Data Science
Term Test 1
Mr. Merrick
December 11, 2025
Time Limit: 70 Minutes

Name (Print): ____________________

This exam contains 6 pages (including this cover page) and 11 problems.

You may use your calculator and R (with the provided datasets) for this exam.

You may not use the internet or AI tools.

You will be provided with the following datasets (as CSV files in R):

- `FastFood.csv`
- `pokemon.csv`
- `steam.csv`
- `lego.csv`

No other datasets may be used.
You are required to show your work on each problem on this exam. The following rules apply:

- Organize your work in a reasonably neat and coherent way in the space provided. Work scattered all over the page without a clear ordering will receive very little credit.

- Mysterious or unsupported answers will not receive full credit. A correct answer, unsupported by calculations, explanation, or code, may receive little or no credit; an incorrect answer supported by substantially correct reasoning might still receive partial credit.

Do not write in the table to the right.

Problem	Points	Score
1	8	
2	6	
3	10	
4	10	
5	8	
6	8	
7	4	
8	4	
9	4	
10	4	
11	1	
Total:	67	

1. (8 points) For each variable below, identify:
   - whether it is categorical or quantitative, and
   - if it is quantitative, whether it is discrete or continuous.

   Give a short explanation (one phrase or sentence) for each.

   (a) `hours` (Steam)

   (b) `is_legendary` (Pokémon)

   (c) `pieces` (LEGO)

   (d) `restaurant` (FastFood)

   (e) `weight_kg` (Pokémon)

   (f) `sodium` (FastFood)

2. (6 points) For each situation, choose the most appropriate plot type from the list and briefly explain your choice.

   Plot types: histogram, boxplot, scatterplot, bar chart.

   (a) You want to show the distribution of `hours` from the Steam dataset.

   (b) You want to compare calorie distributions between restaurants in the FastFood dataset.

   (c) You want to explore the relationship between height and weight of Pokémon.

   (d) You want to show how many LEGO sets come from each theme (e.g., Star Wars, City).

3. (10 points) A sample of calories from the FastFood dataset (9 items) is shown below:

   $$480,\ 510,\ 580,\ 610,\ 640,\ 660,\ 680,\ 740,\ 1120.$$

   (a) Compute the mean calories.

   (b) State the median calories.

   (c) Find the five-number summary: minimum, Q_1, median, Q_3, maximum.

   (d) Compute the interquartile range (IQR).

   (e) Using the $1.5\times$IQR rule, determine whether there are any potential outliers. If so, list them and explain why.

4. (10 points) Assume you have already loaded each dataset into R with appropriate names: `steam`, `pokemon`, and `lego`. Write full `ggplot2` commands (no need to write `library(ggplot2)`). You may run and check your code in R.

   (a) A histogram of `hours` from the Steam dataset with 30 bins.

   (b) A boxplot comparing `total` Pokémon stats for Legendary vs non-Legendary Pokémon. (You may assume there is a variable `is_legendary_factored` with levels `"No"` and `"Yes"`.)

   (c) A scatterplot of LEGO `pieces` (x-axis) vs `usd_msrp` (y-axis), including a smooth curve (`geom_smooth(method = "lm")`) on top of the points.

5. (8 points) A health researcher is studying whether smoking is associated with heart disease. They collect data on 200 people and classify each person as a smoker or non-smoker, and whether or not they have been diagnosed with heart disease. The results are summarized in the table below.

	Heart Disease = Yes	Heart Disease = No	Total
Smoker	20	60	80
Non-smoker	12	108	120
Total	32	168	200

   One person is chosen at random from these 200 people.

   (a) Find P(Heart Disease). (Write your answer as a fraction and/or decimal.)

   (b) Find P(Smoker).

   (c) Find P(Heart Disease AND Smoker).

   (d) Find P(Heart Disease | Smoker). (That is, the probability a person has heart disease given that they are a smoker.)

   (e) Use your answers to decide whether smoking and heart disease appear to be independent in this group. Justify your answer by comparing P(Heart Disease AND Smoker) and $P(\text{Heart Disease}) \cdot P(\text{Smoker})$.

6. (8 points) In the LEGO assignment, you made a scatterplot of `pieces` vs `usd_msrp` (price). Recreate this scatterplot in R using the `lego` dataset, and then answer the following.

   (a) Describe the overall relationship between the number of pieces and price. (Positive or negative? Weak, moderate, or strong?)

   (b) Using the general trend you saw, give a reasonable estimate for the price of a LEGO set with 500 pieces. Explain how you decided.

   (c) Based on the pattern in the scatterplot, would you expect a LEGO set with 1500 pieces to cost more or less than $100? Explain your reasoning in a sentence or two.

7. (4 points) In class, you saw a histogram of Attack values for Fire-type Pokémon. Recreate a histogram of Attack for Fire-type Pokémon in R using the `pokemon` dataset.

   Use the shape and spread of that histogram to answer the following questions:

   (a) Would it be likely to observe a Fire-type Pokémon with an Attack value of 150? Explain in one sentence using the idea of where most values lie.

   (b) Would an Attack value of 40 be unusually low for Fire-type Pokémon? Explain your reasoning based on the histogram.

8. (4 points) In the Pokémon assignment, you explored how total stats and Legendary status vary by type. You may use R and the `pokemon` dataset during this question.

   (a) Based on the data and plots you saw, is there an association between a Pokémon's primary type (`type1`) and whether it is Legendary? Circle one:

   Strong association   Some association   Little or no association

   (b) Explain your choice in 2–3 sentences. Hint: think about which types (e.g., Dragon, Psychic) have many Legendaries and which types (e.g., Normal, Bug) have very few or none.

9. (4 points) You also looked at boxplots of total stats by Pokémon type.

   (a) According to those boxplots, which type or types tend to have the highest total stats on average? (Write one or two types based on what you saw.)

   (b) If you randomly choose a Pokémon with total stats above 550, is it more likely to be Dragon-type or Bug-type? Explain your answer briefly, using what you remember about the distributions.

10. (4 points) In the assignment, you created a scatterplot of Pokémon height (x-axis) vs weight (y-axis). Recreate this scatterplot in R using the `pokemon` dataset.

    (a) Based on that scatterplot, would a Pokémon that is 8 meters tall be expected to weigh more than 300 kg? Explain your reasoning using the trend in the data.

    (b) If two Pokémon have very similar height, do they usually have similar weight? Explain in a sentence or two using how tightly (or loosely) the points were clustered around the trend.

11. (1 point) From the FastFood dataset, which restaurant appears to have the highest median calories-to-fat ratio (calories divided by total fat)? Use R to compute this.