# Final Project

*Gordon Li, Nakrit Manasilp*

*11/21/2018*

This is a case study of the California Housing Prices (Median house prices for California districts derived from the 1990 census) dataset. Each observation represents a district with the following variables:

1. longitude: A measure of how far west a house is; a higher value is farther west
2. latitude: A measure of how far north a house is; a higher value is farther north
3. housingMedianAge: Median age of a house within a block; a lower number is a newer building
4. totalRooms: Total number of rooms within a block
5. totalBedrooms: Total number of bedrooms within a block
6. population: Total number of people residing within a block
7. households: Total number of households, a group of people residing within a home unit, for a block
8. medianIncome: Median income for households within a block of houses (measured in tens of thousands of US Dollars)
9. medianHouseValue: Median house value for households within a block (measured in US Dollars)
10. oceanProximity: Location of the house w.r.t ocean/sea

```
suppressMessages(library(tidyverse))
library(ggplot2)
library(viridis)
```

```
## Loading required package: viridisLite
```

```
library(GGally)
```

```
##
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:dplyr':
##
##     nasa
```

```
df <- read.table('housing.csv', header = TRUE, sep = ",")
```

We first explore the variables of the dataset by viewing the summary statistics and histogram/density plots along with a scatterplot/correlation matrix.

```
summary(df)
```

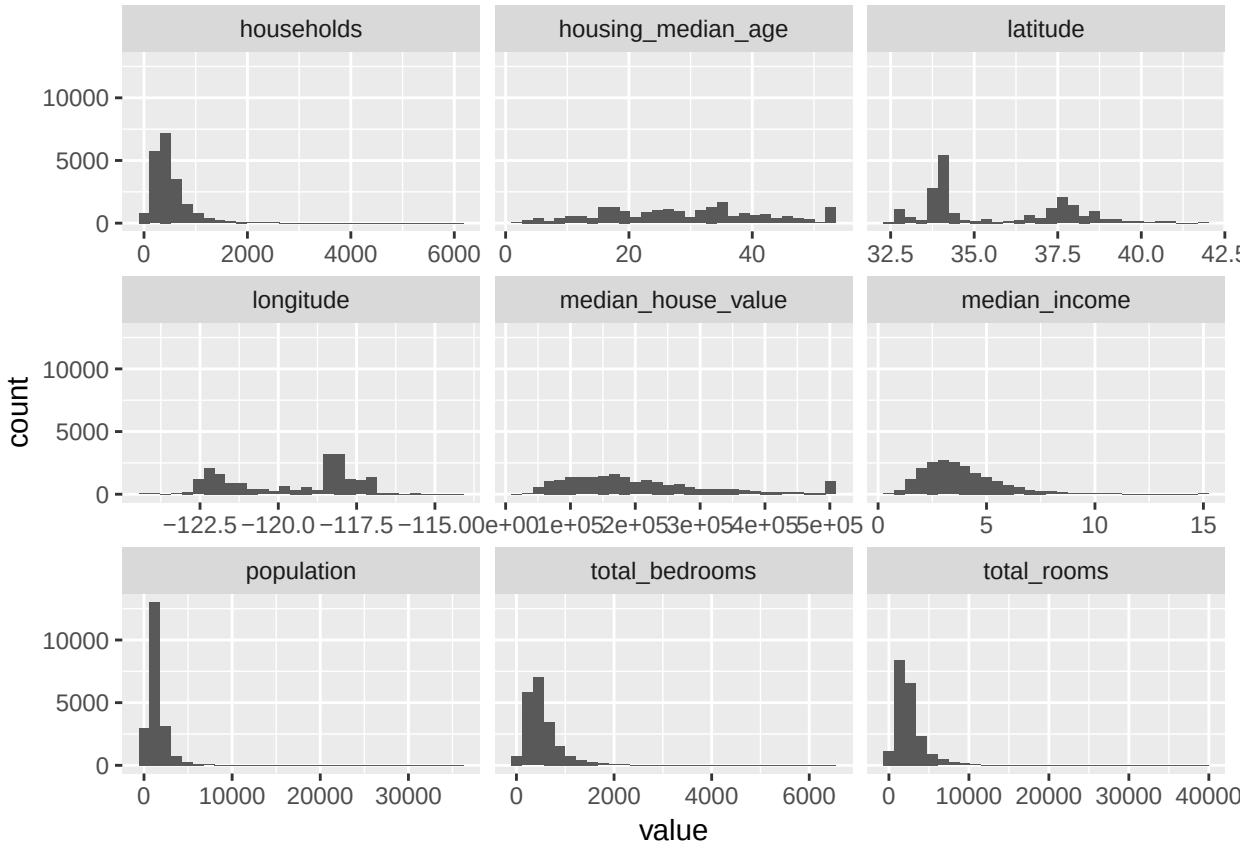
```
##    longitude         latitude     housing_median_age  total_rooms
##  Min.   :-124.3   Min.   :32.54   Min.   : 1.00      Min.   :    2
##  1st Qu.:-121.8   1st Qu.:33.93   1st Qu.:18.00      1st Qu.: 1448
##  Median :-118.5   Median :34.26   Median :29.00      Median : 2127
##  Mean   :-119.6   Mean   :35.63   Mean   :28.64      Mean   : 2636
##  3rd Qu.:-118.0   3rd Qu.:37.71   3rd Qu.:37.00      3rd Qu.: 3148
##  Max.   :-114.3   Max.   :41.95   Max.   :52.00      Max.   :39320
##
##  total_bedrooms     population      households     median_income
##  Min.   :   1.0   Min.   :    3   Min.   :   1.0   Min.   : 0.4999
##  1st Qu.: 296.0   1st Qu.:  787   1st Qu.: 280.0   1st Qu.: 2.5634
##  Median : 435.0   Median : 1166   Median : 409.0   Median : 3.5348
##  Mean   : 537.9   Mean   : 1425   Mean   : 499.5   Mean   : 3.8707
##  3rd Qu.: 647.0   3rd Qu.: 1725   3rd Qu.: 605.0   3rd Qu.: 4.7432
```

```
##  Max.   :6445.0   Max.   :35682   Max.   :6082.0   Max.   :15.0001
##  NA's   :207
##  median_house_value   ocean_proximity
##  Min.   : 14999     <1H OCEAN :9136
##  1st Qu.:119600     INLAND    :6551
##  Median :179700     ISLAND    :   5
##  Mean   :206856     NEAR BAY  :2290
##  3rd Qu.:264725     NEAR OCEAN:2658
##  Max.   :500001
##
```

