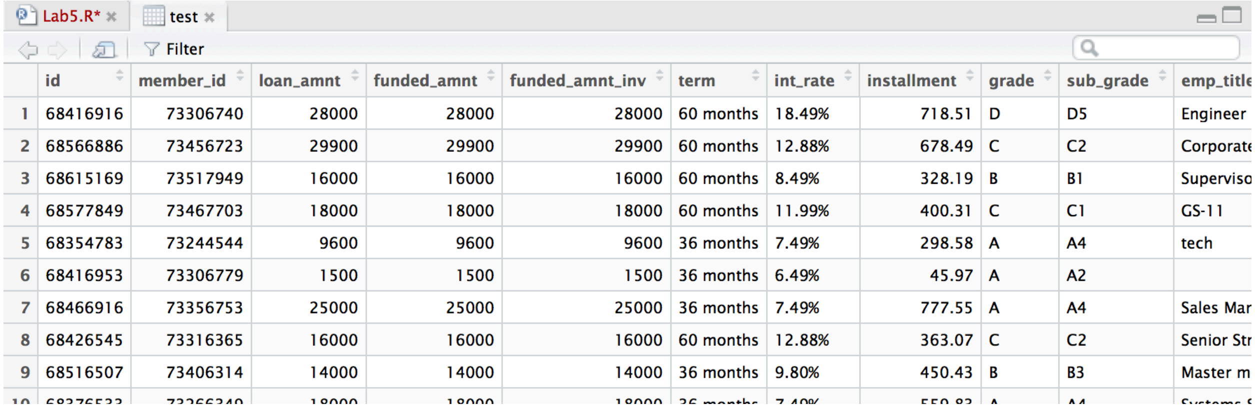
Lab #5 - Peer-to-peer lending

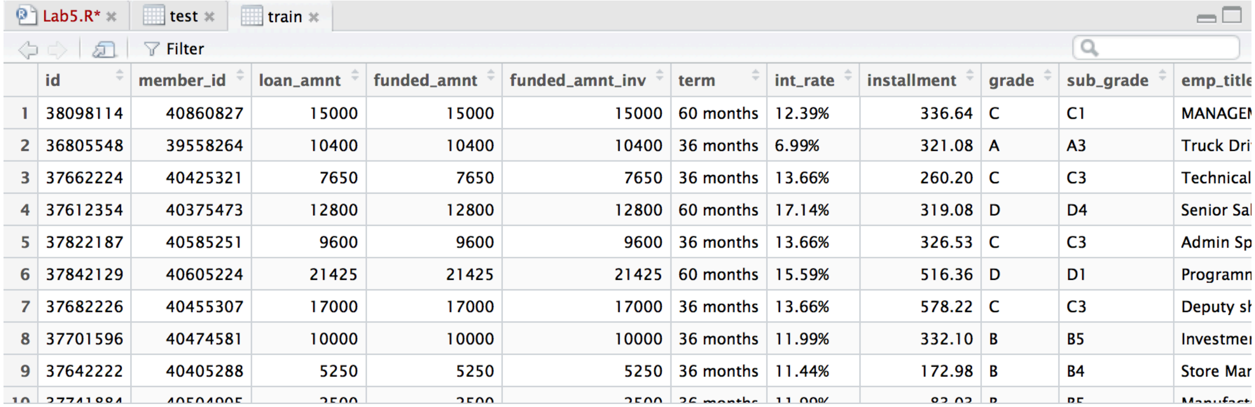
Professor Tambe, Analytics & the Digital Economy

