|  |
| --- |
| Man is reaching for the car door handle. Car rental or car theft concept - temporary car insurance in Ontario |
| Car Insurance Claim Prediction  SAS project report |
| |  |  |  | | --- | --- | --- | | Albert Zihan Zeng | 1/6/24 | DSP SAS Project | |

**Contents**

[1. Introduction 2](#_Toc155366417)

[2. Data import 2](#_Toc155366418)

[3. Exploratory Data Analysis 2](#_Toc155366419)

[3.1. Univariate Analysis 2](#_Toc155366420)

[3.1.1. CRIDET SCORE 4](#_Toc155366421)

[3.1.2. Categorical Variables 6](#_Toc155366422)

[3.1.3. Annual Mileage 8](#_Toc155366423)

[3.2. Bivariate Analysis 8](#_Toc155366424)

[3.2.1. Categorical vs. Categorical Variables 8](#_Toc155366425)

[3.2.2. Continuous vs. Categorical Variables. 10](#_Toc155366426)

[4. Data Preparation 11](#_Toc155366427)

[4.1. Outlier Handling. 11](#_Toc155366428)

[4.2. Feature Engineering. 13](#_Toc155366429)

[4.3. Missing values in ANNUAL\_MILEAGE. 14](#_Toc155366430)

[4.4. Missing values in CREDIT\_SCORE. 17](#_Toc155366431)

[4.5. Split train and test datasets. 19](#_Toc155366432)

[5. Modeling 19](#_Toc155366433)

[5.1. Logistic Regression. 19](#_Toc155366434)

[5.2. Other Models. 21](#_Toc155366435)

[6. Model evaluation 21](#_Toc155366436)

[7. Conclusion 23](#_Toc155366437)

[8. Recommendations 23](#_Toc155366438)

[Appendix 23](#_Toc155366439)

# Introduction

This report investigates a yearly car insurance dataset, mixing real and made-up data to give a balanced picture of how customers behave. With 19 features, including 18 logs, we're exploring how people interact with their car insurance. Our main goal is to figure out how customers decide on insurance claims, marked by 1 for claims and 0 for no claims. The dataset's mix of real and synthetic info helps us understand customer actions better. Using advanced tools, we aim to find patterns in customer behavior and offer straightforward insights for the insurance industry. This report is about making sense of how people deal with car insurance, helping companies make better decisions.

# Data import

The dataset downloaded from Kaggle: <https://www.kaggle.com/datasets/sagnik1511/car-insurance-data/data>. This dataset has 10,000 observations and 19 columns. For importing the dataset in SAS, the following code has been used.

**proc** **import**

datafile="M:\Data Science class sharifa\SAS PROJECT\Group-Project\Car\_Insurance\_Claim.csv"

out=data0

dbms=csv replace;

getnames=Yes;

**run**;

**proc** **print** data=data0 (obs=**50**); **run**;

**proc** **contents** data=data0; **run**;

# Exploratory Data Analysis

In the exploratory data analysis (EDA) phase, we conducted univariate and bivariate analyses on key features. For numerical variables like 'Age,' 'Income,' and 'Speeding Violations,' we examined individual distributions using descriptive statistics and visualizations. Categorical variables such as 'Gender,' 'Marital Status,' and 'Outcome' underwent both univariate and bivariate analyses, exploring frequency distributions and relationships between variables. These analyses provide foundational insights, informing subsequent modeling and a deeper understanding of car insurance dynamics within the dataset.

## Univariate Analysis

The univariate analysis provides insightful information about the distribution and characteristics of key variables in our dataset. For numerical variables, such as 'ID,' 'Credit Score,' 'Annual Mileage,' 'Speeding Violations,' 'DUIs,' and 'Past Accidents,' descriptive statistics including mean, minimum, maximum, and standard deviation were calculated. Notably, variables like 'Annual Mileage' exhibited a diverse range of values, with a notable percentage of missing values.

On the categorical side, the distribution of variables like 'Age,' 'Gender,' 'Race,' 'Driving Experience,' 'Education,' 'Income,' 'Vehicle Ownership,' 'Vehicle Year,' 'Marital Status,' 'Children,' 'Postal Code,' 'Vehicle Type,' 'Speeding Violations,' 'DUIs,' 'Past Accidents,' and the target variable 'Outcome' was explored. These results provide a foundation for understanding the demographic and behavioral aspects of our dataset. For instance, most individuals fall into the age category of 26-39, with an equal split between genders. Additionally, most participants belong to the majority racial group, possess driving experience in the 0-9 years range, and have completed university education.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Univariate Analysis \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

**proc** **means** data=data0 n nmiss mean min q1 median q3 max stddev lclm uclm maxdec=**3**;

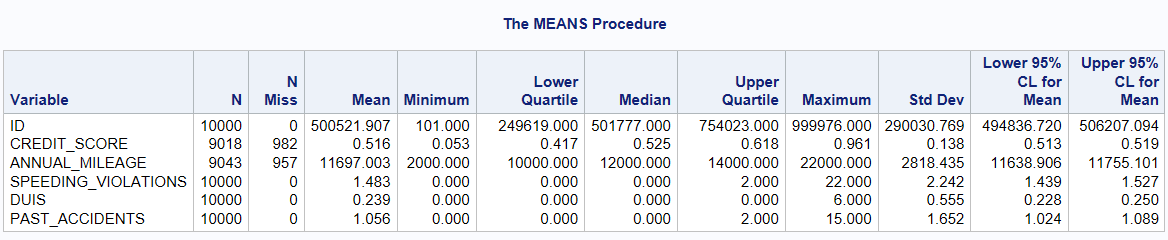
var id credit\_score annual\_mileage speeding\_violations duis past\_accidents;

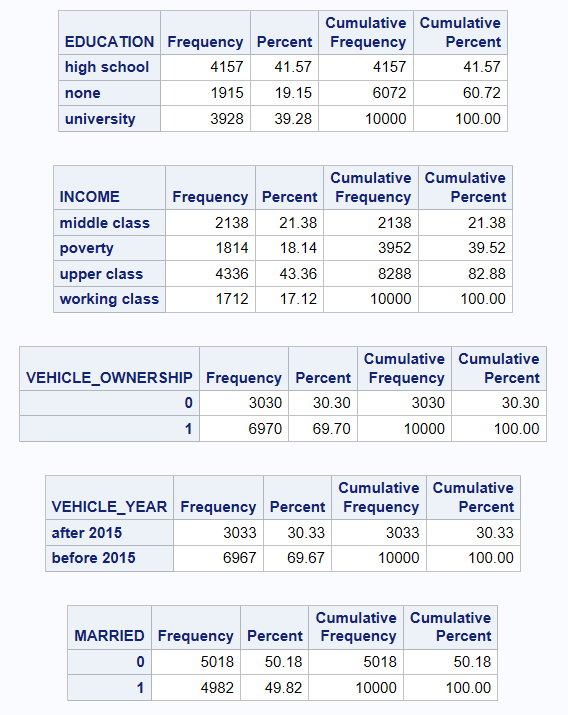
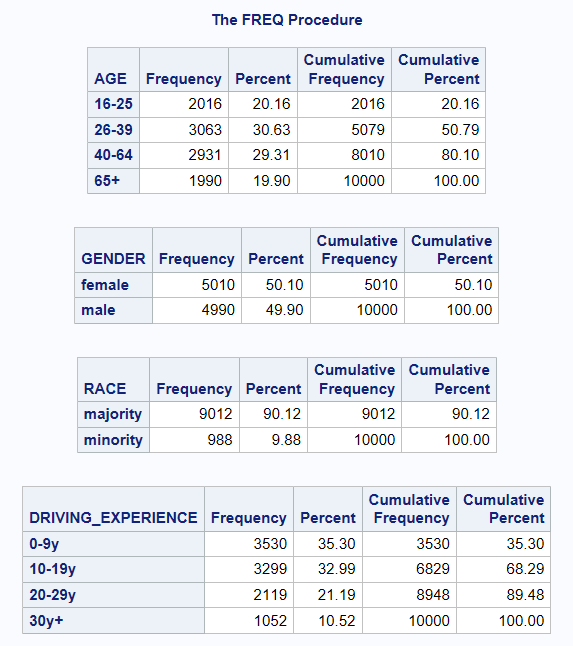
**run**;

**proc** **freq** data=data0 (drop=id credit\_score);

table \_all\_ /missing;

**run**;





### CRIDET SCORE

The histogram and kernel density plot of 'Credit Score' indicate a distribution resembling a normal curve. The visualization suggests that the majority of credit scores are concentrated around an average value, with fewer instances of extreme scores.

