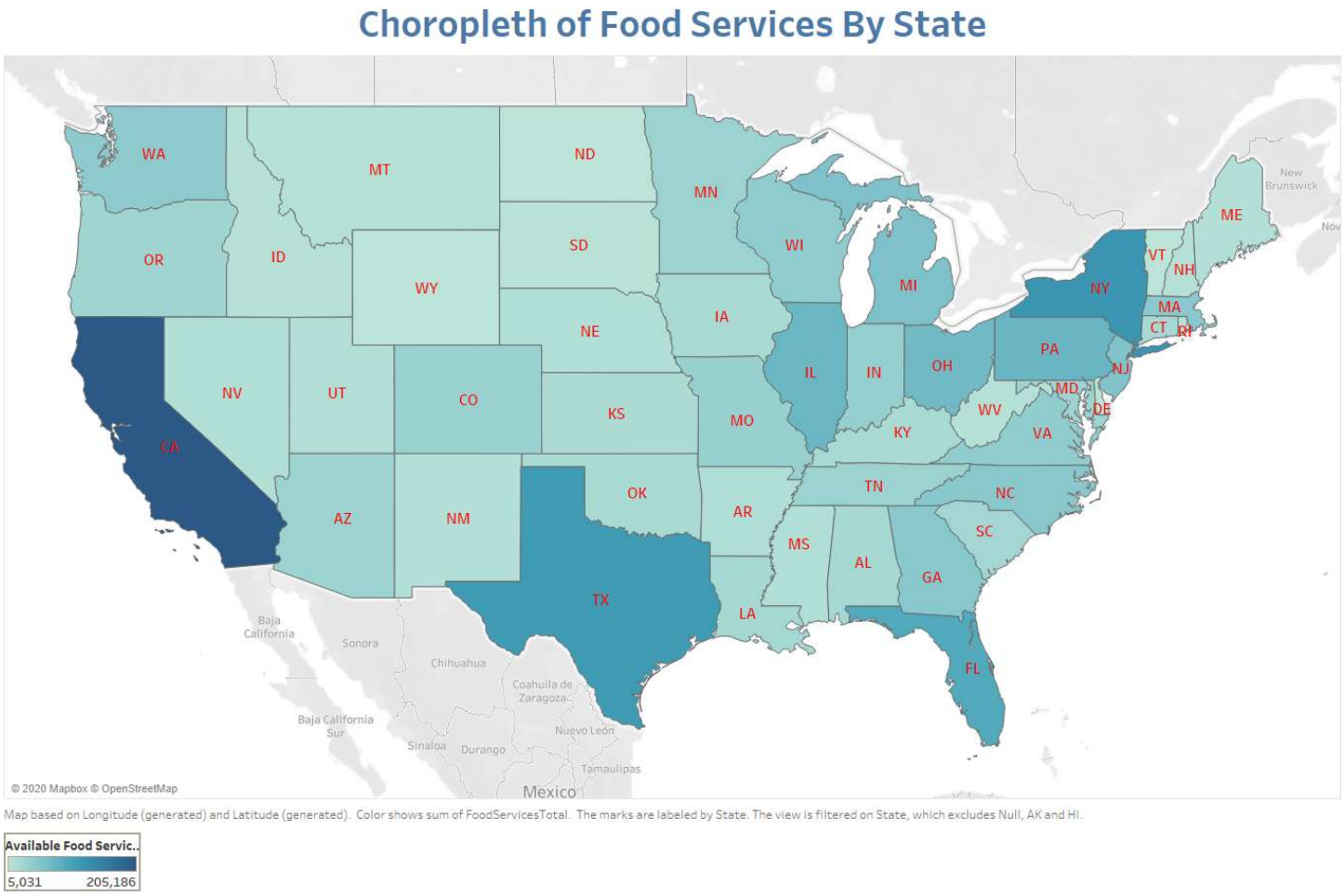
Umair Chaanda
DSC 465 Data Visualization
Homework-2

