Investigation into the Characteristics of Microlenders

Peter Rasmussen

April 15, 2016

Context

- There are an estimated 2 billion unbanked adults who lack access to traditional financial services, including credit[1]
- Microloans are small loans that provide credit to these lower-income, typically unbanked borrowers
- Using platforms like Kiva, individuals can become microlenders and connect to borrowers in other parts of the world
- Since Kiva was founded in 2005, 1.4 million lenders have issued $840 million in loans across 84 different countries[2]
- Other companies and non-profits have taken notice, and are moving into the microlending space
- One of these organizations – Seeds[3] – asked me to investigate the following question: who are microlenders?
- This presentation covers a portion – linear regression – of that investigation

1. http://www.cgap.org/about/faq/who-are-2-billion-unbanked-adults-globally
2. https://www.kiva.org/about
3. http://playseeds.com

Methodology & Workflow

Make a plan

Find data source
- Best data source for microlenders is from Kiva
- Free, extensive, and current

Select label
- Loans / year

Select features
- Gender, region, tech involvement, invites / invited, lending reason

Estimate sample size req'd
- Standard deviation = 0.5
- Error = 2.5%
- Z score = 3.3
- Sample size = 4,277

Get & wrangle data

Download data
- ~1860 json files, each containing 1000 lenders

Make json files valid
- Remove leading and trailing chars

Import & clean data
- Randomly sample json files
- Import list as dataframe
- Encode unicode strings as utf-8
- Standardize entries
- Filter out rows with blank cells

Analyze & present

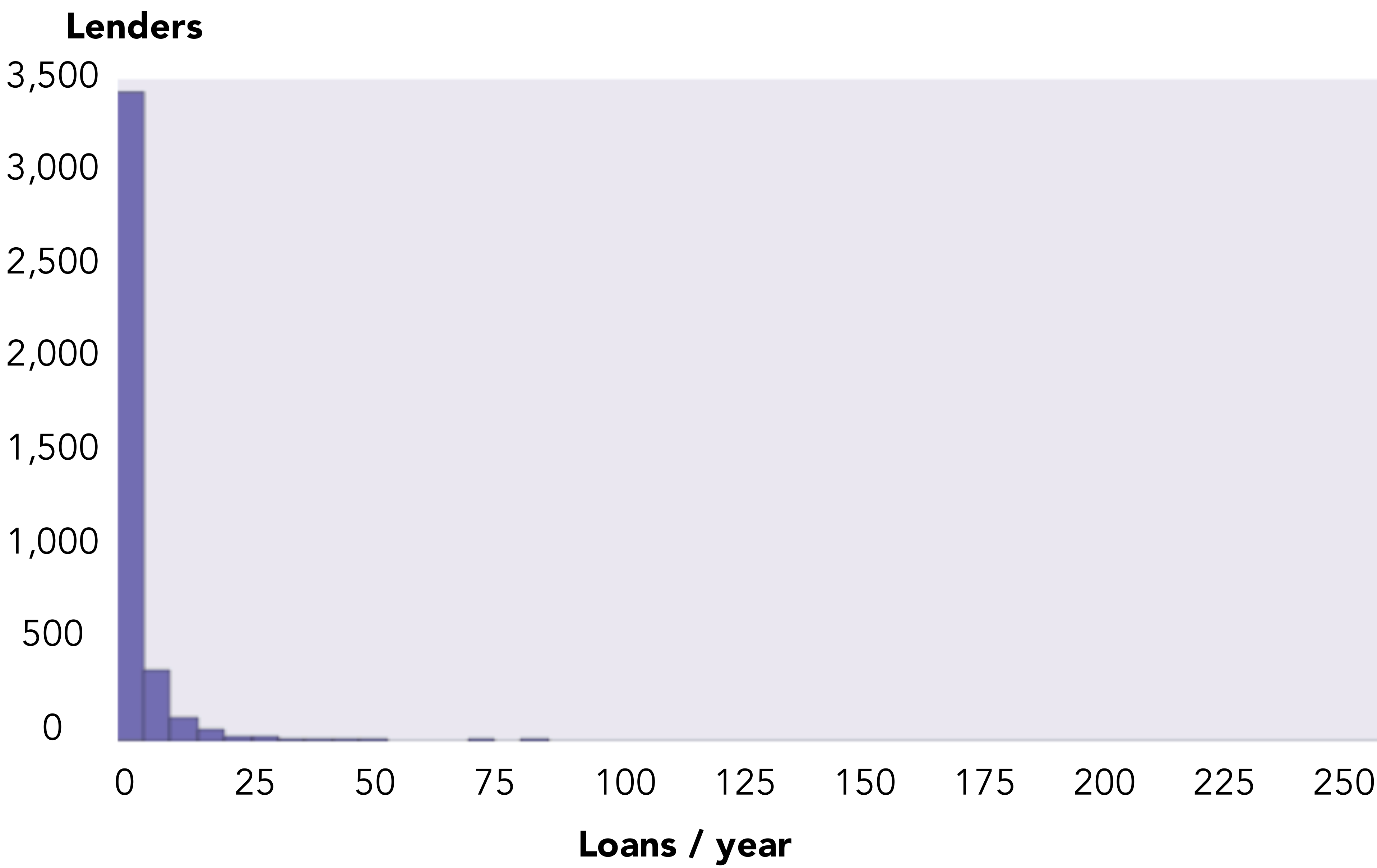
Map and aggregate data
- Assign gender based on name data
- Aggregate locales, occupations

