Answers

# Easy – Restaurant

* Top 10 restaurants have ‘Yes’ for Franchise, everything else is ‘No’
* The Sales column is the Units column reversed (when the dataset is ordered by sales)
* Unit Volume = (51 – Rank)\*100 (if they put that they are perfectly correlated that’s fine)
* New York, N.Y. is changed to New York, Eagles
* Restaurants with a rank divisible by 5 are surrounded in quotation marks
* Any restaurant with a YOY\_Sales percent of over 50 has ‘- mmm tasty’ added to the name

# Easy – Movies

* All movies made later than 2005 are given an IMDB score of 0
* There is a maximum of 1250000 votes for a movie
* All movies have an R rating
* All movie names with more than 3 words are reversed

e.g. (The Silence of the Lambs -> Lambs the of Silence The)

* Run times under 2 hours have in quotation marks. Extra point if they realise the other run times are integers, not strings

# Medium – Laptops

* All laptops with Windows 11 OS are rated 0
* All laptops with no special feature are a shade of red
* The price is normally distributed with a mean of 1000 and a standard deviation of 250
* Screen size of 11 inches is changed to 11 bananas

# Medium – Social Media

* Telegram reviews are all 5 stars
* Facebook review dates are all from '1945-04-30 18:00:00'
* 2% of IDs are made up words - extra point if they get the probabilities within 1%
* Viber reviews have no names
* userLang 'EN' corresponds to French, and userLang 'FR' corresponds to German, and userLang 'DE' corresponds to English

# Hard – Kindle Books

* All books by Matt Schifferle have Rick Astley as their imgUrl (book cover)
* The stars a book has is equal to its price/4 (rounded to 1 dp). Any stars bigger than 5.0 are set at 5.0
* product\_reference is an entirely made up column, made of hex colours
* The category\_id column is a product of 2^bool (false=0,true=1) for all the Boolean columns (isKindleUnlimited, isBestSeller, isEditorsPick, isGoodReadsChoice).

E.g. False, False, False, True = 1 \* 1 \* 1 \* 2 = 2

E.g. True, False, True, False = 2 \* 1 \* 2 \* 1 = 4

# Hard – Tweets

* The number of likes for Jeffrey Dahmer tweets follows a perfect bell curve
* Will Smith is referred to as Will False in all tweets with the keyword ‘Will Smith’
* 1000 of the same fake Osama Bin Laden tweets have been added
* Some likes have negative values
* Some rows have corrupted data

# Hard – Marketing

* All households with an income over 10000 spent 1300 on wine on Halloween 2013
* Everyone who spent more than 1200 on wine is divorced
* Any Deals purchased more than 10 is changed to 255
* If a household has more than one kid, they spent 68 on fruit
* If a household has basic education, they complained (complain = 1)
* The mnt\_total column in wrong
* The marital status row has misspelt words and non-capitalised words (10% chance for both) – extra point if they get the probabilities within 1%