Pandas

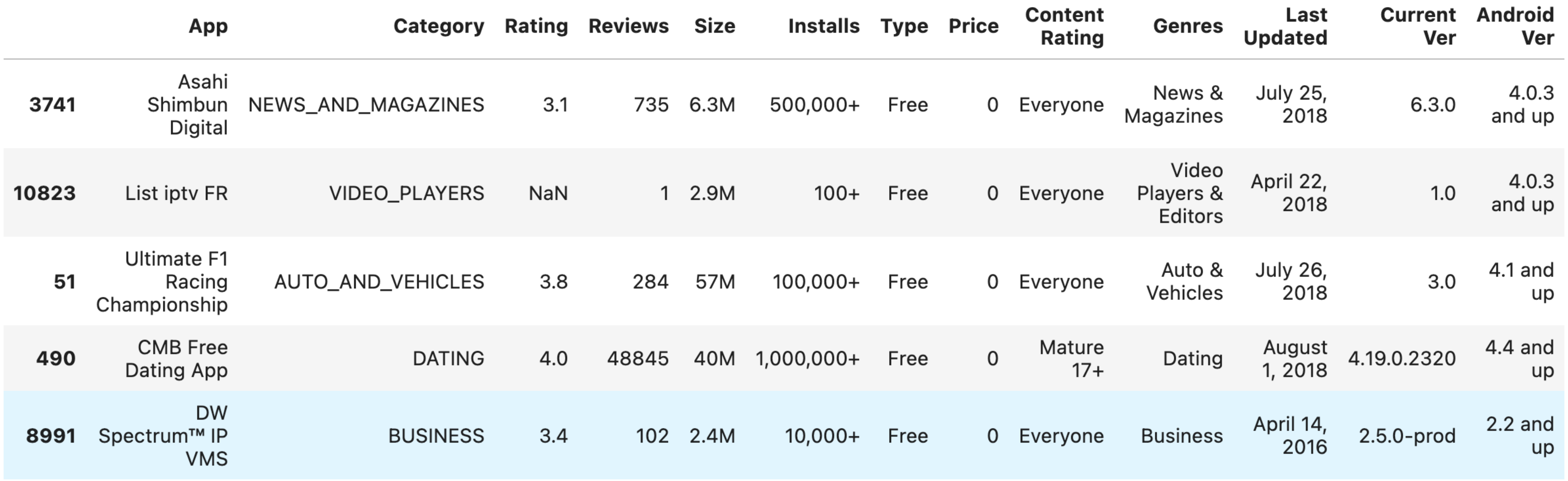
Advanced Pandas

Pandas Advanced: Outline

- Data Preparation and Cleaning
  - Removing unneeded columns
  - Removing the duplicated rows
  - Renaming badly formatted column labels
  - Converting categorical fields into Pandas Category data type
  - Converting numerical fields into numeric values
  - Dealing with missing values
- Utility Pandas functions
- Grouping Dataframes
- Combining Dataframes

Example Dataset: Google Play Store

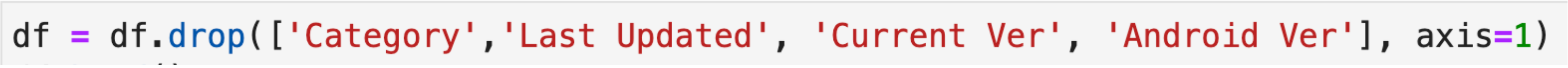
- This dataset has 10000 samples (rows) with 13 features (columns).



