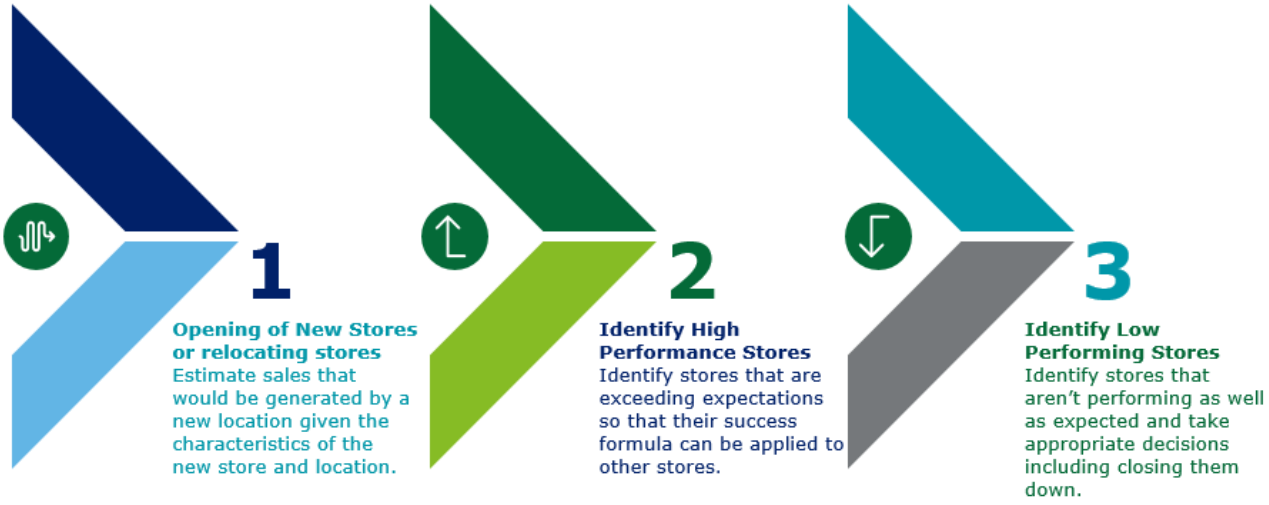
Problem 2:



Findings and insights:

Firstly, I removed few unwanted rows which have NAs and cleared the data. Also, I already have a script (built earlier for other datasets) which use to give same columns in the dataset. So, after running it, I got below output. So, I removed column `CYB07VBASE from the dataset as in variable description also they corresponds to Total households.`

```
[1] 41
[1] "CYB02V001"
[1] 48
[1] "CYB07VBASE"
[1] "##############################################"
[1] "Script Executed"
```

```
summary(Data_Problem_2)
Data_P2 = Data_Problem_2[complete.cases(Data_Problem_2),c(-48)]
```

Now to find out the significant variables which contribute more to Revenue, I ran a linear model.

```
model1 = lm(REVENUE_2013~., data = Data_P2)
summary(model1)
```

Below are the significant variables:
U_CITY, U_STATE, TOT_ATTRITION_2013, NUM_CUST_ACC_REPS, NUM_STORE_MANAGERS, NUM_EMP_PAY_TYPE_H, AVG_PAY_RATE_PAY_TYPE_H, AVG_PAY_RATE_PAY_TYPE_S, FRONTAGE_ROAD, MARKETING_EXP_2012, TOT_NUM_LEADS, NUM_CONVERTED_TO_AGREEMENT, URBANICITY, CYA21V001, CYA21V001, PERC_CYEA07V001, PERC_CYEA07V004

Using the above significant variable, again ran the linear model and check for the adjusted R-squared. Then run the backward step to find the best model and variables.

```
model2 =
lm(REVENUE_2013~U_CITY+U_STATE+TOT_ATTRITION_2013+NUM_CUST_ACC_REPS+NUM_STORE_MANAG
ERS+NUM_EMP_PAY_TYPE_H+AVG_PAY_RATE_PAY_TYPE_H+AVG_PAY_RATE_PAY_TYPE_S+FRONTAGE_ROA
D*MARKETING_EXP_2012+TOT_NUM_LEADS+NUM_CONVERTED_TO_AGREEMENT+URBANICITY+CYA21V00
1+PERC_CYEA07V001+PERC_CYEA07V004, data = Data_P2)
summary(model2)

step(lm(REVENUE_2013~U_CITY+U_STATE+TOT_ATTRITION_2013+NUM_CUST_ACC_REPS+NUM_STORE_MA
NAGERS+NUM_EMP_PAY_TYPE_H+AVG_PAY_RATE_PAY_TYPE_H+AVG_PAY_RATE_PAY_TYPE_S+FRONTAGE_
ROAD+MARKETING_EXP_2012+TOT_NUM_LEADS+NUM_CONVERTED_TO_AGREEMENT+URBANICITY+CYA21
V001+PERC_CYEA07V001+PERC_CYEA07V004, data = Data_P2),direction = "backward")

model3 =
lm(REVENUE_2013~U_CITY+U_STATE+TOT_ATTRITION_2013+NUM_CUST_ACC_REPS+NUM_STORE_MANAG
ERS+NUM_EMP_PAY_TYPE_H+AVG_PAY_RATE_PAY_TYPE_H+AVG_PAY_RATE_PAY_TYPE_S+FRONTAGE_ROA
D+MARKETING_EXP_2012+TOT_NUM_LEADS+NUM_CONVERTED_TO_AGREEMENT+URBANICITY+CYA21V00
1+PERC_CYEA07V001+PERC_CYEA07V004, data = Data_P2)
summary(model3)
```

so, the best model is used to predict the revenue and difference in both revenue is calculated. If Difference is positive, then store has high performance otherwise low.

```
Data_P2$pred_Revenue_2013 = predict(model3, Data_P2)
attach(Data_P2)
Data_P2$diff_Revenue_2013 = REVENUE_2013-pred_Revenue_2013
Data_P2$store_performance = ifelse(Data_P2$diff_Revenue_2013>0,"HIGH","LOW")

length(Data_P2$store_performance[Data_P2$store_performance=="HIGH"])
length(unique(Data_P2$U_CITY[Data_P2$store_performance=="HIGH"]))
length(unique(Data_P2$U_STATE[Data_P2$store_performance=="HIGH"]))
length(unique(Data_P2$SQUARE_FEET[Data_P2$store_performance=="HIGH"]))
```

```
There are 1076 high performing stores and rest are low performing store.

At the end this can be displayed in the form of decision tree to have better understanding.
In Decision tree, U_CITY are not considered as its give unreadable tree but there were 742
unique cities and 47 unique states having high performing stores with 463 unique sizes in s
quare feet.
```

```
data_decision_tree =
Data_P2[,c("CYA21V001","PERC_CYEA07V001","PERC_CYEA07V004","store_performance","TOT_NUM_LEADS","NUM_CONVERTED_TO_AGREEMENT","URBANICITY","NUM_EMP_PAY_TYPE_H","MARKETING_EXP_2012","AVG_PAY_RATE_PAY_TYPE_H","AVG_PAY_RATE_PAY_TYPE_S","FRONTAGE_ROAD","SQUARE_FEET","NUM_CUST_ACC_REPS","TOT_ATTRITION_2013","NUM_STORE_MANAGERS")]

library(rpart)
library(rpart.plot)

#decision
dtm = rpart(store_performance~. , data_decision_tree, method = "class")
rpart.plot(dtm, type = 4, extra = 101)
```