Case Study 3: 1990s Califronia Housing Data

Danielle Angelini

2022-07-13

```
library(knitr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
housing = "https://raw.githubusercontent.com/deangelini/DS501-Case-Study-3/main/housing.csv"
housing_csv = read.csv(housing, header = TRUE)
housing_data = as_tibble(housing_csv)
```

Introduction

The purpose of this case study is to observe different housing variables to see their relationship to ocean proximity in California. I am interested in looking into this because usually people associate high class or higher income populations to living near a coast. Higher income populations can be attested to variables like median house value and median income. I will be also looking into the distribution of the population in California through looking at population within a block and total bedrooms in a block. This will be done through clustering and visualizing the location of these clusters both through the data exploration below and the Shiny Application.

Data Cleaning

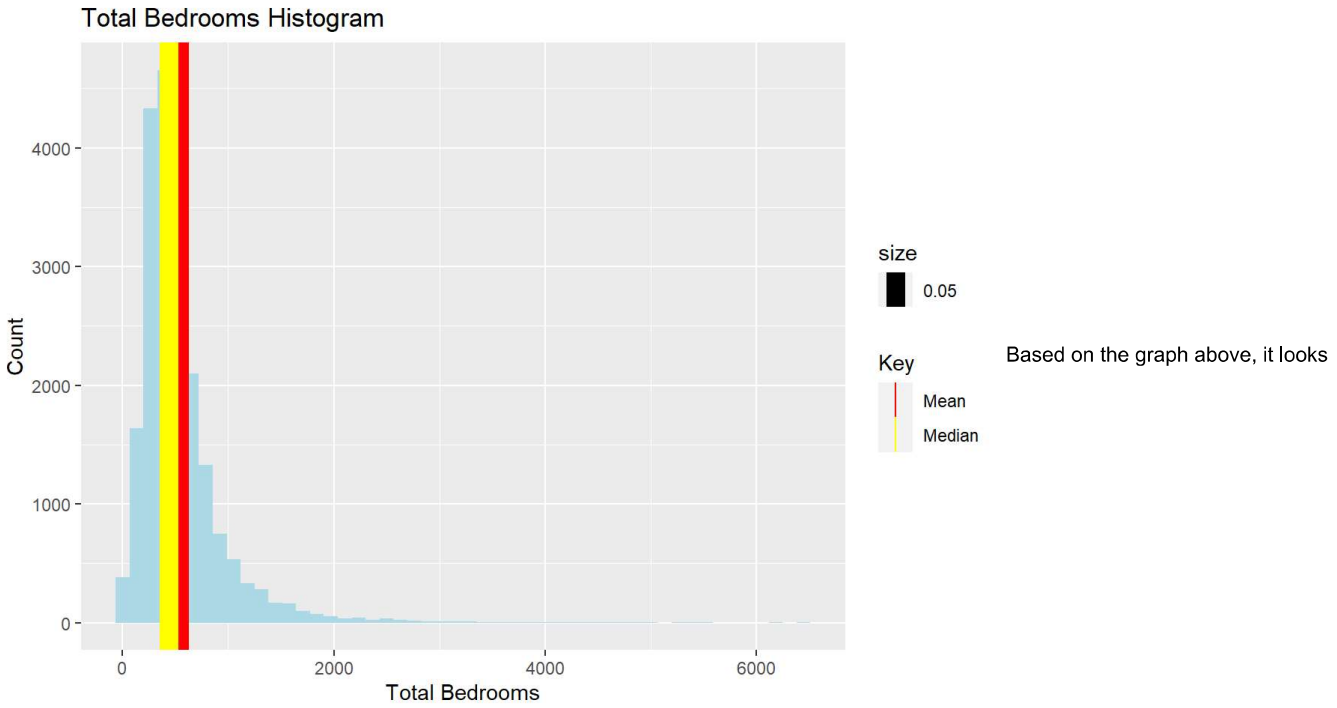
To start data cleaning, let's check for missing variables in the data with a summary.

```
housing = "C:/Users/deang/Documents/DS501/Case Study 3/housing.csv"
housing_csv = read.csv(housing, header = TRUE)
housing_data = as_tibble(housing_csv)
summary(housing_data)
```

```
##    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     Length:20640
##  1st Qu.:119600     Class :character
##  Median :179700     Mode  :character
##  Mean   :206856
##  3rd Qu.:264725
##  Max.   :500001
##
```