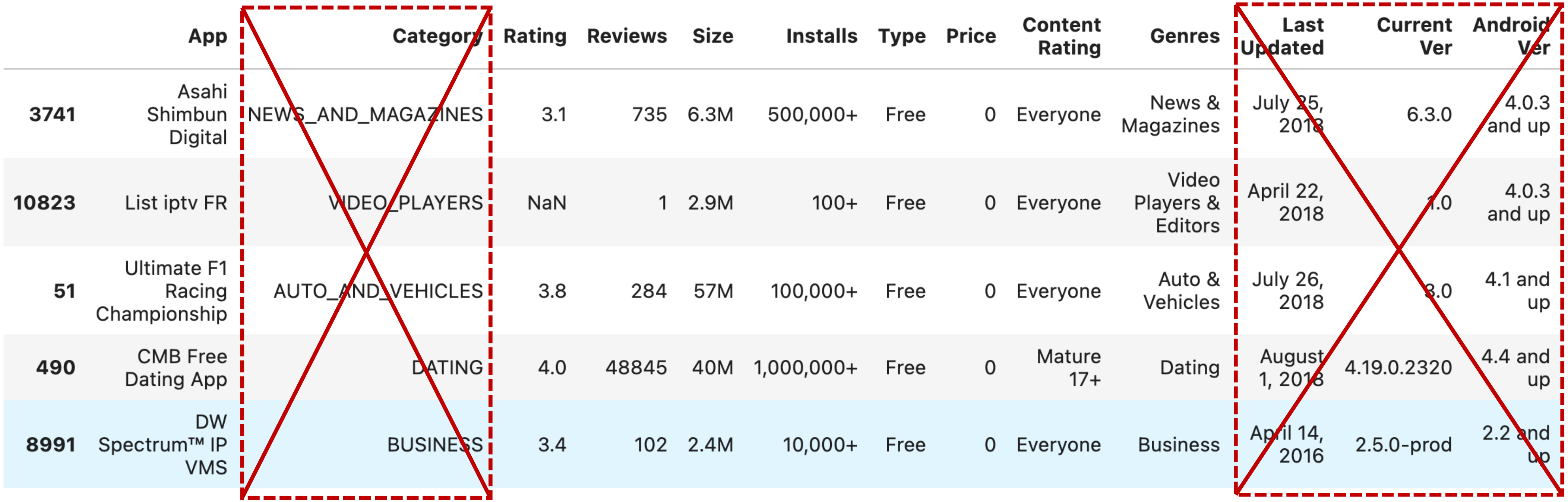
	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
3741	Asahi Shimbun Digital	NEWS_AND_MAGAZINES	3.1	735	6.3M	500,000+	Free	0	Everyone	News & Magazines	July 25, 2018	6.3.0	4.0.3 and up
10823	List iptv FR	VIDEO_PLAYERS	NaN	1	2.9M	100+	Free	0	Everyone	Video Players & Editors	April 22, 2018	1.0	4.0.3 and up
51	Ultimate F1 Racing Championship	AUTO_AND_VEHICLES	3.8	284	57M	100,000+	Free	0	Everyone	Auto & Vehicles	July 26, 2018	3.0	4.1 and up
490	CMB Free Dating App	DATING	4.0	48845	40M	1,000,000+	Free	0	Mature 17+	Dating	August 1, 2018	4.19.0.2320	4.4 and up
8991	DW Spectrum™ IP VMS	BUSINESS	3.4	102	2.4M	10,000+	Free	0	Everyone	Business	April 14, 2016	2.5.0-prod	2.2 and up

Removing Columns from a Dataframe

- We can use drop function to remove some unneeded columns from the data frame.



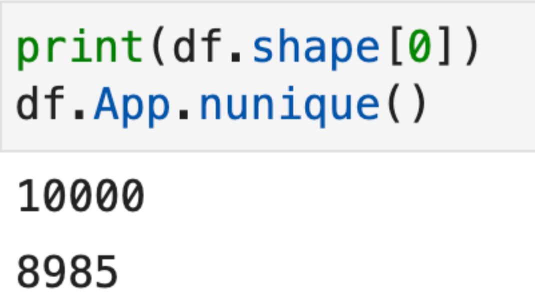
```
df = df.drop(['Category','Last Updated', 'Current Ver', 'Android Ver'], axis=1)
```



	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
3741	Asahi Shimbun Digital	NEWS_AND_MAGAZINES	3.1	735	6.3M	500,000+	Free	0	Everyone	News & Magazines	July 25, 2018	6.3.0	4.0.3 and up
10823	List iptv FR	VIDEO_PLAYERS	NaN	1	2.9M	100+	Free	0	Everyone	Video Players & Editors	April 22, 2018	1.0	4.0.3 and up
51	Ultimate F1 Racing Championship	AUTO_AND_VEHICLES	3.8	284	57M	100,000+	Free	0	Everyone	Auto & Vehicles	July 26, 2018	3.0	4.1 and up
490	CMB Free Dating App	DATING	4.0	48845	40M	1,000,000+	Free	0	Mature 17+	Dating	August 1, 2018	4.19.0.2320	4.4 and up
8991	DW Spectrum™ IP VMS	BUSINESS	3.4	102	2.4M	10,000+	Free	0	Everyone	Business	April 14, 2016	2.5.0-prod	2.2 and up

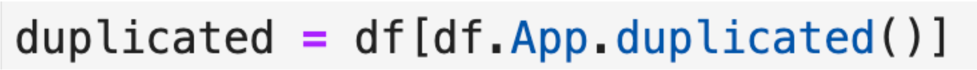
Spotting and Removing Duplicated Samples

- We can use ‘nunique’ and ‘duplicated’ functions to spot duplicated values in a specific column.



```
print(df.shape[0])
df.App.nunique()
```

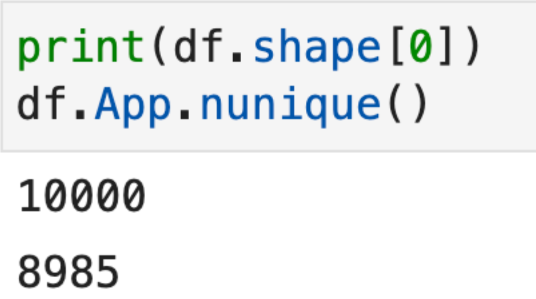
```
10000
8985
```



```
duplicated = df[df.App.duplicated()]
```

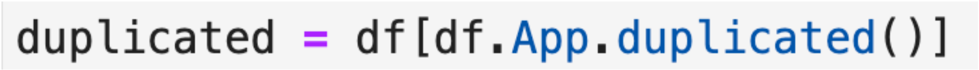
Spotting and Removing Duplicated Samples

- We can use ‘nunique’ and ‘duplicated’ functions to spot duplicated values in a specific column.



