Ανάλυση Δεδομένων καθυστερήσεων πτήσεων για το έτος 2015 στις Ηνωμένες Πολιτείες Αμερικής

Επιχειρηματική Ευφυΐα και Ανάλυση Μεγάλων δεδομένων

Υποβληθείσα στον καθηγητή:

Χατζηαντωνίου Δαμιανό

Δημήτρης Μπουρης: 8190119
Φιλιππος Πριόβολος: 8190147

Περιεχόμενα

Σκοπός εργασίας

Σκοπός της συγκεκριμένης εργασίας είναι η διενέργεια ενός ολοκληρωμένου data analysis task. Πιο συγκεκριμένα, έπρεπε να βρεθεί ένα μεγάλο data set, το οποίο θα πρέπει να καθαριστεί και να εισαχθεί σε μία αποθήκη δεδομένων. Στην συνέχεια, να δημιουργηθεί ένας κύβος δεδομένων με διάφορες μετρικές και να χρησιμοποιηθεί ένα εργαλείο οπτικοποίησης (Tableau ή Power BI) για να δημιουργηθούν διάφορες περιπτώσεις οπτικοποίησης δεδομένων. Τέλος, τα δεδομένα της αποθήκης να χρησιμοποιηθούν για κάποιες λειτουργίες εξόρυξης δεδομένων, όπως για παράδειγμα κατηγοριοποίηση, κανόνες συσχέτισης, συσταδοποίηση, κ.ο.κ. χρησιμοποιώντας μεθόδους και μοντέλα εμπορικού συστήματος ή ενός open-source εργαλείου. Στο παρακάτω report, αναλύονται τα βήματα και οι πρακτικές που ακολουθήθηκαν για την υλοποίηση της εργασίας.

Περιγραφή Dataset

Στα πλαίσια της συγκεκριμένης εργασίας, το dataset το οποίο επιλέγχθηκε ονομάζεται **2015 Flight Delays and Cancellations.** Προέρχεται από τον ιστότοπο Kaggle και έχει δημοσιοποιηθεί απο το U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics. Πιο συγκεκριμένα, περιέχει πτήσεις εσωτερικού (USA) που εκτελούνται κυρίως από μεγάλους αερομεταφορείς. Περιέχει πληροφορίες σχετικά με τον αριθμό των έγκαιρων, καθυστερημένων, ακυρωμένων και εκτροπής πτήσεων δημοσιεύονται στη μηνιαία Έκθεση Καταναλωτών αεροπορικών ταξιδιών της DOT για την ετήσια περίοδο του 2015. Κύριοι στόχοι σχετικά με την ανάλυση των συγκεκριμένων δεδομένων είναι εύρεση απαντήσεων για τα παρακάτω ερωτήματα με την χρήση κυρίως των δεδομένων καθυστέρησης:

1. Ποια είναι η καλύτερη αεροπορική εταιρεία ως προς καθυστερήσεις και ακυρώσεις πτήσεων
2. Ποιο είναι το καλύτερο αεροδρόμιο να προσγειώνεται κάποιος
3. Ποιο είναι το καλύτερο αεροδρόμιο για απογείωση
4. Ποιος είναι ο καλύτερος μήνας για πτήσεις
5. Ποια είναι η καλύτερη ημέρα της εβδομάδας
6. Ποιές ώρες είναι οι καλύτερες τόσο για απογείωση όσο και για προσγείωση
7. Πόσο σημαντικός είναι ο παράγοντας του καιρού για μια πτήση
8. Ποιοί είναι οι σημαντικότεροι παράγοντες που προκαλούν καθυστερήσεις πτήσεων
9. Κατά πόσο χαρακτηριστικά του αεροπλάνου επηρεάζουν τον χρόνο άφιξης του

Για την εύρεση απαντήσεων στα παρακάτω ερωτήματα χρησιμοποιήθηκαν όλα τα αρχεία που παρέχονταν στο συγκεκριμένο dataset. Αναλυτικότερα, τα δεδομένα ήταν οργανωμένα σε 3 διαφορετικά αρχεία. Το **flights.csv**, **airport.csv** και **airline.csv.**
Το πρώτο αρχείο (flights.csv) περιέχει αναλυτικά στοιχεία για 5.819.079 πτήσεις που πραγματοποιήθηκαν το 2015 καθώς και 31 πεδία που τις περιγράφουν. Τα σημαντικότερα πεδία του συγκεκριμένου αρχείου είναι:

- **FLIGHT_NUMBER**: μοναδικός κωδικός πτήσης
- **TAIL_NUMBER**: αναγνωριστικός κωδικός αεροσκάφους
- **YEAR, MONTH, DAY, DAY_OF_WEEK**: πεδία που περιγράφουν αναλυτικά την ημέρα αναχώρησης της κάθε πτήσης
- **AIRLINE**: μοναδικός κωδικός αεροπορικής εταιρείας (2 χαρακτήρες)
- **ORIGIN - DESTINATION AIRPORT**: μοναδικός κωδικός αεροδρομίου αναχώρησης και προορισμού (3 χαρακτήρες)
- **SCEDULED_DEPARTURE**: ώρα προγραμματισμένης αναχώρησης
- **DEPARTURE_TIME**: πραγματική ώρα αναχώρησης
- **DEPARTURE_DELAY**: καθυστέρηση αναχώρησης
- **SCHEDULED_TIME - DISTANCE:** προγραμματισμένη διάρκεια πτήσης και απόσταση
- **SCHEDULED_ARRIVAL**: προγραμματισμένη ώρα άφιξης στην πύλη του αεροδρομίου προορισμού
- **ARRIVAL_TIME**: πραγματική ώρα άφιξης στην πύλη του αεροδρομίου προορισμού
- **DIVERTED - CANCELED:** boolean τιμές που προσδιορίζουν αν η πτήση καθυστέρησε ή προσγειώθηκε σε άλλο αεροδρόμιο
- Τέλος, παρέχονται και τιμές που προσδιορίζουν την αιτία της καθυστέρησης (αν υπήρχε)
    - **AIR_SYSTEM_DELAY**: καθυστέρηση σε λεπτά που προέκυψε από το σύστημα ελέγχου της εναέριας κυκλοφορίας του αεροδρομίου προσγείωσης ή απογείωσης
    - **SECURITY_DELAY**: καθυστέρηση σε λεπτά που οφείλεται στον έλεγχο των επιβατών
    - **AIRLINE_DELAY**: καθυστέρηση σε λεπτά που οφείλεται στην αεροπορική εταιρία (αεροσυνοδοί και προμήθειες)
    - **LATE_AIRCRAFT_DELAY**: καθυστέρηση λόγω της αργοπορημένης αφιξης του αεροπλάνου λόγω κάποιας προηγούμενης πτήσης
    - **WEATHER_DELAY**: καθυστέρηση εξαιτίας καιρικών συνθηκών

Σημείωση: οι τιμές SCHEDULED_DEPARTURE και SCHEDULED_ARRIVAL αφορούν την ώρα αναχώρησης και άφιξης από την πύλη του αεροδρομίου αναχώρησης και προσγείωσης αντίστοιχα.

Συνεχίζοντας, το αρχείο airport.csv περιέχει δεδομένα που περιγράφουν τα αεροδρόμια των ΗΠΑ. Περιέχονται συνολικά 322 αεροδρόμια και 7 πεδία περιγραφής τους τα οποία είναι:

- **IATA_CODE**: μοναδικός αναγνωριστικός κωδικός αεροδρομίου
- **AIRPORT**: πλήρης ονομασία του αεροδρομίου
- **CITY-STATE-COUNTRY**: πόλη, πολιτεία και χώρα που βρίσκεται το αεροδρόμιο
- **LATITUDE**: γεωγραφικό μήκος
- **LONGITUDE**: γεωγραφικό πλάτος

Τέλος, το αρχείο airline.csv περιέχει δεδομένα για τις αεροπορικές εταιρείες που πραγματοποίησαν τις συγκεκριμένες πτήσεις. Πιο συγκεκριμένα, περιέχονται 14 αεροπορικές εταιρείες που περιγράφονται με 2 πεδία:

- **IATA_CODE**: μοναδικός αναγνωριστικός κωδικός αεροπορικής εταιρείας

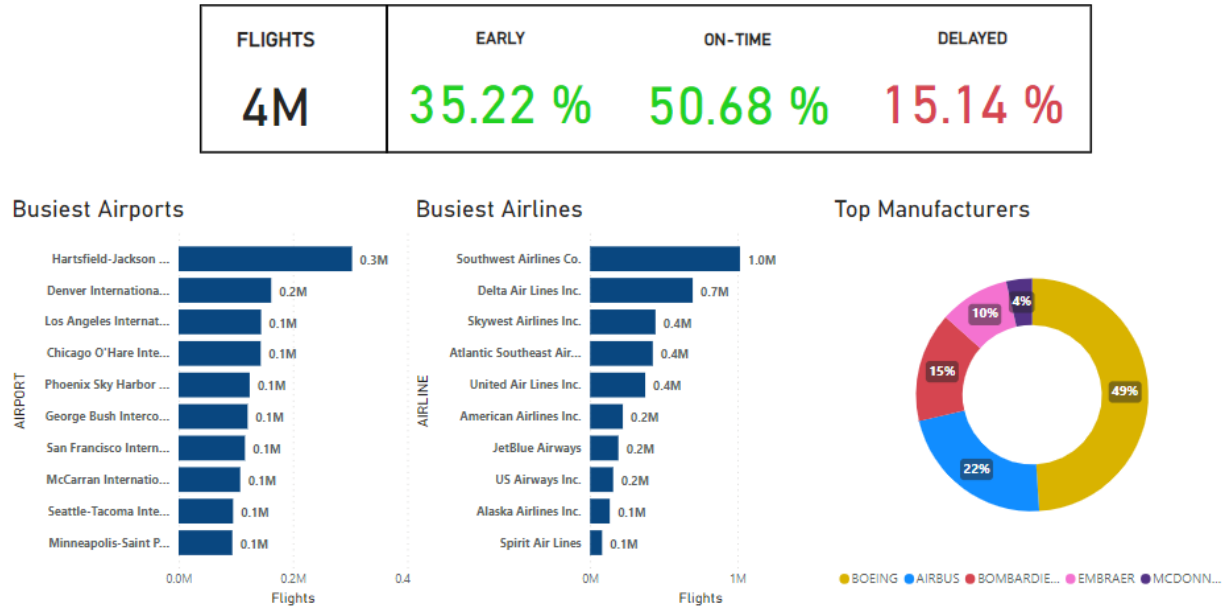
- **AIRLINE**: πλήρης ονομασία της εταιρείας

Στο παραπάνω dataset, προστέθηκαν και πληροφορίες σχετικά με το μοντέλο του αεροπλάνου. Το σκεπτικό πίσω απο αυτήν την προσθήκη είναι η διερεύνηση του κατά πόσο το αεροπλάνο αποτελεί σημαντικό παράγοντα που συμβάλει στην καθυστερημένη άφιξή του. Πιο συγκεκριμένα, τα αναλυτικά δεδομένα αντλήθηκαν από την Federal Aviation Administration και την βάση δεδομένων Aircraft Registration όπου δεδομένα σχετικά με την κατασκευή, συντήρηση και αναβάθμιση κάθε αεροσκάφους για κάθε χρόνο. Για τις ανάγκες της εργασίας αξιοποιήθηκε η βάση για το έτος 2015.
Τα πεδία που χρησιμοποιήθηκαν είναι:

- **N-Number:** μοναδικός κωδικός αναγνώρισης αεροσκάφους
- **AircraftMFRModelCode**: κωδικός αναγνώρισης αεροσκάφους κατασκευαστικής εταιρείας
- **YearMfr**: χρονολογία κατασκευής
- **TypeArcft**: τύπος αεροσκάφους
- **Mfr_name**: όνομα κατασκευαστικής εταιρείας
- **Model_name**: όνομα μοντέλου αεροσκάφους
- **Num_engines**: αριθμός κινητήρων
- **Num_passengers**: μέγιστος αριθμός επιβατών

Για τον περαιτέρω εμπλουτισμό του dataset, θα μπορούαν να προστεθούν επιπλέον και δεδομένα σχετικά με τις καιρικές συνθήκες. Προκύπτει όμως πως η διαδικασία συλλογής δεδομένων καιρού για 340 διαφορετικά αεροδρόμια με ακρίβεια ώρας είναι αρκετά περίπλοκη. Επιπλέον, σύμφωνα με το US Bureau of Transportation Statistics, μόλις το 5% των συνολικών πτήσεων για το 2015 καθυστέρησαν εξαιτίας καιρικών φαινομένων οπότε η πληροφορία για αυτά δεν θα εξηγούσε τους λόγους των καθυστερήσεων των πτήσεων.

Παρακάτω, παρουσιάζεται ένα συγκεντρωτικό dashboard με περιγραφικά στατιστικά για την καλύτερη κατανόηση των δεδομένων και την εξαγωγή μερικών συμπερασμάτων. Τα δεδομένα εκ πρώτης όψης δείχνουν πως εκ των 4 εκατ. πτήσεων που καταγράφησαν, το 15% αυτών καθυστέρησαν, το 50% αυτών έφτασαν στην ώρα τους ενώ το 35% έφτασε νωρίτερα από τον προγραμματισμένο χρόνο άφιξης. Επιπλέον, με βάση την επιβατική κίνηση και τον αριθμό των πτήσεων, σημαντικότερο αεροδρόμιο αναδεικνύεται το Hartsfield–Jackson Atlanta International Airport. Η αεροπορική εταιρία που εκτέλεσε τις περισσότερες πτήσεις είναι η Southwest Airlines ενώ τα περισσότερα αεροπλάνα είναι κατασκευής της Αμερικανικής Boeing.



