# Problem Set #04 Solutions Statistics 506

Problem Set #04

### Problem 1 Solutions - Tidyverse

```
library(tidyverse)
library(nycflights13)
```

a.

```
# Departure
flights %>%
  group_by(origin) %>%
  summarize(mean_delay = mean(dep_delay, na.rm = TRUE),
            med_delay = median(dep_delay, na.rm = TRUE),
            numflights = n()) %>%
  ungroup() %>%
  filter(numflights > 10) %>%
  rename(faa = origin) %>%
  left_join(airports, by = "faa") %>%
  select(name, mean_delay, med_delay) %>%
  arrange(desc(mean_delay))
```

```
# A tibble: 3 x 3
  name                mean_delay med_delay
  <chr>                    <dbl>     <dbl>
1 Newark Liberty Intl       15.1        -1
2 John F Kennedy Intl       12.1        -1
3 La Guardia                10.3        -3
```

```r
# Arrival
flights %>%
  group_by(dest) %>%
  summarize(mean_delay = mean(arr_delay, na.rm = TRUE),
            med_delay = median(arr_delay, na.rm = TRUE),
            numflights = n()) %>%
  ungroup() %>%
  filter(numflights > 10) %>%
  rename(faa = dest) %>%
  left_join(airports, by = "faa") %>%
  mutate(name = coalesce(name, faa)) %>%
  select(name, mean_delay, med_delay) %>%
  arrange(desc(mean_delay)) %>%
  print(n = count(.))
```

```
# A tibble: 101 x 3
    name                                  mean_delay med_delay
    <chr>                                      <dbl>     <dbl>
  1 "Columbia Metropolitan"                     41.8        28
  2 "Tulsa Intl"                                33.7        14
  3 "Will Rogers World"                         30.6        16
  4 "Jackson Hole Airport"                      28.1        15
  5 "Mc Ghee Tyson"                             24.1         2
  6 "Dane Co Rgnl Truax Fld"                    20.2         1
  7 "Richmond Intl"                             20.1         1
  8 "Akron Canton Regional Airport"             19.7         3
  9 "Des Moines Intl"                           19.0         0
 10 "Gerald R Ford Intl"                        18.2         1
 11 "Birmingham Intl"                           16.9        -2
 12 "Theodore Francis Green State"              16.2         1
 13 "Greenville-Spartanburg International"      15.9        -0.5
 14 "Cincinnati Northern Kentucky Intl"         15.4        -3
 15 "Savannah Hilton Head Intl"                 15.1        -1
 16 "Manchester Regional Airport"               14.8        -3
 17 "Eppley Afld"                               14.7        -2
 18 "Yeager"                                    14.7        -1.5
 19 "Kansas City Intl"                          14.5         0
 20 "Albany Intl"                               14.4        -4
 21 "General Mitchell Intl"                     14.2         0
 22 "Piedmont Triad"                            14.1        -2
 23 "Washington Dulles Intl"                    13.9        -3
 24 "Cherry Capital Airport"                    13.0       -10
```

