EDA Presentation

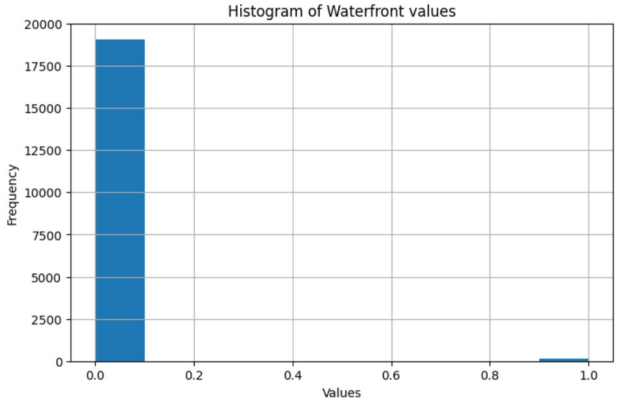
by David Cooper

Jennifer Montgomery

- Has a high budget
- Likes to show off
- Timing within a month
- Waterfront property
- Renovated
- High grade
- Can be resold within a year
- I have formulated 3 hypotheses from these requirements with the data given

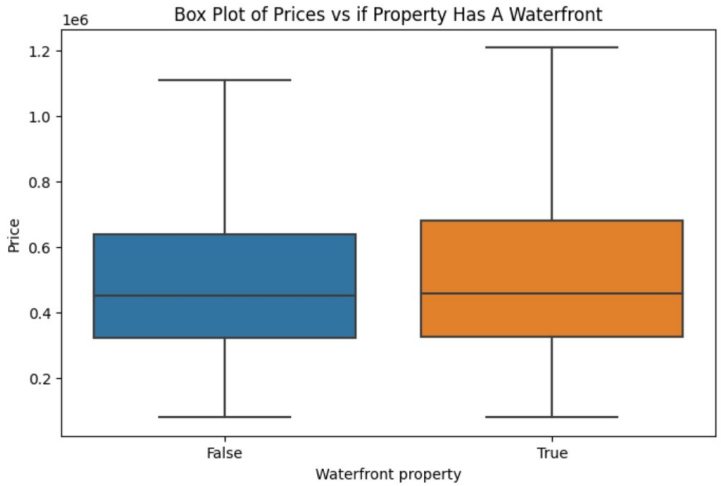
Waterfront properties

- My first hypothesis was that waterfront properties would be more expensive
- After a couple of simple plots (below), it was clear the waterfront values had to be changed
- I changed the values to booleans in order to represent the waterfront vs price comparison effectively
- However a useful insight from the histogram is that there are a very low number of properties in this data set that are waterfront properties



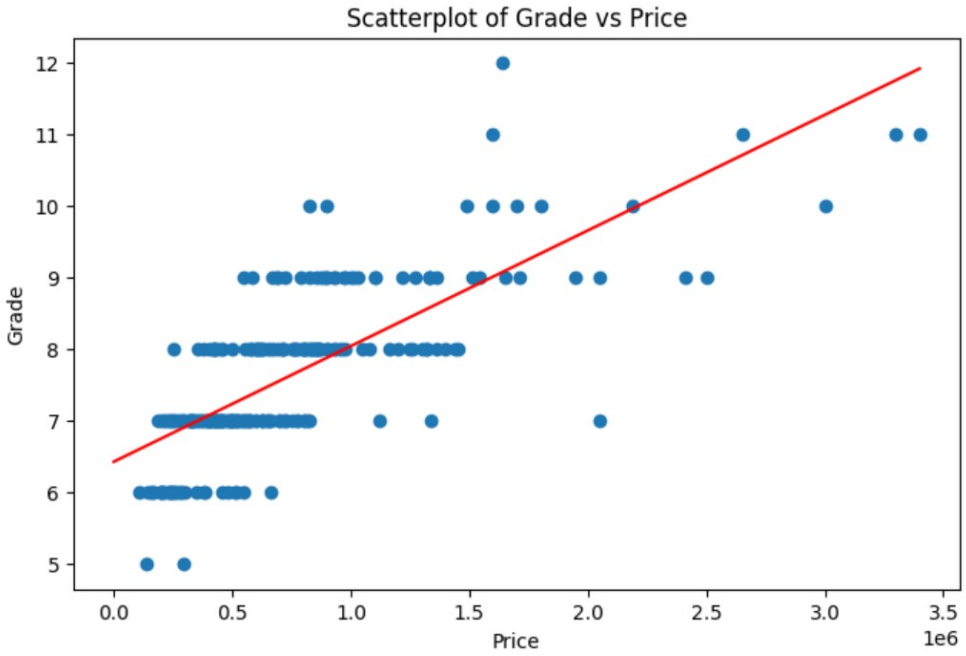
Waterfront properties more expensive?