Κατασκευή Star σχήματος

Το star σχήμα είναι ο δημοφιλέστερος τρόπος οργάνωσης και μοντελοποίησης των δεδομένων σε μια αποθήκη δεδομένων (data warehouse). Αποτελείται από έναν ή περισσότερους πίνακες γεγονότων (fact tables) και έναν ή περισσότερους πίνακες διαστάσεων (dimensions). Ένας πίνακας γεγονότων περιέχει στήλες κλειδιών που σχετίζονται με πίνακες διαστάσεων και στήλες αριθμητικών μετρικών. Οι διαστάσεις χρησιμοποιούνται για την κατηγοριοποίηση και περιγραφή των γεγονότα με τρόπους ουσιαστικούς για την απάντηση ερωτημάτων.
Στα πλαίσια της συγκεκριμένης εργασίας και με βάση το dataset που επιλέχθηκε, το star schema αποτελείται από τον παρακάτω πίνακα fact και τις αντίστοιχες διαστάσεις:

Fact table: Πτήση
Περιέχει δεδομένα και μετρικές που προσδιορίζουν την πτήση και την καθυστέρηση.
Οι μετρικές στον πίνακα fact ενδεικτικά είναι οι παρακάτω:

- **DEPARTURE_DELAY**: καθυστέρηση αναχώρησης
- **ARRIVAL_DELAY**: καθυστέρηση άφιξης
- **TAXI_OUT**: Ο χρόνος που μεσολάβησε μεταξύ της αναχώρησης από την πύλη του αεροδρομίου προέλευσης και της απογείωσης
- **TAXI_IN**: Ο χρόνος που μεσολάβησε μεταξύ της προσγείωσης και της άφιξης στην πύλη του αεροδρομίου προορισμού
- **AIRTIME**: πραγματική διάρκεια πτήσης
- **Delayed**: προσδιορίζει το αν η πτήση καθυστέρησε η όχι (boolean)

- **On-time**: προσδιορίζει το αν η πτήση έφτασε στην ώρα της ή όχι (boolean)
- **Early**: προσδιορίζει το αν η πτήση έφτασε νωρίτερα η όχι (boolean)

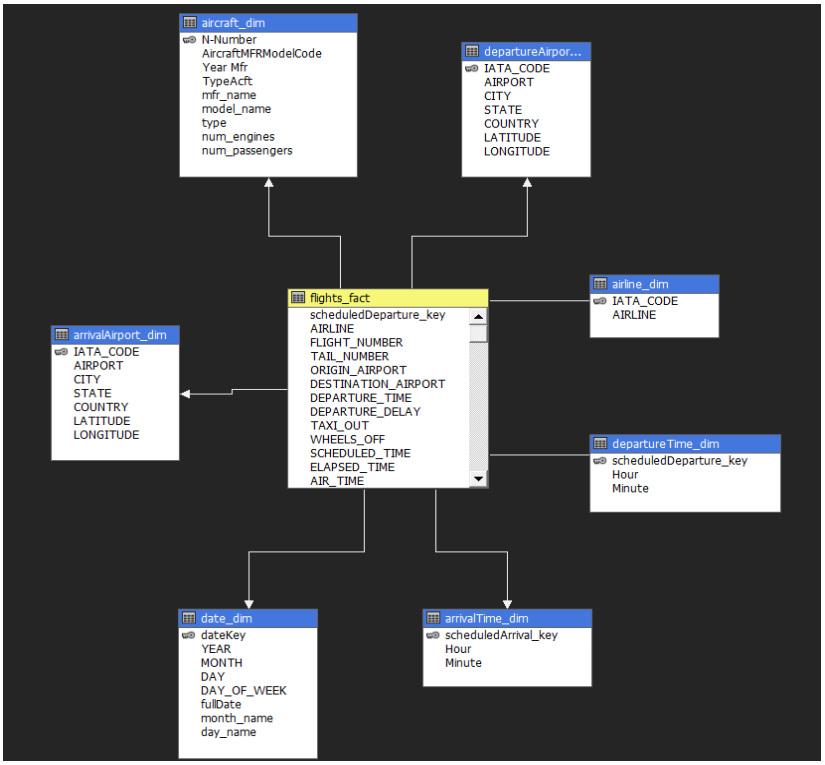
Διαστάσεις:
Παρακάτω καταγράφονται οι πίνακες των διαστάσεων και τα πεδία τους

- **Αεροδρόμιο Αναχώρησης : departureAirport_dim**
  - IATA_CODE
  - AIRPORT
  - CITY
  - STATE
  - COUNTRY
  - LATITUDE
  - LONGITUDE
- **Αεροδρόμιο Προσγείωσης : arrivalAirport_dim**
  - IATA_CODE
  - AIRPORT
  - CITY
  - STATE
  - COUNTRY
  - LATITUDE
  - LONGITUDE
- **Αεροπορική εταιρεία : airline_dim**
  - IATA_CODE
  - AIRLINE
- **Αεροσκάφος : aircraft_dim**
  - N-Number
  - AircraftMFRModelCode
  - YearMfr
  - TypeArcft
  - Mfr_name
  - Model_name
  - Num_engines
  - Num_passengers
- **Ημερομηνία : date_dim**
  - dateKey
  - YEAR
  - MONTH
  - DAY
  - DAY_OF_WEEK
  - Full_date
  - Month_name
  - day_name
- **Προγραμματισμένη Ώρα αναχώρησης : departureTime_dim**

    - scheduledDeparture_key
    - Hour
    - Minute
- **Προγραμματισμένη Ώρα προσγείωσης : arrivalTime_dim**
    - scheduledArrival_key
    - Hour
    - Minute

Η δημιουργία των διαφορετικών πινάκων fact και dimension έγινε στα πλαίσια της διαδικασίας ETL η οποία παρουσιάζεται αναλυτικά σε επόμενο κεφάλαιο.

Σημείωση: Με σκοπό την δημιουργία ενός πιο απλουστευμένου μοντέλου, ήταν εφικτή η δημιουργία μιας ενιαίας διάστασης για το αεροδρόμιο προσγείωσης και απογείωσης. Όμως, η επιλογή της δημιουργίας 2 διαφορετικών διαστάσεων για το αεροδρόμιο (αναχώρησης - άφιξης) έγινε με σκοπό την μελλοντικά ευκολότερη κατηγοριοποίηση και ανάλυση των δεδομένων των πτήσεων τόσο ως προς το αεροδρόμιο αναχώρησης και άφιξης αλλά και ως προς την ώρα δεδομένου του ότι και οι 2 αυτοί παράγοντες επηρεάζουν την καθυστέρηση. Η μοντελοποίηση αυτή μας επιτρέπει να μελετήσουμε με μεγαλύτερη ευκολία τα ερωτήματα όπως: ποιο αεροδρόμιο προσγείωσης (απογείωσης) οδηγεί σε μικρότερη καθυστέρηση, τι μερίδιο ευθύνης έχει το αεροδρόμιο αναχώρησης (άφιξης) στην καθυστέρηση.
Παρόμοια λογική ακολουθήθηκε και για την διάσταση της ώρας (απογείωσης - προσγείωσης).



Διαδικασία ETL

Στον χώρο της ανάλυσης δεδομένων, ETL ορίζεται ως μια διαδικασία τριών φάσεων (εξαγωγή - μετασχηματισμός - φόρτωση) με απώτερο σκοπό την δημιουργία μιας αποθήκης δεδομένων. Στην συγκεκριμένη εργασία, η διαδικασία ETL πραγματοποιήθηκε με την χρήση των εργαλείων Python - Pandas (extract - transform) και Visual Studio (load).

Εξαγωγή - Extract (Python - Pandas)

Για την εξαγωγή των δεδομένων, δεν χρησιμοποιήθηκε κάποια συγκεκριμένη διαδικασία καθώς τα δεδομένα αποκτήθηκαν από τους ιστότοπους στους οποίους διανείμονται.

Η μόνη διαδικασία εξαγωγής πραγματοποιήθηκε στο dataset με τις πληροφορίες για το αεροσκάφος. Πιο συγκεκριμένα, το πλήρες dataset αποτελούνταν απο 7 διαφορετικά αρχεία

- AIRCRAFT REGISTRATION MASTER FILE
    - Περιέχει τα αρχεία όλων των Πολιτικών Αεροσκαφών των ΗΠΑ που τηρούνται από την FAA και Μητρώο Πολιτικής Αεροπορία
- AIRCRAFT REFERENCE FILE
    - Περιέχει όλα τα σημαντικά χαρακτηριστικά του αεροσκάφους όπως κατασκευαστής, μοντέλο και έτος κατασκευής
- ENGINE REFERENCE FILE
    - Περιέχει στοιχεία σχετικά με τους κινητήρες και την συντήρηση τους
- AIRCRAFT DOCUMENT INDEX FILE
- RESERVE N-NUMBER FILE
- AIRCRAFT DEALER APPLICANT FILE
- AIRCRAFT DEREGISTERED FILE

Για την εξαγωγή των πεδίων που αναφέρονται στο κεφάλαιο “Περιγραφή Dataset”, χρησιμοποιήθηκαν τα αρχεία AIRCRAFT REGISTRATION MASTER FILE, AIRCRAFT REFERENCE FILE. Η συγχώνευση των αρχείων έγινε με την χρήση του πεδίου mfr_code, ενός κωδικού που έχει αποδοθεί με βάση το αεροσκάφος, τον κατασκευαστή και το μοντέλο.

Python κώδικας για την εξαγωγή των στοιχείων των αεροσκαφών από την βάση του Federal Aviation Administration.

```
# read the aircraft master file
aircraft_master = pd.read_csv('data/Aircraft/MASTER.txt',
usecols=['N-Number', 'AircraftMFRModelCode',
         'TypeAcft', 'Year Mfr'])
# strip all the values as they had lots of trailing spaces
for i in aircraft_master.columns:
```

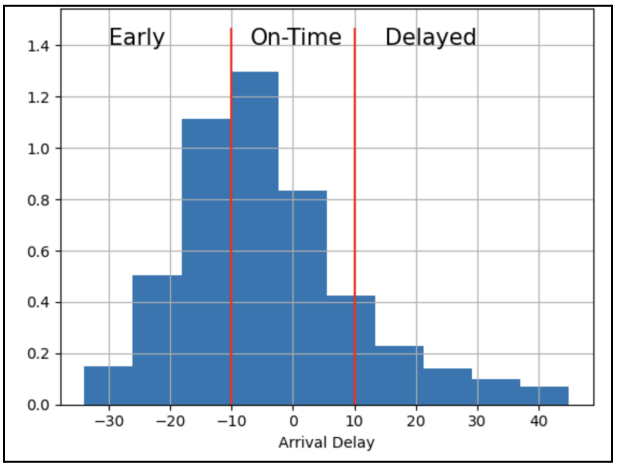
```
    aircraft_master[i] = aircraft_master[i].astype('str').str.strip()

# read the aircraft manufacturer file
mfr_ref = pd.read_csv('data/Aircraft/AcftRef.txt', header=None,
usecols=[0,1,2,6,7,10], skipinitialspace=True)

# again strip the values
for i in mfr_ref.columns:
    mfr_ref[i] = mfr_ref[i].astype('str').str.strip()

# make some renamings
col_names = {
    0 : 'mfr_name',
    1 : 'model_name',
    2 : 'type',
    6 : 'num_engines',
    7 : 'num_passengers',
    10 : 'mfr_code'
}
mfr_ref.rename(columns=col_names, inplace=True)

# merge the two dataframes using the mfr_code attribute
aircraft_info = pd.merge(aircraft_master, mfr_ref,
left_on='AircraftMFRModelCode', right_on='mfr_code')
aircraft_info.drop(columns='mfr_code', inplace=True)

# add the "N" as prefix to the N-Number to match the main data's
TAIL_NUMBER
aircraft_info['N-Number'] = "N" + aircraft_info['N-Number']
```

Μετασχηματισμός - Transform (Python - Pandas)

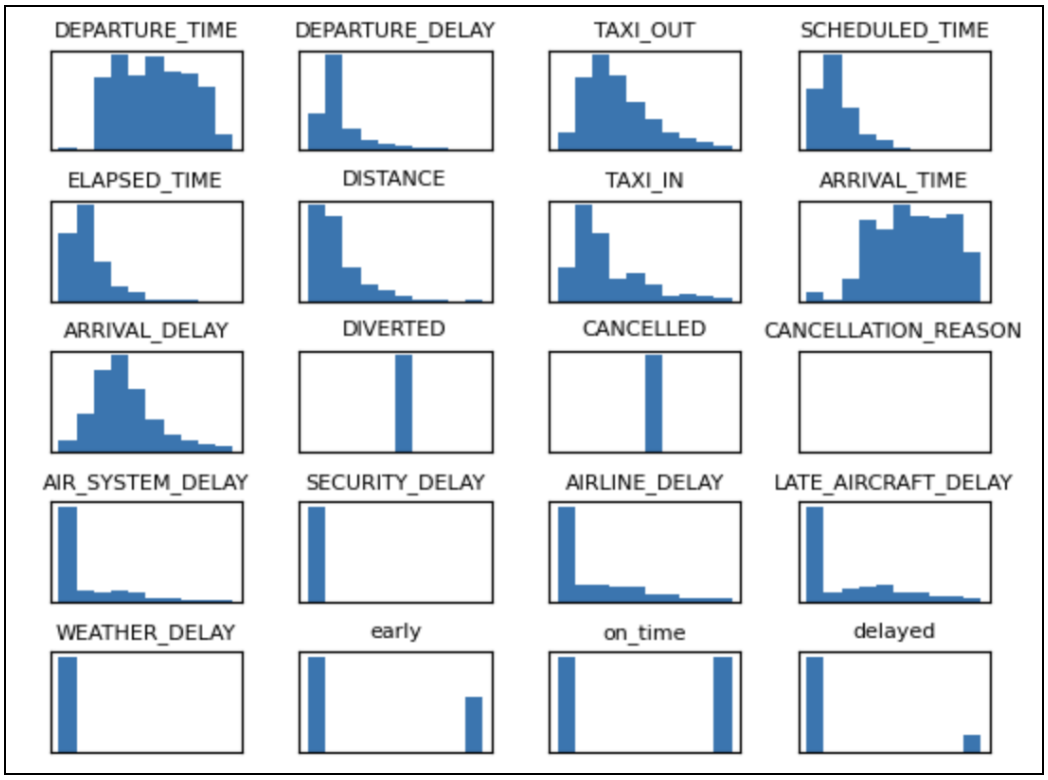
Στο στάδιο του μετασχηματισμού, πραγματοποιήθηκαν αρκετές αλλαγές τόσο στην μορφή των δεδομένων όσο και την ποσότητα τους αφού χρησιμοποιήθηκαν τεχνικές καθαρισμού δεδομένων για την αφαίρεση outlier τιμών.
Το πρώτο στάδιο της διαδικασίας ήταν ο καθαρισμός των δεδομένων. Στο στάδιο αυτό, πτήσεις οι οποίες είχαν υπερβολικά μεγάλη καθυστέρηση είτε δεν ήταν σωστά ορισμένες (missing values) αφαιρέθηκαν απο τα δεδομένα. Πιο συγκεκριμένα, αφαιρέθηκαν από τα δεδομένα πτήσεις των οποίων οι τιμές DEPARTURE_DELAY, ARRIVAL_DELAY, TAXI_IN, TAXI_OUT ήταν πάνω απο το ποσοστημόριο 0.95 (95%) και κάτω απο το ποσοστημόριο 0.01 (1%). Αυτό οδήγησε στην μείωση του συνολικού αριθμού των πτήσεων κατά 950.000.

Επιπλέον, κατά την διάρκεια της διαδικασίας του μετασχηματισμού των δεδομένων δημιουργήθηκαν και οι τιμές delayed, on-time και early. Σύμφωνα με τον United States Federal Aviation Administration (FAA) μία πτήση θεωρείται πως έχει καθυστερήσει αν προσγειωθεί τουλάχιστον 10 λεπτά μετά τον προγραμματισμένο χρόνο άφιξης. Αντίστοιχα, μία πτήση θεωρείται πως αφίχθη νωρίτερα απο την προγραμματισμένη ώρα άφιξης αν προσγειωθεί νωρίτερα τουλάχιστον κατά 10 λεπτά. Δημιουργήθηκαν έτσι οι παραπάνω μεταβλητές με τις ακόλουθες συνθήκες:

- ARRIVAL_DELAY < -10 : **Early**
- -10 ≤ ARRIVAL_DELAY ≤ 10 : **On-time**
- ARRIVAL_DELAY > 10 : **Delayed**



Έπειτα από το πέρας της διαδικασίας του καθαρισμού των δεδομένων, τα κύρια πεδία του dataset διαμορφώθηκαν ως εξής (histogram plot). Συνολικά το dataset μας περιέχει δεδομένα για 4 εκατ. πτήσεις, 14 αεροπορικές εταιρίες, 322 αεροδρόμια και 4099 αεροσκάφη.



Στην συνέχεια, αφού τα δεδομένα καθαρίστηκαν και δημιουργήθηκαν και οι απαραίτητες μεταβλητές, το επόμενο βήμα ήταν η κατασκευή του fact table και των πινάκων των διαστάσεων. Οι διαστάσεις aircraft και airport (departure και arrival) ήταν ήδη δημιουργημένες σε ξεχωριστό πίνακα από το dataset του Kaggle. Έτσι, οι διαστάσεις που χρειάστηκε να δημιουργηθούν ήταν το date_dim και το departure και arrival time.

Σχετικά με την διάσταση time, η αναμενόμενη ώρα αναχώρησης και άφιξης περιέχεται στην μεταβλητή SCHEDULED_DEPARTURE και SCHEDULED_ARRIVAL. Και για τις 2 αυτές τιμές ισχύει πως τα 2 πρώτα ψηφία αντιπροσωπεύουν την ώρα και τα 2 τελευταία τα λεπτά. Έτσι, με την χρήση του pandas η τιμή αυτή χωρίστηκε σε 2 διαφορετικές μεταβλητές δημιουργώντας έτσι τις στήλες Hour και Minute. Η διαδικασία αυτή εφαρμόστηκε και για τις 2 τιμές SCHEDULED_DEPARTURE και SCHEDULED_ARRIVAL. Οι παραγόμενες αυτές τιμές αποθηκεύτηκαν σε έναν νέο πίνακα μαζί με το κλειδί σύνδεσης με τον κεντρικό πίνακα. Το κλειδί που επιλέχθηκε είναι το SCHEDULED_DEPARTURE και SCHEDULED_ARRIVAL αντίστοιχα ως αλφαριθμητικός χαρακτήρας. Τέλος, απο τους 2 νέους πίνακες (departureTime_dim και arrivalTime_dim) αφαιρέθηκαν οι διπλοεγγραφές.

Σχετικά με την διάσταση date, αν παρατηρήσουμε τα δεδομένα μπορούμε να διακρίνουμε πως είναι μερικώς κατασκευασμένη. Οι τιμές YEAR, MONTH, DAY και DAY_OF_WEEK είναι ήδη διαθέσιμες και μένει να προσθέσουμε εμείς περεταίρω πληροφορία. Οι τιμές αυτές τοποθετούνται σε έναν νέο πίνακα και για κλειδί χρησιμοποιούμε τον αλφαριθμητικό χαρακτήρα dateKey του οποίου τα πρώτα 4 ψηφία αντιπροσωπεύουν το έτος, τα επόμενα 2 τον μήνα και τα τελευταία 2 την ημέρα. Όπως και με την διάσταση του χρόνου, αφαιρούνται οι διπλοεγγραφές.

Επιπλέον, προστέθηκαν και οι τιμές day_name, month_name και fulldate με σκοπό την αποτελεσματικότερη χρήση των ημερομηνιών σε γραφήματα.

Τέλος, η διάσταση aircraft έχει κατασκευαστεί στο στάδιο transform αφού τα απαιτούμενα πεδία βρίσκονται σε έναν ενιαίο πίνακα. Η σύνδεση της διάστασης με τον κεντρικό πίνακα γίνεται μέσω της τιμής TAIL_NUMBER στο fact και του N-Number στο aircraft_dim. Στην πραγματικότητα, οι τίτλοι TAIL_NUMBER και N-Number αντιπροσωπεύουν την ίδια τιμή αφού όλοι οι κωδικοί αεροσκαφών (tail number) των ΗΠΑ ξεκινούν με τον χαρακτήρα “N” και για αυτό συχνά αποκαλούνται και “N-Numbers”. Για την επιτυχημένη σύνδεση των 2 πινάκων χρειάστηκε να προστεθεί το πρόθημα “N” σε κάθε TAIL_NUMBER.

Σημείωση: Όσα πεδία τοποθετήθηκαν στους πίνακες των διαστάσεων αφαιρέθηκαν από τον κεντρικό πίνακα fact

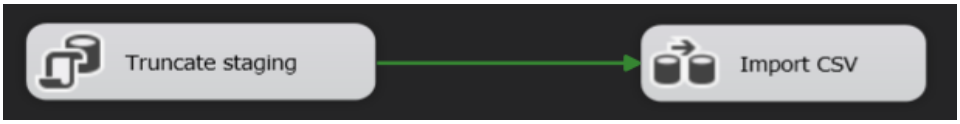
Python κώδικας για την κατασκευή της διάστασης Date:

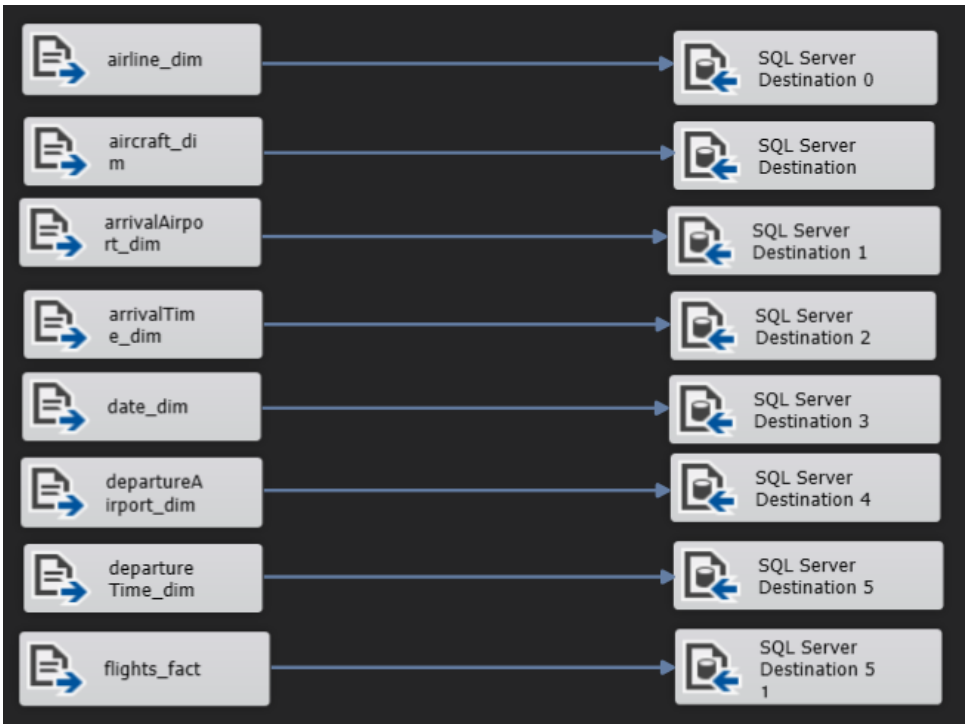
```
# get the full date
flights['fullDate'] = pd.to_datetime(flights[['YEAR', 'MONTH', 'DAY']])

# bring the month and day in the desired 2 characters format
flights['MONTH'] = flights['MONTH'].str.zfill(2)
flights['DAY'] = flights['DAY'].str.zfill(2)

# create the date key
flights['dateKey'] = flights['YEAR'].astype('str') +
flights['MONTH'].astype('str') + flights['DAY'].astype('str')

# add those attributes in a new df and remove them from the main table
date_dim = flights.loc[:,['YEAR','MONTH','DAY','DAY_OF_WEEK', 'fullDate',
'dateKey']]
flights.drop(columns=['YEAR','MONTH','DAY','DAY_OF_WEEK', 'fullDate'],
inplace=True)

# drop the duplicates
date_dim.drop_duplicates(inplace=True)

# generate the day_name and month_name attributes
date_dim['month_name'] = date_dim['fullDate'].dt.month_name()
date_dim['day_name'] = date_dim['fullDate'].dt.day_name()
```

Φόρτωση - Load (Visual Studio)

Μετά το στάδιο του μετασχηματισμού του dataset, την δημιουργία δηλαδή των κατάλληλων αρχείων δεδομένων που αντανακλούν το star σχήμα, μπορούμε να το εισάγουμε στην αποθήκη δεδομένων. Για την διαδικασία αυτή, χρησιμοποιήθηκε το Visual Studio και ως αποθήκη δεδομένων ο Sql Server. Πιο συγκεκριμένα, δημιουργήσαμε ένα Data Integration Project με μία κύρια λειτουργία. Η λειτουργία αυτή είναι η ανάγνωση κάθε αρχείου διάστασης (.csv) καθώς και του fact table με σκοπό την εισαγωγή τους στην βάση δεδομένων. Μέσω του συγκεκριμένου project, μπορούμε να εξασφαλίσουμε την ομαλή εισαγωγή των δεδομένων στους αντίστοιχους πίνακες με τους επιλεγμένους τύπους για κάθε πεδίο, καθώς και έναν χρηστικό τρόπο συνεχούς ανανέωσης των δεδομένων. Πριν από κάθε εισαγωγή σε οποιοδήποτε πίνακα στην αποθήκη δεδομένων, εκτελείται ένα βήμα truncate κατά το οποίο όλες οι τιμές του πίνακα διαγράφονται με σκοπό να εισαχθούν οι νέες. Πριν από την εκτέλεση της παραπάνω διαδικασίας, οι πίνακες των διαστάσεων αλλά και ο fact έχουν δημιουργηθεί και οι σχέσεις ξένου κλειδιού μεταξύ τους έχουν οριστεί με τον ακόλουθο τρόπο.

Dimension	**Fact Table key**	**Dimension Key**
departureAirport_dim	ORIGIN_AIRPORT	IATA_CODE
arrivalAirport_dim	DESTINATION_AIRPORT	IATA_CODE
airline_dim	AIRLINE	AIRLINE_CODE
aircraft_dim	TAIL_NUMBER	N-Number
date_dim	dateKey	dateKey
departureTime_dim	scheduledDeparture_key	scheduledDeparture_key
arrivalTime_dim	scheduledArrival_key	scheduledArrival_key





Κατασκευή Κύβου Δεδομένων

Ο κύβος δεδομένων αποτελεί μια πολυδιάστατη δομή δεδομένων που χρησιμοποιείται για να αποθηκεύει και να αναλύει μεγάλους όγκους δεδομένων. Χρησιμοποιείται συνήθως σε εφαρμογές επιχειρηματικής ευφυΐας και δημιουργείται συγκεντρώνοντας δεδομένα από πολλαπλές πηγές και οργανώνοντας τα σε διαστάσεις. Η χρησιμότητα της συγκεκριμένης δομής ανάγεται στη γρήγορη και αποτελεσματική αναζήτηση και ανάλυση των δεδομένων. Οργανώνοντας τα δεδομένα με δομημένο τρόπο, οι κύβοι επιτρέπουν στους χρήστες να παρατηρούν τα δεδομένα σε διαφορετικές οπτικές γωνίες (π.χ τοποθεσία, χρόνος) και βοηθά στον εντοπισμό τάσεων και μοτίβων που θα ήταν δύσκολο να παρατηρηθούν σε έναν επίπεδο πίνακα δεδομένων.
Στα πλαίσια της συγκεκριμένης εργασίας, για την δημιουργία του κύβου συγκεντρώθηκαν δεδομένα από διάφορες πηγές (Kaggle, U.S. Department of Transportation) και συνδυάστηκαν προκειμένου να συμπεριληφθούν στον κύβο. Τα δεδομένα αυτά, ‘καθαρίστηκαν’ και μετασχηματιστηκαν σε μορφή κατάλληλη για την χρήση τους στον κύβο. Συγκεκριμένα, τα δεδομένα χωριστηκαν σε 7 διαφορετικές διαστάσεις γύρω από ένα κεντρικό πίνακα γεγονότων ακολουθώντας την μεθοδολογία του ‘σχήματος αστέρα. Στη συνέχεια, για την κατασκευή του

κύβου χρησιμοποιήθηκαν τα SQL Server Data Tools (SSDT) τα οποία συμπεριλαμβάνονται στο Microsoft Visual Studio το οποίο αποτελεί μία πλατφόρμα ανάπτυξης λογισμικού. Ύστερα, από τον σχεδιασμό του ‘σχήματος αστέρα’ ακολουθεί η δημιουργία ενός νέου SASS project για την δημιουργία του κύβου δεδομένων. Αυτό υλοποιείται μέσω του Microsoft Visual Studio και συγκεκριμένα μέσω του "Analysis Services Multidimensional and Data Mining Project". Κατά τη δημιουργία του νέου project ορίστηκε ως πηγή δεδομένων ο Microsoft SQL Server ο οποίος τρέχει τοπικά και επιλέχθηκαν οι διαστάσεις και ο κεντρικός πίνακας που αποτελούν τον κύβο τα οποία έχουν οριστεί σε προηγούμενο κεφάλαιο “Κατασκευή Star σχήματος”. Στη συνέχεια, οριστηκαν οι διαστάσεις και τα μέτρα, κατασκευάζοντας ετσι τον κύβο.Ύστερα από την κατασκευή του, ο κύβος αναπτύχθηκε στη βάση δεδομένων του Microsoft SQL Server όπου χρησιμοποιήθηκε για ανάλυση των δεδομένων.