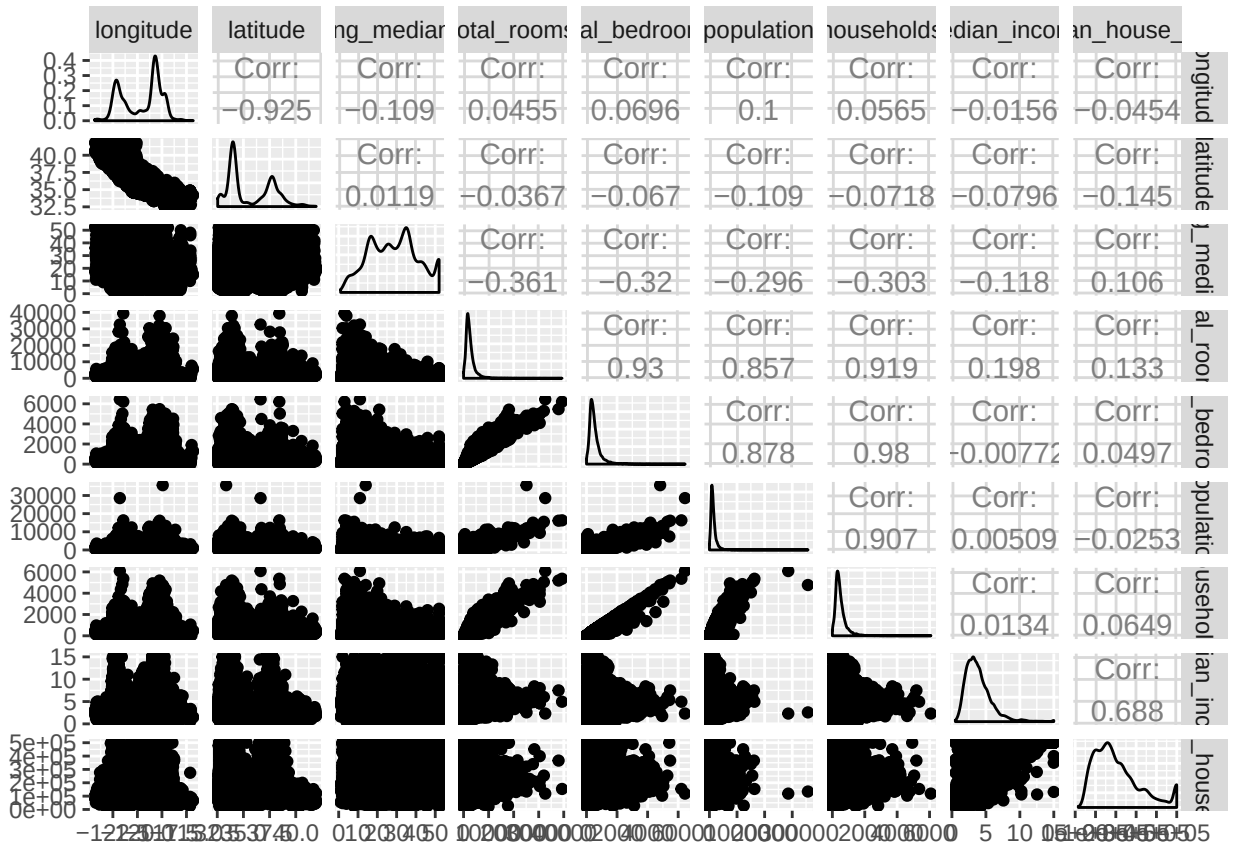
```
df_g = df %>%
  filter(complete.cases(.)) %>%
  select_if(is.numeric) %>%
  gather()

ggplot(df_g, aes(value)) +
    geom_histogram() +
    facet_wrap(~key, scales = 'free_x')
```