By: Christian Miljkovic
Group: Zachary Fineberg, Suchit Sadan
December 16, 2016

1. Data loading and cleanup.



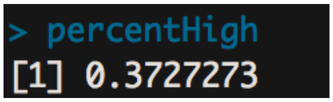
	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	emp_title
1	68416916	73306740	28000	28000	28000	60 months	18.49%	718.51	D	D5	Engineer
2	68566886	73456723	29900	29900	29900	60 months	12.88%	678.49	C	C2	Corporate
3	68615169	73517949	16000	16000	16000	60 months	8.49%	328.19	B	B1	Superviso
4	68577849	73467703	18000	18000	18000	60 months	11.99%	400.31	C	C1	GS-11
5	68354783	73244544	9600	9600	9600	36 months	7.49%	298.58	A	A4	tech
6	68416953	73306779	1500	1500	1500	36 months	6.49%	45.97	A	A2	
7	68466916	73356753	25000	25000	25000	36 months	7.49%	777.55	A	A4	Sales Mar
8	68426545	73316365	16000	16000	16000	60 months	12.88%	363.07	C	C2	Senior Str
9	68516507	73406314	14000	14000	14000	36 months	9.80%	450.43	B	B3	Master m
10	68376533	73266349	18000	18000	18000	36 months	7.49%	559.83	A	A4	Systems S



	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	emp_title
1	38098114	40860827	15000	15000	15000	60 months	12.39%	336.64	C	C1	MANAGEM
2	36805548	39558264	10400	10400	10400	36 months	6.99%	321.08	A	A3	Truck Dri
3	37662224	40425321	7650	7650	7650	36 months	13.66%	260.20	C	C3	Technical
4	37612354	40375473	12800	12800	12800	60 months	17.14%	319.08	D	D4	Senior Sal
5	37822187	40585251	9600	9600	9600	36 months	13.66%	326.53	C	C3	Admin Sp
6	37842129	40605224	21425	21425	21425	60 months	15.59%	516.36	D	D1	Programn
7	37682226	40455307	17000	17000	17000	36 months	13.66%	578.22	C	C3	Deputy sh
8	37701596	40474581	10000	10000	10000	36 months	11.99%	332.10	B	B5	Investmer
9	37642222	40405288	5250	5250	5250	36 months	11.44%	172.98	B	B4	Store Mar
10	37741884	40504905	2500	2500	2500	36 months	11.99%	83.03	B	B5	Manufact

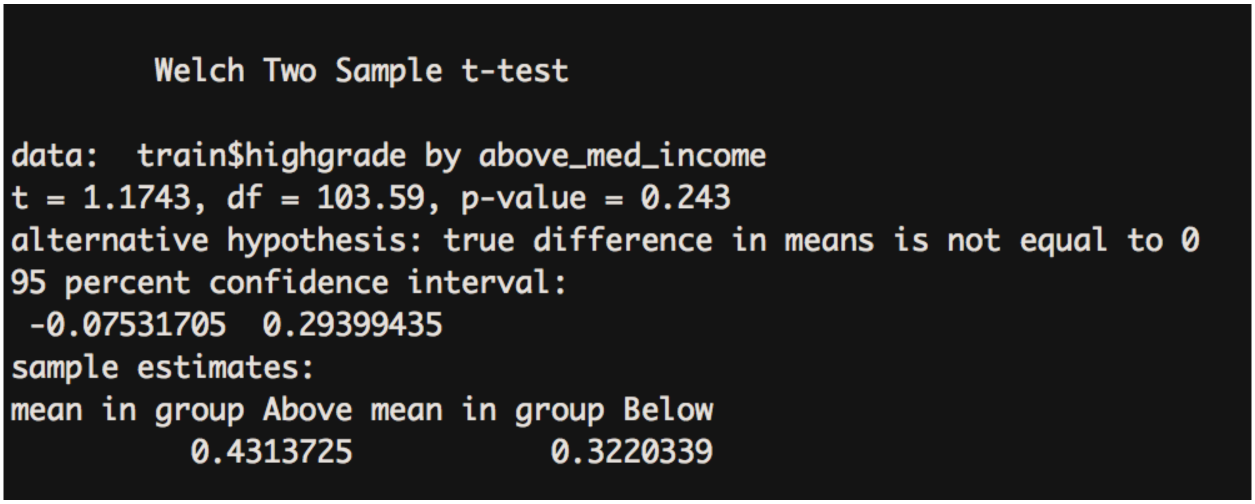
2. Descriptive statistics.

