Predictive Monitoring and Optimization of Transport and Logistic Processes:

Cargo 2000 as Case Study

University of L'Aquila

Data Analytics And Data Driven Decision Making Project submitted to :

Prof. Fabrizio Rossi and Prof. Felici Giovanni

Submitted by :

1. Eze Grace Anulika: 254872
2. Prince Amoateng: 255090
3. Emmanuel Osei Addae: 255103
4. Ogala Wisdom Ifeanyi: 255093

On the 13th of July, 2018

MOTIVATION

The movement of goods and services from place to place has been of ageless economic importance to both developing and developed nations. Our quest is to predict and optimize the transportation processes

The economic importance of these are:

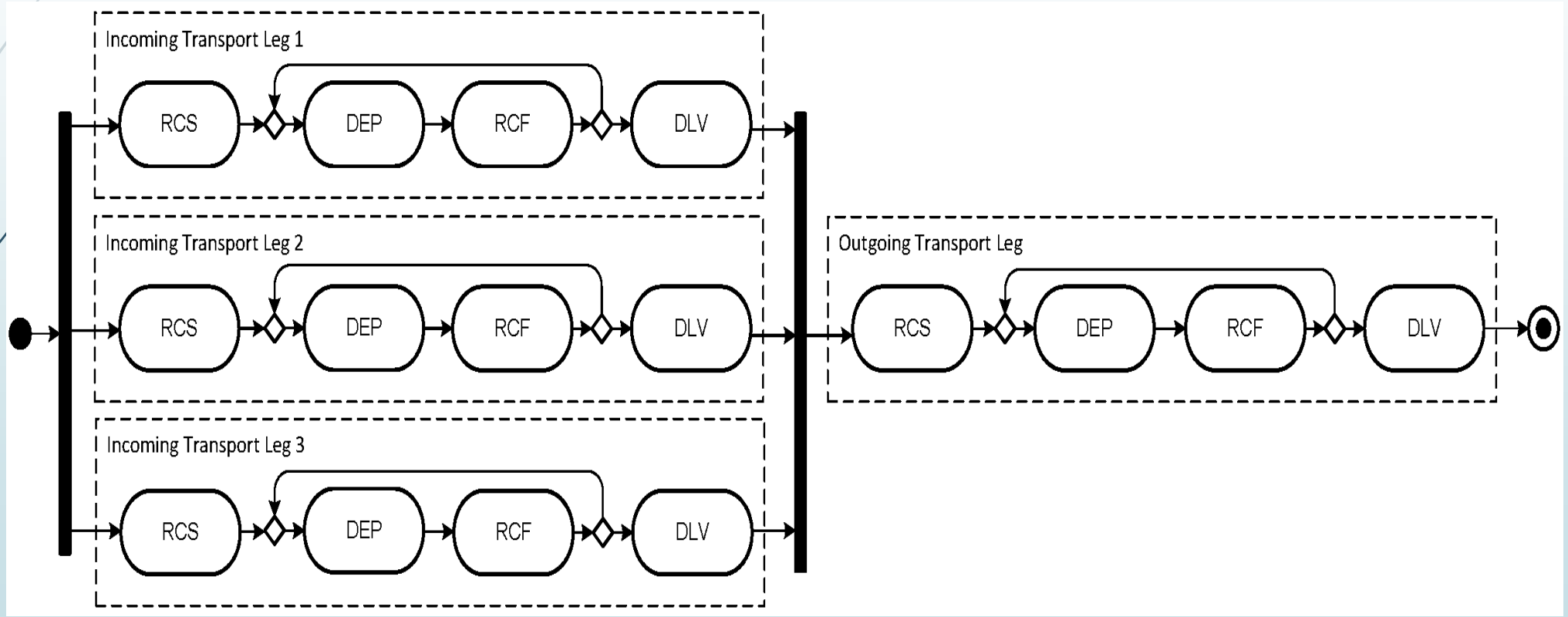
- To salvage up to 15 % of annual expenditure
- To reduce the (Carbondioxide) emission accountable to the transportation industry.
- To predict the delivery time for a particular freight transportation process.

INTRODUCTION: WHY THIS DATA SET?

- In this work, we chose the IATA Cargo 2000 data set obtained from S-CUBE as case studies for our analysis.
- This data set shows the ideal processes involved in a typical freight forwarding cargo company which is most suitable for our analysis.
- Recent papers on this topic, such as "**Comparing and combining predictive business process monitoring techniques**" published by A. Metzger et al, have drawn their analysis for this same data set.
- We seek to use this data set to develop an optimal prediction model to predict the outcome of a planned logistic process.

DESCRIPTION OF THE DATA SET

The diagram below shows a model of the business processes covered in the case study.



DEFINITION OF TERMS IN THE DATA SET

Each of the transport legs involves the following physical transport services:

- **RCS:** Check in freight at departure airline. Shipment is checked in and a receipt is produced at departure airport.
- **DEP:** Confirm goods on board. Aircraft has departed with shipment on board.
- **RCF:** Accept freight at arrival airline. Shipment is checked in according to the documents and stored at arrival warehouse.
- **DLV:** Deliver freight. Receipt of shipment was signed at destination airport.

Content of the Data Set

- Our data set consist of
- execution traces of 3,942 actual business process instances,
- comprising 7,932 transport legs and
- 56,082 service indications.
- Each execution trace includes planned and effective durations (in minutes) for each of the services of the business process

