Department of Computer Science
MSc Data Science and Analytics
Academic Year 2016-17

C5609 Learning Developing Project Coursework

1442467, 1639420, 1119638

April 23rd, 2017

Contents

1. Introduction

The cost of domestic flight delays is estimated at $33 billion (Ball et al., 2010), with about half that cost being borne by airline passengers. The operation of a passenger airline requires the allocation of resources and development of schedule plans over complex networks. The efficient management of such costly resources is one of the key challenges faced by airlines looking to control operating expenses and generate profits (AhmadBeygi et al., 2008).

Needless to say, analysis leading to any reduction in delays is of immense potential interest to both passengers and airlines.

Our focus on this project has been on arrival delays as this has a much bigger impact on customer satisfaction.

2. Literature Review

In their paper on departure delays Yufeng et al. (2008) argue that many existing decision support tools for air traffic flow management take a deterministic rather than a stochastic approach to problem solving. The Federal Aviation Administration for example have used a module known as 'Monitor Alert' to estimate departure time based on the flight's scheduled departure time. This deterministic approach fails to capture the stochastic factors such as uncertainty in a flight's departure time, changes in a flight's route and the impact of queuing at departure. They argue that it is more informative to provide estimates of the entire distribution rather than simple point estimates, which can be misleading.

Yufeng et al. (2008) also state that the delay propagation effect, caused by the cumulative impact of earlier delays is a key reason for delays. Other factors cited include weather, demand surges, mechanical failure and congestion. Schaefer et al. (2001) discuss the impact of inclement weather in much more detail and claim that the main reason for delay is not the bad weather per se but the reduced airport capacity arising from increased aircraft separations (distance between aircraft) during inclement weather. They go on to make a clear distinction between two situations, one known as Instrument Meteorological Conditions (IMC) in which instrument landing systems are required for aircraft navigation and another known as Visual Meteorological Conditions (VMC) in which they are not required. Based on their analysis at three different airports they conclude that IMC results in a drop in capacity of around 50%, depending on the runway geometry and visibility. They conclude delays are not significant when VMC exists at all airports and that the propagation effect is unobservable in airports with a high capacity-to-demand ratio. Locally, delay increases with increasing duration of IMC and propagated effects are significant for the 1st leg after leaving an IMC airport and diminish from leg to leg through to the 5th leg.

AhmadBeygi et al. (2008) state that keeping aircraft and crews together significantly reduces delay propagation, but is insufficient to avoid it altogether. Furthermore, they claim that contrary to the commonly held assumption that the propagation of a delay in a flight network has a greater impact than the root delay itself that based on their analysis many flights do not propagate root delays and even with root delays of up to 180 min, nearly 40% of flights have no propagating effect. They also make an interesting point about incorporating slack into the schedule claiming that all things being equal, the optimal location for the slack is in the middle of the chain. This is where the expected delay is minimised, because it is the trade-off point between the length of the propagation and the probability of the root delay.

Yufeng et al. (2008) also identify a seasonal trend with higher delays in winter and summer and fewer delays in spring and autumn. This is a hypothesis that we would like to test. By using a smoothing spline, time can be treated as continuous factor, which makes sense because the delay at the end of one month will not vary significantly from the delay at the beginning of the next month. A similar argument can be applied to fluctuation in delay over the course of one entire day, making the smoothing spline also an advantageous approach for addressing the daily propagation effect.

In summary, it would be interesting investigate stochastic approaches to flight delay, investigate the impact of inclement weather and finally cyclical trends such as seasonality and trends through the course of the day.

Data Analysis Approach

John Tukey developed a statistical approach known as Exploratory data analysis (EDA) in which graphical methods and summary statistics are used to explore data and gain insight as opposed to answering specific questions (Barsalou, 2015). Brillinger states that Tukey recognised two types of data analysis: EDA together with confirmatory data analysis (CDA). In EDA, the data is sacred and the objective is to look for unexpected patterns in data in order to formulate hypotheses, whereas in CDA the model is sacred and one is looking to confirm or reject specific hypotheses (Brillinger, 2002). Given that this assignment is using a fresh dataset about which little is known in advance it makes sense to resist the temptation to move straight the analysis phases without first carrying-out EDA in order to better understand the data and identify hypotheses.

3 Project Management

Project Management Approach

Project management involves the application of knowledge, skills, tools, and techniques to project activities to meet requirements (The Project Management Institute, 2014). A project is always subject to constraints such as time, scope and resources, all of which must be managed to deliver on time and meet the requirements. This project was no different in many respects, but is unusual in that the requirements cannot be clearly identified until some initial data analysis has been undertaken. For this reason, a flexible iterative approach is illustrated by the project management cycle in figure 3.1 below. Project planning and monitoring are undertaken in light of the triple constraints of time, quality and scope.

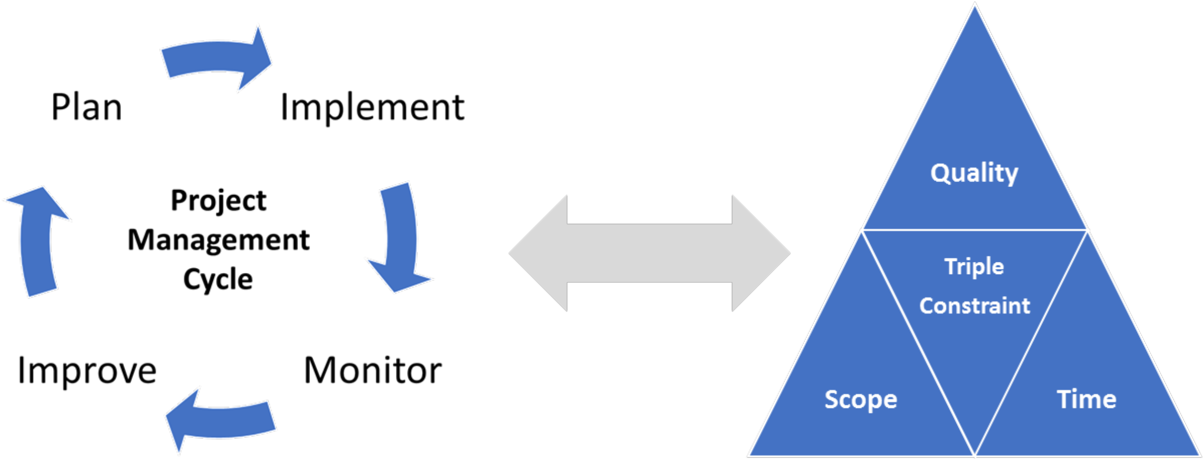


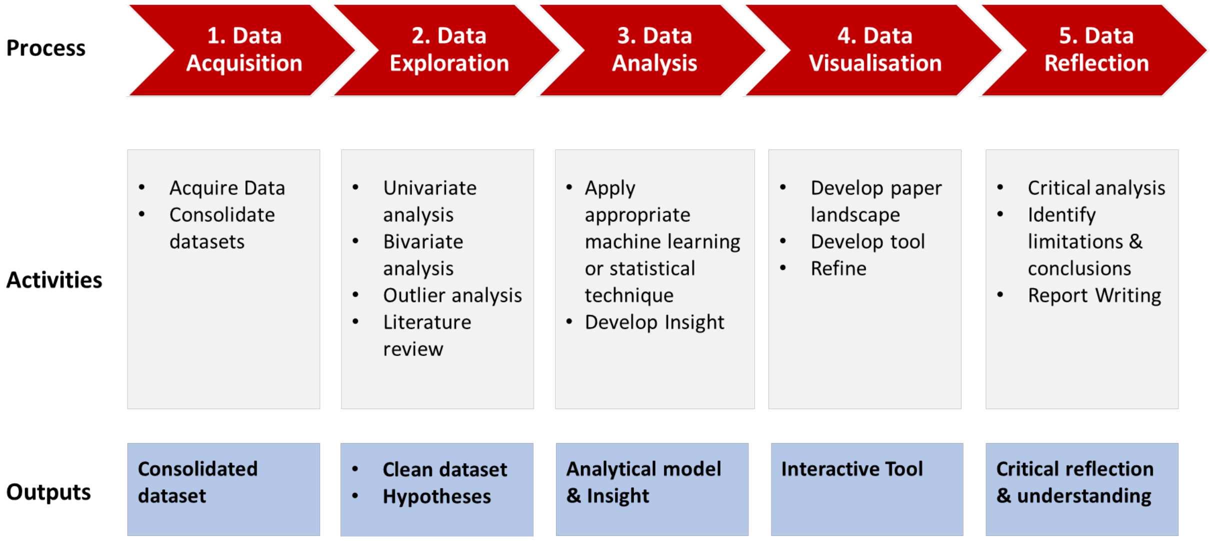
Figure 3.1 Project Management Approach and Constraints

Planning

1) **Process Mapping**. The data analysis process was mapped out and is illustrated in figure 3.2 below. This was incorporated into a broader project plan.
2) **Requirements identification**. This which was not possible until the literature review and initial data analysis had been undertaken.
3) **Assigning roles to team members**. Given the time constraints and other demands on people's time it was important to assign different roles to different team members and to work on tasks in parallel to aid efficiency. The three major roles included project management, data analysis and visualisation.
4) **Coordinating Literature review**. This involved identifying and reviewing the relevant literature.
5) **Tool Selection**. The selection was based on experience working with tools.
   a) Exploratory data analysis: R and Python.
   b) Data processing: Python.
   c) Model Development: SAS Enterprise Miner.
   d) Visualisation: Tableau.

Implementation

The five-step data analysis process outlined in figure 3.2 below was used to structure the implementation. A more detailed discussion on each of these elements is provided in latter sections of the report.



Process	1. Data Acquisition	2. Data Exploration	3. Data Analysis	4. Data Visualisation	5. Data Reflection
Activities	• Acquire Data • Consolidate datasets	• Univariate analysis • Bivariate analysis • Outlier analysis • Literature review	• Apply appropriate machine learning or statistical technique • Develop Insight	• Develop paper landscape • Develop tool • Refine	• Critical analysis • Identify limitations & conclusions • Report Writing
Outputs	Consolidated dataset	• Clean dataset • Hypotheses	Analytical model & Insight	Interactive Tool	Critical reflection & understanding

Figure 3.2: The data analysis process used to structure the project

Monitoring and Improvement

Constant monitoring of the progress of the project was required in order to ensure that:

- Deliverables were on track
- Roadblocks were identified
- That the team was working in a cohesive way
- Analysis was consistent with any conclusions drawn from the literature review
- The scope of the project was realistic

It should be pointed-out that monitoring was very much an active process with tangible consequences. There are three key examples where decisions were taken that had a major impact on the direction of the project. First, fairly early on in the project it be became clear that a team of six was going to be too large to manage effectively and the decision was made to split into two teams of three, which unfortunately meant that we started work later than other groups than we would have liked. Second, more specific roles were assigned to different team members to avoid duplication of effort and ensure effective time management. Third, it became clear that the scope of the problem needed to be reduced given the size of the initial dataset and limited amount of time. This was achieved by focusing on just two years: 2007 and 2008, and only on flights going between ten major airports in the USA.

Project Requirements

Key requirements were identified as follows:

1) To provide insight into flight delays based on:
    a) Literature review
    b) Exploratory data analysis.
    c) Formulated hypotheses
    d) Development of predictive model.
2) To develop an interactive visualisation to explore and provide insight on delays.
3) To effectively evaluate the solution and provide a coherent report, detailing the approach.

4 Data Acquisition

The objective of the project was to undertake analysis on flight delays in order to better understand the causes of delay. The following data sources were used during the course of this project.

1) **Flight data:** Data for domestic flights during 1987 until 2008 in the US were downloaded from the Statistical Computing webiste[1], from the American Statistical Association (ASA). It was later decided that only datasets pertaining to 2007 and 2008 should be used for analysis in order to narrow the scope.
2) **Weather data:** Daily weather data for the ten selected airports were scraped from weather underground website[2], who are a commercial organisation providing real-time weather data. The data was accessed using a bespoke Python script, which is included in the appendix (section 9.4). The dataset includes variables describing mean visibility, mean temperature, mean wind speed, and special weather conditions such as fog or thunder.

Details of how the flight and weather datasets were consolidated is provided in the analysis section 9.1. A data dictionary is also provided in section 10 of the appendices.

Scope

The datasets for 2007 and 2008 each contained about 7.5 million records, over 300 airports and over 5,000 routes (origin and destinations). Such a large and complex dataset presented two main challenges, firstly the sheer size of the dataset presented some manageability challenges and secondly the complexity generated by the large number of airports.

In order to make better use of the time available the decision was made to reduce the scope, by focusing only on key airports. It should be noted about 20% of airports accounted for 80% of flights. Flights between ten airports were selected, which reduced the size of the dataset to around 350 thousand records per year (about 5% of the original size). Eight of these airports were selected on the basis because they accounted for a large portion of flights. An additional two airports (Honolulu, Hawaii and Anchorage, Alaska) were selected as they corresponded to a disproportionate number of delays which it was felt warranted further investigation.

The final weather dataset contained 7,310 records, with no missing values.

5 Data Exploration and Pre-Processing

It is important to resist the temptation to move straight for data acquisition to data analysis without first gaining a deeper understanding of your data. This is also consistent with the Tukey's rationale for undertaking exploratory data analysis as discussed in the literature review.

