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 [12]:
```
from pyspark.sql import SparkSession
import pandas as pd
import numpy as np
import pyspark.sql as sparksql
import warnings
import matplotlib.pyplot as plt
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', '5gb') \
    .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\\pa
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|LiquidPrecOneH
+--------------------+-----+---+----+-----+-----+--------+----------+------+-------+-------------
|file:/C:/Scala_ID...|   01| 01|  00|  -57|  -62|   10301|       180|    15|      9|            -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|            -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|            -9
|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|            -9
|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|            -9
|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|            -9
|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|            -9
|file:/C:/Scala_ID...|   01| 01|  19|    1|    1|   10224|       240|    26|      8|
+--------------------+-----+---+----+-----+-----+--------+----------+------+-------+-------------
only showing top 20 rows
```

In [13]:
```
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[13]:

	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 [14]:
```
# 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 [15]:
```
# number of rows
len(dfIsdData)
```

Out[15]: 78231

In [17]:
```
#Filter for year = 1935
dfIsdData[dfIsdData.Year == 2016]
```

Out[17]:

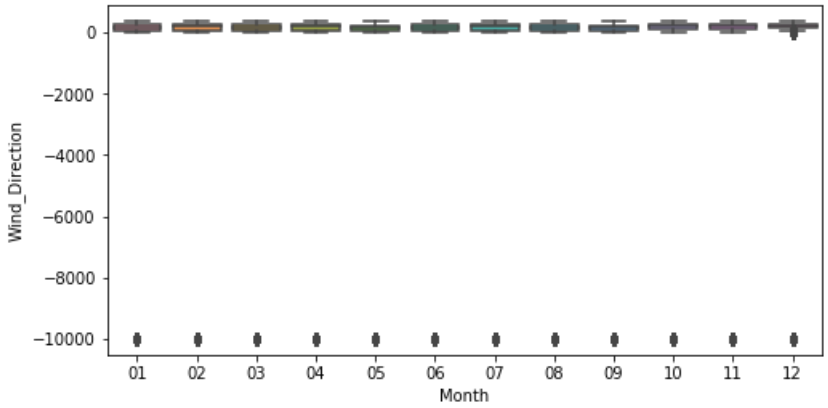
	Station_Code	Month	Day	Hour	Air_Temp	Dew_Point_Temp	Sea_Pressure	Wind_Direction	Wind_Speed	Sky_C
34905	317350	01	01	00	-222	-248	10249	210	20	
34906	317350	01	01	01	-200	-230	-9999	170	20	
34907	317350	01	01	02	-180	-200	-9999	-9999	10	
34908	317350	01	01	03	-163	-187	10229	150	20	
34909	317350	01	01	04	-140	-180	-9999	150	20	
34910	317350	01	01	05	-140	-180	-9999	0	0	
34911	317350	01	01	06	-138	-179	10213	150	10	
34912	317350	01	01	07	-140	-180	-9999	-9999	-9999	
34913	317350	01	01	08	-140	-170	-9999	70	20	
34914	317350	01	01	09	-143	-173	10199	70	20	
34915	317350	01	01	10	-140	-170	-9999	90	20	
34916	317350	01	01	11	-160	-180	-9999	-9999	10	
34917	317350	01	01	12	-160	-179	10183	350	10	
34918	317350	01	01	13	-150	-170	-9999	20	20	
34919	317350	01	01	14	-150	-170	-9999	10	20	
34920	317350	01	01	15	-157	-174	10162	350	20	
34921	317350	01	01	16	-160	-170	-9999	-9999	10	
34922	317350	01	01	17	-170	-180	-9999	340	20	
34923	317350	01	01	18	-169	-188	10146	340	20	
34924	317350	01	01	19	-180	-200	-9999	350	20	
34925	317350	01	01	20	-180	-210	-9999	350	30	
34926	317350	01	01	21	-189	-210	10134	340	20	
34927	317350	01	01	22	-190	-210	-9999	320	20	
34928	317350	01	01	23	-190	-210	-9999	310	20	
34929	317350	01	02	00	-183	-203	10137	310	10	
34930	317350	01	02	01	-180	-200	-9999	340	10	
34931	317350	01	02	02	-160	-190	-9999	280	40	
34932	317350	01	02	03	-156	-183	10135	270	40	
34933	317350	01	02	04	-140	-170	-9999	270	40	
34934	317350	01	02	05	-140	-170	-9999	260	50	
...	...	...	...	...	...	...	...	...	...	
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	
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	

