Table of Contents

WiDS Datathon 2023 - Introduction

Adapting to Climagte Change by Improving Extreme Weather Forecasts

From WiDS Website: **Problem Statement:** "Extreme weather events are sweeping the globe and range from heat waves, wildfires and drought to hurricanes, extreme rainfall and flooding. These weather events have multiple impacts on agriculture, energy, transportation, as well as low resource communities and disaster planning in countries across the globe.

Accurate long-term forecasts of temperature and precipitation are crucial to help people prepare and adapt to these extreme weather events. Currently, purely physics-based models dominate short-term weather forecasting. But these models have a limited forecast horizon. The availability of meteorological data offers an opportunity for data scientists to improve sub-seasonal forecasts by blending physics-based forecasts with machine learning. Sub-seasonal forecasts for weather and climate conditions (lead-times ranging from 15 to more than 45 days) would help communities and industries adapt to the challenges brought on by climate change."

Data Overview

From WiDS Website "The WiDS Datathon 2023 focuses on a prediction task involving forecasting sub-seasonal temperatures (temperatures over a two-week period, in our case) within the United States. We are using a pre-prepared dataset consisting of weather and climate information for a number of US locations, for a number of start dates for the two-week observation, as well as the forecasted temperature and precipitation from a number of weather forecast models (we will reveal the source of our dataset after the competition closes). Each row in the data corresponds to a single location and a single start date for the two-week period. Your task is to predict the arithmetic mean of the maximum and minimum temperature over the next 14 days, for each location and start date.

You are provided with two datasets:

- train_data.csv: the training dataset, where contest-tmp2m-14d__tmp2m, the arithmetic mean of the max and min observed temperature over the next 14 days for each location and start date, is provided
- test_data.csv: the test dataset, where we withhold the true value of contest-tmp2m-14d__tmp2m for each row."

Target: `contest-tmp2m-14d__tmp2m`

Notebook 1 - Introduction

Workflow Initital EDA --> Data Cleaning --> More EDA --> First Modeling Attempt - Random Forest Regressor

In notebook 1, we will explore the training data, `training_data`, and test data, `test_data` to gain insights into the problem space and strategize forecasting methods. We will then clean the data to prepare it for modeling. Some more analysis will be completed

with the clean training data. We will attempt our first model and set our baseline error metrics. To conclude the notebook, we will summarize our findings and establish next steps. In Notebook 2, we will take on a more comprehensive modeling with feature engineering and strategize how to reduce run time to optimize model performance.

In [9]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from datetime import datetime
```

In [10]:
```
#Read in training set
training_data = pd.read_csv('data/train_data.csv')
```

Initial EDA of training_data

In [11]:
```
#How big is this data set
training_data.shape
```

Out[11]:
```
(375734, 246)
```

training_data Size

375,734 rows and 246 columns

In [12]:
```
#Check out training Data
pd.set_option('display.max_columns', 500)
pd.set_option('display.max_rows', 100)
training_data.head(100)
```

Out[12]:

	index	lat	lon	startdate	contest-pevpr-sfc-gauss-14d__pevpr	nmme0-tmp2m-34w__cancm30	nmme0-tmp2m-34w__cancm40	nmme tmp2r 34w__ccsm3
0	0	0.0	0.833333	9/1/14	237.00	29.02	31.64	29.5
1	1	0.0	0.833333	9/2/14	228.90	29.02	31.64	29.5
2	2	0.0	0.833333	9/3/14	220.69	29.02	31.64	29.5
3	3	0.0	0.833333	9/4/14	225.28	29.02	31.64	29.5
4	4	0.0	0.833333	9/5/14	237.24	29.02	31.64	29.5
5	5	0.0	0.833333	9/6/14	237.87	29.02	31.64	29.5
6	6	0.0	0.833333	9/7/14	236.36	29.02	31.64	29.5
7	7	0.0	0.833333	9/8/14	233.36	29.02	31.64	29.5
8	8	0.0	0.833333	9/9/14	233.82	29.02	31.64	29.5
9	9	0.0	0.833333	9/10/14	229.74	29.02	31.64	29.5
10	10	0.0	0.833333	9/11/14	220.59	29.02	31.64	29.5
11	11	0.0	0.833333	9/12/14	208.32	29.02	31.64	29.5
12	12	0.0	0.833333	9/13/14	198.76	29.02	31.64	29.5
13	13	0.0	0.833333	9/14/14	196.75	29.02	31.64	29.5
14	14	0.0	0.833333	9/15/14	195.16	29.02	31.64	29.5
15	15	0.0	0.833333	9/16/14	195.87	29.02	31.64	29.5
16	16	0.0	0.833333	9/17/14	197.96	29.02	31.64	29.5
17	17	0.0	0.833333	9/18/14	201.64	29.02	31.64	29.5
18	18	0.0	0.833333	9/19/14	201.59	29.02	31.64	29.5
19	19	0.0	0.833333	9/20/14	204.63	29.02	31.64	29.5
20	20	0.0	0.833333	9/21/14	216.39	29.02	31.64	29.5
21	21	0.0	0.833333	9/22/14	228.88	29.02	31.64	29.5
22	22	0.0	0.833333	9/23/14	230.29	26.87	27.15	28.
23	23	0.0	0.833333	9/24/14	232.52	26.87	27.15	28.
24	24	0.0	0.833333	9/25/14	238.10	26.87	27.15	28.
25	25	0.0	0.833333	9/26/14	246.83	26.87	27.15	28.
26	26	0.0	0.833333	9/27/14	265.05	26.87	27.15	28.
27	27	0.0	0.833333	9/28/14	272.41	26.87	27.15	28.
28	28	0.0	0.833333	9/29/14	271.33	26.87	27.15	28.
29	29	0.0	0.833333	9/30/14	283.48	26.87	27.15	28.
30	30	0.0	0.833333	10/1/14	295.56	26.87	27.15	28.
31	31	0.0	0.833333	10/2/14	296.27	26.87	27.15	28.
32	32	0.0	0.833333	10/3/14	296.27	26.87	27.15	28.

	index	lat	lon	startdate	contest-pevpr-sfc-gauss-14d__pevpr	nmme0-tmp2m-34w__cancm30	nmme0-tmp2m-34w__cancm40	nmme tmp2r 34w__ccsm3
33	33	0.0	0.833333	10/4/14	302.29	26.87	27.15	28.
34	34	0.0	0.833333	10/5/14	297.77	26.87	27.15	28.
35	35	0.0	0.833333	10/6/14	284.35	26.87	27.15	28.
36	36	0.0	0.833333	10/7/14	274.58	26.87	27.15	28.
37	37	0.0	0.833333	10/8/14	263.68	26.87	27.15	28.
38	38	0.0	0.833333	10/9/14	251.84	26.87	27.15	28.
39	39	0.0	0.833333	10/10/14	239.36	26.87	27.15	28.
40	40	0.0	0.833333	10/11/14	226.89	26.87	27.15	28.
41	41	0.0	0.833333	10/12/14	228.63	26.87	27.15	28.
42	42	0.0	0.833333	10/13/14	237.68	26.87	27.15	28.
43	43	0.0	0.833333	10/14/14	233.77	26.87	27.15	28.
44	44	0.0	0.833333	10/15/14	225.43	26.87	27.15	28.
45	45	0.0	0.833333	10/16/14	217.21	26.87	27.15	28.
46	46	0.0	0.833333	10/17/14	208.81	26.87	27.15	28.
47	47	0.0	0.833333	10/18/14	199.79	26.87	27.15	28.
48	48	0.0	0.833333	10/19/14	203.79	26.87	27.15	28.
49	49	0.0	0.833333	10/20/14	216.33	26.87	27.15	28.
50	50	0.0	0.833333	10/21/14	233.12	26.87	27.15	28.
51	51	0.0	0.833333	10/22/14	244.63	26.87	27.15	28.
52	52	0.0	0.833333	10/23/14	243.42	22.82	26.38	23.9
53	53	0.0	0.833333	10/24/14	235.14	22.82	26.38	23.9
54	54	0.0	0.833333	10/25/14	226.37	22.82	26.38	23.9
55	55	0.0	0.833333	10/26/14	220.72	22.82	26.38	23.9
56	56	0.0	0.833333	10/27/14	215.35	22.82	26.38	23.9
57	57	0.0	0.833333	10/28/14	207.04	22.82	26.38	23.9
58	58	0.0	0.833333	10/29/14	201.30	22.82	26.38	23.9
59	59	0.0	0.833333	10/30/14	207.50	22.82	26.38	23.9
60	60	0.0	0.833333	10/31/14	215.13	22.82	26.38	23.9
61	61	0.0	0.833333	11/1/14	221.10	22.82	26.38	23.9
62	62	0.0	0.833333	11/2/14	213.28	22.82	26.38	23.9
63	63	0.0	0.833333	11/3/14	203.98	22.82	26.38	23.9
64	64	0.0	0.833333	11/4/14	205.92	22.82	26.38	23.9
65	65	0.0	0.833333	11/5/14	201.39	22.82	26.38	23.9

	index	lat	lon	startdate	contest-pevpr-sfc-gauss-14d__pevpr	nmme0-tmp2m-34w__cancm30	nmme0-tmp2m-34w__cancm40	nmme tmp2r 34w__ccsm3
66	66	0.0	0.833333	11/6/14	200.26	22.82	26.38	23.9
67	67	0.0	0.833333	11/7/14	203.77	22.82	26.38	23.9
68	68	0.0	0.833333	11/8/14	203.79	22.82	26.38	23.9
69	69	0.0	0.833333	11/9/14	200.77	22.82	26.38	23.9
70	70	0.0	0.833333	11/10/14	202.40	22.82	26.38	23.9
71	71	0.0	0.833333	11/11/14	206.77	22.82	26.38	23.9
72	72	0.0	0.833333	11/12/14	206.18	22.82	26.38	23.9
73	73	0.0	0.833333	11/13/14	198.71	22.82	26.38	23.9
74	74	0.0	0.833333	11/14/14	189.41	22.82	26.38	23.9
75	75	0.0	0.833333	11/15/14	183.12	22.82	26.38	23.9
76	76	0.0	0.833333	11/16/14	186.39	22.82	26.38	23.9
77	77	0.0	0.833333	11/17/14	187.13	22.82	26.38	23.9
78	78	0.0	0.833333	11/18/14	171.31	22.82	26.38	23.9
79	79	0.0	0.833333	11/19/14	164.24	22.82	26.38	23.9
80	80	0.0	0.833333	11/20/14	162.15	22.82	26.38	23.9
81	81	0.0	0.833333	11/21/14	162.79	22.82	26.38	23.9
82	82	0.0	0.833333	11/22/14	166.86	22.82	26.38	23.9
83	83	0.0	0.833333	11/23/14	168.24	16.91	21.65	13.1
84	84	0.0	0.833333	11/24/14	162.08	16.91	21.65	13.1
85	85	0.0	0.833333	11/25/14	148.95	16.91	21.65	13.1
86	86	0.0	0.833333	11/26/14	141.57	16.91	21.65	13.1
87	87	0.0	0.833333	11/27/14	137.39	16.91	21.65	13.1
88	88	0.0	0.833333	11/28/14	133.33	16.91	21.65	13.1
89	89	0.0	0.833333	11/29/14	130.50	16.91	21.65	13.1
90	90	0.0	0.833333	11/30/14	125.57	16.91	21.65	13.1
91	91	0.0	0.833333	12/1/14	123.22	16.91	21.65	13.1
92	92	0.0	0.833333	12/2/14	121.44	16.91	21.65	13.1
93	93	0.0	0.833333	12/3/14	120.95	16.91	21.65	13.1
94	94	0.0	0.833333	12/4/14	121.83	16.91	21.65	13.1
95	95	0.0	0.833333	12/5/14	117.54	16.91	21.65	13.1
96	96	0.0	0.833333	12/6/14	111.51	16.91	21.65	13.1
97	97	0.0	0.833333	12/7/14	105.98	16.91	21.65	13.1
98	98	0.0	0.833333	12/8/14	101.69	16.91	21.65	13.1

	index	lat	lon	startdate	contest-pevpr-sfc-gauss-14d__pevpr	nmme0-tmp2m-34w__cancm30	nmme0-tmp2m-34w__cancm40	nmme tmp2r 34w__ccsm3
99	99	0.0	0.833333	12/9/14	104.63	16.91	21.65	13.

In [13]:
```
training_data.tail(100)
```

Out[13]:

	index	lat	lon	startdate	contest-pevpr-sfc-gauss-14d__pevpr	nmme0-tmp2m-34w__cancm30	nmme0-tmp2m-34w__cancm40	34w_
375634	375634	1.0	0.866667	5/24/16	143.28	13.64	18.50	
375635	375635	1.0	0.866667	5/25/16	144.91	13.64	18.50	
375636	375636	1.0	0.866667	5/26/16	154.90	13.64	18.50	
375637	375637	1.0	0.866667	5/27/16	172.81	13.64	18.50	
375638	375638	1.0	0.866667	5/28/16	191.11	13.64	18.50	
375639	375639	1.0	0.866667	5/29/16	212.07	13.64	18.50	
375640	375640	1.0	0.866667	5/30/16	224.08	13.64	18.50	
375641	375641	1.0	0.866667	5/31/16	235.20	13.64	18.50	
375642	375642	1.0	0.866667	6/1/16	244.87	13.64	18.50	
375643	375643	1.0	0.866667	6/2/16	258.54	13.64	18.50	
375644	375644	1.0	0.866667	6/3/16	273.68	13.64	18.50	
375645	375645	1.0	0.866667	6/4/16	292.77	13.64	18.50	
375646	375646	1.0	0.866667	6/5/16	309.35	13.64	18.50	
375647	375647	1.0	0.866667	6/6/16	323.44	13.64	18.50	
375648	375648	1.0	0.866667	6/7/16	352.26	13.64	18.50	
375649	375649	1.0	0.866667	6/8/16	367.56	13.64	18.50	
375650	375650	1.0	0.866667	6/9/16	369.23	13.64	18.50	
375651	375651	1.0	0.866667	6/10/16	371.74	13.64	18.50	
375652	375652	1.0	0.866667	6/11/16	375.35	13.64	18.50	
375653	375653	1.0	0.866667	6/12/16	371.36	13.64	18.50	
375654	375654	1.0	0.866667	6/13/16	380.78	13.64	18.50	
375655	375655	1.0	0.866667	6/14/16	377.51	13.64	18.50	
375656	375656	1.0	0.866667	6/15/16	382.44	13.64	18.50	
375657	375657	1.0	0.866667	6/16/16	384.24	13.64	18.50	
375658	375658	1.0	0.866667	6/17/16	377.20	13.64	18.50	
375659	375659	1.0	0.866667	6/18/16	371.78	13.64	18.50	
375660	375660	1.0	0.866667	6/19/16	372.32	13.64	18.50	
375661	375661	1.0	0.866667	6/20/16	369.31	13.64	18.50	
375662	375662	1.0	0.866667	6/21/16	354.84	13.64	18.50	
375663	375663	1.0	0.866667	6/22/16	344.44	13.64	18.50	
375664	375664	1.0	0.866667	6/23/16	344.21	19.84	23.72	
375665	375665	1.0	0.866667	6/24/16	332.64	19.84	23.72	
375666	375666	1.0	0.866667	6/25/16	317.68	19.84	23.72	

	index	lat	lon	startdate	contest-pevpr-sfc-gauss-14d__pevpr	nmme0-tmp2m-34w__cancm30	nmme0-tmp2m-34w__cancm40	34w_
375667	375667	1.0	0.866667	6/26/16	313.54	19.84	23.72	
375668	375668	1.0	0.866667	6/27/16	296.65	19.84	23.72	
375669	375669	1.0	0.866667	6/28/16	287.36	19.84	23.72	
375670	375670	1.0	0.866667	6/29/16	279.64	19.84	23.72	
375671	375671	1.0	0.866667	6/30/16	272.17	19.84	23.72	
375672	375672	1.0	0.866667	7/1/16	263.06	19.84	23.72	
375673	375673	1.0	0.866667	7/2/16	250.71	19.84	23.72	
375674	375674	1.0	0.866667	7/3/16	237.00	19.84	23.72	
375675	375675	1.0	0.866667	7/4/16	223.86	19.84	23.72	
375676	375676	1.0	0.866667	7/5/16	212.37	19.84	23.72	
375677	375677	1.0	0.866667	7/6/16	209.50	19.84	23.72	
375678	375678	1.0	0.866667	7/7/16	207.04	19.84	23.72	
375679	375679	1.0	0.866667	7/8/16	217.68	19.84	23.72	
375680	375680	1.0	0.866667	7/9/16	236.14	19.84	23.72	
375681	375681	1.0	0.866667	7/10/16	236.92	19.84	23.72	
375682	375682	1.0	0.866667	7/11/16	243.39	19.84	23.72	
375683	375683	1.0	0.866667	7/12/16	256.87	19.84	23.72	
375684	375684	1.0	0.866667	7/13/16	266.98	19.84	23.72	
375685	375685	1.0	0.866667	7/14/16	282.81	19.84	23.72	
375686	375686	1.0	0.866667	7/15/16	295.20	19.84	23.72	
375687	375687	1.0	0.866667	7/16/16	306.61	19.84	23.72	
375688	375688	1.0	0.866667	7/17/16	318.07	19.84	23.72	
375689	375689	1.0	0.866667	7/18/16	336.61	19.84	23.72	
375690	375690	1.0	0.866667	7/19/16	336.94	19.84	23.72	
375691	375691	1.0	0.866667	7/20/16	337.30	19.84	23.72	
375692	375692	1.0	0.866667	7/21/16	346.95	19.84	23.72	
375693	375693	1.0	0.866667	7/22/16	355.50	19.84	23.72	
375694	375694	1.0	0.866667	7/23/16	346.58	24.85	29.11	
375695	375695	1.0	0.866667	7/24/16	347.12	24.85	29.11	
375696	375696	1.0	0.866667	7/25/16	351.16	24.85	29.11	
375697	375697	1.0	0.866667	7/26/16	365.32	24.85	29.11	
375698	375698	1.0	0.866667	7/27/16	376.73	24.85	29.11	
375699	375699	1.0	0.866667	7/28/16	366.72	24.85	29.11	

	index	lat	lon	startdate	contest-pevpr-sfc-gauss-14d__pevpr	nmme0-tmp2m-34w__cancm30	nmme0-tmp2m-34w__cancm40	34w_
375700	375700	1.0	0.866667	7/29/16	352.33	24.85	29.11	
375701	375701	1.0	0.866667	7/30/16	340.12	24.85	29.11	
375702	375702	1.0	0.866667	7/31/16	324.19	24.85	29.11	
375703	375703	1.0	0.866667	8/1/16	305.55	24.85	29.11	
375704	375704	1.0	0.866667	8/2/16	303.35	24.85	29.11	
375705	375705	1.0	0.866667	8/3/16	298.66	24.85	29.11	
375706	375706	1.0	0.866667	8/4/16	289.68	24.85	29.11	
375707	375707	1.0	0.866667	8/5/16	273.48	24.85	29.11	
375708	375708	1.0	0.866667	8/6/16	273.93	24.85	29.11	
375709	375709	1.0	0.866667	8/7/16	269.26	24.85	29.11	
375710	375710	1.0	0.866667	8/8/16	264.08	24.85	29.11	
375711	375711	1.0	0.866667	8/9/16	267.95	24.85	29.11	
375712	375712	1.0	0.866667	8/10/16	263.76	24.85	29.11	
375713	375713	1.0	0.866667	8/11/16	277.68	24.85	29.11	
375714	375714	1.0	0.866667	8/12/16	281.80	24.85	29.11	
375715	375715	1.0	0.866667	8/13/16	292.04	24.85	29.11	
375716	375716	1.0	0.866667	8/14/16	305.06	24.85	29.11	
375717	375717	1.0	0.866667	8/15/16	311.27	24.85	29.11	
375718	375718	1.0	0.866667	8/16/16	318.92	24.85	29.11	
375719	375719	1.0	0.866667	8/17/16	327.84	24.85	29.11	
375720	375720	1.0	0.866667	8/18/16	328.07	24.85	29.11	
375721	375721	1.0	0.866667	8/19/16	338.59	24.85	29.11	
375722	375722	1.0	0.866667	8/20/16	360.08	24.85	29.11	
375723	375723	1.0	0.866667	8/21/16	381.64	24.85	29.11	
375724	375724	1.0	0.866667	8/22/16	375.94	24.85	29.11	
375725	375725	1.0	0.866667	8/23/16	350.53	23.13	27.20	
375726	375726	1.0	0.866667	8/24/16	332.11	23.13	27.20	
375727	375727	1.0	0.866667	8/25/16	310.16	23.13	27.20	
375728	375728	1.0	0.866667	8/26/16	314.68	23.13	27.20	
375729	375729	1.0	0.866667	8/27/16	312.05	23.13	27.20	
375730	375730	1.0	0.866667	8/28/16	305.82	23.13	27.20	
375731	375731	1.0	0.866667	8/29/16	311.62	23.13	27.20	
375732	375732	1.0	0.866667	8/30/16	304.54	23.13	27.20	

	index	lat	lon	startdate	contest-pevpr-sfc-gauss-14d__pevpr	nmme0-tmp2m-34w__cancm30	nmme0-tmp2m-34w__cancm40	34w_
375733	375733	1.0	0.866667	8/31/16	295.29	23.13	27.20	

First 100 rows are metrics starting for 9/1/14 collected each day at the same location. lat 0.0 and lon 0.83333. We will need to rename columns or really get to know data dictionary.

In [14]:
```
with pd.option_context('display.max_rows', None,
'display.max_columns', None):
    print(training_data.dtypes)
```

```
index                                    int64
lat                                    float64
lon                                    float64
startdate                               object
contest-pevpr-sfc-gauss-14d__pevpr     float64
nmme0-tmp2m-34w__cancm30               float64
nmme0-tmp2m-34w__cancm40               float64
nmme0-tmp2m-34w__ccsm30                float64
nmme0-tmp2m-34w__ccsm40                float64
nmme0-tmp2m-34w__cfsv20                float64
nmme0-tmp2m-34w__gfdlflora0            float64
nmme0-tmp2m-34w__gfdlflorb0            float64
nmme0-tmp2m-34w__gfdl0                 float64
nmme0-tmp2m-34w__nasa0                 float64
nmme0-tmp2m-34w__nmme0mean             float64
contest-wind-h10-14d__wind-hgt-10      float64
nmme-tmp2m-56w__cancm3                 float64
nmme-tmp2m-56w__cancm4                 float64
nmme-tmp2m-56w__ccsm3                  float64
nmme-tmp2m-56w__ccsm4                  float64
nmme-tmp2m-56w__cfsv2                  float64
nmme-tmp2m-56w__gfdl                   float64
nmme-tmp2m-56w__gfdlflora              float64
nmme-tmp2m-56w__gfdlflorb              float64
nmme-tmp2m-56w__nasa                   float64
nmme-tmp2m-56w__nmmemean               float64
contest-rhum-sig995-14d__rhum          float64
nmme-prate-34w__cancm3                 float64
nmme-prate-34w__cancm4                 float64
nmme-prate-34w__ccsm3                  float64
nmme-prate-34w__ccsm4                  float64
nmme-prate-34w__cfsv2                  float64
nmme-prate-34w__gfdl                   float64
nmme-prate-34w__gfdlflora              float64
nmme-prate-34w__gfdlflorb              float64
nmme-prate-34w__nasa                   float64
nmme-prate-34w__nmmemean               float64
contest-wind-h100-14d__wind-hgt-100    float64
nmme0-prate-56w__cancm30               float64
nmme0-prate-56w__cancm40               float64
nmme0-prate-56w__ccsm30                float64
nmme0-prate-56w__ccsm40                float64
nmme0-prate-56w__cfsv20                float64
nmme0-prate-56w__gfdlflora0            float64
nmme0-prate-56w__gfdlflorb0            float64
nmme0-prate-56w__gfdl0                 float64
nmme0-prate-56w__nasa0                 float64
nmme0-prate-56w__nmme0mean             float64
nmme0-prate-34w__cancm30               float64
nmme0-prate-34w__cancm40               float64
nmme0-prate-34w__ccsm30                float64
nmme0-prate-34w__ccsm40                float64
```

```
nmme0-prate-34w__cfsv20                      float64
nmme0-prate-34w__gfdlflora0                  float64
nmme0-prate-34w__gfdlflorb0                  float64
nmme0-prate-34w__gfdl0                       float64
nmme0-prate-34w__nasa0                       float64
nmme0-prate-34w__nmme0mean                   float64
contest-tmp2m-14d__tmp2m                     float64
contest-slp-14d__slp                         float64
contest-wind-vwnd-925-14d__wind-vwnd-925     float64
nmme-prate-56w__cancm3                       float64
nmme-prate-56w__cancm4                       float64
nmme-prate-56w__ccsm3                        float64
nmme-prate-56w__ccsm4                        float64
nmme-prate-56w__cfsv2                        float64
nmme-prate-56w__gfdl                         float64
nmme-prate-56w__gfdlflora                    float64
nmme-prate-56w__gfdlflorb                    float64
nmme-prate-56w__nasa                         float64
nmme-prate-56w__nmmemean                     float64
contest-pres-sfc-gauss-14d__pres             float64
contest-wind-uwnd-250-14d__wind-uwnd-250     float64
nmme-tmp2m-34w__cancm3                       float64
nmme-tmp2m-34w__cancm4                       float64
nmme-tmp2m-34w__ccsm3                        float64
nmme-tmp2m-34w__ccsm4                        float64
nmme-tmp2m-34w__cfsv2                        float64
nmme-tmp2m-34w__gfdl                         float64
nmme-tmp2m-34w__gfdlflora                    float64
nmme-tmp2m-34w__gfdlflorb                    float64
nmme-tmp2m-34w__nasa                         float64
nmme-tmp2m-34w__nmmemean                     float64
contest-prwtr-eatm-14d__prwtr                float64
contest-wind-vwnd-250-14d__wind-vwnd-250     float64
contest-precip-14d__precip                   float64
contest-wind-h850-14d__wind-hgt-850          float64
contest-wind-uwnd-925-14d__wind-uwnd-925     float64
contest-wind-h500-14d__wind-hgt-500          float64
cancm30                                      float64
cancm40                                      float64
ccsm30                                       float64
ccsm40                                       float64
cfsv20                                       float64
gfdlflora0                                   float64
gfdlflorb0                                   float64
gfdl0                                        float64
nasa0                                        float64
nmme0mean                                    float64
climateregions__climateregion                 object
elevation__elevation                         float64
wind-vwnd-250-2010-1                         float64
wind-vwnd-250-2010-2                         float64
wind-vwnd-250-2010-3                         float64
```

```
wind-vwnd-250-2010-4                float64
wind-vwnd-250-2010-5                float64
wind-vwnd-250-2010-6                float64
wind-vwnd-250-2010-7                float64
wind-vwnd-250-2010-8                float64
wind-vwnd-250-2010-9                float64
wind-vwnd-250-2010-10               float64
wind-vwnd-250-2010-11               float64
wind-vwnd-250-2010-12               float64
wind-vwnd-250-2010-13               float64
wind-vwnd-250-2010-14               float64
wind-vwnd-250-2010-15               float64
wind-vwnd-250-2010-16               float64
wind-vwnd-250-2010-17               float64
wind-vwnd-250-2010-18               float64
wind-vwnd-250-2010-19               float64
wind-vwnd-250-2010-20               float64
wind-uwnd-250-2010-1                float64
wind-uwnd-250-2010-2                float64
wind-uwnd-250-2010-3                float64
wind-uwnd-250-2010-4                float64
wind-uwnd-250-2010-5                float64
wind-uwnd-250-2010-6                float64
wind-uwnd-250-2010-7                float64
wind-uwnd-250-2010-8                float64
wind-uwnd-250-2010-9                float64
wind-uwnd-250-2010-10               float64
wind-uwnd-250-2010-11               float64
wind-uwnd-250-2010-12               float64
wind-uwnd-250-2010-13               float64
wind-uwnd-250-2010-14               float64
wind-uwnd-250-2010-15               float64
wind-uwnd-250-2010-16               float64
wind-uwnd-250-2010-17               float64
wind-uwnd-250-2010-18               float64
wind-uwnd-250-2010-19               float64
wind-uwnd-250-2010-20               float64
mjo1d__phase                          int64
mjo1d__amplitude                    float64
mei__mei                            float64
mei__meirank                          int64
mei__nip                              int64
wind-hgt-850-2010-1                 float64
wind-hgt-850-2010-2                 float64
wind-hgt-850-2010-3                 float64
wind-hgt-850-2010-4                 float64
wind-hgt-850-2010-5                 float64
wind-hgt-850-2010-6                 float64
wind-hgt-850-2010-7                 float64
wind-hgt-850-2010-8                 float64
wind-hgt-850-2010-9                 float64
wind-hgt-850-2010-10                float64
```

