تحلیل تصادفات جاده ای

دانشجو : ساجده عکاف

استاد راهنما : دکترمریم کلکین نما

بهمن ماه 1403

تقدیم به :

بازماندگان تصادفات جاده ای.

چکیده :

در این پروژه، تلاش شده است تا با هدف شناسایی عواملی که روی طول جاده‌های متاثر از تصادفات تاثیر دارن و کمک به بهبود ایمنی جاده‌ها اجرا شود .این دادگان شامل اطلاعات بیش از 246000 رکورد از تصادفات در 47ستون میباشد که از سایت kaggel برداشته شده است.
هدف این پروژه شناسایی عواملی که در جاده ها بر تصادفات تاثیرگزار هستند و کمک به بهبود ایمنی جاده‌ها است.

شرح ستون های ویژگی این دیتاست عبارت است از؛

IDشناسه

Sourceمنبع

Severityشدت

Start_Timeزمان شروع

** End_Time**زمان پایان

Start_Latعرض جغرافیایی شروع

Start_Lngطول جغرافیایی شروع

End_Latعرض جغرافیایی پایان

End_Lngطول جغرافیایی پایان

Distance(mi)فاصله (مایل)

Descriptionتوضیحات

Streetخیابان

Cityشهر

Countyشهرستان

Stateایالت

Zipcodeکد پستی

Countryکشور

Timezoneمنطقه زمانی

Airport_Codeکد فرودگاه

Weather_Timestampزمان ثبت اطلاعات هواشناسی

Temperature_Range(F)دامنه دما (فارنهایت)

** (F)Wind_Chill**سرمازدگی باد (فارنهایت)

** (%) Humidity**رطوبت(%)

(in)Pressureفشار (اینچ جیوه)

(mi)Visibilityدید (مایل)

Wind_Directionجهت باد

(mph)Wind_Speedسرعت باد (مایل بر ساعت)

(in)Precipitationبارندگی (اینچ)

Weather_Conditionشرایط آب و هوایی

Amenityامکانات

Bumpبرآمدگی

Crossingعبور عابر پیاده

Give_Wayبدهید راه

Junctionتقاطع

No_Exitخروجی ممنوع

Railwayراه آهن

Roundaboutدوربرگردان

Stationایستگاه

Stopتوقف

Traffic_Calmingکاهش سرعت ترافیک

Traffic_Signalچراغ راهنمایی

Turning_Loopحلقه چرخش

Sunrise_Sunsetطلوع و غروب خورشید

Civil_Twilightنیمه روشنایی

Nautical_Twilightنیمه دریایی

Astronomical_Twilightنیمه نجومی

فصل اول ، درک داده ها :

1-1 مقدمه:

دیتاست ارائه شده حاوی اطلاعات ارزشمندی در مورد 460,000 تصادف جاده‌ای در آمریکا است. با تحلیل دقیق این ستون‌ها می‌توان به بینش‌های عمیقی در مورد دلایل، مکان‌ها و الگوهای تصادفات دست یافت. در این تحلیل، به توضیح هر ستون و شناسایی احتمالی خطاها و نویز در داده‌ها خواهیم پرداخت.

2-1 شناسایی داده ها :

تحلیل ساختار داده‌های تصادفات

۱ .متغیرهای کمی-عددی(Quantitative)

این متغیرها شامل مقادیر عددی (عدد صحیح یا اعشاری) هستند و می‌توانند **پیوسته (Continuous)** یا **گسسته (Discrete)**باشند.

نام متغیر	نوع عددی	توضیح
Severity	گسسته(Discrete)	شدت تصادف
Start_Lat	پیوسته(Continuous)	عرض جغرافیایی نقطه شروع
Start_Lng	پیوسته(Continuous)	طول جغرافیایی نقطه شروع
End_Lat	پیوسته(Continuous)	عرض جغرافیایی نقطه پایان
End_Lng	پیوسته(Continuous)	طول جغرافیایی نقطه پایان
Distance.mi.	پیوسته(Continuous)	فاصله تصادف بر حسب مایل
Wind_Chill.F.	پیوسته(Continuous)	دمای احساسی (درجه فارنهایت)
Humidity...	پیوسته(Continuous)	درصد رطوبت
Pressure.in.	پیوسته(Continuous)	فشار هوا (اینچ جیوه)
Visibility.mi.	پیوسته(Continuous)	میزان دید (مایل)
Wind_Speed.mph.	پیوسته(Continuous)	سرعت باد (مایل بر ساعت)
Precipitation.in.	پیوسته(Continuous)	میزان بارندگی (اینچ)

۲ .متغیرهای کیفی(Categorical)

این متغیرها شامل **متغیرهای اسمی (Nominal)** و **رتبه‌ای (Ordinal)** هستند.

الف) متغیرهای کیفی اسمی - (Nominal) بدون ترتیب خاص

این متغیرها فقط دسته‌بندی شده‌اند و ترتیب خاصی ندارند.

نام متغیر	توضیح
ID	شناسه تصادف (کد یکتا)
Source	منبع گزارش تصادف
Description	توضیحات تصادف
Street	نام خیابان
City	شهر
County	شهرستان
State	ایالت
Zipcode	کد پستی
Country	کشور
Timezone	منطقه زمانی
Airport_Code	کد فرودگاه نزدیک
Weather_Condition	وضعیت آب‌وهوا
Wind_Direction	جهت باد
Sunrise_Sunset	وضعیت روز یا شب
Civil_Twilight	مرحله‌ی گرگ‌ومیش مدنی
Nautical_Twilight	مرحله‌ی گرگ‌ومیش دریایی
Astronomical_Twilight	مرحله‌ی گرگ‌ومیش نجومی

ب) متغیرهای کیفی رتبه‌ای - (Ordinal) با ترتیب خاص

این متغیرها دارای ترتیب منطقی هستند.

نام متغیر	توضیح
Temperature_Range.F.	محدوده دما (کمینه-بیشینه) بر حسب فارنهایت

ج) متغیرهای دودویی(Binary - True/False)

این متغیرها فقط می‌توانند دو مقدار داشته باشند. (`True/False or Yes/No`)

نام متغیر	توضیح
Amenity	نزدیکی به امکانات شهری
Bump	وجود سرعت‌گیر
Crossing	وجود تقاطع عابر پیاده
Give_Way	وجود تابلو "حق تقدم"
Junction	وجود تقاطع
No_Exit	بن‌بست بودن
Railway	نزدیکی به خط راه‌آهن
Roundabout	وجود میدان
Station	نزدیکی به ایستگاه حمل‌ونقل
Stop	وجود تابلو ایست
Traffic_Calming	وجود تجهیزات کاهش سرعت
Traffic_Signal	وجود چراغ راهنمایی
Turning_Loop	وجود حلقه‌ی دوربرگردان

جمع‌بندی

- **متغیرهای کمی** :۱۲ متغیر (۵ گسسته، ۷ پیوسته)
- **متغیرهای کیفی** :
  - **اسمی (بدون ترتیب)** :۱۵ متغیر
  - **رتبه‌ای (با ترتیب)** :۱ متغیر
  - **دودویی :(True/False)** ۱۳ متغیر

این تقسیم‌بندی به ما کمک می‌کند تا روش‌های تحلیل مناسب را برای هر متغیر انتخاب کنیم.

3-1 خطا یا نویز در داده ها :

- **مقادیر غیرمنتظره(Unexpected Values)**

این دسته شامل مقادیری است که به‌صورت غیرعادی، غیرمنطقی یا خارج از دامنه‌ی معمول خود ظاهر می‌شوند، اما ممکن است از نظر فنی نامعتبر نباشند.

نام متغیر	مقدار غیرمنتظره (مثال‌ها)	توضیح
Severity	مقدار 0یا مقدار بیش از 4	شدت تصادف معمولاً بین 1تا 4است، مقدار خارج از این محدوده غیرمنتظره است.
Start_Lat / Start_Lng / End_Lat / End_Lng	نقاط جغرافیایی خارج از محدوده آمریکا	مثلاً نقاطی با عرض 90 <یا 90- >یا طول 180 <یا 180- >غیرواقعی هستند.
Distance.mi.	مقدار 0برای تصادفاتی که در بزرگراه رخ داده	اگر تصادفی گزارش شده ولی مسافت صفر است، ممکن است داده ناقص باشد.
Temperature_Range.F.	دمایی مثل 100-یا 150 فارنهایت	این مقادیر معمولاً در شرایط طبیعی آمریکا رخ نمی‌دهند.
Wind_Speed.mph.	مقدار 150 mph >	سرعت باد بیش از ۱۵۰ مایل بر ساعت غیرمعمول و احتمالاً خطای ورودی است.
Visibility.mi.	مقدار 0همراه با شرایط آب‌وهوایی مناسب	اگر دید برابر صفر باشد ولی شرایط آب‌وهوایی مناسب باشد، مشکوک است.
Precipitation.in.	مقدار 10 in >	بارش بیش از ۱۰ اینچ در یک حادثه نادر است.
Timezone	مقدار نامعتبر مثل "Unknown"	مقدار ناشناخته یا ناسازگار با مکان تصادف، نشانه‌ی خطاست.