	Station_Code	Month	Day	Hour	Air_Temp	Dew_Point_Temp	Sea_Pressure	Wind_Direction	Wind_Speed	Sky_C
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	

25932 rows × 13 columns

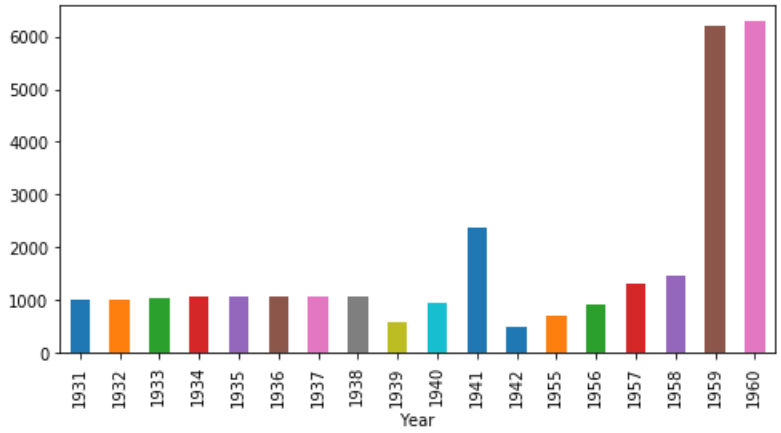
In [18]:
```
#Distribution
#vis1 = sns.distplot(pd.to_numeric(dfIsdData["Year"]))
vis1Plt = plt.hist(pd.to_numeric(dfIsdData["Year"]), bins=20)
```

In [21]:
```
#BoxPlots
#dfIsdData.boxplot(column='Sky Condition',layout=(1,9), figsize=(20,10), whis=[5,95])
#dfIsdData["Month"] = pd.to_numeric(dfIsdData["Month"])
#dfIsdData["Year"] = pd.to_numeric(dfIsdData["Year"])
#dfIsdData["Hour"] = pd.to_numeric(dfIsdData["Hour"])
dfIsdData["Hour"] = pd.to_numeric(dfIsdData["Hour"])
dfIsdData["Wind_Direction"] = pd.to_numeric(dfIsdData["Wind_Direction"])
#dfIsdData[ (dfIsdData.Wind_Direction > 0) and (dfIsdData.Year == 1957) ]
dfIsdData[ (dfIsdData.Wind_Direction > 0)]
#dfIsdData = dfIsdData[dfIsdData.Month == 1]
dfIsdData.groupby(['Year'])['Month'].max()
vis2 = sns.boxplot(data=dfIsdData, x=dfIsdData["Month"], y=dfIsdData["Wind_Direction"])
#vis2.set(xlabel='Max Value Hourly basis', ylabel='Wind Direction')
```



