Homework 1\_Samanta

Aritra Samanta

1/13/2022

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✓ ggplot2 3.3.5 ✓ purrr 0.3.4  
## ✓ tibble 3.1.6 ✓ dplyr 1.0.7  
## ✓ tidyr 1.1.4 ✓ stringr 1.4.0  
## ✓ readr 2.1.1 ✓ forcats 0.5.1

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(nycflights13)  
library(lubridate)

##   
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':  
##   
## date, intersect, setdiff, union

library(dplyr)

1. Use and show R coding to find the number of days from June 6th 2020 to July 14th 2021

(("2020-06-06") %--% ("2021-07-14"))%/%days(1)

## [1] 403

1. Use and show R coding to confirm that the year 1988 was a leap year.

(("1988-01-01") %--% ("1989-01-01"))%/%days(1)

## [1] 366

#As there are 366 days between 1 Jan 1988 and 1989, 1988 is a leap year

1. Use and show R coding to confirm that the year 1989 was not a leap year.

(("1989-01-01") %--% ("1990-01-01"))%/%days(1)

## [1] 365

#As there are 365 days between 1 Jan 1989 and 1990, 1989 is not a leap year

1. Why is there months() but no dmonths ? (Answer in 3 to four sentences)

This is because months do not have a specified value that is true every time. Some months have 30 days,others have 31, and February has 28 or 29 depending on whether it is a leap year or not. There is no exact duration of a month that can be used in duration calculations–it changes based on which month and which year.

1. John was born April 11th, 1962. Use and show R coding to determine how old John is in years

(("1962-04-11") %--% ("2022-01-13"))%/%years(1)

## [1] 59

#John is 59 years old

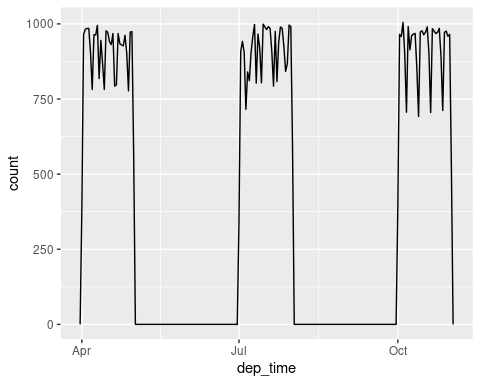
1. Modify the flights\_dt coding in the notes or the book to obtain the following partial data table shown below. Show all required coding. (Most of the coding needed is provided in the notes)

make\_datetime\_100 <- function(year, month, day, time) {  
 make\_datetime(year, month, day, time %/% 100, time %% 100)  
}  
  
flights %>%   
 filter(!is.na(dep\_time), !is.na(arr\_time)) %>%   
 mutate(  
 dep\_time = make\_datetime\_100(year, month, day, dep\_time),  
 arr\_time = make\_datetime\_100(year, month, day, arr\_time),  
 ) %>%   
 select(origin, dest, carrier, arr\_time, dep\_time) ->  
   
 flights\_dt  
flights\_dt

## # A tibble: 328,063 × 5  
## origin dest carrier arr\_time dep\_time   
## <chr> <chr> <chr> <dttm> <dttm>   
## 1 EWR IAH UA 2013-01-01 08:30:00 2013-01-01 05:17:00  
## 2 LGA IAH UA 2013-01-01 08:50:00 2013-01-01 05:33:00  
## 3 JFK MIA AA 2013-01-01 09:23:00 2013-01-01 05:42:00  
## 4 JFK BQN B6 2013-01-01 10:04:00 2013-01-01 05:44:00  
## 5 LGA ATL DL 2013-01-01 08:12:00 2013-01-01 05:54:00  
## 6 EWR ORD UA 2013-01-01 07:40:00 2013-01-01 05:54:00  
## 7 EWR FLL B6 2013-01-01 09:13:00 2013-01-01 05:55:00  
## 8 LGA IAD EV 2013-01-01 07:09:00 2013-01-01 05:57:00  
## 9 JFK MCO B6 2013-01-01 08:38:00 2013-01-01 05:57:00  
## 10 LGA ORD AA 2013-01-01 07:53:00 2013-01-01 05:58:00  
## # … with 328,053 more rows

1. Now, using your table from #6, produce the frequency plot shown which conveys frequency counts for the months of April, July, and October for the year 2013

flights\_dt%>%  
 mutate(month = month(dep\_time, label = TRUE))%>%  
 filter(month=="Apr"|month=="Jul"|month=="Oct")%>%  
 ggplot(aes(dep\_time)) +  
 geom\_freqpoly(binwidth=86400)



1. Now use dplyr functions to produce a data table that shows arrival times for American Airlines at the Dallas Fort Worth Airport from the LaGuardia airport in New York. Your output should show rows 115 to 125. A partial table is provided below.

flights %>%   
 filter(!is.na(arr\_time)) %>%   
 mutate(arr\_time = make\_datetime\_100(year, month, day, arr\_time)) %>%   
 filter(origin=="LGA")%>%  
 filter(dest=="DFW")%>%  
 select(origin, dest, carrier, arr\_time)%>%  
 slice(115:125)-> flights\_dt3  
flights\_dt3

## # A tibble: 11 × 4  
## origin dest carrier arr\_time   
## <chr> <chr> <chr> <dttm>   
## 1 LGA DFW AA 2013-01-09 16:16:00  
## 2 LGA DFW AA 2013-01-09 19:17:00  
## 3 LGA DFW AA 2013-01-09 19:36:00  
## 4 LGA DFW AA 2013-01-09 20:53:00  
## 5 LGA DFW AA 2013-01-09 22:24:00  
## 6 LGA DFW AA 2013-01-10 08:37:00  
## 7 LGA DFW AA 2013-01-10 10:20:00  
## 8 LGA DFW AA 2013-01-10 11:22:00  
## 9 LGA DFW AA 2013-01-10 12:16:00  
## 10 LGA DFW AA 2013-01-10 13:19:00  
## 11 LGA DFW AA 2013-01-10 13:23:00

1. Using the first two observational date time designations from your #8 table, Use and show R code to confirm that there are 181 minutes time intervals between them.

(("2013-01-09 16:16:00") %--% ("2013-01-09 19:17:00"))%/%minutes(1)

## [1] 181