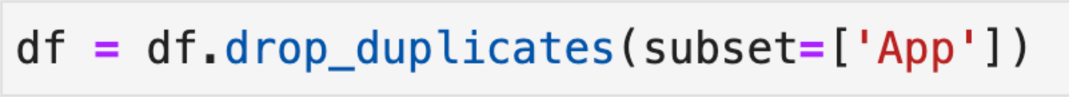
```
print(df.shape[0])
df.App.nunique()
```

```
10000
8985
```



```
duplicated = df[df.App.duplicated()]
```

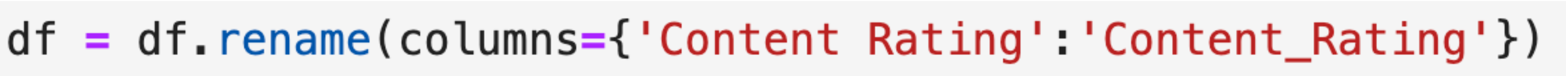
- Then we can use ’drop_duplicates’ function to remove the duplicated samples.



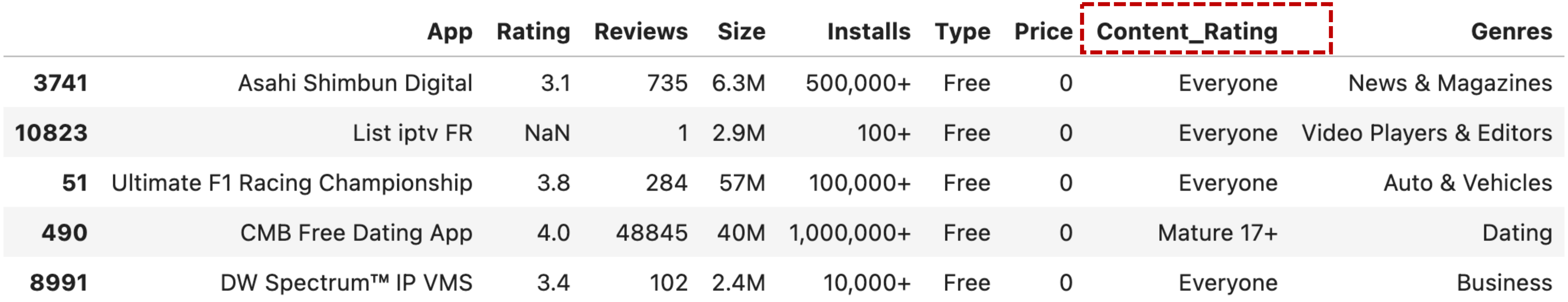
```
df = df.drop_duplicates(subset=['App'])
```

Renaming Columns

- The ‘rename’ function can be used to rename badly formatted column names.



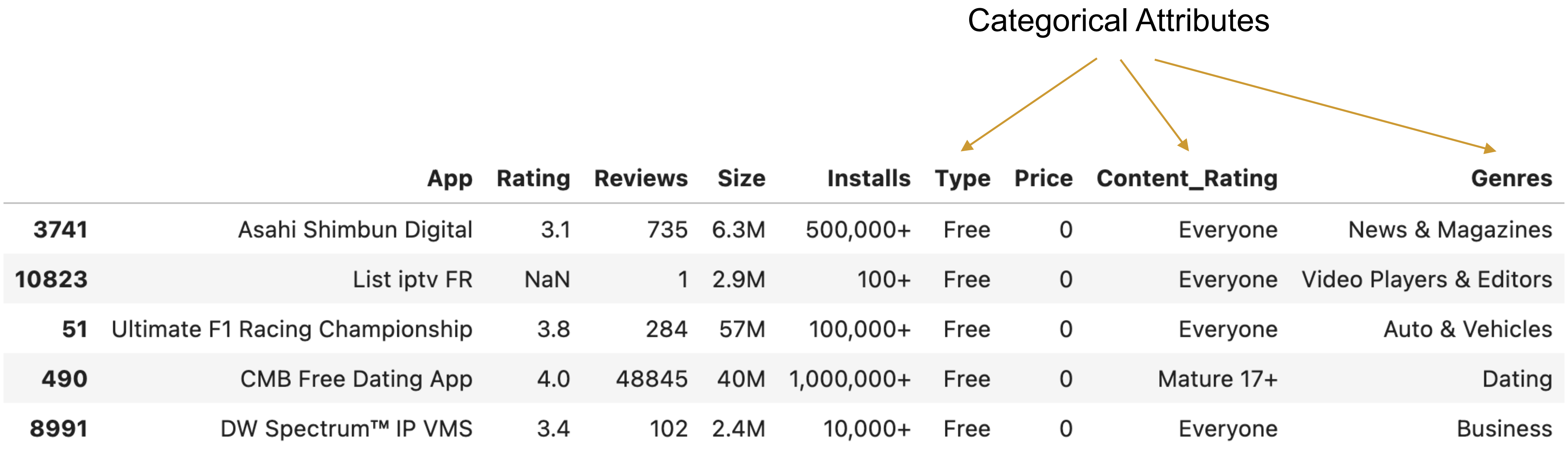
```
df = df.rename(columns={'Content Rating':'Content_Rating'})
```



	App	Rating	Reviews	Size	Installs	Type	Price	Content_Rating	Genres
3741	Asahi Shimbun Digital	3.1	735	6.3M	500,000+	Free	0	Everyone	News & Magazines
10823	List iptv FR	NaN	1	2.9M	100+	Free	0	Everyone	Video Players & Editors
51	Ultimate F1 Racing Championship	3.8	284	57M	100,000+	Free	0	Everyone	Auto & Vehicles
490	CMB Free Dating App	4.0	48845	40M	1,000,000+	Free	0	Mature 17+	Dating
8991	DW Spectrum™ IP VMS	3.4	102	2.4M	10,000+	Free	0	Everyone	Business

Converting categorical fields into Pandas Category

- Casting categorical fields from ‘object’ to ‘categorical’ datatype gives Pandas operations huge boost in processing speed.

