UNIVERSITY OF LONDON

Programming for Data Science
Coursework Project 2024

by Shuayb Ibrahim

Student Number: 220298953
Course Code: ST2195

Contents

Introduction

In this analysis report, we aim to present and analyse the results of a data analysis in the form of a plethora of visualisations, addressing five distinct questions across two chapters.

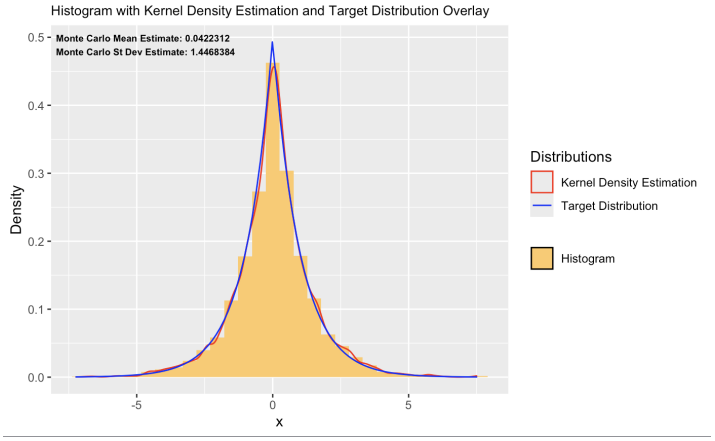
The first chapter consists of two parts, (a) and (b). In part (a) we will be showing the outcome of a Markov Chain Monte Carlo simulation represented by a histogram with a Kernel Density Estimation Overlay. In part (b), we will be assessing the convergence in Markov Chain Monte Carlo simulations by analysing popular convergence diagnostic $\hat{R}$.

The second chapter will be using the Airline On-Time dataset provided by the 2009 ASA Statistical Computing and Graphics Data Expo between the years 1995 and 2004. This chapter also consists of 3 parts, (a), (b) and (c). In part (a) based on the findings, we will look at what are the best times of the day and the day of the week across the years in order to minimise delays as a traveller. In part (b), we will attempt to find out whether or not older planes suffer more delays on a year-to-year basis by conducting a Pearson's Product-Moment Correlation test. Finally, in part (c), we will be fitting a Logistic Regression Model for the probability of diverted United States flights and evaluate the model to understand the best predictors of diverted flights.

Chapter 1

Markov Chain Monte Carlo

1.1 Random Walk Metropolis Algorithm

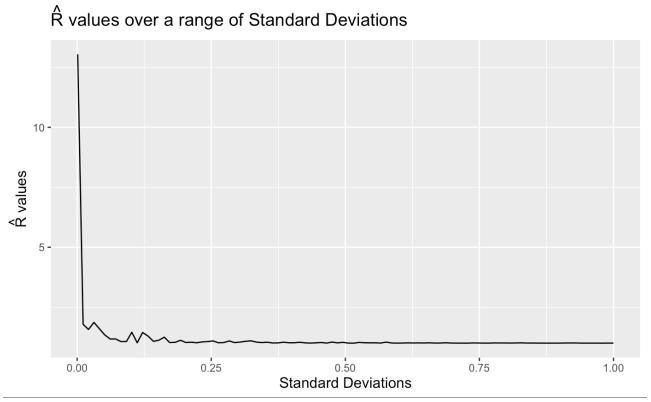


Random Walk Metropolis Algorithm

```r
{r}
metropolis <- function(N,s) {
  samples <- list(runif(1))

  for (i in 1:N-1) {
    #sample candidate from normal distribution
    last_sample <- samples[[length(samples)]]
    candidate <- rnorm(1,mean = last_sample,sd = s)
    log_r <- log(f(candidate)) - log(f(last_sample))
    log_u <- log(runif(1))

    #accept or reject calculated probability
    if (log_u < log_r) {
      samples <- c(samples,candidate)
    }
    else{
      samples <- c(samples,last_sample)
    }
  }

  return(samples)
}
```

Markov Chain Monte Carlo is essence an algorithm that is used in numerous fields to estimate values that aren't straightforward to calculate. In the plot above, you can see that I used the normal distribution to generate sample candidates which will be used in the estimation. Looking at the mean estimate and standard deviation, we can see the normal distribution approximates it well.