For the total bedrooms field, there are 207 missing values. To address these missing values, I will be looking to see whether the missing data can be replaced by the median or mean total bedrooms.

```
## Warning: Removed 207 rows containing non-finite values (stat_bin).
```



Based on the graph above, it looks like the median is a better representation of the total bedroom data. I will be replacing the missing 207 data points with the median total bedrooms.

```
housing_clean = housing_data
housing_clean$total_bedrooms[is.na(housing_clean$total_bedrooms)] = br_median
summary(housing_clean)
```

```
##    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.: 297.0   1st Qu.:  787   1st Qu.: 280.0   1st Qu.: 2.5634
##  Median : 435.0   Median : 1166   Median : 409.0   Median : 3.5348
##  Mean   : 536.8   Mean   : 1425   Mean   : 499.5   Mean   : 3.8707
##  3rd Qu.: 643.2   3rd Qu.: 1725   3rd Qu.: 605.0   3rd Qu.: 4.7432
##  Max.   :6445.0   Max.   :35682   Max.   :6082.0   Max.   :15.0001
##  median_house_value ocean_proximity
##  Min.   : 14999     Length:20640
##  1st Qu.:119600     Class :character
##  Median :179700     Mode  :character
##  Mean   :206856
##  3rd Qu.:264725
##  Max.   :500001
```

Now that there are no more missing data points, the data modeling can proceed through using k-means clustering. This part of the data science life cycle is done through Shiny App.

Data Exploration

With the California housing dataset, there are many options of variables to explore. To help focus analysis on a selection of these variables, let's look to see if any variables are highly correlated with one another.

```
library(kableExtra)
```

```
## Warning in !is.null(rmarkdown::metadata$output) && rmarkdown::metadata$output
## %in% : 'length(x) = 2 > 1' in coercion to 'logical(1)'
```

```
##
## Attaching package: 'kableExtra'
```

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

```
housing_clean_nc = housing_clean[, -10]
corrmatrix = round(cor(housing_clean_nc),3)
kable(t(corrmatrix))
```

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_va
longitude	1.000	-0.925	-0.108	0.045	0.069	0.100	0.055	-0.015	-0
latitude	-0.925	1.000	0.011	-0.036	-0.066	-0.109	-0.071	-0.080	-0
housing_median_age	-0.108	0.011	1.000	-0.361	-0.319	-0.296	-0.303	-0.119	0
total_rooms	0.045	-0.036	-0.361	1.000	0.927	0.857	0.918	0.198	0
total_bedrooms	0.069	-0.066	-0.319	0.927	1.000	0.874	0.974	-0.008	0
population	0.100	-0.109	-0.296	0.857	0.874	1.000	0.907	0.005	-0
households	0.055	-0.071	-0.303	0.918	0.974	0.907	1.000	0.013	0
median_income	-0.015	-0.080	-0.119	0.198	-0.008	0.005	0.013	1.000	0
median_house_value	-0.046	-0.144	0.106	0.134	0.049	-0.025	0.066	0.688	1

Based on the correlation matrix, it looks like total bedrooms and households are highly correlated, total rooms and households are highly correlated, and total bedrooms and total rooms are highly correlated. With that being said, the variables I will be analyzing with location (latitude and longitude) are:

- Total Bedrooms
- Population
- Median Income
- Median House Value

With an interactive k-means clustering model, let's look at the most fitting number of clusters for each field of the California housing data. This can be done by finding the SSE, which is defined as the sum of the squared distance between each member of a cluster and its cluster center.

