Learner Data Management

University of Oklahoma

FAA COE TTHP: Learner Data Management (ANN 001)

Authors:
Christan Grant, Ph.D.
Dean Hougen, Ph.D.

Contact:
cgrant@ou.edu +1-405-325-5408
hougen@ou.edu +1-405-325-3150

April 9, 2019

Contents

1 Executive Summary

Learner data for different sets of FAA trainings is stored across different systems at different locations. In the 1st phase of our project, we explored and analyzed existing database management systems in the academy. We looked in detail at NTD, TRAX, CEDAR, and the Training Database Management systems. Initially, started identified and created an open-source document storage and search system to support a safe and efficient airspace system. In particular, we determined that the Comprehensive Knowledge Archive Network (CKAN) system was suitable to move forward with our work. Using CKAN, we installed and created a document storage website. We continued our analysis of the existing database to be able to understand the different training systems used across the FAA. We also analyzed the course documents required for training air traffic controllers for the FAA. We analyzed the existing schema of the database management system and the relationships between the different systems for updating and transferring data. This helped us understand the types of information stored and maintained for trainees for the FAA. There have been multiple updates in the requirements of the project which created a diversion in our analysis and approach for accomplishment of the project goals. With the help of project coordinators, we learned about the developing system CEDAR and Training Database Management system. These discussions were very informative and helped us in understanding the most recent changes in the FAA database management system. These discussions changed the directions of our plans, as we learned about the database systems in development and affected our plans. Late in the award we were alerted of upcoming changes to systems involved with storing training data and the development of current systems. We then pivoted our exploration to demonstrate method of performing principled analysis of general data types. In summary, we delivered example methods of detecting statistical anomalies in flight data (§ A) and also a thorough analysis of an available flight data set (§ D and § C). This report details our analysis and our inputs on the existing FAA training databases. This report also includes the details of our schedule and milestones.

2 Accomplished Task

We started our project by creating a document storage website. We analyzed the use of the Comprehensive Knowledge Archive Network (CKAN) system, an open-source data portal system. CKAN is the most suitable data management system available for storing and accessing large amounts of data. With our goal to make the scanned document accessible to the user in an electronic form, we started working on optical character recognition (OCR) technology. For effective and efficient use of technologies like CKAN and OCR for accessing different sets of documents, our first step was to have a good understanding of the course documents. To achieve this, we looked at 105 course documents from 12051 V.2016-11 to 6005251. We identified learning evaluation types in courses (e.g., end-of-lesson tests, performance checks, block tests, etc.) and the type of tests, where available (e.g., multiple choice, fill in the blank, observations, etc.). We used tesseract OCR package for extracting text from scans. When working on software for identifying and extracting the questions and answers of a test of a training course, we used a the Pytesseract software and tested it on an example unmanned aircraft general (UAG) training test multiple choice exam; an exam that is part of the Remote Pilot Certificate.

We started reviewing the training requirements of the FAA and the rules and regulations of the performance evaluation process of trainees in the FAA. We had meetings with system experts Ray Sorise (Computer System Designers; Majority Tech Ops) and Randi Schmicking (FAA; Majority Air Traffic) discussing the current learner management infrastructure and opportunities (respectively). During the course of our project, we had a change in the plans for our project goals. With the help of these meetings, we started our next step towards learning and understanding the existing system. In this step, we learned details about NTD (National Training Database), TRAX and CEDAR (for CEDAR, the information was changing as the system is in development). We learned that TRAX doesn't maintain the data at a central location and that it has 300 individual databases. The TRAX system is all encompassing and only qualification entries from lab, classroom, and on-the-job training (OJT) are entered into TRAX and printed out as a part of an official record of training. During our meeting with the experts, we learned that TRAX and NTD will get sunset by a database system known as the Training Database Management system. It was under development and the technical details for the Training Data Management system were changed.

With our focus on analyzing the existing database management system and provide insights about the pain points and gaps in the relationship between different database systems used by the FAA, we have completed the following set of tasks (in brief):

1. Analyzed the information needs of the FAA with respect to technical training and human performance. This includes review of course documents.

2. Analyzed the training documents and designed the scheme for storing documents.

3. Explored the CKAN open-source document storage and search system framework to store training documents and built a prototype system.

4. Reviewed current documents describing learner data management storage systems and federal storage requirements. This includes having reviewed the current TRAX, e-LMS, NTD, and CEDAR methods.

5. Met with system experts and development teams (e.g., for CEDAR) to understand decisions and information going into the new system.

6. Identified pain points and gaps of information sharing between CEDAR (NTD) and the Academy through discussion with experts.

7. Worked with experts to discuss desired system design implications.

8. Summarized findings including data stored (course materials, training structure, performance records) and longevity, uses, data transfers, and plans.

9. Explored the analytical pathways to provide a solution to improve the success rate for on-the-job training of air traffic controllers.

2.1 Our Progress

1. As per the plan for our project, our first milestone was to analyze FAA technical training and human performance (TTHP) information needs and the existing databases used by the FAA. We analyzed the existing database systems NTD (National training database) and TRAX (distributed system for storing the training progress). With the help of experts, we gathered information on the database systems under development including CEDAR and Training Database Management system. We found that the Training Database Management system is going to sunset TRAX and NTD. With the changing details and updates, we have changed the direction of our initial plan which affected the completion of the subsequent milestones.

2. The second milestone of the project plan was to set up a secure data environment, make it operational, and demonstrate it. To achieve this milestone, we installed and created a document storage website. We developed our understanding of CKAN, which is data management system for accessing large sets of data. We tested CKAN with different databases to understand the scope and found that CKAN is a highly suitable data management system for our goal. We were able to reach the 1st two milestones on time.

3. Our next milestone was to have a first-generation data parsing software package written, operating and demonstrated. To achieve this goal, we started working on OCR technology. With this technology, we planned to have a system which could read hand-written forms and convert scanned forms to electronic data. We tested our code on an example unmanned aircraft general (UAG) training multiple-choice exam; an exam that is part of the Remote Pilot Certificate.

4. For our next milestone, we reviewed training course documents used by the FAA. When reviewing storage of course documents and their maintenance, we analyzed course documents that are stored and updated locally. We looked at 105 course documents from 12051 V.2016-11 to 6005251. We also started reviewing the existing database management systems used by the FAA, which are TRAX, NTD, CEDAR, and e-LMS. To gain insight into these systems, we had meetings with system experts. The meetings were informative and we learned about changes in progress for these systems. We learned that CEDAR is in a development phase and is going to sunset TRAX and NTD. To understand more about CEDAR, we met with an expert in DC during the COE SOAR conference. This is the point where our plan started to change.

5. Our biggest challenge is to characterize the existing systems along with all the changes that are being done to the existing systems. But, as mentioned earlier, in the development phase, there are many changes being made. We found that a new system known as the Training Database Management system is being developed which is now going to sunset TRAX and NTD instead of CEDAR. With all the updates, it was challenging to start analyzing machine learning opportunities within the FAA. As per the planned milestone, our work should be continued in the direction of creating an online and mobile interface design but after the meeting with the experts and feedback gained at the COE SOAR conference, this aspect of the project was put on hold.

6. With all the changes in the requirements of the project, we started with in-depth analysis of the relationship between the existing databases and worked to learn more about both CEDAR and the Training Data Management system.

2.2 Limitations

Limitations we found with the existing system are as follows:

1. Trainee performance information is stored with very few details, often pass or fail without any details on the strengths and weakness of the trainee. As per our understanding, the progress of the trainee at any stage of the training progress will help in analyzing the areas where improvement is needed.

2. There is a limitation when analyzing a student's progress in the academy, as the TRAX database system is distributed across locations. Each location has its own complexity and characteristics when it comes to air traffic control and reallocation of the trainees will be easier and more effective if one central database system is used for storing student performance details.

3. There is substantial variation in the success rate of OJT and there is broad scope for improvement in the success rate of trainees. When more students fail after completing OJT, it reduces the efficiency of FAA funds and investment in OJT trainees.

4. The existing FAA database is distributed across the various locations and a large portion of the training details are stored locally.

3 Data Analysis Summary

In Appendix § A we give an walkthrough of how to a particular aggregation bias anomaly called Simpson's Paradox. As we describe in our conference paper, Simpson's Paradox can exist and can introduce fallacies in the the understanding of learner trends. Detection of this paradox should be integrated into the developing storage and analysis stystems.

In Appendix § D and § C we describe the our analysis of open source flight data. The most interesting discovery is that there is a statistically meaningful negative correlation between flight delays and accidents at US airports in the data sets. Because there is some uncertainty in how to interpret missing data, we can't say for certain whether this correlation is strong or weak, but it is almost certainly real.

What we wondered going into this data is, should safety-conscious people avoid airports with a higher percentage of delays because they are likely to be overloaded and dangerous? Or, alternately, should safety-conscious people actually prefer airports with a higher percentage of delays because those airports are willing to incur delays in order to preserve safety? The answer appears to be the latter!

Of course, correlation does not imply causation, but it may well be the case that some airports and rushing to avoid delays and this causes them to be less safe. Note that, if the correlation had gone the other way, an alternative explanation would have been that accidents cause delays, which is probably true. However, with a negative correlation, it is hard to imagine

how a lack of accidents could cause delays, so it seems more reasonable to go the other way and conclude that delays allow airports to avoid accidents. We recommend further analysis, using these techniques on learner data. These techniques should be integrated into data storage systems.

References

[1] Chenguang Xu, Sarah Brown, and Christan Grant. Detecting simpson's paradox, 2018.

Appendices

A Investigator Bios

A.1 Christan Grant, Ph. D.

Christan Grant, Ph.D.

Assistant Professor · University of Oklahoma

Databases, Text Analytics, and Interactive Data Mining · christangrant.com

Education

August 2015 — Ph.D. in Computer Science, **University of Florida**, Gainesville
Dissertation: *Query-Driven Text Analytics for Knowledge Extraction, Resolution, and Inference*
Advisors: Daisy Zhe Wang and Joseph N Wilson

May 2008 — B.S. in Computer Engineering, **University of Florida**, Gainesville

Appointments

Current / Aug 2015 — Assistant Professor at University of Oklahoma
School of Computer Science, Norman, Oklahoma
Faculty Fellow, Dunham College

Publications

Journals

[J 5] Jared Bond, **Christan Grant**, Joshua Imbriani, Erik Holbrook. *A Framework for Interactive t-SNE Clustering.* To appear in the International Journal of Software Informatics (IJSI) special issue on Visual Analytics. 2017

[J 3] Sean Goldberg, Daisy Zhe Wang, **Christan Grant**. *A Probabilistically Integrated System for Crowd-Assisted Text Labeling and Extraction.* ACM Journal of Data and Information Quality (JDIQ). 2017.

[J 4] Kun Li, Xiaofeng Zhou, Daisy Zhe Wang, **Christan Grant**, Alin Dobra, Christopher Dudley. *In-Database Batch and Query-time Inference over Probabilistic Graphical Models using UDA-GIST.* The VLDB Journal (VLDBJ). 2016.