- **مقادیر غیرقابل پذیرش(Invalid Values)**

این دسته شامل مقادیری است که به‌هیچ‌وجه نباید در مجموعه داده‌ها باشند، زیرا از نظر منطقی یا فنی نامعتبر هستند.

نام متغیر	مقدار غیرقابل پذیرش (مثال‌ها)	توضیح
ID	مقدار NULL یا تکراری	شناسه تصادف باید یکتا باشد، مقدار خالی یا تکراری مشکل‌ساز است.
Start_Time / End_Time	مقدار NULL یا ترتیب زمانی نادرست	اگر زمان پایان قبل از زمان شروع باشد، داده اشتباه است.
City / State / County	مقدار NULL یا ترکیب ناسازگار	مثلاً اگر ایالت "CA"باشد ولی شهر "New York"باشد، داده اشتباه است.
Weather_Condition	مقدار NULL یا مقدار نامعتبر مثل "Unknown"	مقدار نامشخص یا خارج از دسته‌های شناخته‌شده غیرقابل قبول است.
Wind_Direction	مقدار غیرواقعی مثل "XYZ"	مقادیر باید شامل جهت‌های معتبر مثل "N", "S", "NE", "SW"باشند.
Amenity, Bump, Crossing, ...	مقدار NULL به جای True/False	این متغیرها باید فقط True یا Falseباشند، مقدار NULLنشان‌دهنده‌ی داده ناقص است.
Zipcode	مقدار NULL یا مقدار غیرعددی	کد پستی باید عددی باشد، وجود کاراکتر غیرعددی اشتباه است.
Pressure.in.	مقدار NULL یا فشار کمتر از 0	فشار هوا نمی‌تواند منفی باشد.

- **مقادیر گمشده :**
  - مقادیر خالی در ستون‌های ضروری مانند زمان، مکان یا شدت تصادف
  - عدم تطابق بین اطلاعات مکانی (مثلاً شهر و ایالت)
- **نقص در داده‌های کیفی :**
  - املای اشتباه در نام شهرها، ایالت‌ها و خیابان‌ها
  - توصیفات مبهم یا متناقض در ستونDescription
- **تضاد در داده‌ها :**
  - اختلاف بین زمان وقوع تصادف و زمان ثبت اطلاعات آب و هوایی
  - عدم تطابق بین شدت تصادف و توصیف آن

فصل دوم، وارد کردن و آماده سازی داده ها :

در ابتدا داده ها را در محیط R فراخوانی می کنیم :

```
1  #فراخوانی داده ها
2  accidents <- read.csv("C:/Users/Saj/onedrive/Desktop/US_Accidents_March23.csv")
```

```
Data
accidents          246633 obs. of 47 variables
```

سپس یک کپی از DataFrame accidents ایجاد کرده و آن را در متغیری به نام accidents_copy ذخیره می‌کنیم. این کار به این دلیل انجام می‌شود که اگر تغییراتی در DataFrame اصلی ایجاد کنیم، کپی دست‌نخورده باقی بماند و بتوانیم به داده‌های اولیه دسترسی داشته باشیم و پنج سطر اول DataFrame را نمایش می‌دهیم تا بتوانیم یک نمای کلی از داده‌ها داشته باشیم.

```
4  # برای استفاده از تابع copy
5  install.packages("dplyr")
6  library(dplyr)
7  # ایجاد یک کپی
8  accidents_copy <- (accidents)
```

پیش پردازش داده ها :

روش پاکسازی داده‌ها که در کد پیاده‌سازی شده است، مجموعه‌ای از مراحل مختلف برای حذف داده‌های گمشده، تکراری و مدیریت داده‌های پرت است. در اینجا به تفصیل توضیح داده‌ام که هر بخش از کد چه کاری انجام می‌دهد و مزایای آن‌ها چیست:

.1حذف ستون‌های اول و آخر:

در این مرحله ابتدا به کمک کد زیر دو ستون اول و آخر را حذف می‌کنیم. دلیل حذف ستون آخر این است که بیش از 90 درصد مقادیر آن تهی هستند. تعداد سطرها و ستون‌ها را نیز محاسبه کرده و سپس این اعداد را به همراه یک متن توضیحی در کنسول چاپ می‌کنیم.

```
> # حذف ستون اول و آخر
> accidents <- accidents %>%
+   select(-1, -ncol(.))
> shape <- dim(accidents)
> print(paste0("Number of columns: ", shape[2]))
[1] "Number of columns: 45"
> print(paste0("Number of rows: ", shape[1]))
[1] "Number of rows: 246633"
```

2.بررسی تعداد سطرها و ستون‌ها قبل از پاکسازی:

- قبل از انجام هر تغییراتی، تعداد سطرها و ستون‌ها ذخیره می‌شود تا بعداً بتوان تأثیر تغییرات را ارزیابی کرد.

3.بررسی داده‌های گمشده(Missing Values)

4.بررسی تعداد مقادیر NA در هر ستون

5.حذف سطرهایی که در ستون‌های کلیدی مقدار NA دارند

مزایا:

- جلوگیری از تأثیرات منفی داده‌های گمشده بر نتایج تحلیل‌های بعدی.
- حفظ داده‌های معتبر و کامل برای تحلیل‌های دقیق‌تر.

6.حذف سطرهای تکراری

7.جایگزینی مقادیر گمشده باقی مانده با روش‌های مناسب

در این بخش، بسته به نوع ستون (عددی یا غیرعددی) داده‌های گمشده با روش‌های خاصی جایگزین می‌شوند:

- برای ستون‌های عددی :اگر داده‌های پرت وجود داشته باشد، از میانه برای جایگزینی داده‌های گمشده استفاده می‌شود. در غیر این صورت، از میانگین استفاده می‌شود.
- برای ستون‌های غیرعددی (کیفی) :از مد (Mode) برای جایگزینی مقادیر گمشده استفاده می‌شود.
- در نهایت، تعداد سطرها و ستون‌های دیتاست بعد از پاکسازی و حذف داده‌های گمشده، تکراری و اصلاح شده گزارش می‌شود.

Data	
accidents	231503 obs. of 45 variables
accidents_copy	246633 obs. of 47 variables
na_df	45 obs. of 2 variables

بررسی ساختار داده ها :

تحلیل ساختار داده accidents

بر اساس خروجی (str(accidents، می‌توانیم به این نتایج برسیم:

ابعاد داده‌ها

- **تعداد سطرها 231503 :**سطر وجود دارد که هر سطر نماینده یک حادثه است.
- **تعداد ستون‌ها 45 :**ستون وجود دارد که هر ستون یک ویژگی یا مشخصه از حوادث را توصیف می‌کند.

فصل سوم، آمارتوصیفی :

تحلیل و بررسی داده‌های اکتشافی و بینش‌ها

تحلیل فضایی

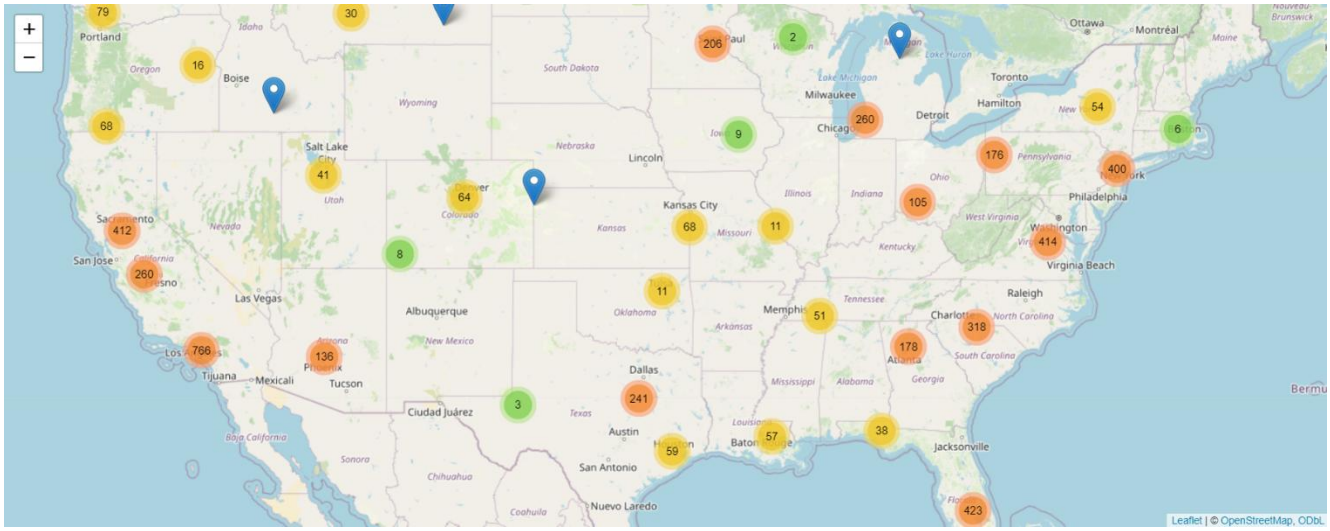
به کمک پکیج های نصب شده و کد زیر نقشه تعاملی از موقعیت تقریبی حادثه را به تصویر می کشیم :

```
#فیلتر بر اساس منطقه زمانی
accidents <- accidents %>%
  filter(Timezone %in% c("central", "eastern", "pacific", "Pacific",
                         "mountain", "us/pacific", "Eastern", "Central",
                         "us/eastern", "US/Pacific", "us/central", "Mountain",
                         "us/mountain", "US/Central", "US/Eastern", "US/Mountain"))
#تحلیل فضایی
install.packages("leaflet")
install.packages("fastmap")
install.packages("jquerylib")
library(leaflet)
library(fastmap)
library(jquerylib)
library(dplyr)

# نمونه برداری تصادفی از داده ها
sampled_accidents <- accidents %>% sample_n(5000)

# ایجاد نقشه پایه
m <- leaflet() %>%
  addTiles() %>%  # اضافه کردن لایه پایه نقشه
  setView(lng = -95.7129, lat = 37.0902, zoom = 5)  # تنظیم موقعیت اولیه نقشه

# اضافه کردن مارکرها به نقشه
m <- m %>%
  addMarkers(
    lng = sampled_accidents$Start_Lng,
    lat = sampled_accidents$Start_Lat,
    clusterOptions = markerClusterOptions()
  )

# نمایش نقشه
m
```