1.2 $\hat{R}$ Convergence Diagnostic

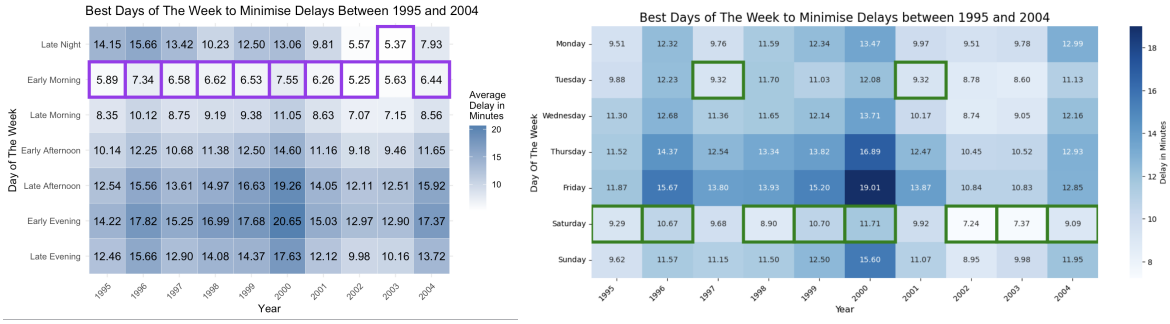


To test the assumption that the algorithm converges, we calculate the Gelman-Rubin diagnostic ($\hat{R}$ diagnostic) over a range of standard deviations. Whilst keeping the sample size and sequence length fixed, in the plot you can see it does converge towards 1 approximately.

Chapter 2

On-time Airlines Data

2.1 What are the best times and days of the week to minimise delays each year?



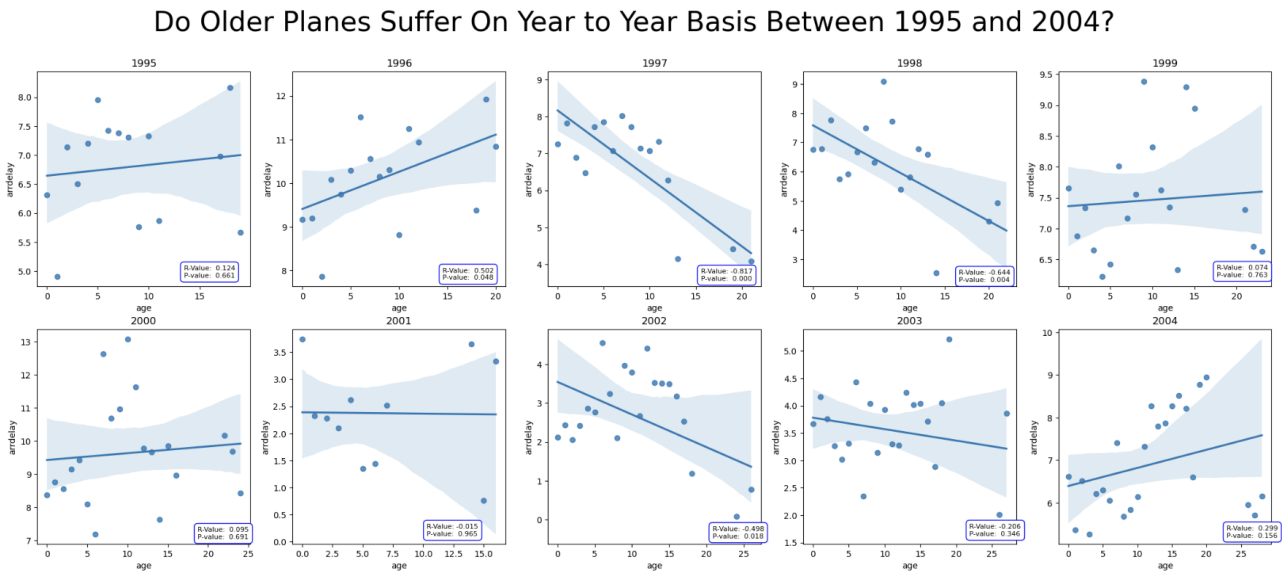
From the traveller's perspective when we think about delay, it primarily has to do with respect to arrival time as the destination is of most importance so we focus on arrival delay. To specifically focus on delays only and not have any skewed data, I set all early arrivals to just no delay. I categorised the booking times to the time of the day to give the traveler a time window instead of a specific time.