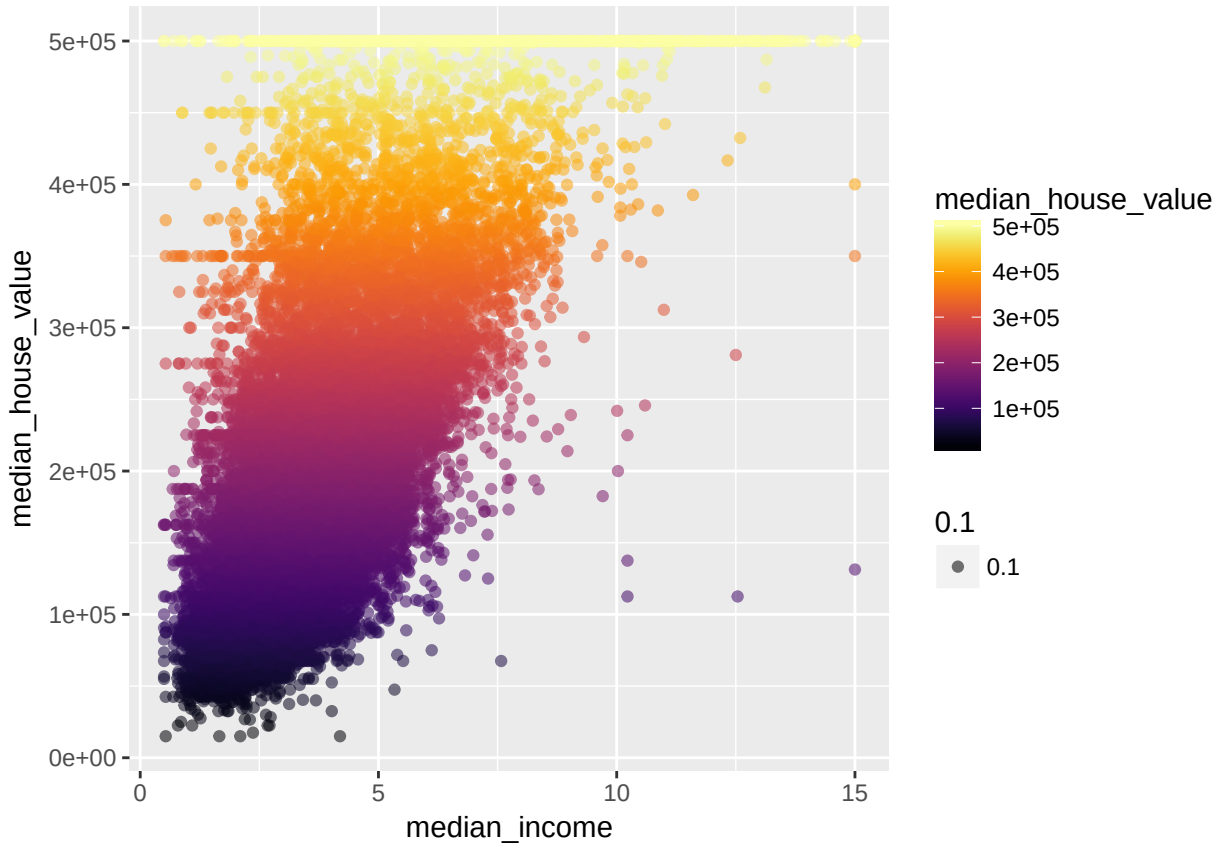
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
