week_01

Primary Findings

Mark Russeff, Renato Albolea, Shuai Ma

Research Questions: Controlling for relevant characteristics, is race/ethnicity associated with the outcome of a mortgage loan application?

Background The following abstract appeared in Alicia H. Munnell, Geoffrey M.B. Tootell, Lynn E. Browne, and James McEneaney (1996), "Mortgage Lending in Boston: Interpreting HMDA Data," American Economic Review 86, 25-53.

> The Home Mortgage Disclosure Act was enacted to monitor minority and low-income access to the mortgage market. The data collected for this purpose show that minorities are more than twice as likely to be denied a mortgage as whites. Yet variables correlated with both race and creditworthiness were omitted from these data, making any conclusion about race's role in mortgage lending impossible. The Federal Reserve Board of Boston collected additional variables important to the mortgage lending decision. . . .

As discussed in Munnell et al (1996), the HMDA data indicate whether an applicant's mortgage application was approved and provide several demographic characteristics. In 1990, following the request of the Federal Reserve Board of Boston, lending institutions in the Boston area provided additional information relevant to mortgage lending decisions. In light of the relatively small number of mortgage loan applications made by minorities, these extra variables were collected for all applications by blacks and Hispanics and for a random sample of those by whites.

> All applicants are non-Hispanic white, non Hispanic black, or Hispanic. In 1990 about 94% of Boston residents were white, Black, or Hispanic.

Data

Loading the data and looking the data structure

```
#load the persional data to variable Base
base <- read_csv(here('raw_data','MLD Data File-1.csv'))
```

```
## Parsed with column specification:
## cols(
##   MARRIED = col_character(),
##   GDLIN = col_double(),
##   OBRAT = col_double(),
##   BLACK = col_double(),
##   HISPAN = col_double(),
##   MALE = col_character(),
##   APPROVE = col_double(),
##   LOANPRC = col_double()
## )
```

```
base %>% str()
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 1989 obs. of  8 variables:
##  $ MARRIED: chr  "0" "1" "0" "1" ...
##  $ GDLIN  : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ OBRAT  : num  34.5 34.1 26 37 32.1 33 36 37 30.7 49 ...
##  $ BLACK  : num  0 0 0 0 0 0 0 0 0 0 ...
```

```
##  $ HISPAN : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ MALE   : chr  "." "1" "1" "1" ...
##  $ APPROVE: num  1 0 1 1 1 1 1 1 1 1 ...
##  $ LOANPRC: num  0.754 0.8 0.895 0.6 0.896 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   MARRIED = col_character(),
##   ..   GDLIN = col_double(),
##   ..   OBRAT = col_double(),
##   ..   BLACK = col_double(),
##   ..   HISPAN = col_double(),
##   ..   MALE = col_character(),
##   ..   APPROVE = col_double(),
##   ..   LOANPRC = col_double()
##   .. )
```

We can see that the features Married, Male, and Guide Line have wrong types of data.

```
base %>% group_by(MALE) %>% summarise(n())
```

```
## # A tibble: 3 x 2
##   MALE  `n()`
##   <chr> <int>
## 1 .        15
## 2 0       369
## 3 1      1605
```

```
base %>% group_by(MARRIED) %>% summarise(n())
```

```
## # A tibble: 3 x 2
##   MARRIED `n()`
##   <chr>   <int>
## 1 .           3
## 2 0         678
## 3 1        1308
```

```
base %>% group_by(GDLIN) %>% summarise(n())
```

```
## # A tibble: 3 x 2
##   GDLIN `n()`
##   <dbl> <int>
## 1     0   171
## 2     1  1816
## 3   666     2
```

Since those features will be important in our analysis, we decided to exclude the wrong data.

```
base <- base %>%
  mutate(MARRIED = as.numeric(MARRIED),
         MALE = as.numeric(MALE)
)
```

```
## Warning: NAs introduced by coercion
```

```
## Warning: NAs introduced by coercion
```

```
base <- base %>% filter(GDLIN<=1 & is.na(MALE)==FALSE)
```

```
base <- base %>% filter(is.na(MARRIED)==FALSE)
```

