Modeling Problem I

Karol Orozco & Charles Hanks

Predicting Province

```
knitr::opts_chunk$set(echo = TRUE, message = FALSE, warning = FALSE)

library(tidyverse)
library(formatR)
library(moderndive)
library(skimr)

wine_pinot <- readRDS(gzcon(url(
  "https://github.com/karolo89/machine_learning_assignment/raw/main/pinot.rds")))
```

```
#adding log price column
pinot <- wine_pinot %>%
  mutate(lprice = log(price))

pinot <- pinot %>%
  mutate(id = as.factor(id))%>%
  mutate(year = as.factor(year))

summary(pinot)
```

```
      id          province             price             points     
1      :   1   Length:8380        Min.   :   7.00   Min.   :80.00  
2      :   1   Class :character   1st Qu.:  31.00   1st Qu.:88.00  
3      :   1   Mode  :character   Median :  45.00   Median :90.00  
4      :   1                      Mean   :  52.52   Mean   :89.98  
5      :   1                      3rd Qu.:  60.00   3rd Qu.:92.00  
6      :   1                      Max.   :2500.00   Max.   :98.00  
(Other):8374                                                        
```

```
     year      description            lprice     
 2014   :2046   Length:8380        Min.   :1.946  
 2013   :1819   Class :character   1st Qu.:3.434  
 2012   :1505   Mode  :character   Median :3.807  
 2015   : 815                      Mean   :3.779  
 2011   : 582                      3rd Qu.:4.094  
 2010   : 502                      Max.   :7.824  
 (Other):1111                                     
```

Preliminary EDA, Feature Engineering Brainstorm, Initial Thoughts

```
pinot %>%
  group_by(province) %>%
  summarize(prov_freq = n(),
            percent_of_ds = round(prov_freq/8380,2))
```

```
# A tibble: 6 x 3
  province          prov_freq percent_of_ds
  <chr>                 <int>         <dbl>
1 Burgundy               1193          0.14
2 California             3959          0.47
3 Casablanca_Valley       131          0.02
4 Marlborough             229          0.03
5 New_York                131          0.02
6 Oregon                 2737          0.33
```

```
#nearly half of wines are californian, good to know...

pinot %>%
  filter(str_detect(description, "[Oo]ak")) %>%
  nrow()
```

```
[1] 1301
```

```
#1301/8380 have the work oak in description

pinot %>% filter(str_detect(description, "[Oo]ak")) %>%
  group_by(province) %>% summarize(prov_freq = n(),
                                   oak_perc = round(prov_freq/1301,2))
```

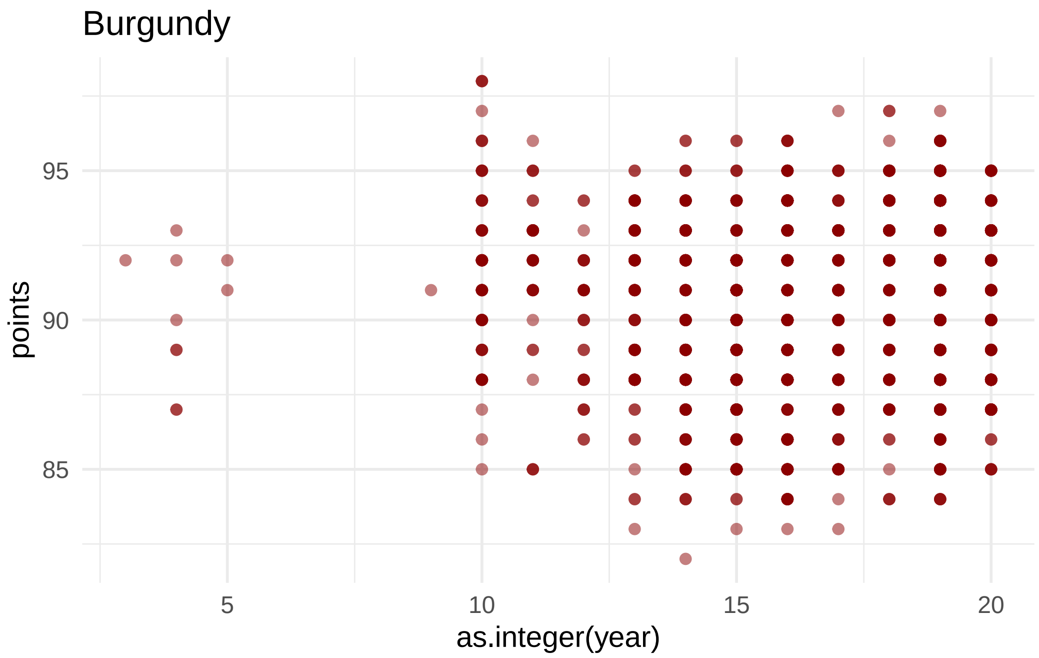
```
# A tibble: 6 x 3
  province          prov_freq oak_perc
  <chr>                 <int>    <dbl>
1 Burgundy                  8     0.01
2 California              739     0.57
3 Casablanca_Valley        64     0.05
4 Marlborough              32     0.02
5 New_York                  9     0.01
6 Oregon                  449     0.35
```

