This file is a mapping of key questions, to appropriate analytic approaches to addressing those questions (with distractor options). The distractor options + appropriate options represent key concepts for analysis. The distractors represent common errors in analysis grounded in either statistical and visualisation misconceptions, or/and in taking default settings in chart selection (particularly in excel).

The two first columns are from an earlier version of this task based on weather data. That version combined too many concerns, including data cleaning issues. While those are of course important, it didn’t address the target construct (statistical literacy) clearly enough, and some students struggled.

This version is much cleaner, the data is tidy, and the errors in analysis/representation are based on the selection of those analyses, rather than on issues in the underlying data. I have also mapped distractors to questions, and one approach to reframe this task would be to present a choice of the 3 blocks for each question, rather than presenting all the blocks together (i.e., limit the selection per-question).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Block:** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** | **17** |
| **Q-cor** |  |  | 5 |  |  | 3 |  | 1 |  |  |  | 4 | 2 |  |  | 6 |  |
| **Q-dis** | 3 | 2 |  | 5 | 3 |  | 1 |  | 5 | 5 | 4 |  |  | 4 | 6 |  | 1 |

Block 16 includes some extra information it’d be good to display when they get the right answer. Check the learnr notes for how to do this

grade\_code(correct = XXXXX is where I can put that

grade\_code(incorrect = if they use

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Question in the weather version** | **Chunk / data in the weather version** | **Key concept** | **Rephrase question** | **Data** |
| Q1. What is the relationship between dew point and humidity? | 8: Scatterplot | Relationships | What is the relationship between retweets and likes? | Retweets and likes scatter+cor Block 8  **DISTRACTORS:** 7 (line chart over time)  **DISTRACTORS:** total likes by whether tweeted or not (block 17) |
| Q2. Where do the fastest winds flow in Sydney? | 13: Or (actually not v nice) boxplots for the same | Defining terms.  Fastest as in the maximum, fastest as in Mean, or fastest as in Median? | Which source has the highest sentiment? | Boxplots, and summary  Most as in max? Or as in mean? Or Median?  Block 13  **DISTRACTORS:** Max  **DISTRACTORS:** Bar chart of averages (block 2) |
| Q3. How does temperature compare in melbourne vs sydney? | 2, 4: Comparative histogram (by percent)  6: Side-by-side boxplots | 2-factor comparison | How do iphone vs android compare in terms of number of words? | Boxplot and hist (block 6)  **DISTRACTORS:** Bar chart of total words (no baseline control)  Block 5  Block 1 (unnormalised hist) |
| Q5. What weather conditions are seen in melbourne and sydney? | 12: Or, a count of a similarly stupid number of conditions | Too many things to meaningfully compare | Q4. What NRC sentiments do we see in each source ? | Block 12 (stacked proportions)  **DISTRACTORS:** Stacked bars without normalisation: 11. And 14: Side-by-side bars for too many categories  Stacked bars (raw n only) – this is a good case for using a split |
| Q4. What kind of relationship is there between visibility and humidity? | Scatterplot again | Association vs relationship  Levels of data (treating mood/emotion as interval)  “pick your data”/don’t show everything | Q5. How are NRC sentiment, source, and RTs associated? | **Block 3**  **Distractor** here is a scatterplot (NRC is at best ordinal data) – in 9  Boxplots in block 10  There’s also a random heatmap (block 4) |
| ~~Q6. How does humidity vary in melbourne vs sydney across months?~~ | ~~10~~ | ~~Temporal change (not just the average, so what should you plot?)~~ |  | ~~SCRAP THESE~~  ~~But keep them at the bottom in case I come back to it~~ |
| ~~Q7. What weather condition has the lowest humidity?~~ | ~~11~~ | ~~Minimum and comparison~~ |  |  |
| Q6 |  | Form over function  (wordclouds) | What kind of language is used in the two main sources? | Block 16 is correct(ish)  Block 15 distractor |
| **DISTRACTORS** | 3: Comparative histogram (by count)  5: A dumb 2 column bar chart for the mean (but it doesn't demo what I want)  7: Two variables on a line chart over time (where we care about relationship)  9: Summary table by a stupid number of ordinal categories (clear, drizzle, etc.) |  |  | trump\_tweets$source |