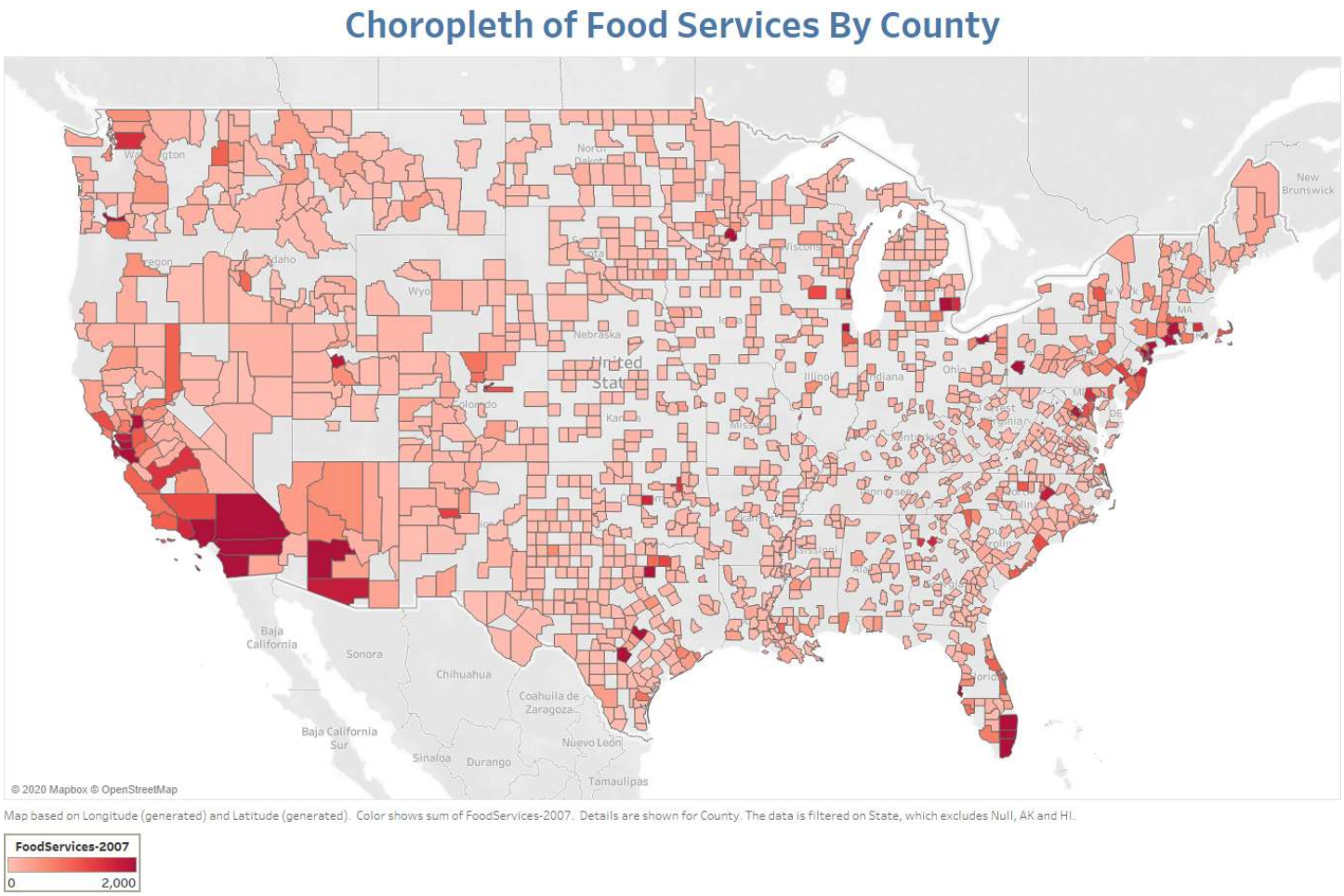
Problem 1: Download the FoodSrvcByCounty.txt file and create the following visualizations for this geographical data. The data is for the availability of food services by county in the U.S. It also has data by state (in the county field, some of them have the state names, and those rows hold the state totals, or you can aggregate by state).

a) Food services by state with an appropriate geographic visualization.



