# PySpark Functions Cheatsheet

## Column & Row Operations

| Function | Description | Example |
|---|---|---|
| withColumn() | Add/modify a column | df.withColumn("new", df["col"] + 1) |
| select() | Select specific columns | df.select("col1", "col2") |
| filter() / where() | Filter rows | df.filter(df["age"] > 30) |
| drop() | Drop a column | df.drop("col") |
| dropDuplicates() | Remove duplicates | df.dropDuplicates(["col"]) |
| distinct() | Unique rows | df.distinct() |
| withColumnRenamed() | Rename column | df.withColumnRenamed("old", "new") |
| orderBy() / sort() | Sort rows | df.orderBy("col", ascending=False) |
| cache() / persist() | Store DataFrame in memory/disk | df.cache() |

## Join & Combine

| Function | Description | Example |
|---|---|---|
| join() | Join two DataFrames | df1.join(df2, "id", "inner") |
| union() | Append DataFrames | df1.union(df2) |
| unionByName() | Append by matching column names | df1.unionByName(df2) |
| repartition() | Increase partitions | df.repartition(10) |
| coalesce() | Reduce partitions | df.coalesce(1) |

## Aggregation & Grouping

| Function | Description | Example |
|---|---|---|
| groupBy() | Group by column(s) | df.groupBy("col") |
| agg() | Multiple aggregations | df.agg(F.max("col"), F.min("col")) |
| describe() | Summary stats | df.describe().show() |
| collect_list() | List aggregation | collect_list("name") |
| collect_set() | Set (distinct) aggregation | collect_set("name") |

## Null & Conditional Handling

| Function | Description | Example |
|---|---|---|
| fillna() | Replace nulls | df.fillna(0) |
| isNull() | Check nulls | df.filter(df["col"].isNull()) |
| when() | If-else logic | when(df.age < 18, "minor").otherwise("adult") |
| lit() | Add literal value | lit("USA") |

| Function | Description | Example |
|---|---|---|
| row_number() | Row number per window | row_number().over(Window.partitionBy("col")) |
| rank() | Rank with gaps | rank().over(Window.partitionBy("col")) |

| | | |
|---|---|---|
| dense_rank() | Rank without gaps | dense_rank().over(Window.partitionBy("col")) |

## Window & Ranking Functions

## Date & Time

| Function | Description | Example |
|---|---|---|
| current_date() | Today's date | current_date() |
| current_timestamp() | Current timestamp | current_timestamp() |
| to_date() | Convert to date | to_date(df["ts"]) |
| date_format() | Format date | date_format("date", "yyyy-MM") |

## Complex Types & JSON

| Function | Description | Example |
|---|---|---|
| explode() | Flatten arrays | explode("skills") |
| array() / struct() | Create complex types | array("c1", "c2"), struct("c1", "c2") |
| get_json_object() | Extract JSON field | get_json_object("json_col", "$.name") |
| from_json() / to_json() | Parse or stringify JSON | from_json("col", schema) |