



*Spring 2019*

**BUAN 6337: Predictive Analytics using SAS**  
*A Report of*  
**Group Project**

*Submitted by Group 7*

**Vinay Singh 2021441554**

**Vaibhav Shrivastava 2021434681**

**Megan Malisani 2021440151**

**Pragati Mishra 2021434655**

**Ishan Jain 2021426222**

**Erhao Liang 2021435949**

*Under the Guidance of*

**Prof. Shervin Shahrokhi Tehrani**

## Table of Contents

1. Executive Summary.....	3
2. Project Background.....	4
3. Data Description.....	5
4. Exploratory Data Analysis.....	6
5. Models and Analysis.....	12
6. Findings and Managerial Implications.....	18
7. Conclusion.....	18
8. SAS Code.....	19
9. References.....	26

## Executive Summary

Kiva is a microfinance organization working to alleviate poverty through microfinance loans. These loans are funded by individuals from around the world who lend in \$25 increments. Occasionally, the loan amount requested by the borrower is not fully funded by the lenders before the expiration date. Our research investigates factors associated with expired loans. Kiva wants to develop promotion and marketing strategies that minimize the proportion of loans that expire before obtaining full funding.

We have used predictive analytics techniques like linear and logistic, and probit regression to investigate this research question. We have also used exploratory data analysis to get a glimpse of the data which will validate our findings in the model building phase.

Through our initial consideration of the problem, we identified many questions we wanted to investigate further. A few of those questions include:

- Who is more likely to get their loan sponsored : Men or Women ?
- Which particular areas, countries or continents should we concentrate on for better results?
- What is average loan amount requested ?
- How does the number of lenders on a loan relate to its likelihood of expiring before funding?

We have used the dataset from Kaggle comprising of more than 1 million records about borrowers and their loans such as borrower country, continent, loan amount, and so on ("Data Science for Good"). We have used SAS to build model and do our exploratory data analysis.

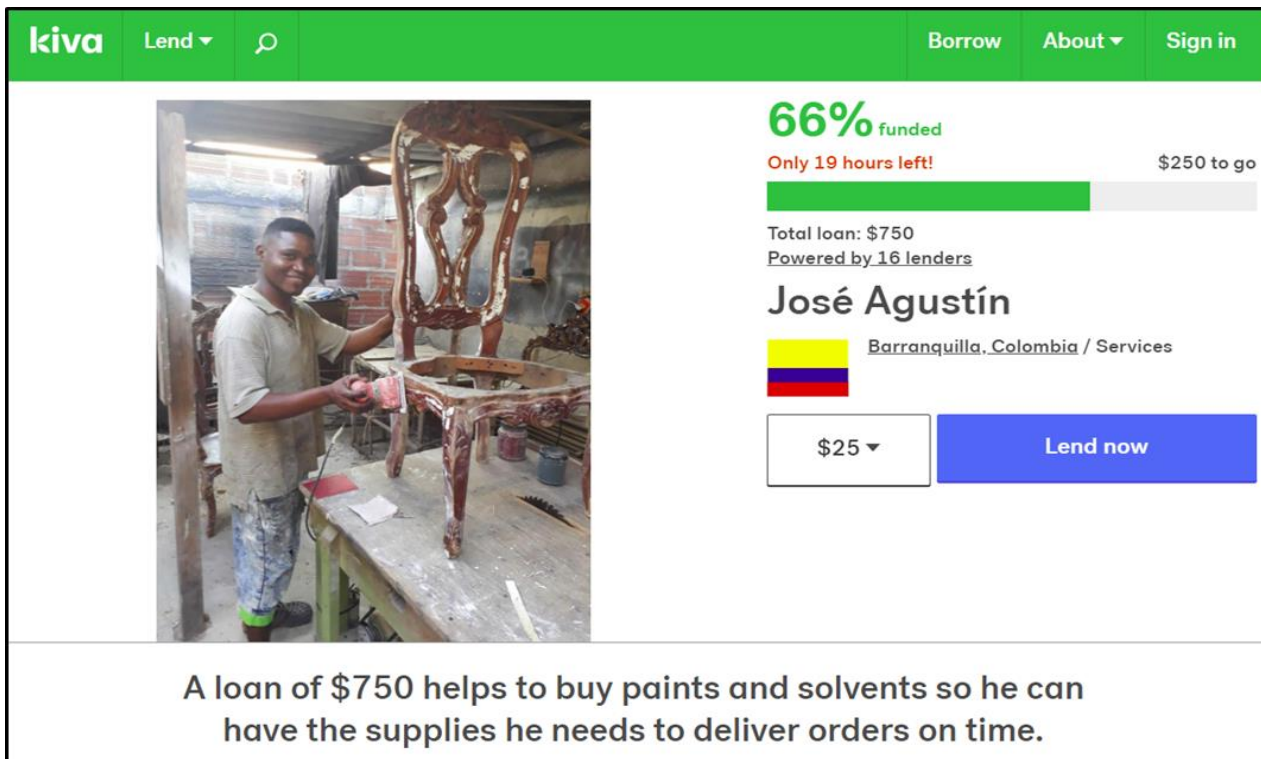
## Project Background

Our goal is to understand factors associated with Kiva loans that expire before receiving full funding. Before doing so, we must gain a deeper understanding of Kiva.

### What is Kiva?

- Kiva facilitates peer-to-peer microloans through a website platform.
- Kiva works with field partners around the world who have in-person contact with borrowers.
- Field partners work with borrowers to create loan requests and post them to the site.
- Kiva's lenders then donate money to fund specific loans.
- The borrower then pays the loan back over a period of months (or years).
- Loans are typically made to borrowers in poorer areas; however, some loans are also made to Americans.