[J 2] **Christan Grant**, Daisy Zhe Wang. *A Challenge for Long-term Knowledge Base Maintenance.* ACM Journal of Data and Information Quality (JDIQ). 2015.

[J 1] Daisy Zhe Wang, Yang Chen, **Christan Earl Grant**, Kun Li. *Efficient In-Database Analytics with Graphical Models.* IEEE Data Engineering Bulletin 37(3). 2014.

Conferences

[C 19] Ashwin Krishna Gunasekaran, Maryan Bahojb Imani, Latifur Khan, **Christan Grant**, Patrick T. Brant, Jennifer Holmes. *SPERG: Scalable Political Event Report Geoparsing in Big Data.* IEEE Intelligence and Security Informatics (ISI). Miami, Florida. 2018.

[C 18] Yan Liang, **Christan Grant**, Andrew Halterman, Jill Irvine, Khaled Jabr. New Techniques for *Coding Political Events Across Languages.* IEEE 18th International Conference on Information Reuse and Integration (IRI). Salt Lake City, UT. 2018.

[C 17] Jasmine DeHart, **Christan Grant**. *Visual Content Privacy Leaks on Social Media Networks.* The IEEE Symposium on Security and Privacy (S&P). San Francisco, California. 2018.

[C 16] Chenguang Xu, Sarah M. Brown, **Christan Grant**. *Detecting Simpsons Paradox.* The 31st International Florida Artificial Intelligence Research Society (FLAIRS) Conference. Melbourne, Florida. 2018.

[C 15] Nina Cesare, **Christan Grant**, Elaine O. Nsoesie. *Estimating Obesity Prevalence By Age and Gender Using Social Media Data* The Annual Meeting for the Population Association of America (PAA). Denver, CO. 2018.

Workshop

[W 8] Keerti Banweer, Austin Graham, Nina Cesare, Elaine O. Nsoesie, Joseph T. Ripberger, **Christan Grant**. *Multi-stage Collaborative filtering for Tweet Geolocation*. The 2nd ACM SIGSPATIAL Workshop on Recommendations for Location-based Services and Social Networks (LocalRec). Seattle, Washington, USA. 2018.

[W 8] Chenguang Xu, Sarah M. Brown, **Christan Grant**, Chris Weaver. *Interactive Visual Analytics for Simpson's Paradox Detection*. The 3rd Workshop on Human-In-the-Loop Data Analytics (HILDA). Houston, Texas. 2018.

Service

2019	Panelist	National Science Foundation
2019	Reviewer	North American Chapter of the Association for Computational Linguistics (NAACL)
2019	Reviewer	IEEE Transactions on Knowledge and Data Engineering (TKDE)
2018	Reviewer	Women in Machine Learning (WIML)
2018	Reviewer	Black in Artificial Intelligence (BAI)
2018	Reviewer	BMJ Open
2018	Reviewer	IEEE International Conference on Information Reuse and Integration (IRI)
2018	Chair for Doctoral Consortium	Richard Tapia Conference
2018	General Chair	Broadening Participation in Data Mining (BPDM)
2017	National Science Foundation Review Panelist	
2017	Program Committee	Black in AI
2017	OU McNair Advisory Board	Ronald E. McNair Program
2017	Technical Program Committee	IEEE International Conference on Information Reuse and Integration (IRI)
2017	Travel Awards Co-Chair	International Conference on Information and Knowledge Management (CIKM)
2017	General Co-Chair	Broadening Participation in Data Mining (BPDM)
2017	Deputy Chair for Doctoral Consortium	Richard Tapia Conference
2017	External Reviewer	ACM Symposium of Applied Computing
2017	Organizing Committee	The Pipeline Workshop:Diversifying the HPC Workforce

Teaching

Spring-B '19	University of Oklahoma	DSA 5970 Python for Data Analysis
Spring '19	University of Oklahoma	CS 5293 Text Analytics
Fall '18	University of Oklahoma	CS 3113 Introduction to Operating Systems
Spring-B '18	University of Oklahoma	UNIV Weapons of Math Destruction
Spring '18	University of Oklahoma	CS 5970 Text Analytics
Fall '17	University of Oklahoma	CS 3113 Introduction to Operating Systems
Spring '17	University of Oklahoma	CS 5970 Introduction to Text Analytics
Fall '16	University of Oklahoma	CS 3113 Introduction to Operating Systems
Spring '16	University of Oklahoma	CS 5970 Text Analytics
Spring '14	University of Florida	CIS 4301 Introduction to Databases

A.2 Dean Hougen, Ph. D.

Dean F. Hougen, Associate Professor, University of Oklahoma, School of Computer Science

Professional Preparation

- Undergraduate Institution: Iowa State University, Major: Computer Science, Degree and Year: B.S. 1988.
- Graduate Institution: University of Minnesota, Major: Computer Science, Degree and Year: Ph.D. 1998.
- Postdoctoral Institution: University of Minnesota, Area: Robotics, 1998-2001.

Professional Appointments

- Associate Professor, School of Computer Science, University of Oklahoma, 2007–present.
- Graduate Faculty Member, School of Electrical and Computer Engineering, University of Oklahoma, 2006–present.
- Assistant Professor, School of Computer Science, University of Oklahoma, 2001–2007.
- Visiting Assistant Professor, Department of Computer Science and Engineering, University of Minnesota, 1998–2001.
- Associate Director, Center for Distributed Robotics (Nikolaos Papanikolopoulos, Director), Department of Computer Science and Engineering, University of Minnesota, 1998–2001.

Products

Five Recent

- Will Booker and Dean Frederick Hougen. "Meiotic Inheritance and Gene Dominance in Synthetic Sympatric Speciation." *IEEE Congress on Evolutionary Computation*, 8 pages, July 2018. `https://doi.org/10.1109/CEC.2018.8477761` **Best Student Paper Award.**
- Byran Hoke and Dean Frederick Hougen. "Nurturing Promotes the Evolution of Generalized Supervised Learning." *IEEE Congress on Evolutionary Computation*, 8 pages, July 2018. `https://doi.org/10.1109/CEC.2018.8477786`
- Joohee Suh and Dean Frederick Hougen. "The Context-Aware Learning Model: neuro-experience-powered Logistic Regression Backpropagation (CALM-nepLRB)." *IEEE/INNS International Joint Conference on Neural Networks*, 8 pages, July 2018. `https://doi.org/10.1109/IJCNN.2018.8489617`
- Syed Naveed Hussain Shah and Dean F. Hougen. "Nurturing Promotes the Evolution of Reinforcement Learning in Changing Environments." *IEEE Symposium Series on Computational Intelligence*, 8 pages, November 2017. `https://doi.org/10.1109/SSCI.2017.8285400`
- Syed Naveed Hussain Shah and Dean F. Hougen. "Stochastic Synapse Reinforcement Learning (SSRL)." *IEEE Symposium Series on Computational Intelligence*, 8 pages, November 2017. `https://doi.org/10.1109/SSCI.2017.8285425`

Five Other Significant

- David R. Peterson, Jamie D. Barrett, Kimberly S. Hester, Issac C. Robledo, Dean F. Hougen, Eric A. Day, and Michael D. Mumford. "Teaching People to Manage Constraints: Effects on Creative Problem-Solving." *Creativity Research Journal*, 25(3): 335-347, August 2013.
- Jamie D. Barrett, David R. Peterson, Kimberly S. Hester, Issac C. Robledo, Eric A. Day, Dean F. Hougen, and Michael D. Mumford. "Thinking About Applications: Effects on Mental Models and Creative Problem-Solving." Creativity Research Journal, 25(2): 199-212, May 2013.
- Michael D. Mumford, Kimberly S. Hester, Issac C. Robledo, David R. Peterson, Eric A. Day, Dean F. Hougen, and Jamie D. Barrett. "Mental Models and Creative Problem-Solving: The relationship of Objective and Subjective Model Attributes." *Creativity Research Journal*, 24(4):311-330, November 2012.
- Kimberly S. Hester, Issac C. Robledo, Jamie D. Barrett, David R. Peterson, Dean F. Hougen, Eric A. Day, and Michael D. Mumford. "Causal Analysis to Enhance Creative Problem-Solving: Performance and Effects on Mental Models." *Creativity Research Journal*, 24(2-3):115-133, April 2012.
- Issac C. Robledo, Kimberly S. Hester, David R. Peterson, Jamie D. Barrett, Eric A. Day, Dean F. Hougen, and Michael D. Mumford. "Errors and Understanding: The Effects of Error Management Training on Creative Problem-Solving." *Creativity Research Journal*, 24(2-3):220-234, April 2012.

Synergistic Activities

- Mentorship: Member, International Student Services Advisory Board, University of Oklahoma, 2009–present; Member, OU2Go Advisory Board, University of Oklahoma, 2009–2010; Faculty Mentor, NSF REU Site on Integrated Machine Learning Systems, 2008–2010; Faculty Mentor, Student Chapter of ACM, University of Oklahoma, 2007–2008; Chair Graduate Committee, School of Computer Science, University of Oklahoma, 2006–2007; Co-PI,

NSF REU Site on Embedded Machine Learning Systems, 2005–2007; Graduate Committee Member, School of Computer Science, University of Oklahoma, 2004–2007; Chair, Graduate Enrollment Management and Fellowship Committee, School of Computer Science, University of Oklahoma, 2004–present; Faculty Mentor, NSF Site on Human Technology Interaction, 2002–2005; Faculty Advisor, Robotics Club, University of Oklahoma, 2001–present.

- Outreach: Member, Steering Committee, "This View of Life": Darwin 2009 at OU, University of Oklahoma, 2008–2010; Planning Committee, Head Referee, and Judge Coordinator, Oklahoma FIRST Regional Robotics Competition, 2007–present; Chair, Publicity, Recruitment and Special Events Committee, School of Computer Science, University of Oklahoma, 2008–2009; Member, Publicity, Recruitment and Special Events Committee, School of Computer Science, University of Oklahoma, 2007–2008; Judge, Oklahoma Regional Tournament, Botball Program, KISS Institute for Practical Robotics, 2002–2003, 2007–2008; Panel Discussion Leader/Participant, Graduate School in Computer Science, Oklahoma Computing Conference, Computing Research Association Committee on the Status of Women in Computing, 2006; Judge, National Tournament, Botball Program, KISS Institute for Practical Robotics, 2002–2003; Instructor, Gifted and Talented Enrichment Classes–Robotics, Norman Public Schools & Precollegiate Programs, University of Oklahoma, 2002.
- Ethics: Investigator, "Development and Evaluation of a Work Practices Approach for Ethics Education in Science and Engineering," NSF Ethics Education in Science and Engineering (EESE) program, $209,776, October 2005–September 2008; Co-PI on Proposal, "Developing Mental Models to Enhance Ethical Decision Making," NSF Science of Science and Innovation Policy (SciSIP) program, $299,929, October 2008–October 2011.
- Writing: Panel Participant, Technical Writing for Theses and Dissertations, Second Annual Computer Science Research Conference, Computer Science Graduate Student Association, School of Computer Science, University of Oklahoma, 2006; Author (with Maria Gini), "The Assessment of Student Writing in Computer Science Classrooms," presented at the Fiftieth Annual Convention, Conference on College Composition and Communication, Workshop: Visible Results in WAC: Assessing Students and Programs, 1999; Author (with Maria Gini), " Writing for Computer Science Students," Research Report, Center for Interdisciplinary Studies of Writing, University of Minnesota, 1997.
- Service: Program Committee, Association for the Advancement of Artificial Intelligence, 2013; Associate Editor, Conference Editorial Board, International Conference on Robotics and Automation, Robotics and Automation Society, Institute of Electrical and Electronics Engineers, 2008–2009; Site Review Team Member, Alberta Ingenuity Centre for Machine Learning, 2007; Program Committee Member, IEEE International Conference on Robotics and Automation, 2006; Review Panel Member, NSF Integrative Graduate Education and Research Traineeship program, 2005; Program Committee Member, IASTED International Conference on Robotics and Applications, 2003–2004; Reviewer for journals including *Artificial Life*, *Autonomous Robots*, *Control and Intelligent Systems*, three *IEEE Transactions*, and the *Journal of Intelligent and Robotic Systems*; numerous conferences; and funding agencies including NSF, the Army Research Office, and US CRDF.

Recent Collaborators and Other Recent Affiliations

Collaborators (Co-authors listed above, foreign collaborators, and collaborators at the University of Oklahoma not listed.)

- Suleyman Karabuk. Zillow

Graduate and Post Doctoral Advisors

- Maria Gini, University of Minnesota
- Nikolaos P. Papanikolopoulos, University of Minnesota
- James Slagle, University of Minnesota (retired)

Thesis Advisor and Postgraduate-Scholar Sponsor

- Brent Eskridge, Southern Nazarene University
- Joshua Beitelspacher, National Instruments
- Sreedevi Chandrasekaran, Virginia Commonwealth University
- Pedro Diaz-Gomez, Cameron University
- Gerardo Gonzalez, Independent Software Consultant
- Shawn McCarroll, National Security Agency
- Cason Clagg, Independent Software Consultant
- Mark Woehrer, Federal Aviation Administration
- Christopher Fenner, Microsoft Corporation
- Syed Naveed Hussain Shah, Microsoft Corporation
- Benjamin P. Carlson, North Point Defense
- Ryan Ralston, K20 Center
- Joohee Suh, Microsoft Corporation
- Bryan Hoke, NextThought
- Jeremy Rand, NameCoin

A total of 10 MS and 7 PhD students have been advised.

B Simpsons Paradox in Flights Data Set [1]

Simpson's Paradox in Flights Data Set

AN001 Lerner Data Management

The document describes discovered instances of Simpson's paradox (SP) in an openly available Aviation data set. The inclusion of this document speaks to possible other instances in FAA and learner data. Discovering SP residing in data sets is a prudent step before any policy information decisions are recommended (Xu et al., 2018).

Data set

The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics tracks the on-time performance of domestic flights operated by large air carriers. Summary information on the number of on-time, delayed, canceled, and diverted flights is published in DOT's monthly Air Travel Consumer Report and in this dataset of 2015 flight delays and cancellations.

- https://www.kaggle.com/usdot/flight-delays/home

We use the following attributes:
AIRLINE: Name of the airline company
ORIGIN_AIRPORT: Airport the originated the delay
CANCELLED: Cancellation

Below are four discovered cases of SP:

Case 1:

2 AIRLINE: AA, UA
12 ORIGIN_AIRPORT: ATL, BOS, COS, DEN, IAH, JFK, LAX, MFE, PHL, ROC, SEA, SFO
Description about SP in this dataset:
AA(0.047) has a higher CANCELLED rate than UA(0.023) for the whole population in this dataset.
However, when we partition the dataset by airport, ATL airport has a reverse trend which is that AA in ATL(0.038) has a lower CANCELLED rate than UA in ATL(0.075). This is the SP.

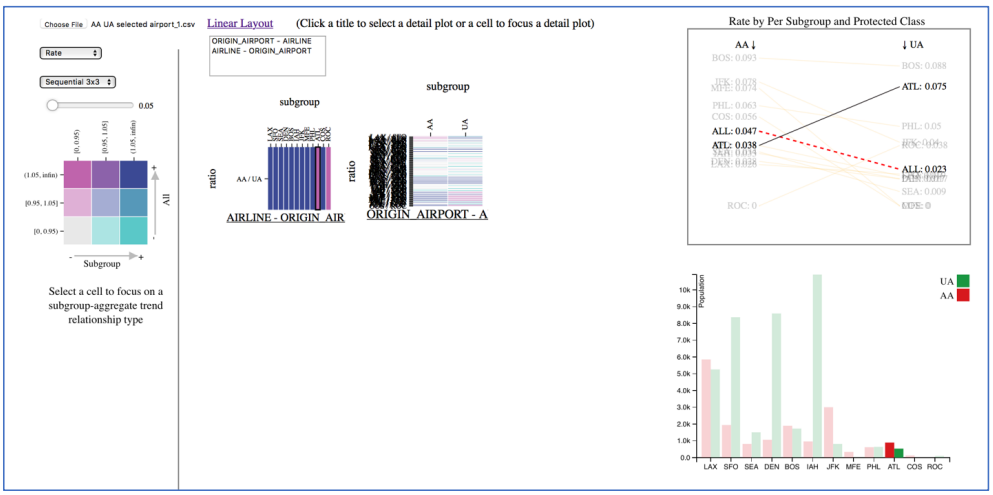

Figure 1: Case1

Case 2:

6 AIRLINE: AA, UA, DL, F9, B6, WN

12 ORIGIN_AIRPORT: ATL, BOS, COS, DEN, IAH, JFK, LAX, MFE, PHL, ROC, SEA, SFO

Description about SP in this dataset:

AA(0.047) has a higher CANCELLED rate than DL(0.022) for the whole population in this dataset.

However, when we partition the dataset by airport, BOS airport has a reverse trend which is that AA in BOS(0.093) has a lower CANCELLED rate than DL in BOS(0.1). This is the SP.

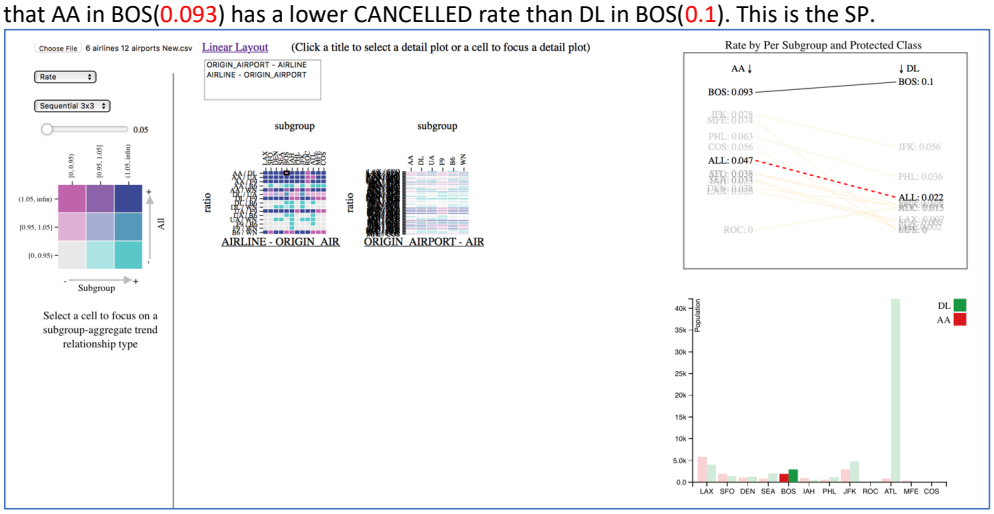

Figure 2: Case 2

Case 3:

6 AIRLINE: AA, UA, DL, F9, B6, WN

6 ORIGIN_AIRPORT: ATL, BOS, COS, DEN, IAH, JFK

Description about SP in this dataset (similar to case 2):

AA(0.064) has a higher CANCELLED rate than DL(0.024) for the whole population in this dataset.

However, when we partition the dataset by airport, BOS airport has a reverse trend which is that AA in BOS(0.093) has a lower CANCELLED rate than DL in BOS(0.1). This is the SP.

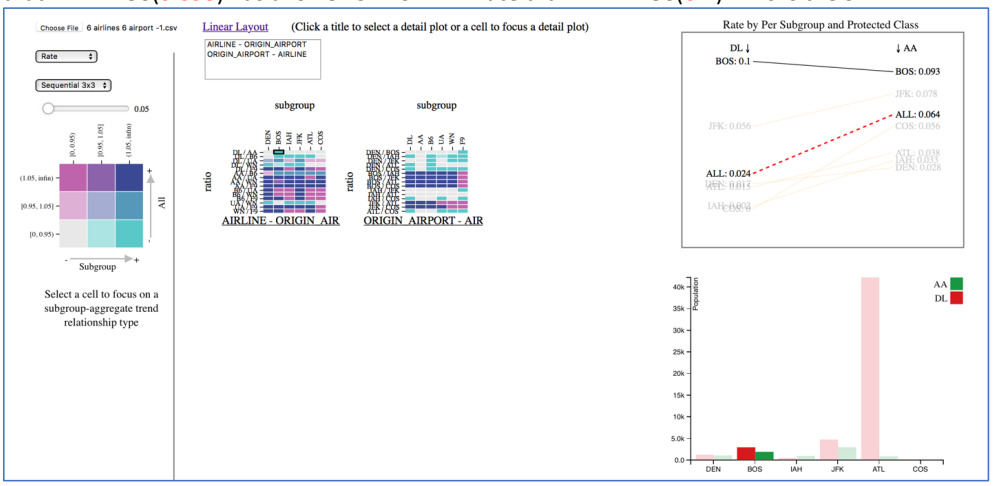

Figure 3: Case 3

Case 4:

6 AIRLINE: AA, UA, DL, F9, B6, WN

6 ORIGIN_AIRPORT: LAX, MFE, PHL, ROC, SEA, SFO

Description about SP in this dataset:

WN(0.032) has a lower CANCELLED rate than B6(0.044) for the whole population in this dataset.

However, when we partition the dataset by airport, SFO airport has a reverse trend which is that WN in SFO(0.038) has a higher CANCELLED rate than B6 in SFO(0.029). This is the SP.

