# Information Visualisation – Coursework Report

For this Information Visualisation project, a multi-variate regression data set about weather forecasting during summer in Seoul, South Korea, was chosen (UCI Machine Learning Repository, 2021). It contains 7750 data points which are made up of 25 attributes (see Appendix, figure 5.1 for a list of the 13 attributes used in this project, listed by name, definition, interval of values and unit).

The purpose of this project was to discover the ways in which these weather parameters are interrelated and to use those relationships to uncover hidden trends in the weather. With this in mind, a set of 4 initial questions listed below were formulated to give a basic understanding of the data:

1. Which weather station reached on average the highest daily *max* air temperature and which one reached the lowest daily *min* air temperature?
2. How did the temperature change over summer from 2013-2017 and how accurate were the weather forecasts?
3. Which stations were likely to be windier during summer?
4. Was there a relationship between the forecasted cloud coverage and the forecasted precipitation in the mornings from 6-11am?

The first step after importing the data was to perform data cleaning and filtering by removing the rows with null values and the redundant columns. Subsequently, the data had to be manipulated to answer each specific question. In doing this, the most suitable solution that created robust and maintainable code, was to copy the relevant data from the main data set and create a separate table for each new question where values could be added and removed safely.

For question 1, the x-axis mapped discrete values representing a nominal variable, i.e., the number of the weather stations. Therefore, the initial thought was to visualise the data as a bar chart which also seemed suitable since the y-axis had continuous values representing quantitative data, i.e., daily max and min temperatures. With reference to data pre-processing and manipulation, an average daily max and min temperature was calculated from the dates available to give a generic overview of the summer temperatures for each station from 2013-2017. In support of this, the visual representation of a bar chart emphasises individual values since it can be grouped to enable visualisation of both the max and the min temperature for each station. In the main data set, the max temperatures were given in a separate column from the min temperatures so they had to be merged with a new column added for a variable that showed whether the value was of max or min type. However, after implementing the idea, it became evident that the visualisation fell victim to the Moiré effect with a cluttered chart of similarly high values that made it difficult for the reader to tell values apart. As a solution, a stacked lollipop chart was selected for the final visualisation as it follows a more minimalistic design and thereby simplifies the visualisation for the reader (see Appendix, figure 1.1). To read the values of each point, the reader would simply have to check its position, which is an effective way of reading quantitative data. Furthermore, the data was reconfigured to show an arrangement that would provide the user with an additional perspective – the stations on the x-axis were sorted by their max temperature in an ascending order. Doing this is in alignment with the design criteria of expressiveness to enable the reader to accurately compare similar values and therefore find real patterns in the data. It is also an effective way to visualise data to be *compared* since ratios are more important than magnitude. Lastly, to visualise the variable determining whether a temperature value was a max or a min point, visual encoding in the form of colour hues was used. Using different colours hues in this way, with conventional ones representing heat (red) and cold (blue), for an unordered nominal variable (Max or Min), is highly recommended according to Jacques Bertin’s levels of organisation and is an effective way of grouping data. Finally, to ensure that the data was being perceived in the correct way, the graph was made interactive using the plotly library. It allowed the user to hover over data points to see the station number, temperature and type, and to filter the data by double-clicking on a type in the legend, helping the user focus. From the plot, it is evident that station 1 is the coldest and that stations 18 and 23 are the hottest. Furthermore, there seems to be a relationship between the min and max temperature where a lower min temperature usually is paired with a lower max temperature.

For question 2, the purpose was to visualise the average measured and forecasted max temperature per day in Seoul, and the difference between the two max temperatures. The first step for the data manipulation was to extract all unique dates into a new data frame where the year would be in a separate column, enabling grouping of data points by year to see how the annual changes in the summer weather. The second step was to calculate the measured and forecasted temperatures. Finally, the average difference between the measured and the forecasted temperatures was calculated to measure the accuracy of the forecasting model used in the data set. Since the data was a time series of quantitative values changing over a continuous period, a line chart was deemed the most suitable to visually represent it (see Appendix, figure 2.1). To avoid creating a spaghetti chart, cluttered with too many lines, the plot was faceted by year into a trellis display, creating 5 different plots, each visualising a measured temperature value and a forecasted one. The average difference between the two lines was added as a text annotation to the start of the first graph with its definition declared in the legend. Whether a temperature was measured or forecasted is a variable of unordered nominal values and therefore, it was visually encoded by colour hues, where the selected hues (green and grey) were neutral in the context of temperature (i.e., not red or blue). The definitions of the colours can be seen in the legend or by hovering over a line which would also display the exact temperature and date of that point on the line in an interactive manner (see Appendix, figure 5.2). Reading the chart, it is clear that the temperature started at around 30 °C each summer, reaching a peak of 35 °C around the start of August, with a small average of 0.5 °C difference between the measured and forecasted values every day, indicating that the weather forecasting model used was very strong.