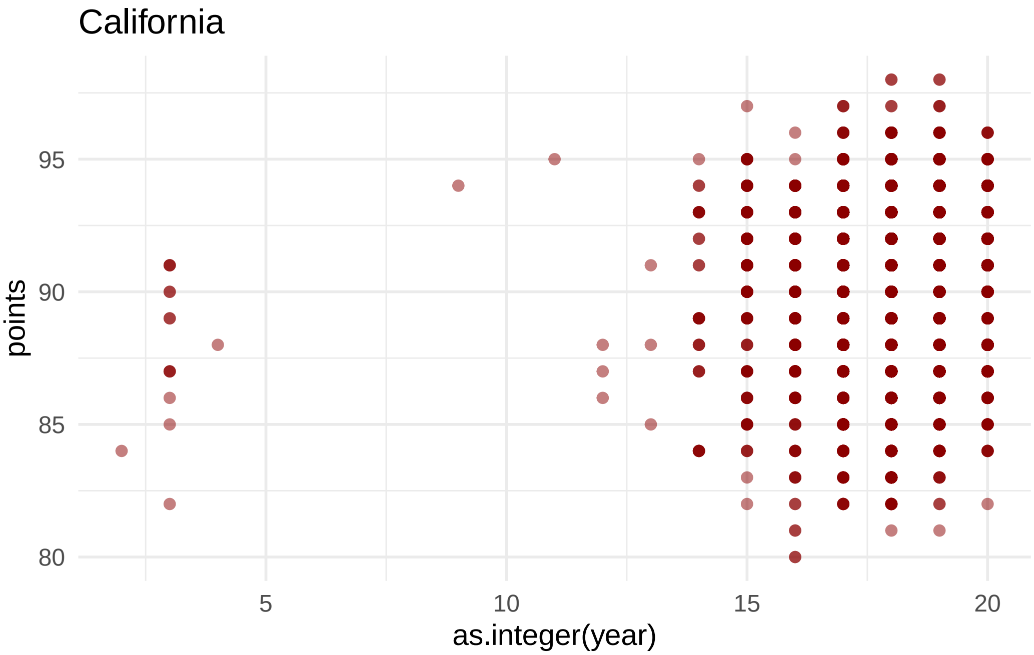
```
#it is likely California or Oregon if there is oak in the description

#some french language patterns to think about developing a regex from:
# "_de_" / "d'"
# "name-name"
# accented letters: "é","ô",
# "St."

pinot %>%
  group_by(province) %>%
  summarize(avgPrice = mean(price),
            avgPoints = mean(points))
```