Assume linear relationship between features & label
- Employ OLS to estimate parameters

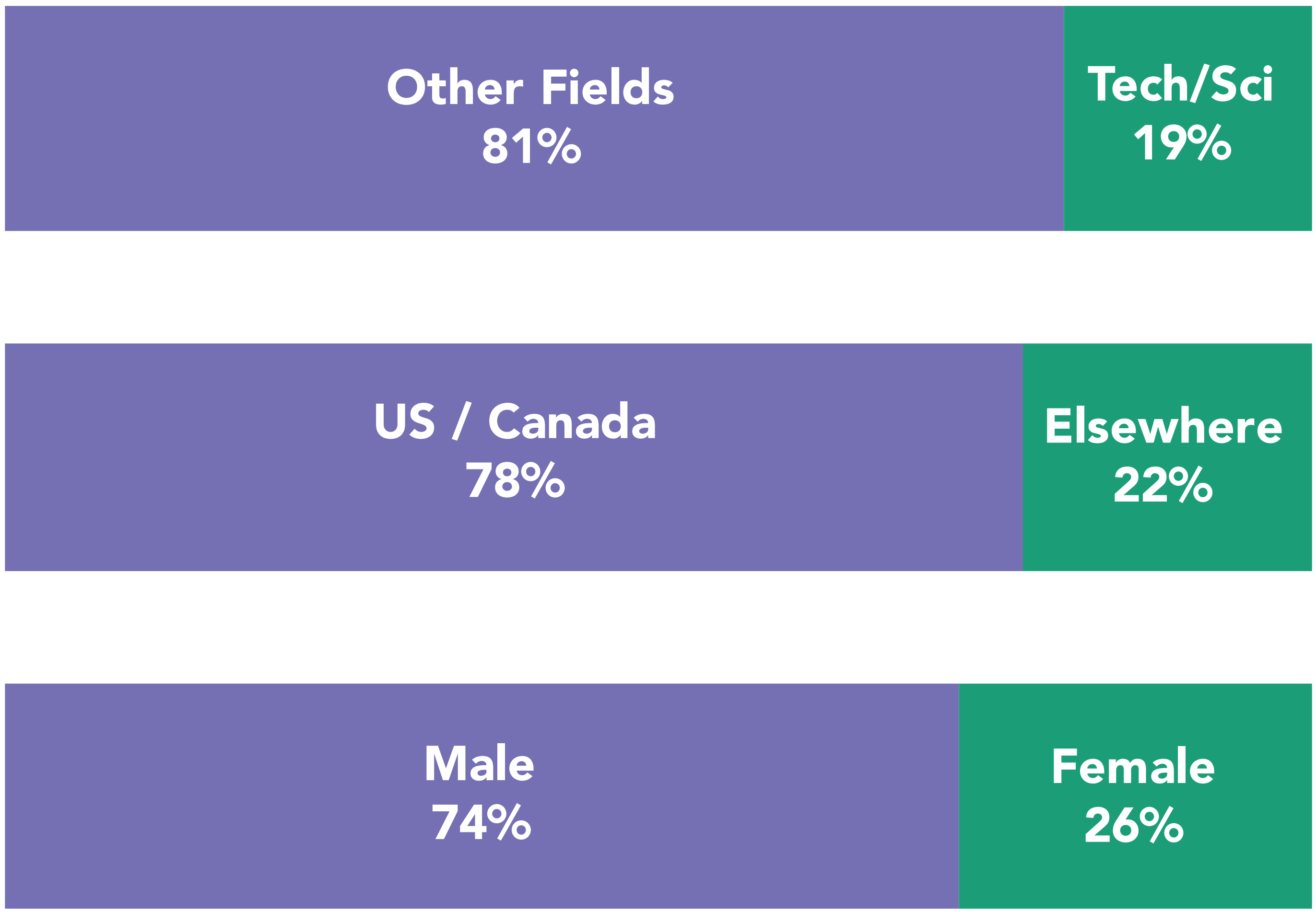
Run the model, tweak segs
- Re-segment region & occupation

