On-time Performance of Commercial Air Travel

Christos Dimopoulos
s1687053
s1687053@sms.ed.ac.uk

Lefteris Manousakis
s1671145
s1671145@sms.ed.ac.uk

Georgios Pligoropoulos
s1687568
s1687568@sms.ed.ac.uk

Abstract

The current report presents an analysis on flights delays using data from the United States Department of Transportation. Initially, extensive exploratory analysis and visualization on the on-time statistics of various commercial flights is performed. The problem is broken down into two main questions: *whether* a flight was delayed and *how much* was this delay. The most important features are identified and used to perform binary and multi-class classification to solve these two problems. Different classifiers are used and feature engineering, as well as, hyper-parameter tuning are employed to improve the results. We manage to predict with 85.4% accuracy if the flight is going to be delayed and with 56.0% the delay time interval.

1 Introduction

Air transportation has become one of the most crucial ways of transport [1]. Each year new flights are introduced to the air traffic system, but the available resources do not increase relatively to the growing demand. The desire to maximize the utilization of resources such as airports, aircrafts, airline companies employees has lead to significantly reduces time between arrival and departure for each aircraft. As it is expected, this phenomenon increases the chances of flight delay, which will be propagated to other flights and airports, affecting a relatively large part of the air transportation network[2].

At the same time, delays are becoming one of the biggest concerns of airline companies as they are the basic factor responsible for customer disappointment. As recent research suggests ([3]), the consequences of a few minutes flight delay can be really important, as they can result in canceled flights, airport congestion, environmentally fuel wastage and of course serious under-utilization of the scarce available resources. Even if the delays cannot be avoided, airline companies have a great interest in providing accurate predictions to the passengers about whether their flight is going to be delayed or not.

1.1 Objective

This project focuses on providing useful techniques for predicting flight delays. This objective is divided to two sub-goals. The first is to predict *whether* a flight will be delayed and the second is to predict *how much* it will be delayed. We begin by performing an extensive exploratory analysis to the available data. This aims to provide useful insights about which factors have a determinative role to delayed flights. Feature engineering will be used to boost the ability of the data to discriminate the characteristics of delayed and non-delayed flights. The tasks end by testing numerous classifiers on the dataset and finding the one that better solves the problem. Hyper-parameter tuning will be employed to further improve their performance.

1.2 Background

The business implications of the problem are vast and determinant for the airline companies. Customer satisfaction and airport utilization depends on the delays. As it is expected, several approaches have been adopted to solve the problem. Some of the most effective classifiers have been used to tackle the problem of producing trustworthy delay predictions. Briefly, among others, the recent state-of-the-art techniques that have been used include Naive Bayes [4], Random Forests [5], SVMs [4], K-Means clustering [5] and recently (2016) even Artificial Neural Networks [6],

1.3 Motivation & Importance

Undoubtedly, the discussed problem is constantly growing and the goal of proposing a good solution is not only interesting because it is a challenging task but because it can massively facilitate the better resource allocation and utilization. Additionally, it can improve the provided customer service level. The importance of the problem -our main motive- is highlighted by the continuous expansion of the problem related literature and the fact that powerful Neural Networks have been proposed ([6]) to solve the delay prediction problem.

2 Data preparation

In order to build predictive models for the flight delays, it is a common practice for the literature (e.g. [4]) to use data originating from the Bureau of Transportation Statistics [7]. The service contains all the details about the commercial flights of United States for the period 1987 to 2016. We manually collected and concatenated the flights for all the months of 2016. The sample was unable to be effectively processed and thus we had to reduce the number. Traditional approaches take into consideration the seasonality and in order to avoid looking into specific months and losing this information, they spatially filter the dataset [6]. In our case, we choose to preserve all the flights originating or heading to any airport of the New York state which is one of the most common US destinations.

The acquired dataset enumerates 492.181 instances with 29 features. Briefly the features contain information about the day, week and month of the flight, the flight number and destination and origin information, including airport, city and state. Additionally, information about the carrier, the departure delay, arrival delay, distance and actual and supposed flight time are provided.

Before proceeding, we remove duplicate information, because some features have both abbreviations and the actual names (states, airports, carriers). Additionally, year attribute does not vary, so it is dropped. Week of the year is generated and then date flight is dropped as the information is contained separately in day and month features.