```
sst-2010-1                          float64
sst-2010-2                          float64
sst-2010-3                          float64
sst-2010-4                          float64
sst-2010-5                          float64
sst-2010-6                          float64
sst-2010-7                          float64
sst-2010-8                          float64
sst-2010-9                          float64
sst-2010-10                         float64
wind-hgt-500-2010-1                 float64
wind-hgt-500-2010-2                 float64
wind-hgt-500-2010-3                 float64
wind-hgt-500-2010-4                 float64
wind-hgt-500-2010-5                 float64
wind-hgt-500-2010-6                 float64
wind-hgt-500-2010-7                 float64
wind-hgt-500-2010-8                 float64
wind-hgt-500-2010-9                 float64
wind-hgt-500-2010-10                float64
icec-2010-1                         float64
icec-2010-2                         float64
icec-2010-3                         float64
icec-2010-4                         float64
icec-2010-5                         float64
icec-2010-6                         float64
icec-2010-7                         float64
icec-2010-8                         float64
icec-2010-9                         float64
icec-2010-10                        float64
wind-uwnd-925-2010-1                float64
wind-uwnd-925-2010-2                float64
wind-uwnd-925-2010-3                float64
wind-uwnd-925-2010-4                float64
wind-uwnd-925-2010-5                float64
wind-uwnd-925-2010-6                float64
wind-uwnd-925-2010-7                float64
wind-uwnd-925-2010-8                float64
wind-uwnd-925-2010-9                float64
wind-uwnd-925-2010-10               float64
wind-uwnd-925-2010-11               float64
wind-uwnd-925-2010-12               float64
wind-uwnd-925-2010-13               float64
wind-uwnd-925-2010-14               float64
wind-uwnd-925-2010-15               float64
wind-uwnd-925-2010-16               float64
wind-uwnd-925-2010-17               float64
wind-uwnd-925-2010-18               float64
wind-uwnd-925-2010-19               float64
wind-uwnd-925-2010-20               float64
wind-hgt-10-2010-1                  float64
wind-hgt-10-2010-2                  float64
```

```
wind-hgt-10-2010-3                  float64
wind-hgt-10-2010-4                  float64
wind-hgt-10-2010-5                  float64
wind-hgt-10-2010-6                  float64
wind-hgt-10-2010-7                  float64
wind-hgt-10-2010-8                  float64
wind-hgt-10-2010-9                  float64
wind-hgt-10-2010-10                 float64
wind-hgt-100-2010-1                 float64
wind-hgt-100-2010-2                 float64
wind-hgt-100-2010-3                 float64
wind-hgt-100-2010-4                 float64
wind-hgt-100-2010-5                 float64
wind-hgt-100-2010-6                 float64
wind-hgt-100-2010-7                 float64
wind-hgt-100-2010-8                 float64
wind-hgt-100-2010-9                 float64
wind-hgt-100-2010-10                float64
wind-vwnd-925-2010-1                float64
wind-vwnd-925-2010-2                float64
wind-vwnd-925-2010-3                float64
wind-vwnd-925-2010-4                float64
wind-vwnd-925-2010-5                float64
wind-vwnd-925-2010-6                float64
wind-vwnd-925-2010-7                float64
wind-vwnd-925-2010-8                float64
wind-vwnd-925-2010-9                float64
wind-vwnd-925-2010-10               float64
wind-vwnd-925-2010-11               float64
wind-vwnd-925-2010-12               float64
wind-vwnd-925-2010-13               float64
wind-vwnd-925-2010-14               float64
wind-vwnd-925-2010-15               float64
wind-vwnd-925-2010-16               float64
wind-vwnd-925-2010-17               float64
wind-vwnd-925-2010-18               float64
wind-vwnd-925-2010-19               float64
wind-vwnd-925-2010-20               float64
dtype: object
```

Data Types

- All columms are integers or floats minus the start date and climate region.
- We will convert `startdate` to datatime now. -
- We will also need to make dummy variables of the `climateregion__climateregion`.

In [15]:

```
# Convert to startdate Datetime
training_data['startdate'] =
pd.to_datetime(training_data['startdate'])
```

In [16]:
```
#Check
training_data.dtypes
```

Out[16]:
```
index                                    int64
lat                                    float64
lon                                    float64
startdate                       datetime64[ns]
contest-pevpr-sfc-gauss-14d__pevpr     float64
                                        ...
wind-vwnd-925-2010-16                  float64
wind-vwnd-925-2010-17                  float64
wind-vwnd-925-2010-18                  float64
wind-vwnd-925-2010-19                  float64
wind-vwnd-925-2010-20                  float64
Length: 246, dtype: object
```

In [17]:
```
print(training_data.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 375734 entries, 0 to 375733
Columns: 246 entries, index to wind-vwnd-925-2010-20
dtypes: datetime64[ns](1), float64(240), int64(4), object(1)
memory usage: 705.2+ MB
None
```

Distribution of Dates and Places

In [18]:
```
#First Date
min = training_data['startdate'].min()
print(min)
```

```
2014-09-01 00:00:00
```

In [19]:
```
#Last Date
max = training_data['startdate'].max()
print(max)
```

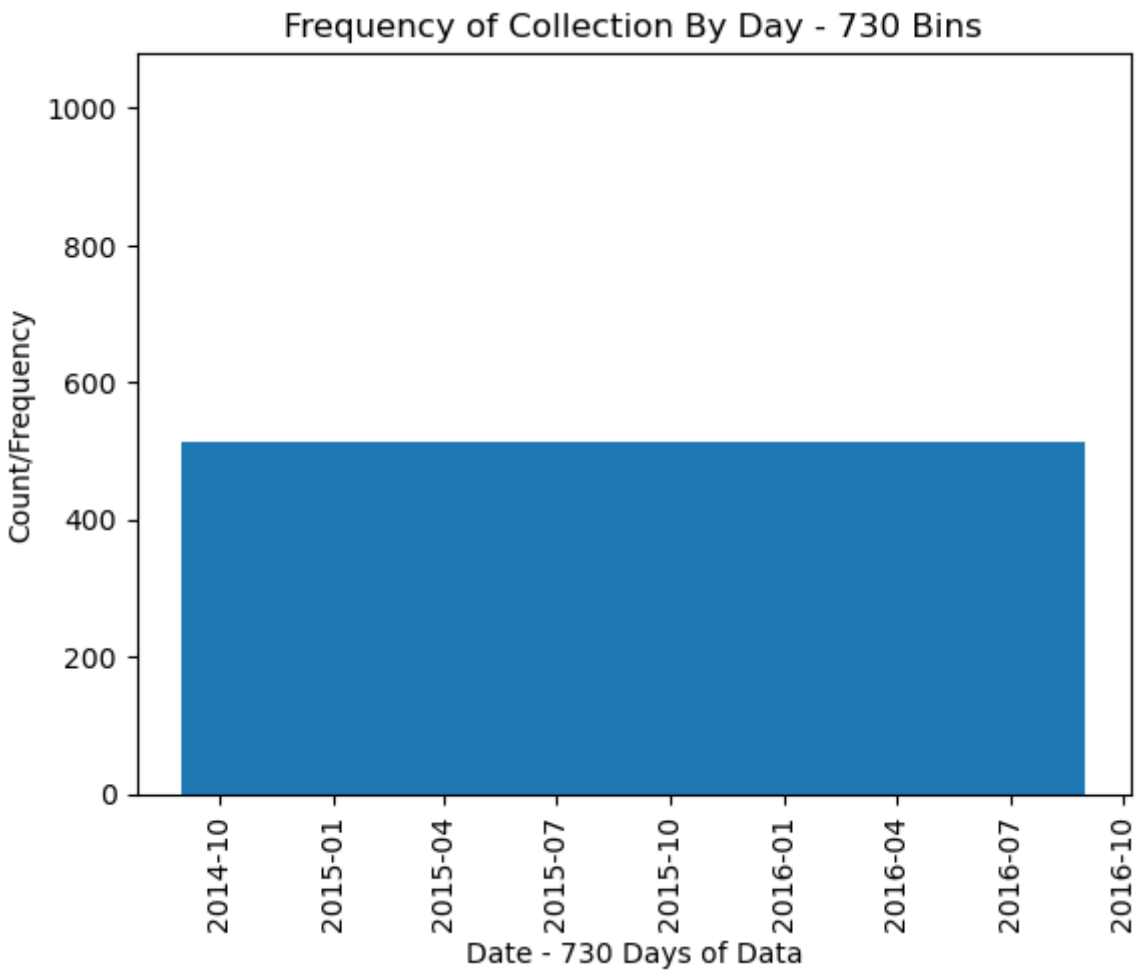
```
2016-08-31 00:00:00
```

In [20]:
```
# Number of days tested - Number of Bins in Histogram
days = max - min
print(days)
```

```
730 days 00:00:00
```

In [21]:
```
# Plot rows of data by start date
plt.figure()
```

```
plt.hist(training_data['startdate'], bins = 730)
plt.title("Frequency of Collection By Day - 730 Bins")
plt.xlabel('Date - 730 Days of Data')
plt.ylabel('Count/Frequency')
plt.xticks(rotation = 90)
plt.show()
```



Approximate - Over 500 data collections per day. Uniform distribution.

In [22]:

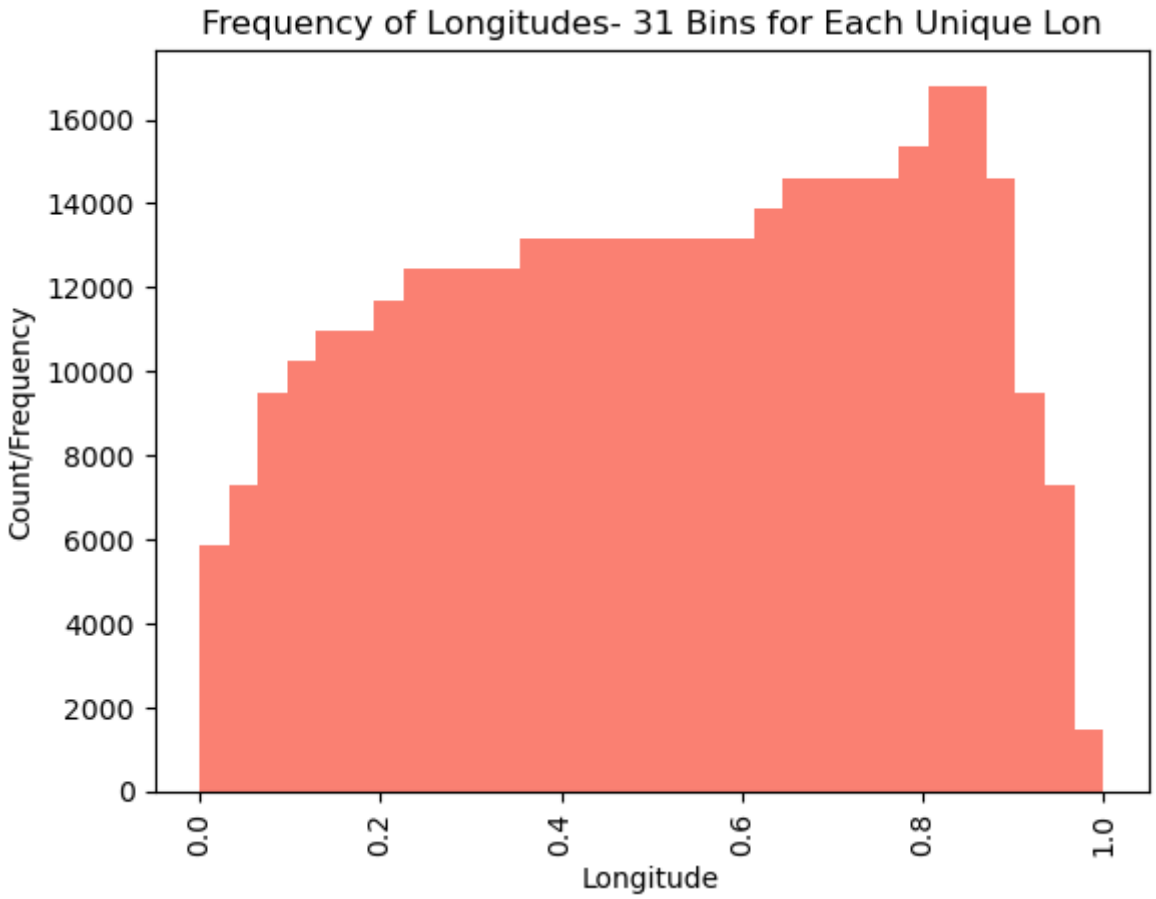
```
# Let's just get the unique lons first
training_data['lon'].value_counts()
```

Out[22]:
```
0.833333    16813
0.866667    16813
0.800000    15351
0.900000    14620
0.666667    14620
0.700000    14620
0.733333    14620
0.766667    14620
0.633333    13889
0.433333    13158
0.600000    13158
0.566667    13158
0.533333    13158
0.500000    13158
0.466667    13158
0.400000    13158
0.366667    13158
0.233333    12427
0.266667    12427
0.300000    12427
0.333333    12427
0.200000    11696
0.133333    10965
0.166667    10965
0.100000    10234
0.933333     9503
0.066667     9503
0.966667     7310
0.033333     7310
0.000000     5848
1.000000     1462
Name: lon, dtype: int64
```

In [23]:
```
# Let's just get the number of unique lons
training_data['lon'].nunique()
```

Out[23]:
```
31
```

In [24]:
```
#Plot Rows of data by longitude
plt.figure()
plt.hist(training_data['lon'], bins = 31, color = 'salmon')
plt.title("Frequency of Longitudes- 31 Bins for Each Unique Lon")
plt.xlabel('Longitude')
plt.ylabel('Count/Frequency')
plt.xticks(rotation = 90)
plt.show()
```



31 unique longitudes not of uniform distribution

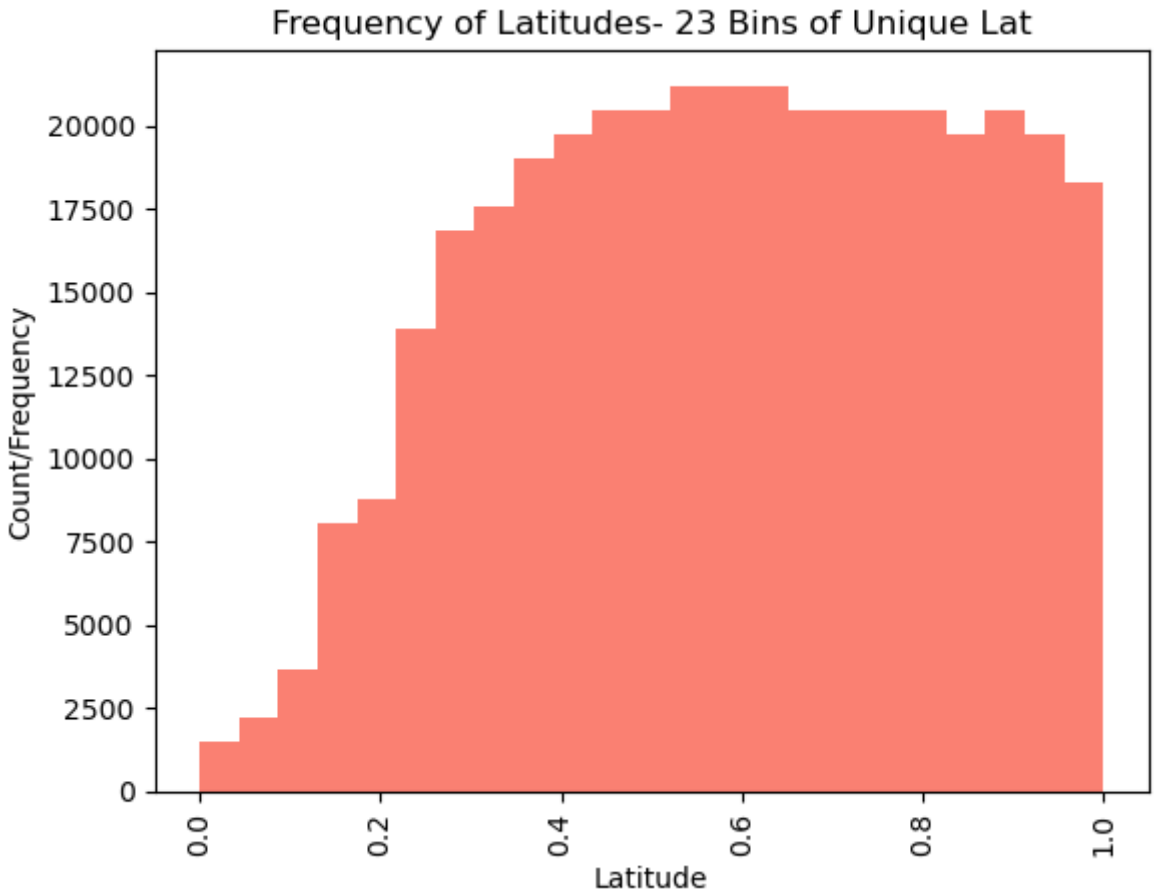
In [25]:
```
# Let's get the unique lats
training_data['lat'].value_counts()
```

Out[25]:
```
0.636364    21199
0.545455    21199
0.590909    21199
0.500000    20468
0.909091    20468
0.818182    20468
0.772727    20468
0.727273    20468
0.454545    20468
0.681818    20468
0.954545    19737
0.863636    19737
0.409091    19737
0.363636    19006
1.000000    18275
0.318182    17544
0.272727    16813
0.227273    13889
0.181818     8772
0.136364     8041
0.090909     3655
0.045455     2193
0.000000     1462
Name: lat, dtype: int64
```

In [26]:
```
# Let's just get the unique lats first
training_data['lat'].nunique()
```

Out[26]:
```
23
```

In [27]:
```
#Plot Rows of data by longitude
plt.figure()
plt.hist(training_data['lat'], bins = 23, color = 'salmon')
plt.title("Frequency of Latitudes- 23 Bins of Unique Lat")
plt.xlabel('Latitude')
plt.ylabel('Count/Frequency')
plt.xticks(rotation = 90)
plt.show()
```



23 unique latitudes of non-uniform distribution.

This doesn't tell us much. Let's map out these data collection sites.

- Isolate unique lat and lon combinations in single dataframe, locations_df, using drop duplicates
- Use Tableau map function map these locations.

In [28]:
```
#Make df of locations
location_df = training_data[['lat','lon']]
```

In [30]:
```
#Check
location_df.head()
```

Out[30]:

	lat	lon
0	0.0	0.833333
1	0.0	0.833333
2	0.0	0.833333
3	0.0	0.833333
4	0.0	0.833333

In [ ]:
```
#Drop Duplicates to get unique locations
location_df = location_df.drop_duplicates()
```

In [37]:
```
#Check
location_df.shape
```

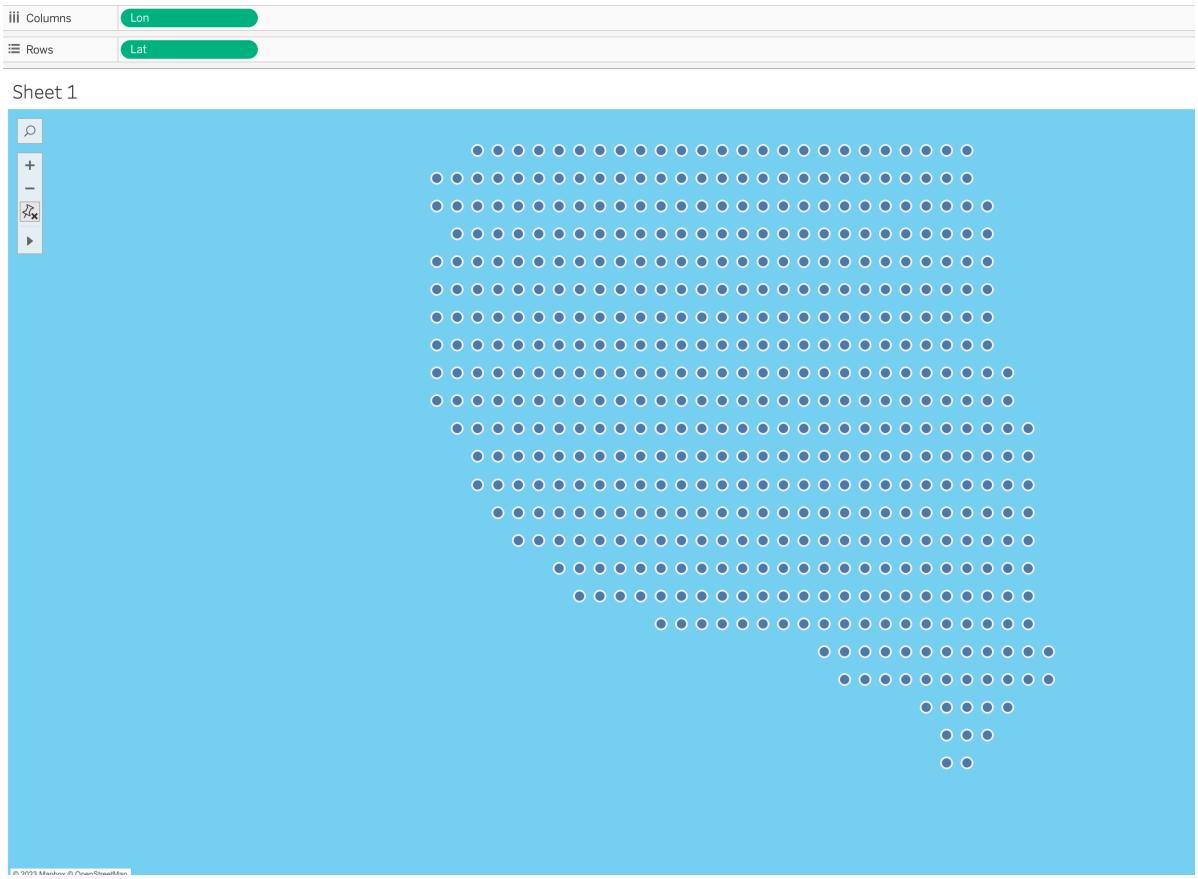
Out[37]:
```
(514, 2)
```

In [38]:
```
#Export location_df to map in Tableau
location_df.to_csv('data/location_df.csv')
```

Data Collection Locations

- We will have to disregard locations or just take into account locations in relation to themselves as we have been given latitudes and longitudes that don't add up.
- The Tableau map of the locations looks like it could be the Western US. We know that the data collection occurred within the US from the data dictionary.
- But all latitudes and longitudes given point to a region of the Atlantic Ocean off the West Coast of Africa. This area has come to be called "Null Island." This is not an actual island but a place where null locations or 0 latitudes and 0 longitudes fall. https://www.atlasobscura.com/articles/null-island-is-one-of-the-most-visited-places-on-earth-too-bad-it-doesnt-exist
- We can look into this more if we really need to figure out were these locations are actually at. The latitudes and longitudes might be able to be scaled to find their actual location, but WiDS gave us these locations by error or on purpose to keep anonymity of location or to create a data processing puzzle.

Tableau Map of Locations



Null Island

Understanding 14 Day Time Window

I think before we move on with more EDA we need to understand more just what this challenge is asking us to do. We are looking for the mean of the min and max temp for each given location 14 days out. This target variable in the training set is labeled `contest-tmp2m-14d__tmp2m` This is between the 40th and 50th column in the training_data.

Let's look at the test_data and the sample_solution to see how our eventul prediction should look.

In [152…
```
#Read in test set
test_data = pd.read_csv('data/test_data.csv')
```

In [153…
```
test_data.shape
```

Out[153]:
```
(31354, 245)
```

Test_data has significantly less weather/data collection rows, and it has one less feature. This one less feature is the target of `contest-tmp2m-14d__tmp2m`.

In [154…
```
#Read in sample solution
sample_solution = pd.read_csv('data/sample_solution.csv')
```

In [155…
```
sample_solution.shape
```

Out[155]:
```
(31354, 2)
```

So, indeed we need to submit a complete column of predicted `contest-tmp2m-14d__tmp2m` for each row or index with in the test_data. %50 of our predictions will be analyzed for a current leaderboard and 50% will then be analyzed for the final leaderboard, but we will not know which indices will be analyzed by both. So, comprehensive validation of our models using the a variety of validation sets accross the training set will be needed to make sure we don't have much variance on our scoring metrics.

Clean training_data and test_data

Let's deal with:

- null values
- non-numerical parameters
- duplicates

Null Values

- We will find out how many null values we have for the train and test set and then begin to deal with filling these in.

In [156…
```
#Sum nulls for training set - 101772 nulls in training
training_data.isnull().sum().sum()
```

Out[156]:
```
0
```

In [157…
```
#Sum null for test set - 0 nulls in test
test_data.isnull().sum().sum()
```

Out[157]:
```
0
```

The training set has a significant number of nulls that we will have to impute. The test data has no nulls.

- Let's see which columns have nulls
- We will have to reconsider our conversion to datetime with start_date if there are null's here.

In [47]:
```
# See what percentage of these nulls make up each column.
# Calculate percentage in each column
null_df = training_data.isnull().sum() / training_data.shape[0] *100
```

In [72]:
```
# Convert to dataframe
null_df=pd.DataFrame(null_df)
```

In [83]:
```
# Check
null_df.head()
```

Out[83]:

	0
index	0.0
lat	0.0
lon	0.0
startdate	0.0
contest-pevpr-sfc-gauss-14d__pevpr	0.0

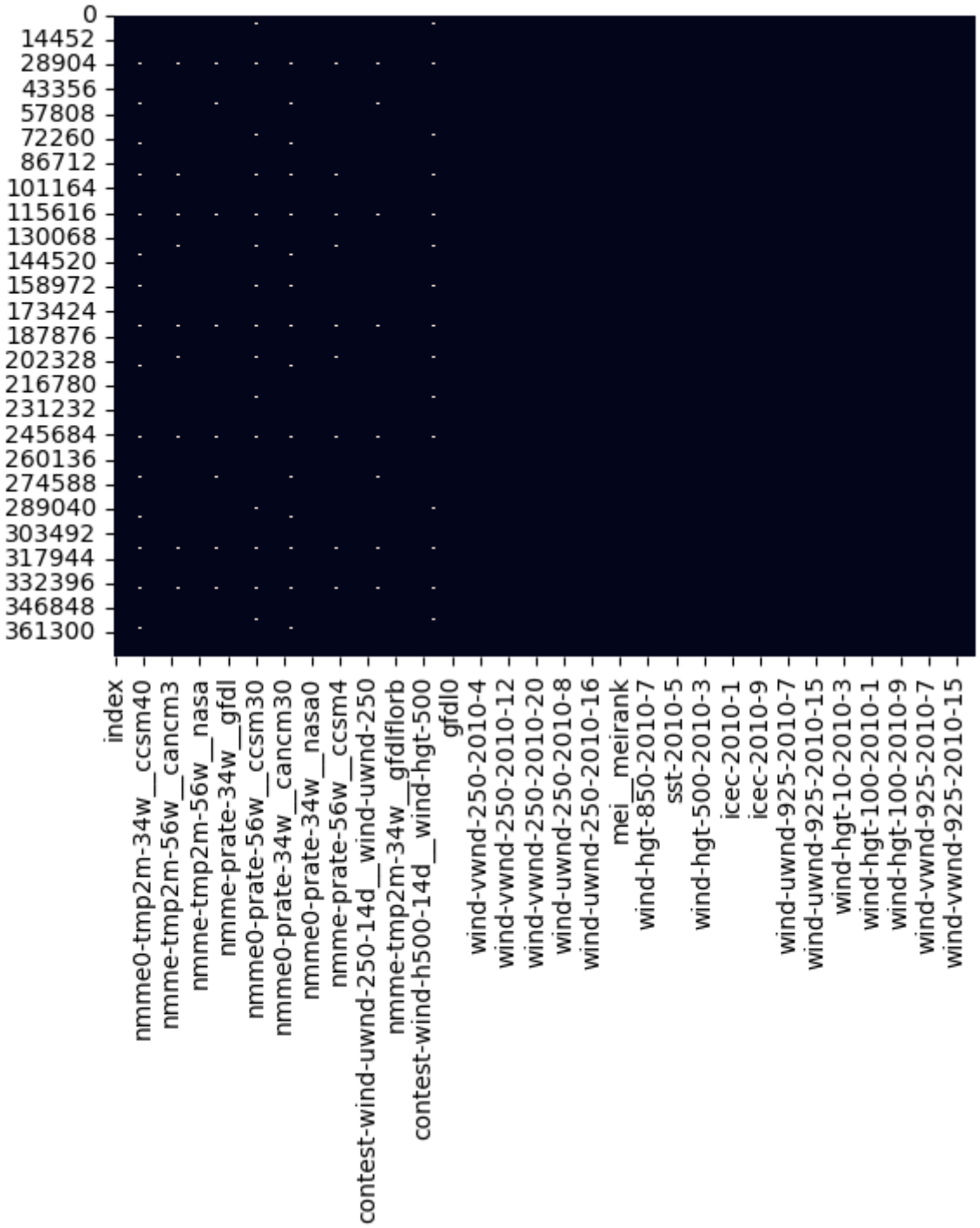
In [82]:
```
# Print columns with null values
for index, row in null_df.iterrows():
    if row.iloc[0] != 0:
        print(index, row)
```

```
nmme0-tmp2m-34w__ccsm30 0    4.240766
Name: nmme0-tmp2m-34w__ccsm30, dtype: float64
nmme-tmp2m-56w__ccsm3 0    2.735978
Name: nmme-tmp2m-56w__ccsm3, dtype: float64
nmme-prate-34w__ccsm3 0    2.325581
Name: nmme-prate-34w__ccsm3, dtype: float64
nmme0-prate-56w__ccsm30 0    4.240766
Name: nmme0-prate-56w__ccsm30, dtype: float64
nmme0-prate-34w__ccsm30 0    4.240766
Name: nmme0-prate-34w__ccsm30, dtype: float64
nmme-prate-56w__ccsm3 0    2.735978
Name: nmme-prate-56w__ccsm3, dtype: float64
nmme-tmp2m-34w__ccsm3 0    2.325581
Name: nmme-tmp2m-34w__ccsm3, dtype: float64
ccsm30 0    4.240766
Name: ccsm30, dtype: float64
```

Eight columns of data with null's ranging from 2.3% to 4.24% null. All null's are numerical columns. Let's see where these are in the data set with a heatmap.

In [85]:

```
plt.figure()
sns.heatmap(training_data.isnull(), cbar=False)
plt.xlabel('Index')
plt.ylabel('Parameters/Columns')
plt.show()
```



Nulls are spread out for each column over 8 - 15 places. So there are not super wide gaps of missing data. The heatmap labels do not match up but we get an idea of how the nulls are dispersed.

- We will fill these null values with the median value for each column. ***We can impute with mean or forward fill or delete null rows later, if we believe this will yield better results**

In [86]:
```
#Fill null values with median values of each column
training_data['nmme0-prate-34w__ccsm30'] = training_data['nmme0-
```

```
prate-34w__ccsm30'].fillna(training_data['nmme0-prate-
34w__ccsm30'].median())
training_data['nmme0-tmp2m-34w__ccsm30'] = training_data['nmme0-
tmp2m-34w__ccsm30'].fillna(training_data['nmme0-tmp2m-
34w__ccsm30'].median())
training_data['ccsm30'] =
training_data['ccsm30'].fillna(training_data['ccsm30'].median())
training_data['nmme0-prate-56w__ccsm30'] = training_data['nmme0-
prate-56w__ccsm30'].fillna(training_data['nmme0-prate-
56w__ccsm30'].median())
training_data['nmme-tmp2m-56w__ccsm3'] = training_data['nmme-tmp2m-
56w__ccsm3'].fillna(training_data['nmme-tmp2m-56w__ccsm3'].median())
training_data['nmme-prate-56w__ccsm3'] = training_data['nmme-prate-
56w__ccsm3'].fillna(training_data['nmme-prate-56w__ccsm3'].median())
training_data['nmme-tmp2m-34w__ccsm3'] = training_data['nmme-tmp2m-
34w__ccsm3'].fillna(training_data['nmme-tmp2m-34w__ccsm3'].median())
training_data['nmme-prate-34w__ccsm3'] = training_data['nmme-prate-
34w__ccsm3'].fillna(training_data['nmme-prate-34w__ccsm3'].median())
```

In [87]:
```
#Check - 0 nulls
# Sum null for training set after medean values imputed
training_data.isnull().sum().sum()
```

Out[87]:
```
0
```

Dummy Variables for Climage Regions for Test/Train and Convert start_date in test_data to Datetime

- We will create dummy variables of each climate region for test and train data.
- We will also set the start date for the test set to date time.
- Both the train and test will then have all numerical data with no nulls.

In [158…
```
# Check data types of test_data
with pd.option_context('display.max_rows', None,
'display.max_columns', None):
    print(test_data.dtypes)
```

```
index                                    int64
lat                                    float64
lon                                    float64
startdate                               object
contest-pevpr-sfc-gauss-14d__pevpr     float64
nmme0-tmp2m-34w__cancm30               float64
nmme0-tmp2m-34w__cancm40               float64
nmme0-tmp2m-34w__ccsm30                float64
nmme0-tmp2m-34w__ccsm40                float64
nmme0-tmp2m-34w__cfsv20                float64
nmme0-tmp2m-34w__gfdlflora0            float64
nmme0-tmp2m-34w__gfdlflorb0            float64
nmme0-tmp2m-34w__gfdl0                 float64
nmme0-tmp2m-34w__nasa0                 float64
nmme0-tmp2m-34w__nmme0mean             float64
contest-wind-h10-14d__wind-hgt-10      float64
nmme-tmp2m-56w__cancm3                 float64
nmme-tmp2m-56w__cancm4                 float64
nmme-tmp2m-56w__ccsm3                  float64
nmme-tmp2m-56w__ccsm4                  float64
nmme-tmp2m-56w__cfsv2                  float64
nmme-tmp2m-56w__gfdl                   float64
nmme-tmp2m-56w__gfdlflora              float64
nmme-tmp2m-56w__gfdlflorb              float64
nmme-tmp2m-56w__nasa                   float64
nmme-tmp2m-56w__nmmemean               float64
contest-rhum-sig995-14d__rhum          float64
nmme-prate-34w__cancm3                 float64
nmme-prate-34w__cancm4                 float64
nmme-prate-34w__ccsm3                  float64
nmme-prate-34w__ccsm4                  float64
nmme-prate-34w__cfsv2                  float64
nmme-prate-34w__gfdl                   float64
nmme-prate-34w__gfdlflora              float64
nmme-prate-34w__gfdlflorb              float64
nmme-prate-34w__nasa                   float64
nmme-prate-34w__nmmemean               float64
contest-wind-h100-14d__wind-hgt-100    float64
nmme0-prate-56w__cancm30               float64
nmme0-prate-56w__cancm40               float64
nmme0-prate-56w__ccsm30                float64
nmme0-prate-56w__ccsm40                float64
nmme0-prate-56w__cfsv20                float64
nmme0-prate-56w__gfdlflora0            float64
nmme0-prate-56w__gfdlflorb0            float64
nmme0-prate-56w__gfdl0                 float64
nmme0-prate-56w__nasa0                 float64
nmme0-prate-56w__nmme0mean             float64
nmme0-prate-34w__cancm30               float64
nmme0-prate-34w__cancm40               float64
nmme0-prate-34w__ccsm30                float64
nmme0-prate-34w__ccsm40                float64
```

```
nmme0-prate-34w__cfsv20                     float64
nmme0-prate-34w__gfdlflora0                 float64
nmme0-prate-34w__gfdlflorb0                 float64
nmme0-prate-34w__gfdl0                      float64
nmme0-prate-34w__nasa0                      float64
nmme0-prate-34w__nmme0mean                  float64
contest-slp-14d__slp                        float64
contest-wind-vwnd-925-14d__wind-vwnd-925    float64
nmme-prate-56w__cancm3                      float64
nmme-prate-56w__cancm4                      float64
nmme-prate-56w__ccsm3                       float64
nmme-prate-56w__ccsm4                       float64
nmme-prate-56w__cfsv2                       float64
nmme-prate-56w__gfdl                        float64
nmme-prate-56w__gfdlflora                   float64
nmme-prate-56w__gfdlflorb                   float64
nmme-prate-56w__nasa                        float64
nmme-prate-56w__nmmemean                    float64
contest-pres-sfc-gauss-14d__pres            float64
contest-wind-uwnd-250-14d__wind-uwnd-250    float64
nmme-tmp2m-34w__cancm3                      float64
nmme-tmp2m-34w__cancm4                      float64
nmme-tmp2m-34w__ccsm3                       float64
nmme-tmp2m-34w__ccsm4                       float64
nmme-tmp2m-34w__cfsv2                       float64
nmme-tmp2m-34w__gfdl                        float64
nmme-tmp2m-34w__gfdlflora                   float64
nmme-tmp2m-34w__gfdlflorb                   float64
nmme-tmp2m-34w__nasa                        float64
nmme-tmp2m-34w__nmmemean                    float64
contest-prwtr-eatm-14d__prwtr               float64
contest-wind-vwnd-250-14d__wind-vwnd-250    float64
contest-precip-14d__precip                  float64
contest-wind-h850-14d__wind-hgt-850         float64
contest-wind-uwnd-925-14d__wind-uwnd-925    float64
contest-wind-h500-14d__wind-hgt-500         float64
cancm30                                     float64
cancm40                                     float64
ccsm30                                      float64
ccsm40                                      float64
cfsv20                                      float64
gfdlflora0                                  float64
gfdlflorb0                                  float64
gfdl0                                       float64
nasa0                                       float64
nmme0mean                                   float64
climateregions__climateregion                object
elevation__elevation                          int64
wind-vwnd-250-2010-1                        float64
wind-vwnd-250-2010-2                        float64
wind-vwnd-250-2010-3                        float64
wind-vwnd-250-2010-4                        float64
```

```
wind-vwnd-250-2010-5                    float64
wind-vwnd-250-2010-6                    float64
wind-vwnd-250-2010-7                    float64
wind-vwnd-250-2010-8                    float64
wind-vwnd-250-2010-9                    float64
wind-vwnd-250-2010-10                   float64
wind-vwnd-250-2010-11                   float64
wind-vwnd-250-2010-12                   float64
wind-vwnd-250-2010-13                   float64
wind-vwnd-250-2010-14                   float64
wind-vwnd-250-2010-15                   float64
wind-vwnd-250-2010-16                   float64
wind-vwnd-250-2010-17                   float64
wind-vwnd-250-2010-18                   float64
wind-vwnd-250-2010-19                   float64
wind-vwnd-250-2010-20                   float64
wind-uwnd-250-2010-1                    float64
wind-uwnd-250-2010-2                    float64
wind-uwnd-250-2010-3                    float64
wind-uwnd-250-2010-4                    float64
wind-uwnd-250-2010-5                    float64
wind-uwnd-250-2010-6                    float64
wind-uwnd-250-2010-7                    float64
wind-uwnd-250-2010-8                    float64
wind-uwnd-250-2010-9                    float64
wind-uwnd-250-2010-10                   float64
wind-uwnd-250-2010-11                   float64
wind-uwnd-250-2010-12                   float64
wind-uwnd-250-2010-13                   float64
wind-uwnd-250-2010-14                   float64
wind-uwnd-250-2010-15                   float64
wind-uwnd-250-2010-16                   float64
wind-uwnd-250-2010-17                   float64
wind-uwnd-250-2010-18                   float64
wind-uwnd-250-2010-19                   float64
wind-uwnd-250-2010-20                   float64
mjo1d__phase                            float64
mjo1d__amplitude                        float64
mei__mei                                float64
mei__meirank                            float64
mei__nip                                float64
wind-hgt-850-2010-1                     float64
wind-hgt-850-2010-2                     float64
wind-hgt-850-2010-3                     float64
wind-hgt-850-2010-4                     float64
wind-hgt-850-2010-5                     float64
wind-hgt-850-2010-6                     float64
wind-hgt-850-2010-7                     float64
wind-hgt-850-2010-8                     float64
wind-hgt-850-2010-9                     float64
wind-hgt-850-2010-10                    float64
sst-2010-1                              float64
```

```
sst-2010-2                          float64
sst-2010-3                          float64
sst-2010-4                          float64
sst-2010-5                          float64
sst-2010-6                          float64
sst-2010-7                          float64
sst-2010-8                          float64
sst-2010-9                          float64
sst-2010-10                         float64
wind-hgt-500-2010-1                 float64
wind-hgt-500-2010-2                 float64
wind-hgt-500-2010-3                 float64
wind-hgt-500-2010-4                 float64
wind-hgt-500-2010-5                 float64
wind-hgt-500-2010-6                 float64
wind-hgt-500-2010-7                 float64
wind-hgt-500-2010-8                 float64
wind-hgt-500-2010-9                 float64
wind-hgt-500-2010-10                float64
icec-2010-1                         float64
icec-2010-2                         float64
icec-2010-3                         float64
icec-2010-4                         float64
icec-2010-5                         float64
icec-2010-6                         float64
icec-2010-7                         float64
icec-2010-8                         float64
icec-2010-9                         float64
icec-2010-10                        float64
wind-uwnd-925-2010-1                float64
wind-uwnd-925-2010-2                float64
wind-uwnd-925-2010-3                float64
wind-uwnd-925-2010-4                float64
wind-uwnd-925-2010-5                float64
wind-uwnd-925-2010-6                float64
wind-uwnd-925-2010-7                float64
wind-uwnd-925-2010-8                float64
wind-uwnd-925-2010-9                float64
wind-uwnd-925-2010-10               float64
wind-uwnd-925-2010-11               float64
wind-uwnd-925-2010-12               float64
wind-uwnd-925-2010-13               float64
wind-uwnd-925-2010-14               float64
wind-uwnd-925-2010-15               float64
wind-uwnd-925-2010-16               float64
wind-uwnd-925-2010-17               float64
wind-uwnd-925-2010-18               float64
wind-uwnd-925-2010-19               float64
wind-uwnd-925-2010-20               float64
wind-hgt-10-2010-1                  float64
wind-hgt-10-2010-2                  float64
wind-hgt-10-2010-3                  float64
```

```
wind-hgt-10-2010-4                  float64
wind-hgt-10-2010-5                  float64
wind-hgt-10-2010-6                  float64
wind-hgt-10-2010-7                  float64
wind-hgt-10-2010-8                  float64
wind-hgt-10-2010-9                  float64
wind-hgt-10-2010-10                 float64
wind-hgt-100-2010-1                 float64
wind-hgt-100-2010-2                 float64
wind-hgt-100-2010-3                 float64
wind-hgt-100-2010-4                 float64
wind-hgt-100-2010-5                 float64
wind-hgt-100-2010-6                 float64
wind-hgt-100-2010-7                 float64
wind-hgt-100-2010-8                 float64
wind-hgt-100-2010-9                 float64
wind-hgt-100-2010-10                float64
wind-vwnd-925-2010-1                float64
wind-vwnd-925-2010-2                float64
wind-vwnd-925-2010-3                float64
wind-vwnd-925-2010-4                float64
wind-vwnd-925-2010-5                float64
wind-vwnd-925-2010-6                float64
wind-vwnd-925-2010-7                float64
wind-vwnd-925-2010-8                float64
wind-vwnd-925-2010-9                float64
wind-vwnd-925-2010-10               float64
wind-vwnd-925-2010-11               float64
wind-vwnd-925-2010-12               float64
wind-vwnd-925-2010-13               float64
wind-vwnd-925-2010-14               float64
wind-vwnd-925-2010-15               float64
wind-vwnd-925-2010-16               float64
wind-vwnd-925-2010-17               float64
wind-vwnd-925-2010-18               float64
wind-vwnd-925-2010-19               float64
wind-vwnd-925-2010-20               float64
dtype: object
```

In [159…

```
# Convert start date to datetime data type for test data
test_data['startdate'] = pd.to_datetime(test_data['startdate'])
```

In [160…

```
# Check
test_data.dtypes
```

Out[160]:

```
index                                          int64
lat                                          float64
lon                                          float64
startdate                              datetime64[ns]
contest-pevpr-sfc-gauss-14d__pevpr           float64
nmme0-tmp2m-34w__cancm30                     float64
nmme0-tmp2m-34w__cancm40                     float64
nmme0-tmp2m-34w__ccsm30                      float64
nmme0-tmp2m-34w__ccsm40                      float64
nmme0-tmp2m-34w__cfsv20                      float64
nmme0-tmp2m-34w__gfdlflora0                  float64
nmme0-tmp2m-34w__gfdlflorb0                  float64
nmme0-tmp2m-34w__gfdl0                       float64
nmme0-tmp2m-34w__nasa0                       float64
nmme0-tmp2m-34w__nmme0mean                   float64
contest-wind-h10-14d__wind-hgt-10            float64
nmme-tmp2m-56w__cancm3                       float64
nmme-tmp2m-56w__cancm4                       float64
nmme-tmp2m-56w__ccsm3                        float64
nmme-tmp2m-56w__ccsm4                        float64
nmme-tmp2m-56w__cfsv2                        float64
nmme-tmp2m-56w__gfdl                         float64
nmme-tmp2m-56w__gfdlflora                    float64
nmme-tmp2m-56w__gfdlflorb                    float64
nmme-tmp2m-56w__nasa                         float64
nmme-tmp2m-56w__nmmemean                     float64
contest-rhum-sig995-14d__rhum                float64
nmme-prate-34w__cancm3                       float64
nmme-prate-34w__cancm4                       float64
nmme-prate-34w__ccsm3                        float64
nmme-prate-34w__ccsm4                        float64
nmme-prate-34w__cfsv2                        float64
nmme-prate-34w__gfdl                         float64
nmme-prate-34w__gfdlflora                    float64
nmme-prate-34w__gfdlflorb                    float64
nmme-prate-34w__nasa                         float64
nmme-prate-34w__nmmemean                     float64
contest-wind-h100-14d__wind-hgt-100          float64
nmme0-prate-56w__cancm30                     float64
nmme0-prate-56w__cancm40                     float64
nmme0-prate-56w__ccsm30                      float64
nmme0-prate-56w__ccsm40                      float64
nmme0-prate-56w__cfsv20                      float64
nmme0-prate-56w__gfdlflora0                  float64
nmme0-prate-56w__gfdlflorb0                  float64
nmme0-prate-56w__gfdl0                       float64
nmme0-prate-56w__nasa0                       float64
nmme0-prate-56w__nmme0mean                   float64
nmme0-prate-34w__cancm30                     float64
nmme0-prate-34w__cancm40                     float64
nmme0-prate-34w__ccsm30                      float64
nmme0-prate-34w__ccsm40                      float64
```

```
nmme0-prate-34w__cfsv20                          float64
nmme0-prate-34w__gfdlflora0                      float64
nmme0-prate-34w__gfdlflorb0                      float64
nmme0-prate-34w__gfdl0                           float64
nmme0-prate-34w__nasa0                           float64
nmme0-prate-34w__nmme0mean                       float64
contest-slp-14d__slp                             float64
contest-wind-vwnd-925-14d__wind-vwnd-925         float64
nmme-prate-56w__cancm3                           float64
nmme-prate-56w__cancm4                           float64
nmme-prate-56w__ccsm3                            float64
nmme-prate-56w__ccsm4                            float64
nmme-prate-56w__cfsv2                            float64
nmme-prate-56w__gfdl                             float64
nmme-prate-56w__gfdlflora                        float64
nmme-prate-56w__gfdlflorb                        float64
nmme-prate-56w__nasa                             float64
nmme-prate-56w__nmmemean                         float64
contest-pres-sfc-gauss-14d__pres                 float64
contest-wind-uwnd-250-14d__wind-uwnd-250         float64
nmme-tmp2m-34w__cancm3                           float64
nmme-tmp2m-34w__cancm4                           float64
nmme-tmp2m-34w__ccsm3                            float64
nmme-tmp2m-34w__ccsm4                            float64
nmme-tmp2m-34w__cfsv2                            float64
nmme-tmp2m-34w__gfdl                             float64
nmme-tmp2m-34w__gfdlflora                        float64
nmme-tmp2m-34w__gfdlflorb                        float64
nmme-tmp2m-34w__nasa                             float64
nmme-tmp2m-34w__nmmemean                         float64
contest-prwtr-eatm-14d__prwtr                    float64
contest-wind-vwnd-250-14d__wind-vwnd-250         float64
contest-precip-14d__precip                       float64
contest-wind-h850-14d__wind-hgt-850              float64
contest-wind-uwnd-925-14d__wind-uwnd-925         float64
contest-wind-h500-14d__wind-hgt-500              float64
cancm30                                          float64
cancm40                                          float64
ccsm30                                           float64
ccsm40                                           float64
cfsv20                                           float64
gfdlflora0                                       float64
gfdlflorb0                                       float64
gfdl0                                            float64
nasa0                                            float64
nmme0mean                                        float64
climateregions__climateregion                     object
elevation__elevation                               int64
wind-vwnd-250-2010-1                             float64
wind-vwnd-250-2010-2                             float64
wind-vwnd-250-2010-3                             float64
wind-vwnd-250-2010-4                             float64
```

```
wind-vwnd-250-2010-5                    float64
wind-vwnd-250-2010-6                    float64
wind-vwnd-250-2010-7                    float64
wind-vwnd-250-2010-8                    float64
wind-vwnd-250-2010-9                    float64
wind-vwnd-250-2010-10                   float64
wind-vwnd-250-2010-11                   float64
wind-vwnd-250-2010-12                   float64
wind-vwnd-250-2010-13                   float64
wind-vwnd-250-2010-14                   float64
wind-vwnd-250-2010-15                   float64
wind-vwnd-250-2010-16                   float64
wind-vwnd-250-2010-17                   float64
wind-vwnd-250-2010-18                   float64
wind-vwnd-250-2010-19                   float64
wind-vwnd-250-2010-20                   float64
wind-uwnd-250-2010-1                    float64
wind-uwnd-250-2010-2                    float64
wind-uwnd-250-2010-3                    float64
wind-uwnd-250-2010-4                    float64
wind-uwnd-250-2010-5                    float64
wind-uwnd-250-2010-6                    float64
wind-uwnd-250-2010-7                    float64
wind-uwnd-250-2010-8                    float64
wind-uwnd-250-2010-9                    float64
wind-uwnd-250-2010-10                   float64
wind-uwnd-250-2010-11                   float64
wind-uwnd-250-2010-12                   float64
wind-uwnd-250-2010-13                   float64
wind-uwnd-250-2010-14                   float64
wind-uwnd-250-2010-15                   float64
wind-uwnd-250-2010-16                   float64
wind-uwnd-250-2010-17                   float64
wind-uwnd-250-2010-18                   float64
wind-uwnd-250-2010-19                   float64
wind-uwnd-250-2010-20                   float64
mjo1d__phase                            float64
mjo1d__amplitude                        float64
mei__mei                                float64
mei__meirank                            float64
mei__nip                                float64
wind-hgt-850-2010-1                     float64
wind-hgt-850-2010-2                     float64
wind-hgt-850-2010-3                     float64
wind-hgt-850-2010-4                     float64
wind-hgt-850-2010-5                     float64
wind-hgt-850-2010-6                     float64
wind-hgt-850-2010-7                     float64
wind-hgt-850-2010-8                     float64
wind-hgt-850-2010-9                     float64
wind-hgt-850-2010-10                    float64
sst-2010-1                              float64
```

```
sst-2010-2                                    float64
sst-2010-3                                    float64
sst-2010-4                                    float64
sst-2010-5                                    float64
sst-2010-6                                    float64
sst-2010-7                                    float64
sst-2010-8                                    float64
sst-2010-9                                    float64
sst-2010-10                                   float64
wind-hgt-500-2010-1                           float64
wind-hgt-500-2010-2                           float64
wind-hgt-500-2010-3                           float64
wind-hgt-500-2010-4                           float64
wind-hgt-500-2010-5                           float64
wind-hgt-500-2010-6                           float64
wind-hgt-500-2010-7                           float64
wind-hgt-500-2010-8                           float64
wind-hgt-500-2010-9                           float64
wind-hgt-500-2010-10                          float64
icec-2010-1                                   float64
icec-2010-2                                   float64
icec-2010-3                                   float64
icec-2010-4                                   float64
icec-2010-5                                   float64
icec-2010-6                                   float64
icec-2010-7                                   float64
icec-2010-8                                   float64
icec-2010-9                                   float64
icec-2010-10                                  float64
wind-uwnd-925-2010-1                          float64
wind-uwnd-925-2010-2                          float64
wind-uwnd-925-2010-3                          float64
wind-uwnd-925-2010-4                          float64
wind-uwnd-925-2010-5                          float64
wind-uwnd-925-2010-6                          float64
wind-uwnd-925-2010-7                          float64
wind-uwnd-925-2010-8                          float64
wind-uwnd-925-2010-9                          float64
wind-uwnd-925-2010-10                         float64
wind-uwnd-925-2010-11                         float64
wind-uwnd-925-2010-12                         float64
wind-uwnd-925-2010-13                         float64
wind-uwnd-925-2010-14                         float64
wind-uwnd-925-2010-15                         float64
wind-uwnd-925-2010-16                         float64
wind-uwnd-925-2010-17                         float64
wind-uwnd-925-2010-18                         float64
wind-uwnd-925-2010-19                         float64
wind-uwnd-925-2010-20                         float64
wind-hgt-10-2010-1                            float64
wind-hgt-10-2010-2                            float64
wind-hgt-10-2010-3                            float64
```

```
wind-hgt-10-2010-4                      float64
wind-hgt-10-2010-5                      float64
wind-hgt-10-2010-6                      float64
wind-hgt-10-2010-7                      float64
wind-hgt-10-2010-8                      float64
wind-hgt-10-2010-9                      float64
wind-hgt-10-2010-10                     float64
wind-hgt-100-2010-1                     float64
wind-hgt-100-2010-2                     float64
wind-hgt-100-2010-3                     float64
wind-hgt-100-2010-4                     float64
wind-hgt-100-2010-5                     float64
wind-hgt-100-2010-6                     float64
wind-hgt-100-2010-7                     float64
wind-hgt-100-2010-8                     float64
wind-hgt-100-2010-9                     float64
wind-hgt-100-2010-10                    float64
wind-vwnd-925-2010-1                    float64
wind-vwnd-925-2010-2                    float64
wind-vwnd-925-2010-3                    float64
wind-vwnd-925-2010-4                    float64
wind-vwnd-925-2010-5                    float64
wind-vwnd-925-2010-6                    float64
wind-vwnd-925-2010-7                    float64
wind-vwnd-925-2010-8                    float64
wind-vwnd-925-2010-9                    float64
wind-vwnd-925-2010-10                   float64
wind-vwnd-925-2010-11                   float64
wind-vwnd-925-2010-12                   float64
wind-vwnd-925-2010-13                   float64
wind-vwnd-925-2010-14                   float64
wind-vwnd-925-2010-15                   float64
wind-vwnd-925-2010-16                   float64
wind-vwnd-925-2010-17                   float64
wind-vwnd-925-2010-18                   float64
wind-vwnd-925-2010-19                   float64
wind-vwnd-925-2010-20                   float64
dtype: object
```

In [161…

```
# First Date
min = test_data['startdate'].min()
print(min)
```

```
2022-11-01 00:00:00
```

In [162…

```
# Last Date
max = test_data['startdate'].max()
print(max)
```

```
2022-12-31 00:00:00
```

In [163…

```
# Number of days tested - 61 days or 60 day range
range_days= max - min
print(range_days)
```

```
60 days 00:00:00
```

In [340…

```
# Check number of data collections on each day
test_data['startdate'].value_counts()
```

Out[340]:

```
2022-11-01    514
2022-12-02    514
2022-12-04    514
2022-12-05    514
2022-12-06    514
2022-12-07    514
2022-12-08    514
2022-12-09    514
2022-12-10    514
2022-12-11    514
2022-12-12    514
2022-12-13    514
2022-12-14    514
2022-12-15    514
2022-12-16    514
2022-12-17    514
2022-12-18    514
2022-12-19    514
2022-12-20    514
2022-12-21    514
2022-12-22    514
2022-12-23    514
2022-12-24    514
2022-12-25    514
2022-12-26    514
2022-12-27    514
2022-12-28    514
2022-12-29    514
2022-12-30    514
2022-12-03    514
2022-12-01    514
2022-11-02    514
2022-11-30    514
2022-11-03    514
2022-11-04    514
2022-11-05    514
2022-11-06    514
2022-11-07    514
2022-11-08    514
2022-11-09    514
2022-11-10    514
2022-11-11    514
2022-11-12    514
2022-11-13    514
2022-11-14    514
2022-11-15    514
2022-11-16    514
2022-11-17    514
2022-11-18    514
2022-11-19    514
2022-11-20    514
2022-11-21    514
```

```
2022-11-22    514
2022-11-23    514
2022-11-24    514
2022-11-25    514
2022-11-26    514
2022-11-27    514
2022-11-28    514
2022-11-29    514
2022-12-31    514
Name: startdate, dtype: int64
```

Test data - Time frame

61 day window on test data. Everyday of November and December 2022. 514 data entries on each day of test data.

Dummy Variables of Climate Region

In [113…
```
# How many unique climate regiions? --> 15 unique climate regions for train and test
training_data['climateregions__climateregion'].unique()
```

Out[113]:
```
array(['BSh', 'Cfa', 'BSk', 'BWk', 'BWh', 'Csa', 'Csb', 'Cfb', 'Dfb',
       'Dsc', 'Dfc', 'Dfa', 'Dsb', 'Dwa', 'Dwb'], dtype=object)
```

In [165…
```
# Same for test data
test_data['climateregions__climateregion'].unique()
```

Out[165]:
```
array(['BSh', 'Cfa', 'BSk', 'BWk', 'BWh', 'Csa', 'Csb', 'Cfb', 'Dfb',
       'Dsc', 'Dfc', 'Dfa', 'Dsb', 'Dwa', 'Dwb'], dtype=object)
```

Both the test and training sets have 15 different climate regions. We will build out columns to house each dummy variable for these climate region, binary 0 or 1. A "1" will inidicate that data entry is in that climate region.

In [118…
```
# We will use pd.get_dummies method to set up a dummy variables for climate region
train_dummy_climate =
pd.get_dummies(training_data['climateregions__climateregion'])
```

In [119…
```
# Check
train_dummy_climate.head()
```

Out[119]:

	BSh	BSk	BWh	BWk	Cfa	Cfb	Csa	Csb	Dfa	Dfb	Dfc	Dsb	Dsc	Dwa	Dwb
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0

In [120…

```
# Combine dummy variable df with training data then we will drop the
original climate region column
training_data = pd.concat((training_data, train_dummy_climate), axis
= 1)
```

In [123…

```
# Drop original climate region column
training_data =
training_data.drop(['climateregions__climateregion'], axis = 1)
```

In [124…

```
# Check data types of test_data
with pd.option_context('display.max_rows', None,
'display.max_columns', None):
    print(training_data.dtypes)
```

