# Stat 420 Final Project Work

STAT 420, Summer 2022, D. Unger

## Analysis Workbook

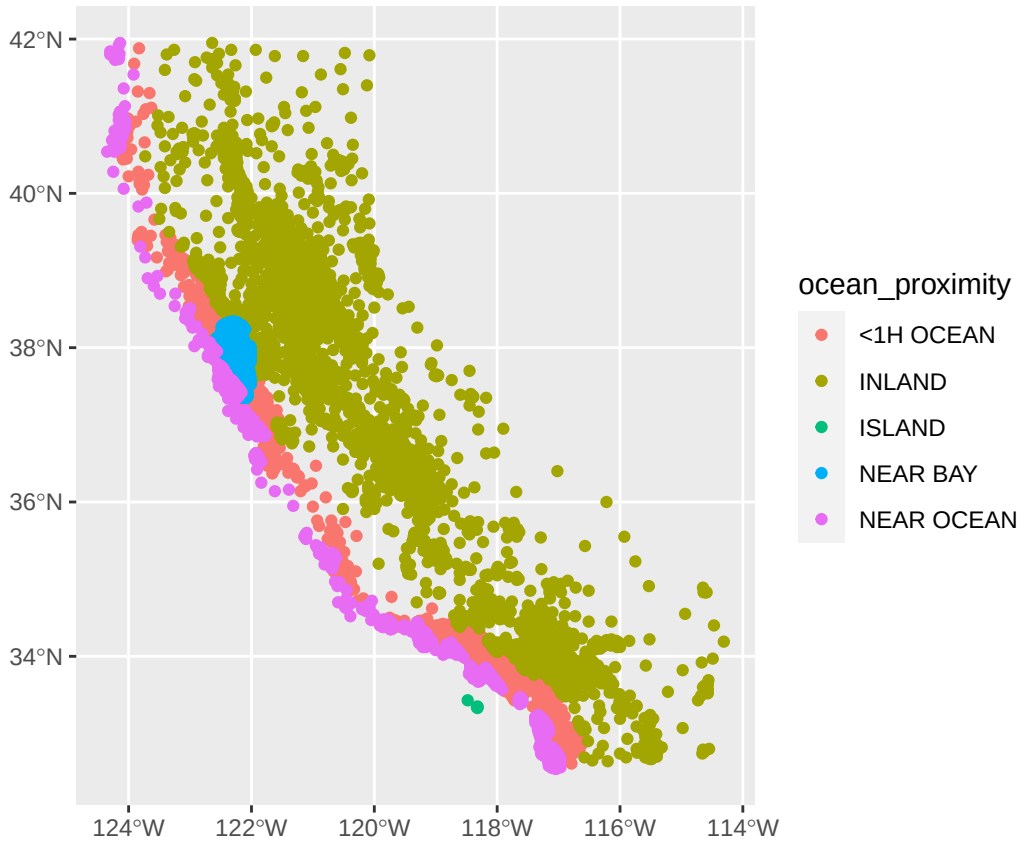
### Define Helper Functions

```
ca_housing_data = read.csv('../000_Data/california-housing-prices/housing.csv')

#convert_to_zip <- function(latitude, longitude){
 #https://stackoverflow.com/questions/11280145/convert-lat-lon-to-zipcode-neighborhood-name
#}
```

### EDA

-shared by 7/13

```
## Map the data
#Source: https://stackoverflow.com/questions/65233613/plot-a-map-using-lat-and-long-with-r
my_sf <- st_as_sf(ca_housing_data, coords = c('longitude', 'latitude'))
my_sf <- st_set_crs(my_sf, value = 4326)

ggplot(my_sf) +
  geom_sf(aes(color = ocean_proximity))
```




## Data cleaning

-agreed/coordinate -shared by 7/13

```
summary(ca_housing_data)
```

```
##    longitude       latitude    housing_median_age  total_rooms
##  Min.   :-124   Min.   :32.5   Min.   : 1.0       Min.   :    2
##  1st Qu.:-122   1st Qu.:33.9   1st Qu.:18.0       1st Qu.: 1448
##  Median :-118   Median :34.3   Median :29.0       Median : 2127
##  Mean   :-120   Mean   :35.6   Mean   :28.6       Mean   : 2636
##  3rd Qu.:-118   3rd Qu.:37.7   3rd Qu.:37.0       3rd Qu.: 3148
##  Max.   :-114   Max.   :42.0   Max.   :52.0       Max.   :39320
##
##  total_bedrooms   population      households   median_income
##  Min.   :   1   Min.   :    3   Min.   :   1   Min.   : 0.50
##  1st Qu.: 296   1st Qu.:  787   1st Qu.: 280   1st Qu.: 2.56
##  Median : 435   Median : 1166   Median : 409   Median : 3.54
##  Mean   : 538   Mean   : 1425   Mean   : 500   Mean   : 3.87
##  3rd Qu.: 647   3rd Qu.: 1725   3rd Qu.: 605   3rd Qu.: 4.74
##  Max.   :6445   Max.   :35682   Max.   :6082   Max.   :15.00
##  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
##
```

```
str(ca_housing_data)
```

```
## 'data.frame':    20640 obs. of  10 variables:
##  $ longitude         : num  -122 -122 -122 -122 -122 ...
##  $ latitude          : num  37.9 37.9 37.9 37.9 37.9 ...
##  $ housing_median_age: num  41 21 52 52 52 52 52 52 42 52 ...
##  $ total_rooms       : num  880 7099 1467 1274 1627 ...
##  $ total_bedrooms    : num  129 1106 190 235 280 ...
##  $ population        : num  322 2401 496 558 565 ...
##  $ households        : num  126 1138 177 219 259 ...
##  $ median_income     : num  8.33 8.3 7.26 5.64 3.85 ...
##  $ median_house_value: num  452600 358500 352100 341300 342200 ...
##  $ ocean_proximity   : chr  "NEAR BAY" "NEAR BAY" "NEAR BAY" "NEAR BAY" ...
```

```
#turn ocean_proximity to factor
ca_housing_data$ocean_proximity <- as.factor(ca_housing_data$ocean_proximity)

#create logged version of house value
ca_housing_data$logged_house_value <- log(ca_housing_data$median_house_value)

#How many missing values do we have?
colSums(is.na(ca_housing_data))
```

```
##          longitude           latitude housing_median_age        total_rooms
##                  0                  0                  0                  0
##     total_bedrooms         population         households      median_income
##                207                  0                  0                  0
## median_house_value    ocean_proximity logged_house_value
##                  0                  0                  0
```

```
#Yes, let's drop those with missing total_bedrooms
ca_housing_data_cln <- ca_housing_data[!is.na(ca_housing_data$total_bedrooms),]

#Do any have any with total_bedrooms > total_rooms?
sum(ca_housing_data_cln$total_bedrooms > ca_housing_data_cln$total_rooms)
```

```
## [1] 0
```

```
#Why is there a population of 3?
ca_housing_data_cln[ca_housing_data_cln$population == 3,]
```

```
##      longitude latitude housing_median_age total_rooms total_bedrooms
## 5343    -118.4    34.04                 16          18              6
##      population households median_income median_house_value ocean_proximity
## 5343          3          4         0.536             350000       <1H OCEAN
##      logged_house_value
## 5343              12.77
```

## Variable Creation

-shared 7/24

## Model Building

```
initial_model <- lm(logged_house_value ~ . - median_house_value, ca_housing_data_cln)
summary(initial_model)
```

```
##
## Call:
## lm(formula = logged_house_value ~ . - median_house_value, data = ca_housing_data_cln)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.384 -0.199 -0.009  0.191  3.371
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -2.31e+00   4.22e-01   -5.48  4.3e-08 ***
## longitude                 -1.62e-01   4.89e-03  -33.05  < 2e-16 ***
## latitude                  -1.57e-01   4.82e-03  -32.60  < 2e-16 ***
## housing_median_age         2.50e-03   2.11e-04   11.87  < 2e-16 ***
## total_rooms               -1.41e-05   3.80e-06   -3.72   0.0002 ***
## total_bedrooms             3.85e-04   3.30e-05   11.68  < 2e-16 ***
## population                -1.73e-04   5.16e-06  -33.47  < 2e-16 ***
## households                 2.53e-04   3.57e-05    7.07  1.6e-12 ***
## median_income              1.69e-01   1.62e-03  104.25  < 2e-16 ***
## ocean_proximityINLAND     -3.08e-01   8.37e-03  -36.82  < 2e-16 ***
## ocean_proximityISLAND      5.90e-01   1.47e-01    4.00  6.4e-05 ***
## ocean_proximityNEAR BAY   -3.83e-02   9.18e-03   -4.17  3.0e-05 ***
## ocean_proximityNEAR OCEAN -3.41e-02   7.53e-03   -4.53  6.0e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.329 on 20420 degrees of freedom
## Multiple R-squared:  0.665,  Adjusted R-squared:  0.665
## F-statistic: 3.38e+03 on 12 and 20420 DF,  p-value: <2e-16
```

```
smaller_model <- lm(logged_house_value ~ longitude + latitude + ocean_proximity + population, ca_housing
summary(smaller_model)
```

```
##
## Call:
## lm(formula = logged_house_value ~ longitude + latitude + ocean_proximity +
##     population, data = ca_housing_data_cln)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5009 -0.2958 -0.0096  0.3036  1.8962
##
```

```
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -7.09e+00   5.69e-01  -12.45  < 2e-16 ***
## longitude                  -2.26e-01   6.57e-03  -34.38  < 2e-16 ***
## latitude                   -2.16e-01   6.49e-03  -33.27  < 2e-16 ***
## ocean_proximityINLAND      -4.17e-01   1.14e-02  -36.58  < 2e-16 ***
## ocean_proximityISLAND       3.83e-01   2.04e-01    1.88    0.060 .
## ocean_proximityNEAR BAY    -2.35e-02   1.24e-02   -1.89    0.059 .
## ocean_proximityNEAR OCEAN -7.12e-02   1.03e-02   -6.90  5.4e-12 ***
## population                 -1.53e-06   2.83e-06   -0.54    0.590
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.455 on 20425 degrees of freedom
## Multiple R-squared:  0.36,   Adjusted R-squared:  0.36
## F-statistic: 1.64e+03 on 7 and 20425 DF,  p-value: <2e-16
```

## Model Selection

- share 7/24 -coordinate 7/24

## Graphs and Tables

## QA: How do we know what we did makes sense?.

## Move to final report

August 2nd (remember to knit often!!) Report Due August 5th