Weather Data Report

Phil Alderman, Ryan Earp, and Katherine Haile

12/9/2021

Data Consolidation

The first task to combine the excel sheets created by the library staff from the paper documents was achieved using an function we created in R. The file name of the document was used as the argument of the function, and the read_excel and read_ods functions were used to assign the file to a temporary data frame called raw. The raw data frames were then subset to exclude the summary data and column names used in the original documents and clean variable names were assigned to the columns. Due to the variation in characters in some of the numerical columns, these were coerced to a numeric form and the original columns were labeled as such. Columns to record the year, month, observer were assigned to the corresponding cells in the original document. The overall date was also combined and the times for the beginning and ending time of rain events were set as such. Finally, the date was set as the first column and the year, month, and day columns were removed.

This function was then used to read in all of the original files and create a single .csv file. The file names were extracted from the raw data folder and ensured only the correct files were selected. This list was then fed into the read_wth_data function and arranged by date.

Fixed Errors in Import Process

During the import process, some challenges were encountered in reading in the files. Two different types of files were used to document the written documents, excel and ods. Due to this, two different functions were needed to read in the files to R. Additionally, varying naming conventions for the files were used throughout the years with some duplicated, so the file list needed to be subset to only include the desired files.

Additionally, some errors were reported when trying to read in some of the files. One file had the sheet name labeled as the month and year instead of sheet 1 like the rest, so an exception was needed for this file. Another file had the summary titles in column A repeated in column B so the information collected from this needed to be shifted to the right one.

After an initial summary of the data set, it was observed that there were several entries were recorded after the year 2000 and others the date was NA. Further investigation found errors that were corrected in code. One file contained the incorrect year in the year cell. Others had varying errors in the month column of the document. This occurred from either misspelling of the month or using a misspelled abbreviation caused the function to incorrectly assign the year. |* Note, this section of functions must be run before the import function.

Other Observations

While going through the dataset, there were several common data entry errors we noticed. For instance, a quotation mark (") implies that the entry is the same as the row above, a dash (-) implies that the data is missing/NA,

Specific Column Problems

Wind Direction at Time of Observation (wind_dir_tobs)

To check for any errors within the wind direction column, the unique function was used to show all unique entries within the column. This resulted in a list that contained variations on how the directions were inputted as capitalization changes would result in a unique entry. Most of these were able to be resolved using code as we could input all the variations and spelling errors and say what it needed to be. This leaves four data entries where the entry is uncertain and will need to be checked.

```
wth_data <- read_csv("../data/processed\\imported_weather_data_1893-1940.csv")
```

```
## Rows: 17103 Columns: 25

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (16): file_path, temp_maximum_orig, temp_minimum_orig, temp_range_orig,...
## dbl   (8): temp_set_max_orig, temp_maximum, temp_minimum, temp_range, temp_s...
## date  (1): date

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
unique_wind_dir <- unique(wth_data$wind_dir_tobs)

wth_data %>% filter(wind_dir_tobs == "?" | wind_dir_tobs == "NE; N?")
```

```
## # A tibble: 4 x 25
##   file_path         date       temp_maximum_or~ temp_minimum_or~ temp_range_orig
##   <chr>             <date>     <chr>            <chr>            <chr>
## 1 data/raw/Weather~ 1893-04-09 55               47               8
## 2 data/raw/Weather~ 1921-11-25 65               26               39
## 3 data/raw/Weather~ 1923-03-01 78               52               26
## 4 data/raw/Weather~ 1923-03-04 78               38               40
## # ... with 20 more variables: temp_set_max_orig <dbl>,
## #   precip_time_of_beginning_orig <chr>, precip_time_of_ending_orig <chr>,
## #   precip_amount_orig <chr>, precip_snowfall_in_inches_orig <chr>,
## #   precip_snow_depth_tobs_orig <chr>, wind_dir_tobs <chr>,
## #   weather_state_tobs <chr>, wind_dir_day <chr>, weather_state_day <chr>,
## #   temp_maximum <dbl>, temp_minimum <dbl>, temp_range <dbl>,
## #   temp_set_max <dbl>, precip_amount <dbl>, ...
```