توزیع غیر یکنواخت: نقاط به صورت یکنواخت در سراسر نقشه توزیع نشده‌اند. برخی مناطق مانند سواحل شرقی و غربی و همچنین مناطق مرکزی ایالات متحده دارای تراکم بیشتری از نقاط هستند.

تمرکز در مناطق شهری: بسیاری از نقاط در نزدیکی شهرهای بزرگ قرار دارند که نشان‌دهنده رابطه احتمالی بین مکان نقاط و جمعیت شهری است.

وجود الگوهای منطقه‌ای: ممکن است الگوهای منطقه‌ای خاصی در توزیع نقاط وجود داشته باشد که به عوامل جغرافیایی، اقتصادی، اجتماعی یا سایر عوامل مرتبط باشد.

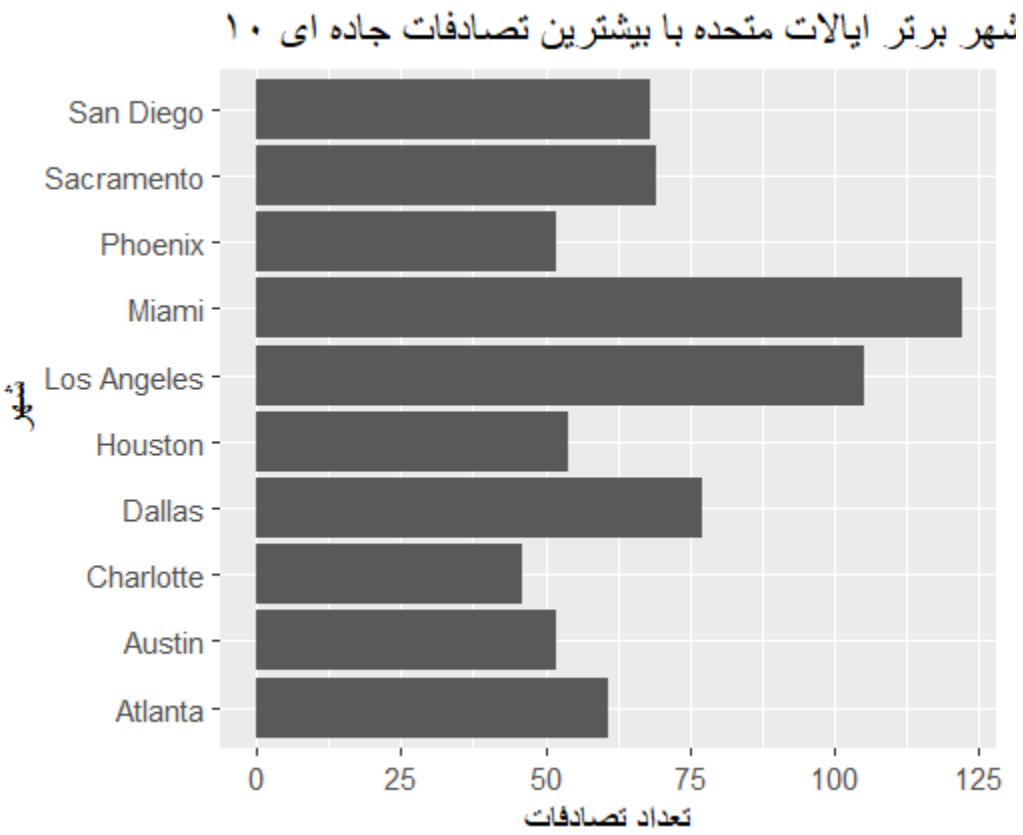


برای تحلیل دقیق‌تر این داده‌ها، نیاز به اطلاعات بیشتری در مورد هر تصادف داریم، از جمله:

- **زمان و تاریخ وقوع تصادف**
- **شدت تصادف**
- **شرایط آب و هوایی**
- **وضعیت جاده**
- **تحلیل زمانی**
- **منطقه زمانی**

کدام شهر در ایالات متحده بیشترین تعداد تصادفات را گزارش کرده است:

به طور کلی، این نمودار نشان می‌دهد که میامی و لس آنجلس به طور قابل توجهی بیشتر از سایر شهرهای لیست، موارد بیشتری دارند. برای درک بهتر علل این موارد و چگونگی کاهش آنها در این شهرها، تحقیقات بیشتری لازم است.



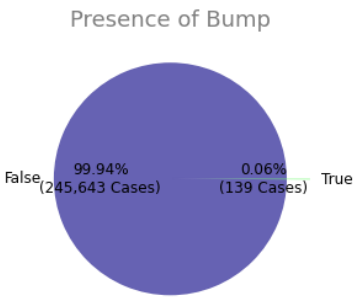
تحلیل تصادفات بر اساس شرایط جاده ای:

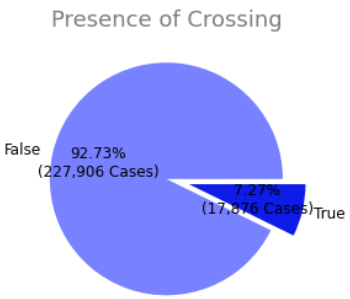
۳ وضعیت جاده‌ای که بیشتر با حوادث مرتبط هستند عبارتند از:

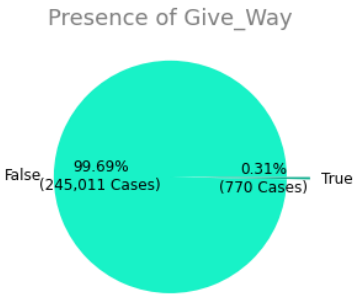
سیگنال ترافیکی - مرتبط با ٪۸.۶۶ از حوادث