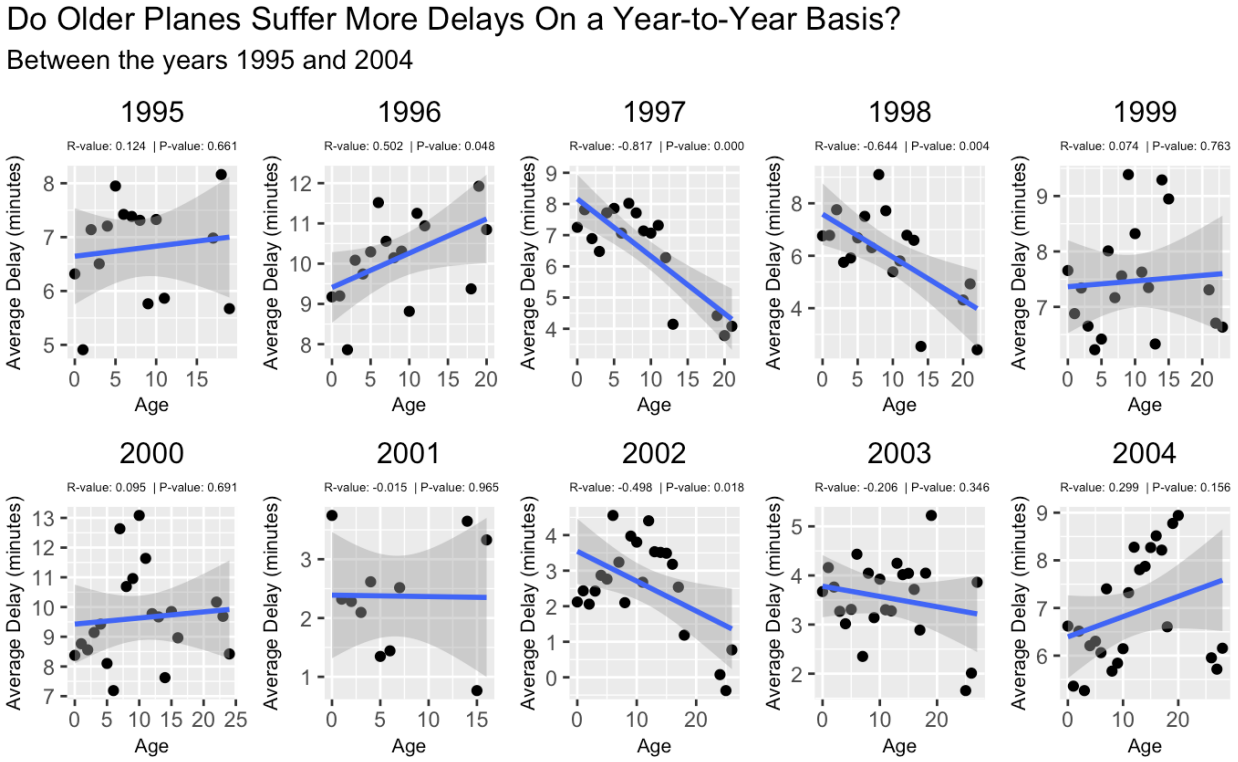
From the results, we can see that over the years the early morning time window between 4am and 8am has consistently been, outside of 2003, the time slot with the least amount of delay.

Looking best day of the week over the course of 10 years, Saturday has been the best day for 80% of the time frame. It has also been the best day to travel for the last 3 years of that time frame.

Saturday would be the recommended day with Tuesday as a solid alternative as it has been the best day for 2 out of the 10 years.