Percent of loans that got high ratings



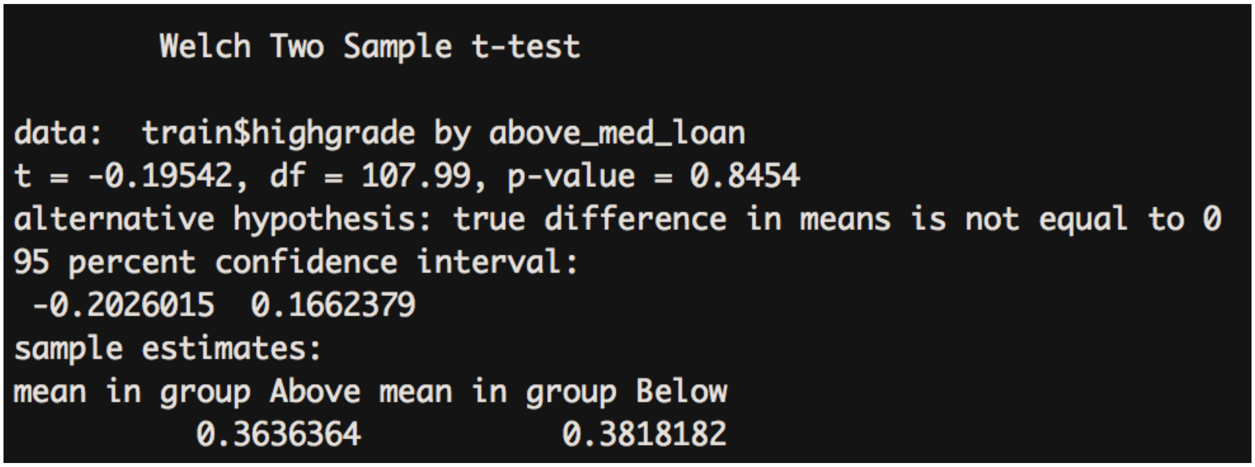
```
> percentHigh
[1] 0.3727273
```

Whether the debtor is above or below the median income level



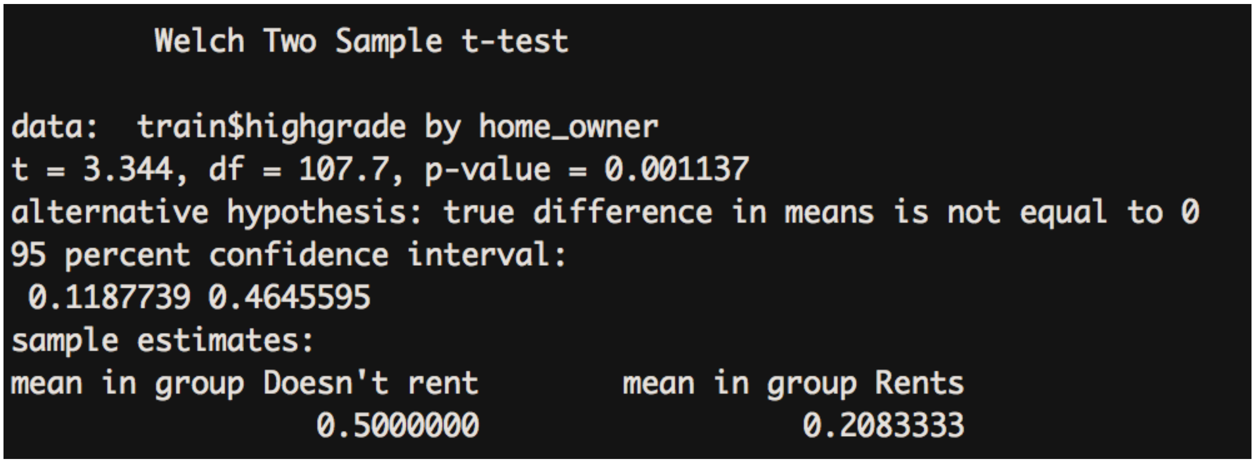
```
        Welch Two Sample t-test

data:  train$highgrade by above_med_income
t = 1.1743, df = 103.59, p-value = 0.243
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.07531705  0.29399435
sample estimates:
mean in group Above mean in group Below
          0.4313725           0.3220339
```