```
index                                          int64
lat                                          float64
lon                                          float64
startdate                              datetime64[ns]
contest-pevpr-sfc-gauss-14d__pevpr           float64
nmme0-tmp2m-34w__cancm30                     float64
nmme0-tmp2m-34w__cancm40                     float64
nmme0-tmp2m-34w__ccsm30                      float64
nmme0-tmp2m-34w__ccsm40                      float64
nmme0-tmp2m-34w__cfsv20                      float64
nmme0-tmp2m-34w__gfdlflora0                  float64
nmme0-tmp2m-34w__gfdlflorb0                  float64
nmme0-tmp2m-34w__gfdl0                       float64
nmme0-tmp2m-34w__nasa0                       float64
nmme0-tmp2m-34w__nmme0mean                   float64
contest-wind-h10-14d__wind-hgt-10            float64
nmme-tmp2m-56w__cancm3                       float64
nmme-tmp2m-56w__cancm4                       float64
nmme-tmp2m-56w__ccsm3                        float64
nmme-tmp2m-56w__ccsm4                        float64
nmme-tmp2m-56w__cfsv2                        float64
nmme-tmp2m-56w__gfdl                         float64
nmme-tmp2m-56w__gfdlflora                    float64
nmme-tmp2m-56w__gfdlflorb                    float64
nmme-tmp2m-56w__nasa                         float64
nmme-tmp2m-56w__nmmemean                     float64
contest-rhum-sig995-14d__rhum                float64
nmme-prate-34w__cancm3                       float64
nmme-prate-34w__cancm4                       float64
nmme-prate-34w__ccsm3                        float64
nmme-prate-34w__ccsm4                        float64
nmme-prate-34w__cfsv2                        float64
nmme-prate-34w__gfdl                         float64
nmme-prate-34w__gfdlflora                    float64
nmme-prate-34w__gfdlflorb                    float64
nmme-prate-34w__nasa                         float64
nmme-prate-34w__nmmemean                     float64
contest-wind-h100-14d__wind-hgt-100          float64
nmme0-prate-56w__cancm30                     float64
nmme0-prate-56w__cancm40                     float64
nmme0-prate-56w__ccsm30                      float64
nmme0-prate-56w__ccsm40                      float64
nmme0-prate-56w__cfsv20                      float64
nmme0-prate-56w__gfdlflora0                  float64
nmme0-prate-56w__gfdlflorb0                  float64
nmme0-prate-56w__gfdl0                       float64
nmme0-prate-56w__nasa0                       float64
nmme0-prate-56w__nmme0mean                   float64
nmme0-prate-34w__cancm30                     float64
nmme0-prate-34w__cancm40                     float64
nmme0-prate-34w__ccsm30                      float64
nmme0-prate-34w__ccsm40                      float64
```

```
nmme0-prate-34w__cfsv20                          float64
nmme0-prate-34w__gfdlflora0                      float64
nmme0-prate-34w__gfdlflorb0                      float64
nmme0-prate-34w__gfdl0                           float64
nmme0-prate-34w__nasa0                           float64
nmme0-prate-34w__nmme0mean                       float64
contest-tmp2m-14d__tmp2m                         float64
contest-slp-14d__slp                             float64
contest-wind-vwnd-925-14d__wind-vwnd-925         float64
nmme-prate-56w__cancm3                           float64
nmme-prate-56w__cancm4                           float64
nmme-prate-56w__ccsm3                            float64
nmme-prate-56w__ccsm4                            float64
nmme-prate-56w__cfsv2                            float64
nmme-prate-56w__gfdl                             float64
nmme-prate-56w__gfdlflora                        float64
nmme-prate-56w__gfdlflorb                        float64
nmme-prate-56w__nasa                             float64
nmme-prate-56w__nmmemean                         float64
contest-pres-sfc-gauss-14d__pres                 float64
contest-wind-uwnd-250-14d__wind-uwnd-250         float64
nmme-tmp2m-34w__cancm3                           float64
nmme-tmp2m-34w__cancm4                           float64
nmme-tmp2m-34w__ccsm3                            float64
nmme-tmp2m-34w__ccsm4                            float64
nmme-tmp2m-34w__cfsv2                            float64
nmme-tmp2m-34w__gfdl                             float64
nmme-tmp2m-34w__gfdlflora                        float64
nmme-tmp2m-34w__gfdlflorb                        float64
nmme-tmp2m-34w__nasa                             float64
nmme-tmp2m-34w__nmmemean                         float64
contest-prwtr-eatm-14d__prwtr                    float64
contest-wind-vwnd-250-14d__wind-vwnd-250         float64
contest-precip-14d__precip                       float64
contest-wind-h850-14d__wind-hgt-850              float64
contest-wind-uwnd-925-14d__wind-uwnd-925         float64
contest-wind-h500-14d__wind-hgt-500              float64
cancm30                                          float64
cancm40                                          float64
ccsm30                                           float64
ccsm40                                           float64
cfsv20                                           float64
gfdlflora0                                       float64
gfdlflorb0                                       float64
gfdl0                                            float64
nasa0                                            float64
nmme0mean                                        float64
elevation__elevation                             float64
wind-vwnd-250-2010-1                             float64
wind-vwnd-250-2010-2                             float64
wind-vwnd-250-2010-3                             float64
wind-vwnd-250-2010-4                             float64
```

```
wind-vwnd-250-2010-5                    float64
wind-vwnd-250-2010-6                    float64
wind-vwnd-250-2010-7                    float64
wind-vwnd-250-2010-8                    float64
wind-vwnd-250-2010-9                    float64
wind-vwnd-250-2010-10                   float64
wind-vwnd-250-2010-11                   float64
wind-vwnd-250-2010-12                   float64
wind-vwnd-250-2010-13                   float64
wind-vwnd-250-2010-14                   float64
wind-vwnd-250-2010-15                   float64
wind-vwnd-250-2010-16                   float64
wind-vwnd-250-2010-17                   float64
wind-vwnd-250-2010-18                   float64
wind-vwnd-250-2010-19                   float64
wind-vwnd-250-2010-20                   float64
wind-uwnd-250-2010-1                    float64
wind-uwnd-250-2010-2                    float64
wind-uwnd-250-2010-3                    float64
wind-uwnd-250-2010-4                    float64
wind-uwnd-250-2010-5                    float64
wind-uwnd-250-2010-6                    float64
wind-uwnd-250-2010-7                    float64
wind-uwnd-250-2010-8                    float64
wind-uwnd-250-2010-9                    float64
wind-uwnd-250-2010-10                   float64
wind-uwnd-250-2010-11                   float64
wind-uwnd-250-2010-12                   float64
wind-uwnd-250-2010-13                   float64
wind-uwnd-250-2010-14                   float64
wind-uwnd-250-2010-15                   float64
wind-uwnd-250-2010-16                   float64
wind-uwnd-250-2010-17                   float64
wind-uwnd-250-2010-18                   float64
wind-uwnd-250-2010-19                   float64
wind-uwnd-250-2010-20                   float64
mjo1d__phase                              int64
mjo1d__amplitude                        float64
mei__mei                                float64
mei__meirank                              int64
mei__nip                                  int64
wind-hgt-850-2010-1                     float64
wind-hgt-850-2010-2                     float64
wind-hgt-850-2010-3                     float64
wind-hgt-850-2010-4                     float64
wind-hgt-850-2010-5                     float64
wind-hgt-850-2010-6                     float64
wind-hgt-850-2010-7                     float64
wind-hgt-850-2010-8                     float64
wind-hgt-850-2010-9                     float64
wind-hgt-850-2010-10                    float64
sst-2010-1                              float64
```

```
sst-2010-2                                      float64
sst-2010-3                                      float64
sst-2010-4                                      float64
sst-2010-5                                      float64
sst-2010-6                                      float64
sst-2010-7                                      float64
sst-2010-8                                      float64
sst-2010-9                                      float64
sst-2010-10                                     float64
wind-hgt-500-2010-1                             float64
wind-hgt-500-2010-2                             float64
wind-hgt-500-2010-3                             float64
wind-hgt-500-2010-4                             float64
wind-hgt-500-2010-5                             float64
wind-hgt-500-2010-6                             float64
wind-hgt-500-2010-7                             float64
wind-hgt-500-2010-8                             float64
wind-hgt-500-2010-9                             float64
wind-hgt-500-2010-10                            float64
icec-2010-1                                     float64
icec-2010-2                                     float64
icec-2010-3                                     float64
icec-2010-4                                     float64
icec-2010-5                                     float64
icec-2010-6                                     float64
icec-2010-7                                     float64
icec-2010-8                                     float64
icec-2010-9                                     float64
icec-2010-10                                    float64
wind-uwnd-925-2010-1                            float64
wind-uwnd-925-2010-2                            float64
wind-uwnd-925-2010-3                            float64
wind-uwnd-925-2010-4                            float64
wind-uwnd-925-2010-5                            float64
wind-uwnd-925-2010-6                            float64
wind-uwnd-925-2010-7                            float64
wind-uwnd-925-2010-8                            float64
wind-uwnd-925-2010-9                            float64
wind-uwnd-925-2010-10                           float64
wind-uwnd-925-2010-11                           float64
wind-uwnd-925-2010-12                           float64
wind-uwnd-925-2010-13                           float64
wind-uwnd-925-2010-14                           float64
wind-uwnd-925-2010-15                           float64
wind-uwnd-925-2010-16                           float64
wind-uwnd-925-2010-17                           float64
wind-uwnd-925-2010-18                           float64
wind-uwnd-925-2010-19                           float64
wind-uwnd-925-2010-20                           float64
wind-hgt-10-2010-1                              float64
wind-hgt-10-2010-2                              float64
wind-hgt-10-2010-3                              float64
```

```
wind-hgt-10-2010-4                      float64
wind-hgt-10-2010-5                      float64
wind-hgt-10-2010-6                      float64
wind-hgt-10-2010-7                      float64
wind-hgt-10-2010-8                      float64
wind-hgt-10-2010-9                      float64
wind-hgt-10-2010-10                     float64
wind-hgt-100-2010-1                     float64
wind-hgt-100-2010-2                     float64
wind-hgt-100-2010-3                     float64
wind-hgt-100-2010-4                     float64
wind-hgt-100-2010-5                     float64
wind-hgt-100-2010-6                     float64
wind-hgt-100-2010-7                     float64
wind-hgt-100-2010-8                     float64
wind-hgt-100-2010-9                     float64
wind-hgt-100-2010-10                    float64
wind-vwnd-925-2010-1                    float64
wind-vwnd-925-2010-2                    float64
wind-vwnd-925-2010-3                    float64
wind-vwnd-925-2010-4                    float64
wind-vwnd-925-2010-5                    float64
wind-vwnd-925-2010-6                    float64
wind-vwnd-925-2010-7                    float64
wind-vwnd-925-2010-8                    float64
wind-vwnd-925-2010-9                    float64
wind-vwnd-925-2010-10                   float64
wind-vwnd-925-2010-11                   float64
wind-vwnd-925-2010-12                   float64
wind-vwnd-925-2010-13                   float64
wind-vwnd-925-2010-14                   float64
wind-vwnd-925-2010-15                   float64
wind-vwnd-925-2010-16                   float64
wind-vwnd-925-2010-17                   float64
wind-vwnd-925-2010-18                   float64
wind-vwnd-925-2010-19                   float64
wind-vwnd-925-2010-20                   float64
BSh                                       uint8
BSk                                       uint8
BWh                                       uint8
BWk                                       uint8
Cfa                                       uint8
Cfb                                       uint8
Csa                                       uint8
Csb                                       uint8
Dfa                                       uint8
Dfb                                       uint8
Dfc                                       uint8
Dsb                                       uint8
Dsc                                       uint8
Dwa                                       uint8
```

```
Dwb                                                    uint8
dtype: object
```

Training_data is numerical with no null values.

In [166…
```
# We will use pd.get_dummies method to set up a dummy variables for climate region
test_dummy_climate =
pd.get_dummies(test_data['climateregions__climateregion'])
```

In [167…
```
# Check
test_dummy_climate.head()
```

Out[167]:

	BSh	BSk	BWh	BWk	Cfa	Cfb	Csa	Csb	Dfa	Dfb	Dfc	Dsb	Dsc	Dwa	Dwb
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0

In [168…
```
# Combine dummy variable df with test data then we will drop the original climate region column
test_data = pd.concat((test_data, test_dummy_climate), axis = 1)
```

In [169…
```
# Drop original climate region column
test_data = test_data.drop(['climateregions__climateregion'], axis = 1)
```

In [170…
```
# Check data types of test_data
with pd.option_context('display.max_rows', None,
'display.max_columns', None):
    print(test_data.dtypes)
```

```
index                                          int64
lat                                          float64
lon                                          float64
startdate                              datetime64[ns]
contest-pevpr-sfc-gauss-14d__pevpr           float64
nmme0-tmp2m-34w__cancm30                     float64
nmme0-tmp2m-34w__cancm40                     float64
nmme0-tmp2m-34w__ccsm30                      float64
nmme0-tmp2m-34w__ccsm40                      float64
nmme0-tmp2m-34w__cfsv20                      float64
nmme0-tmp2m-34w__gfdlflora0                  float64
nmme0-tmp2m-34w__gfdlflorb0                  float64
nmme0-tmp2m-34w__gfdl0                       float64
nmme0-tmp2m-34w__nasa0                       float64
nmme0-tmp2m-34w__nmme0mean                   float64
contest-wind-h10-14d__wind-hgt-10            float64
nmme-tmp2m-56w__cancm3                       float64
nmme-tmp2m-56w__cancm4                       float64
nmme-tmp2m-56w__ccsm3                        float64
nmme-tmp2m-56w__ccsm4                        float64
nmme-tmp2m-56w__cfsv2                        float64
nmme-tmp2m-56w__gfdl                         float64
nmme-tmp2m-56w__gfdlflora                    float64
nmme-tmp2m-56w__gfdlflorb                    float64
nmme-tmp2m-56w__nasa                         float64
nmme-tmp2m-56w__nmmemean                     float64
contest-rhum-sig995-14d__rhum                float64
nmme-prate-34w__cancm3                       float64
nmme-prate-34w__cancm4                       float64
nmme-prate-34w__ccsm3                        float64
nmme-prate-34w__ccsm4                        float64
nmme-prate-34w__cfsv2                        float64
nmme-prate-34w__gfdl                         float64
nmme-prate-34w__gfdlflora                    float64
nmme-prate-34w__gfdlflorb                    float64
nmme-prate-34w__nasa                         float64
nmme-prate-34w__nmmemean                     float64
contest-wind-h100-14d__wind-hgt-100          float64
nmme0-prate-56w__cancm30                     float64
nmme0-prate-56w__cancm40                     float64
nmme0-prate-56w__ccsm30                      float64
nmme0-prate-56w__ccsm40                      float64
nmme0-prate-56w__cfsv20                      float64
nmme0-prate-56w__gfdlflora0                  float64
nmme0-prate-56w__gfdlflorb0                  float64
nmme0-prate-56w__gfdl0                       float64
nmme0-prate-56w__nasa0                       float64
nmme0-prate-56w__nmme0mean                   float64
nmme0-prate-34w__cancm30                     float64
nmme0-prate-34w__cancm40                     float64
nmme0-prate-34w__ccsm30                      float64
nmme0-prate-34w__ccsm40                      float64
```

```
nmme0-prate-34w__cfsv20                          float64
nmme0-prate-34w__gfdlflora0                      float64
nmme0-prate-34w__gfdlflorb0                      float64
nmme0-prate-34w__gfdl0                           float64
nmme0-prate-34w__nasa0                           float64
nmme0-prate-34w__nmme0mean                       float64
contest-slp-14d__slp                             float64
contest-wind-vwnd-925-14d__wind-vwnd-925         float64
nmme-prate-56w__cancm3                           float64
nmme-prate-56w__cancm4                           float64
nmme-prate-56w__ccsm3                            float64
nmme-prate-56w__ccsm4                            float64
nmme-prate-56w__cfsv2                            float64
nmme-prate-56w__gfdl                             float64
nmme-prate-56w__gfdlflora                        float64
nmme-prate-56w__gfdlflorb                        float64
nmme-prate-56w__nasa                             float64
nmme-prate-56w__nmmemean                         float64
contest-pres-sfc-gauss-14d__pres                 float64
contest-wind-uwnd-250-14d__wind-uwnd-250         float64
nmme-tmp2m-34w__cancm3                           float64
nmme-tmp2m-34w__cancm4                           float64
nmme-tmp2m-34w__ccsm3                            float64
nmme-tmp2m-34w__ccsm4                            float64
nmme-tmp2m-34w__cfsv2                            float64
nmme-tmp2m-34w__gfdl                             float64
nmme-tmp2m-34w__gfdlflora                        float64
nmme-tmp2m-34w__gfdlflorb                        float64
nmme-tmp2m-34w__nasa                             float64
nmme-tmp2m-34w__nmmemean                         float64
contest-prwtr-eatm-14d__prwtr                    float64
contest-wind-vwnd-250-14d__wind-vwnd-250         float64
contest-precip-14d__precip                       float64
contest-wind-h850-14d__wind-hgt-850              float64
contest-wind-uwnd-925-14d__wind-uwnd-925         float64
contest-wind-h500-14d__wind-hgt-500              float64
cancm30                                          float64
cancm40                                          float64
ccsm30                                           float64
ccsm40                                           float64
cfsv20                                           float64
gfdlflora0                                       float64
gfdlflorb0                                       float64
gfdl0                                            float64
nasa0                                            float64
nmme0mean                                        float64
elevation__elevation                               int64
wind-vwnd-250-2010-1                             float64
wind-vwnd-250-2010-2                             float64
wind-vwnd-250-2010-3                             float64
wind-vwnd-250-2010-4                             float64
wind-vwnd-250-2010-5                             float64
```

```
wind-vwnd-250-2010-6                    float64
wind-vwnd-250-2010-7                    float64
wind-vwnd-250-2010-8                    float64
wind-vwnd-250-2010-9                    float64
wind-vwnd-250-2010-10                   float64
wind-vwnd-250-2010-11                   float64
wind-vwnd-250-2010-12                   float64
wind-vwnd-250-2010-13                   float64
wind-vwnd-250-2010-14                   float64
wind-vwnd-250-2010-15                   float64
wind-vwnd-250-2010-16                   float64
wind-vwnd-250-2010-17                   float64
wind-vwnd-250-2010-18                   float64
wind-vwnd-250-2010-19                   float64
wind-vwnd-250-2010-20                   float64
wind-uwnd-250-2010-1                    float64
wind-uwnd-250-2010-2                    float64
wind-uwnd-250-2010-3                    float64
wind-uwnd-250-2010-4                    float64
wind-uwnd-250-2010-5                    float64
wind-uwnd-250-2010-6                    float64
wind-uwnd-250-2010-7                    float64
wind-uwnd-250-2010-8                    float64
wind-uwnd-250-2010-9                    float64
wind-uwnd-250-2010-10                   float64
wind-uwnd-250-2010-11                   float64
wind-uwnd-250-2010-12                   float64
wind-uwnd-250-2010-13                   float64
wind-uwnd-250-2010-14                   float64
wind-uwnd-250-2010-15                   float64
wind-uwnd-250-2010-16                   float64
wind-uwnd-250-2010-17                   float64
wind-uwnd-250-2010-18                   float64
wind-uwnd-250-2010-19                   float64
wind-uwnd-250-2010-20                   float64
mjo1d__phase                            float64
mjo1d__amplitude                        float64
mei__mei                                float64
mei__meirank                            float64
mei__nip                                float64
wind-hgt-850-2010-1                     float64
wind-hgt-850-2010-2                     float64
wind-hgt-850-2010-3                     float64
wind-hgt-850-2010-4                     float64
wind-hgt-850-2010-5                     float64
wind-hgt-850-2010-6                     float64
wind-hgt-850-2010-7                     float64
wind-hgt-850-2010-8                     float64
wind-hgt-850-2010-9                     float64
wind-hgt-850-2010-10                    float64
sst-2010-1                              float64
sst-2010-2                              float64
```

```
sst-2010-3                                      float64
sst-2010-4                                      float64
sst-2010-5                                      float64
sst-2010-6                                      float64
sst-2010-7                                      float64
sst-2010-8                                      float64
sst-2010-9                                      float64
sst-2010-10                                     float64
wind-hgt-500-2010-1                             float64
wind-hgt-500-2010-2                             float64
wind-hgt-500-2010-3                             float64
wind-hgt-500-2010-4                             float64
wind-hgt-500-2010-5                             float64
wind-hgt-500-2010-6                             float64
wind-hgt-500-2010-7                             float64
wind-hgt-500-2010-8                             float64
wind-hgt-500-2010-9                             float64
wind-hgt-500-2010-10                            float64
icec-2010-1                                     float64
icec-2010-2                                     float64
icec-2010-3                                     float64
icec-2010-4                                     float64
icec-2010-5                                     float64
icec-2010-6                                     float64
icec-2010-7                                     float64
icec-2010-8                                     float64
icec-2010-9                                     float64
icec-2010-10                                    float64
wind-uwnd-925-2010-1                            float64
wind-uwnd-925-2010-2                            float64
wind-uwnd-925-2010-3                            float64
wind-uwnd-925-2010-4                            float64
wind-uwnd-925-2010-5                            float64
wind-uwnd-925-2010-6                            float64
wind-uwnd-925-2010-7                            float64
wind-uwnd-925-2010-8                            float64
wind-uwnd-925-2010-9                            float64
wind-uwnd-925-2010-10                           float64
wind-uwnd-925-2010-11                           float64
wind-uwnd-925-2010-12                           float64
wind-uwnd-925-2010-13                           float64
wind-uwnd-925-2010-14                           float64
wind-uwnd-925-2010-15                           float64
wind-uwnd-925-2010-16                           float64
wind-uwnd-925-2010-17                           float64
wind-uwnd-925-2010-18                           float64
wind-uwnd-925-2010-19                           float64
wind-uwnd-925-2010-20                           float64
wind-hgt-10-2010-1                              float64
wind-hgt-10-2010-2                              float64
wind-hgt-10-2010-3                              float64
wind-hgt-10-2010-4                              float64
```

```
wind-hgt-10-2010-5                          float64
wind-hgt-10-2010-6                          float64
wind-hgt-10-2010-7                          float64
wind-hgt-10-2010-8                          float64
wind-hgt-10-2010-9                          float64
wind-hgt-10-2010-10                         float64
wind-hgt-100-2010-1                         float64
wind-hgt-100-2010-2                         float64
wind-hgt-100-2010-3                         float64
wind-hgt-100-2010-4                         float64
wind-hgt-100-2010-5                         float64
wind-hgt-100-2010-6                         float64
wind-hgt-100-2010-7                         float64
wind-hgt-100-2010-8                         float64
wind-hgt-100-2010-9                         float64
wind-hgt-100-2010-10                        float64
wind-vwnd-925-2010-1                        float64
wind-vwnd-925-2010-2                        float64
wind-vwnd-925-2010-3                        float64
wind-vwnd-925-2010-4                        float64
wind-vwnd-925-2010-5                        float64
wind-vwnd-925-2010-6                        float64
wind-vwnd-925-2010-7                        float64
wind-vwnd-925-2010-8                        float64
wind-vwnd-925-2010-9                        float64
wind-vwnd-925-2010-10                       float64
wind-vwnd-925-2010-11                       float64
wind-vwnd-925-2010-12                       float64
wind-vwnd-925-2010-13                       float64
wind-vwnd-925-2010-14                       float64
wind-vwnd-925-2010-15                       float64
wind-vwnd-925-2010-16                       float64
wind-vwnd-925-2010-17                       float64
wind-vwnd-925-2010-18                       float64
wind-vwnd-925-2010-19                       float64
wind-vwnd-925-2010-20                       float64
BSh                                           uint8
BSk                                           uint8
BWh                                           uint8
BWk                                           uint8
Cfa                                           uint8
Cfb                                           uint8
Csa                                           uint8
Csb                                           uint8
Dfa                                           uint8
Dfb                                           uint8
Dfc                                           uint8
Dsb                                           uint8
Dsc                                           uint8
Dwa                                           uint8
Dwb                                           uint8
dtype: object
```

Test_data is fully numerical and has no null values.

Duplicated Data or Rows

In [171…
```
# Check if there are any duplicated data points in training set
training_data.duplicated().sum()
```

Out[171]:
```
0
```

In [172…
```
# Check if there are any duplicated data points in test set
test_data.duplicated().sum()
```

Out[172]:
```
0
```

No duplicates to take care of.

Make columns of Month and Season

- Use datetime to make a date column for month.
- Come up with algorithim for season of the year, summer, fall, winter, spring.

In [147…
```
# Add column of month_number on training_data
training_data['month_number'] = training_data['startdate'].dt.month
```

In [173…
```
# Add column of month_number on test_data
test_data['month_number'] = test_data['startdate'].dt.month
```

In [175…
```
#check
training_data.head()
```

Out[175]:

	index	lat	lon	startdate	contest-pevpr-sfc-gauss-14d__pevpr	nmme0-tmp2m-34w__cancm30	nmme0-tmp2m-34w__cancm40	nmme0 tmp2m 34w__ccsm3
0	0	0.0	0.833333	2014-09-01	237.00	29.02	31.64	29.5
1	1	0.0	0.833333	2014-09-02	228.90	29.02	31.64	29.5
2	2	0.0	0.833333	2014-09-03	220.69	29.02	31.64	29.5
3	3	0.0	0.833333	2014-09-04	225.28	29.02	31.64	29.5
4	4	0.0	0.833333	2014-09-05	237.24	29.02	31.64	29.5

In [176…

```
# Check
test_data.tail()
```

Out[176]:

	index	lat	lon	startdate	contest-pevpr-sfc-gauss-14d__pevpr	nmme0-tmp2m-34w__cancm30	nmme0-tmp2m-34w__cancm40	34w__
31349	407083	1.0	0.866667	2022-12-27	62.72	4.6	8.71	
31350	407084	1.0	0.866667	2022-12-28	73.41	4.6	8.71	
31351	407085	1.0	0.866667	2022-12-29	70.00	4.6	8.71	
31352	407086	1.0	0.866667	2022-12-30	79.81	4.6	8.71	
31353	407087	1.0	0.866667	2022-12-31	86.17	4.6	8.71	

In [186…

```
# Set up season columns with a season number (Winter = 1, Spring =
2, Summer = 3, Fall = 4)
# Test algorithm from
https://stackoverflow.com/questions/44124436/python-datetime-to-
season
for month_num in range (1,13):
    season_number = month_num %12 // 3 + 1
    print(f'Month number, {month_num}, gives you this season number,
{season_number}.')
```

```
Month number, 1, gives you this season number, 1.
Month number, 2, gives you this season number, 1.
Month number, 3, gives you this season number, 2.
Month number, 4, gives you this season number, 2.
Month number, 5, gives you this season number, 2.
Month number, 6, gives you this season number, 3.
Month number, 7, gives you this season number, 3.
Month number, 8, gives you this season number, 3.
Month number, 9, gives you this season number, 4.
Month number, 10, gives you this season number, 4.
Month number, 11, gives you this season number, 4.
Month number, 12, gives you this season number, 1.
```

In [188…

```
# Set up season column with season numberr in training_data
training_data['season_number'] = training_data['month_number']%12 //
3 + 1
```

In [190…

