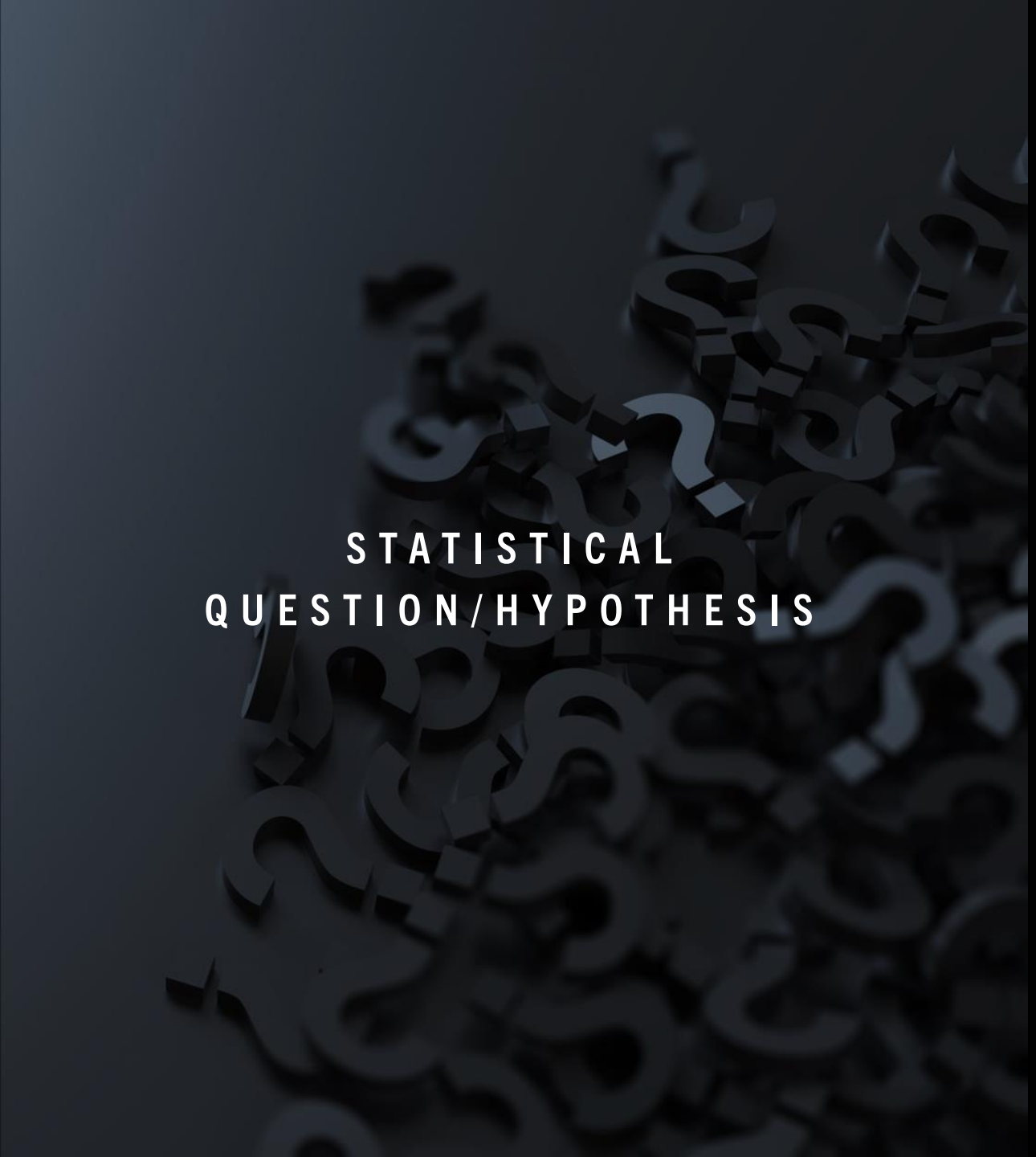
DSC530 TERM PROJECT

EDA and Analysis on Housing

Eli Forta



STATISTICAL QUESTION/HYPOTHESIS

Are houses with more bathrooms more expensive?

VARIABLES USED FOR ANALYSIS AND WHAT THEY MEAN

salePrice is how much each house was sold for in USD. This is our metric for determining the value of a home.

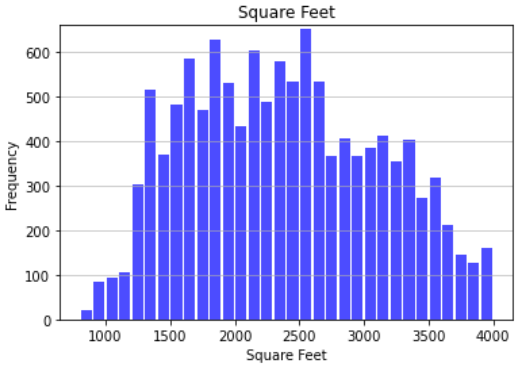
bathrooms is the total amount of bathrooms per home. This is key to our hypothesis that bathrooms have an effect on price.