suppressMessages(ggpairs(dplyr::filter(df, complete.cases(df)), columns=1:9))
```

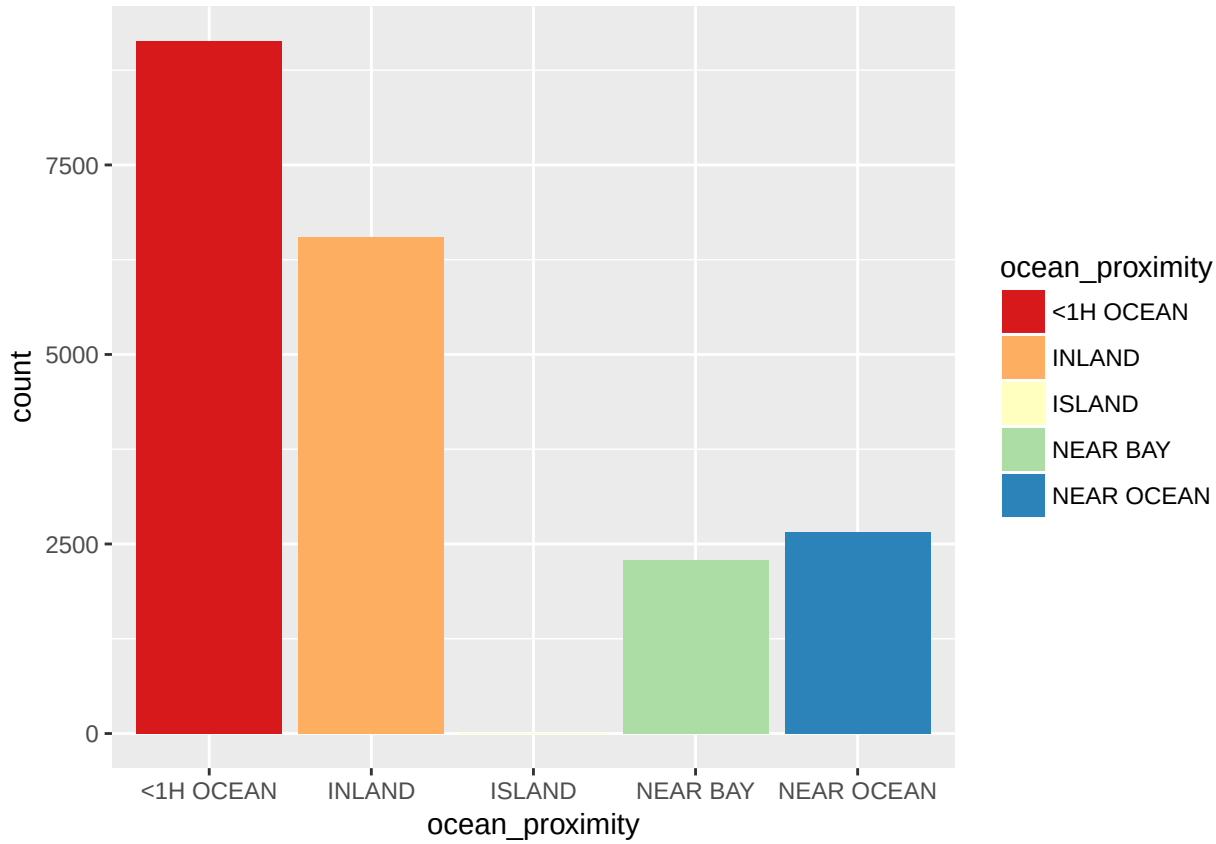