Wind Direction for the Day (wind_dir_day)

The column for wind direction for the day was evaluated for errors and some problems were encountered. Though the instructions only called for one direction, many entries have two or more wind directions for the single day. Also, it seems that on some data sheets the only wind direction recorded for the day was in the Wind Direction at Time of Observation column, and the Wind Direction for the Day column includes extra notes related to the State of Weather column by including entries such as snow, thunderstorm, misting, ect. This is a problem that will need to be corrected.

```
wind_dir <- wth_data %>%
  select(wind_dir_day) %>%
  unique()
```

Weather State at Time of Observation (weather_state_tobs)

Similarly to the the wind direction, there are several errors within the weather state due to either misspelling or different ways of entering the same data entry. A similar code will need to be developed for the spelling errors identified within in this column. There are seven data entries containing question marks indicating the value is uncertain and will need to be manually checked.

```
unique_weather <- unique(wth_data$weather_state_tobs)

wth_data %>%
filter(weather_state_tobs == "Pt????" | weather_state_tobs == "?")
```

```
## # A tibble: 7 x 25
##   file_path         date       temp_maximum_or~ temp_minimum_or~ temp_range_orig
##   <chr>             <date>     <chr>            <chr>            <chr>
## 1 data/raw/Weather~ 1898-03-31 54               33               21
## 2 data/raw/Weather~ 1914-12-02 55               39               <NA>
## 3 data/raw/Weather~ 1921-07-13 90               68               <NA>
## 4 data/raw/Weather~ 1921-07-21 88               65               <NA>
## 5 data/raw/Weather~ 1921-08-15 86               71               <NA>
## 6 data/raw/Weather~ 1922-02-05 56               25               31
## 7 data/raw/Weather~ 1940-11-22 56               37               22
## # ... with 20 more variables: temp_set_max_orig <dbl>,
## #   precip_time_of_beginning_orig <chr>, precip_time_of_ending_orig <chr>,
## #   precip_amount_orig <chr>, precip_snowfall_in_inches_orig <chr>,
## #   precip_snow_depth_tobs_orig <chr>, wind_dir_tobs <chr>,
## #   weather_state_tobs <chr>, wind_dir_day <chr>, weather_state_day <chr>,
## #   temp_maximum <dbl>, temp_minimum <dbl>, temp_range <dbl>,
## #   temp_set_max <dbl>, precip_amount <dbl>, ...
```

Weather State for the Day (weather_state_day)

There were 49 factors of this variable in the final data set. The original instructions call for only using Cloudy, Partly cloudy, and Clear for the state of the weather. Several of the unique values were just misspellings of one of these three factors. There are several other factors that describe the state of the weather but do not fall under one of the three categories. There are also multiple entries that classify how Partly Cloudy the weather is in different fractions.

Maximum and Minimum Temperature (temp_maximum and temp_minimum)

When evaluating the minimum and maximum temperature the data frame was first checked for NA values, which resulted 191 observations. These were checked against the original documents, and the majority were the result of missing values in the original data set. However, the months of May through August in 1929 were not recorded at all and will need to be filled in.

Additionally, the data frame was check to see if any observations had recorded values where the minimum temperature for the day was greater than the maximum temperature for the day. There are 27 observations that fit this, and after some further evaluation, some are the result of errors in the original data set and others from misreading the documents. It is recommended that these are examined individually to determine the proper action needed.

```
na_temp <- wth_data %>%
  filter(is.na(temp_maximum), is.na(temp_minimum))
```

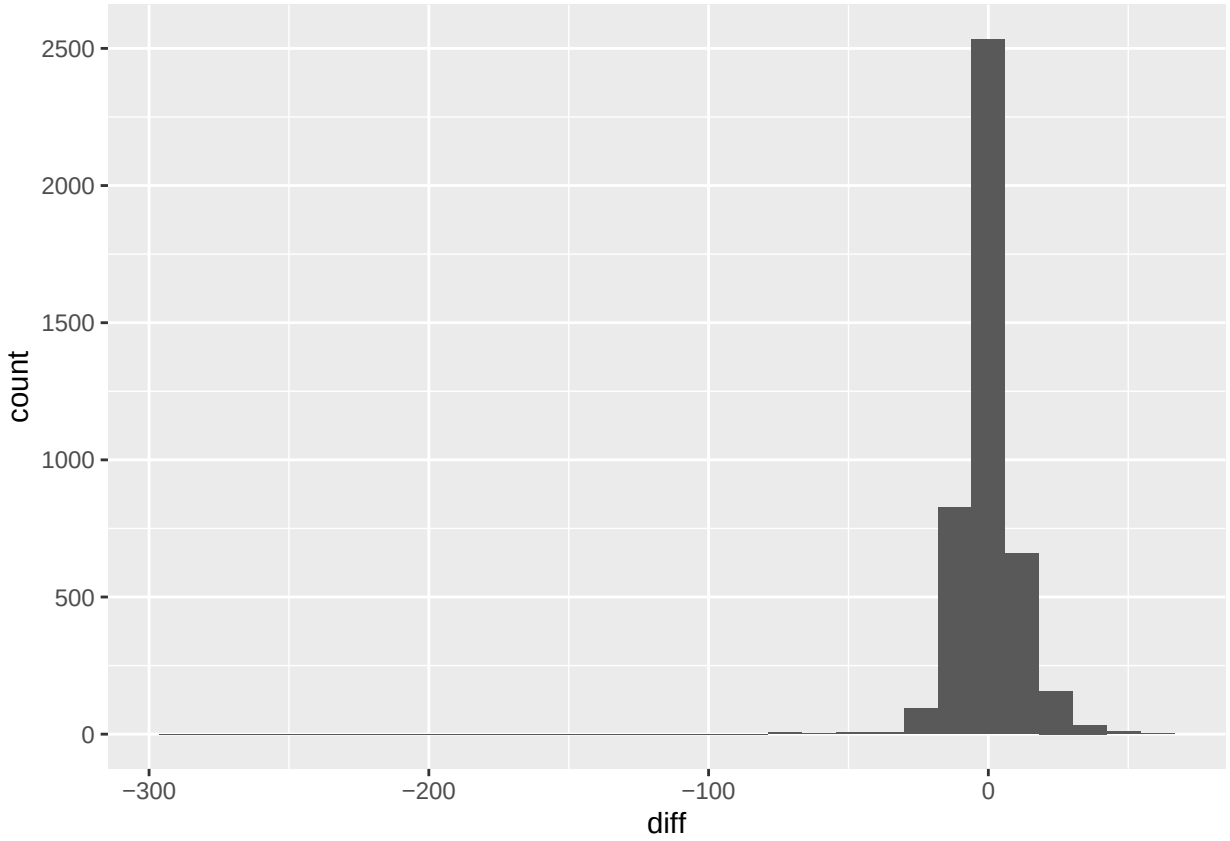
Temperature Range for the Day (temp_range)