Data exploration will involve two steps:

1) Univariate Analysis – to understand the structure of the data, understand integrity issues and identify outliers
2) Bivariate analysis – correlations and trends

5.1 Univariate Analysis

Key Findings

Issue	Details	Action
Data Integrity	Air time greater than journey time	Set air time to null
Data Integrity	Some records in the dataset relate to cancellations	Exclude from main analyses
Data Integrity	Some flights (0.2%) are diverted	Exclude from main analyses
Skewed Distribution	Arrival Delay positive right skew, with a large number of early arrivals	Investigate early arrivals Run analysis with early arrivals recoded as zero minutes late
Skewed Distribution	20% of airports (origins) account for 80% of flights.	Main focus of analysis to be on key airports. This will simplify the analysis.
Data Format	Time values have the wrong format. For example, '02:03' was written as '23'.	Use Python script to correct the format.
Data Format	*Year*, *Month* and *DayofMonth* are separate variables in the airline delay dataset.	Use Python script to combine them into a new variable 'Date'.
Analysis	More flights on Monday and fewer on Saturday otherwise the same	Compare delays on Monday and Saturday to delays on other days.
Assumptions	No reason for delay provided in 60% of cases where an arrival delay existed. Investigations have concluded that these records to not correspond to particular delay durations or airlines.	State Assumption that those flights where a delay incurred a reason provided are representative of all flights

Table 5.1 Key Findings from univariate analysis

5.2 Bivariate Analysis

Issue	Analysis
Delay Variation by Season	Delays are longer in summer and winter and shorter in spring and autumn. This is also consistent with analysis examined in the literature review
Delay Variation by Carrier	Significant variation by carrier
Delay Variation by Reason	A reason for delay was only stated in 40% of cases, which may mean that analysis is misrepresentative. No reason could be established as to no reason was provided in 60% of cases. A summary report is provided in table 5.1 below. Interestingly, weather was cited as being the delay in only 7% of cases, a finding that conflicted with analysis undertaken in the literature review.
Delay Variation by Origin	Huge variation in delay by origin, especially for airports in Alaska and Hawaii. This warrants further investigation.
Delay Variation by Time of Day	Spike in arrival & departure delays for those planes arriving betwenen2 and 5am - need to check for outliers

Table 5.2 Key findings from bivariate analysis

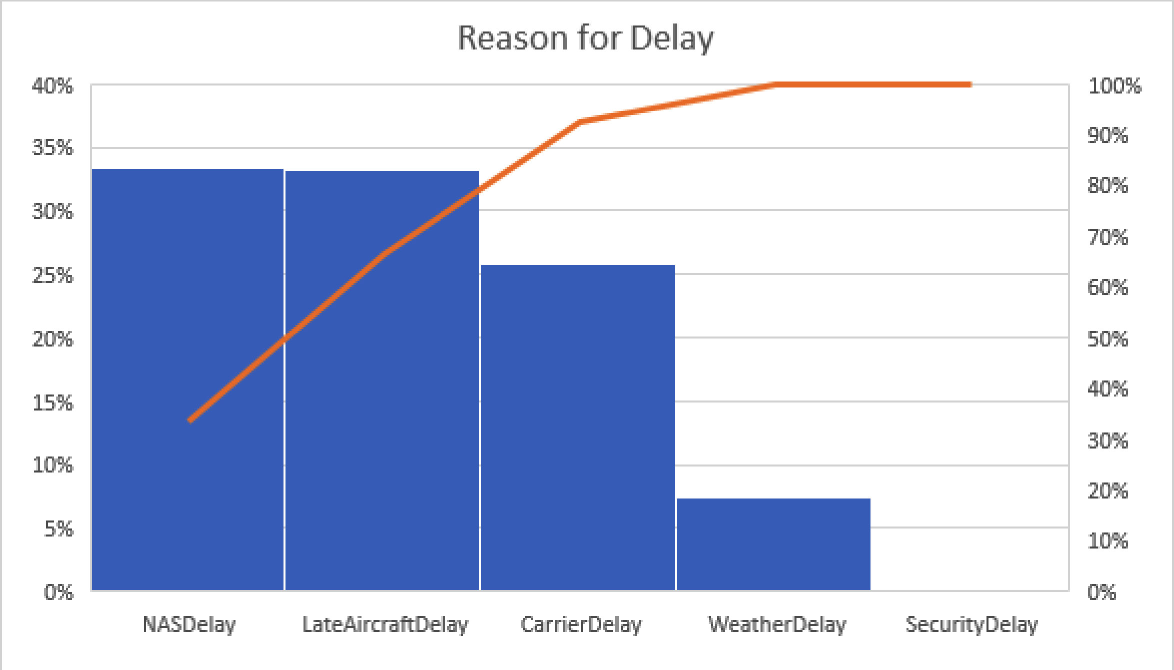


Figure 5.1 Delayed flights broken down by reason, where a reason is provided

6 Data Analysis

6.1 Data Preparation and Classification

In order to prepare the data for the classification models it was necessary to classify exactly what constituted a delay in binary terms. The average delay time for all the flights was 10 min, therefore, flights were classified as being delayed or on time based on the following criteria:

- Delay (=1): arrival delay > 10 min (67%)
- Non-delay (=0): arrival delay =< 10 min (33%)

Training, validation and testing dataset were created as per table 6.1 below.

Dataset	**Details**	**year**
Training Dataset	Random sample of 80% of data	2007
Validation Dataset	Random sample of 20% of data	2007
Testing Dataset	All the data, equivalent to approximately 700 thousand records	2008

Table 6.1 Details of training, validation and testing datasets

6.2 Analysis Summary

Step 1: Run predictive models for arrival delay based on the flight dataset

Analysis: Based on the ROC curve the regression model is statistically better than the decision tree. Therefore, the regression model will be adopted. Key steps were as follows:

1) Key variables: Departure delay, departure time, day of week, destination, origin, distance, month, carrier
2) Model types: decision tree, gradient boosting, regression
3) Review of prediction performance by analysing receiver operating characteristic (ROC) curves
4) Review of misclassification statistics
5) Review of event classification rates for
   a) False negative (FN)
   b) True negative (TN)
   c) False positive (FP)
   d) True positive(TP)
6) Review of assessment report, which details accuracy and precision
   a) Sensitivity – true positive rate or recall.
   b) Specificity - true negative rate
   c) Precision – positive predictive value
   d) Accuracy – measures the statistical bias or difference between measured and true value.

Step 2: Repetition of above, but without departure delay.

Analysis**:** Regression statistically better, but specificity still very low.

Step 3: Combined the weather data with the flight delays and undertook exploratory data analysis

Surprisingly, special weather conditions at the destination airport did not correlate with longer delays.

Step 4: Analysis using consolidated dataset using both destination and origin weather data.

Using destination weather data an ensemble (based on Gradient Boosting and Regression) identified accuracy of 58%

Using origin weather data an ensemble model, which was based on gradient boosting and regression identified accuracy: 72.4%. An option would be to use the data in the original dataset which claims weather delay only accounted for about 6.7% of the delays to explain the prediction result. A model built from Origin weather has higher accuracy than destination weather

Note that the ensemble model was used in order to run two related but different analytical models and then synthesise results into a single score or spread in order to improve the accuracy of predictive analytics. This is because a single model based on one data sample can have biases, high variability or outright inaccuracies that affect the reliability of its analytical findings. Combining different models or analysing multiple samples, helps to reduce the effects of those limitations and provide better information to business decision makers.

Step 5: Combined airline delay dataset with destination weather data

Key variables rejected: Mean visibility, Mean wind speed, Mean temperature, precipitation and all the binary variables. A regression accuracy of 69% was identified.

Step 6:   Combined airline delay dataset with origin weather data   Again, Mean visibility, Mean wind speed, Mean temperature, precipitation and all the binary variables.   Regression accuracy of 70.4% was obtained.   Using destination or origin weather results in similar accuracy.   In summary weather data, did not improve the model accuracy

7 Data Visualisation

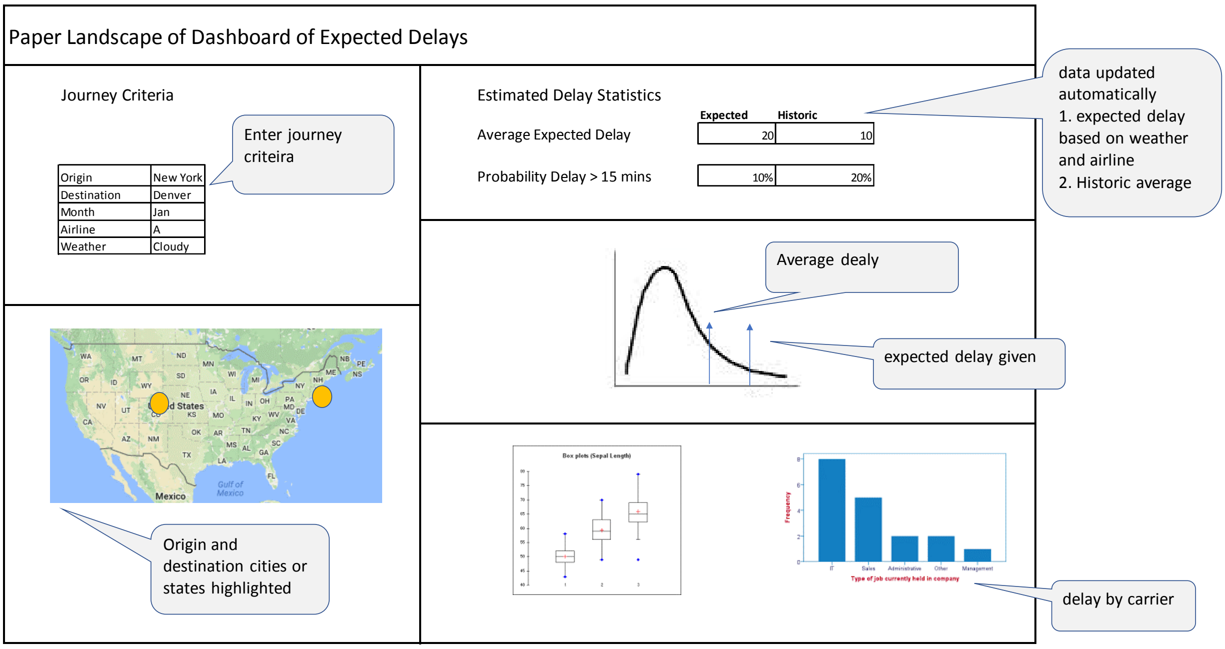
7.1 Data Visualisation

Data visualization is the method of consolidating data into one collective, illustrative graphic. (Bridgeable, 2017). “Tableau produces various visual artefacts to display data – charts, maps, dashboards, tables and so on. Tableau however means ‘visual analytics’ in a different way, and specifically it means that the visualization of data is achieved using a wholly visual interface.” (Butler Analytics, 2017)

The purpose of this visualisation is to be able to show a passenger which airline carriers would experience delays and in doing so will allow them to make a choice as to who they would like to travel with. In order to achieve this task we have created a story within Tableau containing two dashboards. The first dashboard contains information on Arrival Delays and the second contains information on Departure Delays. The layout and the visualisations for both dashboards have been kept the same so that it is easy for the user to navigate from the Arrival page to the Departure page.

7.2 Paper Landscape of Data Visualisation

Paper landscape of the visualisation is undertaken in order to map out the visualisation and to act as framework for discussion with other team members.



Paper Landscape of Dashboard of Expected Delays

Journey Criteria

Origin	New York
Destination	Denver
Month	Jan
Airline	A
Weather	Cloudy

Enter journey criteira

Estimated Delay Statistics

	Expected	Historic
Average Expected Delay	20	10
Probability Delay > 15 mins	10%	20%

data updated automatically
1. expected delay based on weather and airline
2. Historic average

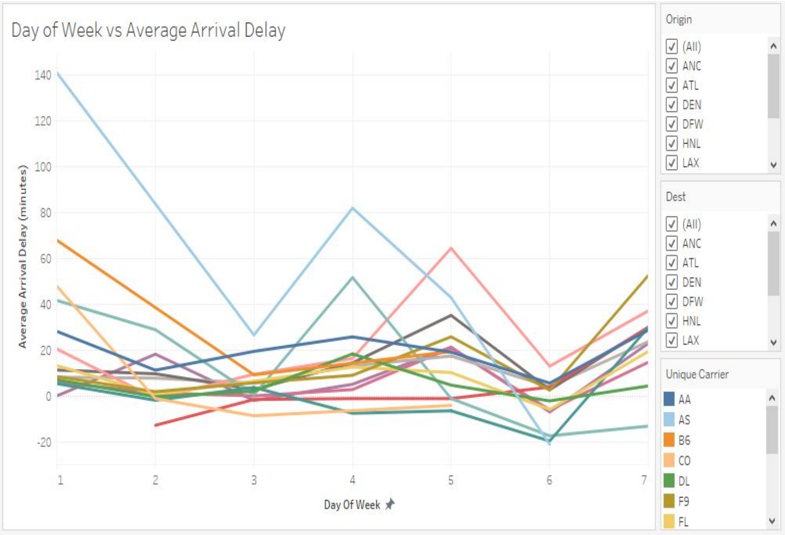
Average dealy

expected delay given

Origin and destination cities or states highlighted

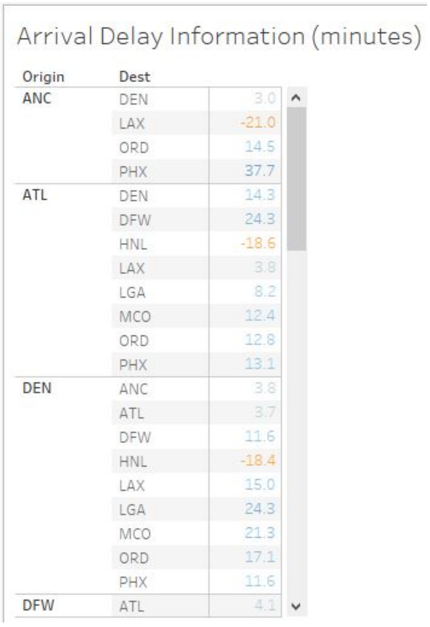
delay by carrier

7.3 Visualisation in Tableau



An interesting visualisation would be one that shows the best day to travel in order to avoid delays. By using the Marks tab we're able to look at delays across the week for each carrier. The reason why this visualisation is included, is so that the user is able to compare carriers to minimise their delay. As you can see the user is able to select an origin and destination and we are able to look at the average delays across the week for the relevant carriers.