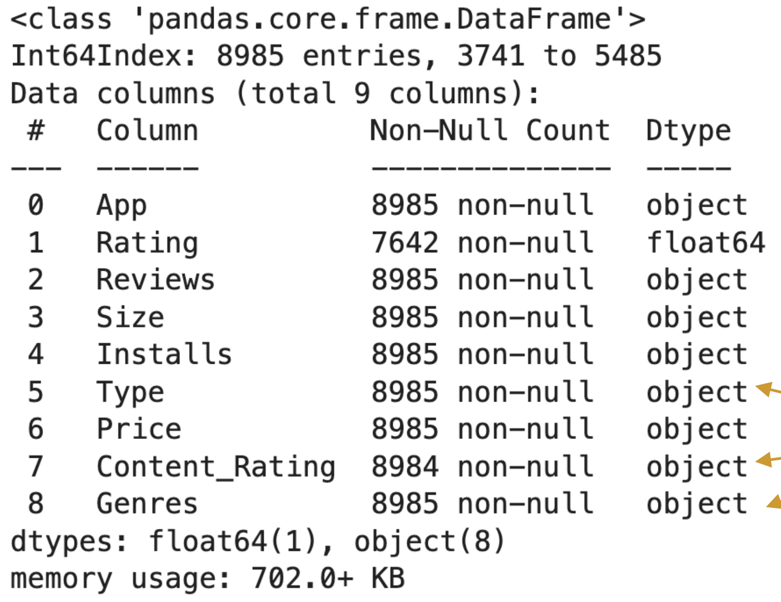


Categorical Attributes

	App	Rating	Reviews	Size	Installs	Type	Price	Content_Rating	Genres
3741	Asahi Shimbun Digital	3.1	735	6.3M	500,000+	Free	0	Everyone	News & Magazines
10823	List iptv FR	NaN	1	2.9M	100+	Free	0	Everyone	Video Players & Editors
51	Ultimate F1 Racing Championship	3.8	284	57M	100,000+	Free	0	Everyone	Auto & Vehicles
490	CMB Free Dating App	4.0	48845	40M	1,000,000+	Free	0	Mature 17+	Dating
8991	DW Spectrum™ IP VMS	3.4	102	2.4M	10,000+	Free	0	Everyone	Business

Converting categorical fields into Pandas Category

- Casting categorical fields from ‘object’ to ‘categorical’ datatype gives Pandas operations huge boost in processing speed.

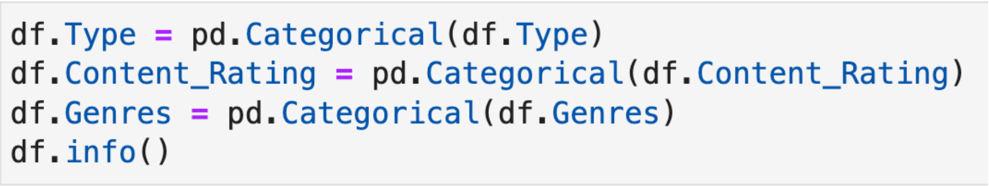


```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 8985 entries, 3741 to 5485
Data columns (total 9 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   App             8985 non-null   object 
 1   Rating          7642 non-null   float64
 2   Reviews         8985 non-null   object 
 3   Size            8985 non-null   object 
 4   Installs        8985 non-null   object 
 5   Type            8985 non-null   object 
 6   Price           8985 non-null   object 
 7   Content_Rating  8984 non-null   object 
 8   Genres          8985 non-null   object 
dtypes: float64(1), object(8)
memory usage: 702.0+ KB
```

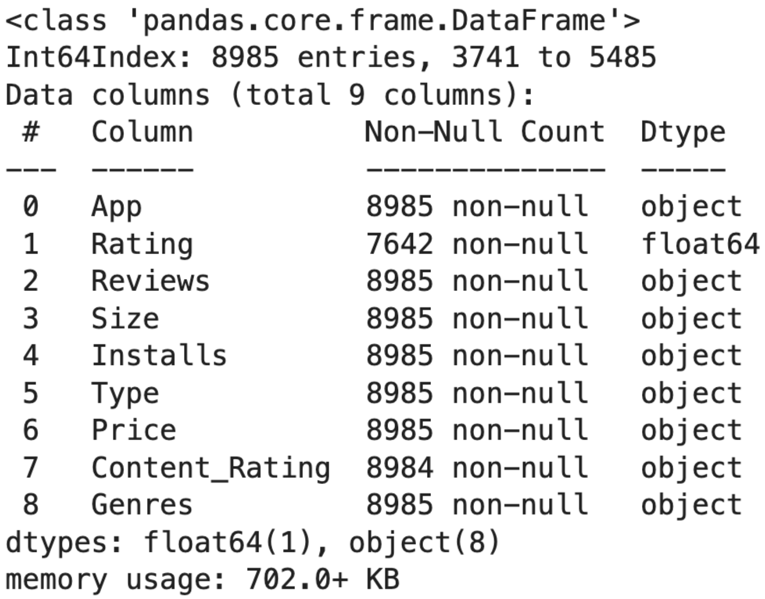
Categorical attributes
with object data type