\*\*\*\*\* Credit Score;

**proc** **sgplot** data=data0;

title "Credit Score Distribution";

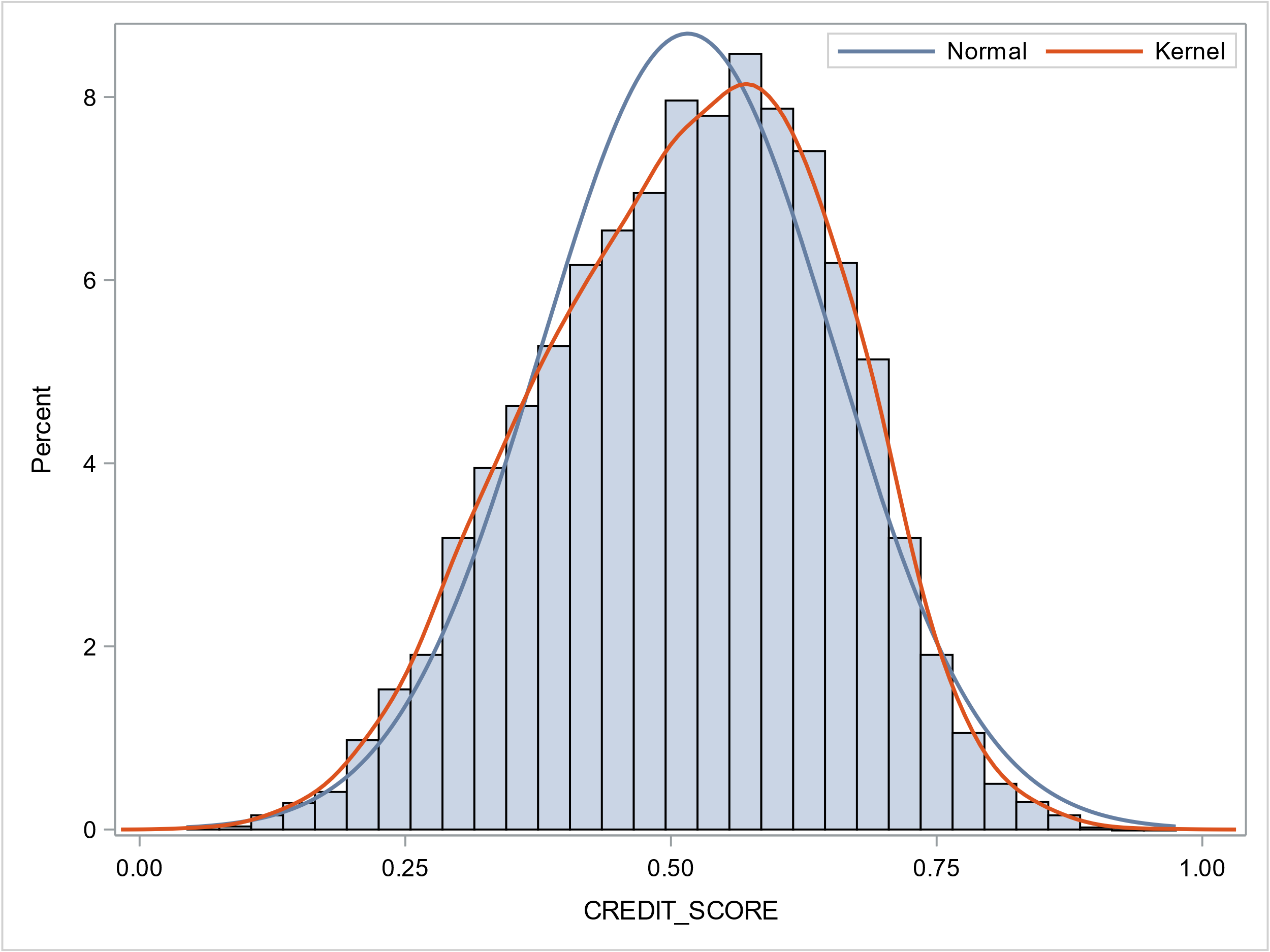
histogram credit\_score;

density credit\_score;

density credit\_score / type=kernel;

keylegend / location=inside position=topright;

**run**; title;



**proc** **sgplot** data=data0;

title "Credit Score by Outcome";

vbox credit\_score / category=outcome;

**run**; title;

**proc** **univariate** data=data0 (drop=id) normal plot;

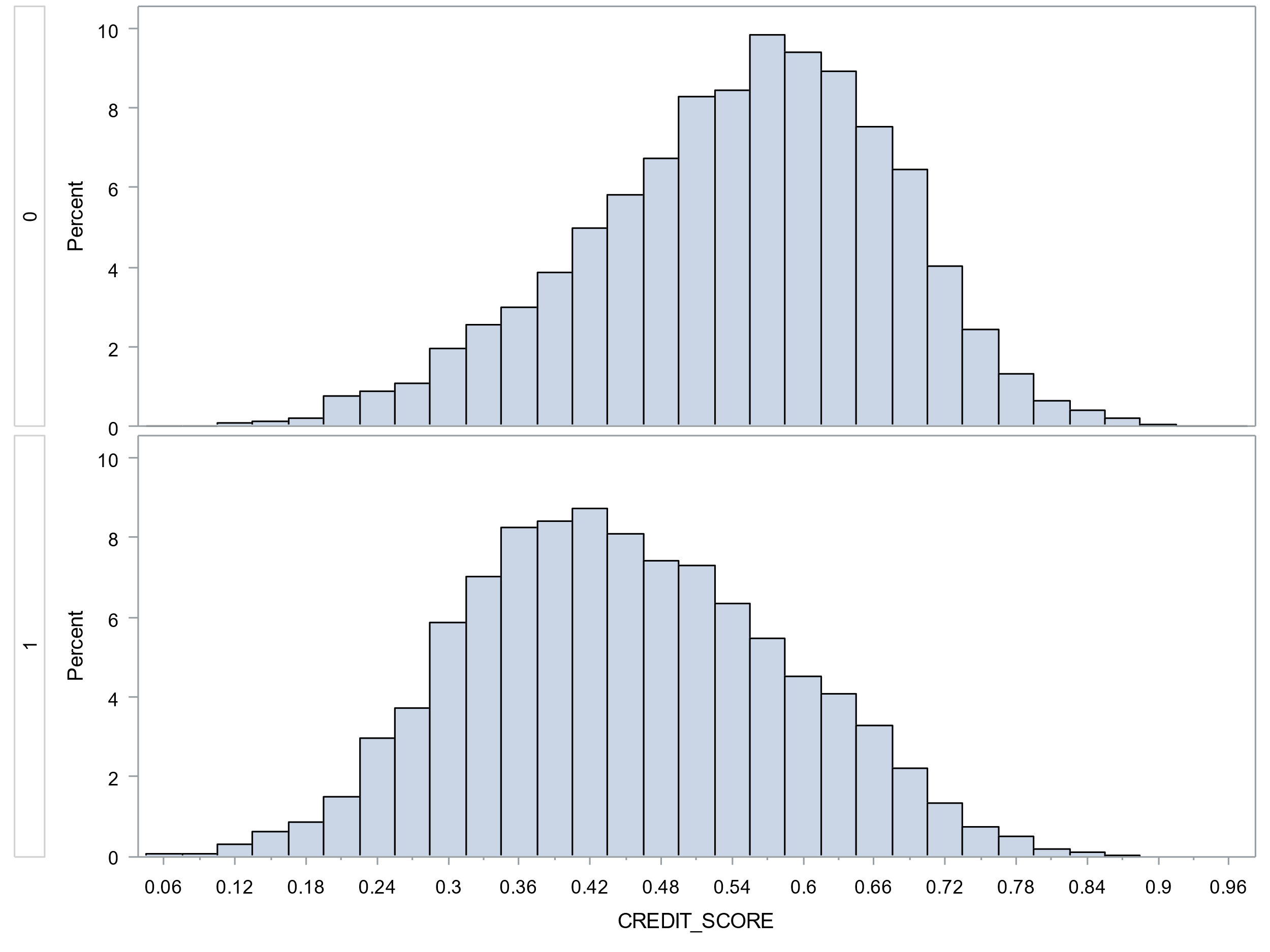
class outcome;

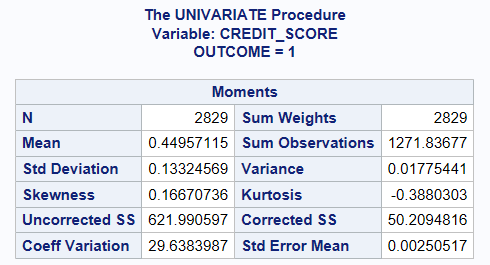
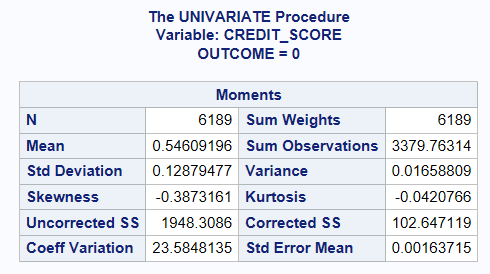
var credit\_score;

histogram credit score;

**run**;

The plots illustrate the distribution of 'Credit Score' based on two outcomes: 0 (no insurance claim) and 1 (insurance claim). For Outcome 0, the scores form a normal distribution around a mean of 0.546, indicating a balanced spread. In Outcome 1, representing insurance claims, the distribution shifts slightly with a lower mean of 0.450, suggesting a different pattern. These visuals offer a quick overview of how credit scores vary between the two outcomes.





### Categorical Variables

\*\*\*\*\* Categorical Variables.