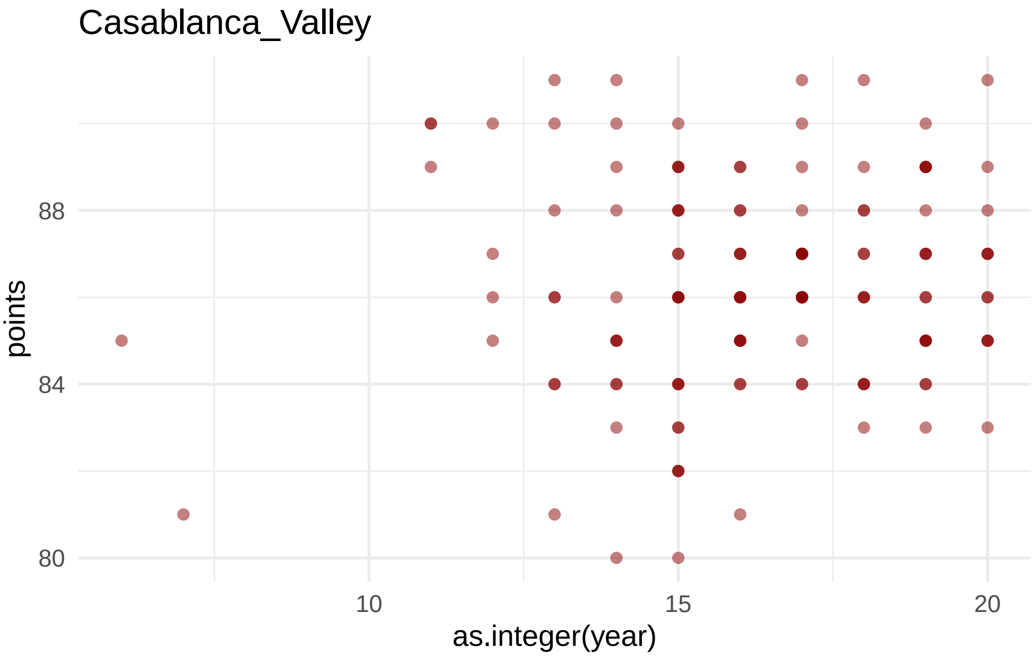
```
# A tibble: 6 x 3
  province          avgPrice avgPoints
  <chr>                <dbl>     <dbl>
1 Burgundy              98.0      90.4
2 California            47.5      90.5
3 Casablanca_Valley     21.1      86.3
4 Marlborough           27.7      87.6
5 New_York              25.7      87.7
6 Oregon                44.9      89.5
```

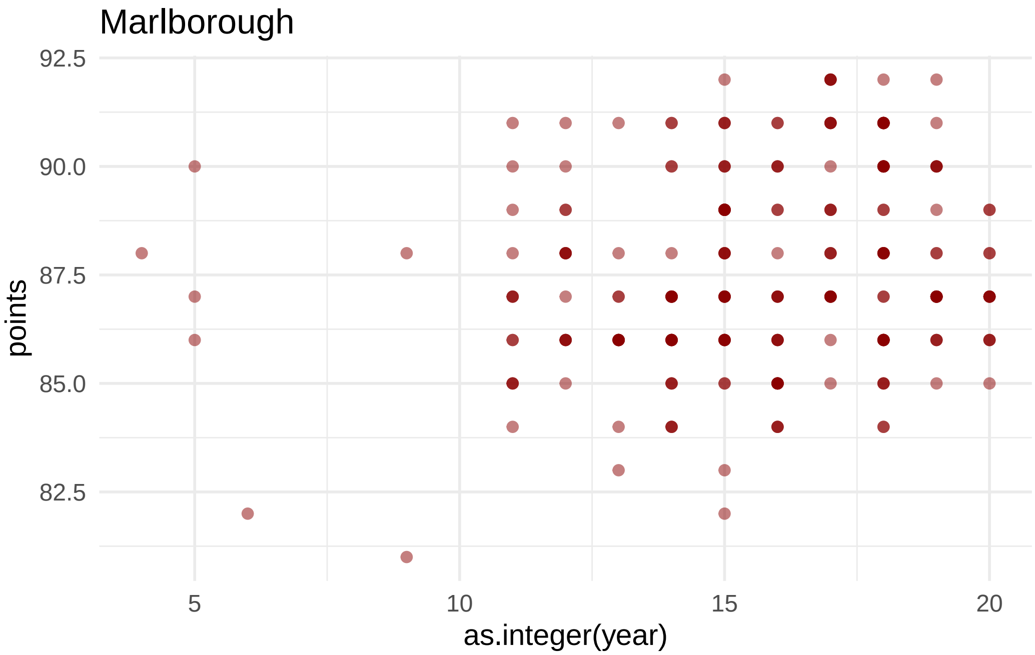
```
# Burgundy wines are on average significantly more expensive...
# and casablanca valley wines on average have the lowest price and score.

#which wines do people recommend waiting before drinking? i.e "drink from XXXX"

#some words to check out: "edge","tannins","dense","firm", oregon pinot is fruity.
```

