app_my

July 21, 2020

```
[1]: import pandas as pd
```

```
[2]: #Wszytanie pliku csv przy pomocy biblioteki Pandas
     flights = pd.read_csv('flight_data_2016.csv')
```

```
[3]: #Sprawdzenie kolumn w danych
     flights.columns
```

```
[3]: Index(['QUARTER', 'MONTH', 'DAY_OF_MONTH', 'DAY_OF_WEEK', 'FL_DATE',
            'UNIQUE_CARRIER', 'FL_NUM', 'ORIGIN_AIRPORT_ID',
            'ORIGIN_AIRPORT_SEQ_ID', 'ORIGIN_CITY_MARKET_ID', 'ORIGIN',
            'ORIGIN_CITY_NAME', 'ORIGIN_STATE_ABR', 'ORIGIN_STATE_NM',
            'DEST_AIRPORT_ID', 'DEST_AIRPORT_SEQ_ID', 'DEST_CITY_MARKET_ID', 'DEST',
            'DEST_CITY_NAME', 'DEST_STATE_ABR', 'DEST_STATE_NM', 'CRS_DEP_TIME',
            'DEP_TIME', 'DEP_DELAY', 'DEP_DELAY_NEW', 'WHEELS_ON', 'TAXI_IN',
            'CRS_ARR_TIME', 'ARR_TIME', 'ARR_DELAY', 'ARR_DELAY_NEW',
            'CRS_ELAPSED_TIME', 'ACTUAL_ELAPSED_TIME', 'AIR_TIME', 'FLIGHTS',
            'DISTANCE', 'CARRIER_DELAY', 'WEATHER_DELAY', 'NAS_DELAY',
            'SECURITY_DELAY', 'LATE_AIRCRAFT_DELAY', 'FIRST_DEP_TIME',
            'Unnamed: 42'],
           dtype='object')
```

```
[4]: #Wyświetlenie pierwszych 5 pozycji
     flights.head()
```

```
[4]:    QUARTER  MONTH  DAY_OF_MONTH  DAY_OF_WEEK     FL_DATE UNIQUE_CARRIER  \
     0        1      1             3            7  2016-01-03             F9
     1        1      1             3            7  2016-01-03             F9
     2        1      1             3            7  2016-01-03             F9
     3        1      1             3            7  2016-01-03             F9
     4        1      1             3            7  2016-01-03             F9

        FL_NUM  ORIGIN_AIRPORT_ID  ORIGIN_AIRPORT_SEQ_ID  ORIGIN_CITY_MARKET_ID  \
     0     694              11292                1129202                  30325
     1     809              14027                1402702                  34027
     2     907              15356                1535602                  35356
     3     908              14492                1449202                  34492
```

```
4     915              15356                1535602                  35356

   … AIR_TIME FLIGHTS DISTANCE CARRIER_DELAY  WEATHER_DELAY  NAS_DELAY  \
0  …     87.0     1.0    692.0           NaN            NaN        NaN
1  …    224.0     1.0   1679.0          19.0            0.0        0.0
2  …     60.0     1.0    373.0           NaN            NaN        NaN
3  …     57.0     1.0    373.0           NaN            NaN        NaN
4  …    107.0     1.0    693.0           2.0            0.0       18.0

   SECURITY_DELAY LATE_AIRCRAFT_DELAY FIRST_DEP_TIME Unnamed: 42
0             NaN                 NaN            NaN         NaN
1             0.0                 0.0            NaN         NaN
2             NaN                 NaN            NaN         NaN
3             NaN                 NaN            NaN         NaN
4             0.0                 0.0            NaN         NaN

[5 rows x 43 columns]
```

[5]:

```
#Wyciagniecie interesujacych danych
flights =␣
 ↪flights[['QUARTER','MONTH','DAY_OF_MONTH','DAY_OF_WEEK','UNIQUE_CARRIER','ARR_DELAY','ORIGI
 ↪'DEST_CITY_NAME', 'DISTANCE', 'AIR_TIME']]
flights
```

