Discussion_Wk2

Discussion Activity: Continue working on this data set to discern any trends in the bird/wildlife dataset during time period (1995- March 2022) at DIA. You can pick a different time frame and parameters if you are interested using the faa's query tool (https://wildlife.faa.gov/search).

Create a box plot showing a relationship between year and number of birds struck Create a plot with ggplot() showing the extent of damage or aircraft by bird collisions Use a subset of data where the incident occurred after 2015 and compute the means and SDs of distance by the number of birds struck. The function tapply() is useful for this. Create a final plot of bird collisions by engine type and distance (nautical miles from airport). You'll need to revalue engine type and do some class conversions to create a good-looking plot. Be sure to include labels and title.

Post to the discussion thread: 1. Any problems/concerns with data. 2. Plots that you think may be relevant to support your assertions. 3. Summary and interpretation of results.

Set working directory

```
#set working directory
setwd("C:\\Users\\adamg\\Documents\\MSDS_660\\Week_2")
```

Load Libraries

Load the data

```
#read data from csv
df<-read_csv("wildlife.csv")

#convert to data.table
df<-as.data.table(df)

#convert column names to lower case
names(df) <- tolower(names(df)) #convert column names to lower case
summary(df)
```

```
##     indx_nr        incident_date        incident_month   incident_year
##  Min.   : 608243   Min.   :1994-09-26   Min.   : 1.000   Min.   :1994
##  1st Qu.: 692887   1st Qu.:2008-03-31   1st Qu.: 6.000   1st Qu.:2008
##  Median : 742855   Median :2013-04-28   Median : 7.000   Median :2013
##  Mean   : 790259   Mean   :2012-06-24   Mean   : 6.862   Mean   :2012
##  3rd Qu.: 806821   3rd Qu.:2018-04-17   3rd Qu.: 9.000   3rd Qu.:2018
##  Max.   :1211589   Max.   :2022-03-01   Max.   :12.000   Max.   :2022
##
##      time           time_of_day         airport_id          airport
##  Length:8456        Length:8456        Length:8456        Length:8456
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
```

```
##
##
##     latitude       longitude         runway             state
##  Min.   :39.86   Min.   :-104.7   Length:8456        Length:8456
##  1st Qu.:39.86   1st Qu.:-104.7   Class :character   Class :character
##  Median :39.86   Median :-104.7   Mode  :character   Mode  :character
##  Mean   :39.86   Mean   :-104.7
##  3rd Qu.:39.86   3rd Qu.:-104.7
##  Max.   :39.86   Max.   :-104.7
##
##   faaregion           location         enroute_state      opid
##  Length:8456        Length:8456        Mode:logical   Length:8456
##  Class :character   Class :character   NA's:8456      Class :character
##  Mode  :character   Mode  :character                  Mode  :character
##
##
##
##
##    operator             reg                flt              aircraft
##  Length:8456        Length:8456        Length:8456        Length:8456
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##      ama                amo                ema                emo
##  Length:8456        Length:8456        Length:8456        Length:8456
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    ac_class            ac_mass        type_eng            num_engs
##  Length:8456        Min.   :1.000   Length:8456        Min.   :1.000
##  Class :character   1st Qu.:4.000   Class :character   1st Qu.:2.000
##  Mode  :character   Median :4.000   Mode  :character   Median :2.000
##                     Mean   :3.877                      Mean   :2.042
##                     3rd Qu.:4.000                      3rd Qu.:2.000
##                     Max.   :5.000                      Max.   :4.000
##                     NA's   :4193                       NA's   :4192
##    eng_1_pos       eng_2_pos       eng_3_pos       eng_4_pos
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1
##  Median :1.000   Median :1.000   Median :5.000   Median :1
##  Mean   :1.807   Mean   :1.871   Mean   :3.793   Mean   :1
##  3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:5.000   3rd Qu.:1
##  Max.   :7.000   Max.   :6.000   Max.   :5.000   Max.   :1
##  NA's   :4195    NA's   :4200    NA's   :8287    NA's   :8441
##  phase_of_flight        height           speed          distance
##  Length:8456        Min.   :    0.0   Min.   :  0.0   Min.   : 0.0000
##  Class :character   1st Qu.:    0.0   1st Qu.:120.0   1st Qu.: 0.0000
##  Mode  :character   Median :    0.0   Median :140.0   Median : 0.0000
```

```