Converting categorical fields into Pandas Category

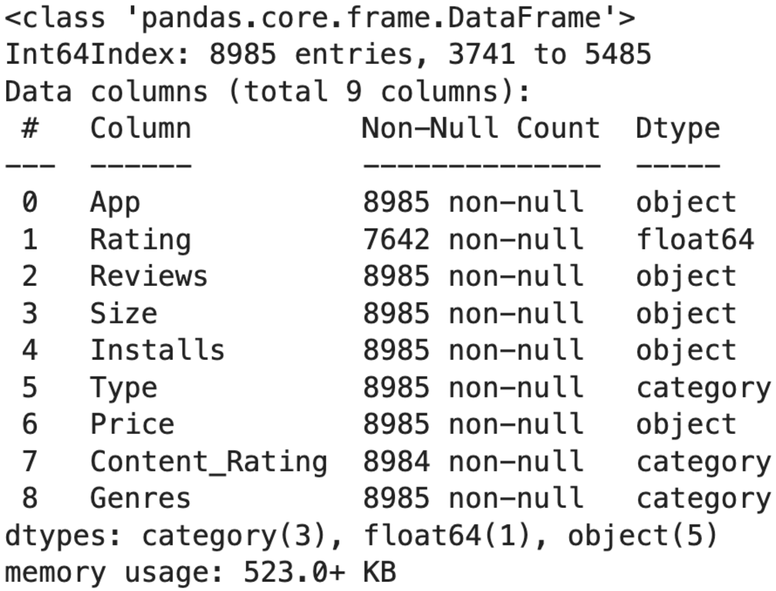
- Casting categorical fields from ‘object’ to ‘categorical’ datatype gives Pandas operations huge boost in processing speed.



```
df.Type = pd.Categorical(df.Type)
df.Content_Rating = pd.Categorical(df.Content_Rating)
df.Genres = pd.Categorical(df.Genres)
df.info()
```



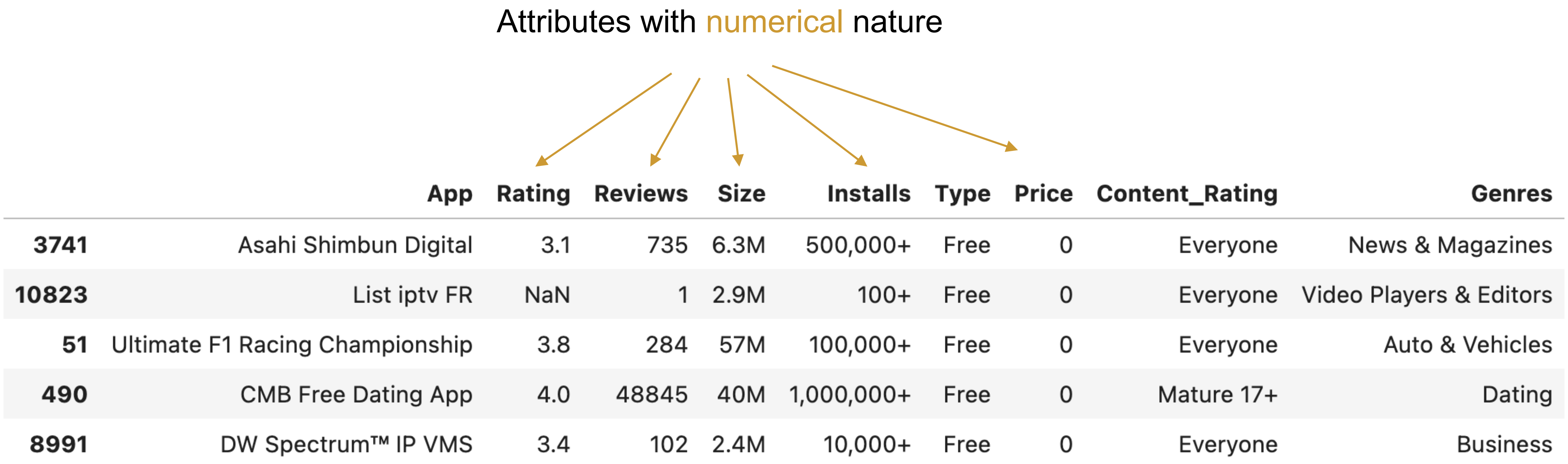
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 8985 entries, 3741 to 5485
Data columns (total 9 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   App             8985 non-null   object
 1   Rating          7642 non-null   float64
 2   Reviews         8985 non-null   object
 3   Size            8985 non-null   object
 4   Installs        8985 non-null   object
 5   Type            8985 non-null   object
 6   Price           8985 non-null   object
 7   Content_Rating  8984 non-null   object
 8   Genres          8985 non-null   object
dtypes: float64(1), object(8)
memory usage: 702.0+ KB
```



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 8985 entries, 3741 to 5485
Data columns (total 9 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   App             8985 non-null   object
 1   Rating          7642 non-null   float64
 2   Reviews         8985 non-null   object
 3   Size            8985 non-null   object
 4   Installs        8985 non-null   object
 5   Type            8985 non-null   category
 6   Price           8985 non-null   object
 7   Content_Rating  8984 non-null   category
 8   Genres          8985 non-null   category
dtypes: category(3), float64(1), object(5)
memory usage: 523.0+ KB
```