[5]:

```
         QUARTER  MONTH  DAY_OF_MONTH  DAY_OF_WEEK UNIQUE_CARRIER  ARR_DELAY  \
0              1      1             3            7             F9       -5.0
1              1      1             3            7             F9       19.0
2              1      1             3            7             F9       -2.0
3              1      1             3            7             F9       -5.0
4              1      1             3            7             F9       20.0
…            …    …           …          …            …          …
1856056        4     12            30            5             DL       -5.0
1856057        4     12            30            5             DL        3.0
1856058        4     12            30            5             DL      -29.0
1856059        4     12            30            5             DL       -3.0
1856060        4     12            30            5             DL      -10.0

                     ORIGIN_CITY_NAME             DEST_CITY_NAME  DISTANCE  \
0                          Denver, CO  Cedar Rapids/Iowa City, IA     692.0
1         West Palm Beach/Palm Beach, FL              Denver, CO    1679.0
2                         Trenton, NJ          Raleigh/Durham, NC     373.0
3                  Raleigh/Durham, NC                 Trenton, NJ     373.0
4                         Trenton, NJ                 Chicago, IL     693.0
…                                 …                          …       …
1856056            Fort Lauderdale, FL                Atlanta, GA     581.0
1856057                    Atlanta, GA              Milwaukee, WI     669.0
1856058                  Milwaukee, WI                Atlanta, GA     669.0
```

```
1856059                       Atlanta, GA              Fort Myers, FL     515.0
1856060                    Fort Myers, FL                 Atlanta, GA     515.0

         AIR_TIME
0            87.0
1           224.0
2            60.0
3            57.0
4           107.0
…           …
1856056      88.0
1856057     105.0
1856058      89.0
1856059      71.0
1856060      85.0

[1856061 rows x 10 columns]
```

[6]:

```python
#Sprawdzenie ilosci wartosci NULL w kazdej kolumnie
flights.isna().sum(axis = 0)
```

[6]:

```
QUARTER                  0
MONTH                    0
DAY_OF_MONTH             0
DAY_OF_WEEK              0
UNIQUE_CARRIER           0
ARR_DELAY            31658
ORIGIN_CITY_NAME         0
DEST_CITY_NAME           0
DISTANCE                 0
AIR_TIME             31658
dtype: int64
```

[7]:

```python
import numpy as np
```

[8]:

```python
#Usuniecie wartosci NULL
flights['ARR_DELAY'].replace('', np.nan, inplace=True)
flights.dropna(subset=['ARR_DELAY'], inplace=True)

flights['AIR_TIME'].replace('', np.nan, inplace=True)
flights.dropna(subset=['AIR_TIME'], inplace=True)
```

[9]:

```python
#Sprawdzenie wartosci NULL w kazdej kolumnie
flights.isnull().sum(axis = 0)
```

[9]:

```
QUARTER             0
MONTH               0
```

```
DAY_OF_MONTH        0
DAY_OF_WEEK         0
UNIQUE_CARRIER      0
ARR_DELAY           0
ORIGIN_CITY_NAME    0
DEST_CITY_NAME      0
DISTANCE            0
AIR_TIME            0
dtype: int64
```

[10]:
```
#Sprawdzenie ilosci wierszy
count_row = flights.shape[0]
count_row
```

[10]:
```
1824403
```

[11]:
```
#dodanie kolumny indeksow na podstawie ilosci wierszy
flights['New_ID'] = range(0, 0 +flights.shape[0])
#Wyciagniecie interesujacych danych w odpowiedniej kolejnosci (indeks na␣
 ↪początku)
flights =␣
 ↪flights[['New_ID','QUARTER','MONTH','DAY_OF_MONTH','DAY_OF_WEEK','UNIQUE_CARRIER','ARR_DELA
 ↪'DEST_CITY_NAME', 'DISTANCE', 'AIR_TIME']]
flights
```