##                    Mean   :  487.8   Mean   :142.9   Mean   : 0.1919
##                    3rd Qu.:   10.0   3rd Qu.:150.0   3rd Qu.: 0.0000
##                    Max.   :16500.0   Max.   :354.0   Max.   :50.0000
##                    NA's   :4952      NA's   :7044    NA's   :1058
##      sky           precipitation          aos          cost_repairs
##  Length:8456        Length:8456        Min.   :  0.10   Min.   :     200
##  Class :character   Class :character   1st Qu.:  1.00   1st Qu.:    9750
##  Mode  :character   Mode  :character   Median :  1.00   Median :   29750
##                                        Mean   : 12.25   Mean   :  787239
##                                        3rd Qu.:  4.00   3rd Qu.:  120000
##                                        Max.   :504.00   Max.   :14000000
##                                        NA's   :8303     NA's   :8415
##    cost_other     cost_repairs_infl_adj cost_other_infl_adj  ingested
##  Min.   :     9   Min.   :     241      Min.   :    10      Mode :logical
##  1st Qu.:   100   1st Qu.:   11154      1st Qu.:   118      FALSE:8241
##  Median :   300   Median :   34625      Median :   375      TRUE :215
##  Mean   : 12836   Mean   :  973487      Mean   : 15361
##  3rd Qu.:   525   3rd Qu.:  149160      3rd Qu.:   653
##  Max.   :560700   Max.   :17668000      Max.   :675083
##  NA's   :8400     NA's   :8415          NA's   :8400
##  indicated_damage damage_level        str_rad         dam_rad
##  Mode :logical    Length:8456        Mode :logical   Mode :logical
##  FALSE:8252       Class :character   FALSE:7887      FALSE:8423
##  TRUE :204        Mode  :character   TRUE :569       TRUE :33
##
##
##
##
##  str_windshld    dam_windshld     str_nose        dam_nose
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:7848      FALSE:8446      FALSE:7855      FALSE:8440
##  TRUE :608       TRUE :10        TRUE :601       TRUE :16
##
##
##
##
##   str_eng1        dam_eng1        ing_eng1        str_eng2
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:8236      FALSE:8417      FALSE:8451      FALSE:8255
##  TRUE :220       TRUE :39        TRUE :5         TRUE :201
##
##
##
##
##   dam_eng2        ing_eng2        str_eng3        dam_eng3
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:8422      FALSE:8449      FALSE:8454      FALSE:8455
##  TRUE :34        TRUE :7         TRUE :2         TRUE :1
##
##
##
##
##   ing_eng3        str_eng4        dam_eng4        ing_eng4
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
```

```
##  FALSE:8456      FALSE:8456      FALSE:8456      FALSE:8456
##
##
##
##
##
##   str_prop        dam_prop       str_wing_rot    dam_wing_rot
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:8434      FALSE:8456      FALSE:8008      FALSE:8406
##  TRUE :22                        TRUE :448       TRUE :50
##
##
##
##
##   str_fuse        dam_fuse         str_lg          dam_lg
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:7872      FALSE:8448      FALSE:8246      FALSE:8447
##  TRUE :584       TRUE :8         TRUE :210       TRUE :9
##
##
##
##
##   str_tail        dam_tail       str_lghts       dam_lghts
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:8429      FALSE:8447      FALSE:8441      FALSE:8450
##  TRUE :27        TRUE :9         TRUE :15        TRUE :6
##
##
##
##
##  str_other       dam_other       other_specify         effect
##  Mode :logical   Mode :logical   Length:8456        Length:8456
##  FALSE:7198      FALSE:8434      Class :character   Class :character
##  TRUE :1258      TRUE :22        Mode  :character   Mode  :character
##
##
##
##
##  effect_other        species_id          species            remarks
##  Length:8456        Length:8456        Length:8456        Length:8456
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  remains_collected remains_sent    bird_band_number    warned
##  Mode :logical     Mode :logical   Mode:logical     Length:8456
##  FALSE:2669        FALSE:7686      NA's:8456        Class :character
##  TRUE :5787        TRUE :770                        Mode  :character
##
##
##
##
```

```
##    num_seen          num_struck           size           nr_injuries
##  Length:8456        Length:8456        Length:8456        Mode:logical
##  Class :character   Class :character   Class :character   NA's:8456
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  nr_fatalities    comments         reporter_name      reporter_title
##  Mode:logical   Length:8456        Length:8456        Length:8456
##  NA's:8456      Class :character   Class :character   Class :character
##                 Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     source             person            lupdate            transfer
##  Length:8456        Length:8456        Min.   :1996-03-18   Mode :logical
##  Class :character   Class :character   1st Qu.:2008-11-20   FALSE:8456
##  Mode  :character   Mode  :character   Median :2013-09-26
##                                        Mean   :2013-03-27
##                                        3rd Qu.:2018-10-24
##                                        Max.   :2022-03-11
##
```

