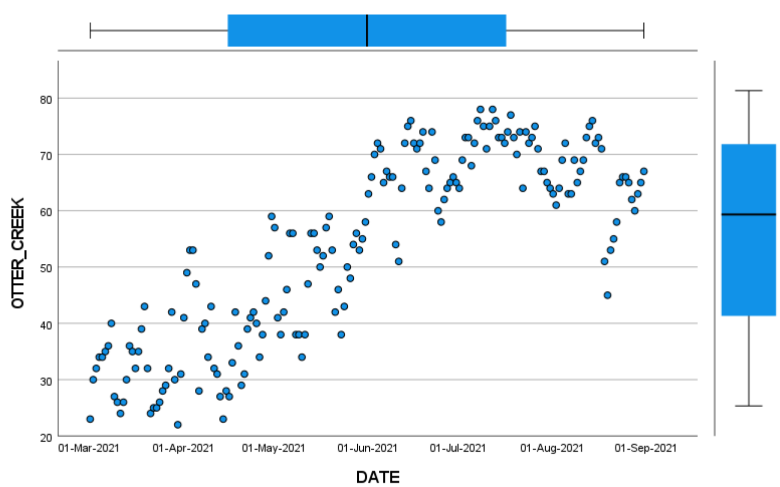
Ann Despain
6010 Linear Regression Project
December 1, 2021

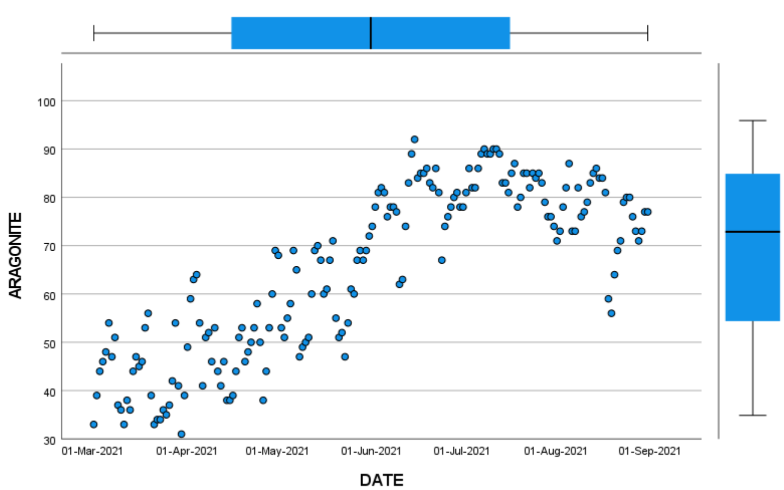
Objective: Predict the average temperature of Salt Lake City, Utah from a linear model of 5 surrounding cities.

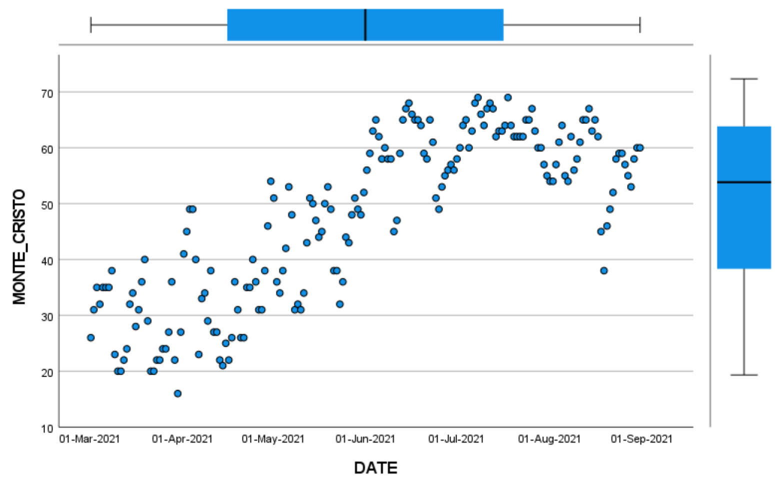
I have selected 5 cities in Utah and collected their daily temperature averages from the National Centers for Environmental Information from **March 1, 2021 to August 31, 2021.** The cities I have chosen are

1. **Aragonite**, located directly west of SLC
2. **Monte Cristo**, located directly north of SLC
3. **Otter Creek**, located directly south in central Utah
4. **Joe's Valley**, located directly south in central Utah
5. **Hewinta**, located north east, in the Uinta Mountain Range.

