SYS 5581 Project - Extract, Transform, and Load Data - 2nd Attempt

Nick Coronato

Version of 2021-02-28 | Due 2021-03-01

Step 1: Identify a Time Series data set that you want to work with

For this project, I will be analyzing a set of airline passenger data from the Bureau of Transportation Statistics (BTS). The data set includes air travel data from September 2018 - August 2020 (24 months).

Step 2: Acquire the data from its source location, reproducibly.

For this project, my data is stored on my local machine. The file name is "2year_t100_data.csv"

Data was obtained from the BTS website: https://www.transtats.bts.gov/DL_SelectFields.asp?gnoyr_VQ=FIL&QO_fu146_anzr=Nv4%20Pn44vr45

Note: Ideally the data will be stored at and read from a Github repository.

```
t100_raw  <-  read_delim("2year_t100_data.csv",",",
                         col_types = cols(.default = col_character(),
                                          "YEAR" = col_integer(),
                                          "MONTH" = col_integer(),
                                          "UNIQUE_CARRIER" = col_factor(),
                                          "PASSENGERS" = col_integer(),
                                          "DISTANCE" = col_integer(),
                                          "ORIGIN_AIRPORT_ID" = col_factor(),
                                          "ORIGIN_STATE_NM" = col_factor(),
                                          "DEST_AIRPORT_ID" = col_factor(),
                                          "DEST_STATE_NM" = col_factor(),
                                          "SEATS" = col_integer(),
                                          "AIR_TIME" = col_integer()
                                                  ))
```

Step 3: Organize your data into a *tidy* data frame.

Organize by taking out the non-useful variables. Make another new variable called RPM (revenue passenger miles), which is simply Passengers X Miles for each observation. Put the Index variable in the first column (eventually it will be 4dig year, 2dig month). Put the Key variables next (unique_carrier, Passengers, Distance, RPM). Then all the other (possibly) relevant variables.

```
t100_raw %>%
  mutate(., RPM = PASSENGERS * DISTANCE) %>%
  select(YEAR, MONTH, UNIQUE_CARRIER, PASSENGERS, DISTANCE, RPM, -ORIGIN_AIRPORT_ID, -ORIGIN_CITY_NAME,
```

Make a new variable called YRMO that concatenates Year and Month into a YR-MO format. Make it a time series column.

```
# Put together YEAR and MONTH into a single column called YRMO
t100_raw %>%
  mutate(YRMO = paste(t100_raw$YEAR, t100_raw$MONTH)) ->t100_raw

# Convert YRMO to a time series object column (year month) and re-arrange
t100_raw %>%
  mutate(YRMO = yearmonth(YRMO))  %>%
  select(YRMO, everything(),-MONTH, -YEAR)-> t100_raw2

# Create summary values for Passengers, RPM, Distance, Seats, and Air_time that will prevent us from ha
t100_raw2 %>%
  group_by(YRMO, UNIQUE_CARRIER) %>%
  summarize(TotalPax=sum(PASSENGERS), TotalRPM = sum(RPM), TotalDistance = sum(DISTANCE), TotalSeats = s

# convert to tsibble
as_tsibble(t100_raw3, index = YRMO, key = UNIQUE_CARRIER) -> t100_ts_tbl
```

