**İçindekiler**

[**Giriş**](#_yrybshm14lvw) **4**

[**Veri Seti ve Ön İşleme**](#_eqnapxs87o6) **8**

[Havayolu Şirketlerinin İncelenmesi](#_dc8jpeujjxya) 8

[Havaalanı Bilgilerinin İncelenmesi](#_2boc08mpa6xb) 9

[Uçuş Verilerinin İncelenmesi](#_7hv9dbtwh2r9) 10

[Gecikme Kabul Edilmeyen Verilerin Temizlenmesi](#_z0z35do2xw53) 12

[Boş Verilerin Temizlenmesi](#_k3sq9iilbm9n) 12

[Temel İstatistik Bilgilerin İncelenmesi](#_gfxo6u51jsyd) 13

[Uçuş Verileri Arasındaki Korelasyonun İncelenmesi](#_19ryqdsm614w) 15

[**Yöntem**](#_7itp3dsns39w) **16**

[Uçuş Gecikmesini Etkileyen Faktörlerin Belirlenmesi](#_v6ew8vrdse2f) 16

[Uçuşların Genel Gecikme Durumu](#_vw89oza5hvot) 17

[Gecikmeler kalkışta mı varışta mı gerçekleşiyor?](#_dvdtaqas7tvu) 17

[Gecikmelerle ilişkili değişkenlerin belirlenmesi](#_9z7p4s667oa0) 18

[Sayısal Değişkenlere Göre Uçuşların ve Gecikmelerin İncelenmesi](#_igcz5tw6owzf) 19

[Kategorik Değişkenlere Göre Uçuşların ve Gecikmelerin İncelenmesi](#_ec98fpxd2plj) 21

[Havaalanlarındaki Uçuş Yoğunluğunun Dağılımının İncelenmesi](#_2tg5ntg3fnu9) 21

[Havayolu Şirketlerine Göre Uçuş Sayılarının ve Gecikmelerin İncelenmesi](#_6j2nlm9zmzsw) 21

[Havaalanları ve Havayolu Şirketleri Arasındaki Gecikme Korelasyonu](#_aconlcp67tjy) 23

[Makine Öğrenmesi Metodlarıyla Tahminleme](#_yw3u03wuyq5k) 25

[Modelleme İçin Feature Selection](#_o5xz5y4b1nbp) 25

[Ön Hazırlık](#_2t8jqhf15qwm) 26

[Tahminleme](#_49dviug8bgld) 28

[Model Doğruluk Analizi](#_goiilwc6x46q) 32

[**Sonuç**](#_vqdbprmwiqmp) **41**

[**Kaynakça**](#_95rio4fel6ms) **43**

**Uçuş Gecikmelerini Etkileyen Faktörler ve Gecikme Tahminleme Üzerine Bir Çalışma: 2015 Yılı Amerika Uçuşları Örneği**

# ÖZET (ABSTRACT)

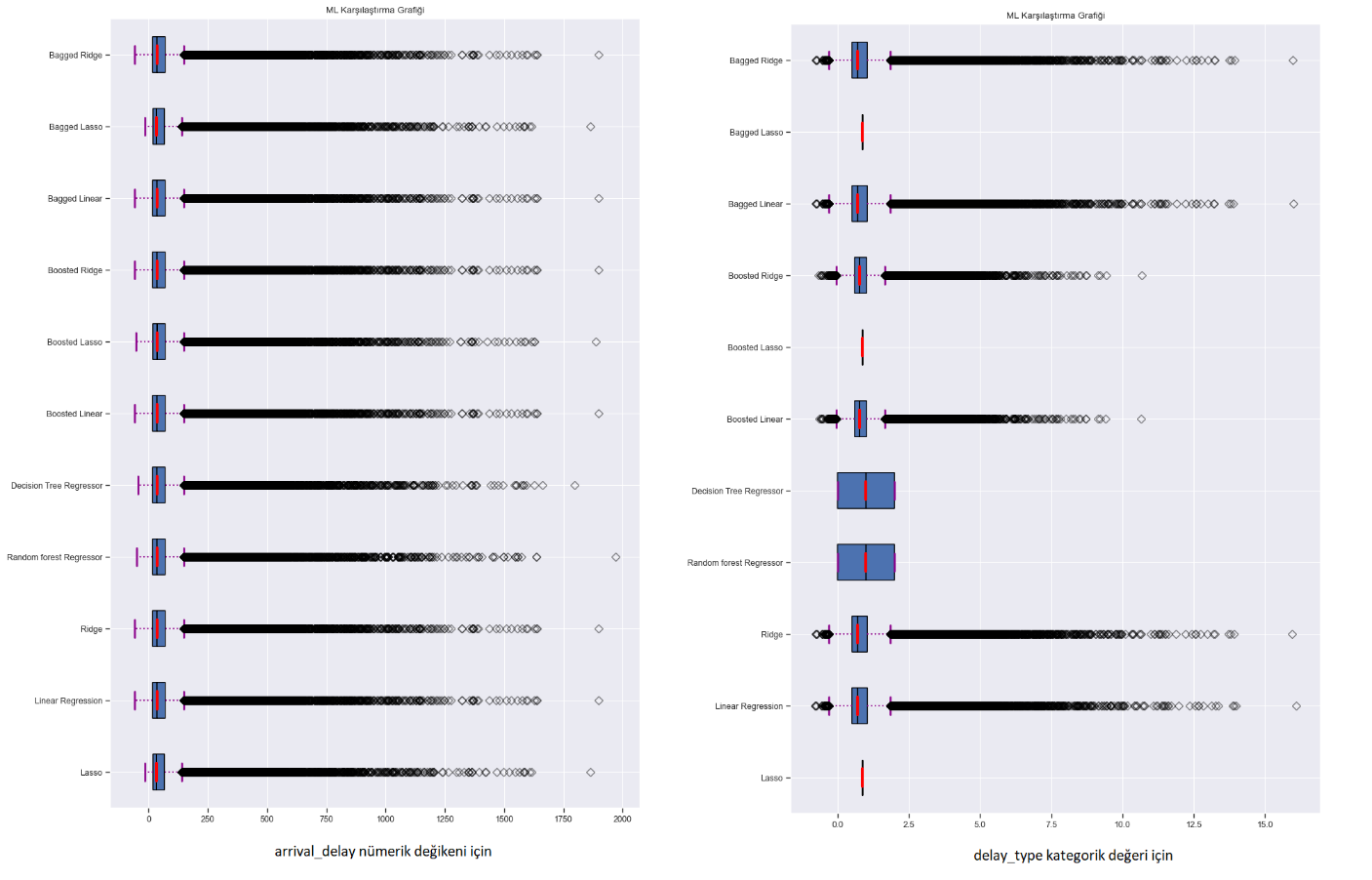
Bu çalışmanın amacı uçuş gecikmelerinin tahmin edilebilmesi için uçuş gecikmelerini etkileyen değişkenleri ve bunların etkilerini belirlemek, detaylı olarak incelemek ve elde edilen bulgulara uygun şekilde ders kapsamında işlenen makine öğrenmesi yaklaşımları ve uygulamalarda görülen diğer farklı modeller de kullanılarak gecikmeler üzerine tahmin modellemeleri yapmaktır. Çalışmanın amacına ulaşılabilmesi için ABD Ulaştırma Bakanlığı'nın (DOT) Ulaştırma İstatistikleri Bürosu tarafından yayınlanan 2015 yılına ait uçuş veri seti kullanılmıştır.

Çalışmada ilk olarak keşifçi veri analizi yöntemleriyle veri ön işleme aşamasında veri seti temizlenmiş, organize edilmiş ve temel istatistiksel değişkenler incelenerek temel çıkarımlar yapılmıştır. Daha sonra temizlenen veri seti üzerinde seçilen yöntemlerle analizler yapılmıştır. Yöntem bölümü 2 parçadan oluşmaktadır. İlk aşamada uçuşlarda yaşanan gecikmelerin hangi değişkenlere ve durumlara bağlı olarak gerçekleştiği incelenmiştir. Bulgular kalkış gecikmelerinin varış gecikmelerini tetiklediğini ve genel gecikmelerin de kalkılan ve varılan havaalanı ile havayolu şirketleri ile yakından ilişkili olduğunu göstermiştir. Tüm değişkenler Jupyter Notebook üzerinde detaylı olarak incelenmiş olup oluşturulan tablo ve grafiklerden örnekleme yoluyla rapora ekleme yapılmıştır.

İlk aşamadaki bulgular ışığında taminleme için veri seti training ve test olarak ayrıldıktan ve modelleme yapıldıktan sonra 'Lasso', 'Linear Regression', 'Ridge', 'Random forest Regressor', 'Decision Tree Regressor', ‘Boost regression', ve 'Bagged Regression’' yöntemleri ile tahminlemeler yapılmıştır. Modellerin sistematik olarak analiz edilebilmesi ve birbirleriyle karşılaştırmaların yapılabilmesi için Katardjiev, McKeever, ve Andreas Hamfelt’in (2019) model karşılaştırmalarını yapmak için önerdiği Box Whisker Plot yöntemi uygulanmıştır.

İlk olarak varış gecikmesi üzerine modelleme yapılmıştır. Bu tahminlemelere bakıldığı zaman en iyi modelin hem hız hem de tahmin başarısı açısından Linear Regression modeli olduğu görülmektedir. Boosted Linear, Boosted Lassoi Boosted Ridge ve Bagged Linear modelleri de Linear Regression modeline yakın başarı göstermişlerdir ancak algoritmaların yüksek çalışma süreleri nedeniyle optimal olmadıkları görülmüştür.

Gecikme süre aralıklarına göre oluşturulan delay\_type kategorik değerlerine göre tahminleme yapıldğında ise iki model dışındaki tüm modellerin başarılarının oldukça düştüğü görülmektedir. Box Whisker Plot ile bunu kontrol edildiği zaman, Random forest Regressor ve Decision Tree Regressor modellerinin yüzde yüz başarı ile tahminleme yaptığı görülmektedir. Bu durum feature selection adımında seçtiğimiz değişkenlerin optimal şekilde seçildiğini göstermektedir.

Yapılan 2 farklı modele göre kullanılan metodların başarı grafikleri aşağıda verilmiştir:

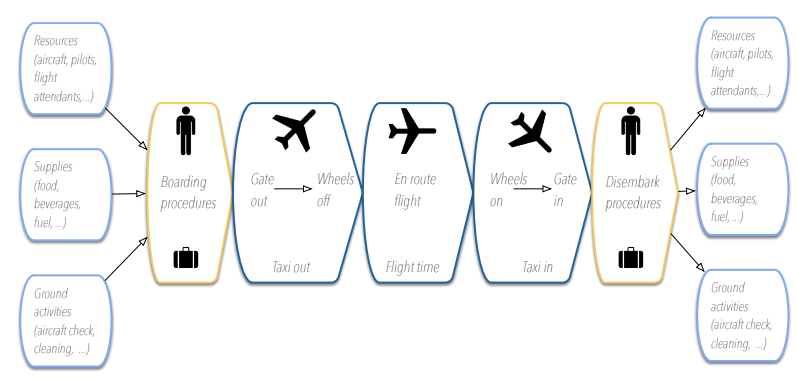
Çalışmadaki analiz ve tahminlemelerden yola çıkılarak havayolu şirketlerinin ve havaalanlarının tahmine dayalı veri analitiği ve makine öğrenmesi metodlarına yönelerek gecikme riskini azaltarak müşterileri için daha iyi bir deneyim sağlamanın yanı sıra zamandan, paradan ve kaynaklardan tasarruf sağlayabileceği söylenebilir. Ayrıca çalışmadaki analiz sonuçları aynı zamanda havayou şirketlerinin check-in işlemi, bagaj taşıma ve güvenlik kontrolleri için havaalanına daha fazla personel ve kaynak sağlayarak gecikmeleri azaltabileceğini işaret etmektedir. İncelenen veri setinin sadece bir seneye ve tek bir ülkeye ait olması, temizleme aşamasında verilerin doldurulması/kurtarılması yerine boş değerlerin silinmesi yolunun seçilmesi çalışmanın kısıtlılıkları arasındadır. Bu durum göz önüne alınarak ilerleyen çalışmalarda daha kapsamlı veri setleri kullanılarak çalışmalar yapılmalıdır.

# **Giriş**

Hava ulaşımı, ulusal ve uluslararası sınırları birbirine bağlayan en hızlı toplu taşıma modlarından biridir ve farklı coğrafi konumlardan insanların mesafe sınırlarını aşmasına ve diğer coğrafi bölgelere kişisel, iş, tıbbi ve turizm amaçlı seyahat etmesine olanak tanır. Havayolu ulaşımı, yolculuk süresinden tasarruf ederek en hızlı ulaşımı sağlasa da önemli bir diğer özelliği de yolculara sağladığı kolaylıklar ve konfor düzeyidir (Tiwari et al., 2019).

Hava ulaşımı sistemi üç ana bileşenden oluşur: havaalanları, ATC (hava trafik kontrol) sistemi ve havayolları. Haavaalanları, uçuşlarının başında ve sonunda ve uçuş aralarında yolcu ve yük kargo talebine hizmet veren havayolu uçaklarını barındırır. ATC, havaalanlarındaki ve havaalanları arasındaki havadaki uçuşların güvenliğini, verimliliğini ve etkinliğini sağlamak için izler ve kontrol eder. ATC, hava trafiği olaylarına ve kazalarına (emniyet), önemli uçak gecikmelerine (etkinlik) ve ek havayolu ve yolcu maliyetlerine (verimlilik) neden olabilecek çatışmaları önlemelidir (Teodorović & Janić, 2022).

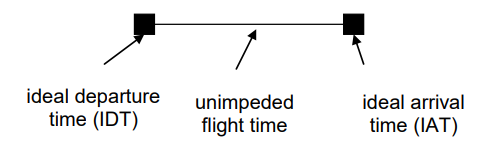
Bununla beraber havaalanları içindeki operasyonlar da bu süreçlerin en iyi şekilde yönetilmesine etki eden en önemli etmenlerden biridir. Hava taşımacılığı ile uğraşan ticari bir uçuş şirketi için yer operasyonları, yolcuların ve kargonun kalkıştan önce uçağa güvenli bir şekilde yüklenmesini sağlayan havaalanı işlem prosedürleriyle ilgilenen süreçleri kapsar. Bu departman: yolcuların biletleme, karşılama, kontrol, yönlendirme ve müşteri hizmetleri süreçlerini yönetirken aynı zamanda bagajların ve kargoların da elleçleme, kontrol, ULD (yükleme için bagaj hazırlama), uygun yük dağılımına göre uçağa yerleştirilmesi; yer destek ekipmanlarını kullanma ve uçuş öncesi ve sonrasındaki uçak hareketlerinin ve park işlemlerinin yürütülmesinden (pushback and headset), uçuş zemininin uygun hale getirilmesinden (ör: buzlu zeminin temizlenmesi) ve personel eğitiminden de sorumludur (The Irish Aviation Authorit, n.d.). Özetle, ticari havacılık karmaşık bir dağıtılmış ulaşım sistemidir. Sorunsuz ve güvenli operasyonlar sağlamak için düzenlemeye ihtiyaç duyan değerli kaynaklar, talep dalgalanmaları ve karmaşık bir başlangıç-varış dengesi ile ilgilenir. Şekil 1, bir ticari uçuşun tipik bir operasyonunu göstermektedir (Carvalho et al., 2020).



Şekil 1: Tipik bir ticari uçuş operasyonu (Carvalho et al., 2020)

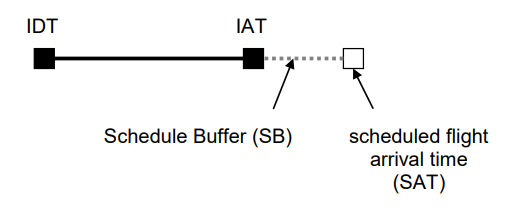
Uçaklarda seyahate olan talep nedeniyle uçuşların kullanımındaki olağanüstü artış ve beklenmedik durumlar, bu karmaşık ve yoğun operasyonda kolayca kesintilere ve gecikmelere neden olabilir. Gecikme, herhangi bir ulaşım sisteminin en çok hatırlanan performans göstergelerinden biridir. Uçuş gecikmesi, bir havayolunun kalkış varış saatinden sonra inmesi veya kalkması veya varış saatinin verilen zamandan daha büyük olması durumunda iniş saatinden sonra meydana gelmesidir. Kısacası gecikme, bir uçuşun geciktiği veya ertelendiği bir süredir. Dolayısıyla, bir gecikme, bir uçağın planlanan kalkış veya varış zamanları ile gerçek zamanları arasındaki farkla temsil edilebilir (Hemadri & Kumar Raja, 2022; Shruti et.al., 2022).

Gecikme süreci aşağıdaki gibi özetlenebilir (Ball et al., 2010). İdeal süreç, tarifeli uçuşun ideal kalkış saati (IDT) ile başlar. IDT, verimli ekip programları ve filo planları oluşturmak için yalnızca tercih edilen yolcu seyahat sürelerini değil, aynı zamanda dahili havayolu kısıtlamaları da dikkate alınarak hesaplanır. Bu sürecin bir parçası olarak havayolu, uçuş için filosundan en uygun uçak tipini seçer. Bu uçağın özelliklerini kullanarak ve kalkıştan varış noktasına en uygun, engelsiz yörüngede uçabileceğini varsayarak, Şekil 2.1.'de gösterildiği gibi ideal bir varış zamanı (IAT) hesaplanabilir.



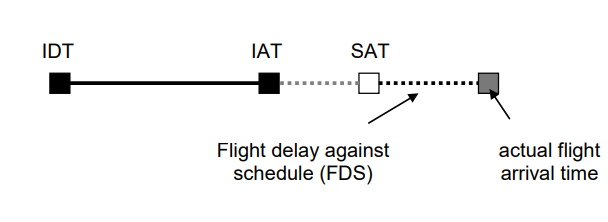
Şekil 2.1. İdeal Uçuş (Ball et al., 2010)

Şekil 2.2.'de gösterildiği gibi, havayolları olası gecikmeleri hesaba katarak planlanan ideal varış zamanına ek bir süre ekleyerek planlanan uçuş varış zamanını (SAT) belirler. olduğunda tarifeli uçuş süresi artar. Aradaki ek süreye schedule buffer (SB) adı verilmektedir.



Şekil 2.2. Planlanmış Uçuş (Ball et al., 2010)

Planlanan uçuşa ek yaşanan gecikmeler Şekil 2.3.’te verilmiştir. Program dışı uçuş gecikmeleri (FDS), çoğu hava ulaşım sistemindeki sıkışıklıkla ilgili olan fazla seyahat süresini yansıtır. Bununla birlikte, belirli bir uçuş için SB önceden bilinirken, FDS bilinmez.



