Slide 4:

* Thank you, Ellie. Moving on to descriptive analysis, we performed some analysis on the features in consideration using Tableau and Excel. The next three slides brief on the findings of the analysis.
* Both the geo plots here, show the density of households given a Lat LONG (Geo Info). In the first plot, we have highlighted the median house price of this density and it can be observed that the units along the coast are generally expensive and moving away from the coast, we can see that price starts to fall. There are areas that are clustered with expensive houses, namely in and around San Francisco and LA.
* From the second plot, we can observe that the number of Inland houses is the highest. This can be explained with the distribution of the general population around the areas that are not expensive. There is an interesting category of houses situated on ISLANDS and all of these are expensive.

Slide 5:

* The first plot shows the HL of MedHValue of houses situated in varying proximity from the ocean. Each box represents the interquartile range and it can be seen that units nearer to the ocean have wider ranges compared to inland units (diverse pricing options).
* From the second plot we can infer that there’s a strong positive correlation between median income and median house value, meaning big earners invest big on houses. Highlighting the data points using Ocean proximity shows us how income affects the house choices in different locations.
* From the third plot, we can see that new houses are found mostly inland. This shows recent shift in trend to develop inland area as the population around these places are on the rise. Bay area happens have the greatest number of old homes, giving us an idea about the initial settlement in the state.

Slide 6:

* The correlation plot gives us insights on the significance of the predictors. Median income is positively correlated with the dependent variable, hinting to give the highest positive impact in the regression model.
* We can find Four independent variables highly correlated with each other (Total rooms, Total Bedrooms, Population and Households). The high correlation between them is explainable, high population in a block would result in more homes in a block which would result in more rooms and bedrooms in the same block. We cannot use all of these features to train our model.
* We did some feature engineering, and calculated 3 new features from the existing set of features (rooms per household, Bedrooms per household and Mean occupation) to have sufficient number of features to perform regression analysis.
* The bar plot on the right shows the distribution of median house value. Its slightly shifted to the left with a long tail. However, a normal distribution can be observed around the peak. Theres a noticeable spike at the end indicating a cap on the feature. (handle outliers, normalization)

Slide 7:

* The Reg. analysis produced a fairly strong model. We can see that from the Adjusted R2 value, 62% of the variance in the house value can be explained.
* The high F-statistic and low p values of the features confirm the statistical significance assuring a reliable house price prediction.
* From the coefficient column, we can infer that Geo location, Ocean proximity, median income and bedrooms per household are the strongest predictors.
* The negative coefficient of the geo location indicates that as we move further south east, house values will decrease.
* Median income has the highest positive impact, proving our initial analysis, high earners invest more in houses.
* It can also be seen that older properties with more rooms/bedrooms will increase in value.
* I will hand it over to Ellie to give us the