

TRAFİK KAZA VERİLERİNİN ANALİZİ & KAZA ŞİDDETLERİNİN TAHMİN EDİLMESİ

Hazırlayan: MERT AÇIKGÖZ (318631062)

İSTANBUL ŞEHİR ÜNİVERSİTESİ

HAZİRAN 2020

T.C.

İSTANBUL ŞEHİR ÜNİVERSİTESİ

İŞLETME ENSTİTÜSÜ

İŞLETME ANABİLİM DALI

İŞ ANALİTİĞİ TEZSİZ YÜKSEK LİSANS PROGRAMI

BİTİRME PROJESİ

**TRAFİK KAZA VERİLERİNİN ANALİZİ & KAZA ŞİDDETLERİNİN TAHMİN EDİLMESİ**

Hazırlayan: **MERT AÇIKGÖZ** (318631062)

Danışman: Doç. Dr. **ENES ERYARSOY**

HAZİRAN 2020

ONAY SAYFASI

Bu proje tarafımca okunmuş olup kapsam ve nitelik açısından, iş analitiği alanında Yüksek Lisans Derecesini alabilmek için yeterli olduğuna karar verilmiştir.

Doç. Dr. ENES ERYARSOY

Bu projenin İstanbul Şehir Üniversitesi İşletme Enstitüsü tarafından konulan tüm standartlara uygun şekilde yazıldığı teyit edilmiştir.

Tarih Mühür/İmza

İNTİHAL TUTANAĞI

Bu çalışmada yer alan tüm bilgilerin akademik kurallara ve etik ilkelere uygun olarak toplanıp sunulduğunu, söz konusu kurallar ve ilkelerin zorunlu kıldığı çerçevede, çalışmada özgün olmayan tüm bilgi ve belgelere, alıntılama standartlarına uygun olarak referans verilmiş olduğunu beyan ederim.

Mert AÇIKGÖZ

İmza:

ÖZ

TRAFİK KAZA VERİLERİNİN ANALİZİ & KAZA ŞİDDETLERİNİN TAHMİN EDİLMESİ

Açıkgöz, Mert:

Yüksek Lisans, İş Analitiği Bölümü

Proje Danışmanı: Doç. Dr. Enes Eryarsoy

Haziran 2020, 40 Sayfa

Bu proje iki aşamadan oluşmaktadır. Birinci aşamada aşağıda detayları açıklanacak verinin, tanımlayıcı analitik kapsamında oldukça kapsamlı bir şekilde analiz edilmesi ve analiz sonuçlarının grafiksel olarak gösterimi, ikinci aşamada ise veri çeşitli işlemlerden geçirildikten sonra, kazaların şiddet seviyelerinin çeşitli makine öğrenmesi sınıflandırma algoritmaları kullanılarak tahmin edilmesi yer alacaktır. Her iki aşamada da programlama dili olarak python kullanılacaktır. İkinci aşama olan tahmin bölümünde analitik platform olarak knime, python ile entegre edilerek kullanılmıştır. Projenin ikinci aşamasında CRISP-DM metodolojisi uygulanarak modelleme ve değerlendirme gerçekleştirilmiştir.

**Anahtar Kelimeler:** Sıcaklık, Nemlilik, Ülke, Zaman aralığı, Eyalet

İÇİNDEKİLER

ONAY SAYFASI …………………………………………………………………………………………………………iii

İNTİHAL TUTANAĞI......................................................................................................iv

ÖZ……………………………………………………………………………………………………………………………v

TRAFİK KAZA VERİLERİNİN ANALİZİ & KAZA ŞİDDETLERİNİN TAHMİN EDİLMESİ…………v

İÇİNDEKİLER TABLOSU ……………………………………………………….......................................1

TABLO VE GÖRSELLER LİSTESİ…………………………………………………………………..................3

1.1 Verinin Tanımı ……………………………………………………………………………………..........6

1.2 Verinin Temel Özelliklerinin İncelenmesi………………………………………………………6

1.2.1 Eksik Veri Kontrolü ve Tamamlanması………………………………………………............7

1.2.2 İstatistikler ve Korelasyon Analizi………………………………………………………….........9

1.3 Trafik Kazaları ile İlgili Analizleri İçeren Grafikler…………………………………………10

1.3.1 Kaza Şiddetleri ve Gün İçindeki Dağılımları………………………............................10

1.3.2 Şehir ve Eyaletlere Göre Kaza Sayıları.........………………………………………...........12

1.3.3 Zaman Bağlantılı Analizler………………………………………………………………………...14

1.3.3.1 Aylara Göre Kaza Sayıları………………………………………………………………..........…14

1.3.3.2 Günlere Göre Kaza Sayılar…………………………………………………………………........15

1.3.3.3 Yıllara Göre Kaza Sayıları…………………………………………….............................…16

1.3.3.4 Kazaların Saatlik Dağılımı……………………………………………………………….........…17

1.3.3.5 Trafik Kazası Olaylarının Aldığı Sürelerin Analizi………………...........................18

1.3.4 Hava Koşulları ile İlgili Analizler……………………… ………………….................…...19

1.3.4.1 Hava Koşullarına Göre Kaza Analizi 1…………………….................……................19

1.3.4.2 Hava Koşullarına Göre Kaza Analizi 2....................……...........………..............20

1.3.4.3 Hava Koşullarına Göre Kaza Yoğunluğunun Gösterilmesi…………….............22

1.3.5 Kaza Kodları Dağılımı……………………………………………………………………………...23

2.1 CRISP-DM Metodolojisinin Aşamaları...……………………........………………........21

2.2 Verinin Tanımı & İkinci Aşamanın Amacı……………….....................................25

2.3 Verinin Okunması ve Birleştirilmesi………………………....................................25

2.4 Veri Ön İşleme………………………………............................................................28

2.5 Modeldeki Bağlı Değişken Kayıt Sayılarının Dengelenmesi.........................31

2.6 Verinin Bölünmesi….................................................................................…33

2.7 Modelleme…………………......................................................................………33

2.7.1 Random Forest…………………………………………………………………………………….…33

2.7.2 Gradient Boosted Trees……………………………………………………………………….….35

2.7.3 Naive Bayes…………………………………………………………………………………………….35

2.8 Model Değerlendirme…………………………………………………………………………….36

2.8.1 ROC Curve High……………………………………………………………………………………….37

2.8.2 ROC Curve Low……………………………………………………………………………………….38

2.9 KNIME Workflow…………………………………………………………………………………….39

3. Kaynakça.....................................................................................................40

ÖZGEÇMİŞ..................................................................................................................40

TABLO VE GÖRSELLER LİSTESİ

şekil 1.2.1 dataframe load ………………………………………….......…………………………………………………………………6

şekil 1.2.2 head…………………………………………………………………..……………………………………………….………………6

şekil 1.2.3 head 2…………………………………………………………………..……………………………………………….……………7

şekil 1.2.4 shape…………………………………………………………………….………………………………………………..………….7

şekil 1.2.1.1 mıssıng value check………………………………………………....…………………………………………..………….7

şekil 1.2.1.2 ımputatıon………………………………………………………………......………………………………………..……….8

şekil 1.2.1.3 ımputatıon 2…………………………………………………………………......……………….……………………..……8

şekil 1.2.1.4 mıssıng value check 2...……………………………………………………….……………………………………………8

şekil 1.2.2.1 descrıbe...…………………………………………………………………………………………………………………….….9

şekil 1.2.2.2 columns………………………………………………………………………………………………………………………..…9

şekil 1.2.2.3 column type…………………………………………………………………………………………………………………....9

şekil 1.2.2.4 shape 2…………………………………………………………………………………………………………………………....9

şekil 1.2.2.5 correlatıon matrix……………………………………………………………………………………………...…………10

şekil 1.3.1.1 severıty types…………………………………………………………………………………………………..…..…………10

şekil 1.3.1.2 accıdent dıstrıbutıon………………………………………………………………………………….…………..………11

şekil 1.3.1.3 state accıdent……………………………………………………………………………………….…………………..…..11

şekil 1.3.2.1 state accıdent graph…………………………………………………………………………………………………..….12

şekil 1.3.2.2 state accıdent 2..…………………………………………………………………………...……………………………...12

şekil 1.3.2.3 state accıdent ınteractıve……………………………………………………………..………………………..………13

şekil 1.3.2.4 state accıdent interactive 2……………………………………………………………..………………………..……13

şekil 1.3.2.5 cıty accıdent ………..…………………………………………………………………………..………………………..….13

şekil 1.3.3.1.1 monthly accıdents …………………………………………………………………………...……………..…………14

şekil 1.3.3.1.2 monthly accıdents 2………………………………………………………………………….……………….……….14

şekil 1.3.3.2.1 daıly accıdents…………………………………..………………………………………………………………………..15

şekil 1.3.3.2.2 daıly accıdents 2…………………………………..……………………………………………………………………..15

şekil 1.3.3.3.1 yearly 1………………………………………………….…………………………………………………………………..16

şekil 1.3.3.3.2 yearly 2………………………………………………….…………………………………………….…………………….16

şekil 1.3.3.4.1 hourly……………………………………………………..……………………………………………..………………….17

şekil 1.3.3.4.2 day part……………………………………………………..……………………………………………..………………..17

şekil 1.3.3.5.1 tıme……………………………………………………………..…………………………………………………………..…18

şekil 1.3.3.5.2 tıme 2……………………………………………………………..………………………………………………….……….18

şekil 1.3.4.1.1 weather 1… ……………………………………………………..………………………………………………………….19

şekil 1.3.4.1.2 weather 2…………………………………………………………..……………………………………………………….19

şekil 1.3.4.1.2 weather 3……………………………………………………………..…………………………………………………….19

şekil 1.3.4.1.4 weather 4………………………………………………………………..………………………………………………….20

şekil 1.3.4.1.5 weather 5………………………………………………………………………………..…………………….……….….20

şekil 1.3.4.2.1 weather 6……………………………………………………………..…………………………………………….………20

şekil 1.3.4.2.2 weather 7………………………………………………………………..………………………………………………….21

şekil 1.3.4.2.3 weather 8…………………………………………………………………..……..………………………………..……..21

şekil 1.3.4.2.4 weather 9 …………………………………………………..………………………………………………………………22

şekil 1.3.4.3.1 weather 10…………………………………………………....…………………………………………………………..22

şekil 1.3.4.3.2 weather 11……………………………………………………....…………………………………………………..……23

şekil 1.3.5.1 accıdent codes dıst……………………………………………….………………………………………………………..23

şekil 1.3.5.2 accıdent codes dıst graph……………………………………….…………………………………………….………..23

şekil 2.1.1 crısp-dm...…………………………………………………………………..…………………………………………….………24

şekil 2.3.1 data read……………………………………………………………………..…………………………………………………..25

şekil 2.3.2 data write to excel………………………………………………………….…………………………………………..……26

şekil 2.3.3 data read and concatanatıon……………………………………………..…………………………………..…………27

şekil 2.3.4 data concatanatıon conf ……………………………………………………..…………………………………….…….27

şekil 2.3.5 python mappıng……………………………………………………………………..………………………………………….28

şekil 2.3.6 knıme concatanatıon………………………………………………………………....…………………………………....28

şekil 2.4.1 knıme meta node…………………………………………………………………………....……………………………..….29

şekil 2.4.2 knıme row fılter………………………………………………………………………………....…………………..……….29

şekil 2.4.3 knıme row fılter conf……………………………………………………………………....…....…………………………29

şekil 2.4.4 knıme numerıc bınner………………………………………………………………………………....…………………….30

şekil 2.4.5 knıme numerıc bınner conf…………………………………………………………………………….....……………….30

şekil 2.4.6 knıme rule engıne……………………………………………………………………………………………….....……….…30

şekil 2.4.7 knıme rule engıne conf………………………………………………………………………………………………..…….30

şekil 2.4.8 knıme column fılter conf…………………………………………………………………………………………..…..….31

şekil 2.4.9 knıme table output……………………………………………………………………………………………………..….…31

şekil 2.5.1 knıme balance meta node………………………............………………………………………………………….…….32

şekil 2.5.2 knıme balance meta node content………………………….............…………………………..............……….32

şekil 2.6.1 knıme partıtıon conf..............................................................................................................33

şekil 2.7.1.1 knıme random forest…………………………………………………………………….……………………..………..33

şekil 2.7.1.2 knıme random forest confusıon matrıx………………………..………………………………………….………34

şekil 2.7.1.3 knıme random forest accuracy statıstıcs……………………..………………………………………….………34

şekil 2.7.1.4 knıme random forest attribute statıstıcs ………………………..………………………………………..…....34

şekil 2.7.2.1 gradıent Boosted Trees…………………………………………………...……………………………………………..35

şekil 2.7.2.2 gradıent Boosted Trees confusıon matrıx…………………………..…………………………………………..35

şekil 2.7.2.3 gradıent Boosted Trees accuracy statıstıcs………………………….. ………………………………..……….35

şekil 2.7.3.1 naıve bayes………………………………………………………………………..………………………………….……….35

şekil 2.7.3.2 naıve bayes confusıon matrıx………………………………………………..……………………………..…………36

şekil 2.7.3.3 naıve bayes acucuracy statıstıcs…………………………………………………….………………………………..36

şekil 2.8.1.1 knıme roc curve hıgh……………………………………………………………………………………………........….37

şekil 2.8.2.1 knıme roc curve low……………………………………………………………………………………………….........38

şekil 2.9.1 knıme workflow....................................................................................................................39

# 1.1 Verinin Tanımı

Projenin İlk aşaması olan Trafik kaza verisinin analiz edilmesi kapsamında, öncelikle verinin tanımı yapalım. Bu veri seti Amerika ülkesi genelinde 49 eyaleti kapsamaktadır ve 2016 ile 2019 yıllarını içermektedir. Amerika ve eyaletleri taşımacılık, hukuk departmanları, yollardaki trafik kamera ve algılayıcılarından sağlanan verileri içerir. Veri setinde 3 milyon kayıt bulunmaktadır, ancak analiz çalışması kapsamında 1 milyon kayıt kullanılmıştır. Veri içerisinde kaza şiddeti, kaza başlangıç ve bitiş zamanları, cadde, şehir, eyalet, sıcaklık, nemlilik, rüzgar hızı, görüş astronomik kısıtlar gibi 49 adet öznitelik bulunmaktadır.

