**American Heart  
Association   
Heart Walk Events**

-Identify Potential Zero Dollar Walker

Group 3: Pei-Chun Lin, Vashisth Sorathia, Sharon Wu

**Contents**

[Introduction 2](#_Toc500365027)

[Objective 2](#_Toc500365028)

[Data Overview 2](#_Toc500365029)

[Zero Dollar Walker Definition 3](#_Toc500365030)

[Data Exploration 4](#_Toc500365031)

[Data Cleansing 5](#_Toc500365032)

[Variables Selection 7](#_Toc500365033)

[Solution Approach 8](#_Toc500365034)

[Segmentation 8](#_Toc500365035)

[Logistic Regression Model 9](#_Toc500365036)

[Linear Regression Model 14](#_Toc500365037)

[Financial Implication 16](#_Toc500365038)

[Recommendations & Conclusion 18](#_Toc500365039)

# **Introduction**

American Heart Association (AHA) is the largest national non-profit voluntary health organization dedicated to fighting heart disease and stroke. The organization holds around 300 Heart Walk Event annually to collect donation to help heart disease patients. The Heart Walk Event touches 1 million participants nationwide and it is the largest fundraiser with raising $130M donations.

However, the organization noticed a significant issue that there are a huge amount of Zero Dollar Walker (ZDW) in each event, participant who did not collect donation in a event year. Each year approximately 58% of Heart Walk participants failed to raise funds, personal or peer-to-peer. If AHA can convert this chunk of participants, there is a 5% increase in donations would mean 2M in revenue. Therefore, “How to identify Zero Dollar Walker and increase the revenue in donation?” becomes one of the first priority problems that AHA wants to solve.

# **Objective**

To help American Heart Association to identify Zero Dollar Walker and increase the revenue, the followings are the objectives for this project:

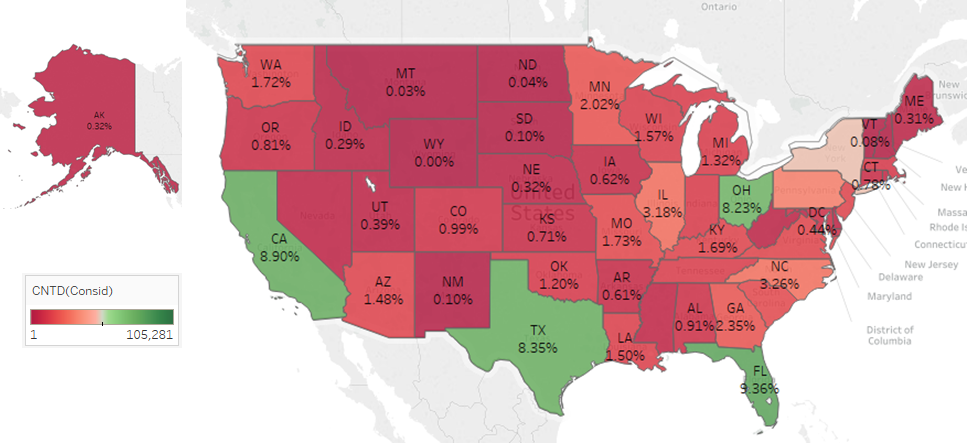
* Leverage data science to identify factors that affect a participant’s propensity of donating
* Using historic data, study common characteristics of a Zero Dollar Walker and utilize this learning to hint if a given participant is potentially one
* Deploying a real-time scalable learning model will help AHA to deal with such participants in advance and present an opportunity to proactively convert them

# **Data Overview**

The dataset provided by American Health Association has 1,624,391 records which mostly are transaction-based, and each participant may have several records. There are 43 variables which include:

* Geographic Information: Stdhomeaddress.City, Stdhomeaddress.State
* Income: Meddi.Cy, Medhinc.Cy, Mednw.Cy
* Lifestyle: Tapsegnum,Tap.Desc, Tap.Lifed, Tap.Life
* Historical Behavior: Lifetimetotal.Transamount, Prevyeartotal.Transamount
* Transaction Information: Giftamount, Registration.Gift

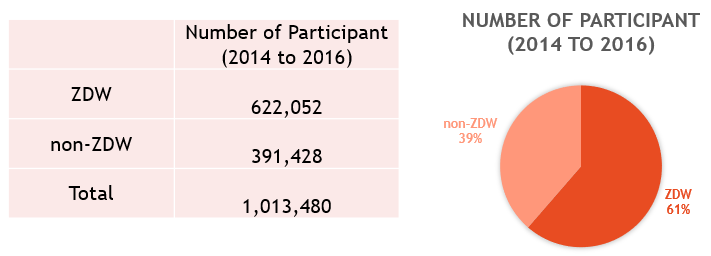
(Figure 1)



This is a geographic visualization. The numbers of participants are filled in each country of the map. The high percentage of participants appears green and low percentage appears red since green makes us think about positive and red is negative. We can get the exact value of percentage by click each country. By compare the level of color, we can also easily compare the percentage of participants. For example, Florida (9.36%), California (8.9%), Ohio (8.23%), Texas (8.35%) and New York (5.12%) have high percentage of participants.

