745098 10.62745098 10.62745098 10.62745098  9.62745098
                 9.62745098 10.62745098 10.62745098 10.62745098 10.62745098 10.62745098
                 11.62745098 11.62745098 11.62745098 11.62745098 11.62745098 12.62745098
                 11.62745098 12.62745098 13.62745098]
FastICA Percentage of error: 22.7 Root Mean Squared Logarithmic Error (RMSLE): 0.2907674815926446
```

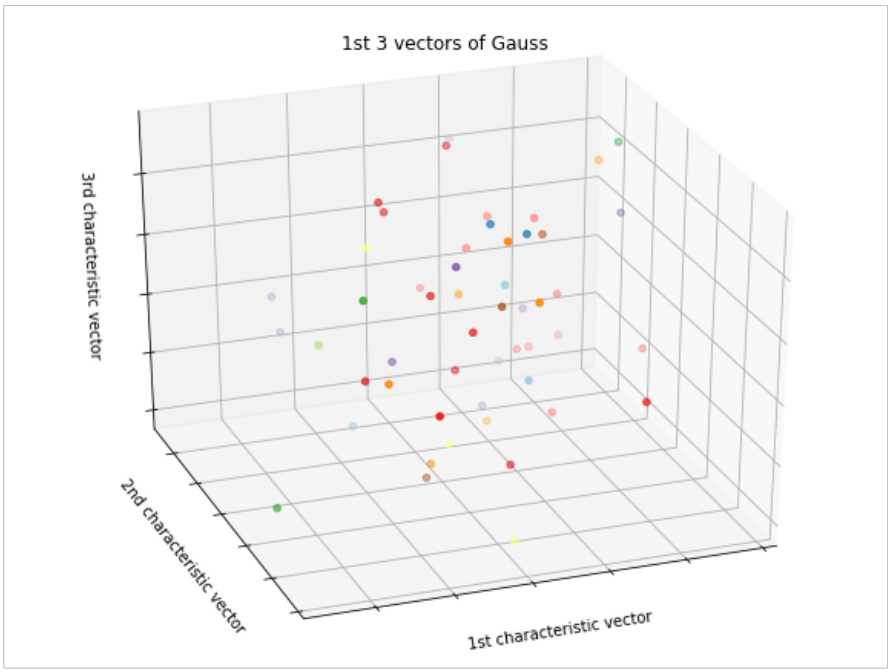

Fig3: Gauss model

```
predicted Model [12.  8. 10. 10. 11. 10. 11. 17. 17. 11. 13. 10. 13. 15. 15. 10. 12. 13.
                 15. 14. 11.  8. 12.  9. 14. 11.  8.  8. 11. 11. 12. 10. 13.  9. 12. 13.
                  9.  8. 12.  8. 12.  9. 10. 10. 13. 11. 12. 10. 10. 12. 11.]
Gauss Percentage of error: 26.6 Root Mean Squared Logarithmic Error (RMSLE): 0.31911398362518656
```

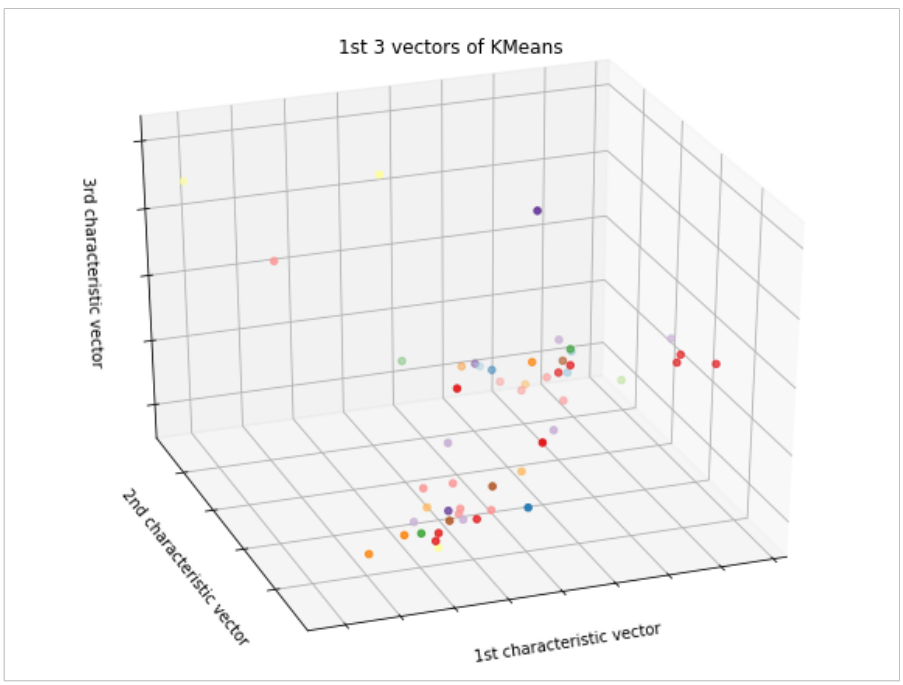

Fig4: KMeans model

