CSC/ST 442 (Fall 2021): Assignment 2

Instruction

This assignment consists of 3 problems. The assignment is due on **Thursday, September 16** at 11:59pm EDT. Please submit your assignment electronically through the moodle webpage. You are encouraged (but not required) to use RMarkdown to write up your homework solution. To start using Rmarkdown read

- Section 40.2 of Introduction to Data Science
- the RStudio tutorial
- the Rmarkdown cheatsheet.

Problem 1 (30pts)

This problem uses the *flights* and *weather* dataset from the **nycflights13** library. A snippet of the data is as follows.

```
library(nycflights13)
flights
```

```
## # A tibble: 336,776 x 19
##    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
## 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
## # ... with 336,770 more rows, and 11 more variables: arr_delay <dbl>,
## #   carrier <chr>, flight <int>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>
```

```
weather
```

```
## # A tibble: 26,115 x 15
##   origin  year month   day  hour  temp  dewp humid wind_dir wind_speed wind_gust
##   <chr>  <int> <int> <int> <int> <dbl> <dbl> <dbl>    <dbl>      <dbl>     <dbl>
## 1 EWR     2013     1     1     1  39.0  26.1  59.4      270      10.4         NA
## 2 EWR     2013     1     1     2  39.0  27.0  61.6      250       8.06        NA
## 3 EWR     2013     1     1     3  39.0  28.0  64.4      240      11.5         NA
## 4 EWR     2013     1     1     4  39.9  28.0  62.2      250      12.7         NA
## 5 EWR     2013     1     1     5  39.0  28.0  64.4      260      12.7         NA
## 6 EWR     2013     1     1     6  37.9  28.0  67.2      240      11.5         NA
## # ... with 26,109 more rows, and 4 more variables: precip <dbl>,
## #   pressure <dbl>, visib <dbl>, time_hour <dttm>
```

Using these two datasets, answer the following question.

1. (5pts) Visualize the departure times of cancelled versus non-cancelled flights.
2. (5pts) Draw a boxplot of the temperatures, grouped by months.

3. (10pts) Visualize the proportion of cancelled flights each day against the average daily temperature, grouped by the airports of origin (EWR, JFK, LGA).

Hint: you need to create a new column that maps the month and day to a number corresonding to the number of days since January 1. Try this

```
as.numeric(as.Date(paste(2013,2,24,sep="/"),format="%Y/%m/%d") -
             as.Date("2013/01/01",format="%Y/%m/%d"))
```

```
## [1] 54
```

A more elegant handling of dates and time is available via the lubridate library, but the above is sufficient for our current purpose.

4. (10pts) Visualize the arrival delay for flights to the top twenty most popular destinations, grouped by destinations.

Problem 2 (20pts)

This problem uses the **us_contagious_diseases** dataset from the dslabs library. This library accompanies the book Introduction to Data Science by R. Irizarry. A snippet of the data is as follows.

```
library(dslabs)
data(us_contagious_diseases)
us_contagious_diseases
```

```
## # A tibble: 16,065 x 6
##   disease     state    year weeks_reporting count population
##   <fct>       <fct>   <dbl>           <dbl> <dbl>      <dbl>
## 1 Hepatitis A Alabama  1966              50   321    3345787
## 2 Hepatitis A Alabama  1967              49   291    3364130
## 3 Hepatitis A Alabama  1968              52   314    3386068
## 4 Hepatitis A Alabama  1969              49   380    3412450
## 5 Hepatitis A Alabama  1970              51   413    3444165
## 6 Hepatitis A Alabama  1971              51   378    3481798
## # ... with 16,059 more rows
```

Each row of the data frame `us_contagious_diseases` record the **yearly** total number of cases of a specific disease (in each of the 50 states) during the period 1928 to 2011. For example the first row say that there was 321 reported cases of Hepatitis A in Alabama during the year 1966.