[11]:
```
          New_ID  QUARTER  MONTH  DAY_OF_MONTH  DAY_OF_WEEK UNIQUE_CARRIER  \
0              0        1      1             3            7             F9
1              1        1      1             3            7             F9
2              2        1      1             3            7             F9
3              3        1      1             3            7             F9
4              4        1      1             3            7             F9
…            …      …    …            …          …            …
1856056  1824398        4     12            30            5             DL
1856057  1824399        4     12            30            5             DL
1856058  1824400        4     12            30            5             DL
1856059  1824401        4     12            30            5             DL
1856060  1824402        4     12            30            5             DL

         ARR_DELAY               ORIGIN_CITY_NAME  \
0             -5.0                     Denver, CO
1             19.0  West Palm Beach/Palm Beach, FL
2             -2.0                    Trenton, NJ
3             -5.0              Raleigh/Durham, NC
4             20.0                    Trenton, NJ
…              …                             …
1856056       -5.0             Fort Lauderdale, FL
1856057        3.0                    Atlanta, GA
```

```
1856058      -29.0                    Milwaukee, WI
1856059       -3.0                      Atlanta, GA
1856060      -10.0                   Fort Myers, FL

                     DEST_CITY_NAME  DISTANCE  AIR_TIME
0        Cedar Rapids/Iowa City, IA     692.0      87.0
1                        Denver, CO    1679.0     224.0
2                Raleigh/Durham, NC     373.0      60.0
3                       Trenton, NJ     373.0      57.0
4                       Chicago, IL     693.0     107.0
…                             …       …       …
1856056                 Atlanta, GA     581.0      88.0
1856057               Milwaukee, WI     669.0     105.0
1856058                 Atlanta, GA     669.0      89.0
1856059              Fort Myers, FL     515.0      71.0
1856060                 Atlanta, GA     515.0      85.0

[1824403 rows x 11 columns]
```

[12]:

```
#Wyliczenie dodatkowo szybkosci samolotu
flights['air_speed (mph)'] = flights['DISTANCE'] / (flights['AIR_TIME'] / 60)
flights
```

```
/Users/monikajanocha/opt/anaconda3/lib/python3.7/site-
packages/ipykernel_launcher.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

[12]:

```
          New_ID  QUARTER  MONTH  DAY_OF_MONTH  DAY_OF_WEEK UNIQUE_CARRIER  \
0              0        1      1             3            7             F9
1              1        1      1             3            7             F9
2              2        1      1             3            7             F9
3              3        1      1             3            7             F9
4              4        1      1             3            7             F9
…          …      …    …           …          …           …
1856056  1824398        4     12            30            5             DL
1856057  1824399        4     12            30            5             DL
1856058  1824400        4     12            30            5             DL
1856059  1824401        4     12            30            5             DL
1856060  1824402        4     12            30            5             DL

         ARR_DELAY                 ORIGIN_CITY_NAME  \
0             -5.0                       Denver, CO
```

```
1             19.0  West Palm Beach/Palm Beach, FL
2             -2.0                     Trenton, NJ
3             -5.0              Raleigh/Durham, NC
4             20.0                     Trenton, NJ
…            …                              …
1856056       -5.0             Fort Lauderdale, FL
1856057        3.0                     Atlanta, GA
1856058      -29.0                   Milwaukee, WI
1856059       -3.0                     Atlanta, GA
1856060      -10.0                  Fort Myers, FL

                     DEST_CITY_NAME  DISTANCE  AIR_TIME  air_speed (mph)
0        Cedar Rapids/Iowa City, IA     692.0      87.0       477.241379
1                        Denver, CO    1679.0     224.0       449.732143
2                Raleigh/Durham, NC     373.0      60.0       373.000000
3                       Trenton, NJ     373.0      57.0       392.631579
4                       Chicago, IL     693.0     107.0       388.598131
…                             …       …       …              …
1856056                 Atlanta, GA     581.0      88.0       396.136364
1856057               Milwaukee, WI     669.0     105.0       382.285714
1856058                 Atlanta, GA     669.0      89.0       451.011236
1856059              Fort Myers, FL     515.0      71.0       435.211268
1856060                 Atlanta, GA     515.0      85.0       363.529412