\*store all categorical variables into list VAR\_CAT;

**proc** **sql** noprint;

select name into :VAR\_CAT separated by " "

from dictionary.columns

where libname = 'WORK' and memname = 'DATA0'

and name not in ('ID','CREDIT\_SCORE')

;**quit**;

%put &VAR\_CAT;

\*macro to plot each variable;

**%macro** univar\_plot(var,table=data0);

proc sgplot data=&table;

title "&var Distribution";

vbar &var / group=outcome groupdisplay=stack;

run; title;

**%mend**;

\*macro to loop over variables;

**%macro** univar(vlist/\*list by space\*/,table=data0/\*table name\*/);

%let nvar=%sysfunc(countw(&vlist));

%do i=**1** %to &nvar;

%let var=%scan(&vlist,&i);

%***univar\_plot***(&var,table=&table);

%end;

**%mend**;

|  |
| --- |
|  |

### Annual Mileage

**proc** **univariate** data=data0 normal;

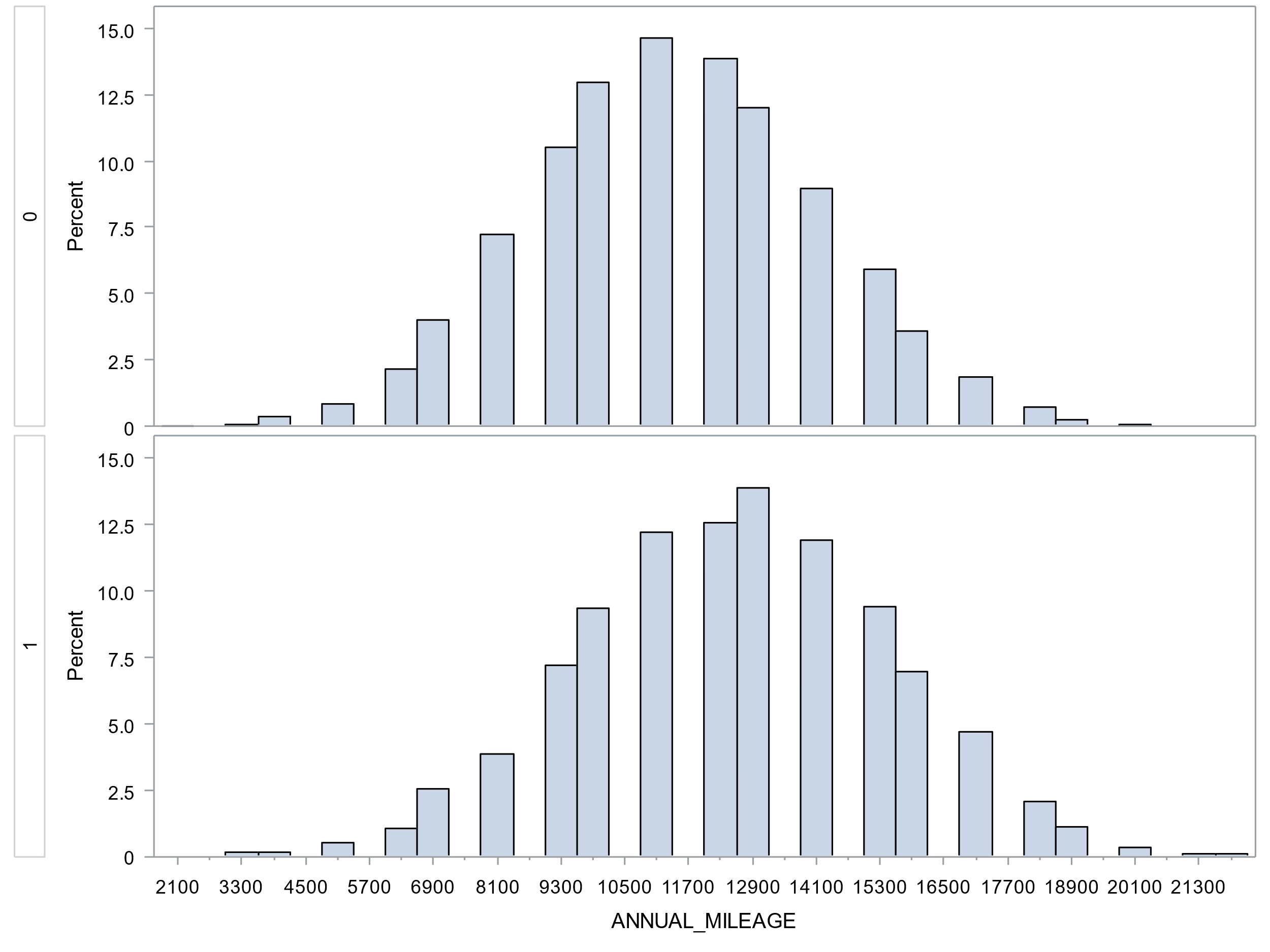
var annual\_mileage;

class outcome;

histogram annual\_mileage;

**run**;

The analysis of 'Annual Mileage' for insurance claim outcomes (0 and 1) reveals key patterns. For Outcome 0 (no claim), the average annual mileage is around 11,343, with a concentration between 9,000 and 13,000. Outcome 1 (claim) shows a higher mean mileage of approximately 12,483, with a broader range of 8,000 to 14,000. These findings suggest a potential correlation between higher annual mileage and the likelihood of insurance claims.



## Bivariate Analysis

### Categorical vs. Categorical Variables

**%macro** bivar\_chisq(var1,var2,table=data0);

proc freq data=&table;

title "Chi-Square Test for &var1 vs. &var2";

table &var1\*&var2 /chisq nopercent norow;

run; title;

**%mend**;

**%macro** bivar\_cat(var1,vlist=&var\_cat,table=data0);

%let nvar=%sysfunc(countw(&vlist));

%do i=**1** %to &nvar;

%let var2=%scan(&vlist,&i);

%if %sysfunc(prxmatch(s/^&var1$//,&var2))=**0**

%then %***bivar\_chisq***(&var1,&var2,table=&table);

%end;

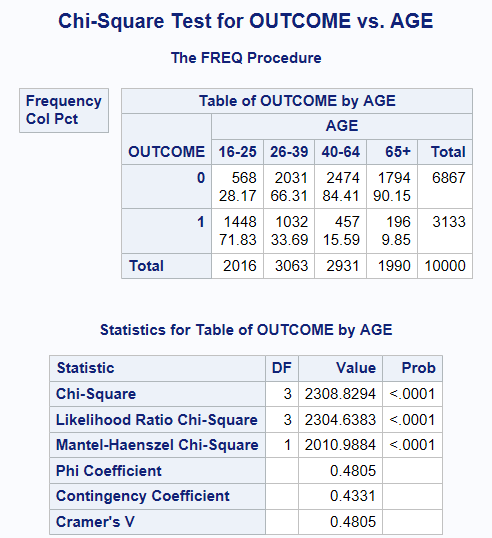
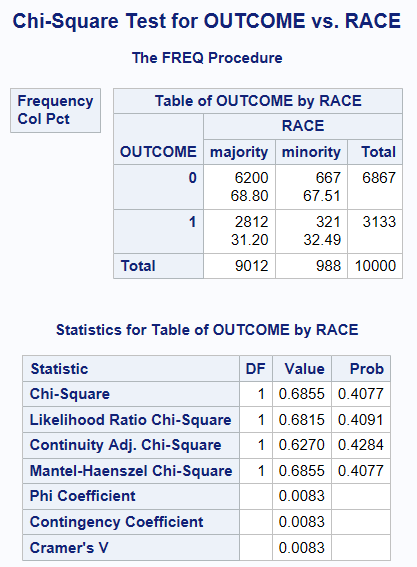
**%mend**;

%***bivar\_cat***(OUTCOME);

\* OUTCOME has no significant relationship with RACE by chi-square test;

%***bivar\_cat***(RACE);

%***bivar\_cat***(GENDER);

By running chi-square tests for pairs of categorical variables, significant associations were found between "OUTCOME" and categorical variables like AGE (p < 0.0001), GENDER (p < 0.0001), DRIVING\_EXPERIENCE (p < 0.0001), EDUCATION (p < 0.0001), and INCOME (p < 0.0001). These results indicate that these factors may influence insurance claim outcomes. However, the Chi-Square test did not reveal a significant relationship between "OUTCOME" and "RACE" (p = 0.4077), suggesting race may not substantially impact insurance claim outcomes in this analysis.

