```sas
options nodate nonumber;
ods graphics on / width = 2.5 in height = 2.125 in;
ods pdf file = "/home/u64313648/Personal Work Folder/Progress Report - Final Project/Project.pdf" pdftoc = 3 startpage =

proc import datafile="/home/u64313648/Personal Work Folder/Progress Report - Final Project/home_loan.csv"
    out=home_raw
    dbms=csv
    replace;
    guessingrows=max;
    getnames=yes;
run;

/* Data Preparation: */

data loan;
    set home_raw;
    drop Loan_ID;

    length Loan_Status_YN $3;
    if upcase(Loan_Status) in ("Y","YES") then Loan_Status_YN = "Yes";
    else if upcase(Loan_Status) in ("N","NO") then Loan_Status_YN = "No";
    else Loan_Status_YN = "NA";

    length DepCat $4;
    if Dependents in ("0","1","2","3+") then DepCat = Dependents;
    else DepCat = "3+";

    length Gender2 $6;
    if upcase(Gender) in ("MALE","M") then Gender2 = "Male";
    else if upcase(Gender) in ("FEMALE","F") then Gender2 = "Female";
    else Gender2 = "Other";

    length Married2 $3;
    if upcase(Married) in ("YES","Y") then Married2 = "Yes";
    else if upcase(Married) in ("NO","N") then Married2 = "No";
    else Married2 = "NA";

    length Educ2 $12;
    if upcase(Education) = "GRADUATE" then Educ2 = "Graduate";
    else if upcase(Education) = "NOT GRADUATE" then Educ2 = "Not Graduate";
    else Educ2 = "Other";

    length SelfEmp2 $3;
    if upcase(Self_Employed) in ("YES","Y") then SelfEmp2 = "Yes";
    else if upcase(Self_Employed) in ("NO","N") then SelfEmp2 = "No";
    else SelfEmp2 = "NA";

    TotalIncome = ApplicantIncome + CoapplicantIncome;
run;

proc format;
    value $loanfmt
        "Yes" = "Approved"
        "No"  = "Not Approved"
        "NA"  = "Missing/Other";

    value creditfmt
        0 = "No / Poor"
        1 = "Good";

    value $depFmt
        "0" = "0"
        "1" = "1"
        "2" = "2"
        "3+" = "3+";

    value $genderFmt
        "Male"   = "Male"
        "Female" = "Female"
        "Other"  = "Other";

    value $marFmt
        "Yes" = "Married"
        "No"  = "Not Married"
        "NA"  = "Unknown";

```

```
    value $eduFmt
        "Graduate"     = "Graduate"
        "Not Graduate" = "Not Graduate"
        "Other"        = "Other";

    value $selfFmt
        "Yes" = "Self-Employed"
        "No"  = "Not Self-Employed"
        "NA"  = "Unknown";
run;



/* Table 1: Loan_Status distribution */

title2 "Loan Status Prediction Project in SAS" justify = C;

proc freq data=loan;
    tables Loan_Status_YN / nocum norow nocol;
    format Loan_Status_YN $loanfmt.;
    title2 "Distribution of Loan Approval Status";
run;

/* Table 2: Overall categorical distributions */

proc freq data=loan;
    tables Married2 Credit_History Property_Area;
    format Gender2      $genderFmt.
           Married2     $marFmt.
           DepCat       $depFmt.
           Educ2        $eduFmt.
           SelfEmp2     $selfFmt.
           Credit_History creditfmt.;
    title2 "Overall Distribution of Socio-Economic Categorical Predictors";
run;

data loan_adjusted;
    set loan;
    if (Married2= "NA" OR Gender2 = "Other") then delete;
run;

proc odstext;
    p 'We removed values from the variables Married2 and Gender2 where we found that there were lack of observations for
run;

/* Table 3: Summary statistics overall (continuous predictors) */

proc means data=loan_adjusted n mean std median min max maxdec=1;
    var ApplicantIncome CoapplicantIncome TotalIncome
        LoanAmount Loan_Amount_Term;
    title2 "Summary Statistics for Continuous Predictors (Overall Sample)";
run;

/* Histogram: ApplicantIncome */
proc sgplot data=loan_adjusted;
    histogram ApplicantIncome;
    density ApplicantIncome;
    title2 "Distribution of Applicant Income";
run;

/* Histogram: LoanAmount */
proc sgplot data=loan_adjusted;
    histogram LoanAmount;
    density LoanAmount;
    title2 "Distribution of Loan Amount";
run;

proc odstext;
    p 'We can observe from the Distribution of Applicant Income vs Distribution of Loan Amount, that the first graph ter
run;

/* Table 4: Cross-tabs for key categorical predictors */

proc freq data=loan_adjusted;
    tables Loan_Status_YN * Credit_History
           Loan_Status_YN * Property_Area
           Loan_Status_YN * Married2
```

```
          / nocol norow;
    format Loan_Status_YN $loanfmt.