Using this data, answer the four questions in Section 10.15 of the book Introduction to Data Science. You might want to read through the code examples in Section 10.14 of that book.

Problem 3 (25pts)

This problem uses the obesity data set from the CDC. The original data is in Excel format and we had uploaded a copy of this file onto Moodle. First download the data onto your desktop/laptop work space and extract the data as follows.

```
library(readxl) ## install.packages("readxl") if the readxl library is missing.
fname <- "obesity_data.xlsx"
wrkbk <- read_excel(fname)
obesity_2012 <- setNames(wrkbk[-1, c(2, 61)], c("fips", "pct"))
obesity_2012$pct <- as.numeric(obesity_2012$pct) / 100
obesity_2012
```

```
## # A tibble: 3,224 x 2
##   fips    pct
```

```
##   <chr> <dbl>
## 1 01001 0.309
## 2 01003 0.267
## 3 01005 0.408
## 4 01007 0.401
## 5 01009 0.324
## 6 01011 0.445
## # ... with 3,218 more rows
```

We next load the socviz library and get access to the boundaries line for the US map.

```
library(socviz) ## install.packages("socviz") if the socviz library is missing.
county_map
```

```
## # A tibble: 191,382 x 7
##       long       lat order hole  piece group            id
##      <dbl>     <dbl> <int> <lgl> <fct> <fct>            <chr>
## 1 1225889. -1275020.     1 FALSE 1     0500000US01001.1 01001
## 2 1235324. -1274008.     2 FALSE 1     0500000US01001.1 01001
## 3 1244873. -1272331.     3 FALSE 1     0500000US01001.1 01001
## 4 1244129. -1267515.     4 FALSE 1     0500000US01001.1 01001
## 5 1272010. -1262889.     5 FALSE 1     0500000US01001.1 01001
## 6 1276797. -1295514.     6 FALSE 1     0500000US01001.1 01001
## # ... with 191,376 more rows
```

```
county_data
```

```
## # A tibble: 3,195 x 32
##   id    name   state census_region pop_dens pop_dens4 pop_dens6 pct_black    pop
##   <chr> <chr>  <fct> <fct>         <fct>    <fct>     <fct>     <fct>      <int>
## 1 0     <NA>   <NA>  <NA>          [   50,~ [ 45,  1~ [ 82,  2~ [10.0,15~ 3.19e8
## 2 01000 1      AL    South         [   50,~ [ 45,  1~ [ 82,  2~ [25.0,50~ 4.85e6
## 3 01001 Autau~ AL    South         [   50,~ [ 45,  1~ [ 82,  2~ [15.0,25~ 5.54e4
## 4 01003 Baldw~ AL    South         [  100,~ [118,716~ [ 82,  2~ [ 5.0,10~ 2.00e5
## 5 01005 Barbo~ AL    South         [   10,~ [ 17,   ~ [ 25,   ~ [25.0,50~ 2.69e4
## 6 01007 Bibb ~ AL    South         [   10,~ [ 17,   ~ [ 25,   ~ [15.0,25~ 2.25e4
## # ... with 3,189 more rows, and 23 more variables: female <dbl>, white <dbl>,
## #   black <dbl>, travel_time <dbl>, land_area <dbl>, hh_income <int>,
## #   su_gun4 <fct>, su_gun6 <fct>, fips <dbl>, votes_dem_2016 <int>,
## #   votes_gop_2016 <int>, total_votes_2016 <int>, per_dem_2016 <dbl>,
## #   per_gop_2016 <dbl>, diff_2016 <int>, per_dem_2012 <dbl>,
## #   per_gop_2012 <dbl>, diff_2012 <int>, winner <chr>, partywinner16 <chr>,
## #   winner12 <chr>, partywinner12 <chr>, flipped <chr>
```

Combining the `obesity_2012` data frame and either the `county_map` or `county_data` data, generate a visualization of the US Obesity Rate by County. Using the `county_data`, find the variables that are "correlated" with obesity rates.