### Continuous vs. Categorical Variables.

%let VAR\_CAT\_BI =GENDER RACE VEHICLE\_OWNERSHIP VEHICLE\_YEAR

MARRIED CHILDREN VEHICLE\_TYPE OUTCOME;

%let VAR\_CAT\_MULTI =AGE DRIVING\_EXPERIENCE EDUCATION INCOME POSTAL\_CODE

ANNUAL\_MILEAGE SPEEDING\_VIOLATIONS DUIS PAST\_ACCIDENTS;

%let VAR\_CONT =CREDIT\_SCORE;

**%macro** bivar\_ttest(var1,var2,table=data0);

proc ttest data=&table;

class &var2;

var &var1;

title "T-test for &var1 by &var2";

run; title;

**%mend**;

**%macro** bivar\_anova(var1,var2,table=data0);

proc glm data=&table;

class &var2;

model &var1=&var2;

means &var2;

title "Anova Test for &var1 by &var2";

run; title;

**%mend**;

**%macro** bivar\_cont(var1,vlist2=&var\_cat\_bi /\*variable list for ttest\*/,

vlist3=&var\_cat\_multi /\*variable list for anova test\*/,table=data0,graph=off);

ods graphics &graph;

%if %length(&vlist2)=**0** %then %let nvar2=0;

%else %let nvar2=%sysfunc(countw(&vlist2));

%do i=**1** %to &nvar2;

%let var2=%scan(&vlist2,&i);

%***bivar\_ttest***(&var1,&var2,table=&table);

%end;

%if %length(&vlist3)=**0** %then %let nvar3=0;

%else %let nvar3=%sysfunc(countw(&vlist3));

%do j=**1** %to &nvar3;

%let var3=%scan(&vlist3,&j);

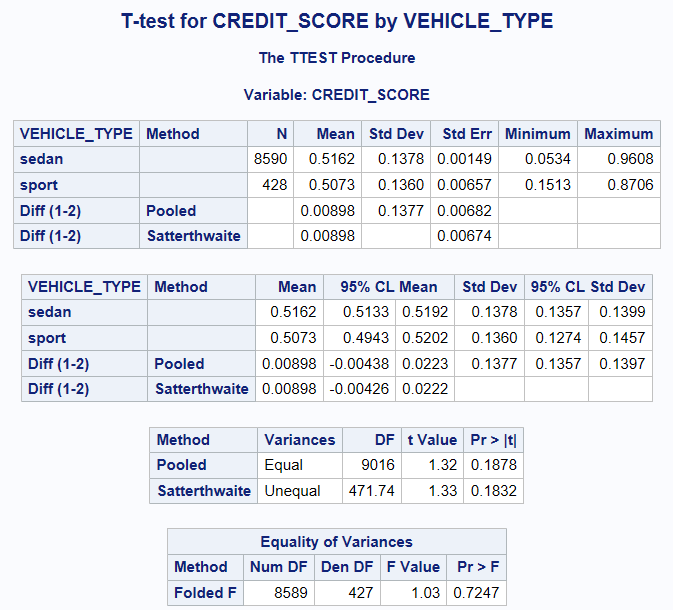
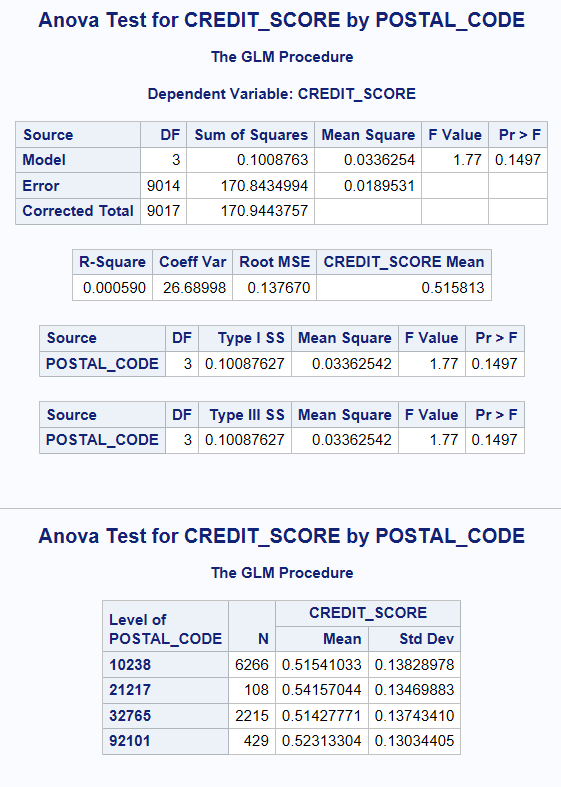
%***bivar\_anova***(&var1,&var3,table=&table);

%end;

ods graphics off;

**%mend**;

%***bivar\_cont***(CREDIT\_SCORE);

The only continuous variable "CREDIT\_SCORE" was tested against other categorical variables using either t-test or anova test depending on the number of levels of each variable. The result shows that "CREDIT\_SCORE" has no significant relationship with variables "VIHECLE\_TYPE" and "POSTAL\_CODE", with a t-test P-value of 0.1878 and anova P-value of 0.1497, respectively.

# Data Preparation

In this data preparation step, our primary objective is to address and enhance the dataset. We perform outlier handling and feature engineering. Due to the inaccuracy nature of the annual mileage data, we grouped annual mileage data into intervals as a new variable named "ANNUAL\_MILEAGE\_GRP". Additionally, we handle missing values by building simple logistic/regression models to understand the correlation with other variables and impute the missing observations by using the correlated variables.

## Outlier Handling.

%***univar***(DUIS PAST\_ACCIDENTS SPEEDING\_VIOLATIONS);

\* Reducing levels in DUIS PAST\_ACCIDENTS SPEEDING\_VIOLATIONS;

**data** data1;

set data0;

if duis>=**2** then DUIS\_GRP="2+";

else duis\_grp=duis;

if past\_accidents>=**4** then PAST\_ACCIDENTS\_GRP="4+";

else past\_accidents\_grp=past\_accidents;

if speeding\_violations>=**3** then SPEEDING\_VIOLATIONS\_GRP="3+";

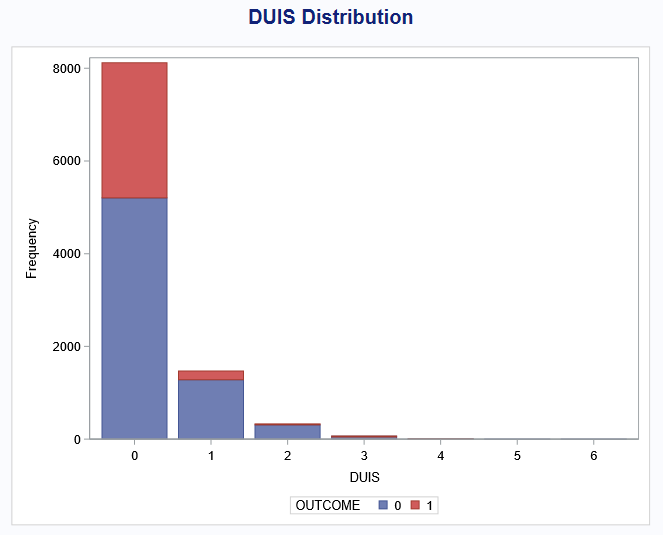
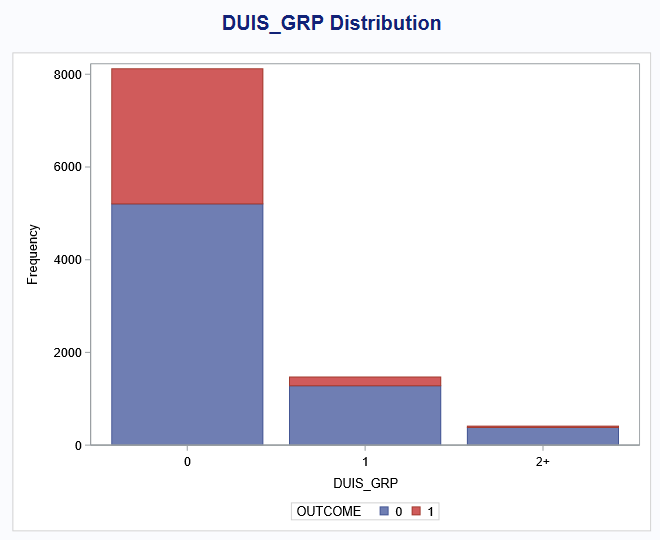
else speeding\_violations\_grp=speeding\_violations;

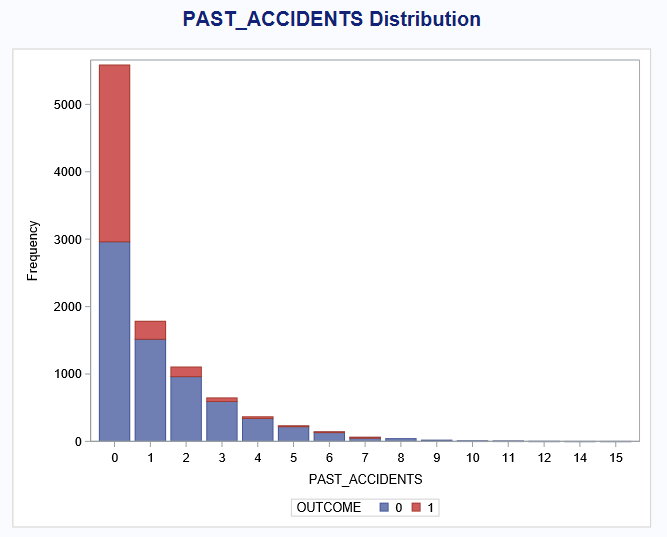
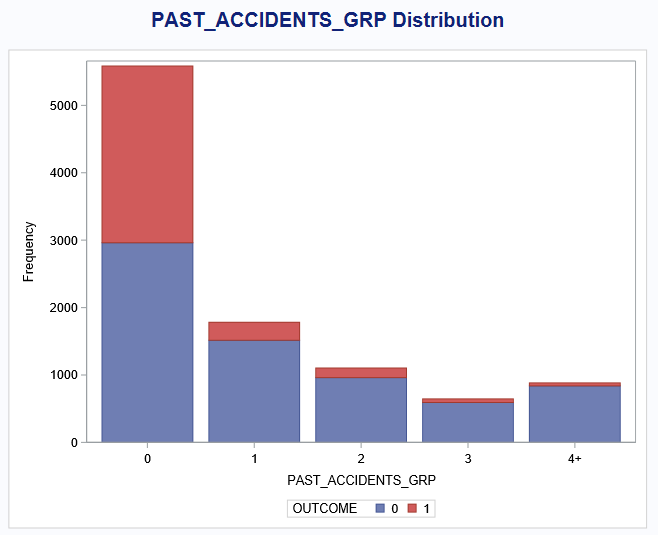
drop duis past\_accidents speeding\_violations;

