Group project on weather data for St. Andrews and Edinburgh

Muntasir Akash Anastasia Akchurina Joshua Arrabaca
Ivan Ben Kristina Haarbeck

2023-11-24

University of
St Andrews

1 Abstract

This report presents a comprehensive statistical analysis of weather information for St Andrews and Edinburgh. Employing time-stamped predictions (November 21 — December 06, 2023; extracted on November 21, 2023), the study aims to understand patterns, correlations, and significant differences between these two geographically proximate localities.

As a first step, we coded and published a Weather Forecast Shiny App based on data from www.open-meteo.com. Secondly, we carried out a Jackknife test to check the reliability of the hourly temperature forecast in St Andrews. The test did not deviate from the forecasted values. Furthermore, the relationship of precipitation and temperature was examined for the location St. Andrews. It was found that higher precipitation is associated with elevated temperatures. A two-sample t-test, strengthened with a permutation test, was run to check any difference in temperature forecasts between St. Andrews and Edinburgh. A subtle difference was found, as temperatures were found to be slightly higher in Edinburgh.

This report and the published Shiny App can offer valuable insights that could inform local weather forecasting, as well as environmental research contributing to the broader understanding of regional climate variations.

This report can be found in the following Github Repository.

2 Introduction

This analysis explores the complex nature of weather patterns and how they affect St. Andrews and Edinburgh. The report focuses on two main outcomes:

1) A **Forecast Weather App** In the context of this report a Weather Forecast app that is focused on St. Andrews and Edinburgh was developed. The app aims to be a user-friendly platform for accessing real-time and forecast weather information. The App aims to provide residents and visitors with reliable forecasts, enhancing their daily planning and activities.

2) A **detailed analysis** Furthermore, the study analyzes in detail how weather patterns in St. Andrews and Edinburgh behave. In detail, it looks at whether temperature in St. Andrews and Edinburgh is significantly different and how precipitation and temperature are linked.

3 Methodology

3.1 Shiny App

3.1.1 Data retrieval

The weather data is retrieved via an API address generated from open-meteo.com, a forecasting service providing free weather data that updates every hour. The API address is automatically configured on the website based on the selected parameters. For the present Shiny app, these are *Hourly Temperature (in °C), Precipitation (in mm), and Wind Speed (in m/s).* Forecast data is available up to 16 days into the future. Weather data is obtained for the locations St. Andrews and Edinburgh in Scotland.

The data is initially retrieved as a JSON file. Then, with the help of the jsonlite and tidyverse libraries in R, it is transformed into an R data frame. The weather variables are vertically combined using the cbind() function. The *date* variable is parsed using the lubridate library's as.POSIXct() to preserve the date and times, while the rest of the fields are parsed into numeric. This process is done separately for each of the two locations (St. Andrews and Edinburgh).

3.1.2 Refreshing and download of data

Refreshing the data is done via observeEvent which calls on the above to retrieve the weather JSON for St. Andrews and Edinburgh, as well as display the current time via RefreshTime function. Downloading the data is possible with the downloadHandler() and write.csv().

3.1.3 Choice of location

The location data displayed is determined by the user's selection. The dataset (St. Andrews vs. Edinburgh) used for the Shiny App is reactively configured based on the user's choice through a reactive expression.

3.1.4 Display of current weather data

To display the current hour's weather information, the dataset is filtered to include only the entry that corresponds to the current hour and date by subsetting the entry which matches Sys.time(), rounded to the nearest hour.

3.1.5 Display of forecast weather data

The Weather Forecast tab incorporates input fields for users to select a forecast date and time. These input fields include a date picker dateInput and a dropdown menu selectInput. To display weather information at the selected date and time, the dataset is subsetted to include only the entry that corresponds to the chosen point in time.

An overview over the 16-day hourly weather forecast is given in line plots which are integrated into a tabsetPanel with three tabs: "Temperature", "Precipitation" and "Wind Speed".

The plots tempPlot, precPlot, and windPlot are defined using the renderPlot function. ggplot2 package is used to create line plots. The chosen_data() function is employed to dynamically select the relevant data. The plots are customised with the aes function for aesthetics. The line colours were chosen to match the app colour scheme, the plot also includes axis labels, grid lines, and background changes for improved visualisation.

3.1.6 App aesthetics

The Shiny App's layout is established using the "sandstone" theme within the shinythemes package.

The dashboardHeader function is used to set the app's title to "Weather Forecast for East Scotland" with a specified title width.

Within the dashboardSidebar function a drop-down menu selectInput is created to enable users to choose between the locations St. Andrews and Edinburgh. Menu items for Current Weather and Weather Forecast are implemented using sidebarMenu.

The sidebar includes a refresh button actionButton and a download button downloadButton. The Refresh Data button triggers the observeEvent function updating the weather data for St Andrews and Edinburgh. The last refresh time is displayed through textOutput.

The tabItems function organises two-tab items: Current Weather (tabName = c_weather) and Weather Forecast (tabName = f_weather). These tabs organize the user-friendly interface, providing intuitive access to current and future weather information.

To present current and future date, temperature, wind speed, time, and precipitation as well as the forecast plots in a visually appealing way the infoBox function is used.

3.2 Data set (used for Tasks 2-4)

For Tasks 2 to Task 4, common data sets were used, called "WeatherData.csv" (for St Andrews) and "WeatherData2.csv" (for Edinburgh). These datasets were generated on 21-November 2023, with rows for hourly forecasts from 21-Nov 00:00 up to 06-Dec 23:00, and each contain 384 rows of forecasts (excluding the headers).

3.3 Jackknife

For Jackknife, the pandas and numpy packages were required for data analysis, while matplotlib and plotnine used for visualisation.

The Python script reads 'WeatherData.csv' (for St Andrews Weather), and takes the csv Hourly Temperatures as the value of interest. Any non-numeric values are removed with ~no.isnan(), then the mean is calculated with np.mean().

The jackknife function creates an empty array to store the mean of samples. Then, it iterates over the length of x, and creates a sample with the value at the ith index removed (McIntosh (2016)). The mean for this sample is stored in the array, then the process is repeated until the end of x. The array of means is then returned.

Another function computes for the residuals of the array of means. This function iterates over the length of x, then finds the difference of the mean of x, over each mean in the array. This difference (the residual) is then stored in the residuals array.

Both the jackknife sample means and residuals are visualised in a histogram using plotnine.

3.4 Bootstrap Regression

The task Bootstrap Regression comprises two aspects: 1) *Exploration of the relationship* between hourly precipitation (explanatory variable) and hourly temperature (outcome variable) through visualisation and linear model fitting and 2) Obtainment of *bootstrap confidence interval* for the regression parameters of this linear relationship. For the Bootstrap Regression the St. Andrews data set is used.

3.4.1 Exploration of the relationship between hourly precipitation and hourly temperature

As a first step, required libraries (Pandas, Statsmodels, Plotnine, and Scipy) are imported.

Secondly, a scatter plot is created to visually explore the relationship and a line is fitted to the data.

Finally, a linear regression model of the following format is built using the Statsmodels library:

Hourly Temperature $= m \times$ Hourly Precipitation $+ n$ (m is the regression slope coefficient and n the intercept)

To assess the model's meaningfulness, assumptions of the model are assessed.

- *Linear relationship of dependent and independent variable*
- *Normal distribution of residuals*

- *Constant variance of residuals*
- *Uncorrelatedness of residuals*

Figure 3 is used to assess the linearity assumption. As a clear pattern is visible it can be concluded that the assumption is not met.

The normality assumption is assessed by the Shapiro-Wilk test as well as the histogram and the quantile-quantile plot of residuals. The Shapiro-Wilk test yiels a highly significant p-value of $6.816 * 10^{-}8$ which indicates non-normality. This assumption is supported by the histogram of residuals (Figure 1) and the quantile-quantile plot (Figure 2) of residuals.

