MAS286 Assignment 3: Data Visualization

Data Origins:

We found our data from an Altair book authored by By Jeffrey Heer, Dominik Moritz, Jake VanderPlas, and Brock Craft regarding data visualisation techniques and methods (https://uwdata.github.io/visualization-curriculum/intro.html). In section 5.1. “Weather data” there was a hyperlink (https://vega.github.io/vega-lite/data/weather.csv) with a csv file containing weather statistics for the U.S. cities of Seattle and New York. The data contains 7 columns about several observational indicators for instance, precipitation, max temperature, min temperature, wind and the weather type. The data spanned a 4 year timeframe with the first data point being on the 1st of January 2012 and ending on the 31st December 2015, providing us with 2923 total observations.
The first three rows are as follows:

Location	Date	Precipitation	Temp_max	Temp_min	Wind	Weather
Seattle	01/01/2012	0	12.8	5	4.7	Drizzle
Seattle	02/01/2012	10.9	10.6	2.8	4.5	rain
Seattle	03/01/2012	0.8	11.7	7.2	2.3	rain

Research Questions:

- Has there been an increase in higher temperature days in US cities?
- If there are significant temperature changes in US cities, what is this likely to look like in the future?
- Is there a correlation between type of weather and wind in relation to the Beaufort scale for cities in the US?
- Which weather patterns are responsible for the strongest winds for cities in the US?

Data Preparation:

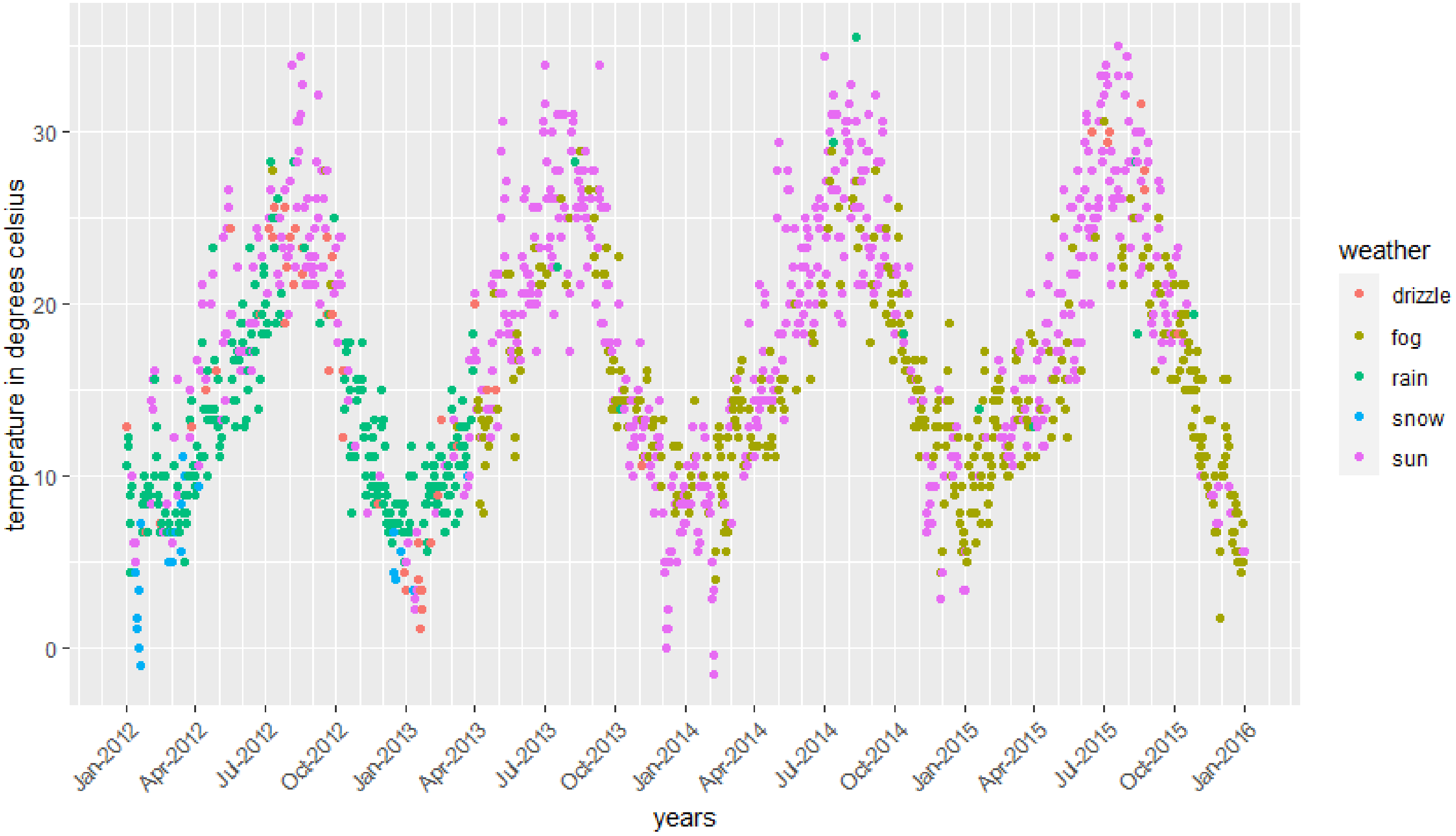
Our group aimed to focus on the change of weather in Seattle, so we removed all the data relating to New York by creating a new dataset labelled “weather1”. From there we selected the variables that were useful in answering our research questions and made plots to visualize the data. For instance for figure 1, we took the variables of maximum temperatures, date and weather to find whether the hottest days in Seattle had increased. We also used the same dataset to form our boxplot for figure 2 but instead utilising the wind and weather types variables.