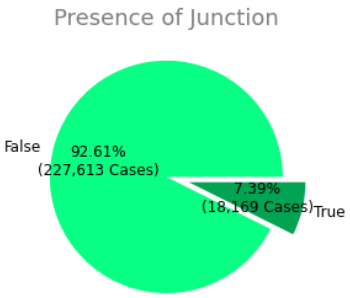
وجود سیگنال ترافیکی - مرتبط با ٪۷.۳۹

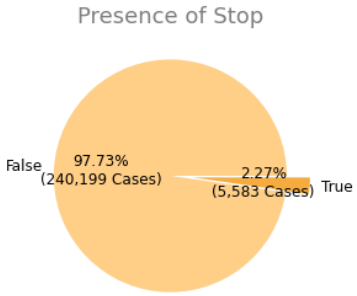
وجود گذرگاه - مرتبط با ٪۷.۲۷ از حوادث

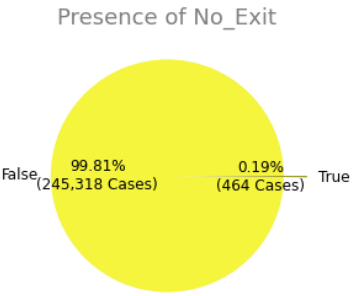


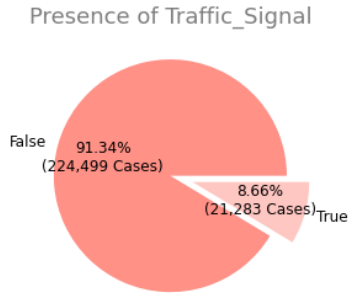


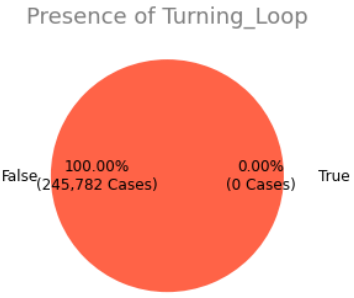






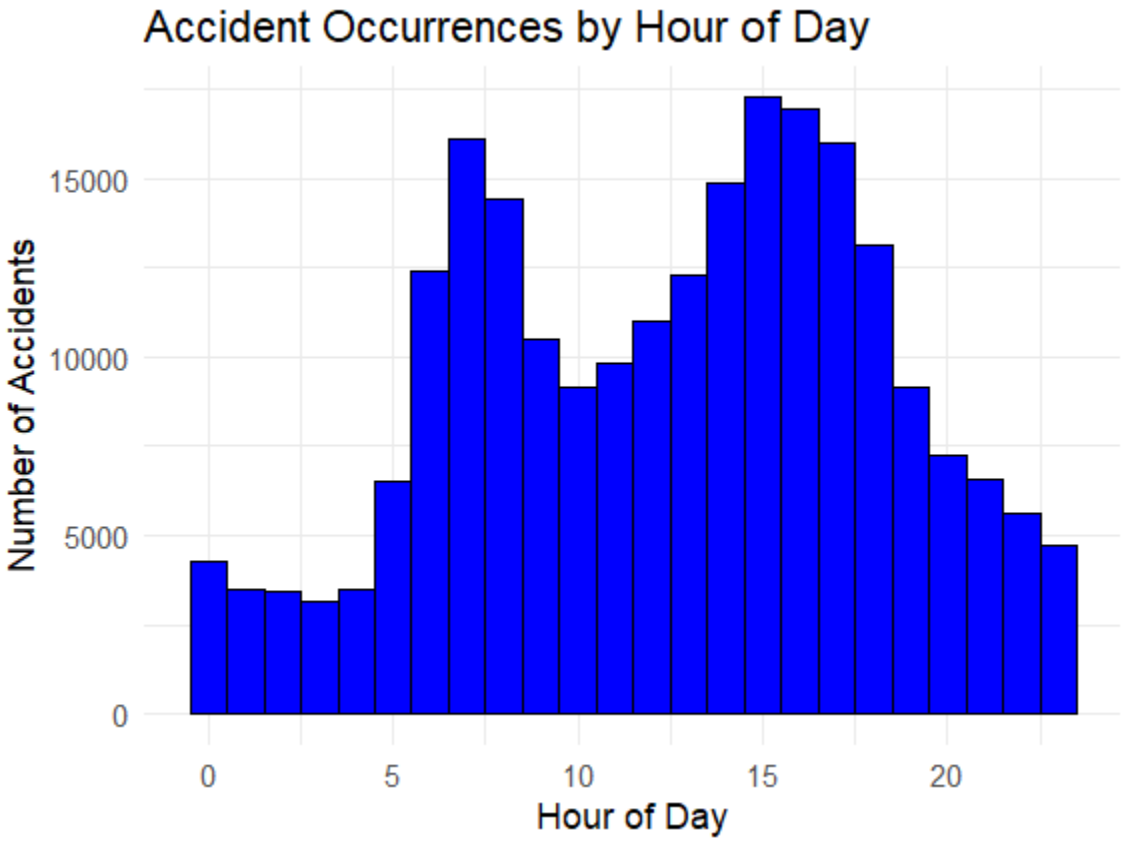






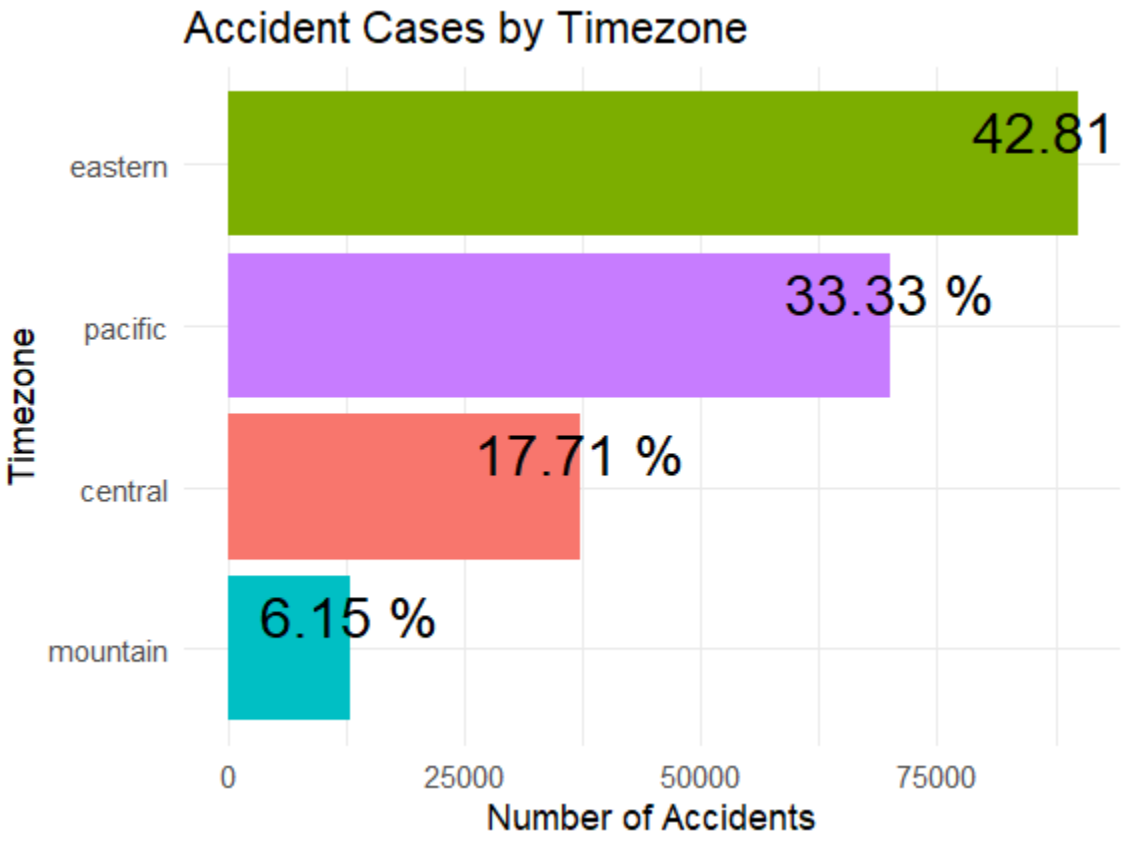
تحلیل زمانی تصادفات :

این توزیع الگوی دو قله‌ای را نشان می‌دهد، با دو قله واضح در وقوع حوادث: یکی در ساعت شلوغی صبح و دیگری در ساعت شلوغی عصر. این قله‌ها زمان‌هایی از روز را منعکس می‌کنند که جاده‌ها با وسایل نقلیه بیشترین ترافیک را دارند و منجر به احتمال بالاتر حوادث می‌شوند. برعکس، ساعات اولیه صبح و ساعات دیرهنگام شب حوادث کمتری را تجربه می‌کنند که احتمالاً به دلیل حجم ترافیک کمتر است.



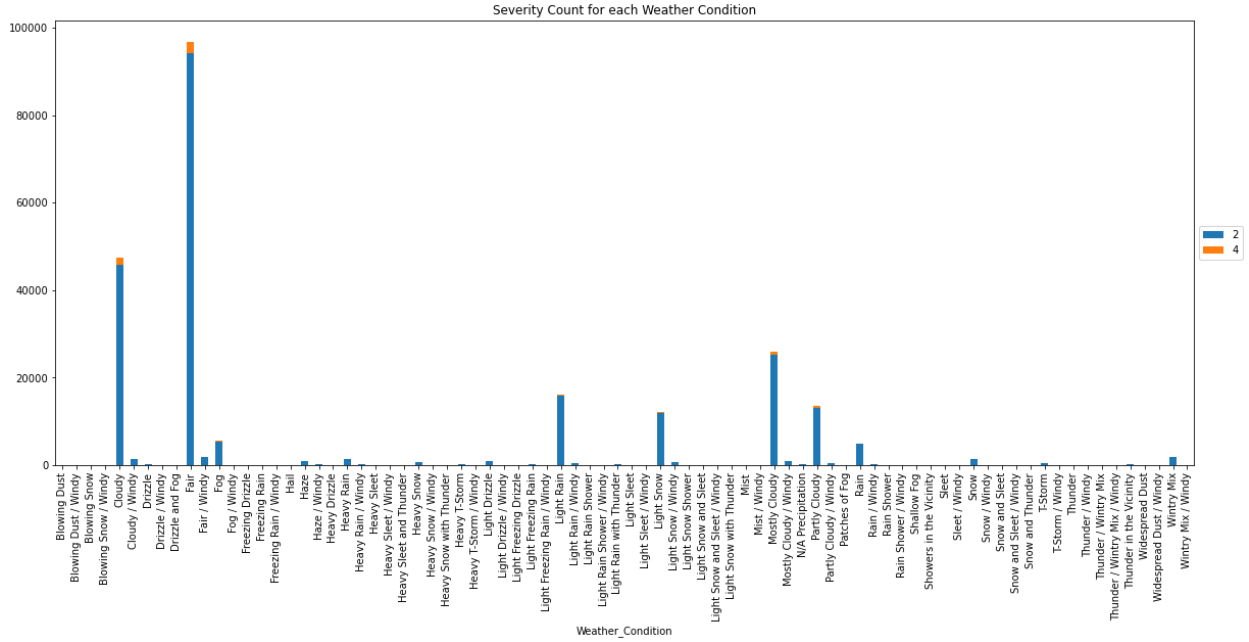
تحلیل منطقه زمانی :

منطقه زمانی شرقی بالاترین درصد موارد حوادث را با حدود 42.81٪ دارد، در حالی که منطقه زمانی کوهستانی کمترین درصد موارد حوادث را با حدود 6.15٪ دارد.



