Final Project Presentation

CUNY MSDA Bridge Courses:

SQL

R Programming Basics

Probability and Linear Algebra Basics

Randi Skrelja

July 26, 2015

Scope of Project

This project seeks to combine the concepts learned from all three bridges – R, SQL and Data Science Math. The first step involved downloading data from both Lending Club and Zipcodes.com in csv format. The data files were then scrubbed, imported and combined within SQL. Then within R, the data was transformed and analyzed with descriptive charts and some conditional probability.

Data Sources:

Lending Club – Investors are provided with historical peer-to-peer lending stats; data downloaded for the 2013/2014 period as a csv file.

Zip-Codes.com – Subscription data which offers demographic information by zip code for the nation; data downloaded as a csv file.

SQL: Scrubbing, Creating, Joining and Exporting Data

```sql
-- Create table from lending club csv

DROP TABLE lendingclub;

CREATE TABLE lendingclub
(
id int,
term int,
intrate numeric,
grade varchar,
homeownership varchar,
annualinc numeric,
issued date,
loanstatus varchar,
title varchar,
zipcode varchar,
addrstate varchar,
dti numeric,
ficorangelow int,
totalpymnt numeric,
totalpymntinv numeric,
lastficorangelow int
)
WITH (
  OIDS=FALSE
);
ALTER TABLE lendingclub
  OWNER TO postgres;
```

```sql
-- Create table from zipcodes csv

DROP TABLE zipcodes;

CREATE TABLE zipcodes
(
zipcode varchar,
AverageHouseValue int,
IncomePerHousehold int,
State varchar,
NumberOfBusinesses int,
NumberOfEmployees int,
BusinessAnnualPayroll int,
PopulationEstimate int
)
WITH (
  OIDS=FALSE
);
ALTER TABLE zipcodes
  OWNER TO postgres;
```

```sql
-- Created new zip code table to transform zipcode column.

Drop table zipcodes_grouped;

SELECT substring(zipcode,1,3)||'00' zipcode,
       avg(averagehousevalue) averagehousevalue,
       avg(incomeperhousehold) incomeperhousehold,
       state, avg(numberofbusinesses) numberofbusinesses,
       avg(numberofemployees) numberofemployees,
       avg(businessannualpayroll) businessannualpayroll,
       avg(populationestimate) populationestimate
INTO zipcodes_grouped
FROM zipcodes
GROUP BY substring(zipcodes.zipcode,1,3),state;


-- Join lending club table with new zipcodes table (from code above).
-- Used "Execute to File" in Query menu to save joined results
--    to csv file named lendingclub_zipcodes.

SELECT l.*, z.*
FROM lendingclub l
     LEFT JOIN zipcodes_grouped z
     ON l.zipcode = z.zipcode
     and l.addrstate = z.state
ORDER BY id;
```

Output pane

Data Output | Explain | Messages | History

	id integer	term integer	intrate numeric	grade character varying	homeownership character varying	annualinc numeric	issued date	loanstatus character varying	title character varying	zipcode character varying	addrstate character varying	dti numeric	ficorangelow integer	totalpymnt numeric	totalpymntinv numeric	lastficorangelow integer	zipcode text	aver
1	57167	36	0.1699	D	RENT	70000	2015-08-14	ChargedOff	mlue	10000	NY	10.5	660	2718.16	2688.26	535	10000	
2	300390	36	0.0649	A	RENT	165000	2015-12-14	Current	Debtconsolidation	07000	NJ	4.45	715	2212.11	2212.11	755	07000	
3	361542	36	0.0699	A	MORTGAGE	250000	2015-12-14	Current	Business	14000	NY	3.25	820	5175.77	5175.77	770	14000	
4	367050	36	0.0712	A	MORTGAGE	100000	2015-10-14	Current	Gettingoutofdebt	98100	WA	19.13	715	6917.72	6917.72	685	98100	
5	377140	36	0.0649	A	MORTGAGE	102000	2015-12-14	Current	Debtconsolidation	33000	FL	12.4	765	5378.56	5378.56	820	33000	