Whether the loan request is above or below the median loan amount



```
        Welch Two Sample t-test

data:  train$highgrade by above_med_loan
t = -0.19542, df = 107.99, p-value = 0.8454
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2026015  0.1662379
sample estimates:
mean in group Above mean in group Below
          0.3636364           0.3818182
```

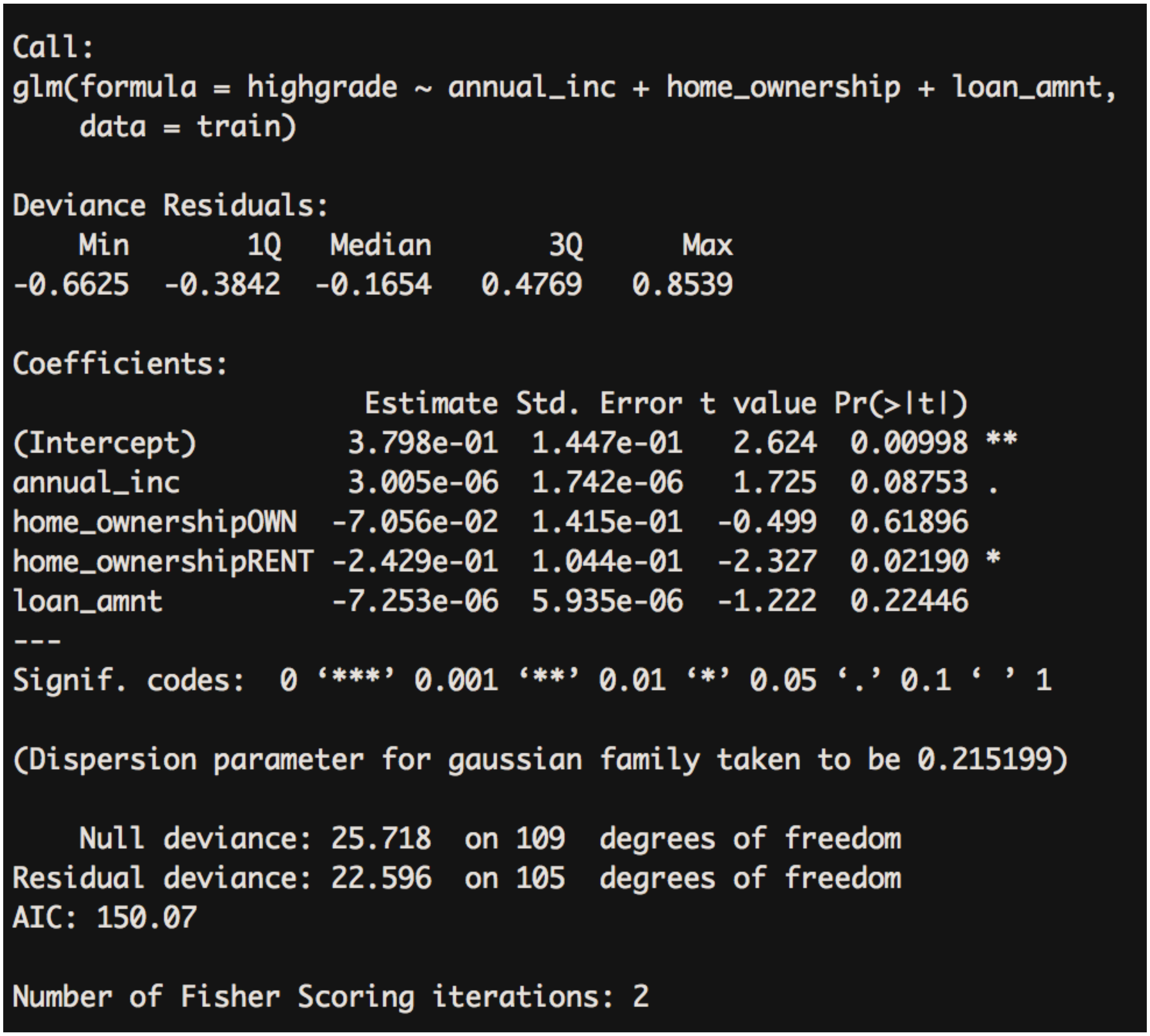
Whether the debtor rents their home or not



```
        Welch Two Sample t-test

data:  train$highgrade by home_owner
t = 3.344, df = 107.7, p-value = 0.001137
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1187739 0.4645595
sample estimates:
mean in group Doesn't rent        mean in group Rents
                 0.5000000                  0.2083333
```