Casting to Numerical Values

- Sometimes we need to cast object (string) fields into numerical values before we can perform some statistical or mathematical operations.

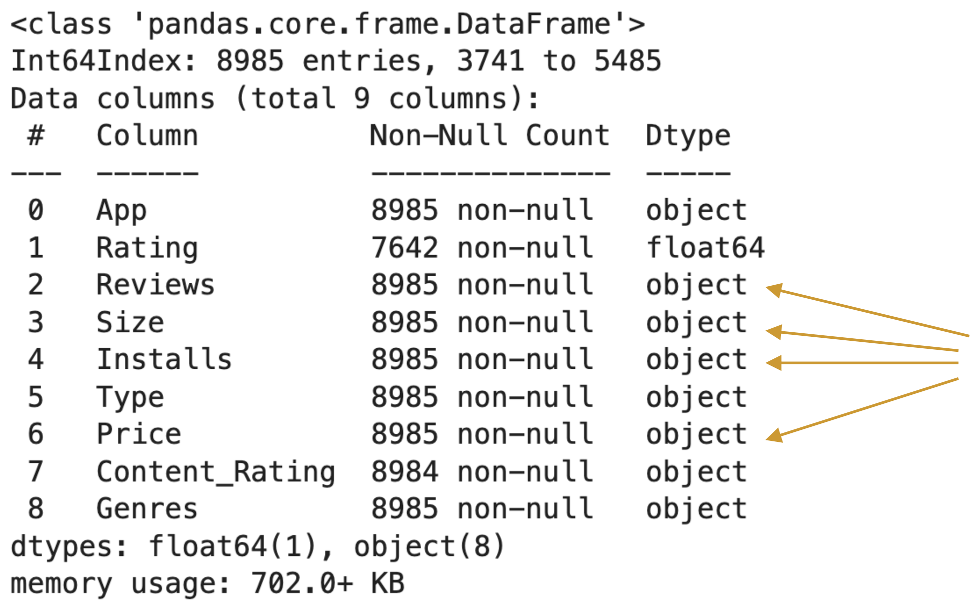


Attributes with numerical nature

	App	Rating	Reviews	Size	Installs	Type	Price	Content_Rating	Genres
3741	Asahi Shimbun Digital	3.1	735	6.3M	500,000+	Free	0	Everyone	News & Magazines
10823	List iptv FR	NaN	1	2.9M	100+	Free	0	Everyone	Video Players & Editors
51	Ultimate F1 Racing Championship	3.8	284	57M	100,000+	Free	0	Everyone	Auto & Vehicles
490	CMB Free Dating App	4.0	48845	40M	1,000,000+	Free	0	Mature 17+	Dating
8991	DW Spectrum™ IP VMS	3.4	102	2.4M	10,000+	Free	0	Everyone	Business

Casting to Numerical Values

- Sometimes we need to cast object (string) fields into numerical values before we can perform some statistical or mathematical operations.

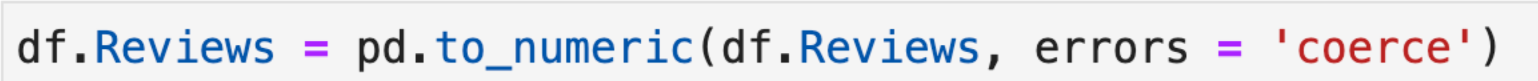


```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 8985 entries, 3741 to 5485
Data columns (total 9 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   App             8985 non-null   object 
 1   Rating          7642 non-null   float64
 2   Reviews         8985 non-null   object 
 3   Size            8985 non-null   object 
 4   Installs        8985 non-null   object 
 5   Type            8985 non-null   object 
 6   Price           8985 non-null   object 
 7   Content_Rating  8984 non-null   object 
 8   Genres          8985 non-null   object 
dtypes: float64(1), object(8)
memory usage: 702.0+ KB
```

Numerical attributes with object data type

Casting to Numerical Values

- ‘to_numeric’ function can be used to cast numerical fields with object data types into numerical data types such as float or int:

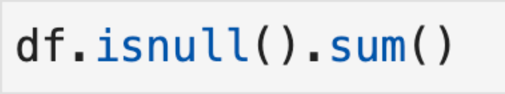


```
df.Reviews = pd.to_numeric(df.Reviews, errors = 'coerce')
```

- But sometimes such a casting is not simply possible, for example when values are mixture of numbers and characters. In these cases, we need to modify the entries by removing extra characters.