- I chose Choropleth graph to map this data. This map is based on Longitude and Latitude coordinates. It uses color to fill in the map with availability of food services.
- I created a new calculated field "**FoodServicesTotal"** which is the sum of three years [FoodServices-97] + [FoodServices-2002] + [FoodServices-2007].
- Color shows SUM of Food Services Total. The marks are labeled by State.
- The view is filtered on State which excludes **Hawaii (HI)**, **Alaska (AK)** states and **Null** values.
- The graph shows that **California**, **Texas**, and **New York** states have higher food service availability.

b) Food services by county with the same type of visualization.

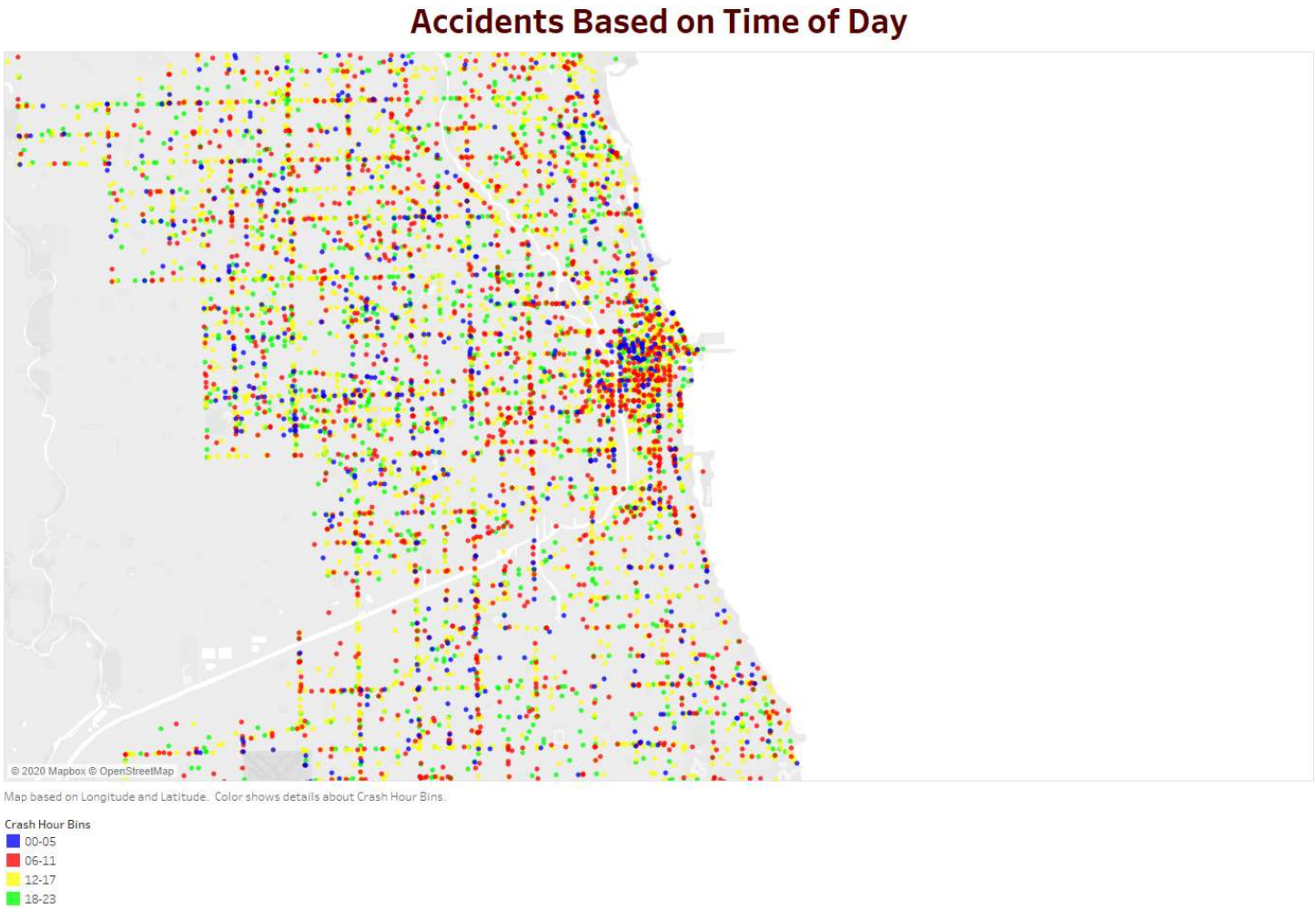


- Color shows SUM of Food Service-2007. Details are shown for each county.
