Educ 260B - Problem Set 2

Paula Nazario

January 20, 2022

About the data

According to Pace & Barry (1997), the California Housing Prices data accounts for the median house prices for California from the 1990 census. In this data set, the column names are as follows: longitude; latitude; housingmedianage; total_rooms; total_bedrooms; population; households; median_income; medianhouse-value; and ocean_proximity. According to the California Housing Prices data, the average housing media age is 28.6. In terms of median income, the average was 3.87K, with a standard deviation of 1.9K. Furthermore, in terms of median house value, the average value seems to be 207K, with a standard deviation of 115K (Pace & Barry, 1997).

```
## Part III, Question 6
list.files(data_dir)
```

```
## [1] "housing.RDS"
```

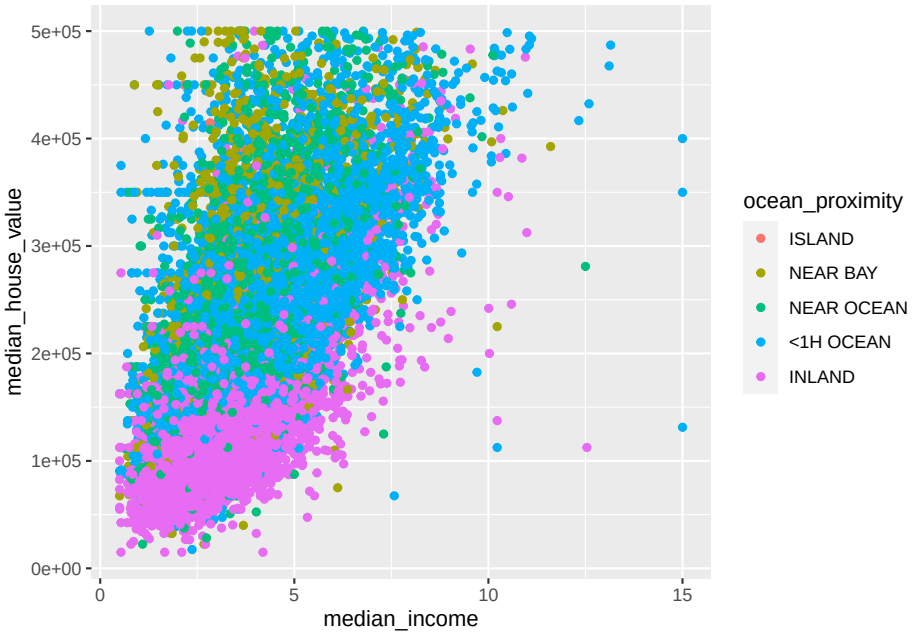
```
housing_df <- readRDS(file = file.path(data_dir, 'housing.RDS'))
housing_df %>%
      head(10) %>%
      kable() %>%
      kable_styling(latex_options = c('scale_down', 'HOLD_position'))
```

longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
-122.23	37.88	41	880	129	322	126	8.3252	452600	NEAR BAY
-122.22	37.86	21	7099	1106	2401	1138	8.3014	358500	NEAR BAY
-122.24	37.85	52	1467	190	496	177	7.2574	352100	NEAR BAY
-122.25	37.85	52	1274	235	558	219	5.6431	341300	NEAR BAY
-122.25	37.85	52	1627	280	565	259	3.8462	342200	NEAR BAY
-122.25	37.85	52	919	213	413	193	4.0368	269700	NEAR BAY
-122.25	37.84	52	2535	489	1094	514	3.6591	299200	NEAR BAY
-122.25	37.84	52	3104	687	1157	647	3.1200	241400	NEAR BAY
-122.26	37.84	42	2555	665	1206	595	2.0804	226700	NEAR BAY
-122.25	37.84	52	3549	707	1551	714	3.6912	261100	NEAR BAY