Including Plots

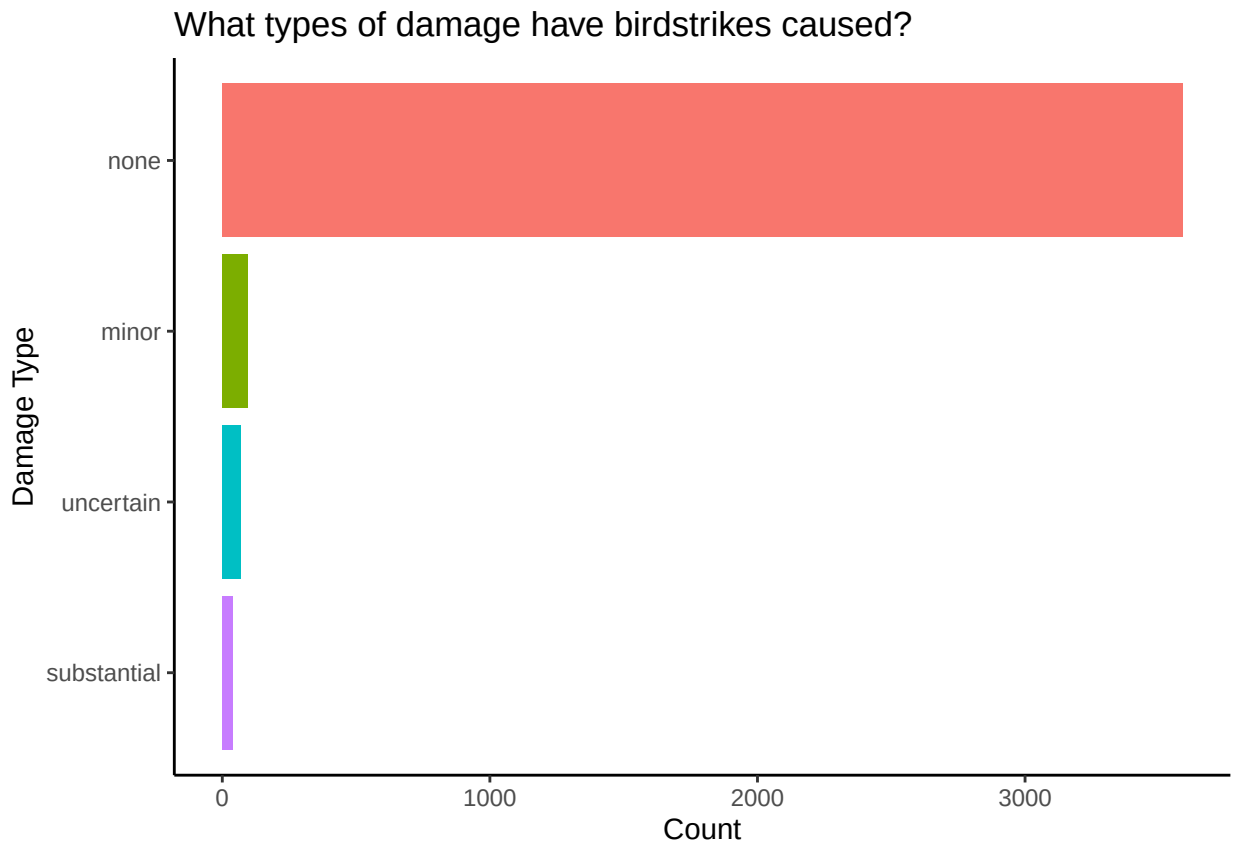
Create a boxplot that shows the number of birds struck by aircraft by year at DIA. You may need to reorder categories of number of birds struck by using factor.

```
#Convert character column to factor
df$num_struck<-factor(df$num_struck, levels = c("0","1","2-10","11-100","over 100"))

#set title font size
par(cex.main=1)

#generate boxplot
boxplot(incident_year~num_struck,
        main="Bird Strike Size by Year",
        xlab="Size of Bird Strike",
        ylab="Year",
        data=df)
```