- The view is filtered on State which excludes **Hawaii (HI)**, **Alaska (AK)** states and **Null** values.
- The graph shows that some of the counties in **California**, **Florida**, **Washington** and **NY** states have higher food service availability.

Problem 2: The Chicago_crashes.csv file contains information on every crash recorded in Chicago in June2019 (see Chicago's portal at https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if for the latest data. I chose a random month because the data get dense quickly).

a) Geographic plot to show where all the accidents in this data occur.

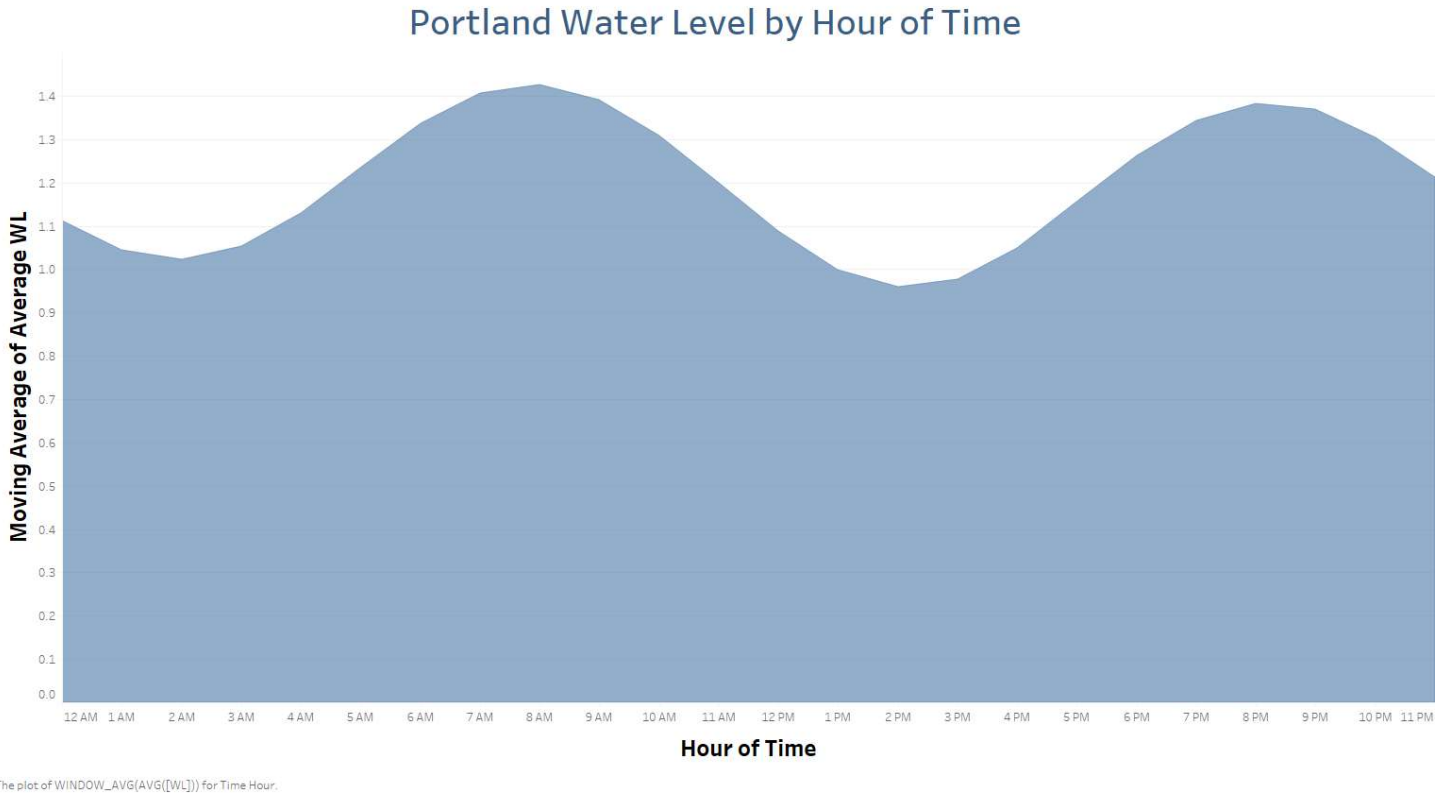
b) Visualization that shows how common crashes are in different parts of the city based on time of day.