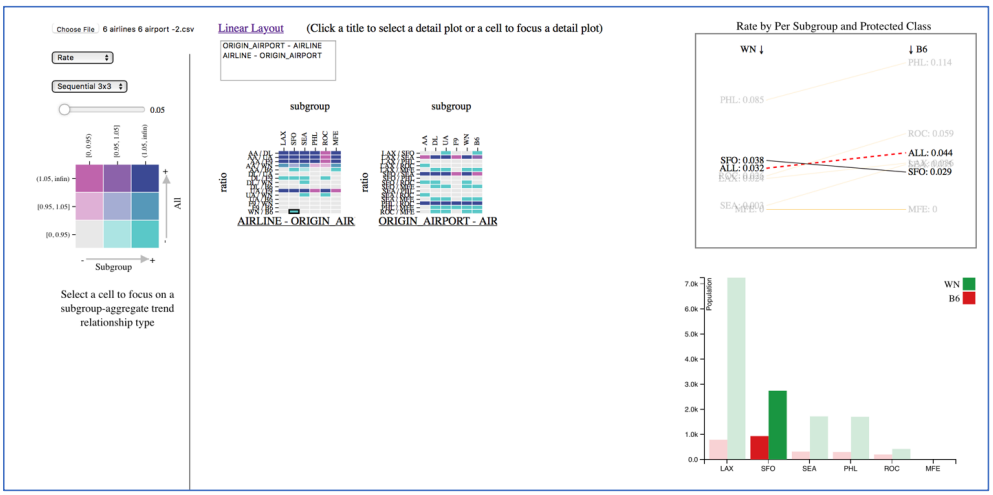

Figure 4: Case 4

C Accident and Delays

Accident and Delays

In this work we tried to find if there exists a correlation between delays and accidents. For this we have used two datasets, one for Delays and other for Accidents.

Delays:

- Initially, delays dataset has 5819079 observations and 31 features.
- Later, considered only observations having delay at the departure airport. Then delay dataset has 2125618 observations.
- There are 626 Unique number of Origin airports in delays dataset.
- For each airport calculated number of delays and % of delays that airport accounted for.

Accidents:

- Initially, accidents dataset has 81,013 observations and 31 features.
- There are 734 Unique number of Origin airports in accidents dataset.
- Later, considered only observations where country is “United States” and year is 2015. Since, Delays dataset is limited to United States and has observations for 2015. Also removed observations where airport code is missing. Then accident dataset is reduced to 901 observations.
- Later, for each airport calculated number of accidents and % of accidents that airport accounted for.

Merging:

- Merging Delays and Accidents dataset based on Airport code.
- Total number of airports present in both Delay and Accident dataset are 86
- Now, we have 86 airports and each airport has number of delays, % of delays airport accounted for (Delay factor), number of accidents, % of accidents airport accounted for (Accident factor).

Correlation plots:

- First tried to observe correlation between number of delays and number of accidents with the help of scatter plot.

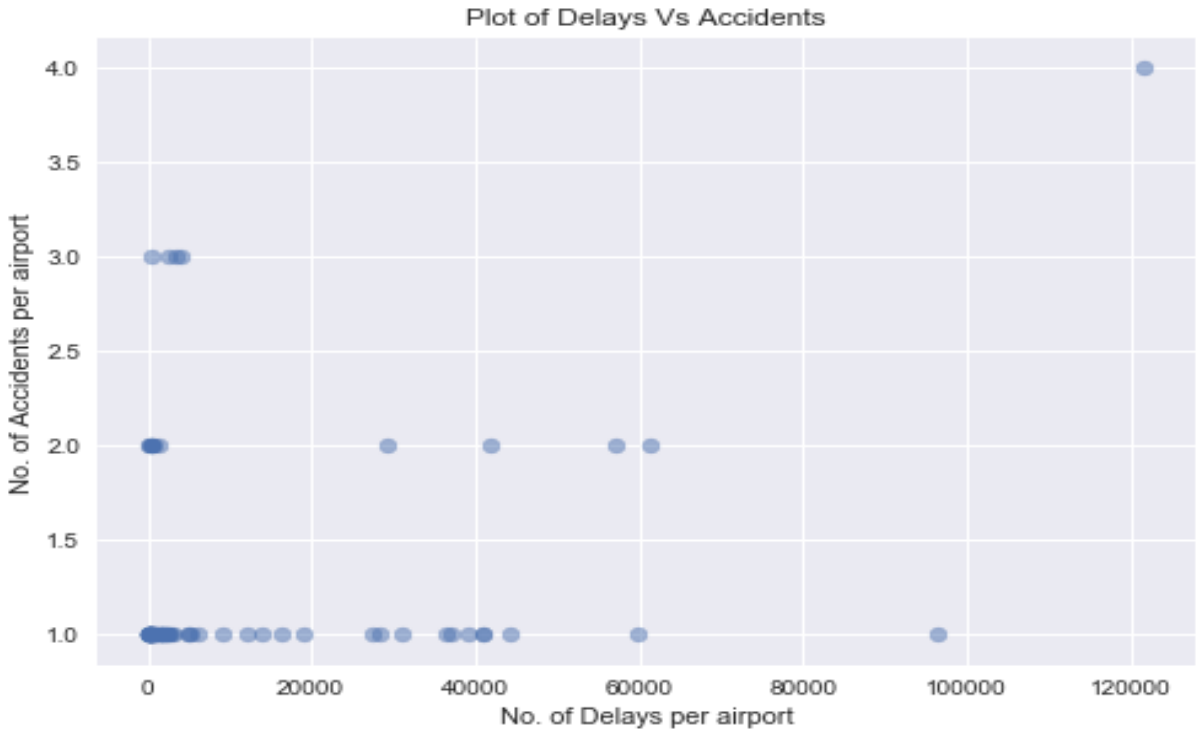

Fig1: Scatter plot between Number of accidents per airport Vs Number of Delays per airport.

Calculating Pearson and Spearman Correlations between number of delays and accidents:

- From Fig1, we couldn't observe correlations. So calculated Pearson and Spearman correlations.
- The Pearson correlation coefficient measures the linear relationship while spearman correla tion considers ranks. Both of these return correlations along with p-value.
- Initially, Pearson coefficients are (correlation = 0.3054, pvalue = 0.0042)
  Spearman coefficients are (correlation = 0.0878, pvalue=0.4214)
- From Fig.1 we could notice that there are two outliers.
- So, calculated z-score on number of delays. Observations with z > 3 are generally considered outliers.
- There are two outliers (with z > 3) Chicago (ORD) has a z-score of 5.11 and Dallas (DFW) has a z-score of 3.95.
- Calculated correlations after removing univariate outliers.
- After removing univariate outliers,
  Pearson coefficients are (correlation = 0.0737, pvalue = 0.5047)
  Spearman coefficients are (correlation = 0.0494, pvalue=0.6553)

Note: In above case we have observed correlation between No. of delays and No. of accidents for each possible airport. But, finding correlation between No. of delays and No. of accidents is a statistical error. Since, we are considering No. of delays and No. of accidents but not considering frequency of journeys that are causing delays and accidents.

So, it is better to find correlations between percentage of flights that are delayed and percentage of flights for which accidents occurred for all possible airports.

Note: We know that accident dataset will not have observations of all the airports in delay dataset. (Accidents mayn't occur for flights flying from all the airports). In this case accident count will be zero for all the airports that are absent in accident dataset and present in delay dataset. But accident count might go wrong if airport codes are not similar in both the datasets (We are not sure of this). So, we computed correlations for both the cases, i.e. also computed correlation by considering only airports that are common in both datasets.

1. **Computing correlation between percentage of flights delayed and percentage of flights met accident by considering only airports that are common in both datasets:**

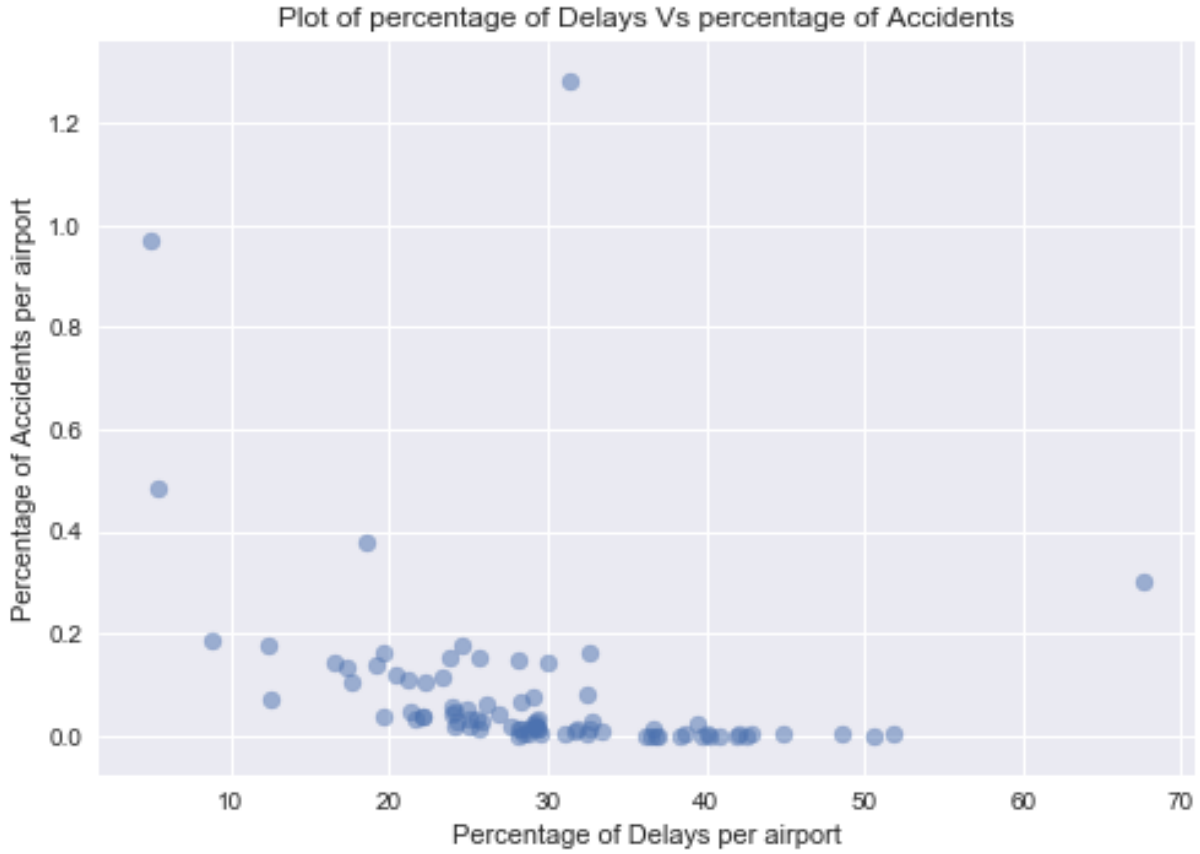

Fig2: Scatter plot between percentage of accidents Vs percentage of Delays by considering only airports that are common in both accident and delay dataset (86 airports).