The constant variance assumption is examined by the Breusch-Pagan test and by the variance plot (Figure 3). The Breusch-Pagan test results in a very significant p-value of $7.534 * 10^{-}5$, suggesting absence of constant variance. This observation is supported by the variance plot which shows that the variance of residuals changes along the fitted values.

The Durbin-Watson test was conducted to assess the assumption of independent residuals. The resulting test statistic of 0.029 suggests the presence of positive autocorrelation. Consequently, it indicates a deviation from the assumption.

```
[('Lagrange multiplier statistic', 15.671824287384482),
 ('p-value', 7.53381611302326e-05),
 ('f-value', 16.25354038202031),
 ('f p-value', 6.69017677511197e-05)]
```

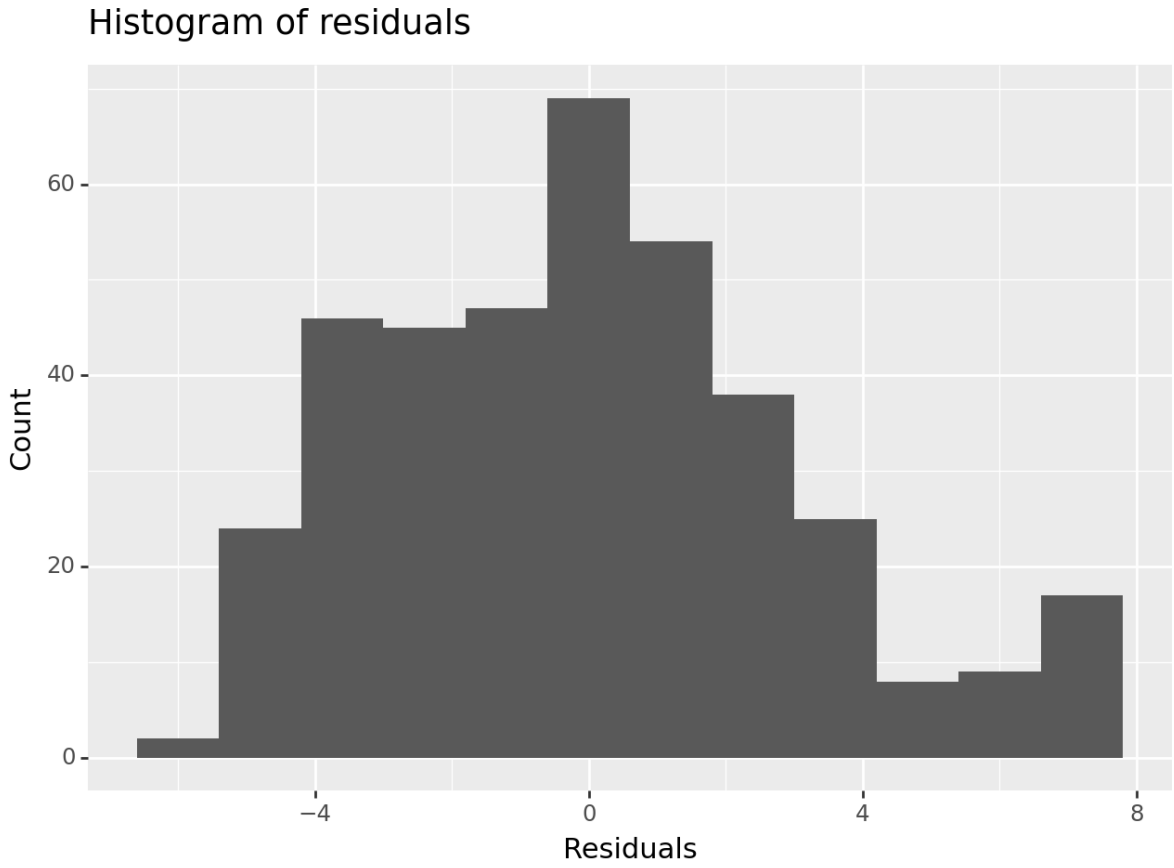

Figure 1: Histogram of residuals

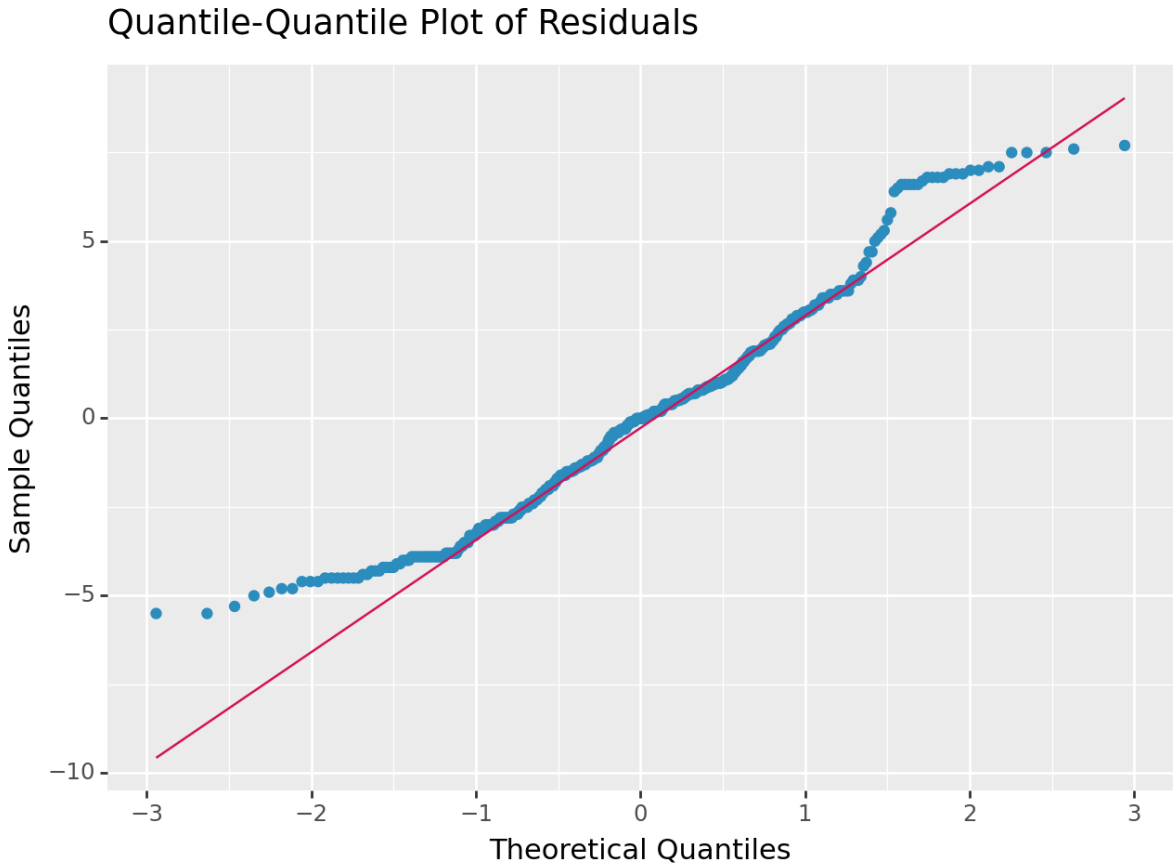

Figure 2: Quantile-Quantile Plot of Residuals

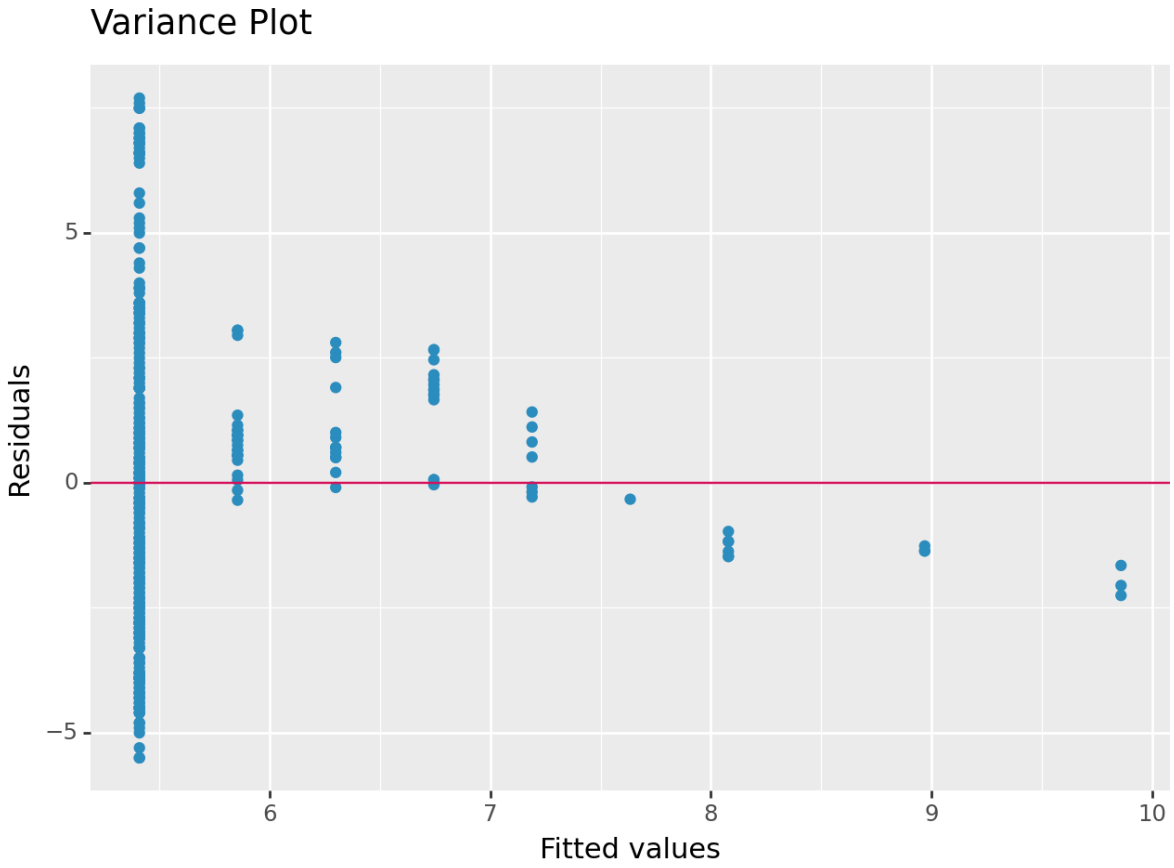

Figure 3: Fitted values vs. Residuals

3.4.2 Obtainment of bootstrap confidence interval for the regression parameters

As the assumptions for the linear model were not met, bootstrapping was applied to obtain reliable insights into the relationship of precipitation and temperature. Bootstrapping was conducted in five steps.

- Firstly, the number of bootstrap samples was set to 1000
- Secondly, an array called *outboot* was obtained which contains a set of 1000 shuffled indices
- Thirdly, a function called *bootstrap_log_reg* was set up which draws bootstrap samples with replacement from the precipitation and temperature values and fits a lineaer model to the data using the shuffled indices created in step 2
- Fourthly, the resulting coefficients and intercepts are stored in an array which is called boot_stats
- Finally, the np.quantile() function is used to obtain the 95% confidence interval for the coefficient and intercept

3.5 T-Tests

Within the T-Test task the following hypotheses are tested:

$$H0 : HourlyTemperature_{St.Andrews} = HourlyTemperature_{Edinburgh}$$

$$H1 : HourlyTemperature_{St.Andrews} \neq HourlyTemperature_{Edinburgh}$$

To initial insights, the temperature data at the two locations is compared visually using a boxplot.

For formal assessment of the hypotheses, two methods are applied: 1) Permutation test and 2) Obtainment of bootstrap estimates of the difference in means.

3.5.1 Visual exploration

First, the temperatures in St. Andrews and Edinburgh are explored visually using a histogram and boxplot.

3.5.2 Assumption checking t-test

Normality assumption The Shapiro-Wilk test was conducted to assess the normality of a dataset. This is important as most statistical techniques will assume that the dataset follows a normal distribution. Should the data vary significantly from the normal distribution, the validity of these models would be affected. Hence, before proceeding with any parametric tests (which would assume normality)- the Shapiro Wilk Test is conducted. For the temp data of St. Andrews and Edinburgh, the test was deployed to test the null hypothesis that the data are normally distributed against alternative hypothesis that the data doesnt follow a normal distribution. The results for St. Andrews shows a Shapiro-Wilk statistic of 0.975 with a p-value of $4.359 * 10^{-}6$, implying that the temperature data from the St. Andrews significantly deviate from a normal distribution. Likewise, for Edinburgh, the Shapiro-Wilk statistic is 0.972 with a p-value of $8.138 * 10^{-}7$, again showing deviation from normality. As both the p-values are less than the common alpha level of 0.05, the null hypothesis can be rejected for both locations. This means that the hourly temp data for St. Andrews and Edinburgh doesnt fit the normal distribution model.