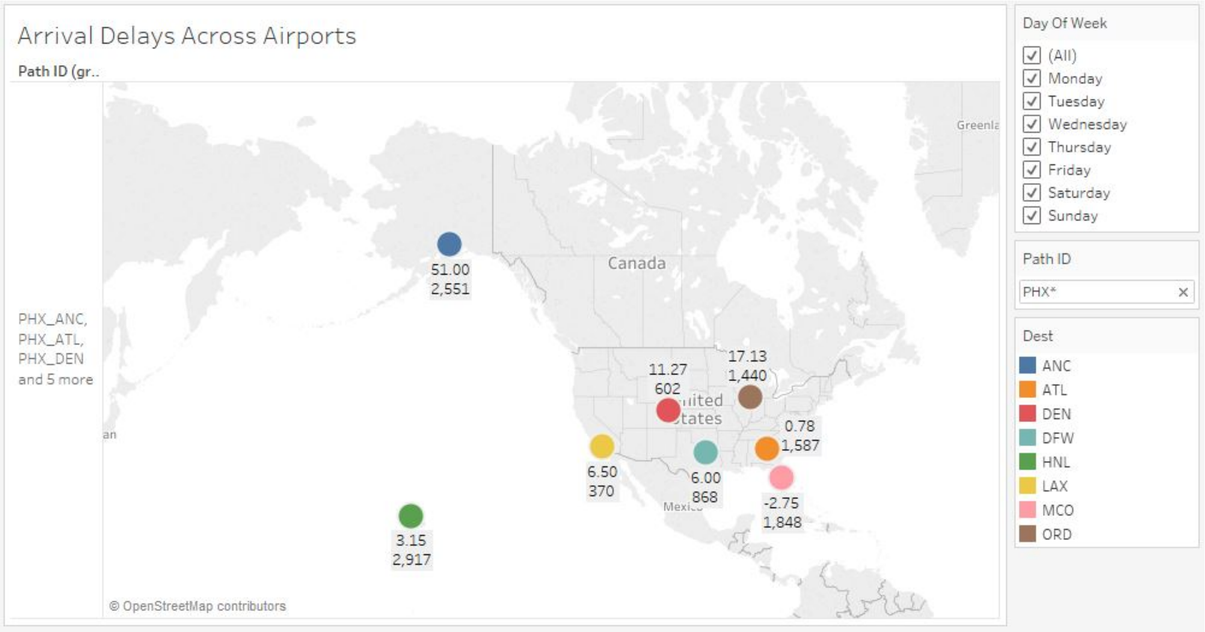
The Average Arrival Delay Information of each destination is also presented on the dashboard. There is a filter on each of the worksheets and this is updated based on the month of travel that the user decides to select. This data is presented with some conditional formatting as there are marks which are superimposed on the data which shows the colour of the text corresponds with the severity of the delay, for example, if the delay is severe then the colour of the text would be red, however if it were early it would be blue.



Arrival Delay Information (minutes)

Origin	Dest	
ANC	DEN	3.0
	LAX	-21.0
	ORD	14.5
	PHX	37.7
ATL	DEN	14.3
	DFW	24.3
	HNL	-18.6
	LAX	3.8
	LGA	8.2
	MCO	12.4
	ORD	12.8
	PHX	13.1
DEN	ANC	3.8
	ATL	3.7
	DFW	11.6
	HNL	-18.4
	LAX	15.0
	LGA	24.3
	MCO	21.3
	ORD	17.1
	PHX	11.6
DFW	ATL	4.1

The visualisation below shows the location of each of the airports that we have used for this analysis below. As mentioned above there is also a date filter applied to this visualisation so when the user selects the month June, all the data within the map is updated for the arrival delays for all of these locations.



Arrival Delays Across Airports

Day Of Week

- ☑ (All)
- ☑ Monday
- ☑ Tuesday
- ☑ Wednesday
- ☑ Thursday
- ☑ Friday
- ☑ Saturday
- ☑ Sunday

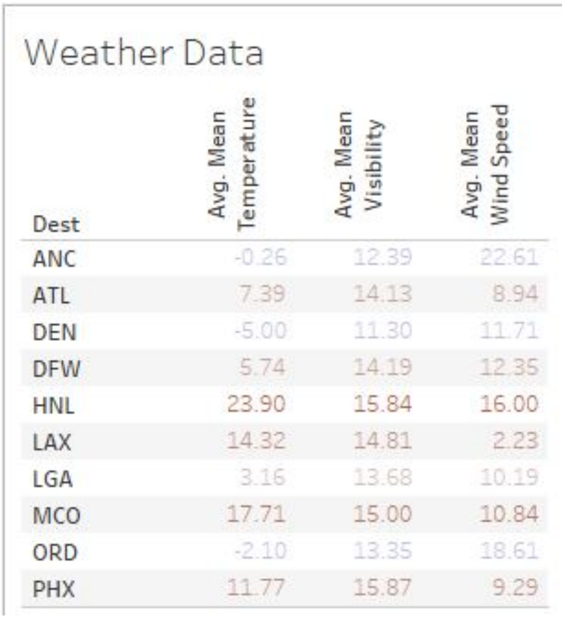
Path ID

PHX* ×

Dest

- ■ ANC
- ■ ATL
- ■ DEN
- ■ DFW
- ■ HNL
- ■ LAX
- ■ MCO
- ■ ORD

The user is required to enter in their origin into the Path ID box and all destinations are shown with the delay and the distance (in miles) from the origin to the destination. This allows the users to understand how much time to allow in terms of expected delay i.e. when a user wants to request for a taxi to wait for them at the airport.

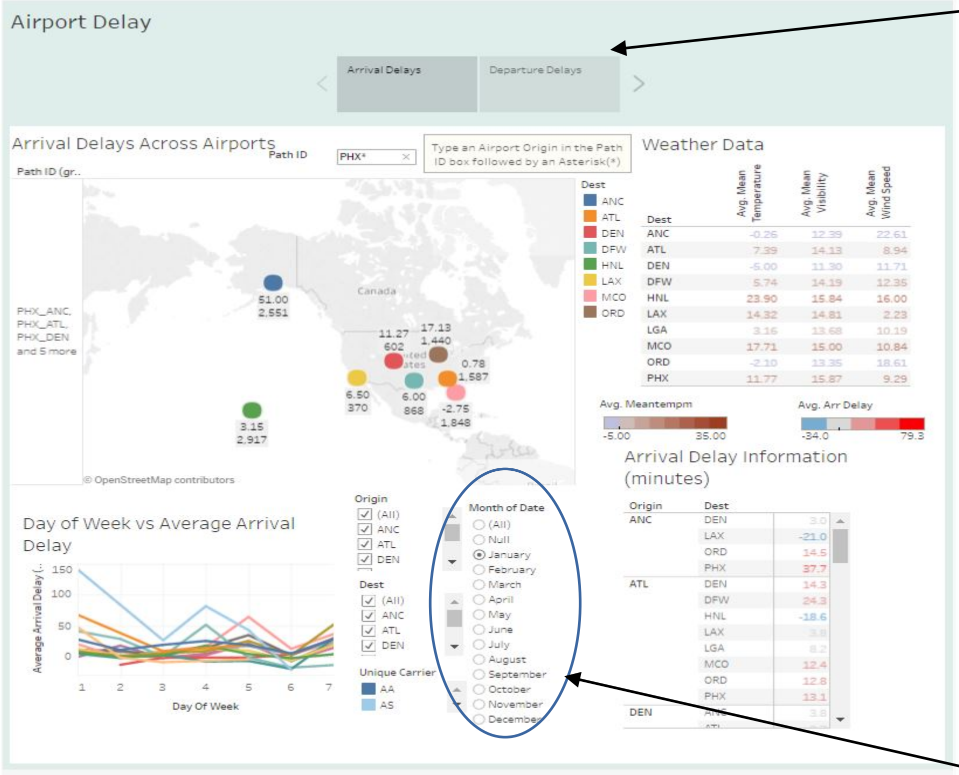


Weather Data

Dest	Avg. Mean Temperature	Avg. Mean Visibility	Avg. Mean Wind Speed
ANC	-0.26	12.39	22.61
ATL	7.39	14.13	8.94
DEN	-5.00	11.30	11.71
DFW	5.74	14.19	12.35
HNL	23.90	15.84	16.00
LAX	14.32	14.81	2.23
LGA	3.16	13.68	10.19
MCO	17.71	15.00	10.84
ORD	-2.10	13.35	18.61
PHX	11.77	15.87	9.29

The data was blended to weather data for each of the airports. Here we are able to produce a table with the average mean temperature, average mean visibility and average mean wind speed. This may help an individual understand which airline carriers may experience more of a delay than others in terms of priority given the weather data presented.

Dashboard: Arrival Delay



Origin
☑ (All)
☑ ANC
☑ ATL
☑ DEN

Dest
☑ (All)
☑ ANC
☑ ATL
☑ DEN

Unique Carrier
🟦 AA
🟦 AS

Month of Date
○ (All)
○ Null
● January
○ February
○ March
○ April
○ May
○ June
○ July
○ August
○ September
○ October
○ November
○ December

Weather Data

Dest	Avg. Mean Temperature	Avg. Mean Visibility	Avg. Mean Wind Speed
ANC	-0.26	12.39	22.61
ATL	7.39	14.13	8.94
DEN	-5.00	11.30	11.71
DFW	5.74	14.19	12.35
HNL	23.90	15.84	16.00
LAX	14.32	14.81	2.23
LGA	3.16	13.68	10.19
MCO	17.71	15.00	10.84
ORD	-2.10	13.35	18.61
PHX	11.77	15.87	9.29

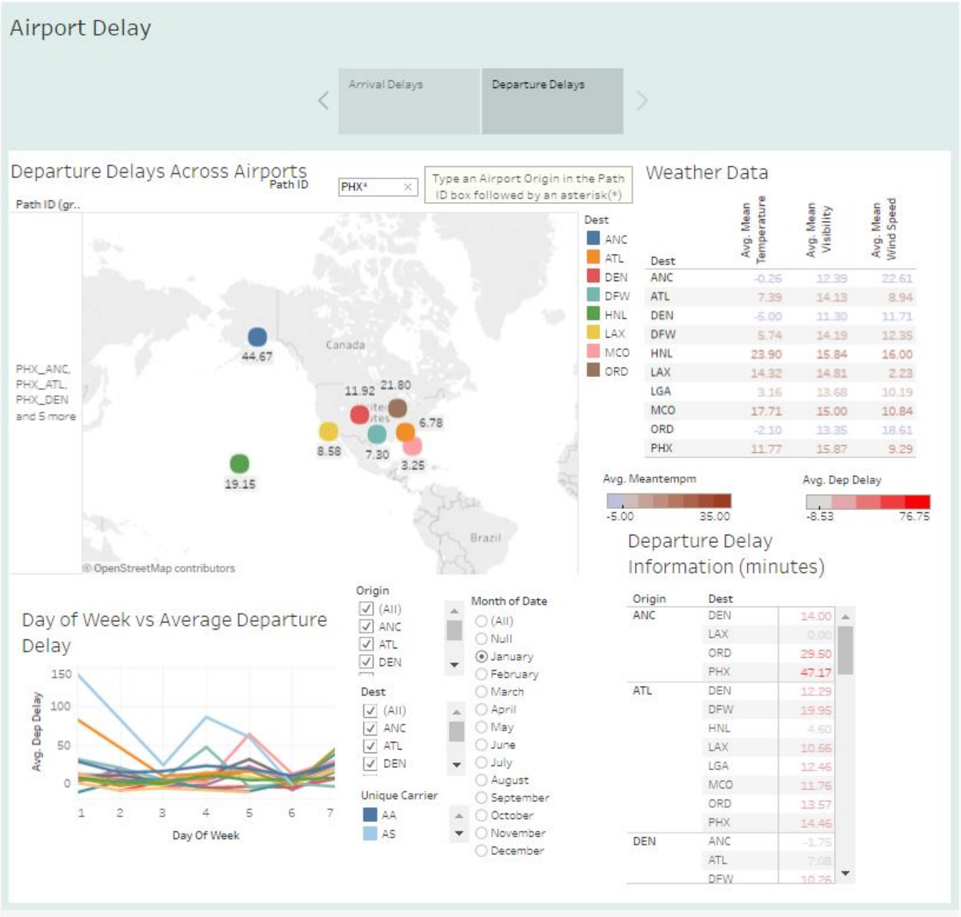
Arrival Delay Information (minutes)

Origin	Dest	
ANC	DEN	3.0
	LAX	-21.0
	ORD	14.5
	PHX	37.7
ATL	DEN	14.3
	DFW	24.3
	HNL	-18.6
	LAX	3.8
	LGA	8.2
	MCO	12.4
	ORD	12.8
	PHX	13.1
DEN	ANC	3.8
	ATL	[illegible]

Here we have used the "Story" element for this visualisation so that we are able to look at both arrival delays and departure delays for the chosen airports.

