

Week 9 Assignment Part 2 (R/RMarkdown)

Due Nov 16, 2022 by 10am **Points** 5 **Submitting** a file upload

Available Nov 7, 2022 at 12am - Nov 16, 2022 at 10am


This assignment was locked Nov 16, 2022 at 10am.

This assignment returns to the crime data that we worked with earlier in the term, and will give you some practice using pivots and joins to enrich your data with additional information.

As usual, follow the instructions below in an RMarkdown document. Don't forget to include your name, student number, and an informative title. Please make sure that your knitted document includes the R code chunks (so don't use the "echo=FALSE" option). We want to see your code and the results! You should suppress unnecessary messages and warnings using the appropriate code chunk options. When you've answered the questions below, save your .Rmd file and knit the RMarkdown document to produce a nicely formatted html document. To complete the assignment, submit BOTH your RMarkdown (.Rmd) file AND the knitted .html file via canvas.

Your instructions are:


Part 1

- Download the incident-level Vancouver crime data
[crimedata_csv_AllNeighbourhoods_AllYears.csv](https://canvas.sfu.ca/courses/73572/files/20033041?wrap=1)
(<https://canvas.sfu.ca/courses/73572/files/20033041?wrap=1>) 
(https://canvas.sfu.ca/courses/73572/files/20033041/download?download_frd=1) and import them into R.
- Subset the data to only keep observations between 2006 and 2021.
- Use the `lubridate` package to create a properly formatted date variable.
- Compute a grouped summary of your data to count the total number of incidents of each type on each date.
- Pivot your counts of the total number of incidents of each type on each date so that you end up with a data object where there is one row for each date, and the columns measure the number of crimes of each type.




If you're done everything correctly to this point, your data should have 5844 observations (there are 5844 days in the 16 years covered by your data). There also will be lots of missing values ("NA") in the columns of some crimes! Specifically, there should be a missing value if no crimes of a particular type were reported on a given day (e.g., the Homicide variable should take value "NA" on January 1 2006 because no homicides were reported that day).

- Replace the NAs in your data with zeros. You can do this variable-by-variable using `mutate()`. However, if you're looking for a more compact method, check out the `dplyr` function `replace()` and the `is.na()` function.

Note that replacing the NAs with zeros is important for what follows! Lots of crimes occur infrequently; when we compute statistics like the average daily count of some type of crime, we want to make sure that our average includes the days with zero crimes!

- Install the `vtable` package and use the `sumtable()` function to create a nicely formatted table of summary statistics for the daily number of crimes of each type. See [here](https://cran.r-project.org/web/packages/vtable/vignettes/sumtable.html)  (<https://cran.r-project.org/web/packages/vtable/vignettes/sumtable.html>) for a description and some examples.
- Calculate the annual average of the daily count of each type of crime for each year between 2006-2021, and then plot the annual averages. Comment on any trends you observe. Note your plot should clearly show all crime types in a single graph, and don't forget to label axes, give the plot a title, and do anything else you think improves the presentation of your visualization.

Part 2

- Import the [BCHolidays.csv](https://canvas.sfu.ca/courses/73572/files/20302281?wrap=1) (<https://canvas.sfu.ca/courses/73572/files/20302281?wrap=1>)  (https://canvas.sfu.ca/courses/73572/files/20302281/download?download_frd=1) and [weatherstats_vancouver_daily.csv](https://canvas.sfu.ca/courses/73572/files/20263397?wrap=1) (<https://canvas.sfu.ca/courses/73572/files/20263397?wrap=1>)  (https://canvas.sfu.ca/courses/73572/files/20263397/download?download_frd=1) files into R. The first of these is a list of holidays in BC from 2003-2021. The second includes daily weather information for Vancouver from the 1930s through November 2022, downloaded from <https://vancouver.weatherstats.ca>.  (<https://vancouver.weatherstats.ca>)
- Join the holiday data to your daily crime data by date, ensuring that the joined data cover the same 2006-2021 time period as your crime data!
- Create a logical variable that equals `TRUE` if the day is a BC Holiday and `FALSE` otherwise. *Hint: this is what you'll end up with if you use a logical test in a `mutate()` command. Most of the time, R will treat a logical TRUE/FALSE variable the same as a binary 1/0 dummy variable.*
- Use a mutating join to add the `max_temperature`, `min_temperature`, `avg_temperature` and `precipitation` variables from the weatherstats data to your daily crime data. Again, take care to ensure that the joined data cover the same 2006-2021 time period as your crime data!
- Create a variable measuring the day of the week
- Create a variable measuring month of the year
- Use `sumtable()` again to create a nicely formatted table of summary statistics for subset of your data that are holidays. Are there any notable differences between the average number of crimes on holidays vs. all days?

Part 3

- Choose one type of crime that interests you (but please don't choose one of the crimes whose average daily count is less than one!). Use `ggplot` to visualize the relationship between your chosen

crime and each of the four weather variables (`max_temperature`, `min_temperature`, `avg_temperature`, `precipitation`) in a scatter plot with overlaid linear regression line. This should be four separate plots, or perhaps four facets, unless you have a better idea! Each plot/facet should measure the crime variable on the vertical axis and the weather variable on the horizontal axis. Be sure to remove outliers if there are obvious ones skewing the relationship or making it hard to see the linear fit. Comment on any pattern you observe.

- Use `lm()` to estimate the following two regressions and add the residuals to your data.
 1. your crime variable ~ `max_temperature`
 2. your crime variable ~ `max_temperature` + day of week + month of year + holiday indicator

Note: you'll want to ensure that your day-of-week and month-of-year variables are unordered factor variables so that R automatically creates dummy variables in the regression for you; OR, if you prefer you can manually create dummy variables for each month and day of the week and include those in the regression.

- Calculate the mean of the squared residuals from the two regressions and comment on which has the smallest MSFE.
- Use the `broom` package to plot the regression coefficients and confidence intervals from the second regression above. Comment on the results. Does the confidence interval for maximum daily temperature include zero? What does that tell you? Which months and days of the week have the highest incidence of your chosen crime variable?