Make sense of outputs
- Conclusions and next steps

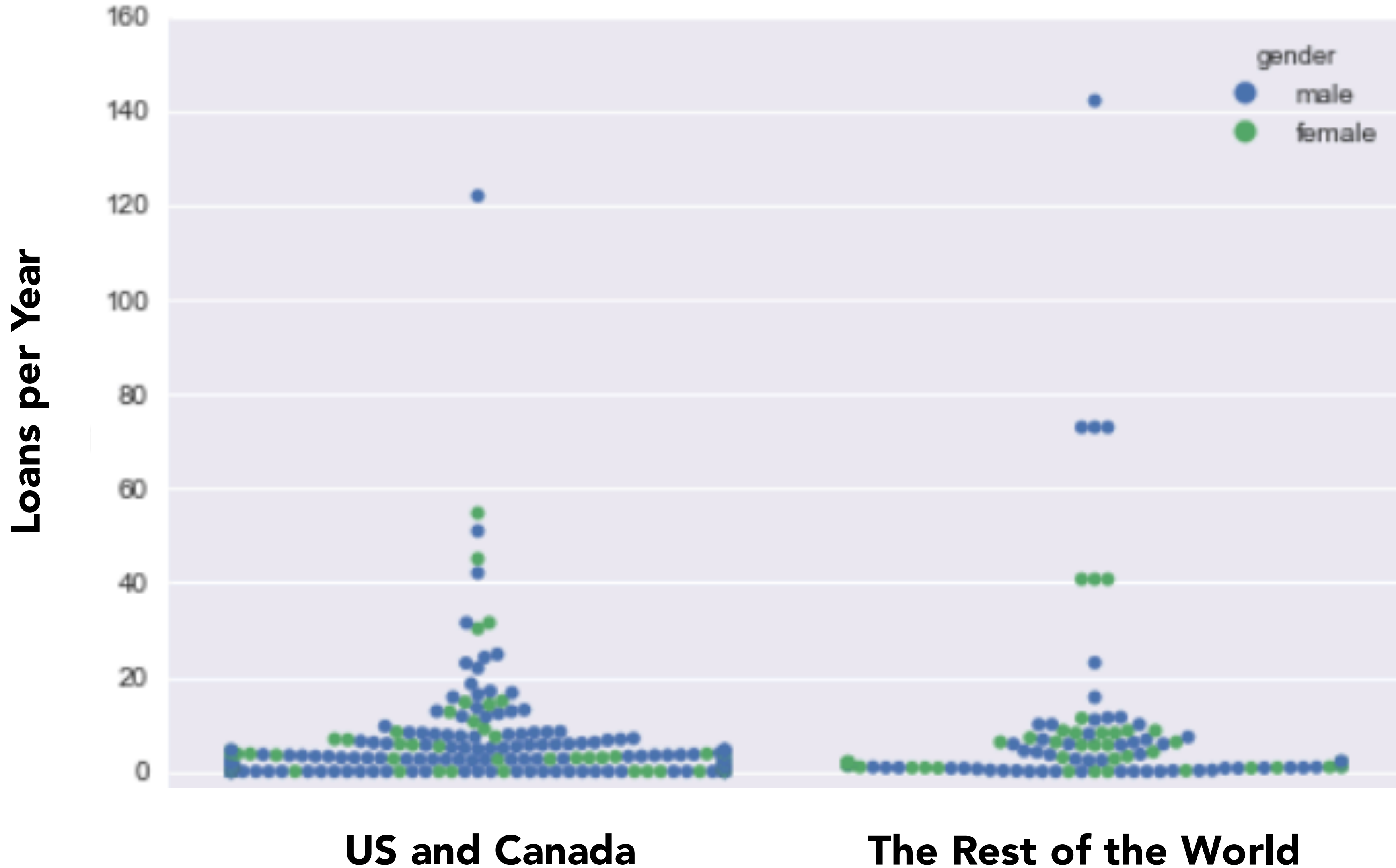
Analysis: PMF of loans / year



Analysis: Demographics Stats



Analysis: Categorical Scatter Plot



Analysis: Regression Statistics

Dep. Variable:	loans_year	**R-squared:**	0.028
Model:	OLS	**Adj. R-squared:**	0.027
Method:	Least Squares	**F-statistic:**	23.70
Date:	Thu, 14 Apr 2016	**Prob (F-statistic):**	1.46e-23
Time:	16:49:04	**Log-Likelihood:**	-18694.
No. Observations:	4125	**AIC:**	3.740e+04
Df Residuals:	4119	**BIC:**	3.744e+04
Df Model:	5		
Covariance Type:	nonrobust		

Model's predictive power is negligible

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	4.2551	0.966	4.407	0.000	2.362 6.148
gender[T.male]	1.8672	0.798	2.339	0.019	0.302 3.432
region[T.US & Canada]	-3.5062	0.852	-4.116	0.000	-5.176 -1.836
tech_science[T.yes]	1.2611	0.896	1.407	0.159	-0.496 3.018
invites_year	5.0628	0.772	6.559	0.000	3.550 6.576
lending_reason	0.0323	0.005	6.408	0.000	0.022 0.042

Omnibus:	10969.715	**Durbin-Watson:**	1.943
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	324329856.778
Skew:	31.262	**Prob(JB):**	0.00
Kurtosis:	1375.261	**Cond. No.**	272.

Analysis: Regression Statistics

Dep. Variable:	loans_year	**R-squared:**	0.028
Model:	OLS	**Adj. R-squared:**	0.027
Method:	Least Squares	**F-statistic:**	23.70
Date:	Thu, 14 Apr 2016	**Prob (F-statistic):**	1.46e-23
Time:	16:49:04	**Log-Likelihood:**	-18694.
No. Observations:	4125	**AIC:**	3.740e+04
Df Residuals:	4119	**BIC:**	3.744e+04
Df Model:	5		
Covariance Type:	nonrobust		

Even so, F-statistics show that it's highly unlikely that the model outputs could be reproduced if null hypothesis were true

