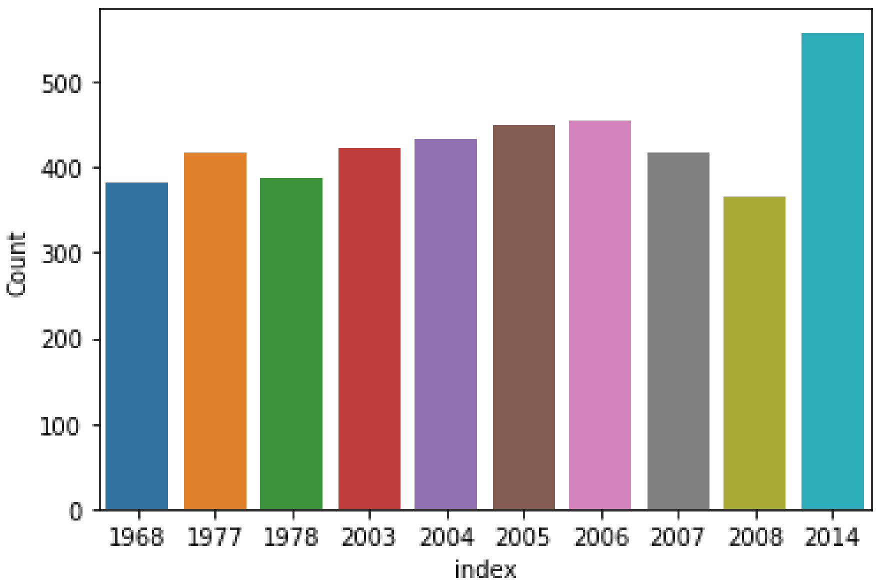
Q(B)

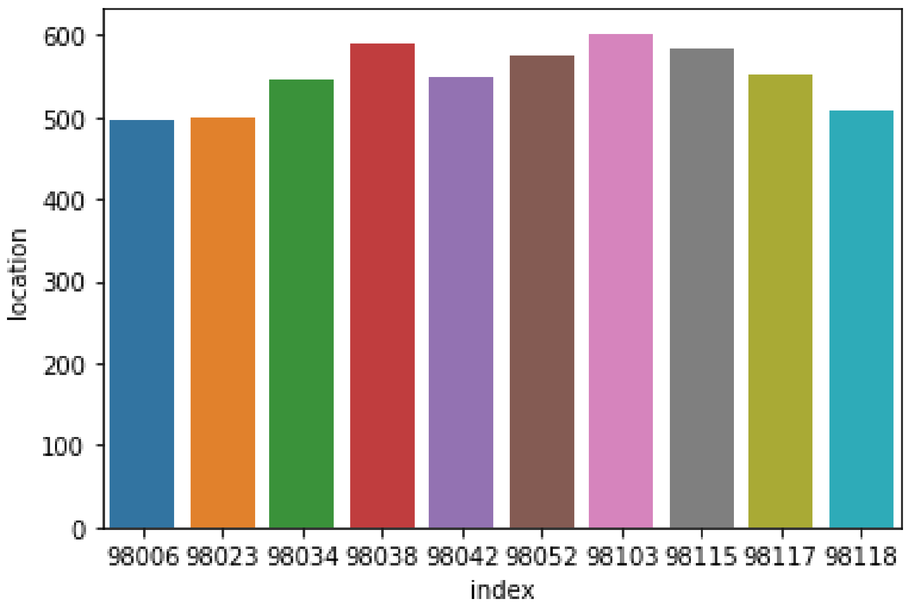
1. Document 5-6 key insights from EDA and support each point with a visualization.

#1


YEAR AND COUNT

2014 shows the best sale.

