Assignment07 Sutow Brett

July 17, 2021

```
[]: #Assignment 7#
      \#I believe my code is correct, however I got an error throughout dealing with
      →opening the file#
      #I used what was provided to open the file#
[56]: #7A#
      import pandas as pd
      from pathlib import Path
      import pyarrow as pa
      import pyarrow.parquet as parq
      def get_key(val):
          for key, value in parts.items():
              #if val == str(value):
              if val in value:
                  return kev
          return "key doesn't exist"
      partitions = (
              ('A', 'A'), ('B', 'B'), ('C', 'D'), ('E', 'F'),
              ('G', 'H'), ('I', 'J'), ('K', 'L'), ('M', 'M'),
              ('N', 'N'), ('O', 'P'), ('Q', 'R'), ('S', 'T'),
              ('U', 'U'), ('V', 'V'), ('W', 'X'), ('Y', 'Z')
      partitions_val_keys = (
              'A', 'B', 'C-D', 'E-F',
              'G-H', 'I-J', 'K-L', 'M',
              'N', 'O-P', 'Q-R', 'S-T',
              'U', 'V', 'W-X', 'Y-Z'
      parts = dict(zip(partitions_val_keys, partitions))
      print(parts)
      import os
      import json
```

from pathlib import Path

import gzip

```
import hashlib
import shutil
import pandas as pd
import pygeohash
import s3fs
endpoint_url='https://storage.budsc.midwest-datascience.com'
current_dir = Path(os.getcwd()).absolute()
results_dir = current_dir.joinpath('results')
if results dir.exists():
    shutil.rmtree(results dir)
results dir.mkdir(parents=True, exist ok=True)
def read_jsonl_data():
   s3 = s3fs.S3FileSystem(
       anon=True,
       client_kwargs={
            'endpoint_url': endpoint_url
        }
    src_data_path = 'data/processed/openflights/routes.jsonl.gz'
   with s3.open(src_data_path, 'rb') as f_gz:
       with gzip.open(f_gz, 'rb') as f:
            records = [json.loads(line) for line in f.readlines()]
   return records
def flatten record(record):
   flat record = dict()
   for key, value in record.items():
        if key in ['airline', 'src_airport', 'dst_airport']:
            if isinstance(value, dict):
                for child_key, child_value in value.items():
                    flat_key = '{}_{}'.format(key, child_key)
                    flat_record[flat_key] = child_value
        else:
            flat_record[key] = value
   return flat_record
def create_flattened_dataset():
   records = read_jsonl_data()
   parquet_path = results_dir.joinpath('routes-flattened.parquet')
   return pd.DataFrame.from_records([flatten_record(record) for record in_u
→records])
df = create_flattened_dataset()
df['key'] = df['src_airport_iata'].astype(str) + df['dst_airport_iata'].
→astype(str) + df['airline_iata'].astype(str)
ndf = pd.read_parquet(df, engine='pyarrow')
print(list(ndf.columns.values))
```

```
ndf['key'] = ndf['src_airport.iata']+ndf['dst_airport.iata']+ndf['airline.icao']
ndf['partition_value'] = ndf['key'].str[:1]
ndf['kv_key'] = ndf.apply(lambda x: get_key(x.partition_value), axis=1)
table = pa.Table.from_pandas(ndf)
print(table)
{'A': ('A', 'A'), 'B': ('B', 'B'), 'C-D': ('C', 'D'), 'E-F': ('E', 'F'), 'G-H':
('G', 'H'), 'I-J': ('I', 'J'), 'K-L': ('K', 'L'), 'M': ('M', 'M'), 'N': ('N',
'N'), 'O-P': ('O', 'P'), 'Q-R': ('Q', 'R'), 'S-T': ('S', 'T'), 'U': ('U', 'U'),
'V': ('V', 'V'), 'W-X': ('W', 'X'), 'Y-Z': ('Y', 'Z')}
 TypeError
                                             Traceback (most recent call last)
 TypeError: cannot construct a FileSource from
                                                        airline_airline_id
                           airline_alias \
  →airline_name
                        410
 0
                                   Aerocondor ANA All Nippon Airways
 1
                        410
                                   Aerocondor ANA All Nippon Airways
 2
                        410
                                   Aerocondor ANA All Nippon Airways
 3
                                   Aerocondor ANA All Nippon Airways
                        410
                                   Aerocondor ANA All Nippon Airways
 4
                        410
 67658
                       4178
                             Regional Express
                                                        Qantas Airways
 67659
                      19016
                                   Apache Air
                                                                Apache
 67660
                                   Apache Air
                                                                Apache
                      19016
 67661
                                   Apache Air
                                                                Apache
                      19016
 67662
                      19016
                                   Apache Air
                                                                Apache
       airline_iata airline_icao airline_callsign airline_country
 0
                  2B
                              AR.D
                                        AEROCONDOR
                                                           Portugal
 1
                  2B
                              ARD
                                        AEROCONDOR
                                                           Portugal
 2
                  2B
                              ARD
                                        AEROCONDOR
                                                           Portugal
 3
                              ARD
                                         AEROCONDOR
                                                           Portugal
                  2B
 4
                  2B
                              ARD
                                        AEROCONDOR
                                                           Portugal
 67658
                  ZL
                              RXA
                                                REX
                                                          Australia
 67659
                  ZM
                              IWA
                                             APACHE
                                                      United States
                                                      United States
                  ZM
                              IWA
                                             APACHE
 67660
 67661
                  ZM
                              IWA
                                             APACHE
                                                      United States
                                                      United States
 67662
                  ZM
                              IWA
                                             APACHE
                         src_airport_airport_id
        airline_active
```