```
predicted Model [10. 11. 12. 12. 12. 12. 12. 12. 13. 11. 13. 12. 12. 12. 12. 13. 13. 13.
                 13. 13. 13. 11. 13. 13. 16. 17. 12. 11. 10. 11. 11. 11. 10. 11. 10. 10.
```

```
                 11. 10. 10. 10. 10. 10. 10. 10. 10. 10. 11. 11. 10. 12. 14.]
KMeans Percentage of error: 23.2 Root Mean Squared Logarithmic Error (RMSLE): 0.296136189421982
```

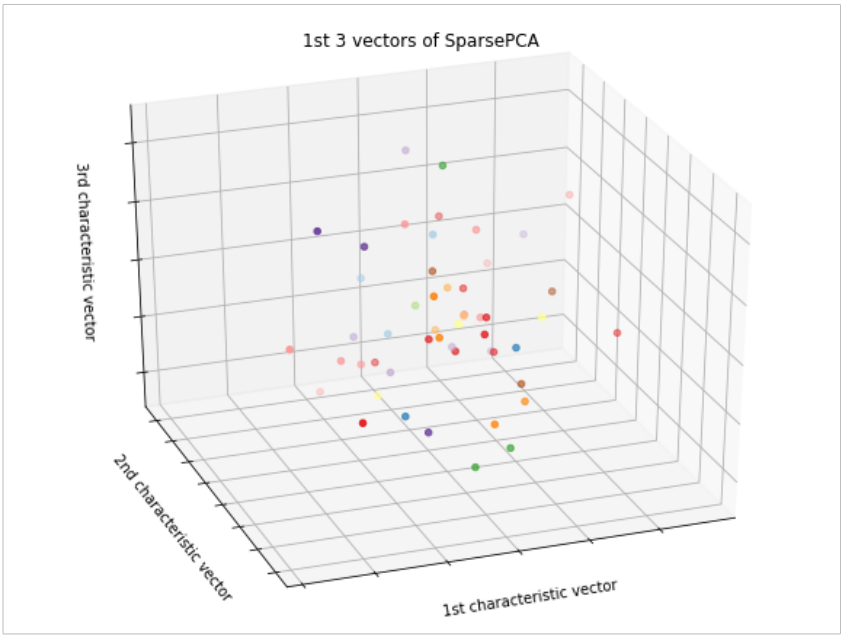

Fig5: SparsePCA model

```
predicted Model [11. 12. 14. 14. 12. 13. 12. 12. 14. 13. 12. 14. 13. 11. 14. 12. 11. 12.
                 10. 12. 16. 13. 15. 10. 14.  9. 10. 10.  9.  9.  9.  9. 11.  9. 12.  8.
                 11. 13.  6. 10.  8. 10.  9. 11. 11. 12. 12.  9. 11. 12. 13.]
SparsePCA Percentage of error: 26.3 Root Mean Squared Logarithmic Error (RMSLE):
0.3229564833563979
```

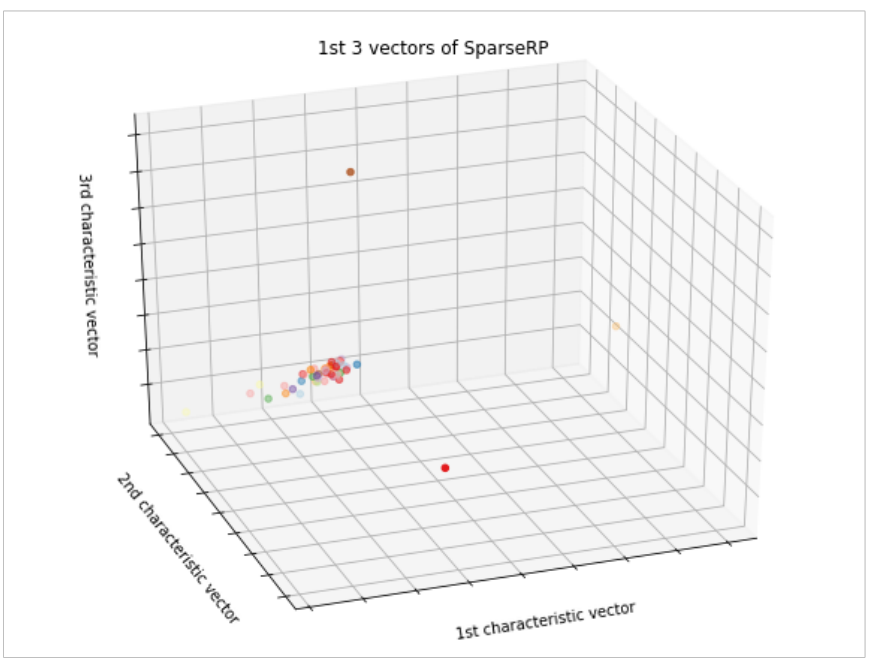

Fig6: SparseRP model

```
predicted Model [15. 12. 11. 17. 10. 18. 11. 16.  8.  6. 11. 10. 18. 10. 14. 13. 13. 12.
```