Şekil 2.3. Program Dışı Uçuş Gecikmesi (Ball et al., 2010)

Program dışı uçuş gecikmeleri hava durumu (ör. kış fırtınaları, şiddetli rüzgarlar vb.); havacılık sistemi (ör. hava trafik kontrolü, yoğun trafik hacmi vb.); özel günler ve tatil dönemleri gerçekleşen aşırı yolcu ve uçuş yoğunluğu sebebiyle terminal sınırları, pistler ve hava sahasındaki yoğunluk ve personelin bu yoğunluğa yetişememesi; uçaklarda ihtiyaç duyulan bakım ve temizlik operasyonları; güvenlik ile ilgili durumlar; önceki uçuş programları sebebiyle gerçekleşen kaymalar gibi bir ço farklı sebepten meydana gelebilir (Bureau of Transportation Statistics (BTS), 2009; Hemadri & Kumar Raja, 2022; Shruti et.al., 2022).

Uçuş gecikmeleri hava ulaşımının tüm paydaşlarını negatif olarak etkilemektedir. Uçağın rötar yapması yolcular için seyahatin zorlaşmasına, programların aksamasına, zaman ve para kaybına neden olmaktadır. Gecikmelerin belirsizliği göz önüne alındığında, yolcular genellikle uçuşları için zamanında varmalarını sağlamak için saatler önceden havaalanı seyahatlerini planlarlar, bu da yolculuk maliyetlerini artırır ve gecikmeler yolcuların zamanlarını boşa harcamalarına sebep olur.

Havayolu şirketleri için uçuş gecikmeleri büyük ekonomik kayıplar getirir. Öte yandan, havayolları cezalara, para cezalarına ve mürettebatın ve uçakların havaalanlarında tutulması gibi ek işletme maliyetlerine maruz kalmaktadır. Ayrıca sürdürülebilirlik açısından gecikmeler, yakıt tüketimini ve gaz emisyonlarını artırarak çevresel zararlara da neden olabilir. Gecikmeler aynı zamanda havayollarının pazarlama stratejilerini de negatif etkiler. Gecikme seviyeleri ve ücretler, uçak boyutları, uçuş sıklığı ve havayolu hizmeti ile ilgili şikayetler arasında bir ilişki olduğu bilinmektedir. Havaalanı için uçuşun gecikmesi havaalanının normal işleyişini ciddi şekilde etkiliyor. Uçuş gecikmeleri, havaalanları ve havayolu yöneticilerinin taktiksel ve operasyonel kararlarını zorlaştırmaktadır (Shruti et.al., 2022; Carvalho et al., 2020).

Hava taşımacılığı yönetimini iyileştirmek için gecikmeleri neyin tetiklediğini anlamak günümüzde bir gereklilik haline gelmiştir. Havayolu şirketleri uçuş gecikmelerini öngörebilirse aksaklıkları daha iyi yönetebilir, son dakika yeniden planlamasını azaltabilir ve seyahat deneyimini kolaylaştırabilir. Bu nedenle havayolu şirketleri, uçuş gecikmelerini tahmin etmek için veri Analitiğine yönelmektedir. Bunun sağlanabilmesi ve tüm uçuş ekosistemlerinin daha iyi anlaşılabilmesi için ticari havacılıkta her an büyük miktarda veri toplanır ve veritabanlarında saklanır. Bu, havayollarının bilet satın alma, koltuk seçimi, bagaj, biniş, kara taşımacılığı vb. dahil olmak üzere uçuş öncesinden uçuş sonrasına tüm süreçleri dönüştürmeye ve gecikmelerden kaynaklı zararları azaltma çabalarına yardımcı olmaktadır.

Bu çalışmanın amacı uçuş gecikmelerinin tahmin edilebilmesi için uçuş gecikmelerini etkileyen değişkenleri ve bunların etkilerini belirlemek, detaylı olarak incelemek ve elde edilen bulgulara uygun şekilde ders kapsamında işlenen makine öğrenmesi yaklaşımları ve uygulamalarda görülen diğer farklı modeller de kullanılarak gecikmeler üzerine tahmin modellemeleri yapmaktır.

# **Veri Seti ve Ön İşleme**

* Veri Seti: 2015 Flight Delays and Cancellations
* Kaynak: <https://www.kaggle.com/datasets/usdot/flight-delays?datasetId=810>

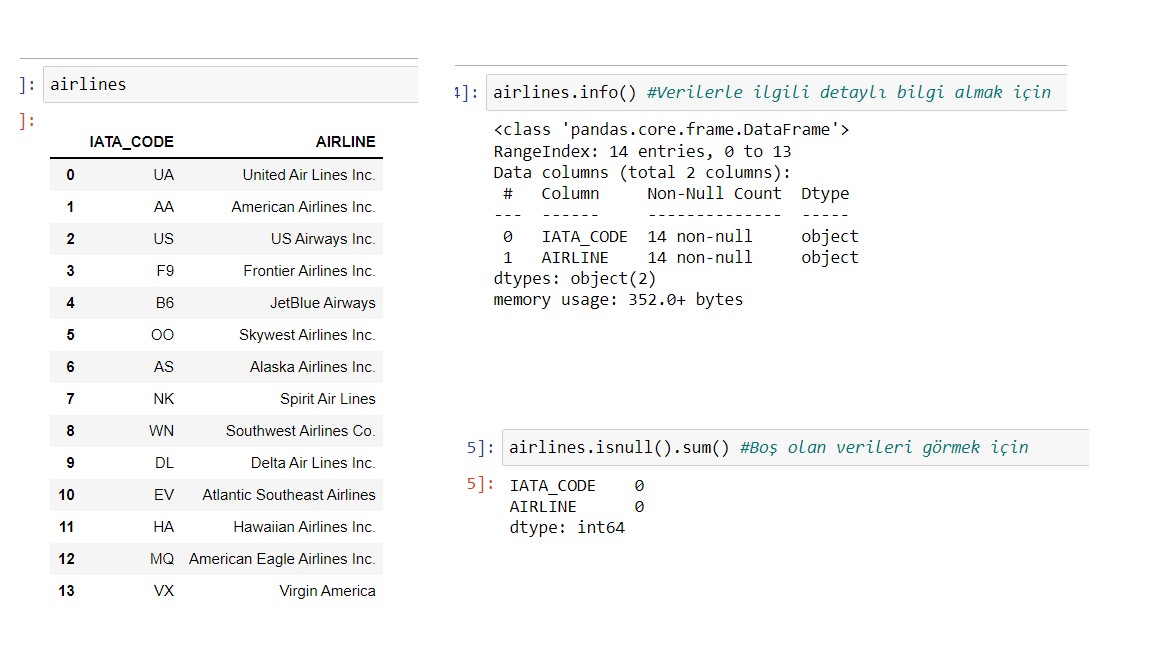
ABD Ulaştırma Bakanlığı'nın (DOT) Ulaştırma İstatistikleri Bürosu, büyük hava taşıyıcıları tarafından gerçekleştirilen iç hat uçuşlarının dakik performansını izleyerek kayıt altına almaktan sorumludur. DOT'un aylık Hava Yolculuğu Tüketici Raporu'na ek olarak sağlanan uçuş rötar ve iptal veri setinde zamanında yetişen, rötar yapan, iptal edilen ve rotası değiştirilen uçuşların sayısına ilişkin özet bilgiler yayınlanmaktadır. Bu çalışma kapsamında 2015 yılına ait veri seti kullanılmıştır. Veri seti Kaggle sitesinden alınmış, temizlenmiş ve analiz edilmiştir.

Veri seti 3 farklı dosyadan oluşmaktadır: havayolu şirketlerine ait bilgilerin bulunduğu csv dosyası, havaalanlarının bilgilerinin bulunduğu csv dosyası ve uçuşlara ilşikin kayıtların tutulduğu csv dosyası.

İlgili veri setleri kaynaktan indirilerek localden çağırılabilir hale getirilmiştir. Veri setinin ön işlemesi ve veri üzerinde gerçekleştirilecek keşifçi veri analizleri için matplolib, seaborn, pandas ve numpy kütüphaneleri kullanılmıştır. Bu kütüphaneler yardımı ile Jupyter Notebook üzerinde veri manipülasyonu ve analizleri gerçekleştirilmiştir.

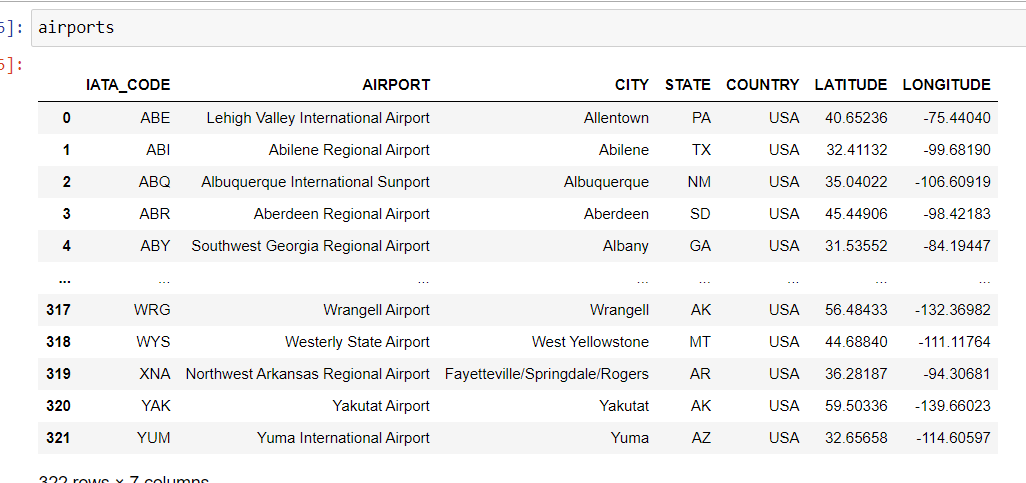
## **Havayolu Şirketlerinin İncelenmesi**

Airlines veri setinde 14 havayolunun adı ve IATA kodlarının yer aldığı görülmektedir. Veri setinde 2 sütun bulunmaktadır, 14 satırın tamamı için her 2 sütunda da boş alan yoktur.



## **Havaalanı Bilgilerinin İncelenmesi**

Veri setinde 322 adet satır veriye ait bilgiler havaalanı adı, IATA kodu, bulunduğu şehir bilgisi, eyaleti, ülkesi, enlem ve boylamı olmak üzere 7 sütunda tutulmaktadır. 3 havaalanına ait enlem ve boylam verilerinin eksik olduğu görülmüştür. Veriler manuel olarak Google maps üzerinden elde edilmeye çalışılmış ancak alınamamıştır. Verilerin eksikliği analizleri etkilemeyeceğinden bu şekilde bırakılarak devam edilmiştir.

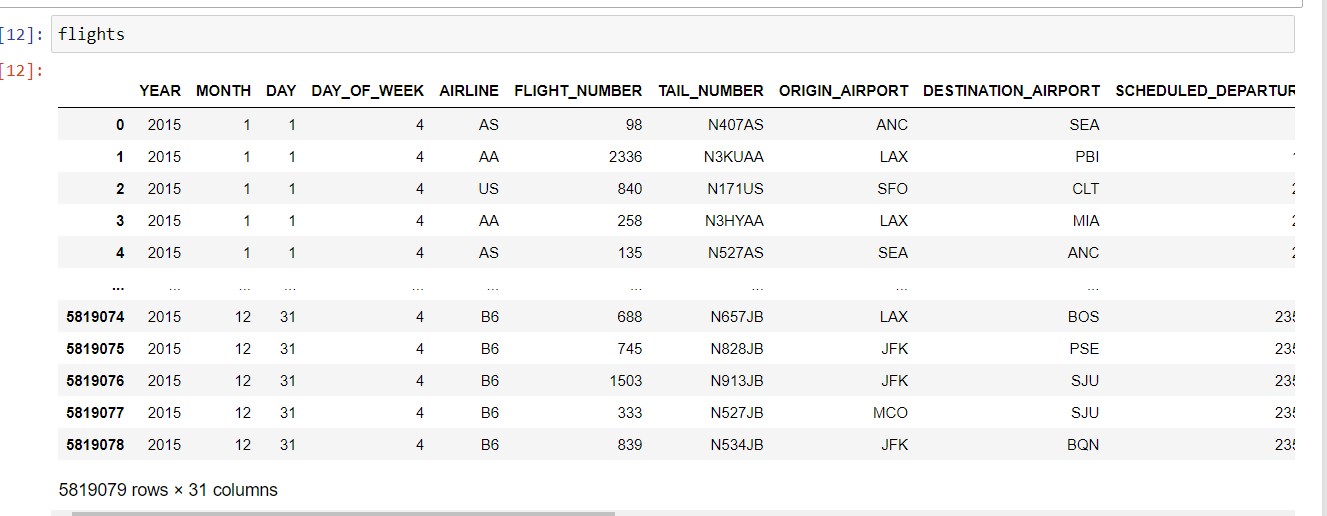




## **Uçuş Verilerinin İncelenmesi**

flights.info() kodu ile veriler hakkında bilgi alınmıştır. Buna göre 5819079 satır veri ve 39 sütun bulunmaktadır. Veri setindeki sütun alanları ise şu şekildedir:

* **YEAR**: Year of the Flight Trip
* **MONTH**: Month of the Flight Trip
* **DAY**: Day of the Flight Trip
* **DAY\_OF\_WEEK**: Day of week of the Flight Trip
* **AIRLINE**: Airline Identifier
* **FLIGHT\_NUMBER**: Flight Identifier
* **TAIL\_NUMBER**: Aircraft Identifier
* **ORIGIN\_AIRPORT**: Starting Airport
* **DESTINATION\_AIRPORT**: Destination Airport
* **SCHEDULED\_DEPARTURE**: Planned Departure Time
* **DEPARTURE\_TIME**: WHEEL\_OFF - TAXI\_OUT
* **DEPARTURE\_DELAY**: Total Delay on Departure
* **TAXI\_OUT**: The time duration elapsed between departure from the origin airport gate and wheels off
* **WHEELS\_OFF**: The time point that the aircraft's wheels leave the ground
* **SCHEDULED\_TIME**: Planned time amount needed for the flight trip
* **ELAPSED\_TIME**: AIR\_TIME + TAXI\_IN + TAXI\_OUT
* **AIR\_TIME**: The time duration between wheels\_off and wheels\_on time
* **DISTANCE**: Distance between two airports
* **WHEELS\_ON**: The time point that the aircraft's wheels touch on the ground
* **TAXI\_IN**: The time duration elapsed between wheels-on and gate arrival at the destination airport
* **SCHEDULED\_ARRIVAL**: Planned arrival time
* **ARRIVAL\_TIME**: WHEELS\_ON + TAXI\_IN
* **ARRIVAL\_DELAY**: ARRIVAL\_TIME - SCHEDULED\_ARRIVAL
* **DIVERTED**: Aircraft landed on airport that out of schedule
* **CANCELLED**: Flight Cancelled (1 = cancelled)
* **CANCELLATION\_REASON**: Reason for Cancellation of flight: A - Airline/Carrier; B - Weather; C - National Air System; D - Security
* **AIR\_SYSTEM\_DELAY**: Delay caused by air system
* **SECURITY\_DELAY**: Delay caused by security
* **AIRLINE\_DELAY**: Delay caused by the airline
* **LATE\_AIRCRAFT\_DELAY**: Delay caused by aircraft
* **WEATHER\_DELAY**: Delay caused by weather

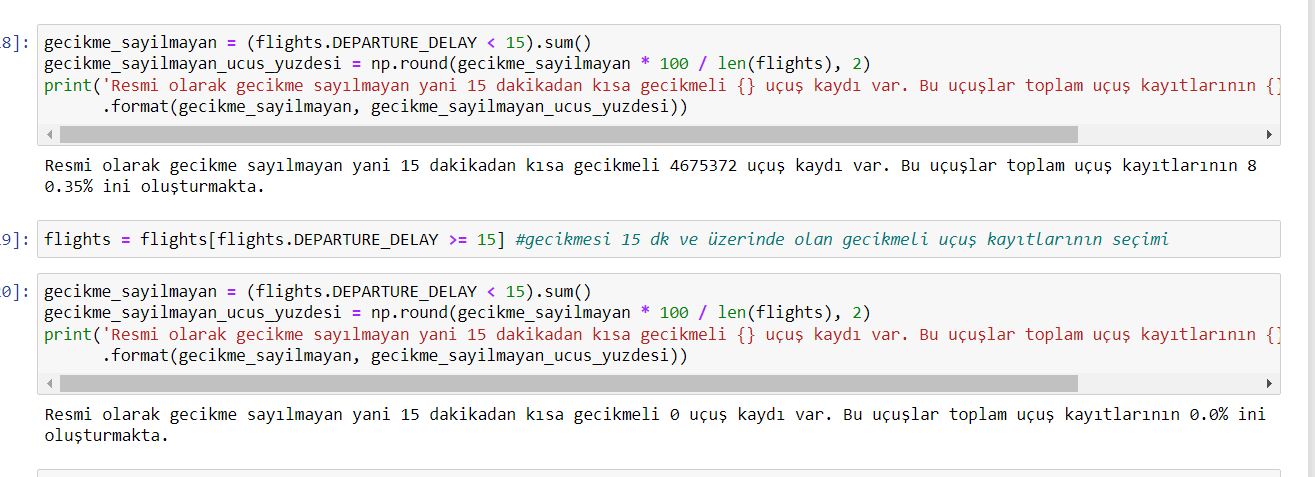


İlk olarak bu verilerin türlerine bakılarak kategorik olması gereken "AIRLINE", "DESTINATION\_AIRPORT", "ORIGIN\_AIRPORT" sütunlarının türleri değiştirilmiştir. Bu düzeltme sonrasında analizlerde kullanılmayacak olan 'TAXI\_OUT', 'TAXI\_IN', 'WHEELS\_ON', 'WHEELS\_OFF', 'YEAR', 'AIR\_SYSTEM\_DELAY', 'SECURITY\_DELAY', 'AIRLINE\_DELAY', 'LATE\_AIRCRAFT\_DELAY', 'WEATHER\_DELAY', 'DIVERTED', 'CANCELLED', 'CANCELLATION\_REASON', 'FLIGHT\_NUMBER', 'TAIL\_NUMBER', 'AIR\_TIME' sütunları silinmiştir.

Daha sonra flights.isnull().sum() sütunuyla boş olan veriler incelenmiştir. Boş olan verilerin yüksek olduğu sütunlarda anlamlı tamamlamalar yapılamayacağından verilerinin %25’inden fazlası boş olan sütunlar silinmiştir.