The Month of Date filter is applied to all worksheets within the dashboard. So the weather data table, the Delay Information Table, the day of the week vs. Arrival Delay graph and the map all contain information for January

Dashboard: Departure Delay



Origin
☑ (All)
☑ ANC
☑ ATL
☑ DEN

Dest
☑ (All)
☑ ANC
☑ ATL
☑ DEN

Unique Carrier
🟦 AA
🟦 AS

Month of Date
○ (All)
○ Null
● January
○ February
○ March
○ April
○ May
○ June
○ July
○ August
○ September
○ October
○ November
○ December

Weather Data

Dest	Avg. Mean Temperature	Avg. Mean Visibility	Avg. Mean Wind Speed
ANC	-0.26	12.39	22.61
ATL	7.39	14.13	8.94
DEN	-5.00	11.30	11.71
DFW	5.74	14.19	12.35
HNL	23.90	15.84	16.00
LAX	14.32	14.81	2.23
LGA	3.16	13.68	10.19
MCO	17.71	15.00	10.84
ORD	-2.10	13.35	18.61
PHX	11.77	15.87	9.29

Departure Delay Information (minutes)

Origin	Dest	
ANC	DEN	14.00
	LAX	0.00
	ORD	29.50
	PHX	47.17
ATL	DEN	12.29
	DFW	19.95
	HNL	4.60
	LAX	10.66
	LGA	12.46
	MCO	11.76
	ORD	13.57
	PHX	14.46
DEN	ANC	-1.75
	ATL	7.08
	DFW	10.26

7.4 Evaluation on visualisation

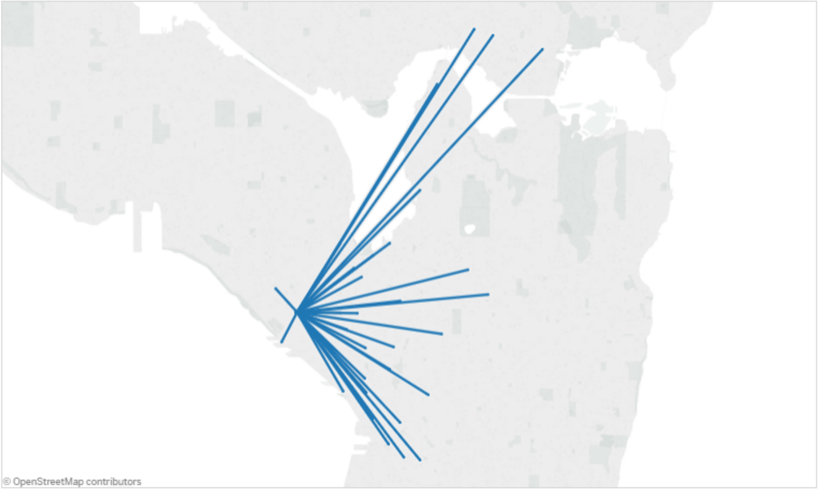
In order for this visualisation to have more value, more airports should have been used. This visualisation contains ten percent of the data used for this analysis. If I were to have done this assignment again it would have been sufficient to limit the number of rows of data for each carrier for each origin and destination pair. We could have performed some data aggregation so that more data could have been analysed and we would have been able to use this to create a more effective model containing data for most airports across the States of America. A reason for this would be if the user wanted to travel to an alternative airport, they would be able to see more options about getting to the same place.

An interesting feature that could have added onto the visualisation would have been the average price of the flights for the destination and origin pairs. This would allow customers to compare the average delay per carrier and the respective price per ticket.

Given the time constraints we were unable to show the map in a more effective way, it would have been useful to show Hawaii and Alaska closer to continental United States. This meant that in order to include Alaska within the visualisation we have had to zoom out of the map. This visualisation may cause issues for users that are visually challenged.

Initially we were trying to achieve a "spider map" for this visualisation, which in essence shows all the possible destinations you can travel to from any given origin as shown below. However this was not possible to do due to the nature of our dataset. In order for this to be achieved there needed to be single Origin-Destination pairs of data. Our data set included multiple Origin-Destination pairs which allowed us to show the averages of flight delays, which meant that there was a trade-off between the aforementioned average of flight delays and the visual shown below.

The goal was to get the user to only enter in the information at once and both dashboards are updated with this input. As both dashboards contained within the story contain similar information, both still require input from the user. A user-friendly way of achieving this was to get the user to enter this data on the "Arrival Delays" Dashboard and the "Departure Delays" Dashboard would also get updated.



8 Critical Reflection

1) Project management

Efficiency could have been improved to by agreeing a project manager at an earlier stage in order to provide more structure and avoid duplication of effort. One such example is the disproportionate amount of time was spent exploring and formulating problems at the very beginning. Once, a project manager and other time roles had been agreed progress was much more efficiently.

Face-to-face communication is important in the early stages of the project, while the team are learning how to work together. In the latter stages, it is more efficient to work independently. This finding is consistent with work carried-out by Bruce Tuckman on the four stages of team growth (Abudi, 2009).

Throughout the project, sharing and discussing individual working plans are as important as sharing and discussing individual working progresses and outcomes.

Documentary is important for keeping all the plans and results in record, and keeping all group members on track. We did well in this part since we kept making documentaries for all group meetings and project progresses.

2) Exploratory Data Analysis and Cleaning

Exploratory data analysis and data cleaning was key to the success of the project, and the work carried-out by John Tukey in this area is still relevant today. Once such example is the identification and correction of the time relevant variables in the flight dataset. It is important therefore that sufficient team is incorporated into the project plan for such tasks.

3) Data analysis and Insight

Correlations between different variables were measured and predictive models developed.

With/without departure delay

The predictive model was statistically better with departure delay data, indicating that departure delay correlated very well with arrival delays. If there was a departure delay, there would very likely be an arrival delay as well. This is perhaps an unsurprising finding.

Weather Conditions

It is commonly known that extreme weather conditions such as thunder or storm would usually lead to flight delays. However, surprisingly, the statistical evaluation of the predictive models based on weather alone and the consolidated dataset indicated that weather conditions did not play a crucial role in predicting arrival delays. This might due to the very few days of extreme weather throughout the year and extreme weather might only last for a few hours instead of for the whole day. Weather at origin is slightly more important in building predictive models for arrival delays than the weather data at the destination.

Classification

For the airline delay data, the distribution of number of instances in the two classes are biased (class 0: class 1 = 2:1), and this in turn might cause the machine learning output to be biased. For example, some outputs based on the Gradient Boosting or Random Forest algorithms have no predictive output for class 1. This could have been adjusted by changing the weights of the two classes.

4) **Suggestions for Further Work**

a) Applying the predictive models from the selected ten airports to all the airports and even on a global scale. This would allow us to compare how airports handle delays across multiple countries.

b) To repeat the analysis with a richer dataset with the following information
   - Data on priority assigned to different carriers. This could provide some information about NAS delays, which was the reason accounted for 33% of the delays in 2007.
   - Ticket price information. This would have facilitated analysis of any correlations between price and delay. In addition, if we want to extend the aim of our project to help customers choose the best flight – cheap price, fewer delays – in a specified airport at a specified season or even date, this information would be necessary.

c) Developing analysis that distinguishes between different classes of delay. Clearly a delay of 4 hours is much more serious that a delay of 15 minutes. It can also be hypothesised that the root causes are very different and this warrants further investigation.

d) Undertaking other assessments such as a more detailed analysis of residuals for more complicated analyses.

e) To extend our analysis to more airports and see how to optimise the analysis to make it universally applicable. For example, after the SAS training, we would be able to use SAS code to adjust the parameters or algorithms of the predictive models to make them more suitable for our problem.

f) Increasing the number of variables rather than looking at the typical factors concerning delays. Situations that occur on board prior to take-off that may not have been anticipated, such as health issues, security breaches or defects with the aircraft are recorded by airport staff that log this data onto a system.

g) Linking Twitter to the data set that could show potential factors contributing to airline delays in real-time. Most News Organisations have a Twitter account, by following certain organisations, there is scope to use this information to be able to determine delays, if for example, there are unforeseen weather conditions, or terrorist-linked activities, etc.

h) It would be useful to incorporate natural disaster data (earthquakes, hurricanes etc.) which can be used to predict weather conditions that may not usually be picked up by companies that produce weather forecast services.

9 Appendices

9.1 Details of Analysis

Step 1: Run predictive models for arrival delay based on the flight dataset.

Variables: Departure Delay, Departure Time, Day of Week, Destination, Origin, Distance, Month, Carrier

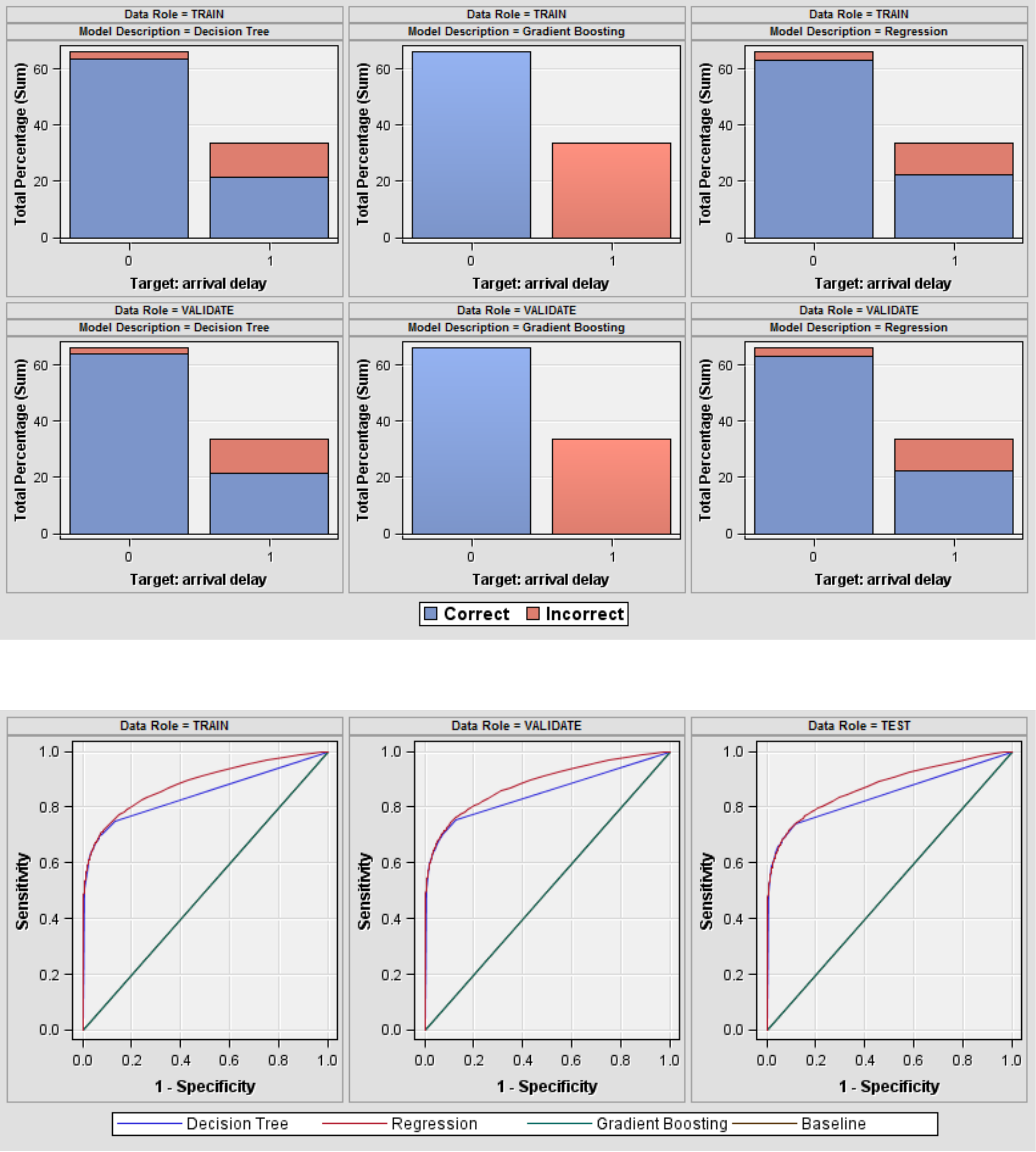


Figure 9.1 Misclassification chart and ROC curves for predictive models for arrival delay based on the flight dataset

Report:

Fit Statistics

Model Selection based on Valid: Misclassification Rate (_VMISC_)

Model	Valid: misclassification rate	Train: Average squared error	Train: misclassification rate	Valid: Average square error
Regression	0.14400	0.11057	0.14606	0.10960
Decision Tree	0.14493	0.11413	0.14549	0.11328
Gradient Boosting	0.33599	0.22310	0.33598	0.22310

Event Classification Table

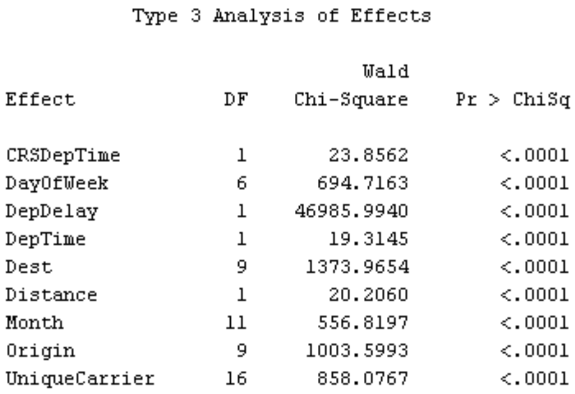
Model	Data Role	FN	TN	FP	TP
Decision tree	Train	33198	177794	7372	60494
Decision tree	Validate	8325	44514	1779	15099
Gradient boosting	Train	93692	185166	0	0
Gradient boosting	Validate	23424	46293	0	0
Regression	Train	31686	176121	9045	62006
Regression	Validate	7878	44132	2161	15546

Assessment report

Model	Data Role	Sensitivity	Specificity	Precision	Accuracy
Decision tree	Train	0.66	0.95	0.89	0.85
Decision tree	Validate	0.66	0.95	0.90	0.86
Regression	Train	0.66	0.95	0.87	0.85
Regression	Validate	0.66	0.95	0.88	0.86

ROC curve chart shows that the regression model is statistically better than the decision tree. Therefore, we adopt Regression model in this analysis.