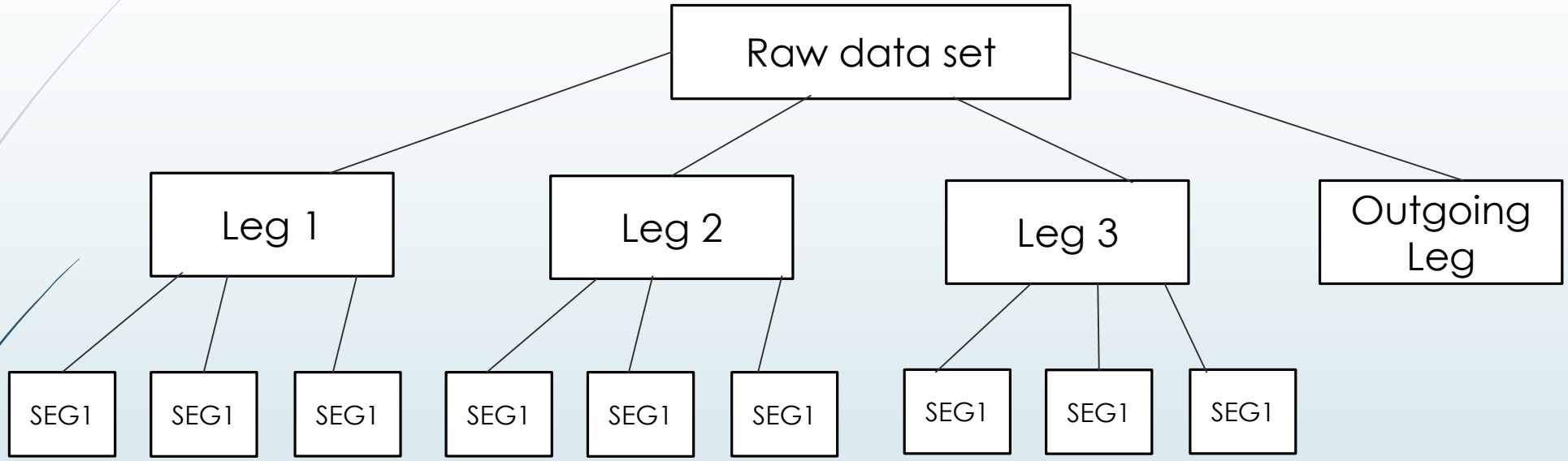
DATA CLEANING

- Our raw data set looks like this

	nr	i1_legid	i1_rcs_p	i1_rcs_e	i1_dep_1_p	i1_dep_1_e	i1_dep_1_place	i1_rcf_1_p	i1_rcf_1_e	i1_rcf_1_place	...	o_dep_3_p	o_dep_3_e	o_dep_3_place
0	0.0	5182.0	199.0	218.0	210.0	215.0	609.0	935.0	736.0	256.0	...	?	?	?
1	1.0	6523.0	844.0	584.0	90.0	297.0	700.0	1935.0	1415.0	431.0	...	?	?	?
2	2.0	5878.0	4380.0	4119.0	90.0	280.0	456.0	905.0	547.0	700.0	...	?	?	?
3	3.0	1275.0	759.0	169.0	240.0	777.0	173.0	340.0	577.0	349.0	...	?	?	?
4	4.0	8117.0	1597.0	1485.0	150.0	241.0	411.0	585.0	612.0	128.0	...	?	?	?

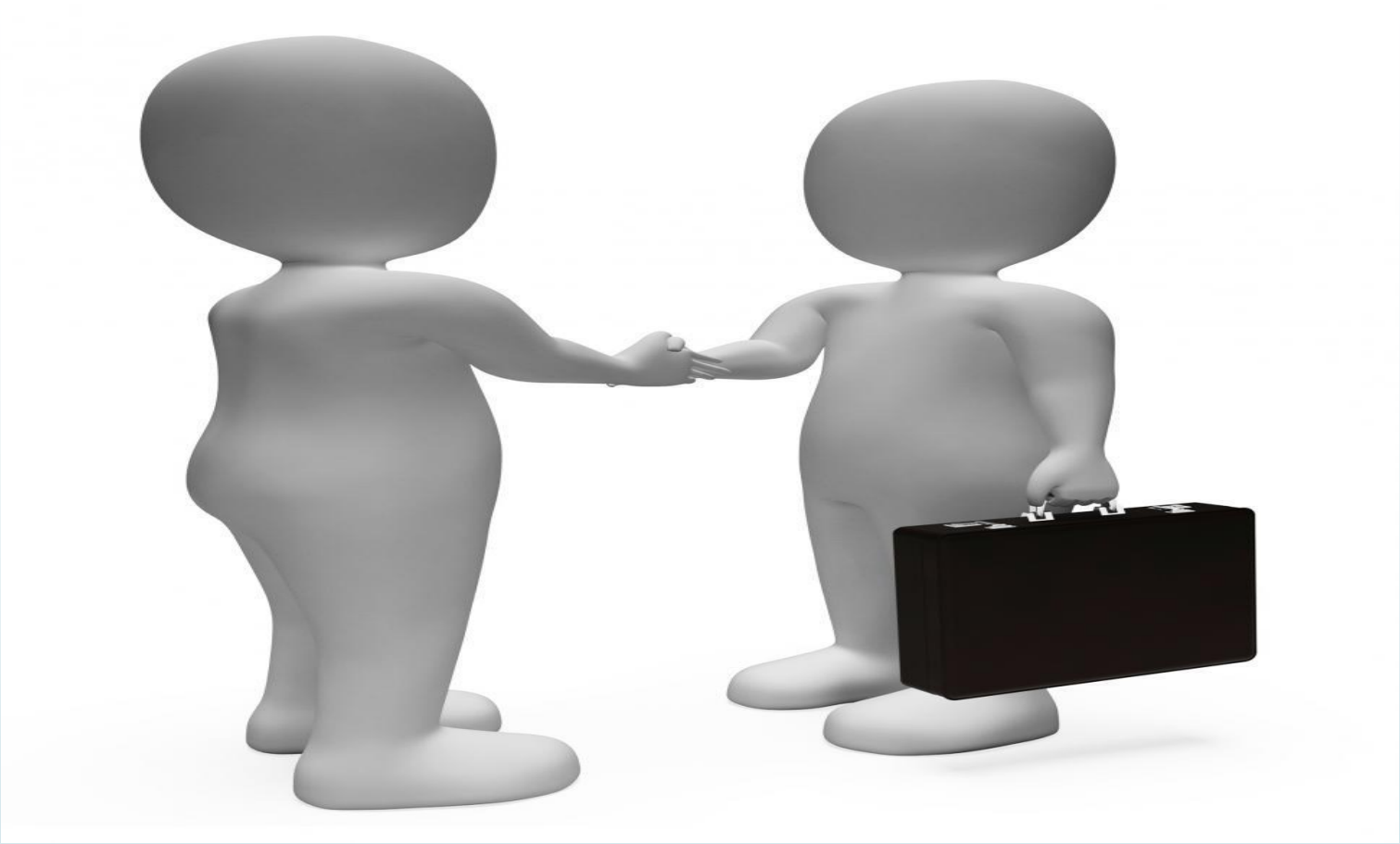
Customers whose goods did not pass through a segment or leg results into “?”

- We applied method of **DATA CLASSIFICATION** to classify the data set into categories as shown in the diagram below:



We use the segment attractor python code to split the Legs into segments

```
def segment_extractor(name, start, end):
    col_header = list(name)
    first_seg = name[col_header[start: end]]
    return first_seg
```



