

## ▼ Start

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import matplotlib as mpl
import matplotlib.patches as patches
from matplotlib.patches import ConnectionPatch
from collections import OrderedDict
from matplotlib.gridspec import GridSpec
```

```
dataset = pd.read_csv("dataset.csv")
carr=pd.read_csv('carriers.csv')
```

```
dataset.head()
```

	year	month	day	dep_time	arr_delay	carrier	origin	dest	distance	temp
<b>0</b>	2013	1	1	554.0	-25.0	DL	LGA	ATL	762.0	39.0%
<b>1</b>	2013	1	1	555.0	19.0	B6	EWR	FLL	1065.0	39.0%
<b>2</b>	2013	1	1	557.0	-14.0	EV	LGA	IAD	229.0	39.0%
<b>3</b>	2013	1	1	557.0	-8.0	B6	JFK	MCO	944.0	39.0%

## ▼ Quick Look at the data

```
dataset.head()
```

	year	month	day	dep_time	arr_delay	carrier	origin	dest	distance	temp
<b>0</b>	2013	1	1	554.0	-25.0	DL	LGA	ATL	762.0	39.0%
<b>1</b>	2013	1	1	555.0	19.0	B6	EWR	FLL	1065.0	39.0%
<b>2</b>	2013	1	1	557.0	-14.0	EV	LGA	IAD	229.0	39.0%
<b>3</b>	2013	1	1	557.0	-8.0	B6	JFK	MCO	944.0	39.0%

✓ 0s completed at 8:57 PM



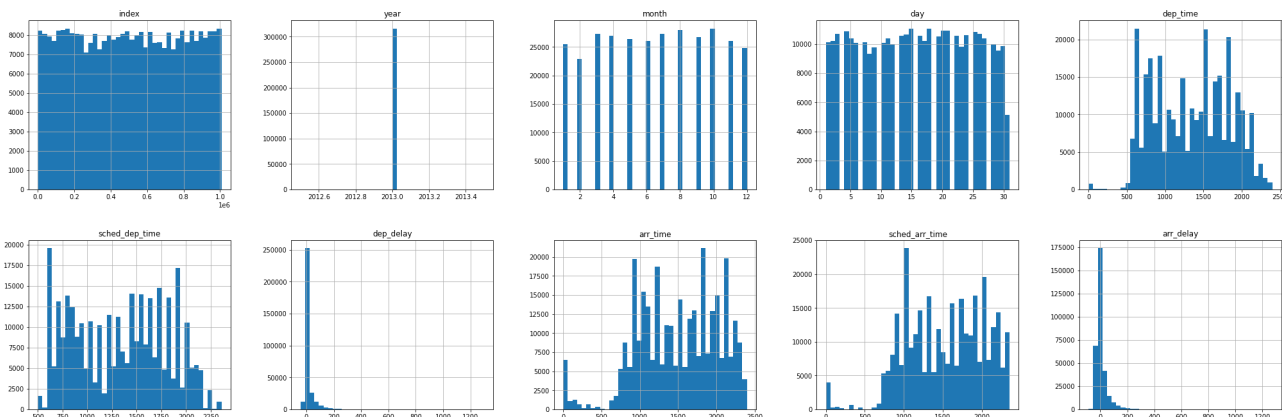
```

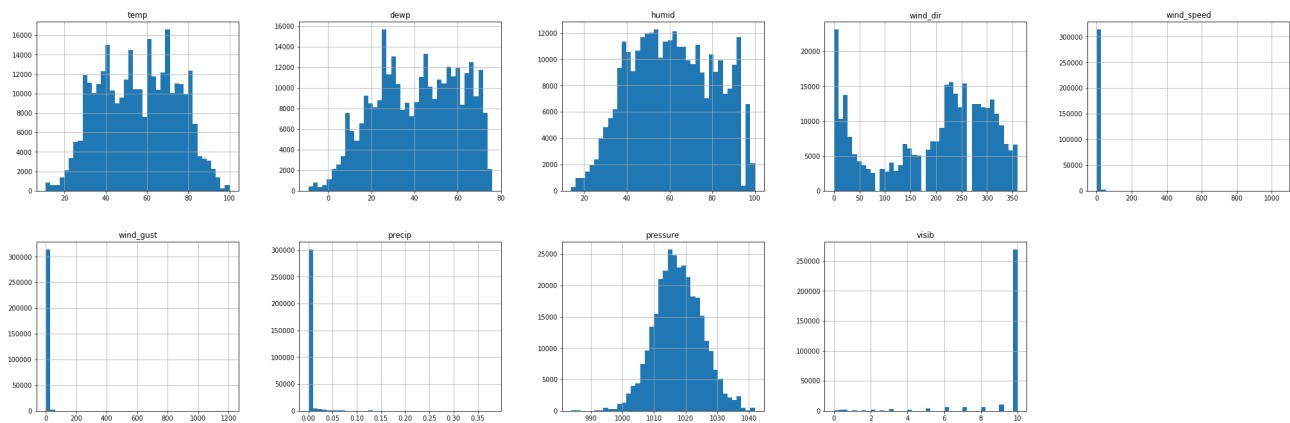
---
0  index      316165 non-null int64
1  year       316165 non-null int64
2  month      316165 non-null int64
3  day        316165 non-null int64
4  dep_time   316165 non-null float64
5  sched_dep_time 316165 non-null float64
6  dep_delay   316165 non-null float64
7  arr_time    316165 non-null float64
8  sched_arr_time 316165 non-null float64
9  arr_delay   316165 non-null float64
10 carrier     316165 non-null object
11 flight      316165 non-null float64
12 tailnum     316165 non-null object
13 origin      316165 non-null object
14 dest        316165 non-null object
15 air_time    316165 non-null float64
16 distance    316165 non-null float64
17 hour        316165 non-null int64
18 minute      316165 non-null float64
19 time_hour   316165 non-null object
20 temp        316165 non-null float64
21 dewp        316165 non-null float64
22 humid       316165 non-null float64
23 wind_dir    316165 non-null float64
24 wind_speed  316165 non-null float64
25 wind_gust   316165 non-null float64
26 precip      316165 non-null float64
27 pressure    316165 non-null float64
28 visib       316165 non-null float64
dtypes: float64(19), int64(5), object(5)
memory usage: 70.0+ MB

```

## Visualizing the data

