Homework #2

The Houston Flights Data

The dataset that we will be using for this assignment is on flights from Houston, Texas in 2011. The dataset contains all flights departing from Houston airports IAH (George Bush Intercontinental) and HOU (Houston Hobby) in the year 2011. The data comes from the Research and Innovation Technology Administration at the Bureau of Transporation statistics. The dataset is contained in the `hflights` package and contains 227,496 flights and 21 variables. The variables are:

Variable	Description
year	The year of the flight
month	the month of the flight
day_of_month	the day of the month of the flight
day_of_week	day of week of departure (useful for removing weekend effects)
dep_time	time of departure in local time (hhmm)
arr_time	time of arrival in local time (hhmm)
unique_carrier	unique abbreviation for a carrier (ex. wn = Southwest Airlines weirdly)
flight_num	flight number
tail_num	airplane tail number
actual_elapsed_time	elapsed time of flight, in minutes
air_time	flight time, in minutes
arr_delay	arrival delay, in minutes
dep_delay	departure delay, in minutes
origin	origin airport code
dest	destination airport code
distance	distance of flight, in miles
taxi_in	taxi in times, in minutes
taxi_out	taxi out times, in minutes
cancelled	cancelled indicator: 1 = Yes, 0 = No
cancel_code	reason for cancellation: A = carrier, B = weather, C = national air system, D = security
diverted	diverted indicator: 1 = Yes, 0 = No

The first thing that you need to do is to install the `hflights` data and load the package into your session.

```
# installing the hlfights data. Run the install.packages()
# lines in your consle, NOT inside the RMarkdown document.
# install.packages("hflights")
# install.packages("tidyverse")
# install.packages("magrittr")

library("hflights")
library("tidyverse");theme_set(theme_bw())
library("magrittr")
```

The `hflights` package holds one dataset `hflights`. In the code below, I modify the original dataset names to make them better. Do not modify this code!!! We don't know how to do this yet, but we will soon and this type of data "cleaning" is very valuable!

```
hflights %<>% as_tibble %>% purrr::set_names(~ str_replace_all(.x, "([a-z])([A-Z])", "\\1_\\2")) %>%
    purrr::set_names(~ str_to_lower(.x)) %>%
    rename("day_of_month" = dayof_month) %>%
    rename("cancel_code" = cancellation_code)
```

The code above does modify the `hflights` but does not change the name of the dataset. This is done with the pipe-assign operator `%<>%`, which will act as a pipe, but at the end will assign the result back to the object at the top of the "pipeline". All that to say, use `hflights` for all subsequent questions.

Problems

Problem 1: Do flights departing get delayed more often on certain days? I want you to calculate the proportion of flights that were delayed (departing) for each day of the week. Then, comment on your findings. Does it seem like one day is better or worse than the other? Are the differences just "random noise" or do you think there is really an effect there? Hint: `group_by()` and `summarize()`. Also watch out for missing data!! Use the function `drop_na()` before your `group_by()` and `summarize()` calls.

Comment: This is easy!

Problem 2: Maybe I should look at the how long the delays were for each day of the week! First, find the average departure delay time for each day of the week! Then, make a side-by-side boxplot of departure delays for each day that were more than 30 minutes long (there is too much overplotting if we look at all of the data). Comment on your findings. Hint: For the plot, you need to do the following in a similar order: change the `day_of_week` variable into a factor (data type for categorical variables) using `as_factor()` and `mutate()`, filter out the `NA`'s with the `drop_na()` function, then filter so that only use delays of more than 30 minutes. You will need to use the `drop_na()` for almost every problem here.

Comment:

Problem 3: Now that I know something about days of the week, maybe I should also think about what airline I should choose? I want you to count the number of departure delayed flights over the year for each airline carrier.

Problem 4: Wow that's a lot of flights for some carriers! But maybe they just fly out of the airports more often! Check by finding the proportion of departure delayed flights over the year. Comment on the differences/similarity in the summaries from problem 3 and problem 4.

Comment:

Problem 5: Now make a similar boxplot as in problem 2 for each of the carriers and still filter for greater than 30 minutes. You should follow the same general hints for this question. Comment on your findings:

Comments:

Problem 6: How much time do the airports gain or lose between arrival? To answer this question, find the mean amount of time between the arrival delay and the departure delay. Hint: use mutate to create a new variable that holds the difference column. Comment on if the airports are gaining or losing time.

Comments:

Problem 7: Maybe that number above is misleading! Airlines will try to stay on schedule and thus will make the departure on time rather than early. I want you to find the same statistic as above, but this time only for flights that were delayed on arrival. Comment on the degree to which the average changed / stayed the same.

Comments:

Problem 8: Now that I know what day to travel and what airline to travel, I need to find out which airport to fly out of!! There are two airports in this data. IAH and HOU. They are in the `origin` variable. Along the same lines as the other problems, find the average departure delay time for each airport. Also, this time make histograms for departure delays greater than 30 minutes and use `facet_wrap()` to plot them beside each other. Also, try adjusting the number of bins to see if the relationship becomes any clearer. Lastly, I want you to copy and paste the exact same code that you used to make the histograms in the last chunk and add a y aesthetic of '`..density...` Comment on the differences between the plots and what you think the `..density..` did to the plots.

Comments:

Problem 9: This is where you get to be creative (Hopefully that's exciting and not scary)! I want you to come up with a question from the data that I did not already ask and solve it! It can be any question that you want as long as you have to use some `dplyr` function to solve it! Feel free to consult me if you can't come up with a question!