Quiz 3: Cleaning Your Data

TOTAL POINTS 6

1. To complete this quiz, you need to import the flights data for January 2015 using the provided import function. 1 point

```
flights = importFlightsData("flightsJan.csv");
```

How many variables in the flights table have at least one missing entry?

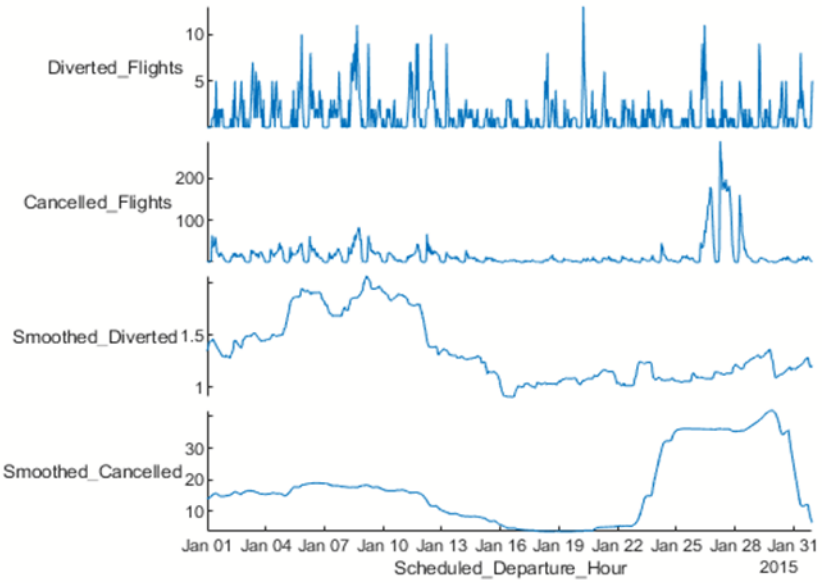
10

2. You can verify that CANCELLATION_CODE is missing whenever CANCELLED = 0, and that ACTUAL_ARRIVAL_TIME is missing whenever CANCELLED = 1. 1 point

What mechanism describes the missing entries in CANCELLATION_CODE and ACTUAL_ARRIVAL_TIME?

- (●) Missing at Random (MAR)
- ( ) Missing Completely at Random (MCAR)
- ( ) Missing not at Random (MNAR)
- ( ) Need more information

3. The next questions investigate if there is a correlation between diverted and cancelled flights. The figure below shows the number of diverted and cancelled flights for each hour (top two plots). The bottom two plots were created using the "Smooth Data" live task using a moving mean with a centered window of 7 days. 1 point



It looks like the smoothed diversions and cancellations may trend together up until the later part of the month. The original cancellations data appears to have significant outliers late in the month, which have a significant impact on the smoothing results. Which of the following approaches would reduce the effect caused by the outliers? (Select all correct answers.)

- [x] Before smoothing, remove the outliers from the cancelled data and the corresponding points from the diverted data
- [ ] Reduce the window size when smoothing
- [ ] Before smoothing, normalize the cancellations data.
- [x] Change the smoothing method to be a moving median rather than a moving mean.

4. Use the code below to create a table with the total flights, diverted flights, and cancelled flights each hour over the month. 1 point

```
flights.SCHEDULED_DEPARTURE_HOUR = dateshift(flights.SCHEDULED_DEPARTURE_TIME,"start","hour");
flightsNotArriving = groupsummary(flights,"SCHEDULED_DEPARTURE_HOUR","sum",["DIVERTED" , "CANCE
flightsNotArriving.Properties.VariableNames = [ "Scheduled_Departure_Hour" "Total_Flights" "Div
```

This table contains the data in the first two plots in Question 3.

Remove the outliers in the cancelled flights using the "quartiles" method with the default threshold factor. Make sure the corresponding points are also removed from the diverted and scheduled time.

How many outliers are removed?

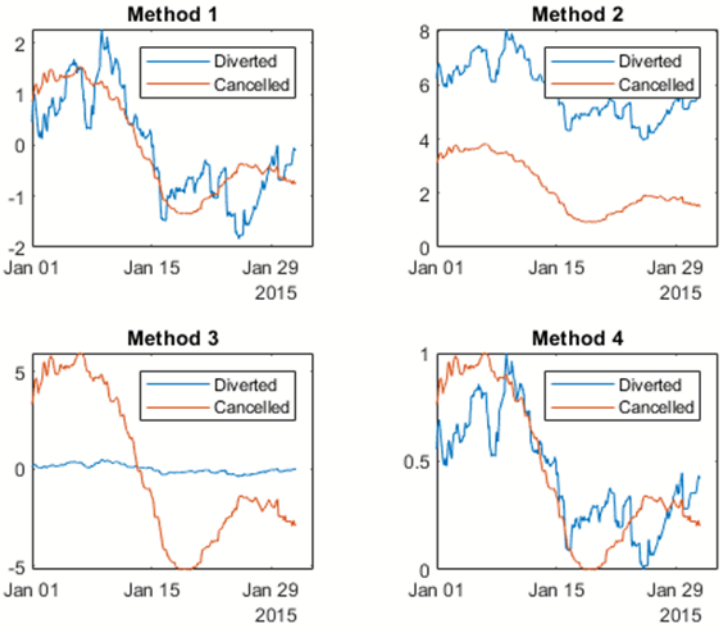
0

5. Using the results from the previous question (Diverted_Flights and Cancelled_Flights with outliers removed) apply a moving mean smoothing with a centered window of 7 days. 1 point
If you visualize the outcome, you should see the following.

There appears to be a stronger correlation. What is the correlation coefficient for cancelled and diverted flights?

- 0.20
- 0.91
- 0.83
- 0.75
- 0.06

6. Below, the previous smoothed results have been normalized using four different normalization methods. 1 point



Examine the plots closely and/or try each method to see the results. For example:

```
1   normalize(dataToNomalize, "scale")
```

In order from 1 to 4, which normalization method was used?

- ○ zscore , range , scale , center
- ◉ zscore , scale , center, range
- ○ scale , range , center , zscore
- ○ center , scale , range , zscore