Ryan Kulyassa
Data 502 Midterm
3/18/25

Part 2: Short Answer

Questions:

1. How many rows and columns are in the dataframe?
    a. 27901 rows and 18 columns
2. How many categorical (text-based) variables are in the dataframe? For each of these variables provide a brief description of what it means / refers to.
    a. There are 8 categorical/text-based variables.
        i. "Gender" - the gender of the student (for this dataset, just Male/Female)
        ii. "City" - the city the student lives
        iii. "Profession" - the vast majority are just "Student"s, but some are "Architect"s, "Teacher"s, etc
        iv. "Sleep Duration" - a nominal grouping of how many hours of sleep the student gets, ranging from <5 to >8 hours
        v. "Dietary Habits" - the dietary habit of the student (Unhealthy, Moderate, Healthy)
        vi. "Degree" - the degree the student has attained, lots of different kinds in this data
        vii. "Have you ever had suicidal thoughts?" - whether the student has had suicidal thoughts (Yes/No)
        viii. "Family history of Mental Illness" - whether the student has a family history of mental illness (Yes/No)
3. How many female students over the age of 21 are depressed?
    a. 4,921 students
4. Which degree has the highest mean Academic Pressure?
    a. "Class 12" with mean Academic Pressure of 3.359375.
    b. I am unsure as to whether "Class 12" is actually a degree, so otherwise it would be "B.Arch" with mean Academic Pressure of 3.063599.
5. Which city with at least 100 students has the highest percentage of depressed students?
   Hint: Add the following code to create a user ID column
   `df['User ID'] = df.index`
    a. Ahmedabad with 67.2976% of students depressed
6. What is the maximum satisfaction among students in the city of Srinagar?
   Hint: `'max'` denotes the maximum.
    a. 5.0
7. Among students, which combinations of sleep duration and dietary habits were the most and least depressed on average?

    - Exclude the Others dietary habit.
    - Use the *percentage depressed* aggregation.
    - Are these results surprising to you?'

    a. The combinations that were most depressed on average:
        i. "Less than 5 hours" of sleep with "Unhealthy" dietary habits: 0.760274
        ii. "7-8 hours" of sleep with "Unhealthy" dietary habits: 0.718495

iii. “5-6 hours” of sleep with “Unhealthy” dietary habits: 0.680587

b. Least depressed on average:
   i. “More than 8 hours” of sleep with “Healthy” dietary habits: 0.366385
   ii. “5-6 hours” of sleep with “Healthy” dietary habits: 0.458731
   iii. “7-8 hours” of sleep with “Healthy” dietary habits: 0.467099
c. These results are not surprising to me, because those who were most depressed on average all had unhealthy dietary habits. Also, the most depressed group on average was those with unhealthy dietary habits and little sleep. This makes sense because lack of sleep and poor diet would definitely contribute to depression. Also, those who were least depressed on average all had healthy dietary habits, and all got decent amounts of sleep, which also makes sense because eating healthy and sleeping well will lead to lower rates of depression.

Part 3: Long Answer

8.