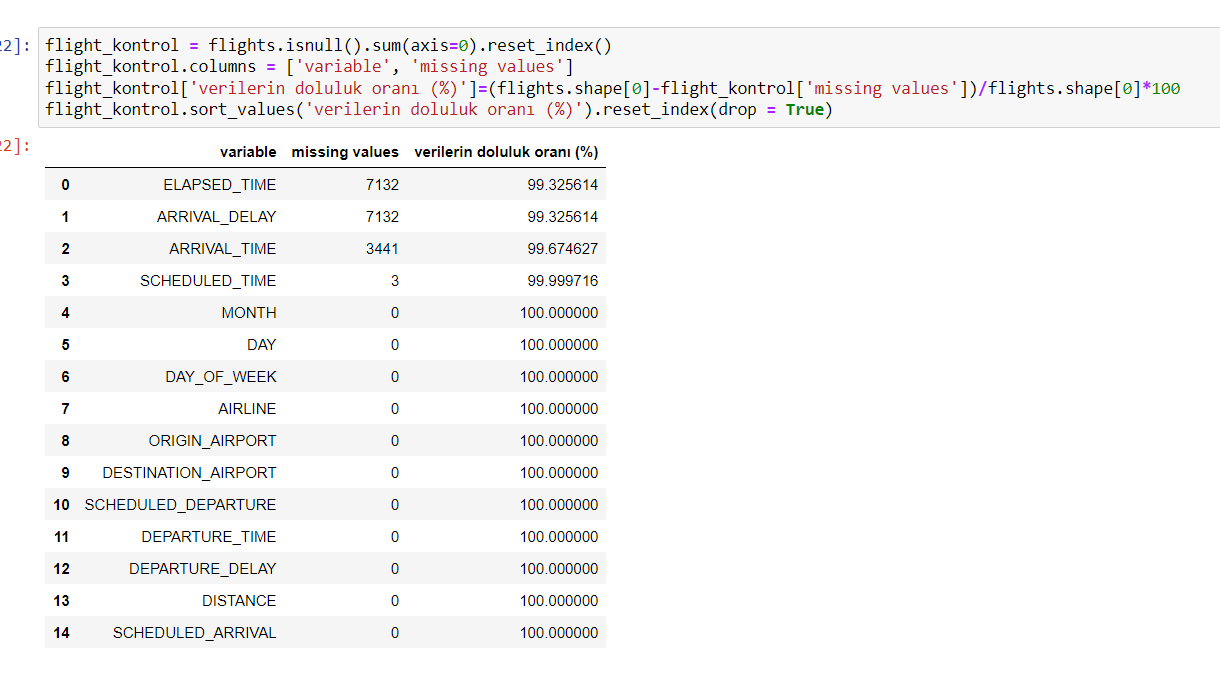
### ***Gecikme Kabul Edilmeyen Verilerin Temizlenmesi***

Uçuş gecikmesi, bir havayolu uçuşunun planlanan süresinden daha geç kalkması ve / veya inmesidir. Federal Havacılık İdaresi (FAA), bir uçuşun planlanan süresinden itibaren en az 15 dakika sapma olması durumunda uçuşun ertelenmiş yani gecikmiş sayılacağını kabul etmektedir. Bu uygulamaya göre 15 dakikanın altındaki sapmalar resmi olarak gecikme sayılmadığından bunlar temizlenmiştir. Resmi olarak gecikme sayılmayan yani 15 dakikadan kısa gecikmeli 4675372 uçuş kaydı var. Bu uçuşlar toplam uçuş kayıtlarının 80.35% ini oluşturmaktadır. Bu veriler silinmiştir.



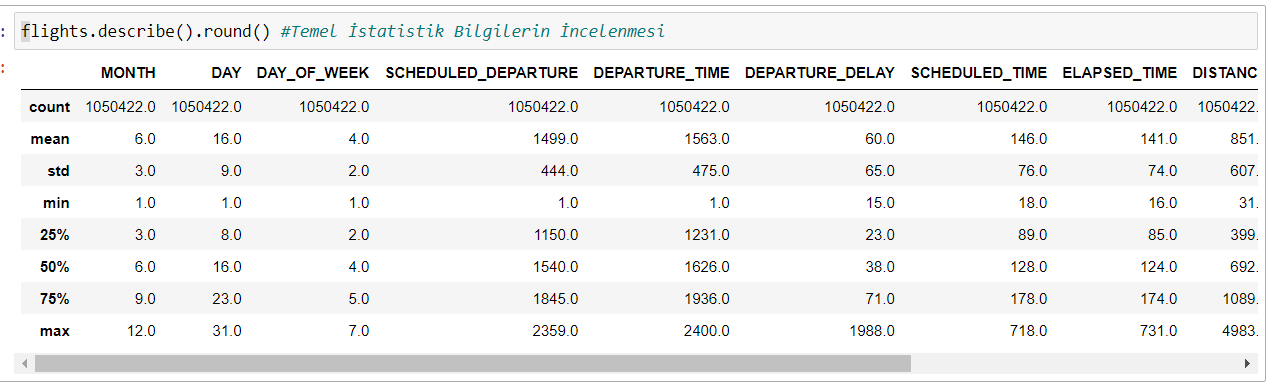
### ***Boş Verilerin Temizlenmesi***

flights.isnull().sum() ile bakıldığında veri setinde hala boş değere sahip sütunların olduğu görülmekte. Bu verilerin silinmesine karar vermeden önce veri setinin doluluğunun incelenmesi gerekiyor.

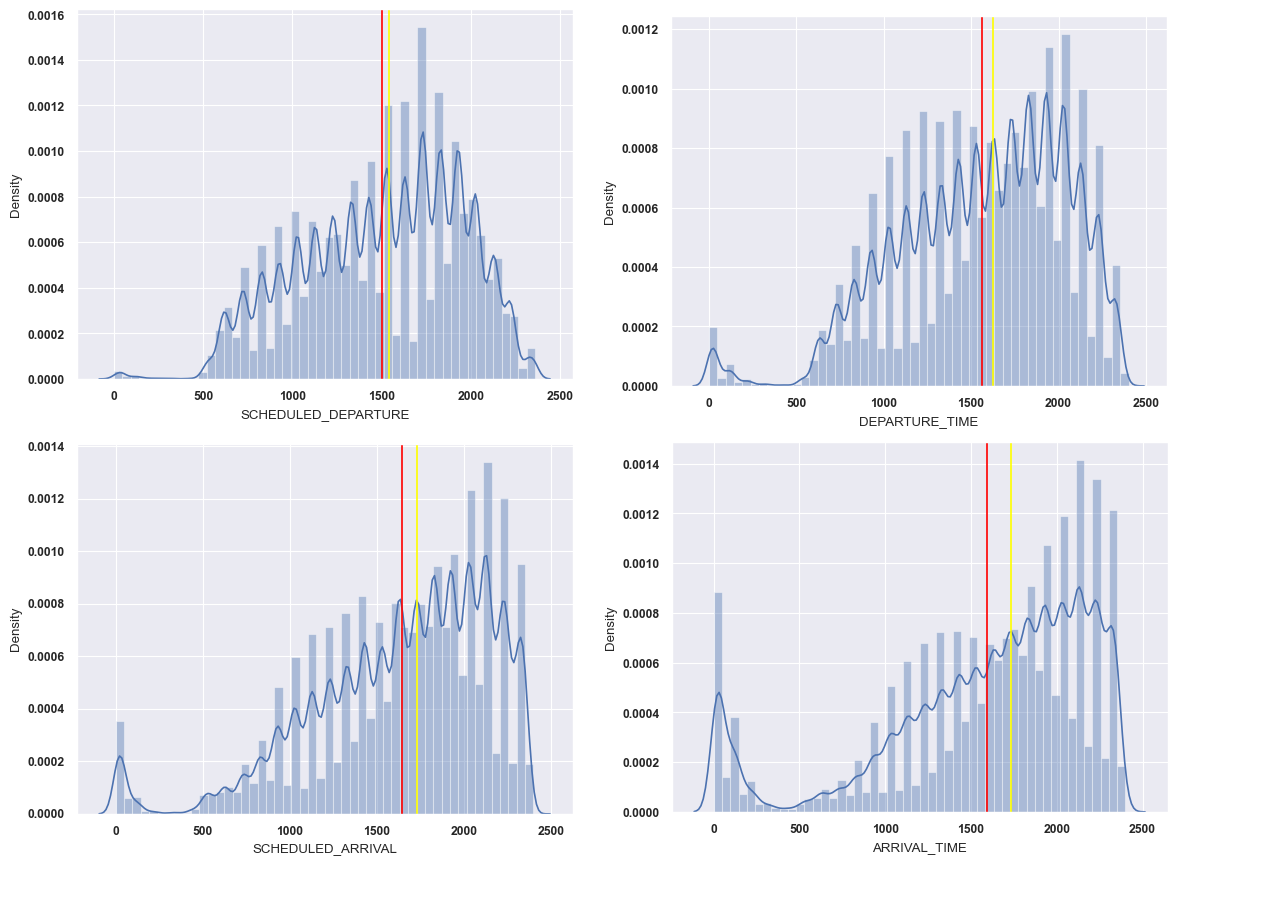


Değişken bazında verilerin doluluk oranlarının oldukça iyi (> %99) olduğunu görüyoruz. Bu çalışma kapsamında eksik verilerin doldurulma ihtimali olmadığından, %99 doluluk oranı yeterli kabul edilecek ve eksik değerleri içeren kayıtlar flights.dropna(inplace = True) kodu ile silinmiştir.

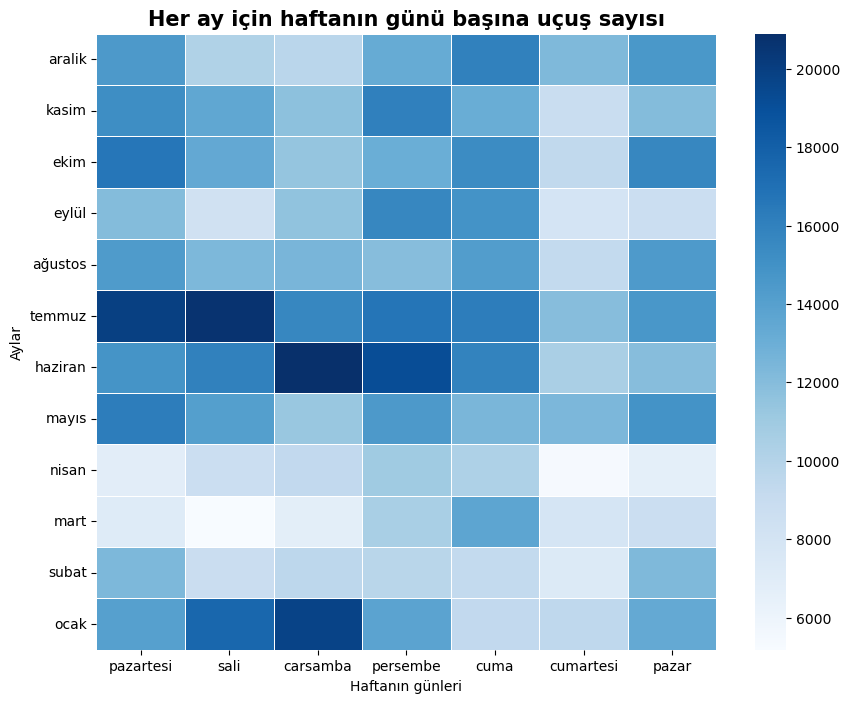
### ***Temel İstatistik Bilgilerin İncelenmesi***



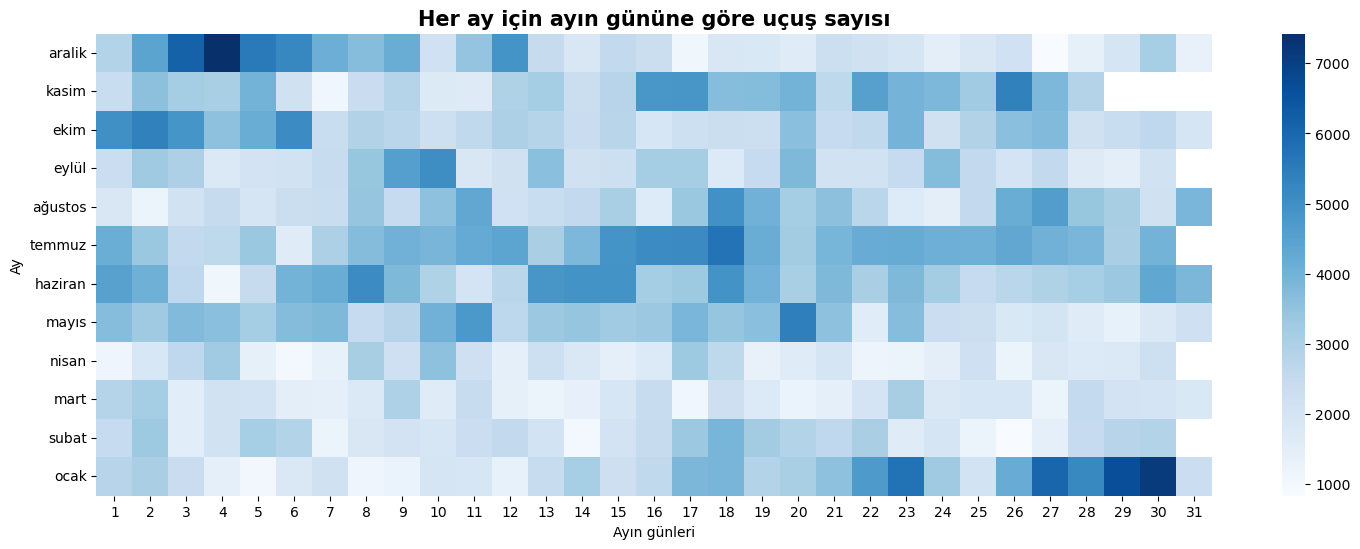
Sayısal sütunlar seçilerek histogramları çıkarılmış ve dağılımları incelenmiştir. Ağapıda bazı değişkenlerin örnek dağılımları verilmiştir. Verilerde istatistiksel olarak sorun görülmediğinden analizlere devam edilmiştir. Dağılımların tümü dijital ortamda teslim edilen Jupyter Notebook dosyasında bulunmaktadır.



Uçuşların aylar ve günler üzerine dağılımlarının genel görünümününe bakılmıştır. Bu sayede belirli dönemler arasında analizleri etkileyece aşırı farkların olup olmadığı da gözlemlenecektir.



Bu grafik, yıl boyunca hafta içi uçuş sayısının genel olarak sabit olduğunu, Cumartesi ve Pazar günleri ise azaldığını göstermektedir. Her ayın kendi yopun dönemleri olduğu fark edilebilir, örneğin, Ocak ayında Cuma günleri ve hafta sonları daha fazla uçuş varken, Mart ayında Pazartesi günleridir. Ayrıca ısı haritası, verilerden gelen ilgili bilgileri aramak için daha iyi bir veri görselleştirmesi sunabilir. Harita, hafta içi günler ve hafta sonları arasındaki farkı daha iyi gösterir, Noel, Yaz veya bahar tatili gibi tatillerin olduğu aylarda bazı hafta sonları artan değerler gösterse bile, en yüksek uçuş sayısı her zaman hafta içinde, minimum uçuşlar ise hafta sonlarındadır.

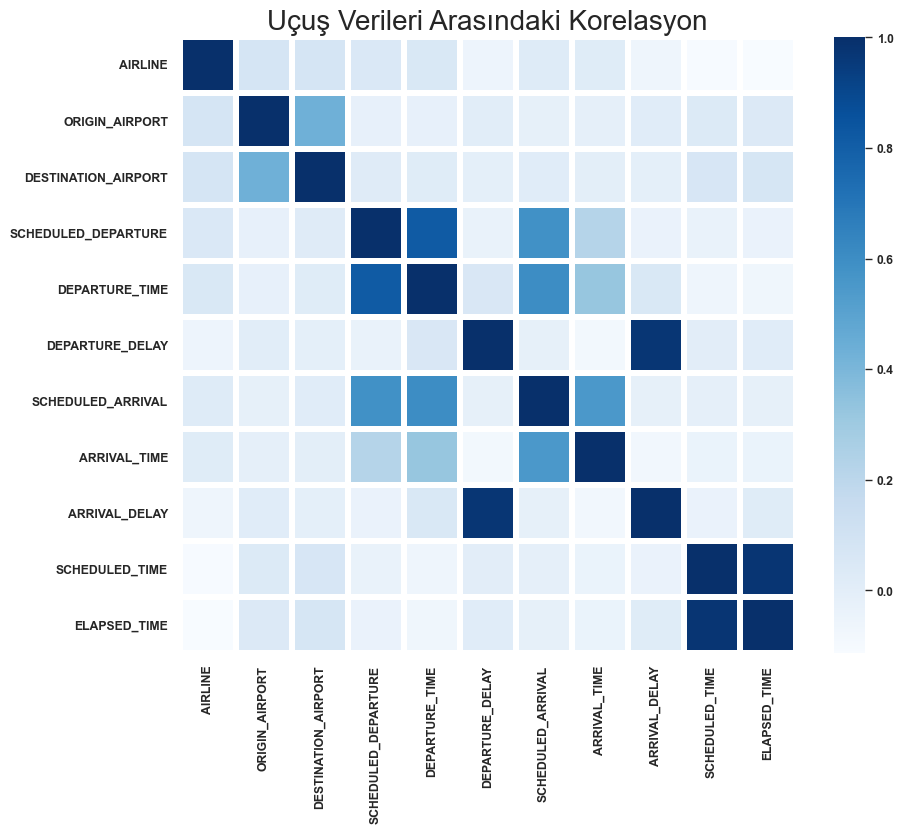


Ayların son günlerinde uçuş sayısı azalmaktadır, bunun nedeni tüm ayların aynı gün sayısına sahip olmamasıdır (bir sonraki grafikte beyaz kutulardan da anlaşılacağı gibi). Benzer şekilde Şubat ayının diğer aylara göre daha az uçuş sayısına sahip olduğunu da görebiliriz.

Genel olarak dönemsel farklılıklar çok fazla olmadığından herhangi bir ek işleme gerek duyulmamaktadır.

### ***Uçuş Verileri Arasındaki Korelasyonun İncelenmesi***

Aşağıdaki korelasyon ısı haritasına bakıldığı zaman değişkenlerin büyük çoğunluğu arasında 0’a yakın bir korelasyon olduğu yani aralarında anlamlı bir ilişki olmadığı görülmektedir. Bu ilişkiler gecikmeler üzerinden sonraki bölümlerde detaylı olarak incelenecektir.



# **Yöntem**

Keşifçi veri analizi yöntemleriyle veri ön işleme aşamasında veri seti temizlenmiş, organize edilmiş ve temel değişkenler incelenerek temel çıkarımlar yapılmıştır.

