

NOAA_ISD_Analysis (/github/Sri-Kusampudi/NOAA_ISD_Analysis/tree/master)

/

Jupyter_PySpark_Program_Data_Analysis (/github/Sri-Kusampudi/NOAA_ISD_Analysis/tree/master/Jupyter_PySpark_Program_Data_Analysis)

In [98]:

```

from pyspark.sql import SparkSession
import pandas as pd
import numpy as np
import pyspark.sql as sparksql
import warnings
import matplotlib as mpl
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
import seaborn as sns
%matplotlib inline
plt.rcParams['figure.figsize']=8,4

warnings.filterwarnings('ignore')
#from pyspark_dist_explore import pandas_histogram

# initialise sparkContext
spark = SparkSession.builder \
    .master('local') \
    .appName('isd_lite_data') \
    .config('spark.executor.memory', '14gb') \
    .config("spark.cores.max", "6") \
    .getOrCreate()

sc = spark.sparkContext

# using SQLContext to read parquet file
from pyspark.sql import SQLContext
sqlContext = SQLContext(sc)

# to read parquet file
#df = sqlContext.read.parquet('C:\Scala_IDE_Eclipse\eclipse\Scala_projects\ATT_Interview_Project\ATT\
#df = pd.read_parquet('C:\Scala_IDE_Eclipse\eclipse\Scala_projects\ATT_Interview_Project\ATT\pc
df = sqlContext.read.parquet('C:\Scala_IDE_Eclipse\eclipse\Scala_projects\ISD_Data_Extract\ATT\parque
df.show()

```

```

+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|      FileName|Month|Day|Hour|ATemp|DTemp|SeaPress|WDirection|WSpeed|SkyCond|LiquidPrecOne|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|file:/C:/Scala_ID...|01|01|00|-57|-62|10301|180|15|9|-|
|file:/C:/Scala_ID...|01|01|01|-59|-64|10293|170|5|9|-|
|file:/C:/Scala_ID...|01|01|02|-57|-61|10286|190|10|9|-|
|file:/C:/Scala_ID...|01|01|03|-54|-58|10282|180|10|9|-|
|file:/C:/Scala_ID...|01|01|04|-49|-53|10275|230|15|9|-|
|file:/C:/Scala_ID...|01|01|05|-47|-51|10268|200|15|8|-|
|file:/C:/Scala_ID...|01|01|06|-47|-51|10265|210|15|8|-|
|file:/C:/Scala_ID...|01|01|07|-48|-52|10259|200|15|8|-|
|file:/C:/Scala_ID...|01|01|08|-52|-56|10256|190|15|6|-|
|file:/C:/Scala_ID...|01|01|09|-42|-45|10251|180|15|8|-|
|file:/C:/Scala_ID...|01|01|10|-31|-32|10247|210|21|8|-|
|file:/C:/Scala_ID...|01|01|11|-23|-24|10238|220|26|8|-|
|file:/C:/Scala_ID...|01|01|12|-20|-20|10233|230|26|8|-|
|file:/C:/Scala_ID...|01|01|13|-21|-22|10226|230|36|8|-|
|file:/C:/Scala_ID...|01|01|14|-18|-18|10222|220|41|8|-|
|file:/C:/Scala_ID...|01|01|15|-8|-8|10222|230|31|8|-|
|file:/C:/Scala_ID...|01|01|16|-4|-4|10224|230|31|8|-|
|file:/C:/Scala_ID...|01|01|17|-2|-2|10224|240|31|8|-|
|file:/C:/Scala_ID...|01|01|18|0|0|10225|250|31|8|-|
|file:/C:/Scala_ID...|01|01|19|1|1|10224|240|26|8|-|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 20 rows

```

In [99]:

```
#Trimming the filename column to extract the station code. Prepping the Data Frame for analysis.
dfIsdData = df.toPandas()
dfIsdData.columns = ['Station_Code', 'Month', 'Day', 'Hour', 'Air_Temp' \
, 'Dew_Point_Temp', 'Sea_Pressure', 'Wind_Direction', 'Wind_Speed', 'Sky_Condition' \
, 'Liquid_Precipitation_One_Hour', 'Liquid_Precipitation_Six_Hour', 'Year']
dfIsdData['Station_Code'] = dfIsdData['Station_Code'].str[-20:].str[:6] # - Station code extraction fr
dfIsdData
```

Out[99]:

	Station_Code	Month	Day	Hour	Air_Temp	Dew_Point_Temp	Sea_Pressure	Wind_Direction	Wind_Speed	Sky_C
0	071810	01	01	00	-57	-62	10301	180	15	
1	071810	01	01	01	-59	-64	10293	170	5	
2	071810	01	01	02	-57	-61	10286	190	10	
3	071810	01	01	03	-54	-58	10282	180	10	
4	071810	01	01	04	-49	-53	10275	230	15	
5	071810	01	01	05	-47	-51	10268	200	15	
6	071810	01	01	06	-47	-51	10265	210	15	
7	071810	01	01	07	-48	-52	10259	200	15	
8	071810	01	01	08	-52	-56	10256	190	15	
9	071810	01	01	09	-42	-45	10251	180	15	
10	071810	01	01	10	-31	-32	10247	210	21	
11	071810	01	01	11	-23	-24	10238	220	26	
12	071810	01	01	12	-20	-20	10233	230	26	
13	071810	01	01	13	-21	-22	10226	230	36	
14	071810	01	01	14	-18	-18	10222	220	41	
15	071810	01	01	15	-8	-8	10222	230	31	
16	071810	01	01	16	-4	-4	10224	230	31	
17	071810	01	01	17	-2	-2	10224	240	31	
18	071810	01	01	18	0	0	10225	250	31	
19	071810	01	01	19	1	1	10224	240	26	
20	071810	01	01	20	1	1	10227	260	26	
21	071810	01	01	21	-2	-2	10231	290	26	
22	071810	01	01	22	-3	-3	10231	240	26	
23	071810	01	01	23	-2	-2	10233	230	36	
24	071810	01	02	00	-14	-14	10233	240	10	
25	071810	01	02	01	-7	-7	10231	270	15	
26	071810	01	02	02	-5	-5	10233	260	15	
27	071810	01	02	03	-4	-4	10231	240	21	
28	071810	01	02	04	-4	-4	10229	250	31	
29	071810	01	02	05	-7	-7	10233	260	26	
...
78201	318660	12	28	06	-83	-207	10182	230	20	
78202	318660	12	28	09	-98	-232	10162	290	80	
78203	318660	12	28	12	-122	-227	10143	220	50	
78204	318660	12	28	15	-112	-241	10114	270	90	
78205	318660	12	28	18	-121	-249	10117	350	20	
78206	318660	12	28	21	-120	-248	10113	300	30	
78207	318660	12	29	00	-125	-258	10121	320	30	
78208	318660	12	29	03	-92	-241	10107	250	10	

	Station_Code	Month	Day	Hour	Air_Temp	Dew_Point_Temp	Sea_Pressure	Wind_Direction	Wind_Speed	Sky_C
78209	318660	12	29	06	-98	-237	10115	280	60	
78210	318660	12	29	09	-152	-242	10131	200	10	
78211	318660	12	29	12	-175	-244	10145	290	30	
78212	318660	12	29	15	-191	-243	10141	300	30	
78213	318660	12	29	18	-155	-258	10144	270	20	
78214	318660	12	29	21	-137	-257	10135	240	20	
78215	318660	12	30	00	-147	-232	10142	240	20	
78216	318660	12	30	03	-105	-242	10137	290	30	
78217	318660	12	30	06	-93	-218	10128	0	-9999	
78218	318660	12	30	09	-94	-202	10138	240	30	
78219	318660	12	30	12	-97	-219	10142	10	10	
78220	318660	12	30	15	-116	-208	10141	190	30	
78221	318660	12	30	18	-158	-214	10166	300	30	
78222	318660	12	30	21	-192	-225	10186	250	20	
78223	318660	12	31	00	-156	-207	10206	230	40	
78224	318660	12	31	03	-47	-201	10195	250	10	
78225	318660	12	31	06	-44	-196	10197	290	40	
78226	318660	12	31	09	-71	-188	10209	270	60	
78227	318660	12	31	12	-76	-189	10207	260	50	
78228	318660	12	31	15	-72	-195	10200	270	70	
78229	318660	12	31	18	-76	-199	10193	280	70	
78230	318660	12	31	21	-120	-194	10185	220	30	

78231 rows × 13 columns

In [53]:

```
# information on the columns
dfIsdData.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 78231 entries, 0 to 78230
Data columns (total 13 columns):
Station_Code      78231 non-null object
Month             78231 non-null object
Day              78231 non-null object
Hour             78231 non-null object
Air_Temp          78231 non-null object
Dew_Point_Temp    78231 non-null object
Sea_Pressure      78231 non-null object
Wind_Direction    78231 non-null object
Wind_Speed        78231 non-null object
Sky_Condition     78231 non-null object
Liquid_Precipitation_One_Hour 78231 non-null object
Liquid_Precipitation_Six_Hour 78231 non-null object
Year              78231 non-null int32
dtypes: int32(1), object(12)
memory usage: 7.5+ MB
```

In [54]:

```
# number of rows
len(dfIsdData)
```

Out[54]:

```
78231
```

In [55]:

```
#Filter for year = 2018
dfIsdData[dfIsdData.Year == 2017]
```

Out[55]:

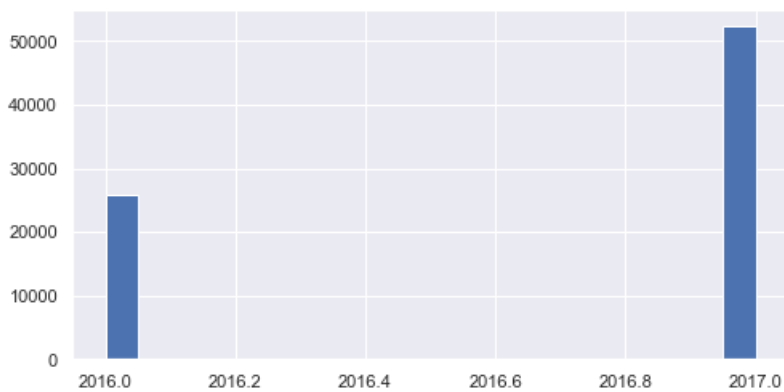
	Station_Code	Month	Day	Hour	Air_Temp	Dew_Point_Temp	Sea_Pressure	Wind_Direction	Wind_Speed	Sky_C
0	071810	01	01	00	-57	-62	10301	180	15	
1	071810	01	01	01	-59	-64	10293	170	5	
2	071810	01	01	02	-57	-61	10286	190	10	
3	071810	01	01	03	-54	-58	10282	180	10	
4	071810	01	01	04	-49	-53	10275	230	15	
5	071810	01	01	05	-47	-51	10268	200	15	
6	071810	01	01	06	-47	-51	10265	210	15	
7	071810	01	01	07	-48	-52	10259	200	15	
8	071810	01	01	08	-52	-56	10256	190	15	
9	071810	01	01	09	-42	-45	10251	180	15	
10	071810	01	01	10	-31	-32	10247	210	21	
11	071810	01	01	11	-23	-24	10238	220	26	
12	071810	01	01	12	-20	-20	10233	230	26	
13	071810	01	01	13	-21	-22	10226	230	36	
14	071810	01	01	14	-18	-18	10222	220	41	
15	071810	01	01	15	-8	-8	10222	230	31	
16	071810	01	01	16	-4	-4	10224	230	31	
17	071810	01	01	17	-2	-2	10224	240	31	
18	071810	01	01	18	0	0	10225	250	31	
19	071810	01	01	19	1	1	10224	240	26	
20	071810	01	01	20	1	1	10227	260	26	
21	071810	01	01	21	-2	-2	10231	290	26	
22	071810	01	01	22	-3	-3	10231	240	26	
23	071810	01	01	23	-2	-2	10233	230	36	
24	071810	01	02	00	-14	-14	10233	240	10	
25	071810	01	02	01	-7	-7	10231	270	15	
26	071810	01	02	02	-5	-5	10233	260	15	
27	071810	01	02	03	-4	-4	10231	240	21	
28	071810	01	02	04	-4	-4	10229	250	31	
29	071810	01	02	05	-7	-7	10233	260	26	
...
61005	071825	12	30	18	120	110	-9999	220	88	
61006	071825	12	30	19	120	110	-9999	230	103	
61007	071825	12	30	20	120	110	-9999	230	108	
61008	071825	12	30	21	120	100	-9999	220	103	
61009	071825	12	30	22	120	100	-9999	220	98	
61010	071825	12	30	23	120	90	-9999	220	88	
61011	071825	12	31	00	110	90	-9999	220	82	
61012	071825	12	31	01	110	90	-9999	210	88	
61013	071825	12	31	02	110	90	-9999	220	93	
61014	071825	12	31	03	110	90	-9999	210	82	
61015	071825	12	31	04	110	90	-9999	210	77	
61016	071825	12	31	05	110	80	-9999	210	88	

	Station_Code	Month	Day	Hour	Air_Temp	Dew_Point_Temp	Sea_Pressure	Wind_Direction	Wind_Speed	Sky_C
61017	071825	12	31	06	100	80	-9999	200	72	
61018	071825	12	31	07	100	80	-9999	190	88	
61019	071825	12	31	08	100	70	-9999	200	93	
61020	071825	12	31	09	100	80	-9999	200	88	
61021	071825	12	31	10	110	80	-9999	200	93	
61022	071825	12	31	11	120	80	-9999	190	88	
61023	071825	12	31	12	120	70	-9999	190	103	
61024	071825	12	31	13	120	70	-9999	200	82	
61025	071825	12	31	14	120	70	-9999	190	72	
61026	071825	12	31	15	110	60	-9999	180	67	
61027	071825	12	31	16	110	60	-9999	180	62	
61028	071825	12	31	17	110	50	-9999	190	62	
61029	071825	12	31	18	100	60	-9999	190	72	
61030	071825	12	31	19	100	60	-9999	190	88	
61031	071825	12	31	20	100	60	-9999	200	82	
61032	071825	12	31	21	90	60	-9999	250	129	
61033	071825	12	31	22	80	60	-9999	230	72	
61034	071825	12	31	23	80	50	-9999	220	88	