Model summary:



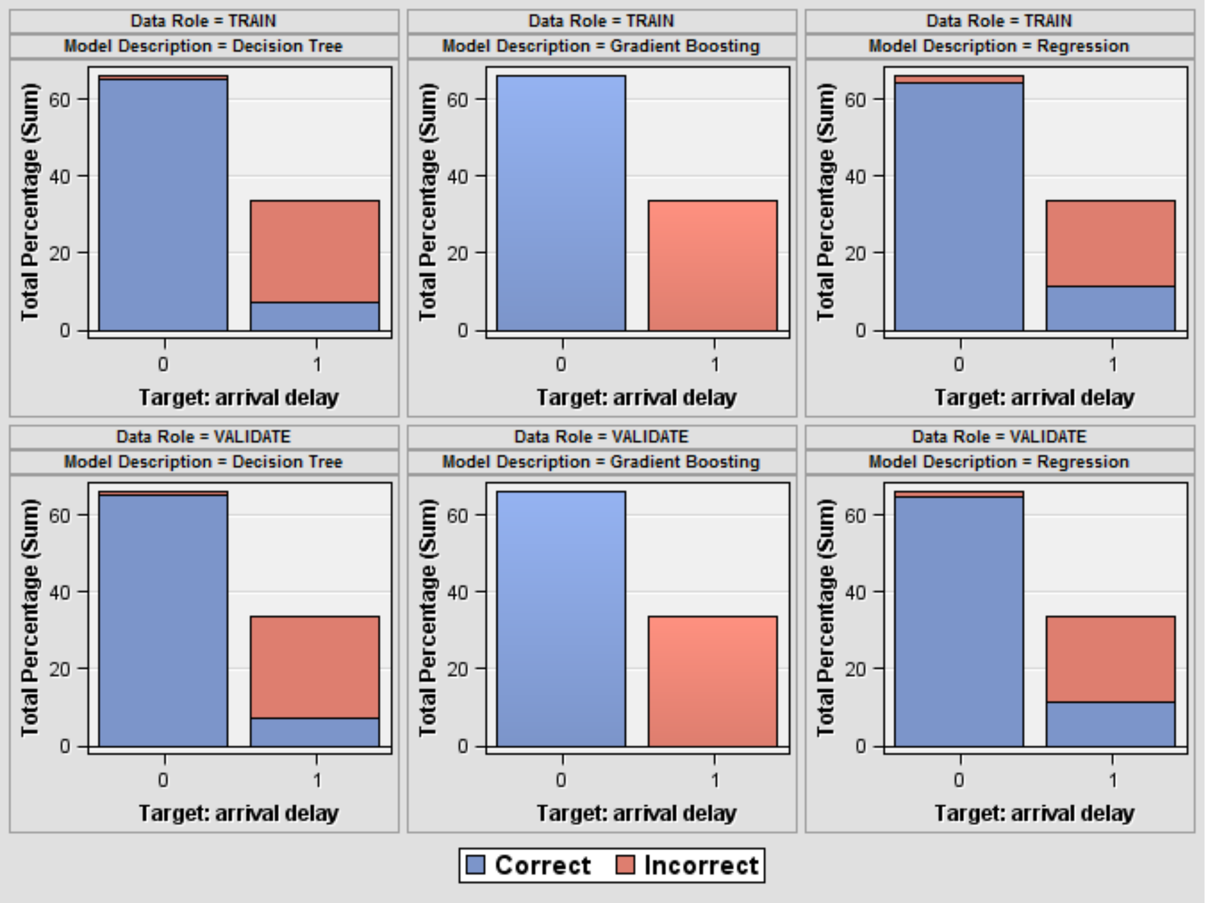
```
          Type 3 Analysis of Effects

                              Wald
Effect              DF    Chi-Square    Pr > ChiSq

CRSDepTime           1       23.8562        <.0001
DayOfWeek            6      694.7163        <.0001
DepDelay             1    46985.9940        <.0001
DepTime              1       19.3145        <.0001
Dest                 9     1373.9654        <.0001
Distance             1       20.2060        <.0001
Month               11      556.8197        <.0001
Origin               9     1003.5993        <.0001
UniqueCarrier       16      858.0767        <.0001
```

Step 2: Repetition of above, but without departure delay.

Misclassification chart



ROC curve

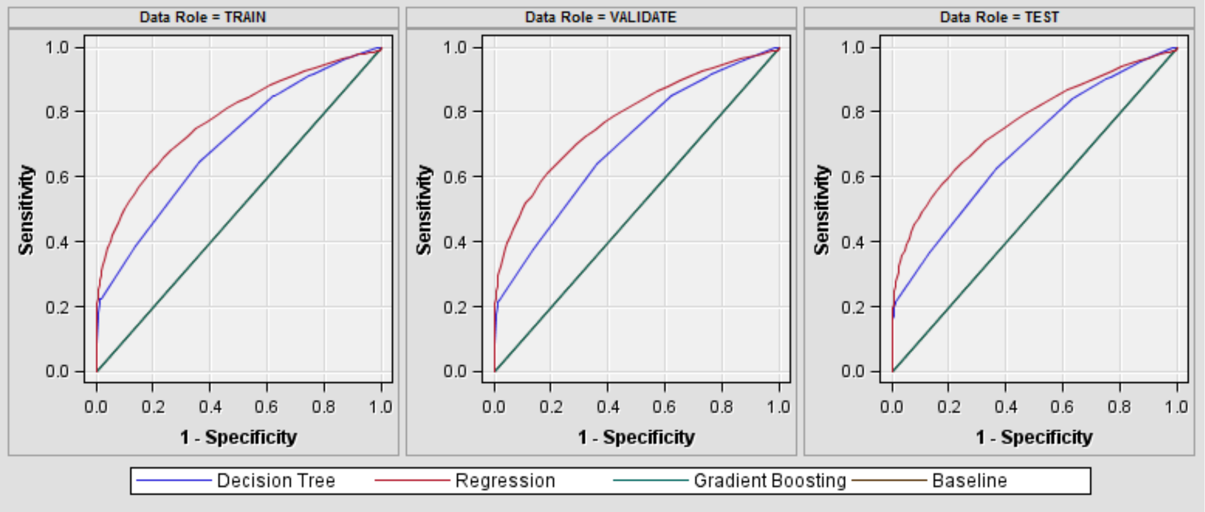
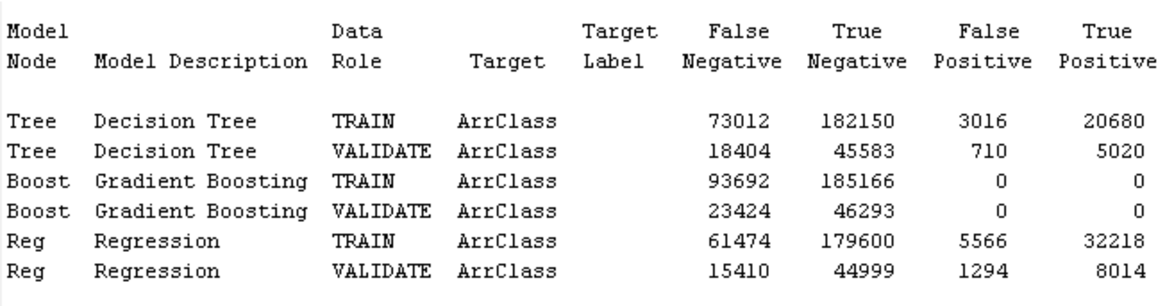


Figure 9.2 Misclassification chart and ROC curves for predictive models for arrival delay based on the flight dataset, with departure delay excluded

Report:

Model	Valid: Misclassification rate	Train: Average squared error	Train: Misclassification rate	Valid: Average squared error
Regression	0.23960	0.17484	0.24041	0.17486
Decision tree	0.27417	0.18767	0.27264	0.18837
Gradient boosting	0.33599	0.22310	0.33598	0.22310

Assessment report:



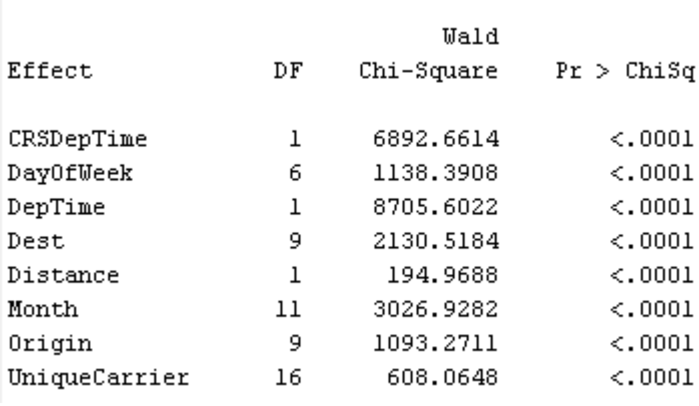
```
Model                         Data                   Target    False      True     False     True
Node   Model Description      Role        Target     Label   Negative  Negative  Positive  Positive

Tree   Decision Tree          TRAIN      ArrClass              73012    182150     3016     20680
Tree   Decision Tree          VALIDATE   ArrClass              18404     45583      710      5020
Boost  Gradient Boosting      TRAIN      ArrClass              93692    185166        0         0
Boost  Gradient Boosting      VALIDATE   ArrClass              23424     46293        0         0
Reg    Regression             TRAIN      ArrClass              61474    179600     5566     32218
Reg    Regression             VALIDATE   ArrClass              15410     44999     1294      8014
```

Model	Sensitivity	Specificity	Precision	Accuracy
Decision tree	0.22	0.98	0.87	0.73
Decision tree	0.21	0.98	1.00	0.73
Regression	0.34	0.97	0.85	0.76
Regression	0.34	0.97	0.86	0.76

- ➔ Regression statistically better, but Specificity still very low

Model summary:



```
                                  Wald
Effect               DF     Chi-Square     Pr > ChiSq

CRSDepTime            1      6892.6614         <.0001
DayOfWeek             6      1138.3908         <.0001
DepTime               1      8705.6022         <.0001
Dest                  9      2130.5184         <.0001
Distance              1       194.9688         <.0001
Month                11      3026.9282         <.0001
Origin                9      1093.2711         <.0001
UniqueCarrier        16       608.0648         <.0001
```

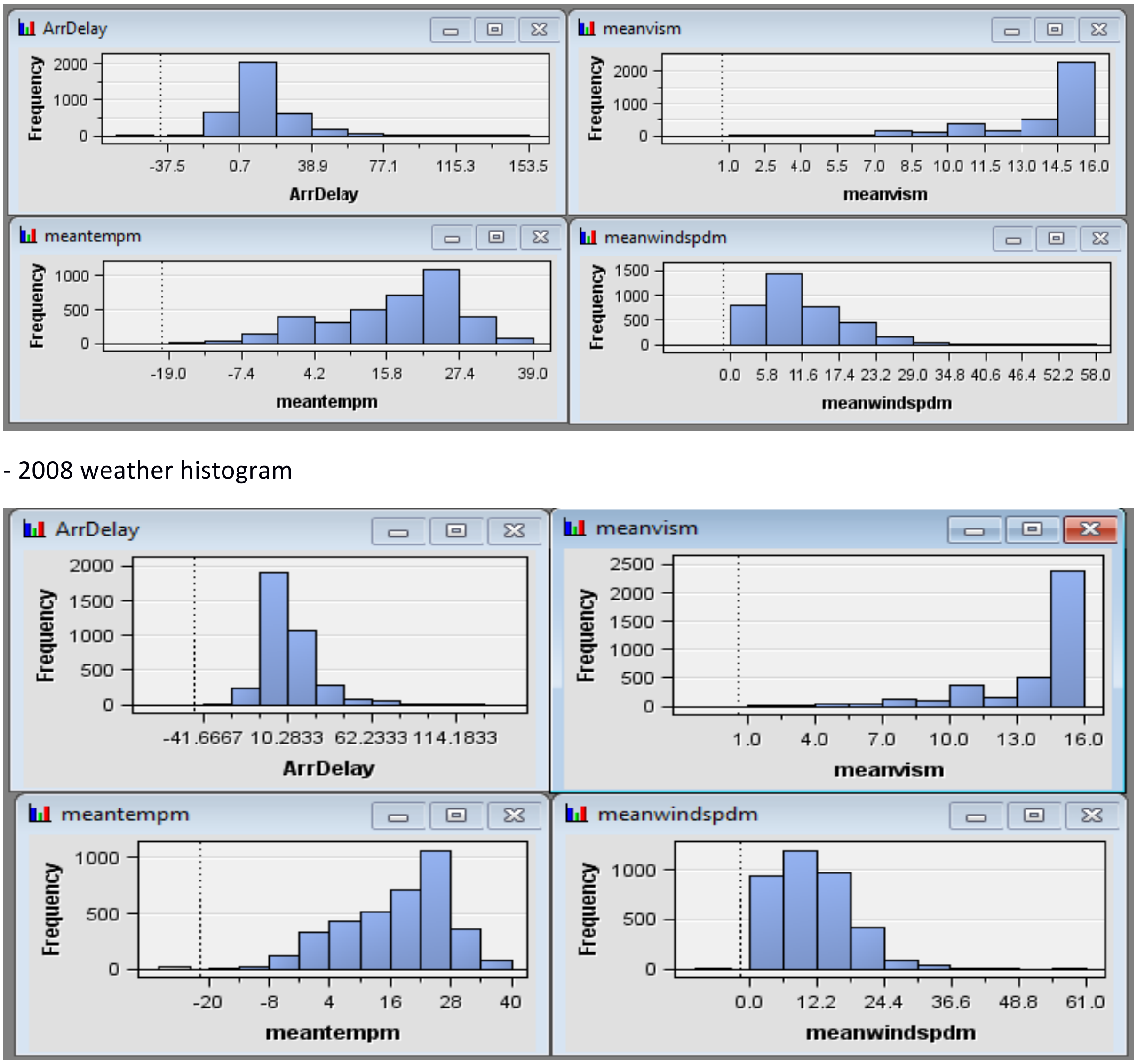
Step 3: Combined the weather data with the flight delays and undertook exploratory data analysis

The daily average arrival delay time was concatenated to the weather data via *Date*. Summary of the binary data in the weather dataset is shown in the table below.

