Explanatory Data Analysis on 2018 US flight data

Out[57]:

The raw code for this IPython notebook is by default hidden for easier reading. To toggle on/off the raw code, click here.

```
(2319612, 19)

There is a total of 2319612 flights in the dataset.

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2319612 entries, 0 to 2319611
Data columns (total 19 columns):
Year                int64
Month               int64
DayofMonth          int64
DayOfWeek           int64
UniqueCarrier       object
Origin              object
Dest                object
Distance            int64
Cancelled           int64
CancellationCode    object
Code                object
Description         object
iata                object
airport             object
city                object
state               object
country             object
lat                 float64
long                float64
dtypes: float64(2), int64(6), object(11)
memory usage: 336.2+ MB
```

The data is composed of integers, objects and floats. The memory usage is 336.2+ MB.

Out[123]:

	Year	Month	DayofMonth	DayOfWeek	UniqueCarrier	Origin	Dest	Distance	Cancelle
0	2008	1	1	2	XE	EWR	MYR	550	
1	2008	1	1	2	XE	AUS	ONT	1197	
2	2008	1	1	2	XE	ONT	MCI	1318	
3	2008	1	1	2	XE	FAT	ONT	222	
4	2008	1	1	2	XE	ONT	ELP	670	

Out[124]:

```
Year                False
Month               False
DayofMonth          False
DayOfWeek           False
UniqueCarrier       False
Origin              False
Dest                False
Distance            False
Cancelled           False
CancellationCode     True
Code                False
Description         False
iata                False
airport             False
city                 True
state                True
country             False
lat                 False
long                False
dtype: bool
```

The data contains Null values in the columns Cancellation Code, city and state. This is acceptable for the intended purpose so I will keep them.

Out[125]:

	count	mean	std	min	25%	50%	
Year	2319612.0	2008.000000	0.000000	2008.000000	2008.000000	2008.000000	2
Month	2319612.0	6.343666	3.433463	1.000000	3.000000	6.000000	
DayofMonth	2319612.0	15.727206	8.807478	1.000000	8.000000	16.000000	
DayOfWeek	2319612.0	3.979993	1.997027	1.000000	2.000000	4.000000	
Distance	2319612.0	812.030221	622.321614	11.000000	365.000000	640.000000	1
Cancelled	2319612.0	0.047514	0.212735	0.000000	0.000000	0.000000	
lat	2319612.0	37.206219	5.924640	17.701889	33.640444	38.533963	
long	2319612.0	-93.902916	17.520690	-176.646031	-104.667002	-87.904464	

To summarize the first investigation, the data is very tidy and can easily be used for further analyis.

```
Year                                2008
Month                                  1
DayofMonth                             1
DayOfWeek                              2
UniqueCarrier                         XE
Origin                               IAH
Dest                                 BRO
Distance                             308
Cancelled                              0
CancellationCode                     NaN
Code                                  XE
Description           Expressjet Airlines Inc.
iata                                 IAH
airport          George Bush Intercontinental
city                             Houston
state                                 TX
country                              USA
lat                              29.9805
long                            -95.3397
Name: 358, dtype: object
```

The Cancelation Code contains not NaN values. This needs to be fixed.

```
Year                                    2008
Month                                      1
DayofMonth                                 1
DayOfWeek                                  2
UniqueCarrier                             XE
Origin                                   IAH
Dest                                     BRO
Distance                                 308
Cancelled                                  0
CancellationCode                           0
Code                                      XE
Description            Expressjet Airlines Inc.
iata                                     IAH
airport            George Bush Intercontinental
city                                 Houston
state                                     TX
country                                  USA
lat                                  29.9805
long                                -95.3397
Name: 358, dtype: object
```

Ok, looks good now. After I was using fillna I still got NaN, replace is the better choice in this case.

Investigation on the questions:

```
What are the airliners with the least cancellations?
Which airports are the busiest ones?
Is there a timly insight on flight frequency over the months?
```

