ST2195 PROGRAMING FOR DATA SCIENCE COURSEWORK

Gordon Lin (220572956)

Contents

1. Introduction

There are two parts to this report the first part on random walk metropolis and the second part on using flight data for all commercial flights in the USA from 1987 – 2008 to provide answers to the questions outlined below:

- What are the best times and day of the week to minimise delays each year?
- Evaluate whether older planes suffer more delays on a year-to-year basis.
- Use Logistic Regression Model for the probability of diverted flight using as many features as possible and visualise the co-efficient across the years.

Language used are R & Python. And the database "**FlightStats.db**" was created using DB-Browser. The four tables in the database are:

1) "**airports**" table consisting of data on the geographic location of each airport and their respective IATA code.
2) "**carriers**" table consisting of data on the major carriers that operated within the USA.
3) "**flights**" tables consisting of the data on commercial flights from (1998 – 2007 chosen 10 years.)
4) "**planes**" table consisting of data on different planes.

Other notable things to note. All codes between R & Python are created to serve the same purpose. But they might look different due to the restrictions of the specific language. And For part 2C two different random samples of **N = 10000** are chosen from the 10 years due to the overly large dataset.

NOTE: IF BOTH GRAPH ARE USED TO SERVE THE SAME PURPOSE AND HAVE SIMILAR RESULTS THE NICER ONE WOULD BE SELECTED FOR ILLUSTRATION PURPOSES

1.1. Formula Used

Probability = Favorable cases / possible cases x 100
(Used to calculate Percentage Chance in part 2)

EV=∑P(Xi)×Xi
(Used to calculate Expected Value in part 2)

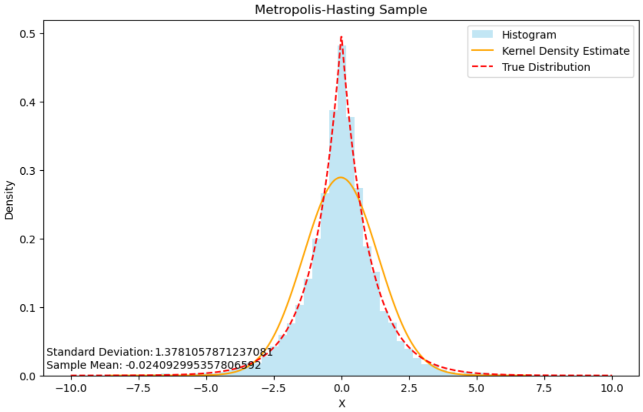
1.2. Bibliography

https://www.qualtrics.com/au/experience-management/research/determine-sample-size/

(Used their calculator to choose appropriate sample size.)

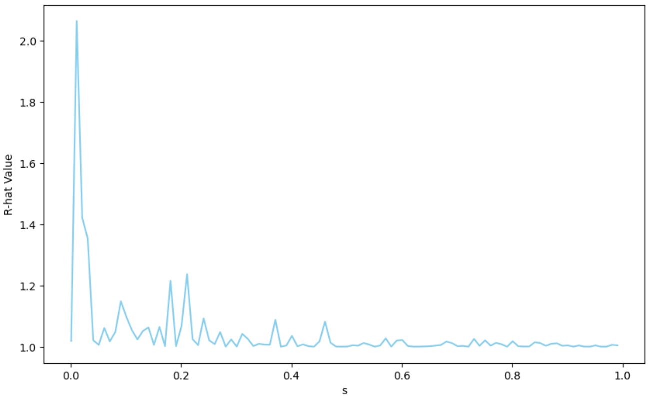
2. Part 1 of the Coursework

2.1. Part a



For this graph it shows the sample values of parameter ‘x’ in the metropolis hasting algorithm and ‘y’ shows the density of the sampled values. The metropolis hasting algorithm plots the histogram and that the kernel density estimate is a smoother estimate of the distribution. And the closer the true distribution is to the histogram it means the algorithm is sampling the target accurately.