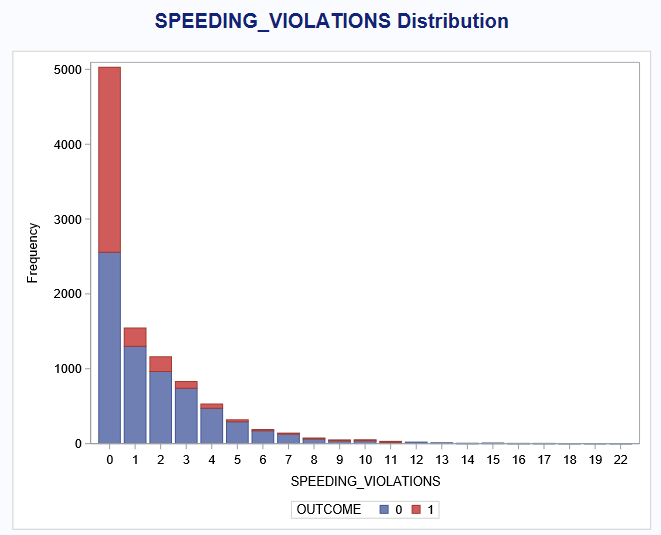
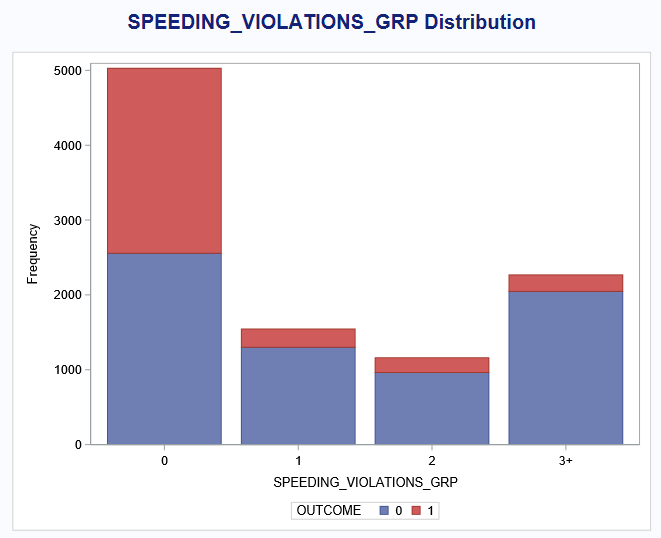
**run**;

**proc** **print** data=data1 (obs=**50**); **run**;

%***univar***(DUIS\_GRP PAST\_ACCIDENTS\_GRP SPEEDING\_VIOLATIONS\_GRP, table=data1);

## Feature Engineering.

\* Grouping ANNUAL\_MILEAGE;

**proc** **sql**;

create table data\_am as

select annual\_mileage, count(\*) as n, sum(outcome) as n\_claim,

sum(outcome)/count(\*) as claim\_rate

from data1

where annual\_mileage>**0**

group by annual\_mileage

order by annual\_mileage

;**quit**;

**proc** **print** data=data\_am; title "Claims by Annual Mileage";**run**;title;

**proc** **sgplot** data=data\_am;

title 'Claim Rate by Annual Mileage';

vbar annual\_mileage /response=claim\_rate;

**run**;title;

**data** data1;

set data1;

length ANNUAL\_MILEAGE\_GRP $**12**;

if missing(annual\_mileage) then ANNUAL\_MILEAGE\_GRP="";

else if annual\_mileage<=**8000** then ANNUAL\_MILEAGE\_GRP="0-8000";

else if annual\_mileage<=**12000** then ANNUAL\_MILEAGE\_GRP="9000-12000";

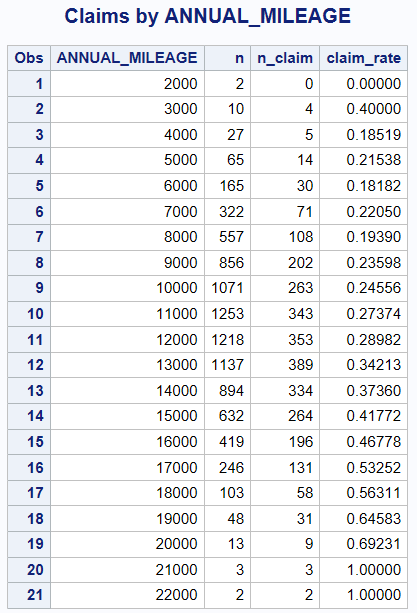
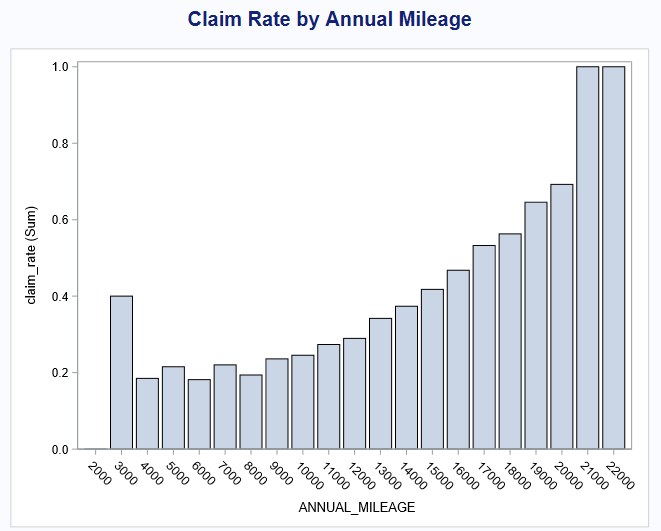
else if annual\_mileage<=**16000** then ANNUAL\_MILEAGE\_GRP="13000-16000";

else ANNUAL\_MILEAGE\_GRP="16000+";

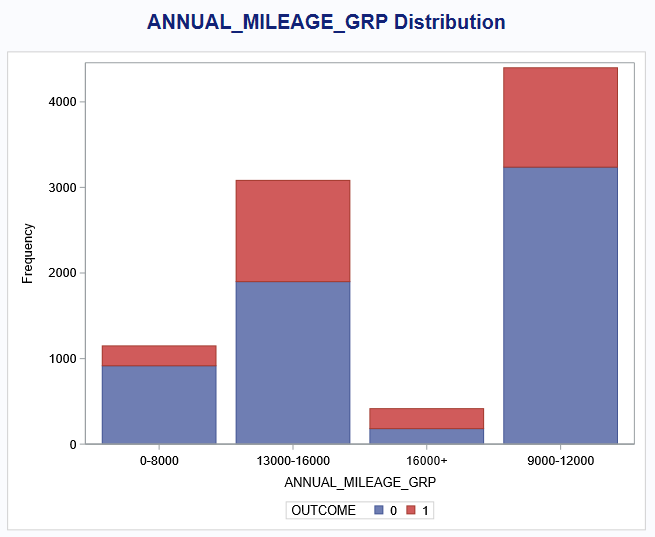
drop annual\_mileage;

**run**;

%***univar***(ANNUAL\_MILEAGE\_GRP, table=data1);

Based on the number of observations and the claim rate in each annual mileage, the annual mileage was grouped into 4 intervals, "0-8000" / "9000-12000" / "13000-16000" / "16000+".



## Missing values in ANNUAL\_MILEAGE.

**proc** **sql** noprint;

select name into :VAR\_CATX1 separated by " "

from dictionary.columns

where libname = 'WORK' and memname = 'DATA1'

and name not in ('ID','CREDIT\_SCORE','OUTCOME')

;**quit**;

%let VAR\_CATX1\_AM=%sysfunc(prxchange(s/ANNUAL\_MILEAGE\_GRP//,-1,&var\_catx1));

%put &var\_catx1;

%put &var\_catx1\_am;

\* check correlations;

ods graphics on;

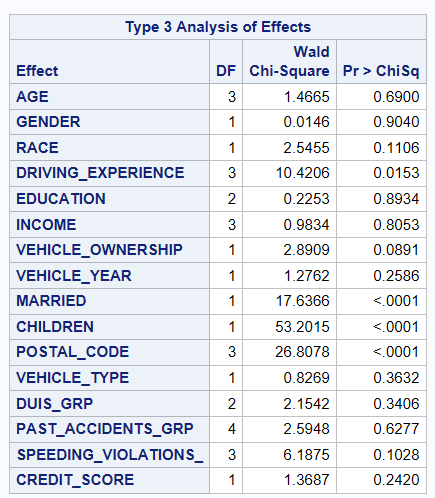
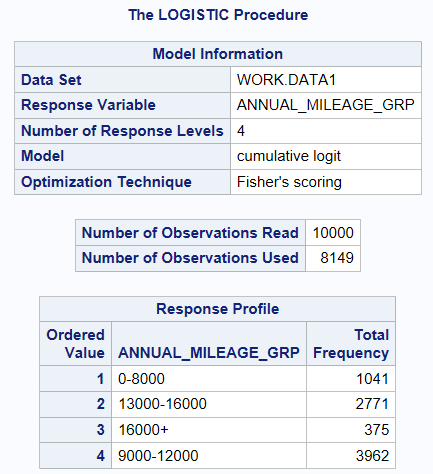
**proc** **logistic** data=data1 plots(only)=(effect oddsratio);