Veri aşağıdaki adresten indirilebilir:

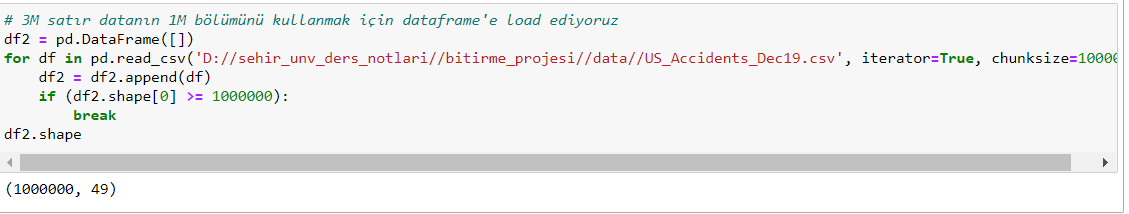
<https://drive.google.com/open?id=1Jv7yHBumRavNIOLootzvvkKQ_WqoA8OT>

Not: Kodların tamamı data\_analysis\_visualization\_318631062.ipynb dosyasında mevcuttur.

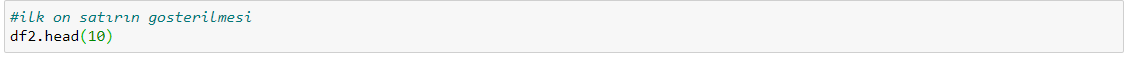
* 1. **Verinin Temel özelliklerinin İncelenmesi**

Verinin 1M satırlık bölümünü bir dataframe’e import ederek analiz etmek için aşağıdaki kod kullanılıyor.

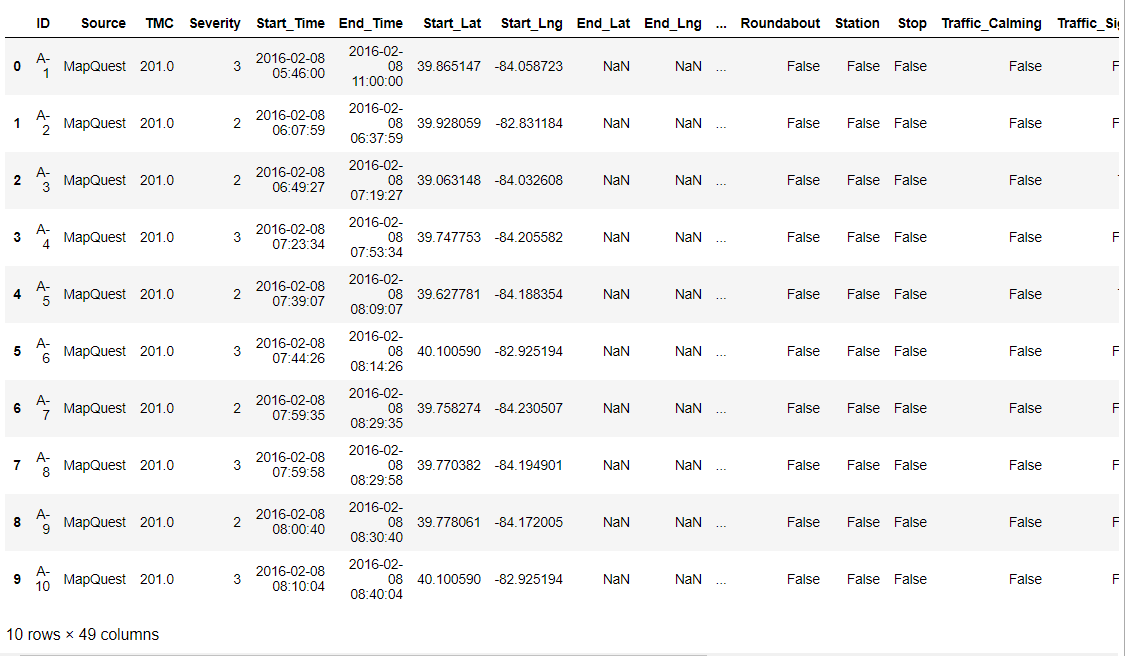
Not: Veri bilgisayarda nereye indirildi ise path o şekilde güncellenmelidir.



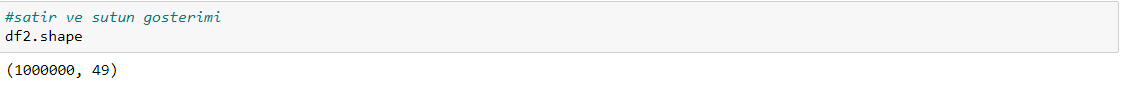
Şekil 1.2.1 datafame load



Şekil 1.2.2 head

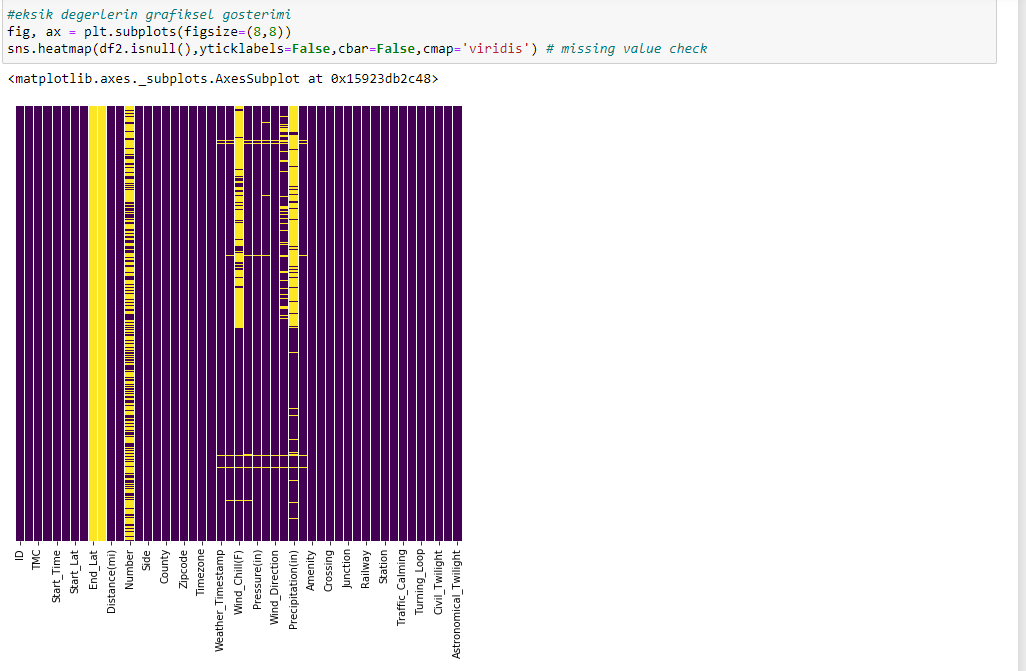


Şekil 1.2.3 head 2

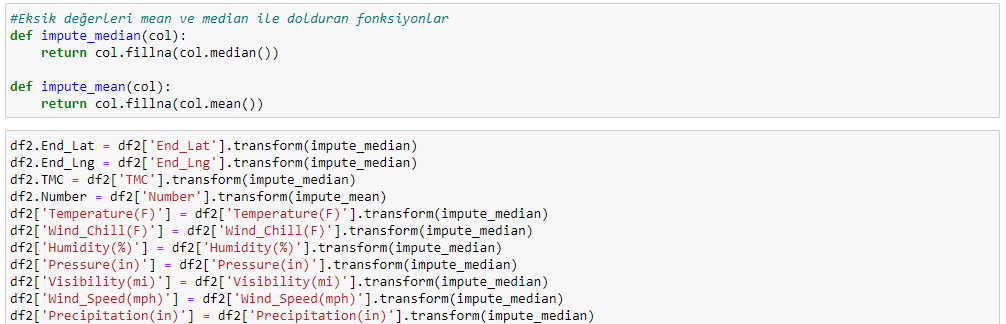


Şekil 1.2.4 shape

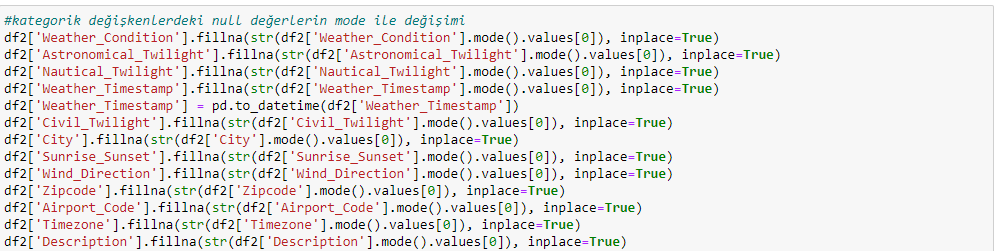
**1.2.1 Eksik Veri Kontrolü ve Tamamlanması**



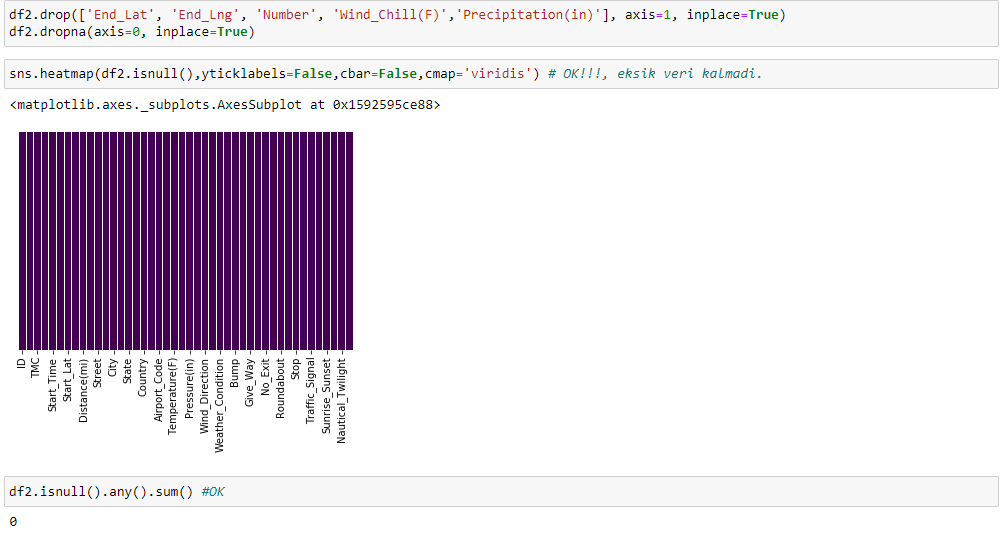
Şekil 1.2.1.1 mising value check



Şekil 1.2.1.2 imputation



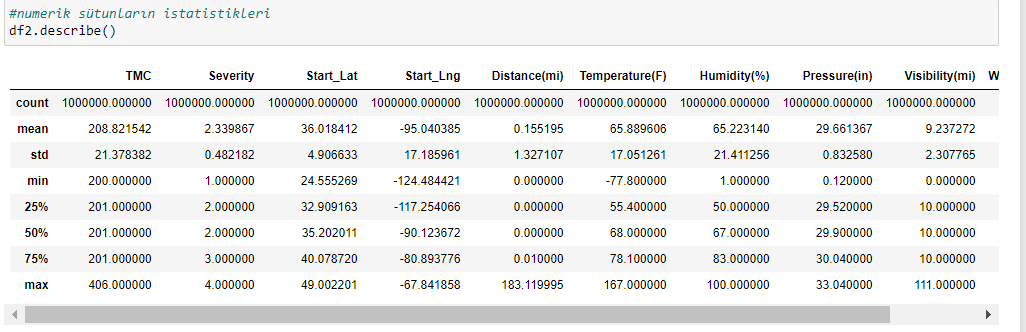
Şekil 1.2.1.3 imputation 2



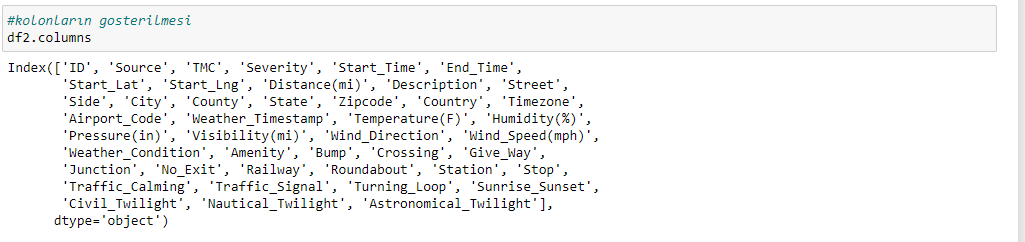
Şekil 1.2.1.4 missing value check 2

Görüldüğü gibi eksik veriler tamamlandıktan ve veride zaten eksik olan kolonlar atıldıktan sonra, eksik verimiz kalmadı.