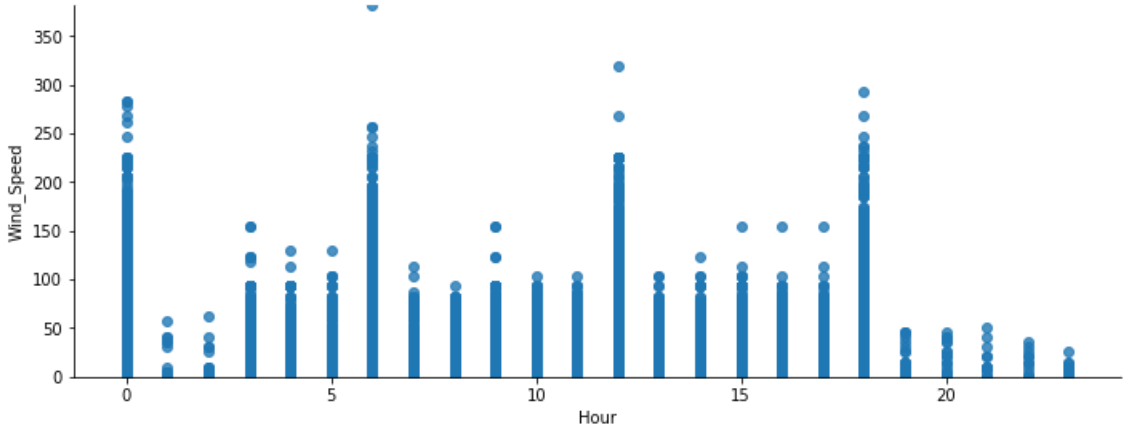
In [351]:
```
dfIsdData.groupby(['Year'])['Year'].count().plot(kind='bar')
```