52299 rows × 13 columns

In [56]:

```
#This below shows the number of rows per year.
#vis1 = sns.distplot(pd.to_numeric(dfIsdData["Year"]))
vis1Plt = plt.hist(pd.to_numeric(dfIsdData["Year"]), bins=20)
```

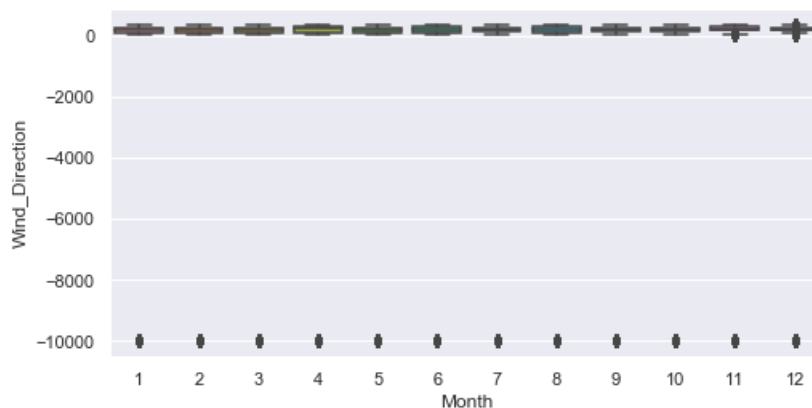


In [100]:

```
#Prepping the Data
dfIsdData["Year"] = pd.to_numeric(dfIsdData["Year"])
dfIsdData["Station_Code"] = pd.to_numeric(dfIsdData["Station_Code"])
dfIsdData["Month"] = pd.to_numeric(dfIsdData["Month"])
dfIsdData["Day"] = pd.to_numeric(dfIsdData["Day"])
dfIsdData["Hour"] = pd.to_numeric(dfIsdData["Hour"])
dfIsdData["Air_Temp"] = pd.to_numeric(dfIsdData["Air_Temp"])
dfIsdData["Dew_Point_Temp"] = pd.to_numeric(dfIsdData["Dew_Point_Temp"])
dfIsdData["Sea_Pressure"] = pd.to_numeric(dfIsdData["Sea_Pressure"])
dfIsdData["Wind_Direction"] = pd.to_numeric(dfIsdData["Wind_Direction"])
dfIsdData["Wind_Speed"] = pd.to_numeric(dfIsdData["Wind_Speed"])
dfIsdData["Sky_Condition"] = pd.to_numeric(dfIsdData["Sky_Condition"])
dfIsdData["Liquid_Precipitation_One_Hour"] = pd.to_numeric(dfIsdData["Liquid_Precipitation_One_Hour"])
dfIsdData["Liquid_Precipitation_Six_Hour"] = pd.to_numeric(dfIsdData["Liquid_Precipitation_Six_Hour"])
dfIsdData = dfIsdData[((dfIsdData.Air_Temp != -9999) & dfIsdData.Wind_Speed != -9999 & \
                        (dfIsdData.Wind_Direction != -9999) & (dfIsdData.Dew_Point_Temp != -9999) & \
                        (dfIsdData.Sea_Pressure != -9999) & (dfIsdData.Sky_Condition != -9999) & \
                        (dfIsdData.Liquid_Precipitation_One_Hour != -9999) & (dfIsdData.Liquid_Precipit
```

In [58]:

```
#BoxPlots
#dfIsdData.boxplot(column='Sky Condition',layout=(1,9), figsize=(20,10), whis=[5,95])
dfIsdDataBP = dfIsdData[(dfIsdData.Year == 2017)]
dfIsdDataBP.groupby(['Year', 'Month'])['Wind_Direction'].max()
vis2 = sns.boxplot(data=dfIsdDataBP, x=dfIsdDataBP["Month"], y=dfIsdDataBP["Wind_Direction"])
#vis2.set(xlabel='Max Value Hourly basis', ylabel='Wind Direction')
```



In [59]:

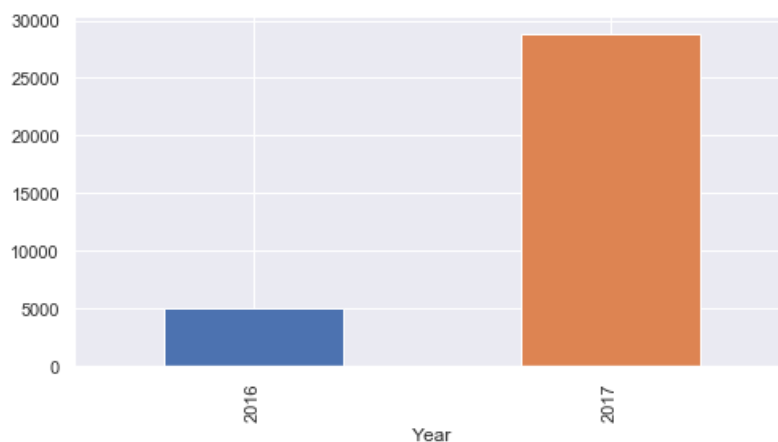
```
#BoxPlots
#dfIsdData["Station_Code"] = pd.to_numeric(dfIsdData["Station_Code"])
#dfIsdData["Year"] = pd.to_numeric(dfIsdData["Year"])

#f, (ax) = plt.subplots(1, 1, figsize=(12, 4))
#f.suptitle('Year - Station_Code', fontsize=14)

#sns.boxplot(x="Station_Code", y="Year", data=dfIsdData, ax=ax)
#ax.set_xlabel("Station_Code",size = 12,alpha=0.8)
#ax.set_ylabel("Year",size = 12,alpha=1.0)
```

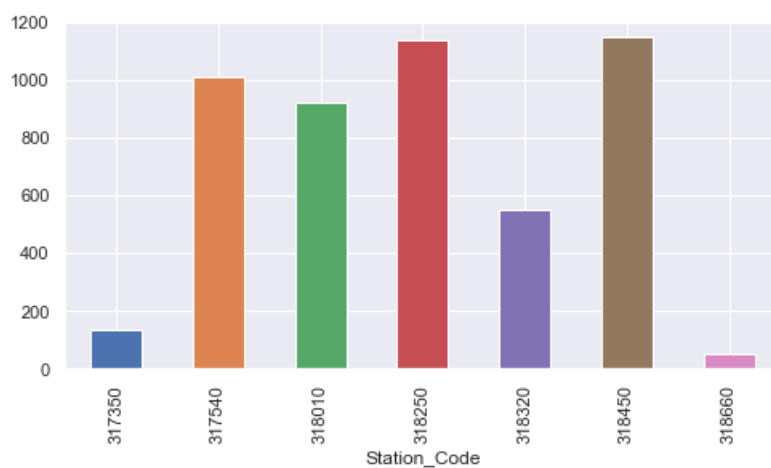
```
In [60]: #This below shows the number of stations data captured per year.  
dfIsdData.groupby(['Year'])['Year'].count().plot(kind='bar')
```

```
Out[60]: <matplotlib.axes._subplots.AxesSubplot at 0x282fa164eb8>
```



```
In [61]: #In a particular year, how much data is captured at each station  
dfIsdDataBar = dfIsdData[(dfIsdData.Year == 2016)]  
dfIsdDataBar.groupby(['Station_Code'])['Station_Code'].count().plot(kind='bar')
```

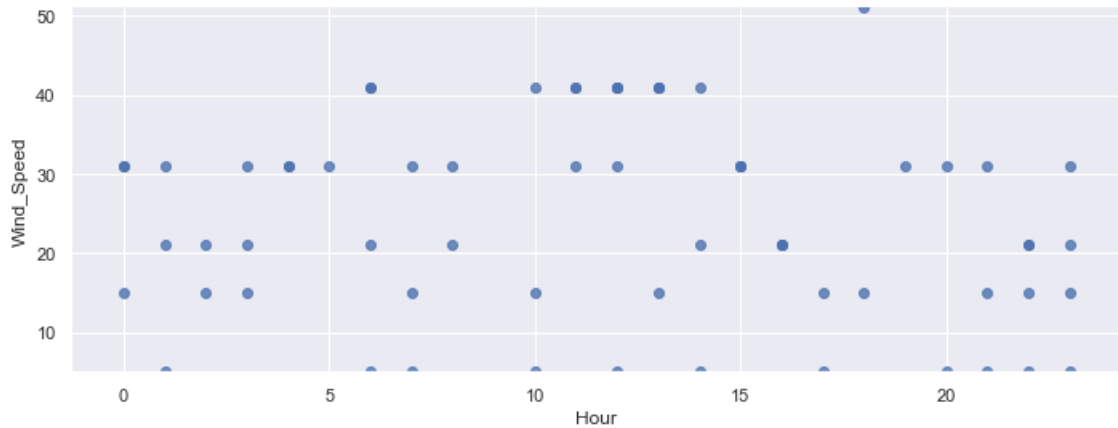
```
Out[61]: <matplotlib.axes._subplots.AxesSubplot at 0x282f9c222b0>
```



In [62]:

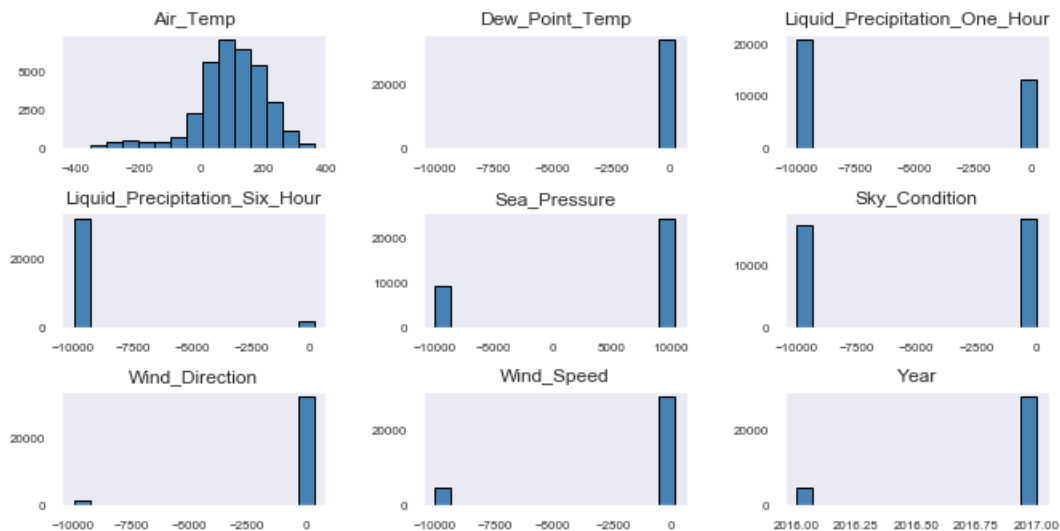
```
#Wind_Speed hourly
dfIsdDataCompare = dfIsdData[(dfIsdData.Year == 2017) & (dfIsdData.Month == 1) & (dfIsdData.Day == 26)]
vis3 = sns.lmplot(data=dfIsdDataCompare, x='Hour', y='Wind_Speed', fit_reg=False, size = 4, aspect = 2.
vis3.axes[0,0].set_ylim(min(dfIsdDataCompare.Wind_Speed), max(dfIsdDataCompare.Wind_Speed))
```

Out[62]: (5, 51)



In [63]:

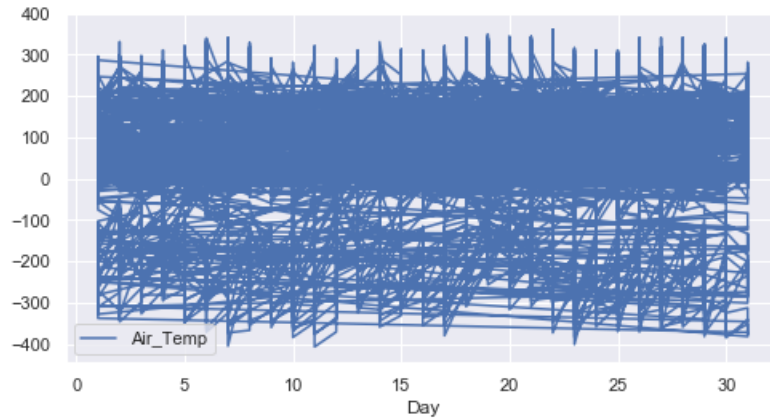
```
cols = ['Air_Temp' \
, 'Dew_Point_Temp', 'Sea_Pressure', 'Wind_Direction', 'Wind_Speed', 'Sky_Condition' \
, 'Liquid_Precipitation_One_Hour', 'Liquid_Precipitation_Six_Hour', 'Year']
dfIsdDataBar2 = dfIsdData[cols]
dfIsdDataBar2[cols].hist(bins=15, color='steelblue', edgecolor='black', linewidth=1.0,
                        xlabelsize=8, ylabelsize=8, grid=False)
plt.tight_layout(rect=(0, 0, 1.2, 1.2))
```



```
In [126]: #Time Series for a month in a year - Air_Temp
dfIsdDataTS = dfIsdData[(dfIsdData.Year == 2017)]
dfIsdDataTS = dfIsdData[(dfIsdData.Month == 2)]

dfIsdDataTS.groupby(['Station_Code', 'Month', 'Day'])['Air_Temp'].max()
cols = ['Day', 'Air_Temp']
dfIsdDataTS = dfIsdData[cols]
dfIsdDataTS.set_index('Day').plot()
```