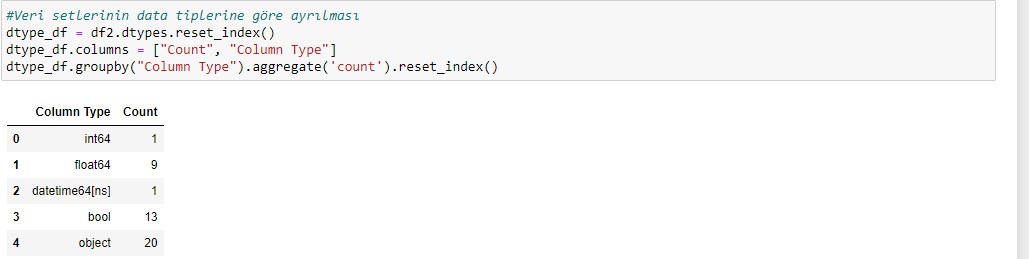
**1.2.2 İstatistikler ve Korelasyon Analizi**



Şekil 1.2.2.1 describe



Şekil 1.2.2.2 columns



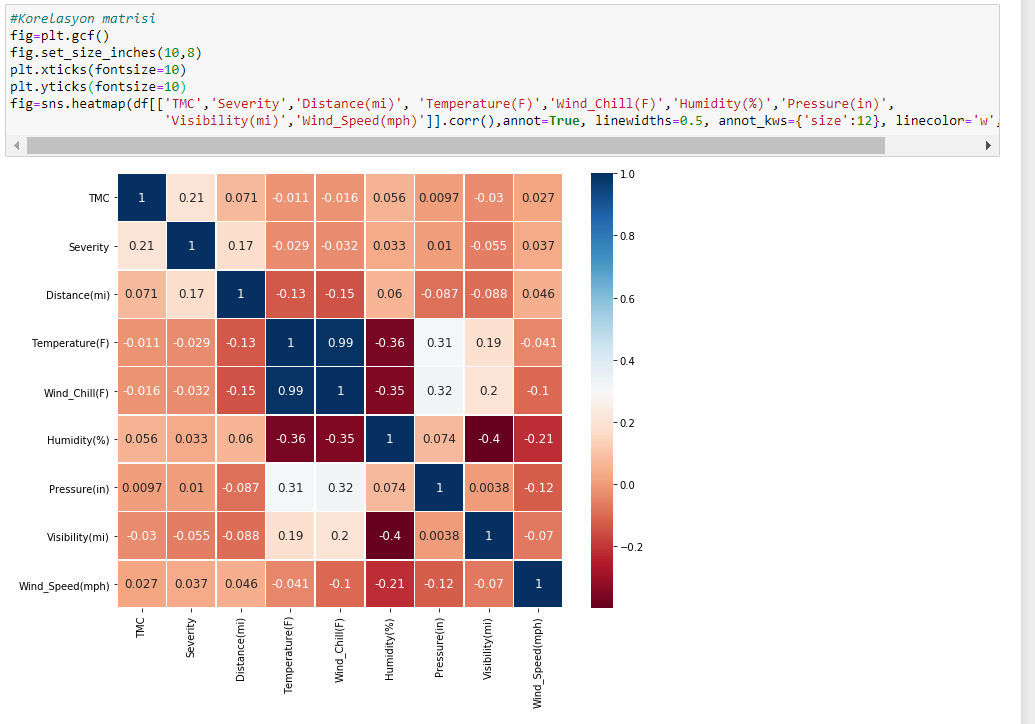
Şekil 1.2.2.3 column type

Verinin şu anki satır ve sütun sayıları:



Şekil 1.2.2.4 shape 2

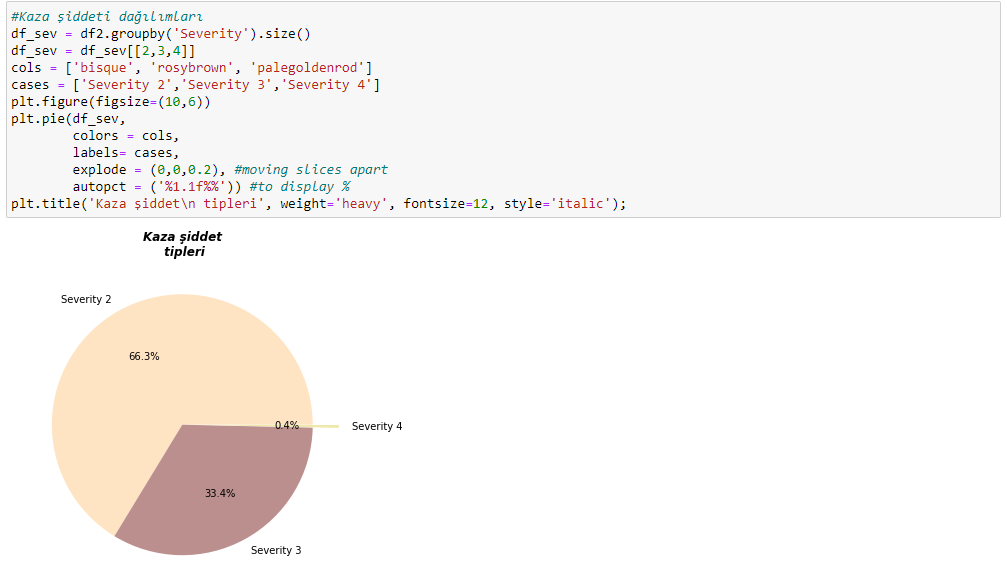
Önemli nümerik sütunların korelasyon matrisi:



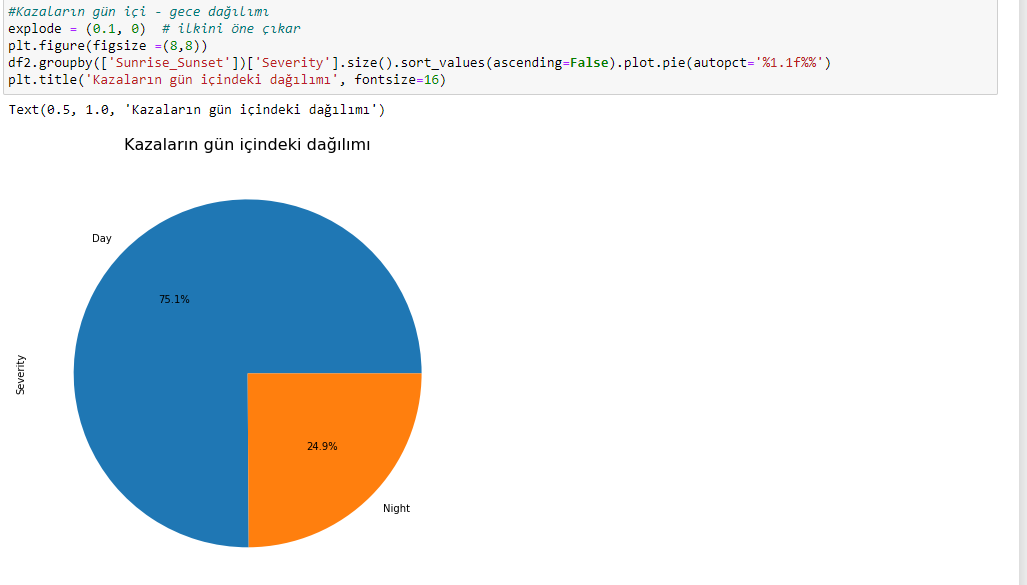
Şekil 1.2.2.5 correlation matrix

**1.3 Trafik Kazaları ile İlgili Analizleri İçeren Grafikler**

**1.3.1 Kaza Şiddetleri ve Gün İçindeki Dağılımları**

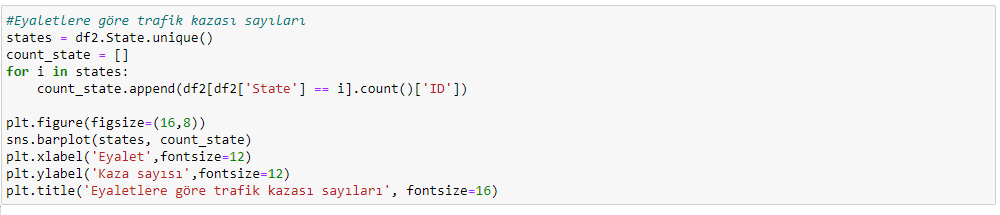


Şekil 1.3.1.1 severity types



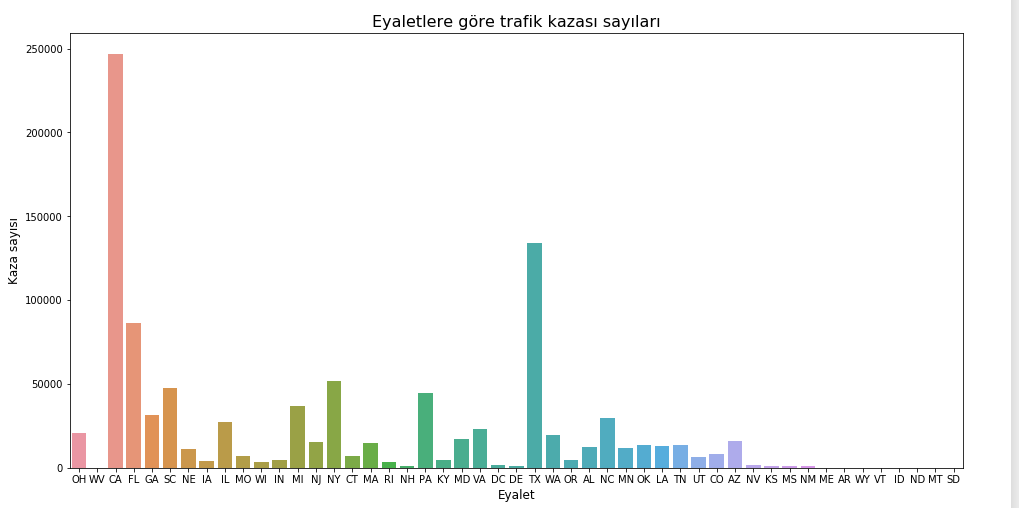
Şekil 1.3.1.2 accident distribution

Kazaların büyük bölümünün gün içinde olduğu görülüyor.



Şekil 1.3.1.3 State accident

**1.3.2 Eyaletlere Göre Kaza Sayıları**



Şekil 1.3.2.1 State accident graph

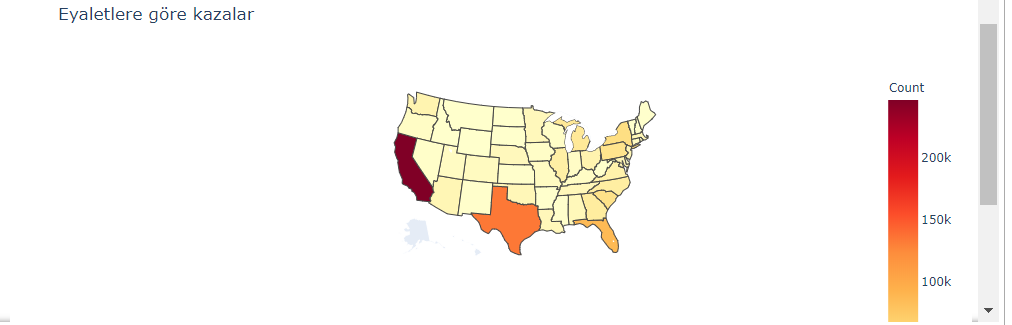
Verilere göre, California en çok trafik kazasının yaşandığı şehir. Şehirlere göre kaza sayıları aşağıda da görülebilir



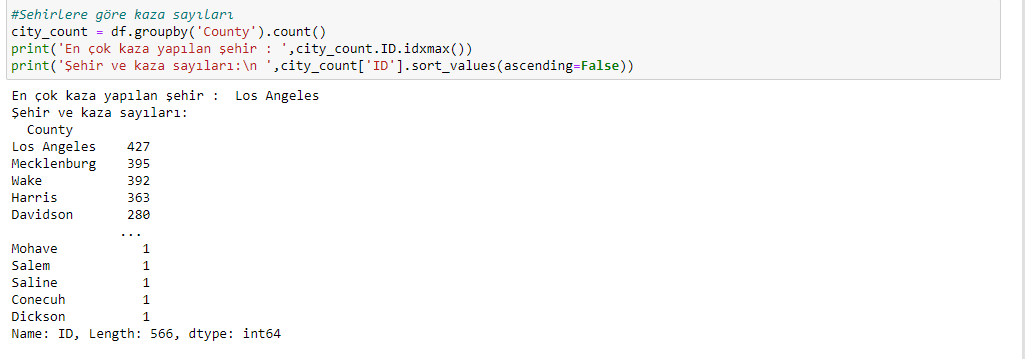
Şekil 1.3.2.2 State accident 2



Şekil 1.3.2.3 State accident interactive



Şekil 1.3.2.4 State accident interactive 2



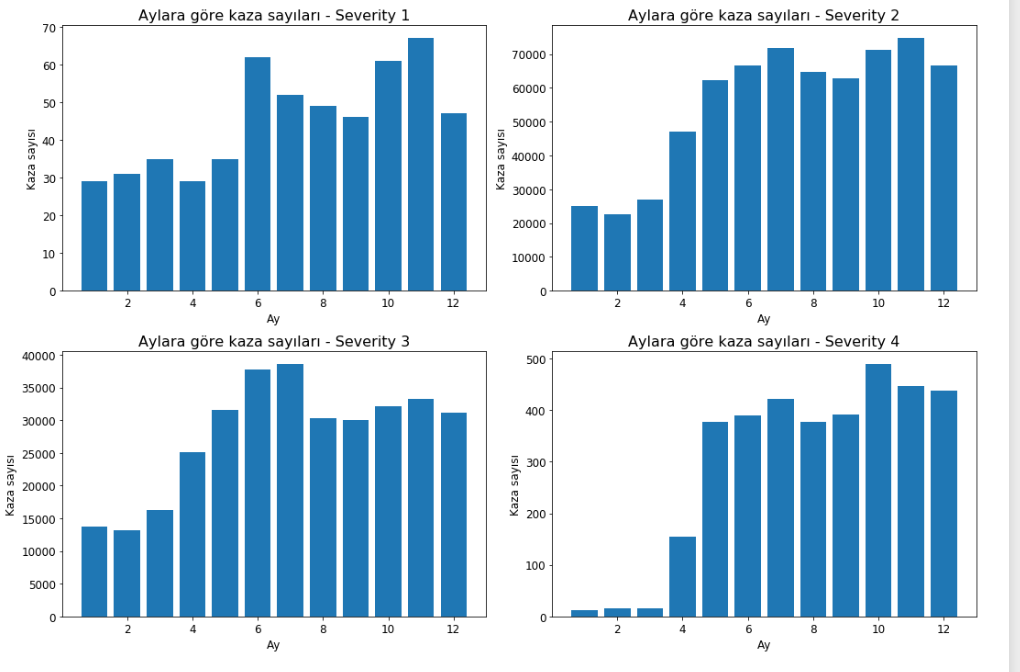
Şekil 1.3.2.5 City accident

**1.3.3 Zaman Bağlantılı Analizler**

**1.3.3.1 Aylara Göre Kaza Sayıları**



Şekil 1.3.3.1.1 monthly accidents



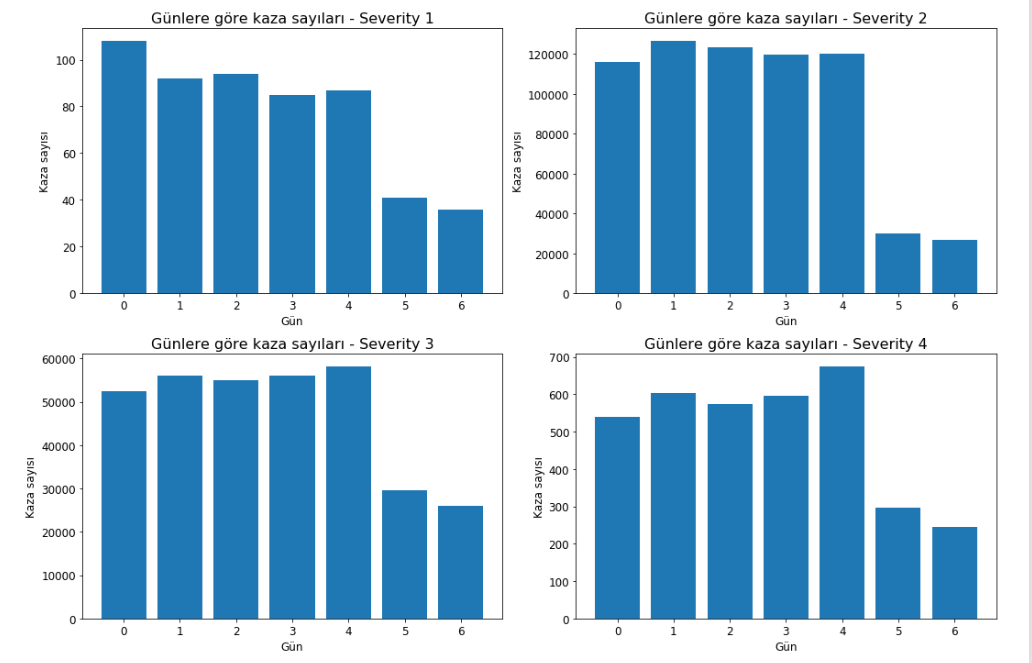
Şekil 1.3.3.1.2 monthly accidents 2

Elimizdeki verilere göre; yaz aylarının başlaması ile beraber, kaza sayılarında bir artış yaşandığı görülüyor.

