**BUAN 6337.501**

**Homework 1**

**Group # 2**

Virajita Gaddam

Neelima Naladimmu

Varun Muralidharan

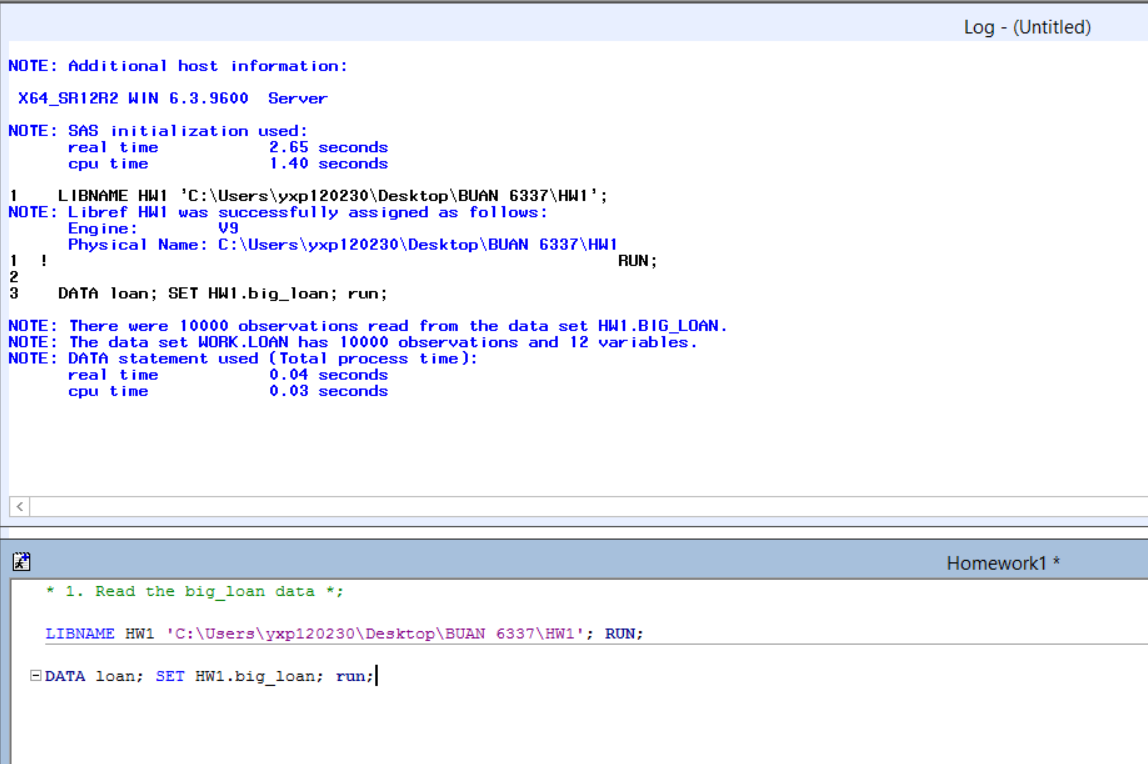
Yohanjohn Park

Zhao Wang

**1. Read the big\_loan data**

LIBNAME HW1 'C:\Users\yxp120230\Desktop\BUAN 6337\HW1'; **RUN**;

