Project 1

Rohit Kamat rgk359

Instructions

For this project, I will be using the data set `Titanic` available in R and data sets `airlines` , `airports` , `flights` , `planes` , and `weather` in the `nycflights13` package.

```
Titanic
```

```
## , , Age = Child, Survived = No
## 
##       Sex
## Class  Male Female
##   1st     0      0
##   2nd     0      0
##   3rd    35     17
##   Crew    0      0
## 
## , , Age = Adult, Survived = No
## 
##       Sex
## Class  Male Female
##   1st   118      4
##   2nd   154     13
##   3rd   387     89
##   Crew  670      3
## 
## , , Age = Child, Survived = Yes
## 
##       Sex
## Class  Male Female
##   1st     5      1
##   2nd    11     13
##   3rd    13     14
##   Crew    0      0
## 
## , , Age = Adult, Survived = Yes
## 
##       Sex
## Class  Male Female
##   1st    57    140
##   2nd    14     80
##   3rd    75     76
##   Crew  192     20
```

```
library(nycflights13)
head(flights)
```

```
## # A tibble: 6 × 19
##    year month   day dep_time sched_dep_time dep_delay arr_time
##   <int> <int> <int>    <int>          <int>     <dbl>    <int>
## 1  2013     1     1      517            515         2      830
## 2  2013     1     1      533            529         4      850
## 3  2013     1     1      542            540         2      923
## 4  2013     1     1      544            545        -1     1004
## 5  2013     1     1      554            600        -6      812
## 6  2013     1     1      554            558        -4      740
## # ... with 12 more variables: sched_arr_time <int>, 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>
```

Problems

Problem 1: (5 pts) The data set `Titanic` is not tidy. Suggest a different way to represent this data set that *would* be tidy. What columns would you need? What would the individual rows represent?

So to make the data set tidy I would have a ‘Sex’,‘Age’, ‘Class’, and ‘Survived’ columns. The observations for each column has to correspond to one persob .

Problem 2: (25 pts) Combine the data sets `flights` and `airlines` such that all the information in the `flights` data set is retained. Using the combined data set and the full name of airline carriers, find the carrier with the longest and the shortest mean departure delay. Perform a statistical test to determine whether there is a significant difference in the departure delays between these two carriers, and interpret your findings.

```
new.data<-left_join(flights,airlines)
```

```
## Joining, by = "carrier"
```

```
new.data %>% select(carrier, dep_delay) %>% group_by(carrier)%>% summarize(mean.dep.dela
y=mean(dep_delay,na.rm=TRUE)) %>% arrange(mean.dep.delay)->mean.data
mean.data
```

```
## # A tibble: 16 × 2
##    carrier mean.dep.delay
##      <chr>          <dbl>
## 1       US       3.782418
## 2       HA       4.900585
## 3       AS       5.804775
## 4       AA       8.586016
## 5       DL       9.264505
## 6       MQ      10.552041
## 7       UA      12.106073
## 8       OO      12.586207
## 9       VX      12.869421
## 10      B6      13.022522
## 11      9E      16.725769
## 12      WN      17.711744
## 13      FL      18.726075
## 14      YV      18.996330
## 15      EV      19.955390
## 16      F9      20.215543
```

```
t.test(new.data$dep_delay[new.data$carrier=="US"],new.data$dep_delay[new.data$carrier=="F9
```

```
## 
##  Welch Two Sample t-test
## 
## data:  new.data$dep_delay[new.data$carrier == "US"] and new.data$dep_delay[new.data$c
arrier == "F9"]
## t = -7.3242, df = 691.84, p-value = 6.704e-13
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -20.83833 -12.02791
## sample estimates:
## mean of x mean of y 
##  3.782418 20.215543
```

With a p-value less than .05, I reject the null hypothesis. I am 95% confident that the true difference in means between the departure delay of carrier “US” and “F9” is not equal to zero and the true mean difference lies within the interval of 12.02791 to 20.83833 minutes.

Problem 3: (40 pts)

a. (30 points) Using the `flights` data set, find the mean distance and mean air time in each month for flights from the three New York airports. Now make one plot that visualizes all this information at once. Your code should be well-commented and describe the various steps you take to create this figure.