Improving data structure and summarizing the data.

```
base <- base %>%
  mutate(MALE = as.factor(MALE),
         GDLIN = as.factor(GDLIN),
         APPROVE = as.factor(APPROVE)
  )

source(here("code","function_convert_CSV_to_vector.R"))
levels(base$MALE) <- convertCSV2Factor("MALE")
levels(base$GDLIN) <- convertCSV2Factor("GDLIN")
levels(base$APPROVE) <- convertCSV2Factor("APPROVE")
base %>% summary()
```

```
##     MARRIED                            GDLIN          OBRAT
##  Min.   :0.0000   Doesn't meet Guide Line: 171   Min.   : 0.00
##  1st Qu.:0.0000   Meet Guide Line        :1798   1st Qu.:28.00
##  Median :1.0000                                 Median :33.00
##  Mean   :0.6592                                 Mean   :32.39
##  3rd Qu.:1.0000                                 3rd Qu.:37.00
##  Max.   :1.0000                                 Max.   :95.00
##      BLACK             HISPAN            MALE          APPROVE
##  Min.   :0.00000   Min.   :0.00000   Female: 368   Rejected: 244
##  1st Qu.:0.00000   1st Qu.:0.00000   Male  :1601   Approved:1725
##  Median :0.00000   Median :0.00000
##  Mean   :0.09903   Mean   :0.05485
##  3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000
##     LOANPRC
##  Min.   :0.02105
##  1st Qu.:0.70000
##  Median :0.80000
##  Mean   :0.77032
##  3rd Qu.:0.89888
##  Max.   :2.57143
```

Excluding Loan percentage to less or equal to 100%.

```
base <- base %>% filter(LOANPRC<=1)
```

Understanding data regarding Blacks

```
base %>% filter(BLACK == 1) %>% summary()
```

```
##     MARRIED                            GDLIN         OBRAT
##  Min.   :0.0000   Doesn't meet Guide Line: 53   Min.   : 5.60
##  1st Qu.:0.0000   Meet Guide Line        :139   1st Qu.:31.00
##  Median :1.0000                                Median :35.00
##  Mean   :0.6094                                Mean   :35.03
##  3rd Qu.:1.0000                                3rd Qu.:38.90
##  Max.   :1.0000                                Max.   :63.00
##      BLACK       HISPAN      MALE          APPROVE       LOANPRC
##  Min.   :1   Min.   :0   Female: 50   Rejected: 64   Min.   :0.2899
##  1st Qu.:1   1st Qu.:0   Male  :142   Approved:128   1st Qu.:0.8000
##  Median :1   Median :0                              Median :0.8606
```

```
##  Mean   :1   Mean   :0                                  Mean   :0.8289
##  3rd Qu.:1   3rd Qu.:0                                  3rd Qu.:0.9023
##  Max.   :1   Max.   :0                                  Max.   :1.0000
```

```
obs <- base %>% filter(BLACK == 1) %>% summarise(n()) %>% as.numeric(.)
base %>% filter(BLACK == 1) %>%
  group_by(MALE,GDLIN) %>%
  summarise(n = n(),
            freq = round(n/obs*100,digits =1),
            perc_approve = round(mean(as.numeric(APPROVE))*100, digits = 2),
            min(OBRAT),
            perc_obrat = round(mean(OBRAT), digits = 2),
            max(OBRAT),
            perc_min_loanprc = round(min(LOANPRC) * 100, digits = 2),
            perc_loanprc = round(mean(LOANPRC) * 100, digits = 2),
            perc_max_loanprc = round(max(LOANPRC) * 100, digits = 2))
```

```
## # A tibble: 4 x 11
## # Groups:   MALE [2]
##   MALE  GDLIN     n  freq perc_approve `min(OBRAT)` perc_obrat `max(OBRAT)`
##   <fct> <fct> <int> <dbl>        <dbl>        <dbl>      <dbl>        <dbl>
## 1 Fema~ Does~    16   8.3         112.         28.7       41.0         63
## 2 Fema~ Meet~    34  17.7         182.         10.6       35.0         47
## 3 Male  Does~    37  19.3         111.          8         35.2         57.6
## 4 Male  Meet~   105  54.7         190.          5.6       34.1         58
## # ... with 3 more variables: perc_min_loanprc <dbl>, perc_loanprc <dbl>,
## #   perc_max_loanprc <dbl>
```