**DATA** loan; SET HW1.big\_loan; **run**;



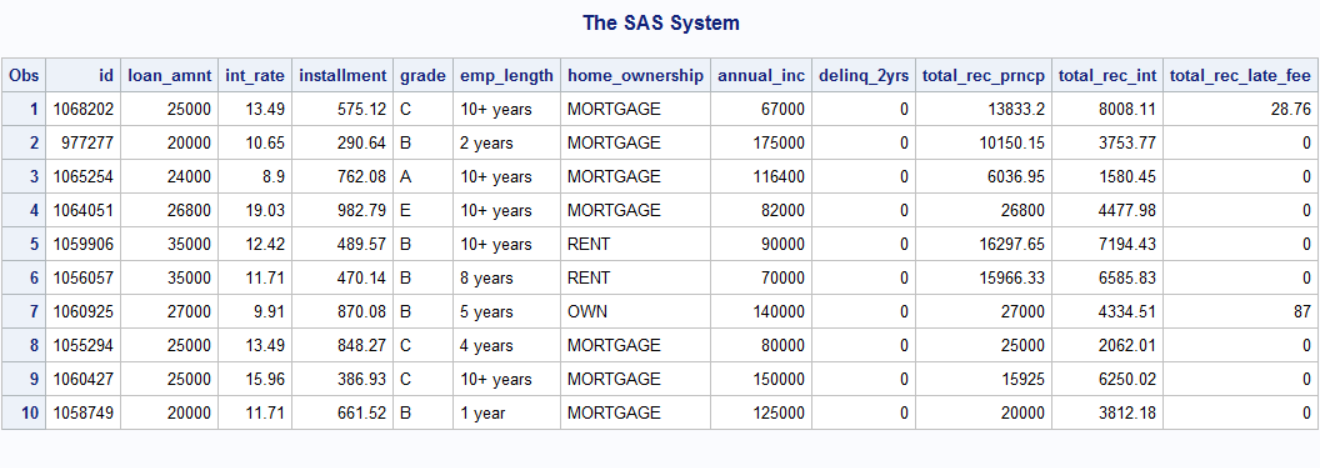
**2. What variables are in the data?**

Id, loan\_amnt, int\_rate, installment, grade, emp\_lenth, home\_ownership, annual\_inco, deling\_2yrs, total\_rec\_prncp, total\_rec\_int, total\_rec\_late\_fee



**3. Write SAS code to print 10 observations?**

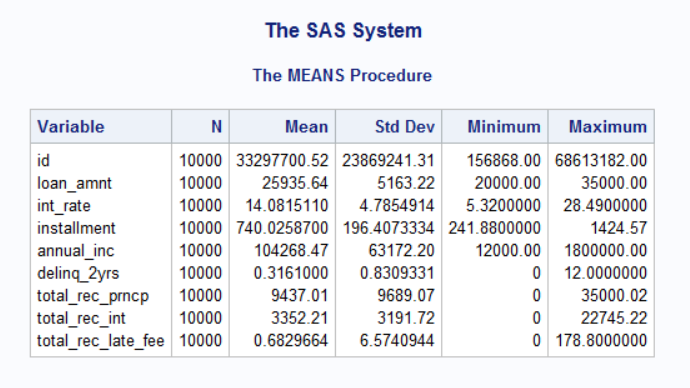
**PROC** **PRINT** DATA=loan (obs=**10**); **RUN**;



**4. What is the average loan amount?**

$25,935.64

**PROC** **MEANS** DATA=loan; **RUN**;



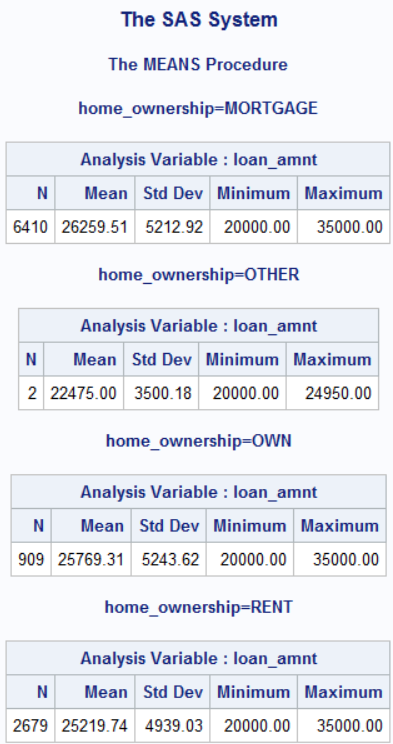
**5. How many home-owners are in the data?**

909 home-owners.

**6. What is the average loan amount across different categories of home-ownership?**

Mortgage : $26259.51, Other : $22475, Own: $25769.31, Rent: $25219.74

**PROC** **SORT** DATA=loan OUT=ownership;BY home\_ownership; **RUN**;

**PROC** **MEANS**; VAR loan\_amnt; BY home\_ownership; **RUN**;

**7. Do a ttest to check if the average loan amount is different between mortgage owners and renters.**

1. Mortgage owners have higher average loan amount than renters: (26,250 > 25,220)
2. Difference in average loan amount: $1,039.8
3. F value is the ratio of the sample variances, so if the variances were equal it would be 1.