2965.0

0

True

```
1
                  True
                                          2966.0
2
                  True
                                          2966.0
3
                  True
                                          2968.0
4
                                          2968.0
                  True
67658
                                          6334.0
                  True
67659
                  True
                                          4029.0
67660
                  True
                                          2912.0
67661
                  True
                                          2912.0
67662
                  True
                                          2913.0
                         src_airport_name
                                            ... dst_airport_longitude
0
             Sochi International Airport
                                                           49.278702
1
                                                           49.278702
                        Astrakhan Airport
2
                        Astrakhan Airport
                                                           43.081902
3
          Chelyabinsk Balandino Airport
                                                           49.278702
4
          Chelyabinsk Balandino Airport
                                                           82.650703
67658
                         Whyalla Airport
                                                          138.531006
67659
       Domodedovo International Airport
                                                           74.477600
67660
             Manas International Airport
                                                           37.906300
             Manas International Airport
67661
                                                           72.793297
67662
                              Osh Airport
                                                           74.477600
      dst_airport_altitude dst_airport_timezone dst_airport_dst
0
                      411.0
                                               3.0
                                                                   N
1
                      411.0
                                               3.0
                                                                   N
2
                     1054.0
                                               3.0
                                                                   N
3
                                               3.0
                      411.0
                                                                   N
4
                      365.0
                                               7.0
                                                                   N
67658
                       20.0
                                               9.5
                                                                   0
                                                                   U
67659
                     2058.0
                                               6.0
67660
                                               3.0
                                                                   N
                      588.0
                                                                   U
67661
                     2927.0
                                               6.0
                                                                   U
67662
                     2058.0
                                               6.0
                                                dst_airport_source
                                                                      codeshare
        dst_airport_tz_id
                             dst_airport_type
0
                                                        OurAirports
                                                                           False
             Europe/Moscow
                                       airport
                                                                          False
1
             Europe/Moscow
                                       airport
                                                        OurAirports
2
             Europe/Moscow
                                                        OurAirports
                                                                          False
                                       airport
3
             Europe/Moscow
                                                        OurAirports
                                                                          False
                                       airport
4
         Asia/Krasnoyarsk
                                                        OurAirports
                                                                          False
                                       airport
67658
       Australia/Adelaide
                                       airport
                                                        OurAirports
                                                                          False
67659
              Asia/Bishkek
                                       airport
                                                        OurAirports
                                                                          False
67660
             Europe/Moscow
                                       airport
                                                        OurAirports
                                                                          False
67661
              Asia/Bishkek
                                       airport
                                                        OurAirports
                                                                          False
```

67662	Asia/Bishkek		airport	OurAirports	False
ed	quipment	key			
0	[CR2]	AERKZN2B			
1	[CR2]	ASFKZN2B			
2	[CR2]	ASFMRV2B			
3	[CR2]	CEKKZN2B			
4	[CR2]	CEKOVB2B			
	•••	•••			
67658	[SF3]	WYAADLZL			
67659	[734]	DMEFRUZM			
67660	[734]	FRUDMEZM			
67661	[734]	FRUOSSZM			
67662	[734]	OSSFRUZM			
[67663 rows x 39 columns]					

Exception ignored in: 'pyarrow._dataset._make_file_source'
Traceback (most recent call last):

File "/opt/conda/lib/python3.8/site-packages/pyarrow/parquet.py", line 1514, in __init__