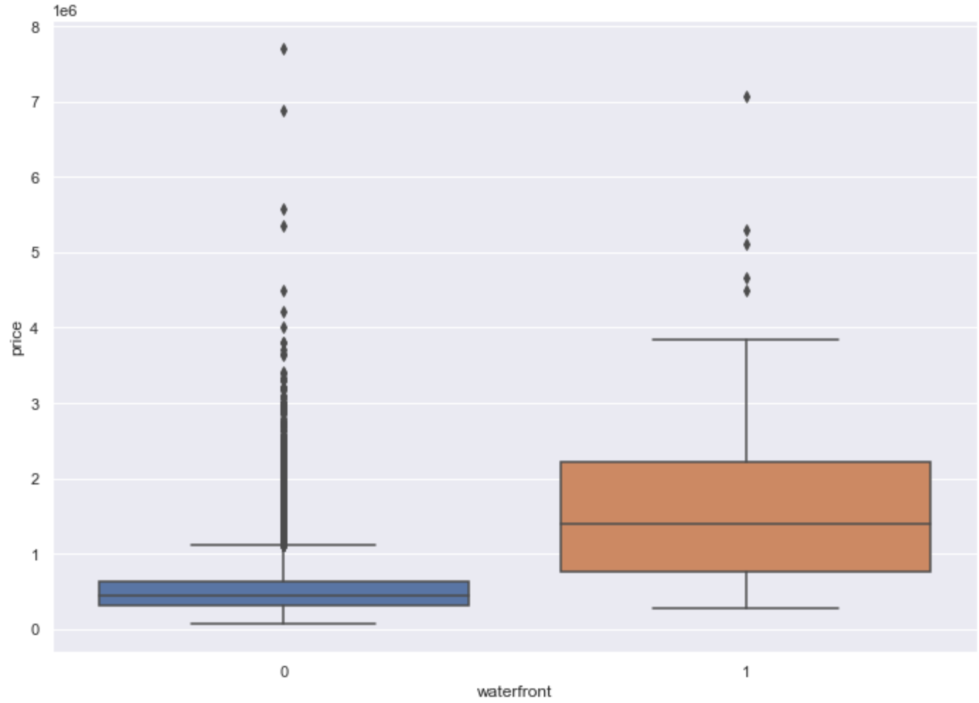
#2


TOP LISTED LOCATION

Top listed locations for home sale.

#3

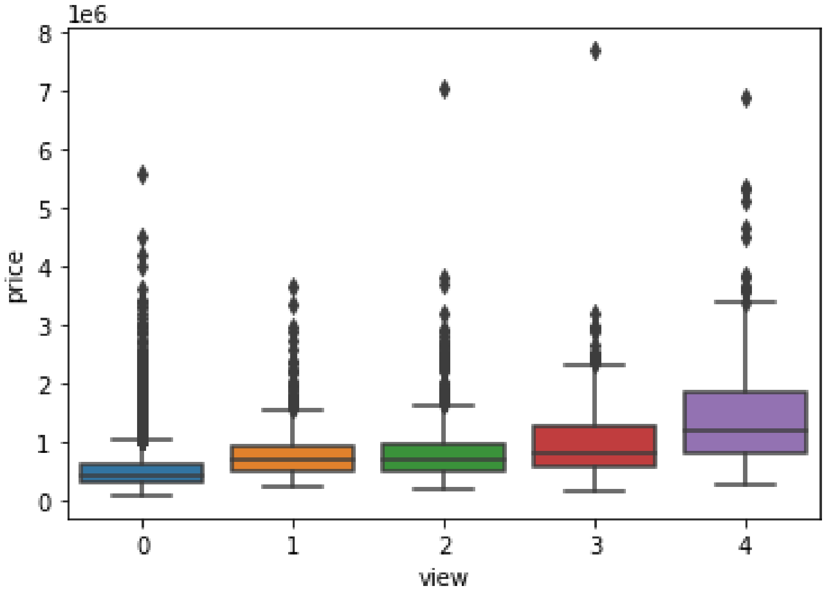
PRICE VS HOME WITH WATERFRONT



Graph shows that houses with waterfront has more demand than without.