Special weather	2007: 0	2007: 1	2008: 0	2008: 1
Fog	3509	141	3529	131
Hail	3649	1	3658	2
Rain	2688	962	2849	811
Thunder	3381	269	3474	186
Snow	3519	131	3530	130
Tornado	3649	1	3658	2

The distribution of the values in the interval variables are summarised by histograms.

- 2007 weather histogram



- 2008 weather histogram

Figure 9.3 Histograms of key weather variables for 2007 and 2008

The correlations between daily delays and special weather conditions were visualised by box plots.

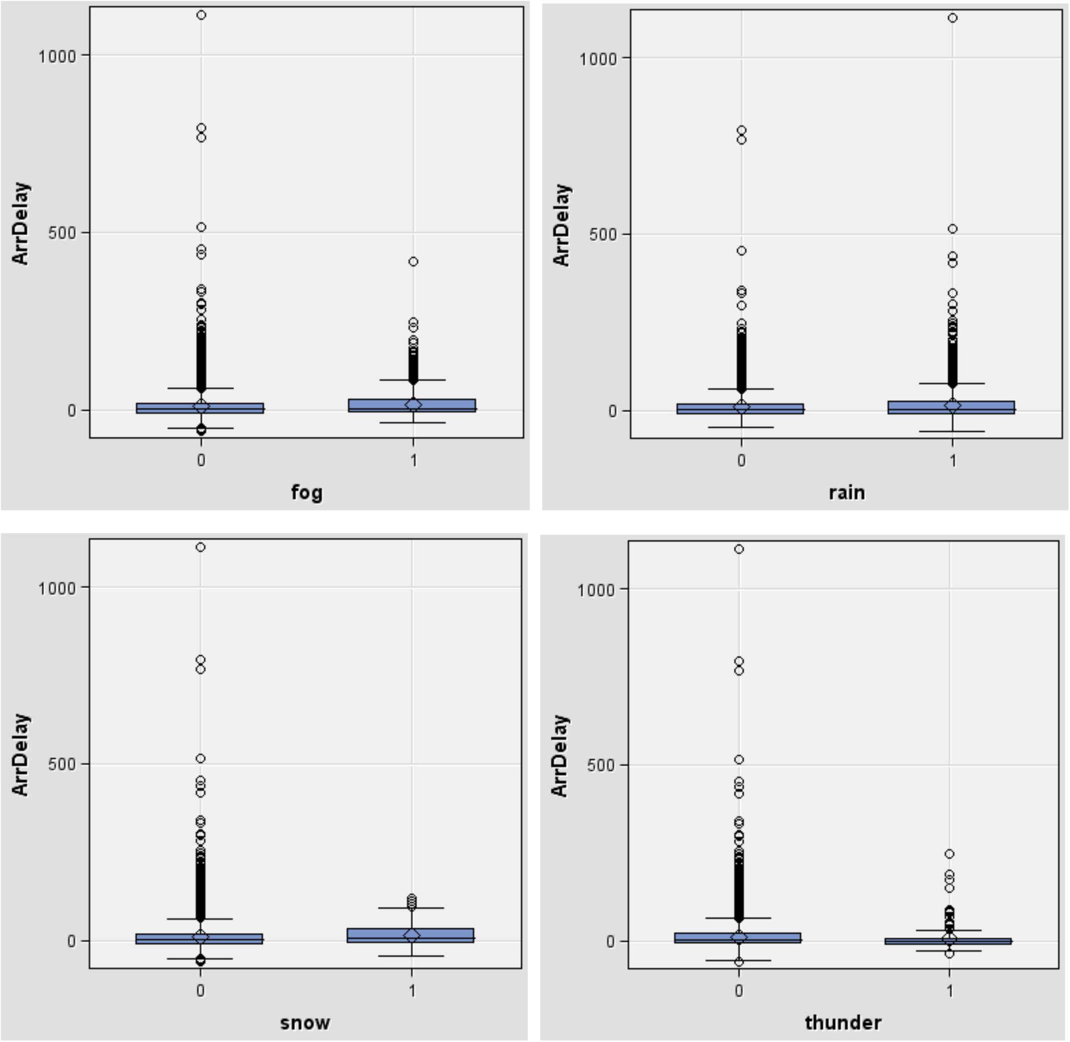


Figure. 9.4 Bivariate exploratory data analysis for key weather variables for 2007 and 2008

- ➔ Surprisingly, special weather conditions in the destination did not seem to be correlated with longer delays
- ➔ Surprisingly, special weather conditions at the destination airport did not correlate with longer delays.

Step 4: Run predictive models with weather data.

Model: SVM, Regression, Gradient boosting, Ensemble (Regression + Gradient boosting)

Class proportion in the target variable:

For 2007, 1:0 = 1891:1758

For 2008, 1:0 = 2135 : 1525

– Weather: for destination.

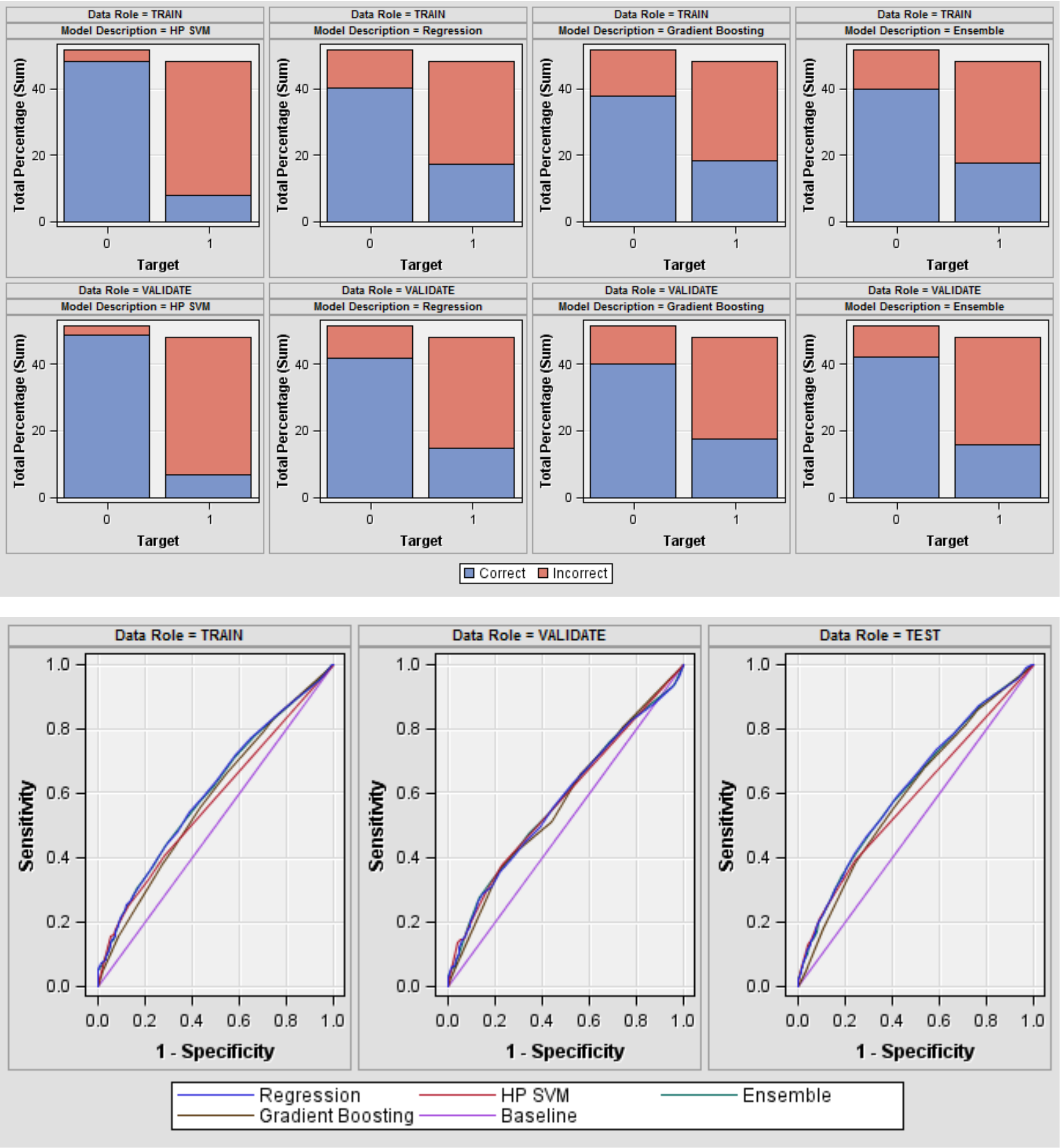
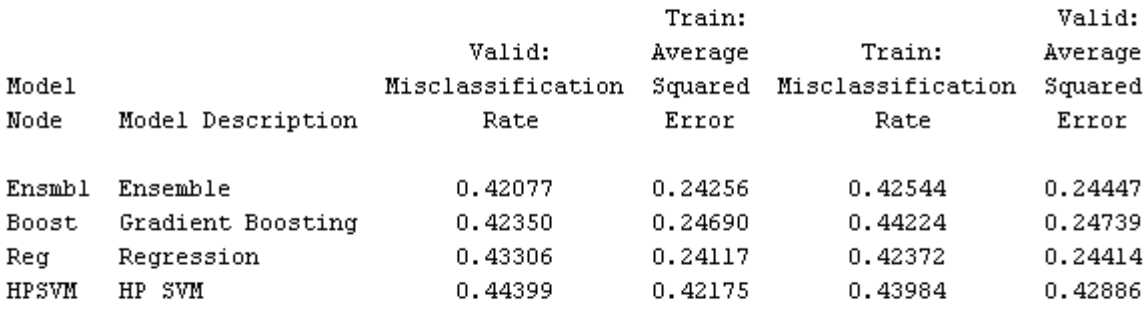


Figure 9.5 Misclassification chart and ROC curves for predictive models for arrival delay based on the flight dataset and destination weather



```
                                                 Train:                     Valid:
                                 Valid:          Average       Train:       Average
Model                       Misclassification   Squared   Misclassification Squared
Node     Model Description        Rate           Error          Rate         Error

Ensmbl   Ensemble               0.42077        0.24256        0.42544       0.24447
Boost    Gradient Boosting      0.42350        0.24690        0.44224       0.24739
Reg      Regression             0.43306        0.24117        0.42372       0.24414
HPSVM    HP SVM                 0.44399        0.42175        0.43984       0.42886
```

➔ Ensemble (based on Gradient Boosting and Regression): accuracy 58%

– Weather: for origin.