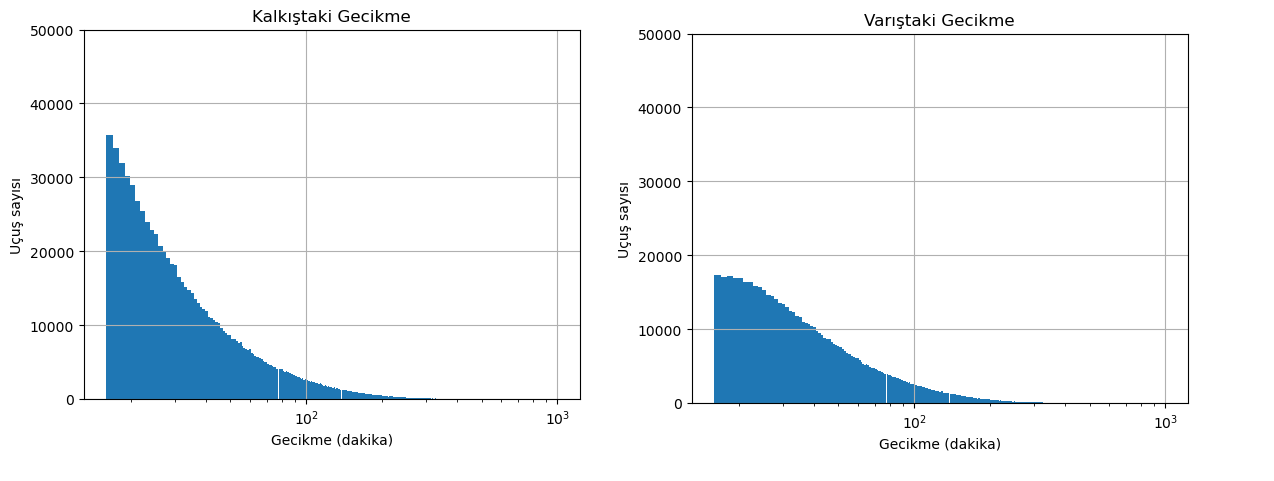
Yöntem bölümü 2 parçadan oluşmaktadır. İlk aşamada uçuşlarda yaşanan gecikmelerin hangi değişkenlere ve durumlara bağlı olarak gerçekleştiği incelenecektir. Bu incelemelerle, ikinci aşama olan makine öğrenmesi ile uçuş gecikmesi tahminlemesi yapılabilmesi için gereken özelliklerin belirlenmesi (feature selection) hedeflenmektedir. Feature selection adımından sonra 'Lasso', 'Linear Regression', 'Ridge', 'Random forest Regressor', 'Decision Tree Regressor', 'Boosted Linear', 'Boosted Lasso', 'Boosted Ridge', 'Bagged Linear', 'Bagged Lasso', 'Bagged Ridge' yöntemleri ile tahminleme yapılacaktır. Bu yöntemlerden ilgili başlık altında kısaca bahsedilecektir.

## **Uçuş Gecikmesini Etkileyen Faktörlerin Belirlenmesi**

Bu bölümde keşifçi veri analizi yöntemleri ve görselleştirmeler kullanılarak veri ön işleme aşamasında gerçekleştirilen temel çıkarımlar derinleştirilecek ve detaylandırılacaktır.

### ***Uçuşların Genel Gecikme Durumu***

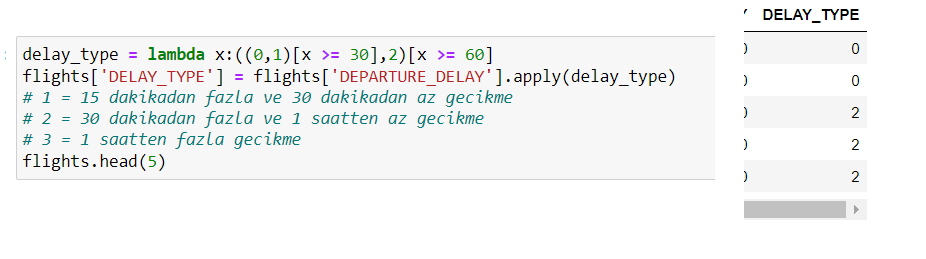
Uçuşların kalkış ve varışlarındaki gecikmeler genel kapsamlı incelenmiştir. Detaylı analizler, bu adımın bulguları kullanılarak sonraki aşamalarda gerçekleştirilecektir.

Gecikmelerin yoğunluklu olarak 15 dakika ve 1 saat arasında olduğu görülmektedir. Bunun tam dağılımı yapılarak gecikme türü atanmalıdır. Bu atama sonraki adımlarda ve özellikle tahminleme aşamasında kullanışlı olacaktır.

# 1 = 15 dakikadan fazla ve 30 dakikadan az gecikme

# 2 = 30 dakikadan fazla ve 1 saatten az gecikme

# 3 = 1 saatten fazla gecikme



#### *Gecikmeler kalkışta mı varışta mı gerçekleşiyor?*

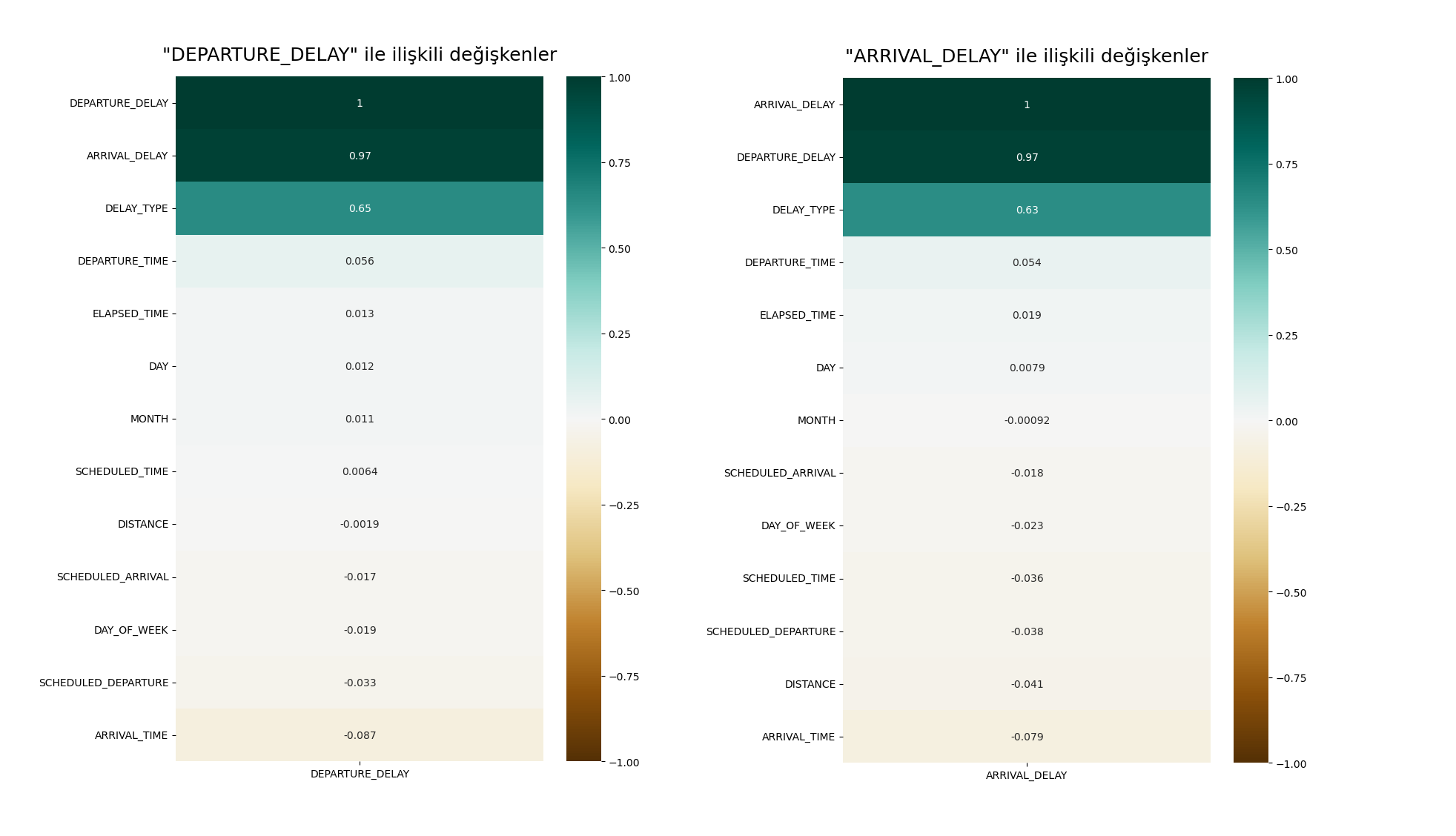
Bu grafikte varıştaki gecikmelerin kalkıştaki gecikmeden genellikle daha kısa olduğu görülmektedir. Bunun sebebi havayolu şirketlerinin gecikmeyi azaltmak için uçuş hızını artırmak gibi düzenlemeler yapmasından kaynaklı olabilir. Alınabilecek aksiyonlar havayolu şirketinin büyüklüğüne ve duruma göre değişebileceğinden gecikme tahminlemeleri sadece kalkıştaki gecikme üzerinden yapılacaktır.

Keşifçi veri analizlerinde ise veri setinin ve durumun tam olarak anlaşılabilmesi için hem kalkış hem de varış gecikmeleri incelenecektir.



#### *Gecikmelerle ilişkili değişkenlerin belirlenmesi*

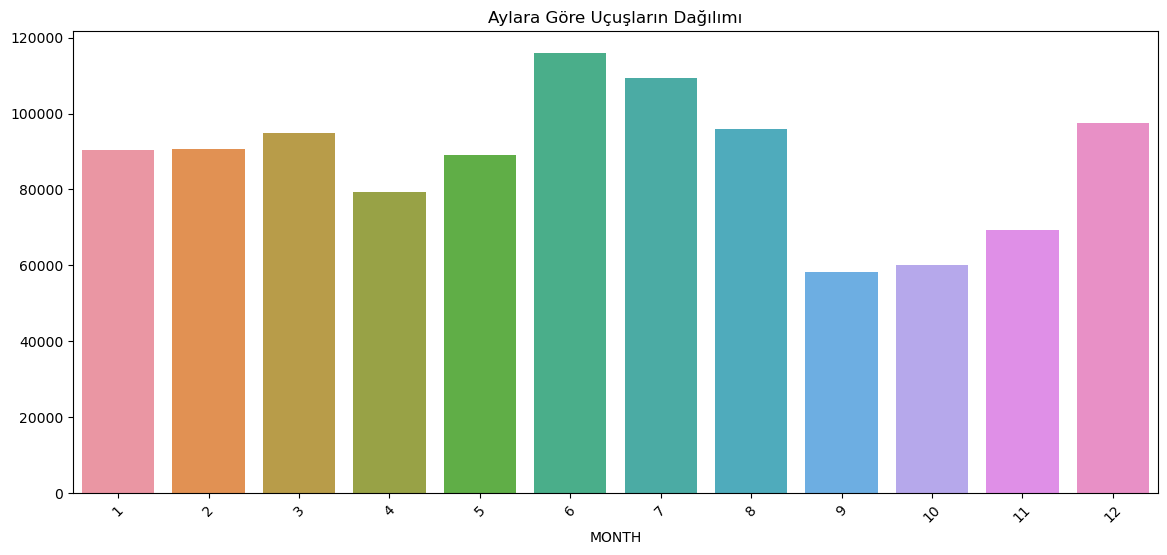
Yapılan genel korelasyon analizinde departure delay ve arrival delay arasında güçlü bağlantılar görülmüş ve diğer değişkenlerin gecikmelerle bağlantısı zayıf olarak ortaya çıkmıştır. Özellikle bu gecikme değişkenleri ile bağıntılı olan faktörler incelenecektir.

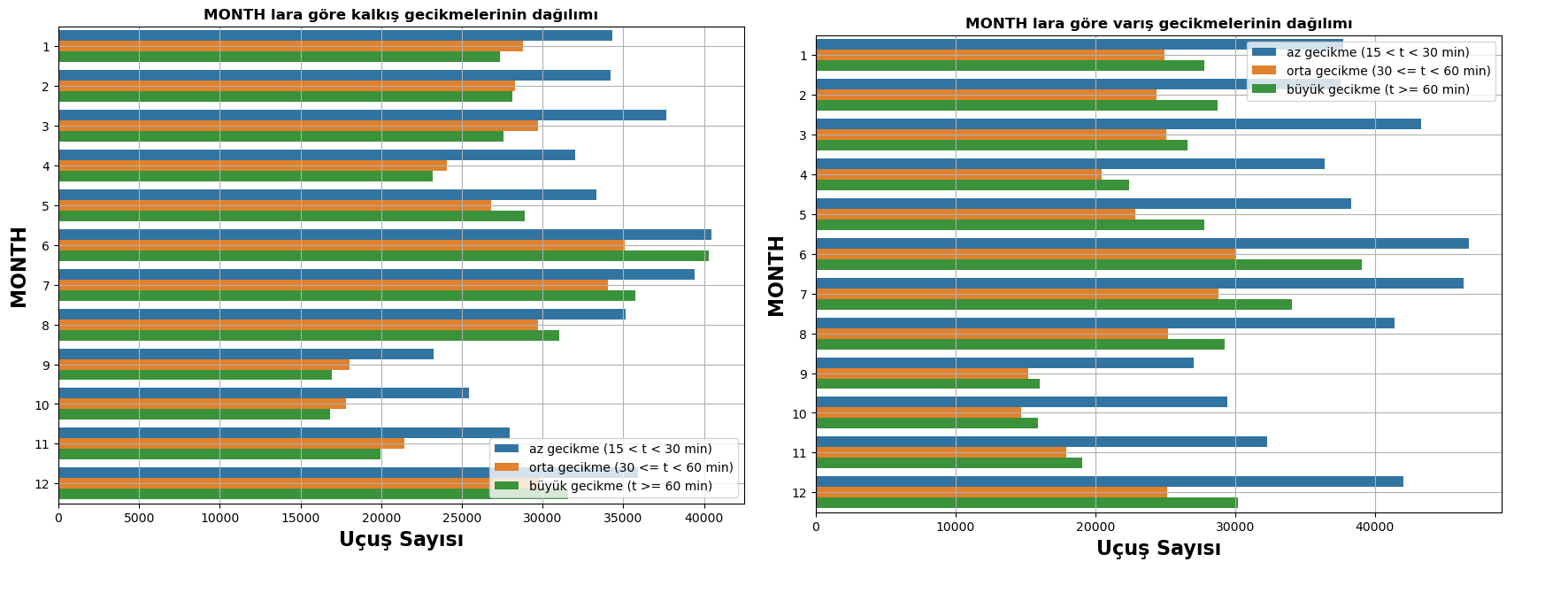


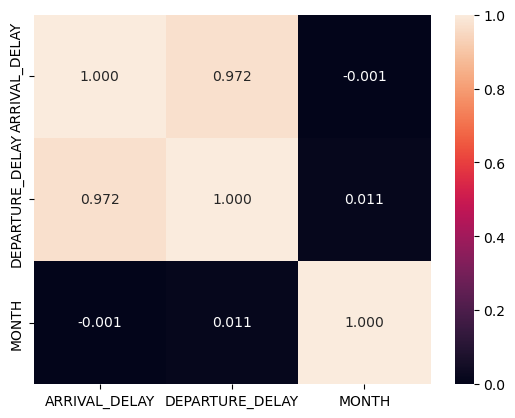
Verilen 2 korelasyon grafiği gecikmelerle ilişkili değişkenleri sayısal olarak göstermiş olsa da, veri setinin ve durumun tam olarak anlaşılması için seçilen bir kaç değişken üzerinden detaylı analizler yapılacaktır. Sonrasında ise havayolu şirketi, havaalanı gibi kategorik değerler üzerinden incelemeler yapılacaktır. Bu incelemelerin yapılabilmesi için gerekli fonksiyonlar yazılarak, aynı kod bloklarının tekrar tekrar yazılmasının önüne geçilmesi sağlanmıştır.

### ***Sayısal Değişkenlere Göre Uçuşların ve Gecikmelerin İncelenmesi***

Oluşturulan fonksiyonlar yardımıyla seçilen ay değişkeni için uçuş dağılımı, kalkış ve varış gecikmelerinin dağılımı görselleştirilmiş ve incelenmiştir. Sonrasında varış ve kalkış gecikmeleri ile aylar arasındaki korelsyon incelenmiştir.





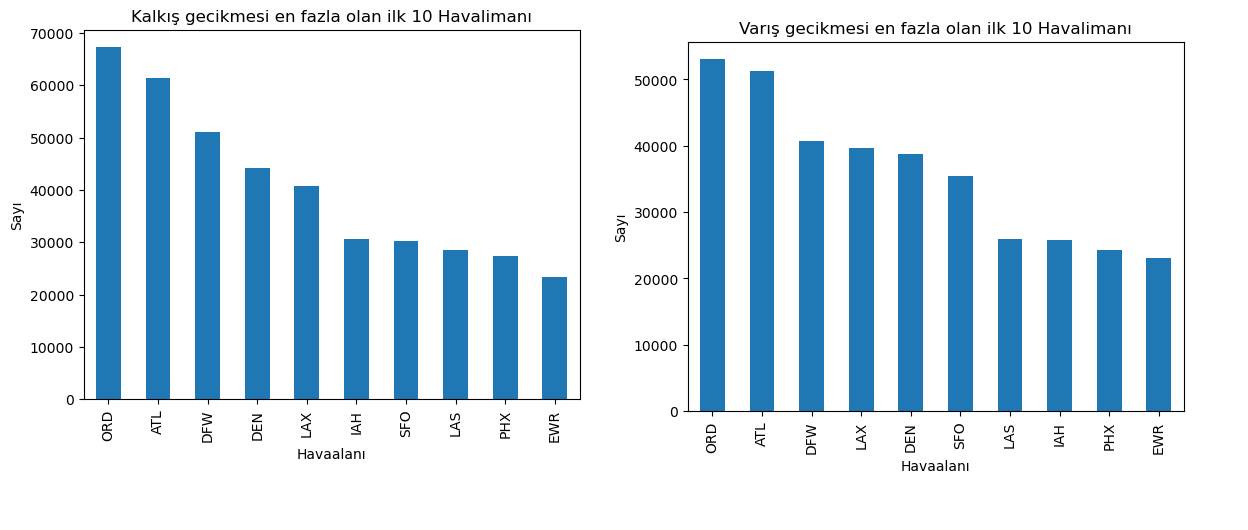


Oluşturulan uçuş dağılımlarına bakıldığında yaz döneminde uçuş sayılarında bir artış olsa da genel olarak aylar ve mevsimler arasında anlamlı bir fark olmadığı görülmektedir. Buna paralel olarak kalkış ve varış gecikmelerinin dağılımlarında da aylar ve mevsimler arasında belirli farklılıklar olsa da dikkate alınması gereken büyüklükte farklar görünmemektedir, bu durum uçuş sayısıyla ilişkilendirilebilir. Son olarak kalkış ve varış gecikmeleri ile aylar arasındaki korelasyona bakıldığında değişkenler arasında bir ilişki bulunmadığı görülmektedir. Yani oluşan gecikmelerin aylarla hernagi bir ilişkisi olmadığı söylenebilir.