**1.3.3.2 Günlere Göre Kaza Sayıları**



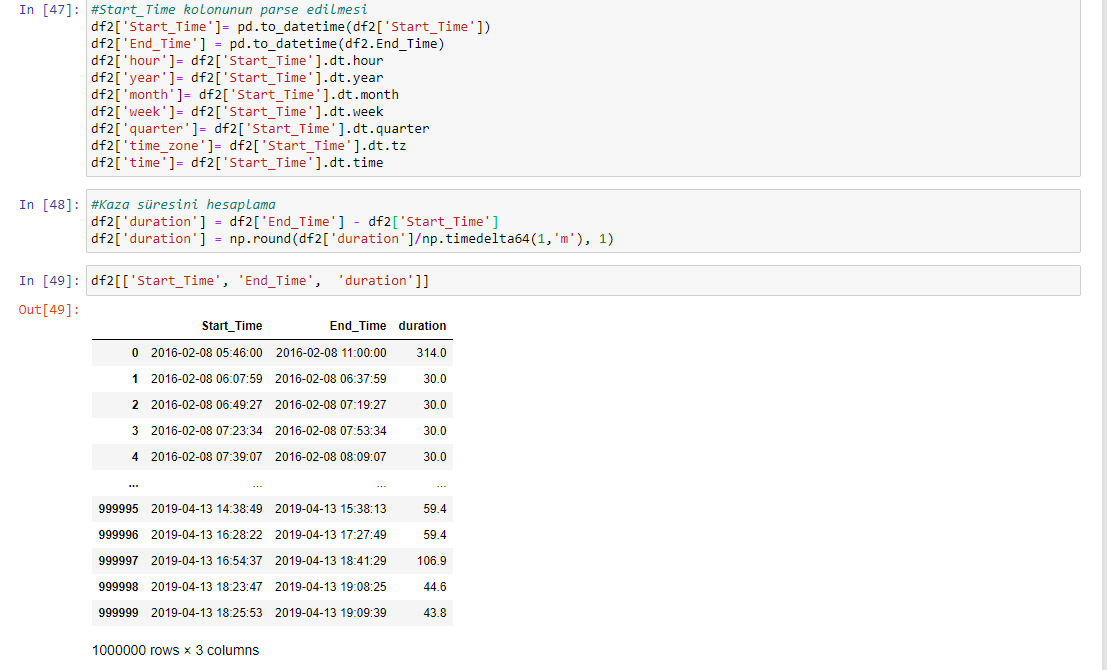
Şekil 1.3.3.2.1 daily accidents



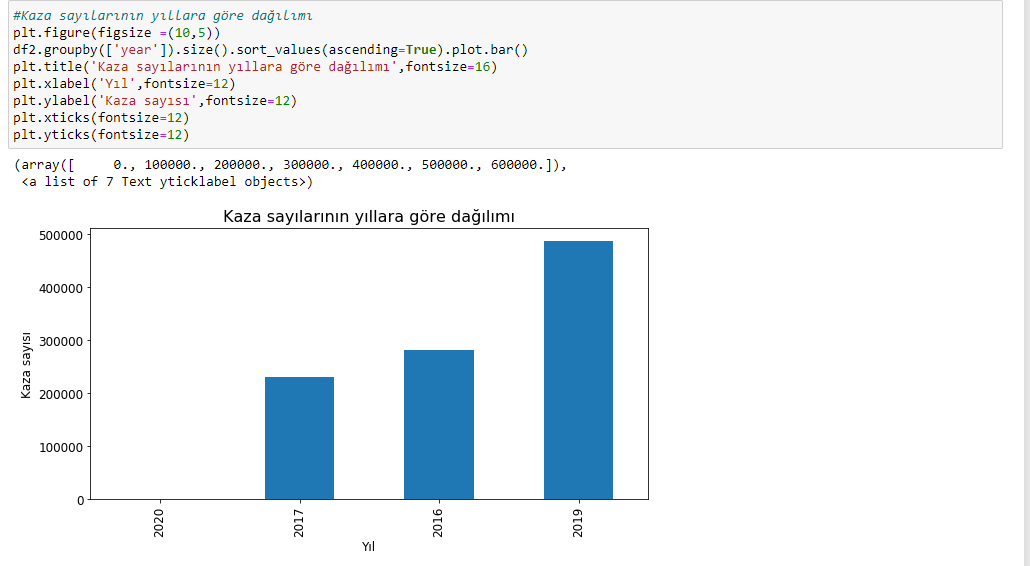
Şekil 1.3.3.2.2 daily accidents 2

Verilere göre, hafta sonları kaza sayılarında önemli bir düşüş görülüyor.

**1.3.3.3 Yıllara Göre Kaza Analizi**



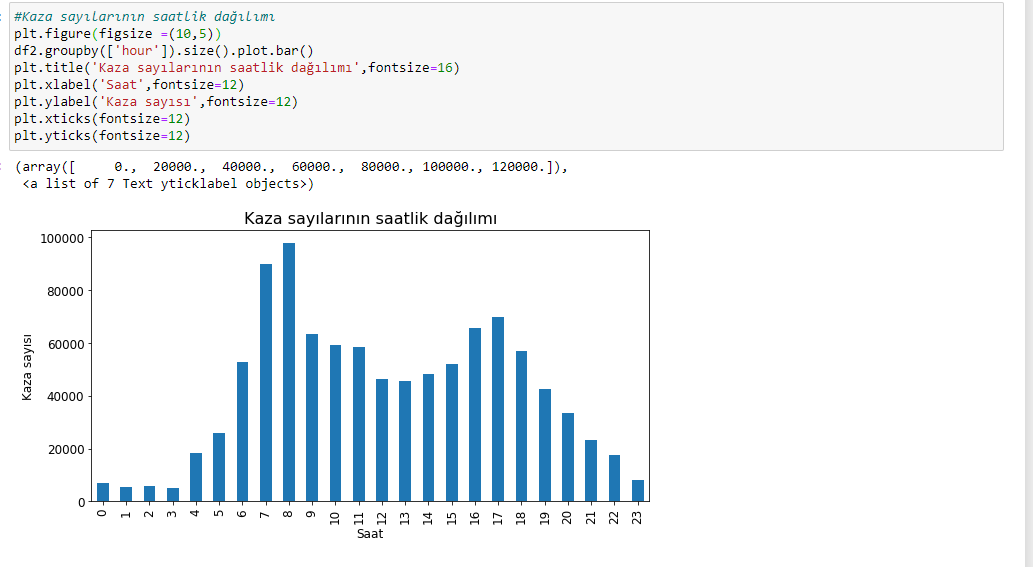
Şekil 1.3.3.3.1 yearly 1



Şekil 1.3.3.3.2 yearly 2

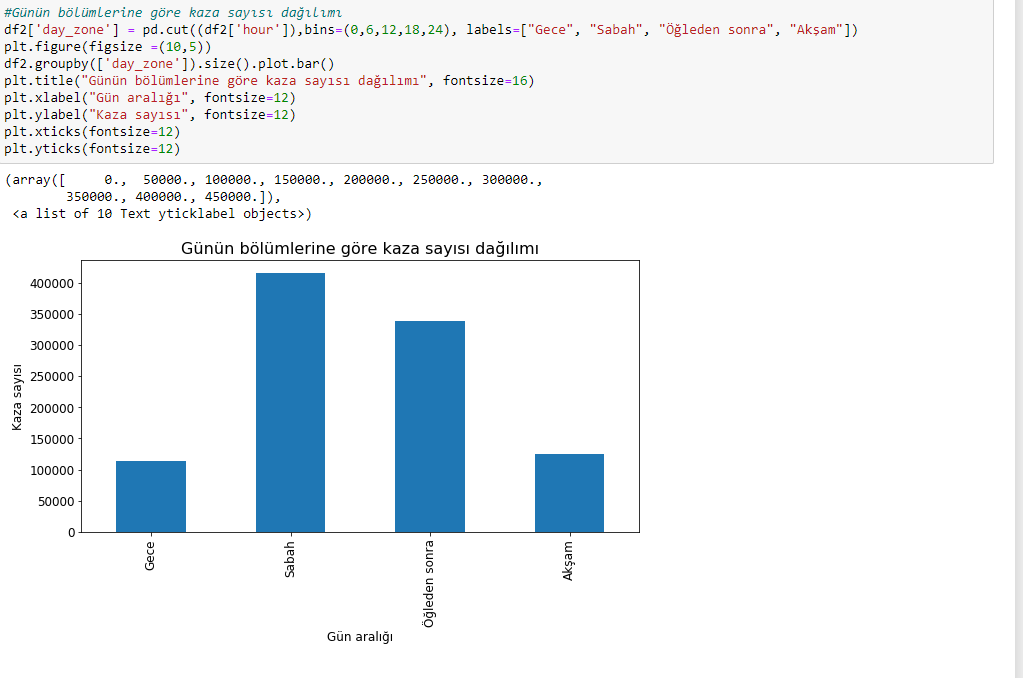
Kaza sayılarının 2019'da arttığı görülüyor.

# 1.3.3.4 Kazaların Saatlik Dağılımı



Şekil 1.3.3.4.1 hourly

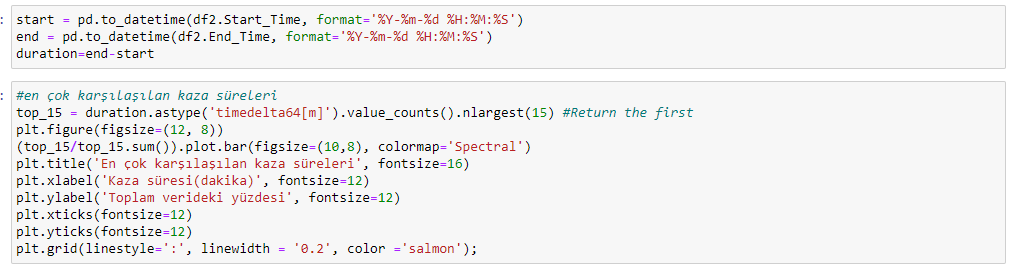
Kazaların en çok mesai başlangıcında işe gidiş sırasında yapıldığı görülüyor.



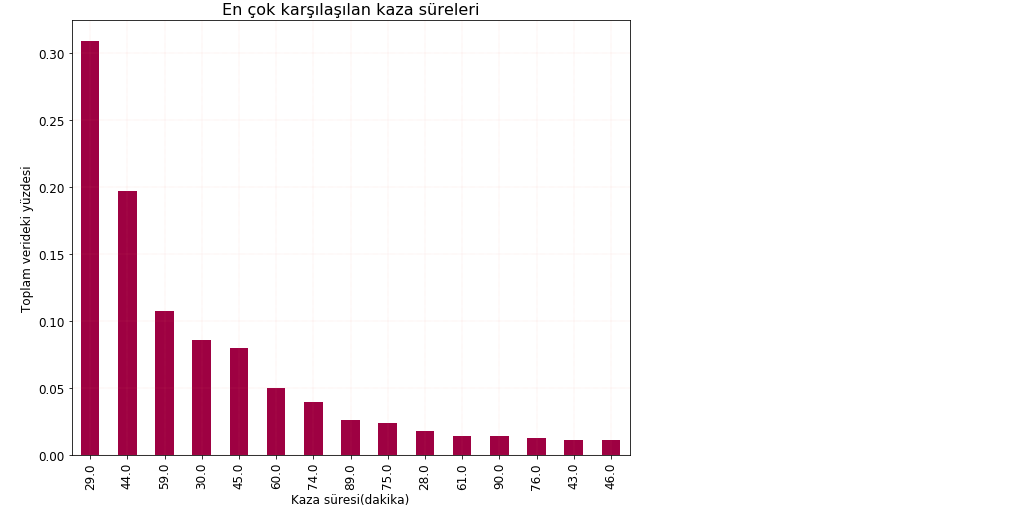
Şekil 1.3.3.4.2 day part

Grafiğe göre, kazaların genellikle 06:00 - 18:00 saatleri arasında yapıldığı görülüyor.

