APA - Data Preparation Template

DATA 3300

Name(s):

Read and follow these assignment instructions carefully! Ordinarily, we'd jump into cleansing data as needed and we'd each do it slightly differently. This is difficult to grade, so please go in the order of this document and follow these instructions.

For this assignment students should submit one .ipynb file, one clean, sorted Excel file, and a PDF of this iPython notebook file. See the assignment background for attribute descriptions and the assignment requirements link on the Canvas assignment page.

Students should alter code in lines with "replace code..." language. Otherwise, students do not need to alter the provided code.

Q1

What does the term "data quality" refer to and why is it important within the context of using data to solve business problems?

The degree to which a dataset can be efficiently and effectively processed and used. It is important because it helps create analyses that are meaningful, trustworthy, and that do not induce mayhem.

Q2

The lecture and textbook discussed five characteristics of data quality: accuracy, completeness, consistency, timeliness, and uniqueness. While "accuracy" in this dataset is hard for us to gauge

without further inquiry and"timeliness" kind of depends on what it is the data are to be used for, there are clear examples of problems with each of the other three characteristics.

For each of the other three (completeness, consistency, uniqueness), identify a specific situation within the *hbdata-orig.csv* dataset (identify record numbers when applicable). Hint: For uniqueness, look at the VisitSpan attribute. Be sure your answers are clearly labelled and described.

This can be done by viewing the data in Excel or importing it into Python. If using Python, begin by importing your libraries and then read in the .csv file.

```
import pandas as pd
import numpy as np

#just run this code block after importing libraries above, no need to alter the belo
from datetime import datetime #you'll also need this method to convert a variable to
import warnings
warnings.filterwarnings("ignore") #this can be removed but will help the code be a t


#replace with code for importing the dataset
df = pd.read_excel('/content/hbdata-orig24.xlsx')
#run this code block after importing dataset above
pd.set_option("display.max_rows", None, "display.max_columns", None) #View full data
df
```

		O'Flaherty				00:00:00	9:39:03.033	
271	2474	Blossom Kim	Rufous	NaN	NaN	2017-08-02 00:00:00	16:03:50.478-16:08:17.875	
272	2475	Clancy McKnight	Prefect	M	no	2017-08-02 00:00:00	17:40:02.809-17:44:13.041	D
273	2476	Malachi Schwalbe	Rufus	NaN	male	2017-08-02 00:00:00	2:56:53.742-3:00:54.181	
274	2477	Bill O'Flaherty	Anna's	M	female	2017-08-02 00:00:00	22:31:59.820-22:33:19.050	Back
275	2478	Bort O'Flaherty	Prefect	M	no	2017-08-03 00:00:00	9:45:22.121-9:48:19.824	NE C
276	2479	Walpurga Schwalbe	Rufous	medium	female	2017-08-03 00:00:00	23:43:03.106-23:43:50.746	NE C
277	2480	Renee Prefect	Prefect	XL	male	2017-08-03 00:00:00	12:45:27.135-12:48:29.478	D
278	2481	Hortence Prefect	Rufous	M	male	2017-08-03 00:00:00	3:08:05.954-3:12:17.969	Back
279	2482	Horace	Rufus	S	male	2017-08-03 00:00:00	9:06:36.376-9:08:40.192	NE C
280	2483	Coco Prefect	Calliope	XL	no	2017-08-03 00:00:00	23:53:06.629-23:54:10.682	Back
281	2484	Gunnar Prefect	Rufus	20 ml	no	2017-08-03 00:00:00	14:22:54.401-14:26:12.481	Back
282	2485	Gunnar Prefect	Anna's	XL	male	2017-08-03 00:00:00	12:24:36.488-12:25:54.007	NE C
283	2486	June McKnight	Prefect	XL	no	2017-08-03 00:00:00	21:52:16.738-21:54:30.061	
284	2487	Wally Sharma	Calliope	M	female	2017-08-03 00:00:00	1:22:54.567-1:23:39.488	Back
285	2488	Horace	Anna's	L	female	2017-08-03 00:00:00	9:48:39.472-9:49:54.954	NE C
286	2489	Virgil Schwalbe	Anna's	S	male	2017-08-03 00:00:00	6:35:10.227-6:35:59.223	Back
287	2490	Wally Sharma	Calliope	medium	male	2017-08-03 00:00:00	10:22:04.093-10:23:09.080	13:14:3 13:15:(
288	2491	Ursula Schwalbe	Prefect	XL	female	2017-08-03 00:00:00	7:15:12.062-7:19:21.388	D
289	2492	Kevin McKnight	Rufus	XL	male	2017-08-03 00:00:00	9:15:54.669-9:16:31.527	De
290	2493	Mitz Johnson	Rufous	M	female	2017-07-16 00:00:00	07:27:24.05-07:27:52.78	Back

- Completeness: VisitIDs 2209, 2395, missing data
- Consistency: VisitIDs 2250-2257, medium instead of M. Lots of oz instead of ml as well
- Uniqueness: 2235-2336: Time is the same, duplicate entry

Q3

Answer the following questions regarding the importance of each of the named data scrubbing steps.

3A

What negative outcome would likely occur during data analysis if duplicated visits were not removed (consolidated)?

The dulicate data would be recorded multiple times, scewing the results of a particular bird or feeder. We would likely find that there are more birds observed than there should be, and their results would skew what we assume about that bird.