Median House Value:

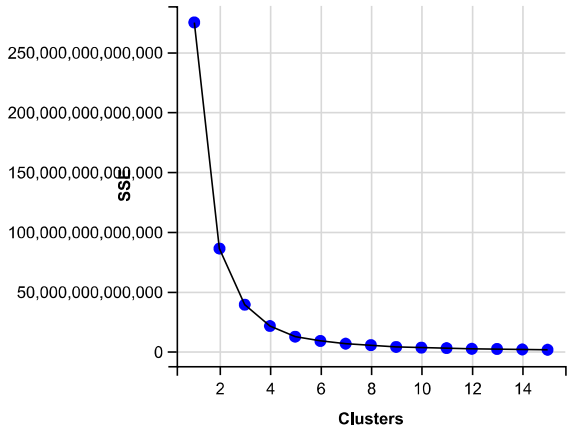
```
library(ggvis)
```

```
## Warning: package 'ggvis' was built under R version 4.2.1
```

```
##
## Attaching package: 'ggvis'
```

```
## The following object is masked from 'package:ggplot2':
##
##     resolution
```

```
#Median House Value
m_h_value = subset(housing_clean, select = c(longitude, latitude, median_house_value))
wss_1 = kmeans(m_h_value, centers=1)$tot.withinss
for (i in 2:15)
  wss_1[i] = kmeans(m_h_value, centers=i)$tot.withinss
library(ggvis)

sse_1 = data.frame(c(1:15), c(wss_1))
names(sse_1)[1] = 'Clusters'
names(sse_1)[2] = 'SSE'
sse_1 %>%
  ggvis(~Clusters, ~SSE) %>%
  layer_points(fill := 'blue') %>%
  layer_lines() %>%
  set_options(height = 300, width = 400)
```



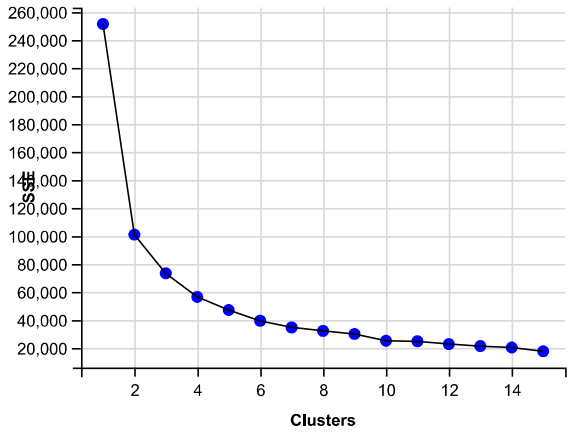
Median Income:

```
library(ggvis)

#Median Income
m_i_value = subset(housing_clean, select = c(longitude, latitude, median_income))
wss_2 = kmeans(m_i_value, centers=1)$tot.withinss
for (i in 2:15)
  wss_2[i] = kmeans(m_i_value, centers=i)$tot.withinss
```

```
## Warning: Quick-TRANSfer stage steps exceeded maximum (= 1032000)
```

```
## Warning: did not converge in 10 iterations
```

```
library(ggvis)

sse_2 = data.frame(c(1:15), c(wss_2))
names(sse_2)[1] = 'Clusters'
names(sse_2)[2] = 'SSE'
sse_2 %>%
  ggvis(~Clusters, ~SSE) %>%
  layer_points(fill := 'blue') %>%
  layer_lines() %>%
  set_options(height = 300, width = 400)
```

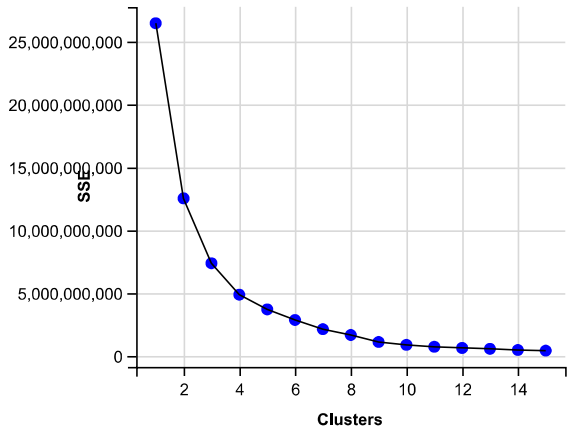


Population:

```
library(ggvis)

#Population
p_value = subset(housing_clean, select = c(longitude, latitude, population))
wss_3 = kmeans(p_value, centers=1)$tot.withinss
for (i in 2:15)
  wss_3[i] = kmeans(p_value, centers=i)$tot.withinss
```