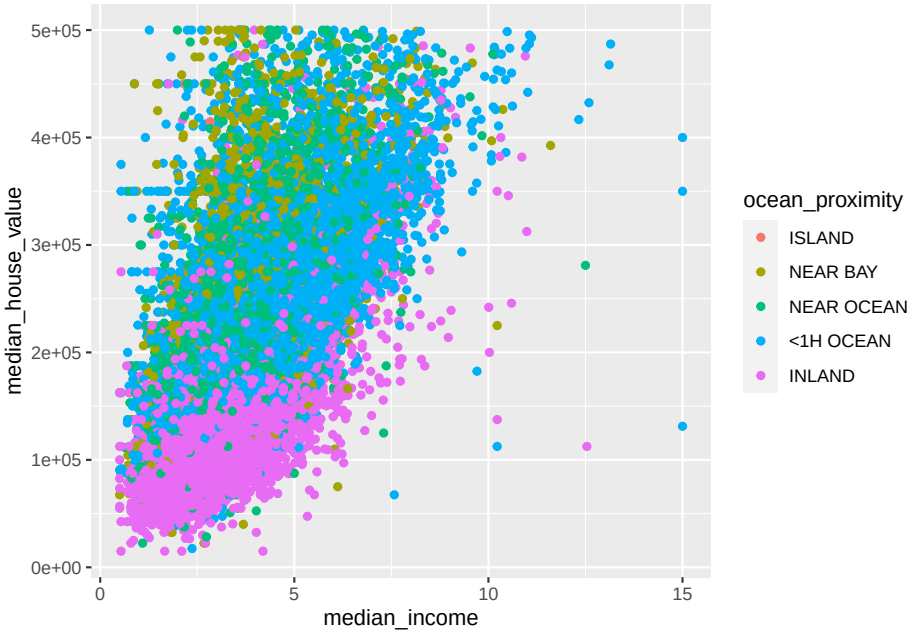
Data analysis

```
## Part IV, Question 1

## Investigations
## housing_df$ocean_proximity
## housing_df$ocean_proximity %>% str()
```

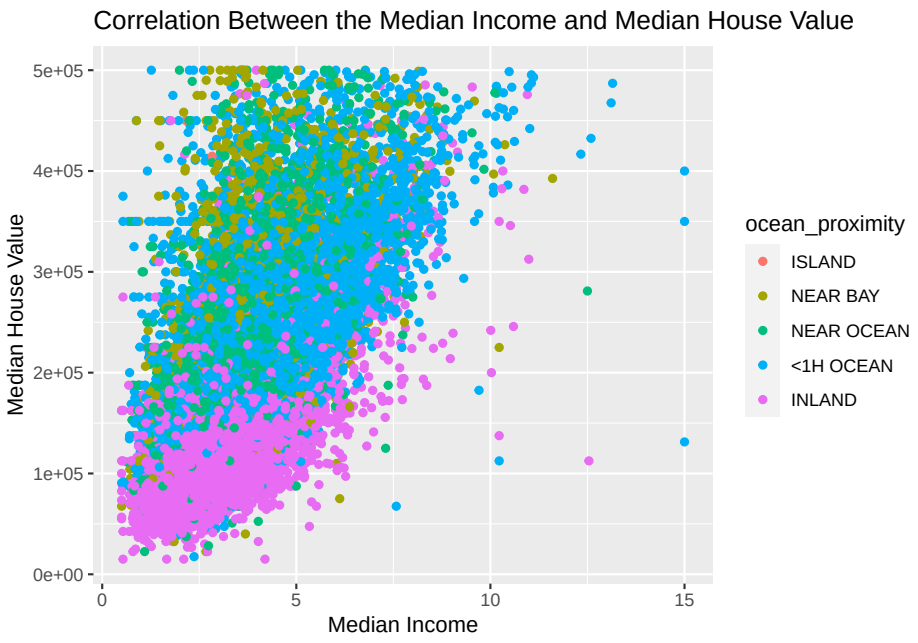
```
## housing_df$ocean_proximity %>% typeof()
## housing_df$ocean_proximity %>% class()
## housing_df %>% select(ocean_proximity) %>% count()
## housing_df$ocean_proximity %>% summary()

## Plots from Part II, Question 5
ggplot(data = housing_df, mapping = aes(x = median_income, y = median_house_value, color = ocean_proximi
```




```
ggplot(data = housing_df, aes(x = median_income, y = median_house_value, color = ocean_proximity)) + ge
```