### Kiva Website Layout:



The screenshot displays the Kiva website interface for a specific loan. At the top, a green navigation bar contains the Kiva logo, a 'Lend' dropdown menu, a search icon, and links for 'Borrow', 'About', and 'Sign in'. The main content area features a photograph of José Agustín, a man in a workshop, on the left. To the right of the photo, the loan details are shown: '66% funded' in green, 'Only 19 hours left!' in red, and '\$250 to go' in grey. A green progress bar is partially filled. Below this, it states 'Total loan: \$750' and 'Powered by 16 lenders'. The borrower's name 'José Agustín' is prominently displayed, followed by a Colombian flag icon and the text 'Barranquilla, Colombia / Services'. At the bottom of the loan details, there is a '\$25' dropdown menu and a blue 'Lend now' button. A white box at the bottom of the page contains the text: 'A loan of \$750 helps to buy paints and solvents so he can have the supplies he needs to deliver orders on time.'

## What if a Kiva Loan does not get funded?

- After 30 days, if the Kiva loan is not fully funded, it is removed from the site.
- Kiva's partner assumes the risk on the loan, which makes them less likely to work with Kiva in the future.
- For this reason, Kiva is striving to achieve a higher rate of full funding.

## Dataset Description

This data set contains information on a random sample of 1,016,534 Kiva loans and their repayment structures ("Data Science for Good"). The dataset has 12 variables. Each variable is explained below:

Variable Name	Type	Explanation
Loan ID	Categorical	Loan ID
Loan Amount	Numerical	Loan amount requested by borrower
Expired	Categorical	Expired=1 if a loan expires before receiving full funding. Expired=0 if a loan is fully funded.
Activity Name	Categorical	Activity for which loan is requested
Sector Name	Categorical	Sector for which loan is requested
Country code	Categorical	ISO country code of country in which loan was disbursed
Continent	Categorical	Continent in which the country belongs
Country Name	Categorical	Name of the country for the loan
Partner ID	Categorical	Field partner ID for local lending institutions
Month	Numerical	Duration of repayment period in months
Number of Lenders in Total	Numerical	The total number of lenders that contributed to this loan
Gender	Categorical	Gender of borrower(Male/Female)

## Exploratory Data Analysis

### Summary statistics of loan amount and loan duration (months):

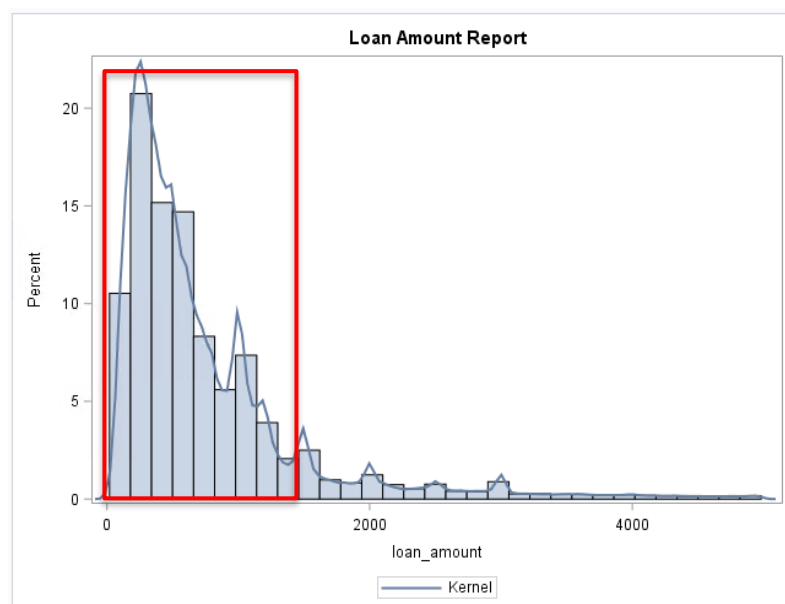
#### The MEANS Procedure

##### Expired=0

Variable	N	Mean	Std Dev	Minimum	25th Pctl	Median	75th Pctl	Maximum
loan_amount	969993	800.23	1048.42	25.00	275.00	500.00	950.00	100000.00
Month	969993	12.74	7.39	1.00	8.00	12.00	14.00	156.00
num_lenders_total	969993	22.45	28.17	1.00	8.00	15.00	27.00	2986.00

##### Expired=1

Variable	N	Mean	Std Dev	Minimum	25th Pctl	Median	75th Pctl	Maximum
loan_amount	46541	1497.20	1441.18	25.00	700.00	1050.00	1650.00	50000.00
Month	46541	18.37	8.11	3.00	14.00	15.00	21.00	145.00
num_lenders_total	46541	17.40	18.16	0.00	7.00	12.00	22.00	905.00



