# MIDTERM 1

## PART 1: BASIC DATA CLEANING

1. Read the listings.csv file

2. Clean the “available” column: replaces all “y” and “Y” by “yes” and all “n” and “NO” BY ‘no’.

3. Removes all non-digit characters from the num_rooms column.

## PART 2: STRING MANIPULATION

Data: hacker_news.csv

1. Use a regex pattern and the ignorecase flag to count the number of mentions of SQL in the title column.

2. Extract the mentions of different SQL flavors into a new column and clean those duplicates by making them all lowercase

- Create a new dataframe named hn_sql, including only rows that mention a SQL flavor

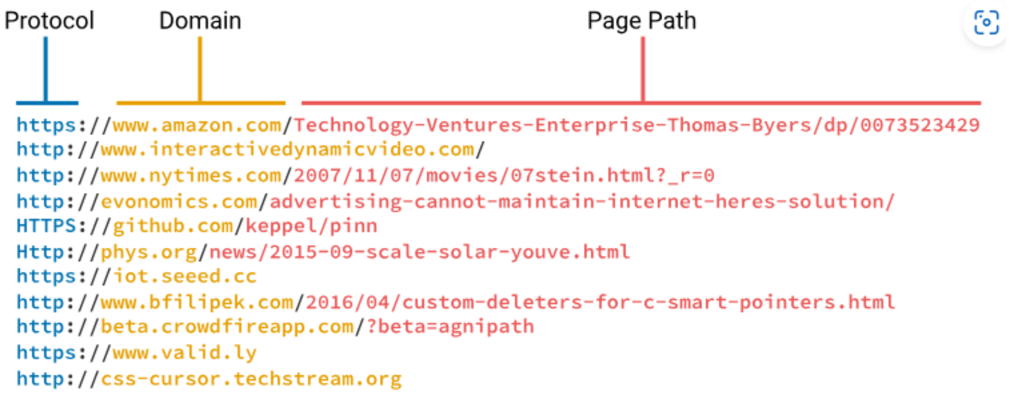
- Create a new column called “flavor” in the hn_sql dataframe, containing extracted mentions of SQL flavors, defined as:

  - Anytime ‘SQL’ is preceded by one or more word characters
  - Ingoring all case variation

- Use the Series.str.lower() method to clean the values in the flavor columns by converting them to lowercase. Assign the values back to the column in hn_sql

- Use the DataFrame.pivot_tables() method to create a pivot table, sql_pivot:

  - The index of the pivot table should be the flavor column
  - The values of the pivot tables should be the mean of the num_comments column, aggregated by SQL flavor.



```
https://www.amazon.com/Technology-Ventures-Enterprise-Thomas-Byers/dp/0073523429
http://www.interactivedynamicvideo.com/
http://www.nytimes.com/2007/11/07/movies/07stein.html?_r=0
http://evonomics.com/advertising-cannot-maintain-internet-heres-solution/
HTTPS://github.com/keppel/pinn
Http://phys.org/news/2015-09-scale-solar-youve.html
https://iot.seeed.cc
http://www.bfilipek.com/2016/04/custom-deleters-for-c-smart-pointers.html
http://beta.crowdfireapp.com/?beta=agnipath
https://www.valid.ly
http://css-cursor.techstream.org
```

3. Extracting Domains from URLs

- Use the regular expression pattern to extract the URL components from the url column of the hn dataframe. Assign the results to url_parts. Add names to each capture group:

- The first capture group should be call “protocol”
- The second capture group should be called “domain”
- The third capture group should be called “path”

**PART 3: GroupBy Operations**

We'll compare two different types of posts from Hacker News: `Ask HN` or `Show HN`.

- Users submit `Ask HN` posts to ask the Hacker News community a specific question: the lowercase version of title starts with “ask hn”.

- Users submit `Show HN` posts to show the Hacker News community a project, product, or just generally something interesting: the lowercase version of title starts with “show hn”:

1. Do `Ask HN` or `Show HN` receive more comments on average?

2. Finding the Number of Ask Posts and Comments by Hour Created?

**PART 4: DUPLICATE DATA**

Revenue dataset
This dataset provides the revenue created by each customer for each month

1. For typing errors, in this dataset, some customers' revenue appears more than once for a specific month. We need to delete one and keep the row with a smaller income.

2. In this dataset, we would like to know the earliest revenue created by each customer and the time. We would remove rows that contain a CustomerID already listed later. Do it using the drop_duplicates method.

**PART 5: MISSING VALUE AND DATA TRANSFORMATION**

READ bank-additional-full.csv FILE

1. Encode each value of the month variable with a corresponding number: We would like to encode the Jan value with the number 0, the Feb value with the number 1…

a. Using the apply function. Assign the result to a new attribute named enc_month1

b. Using the OrdinalEncoder class. Assign the result to a new attribute named enc_month2

Hint: By default, OrdinalEncoder will assign integers to labels in the order that is observed in the data. If a specific order is desired, it can be specified via the categories argument as a list with the rank order of all expected labels

c. Check if the value of two attributes (enc_month1, enc_month2) is equal

2. Using the PowerTransformer class to transform the duration attribute (choose method = ‘yeo-johnson’, standardize = False) and then MinMaxScaler() for this attribute. Assign the result to a new feature named ‘duration_T’.

-