Equal variance assumption The Levene's test was conducted to evaluate the assumption of equal variances or homoscedasticity between two or more independent groups. For the t-tests, it is valuable to verify this assumption because the validity of the test results depends on it. Generally, t-tests assumes that the variances of the groups are equal- and if this rule is not met, the test results might not be reliable and another version of the t-test would be required (one that doesn't assume equal variances) The test was conducted for hourly temp of St. Andrews and Edinburgh yielding a p-value of 0.211. This value is greater than the usual alpha (0.05), indicating there is no significant difference in the variances between the two groups. The test statistic value of 1.569 further confirming that the variances are similar, as a larger test statistic would imply a greater deviation from equal variances. As the p-value is above the threshold, we fail to reject the null hypothesis of Levene's test. This means it is appropriate to proceed with a standard t-test for comparing the means of St. Andrews and Edin's hourly temp, as the assumption of homoscedasticity is not violated.

3.5.3 Permutation test

To generate reliable insights a permutation testing is applied. The number of permutations is set to 1000.
The operation of the test is as follows:

- Firstly, temperature data from St. Andrews and Edinburgh is combined and randomly permuted
- Secondly, for each of the permuted samples the t-statistic is calculated

- The totality of these values forms a sampling distribution of t-statistics under the assumption that H0 (no difference in temperature between St. Andrews and Edinburgh) is true

Finally, the observed test statistic is placed within this sampling distribution. The p-value is obtained by assessing the proportion of random t-statistics within the obtained distribution which are more extreme then the observed statistic.

3.5.4 Bootstrap estimates of the difference in means

In addition bootstrapping is used to obtain estimates of the difference in means.

First, bootstrap estimates for the mean in each location are calculated. For this purpose, the following steps are conducted for both St. Andrews and Edinburgh temperature data: * Firstly, the number of bootstrap samples was set to 1000 * Secondly, a set of sample means is derived from a normal distribution using the mean and standard deviation obtained from the actual data. * Thirdly, bootstrap samples are generated from the sample obtained in step 2 and stored in an array called "outboot" * Fourthly, the mean of each bootstrap is calculated and stored in an array called "bootmeans"

To compute bootstrap estimates of the difference in means, an array named "bootmeans_difference" is created by subtracting the bootstrap temperature means obtained for Edinburgh from those in St. Andrews.

Lastly, the confint() function is used to obtain the 95% confidence interval for the difference in bootstrap means.

4 Results

4.1 Shiny App

Please find the code to the Shiny App "Weather Forecast for East Scotland" in the following GitHub repository.

The Shiny App is published here.

The Shiny App "Weather Forecast for East Scotland" (as shown in Figure 4) allows the user to access real-time and forecast weather information for the locations Edinburgh and St. Andrews. It provides insights into temperature, wind speed and precipitation, all displayed on an hourly level. The forecast is available for up to 16 days into the future.