**Observation: Median of loan amount is \$500 and mean is \$800. Median of lending months is 12.**

## Summary statistics of gender and continent of borrower:

The FREQ Procedure

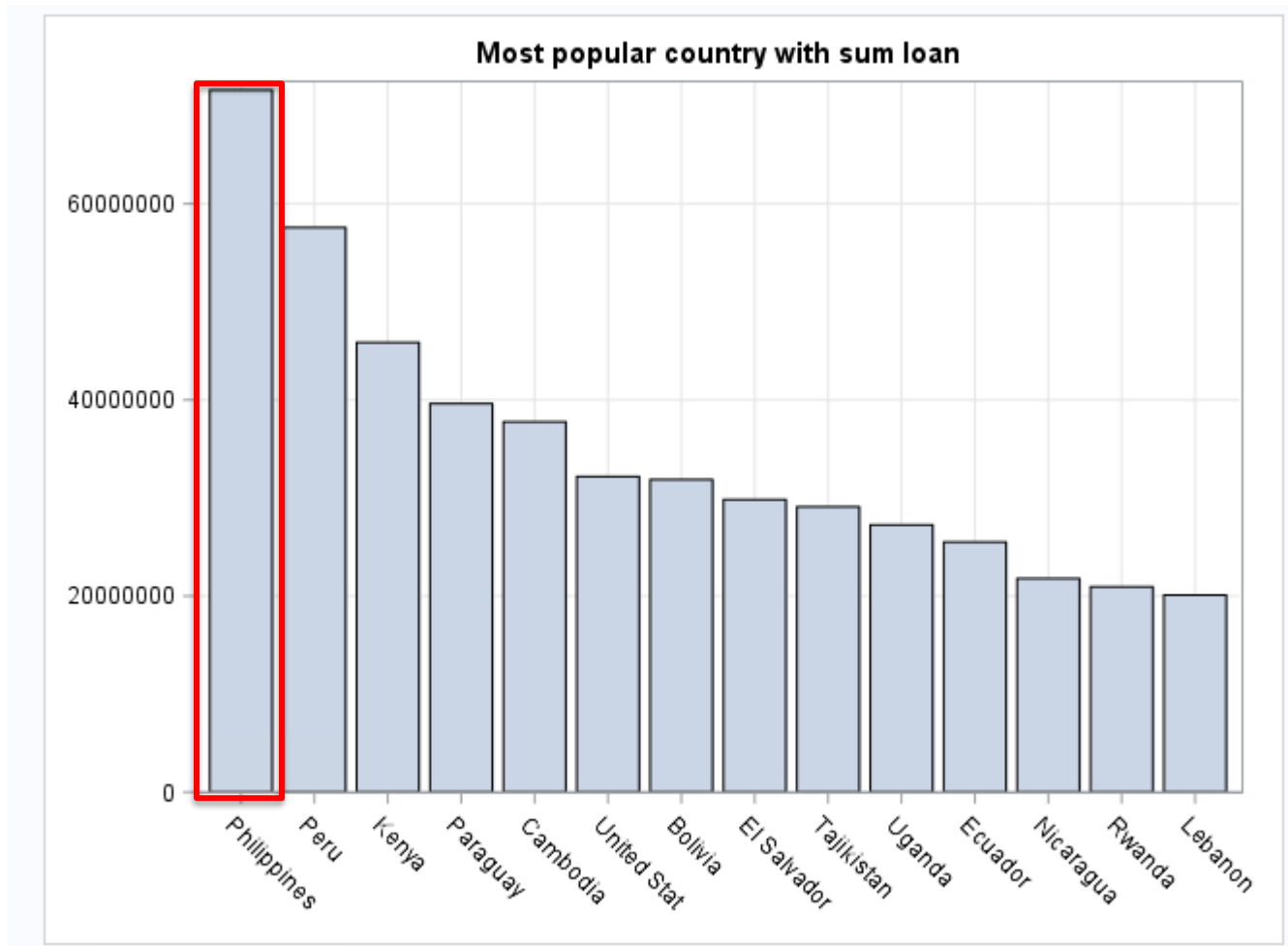
Expired	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	969993	95.42	969993	95.42
1	46541	4.58	1016534	100.00

Gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Group	148732	14.63	148732	14.63
female	646770	63.63	795502	78.26
male	221032	21.74	1016534	100.00

Continent	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Africa	269424	26.50	269424	26.50
Americas	273137	26.87	542561	53.37
Asia	452733	44.54	995294	97.91
Europe	8554	0.84	1003848	98.75
Oceania	12686	1.25	1016534	100.00

**Observation:** Number of loans from Africa and Asia is highest. In the dataset, it is shown that females apply for more loans than males.

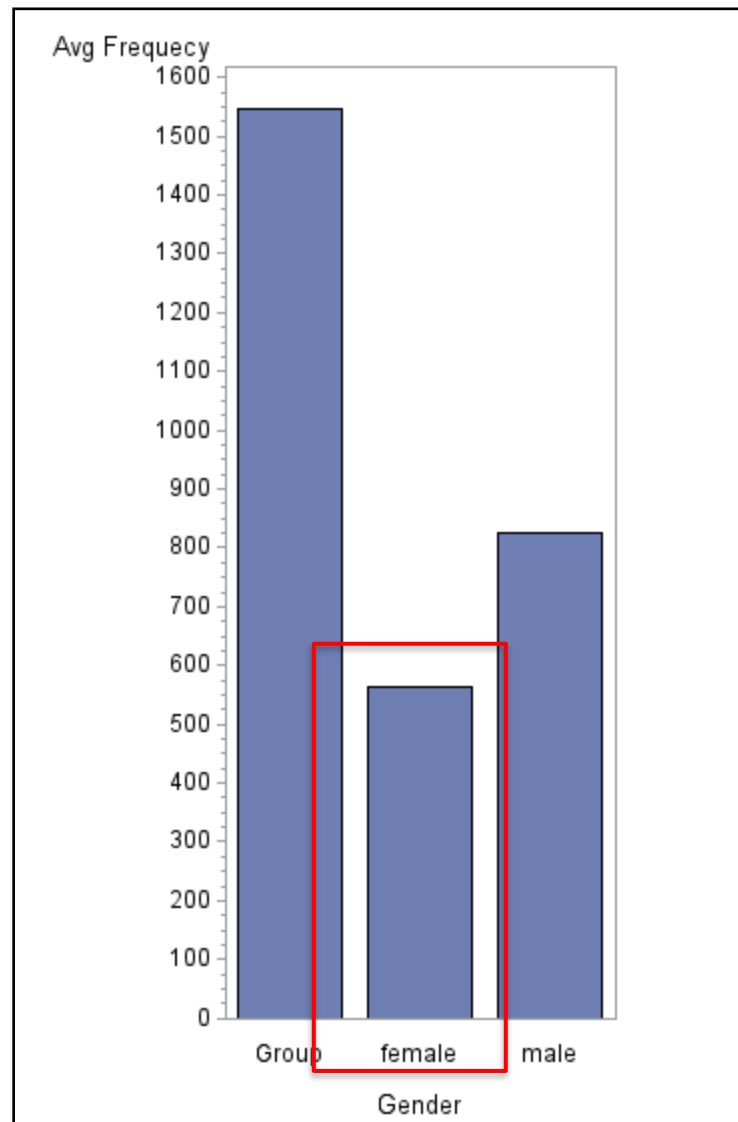
### Bar chart of most popular country by sum of loan amounts:



**Observation: Most popular country by loan sum amount is the Philippines.**

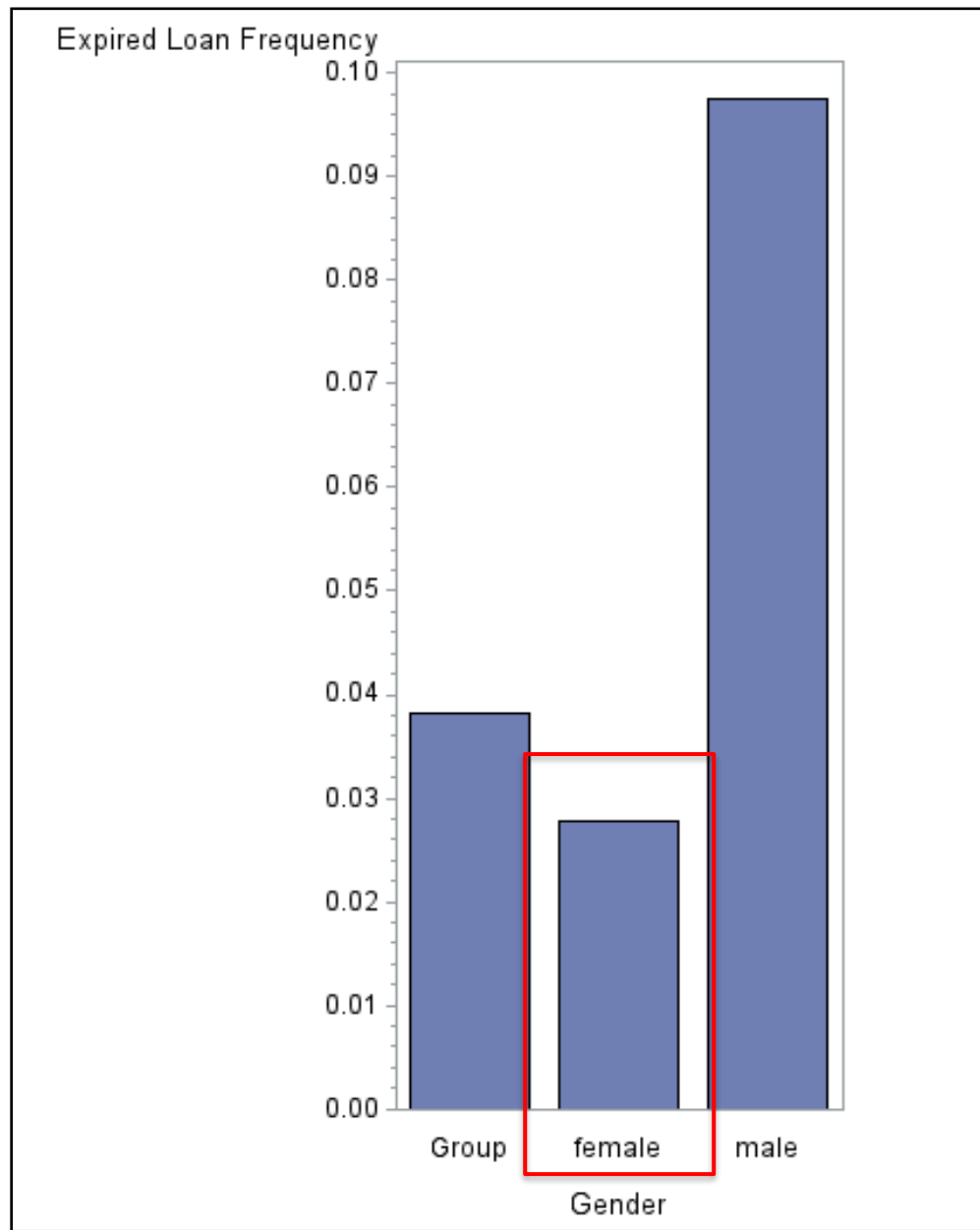


### Bar chart of average loan amount by gender or combined:



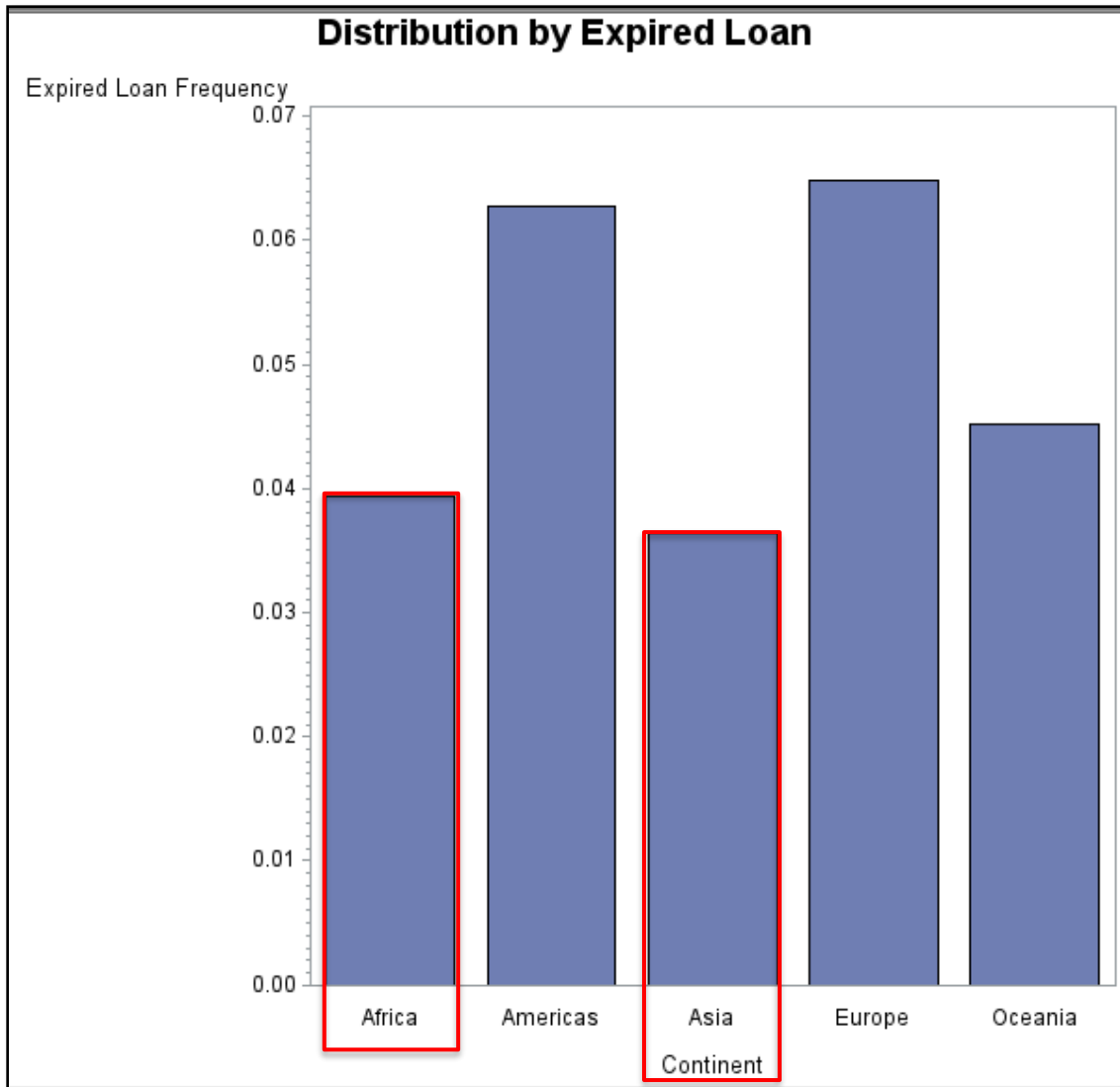
**Observation:** We can observe that groups request the highest average loan amount and individual females request the lowest average loan amount.