[1824403 rows x 12 columns]
```

[13]:

```
#wygenerowanie nowego pliku csv
flights.to_csv('flight_data_2016_nowe.csv', index=False)
```

[13]:

```
#Wszytanie pliku csv przy pomocy biblioteki Pandas
flights2 = pd.read_csv('flight_data_2016_nowe.csv', index_col='New_ID')
```

```
/Users/monikajanocha/opt/anaconda3/lib/python3.7/site-
packages/numpy/lib/arraysetops.py:569: FutureWarning: elementwise comparison
failed; returning scalar instead, but in the future will perform elementwise
comparison
  mask |= (ar1 == a)
```

[14]:

```
flights2
```

[14]:

```
         QUARTER  MONTH  DAY_OF_MONTH  DAY_OF_WEEK UNIQUE_CARRIER  ARR_DELAY   \
New_ID
0              1      1             3            7             F9       -5.0
1              1      1             3            7             F9       19.0
2              1      1             3            7             F9       -2.0
3              1      1             3            7             F9       -5.0
4              1      1             3            7             F9       20.0
```

```
…          …    …           …          …            …        …
1824398        4     12            30            5             DL       -5.0
1824399        4     12            30            5             DL        3.0
1824400        4     12            30            5             DL      -29.0
1824401        4     12            30            5             DL       -3.0
1824402        4     12            30            5             DL      -10.0

                       ORIGIN_CITY_NAME              DEST_CITY_NAME  DISTANCE  \
New_ID
0                            Denver, CO  Cedar Rapids/Iowa City, IA     692.0
1        West Palm Beach/Palm Beach, FL                  Denver, CO    1679.0
2                           Trenton, NJ          Raleigh/Durham, NC     373.0
3                    Raleigh/Durham, NC                 Trenton, NJ     373.0
4                           Trenton, NJ                 Chicago, IL     693.0
…                                 …                         …       …
1824398             Fort Lauderdale, FL                 Atlanta, GA     581.0
1824399                     Atlanta, GA               Milwaukee, WI     669.0
1824400                   Milwaukee, WI                 Atlanta, GA     669.0
1824401                     Atlanta, GA              Fort Myers, FL     515.0
1824402                  Fort Myers, FL                 Atlanta, GA     515.0

         AIR_TIME  air_speed (mph)
New_ID
0            87.0       477.241379
1           224.0       449.732143
2            60.0       373.000000
3            57.0       392.631579
4           107.0       388.598131
…           …              …
1824398      88.0       396.136364
1824399     105.0       382.285714
1824400      89.0       451.011236
1824401      71.0       435.211268
1824402      85.0       363.529412

[1824403 rows x 11 columns]
```

[15]:

```
flights2_okrojone = flights2.iloc[:,4:]
```

[16]:

```
flights2_okrojone
```

[16]:

```
        UNIQUE_CARRIER  ARR_DELAY                ORIGIN_CITY_NAME  \
New_ID
0                   F9       -5.0                      Denver, CO
1                   F9       19.0  West Palm Beach/Palm Beach, FL
2                   F9       -2.0                     Trenton, NJ
3                   F9       -5.0              Raleigh/Durham, NC
```

```
4                   F9       20.0                 Trenton, NJ
…                 …        …                        …
1824398             DL       -5.0         Fort Lauderdale, FL
1824399             DL        3.0                 Atlanta, GA
1824400             DL      -29.0               Milwaukee, WI
1824401             DL       -3.0                 Atlanta, GA
1824402             DL      -10.0              Fort Myers, FL