```
# Check
training_data.head(100)
```

Out[190]:

	index	lat	lon	startdate	contest-pevpr-sfc-gauss-14d__pevpr	nmme0-tmp2m-34w__cancm30	nmme0-tmp2m-34w__cancm40	nmme tmp2 34w__ccsm
0	0	0.0	0.833333	2014-09-01	237.00	29.02	31.64	29
1	1	0.0	0.833333	2014-09-02	228.90	29.02	31.64	29
2	2	0.0	0.833333	2014-09-03	220.69	29.02	31.64	29
3	3	0.0	0.833333	2014-09-04	225.28	29.02	31.64	29
4	4	0.0	0.833333	2014-09-05	237.24	29.02	31.64	29
5	5	0.0	0.833333	2014-09-06	237.87	29.02	31.64	29
6	6	0.0	0.833333	2014-09-07	236.36	29.02	31.64	29
7	7	0.0	0.833333	2014-09-08	233.36	29.02	31.64	29
8	8	0.0	0.833333	2014-09-09	233.82	29.02	31.64	29
9	9	0.0	0.833333	2014-09-10	229.74	29.02	31.64	29
10	10	0.0	0.833333	2014-09-11	220.59	29.02	31.64	29
11	11	0.0	0.833333	2014-09-12	208.32	29.02	31.64	29
12	12	0.0	0.833333	2014-09-13	198.76	29.02	31.64	29
13	13	0.0	0.833333	2014-09-14	196.75	29.02	31.64	29
14	14	0.0	0.833333	2014-09-15	195.16	29.02	31.64	29
15	15	0.0	0.833333	2014-09-16	195.87	29.02	31.64	29
16	16	0.0	0.833333	2014-09-17	197.96	29.02	31.64	29
17	17	0.0	0.833333	2014-09-18	201.64	29.02	31.64	29
18	18	0.0	0.833333	2014-09-19	201.59	29.02	31.64	29
19	19	0.0	0.833333	2014-09-20	204.63	29.02	31.64	29
20	20	0.0	0.833333	2014-09-21	216.39	29.02	31.64	29

	index	lat	lon	startdate	contest-pevpr-sfc-gauss-14d__pevpr	nmme0-tmp2m-34w__cancm30	nmme0-tmp2m-34w__cancm40	nmme tmp2 34w__ccsm
21	21	0.0	0.833333	2014-09-22	228.88	29.02	31.64	29
22	22	0.0	0.833333	2014-09-23	230.29	26.87	27.15	28
23	23	0.0	0.833333	2014-09-24	232.52	26.87	27.15	28
24	24	0.0	0.833333	2014-09-25	238.10	26.87	27.15	28
25	25	0.0	0.833333	2014-09-26	246.83	26.87	27.15	28
26	26	0.0	0.833333	2014-09-27	265.05	26.87	27.15	28
27	27	0.0	0.833333	2014-09-28	272.41	26.87	27.15	28
28	28	0.0	0.833333	2014-09-29	271.33	26.87	27.15	28
29	29	0.0	0.833333	2014-09-30	283.48	26.87	27.15	28
30	30	0.0	0.833333	2014-10-01	295.56	26.87	27.15	28
31	31	0.0	0.833333	2014-10-02	296.27	26.87	27.15	28
32	32	0.0	0.833333	2014-10-03	296.27	26.87	27.15	28
33	33	0.0	0.833333	2014-10-04	302.29	26.87	27.15	28
34	34	0.0	0.833333	2014-10-05	297.77	26.87	27.15	28
35	35	0.0	0.833333	2014-10-06	284.35	26.87	27.15	28
36	36	0.0	0.833333	2014-10-07	274.58	26.87	27.15	28
37	37	0.0	0.833333	2014-10-08	263.68	26.87	27.15	28
38	38	0.0	0.833333	2014-10-09	251.84	26.87	27.15	28
39	39	0.0	0.833333	2014-10-10	239.36	26.87	27.15	28
40	40	0.0	0.833333	2014-10-11	226.89	26.87	27.15	28
41	41	0.0	0.833333	2014-10-12	228.63	26.87	27.15	28

	index	lat	lon	startdate	contest-pevpr-sfc-gauss-14d__pevpr	nmme0-tmp2m-34w__cancm30	nmme0-tmp2m-34w__cancm40	nmme tmp2 34w__ccsm
42	42	0.0	0.833333	2014-10-13	237.68	26.87	27.15	28
43	43	0.0	0.833333	2014-10-14	233.77	26.87	27.15	28
44	44	0.0	0.833333	2014-10-15	225.43	26.87	27.15	28
45	45	0.0	0.833333	2014-10-16	217.21	26.87	27.15	28
46	46	0.0	0.833333	2014-10-17	208.81	26.87	27.15	28
47	47	0.0	0.833333	2014-10-18	199.79	26.87	27.15	28
48	48	0.0	0.833333	2014-10-19	203.79	26.87	27.15	28
49	49	0.0	0.833333	2014-10-20	216.33	26.87	27.15	28
50	50	0.0	0.833333	2014-10-21	233.12	26.87	27.15	28
51	51	0.0	0.833333	2014-10-22	244.63	26.87	27.15	28
52	52	0.0	0.833333	2014-10-23	243.42	22.82	26.38	23
53	53	0.0	0.833333	2014-10-24	235.14	22.82	26.38	23
54	54	0.0	0.833333	2014-10-25	226.37	22.82	26.38	23
55	55	0.0	0.833333	2014-10-26	220.72	22.82	26.38	23
56	56	0.0	0.833333	2014-10-27	215.35	22.82	26.38	23
57	57	0.0	0.833333	2014-10-28	207.04	22.82	26.38	23
58	58	0.0	0.833333	2014-10-29	201.30	22.82	26.38	23
59	59	0.0	0.833333	2014-10-30	207.50	22.82	26.38	23
60	60	0.0	0.833333	2014-10-31	215.13	22.82	26.38	23
61	61	0.0	0.833333	2014-11-01	221.10	22.82	26.38	23
62	62	0.0	0.833333	2014-11-02	213.28	22.82	26.38	23

	index	lat	lon	startdate	contest-pevpr-sfc-gauss-14d__pevpr	nmme0-tmp2m-34w__cancm30	nmme0-tmp2m-34w__cancm40	nmme tmp2 34w__ccsm
63	63	0.0	0.833333	2014-11-03	203.98	22.82	26.38	23
64	64	0.0	0.833333	2014-11-04	205.92	22.82	26.38	23
65	65	0.0	0.833333	2014-11-05	201.39	22.82	26.38	23
66	66	0.0	0.833333	2014-11-06	200.26	22.82	26.38	23
67	67	0.0	0.833333	2014-11-07	203.77	22.82	26.38	23
68	68	0.0	0.833333	2014-11-08	203.79	22.82	26.38	23
69	69	0.0	0.833333	2014-11-09	200.77	22.82	26.38	23
70	70	0.0	0.833333	2014-11-10	202.40	22.82	26.38	23
71	71	0.0	0.833333	2014-11-11	206.77	22.82	26.38	23
72	72	0.0	0.833333	2014-11-12	206.18	22.82	26.38	23
73	73	0.0	0.833333	2014-11-13	198.71	22.82	26.38	23
74	74	0.0	0.833333	2014-11-14	189.41	22.82	26.38	23
75	75	0.0	0.833333	2014-11-15	183.12	22.82	26.38	23
76	76	0.0	0.833333	2014-11-16	186.39	22.82	26.38	23
77	77	0.0	0.833333	2014-11-17	187.13	22.82	26.38	23
78	78	0.0	0.833333	2014-11-18	171.31	22.82	26.38	23
79	79	0.0	0.833333	2014-11-19	164.24	22.82	26.38	23
80	80	0.0	0.833333	2014-11-20	162.15	22.82	26.38	23
81	81	0.0	0.833333	2014-11-21	162.79	22.82	26.38	23
82	82	0.0	0.833333	2014-11-22	166.86	22.82	26.38	23
83	83	0.0	0.833333	2014-11-23	168.24	16.91	21.65	13

	index	lat	lon	startdate	contest-pevpr-sfc-gauss-14d__pevpr	nmme0-tmp2m-34w__cancm30	nmme0-tmp2m-34w__cancm40	nmme tmp2 34w__ccsm
84	84	0.0	0.833333	2014-11-24	162.08	16.91	21.65	13
85	85	0.0	0.833333	2014-11-25	148.95	16.91	21.65	13
86	86	0.0	0.833333	2014-11-26	141.57	16.91	21.65	13
87	87	0.0	0.833333	2014-11-27	137.39	16.91	21.65	13
88	88	0.0	0.833333	2014-11-28	133.33	16.91	21.65	13
89	89	0.0	0.833333	2014-11-29	130.50	16.91	21.65	13
90	90	0.0	0.833333	2014-11-30	125.57	16.91	21.65	13
91	91	0.0	0.833333	2014-12-01	123.22	16.91	21.65	13
92	92	0.0	0.833333	2014-12-02	121.44	16.91	21.65	13
93	93	0.0	0.833333	2014-12-03	120.95	16.91	21.65	13
94	94	0.0	0.833333	2014-12-04	121.83	16.91	21.65	13
95	95	0.0	0.833333	2014-12-05	117.54	16.91	21.65	13
96	96	0.0	0.833333	2014-12-06	111.51	16.91	21.65	13
97	97	0.0	0.833333	2014-12-07	105.98	16.91	21.65	13
98	98	0.0	0.833333	2014-12-08	101.69	16.91	21.65	13
99	99	0.0	0.833333	2014-12-09	104.63	16.91	21.65	13

In [191…

```
# Set up season column with season numberr in test_data
test_data['season_number'] = test_data['month_number']%12 // 3 + 1
```

In [194…

```
# Check
test_data.tail()
```

Out[194]:

	index	lat	lon	startdate	contest-pevpr-sfc-gauss-14d__pevpr	nmme0-tmp2m-34w__cancm30	nmme0-tmp2m-34w__cancm40	34w__
31349	407083	1.0	0.866667	2022-12-27	62.72	4.6	8.71	
31350	407084	1.0	0.866667	2022-12-28	73.41	4.6	8.71	
31351	407085	1.0	0.866667	2022-12-29	70.00	4.6	8.71	
31352	407086	1.0	0.866667	2022-12-30	79.81	4.6	8.71	
31353	407087	1.0	0.866667	2022-12-31	86.17	4.6	8.71	

More EDA - Visualize Target by Month and Corelation Analysis

Visualize Monthly Temperature Variation

- Set up line plot to visualize temp change by month

In [204…

```
# Aggregate training data with groupby of month_number to pull out
min, max, and mean
monthly_temp_df = training_data.groupby('month_number').agg(['min',
'max','mean'])
```

In [205…

```
# Isolate just target variable - or mean temp of next 14 days
monthly_temp_df = monthly_temp_df['contest-tmp2m-14d__tmp2m']
```

In [207…

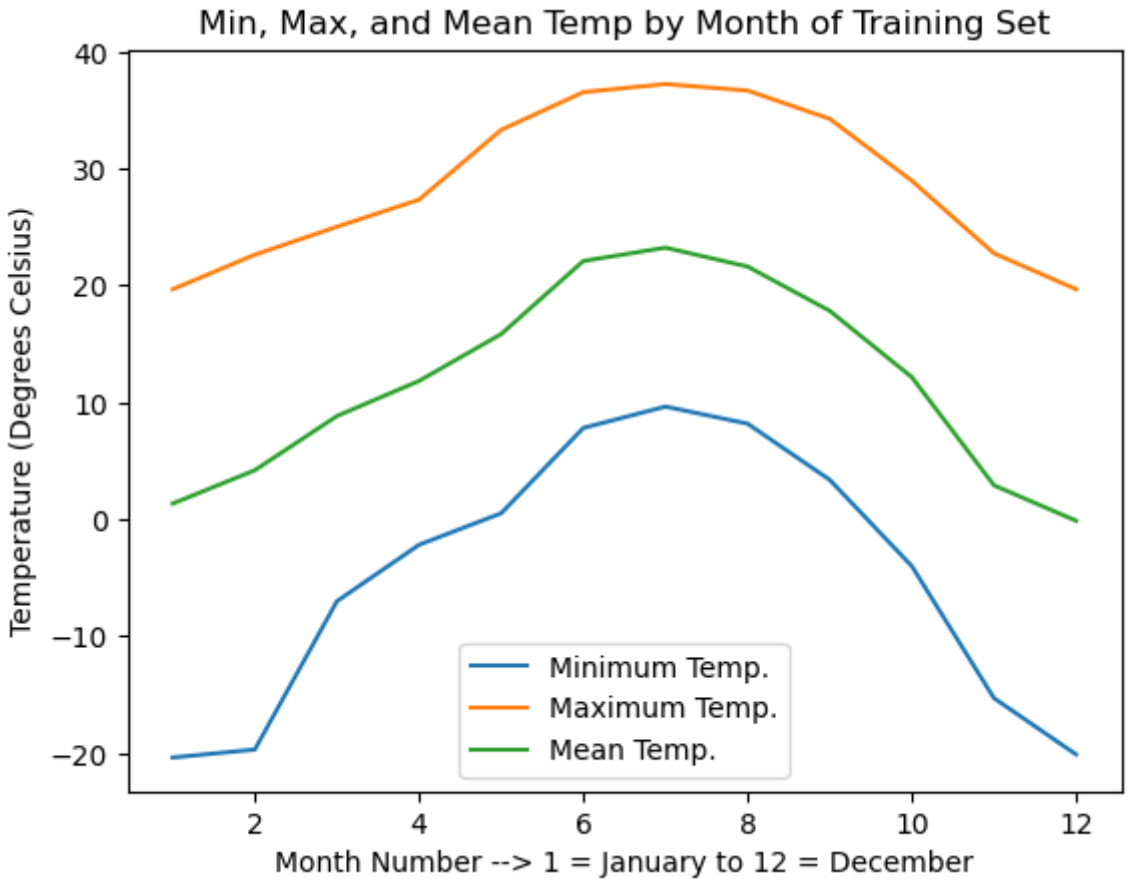
```
# Check
monthly_temp_df
```

Out[207]:

month_number	min	max	mean
1	-20.358963	19.687350	1.356244
2	-19.675203	22.640816	4.209260
3	-6.987536	25.033540	8.846492
4	-2.164256	27.350976	11.846632
5	0.532339	33.317022	15.853336
6	7.826261	36.546187	22.103172
7	9.654518	37.238782	23.245233
8	8.192403	36.679498	21.630298
9	3.393432	34.264141	17.844833
10	-3.988397	28.967995	12.167696
11	-15.278934	22.763140	2.922867
12	-20.096473	19.689836	-0.105181

In [214…

```
plt.figure()
x = monthly_temp_df.index
plt.plot(x,monthly_temp_df['min'], label = 'Minimum Temp.')
plt.plot(x,monthly_temp_df['max'], label = 'Maximum Temp.')
plt.plot(x,monthly_temp_df['mean'], label = 'Mean Temp.')
plt.title('Min, Max, and Mean Temp by Month of Training Set')
plt.xlabel('Month Number --> 1 = January to 12 = December')
plt.ylabel('Temperature (Degrees Celsius)')
plt.legend()
plt.show()
```



Nothing surprising here. Appears to be what one would see in a climate graph of the US. It would be interesting to do by climate region. We can save that for later, if timeworthy.

Run Pearson Correlation test

- We will get a coefficient on each parameter compared to target.
  - This will give us an idea of parameter influence.
- Plan on starting out with models were multicolinearity is not a problem
  - So, we will not concern ourselves with multicollinearity for now.

In [215…

```
# We will view the correlations through sorted output as there are
too many variables to see nicely
# Syntax found here-->
# https://www.kaggle.com/code/chrismat10/wind-turbine-sarima-
xgboost-randomforest-lstm
# Unstack correlations matrix
correlations =
training_data.corr().unstack().sort_values(ascending=False)

# Pull out just correlations with our target variable as a dataframe
# Drop duplicates of matrix
```

```
correlations_df = pd.DataFrame(correlations['contest-tmp2m-
14d__tmp2m'].drop_duplicates())
```

In [216…

```
# Display
pd.set_option('display.max_rows', 500)
correlations_df
```

Out[216]:

	0
contest-tmp2m-14d__tmp2m	1.000000
nmme-tmp2m-56w__cfsv2	0.954668
nmme-tmp2m-34w__cfsv2	0.954483
nmme-tmp2m-56w__nmmemean	0.950865
nmme-tmp2m-34w__nmmemean	0.950187
nmme-tmp2m-56w__gfdlflora	0.949846
nmme-tmp2m-34w__gfdlflorb	0.949346
nmme-tmp2m-56w__gfdlflorb	0.949016
nmme-tmp2m-34w__gfdlflora	0.947379
nmme-tmp2m-56w__ccsm4	0.944435
nmme-tmp2m-34w__ccsm4	0.943234
nmme-tmp2m-56w__nasa	0.941429
nmme-tmp2m-34w__nasa	0.940128
nmme-tmp2m-56w__gfdl	0.937179
nmme-tmp2m-34w__gfdl	0.936684
nmme-tmp2m-34w__cancm3	0.931658
nmme-tmp2m-56w__cancm3	0.930164
nmme-tmp2m-34w__cancm4	0.928570
nmme-tmp2m-56w__cancm4	0.927753
nmme-tmp2m-56w__ccsm3	0.908332
contest-wind-h100-14d__wind-hgt-100	0.898187
nmme-tmp2m-34w__ccsm3	0.890699
contest-wind-h500-14d__wind-hgt-500	0.884177
nmme0-tmp2m-34w__cfsv20	0.862264
nmme0-tmp2m-34w__nasa0	0.852478
nmme0-tmp2m-34w__nmme0mean	0.851105
nmme0-tmp2m-34w__gfdlflora0	0.848912
nmme0-tmp2m-34w__gfdlflorb0	0.848642
nmme0-tmp2m-34w__cancm30	0.834107
nmme0-tmp2m-34w__cancm40	0.830026
nmme0-tmp2m-34w__ccsm40	0.828936
nmme0-tmp2m-34w__gfdl0	0.824343
nmme0-tmp2m-34w__ccsm30	0.823158
contest-pevpr-sfc-gauss-14d__pevpr	0.805301
wind-uwnd-925-2010-1	0.798416

	0
wind-uwnd-250-2010-1	0.782773
contest-prwtr-eatm-14d__prwtr	0.772465
contest-wind-h10-14d__wind-hgt-10	0.763524
cfsv20	0.758914
nasa0	0.748423
gfdlflora0	0.746567
gfdlflorb0	0.746312
nmme0mean	0.745223
cancm30	0.729594
cancm40	0.728972
ccsm30	0.725942
ccsm40	0.718877
gfdl0	0.716059
sst-2010-1	0.586790
icec-2010-9	0.517935
contest-wind-h850-14d__wind-hgt-850	0.497195
sst-2010-3	0.438532
contest-wind-vwnd-250-14d__wind-vwnd-250	0.432445
season_number	0.427690
icec-2010-2	0.421824
wind-hgt-850-2010-7	0.381818
icec-2010-8	0.368287
wind-vwnd-250-2010-7	0.349429
wind-uwnd-925-2010-5	0.327369
icec-2010-10	0.321729
wind-uwnd-250-2010-18	0.307128
wind-hgt-100-2010-2	0.285676
wind-vwnd-925-2010-18	0.284506
sst-2010-5	0.283539
wind-vwnd-925-2010-15	0.279304
wind-hgt-500-2010-10	0.278332
wind-uwnd-925-2010-14	0.275827
wind-uwnd-925-2010-12	0.275036
contest-wind-vwnd-925-14d__wind-vwnd-925	0.273326
sst-2010-9	0.254034

	0
wind-hgt-500-2010-9	0.245759
Cfa	0.245350
contest-pres-sfc-gauss-14d__pres	0.244318
wind-hgt-100-2010-9	0.238118
icec-2010-4	0.221640
wind-vwnd-925-2010-16	0.213611
sst-2010-4	0.203734
mei__nip	0.197470
wind-hgt-10-2010-4	0.192229
wind-hgt-100-2010-7	0.180696
wind-hgt-10-2010-2	0.179335
wind-uwnd-250-2010-4	0.169979
wind-vwnd-250-2010-12	0.169653
BWh	0.168750
wind-hgt-10-2010-10	0.165748
wind-vwnd-250-2010-14	0.165271
wind-hgt-10-2010-5	0.162575
wind-uwnd-925-2010-9	0.154607
wind-vwnd-250-2010-10	0.139548
wind-uwnd-925-2010-20	0.129687
nmme0-prate-56w__cfsv20	0.121430
wind-uwnd-925-2010-10	0.118718
nmme0-prate-56w__nasa0	0.115602
wind-uwnd-925-2010-7	0.113698
wind-vwnd-250-2010-9	0.110772
BSh	0.107461
nmme0-prate-34w__cfsv20	0.098975
wind-vwnd-250-2010-19	0.094738
nmme0-prate-34w__nasa0	0.094668
lon	0.092923
wind-hgt-500-2010-4	0.092251
wind-hgt-500-2010-2	0.087268
nmme0-prate-56w__cancm30	0.085882
wind-vwnd-250-2010-11	0.083944
wind-uwnd-925-2010-4	0.083939

	0
mei__meirank	0.081722
wind-uwnd-250-2010-11	0.080493
wind-uwnd-250-2010-8	0.080478
BWk	0.079562
contest-precip-14d__precip	0.079041
wind-uwnd-925-2010-16	0.075444
wind-hgt-10-2010-9	0.074434
wind-uwnd-925-2010-15	0.072228
nmme0-prate-56w__cancm40	0.065023
nmme0-prate-56w__gfdl0	0.063657
wind-uwnd-925-2010-2	0.062602
Csa	0.062486
mei__mei	0.061639
wind-hgt-100-2010-6	0.056100
wind-vwnd-250-2010-18	0.053416
wind-vwnd-925-2010-5	0.052884
wind-vwnd-925-2010-13	0.052686
wind-hgt-10-2010-3	0.051787
month_number	0.050412
wind-uwnd-250-2010-13	0.043952
wind-vwnd-925-2010-8	0.043943
nmme0-prate-34w__cancm30	0.041223
wind-uwnd-925-2010-17	0.037641
wind-hgt-100-2010-3	0.025265
wind-vwnd-250-2010-8	0.022375
wind-vwnd-925-2010-17	0.022358
wind-uwnd-925-2010-3	0.018768
nmme-prate-56w__cfsv2	0.015907
nmme-prate-34w__cfsv2	0.014493
wind-vwnd-925-2010-14	0.011794
wind-hgt-10-2010-8	0.011326
nmme0-prate-34w__gfdl0	0.010788
wind-hgt-850-2010-4	0.009322
wind-uwnd-250-2010-6	0.006462
wind-hgt-100-2010-8	0.004020

	0
nmme0-prate-56w__nmme0mean	0.003193
wind-hgt-100-2010-5	0.002489
wind-vwnd-250-2010-15	0.002287
nmme-prate-34w__nasa	0.002206
wind-hgt-850-2010-8	0.000771
wind-uwnd-250-2010-14	-0.001884
nmme0-prate-34w__cancm40	-0.003955
nmme-prate-56w__nasa	-0.004842
wind-uwnd-925-2010-19	-0.005322
Cfb	-0.011304
Csb	-0.011888
wind-hgt-100-2010-4	-0.013188
Dwa	-0.013410
wind-vwnd-925-2010-11	-0.014704
Dwb	-0.017207
wind-uwnd-250-2010-10	-0.024601
wind-hgt-850-2010-10	-0.026426
wind-vwnd-250-2010-20	-0.033871
wind-vwnd-250-2010-6	-0.034476
wind-uwnd-250-2010-7	-0.034726
Dfa	-0.035016
wind-vwnd-925-2010-6	-0.036455
Dsc	-0.040514
wind-hgt-10-2010-6	-0.040606
Dsb	-0.042262
nmme0-prate-34w__nmme0mean	-0.049191
wind-hgt-850-2010-6	-0.053558
wind-vwnd-250-2010-16	-0.054980
wind-uwnd-250-2010-20	-0.056983
BSk	-0.058873
wind-uwnd-925-2010-18	-0.063945
wind-hgt-500-2010-7	-0.066340
nmme-prate-56w__gfdlflorb	-0.066928
nmme0-prate-56w__gfdlflora0	-0.067509
wind-vwnd-925-2010-7	-0.068578

	0
wind-uwnd-250-2010-2	-0.070978
nmme0-prate-56w__gfdlflorb0	-0.071111
wind-vwnd-925-2010-19	-0.072492
wind-uwnd-250-2010-12	-0.075881
wind-vwnd-925-2010-2	-0.076117
nmme-prate-56w__gfdlflora	-0.085523
nmme0-prate-56w__ccsm40	-0.086109
wind-uwnd-250-2010-9	-0.093002
wind-hgt-500-2010-6	-0.096707
nmme-prate-34w__gfdlflorb	-0.096816
wind-vwnd-250-2010-5	-0.104768
nmme-prate-34w__gfdlflora	-0.105720
wind-uwnd-925-2010-13	-0.107279
nmme0-prate-34w__gfdlflora0	-0.107452
nmme0-prate-34w__gfdlflorb0	-0.110909
wind-uwnd-250-2010-17	-0.113461
mjo1d__amplitude	-0.115210
wind-vwnd-925-2010-12	-0.118850
wind-hgt-500-2010-3	-0.121509
wind-vwnd-250-2010-2	-0.124070
wind-vwnd-925-2010-20	-0.127422
nmme0-prate-34w__ccsm40	-0.127728
mjo1d__phase	-0.130790
wind-vwnd-925-2010-4	-0.132691
wind-hgt-850-2010-2	-0.137594
Dfc	-0.139666
wind-vwnd-250-2010-4	-0.144695
nmme0-prate-56w__ccsm30	-0.147092
nmme-prate-56w__gfdl	-0.158220
wind-uwnd-250-2010-15	-0.160026
nmme-prate-56w__nmmemean	-0.163111
nmme-prate-34w__cancm3	-0.167068
wind-hgt-850-2010-9	-0.169601
nmme-prate-34w__nmmemean	-0.170152
wind-hgt-10-2010-7	-0.170252

	0
nmme-prate-34w__gfdl	-0.171759
nmme-prate-56w__cancm3	-0.174841
wind-vwnd-250-2010-3	-0.181221
sst-2010-10	-0.181708
sst-2010-6	-0.184491
wind-vwnd-250-2010-13	-0.184754
wind-vwnd-250-2010-17	-0.185391
wind-hgt-100-2010-10	-0.187325
wind-vwnd-925-2010-10	-0.197359
nmme-prate-34w__cancm4	-0.201921
Dfb	-0.203071
nmme-prate-56w__cancm4	-0.206909
elevation__elevation	-0.207810
wind-hgt-850-2010-5	-0.207869
wind-uwnd-250-2010-16	-0.211986
wind-hgt-500-2010-5	-0.212697
nmme0-prate-34w__ccsm30	-0.216063
icec-2010-3	-0.218507
wind-uwnd-250-2010-3	-0.219694
wind-vwnd-925-2010-9	-0.225632
wind-uwnd-925-2010-8	-0.235189
wind-vwnd-925-2010-3	-0.244246
wind-uwnd-250-2010-19	-0.249227
wind-hgt-850-2010-3	-0.253531
wind-uwnd-925-2010-6	-0.256242
nmme-prate-34w__ccsm4	-0.263720
nmme-prate-56w__ccsm4	-0.263803
sst-2010-8	-0.285873
nmme-prate-56w__ccsm3	-0.286250
icec-2010-6	-0.289349
nmme-prate-34w__ccsm3	-0.298928
contest-wind-uwnd-250-14d__wind-uwnd-250	-0.326666
contest-wind-uwnd-925-14d__wind-uwnd-925	-0.365888
wind-hgt-500-2010-8	-0.387288
index	-0.394012

	0
wind-uwnd-925-2010-11	-0.396562
lat	-0.398388
wind-uwnd-250-2010-5	-0.415874
icec-2010-7	-0.448767
sst-2010-2	-0.453613
icec-2010-1	-0.511502
sst-2010-7	-0.517767
contest-rhum-sig995-14d__rhum	-0.565127
icec-2010-5	-0.603640
contest-slp-14d__slp	-0.707640
wind-vwnd-250-2010-1	-0.731919
wind-hgt-850-2010-1	-0.779923
wind-hgt-100-2010-1	-0.802240
wind-hgt-500-2010-1	-0.806078
wind-vwnd-925-2010-1	-0.807371
wind-hgt-10-2010-1	-0.815701

In [222…

```
correlations_df.iloc[:,0]
```

Out[222]:

```
contest-tmp2m-14d__tmp2m                  1.000000
nmme-tmp2m-56w__cfsv2                     0.954668
nmme-tmp2m-34w__cfsv2                     0.954483
nmme-tmp2m-56w__nmmemean                  0.950865
nmme-tmp2m-34w__nmmemean                  0.950187
nmme-tmp2m-56w__gfdlflora                 0.949846
nmme-tmp2m-34w__gfdlflorb                 0.949346
nmme-tmp2m-56w__gfdlflorb                 0.949016
nmme-tmp2m-34w__gfdlflora                 0.947379
nmme-tmp2m-56w__ccsm4                     0.944435
nmme-tmp2m-34w__ccsm4                     0.943234
nmme-tmp2m-56w__nasa                      0.941429
nmme-tmp2m-34w__nasa                      0.940128
nmme-tmp2m-56w__gfdl                      0.937179
nmme-tmp2m-34w__gfdl                      0.936684
nmme-tmp2m-34w__cancm3                    0.931658
nmme-tmp2m-56w__cancm3                    0.930164
nmme-tmp2m-34w__cancm4                    0.928570
nmme-tmp2m-56w__cancm4                    0.927753
nmme-tmp2m-56w__ccsm3                     0.908332
contest-wind-h100-14d__wind-hgt-100       0.898187
nmme-tmp2m-34w__ccsm3                     0.890699
contest-wind-h500-14d__wind-hgt-500       0.884177
nmme0-tmp2m-34w__cfsv20                   0.862264
nmme0-tmp2m-34w__nasa0                    0.852478
nmme0-tmp2m-34w__nmme0mean                0.851105
nmme0-tmp2m-34w__gfdlflora0               0.848912
nmme0-tmp2m-34w__gfdlflorb0               0.848642
nmme0-tmp2m-34w__cancm30                  0.834107
nmme0-tmp2m-34w__cancm40                  0.830026
nmme0-tmp2m-34w__ccsm40                   0.828936
nmme0-tmp2m-34w__gfdl0                    0.824343
nmme0-tmp2m-34w__ccsm30                   0.823158
contest-pevpr-sfc-gauss-14d__pevpr        0.805301
wind-uwnd-925-2010-1                      0.798416
wind-uwnd-250-2010-1                      0.782773
contest-prwtr-eatm-14d__prwtr             0.772465
contest-wind-h10-14d__wind-hgt-10         0.763524
cfsv20                                    0.758914
nasa0                                     0.748423
gfdlflora0                                0.746567
gfdlflorb0                                0.746312
nmme0mean                                 0.745223
cancm30                                   0.729594
cancm40                                   0.728972
ccsm30                                    0.725942
ccsm40                                    0.718877
gfdl0                                     0.716059
sst-2010-1                                0.586790
icec-2010-9                               0.517935
contest-wind-h850-14d__wind-hgt-850       0.497195
sst-2010-3                                0.438532
```

```
contest-wind-vwnd-250-14d__wind-vwnd-250    0.432445
season_number                               0.427690
icec-2010-2                                 0.421824
wind-hgt-850-2010-7                         0.381818
icec-2010-8                                 0.368287
wind-vwnd-250-2010-7                        0.349429
wind-uwnd-925-2010-5                        0.327369
icec-2010-10                                0.321729
wind-uwnd-250-2010-18                       0.307128
wind-hgt-100-2010-2                         0.285676
wind-vwnd-925-2010-18                       0.284506
sst-2010-5                                  0.283539
wind-vwnd-925-2010-15                       0.279304
wind-hgt-500-2010-10                        0.278332
wind-uwnd-925-2010-14                       0.275827
wind-uwnd-925-2010-12                       0.275036
contest-wind-vwnd-925-14d__wind-vwnd-925    0.273326
sst-2010-9                                  0.254034
wind-hgt-500-2010-9                         0.245759
Cfa                                         0.245350
contest-pres-sfc-gauss-14d__pres            0.244318
wind-hgt-100-2010-9                         0.238118
icec-2010-4                                 0.221640
wind-vwnd-925-2010-16                       0.213611
sst-2010-4                                  0.203734
mei__nip                                    0.197470
wind-hgt-10-2010-4                          0.192229
wind-hgt-100-2010-7                         0.180696
wind-hgt-10-2010-2                          0.179335
wind-uwnd-250-2010-4                        0.169979
wind-vwnd-250-2010-12                       0.169653
BWh                                         0.168750
wind-hgt-10-2010-10                         0.165748
wind-vwnd-250-2010-14                       0.165271
wind-hgt-10-2010-5                          0.162575
wind-uwnd-925-2010-9                        0.154607
wind-vwnd-250-2010-10                       0.139548
wind-uwnd-925-2010-20                       0.129687
nmme0-prate-56w__cfsv20                     0.121430
wind-uwnd-925-2010-10                       0.118718
nmme0-prate-56w__nasa0                      0.115602
wind-uwnd-925-2010-7                        0.113698
wind-vwnd-250-2010-9                        0.110772
BSh                                         0.107461
nmme0-prate-34w__cfsv20                     0.098975
wind-vwnd-250-2010-19                       0.094738
nmme0-prate-34w__nasa0                      0.094668
lon                                         0.092923
wind-hgt-500-2010-4                         0.092251
wind-hgt-500-2010-2                         0.087268
nmme0-prate-56w__cancm30                    0.085882
wind-vwnd-250-2010-11                       0.083944
```