```
ggplot(data = df) +
  geom_point(mapping = aes(x = median_income, y = median_house_value,
                           color = median_house_value, alpha = 0.1)) +
  scale_colour_viridis(option = "B")
```




We take a closer look at the relationship between median income and house value, noting a positive correlation. This makes sense as higher income districts would probably have more expensive houses.

```
ggplot(data = df) +
  geom_bar(mapping = aes(x = ocean_proximity, fill = ocean_proximity)) +
  scale_fill_brewer(palette = "Spectral")
```



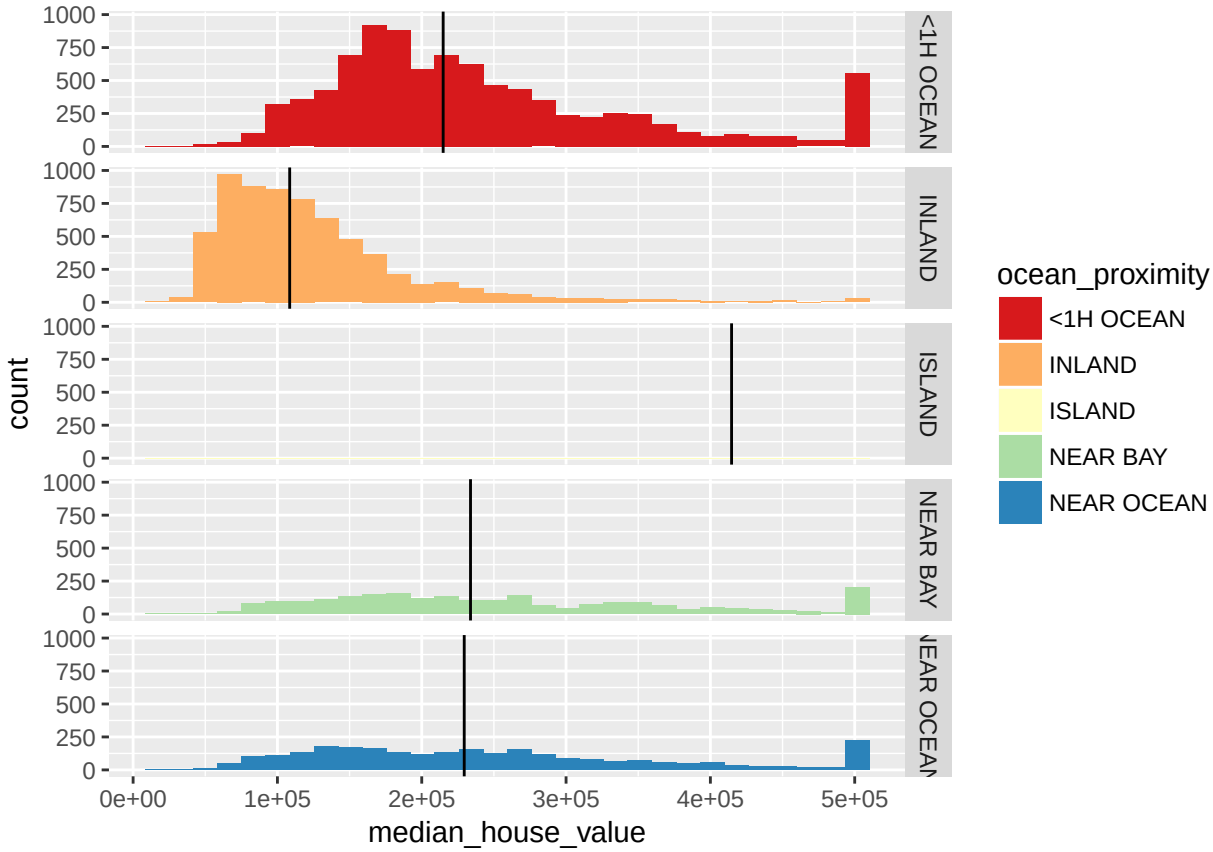

We look at the distribution of districts with respect to their proximity to the ocean. We also try to relate this to house value, hypothesizing that houses nearer the ocean may be on average more expensive.