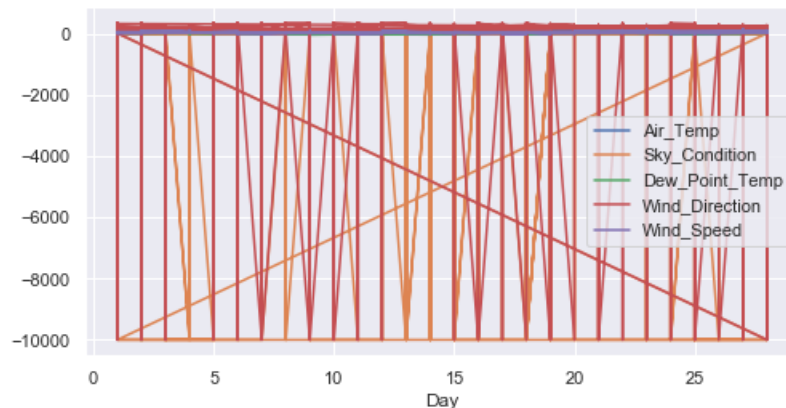
Out[126]: <matplotlib.axes._subplots.AxesSubplot at 0x282828c3780>



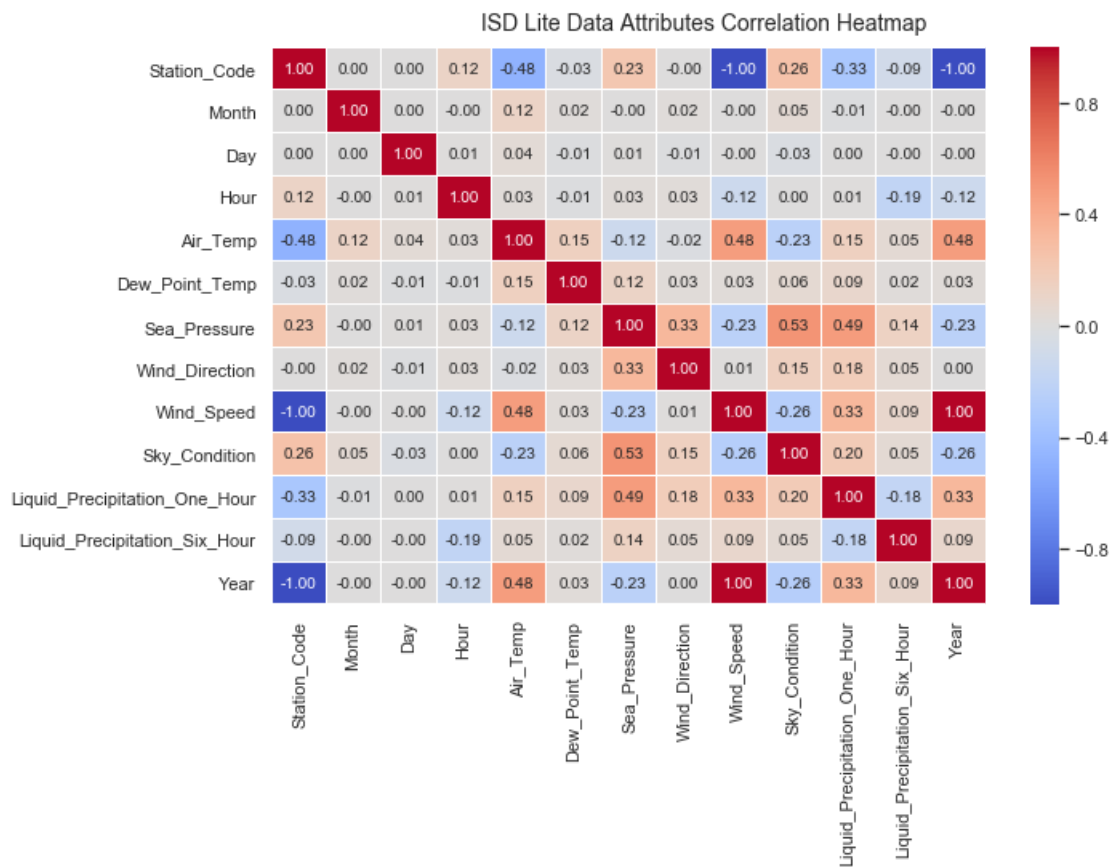
```
In [127]: #Time Series for a month in a year
dfIsdDataTS2 = dfIsdData[(dfIsdData.Year == 2017)]
dfIsdDataTS2 = dfIsdDataTS2[(dfIsdDataTS2.Month == 2)]

dfIsdDataTS2.groupby(['Station_Code', 'Month', 'Day'])['Air_Temp'].max()
cols = ['Day', 'Air_Temp', 'Sky_Condition', 'Dew_Point_Temp', 'Wind_Direction', 'Wind_Speed']
dfIsdDataTS2 = dfIsdDataTS2[cols]
dfIsdDataTS2.set_index('Day').plot()
```