# **Zero Dollar Walker Definition**

(Figure 2)

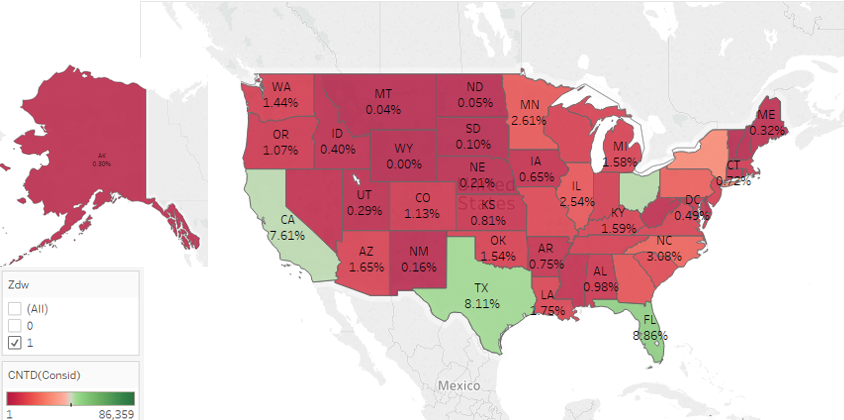


A participant has an aggregate total gift amount of $0 for an event year is a Zero Dollar Walker. A participant may be a Zero Dollar Walker for a given year but not for another year based on gift amounts. Therefore, we took out event year from event date in the dataset which provided by AHA and aggregated Gift Amount as one record per Participant ID and Event Year. If total gift amount equal to $0 for an event year, we set ZDW flag = 1 to the record. On the other hand, ZDW = 0 means that the participant has collected some donations and is not a Zero Dollar Walker in the event year.

After we aggregated the gift amount based on Participant ID and Event Year, total record numbers decrease from 1,624,391 to 1,013,480. Total Zero Dollar Walker from year 2014 to year 2016 had 622,052 participants which is 61% of the total aggregated records, and AHA had 39% non-Zero Dollar Walker which indicates 391,428 participants. The summary is shown in Figure 2.

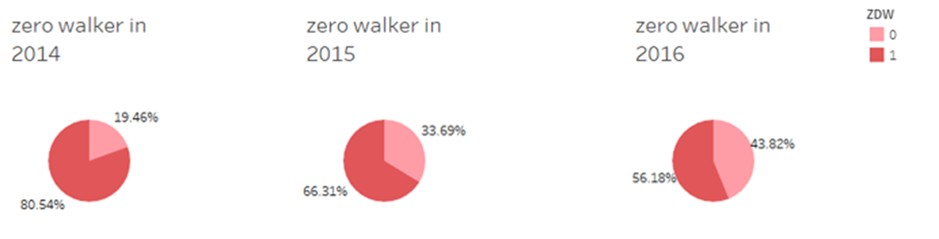
# **Data Exploration**

(Figure 3)



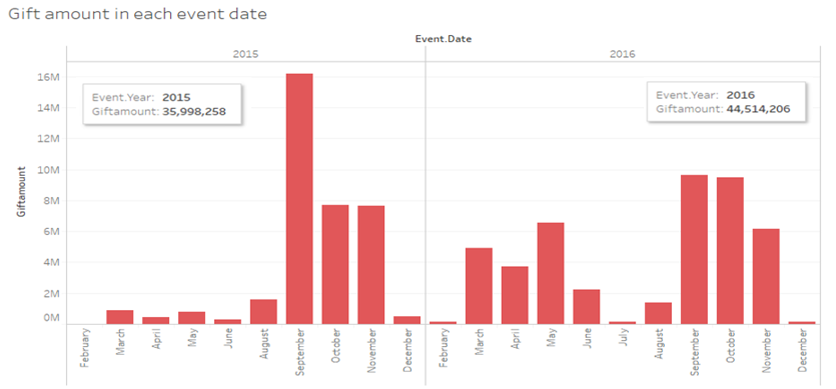
This geographic visualization shows that Florida (8.86%), Texas (8.11%), California (7.61%), Ohio (7.71%), and New York (4.69%) have high percentage of Zero Dollar Walker.