Out[351]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x25ecefd3b70>
```

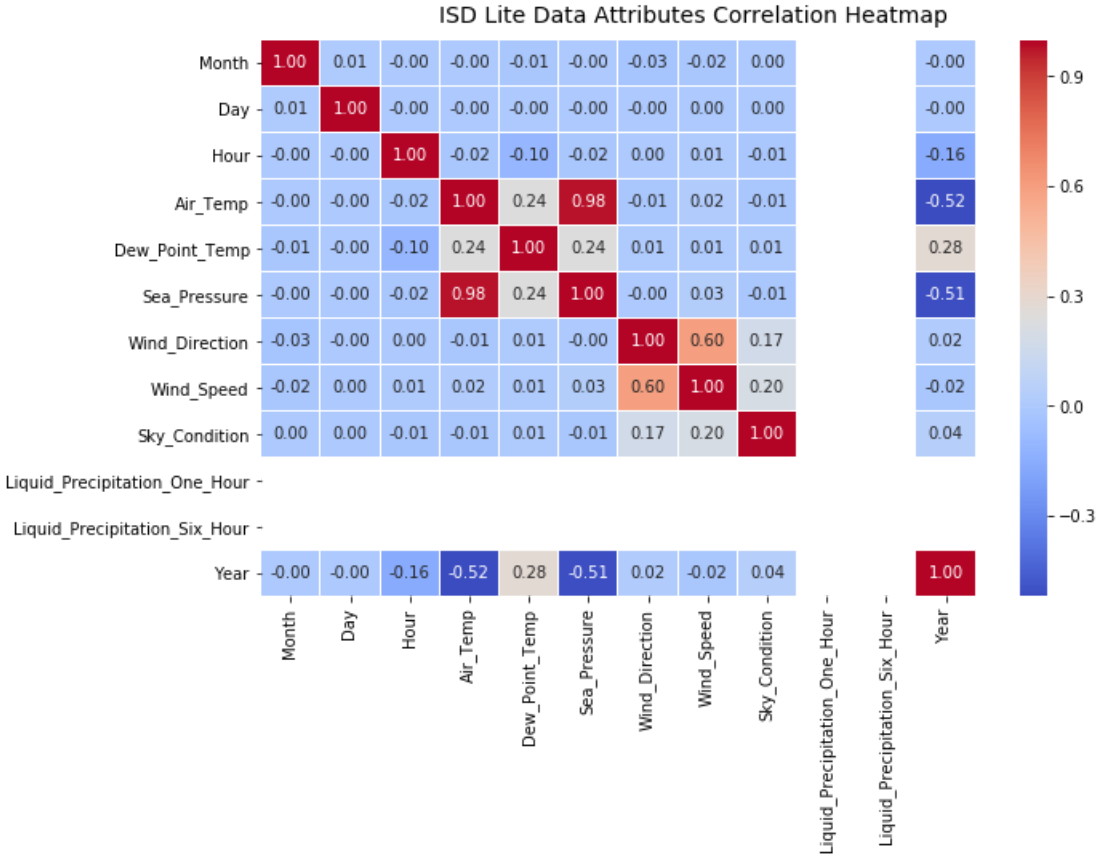


In [416]:
```
dfIsdData["Hour"] = pd.to_numeric(dfIsdData["Hour"])
dfIsdData["Year"] = pd.to_numeric(dfIsdData["Year"])
dfIsdData["Month"] = pd.to_numeric(dfIsdData["Month"])
dfIsdData["Day"] = pd.to_numeric(dfIsdData["Day"])
dfIsdData["Wind_Speed"] = pd.to_numeric(dfIsdData["Wind_Speed"])
dfIsdData = dfIsdData[dfIsdData.Wind_Speed != -9999]
#dfIsdData[ (dfIsdData.Wind_Direction > 0) and
dfIsdData[(dfIsdData.Year == 1957) & (dfIsdData.Month == 1) & (dfIsdData.Day == 20)]
vis3 = sns.lmplot(data=dfIsdData,x='Hour',y='Wind_Speed', fit_reg=False, size = 4, aspect = 2.5)