Dealing with Missing Values

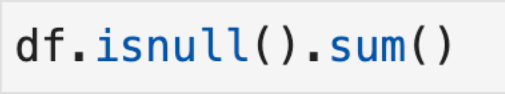
- Detecting missing values: isnull(), isna()



```
df.isnull().sum()
```

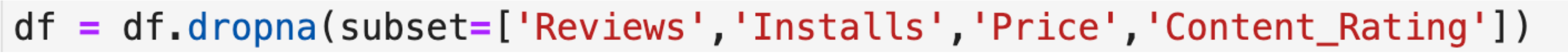
Dealing with Missing Values

- Detecting missing values: isnull(), isna()



```
df.isnull().sum()
```

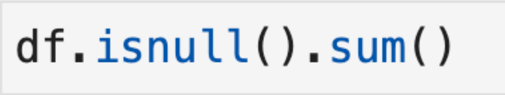
- Removing missing values: a good solution when we have few rows with missing values. We can use ‘dropna’ function to do so.



```
df = df.dropna(subset=['Reviews','Installs','Price','Content_Rating'])
```

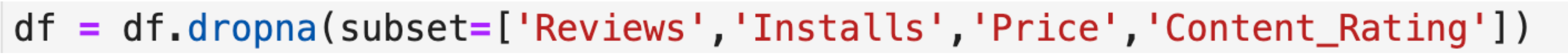
Dealing with Missing Values

- Detecting missing values: isnull(), isna()



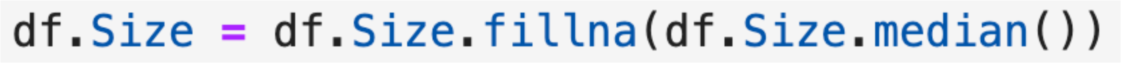
```
df.isnull().sum()
```

- Removing missing values: a good solution when we have few rows with missing values. We can use ‘dropna’ function to do so.



```
df = df.dropna(subset=['Reviews','Installs','Price','Content_Rating'])
```

- Filling in the missing values: useful when having many missing entries in a specific column. We can use ‘fillna’ function to fill in the missing values with specific value.



```
df.Size = df.Size.fillna(df.Size.median())
```

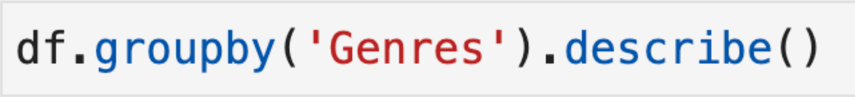
Questions?

Utility Pandas Functions

- count: Counts non-NA cells for each column or row.
- mean: Returns the mean of the values over the requested axis.
- median: Returns the median of the values over the requested axis.
- max: Returns the maximum of the values over the requested axis.
- min: Returns the minimum of the values over the requested axis.
- std: Returns sample standard deviation over requested axis.
- sum: Returns the sum of the values over the requested axis.
- idxmax: Returns index of first occurrence of maximum over an axis.
- idxmin: Returns index of first occurrence of minimum over an axis.
- nlargest: Returns the first n rows ordered by columns in descending order.
- nsmallest: Returns the first n rows ordered by columns in ascending order.
- sort_values: Sorts the dataframe based on the specified column(s).

Grouping Dataframes

- We can use the ‘groupby’ function to group a dataframe based on the values of a column or columns. For example, we can group the applications in our dataset by their genres.

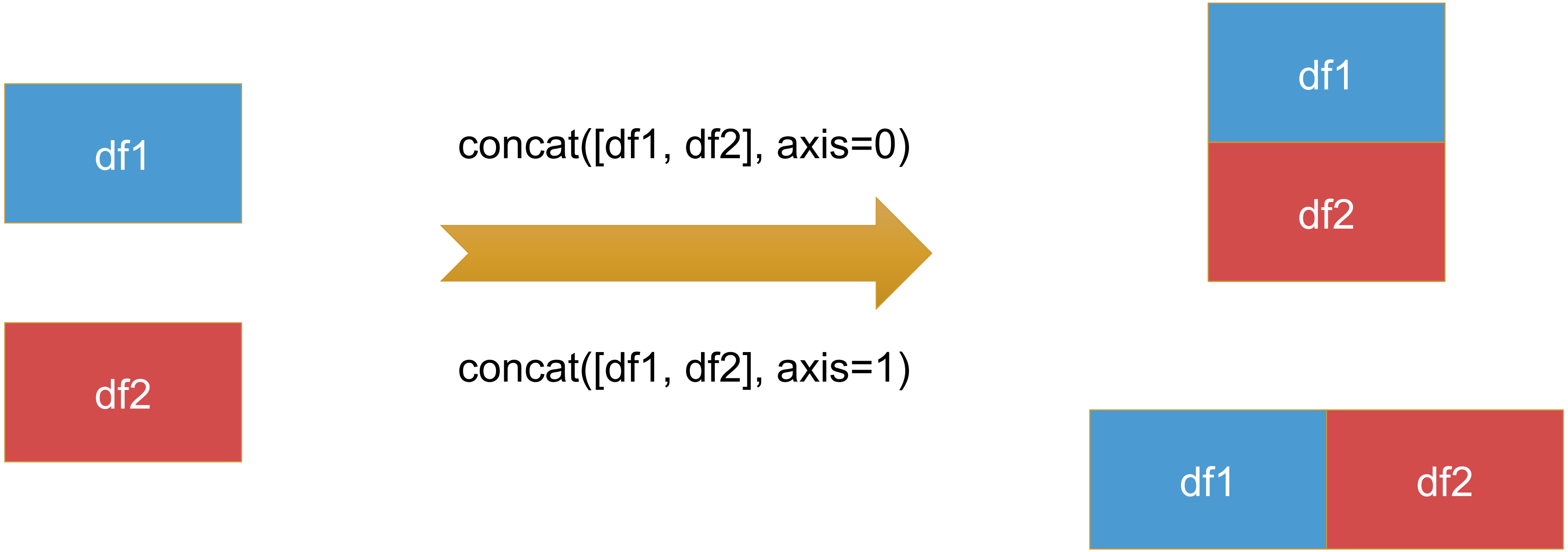


```
df.groupby('Genres').describe()
```