          Credit_History creditfmt.
          Married2 $marFmt.;
    title2 "Loan Approval Status by Selected Categorical Predictors";
run;

/* Fig. 2: Approval % by Credit History */

proc sgplot data=loan_adjusted;
    vbar Credit_History / group=Loan_Status_YN stat=percent
                          groupdisplay=cluster datalabel;
    format Credit_History creditfmt.
          Loan_Status_YN $loanfmt.;
    yaxis label="Percentage within Credit History Group";
    xaxis label="Credit History";
    title2 "Loan Approval by Credit History";
run;

proc odstext;
    p 'For those who were not approved, No/Poor credit history vs Good seemed to not have a major impact on the proporti
run;

/* Fig. 3: Approval % by Property Area */

proc sgplot data=loan_adjusted;
    vbar Property_Area / group=Loan_Status_YN stat=percent
                         groupdisplay=cluster datalabel;
    format Loan_Status_YN $loanfmt.;
    yaxis label="Percentage within Property Area";
    xaxis label="Property Area";
    title2 "Loan Approval by Property Area";
run;

/* Table 5: Continuous predictors by Loan_Status */

proc means data=loan_adjusted n mean std median min max maxdec=1;
    class Loan_Status_YN;
    var ApplicantIncome CoapplicantIncome TotalIncome
        LoanAmount Loan_Amount_Term;
    format Loan_Status_YN $loanfmt.;
    title2 "Continuous Predictors by Loan Approval Status";
run;

/* Fig. 4: Boxplot of ApplicantIncome by Loan_Status */

proc sgplot data=loan_adjusted;
    vbox ApplicantIncome / category=Loan_Status_YN;
    format Loan_Status_YN $loanfmt.;
    yaxis label="Applicant Monthly Income";
    xaxis label="Loan Approval Status";
    title2 "Applicant Income by Loan Approval Status";
run;


proc sgplot data=loan_adjusted;
    vbox LoanAmount / category=Loan_Status_YN;
    format Loan_Status_YN $loanfmt.;
    yaxis label="Requested Loan Amount";
    xaxis label="Loan Approval Status";
    title2 "Loan Amount by Loan Approval Status";
run;

proc odstext;
    p "In the Loan Amount by Loan Approval Status we see that the two boxplots look similar but that those who have the
run;


ods select none;
ods output ChiSq = chi_all;

proc freq data=loan;
    tables Loan_Status_YN *
          (Gender2 Married2 DepCat Educ2 SelfEmp2
           Credit_History Property_Area)
    / chisq;
run;
```

```
ods output close;
ods select all;


/* Keep only Pearson Chi-Square for each predictor */
data chi_summary;
    set chi_all;
    where Statistic = "Chi-Square";
    length Predictor $20;
    Predictor = scan(Table, 2, '');
run;

proc print data=chi_summary noobs;
    var Predictor DF Value Prob;
    format Prob pvalue8.4;
    title2 "Chi-Square Tests: Association with Loan Approval (Categorical Predictors)";
run;


/* T-tests */

ods select none;
ods output TTests = t_all;

proc ttest data=loan plots=none;
    class Loan_Status_YN;
    var ApplicantIncome CoapplicantIncome TotalIncome
        LoanAmount Loan_Amount_Term;
run;

ods output close;
ods select all;

data t_summary;
    set t_all;
    where Method = "Pooled";
run;

proc print data=t_summary noobs;
    var Variable DF tValue Probt;
    format Probt pvalue8.4;
    title2 "T-Test p-values for Continuous Predictors by Loan Status";
run;


/* Wilcoxon rank-sum nonparametric */
ods select none;
ods output WilcoxonTest = w_all;

proc npar1way data=loan wilcoxon plots=none;
    class Loan_Status_YN;
    var ApplicantIncome CoapplicantIncome TotalIncome
        LoanAmount Loan_Amount_Term;
run;

ods output close;
ods select all;

/* Keep only the Pr > |Z| line for each variable */
data w_summary;
    set w_all;
    *where Label1 = "Pr > |Z|";
run;

proc print data=w_summary noobs;
    *var Variable nValue1;
    *format nValue1 pvalue8.4;
    title2 "Wilcoxon Rank-Sum p-values for Continuous Predictors by Loan Status";
run;



proc odstext;
    p 'This dataset contains 480 loan applicants and includes socio-economic, demographic, and financial predictors rela