vis3.axes[0,0].set_ylim(min(dfIsdData.Wind_Speed), max(dfIsdData.Wind_Speed))
```

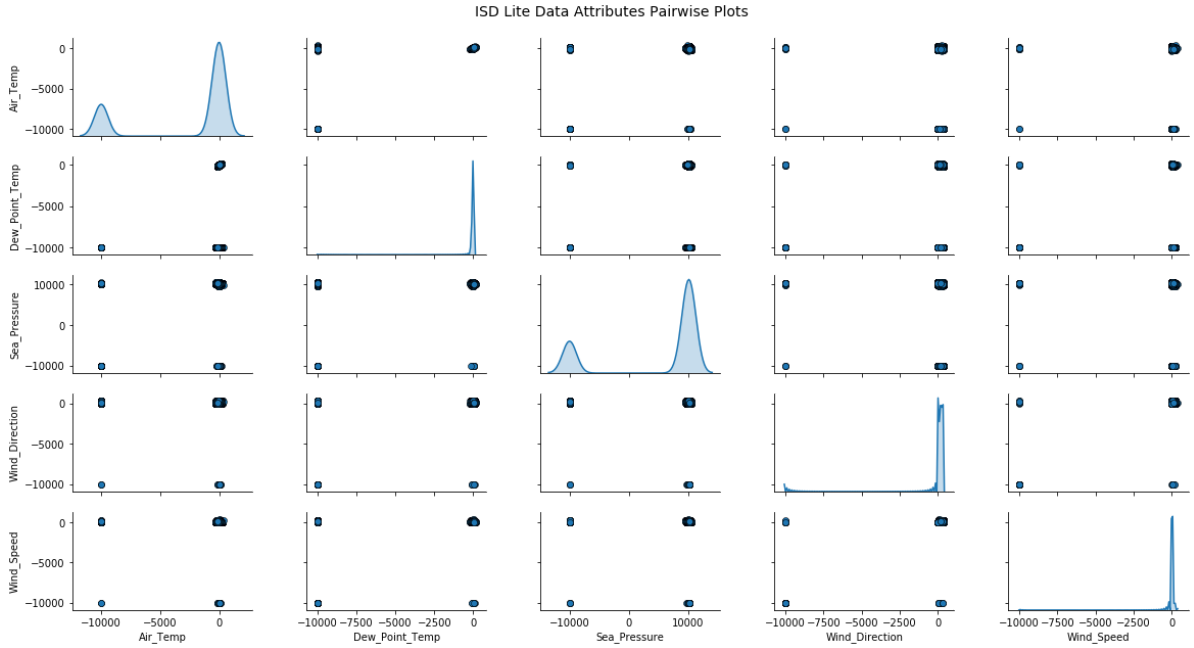
Out[416]: (0, 381)



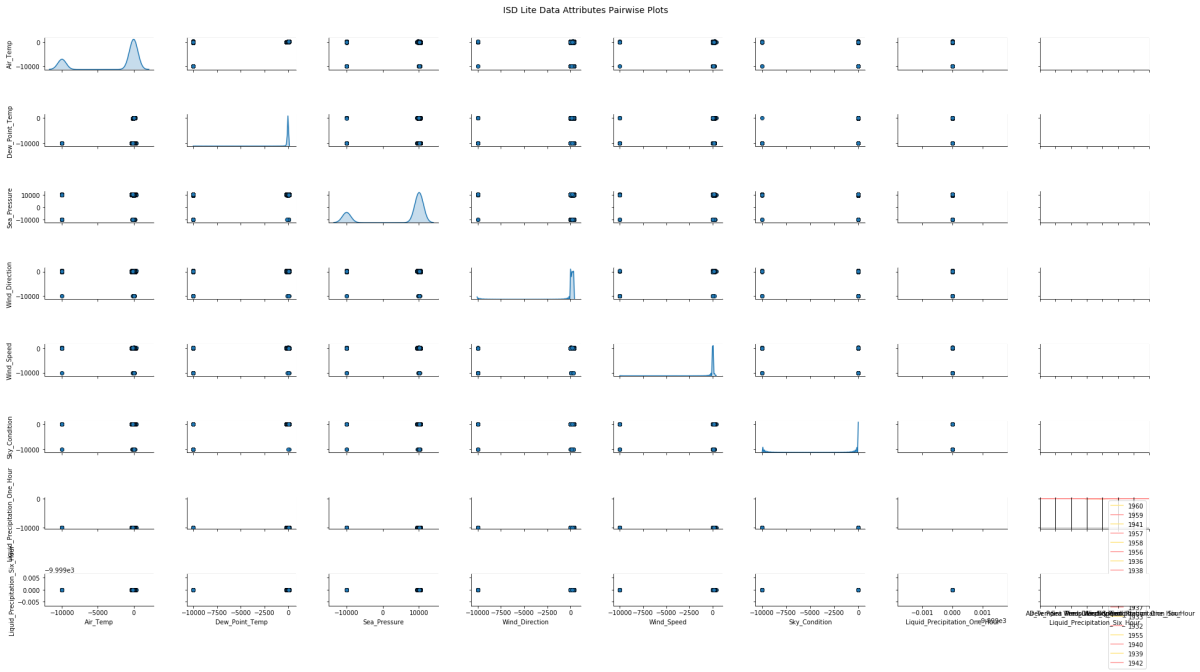
In [359]:
```
# Correlation Matrix Heatmap
dfIsdData["Year"] = pd.to_numeric(dfIsdData["Year"])
dfIsdData["Month"] = pd.to_numeric(dfIsdData["Month"])
dfIsdData["Day"] = pd.to_numeric(dfIsdData["Day"])
dfIsdData["Hour"] = pd.to_numeric(dfIsdData["Hour"])
dfIsdData["Air_Temp"] = pd.to_numeric(dfIsdData["Air_Temp"])
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"]
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 [368]:
```
# 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)
```