2.2 Do Older Planes Suffer More Delays on a Year to Year Basis?



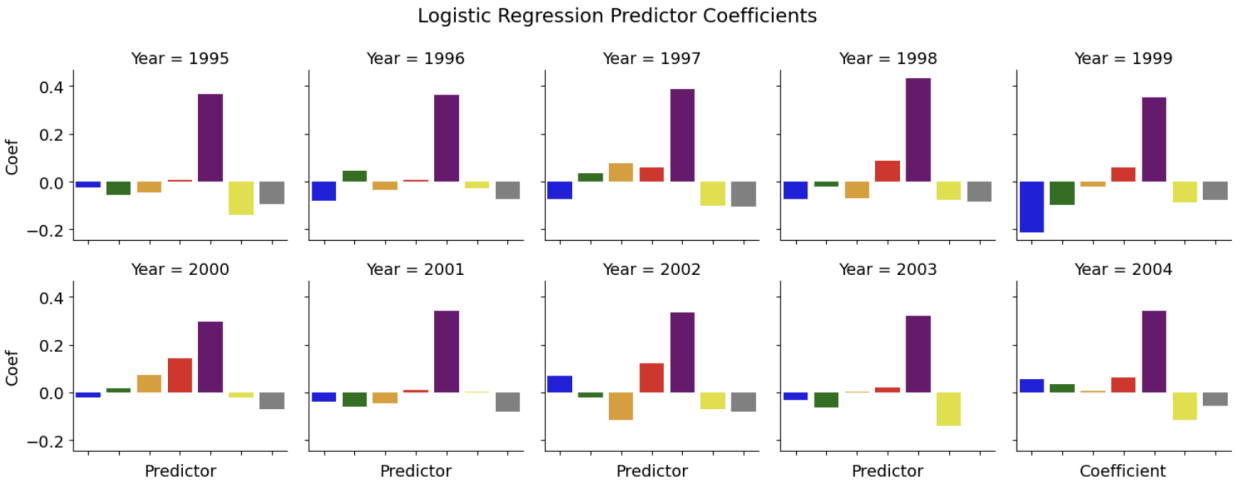


To answer this question, I used Pearsons Correlation test. We use this test to determine whether the age of a plane is correlated with an increase in average delay and whether the correlation is significant enough using the p-value.

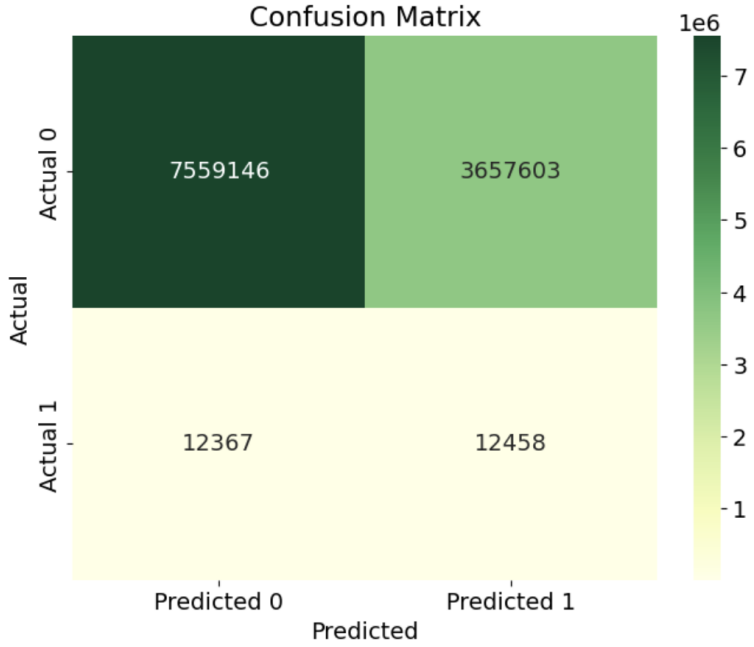
In the graphic above, we can see the scatterplot of airplane age and average delay on a year to year basis with a regression line as well as the r-value and p-value. Across the board, we cannot spot any pattern and that the relationship changed on a year to year basis.

To conclude, there is very little/close to no evidence to say that older planes suffer more delays on a year-to-year basis compared to younger planes.

2.3 Predicting Diverted US Flights




```
{'Month': 'blue', 'DayofMonth': 'green', 'DayOfWeek': 'orange', 'Hour': 'red', 'Distance': 'purple', 'CarrierDelayRate': 'yellow', 'Airport
DelayRate': 'grey'}
```

	precision	recall	f1-score	support
0	1.00	0.67	0.80	11216749
1	0.00	0.50	0.01	24825
accuracy			0.67	11241574
macro avg	0.50	0.59	0.41	11241574
weighted avg	1.00	0.67	0.80	11241574

To predict Diverted US Flights, I trained a logistic regression model using the following predictors: Month, DayOfWeek, TimeOfDay, Distance, CarrierDelayRate and AirportDelayRate.

The model produced an accuracy score of 0.67 which makes it accurate 2/3 times. Looking at the confusion matrix, we can see that the model done a great job predicting flights that won't be diverted in particular. It predicted a lot of false positives too.

The best predictor is by far CarrierDelayRate, which is the delay rate of a given airline service suggesting they might have a bigger role in diverted flights which should be explored.