Problems:

Conducted data analysis on flight on-time performance records provided by the Bureau of Transportation Statistics. Due to the enormous value of flight records, we will focus on domestic flights going in and out of major US airports (with more than 5,000 flight records in January and July 2019).

The data set consists of 265344 records (117124 in January and 148220 in July).

Questions:

1. Which day in January has the most completed flights? for July?

2. What are the three busiest airports measured by total movements (the number of inbound flights + the number of outbound flights) in January? in July?

3. Which month (January or July) has a higher percentage of cancelled flights?

4. What are the three (origin) airports with the highest percentage of cancelled flights?

5. The Federal Aviation Administration (FAA) states that "a flight delay is when an airline flight takes off and/or lands 15 minutes or more later than its scheduled time". What are the three airports with the highest percentage of delayed flights?

6. Which month (January or July) has a higher percentage of delayed flights?

7. From the passenger's point of view, what matters the most is the total delay (departure delay + arrival delay). What percentage of the flights in the data set have a total delay time greater than 0 minutes? 15 minutes?

8. Perform the following steps in an attempt to answer the question, "Which day of the week is the best day to fly (domestically)"?

a. Create a new column called DAY_OF_WEEK that indicates the day of the week the travel was recorded. Use the following values for this variable Monday, Tuesday, .., Saturday, and Sunday. Use the real calendar dates so Jan- uary 1, 2019, is a Tuesday and July 1, 2019, is a Monday. If done correctly, the code head(flights$DAY_OF_WEEK) should return a vector: [1] "Sunday" "Tuesday" "Wednesday" "Wednesday" "Thursday" "Friday". Print table(flights$DAY_OF_WEEK).

b. Which day of the week has the highest percentage of cancelled flights?

c. Which day of the week has the highest percentage of flights with a total delay time greater than 15

minutes?

9. Let X be the random variable that represents the number of cancelled flights scheduled to leave Seattle (airport code = SEA) in a day in January. Note that even though X is fully observed in 2019, we pretend X is a random variable as it could be used to estimate/predict the number of cancelled flights in other unobserved years. Use hanging rootograms (from the vcd package) to determine the distribution that fits X the best. To receive full credit, display the rootograms, state the hypotheses, p-value, and explain your conclusion.

10. Use the best-fit distribution from #9 and the corresponding parameter(s) to find the following probabilities:

a. $P(X = 0)$

b. $P(X \geq 1)$

c. $P(X \leq 5)$

d. $P(0 \leq X \leq 3)$

e. $P(0 \leq X < 3)$

11. Let Y be a random variable that represents the number of cancelled flights scheduled to leave Seattle in July. Identify the best-fit distribution and the corresponding parameter(s) as in #9. Based on the best-fit distributions, compare E(X) and E(Y). Which month (January or July) is expected to have more cancellation in Seattle?

12. Use simulation with at least 10,000 replications to approximate $P(X < Y)$.