OK. Unix Ln 79, Col 13, Ch 1590 145 chars 235629 rows. 36556 ms

R: Importing & Transforming Data

```
# I. Reading in data / creating data frames and subsets:
lendingclub_zipcodes <- read.csv("C:/Users/Randi Skrelja/Desktop/MSDA/Final/lendingclub_zipcodes.csv", header=T
RUE, sep=",")
loans <- data.frame(lendingclub_zipcodes)
numloans <- nrow(loans)
defaults <- nrow(loans[loans$loanstatus=="ChargedOff",])
homeowner <- data.frame(loans[loans$homeownership=="OWN"|loans$homeownership=="MORTGAGE",])
homeownercount <- nrow(homeowner)
defaults_homeowner <- nrow(homeowner[homeowner$loanstatus=="ChargedOff",])
defaults_df <- data.frame(loans[loans$loanstatus=="ChargedOff",])
```

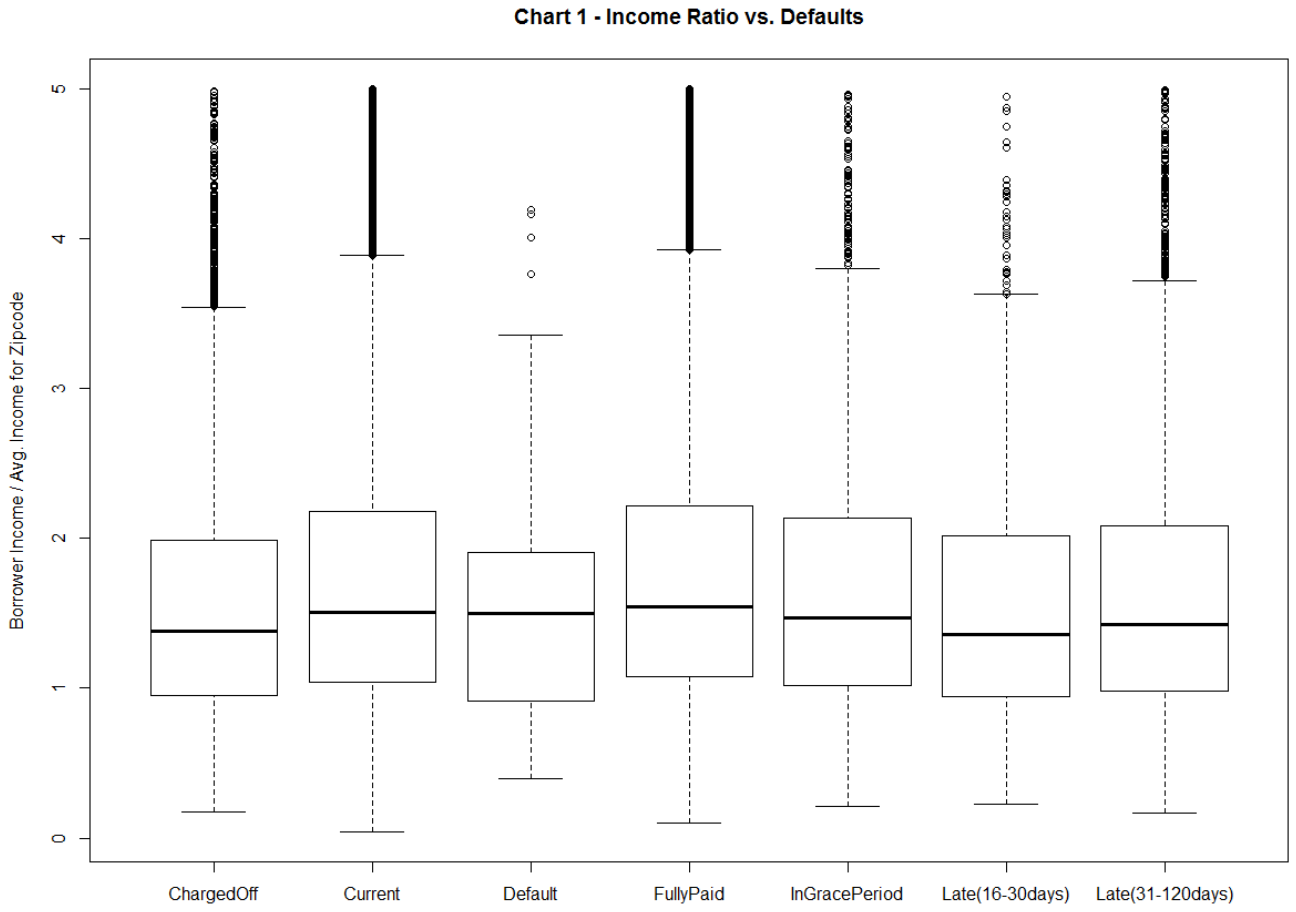
```
# II. Transforming dataframe to add new column (Income Ratio vs. Defaults Chart 1):
loans$incomeperhousehold[loans$incomeperhousehold==0] <- NA  #replace missing values with NA to exclude from ca
lcuation
loans["income_ratio"] <- NA  # Adds a new column with NA placeholders
loans$income_ratio <- loans$annualinc/loans$incomeperhousehold  # populates new column with calculated values
summary(loans$income_ratio)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##   0.0429   1.0560   1.5390   1.9020   2.2720 269.8000      464
```

```
loans$income_ratio[loans$income_ratio > 5] <- NA  # Ignoring outliers > 5 in ratios
summary(loans$income_ratio)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.043   1.040   1.505   1.715   2.178   5.000    7737
```