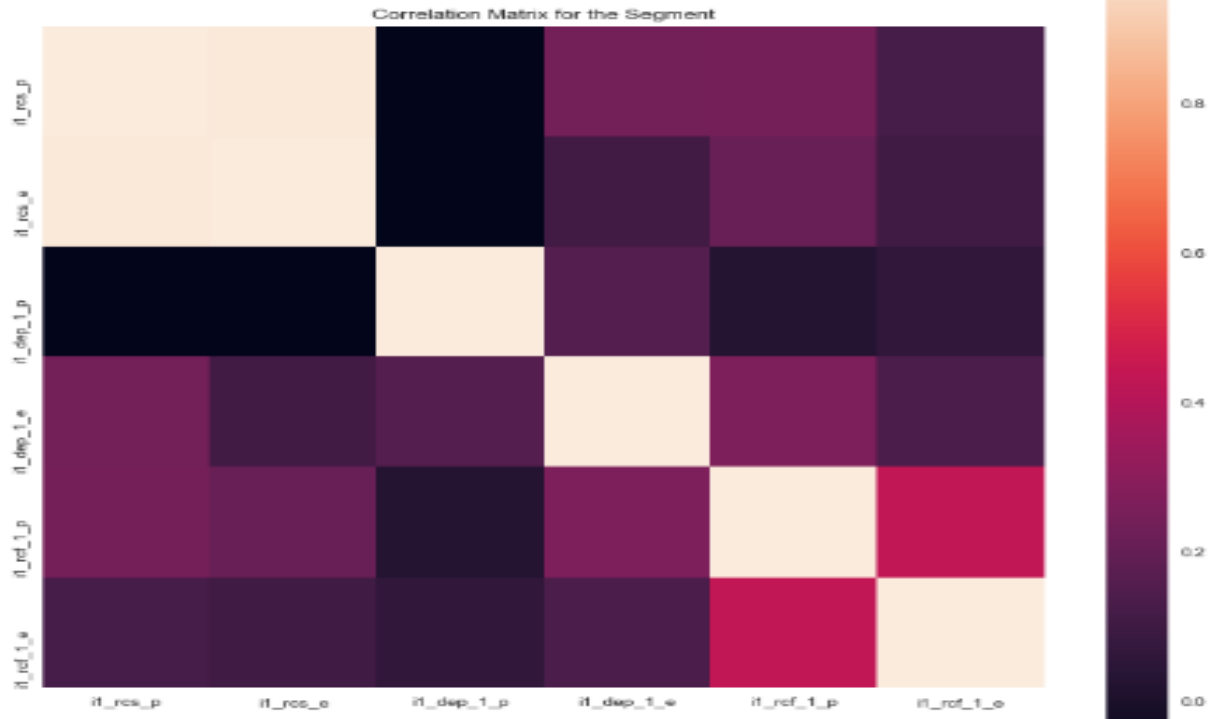
No missing data after the Classification into categories

Data Classification worked perfectly

Exploratory Analysis

- We ran exploratory analysis for the segments of each legs
- We drop all columns with header "_place" and "i.d's" of the Airports, since they are not occurrence.
- We concentrated our analysis only on **Leg 1 (Segment 1, 2, 3)** which is our Leg of interest for this work
- The Legs are independent so we can use one Leg to generalize for the entire processes

Correlation Matrix for Seg. 1, Leg 1



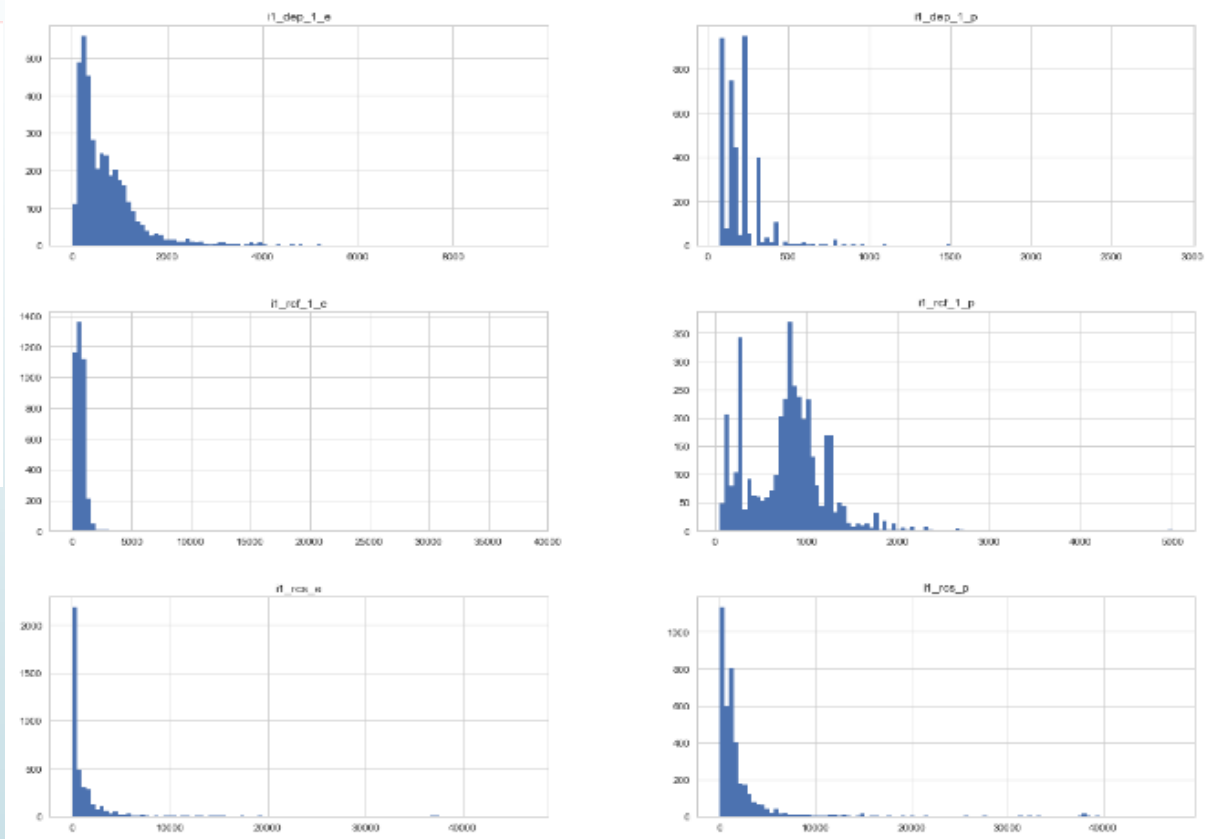

```
i1_rcs_e    i1_rcs_p      0.988858
i1_rcf_1_e  i1_rcf_1_p    0.434936
i1_rcf_1_p  i1_dep_1_e    0.266012
dtype: float64
```

In an ideal setting, we expect the planned time to be strongly correlated with the execution time in general, but this segment reveal a total contradiction to this fact. This is expected because the segments (including this) of the data set is a real life situation and the data set was collected independently. Here, only "i1_rcs_e" and "i1_rcs_p" shows strong correlation, that is an increase or decrease in planned time leads to an increase or decrease in execution time

Descriptive Statistics and Histogram for Seg 1, Leg 1

	l1_rcs_p	l1_rcs_e	l1_dep_1_p	l1_dep_1_e	l1_rcf_1_p	l1_rcf_1_e
count	3942.000000	3942.000000	3942.000000	3942.000000	3942.000000	3942.000000
mean	2203.533486	1653.323440	205.891933	711.408929	796.002790	666.358447
std	4683.333105	4553.855588	140.283173	684.789184	439.991579	764.145906
min	5.000000	1.000000	75.000000	4.000000	50.000000	13.000000
25%	377.750000	113.000000	120.000000	263.000000	455.000000	274.000000
50%	1085.000000	340.000000	180.000000	516.000000	820.000000	657.500000
75%	1946.500000	1375.000000	240.000000	949.750000	1020.000000	883.000000
max	47190.000000	46357.000000	2876.000000	9513.000000	5001.000000	38116.000000

For this segment, the variable that exhibits the highest std is the "l1_rcs_p" showing that the occurences show high variation or spread from the mean value. It has the maximum time, highest mean time, which means that on the average, this section of the segment takes more time to come into completion than the others. While the "l1_dep_1_p" has the minimum std and mean showing less variation than others.