2.2. Part b



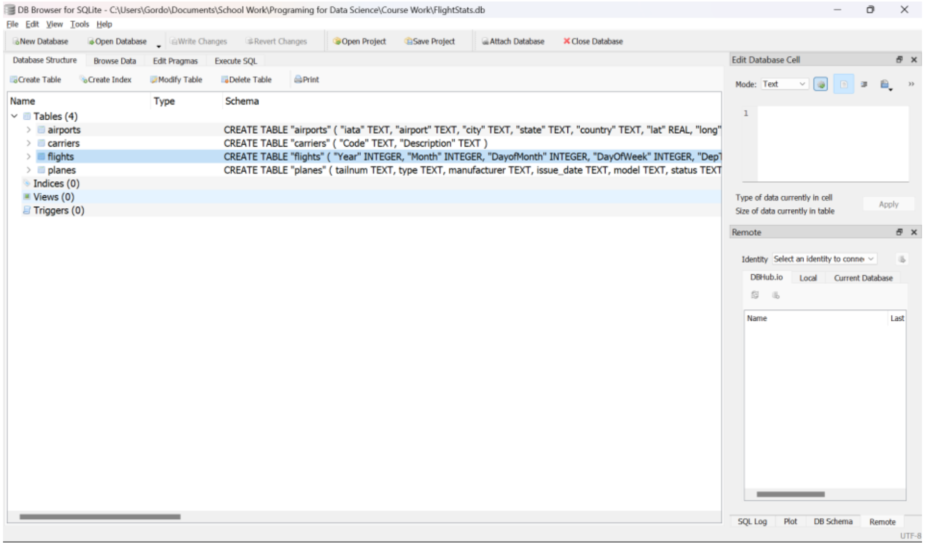
This graph shows the relations between R-hat Value and Standard Deviation and that when R-hat is close to 1 indicates convergence while 2 indicates lack of convergence in the metropolis hasting algorithm. So, this graph shows the if the standard deviation is too small there will be a complete lack of convergence however if standard deviation is too big R-hat converges showing less exploration of the parameter space and poor mixing of the chain. So this graph helps you to visualize the amount of standard deviation is needed to find the perfect balance of coverage.

3. Part 2 of the Coursework

3.1. Database setup

The Database setup involves extracting data from https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/HG7NV7

And then adding the Different CSV files into on Database. In my case I used DB-Browser.



Name	Type	Schema
Tables (4)		
airports		CREATE TABLE "airports" ("iata" TEXT, "airport" TEXT, "city" TEXT, "state" TEXT, "country" TEXT, "lat" REAL, "long"
carriers		CREATE TABLE "carriers" ("Code" TEXT, "Description" TEXT)
flights		CREATE TABLE "flights" ("Year" INTEGER, "Month" INTEGER, "DayofMonth" INTEGER, "DayOfWeek" INTEGER, "DepT
planes		CREATE TABLE "planes" (tailnum TEXT, type TEXT, manufacturer TEXT, issue_date TEXT, model TEXT, status TEXT
Indices (0)		
Views (0)		
Triggers (0)		

As mentioned in the introduction the 4 tables are “airports”, “carriers”, “flights”, “planes”.

3.2. What are the best times and day of the week to minimize delays each year?

Before we can find best time of day we need to define Morning, Afternoon, Evening and Night. As they are not a variable in any of the table, so I created a sub query named TimePeriod.

WHEN TIME >= 500 AND TIME < 1200 THEN 'Morning'

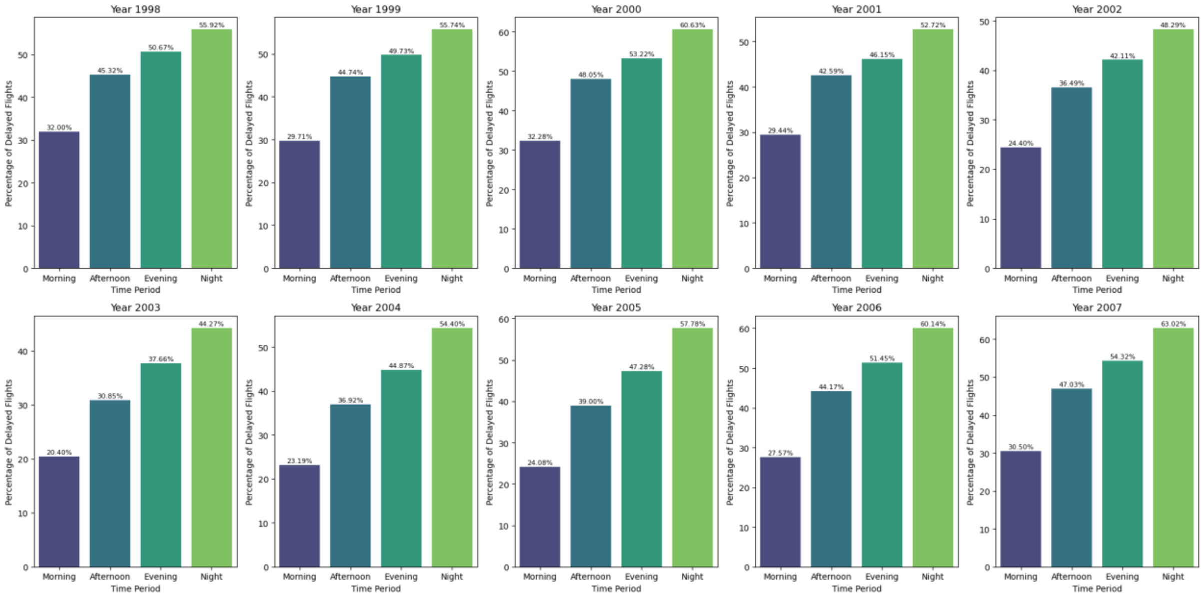
WHEN TIME >= 1200 AND TIME < 1700 THEN 'Afternoon'

WHEN TIME >= 1700 AND TIME < 2100 THEN 'Evening'

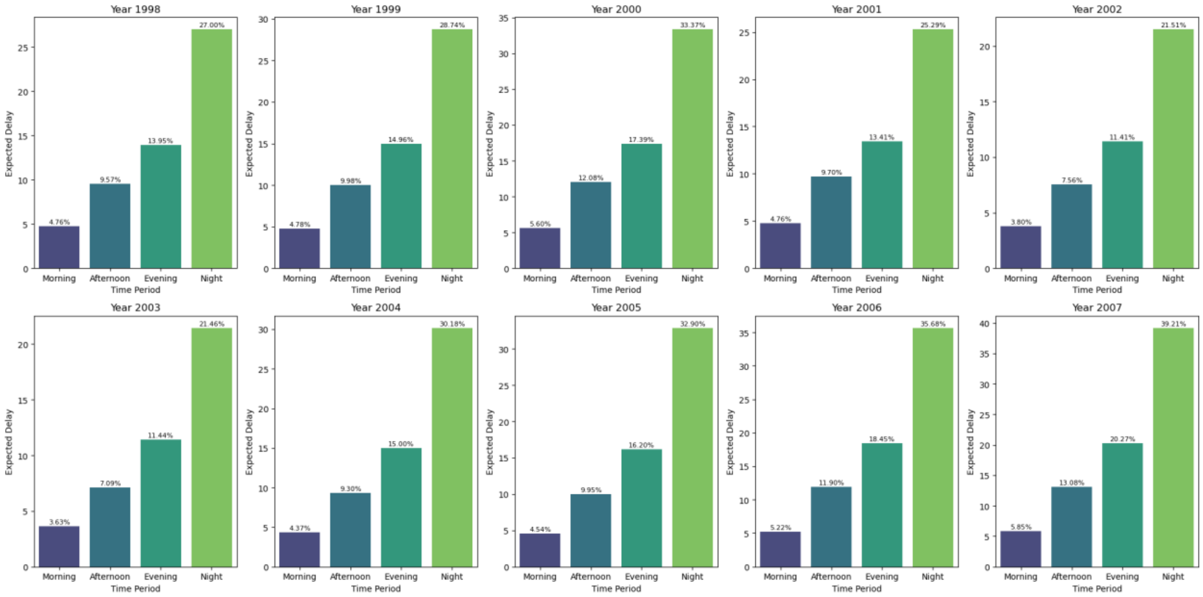
ELSE 'Night'

Essentially any time between 5am and 12pm as Morning and 12pm till 5pm as Afternoon 5pm till 9pm as Evening and 9pm till 5am as Night.

To tackle this problem, I initially counted percentage chance of delay using the probability formula which is **Probability = Favorable cases / possible cases x 100** in this scenario Favorable case being Delayed Flights and Possible Case being Total Number of Flights. Per TimePeriod to get the graph below.

