The Objective

The purpose of this exercise is to import, examine, analyze and report on real estate data to answer the questions below.

The Details

- Use any code-based or point-and-click analysis technology (such as SQL, R, Tableau, Excel) to solve these problems. We understand that access to your preferred tool may be limited.
- Deliver your solutions in a format that you are comfortable with. We recommend MS Word, MS PowerPoint, Tableau, PDF, or something similar.
- External data sources can be used for analysis or validation, but they are neither required nor necessary.

The Data

There are three files used to solve these problems. Here are their schemas.

- **Homes.csv** - Sample home set
  - PropertyID - Unique identifier for the home
  - Street - The street for the home address
- **Transaction.csv** - All property sales over a 10-year period for homes in Homes.csv
  - PropertyID - Unique identifier for the home
  - SaleDate - Date of the sale of the property
  - SaleAmount - Sale price of the property
  - ZestimateOnSaleDate - Zillow's estimated home value on the sale date
- **ZestimateHistory.csv** - Annual Zestimate history for homes in Homes.csv
  - ZestimateDate - Date of the Zestimate
  - PropertyID - Unique home identifier
  - ZestimateAmount - Dollar amount of Zestimate on the Zestimate date

The Questions

1. Tell us which tool(s) you chose to organize/analyze the data and why?
2. What are the 5 most and least expensive homes sold in 2007? Do these values seem realistic? Please explain your reasoning both qualitatively and quantitatively.
3. For each street, what are the average, minimum, and maximum Zestimate values in 2007?
4. What is the median Zestimate (aka Zillow Home Value Index, or ZHVI) for all homes by year? Why would Zillow calculate the ZHVI using the median instead of the mean?
5. Using the Transaction.csv file, calculate the annual median Zestimate percentage error and provide us with the results (the Zestimate percentage error is defined as the difference between the Zestimate and actual sale price, stated as a percentage of the sale price). Do these results suggest we have any systematic errors in our Zestimates? Explain.
6. Combine your results from #3, #4, and #5 to calculate a new set of ZHVIs for all homes, by year, that takes into account any systematic errors you may have uncovered, and explain what you did/why you did it.
7. Why should Zillow be concerned about the median Zestimate percentage error, and how could you apply your findings from 5 to help improve the prediction?

These data are simulated for the purposes of candidate screening only. Any anomalies or issues with these data should not be assumed to apply to actual housing data.