Our F value is 1.11, it indicates that the two groups have similar sample variances.

1. The p-value 0.001 (Pr>F) is less than 0.05, we fail to reject equality of variance;

therefore, we assume unequal variances and use Satterthwaite t value.

**DATA** q7; SET loan; IF home\_ownership='MORTGAGE' or home\_ownership='RENT' THEN output; **RUN**;

**PROC** **TTEST** DATA=q7 sides=**2** alpha = **0.05** h0 = **0**;

CLASS home\_ownership; VAR loan\_amnt;

title "Two sample t-test"; **RUN**;



**8. Do a ttest to check if the delinquency in 2 years is different between mortgage owners and renters.**

1. Mortgage owners have a higher delinquency rate.
2. Difference in average delinquency rate: 0.0532
3. F value: 1.44 indicates the two variances are close.
4. The p-value 0.0001 (Pr>F) is less than 0.05, we fail to reject equality of variance;

therefore, we assume unequal variances and use Satterthwaite t value.

**PROC** **TTEST** DATA=q7 sides=**2** alpha = **0.05** h0 = **0**;

CLASS home\_ownership; VAR delinq\_2yrs;

title "Two sample t-test Q8"; **RUN**;



**9. Check for significant differences in principal, interest, and late fee between the two groups. What do you conclude?**

1. For principal, assumed unequal variances, there is a statistically significant evidence (p=0.006) that the mean for mortgage owners (9689.8) is larger than renters (9089.1).
2. For interest, assumed equal variances, there is not a statistically significant evidence (p=0.387) that the mean interest of mortgage owners (3406) is different from renters (3343).
3. For late fee, assumed unequal variances, there is statistically significant evidence (p=0.029) that the mean late fee of mortgage owners (0.577) is smaller than renters (0.951).

**PROC** **TTEST** DATA=q7 sides=**2** alpha = **0.05** h0 = **0**;

CLASS home\_ownership; VAR total\_rec\_prncp;

title "Two sample t-test - principal"; **RUN**;

**PROC** **TTEST** DATA=q7 sides=**2** alpha = **0.05** h0 = **0**;

CLASS home\_ownership; VAR total\_rec\_int;

title "Two sample t-test - interest"; **RUN**;

**PROC** **TTEST** DATA=q7 sides=**2** alpha = **0.05** h0 = **0**;

CLASS home\_ownership; VAR total\_rec\_late\_fee;

title "Two sample t-test - late fee"; **RUN**;

**10. Use PROC ANOVA to test if there is a difference in mean loan amount between the three groups in home-ownership? First delete “Other” since there are only 2 observations in this group.**

The p-value (0.0001) for the overall ANOVA test indicates there is evidence that at least one pair of means is not equal.

We used Tukey's test as a post hoc test and as a result, there are significant differences in mean loan amount between all three groups.

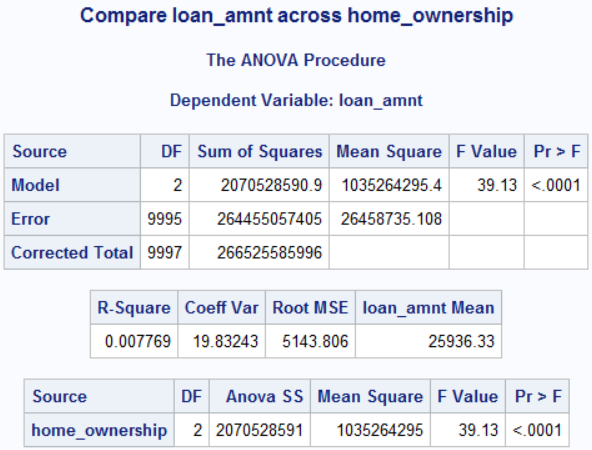
**DATA** q10; SET loan; IF home\_ownership='OTHER' THEN delete; **RUN**;