(Figure 4)



This is a pie chart for viewing the percentage of Zero Dollar Walker from 2014 to 2016. Dark pink represents Zero Dollar Walker and light pink represents non-Zero Dollar Walker. Based on this chart, we figured out that Zero Dollar Walker decreases every year, but it still has more than half of the participant population in 2016.

(Figure 5)



This is a bar chart. The degree of bar color reflects the gift amount in each event date. We found that the highest gift amount AHA earned from donors was in September 2015. Total gift amount in 2015 was 35,998,258 and 2016 was 44,914,206 which increased 20% within two years.

# **Data Cleansing**

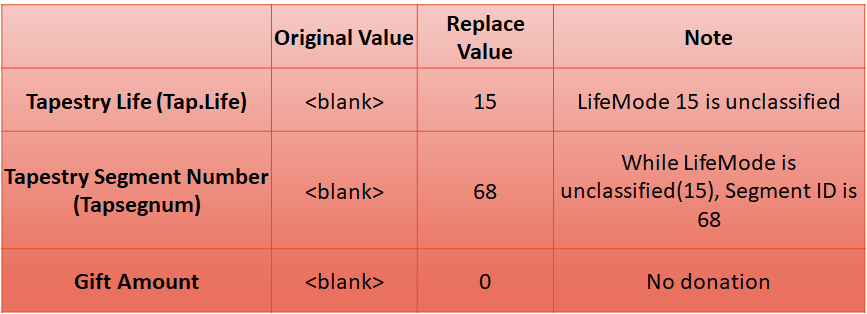
In AHA dataset, there are 9 variables with more than 99% missing value will be ignored which include:

* Current Year Site Visit Count
* Total Count Donated Offline
* Total Count Donated Online
* Total Count Raised Offline
* Total Count Raised Online
* Total Dollars Donated Offline
* Total Dollars Donated Online
* Total Dollars Raised Offline
* Total Dollars Raised Online

On the other hand, there are some variables with missing value less than 50% which is replaced based on different criteria:

* When Tapestry Life is missing, the field is filled in value 15 since the original dataset has a life mode in 15 with description “unclassified” which has the same meaning as missing value.
* When Tapestry Segment Number is missing, the field is filled in value 68 since while life mode is unclassified with value 15, segment id for the participant has a value of 68
* When Gift Amount is missing, we assumed that the donor did not donate and the value should be filled in value 0

(Figure 6)

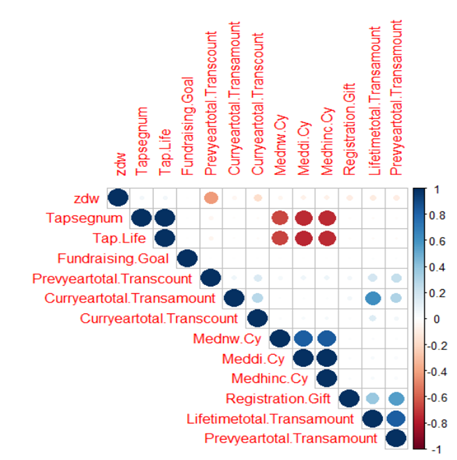


# 

# 

# **Variables Selection**

(Figure 7)



By running correlation analysis, we are able to identify whether any multicollinearity problem will be happened in our regression model. The high correlation appears blue and low correlation appears red. However, in the result, Meddi Cy, Medhinc Cy and Mednw.Cy have highly correlation. We selected only median household income (Medhinc Cy). TapLife and Tapsegnum are highly correlated, and both are correlation with income as well, so we created segmentation based on tapestry life (TapLife). Lifetimetotal.Transamount has highly correlated with Preyeartotal.Transamount. Since there is no missing value in the previous year total transaction amount, we kept it into our regression model rather than Lifetimetotal.Transamount.