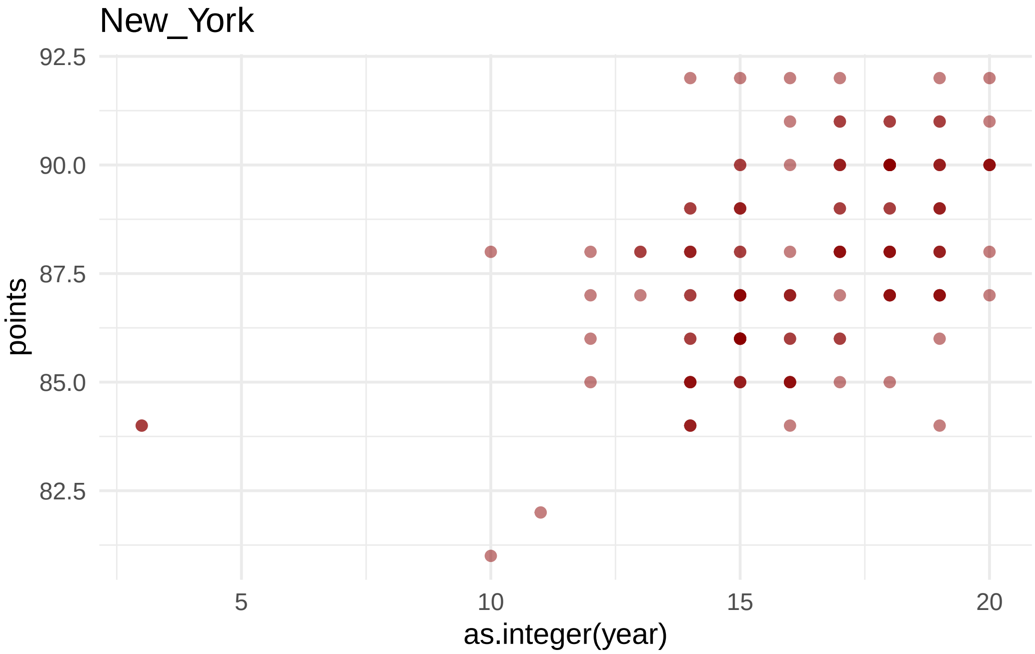
```
province_vec = c("Burgundy", "California", "Casablanca_Valley","Marlborough",
                 "New_York", "Oregon")

for(i in province_vec){
  plot = ggplot(pinot %>%
                  filter(province == i), aes(x = as.integer(year), y = points)) +
    geom_point(alpha =.5, color = "red4") +
    ggtitle(i)+
    theme_minimal()

  print(plot)
}
```