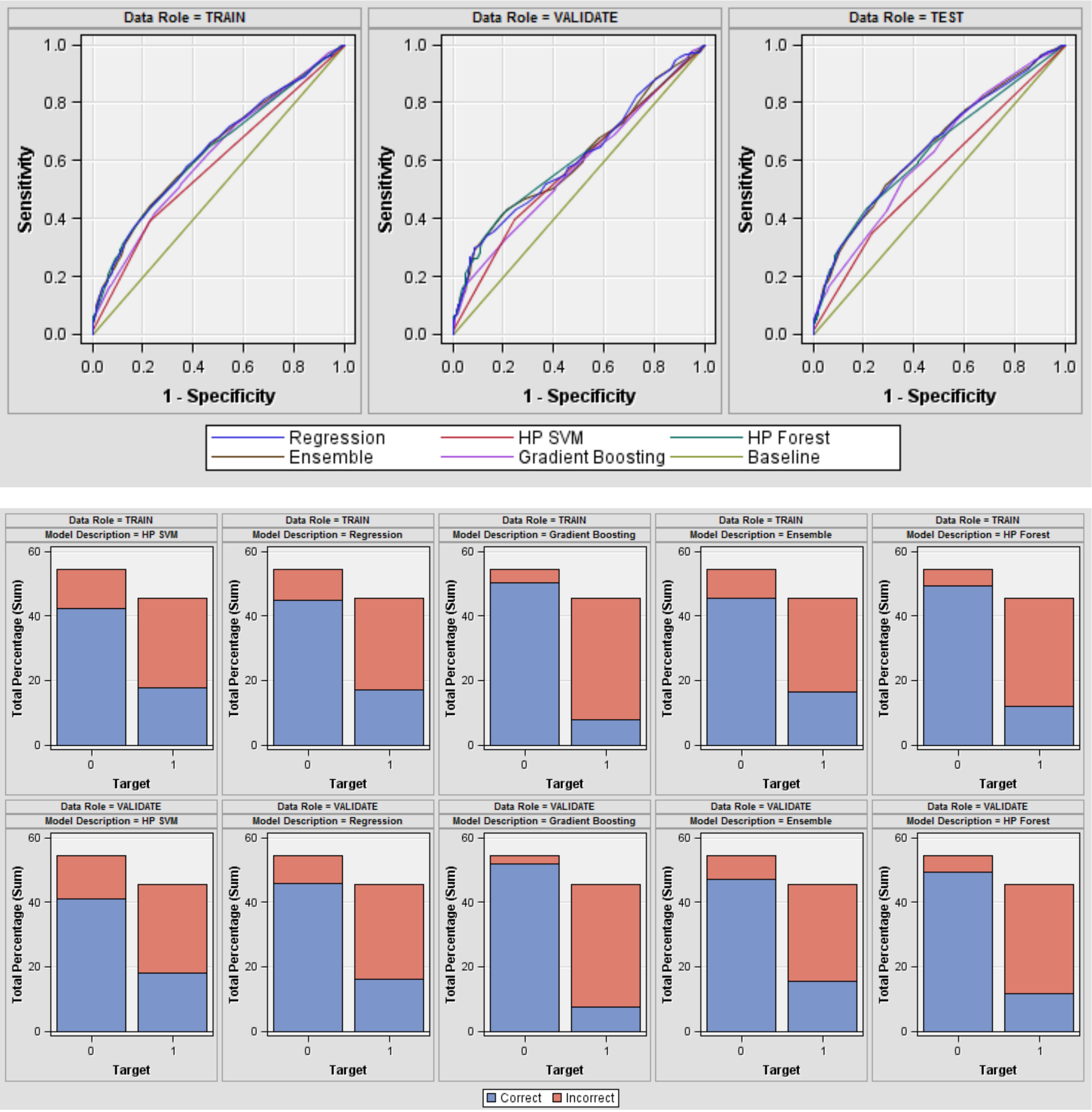
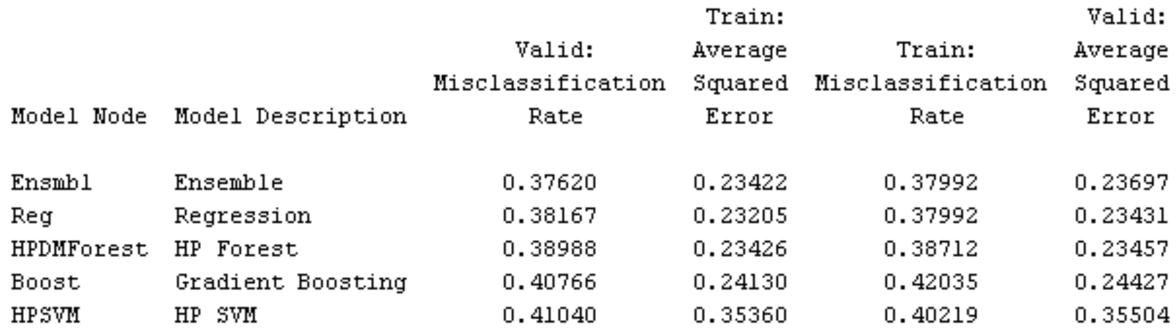
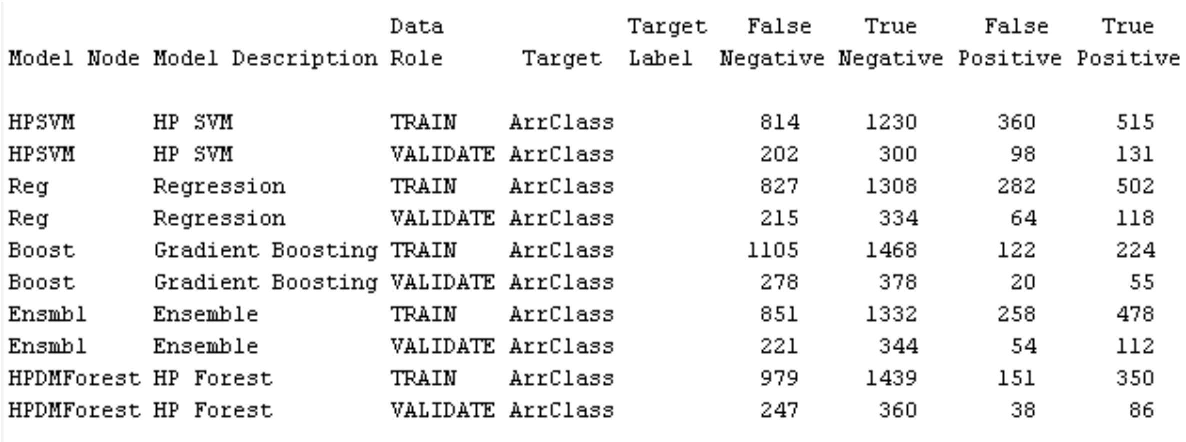


Figure 9.6 Misclassification chart and ROC curves for predictive models for arrival delay based on the flight dataset and origin weather



Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Ensmbl	Ensemble	0.37620	0.23422	0.37992	0.23697
Reg	Regression	0.38167	0.23205	0.37992	0.23431
HPDMForest	HP Forest	0.38988	0.23426	0.38712	0.23457
Boost	Gradient Boosting	0.40766	0.24130	0.42035	0.24427
HPSVM	HP SVM	0.41040	0.35360	0.40219	0.35504



Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
HPSVM	HP SVM	TRAIN	ArrClass		814	1230	360	515
HPSVM	HP SVM	VALIDATE	ArrClass		202	300	98	131
Reg	Regression	TRAIN	ArrClass		827	1308	282	502
Reg	Regression	VALIDATE	ArrClass		215	334	64	118
Boost	Gradient Boosting	TRAIN	ArrClass		1105	1468	122	224
Boost	Gradient Boosting	VALIDATE	ArrClass		278	378	20	55
Ensmbl	Ensemble	TRAIN	ArrClass		851	1332	258	478
Ensmbl	Ensemble	VALIDATE	ArrClass		221	344	54	112
HPDMForest	HP Forest	TRAIN	ArrClass		979	1439	151	350
HPDMForest	HP Forest	VALIDATE	ArrClass		247	360	38	86

- ➔ Ensemble (based on Gradient Boosting and Regression): accuracy: 72.4%
- ➔ Could use the data in the original dataset which claims weather delay only accounted for about 6.7% of the delays to explain the prediction result
- ➔ Model built from Origin weather has higher accuracy than destination weather

Step 5: Combined airline delay dataset with destination weather data

The weather data were right merged with the airline delay data via the variables *Date* and *Airport* from weather data, and *Date* and *Dest from* airline delay data.

Variables from the weather data used in the predictive model: Mean visibility, Mean wind speed.

(All the binary variables and meantempm, precipm were rejected)

A regression accuracy of 69% was identified.

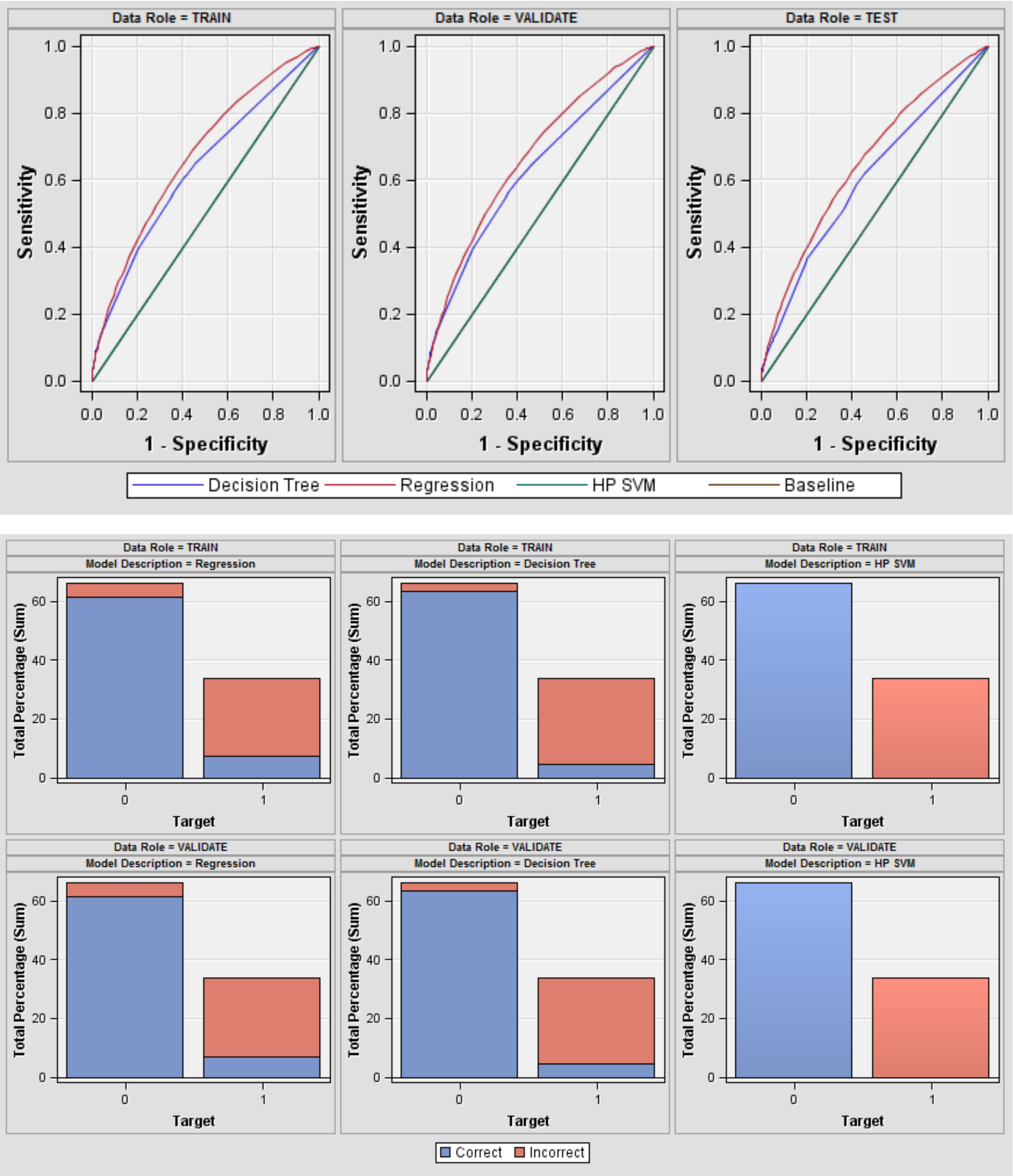
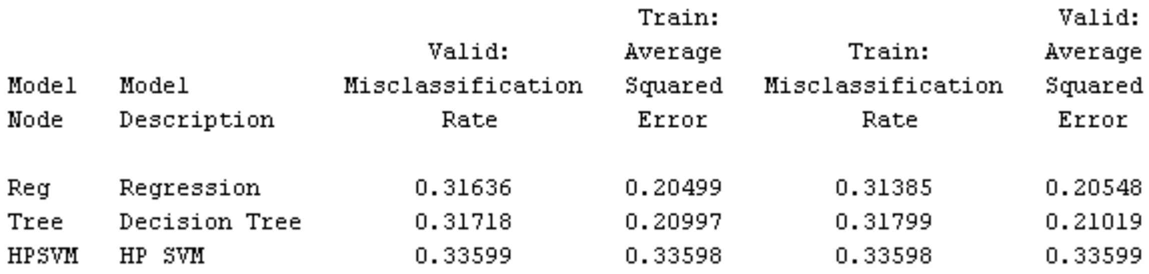
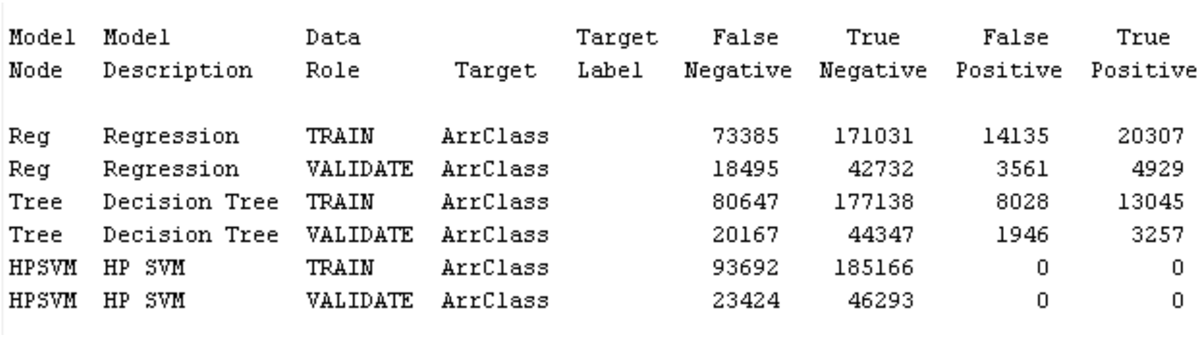


Figure 9.7 Misclassification chart and ROC curves for predictive models for arrival delay based on combined airline delay dataset with destination weather data



```
                                        Train:                          Valid:
                         Valid:         Average         Train:          Average
Model   Model       Misclassification   Squared   Misclassification     Squared
Node    Description       Rate           Error          Rate             Error

Reg     Regression       0.31636        0.20499        0.31385          0.20548
Tree    Decision Tree    0.31718        0.20997        0.31799          0.21019
HPSVM   HP SVM           0.33599        0.33598        0.33598          0.33599
```



```
Model  Model          Data                 Target    False     True    False    True
Node   Description    Role        Target   Label  Negative Negative Positive Positive

Reg    Regression     TRAIN     ArrClass             73385   171031    14135    20307
Reg    Regression     VALIDATE  ArrClass             18495    42732     3561     4929
Tree   Decision Tree  TRAIN     ArrClass             80647   177138     8028    13045
Tree   Decision Tree  VALIDATE  ArrClass             20167    44347     1946     3257
HPSVM  HP SVM         TRAIN     ArrClass             93692   185166        0        0
HPSVM  HP SVM         VALIDATE  ArrClass             23424    46293        0        0
```

- ➔ Regression: accuracy 69%

Step 6: Combined airline delay dataset with origin weather data

The weather data were right merged with the airline delay data via the variables *Date* and *Airport* from weather data, and *Date* and *Origin from* airline delay data.

Variables used in the predictive model from weather data: meanvism & meanwindspdm.

(All the binary variables and meantempm, precipm were rejected)

A regression accuracy of 70.4% was obtained.