From the Histogram plots above, we observe that the data values are skewed to the right for each of the processes. Also, we do not observe lots of extreme outliers in the processes which could have an important effect on our analysis

Summary for Seg. 2 and Seg 3.

- For Seg2. the variable that exhibits the highest std is the "i1_dep_2_e" showing that the occurences show high variation or spread from the mean value 1018.903766.
- For Seg.2 "i1_dep_2_e" and "i1_dep_2_p" shows strong correlation as well as "i1_rcf_2_e" and "i1_rcf_2_p“
- For Seg.3 "i1_dep_3_e" has the highest standard deviation from the mean value 978
- For Seg.3. “i1_dep_3_e” and “ i1_dep_3_p “ shows the largest correlation

Supervised Learning on Leg 1 (Seg. 1)

- We carried out a **Logistic Regression** on Leg 1, Seg. 1
- We constructed a binomial distribution on our data set by comparing planned time and execution time – **on-time checker**
- We used the binomial distribution of the on-time checker as a independent variable and the total delivery on-time checker as a dependent variable
- We applied the Logistic regression algorithm on the data set consisting of 3941 occurrences
- The algorithm auto splitted the data set into 2956 training data and 986 testing data
- And obtained 88% accuracy rate

Confusion Matrix for Seg.1

	Prediction:No	Prediction: Yes
Actual: No	TN = 78	FP = 27
Actual: Yes	FN = 92	TP = 789
	Total: 170	**816**

FP=False Positive
(type 1 – error)

FN=False Negative
(type 2 – error)

We applied the same algorithm and same process to Seg. 2 and obtained the following:

	Prediction:No	Prediction: Yes
Actual: No	TN = 52	FP = 87
Actual: Yes	FN = 10	TP = 150

Accuracy rate: 68%
Total Data: 1195
Training Data: 896
Testing Data: 299

Linear Regression Analysis on Leg 1. Seg1

- Y = a0 + b1*x1 +b2*x2 +b3*x3 +Residuals
- We applied an OLS(Ordinary Least Square) regression on Leg 1, Seg 1 and obtained the table:

OLS Regression Results

Dep. Variable:	total_dlv_e	R-squared:	0.978
Model:	OLS	Adj. R-squared:	0.978
Method:	Least Squares	F-statistic:	5.931e+04
Date:	Thu, 12 Jul 2018	Prob (F-statistic):	0.00
Time:	10:31:25	Log-Likelihood:	-31477.
No. Observations:	3942	AIC:	6.296e+04
Df Residuals:	3938	BIC:	6.299e+04
Df Model:	3		
Covariance Type:	nonrobust		

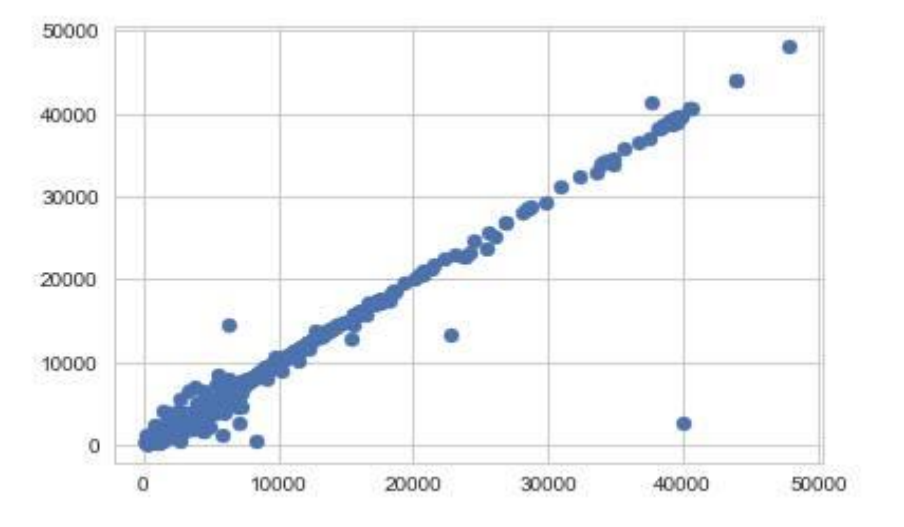
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-32.8722	28.479	-1.154	0.248	-88.707	22.962
i1_rcs_p	1.0025	0.003	400.754	0.000	0.998	1.007
i1_dep_1_p	1.2160	0.081	15.001	0.000	1.057	1.375
i1_rcf_1_p	0.7594	0.027	28.598	0.000	0.707	0.811

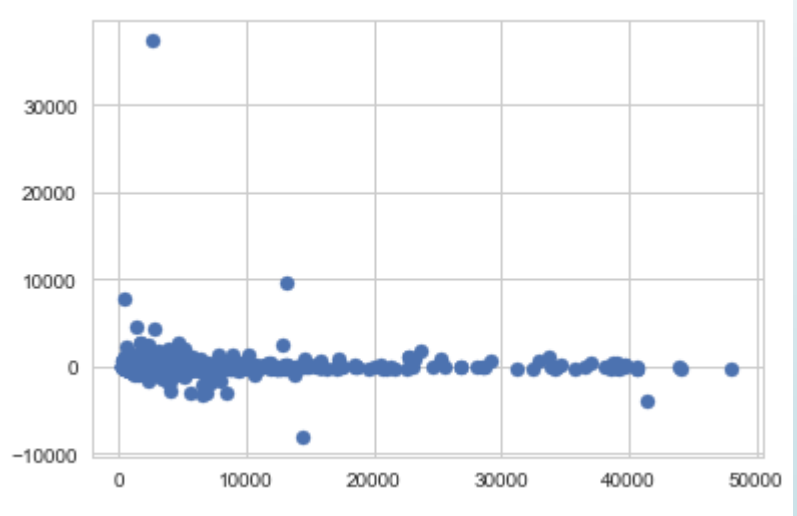
Omnibus:	11272.209	Durbin-Watson:	1.997
Prob(Omnibus):	0.000	Jarque-Bera (JB):	631609309.175
Skew:	37.673	Prob(JB):	0.00
Kurtosis:	1962.525	Cond. No.	1.31e+04

- We obtained the following model equation:

$$total\_dlv\_e = -32.8722 + 1.0025 \times (i1\_rcs\_p) + 1.2160 \times (i1\_dep\_1\_p) + 0.7594 \times (i1\_rcf\_1\_p)$$

- The plot of **y against ypredict –left and ypredict against err –Right** is shown below:





Linear Regression Analysis on Leg 1, seg. 2

- Y = a0 + b1*x1 +b2*x2 +Residuals
- We applied an OLS(Ordinary Least Square) regression on Leg 1, Seg 2 and obtained the table:

OLS Regression Results

Dep. Variable:	total_dlv_e_2	R-squared:	0.778
Model:	OLS	Adj. R-squared:	0.778
Method:	Least Squares	F-statistic:	2090.
Date:	Thu, 12 Jul 2018	Prob (F-statistic):	0.00
Time:	10:32:10	Log-Likelihood:	-9190.2
No. Observations:	1195	AIC:	1.839e+04
Df Residuals:	1192	BIC:	1.840e+04
Df Model:	2		
Covariance Type:	nonrobust		