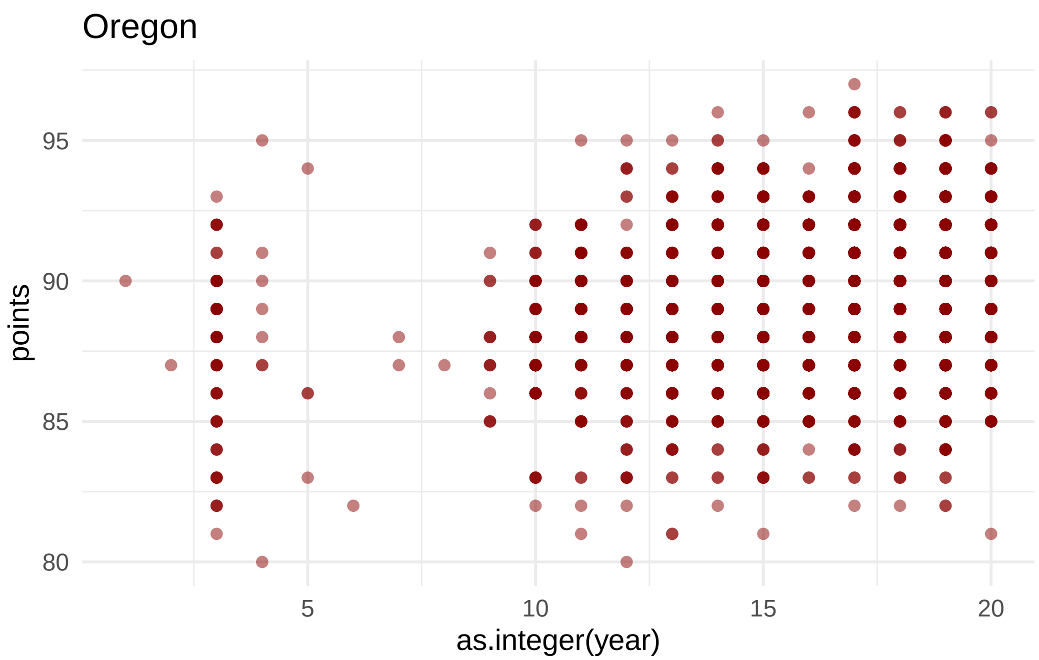



California
95
90
85
80
points
5
10
15
20
as.integer(year)

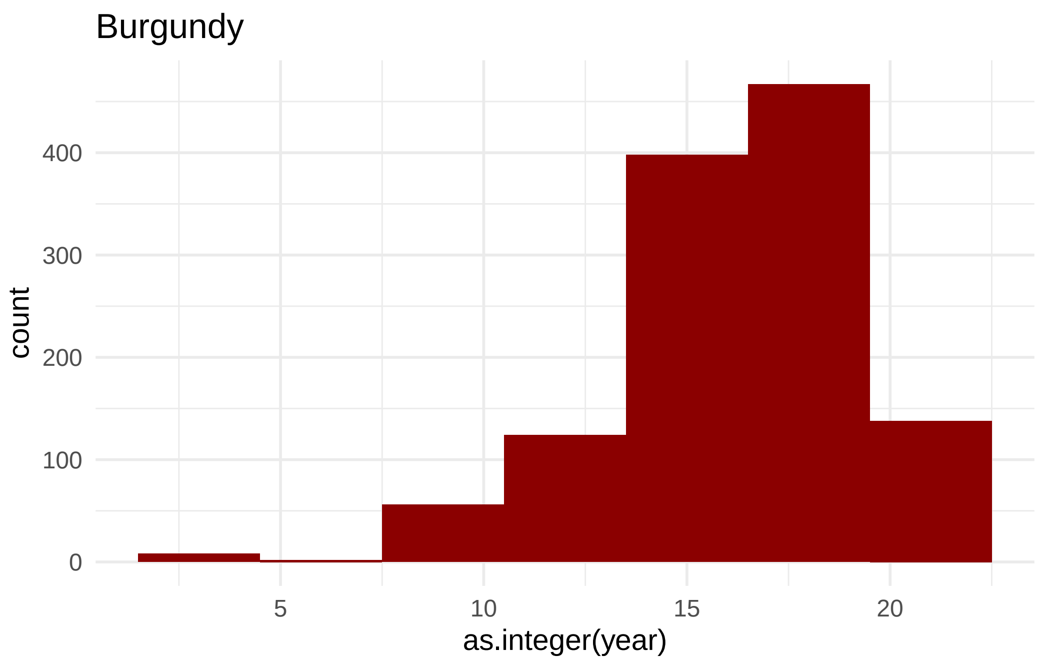

Casablanca_Valley
88
84
80
points
10
15
20
as.integer(year)