To check the temperature range column, the dataframe was filtered to include any observations where the range did not equal the maximum minus the minimum temperature. A total of 4340 observations fit this description. No further cleaning was done for this as the error could occur within any of the three columns.

```
wrong_range <- wth_data %>%
  filter(temp_range != temp_maximum - temp_minimum) %>%
  mutate(diff =(temp_maximum - temp_minimum)-temp_range) %>%
  select(file_path, date, temp_minimum, temp_maximum, temp_range, diff)

wrong_range %>%
  ggplot(aes(x = diff))+
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Precipitation Amount (precip_amount)

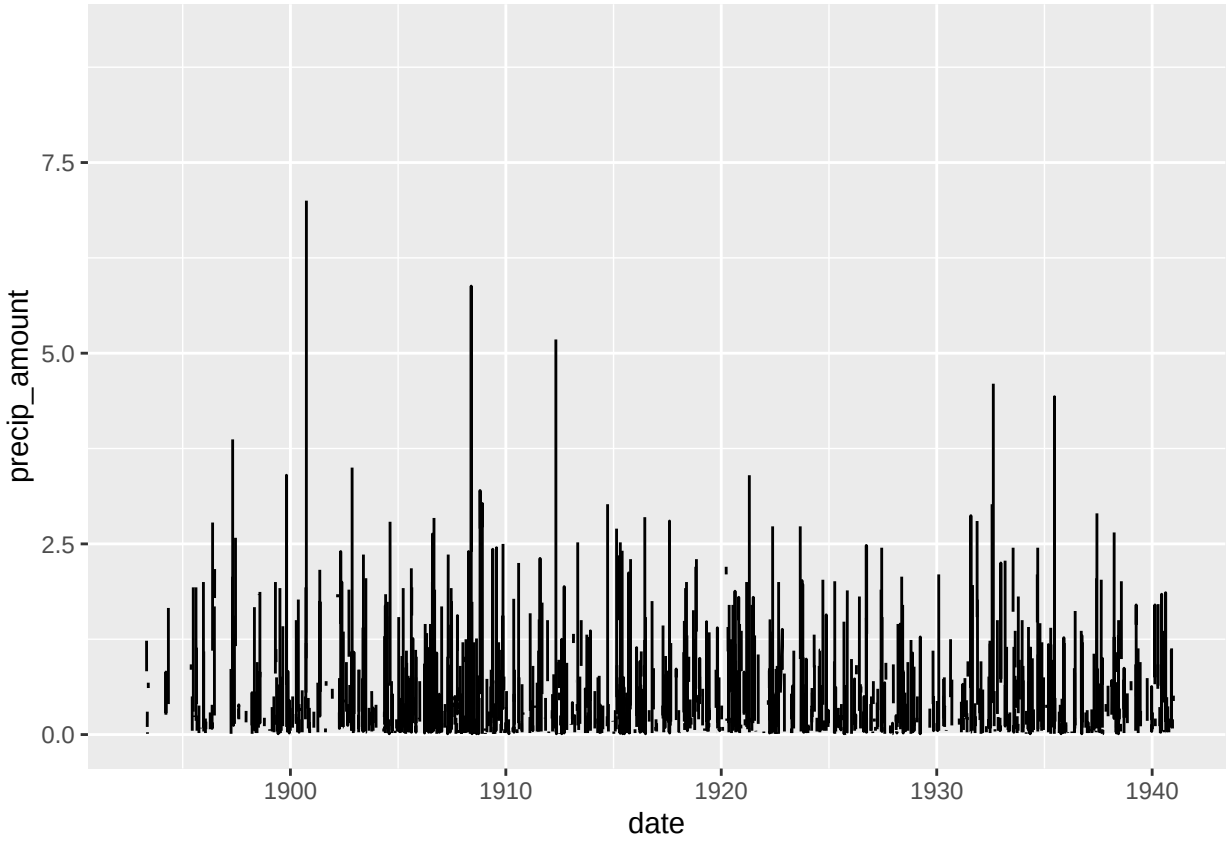
There are six values found that were likely transcribed incorrectly from the original data sheet. Four of the highest values simply need a decimal placed before the first number to make them correct. The fifth and sixth incorrect values, which are the next highest ones in ranking order, are difficult to interpret from the original handwriting. However, these values are not correct when compared to the monthly totals that are listed. It may also be worthwhile to import the monthly totals entered in the data sheets and compare them to the calculated totals from each entry for each respective month, which would help ensure that there are no other mistyped data entries that are not as easily noticeable.