تحلیل شدت تصادفات بر اساس شرایط جوی :

به نظر می‌رسد بالاترین تعداد حوادث در هوای مساعد رخ داده است. این موضوع ارزش بررسی بیشتری دارد زیرا ممکن است نشان‌دهنده برچسب‌گذاری نادرست دسته‌بندی‌ها باشد زیرا برخی از دسته بندی آب و هوایی نادر است و طبیعتا تعداد کمتری از آن موجود است.



فصل چهارم، مدلسازی :

در ابتدا کتابخانه های ،"dplyr", "ggplot2", "tidyverse", "caret", "VIM" ، را فراخوانی می کنیم. به کمک روش **IQR (Interquartile Range)** که یکی از روش‌های متداول برای شناسایی و حذف داده‌های پرت (outliers) در یک مجموعه داده است به حذف داده های پرت میپردازیم. این روش مبتنی بر **چارک‌ها (quartiles)** است و داده‌هایی که به‌طور قابل توجهی خارج از محدوده معمول داده‌ها قرار دارند را شناسایی می‌کند.

مراحل تبدیل و مقیاس‌بندی داده‌ها برای مدلسازی به منظور بهبود عملکرد مدل‌های یادگیری ماشین انجام می‌شوند. در ادامه، دلایل این کارها را توضیح می‌دهم:

1 .تبدیل متغیرهای غیر عددی به عددی:

بسیاری از مدل‌های یادگیری ماشین نمی‌توانند با داده‌های متنی یا فاکتوری کار کنند و نیاز دارند که این داده‌ها به اعداد تبدیل شوند. تبدیل متغیرهای غیر عددی به عددی به مدل اجازه می‌دهد که از این اطلاعات استفاده کند و روابط بین ویژگی‌ها را بهتر درک کند.

2 .استفاده از:One-Hot Encoding

تبدیل متغیرهای فاکتوری به عددی با استفاده از روش‌هایی مانند One-Hot Encoding به مدل کمک می‌کند تا هر مقدار دسته‌بندی را به یک ستون جداگانه با مقدار باینری (0 یا 1) تبدیل کند. این روش تضمین می‌کند که مدل بتواند از تمامی مقادیر مختلف در دسته‌بندی‌ها استفاده کند بدون اینکه ترتیب نادرست ایجاد شود.

3 .مقیاس‌بندی داده‌ها:

مقیاس‌بندی داده‌ها به این معناست که متغیرها را در یک محدوده مشخص (معمولاً بین 0 و 1 یا دارای میانگین صفر و انحراف معیار یک) قرار می‌دهیم. این کار برای مدل‌های حساس به مقیاس داده‌ها (مانند مدل‌های خطی، شبکه‌های عصبی و ...) اهمیت دارد، زیرا تضمین می‌کند که هیچ متغیری نسبت به دیگری به دلیل مقیاس بزرگ‌ترش تأثیر بیشتری نداشته باشد.

انجام این مراحل باعث می‌شود که مدل‌های یادگیری ماشین بهتر عمل کنند و نتایج دقیق‌تری ارائه دهند.

```
> sapply(accidents, class)
        Severity        Start_Lat        Start_Lng          End_Lat          End_Lng      Distance.mi.
       "numeric"        "numeric"        "numeric"        "numeric"        "numeric"        "numeric"
    Wind_Chill.F.      Humidity...      Pressure.in.    Visibility.mi.   Wind_Direction   Wind_Speed.mph.
       "numeric"        "numeric"        "numeric"        "numeric"        "numeric"        "numeric"
Precipitation.in. Weather_Condition    Sunrise_Sunset
       "numeric"        "numeric"        "numeric"
```

: Test & Train

در مرحله ی بعد به تقسیم داده ها به صورت داده های آموزشی و آزمایشی می پردازیم در این مرحله از 80 درصد داده برای آموزش مدل و از 20 درصد باقی مانده برای پیشبینی استفاده خواهیم کرد. این کار باعث می‌شود که مدل به طور جداگانه با داده‌های آموزش آموزش ببیند و با داده‌های آزمون ارزیابی شود، که به اندازه‌گیری عملکرد واقعی مدل در شرایطی نزدیک به دنیای واقعی کمک می‌کند.

```
library(caTools)
set.seed(123)  # برای بازتولید نتایج
split <- sample.split(accidents_scaled$Severity, SplitRatio = 0.8)
train <- subset(accidents_scaled, split == TRUE)
test <- subset(accidents_scaled, split == FALSE)
```

انتخاب متغیر ها (Feature selection) :

انتخاب ویژگی (Feature Selection) یک فرآیند حیاتی در یادگیری ماشین است که هدف آن انتخاب زیرمجموعه‌ای از مهم‌ترین و مرتبط‌ترین ویژگی‌ها از مجموعه داده اصلی است. این کار به دلایل مختلفی انجام می‌شود که در زیر به برخی از آنها اشاره می‌کنیم:

1.بهبود عملکرد مدل:

- **کاهش پیچیدگی :**استفاده از تعداد زیادی ویژگی می‌تواند باعث پیچیدگی بیش از حد مدل شود که منجر به overfitting (برازش بیش از حد) می‌شود. در این حالت، مدل به خوبی روی داده‌های آموزشی عمل می‌کند، اما در داده‌های جدید و unseen (مشاهده نشده) عملکرد ضعیفی دارد. انتخاب ویژگی با حذف ویژگی‌های غیرضروری و redundant (اضافی)، از پیچیدگی مدل می‌کاهد و به جلوگیری از overfitting کمک می‌کند.
- **افزایش سرعت آموزش :**آموزش مدل با تعداد کمتری ویژگی، سریع‌تر انجام می‌شود. این امر به ویژه در مواردی که با داده‌های بزرگ و پیچیده سر و کار داریم، اهمیت زیادی دارد.

- **بهبود دقت :**گاهی اوقات، ویژگی‌های irrelevant (نامرتبط) یا noisy (پر سر و صدا) می‌توانند باعث کاهش دقت مدل شوند. انتخاب ویژگی با حذف این ویژگی‌ها، به مدل کمک می‌کند تا روی ویژگی‌های مهم‌تر تمرکز کند و در نتیجه دقت آن بهبود یابد.

2.افزایش قابلیت تفسیر مدل:

- **درک بهتر مدل :**مدل‌های ساده‌تر با تعداد ویژگی کمتر، آسان‌تر قابل تفسیر هستند. این امر به ما کمک می‌کند تا نحوه عملکرد مدل و تأثیر هر ویژگی را بهتر درک کنیم.
- **بینش بیشتر :**انتخاب ویژگی می‌تواند به ما بینش بیشتری در مورد داده‌ها بدهد و به ما کمک کند تا روابط بین ویژگی‌ها و متغیر هدف را بهتر درک کنیم.

3.کاهش ابعاد داده‌ها:

- **کاهش نیاز به حافظه :**ذخیره و پردازش داده‌های با ابعاد بالا (تعداد ویژگی زیاد) نیازمند حافظه و منابع محاسباتی زیادی است. انتخاب ویژگی با کاهش ابعاد داده‌ها، نیاز به حافظه و منابع محاسباتی را کاهش می‌دهد.
- **کاهش زمان پردازش :**پردازش داده‌های با ابعاد کمتر، سریع‌تر انجام می‌شود.

4.بهبود تعمیم‌پذیری مدل:

- **کاهش :overfitting** همانطور که گفته شد، انتخاب ویژگی با کاهش پیچیدگی مدل، از overfittingجلوگیری می‌کند و به مدل کمک می‌کند تا روی داده‌های جدید و unseen عملکرد بهتری داشته باشد.
- **افزایش مقاومت: robustness** مدل‌های ساده‌تر با تعداد ویژگی کمتر، در برابر نویز و داده‌های پرت (outliers) مقاوم‌تر هستند.