Calculating Pearson and Spearman Correlations:

- From Fig2, we couldn't observe correlations. So calculated Pearson and Spearman correlations.
- Initially,
  Pearson coefficients are (correlation = -0.3031, pvalue = 0.0045)
  Spearman coefficients are (correlation = -0.7138, pvalue = 1.2135 e-14)
- From Fig.2 we could notice that there are few outliers.
- To remove outliers, we performed both univariate and multivariate outlier analysis.
- To remove outliers by univariate analysis, calculated z-score on percentage of delays and on percentage of accidents. Observations with z > 3 are generally considered outliers.
- There are total of three univariate outliers:
  Observation is represented as City, State (Airport Code) – (percentage of flights delayed, percentage of flights met accident), Number of flights departure from this airport.
  - Outliers in terms of Delay Percentage:
    1. Agana, GU (GUM) – (67.66, 0.29), 334
  - Outliers in terms of Accident Percentage:
    1. Mammoth Lakes, CA (MMH) – (31.41, 1.28), 156
    2. Vernel, UT (VEL) – (4.85, 0.97), 206
- Performed multivariate outlier analysis using mahanalobis distance and outliers detected are same as univariate analysis.

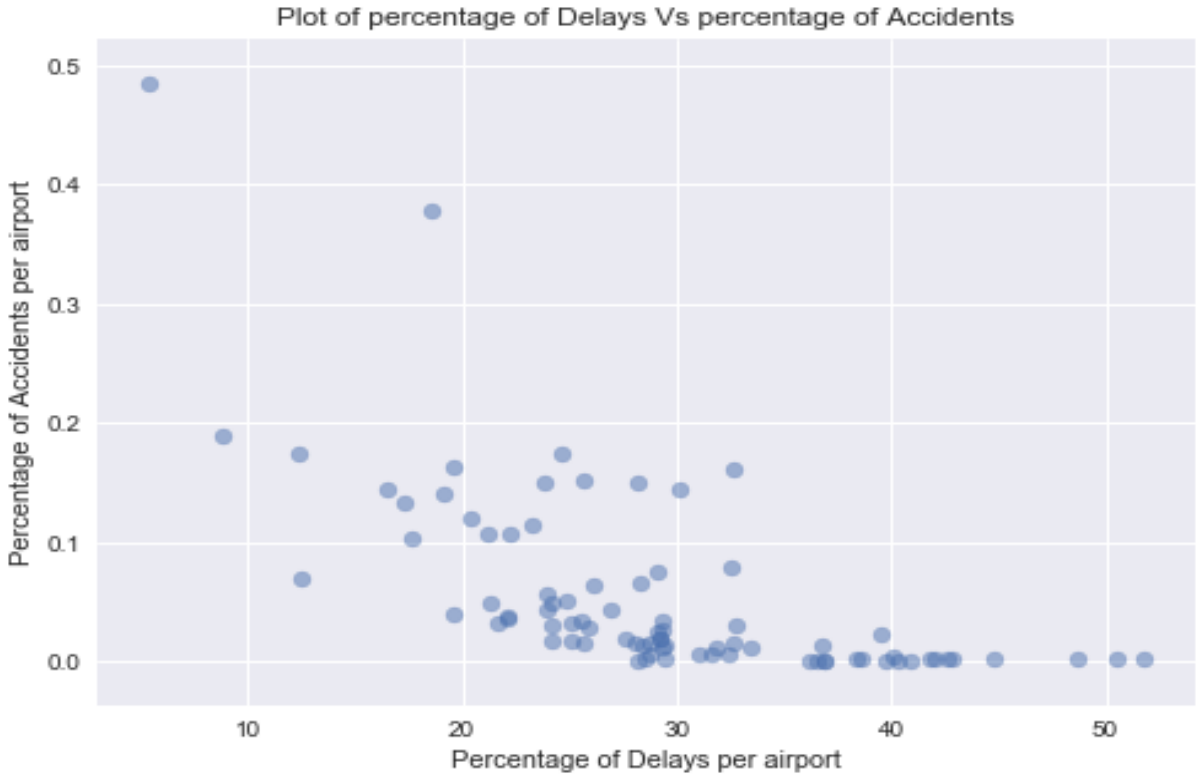

Fig3: After removing outliers, Scatter plot between percentage of accidents Vs percentage of Delays by considering only airports that are common in both accident and delay dataset.

- Calculated correlations after removing outliers.
- After removing outliers,

  Pearson coefficients are (correlation = -0.6344, pvalue = 1.1987 e-10)

  Spearman coefficients are (correlation = -0.7898, pvalue=0.6553 e-19)

2. Computing correlation between percentage of flights delayed and percentage of flights met accident by considering all airports that are in delay dataset.

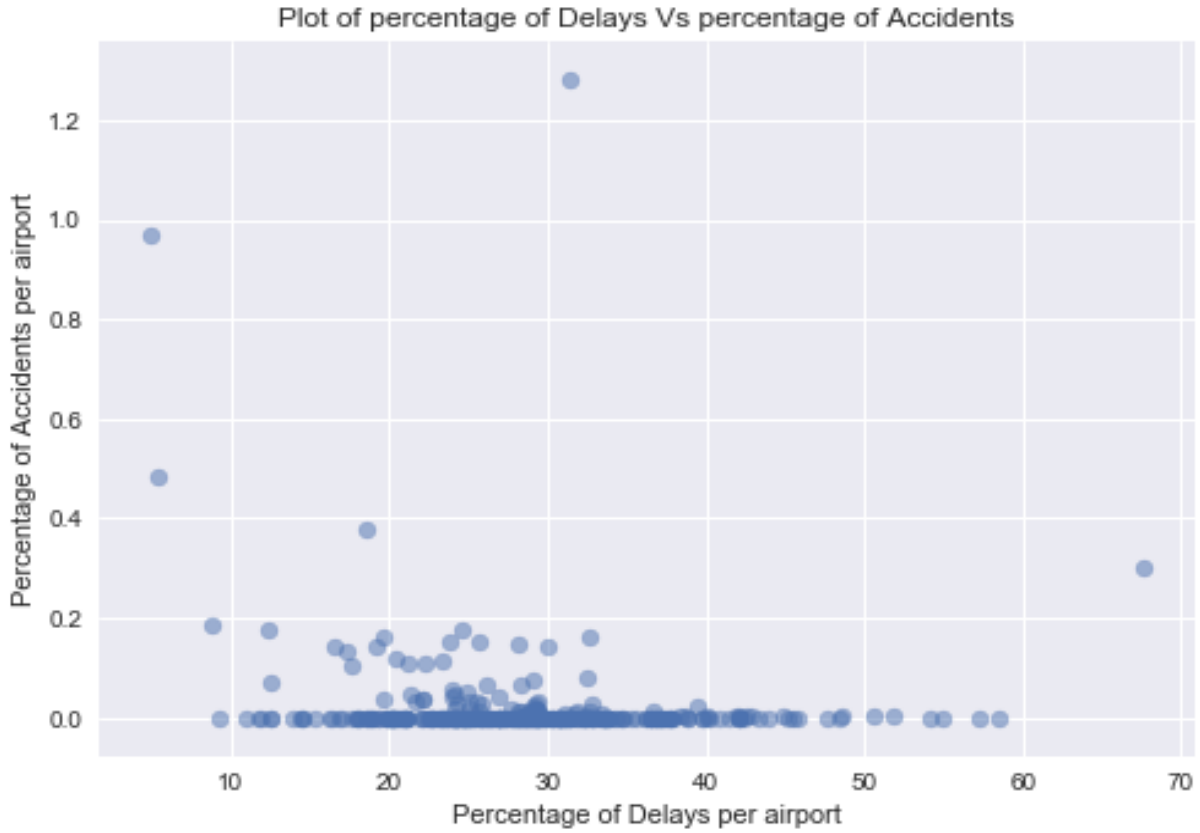

Fig4: Scatter plot between percentage of accidents Vs percentage of Delays by considering all airports that are in Delay dataset (322 airports).

Calculating Pearson and Spearman Correlations:

- From Fig4, we couldn't observe correlations. So calculated Pearson and Spearman correlations.
- Initially,

  Pearson coefficients are (correlation = -0.1472, pvalue = 0.0081)

  Spearman coefficients are (correlation = -0.0141, pvalue = 0.8001)
- From Fig.4 we could notice that there are few outliers.
- To remove outliers, we performed both univariate and multivariate outlier analysis.

- To remove outliers by univariate analysis, calculated z-score on percentage of delays and on percentage of accidents. Observations with z > 3 are generally considered outliers.
- There are total of eight univariate outliers:
  Observation is represented as City, State (Airport Code) – (percentage of flights delayed, percentage of flights met accident), Number of flights departure from this airport.
    - Outliers in terms of Delay Percentage:
        1. Adak, AK (ADK) – (57.29, 0), 96
        2. Gustavus, AK (GST) – (58.44, 0), 77
        3. Agana, GU (GUM) – (67.66, 0.29), 334
        4. Wilmington, DE (ILG) – (55, 0), 100
    - Outliers in terms of Accident Percentage:
        1. Moab, UT (CNY) – (5.33, 0.48), 206
        2. Pueblo, CO (PUB) – (18.56, 0.38), 264
        3. Mammoth Lakes, CA (MMH) – (31.41, 1.28), 156
        4. Vernel, UT (VEL) – (4.85, 0.97), 206
- Performed multivariate outlier analysis using mahanalobis distance and outliers detected are same as univariate analysis.

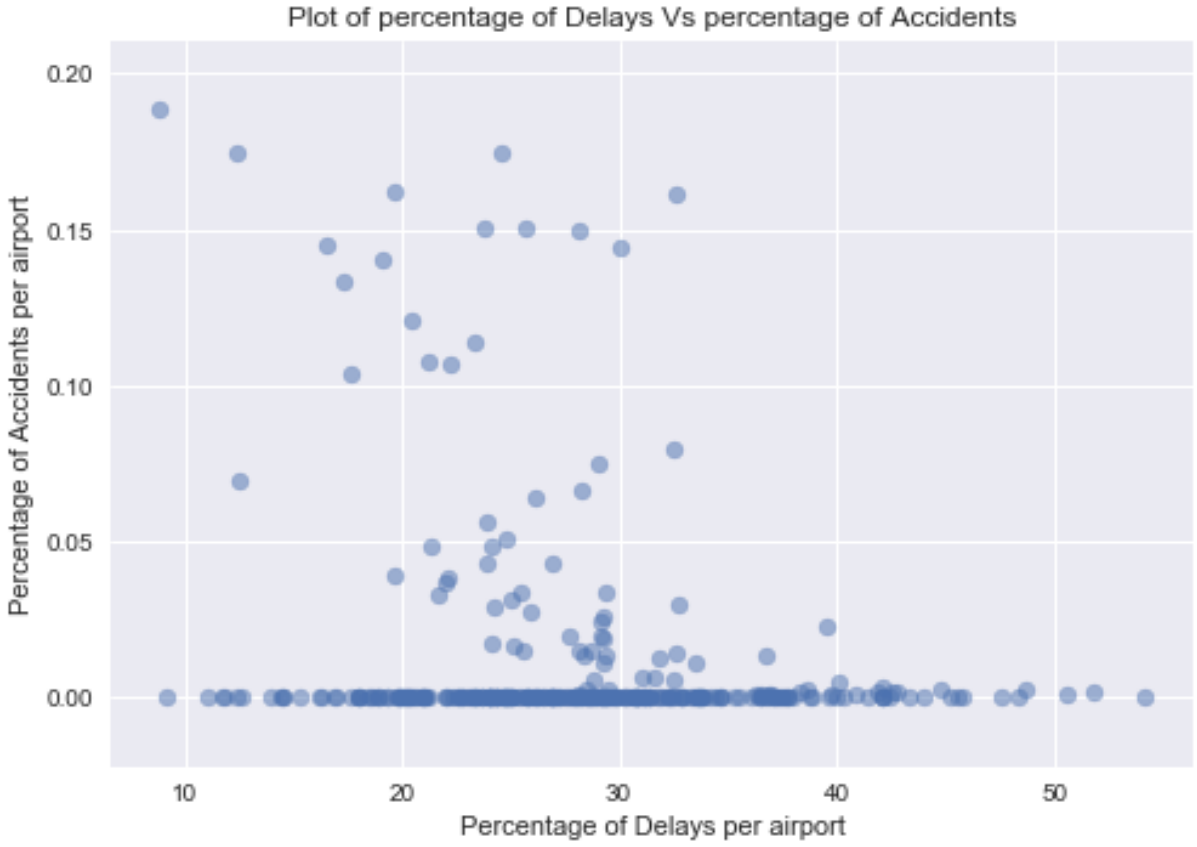

Fig5: After removing outliers, Scatter plot between percentage of accidents Vs percentage of Delays by considering all airports that are in Delay dataset.

