In [1]:
```
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"
```

In [2]:
```
import pandas as pd
from pathlib import Path
import pyarrow.parquet as pq

month = 1
year = 2023
path = Path("..") / "data" / "raw" / f"rides_{year}_{month:02}.parquet"

table = pq.read_table(path)
rides = table.to_pandas()
rides.head()
```

```
A module that was compiled using NumPy 1.x cannot be run in
NumPy 2.2.3 as it may crash. To support both 1.x and 2.x
versions of NumPy, modules must be compiled with NumPy 2.0.
Some module may need to rebuild instead e.g. with 'pybind11>=2.12'.

If you are a user of the module, the easiest solution will be to
downgrade to 'numpy<2' or try to upgrade the affected module.
We expect that some modules will need time to support NumPy 2.

Traceback (most recent call last):  File "<frozen runpy>", line 198, in _run_module_as_m
ain
  File "<frozen runpy>", line 88, in _run_code
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\ipykernel_launcher.py", line 17, in
<module>
    app.launch_new_instance()
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\traitlets\config\application.py", l
ine 992, in launch_instance
    app.start()
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\ipykernel\kernelapp.py", line 711,
in start
    self.io_loop.start()
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\tornado\platform\asyncio.py", line
195, in start
    self.asyncio_loop.run_forever()
  File "C:\Users\SUMANTH\anaconda3\Lib\asyncio\base_events.py", line 607, in run_forever
    self._run_once()
  File "C:\Users\SUMANTH\anaconda3\Lib\asyncio\base_events.py", line 1922, in _run_once
    handle._run()
  File "C:\Users\SUMANTH\anaconda3\Lib\asyncio\events.py", line 80, in _run
    self._context.run(self._callback, *self._args)
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\ipykernel\kernelbase.py", line 510,
in dispatch_queue
    await self.process_one()
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\ipykernel\kernelbase.py", line 499,
in process_one
    await dispatch(*args)
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\ipykernel\kernelbase.py", line 406,
in dispatch_shell
    await result
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\ipykernel\kernelbase.py", line 729,
in execute_request
    reply_content = await reply_content
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\ipykernel\ipkernel.py", line 411, i
n do_execute
    res = shell.run_cell(
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\ipykernel\zmqshell.py", line 531, i
n run_cell
    return super().run_cell(*args, **kwargs)
```

```
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\IPython\core\interactiveshell.py",
line 3006, in run_cell
    result = self._run_cell(
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\IPython\core\interactiveshell.py",
line 3061, in _run_cell
    result = runner(coro)
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\IPython\core\async_helpers.py", lin
e 129, in _pseudo_sync_runner
    coro.send(None)
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\IPython\core\interactiveshell.py",
line 3266, in run_cell_async
    has_raised = await self.run_ast_nodes(code_ast.body, cell_name,
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\IPython\core\interactiveshell.py",
line 3445, in run_ast_nodes
    if await self.run_code(code, result, async_=asy):
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\IPython\core\interactiveshell.py",
line 3505, in run_code
    exec(code_obj, self.user_global_ns, self.user_ns)
  File "C:\Users\SUMANTH\AppData\Local\Temp\ipykernel_26552\2593841401.py", line 1, in <
module>
    import pandas as pd
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\pandas\__init__.py", line 62, in <m
odule>
    from pandas.core.api import (
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\pandas\core\api.py", line 28, in <m
odule>
    from pandas.core.arrays import Categorical
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\pandas\core\arrays\__init__.py", li
ne 1, in <module>
    from pandas.core.arrays.arrow import ArrowExtensionArray
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\pandas\core\arrays\arrow\__init__.p
y", line 5, in <module>
    from pandas.core.arrays.arrow.array import ArrowExtensionArray
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\pandas\core\arrays\arrow\array.py",
line 50, in <module>
    from pandas.core import (
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\pandas\core\ops\__init__.py", line
8, in <module>
    from pandas.core.ops.array_ops import (
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\pandas\core\ops\array_ops.py", line
56, in <module>
    from pandas.core.computation import expressions
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\pandas\core\computation\expression
s.py", line 21, in <module>
    from pandas.core.computation.check import NUMEXPR_INSTALLED
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\pandas\core\computation\check.py",
line 5, in <module>
    ne = import_optional_dependency("numexpr", errors="warn")
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\pandas\compat\_optional.py", line 1
35, in import_optional_dependency
    module = importlib.import_module(name)
  File "C:\Users\SUMANTH\anaconda3\Lib\importlib\__init__.py", line 126, in import_modul
e