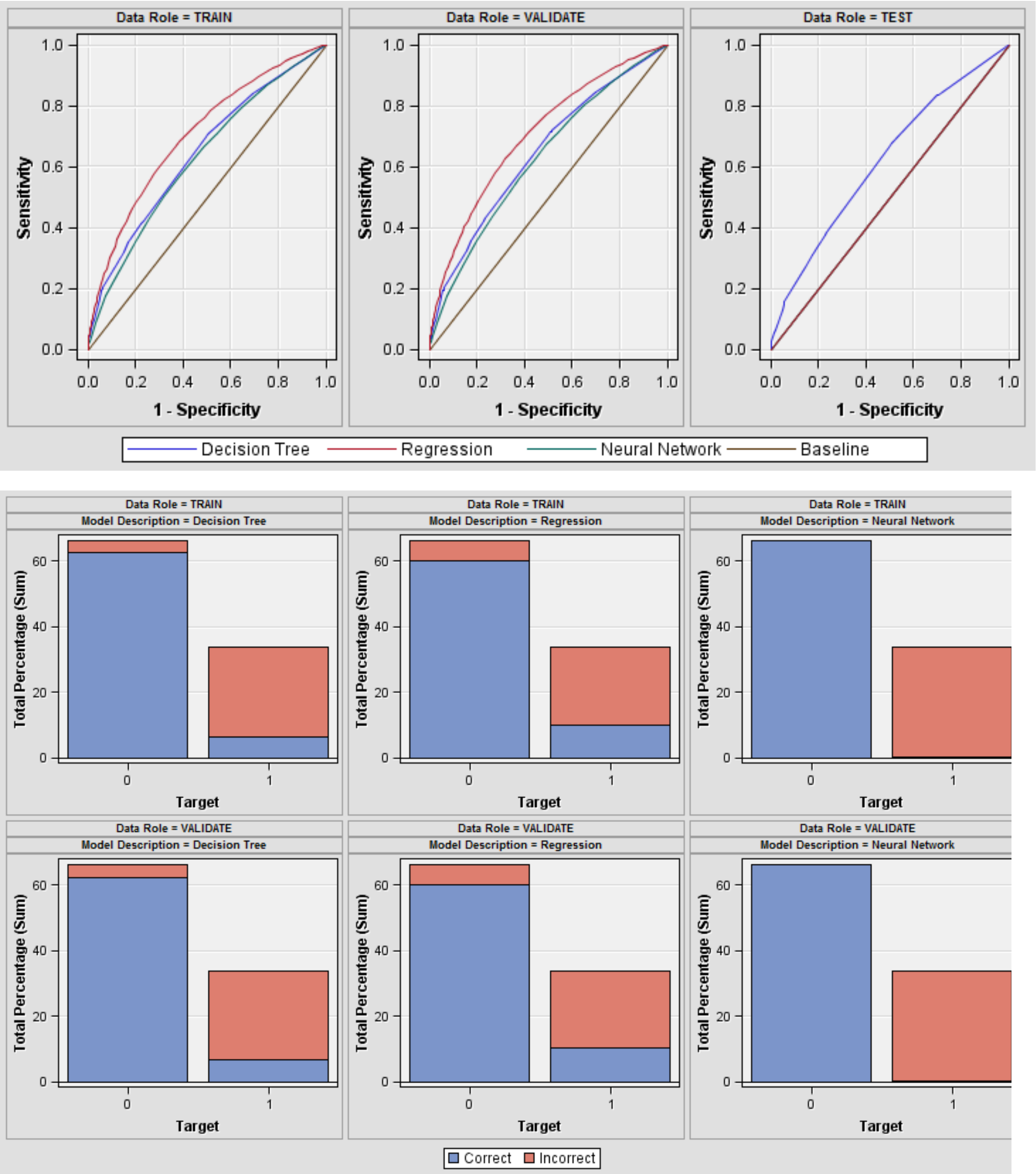
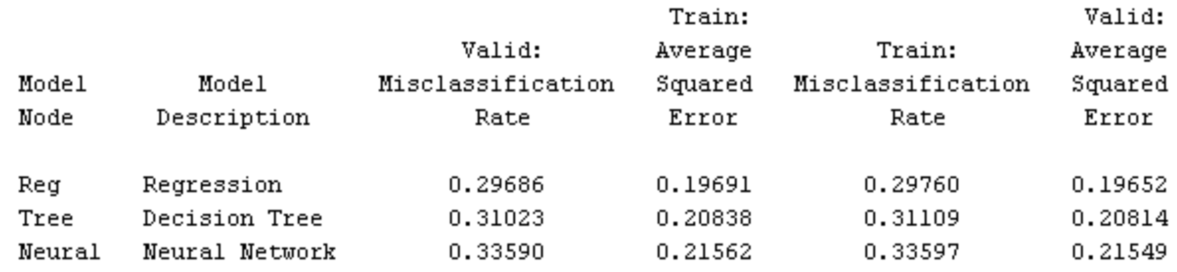
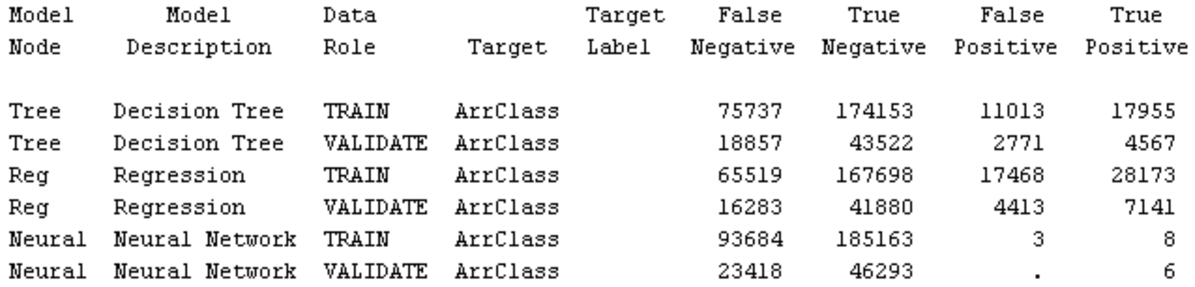


Figure 9.8 Misclassification chart and ROC curves for predictive models for arrival delay based on combined airline delay dataset with origin weather data



Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Reg	Regression	0.29686	0.19691	0.29760	0.19652
Tree	Decision Tree	0.31023	0.20838	0.31109	0.20814
Neural	Neural Network	0.33590	0.21562	0.33597	0.21549



Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Tree	Decision Tree	TRAIN	ArrClass		75737	174153	11013	17955
Tree	Decision Tree	VALIDATE	ArrClass		18857	43522	2771	4567
Reg	Regression	TRAIN	ArrClass		65519	167698	17468	28173
Reg	Regression	VALIDATE	ArrClass		16283	41880	4413	7141
Neural	Neural Network	TRAIN	ArrClass		93684	185163	3	8
Neural	Neural Network	VALIDATE	ArrClass		23418	46293	.	6

- ➔ Regression, accuracy: 70.4%
- ➔ Using Destination/Origin weather results in similar accuracy
- ➔ Weather data did not improve the model accuracy

9.2 Assumptions

1) No reason for delay has been provided in 60% of cases where an arrival delay existed. Investigations have concluded that these records to not correspond to particular delay durations or airlines. It is therefore assumed that flights where a delay incurred and a reason provided are representative of all flights

9.3 Data Dictionary

– Airline delay dataset variable descriptions

	Name	Data Type	Units	Description
1	Year	integer	Year	1987-2008
2	Month	integer	Months	1-12
3	DayofMonth	integer	Days	1-31
4	DayOfWeek	integer	Days	1 (Monday) - 7 (Sunday)
5	DepTime	Data/time	hhmm	actual departure time (local, hhmm)
6	CRSDepTime	Data/time	hhmm	Departure Time (as scheduled by the computer reservation system) (local, hhmm)
7	ArrTime	Data/time	hhmm	actual arrival time (local, hhmm)
8	CRSArrTime	Data/time	hhmm	scheduled arrival time (local, hhmm) (as scheduled by the computer reservation system)
9	UniqueCarrier	string		unique carrier code
10	FlightNum	string		flight number
11	TailNum	string		plane tail number
12	ActualElapsedTime	Data/time	minutes	
13	CRSElapsedTime	Data/time	minutes	
14	AirTime	Data/time	minutes	
15	ArrDelay	Data/time	minutes	arrival delay, in minutes
16	DepDelay	Data/time	minutes	departure delay, in minutes
17	Origin	string		origin IATA airport code
18	Dest	string		destination IATA airport code
19	Distance	Integer	miles	
20	TaxiIn	Data/time	minutes	taxi in time, in minutes
21	TaxiOut	Data/time	minutes	taxi out time in minutes
22	Cancelled	Boolean	0 or 1	was the flight cancelled?
23	CancellationCode	string		reason for cancellation (A = carrier, B = weather, C = NAS, D = security)
24	Diverted	Boolean	0 or 1	1 = yes, 0 = no

25	CarrierDelay	Data/time	minutes	
26	WeatherDelay	Data/time	minutes	
27	NASDelay	Data/time	minutes	
28	SecurityDelay	Data/time	minutes	
29	LateAircraftDelay	Data/time	minutes	

– Weather dataset variable descriptions

	Name	Data Type	Units	Description
1	Date			Date of the year (for example, '2007-01-01')
2	Airport			Three letter airport abbreviation
3	meantempm			Mean temperature of the day (Celsius degree)
4	meanvism			Mean visibility of the day (1-16)
5	meanwindspdm		Knots	Mean wind speed (knots)
6	precipm		cm	Mean precipitation (cm)
7	rain	Boolean	0 or 1	Binary (1: yes, 0: no)
8	fog	Boolean	0 or 1	Binary (1: yes, 0: no)
9	hail	Boolean	0 or 1	Binary (1: yes, 0: no)
10	snow	Boolean	0 or 1	Binary (1: yes, 0: no)
11	thunder	Boolean	0 or 1	Binary (1: yes, 0: no)
12	tornado	Boolean	0 or 1	Binary (1: yes, 0: no)
13	city			City name where the airport locates
14	state			State name where the airport locates

9.4 Python script for weather scrape

```
import urllib2
import json
import re
import pandas as pd
from datetime import date, datetime, timedelta as td

# columns we are intereseted in are listed here
col_names_full = ['Date', 'state', 'city','fog', 'rain', 'snow', 'hail', 'thunder', 'tornado', 'meantempm',
'maxtempm', 'mintempm', 'meanwindspdm', 'meanvism', 'precipm']

# columns we will scrape from the web
col_names = ['fog', 'rain', 'snow', 'hail', 'thunder', 'tornado', 'meantempm',
'maxtempm', 'mintempm', 'meanwindspdm', 'meanvism', 'precipm']

# create a dataframe to store data
weather = pd.DataFrame(columns = col_names_full)
weather = weather.fillna(0)

# set the format of the url, 'cf4e99ce6a5a4f52' is my api key
url = "http://api.wunderground.com/api/cf4e99ce6a5a4f52/history_date/q/statecode/city.json"

# set the time period here
d1 = date(2007, 1, 1)
d2 = date(2008, 12, 31)
delta = d2 - d1

# start to scrape
for i in range(delta.days + 1):
    date_1 = d1 + td(days=i)
```

```
        date_2 = datetime.strftime(date_1, "%Y%m%d")

# edit state and city name here
        state = 'FL'  #two letter code of state
        city = 'Orlando'     #city name

# set the correct url
        url_date = re.sub('history_date/q/statecode/city', 'history_%s/q/%s/%s' %(date_2, state, city), url)

        f = urllib2.urlopen(url_date)
        json_string = f.read()
        parsed_json = json.loads(json_string)

# get the daily summary
        daily_summary= parsed_json['history']['dailysummary']
        daily = pd.DataFrame(daily_summary, columns = col_names)

        daily['Date'] = date_1
        daily['state'] = state
        daily['city'] = city

# append the data to the dataframe
        weather = pd.concat([weather, daily])

        f.close()

# save the data to a csv file
weather.to_csv('weather_AirportCode_City_0708.csv')
```

10 References

Abudi, G. (2009) The Five Stages of Team Development – Every Team Goes Through Them, retrieved from http://www.ginaabudi.com/the-five-stages-of-team-development-part-i/

AhmadBeygi, S., Cohn, A., Guan, Y. and Belobaba, P. (2008) Analysis of the potential for delay propagation in passenger airline networks, Elsevier Ltd, Journal of Air Transport Management 14 221– 236

Ball, M., Barnhart, C., Dresner, M., Hansen, M., Neels, K., Odoni, A., Peterson, E., Sherry, L., Trani, A., and Zou. (2010) Total Delay Impact Study - A Comprehensive Assessment of the Costs and Impacts of Flight Delay in the United States, Washington DC, USA, Federal Aviation Administration

Barsalou, M. (2015), Exploring Tukey's Exploratory Data Analysis, Quality Digest

Brillinger, D. John W. Tukey: His life and professional contributions. Ann. Statist. 30 (2002), no. 6, 1535--1575. doi:10.1214/aos/1043351246. http://projecteuclid.org/euclid.aos/1043351246.

Cadle, J. and Yates, D. (2004) Project Management for Information Systems, Prentice Hall

Schaefer, L. and Millner, D. (2001) Flight Delay Propagation Analysis with Detailed Policy Assessment Tool, Center for Advanced Aviation System Development, The Mitre Corporation, Virginia, USA

Yufeng, T., Ball, O. and Jank, W (2008) Estimating Flight Departure Delay Distributions—A Statistical Approach with Long-Term Trend and Short-Term Pattern, Journal of the American Statistical Association, 103:481, 112-125, doi: 10.1198/016214507000000257

Business Analytics.com, (2015) Ensemble Modelling, [online] Available at http://searchbusinessanalytics.techtarget.com/definition/Ensemble-modeling

Project Management Institute (2014). What is Project Management? | Project Management Institute. [online] Available at http://www.pmi.org/about/learn-about-pmi/what-is-project-management

Bridgeable. (2017). The Importance of Data Visualization. [online] Available at: http://bridgeable.com/the-importance-of-data-visualization/ [Accessed 22 Apr. 2017].

Butler Analytics. (2017). What is Tableau? - Butler Analytics. [online] Available at: http://www.butleranalytics.com/what-is-tableau/ [Accessed 22 Apr. 2017].

Data Sources

1) http://stat-computing.org/dataexpo/2009/the-data.html
2) https://www.wunderground.com/