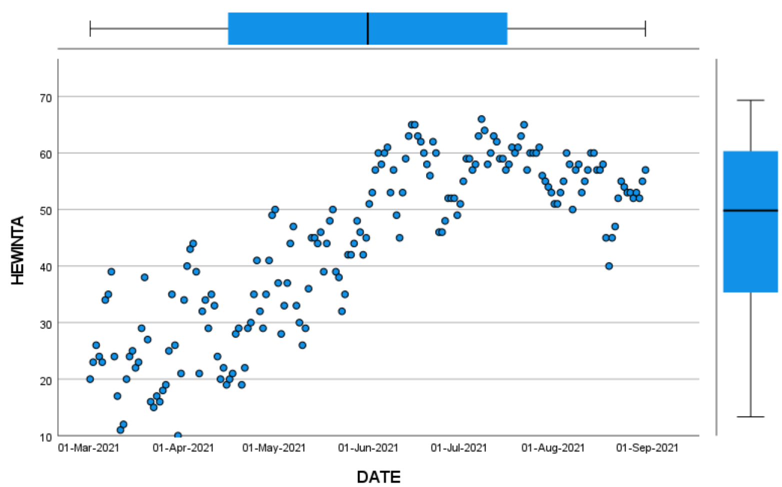
I.PLOT THE DATA:



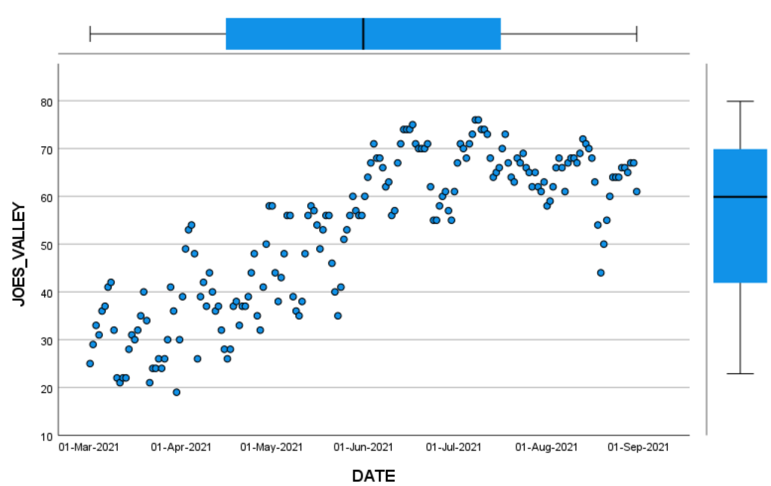




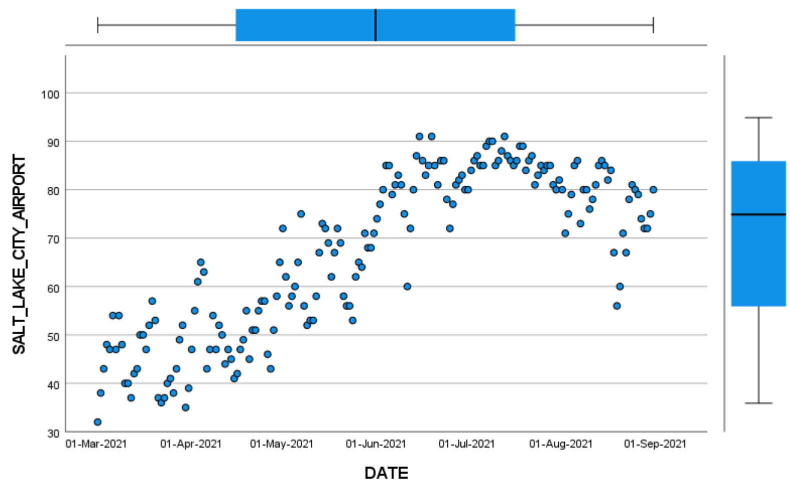
MONTE_CRISTO
70
60
50
40
30
20
10
01-Mar-2021
01-Apr-2021
01-May-2021
01-Jun-2021
01-Jul-2021
01-Aug-2021
01-Sep-2021
DATE



HEWNTA
70
60
50
40
30
20
10
01-Mar-2021
01-Apr-2021
01-May-2021
01-Jun-2021
01-Jul-2021
01-Aug-2021
01-Sep-2021
DATE



JOES_VALLEY
80
70
60
50
40
30
20
10
01-Mar-2021
01-Apr-2021
01-May-2021
01-Jun-2021
01-Jul-2021
01-Aug-2021
01-Sep-2021
DATE



II.DESCRIPTIVE STATISTICS: Upon initial inspection, I ran some quick descriptive statistics, as shown in the chart below. Looking at the mean, median, and mode, we can see some similarities and differences when compared to SLC, though nothing is too extreme. I would expect to find some strong correlations in temperature for these cities that show similarities in the mean which would possibly suggest that our model would appear decent at predicting the weather in Salt Lake City.