Συγκεκριμενα, μέσω του κύβου δημιουργήθηκαν οι παρακάτω μετρικές και ιεραρχίες:

- Μετρικές

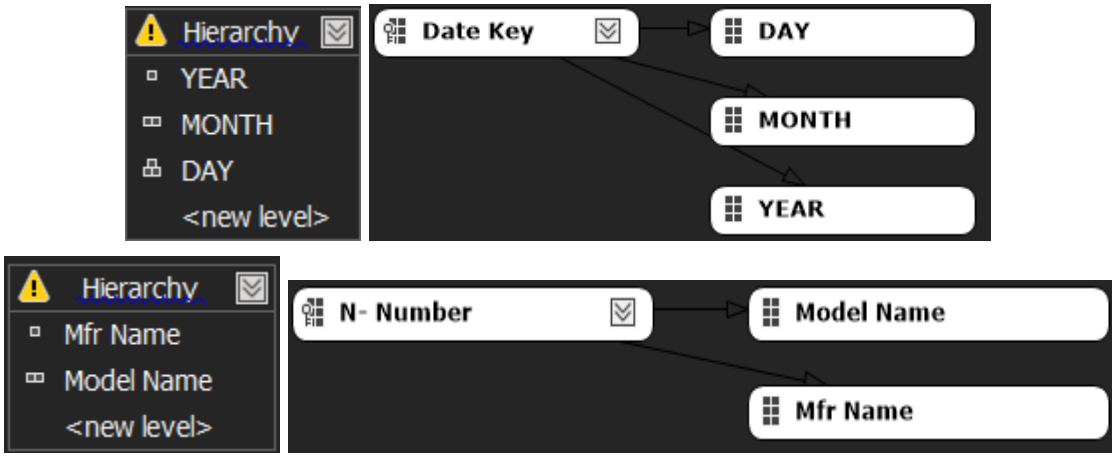
    - **AVG_Delayed**: Μέσος όρος των καθυστερημένων πτήσεων
    - **Average_delay_time**: Μέσος όρος του χρόνου καθυστέρησης των πτήσεων
    - **AVG_On-time:** Μέσος όρος των πτήσεων που έφτασαν στην ώρα τους
    - **AVG_Early**: Μέσος όρος των πτήσεων που έφτασαν νωρίτερα από την προγραμματισμένη ώρα

- Ιεραρχίες

    - Date
        1. Year
        2. Month
        3. Day

    - Aircraft
        1. Manufacturer name
        2. Model name

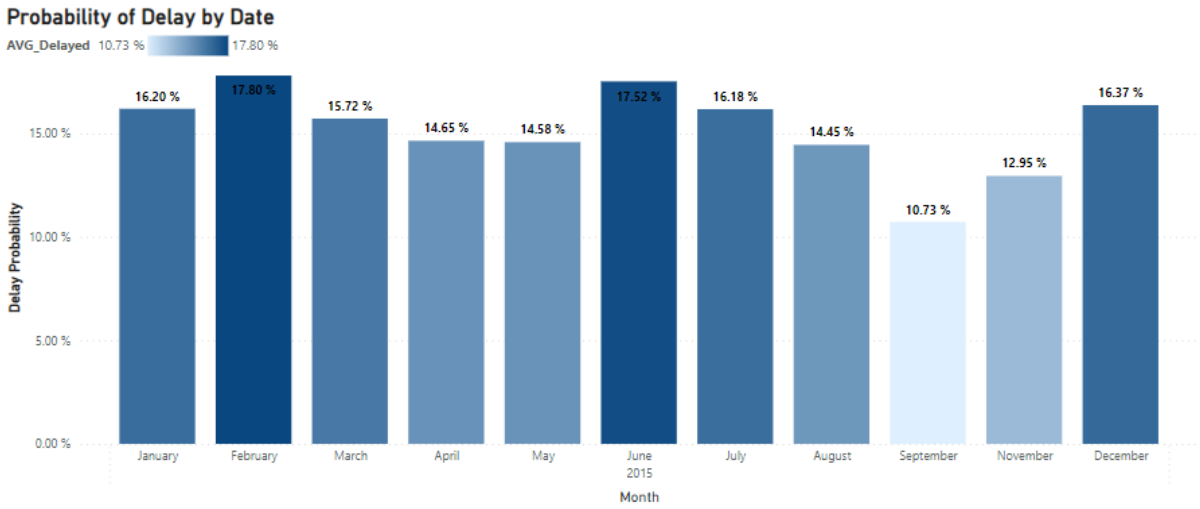
Οι παραπάνω μετρήσεις και ιεραρχίες χρησιμοποιήθηκαν στη συνέχεια για την οπτικη αναπαράσταση των δεδομένων και την επισήμανση σημαντικών τάσεων και μοτίβων που εντοπίστηκαν σε αυτά. Συγκεκριμένα οι μετρικές και οι ιεραρχίες χρησιμοποιήθηκαν για την αναπαράσταση διαδραστικών γραφημάτων με λειτουργίες drill-up και drill-down.



Visual reports-Dashboards

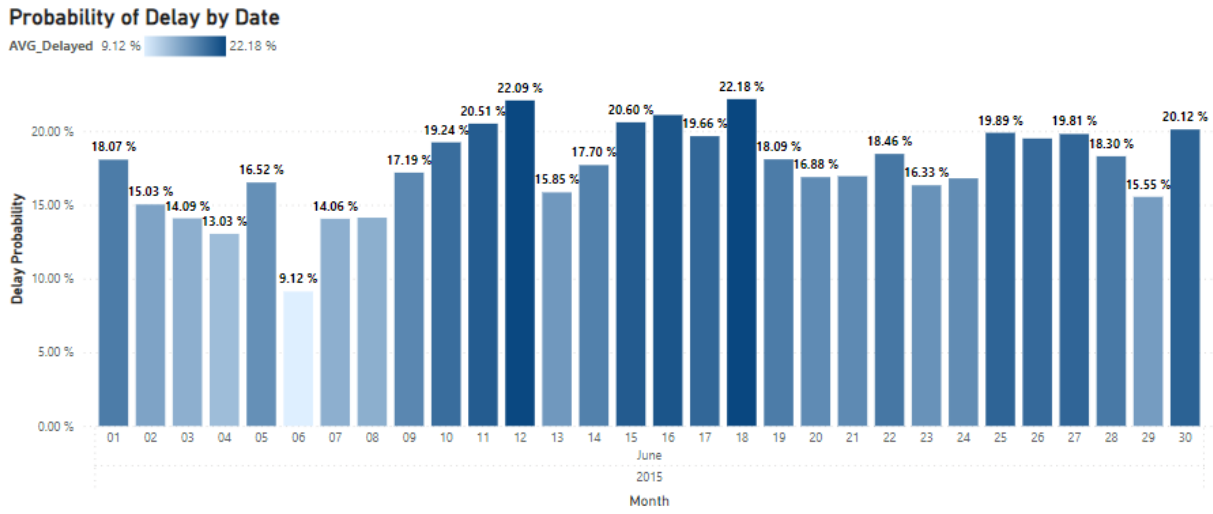
΄Υστερα από τη κατασκευή του κύβου δεδομένων χρησιμοποιήθηκε το Microsoft Power BI για την δημιουργία διαδραστικών γραφημάτων. Το Microsoft Power BI αποτελεί μια πλατφόρμα επιχειρηματικής ευφυΐας και οπτικοποίησης δεδομένων που είναι συμβατή με "SQL Server Analysis Services" και επιτρέπει την σύνδεση με τον κύβο δεδομένων.
Στο συγκεκριμένο dataset συνδυάστηκαν διάφορες μετρικές που υπολογιστήκαν μέσω αριθμητικών πράξεων στο κύβο με άλλα πραγματικά δεδομένα εξάγοντας σημαντικά μοτίβα και ευρήματα όσον αφορά τις πτήσεις που διεξήχθησαν το 2015 στην Ηνωμενες Πολιτειες Αμερικής. Παρακάτω απεικονίζονται τα σημαντικότερα γραφήματα:

1. Πιθανότητα καθυστέρησης ανά Χρονιά - Μήνα - Ημέρα

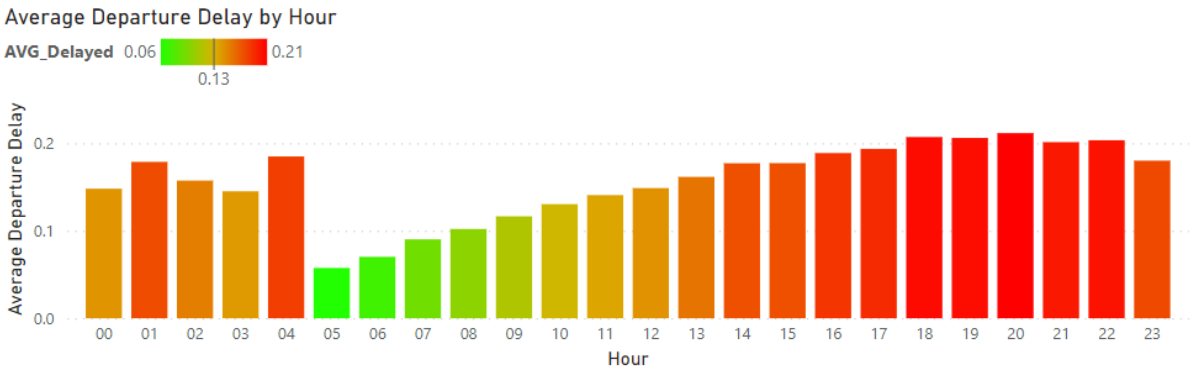


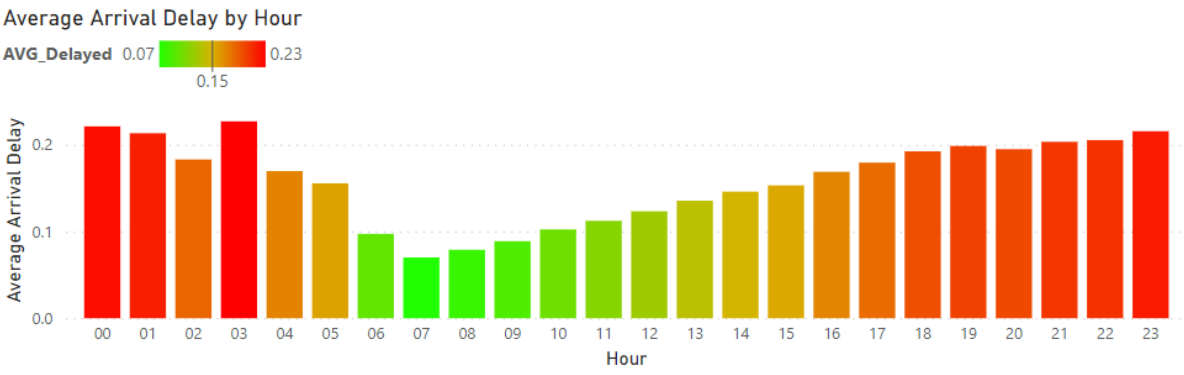
Στο παραπάνω διάγραμμα απεικονίζεται η πιθανότητα καθυστέρησης ανά μήνα. Παρατηρούμε πως ο καλύτερος μήνας για την αποφυγή καθυστερήσεων είναι ο Σεπτέμβρης (10.7%) ενώ ο χειρότερος ο Φεβρουάριος (17.8%). Οι διαπιστώσεις μας επιβεβαιώνονται και απο τον FDA καθώς και αυτός αναδεικνύει τον Σεπτέμβρη ώς τον καλύτερο μήνα με βάση τις καθυστερήσεις και τον Φεβρουάριο ως τον χειρότερο. Οι καθυστερήσεις στους μήνες Ιανουάριος - Φεβρουάριος αποδίδονται κυρίως στις καιρικές συνθήκες ενώ για τον μήνα Δεκέμβριο, αποδίδονται στις καιρικές συνθήκες και την αυξημένη επιβατική κίνηση εξαιτίας των εορτών των χριστουγέννων. Επιπλέον, καθυστερήσεις στους μήνες Ιούνιος - Ιούλιος - Αύγουστος αποδίδονται κυρίως σε έντονη αεροπορική κινητικότητα λόγω των καλοκαιρινών διακοπών. Ο μικρός αριθμός καθυστερήσεων τον Σεπτέμβριο - Οκτώβρη οφείλεται στον γενικότερα καλό καιρό και την χαμηλή επιβατική κίνηση (“Shoulder” season).

Στα πλαίσια του συγκεκιμένου διαγράμματος αξιοποιείται και η ιεραρχία που έχει δημιουργηθεί στην διάσταση date. Αυτό μας δίνει την δυνατότητα drill-down και drill-up. Έτσι, κάνοντας διπλό κλικ στο διάγραμμα στο Power Bi σε οποιονδήποτε μήνα θα μας δείξει αναλυτικά την πιθανότητα καθυστέρησης ανά ημέρα του συγκεκριμένου μήνα. Επιλέγοντας για παράδειγμα τον μήνα Ιούνιο, παίρνουμε το παρακάτω διάγραμμα.



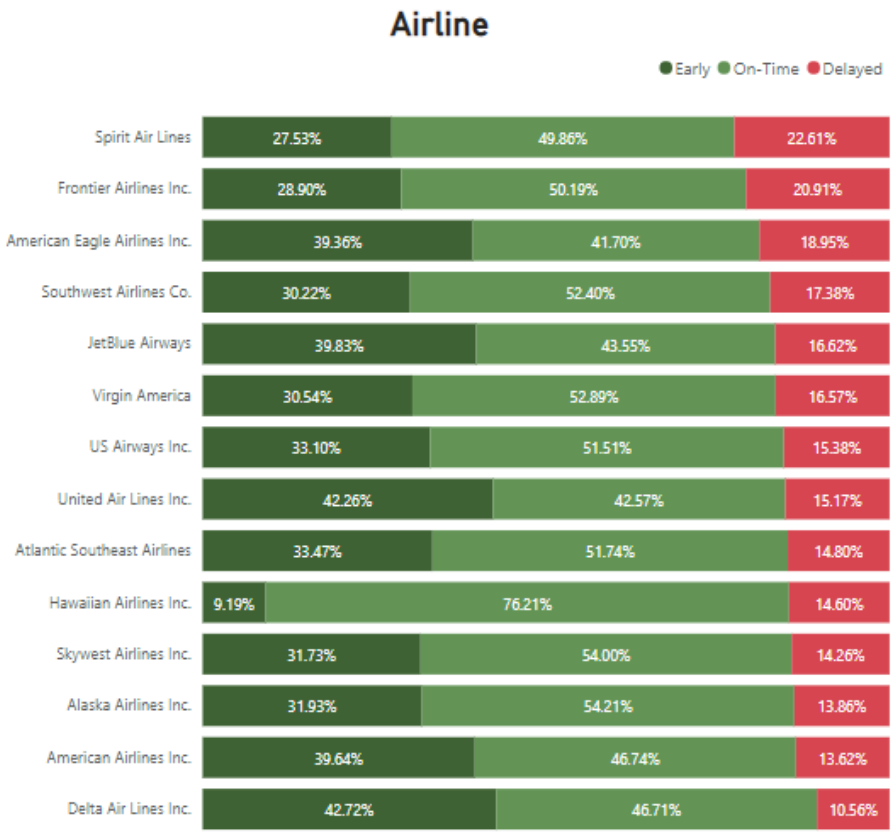
2. Δείκτης μέσης χρονικής καθυστέρησης με βάση την ώρα αναχώρησης-άφιξης