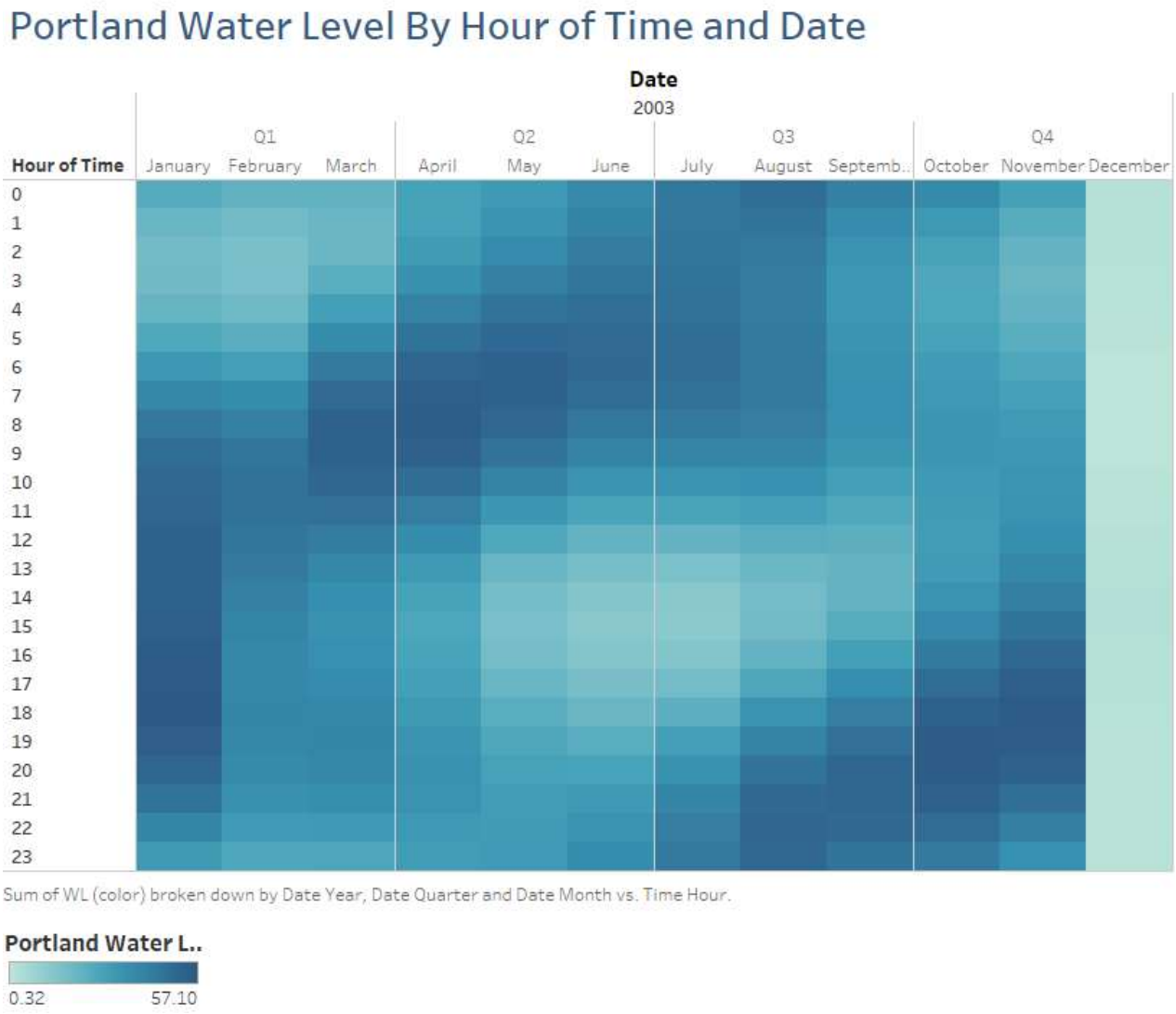

```
IF [Crash Hour] <=5 THEN "00-05"
ELSEIF [Crash Hour] >5 AND [Crash Hour] <=11 THEN "06-11"
ELSEIF [Crash Hour] >11 AND [Crash Hour] <=17 THEN "12-17"
ELSEIF [Crash Hour] >17 AND [Crash Hour] <=23 THEN "18-23"
END
```

