Data Wrangling - 3.3

Hamed

2/23/2020

## (a). Load the datasets/tables flights and airlines

# (a). Install and load the library nycflights13  
library(tidyverse)

## -- Attaching packages ------------------------------------------------- tidyverse 1.3.0 --

## <U+2713> ggplot2 3.2.1 <U+2713> purrr 0.3.3  
## <U+2713> tibble 2.1.3 <U+2713> dplyr 0.8.3  
## <U+2713> tidyr 1.0.0 <U+2713> stringr 1.4.0  
## <U+2713> readr 1.3.1 <U+2713> forcats 0.4.0

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

library(dplyr)  
library(nycflights13)

# (b). Load the datasets/tables flights and airlines

airlines=data.frame(airlines)  
flights=data.frame(flights)  
airports=data.frame(airports)  
planes=data.frame(planes)  
weather=data.frame(weather)  
  
head(airlines)

## carrier name  
## 1 9E Endeavor Air Inc.  
## 2 AA American Airlines Inc.  
## 3 AS Alaska Airlines Inc.  
## 4 B6 JetBlue Airways  
## 5 DL Delta Air Lines Inc.  
## 6 EV ExpressJet Airlines Inc.

head(flights)

## year month day dep\_time sched\_dep\_time dep\_delay arr\_time sched\_arr\_time  
## 1 2013 1 1 517 515 2 830 819  
## 2 2013 1 1 533 529 4 850 830  
## 3 2013 1 1 542 540 2 923 850  
## 4 2013 1 1 544 545 -1 1004 1022  
## 5 2013 1 1 554 600 -6 812 837  
## 6 2013 1 1 554 558 -4 740 728  
## arr\_delay carrier flight tailnum origin dest air\_time distance hour minute  
## 1 11 UA 1545 N14228 EWR IAH 227 1400 5 15  
## 2 20 UA 1714 N24211 LGA IAH 227 1416 5 29  
## 3 33 AA 1141 N619AA JFK MIA 160 1089 5 40  
## 4 -18 B6 725 N804JB JFK BQN 183 1576 5 45  
## 5 -25 DL 461 N668DN LGA ATL 116 762 6 0  
## 6 12 UA 1696 N39463 EWR ORD 150 719 5 58  
## time\_hour  
## 1 2013-01-01 05:00:00  
## 2 2013-01-01 05:00:00  
## 3 2013-01-01 05:00:00  
## 4 2013-01-01 05:00:00  
## 5 2013-01-01 06:00:00  
## 6 2013-01-01 05:00:00

head(airports)

## faa name lat lon alt tz dst  
## 1 04G Lansdowne Airport 41.13047 -80.61958 1044 -5 A  
## 2 06A Moton Field Municipal Airport 32.46057 -85.68003 264 -6 A  
## 3 06C Schaumburg Regional 41.98934 -88.10124 801 -6 A  
## 4 06N Randall Airport 41.43191 -74.39156 523 -5 A  
## 5 09J Jekyll Island Airport 31.07447 -81.42778 11 -5 A  
## 6 0A9 Elizabethton Municipal Airport 36.37122 -82.17342 1593 -5 A  
## tzone  
## 1 America/New\_York  
## 2 America/Chicago  
## 3 America/Chicago  
## 4 America/New\_York  
## 5 America/New\_York  
## 6 America/New\_York

head(planes)

## tailnum year type manufacturer model engines seats  
## 1 N10156 2004 Fixed wing multi engine EMBRAER EMB-145XR 2 55  
## 2 N102UW 1998 Fixed wing multi engine AIRBUS INDUSTRIE A320-214 2 182  
## 3 N103US 1999 Fixed wing multi engine AIRBUS INDUSTRIE A320-214 2 182  
## 4 N104UW 1999 Fixed wing multi engine AIRBUS INDUSTRIE A320-214 2 182  
## 5 N10575 2002 Fixed wing multi engine EMBRAER EMB-145LR 2 55  
## 6 N105UW 1999 Fixed wing multi engine AIRBUS INDUSTRIE A320-214 2 182  
## speed engine  
## 1 NA Turbo-fan  
## 2 NA Turbo-fan  
## 3 NA Turbo-fan  
## 4 NA Turbo-fan  
## 5 NA Turbo-fan  
## 6 NA Turbo-fan

head(weather)

