Practice Quiz 1: EDA & Data Munging

MSE 125 — Lectures 1–3

Use this practice quiz to prepare for Quiz 1 (Wednesday, April 8). The real quiz will have 2 questions in 10 minutes, closed-book. This practice set covers the full range of topics from Lectures 1–3.

Question 1. A dataset of home sale prices has mean = \$320,000 and median = \$185,000.

(a) What does this gap tell you about the shape of the distribution?

(b) A real estate agent advises a seller: "The average home in this area sells for \$320K." Is this helpful advice? What would you say instead?

Solution

(a) The distribution is right-skewed (a long tail of expensive homes). A small number of high-priced sales pull the mean well above the median.

(b) Not helpful — \$320K overstates what a typical home sells for. The median (\$185K) better represents the typical sale. The seller should look at the median and the price range for comparable homes (similar size, location, condition), not the overall mean.

Question 2. A student loads a CSV of employee records and runs `df.info()`:

```
Column          Non-Null Count  Dtype
------          --------------  -----
employee_id     5000            int64
name            5000            object
department      5000            object
salary          5000            object
start_date      4812            object
```

The student is surprised that `salary` has dtype `object` instead of `float64`, even though it has no missing values.

(a) What is the most likely explanation?

(b) The student runs `pd.to_numeric(df['salary'], errors='coerce')` and finds that 342 values become `NaN`. Should they just drop those rows? What should they do first?

Solution

(a) The salary column contains non-numeric characters — likely dollar signs (`$42,000`), commas, or text entries like "Contractor" or "N/A." Even one non-numeric value forces pandas to store the whole column as `object`.

(b) Don't drop yet. First, inspect the non-numeric values with `df[pd.to_numeric(df['salary'], errors='coerce').isna()]['salary'].value_counts()` to see what they actually are. They might be fixable (strip `$` and commas), or they might be informative (e.g., "Contractor" indicates a different employment type). Dropping without inspecting risks silent data loss.

Question 3. An analyst compares average Airbnb prices across NYC boroughs and concludes that Manhattan is the most expensive place to stay.

(a) Name a confounder that could explain at least part of this difference.

(b) How would you adjust the comparison to account for this confounder?

Solution

(a) Room type. Manhattan has a higher proportion of entire home/apartment listings compared to other boroughs. Entire homes are more expensive than private or shared rooms regardless of borough, so Manhattan's higher average partly reflects its listing mix.

(b) Compare prices *within* each room type across boroughs (e.g., compare entire-home prices in Manhattan vs. Brooklyn). This can be done with a grouped summary (`groupby(['neighbourhood_group', 'room_type']).median()`) or with faceted box plots.

Question 4. A student joins institution-level data (7,000 rows) to program-level data (120,000 rows) on `UNITID`:

```
Rows before join: 7,000
Rows after join:  98,412
```

The student says: "The join created 91,000 extra rows — something went wrong."

(a) Did something go wrong? Explain why the result has more rows than either input table.

(b) After the join, the student computes `merged['UGDS'].mean()` (average undergraduate enrollment). Why might this number be misleading?

Solution

(a) Nothing went wrong. This is a one-to-many join: each institution matches multiple programs, so the institution row is replicated for each match. A university with 50 programs appears 50 times. The result has fewer rows than the programs table because

some programs come from institutions not in the scorecard.

(b) Universities with more programs get counted more times. A large university with 80 programs contributes 80 copies of its enrollment to the mean, while a small school with 3 programs contributes only 3. The mean is weighted by program count, not treating each institution equally.

Question 5. A hospital's patient-tracking system records blood oxygen levels (SpO2, normally 95–100%). The dataset contains: 97, 99, -1, 96, 98, -1, 95, 99, -1, 97.

(a) Three readings are recorded as -1. What is the likely meaning, and what is the general term for this kind of encoding?

(b) An analyst reports "average SpO2 = 68.0%." A nurse flags this as impossibly low. What went wrong, and what should the analyst have done?

Solution

(a) -1 is a sentinel value — a special number used to indicate missing or invalid data (e.g., sensor disconnected, probe fell off). It is not a real oxygen reading.

(b) The analyst computed the mean without replacing the sentinel values: (97 + 99 + (-1) + 96 + 98 + (-1) + 95 + 99 + (-1) + 97) / 10 = 678 / 10 = 67.8, which rounds to 68.0%. The three -1 values dragged the average far below the true range. The analyst should first replace -1 with `NaN` and then compute the mean of the remaining 7 values: (97 + 99 + 96 + 98 + 95 + 99 + 97) / 7 = 681 / 7 = 97.3%.

