Data Cleaning in Databricks

Removing Duplicate Rows

id	first_name	last_name	email	gender	ip_address
1	Karita	Sendley	ksendley0@parallels.com	Female	114.8.222.223
2	Forbes	Wardel	fwardel1@list-manage.com	Male	224.93.24.171
2	Forbes	Wardel	fwardel1@list-manage.com	Male	224.93.24.171

```
df = df.dropDuplicates()
```

id	first_name	last_name	email	gender	ip_address
1	Karita	Sendley	ksendley0@parallels.com	Female	114.8.222.223
2	Forbes	Wardel	fwardel1@list-manage.com	Male	224.93.24.171

Shwetank Singh

GritSetGrow - GSGLearn.com

Data Cleaning in Databricks

Filtering Rows

id	first_name	last_name	email	gender	ip_address
1	Karita	Sendley	ksendley0@parallels.com	Female	114.8.222.223
2	Forbes	Wardel	fwardel1@list-manage.com	Male	224.93.24.171
3	Nesta	Beamond	nbeamond2@gov.uk	Female	124.97.188.174

```
df = df.filter(df.id> 2)
```

id	first_name	last_name	email	gender	ip_address
1	Karita	Sendley	ksendley0@parallels.com	Female	114.8.222.223
2	Forbes	Wardel	fwardel1@list-manage.com	Male	224.93.24.171

Shwetank Singh

GritSetGrow - GSGLearn.com

Data Cleaning in Databricks

Filling or Replacing Null Values

id	first_name	last_name	email	gender	ip_address
1	Karita	Sendley	ksendley0@parallels.com	Female	114.8.222.223
2	Forbes	Wardel	*NULL*	Male	224.93.24.171
3	Nesta	Beamond	nbeamond2@gov.uk	Female	124.97.188.174

```
df = df.na.fill(value="unknown", subset=["email"])
```

id	first_name	last_name	email	gender	ip_address
1	Karita	Sendley	ksendley0@parallels.com	Female	114.8.222.223
2	Forbes	Wardel	unknown	Male	224.93.24.171
3	Nesta	Beamond	nbeamond2@gov.uk	Female	124.97.188.174

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Cleaning in Databricks

Trimming Strings

id	first_name	last_name	email	gender	ip_address
1	Karita	Sendley	ksendley0@parallels.com	Female	114.8.222.223
2	Forbes ___	Wardel	*fwardel1@list-manage.com*	Male	224.93.24.171
3	___Nesta	Beamond	nbeamond2@gov.uk	Female	124.97.188.174

```
from pyspark.sql.functions import trim
df = df.withColumn("first_name", trim(df.first_name))
```

id	first_name	last_name	email	gender	ip_address
1	Karita	Sendley	ksendley0@parallels.com	Female	114.8.222.223
2	Forbes	Wardel	fwardel1@list-manage.com	Male	224.93.24.171
3	Nesta	Beamond	nbeamond2@gov.uk	Female	124.97.188.174

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Cleaning in Databricks

Type Casting

id	first_name	last_name	email	gender	age
1	Karita	Sendley	ksendley0@parallels.com	Female	35
2	Forbes	Wardel	fwardel1@list-manage.com	Male	45
3	Nesta	Beamond	nbeamond2@gov.uk	Female	23

```
df = df.withColumn("age", df["age"].cast("integer"))
```

id	first_name	last_name	email	gender	age
1	Karita	Sendley	ksendley0@parallels.com	Female	35
2	Forbes	Wardel	fwardel1@list-manage.com	Male	45
3	Nesta	Beamond	nbeamond2@gov.uk	Female	23

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Cleaning in Databricks

Renaming Columns

id	first_name	last_name	email	gender	age
1	Karita	Sendley	ksendley0@parallels.com	Female	35
2	Forbes	Wardel	fwardel1@list-manage.com	Male	45
3	Nesta	Beamond	nbeamond2@gov.uk	Female	23

```
df = df.withColumnRenamed("id", "cust_id")
```

cust_id	first_name	last_name	email	gender	age
1	Karita	Sendley	ksendley0@parallels.com	Female	35
2	Forbes	Wardel	fwardel1@list-manage.com	Male	45
3	Nesta	Beamond	nbeamond2@gov.uk	Female	23

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Cleaning in Databricks

Dropping Columns

id	first_name	last_name	email	gender	age
1	Karita	Sendley	ksendley0@parallels.com	Female	35
2	Forbes	Wardel	fwardel1@list-manage.com	Male	45
3	Nesta	Beamond	nbeamond2@gov.uk	Female	23

```
df = df.drop("age")
```

cust_id	first_name	last_name	email	gender
1	Karita	Sendley	ksendley0@parallels.com	Female
2	Forbes	Wardel	fwardel1@list-manage.com	Male
3	Nesta	Beamond	nbeamond2@gov.uk	Female

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Cleaning in Databricks

Splitting Columns

id	full_name	email	gender
1	Karita Sendley	ksendley0@parallels.com	Female
2	Forbes Wardel	fwardel1@list-manage.com	*Male*
3	Nesta Beamond	nbeamond2@gov.uk	Female

```
from pyspark.sql.functions import split
df = df.withColumn("full_name", split(df["full_name"],
" ")).select("full_name.*")
```

cust_id	full_name[0]	full_name[1]	email	gender
1	Karita	Sendley	ksendley0@parallels.com	Female
2	Forbes	Wardel	fwardel1@list-manage.com	Male
3	Nesta	Beamond	nbeamond2@gov.uk	Female

Shwetank Singh

GritSetGrow - GSGLearn.com

Data Cleaning in Databricks

Merging Columns

id	first_name	last_name	email	gender	age
1	Karita	Sendley	ksendley0@parallels.com	Female	35
2	Forbes	Wardel	*fwardel1@list-manage.com*	Male	45
3	Nesta	Beamond	nbeamond2@gov.uk	Female	23

```
from pyspark.sql.functions import concat_ws
df = df.withColumn("full_name",
concat_ws(" ", df["first_name"], df["last_name"]))
```

id	full_name	email	gender	age
1	Karita Sendley	ksendley0@parallels.com	Female	35
2	Forbes Wardel	fwardel1@list-manage.com	*Male*	45
3	Nesta Beamond	nbeamond2@gov.uk	Female	23

Shwetank Singh
GritSetGrow - GSGLearn.com