Understanding data regarding Hispanics

```
base %>% filter(HISPAN == 1) %>% summary()
```

```
##     MARRIED                            GDLIN        OBRAT           BLACK
##  Min.   :0.0000   Doesn't meet Guide Line:14   Min.   :14.60   Min.   :0
##  1st Qu.:0.0000   Meet Guide Line        :90   1st Qu.:29.00   1st Qu.:0
##  Median :1.0000                               Median :33.00   Median :0
##  Mean   :0.7115                               Mean   :33.32   Mean   :0
##  3rd Qu.:1.0000                               3rd Qu.:38.05   3rd Qu.:0
##  Max.   :1.0000                               Max.   :62.00   Max.   :0
##      HISPAN      MALE        APPROVE      LOANPRC
##  Min.   :1   Female:20   Rejected:23   Min.   :0.4009
##  1st Qu.:1   Male  :84   Approved:81   1st Qu.:0.8000
##  Median :1                             Median :0.8931
##  Mean   :1                             Mean   :0.8391
##  3rd Qu.:1                             3rd Qu.:0.9027
##  Max.   :1                             Max.   :1.0000
```

```
obs <- base %>% filter(HISPAN == 1) %>% summarise(n()) %>% as.numeric(.)
base %>% filter(HISPAN == 1) %>%
  group_by(MALE,GDLIN) %>%
  summarise(n = n(),
            freq = round(n/obs*100,digits =1),
            perc_approve = round(mean(as.numeric(APPROVE))*100, digits = 2),
            min(OBRAT),
            perc_obrat = round(mean(OBRAT), digits = 2),
            max(OBRAT),
```

```
            perc_min_loanprc = round(min(LOANPRC) * 100, digits = 2),
            perc_loanprc = round(mean(LOANPRC) * 100, digits = 2),
            perc_max_loanprc = round(max(LOANPRC) * 100, digits = 2))
```

```
## # A tibble: 4 x 11
## # Groups:   MALE [2]
##   MALE  GDLIN     n  freq perc_approve `min(OBRAT)` perc_obrat `max(OBRAT)`
##   <fct> <fct> <int> <dbl>        <dbl>        <dbl>      <dbl>        <dbl>
## 1 Fema~ Does~     2   1.9         150          32         43.0         53.9
## 2 Fema~ Meet~    18  17.3         189.         16.8       32.3         47
## 3 Male  Does~    12  11.5         117.         22         36.5         49
## 4 Male  Meet~    72  69.2         186.         14.6       32.8         62
## # ... with 3 more variables: perc_min_loanprc <dbl>, perc_loanprc <dbl>,
## #   perc_max_loanprc <dbl>
```

Understanding data regarding White

```
base %>% filter(HISPAN ==0 & BLACK ==0) %>% summary()
```

```
##     MARRIED                              GDLIN          OBRAT
##  Min.   :0.0000   Doesn't meet Guide Line: 100   Min.   : 0.00
##  1st Qu.:0.0000   Meet Guide Line        :1541   1st Qu.:27.60
##  Median :1.0000                                 Median :32.50
##  Mean   :0.6606                                 Mean   :31.99
##  3rd Qu.:1.0000                                 3rd Qu.:36.50
##  Max.   :1.0000                                 Max.   :95.00
##      BLACK       HISPAN      MALE          APPROVE        LOANPRC
##  Min.   :0   Min.   :0   Female: 291   Rejected: 148   Min.   :0.02105
##  1st Qu.:0   1st Qu.:0   Male  :1350   Approved:1493   1st Qu.:0.67708
##  Median :0   Median :0                                 Median :0.79861
##  Mean   :0   Mean   :0                                 Mean   :0.74782
##  3rd Qu.:0   3rd Qu.:0                                 3rd Qu.:0.89147
##  Max.   :0   Max.   :0                                 Max.   :1.00000
```