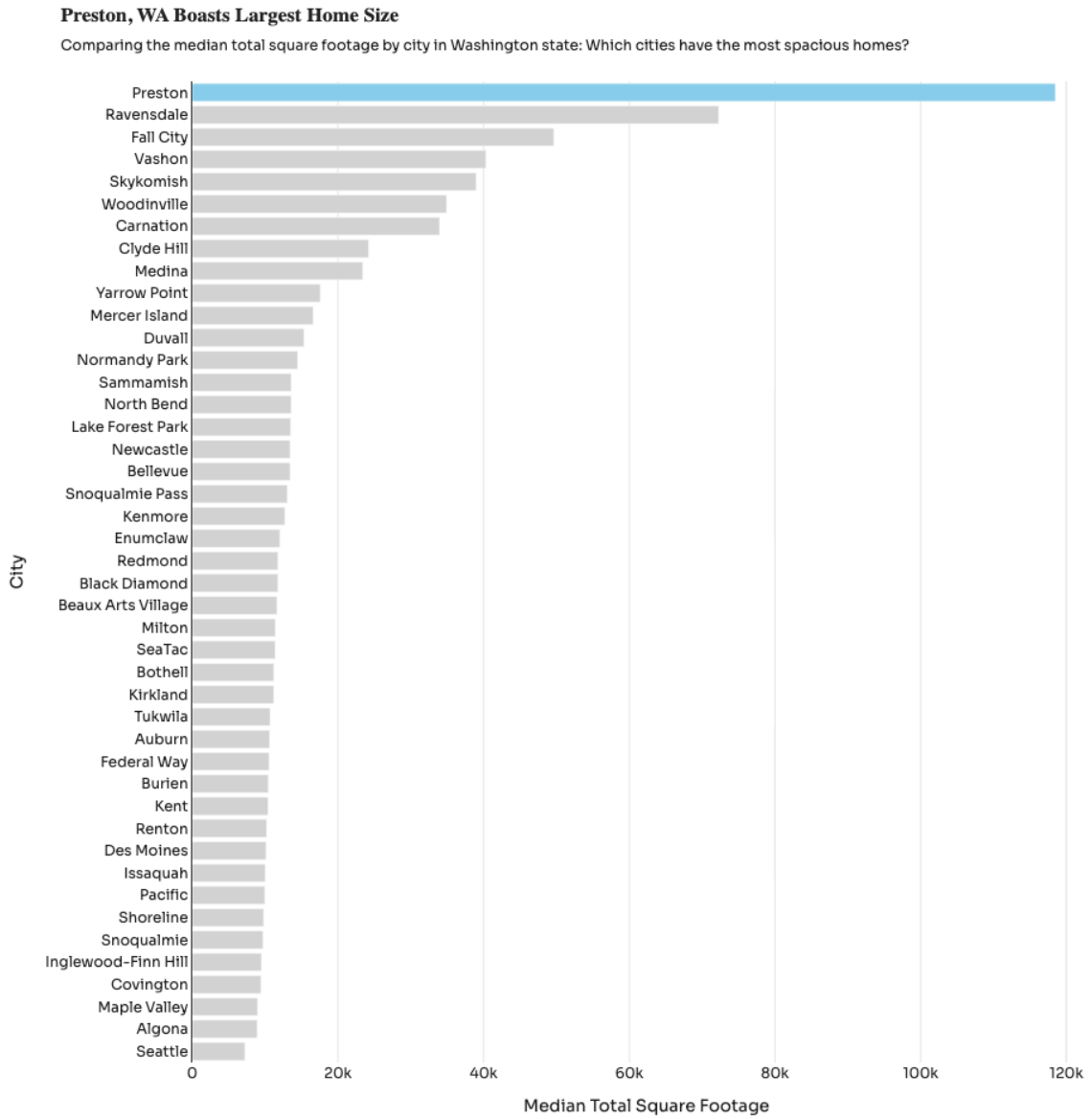
Prices per address in the city of Black Diamond

A comparison of the top home sales in the Black Diamond housing market.

Address	Price
28815 237th Pl SE	650000.0
22426 SE 300th St	531500.0
30734 229th Pl SE	510000.0
29613-29615 232nd Ave SE	398000.0
22821 SE 288th St	255000.0
25502 Kanasket Dr	253000.0
21612 SE 290th Ct	234950.0
32428-32598 5th Ave	224000.0

Source: Adil Shamim

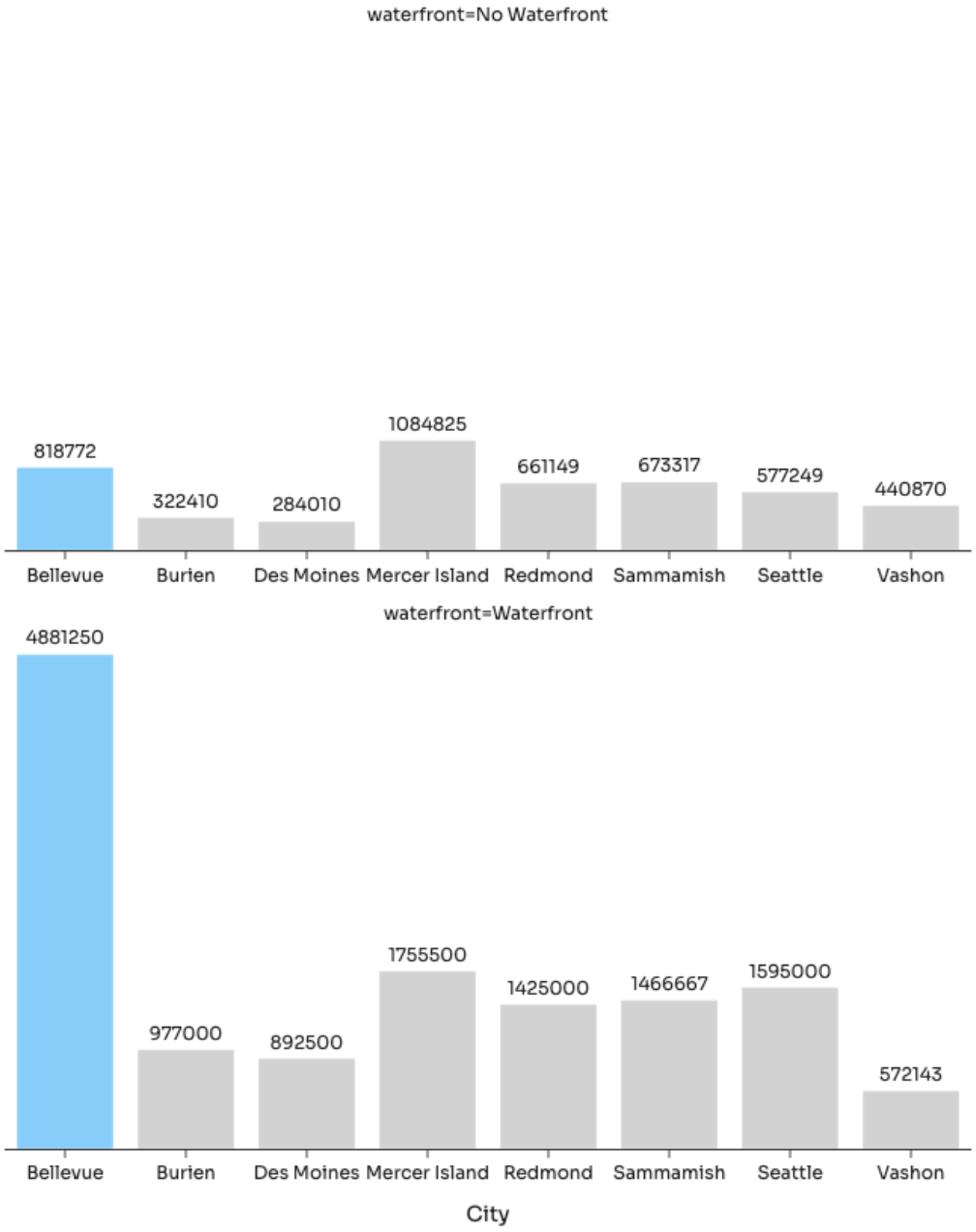
9.