### Distribution of expired loan frequency



**Observation: Males have the highest frequency of expired loans and female have the lowest frequency of expired loans.**

### Distribution of expired loan frequency as per continent:



**Observation: Africa and Asia have lowest expired loan frequency**

### Conclusion of Exploratory Data Analysis:

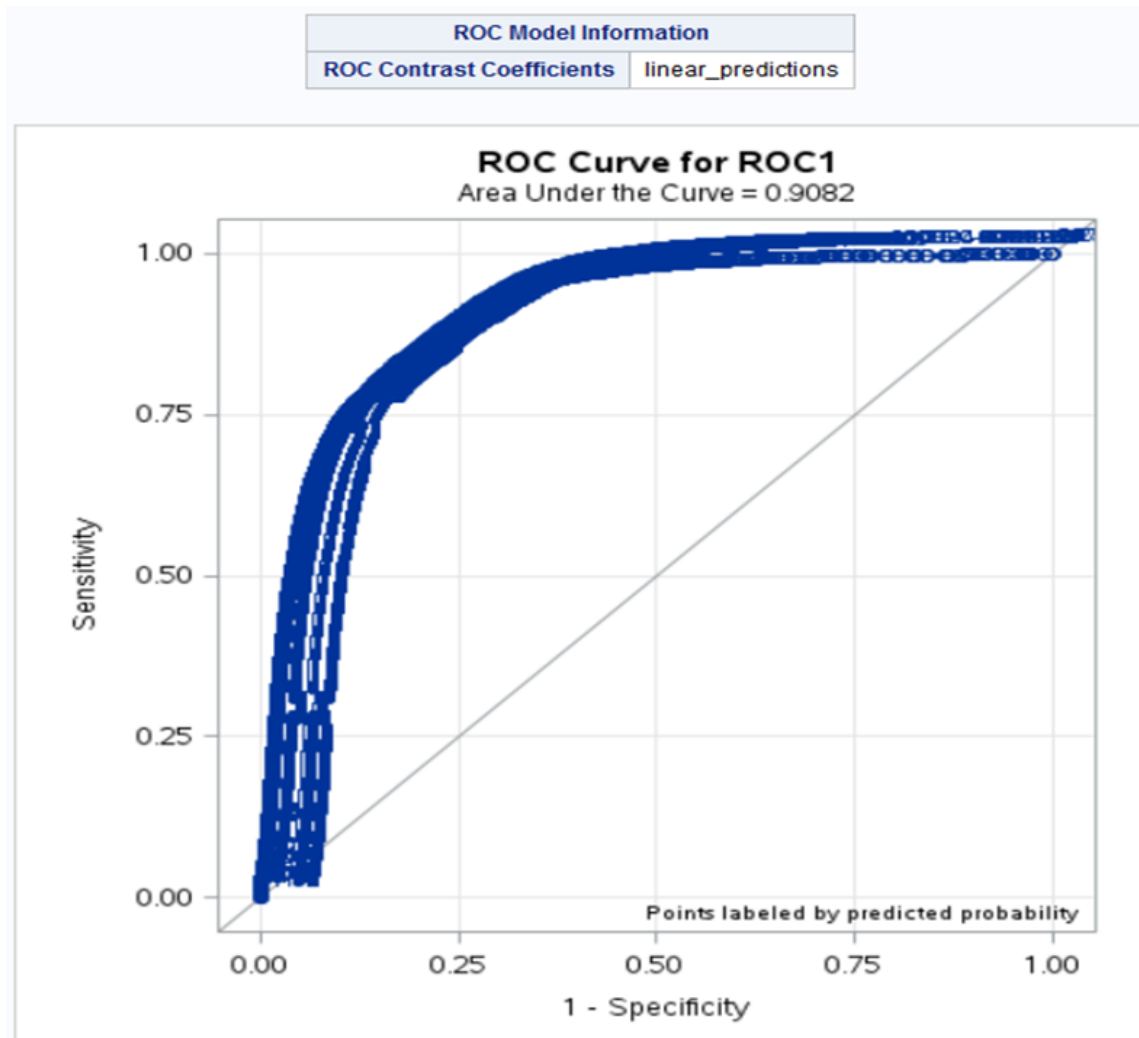
Title	Conclusion
Most loans by Continent	Africa and Asia
Most loan by Gender	Female
Lowest Average Amount by Gender	Female
Lowest Expiration Rate by Gender	Female
Lowest Expiration Rate by Continent	Africa and Asia

## 5. Models and Analysis:

### 5.1 Linear Regression:

- Requests for higher loan amounts are more likely to expire without full funding.
- Compared to males, females and groups have a lower possibility of loan expiration i.e. a higher chance of loan being funded from Kiva.
- Loans with longer repayment durations are more likely to expire.
- Loans from the continent of Asia have a higher chance of being fully funded, as compared with the Americas.
- Africa was not shown to be significant at the 5% level and has a coefficient estimate near 0. This indicates that loans from Africa aren't significantly more or less likely to be fully funded when compared to the Americas.
- Loans from Europe and Oceania are more likely to expire without full funding when compared to the Americas.