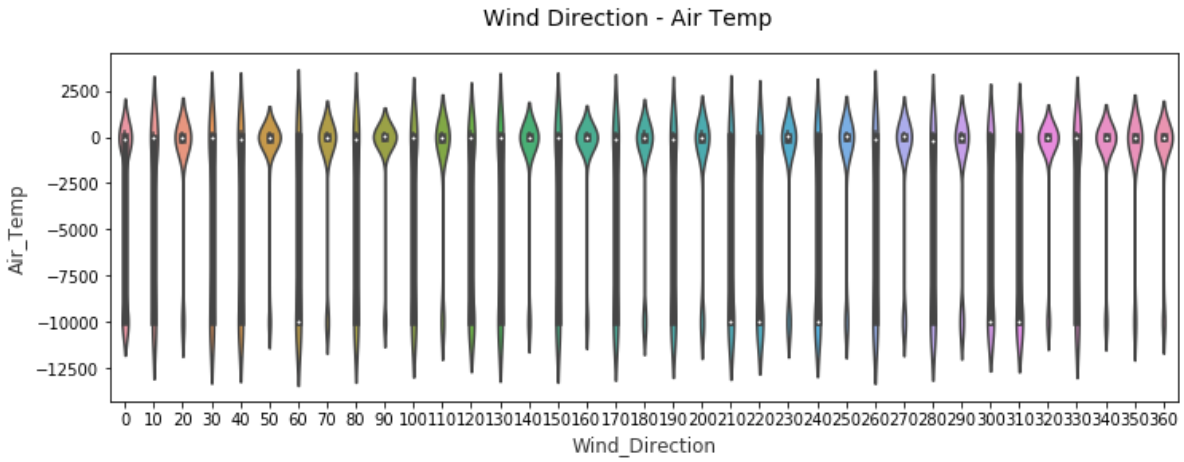


In [370]:
```
# Scaling attribute values to avoid few outiers
cols = ['Air_Temp' \
,'Dew_Point_Temp','Sea_Pressure', 'Wind_Direction', 'Wind_Speed', 'Sky_Condition' \
,'Liquid_Precipitation_One_Hour', 'Liquid_Precipitation_Six_Hour']
pp = sns.pairplot(dfIsdData[cols], size=1.8, aspect=1.8,
                  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)
subset_df = dfIsdData[cols]

from sklearn.preprocessing import StandardScaler
ss = StandardScaler()

scaled_df = ss.fit_transform(subset_df)
scaled_df = pd.DataFrame(scaled_df, columns=cols)
final_df = pd.concat([scaled_df, dfIsdData['Year']], axis=1)
final_df.head()

# plot parallel coordinates
from pandas.plotting import parallel_coordinates
pc = parallel_coordinates(final_df, 'Year', color=('#FFE888', '#FF9999'))
```



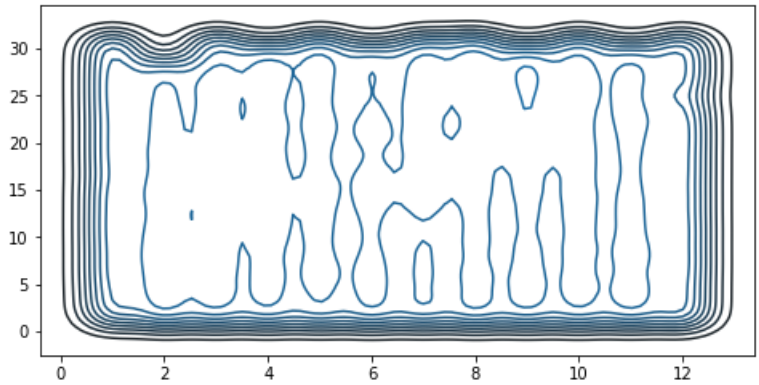
In [402]:
```
# Violin Plots
dfIsdData = dfIsdData[dfIsdData.Wind_Direction != -9999]
f, (ax) = plt.subplots(1, 1, figsize=(12, 4))
f.suptitle('Wind Direction - Air Temp', fontsize=14)

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

Out[402]: Text(0, 0.5, 'Air_Temp')



In [401]:
```
# Visualizing 3-D mix data using kernel density plots
#dfIsdData = dfIsdData[dfIsdData.Wind_Direction != -9999]
#ax = sns.kdeplot(dfIsdData['Air_Temp'], dfIsdData['Dew_Point_Temp'],
#                 cmap="YlOrBr", shade=True, shade_lowest=False)
#ax = sns.kdeplot(dfIsdData['Air_Temp'], dfIsdData['Wind_Direction'],
#                  cmap="Reds", shade=True, shade_lowest=False)
#ax = sns.kdeplot(dfIsdData['Air_Temp'], dfIsdData['Sky_Condition'],
#                  cmap="Reds", shade=True, shade_lowest=False)
```

In [392]:
```
sns.kdeplot(dfIsdData);
```



In [396]:
```
with sns.axes_style('white'):
    g = sns.jointplot("Air_Temp", "Dew_Point_Temp", dfIsdData, kind='hex')
    g.ax_joint.plot(np.linspace(4000, 16000),
                    np.linspace(8000, 32000), ':k')
```