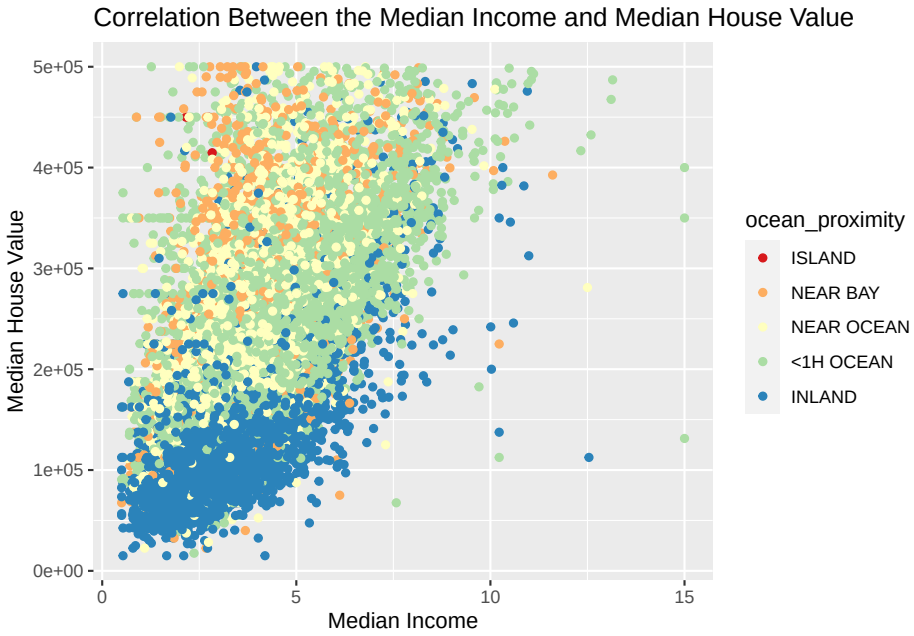


```
## Title and Labels
ggplot(data = housing_df, mapping = aes(x = median_income, y = median_house_value, color = ocean_proximi
```

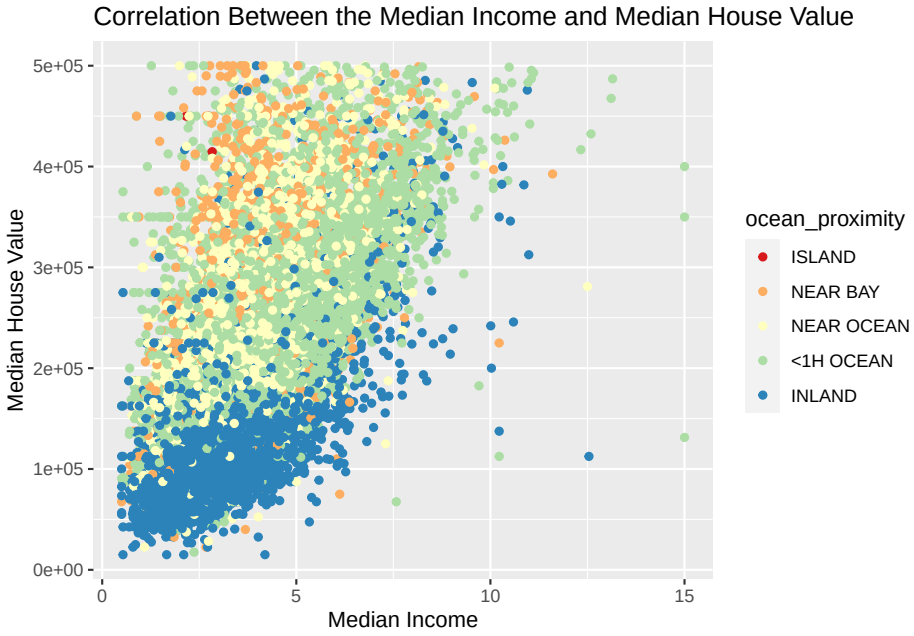



```
ggplot(data = housing_df, aes(x = median_income, y = median_house_value, color = ocean_proximity)) + geo
```