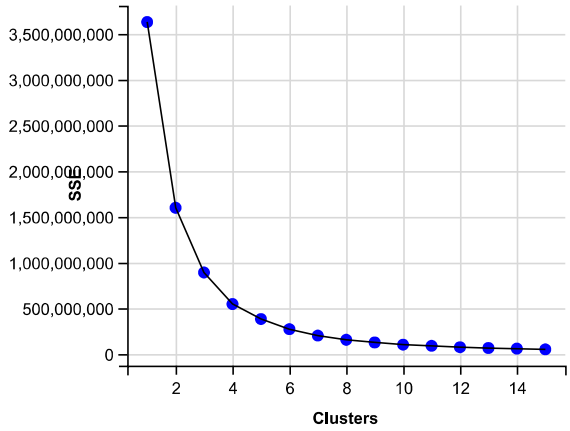
```
## Warning: Quick-TRANSfer stage steps exceeded maximum (= 1032000)

## Warning: Quick-TRANSfer stage steps exceeded maximum (= 1032000)
```

```
library(ggvis)

sse_3 = data.frame(c(1:15), c(wss_3))
names(sse_3)[1] = 'Clusters'
names(sse_3)[2] = 'SSE'
sse_3 %>%
  ggvis(~Clusters, ~SSE) %>%
  layer_points(fill := 'blue') %>%
  layer_lines() %>%
  set_options(height = 300, width = 400)
```



Total Bedrooms:

```
library(ggvis)

#Total Bedrooms
b_value = subset(housing_clean, select = c(longitude, latitude, total_bedrooms))
wss_4 = kmeans(b_value, centers=1)$tot.withinss
for (i in 2:15)
  wss_4[i] = kmeans(b_value, centers=i)$tot.withinss
library(ggvis)

sse_4 = data.frame(c(1:15), c(wss_4))
names(sse_4)[1] = 'Clusters'
names(sse_4)[2] = 'SSE'
sse_4 %>%
  ggvis(~Clusters, ~SSE) %>%
  layer_points(fill := 'blue') %>%
  layer_lines() %>%
  set_options(height = 300, width = 400)
```

