Introduction to Stochastics
Exercise 5: Group Work

1 Reading Plots

The following plots visualize certain characteristics of a synthetic dataset about student performance factors. Answer the questions about the plots.

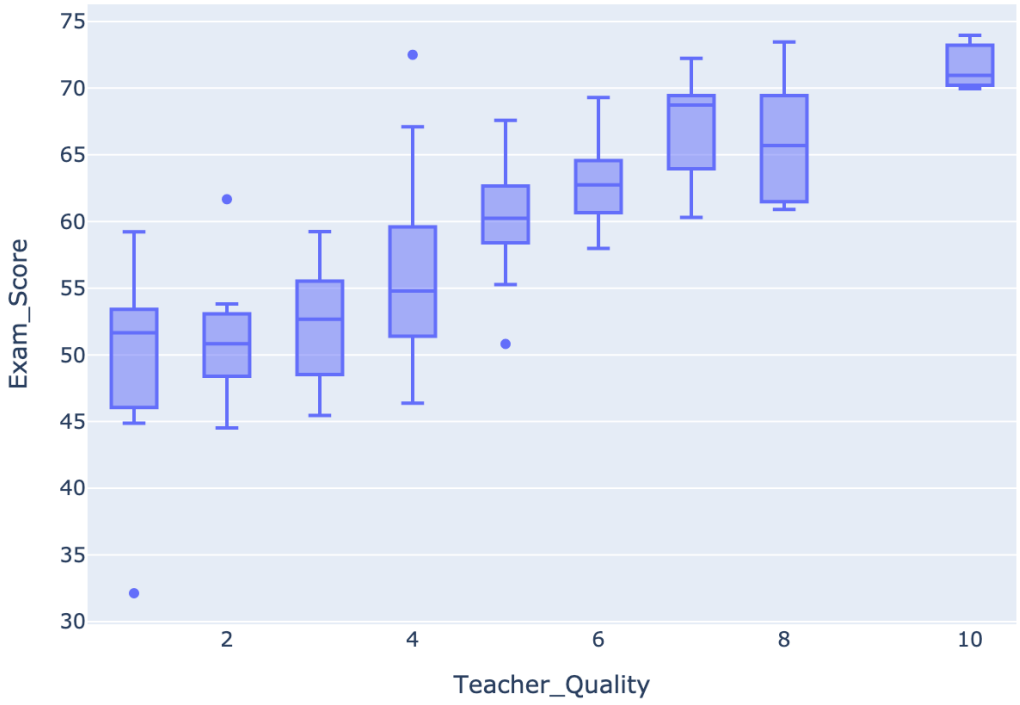

Figure 1: Boxplot

- What is the median exam score (approximately) for a teacher quality level of 4?
- Which teacher quality level contains outliers in its exam score distribution?

- Between teacher quality levels 2 and 8, which has a higher maximal exam score?
- Estimate the five-number summary for teacher quality level 8 and 4.

Solutions and Explanation

- **Median exam score for teacher quality level 4:** To determine the median, locate the horizontal line inside the box corresponding to level 4 on the x-axis. The median is approximately **55**.
- **Teacher quality levels with outliers:** Outliers appear as individual points beyond the whiskers of a boxplot. Based on the plot, levels **1, 2, 4, and 5** show clear outliers in their exam score distributions.
- **Higher maximum exam score between levels 2 and 8:** The maximum value is indicated by the top whisker or the highest point if an outlier exceeds it. Level 8 has a higher maximum score than level 2, with an estimated max of approximately **74**.
- **Five-number summaries:** To estimate these values, observe the bottom whisker (minimum), bottom of the box (Q1), median line, top of the box (Q3), and top whisker (maximum):
  - **Level 8:** [61, 62, 66, 69, 74]
  - **Level 2:** [44.5, 48.5, 51, 53, 61]