Descriptive Statistics

		ARAGONITE	OTTER_CREEK	MONTE_CRISTO	JOES_VALLEY	HEWINTA	SALT_LAKE_CITY_AIRPORT
N	Valid	184	184	184	184	184	184
	Missing	0	0	0	0	0	0
Mean		65.07	53.59	47.44	52.52	43.80	66.96
Median		69.00	56.00	50.50	56.00	46.50	71.00
Mode		78[a]	72	65	56[a]	60	85
Std. Deviation		17.213	16.630	14.866	15.740	14.886	16.633
Variance		296.296	276.549	221.013	247.759	221.582	276.654
Skewness		-.307	-.302	-.356	-.428	-.460	-.311

Std. Error of Skewness		.179	.179	.179	.179	.179	.179
Range		61	56	53	57	56	59
Minimum		31	22	16	19	10	32
Maximum		92	78	69	76	66	91
Sum		11973	9861	8729	9664	8059	12320
Percentiles	25	50.25	38.00	35.00	38.00	32.00	52.00
	50	69.00	56.00	50.50	56.00	46.50	71.00
	75	81.00	68.75	60.75	66.00	57.00	82.00

a. Multiple modes exist. The smallest value is shown

II.CORRELATION: I next looked at the correlations. All of the cities daily average temperatures are highly correlated with Aragonite and Otter Creek the highest, and Joe's Valley the lowest. This is consistent with the cities' distance from SLC.

Correlations

		ARAGONITE	OTTER_CREEK	MONTE_CRISTO	JOES_VALLEY	HEWINTA	SALT_LAKE_CITY_AIRPORT
ARAGONITE	Pearson Correlation	1	.988**	.985**	.974**	.969**	.975**
	Sig. (2-tailed)		.000	.000	.000	.000	.000
	N	184	184	184	184	184	184
OTTER_CREEK	Pearson Correlation	.988**	1	.993**	.973**	.978**	.975**
	Sig. (2-tailed)	.000		.000	.000	.000	.000
	N	184	184	184	184	184	184
MONTE_CRISTO	Pearson Correlation	.985**	.993**	1	.975**	.979**	.965**
	Sig. (2-tailed)	.000	.000		.000	.000	.000
	N	184	184	184	184	184	184
JOES_VALLEY	Pearson Correlation	.974**	.973**	.975**	1	.981**	.964**
	Sig. (2-tailed)	.000	.000	.000		.000	.000
	N	184	184	184	184	184	184
HEWINTA	Pearson Correlation	.969**	.978**	.979**	.981**	1	.970**
	Sig. (2-tailed)	.000	.000	.000	.000		.000
	N	184	184	184	184	184	184
SALT_LAKE_CITY_AIRPORT	Pearson Correlation	.975**	.975**	.965**	.964**	.970**	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	
	N	184	184	184	184	184	184

**. Correlation is significant at the 0.01 level (2-tailed).

III. COEFFICIENTS: Now we look at the coefficients for each city and the intercept to develop our model.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	10.661	1.203		8.863	.000
	ARAGONITE	.504	.091	.522	5.529	.000
	OTTER_CREEK	.682	.133	.682	5.112	.000
	MONTE_CRISTO	-.782	.140	-.699	-5.592	.000
	JOES_VALLEY	.065	.084	.062	.776	.439
	HEWINTA	.470	.094	.421	4.997	.000

a. Dependent Variable: SALT_LAKE_CITY_AIRPORT

Our predictive model for the weather in SLC is:

$$y = 10.661 + 0.504x_1 + 0.682x_2 - 0.782x_3 + 0.065x_4 + 0.470x_5$$

From the coefficient chart calculation, we see that all of the cities have a significant p-value with the exception of Joe's Valley, which is greater than 0.05 with a value of 0.439. This is not surprising as Joe's Valley had the lowest correlation, is the greatest distance from Salt Lake, and the greatest difference in mean from Salt Lake's mean.

We can interpret these coefficients and their relationship to be as follows: for every degree change in temperature in Aragonite, the temperature will increase 0.504 degrees in Salt Lake. For every degree change in temperature in Monte Cristo, the temperature will decrease 0.782 degrees in Salt Lake, and so on.

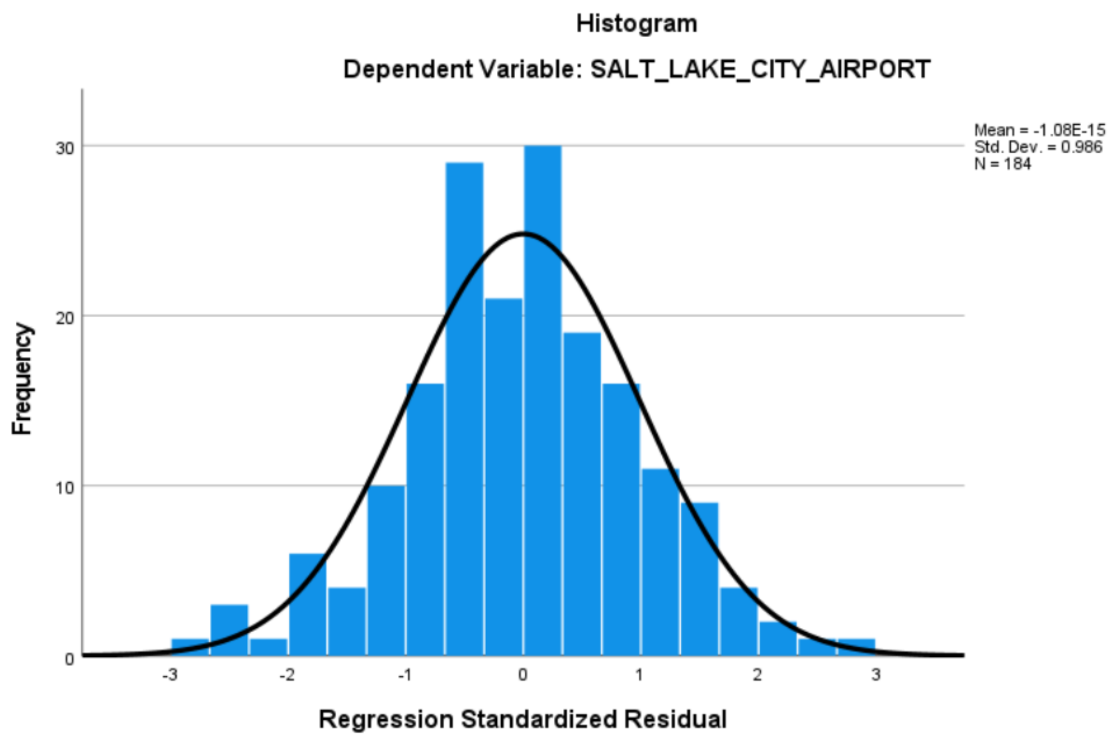
We see that there exists multicollinearity between the predictor variables as the VIF is larger than 10.00 for all of the variables in the regression model. This tells us that each of the variables are ultimately telling us the same thing, which is expected when they are all so highly correlated. Picking cities farther apart with more extremes in temperature would probably change that.

IV: PLOT RISIDUALS: In the chart below, we see information about the residuals followed by plots.

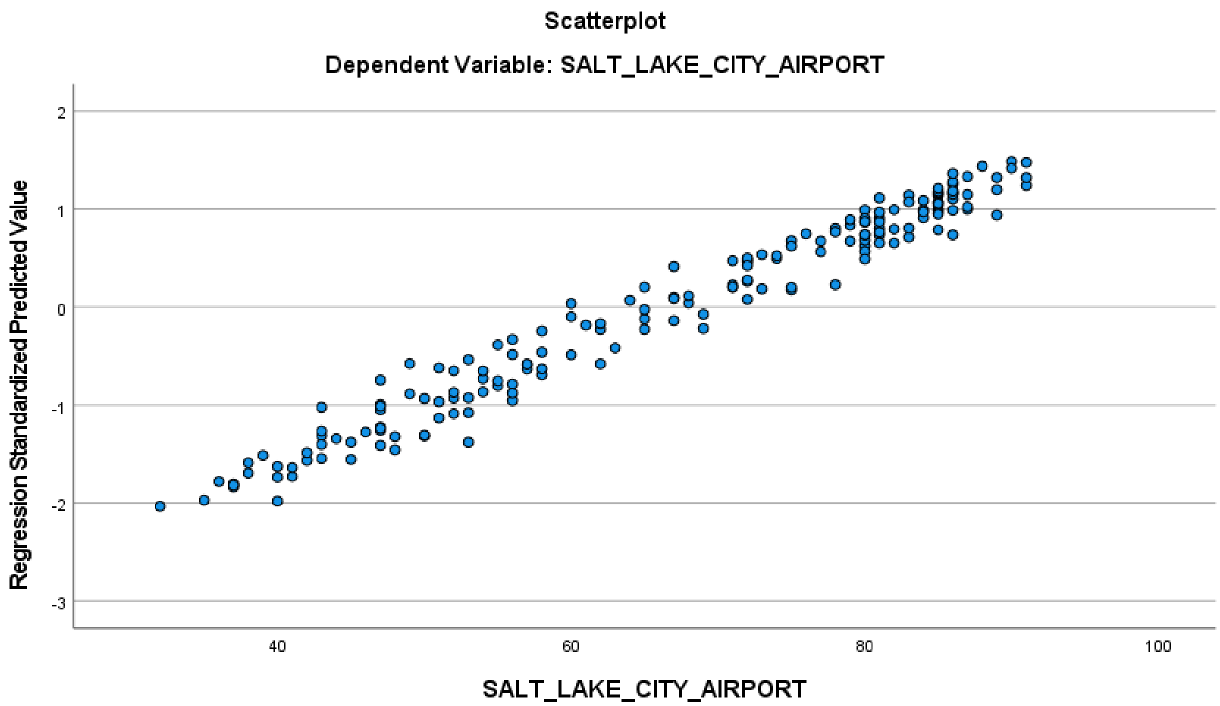
Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	33.70	91.29	66.96	16.358	184
Residual	-8.523	8.607	.000	3.014	184
Std. Predicted Value	-2.033	1.487	.000	1.000	184
Std. Residual	-2.789	2.817	.000	.986	184

a. Dependent Variable: SALT_LAKE_CITY_AIRPORT

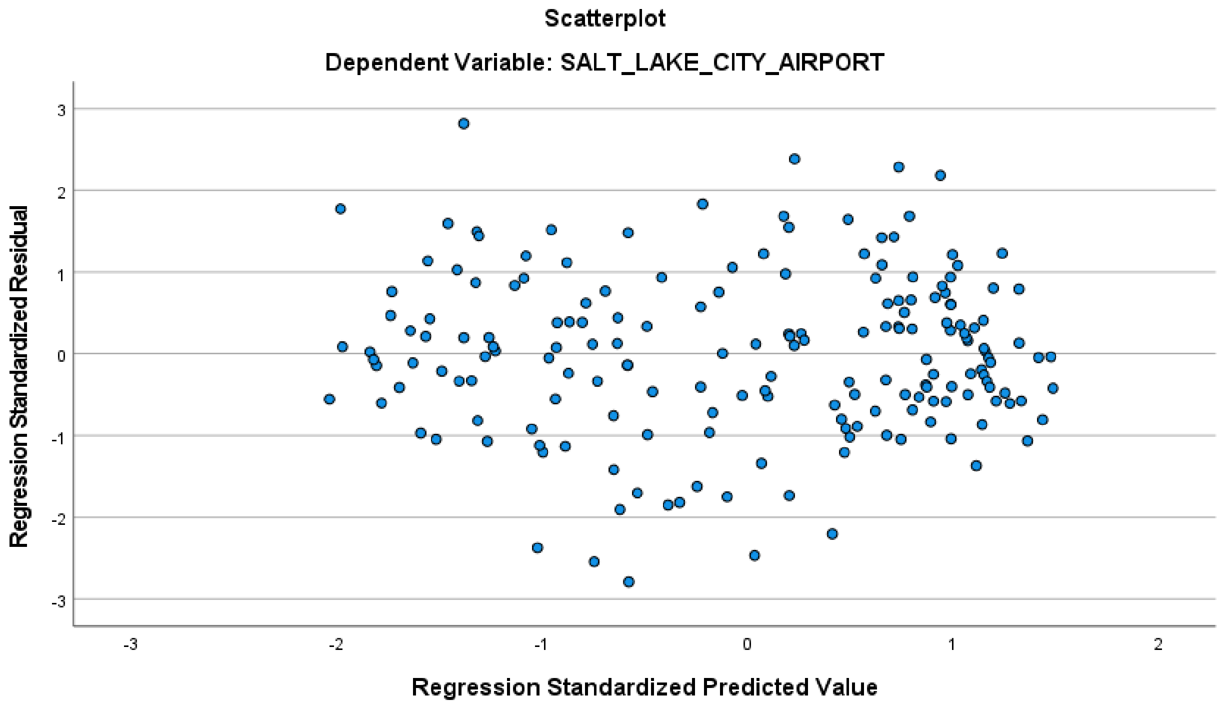


Here we see the daily average temperatures in Salt Lake City plotted with the fitted values.

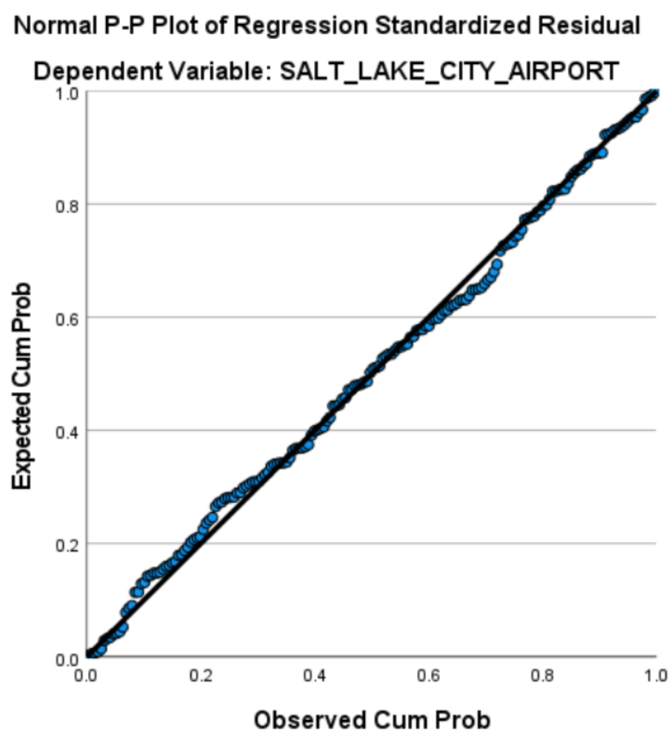


V: NORMALITY:

To test normality, we have the assumption of **independence** and **homoscedasticity** of the residuals. They are both tested in the same way. We want to plot the standardized residuals on the y and the standardized predicted values on the x axis.



Our plot looks like we have met the assumptions of independence and homoscedasticity. The following plot shows that we have very close to normal conditions.



Furthermore, another test shows

Tests of Normality

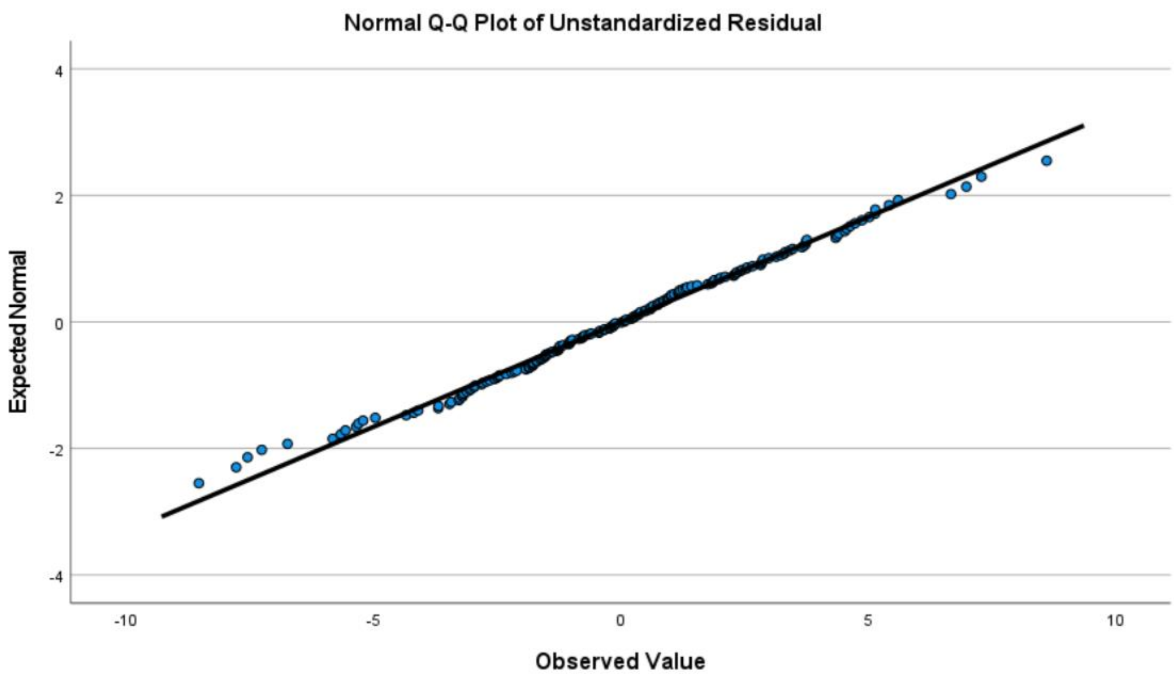
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Unstandardized Residual	.041	184	.200*	.995	184	.753
Standardized Residual	.041	184	.200*	.995	184	.753

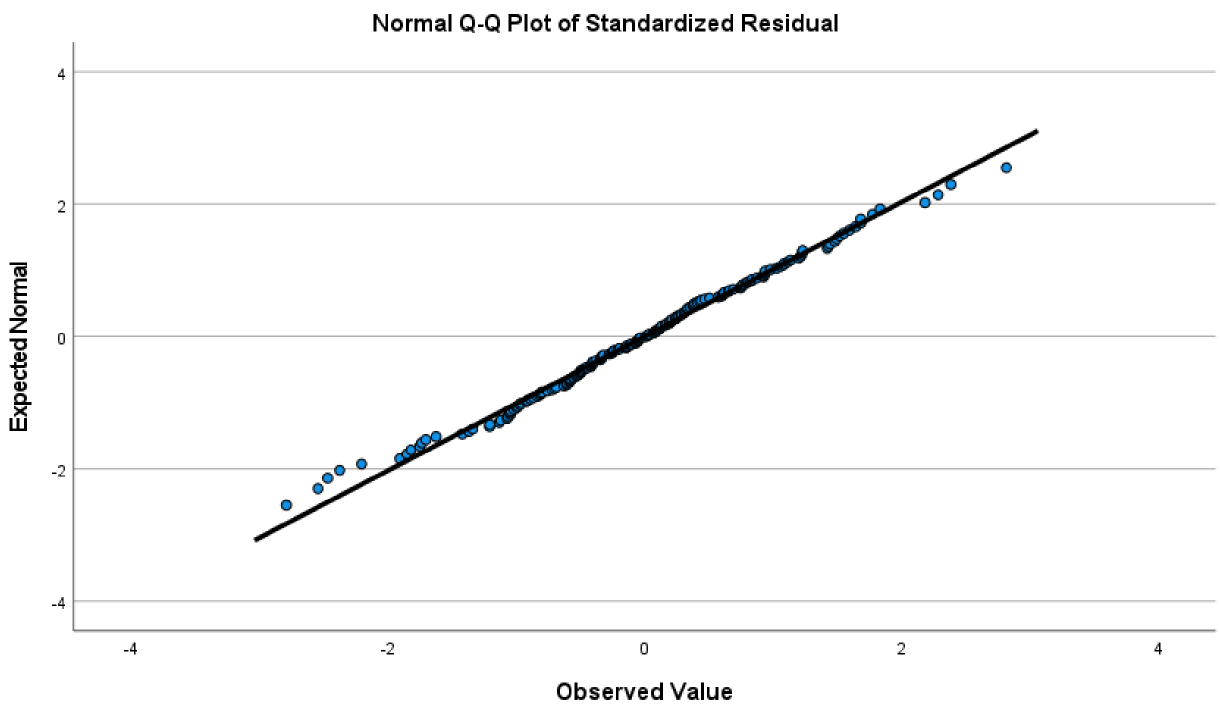
*. This is a lower bound of the true significance.

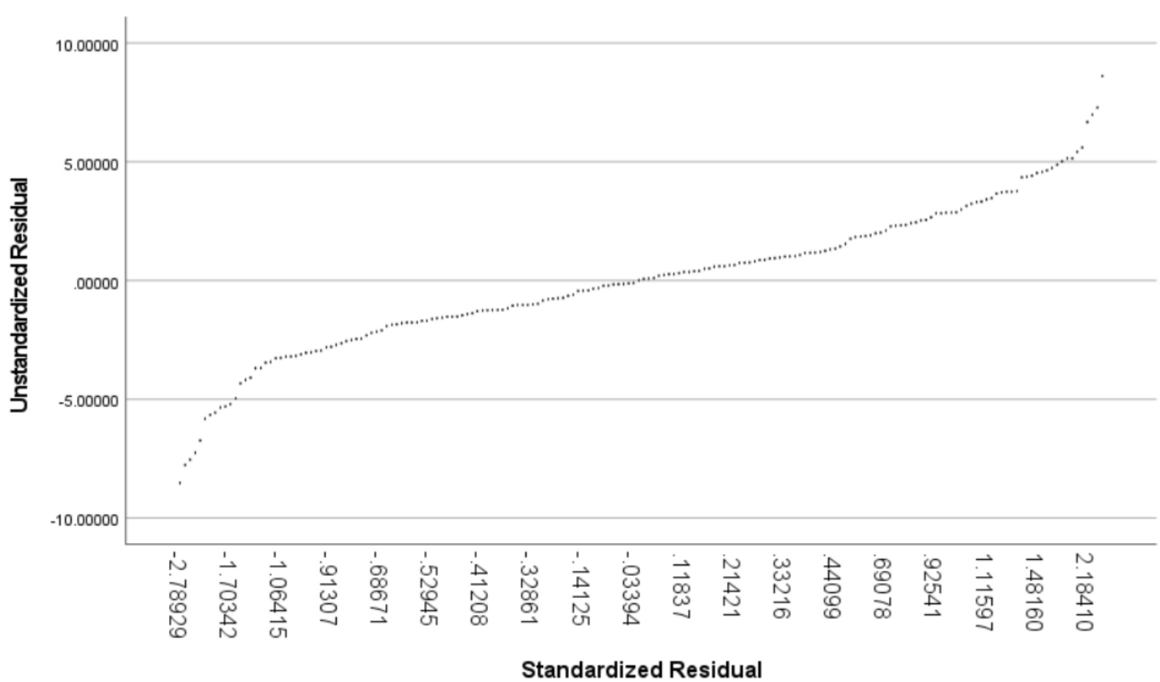
a. Lilliefors Significance Correction

Because we have huge p-values for both Standardized and Unstandardized Residuals regarding the Kolmogorov-Smirnov and Shapiro-Wilk tests, we can assume normality.

And finally, Q-Q plots to confirm.




Normal Q-Q Plot of Standardized Residual
4
2
0
-2
-4
Expected Normal
-4
-2
0
2
4
Observed Value

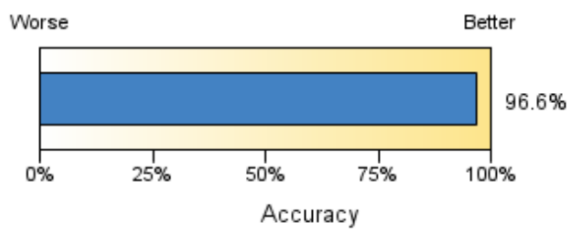

10.00000
5.00000
.00000
-5.00000
-10.00000
Unstandardized Residual
-2.78929
-1.70342
-1.06415
-.91307
-.68671
-.52945
-.41208
-.32861
-.14125
-.03394
.11837
.21421
.33216
.44099
.69078
.92541
1.11597
1.48160
2.18410
Standardized Residual

V: ACCURACY: Even though weather is random and hard to predict, the model summary shown below suggests that the model accuracy is quite good at predicting the average weather in Salt Lake City.

Model Summary

Target	SALT_LAKE_CITY_AIRPORT
Automatic Data Preparation	On
Model Selection Method	Forward Stepwise
Information Criterion	415.919

The information criterion is used to compare to models. Models with smaller information criterion values fit better.



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.983[a]	.967	.966	3.056

a. Predictors: (Constant), HEWINTA, ARAGONITE, JOES_VALLEY , MONTE_CRISTO, OTTER_CREEK

b. Dependent Variable: SALT_LAKE_CITY_AIRPORT

Data Found at https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00024127/detail

Code: (excluding what I forgot to paste)

```
GET DATA
  /TYPE=XLSX
```

```
/FILE='\\Client\C$\Users\ahdes\Documents\6010\WeatherRegression.xlsx'
  /SHEET=name 'Sheet1'
  /CELLRANGE=FULL
  /READNAMES=ON
  /DATATYPEMIN PERCENTAGE=95.0
  /HIDDEN IGNORE=YES.
EXECUTE.
DATASET NAME DataSet1 WINDOW=FRONT.

DATASET ACTIVATE DataSet1.
FREQUENCIES VARIABLES=ARAGONITE OTTER_CREEK MONTE_CRISTO JOES_VALLEY
HEWINTA SALT_LAKE_CITY_AIRPORT
  /NTILES=4
  /STATISTICS=STDDEV VARIANCE RANGE MINIMUM MAXIMUM MEAN MEDIAN MODE
SUM SKEWNESS SESKEW
  /ORDER=ANALYSIS.

CORRELATIONS
  /VARIABLES=ARAGONITE OTTER_CREEK MONTE_CRISTO JOES_VALLEY HEWINTA
SALT_LAKE_CITY_AIRPORT
  /PRINT=TWOTAIL NOSIG FULL
  /MISSING=PAIRWISE.


*Automatic Linear Modeling.
LINEAR
  /FIELDS TARGET=SALT_LAKE_CITY_AIRPORT INPUTS=DATE
  /BUILD_OPTIONS OBJECTIVE=STANDARD USE_AUTO_DATA_PREPARATION=TRUE
CONFIDENCE_LEVEL=95
    MODEL_SELECTION=FORWARDSTEPWISE CRITERIA_FORWARD_STEPWISE=AICC
REPLICATE_RESULTS=TRUE SEED=54752075
  /ENSEMBLES COMBINING_RULE_CONTINUOUS=MEAN COMPONENT_MODELS_N=10.
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS BCOV R ANOVA COLLIN TOL ZPP
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT SALT_LAKE_CITY_AIRPORT
  /METHOD=ENTER ARAGONITE OTTER_CREEK MONTE_CRISTO JOES_VALLEY HEWINTA
  /SCATTERPLOT=(*ZRESID ,*DRESID) (*ZRESID ,*SRESID) (*ZRESID
,*DRESID) (*ZRESID ,*SDRESID)
  /RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID).

STATS WEIBULL PLOT TIME=DATE FAILURE="F" SUSPENSION="S"
/OPTIONS TITLE="Weibull Probability Plot"
/SAVE  FILEMODE=OVERWRITE.
```

```
STATS REGRESS PLOT YVARS=ARAGONITE OTTER_CREEK MONTE_CRISTO
JOES_VALLEY HEWINTA
    SALT_LAKE_CITY_AIRPORT XVARS=DATE
/OPTIONS CATEGORICAL=BARS TITLE="Actual Data" GROUP=1 BOXPLOTS
INDENT=15 YSCALE=75
/FITLINES APPLYTO=TOTAL.

STATS REGRESS PLOT YVARS=ARAGONITE OTTER_CREEK MONTE_CRISTO
JOES_VALLEY HEWINTA
    SALT_LAKE_CITY_AIRPORT XVARS=DATE COLOR=DATE
/OPTIONS CATEGORICAL=BARS TITLE="Actual Data" GROUP=1 BOXPLOTS
INDENT=15 YSCALE=75
/FITLINES APPLYTO=TOTAL.

LINEAR
  /FIELDS TARGET=SALT_LAKE_CITY_AIRPORT INPUTS=ARAGONITE OTTER_CREEK
MONTE_CRISTO JOES_VALLEY
    HEWINTA
  /BUILD_OPTIONS OBJECTIVE=STANDARD USE_AUTO_DATA_PREPARATION=TRUE
CONFIDENCE_LEVEL=95
    MODEL_SELECTION=FORWARDSTEPWISE CRITERIA_FORWARD_STEPWISE=AICC
REPLICATE_RESULTS=TRUE SEED=54752075
  /ENSEMBLES COMBINING_RULE_CONTINUOUS=MEAN COMPONENT_MODELS_N=10.

UNIANOVA SALT_LAKE_CITY_AIRPORT BY ARAGONITE OTTER_CREEK MONTE_CRISTO
JOES_VALLEY HEWINTA PRE_1
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /PLOT=PROFILE(PRE_1) TYPE=LINE ERRORBAR=CI MEANREFERENCE=NO
YAXIS=AUTO
  /PRINT DESCRIPTIVE LOF HOMOGENEITY
  /PLOT=RESIDUALS
  /CRITERIA=ALPHA(.05)
  /DESIGN=ARAGONITE OTTER_CREEK MONTE_CRISTO JOES_VALLEY HEWINTA PRE_1

EXAMINE VARIABLES=RES_1 ZRE_1
  /PLOT BOXPLOT STEMLEAF NPPLOT
  /COMPARE VARIABLES
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.
```