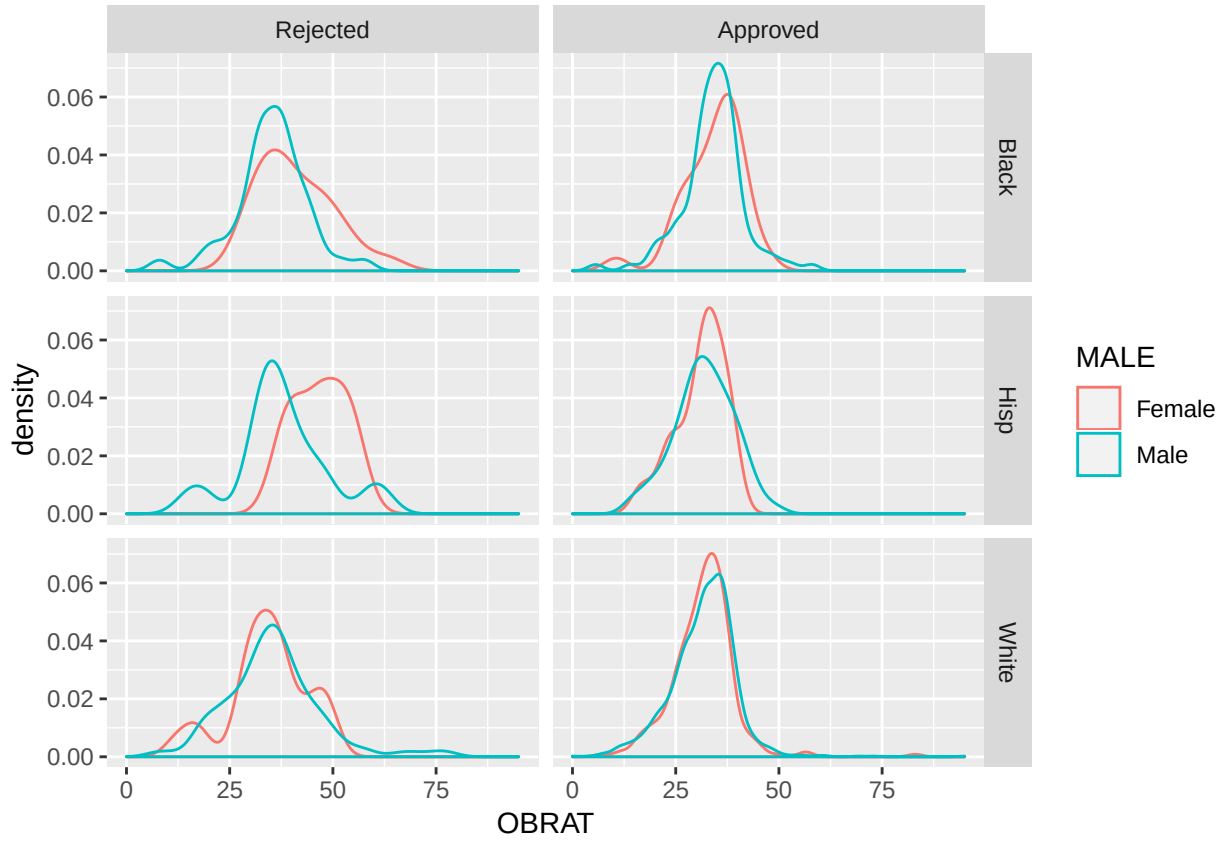
```
obs <- base %>% filter(HISPAN ==0 & BLACK ==0) %>% summarise(n()) %>% as.numeric(.)
base %>% filter(HISPAN ==0 & BLACK ==0) %>%
  group_by(MALE,GDLIN) %>%
  summarise(n = n(),
            freq = round(n/obs*100,digits =1),
            perc_approve = round(mean(as.numeric(APPROVE))*100, digits = 2),
            min(OBRAT),
            perc_obrat = round(mean(OBRAT), digits = 2),
            max(OBRAT),
            perc_min_loanprc = round(min(LOANPRC) * 100, digits = 2),
            perc_loanprc = round(mean(LOANPRC) * 100, digits = 2),
            perc_max_loanprc = round(max(LOANPRC) * 100, digits = 2))
```

```
## # A tibble: 4 x 11
## # Groups:   MALE [2]
##   MALE  GDLIN     n  freq perc_approve `min(OBRAT)` perc_obrat `max(OBRAT)`
##   <fct> <fct> <int> <dbl>        <dbl>        <dbl>      <dbl>        <dbl>
## 1 Fema~ Does~    12   0.7         117.         12         32.5           49
## 2 Fema~ Meet~   279  17           195.          6.99      31.8           83
## 3 Male  Does~    88   5.4         130.         16         36.6           95
## 4 Male  Meet~  1262  76.9         195.          0         31.7           78
## # ... with 3 more variables: perc_min_loanprc <dbl>, perc_loanprc <dbl>,
```

