Assignment 11

Q1. 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).

Answer:

import pandas as pd

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]],

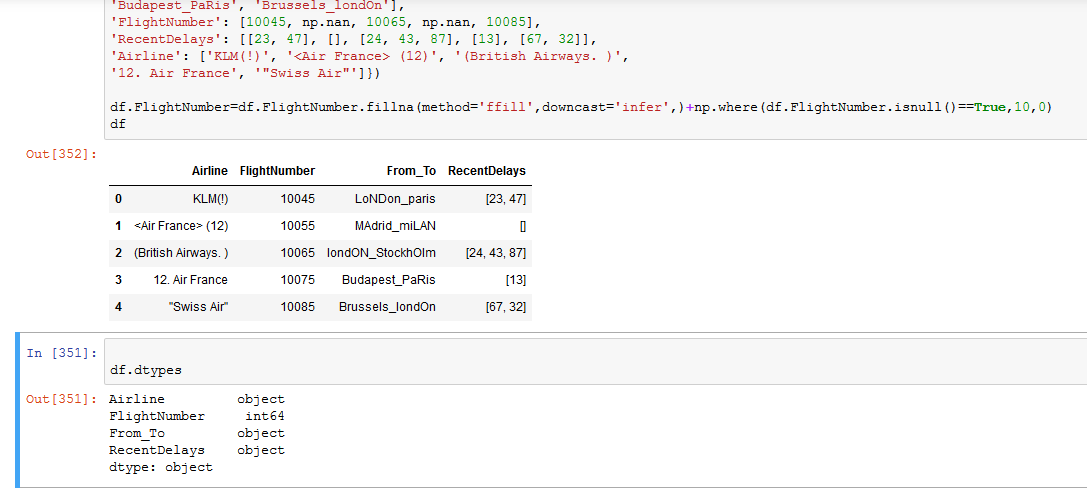
'Airline': ['KLM(!)', '<Air France> (12)', '(British Airways. )',

'12. Air France', '"Swiss Air"']})

df.FlightNumber=df.FlightNumber.fillna(method='ffill',downcast='infer',)+np.where(df.FlightNumber.isnull()==True,10,0)

df

df.dtypes



Q2: 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.

Answer:

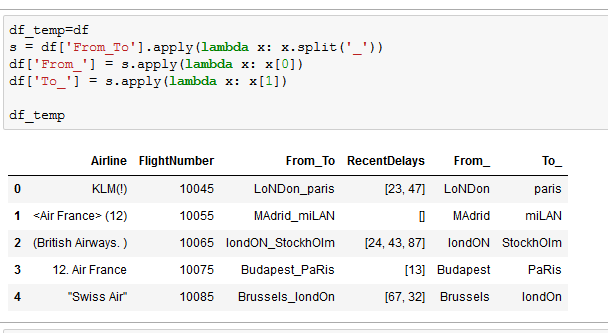
df\_temp=df

s = df['From\_To'].apply(lambda x: x.split('\_'))

df['From\_'] = s.apply(lambda x: x[0])

df['To\_'] = s.apply(lambda x: x[1])

df\_temp



Q3: 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.

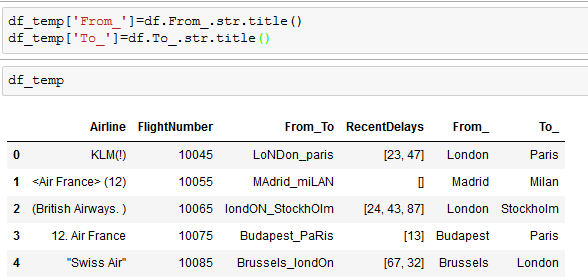
"londON" should become "London".)

Answer:

df\_temp['From\_']=df.From\_.str.title()

df\_temp['To\_']=df.To\_.str.title()

df\_temp



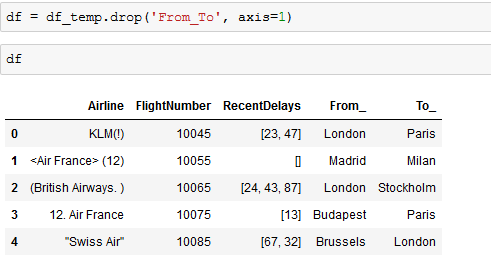
Q4: Delete the From\_To column from df and attach the temporary DataFrame from the

previous questions.

Answer:

df = df\_temp.drop('From\_To', axis=1)

df



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.

Answer:

delays = pd.DataFrame(df['RecentDelays'].values.tolist(), columns=['delay1','delay2','delay3'])

df=pd.concat([df,delays],axis=1, join\_axes=[df.index])

df = df.drop('RecentDelays', axis=1)

df