The continuous variables show the usual right-skewness seen in financial data. Applicant and co-applicant incomes vary w
```

```
To formally assess these patterns, we applied chi-square tests to each categorical predictor. Credit history, property a
run;



/* PROGRESS REPORT END */



title2 'Correlation Table Examining Collinearity Between Numerical Variables';

/* Reduced output from ods select */
proc corr data = loan_adjusted cov spearman plots =matrix(histogram);
    var ApplicantIncome CoapplicantIncome LoanAmount Loan_Amount_Term TotalIncome;
    ods trace on;
    ods select Corr.VarInformation Corr.PearsonCorr
run;



proc odstext;
    p 'We can see from this correlation matrix that TotalIncome and ApplicantIncome are both highly correlated with a Pe
run;


/* This will create a Loan_Status variable which is binary that can be used instead of the character version, but same i
Most up to date data that should be used is logistic_loan. No output created here*/
data logistic_loan;
    set loan_adjusted;
    if Loan_Status_YN = "Yes" then Loan_Status_Num = 1;
    else Loan_Status_Num = 0;
run;

/* Uses Reg procedure to further show collinearity diagnostics from proc reg reduced output from ods select*/
title2 'VIF and Collinearity Diagnostics Displayed From the REG Procedure';

proc reg data = logistic_loan;
    model Loan_Status_Num = ApplicantIncome CoapplicantIncome LoanAmount Loan_Amount_Term TotalIncome / VIF collin;
    ods trace on;
    ods select Reg.MODEL1.Fit.Loan_Status_Num.DependenceEquations Reg.MODEL1.Fit.Loan_Status_Num.ParameterEstimates Reg
run;

proc odstext;
    p 'Using proc reg to see the VIF between our numerical continuous variables we see that TotalIncome is a linear comb
        When looking at the collinearity diagnostics we see that TotalIncome has a condition index of 2125178 which is c
run;

/* Remember when fitting the models to not include TotalIncome in our model because of collinearity concerns */
Title2  'Stepwise Regression From LOGISTIC Procedure' justify=center;

/* Output Reduced*/
ods graphics on;

proc logistic data = logistic_loan;
    class Credit_History (ref='0')
          Property_Area  (ref='Rural')
          Married2       (ref='No')
          SelfEmp2 Educ2 Gender2 DepCat
          / param = ref;

    model Loan_Status_Num(event='1') =
          ApplicantIncome CoapplicantIncome LoanAmount Loan_Amount_Term
          SelfEmp2 Educ2 Gender2 DepCat
          Property_Area Credit_History Married2
          / selection = stepwise details lackfit influence;

    output out = logistic_loan2
           cbar = cbar
           DFBetas = DfBetas;

    ods exclude
        Influence
        InfluencePlots
        ROC
        ROCCurve
        AssociationPlot
```

```
        EffectPlot
        ParameterEstimates;

    ods select
        Logistic.ModelFit
        Logistic.StepwiseSummary
        Logistic.Step3.OddsRatios
        Logistic.Step3.Association
        Logistic.LackFit.LackFitChiSq;
run;

ods graphics off;



proc odstext;
    p "Output reduced to step 3 fit statistics. We have an AIC value of 479.972, a c statistic or AUC value of .789 whic
        We see that our Pearson residuals have positive values as high as around 4, and an evident trend of cases above
        When looking at our leverage we see that most of our higher leverage points tend to be cases where loan status i
run;

title2 'Meaningful Observations';

/* Will only print a few rows that are influential*/
proc print data = logistic_loan2;
    where cbar >= 1;
run;

proc print data = logistic_loan2;
    where DfBetas >= (2 / sqrt(515));
run;

proc odstext;
    p "These were influential observations for our model for one reason or another, observation 153 has a high applicant
run;

/* Genmod reduced output*/
title2 'Genmod Procedure Fit and Estimates';

proc genmod data = logistic_loan;