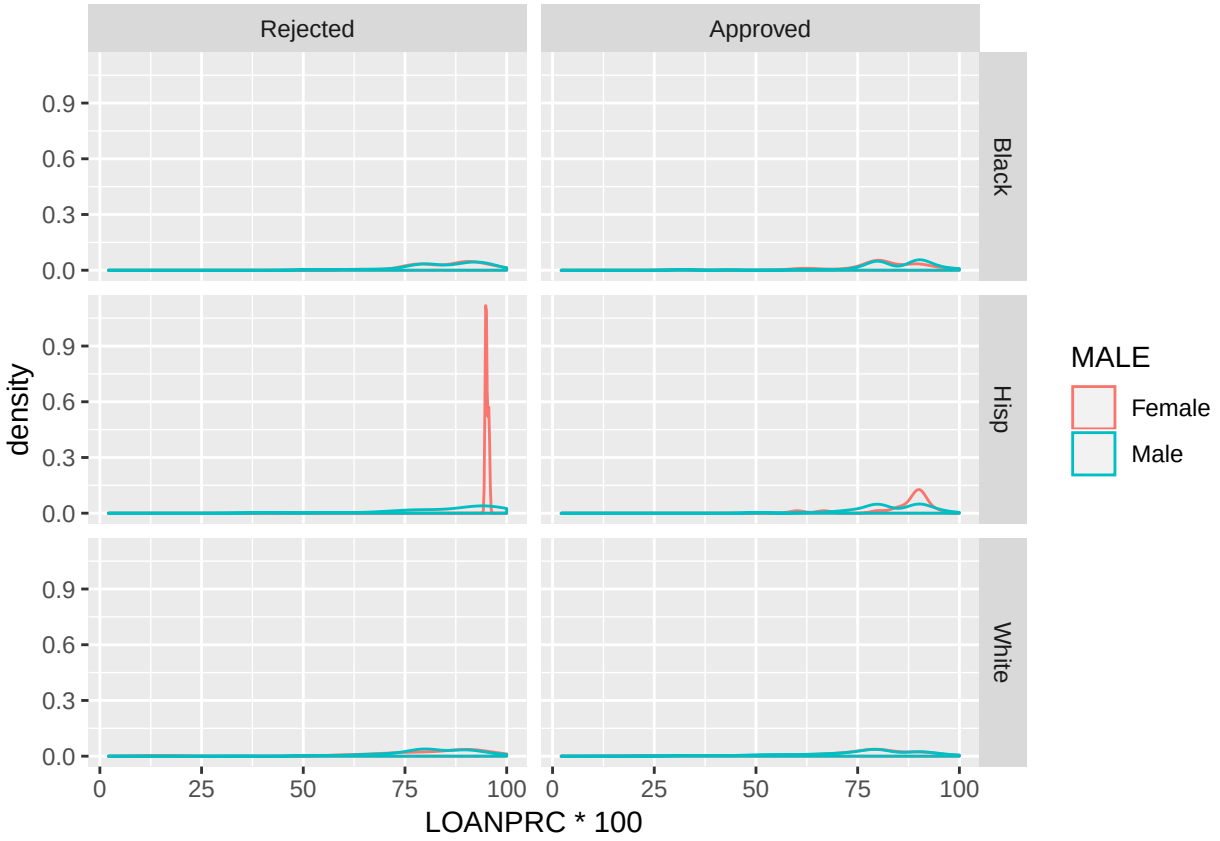
```
## #   perc_max_loanprc <dbl>
```

Graphical Analysis

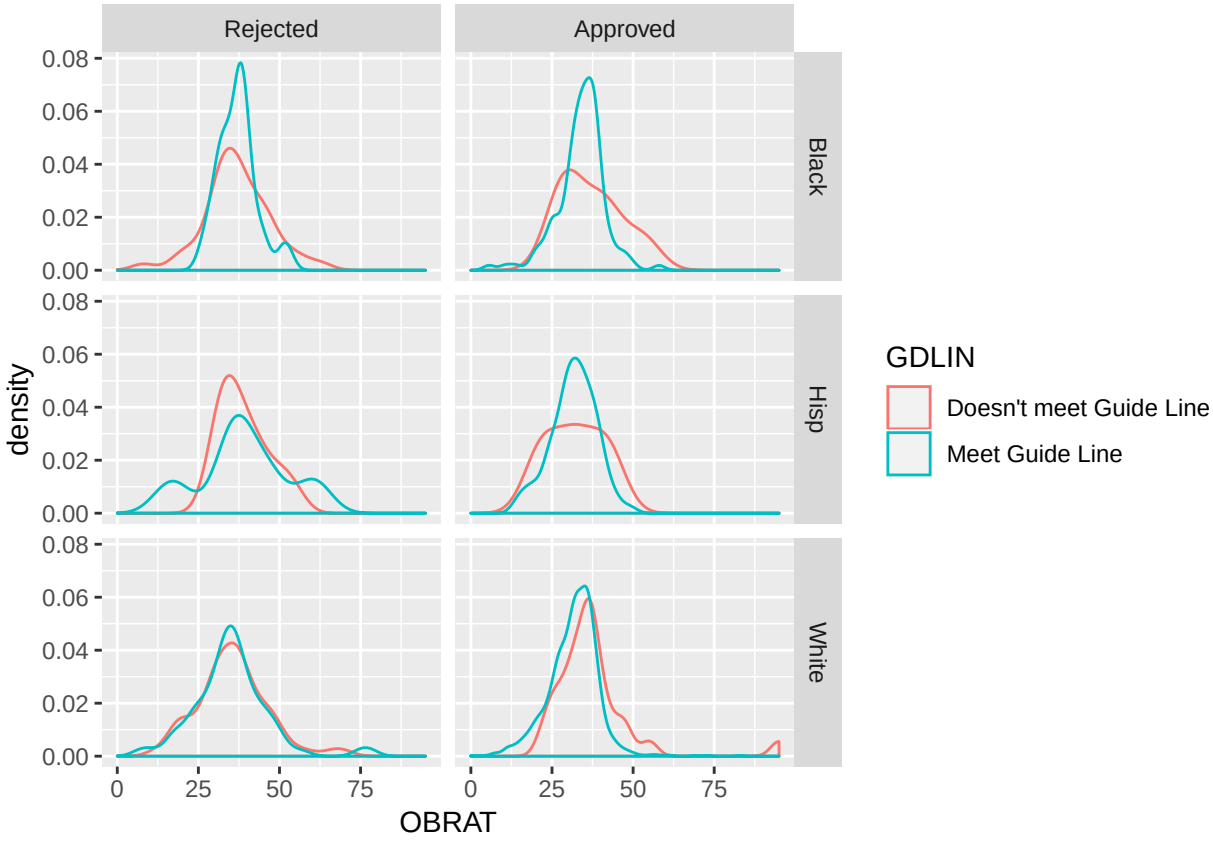
```
base <- base %>%
  mutate(race = case_when(
    BLACK == 1 ~ "Black",
    HISPAN == 1 ~ "Hisp",
    TRUE ~"White" ) )


base %>%
  ggplot(aes(x=OBRAT,color = MALE)) +
  geom_density() +
  facet_grid(race ~ APPROVE)
```