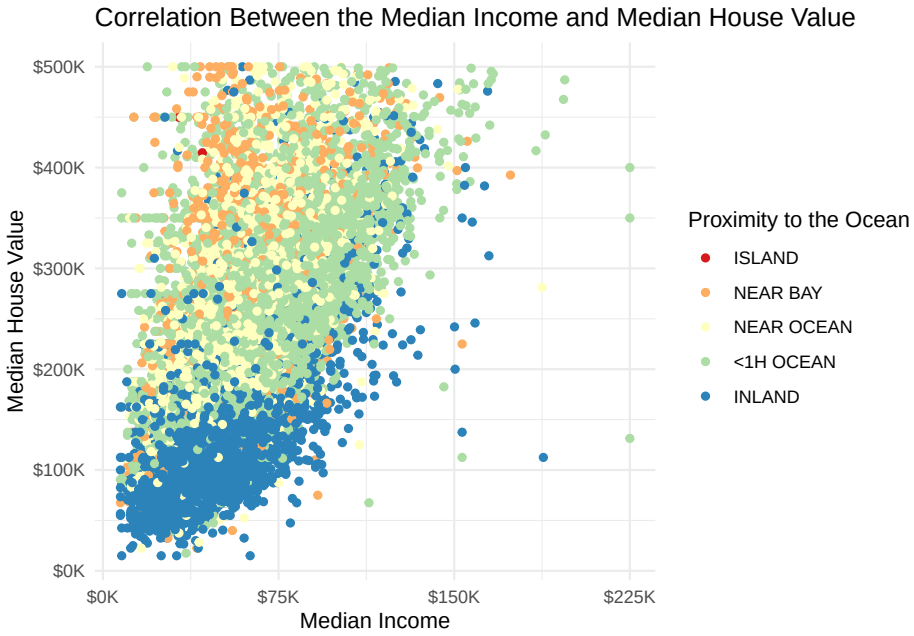

```
## Palettes
ggplot(data = housing_df, mapping = aes(x = median_income, y = median_house_value, color = ocean_proximi
```



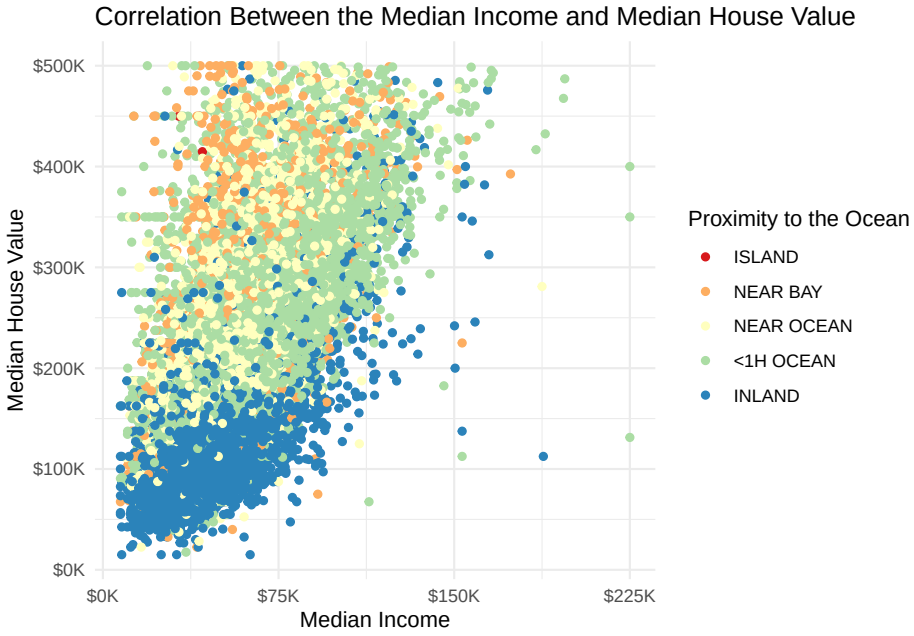

```
ggplot(data = housing_df, aes(x = median_income, y = median_house_value, color = ocean_proximity)) + ge
```




```
## Legend Title
ggplot(data = housing_df, mapping = aes(x = median_income, y = median_house_value, color = ocean_proxim
```

