

Advanced Statistical Programming

Module code: ACCA7022

University of Wales Trinity Saint David

Sriskantharaja Mithushan (2232358)

Task 2 – Documentation

Q1

Task 1.1: Average Seasonal Maximum Temperatures Over Time

The code extracts relevant data from the "Max Temp" sheet of the Excel file. It selects the columns for the year, winter (win), spring (spr), summer (sum), and autumn (aut) temperatures. The data is then filtered to include every tenth year. The temperature columns are converted to numeric values, and the row means are calculated. Finally, a line plot with points is created using ggplot to visualize the average seasonal maximum temperatures over every tenth year.

Import Data:

The process begins by loading the Excel file using the `read_excel` function from the `readxl` library. The data is stored in the workbook object (wb), specifically reading the "Max Temp" sheet.

Extract Relevant Columns:

The relevant data for this task includes the columns for the year, winter (win), spring (spr), summer (sum), and autumn (aut) temperatures. These columns are extracted and stored in the `seasons_data` data frame.

Filter Data (Every Tenth Year):

To focus on a more manageable and representative dataset, the script filters the data to include observations for every tenth year using the `seq` function. This new subset is stored in the `every_tenth_year` data frame.

Calculate Row Means:

Numeric conversion is applied to the temperature columns, and the script then calculates the row means for each row of the subset, representing the average seasonal maximum temperature for every tenth year. The results are stored in a new column named 'average.'

Plot Data (ggplot):

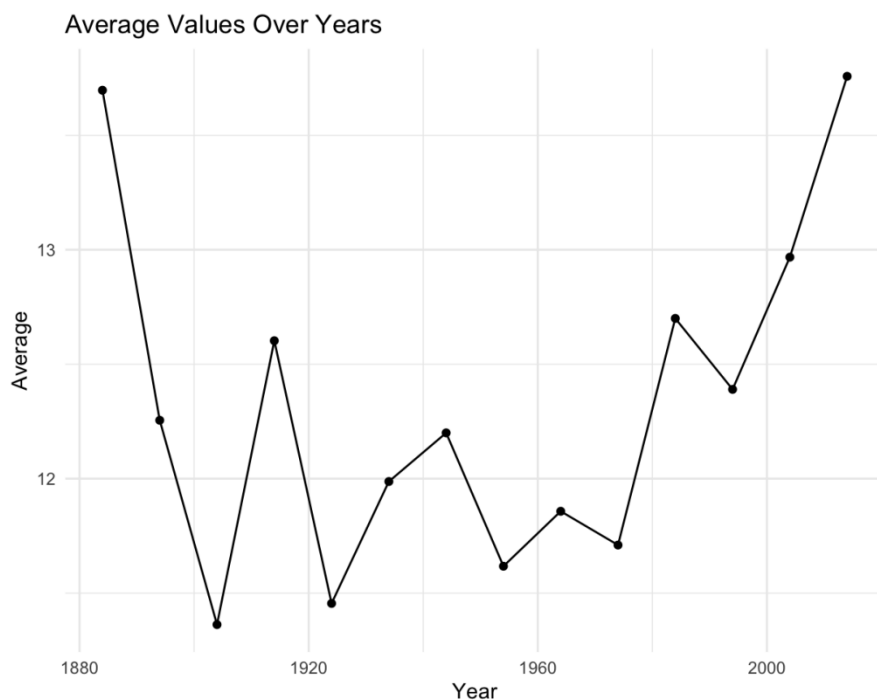
The final step involves visualizing the data using the `ggplot2` library. A line plot with points is created, depicting how the average seasonal maximum temperatures have changed from 1892 to 2022. The x-axis represents the years, while the y-axis represents the calculated average temperatures.

Data Extraction:

`seasons_data <- wb[, c("year", "win", "spr", "sum", "aut")]`: Extracts relevant columns.

`every_tenth_year <- seasons_data[seq(1, nrow(seasons_data), by = 10), ]`: Filters data for every tenth year.

Plot



Task 1.2: Correlation between Annual Rainfall and Annual Maximum Temperature

This part of the code reads the "Rainfall" sheet and extracts total annual rainfall and annual maximum temperature data. It then calculates the correlation between these two variables and prints the result.

Import Data:

Similar to Task 1.1, the script imports the Excel file, focusing this time on the "Rainfall" sheet.

Extract Rainfall and Max Temp:

The relevant data for this task involves the total annual rainfall (rainfall) and annual maximum temperature (max_temp). These variables are extracted from the respective columns in the workbook.

Calculate Correlation:

The script uses the cor function to calculate the correlation coefficient between total annual rainfall and annual maximum temperature, focusing on data from the 50th year onward.

Print Correlation:

The calculated correlation coefficient is then printed to the console using cat, providing insight into the relationship between total annual rainfall and annual maximum temperature.

Data Extraction:

rainfall <- wb2\$ann: Extracts total annual rainfall data.

max_temp <- wb\$ann: Extracts annual maximum temperature data.

Result

Correlation between Total Annual Rainfall and Annual Maximum Temperature:
0.0725137

The correlation coefficient of 0.0725137 indicates a weak positive relationship between Total Annual Rainfall and Annual Maximum Temperature. While there is a small connection, the low correlation value suggests that other factors may play a more substantial role in influencing these variables.

Task 1.3: T-test for Minimum Temperatures in Different Time Periods

The code reads the "Min Temp" sheet and creates two vectors representing two groups based on specified time periods. It then performs an independent samples t-test to identify if there is a significant difference in minimum temperatures between the two periods.

Import Data:

The "Min Temp" sheet of the Excel file is loaded to analyze minimum temperatures.

Create Groups:

Two groups are created, representing different time periods: group1 (August 1890-1910) and group2 (August 2000-2020).

Perform t-test:

The script conducts an independent samples t-test using the `t.test` function to identify whether there is a significant difference in minimum temperatures between the two specified time periods.

Print Result:

The t-test result, including statistics such as t-value and p-value, is printed to the console.

Data Extraction:

group1 and group2: Create vectors for two groups based on time periods.

In the context of a t-test comparing minimum temperatures between two time periods (August 1890-1910 and August 2000-2020), the null and alternative hypotheses can be stated as follows:

Null Hypothesis (H_0): $\mu_1 = \mu_2$

The null hypothesis asserts that there is no significant difference in the mean minimum temperatures between August 1890-1910 (μ_1) and August 2000-2020 (μ_2).

Alternative Hypothesis (H_1): $\mu_1 \neq \mu_2$

The alternative hypothesis suggests that there is a significant difference in the mean minimum temperatures between the two time periods.

To determine which t-test to use, we need to consider whether the samples are independent or paired.

Independent Samples T-test: This test is used when we have two separate groups, and the observations in one group are not related to the observations in the other group. Each observation in one group is independent of the observations in the other group.

Paired Samples T-test: This test is used when we have two sets of observations that are related in some way. Typically, the observations are paired because they represent

measurements taken on the same subjects or at the same experimental unit under different conditions.

In our case, we are comparing minimum temperatures between two time periods (August 1890-1910 and August 2000-2020). The temperatures in each period are not likely to be paired observations, as they represent different years. Therefore, we should use an **Independent Samples T-test**. So, this is a two-sample independent t-test with a two-tailed alternative hypothesis ($\neq$), indicating that we are interested in whether the means are significantly different in either direction.

Result

Welch Two Sample t-test

data: group1 and group2

$t = -4.6707$, $df = 39.819$, $p\text{-value} = 3.388e-05$

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-1.5555841 -0.6158445

sample estimates:

mean of x mean of y

10.24286 11.32857

Test Statistics:

$t = -4.6707$: This is the t-statistic, a measure of how many standard errors the sample mean is from the null hypothesis mean. In this case, the negative sign indicates that group1 has a lower mean than group2.

$df = 39.819$: This is the degrees of freedom associated with the t-test. It's calculated using Welch's correction, which is applied when variances between the two groups are assumed to be unequal.

$p\text{-value} = 3.388e-05$: This is a very small p-value ($3.388 * 10^{-5}$), indicating strong evidence against the null hypothesis. The null hypothesis in this context is that there is

no difference in mean temperatures between the two time periods. The small p-value suggests that you can reject the null hypothesis.

Confidence Interval:

95 percent confidence interval: -1.5555841 to -0.6158445: This interval provides a range of values within which we can be 95% confident that the true difference in means lies. In this case, the interval does not include 0, supporting the conclusion that there is a significant difference between the means.

Sample Estimates:

mean of x (group1): 10.24286

mean of y (group2): 11.32857

These are the sample means for the two groups. The mean temperature in August for group1 (presumably the earlier time period) is 10.24286, while for group2 (presumably the later time period), it is 11.32857.

Conclusion:

Based on the small p-value ($p < 0.05$) and the confidence interval not containing 0, we have statistically significant evidence to reject the null hypothesis. Therefore, we can conclude that there is a significant difference in minimum temperatures between August 1890-1910 and August 2000-2020, with temperatures being lower in the earlier time period.

Task 1.4: ANOVA and Ad-hoc Tests for Rain Days Across Decades

This part of the code reads the "Rain Days" sheet and adds a new variable 'decade' to represent complete decades. It performs an ANOVA test to identify if there are statistically significant differences in the annual numbers of wet days across decades. If significant, ad-hoc tests can be performed to identify where the differences lie.

Import Data:

The script loads the "Rain Days" sheet of the Excel file.

Create Decade Variable:

A new variable, 'decade,' is created to represent complete decades. The cut function is used to categorize the years into the corresponding decades.

Perform ANOVA:

The script performs an Analysis of Variance (ANOVA) using the aov function to identify statistically significant differences in the annual numbers of wet days across

the complete decades.

Print ANOVA Summary:

The summary of the ANOVA result is printed to the console, providing information on factors such as the F-statistic, degrees of freedom, and p-value. If a significant difference is found, additional ad-hoc tests can be performed to pinpoint where the differences lie.

Data Extraction:

`wb4$decade <- cut(wb4$year, breaks = seq(1890, 2020, by = 10), labels = FALSE):`
Creates a new variable representing complete decades.

`anova_result <- aov(ann ~ decade, data = wb4):` Performs ANOVA.

In the context of an ANOVA test comparing the annual numbers across different decades, the null and alternative hypotheses can be stated as follows:

Null Hypothesis (H_0):

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

The null hypothesis asserts that there is no significant difference in the mean annual numbers across the decades represented by groups 1, 2, ..., k.

Alternative Hypothesis (H_a):

H_a : At least one pair of means is different

The alternative hypothesis suggests that there is a significant difference in the mean annual numbers between at least two decades.

In mathematical terms, if $\mu_1, \mu_2, \dots, \mu_k$ represent the population means of the annual numbers for each decade, the null hypothesis is that all of these population means are equal, while the alternative hypothesis is that at least one pair of them is different.

The ANOVA test is designed to determine whether the observed variation between the

group means is statistically significant or whether it could reasonably be attributed to random sampling variability. If the p-value is small (typically less than the chosen significance level, e.g., 0.05), we would reject the null hypothesis and conclude that there is evidence of a significant difference in at least one pair of means.

Results

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
decade	1	892	892.1	3.686	0.0571
Residuals	128	30976	242.0		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Df (Degrees of Freedom):

decade: 1 degree of freedom for the 'decade' variable, as it represents the factor you are testing.

Residuals: 128 degrees of freedom for the residuals, which represent the unexplained variability.

Sum Sq (Sum of Squares):

decade: 892.1 is the sum of squares associated with the 'decade' variable. It represents the variability between the means of the decades.

Residuals: 30976 is the sum of squares for the residuals. It represents the unexplained variability within each decade.

Mean Sq (Mean Square):

decade: $892.1 / 1 = 892.1$. This is the mean square associated with the 'decade' variable and is a measure of the average variability between the means of the decades.

Residuals: 242.0. This is the mean square for the residuals, representing the average unexplained variability within each decade.

F value:

F value = Mean Sq (decade) / Mean Sq (Residuals) = $892.1 / 242.0 = 3.686$. The F value is a ratio of the variance between groups to the variance within groups. It's used to test the null hypothesis that all group means are equal. In this case, a higher F value suggests more variability between groups relative to within groups.

Pr(>F) (p-value):

p-value = 0.0571. This is the probability associated with the F value. It indicates the likelihood of observing an F value as extreme as the one computed, assuming the null hypothesis is true. In this case, the p-value is greater than the commonly used significance level of 0.05, suggesting that we do not have enough evidence to reject the null hypothesis. Therefore, no significant differences detected between groups.

Flowchart for all tasks

Start

|

|---> Task 1.1

| |

| |---> Import Data

| |

| |---> Extract Relevant Columns

| |

| |---> Filter Data (Every Tenth Year)

```
|  
|  
|    |---> Calculate Row Means
```

```
|  
|  
|    |---> Plot Data (ggplot)
```

```
|  
|---> Task 1.2
```

```
|  
|  
|    |---> Import Data
```

```
|  
|  
|    |---> Extract Rainfall and Max Temp
```

```
|  
|  
|    |---> Calculate Correlation
```

```
|  
|  
|    |---> Print Correlation
```

```
|  
|---> Task 1.3
```

```
|  
|  
|    |---> Import Data
```

```
|  
|  
|    |---> Create Groups
```

```
|  
|  
|    |---> Perform t-test
```

```
|  
|  
|    |---> Print Result
```

```
|  
|---> Task 1.4
```

```
|  
|  
|    |---> Import Data
```

```
|  
|---> Create Decade Variable  
|  
|---> Perform ANOVA  
|  
|---> Print ANOVA Summary  
|  
End
```

Q2

Challenges Encountered:

Data Understanding:

Understanding the details of the climate dataset posed a challenge. Going through the dataset to identify and extract the relevant information for analysis required careful analytical skills.

Statistical Analysis:

The application of statistical analyses, such as t-tests and ANOVA, brought to light the complexities of statistical concepts. Ensuring the appropriate selection of tests and accurately interpreting results was challenging but insightful.

Data Visualization:

Crafting meaningful and informative visualizations presented its set of challenges. Selecting the right visualization technique, ensuring clarity, and effectively communicating insights through visuals demanded a combination of statistical and graphical skills.

Learnings:

Data Wrangling Techniques:

I learned various data wrangling techniques, such as subsetting, filtering, and creating new variables. These skills are fundamental for preparing data for analysis and gaining insights.

Statistical Analysis in R:

The tasks provided an opportunity to apply and deepen my understanding of statistical concepts like correlation, t-tests, and ANOVA in the R programming language. This practical application enhanced my statistical analysis skills.

Effective Code Organization:

Through this project, I gained insights into the importance of clear and well-commented code for easy understanding. Proper code organization, with comments explaining each step, contributes to code readability and future reference.

What I Would Do Differently Next Time:

Enhance Documentation:

Looking ahead, I aim to enhance documentation by providing more context for each task. This involves explaining the reasoning behind certain decisions and including comments within complex functions to improve overall code readability.

Error Handling:

In future projects, I intend to implement more robust error-handling mechanisms. This includes checks for potential issues during data import, ensuring columns exist before extraction, and handling missing or inconsistent data more gracefully.

Exploratory Data Analysis (EDA):

I realized the significance of conducting a more thorough exploratory data analysis before delving into specific tasks. Future projects will involve a deeper examination of key variables, identifying outliers, and discerning patterns that can guide analytical methods.

Modularization:

To enhance code maintainability, I plan to explore the separation of tasks into modular functions. Each task could be encapsulated within a function, facilitating easier reuse or modification of individual components without affecting the entire script.