Data Analysis Exercise

Please answer the questions below to the best of your abilities. Please submit your answers as a written document. Please submit the document as a PDF file. The document should not exceed two pages of text. Please include all relevant tables and figures used in answering the questions, though these will not count towards the maximum page count. Please also provide any code you may have used to solve the exercise.

Data

A google drive folder has been shared with a data set containing the UK Cyber Security Breaches Survey from 2017

- 2017_microdata.csv contains the raw data.
- 2017_labels.xlsx contains a mapping between variable names and survey questions.
- Cyber_Security_Breaches_Survey_2017_annex_PUBLIC.pdf contains study appendix with more on questions and survey methodology.
  - Purely optional document for this exercise.

Task

- Produce summary table of i) cyber sec investment as a share of revenue, and ii) number of incidents, broken down by the following revenue bands:
  - Micro: <$10M
  - Small: $10 to $250M (not included)
  - Medium: $250M to $1B (not included)
  - Large: $\geq$ $1B
- Evaluate the following statement based on the data:

  Medium sized companies spend less on cyber security because they have fewer breaches.
- Comment (2-3 lines) on challenges in evaluating the claim using this data.

Hints

- Assume 1pound = 1.3 USD
- Values encoded as "-1" means "Refused" or "Don't know"
- *investn* has numeric (imputed) cyber sec investment values.
- Use average (numeric) value of *salesa* within revenue bands to impute categorical answers in *salesb* before converting to USD.
- Is there a better method?

  [don't do it, but do write 1–2 lines if you believe there is]
- *numba* and *numbb* asks about number of "breaches or attacks" experienced in past 12 months.