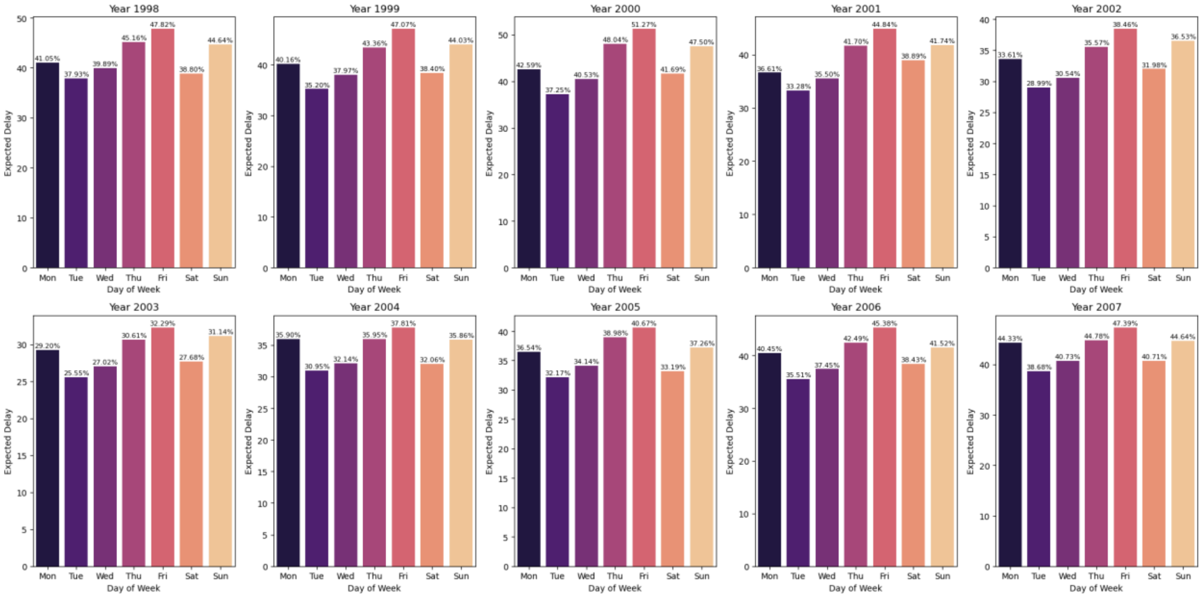


After that I used the formula **EV=∑P(Xi)×Xi** to find a more accurate Estimate for delay and moving forward all my steps will only plot Expected Value due to me realising it's a much better estimate for Delay as it takes into account not just chance of delay but how long each delay is as well.



Using Expected Delay I managed to plot this for delayed based of TimePeriod and the comparison between this two shows the vast difference between expected delay and percentage chance of delay.

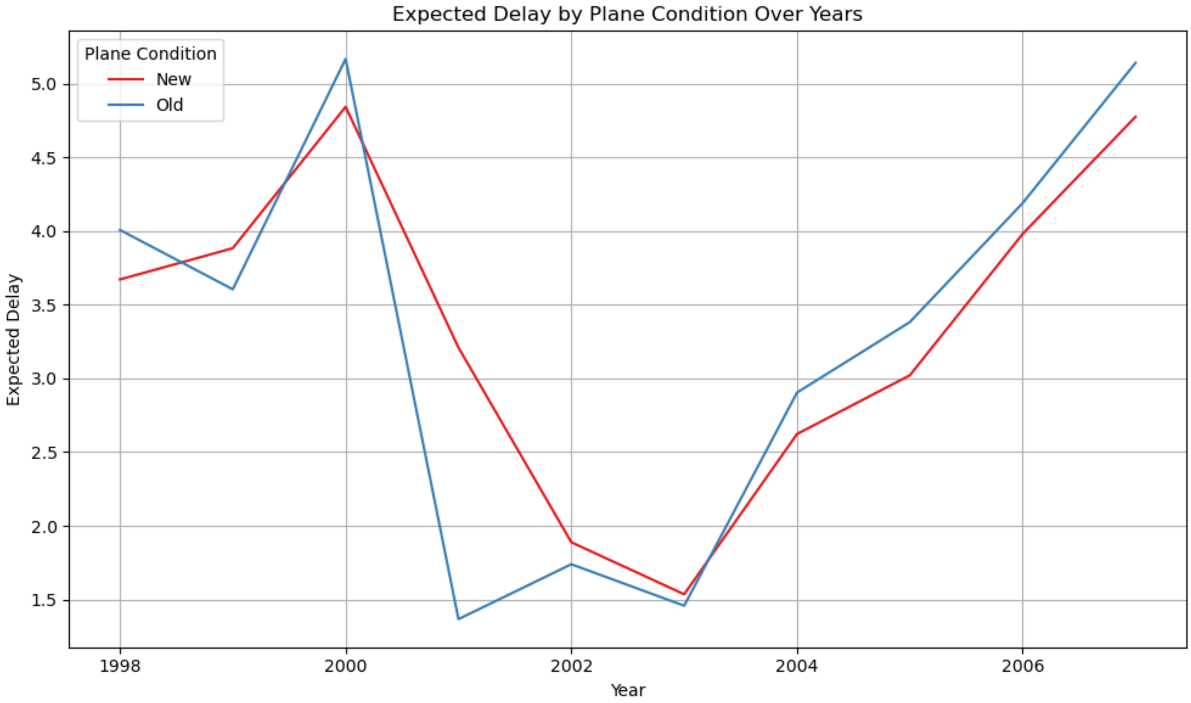
Next up is day of the week it is using a similar query just that instead of using TimePeriod I am using DayOfWeek which is provided however I renamed the values in DayOfWeek as its listed as 1 – 7 in the CSVs so it's now Mon – Sun respectively.