Preston, WA Boasts Largest Home Size
Comparing the median total square footage by city in Washington state: Which cities have the most spacious homes?
Preston
Ravensdale
Fall City
Vashon
Skykomish
Woodinville
Carnation
Clyde Hill
Medina
Yarrow Point
Mercer Island
Duvall
Normandy Park
Sammamish
North Bend
Lake Forest Park
Newcastle
Bellevue
Snoqualmie Pass
Kenmore
Enumclaw
Redmond
Black Diamond
Beaux Arts Village
Milton
SeaTac
Bothell
Kirkland
Tukwila
Auburn
Federal Way
Burien
Kent
Renton
Des Moines
Issaquah
Pacific
Shoreline
Snoqualmie
Inglewood-Finn Hill
Covington
Maple Valley
Algona
Seattle
City
0
20k
40k
60k
80k
100k
120k
Median Total Square Footage

10.

Where Are Properties Most Expensive?

Waterfront properties regularly sell for much higher prices than inland properties, notably in Bellevue, WA.



11. Pick your best plot and answer the following questions.
I am choosing my 2nd plot from question 9, the horizontal bar chart.

- What Gestalt Principles of Visualization apply to your visualization? How?

- Continuity: the columns are sorted in height in descending order
- Color: Preston is highlighted light blue, while the rest are light gray, drawing attention to this column
- Length: the length of the bars represent the median total square footage of the homes in each city
- The textbook states on page 27, “Preattentive attributes are effects that seem to pop out from their surroundings. There are many we can use to tap into our reader's visual processing network to draw their attention: shape, line width, color, position, length, and more.” Through incorporating multiple of these preattentive attributes into my plot, I am able to effectively highlight data in my plot and convey my message.

- How does the scale of elementary perceptual tasks apply to your visualization?

Since we are using a horizontal bar chart, we are able to utilize position along common scales, enabling the most accurate estimates according to the scale of elementary perceptual tasks. The textbook reinforces this on page 15 by stating “Standard graphs, like bar and line charts, are so common because they are perceptually more accurate, familiar to people, and easy to create.” My plot also incorporates some other elements of the scale, such as length and color, however they do not significantly contribute to the viewer’s ability to get accurate estimates as much as the common scales.

- How has clutter been mitigated in your visualization?

Clutter has been mitigated in many ways in my plot, such as:

- Removing extraneous ticks on both axes. From textbook page 76: “Bar charts don't need tick marks between the bars. White space is an effective separator and deleting the tick marks reduces clutter.”
- Removing unnecessary line on x axis
- Removing the legend used for the indicator variable
- Rotating the graph so that the axis labels are oriented normally. From textbook page 79: “The most elegant solution is to simply rotate the entire graph. This still uses the same pre- attentive attribute— the length of the bars-but the axis labels are now aligned horizontally; they are easy to read with no effect on data comprehension.”
- The quantity of labels and text in general is minimized, as the textbook on page 31 also notes that otherwise, “too much text and too many labels, cluttering the space and crowding out the data”