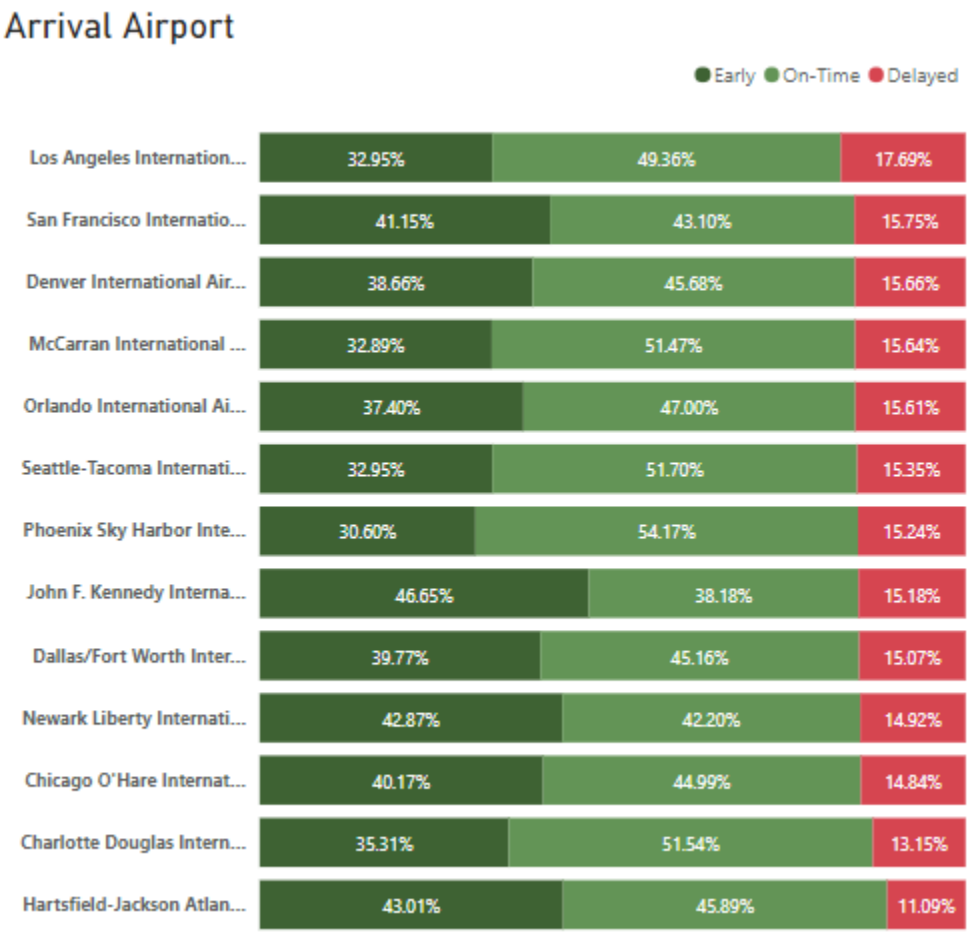
Στα παραπάνω γραφήματα απεικονίζονται η μέση χρονική καθυστέρηση αναχώρησης των πτήσεων(πάνω σχήμα) καθώς και η μέση χρονική καθυστέρηση άφιξης των πτήσεων(κάτω σχήμα) σε συνδυασμό με την ώρα της ημέρας. Τα παραπάνω γραφήματα είναι ταξινομημένα κατά αύξουσα σειρά με βάση την ώρα της ημέρας. Είναι φανερό τα δύο παραπάνω γραφήματα ακολουθούν μια παρόμοια κατανομή. Συνεπώς η ανάλυση θα αναφέρεται στην γενικη χρονική καθυστέρηση μιας πτήσης και οχι συγκεκριμενα στην άφιξη ή στην αναχώρηση. Παρατηρώντας τα διαγράμματα γίνεται αντιληπτό ότι τις πολύ πρωινές ώρες υπάρχει μειωμένη καθυστέρηση των πτήσεων ενώ όσο προχωράει η μέρα τόσο και αυξάνεται η καθυστέρηση των πτήσεων η οποία μεγιστοποιείται στο διάστημα 19:00-21:00. Αυτό εξηγείται λογικά καθώς τα πρωινά πολύ λίγη εναέρια κυκλοφορία έχει συσσωρευτεί γύρω από τα αεροδρόμια, κάνοντας τις αναχωρήσεις και τις αφίξεις πιο ομαλές και γρήγορες. Καθώς ο εναέριος χώρος γίνεται πιο γεμάτος, οι ελεγκτές εναέριας κυκλοφορίας καθυστερούν τις αναχωρήσεις πτήσεων και χρειάζονται περισσότερο χρόνο για να προετοιμαστούν για τις πτήσεις που φτάνουν. Επομένως, από τα παραπάνω διαγράμματα εξάγονται σημαντικά δεδομένα που μπορεί να επηρεάσουν τον χρονοπρογραμματισμό και την επιλογή της ώρας που κάποιος ενδιαφερόμενος θα επιλέξει για την πτήση του. Σημαντικό είναι επίσης το συμπέρασμα πως πτήσεις που αναχωρούν και προσγειώνονται στο διάστημα 12 - 04 το βράδυ έχουν σημαντική πιθανότητα να καθυστερήσουν παρόλο που η επιβατική κίνηση στα αεροδρόμια εκείνη την ώρα είναι αρκετά μειωμένη. Αξίζει να σημειωθεί ότι τα αποτελέσματα που προκύπτουν συμβαδίζουν με αυτά που προβλέπει το Department of Transportation σύμφωνα με τη Forbes.

3. Διάγραμμα early, on-time και delayed ανά Αεροπορική εταιρεία



Στο παραπάνω γράφημα απεικονίζονται οι 14 πιο γνωστες αεροπορικές εταιρείες(με βαση τον αριθμο των πτήσεων που πραγματοποιηθηκαν) σε συνδυασμο με το ποσοστό των πτήσεων τους που ηρθαν νωριτερα απο την προγραμματισμενη ωρα, ήρθαν στην ώρα τους και που ήρθαν καθυστερημένα. Το παραπάνω διάγραμμα είναι ταξινομημένο σε φθίνουσα σειρά ως πρός το ποσοστό των καθυστερημένων πτήσεων. Από το διάγραμμα αυτό εξάγονται σημαντικές πληροφορίες όσον αφορά την αξιοπιστία των αεροπορικών εταιρειών.Συγκεκριμένα φαίνεται ότι οι αεροπορικές εταιρείες: Spirit Airlines, Frontier Airlines Inc , American Eagle Airlines Inc παρουσιάζουν σημαντικά ποσοστά καθυστερημένων πτήσεων(≈20%) ενώ οι:
Alaska Airlines Inc, American Airlines Inc παρουσίαζουν χαμηλά ποσοστά καθυστερημένων πτήσεων(≈10%). Ως πιο αξιόπιστη αεροπορική αεροπορική εταιρεία παρατηρείται η , Delta Airlines Inc το 89% των συνολικών πτήσεων της φτάνουν νωρίτερα ή στην ώρα τους. Οι διαπιστώσεις μας επιβεβαιώνονται και από το Forbes το οποίο στις αξιολογήσεις των αεροπορικών εταιριών τα τελευταία χρόνια, αναδεικνύει την Delta Airlines.

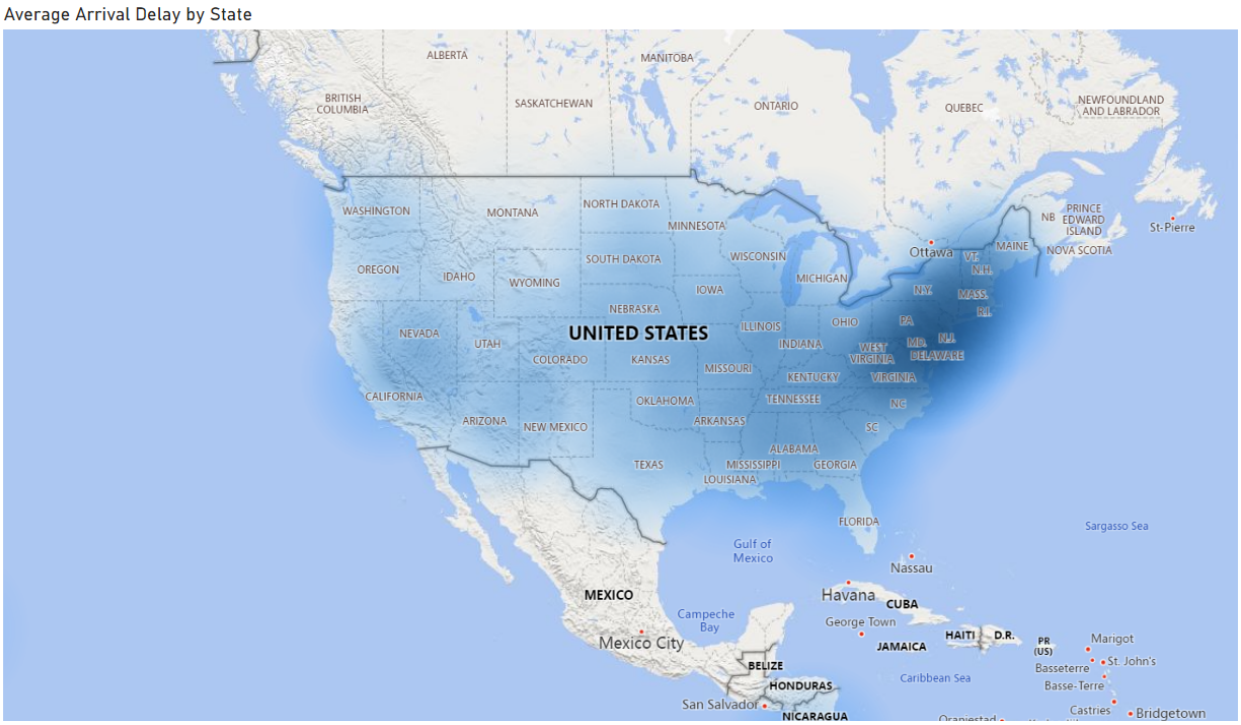
4. Διάγραμμα early, on-time και delayed ανά Αεροδρόμιο Άφιξης



Το παραπάνω διάγραμμα απεικονίζει το ποσοστό των πτήσεων που προσγειώθηκαν νωρίτερα, έφτασαν στην ώρα τους και καθυστέρησαν ανά αεροδρόμιο άφιξης. Το αεροδρόμιο με τις περισσότερες καθυστερήσεις (17% των συνολικών πτήσεων) είναι το Los Angeles International Airport (LAX). Το αεροδρόμιο με την μεγαλύτερη συνέπεια σε on-time αφίξεις αναδεικνύεται το Phoenix Sky Harbor Airport της Arizona. Τέλος, το αεροδρόμιο με τις περισσότερες αφίξεις νωρίτερα του προγραμματισμένου είναι το John F. Kennedy International Airport στην Νεα Υορκη. Θεωρώντας πως σκοπός των επιβατών είναι να φτάνουν στην ώρα τους ή ακόμα και νωρίτερα, το Hartsfield–Jackson Atlanta International Airport αναδεικνύεται νικητής με το 89% (43%[early] + 46%[on-time]) των πτήσεων να φτάνουν νωρίτερα ή στην ώρα τους.

Σημείωση: *στο παραπάνω διάγραμμα απεικονίζονται οι μετρικές της καθυστέρησης μόνο για τα 13 μεγαλύτερα αεροδρόμια με βάση την επιβατική κίνηση.*

5. Θερμικός χάρτης απεικόνισης των πολιτειών της Αμερικής σε συνδυασμό με την πιθανότητα χρονικής καθυστέρηση στις πτήσεις τους



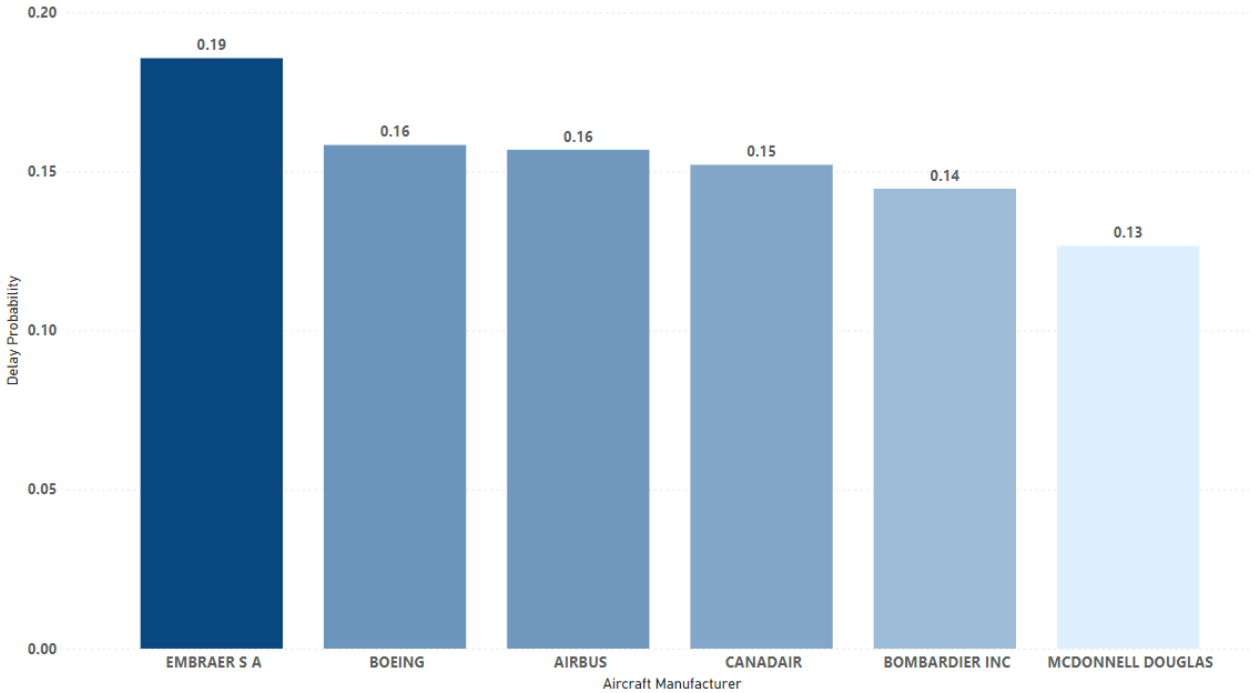
Το παραπάνω γράφημα αποτελεί ένα θερμικό χάρτη που αφορα τις πολιτείες της Αμερικής. Συγκεκριμένα, η θερμότητα(χρώμα) του χάρτη κυμαίνεται στις αποχρώσεις του μπλε με την πιο ανοιχτή απόχρωση να δείχνει τις πολιτείες που δεν έχουν καθυστέρηση ενώ με την πιο σκούρη απόχρωση να φαίνονται οι πολιτείες που κατα μέσο όρο έχουν μεγάλη καθυστέρηση στις πτήσεις τους. Με μια πρώτη ματιά γίνεται αντιληπτό ότι το βορειοανατολικό κομμάτι του χάρτη προβάλει πιο σκούρη απόχρωση σε σχέση με τον υπόλοιπο χάρτη. Συνεπώς, παρατηρείται μια συσσώρευση καθυστερημένων πτήσεων στις βορειοανατολικες πολιτείες της Αμερικής. Πιο αναλυτικά ζουμάροντας στον χάρτη παρατηρείται ότι το σκούρο χρώμα συγκεντρώνεται κατα κύριο λόγο στην πολιτεία της Νέας Υόρκης. Η συγκεκριμένη πολιτεία δεν είναι τυχαία καθώς το πρόβλημα του μεγάλου αριθμού καθυστερημένων πτήσεων στη Νέας Υόρκη υπάρχει εδώ και πολλά χρόνια και μάλιστα πρόσφατα το Aviation Rulemaking Committee (ARC) έλαβε νέα μέτρα για την αντιμετώπιση αυτού του προβλήματος.
Οι κύριοι λόγοι έντονης συσσώρευσης καθυστερημενων πτησεων στην Νέα Υόρκη είναι:

- **Συμφόρηση εναέριας κυκλοφορίας**: Η Νέα Υόρκη έχει μερικά από τα πιο πολυσύχναστα αεροδρόμια της χώρας και τα υψηλά επίπεδα εναέριας κυκλοφορίας μπορεί να οδηγήσουν σε καθυστερήσεις λόγω συμφόρησης στον εναέριο χώρο γύρω από το αεροδρόμιο.
- **Χωρητικότητα αεροδρομίου**: Μερικά από τα αεροδρόμια της Νέας Υόρκης, όπως το LaGuardia και το JFK, λειτουργούν κοντά στη πλήρη χωρητικότητα τους, γεγονός που μπορεί να συμβάλει σε καθυστερήσεις, καθώς υπάρχει λιγότερη ευελιξία για την αντιμετώπιση απροσδόκητων διακοπών.