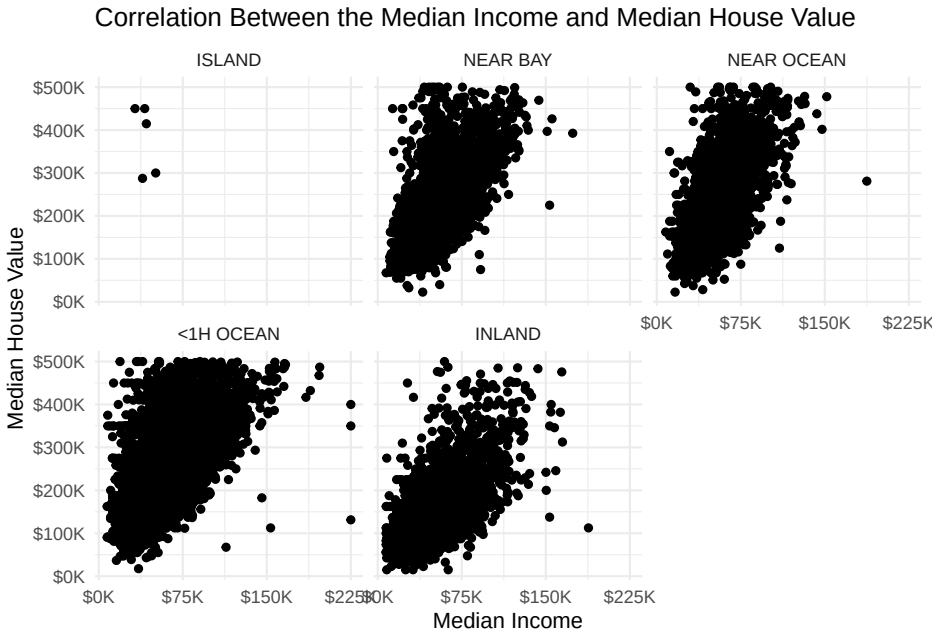



```
ggplot(data = housing_df, aes(x = median_income, y = median_house_value, color = ocean_proximity)) + ge
```

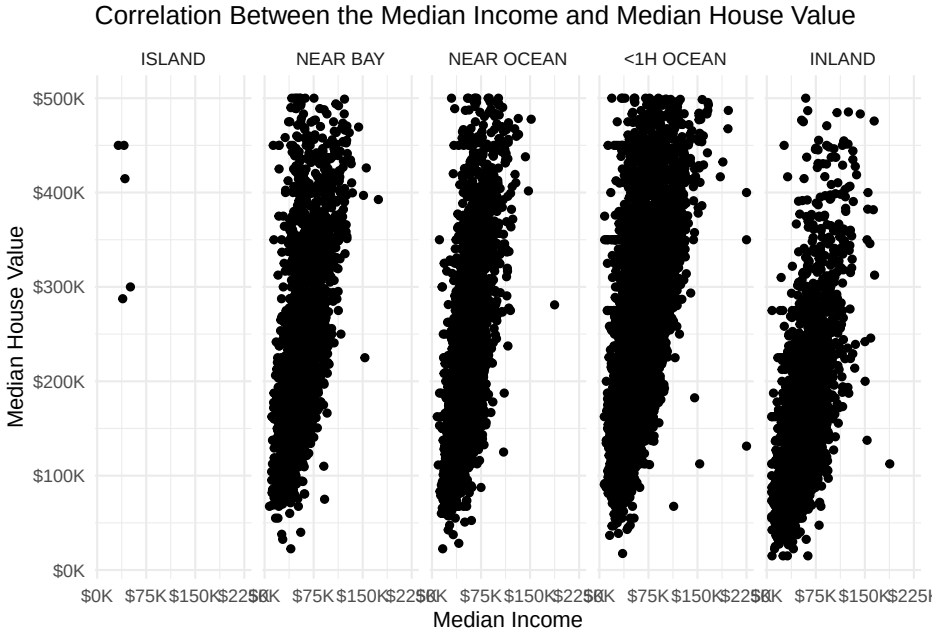



```
## Based on this plot, we can assume that there is a strong upward correlation between the median incom
```

```
## Part IV, Question 2

## facet_wrap()
ggplot(data = housing_df, mapping = aes(x = median_income, y = median_house_value)) + geom_point() + fa
```