Parameter	Estimate		Standard Error	t Value	Pr >  t
Intercept	0.0399372331	B	0.00078487	50.88	<.0001
loan_amount	0.0001088676		0.00000041	265.76	<.0001
Gender Group	-.0612741697	B	0.00083861	-73.07	<.0001
Gender female	-.0556619249	B	0.00058897	-94.51	<.0001
Gender male	0.0000000000	B	.	.	.
num_lenders_total	-.0041043263		0.00001545	-265.69	<.0001
Month	0.0040920735		0.00003256	125.69	<.0001
Continent Africa	0.0003863586	B	0.00064120	0.60	0.5468
Continent Asia	-.0063660416	B	0.00057857	-11.00	<.0001
Continent Europe	0.0101980707	B	0.00257308	3.96	<.0001
Continent Oceania	0.0106539864	B	0.00212652	5.01	<.0001
Continent Americas	0.0000000000	B	.	.	.

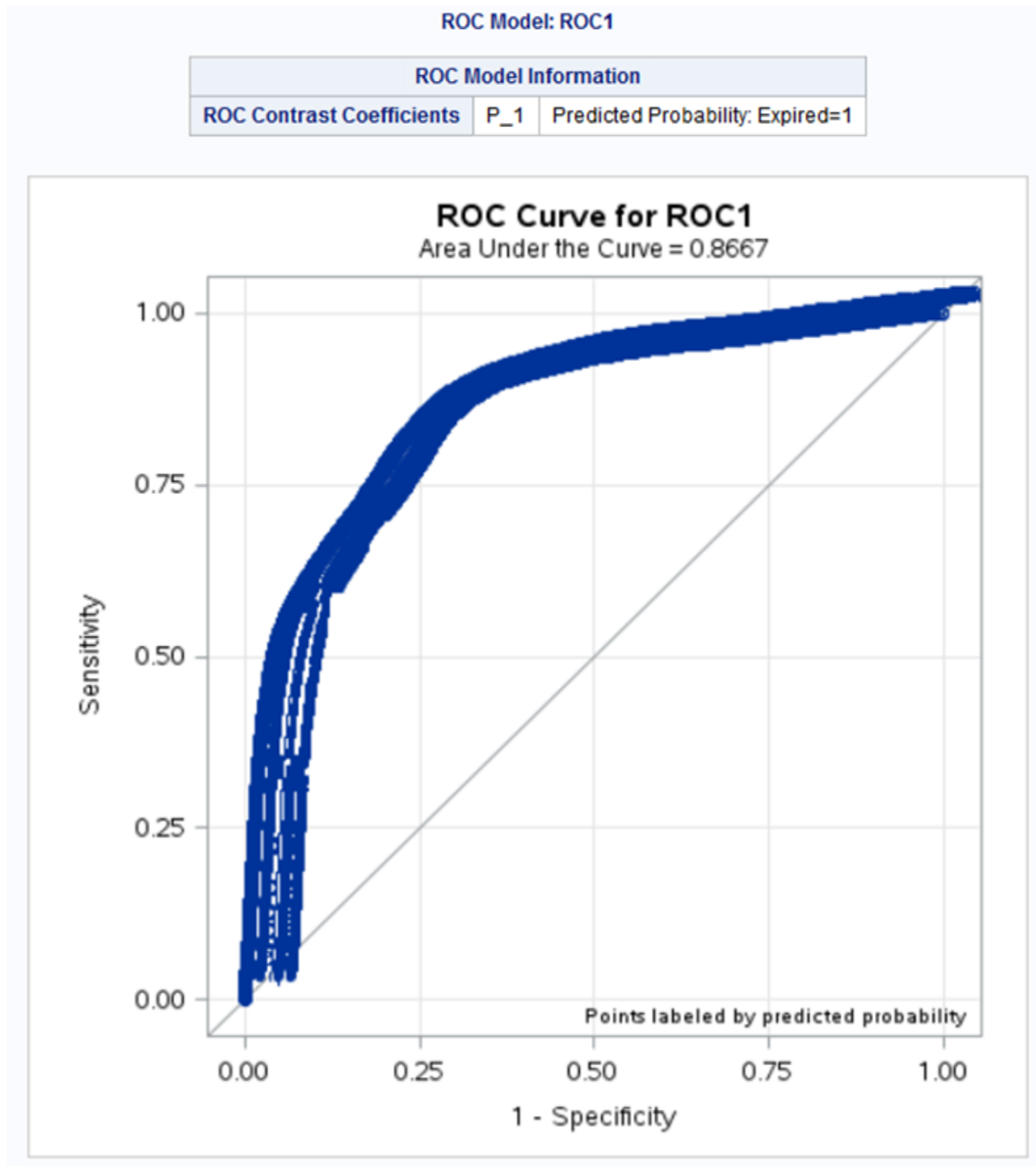
**ROC Curve of the Linear Regression Model: AUC= 0.9082**

## 5.2 Logistic Regression

Inferences from the logistic regression model are largely the same. The notable exception is that loans from Africa are now shown to be significantly more likely to receive funding when compared to the Americas at the 5% level.

- Requests for higher loan amounts are more likely to expire without full funding.
- Compared to males, females and groups have a lower possibility of loan expiration i.e. a higher chance of loan being funded from Kiva.
- Loans with longer repayment durations are more likely to expire.
- Loans from the continents of Africa and Asia have a higher chance of being fully funded, as compared with the Americas.
- Loans from Europe and Oceania are more likely to expire without full funding when compared to the Americas.