```
wind-uwnd-925-2010-4                  0.083939
mei__meirank                          0.081722
wind-uwnd-250-2010-11                 0.080493
wind-uwnd-250-2010-8                  0.080478
BWk                                   0.079562
contest-precip-14d__precip            0.079041
wind-uwnd-925-2010-16                 0.075444
wind-hgt-10-2010-9                    0.074434
wind-uwnd-925-2010-15                 0.072228
nmme0-prate-56w__cancm40              0.065023
nmme0-prate-56w__gfdl0                0.063657
wind-uwnd-925-2010-2                  0.062602
Csa                                   0.062486
mei__mei                              0.061639
wind-hgt-100-2010-6                   0.056100
wind-vwnd-250-2010-18                 0.053416
wind-vwnd-925-2010-5                  0.052884
wind-vwnd-925-2010-13                 0.052686
wind-hgt-10-2010-3                    0.051787
month_number                          0.050412
wind-uwnd-250-2010-13                 0.043952
wind-vwnd-925-2010-8                  0.043943
nmme0-prate-34w__cancm30              0.041223
wind-uwnd-925-2010-17                 0.037641
wind-hgt-100-2010-3                   0.025265
wind-vwnd-250-2010-8                  0.022375
wind-vwnd-925-2010-17                 0.022358
wind-uwnd-925-2010-3                  0.018768
nmme-prate-56w__cfsv2                 0.015907
nmme-prate-34w__cfsv2                 0.014493
wind-vwnd-925-2010-14                 0.011794
wind-hgt-10-2010-8                    0.011326
nmme0-prate-34w__gfdl0                0.010788
wind-hgt-850-2010-4                   0.009322
wind-uwnd-250-2010-6                  0.006462
wind-hgt-100-2010-8                   0.004020
nmme0-prate-56w__nmme0mean            0.003193
wind-hgt-100-2010-5                   0.002489
wind-vwnd-250-2010-15                 0.002287
nmme-prate-34w__nasa                  0.002206
wind-hgt-850-2010-8                   0.000771
wind-uwnd-250-2010-14                -0.001884
nmme0-prate-34w__cancm40             -0.003955
nmme-prate-56w__nasa                 -0.004842
wind-uwnd-925-2010-19                -0.005322
Cfb                                  -0.011304
Csb                                  -0.011888
wind-hgt-100-2010-4                  -0.013188
Dwa                                  -0.013410
wind-vwnd-925-2010-11                -0.014704
Dwb                                  -0.017207
wind-uwnd-250-2010-10                -0.024601
```

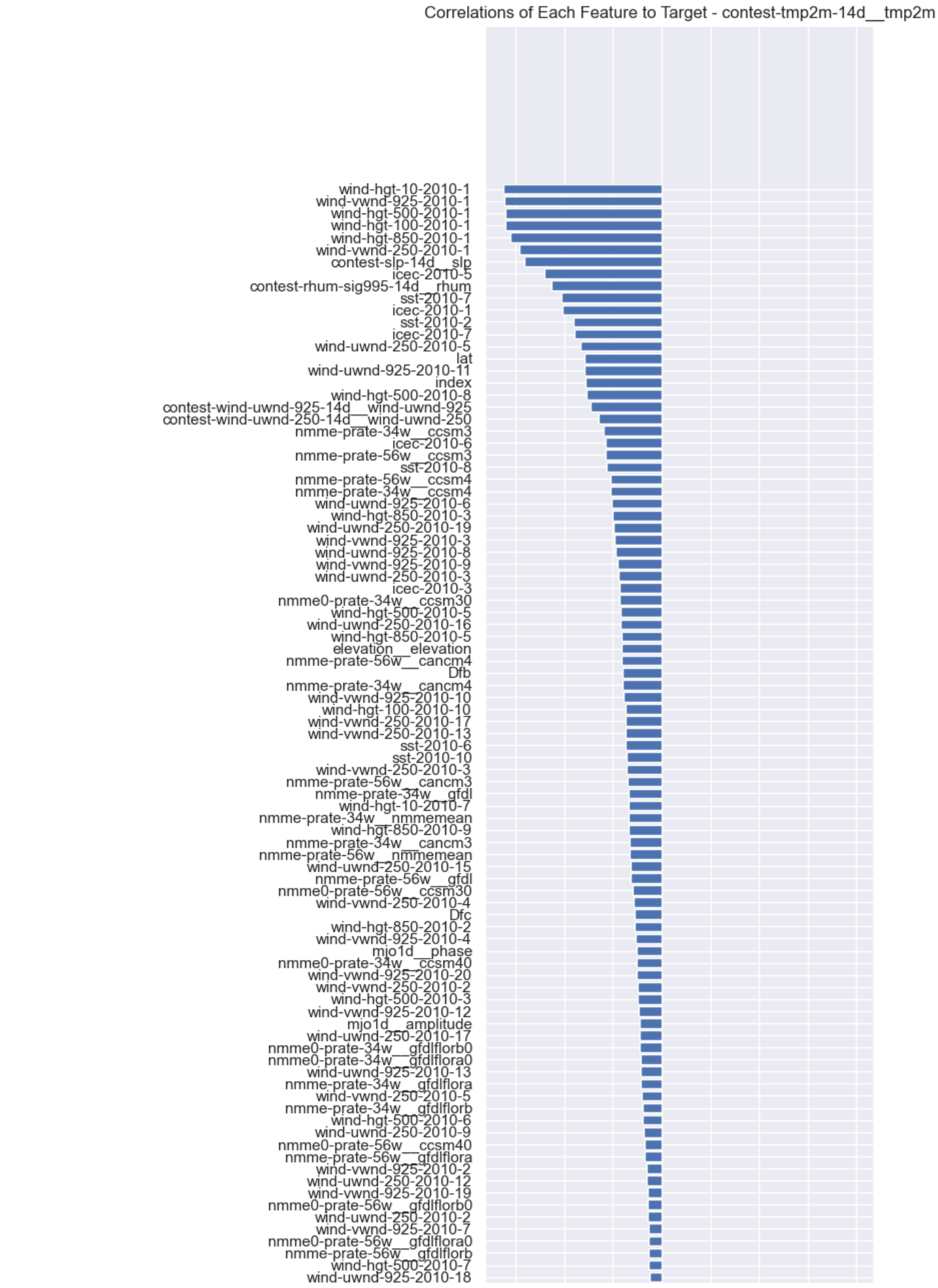
```
wind-hgt-850-2010-10                -0.026426
wind-vwnd-250-2010-20               -0.033871
wind-vwnd-250-2010-6                -0.034476
wind-uwnd-250-2010-7                -0.034726
Dfa                                 -0.035016
wind-vwnd-925-2010-6                -0.036455
Dsc                                 -0.040514
wind-hgt-10-2010-6                  -0.040606
Dsb                                 -0.042262
nmme0-prate-34w__nmme0mean          -0.049191
wind-hgt-850-2010-6                 -0.053558
wind-vwnd-250-2010-16               -0.054980
wind-uwnd-250-2010-20               -0.056983
BSk                                 -0.058873
wind-uwnd-925-2010-18               -0.063945
wind-hgt-500-2010-7                 -0.066340
nmme-prate-56w__gfdlflorb           -0.066928
nmme0-prate-56w__gfdlflora0         -0.067509
wind-vwnd-925-2010-7                -0.068578
wind-uwnd-250-2010-2                -0.070978
nmme0-prate-56w__gfdlflorb0         -0.071111
wind-vwnd-925-2010-19               -0.072492
wind-uwnd-250-2010-12               -0.075881
wind-vwnd-925-2010-2                -0.076117
nmme-prate-56w__gfdlflora           -0.085523
nmme0-prate-56w__ccsm40             -0.086109
wind-uwnd-250-2010-9                -0.093002
wind-hgt-500-2010-6                 -0.096707
nmme-prate-34w__gfdlflorb           -0.096816
wind-vwnd-250-2010-5                -0.104768
nmme-prate-34w__gfdlflora           -0.105720
wind-uwnd-925-2010-13               -0.107279
nmme0-prate-34w__gfdlflora0         -0.107452
nmme0-prate-34w__gfdlflorb0         -0.110909
wind-uwnd-250-2010-17               -0.113461
mjo1d__amplitude                    -0.115210
wind-vwnd-925-2010-12               -0.118850
wind-hgt-500-2010-3                 -0.121509
wind-vwnd-250-2010-2                -0.124070
wind-vwnd-925-2010-20               -0.127422
nmme0-prate-34w__ccsm40             -0.127728
mjo1d__phase                        -0.130790
wind-vwnd-925-2010-4                -0.132691
wind-hgt-850-2010-2                 -0.137594
Dfc                                 -0.139666
wind-vwnd-250-2010-4                -0.144695
nmme0-prate-56w__ccsm30             -0.147092
nmme-prate-56w__gfdl                -0.158220
wind-uwnd-250-2010-15               -0.160026
nmme-prate-56w__nmmemean            -0.163111
nmme-prate-34w__cancm3              -0.167068
wind-hgt-850-2010-9                 -0.169601
```

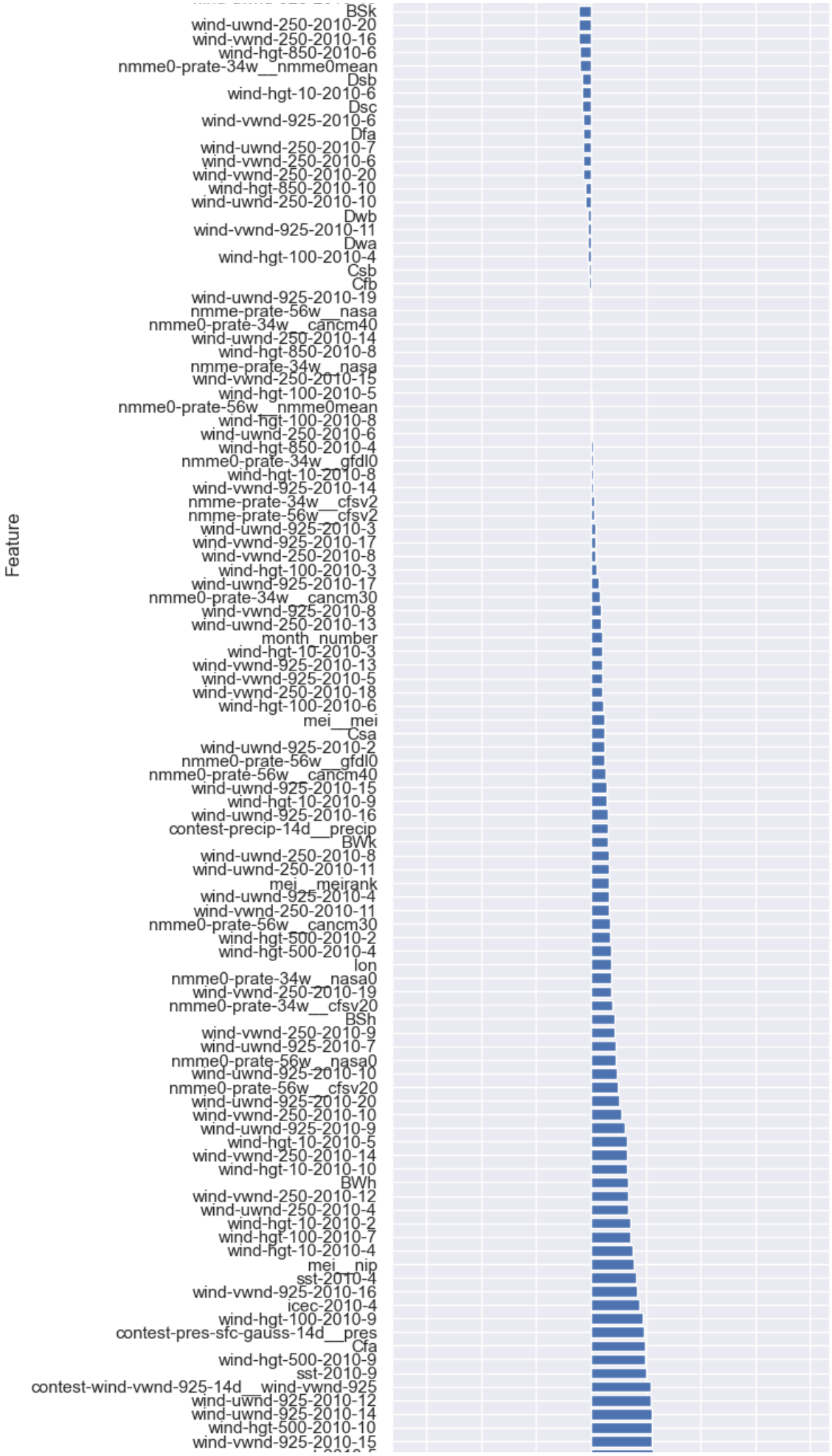
```
nmme-prate-34w__nmmemean                    -0.170152
wind-hgt-10-2010-7                          -0.170252
nmme-prate-34w__gfdl                        -0.171759
nmme-prate-56w__cancm3                      -0.174841
wind-vwnd-250-2010-3                        -0.181221
sst-2010-10                                 -0.181708
sst-2010-6                                  -0.184491
wind-vwnd-250-2010-13                       -0.184754
wind-vwnd-250-2010-17                       -0.185391
wind-hgt-100-2010-10                        -0.187325
wind-vwnd-925-2010-10                       -0.197359
nmme-prate-34w__cancm4                      -0.201921
Dfb                                         -0.203071
nmme-prate-56w__cancm4                      -0.206909
elevation__elevation                        -0.207810
wind-hgt-850-2010-5                         -0.207869
wind-uwnd-250-2010-16                       -0.211986
wind-hgt-500-2010-5                         -0.212697
nmme0-prate-34w__ccsm30                     -0.216063
icec-2010-3                                 -0.218507
wind-uwnd-250-2010-3                        -0.219694
wind-vwnd-925-2010-9                        -0.225632
wind-uwnd-925-2010-8                        -0.235189
wind-vwnd-925-2010-3                        -0.244246
wind-uwnd-250-2010-19                       -0.249227
wind-hgt-850-2010-3                         -0.253531
wind-uwnd-925-2010-6                        -0.256242
nmme-prate-34w__ccsm4                       -0.263720
nmme-prate-56w__ccsm4                       -0.263803
sst-2010-8                                  -0.285873
nmme-prate-56w__ccsm3                       -0.286250
icec-2010-6                                 -0.289349
nmme-prate-34w__ccsm3                       -0.298928
contest-wind-uwnd-250-14d__wind-uwnd-250    -0.326666
contest-wind-uwnd-925-14d__wind-uwnd-925    -0.365888
wind-hgt-500-2010-8                         -0.387288
index                                       -0.394012
wind-uwnd-925-2010-11                       -0.396562
lat                                         -0.398388
wind-uwnd-250-2010-5                        -0.415874
icec-2010-7                                 -0.448767
sst-2010-2                                  -0.453613
icec-2010-1                                 -0.511502
sst-2010-7                                  -0.517767
contest-rhum-sig995-14d__rhum               -0.565127
icec-2010-5                                 -0.603640
contest-slp-14d__slp                        -0.707640
wind-vwnd-250-2010-1                        -0.731919
wind-hgt-850-2010-1                         -0.779923
wind-hgt-100-2010-1                         -0.802240
wind-hgt-500-2010-1                         -0.806078
wind-vwnd-925-2010-1                        -0.807371
```

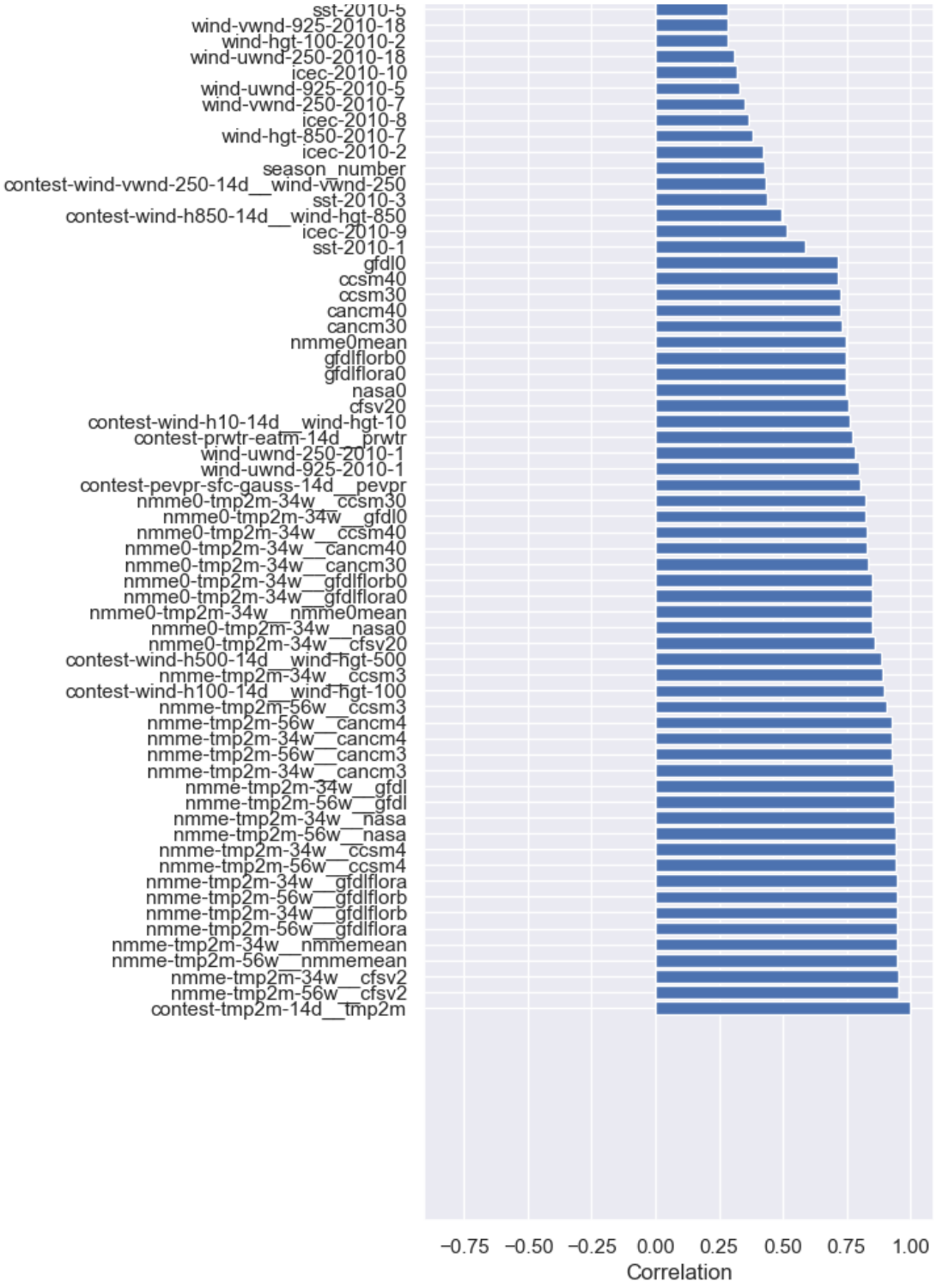
```
wind-hgt-10-2010-1                          -0.815701
Name: 0, dtype: float64
```

In [230…