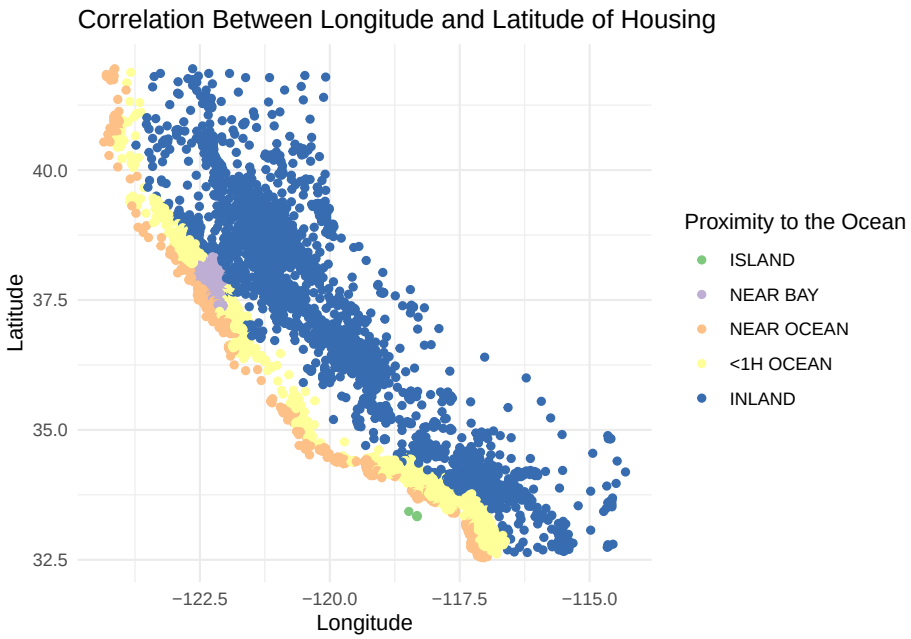



```
## facet_grid()
ggplot(data = housing_df, aes(x = median_income, y = median_house_value)) + geom_point() + facet_grid(~
```

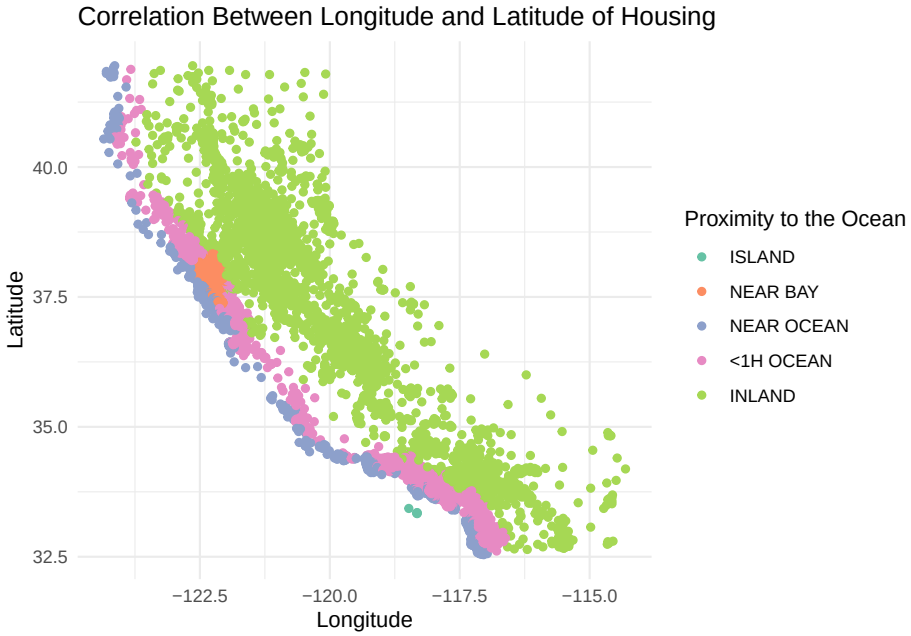



```
## According to these facet_wrap and facet_grid plots, it looks like there are strong correlations betw
```

```
## Part IV, Question 3

## coord_fixed()
ggplot(data = housing_df, mapping = aes(x = longitude, y = latitude, color = ocean_proximity)) + geom_p
```