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	4.2551	0.966	4.407	0.000	2.362 6.148
gender[T.male]	1.8672	0.798	2.339	0.019	0.302 3.432
region[T.US & Canada]	-3.5062	0.852	-4.116	0.000	-5.176 -1.836
tech_science[T.yes]	1.2611	0.896	1.407	0.159	-0.496 3.018
invites_year	5.0628	0.772	6.559	0.000	3.550 6.576
lending_reason	0.0323	0.005	6.408	0.000	0.022 0.042

Omnibus:	10969.715	**Durbin-Watson:**	1.943
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	324329856.778
Skew:	31.262	**Prob(JB):**	0.00
Kurtosis:	1375.261	**Cond. No.**	272.

Analysis: Regression Statistics

Number of observations was perhaps high enough to offset affect of high variability of label

Dep. Variable:	loans_year	**R-squared:**	0.028
Model:	OLS	**Adj. R-squared:**	0.027
Method:	Least Squares	**F-statistic:**	23.70
Date:	Thu, 14 Apr 2016	**Prob (F-statistic):**	1.46e-23
Time:	16:49:04	**Log-Likelihood:**	-18694.
No. Observations:	4125	**AIC:**	3.740e+04
Df Residuals:	4119	**BIC:**	3.744e+04
Df Model:	5		
Covariance Type:	nonrobust		

Even so, F-statistics show that it's highly unlikely that the model outputs could be reproduced if null hypothesis were true

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	4.2551	0.966	4.407	0.000	2.362 6.148
gender[T.male]	1.8672	0.798	2.339	0.019	0.302 3.432
region[T.US & Canada]	-3.5062	0.852	-4.116	0.000	-5.176 -1.836
tech_science[T.yes]	1.2611	0.896	1.407	0.159	-0.496 3.018
invites_year	5.0628	0.772	6.559	0.000	3.550 6.576
lending_reason	0.0323	0.005	6.408	0.000	0.022 0.042

Omnibus:	10969.715	**Durbin-Watson:**	1.943
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	324329856.778
Skew:	31.262	**Prob(JB):**	0.00
Kurtosis:	1375.261	**Cond. No.**	272.