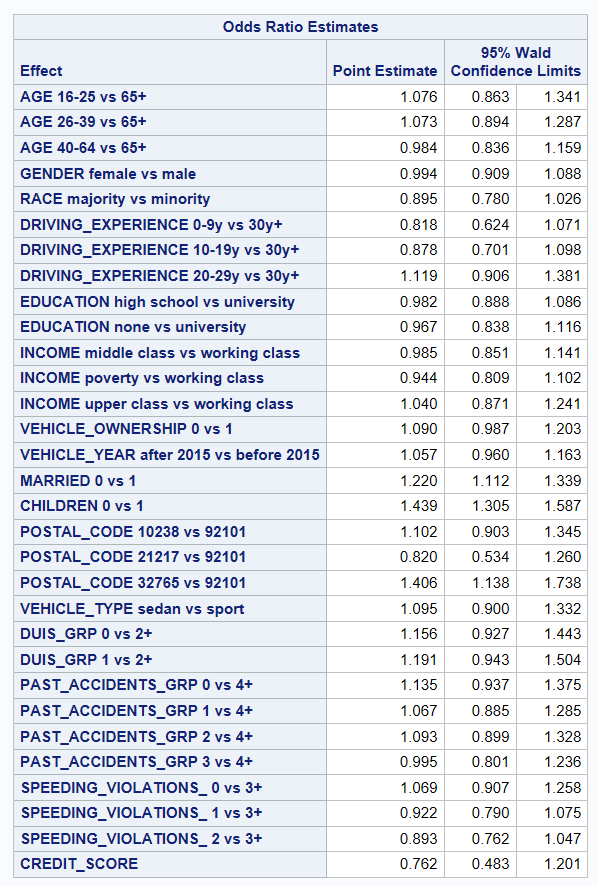
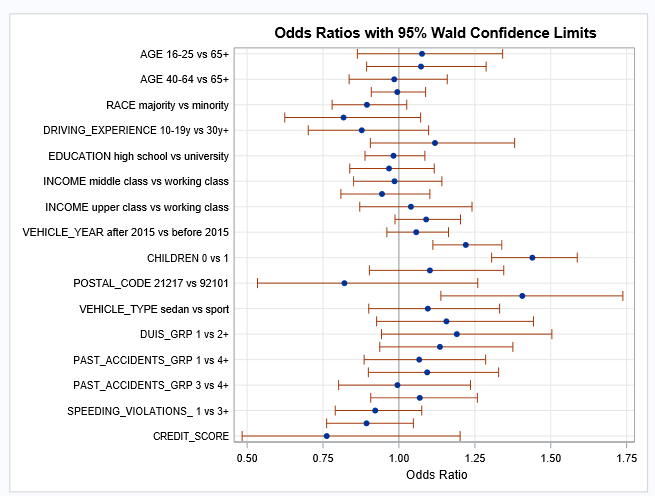
class &var\_catx1\_am;

model annual\_mileage\_grp = &var\_catx1\_am credit\_score;

**run**; **quit**;

ods graphics off;



From the logistic model for ANNUAL\_MILEAGE\_GRP, we can conclude that it is strongly correlated with variables DRIVING\_EXPERIENCE, MARRID, CHILDREN, and POSTAL\_CODE. We will then impute the missing values in annual mileage group by using these variables.

\* impute missing values;

ods select misspattern;

**proc** **mi** data=data1 nimpute=**1** out=data1\_imp1 seed=**42**;

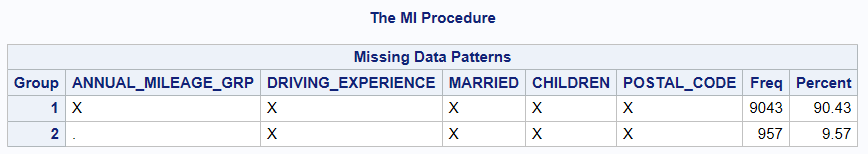
class annual\_mileage\_grp driving\_experience married children postal\_code;

var annual\_mileage\_grp driving\_experience married children postal\_code;

fcs logistic;

**run**;

ods select all;



\* result compare;

**proc** **freq** data=data1;

table annual\_mileage\_grp;

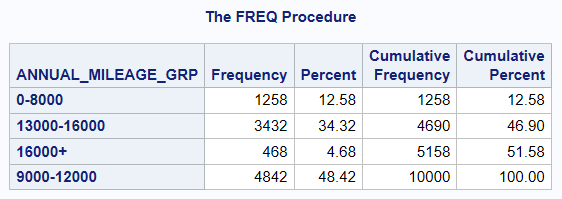
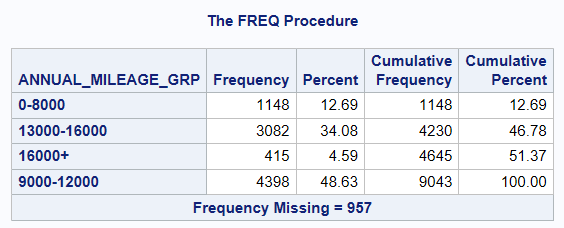
**run**;

**proc** **freq** data=data1\_imp1;

table annual\_mileage\_grp;

**run**;

%***univar***(annual\_mileage\_grp,table=data1\_imp1);



## Missing values in CREDIT\_SCORE.

\* check credit\_score correlations;

ods graphics on;

**proc** **glm** data=data1\_imp1;

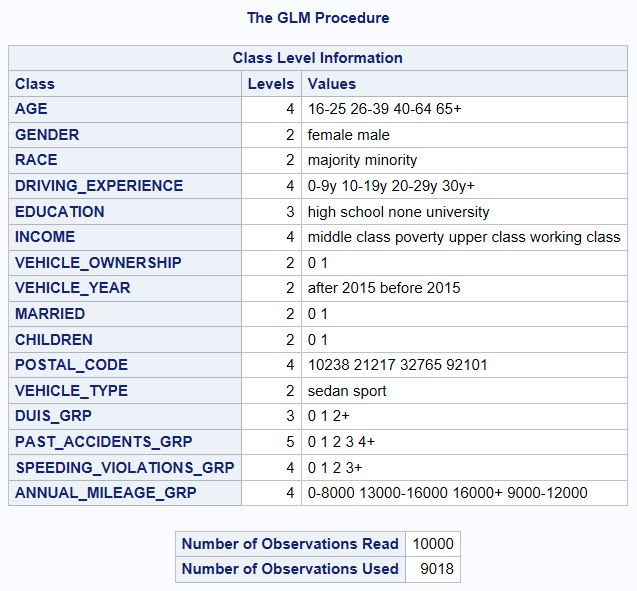
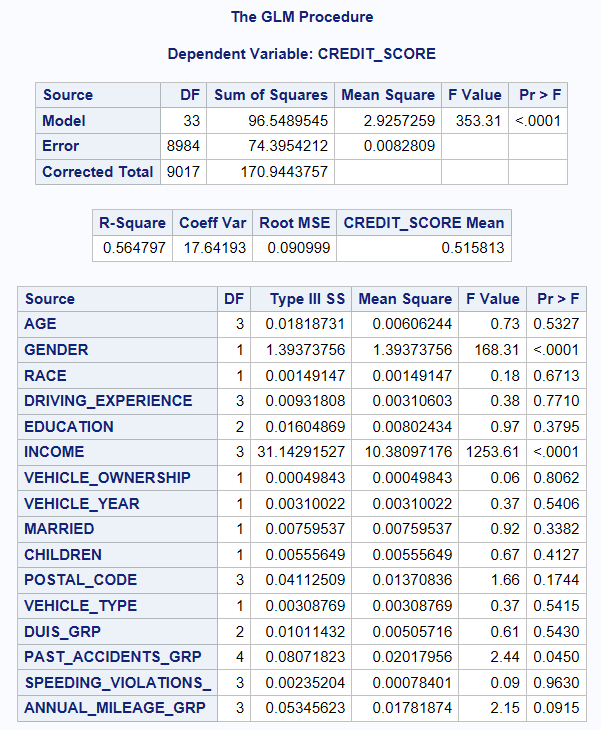
class &var\_catx1;

model credit\_score = &var\_catx1 / solution ss3;

\*lsmeans &var\_catx1 /pdiff stderr cl;

**run**; **quit**;

ods graphics off;

The GLM model indicates that credit score is highly correlated with GENDER, INCOME, and PAST\_ACCIDENTS\_GRP, with a P-value < 0.05. These variables will be used to impute the missing values in credit score.