**1.3.3.5 Trafik Kazası Olaylarının Aldığı Sürelerin Analizi**



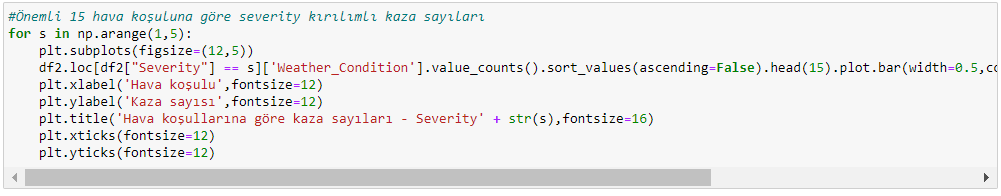
Şekil 1.3.3.5.1 time



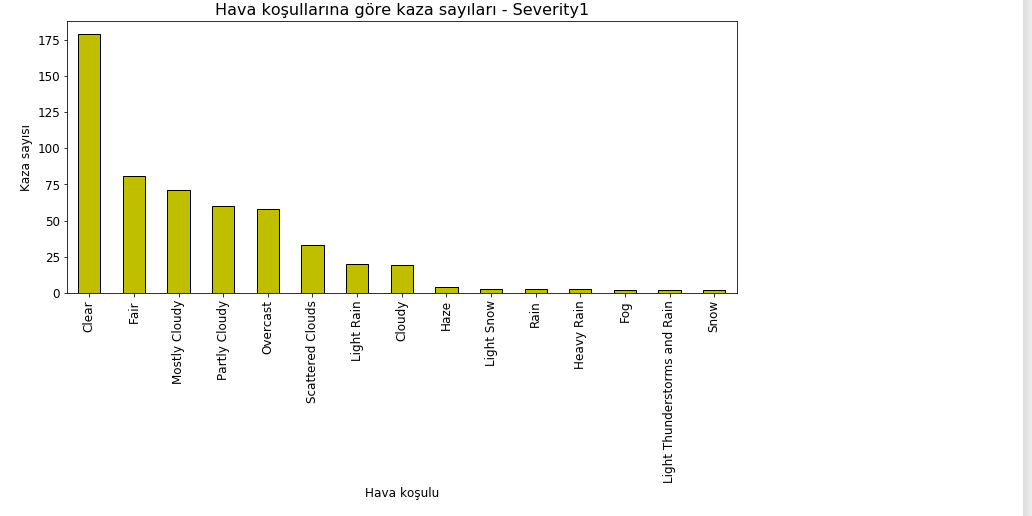
Şekil 1.3.3.5.2 time 2

**1.3.4 Hava Koşulları ile İlgili Analizler**

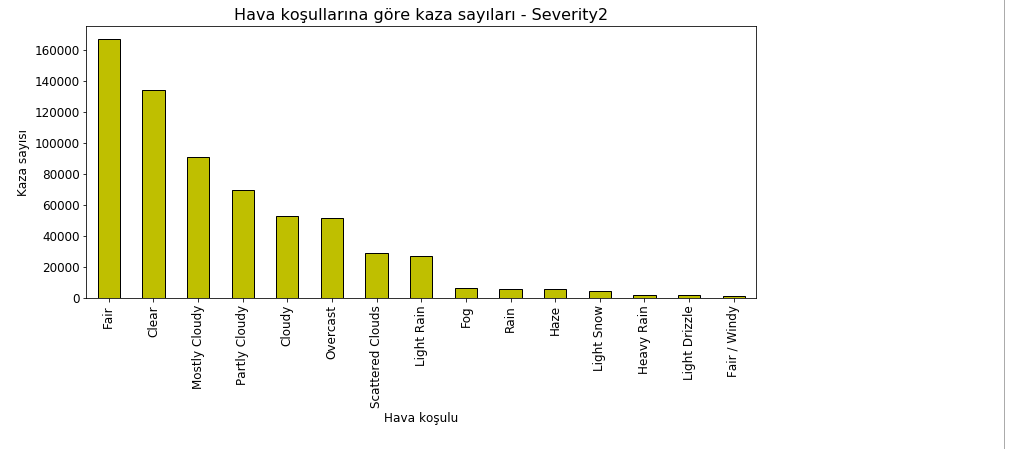
**1.3.4.1 Hava Koşullarına Göre Kaza Analizi 1**



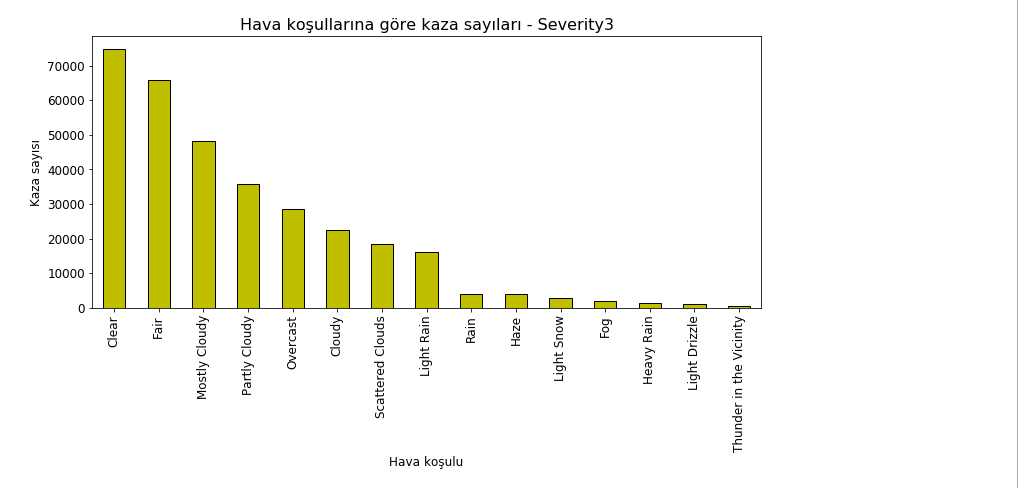
Şekil 1.3.4.1.1 weather 1



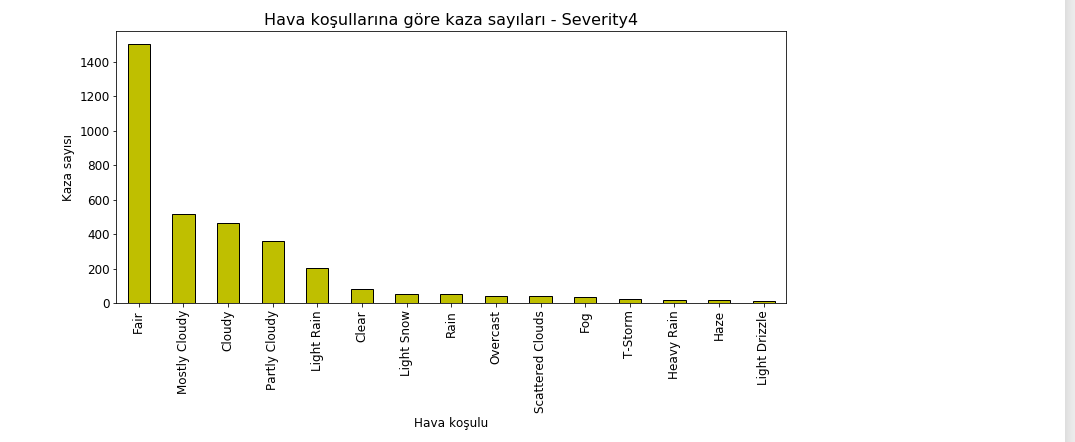
Şekil 1.3.4.1.2 weather 2



Şekil 1.3.4.1.3 weather 3



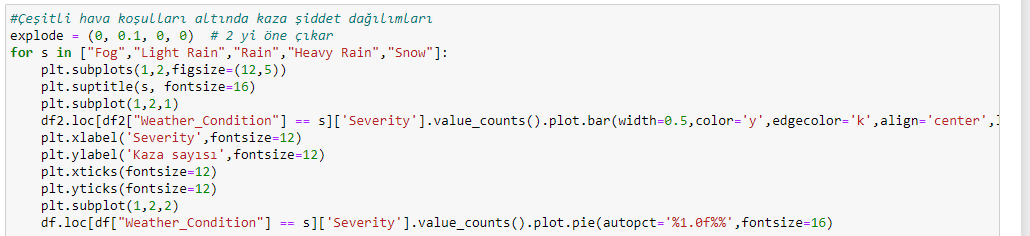
Şekil 1.3.4.1.4 weather 4



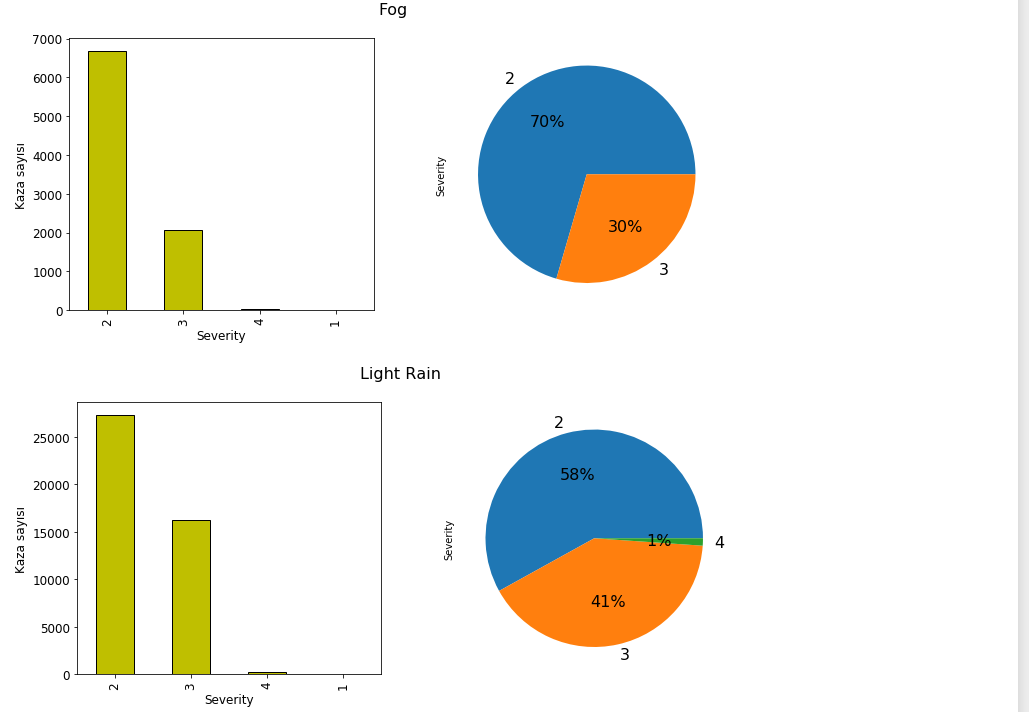
Şekil 1.3.4.1.5 weather 5

Görüldüğü gibi, kazaların büyük çoğunluğu açık veya bulutlu havalarda meydana gelmiş. Bu hava koşulları yağmur veya kar ile karşılaştırıldığında daha tehlikesizler, ancak en sık karşılaşılan hava koşulları bunlar. Hafif yağmur ve hafif kar durumlarında ise kaza sayıları yağmur / kara göre daha yüksek, bunun nedeni ise, yolun daha kaygan olması, ancak buna rağmen sürücülerin pek dikkatli olmamaları diye düşünülebilir.

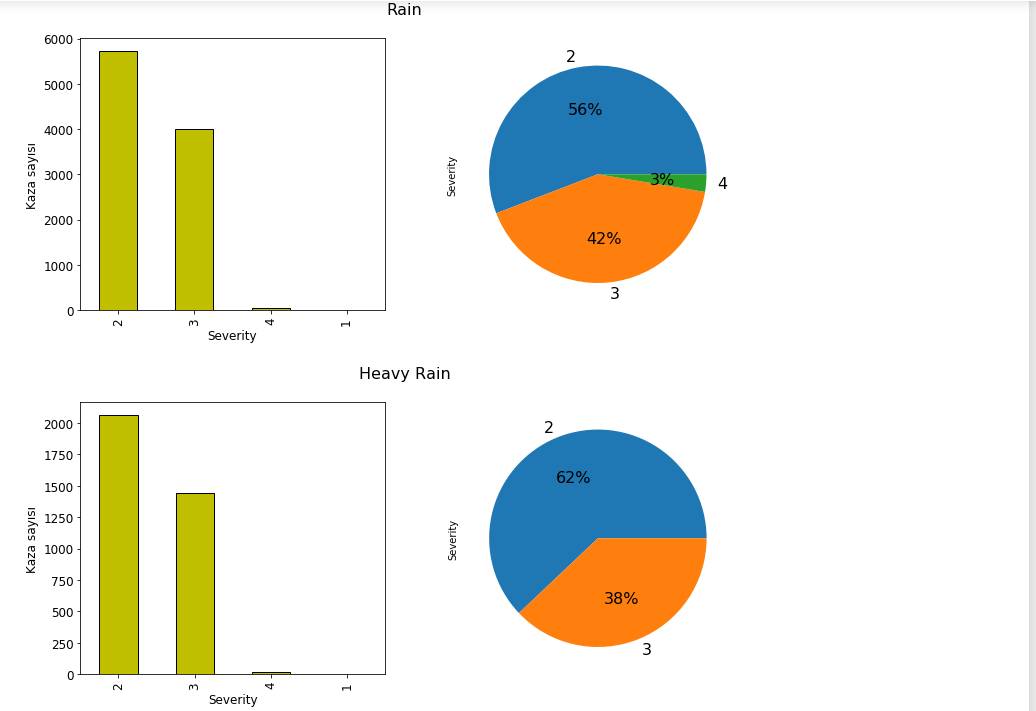
**1.3.4.2 Hava Koşullarına Göre Kaza Analizi 2**



