COMP20008 Elements of Data Processing Project Phase 1 Plan

1.1

Iterate through the data structure for pickup date time - drop off date time to keep a list of the difference. Store as a new column as duration

Create two new data spreads with (Both Jan and Jun)

Index, Duration, Trip distance, Passenger count, Payment amount

1.2

Calculate mean of duration, distance and count for both

1.3

Plot the statistics of Duration, Distance and Trip Fare

Two boxplots for each feature, one for June, one for Jan

1.4

Use IQR to create a justifiable range for data collected, everything outside is outlier

Find the ratio of outliers = (undervalue outliers + overvalue outliers)/ total entries

X% of instances in “Feature” are noisy! Normal range is between IQR -3 and IQR3

Code to replace noisy values in data

While i < n,

If A[i] <lower limit or i>upper limit

A[i] = mean value

i++

2.

Use Datetimeindex to create isweekend column (function)

0 for weekday, 1 for weekend

% of weekend trips=XXX

Column for hour in which where the trip began

Just look at hour value and use counter variable (can reset to zero after every number change)

Value 0 for 00:00 to value 23

Graph histogram, one for weekend, one for weekdays

Ranges [x, y), so including x but excluding y

Y axis for frequency

Column for income efficiency

Payment/ Duration

Barplot showing mean efficiency (y) and hours of day (x), [0,24)