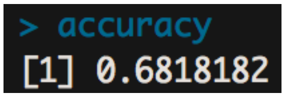
3. Build a logistic classifier on the training data.

Cut and paste the output produced by the *summary* command.



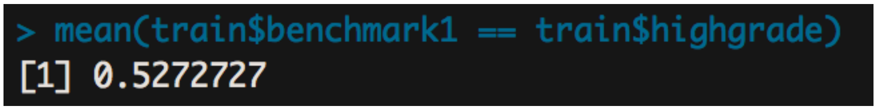
```
Call:
glm(formula = highgrade ~ annual_inc + home_ownership + loan_amnt,
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6625  -0.3842  -0.1654   0.4769   0.8539

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.798e-01  1.447e-01   2.624  0.00998 **
annual_inc          3.005e-06  1.742e-06   1.725  0.08753 .
home_ownershipOWN  -7.056e-02  1.415e-01  -0.499  0.61896
home_ownershipRENT -2.429e-01  1.044e-01  -2.327  0.02190 *
loan_amnt          -7.253e-06  5.935e-06  -1.222  0.22446
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.215199)

    Null deviance: 25.718  on 109  degrees of freedom
Residual deviance: 22.596  on 105  degrees of freedom
AIC: 150.07

Number of Fisher Scoring iterations: 2
```

What is the accuracy of this classifier on the training data?



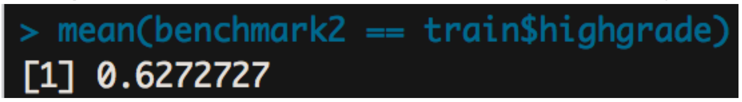
```
> accuracy
[1] 0.6818182
```

As a benchmark, what would be the accuracy of a classifier that randomly assigns 0 and 1 values as the predicted class?



```
> mean(train$benchmark1 == train$highgrade)
[1] 0.5272727
```

As another benchmark, what is the accuracy of a classifier that simply assigns a value of 0 to all rows for the predicted class?



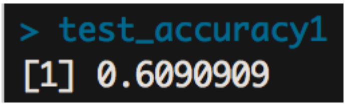
```
> mean(benchmark2 == train$highgrade)
[1] 0.6272727
```

4. Supervised learning.

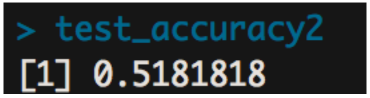
The machine learning based classifier has an accuracy of 0.7545 while the regression based approach has an accuracy of 0.6818

5. Model performance on the test data.

Evaluate the accuracy of both of the classifiers you built above (logistic regression + machine learning) on the test data.

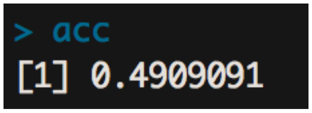


```
> test_accuracy1
[1] 0.6090909
```



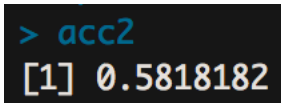
```
> test_accuracy2
[1] 0.5181818
```

As a benchmark, what is the accuracy of a classifier that randomly assigns 0 and 1 values to the test data?



```
> acc
[1] 0.4909091
```

As another benchmark, what is the accuracy of a classifier that simply assigns a value of 0 to all rows of the test data?



```
> acc2
[1] 0.5818182
```