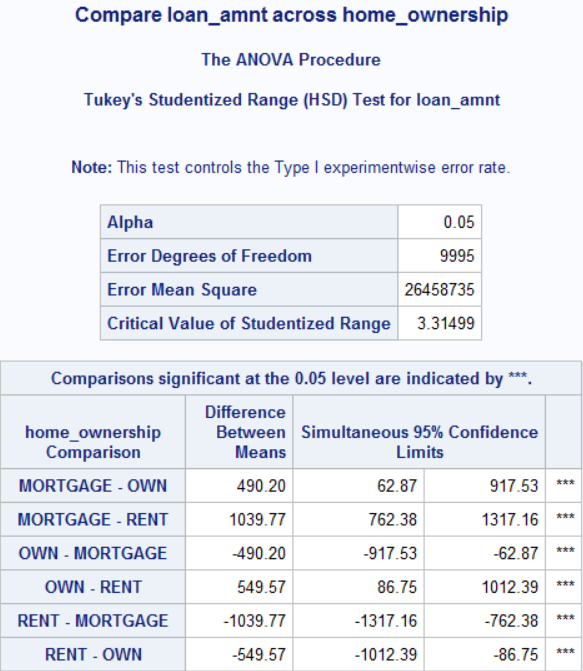
**PROC** **ANOVA** DATA=q10;

CLASS home\_ownership; MODEL loan\_amnt = home\_ownership;

MEANS home\_ownership / tukey;

TITLE 'Compare loan\_amnt across home\_ownership'; **RUN**;





**11. Create a new variable INCOME that divides all consumers into three groups based on whether they are high income, medium income or low income.**

1. According to quartiles. Q1, IQR, Q3 for low, medium, high income respectively.
2. Median: 90,000
3. 25th percentile: 70,000
4. 75th percentile: 120,000

**PROC** **UNIVARIATE** DATA=q10; VAR annual\_inc; TITLE 'Annual Income'; **RUN**;

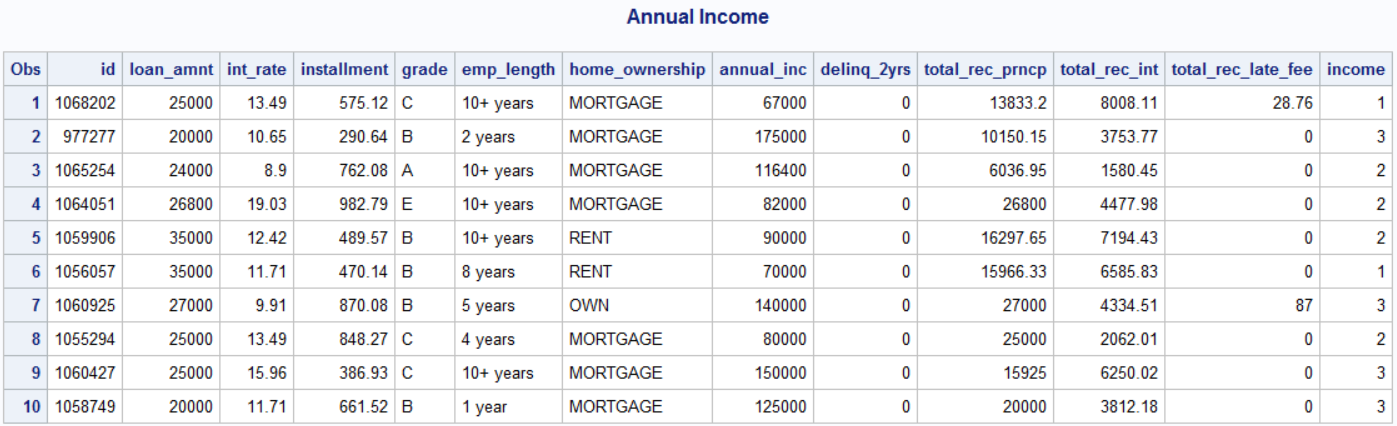
**DATA** income; SET q10;

IF annual\_inc <= **70000** THEN income = **1**;

IF annual\_inc > **70000** AND annual\_inc <= **120000** THEN income = **2**;

IF annual\_inc > **120000** THEN income = **3**;