The dataset was further preprocessed before the explanatory analysis and the classification task. First of all the canceled flights were removed, because the prediction must be based on whether he flights are going to be late and flight cancellation cannot provide helpful information. In total, 8949 canceled flights, which is 1.85% of the total dataset. Furthermore, a portion smaller than 0.4% or 1570 flights was removed because it contains null values. We end up with a clean dataset of 481.662 commercial flights, where we have binary delay variable and a delay group with 13 classes to be the targets for binary and multi-class classification.

3 Exploratory Analysis

The dataset contains hidden information that can have a determinant role on building a good classification model. As the first sub-goal, we aim to perform extensive exploratory analysis in order to extract useful insights for the structure of the dataset and acquire field knowledge that will enable us to construct accurate classification methods.

3.1 Outliers detection

The first step before classification is to detect possible outliers that will distract the classifier. A small number of flights had departure delays more than 1200 minutes. We managed to decrease the top

delay from around 1200 to 300 with only removing 1828 flights. We further elaborate on outliers removal by using the quantiles Q_1 and Q_3 to get the boundaries for each attribute as, $lowerboundary$ = Q_1 - $k(Q_3 - Q_1)$ and $upperboundary$ = Q_3 + $k(Q_3 - Q_1)$. Q_1 and Q_3 are the medians of the first half part of the array of values of an attribute and the second half part correspondingly and k is a hyperparameter for the boundaries normally ranging between 1.5 and 3.0. A version was also created where instead of the medians the mean values are used which is useful for binary features. This is useful for features that are binary. Since most attributes (such as month, day of week etc.) are features that can be straightforwardly be checked for whether they contain valid values, we put the emphasis on calculating distance and departure delay features.

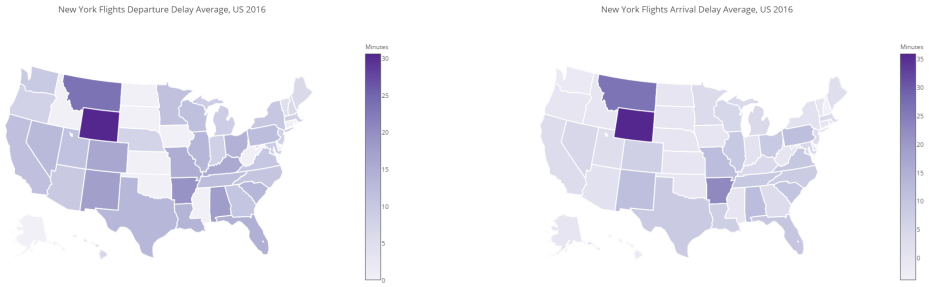
Feature	Lower bound	Upper bound
Departure delay	-107.63	132.35
Distance	-4493.0	7701.0

Table 1: Outliers boundaries.

Based on these reported boundaries (1), 2123 outliers with respect to departure delay and none for distance are found. It makes sense to neglect flights with departure delay of more than the higher boundary of the outlier system, as a departure delay of larger than approximately 132 minutes is always going to be late. Another really interesting fact is that 100% of all outliers are being classified as delayed to depart which makes sense since all of these outliers have already a very large departure delay and it would be impossible to have arrived on time to avoid the arrival delay. This highlights the importance of departure delay for out task.

3.2 Feature Extraction

To begin with, we would like to determine the importance of each factor to the classification task and present the most important of them. So, we dive into the categorical and numerical features and try to find whether some are indicators of delayed flights. Firstly, we identify the importance of spatial distribution of the flights.



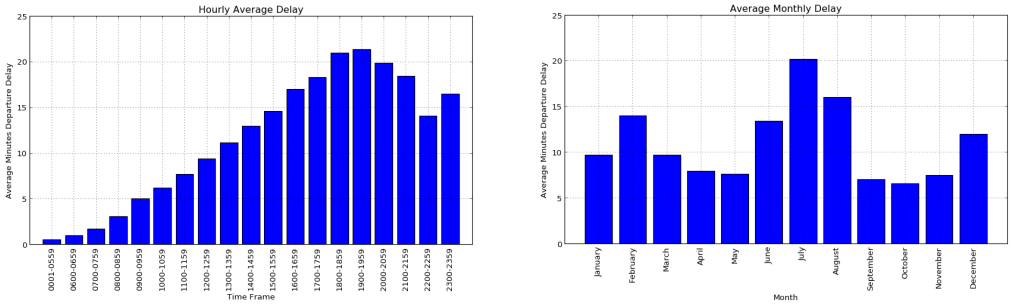
(a) Departure Map (b) Arrival Map

Figure 1: Average flight delay of each US state.

The figure 1 shows that the state of origin of a flight going to New York or the destination of a flight originating from New York is actually important. States like Arkansas and Wyoming seem to have the highest delays when traveling to New York or when a flight from New York arrives.

Continuing with time analysis, we find that there is a really smooth distribution of the hours intervals with respect to delays (2). Confirming literature [2, 1] we spot that delays do actually propagate. The delay during the afternoon seems to always be affected by earlier delays which force flights to leave later. The month 2 also adds to the analysis. Summer months tend to have significantly more delays. Additionally, we must mention that week day and day of month seems to affect the average delay. Delays are more probable in the middle of the month and during Thursday and Friday.

A good approach to the problem is to investigate the role of holidays into the problem 4. In order to do that, we choose the 12 biggest US holidays (such as Independence Day, Thanksgiving, Christmas

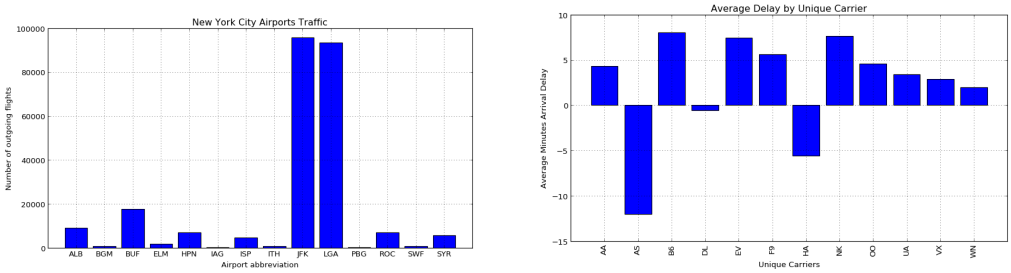


(a) Average hourly delay

(b) Average monthly delay

Figure 2: Time effect on delays.

and New Year). By observing the average delay, we notice an important fact. The days with most delayed flights are those before or after each holiday and during the each particular holiday. Looking into the volume of flights we also get the same results.



(a) NY airport traffic volume.

(b) Average delay per carrier.

Figure 3: Airport and carrier information.

Last but not least, by plotting the average delay with respect to the carrier 3 we can spot different policies. AS, HA and DL are always before time which means that they predict a longer flight time that originally needed, This policy leads to having zero delays and maybe enhancing their good image to the public. Also, the analysis show that only two (namely John F. Kennedy International Airport and LaGuardia Airport) concentrate almost the whole volume of air traffic originating or heading to NY state.

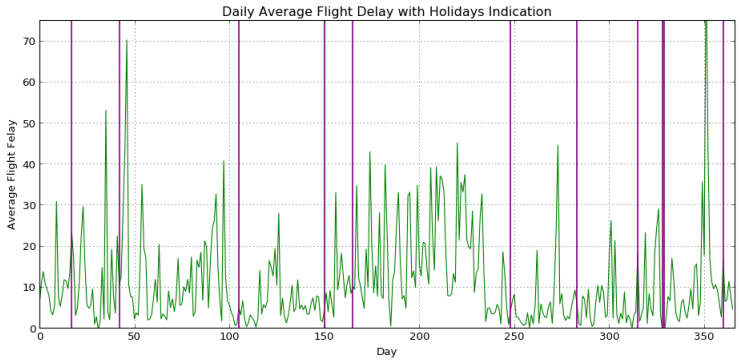


Figure 4: Daily average flight delay with holiday indications (purple lines).

4 On-time Prediction of Commercial Air Travel

The first goal of this paper is to predict whether or not a flight will be delayed. As it was mentioned in section 2, the acquired dataset refers to all the flights, on-going and off-going, from New York state. To solve this task, a binary classification approach was made. First, a description is given on the feature creation based on the explanatory analysis of the previous section. At the same time, a brief discussion on how the data was further processed for the particular task is being made. The section ends with the report of the findings and some important conclusions.