## origin year month day hour temp dewp humid wind\_dir wind\_speed wind\_gust  
## 1 EWR 2013 1 1 1 39.02 26.06 59.37 270 10.35702 NA  
## 2 EWR 2013 1 1 2 39.02 26.96 61.63 250 8.05546 NA  
## 3 EWR 2013 1 1 3 39.02 28.04 64.43 240 11.50780 NA  
## 4 EWR 2013 1 1 4 39.92 28.04 62.21 250 12.65858 NA  
## 5 EWR 2013 1 1 5 39.02 28.04 64.43 260 12.65858 NA  
## 6 EWR 2013 1 1 6 37.94 28.04 67.21 240 11.50780 NA  
## precip pressure visib time\_hour  
## 1 0 1012.0 10 2013-01-01 01:00:00  
## 2 0 1012.3 10 2013-01-01 02:00:00  
## 3 0 1012.5 10 2013-01-01 03:00:00  
## 4 0 1012.2 10 2013-01-01 04:00:00  
## 5 0 1011.9 10 2013-01-01 05:00:00  
## 6 0 1012.4 10 2013-01-01 06:00:00

## (c). What is the primary key of the planes table

#Planes$tailnum this is the primary key as it uniquely identifies each plane in the planes table  
planes%>%count(tailnum)%>%filter(n>1)

## # A tibble: 0 x 2  
## # … with 2 variables: tailnum <chr>, n <int>

## (d). Add full airline name from the airlines table to the flights table that keeps all the records in the flights table by using the appropriate join

flights\_with\_names<-left\_join(x=flights,y=airlines,by='carrier',all.x=TRUE)  
  
head(flights\_with\_names)

## year month day dep\_time sched\_dep\_time dep\_delay arr\_time sched\_arr\_time  
## 1 2013 1 1 517 515 2 830 819  
## 2 2013 1 1 533 529 4 850 830  
## 3 2013 1 1 542 540 2 923 850  
## 4 2013 1 1 544 545 -1 1004 1022  
## 5 2013 1 1 554 600 -6 812 837  
## 6 2013 1 1 554 558 -4 740 728  
## arr\_delay carrier flight tailnum origin dest air\_time distance hour minute  
## 1 11 UA 1545 N14228 EWR IAH 227 1400 5 15  
## 2 20 UA 1714 N24211 LGA IAH 227 1416 5 29  
## 3 33 AA 1141 N619AA JFK MIA 160 1089 5 40  
## 4 -18 B6 725 N804JB JFK BQN 183 1576 5 45  
## 5 -25 DL 461 N668DN LGA ATL 116 762 6 0  
## 6 12 UA 1696 N39463 EWR ORD 150 719 5 58  
## time\_hour name  
## 1 2013-01-01 05:00:00 United Air Lines Inc.  
## 2 2013-01-01 05:00:00 United Air Lines Inc.  
## 3 2013-01-01 05:00:00 American Airlines Inc.  
## 4 2013-01-01 05:00:00 JetBlue Airways  
## 5 2013-01-01 06:00:00 Delta Air Lines Inc.  
## 6 2013-01-01 05:00:00 United Air Lines Inc.

## (e). Now add the destination latitude and longitude to the flights table from the airports table by using the appropriate join

# Rename "faa" to "dest" in order to use it as joining key  
names(airports)[names(airports)=="faa"] <- "dest"  
  
flights\_dest<-left\_join(x=flights,y=airports,by='dest',all.x=TRUE)[c("year","month",  
 "day","dep\_time","sched\_dep\_time","dep\_delay","arr\_time",  
 "sched\_arr\_time","arr\_delay","carrier","flight","tailnum","origin",  
 "dest","air\_time" ,"distance","hour","minute","time\_hour","lat","lon")]  
  
head(flights\_dest)

## year month day dep\_time sched\_dep\_time dep\_delay arr\_time sched\_arr\_time  
## 1 2013 1 1 517 515 2 830 819  
## 2 2013 1 1 533 529 4 850 830  
## 3 2013 1 1 542 540 2 923 850  
## 4 2013 1 1 544 545 -1 1004 1022  
## 5 2013 1 1 554 600 -6 812 837  
## 6 2013 1 1 554 558 -4 740 728  
## arr\_delay carrier flight tailnum origin dest air\_time distance hour minute  
## 1 11 UA 1545 N14228 EWR IAH 227 1400 5 15  
## 2 20 UA 1714 N24211 LGA IAH 227 1416 5 29  
## 3 33 AA 1141 N619AA JFK MIA 160 1089 5 40  
## 4 -18 B6 725 N804JB JFK BQN 183 1576 5 45  
## 5 -25 DL 461 N668DN LGA ATL 116 762 6 0  
## 6 12 UA 1696 N39463 EWR ORD 150 719 5 58  
## time\_hour lat lon  
## 1 2013-01-01 05:00:00 29.98443 -95.34144  
## 2 2013-01-01 05:00:00 29.98443 -95.34144  
## 3 2013-01-01 05:00:00 25.79325 -80.29056  
## 4 2013-01-01 05:00:00 NA NA  
## 5 2013-01-01 06:00:00 33.63672 -84.42807  
## 6 2013-01-01 05:00:00 41.97860 -87.90484