                    DEST_CITY_NAME  DISTANCE  AIR_TIME  air_speed (mph)
New_ID
0        Cedar Rapids/Iowa City, IA     692.0      87.0       477.241379
1                        Denver, CO    1679.0     224.0       449.732143
2                 Raleigh/Durham, NC     373.0      60.0       373.000000
3                       Trenton, NJ     373.0      57.0       392.631579
4                       Chicago, IL     693.0     107.0       388.598131
…                               …       …       …               …
1824398                 Atlanta, GA     581.0      88.0       396.136364
1824399               Milwaukee, WI     669.0     105.0       382.285714
1824400                 Atlanta, GA     669.0      89.0       451.011236
1824401              Fort Myers, FL     515.0      71.0       435.211268
1824402                 Atlanta, GA     515.0      85.0       363.529412

[1824403 rows x 7 columns]
```

[17]:

```
#Wyliczenie jakie sa wartosci statystyczne dla kazdej z linii lotniczej o␣
 ↪konkretnej trasie
flights2_grouped = flights2_okrojone.
 ↪groupby(['UNIQUE_CARRIER','ORIGIN_CITY_NAME', 'DEST_CITY_NAME']).
 ↪agg(['count', 'mean', 'min', 'max']).reset_index()
flights2_grouped
```

[17]:

```
     UNIQUE_CARRIER ORIGIN_CITY_NAME         DEST_CITY_NAME ARR_DELAY  \
                                                                count
0                AA      Albany, NY         Charlotte, NC       324
1                AA  Albuquerque, NM  Dallas/Fort Worth, TX       533
2                AA     Amarillo, TX  Dallas/Fort Worth, TX        56
3                AA    Anchorage, AK  Dallas/Fort Worth, TX        30
4                AA    Anchorage, AK        Los Angeles, CA        22
…              …              …                    …      …
7417             WN      Wichita, KS            Chicago, IL        53
7418             WN      Wichita, KS             Dallas, TX        57
7419             WN      Wichita, KS          Las Vegas, NV       122
7420             WN      Wichita, KS            Phoenix, AZ        91
7421             WN      Wichita, KS          St. Louis, MO       171

                              DISTANCE                       AIR_TIME  \
            mean   min     max    count     mean     min     max    count
```

```
0      -7.904321 -49.0   206.0      324   646.0   646.0   646.0      324
1       6.744841 -35.0  1260.0      533   569.0   569.0   569.0      533
2      -1.625000 -31.0   130.0       56   312.0   312.0   312.0       56
3     114.733333 -26.0  1295.0       30  3043.0  3043.0  3043.0       30
4     -33.318182 -54.0    43.0       22  2345.0  2345.0  2345.0       22
…          …   …     …      …     …     …     …      …
7417    4.320755 -21.0   178.0       53   589.0   589.0   589.0       53
7418    4.245614 -21.0   313.0       57   333.0   333.0   333.0       57
7419   -6.229508 -43.0   298.0      122   986.0   986.0   986.0      122
7420   -7.494505 -45.0   108.0       91   870.0   870.0   870.0       91
7421   -2.192982 -31.0   325.0      171   392.0   392.0   392.0      171

                               air_speed (mph)                           \
            mean   min    max           count        mean         min
0     106.395062  87.0  138.0             324  367.229228  280.869565
1      77.645403  67.0  112.0             533  441.701742  304.821429
2      49.071429  42.0   67.0              56  384.090308  279.402985
3     351.800000 332.0  375.0              30  519.408818  486.880000
4     277.136364 264.0  292.0              22  508.006802  481.849315
…          …    …    …             …         …         …
7417   82.679245  75.0  103.0              53  428.576997  343.106796
7418   56.368421  50.0   63.0              57  355.535563  317.142857
7419  138.237705 115.0  166.0             122  430.175175  356.385542
7420  126.340659 103.0  147.0              91  415.328312  355.102041
7421   54.099415  47.0   93.0             171  437.516893  252.903226


             max
0     445.517241
1     509.552239
2     445.714286
3     549.939759
4     532.954545
…          …
7417  471.200000
7418  399.600000
7419  514.434783
7420  506.796117
7421  500.425532