4.1 Methodology

The actual interest of the challenge lies in the ability to accurately predict whether or not a flight will be delayed. In order to achieve this we utilize the insight gained from the exploratory analysis to transform our features or create new ones. We prepare the dataset by performing one-hot-encoding to all the categorical features that we are interested in. Using the information provided from the previous analysis, we use the time that the flight departs separated in classes of one hour. Information for the day of week, day of month and actual month is also incorporated as it is realized that specific days of the week and the month or even months (i.e. summer months) have different delay distributions. Utilizing the result of the maps 1, the state of origin is also added to the features. Further more, the carrier, the departure delay group and the distance group are one-hot-encoded. During this procedure, one of the classes of each encoded feature is dropped to avoid correlation of the features, as one column is fully defined by the rest. Lastly, the actual numerical distance of the flight is added. The features are enhanced by performing feature engineering. Specifically, with respect to 4, we consider helpful to construct 4 holiday related features: one binary of whether it is a holiday or not, the distance from closest upcoming holiday, the distance from closest previous holiday and the minimum distance from any holiday.

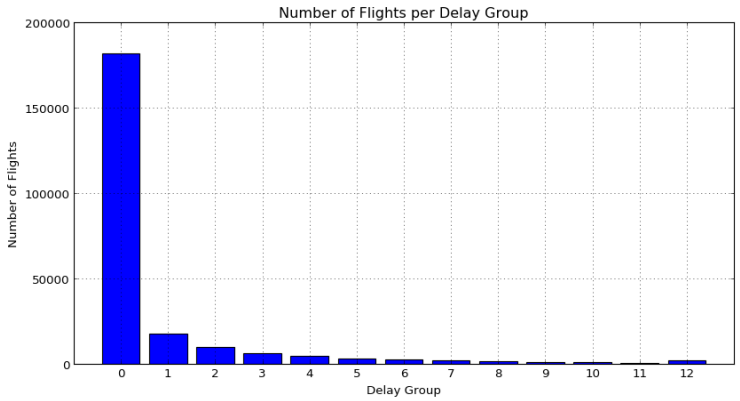


Figure 5: Flights per delay group distribution.

The size of the acquired preprocessed dataset to perform classification is constrained by the computational resources, as the vast majority of classifiers fail to produce results in a logical computation time. Hence, a sub-sampling procedure is used to reduce the dataset into a manageable size for our main task. The distribution of classes 5 indicates that the classes are highly unbalanced, as more than 80% of the flights arrive on time. In order to avoid low accuracy due to the imbalance and be able to process the dataset, we split the data into two groups: the on-going flights to New York an the out-going from New York. This splits the dataset into an almost equal number of flights of about 180 thousand instances. Only 50 of the 180 thousands belong to the classes that actually were delayed as the flights group delay distribution of figure 5 presents. Then, the non-delayed class is subsampled in order to make the delayed class account for the 30% of the total instances. Finally, in order to be able to actually investigate the generalization of our models, the final datasets for the classification tasks were split into 70% training set, 15% validation set and 15% test set respectively.

4.2 Results and Discussion

The approach that was taken to tackle the binary classification problem of predicting whether a flight will be delayed or not, was to experiment with a set of available classifiers and keep the one that produced the best performance on the validation set. To realize this, the off-going flights subset was used for this first experiment. Table 2 presents the classifiers that were used in this section and the results for a set of metrics that were used for evaluation. Besides the classification accuracy and the log loss, Cohen's kappa metric was used as supportive evidence for the decision of the classifier. This metric was particularly chosen because it is good indicator and robust measure for classification tasks with imbalanced classes since it takes into account whether a prediction was occurred by chance.

From the evaluation metrics of table 2, it can be observed that the logistic regression classifier shows the best performance. It particular, the classifier achieved the 85.5% accuracy, 0.342 log loss and 0.684 kappa score. These results, however, is only a first indication and insight that logistic regression classifier behaves well for the binary classification task and we need to conduct more experiments be certain on the results.

Classifier	Accuracy (%)	Log-loss	Kappa score
Logistic Regression	85.5	0.342	0.684
k Nearest Neighbors	76.1	0.546	0.675
Decision Tree	85.0	0.600	0.514
Random Forest	83.3	0.403	0.609
Multi-layer Neural Net	83.4	0.961	0.620
Gaussian Naive Bayes	64.2	12.328	0.364
LDA	85.4	0.363	0.689
QDA	67.3	11.199	0.075

Table 2: Evaluation of classifiers on validation set of off-going flights.

