**PROJECT ONE DATA MANAGEMENT—ANY COMMENTS NEEDED PASTE THEM HERE**

THE FIRST THING WE DO IS LOOKING AT VISUALIZATIONS. WE START WITH EXAMINING

--HOW PREVALENT DEFAULTS ARE,

--WHETHER THE DEFAULT RATE CHANGES BY LOAN GRADE

--NUMBER OF DELINQUENCIES

--A COUPLE OF SCATTER PLOTS OF DEFAULTS AGAINST LOAN AMOUNT AND INCOME

Defaults is set to no 86% of the time.

Defaults is set to yes 14% of the time.

Hence, defaults are only 14% prevalent

Does the default prevalence change by loan grade

Indeed, they change. We need to bear in mind, however, that we have more data for grade A and B and progressively less data for the other grades.

The same scenario applies to defaults by number of delinquencies.

There is less data as the number of delinquencies increases.

For the fourth plot, what we can see is that in general there are less people with a higher loan amount and as the loan amount increases so does the number of defaults, but this is due to the fact that there are less people in this category.

As the annual income increases, the number of default decreases but so does the number of points overall. I.E. there are less people defaulting with a high annual income but this is due to less people having a higher annual income.

dti🡪 debt to income ratio

**Add one more visualization of your own**

# In the same axes, produce box plots of the interest rate for every value of delinquencies

# boxplot with colour for different home\_ownership

ggplot(lc\_clean, aes(x = home\_ownership, y = int\_rate, colour = home\_ownership)) +

geom\_boxplot() +

coord\_flip() +

scale\_y\_continuous(labels=scales::percent) +

labs(y = "Interest Rate", x = "", title = "Home Ownership to a certain extend influences awarded interest rates", subtitle = "Distribution of interest rates by home ownership type") +

theme\_bw() +

theme(legend.position = "none")

WHAT CAN SEE ABOUT THIS, WE CAN SEE SEVERAL THINGS

In the first place, we see that most of the data corresponds to people that either have rented or have a mortgage for the property. The relative percentage of default for people that have rented the property is 15.17% whereas for those that have the property on a mortgage basis, the relative percent is 13.34%. Hence, the values are very similar and we cannot really conclude that home ownership has an impact on default.

>Q7. Using the model logistic 2 produce the ROC curve and calculate the AUC measure. Explain what the ROC shows and what the AUC measure means. Why do we expect the AUC of any predictive model to be between 0.5 and 1? Could the AUC ever be below 0.5 or above 1?

PROJECT1—DATA MANAGEMENT

#We will