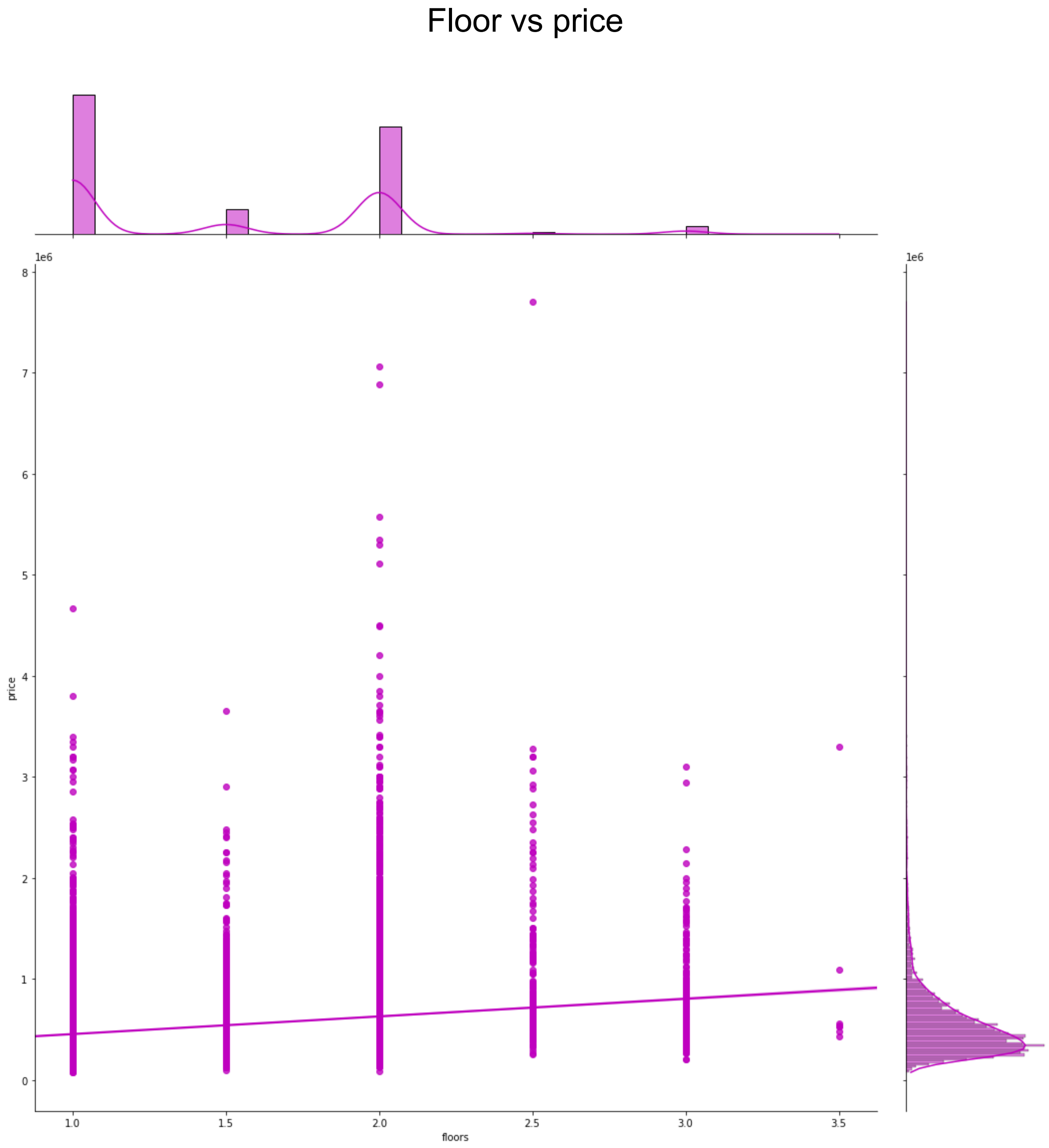
#4

VIEW VS PRICE



House with view is on demand so had more price.

#5



Houses with 2 floors had more demand hence the price.

2.

i. What are the assumptions of linear regression?

1. Independence of observations
2. No Hidden or Missing Variables
3. Linear relationship - relations between the independent and dependent variables must be linear.
4. Normality of the residuals - For any fixed value of X, Y is normally distributed.