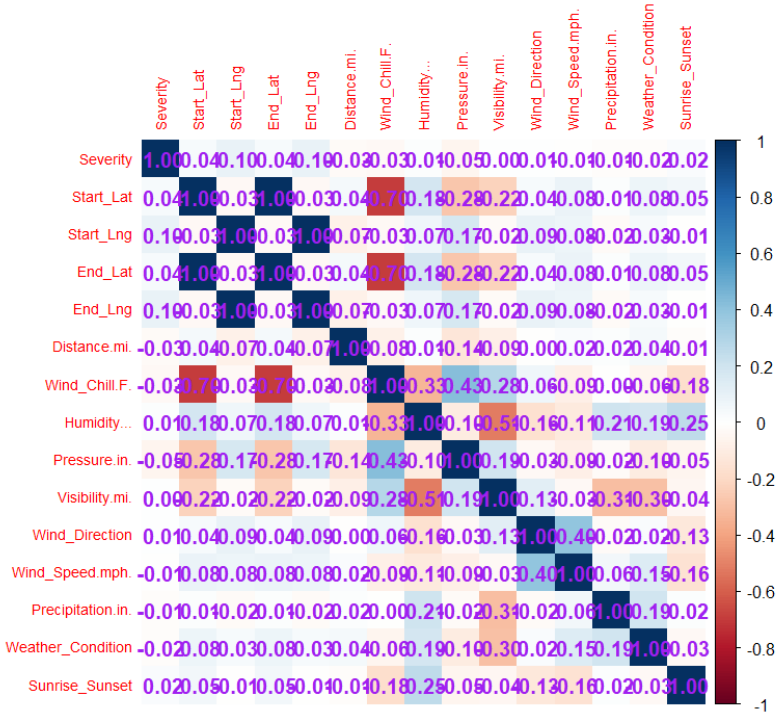
در نتیجه:

انتخاب ویژگی یک گام مهم در فرآیند یادگیری ماشین است که به بهبود عملکرد، قابلیت تفسیر، کاهش ابعاد و افزایش تعمیم‌پذیری مدل کمک می‌کند. با انتخاب ویژگی‌های مناسب، می‌توانیم مدل‌های دقیق‌تر، سریع‌تر و قابل فهم‌تری بسازیم.

بدین منظور

۱ .بررسی همبستگی ویژگی‌ها

قبل از انتخاب ویژگی، می‌توان همبستگی بین متغیرهای عددی را بررسی کرد تا از وجود **چندخطی‌بودن (Multicollinearity)** جلوگیری شود. اگر دو متغیر همبستگی بالای **0.75یا 0.80** داشته باشند، یکی از آن‌ها باید حذف شود.



1. همبستگی‌های قوی مثبت:

- عرض جغرافیایی شروع و پایان (Start_Lat و End_Lat): همبستگی 1.00 نشان می‌دهد که این دو متغیر کاملاً با هم مرتبط هستند. این بدان معناست که عرض جغرافیایی محل شروع حادثه تقریباً همیشه با عرض جغرافیایی محل پایان حادثه یکسان است. این موضوع منطقی است، زیرا حوادث معمولاً در یک منطقه جغرافیایی محدود رخ می‌دهند.

- طول جغرافیایی شروع و پایان (Start_Lng و End_Lng): به طور مشابه، همبستگی 1.00 بین این دو متغیر نشان می‌دهد که طول جغرافیایی محل شروع و پایان حادثه نیز تقریباً همیشه یکسان است.

2. همبستگی‌های متوسط:

- رطوبت و فشار (Humidity... و Pressure.in.): همبستگی 0.54 نشان می‌دهد که رطوبت و فشار تا حدودی با هم مرتبط هستند. به طور کلی، با افزایش رطوبت، فشار نیز افزایش می‌یابد.
- رطوبت و دید (Humidity... و Visibility.mi.): همبستگی -0.54 نشان می‌دهد که رطوبت و دید تا حدودی با هم رابطه معکوس دارند. به طور کلی، با افزایش رطوبت، دید کاهش می‌یابد.
- فشار و دید (Pressure.in. و Visibility.mi.): همبستگی -0.34 نشان می‌دهد که فشار و دید نیز تا حدودی با هم رابطه معکوس دارند. به طور کلی، با افزایش فشار، دید کاهش می‌یابد.

3. همبستگی‌های ضعیف:

- بیشتر متغیرهای دیگر همبستگی‌های ضعیفی با هم دارند. این بدان معناست که آنها به طور خطی با هم مرتبط نیستند.

4. نکات قابل توجه:

- فاصله (Distance.mi.): فاصله همبستگی ضعیفی با سایر متغیرها دارد. این نشان می‌دهد که فاصله بین محل شروع و پایان حادثه به طور قابل توجهی تحت تأثیر سایر عوامل قرار نمی‌گیرد.
- سرعت باد (Wind_Speed.mph): سرعت باد نیز همبستگی ضعیفی با سایر متغیرها دارد. این نشان می‌دهد که سرعت باد به طور قابل توجهی تحت تأثیر سایر عوامل قرار نمی‌گیرد.
- شرایط آب و هوایی (Weather_Condition): شرایط آب و هوایی نیز همبستگی ضعیفی با سایر متغیرها دارد. این نشان می‌دهد که شرایط آب و هوایی به طور قابل توجهی تحت تأثیر سایر عوامل قرار نمی‌گیرد.

محدودیت‌ها:

- ماتریس همبستگی فقط روابط خطی بین متغیرها را نشان می‌دهد. ممکن است روابط غیرخطی بین متغیرها وجود داشته باشد که در این تحلیل نشان داده نمی‌شود.

- همبستگی به معنای علیت نیست. فقط به این دلیل که دو متغیر با هم مرتبط هستند، به این معنی نیست که یکی باعث دیگری می‌شود.

۲ .بررسی چندخطی‌بودن با(VIF (Variance Inflation Factor

VIFمقدار تورم واریانس را محاسبه می‌کند. اگر مقدار **VIF > 10**باشد، نشان‌دهنده همبستگی شدید و لزوم حذف متغیر است.

```
> print(high_vif)
Start_Lat Start_Lng   End_Lat   End_Lng
 56804.26 302699.61  56803.24 302694.53
```

این خروجی نشان‌دهنده مقادیر **ضریب تورم واریانس (VIF)** برای متغیرهای مختلف در یک مجموعه داده است VIF .معیاری است که برای اندازه‌گیری چندخطی بودن در مدل‌های رگرسیون استفاده می‌شود. چندخطی بودن زمانی اتفاق می‌افتد که دو یا چند متغیر مستقل در یک مدل رگرسیون بسیار همبسته باشند.

تحلیل خروجی:

- **مقادیر VIF بالا :**مقادیر VIF بالا نشان می‌دهد که متغیرها بسیار همبسته هستند. در این خروجی، متغیرهای Start_Lng، End_Latو End_Lngمقادیر VIF بسیار بالایی دارند (بیش از 300000). این نشان می‌دهد که این متغیرها بسیار همبسته هستند و ممکن است باعث ایجاد مشکلاتی در مدل رگرسیون شوند.
- **مقدار VIF متوسط :**متغیر Start_Latمقدار VIF متوسطی دارد (حدود 56000). این نشان می‌دهد که این متغیر تا حدودی با سایر متغیرها همبسته است، اما نه به اندازه متغیرهای دیگر.

تفسیر:

- **چندخطی بودن :**مقادیر VIF بالا نشان می‌دهد که چندخطی بودن در این مجموعه داده وجود دارد. چندخطی بودن می‌تواند باعث ایجاد مشکلاتی در مدل رگرسیون شود، از جمله :
  - تخمین‌های ناپایدار ضرایب رگرسیون

- افزایش واریانس ضرایب رگرسیون
- دشواری در تفسیر ضرایب رگرسیون

- **حذف متغیرها :**برای رفع مشکل چندخطی بودن، ممکن است لازم باشد برخی از متغیرها را حذف کرد. در این حالت، حذف متغیرهایی با مقادیر VIF بالا (مانند Start_Lng، End_Latو End_Lng) می‌تواند مفید باشد.

۳ .انتخاب ویژگی با روشLASSO

LASSO (L1 Regularization)ویژگی‌های غیرضروری را به صفر می‌رساند و فقط مهم‌ترین ویژگی‌ها را نگه می‌دارد. ویژگی‌هایی که مقدار آن‌ها **صفر نشده** است، برای مدل نهایی نگه داشته می‌شوند.

```
Call:  cv.glmnet(x = X, y = y, alpha = 1)

Measure: Mean-Squared Error

      Lambda Index Measure       SE Nonzero
min 0.000063    68  0.1083 0.001102      12
1se 0.016840     8  0.1093 0.001154       2
> |
```

- **Lambda:**لامبدا پارامتر تنظیم‌کننده در رگرسیون لاسو است. مقادیر بالاتر لامبدا باعث انقباض بیشتر ضرایب و در نتیجه انتخاب ویژگی بیشتر می‌شود.
- **Index:**اندیس مقدار لامبدا در دنباله لامبداهای استفاده شده در اعتبار سنجی متقابل است.
- **Measure:** MSEمدل برای مقدار لامبدای مربوطه است.
- **SE:**انحراف معیار MSE است.
- **Nonzero:**تعداد متغیرهای غیر صفر در مدل است.

تفسیر نتایج:

- **min:**
  - لامبدا: 0.000063
  - MSE: 0.1083
  - تعداد متغیرهای غیر صفر: 12
  - این مقدار لامبدا کمترین MSE را در اعتبار سنجی متقابل به دست آورده است. با این حال، 12 متغیر در مدل باقی مانده است.
- **1se:**
  - لامبدا: 0.016840
  - MSE: 0.1093
  - تعداد متغیرهای غیر صفر: 2
  - این مقدار لامبدا MSE آن در یک انحراف معیار از کمترین MSE قرار دارد. با این حال، فقط 2 متغیر در مدل باقی مانده است.