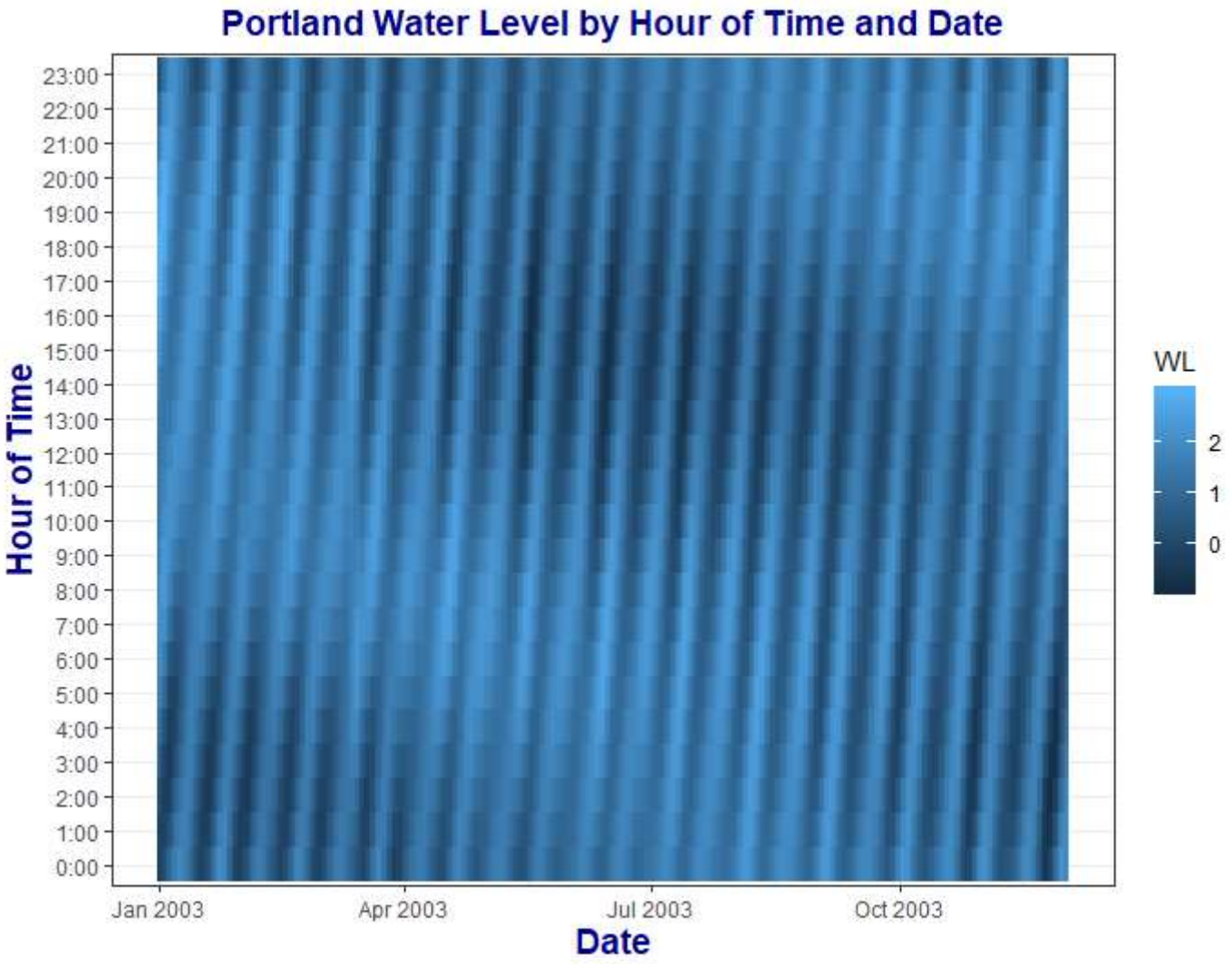
Problem 3: Download the Portland Water Level dataset and explore it by creating the following visualizations of the time series from the techniques described in lecture. Use both R and Tableau for at least one question part. They should, of course, adhere to the design criteria that we've learned, and should clearly display the information described in each part.