Visualizations:

To start we wanted to answer our query regarding whether there were any large changes in temperature in Seattle. We chose to represent the data with a scatter graph because we thought that it would be the easiest way to spot an emerging pattern if it were present. We believed that this representation would be the best way to show the data, so that we could find an answer to the question.

Figure 1: Graph showing maximum temperatures in Seattle for a single day from 2012 to 2016.

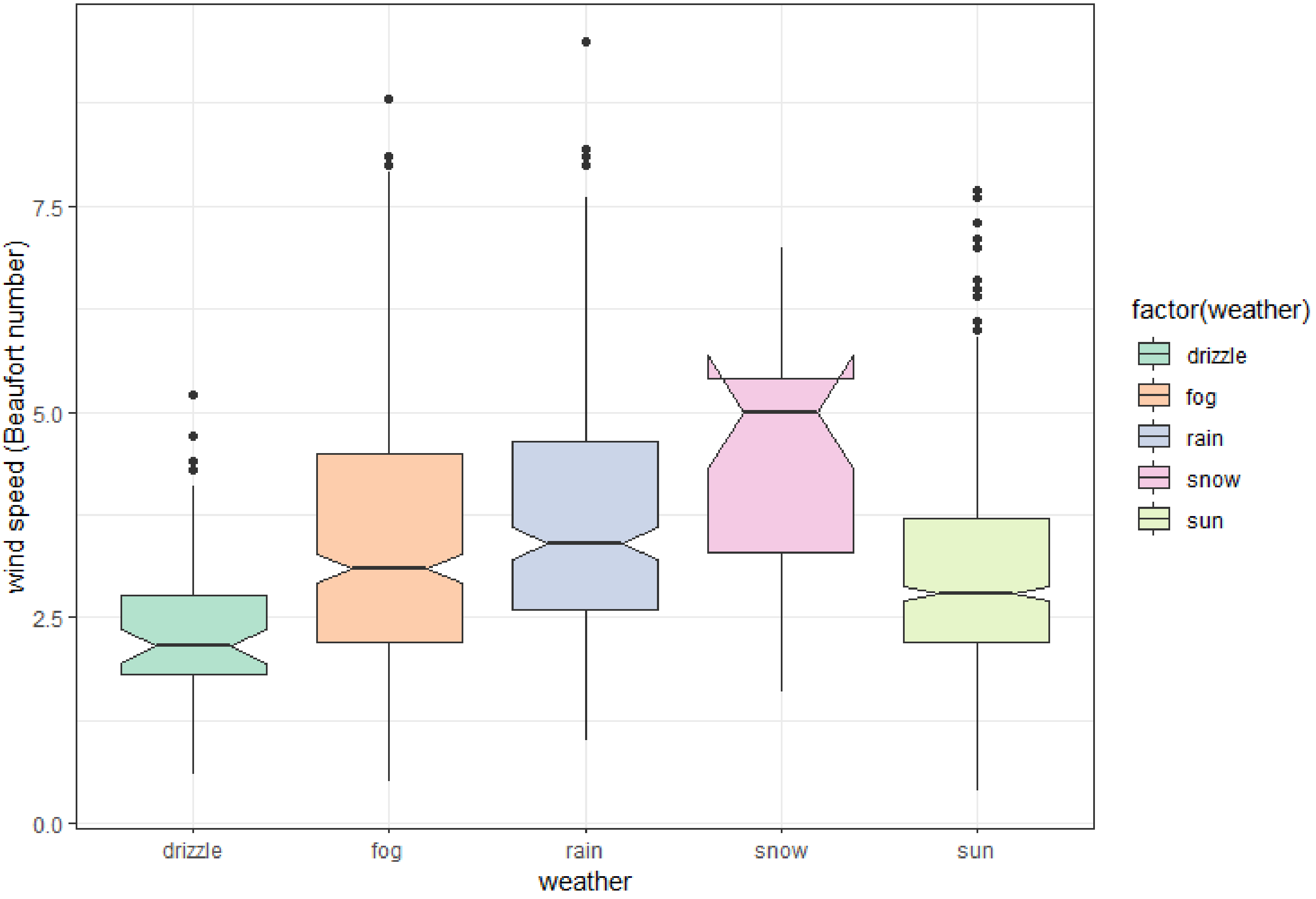
Data points are maximum temperatures for a specific day.