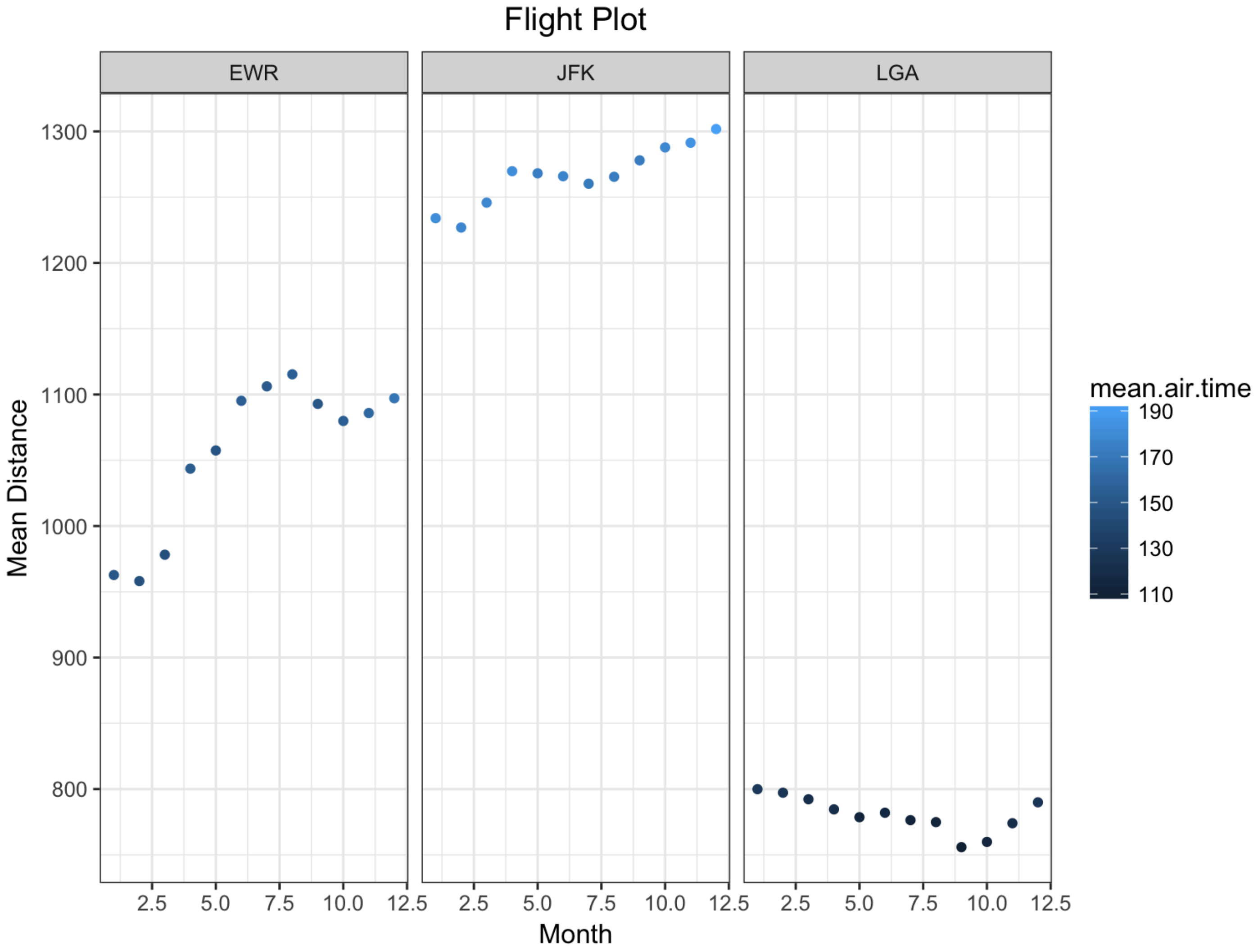
```
#select the month, distance, air time, and origin of the data set
flights %>% select(month,distance,air_time,origin)%>%

#group the data set by origin and month
group_by(origin,month) %>%

# find the mean distance and mean air time of the graph
summarize(mean.dis= mean(distance, na.rm=TRUE),
mean.air.time=mean(air_time,na.rm=TRUE))-> dis.time
dis.time
```

```
## Source: local data frame [36 x 4]
## Groups: origin [?]
##
##    origin month  mean.dis mean.air.time
##     <chr> <int>     <dbl>         <dbl>
## 1     EWR     1  962.7536      149.7083
## 2     EWR     2  958.1264      146.4167
## 3     EWR     3  978.1763      145.0410
## 4     EWR     4 1043.5987      153.9682
## 5     EWR     5 1057.4337      147.8845
## 6     EWR     6 1095.1776      155.0507
## 7     EWR     7 1106.1806      152.1566
## 8     EWR     8 1115.3224      155.0618
## 9     EWR     9 1092.8347      150.0513
## 10    EWR    10 1079.8628      154.9255
## # ... with 26 more rows
```

```
#graph the plot using ggplot, assigning the x-axis, y-axis, and the color for the extra
 dependent variable. I seperated the graph of each of the three airports
dis.time %>% ggplot(aes(x=month, y=mean.dis,color=mean.air.time)) + geom_point() + facet
_wrap(~origin) +xlab("Month") + ylab("Mean Distance") + ggtitle("Flight Plot") + theme(p
lot.title = element_text(hjust = 0.5))
```