Question 6. A data scientist at a ride-sharing company computes:

```
print(f"Average driver rating: {df['driver_rating'].mean():.2f}")
```

Output: `Average driver rating: 4.71`

The minimum possible rating is 1 and the maximum is 5. The company concludes: "Our drivers are excellent — the average rating is 4.71 out of 5."

(a) Why might 4.71 overstate true rider satisfaction?

(b) Is this an example of MCAR, MAR, or MNAR missingness? Explain.

Solution

(a) Most riders don't leave a rating at all — they only rate when they had a notably good or bad experience. Satisfied-but-unremarkable rides are underrepresented. Since very positive experiences are more common than very negative ones (most rides are fine), the rated subset skews high.

(b) MNAR. The decision to rate depends on the rating itself: riders with strong opinions (very happy or very unhappy) are more likely to submit a rating. The missing ratings are not random — they come disproportionately from riders whose experience was middling.

Question 7. An online tutoring company wants to show that students who use their platform get better grades. They compare 500 students who signed up for tutoring to 500 who did not. The tutoring group's average GPA is 0.3 points higher.

(a) Name the most likely confounder and explain how it creates a misleading association.

(b) A skeptical analyst points out that the company only has grade data for students who completed the course. 15% of the non-tutoring group dropped out vs. 5% of the tutoring group. What type of missingness is this, and in which direction does it bias the GPA comparison?

Solution

(a) Motivation (or academic engagement). Students who voluntarily seek tutoring are likely more motivated, better-organized, or more concerned about their grades than those who don't sign up. This motivation independently improves grades, so the 0.3-point gap reflects the *type of student* who signs up, not (only) the effect of tutoring.

(b) MNAR — the students who dropped out likely had lower grades, and their grades are missing precisely because they left. Since more non-tutoring students dropped out, the surviving non-tutoring group is artificially stronger than the full group would have been. This makes the two groups look *more similar* than they really are, so the true GPA gap (with dropouts included) would likely be *larger*. However, this larger gap still cannot be attributed to tutoring because of the confounding from part (a).

Question 8. A city's open-data portal lists restaurant health inspections. Two entries appear:

Name	Address	Violation count
TACO PALACE	412 Main St	3
Taco Palace LLC	412 Main Street	1

(a) Are these the same restaurant or two different restaurants? What information would help you decide?

(b) If you merge them, the combined violation count is 4. If you keep them separate, each appears to have fewer violations. Why is this decision a value judgment, and what are the risks of each choice?

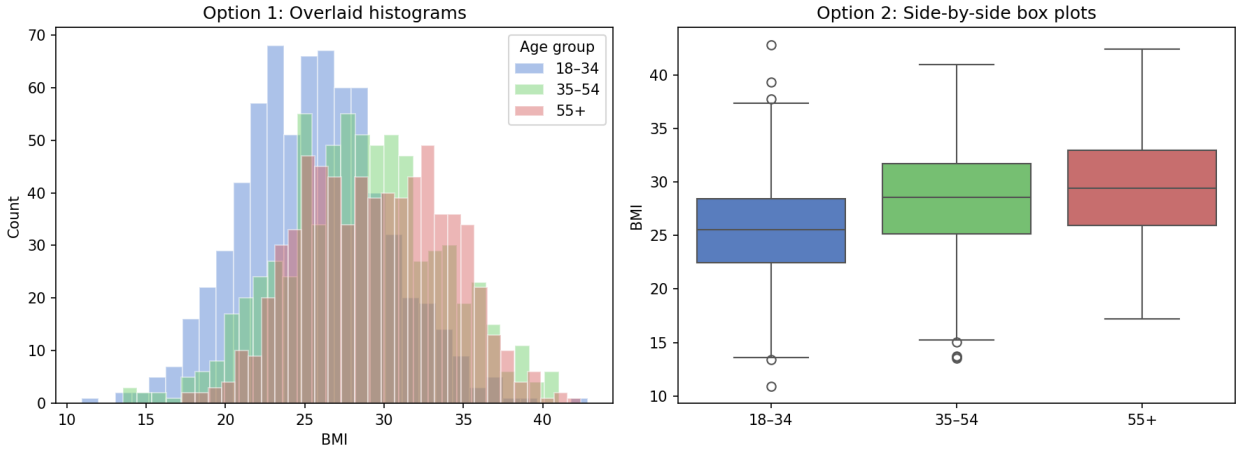
Solution

(a) They are likely the same restaurant — same address (with minor formatting differences) and similar name (capitalization and "LLC" suffix differ). To confirm, you could check: business license numbers, inspection dates (are they interleaved or non-overlapping?), phone number, or owner name. Without a unique identifier, this is ambiguous.