```
dataset.hist(figsize = (35,30), bins = 40)
plt.show()
```





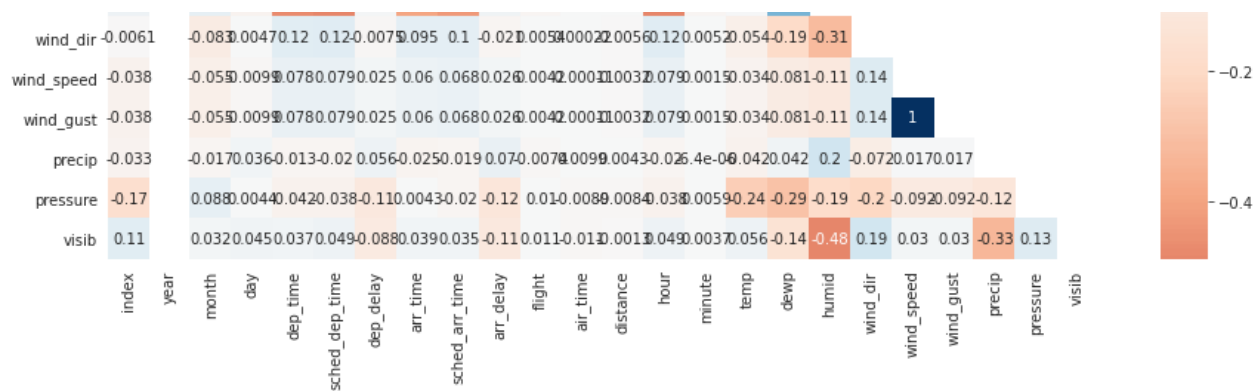
```
dataset.corr()
```

<b>flight</b>	-0.000428	NaN	0.001586	-0.000649	0.041783	0.029254
<b>air_time</b>	-0.032135	NaN	0.008980	0.001380	-0.015228	-0.016261
<b>distance</b>	0.015400	NaN	0.019959	0.003220	-0.014283	-0.013319
<b>hour</b>	0.003435	NaN	-0.007545	-0.002994	0.955943	0.999151
<b>minute</b>	0.022402	NaN	0.015576	0.000847	0.091805	0.082640
<b>temp</b>	0.654072	NaN	0.242402	-0.003485	0.227226	0.231091
<b>dewp</b>	0.605903	NaN	0.260007	-0.018968	-0.021950	-0.027911
<b>humid</b>	0.054889	NaN	0.089136	-0.033570	-0.467822	-0.488281
<b>wind_dir</b>	-0.006084	NaN	-0.083304	0.004725	0.115949	0.120701
<b>wind_speed</b>	-0.038182	NaN	-0.054939	-0.009858	0.077815	0.078614
<b>wind_gust</b>	-0.038182	NaN	-0.054939	-0.009858	0.077815	0.078614
<b>precip</b>	-0.032558	NaN	-0.017147	0.035895	-0.012682	-0.019620
<b>pressure</b>	-0.165455	NaN	0.088488	0.004376	-0.042120	-0.037791
<b>visib</b>	0.111793	NaN	0.031926	0.044819	0.037271	0.048601

```

mask = np.zeros_like(dataset.corr(), dtype=np.bool)
mask[np.triu_indices_from(mask)] = True
sns.set_style('whitegrid')
plt.subplots(figsize = (15,12))
sns.heatmap(dataset.corr(),
             annot=True,

```



Observation

<b>max</b>	1.007588e+06	2013.0	12.000000	31.000000	2400.000000
------------	--------------	--------	-----------	-----------	-------------

```
#plotting mean delays by airlines
carriers = pd.read_csv("carriers.csv")
carrier_code=carriers.set_index('IATA_CODE')['AIRLINE'].to_dict()
mpl.rc('patch', edgecolor = 'dimgray', linewidth = 1)
mpl.rcParams.update(mpl.rcParamsDefault)
mpl.rcParams['hatch.linewidth'] = 2.0

fig = plt.figure(1, figsize = (11, 6))
ax = sns.barplot(x = 'dep_delay', y = 'carrier', data = dataset, color = 'lightblue')
ax = sns.barplot(x = 'arr_delay', y = 'carrier', data = dataset, color = 'red',
labels = [carrier_code[item.get_text()] for item in ax.get_yticklabels()])
ax.set_yticklabels(labels)
ax.yaxis.label.set_visible(False)
plt.xlabel("Mean delay [min] (@departure: blue, @arrival: hatch lines)", font
```

- **Hawaiian Airlines- HA**
- **JetBlue- B6**

```
mpl.rcParams['patch', edgecolor = 'dimgray', linewidth = 1)
mpl.rcParams.update(mpl.rcParamsDefault)
mpl.rcParams['hatch.linewidth'] = 2.0
```

```
fig = plt.figure(1, figsize = (10, 6))
```

```
#Subset 4 major airlines
```

```
ax = sns.barplot(x = 'dep_delay', y = 'carrier', data = dataset, order = ['AA',
```