Ay değişkeni sonrasında haftanın günleri ve ayın günleri değişkenleri için de aynı dağılımlar yapılmıştır. Her ayın 31. günü olmadığı düşünülerek ayın günlerine göre olan dağılım değerlendirildiğinde ayın günleri arasındaki uçuş dağılımlarında dikkate alınması gereken farklılıklar gözlenmemektedir. Kalkış ve varış gecikmelerinin dağılımlarında da ayın günlerine göre bir farklılık bulunmamaktadır. Bu duruma uygun şekilde haftanın günlerine göre uçuş sayısı dağılımları ve gecikme dağılımları da uyumludur. Ayın ve haftanın günleri ve kalkış ve varış gecikmeleri arasındaki ilişki korelasyon grafiğiyle incelendiğinde; aylara göre dağılımda olduğu gibi bu değişkenler arasında herhangi bir bağlantı bulunmadığı gözlenmiştir. Benzer şekilde diğer sayısal değişkenlerden biri olan mesafe değişkeni üzerinden de aynı dağılımlar incelenerek benzer gözlemler elde edilmiştir. Bu değişkenlere ait görsel dağılımlara, anlamlı bir katkı sağlamamaları sebebiyle raporda yer verilmemiş olup iletilen Jupyter Notebook dosyasında bu görseller ve dağılımlar mevcuttur.

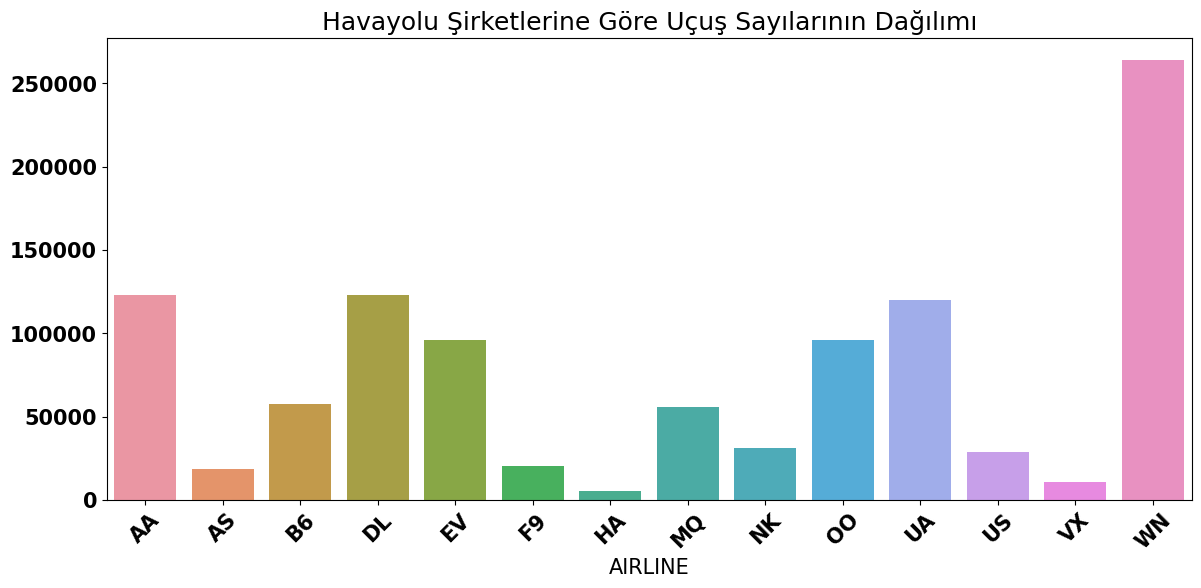
### ***Kategorik Değişkenlere Göre Uçuşların ve Gecikmelerin İncelenmesi***

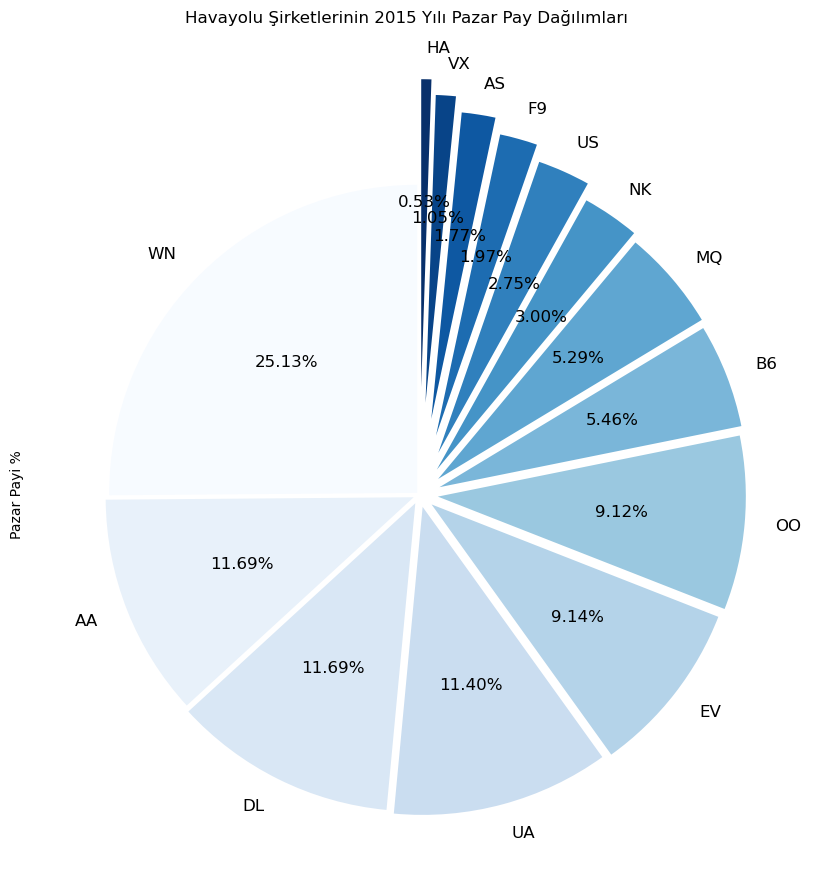
#### *Havaalanlarındaki Uçuş Yoğunluğunun Dağılımının İncelenmesi*



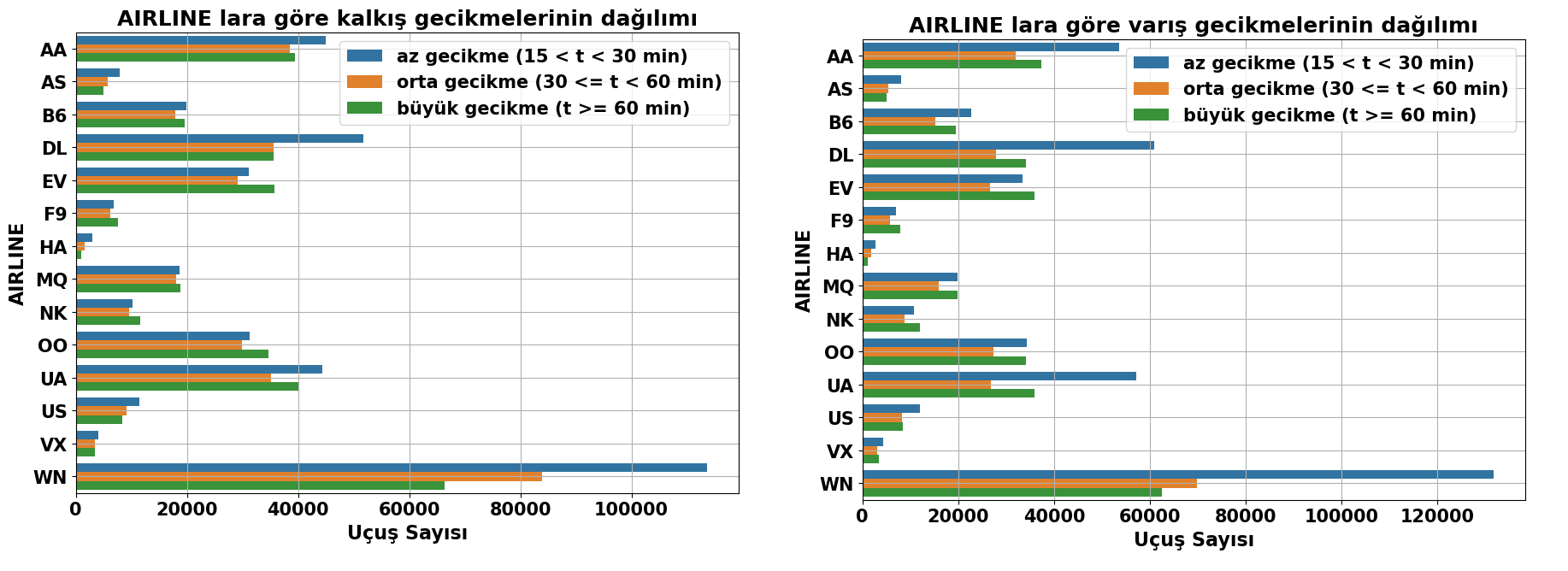
#### *Havayolu Şirketlerine Göre Uçuş Sayılarının ve Gecikmelerin İncelenmesi*

Her havayolu şirketi aynı büyüklükte bir filoya ve aynı sayıda/çeşitlilikte uçuş hattına sahip değildir. Doğru değerlendirmenin yapılabilmesi için bu şirketlerin pazar paylarının ve havayolu bazlı gecikmelerinin incelenmesi gereklidir.

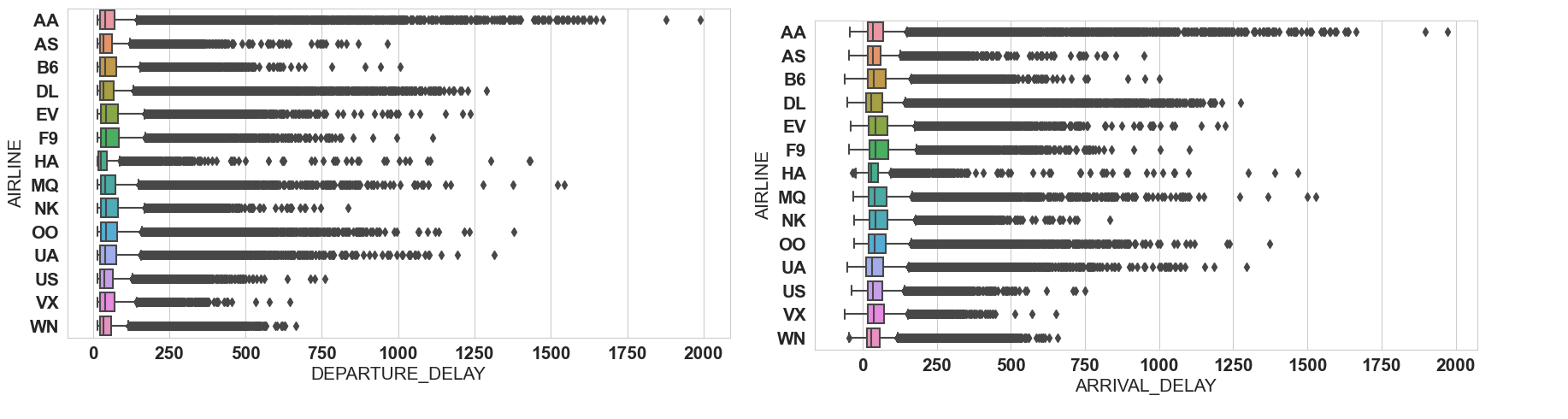




Havayolu başına uçuş yüzdesini veren ilk pasta grafiğine baktığımızda, taşıyıcılar arasında bir miktar eşitsizlik olduğunu görüyoruz. Örneğin Southwest Airlines, uçuşların yaklaşık %20'sini gerçekleştiriyor ve bu da neredeyse en küçük 7 havayolunun düzenlediği uçuş sayısına eşit. Sonrasında American Airlines, Delta Air Lines ve United Airlines çok yakın pazar paylarına sahip olarak sıralanmakta. Son sıralarda ise daha yerel olan Hawaian Airlines,Virgin America gibi bölgesel firmalar gelmekte.



Bu grafikte ise, 30 dakikadan az gecikmelerin, 30 < t < 60 dakika aralığındaki gecikmelerin ve son olarak 60 dakikadan büyük gecikmelerin sayıları verilmiştir. Gemel olarak havayolundan bağımsız olarak 60 dakikayı aşan rötarların sadece birkaç yüzdelik paya sahip olduğunu görüyoruz. Bununla birlikte, bu üç gruptaki gecikmelerin oranı havayoluna bağlıdır: örnek olarak, SkyWest Havayolları örneğinde, 60 dakikadan uzun gecikmeler, 30 < t < 60 aralığındaki gecikmelere göre yalnızca ∼ 15 daha düşüktür. dk. SoutWest Airlines için işler daha iyi çünkü 60 dakikadan uzun gecikmeler 15 < t < 30 dakika aralığındaki gecikmelerden 4 kat daha az sıklıkta yaşanıyor. Veya UA (United Air Lines Inc.) havayolunun uçuşlarının yaklaşık %50'sinde büyük gecikme varken, DL (Delta Air Lines Inc.) havayolunun uçuşlarının ~%25'inde büyük gecikme var.



Son derece düşük ortalama gecikmeler bildiren Hawaiian Airlines ve Alaska Airlines hariç, ∼ 60 ± 5 dakikalık bir değerin tüm ortalama gecikmeleri doğru bir şekilde temsil ettiğini elde ettik. Bu değerin oldukça düşük olduğunu unutmayın, bu da her havayolu için standardın programa uymak olduğu anlamına gelir.

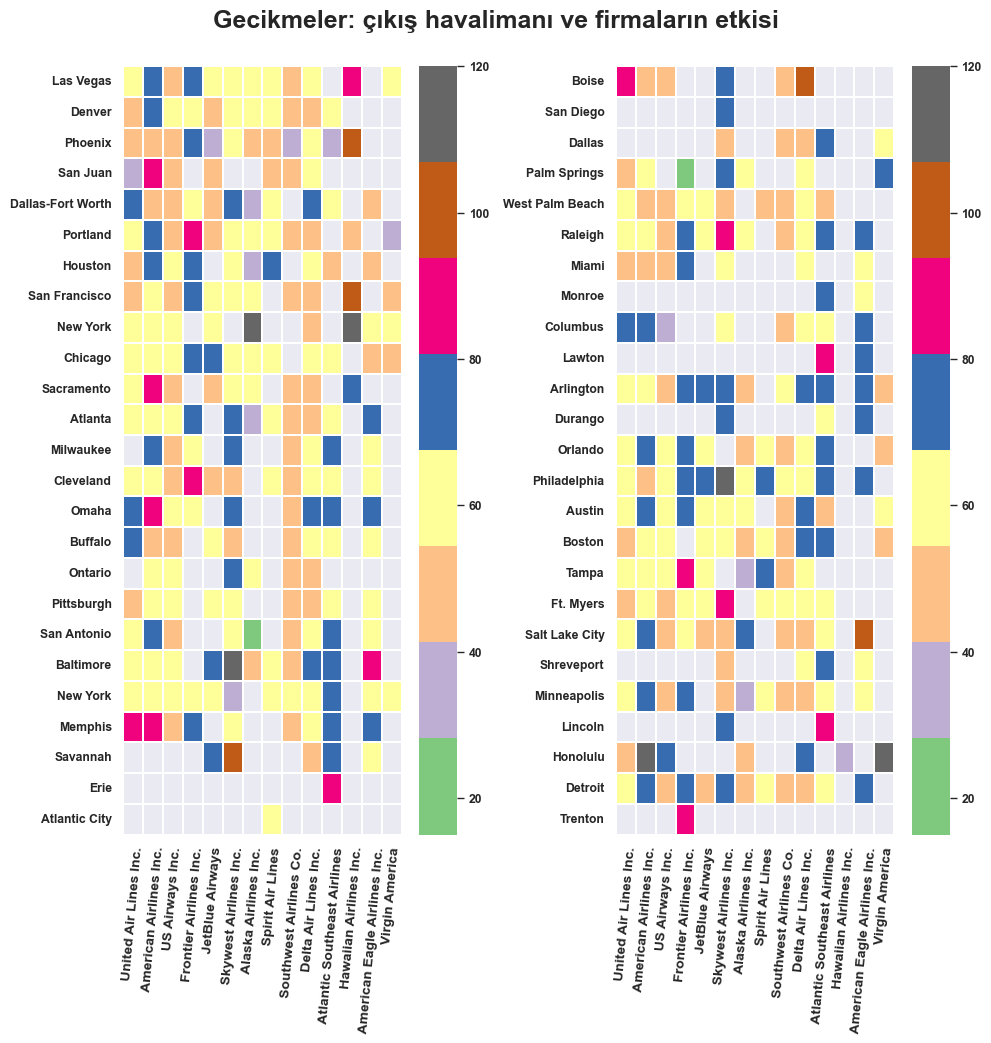
Son olarak, alttaki şekil, Ocak 2015'te ölçülen tüm gecikmelerin bir sayımını göstermektedir. Bu gösterim, pasta grafikte görünen göreli homojenliği perspektife yerleştirmektedir. Nitekim tüm ortalama rötarların 50 dakika civarında olmasına rağmen, bu değerin uçuşların büyük çoğunluğunun zamanında kalkmamasının bir sonucu olduğunu görüyoruz.

#### *Havaalanları ve Havayolu Şirketleri Arasındaki Gecikme Korelasyonu*

Havaalanlarının sayısı oldukça fazla olduğundan, tüm bilgileri bir kerede gösteren bir grafik, yaklaşık 4400 değeri temsil edeceğinden (yani 312 hava alanı × 14 hava yolu) zor olacaktır. Bu nedenle, verilerin bir alt kümesi temsili olarak raporda gösterilecektir ancak tüm havayolu dağılımları Jupyter Notebook dosyasında görülebilir.

Bu grafik bazı temel sonuçlar çıkarmamıza yardımcı olacaktır. İlk olarak, farklı havayolları ile ilişkili verilere bakarak, daha önce gözlemlediğimiz çıkarımların bir tekrarını buluyoruz: örneğin, sağ paneli ele alırsak, American Eagle Airlines ile ilişkili sütunun çoğunlukla büyük gecikmeler içerdiği görülecektir. Delta Airlines ile ilişkilendirilen sütunda, esas olarak 30 dakikadan kısa gecikmeler daha yoğun görülmektedir. Şimdi menşe havaalanlarına bakarsak Denver, Chicago veya New York gibi bazı havaalanlarında geç kalkışların yaygın olduğu görülmektedir. Portland veya Oakland gibi diğer havaalanlarında ise çoğunlukla zamanında kalkışlar bulunmaktadır.