- **Καιρός**: Οι δυσμενείς καιρικές συνθήκες, όπως καταιγίδες, έντονες χιονοπτώσεις,ισχυροί άνεμοι, μπορεί να διαταράξουν τις πτήσεις και να προκαλέσουν καθυστερήσεις.

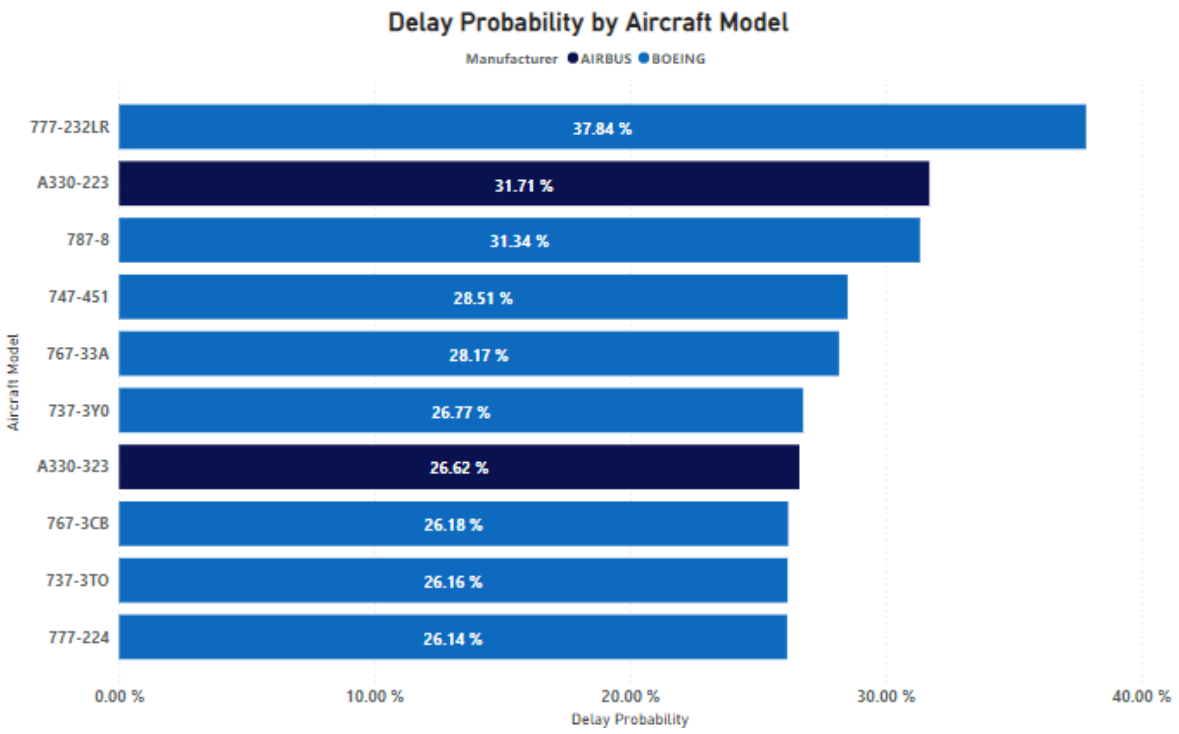
6. Διάγραμμα πιθανότητας καθυστέρησης ανά Κατασκευαστή

Σύμφωνα με το Bereau of Transportation Statistics, η καθυστέρηση μιας πτήσεις μπορεί να οφείλεται και στο αεροσκάφος. Πιο συγκεκριμένα, ένα αεροσκάφος μπορεί να προκαλέσει καθυστέρηση λόγω της αργοπορημένης άφιξης του λόγω προηγούμενης πτήσης, εξαιτίας της αδυναμίας αποδοτικού ανεφοδιασμού του ή τέλος τεχνικού προβλήματος. Στο παρακάτω διάγραμμα φαίνεται η πιθανότητα καθυστέρησης ανά κατασκευαστική εταιρεία. Η μεγαλύτερη τιμή παρατηρείται για την Embraer (19%). Η Embraer S.A. είναι ένας βραζιλιάνικος πολυεθνικός κατασκευαστής που παράγει εμπορικά και στρατιωτικά αεροσκάφη. Οι 2 μεγαλύτεροι κατασκευαστές (Boeing και Airbus) παρουσιάζουν παρόμοια πιθανότητα καθυστέρησης για τα αεροσκάφη που παράγουν (16%). Η εταιρεία της οποίας τα αεροσκάφη καθυστερούν με την μικρότερη πιθανότητα είναι η McDonnell Douglas (13%). Η σειρά πολιτκών αεροσκαφών της (MD) χρησιμοποιείται κυρίως για πτήσεις μικρών αποστάσεων και παρουσιάζει εξαιρετικά ποσοστά early και on-time αφίξεων. Παρόλαυτα, λόγω του υψηλού ανταγωνισμού από τα αεροσκάφη A320-ceo και B737-800, η Delta πραγματοποίησε την τελευταία πτήση αεροσκάφους της McDonnell τον Ιούνιο 2020.



** Στο δίαγραμμα απεικονίζονται κατασκευαστές με αεροσκάφη τουλάχιστον 20 επιβατών και τουλάχιστων 2 κινητήρων.*

7. Διάγραμμα της πιθανότητας καθυστέρησης ανά μοντέλο αεροσκάφους



Στο παραπάνω διάγραμμα απεικονίζεται η πιθανότητα μια πτήση να καθυστερήσει (delay > 10min) για κάθε μοντέλο αεροσκάφους. Για την καλύτερη απεικόνιση των αποτελεσμάτων, στο διάγραμμα περιέχονται τα 10 μοντέλα αεροσκάφους με την μεγαλύτερη πιθανότητα καθυστέρησης. Η κάθε μπάρα είναι χρωματισμένη με βάση τον κατασκευαστή του αεροσκάφους. Προκύπτει λοιπόν πως το μοντέλο της Boeing 777-232LR (290 επιβάτες) έχει την μεγαλύτερη πιθανότητα καθυστέρησης (38%). Απο την μεριά του μεγάλου ανταγωνιστή της, το αεροσκάφος της Airbus με την μεγαλύτερη πιθανότητα καθυστέρησης είναι το A330-223 (277 επιβάτες). Ενδιαφέρον έχει επίσης και το γεγονός πως και τα 2 παραπάνω μοντέλα αεροσκαφών χρησιμοποιούνται κυρίως από την αεροπορική εταιρεία Delta Airlines. Γενικότερα, ένα αεροσκάφος μπορεί να προκαλέσει καθυστέρηση αν προκύψει κάποιο τεχνικό πρόβλημα ή ο ανεφοδιασμός του σε καύσιμα ή πρώτες ύλες είναι χρονοβόρα διαδικασία.

Μοντέλα Εξόρυξης Δεδομένων

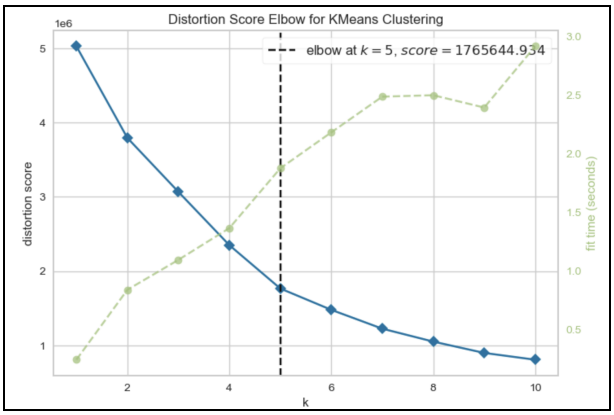
Στα πλαίσια της συγκεκριμένης ανάλυσης των δεδομένων πτήσεων για το 2015 στις ΗΠΑ, υλοποιήθηκαν 3 διαφορετικά μοντέλα εξόρυξης δεδομένων:

1. Συσταδοποίηση των πτήσεων και των αιτιών καθυστέρησης
2. Μοντέλο παλινδρόμησης για την πρόβλεψη του μεγέθους της καθυστέρησης
3. Μοντέλο κατηγοριοποίησης για την πρόβλεψη του αν μια πτήση θα φτάσει νωρίτερα, στην ώρα της ή θα καθυστερήσει

1. Συσταδοποίηση των πτήσεων και των αιτιών καθυστέρησης

Σκοπός της συγκεκριμένης ανάλυσης είναι η διερεύνηση των διαφορετικών λόγων που οδηγούν σε καθυστερήσεις πτήσεων καθώς και πόσες πτήσεις επηρεάστηκαν από αυτούς το έτος 2015. Χρησιμοποιήθηκε ο αλγόριθμος K-Means και ο καταλληλότερος αριθμός συστάδων επιλέχθηκε με την μεθόδου Elbow. Σαν input στον αλγόριθμο χρησιμοποιήθηκαν όσες πτήσεις έχουν κατηγοριοποιηθεί ως delayed στο στάδιο του ETL με τα παρακάτω features:

- DEPARTURE_DELAY
- ARRIVAL_DELAY
- AIR_SYSTEM_DELAY
- SECURITY_DELAY
- AIRLINE_DELAY
- LATE_AIRCRAFT_DELAY
- WEATHER_DELAY



Κώδικας υλοποίησης:

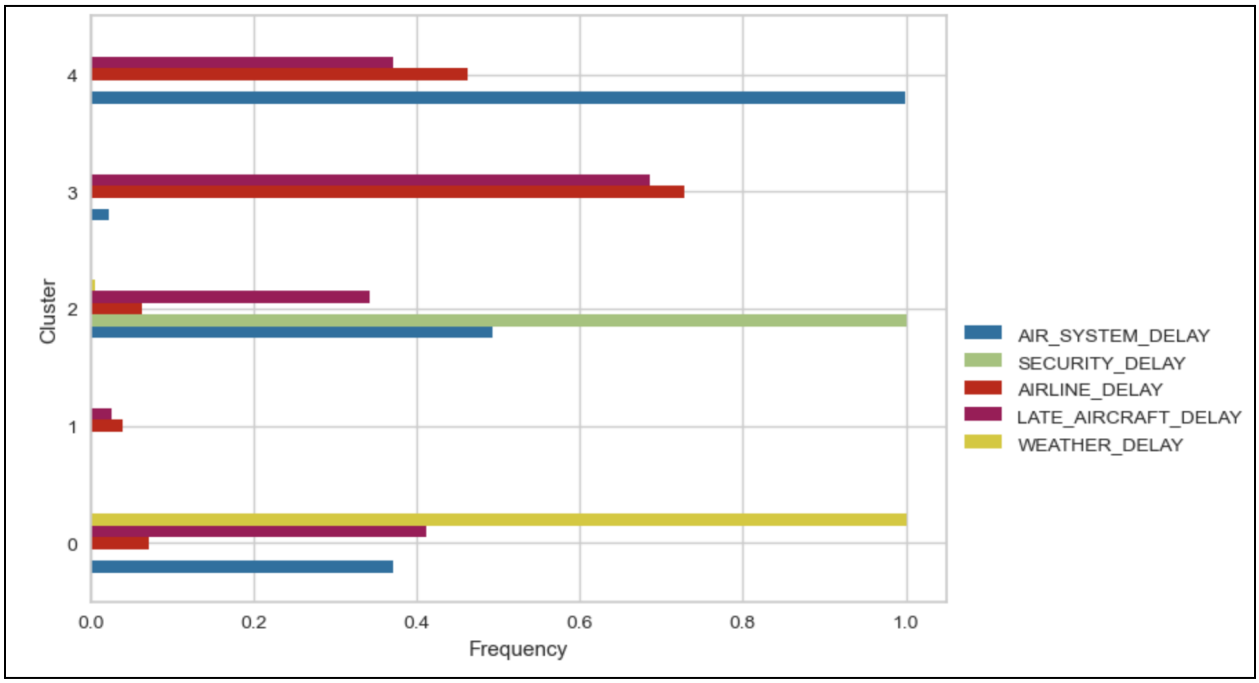
```
# read the needed columns from the fact table
fact = pd.read_csv('flights_fact.csv', usecols=['AIR_SYSTEM_DELAY',
'SECURITY_DELAY', 'AIRLINE_DELAY',
'LATE_AIRCRAFT_DELAY', 'WEATHER_DELAY', 'delayed', 'ARRIVAL_DELAY',
'DEPARTURE_DELAY'])

# keep the delayed flights only
fact = fact[fact['delayed']==1]
```

```
# turn the AIR_SYSTEM_DELAY, SECURITY_DELAY, AIRLINE_DELAY,
LATE_AIRCRAFT_DELAY AND WEATHER_DELAY in boolean
fact['AIR_SYSTEM_DELAY'] =
fact['AIR_SYSTEM_DELAY'].mask(fact['AIR_SYSTEM_DELAY'] > 1, 1)
fact['SECURITY_DELAY'] = fact['SECURITY_DELAY'].mask(fact['SECURITY_DELAY']
> 1, 1)
fact['AIRLINE_DELAY'] = fact['AIRLINE_DELAY'].mask(fact['AIRLINE_DELAY'] >
1, 1)
fact['LATE_AIRCRAFT_DELAY'] =
fact['LATE_AIRCRAFT_DELAY'].mask(fact['LATE_AIRCRAFT_DELAY'] > 1, 1)
fact['WEATHER_DELAY'] = fact['WEATHER_DELAY'].mask(fact['WEATHER_DELAY'] >
1, 1)

# drop the delayed variable
fact.drop(columns=['delayed'], inplace=True)

# scale the data
scaler = StandardScaler()
X = scaler.fit_transform(fact)
fact_scale = pd.DataFrame(X, index=fact.index,
                          columns=fact.columns)

# Create the model
kmeans = KMeans(n_clusters=5, random_state=13)

# Fit the model to the data and add the labels attribute to get the results
kmeans.fit(fact_scale)
fact['labels'] = kmeans.labels_
res = fact.groupby(by='labels').mean()
```

Ο παραπάνω κώδικας παράγει τα παρακάτω αποτελέσματα για τις συστάδες που δημιουργήθηκαν:

labels	DEPARTURE_DELAY	ARRIVAL_DELAY	AIR_SYSTEM_DELAY	SECURITY_DELAY	AIRLINE_DELAY	LATE_AIRCRAFT_DELAY	WEATHER_DELAY	% Flights
0	29.075467	28.138822	0.370872	0.0	0.071447	0.411684	1.000000	0.027361
1	10.874411	12.126160	0.000000	0.0	0.038123	0.025856	0.000000	0.319064
2	25.417989	25.443001	0.493506	1.0	0.062049	0.341510	0.005772	0.002895
3	36.541711	27.502474	0.021744	0.0	0.728671	0.685079	0.000000	0.345216
4	14.930793	25.152690	0.999631	0.0	0.461481	0.371657	0.000000	0.305464

Ο παραπάνω πίνακας ερμηνεύεται ως εξής: σε κάθε γραμμή αναφέρονται τα χαρακτηριστικά μιας συστάδας. Τα ποσοστά στις μεταβλητές AIR_SYSTEM_DELAY έως WEATHER_DELAY εκφράζουν το ποσοστό των πτήσεων εντός της συστάδας που καθυστέρησαν εξαιτίας αυτού του παράγοντα. Πιο ξεκάθαρα απεικονίζονται τα αποτελέσματα στο παρακάτω διάγραμμα.



