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/r
```

```
summary(wine_pinot)
```

```
       id          province             price             points     
 Min.   :   1   Length:8380        Min.   :   7.00   Min.   :80.00  
 1st Qu.:2096   Class :character   1st Qu.:  31.00   1st Qu.:88.00  
 Median :4190   Mode  :character   Median :  45.00   Median :90.00  
 Mean   :4190                      Mean   :  52.52   Mean   :89.98  
 3rd Qu.:6285                      3rd Qu.:  60.00   3rd Qu.:92.00  
 Max.   :8380                      Max.   :2500.00   Max.   :98.00  
      year      description       
 Min.   :1996   Length:8380       
 1st Qu.:2011   Class :character  
 Median :2013   Mode  :character  
 Mean   :2012                     
 3rd Qu.:2014                     
 Max.   :2015                     
```

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

pinot <- pinot %>%
  mutate(id = as.factor(id))%>%
  mutate(year = as.factor(year))%>%
  select(id, year, province, lprice)

skim(pinot)
```

Table 1: Data summary

Name	pinot
Number of rows	8380
Number of columns	4
Column type frequency:	
character	1
factor	2
numeric	1
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
province	0	1	6	17	0	6	0

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
id	0	1	FALSE	8380	1: 1, 2: 1, 3: 1, 4: 1
year	0	1	FALSE	20	201: 2046, 201: 1819, 201: 1505, 201: 815

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
lprice	0	1	3.78	0.55	1.95	3.43	3.81	4.09	7.82	