| | | | |
|---|---|---|---|
| 25 | "James M Cox Dayton Intl" | 12.7 | -3 |
| 26 | "Louisville International Airport" | 12.7 | -2 |
| 27 | "Chicago Midway Intl" | 12.4 | -1 |
| 28 | "Sacramento Intl" | 12.1 | 4 |
| 29 | "Jacksonville Intl" | 11.8 | -2 |
| 30 | "Nashville Intl" | 11.8 | -2 |
| 31 | "Portland Intl Jetport" | 11.7 | -4 |
| 32 | "Greater Rochester Intl" | 11.6 | -5 |
| 33 | "Hartsfield Jackson Atlanta Intl" | 11.3 | -1 |
| 34 | "Lambert St Louis Intl" | 11.1 | -3 |
| 35 | "Norfolk Intl" | 10.9 | -4 |
| 36 | "Baltimore Washington Intl" | 10.7 | -5 |
| 37 | "Memphis Intl" | 10.6 | -2.5 |
| 38 | "Port Columbus Intl" | 10.6 | -3 |
| 39 | "Charleston Afb Intl" | 10.6 | -4 |
| 40 | "Philadelphia Intl" | 10.1 | -3 |
| 41 | "Raleigh Durham Intl" | 10.1 | -3 |
| 42 | "Indianapolis Intl" | 9.94 | -3 |
| 43 | "Charlottesville-Albemarle" | 9.5 | -5 |
| 44 | "Cleveland Hopkins Intl" | 9.18 | -5 |
| 45 | "Ronald Reagan Washington Natl" | 9.07 | -2 |
| 46 | "Burlington Intl" | 8.95 | -4 |
| 47 | "Buffalo Niagara Intl" | 8.95 | -5 |
| 48 | "Syracuse Hancock Intl" | 8.90 | -5 |
| 49 | "Denver Intl" | 8.61 | -2 |
| 50 | "Palm Beach Intl" | 8.56 | -3 |
| 51 | "BQN" | 8.25 | -1 |
| 52 | "Bob Hope" | 8.18 | -3 |
| 53 | "Fort Lauderdale Hollywood Intl" | 8.08 | -3 |
| 54 | "Bangor Intl" | 8.03 | -9 |
| 55 | "Asheville Regional Airport" | 8.00 | -1 |
| 56 | "PSE" | 7.87 | 0 |
| 57 | "Pittsburgh Intl" | 7.68 | -5 |
| 58 | "Gallatin Field" | 7.6 | -2 |
| 59 | "NW Arkansas Regional" | 7.47 | -2 |
| 60 | "Tampa Intl" | 7.41 | -4 |
| 61 | "Charlotte Douglas Intl" | 7.36 | -3 |
| 62 | "Minneapolis St Paul Intl" | 7.27 | -5 |
| 63 | "William P Hobby" | 7.18 | -4 |
| 64 | "Bradley Intl" | 7.05 | -10 |
| 65 | "San Antonio Intl" | 6.95 | -9 |
| 66 | "Louis Armstrong New Orleans Intl" | 6.49 | -6 |
| 67 | "Key West Intl" | 6.35 | 7 |

```
 68 "Eagle Co Rgnl"                          6.30        -4
 69 "Austin Bergstrom Intl"                  6.02        -5
 70 "Chicago Ohare Intl"                     5.88        -8
 71 "Orlando Intl"                           5.45        -5
 72 "Detroit Metro Wayne Co"                 5.43        -7
 73 "Portland Intl"                          5.14        -5
 74 "Nantucket Mem"                          4.85        -3
 75 "Wilmington Intl"                        4.64        -7
 76 "Myrtle Beach Intl"                      4.60       -13
 77 "Albuquerque International Sunport"      4.38        -5.5
 78 "George Bush Intercontinental"           4.24        -5
 79 "Norman Y Mineta San Jose Intl"          3.45        -7
 80 "Southwest Florida Intl"                 3.24        -5
 81 "San Diego Intl"                         3.14        -5
 82 "Sarasota Bradenton Intl"                3.08        -5
 83 "Metropolitan Oakland Intl"              3.08        -9
 84 "General Edward Lawrence Logan Intl"     2.91        -9
 85 "San Francisco Intl"                     2.67        -8
 86 "SJU"                                    2.52        -6
 87 "Yampa Valley"                           2.14         2
 88 "Phoenix Sky Harbor Intl"                2.10        -6
 89 "Montrose Regional Airport"              1.79       -10.5
 90 "Los Angeles Intl"                       0.547       -7
 91 "Dallas Fort Worth Intl"                 0.322       -9
 92 "Miami Intl"                             0.299       -9
 93 "Mc Carran Intl"                         0.258       -8
 94 "Salt Lake City Intl"                    0.176       -8
 95 "Long Beach"                            -0.0620     -10
 96 "Martha\\\\'s Vineyard"                 -0.286      -11
 97 "Seattle Tacoma Intl"                   -1.10       -11
 98 "Honolulu Intl"                         -1.37        -7
 99 "STT"                                   -3.84        -9
100 "John Wayne Arpt Orange Co"             -7.87       -11
101 "Palm Springs Intl"                    -12.7        -13.5
```

b.

```r
flights %>%
  left_join(planes, by = "tailnum") %>%
  mutate(time = air_time/60,
         mph = distance/time) %>%
  group_by(model) %>%
```

```
  summarize(avgmph = mean(mph, na.rm = TRUE),
            nflights = n()) %>%
  arrange(desc(avgmph)) %>%
  slice(1)
```

```
# A tibble: 1 x 3
  model   avgmph nflights
  <chr>    <dbl>    <int>
1 777-222   483.        4
```