Analysis: Regression Statistics

Per model coefficients

- Males make 1.9 more loans/yr
- US & Canada make 3.5 loans/yr less than elsewhere
- Tech & science people tend to make 1.3 more loans/yr

Dep. Variable:	loans_year	**R-squared:**	0.028
Model:	OLS	**Adj. R-squared:**	0.027
Method:	Least Squares	**F-statistic:**	23.70
Date:	Thu, 14 Apr 2016	**Prob (F-statistic):**	1.46e-23
Time:	16:49:04	**Log-Likelihood:**	-18694.
No. Observations:	4125	**AIC:**	3.740e+04
Df Residuals:	4119	**BIC:**	3.744e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	4.2551	0.966	4.407	0.000	2.362 6.148
gender[T.male]	1.8672	0.798	2.339	0.019	0.302 3.432
region[T.US & Canada]	-3.5062	0.852	-4.116	0.000	-5.176 -1.836
tech_science[T.yes]	1.2611	0.896	1.407	0.159	-0.496 3.018
invites_year	5.0628	0.772	6.559	0.000	3.550 6.576
lending_reason	0.0323	0.005	6.408	0.000	0.022 0.042

Omnibus:	10969.715	**Durbin-Watson:**	1.943
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	324329856.778
Skew:	31.262	**Prob(JB):**	0.00
Kurtosis:	1375.261	**Cond. No.**	272.

Analysis: Regression Statistics

Dep. Variable:	loans_year	**R-squared:**	0.028
Model:	OLS	**Adj. R-squared:**	0.027
Method:	Least Squares	**F-statistic:**	23.70
Date:	Thu, 14 Apr 2016	**Prob (F-statistic):**	1.46e-23
Time:	16:49:04	**Log-Likelihood:**	-18694.
No. Observations:	4125	**AIC:**	3.740e+04
Df Residuals:	4119	**BIC:**	3.744e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	4.2551	0.966	4.407	0.000	2.362 6.148
gender[T.male]	1.8672	0.798	2.339	0.019	0.302 3.432
region[T.US & Canada]	-3.5062	0.852	-4.116	0.000	-5.176 -1.836
tech_science[T.yes]	1.2611	0.896	1.407	0.159	-0.496 3.018
invites_year	5.0628	0.772	6.559	0.000	3.550 6.576
lending_reason	0.0323	0.005	6.408	0.000	0.022 0.042

Omnibus:	10969.715	**Durbin-Watson:**	1.943
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	324329856.778
Skew:	31.262	**Prob(JB):**	0.00
Kurtosis:	1375.261	**Cond. No.**	272.

p values for gender, region, invites / year, and lending indicate that data for these features are inconsistent with what the null hypothesis would predict → **reject null**

Analysis: Regression Statistics

Dep. Variable:	loans_year	**R-squared:**	0.028
Model:	OLS	**Adj. R-squared:**	0.027
Method:	Least Squares	**F-statistic:**	23.70
Date:	Thu, 14 Apr 2016	**Prob (F-statistic):**	1.46e-23
Time:	16:49:04	**Log-Likelihood:**	-18694.
No. Observations:	4125	**AIC:**	3.740e+04
Df Residuals:	4119	**BIC:**	3.744e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**	
Intercept	4.2551	0.966	4.407	0.000	2.362	6.148
gender[T.male]	1.8672	0.798	2.339	0.019	0.302	3.432
region[T.US & Canada]	-3.5062	0.852	-4.116	0.000	-5.176	-1.836
tech_science[T.yes]	1.2611	0.896	1.407	0.159	-0.496	3.018
invites_year	5.0628	0.772	6.559	0.000	3.550	6.576
lending_reason	0.0323	0.005	6.408	0.000	0.022	0.042

p value for tech & science indicates weak evidence against the null hypothesis, so **fail to reject**

Omnibus:	10969.715	**Durbin-Watson:**	1.943
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	324329856.778
Skew:	31.262	**Prob(JB):**	0.00
Kurtosis:	1375.261	**Cond. No.**	272.

Conclusions and Next Steps

- This is a preliminary analysis, and more insights could be gleaned from Kiva data
- Even so, we can draw the following conclusions from this analysis:
  - Microlending activity across microlenders seems to be exponentially distributed
  - While the model developed for this analysis is not predictive, it does explain a subset of the types of people that tends to make more microloans
  - Within the sample data, males and people living outside the US and Canada tend to be more active microlenders
  - More analysis is needed before we can say the same for people involved in tech and the sciences
- Further analysis would may show that the selected features for this model should be refined, removed, or added
- Once we obtain a more predictive model, could move on to test the training data with test data to see if model is over or under fitted