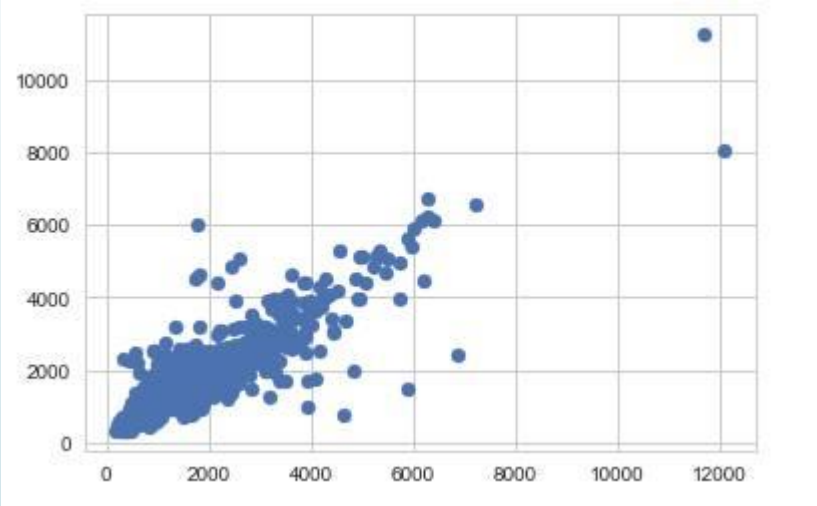
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	216.2644	30.274	7.144	0.000	156.868	275.661
i1_dep_2_p	0.9709	0.018	54.763	0.000	0.936	1.006
i1_rcf_2_p	0.7127	0.026	27.459	0.000	0.662	0.764

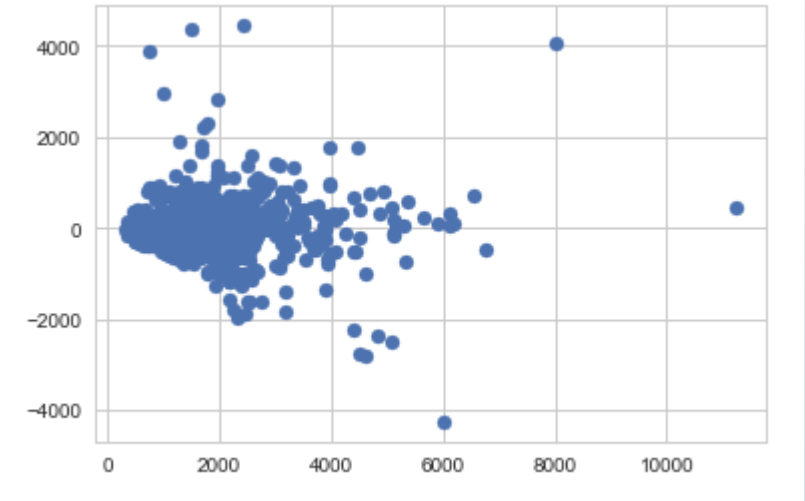
Omnibus:	499.019	Durbin-Watson:	2.030
Prob(Omnibus):	0.000	Jarque-Bera (JB):	22011.013
Skew:	1.190	Prob(JB):	0.00
Kurtosis:	23.890	Cond. No.	2.94e+03

- We obtained the following model equation:

total_dlv_e_2 = 216.2644 + 0.9709*(i1_dep_2_p) + 0.7127*(i1_rcf_2_p)

- The plot of **y against ypredict –left and ypredict against err –Right** is shown below:

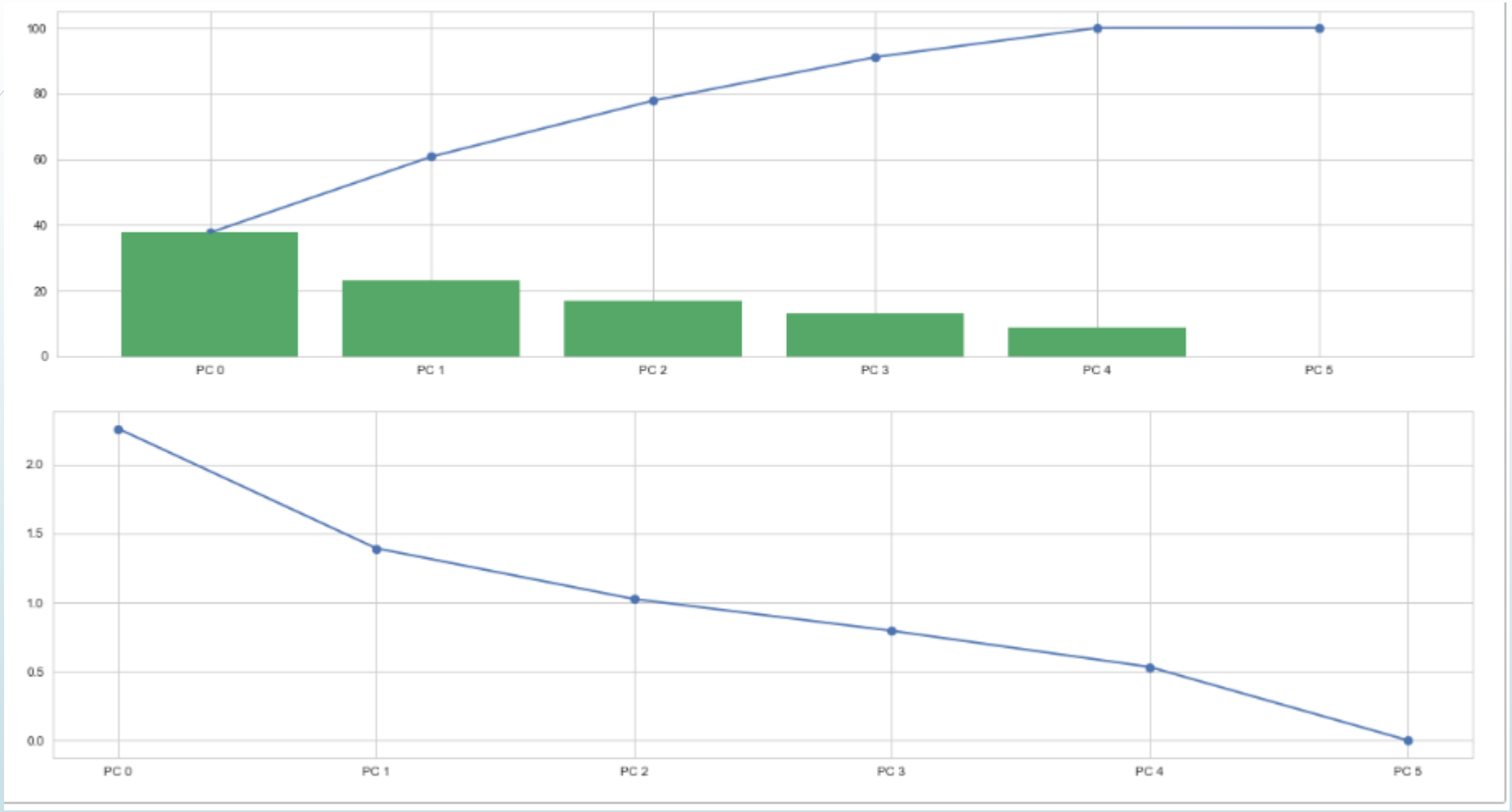




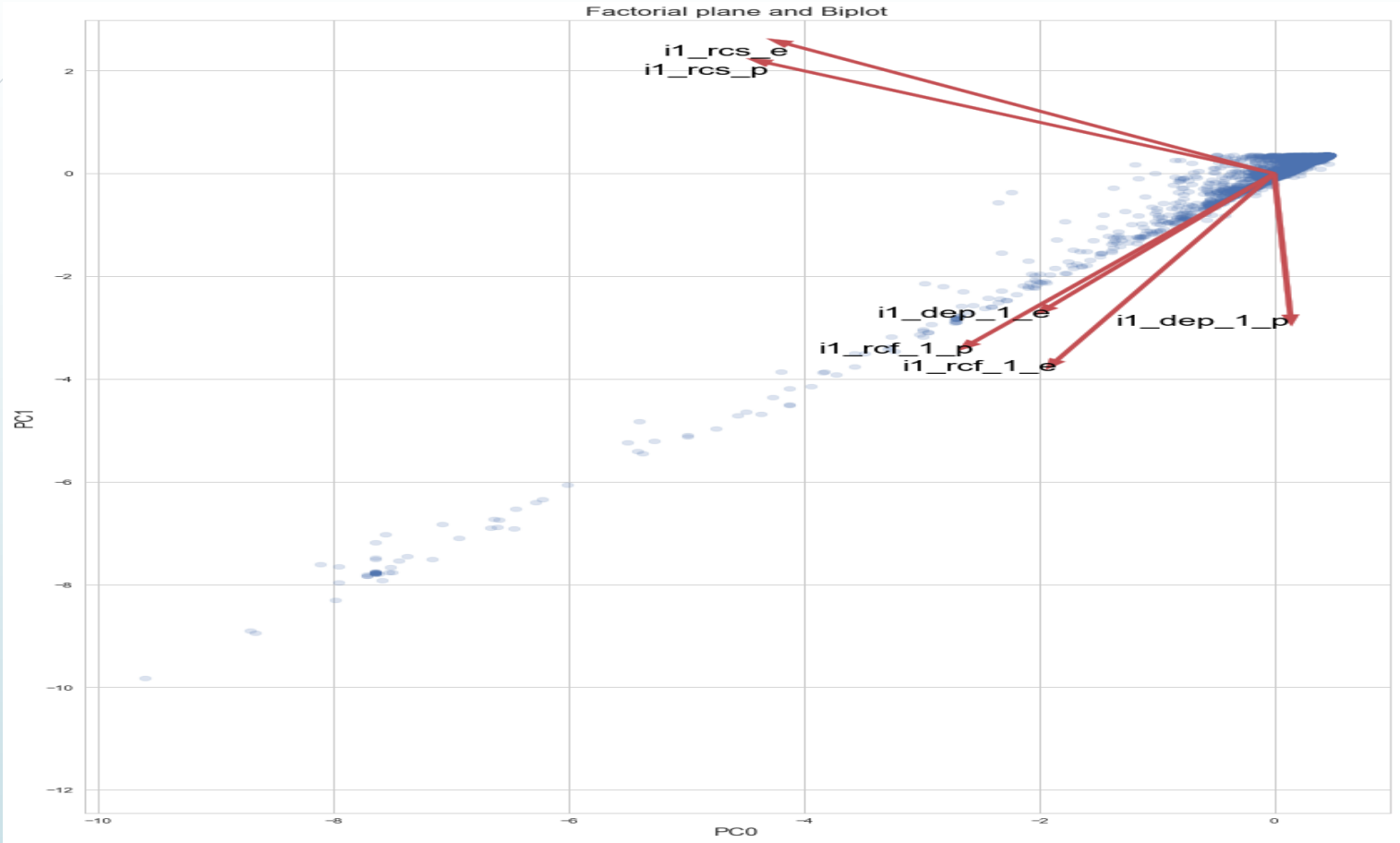
Unsupervised Learning on Leg 1, Seg. 1

- We applied **Principal Component Analysis (PCA)** and **K-Means clustering** algorithm to study Leg 1, Seg. 1
- Principal Component Analysis allows us to find the directions of maximum variance in high-dimensional data and project it onto a smaller dimensional subspace while retaining most of the information.
- This is achieved via the highest eigenvalues in the covariance matrix