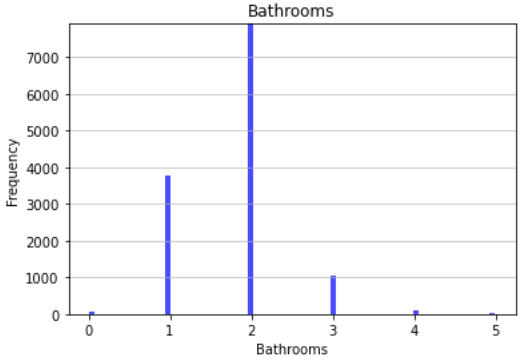
squareft is the total square footage of each home. This will tell us what portion of a house's price is determined by size.

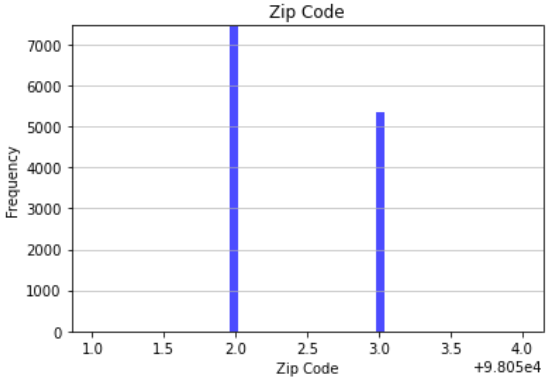
zipcode is the zip code of each home. This will help us compare homes within similar geographic and socioeconomic areas.

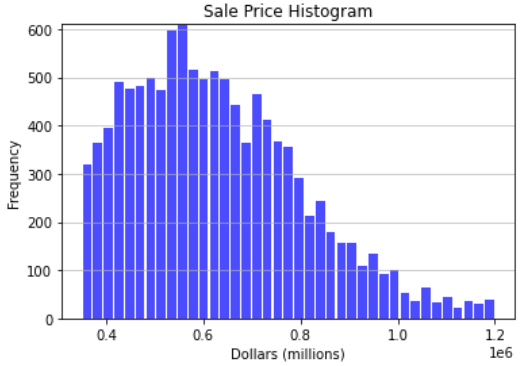
grade is a score given for the quality of a given structure out of eleven. This will help us with comparing homes with similar structural integrity.

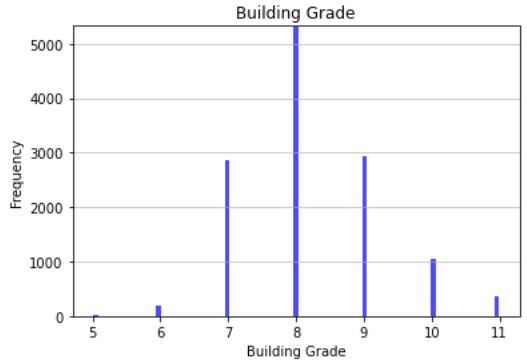
HISTOGRAMS OF VARIABLES











SUMMARY AND ANALYSIS: OUTLIERS

- Outliers: All houses sold for more than 1.4 million were removed from the analysis as they are extreme outliers in this dataset. They are 5.1 Standard Deviations from the mean.
- Additionally, houses which sold for less than 250 thousand were removed for the same reason.

SUMMARY AND ANALYSIS: MEAN AND MODE

- Mean: The mean price is 660,000, mean of bathrooms is 1.8, mean square footage is 2,540, the other variables are discrete values.
- Mode: Price is right-skewed with one cluster, Bathrooms has a Gaussian distribution, Square Feet has many clusters, and its utility is questionable, Zip Code only has two values, and Grade is left-skewed with a cluster at the mean.

SUMMARY STATISTICS: SPREAD AND TAILS

Price and Square Feet appear to have high variability. The others do not.

Square Feet does not drop off at any point, it is evenly spaced. Price is right-skewed and drops off around the mean. Grade is left-skewed and is smaller at values lower than the mean.

CDF RESULTS

The CDF tells us that 90% of homes are worth less than a million dollars. This will help us key in on those homes as they are more representative of the sample.

Additionally, given this new information, we may have to normalize the distribution.

ANALYTICAL DISTRIBUTION

WE PLOTTED A NORMAL PROBABILITY PLOT FOR SALEPRICE.

THE RESULTS SEEM TO INDICATE A POOR FIT, OR PERHAPS A MISUSE OF THE NORMAL DISTRIBUTION.

CORRELATION AND CAUSATION

- Our scatter plot does not seem to indicate any correlation between bathrooms and price.
- Covariance between price and bathrooms is undoubtedly positive, this alone indicates that the number of bathrooms could have a causal relationship with price. However, it is also possible that there is another factor causing both to rise.
- Pearson's backs up our basic correlation function; namely, that the correlation between price and bathrooms is relatively low.
- Spearman's, while still relatively low, is far higher than Pearson's. This, coupled with the high covariance indicates the possibility of a non-linear relationship between price and bathrooms.

CONCLUSION

- Based on our tests, we can say that while there exists positive covariance between bathrooms and price, there is certainly no linear correlation.
- Both factors are directly affected by Square Feet, so that is a possible factor.
- Another possibility is that the tertiary cause is outside the scope of this dataset.
- Either way, we fail to reject or prove the hypothesis