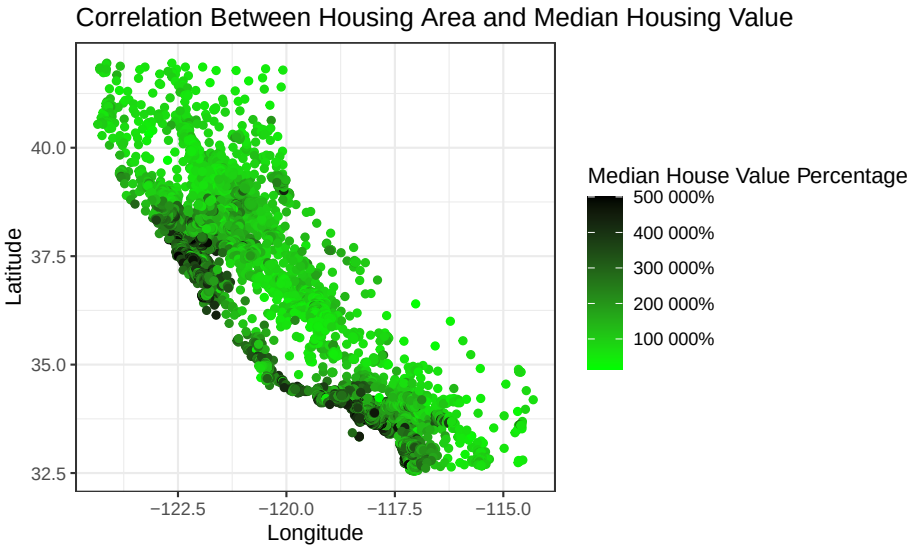



```
ggplot(data = housing_df, aes(x = longitude, y = latitude, color = ocean_proximity)) + geom_point() + gg
```

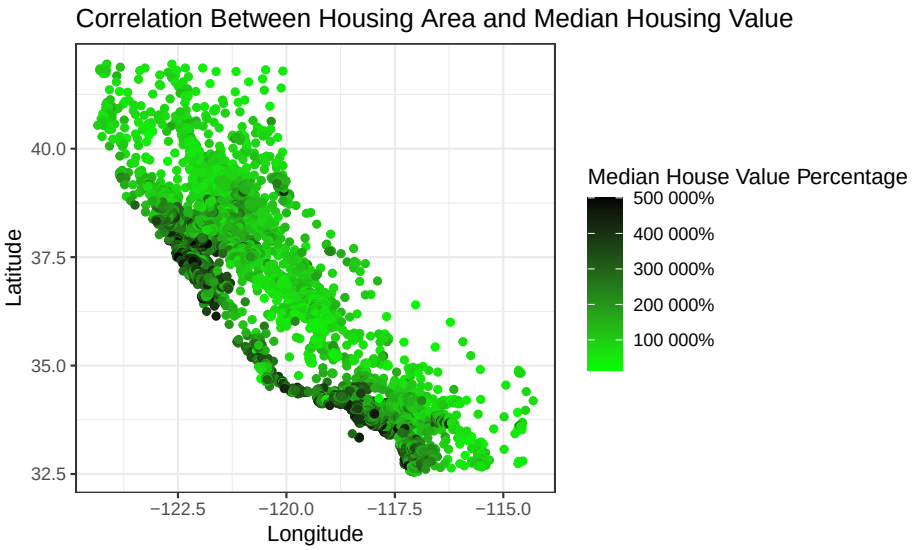



```
## In this plot, there is a clear distinction between location of housing units in the state of Califor
```

```
## Part IV, Question 4

## Gradients
ggplot(data = housing_df, mapping = aes(x = longitude, y = latitude, color = median_house_value)) + geo
```

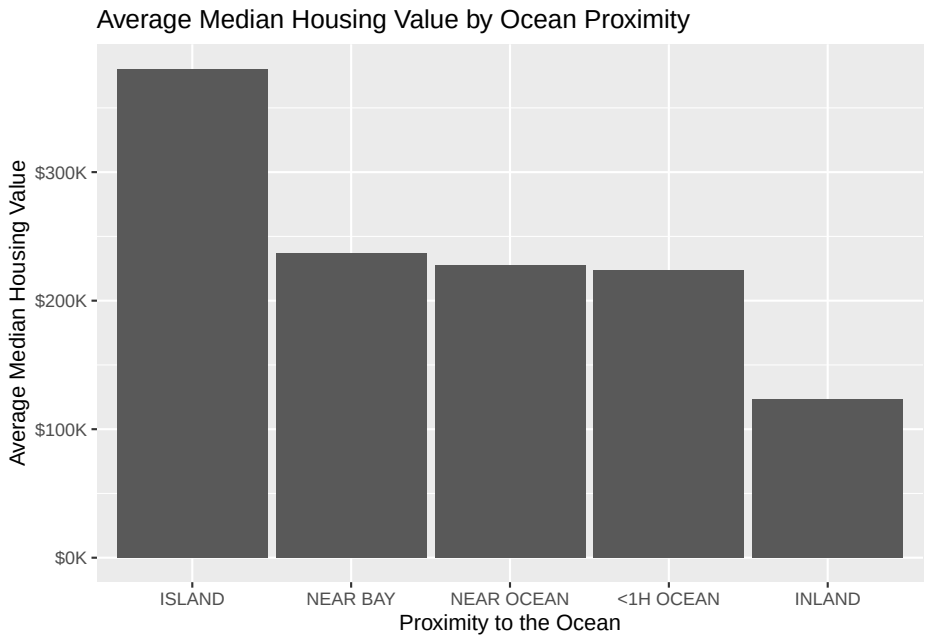



```
ggplot(data = housing_df, aes(x = longitude, y = latitude, color = median_house_value)) + geom_point() +
```



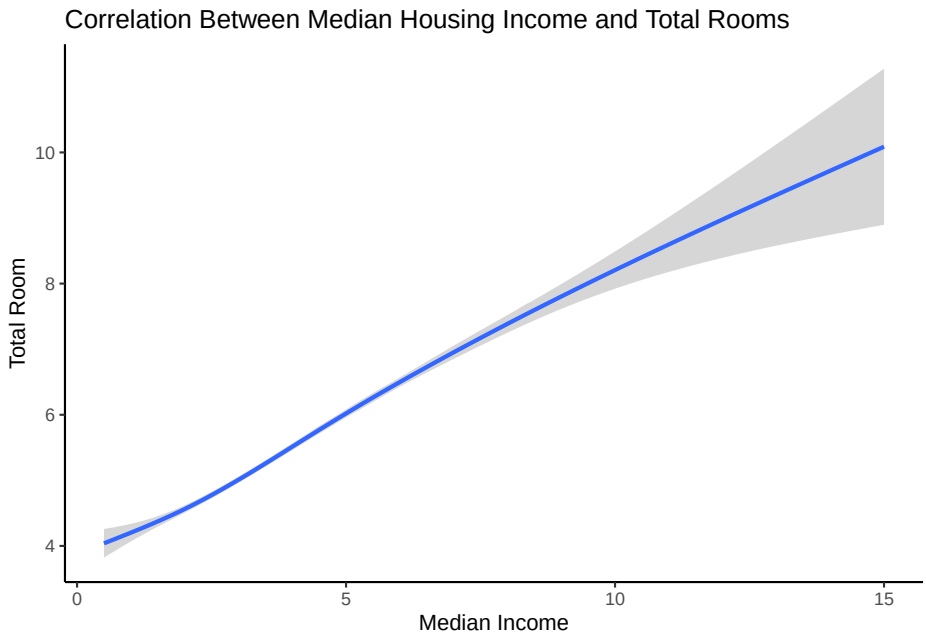

```
## This plot shows the correlation between the housing areas and median housing values in the state of
```

```
## Part IV, Question 5

## Bar plot
housing_df %>% group_by(ocean_proximity) %>% summarise(mean = mean(median_house_value)) %>%
  ggplot(mapping = aes(x = ocean_proximity, y = mean)) +  geom_col(width = .95) + ggtitle('Average Media
```




```
## According to this bar plot, it seems that the closer the homes are to the inland area, the more affo
```

```
## Part V: Bonus question
## Create a plot of your choosing showing any variables you'd like from the housing dataset. However, t

ggplot(data = housing_df, aes(x = median_income, y = total_rooms / households )) + geom_smooth() + ggti
```




```
## According to the smoothed prediction lines, it seems like there is a strong correlation between a hi
```

```
## Part VI: Create a GitHub issue
## Paste the url to your issue here: https://github.com/anyone-can-cook/rclass2_w22_student_issues/issu
## Paste the url to the issue you responded to here: https://github.com/anyone-can-cook/rclass2_w22_stu
```

References

Pace, Kelley R., & Barry, R. (1997). Sparse spatial autoregressions. *Statistics and Probability Letters*, *33*(3), 291–297.