DATA CLEANING

Assignment 1 Question

DATA CLEANING: Assignment 1

Table of Contents

1.Introduction

This assignment will help you to consolidate the concepts learnt in the session.

2.Problem Statement

It happens all the time: someone gives you data containing malformed strings, Python, lists and missing data. How do you tidy it up so you can get on with the analysis?

Take this monstrosity as the DataFrame to use in the following puzzles:

```
df = pd.DataFrame({'From_To': ['LoNDon_paris', 'MAdrid_miLAN',
'londON_StockhOlm',
                    'Budapest_PaRis', 'Brussels_londOn'],
     'FlightNumber': [10045, np.nan, 10065, np.nan, 10085],
     'RecentDelays': [[23, 47], [], [24, 43, 87], [13], [67, 32]],
        'Airline': ['KLM(!)', '<Air France> (12)', '(British Airways. )',
                    '12. Air France', '"Swiss Air"']})
```

1. Some values in the the FlightNumber column are missing. These numbers are meant to increase by 10 with each row so 10055 and 10075 need to be put in place. Fill in these missing numbers and make the column an integer column (instead of a float column).

2. The From_To column would be better as two separate columns! Split each string on the underscore delimiter _ to give a new temporary DataFrame with the correct values. Assign the correct column names to this temporary DataFrame.

3. Notice how the capitalisation of the city names is all mixed up in this temporary DataFrame. Standardise the strings so that only the first letter is uppercase (e.g. "londON" should become "London".)

4. Delete the From_To column from df and attach the temporary DataFrame from the previous questions.

5. In the RecentDelays column, the values have been entered into the DataFrame as a list. We would like each first value in its own column, each

second value in its own column, and so on. If there isn't an Nth value, the value should be NaN.

Expand the Series of lists into a DataFrame named delays, rename the columns delay_1, delay_2, etc. and replace the unwanted RecentDelays column in df with delays.

Note: Solution submitted via github must contain all the source code and output.

3. Expected Output

This assignment consists of 200 marks and has to be submitted in .ipynb format in the upcoming session for evaluation.