- Calculated correlations after removing outliers.
- After removing outliers,

  Pearson coefficients are (correlation = -0.2250, pvalue = 5.7122 e-05)

  Spearman coefficients are (correlation = 0.011, pvalue= 0.8354)

- **Conclusion:**

There exists a significant negative correlation between percentage of flights that are delayed and percentage of flights that met accident. This suggests that airports with high number of delays resulted in lower accidents.

D Data Analysis on Delays and Cancellations

Data Analysis on Delays and Cancellations:

In this work data analysis is done on Flight Delays and Cancellations. The first dataset "airlines.csv" gives information about airlines and the second dataset "airports.csv" gives information about the airports used in flights.csv, flights.csv has data about delays and cancellations of flights in 2015. So, in this work we will be focused on flights.csv.

- Flights dataset has 5819079 observations and 31 features.
- **Airlines:**
  - Total number of unique flights used by all airlines in 2015 are 4898.

	AIRLINE	Total Journeys	Total Flights
0	American Airlines (AA)	725984	707
1	Alaska Airlines (AS)	172521	147
2	JetBlue Airways (B6)	267048	215
3	Delta Air Lines (DL)	875881	828
4	Atlantic Southeast Airlines (EV)	571977	390
5	Frontier Airlines (F9)	90836	63
6	Hawaiian Airlines (HA)	76272	50
7	American Eagle Airlines (MQ)	294632	203
8	Spirit Air Lines (NK)	117379	79
9	Skywest Airlines (OO)	588353	383
10	United Air Lines (UA)	515723	721
11	US Airways (US)	198715	351
12	Virgin America (VX)	61903	57
13	Southwest Airlines (MN)	1261855	704

Table1: Total number of journeys and Flights owned by different airlines in 2015

- **Day of a week:**

	DAY_OF_WEEK	Number of Journeys
0	Monday	865543
1	Tuesday	844600
2	Wednesday	855897
3	Thursday	872521
4	Friday	862209
5	Saturday	700545
6	Sunday	817764

Table2: Frequency of flights for different days of a week in 2015

- We could notice that number of flights operated on weekends is less when compared to number of flights operated on weekdays. While, Saturday recording least number of flights departed. This can be observed from bar graph below.

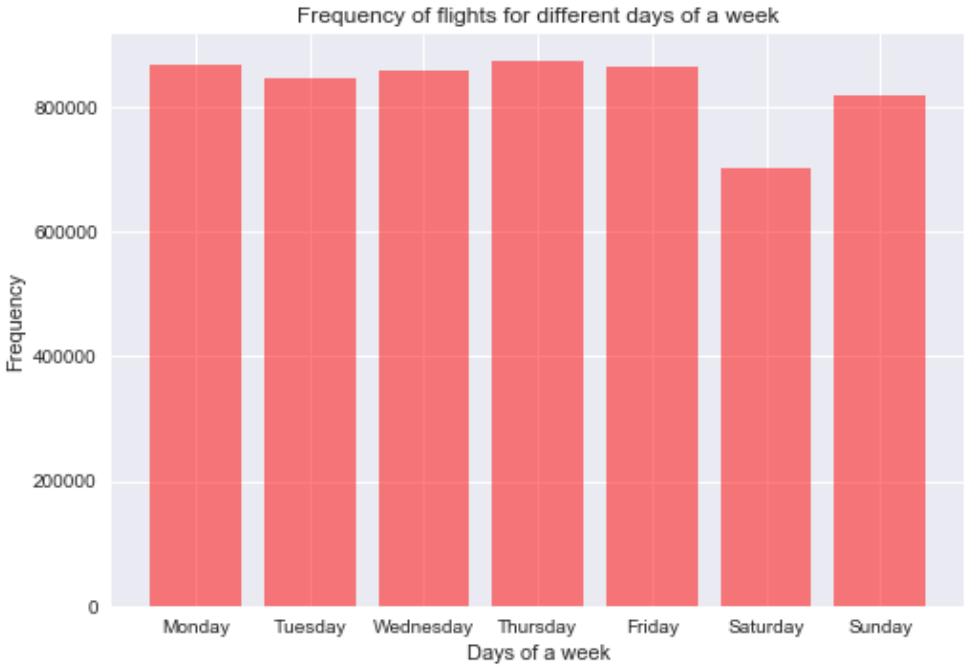

Fig1: Plot of frequency of flights for different days of a week.

- **Distance:**
  - Maximum distance is between Honolulu (HNL) and New York (JFK) and the distance is 4983 miles.
  - Minimum distance recorded is between Newark Liberty International Airport (EWR) and New York (JFK) and the distance is 21 miles, but this journey occurred only once in the dataset and this journey is a cancelled trip.
  - Minimum distance that flights regularly fly is between Gustavas – AK (GST) and Juneau – AK (JNU) and the distance is 41 miles.
  - Spread of Distance between departure and arrival airports is shown with the help of histogram below:

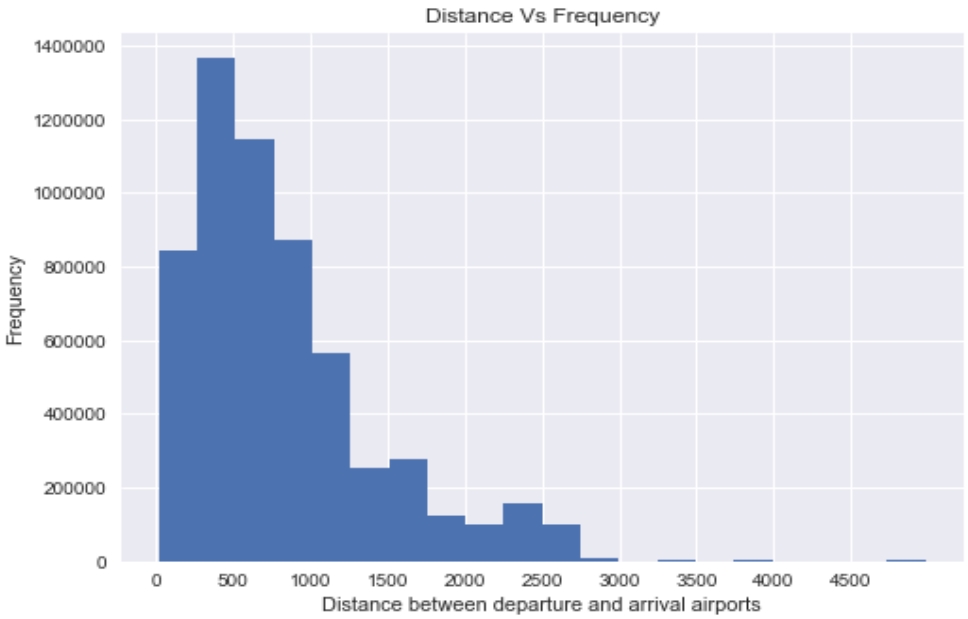

Fig2: Histogram of distance between departure and arrival airports

- **Arrival Delay:**

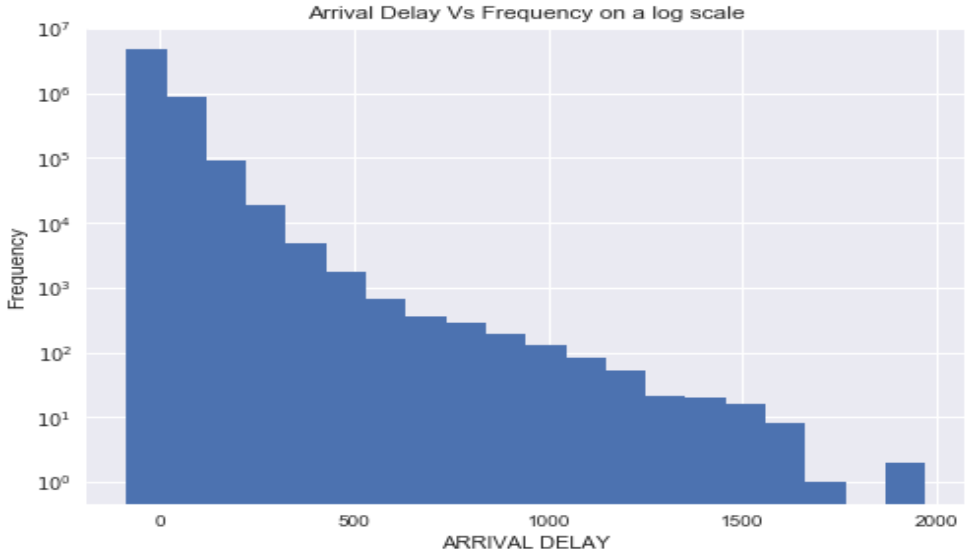

Fig3: Histogram of Arrival Delay and frequency represented on a log scale.

- Minimum Delay recorded is -87 minutes and maximum delay recorded is 1971 minutes.
- To respond to skewness in arrival delays, plot is represented in log scale.