```
#Visualize with horizontal bar plot
sns.set(rc={'figure.figsize':(5,45)})
y = correlations_df.index
x = correlations_df.iloc[:,0]
plt.barh(y, x)
plt.title('Correlations of Each Feature to Target - contest-tmp2m-14d__tmp2m')
plt.ylabel('Feature')
plt.xlabel('Correlation')
plt.show()
```


Correlations of Each Feature to Target - contest-tmp2m-14d__tmp2m
wind-hgt-10-2010-1
wind-vwnd-925-2010-1
wind-hgt-500-2010-1
wind-hgt-100-2010-1
wind-hgt-850-2010-1
wind-vwnd-250-2010-1
contest-slp-14d__slp
icec-2010-5
contest-rhum-sig995-14d__rhum
sst-2010-7
icec-2010-1
sst-2010-2
icec-2010-7
wind-uwnd-250-2010-5
lat
wind-uwnd-925-2010-11
index
wind-hgt-500-2010-8
contest-wind-uwnd-925-14d__wind-uwnd-925
contest-wind-uwnd-250-14d__wind-uwnd-250
nmme-prate-34w__ccsm3
icec-2010-6
nmme-prate-56w__ccsm3
sst-2010-8
nmme-prate-56w__ccsm4
nmme-prate-34w__ccsm4
wind-uwnd-925-2010-6
wind-hgt-850-2010-3
wind-uwnd-250-2010-19
wind-vwnd-925-2010-3
wind-uwnd-925-2010-8
wind-vwnd-925-2010-9
wind-uwnd-250-2010-3
icec-2010-3
nmme0-prate-34w__ccsm30
wind-hgt-500-2010-5
wind-uwnd-250-2010-16
wind-hgt-850-2010-5
elevation__elevation
nmme-prate-56w__cancm4
Dfb
nmme-prate-34w__cancm4
wind-vwnd-925-2010-10
wind-hgt-100-2010-10
wind-vwnd-250-2010-17
wind-vwnd-250-2010-13
sst-2010-6
sst-2010-10
wind-vwnd-250-2010-3
nmme-prate-56w__cancm3
nmme-prate-34w__gfdl
wind-hgt-10-2010-7
nmme-prate-34w__nmmemean
wind-hgt-850-2010-9
nmme-prate-34w__cancm3
nmme-prate-56w__nmmemean
wind-uwnd-250-2010-15
nmme-prate-56w__gfdl
nmme0-prate-56w__ccsm30
wind-vwnd-250-2010-4
Dfc
wind-hgt-850-2010-2
wind-vwnd-925-2010-4
mjo1d__phase
nmme0-prate-34w__ccsm40
wind-vwnd-925-2010-20
wind-vwnd-250-2010-2
wind-hgt-500-2010-3
wind-vwnd-925-2010-12
mjo1d__amplitude
wind-uwnd-250-2010-17
nmme0-prate-34w__gfdlflorb0
nmme0-prate-34w__gfdlflora0
wind-uwnd-925-2010-13
nmme-prate-34w__gfdlflora
wind-vwnd-250-2010-5
nmme-prate-34w__gfdlflorb
wind-hgt-500-2010-6
wind-uwnd-250-2010-9
nmme0-prate-56w__ccsm40
nmme-prate-56w__gfdlflora
wind-vwnd-925-2010-2
wind-uwnd-250-2010-12
wind-vwnd-925-2010-19
nmme0-prate-56w__gfdlflorb0
wind-uwnd-250-2010-2
wind-vwnd-925-2010-7
nmme0-prate-56w__gfdlflora0
nmme-prate-56w__gfdlflorb
wind-hgt-500-2010-7
wind-uwnd-925-2010-18


Feature
BSk
wind-uwnd-250-2010-20
wind-vwnd-250-2010-16
wind-hgt-850-2010-6
nmme0-prate-34w__nmme0mean
Dsb
wind-hgt-10-2010-6
Dsc
wind-vwnd-925-2010-6
Dfa
wind-uwnd-250-2010-7
wind-vwnd-250-2010-6
wind-vwnd-250-2010-20
wind-hgt-850-2010-10
wind-uwnd-250-2010-10
Dwb
wind-vwnd-925-2010-11
Dwa
wind-hgt-100-2010-4
Csb
Cfb
wind-uwnd-925-2010-19
nmme-prate-56w__nasa
nmme0-prate-34w__cancm40
wind-uwnd-250-2010-14
wind-hgt-850-2010-8
nmme-prate-34w__nasa
wind-vwnd-250-2010-15
wind-hgt-100-2010-5
nmme0-prate-56w__nmme0mean
wind-hgt-100-2010-8
wind-uwnd-250-2010-6
wind-hgt-850-2010-4
nmme0-prate-34w__gfdl0
wind-hgt-10-2010-8
wind-vwnd-925-2010-14
nmme-prate-34w__cfsv2
nmme-prate-56w__cfsv2
wind-uwnd-925-2010-3
wind-vwnd-925-2010-17
wind-vwnd-250-2010-8
wind-hgt-100-2010-3
wind-uwnd-925-2010-17
nmme0-prate-34w__cancm30
wind-vwnd-925-2010-8
wind-uwnd-250-2010-13
month_number
wind-hgt-10-2010-3
wind-vwnd-925-2010-13
wind-vwnd-925-2010-5
wind-vwnd-250-2010-18
wind-hgt-100-2010-6
mei__mei
Csa
wind-uwnd-925-2010-2
nmme0-prate-56w__gfdl0
nmme0-prate-56w__cancm40
wind-uwnd-925-2010-15
wind-hgt-10-2010-9
wind-uwnd-925-2010-16
contest-precip-14d__precip
BWk
wind-uwnd-250-2010-8
wind-uwnd-250-2010-11
mei__meirank
wind-uwnd-925-2010-4
wind-vwnd-250-2010-11
nmme0-prate-56w__cancm30
wind-hgt-500-2010-2
wind-hgt-500-2010-4
lon
nmme0-prate-34w__nasa0
wind-vwnd-250-2010-19
nmme0-prate-34w__cfsv20
BSh
wind-vwnd-250-2010-9
wind-uwnd-925-2010-7
nmme0-prate-56w__nasa0
wind-uwnd-925-2010-10
nmme0-prate-56w__cfsv20
wind-uwnd-925-2010-20
wind-vwnd-250-2010-10
wind-uwnd-925-2010-9
wind-hgt-10-2010-5
wind-vwnd-250-2010-14
wind-hgt-10-2010-10
BWh
wind-vwnd-250-2010-12
wind-uwnd-250-2010-4
wind-hgt-10-2010-2
wind-hgt-100-2010-7
wind-hgt-10-2010-4
mei__nip
sst-2010-4
wind-vwnd-925-2010-16
icec-2010-4
wind-hgt-100-2010-9
contest-pres-sfc-gauss-14d__pres
Cfa
wind-hgt-500-2010-9
sst-2010-9
contest-wind-vwnd-925-14d__wind-vwnd-925
wind-uwnd-925-2010-12
wind-uwnd-925-2010-14
wind-hgt-500-2010-10
wind-vwnd-925-2010-15



To properly breakdown each of these features, we will need to spend some time with the data dictionary. **Correlations by Feature Category**

- North American Multi-Model Ensemble (NMME): The North American Multi-Model Ensemble (NMME) is a collection of physics-based forecast models from various modeling centers in North America. Forecasts issued monthly from the Cansips, CanCM3, CanCM4, CCSM3, CCSM4, GFDL-CM2.1-aer04, GFDL-CM2.5, FLOR-A06 and FLOR-B01, NASA-GMAO-062012, and NCEP-CFSv2 models were downloaded from the IRI/LDEO Climate Data Library. Each forecast contains monthly mean predictions from 0.5 to 8.5 months ahead. https://iridl.ldeo.columbia.edu/SOURCES/.Models/.NMME/

- Weeks 3-4 weighted average of most recent monthly NMME model forecasts for target label(Temperature)
- North American Multi-Model Ensemble (NMME): The North American Multi-Model Ensemble (NMME) is a collection of physics-based forecast models from various modeling centers in North America. :
  - nmme-tmp2m-56w__cfsv2 0.954668
  - nmme-tmp2m-34w__cfsv2 0.954483
  - nmme-tmp2m-56w__nmmemean 0.950865
  - nmme-tmp2m-34w__nmmemean 0.950187
  - nmme-tmp2m-56w__gfdlflora 0.949846
  - nmme-tmp2m-34w__gfdlflorb 0.949346
  - nmme-tmp2m-56w__gfdlflorb 0.949016
  - nmme-tmp2m-34w__gfdlflora 0.947379
  - nmme-tmp2m-56w__ccsm4 0.944435
  - nmme-tmp2m-34w__ccsm4 0.943234
  - nmme-tmp2m-56w__nasa 0.941429
  - nmme-tmp2m-34w__nasa 0.940128
  - nmme-tmp2m-56w__gfdl 0.937179
  - nmme-tmp2m-34w__gfdl 0.936684
  - nmme-tmp2m-34w__cancm3 0.931658
  - nmme-tmp2m-56w__cancm3 0.930164
  - nmme-tmp2m-34w__cancm4 0.928570
  - nmme-tmp2m-56w__cancm4 0.927753
  - nmme-tmp2m-56w__ccsm3 0.908332
  - nmme-tmp2m-34w__ccsm3 0.890699
- Geopotential height, zonal wind, and longitudinal wind: To capture polar vortex variability, obtained daily mean geopotential height were obtained at 10mb from the NCEP Reanalysis dataset.
  ftp://ftp.cdc.noaa.gov/Datasets/ncep.reanalysis.dailyavgs/pressure/
- geopotential height at 100 millibars and 500 millibars and 10 millibars and 850 millibars (Windspeed??)
  - contest-wind-h100-14d__wind-hgt-100 0.898187
  - contest-wind-h500-14d__wind-hgt-500 0.884177
  - contest-wind-h10-14d__wind-hgt-10 0.763524
  - contest-wind-h850-14d__wind-hgt-850 0.497195
  - wind-hgt-850-2010-7 0.381818
  - wind-hgt-100-2010-2 0.285676
  - wind-hgt-500-2010-10 0.278332
  - wind-hgt-500-2010-9 0.245759
  - wind-hgt-100-2010-9 0.238118
  - wind-hgt-10-2010-4 0.192229
  - wind-hgt-100-2010-7 0.180696
  - wind-hgt-10-2010-2 0.179335

    - wind-hgt-10-2010-10 0.165748
    - wind-hgt-10-2010-5 0.162575
    - wind-hgt-500-2010-4 0.092251
    - wind-hgt-500-2010-2 0.087268
    - wind-hgt-100-2010-6 0.056100
    - wind-hgt-10-2010-3 0.051787
    - wind-hgt-100-2010-3 0.025265
    - wind-hgt-10-2010-8 0.011326
    - wind-hgt-850-2010-4 0.009322
    - wind-hgt-100-2010-8 0.004020
    - wind-hgt-100-2010-5 0.002489
    - wind-hgt-850-2010-8 0.000771
    - wind-hgt-100-2010-4 -0.013188
    - wind-hgt-850-2010-10 -0.026426
    - wind-hgt-10-2010-6 -0.040606
    - wind-hgt-850-2010-6 -0.053558
    - wind-hgt-500-2010-7 -0.066340
    - wind-hgt-500-2010-6 -0.096707
    - wind-hgt-500-2010-3 -0.121509
    - wind-hgt-850-2010-2 -0.137594
    - wind-hgt-850-2010-9 -0.169601
    - wind-hgt-10-2010-7 -0.170252
    - wind-hgt-100-2010-10 -0.187325
    - wind-hgt-850-2010-5 -0.207869
    - wind-hgt-500-2010-5 -0.212697
    - wind-hgt-850-2010-3 -0.253531
    - wind-hgt-500-2010-8 -0.387288
    - wind-hgt-850-2010-1 -0.779923
    - wind-hgt-100-2010-1 -0.802240
    - wind-hgt-500-2010-1 -0.806078
    - wind-hgt-10-2010-1 -0.815701
- most recent monthly NMME model forecasts for tmp2m (temperature)
    - nmme0-tmp2m-34w__cfsv20 0.862264
    - nmme0-tmp2m-34w__nasa0 0.852478
    - nmme0-tmp2m-34w__nmme0mean 0.851105
    - nmme0-tmp2m-34w__gfdlflora0 0.848912
    - nmme0-tmp2m-34w__gfdlflorb0 0.848642
    - nmme0-tmp2m-34w__cancm30 0.834107
    - nmme0-tmp2m-34w__cancm40 0.830026
    - nmme0-tmp2m-34w__ccsm40 0.828936
    - nmme0-tmp2m-34w__gfdl0 0.824343
    - nmme0-tmp2m-34w__ccsm30 0.823158
- potential evaporation

    - contest-pevpr-sfc-gauss-14d__pevpr 0.805301
- zonal wind at 925 millibars and 250 millibars (Windspeed??)
    - wind-uwnd-925-2010-1 0.798416
    - wind-uwnd-250-2010-1 0.782773
    - wind-uwnd-925-2010-5 0.327369
    - wind-uwnd-250-2010-18 0.307128
    - wind-uwnd-925-2010-14 0.275827
    - wind-uwnd-925-2010-12 0.275036
    - wind-uwnd-250-2010-4 0.169979
    - wind-uwnd-925-2010-9 0.154607
    - wind-uwnd-925-2010-20 0.129687
    - wind-uwnd-925-2010-10 0.118718
    - wind-uwnd-925-2010-7 0.113698
    - wind-uwnd-925-2010-4 0.083939
    - wind-uwnd-250-2010-11 0.080493
    - wind-uwnd-250-2010-8 0.080478
    - wind-uwnd-925-2010-16 0.075444
    - wind-uwnd-925-2010-15 0.072228
    - wind-uwnd-925-2010-2 0.062602
    - wind-uwnd-250-2010-13 0.043952
    - wind-uwnd-925-2010-17 0.037641
    - wind-uwnd-925-2010-3 0.018768
    - wind-uwnd-250-2010-6 0.006462
    - wind-uwnd-250-2010-14 -0.001884
    - wind-uwnd-925-2010-19 -0.005322
    - wind-uwnd-250-2010-10 -0.024601
    - wind-uwnd-250-2010-7 -0.034726
    - wind-uwnd-250-2010-20 -0.056983
    - wind-uwnd-925-2010-18 -0.063945
    - wind-uwnd-250-2010-2 -0.070978
    - wind-uwnd-250-2010-12 -0.075881
    - wind-uwnd-250-2010-9 -0.093002
    - wind-uwnd-925-2010-13 -0.107279
    - wind-uwnd-250-2010-17 -0.113461
    - wind-uwnd-250-2010-15 -0.160026
    - wind-uwnd-250-2010-16 -0.211986
    - wind-uwnd-250-2010-3 -0.219694
    - wind-uwnd-925-2010-8 -0.235189
    - wind-uwnd-250-2010-19 -0.249227
    - wind-uwnd-925-2010-6 -0.256242
    - contest-wind-uwnd-250-14d__wind-uwnd-250 -0.326666
    - contest-wind-uwnd-925-14d__wind-uwnd-925 -0.365888
    - wind-uwnd-925-2010-11 -0.396562

    - wind-uwnd-250-2010-5 -0.415874
    -
- precipitable water for entire atmosphere
    - contest-prwtr-eatm-14d__prwtr 0.772465
- most recent forecasts from weather models
    - cfsv20 0.758914
    - nasa0 0.748423
    - gfdlflora0 0.746567
    - gfdlflorb0 0.746312
    - nmme0mean 0.745223
    - cancm30 0.729594
    - cancm40 0.728972
    - ccsm30 0.725942
    - ccsm40 0.718877
    - gfdl0 0.716059
- Sea surface temperature and sea ice concentration: NOAA's Optimum Interpolation Sea Surface Temperature (SST) dataset provides SST and sea ice concentration data, daily from 1981 to the present. ftp://ftp.cdc.noaa.gov/Projects/Datasets/noaa.oisst.v2.highres/
- Sea surface temperature
    - sst-2010-1 0.586790
    - sst-2010-3 0.438532
    - sst-2010-5 0.283539
    - sst-2010-9 0.254034
    - sst-2010-4 0.203734
    - sst-2010-10 -0.181708
    - sst-2010-6 -0.184491
    - sst-2010-8 -0.285873
    - sst-2010-2 -0.453613
    - sst-2010-7 -0.517767
- Sea ice concentration
    - icec-2010-9 0.517935
    - icec-2010-2 0.421824
    - icec-2010-8 0.368287
    - icec-2010-10 0.321729
    - icec-2010-4 0.221640
    - icec-2010-3 -0.218507
    - icec-2010-6 -0.289349
    - icec-2010-7 -0.448767
    - icec-2010-1 -0.511502
    - icec-2010-5 -0.603640
- longitudinal wind at 250 millibars (wind speed?)
    - contest-wind-vwnd-250-14d__wind-vwnd-250 0.432445
    - wind-vwnd-250-2010-7 0.349429

    - wind-vwnd-925-2010-18 0.284506
    - wind-vwnd-925-2010-15 0.279304
    - contest-wind-vwnd-925-14d__wind-vwnd-925 0.273326
    - wind-vwnd-925-2010-16 0.213611
    - wind-vwnd-250-2010-12 0.169653
    - wind-vwnd-250-2010-14 0.165271
    - wind-vwnd-250-2010-10 0.139548
    - wind-vwnd-250-2010-9 0.110772
    - wind-vwnd-250-2010-19 0.094738
    - wind-vwnd-250-2010-11 0.083944
    - wind-vwnd-250-2010-18 0.053416
    - wind-vwnd-925-2010-5 0.052884
    - wind-vwnd-925-2010-13 0.052686
    - wind-vwnd-925-2010-8 0.043943
    - wind-vwnd-250-2010-8 0.022375
    - wind-vwnd-925-2010-17 0.022358
    - wind-vwnd-925-2010-14 0.011794
    - wind-vwnd-250-2010-15 0.002287
    - wind-vwnd-925-2010-11 -0.014704
    - wind-vwnd-250-2010-20 -0.033871
    - wind-vwnd-250-2010-6 -0.034476
    - wind-vwnd-925-2010-6 -0.036455
    - wind-vwnd-250-2010-16 -0.054980
    - wind-vwnd-925-2010-7 -0.068578
    - wind-vwnd-925-2010-19 -0.072492
    - wind-vwnd-925-2010-2 -0.076117
    - wind-vwnd-250-2010-5 -0.104768
    - wind-vwnd-925-2010-12 -0.118850
    - wind-vwnd-250-2010-2 -0.124070
    - wind-vwnd-925-2010-20 -0.127422
    - wind-vwnd-925-2010-4 -0.132691
    - wind-vwnd-250-2010-4 -0.144695
    - wind-vwnd-250-2010-3 -0.181221
    - wind-vwnd-250-2010-13 -0.184754
    - wind-vwnd-250-2010-17 -0.185391
    - wind-vwnd-925-2010-10 -0.197359
    - wind-vwnd-925-2010-9 -0.225632
    - wind-vwnd-925-2010-3 -0.244246
    - wind-vwnd-250-2010-1 -0.731919
    - wind-vwnd-925-2010-1 -0.807371
- Season
    - season_number 0.427690

- Pressure and potential evaporation: ftp://ftp.cdc.noaa.gov/Datasets/ncep.reanalysis/surface_gauss/
- Pressure
  - contest-pres-sfc-gauss-14d__pres 0.244318
- öppen-Geiger climate classifications: http://koeppen-geiger.vu-wien.ac.at/present.htm
- Climate Region
  - Cfa 0.245350
  - BWh 0.168750
  - BSh 0.107461
  - BWk 0.079562
  - Csa 0.062486
  - Cfb -0.011304
  - Csb -0.011888
  - Dwa -0.013410
  - Dwb -0.017207
  - Dfa -0.035016
  - Dsc -0.040514
  - Dsb -0.042262
  - BSk -0.058873
  - Dfc -0.139666
  - Dfb -0.203071
- Multivariate ENSO index (MEI): Bimonthly MEI values (MEI) from 1949 to the present, were obtained from NOAA/Earth System Research Laboratory. The MEI is a scalar summary of six variables (sea-level pressure, zonal and meridional surface wind components, SST, surface air temperature, and sky cloudiness) associated with El Niño/Southern Oscillation (ENSO), an ocean-atmosphere coupled climate mode. https://www.esrl.noaa.gov/psd/enso/mei/
- MEI NIP (Nino Index Phase)
  - mei__nip 0.197470
- MEI Rank
  - mei__meirank 0.081722
- MEI
  - mei__mei 0.061639 -Global precipitation: Daily precipitation data from 1979 onward were obtained from NOAA's CPC Gauge-Based Analysis of Global Daily Precipitation [42] and converted to mm. ftp://ftp.cpc.ncep.noaa.gov/precip/CPC_UNI_PRCP/GAUGE_GLB/RT/
- U.S. precipitation: Daily U.S. precipitation data in mm were collected from the CPC Unified Gauge-Based Analysis of Daily Precipitation over CONUS. Measurements were replaced with sums over the ensuing two-week period. https://www.esrl.noaa.gov/psd/thredds/catalog/Datasets/cpc_us_precip/catalog.html

- weeks 5-6 weighted average of most recent monthly NMME model forecasts for precipitation

    - nmme0-prate-56w__cfsv20 0.121430
    - nmme0-prate-56w__nasa0 0.115602
    - nmme0-prate-34w__cfsv20 0.098975
    - nmme0-prate-34w__nasa0 0.094668
    - nmme0-prate-56w__cancm30 0.085882
    - nmme0-prate-56w__cancm40 0.065023
    - nmme0-prate-56w__gfdl0 0.063657
    - nmme0-prate-34w__cancm30 0.041223
    - nmme-prate-56w__cfsv2 0.015907
    - nmme-prate-34w__cfsv2 0.014493
    - nmme0-prate-34w__gfdl0 0.010788
    - nmme0-prate-56w__nmme0mean 0.003193
    - nmme-prate-34w__nasa 0.002206
    - nmme0-prate-34w__cancm40 -0.003955
    - nmme-prate-56w__nasa -0.004842
    - nmme0-prate-34w__nmme0mean -0.049191
    - nmme-prate-56w__gfdlflorb -0.066928
    - nmme0-prate-56w__gfdlflora0 -0.067509
    - nmme0-prate-56w__gfdlflorb0 -0.071111
    - nmme-prate-56w__gfdlflora -0.085523
    - nmme0-prate-56w__ccsm40 -0.086109
    - nmme-prate-34w__gfdlflorb -0.096816
    - nmme-prate-34w__gfdlflora -0.105720
    - nmme0-prate-34w__gfdlflora0 -0.107452
    - nmme0-prate-34w__gfdlflorb0 -0.110909
    - nmme0-prate-34w__ccsm40 -0.127728
    - nmme0-prate-56w__ccsm30 -0.147092
    - nmme-prate-56w__gfdl -0.158220
    - nmme-prate-56w__nmmemean -0.163111
    - nmme-prate-34w__cancm3 -0.167068
    - nmme-prate-34w__nmmemean -0.170152
    - nmme-prate-34w__gfdl -0.171759
    - nmme-prate-56w__cancm3 -0.174841
    - nmme-prate-34w__cancm4 -0.201921
    - nmme-prate-56w__cancm4 -0.206909
    - nmme0-prate-34w__ccsm30 -0.216063
    - nmme-prate-34w__ccsm4 -0.263720
    - nmme-prate-56w__ccsm4 -0.263803
    - nmme-prate-56w__ccsm3 -0.286250
    - nmme-prate-34w__ccsm3 -0.298928
- longitude
    - lon 0.092923
- latitude

    - lat -0.398388
- measured precipitation
    - contest-precip-14d__precip 0.079041
- month
    - month_number 0.050412
- Madden-Julian oscillation (MJO): Daily MJO values since 1974 are provided by the Australian Government Bureau of Meteorology. MJO is a metric of tropical convection on daily to weekly timescales and can have a significant impact on the United States sub-seasonal climate. Measurements of phase and amplitude on the target date were extracted over the two-week period. http://www.bom.gov.au/climate/mjo/graphics/rmm.74toRealtime.txt
- MJO phase and amplitude
    - mjo1d__amplitude -0.115210
    - mjo1d__phase -0.130790
- Elevation: http://research.jisao.washington.edu/data_sets/elevation/elev.1-deg.nc
- Elevation
    - elevation__elevation -0.207810
- Index
    - index -0.394012 Relative humidity, sea level pressure, and precipitable water for the entire atmosphere: NOAA's National Center for Environmental Prediction (NCEP)/National Center for Atmospheric Research Reanalysis dataset contains daily relative humidity (rhum) near the surface (sigma level 0.995) from 1948 to the present and daily pressure at the surface (pres) from 1979 to the present. ftp://ftp.cdc.noaa.gov/Datasets/ncep.reanalysis/surface/
- Relative Humidity
    - contest-rhum-sig995-14d__rhum -0.565127
- Sea Level Pressure
    - contest-slp-14d__slp -0.707640

First Modeling Attempt - Random Forest Regressor

- Set up independent and dependent variables.
- We will need to do a TimesSeriesSplit to set up our training and testing data within the training_data to evaluate models without data leakage due to the time series.
- We will attempt a basic Random Forest Regressor or RFR to see what our timing on one of the splits looks like and get a baseline.
- We will then try and tune hyperparameters with a gridsearch.
- We will then test our optimal model.

Set up X and y

- X - All features but our target
- y - Our target `contest-tmp2m-14d__tmp2m`

In [245…
```
# Set index to datetime feature of start date and sort by startdate
# Make copy to keep training data with original index
time_training_data = training_data.copy()
time_training_data.set_index('startdate', inplace=True)
time_training_data.sort_index(inplace=True)
```

In [246…
```
# Display
time_training_data.head()
```

Out[246]:

startdate	index	lat	lon	contest-pevpr-sfc-gauss-14d__pevpr	nmme0-tmp2m-34w__cancm30	nmme0-tmp2m-34w__cancm40	nm tr 34w__c
2014-09-01	0	0.000000	0.833333	237.00	29.02	31.64	
2014-09-01	290938	0.818182	0.633333	323.63	24.18	26.75	
2014-09-01	35819	0.227273	0.900000	385.92	31.16	32.19	
2014-09-01	290207	0.818182	0.600000	303.36	23.34	25.66	
2014-09-01	289476	0.818182	0.566667	319.97	22.50	24.57	

In [247…
```
#Check shape of training data before drop for X
time_training_data.shape
```

Out[247]:
```
(375734, 261)
```

In [301…
```
# Set up independent variables of all columns except target
X = time_training_data.drop(['contest-tmp2m-14d__tmp2m'], axis = 1)
```

In [302…
```
#Check shape after drop - looks good
X.shape
```

Out[302]:
```
(375734, 260)
```

In [303…
```
# Check
X.head()
```

Out[303]:

startdate	index	lat	lon	contest-pevpr-sfc-gauss-14d__pevpr	nmme0-tmp2m-34w__cancm30	nmme0-tmp2m-34w__cancm40	nr tr 34w__c
2014-09-01	0	0.000000	0.833333	237.00	29.02	31.64	
2014-09-01	290938	0.818182	0.633333	323.63	24.18	26.75	
2014-09-01	35819	0.227273	0.900000	385.92	31.16	32.19	
2014-09-01	290207	0.818182	0.600000	303.36	23.34	25.66	
2014-09-01	289476	0.818182	0.566667	319.97	22.50	24.57	

In [306…
```
# Set up array for target data - dependent variable
y = time_training_data['contest-tmp2m-14d__tmp2m']
```

In [307…
```
# Check
y.shape
```

Out[307]:
```
(375734,)
```

In [308…
```
# Check
y.head()
```

Out[308]:
```
startdate
2014-09-01    28.744480
2014-09-01    12.515442
2014-09-01    27.652504
2014-09-01    12.346471
2014-09-01    11.736088
Name: contest-tmp2m-14d__tmp2m, dtype: float64
```

In [309…
```
# Check dependent variable data types --> No Change
X.dtypes
```

Out[309]:

```
index                                    int64
lat                                    float64
lon                                    float64
contest-pevpr-sfc-gauss-14d__pevpr     float64
nmme0-tmp2m-34w__cancm30               float64
nmme0-tmp2m-34w__cancm40               float64
nmme0-tmp2m-34w__ccsm30                float64
nmme0-tmp2m-34w__ccsm40                float64
nmme0-tmp2m-34w__cfsv20                float64
nmme0-tmp2m-34w__gfdlflora0            float64
nmme0-tmp2m-34w__gfdlflorb0            float64
nmme0-tmp2m-34w__gfdl0                 float64
nmme0-tmp2m-34w__nasa0                 float64
nmme0-tmp2m-34w__nmme0mean             float64
contest-wind-h10-14d__wind-hgt-10      float64
nmme-tmp2m-56w__cancm3                 float64
nmme-tmp2m-56w__cancm4                 float64
nmme-tmp2m-56w__ccsm3                  float64
nmme-tmp2m-56w__ccsm4                  float64
nmme-tmp2m-56w__cfsv2                  float64
nmme-tmp2m-56w__gfdl                   float64
nmme-tmp2m-56w__gfdlflora              float64
nmme-tmp2m-56w__gfdlflorb              float64
nmme-tmp2m-56w__nasa                   float64
nmme-tmp2m-56w__nmmemean               float64
contest-rhum-sig995-14d__rhum          float64
nmme-prate-34w__cancm3                 float64
nmme-prate-34w__cancm4                 float64
nmme-prate-34w__ccsm3                  float64
nmme-prate-34w__ccsm4                  float64
nmme-prate-34w__cfsv2                  float64
nmme-prate-34w__gfdl                   float64
nmme-prate-34w__gfdlflora              float64
nmme-prate-34w__gfdlflorb              float64
nmme-prate-34w__nasa                   float64
nmme-prate-34w__nmmemean               float64
contest-wind-h100-14d__wind-hgt-100    float64
nmme0-prate-56w__cancm30               float64
nmme0-prate-56w__cancm40               float64
nmme0-prate-56w__ccsm30                float64
nmme0-prate-56w__ccsm40                float64
nmme0-prate-56w__cfsv20                float64
nmme0-prate-56w__gfdlflora0            float64
nmme0-prate-56w__gfdlflorb0            float64
nmme0-prate-56w__gfdl0                 float64
nmme0-prate-56w__nasa0                 float64
nmme0-prate-56w__nmme0mean             float64
nmme0-prate-34w__cancm30               float64
nmme0-prate-34w__cancm40               float64
nmme0-prate-34w__ccsm30                float64
nmme0-prate-34w__ccsm40                float64
nmme0-prate-34w__cfsv20                float64
```

```
nmme0-prate-34w__gfdlflora0                float64
nmme0-prate-34w__gfdlflorb0                float64
nmme0-prate-34w__gfdl0                     float64
nmme0-prate-34w__nasa0                     float64
nmme0-prate-34w__nmme0mean                 float64
contest-slp-14d__slp                       float64
contest-wind-vwnd-925-14d__wind-vwnd-925   float64
nmme-prate-56w__cancm3                     float64
nmme-prate-56w__cancm4                     float64
nmme-prate-56w__ccsm3                      float64
nmme-prate-56w__ccsm4                      float64
nmme-prate-56w__cfsv2                      float64
nmme-prate-56w__gfdl                       float64
nmme-prate-56w__gfdlflora                  float64
nmme-prate-56w__gfdlflorb                  float64
nmme-prate-56w__nasa                       float64
nmme-prate-56w__nmmemean                   float64
contest-pres-sfc-gauss-14d__pres           float64
contest-wind-uwnd-250-14d__wind-uwnd-250   float64
nmme-tmp2m-34w__cancm3                     float64
nmme-tmp2m-34w__cancm4                     float64
nmme-tmp2m-34w__ccsm3                      float64
nmme-tmp2m-34w__ccsm4                      float64
nmme-tmp2m-34w__cfsv2                      float64
nmme-tmp2m-34w__gfdl                       float64
nmme-tmp2m-34w__gfdlflora                  float64
nmme-tmp2m-34w__gfdlflorb                  float64
nmme-tmp2m-34w__nasa                       float64
nmme-tmp2m-34w__nmmemean                   float64
contest-prwtr-eatm-14d__prwtr              float64
contest-wind-vwnd-250-14d__wind-vwnd-250   float64
contest-precip-14d__precip                 float64
contest-wind-h850-14d__wind-hgt-850        float64
contest-wind-uwnd-925-14d__wind-uwnd-925   float64
contest-wind-h500-14d__wind-hgt-500        float64
cancm30                                    float64
cancm40                                    float64
ccsm30                                     float64
ccsm40                                     float64
cfsv20                                     float64
gfdlflora0                                 float64
gfdlflorb0                                 float64
gfdl0                                      float64
nasa0                                      float64
nmme0mean                                  float64
elevation__elevation                       float64
wind-vwnd-250-2010-1                       float64
wind-vwnd-250-2010-2                       float64
wind-vwnd-250-2010-3                       float64
wind-vwnd-250-2010-4                       float64
wind-vwnd-250-2010-5                       float64
wind-vwnd-250-2010-6                       float64
```

```
wind-vwnd-250-2010-7                    float64
wind-vwnd-250-2010-8                    float64
wind-vwnd-250-2010-9                    float64
wind-vwnd-250-2010-10                   float64
wind-vwnd-250-2010-11                   float64
wind-vwnd-250-2010-12                   float64
wind-vwnd-250-2010-13                   float64
wind-vwnd-250-2010-14                   float64
wind-vwnd-250-2010-15                   float64
wind-vwnd-250-2010-16                   float64
wind-vwnd-250-2010-17                   float64
wind-vwnd-250-2010-18                   float64
wind-vwnd-250-2010-19                   float64
wind-vwnd-250-2010-20                   float64
wind-uwnd-250-2010-1                    float64
wind-uwnd-250-2010-2                    float64
wind-uwnd-250-2010-3                    float64
wind-uwnd-250-2010-4                    float64
wind-uwnd-250-2010-5                    float64
wind-uwnd-250-2010-6                    float64
wind-uwnd-250-2010-7                    float64
wind-uwnd-250-2010-8                    float64
wind-uwnd-250-2010-9                    float64
wind-uwnd-250-2010-10                   float64
wind-uwnd-250-2010-11                   float64
wind-uwnd-250-2010-12                   float64
wind-uwnd-250-2010-13                   float64
wind-uwnd-250-2010-14                   float64
wind-uwnd-250-2010-15                   float64
wind-uwnd-250-2010-16                   float64
wind-uwnd-250-2010-17                   float64
wind-uwnd-250-2010-18                   float64
wind-uwnd-250-2010-19                   float64
wind-uwnd-250-2010-20                   float64
mjo1d__phase                              int64
mjo1d__amplitude                        float64
mei__mei                                float64
mei__meirank                              int64
mei__nip                                  int64
wind-hgt-850-2010-1                     float64
wind-hgt-850-2010-2                     float64
wind-hgt-850-2010-3                     float64
wind-hgt-850-2010-4                     float64
wind-hgt-850-2010-5                     float64
wind-hgt-850-2010-6                     float64
wind-hgt-850-2010-7                     float64
wind-hgt-850-2010-8                     float64
wind-hgt-850-2010-9                     float64
wind-hgt-850-2010-10                    float64
sst-2010-1                              float64
sst-2010-2                              float64
sst-2010-3                              float64
```

```
sst-2010-4                          float64
sst-2010-5                          float64
sst-2010-6                          float64
sst-2010-7                          float64
sst-2010-8                          float64
sst-2010-9                          float64
sst-2010-10                         float64
wind-hgt-500-2010-1                 float64
wind-hgt-500-2010-2                 float64
wind-hgt-500-2010-3                 float64
wind-hgt-500-2010-4                 float64
wind-hgt-500-2010-5                 float64
wind-hgt-500-2010-6                 float64
wind-hgt-500-2010-7                 float64
wind-hgt-500-2010-8                 float64
wind-hgt-500-2010-9                 float64
wind-hgt-500-2010-10                float64
icec-2010-1                         float64
icec-2010-2                         float64
icec-2010-3                         float64
icec-2010-4                         float64
icec-2010-5                         float64
icec-2010-6                         float64
icec-2010-7                         float64
icec-2010-8                         float64
icec-2010-9                         float64
icec-2010-10                        float64
wind-uwnd-925-2010-1                float64
wind-uwnd-925-2010-2                float64
wind-uwnd-925-2010-3                float64
wind-uwnd-925-2010-4                float64
wind-uwnd-925-2010-5                float64
wind-uwnd-925-2010-6                float64
wind-uwnd-925-2010-7                float64
wind-uwnd-925-2010-8                float64
wind-uwnd-925-2010-9                float64
wind-uwnd-925-2010-10               float64
wind-uwnd-925-2010-11               float64
wind-uwnd-925-2010-12               float64
wind-uwnd-925-2010-13               float64
wind-uwnd-925-2010-14               float64
wind-uwnd-925-2010-15               float64
wind-uwnd-925-2010-16               float64
wind-uwnd-925-2010-17               float64
wind-uwnd-925-2010-18               float64
wind-uwnd-925-2010-19               float64
wind-uwnd-925-2010-20               float64
wind-hgt-10-2010-1                  float64
wind-hgt-10-2010-2                  float64
wind-hgt-10-2010-3                  float64
wind-hgt-10-2010-4                  float64
wind-hgt-10-2010-5                  float64
```

```
wind-hgt-10-2010-6                  float64
wind-hgt-10-2010-7                  float64
wind-hgt-10-2010-8                  float64
wind-hgt-10-2010-9                  float64
wind-hgt-10-2010-10                 float64
wind-hgt-100-2010-1                 float64
wind-hgt-100-2010-2                 float64
wind-hgt-100-2010-3                 float64
wind-hgt-100-2010-4                 float64
wind-hgt-100-2010-5                 float64
wind-hgt-100-2010-6                 float64
wind-hgt-100-2010-7                 float64
wind-hgt-100-2010-8                 float64
wind-hgt-100-2010-9                 float64
wind-hgt-100-2010-10                float64
wind-vwnd-925-2010-1                float64
wind-vwnd-925-2010-2                float64
wind-vwnd-925-2010-3                float64
wind-vwnd-925-2010-4                float64
wind-vwnd-925-2010-5                float64
wind-vwnd-925-2010-6                float64
wind-vwnd-925-2010-7                float64
wind-vwnd-925-2010-8                float64
wind-vwnd-925-2010-9                float64
wind-vwnd-925-2010-10               float64
wind-vwnd-925-2010-11               float64
wind-vwnd-925-2010-12               float64
wind-vwnd-925-2010-13               float64
wind-vwnd-925-2010-14               float64
wind-vwnd-925-2010-15               float64
wind-vwnd-925-2010-16               float64
wind-vwnd-925-2010-17               float64
wind-vwnd-925-2010-18               float64
wind-vwnd-925-2010-19               float64
wind-vwnd-925-2010-20               float64
BSh                                   uint8
BSk                                   uint8
BWh                                   uint8
BWk                                   uint8
Cfa                                   uint8
Cfb                                   uint8
Csa                                   uint8
Csb                                   uint8
Dfa                                   uint8
Dfb                                   uint8
Dfc                                   uint8
Dsb                                   uint8
Dsc                                   uint8
Dwa                                   uint8
Dwb                                   uint8
month_number                          int64
```