Συστάδα 0: Το 2.7% των καθυστερημένων πτήσεων οφείλεται στις καιρικές συνθήκες (100%). Η μέση καθυστέρηση άφιξης είναι 28 λεπτά ενώ η μέση καθυστέρηση αναχώρισης είναι 29 λεπτά. Παρατηρούμε λοιπόν πως τα καιρικά φαινόμενα προκαλούν σημαντικές καθυστερήσεις. Επιπλέον, αν θεωρήσουμε πως η πτήση είχε τον αναμενόμενο χρόνο στον αέρα, το γεγονός ότι DEPARTURE_DELAY ≃ ARRIVAL_DELAY μας παραπέμπει στο ότι η καθυστέρηση οφείλεται περισσότερο στο αεροδρόμιο αναχώρησης.

Συστάδα 1: Το 31% των καθυστερημένων πτήσεων οφείλεται σε απροσδιόριστους παράγοντες. Η μέση καθυστέρηση άφιξης είναι 12 λεπτά ενώ η μέση καθυστέρηση αναχώρησης είναι 10 λεπτά. Παρατηρούμε πως οι απροσδιόριστοι αυτοί παράγοντες δεν προκαλούν σημαντικές καθυστερήσεις αλλά από την άλλη εμφανίζονται σε μεγάλο αριθμό πτήσεων.

Συστάδα 2: Το 0.2% των καθυστερημένων πτήσεων οφείλεται σε καθυστερήσεις στα συστήματα ελέγχου ασφαλείας των επιβατών. Η μέση καθυστέρηση άφιξης είναι 25 λεπτά ενώ η μέση καθυστέρηση αναχώρισης είναι 25 λεπτά. Παρατηρούμε λοιπόν τα συστήματα αυτά δεν προκαλούν συχνά καθυστερήσεις αλλά αν προκύψουν οι καθυστερήσεις είναι σημαντικές. Στην περίπτωση αυτή η καθυστέρηση οφείλεται αποκλειστικά στο αεροδρόμιο αναχώρησης.

Συστάδα 3: Το 34% των καθυστερημένων πτήσεων οφείλεται σε καθυστερήσεις της αεροπορικής εταιρείας καθώς και της καθυστερημένης άφιξης του αεροσκάφους. Η μέση καθυστέρηση άφιξης είναι 27 λεπτά ενώ η μέση καθυστέρηση αναχώρισης είναι 36 λεπτά. Παρατηρούμε λοιπόν πως οι αεροπορικές εταιρίες προκαλούν συχνά καθυστερήσεις οι οποίες είναι τις περισσότερες φορές μεγάλες. Στην συγκεκριμένη περίπτωση παρατηρούμε επιπλέον πως DEPARTURE_DELAY >> ARRIVAL_DELAY το οποίο παραπέμπει σε έναν συντομότερο χρόνο στον αέρα. Μια πιθανή εξήγηση είναι πως η αεροπορική εταιρεία ήθελε να περιορίσει την καθυστέρηση που η ίδια προκάλεσε μειώνοντας τον χρόνο πτήσης (αυξάνοντας την ταχύτητα). Μια άλλη εξήγηση είναι πως το αεροπλάνο μετά την προσγείωση έφτασε σημαντικά νωρίτερα στην πύλη του αεροδρομίου άφιξης.

Συστάδα 4: Το 30% των καθυστερημένων πτήσεων οφείλεται σε προβλήματα που προέκυψαν από τα συστήματα διαχείρισης εναέριας κυκλοφορίας. Η μέση καθυστέρηση άφιξης είναι 25 λεπτά ενώ η μέση καθυστέρηση αναχώρισης είναι 14 λεπτά. Παρατηρούμε λοιπόν πως και τα συστήματα αυτά προκαλούν συχνά καθυστερήσεις. Στην συγκεκριμένη περίπτωση παρατηρούμε επιπλέον πως DEPARTURE_DELAY << ARRIVAL_DELAY το οποίο παραπέμπει σε καθυστέρηση τόσο στο αεροδρόμιο αναχώρησης αλλά και προσγείωσης. Η εξήγηση είναι πως η κίνηση τόσο στον εναέριο όσο και τον επίγειο χώρο (taxiing) των αεροδρομίων άφιξης και αναχώρησης ήταν μεγάλη με αποτέλεσμα το αεροσκάφος να καθυστερήσει να αναχωρήσει και επιπλέον να φτάσει στην πύλη του αεροδρομίου άφιξης. Το πρόβλημα στην συγκεκριμένη περίπτωση αφορά και τα 2 αεροδρόμια.

Συμπεράσματα: Οι σημαντικότεροι παράγοντες που προκαλούν καθυστερήσεις είναι οι αεροπορικές εταιρείες και η καθυστερημένη άφιξη του αεροσκάφους (34%) και η κίνηση στον εναέριο χώρο κυκλοφορίας (30%). Σημαντική είναι και η συνεισφορά των απροσδιόριστων παραγόντων. Τέλος, η ευθύνη για την καθυστέρηση συχνά επιβαρύνει τόσο το αεροδρόμιο αναχώρησης όσο και της άφιξης.

2. Μοντέλο παλινδρόμησης για την πρόβλεψη του μεγέθους της καθυστέρησης

Σκοπός του συγκεκριμένου μοντέλου είναι η πρόβλεψη του μεγέθους της καθυστέρησης (σε λεπτά) μιας πτήσης δεδομένης της καθυστέρησης αναχώρησης της. Το μοντέλο αυτό είναι χρήσιμο για τα αεροδρόμια άφιξης καθώς η πληροφορία αυτή θα συμβάλλει στον καλύτερο προγραμματισμό των τροχοδρομήσεων και κατ' επέκταση των αφίξεων στις πύλες. Για την κατασκευή του μοντελου χρησιμοποιήθηκαν τα δέντρα αποφάσεων και πιο συγκεκριμένα ο XGBoost Regressor. Τα πεδία που χρησιμοποιήθηκαν για την εκπαίδευση του μοντέλου είναι τα εξής:

- DEPARTURE_DELAY
- SCHEDULED_TIME
- AIR_TIME
- DISTANCE

- num_passengers
- AIRLINE
- ORIGIN_AIRPORT
- DESTINATION_AIRPORT

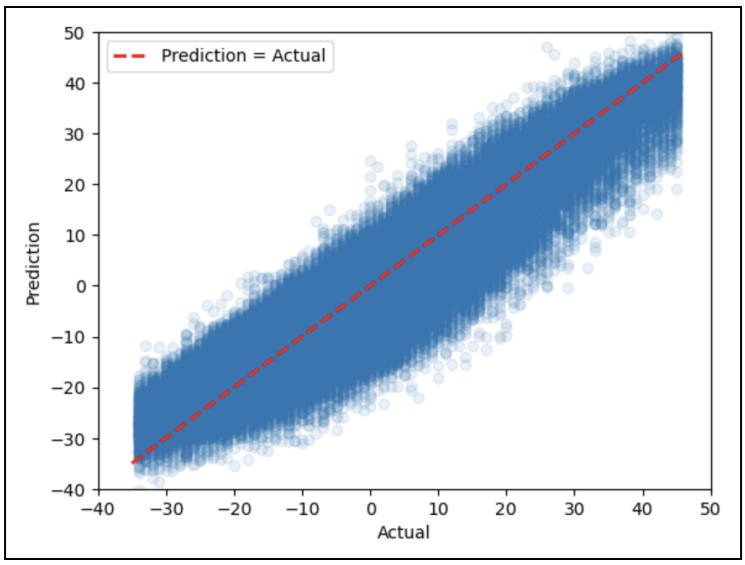
Σημείωση: Για τις κατηγορικές μεταβλητές AIRLINE, ORIGIN_AIRPORT και DESTINATION_AIRPORT χρησιμοποιήθηκαν dummy_variables, δημιουργήθηκε δηλαδή μια νέα στήλη για κάθε μοναδική τιμή τους. Επιπλέον, για την μείωση του υπολογιστικού χρόνου που απαιτείται για την εκπαίδευση του μοντέλου, χρησιμοποιήθηκαν δεδομένα πτήσεων από και πρός τα 15 μεγαλύτερα αεροδρόμια με βάση την επιβατική κίνηση.

Κώδικας υλοποίησης:

```
# read the needed columns from the fact table
fact = pd.read_csv('flights_fact.csv', usecols=['AIRLINE',
'scheduledDeparture_key', 'AIR_TIME', 'DISTANCE', 'ARRIVAL_DELAY',
'ORIGIN_AIRPORT', 'DESTINATION_AIRPORT', 'dateKey',
'TAIL_NUMBER', 'DEPARTURE_DELAY', 'scheduledArrival_key',
'SCHEDULED_TIME'],  type={'dateKey':'str'})

# read the aircraft and time (departure and arrival) dimensions
aircraft = pd.read_csv('aircraft_dim.csv', usecols=['N-Number',
'num_passengers', 'mfr_name'])
time1 = pd.read_csv('departureTime_dim.csv', usecols=['Hour', 'Minute',
'scheduledDeparture_key'])
time2 = pd.read_csv('arrivalTime_dim.csv', usecols=['Hour', 'Minute',
'scheduledArrival_key'])

# merge the above tables with the fact
fact = pd.merge(fact, time1, on='scheduledDeparture_key')
fact = pd.merge(fact, time2, on='scheduledArrival_key')

# keep only the top 15 airports and airlines to make the dataset manageable
top_15_airports = ['ATL', 'LAX', 'ORD', 'DFW', 'DEN', 'JFK', 'SFO', 'SEA',
'MCO', 'LAS', 'CLT', 'EWR', 'PHX', 'IAH',
                   'MIA']
fact = fact[(fact['ORIGIN_AIRPORT'].isin(top_15_airports)) &
(fact['DESTINATION_AIRPORT'].isin(top_15_airports))]

# get dummy variables for the categorical ones
fact = pd.get_dummies(fact, columns=['AIRLINE', 'ORIGIN_AIRPORT',
'DESTINATION_AIRPORT', 'mfr_name'])
```

```
# split the dataset in dependent and independent variables
Y = fact['ARRIVAL_DELAY']
fact.drop(columns='ARRIVAL_DELAY', inplace=True)
X = fact

# get a train and test dataset
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.25)

# train an xgb regressor
xgb_reg = xgb.XGBRegressor(max_depth=8)
xgb_reg.fit(X_train, y_train)
```

Το μοντέλο προβλέπει με **ακρίβεια 4.5 λεπτών (mean_absolute_error)** την καθυστέρηση άφιξης του αεροσκάφους στην πύλη. Ο δείκτης που δείχνει το ποσοστό της διακύμανσης της εξαρτημένης που εξηγείται από το μοντέλο μας είναι 0.85% **(R^2 = 0.85).** Παρακάτω φαίνεται και η σχέση μεταξύ των πραγματικών τιμών και των προβλεπόμενων κατά το στάδιο της δοκιμής του μοντέλου σε ενα test dataset. Παρατηρούμε πως οι αποκλίσεις είναι μικρές, της τάξης του mean_absolute_error που αναφέρθηκε παραπάνω.



3. Μοντέλο κατηγοριοποίησης για την πρόβλεψη του αν μια πτήση θα καθυστερήσει ή οχι

Σκοπός του παραπάνω μοντέλου είναι η πρόβλεψη και κατηγοριοποίηση μελλοντικών πτήσεων σχετικά με την χρονική τους απόδοση σε καθυστερημένες και μη.
Για την κατασκευή του παραπάνω μοντέλου χρησιμοποιήθηκαν τα παρακάτω πεδία:

- **DISTANCE**
- **dateKey**
- **LONGITUDE_x**
- **LONGITUDE_y**
- **AIRLINE**
- **total_minutes_dep**
- **total_minutes_arr**

Σημείωση: Για τη κατηγορική μεταβλητή AIRLINE χρησιμοποιήθηκε η μέθοδος one-hot encoding δημιουργήθηκε δηλαδή μια νέα στήλη για κάθε μοναδική τιμή δηλαδή για κάθε αεροπορική εταιρεία. Συγκεκριμένα, το dataset αποτελείται από 13 διαφορετικές αεροπορικές με αποτέλεσμα την δημιουργία 13 νέων στηλών τύπου binary. Επιπρόσθετα, οι στηλες **total_minutes_dep** και **total_minutes_arr** περιέχουν τις ώρες αναχώρησης και άφιξης στο αεροδρόμιο και δημιουργήθηκαν ύστερα από επεξεργασία προκειμένου τα δεδομένα τους να αναπαρασταθούν στην βέλτιστη μορφή για την εκπαίδευση του μοντέλου. Τα δεδομένα του dataset περιείχαν τα γνωρίσματα **scheduledDeparture_time, scheduledArrival_time** τα οποία αναφέρονταν στην προγραμματισμένη ώρα και λεπτό αναχώρησης και άφιξης στο αεροδρόμιο. Ωστόσο τα δεδομένα ήταν σε μορφή ακέραιου αριθμού απεικονίζοντας την ακριβή ώρα ως τετραψήφιο νούμερο το οποίο κυμαίνεται σε κλίμακα [0,2359]. ‘Ετσι, προκειμένου να γίνει σωστή διαχείριση των δύο αυτών γνωρισμάτων έγινε διαχωρισμός του τετραψήφιου αριθμού σε δύο στήλες **Hours, Minutes** αντιστοιχίζοντας τα 2 πρώτα ψηφία με την ώρα και τα δύο τελευταία με τα λεπτά. Στη συνέχεια ακολουθήσαμε την παρακάτω πρακτική για να συνενώσουμε τις δύο αυτές διαστάσεις σε 1:

- ➢ Πολλαπλασιάσαμε κάθε ώρα επι 60
- ➢ Προσθέσαμε την νέα ώρα με τα αντίστοιχα λεπτά
- ➢ Αποθηκεύσαμε τον νέο αποτέλεσμα στις αντίστοιχες νέες στήλες

Παράδειγμα:

Hour	Minute	Total Minutes
1	30	90
2	45	165
3	15	95

Το αποτέλεσμα είναι η δημιουργία 2 νέων στηλών (**total_minutes_dep ,total_minutes_arr)** που περιέχουν την πληροφορία για τον χρόνο αναχώρησης και άφιξης εκφρασμένη σε μικρότερη κλίμακα και κατανεμημένη καλύτερα για να αξιοποιηθεί από το μοντέλο.