### Problem 2 Solutions - `get_temp()`

```
library(tidyverse)
```

```
##' Get average monthly temperature
##' @param month Numeric or string month
##' @param year Year
##' @param data Data set containing `month_numeric`, `year`, and `temp` columns
##' @param average_fn Function to compute average. Default is `mean`.
##' @param celsius Logical, default `FALSE` (return fahrenheit instead)
##' @return Average temperature
get_temp <- function(month, year, data, average_fn = mean, celsius = FALSE) {

  if (month %>% is.numeric) {
    # Ifmonth` is numeric, make sure it is valid
    if (month < 1 | month > 12) {
      stop("Invalid month")
    }
  } else if (month %>% is.character) {
    # If `month` is a string, use `match.arg` to handle abbrevations
    months <- c("January", "February", "March", "April", "May", "June", "July",
                "August", "September", "October", "November", "December")
    month %>%
      match.arg(months) %>%
      `==`(months) %>%
      which -> month
  } else {
    stop("Month must be numeric or character")
  }
```

```
  if (!year %>% is.numeric) {
    stop("year must be numeric")
  }
  if (year < 1997 | year > 2000) {
    stop("year out of range")
  }

  if (!(average_fn %>% is.function)) {
    stop("average_fn must be a function")
  }

  data %>%
    select(temp, month_numeric, year) %>%
    rename(year_col = year) %>% # Rename to avoid conflict between `year` and
                                # `data$year`
    filter(year_col == year,
           month_numeric == month) %>%
    summarize(avgtmp = average_fn(temp)) %>%
    mutate(avgtmp = ifelse(isTRUE(celsius), 5/9*(avgtmp - 32), avgtmp)) %>%
    as.numeric -> out # convert to numeric for result

  return(out)
}
```

```
nnmaps <- read_csv("data/chicago-nmmaps.csv")
```

```
get_temp("Apr", 1999, data = nnmaps)
```

```
[1] 49.8
```

```
get_temp("Apr", 1999, data = nnmaps, celsius = TRUE)
```

```
[1] 9.888889
```

```
get_temp(10, 1998, data = nnmaps, average_fn = median)
```

```
[1] 55
```

```r
get_temp(13, 1998, data = nnmaps)
```

```
Error in get_temp(13, 1998, data = nnmaps): Invalid month
```

```r
get_temp(2, 2005, data = nnmaps)
```

```
Error in get_temp(2, 2005, data = nnmaps): year out of range
```

```r
get_temp("November", 1999, data =nnmaps, celsius = TRUE,
         average_fn = function(x) {
           x %>% sort -> x
           x[2:(length(x) - 1)] %>% mean %>% return
         })
```

```
[1] 7.301587
```

### Problem 3 Solutions - SAS

Complete .sas script can be found here.

Results can be found here.

```
* Read in data;
DATA recs;
    SET "~/recs2020_public_v5.sas7bdat";
RUN;
```

a.

```
PROC FREQ DATA=recs ORDER=FREQ;
    WEIGHT nweight;
    TABLES state_name;
RUN;
```

From the results, we see that California has the highest proportion of records, and Michigan accounts for 3.17% of records.

b.

```
** Restrict data to non-zero values for visibility;
DATA recs2;
    SET recs;
    IF dollarfo > 0;
RUN;

PROC SGPLOT DATA=work.recs2;
   histogram dollarfo ;
run;
```

c.

```
** Take the log transformation;
DATA RECS3;
    SET RECS;
    ldollarfo = LOG(dollarfo);
RUN;

PROC SGPLOT DATA=recs3;
   HISTOGRAM ldollarfo;
RUN;
```

d.

```
** prkgplc1 contains -2 for missing values, drop;
DATA recs4;
    SET recs3;
    WHERE prkgplc1 >= 0;
RUN;

PROC GLM DATA=recs4;
    CLASS prkgplc1;
    MODEL ldollarfo=totrooms prkgplc1 / SOLUTION;
    WEIGHT nweight;
    OUTPUT OUT=predresults PREDICTED=pred;
RUN;
```

e.

```
** Exponentiate predicted value;
DATA predresults2;
    SET predresults;
```

```
    predexp = EXP(pred);
RUN;

PROC SGPLOT DATA=predresults2;
    SCATTER X=dollarfo Y=predexp;
RUN;
```

### Problem 4 Solutions - Multiple tools

a.