To gain a more robust measure of the accuracy over the dataset, a k-fold cross validation with 3 random shuffled folds was implemented. At the same time, to be more certain that the model did not over-fit and generalizes broadly, L2 regularization was incorporated. Along with cross validation, we experiment with different values of the L2 penalty hyper-parameter ranged from 10^{-4} to 10^{-1}. The experiment resulted in 85.5% accuracy with the best setting for the L2 penalty at 0.1.

Flights	Accuracy (%)	Log-loss	Precision	Recall	F1 score
Off-going	85.4	0.347	0.833	0.856	0.857
On-going	83.2	0.378	0.809	0.828	0.835

Table 3: Final model's results on test set.

Finally, it is sensible to observe hoe good the final model does on unseen data. Table 3 presents the results of the final logistic regression classifier on the test data. The model achieved 85.4% and 83.2% accuracy for the off-going and on-going flights respectively. That means that we are fairly accurate on predicting whether a flight will be delayed or not. The high precision and recall results of table 3 confirms this statement since we are not only able to predict accurately the flights that were actually delayed, but also to be confident on the model since the ratio of predictions that have falsely been predicted correctly is small.

5 Time Interval Delay Prediction of Commercial Air-Travel

5.1 Methodology

It is very useful information to be aware of whether a flight will be delayed or not and the binary classification task of the previous section predicted this fairly accurately. However, it is also extremely important to know *how much* a flight will be delayed. This would require to actually regress on the actual delay but intuitively this would be extremely difficult and would require far more extra data to make accurate predictions. For this reason, a different approach will be considered which will essentially solve the task in a similar way. The data set provide us with the actual flight delay grouped into 15 minutes intervals between 0 and > 180. By using the aforementioned time intervals

transformed into 13 delay classes, we will work on a multi-classification task which will produce the 15 minute interval in which a particular flight delay belongs to. This could show a good indication of *how much* a flight will be delayed in a similar way as predicting the actual delay time.

To realize this task and solve the aforementioned multi-class classification problem, the same features as described in section 3.2 will be used. These features led to fairly accurate predictions on the binary classification task of the previous section and thus, they will remain the same. In order, to reduce the larger number of instances of the data and to make more sensible hypothesis we split the data into two groups. The flights that arrived to New York and those that departed from it. We implement the same classification task for both upcoming and ongoing groups of flights. This splits the dataset into an almost equal number of flights of about 180 thousand instances. In addition, out of the 180 thousand flights only the 50 thousand belong to the classes that actually were delayed as the flights group delay distribution of figure 5 presents. To treat such a matter we subsample the imbalanced non-delayed class 0 and make it about 30% of the distribution of the number of the other classes together. Lastly, the final datasets for the classification tasks were split into 70% training set, 15% validation set and 15% test set.

5.2 Results and Discussion

The first set of experiments were conducted by training a variety of classifiers and evaluate their results on the validation set as table 4 presents. It is important to mention, that along with the accuracy and the log loss error, the f1 score is reported as supporting evidence on choosing the best algorithm since it is a good indication of the quality of the predictions. As it can be derived from table 4, the logistic regression classifier produced the highest accuracy at 55.1% and f1 score at 0.53 as well as the lowest log loss at 1.171 than all the other classifiers that we experimented with for the off-going New York flights.

Classifier	Accuracy (%)	Log-loss	F1 score
Logistic Regression	55.1	1.171	0.530
k Nearest Neighbors	41.7	5.583	0.357
Decision Tree	54.5	1.721	0.514
Random Forest	53.3	1.395	0.484
Multi-layer Neural Net	52.1	1.383	0.510
Gaussian Naive Bayes	41.0	12.843	0.323
LDA	54.6	1.599	0.526
QDA	4.7	32.580	0.047

Table 4: Evaluation of classifiers on validation set of off-going flights.

In order to gain a more robust insight and be more confident on our predictions, we implement a k-fold (3 random shuffled folds) cross-validation on the best aforementioned classifier for both ongoing and off-going from New York flights data sets. At the same time, we would like to avoid the possibility of over-fitting by regularizing our model with L2 regularization and fine tune the regularization penalty. This experiment resulted in 55.3% for the off-going and 53.2% for the on-going flights average accuracy on the validation set with the best value of L2 penalty hyper-parameter at 0.001. Last but not least, we use our final best model and report our results on the test set which are summarized on table 5.