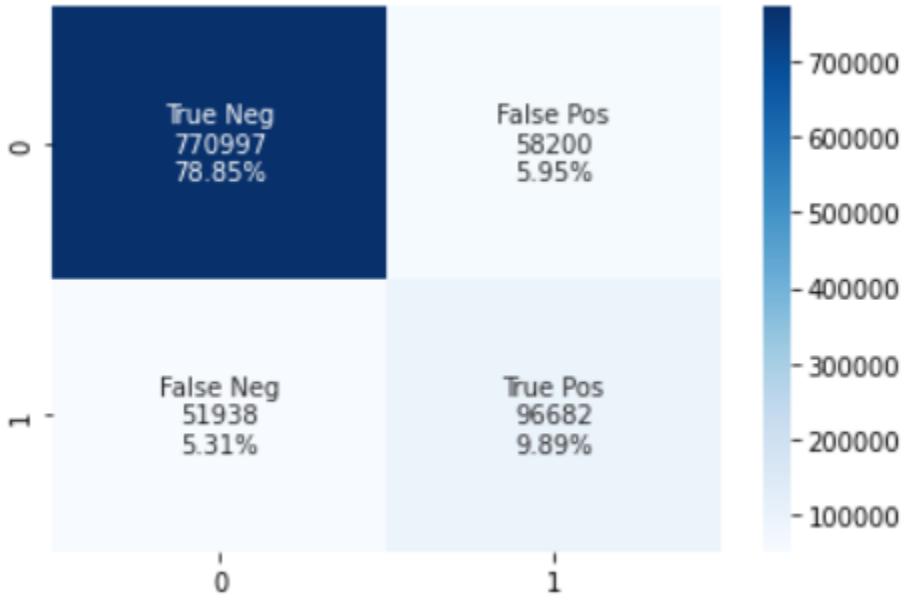
Κώδικας υλοποίησης:

```
# Remove duplicate columns
ex_df = df.loc[:, ~df.columns.duplicated()]
ex_df = ex_df.sample(frac=0.7, random_state=0)
x= ex_df.drop(['delayed','DEPARTURE_TIME',
'DEPARTURE_DELAY','LATITUDE_x','LATITUDE_y','AA','OO','SCHEDULED_TIME','B6'
,'AS', 'EV' ,'HA', 'MQ','UA', 'US', 'VX'],axis=1)
y= ex_df['delayed']
x_train, x_test, y_train, y_test= train_test_split(x, y, test_size=
0.25,random_state=0)
y_train.value_counts()
# define the SMOTE model
smote = SMOTE()
# fit and transform the training data
X_train_resampled, y_train_resampled = smote.fit_resample(x_train, y_train)
X_train_resampled
%%time
# Create the classifier
clf = XGBClassifier(max_depth=25,n_estimators=100)
# Fit the classifier to the training data
clf.fit(X_train_resampled, y_train_resampled)
# Make predictions on the test data
predictions = clf.predict(x_test)
# Print the classification report(Gini)
report = classification_report(y_test, predictions)
print(report)
# Print the F1 score
f1 = f1_score(y_test, predictions, average="micro")
print("F1 score:", f1)
# Plot the feature importance scores, excluding the first two columns
xgb.plot_importance(clf,importance_type='weight', max_num_features=11)
import seaborn as sns
import numpy as np
# compute the confusion matrix
confusion_matrix1 = confusion_matrix(y_test, predictions)
group_names = ['True Neg','False Pos','False Neg','True Pos']
group_counts = ['{0:0.0f}'.format(value) for value in
                confusion_matrix1.flatten()]
group_percentages = ['{0:.2%}'.format(value) for value in
                     confusion_matrix1.flatten()/np.sum(confusion_matrix1)]
labels = [f'{v1}\n{v2}\n{v3}' for v1, v2, v3 in
          zip(group_names,group_counts,group_percentages)]
```

```
labels = np.asarray(labels).reshape(2,2)
sns.heatmap(confusion_matrix1, annot=labels, fmt='', cmap='Blues')
# Compute the false positive rate and true positive rate at different
classification thresholds
fpr, tpr, thresholds = roc_curve(y_test,predictions)
# Compute the area under the curve (AUC)
auc = roc_auc_score(y_test, predictions)
# Create a figure and axis object with a specific size and aspect ratio
fig, ax = plt.subplots(figsize=(8, 6), dpi=80)
```

Ύστερα από διάφορους στατιστικούς ελέγχους, μέσα από ένα σύνολο περίπου 100 διαφορετικών γνωρισμάτων χρησιμοποιήθηκαν τελικά μόνο τα 7 παραπάνω καθώς ήταν αυτά που εμφάνιζαν την μεγαλύτερη συσχέτιση ως προς το πεδίο που έπρεπε να προβλεφθεί. Επιπρόσθετα, ένας ακόμα κύριος λόγος που επιλέχθηκαν μόνο τα 7 σημαντικότερα γνωρίσματα είναι για την αποφυγή της υπερμοντελοποίησης και την δημιουργία ενός προβλεπτικού μοντέλου το οποίο μπορεί να γενικευθεί σε άλλες περιπτώσεις. Για την δημιουργία του προβλεπτικού μοντέλου αρχικά το dataset χωρίστηκε σε δύο κατηγορίες το training set(75%) και το testing set(25%) προκειμένου το μοντέλο να εκπαιδευτεί και στη συνέχεια να εξεταστεί σε νέα δεδομένα ελέγχοντας την απόδοση του. Για την εκπαίδευση χρησιμοποιήθηκαν διάφορες προσεγγίσεις επιβλεπόμενης μάθησης με κυρίαρχη να αποτελεί αυτή των δέντρων αποφάσεων. Συγκεκριμένα υλοποιήθηκαν 3 διαφορετικά μοντέλα κατηγοριοποίησης χρησιμοποιώντας το DecisionTreeClassifier και το XGBClassifier(eXtreme Gradient Boosting). Τα πρώτα δύο μοντέλα αφορούσαν το DecisionTreeClassifier και εκπαιδεύτηκαν πάνω στο ίδιο training set με μόνη διαφορά το κριτήριο εκπαίδευσης το οποίο στο πρωτο ορίστηκε με τον δείκτη gini ενώ στο δεύτερο με την εντροπία. Ως αποτέλεσμα τα δύο αυτά μοντέλα εμφάνισαν πανομοιότυπα σχεδόν αποτελέσματα προβλέποντας την σωστή κατηγορία με ποσοστό 75% κατά μέσο όρο. Ωστόσο παρά το σχετικά καλό ποσοστό πρόβλεψης παρατηρήθηκε μεγάλη ανισορροπία στο ποσοστό σωστών προβλέψεων μεταξύ των 2 κλάσεων. Συγκεκριμένα, παρατηρήθηκε σχεδόν άριστη πρόβλεψη(≈90%) για την κλάση με τιμή 0 που υποδεικνύει ότι η πτήση δεν θα αργήσει ενώ η κλάση με τιμή 1 εμφάνισε πολύ χαμηλά ποσοστά σωστών προβλέψεων(=25%) με αποτέλεσμα να εξισορροπείται η μέση απόδοση στο 75% ενώ το μοντέλο στην πραγματικότητα είχε πολύ χαμηλή ισχύ. Κατά συνέπεια, παρατηρήθηκε η τεράστια ανισορροπία μεταξύ του αριθμού των καθυστερημένων πτήσεων και των μη σε ποσοστό 15% έναντι 85% με αποτέλεσμα το μοντέλο να εκπαιδευεται με ‘imbalanced data’ και να μην αποδίδει σωστά για τις καθυστερημένες πτήσεις. Για την αντιμετώπιση του προβλήματος χρησιμοποιήθηκε η μέθοδος **SMOTE** (Synthetic Minority Oversampling Technique) η οποία χρησιμοποιεί την μέθοδο oversampling δημιουργώντας τεχνητά δεδομένα για την κλάση που υστερεί εξισορροπώντας τις δύο κλάσεις.Τα συνθετικά δεδομένα δημιουργούνται μέσω του αλγορίθμου K-Nearest Neighbors προκειμένου να ταυτίζονται και να ακολουθούν παρόμοια κατανομή με τα πραγματικά δεδομένα. Όσον αφορά το τελευταίο μοντέλο χρησιμοποιήθηκε ο XGBClassifier ο οποίος αποτελεί μία βελτιωμένη προσέγγιση των απλών δέντρων αποφάσεων χρησιμοποιώντας ένα σύνολο

πολλαπλών δένδρων συνδυάζοντας τα αποτελέσματα τους για να κάνει προβλέψεις υψηλής ακρίβειας.
Στη συγκεκριμένη περίπτωση το μοντέλο εκπαιδεύτηκε με ένα σύνολο 100 διαφορετικών δέντρων που το καθένα έχει μέγιστο βάθος 25 κόμβων. Για την εκπαίδευση του μοντέλου πάρθηκε ένα τυχαίο δείγμα που αποτελούνταν από το 70% των δεδομένων πάνω στο οποίο εφαρμόστηκε ο αλγόριθμος SMOTE.Τα αποτελέσματα στη περίπτωση αυτή ήταν βελτιωμένα προβλέποντας την σωστή κατηγορία με ποσοστό 89% κατά μέσο όρο και συγκεκριμένα την κατηγορία 0 με ποσοστο 94% και την κατηγορία 1 με ποσοστό 65%.
Παρακάτω απεικονίζεται καλύτερα το ‘confusion matrix’ :



	precision	recall	f1-score	support
0	0.94	0.93	0.93	829197
1	0.62	0.65	0.64	148620
accuracy			0.89	977817
macro avg	0.78	0.79	0.79	977817
weighted avg	0.89	0.89	0.89	977817

F1 score: 0.8873633819007033

Το παραπάνω ‘confusion matrix’ αποτελεί ένα θερμικό χάρτη ο οποίος δείχνει την απόδοση κάθε κατηγορίας. Για παράδειγμα, στο πρώτο τεταρτημόριο εμφανίζεται συνολική συγκέντρωση του 79% των δεδομένων το οποιο σημαίνει ότι στην κατηγορία 0 (πτήσεις που δεν αργούν) το μοντέλο προέβλεψε το 79% των δεδομένων της κατηγορίας σωστά.

ΣΥΜΠΕΡΑΣΜΑΤΑ: Ένα τέτοιο μοντέλο πρόβλεψης της κατάστασης των πτήσεων μπορεί να χρησιμοποιηθεί και να ωφελήσει διάφορους φορείς. Οι ταξιδιώτες μπορούν να χρησιμοποιήσουν αυτό το μοντέλο για να προγραμματίσουν τα ταξίδια τους και να κάνουν εναλλακτικές ρυθμίσεις σε περίπτωση που η πτήση τους προβλέπεται να καθυστερήσει. Παράλληλα, οι αεροπορικές εταιρείες μπορούν να χρησιμοποιήσουν αυτό το μοντέλο για να βελτιστοποιήσουν τις δραστηριότητές τους και να προγραμματίσουν τη συντήρηση ή άλλες δραστηριότητες για αεροπλάνα που προβλέπεται ότι θα έχουν μεγαλύτερες διακοπές. Τέλος, από το συγκεκριμένο μοντέλο μπορούν να επωφεληθούν και τα ταξιδιωτικά γραφεία ενημερώνοντας τους πελάτες τους σχετικά με την κατάσταση των πτήσεων τους με σκοπό να κάνουν εναλλακτικές ταξιδιωτικές ρυθμίσεις εάν αυτό είναι απαραίτητο.

Γενικά Συμπεράσματα

Μετά το πέρας της παραπάνω ανάλυσης προκύπτουν τα εξής συμπεράσματα. Το έτος 2015 το 35% των συνολικών πτήσεων αφίχθη νωρίτερα από την προγραμματισμένη ώρα, το 50% έφτασε την προγραμματισμένη ώρα ενώ το 15% καθυστέρησε (>15 λεπτά). Οι μεγαλύτερες καθυστερήσεις συσσωρεύονται σε πτήσεις μεταξύ πολιτειών του βορειοανατολικού τμήματος της χώρας. Ο καλύτερος μήνας πτήσεων είναι ο Σεπτέμβρης ενώ οι περισσότερες καθυστερήσεις παρατηρούνται τους μήνες Φεβρουάριος και Ιούνιος εξαιτίας του καιρού και των καλοκαιρινών διακοπών αντίστοιχα. Παράλληλα, παρατηρήθηκε πως οι πρώτες πρωινές ώρες (αναχώρησης και άφιξης) είναι ιδανικές για την αποφυγή καθυστερήσεων. Όσον αφορά τις αεροπορικές εταιρείες, η Delta Airlines είναι η πιο χρονικά αξιόπιστη με μόλις το 10% των πτήσεων της να καθυστερούν σε αντίθεση με την Spirit Air της οποίας οι πτήσεις καθυστερούν με πιθανότητα 22%. Απο την πλευρά των αεροδρομίων καλύτερη απόδοση παρουσιάζει το Hartsfield Jackson Atlanta International Airport (11%) ενώ την χειρότερη το Los Angeles International Airport (18%). Το αεροσκάφος με τις περισσότερες καθυστερήσεις είναι το Boeing 777-232LR με 38% πιθανότητα καθυστέρησης ενώ το καλύτερο είναι το Boeing 737-732 με πιθανότητα 7%.