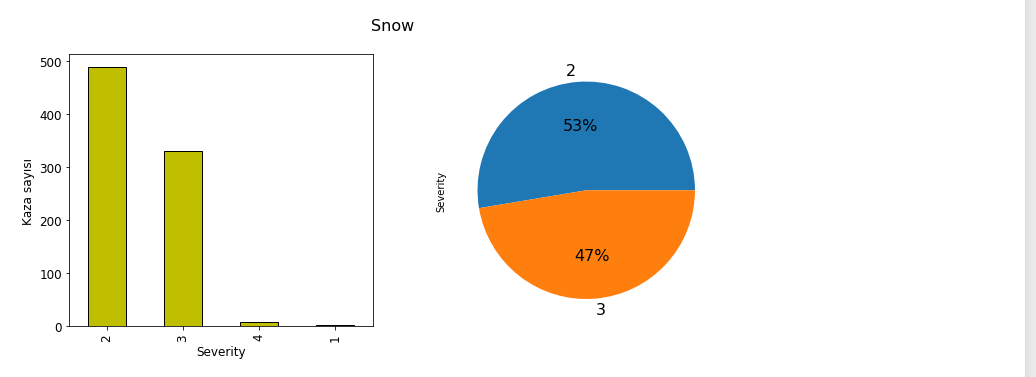
Şekil 1.3.4.2.1 weather 6



Şekil 1.3.4.2.2 weather 7



Şekil 1.3.4.2.3 weather 8

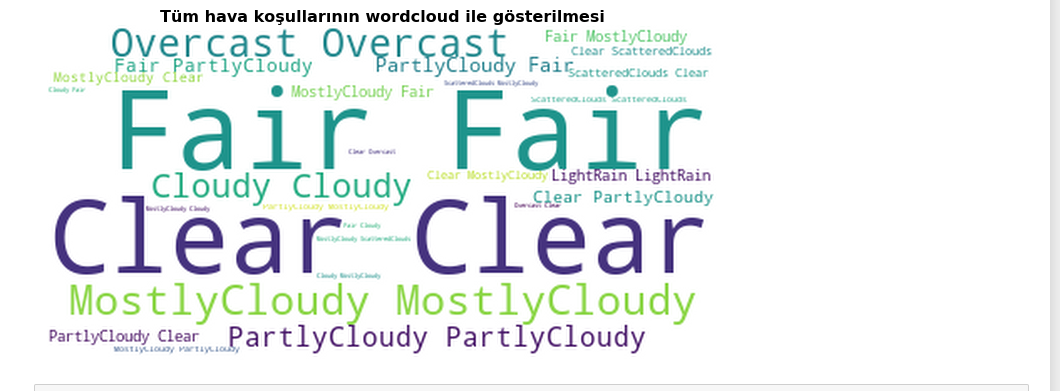


Şekil 1.3.4.2.4 weather 9

**1.3.4.3 Hava Koşullarına Göre Kaza Yoğunluğunun Gösterimi**

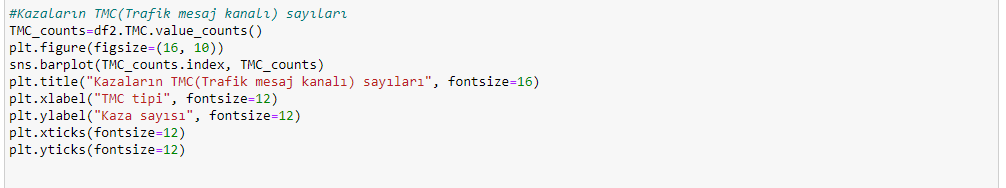


Şekil 1.3.4.3.1 weather 10

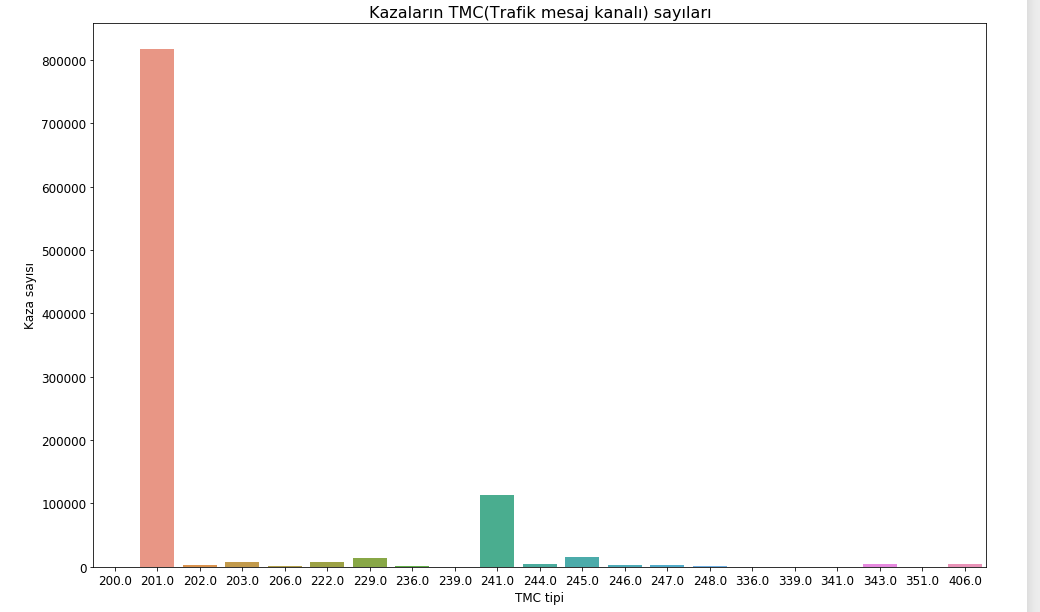


Şekil 1.3.4.3.2 weather 11

**1.3.5 Kaza Kodları Dağılımı**



Şekil 1.3.5.1 accident codes dist



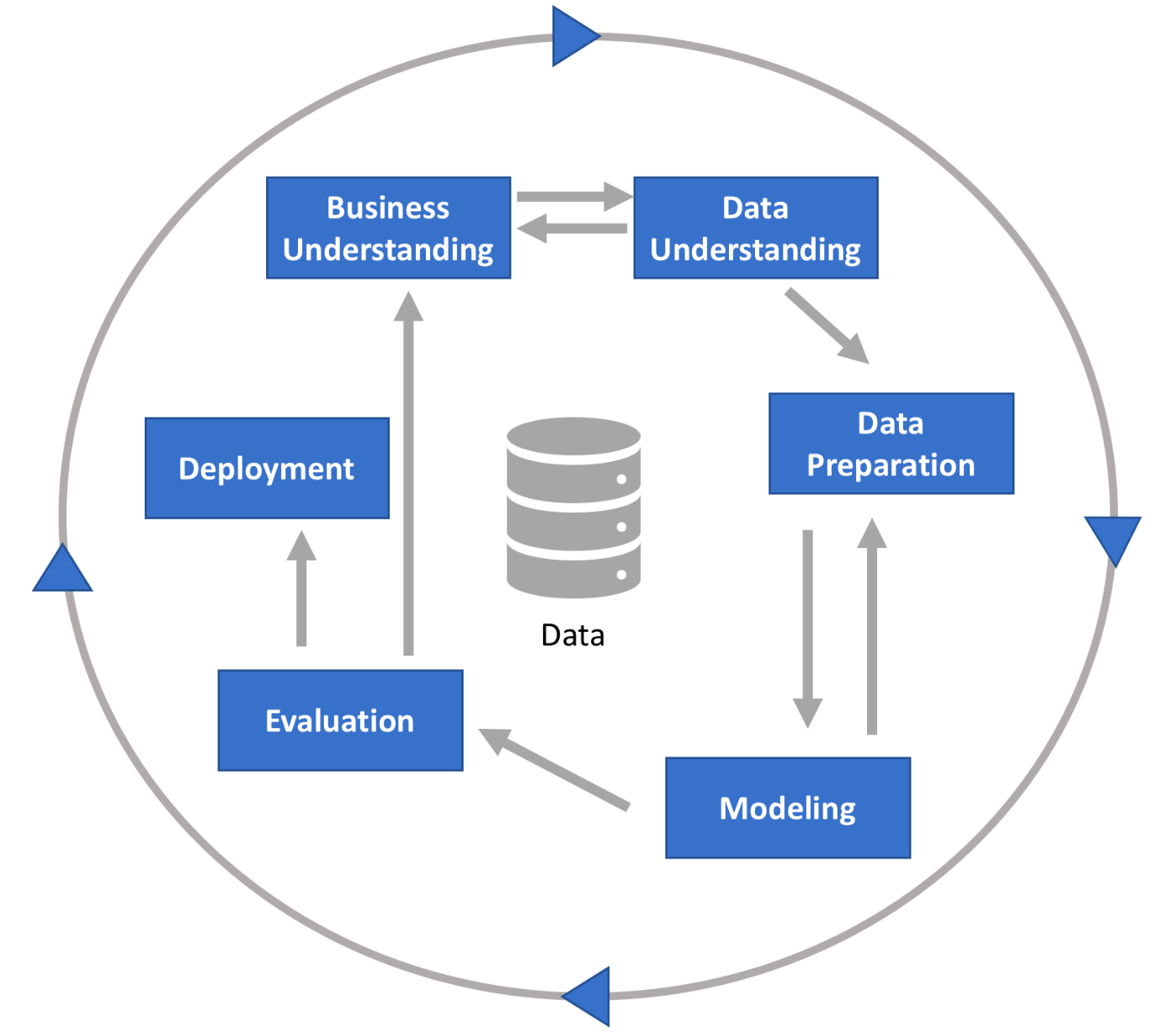
Şekil 1.3.5.2 accident codes dist graph

En çok görülen mesaj kanalının 201 yani (Q) accident olduğu görülüyor.

**2.1 CRISP-DM Metodolojisinin Aşamaları**

Projenin ikinci aşamasında CRISP-DM metodolojisi uygulanarak modelleme ve değerlendirme gerçekleştirilmiştir.

Özetle, İşi anlama, Veriyi anlama, Veriyi hazırlama, Modelleme, Değerlendirme ve Uygulama.



Şekil 2.1.1 CRISP-DM

**2.2 Verinin Tanımı & İkinci Aşamanın Amacı**

Projenin ilk aşaması olan detaylı analiz bölümünde veriyi zaten tanımış olduk. İkinci aşamada veriyi KNIME platformuna eklememiz gerekiyor. Veri KNIME’a eklendikten sonra, python’da ve KNIME node’ları ile çeşitli ön işlemlerden geçirildikten sonra birkaç makine öğrenmesi algoritması kullanılarak kaza şiddeti tahmin edilmeye çalışılacak ve sonuçlar değerlendirilecektir.

Not 1: KNIME platformunda python kullanabilmek için python extension’ı eklenmelidir.

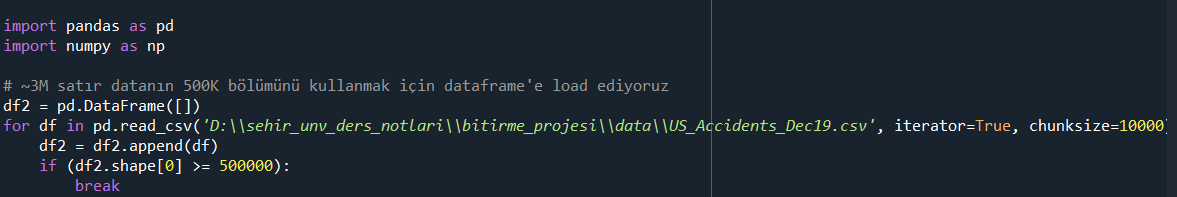
Not 2: Workflow’un tamamı Bitirme\_projesi\_v2\_Mert\_Acikgoz.knwf dosyasında mevcuttur.

**2.3 Verinin Okunması ve Birleştirilmesi**

KNIME analitik platformunda veriyi excel dosyaları ile okumamız gerekiyor. Veri aşağıdaki adresten kişisel bilgisayara indirilmeli: (csv formatında)

<https://drive.google.com/open?id=1Jv7yHBumRavNIOLootzvvkKQ_WqoA8OT>

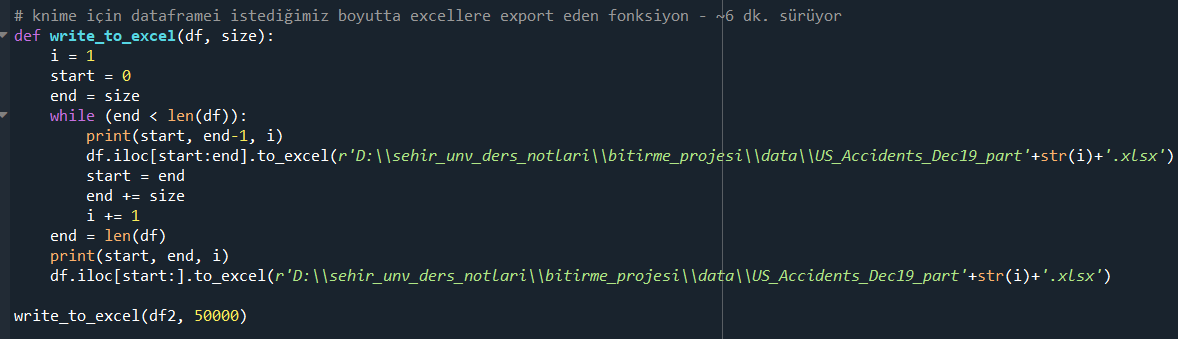
Veri 1 GB boyutunda ve 3M satırdan oluşuyor. Bu noktada KNIME analitik platformunun mümkün olduğu kadar hızlı çalışabilmesi için python’da çalışacak aşağıdaki kod ile verinin 500 K bölümünü önce bir dataframe’e import etmemiz gerekiyor. (Kod bitirme\_318631062.py dosyasında mevcuttur.)



Şekil 2.3.1 data read

Not: csv dosyası bilgisayarınızdaki path ile değiştirilmelidir.

Sonrasında, excel programının sınırları nedeniyle dosyayı 10 adet 50K dosyaya bölecek aşağıdaki kod çalıştırılmalı. (Kod bitirme\_318631062.py dosyasında mevcuttur.)

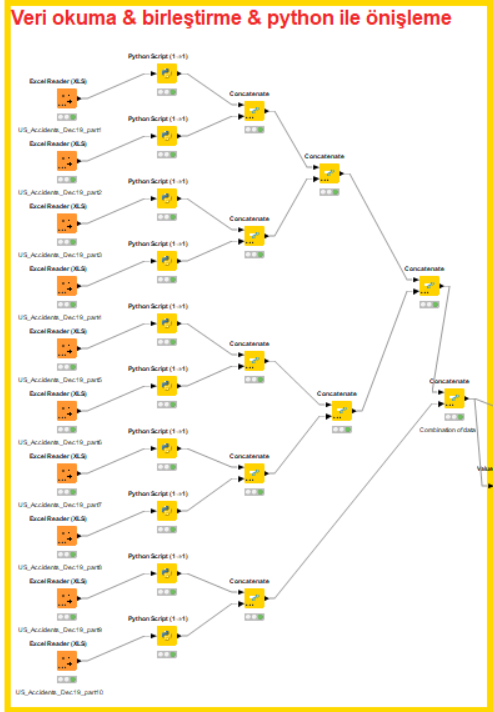


Şekil 2.3.2 data write to excel

Not: excel dosyalarının nerede oluşturulması isteniyor ise dosya path’i o şekilde güncellenmelidir.

