Pandas (Again)

Meeting #7


CLEAN UP!
DINO SALLY

Dataframes & Data Cleaning

Meeting #7

DataFrame: The pandas DataFrame is a structure that contains two-dimensional data and its corresponding labels.

Let’s first LOOK at some data...

CustomerID	First_Name	Last_Name	Phone_Number	Address	Paying Customer	Do_Not_Contact	Not_Useful_Column
1001	Frodo	Baggins	123-545-5421	123 Shire Lane, Shire	Yes	No	TRUE
1002	Abed	Nadir	123/643/9775	93 West Main Street	No	Yes	FALSE
1003	Walter	/White	7066950392	298 Drugs Driveway	N		TRUE
1004	Dwight	Schrute	123-543-2345	980 Paper Avenue, Pennsylvania, 18503	Yes	Y	TRUE
1005	Jon	Snow	876|678|3469	123 Dragons Road	Y	No	TRUE
1006	Ron	Swanson	304-762-2467	768 City Parkway	Yes	Yes	TRUE
1007	Jeff	Winger		1209 South Street	No	No	FALSE
1008	Sherlock	Holmes	876|678|3469	98 Clue Drive	N	No	FALSE
1009	Gandalf		N/a	123 Middle Earth	Yes		FALSE
1010	Peter	Parker	123-545-5421	25th Main Street, New York	Yes	No	TRUE
1011	Samwise	Gamgee		612 Shire Lane, Shire	Yes	No	TRUE
1012	Harry	...Potter	7066950392	2394 Hogwarts Avenue	Y		TRUE
1013	Don	Draper	123-543-2345	2039 Main Street	Yes	N	FALSE
1014	Leslie	Knope	876|678|3469	343 City Parkway	Yes	No	FALSE
1015	Toby	Flenderson_	304-762-2467	214 HR Avenue	N	No	FALSE
1016	Ron	Weasley	123-545-5421	2395 Hogwarts Avenue	No	N	FALSE
1017	Michael	Scott	123/643/9775	121 Paper Avenue, Pennsylvania	Yes	No	FALSE
1018	Clark	Kent	7066950392	3498 Super Lane	Y		TRUE
1019	Creed	Braton	N/a	N/a	N/a	Yes	TRUE
1020	Anakin	Skywalker	876|678|3469	910 Tatooine Road, Tatooine	Yes	N	TRUE

Purpose of Cleaning Data

Data cleaning is essential to **remove** unnecessary or corrupted data. It also helps to **fix** data that is incorrectly formatted, duplicated, or incomplete.

1. Removing Columns

The drop() function simply removes the column from the dataframe

```
df = df.drop(columns = "Column_we_dont_want")
```

2. Removing Values around Data

The strip() function removes unnecessary characters around our data

```
df = df["Column_Name"].strip("123./\|-$*")

df = df["Column Name"].lstrip("/")

df = df["Column Name"].rstrip(".")
```

strip() - removes from both sides

lstrip() - removes from left side

rstrip() - removes from right side

3. Replacing Values

The replace() function replaces poorly-formatted values to make them more descriptive

```
df["Pumpkin Lover"] = df["Pumpkin Lover"].str.replace("Y", "Yes")

df.replace("NA", "")
```

Before	After
Yes	Yes
Y	Yes
NA	

4. inplace Attribute

The inplace attribute changes the original dataframe

```
df["Ramen_Lover"].drop(inplace=True)
```

inplace=True will change the original dataframe

By default, inplace=False

5. Resetting Indices

The reset_index() function makes the updated columns begin from 0

```
df.reset_index(drop=True)
```

When you reset the index, the older indices gets saved in a new column.

drop=True deletes that new column created

UPCOMING NEXT...

LET'S CODE!!

https://shorturl.at/cAFHJ