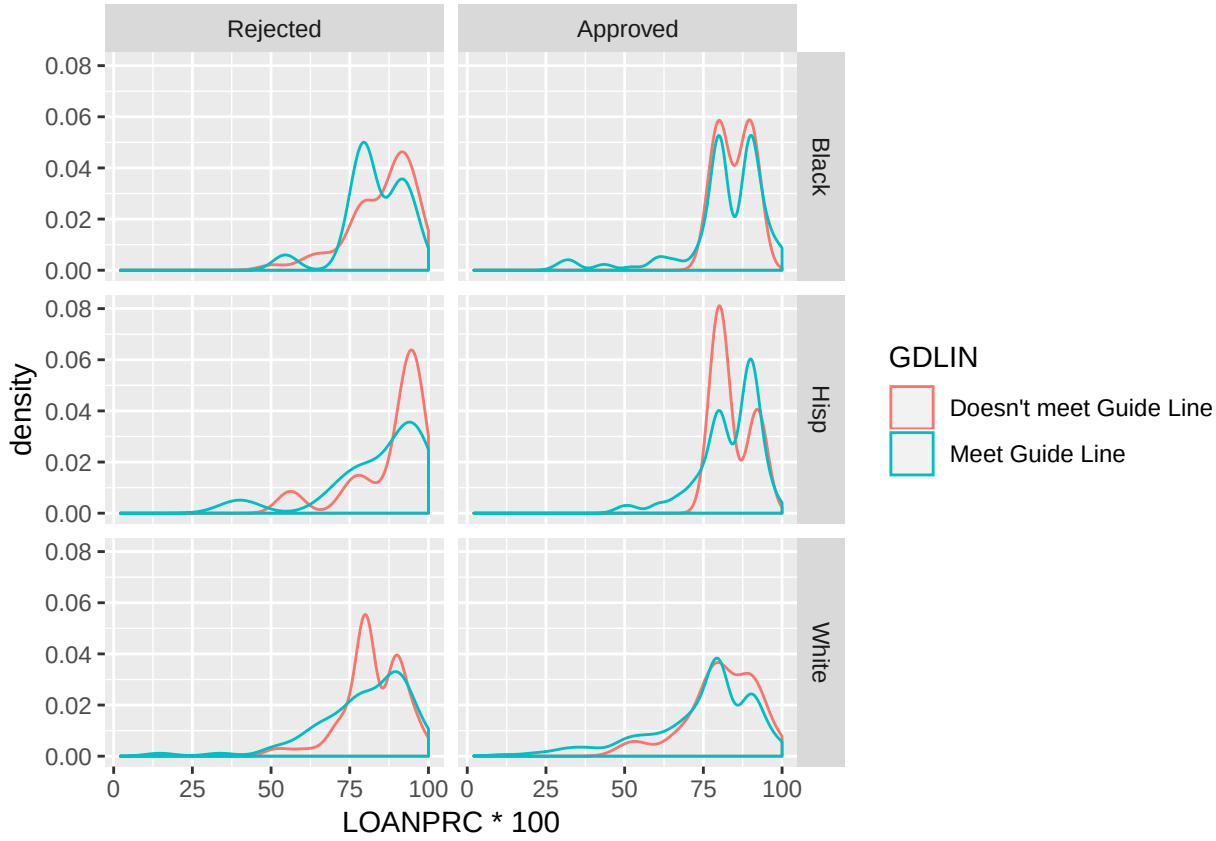

```
base %>%
  ggplot(aes(x=LOANPRC*100,color = MALE)) +
  geom_density() +
  facet_grid(race ~ APPROVE)
```




```
base %>%
  ggplot(aes(x=OBRAT,color = GDLIN)) +
  geom_density() +
  facet_grid(race ~ APPROVE)
```




```
base %>%
  ggplot(aes(x=LOANPRC*100,color = GDLIN)) +
  geom_density() +
  facet_grid(race ~ APPROVE)
```