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-3.7315	0.0216	29845.6515	<.0001
loan_amount		1	0.00128	8.88E-6	20709.9108	<.0001
Gender	Group	1	-0.8194	0.0187	1924.2727	<.0001
Gender	female	1	-0.2067	0.0118	305.0749	<.0001
num_lenders_total		1	-0.0759	0.000559	18395.1891	<.0001
Month		1	0.0675	0.000604	12490.0599	<.0001
Continent	Africa	1	-0.2133	0.0195	120.0226	<.0001
Continent	Asia	1	-0.3001	0.0184	266.3412	<.0001
Continent	Europe	1	0.2693	0.0499	29.1210	<.0001
Continent	Oceania	1	0.3790	0.0444	72.8239	<.0001

**ROC Curve for Logistic Regression Model: AUC= 0.8667**

**Probability Table:**

The following table summarizes how the probability of receiving full funding changes given a 1 unit change in each factor. Estimated Probability is based on odds ratio of loan expiration and calculated probability of loan getting funded.

Factor	Estimates	Probability of expire loan	Probability of getting loan
Loan Amount	0.00128	1.001	-0.001
Group	-0.8194	0.441	0.56
Female	-0.2067	0.813	0.19
Number of Lenders	-0.0759	0.927	0.07
Month	0.0675	1.070	-0.07
Africa	-0.2133	0.808	0.19
Asia	-0.3001	0.741	0.26
Europe	0.2693	1.309	-0.31
Oceania	0.379	1.461	-0.46



### 5.3 Probit Model:

The probit model had the same inferences as logistic regression:

- Requests for higher loan amounts are more likely to expire without full funding.
- Compared to males, females and groups have a lower possibility of loan expiration i.e. a higher chance of loan being funded from Kiva.
- Loans with longer repayment durations are more likely to expire.
- Loans from the continents of Africa and Asia have a higher chance of being fully funded, as compared with the Americas.
- Loans from Europe and Oceania are more likely to expire without full funding when compared to the Americas.

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-3.7315	0.0216	29845.6515	<.0001
loan_amount		1	0.00128	8.88E-6	20709.9108	<.0001
Gender	Group	1	-0.8194	0.0187	1924.2727	<.0001
Gender	female	1	-0.2067	0.0118	305.0749	<.0001
num_lenders_total		1	-0.0759	0.000559	18395.1891	<.0001
Month		1	0.0675	0.000604	12490.0599	<.0001
Continent	Africa	1	-0.2133	0.0195	120.0226	<.0001
Continent	Asia	1	-0.3001	0.0184	266.3412	<.0001
Continent	Europe	1	0.2693	0.0499	29.1210	<.0001
Continent	Oceania	1	0.3790	0.0444	72.8239	<.0001

## 6. Findings and Implications:

- Females borrowing from Kiva have a higher chance that their loan gets funded.
- If a person is from developing countries in continents like Asia and Africa, their chance of loan getting funded is higher.
- If the loan amount and the number of months taken by a person to repay the amount increases, the chances of a loan being funded decreases.
- If the number of lenders increases, the chances of loan being funded also increases.

## 7. Conclusion and recommendation

- Kiva should focus its marketing strategy towards a women-centric campaign for effective utilization of resources.
- Kiva should understand that higher loan acceptance can likely be achieved by concentrating on Asian and African countries.
- Kiva should encourage borrowers to take shorter loans for smaller amounts. Imagine a borrower is working on a long-term farm expansion project. Rather than taking one large loan, they are more likely to be successful if they take an initial loan for fertilizer and seed, followed by a second loan for chickens once the first is repaid.

## 8. SAS Code:

```
LIBNAME Proj 'H:\My SAS Files\Project';

/* This imports the csv dataset into SAS. */
/* You can do it by using the "Import Data" option in File on the main menu */
PROC IMPORT OUT= Proj.Kiva_Loan
    DATAFILE= "H:\My SAS Files\Project\KIVA_Loans_Funding.csv"
    DBMS=CSV REPLACE;
    GETNAMES=YES;
    DATAROW=2;
RUN;

/* generating the working dataset in Work library */
data Kiva_Loan;
set Proj.Kiva_Loan;
run;

data Kiva_Loan_Hist;
set Proj.Kiva_Loan;
where loan_amount <5000;
run;

/*Summary Statistics for Expired Loan Numbers */
proc sort data = Kiva_Loan;
by Expired;
run;

/*Getting the summary statistics for loan amount, month, number of lenders in total*/
proc means data= Kiva_Loan n mean stddev min p25 median p75 max maxdec= 2;
var Loan_Amount month_num_lenders_total;
title 'Summary Statistics by gender';
by Expired;
run;

/*Categorical Variables : Summary Statistics*/
proc freq data= Kiva_Loan;
table Expired gender continent;
title 'Summary Statistics of categorical variables';
run;

/*Qualitative Variables : Summary Statistics*/
proc means data= Kiva_Loan n mean stddev min p25 median p75 max maxdec= 2;
var Loan_Amount month_num_lenders_total;
title 'Summary Statistics';
run;

/*Creating sql for Most Popular country */
```

```

proc sql;
Create table popular_country as
SELECT SUM(Loan_Amount) as Sum_Loan, AVG(Loan_Amount) as Avg_Loan, Country_name
from
Kiva_Loan
group by Country_name
order by Sum_Loan DESC;
run;

proc sql;
Create table popular_country_TOP as
SELECT * FROM POPULAR_COUNTRY WHERE Sum_Loan > 18234200;
run;

/*Creating the bar chart for most popular country with sum loan*/
title 'Most popular country with sum loan';
proc sgplot data=Popular_country_top;
vbar country_name / response=sum_loan
categoryorder=respdesc nostatlabel;
xaxis grid display=(nolabel);
yaxis grid discreteorder=data display=(nolabel);
run;

/* Loan Size Distribution */
/*Creating histogram for loan size distribution*/
proc sgplot data= Kiva_Loan_Hist;
histogram loan_amount / binstart = 100 binwidth = 50 ;
density loan_amount / type = kernel;
title 'Loan Amount Report';
run;

/* Proportion of Gender by mean amount */
Proc SQL;
CREATE TABLE AVG_GENDER AS
SELECT AVG(LOAN_AMOUNT) AS Average,GENDER FROM
Kiva_Loan_Hist
group by Gender;
RUN;
title 'Distribution by Gender';
axis1 label=('Gender');
axis2 label=('Avg Frequency');
format height width 50;

/* Create space at the bottom of the graph */
/*Creating bar chart for gender distribution by mean amount*/
footnote h=.01 in ' ';
proc gchart data=AVG_GENDER;
vbar gender / sumvar=average maxis=axis1 raxis=axis2;
run;

```

```

quit;

/* Proportion of Expired Loan by Gender */
Proc SQL;
CREATE TABLE EXPIRED_LOAN AS
SELECT COUNT(LOAN_ID) AS Count_Expired, Gender FROM
Kiva_Loan_Hist
where expired=1
group by Gender;
RUN;

CREATE TABLE EXPIRED_LOAN_Tot AS
SELECT Count(Loan_ID) AS TOTAL, Gender from
Kiva_Loan_Hist
group by Gender;
run;

create table proportion as
select A.Count_Expired/B.TOTAL AS Proportion,A.GENDER from
EXPIRED_LOAN AS A
INNER JOIN
EXPIRED_LOAN_Tot AS B
ON A.Gender = B.gender;
RUN;
title 'Distribution by Expired Loan';
axis1 label=('Gender');
axis2 label=('Expired Loan Frequency');
format height width 50;

/* Create space at the bottom of the graph */
/*Creating bar chart for gender distribution */
footnote h=.01 in ' ';
proc gchart data=Proportion;
vbar gender / sumvar=proportion maxis=axis1 raxis=axis2;
run;
quit;

/* Proportion of Expired Loan by Continent */
Proc SQL;
CREATE TABLE EXPIRED_LOAN_Cont AS
SELECT COUNT(LOAN_ID) AS Count_Expired, Continent FROM
Kiva_Loan_Hist
where expired=1
group by Continent;
RUN;

CREATE TABLE EXPIRED_LOAN_Tot_cont AS
SELECT Count(Loan_ID) AS TOTAL,Continent from
Kiva_Loan_Hist

```

```

group by Continent;
run;

create table proportion_cont as
select A.Count_Expired/B.TOTAL AS Proportion,A.Continent from
EXPIRED_LOAN_cont AS A
INNER JOIN
EXPIRED_LOAN_Tot_cont AS B
ON A.Continent = B.Continent;
RUN;
title 'Distribution by Expired Loan';
axis1 label=('Continent');
axis2 label=('Expired Loan Frequency ');
format height width 50;

/* Create space at the bottom of the graph */
footnote h=.01 in ' ';
proc gchart data=Proportion_cont;
vbar continent / sumvar=proportion maxis=axis1 raxis=axis2;
run;
quit;

/* Model Building */
/* choosing the 70% of sample. Seed = 2 will help you have same random samples
if you repeat the analysis */
/* Create training and test datasets. 70% of sample in training */
proc surveyselect data=Kiva_Loan out=Kiva_sampled outall samprate=0.7 seed=2;
run;

data Kiva_training Kiva_test;
set Kiva_sampled;
if selected then output kiva_training; /* Tell SAS that only keep the 70%
selected one in sample. The rest will be in test data */
else output kiva_test;
run;

/* Linear probability model using linear regression */
proc glm data=kiva_sampled ;
class Continent(ref='Americas') Gender(ref='male');
model Expired = loan_amount gender num_lenders_total month continent /solution;
weight selected;
output out=kiva_lin_predict p=linear_predictions; /*only training sample is used
for estimation, since selected=0 for test sample */
run;
quit;

/* To plot ROC curve based on predictions from linear model */
proc logistic data=kiva_lin_predict plots=roc(id=prob);

```

```

class Continent(ref='Americas') Gender(ref='male');
logit:model Expired (event='1')= loan_amount gender num_lenders_total month
continent/nofit;
roc pred=linear_predictions;
roc pred=linear_predictions;
where selected=0;
run;

/* Logistic Regression */
proc logistic data=kiva_sampled ;
class Continent(ref='Americas') Gender(ref='male');
logit: model Expired (event='1')= loan_amount gender num_lenders_total month
continent;
weight selected; /*only training sample is used for estimation, since selected =
0 for test sample */
run;
quit;

/* Logistic regression */
/* Make predictions on test data */
proc logistic data=kiva_training ;
class Continent(ref='Americas') Gender(ref='male');
logit: model Expired (event='1')= loan_amount gender num_lenders_total month
continent;
score data=kiva_test out=kiva_logit_predict; /* predictions are made only for
the dataset specified*/
run;

ods graphics on;
/*ROC curve on test data */
proc logistic data=kiva_logit_predict plots=roc(id=prob);
class Continent(ref='Americas') Gender(ref='male');
model Expired (event='1')= loan_amount gender num_lenders_total month
continent/nofit;
roc pred=P_1;
roc pred=P_1;
run;

/* Probit Regression */
proc logistic data=kiva_sampled outmodel=Probitmodel;
class Continent(ref='Americas') Gender(ref='male');
probit: model Expired (event='1')= loan_amount gender num_lenders_total month
continent;
weight selected; /*only training sample is used for estimation, since selected =
0 for test sample */
run;
quit;

```

## REFERENCES

"Data Science for Good: Kiva Crowdfunding." Kaggle, [www.kaggle.com/kiva/data-science-for-good-kiva-crowdfunding](http://www.kaggle.com/kiva/data-science-for-good-kiva-crowdfunding).

Holmes, Christian. "Will My Kiva Loan Get Funded?" NYC Data Science Academy Blog, 17 Aug. 2016, [nycdatascience.com/blog/student-works/kiva-loans/](http://nycdatascience.com/blog/student-works/kiva-loans/).

"Kiva (Organization)." Wikipedia, Wikimedia Foundation, 25 Feb. 2019, [en.wikipedia.org/wiki/Kiva\\_\(organization\)](http://en.wikipedia.org/wiki/Kiva_(organization)).

"What Factors Affect Loan Funding Times?" Kiva, [www.kiva.org/blog/what-factors-affect-loan-funding-times](http://www.kiva.org/blog/what-factors-affect-loan-funding-times).

"What Makes Us Unique." Kiva, [www.kiva.org/borrow](http://www.kiva.org/borrow).