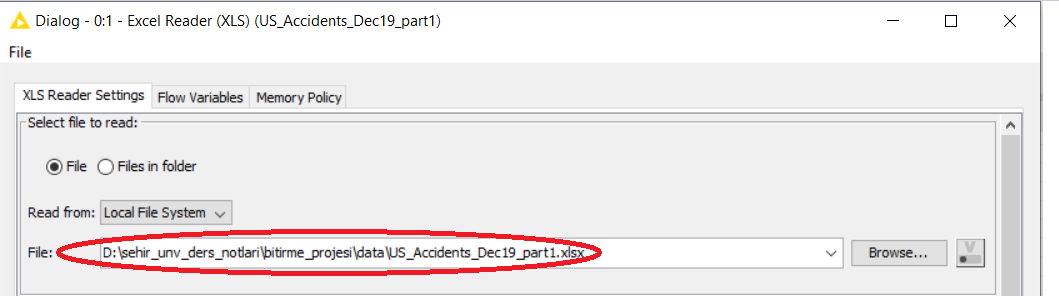
Bu noktada elimizde 10 adet 50K satırlık excel dosyası mevcut.

Veri KNIME’da aşağıdaki şekilde belirtildiği gibi excel dosyalarına atılıyor.



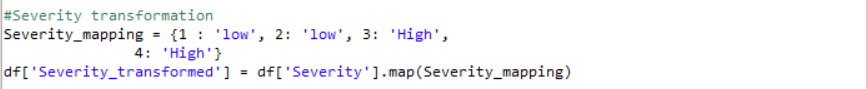
Şekil 2.3.3 data read and concatanation

Excel reader node’larının konfigürasyonunda aşağıda belirtildiği gibi bilgisayardaki path’ler değiştirilmelidir:



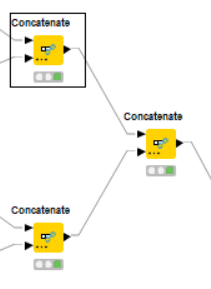
Şekil 2.3.4 data concatanation conf

Python Script node’larının içeriği proje ile gönderilecek KNIME workflow dosyasından alınabilecektir. Python ile yapılan önemli işlemlerden birisi Severity kolonunun low ve high olarak kategorize edildiği aşağıdaki işlemdir:



Şekil 2.3.5 python mapping

Sonrasında Concatanate node’ları ile veri birleştiriliyor.

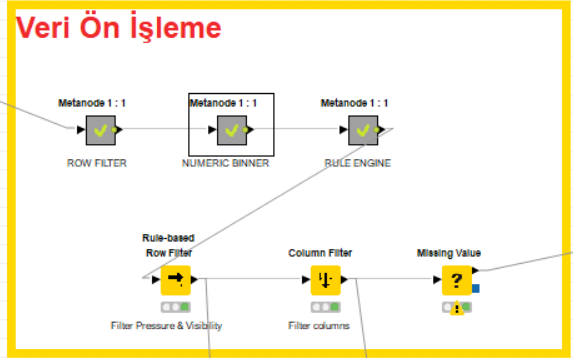


Şekil 2.3.6 KNIME concatanation

**2.4 Veri Ön İşleme**

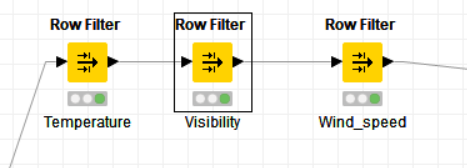
Birleştirilen veri numeric binner, rule engine, row filter meta node’ları içinde, column filter, rule based row filter ve missing value node’ları ile çeşitli

ön işlemlerden geçiriliyor.



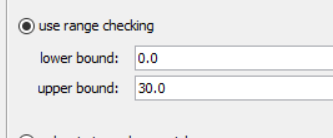
Şekil 2.4.1 KNIME meta node

Row filter meta node’u içerisinde çeşitli kolonlardaki outlier değerlerin alınmaması sağlanıyor:



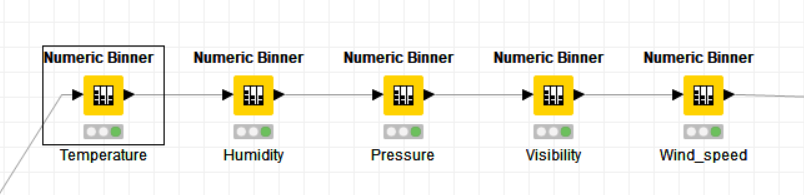
Şekil 2.4.2 KNIME row filter

Örneğin wind speed bağımsız değişkeni:



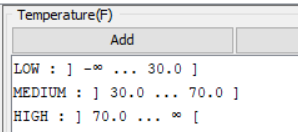
Şekil 2.4.3 KNIME row filter conf

Numeric binner meta node’u içerisinde çeşitli kolonlar kategorize ediliyor:



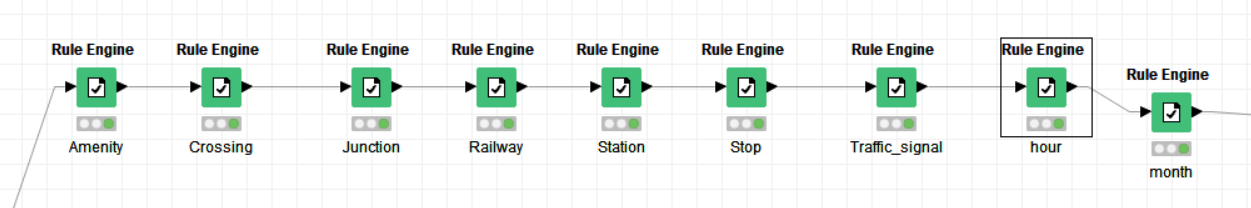
Şekil 2.4.4 KNIME numeric binner

Örneğin temperature bağımsız değişkeni:



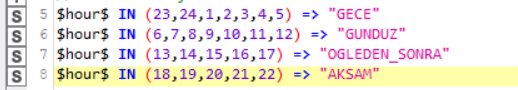
Şekil 2.4.5 KNIME numeric binner conf

Rule engine meta node’u ile bazı kolonlar için belirli kurallar koyuluyor ve bu kurallar neticesinde kategorize ediliyorlar:



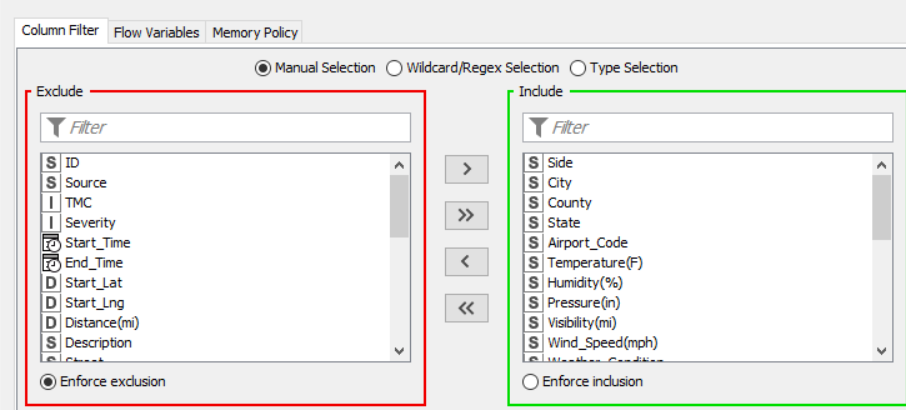
Şekil 2.4.6 KNIME rule engine

Örneğin hour bağımsız değişkeni:



Şekil 2.4.7 KNIME rule engine conf

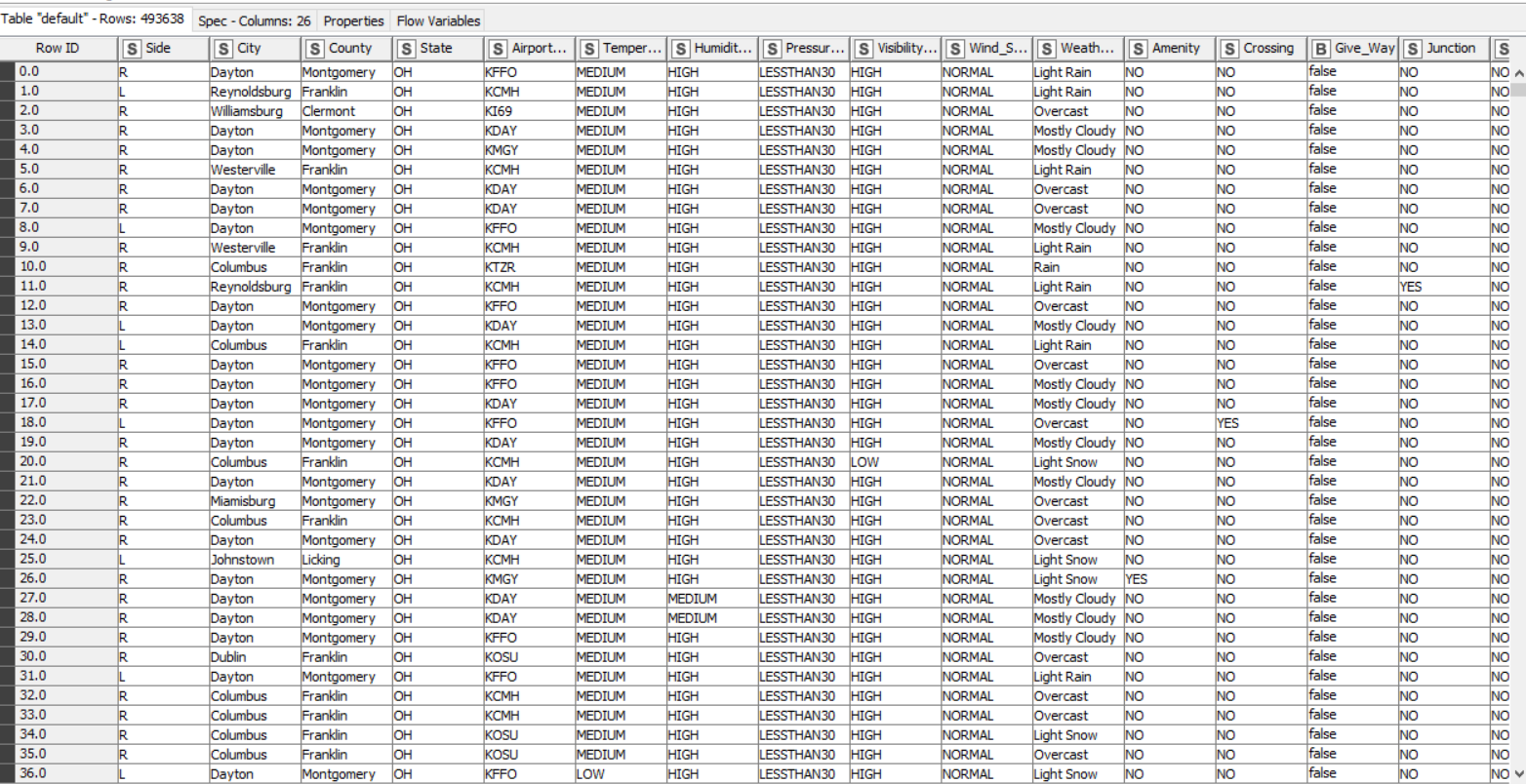
Column filter node’u ile bazı bağımsız değişkenler model harici bırakılıyor:



Şekil 2.4.8 KNIME column filter conf

Missing value node’u ile de, öncesinde zaten python ile eksik değerler tamamlandığı ve gerekli ön işlemler uygulandığı için, kalan satırlar siliniyor.

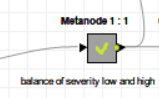
Bu noktada aşağıdaki gibi 493.638 satır ve 26 sütunumuz mevcut.



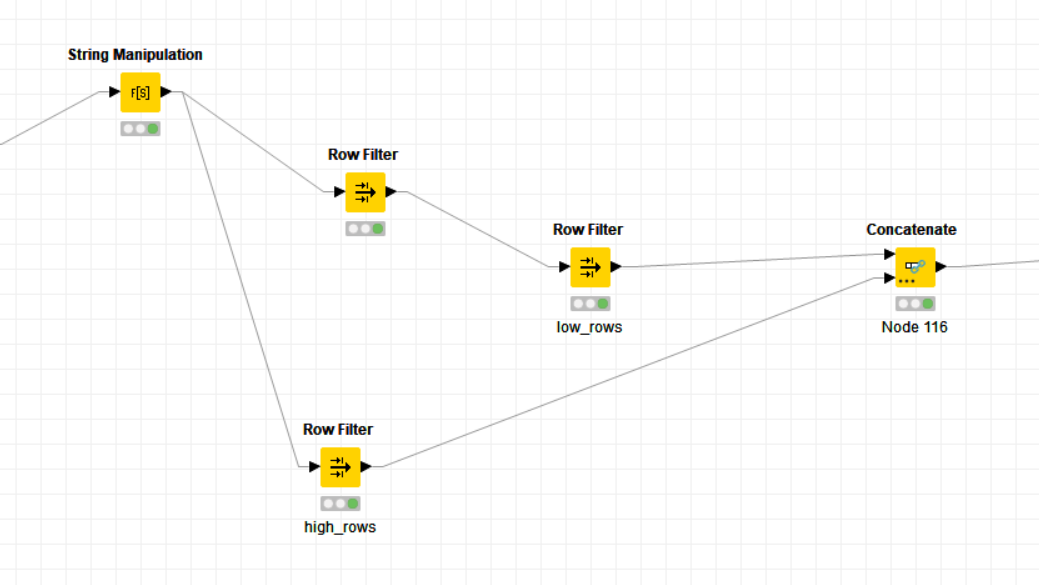
Şekil 2.4.9 KNIME table output

**2.5 Modeldeki Bağlı Değişken Kayıt Sayılarının Dengelenmesi**

Modelde kullanacağımız / tahmin etmeye çalışacağımız bağlı değişkenin (Severity\_transformed) verimizde her iki sınıftan da (low ve high) eşit sayıda olmasını istiyoruz, bu nedenle düşük sayıda kayıdı olan sınıf için SMOTE ile sentetik veri oluşturmak amaçlandı ilk olarak, ancak kayıt sayısı fazla olduğu ve işlemin sonlanmadığı görüldüğü için, her iki sınıfa ait kayıt sayısının da yaklaşık olarak aynı olmasını sağlamak amacıyla aşağıdaki meta node içindeki işlemler uygulandı.