a) This data contains a year of data with water level (WL) measurements every hour as a function of Time (i.e. 365 x 24 data points!). Since there is a lot of data, clean it up by smoothing the data by calculating a moving average. Use a window approach with window size that covers a range of days (remember, the data is hourly) and graph the smoothed result. Work with the window to see what size window gives you the best view of the changes in the data while still smoothing the noise well.



b) Graph the cycles that happen each day (because of tides). You might try overlapping many days' data as separate overlapping time series, using a level plot, a horizon graph, etc. The point of this exercise is to try to come up with a way of showing the progression of the tides over some period of time that is rich and detailed and which shows the pattern, but which is still readable and which doesn't clutter the graph.

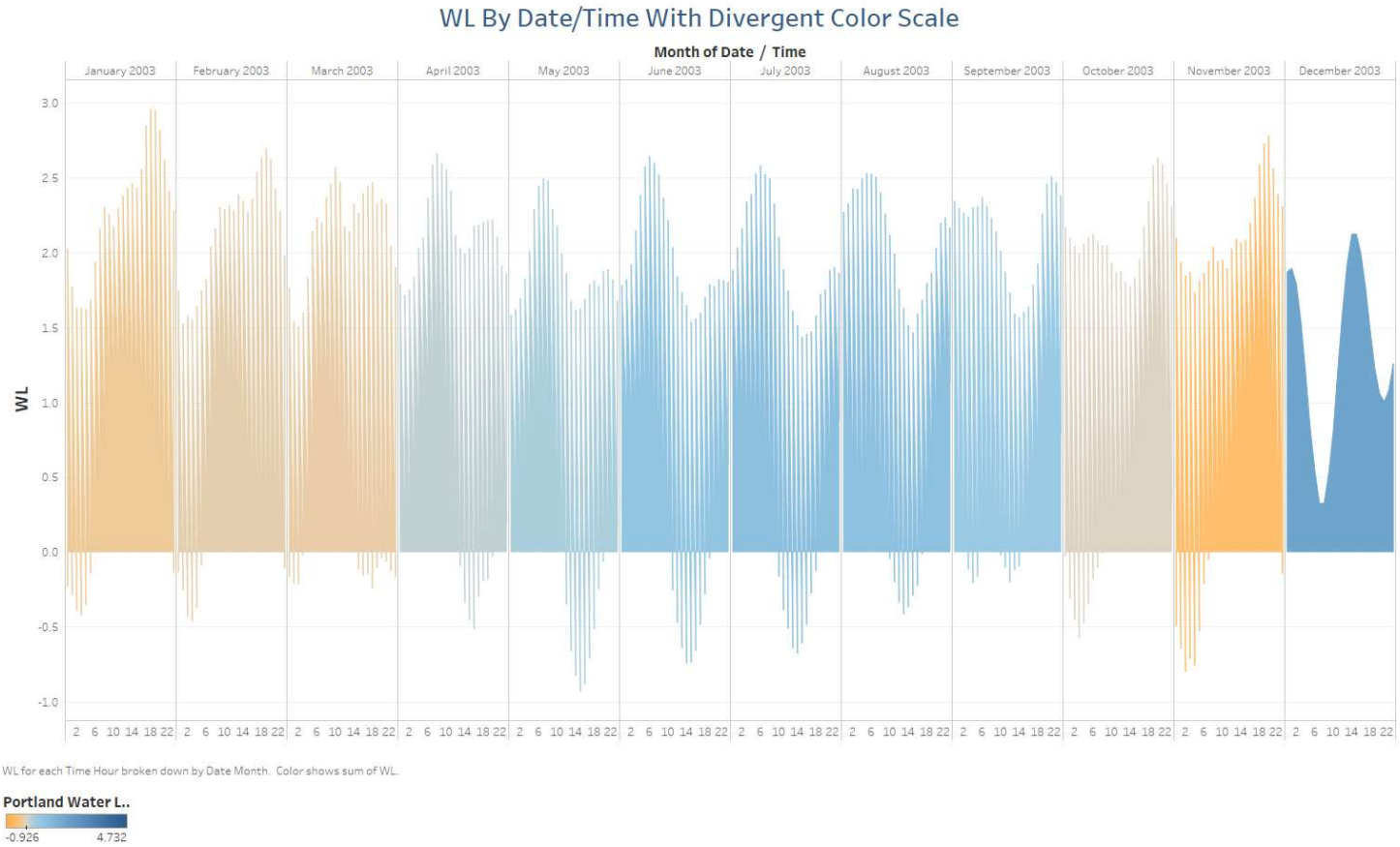




c) Then write a single paragraph outlining the differences between the information that each graph communicates.

In graph (a), I did smooth the data by calculating the moving average of average water level using window approach. The first graph shows the average fluctuation in water levels over hour of time. We can see that the water levels are higher during morning and evening times, whereas water levels are much lower during afternoon. The graph (b) shows the cycles that happen by hour of time and date. It shows the overlapping time series data pattern which is good way of showing the progression of the tides over some period. We can see that the water levels are higher during some months and time of the day. The graph (b1) was built in Tableau whereas graph (b2) was built in R using ggplot.

Problem 4: Return to the Portland Water Level dataset. Recreate one of your plots from Question 3 with a custom color scale. Specifically, create a divergent color scale with the average water level at the midpoint and two separate colors used to show when the water is getting very high and very low. The point of this exercise is to experiment with creating a color scale, so choose your own distinctive colors to use for the endpoints and center. Make sure that they are reasonable choices given what you know about color scales. Use HSV space to choose the colors and explain how you made your decision. In Tutorial 4, you can see how to create color scale in ggplot that is interpolated in Lab space.



- I recreated this plot using a custom color scale. Specifically, a diverging color scale of **Orange - Blue**.
- I created the color contrast by using colors from opposite sides of the HSV cone which emphasize both extremes.
- It shows the average water level at the midpoint with two separate colors **Orange** and **Blue** to show endpoints e.g. when the water is getting very low and very high respectively.