\* impute missing values;

ods select misspattern;

**proc** **mi** data=data1\_imp1 nimpute=**1** out=data1\_imp2 seed=**42**;

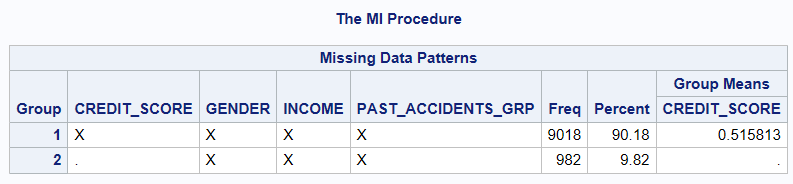
class gender income past\_accidents\_grp;

var credit\_score gender income past\_accidents\_grp;

fcs logistic regpmm;

**run**;

ods select all;



\* result compare;

**proc** **means** data=data1 missing;

var credit\_score;

**run**;

**proc** **means** data=data1\_imp2 missing;

var credit\_score;

**run**;

**proc** **univariate** data=data1\_imp2;

var credit\_score;

histogram credit\_score;

class outcome;

**run**;

|  |  |
| --- | --- |
|  |  |

## Split train and test datasets.

**proc** **sort** data=data1\_imp2;

by outcome;

**run**;

**proc** **surveyselect**

data=data1\_imp2 rate=**0.8** seed=**42**

out=surv\_result outall;

strata outcome;

**run**;

**data** traindata testdata;

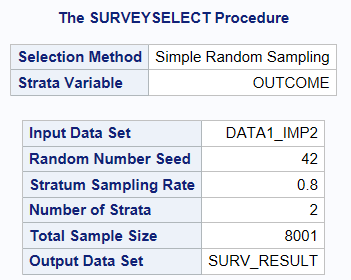
set surv\_result;

if selected=**1** then output traindata;

else output testdata;

drop selected;

**run**;



# Modeling

## Logistic Regression.

ods graphics on;

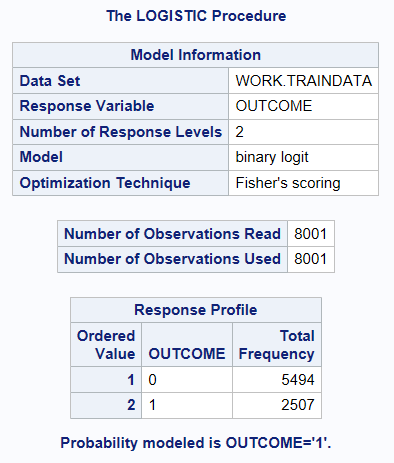
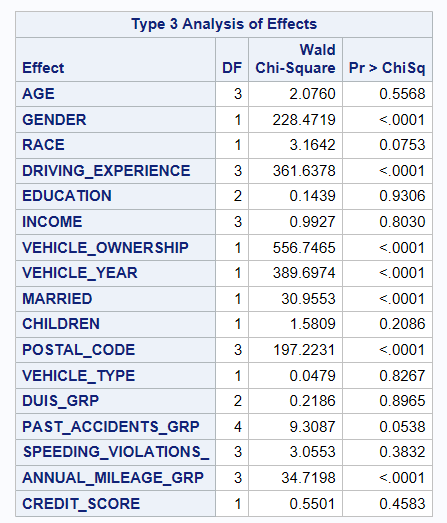
**proc** **logistic** data=traindata plots(only)=(effect oddsratio);

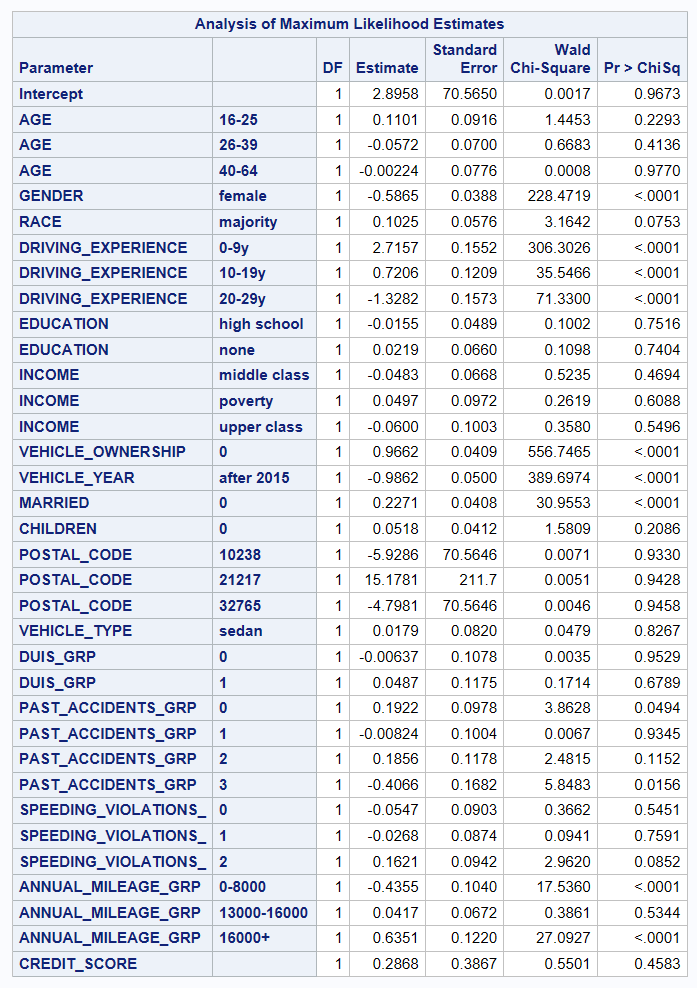
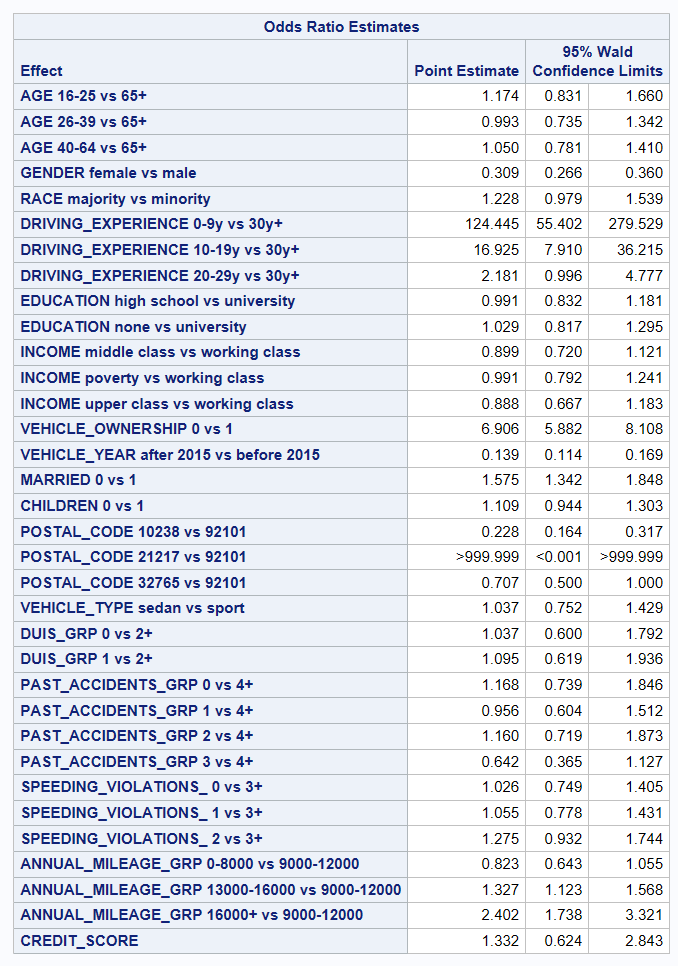
class &var\_catx1;

model outcome(event="1")=&var\_catx1 credit\_score;

**run**; **quit**;

ods graphics off;

A logistic regression model was used to predict if an event ("OUTCOME" being 1) would occur based on various factors. The model showed stable results, with certain factors like "GENDER," "DRIVING\_EXPERIENCE," and "VEHICLE\_OWNERSHIP" significantly impacting the outcome. Odds ratios help understand the likelihood of the event based on these factors. The model performed well, with high accuracy and reliable predictions. The predicted probabilities show how different factors influence the chance of the event happening.

## Other Models.

\*lasso;

ods graphics on;

**proc** **hpgenselect** data=traindata;

class &var\_catx1;

model outcome(event="1")=&var\_catx1 credit\_score /distribution=binary cl;

selection method=lasso (choose=SBC) details=all;

output out=output1 p=prodlasso;

**run**; **quit**;

ods graphics off;

\*gam;

ods graphics on;

**proc** **gam** data=traindata plots=all;

class &var\_catx1;

model outcome(event="1")=param(&var\_catx1 credit\_score) /dist=binomial;

score data=testdata out=output\_gam;

output out=outputdata\_gam p=prob\_predicted\_gam all;

**run**; **quit**;

ods graphics off;

# Model evaluation

**proc** **sort** data=testpred;

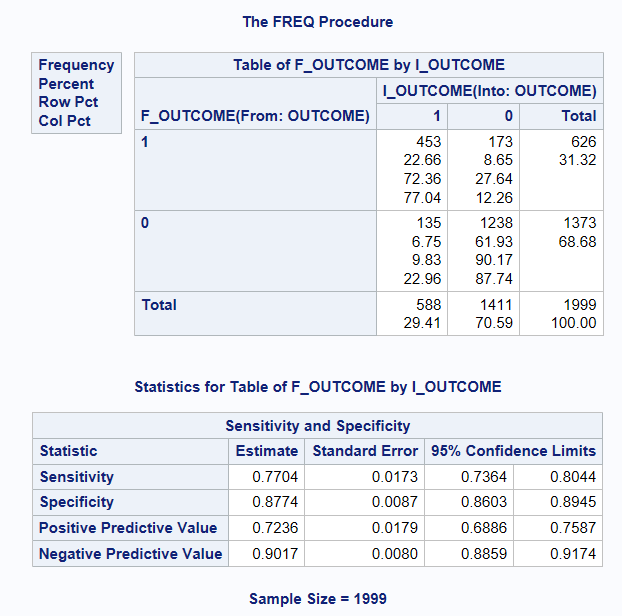
by descending i\_outcome descending f\_outcome;