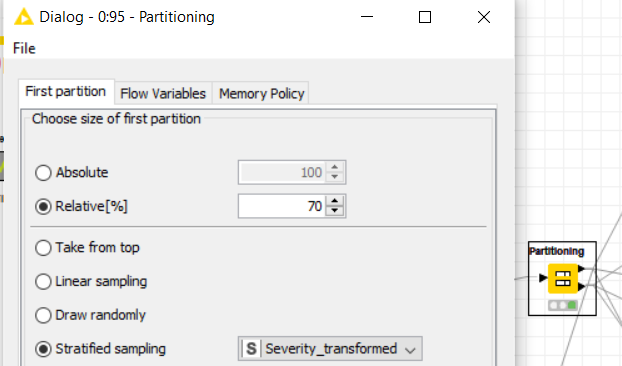
Şekil 2.5.1 KNIME balance meta node



Şekil 2.5.2 KNIME balance meta node content

**2.6 Verinin Bölünmesi**

Veriyi aşağıdaki gibi %70 eğitim ve %30 test olmak üzere iki bölüme ayırıyoruz.

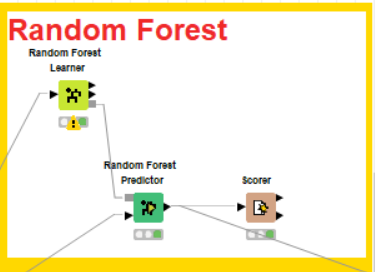


Şekil 2.6.1 KNIME partition conf

**2.7 Modelleme**

Tahminleme üç farklı model ile gerçekleştirildi. Bölünen verinin eğitim bölümü ile modeller eğitilip, test bölümü ile de tahminleme yapıldı:

**2.7.1 Random Forest**

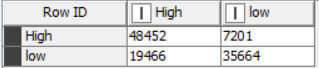


Şekil 2.7.1.1 KNIME random forest

Eğitim bölümünde, ağaç kriterlerinde üç farklı opsiyon bulunuyor:

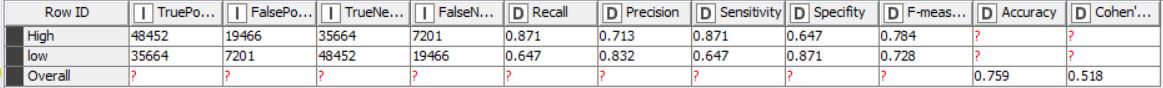
Gini index, information gain ve information gain ratio. Bu opsiyonlardan en verimlisi gini index olduğundan bu opsiyon kullanıldı.

Confusion matrix aşağıdaki gibi oluştu:



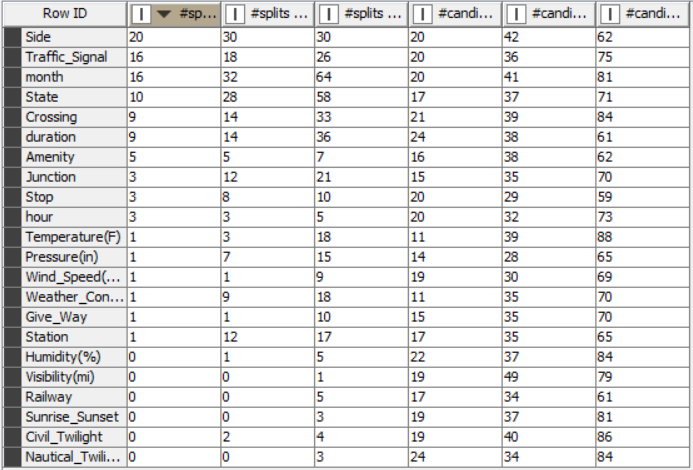
Şekil 2.7.1.2 KNIME random forest confusion matrix

Accuracy değerleri aşağıdaki gibi oluştu:



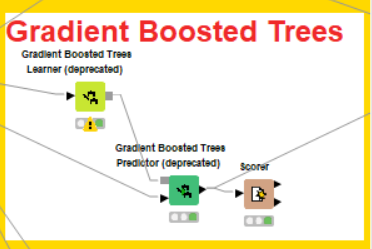
Şekil 2.7.1.3 KNIME random forest accuracy statistics

Bu model için; değişkenlerin model oluşumundaki önemleri ve ağaç oluşturulurken değişkenlerin hangi seviyelerde ve hangi sıklıkla kullanıldıklarını gösteren istatistik verileri aşağıdaki gibi oluştu.



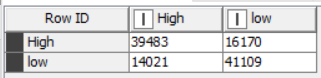
Şekil 2.7.1.4 KNIME random forest attribute statistics

**2.7.2 Gradient Boosted Trees**



Şekil 2.7.2.1 Gradient Boosted Trees

Confusion matrix aşağıdaki gibi oluştu:



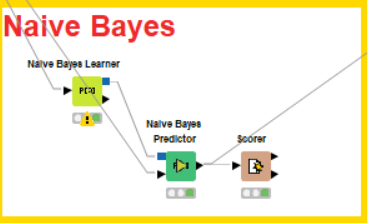
Şekil 2.7.2.2 Gradient Boosted Trees confusion matrix

Accuracy değerleri aşağıdaki gibi oluştu:

****

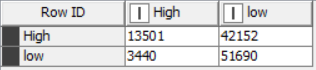
Şekil 2.7.2.3 Gradient Boosted Trees accuracy statistics

**2.7.3 Naive Bayes**



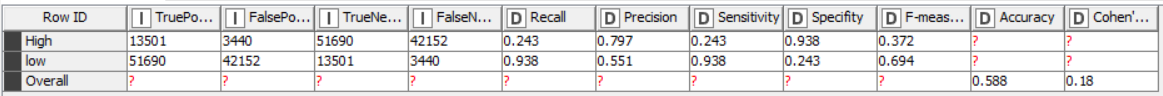
Şekil 2.7.3.1 Naive Bayes

Confusion matrix aşağıdaki gibi oluştu:



Şekil 2.7.3.2 Naive Bayes confusion matrix

Accuracy değerleri aşağıdaki gibi oluştu:

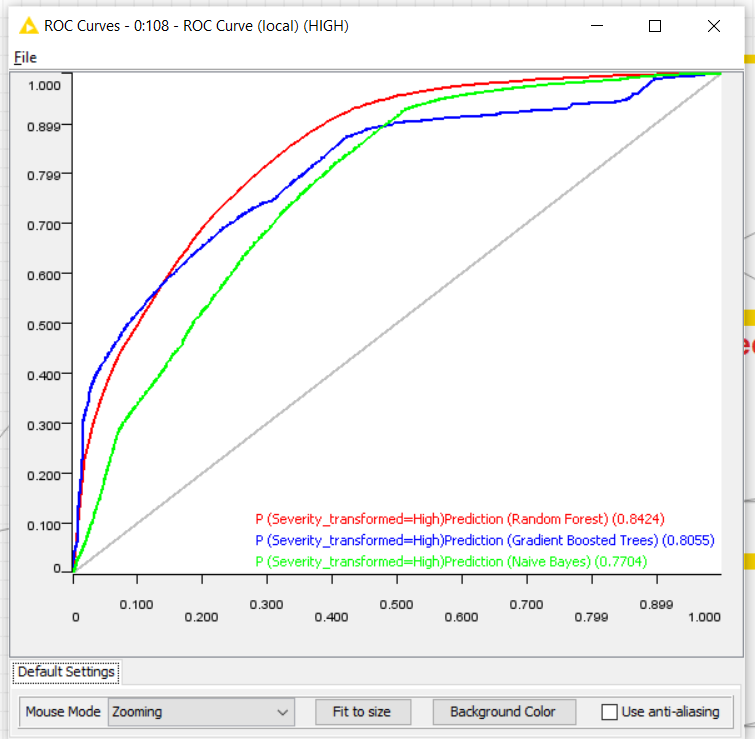
****

Şekil 2.7.3.3 Naive Bayes accuracy statistics

**2.8 Model Değerlendirme**

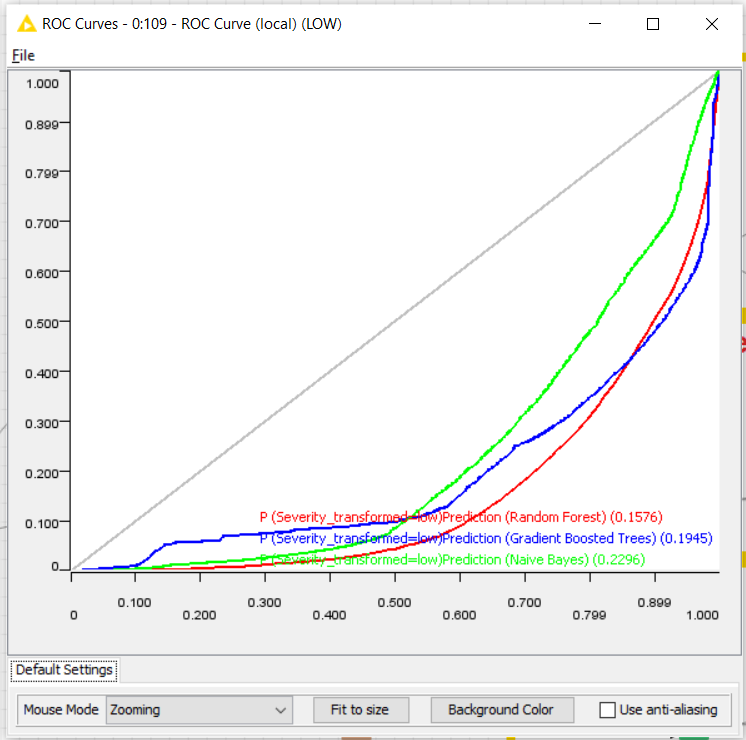
Kullanılan modellerde en iyi sonucu **Random Forest** modelinin verdiği (~%76 accuracy) , ancak bununla beraber **Gradient Boosted** modelinin de yakın sonuç verdiği görüldü. Performansı en düşük model ise **Naive Bayes** oldu.

**2.8.1 ROC Curve High**

****

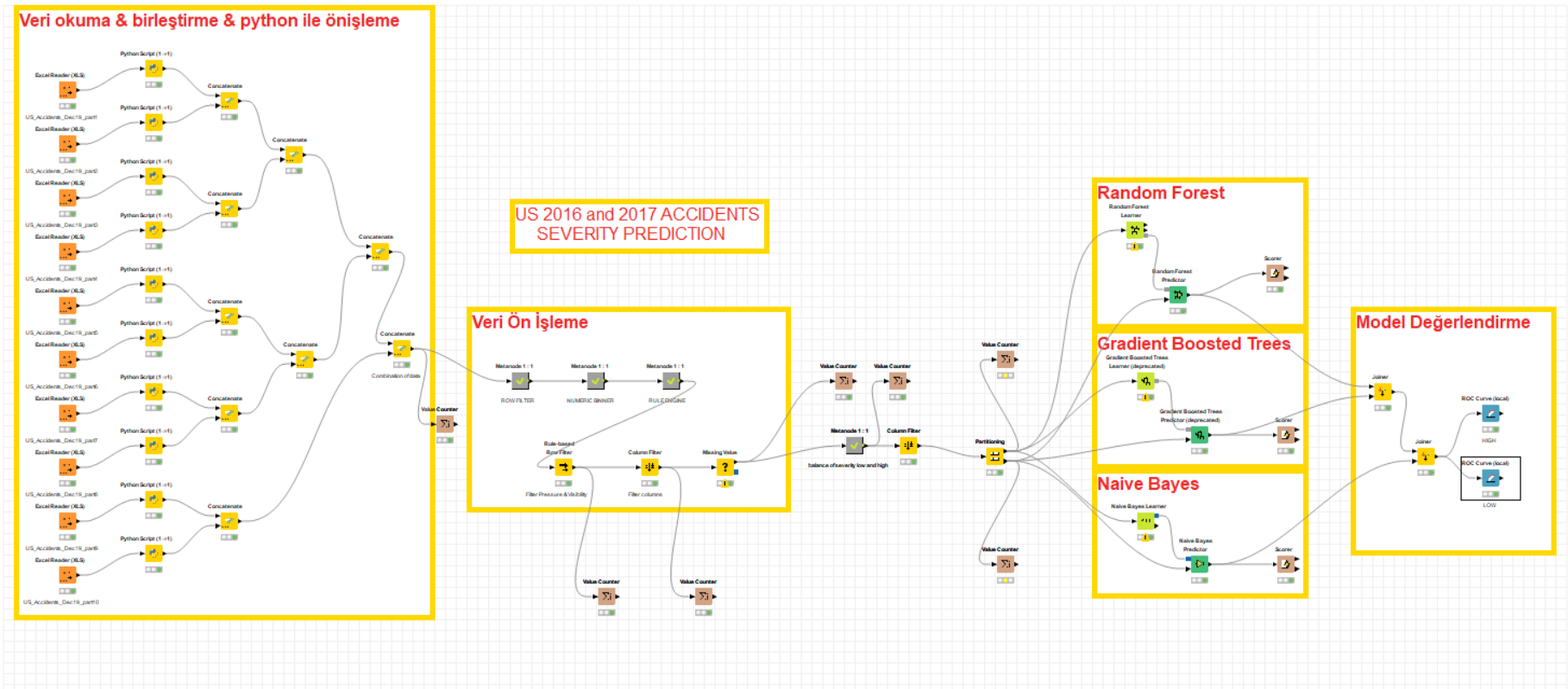
Şekil 2.8.1.1 KNIME ROC Curve High

**2.8.2 ROC Curve Low**

****

Şekil 2.8.2.1 KNIME ROC Curve Low

**2.9 KNIME Workflow**

****

Şekil 2.9.1 KNIME workflow

**3. Kaynakça**

<https://www.kaggle.com/sobhanmoosavi/us-accidents>

**ÖZGEÇMİŞ**

Kişisel Bilgiler:

Adı, Soyadı: Mert Açıkgöz

E-mail: [mert.acikgoz@ymail.com](mailto:mert.acikgoz@ymail.com)

Web Sitesi: [www.mertacikgoz.com](http://www.mertacikgoz.com)

Eğitim:

**Yüksek Lisans:** Yeditepe Üniversitesi, MBA, İstanbul **09.2007 – 07.2010**

**Lisans:** Yıldız Teknik Üniversitesi, Matematik Mühendisliği, İstanbul **09.1998 – 09.2003**

**Lise:** Kdz. Ereğli Anadolu Lisesi,Matematik, Zonguldak **1991 - 1998**

İş Deneyimi:

**01.2018 - 08.2018** Sözcü Gazetesi

**03.2013 - 05.2016** PIA (People in Action)

**12.2010 - 03.2013** Vodafone

**02.2007 - 03.2010** Turkcell

**10.2005 - 01.2007** Experteam Consulting

**04.2005 - 10.2005** Class Software & Consulting

**10.2004 - 04.2005** Aras Kargo

**07.2001 - 06.2002** Time Communications