Specifically, the App incorporates the following interactive features:

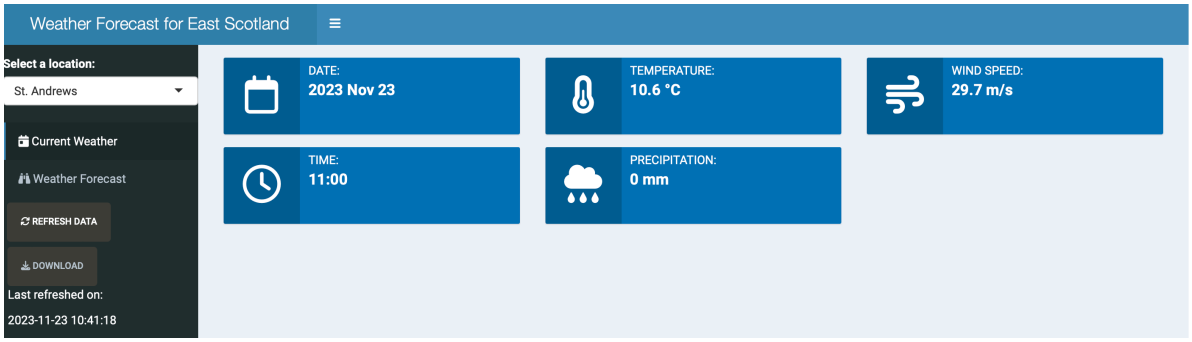

Figure 4: App Overview

- **Selection of location via drop-down**: Users can choose whether they are interested in weather information about St. Andrews or Edinburgh

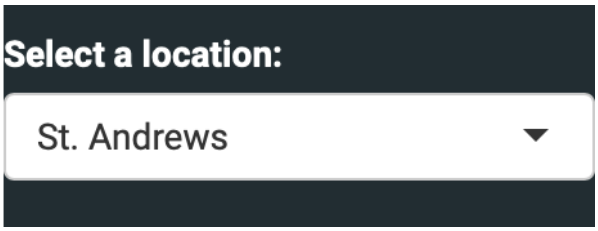

Figure 5: Selection of location

- **Selection of current vs. forecast data via sidebar**: Users can navigate via the sidebar whether they are interested in real-time or forecast information

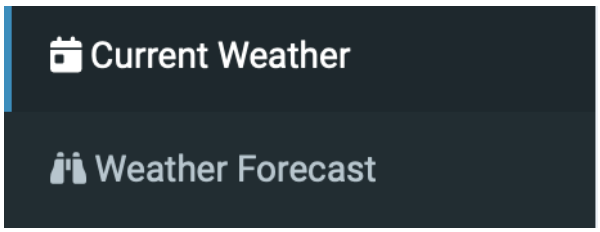

Figure 6: Current vs. forecast data

- **Download button**: Users can download current and forecast data for the chosen location via the respective button which will then be saved to csv with the default filename "WeatherData.csv"
- **Refresh button**: Users can update the data to the most recent information via the designated refresh button. Additionally, the app displays the timestamp of the latest refresh right below the button.

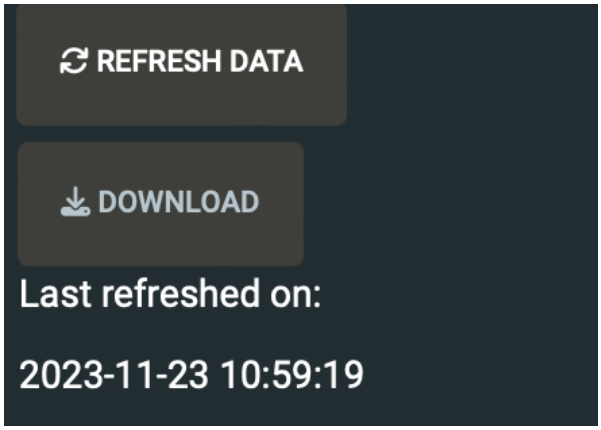

Figure 7: Selection of date and time

- **Selection of forecast data and time**: In the "Weather Forecast" section users can choose a specific date and hour in the future via drop-down menus for which they want to obtain forecast weather information.

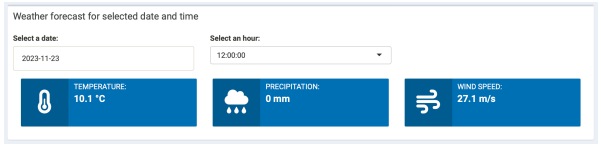

Figure 8: Selection of date and time

- **Selection weather data's type for forecast overview**: The "Weather Forecast" section includes plots that offer a comprehensive view of the 16-day forecast, focusing on temperature, wind speed, or precipitation. Users can switch between these data types using a tab menu.

The information displayed in the App is based on data by open-meteo.com, a service that provides free hourly weather data.

4.2 Jackknife

The Jackknife test is able to take in a vector of values, iterates through the length of the vector, and creates samples with the ith value removed. The test then computes for the mean of each new sample. In this case, the vector is hourly temperatures of St Andrews, and the mean is 5.66° C.

When visualizing the Jackknife sample means, the distribution is slightly skewed (Figure 10). According to Figure Figure 10, the spread ranges from 5.64° C to 5.68° C, with most estimates

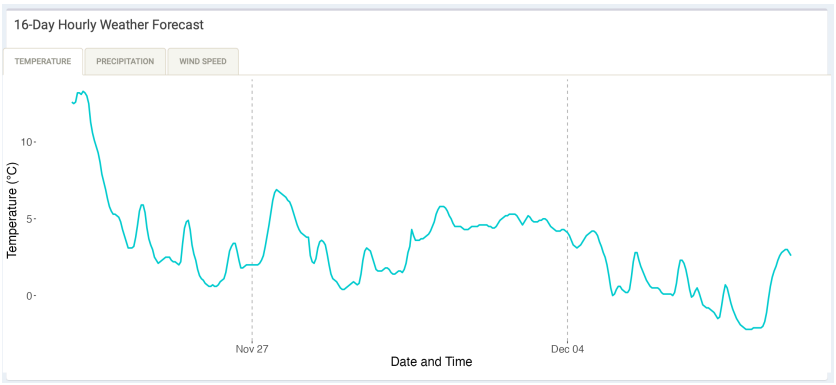

Figure 9: Forecast plots

centered around the true mean of 5.66° C. When looking at the residuals in Figure 11 (i.e., true mean - sample mean), the Jackknife means lie close to the actual mean obtained from the dataset, The spread of the residuals is low (-0.02 to 0.02), with most around 0—indicating, the distribution followed a normal spread.

4.3 Bootstrap Regression

4.3.1 Exploration of the relationship between hourly precipitation and hourly temperature

Figure 12 illustrates the correlation between hourly precipitation and hourly temperature. It becomes evident that 0 m/s precipitation is associated with all temperature levels. However, as precipitation intensifies, the range of associated temperatures narrows. It is evident that precipitation predominantly occurs at moderate temperature levels. Specifically, very high precipitation rates (1 m/s) are observed only within the temperature range of 6 to 8 °C.

To further explore the relationship, a linear model is fitted to the data. The resulting coefficient estimate for precipitation is 4.451 which is significant at the 1% level. The coefficient's corresponding 95% confidence interval is [2.497; 6.405]. The intercept is estimated to be 5.408 with a 95% confidence interval of [5.085; 5.732].

As outlined under "Methodology", the assumptions for a linear model are not met. Therefore, bootstrapping is used to obtain robust insights into the relationship of precipitation and temperature.

4.3.2 Bootstrap confidence interval for the regression parameters

The bootstrapping results in a 95% confidence interval of [3.458; 6.077] for the coefficient (as shown in Figure 13 and [5.083; 5.757] for the intercept (as shown in Figure 14).