b. (10 points) Discuss the information (overarching trends, patterns, etc.) your final plot reveals. Be sure to include in your discussion the similarities/differences among the three New York airports and a clear, logical justification for why you selected the particular geom(s) used to represent this data. Please limit your full response to a maximum of 6 sentences.

My final plot reveals that JFK had flights that had the highest mean distance, followed by EWR, while LGA had flights with the lowest mean distance. In terms of flights mean air time, JFK seemed to have the highest mean air time with the lightest shade of blue, followed by EWR, with LGA having the lowest mean air time. From this plot I also notice that increase mean distance increases with higher mean air time. I used geom point and not a geom line for my ggplot, because month is a nominal variable not a continous variable so geom point would best explain the data between the different months. With mean distance having a wider range of values compared to mean air time, I used the mean distance as my y-axis, with the range of color for my mean air time. For EWR and JFK mean distance increases throughout the year, but LGA did not have a large change in altitude during the year. There was not a large increase or decrease for mean airtime throughout the year within the three airports.

Problem 4: (30 pts)

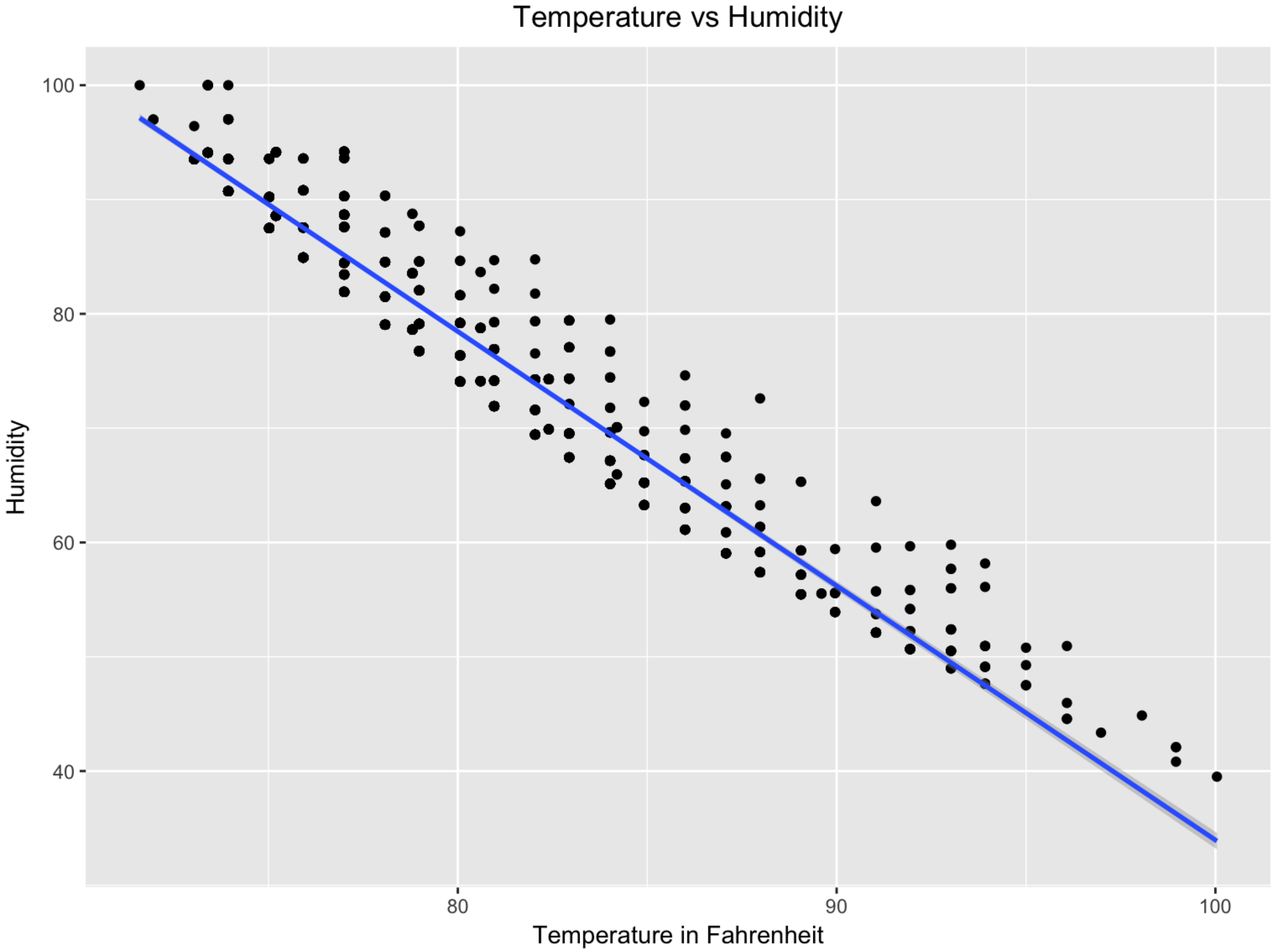
Using the ‘weather’ data set, is there a correlation between temperature and humidity when the dewpoint is above 70 and if so describe the trend of the correlation? Make a ggplot that visualizes the data set in a graph, discuss what your final plot reveals and why you used this type of plot?