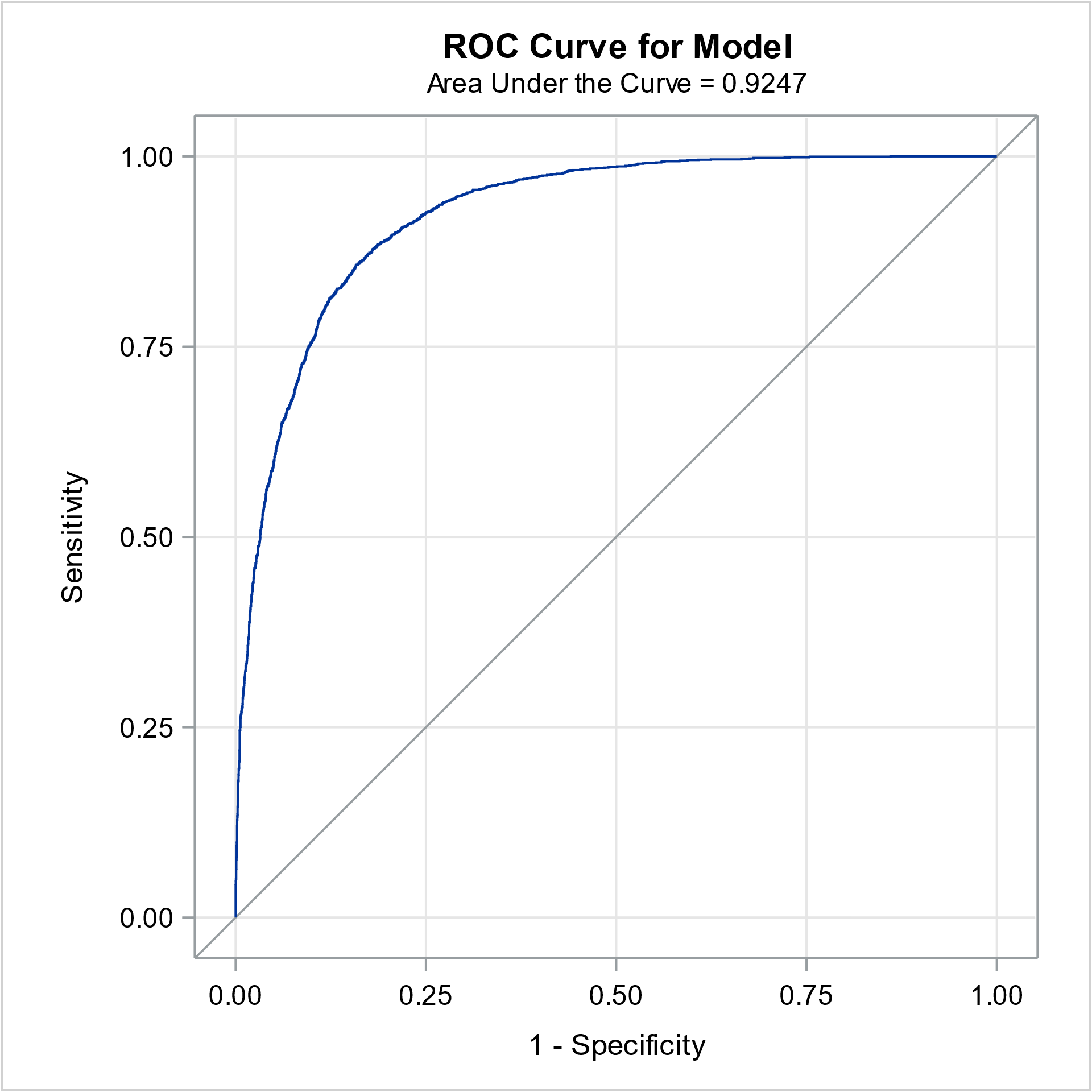
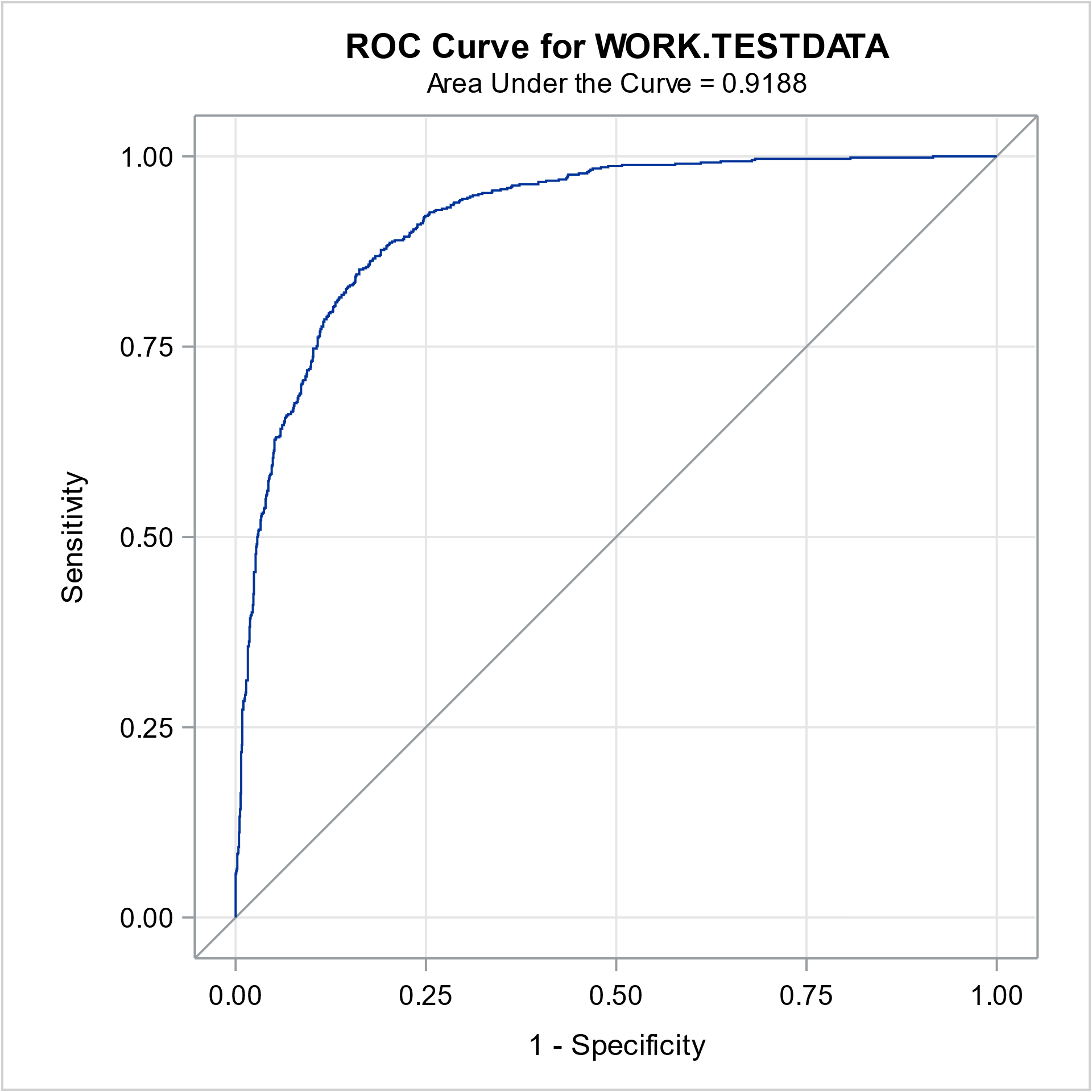
**run**;

**proc** **freq** data=testpred order=data;

tables f\_outcome\*i\_outcome /senspec;

**run**;



The logistic model gives an overall ROC-AUC score of 0.9247 on training data and 0.9188 on testing data, which indicates the model is highily accurate with no significant overfitting. The recall rate (sensitivity) on Outcome 1 class (insurance claimed) is 0.77, which indicates this logistic regression model can correctly predict 77% of the actual insurance claims.

# Conclusion

In conclusion, the analysis of your data set reveals meaningful insights into the factors influencing insurance claim outcomes. Age, gender, driving experience, and vehicle ownership emerged as significant predictors, demonstrating their impact on the likelihood of an insurance claim. Interestingly, race did not show a substantial relationship with claim outcomes. The logistic regression model achieved good predictive performance, providing a reliable understanding of the studied variables.

# Recommendations

1. **Focus on Age and Gender:** Pay extra attention to a person's age and gender when setting insurance rates and assessing risk. These factors strongly influence the likelihood of filing an insurance claim.
2. **Consider Driving Experience and Ownership:** Take into account a driver's experience and whether they own the vehicle. These aspects play a crucial role in predicting insurance claim outcomes and should be key factors in decision-making.
3. **Regularly Update the Model:** Keep your insurance prediction model up to date. The insurance landscape changes, so regularly monitor, and adjust your model to ensure it stays accurate and relevant for making informed decisions.

# Appendix

1. Dataset
2. SAS notebook
3. Project report
4. PowerPoint