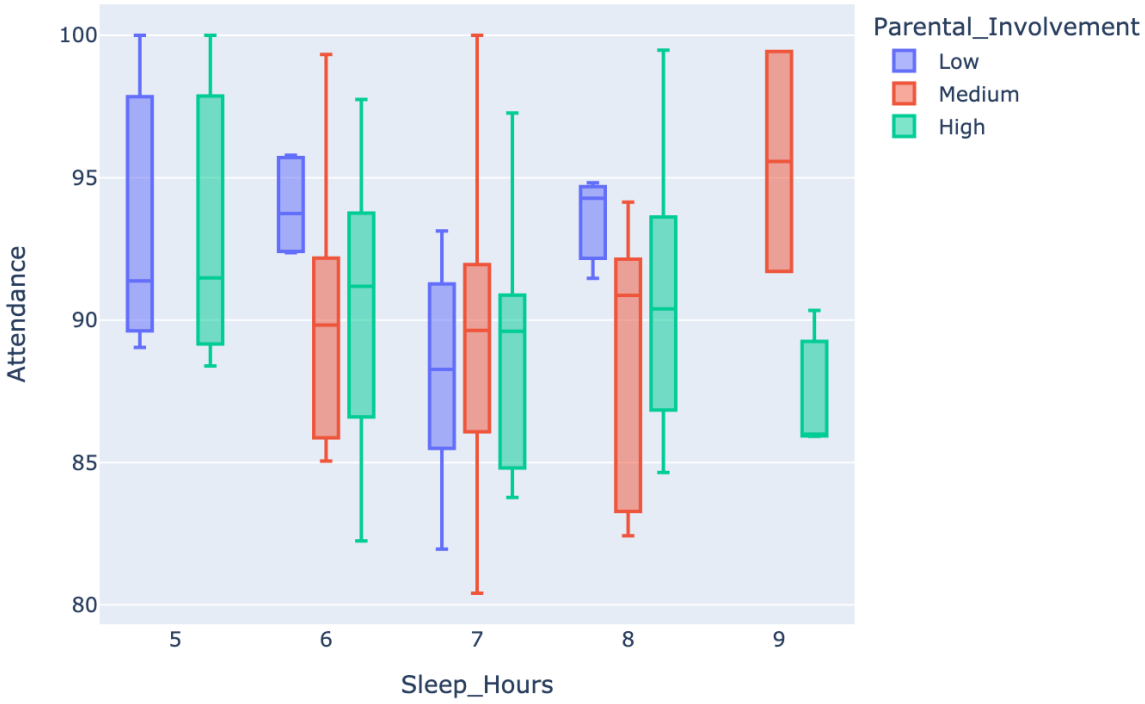

Figure 2: Boxplot 2

- Which parental involvement class includes students who sleep 5 hours?
- Which amount of sleep corresponds to the lowest median attendance rate when parental involvement is high?

Solutions and Explanation

- **Parental involvement categories with students who sleep 5 hours:** From the x-axis labels and corresponding boxplots, both the **"Low"** and **"High"** parental involvement groups contain data for students who report sleeping 5 hours.
- **Sleep amount with lowest median attendance (High parental involvement):** In the "High" parental involvement group (color), compare the median lines for each sleep duration. The lowest median attendance rate occurs at **9 hours** of sleep.