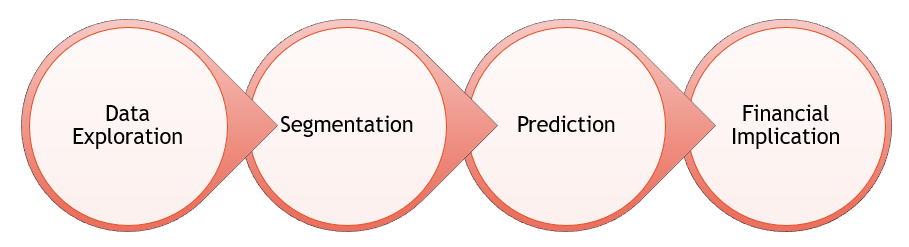
# 

# 

# **Solution Approach**

First, we created some graphs to explore AHA dataset and got ideas for our future analysis. Second, we divided the participants to diverse groups based on their historical behaviors to have different strategies for each group. Third, we built logistic and linear regression models to have better prediction for Zero Dollar Walker and gift amount. In the end, we had our financial implication and recommendations for AHA to help the organization improve future donation gift amounts. (Figure 8)

(Figure 8)

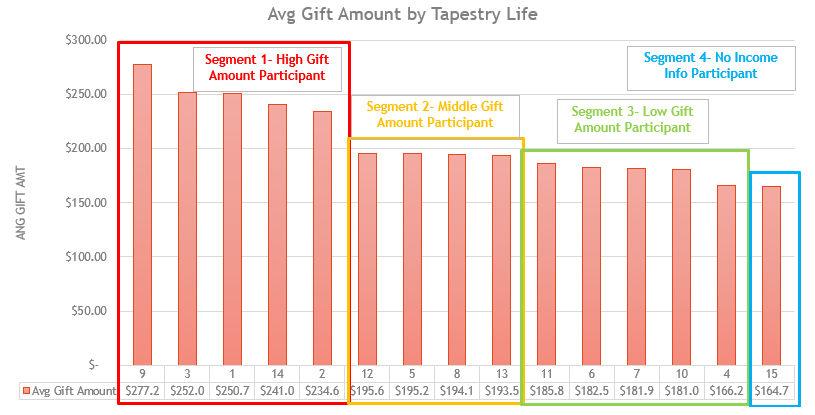


# **Segmentation**

To have a better strategy for the Zero Dollar Walker forecast, we decided to divide the participants to several groups. We tried to segment participants based on income, geographic information, lifestyle, historical behavior or transaction information, and figured out that we had the best segmentation when we segmented each group by Tapestry Life and level of average gift amount. Therefore, our final segments have 4 groups as following:

* Segment 1 includes Tapestry Life 1, 2, 3, 9 and 14 which have average gift amount greater than $200 -> High Gift Amount Participant
* Segment 2 includes Tapestry Life 5, 8, 12 and 13 which have average gift amount between $190 and $200 -> Middle Gift Amount Participant
* Segment 3 includes Tapestry Life 4, 6, 7, 10 and 11 which have average gift amount less than $190 -> Low Gift Amount Participant
* Segment 4 includes Tapestry Life 15 which participants did not provide income information -> No Income Information Participant

(Figure 9)

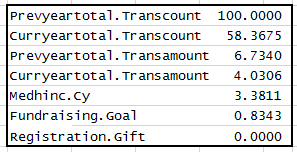


# **Logistic Regression Model**

Logistic Regression is selected as a predictive model in this project because the responding variable, Zero Dollar Walker and non-Zero Dollar Walker, is binary. We split each segment into 70% training and 30% testing dataset.

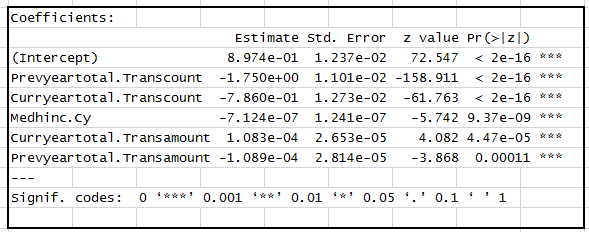
In the segment 1, the following graphic shows the important percentage of variables:

(Figure 10)



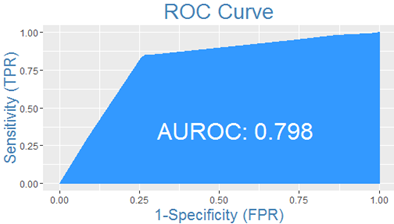
We chose five variables which important percentage is more than 1 into our predictive model and those variables are significant at 5% level.

(Figure 11)



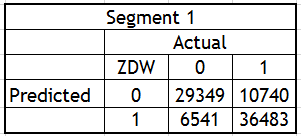
In the ROC curve below, area under the curve shows that this model is covering 0.798 of the area, which indicates that this model is an acceptable discrimination.