- The Variance Explained diagram

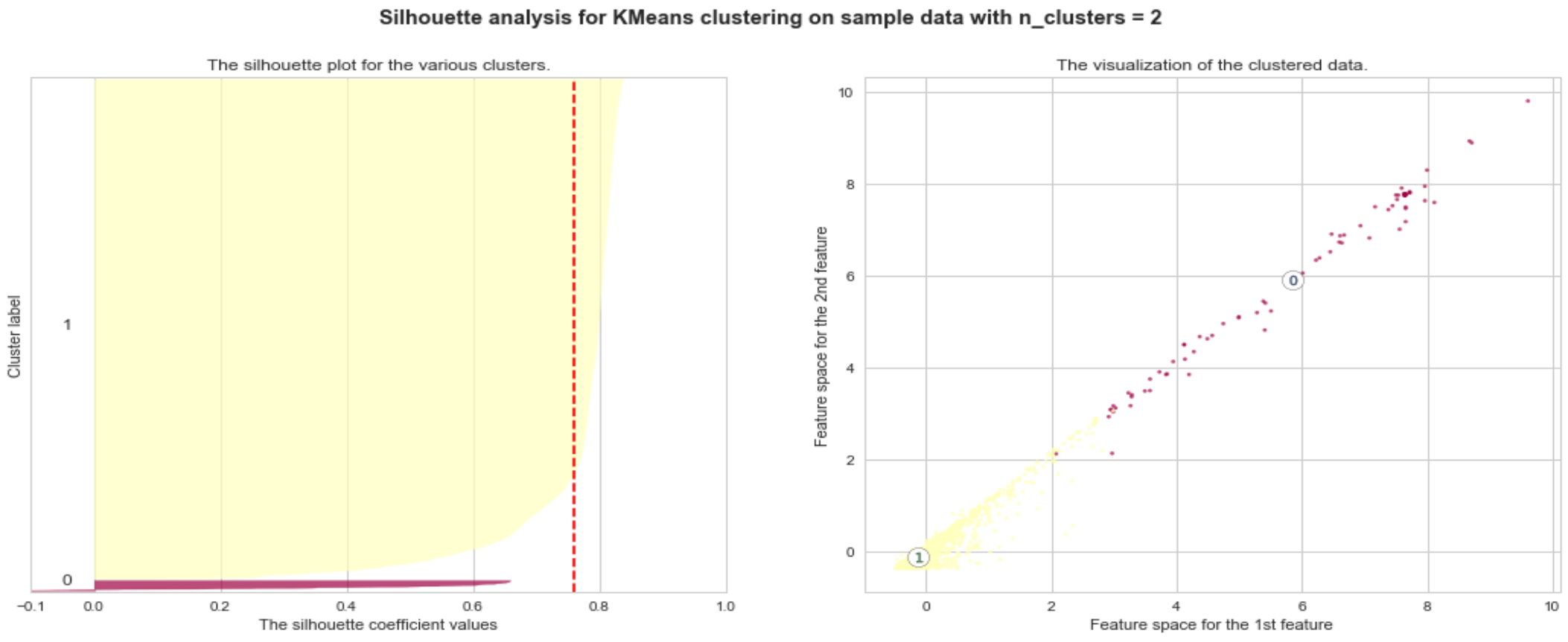


- The Biplot is shown below



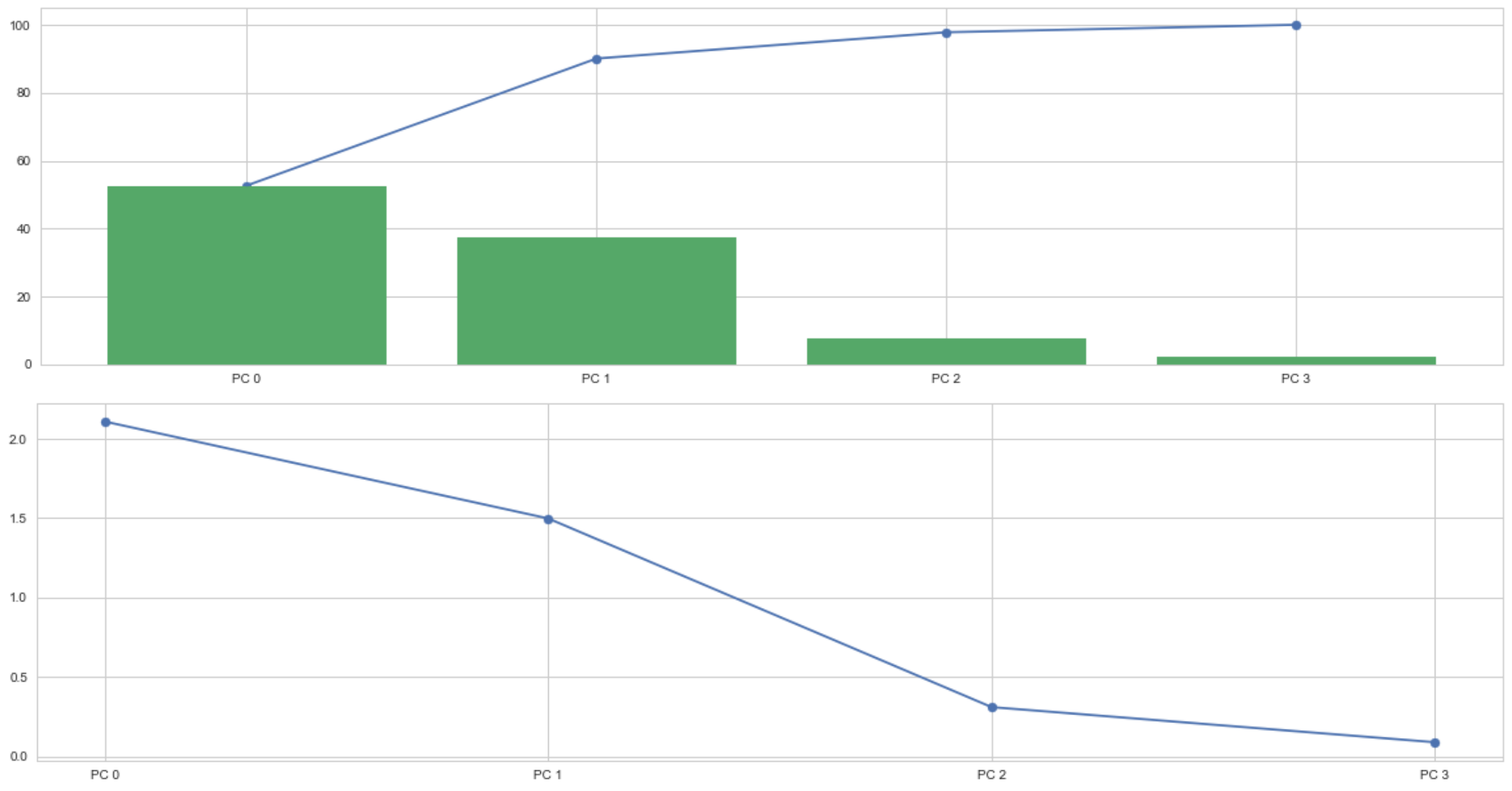
- we can say that i1_rcs_p and i1_rcs_e has greater influence on the Logistic process for the first segment of Leg 1 because they higher score in pc0