```
                 14. 10. 10. 11. 14. 11. 10. 17.  8. 13.  6. 16.  9.  4. 12. 10. 19.  9.
                 7.  5. 10.  8. 11.  5. 11. 13. 11. 11. 11. 14. 15. 16. 17.]
SparseRP Percentage of error: 0.0 Root Mean Squared Logarithmic Error (RMSLE):
1.6998793374513743e-15
```

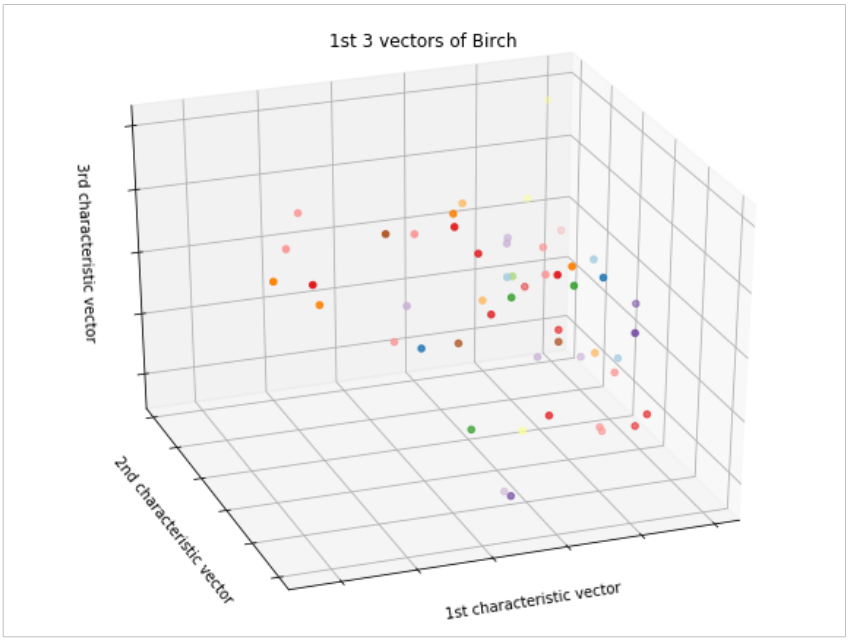

Fig7: Birch model

```
predicted Model [12. 11. 10. 12. 12. 12. 12. 13. 13. 13. 12. 11. 13. 13. 12. 13. 13. 12.
                 14. 13. 13. 13. 13. 14. 16. 13.  9.  8.  9.  9. 10. 10. 10. 10. 10.  9.
                 10. 10. 10. 10. 10. 11. 11. 11. 11. 11. 12. 12. 11. 12. 13.]
Birch Percentage of error: 22.7 Root Mean Squared Logarithmic Error (RMSLE): 0.2894982987879157
```

By comparison of percentage of error and RMSLE in different models, it is clear that SparseRP model with zero percentage of error and least amount of RMSLE is the fitted model of coefficients to predict week 25.

Normalized values {Week 25=
15*W0+12*W1+11*W2+17*W3+10*W4+18*W5+11*W6+16*W7+8*W8+6*W9+11*W10+10
*W11+18*W12+10*W13+14*W14+13*W15+13*W16+12*W17+14*W18+10*W19+10*W20
+11*W21+14*W22+11*W23+10*W24+17*W26+8*W27+13*W28+6*W29+16*W30+9*W31
+4*W32+12*W33+10*W34+19*W35+9*W36+7*W37+5*W38+10*W39+8*W40+11*W41+5
*W42+11*W43+13*W44+11*W45+11*W46+11*W47+14*W48+15*W49+16*W50+17*W51}

Conclusion:

In this project I evaluated descriptive, correlation and regression analysis on Sales quantities. In descriptive analysis, I presented statistical characteristics of data set such as standard deviation, mean, minimum, maximum,1st, 2nd and 3rd quartile values and depicted them in box-plot. In addition, I performed correlation analysis on all weeks. In addition, I did ANOVA function to determine the weeks which were statistically significant with week 25. Finally, I did regression analysis in 2 steps; first step included multiple regression which was based on backward selection method. The resulted coefficients were not accurate because of the limitation which this method had in omitting attributes. The multiple regression generated different equations based on omitting path. In the second part, I performed clustering in 7 methods of classifications to model sales in week 25. SparseRP method had the best result of regression model with 0 percentage of error.

Reference:

Tan, Chuan , Lau & Pei (2014). Time series clustering: A superior alternative for market basket analysis. Retrieved from https://archive.ics.uci.edu/ml/datasets/Sales_Transactions_Dataset_Weekly

Crawford, C. (2017). Weekly Sales Transactions. Retrieved from https://www.kaggle.com/crawford/weekly-sales-transactions#Sales_Transactions_Dataset_Weekly.csv

Prediction vs. Causation in Regression Analysis. (n.d.). Retrieved from https://statisticalhorizons.com/prediction-vs-causation-in-regression-analysis