**RUN**;

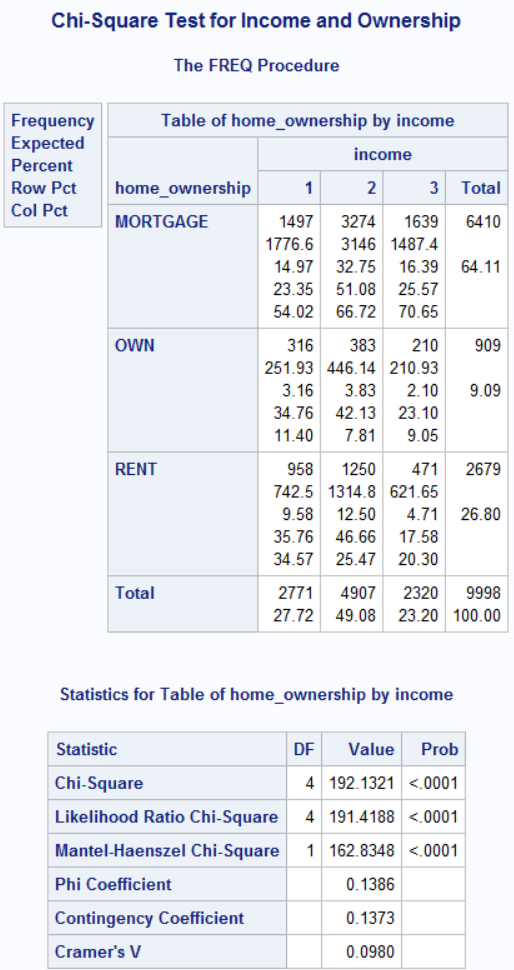


**12. Check whether there is a relationship between income levels and home\_ownership using Chi-square test.**

1. Since the p-value (0.0001) is less than the significance level (0.05), we cannot accept the null hypothesis. Thus, we conclude that there is a relationship between income level and home ownership.
2. Medium income group are more likely to be renters.
3. Medium income group are more likely to be mortgage owners.

**PROC** **FREQ** DATA = income; TABLES home\_ownership\*income / expected chisq;

TITLE 'Chi-Square Test for Income and Ownership'; **RUN**;

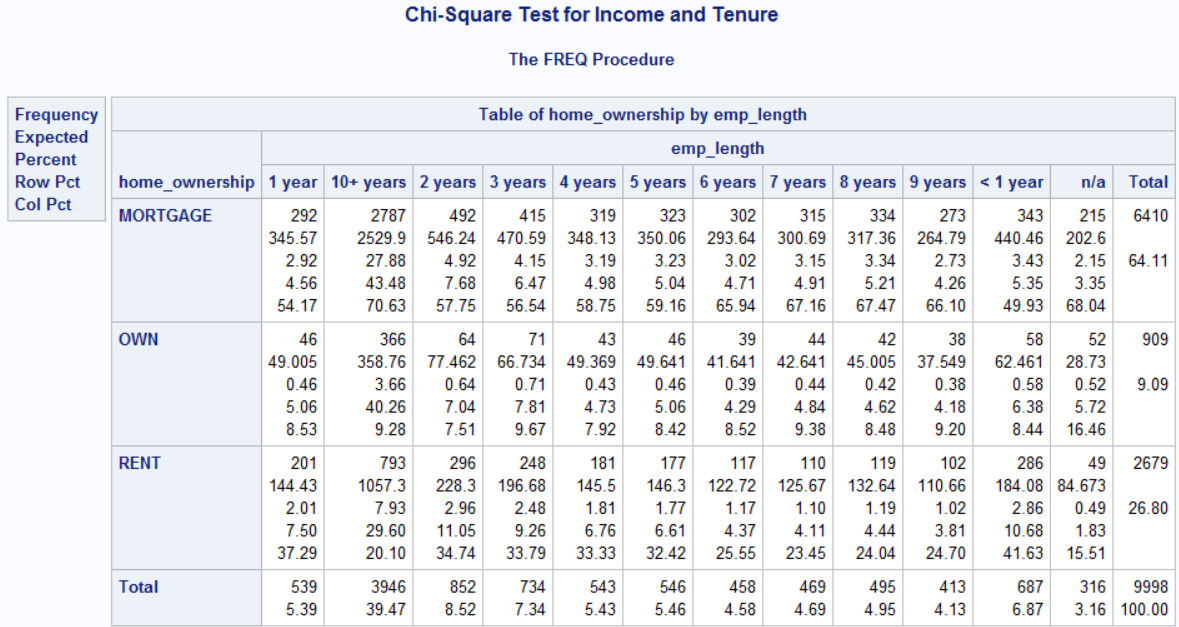


**13. Check if there is a relationship between home\_ownership and empl\_length using chi-square test.**

The p-value (0.0001) is less than the significance level (0.05). We reject the null and conclude that there is a relationship between home ownership and job tenure.