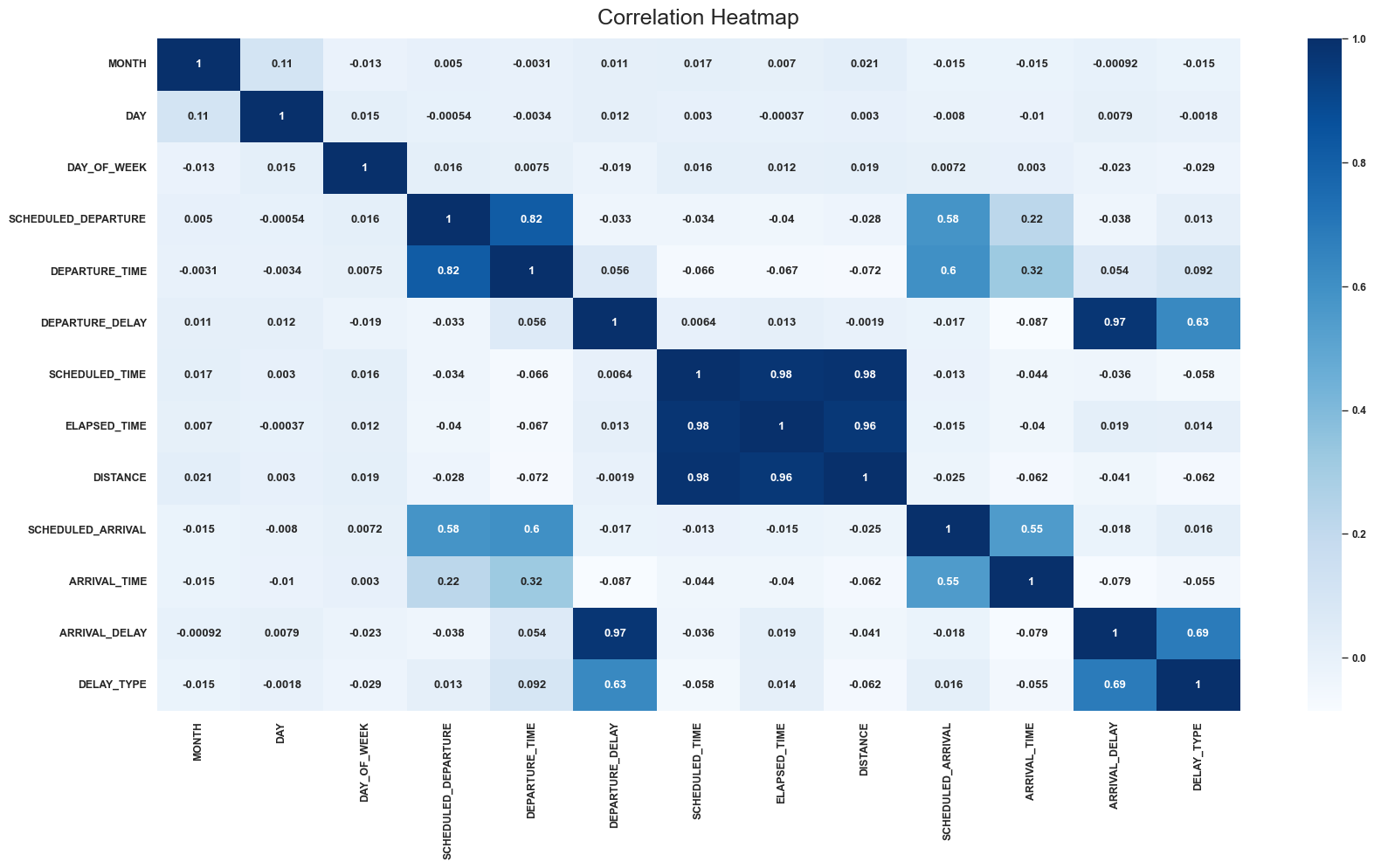
Son olarak, bu gözlemlerden hem farklı havaalanları arasında hem de farklı havayolları arasında ortalama gecikmelerde yüksek bir değişkenlik olduğu sonucuna varabiliriz. Bu çıkarım gecikmeleri doğru bir şekilde modellemek için şirkete ve kalkış yapılan havaalanına özgü bir modelin benimsenmesi gerekeceğini ima ettiğinden önemli bir noktadır.



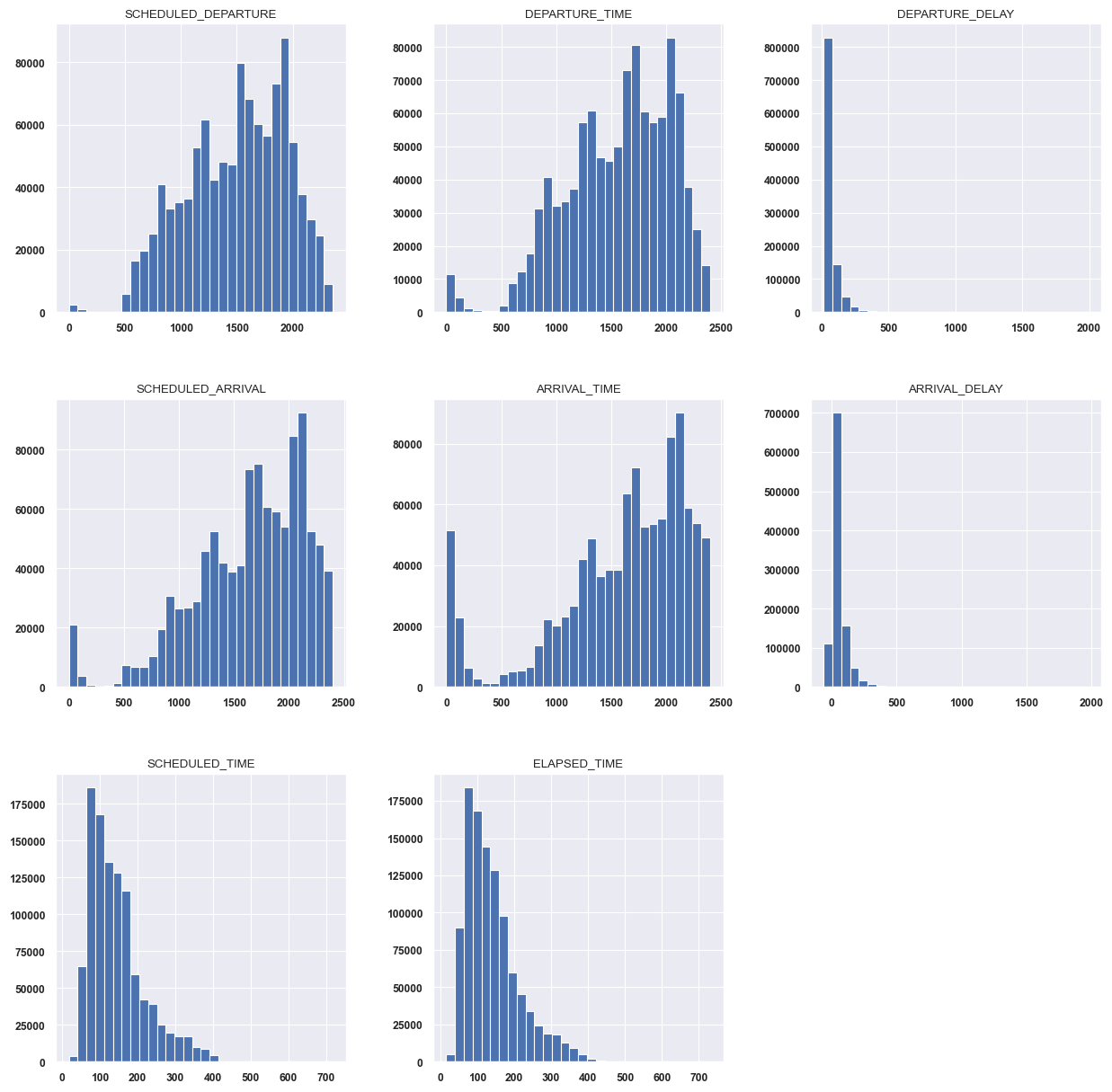
## **Makine Öğrenmesi Metodlarıyla Tahminleme**

### ***Modelleme İçin Feature Selection***

Yukarıda yapılan tüm analizlerle beraber havaalanları, havayolu şirketleri, kalkış ve varış gecikmeleri gibi dedğişkenlerin tahminleme için önem arzettiğini gözlemlemiş olduk. Bu aşamada makine öğrenmesi modelleri kullanmadan önce seçtiğimiz değişkenleri detaylı olarak inceleyip düzenleyeceğiz. Bunun için ilk olarak kalan verilerimizle bir korelasyon grafiği daha çıkarıyoruz.



Buradaki korelasyon grafiğindeki bağlantıları inceleyerek modellemede kullanılmayacak olan sütunları siliyoruz ve kalan değişkenlerin dağılımlarını kontrol etmek için histogram grafiklerini çıkartıyoruz.

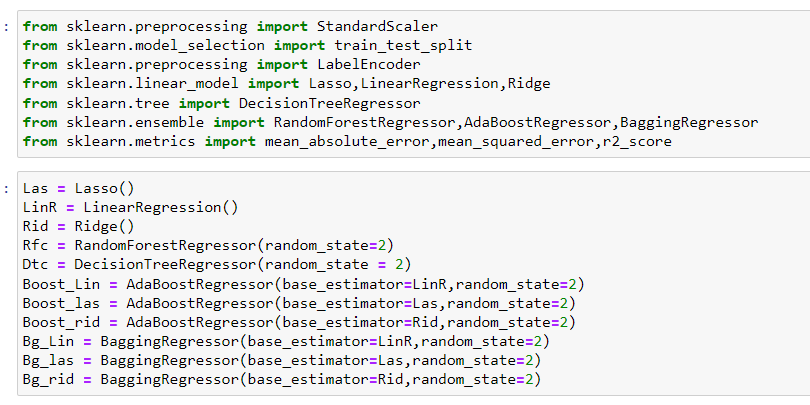


### ***Ön Hazırlık***

Scikit kütüphanesinin kullanıcı kılavuzlarından seçilen makine öğrenmesi modelleri ile ilgili detaylı bilgiler elde edilmiştir (Scikit, n.d.):

* Lasso (en az mutlak büzülme ve seçim operatörü; ayrıca Lasso veya LASSO), ortaya çıkan istatistiksel modelin tahmin doğruluğunu ve yorumlanabilirliğini geliştirmek için hem değişken seçimi hem de düzenlileştirme gerçekleştiren bir regresyon analizi yöntemidir.
* Doğrusal regresyon, skaler bir yanıt ile bir veya daha fazla açıklayıcı değişken (bağımlı ve bağımsız değişkenler olarak da bilinir) arasındaki ilişkiyi modellemek için doğrusal bir yaklaşımdır.
* Ridge regresyonu, çoklu bağlantıdan muzdarip herhangi bir veriyi analiz etmek için kullanılan bir model ayarlama yöntemidir. Bu yöntem, L2 düzenlemesini gerçekleştirir. Çoklu bağlantı sorunu ortaya çıktığında, en küçük kareler yansızdır ve varyanslar büyüktür, bu, tahmin edilen değerlerin gerçek değerlerden çok uzakta olmasına neden olur.
* Karar Ağacı, akış şeması benzeri bir ağaç yapısı kullanan veya sonuçlar, girdi maliyetleri ve fayda da dahil olmak üzere tüm olası sonuçların ve kararların bir modeli olan bir karar verme aracıdır.
* AdaBoost regresörü, orijinal veri kümesine bir regresör yerleştirerek başlayan ve ardından aynı veri kümesine regresörün ek kopyalarını yerleştiren, ancak örneklerin ağırlıklarının mevcut tahminin hatasına göre ayarlandığı bir meta-tahmin edicidir. Bu nedenle, sonraki regresörler daha çok zor durumlara odaklanır.
* Torbalama regresörü, temel regresörleri orijinal veri kümesinin rasgele alt kümelerine yerleştiren ve ardından nihai bir tahmin oluşturmak için bireysel tahminlerini (oylama veya ortalama alarak) toplayan bir topluluk meta-tahmincisidir. Böyle bir meta-tahmin edici, tipik olarak, bir kara kutu tahmin edicisinin (örneğin, bir karar ağacı) varyansını azaltmanın bir yolu olarak, yapım prosedürüne rasgeleleştirmeyi dahil ederek ve ardından bundan bir topluluk oluşturarak kullanılabilir.

Bu modeller ilk olarak import edilerek kısaltmaları tanımlanmıştır.

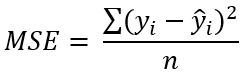


Daha sonra kategorik değerler sayısal değerlere dönüştürülmüştür. Veri seti üzerinden eğitim ve test verileri tanımlanmıştır.



### ***Tahminleme***

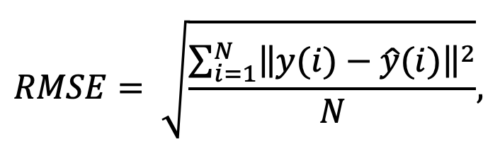
Mean Square Error (MSE) yeniden yapılandırılmış ve orijinal sonuç arasındaki kümülatif kare hatasını temsil eder. MSE , hata karelerinin ortalamasını, yani tahmin edilen değerler ile gerçek değer arasındaki ortalama kare farkını ölçer. Karesi alınmış hata kaybının beklenen değerine karşılık gelen bir risk fonksiyonudur. MSE, hatanın (başlangıç noktasıyla ilgili) ikinci momentidir ve bu nedenle hem tahmin edicinin varyansını hem de yanlılığını içerir. MSE ne kadar düşük olursa, hata o kadar düşük olur.



burada N, veri noktalarının sayısıdır, y(i) i'inci ölçümdür ve y ̂(i) buna karşılık gelen tahmindir.

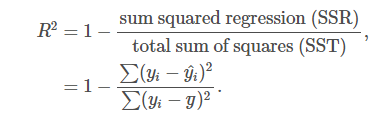
Makine öğreniminde, eğitim, çapraz doğrulama veya dağıtımdan sonra izleme sırasında bir modelin performansını yargılamak için bir referansa sahip olmak son derece yararlıdır.Kök ortalama kare hatası veya kök ortalama kare sapması (RMSE), tahminlerin kalitesini değerlendirmek için en sık kullanılan ölçütlerden biridir. Öklid mesafesini kullanarak tahminlerin ölçülen gerçek değerlerden ne kadar uzak olduğunu gösterir.

RMSE'yi hesaplamak için, her veri noktası için kalıntıyı (tahmin ve gerçek arasındaki fark) hesaplayın, her veri noktası için kalıntı normunu hesaplayın, artıkların ortalamasını hesaplayın ve bu ortalamanın karekökünü alın. RMSE, tahmin edilen her veri noktasında gerçek ölçümler kullandığından ve buna ihtiyaç duyduğundan, RMSE denetimli öğrenme uygulamalarında yaygın olarak kullanılır. Kök ortalama kare hatası şu şekilde ifade edilebilir:

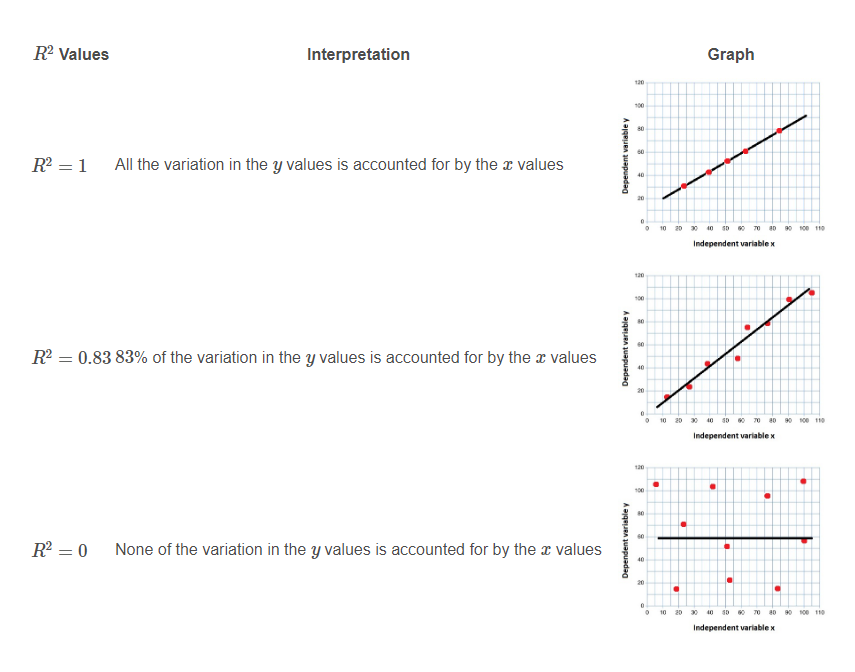


burada N, veri noktalarının sayısıdır, y(i) i'inci ölçümdür ve y ̂(i) buna karşılık gelen tahmindir.

R-kare, bir regresyon modelinin uyum iyiliğini temsil eden istatistiksel bir ölçüdür ve makine öğrenimi modelleri için performans değerlendirmede kullanılan ölçütlerinden biridir. Basitçe ifade etmek gerekirse, veri setindeki örnekler ile model tarafından yapılan tahminler arasındaki farktır. Regresyon modellerinin uyum başarısı, R-kare yöntemi temelinde analiz edilebilir. r-kare için ideal değer 1'dir. r-kare değeri 1'e ne kadar yakınsa, model o kadar uyumludur. r-kare değeri 1'e ne kadar yakınsa, model o kadar iyidir.R-kare formülü şekilde ifade edilebilir: (c3.ai, 2021)

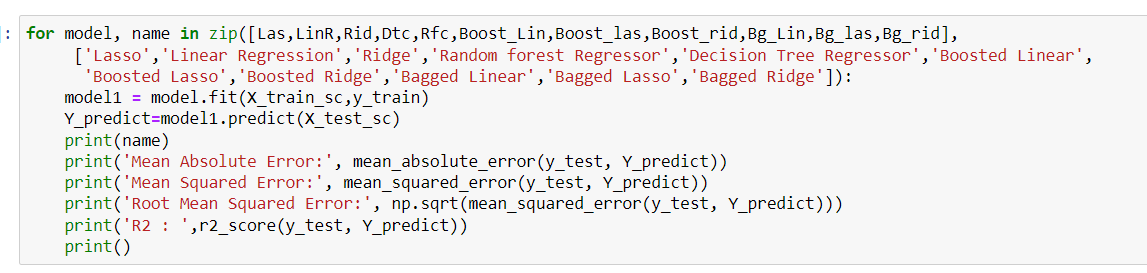


Aşağıda R-kare’nin görsel örnekleri verilmiştir:



Şekil 3. R-kare’nin görsel örnekler üzeirnden açıklanması (Newcastle University, n.d.)