For question 3, the main data set was sufficiently fit to answer the question without any significant data manipulation. The only pre-processing done was reordering the stations by their wind speed to help the reader more accurately compare similar values. To answer the question, a box plot visualising the *distribution* of the data set was initially preferred, since it would tell which stations were likely to be windier than others and which wind speeds could be expected more than others at each station. Box plots are indeed a good way of easily visualising the summary of the variation in a large data set. However, to also show outliers and give more insight into the data distribution in the form of a probability density, a violin plot was selected as the final visualisation strategy (see Appendix, figure 3.1) – a variant of the box plot. The mean value of each distribution was also marked out in each violin so the reader can easily follow the mean points to see a trend. No additional parameters were added to require visual encoding so the colour green was simply chosen for the violins to help them stand out from the white background due to its neutrality in this context, as mentioned before. Similar to the previous plots, this one was also made interactive using the plotly library, which allows the user to hover over parts of the distribution to see exact values of specific data points. From the plot, it is evident that stations 7 and 22 are the top 2nd and 3rd least windy stations respectively, with their wind speeds usually ranging from 4-6 m/s. Interestingly, figure 1.1 (relating to question 1) showed that these stations also had the highest min temperatures recorded among the stations, indicating of a potential relationship between wind speed and temperature. Further supporting this relationship, stations 1 and 12, which were the top 2nd and 3rd windiest ones, were also part of the top 3 coldest stations according to figure 1.1.

For question 4, the purpose was to plot out the percentages of cloud coverage and precipitation in the summer mornings to discover if there was a relationship between the two weather parameters. The data in the mornings was chosen instead of calculating a daily average in an effort to avoid overplotting and the potential relationship between cloud coverage and precipitation was assumed to be independent of time for the purpose of this project. To answer this question, the main data set was fully sufficient without any further data manipulation. A scatter plot was chosen as the visualisation strategy (see Appendix, figure 4.1) since it is suitable to represent quantitative bivariate data. It will visualise any existing relationship and a smoothed line was added using a generalised additive model (gam) to show the trend. This makes it easier for the reader to get a general idea of the relationship regardless of whether it is linear or non-linear, specifically in the presence of a somewhat cluttered graph. In addition, the data points were plotted with 20% opacity in order to visualise which values were more common, i.e., values that were plotted overlapping each other and making the area appear darker. To enable the user to read exact values from specific data points, this plot was also made interactive. From this plot, it is evident that there indeed is a relationship between the level of forecasted cloud coverage and the precipitation in the mornings, where both values increase in parallel at different speeds – the most common values (where the area is the darkest) being between 0-70% cloud coverage and less than 1% precipitation. Nonetheless, the biggest changes seem to occur as cloud coverage rises to approximately 70-90%, and precipitation increases to vales ranging from 2-6%. In some rare cases, precipitation increases to vales ranging from 10-20%.

The results from questions 1 to 4 were developed respectively to further investigate the inter-parametric relationships based on the results obtained. Following are the extended questions derived in order of the initial questions.

1. What was the average daily incoming solar radiation and heat flux per station and was there a relationship between them and the air temperature? Moreover, did the slope of the stations affect this relationship?
2. When did the biggest changes in temperature over summer in 2013-2017 in Seoul happen and what was the general trend?
3. Was there a relationship between the wind speeds and the elevation of the stations?
4. Was the humidity related to the levels of cloud coverage and precipitation?

Due to the previously discovered relationship between the max and the min temperatures in figure 1.1, it seemed fitting to investigate whether solar radiation and heat flux were also a part of that relationship, assuming both values would increase with the temperature. Taking inspiration from a multi-dimensional bubble chart, figure 1.1 was extended by visually encoding the temperature-points by colour and size to include data from the two new variables. Similar to the manipulation of the max and min temperatures for figure 1.1, the solar radiation and heat flux values were also calculated as averages over the entire timespan of the data set per station. Since solar radiation has the biggest interval of values, i.e., the values changed by a larger magnitude between the stations, it was chosen at the colour-encoded variable. This choice was made in the hopes that the large differences would appear a bit clearer than the smaller ones in heat flux since colour-encoding by value is not the most efficient way to visualise changes in a continuous quantitative variable. A continuous colour scale going from light to dark values (yellow to red) was chosen in the end. The heat flux variable, which was encoded by size, was binned by increments of 60 W/m2 to ensure that the stations’ values could be compared accurately at *significant* changes. This technically converted the variable from a quantitative continuous one to a nominal one. Despite nominal variables generally not being advised for encoding size, it was deemed acceptable since the variable in this case is ordered. The two new variables are defined for the reader in the legend for clarification. Contrary to the hypothesis, it seemed as if neither solar radiation nor heat flux depended on temperature, since both variables changed rather randomly as the temperature increased, which was a significant surprise derived from the data (see Appendix, figure 1.2). After some research into the matter, it was discovered that the topology, more specifically the slope of the land, had more influence on solar radiation since it affected how the sun rays hit the ground (Solar Radiation Basics, 2021). This led to refining the question by sorting the stations according to their slope instead of their max temperature, which interestingly visualised a relationship where a higher slope meant more solar radiation (see Appendix, figure 1.3). Nevertheless, a relationship between heat flux and the other parameters still remained undisclosed.