```
#focus on temperature and humidity in the weather data set
weather %>% filter(dewp>70) %>% select(temp, humid) -> relation

#perform a linear model
relation_lm<-lm(humid ~ temp, relation)
summary(relation_lm)
```

```
##
## Call:
## lm(formula = humid ~ temp, data = relation)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4851 -2.3609 -0.5397  1.6418 11.8975
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 256.40122    1.40487   182.5   <2e-16 ***
## temp         -2.22435    0.01756  -126.7   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.232 on 1168 degrees of freedom
## Multiple R-squared:  0.9321, Adjusted R-squared:  0.9321
## F-statistic: 1.605e+04 on 1 and 1168 DF,  p-value: < 2.2e-16
```

```
#ggplot graph with month as x-axis, y-axis mean temperature, and the color as the mean h
umidity
ggplot(relation, aes(x=temp, y=humid)) + geom_point() + geom_smooth(method="lm") + theme
_gray() + xlab("Temperature in Fahrenheit") + ylab("Humidity") + ggtitle("Temperature vs
 Humidity") + theme(plot.title = element_text(hjust = 0.5))
```



Using a linear model from the weather data, there was a strong negative linear relationship between the humidity and temperature when the dew point was above 70. As the temperature increase by one degree the humidity when down by 2.22 which was signisficant with p<.05. The correlation is displayed in the plot as one can tell from looking at the graph that as the temperature increases the humidity decreases. I use this type of plot because a scatterplot is used to compare a correlation between two variables.