Tüm modeller için performans ölçüm skorları hesaplanmış ve basılmıştır:



* Lasso

Mean Absolute Error: 6.5616011731794694

Mean Squared Error: 92.5537243324849

Root Mean Squared Error: 9.620484620458832

R2 : 0.9796133636669103

* Linear Regression

Mean Absolute Error: 1.171746362222105e-13

Mean Squared Error: 3.307328887463935e-26

Root Mean Squared Error: 1.8186063035918288e-13

R2 : 1.0

* Ridge

Mean Absolute Error: 0.0006677719559795821

Mean Squared Error: 9.578576021005984e-07

Root Mean Squared Error: 0.0009787019986188842

R2 : 0.9999999997890144

* Random forest Regressor

Mean Absolute Error: 1.5927324538995389

Mean Squared Error: 14.537500753664396

Root Mean Squared Error: 3.8128074634925375

R2 : 0.9967978518077533

* Decision Tree Regressor

Mean Absolute Error: 0.6999674416981092

Mean Squared Error: 5.695829460503227

Root Mean Squared Error: 2.3865936940550285

R2 : 0.9987453902620986

* Boosted Linear

Mean Absolute Error: 7.048807522938558e-14

Mean Squared Error: 9.793282592961399e-27

Root Mean Squared Error: 9.89610155210697e-14

R2 : 1.0

* Boosted Lasso

Mean Absolute Error: 1.2461952405312477

Mean Squared Error: 2.5321156682537707

Root Mean Squared Error: 1.5912622876992248

R2 : 0.9994422556017674

* Boosted Ridge

Mean Absolute Error: 0.00021922635392724354

Mean Squared Error: 8.791512500867104e-08

Root Mean Squared Error: 0.0002965048482043271

R2 : 0.999999999980635

* Bagged Linear

Mean Absolute Error: 1.1257242115449464e-13

Mean Squared Error: 3.361857160008373e-26

Root Mean Squared Error: 1.8335367899249726e-13

R2 : 1.0

* Bagged Lasso

Mean Absolute Error: 6.556020611328913

Mean Squared Error: 92.41090790666637

Root Mean Squared Error: 9.613059237655117

R2 : 0.979644821574807

* Bagged Ridge

Mean Absolute Error: 0.0006672289635559722

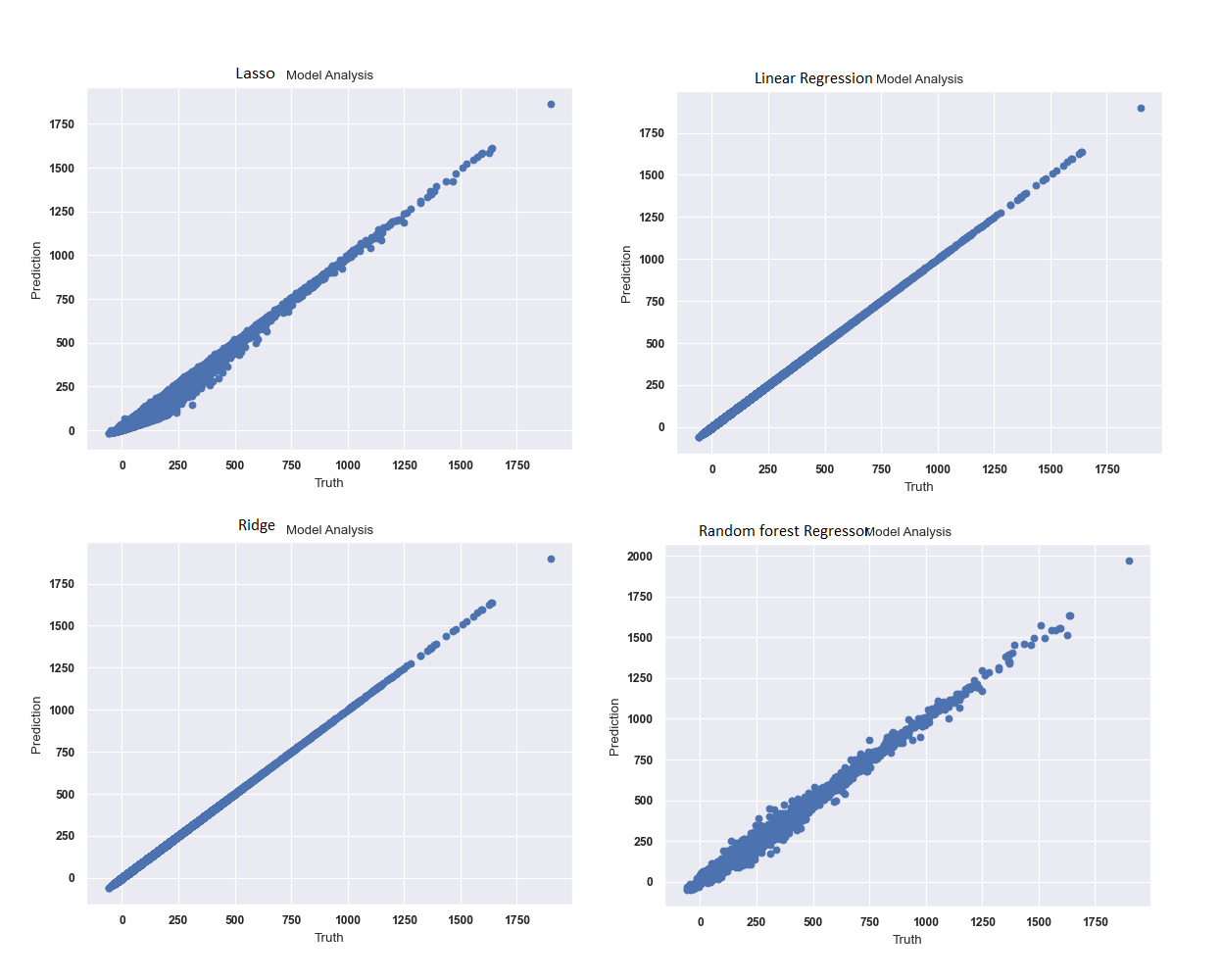
Mean Squared Error: 9.564448622310568e-07

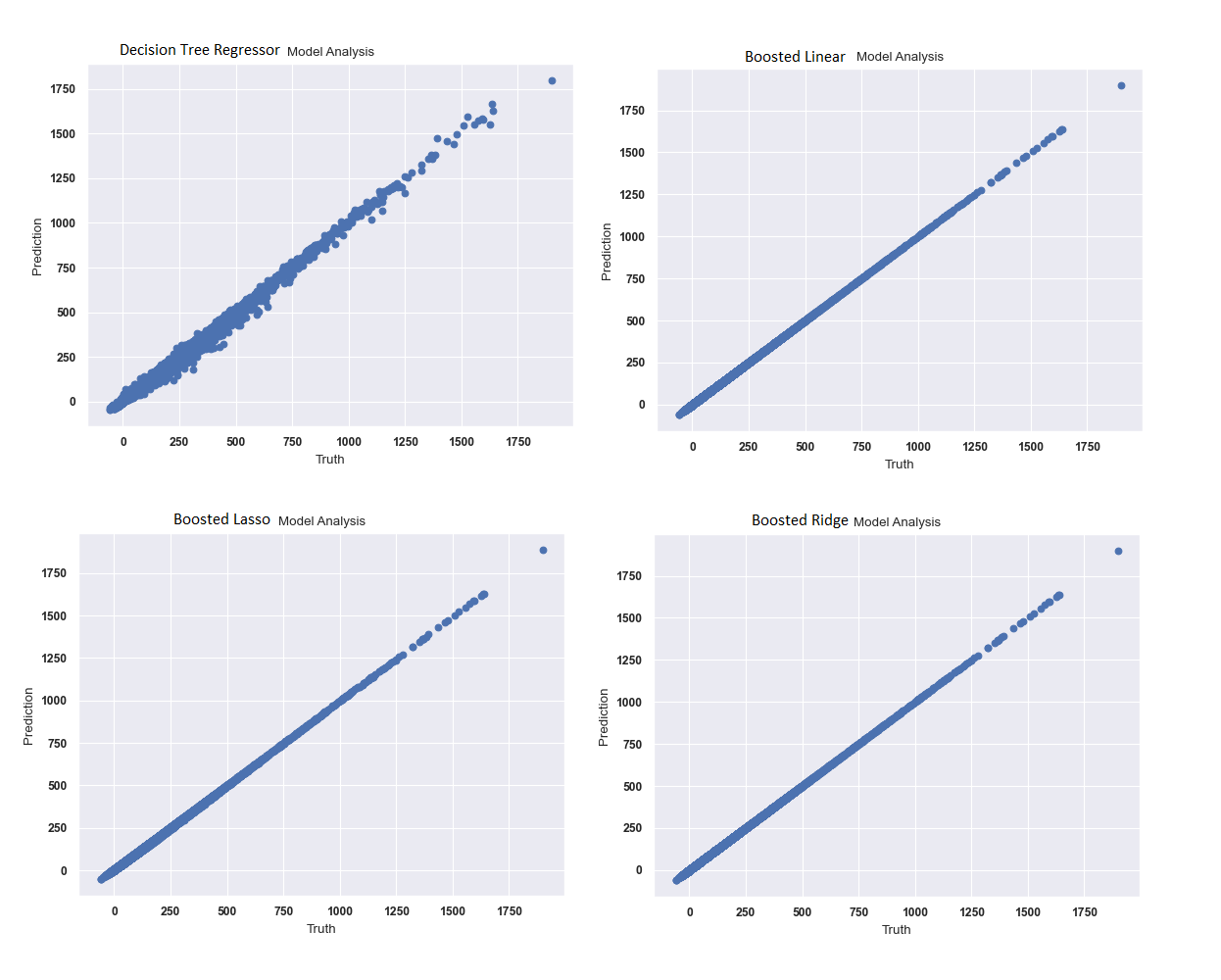
Root Mean Squared Error: 0.0009779799907109843

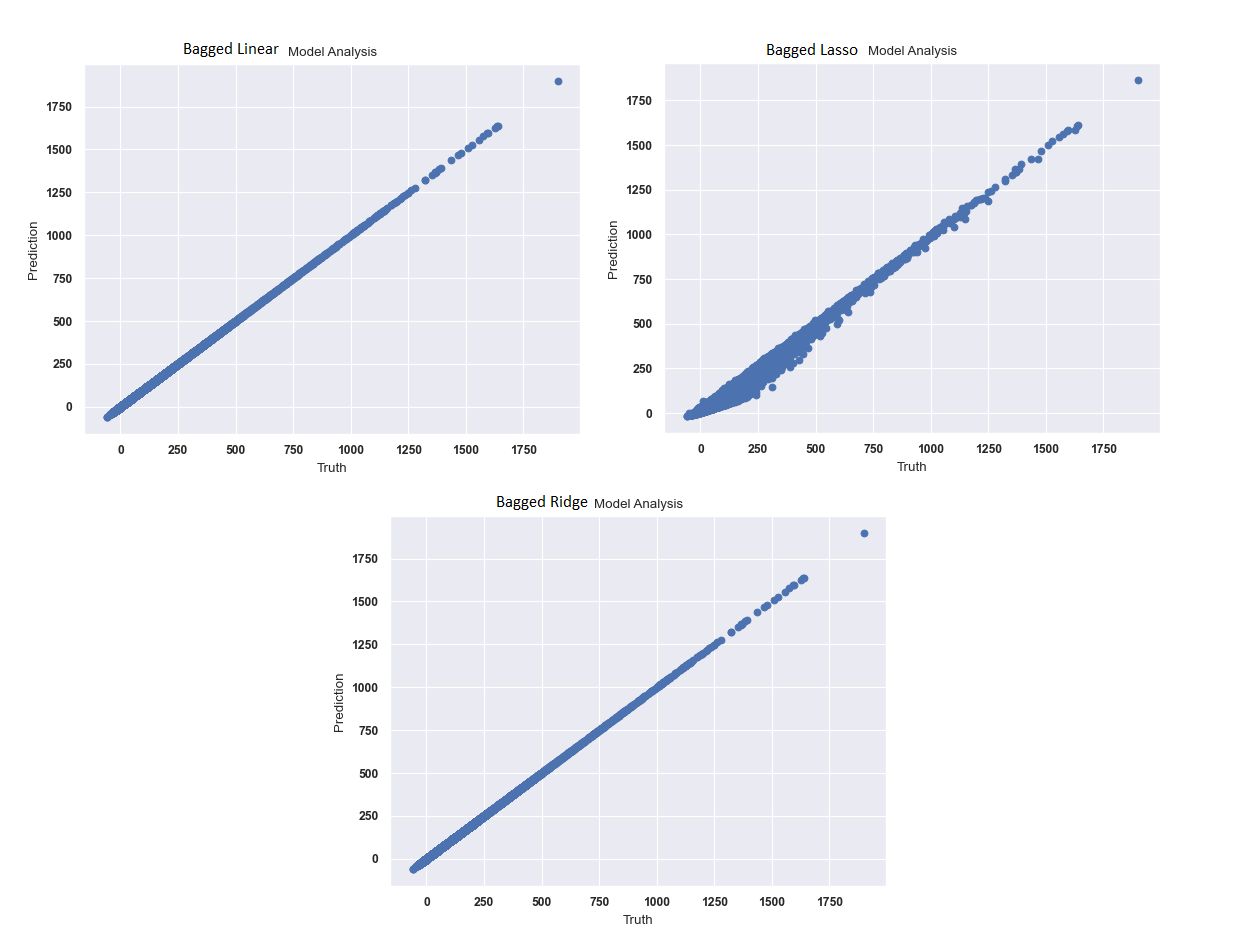
R2 : 0.9999999997893256

### ***Model Doğruluk Analizi***

Değişkenlere baktığımız zaman modellerin başarılı tahminlerde bulunduğu görülmektedir. Görsel olarak bakıldığında bu gözlem doğrulanmıştır.

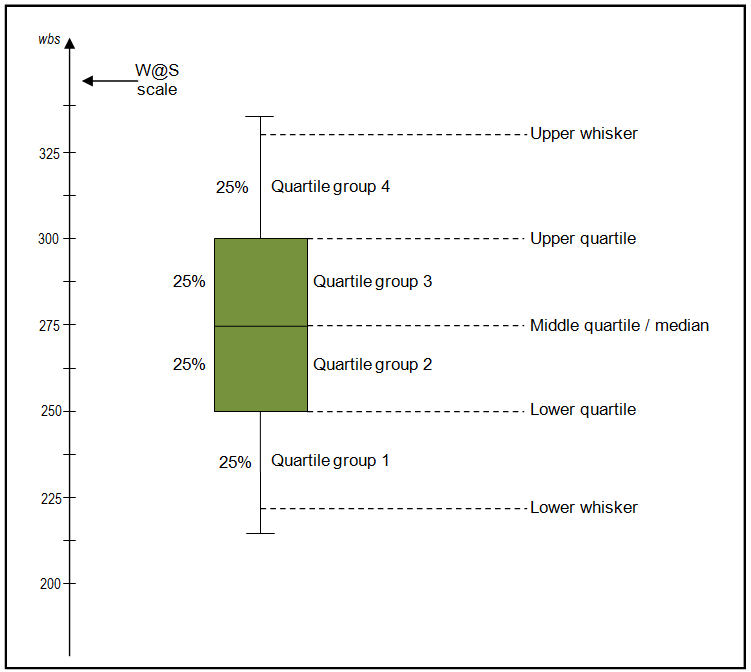






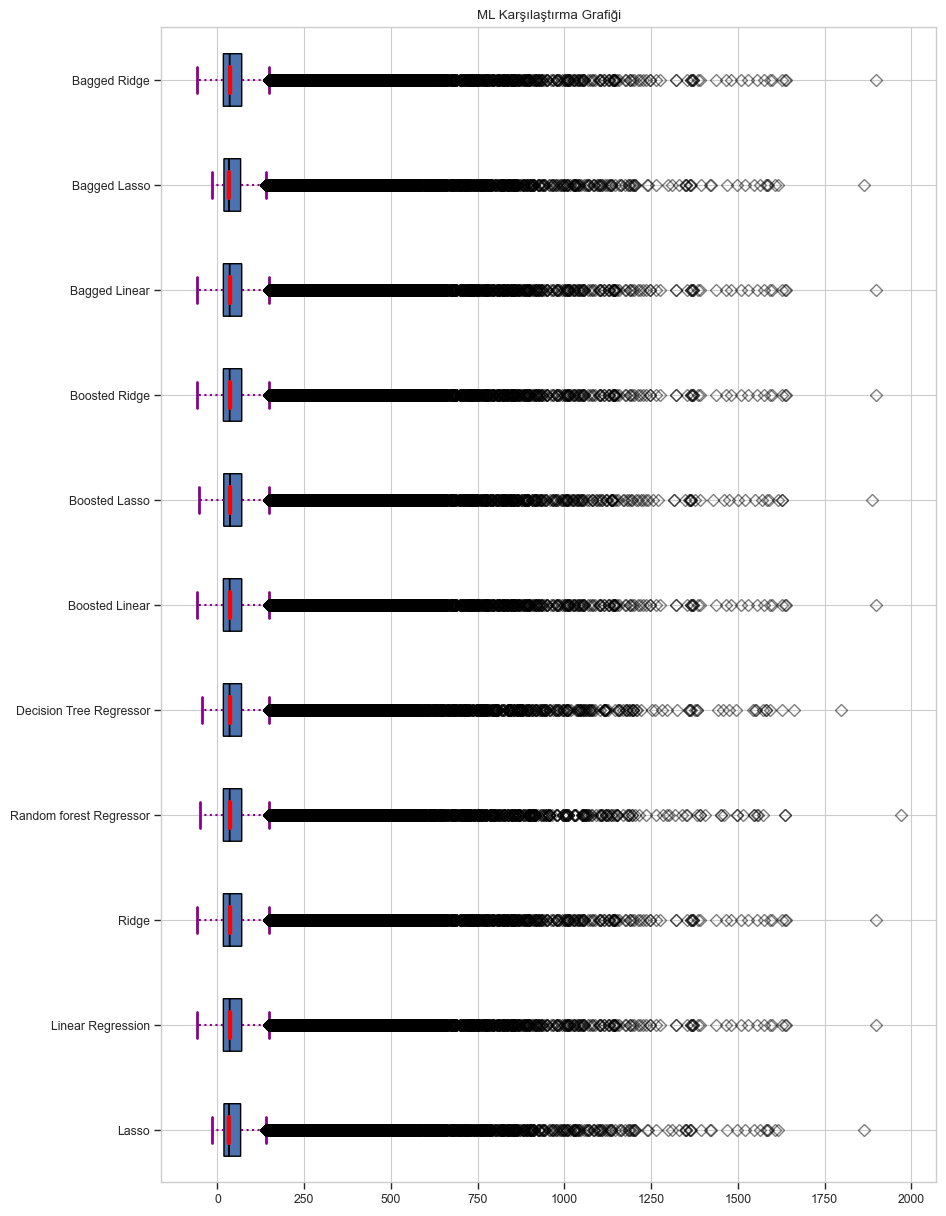
Modellerin daha sistematik olarak analiz edilebilmesi ve birbirleriyle karşılaştırmaların yapılabilmesi için Katardjiev, McKeever, ve Andreas Hamfelt’in (2019) model karşılaştırmalarını yapmak için önerdiği Box Whisker Plot yöntemi uygulanmıştır.