- **Flights that are cancelled:**
  - Number of flights cancelled in 2015 are 89884 and percentage of flights cancelled in 2015 are 1.544 %

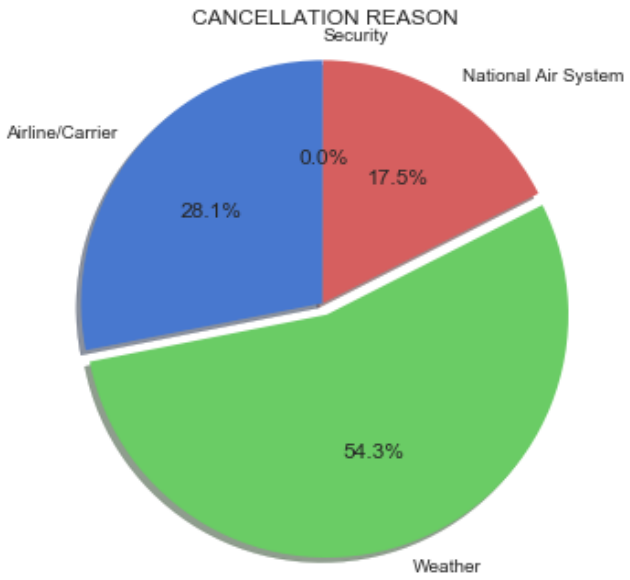

Fig4: Pie chart displaying percentage each factor caused towards cancellation.

  - From Fig.4. we can conclude that more than half of cancellations are done due to weather and a very few cancellations are done due to security reasons.

- **Flights that are diverted:**
  - Flights that are Diverted in 2015 are 15187 and percentage of flights diverted in 2015 are 0.26 %

- **Flights that are delayed:**
  Note:
  1. *Delays referred in dataset are referred to arrival delay but not departure delay.*
  2. *In this section we will be using dataset with only observations having delay.*

- For each delay of an aircraft in the dataset cause of delay is mentioned.
- Total number of significant delays in 2015 are 1063439 and percentage of flights with delay are 18.275 %.

	Departure_Mean	Departure_Median	Number of Delays	Arrival_Mean	Arrival_Median	%Each month Accounted for
MONTH						
JANUARY	50.955561	35.0	95951.0	56.694167	37.0	9.022708
FEBRUARY	52.678637	35.0	95179.0	59.210498	38.0	8.950114
MARCH	52.768921	35.0	95452.0	56.980577	36.0	8.975785
APRIL	51.462303	34.0	82247.0	56.163848	35.0	7.734059
MAY	56.858040	38.0	89645.0	60.654348	38.0	8.429727
JUNE	60.580999	41.0	115742.0	63.683192	41.0	10.883746
JULY	56.892722	40.0	107627.0	59.339236	39.0	10.120656
AUGUST	57.172792	40.0	94113.0	59.677951	39.0	8.849873
SEPTEMBER	50.641365	34.0	60061.0	54.804066	35.0	5.647809
OCTOBER	51.498078	34.0	60079.0	55.181178	33.0	5.649501
NOVEMBER	53.225588	34.0	70571.0	57.257287	34.0	6.636112
DECEMBER	59.487135	39.0	96772.0	62.616418	38.0	9.099911

Table3: Considering all types of Delays during departure and arrival for different months.

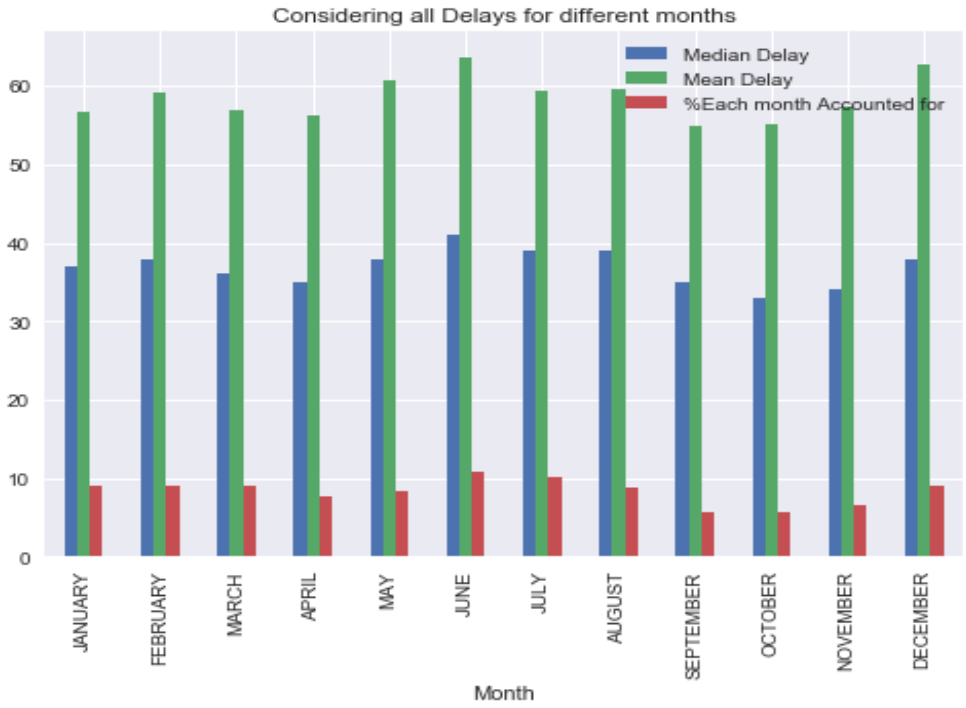

Fig5: Considering all types of arrival delays for different months (values from Table 3)

- Five reasons mentioned in the dataset that caused delay are:
    1.Weather
    2. Air system
    3. Security
    4. Airline
    5. Late aircraft.

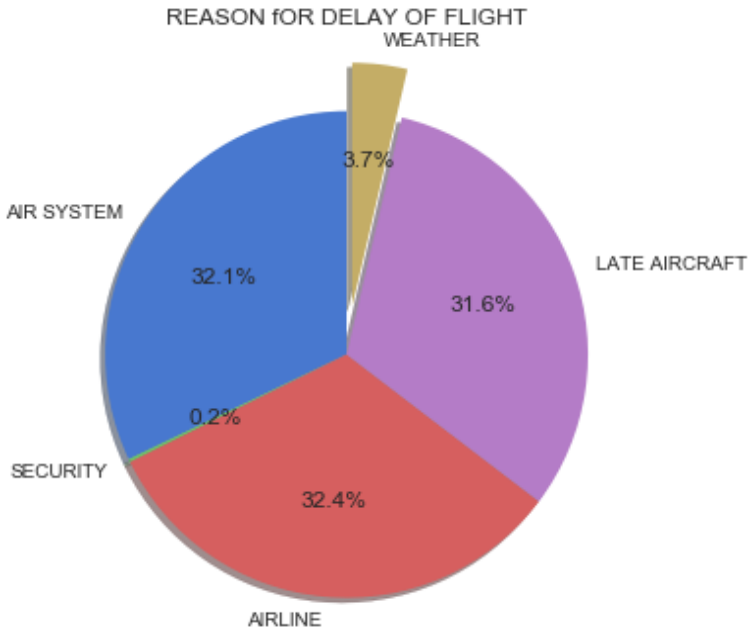

Fig6: Pie chart displaying percentage each factor caused towards delay.

- From figure 4 and 6 we could notice that more than half of flights (54.3%) of flights got cancelled due to weather, where as only 3.7% of flights got delayed due to weather.

- **Delays due to weather:**
    - Total number of weather delays in 2015 are 64,716.
    - Below table (4) and figure (6) gives more details about delays due to weather.

	Number of Delays	Mean Delay	Median Delay	%Each month Accounted for
MONTH				
JANUARY	6383.0	41.216826	22.0	9.863094
FEBRUARY	8940.0	46.002796	21.0	13.814204
MARCH	4510.0	50.700887	24.0	6.968910
APRIL	4978.0	44.513861	26.0	7.692070
MAY	6259.0	53.771369	30.0	9.671488
JUNE	7501.0	50.753100	29.0	11.590642
JULY	5601.0	41.504731	26.0	8.654738
AUGUST	5823.0	45.645200	29.0	8.997775
SEPTEMBER	3246.0	42.219039	26.0	5.015761
OCTOBER	2106.0	51.863723	29.0	3.254218
NOVEMBER	3456.0	54.094329	25.0	5.340256
DECEMBER	5913.0	55.284627	24.0	9.136844

Table:4 Stats regarding weather delays (arrival delays) for different months

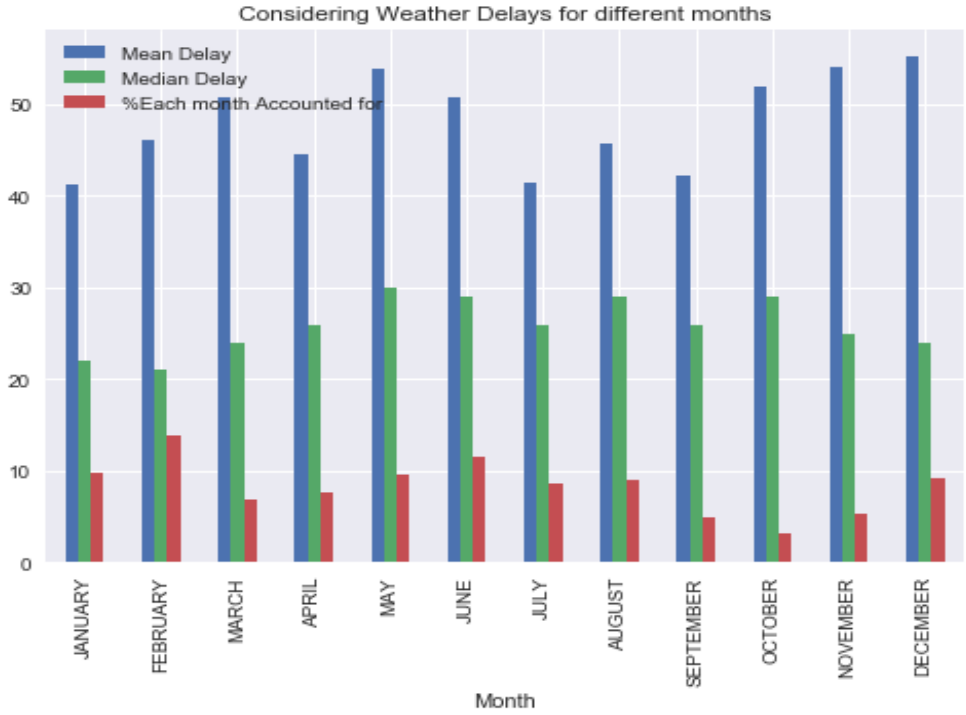

Fig7: Considering weather delays for different months (values from Table 4)

- **Operation of flights:**

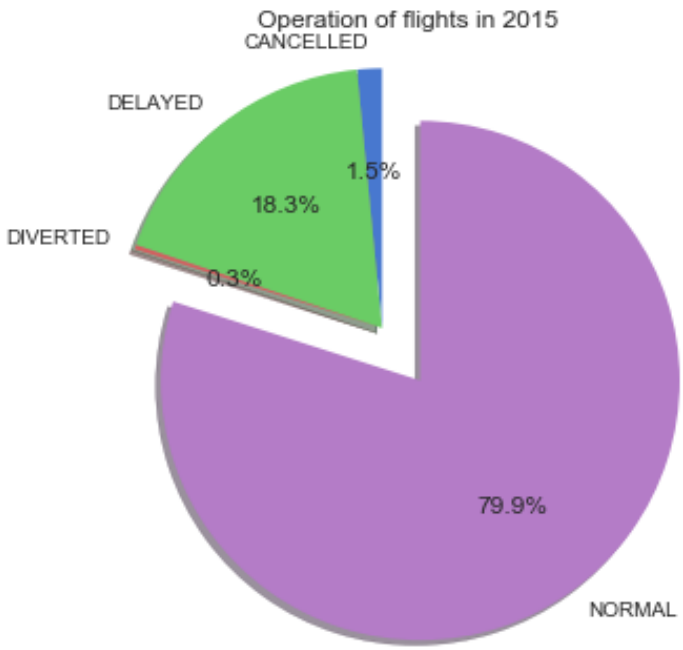

Fig8: Pie chart displaying percentage of operation of flights.