Let's see if airplanes are reporting any damage associated with bird collisions. Load plyr library. To do this follow the prompts in the comments.

```
#subset data
damage_level_df <- subset(df, !is.na(damage_level))

#Convert character column to factor
damage_level_df$damage_level<-factor(damage_level_df$damage_level, levels = c("N","M","M?","S","D"))

#relabel damage level names for clarity with revalue from plyr
damage_level_df$damage_level<-revalue(damage_level_df$damage_level, c(N="none",
                                                                       M="minor",
                                                                      `M?` = "uncertain",
                                                                       S ="substantial",
                                                                       D ="destroyed"
                                                                       ))

#create a plot with ggplot and label the axis and give a title
ggplot(damage_level_df, aes(y = fct_infreq(damage_level), fill=damage_level)) +
  geom_bar() +
  scale_y_discrete(limits=rev)+
  labs(x = "Count",
       y = "Damage Type",
       title="What types of damage have birdstrikes caused?")+
  theme_bw() +
  theme(panel.border = element_blank(),
        panel.grid.major = element_blank(),
```

```
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "black"),
        legend.position="none")
```



Data Slicing We'll now slice and dice the data to look at more recent data.

```
#create a subset of our data and only look at bird collisions occurring in 2015 and later.
post_2015 = df[df$incident_year >= 2015,]

# run summary on year to check we filtered correctly
print("Summary of incident_year column")
```

```
## [1] "Summary of incident_year column"
```

```
summary(post_2015$incident_year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2015    2016    2019    2018    2020    2022
```

```
#use tapply to find means and sd of distance by number of birds struck on the new subset of data
print("Mean and SD of Distance Grouped by Birdstrike")
```

```
## [1] "Mean and SD of Distance Grouped by Birdstrike"
```

```
print("Mean")
```

```
## [1] "Mean"
```

```
tapply(post_2015$distance, post_2015$num_struck, mean, na.rm=TRUE)
```

```
##         0         1      2-10    11-100  over 100
##        NA 0.2319149 0.1377551 0.0000000        NA
```

```
print("Standard Deviation")
```

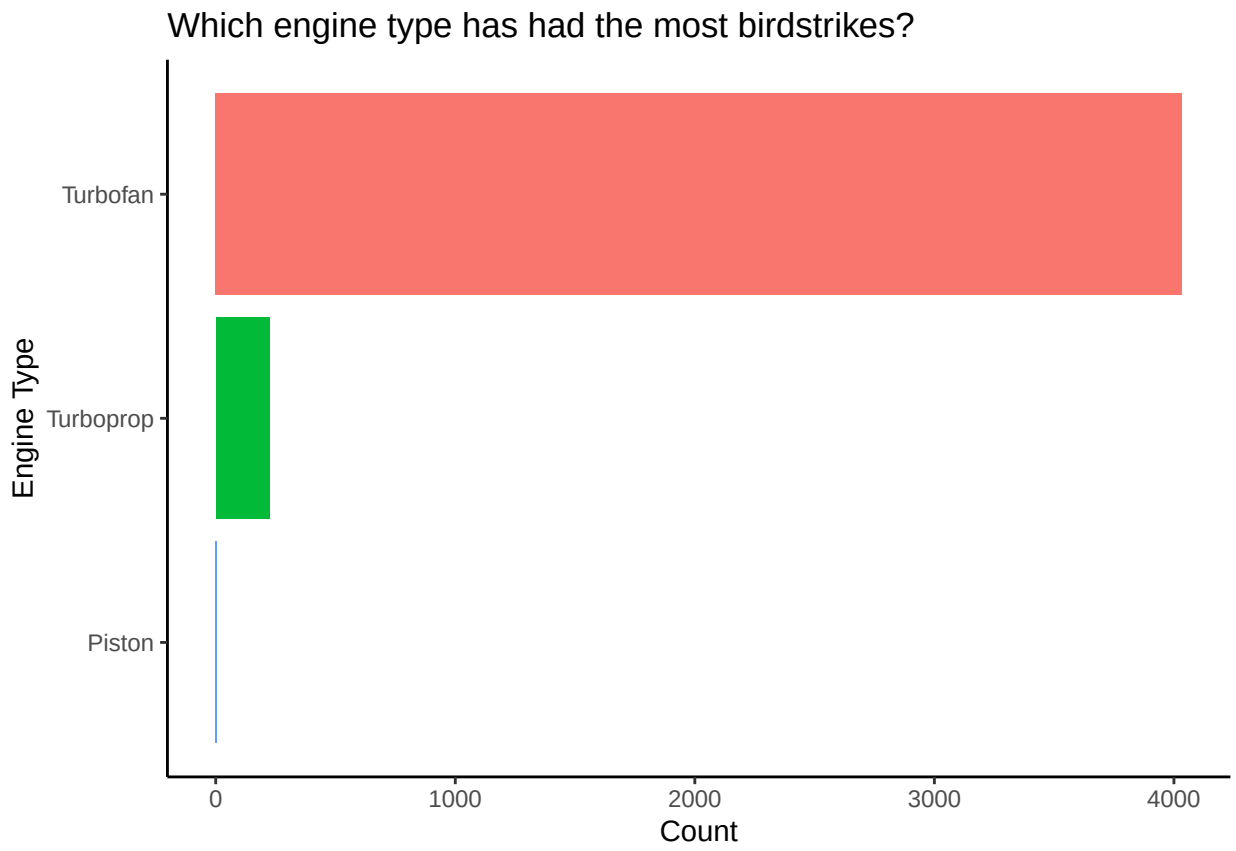
```
## [1] "Standard Deviation"
```

```
tapply(post_2015$distance, post_2015$num_struck, sd, na.rm=TRUE)
```

```
##        0        1     2-10   11-100 over 100
##       NA 1.780096 1.238859 0.000000       NA
```

Here is more practice slicing and dicing data and creating a good looking plot! Use the original dataset.

```
#subset data
type_eng_df <- subset(df, !is.na(type_eng))

#Convert character column to factor
type_eng_df$type_eng<-factor(type_eng_df$type_eng, levels = c("A","B","C","D","E","F","Y"))

#use revalue from plyr to give clear labels to type_eng. Use the data directory to find what the abbrev
type_eng_df$type_eng<-revalue(type_eng_df$type_eng, c("A"="Piston",
                                                      "B"="Turbojet",
                                                      "C"="Turboprop",
                                                      "D"="Turbofan",
                                                      "E"="Glider",
                                                      "F"="Helicopter",
                                                      "Y"="Other"))

#draw ggplot
ggplot(type_eng_df, aes(y = fct_infreq(type_eng), fill=fct_infreq(type_eng))) +
  geom_bar() +
  scale_y_discrete(limits=rev)+
  labs(x = "Count",
       y = "Engine Type",
       title="Which engine type has had the most birdstrikes?")+
  theme_bw() +
  theme(panel.border = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "black"),
        legend.position="none")
```




```
#convert num_struck to integer
df$num_struck<- as.integer(df$num_struck)

summary(df$num_struck)
```

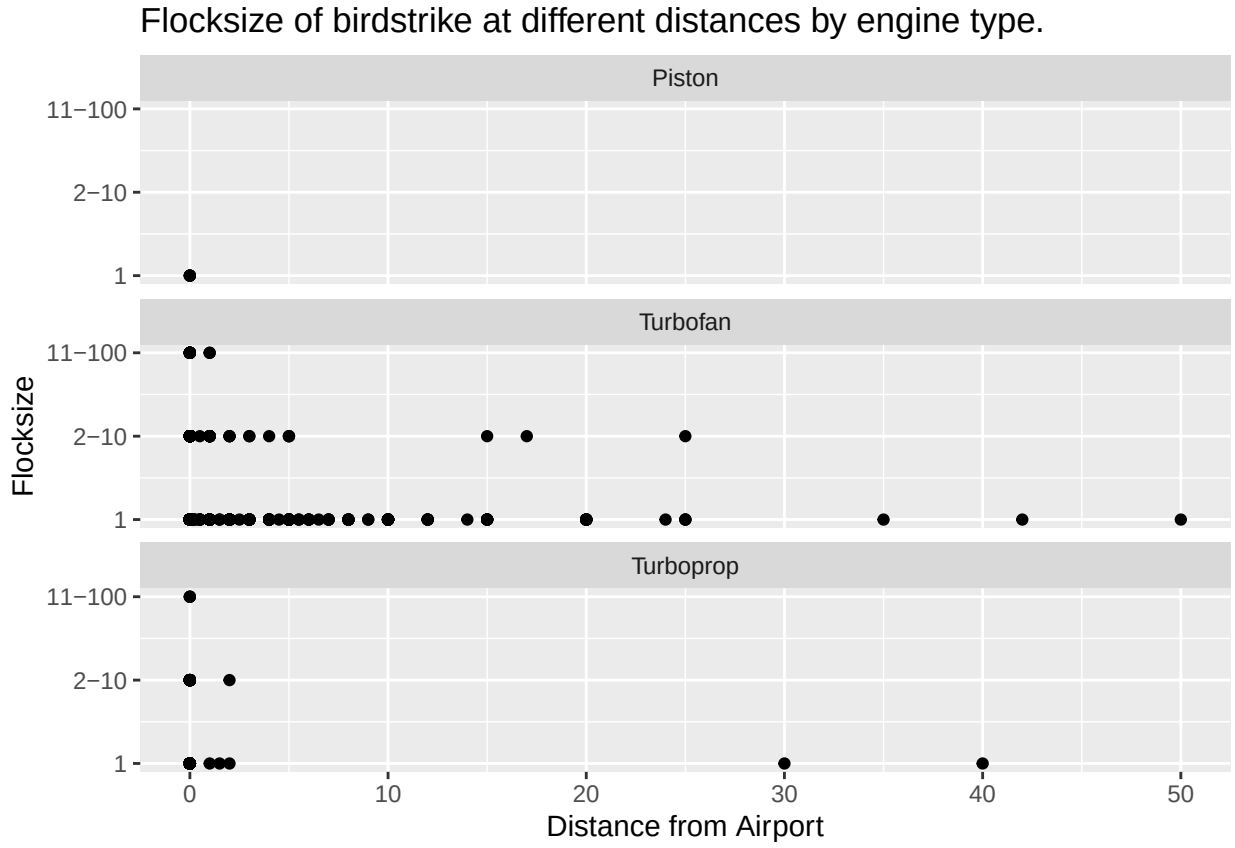
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   2.000   2.000   2.000   2.103   2.000   4.000       5
```