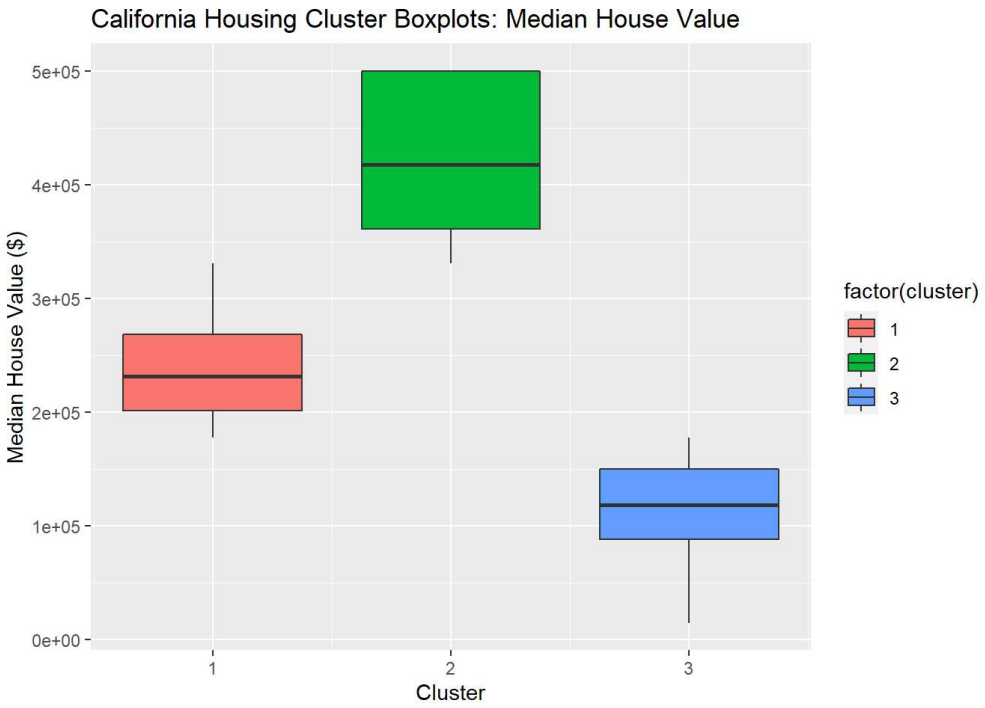


Based on all these graphs, it looks like that there is little change in error in using a higher number of clusters for the following variables and cluster amounts:

Median House Value: 3 Median Income: 4 Population: 4 Total Bedrooms: 4

With this given information, we can specifically look at the clusters formed for each variable and analyze the distribution of the data using box plots. Box plots will help to visualize the clustered data and to look for signs of skewness, to identify mean values, and to observe the dispersion of the data set.

```
library(ggplot2)

#Median House Value
k_clust = kmeans(m_h_value$median_house_value, 3)
m_h_value$cluster = k_clust$cluster
housing_value_clust = as.data.frame(m_h_value)

ggplot(data = housing_value_clust, mapping = aes(x = cluster, y = median_house_value)) + geom_boxplot(aes(fill = factor(clus
ter))) + xlab("Cluster") + ylab("Median House Value ($)") +
  ggtitle("California Housing Cluster Boxplots: Median House Value")
```



Looking at the median house values of each cluster, it seems like 1990s California median housing prices can be clustered into median house prices with averages around 125,000, 230,000, and 410,000. The third cluster looks like it has the largest range of house prices, 350,000 to 500,000. Based on the Shiny Application, the cluster of houses with the highest median value tend to be by the coast of California and not as inland. This can be verified by the following table:

```
library(dplyr)
library(kableExtra)
location_table = housing_clean%>%select(longitude, latitude, ocean_proximity)

m_h_cluster_location = merge(housing_value_clust, location_table, by.x = c("longitude", "latitude"), by.y = c("longitude",
"latitude"))%>%select(cluster, ocean_proximity)%>%group_by(cluster, ocean_proximity)%>%summarise(count = n())
```

```
## `summarise()` has grouped output by 'cluster'. You can override using the
## `.groups` argument.
```

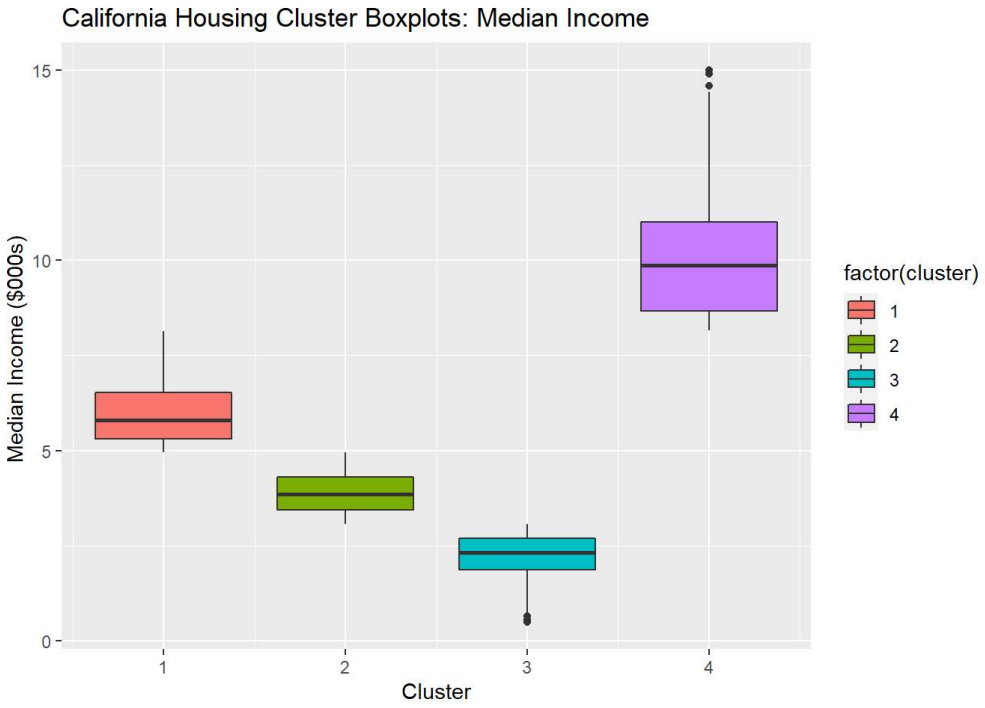
```
kable(m_h_cluster_location, caption = "California Housing Price Location by Cluster")
```