(b) It is a value judgment because reasonable analysts could disagree. Merging risks combining records that are actually different businesses (a false positive), inflating violation counts for one restaurant. Keeping them separate risks undercounting violations for a single

restaurant (a false negative), making it appear safer than it is. The right choice depends on the downstream decision: if the city is flagging dangerous restaurants for closure, a false negative (missing violations) is more harmful; if the city is revoking licenses, a false positive (wrong restaurant penalized) is more harmful.

Question 9. A public health researcher has BMI data for 2,000 adults, split into three age groups (18–34, 35–54, 55+). She wants to compare the distributions across groups and produces two plots of the same data:



(a) Name one advantage of the box plot (Option 2) over the overlaid histogram (Option 1) for comparing these three groups.

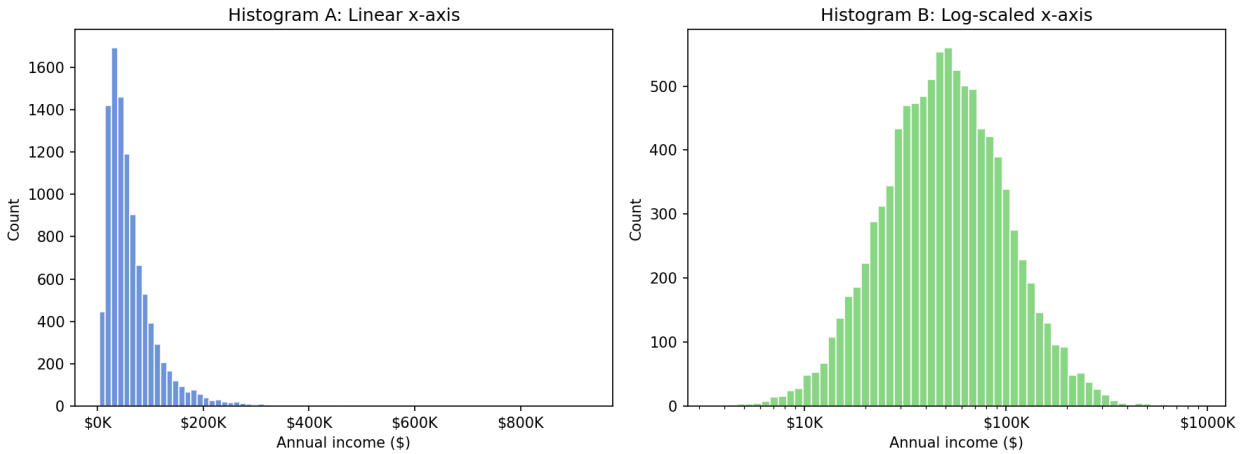
(b) Look carefully at the 55+ group in the histogram. What feature of its distribution does the histogram reveal that the box plot hides?

Solution

(a) The box plots make it easy to compare medians and spreads across groups at a glance. In the overlaid histogram, the three distributions overlap and obscure each other — it is hard to tell where one group's bars end and another's begin, especially in the middle of the range.

(b) The 55+ group (red) is bimodal — it has two peaks, one around BMI 26 and another around BMI 33. The box plot compresses this into a single box with a wider IQR, completely hiding the two-cluster structure. A researcher relying only on the box plot would miss that the 55+ group contains two distinct subpopulations (e.g., healthy-weight and obese adults).

Question 10. Two histograms show the same income data for 10,000 households. Histogram A uses a linear x-axis; Histogram B uses a log-scaled x-axis.



(a) Histogram A shows a sharp spike near \$0 with a long right tail. Histogram B looks roughly symmetric and bell-shaped. What does the symmetric shape on the log scale tell you about the distribution?

(b) A policymaker reads Histogram B and says: "Income is roughly symmetric — there are about as many high earners as low earners." Is this a valid conclusion? What is the policymaker missing?

Solution

(a) The data is approximately lognormal: the logarithm of income is roughly normally distributed. Multiplicative processes (raises, investment returns, compounding) tend to produce lognormal distributions. The log scale spreads out the low values and compresses the high values, revealing a symmetric shape that was hidden by the extreme right skew on the linear scale.

(b) Not valid. The symmetry is in log-dollars, not dollars. On the log scale, "\$20K below the median" and "\$20K above the median" correspond to very different absolute distances. Histogram A shows the true shape: a sharp concentration of low earners and a long right tail extending past \$200K. A household at the 90th percentile earns many times more than a household at the 10th percentile, even though they are equally far from the center in Histogram B. The log scale compresses high values, making extreme inequality look symmetric.