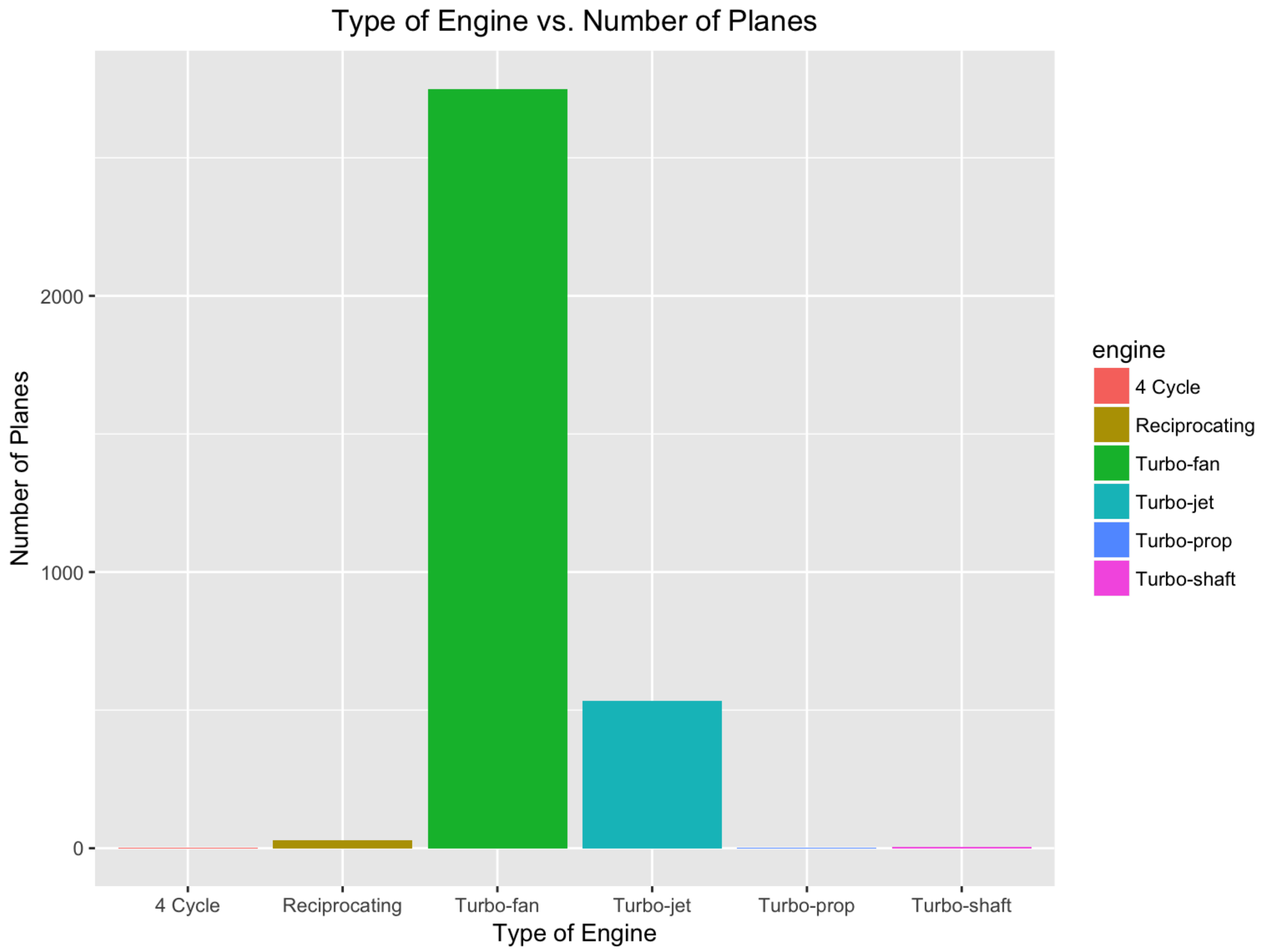
Question 2

From the planes that were used, which engine was the most used by planes? How many planes used the most common engine? Create a ggplot to show the relationship between the two variables and explain why you use this type of plot.

```
#Filter from the planes data to find the amount of planes that were usesd per engine
planes %>% select(engine) %>% group_by(engine) %>% tally()-> new_planes
new_planes
```

```
## # A tibble: 6 × 2
##          engine     n
##           <chr> <int>
## 1       4 Cycle     2
## 2 Reciprocating    28
## 3     Turbo-fan  2750
## 4     Turbo-jet   535
## 5    Turbo-prop     2
## 6   Turbo-shaft     5
```

```
#Use ggplot x-axis is longititude, y-axis is altitude, use a geom smooth to find the overall trend
ggplot(new_planes, aes(engine, n, fill=engine))+ geom_col() + xlab("Type of Engine") + ylab("Number of Planes") + theme_grey() + ggtitle("Type of Engine vs. Number of Planes") +  theme(plot.title = element_text(hjust = 0.5))
```



So after filtering the data based on the type of engine with number of planes that used the engine, Turbo-fan was the most common engine with 2750 planes using it. Then a plot was made using ggplot and geom_col() to show the total amount of planes from each engine. From the plot one can notice that Turbo Fan had the most number of planes with the highest column.