California Housing Price
Location by Cluster

clusterocean_proximity	count
1<1H OCEAN	11622
1INLAND	1220
1ISLAND	2
1NEAR BAY	3321
1NEAR OCEAN	2826
2<1H OCEAN	4238
2INLAND	220
2ISLAND	3
2NEAR BAY	2754
2NEAR OCEAN	1226
3<1H OCEAN	9452
3INLAND	8383
3NEAR BAY	2533
3NEAR OCEAN	2666

Next, let's analyze median income with box-plots and a location table of each cluster.

```
library(ggplot2)
library(kableExtra)

#Median Income
k_clust = kmeans(m_i_value$median_income, 4)
m_i_value$cluster = k_clust$cluster
income_clust = as.data.frame(m_i_value)

ggplot(data = income_clust, mapping = aes(x = cluster, y = median_income)) + geom_boxplot(aes(fill = factor(cluster))) + xla
b("Cluster") + ylab("Median Income ($000s) ") +
  ggtitle("California Housing Cluster Boxplots: Median Income")
```




```
m_i_cluster_location = merge(income_clust, location_table, by.x = c("longitude", "latitude"), by.y = c("longitude", "latitud
e"))%>%select(cluster, ocean_proximity)%>%group_by(cluster, ocean_proximity)%>%summarise(count = n())
```

```
## `summarise()` has grouped output by 'cluster'. You can override using the
## `.groups` argument.
```

```
kable(m_i_cluster_location, caption = "California Median Income Location by Cluster")
```

California Median Income
Location by Cluster

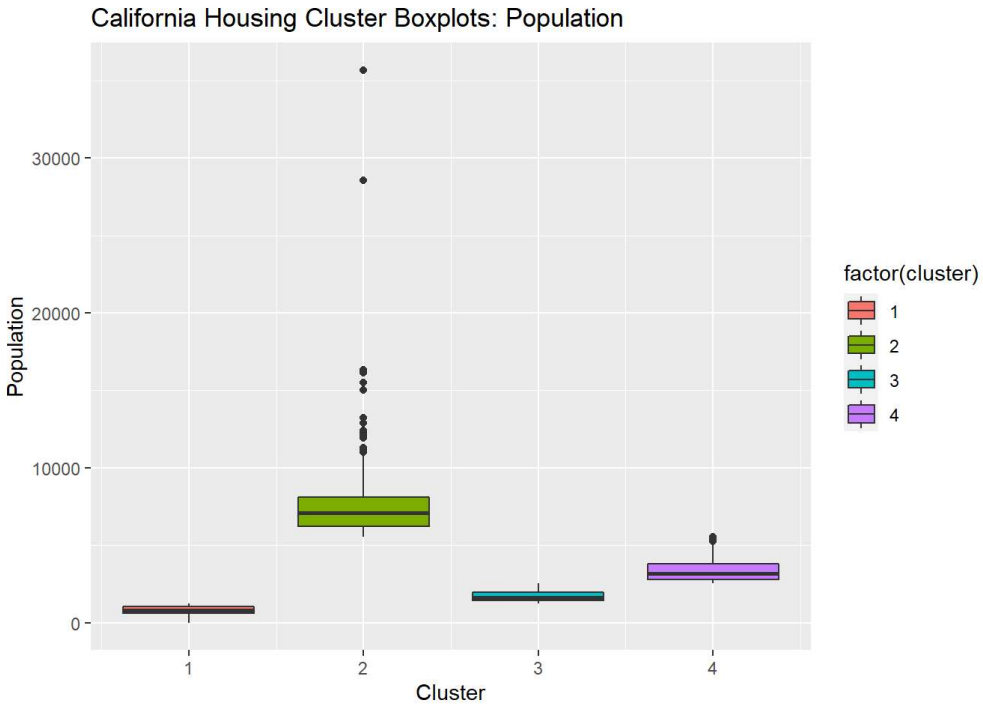
clusterocean_proximity	count
1<1H OCEAN	4853
1INLAND	958
1NEAR BAY	1334
1NEAR OCEAN	927
2<1H OCEAN	10048
2INLAND	3477
2ISLAND	1
2NEAR BAY	3788
2NEAR OCEAN	2714
3<1H OCEAN	9763
3INLAND	5317
3ISLAND	4
3NEAR BAY	3277
3NEAR OCEAN	2911
4<1H OCEAN	648
4INLAND	71