Bir Box Whisker Plot, sürekli bir değişkenin dağılımını, ortalama ve standart sapma ile kaplanmış çeyrekler, medyan ve aykırı değerler gibi bilgilerle göstermenin basit bir görsel yoludur. Kutu bıyık veya çeyreklik çizelgelerin ana çekiciliği, birden fazla özelliğin dağılımlarının yan yana karşılaştırılabilmesi ve aralarındaki örtüşmenin çıkarılabilmesidir. Çeyrekler, %25 kutu boyutuna sahip veri noktalarını gösteren Q1, Q2 ve Q3 noktaları ile gösterilir. Bir dağılımda veri noktalarının %25'i Q1'in altında, %50'si Q2'nin altında ve %75'i Q3'ün altında olacaktır (Kotu & Deshpande, 2019). Q1 ve Q3 noktaları, kutunun kenarlarıyla gösterilir. Dağılımın medyanı olan Q2 noktası, kutunun içinde bir çapraz çizgi ile gösterilir. Bıyık, değer aralığını temsil eder ve çoğu istatistiksel pakette kullanılan varsayılan seçenek, bıyığı çeyrekler arası aralığın 1,5 veya 3 katına çıkarmaktır. Kutu grafiği dikey olarak sunulursa, kutunun üst kenarından alt kenarına kadar olan alan, çeyrekler arası aralığı temsil eder. Aykırı değerler, bıyık çizgisinin sonundaki dairelerle gösterilir. Bazı durumlarda, ortalama nokta katı bir nokta kaplaması ve ardından bir çizgi kaplaması olarak standart sapma ile gösterilir (Kotu & Deshpande, 2019; Forthofer et al., 2007).

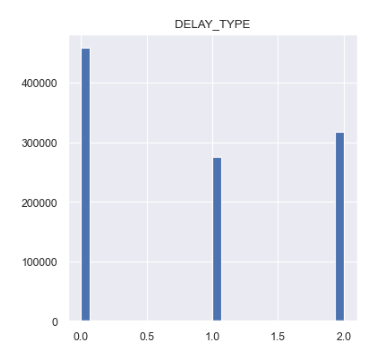


Şekil 4. Box Whisker Plot gösterimi (Wellbeing@School, n.d.)

Aşağıda verilen Box Whisker Plot grafiklerine bakıldığı zaman en iyi modelin Linear Regression modeli olduğu görülmektedir. Bu model hem hız hem de etkinlik açısından başarılı bir modeldir. Boosted Linear, Boosted Lassoi Boosted Ridge ve Bagged Linear modelleri de Linear Regression modeline yakın başarı göstermişlerdir ancak algoritmaların yüksek çalışma süreleri nedeniyle optimal olmadıkları görülmüştür. Random forest Regressor ve Decision Tree Regressor ise en düşük başarı gösteren modeller olmuştur. Feature selection adımında; uçuş zamanında gerçekleşti-gecikti, geciken uçuş küçük, orta, büyük ölçekli gerçekleşti gibi çok katmanlı bir model yapısı oluşturulsaydı bu modeller daha yüksek başarıya ulaşabilirdi.



Benzer şekilde oluşturulan delay\_type sınıflandırması üzerinden tahminleme yaptığımız zaman sonuçlar aşağıdaki gibi olmaktadır:



# 1 = 15 dakikadan fazla ve 30 dakikadan az gecikme

# 2 = 30 dakikadan fazla ve 1 saatten az gecikme

# 3 = 1 saatten fazla gecikme

* Lasso

Mean Absolute Error: 0.7545276880430555

Mean Squared Error: 0.7190060080045525

Root Mean Squared Error: 0.8479422197323073

R2 : -7.447380672376269e-06

* Linear Regression

Mean Absolute Error: 0.48295454022689754

Mean Squared Error: 0.35881938898442184

Root Mean Squared Error: 0.5990153495399111

R2 : 0.5009470612841773

* Ridge

Mean Absolute Error: 0.48463150009976075

Mean Squared Error: 0.35889929818001576

Root Mean Squared Error: 0.5990820462841594

R2 : 0.5008359220310716

* Random forest Regressor

Mean Absolute Error: 0.0

Mean Squared Error: 0.0

Root Mean Squared Error: 0.0

R2 : 1.0

* Decision Tree Regressor

Mean Absolute Error: 0.0

Mean Squared Error: 0.0

Root Mean Squared Error: 0.0

R2 : 1.0

* Boosted Linear

Mean Absolute Error: 0.5351719140180246

Mean Squared Error: 0.38887962499349527

Root Mean Squared Error: 0.6236021367775252

R2 : 0.45913870426846826

* Boosted Lasso

Mean Absolute Error: 0.7546242108249804

Mean Squared Error: 0.7190030464715926

Root Mean Squared Error: 0.8479404734246341

R2 : -3.3284234139951963e-06

* Boosted Ridge

Mean Absolute Error: 0.534518137035917

Mean Squared Error: 0.3884713699841107

Root Mean Squared Error: 0.6232747146997949

R2 : 0.45970651322315015

* Bagged Linear

Mean Absolute Error: 0.48465050724970277

Mean Squared Error: 0.3587773025811282

Root Mean Squared Error: 0.5989802188562893

R2 : 0.5010055958670022

* Bagged Lasso

Mean Absolute Error: 0.7545791497461135

Mean Squared Error: 0.7190042826049495

Root Mean Squared Error: 0.8479412023277024

R2 : -5.047661569168582e-06

* Bagged Ridge

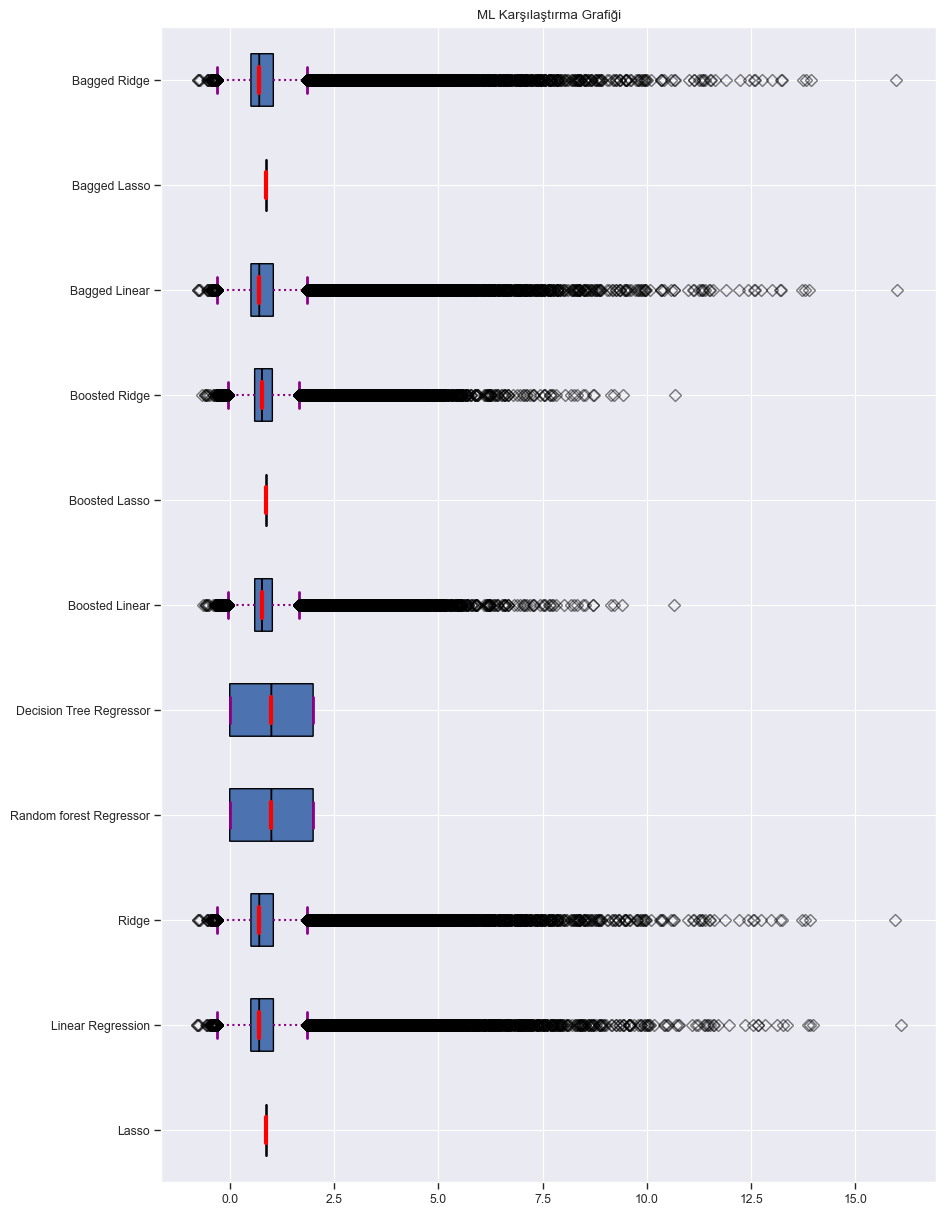
Mean Absolute Error: 0.4845100585545789

Mean Squared Error: 0.35891920478025524

Root Mean Squared Error: 0.5990986603058425

R2 : 0.5008082355468562

Burada iki model dışındaki tüm modellerin başarılarının oldukça düştüğünü görüyoruz. Box Whisker Plot İle bunu kontrol ettiğimiz zaman, Random forest Regressor ve Decision Tree Regressor modellerinin yüzde yüz başarı ile tahminleme yaptığını görüyoruz. Bu durum feature selection adımında seçtiğimiz değişkenlerin optimal şekilde seçildiğini göstermektedir.



# **Sonuç**

Havacılık endüstrisi, küresel ekonominin önemli bir bileşenidir ve insanları ve malları dünyanın her köşesine taşımaktan sorumludur. Ancak sektör, sert hava koşulları, mekanik arızalar veya fazla rezervasyon gibi çeşitli faktörler nedeniyle aksamalara ve gecikmelere eğilimlidir. Uçuş gecikmeleri, genellikle zaman ve üretkenlik kaybına ve ayrıca ek masraflara neden olduğundan, havayolları ve yolcuları üzerinde büyük bir ekonomik etkiye sahip olabilir. Ek olarak, uçuş gecikmeleri uçuş ekosisteminde zincirleme etkiler yaratarak tüm seyahat programını kesintiye uğratabilecek daha fazla gecikmeye ve zarar neden olabilir. Bu nedenle, havacılık endüstrisinin uçuş gecikmelerinin önlenmesine öncelik vermesi esastır. Veri analitiği, havayollarının uçuş gecikmelerini önlemesine yardımcı olarak son yıllarda havacılık endüstrisinde devrim yaratmaktadır. Havayolları tahmine dayalı analitiği kullanarak operasyonlarındaki eğilimleri belirleyebilir, operasyonel verimsizlikleri tespit edebilir ve olası gecikmeleri gerçekleşmeden önce tahmin edebilir. Makine öğrenimi teknikleri, gecikmelerin temel nedenlerinin belirlenmesine de yardımcı olarak havayollarının bunları ele almasını kolaylaştırabilir. Sonuç olarak havayolları, müşteri memnuniyetini artırmaya ve gecikmelerle ilişkili maliyetleri azaltmaya yardımcı olmak için makine öğrenimi tekniklerini kullanabilir.

Havacılık endüstrisindeki uçuş gecikmelerini tahmin etmek için makine öğrenimi tekniklerini kullanmak, yolcuların bekleme süresini ve gecikmeler nedeniyle havayollarının kaybettiği para miktarını azaltmanın harika bir yoludur. Havayolları, makine öğrenimi algoritmalarını kullanarak potansiyel uçuş gecikmelerini hızlı ve doğru bir şekilde tahmin edebilir ve bunları azaltmak veya hafifletmek için önlemler alabilir. Bu, proaktif programlamayı, kaynakların daha iyi yönetimini ve havaalanı personeli arasında geliştirilmiş iletişimi içerebilir. Makine öğrenimi teknikleri, gecikmelerin temel nedenlerinin belirlenmesine de yardımcı olarak havayollarının bunları ele almasını kolaylaştırabilir. Sonuç olarak havayolları, müşteri memnuniyetini artırmaya ve gecikmelerle ilişkili maliyetleri azaltmaya yardımcı olmak için makine öğrenimi tekniklerini kullanabilir.

Bu çalışmada Amerika’daki uçuşlara ait 2015 yılı verileri kullanılarak uçuş gecikmelerini etkileyen faktörler belirlenmiş ve makine öğrenmesi teknikleri kullanılarak uçuşlardaki gecikmeler üzerine bir tahminleme çalışması yapılmıştır. Yapılan analizlerde aylar, ayın güncleri, haftanın günleri, mesafe gibi sayısal değişkenler ve kalkış gecikmeleri ve varış gecikmeleri arasındaki ilişkiler incelenmiş; bu değişkenler arasında herhangi bir bağlantı bulunmadığı gözlenmiştir. Kategorik değişkenler olan varış ve kalkış havaalanı ve havayolu şirketleri incelendiğinde gecikmelerle yakından ilgili değişkenler oldukları belirlenmiştir. Bu analizler sonrasında gecikmelerle ilikili değişkenler feature selection adımında seçilmiş ve veri seti tahminlemeye hazır hale getirilmiştir. 'Lasso', 'Linear Regression', 'Ridge', 'Random forest Regressor', 'Decision Tree Regressor', ‘Boost regression', ve 'Bagged Regression’' yöntemleri ile tahminlemeler yapılmıştır.

İlk olarak varış gecikmesi üzerine modelleme yapılmıştır. Bu modellemeye ait tahminlemelere bakıldığı zaman en iyi modelin hem hız hem de tahmin başarısı açısından Linear Regression modeli olduğu görülmektedir. Boosted Linear, Boosted Lassoi Boosted Ridge ve Bagged Linear modelleri de Linear Regression modeline yakın başarı göstermişlerdir ancak algoritmaların yüksek çalışma süreleri nedeniyle optimal olmadıkları görülmüştür.

Gecikme süre aralıklarına göre oluşturulan delay\_type kategorik değerlerine göre tahminleme yapıldğında ise iki model dışındaki tüm modellerin başarılarının oldukça düştüğü görülmektedir. Box Whisker Plot ile bunu kontrol edildiği zaman, Random forest Regressor ve Decision Tree Regressor modellerinin yüzde yüz başarı ile tahminleme yaptığı görülmektedir. Bu durum feature selection adımında seçtiğimiz değişkenlerin optimal şekilde seçildiğini göstermektedir.

Bu adımda yapılan tahminlemeler en fazla gecikme yaşayan havalimanlarını belirlemek ve bu daha sonra uçuş programlarını ayarlamak ve/veya gecikme sayısını azaltmak için uçuşları yeniden yönlendirmek için kullanılabilir. Buna ek olarak, tahmine dayalı analitik, gecikme yaşama olasılığı daha yüksek olan uçakları belirlemek için kullanılabilir ve bu, gelecekteki gecikme olasılığını azaltmak için personel ve bakım programlarını ayarlamak için kullanılabilir.

Özetle havayolu şirketleri ve havaalanları tahmine dayalı veri analitiği ve makine öğrenmesi metodlarına yönelerek gecikme riskini azaltarak müşterileri için daha iyi bir deneyim sağlamanın yanı sıra zamandan, paradan ve kaynaklardan tasarruf sağlayabilir. Çalışmadaki analiz sonuçları aynı zamanda havayou şirketlerinincheck-in işlemi, bagaj taşıma ve güvenlik kontrolleri için havaalanına daha fazla personel ve kaynak sağlayarak gecikmeleri azaltabileceğini işaret etmektedir. İncelenen veri setinin sadece bir seneye ve tek bir ülkeye ait olması, temizleme aşamasında verilerin doldurulması/kurtarılması yerine boş değerlerin silinmesi yolunun seçilmesi çalışmanın kısıtlılıkları arasındadır. Bu durum göz önüne alınarak ilerleyen çalışmalarda daha kapsamlı veri setleri kullanılarak çalışmalar yapılmalıdır.

# **Kaynakça**

Bureau of Transportation Statistics (BTS) (2009) Airline on-time statistics and delay causes: on-time arrival performance. Available at: [http://www.transtats.bts.gov/OT\_ Delay/ot\_delaycause 1.asp?display=data&pn=1](http://www.transtats.bts.gov/OT_)

Hemadri A D, & Kumar Raja D R. (2022). Flight Delay Prediction Using Machine Learning Algorithm. International Journal of Human Computations & Intelligence, 1(2), 5–9. Retrieved from <https://milestoneresearch.in/JOURNALS/index.php/IJHCI/article/view/28>

Shruti S. Pophale, Purushottam R. Patil, Amol D. Potgantwar, Pawan R. Bhaladhare. (2022). Flight Delay Analysis and Prediction Using Machine Learning Algorithms. Mathematical Statistician and Engineering Applications, 71(4), 6071–6085. Retrieved from <https://philstat.org.ph/index.php/MSEA/article/view/1205>,

Ball, M. *et al.* (2010) *Total Delay Impact Study: A Comprehensive Assessment of the Costs and Impacts of Flight Delay in the United States*. Rep.

c3.ai (2021) *Root mean square error (RMSE)*, *C3 AI*. Available at: <https://c3.ai/glossary/data-science/root-mean-square-error-rmse/> (Accessed: January 11, 2023).

Carvalho, L. *et al.* (2020) “On the relevance of data science for Flight Delay Research: A systematic review,” *Transport Reviews*, 41(4), pp. 499–528. Available at: <https://doi.org/10.1080/01441647.2020.1861123>.

Forthofer, R. N., Lee, E. S., & Hernandez, M. (2007). Descriptive methods. *Biostatistics*, 21–69. <https://doi.org/10.1016/b978-0-12-369492-8.50008-x>

Katardjiev, Nikola & McKeever, Steve & Hamfelt, Andreas. (2019). A machine learning-based approach to forecasting alcoholic relapses. ITISE 2019: International Conference on Time Series and Forecasting; Granada (Spain); 25-27 Sep 2019; Available from <https://itise.ugr.es/ITISE2019_Vol2.pdf>

Kotu, V., & Deshpande, B. (2019). Data exploration. *Data Science*, 39–64. <https://doi.org/10.1016/b978-0-12-814761-0.00003-4>

Newcastle University (n.d.) *Coefficient of Determination, R-squared*, *Numeracy, Maths and statistics - academic skills kit*. Available at: <https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/coefficient-of-determination-r-squared.html> (Accessed: January 11, 2023).

The Irish Aviation Authority (no date) *Ground Operations*, *iaa.ie*. Available at: <https://www.iaa.ie/commercial-aviation/ground-operations#:~:text=For%20an%20aircraft%20operator%20involved,the%20aircraft%20prior%20to%20departure> (Accessed: January 9, 2023).

Scikit (no date) *User guide: Contents*, *https://scikit-learn.org/*. Available at: <https://scikit-learn.org/stable/user_guide.html> (Accessed: January 9, 2023).

Teodorović, D. and Janić, M. (2022) “Traffic control,” *Transportation Engineering*, pp. 293–403. Available at: <https://doi.org/10.1016/b978-0-323-90813-9.00006-0>

Tiwari, P. *et al.* (2019) “Sentiment analysis for airlines services based on the Twitter dataset,” *Social Network Analytics*, pp. 149–162. Available at: <https://doi.org/10.1016/b978-0-12-815458-8.00008-6>

Wellbeing@School (no date) *Understanding and interpreting box plots*. Available at <https://www.wellbeingatschool.org.nz/information-sheet/understanding-and-interpreting-box-plots> (Accessed: January 11, 2023)