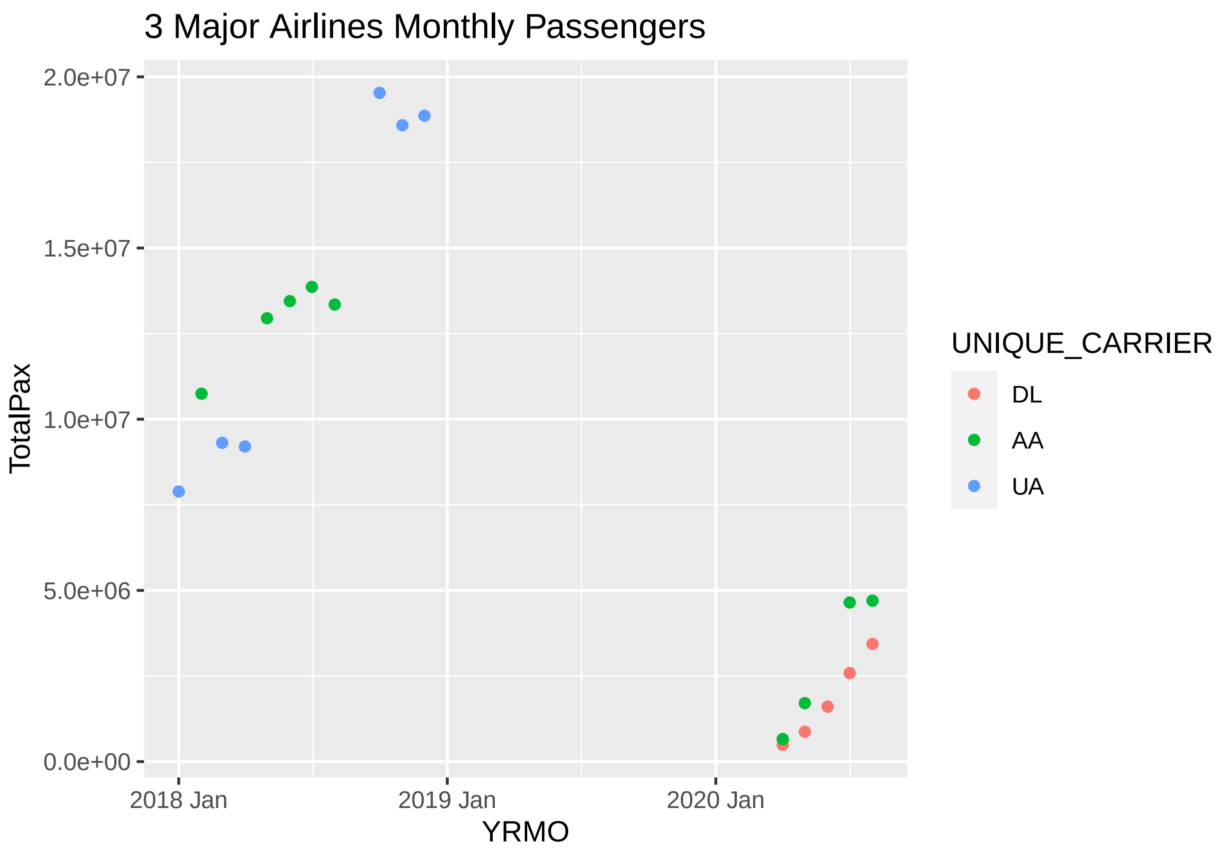
Some quality control checks.

```
#Check that each column got the right Type.

head(t100_ts_tbl)
```

```
## # A tsibble: 6 x 7 [1M]
## # Key:       UNIQUE_CARRIER [1]
## # Groups:    @ YRMO [6]
##       YRMO UNIQUE_CARRIER TotalPax TotalRPM TotalDistance TotalSeats
##      <mth> <fct>             <int>    <dbl>         <int>      <int>
## 1 2018 Jan WN             12380526  9.73e 9       2376696   16634106
## 2 2018 Feb WN             11548517  9.00e 9       2269742   14755049
## 3 2018 Mar WN             14702736  1.17e10       2440579   17740663
## 4 2018 Apr WN             14164211  1.12e10       2545974   17539196
## 5 2018 May WN             14838662  1.17e10       2444958   17936220
## 6 2018 Jun WN             15110630  1.22e10       3414681   17835753
## # ... with 1 more variable: TotalAIR_TIME <int>
```

This chunk is for example purposes; an unrefined ggplot of passnegers over time.



Generate and print the tsibble.

```
## # A tsibble: 2,206 x 7 [1M]
## # Key:       UNIQUE_CARRIER [129]
## # Groups:    @ YRMO [20]
##        YRMO UNIQUE_CARRIER TotalPax TotalRPM TotalDistance TotalSeats
##       <mth> <fct>             <int>    <dbl>         <int>      <int>
##  1 2018 Jan WN             12380526  9.73e 9       2376696   16634106
##  2 2018 Feb WN             11548517  9.00e 9       2269742   14755049
##  3 2018 Mar WN             14702736  1.17e10       2440579   17740663
##  4 2018 Apr WN             14164211  1.12e10       2545974   17539196
##  5 2018 May WN             14838662  1.17e10       2444958   17936220
##  6 2018 Jun WN             15110630  1.22e10       3414681   17835753
##  7 2018 Jul WN             15348967  1.26e10       3325204   18365951
##  8 2018 Aug WN             14350577  1.14e10       3449196   17587240
##  9 2018 Sep WN             25606502  2.00e10       6142526   32776074
## 10 2018 Oct WN             28796640  2.27e10       6358104   35240914
## # ... with 2,196 more rows, and 1 more variable: TotalAIR_TIME <int>
```