```
precip <- wth_data %>%
  select(date, precip_amount) %>%
  drop_na() %>%
  arrange(desc(precip_amount))

# Fixing the large values
wth_data$precip_amount[wth_data$precip_amount == 78.00] <- .78
wth_data$precip_amount[wth_data$precip_amount == 58.00] <- .58
wth_data$precip_amount[wth_data$precip_amount == 50.00] <- .50
wth_data$precip_amount[wth_data$precip_amount == 45.00] <- .45

wth_data %>%
  ggplot()+
  geom_line(aes(x = date, y = precip_amount))
```

```
## Warning: Removed 19 row(s) containing missing values (geom_path).
```

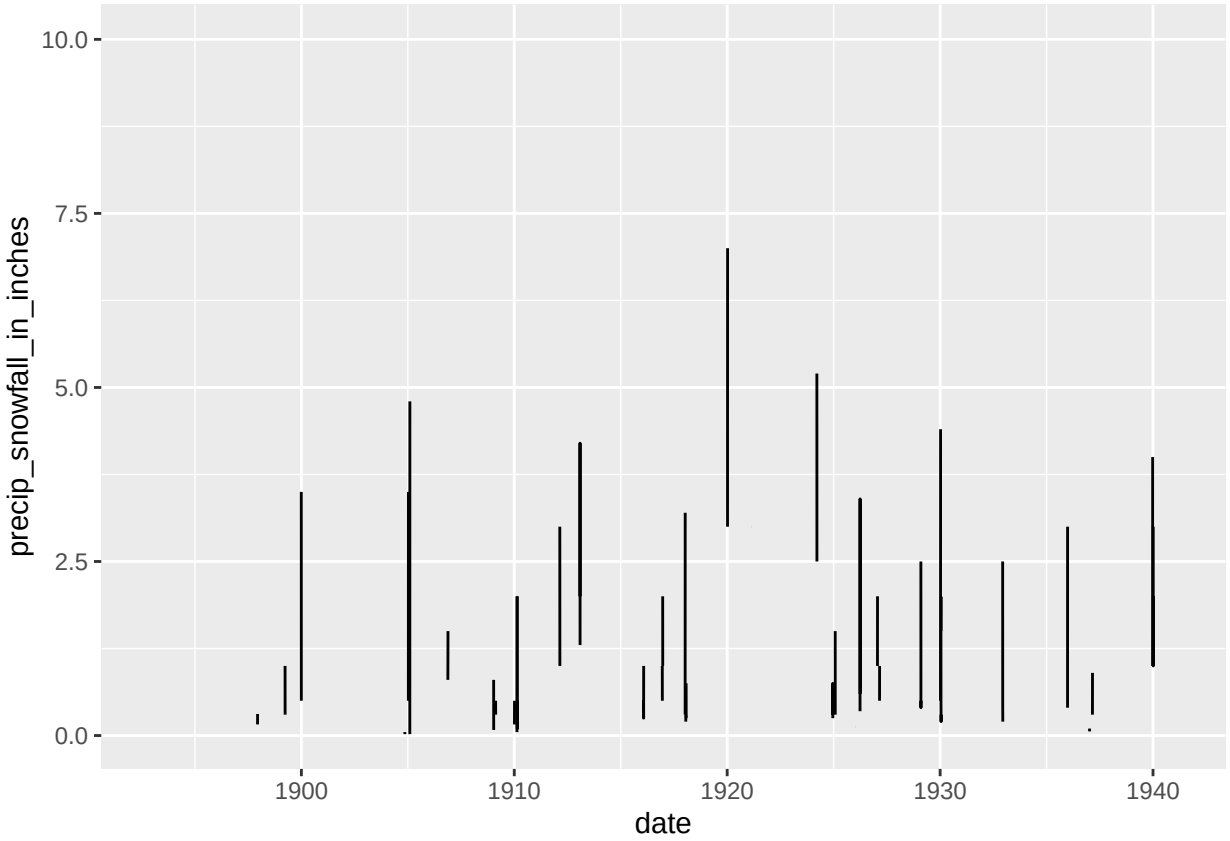


Snowfall in Inches (precip_snowfall_in_inches)

There does not appear to be any noticeable problems in this column.