The codebook was generated with Stata's `codebook` command.

---

### SAS

b.

```
LIBNAME home "~";

/* Read in data */
DATA work.shed;
    SET "~/public2022.sas7bdat";
RUN;

/* Extract relevant columns, and rename */
PROC SQL;
    CREATE TABLE work.shedsmall AS
        SELECT b3 AS financial, nd2 as naturaldisaster,
               b7_b AS economic_condition,
               gh1 AS home, educ_4cat, race_5cat
          FROM work.shed;
QUIT;

/* Convert to binary */
DATA work.shedsmall;
    set work.shedsmall;
    worseoff = financial le 2;
RUN;
```

c.

```
/* Export */
DATA home.shedsmall;
    SET work.shedsmall;
RUN;
```

---

### Stata

The complete Do-file can be found here.

d.

```
. import sas using "data/shedsmall.sas7bdat", clear
(7 vars, 11,667 obs)
. rename _all, lower
  (all newnames==oldnames)
. describe, short
Contains data
 Observations:        11,667
    Variables:             9
Sorted by:
     Note: Dataset has changed since last saved.
```

e.

```
. capture drop worseoff2
. generate worseoff2 = financial <= 2
. * Show that both SAS and Stata produce the same binary
. tabulate worseoff worseoff2

           |       worseoff2
  worseoff |         0          1 |     Total
-----------+----------------------+----------
         0 |     7,371          0 |     7,371
         1 |         0      4,296 |     4,296
-----------+----------------------+----------
     Total |     7,371      4,296 |    11,667
```

f.

```
. svyset caseid [pw=weight_pop]

Sampling weights: weight_pop
             VCE: linearized
     Single unit: missing
        Strata 1: <one>
 Sampling unit 1: caseid
           FPC 1: <zero>
.
. svy: logit worseoff naturaldisaster economic_condition i.home ///
>                    i.educ_4cat i.race_5cat, or
(running logit on estimation sample)

Survey: Logistic regression

Number of strata =      1                        Number of obs    =      11,667
Number of PSUs   = 11,667                        Population size = 255,114,223
                                                 Design df        =      11,666
                                                 F(12, 11655)     =       71.73
                                                 Prob > F         =      0.0000

------------------------------------------------------------------------------
                   |             Linearized
         worseoff | Odds ratio   std. err.      t    P>|t|     [95% conf. interval]
------------------+-----------------------------------------------------------------
  naturaldisaster |   .9649101   .0293842    -1.17   0.241     .9089975    1.024262
economic_condit~n |   .3796709   .0138959   -26.46   0.000     .3533866    .4079101
                  |
             home |
                2 |   1.067679   .0600423     1.16   0.244     .9562412    1.192104
                3 |   .9682945   .0564471    -0.55   0.580     .8637363     1.08551
                4 |   .7018847   .0692285    -3.59   0.000      .578497    .8515899
                  |
        educ_4cat |
                2 |   .8904145   .1032571    -1.00   0.317     .7093691    1.117666
                3 |   .8437971   .0936402    -1.53   0.126     .6788382    1.048841
                4 |   .7244437    .080169    -2.91   0.004     .5831743    .8999344
                  |
        race_5cat |
                2 |    .491115   .0396372    -8.81   0.000     .4192537    .5752937
                3 |   .8486486   .0605296    -2.30   0.021     .7379212    .9759911
```

```
               4  |   .6379569   .0801576    -3.58   0.000      .498688    .8161194
               5  |   1.018366    .166154     0.11   0.911     .7396213    1.402162
                  |
           _cons |   4.493002   .6866228     9.83   0.000     3.329984    6.062213
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

The p-value of .241 indicates that we see no statistically significant evidence of a relationship between whether a person thinks a natural disaster is coming and whether a person believes they will be worse-off.

g.

```
. save data/shedsmall, replace
file data/shedsmall.dta saved
```

---

**R**

h.

```
library(haven)
dat <- read_dta("data/shedsmall.dta")
names(dat) <- tolower(names(dat))

library(survey)
design <- svydesign(id = ~ caseid, weight = ~ weight_pop, data = dat)

mod <- svyglm(worseoff ~ naturaldisaster + economic_condition +
                as.factor(home) + as.factor(educ_4cat) +
                as.factor(race_5cat), data = dat, design = design,
              family = quasibinomial)

psrsq(mod)
```

```
[1] 0.1080233
```