Project Proposal

Annelise Koster

12/3/2020

My blog can be found at https://akoster-dasc-1104-blog.netlify.app/

```
library(here)
library(ggplot2)
library(tidyverse)
library(readxl)
knitr::opts_chunk$set(echo = FALSE, tidy = TRUE)
```

My first data set is hotel bookings

```
## 'data.frame':    119390 obs. of  32 variables:
##  $ hotel                         : Factor w/ 2 levels "City Hotel","Resort Hotel": 2 2 2 2 2 2 2 2 2
##  $ is_canceled                   : int  0 0 0 0 0 0 0 0 1 1 ...
##  $ lead_time                     : int  342 737 7 13 14 14 0 9 85 75 ...
##  $ arrival_date_year             : int  2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...
##  $ arrival_date_month            : Factor w/ 12 levels "April","August",..: 6 6 6 6 6 6 6 6 6 6 ...
##  $ arrival_date_week_number      : int  27 27 27 27 27 27 27 27 27 27 ...
##  $ arrival_date_day_of_month     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ stays_in_weekend_nights       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ stays_in_week_nights          : int  0 0 1 1 2 2 2 2 3 3 ...
##  $ adults                        : int  2 2 1 1 2 2 2 2 2 2 ...
##  $ children                      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ babies                        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ meal                          : Factor w/ 5 levels "BB","FB","HB",..: 1 1 1 1 1 1 1 2 1 3 ...
##  $ country                       : Factor w/ 178 levels "ABW","AGO","AIA",..: 137 137 60 60 60 60 137
##  $ market_segment                : Factor w/ 8 levels "Aviation","Complementary",..: 4 4 4 3 7 7 4 4
##  $ distribution_channel          : Factor w/ 5 levels "Corporate","Direct",..: 2 2 2 1 4 4 2 2 4 4 .
##  $ is_repeated_guest             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ previous_cancellations        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ previous_bookings_not_canceled: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ reserved_room_type            : Factor w/ 10 levels "A","B","C","D",..: 3 3 1 1 1 1 3 3 1 4 ...
##  $ assigned_room_type            : Factor w/ 12 levels "A","B","C","D",..: 3 3 3 1 1 1 3 3 1 4 ...
##  $ booking_changes               : int  3 4 0 0 0 0 0 0 0 0 ...
##  $ deposit_type                  : Factor w/ 3 levels "No Deposit","Non Refund",..: 1 1 1 1 1 1 1 1
##  $ agent                         : Factor w/ 334 levels "1","10","103",..: 334 334 334 157 103 103 33
##  $ company                       : Factor w/ 353 levels "10","100","101",..: 353 353 353 353 353 353
##  $ days_in_waiting_list          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ customer_type                 : Factor w/ 4 levels "Contract","Group",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ adr                           : num  0 0 75 75 98 ...
##  $ required_car_parking_spaces   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ total_of_special_requests     : int  0 0 0 0 1 1 0 1 1 0 ...
```

```
##  $ reservation_status            : Factor w/ 3 levels "Canceled","Check-Out",..: 2 2 2 2 2 2 2 2 1 1
##  $ reservation_status_date       : Factor w/ 926 levels "1/1/2015","1/1/2016",..: 669 669 702 702 735
```

For this project I am examining hotel booking data contained in the Hotel booking demand dataset from Kaggle.com. This dataset contains booking information from two hotels including number of guests, length of stay and how the booking was made. The data consists of 119,390 observations of 32 variables. The variables include the type of hotel, the arrival date, the number of guests, the guest's nationality, and how many special requests the customer had. Based off my observations and assumptions I believe this data to be from a specific geographic area that draws numerous international visitors specifically groups of two adults.

- Question 1: Determine if there is a difference in lead times between the two types of hotels.
  - Will most likely do this by generating different data visualizations and calculating summary statistics like means, medians, and standard deviations.
- Question 2: Determine if there is a change in lead times over the three years of data.
  - Will most likely do this by generating different data visualizations and calculating summary statistics like means, medians, and standard deviations.
- Question 3: Do customers from different countries have different booking behavior? (do some have greater lead times than others?)
  - Will most likely do this by generating different data visualizations and calculating summary statistics like means, medians, and standard deviations.
  - Could also create a spatial map denoting the average lead time for customers from various locations and color the data points by lead time to denote the differences.
- Question 4: Determine which country's citizens require more special attention.
  - Develop a spatial map which shows the average number of special requests from customers from different countries.
- Question 5: Determine which month is the busiest for the hotels.
  - Will most likely do this by generating different data visualizations and calculating summary statistics like means, medians, and standard deviations.

I will probably only end up answering three or four of these questions.

My second data set is 911 emergency calls from Montgomery County

```
## 'data.frame':    663522 obs. of  9 variables:
##  $ lat      : num  40.3 40.3 40.1 40.1 40.3 ...
##  $ lng      : num  -75.6 -75.3 -75.4 -75.3 -75.6 ...
##  $ desc     : Factor w/ 663282 levels ".  ; AMBLER; 2018-04-05 @ 14:35:42;",..: 467158 67539 251193 1
##  $ zip      : int  19525 19446 19401 19401 NA 19446 19044 19426 19438 19462 ...
##  $ title    : Factor w/ 148 levels "EMS: ABDOMINAL PAINS",..: 10 23 104 18 25 42 50 60 69 146 ...
##  $ timeStamp: Factor w/ 543989 levels "1/1/2016 0:10",..: 134451 134454 134442 134447 134449 134443 1
##  $ twp      : Factor w/ 69 levels "","ABINGTON",..: 36 20 37 37 30 23 21 49 32 42 ...
##  $ addr     : Factor w/ 41292 levels ".","10TH AVE",..: 29611 3866 15602 948 6189 5155 19440 7168 211
##  $ e        : int  1 1 1 1 1 1 1 1 1 1 ...
```

For this project I am examining the 911 calls data contained in the Emergency – 911 Calls dataset from Kraggle.com. This dataset contains information about calls made to 911 in Montgomery County Pennsylvania. The data consists of 663,522 observations of 9 variables. The variable lat is a number denoting the latitude of the cal. The variable lng is a number denoting the longitude of the call. The desc variable is a

factor with 663,282 levels which represent the street name, town station number time and date of the call. The zip variable is a 5-digit integer representing the zip code of the location of the emergency. The variable title is a factor with 148 levels indicating the type of emergency. The timestamp variable is a factor with 543,989 levels indicating the time the call was placed. The twp variable is a factor with 69 levels indicating the town from which the call originated. The variable addr is a factor with 41,292 levels giving the street names of the location of the incident. The variable e is an integer which denotes the call was an emergency.

- Question 1: See if a specific type of emergency (EMS, fire, traffic) is more likely to occur at a specific time of day.
  - Not entirely sure how to explain how I want to do this but will most likely generate some data visualizations and summary statistics but I am not entirely sure at this moment.
- Question 2: Determine which area of Montgomery County had the most emergencies.
  - Create a spatial map which shows the number of emergencies across the county.
- Question 3: Determine which type of emergency is the most common.
  - Will most likely do this by generating different data visualizations and calculating summary statistics like means, medians, and standard deviations.
- Question 4: Determine which day of the week has the most traffic accidents.
  - Will most likely do this by generating different data visualizations and calculating summary statistics like means, medians, and standard deviations.