/\*1.Import dataset in the SAS environment and check top 10 record of import dataset\*/  
proc import file="/home/u61780667/Life+Insurance+Dataset.csv"  
 out=work.life\_insurance  
 dbms=csv;  
run;  
  
data work.life\_insurance\_top10\_rows;  
 set work.life\_insurance(obs=10);   
run;  
  
/\*2.Check variable type of the import dataset\*/  
proc contents data=work.life\_insurance;  
run;  
  
/\*3.Checks if any variables have missing values, if yes then do treatment?\*/  
proc means data=life\_insurance nmiss;  
run;  
/\*There are no null values\*/  
  
/\*4.Check summary and percentile distribution of all numerical variables for churners and   
non-churners?\*/  
proc univariate data= life\_insurance;  
var Age Cust\_Tenure Overall\_cust\_satisfation\_score CC\_Satisfation\_score Cust\_Income Agent\_Tenure  
 YTD\_contact\_cnt Due\_date\_day\_cnt Existing\_policy\_count Miss\_due\_date\_cnt;  
class churn;  
run;  
  
proc means data= life\_insurance n nmiss min p1 p5 p10 p25 p50 p75 p90 p95 p99 max;  
var Age Cust\_Tenure Overall\_cust\_satisfation\_score CC\_Satisfation\_score Cust\_Income Agent\_Tenure  
 YTD\_contact\_cnt Due\_date\_day\_cnt Existing\_policy\_count Miss\_due\_date\_cnt;  
run;  
  
  
/\*5.Check for outlier, if yes then do treatment?\*/  
proc sgplot data= life\_insurance;  
vbox Age;  
run;  
  
proc sgplot data= life\_insurance;  
vbox Cust\_Tenure;  
run;  
  
proc sgplot data = life\_insurance;  
vbox Overall\_cust\_satisfation\_score;  
run;  
  
proc sgplot data = life\_insurance;  
vbox CC\_Satisfation\_score;  
run;  
  
proc sgplot data = life\_insurance;  
vbox Cust\_Income;  
run;  
  
proc sgplot data = life\_insurance;  
vbox Agent\_Tenure;  
run;  
  
proc sgplot data = life\_insurance;  
vbox YTD\_contact\_cnt;  
run;  
  
proc sgplot data = life\_insurance;  
vbox Due\_date\_day\_cnt;  
run;  
  
proc sgplot data = life\_insurance;  
vbox Existing\_policy\_count;  
run;  
  
proc sgplot data = life\_insurance;  
vbox Miss\_due\_date\_cnt;  
run;  
  
proc univariate data= life\_insurance;  
var Age Cust\_Tenure Overall\_cust\_satisfation\_score CC\_Satisfation\_score Cust\_Income Agent\_Tenure  
 YTD\_contact\_cnt Due\_date\_day\_cnt Existing\_policy\_count Miss\_due\_date\_cnt;  
run;  
  
  
/\*outliers in Miss\_due\_date\_cnt, Due\_date\_day\_cnt, YTD\_contact\_cnt, Cust\_Income.  
we will use flooring and capping for these variables\*/  
data insurance;  
 set life\_insurance;  
 if Cust\_Income > 31585.5 then Cust\_Income = 31585.5;  
 if YTD\_contact\_cnt > 30.5 then YTD\_contact\_cnt = 30.5;  
 if Due\_date\_day\_cnt > 29.75 then Due\_date\_day\_cnt = 29.75 ;  
 if Miss\_due\_date\_cnt > 5 then Miss\_due\_date\_cnt = 5;   
run;  
  
/\*6.Check the proportion of all categorical variables and extract percentage contribution   
of each class in respective variables?\*/  
proc freq data = insurance;  
table  
churn  
Payment\_period  
Product  
EducationField  
Gender  
Overall\_cust\_satisfation\_score  
Cust\_Designation  
CC\_Satisfation\_score  
Cust\_MaritalStatus  
Complaint  
/ nocum;  
run;  
  
/\*7.Customer service management want you to create a macro where they will just put mobile number  
 and they will get all the important information like Age, Education, Gender, Income and CustID\*/  
%MACRO customer\_information();  
DATA macro\_insurance (keep = Mobile\_num CustID Age EducationField Gender Cust\_Income);  
SET insurance;  
where Mobile\_num in (&Mobile\_num.);  
RUN;  
proc print data=output;  
run;  
%MEND;  
/\*input mobile number\*/  
%let Mobile\_num = 9878913773,9925945763;  
/\*run macro for output\*/  
%customer\_information;  
  
/\*8.Check correlation of all numerical variables before building model, because we cannot add   
correlated variables in model?\*/  
proc corr data= insurance noprob;  
var Age Cust\_Tenure Overall\_cust\_satisfation\_score CC\_Satisfation\_score Cust\_Income Agent\_Tenure  
 Complaint YTD\_contact\_cnt Due\_date\_day\_cnt Existing\_policy\_count Miss\_due\_date\_cnt;  
run;  
  