cluster	ocean_proximity	count
4	NEAR BAY	209
4	NEAR OCEAN	166

The fourth cluster with the highest average median income tends to be the least inland, similar to the cluster with the highest average median housing price. The first cluster with the lowest median income tends to be the most inland with 5,317.

Let's look at population cluster data in California.

```
library(ggplot2)

#Population
k_clust = kmeans(p_value$population, 4)
p_value$cluster = k_clust$cluster
pop_clust = as.data.frame(p_value)

ggplot(data = pop_clust, mapping = aes(x = cluster, y = population)) + geom_boxplot(aes(fill = factor(cluster))) + xlab("Clu
ster") + ylab("Population") +
  ggtitle("California Housing Cluster Boxplots: Population")
```



Between population clusters, there is not too much variation in the averages of each cluster. However, the first cluster has the most outliers with populations reaching over 15,000 within a block. Let's see if this is ocean proximity driven similarly with the previous variables.

```
library(kableExtra)

p_cluster_location = merge(pop_clust, location_table, by.x = c("longitude", "latitude"), by.y = c("longitude", "latitude"))%
>%select(cluster, ocean_proximity)%>%group_by(cluster, ocean_proximity)%>%summarise(count = n())
```

```
## `summarise()` has grouped output by 'cluster'. You can override using the
## `.groups` argument.
```

```
kable(p_cluster_location, caption = "California Population within a Block: Location by Cluster")
```

California Population within a Block: Location by Cluster

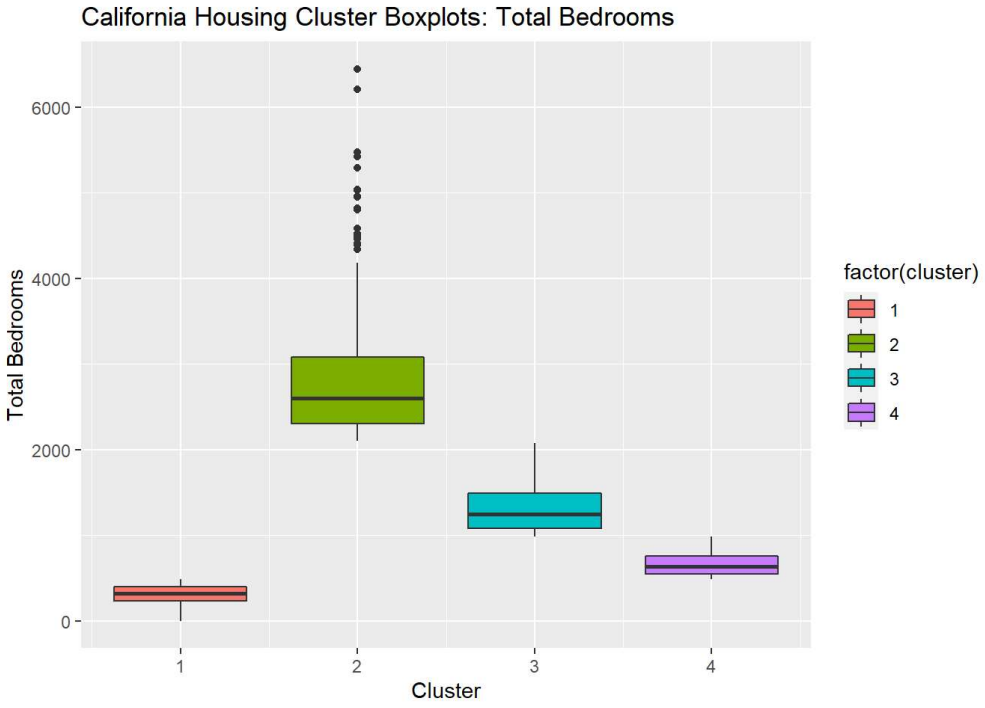
cluster	ocean_proximity	count
1	<1H OCEAN	13455
1	INLAND	5841
1	ISLAND	5
1	NEAR BAY	5680
1	NEAR OCEAN	4015
2	<1H OCEAN	141