Estimating Models

Model considering Race

```
#Estimate Logit Model
LogitModel = glm(APPROVE ~ OBRAT + BLACK + HISPAN, data = base,
                 family = "binomial")
summary(LogitModel)
```

```
##
## Call:
## glm(formula = APPROVE ~ OBRAT + BLACK + HISPAN, family = "binomial",
##     data = base)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7391   0.3497   0.4190   0.4759   1.6139
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.102332   0.308835  13.283  < 2e-16 ***
## OBRAT       -0.053552   0.008528  -6.280 3.39e-10 ***
## BLACK       -1.503975   0.179092  -8.398  < 2e-16 ***
## HISPAN      -1.004770   0.256139  -3.923 8.75e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1431.6  on 1936  degrees of freedom
## Residual deviance: 1308.4  on 1933  degrees of freedom
## AIC: 1316.4
##
## Number of Fisher Scoring iterations: 5
```

```
#stargazer(LogitModel,title="Results",header = FALSE)

#Generate Odds Ratios
exp(coef(LogitModel))
```

```
## (Intercept)       OBRAT       BLACK      HISPAN
##  60.4811395   0.9478570   0.2222449   0.3661287
```

```
#Define prototypical loan applicants (you will need more than 3)
prototype_black <- data.frame(OBRAT=mean(base$OBRAT),BLACK = 1, HISPAN = 0)
prototype_hisp <- data.frame(OBRAT=mean(base$OBRAT),BLACK = 0, HISPAN = 1)
prototype_white <- data.frame(OBRAT=mean(base$OBRAT),BLACK = 0, HISPAN = 0)

#Predict probabilities for prototypical individuals
prototype_black$predictedprob <- round(
  predict (LogitModel,
           newdata = prototype_black,
           type ="response")*100,
  digits = 1)


prototype_hisp$predictedprob <-  round(
  predict (LogitModel,
           newdata = prototype_hisp,
           type ="response")*100,
  digits = 1)
prototype_white$predictedprob <- round(
  predict (LogitModel,
           newdata = prototype_white,
           type ="response")*100,
  digits = 1)
prototype_black
```

```
##      OBRAT BLACK HISPAN predictedprob
## 1 32.36561     1      0          70.4
```

```
prototype_hisp
```

```
##      OBRAT BLACK HISPAN predictedprob
## 1 32.36561     0      1          79.6
```

```
prototype_white
```

```
##      OBRAT BLACK HISPAN predictedprob
## 1 32.36561     0      0          91.4
```

```
#Estimate Probit Model
ProbitModel = glm(APPROVE ~ OBRAT + BLACK + HISPAN, data = base,
                  family = "binomial" (link = "probit"))
```

```r
summary(ProbitModel)
```

```
##
## Call:
## glm(formula = APPROVE ~ OBRAT + BLACK + HISPAN, family = binomial(link = "probit"),
##     data = base)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7853   0.3509   0.4227   0.4790   1.4023
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.227906   0.162911  13.676  < 2e-16 ***
## OBRAT       -0.026830   0.004644  -5.778 7.56e-09 ***
## BLACK       -0.848779   0.104508  -8.122 4.60e-16 ***
## HISPAN      -0.538181   0.145827  -3.691 0.000224 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1431.6  on 1936  degrees of freedom
## Residual deviance: 1311.5  on 1933  degrees of freedom
## AIC: 1319.5
##
## Number of Fisher Scoring iterations: 5
```

```r
#Predict probabilities for prototypical individuals
prototype_black$predictedprob <- round(
  predict (ProbitModel,
           newdata = prototype_black,
           type ="response")*100,
  digits = 1)
prototype_hisp$predictedprob <-  round(
  predict (ProbitModel,
           newdata = prototype_hisp,
           type ="response")*100,
  digits = 1)
prototype_white$predictedprob <- round(
  predict (ProbitModel,
           newdata = prototype_white,
           type ="response")*100,
  digits = 1)

prototype_black
```

```
##      OBRAT BLACK HISPAN predictedprob
## 1 32.36561     1      0          69.5
```

```r
prototype_hisp
```

```
##      OBRAT BLACK HISPAN predictedprob
## 1 32.36561     0      1          79.4
```

```
prototype_white
```

```
##      OBRAT BLACK HISPAN predictedprob
## 1 32.36561     0      0          91.3
```