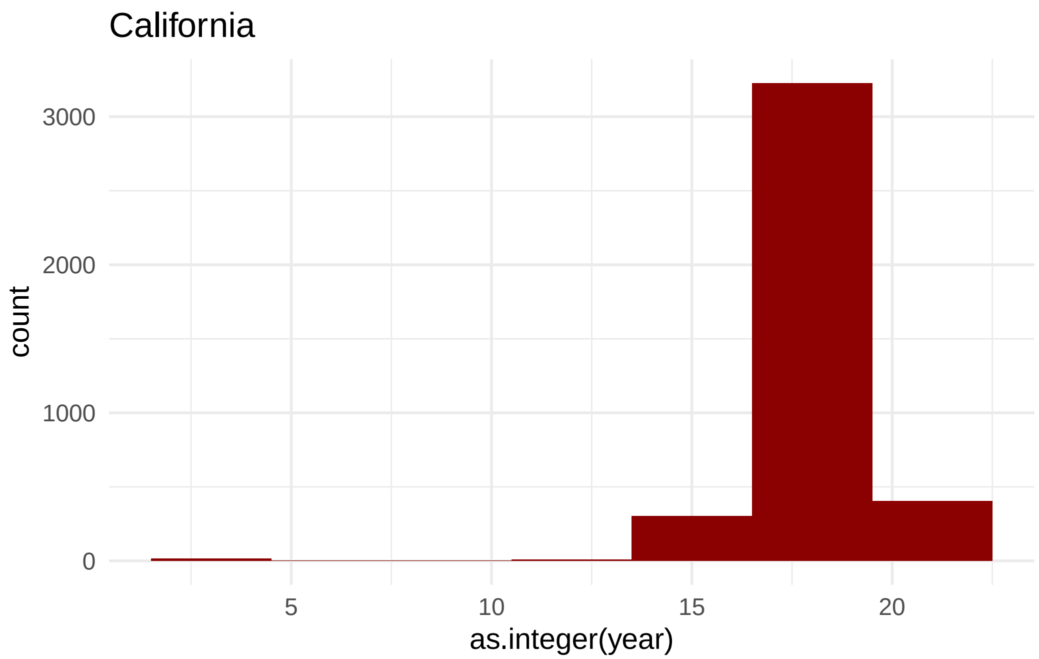

Marlborough
92.5
90.0
87.5
85.0
82.5
points
5
10
15
20
as.integer(year)

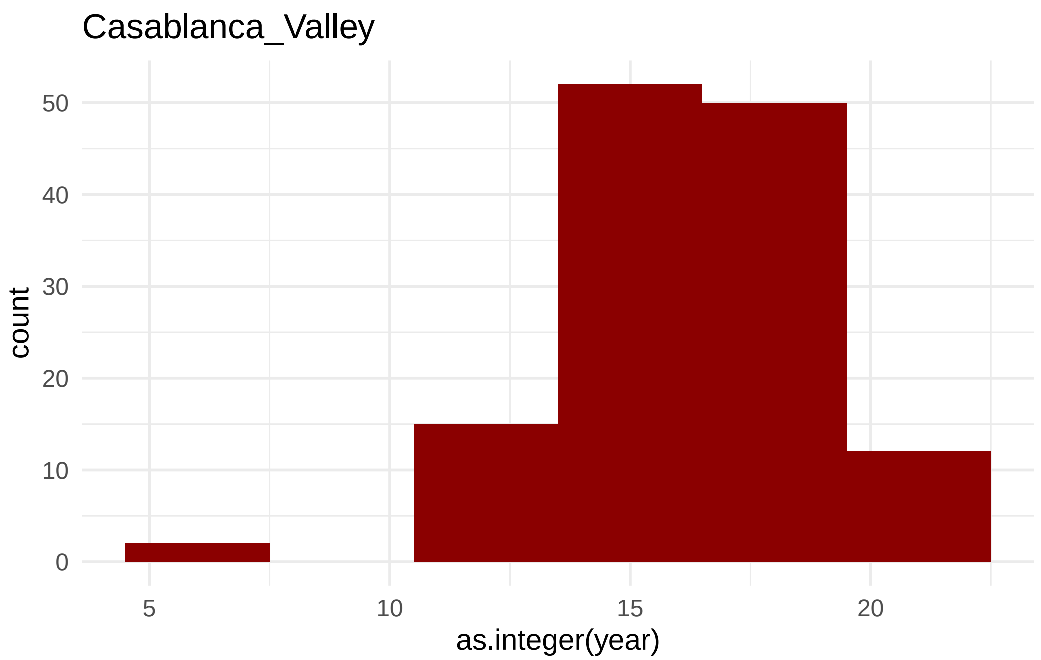

New_York
92.5
90.0
87.5
85.0
82.5
points
5
10
15
20
as.integer(year)

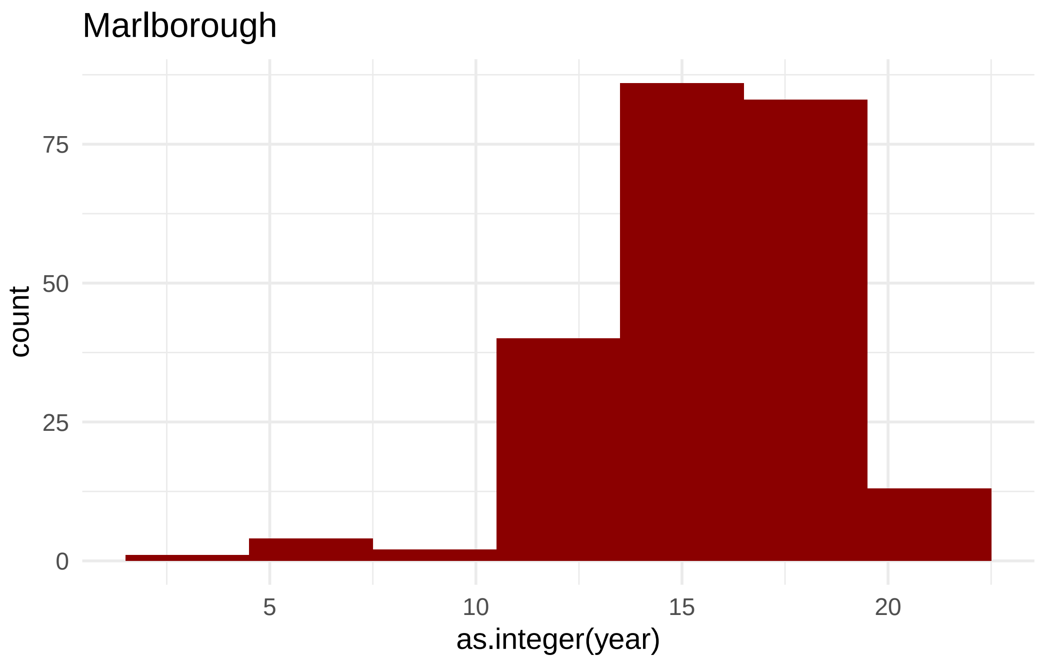


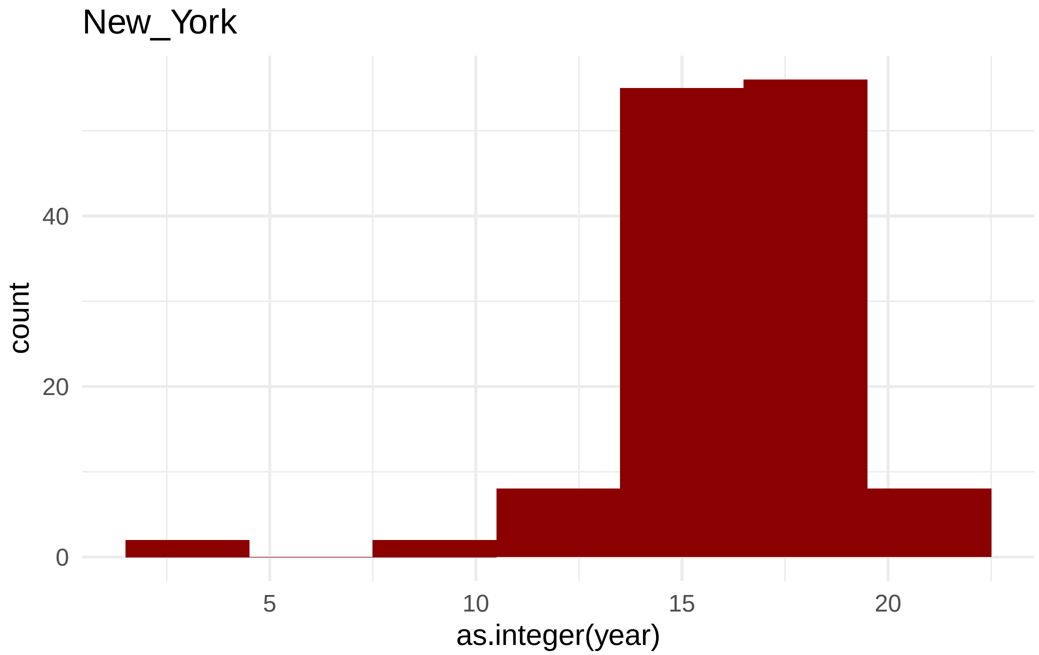
```
for(i in province_vec){
  plot2 = ggplot(pinot %>%
                   filter(province == i), aes(x = as.integer(year))) +
    geom_histogram(binwidth =3, fill = "red4") +
    ggtitle(i)+
    theme_minimal()
  print(plot2)
}
```

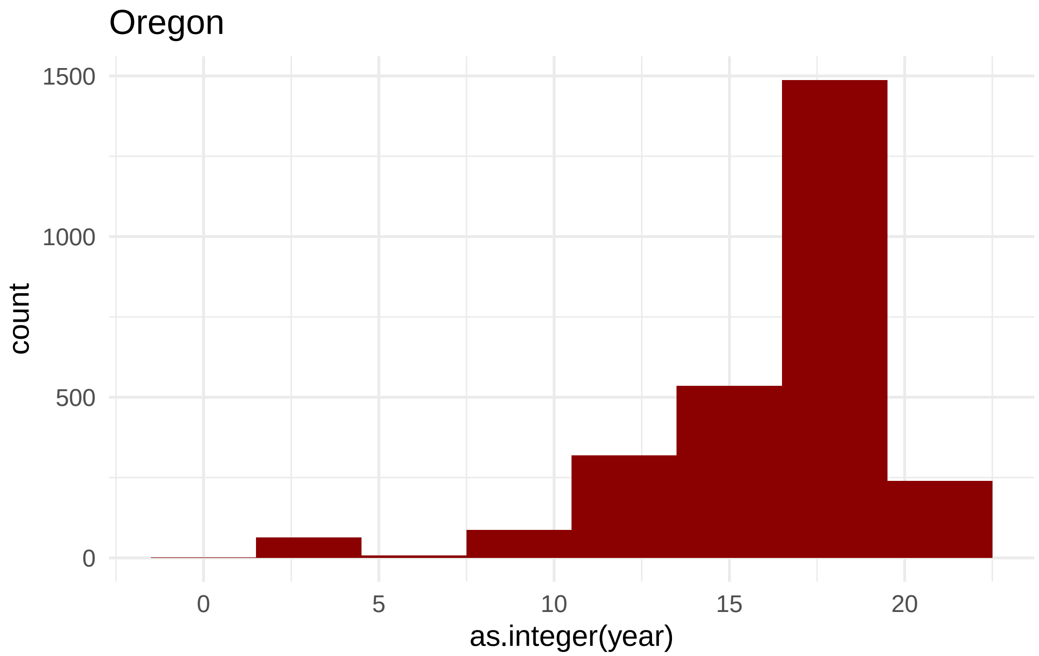

Burgundy
count
400
300
200
100
0
5
10
15
20
as.integer(year)


California
count
3000
2000
1000
0
5
10
15
20
as.integer(year)


Casablanca_Valley
count
0
10
20
30
40
50
5
10
15
20
as.integer(year)


Marlborough
count
0
25
50
75
5
10
15
20
as.integer(year)


New_York
count
0
20
40
5
10
15
20
as.integer(year)


Oregon
count
0
500
1000
1500
0
5
10
15
20
as.integer(year)

```
#Some findings from viz:
#california pinot noir production did not begin until ~2008, then exploded!
#before year 2000, likely to be oregon
#burgundy pinots score high around 2005,
#after almost no burgundy pinots between 2000 and 2005
#California pinot game WAY STRONG between 2010 and 2015
#New York pinot score high between 2008 and 2015
#What happened around 2014?? Counts drop across provinces....
```

Preprocessing (3pts)

1. Preprocess the dataframe that you created in the previous question using centering and scaling of the numeric features
2. Create dummy variables for the year factor column

Running KNN (5pts)

1. Split your data into an 80/20 training and test set
2. Use Caret to run a KNN model that uses your engineered features to predict province

- use 5-fold cross validated subsampling
- allow Caret to try 15 different values for K

3. Display the confusion matrix on the test data

Kappa (2pts)

Is this a good value of Kappa? Why or why not?

Answer: (write your answer here)

Improvement (2pts)

Looking at the confusion matrix, where do you see room for improvement in your predictions?

Answer: (write your answer here)