```
#remove missing data from type_eng
#performed in pipeline above
```

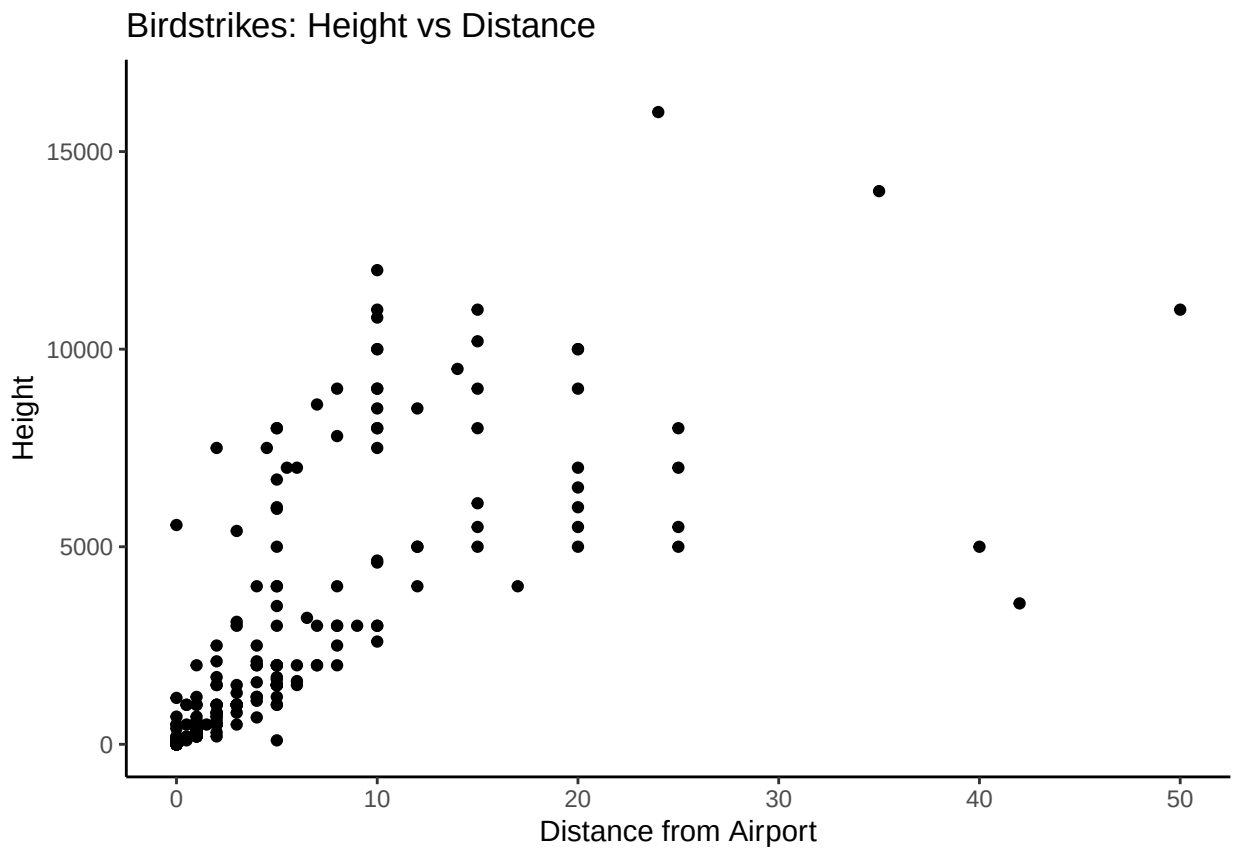
Create a final plot of bird collisions by plane type and distance (nautical miles from airport). Use ggplot and make sure that category labels are appropriate, provide title, axis labels and aesthetics are looking good!

