Dataset schema is:

```
root
 |-- Year: integer (nullable = true)
 |-- Month: integer (nullable = true)
 |-- DayofMonth: integer (nullable = true)
 |-- DayOfWeek: integer (nullable = true)
 |-- DepTime: string (nullable = true)
 |-- CRSDepTime: integer (nullable = true)
 |-- ArrTime: string (nullable = true)
 |-- CRSArrTime: integer (nullable = true)
 |-- UniqueCarrier: string (nullable = true)
 |-- FlightNum: integer (nullable = true)
 |-- TailNum: string (nullable = true)
 |-- ActualElapsedTime: string (nullable = true)
 |-- CRSElapsedTime: integer (nullable = true)
 |-- AirTime: string (nullable = true)
 |-- ArrDelay: string (nullable = true)
 |-- DepDelay: string (nullable = true)
 |-- Origin: string (nullable = true)
 |-- Dest: string (nullable = true)
 |-- Distance: integer (nullable = true)
 |-- TaxiIn: integer (nullable = true)
 |-- TaxiOut: integer (nullable = true)
 |-- Cancelled: integer (nullable = true)
 |-- CancellationCode: string (nullable = true)
 |-- Diverted: integer (nullable = true)
 |-- CarrierDelay: integer (nullable = true)
 |-- WeatherDelay: integer (nullable = true)
 |-- NASDelay: integer (nullable = true)
 |-- SecurityDelay: integer (nullable = true)
 |-- LateAircraftDelay: integer (nullable = true)
```

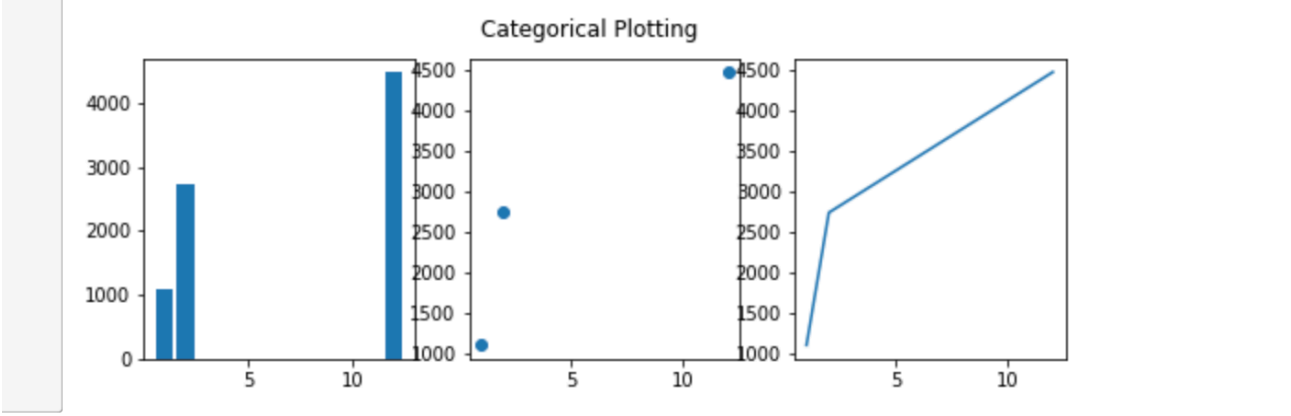
Sample of data:

```
myair_transits:
Row(Year=2007, Month=1, DayofMonth=1, DayOfWeek=1, DepTime='1232', CRSDepTime=1225, ArrTime='1341', CRSArrTime=1340,
UniqueCarrier='WN', FlightNum=2891, TailNum='N351', ActualElapsedTime='69', CRSElapsedTime=75, AirTime='54', ArrDelay
='1', DepDelay='7', Origin='SMF', Dest='ONT', Distance=389, TaxiIn=4, TaxiOut=11, Cancelled=0, CancellationCode=None,
Diverted=0, CarrierDelay=0, WeatherDelay=0, NASDelay=0, SecurityDelay=0, LateAircraftDelay=0)

Row(Year=2007, Month=1, DayofMonth=1, DayOfWeek=1, DepTime='1918', CRSDepTime=1905, ArrTime='2043', CRSArrTime=2035,
UniqueCarrier='WN', FlightNum=462, TailNum='N370', ActualElapsedTime='85', CRSElapsedTime=90, AirTime='74', ArrDelay
='8', DepDelay='13', Origin='SMF', Dest='PDX', Distance=479, TaxiIn=5, TaxiOut=6, Cancelled=0, CancellationCode=None,
Diverted=0, CarrierDelay=0, WeatherDelay=0, NASDelay=0, SecurityDelay=0, LateAircraftDelay=0)

Row(Year=2007, Month=1, DayofMonth=1, DayOfWeek=1, DepTime='2206', CRSDepTime=2130, ArrTime='2334', CRSArrTime=2300,
UniqueCarrier='WN', FlightNum=1229, TailNum='N685', ActualElapsedTime='88', CRSElapsedTime=90, AirTime='73', ArrDelay
='34', DepDelay='36', Origin='SMF', Dest='PDX', Distance=479, TaxiIn=6, TaxiOut=9, Cancelled=0, CancellationCode=Non
e, Diverted=0, CarrierDelay=3, WeatherDelay=0, NASDelay=0, SecurityDelay=0, LateAircraftDelay=31)
```

Explore data per month:



Find the plane with the highest number of flights. Each plane has a unique TailNum

```
print('Explore3: highest number of flights. Each plane has a unique TailNum')
df1 = df.filter(~col('TailNum').isin(['0', '000000']))
df1 = df1.groupBy("TailNum").count().sort('count',ascending=False)

if df1 is not None:
    for item in df1.rdd.collect()[:3]:
        print(item['TailNum'], item['count'])
    print('highest_number_filghts: ', df1.collect()[0]['TailNum'] )
```

```
Explore3: highest number of flights. Each plane has a unique TailNum
N912DE 99
N909DE 88
N911DE 88
highest_number_filghts:  N912DE
```

```
Compute the total flight time of each airplane, sorted by flight time in descending order.
total_flight_times = hiveCtx.sql("SELECT TailNum, SUM (AirTime) as total_fight_time FROM air_transit GROUP BY TailNum o
print('-'*10, 'We make ', '-'*10)
print('total flight time of each airplane: ')
for item in  total_flight_times.collect():
    print(item['TailNum'], ' total fight time is', item['total_fight_time'])

int('Q4:total flight time of each airplane, sorted by flight time in descending order')
1 = df.groupBy("TailNum").agg(F.sum(df['AirTime']).alias('result')).orderBy('result',ascending=False)
 df1 is not None:
  for item in df1.rdd.collect()[:3]:
      print(item['TailNum'], item['result'])
```

```
Q4:total flight time of each airplane, sorted by flight time in descending order
N3767 9977.0
N385DN 8844.0
N6707A 8173.0
```

Find the busiest airport (in terms of number of departures + arrivals of all operated flights)

```
1 ----------
{'SNA': 473, 'SMF': 473, 'STL': 308}
month_max: SNA  ,departures + arrivals of this month: 473
2 ----------
{'LAS': 1771, 'LAX': 1122}
month_max: LAS  ,departures + arrivals of this month: 1771
```

for each month.