```
median_median_house_values <- df %>%
  group_by(ocean_proximity) %>%
  summarise(value = median(median_house_value))
median_median_house_values
```

```
## # A tibble: 5 x 2
##   ocean_proximity  value
##   <fct>            <dbl>
## 1 <1H OCEAN       214850
## 2 INLAND          108500
## 3 ISLAND          414700
## 4 NEAR BAY        233800
## 5 NEAR OCEAN      229450
```

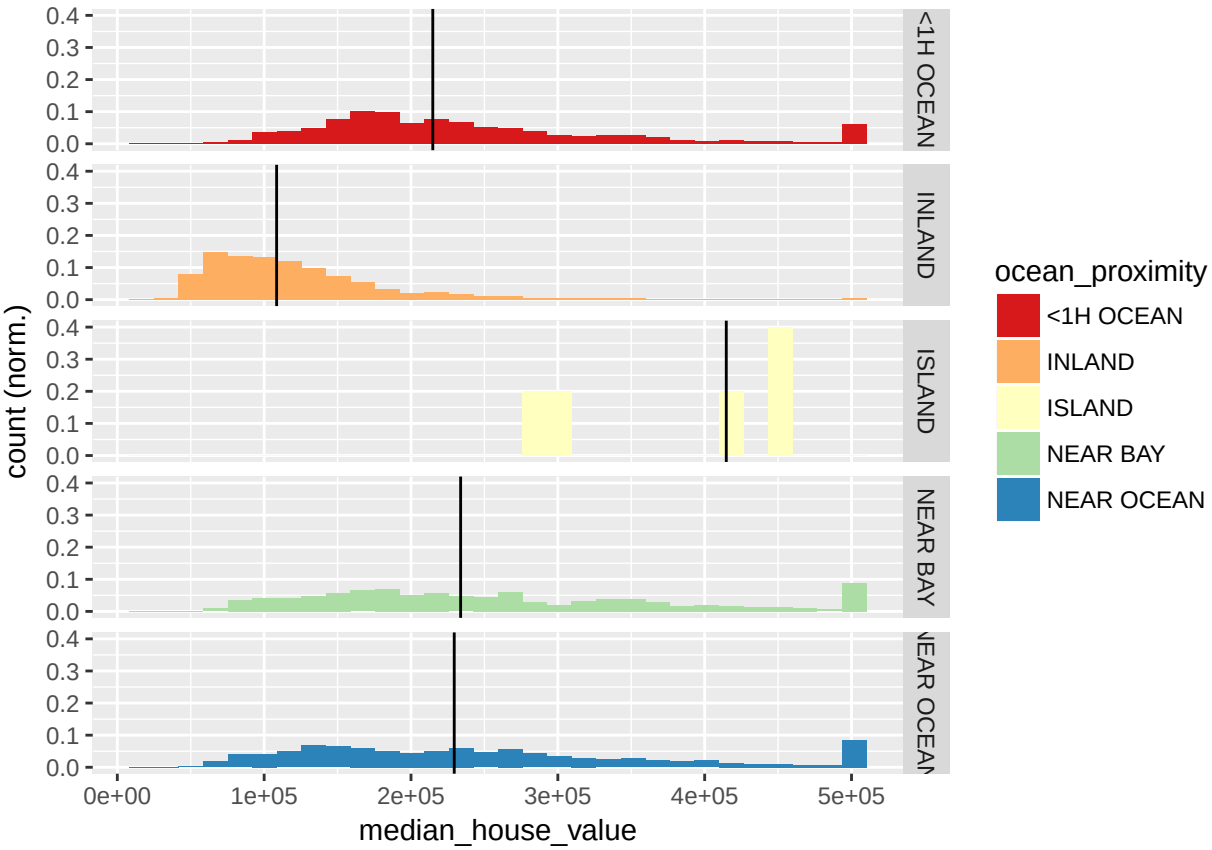
```
ggplot(data = df) +
  stat_bin(aes(median_house_value, fill = ocean_proximity)) +
  scale_fill_brewer(palette = "Spectral") +
  facet_grid(ocean_proximity ~ .) +
  geom_vline(data = median_median_house_values, aes(xintercept=value))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