```
birds_by_aircraft<- df[!is.na(df$type_eng),]

birds_by_aircraft$type_eng<-revalue(birds_by_aircraft$type_eng, c("A"="Piston",
                                                                  #"B"="Turbojet",
                                                                  "C"="Turboprop",
                                                                  "D"="Turbofan"
                                                                  #"E"="Glider",
                                                                  #"F"="Helicopter",
                                                                  #"Y"="Other",
                                                                  ))


plot<-ggplot(birds_by_aircraft, aes(x=distance, y=num_struck)) +
  geom_point()+
```

```
  facet_wrap(~type_eng, nrow=3) +
  scale_y_continuous(breaks = 2:4,labels=c("1", "2-10","11-100"))+
  labs(x = 'Distance from Airport',
       y = 'Flocksize',
       title = 'Flocksize of birdstrike at different distances by engine type.')
plot
```




```
ggplot(birds_by_aircraft, aes(x=distance, y=height)) +
  geom_point()+
  labs(x = 'Distance from Airport',
       y = 'Height',
       title = 'Birdstrikes: Height vs Distance ')+
  theme_bw() +
  theme(panel.border = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "black"),
        legend.position="none")
```



Provide a summary of your process and observations from the data. Post to the discussion thread: 1. Any problems/concerns with data.

The NA values in type_eng and damage_level were troublesome. Applying revalue over rows with NA messed up the data structure which prevented the subset() and fct_infreq() functions in the ggplot from running as advertised in the lab. To get around this I applied filtering before revaluing to remove the NA rows prior to revaluing. This allowed the ggplots to be drawn but doesn't imprint the revalues in the original data set. Did anyone find a way to deal with NA's inside the revalue function itself?

Converting num_struck to an integer is problematic. It uses the factor levels as numeric values, so values of 2-4 are given to each row rather than numbers that represent the count of birds in each bird strike where 2=1, 3=2-10, and 4=11-100. This variable records categorical "flock size" more so than the number of birds the aircraft hit. This led me to create a graph which shows flock size at different distances by engine. Incidents involving large flocks have occurred close to the airport. Large flocks are struck less often further away from the airport.

The method I came up with to deal with this was to mutate a new column with integer values based on the value of the num_struck column. I did not use this code for analysis but a method for generating a column with values based on values in another column will come in handy in the future.

```
# Adding new column to data frame with values based on existing values in an adjacent column
df <- df %>% mutate(birds =case_when(num_struck == 1 ~ 0,
                                     num_struck == 2 ~ 1,
                                     num_struck == 3 ~ 5,
                                     num_struck == 4 ~ 50,
                                     num_struck == 5 ~ 100,
                                     )
                    )
```

```
summary(df$birds)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   1.000   1.000   1.505   1.000  50.000       5
```

2. Plots that you think may be relevant to support your assertions.

3. Summary and interpretation of results.

The data set is biased by the airport. DIA serves large passenger planes, primarily with turbo fan engines. This is reflected by the large number of bird strike incidents seen by turbo fan planes. A municipal airport that does not service large passenger planes would see a different mix of engine types in its bird strike data and a military airport might have turbojet and helicopter strikes in its data.

It is also not surprising that most bird strikes occur close to the airport. Most passenger planes cruise at a height of over 30,000 feet where there are no birds. Therefore, most bird strikes are likely to occur close to the airport during takeoff or descent while the planes are at low altitude. Plotting bird strike occurrences by distance and height with a scatterplot demonstrates the relationship between distance and height during bird strike events with most occurring low and close to the airport.