Flights	Accuracy (%)	Log-loss	Precision	Recall	F1 score
Off-going	56.0	1.198	0.426	0.386	0.531
On-going	52.2	1.301	0.373	0.332	0.491

Table 5: Final model's results on test set.

The final logistic regression classifier achieved 56% and 52.2% accuracy for the off-going and on-going flights respectively. Figure 6 shows the confusion matrix for the final model, which presents the predicted relationship between the different classes. In particular, we could observe that almost all of the classes are being confused, to some extend, mostly by their 'neighbor' classes which are the neighbor time intervals. Finally, another interesting fact that we could derive from the confusion

matrix is that the two time distant classes, that is $0 - 15$ and over 180 minutes delay, could be predicted pretty accurately with probabilities of over 90%.

Figure 6: Confusion matrix of the final model.

6 Conclusions

The present paper attempted to predict the delayed flights. A vast dataset is carefully preprocessed and cleaned, before being used to provide useful insights on the factors that determine whether a flight is delayed or not. The importance of the time of departure, day of week, day of month, carrier and airport are highlighted with plots. Extensive work on outliers detection and feature engineering has been performed. Following, we managed to achieve accuracy as high as 85.4% for the on-time performance task in the test set and 56% for predicting the delay group divided in 15 minutes intervals. For both cases logistic regression outperformed all other classifiers that were tested. Overall, a trustworthy prediction method was proposed for both tasks.

6.1 Future work

The project performed an extensive exploratory analysis on the factors affecting commercial flight delays and managed to achieve high predictive power for both the binary and the multi-class cases. Future work could include analysis on more data, resulting in less filtering, provided the availability of computational resources. The weather data should be incorporated to the dataset, as literature points out weather conditions to be the most accurate and robust predictor of flight delays.

References

[1] Michael Nolan. *Fundamentals of air traffic control*. Cengage learning, 2010.

[2] Shervin AhmadBeygi, Amy Cohn, Yihan Guan, and Peter Belobaba. Analysis of the potential for delay propagation in passenger airline networks. *Journal of air transport management*, 14(5):221–236, 2008.

[3] Chamath Malinda Ariyawansa and Achala Chathuranga Aponso. Review on state of art data mining and machine learning techniques for intelligent airport systems. In *Information Management (ICIM), 2016 2nd International Conference on*, pages 134–138. IEEE, 2016.

[4] RaJ Bandyopadhyay and Rafael Guerrero. Predicting airline delays. 2012.

[5] Juan Jose Rebollo and Hamsa Balakrishnan. Characterization and prediction of air traffic delays. *Transportation research part C: Emerging technologies*, 44:231–241, 2014.

[6] Sina Khanmohammadi, Salih Tutun, and Yunus Kucuk. A new multilevel input layer artificial neural network for predicting flight delays at jfk airport. *Procedia Computer Science*, 95:237 – 244, 2016.

[7] https://www.transtats.bts.gov/Fields.asp?Table_ID=236. Bureau of transformation statistics. 2017.

Work Planing and Contributions:
The main bulk of the work was divided into three parts and was implemented equally by every member of the team. In particular, we begun by downloading, merging and preprocessing the raw data. Then, cleaning (outliers detection) and exploratory analysis was performed with visualizations in order to get meaningful insights on the data. Feature construction and classification tasks were also equally divided and implemented by each of the group members. Finally, each individual work was reported and merged to a unified style report with contributions by all three members.

Bonus fun fact:
By plotting the volume of the flights daily for 2016 we observe high periodicity. The volume patterns for each week seem highly correlated. However, we spotted a great decline for January 22. The first thought questioning the existence of missing data, quickly was rejected, as we found that this day a blizzard affected the whole state! (source: https://en.wikipedia.org/wiki/January_2016_United_States_blizzard)

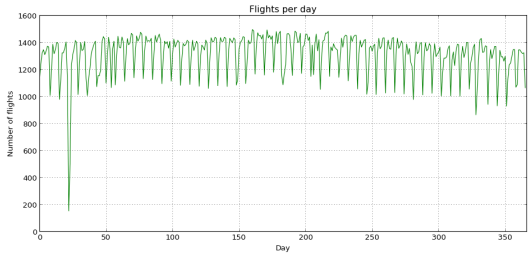


Figure 7: Number of flights per day.