Model considering Gender

```
#Estimate Logit Model
LogitModel = glm(APPROVE ~ OBRAT + MALE, data = base,
                 family = "binomial")
summary(LogitModel)
```

```
##
## Call:
## glm(formula = APPROVE ~ OBRAT + MALE, family = "binomial", data = base)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6971   0.3912   0.4779   0.5348   1.9165
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.910538   0.337325  11.593  < 2e-16 ***
## OBRAT       -0.059923   0.008496  -7.053 1.75e-12 ***
## MALEMale     0.119326   0.176489   0.676    0.499
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1431.6  on 1936  degrees of freedom
## Residual deviance: 1378.3  on 1934  degrees of freedom
## AIC: 1384.3
##
## Number of Fisher Scoring iterations: 5
```

```
#Generate Odds Ratios
exp(coef(LogitModel))
```

```
## (Intercept)       OBRAT    MALEMale
##  49.9258045   0.9418374   1.1267374
```

```
#Define prototypical loan applicants
prototype_woman <- data.frame(OBRAT=mean(base$OBRAT),MALE = 0)
prototype_woman <-prototype_woman %>% mutate(MALE = as.factor(MALE))
levels(prototype_woman$MALE) <- "Female"
prototype_men <- data.frame(OBRAT=mean(base$OBRAT),MALE = 1)
prototype_men <- prototype_men %>% mutate(MALE = as.factor(MALE))
levels(prototype_men$MALE) <- "Male"
#Predict probabilities for prototypical individuals
prototype_woman$predictedprob <- round(
  predict (LogitModel,
           newdata = prototype_woman,
           type ="response")*100,
  digits = 1)
```

```r
prototype_men$predictedprob <-  round(
  predict (LogitModel,
           newdata = prototype_men,
           type ="response")*100,
  digits = 1)

prototype_woman
```

```
##      OBRAT   MALE predictedprob
## 1 32.36561 Female          87.8
```

```r
prototype_men
```

```
##      OBRAT MALE predictedprob
## 1 32.36561 Male            89
```

```r
#Estimate Probit Model
ProbitModel = glm(APPROVE ~ OBRAT + MALE , data = base,
                  family = "binomial" (link = "probit"))
summary(ProbitModel)
```

```
##
## Call:
## glm(formula = APPROVE ~ OBRAT + MALE, family = binomial(link = "probit"),
##     data = base)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7309   0.3967   0.4839   0.5384   1.6647
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.134905   0.177894  12.001  < 2e-16 ***
## OBRAT       -0.030125   0.004577  -6.582 4.65e-11 ***
## MALEMale     0.052976   0.094495   0.561    0.575
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1431.6  on 1936  degrees of freedom
## Residual deviance: 1381.7  on 1934  degrees of freedom
## AIC: 1387.7
##
## Number of Fisher Scoring iterations: 5
```

```r
#Predict probabilities for prototypical individuals
prototype_woman$predictedprob <- round(
  predict (ProbitModel,
           newdata = prototype_woman,
           type ="response")*100,
  digits = 1)
prototype_men$predictedprob <-  round(
  predict (ProbitModel,
           newdata = prototype_men,
```

```
          type ="response")*100,
  digits = 1)


prototype_woman
```

```
##      OBRAT   MALE predictedprob
## 1 32.36561 Female          87.7
```

```
prototype_men
```

```
##      OBRAT MALE predictedprob
## 1 32.36561 Male          88.7
```