And with these two graphs we can see that the pattern for Expect Delay for both TimePeriod and DayOfWeek doesn't change throughout the year of 1998 – 2007 hence we can infer from these two graphs that the best time and day of the week to **MINIMISE** delay is **TUESDAY MORNING** (Answer for Part 2a)

3.3. Evaluate whether older planes suffer more delay on a Year-to-Year basis.

To first do this question we have to determine what is considered Old. In this case I used any plane before 1998 as OLD and anything after 1998 as NEW. I then us the same Method as part 2A to find the Expected Delay between the NEW and OLD planes spanning from 1998 – 2007.

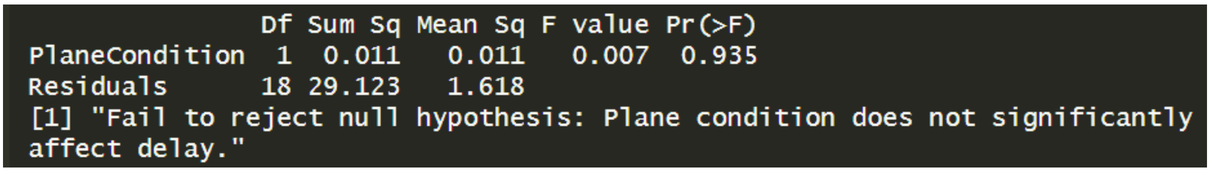


After getting the numbers I did a ANOVA hypothesis test with conditions

H0: Plane Condition does not significantly affect delay.

H1: Plane Condition significantly affect delay.

And the results returned were:

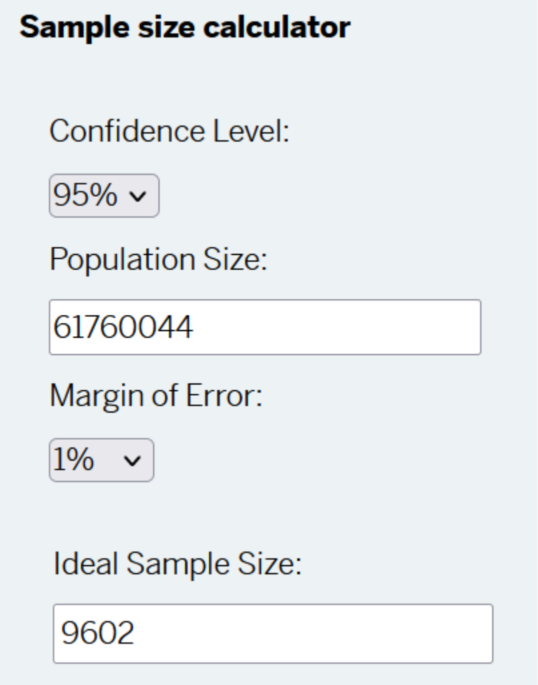


```
               Df Sum Sq Mean Sq F value Pr(>F)
PlaneCondition  1  0.011   0.011   0.007  0.935
Residuals      18 29.123   1.618
[1] "Fail to reject null hypothesis: Plane condition does not significantly
affect delay."
```

Therefore, the answer to Part 2b is **NO, PLANE CONDITION DOES NOT SIGNIFICANTLY AFFECT DELAY.** (Answer for Part 2b)

3.4. Use Logistic Regression Model for the probability of diverted flight using as many features as possible and visualise the co-efficient across the years.