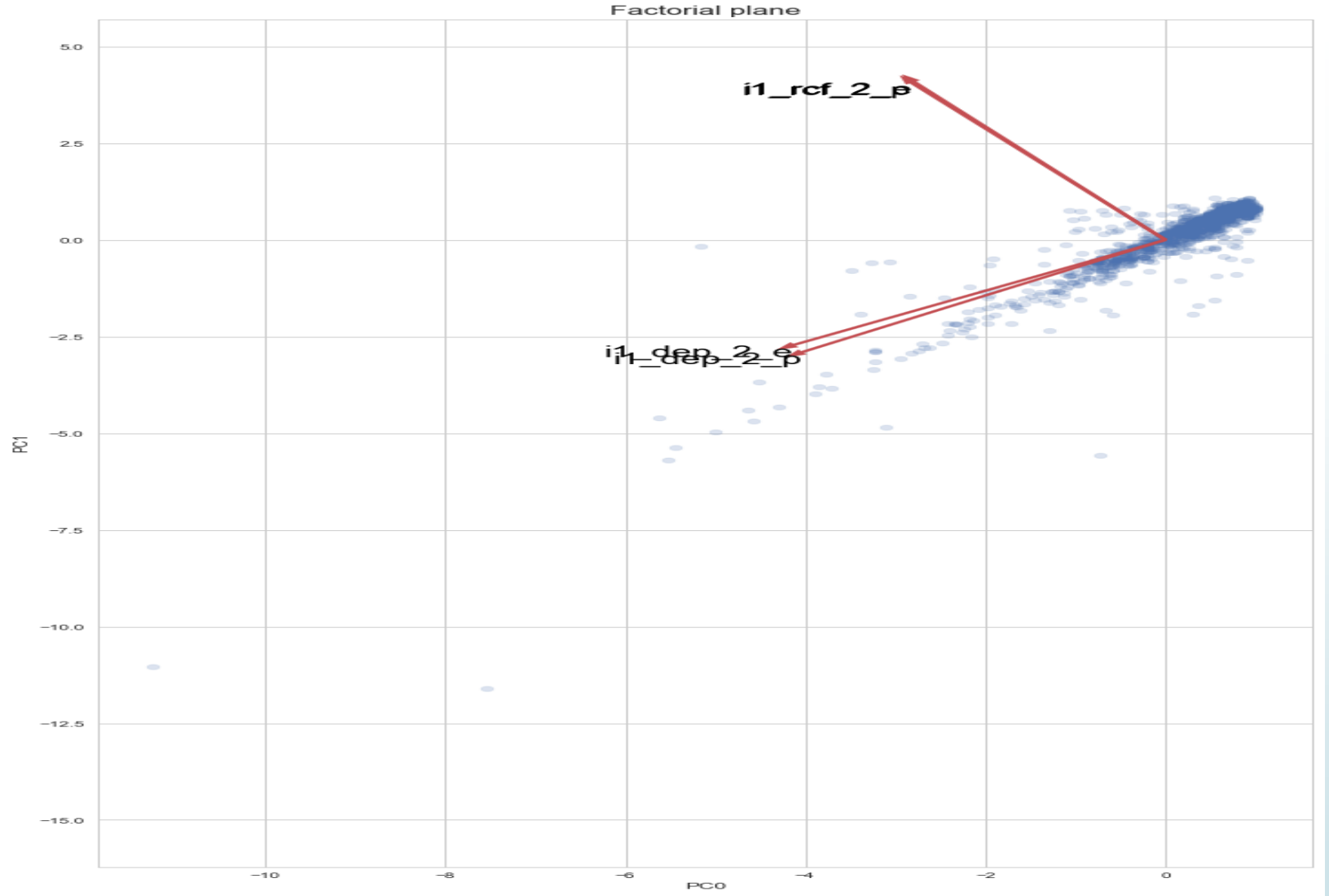
K-Means Clustering of Leg 1, Seg. 1



This clustering can be used to confirm certain business assumption and to understand which execution the Freight Company should focus on

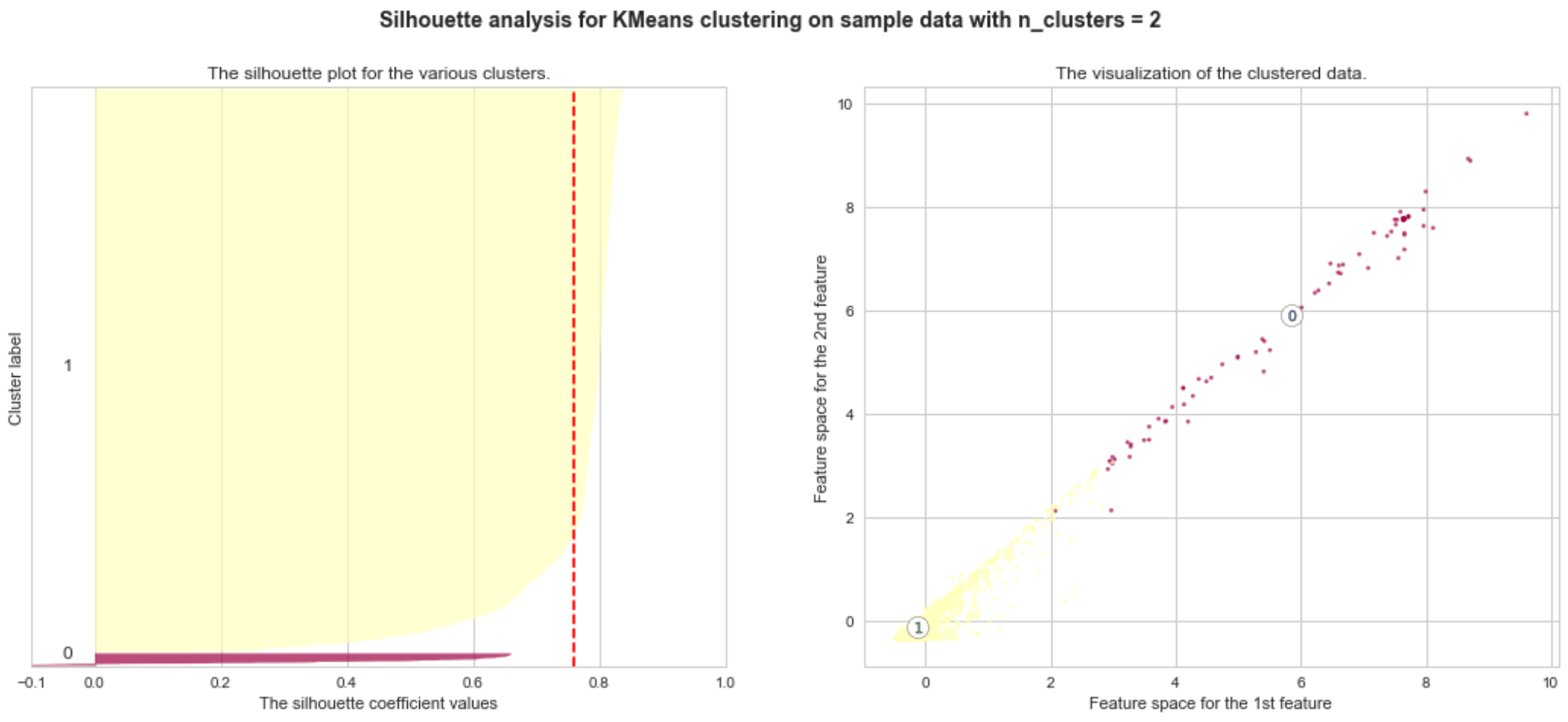
Unsupervised Learning on Leg1, Seg. 2





PCO explain more of the data set than PC1, we conclude that i1_dep_2_e is the most important process in the segment.

Silhouette Plot and K-Means Clustering



Third Segment Leg 1, Using Random Decision Forest (Ensembling Learning)

- Aim: Condsidering that the data set size for the third segment of Leg 1 is very small (23 occurence only), we shall result to an alternative learning modelling called "Ensembling Learning".
- Ensembles combine multiple hypotheses to form a (hopefully) better hypothesis
- This uses classification or regression approach
- In this work, we use the classification approach and construct a binomial structure from the data set.
- We trained with 60% of 23 and test with 40% of 23 using 10,000 estimator (randomization)
- We obtained a varying accuracy score and unreliable model because the data set is insufficient for the learning process.

Conclusion

- The linear regression analysis that we carried out on Leg 1, help us to develop a linear model which when given the planned time for each section of the processes involved can predict accurately the execution time, with an error of 10^{-6} and 10^{-11}respectively.
- From our study of Leg 1 using Logistic Regression Analysis, we developed a model that; when giving the scheduled time and execution time, could predict accurately if the planned time is achieved or not.
- With this Logistic regression we obtained an accuracy of 88% for Segment 1 and 68% for Segment 2 of Leg 1 respectively.
- In seg. 3, small data size resulted in an unreliable model
- The unsupervised learning carried out leg 1, show us that the most significant process in segment 1 is "i1_rcs_p and i1_rcs_e ", while the most significant process in segment 2 is " i1_dep_2_e ",

Deduction From Conclusion

- The freight forwarding company can increase it's efficiency by using the developed model to predict her execution time for planned processes
- The freight forwarding company should focus on optimizing those processes highlight as been significant by the unsupervised learning processes.


FRAGILE
Thank you