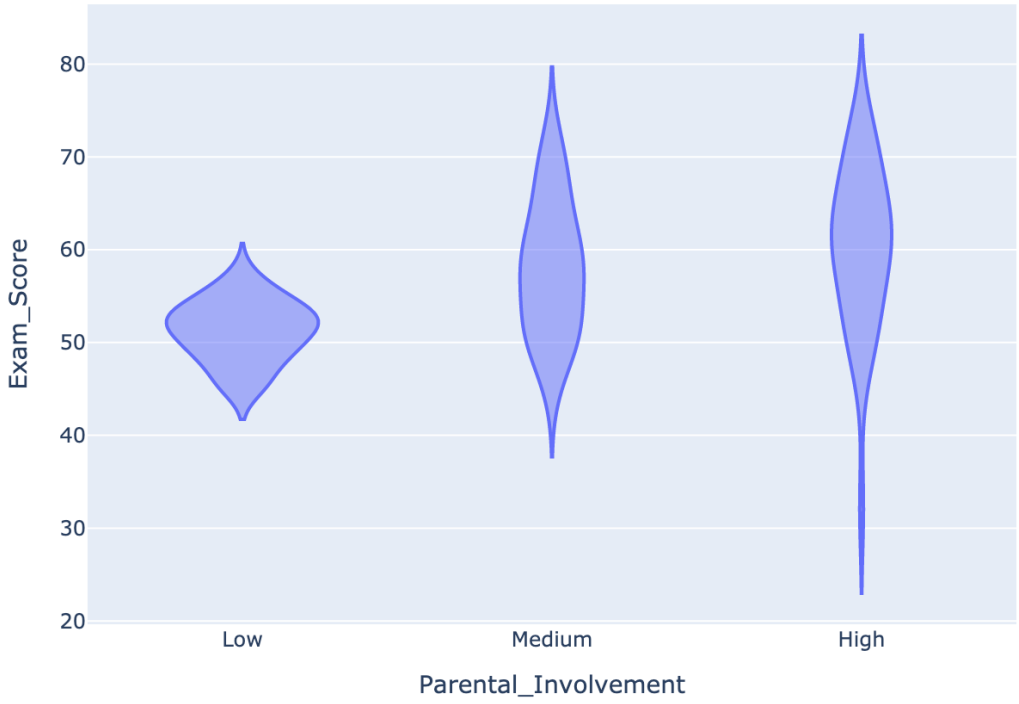

Figure 3: Violin plot

- Which parental involvement level shows the most spread in exam scores, and how can you tell from the shape of the violin plot?
- In which parental involvement category are exam scores most concentrated around a single value, and how does the shape (width) of the violin plot help you see this?
- Do you observe a possible relationship between parental involvement and exam score? Explain briefly.

Solutions and Explanation

- **Greatest spread of exam scores:** The **"High"** parental involvement group shows the widest spread in exam scores. This is evident from the vertical extent of the violin plot, which ranges approximately from **22 to 81**. The overall height of the violin indicates the range of the data, and this group spans the largest interval.
- **Most concentrated scores around a single value:** The **"Low"** parental involvement group has scores most concentrated around a specific value. This is shown by the violin plot being widest near a score of about **51**, indicating that many students have scores near this value. Additionally, the narrower overall shape of this violin suggests less variability.

- **Relationship between parental involvement and exam scores:** Based on visual inspection, there appears to be a **positive correlation** between parental involvement and exam scores. As parental involvement increases from low to high:
  - The **median** exam score increases.
  - The **spread** of scores becomes larger, indicating both higher maximums and lower minimums.
  - This could suggest that greater parental involvement is associated with both improved performance and more variation in outcomes.