Before we begin this question as using the whole population size of 61,760,044 is to large, I used the sample calculator from the website in the bibliography to find an Ideal Sample Size with 95% Confidence Interval Level and 1% Margin of Error before rounding it up to 5000 to create a sample size for each diverted and non-diverted so totalling in **10,000**.



Sample size calculator

Confidence Level:

95%

Population Size:

61760044

Margin of Error:

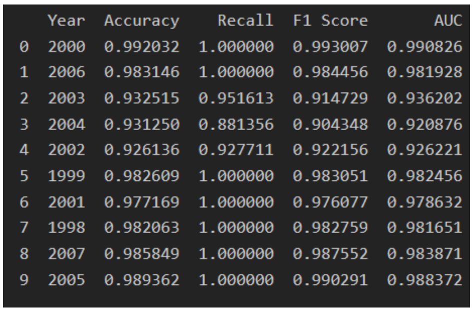
1%

Ideal Sample Size:

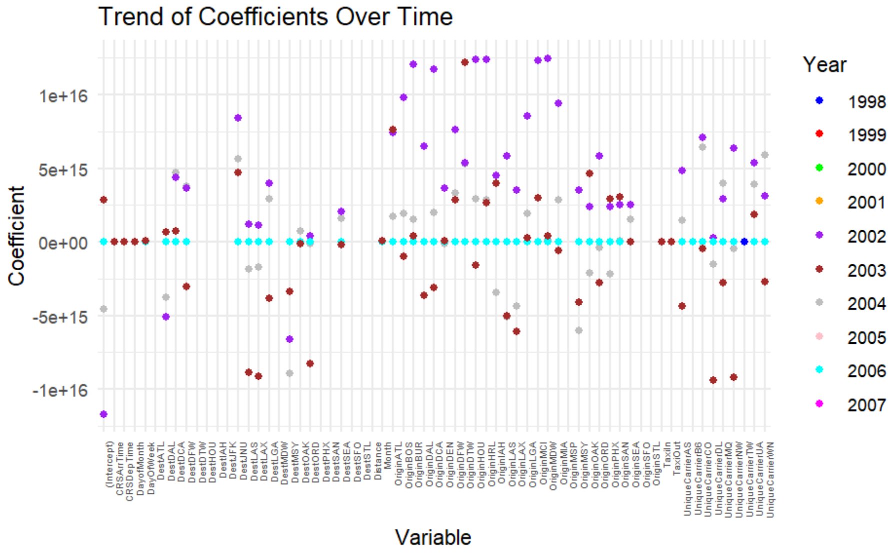
9602

And as the Sample Population are pulled **randomly** in the query with the only exception to Origin Dest and UniqueCarrier where only their top 25 count are called this is so that we can ensure that both the train and the test data have a similar factor level, and the library used to do the logistic regression is different the **answers from R and Python will differ** hence in this portion, I am going to show both graph side by side. The library used for R is (GLMNET) and the library used for Python is (SKLEARN).

3.4.1 R Version

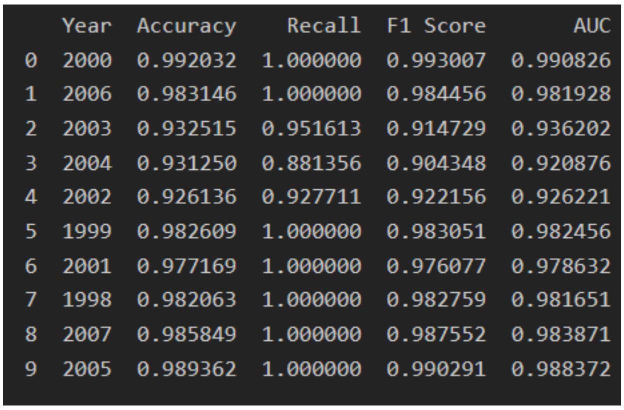


```
   Year  Accuracy    Recall  F1 Score       AUC
0  2000  0.992032  1.000000  0.993007  0.990826
1  2006  0.983146  1.000000  0.984456  0.981928
2  2003  0.932515  0.951613  0.914729  0.936202
3  2004  0.931250  0.881356  0.904348  0.920876
4  2002  0.926136  0.927711  0.922156  0.926221
5  1999  0.982609  1.000000  0.983051  0.982456
6  2001  0.977169  1.000000  0.976077  0.978632
7  1998  0.982063  1.000000  0.982759  0.981651
8  2007  0.985849  1.000000  0.987552  0.983871
9  2005  0.989362  1.000000  0.990291  0.988372
```