```
wth_data %>%
  ggplot()+
  geom_line(aes(x = date, y = precip_snowfall_in_inches))
```

```
## Warning: Removed 332 row(s) containing missing values (geom_path).
```



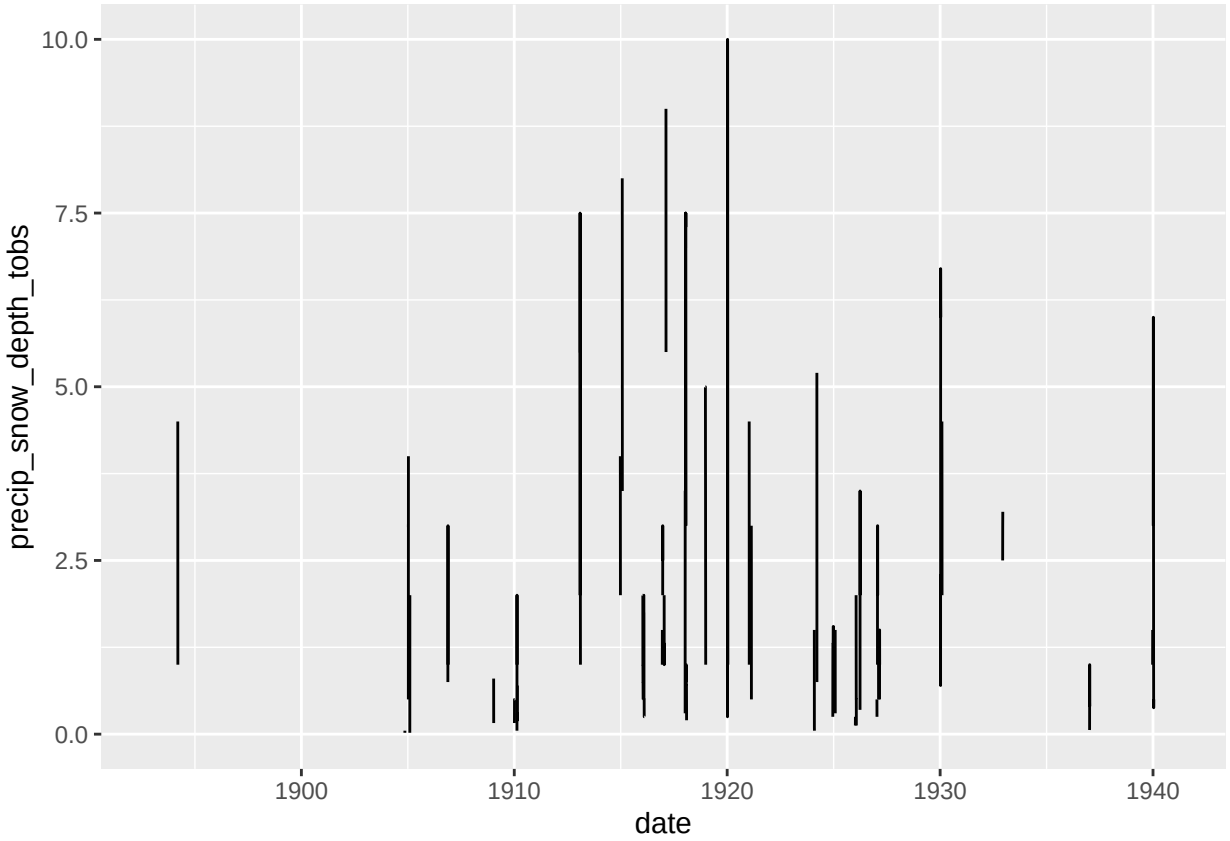

```
snow <- wth_data %>%
  select(date, precip_snowfall_in_inches) %>%
  drop_na() %>%
  arrange(desc(precip_snowfall_in_inches))
```

Snow Depth at Time of Observation (precip_snow_depth_tobs)

All values in this column seem to be reasonable amounts.

```
wth_data %>%
  ggplot()+
  geom_line(aes(x = date, y = precip_snow_depth_tobs))
```

```
## Warning: Removed 569 row(s) containing missing values (geom_path).
```




```
snow <- wth_data %>%
  select(date, precip_snow_depth_tobs) %>%
  drop_na() %>%
  arrange(desc(precip_snow_depth_tobs))
```