[7422 rows x 19 columns]
```

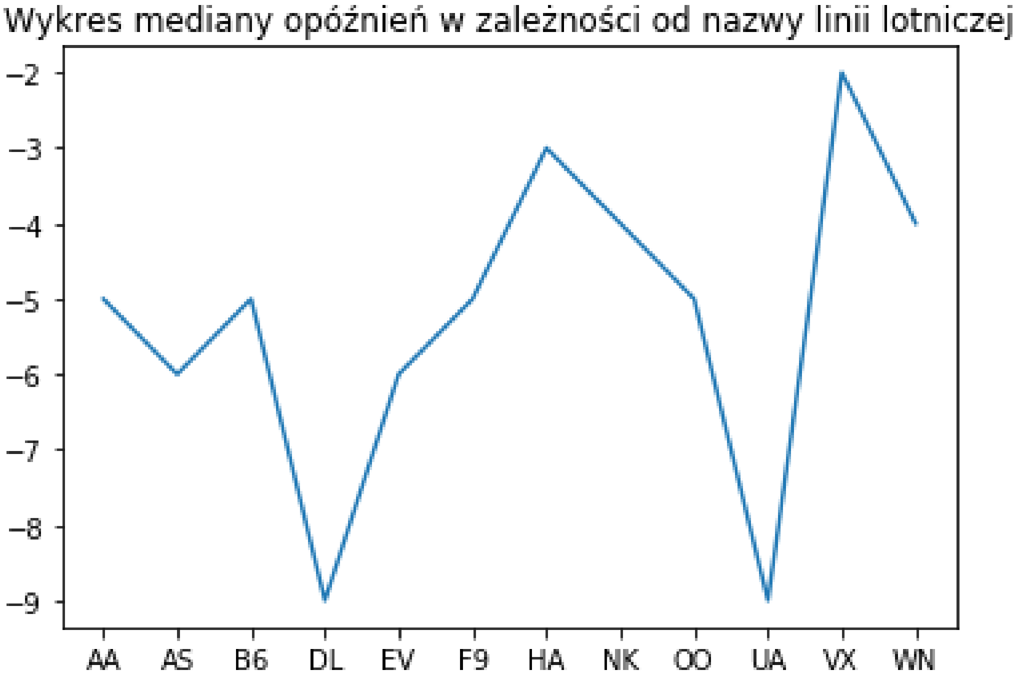
```
[18]: carrier_stats2 = flights2_okrojone.groupby('UNIQUE_CARRIER')['air_speed (mph)'].
       ↪describe().reset_index().rename(columns={'UNIQUE_CARRIER': 'airline',␣
       ↪'count': 'number_of_flights', '50%':'median'})
      carrier_stats2
```

[18]:
```
   airline  number_of_flights        mean        std         min         25%  \
0       AA           295833.0  423.912774  68.590616   70.751445  384.761905
1       AS            58543.0  437.894440  62.692879   89.062500  409.606299
2       B6            93313.0  420.127606  72.145404   97.500000  379.487179
3       DL           296594.0  417.086086  64.261109  115.675676  377.647059
4       EV           154224.0  361.223625  66.158543   50.308530  318.260870
5       F9            31236.0  448.124893  49.197357  165.397590  416.097561
6       HA            25683.0  341.098756  88.909812  149.268293  272.727273
7       NK            46196.0  435.862743  54.175735  131.111111  404.457831
8       OO           194789.0  367.814386  74.208374   70.985915  322.857143
9       UA           178621.0  444.691363  64.347271   95.714286  407.272727
10      VX            23063.0  442.336155  65.358191  190.754717  392.701422
11      WN           426308.0  415.021506  59.775816   77.513514  374.444444