/\*9.Train and test (70:30) dataset from the existing data set. Put seed 1234?\*/  
proc freq data=insurance;  
table Churn / nocum;  
run;  
  
proc surveyselect data=insurance method=srs reps=1 sampsize=500 seed=1234 out=test;  
run;  
  
proc contents data=test varnum; /\* data=test \*/  
run;  
  
proc freq data=test;  
table Churn /nocum;  
run;  
  
proc sql;  
create table train as select tes.\* from insurance as tes  
where CustID not in (select CustID from test);  
quit;  
  
proc freq data=train;  
table Churn /nocum;  
run;  
  
/\*10.Develop linear regression model first on the target variable to extract VIF information   
to check multicollinearity?\*/  
/\* Note: For further questions train data is taken also it's not mentioned which data we have  
 to take in account in regression model (target variable) means which target variable?  
 Only numerical values are taken in consideration\*/  
proc contents data= train;  
run;  
  
proc freq data=train;  
tables Churn \* Overall\_cust\_satisfation\_score;  
run;  
  
proc freq data=train;  
tables (Age Cust\_Tenure Overall\_cust\_satisfation\_score CC\_Satisfation\_score Cust\_Income   
Agent\_Tenure Complaint YTD\_contact\_cnt Due\_date\_day\_cnt Existing\_policy\_count Miss\_due\_date\_cnt)   
\* Churn / chisq;  
run;  
  
data new\_train (keep = CustID Churn Age Cust\_Tenure Overall\_cust\_satisfation\_score   
CC\_Satisfation\_score Cust\_Income Agent\_Tenure Complaint YTD\_contact\_cnt Due\_date\_day\_cnt   
Existing\_policy\_count Miss\_due\_date\_cnt);  
set train;  
run;  
  
proc freq data=new\_train;  
tables ( Age Cust\_Tenure Overall\_cust\_satisfation\_score CC\_Satisfation\_score Cust\_Income   
Agent\_Tenure Complaint YTD\_contact\_cnt Due\_date\_day\_cnt Existing\_policy\_count Miss\_due\_date\_cnt)  
 \* Churn / chisq;  
run;  
  
proc logistic data = new\_train;  
class Churn Overall\_cust\_satisfation\_score / param=ref;  
model Churn = Overall\_cust\_satisfation\_score / lackfit rsq;  
title 'Churn vs Overall\_cust\_satisfation\_score';  
run;  
  
proc logistic data = new\_train;  
class Churn Age Cust\_Tenure Overall\_cust\_satisfation\_score CC\_Satisfation\_score   
Cust\_Income Agent\_Tenure Complaint YTD\_contact\_cnt Due\_date\_day\_cnt   
Existing\_policy\_count Miss\_due\_date\_cnt / param=ref;  
model Churn = Age Cust\_Tenure Overall\_cust\_satisfation\_score CC\_Satisfation\_score Cust\_Income   
Agent\_Tenure Complaint YTD\_contact\_cnt Due\_date\_day\_cnt Existing\_policy\_count Miss\_due\_date\_cnt   
/ lackfit rsq;  
title 'Churn vs Overall\_cust\_satisfation\_score - Multivariable Logistic Regression';  
run;  
  
proc corr data=new\_train;  
var Churn Age Cust\_Tenure Overall\_cust\_satisfation\_score   
CC\_Satisfation\_score Cust\_Income Agent\_Tenure Complaint YTD\_contact\_cnt Due\_date\_day\_cnt   
Existing\_policy\_count Miss\_due\_date\_cnt;  
run;  
  
/\*11.Create clean logistic model on the target variables?\*/  
%let var = Age Cust\_Tenure Overall\_cust\_satisfation\_score CC\_Satisfation\_score Cust\_Income   
Agent\_Tenure Complaint YTD\_contact\_cnt Due\_date\_day\_cnt Existing\_policy\_count Miss\_due\_date\_cnt;  
  
proc logistic data=new\_train descending outmodel=model;  
model Churn = &var / lackfit;  
output out = train\_output xbeta = coeff stdxbeta = stdcoeff predicted = prob;  
run;  
  
/\*12.Create a macro and take a KS approach to take a cut off on the calculated scores?\*/  
ods graphics on;  
proc npar1way edf plots= edfplot data= new\_train;  
class Churn;  
var Overall\_cust\_satisfation\_score;  
exact ks;  
run;  
ods graphics off;  
  
/\*13.Predict test dataset using created model?\*/  
data test;  
set test;  
prob = -0.0226-0.0398\*Age+0.4174\*Overall\_Satisfaction\_Score  
-0.00009\*Premium+0.0930\*Network\_hospital\_nearby+0.0289\*not\_passed\_percent\_claim;  
score = exp(problem)/(1+exp(problem));  
run;