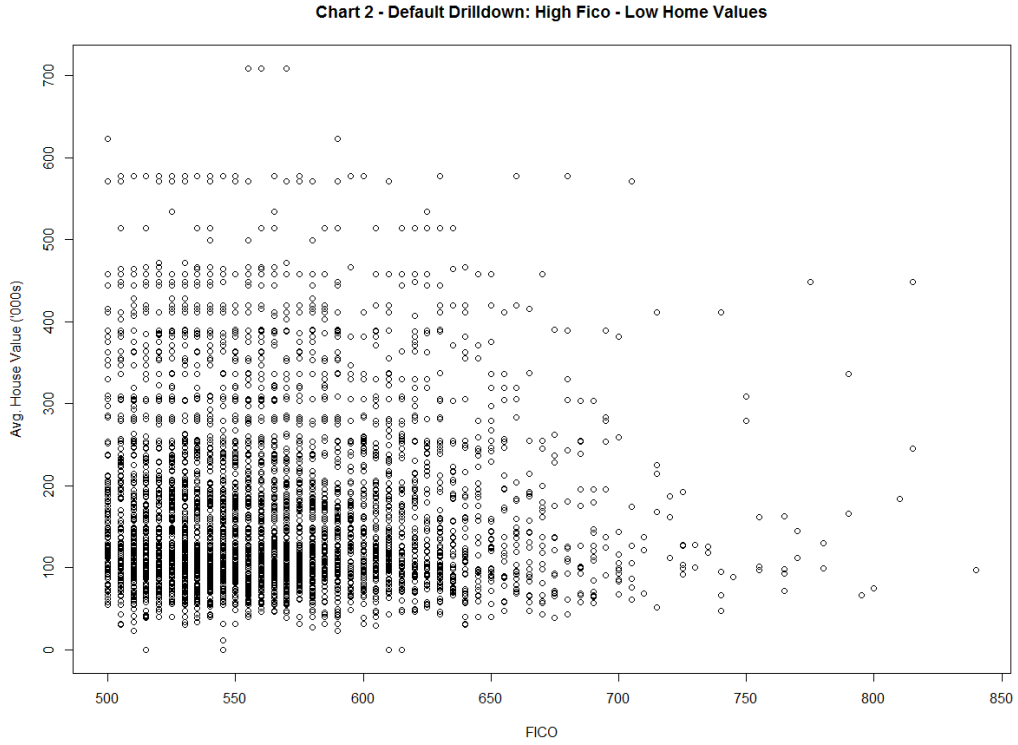
R: Visualizing & Analyzing Data



We charted the ratio of borrower income to average income against loan status and observe the median for Defaults ("ChargedOff") is lower than for performing loans.

```
plot(loans$loanstatus,loans$income_ratio,ylab="Borrower Income / Avg. Income for Zipcode", main = "Chart 1 - In
come Ratio vs. Defaults")
```

