MIS 3335 – Data Analytics Using Python

Homework 4 – Data Assembly and Tidy Data (50 points total)

This is an individual assignment and must be completed separately by each student. Group work is not permitted.

► HERE IS WHAT I WANT YOU TO DO:

Write well-commented Python code in a Jupyter Notebook file that uses pandas to combine two datasets into one and then makes that dataset tidy. There are two files for this assignment: **lhr_visit_dom.csv** and **lhr_visit_intl.csv**. These datasets are based upon actual datasets on the characteristics of passengers ending their travels at London's Heathrow Airport.

► HERE IS WHY I WANT YOU TO DO IT:

In addition to the skills we have practiced to this point, this assignment will give you more experience with the following concepts:

- Accounting for and filling in missing values.
- Combining two dataframes to create a single dataframe.
- Saving intermediate results to a csv file.
- Converting multiple value columns into a single variable column (aka, "melting").
- Separating values in a single column representing multiple variables into values in multiple variable columns.

► HERE'S HOW TO DO IT:

This dataset is a small part of the data from a set of passenger surveys conducted annually in airports in the UK. This data was collected at London's Heathrow Airport during 2018 to determine the characteristics of passengers who end their journeys at Heathrow (airport code: LHR). It divides passengers into categories based upon their place of residence (UK or foreign), purpose of their trip (business or leisure), the type of flight upon which they arrived (international or domestic), and the length of their stay. The numbers in the dataset are the counts of passengers at the intersections of those categories. The domestic flight arrival data is in the file **lhr_visit_dom.csv** and the international flight arrival data is in the file **lhr_visit_intl.csv**.

This is one way you might see data like this presented:

Trip length of terminating passengers at Heathrow Airport in 2018.

	Length of stay											
	Up to 12 hrs	Over 12 hrs to 1 day	Over 1 day to 2 days	Over 2 days to 3 days	Over 3 days to 4 days	Over 4 days to 5 days	Over 5 days to 6 days	Over 6 days to 1 week	Over 1 week to 2 weeks	Over 2 weeks to 3 weeks	Over 3 weeks to 4 weeks	Over 4 weeks
International flights												
UK												
Business	128740	292759	802367	1179401	1182390	565233	209144	802846	698202	153495	107981	204405
Leisure	12033	26915	186414	789046	1869780	1294056	566397	3159862	6132141	2634907	1487191	1481942
Foreign												
Business	259642	227083	541518	1036011	1185938	735945	286072	1045448	787222	186617	87751	187868
Leisure	26326	57952	367979	842326	1875647	1343424	582867	3088085	4980963	1665635	1110165	1164338
Domestic flights												
UK												
Business	154420	125309	280907	195927	77236	33447	10645	3955	13284	13708	4522	3564
Leisure	13627	13148	68630	187788	218805	66226	28552	92861	97781	8545	17695	3821
Foreign												
Business	0	0	10095	10764	2335	8324	0	5980	10654	0	0	0
Leisure	4767	0	25372	7687	10537	38694	10233	18896	26359	4767	5898	0

However, the data you will use is in a different format. Here is a view of the domestic flight data in Excel:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	type	Up to 12 hr	Over 12 hrs	Over 1 day	Over 2 days	Over 3 days	Over 4 days	Over 5 days	Over 6 days	Over 1 wee	Over 2 wee	Over 3 wee	Over 4 weeks
2	uk_bus_dom	154420	125309	280907	195927	77236	33447	10645	3955	13284	13708	4522	3564
3	uk_fun_dom	13627	13148	68630	187788	218805	66226	28552	92861	97781	8545	17695	3821
4	for_bus_dom			10095	10764	2335	8324	none	5980	10654			
5	for_fun_dom	4767	none	25372	7687	10537	38694	10233	18896	26359	4767	5898	
6													
7													

The international flight data has the same structure.

Use pandas in the Jupyter Notebook (overall quality =10 pts) to do the following:

1. (2 pts) Read the data into two dataframes with appropriate names (i.e., better than df1 & df2).
   a. (3 pts) Replace missing values (null and "none") in ***both files*** with NaN in the read statements.
2. (5 pts) Combine the two dataframes into a new dataframe called "lhr_visits." Make sure the row index values are reset rather than duplicated.
3. (5 pts) Fill all missing values with a zero value.
4. (5 pts) Save the data in a csv file called "lhr_visits.csv" using the "to_csv" dataframe method. Set the appropriate option to keep from saving the row numbers. The data in the saved file should look like this:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	type	Up to 12 hr	Over 12 hrs	Over 1 day	Over 2 days	Over 3 days	Over 4 days	Over 5 days	Over 6 days	Over 1 wee	Over 2 wee	Over 3 wee	Over 4 weeks
2	uk_bus_dor	154420	125309	280907	195927	77236	33447	10645	3955	13284	13708	4522	3564
3	uk_fun_dor	13627	13148	68630	187788	218805	66226	28552	92861	97781	8545	17695	3821
4	for_bus_do	0	0	10095	10764	2335	8324	0	5980	10654	0	0	0
5	for_fun_do	4767	0	25372	7687	10537	38694	10233	18896	26359	4767	5898	0
6	uk_bus_int	128740	292759	802367	1179401	1182390	565233	209144	802846	698202	153495	107981	204405
7	uk_fun_intl	12033	26915	186414	789046	1869780	1294056	566397	3159862	6132141	2634907	1487191	1481942
8	for_bus_int	259642	227083	541518	1036011	1185938	735945	286072	1045448	787222	186617	87751	187868
9	for_fun_int	26326	57952	367979	842326	1875647	1343424	582867	3088085	4980963	1665635	1110165	1164338

5. (2 pts) Load the data from "lhr_visits.csv" into a dataframe called "visits."
6. (5 pts) Melt the data columns into a column called "visit_length" and a column called "count."
7. (10 pts) Split the values in the "type" column into values for variables called "residence", "purpose", and "flight_type." Add those three columns to the dataframe.
8. (3 pts) Drop the "type" column when you are confident your results are what you were expecting.

Your final dataframe should look something like this:

```
    visit_length      count residence purpose flight_type
0   Up to 12 hrs   154420.0        uk     bus         dom
1   Up to 12 hrs    13627.0        uk     fun         dom
2   Up to 12 hrs        0.0       for     bus         dom
3   Up to 12 hrs     4767.0       for     fun         dom
4   Up to 12 hrs   128740.0        uk     bus        intl
..           ...        ...       ...     ...         ...
91  Over 4 weeks        0.0       for     fun         dom
92  Over 4 weeks   204405.0        uk     bus        intl
93  Over 4 weeks  1481942.0        uk     fun        intl
94  Over 4 weeks   187868.0       for     bus        intl
95  Over 4 weeks  1164338.0       for     fun        intl

[96 rows x 5 columns]
```

► **HERE IS WHAT YOU SHOULDN'T WORRY ABOUT:**

1. You will notice "fun" is used in the table in place of "leisure." Using either word is fine.
2. The order of the columns is not important.
3. None of the values need to be formatted or changed to other data types.