Out[179]:

```
0.047513549679860254
```

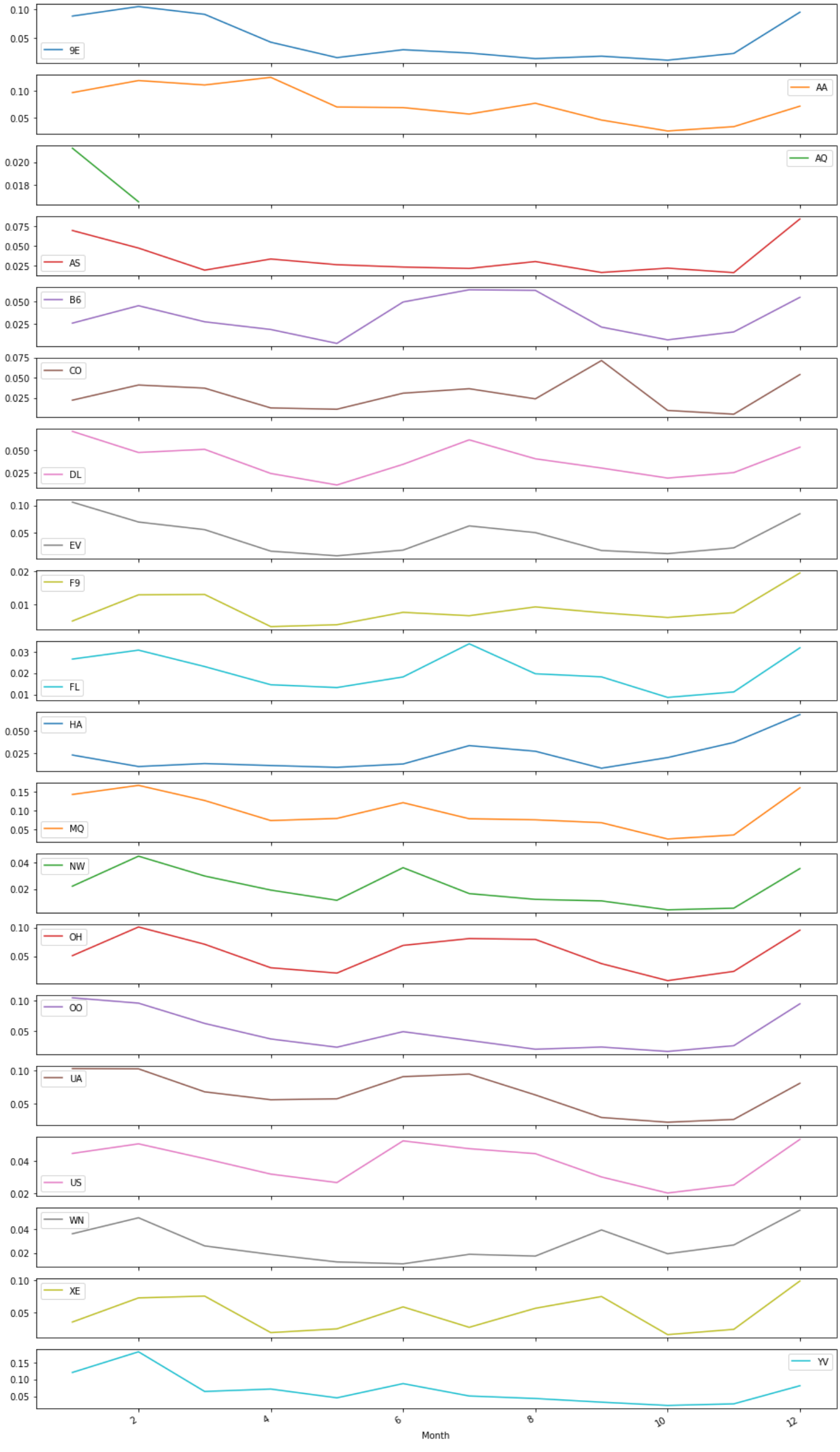
The overall cancellation rate of flights is 0.048 in 2018, nearly 5%.

Out[199]:

UniqueCarrier	Cancelled	Canrate
9E	6666	325409089032
AA	12109	362808193308
AQ	40	4901340156
AS	1791	120387862800
B6	2780	189707147808
CO	2369	186700930656
DL	5368	310651717488
EV	4303	224508286644
F9	243	66981116112
FL	1953	214239364320
HA	329	32548795584
MQ	14345	336357657672
NW	2556	281109139056
OH	5529	225765516348
OO	10439	484302511032
UA	9118	311236259712
US	5470	323286644052
WN	8874	744386686920
XE	7559	364765945836
YV	8372	270545626008

Out[183]:

```
array([<matplotlib.axes._subplots.AxesSubplot object at 0x14cc087f
0>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x151e0751
8>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x14bb6c97
8>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x14bb79ef
0>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x14e7dd4a
8>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x14e7dd4e
0>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x1cc378f9
8>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x141e4a55
0>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x141e4fac
8>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x146734c5
0>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x1521a263
0>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x142dceb7
0>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x14d9ba12
8>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x14d9a96a
0>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x150d5cb7
0>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x14590ce1
0>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x1483bf16
0>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x1483c366
8>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x1a854fc5
0>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x15141e20
8>],
      dtype=object)
```