fragment = parquet_format.make_fragment(single_file, filesystem) TypeError: cannot construct a FileSource from airline_airline_id airline name airline alias \ 410 Aerocondor ANA All Nippon Airways 1 410 Aerocondor ANA All Nippon Airways 2 410 ANA All Nippon Airways Aerocondor 3 ANA All Nippon Airways 410 Aerocondor 4 410 Aerocondor ANA All Nippon Airways 67658 4178 Regional Express Qantas Airways 67659 19016 Apache Air Apache 67660 19016 Apache Air Apache 67661 19016 Apache Air Apache 67662 Apache Air Apache 19016 airline_iata airline_icao airline_callsign airline_country 0 2B ARD AEROCONDOR Portugal 1 2B ARD **AEROCONDOR** Portugal 2 2B ARD Portugal **AEROCONDOR** 2B ARD 3 **AEROCONDOR** Portugal 4 2B ARD **AEROCONDOR** Portugal 67658 ZLRXA REX Australia 67659 ZMIWA **APACHE** United States United States 67660 ZMIWA **APACHE** 67661 ZMIWA APACHE United States

```
67662
                 ZM
                              IWA
                                             APACHE
                                                      United States
       airline_active
                       src_airport_airport_id \
0
                  True
                                         2965.0
1
                  True
                                         2966.0
2
                  True
                                         2966.0
3
                  True
                                         2968.0
4
                  True
                                         2968.0
67658
                  True
                                         6334.0
67659
                  True
                                         4029.0
                  True
67660
                                         2912.0
                  True
67661
                                         2912.0
                  True
67662
                                         2913.0
                        src_airport_name ... dst_airport_longitude
0
            Sochi International Airport
                                                           49.278702
1
                       Astrakhan Airport
                                                           49.278702
2
                       Astrakhan Airport
                                                          43.081902
3
          Chelyabinsk Balandino Airport
                                                           49.278702
          Chelyabinsk Balandino Airport
4
                                                           82.650703
                                                             •••
67658
                         Whyalla Airport
                                                          138.531006
67659
       Domodedovo International Airport
                                                          74.477600
67660
            Manas International Airport
                                                          37.906300
            Manas International Airport
67661
                                                          72.793297
67662
                              Osh Airport
                                                          74.477600
      dst_airport_altitude dst_airport_timezone dst_airport_dst
0
                      411.0
                                               3.0
                                               3.0
1
                      411.0
                                                                  N
2
                     1054.0
                                               3.0
                                                                  N
3
                      411.0
                                               3.0
                                                                  N
4
                      365.0
                                               7.0
                                                                  N
67658
                       20.0
                                               9.5
                                                                  0
                                               6.0
                                                                  U
67659
                     2058.0
67660
                      588.0
                                               3.0
                                                                  N
67661
                     2927.0
                                               6.0
                                                                  U
67662
                     2058.0
                                               6.0
                                                                  IJ
        dst_airport_tz_id
                            dst_airport_type
                                               dst_airport_source
                                                                     codeshare
0
             Europe/Moscow
                                                       OurAirports
                                      airport
                                                                         False
1
             Europe/Moscow
                                                       OurAirports
                                      airport
                                                                         False
2
            Europe/Moscow
                                      airport
                                                       OurAirports
                                                                         False
3
             Europe/Moscow
                                      airport
                                                       OurAirports
                                                                         False
4
         Asia/Krasnoyarsk
                                      airport
                                                       OurAirports
                                                                         False
```

```
67658 Australia/Adelaide
                                     airport
                                                      OurAirports
                                                                        False
67659
             Asia/Bishkek
                                                                        False
                                     airport
                                                      OurAirports
67660
            Europe/Moscow
                                     airport
                                                      OurAirports
                                                                        False
67661
             Asia/Bishkek
                                     airport
                                                      OurAirports
                                                                        False
67662
             Asia/Bishkek
                                                      OurAirports
                                     airport
                                                                        False
      equipment
                      key
0
          [CR2]
                 AERKZN2B
1
          [CR2]
                ASFKZN2B
2
          [CR2] ASFMRV2B
3
          [CR2] CEKKZN2B
4
                CEKOVB2B
          [CR2]
67658
          [SF3]
                 WYAADLZL
67659
          [734]
                 DMEFRUZM
67660
          [734] FRUDMEZM
67661
          [734]
                 FRUOSSZM
67662
          [734] OSSFRUZM
[67663 rows x 39 columns]
```

```
ArrowInvalid
                                          Traceback (most recent call last)
<ipython-input-56-23d0ab08b002> in <module>
     74 df['key'] = df['src_airport_iata'].astype(str) + df['dst_airport_iata']
→astype(str) + df['airline_iata'].astype(str)
---> 76 ndf = pd.read parquet(df, engine='pyarrow')
    77 print(list(ndf.columns.values))
     78
/opt/conda/lib/python3.8/site-packages/pandas/io/parquet.py in |
 →read_parquet(path, engine, columns, use_nullable_dtypes, **kwargs)
    457
    458
            impl = get_engine(engine)
            return impl.read(
--> 459
    460
                path, columns=columns, use_nullable_dtypes=use_nullable_dtypes,
→**kwargs
            )
    461
/opt/conda/lib/python3.8/site-packages/pandas/io/parquet.py in read(self, path,
 →columns, use nullable dtypes, storage options, **kwargs)
    219
                )
    220
                try:
                    return self.api.parquet.read_table(
--> 221
    222
                        path_or_handle, columns=columns, **kwargs
    223
                    ).to_pandas(**to_pandas_kwargs)
```

```
/opt/conda/lib/python3.8/site-packages/pyarrow/parquet.py in read_table(source,
  →columns, use_threads, metadata, use_pandas_metadata, memory_map, use_thread_dictionary, filesystem, filters, buffer_size, partitioning, use_thread_dictionary, filesystem, f
  →use_legacy_dataset, ignore_prefixes)
         1670
         1671
                                               try:
-> 1672
                                                           dataset = _ParquetDatasetV2(
         1673
                                                                       source.
         1674
                                                                       filesystem=filesystem,
/opt/conda/lib/python3.8/site-packages/pyarrow/parquet.py in __init__(self,_
   →path_or_paths, filesystem, filters, partitioning, read_dictionary,
  →buffer size, memory map, ignore prefixes, **kwargs)
         1515
                                                           self. dataset = ds.FileSystemDataset(
         1516
-> 1517
                                                                        [fragment], schema=fragment.physical_schema,
         1518
                                                                       format=parquet format,
         1519
                                                                       filesystem=fragment.filesystem
/opt/conda/lib/python3.8/site-packages/pyarrow/_dataset.pyx in pyarrow._dataset
  →Fragment.physical_schema.__get__()
/opt/conda/lib/python3.8/site-packages/pyarrow/error.pxi in pyarrow.lib.
   →pyarrow internal check status()
/opt/conda/lib/python3.8/site-packages/pyarrow/error.pxi in pyarrow.lib.
  →check status()
ArrowInvalid: Called Open() on an uninitialized FileSource
```

```
[57]: #Assignment 7B#
import hashlib

def hash_key(key):
    m = hashlib.sha256()
    m.update(str(key).encode('utf-8'))
    return m.hexdigest()

ndf = pd.read_parquet(df, engine='pyarrow')

ndf['key'] = ndf['src_airport.iata']+ndf['dst_airport.iata']+ndf['airline.icao']
    ndf['hashed'] = ndf.apply(lambda x: hash_key(x))
    ndf['hash_key'] = ndf['hashed'].str[:1]

# For troubleshooting
table = pa.Table.from_pandas(ndf)
```

```
ndftable = parq.read_table(partitioned_parquet_file)
print(ndftable)
```

```
Traceback (most recent call last)
TypeError
TypeError: cannot construct a FileSource from
                                                        airline_airline_id
→airline_name
                           airline_alias \
0
                       410
                                  Aerocondor ANA All Nippon Airways
1
                       410
                                  Aerocondor ANA All Nippon Airways
2
                       410
                                  Aerocondor ANA All Nippon Airways
3
                       410
                                  Aerocondor ANA All Nippon Airways
4
                       410
                                  Aerocondor ANA All Nippon Airways
                            Regional Express
67658
                      4178
                                                        Qantas Airways
                                  Apache Air
67659
                     19016
                                                                Apache
67660
                     19016
                                  Apache Air
                                                                Apache
                     19016
                                  Apache Air
                                                                Apache
67661
67662
                     19016
                                  Apache Air
                                                                Apache
      airline_iata airline_icao airline_callsign airline_country
0
                2B
                             ARD
                                        AEROCONDOR
                                                           Portugal
                2B
                             ARD
1
                                        AEROCONDOR
                                                           Portugal
2
                             ARD
                                                           Portugal
                2B
                                        AEROCONDOR
                             ARD
3
                2B
                                        AEROCONDOR
                                                           Portugal
                             ARD
                                        AEROCONDOR
4
                2B
                                                           Portugal
67658
                             RXA
                                               REX
                                                          Australia
                ZL
                             IWA
                                                     United States
67659
                ZM
                                            APACHE
                                                     United States
67660
                ZM
                             IWA
                                            APACHE
                                                     United States
67661
                ZM
                             IWA
                                            APACHE
                                                     United States
67662
                ZM
                             IWA
                                            APACHE
                        src_airport_airport_id
       airline_active
0
                  True
                                         2965.0
                  True
1
                                         2966.0
2
                  True
                                         2966.0
3
                  True
                                         2968.0
4
                                         2968.0
                  True
                                         6334.0
67658
                  True
67659
                 True
                                         4029.0
67660
                 True
                                         2912.0
67661
                 True
                                         2912.0
                  True
                                         2913.0
67662
                        src_airport_name ... dst_airport_longitude \
```

```
0
             Sochi International Airport
                                                           49.278702
1
                        Astrakhan Airport
                                                           49.278702
2
                                                           43.081902
                        Astrakhan Airport
3
          Chelyabinsk Balandino Airport
                                                           49.278702
4
          Chelyabinsk Balandino Airport
                                                           82.650703
67658
                          Whyalla Airport
                                                          138.531006
67659
       Domodedovo International Airport
                                                           74.477600
67660
             Manas International Airport
                                                           37.906300
67661
             Manas International Airport
                                                           72.793297
67662
                                                           74.477600
                              Osh Airport
      dst_airport_altitude dst_airport_timezone dst_airport_dst
0
                      411.0
                                               3.0
                                                                   N
1
                      411.0
                                               3.0
                                                                   N
2
                     1054.0
                                               3.0
                                                                   N
3
                      411.0
                                               3.0
                                                                   N
4
                      365.0
                                               7.0
                                                                   N
67658
                       20.0
                                               9.5
                                                                   0
67659
                     2058.0
                                               6.0
                                                                   U
                                               3.0
67660
                      588.0
                                                                   N
67661
                     2927.0
                                               6.0
                                                                   U
67662
                     2058.0
                                               6.0
                                                                   U
                                                                      codeshare
        dst_airport_tz_id
                             dst_airport_type
                                                dst_airport_source
0
             Europe/Moscow
                                                        OurAirports
                                                                           False
                                       airport
1
             Europe/Moscow
                                       airport
                                                        OurAirports
                                                                           False
2
                                                        OurAirports
                                                                           False
             Europe/Moscow
                                       airport
3
             Europe/Moscow
                                       airport
                                                        OurAirports
                                                                           False
4
                                                        OurAirports
                                                                           False
         Asia/Krasnoyarsk
                                       airport
67658
       Australia/Adelaide
                                       airport
                                                        OurAirports
                                                                          False
67659
              Asia/Bishkek
                                                        OurAirports
                                                                          False
                                       airport
                                       airport
                                                        OurAirports
                                                                          False
67660
             Europe/Moscow
67661
              Asia/Bishkek
                                       airport
                                                        OurAirports
                                                                          False
67662
              Asia/Bishkek
                                       airport
                                                        OurAirports
                                                                          False
      equipment
                       key
0
           [CR2]
                  AERKZN2B
1
           [CR2]
                  ASFKZN2B
2
           [CR2]
                  ASFMRV2B
3
           [CR2]
                  CEKKZN2B
4
           [CR2]
                  CEKOVB2B
67658
           [SF3]
                  WYAADLZL
67659
           [734]
                  DMEFRUZM
67660
           [734]
                  FRUDMEZM
```

```
67662
            [734]
                   OSSFRUZM
 [67663 rows x 39 columns]
Exception ignored in: 'pyarrow._dataset._make_file_source'
Traceback (most recent call last):
  File "/opt/conda/lib/python3.8/site-packages/pyarrow/parquet.py", line 1514,
in __init__
    fragment = parquet_format.make_fragment(single_file, filesystem)
TypeError: cannot construct a FileSource from
                                                        airline airline id
airline_name
                        airline_alias \
                       410
                                   Aerocondor
                                               ANA All Nippon Airways
1
                       410
                                   Aerocondor
                                               ANA All Nippon Airways
2
                       410
                                               ANA All Nippon Airways
                                   Aerocondor
3
                       410
                                   Aerocondor
                                               ANA All Nippon Airways
4
                       410
                                                ANA All Nippon Airways
                                   Aerocondor
67658
                      4178
                            Regional Express
                                                        Qantas Airways
                                   Apache Air
67659
                     19016
                                                                 Apache
67660
                     19016
                                   Apache Air
                                                                 Apache
67661
                     19016
                                   Apache Air
                                                                 Apache
67662
                     19016
                                   Apache Air
                                                                 Apache
      airline_iata airline_icao airline_callsign airline_country
0
                 2B
                              ARD
                                        AEROCONDOR
                                                           Portugal
1
                 2B
                              ARD
                                        AEROCONDOR
                                                           Portugal
2
                 2B
                              ARD
                                        AEROCONDOR
                                                           Portugal
3
                 2B
                              AR.D
                                        AEROCONDOR
                                                           Portugal
                 2B
                              ARD
                                                           Portugal
                                        AEROCONDOR
67658
                 ZL
                             RXA
                                                          Australia
                                                REX
67659
                 ZM
                              IWA
                                            APACHE
                                                      United States
67660
                 ZM
                              IWA
                                            APACHE
                                                      United States
                 ZM
                                                      United States
67661
                              IWA
                                            APACHE
67662
                 ZM
                              IWA
                                            APACHE
                                                      United States
       airline_active
                        src_airport_airport_id
0
                  True
                                         2965.0
1
                  True
                                         2966.0
2
                  True
                                         2966.0
3
                  True
                                         2968.0
4
                  True
                                         2968.0
                                         6334.0
67658
                  True
                                         4029.0
67659
                  True
67660
                  True
                                         2912.0
```

67661

[734]

FRUOSSZM

```
67661
                  True
                                          2912.0
67662
                  True
                                          2913.0
                                            ... dst_airport_longitude
                         src_airport_name
             Sochi International Airport
                                                           49.278702
0
1
                        Astrakhan Airport
                                                           49.278702
2
                        Astrakhan Airport
                                                           43.081902
          Chelyabinsk Balandino Airport
3
                                                           49.278702
4
          Chelyabinsk Balandino Airport
                                                           82.650703
67658
                          Whyalla Airport
                                                          138.531006
       Domodedovo International Airport
67659
                                                           74.477600
67660
             Manas International Airport
                                                           37.906300
             Manas International Airport
67661
                                                           72.793297
67662
                              Osh Airport
                                                           74.477600
      dst_airport_altitude dst_airport_timezone dst_airport_dst
0
                      411.0
                                                3.0
1
                       411.0
                                                3.0
                                                                   N
2
                      1054.0
                                               3.0
                                                                   N
3
                      411.0
                                                3.0
                                                                   N
4
                       365.0
                                               7.0
                                                                   N
67658
                        20.0
                                               9.5
                                                                   0
67659
                     2058.0
                                               6.0
                                                                   IJ
                                                3.0
                                                                   N
67660
                      588.0
67661
                     2927.0
                                                6.0
                                                                   U
                                                6.0
67662
                     2058.0
                                                                   U
        dst_airport_tz_id
                             dst_airport_type
                                                dst_airport_source
                                                                      codeshare
0
             Europe/Moscow
                                                        OurAirports
                                                                           False
                                       airport
1
             Europe/Moscow
                                       airport
                                                        OurAirports
                                                                          False
2
             Europe/Moscow
                                       airport
                                                        OurAirports
                                                                          False
3
             Europe/Moscow
                                       airport
                                                        OurAirports
                                                                          False
4
         Asia/Krasnovarsk
                                                        OurAirports
                                                                          False
                                       airport
       Australia/Adelaide
                                                                          False
67658
                                       airport
                                                        OurAirports
              Asia/Bishkek
67659
                                       airport
                                                        OurAirports
                                                                          False
67660
             Europe/Moscow
                                       airport
                                                        OurAirports
                                                                          False
              Asia/Bishkek
                                                        OurAirports
                                                                          False
67661
                                       airport
67662
              Asia/Bishkek
                                       airport
                                                        OurAirports
                                                                          False
      equipment
                        key
0
           [CR2]
                  AERKZN2B
1
           [CR2]
                  ASFKZN2B
2
           [CR2]
                  ASFMRV2B
3
           [CR2]
                  CEKKZN2B
4
           [CR2]
                  CEKOVB2B
```

[67663 rows x 39 columns]

```
ArrowInvalid
                                              Traceback (most recent call last)
<ipython-input-57-985032d9afbc> in <module>
             return m.hexdigest()
----> 9 ndf = pd.read parquet(df, engine='pyarrow')
     11 ndf['key'] = ndf['src airport.iata']+ndf['dst airport.
 →iata']+ndf['airline.icao']
/opt/conda/lib/python3.8/site-packages/pandas/io/parquet.py in_
 →read parquet(path, engine, columns, use nullable dtypes, **kwargs)
             11 11 11
    457
    458
             impl = get_engine(engine)
--> 459
             return impl.read(
    460
                 path, columns=columns, use_nullable_dtypes=use_nullable_dtypes,
 →**kwargs
    461
             )
/opt/conda/lib/python3.8/site-packages/pandas/io/parquet.py in read(self, path,
 →columns, use_nullable_dtypes, storage_options, **kwargs)
    219
                 )
    220
                 try:
--> 221
                     return self.api.parquet.read_table(
    222
                          path_or_handle, columns=columns, **kwargs
    223
                      ).to pandas(**to pandas kwargs)
/opt/conda/lib/python3.8/site-packages/pyarrow/parquet.py in read_table(source,
→columns, use_threads, metadata, use_pandas_metadata, memory_map, 

→read_dictionary, filesystem, filters, buffer_size, partitioning, use_threads.
 →use_legacy_dataset, ignore_prefixes)
   1670
   1671
                 try:
-> 1672
                     dataset = _ParquetDatasetV2(
   1673
                          source.
   1674
                          filesystem=filesystem,
```

```
/opt/conda/lib/python3.8/site-packages/pyarrow/parquet.py in __init__(self,_
       →path_or_paths, filesystem, filters, partitioning, read_dictionary, ___
       →buffer_size, memory_map, ignore_prefixes, **kwargs)
         1515
                          self. dataset = ds.FileSystemDataset(
         1516
      -> 1517
                               [fragment], schema=fragment.physical_schema,
                              format=parquet format,
         1518
         1519
                              filesystem=fragment.filesystem
      /opt/conda/lib/python3.8/site-packages/pyarrow/_dataset.pyx in pyarrow._dataset
       →Fragment.physical_schema.__get__()
      /opt/conda/lib/python3.8/site-packages/pyarrow/error.pxi in pyarrow.lib.
       →pyarrow_internal_check_status()
      /opt/conda/lib/python3.8/site-packages/pyarrow/error.pxi in pyarrow.lib.
       →check_status()
      ArrowInvalid: Called Open() on an uninitialized FileSource
[58]: #Assignment 7C#
      import pygeohash
      West = pygeohash.encode(45.5945645, -121.1786823)
      print(West)
      Central = pygeohash.encode(41.1544433, -96.0422378)
      print(Central)
      East = pygeohash.encode(39.08344, -77.6497145)
      print(East)
     c21g6s0rs4c7
     9z7dnebnj8kb
     dqby34cjw922
[60]: | ndf = pd.read_parquet(df, engine='pyarrow')
      ndf['key'] = ndf['src_airport.iata']+ndf['dst_airport.iata']+ndf['airline.icao']
      ndf['long'] = ndf['src_airport.longitude']
      ndf['lat'] = ndf['src_airport.latitude']
      ndf['location'] = ndf.apply(lambda x: get data center val(x.lat, x.long, West, ...
      ntable = pa.Table.from_pandas(ndf)
      print(ntable)
```

```
TypeError
                                             Traceback (most recent call last)
                                                        airline_airline_id
TypeError: cannot construct a FileSource from
→airline_name
                           airline_alias \
0
                       410
                                   Aerocondor
                                                ANA All Nippon Airways
1
                       410
                                   Aerocondor
                                                ANA All Nippon Airways
2
                       410
                                                ANA All Nippon Airways
                                   Aerocondor
3
                       410
                                   Aerocondor
                                                ANA All Nippon Airways
4
                       410
                                   Aerocondor
                                                ANA All Nippon Airways
                            Regional Express
67658
                      4178
                                                         Qantas Airways
67659
                     19016
                                   Apache Air
                                                                 Apache
67660
                     19016
                                   Apache Air
                                                                 Apache
67661
                     19016
                                   Apache Air
                                                                 Apache
                                   Apache Air
67662
                     19016
                                                                 Apache
      airline_iata airline_icao airline_callsign airline_country
0
                 2B
                              ARD
                                         AEROCONDOR
                                                            Portugal
                 2B
                              ARD
1
                                        AEROCONDOR
                                                            Portugal
2
                 2B
                              AR.D
                                         AEROCONDOR
                                                            Portugal
3
                 2B
                              ARD
                                         AEROCONDOR
                                                            Portugal
4
                 2B
                              ARD
                                         AEROCONDOR
                                                            Portugal
67658
                 ZL
                              RXA
                                                REX
                                                           Australia
67659
                 ZM
                              IWA
                                             APACHE
                                                      United States
67660
                 ZM
                              IWA
                                             APACHE
                                                      United States
67661
                 ZM
                              IWA
                                             APACHE
                                                      United States
67662
                 ZM
                              IWA
                                             APACHE
                                                      United States
                        src airport airport id
       airline_active
0
                  True
                                          2965.0
1
                  True
                                          2966.0
2
                  True
                                         2966.0
3
                  True
                                         2968.0
4
                                         2968.0
                  True
67658
                                         6334.0
                  True
67659
                  True
                                         4029.0
67660
                  True
                                         2912.0
67661
                  True
                                         2912.0
67662
                  True
                                         2913.0
                        src_airport_name ... dst_airport_longitude
0
            Sochi International Airport
                                                           49.278702
1
                       Astrakhan Airport
                                                           49.278702
2
                       Astrakhan Airport
                                                           43.081902
3
          Chelyabinsk Balandino Airport
                                                           49.278702
```

```
4
          Chelyabinsk Balandino Airport
                                                           82.650703
67658
                         Whyalla Airport
                                                          138.531006
67659
       Domodedovo International Airport
                                                           74.477600
67660
            Manas International Airport
                                                           37.906300
67661
            Manas International Airport
                                                           72.793297
67662
                              Osh Airport
                                                           74.477600
      dst_airport_altitude dst_airport_timezone dst_airport_dst
0
                      411.0
                                               3.0
1
                      411.0
                                               3.0
                                                                  N
2
                     1054.0
                                               3.0
                                                                  N
3
                                               3.0
                                                                  N
                      411.0
4
                      365.0
                                               7.0
                                                                  N
67658
                       20.0
                                               9.5
                                                                  0
67659
                     2058.0
                                               6.0
                                                                  U
                                               3.0
                                                                  N
67660
                      588.0
67661
                     2927.0
                                               6.0
                                                                  IJ
                                                                  IJ
67662
                     2058.0
                                               6.0
        dst airport tz id
                                                                     codeshare
                            dst_airport_type
                                                dst airport source
            Europe/Moscow
                                                        OurAirports
                                                                          False
0
                                      airport
                                                                          False
1
            Europe/Moscow
                                      airport
                                                       OurAirports
2
            Europe/Moscow
                                      airport
                                                        OurAirports
                                                                          False
3
            Europe/Moscow
                                      airport
                                                       OurAirports
                                                                          False
4
         Asia/Krasnoyarsk
                                      airport
                                                        OurAirports
                                                                          False
                                      •••
67658
       Australia/Adelaide
                                                                          False
                                      airport
                                                        OurAirports
67659
              Asia/Bishkek
                                      airport
                                                        OurAirports
                                                                          False
                                                                          False
67660
            Europe/Moscow
                                      airport
                                                       OurAirports
67661
              Asia/Bishkek
                                      airport
                                                        OurAirports
                                                                          False
67662
             Asia/Bishkek
                                                                          False
                                      airport
                                                        OurAirports
      equipment
                       key
           [CR2]
                  AERKZN2B
0
1
           [CR2]
                  ASFKZN2B
2
           [CR2]
                  ASFMRV2B
3
           [CR2]
                  CEKKZN2B
4
           [CR2]
                  CEKOVB2B
67658
           [SF3]
                  WYAADLZL
67659
           [734]
                  DMEFRUZM
67660
           [734]
                  FRUDMEZM
67661
           [734]
                  FRUOSSZM
           [734]
67662
                  OSSFRUZM
```

```
Exception ignored in: 'pyarrow._dataset._make_file_source'
Traceback (most recent call last):
 File "/opt/conda/lib/python3.8/site-packages/pyarrow/parquet.py", line 1514,
in __init__
    fragment = parquet_format.make_fragment(single_file, filesystem)
TypeError: cannot construct a FileSource from
                                                        airline_airline_id
airline name
                        airline_alias \
                       410
                                   Aerocondor
                                               ANA All Nippon Airways
1
                                              ANA All Nippon Airways
                       410
                                   Aerocondor
2
                                               ANA All Nippon Airways
                       410
                                   Aerocondor
3
                                               ANA All Nippon Airways
                       410
                                   Aerocondor
4
                       410
                                   Aerocondor
                                               ANA All Nippon Airways
67658
                      4178
                            Regional Express
                                                        Qantas Airways
67659
                     19016
                                   Apache Air
                                                                Apache
67660
                     19016
                                   Apache Air
                                                                Apache
67661
                                   Apache Air
                                                                Apache
                     19016
67662
                                   Apache Air
                     19016
                                                                Apache
      airline_iata airline_icao airline_callsign airline_country
0
                             ARD
                                        AEROCONDOR
                2B
                                                           Portugal
                2B
                             ARD
1
                                        AEROCONDOR
                                                           Portugal
2
                2B
                             ARD
                                        AEROCONDOR
                                                           Portugal
3
                2B
                             ARD
                                        AEROCONDOR
                                                           Portugal
                2B
                             ARD
                                        AEROCONDOR
                                                           Portugal
4
67658
                ZL
                             RXA
                                               REX
                                                          Australia
67659
                ZM
                                                      United States
                             IWA
                                            APACHE
67660
                ZM
                             IWA
                                            APACHE
                                                      United States
67661
                ZM
                             IWA
                                            APACHE
                                                      United States
                                                      United States
67662
                ZM
                             IWA
                                            APACHE
       airline_active
                        src_airport_airport_id
0
                  True
                                         2965.0
1
                 True
                                         2966.0
2
                                         2966.0
                 True
3
                 True
                                         2968.0
                                         2968.0
4
                 True
67658
                  True
                                         6334.0
                  True
                                         4029.0
67659
67660
                 True
                                         2912.0
67661
                 True
                                         2912.0
67662
                 True
                                         2913.0
```

```
... dst_airport_longitude
                         src_airport_name
0
            Sochi International Airport
                                                           49.278702
1
                       Astrakhan Airport
                                                           49.278702
2
                       Astrakhan Airport
                                                           43.081902
          Chelyabinsk Balandino Airport
3
                                                           49.278702
4
          Chelyabinsk Balandino Airport
                                                           82.650703
67658
                          Whyalla Airport
                                                          138.531006
67659
       Domodedovo International Airport
                                                           74.477600
            Manas International Airport
67660
                                                           37.906300
67661
                                                           72.793297
            Manas International Airport
67662
                              Osh Airport
                                                           74.477600
      dst_airport_altitude dst_airport_timezone dst_airport_dst
0
                      411.0
                                                3.0
                                                                   N
1
                      411.0
                                                3.0
                                                                   N
2
                     1054.0
                                                3.0
                                                                   N
3
                      411.0
                                                3.0
                                                                   N
4
                      365.0
                                               7.0
                                                                   N
67658
                       20.0
                                                9.5
                                                                   0
                                                6.0
                                                                   U
67659
                     2058.0
67660
                      588.0
                                                3.0
                                                                   N
67661
                     2927.0
                                                6.0
                                                                   U
67662
                     2058.0
                                                6.0
                                                                   IJ
                                                dst_airport_source
        dst_airport_tz_id
                             dst_airport_type
                                                                      codeshare
0
             Europe/Moscow
                                       airport
                                                        OurAirports
                                                                           False
1
                                                        OurAirports
             Europe/Moscow
                                       airport
                                                                           False
2
             Europe/Moscow
                                       airport
                                                        OurAirports
                                                                           False
3
             Europe/Moscow
                                                        OurAirports
                                                                           False
                                       airport
4
         Asia/Krasnoyarsk
                                       airport
                                                        OurAirports
                                                                           False
       Australia/Adelaide
67658
                                       airport
                                                        OurAirports
                                                                           False
              Asia/Bishkek
                                                        OurAirports
67659
                                       airport
                                                                           False
67660
             Europe/Moscow
                                       airport
                                                        OurAirports
                                                                           False
              Asia/Bishkek
67661
                                       airport
                                                        OurAirports
                                                                           False
67662
              Asia/Bishkek
                                       airport
                                                        OurAirports
                                                                           False
      equipment
                       key
0
           [CR2]
                  AERKZN2B
1
           [CR2]
                  ASFKZN2B
2
           [CR2]
                  ASFMRV2B
3
           [CR2]
                  CEKKZN2B
4
           [CR2]
                  CEKOVB2B
67658
           [SF3]
                  WYAADLZL
67659
           [734]
                  DMEFRUZM
```

```
67660 [734] FRUDMEZM
67661 [734] FRUDSSZM
67662 [734] OSSFRUZM
```

[67663 rows x 39 columns]

```
ArrowInvalid
                                                                                                           Traceback (most recent call last)
<ipython-input-60-e699375a279a> in <module>
----> 1 ndf = pd.read_parquet(df, engine='pyarrow')
               2
               4 ndf['key'] = ndf['src airport.iata']+ndf['dst airport.
  →iata']+ndf['airline.icao']
               5 ndf['long'] = ndf['src airport.longitude']
/opt/conda/lib/python3.8/site-packages/pandas/io/parquet.py in__
  →read parquet(path, engine, columns, use nullable dtypes, **kwargs)
          457
          458
                              impl = get_engine(engine)
--> 459
                              return impl.read(
                                        path, columns=columns, use nullable dtypes=use nullable dtypes,
          460
  →**kwargs
          461
                              )
/opt/conda/lib/python3.8/site-packages/pandas/io/parquet.py in read(self, path,
  →columns, use nullable dtypes, storage options, **kwargs)
          219
                                        )
          220
                                        try:
--> 221
                                                  return self.api.parquet.read_table(
          222
                                                             path_or_handle, columns=columns, **kwargs
          223
                                                   ).to_pandas(**to_pandas_kwargs)
/opt/conda/lib/python3.8/site-packages/pyarrow/parquet.py in read table(source,
  →columns, use_threads, metadata, use_pandas_metadata, memory_map, use_thread_dictionary, filesystem, filters, buffer_size, partitioning, use_thread_dictionary, filesystem, f
  →use_legacy_dataset, ignore_prefixes)
       1670
       1671
                                        try:
-> 1672
                                                  dataset = _ParquetDatasetV2(
       1673
                                                             source,
       1674
                                                             filesystem=filesystem,
/opt/conda/lib/python3.8/site-packages/pyarrow/parquet.py in __init__(self,_
  →path_or_paths, filesystem, filters, partitioning, read_dictionary, ___
  →buffer size, memory map, ignore prefixes, **kwargs)
       1515
```

```
[113]: #Assignment 7D#
       def balance_partitions (keys, num_partitions):
           unique = sorted(set(keys))
           numbervals = len(unique)
           count = (numbervals / num_partitions)+1
           partitions = []
           parta = 10
           partb = 10
           for i in range(numbervals):
               keyval ={}
               if parta <= count:</pre>
                   keyval[unique[i]] = partb
                   parta = parta + 1
               else:
                   parta = 1
                   partb = partb + 1
                   keyval[unique[i]] = partb
                   parta = parta + 1
               partitions .append(keyval)
           return partitions
```

```
[114]: keys = ['apple', 'grape', 'orange', 'banana']
   nummberpartitions = 5
   partitions = balance_partitions(keys, nummberpartitions)
   print(partitions)

keys = ['green', 'red', 'orange', 'purple']
   nummberpartitions = 4
   partitions = balance_partitions(keys, nummberpartitions)
   print(partitions)
```

```
[{'apple': 11}, {'banana': 12}, {'grape': 13}, {'orange': 14}]
[{'green': 11}, {'orange': 11}, {'purple': 12}, {'red': 12}]
[]:
```