- **Arrival and Departure delays for different airports:**
  1. In this section also we will be using dataset with only observations having delay.
  2. There are total 628 unique airports in Flights (delays) dataset, but only information about 322 airports is given in airports dataset. So, removed observations with airport codes that are not mentioned in airports dataset. Surprising all the data of October doesn't have proper airport codes and all the observations of remaining months have proper airport codes.
  3. City name is used instead of airport codes. They are uniquely distinguished by:
     - If there are more than one city with same name in different states, then they are distinguished by having state code later to city name. Ex: Albany (GA), Albany (NY)
     - If there are multiple airports in same city, then they are distinguished by having their airport code later to the city name. Ex: Chicago (MDW), Chicago (ORD)
  4. Some details regarding delays at the departure airports (see table:5 and Fig:9) and delays at the arrival airports (see table:6 and Fig:10) are given below.

	Number of Delays	Departure Mean	Departure Median
St_City			
Chicago (ORD)	66663.0	58.561271	42.0
Atlanta	56462.0	55.035387	38.0
Dallas-Fort Worth	50478.0	54.451523	39.0
Denver	43331.0	52.170317	37.0
Los Angeles	40281.0	51.368089	36.0
Houston (IAH)	30690.0	55.238938	38.0
San Francisco	29534.0	58.015711	41.0
Phoenix	27427.0	47.305101	33.0
Las Vegas	27225.0	54.042975	37.0
New York (LGA)	22709.0	58.541635	40.0

Table5: stats regarding departure delay for the airports with large number of departure delays

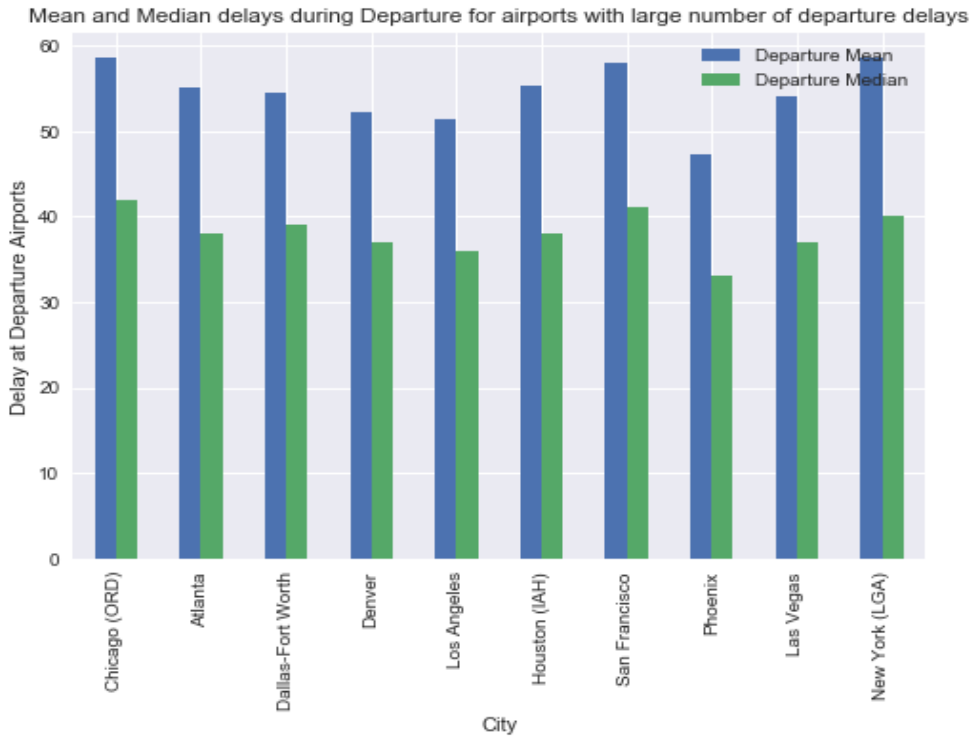

Fig9: Grouped bar chart for mean and median departure delay for the airports with large number of departure delays (Values from table 5)

	Number of Delays	Arrival Mean	Arrival Median
DEST_City			
Chicago (ORD)	58887.0	69.494999	42.0
Atlanta	52078.0	65.797515	38.0
Dallas-Fort Worth	44162.0	69.265794	39.0
Los Angeles	43453.0	54.805284	35.0
Denver	38023.0	61.079636	38.0
San Francisco	32984.0	63.219773	43.0
Houston (IAH)	27995.0	63.627934	38.0
New York (LGA)	24819.0	63.567428	43.0
Phoenix	24615.0	52.861101	33.0
Las Vegas	24496.0	54.383328	36.0

Table6: stats regarding arrival delay for the airports with large number of arrival delays

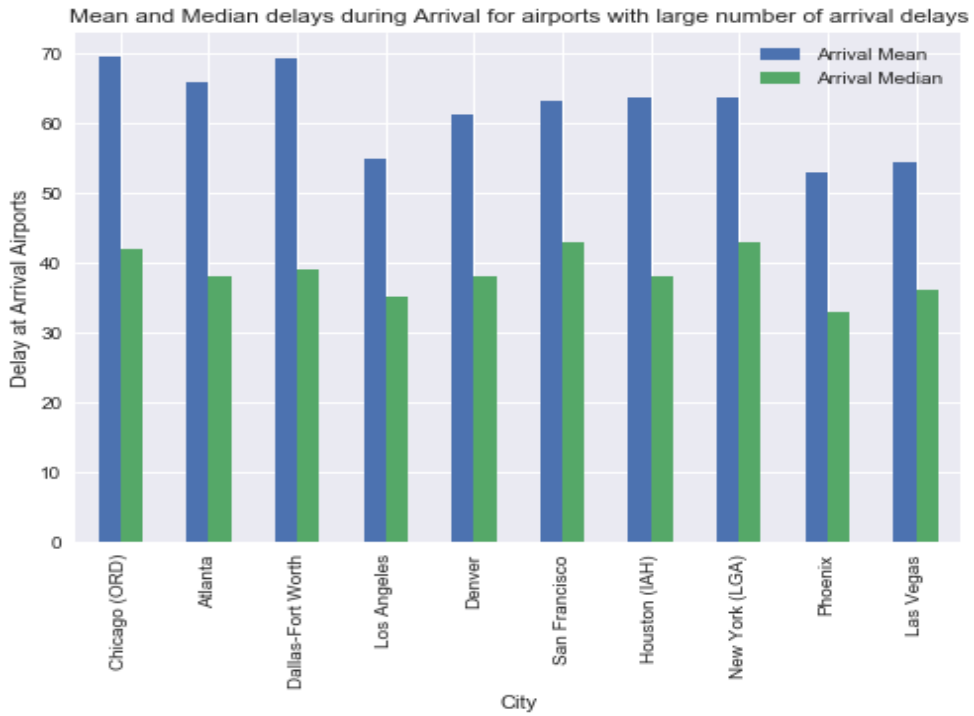

Fig10: Grouped bar chart for mean and median arrival delay for the airports with large number of arrival delays (Values from table 6)

- **Busiest Airports based on flights departing in 2015:**
  Note: These conclusions are based on data of 11 months (excluding October), since dataset d oesn't have proper airport codes for the entire month of October.

	City	Number of Journeys as departure airport	Frequency Percent
0	Atlanta	343506	6.566574
1	Chicago (ORD)	276554	5.286697
2	Dallas-Fort Worth	232647	4.447356
3	Denver	193402	3.697136
4	Los Angeles	192003	3.670392
5	Phoenix	145552	2.782420
6	San Francisco	145491	2.781254
7	Houston (IAH)	144019	2.753115
8	Las Vegas	131937	2.522151
9	Minneapolis	111055	2.122964

Table7: Stats regarding top 10 busiest airports in US in 2015

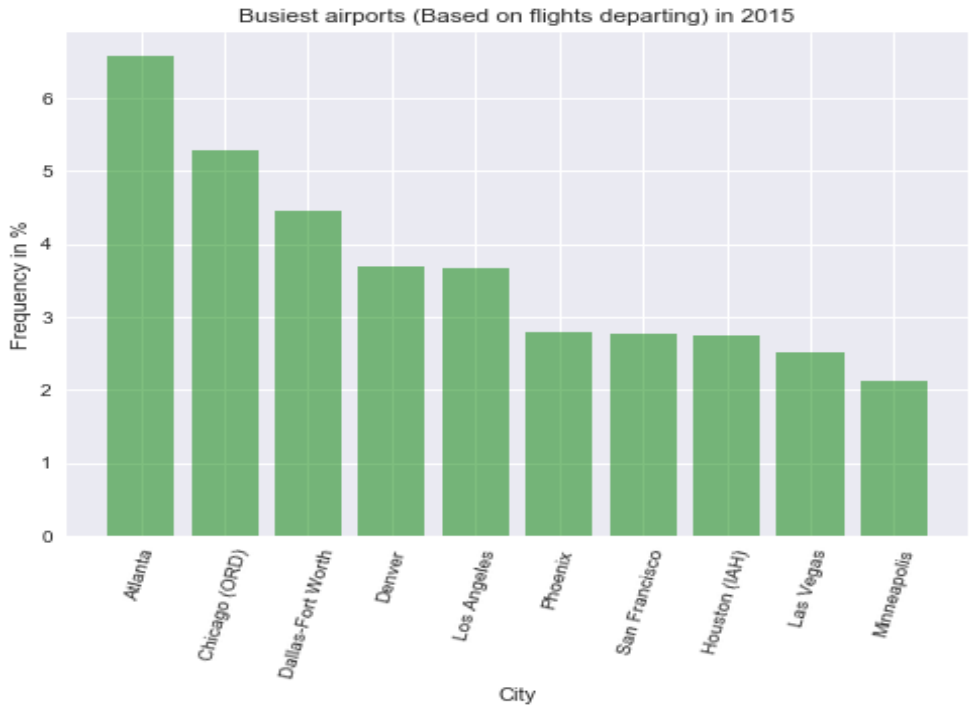

Fig11: Bar chart for busiest airports w.r.t. percentage of flights departed (values from Table 7)