    class SelfEmp2 Educ2 Married2 Gender2 DepCat Property_Area Credit_History;
    model Loan_Status_Num(event = '1') = ApplicantIncome CoapplicantIncome LoanAmount Loan_Amount_Term SelfEmp2 Educ2 Ma
    ods trace on;
    ods select Genmod.ModelFit Genmod.ParameterEstimates;
run;

proc odstext;
    p "We're only displaying model fit statistics and parameter estimates, to show the AIC value (~490) from the genmod
run;

title2 'Final Model Selection and Interpretation';
proc odstext;
    p "This shows that our stepwise logistic regression model is the best model for predicting loan status. It doesn't s
        What this could mean overall is that the information in this training dataset is heavily reliant or based arounc
        factor for determining loan status, and this dataset is reflective of that. Our second most influential variable
        If someone wanted to maximize their likelihood of getting a loan they would want to have a credit history, prope
        As mentioned earlier no difference between urban and rural odds, however between semiurban and rural there is a
        Finally for credit history we have a statistically significant result since 1 isn't within the interval of (19.6
run;




/* Additional analysis */

title2 "Classification Performance of Final Model";

ods select none;
ods output Classification = CT_05;

proc logistic data=logistic_loan;
    class Credit_History (ref='0')
          Property_Area  (ref='Rural')
          Married2       (ref='No')
          / param=ref;

    model Loan_Status_Num(event='1') =
```

```
        Credit_History Property_Area Married2
        / ctable pprob=0.5;
run;

ods output close;
ods select all;

proc print data=CT_05 noobs label;
    title3 "Classification Table (Cutoff = 0.5)";
run;


ods select none;
ods output Classification = CT_CUT;

proc logistic data=logistic_loan;
    class Credit_History (ref='0')
          Property_Area  (ref='Rural')
          Married2       (ref='No')
          / param=ref;

    model Loan_Status_Num(event='1') =
          Credit_History Property_Area Married2
          / ctable pprob=(0.3 0.5 0.7);
run;

ods output close;
ods select all;

proc print data=CT_CUT noobs label;
    title3 "Classification Tables at Different Cutoffs (0.3, 0.5, 0.7)";
run;

proc odstext;
    p "To further evaluate the performance of our final logistic regression model, we examined its classification accura
    To better understand this tradeoff, we also examined alternative probability cutoffs of 0.3 and 0.7. When using a lo
run;

proc odstext;
    p "The code below has output that was created separate from the rest of the analysis above and serves to provide vis
run;


/*

ods output FitStatistics = fitstats;

proc sgplot data=fitstats;
    where Criterion = 'AIC';
    series x=Step y=InterceptAndCovariates / markers;
    xaxis label="Step in Selection Procedure";
    yaxis label="AIC";
    title "AIC Across Stepwise Logistic Regression";
run;

proc logistic data=loan;
    class Credit_History (ref='0')
          Property_Area  (ref='Rural')
          Married2       (ref='No')
          / param=ref;

    model Loan_Status_YN(event='Yes') =
          Credit_History Property_Area Married2;

    output out=diag_out
        pred=phat
        reschi=pearson_resid
        resdev=deviance_resid
        hat=leverage
        c=cooksd;
run;

proc sort data=diag_out;
    by descending cooksd;
run;

title "Top 10 Observations by Cook's Distance (Final Logistic Model)";
```

```
proc print data=diag_out(obs=10);
    var Loan_Status_YN phat pearson_resid deviance_resid leverage cooksd;
run;
title;

proc logistic data=logistic_loan plots(only)=roc;
    class Credit_History (ref='0')
          Property_Area  (ref='Rural')
          Married2       (ref='No')
          / param=ref;

    model Loan_Status_Num(event='1') =
          Credit_History Property_Area Married2;
run;

ods graphics off; */
```