Out[127]: <matplotlib.axes._subplots.AxesSubplot at 0x282835bc198>



```
In [122]: # Correlation Matrix Heatmap - showing the correlation between all attributes
f, ax = plt.subplots(figsize=(10, 6))
corr = dfIsdData.corr()
hm = sns.heatmap(round(corr,2), annot=True, ax=ax, cmap="coolwarm",fmt='.2f',
                    linewidths=.05)
f.subplots_adjust(top=0.93)
t= f.suptitle('ISD Lite Data Attributes Correlation Heatmap', fontsize=14)
```

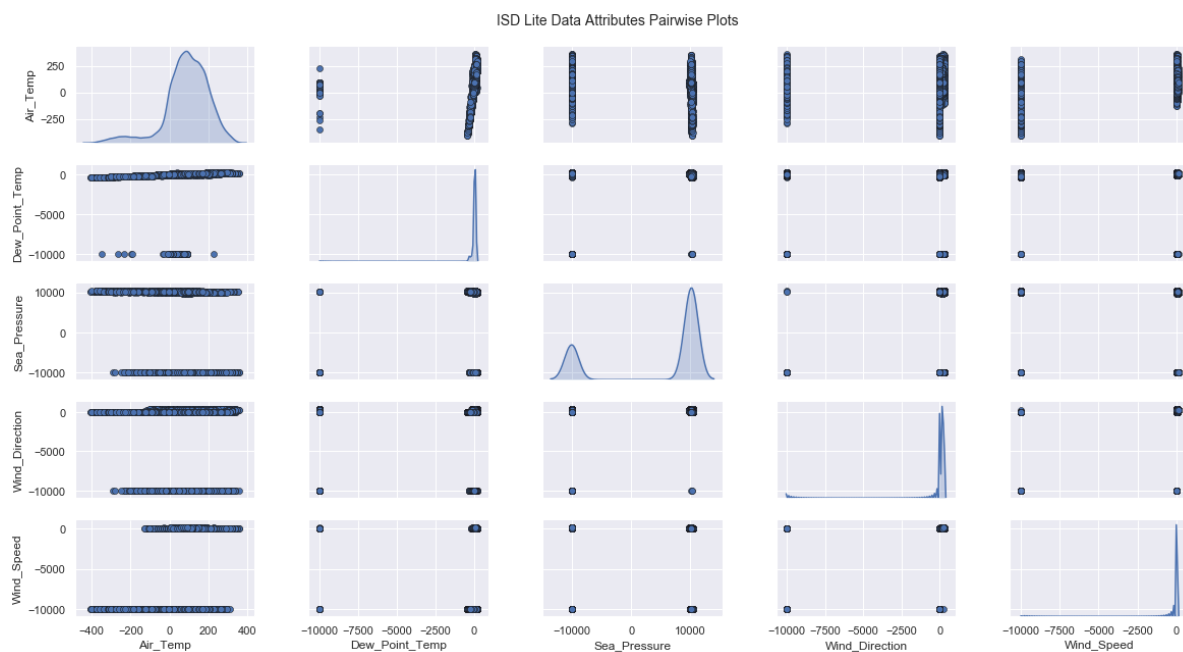


In [123]:

```
# Pair-wise Scatter Plots
cols = ['Air_Temp' \
        , 'Dew_Point_Temp', 'Sea_Pressure', 'Wind_Direction', 'Wind_Speed'] # 'Sky_Condition' \
        #, 'Liquid_Precipitation_One_Hour', 'Liquid_Precipitation_Six_Hour']

pp = sns.pairplot(dfIsdData[cols], size=1.8, aspect=1.8,
                  plot_kws=dict(edgecolor="k", linewidth=0.5),
                  diag_kind="kde", diag_kws=dict(shade=True))

fig = pp.fig
fig.subplots_adjust(top=0.93, wspace=0.3)
t = fig.suptitle('ISD Lite Data Attributes Pairwise Plots', fontsize=14)
```



In [129]:

```
# Violin Plots
dfIsdDataV = dfIsdData[(dfIsdData.Year == 2017)]
dfIsdDataV = dfIsdDataV[(dfIsdDataV.Month == 2)]

dfIsdDataV.groupby(['Station_Code', 'Month', 'Day'])['Air_Temp'].max()
cols = ['Day', 'Air_Temp']
dfIsdDataV = dfIsdDataV[cols]
f, (ax) = plt.subplots(1, 1, figsize=(12, 4))
f.suptitle('Day - Air Temp', fontsize=14)

sns.violinplot(x="Day", y="Air_Temp", data=dfIsdDataV, ax=ax)
ax.set_xlabel("Day", size = 12, alpha=0.8)
ax.set_ylabel("Air_Temp", size = 12, alpha=0.8)
```

Out[129]:

Text(0, 0.5, 'Air_Temp')