```
season_number                                int64
dtype: object
```

TimeSeriesSplit

- Forward Chaining window
- idea gathered from Brad Messer on Competition Kaggle Page
  - https://www.kaggle.com/competitions/widsdatathon2023/discussion/377400

In [260…
```
from sklearn.model_selection import TimeSeriesSplit

for train_idx, test_idx in TimeSeriesSplit(n_splits=5).split(X):
    X_train, X_test = X.iloc[train_idx, :], X.iloc[test_idx,:]
    y_train, y_test = y.iloc[train_idx], y.iloc[test_idx]
```

In [287…
```
# check that index of train is before test
X_train.shape
```

Out[287]:
```
(313112, 259)
```

In [288…
```
X_test.shape
```

Out[288]:
```
(62622, 259)
```

In [265…
```
#Looks good
X_test.index
```

Out[265]:
```
DatetimeIndex(['2016-05-02', '2016-05-02', '2016-05-02', '2016-05-02',
               '2016-05-02', '2016-05-02', '2016-05-02', '2016-05-02',
               '2016-05-02', '2016-05-02',
               ...
               '2016-08-31', '2016-08-31', '2016-08-31', '2016-08-31',
               '2016-08-31', '2016-08-31', '2016-08-31', '2016-08-31',
               '2016-08-31', '2016-08-31'],
              dtype='datetime64[ns]', name='startdate', length=62622, freq=Non
e)
```

In [266…
```
y_test.index
```

Out[266]:
```
DatetimeIndex(['2016-05-02', '2016-05-02', '2016-05-02', '2016-05-02',
               '2016-05-02', '2016-05-02', '2016-05-02', '2016-05-02',
               '2016-05-02', '2016-05-02',
               ...
               '2016-08-31', '2016-08-31', '2016-08-31', '2016-08-31',
               '2016-08-31', '2016-08-31', '2016-08-31', '2016-08-31',
               '2016-08-31', '2016-08-31'],
              dtype='datetime64[ns]', name='startdate', length=62622, freq=Non
e)
```

In [341…
```
#Visualize test set - Make sure testing data is after training data.
plt.figure()
sns.set(rc={'figure.figsize':(10,7)})
y_train.groupby('startdate').mean().plot(label=' training data')
y_test.groupby('startdate').mean().plot(label = 'test data')
plt.xlabel('Start Date')
plt.ylabel('Target - Mean Temp - contest-tmp2m-14d__tmp2m')
plt.legend()
plt.show()
```

In [284…
```
#Import model and scoring metric
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error
```

```
from sklearn.metrics import r2_score
from math import sqrt
```

In [274…

```
# Instatiate
RFRmodel = RandomForestRegressor(n_estimators=10, verbose = 1,
random_state=32)
```

In [276…

```
#Fit Model
%time
RFRmodel.fit(X_train,y_train)
```

```
CPU times: user 2 µs, sys: 1e+03 ns, total: 3 µs
Wall time: 5.72 µs
```

```
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  10 out of  10 | elapsed:  6.5min finished
```

Out[276]:

```
RandomForestRegressor(n_estimators=10, random_state=32, verbose=1)
```

In [279…

```
# Get Predictions from RFR Model
pred = RFRmodel.predict(X_test)
```

```
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  10 out of  10 | elapsed:    1.1s finished
```

In [281…

```
# Gather results in dataframe for visualization of expected vs.
predictions
results_RFR = pd.DataFrame(data = {'Actual':y_test, \
                                   'Predictions':pred},
index=y_test.index)
```

In [282…

```
results_RFR
```

Out[282]:

startdate	Actual	Predictions
2016-05-02	21.080032	21.262204
2016-05-02	12.921798	13.555778
2016-05-02	11.742005	11.453181
2016-05-02	18.386656	18.363113
2016-05-02	10.771266	11.325038
...	...	...
2016-08-31	19.772010	19.855250
2016-08-31	19.998931	19.987436
2016-08-31	20.392470	20.069189
2016-08-31	10.406187	10.030828
2016-08-31	15.910995	16.069659

62622 rows × 2 columns

In [285…

```
# Error Metrics for initial RFR
print('R-squared =
{:.3f}'.format(r2_score(results_RFR['Actual'],results_RFR['Prediction
print('RMSE =
{:.3f}'.format(sqrt(mean_squared_error(results_RFR['Actual'],results_
```

```
R-squared = 0.900
RMSE = 1.772
```

RFR Initial Model Evaluation

- This is promising. With no hypertuning, we have some pretty low error metrics.
  - R-squared = 0.900
  - RMSE = 1.772
- Timing - We will need to get a better method to time running this model as I stepped away from the computer while it ran. I would estimate that it took less than an hour.
- We will now tune hyperparameters this one with a grid search.

RFR Tuning Hyperparameters

- We will need to get more folds in our time series split and then set up a grid search.
- With more estimators and adjusting other hyperparameters, I think we can really improve the predictive power of this model.

In [289…
```
# Exploring timeseriessplit
tscv = TimeSeriesSplit()
print(tscv)
```

```
TimeSeriesSplit(gap=0, max_train_size=None, n_splits=5, test_size=None)
```

In [290…
```
#Print out folds
for i, (train_index, test_index) in enumerate(tscv.split(X)):
    print(f"Fold {i}:")
    print(f"  Train: index={train_index}")
    print(f"  Test:  index={test_index}")
```

```
Fold 0:
  Train: index=[     0      1      2 ... 62621 62622 62623]
  Test:  index=[ 62624  62625  62626 ... 125243 125244 125245]
Fold 1:
  Train: index=[     0      1      2 ... 125243 125244 125245]
  Test:  index=[125246 125247 125248 ... 187865 187866 187867]
Fold 2:
  Train: index=[     0      1      2 ... 187865 187866 187867]
  Test:  index=[187868 187869 187870 ... 250487 250488 250489]
Fold 3:
  Train: index=[     0      1      2 ... 250487 250488 250489]
  Test:  index=[250490 250491 250492 ... 313109 313110 313111]
Fold 4:
  Train: index=[     0      1      2 ... 313109 313110 313111]
  Test:  index=[313112 313113 313114 ... 375731 375732 375733]
```

In [310…
```
for train_idx, test_idx in TimeSeriesSplit(n_splits=5).split(X):
    X_train, X_test = X.iloc[train_idx, :], X.iloc[test_idx,:]
    y_train, y_test = y.iloc[train_idx], y.iloc[test_idx]
```

In [311…
```
X_train.shape
```

Out[311]:
```
(313112, 260)
```

In [312…
```
X_test.shape
```

Out[312]:
```
(62622, 260)
```

In [294…
```
# We will do 5 splits with 10 and 20 estimators and see how long
this takes
# We will do a more intensive grid search with more parameters once
we no this is good to go
```

```
%%time
from sklearn.model_selection import GridSearchCV

model = RandomForestRegressor(verbose = 1, random_state=32)
param_search = {'n_estimators': [10, 20],
                'max_depth': [None]}

tscv = TimeSeriesSplit(n_splits=5)
gsearch = GridSearchCV(estimator=model, cv=tscv,
                       param_grid=param_search)
gsearch.fit(X_train, y_train)
```

```
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  10 out of  10 | elapsed:   52.9s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  10 out of  10 | elapsed:    0.1s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  10 out of  10 | elapsed:  2.0min finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  10 out of  10 | elapsed:    0.1s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  10 out of  10 | elapsed:  3.3min finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  10 out of  10 | elapsed:    0.1s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  10 out of  10 | elapsed:  4.9min finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  10 out of  10 | elapsed:    0.3s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  10 out of  10 | elapsed:  7.0min finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  10 out of  10 | elapsed:    0.4s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  20 out of  20 | elapsed:  2.0min finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  20 out of  20 | elapsed:    0.3s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  20 out of  20 | elapsed:  4.6min finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  20 out of  20 | elapsed:    0.2s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  20 out of  20 | elapsed:  7.3min finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  20 out of  20 | elapsed:    0.7s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  20 out of  20 | elapsed:  9.4min finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
```

```
s.
[Parallel(n_jobs=1)]: Done  20 out of  20 | elapsed:    0.5s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  20 out of  20 | elapsed: 10.4min finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  20 out of  20 | elapsed:    0.8s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
CPU times: user 1h 3min 32s, sys: 40.4 s, total: 1h 4min 13s
Wall time: 1h 5min 9s
```

```
[Parallel(n_jobs=1)]: Done  20 out of  20 | elapsed: 13.1min finished
```

Out[294]:
```
GridSearchCV(cv=TimeSeriesSplit(gap=0, max_train_size=None, n_splits=5, test_s
ize=None),
             estimator=RandomForestRegressor(random_state=32, verbose=1),
             param_grid={'max_depth': [None], 'n_estimators': [10, 20]})
```

In [295…
```
# Obtain best score from grid search; default is R-squared.
gsearch.score(X_test, y_test)
```

```
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  20 out of  20 | elapsed:    0.8s finished
```

Out[295]:
```
0.9067018890795158
```

In [296…
```
# View best model from grid search
gsearch.best_estimator_
```

Out[296]:
```
RandomForestRegressor(n_estimators=20, random_state=32, verbose=1)
```

In [297…
```
#Get Predictions from grid search
pred = gsearch.predict(X_test)
```

```
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  20 out of  20 | elapsed:    0.6s finished
```

In [298…
```
# Gather results in dataframe for visualization of expected vs.
predictions
results_RFR = pd.DataFrame(data = {'Actual':y_test, \
                                   'Predictions':pred},
index=y_test.index)
```

In [299…
```
# Check
results_RFR
```

Out[299]:

startdate	Actual	Predictions
2016-05-02	21.080032	20.550294
2016-05-02	12.921798	13.442718
2016-05-02	11.742005	11.474627
2016-05-02	18.386656	18.415572
2016-05-02	10.771266	11.392774
...	...	...
2016-08-31	19.772010	19.343009
2016-08-31	19.998931	20.263724
2016-08-31	20.392470	19.799303
2016-08-31	10.406187	11.024972
2016-08-31	15.910995	16.072374

62622 rows × 2 columns

In [300…

```
# Error Metrics for Grid Search
print('R-squared =
{:.3f}'.format(r2_score(results_RFR['Actual'],results_RFR['Prediction
print('RMSE =
{:.3f}'.format(sqrt(mean_squared_error(results_RFR['Actual'],results_
```

```
R-squared = 0.907
RMSE = 1.712
```

Initial Grid Search - Evaluation

- We will need to run another grid search as 20 estimators did better than 10, but just slightly.

R-squared = 0.907 RMSE = 1.712

- We will run it with higher number of estimators (20, 100, 200)
- We will do less splits to reduce run time 5 to 3.
- We will also run with max_features of (None, sqrt).

In [315…

```
%%time
from sklearn.model_selection import GridSearchCV

model = RandomForestRegressor(criterion = 'squared_error', verbose =
1, random_state=32)
```

```
param_search = {'n_estimators': [20, 100, 200],
                'max_features': [None, 'sqrt']}

tscv = TimeSeriesSplit(n_splits=3)
gsearch2 = GridSearchCV(estimator=model, cv=tscv,
                        param_grid=param_search)
gsearch2.fit(X_train, y_train)
```

```
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  20 out of  20 | elapsed:  2.5min finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  20 out of  20 | elapsed:    0.3s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  20 out of  20 | elapsed:  5.8min finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  20 out of  20 | elapsed:    0.5s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  20 out of  20 | elapsed:  9.1min finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  20 out of  20 | elapsed:    0.6s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done 100 out of 100 | elapsed: 12.5min finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done 100 out of 100 | elapsed:    2.0s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done 100 out of 100 | elapsed: 29.8min finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done 100 out of 100 | elapsed:    6.6s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done 100 out of 100 | elapsed: 47.3min finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done 100 out of 100 | elapsed:    7.2s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done 200 out of 200 | elapsed: 25.0min finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done 200 out of 200 | elapsed:    5.2s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done 200 out of 200 | elapsed: 55.4min finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done 200 out of 200 | elapsed:    8.8s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done 200 out of 200 | elapsed: 86.8min finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
```

```
s.
[Parallel(n_jobs=1)]: Done 200 out of 200 | elapsed:   13.5s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  20 out of  20 | elapsed:   10.4s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  20 out of  20 | elapsed:    0.3s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  20 out of  20 | elapsed:   22.7s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  20 out of  20 | elapsed:    0.4s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  20 out of  20 | elapsed:   36.7s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done  20 out of  20 | elapsed:    0.5s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done 100 out of 100 | elapsed:   51.3s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done 100 out of 100 | elapsed:    1.4s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done 100 out of 100 | elapsed:  1.9min finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done 100 out of 100 | elapsed:    2.0s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done 100 out of 100 | elapsed:  3.1min finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done 100 out of 100 | elapsed:    2.2s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done 200 out of 200 | elapsed:  1.7min finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done 200 out of 200 | elapsed:    2.9s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done 200 out of 200 | elapsed:  3.8min finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done 200 out of 200 | elapsed:    4.8s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
```

```
[Parallel(n_jobs=1)]: Done 200 out of 200 | elapsed:  6.1min finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done 200 out of 200 | elapsed:    5.0s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
CPU times: user 5h 52min 26s, sys: 1min 39s, total: 5h 54min 5s
Wall time: 5h 55min 42s
```

```
[Parallel(n_jobs=1)]: Done 100 out of 100 | elapsed: 61.3min finished
```

Out[315]:
```
GridSearchCV(cv=TimeSeriesSplit(gap=0, max_train_size=None, n_splits=3, test_s
ize=None),
             estimator=RandomForestRegressor(random_state=32, verbose=1),
             param_grid={'max_features': [None, 'sqrt'],
                         'n_estimators': [20, 100, 200]})
```

In [316…
```
# Best model of our second grid search
gsearch2.best_estimator_
```

Out[316]:
```
RandomForestRegressor(max_features=None, random_state=32, verbose=1)
```

In [326…
```
# Best parameters for that model of second grid search
gsearch2.best_params_
```

Out[326]:
```
{'max_features': None, 'n_estimators': 100}
```

In [317…
```
# Score second grid search - R-squared
gsearch2.score(X_test, y_test)
```

```
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done 100 out of 100 | elapsed:    6.5s finished
```

Out[317]:
```
0.9086886162434469
```

In [320…
```
# Get predictions on test data
pred = gsearch2.predict(X_test)
```

```
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done 100 out of 100 | elapsed:    2.1s finished
```

In [321…
```
# Gather results in dataframe for visualization of expected vs.
predictions
results_RFR = pd.DataFrame(data = {'Actual':y_test, \
                                   'Predictions':pred},
index=y_test.index)
```

In [322…

```
results_RFR
```

Out[322]:

startdate	Actual	Predictions
2016-05-02	21.080032	20.384202
2016-05-02	12.921798	13.349094
2016-05-02	11.742005	11.339025
2016-05-02	18.386656	17.986082
2016-05-02	10.771266	11.314935
...	...	...
2016-08-31	19.772010	19.113198
2016-08-31	19.998931	19.511693
2016-08-31	20.392470	19.606974
2016-08-31	10.406187	11.100920
2016-08-31	15.910995	16.343491

62622 rows × 2 columns

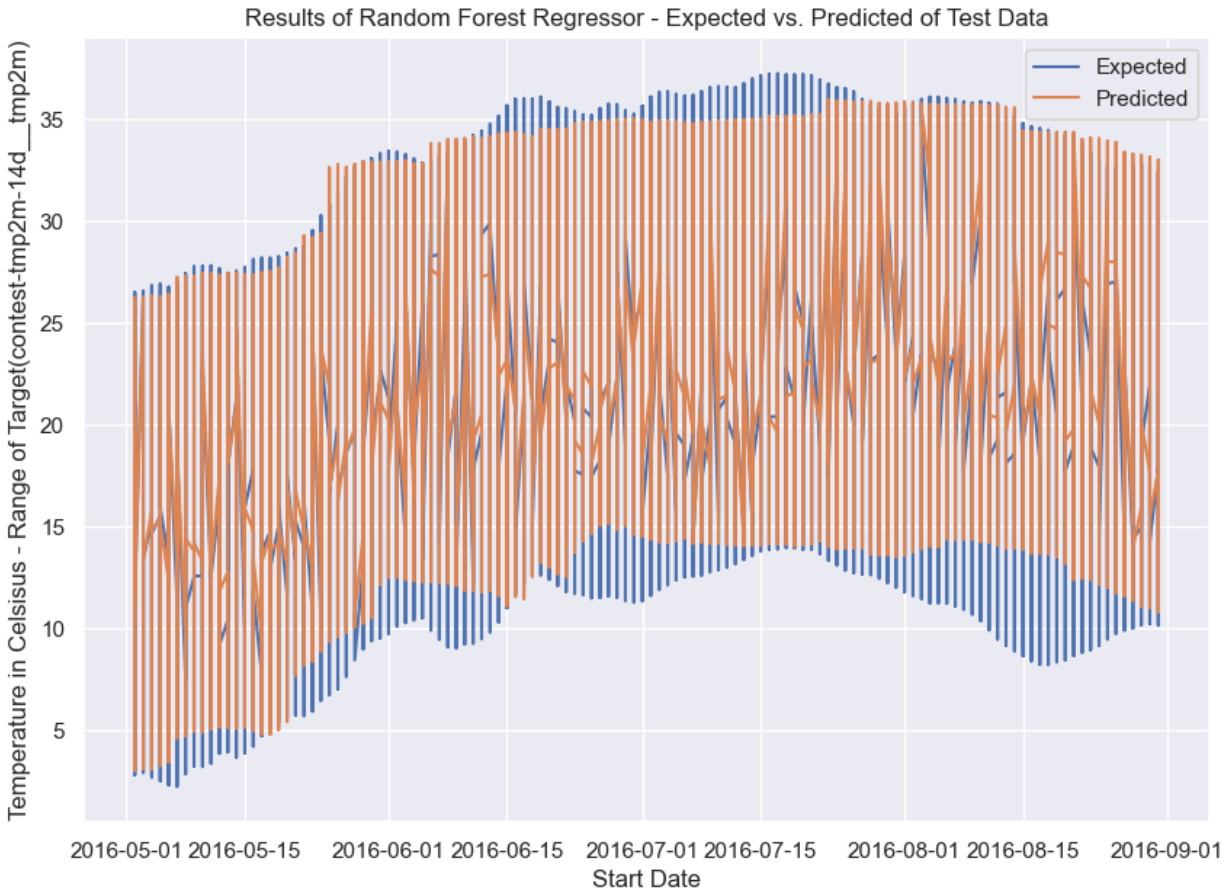
In [323…

```
# Error Metrics for Grid Search2 with RFR {'max_features': None,
'n_estimators': 100}
print('R-squared =
{:.3f}'.format(r2_score(results_RFR['Actual'],results_RFR['Prediction
print('RMSE =
{:.3f}'.format(sqrt(mean_squared_error(results_RFR['Actual'],results_
```

```
R-squared = 0.909
RMSE = 1.694
```

In [338…

```
plt.figure()
plt.plot(results_RFR['Actual'], label = 'Expected')
plt.plot(results_RFR['Predictions'], label ='Predicted')
plt.title("Results of Random Forest Regressor - Expected vs.
Predicted of Test Data")
plt.xlabel('Start Date')
plt.ylabel('Temperature in Celsisus - Range of Target(contest-tmp2m-
14d__tmp2m)')
plt.legend()
plt.show()
```



In [330…

```
# group by startdate and aggregate with mean of each start date
results_mean_df = results_RFR.groupby('startdate').agg(['mean'])
```

In [331…

```
# Check
results_mean_df
```

Out[331]:

	Actual	Predictions
	mean	mean
startdate		
2016-05-02	14.274914	14.310927
2016-05-03	14.461023	14.540965
2016-05-04	14.443859	14.614066
2016-05-05	14.346695	14.630657
2016-05-06	14.211061	14.693191
2016-05-07	14.146367	15.493039
2016-05-08	14.217628	15.524998
2016-05-09	14.305582	15.532885
2016-05-10	14.366480	15.558014
2016-05-11	14.527683	15.586419
2016-05-12	14.713003	15.616100
2016-05-13	14.801056	15.638675
2016-05-14	14.825579	15.633770
2016-05-15	14.991848	15.633594
2016-05-16	15.309252	15.630491
2016-05-17	15.587497	15.630550
2016-05-18	15.830777	15.685049
2016-05-19	16.069144	15.888248
2016-05-20	16.367926	16.377168
2016-05-21	16.734441	16.772342
2016-05-22	17.126479	17.121815
2016-05-23	17.541987	17.927965
2016-05-24	17.993140	18.358063
2016-05-25	18.402610	18.833023
2016-05-26	18.858181	19.174090
2016-05-27	19.341647	19.561069
2016-05-28	19.818440	19.958675
2016-05-29	20.165615	20.441802
2016-05-30	20.371744	20.983861
2016-05-31	20.542292	21.308870
2016-06-01	20.707600	21.386023
2016-06-02	20.817875	21.362522
2016-06-03	20.884483	21.359634

	Actual mean	Predictions mean
startdate		
2016-06-04	20.926638	21.309402
2016-06-05	20.967283	21.272538
2016-06-06	21.005520	21.442642
2016-06-07	21.085344	21.512720
2016-06-08	21.219452	21.634363
2016-06-09	21.255808	21.619854
2016-06-10	21.285071	21.622853
2016-06-11	21.342450	21.753639
2016-06-12	21.404886	21.677452
2016-06-13	21.527032	21.817593
2016-06-14	21.749333	21.889030
2016-06-15	22.059877	22.088860
2016-06-16	22.362804	22.021168
2016-06-17	22.553303	22.104227
2016-06-18	22.609243	22.229726
2016-06-19	22.664623	22.368067
2016-06-20	22.732632	22.510069
2016-06-21	22.774394	22.584579
2016-06-22	22.729320	22.656005
2016-06-23	22.696717	22.685234
2016-06-24	22.655146	22.763800
2016-06-25	22.659306	22.855015
2016-06-26	22.795126	23.054341
2016-06-27	22.838457	23.170720
2016-06-28	22.724302	23.260783
2016-06-29	22.580443	23.391446
2016-06-30	22.514432	23.603313
2016-07-01	22.508604	23.750317
2016-07-02	22.563888	23.754463
2016-07-03	22.586919	23.730150
2016-07-04	22.595789	23.734974
2016-07-05	22.637340	23.787139
2016-07-06	22.735289	23.895172

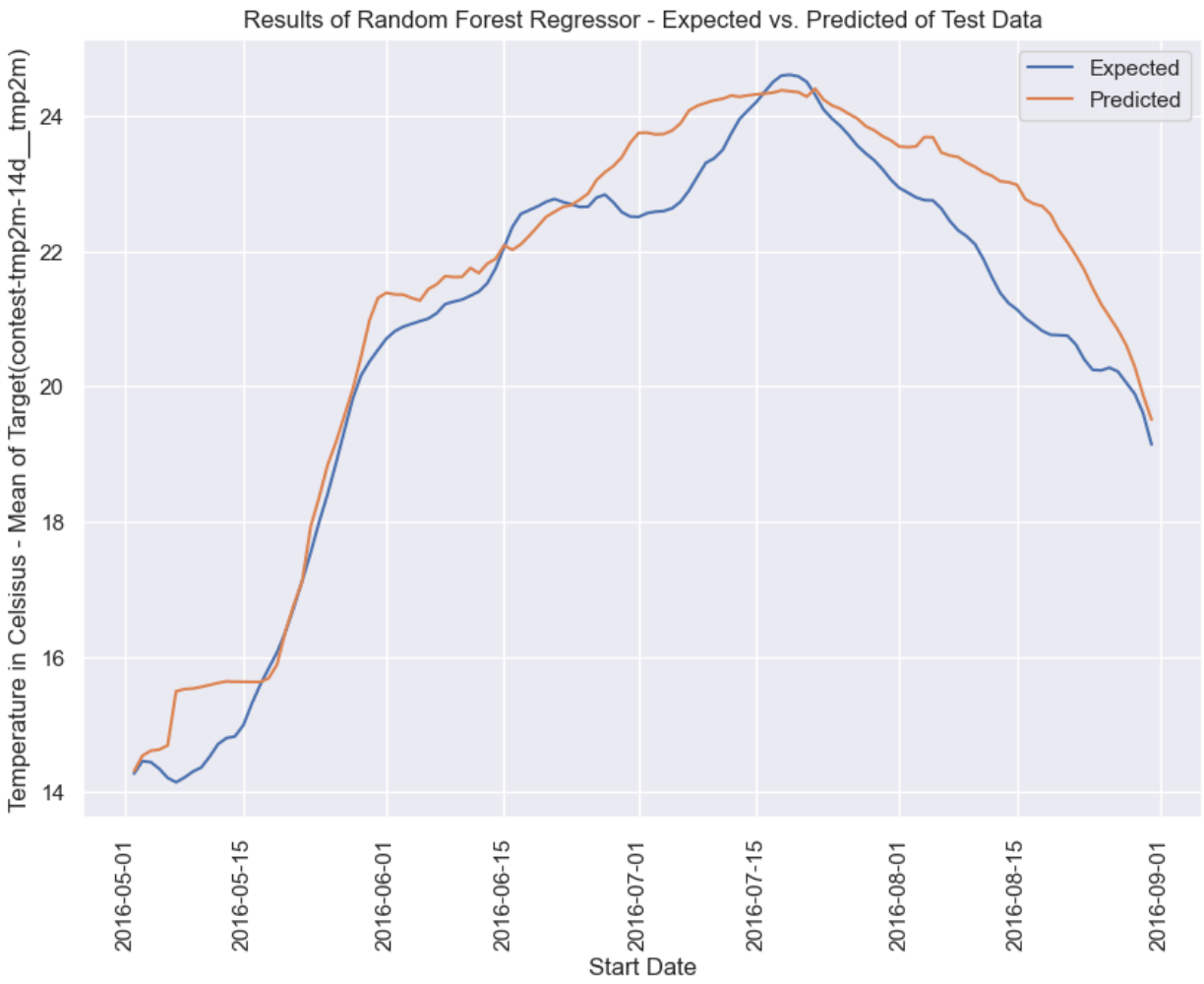
	Actual mean	Predictions mean
startdate		
2016-07-07	22.899492	24.083417
2016-07-08	23.107291	24.153503
2016-07-09	23.308575	24.195053
2016-07-10	23.376422	24.234194
2016-07-11	23.499349	24.256317
2016-07-12	23.737940	24.303000
2016-07-13	23.955960	24.285970
2016-07-14	24.084743	24.305950
2016-07-15	24.207706	24.323003
2016-07-16	24.360605	24.333157
2016-07-17	24.505871	24.346964
2016-07-18	24.600690	24.382195
2016-07-19	24.611658	24.367344
2016-07-20	24.587965	24.354882
2016-07-21	24.505870	24.287252
2016-07-22	24.313727	24.406677
2016-07-23	24.096559	24.242978
2016-07-24	23.963091	24.158331
2016-07-25	23.854308	24.108834
2016-07-26	23.718363	24.033830
2016-07-27	23.564793	23.964308
2016-07-28	23.449897	23.851828
2016-07-29	23.349218	23.792282
2016-07-30	23.216455	23.703803
2016-07-31	23.060571	23.641221
2016-08-01	22.939113	23.550775
2016-08-02	22.869114	23.541603
2016-08-03	22.797587	23.552346
2016-08-04	22.757616	23.690444
2016-08-05	22.752540	23.687160
2016-08-06	22.633763	23.461556
2016-08-07	22.451044	23.418393
2016-08-08	22.310703	23.395170

	Actual mean	Predictions mean
startdate		
2016-08-09	22.224109	23.311834
2016-08-10	22.106072	23.250379
2016-08-11	21.887269	23.168904
2016-08-12	21.622508	23.117612
2016-08-13	21.385587	23.037370
2016-08-14	21.233652	23.021253
2016-08-15	21.139810	22.984461
2016-08-16	21.012116	22.770004
2016-08-17	20.919937	22.702594
2016-08-18	20.824586	22.669502
2016-08-19	20.765758	22.545512
2016-08-20	20.761002	22.309198
2016-08-21	20.752686	22.134454
2016-08-22	20.622889	21.939019
2016-08-23	20.401030	21.726348
2016-08-24	20.247058	21.458705
2016-08-25	20.238387	21.221473
2016-08-26	20.279093	21.031688
2016-08-27	20.222286	20.839554
2016-08-28	20.056948	20.612867
2016-08-29	19.893135	20.295799
2016-08-30	19.606712	19.875769
2016-08-31	19.142702	19.510920

In [339…

```
# Visuallize means of each start date
plt.figure()
plt.plot(results_mean_df['Actual'], label = 'Expected')
plt.plot(results_mean_df['Predictions'], label = 'Predicted')
plt.xticks(rotation = 90)
plt.title("Results of Random Forest Regressor - Expected vs.
Predicted of Test Data")
plt.xlabel('Start Date')
plt.ylabel('Temperature in Celsisus - Mean of Target(contest-tmp2m-
14d__tmp2m)')
```

```
plt.legend()
plt.show()
```



RFR - Final Evaluation

The RFR did a nice job, but after tuning the model we could not get much improvment on our error metrics optimizing our scores on the test set at:

- R-squared = 0.909
- RMSE = 1.694

We will need to test other regression models in our next notebook to improve on this root mean squared error metric.

WiDS Notebook 1 - Conclusion

We gained some valuable insight into the 2023 WiDS Challenge. We now have a good grasp on the size and shape of our data sets. We know how our target variables relates to other features. We saw strong positive and negative correlations between features and our target.

These features will allow us to make nice predictions and have low error metrics when choosing models that allow for multicollinearity.

The random forest regressor model performed well making nice predictions with low errors scores as seen in the the two cells above. Unfortunately, while these scores are in shooting distance of the challenge leaderboard they would not put us on the leaderboard. Competitors on the leaderboard are achieving RMSE scores of 0.79 to 1.14. We will need to move on to other models as the RFR after optimizing did not improve much. Combing some of the competition's discussion posts, teams are having better luck with XGRRegressor, MLPRegressor, and LGBMRegressor models.

We have set ourselves up for a good start with this notebook. We will need to carry out a variety of steps in Notebook 2 to acieve better results.

Next Steps

- Feature/data reduction
  - Run times on our grid searches were in the 5 to 6 hour range.
  - We will carry out some pricipal compenent analysis or PCA to reduce run time.
- Feature engineering
  - We will research steps to boost the predictive power of our features.
- More comprehensive model evaluation
  - We will need to extend our grid searches to other regressor models.

In [ ]: