Predictive Analysis of Temperature Using NOAA Station Data

Milestone: Final Project Submission

Student 1: Tim Terry

414-534-5552 (Tel of Student 1)

terry.ti@northeastern.edu

Percentage of Effort Contributed by Student 1: 100%

Signature of Student 1: Tim Terry

Submission Date: 04/13/2025

PROBLEM DEFINITION:

The project goal was to determine the pre-processing activities required to develop a predictive model for temperatures across the United States using historical data from the National Oceanic and Atmospheric Administration (NOAA). As the project unfolded, it was apparent that the datasets required extensive pre-processing before they could be effectively utilized for model training. The pre-processing stage was critical to properly handle data quality issues, geospatial variations, measurement inconsistencies, and preparing the data for optimal model performance. As a result of the analysis work performed, a new factor became evident that the selection of the appropriate model to use for predictive forecasting would be dictated by the data characteristics, specifically the distribution of the temperature observations in the NOAA collected data for the year 2024.

ANALYSIS:

Documented in the *Project Proposal Second Submission* and the subsequent Python Notebook submitted for *Mid-Project Review*, the data acquisition, data exploration, data cleaning, and feature extraction processes were performed. The resulting datasets were saved and imported for usage in the final Python Notebook and analysis described below.

In assessing the characteristics of the station geospatial data and the temperature observations, specific tests were performed on the data to help validate some of the assertions made by climate researchers in previous papers. In particular, *"Global Trends of Measured Surface Air Temperature"* by James Hansen and Sergej Lebedeff published November 20, 1987, is often cited as providing the fundamental presumptions that have led to core components of historical weather and temperature models by NOAA.

An assertion made by Hansen and Lebedeff is that temperature estimates can be derived for areas of the earth with sparse coverage due to lack of monitoring stations or lower than expected observations. This assertion was also a core component of analyzing temperature change in the paper *"Global Surface Temperature Change"* by Hansen, Ruedy, Sato, and Lo published June 10, 2010.

This is an important element when considering using nearest-neighbor stations performing missing data imputation. In the papers, Hansen states that the temperature correlation coefficient was within the range 0.5 – 0.6 for stations within 1,000km and that the average correlation coefficient was 0.5 at distances up to 1,200km.

Another assertion made by Hansen is that global, hemispheric, and regional temperature trends are based on annual-mean temperature changes. This involves taking the average temperature across all samples for the basis of model inputs. The assumption would have to be that temperature observations follow a normal distribution. This has ramifications for the models that are used to predict temperature changes.

Therefore, tests were performed on the data to determine a method for spatial interpolation to handle missing data, correlational analysis was performed to determine whether nearest-neighbor approaches would be appropriate and/or reliable, and tests for normalcy across all stations was performed to determine whether models requiring an assumption of normally distributed data are appropriate.

Methods:

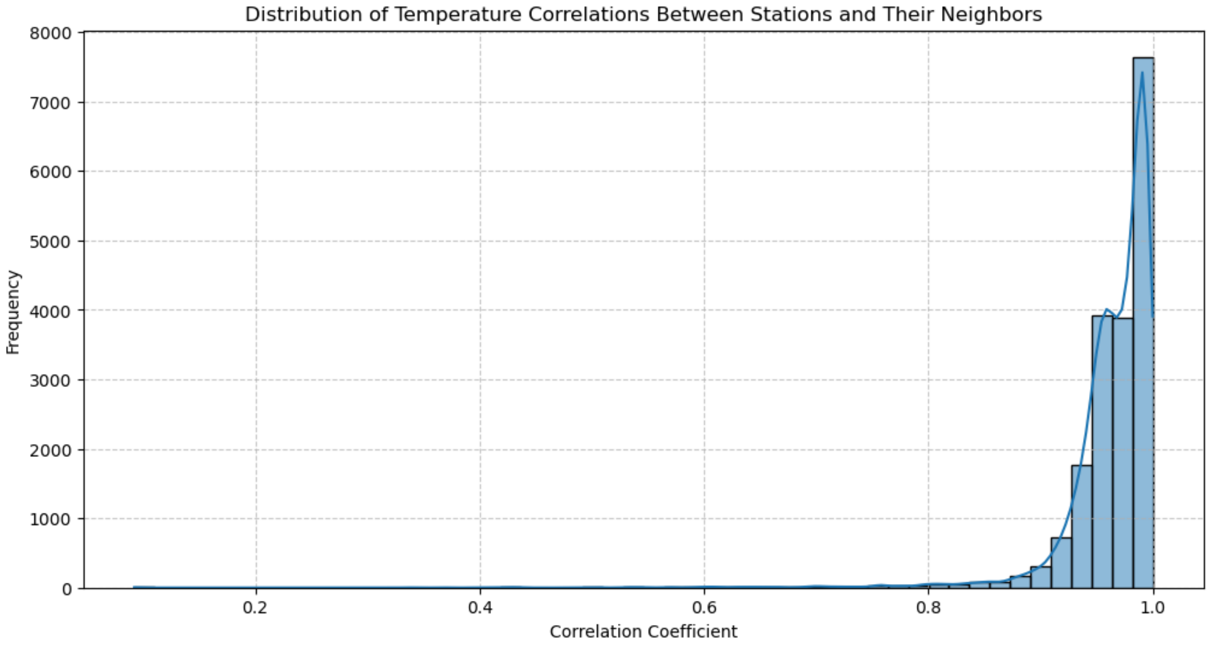
The first task was to develop a strategy for handling missing temperature observations identified in the previous EDA. The actual observations compared to the potential observations were 94.39%, meaning over 5% of the data needed for modeling was missing. The strategy for addressing missing data was to use the individual station's latitude and longitude values to determine nearest-neighbor proximity and assign weights based on distance using the Haversine Distance. The parameters for generating the nearest-neighbors weighted table was k=5 (5 nearest neighbors) and max_distance=500km. The result would provide for the most-likely neighbor stations with the highest correlation coefficient and reliable weights for spatial imputation.

Example of neighbor weights for 3 stations:

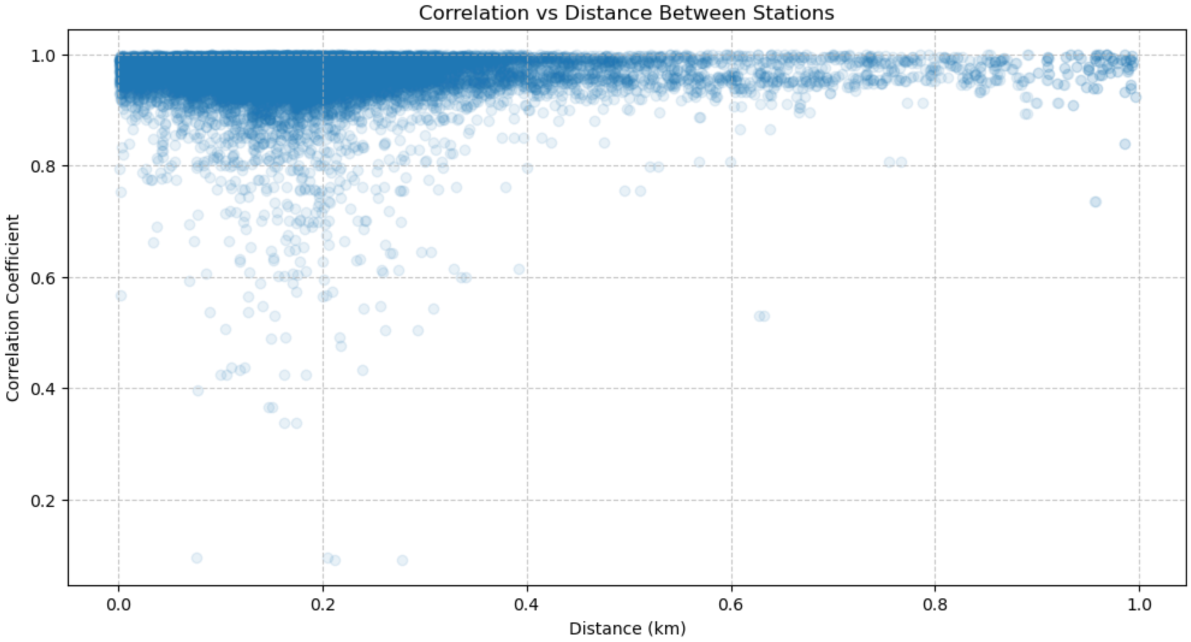
```
Station US009052008 neighbors:
  - USW00004990 (distance: 0.91 km, weight: 0.8678)
  - USW00014944 (distance: 17.93 km, weight: 0.0438)
  - USC00397666 (distance: 17.95 km, weight: 0.0438)
  - USC00392984 (distance: 35.02 km, weight: 0.0224)
  - USC00217012 (distance: 35.44 km, weight: 0.0222)
Station USC00010063 neighbors:
  - USR0000ABAN (distance: 18.66 km, weight: 0.2722)
  - USC00012386 (distance: 22.95 km, weight: 0.2213)
  - USC00018812 (distance: 26.24 km, weight: 0.1936)
  - USC00012840 (distance: 32.17 km, weight: 0.1579)
  - USC00015635 (distance: 32.78 km, weight: 0.1549)
Station USC00010148 neighbors:
  - USW00063866 (distance: 17.65 km, weight: 0.2097)
  - USC00013578 (distance: 18.3 km, weight: 0.2023)
  - USC00013575 (distance: 18.72 km, weight: 0.1977)
  - USC00013573 (distance: 18.73 km, weight: 0.1977)
  - USC00017207 (distance: 19.23 km, weight: 0.1925)
```

This information generated was then used to perform a correlational analysis for each station and its neighbors to determine whether temperature observations followed a positive trend across neighbors. Using the Pearson Correlation Coefficient, 14,269 stations were assessed against its 5 nearest-neighbors using the 2024 temperature data to determine the correlation of temperature observations between stations. This would test Hansen's claim that missing data imputation for sparse regions could be an adequate method when developing models.

Distribution of Correlation Coefficients

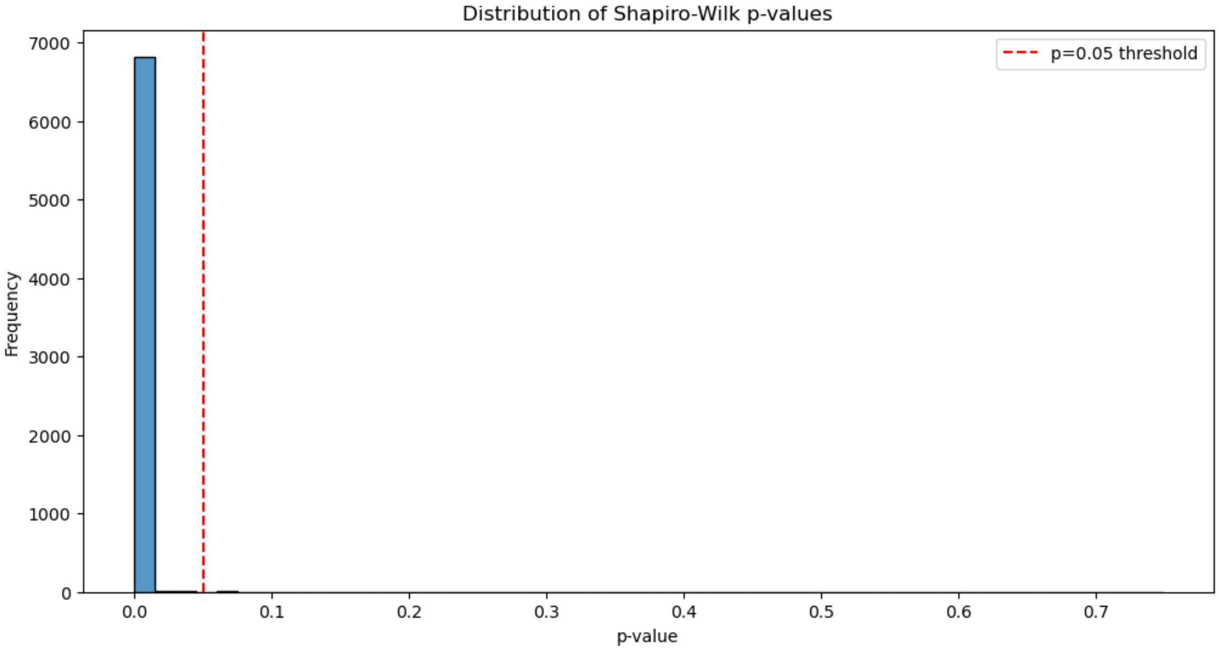


Correlation vs Distance

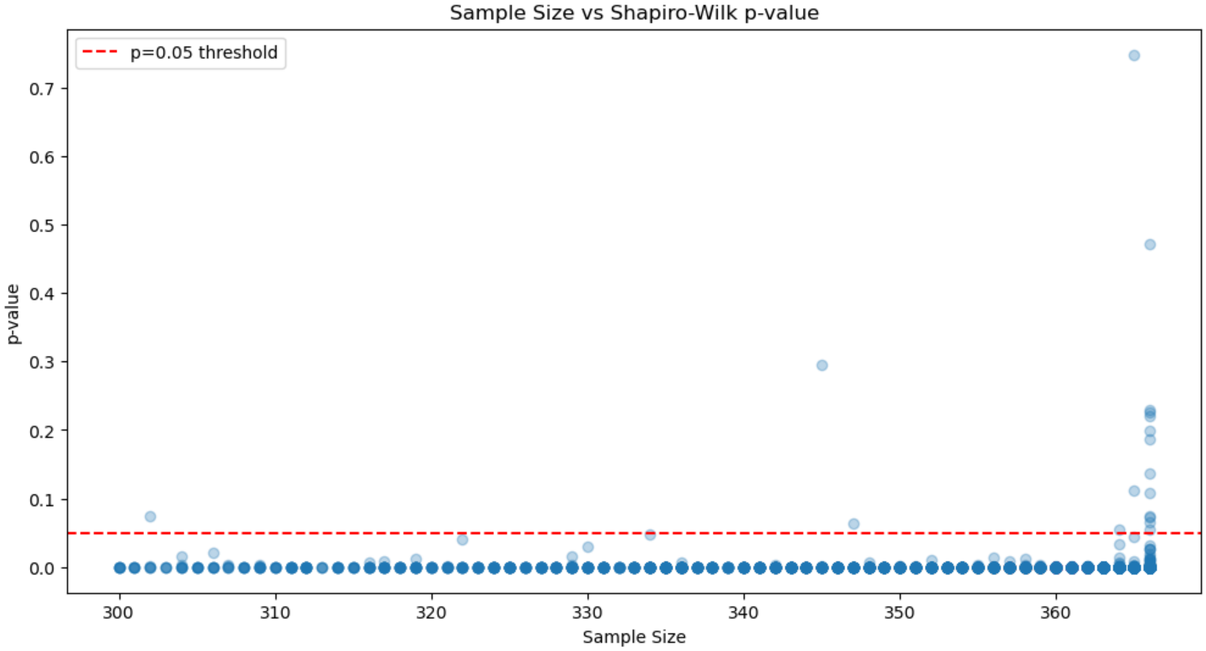


The final test of the temperature data related to the distribution properties of the observations. That is, do the temperature observations follow a normal distribution which allows for the mean value of observations to be appropriate for modeling purposes. For the test of normalcy, a Shapiro-Wilk test was performed on all stations and their respective daily temperature observations.

Distribution of Shapiro-Wilk p-values



Sample Size vs Shapiro-Wilk p-values



Findings:

Nearest Neighbors for Weather Station Data Imputation

The nearest neighbors approach implemented in this project provided a robust method for imputing missing temperature data from weather stations. Here's a summary of how it works:

- Spatial Neighbor Identification: The system identifies nearby weather stations for each station using geographic coordinates (latitude and longitude).
- Distance-Based Weighting: Neighbors are weighted inversely proportional to their distance - closer stations have more influence than distant ones.
- Missing Data Imputation: When a station has missing temperature readings, data from neighboring stations is used to estimate the missing values.
- Preserves Local Climate Patterns: By using nearby stations, the imputation respects local climate variations.
- Handles Sparse Networks: Even stations with few neighbors can benefit from the available data.
- Customizable Parameters: The number of neighbors (k) and maximum distance can be adjusted based on station density.

This approach is particularly valuable for climate research, where complete datasets are essential for trend analysis and climate modeling. The method maintains data integrity by leveraging the spatial correlation of temperature patterns.

Nearest Neighbors Correlation Analysis

The results indicate that nearly all stations have very high average correlations with their nearby stations.

- High Consistency in Temperature Measurements: The fact that the majority of stations have high correlations with their nearby neighbors suggests that the temperature readings are very consistent within a local geographic area. This implies that the spatial patterns in temperature are strong and that neighboring stations tend to exhibit similar temperature trends.
- Reliability for Imputation: In scenarios where data might be missing or incomplete, these high correlations serve as a robust basis for imputing or adjusting values based on spatial neighbors. With nearly all stations showing strong agreement, using information from neighboring stations is likely a reliable approach for estimating missing values.
- Spatial Homogeneity: The high levels of correlation (> 0.9 for over 95% of stations) hint at substantial spatial homogeneity in the temperature signals. Nearly all stations share similar temperature dynamics, which is a promising sign for analyses requiring regional aggregation or interpolation.

The correlation analysis helps support Hansen’s assertion that using nearby neighbors provides a statistically reliable method for handling missing data and addressing sparse observation areas.

Test for Normal Distribution

The results clearly show that the majority of stations (99.7%) fail the Shapiro-Wilk normality test, suggesting that daily average temperatures do not follow a normal distribution at most weather stations. This finding has important implications in terms of proper prediction modeling.

- Mean vs Median: Using the median might be more appropriate than the mean for temperature data. The median is more robust to non-normal distributions and less affected by outliers or skewness in the data.
- Statistical Assumptions: Many statistical methods (like t-tests, ANOVA, and certain regression techniques) assume normality. The results suggest we should be cautious when applying these methods to raw temperature data without transformation.
- Alternative Approaches: If normality is required for prediction modeling, data transformations (log, Box-Cox) may be required. Since seasonality may play a role in the non-normal distribution of observations, models that account for seasonality would be appropriate. Otherwise, non-parametric methods that don't assume normality would be recommended.

Conclusion:

The project identified that while temperature data exhibits strong spatial correlation, making neighbor-based approaches valuable, the underlying distributions are not normal. This suggests that practical applications, especially prediction and imputation efforts, should consider using robust measures like the median and possibly transform or adapt statistical methods to account for the non-normal distribution of observed temperatures.

- High Spatial Consistency: The correlation analysis between neighboring stations revealed extremely high correlations. The majority of station pairs exceed 0.90, indicating that temperature readings tend to be very similar among nearby stations. This strong spatial consistency implies that neighboring observations can be reliably used for imputation or validation purposes.
- Robustness of Neighbor-Based Imputation: Given that nearly all station-neighbor pairs exhibit highly correlated temperature trends even within a radius of 500 km, spatial imputation methods leveraging data from neighboring stations are likely to be effective.
- Non-Normality of Temperature Observations: The normality tests (using the Shapiro-Wilk method) show that almost all stations fail to adhere to a normal distribution (only about 0.30% appear normal). This suggests that the distribution of temperature observations is influenced by factors such as seasonality, skewness, and possibly multimodality, making the use of the mean as a sole metric potentially misleading.
- Implications for Statistical Analysis: Due to the absence of normality in the temperature data, relying on averages might not be optimal. The median, being more robust to skewed or multimodal distributions, could be a better measure for central tendency. Moreover, any statistical analyses or predictive models that assume normality should be approached with caution, and alternatives like non-parametric methods or data transformations might be more appropriate.