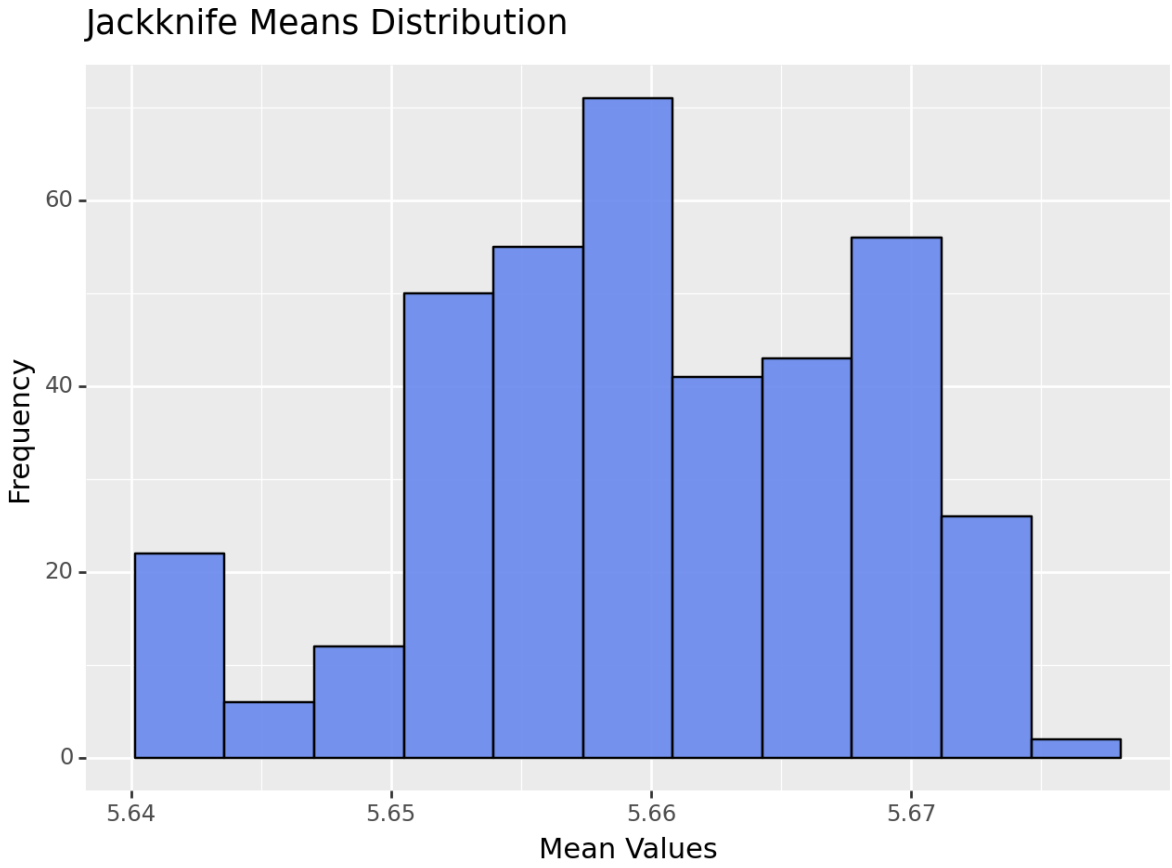

Figure 10: Jackknife Means Distribution

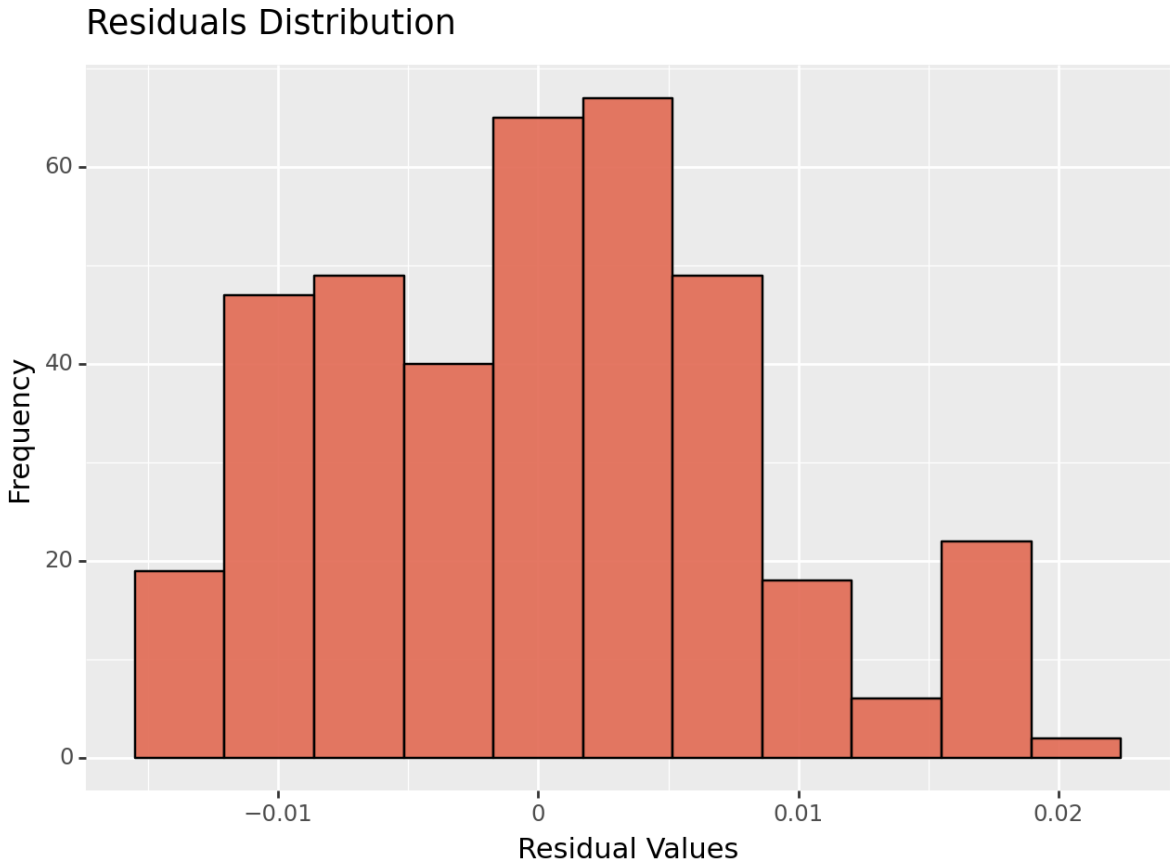

Figure 11: Residuals Distribution

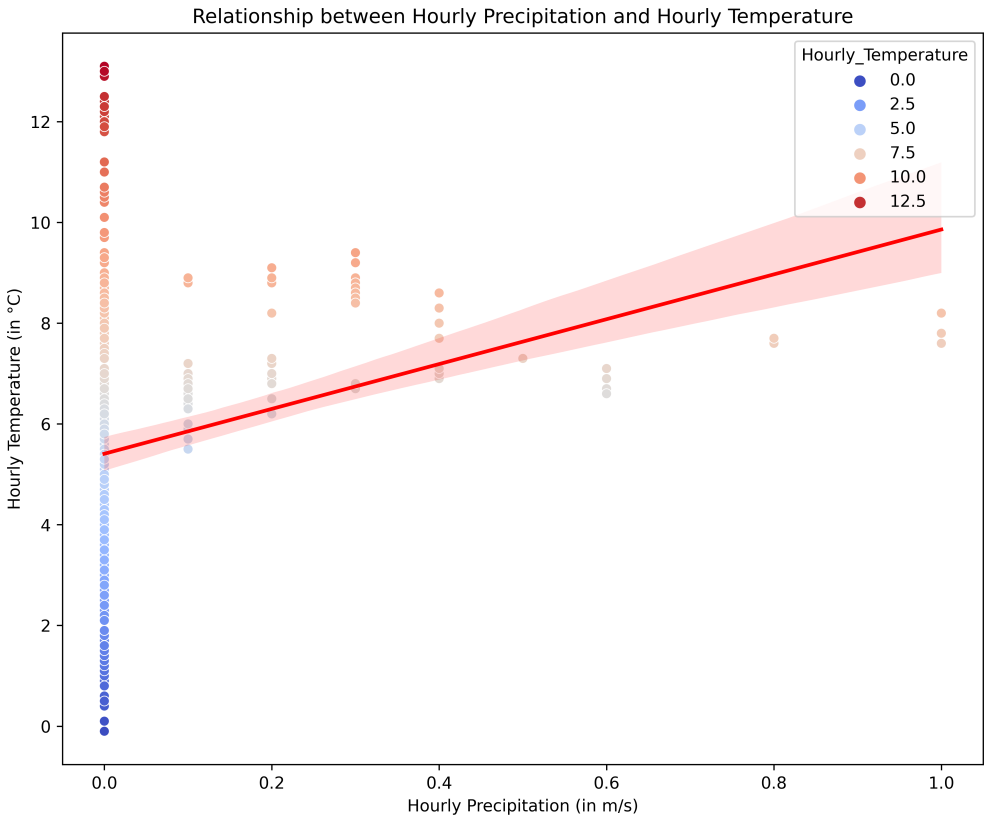

Figure 12: Relationship hourly precipitation and hourly temperature

Please note that the confidence intervals obtained from linear model fitting and the confidence intervals resulting from bootstrapping overlap substantially, indicating that the coefficient and intercept estimate as derived from the linear model are somewhat reliable. Nevertheless, the bootstrap confidence interval is narrower for the coefficient, suggesting that the method is estimating its value more precisely.

The demonstrated positive relationship of precipitation and temperature is in line with prior visual analysis which found that increasing precipitation is associated with moderate levels of temperature. However, there is evidence suggesting that the relationship cannot be adequately explained by a linear model, as there exists an upper (10 °C) and lower (5 °C)limit to temperature when precipitation is present. This proposition is supported by the low adjusted R^2 value of 0.047 found for the fitted value, indicating that precipitation only accounts for a minor proportion of the temperature changes.

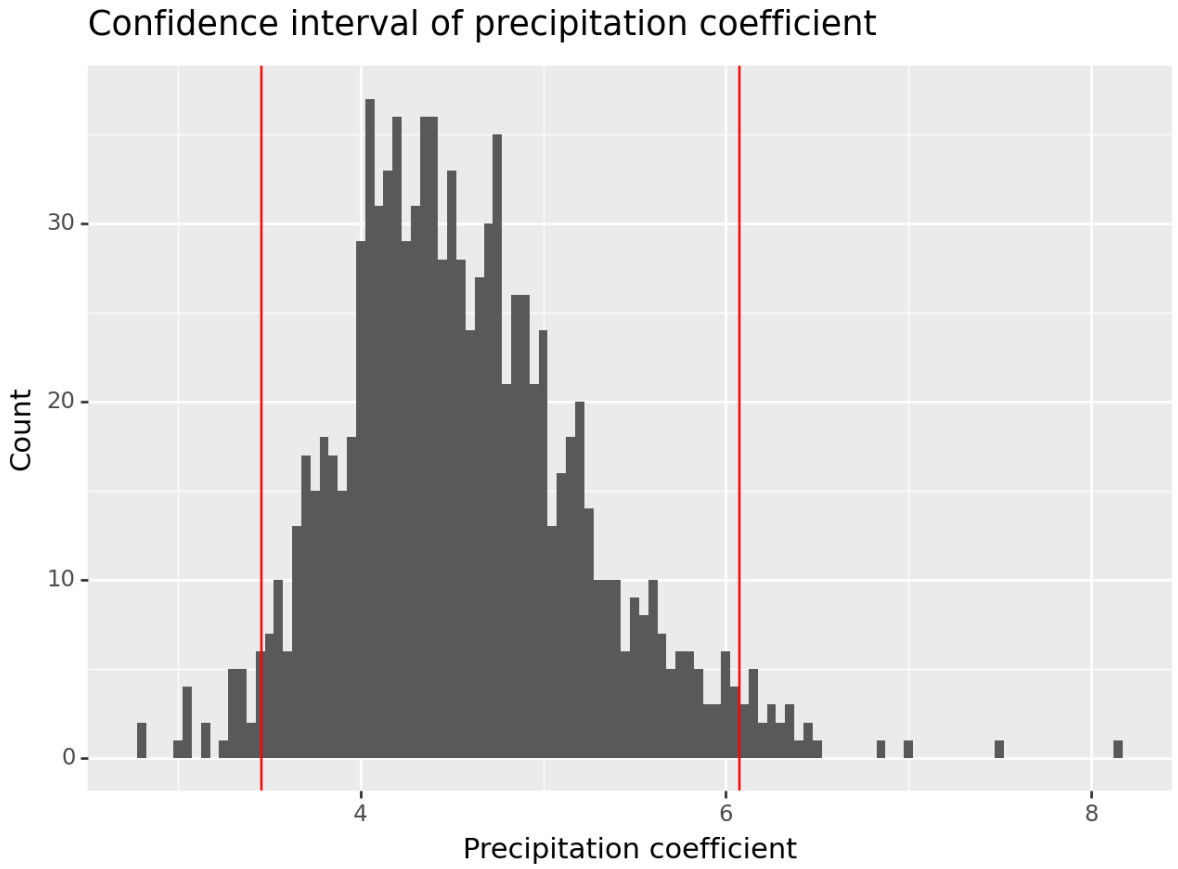

Figure 13: 95% confidence interval for precipitation coefficient

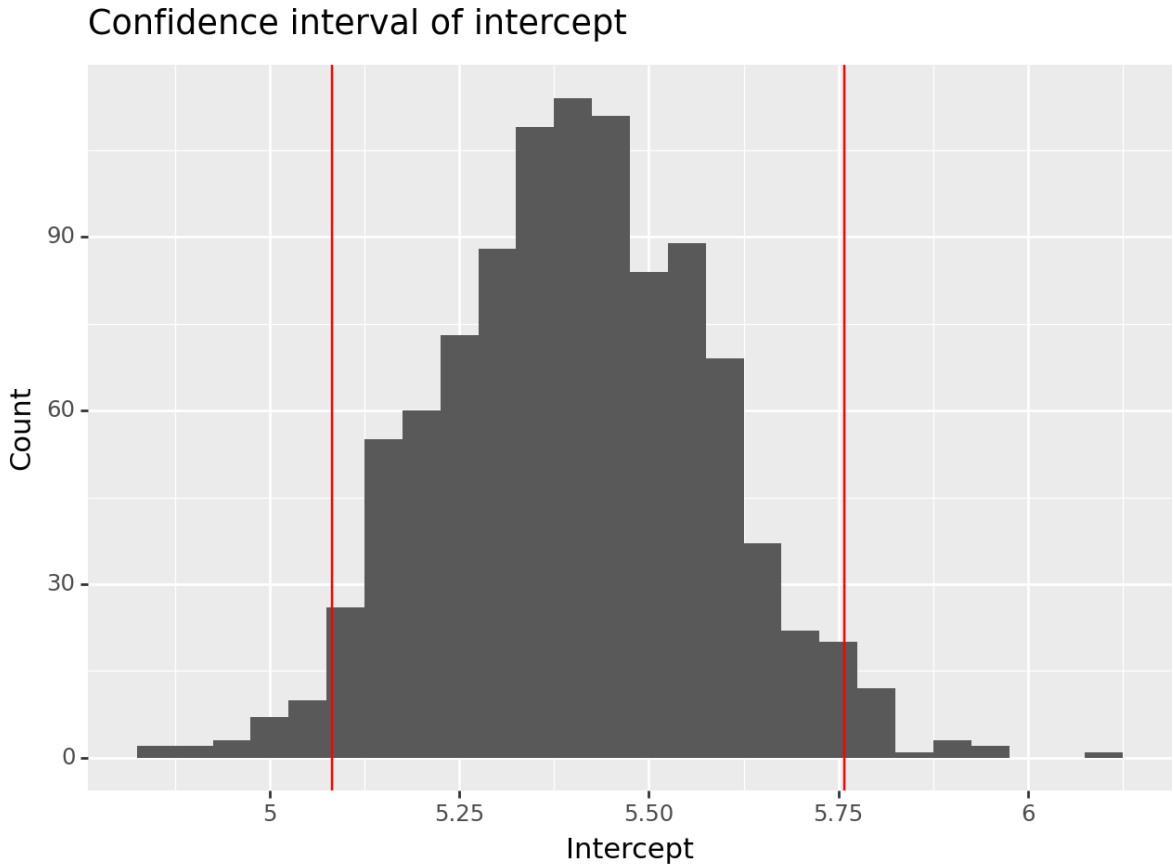

Figure 14: 95% confidence interval for intercept

5 T-test

This section analyzes whether there is a significant difference in temperature between the locations St. Andrews and Edinburgh.

5.1 Visual Exploration

For that purpose, a visual analysis is conducted first. Figure 15 depicts the frequency distribution of hourly temperatures for St Andrews and Edinburgh with a Kernel Distribution Estimation (KDE) overlay for both. It is evident from the histograms that both locations exhibit a similar pattern of temperature distribution, which is reasonable given the location's shared geographical space.

This observation is confirmed by Figure 16. The central line in each box showcases the median temperature, which is similar for St. Andrews and Edinburgh. The height of the boxes represents the interquartile range (IQR) which is slighlty smaller for Edinburgh, suggesting that the middle 50% of the temperature are spread out over a somewhat narrower range for Edinburgh. The whiskers of the box plot, represent the range of data excluding outliers which is marginally smaller at the location Edinburgh. However, please note that these deviations appear so small that they can be neglected.

Generally, the visual exploration implies that there is no significant difference between the locations.

5.2 Formal Analysis using a Permutation and Randomisation test

Formal analysis is conducted to generate robust insights.

Firstly, am independent two-sample t-test is conducted to obtain the observed test statistic from the sampled data which yields a t-value statistic of -0.436 and a p-value of 0.663. As the p-value exceeds the chosen alpha level of 0.05, the result suggests that there is no significant difference in the mean hourly temperatures between St. Andrews and Edinburgh.

However, as outlined in the Methodology section the temperature samples from St. Andrews and Edinburgh do not satisfy the normality assumption. Therefore, the results obtained by the t-test are not fully robust.

To generate reliable insights a permutation testing is applied which results in a sampling distribution of t-statistics under the assumption that H0 (no difference in temperature between St. Andrews and Edinburgh) is true (as shown in Figure 17). The observed test statistic of -0.436 is placed within this sampling distribution. The observed t-value is centrally located within the distribution, suggesting that H0 cannot be rejected. This hypothesis is confirmed by the obtained p-value of 0.673 which strongly exceeds the set alpha level of 0.05. Based

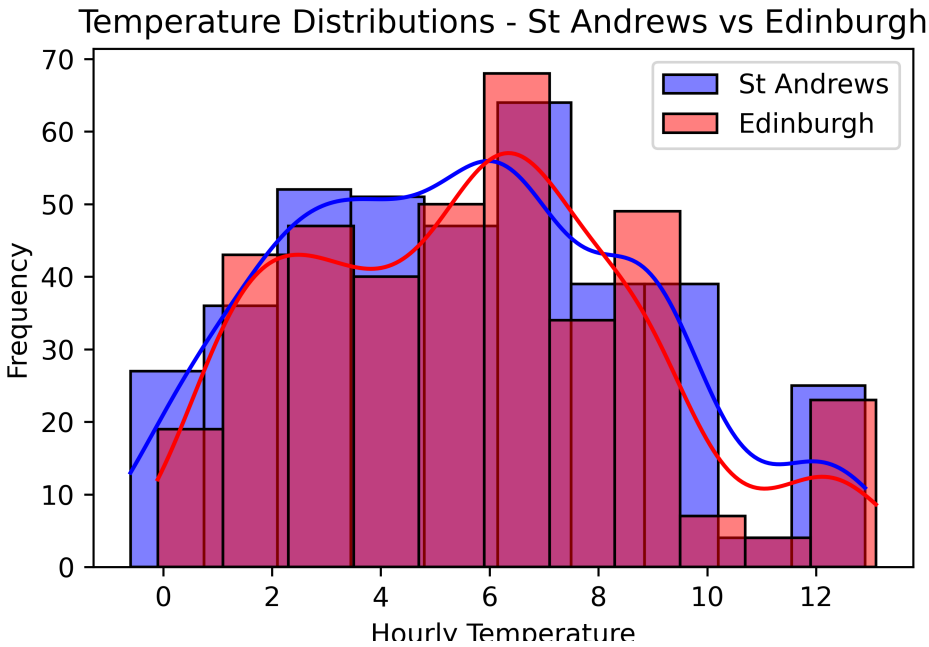

Figure 15: Temperature distributions in St. Andrews and Edinburgh

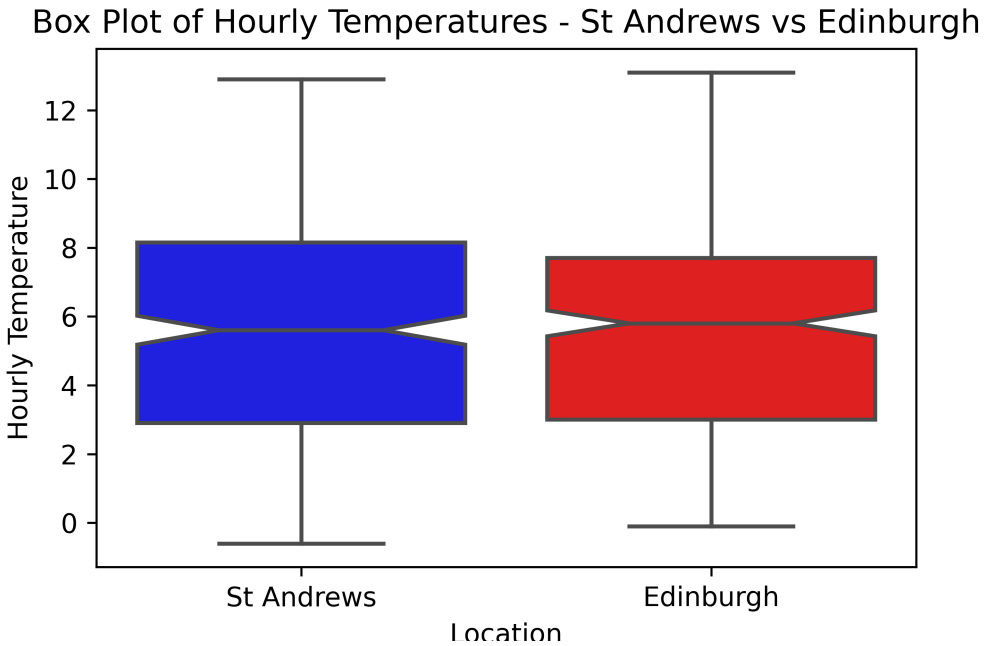

Figure 16: Boxplot for temperature distribution in St. Andrews and Edinburgh

on the permutation test no significant difference between temperatures in St. Andrews and Edinburgh could be established.

The permutation test is chosen over a randomization test as it is more efficient. While the permutation test takes only 0.000166 seconds to run the randomization test takes 0.000260 seconds.

The randomization test is slower because it includes two additional steps that the permutation test does not execute. Firstly, within the permutation test, the test statistic is directly calculated based on the permuted sample. In the randomization test, the means are taken first and based on these the test statistic is examined. Secondly, in the permutation test, the values are directly sampled randomly into the two groups. In the randomization test, an index is first randomly created, along which the values are then assigned to the groups in an additional step.

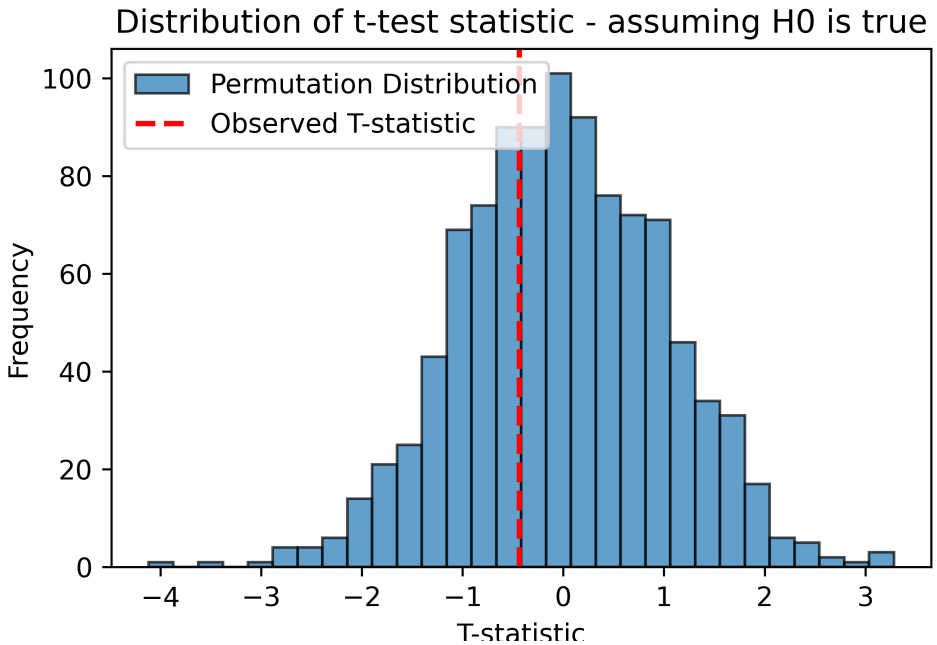

Figure 17: Distribution of t-test statistic - assuming H0 is true

5.2.1 Bootstrap estimates of the difference in means

The 95% confidence interval for the difference in means obtained by bootstrapping is [-.0.100; -0.077] (as shown in Figure 18). As comparison, [-0.552; 0.351] is retrieved as a 95% confidence interval of the difference in means using the statsmodels package.

The comparison between the bootstrapping interval and the statsmodel interval reveals that, while the two intervals overlap, the bootstrapping interval is notably narrower, suggesting greater precision. This difference is likely attributed to the non-normal distribution of the underlying data, as illustrated in the Methodology section. The non-normality of the data makes bootstrapping a more appropriate and reliable method, as it is less sensitive to distribution assumptions and provides a more robust estimation of uncertainty.

It's important to highlight a slight inconsistency between the bootstrap confidence interval and the result obtained from the permutation test. The bootstrap interval, being our primary method due to its robustness in non-parametric scenarios, does not include the value zero. This suggests a subtle yet statistically significant difference between temperatures in St. Andrews and Edinburgh.

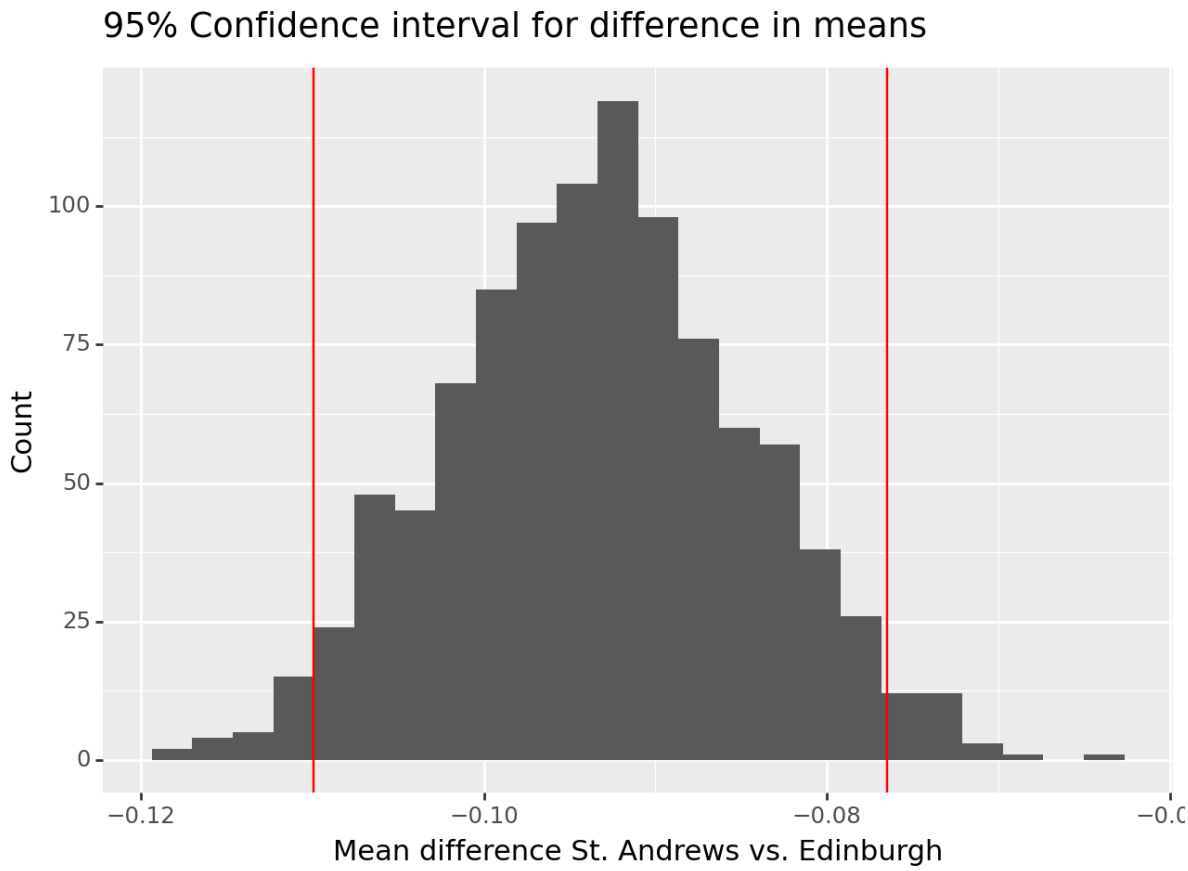

Figure 18: 95% Confidence interval for difference in means

6 Discussion and Conclusion

The report introduces a Shiny App designed to showcase current and forecasted weather details for St. Andrews and Edinburgh. Subsequent analyses were conducted based on the forecast data to gain deeper insights into the weather patterns in these locations.

The findings revealed that the mean temperature in St. Andrews stands at 5.66 °C. The Jackknife results indicate the Jackknife sampling means closely approximate the true mean temperature. Moreover, through regression and bootstrapping analyses, it was established that higher precipitation is significantly associated with elevated temperature levels in the examined dataset. Additionally, bootstrapping of mean temperature differences between St. Andrews and Edinburgh indicated a subtle distinction in temperature between the two locations within the observed dataset.

However, it's crucial to acknowledge several limitations within this report. Firstly, the utilized weather data is forecast-based, lacking historical evidence. Secondly, the dataset itself is relatively small, encompassing weather information for only 16 days. These limitations emphasize the need for cautious interpretation of the results.

References

McIntosh, Avery. 2016. "The Jackknife Estimation Method." https://arxiv.org/abs/1606.00497.