- When plotted without outliers, there is only slight skew in price when comparing with to without waterfront
- Even with outliers included, there is not a strong pattern (perhaps even a weaker correlation)
- From this I drew the conclusion my hypothesis was, for the most part, wrong and that perhaps this should be deprioritized as a requirement, as there are also just not a large quantity of these properties in general
- It is also clear the data needs to be cleaned more



Satisfying price, grade, renovation needs

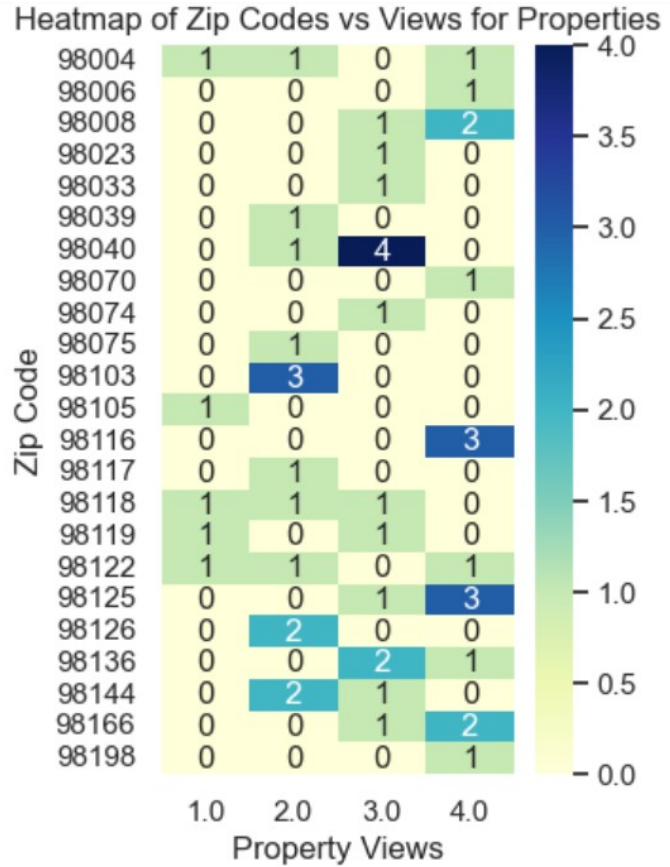
- I cleaned the data by removing the bottom 60% priced properties
- The buyer has a high budget, but not necessarily top 10% so I decided starting with the top 40% would be appropriate
- I then removed the properties that hadn’t been renovated, and plotted a scatter graph to compare price and grade
- This lead me to my second hypothesis: that properties with a high price will have a higher grade



- Graph shows a vaguely positive correlation
- Not many extreme outliers
- Client can be advised that generally, the grade of the property will be reflected in the price
- My hypothesis was mostly true in this case

Location vs Viewings

- Lastly, I will add a geographical insight to my analysis by comparing property views to zip codes
- My hypothesis is that there will be certain zip codes where properties are viewed more
- It can be assumed these areas are more desirable, and that there is more movement in the housing market there generally
- This would satisfy my clients desire to “be flashy” and a property in these areas would likely resell quicker, as client wanted to resell the property in a year
- I removed 0 views properties from my data, and plotted a heat map with the remainder



- This heat map shows that 98008, 98040, 98116, 98118, 98122, 98125, 98136, 98144, 98166 are the zip codes with the most viewing activity
- My hypothesis that there would be zip codes with higher property views appears to be true
- These could be recommended areas for the client to buy as she wants to resell in a year, and these seem like desirable areas

Summary

- From my analysis of the data, and the conclusions of my hypotheses I can recommend the following to the client:
- Discarding or at least deprioritizing buying a waterfront property, as there are simply so few of them. Also, with no strong correlation to price and grade it compromises the clients other requirements
- Many of the properties are renovated, and the customers desire for a high grading will generally be reflected in the price – good news for her as she has a high budget
- There are specific areas she can target where it seems likely the property can be resold quickly, and are considered desirable (as she likes to show off), such as the zip codes outlined in the previous slide where properties are viewed more often.