clusterocean_proximity	count
2INLAND	109
2NEAR BAY	18
2NEAR OCEAN	24
3<1H OCEAN	9439
3INLAND	3170
3NEAR BAY	2622
3NEAR OCEAN	2291
4<1H OCEAN	2277
4INLAND	703
4NEAR BAY	288
4NEAR OCEAN	388

It looks like the first cluster has the least amount of data, so it is hard to come to a defined conclusion specifically for this cluster. However among all the clusters, it looks like most of the population tends to be near the ocean. Given the geography of California and the location of cities, this makes sense as cities tend to be more populous.

Finally, let's see if total bedrooms within a block gives more variation than total population within a block. This will be done by looking at both box-plots and ocean-proximity.

```
library(ggplot2)

#Total Bedrooms
k_clust = kmeans(b_value$total_bedrooms, 4)
b_value$cluster = k_clust$cluster
br_clust = as.data.frame(b_value)

ggplot(data = br_clust, mapping = aes(x = cluster, y = total_bedrooms)) + geom_boxplot(aes(fill = factor(cluster))) + xlab(
"Cluster") + ylab("Total Bedrooms") +
  ggtitle("California Housing Cluster Boxplots: Total Bedrooms")
```




```
br_cluster_location = merge(br_clust, location_table, by.x = c("longitude", "latitude"), by.y = c("longitude", "latitude"))%
>%select(cluster, ocean_proximity)%>%group_by(cluster, ocean_proximity)%>%summarise(count = n())
```

```
## `summarise()` has grouped output by 'cluster'. You can override using the
## `.groups` argument.
```

```
kable(br_cluster_location, caption = "California Total Bedrooms within a Block: Location by Cluster")
```

California Total Bedrooms
within a Block: Location by

Cluster

clusterocean_proximity	count
1<1H OCEAN	15511
1INLAND	6132
1ISLAND	2
1NEAR BAY	5083
1NEAR OCEAN	3771
2<1H OCEAN	175
2INLAND	107
2NEAR BAY	34
2NEAR OCEAN	22
3<1H OCEAN	1805
3INLAND	650
3NEAR BAY	568
3NEAR OCEAN	419
4<1H OCEAN	7821
4INLAND	2934
4ISLAND	3
4NEAR BAY	2923
4NEAR OCEAN	2506

similarly to population, the cluster with the highest total number of bedrooms does not have as much data as the other clusters. This makes sense as there are usually not many houses with more bedrooms than needed by the average sized family. The cluster with the least amount of total bedrooms in a block has the most houses inland than the other clusters. Considering what was found through looking at median income and median house value, this makes sense that lower median income and lower median house value areas also would have less bedrooms in the same area.

Conclusion

Overall, it looks like that higher income and higher median house values tend to be near the coast of California. As for the population distribution in California, this also tends to be more towards the coast. This could be attributed to the opportunity arising in the cities of California and their proximity to resources.