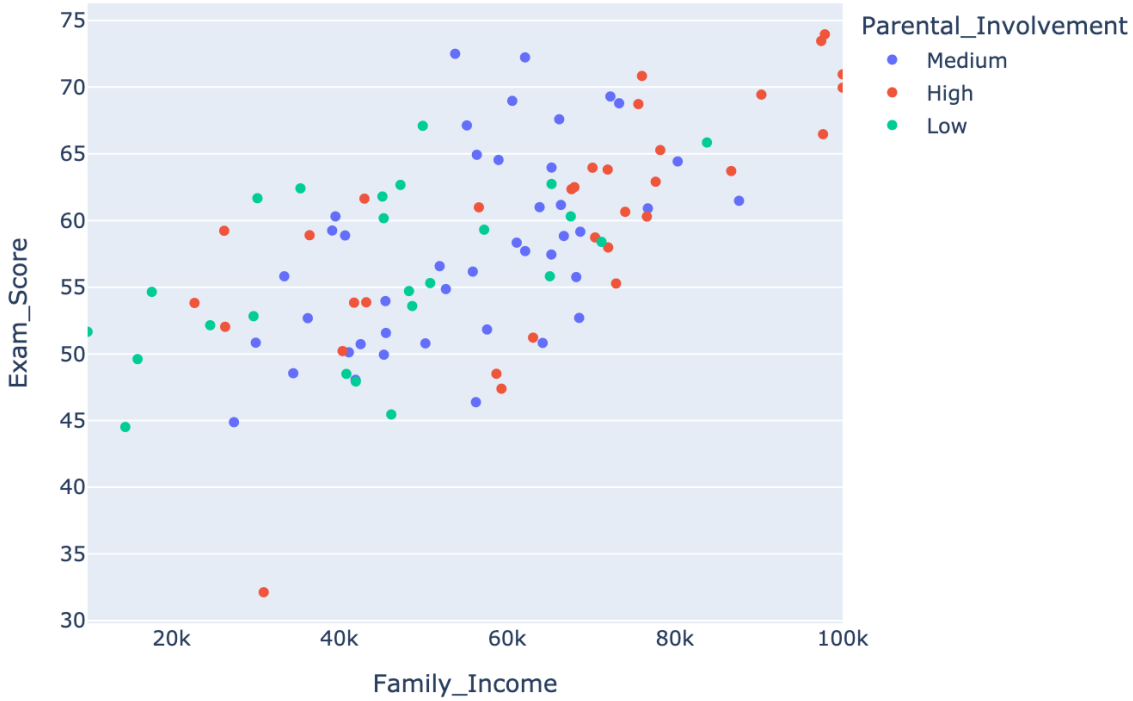

Figure 4: Scatter plot

- What is the highest exam score (approximately) achieved by a student whose parents earn between 20k and 40k?
- What is the highest exam score (approximately) achieved by a student whose parents earn between 80k and 100k, where parental involvement is low?
- Can you suggest a model that could explain the relationship between family income and exam score?

Solutions and Explanation

- **Highest exam score for family income between 20k and 40k:** To determine this, look at the horizontal (x-axis) interval between 20k and 40k. Then identify the point within this range that lies highest on the y-axis (exam score). The highest score in this range is approximately **63**.
- **Highest exam score for income between $80k and $100k with low parental involvement:** Focus on points in the 80k–100k income range and filter only those marked in green (representing **low parental involvement**). Among these, the highest y-axis value corresponds to an exam score of approximately **65.5**.
- **Suggested model to explain the relationship:** From the scatter plot, there appears to be a **positive trend** between family income and exam scores, students from higher-income families tend to score higher on exams. A reasonable starting point for modeling this relationship would be a **simple linear regression**.

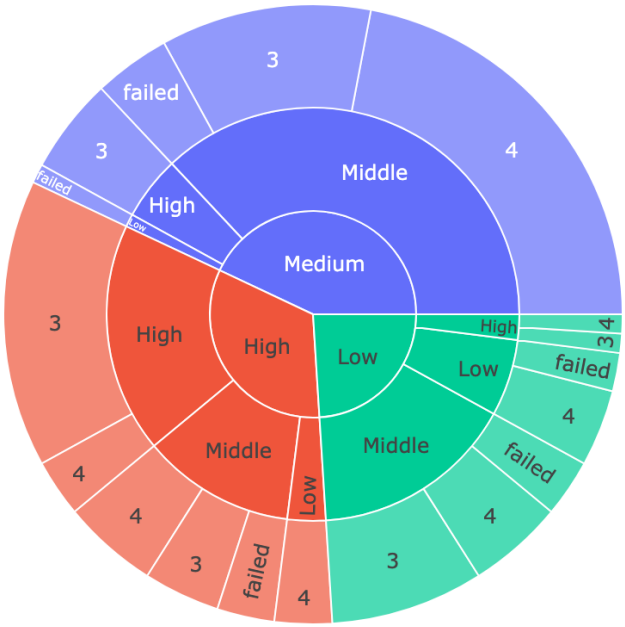

Figure 5: Sunburst chart: Parental_Involvement → Income_Level → Grade

- Which is the most common parental involvement level?
- Which is the most common grade per income level for medium parental involvement?

- List all combinations of parental involvement and income level where students failed the exam.

Solutions and Explanation

- **Most common parental involvement level:** In a sunburst chart, the innermost ring represents the top-level category—in this case, parental involvement. The size of each sector reflects the number of students in that category. The largest central segment corresponds to **medium** parental involvement, making it the most common.
- **Most common grade per income level for medium parental involvement:** The second ring from the center shows income level nested within each parental involvement category. The outermost segments then represent student grades. For students with **medium** parental involvement, the most frequent income group is **middle income**, and the largest grade sector within this group is the **"4"** grade.
- **Combinations of parental involvement and income level where students failed the exam:** A failing grade is labeled as **"failed"** in the chart. By visually identifying all outer segments labeled "F," and tracing their path inward through income level and parental involvement, we find the following combinations:
  - **Medium, Middle**
  - **Medium, Low**
  - **Low, Middle**
  - **Low, Low**
  - **High, Middle**