(Figure 12)



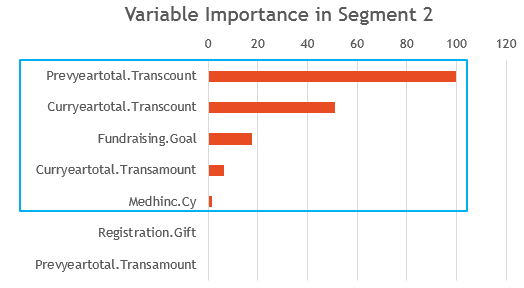
Then we compared our estimation results with actual values of zero dollar walk or non-Zero Dollar Walker in testing and training dataset. The result is 79.21% accuracy to the prediction.

(Figure 13)

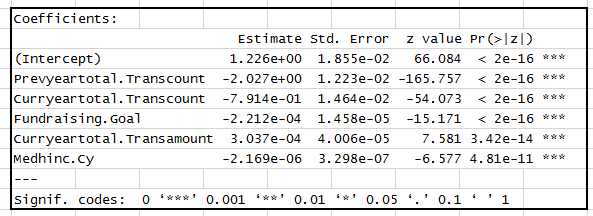


In the segment 2, the importance percentage of variables shows in the Figure 14, we chose top five variables again for our predictive model and those variables are significant at 5% level.

(Figure 14)

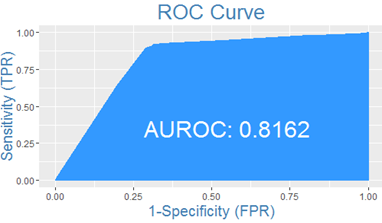


(Figure 15)



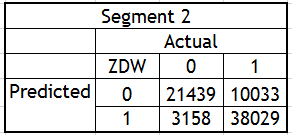
In the ROC curve below, area under the curve shows that this model is covering 0.8162 of the area, which indicates that this model is a good discrimination.

(Figure 16)



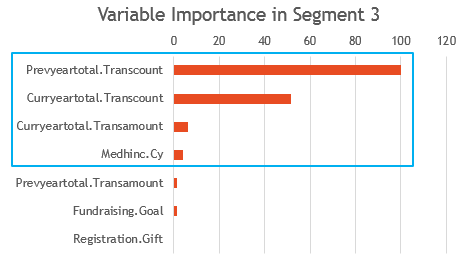
Then we compared our estimation results with actual values of zero dollar walk or non-Zero Dollar Walker in testing and training dataset. The result is 81.85% accuracy to the prediction.

(Figure 17)

****

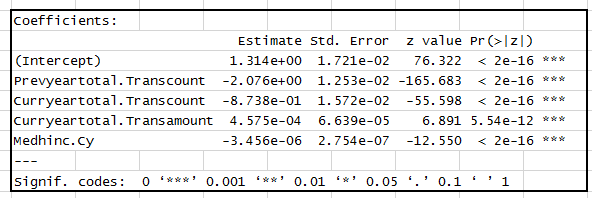
In the segment 3, the importance percentage of variables shows in Figure 18, we chose top four variables which have percentage greater than 2.

(Figure 18)



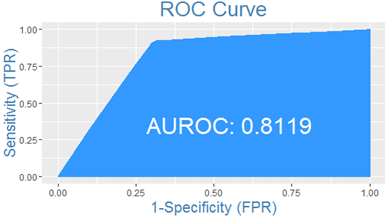
Those variables are significant at 5% level as well.

(Figure 19)



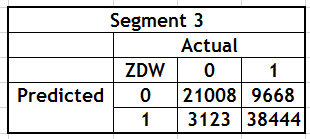
In the ROC curve below, area under the curve shows that this model is covering 0.8119 of the area, which indicates that this model is a good discrimination.

(Figure 20)



Then we compared our estimation results with actual values of zero dollar walk or non-Zero Dollar Walker in testing and training dataset. The result is 82.29% accuracy to the prediction.

(Figure 21)



# 

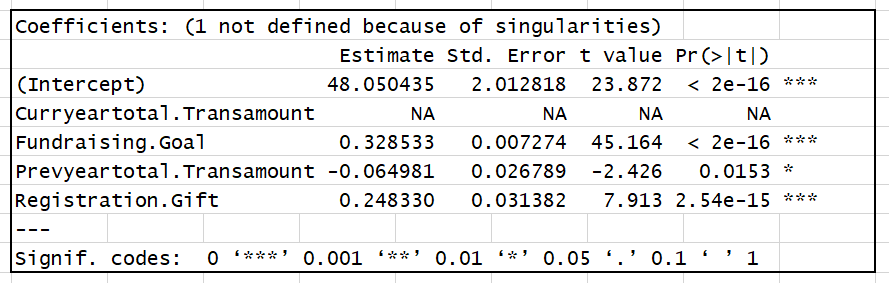
# **Linear Regression Model**

After we ran logistic regression model to understand what variables are important to predict potential ZDW, we built linear regression to forecast what the leading variables impact the gift amount and how. Here we took out only those non-ZDWs to build the model and removed outliers with less and greater than 1.5 interquartile range.