        median         75%         max
0   429.310345  469.629630  760.000000
1   440.571429  474.279221  762.105263
2   432.857143  467.213115  628.358209
3   419.318182  458.365385  628.369565
4   365.357143  407.899160  731.250000
5   445.263158  478.656716  619.072848
6   322.105263  370.285714  616.267606
7   439.800000  471.190476  674.838710
8   375.428571  420.000000  726.000000
9   447.532468  488.000000  647.088608
10  440.597015  493.214286  625.603448
11  414.827586  455.368421  786.346154
```

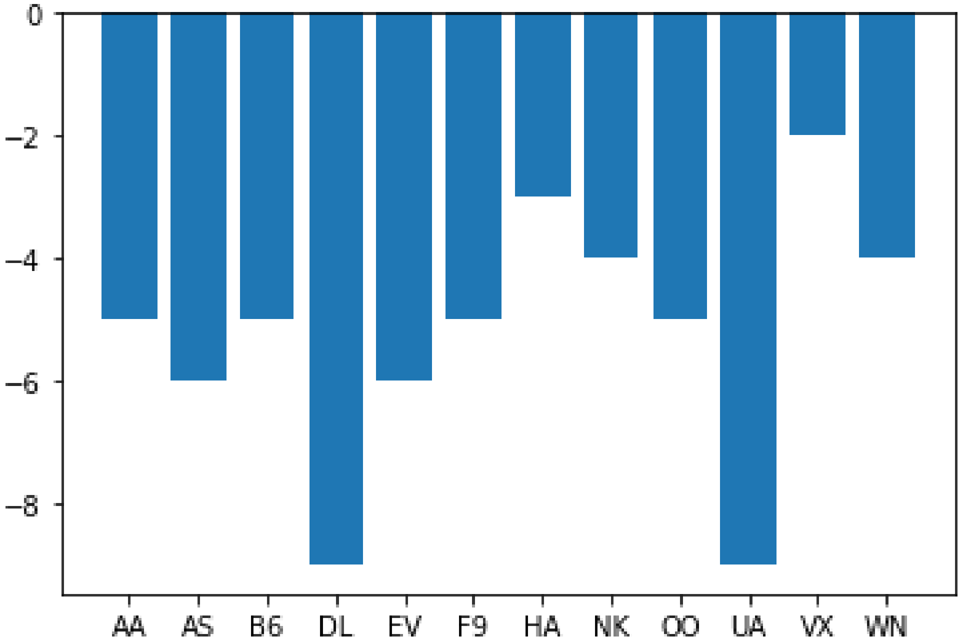
[20]:
```
import matplotlib.pyplot as plt
```

[21]:
```
x = carrier_stats['airline']
y = carrier_stats['median']
plt.plot(x, y)
plt.title('Wykres mediany opóźnień w zależności od nazwy linii lotniczej')
plt.show()
```



```
[22]: x=carrier_stats['airline']
      y=carrier_stats['median']
      plt.bar(x, y)
```

```
[22]: <BarContainer object of 12 artists>
```

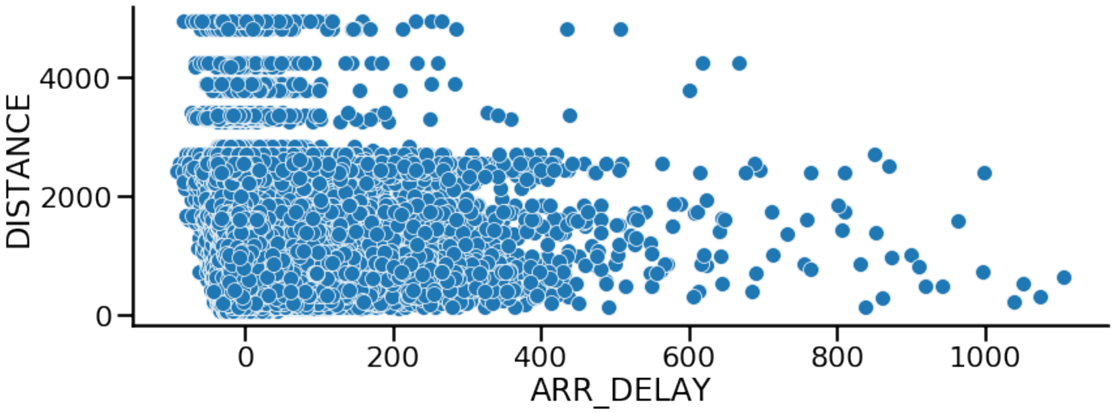


[23]:

```
import seaborn as sns
```

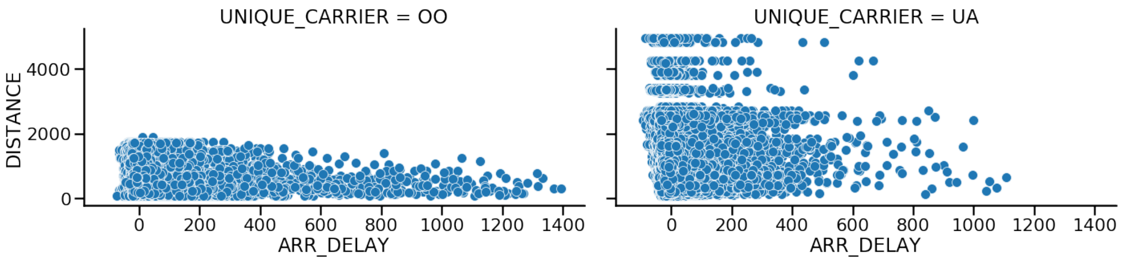
[24]:

```
sns.set_context('poster')
sns.relplot(x="ARR_DELAY",
            y="DISTANCE",
            aspect=2.5,
            data=flights[flights['UNIQUE_CARRIER']=='UA'],
            kind="scatter");
plt.show()
```

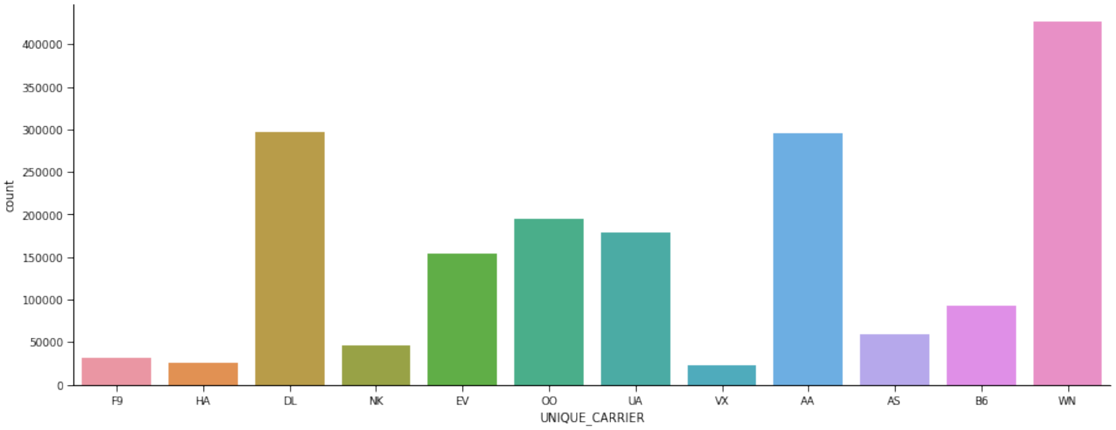


[27]:

```
sns.set_context('poster')
sns.relplot(data=flights[(flights['UNIQUE_CARRIER'] == 'OO') |␣
 ↪(flights['UNIQUE_CARRIER'] == 'UA') ],
            x="ARR_DELAY",
            y="DISTANCE",
            aspect=2,
            kind="scatter",
            col='UNIQUE_CARRIER')
plt.show()
```



```
[26]: sns.set_context('paper')
      sns.catplot(x="UNIQUE_CARRIER",
                  data=flights,
                  aspect=2.5,
                  kind='count')
      plt.show()
```


count
F9
HA
DL
NK
EV
OO
UA
VX
AA
AS
B6
WN
UNIQUE_CARRIER