    return _bootstrap._gcd_import(name[level:], package, level)
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\numexpr\__init__.py", line 24, in <
module>
    from numexpr.interpreter import MAX_THREADS, use_vml, __BLOCK_SIZE1__
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
AttributeError: _ARRAY_API not found

A module that was compiled using NumPy 1.x cannot be run in
NumPy 2.2.3 as it may crash. To support both 1.x and 2.x
versions of NumPy, modules must be compiled with NumPy 2.0.
Some module may need to rebuild instead e.g. with 'pybind11>=2.12'.
```

```
If you are a user of the module, the easiest solution will be to
downgrade to 'numpy<2' or try to upgrade the affected module.
We expect that some modules will need time to support NumPy 2.

Traceback (most recent call last):  File "<frozen runpy>", line 198, in _run_module_as_m
ain
  File "<frozen runpy>", line 88, in _run_code
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\ipykernel_launcher.py", line 17, in
<module>
    app.launch_new_instance()
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\traitlets\config\application.py", l
ine 992, in launch_instance
    app.start()
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\ipykernel\kernelapp.py", line 711,
in start
    self.io_loop.start()
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\tornado\platform\asyncio.py", line
195, in start
    self.asyncio_loop.run_forever()
  File "C:\Users\SUMANTH\anaconda3\Lib\asyncio\base_events.py", line 607, in run_forever
    self._run_once()
  File "C:\Users\SUMANTH\anaconda3\Lib\asyncio\base_events.py", line 1922, in _run_once
    handle._run()
  File "C:\Users\SUMANTH\anaconda3\Lib\asyncio\events.py", line 80, in _run
    self._context.run(self._callback, *self._args)
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\ipykernel\kernelbase.py", line 510,
in dispatch_queue
    await self.process_one()
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\ipykernel\kernelbase.py", line 499,
in process_one
    await dispatch(*args)
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\ipykernel\kernelbase.py", line 406,
in dispatch_shell
    await result
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\ipykernel\kernelbase.py", line 729,
in execute_request
    reply_content = await reply_content
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\ipykernel\ipkernel.py", line 411, i
n do_execute
    res = shell.run_cell(
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\ipykernel\zmqshell.py", line 531, i
n run_cell
    return super().run_cell(*args, **kwargs)
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\IPython\core\interactiveshell.py",
line 3006, in run_cell
    result = self._run_cell(
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\IPython\core\interactiveshell.py",
line 3061, in _run_cell
    result = runner(coro)
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\IPython\core\async_helpers.py", lin
e 129, in _pseudo_sync_runner
    coro.send(None)
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\IPython\core\interactiveshell.py",
line 3266, in run_cell_async
    has_raised = await self.run_ast_nodes(code_ast.body, cell_name,
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\IPython\core\interactiveshell.py",
line 3445, in run_ast_nodes
    if await self.run_code(code, result, async_=asy):
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\IPython\core\interactiveshell.py",
line 3505, in run_code
    exec(code_obj, self.user_global_ns, self.user_ns)
  File "C:\Users\SUMANTH\AppData\Local\Temp\ipykernel_26552\2593841401.py", line 1, in <
module>
    import pandas as pd
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\pandas\__init__.py", line 62, in <m
odule>
```

```
    from pandas.core.api import (
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\pandas\core\api.py", line 28, in <m
odule>
    from pandas.core.arrays import Categorical
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\pandas\core\arrays\__init__.py", li
ne 1, in <module>
    from pandas.core.arrays.arrow import ArrowExtensionArray
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\pandas\core\arrays\arrow\__init__.p
y", line 5, in <module>
    from pandas.core.arrays.arrow.array import ArrowExtensionArray
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\pandas\core\arrays\arrow\array.py",
line 64, in <module>
    from pandas.core.arrays.masked import BaseMaskedArray
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\pandas\core\arrays\masked.py", line
60, in <module>
    from pandas.core import (
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\pandas\core\nanops.py", line 52, in
<module>
    bn = import_optional_dependency("bottleneck", errors="warn")
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\pandas\compat\_optional.py", line 1
35, in import_optional_dependency
    module = importlib.import_module(name)
  File "C:\Users\SUMANTH\anaconda3\Lib\importlib\__init__.py", line 126, in import_modul
e
    return _bootstrap._gcd_import(name[level:], package, level)
  File "C:\Users\SUMANTH\anaconda3\Lib\site-packages\bottleneck\__init__.py", line 7, in
<module>
    from .move import (move_argmax, move_argmin, move_max, move_mean, move_median,
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
AttributeError: _ARRAY_API not found
```