In the segment 1, variable Fundraising Goal and Registration Gift have positive relationship with Gift Amount; Previous Year Total Transaction Amount has negative relationship with Gift Amount. There is no relationship between Current Year Total Transaction Amount and Gift Amount.

If a participant’s fundraising goal goes up by $10, the total gift amount will probably go up by $3.3; if a participant’s registration gift goes up by $10, the total gift amount will probably go up by $2.5.

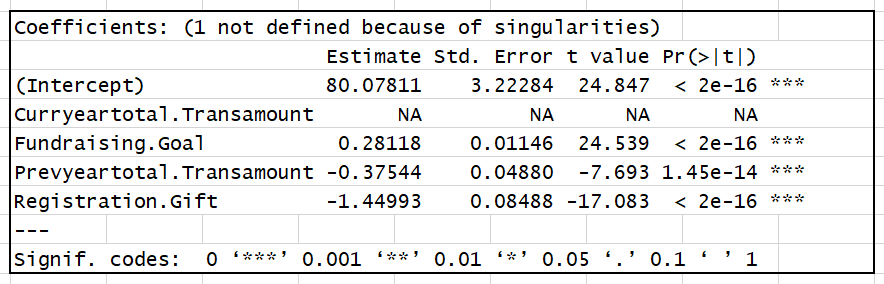
(Figure 22)



In the segment 2, variable Fundraising Goal has positive relationship with Gift Amount; Previous Year Total Transaction Amount and Registration Gift have negative relationship with Gift Amount. There is no relationship between Current Year Total Transaction Amount and Gift Amount.

If a participant’s fundraising goal goes up by $10, the total gift amount will probably go up by $2.8.

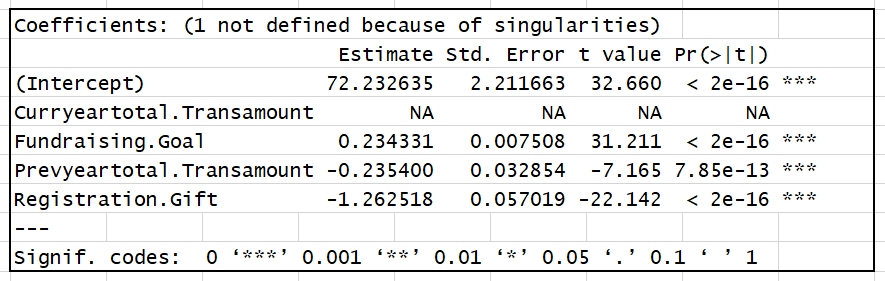
(Figure 23)



In the segment 3, variable Fundraising Goal has positive relationship with Gift Amount; Previous Year Total Transaction Amount and Registration Gift have negative relationship with Gift Amount. There is no relationship between Current Year Total Transaction Amount and Gift Amount.

If a participant’s fundraising goal goes up by $10, the total gift amount will probably go up by $2.3.

(Figure 24)

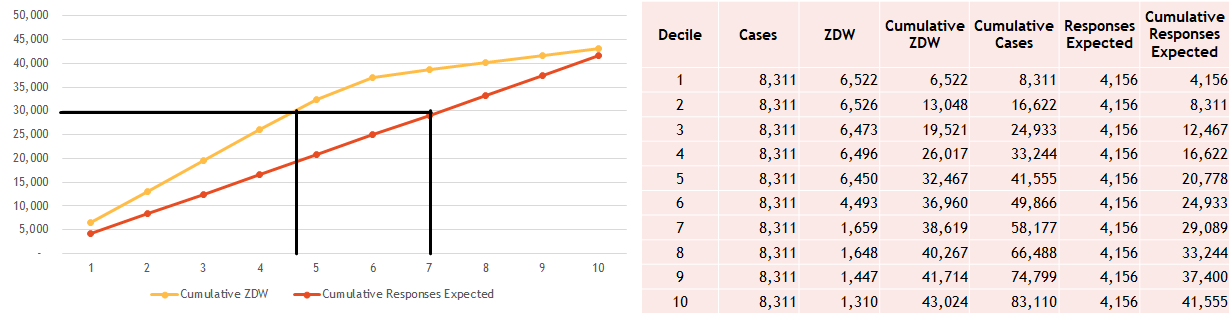


# **Financial Implication**

We created cumulative gains chart to show the percentage of the overall number of cases in each category “response” by targeting a percentage of the total number of cases. First, we divided each segment to 10 equal buckets based on participant records, and had top 10% predicted ZDW rate in the first bucket, top 10% to 20% predicted ZDW rate in the second bucket, and so on. Then we assumed the expected response rate is 50%, and calculated the cumulative ZDW, cumulative cases and cumulative responses expected to have our marketing cost saving.

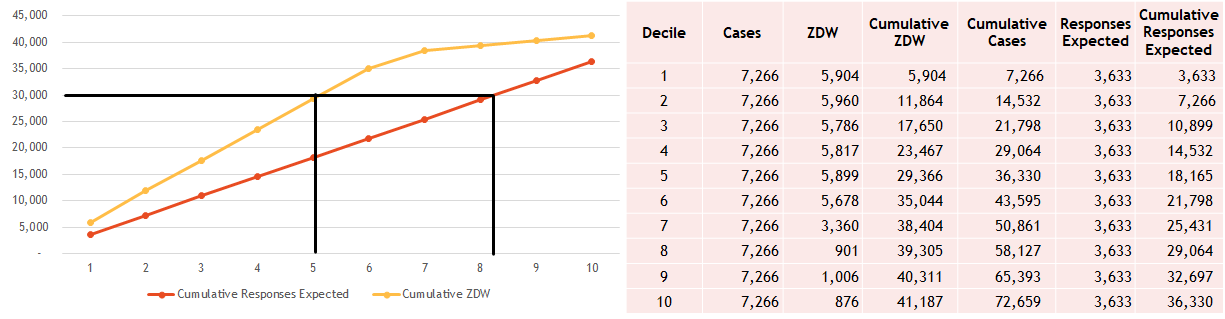
In the segment 1, there are 83.1k participants. According to Figure 25, if we want to reach 30k ZDWs, we need to target only 39k participants based on our logistic regression model result instead of 58.1k participants (without the logistic regression model). By doing this, we save about 33% in marketing costs compared to baseline scenario.

(Figure 25)



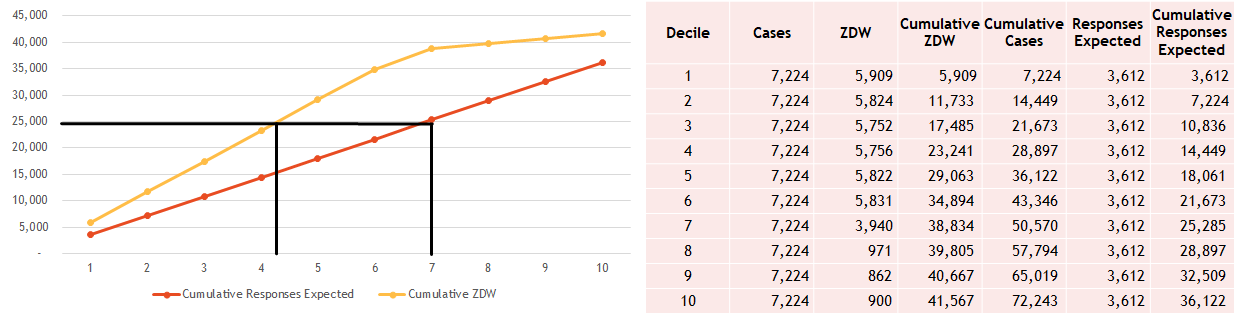
In the segment 2, there are 72.6k participants. According to Figure 26, if we want to reach 30k ZDWs, we need to target only 36k participants based on our logistic regression model result instead of 59k participants (without the logistic regression model). By doing this, we save about 39% in marketing costs compared to baseline scenario.

(Figure 26)



In the segment 3, there are 72.2k participants. According to Figure 27, if we want to reach 25k ZDWs, we need to target only 30.3k participants based on our logistic regression model result instead of 50k participants (without the logistic regression model). By doing this, we save about 39.4% in marketing costs compared to baseline scenario.

(Figure 27)

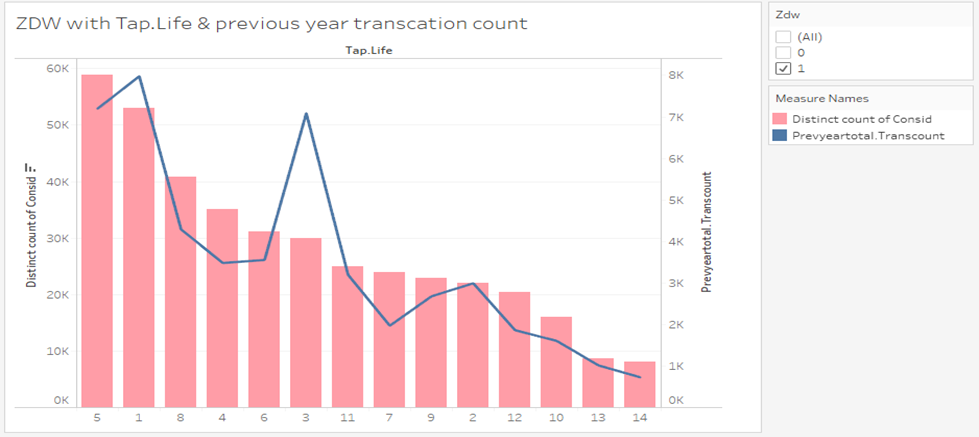


# 

# **Recommendations & Conclusion**

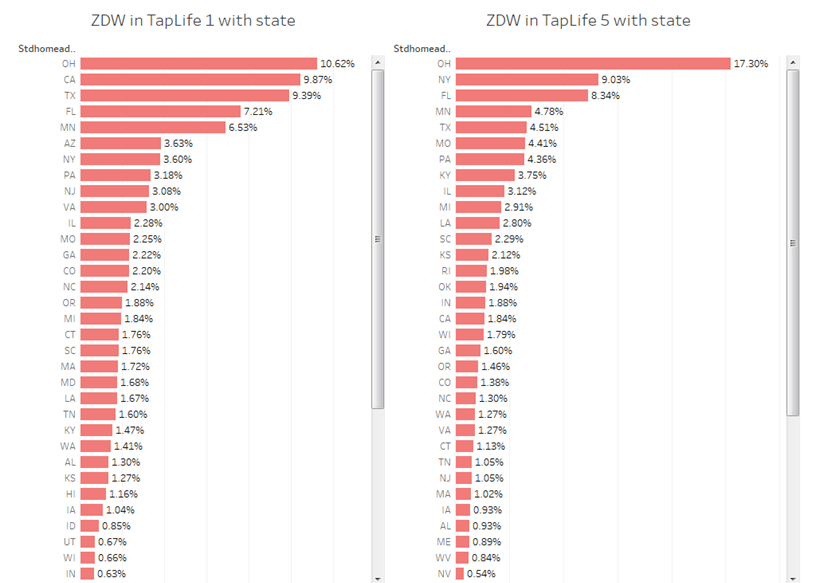
After several statically analyses, we concluded the following recommendations for AHA to decrease their ZDW in Heart Walker Event and increase their revenue:

(Figure 28)



This is a visualization combining bar chart and line chart. The line is total transaction count in previous year and the bar is the numbers of Zero Dollar Walker. For this visualization in Figure 28, we suggest AHA can focus on the participants in Tap life 1 and 5.

(Figure 29)

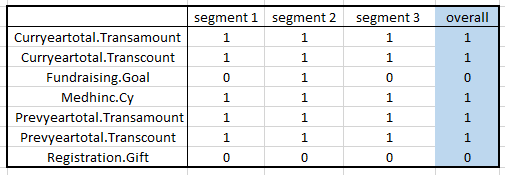


Specifically, we recommend that AHA focus on Zero Dollar Walker s in Tap life 1 in Ohio, California and Texas, and Tap life 5 in Ohio, New York and Florida and try marketing strategies having them make some donations.

AHA can also plan a special marketing strategy on Segment 1 because this segment has high gift amount participant (average gift amount greater than $200). If AHA can lower the percentage of ZDW, there will be a big increase in donation amount.

By looking at the logistic regression model, we are able to find significant variables in each segment. Marked 1, if the variable turns out to be significant in the model, otherwise marked 0. We concluded that Curryeartotal.Transamount, Curryeartotal.Transcount, Medhinc, Preyeartotal.Transamount and Preyeartotal.Transcount are significant variables to identify Zero Dollar Walker s.

(Figure 30)



By looking at the linear regression model, when focus on segment 1, AHA should encourage the participants to have higher fundraising goal and registration gift which every $10 amount higher will cause about $3 higher on revenue. On the other hand, when focus on segment 2 and 3, AHA should encourage the participants to have higher fundraising goal which every $10 amount higher will cause about $2.5 higher on revenue.

According to financial implication, AHA can save approximately 37% marketing cost by targeting 35% participants instead of all participant in each segment.

(Figure 31)