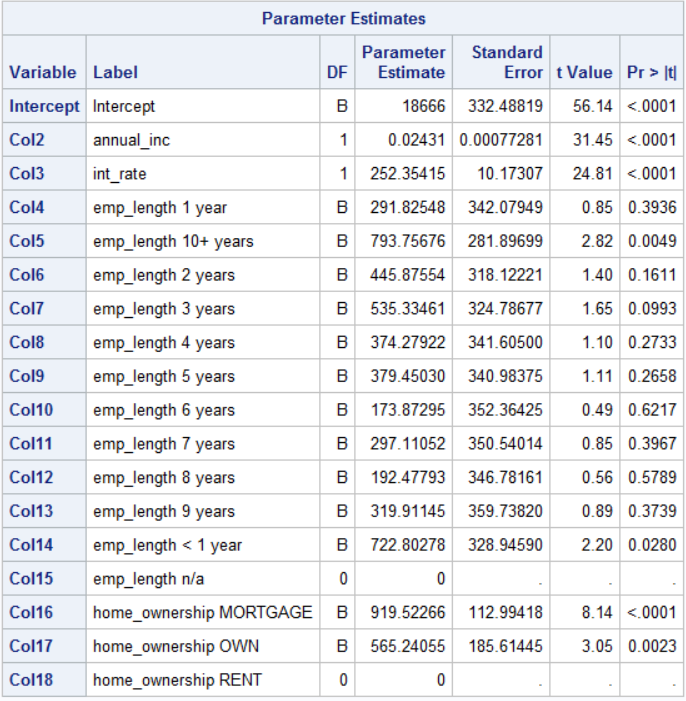
People with a job tenure of 10 years or more are much more likely to be mortgage owners.

**PROC** **FREQ** DATA = income; TABLES home\_ownership\*emp\_length / expected chisq; TITLE 'Chi-Square Test for Income and Tenure'; **RUN**;



**14. Run a regression model to explain the loan amount. Use as independent variables annual\_inc, interest rate, emp\_length, home\_ownership.**

1. F-value (104.37) is the ratio of mean square model and mean square residual; it is used for the null hypothesis that the model does not explain the variance of our response variable; thus, testing for the overall significance of the model. Our p-value (0.0001) associated with F-value is smaller than 0.05, thus we reject the null and conclude our model explains a statistically significant proportion of the variance.
2. 'annual income', 'interest rate', 'emp length 10+ years', 'emp length < 1year', 'Mortgage', 'Own', 'Rent'
3. A $1 increase in annual income would increase the loan amount by $0.02. It is significant with a t-value of 31.35 and a p-value < 0.0001.
4. The coefficient for renter is 0 because it’s the reference category with respect to home\_ownership variable. The intercept estimates the mean of loan amount for renters given emp length = unknown and the rest of the variables are 0.



**15. Run a regression model to explain the delinq\_2yrs. Use as independent variables annual\_inc, interest rate, emp\_length, home\_ownership and loan amount.**

1. F-value (9.48) is the ratio of mean square model and mean square residual.

Our p-value(0.0001) associated with F-value is smaller than 0.05, thus we reject the null and conclude our model explains a statistically significant proportion of the variance.

1. annual\_inc, int\_rate, mortgage, loan\_amnt
2. A $1 increase in loan amount would decrease the 2-year delinquency by 0.000006.

It's significant as the |t-value| is greater than 1.96 and the p-value (0.0004) is smaller than 0.05

1. The coefficient for renter is 0 because it's the reference category of with respect to home\_ownership variable. The intercept estimates the average 2-year delinquency for renters given emp length = unknown and the rest of the variables are 0.

Although I used linear regression to answer the question (because it asks for F-value), I think we need to use Poisson regression as the response variable ‘delinq\_2yrs’ is a discrete count variable and follows the Poisson distribution and zero inflated, thus, failing the assumptions of linear relationship and normality. We could use PROC COUTNREG to achieve that.