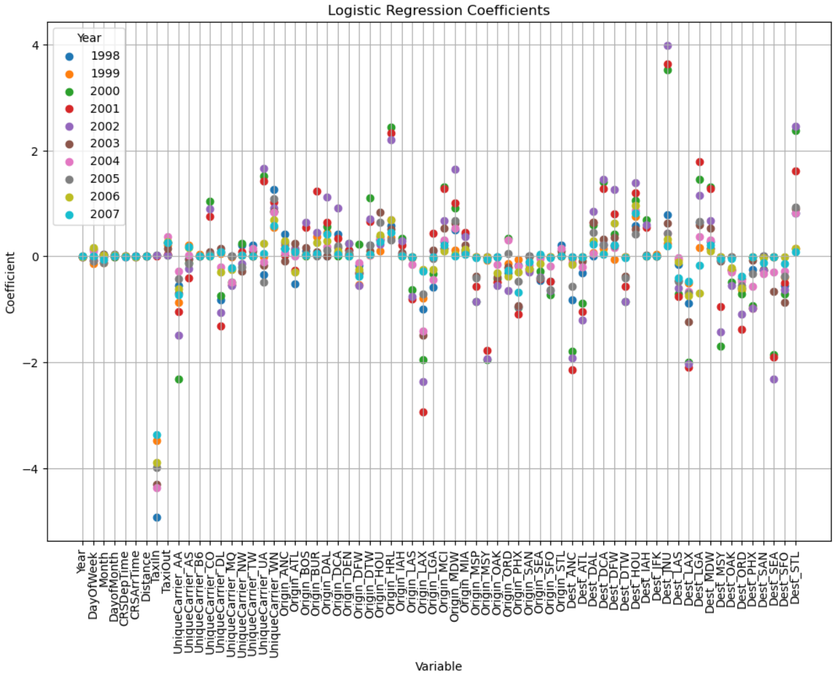


For the Accuracy, Recall, F1_Score and AUC values calculated by using the training data of the portion before it is predicted using the test data that shows the prediction in the dot diagram. Which is split in an 80/20 manner. And from the graph only 2002, 2003 and 2004 have enough variables for the model to predict. This might be due to the singularity of the variable. But using this we can tell that most variables have huge positive or negative relation to Diverted Flights.

3.4.2. Python Version



```
   Year  Accuracy    Recall  F1 Score       AUC
0  2000  0.992032  1.000000  0.993007  0.990826
1  2006  0.983146  1.000000  0.984456  0.981928
2  2003  0.932515  0.951613  0.914729  0.936202
3  2004  0.931250  0.881356  0.904348  0.920876
4  2002  0.926136  0.927711  0.922156  0.926221
5  1999  0.982609  1.000000  0.983051  0.982456
6  2001  0.977169  1.000000  0.976077  0.978632
7  1998  0.982063  1.000000  0.982759  0.981651
8  2007  0.985849  1.000000  0.987552  0.983871
9  2005  0.989362  1.000000  0.990291  0.988372
```



From the Python Portion, the random sample has better values due to lesser singularity of the variables hence there are more plots. And from the graph we can tell that 'TaxiIn' has negative relation to Diverted flight and this makes sense due to if the flight is diverted the plane will not need to Taxi in due to it not even taxing out in the first place. And from the graph we can tell that there are small differences between UniqueCarrier and their relations to diverted flight with UniqueCarrierUA having the highest relation to diverted flight means, that their flight have diverted more times compared to the other UniqueCarrier, hence it the model predicted their higher relation with diverted flight while UniqueCarrierAA has negative relation with diverted flight so showing it has the least chance of flight diversion same for OriginLAX having the least relation while OriginHRL having the highest relation and lastly DestSEA having the lowest relation and DestTNU having the highest relation and. Those with big fluctuation might suggest faults in the model and might be due to outlier variables values or other unaccounted factors. (Answer for Part 2c)

4. Summary

From the flight analysis ranging from 1998 – 2007 it showed the best time and day of week to take a flight whether if planes condition(age) will affect plane delays and logistic regression models where we predict different variables relation to diverted flight. With this information we can improve either operation. Or avoid the disruptions to your flight.