R Notebook

```
library(reticulate)
Sys.which("python")
```

```
##                                python 
## "C:\\PROGRA~3\\ANACON~1\\python.exe"
```

```
py_config()
```

```
## python:         C:\PROGRA~3\ANACON~1\python.exe
## libpython:      C:/PROGRA~3/ANACON~1/python37.dll
## pythonhome:     C:\PROGRA~3\ANACON~1
## version:        3.7.3 (default, Mar 27 2019, 17:13:21) [MSC v.1915 64 bit (AMD64)]
## Architecture:   64bit
## numpy:          C:\PROGRA~3\ANACON~1\lib\site-packages\numpy
## numpy_version:  1.16.2
## 
## python versions found: 
##  C:\PROGRA~3\ANACON~1\python.exe
##  C:\Users\KIM\AppData\Local\Programs\Python\PYTHON~1\\python.exe
##  C:\Users\KIM\AppData\Local\Programs\Python\Python37\\python.exe
##  C:\Python27\\python.exe
##  D:\Anaconda2019\python.exe
##  C:\ProgramData\Anaconda3\python.exe
```

```
use_condaenv("r-reticulate")
use_virtualenv("myenv")
```

Run python

Refer https://www.kaggle.com/kashnitsky/topic-1-exploratory-data-analysis-with-pandas/notebook

```
import numpy as np
import pandas as pd
# we don't like warnings
# you can comment the following 2 lines if you'd like to
```

Read data

```
df = pd.read_csv('flights.csv')
df.head()
```

```
##    year  month  day  dep_time  dep_delay  ...  dest  air_time distance hour  minute
## 0  2013      1    1     517.0        2.0  ...   IAH     227.0     1400  5.0    17.0
## 1  2013      1    1     533.0        4.0  ...   IAH     227.0     1416  5.0    33.0
## 2  2013      1    1     542.0        2.0  ...   MIA     160.0     1089  5.0    42.0
## 3  2013      1    1     554.0       -6.0  ...   ATL     116.0      762  5.0    54.0
## 4  2013      1    1     554.0       -4.0  ...   ORD     150.0      719  5.0    54.0
## 
## [5 rows x 16 columns]
```

see dimension

```
print(df.shape)
```

```
## (160754, 16)
```

```
print(df.columns)
```

```
## Index(['year', 'month', 'day', 'dep_time', 'dep_delay', 'arr_time',
##        'arr_delay', 'carrier', 'tailnum', 'flight', 'origin', 'dest',
##        'air_time', 'distance', 'hour', 'minute'],
##       dtype='object')
```

change column format

```
df.describe()
```

```
##            year          month  ...          hour         minute
## count  160754.0  160754.000000  ...  158418.000000  158418.000000
## mean     2013.0       6.547395  ...      12.837582      32.387847
## std         0.0       3.410001  ...       4.725552      18.687423
## min      2013.0       1.000000  ...       0.000000       0.000000
## 25%      2013.0       4.000000  ...       8.000000      16.000000
## 50%      2013.0       7.000000  ...      13.000000      32.000000
## 75%      2013.0      10.000000  ...      17.000000      51.000000
## max      2013.0      12.000000  ...      24.000000      59.000000
##
## [8 rows x 12 columns]
```

```
df.describe(include=['object', 'bool'])
```

```
##        carrier tailnum  origin    dest
## count   160754  159321  160754  160754
## unique       5    2222       3      59
## top         UA  N328AA     LGA     ORD
## freq     58665     393   59706   13043
```