Out[2]:

	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	RatecodeID	store_and
0	2	2023-01-01 00:32:10	2023-01-01 00:40:36	1.0	0.97	1.0	
1	2	2023-01-01 00:55:08	2023-01-01 01:01:27	1.0	1.10	1.0	
2	2	2023-01-01 00:25:04	2023-01-01 00:37:49	1.0	2.51	1.0	
3	1	2023-01-01 00:03:48	2023-01-01 00:13:25	0.0	1.90	1.0	
4	2	2023-01-01 00:10:29	2023-01-01 00:21:19	1.0	1.43	1.0	

In [3]:
```
rides_cp = rides.copy()
rides_cp["duration"] = rides["tpep_dropoff_datetime"] - rides["tpep_pickup_datetime"]
rides_cp.head()
```

Out[3]:

	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	RatecodeID	store_and
0	2	2023-01-01 00:32:10	2023-01-01 00:40:36	1.0	0.97	1.0	
1	2	2023-01-01 00:55:08	2023-01-01 01:01:27	1.0	1.10	1.0	
2	2	2023-01-01 00:25:04	2023-01-01 00:37:49	1.0	2.51	1.0	
3	1	2023-01-01 00:03:48	2023-01-01 00:13:25	0.0	1.90	1.0	
4	2	2023-01-01 00:10:29	2023-01-01 00:21:19	1.0	1.43	1.0	

In [4]:
```
rides_cp["duration"].describe().T
```

Out[4]:
```
count                     3066766
mean      0 days 00:15:40.139710
std       0 days 00:42:35.661074
min            -1 days +23:30:48
25%              0 days 00:07:07
50%              0 days 00:11:31
75%              0 days 00:18:18
max              6 days 23:09:11
Name: duration, dtype: object
```

In [5]:
```
rides_cp["duration"].quantile(0)
rides_cp["duration"].quantile(0.01)
rides_cp["duration"].quantile(0.995)
rides_cp["duration"].quantile(0.999)
```

Out[5]:
```
Timedelta('-1 days +23:30:48')
```

Out[5]:
```
Timedelta('0 days 00:00:47')
```

Out[5]:
```
Timedelta('0 days 01:05:31')
```

Out[5]:
```
Timedelta('0 days 02:55:49.290000')
```

In [6]:
```
duration_filter = (rides_cp["duration"] > pd.Timedelta(0)) & (rides_cp["duration"] <= pd
sum(~duration_filter)
```

Out[6]:
```
4001
```

In [7]:
```
rides_cp["total_amount"].describe().T
```

Out[7]:
```
count    3.066766e+06
mean     2.702038e+01
std      2.216359e+01
min     -7.510000e+02
25%      1.540000e+01
50%      2.016000e+01
75%      2.870000e+01
max      1.169400e+03
Name: total_amount, dtype: float64
```

In [8]:
```
rides_cp["total_amount"].quantile(0.0)
rides_cp["total_amount"].quantile(0.01)
rides_cp["total_amount"].quantile(0.995)
rides_cp["total_amount"].quantile(0.999)
```

Out[8]:
```
np.float64(-751.0)
```

Out[8]:
```
np.float64(5.5)
```

Out[8]:
```
np.float64(108.9)
```

Out[8]:
```
np.float64(167.01175000001678)
```

In [9]:
```
rides_cp["total_amount"].max()
```

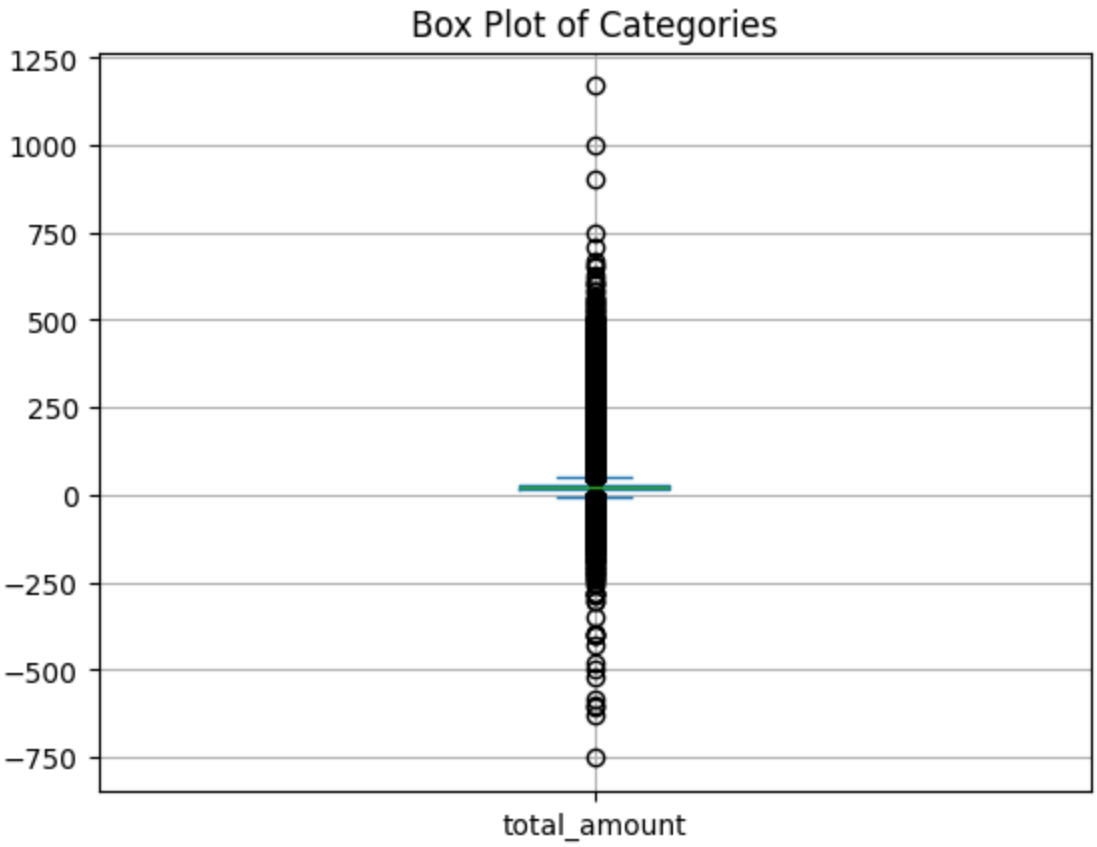
Out[9]:
```
np.float64(1169.4)
```

In [10]:
```
total_amount_filter = (rides_cp["total_amount"]  > 0) & (rides_cp["total_amount"] <= rid
sum(~total_amount_filter) / rides_cp.shape[0] * 100
```

Out[10]:
```
0.9403717140466537
```

In [11]:
```
rides_cp["total_amount"].plot.box(title="Box Plot of Categories", grid=True)
```

Out[11]: <Axes: title={'center': 'Box Plot of Categories'}>



In [12]:
```
nyc_locations = ~rides_cp["PULocationID"].isin((1, 264, 265))
sum(~nyc_locations)
```

Out[12]: 42173

In [13]:
```
sorted_df = rides_cp.sort_values(by="tpep_pickup_datetime", ascending=True)

# Get the top 10 (smallest) and bottom 10 (largest) values
top_10 = sorted_df.head(10)
bottom_10 = sorted_df.tail(10)

top_10

bottom_10
```

Out[13]:

	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	RatecodeID	st
2138036	2	2008-12-31 23:01:42	2009-01-01 14:29:11	1.0	17.76	2.0	
209091	2	2008-12-31 23:04:41	2009-01-01 19:55:36	1.0	0.00	1.0	
10023	2	2022-10-24 17:37:47	2022-10-24 17:37:51	1.0	0.00	5.0	
18219	2	2022-10-24 20:01:46	2022-10-24 20:01:48	1.0	0.00	5.0	
21660	2	2022-10-24 21:45:35	2022-10-24 21:45:38	1.0	0.00	5.0	
22489	2	2022-10-24 23:15:32	2022-10-24 23:15:42	1.0	0.00	5.0	
24577	2	2022-10-25 00:42:10	2022-10-25 00:44:22	1.0	0.97	1.0	

24578	2	2022-10-25 00:59:02	2022-10-25 01:09:02	1.0	2.33	1.0
31916	2	2022-10-25 03:45:46	2022-10-25 03:45:50	1.0	0.02	5.0
47843	2	2022-10-25 07:48:15	2022-10-25 07:48:18	2.0	0.76	5.0

Out[13]:

	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	RatecodeID	st
2993635	2	2023-02-01 00:00:01	2023-02-01 00:33:41	1.0	17.31	2.0	
2993262	2	2023-02-01 00:00:18	2023-02-01 00:08:46	1.0	2.12	1.0	
2993890	2	2023-02-01 00:00:20	2023-02-01 00:13:18	2.0	2.31	1.0	
2992346	2	2023-02-01 00:00:24	2023-02-01 00:07:53	2.0	2.22	1.0	
2994212	2	2023-02-01 00:00:35	2023-02-01 00:17:12	1.0	2.88	1.0	
2994844	2	2023-02-01 00:00:40	2023-02-01 00:23:03	5.0	10.12	1.0	
2993558	2	2023-02-01 00:00:55	2023-02-01 00:06:33	1.0	1.09	1.0	
2992642	2	2023-02-01 00:01:10	2023-02-01 00:14:26	1.0	2.03	1.0	
2929496	2	2023-02-01 00:13:10	2023-02-01 00:29:37	1.0	3.27	1.0	
2929497	2	2023-02-01 00:56:53	2023-02-01 01:06:43	1.0	2.38	1.0	

In [14]:
```
filter_date_range = (rides_cp["tpep_pickup_datetime"] >= "2023-01-01") & (rides_cp["tpep
sum(~filter_date_range)
```

Out[14]:
```
48
```

In [15]:
```
final_filter = duration_filter & total_amount_filter & nyc_locations & filter_date_range
numbers_dropped = final_filter.shape[0] - sum(final_filter) # numbers dropped
numbers_dropped
numbers_dropped/final_filter.shape[0] * 100
```

Out[15]:
```
73626
```

Out[15]:
```
2.400770062013209
```

In [16]:
```
rides = rides[final_filter]
rides = rides[["tpep_pickup_datetime", "PULocationID"]]
rides.rename(columns={
    "tpep_pickup_datetime": "pickup_datetime",
    "PULocationID": "pickup_location_id"
}, inplace=True)
rides.head()
year = 2023
month = 1
```

```
path = Path("..") / "data" / "processed" / f"rides_{year}_{month:02}.parquet"
rides.to_parquet(path, engine="pyarrow", index=False)
```

Out[16]:

	pickup_datetime	pickup_location_id
0	2023-01-01 00:32:10	161
1	2023-01-01 00:55:08	43
2	2023-01-01 00:25:04	48
3	2023-01-01 00:03:48	138
4	2023-01-01 00:10:29	107

In [17]:
```
rides[rides["pickup_location_id"] == 2]
```

Out[17]:

	pickup_datetime	pickup_location_id
2687593	2023-01-28 17:03:38	2

In [ ]:

In [ ]:

In [ ]: