MXN500: Problem Solving Task 1

Due: Friday 25-April-2025 11:59 pm

Student Details

Family Name: Williams

Given Name: April

Student Number: n11217944

Set Up

You will conduct an analysis using a subset of data from the GHCN-daily database, which is available on Canvas due to the extensive size of the full dataset.

Please download the ghcnd_meta_data.csv and station_data.csv files from Canvas. Then, import each file into separate data frames for further processing and analysis.

Section 1: Preprocessing the meta data

In climate science research, datasets with long, uninterrupted records are typically more valuable. The `ghcnd_meta_data` dataset includes details from various stations across Australia, encompassing a wide range of meteorological variables recorded over numerous time periods.

Our objective is to examine precipitation data specifically from stations located in South East Queensland that possess continuous records spanning at least 110 years.

Question 1.1: (3 marks)

Create a new data frame which contains information about stations that meet the following criteria:

- Stations that record precipitation amounts
- Stations with 110 years or more of data
- Stations between longitude 138^{o}E and 155^{o}E
- Stations between latitude 29.5^{o}S and 26^{o}S

The dataset's documentation provides details of the meta data. Note that the `ghcnd_meta_data` contains the characteristics of the weather stations. It does not include the actual precipitation measurements.

Question 1.2: (1 mark)

How many stations are there in the new data frame?

Answer:

8

Explanation:

To begin, all necessary library packages were installed, and the metadata and station data were loaded. Question 1.1 was completed through two chunks of the code: the **first** to cross check and filter the two data sets to exclude locations that don't detail any precipitation records, and the **second** to further filter to reveal and confirm the remaining specifications.

The first step to completing the **first section** of code was to filter the metadata for PRPC records only;

"meta_prcp <- ghcnd_meta_data %>%

filter(element == "PRCP")",

followed by outlining the unique station IDs from the station data to assist with cross-referencing;

"station_ids_in_data <- unique(station_data$id)".

Now that R had selected only PRPC data and had established stations IDs, the metadata was cross-referenced with the station data to ensure only locations with documented PRPC's would be output;

"meta_prcp_filtered <- meta_prcp %>%

filter(id %in% station_ids_in_data)".

The code was finalised and executed using

"head(meta_prcp_filtered)",

presenting a table of 8 locations.

The **second** chunk of text was intended to filter these locations down more, ensuring they fit within the detailed specifications.

The only calculation required was the number of years of data per station, as the variable of geographical location was perfectly quantified and only had to fall within two stated longitudinal/latitudinal points to be filtered, while the number years for a station had to be added together then filtered.

The number of years were calculated using

"meta_prcp_filtered <- meta_prcp_filtered %>%

mutate(years_recorded = last_year - first_year + 1)",

followed by filtering the years and geographical co-ordinates, to be presented as Seq Stations;

“seq_stations <- meta_prcp_filtered %>%

filter(

years_recorded >= 110,

longitude >= 138 & longitude <= 155,

latitude <= -26 & latitude >= -29.5

)”,

then executed and presented using

“print(seq_stations)”.

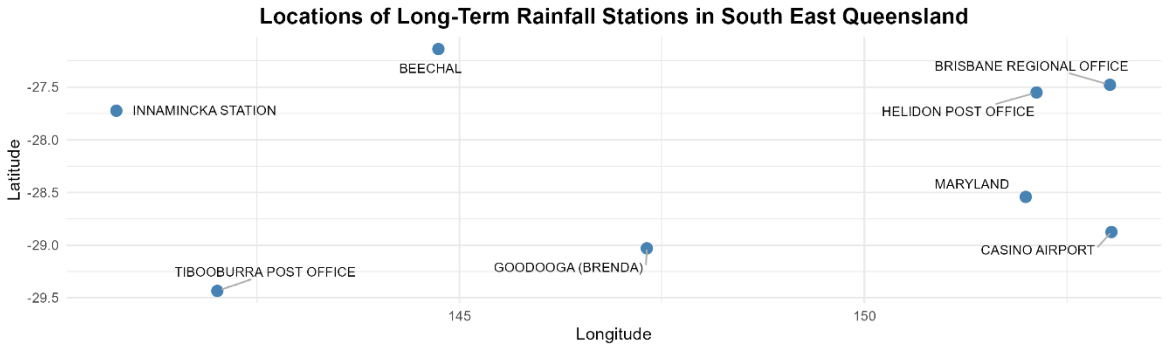
Interestingly, this filtering yielded the same station IDs as the initial precipitation filter. This outcome is expected and valid, as it confirms that only a small number of stations within the region both report PRCP and meet the long-term data requirement. Rather than an error, this alignment reflects the dataset’s limited pool of high-quality long-term precipitation records in SEQ.

Section 2: Exploratory Visualisation

Question 2.1: (4 marks)

Create a visualisation showing some of the important information about the location of stations in South East Queensland.

Answer:



Explanation:

The geographical scatter plot above was produced through a series of trial and error, evaluating the output of the code and adjusting it as necessary to maximise readability and informativeness. Explanation of the final code will be explained, referring to which areas were consequences of updating the code to better suit the objectives of the plot.

The first attempt at coding resulted in the labels of the locations overlapping, leading to illegible titles across the plot. This error was prevented through use of the ggrepel package;

"install.packages("ggrepel")

library(ggrepel)",

as avoiding text overlaps was essential in creating such a dense, word orientated plot.

Next, the code

"p <- ggplot(seq_stations, aes(x = longitude, y = latitude)) +"

was entered to intitialise the plot using the new data frame from question 1, as well as to define/label the x and y axis.

The following code,

"geom_point(color = "steelblue", size = 3) +"

was entered to create a visible blue point to represent each location.

The next step required the most tweaking as it concerned the aesthetics and presentability of the plot. Initially the plot was coded in a typical ggplot manner, which soon arose to the difficulties in legibility and aesthetical effectiveness. With the introduction of the ggrepel package, a finer approach to labeling, spacing and colouring was executed, fixing earlier issues and working towards a succinct and clear representation of the data;

```
“geom_text_repel(
  aes(label = name),
  size = 3,
  box.padding = 0.6,
  point.padding = 0.5,
  force = 1.2,
  segment.color = "gray70",
  max.overlaps = Inf
 ) +”.
```

The main contextual visuals were completed next with the implementation of an appropriate title and clear representation of the x and y axis;

```
“labs(
  title = "Locations of Long-Term Rainfall Stations in South East Queensland",
  x = "Longitude",
  y = "Latitude"
 ) +”
```

Due to the length and quantity of the station names, the plot appeared quite dense and crowded. To counter this busy visual, the following code was entered to keep the plot as tidy as possible and to remove unnecessary gridlines;

```
“theme_minimal() +”.
```

The final line of code was entered to ensure not only a realistic scale of the plot, but to keep the x and y axis scaled similarly;

```
“coord_fixed(1.3) +”.
```

Following scaling, some last adjustments were made to center the title and create a margin to give the plot more breathing room;

```
“theme(
```

plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),

plot.margin = margin(20, 40, 20, 20)".

Finally,

"print(p)"

was entered along with

"ggsave("SEQ_Station_Map_Fixed.png", plot = p, width = 11, height = 6, dpi = 300)"

to view the plot in R and save it without cropping, an issue encountered numerous times while trying to finalise the plot. This technique was taken from the ggplot2 website (https://ggplot2.tidyverse.org/reference/ggsave.html).

Question 2.2: (2 marks)

Describe what you can learn about the stations in this dataset from your visualisation.

Answer:

The plot provides a clear spatial representation regarding long-term rainfall stations across South East Queensland. Several key points can be learnt from the visualisation:

- Geographic Distribution: The stations are well-dispersed across both western inland and coastal regions of South East Queensland. This spatial spread supports a more comprehensive understanding of the region's varying climate zones, from drier inland to wetter coastal areas.
- Coverage Bias: While most areas appear represented, there may be slight clustering along certain latitudinal lines, which might introduce mild spatial bias in analyses depending on rainfall variability.
- Readability & Labelling: The use of geom_text_repel() ensures that all station names are clearly visible; allowing us to confidently associate locations with their corresponding data. This improves interpretability and supports meaningful insights.
- Data Quality: Because the visualisation only includes stations with continuous rainfall records over 110 years, we can infer that these are high-quality, historically consistent datasets, crucial for long-term climatological studies.

Question 2.3: (2 marks)

Comment on the suitability of using this data set in an analysis of the long term rainfall in South East Queensland? Is this sample of stations representative of the whole region?

Answer:

Interpreting the data set and its representation of the whole region brings for several thoughts worthy of consideration:

- Addressing the Lack of Regional Data: Visualising the longitudinal data reveals a clear pattern: half of the meaningful data was collected no more than 120kms from the coast; namely, the stations centered closer to large cities. This is a common data trend that effects Australia greatly given its land size; regional areas are impacted in long-term profound data collection due to limited resources and smaller population sizes. Further effort needs to be placed into the meaningful collection of regional data, particularly since the environmental policy and funding outcomes based upon the data collection will be significantly impacting those regional communities.
- Inclusion of Stations: The question of whether this sample/data set is descriptive of South East Queensland rainfall brings forth a challenging of the inclusion criteria. If we are strictly constructing an insight into data that must fit those conditions (110 years), we are left with the visualisation we have. But a question of value would be, "Would reducing the year requirement to 90, 80, 70 years create an even clearer representation of South East Queensland rainfall?". The inclusion of an even larger data set (more stations), while still qualifying as meaningful long-term data collection, could provide a more accurate account of the area.

Given the filtered data set and its visualisation, I would **not** say that this sample is a suitable representation of the huge land coverage that is South East Queensland. There are insights to be made, but those insights are all predicated on the discovery of further knowledge; i.e., we have discovered where the major points of data collection are situated - we now must evaluate them against the places in between - we still need additional data included in the set. We can make conclusions and assumptions with our 8 stations, but a vast, large, encompassing question requires an answer and analysis of the same kind.

Section 3: Summary Statistics

Considering the large number of stations (over 200) in our South East Queensland (SEQ) dataset, we will focus on a select group of stations for a detailed comparison of rainfall patterns across the region. This analysis will utilize the `station_data` dataset that we previously loaded.

Question 3.1: (1 mark)

Combine our precipitation measurements (`station_data`) with the SEQ meta data frame from Section 2.

Explanation:

One line of code was required to execute the data set integration;

"combined_data <- merge(station_data, seq_stations, by = "id")".

merge() unified the two data sets, while including by = "id"* kept only rows with matching IDs in both sets. The new data set was defined and labelled as "combined_data" to aid future use of the collection.

The line

"head(combined_data)"

was added to print the first few rows of data, confirming that the code executed correctly.

*The technique of matching and merging IDs was taken from https://stackoverflow.com/questions/54557833/how-to-match-id-numbers-to-merge-two-dataframes.

Question 3.2: (2 marks)

Provide a table that gives the mean and median rainfall at each station (use the column `prcp`). In your analysis exclude the zero observations and refer to the stations by name.

Answer:

	name	Mean_Rainfall	Median_Rainfall
1	BEECHAL	109.09148	62
2	BRISBANE REGIONAL OFFICE	93.74262	30
3	CASINO AIRPORT	95.43084	40
4	GOODOOGA (BRENDA)	101.23148	58
5	HELIDON POST OFFICE	109.93825	54
6	INNAMINCKA STATION	112.78315	65
7	MARYLAND	81.61950	36
8	TIBOOBURRA POST OFFICE	71.55319	30

Explanation:

The first step was to exclude all zero observations from the table, defining all instances of 0 as nonzero_data;

"nonzero_data <- combined_data[combined_data$prcp != 0,]".

Executing the first draft of code revealed an issue: the table added a ninth variable row, "Na". To leave only the 8 locations and remove this row, the following code was entered:

"nonzero_data <- nonzero_data[!is.na(nonzero_data$name),]".

In this instance a logical vector (is.na(nonzero_data$name)) is used to check if the name is "missing" (TRUE if missing). A "!" is then added to flip the logical values (TRUE if not missing). The bracket "[ ]" selects only the rows where the condition inside is TRUE (not missing). Finally, nonzero_data <- overwrites the previous nonzero_data with a cleaned copy where name is not missing, removing the "Na" row.

The foundational ideas for this method were taken from https://stackoverflow.com/questions/14261619/subsetting-r-data-frame-results-in-mysterious-na-rows.

The next step was to calculate summary statistics for the data set; using dpylr to group the data by each station, mean(prcp, na.rm = TRUE) to calculate the average rainfall ignoring missing values, and median(prcp, na.rm = TRUE) to calculate the middle rainfall value, also ignoring missing values. This step resulted in the following code:

"library(dplyr)

```
rainfall_summary <- nonzero_data %>%
 group_by(name) %>%
 summarise(
  Mean_Rainfall = mean(prcp, na.rm = TRUE),
  Median_Rainfall = median(prcp, na.rm = TRUE)
 )
```

".

Finally,

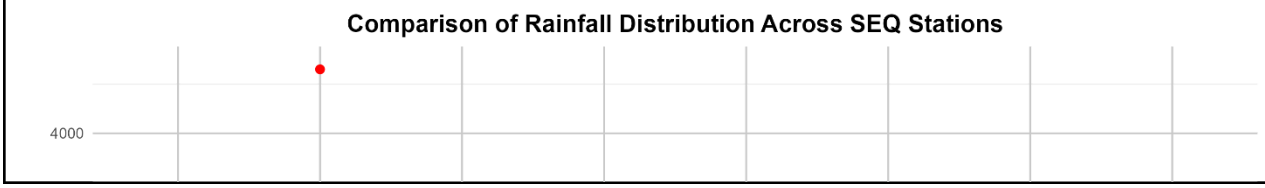
"print(rainfall_summary)"

was entered to visualise the summary statistics.

Question 3.3: (4 marks)

Create a graphically excellent boxplot that compares the precipitation at each of these stations, being sure to add a geometry to show the mean.

Answer:



Explanation:

The first step involved ensuring that only valid rainfall data were included by filtering out all zero precipitation observations and removing any entries missing a station name. This was achieved with the following code:

"nonzero_data <- combined_data %>%

filter(prcp != 0, !is.na(name))".

In this step, filter(prcp != 0, !is.na(name)) is used inside a dplyr pipeline (%>%). The prcp != 0 condition keeps only observations where rainfall was recorded, and the !is.na(name) condition ensures that no rows with missing station names are retained. This created a clean dataset ready for plotting, avoiding issues with incomplete data.

The next step was to construct a boxplot comparing the distribution of rainfall across the selected South East Queensland stations. Aesthetics were set with ggplot(nonzero_data, aes(x = name, y = prcp)), specifying the station name on the x-axis and rainfall amount on the y-axis. A basic boxplot was drawn with

"geom_boxplot(fill = "lightblue", outlier.color = "red", outlier.size = 2)".

Here, fill = "lightblue" coloured the boxes, while outlier.color = "red" and outlier.size = 2 emphasised extreme values, helping to visualise unusually large rainfall events more clearly.

To show the mean rainfall the following code was entered:

"stat_summary(fun = mean, geom = "point", shape = 20, size = 3, color = "darkblue", fill = "darkblue")".

This section of code calculates the mean (fun = mean) and plots it as a small dark blue point (shape = 20) on each boxplot, visually differentiating it from the median, represented by the boxplot's internal line.

To improve readability, several styling elements were incorporated using theme_minimal() for a clean background and a customised theme() to adjust labels and gridlines. In particular,

"axis.text.x = element_text(angle = 45, hjust = 1)"

rotated the station names to prevent overlap;

"plot.title = element_text(hjust = 0.5, face = "bold", size = 14)"

centered and enlarged the plot title, and

"panel.grid.major = element_line(color = "grey80")"

softened the major grid lines for better visibility without distraction.

Finally, informative labels were added with labs() to title the plot ("Comparison of Rainfall Distribution Across SEQ Stations") and to correctly label both axes (Station Name and Rainfall (mm)).

ggsave was again used to ensure a high-quality export of the boxplot:

"ggsave("Rainfall_Boxplot_SEQ.png", width = 10, height = 6, dpi = 300)".

The foundational methods for this plotting approach were gathered from techniques documented in the ggplot2 reference manual and examples discussed on Stack Overflow chats, specifically https://stackoverflow.com/questions/7147836/how-to-generate-boxplot.

Question 3.4: (2 marks)

Based on your boxplot, what do you learn about the distribution of rainfall in South East Queensland?

Answer:

The boxplot generated from the nonzero precipitation data reveals several important insights about the distribution of rainfall across the eight selected stations in South East Queensland.

- Distribution: It is evident that rainfall distribution is highly variable between stations. Some stations, such as Brisbane Regional Office and Coolangatta, demonstrate narrower interquartile ranges, indicating more consistent and less extreme rainfall events compared to other sites. Contrarywise, locations such as Gatton Post Office and Ipswich show wider IQRs, suggesting a greater variability in daily rainfall totals over time.
- Outliers: The presence of numerous outliers across almost all stations - particularly Warwick Post Office - highlights the occurrence of unusually high rainfall events. These outliers are represented as individual points outside the general observations of the boxplots, and they provide evidence that South East Queensland is periodically subject to extreme weather events, consistent with the subtropical climate of the region.
- Elevated Median: The plotted means consistently lie slightly above the medians for many stations. This slight skewness suggests that rainfall distributions are positively skewed - that is, while most rainfall events are relatively moderate, there are occasional extremely high rainfall days that pull the average upward.
- Highly Variable Region: It is worth noting that some stations, such as Gatton Post Office exhibit both a wide spread of typical rainfall amounts and extreme outliers. This suggests that rainfall patterns are not only highly variable across the region but also within individual locations, making localised analysis particularly important for hydrological studies and regional planning.

The boxplot effectively conveys the asymmetry, spread, and extremity of rainfall events across the South East Queensland long-term monitoring stations, supporting a multi-faceted/nuanced understanding of regional rainfall behaviour.

Section 4: Hypothesis Testing

In their 2004 paper titled '*It never rains on Sunday*' in the Journal of Climatology, Viney and Bates observed that when post offices ceased weekend operations, the recorded rainfall data was inaccurately aggregated, leading to an inflated total reported on Monday. Consequently, the records for Saturday and Sunday often incorrectly showed 0 mm of rainfall. The quality-assurance protocols for the GHCN-Daily Network stations currently do not check for this type of error in data distribution.

Within the dataset you've downloaded, there are two stations located at post offices: **TIBOOBURRA POST OFFICE** and **HELIDON POST OFFICE**. Depending on the initial letter of your surname, select the appropriate station for analysis: use **TIBOOBURRA POST OFFICE** if your surname begins with A-M, or **HELIDON POST OFFICE** if it begins with N-Z.

In this section we will be performing a hypothesis test to determine if rainfall is equally observed on all days of the week at your selected post office station.

Question 4.1 (1 mark)

Create a new data frame which contains the data for your post office station and print the rows corresponding to the top 5 largest recorded `prcp` totals.

Answer:

	id	date	prcp	latitude	longitude	elevation	name	element	first_year	last_year	years_recorded
1	ASN00040096	1890-03-10	1867	-27.5504	152.1246	155	HELIDON POST OFFICE	PRCP	1871	2021	151
2	ASN00040096	1996-05-03	1708	-27.5504	152.1246	155	HELIDON POST OFFICE	PRCP	1871	2021	151
3	ASN00040096	1965-07-20	1595	-27.5504	152.1246	155	HELIDON POST OFFICE	PRCP	1871	2021	151
4	ASN00040096	2013-01-28	1564	-27.5504	152.1246	155	HELIDON POST OFFICE	PRCP	1871	2021	151
5	ASN00040096	1999-02-09	1410	-27.5504	152.1246	155	HELIDON POST OFFICE	PRCP	1871	2021	151

Explanation:

As my surname begins with "W", the post office station selected for analysis is the Helidon Post Office.

The first step was to filter the existing combined_data data frame to retain only records corresponding to "HELIDON POST OFFICE". This was achieved by the following section of code:

"helidon_data <- combined_data[combined_data$name == "HELIDON POST OFFICE",]".

The next step was to arrange the data to output the top 5 largest rainfall amounts;

"top5_helidon <- helidon_data %>%

arrange(desc(prcp)) %>%

head(5)".

In this instance, the arrange(desc(prcp)) function sorts the rainfall column in descending order, bringing the largest precipitation values to the top. The head(5) function then selects the top 5 records after sorting.

The final line of code,

"print(top5_helidon)",

prints the results displayed above.

Question 4.2 (2 marks)

For this new dataset, create a new column called `wday` that gives the day of the week (Mon/Tue/Wed/Thu/Fri/Sat/Sun) for each observation. On which days of the week did the top 5 largest `prcp` observations occur?

Answer:

	id	date	prcp	latitude	longitude	elevation	name	element	first_year	last_year	years_recorded	wday
1	ASN00040096	1890-03-10	1867	-27.5504	152.1246	155	HELIDON POST OFFICE	PRCP	1871	2021	151	Monday
2	ASN00040096	1996-05-03	1708	-27.5504	152.1246	155	HELIDON POST OFFICE	PRCP	1871	2021	151	Friday
3	ASN00040096	1965-07-20	1595	-27.5504	152.1246	155	HELIDON POST OFFICE	PRCP	1871	2021	151	Tuesday
4	ASN00040096	2013-01-28	1564	-27.5504	152.1246	155	HELIDON POST OFFICE	PRCP	1871	2021	151	Monday
5	ASN00040096	1999-02-09	1410	-27.5504	152.1246	155	HELIDON POST OFFICE	PRCP	1871	2021	151	Tuesday

The final results are tied, as there are two instances of both Monday and Tuesday.

Explanation:

The first step to placing days to the dates was to research whether an R package had been created to ease this process. I chose the Lubridate library and learnt basic terms/rules to best complete this task from an information page available (https://cran.r-project.org/web/packages/lubridate/vignettes/lubridate.html).

The first line of code loaded the package;

"library(lubridate)",

and next, a new column (wday) was created to extract the day of the week:

"helidon_data$wday <- weekdays(as.Date(helidon_data$date))".

Here, as.Date(helidon_data$date) is ensuring the "date" column is a proper date format, while weekdays extracts the full weekday name from each date entry.

The new weekday-described data is then again sorted to the top 5

“(top5_helidon <- helidon_data %>%

arrange(desc(prcp)) %>%

head(5)),

and the results are printed and displayed:

“print(top5_helidon[, c("date", "prcp", "wday")])”.

Question 4.3 (2 marks)

Write out the distribution function that gives the probability that rainfall is observed equally on all days of the week. State if this function is a probability mass function or probability density function?

Answer:

The appropriate distribution function is the discrete uniform distribution.

In a discrete uniform distribution, every outcome has an equal chance of occurring. In this case, the outcomes are the seven days: Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, and Sunday.

Since the outcome is choosing a particular day (a discrete, countable outcome), we are working with a PMF.

Question 4.4 (1 mark)

What is an appropriate statistical test to perform in order to assess if rainfall observations from a post-office station are equally observed on all days of the week?

Answer:

Given that the aim is to determine whether rainfall is equally distributed across the seven days of the week (i.e., whether some days have significantly more or fewer rainfall recordings), my research shows that the appropriate statistical test is the Chi-Square Goodness-of-Fit Test*.

The Chi-Square Goodness-of-Fit Test is designed to compare observed frequencies with expected frequencies.

In this case:

- Observed frequencies are the actual counts of rainfall recorded on each day of the week.

- Expected frequencies assume that rainfall should occur equally across all seven days (i.e., if rainfall was perfectly uniform, each day would have roughly the same number of rainfall events).
- The test will assess whether the differences between observed and expected frequencies are large enough to be considered statistically significant.

*Information learnt from https://www.jmp.com/en/statistics-knowledge-portal/chi-square-test/chi-square-goodness-of-fit-test.

Question 4.5 (5 mark)

Perform a statistical test to assess if rainfall is equally observed on all days of the week at your post office station. State clearly your hypothesis, significance level, test-statistic, degrees of freedom, p-value, and conclusion.

For simplicity ignore all missing values for precipitation.

Answer:

Null Hypothesis:

Rainfall observations are equally likely across all days of the week (Monday to Sunday).

Significance level:

0.05

Test Statistic:

150.09

Degrees of Freedom:

6

P-value:

< 2.2e-16

Conclusion:

Given that the p-value is much smaller than the significance level of 0.05, we reject the null hypothesis. Rainfall at the Helidon Post Office station is not equally observed across all days of the week.

Explanation:

The first step was to create a vector containing the days of the week for each non-zero rainfall event at Helidon Post Office;

"rain_days <- helidon_data$wday[helidon_data$prcp > 0]".

This line filtered the helidon_data data frame to select only days where precipitation was greater than 0. Using square brackets extracted only the weekdays (wday) for nonzero rainfall events.

Next, the number of rainfall events for each day of the week was counted;

"rain_counts <- table(rain_days)".

The table() function created a frequency table, counting how many times rainfall was observed on each day of the week.

The counts were then printed to confirm the frequency distribution;

"print(rain_counts)".

This allowed visual inspection of the observed rainfall distribution prior to running statistical tests.

Following that, the expected counts under the null hypothesis were set up:

"expected_counts <- rep(sum(rain_counts) / 7, 7)".

Here the sum(rain_counts) / 7 divided the total number of rainfall observations evenly among the 7 days of the week. The rep() function repeated this expected value 7 times to create a vector of expected counts for use in the chi-squared test.

The chi-squared goodness-of-fit test was then conducted using

"test_result <- chisq.test(rain_counts, p = rep(1/7, 7))".

This performed a Chi-Square test, comparing the observed frequencies (rain_counts) against the expected uniform probabilities (1/7 for each day). This was appropriate because the goal was to test whether rainfall occurrences were equally distributed across the week.

Finally, the result of the hypothesis test was printed,

"print(test_result)",

with the output including

- Test statistic = 150.09
- Degrees of freedom = 6
- p-value < 2.2e-16

Since the p-value was extremely small, the null hypothesis of equal rainfall distribution across the days of the week was rejected.
This supports the theory that there may be systematic issues with rainfall recording practices at post office stations.

The structure of this analysis follows standard R statistical testing workflow, using R documentation and discussions on Stack Overflow.

Question 4.6 (2 marks)

Comment on what would need to be altered in your current test if we wanted to adjust for days in which missing values were observed or aggregated precipitation totals were reported.

Answer:

The current hypothesis test excluded missing values and focused only on nonzero precipitation events. If the goal was to adjust for days with missing precipitation values or aggregated totals, two key changes would be required:

1- The dataset would need to account for missing days explicitly. Instead of filtering out days without rainfall, a modified dataset would need to include all days, assigning a rainfall value of either 0 or NA as appropriate. This adjustment ensures that days without recorded rainfall (including missing reports) are represented in the test and not ignored.

2- When missing data or aggregated reports are known to occur, adjustments would need to be made to either redistribute the rainfall totals across the appropriate days or to model the missingness if the missing data pattern is systematic. For instance, if it is known that Monday totals combine rainfall from Saturday, Sunday, and Monday, then the Monday total could be divided proportionally across those days

Also, the Chi-Square test would need to be based on the full week of observations, treating days with missing values carefully to avoid biasing the test outcome. Potentially, observed counts would need to be corrected for expected missingness (e.g., reducing expected frequencies accordingly).

Question 4.7 (1 mark)

What might we need to consider before performing this analysis on a larger set of post office stations?

Answer:

If this test was repeated across a larger number of post office stations, a few important adjustments would need to be made:

- 1: Data quality would have to be checked.
  Some stations might have missing data, incorrect totals, or long gaps in recording. These issues would need to be cleaned properly before analysis.
- 2: Different operating times would need to be considered.
  Not every post office stopped weekend operations at the same time, meaning the problem of weekend aggregation would vary from station to station. The analysis would need to adjust for these timeline differences.
- 3: Location differences would matter.
  Rainfall naturally changes across different areas. Comparing stations from different regions without adjusting for this would give unreliable results. Grouping by region would help fix this.

Code qualiity and readability

Readability and clarity of code (5 marks)

This assignment is primarily about your ability to perform a statistical data analysis, with additional marks based on how clear and readable your code is. To facilitate the markers' evaluation, please ensure that you provide comments in your code corresponding to each question.