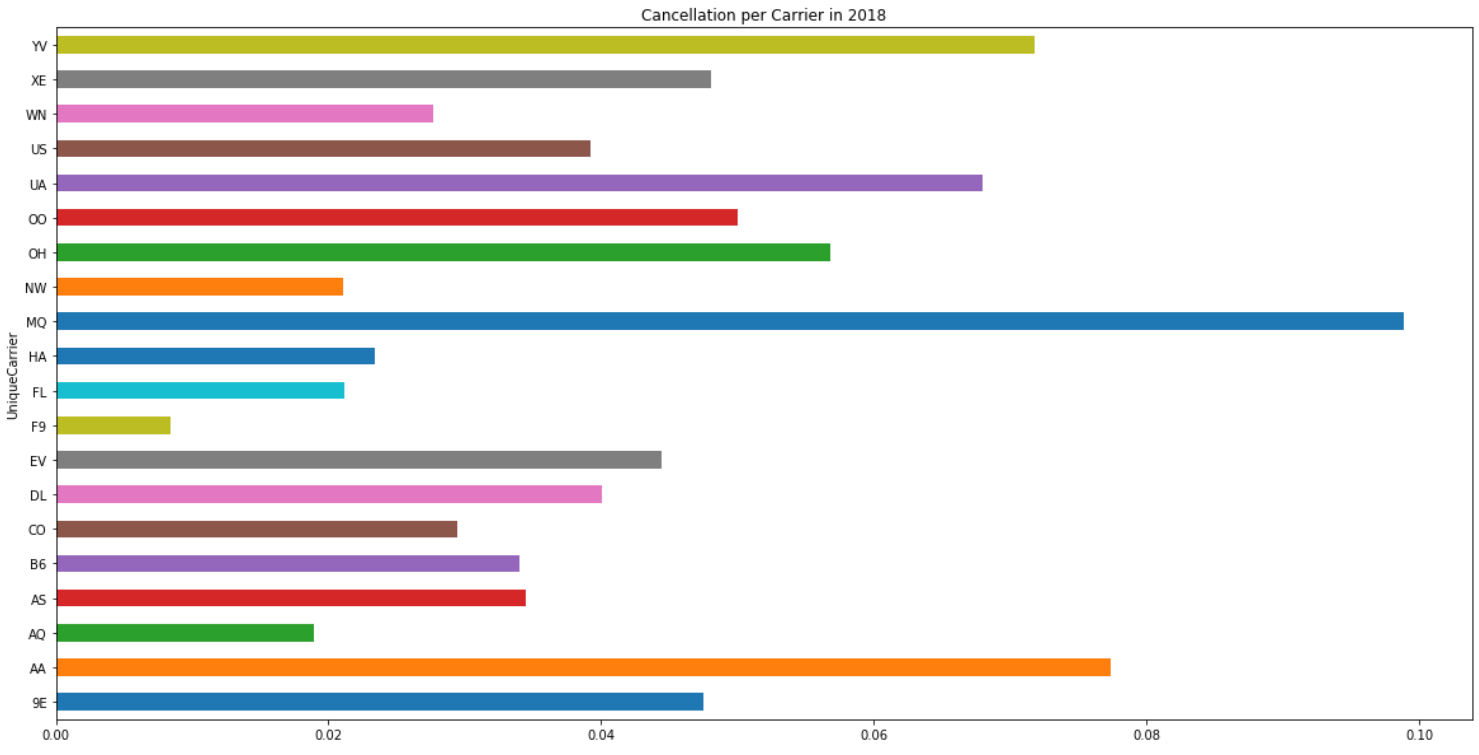

9E
AA
AQ
AS
B6
CO
DL
EV
F9
FL
HA
MQ
NW
OH
OO
UA
US
WN
XE
YV
Month

Out[201]:

```
UniqueCarrier
9E    0.047517
AA    0.077419
AQ    0.018930
AS    0.034509
B6    0.033992
CO    0.029433
DL    0.040082
EV    0.044458
F9    0.008415
FL    0.021146
HA    0.023446
MQ    0.098927
NW    0.021091
OH    0.056807
OO    0.049999
UA    0.067956
US    0.039248
WN    0.027653
XE    0.048069
YV    0.071780
dtype: float64
```

Out[205]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x13a1e6da0>
```



I need to find a way to display the carrier full name, for the moment it does not work with description.