To further explore the changes in temperature over summer viewed in figure 2.1, some measures were added to an extension of that figure (see Appendix, figure 2.2), which would provide more general insight in an effective way. Firstly, an orange dashed line was added, representing the average temperature over summer throughout the included years. This would quickly visualise the general temperature during summer in Seoul. Secondly, in support of the average value, the standard deviation was marked out as a yellow area over the plot. This would visualise the average range of values to be expected and highlight any potential outliers which would be plotted outside the area. Finally, to further refine the average with respect to time and visualise a trend of the temperature, a rolling average with a window of 7 days was calculated and visualised as a dashed grey line. To keep in line with the original plot (figure 2.1) answering the initial question 2, the line representing daily max temperature was kept the same colour, and the line representing the forecasted max temperature was removed as it was redundant to this extended question. All colours were chosen as different hues since they represented different nominal variables and the definitions were added to the legend. The plot was made interactive so they reader can easily hover over specific data points to see exact values and select the lines to be shown by clicking on the legend. The average summer temperature in Seoul was 30°C between the years of 2013 to 2017 with a standard deviation of 3°C. The rolling average shows that the temperature was lower in July than in August, with a general annual increase. The outliers in the plot also seem to be growing further from the average value, indicating on an increase in magnitude of change, with the temperature usually reaching its peak at around the start of August.

Figure 3.1, answering the initial question 3, revealed that weather stations with colder climate were generally windier, and since mountainous regions are known to be colder, an assumption was made that the topology of an area (DEM, i.e., elevation) affects the wind speeds of it. To test this theory out in this extended question 3, the violins, representing the distribution of wind speeds in each station in figure 3.1, were visually encoded according to DEM by colour-values ranging from dark blue (low DEM) to light blue (high DEM) (see Appendix, figure 3.2). In accordance with DEM being a continuous quantitative variable, the colour scale was also made continuous. For reference, there was no need for any additional data manipulation. The specific DEM value of each station may be seen by hovering over the respective violins, also showing the wind speeds of the stations. In conclusion, the theory was correct and the wind speed did in fact decrease with the elevation. Station 17 had the lowest wind speed and also the lowest elevation and interestingly, figure 1.3 showed that it also had one of the highest amounts of average heat flux, which is in alignment with the assumption that triggered the creation of this plot – that mountainous areas are colder, meaning that less elevated areas are warmer.

Since figure 4.1 visualised that precipitation grows with cloud coverage, the matter was further explored by testing if humidity was also influenced in the same way. To do this, the points in the scatter plot in figure 4.1 were encoded by colour with values ranging from light to dark purple, representing the levels of daily max humidity (see Appendix, figure 4.2). To avoid confusing the reader when translating colour values into humidity values, the transparency of the points was discarded since it affects how the colours appear. The results showed that the points with higher cloud coverage and thereby also higher precipitation were darker, indicating that the max humidity levels increased with cloud coverage and precipitation, confirming the hypothesis.

The project was successful in exploring the many relationships between weather parameters and discovering trends in the summer weather in Seoul. The visualisation strategies and encodings used were an effective way of refining the questions and communicating answers in an expressive manner. In conclusion, Seoul appears to be a hot place with not much rain and wind speeds at around 5-7 m/s. Potential future implementations include investigating how precipitation changes within a day and what other parts of the topology affect wind speed and heat flux.

## Sources

Archive.ics.uci.edu. 2021. UCI Machine Learning Repository: Bias correction of numerical prediction model temperature forecast Data Set. [online] Available at: <https://archive.ics.uci.edu/ml/datasets/Bias+correction+of+numerical+prediction+model+temperature+forecast> [Accessed 19 April 2021].

Energy.gov. 2021. Solar Radiation Basics. [online] Available at: <https://www.energy.gov/eere/solar/solar-radiation-basics> [Accessed 19 April 2021].

## Appendix

Figure 1.1

A picture containing text, appliance

Description automatically generated

Figure 1.2

Chart

Description automatically generated

Figure 1.3

Chart, scatter chart

Description automatically generated

Figure 2.1

A picture containing text, indoor

Description automatically generated

Figure 2.2

Chart

Description automatically generated

Figure 3.1

Chart, histogram

Description automatically generated

Figure 3.2

A picture containing chart

Description automatically generated

Figure 4.1

Chart

Description automatically generated

Figure 4.2

Chart, scatter chart

Description automatically generated

Figure 5.1

Text, letter

Description automatically generated

Figure 5.2

Chart

Description automatically generated