Source: uwdata.github.io

We chose to visualise the relationship between wind and weather with a boxplot, as we believed that it would be the clearest way to compare distributions of measurements from multiple different groups. Otherwise having used a different type of graph, with the multiple discrete variables, may have created a more confusing representation of the data. Compared with traditional charts, boxplots contain richer information and it is not affected by outliers; thus accurately and stably depicting the discrete distribution of the data as shown in figure 2 below.

Figure 2: Graph showing the relationship between wind and weather from January 2012 to January 2016 in Seattle.



Source: uwdata.github.io

Summary:

From a macro perspective, the first visualisation depicts a periodic increasing and decreasing pattern of temperatures as the number of days increase. This most likely coincides with the seasons of the year with winter being responsible for the initial lower temperatures then as it turns to spring it gradually increases, peaking around summer, and then decreasing once autumn arrives only for the cycle to repeat again. We can also see that Seattle's weather changes throughout the years, as the first (and the beginning of the second year) was mostly dominated by rainfall during the colder months; whereas in the later years rainfall is minimal and fog is the more abundant weather pattern. Furthermore, there is an increase in sunny days as time passes signifying that the weather pattern is changing to favour a hotter climate. This can also be evidenced by looking more precisely at the concentration of purple dots which are increasing not only for the hotter seasons but also for the colder months. This would suggest that we would see an upward trend in our graph as we would expect average temperatures to increase since higher temperatures are correlated with sunnier weather. However, our graph does not really show this, in fact it retains a strangely similar distribution throughout despite the weather changes.

From figure 2, we can look at the distribution of wind speed depending on the different weather types. It is clear to see that even though days of snow resulted in the greatest winds on average, rainy and foggy days had some of the strongest wind conditions. Moreover, we can estimate that the average wind for the three years would be approximately scale three which is equivalent to a gentle breeze. From the boxplot, wind speeds are lower for days that experience drizzle and sun whereas they seem to be higher during harsher weather patterns such as snow and rain. This suggests that there is a positive correlation between the harshness of the climate types and the wind speed which aligns with our own intuitive expectations from our common knowledge about the weather. On the other hand, there were many outliers for sunny days which would seem to counter the idea that harsher weathers have stronger winds. Although the term “harsh”, that describes the weather, is quite subjective meaning that there is little empirical evidence to support either side of the statement, indicating that we may need to do further data gathering in order to reach a conclusion.

One limitation of our dataset is that four years is not a large enough timespan when measuring something like change in temperature; however, using our plot we could predict what might occur (holding other variables constant). We hypothesise that there will be a continuation in increasingly hotter days. Hence, not only will more days get hotter, but also the temperature for these days will likely be higher as well. Since hotter days are usually drier and more humid we can assume that rain and snow would become more uncommon, compared to sun and fog.

If we had more time we would have liked to find a data set with a longer span of time so that we could depict more representative trends in the change in maximum temperature and understand Seattle's weather better.

Stretch Task:

We used the Shiny packages available to R studio to get an interactive scatter plot of the data, similar to the one shown in figure 1. Now it is possible to change the dependent variable to be any of the 7 in the dataset, including being able to switch the location between New York and Seattle. This allows us to compare the two cities seamlessly for any of the possible variables. Additionally, there is a small calendar date range selector for the start and end date of the plot meaning that we can manually select a specific timeframe in which the data is graphed. To further progress our visualisation, we could have found a larger dataset with other cities (or countries) and replicated our analysis and then compared them to see if we could measure whether this relationship was present additionally in other areas.