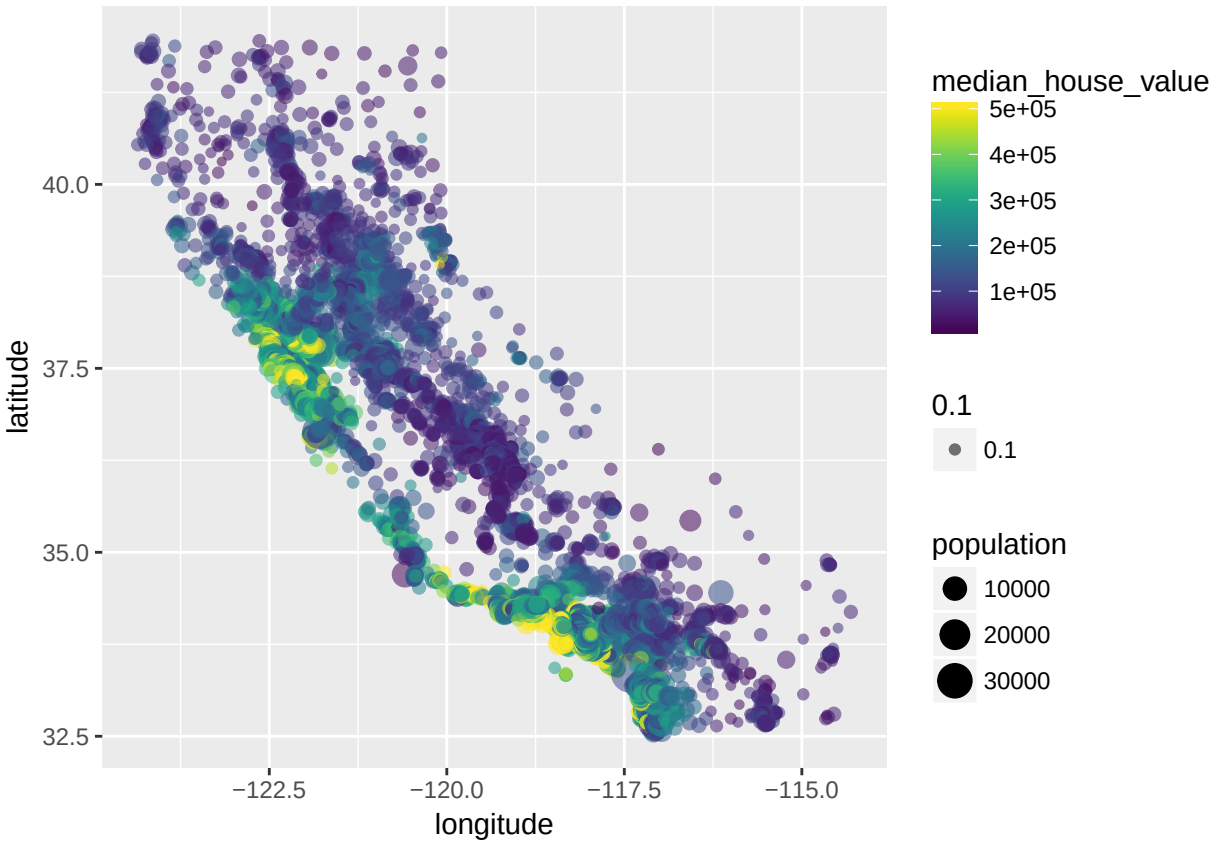

```
ggplot(data = df, aes(x = median_house_value, fill = ocean_proximity)) +
  geom_histogram(aes(y=(..count..)/tapply(..count..,..PANEL..,sum)[..PANEL..])) +
  facet_grid(ocean_proximity ~ .) +
  scale_y_continuous('count (norm.)') +
  scale_fill_brewer(palette = "Spectral") +
  geom_vline(data = median_median_house_values, aes(xintercept=value))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




Above plotted are histograms of median_house_value by each ocean_proximity category. Each subplot shows the median of median_house_value as a vertical line, and the second plot normalizes each categories' count for better comparison of the shape of the distribution. We see that in order of increasing house value are inland, <1h ocean, near ocean, near bay, and island. This confirms that in general, houses more near or more surrounded by ocean are more expensive.

```
ggplot(data = df) +
  geom_point(aes(x = longitude, y = latitude,
                 color = median_house_value, alpha = 0.1, size = population)) +
  scale_colour_viridis()
```




In this graph we plotted longitude and latitude to make a map that color codes each district based on how expensive it is: as the color gets brighter (yellower), the houses are more expensive than the darker (purple). The size of the plot point also shows the population density of the area, so the bigger the dot the more people there are. We can relate this to real life as the data we took are of California's housing: we see that the closer the houses are to the ocean, the more expensive they are. We also see that the most expensive and populous areas on the graph are around San Francisco and Los Angeles, the most popular cities in Northern and Southern California respectively.