R: Visualizing & Analyzing Data



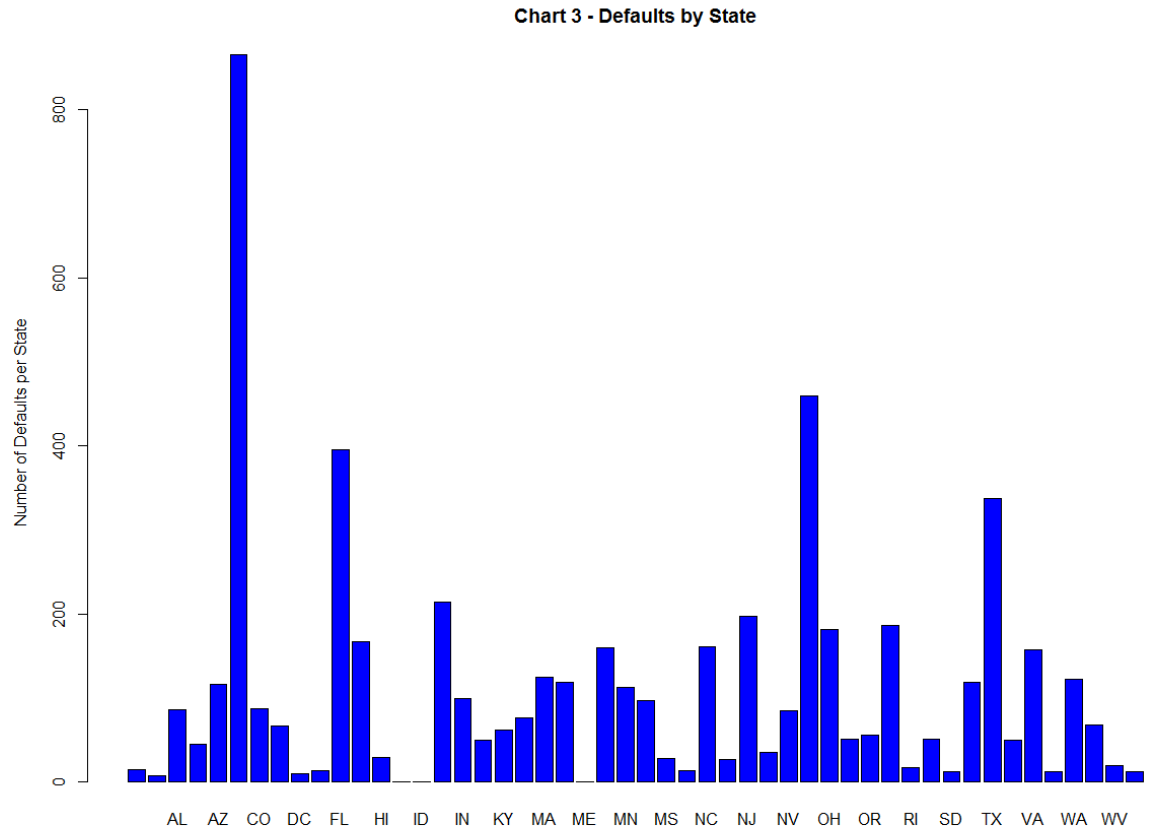
Within defaults, we charted average house values against fico scores, and observe that even with higher fico scores the borrower can still default if the home values are low.

```
# III. Fico Score vs. Avg House Value by Zipcode for Defaults (Chart 2):
defaults_df$lastficorangelow[defaults_df$lastficorangelow==0] <- NA  # Replace missing ficos w/NA to exclude fr
om plot
summary(defaults_df$lastficorangelow)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   500.0   530.0   555.0   565.5   590.0   840.0     725
```

```
plot(defaults_df$lastficorangelow,defaults_df$averagehousevalue/1000,xlab = "FICO",ylab = "Avg. House Value ('0
00s)",main = "Chart 2 - Default Drilldown: High Fico - Low Home Values")
```

R: Visualizing & Analyzing Data



CA has the most loans and the most defaults.

```
# IV. Defaults Distribution by State (Chart 3):
plot(defaults_df$state, col="blue",ylab="Number of Defaults per State", main = "Chart 3 - Defaults by State")
```

Math: Conditional Probability

We derive from the data that the probability of default is 2.3% overall. Applying P(A|B)=P(A and B)/P(B), we add the condition that those defaulting are also homeowners and see the probability of default is lower at 2.0%.

```
# V. Conditional Probability Calculations:
prob_defaults <- defaults / numloans * 100  # Probability of defaulting
print(prob_defaults)
```

```
## [1] 2.329085
```

```
prob_homeowner <- homeownercount / numloans * 100  # Probability of homeownership
print(prob_homeowner)
```

```
## [1] 60.66486
```

```
prob_defaults_homeowner <- defaults_homeowner/homeownercount * 100  #Probability of default given homeownership
print(prob_defaults_homeowner)
```

```
## [1] 2.044157
```