5. No or little Multicollinearity - Multicollinearity is the phenomenon when a number of the explanatory variables are strongly correlated.
6. Homoscedasticity - Homoscedasticity in a model means that the error is constant along the values of the dependent variable i.e The variance of residual is same for any value of X.
The best way for checking homoscedasticity is to make a scatterplot with the residuals against the dependent variable.
7. All independent variables are uncorrelated with the error term - check whether there is correlation between any of the independent variables and the error term.

ii. How can we evaluate a Regression model? Define each metric and its interpretation.

1. **R Square/Adjusted R Square**

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

- R Square measures how much variability in dependent variables can be explained by the model.
- It is the square of the Correlation Coefficient(R)
- R Square value is between 0 to 1 and a bigger value indicates a better fit between prediction and actual value.

2. **Mean Square Error(MSE)/Root Mean Square Error(RMSE)**

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

- Mean Square Error is an absolute measure of the goodness of fit.
- Root Mean Square Error(RMSE) is the square root of MSE

3. **Mean Absolute Error(MAE)**

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|$$

- Mean Absolute Error(MAE) is similar to Mean Square Error(MSE). However, instead of the sum of squares of error in MSE, MAE is taking the sum of the absolute value of error.is a more direct representation of sum of error terms. **MSE gives larger penalization to big prediction errors by square it while MAE treats all errors the same**.

iii. Can R squared be negative?

- R-Squared is a statistical measure of fit that indicates how much variation of a dependent variable is explained by the independent variable(s) in a regression model.
- A negative R2 is only possible with linear regression when either the intercept or the slope are constrained so that the "best-fit" line (given the constraint) fits worse than a horizontal line. With nonlinear regression, the R2 can be negative whenever the best-fit model (given the chosen equation, and its constraints, if any) fits the data worse than a horizontal line.

iv. What is a dummy variable trap?

The Dummy variable trap is a scenario where there are attributes that are highly correlated (Multicollinear) and one variable predicts the value of others. When we use *one-hot encoding* for handling the categorical data, then one dummy variable (attribute) can be predicted with the help of other dummy variables. Hence, one dummy variable is highly correlated with other dummy variables. Using all dummy variables for regression models leads to a ***dummy variable trap***. So, the regression models should be designed to exclude one dummy variable

v. Is One Hot Encoding different from Dummy Variables?

For transforming categorical attributes to numerical attributes to perform regression models, we can use the label encoding procedure (label encoding assigns a unique integer to each category of data).But this procedure is not alone that suitable, hence, ***One hot encoding*** is used in regression models following label encoding. This enables us to create new attributes according to the number of classes present in the categorical attribute i.e if there are *n* number of categories in categorical attribute, *n* new attributes will be created. These attributes created are called ***Dummy Variables***. Hence, dummy variables are “proxy” variables for categorical data in regression models.
These dummy variables will be created with one-hot *encoding* and each attribute will have a value of either 0 or 1, representing the presence or absence of that attribute.

vi. How is polynomial regression different from linear regression?

Polynomial regression is a form of Linear regression where only due to the Non-linear relationship between dependent and independent variables we add some polynomial terms to linear regression to convert it into Polynomial regression.

vii. Interpret the screenshot below from the notebook we discussed in class today

Here score shows the r square value where, R-Squared is a statistical measure of fit that indicates how much variation of a dependent variable is explained by the independent variable(s) in a regression model.
The regression predictions exactly fit the data if the r2 is 1.When the model fits the data worse than the poorest possible least-squares predictor, r2 values outside the range 0 to 1 occur.
When you add more parameters, r2 continues to rise, in which r2 = 1 denotes a perfect fit.

viii. Bonus: We saw Sweetviz as an Automated EDA option. What are the other options? Try a few of them and share which one did you find the best.

Python provides certain open-source modules that can automate the whole process of EDA and save a lot of time.
1.Pandas Profiling
2.Autoviz
3.sweetviz

All these three automated EDA libraries have their own advantages and disadvantages over the others. By using these you could save up a lot of time and get results quickly. The overall purpose of these libraries is to help in:

- Feature engineering: visualize how engineered features perform/correlate relative to other features and the target variable
- Interpretation/communication: the generated graphs can provide insights that are easily interpretable and can be passed amongst a team or to clients quickly, without any extra work
- Testing: confirm the makeup & balance of testing/validation sets