Questions?

Combining Dataframes: Concatenation

- We can use ‘concat’ function to concatenate Pandas objects along a particular axis.



df1

df2

concat([df1, df2], axis=0)

concat([df1, df2], axis=1)

df1

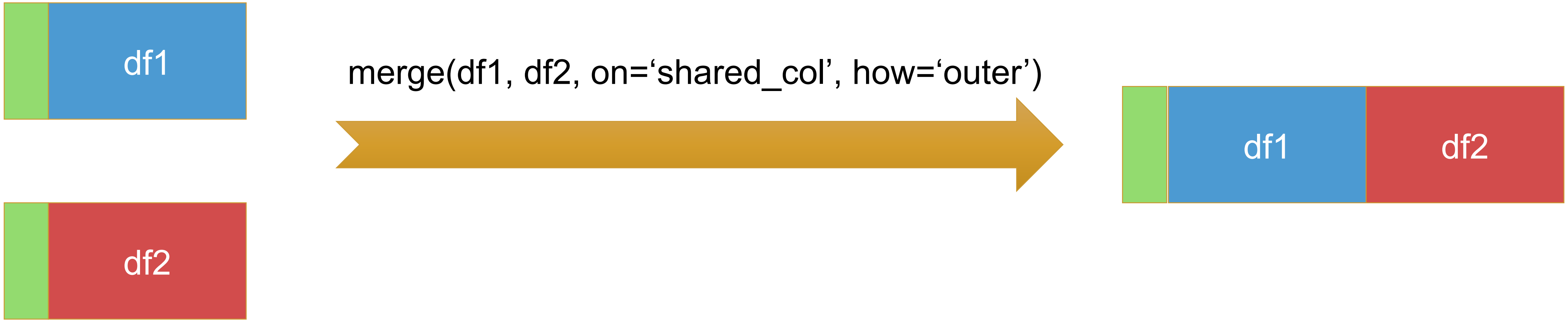
df2

df1

df2

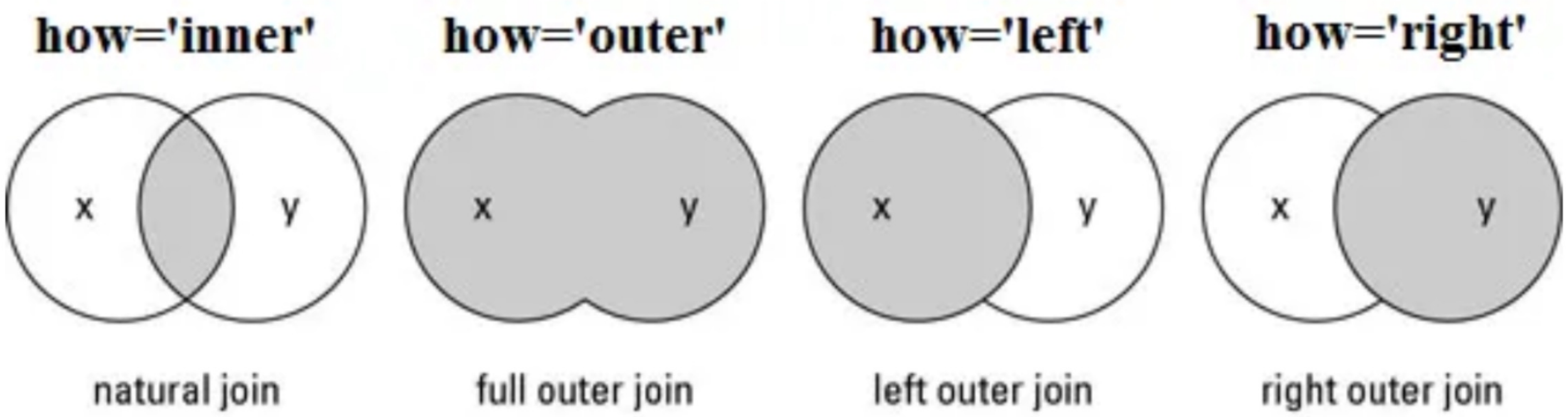
Combining Dataframes: Merging

- The ‘merge’ function is used to combine two dataframes based on a shared column.



Combining Dataframes: Merging

- By specifying the ‘how’ parameter we can choose several strategies to merge two dataframes.



Questions?

Thanks!