Case Studies Using Numpy, Pandas, matplotlib & seaborn

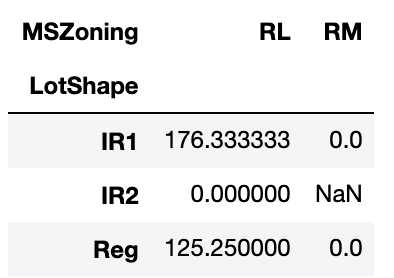
*Case Study 1:*

Dataset Description: The file consists of start-ups investment details.

1. Read the given comma separated values as dataframe (investments.csv)
2. List out all column names.
3. Create a dataframe with numerical columns.
4. Create a dataframe with categorical columns.
5. Get a summary on the data and draw inferences if any.
6. Display duplicate rows.
7. For each column find out the percentage of missing values.
8. Find count of ‘name’ in each ‘country\_code’.
9. What is the percentage of the companies which have status ‘acquired’ ‘operating’?
10. What is the percentage of the companies which have status ‘acquired’ acquired?
11. Filter records having missing values in column ‘year\_founded’.
12. Create a column ‘category\_list\_count’ having count of category lists.
13. Find average funding\_total\_usd for each country\_code.
14. Find total funding\_total\_usd for each country\_code.
15. Find average funding\_total\_usd in each country\_code and region.
16. How many companies have got just 1 round of funding?
17. Perform mapping on status column; acquired -> A, operating -> O and closed -> C.
18. How many companies have ‘debt\_financing’ above zero?
19. Create a column ‘homepage’ to store company name from ‘homepage\_url’: For example: If url is <http://www.waywire.com>, name is waywire.
20. Find the count of companies in each of the markets.
21. Find the count of companies in each of the markets and store the new column ‘cnt\_name’ in the original dataframe.
22. Rename ' funding\_total\_usd ' to 'funding\_total\_usd'
23. Fill missing values in column ‘city’ with ‘other\_city’
24. For each row in column ‘funding\_total\_usd’, calculate actual – average value for each group ‘city’
25. Normalize ‘‘funding\_total\_usd’ at country level.
26. What is the average ‘funding\_total\_usd’ for each city?
27. Plot histogram/distribution of ‘funding\_total\_usd’ and provide insights if any.
28. What is maximum ‘funding\_total\_usd’ for each market status?
29. How many years has it been since each company was founded?
30. Visualize ‘grant’ distribution.
31. Visualize ‘debt\_financing’ distribution.
32. Display proportion of companies status.
33. How many US states are available?
34. create column ‘cmt\_address’ by joining country code, state code, region and city.
35. select columns with underscore in their names.

*Case Study 2:*

Load practice.csv file as a data-frame and perform following operations on the data-frame

1. Display all columns
2. create numerical and categorical columns list
3. display size of the data-frame
4. rename column MSSubClass -> SubClass, MSZoning -> Zones
5. display distinct values for Zoning, LotShape, LotConfig
6. display count of distinct values for Zoning, LotShape, LotConfig
7. max, min of column YearBuilt
8. create a new column “year\_diff’. This will be holding difference of current year and YearBuilt
9. display distinct MSZoning for each OverallQual
10. What is the maximum LotArea where BsmtExposure = Mn?
11. Sort dataframe based on following columns and orders: MSSubClass; ascending, YearBuilt; descending
12. What is average OverallQual.
13. convert column ‘YearBuilt’ into date type.
14. Group by YearBuilt and find maximum OverallQal
15. Load the practice.csv again with MSSubClass as new index
16. Convert LotArea as numpy array
17. In column MasVnrArea replace 0 with -1
18. Check if there is/are any Null values (NaN) in the data given
19. Display percentage of missing values in each column if any
20. Select records where LotConfig is Inside
21. Make a new dataframe with only numeric columns
22. Make a new dataframe with only factorial/string columns
23. Drop column ExterQual
24. Group data on LotShape and find average LotArea
25. Write code to get a pivot table as shown (average of **MSSubClass**)
26. 

Case Study 3:

1. Display details of the songs that have popularity above 80

2. Which is having maximum length, beats per minute?

3. Which is having minimum length, beats per minute?

4. Create new column to store string length

5. Create new column to store first name of each artist

6. How many songs from “dance pop” genre?

7. What is average length for each genre?

8. What is average length for each artist?

9. What is maximum length for each genre?

10. What is average length for each artist?

11. How many songs share same popularity?

12. Rank each song based on popularity.

13. Rank each song based on danceability.

14. Average, Maximum, Minimum, standard deviation of each column

15. Average, Maximum, Minimum, standard deviation of each row

16. Add a new column where each row will represent difference between row value and mean of the each genre group.

17. Take log transform on Loudness DB and add as column

18. Visualize if there is linear correlation between beats per minutes and popularity, danceability and popularity, energy and popularity

19. What is distribution of popularity with respect to energy

20. What is distribution of popularity with respect to beats Per minute

21. Are there any outliers in popularity, danceability, beats per minute?

22. Visual each artist distribution

23. Visual each track name distribution

24. Visual each genre distribution

25. Visual artist, track, genre distribution using subplots