Key findings

- On average, customers who use phones to purchase will generate more revenue compared to those who purchase on the website. Besides, customers who use phones to place orders in cluster 4 generate higher revenue than those in cluster 2. Therefore, we should conduct more campaigns on the phone channel, especially for segments that generate more profits.
- The number of orders has a negative effect on revenue in Cluster 2. It means that the more orders a customer places, the less revenue he or she brings to us. Therefore, we should encourage customers to buy more products in one order, such as setting a threshold for free shipping.

Technical Report

This report is aimed to find out regression models to describe which variables have effect on the total revenue each customer generated, how significantly each variable contributes to this model and especially focused on the differences of models between two segments we found in project 2. Our team generally followed the process as 1) transform dataset, 2) data exploratory (come up with rough models and decide on two clusters), 3) validate models in two clusters.

First, since the former dataset doesn't consist of all the variables we need for project 3, we aggregated the data including cost, product quantity, orders quantity, cancel/return quantity, channel, and payment method. Thus, our team came up with a new data file, trying to find out the effect of these variables (dependent variables) on revenue (independent variables).

Then, since payment method and channel are categorical variables, we created dummy variables for them. For payment methods, we chose other payment methods as the basic variable and code other methods like Visa, America Express, Discover, Master Card, personal check,

PayPal into new values. For Channel, we chose web as our basic value, and create two new variables for phone and mail.

Then our team started to do the data exploratory. With a completed data file, we continued to run the regression model in seven clusters and aimed to find out something interesting. We chose to use Stepwise method to get a clear view on the significance of each variable on the model in different segments. After this process, we found that there are significant differences between two segments, one of which is in the medium-low profit group (Cluster 2) and another is in the low profit group (Cluster 4), and the rough models of these two clusters provided us with delighted ideas. Thus, we decided these two clusters to further develop our findings.

Next, we started to conduct regression in Cluster 2. To make sure that dependent and independent variables have a linear relationship, we checked the mean of residual, which should be 0.00, observed the Residuals Statistics, Normal Probability Plot and Residual Plot to verify the homoscedasticity. Another thing to look into is that VIF should be lower than 4, avoiding multicollinearity problem. According to the Coefficient table below, VIF of quantity and product number is larger than 4. Since quantity value are more significant to explain the data, so we first deleted the product number and reran the regression to see whether the value will change.

TABLE 1. COEIFFICENT TABLE (LINEAR REGRESSION) IN THE CALIVRATION GROUP OF CLUSTER TWO (BEFORE CHECK THE OUTLIERS)

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	76.014	2.594		29.299	0		
COST	1.679	0.007	0.907	241.209	0	0.883	1.133
QUANTITY	5.192	0.666	0.082	7.794	0	0.114	8.776
MONTHS	-0.382	0.033	-0.042	-11.733	0	0.985	1.015
ORDER_NO	-4.068	0.686	-0.023	-5.932	0	0.821	1.218
PHONE	7.553	1.269	0.022	5.952	0	0.954	1.048
PRODUCT_NO	2.566	0.734	0.037	3.495	0	0.112	8.968

Moreover, we deleted some outlier in our model and reran the regression. If the Mahalanobis value of one data point is greater than 0.02, we deleted it to prevent the inaccuracy.

In the end, we found that cost is very significant to revenue in all cluster, therefore we should consider it as the most important one, and here are two regression models for segment 2 and segment 4.

- Cluster 2:

  Revenue=76.612+1.680*COST+7.602*QUANTITY-0.384*MONTHS-4.040*ORDER_NO+6.705*PHONE

- Cluster 4:

  Revenue=215.010+1.627*COST+7.573*QUANTITY-0.783*MONTHS-25.332*PHONE

After data exploratory, we continued to validate two models in these two segments. First, our team broke our sample into two groups and used 60 percent to randomly select data as calibration samples, while the rest samples are validation samples.

Then, we selected samples both in the calibration group and also in Cluster 2. We ran the linear regression process of calibration samples in the Cluster 2 with Stepwise method. Based on the ANOVA table (see *Table.2*) and coefficient table (see *Table.3*), we found out that the cost, the quantity of products, the months after last purchase, the number of orders and the channel to purchase all have linear relationship with revenue.

TABLE 2. ANOVA TABLE (LINEAR REGRESSION) IN THE CALIVRATION GROUP OF CLUSTER TWO (BEFORE CHECK THE OUTLIERS)

	Sum of Squares	df	Mean Square	F	Sig.
Regression	133504162	5	26700832.4	9010.66	.000
Residual	14961445.5	5049	2963.249		
Total	148465608	5054			

TABLE 3. COEFFICENETS (LINEAR REGRESSION) IN THE CALIVRATION GROUP OF CLUSTER TWO (BEFORE CHECK THE OUTLIERS)

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	75.107	3.27		22.968	0		
COST	1.68	0.009	0.906	190.424	0	0.883	1.133
QUANTITY	8.043	0.326	0.125	24.708	0	0.781	1.28
MONTHS	-0.376	0.041	-0.041	-9.144	0	0.99	1.01
ORDER_NO	-4.663	0.856	-0.026	-5.445	0	0.855	1.17
PHONE	7.642	1.597	0.022	4.786	0	0.981	1.019

Our team also ran Collinearity diagnostics, checked the value of VIF (see *Table 1.2*) and found that there is no multicollinearity problem. Besides, our team also checked the Residuals Statistics, drew Normal Probability Plot and Residual Plot to verify the homoscedasticity between dependent variables and independent variables and finally verified the linear relationship.

Next, our team also checked the Mahalanobis distance and Leverage values, and then deleted one outlier. After all these steps, we ran the linear regression process again and came up with the final regression in the calibration group of Cluster 2 (see table) and the result is as follows:

REVENUE=75.037+1.679*COST+8.131*QUANTITY-0.374*MONTH-4.765*ORDER_NO+7.669*PHONE

TABLE 4. COEFFICENETS (LINEAR REGRESSION) IN THE CALIVRATION GROUP OF CLUSTER TWO (AFTER DELETE THE OUTLIERS)

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	75.037	3.27		22.95	0		
COST	1.679	0.009	0.905	190.162	0	0.881	1.136
QUANTITY	8.131	0.329	0.125	24.691	0	0.776	1.288
MONTHS	-0.374	0.041	-0.041	-9.083	0	0.99	1.01
ORDER_NO	-4.765	0.858	-0.027	-5.553	0	0.851	1.176
PHONE	7.669	1.596	0.022	4.804	0	0.981	1.019

Then, we selected samples both in the validation group and also in Cluster 2. We created a new variable and used the regression model to calculate the predicted value ($\widehat{y_i}$). After, we calculate the average of observed value of revenue ($\overline{y} = 426.29$), separately got SSR and SST (SST= $\sum_{i=1}^{n}(y_i - \overline{y})2$, and SSR=$\sum_{i=1}^{n}(\widehat{y_i} - \overline{y})2$) and eventually got the R2 for evaluating the validity of our regression model.

In Cluster 2, SST is equal to 100180026.5, and SSR is equal to 88172773.16, so the R^2 is equal to 0.88, which means that the regression model we got from calibration group is quite fit for the validation group. In another word, this regression model is trustful.

Then, we conducted the same progress on the Cluster 4. Since all the progresses are the same, we won't explain the details in this part and will give the outcome of our analysis with few tables.

TABLE 5. ANOVA TABLE (LINEAR REGRESSION) IN THE CALIVRATION GROUP OF CLUSTER FOUR (AFTER DELETE THE OUTLIERS)

	Sum of Squares	df	Mean Square	F	Sig.
Regression	166988222	4	41747055.6	2057.413	.000
Residual	21366465.7	1053	20291.041		
Total	188354688	1057			

TABLE 6. COEFFICENETS (LINEAR REGRESSION) IN THE CALIVRATION GROUP OF CLUSTER TWO (BEFORE CHECK THE OUTLIERS)

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	222.922	14.401		15.48	0		
COST	1.64	0.019	0.91	85.415	0	0.95	1.053
QUANTITY	6.947	0.772	0.097	9.004	0	0.931	1.074
MONTHS	-0.874	0.182	-0.05	-4.798	0	0.985	1.015
PHONE	19.355	8.851	0.023	2.187	0.029	0.983	1.017

Thus, final regression in the calibration group of Cluster 4 (see *Table.5* and *Table.6*) and the result is as follows:

REVENUE=222.922+1.64*COST+6.947*QUANTITY-0.874*MONTH +19.355*PHONE

As the same process to validate the regression model, in Cluster 4, SST is equal to 138064514, and SSR is equal to 122745936, so the R^2 is equal to 0.889, which means the regression model is also trustful.

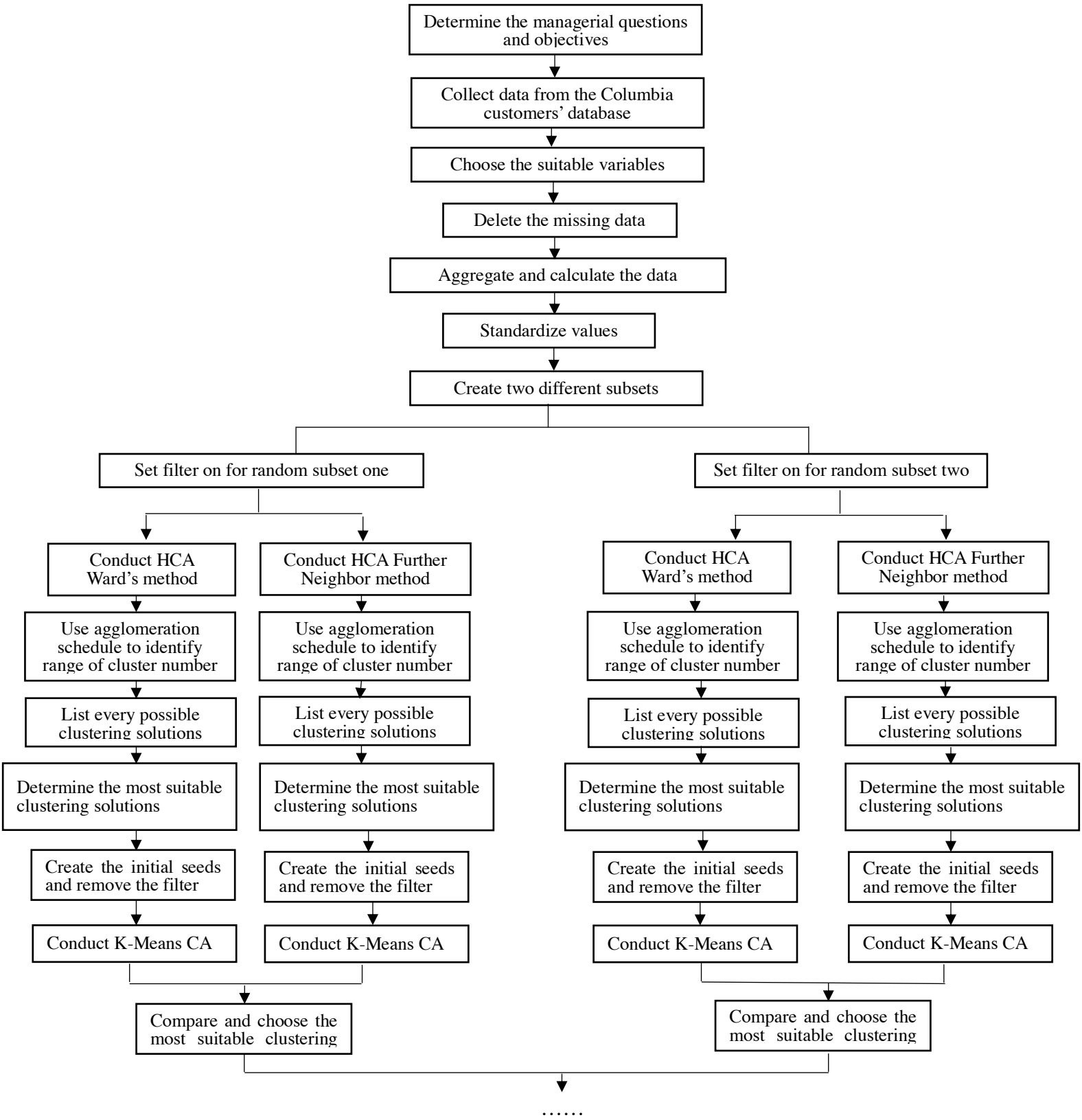

Figure1. Flowchart for project 2 and 3(to be continued)

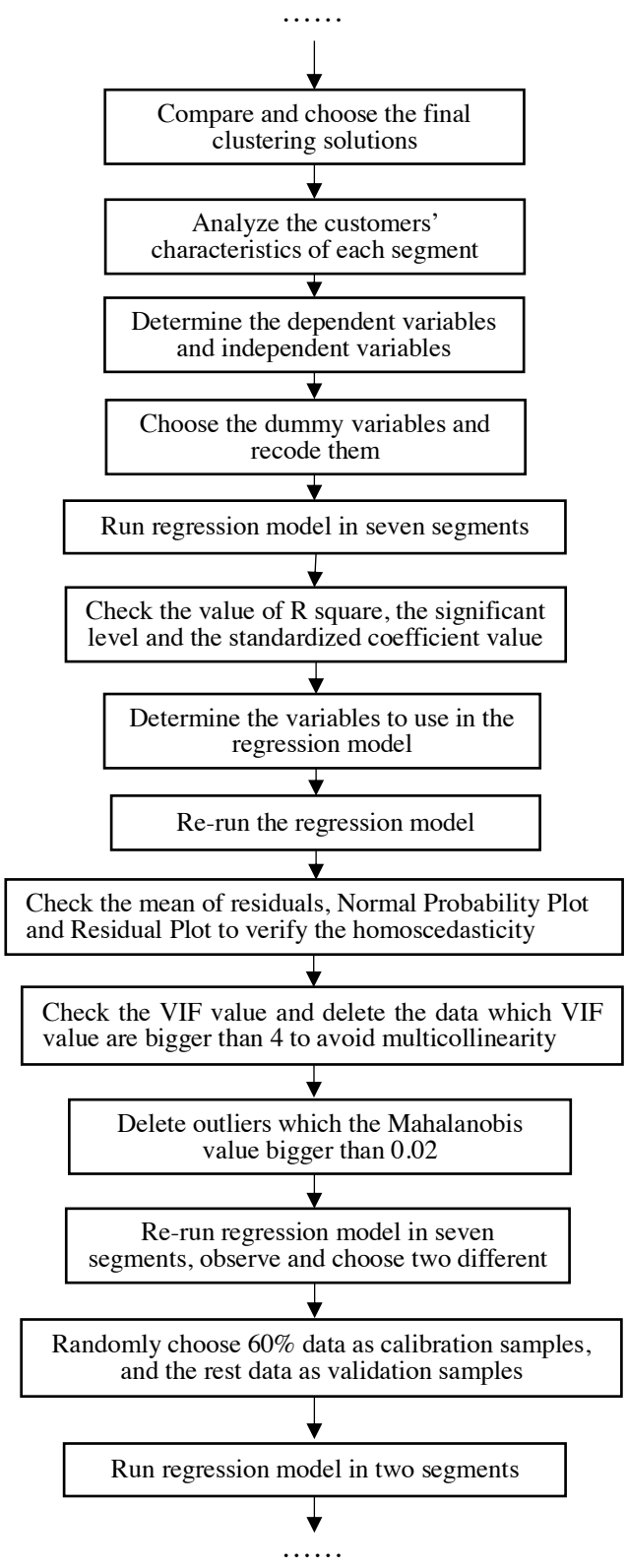

Figure1. Flowchart for project 2 and 3(to be continued)

……

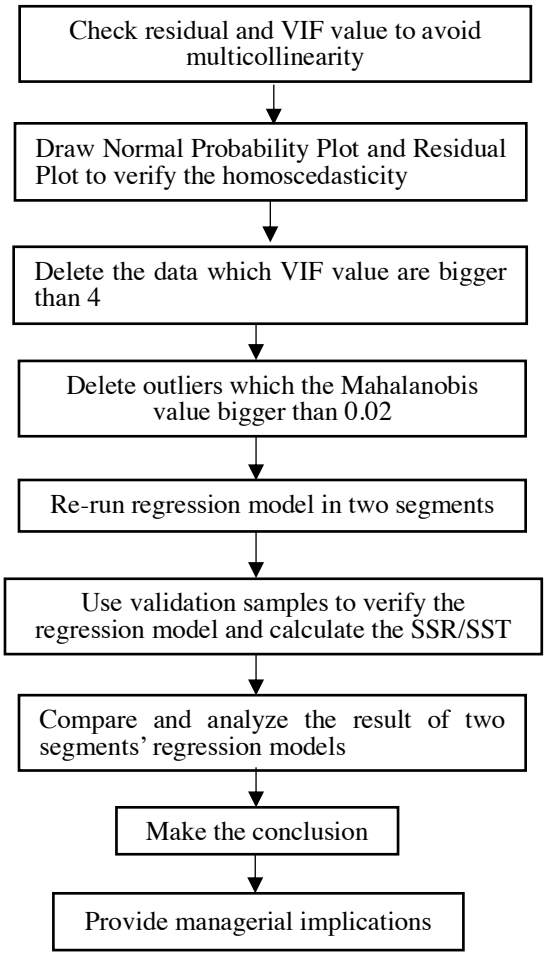

Figure1. Flowchart for project 2 and 3