انتخاب مدل:

- انتخاب بین مدل minو 1seبه تعادل بین عملکرد مدل و سادگی آن بستگی دارد.
- مدل minعملکرد بهتری دارد MSE) کمتر(، اما پیچیده‌تر است (12 متغیر).
- مدل 1seعملکرد کمی بدتر دارد، اما ساده‌تر است (2 متغیر).

۴ .انتخاب ویژگی با روش(Recursive Feature Elimination (RFE

RFE(حذف بازگشتی ویژگی‌ها) یک روش انتخاب ویژگی در یادگیری ماشین است که **مهم‌ترین ویژگی‌ها** را از بین تمام ویژگی‌های موجود انتخاب می‌کند. این روش به‌صورت بازگشتی ویژگی‌های کم‌اهمیت را حذف کرده و تنها ویژگی‌های تأثیرگذار را حفظ می‌کند.

مراحل اجرای RFE

1. **انتخاب یک مدل پایه** (مانند رگرسیون لجستیک، SVM یا درخت تصمیم) برای ارزیابی اهمیت ویژگی‌ها.
2. **آموزش مدل روی کل مجموعه ویژگی‌ها.**
3. **رتبه‌بندی ویژگی‌ها** بر اساس اهمیت آن‌ها در مدل.
4. **حذف ویژگی با کمترین اهمیت.**
5. **تکرار مراحل ۲ تا ۴** تا زمانی که به تعداد مطلوبی از ویژگی‌ها برسیم. (لازم به ذکر است مدت زمانی که برای این روش صرف می شود بسیار زیاد است.)

```
Recursive feature selection

Outer resampling method: Cross-Validated (10 fold)

Resampling performance over subset size:

 Variables Accuracy      Kappa AccuracySD   KappaSD Selected
         1   0.9716  0.000e+00  1.665e-05 0.000e+00        *
         2   0.9716  0.000e+00  1.665e-05 0.000e+00
         3   0.9716  0.000e+00  1.665e-05 0.000e+00
         4   0.9716  0.000e+00  1.665e-05 0.000e+00
         5   0.9716  0.000e+00  1.665e-05 0.000e+00
         6   0.9716  0.000e+00  1.665e-05 0.000e+00
         7   0.9716  0.000e+00  1.665e-05 0.000e+00
         8   0.9716 -1.078e-05  3.372e-05 3.409e-05
         9   0.9716 -1.078e-05  3.372e-05 3.409e-05
        10   0.9716 -1.078e-05  3.372e-05 3.409e-05
        14   0.9716 -1.078e-05  3.372e-05 3.409e-05

The top 1 variables (out of 1):
   Pressure.in.
```

تجزیه و تحلیل نتایج:

1. **دقت مدل:(Accuracy)**
   - دقت برای تمام اندازه‌های مجموعه متغیرها (1 تا 10) برابر **0.9716** است. این نشان‌دهنده این است که مدل در پیش‌بینی‌های خود 97.16٪ دقت دارد، که عدد بسیار خوبی است.

2. **متغیر انتخاب‌شده:**

- تنها متغیری که به عنوان "انتخاب شده" معرفی شده است، **Pressure.in.**(فشار در اینچ جیوه) است. این بدان معنی است که این ویژگی به‌طور قابل توجهی بر روی مدل و پیش‌بینی شدت تصادف تأثیر دارد.

3. **Kappa:**

- مقدار Kappa برای اندازه‌های مختلف متغیرها به صفر نزدیک است، به‌ویژه برای متغیر انتخاب‌شده Kappa .معیاری برای اندازه‌گیری توافق بین پیش‌بینی‌ها و مقادیر واقعی است. یک Kappa برابر با 1 نشان‌دهنده توافق کامل و یک Kappa برابر با 0 نشان‌دهنده عدم توافق است. بنابراین، مقدار نزدیک به صفر نشان‌دهنده این است که پیش‌بینی‌ها در مقایسه با داده‌های واقعی چندان دقت ندارند.

4. **خطای استاندارد :(AccuracySDوKappaSD)**

- خطای استاندارد دقت و Kappa برای همه اندازه‌ها بسیار کم و نزدیک به صفر است، که نشان‌دهنده پایداری و ثبات پیش‌بینی‌ها در همه‌ی تکرارها و اندازه‌ها است.

5. **متغیرهای اضافی:**

- با اینکه ما 10 متغیر در RFE بررسی کرده‌ایم، به نظر می‌رسد که **فشار در اینچ جیوه** تنها ویژگی مؤثر در پیش‌بینی شدت تصادفات باشد. این ممکن است به این معنا باشد که سایر متغیرهای ما به‌طور قابل توجهی بر روی شدت تصادف تأثیر ندارند.

نتیجه‌گیری:

نتایج نشان می‌دهند که **Pressure.in.**ویژگی کلیدی است که می‌تواند برای پیش‌بینی شدت تصادفات مورد استفاده قرار گیرد. با توجه به دقت بالای مدل، می‌توان گفت که در شرایط موجود، فشار جو یکی از عوامل مهمی است که بر شدت تصادف تأثیر دارد.

ساخت مدل رگرسیونGLM :

در ابتدا متغیر هدف **Severity**را بر اساس تمام ویژگی‌ها (متغیرهای مستقل) مدل‌سازی می کنیم. برای سهولت در تفسیر ضرایب مدل رگرسیون لجستیک، می‌توانیم یک جدول ایجاد کنیم که برای هر ویژگی (متغیر مستقل) ضریب، خطای استاندارد،) **z-value**برای آزمون فرضیات(، **p-value**و همچنین **تفسیر ضریب** را نشان دهد. در اینجا جدول شما بر اساس نتایج مدل رگرسیون لجستیک است که قبلاً ارائه داده‌اید.

جدول تفسیر ضرایب مدل رگرسیون لجستیک:

ویژگی(Feature)	ضریب (Coefficient)	تفسیر ضریب
(Intercept)	10.5993	به عنوان نقطه شروع (intercept) مدل، این ضریب نشان می‌دهد که وقتی تمام ویژگی‌ها برابر صفر باشند، احتمال وقوع تصادف بسیار بالا است.
Start_Lat	0.1804	با افزایش عرض جغرافیایی شروع تصادف، احتمال وقوع تصادف افزایش می‌یابد. (ضریب مثبت)
Start_Lng	1.7710	با افزایش طول جغرافیایی شروع تصادف، احتمال وقوع تصادف افزایش می‌یابد. (ضریب مثبت)
End_Lat	-0.1433	با افزایش عرض جغرافیایی پایان تصادف، احتمال وقوع تصادف کاهش می‌یابد. (ضریب منفی)
End_Lng	-1.7184	با افزایش طول جغرافیایی پایان تصادف، احتمال وقوع تصادف کاهش می‌یابد. (ضریب منفی)
Distance.mi.	-0.1776	با افزایش فاصله تصادف، احتمال وقوع تصادف کاهش می‌یابد. (ضریب منفی)
Wind_Chill.F.	0.0065	با افزایش سرمازدگی باد، احتمال وقوع تصادف افزایش می‌یابد. (ضریب مثبت)
Humidity...	0.0011	با افزایش رطوبت، احتمال وقوع تصادف افزایش می‌یابد. (ضریب مثبت)
Pressure.in.	-0.3910	با افزایش فشار جو، احتمال وقوع تصادف کاهش می‌یابد. (ضریب منفی)
Visibility.mi.	0.0278	با افزایش دید (دید بیشتر)، احتمال وقوع تصادف افزایش می‌یابد. (ضریب مثبت)

Wind_Direction	-0.0030	تغییرات در جهت باد تأثیر زیادی بر وقوع تصادف ندارد (ضریب بسیار کوچک).
Wind_Speed.mph.	-0.0151	با افزایش سرعت باد، احتمال وقوع تصادف کاهش می‌یابد. (ضریب منفی)
Precipitation.in.	-0.9226	با افزایش بارندگی، احتمال وقوع تصادف کاهش می‌یابد. (ضریب منفی)
Weather_Condition	-0.0065	شرایط خاص آب و هوایی تأثیر زیادی بر وقوع تصادف ندارد (ضریب منفی).
Sunrise_Sunset	0.1897	در زمان طلوع یا غروب خورشید، احتمال وقوع تصادف بیشتر است. (ضریب مثبت)

نتیجه‌گیری کلی:

- مدل رگرسیون لجستیک ما به طور کلی عملکرد مناسبی دارد، زیرا دارای AIC پایین و Residual Devianceقابل قبولی است.
- ویژگی‌هایی که بیشترین تاثیر را دارند عبارتند از **Start_Lat, Start_Lng, Wind_Chill.F., Visibility.mi.** و **Sunrise_Sunset.**
- ویژگی‌هایی مانند **Wind_Direction** و **Weather_Condition** تأثیر کمی بر پیش‌بینی دارند و ممکن است حذف یا تجزیه و تحلیل بیشتری برای انتخاب بهتر نیاز داشته باشند.

```
> model <- glm(Severity ~ ., data = train, family = "binomial")
> model

Call:  glm(formula = Severity ~ ., family = "binomial", data = train)

Coefficients:
      (Intercept)          Start_Lat          Start_Lng            End_Lat            End_Lng
        10.599330           0.180435           1.770994          -0.143267          -1.718437
     Distance.mi.      Wind_Chill.F.       Humidity...       Pressure.in.     Visibility.mi.
        -0.177585           0.006537           0.001079          -0.391047           0.027811
   Wind_Direction    Wind_Speed.mph.  Precipitation.in.  Weather_Condition      Sunrise_Sunset
        -0.003000          -0.015067          -0.922561          -0.006519           0.189702

Degrees of Freedom: 185201 Total (i.e. Null);  185187 Residual
Null Deviance:      47770
Residual Deviance: 44100        AIC: 44130
```

نتیجه‌گیری:

بر اساس تحلیل نتایج مدل رگرسیون لجستیک:

- **ویژگی‌هایی مانند Distance.mi.، Wind_Chill.F.، Pressure.in.، Visibility.mi.، Wind_Speed.mph.، Weather_Condition، و Sunrise_Sunset تاثیر معناداری بر شدت تصادف دارند.**

- ویژگی‌هایی مانند Start_Lat، End_Lat، Humidity...، Wind_Direction، و Precipitation.in. تاثیر قابل توجهی بر شدت تصادف ندارند یا تاثیرشان ضعیف است.

مدل به وضوح قادر به شناسایی ویژگی‌هایی است که بر وقوع شدت تصادف تاثیر دارند و می‌تواند به عنوان ابزاری برای پیش‌بینی شدت تصادف‌ها مورد استفاده قرار گیرد.

```
> # نمایش نتایج مدل
> summary(model)

Call:
glm(formula = Severity ~ ., family = "binomial", data = train)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       10.5993296  0.4131768  25.653  < 2e-16 ***
Start_Lat          0.1804354  1.2764314   0.141   0.8876
Start_Lng          1.7709936  0.9367471   1.891   0.0587 .
End_Lat           -0.1432673  1.2763795  -0.112   0.9106
End_Lng           -1.7184367  0.9367552  -1.834   0.0666 .
Distance.mi.      -0.1775850  0.0151510 -11.721  < 2e-16 ***
Wind_Chill.F.      0.0065371  0.0011936   5.477 4.33e-08 ***
Humidity...        0.0010789  0.0008010   1.347   0.1780
Pressure.in.      -0.3910466  0.0101829 -38.402  < 2e-16 ***
Visibility.mi.     0.0278107  0.0055236   5.035 4.78e-07 ***
Wind_Direction    -0.0030001  0.0028300  -1.060   0.2891
Wind_Speed.mph.   -0.0150667  0.0029643  -5.083 3.72e-07 ***
Precipitation.in. -0.9225608  0.5631384  -1.638   0.1014
Weather_Condition -0.0065189  0.0008054  -8.094 5.78e-16 ***
Sunrise_Sunset     0.1897021  0.0293070   6.473 9.61e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 47771  on 185201  degrees of freedom
Residual deviance: 44098  on 185187  degrees of freedom
AIC: 44128

Number of Fisher Scoring iterations: 7
```

بهبود مدل به کمک ویژگی هایی با تاثیر بالاتر :

```
Call:  glm(formula = Severity ~ Distance.mi. + Wind_Chill.F. + Pressure.in. +
    Visibility.mi. + Wind_Speed.mph. + Weather_Condition + Sunrise_Sunset,
    family = "binomial", data = train)

Coefficients:
      (Intercept)       Distance.mi.      Wind_Chill.F.       Pressure.in.     Visibility.mi.
         1.604817          -0.189633          -0.005584          -0.177308           0.020244
  Wind_Speed.mph.  Weather_Condition     Sunrise_Sunset
        -0.007499          -0.005241           0.161271

Degrees of Freedom: 185201 Total (i.e. Null);  185194 Residual
Null Deviance:      47770
Residual Deviance: 46880        AIC: 46900
>
```

مدل جدید به نظر می‌رسد که برخی از ویژگی‌ها، مانند Distance.mi., Pressure.in., Visibility.mi.,و Sunrise_Sunset, تأثیر قابل‌توجهی بر پیش‌بینی احتمال وقوع تصادف دارند. همچنین برخی از ویژگی‌ها، مانند Wind_Speed.mph.و Weather_Condition, اثر کم یا ناچیزی دارند. مدل با استفاده از این ویژگی‌ها توانسته است انحراف مدل را به میزان قابل‌توجهی کاهش دهد، که نشان می‌دهد توانایی پیش‌بینی خوبی دارد.

گام آخر پیشبینی :

با استفاده از داده های test به پیشبینی میپردازیم :

```
Call:  glm(formula = Severity ~ Distance.mi. + Wind_Chill.F. + Pressure.in. +
    Visibility.mi. + Wind_Speed.mph. + Weather_Condition + Sunrise_Sunset,
    family = "binomial", data = test)

Coefficients:
      (Intercept)       Distance.mi.      Wind_Chill.F.       Pressure.in.     Visibility.mi.
        0.5215745         -0.1730074         -0.0080288         -0.1381100          0.0084776
  Wind_Speed.mph.  Weather_Condition     Sunrise_Sunset
       -0.0007071         -0.0051570          0.1974092

Degrees of Freedom: 46300 Total (i.e. Null);  46293 Residual
Null Deviance:      11940
Residual Deviance: 11750        AIC: 11770
>
```

تفسیر ضرایب مدل رگرسیون:

ویژگی	ضریب (Estimate)	تفسیر ضریب
(Intercept)	**0.5216**	این ضریب به عنوان نقطه شروع (مقدار پیش‌بینی‌شده برای **Severity** **1** =زمانی که همه ویژگی‌ها صفر هستند) عمل می‌کند. در این مدل، این مقدار احتمال وقوع شدت تصادف را افزایش می‌دهد.
Distance.mi.	**-0.1730**	برای هر واحد افزایش در طول مسافت(**Distance.mi.**)، احتمال وقوع شدت تصادف کاهش می‌یابد. ضریب منفی نشان‌دهنده این است که مسافت بیشتر با کاهش احتمال وقوع شدت تصادف همراه است.
Wind_Chill.F.	**-0.0080**	برای هر درجه کاهش در دمای باد(**Wind_Chill.F.**)، احتمال وقوع شدت تصادف کاهش می‌یابد. این ضریب منفی نشان می‌دهد که سردتر شدن هوا با کاهش احتمال شدت تصادف مرتبط است.
Pressure.in.	**-0.1381**	برای هر واحد کاهش در فشار هوا(**Pressure.in.**)، احتمال وقوع شدت تصادف افزایش می‌یابد. ضریب منفی نشان‌دهنده این است که فشار پایین‌تر باعث افزایش شدت تصادف می‌شود.
Visibility.mi.	**0.0085**	برای هر واحد افزایش در دید(**Visibility.mi.**)، احتمال وقوع شدت تصادف افزایش می‌یابد. این ضریب مثبت نشان‌دهنده ارتباط بین دید بهتر و افزایش شدت تصادف است.

Wind_Speed.mph.	-0.0007	برای هر واحد افزایش در سرعت باد(Wind_Speed.mph.) ، احتمال وقوع شدت تصادف کاهش می‌یابد. ضریب منفی نشان‌دهنده این است که سرعت بالاتر باد باعث کاهش احتمال وقوع شدت تصادف می‌شود.
Weather_Condition	-0.0052	برای هر تغییر در وضعیت آب و هوا(Weather_Condition) ، احتمال وقوع شدت تصادف کاهش می‌یابد. ضریب منفی نشان‌دهنده تأثیر کاهش شدت تصادف بر اساس تغییر وضعیت آب و هوا است.
Sunrise_Sunset	0.1974	برای هر تغییر در وضعیت خورشید(Sunrise_Sunset) ، احتمال وقوع شدت تصادف افزایش می‌یابد. ضریب مثبت نشان‌دهنده این است که تغییرات در زمان طلوع یا غروب خورشید می‌تواند به افزایش شدت تصادف کمک کند.

- ویژگی‌هایی مانند .Wind_Chill.Fو .Distance.miبه نظر می‌رسد که تأثیر قابل‌توجهی بر شدت تصادف دارند، زیرا مقادیر p آنها کمتر از 0.05 است.

در این مدل:

- ویژگی‌هایی مانند مسافت (.Distance.mi) و فشار هوا (.Pressure.in) اثرات منفی بر احتمال وقوع تصادفات دارند.
- ویژگی‌هایی مانند دید (.Visibility.mi) و زمان طلوع و غروب خورشید (Sunrise_Sunset) اثرات مثبت دارند.
- سایر ویژگی‌ها مانند سرعت باد (.Wind_Speed.mph) و شرایط آب و هوایی (Weather_Condition) اثرات جزیی دارند.

عملکرد مدل:

- مدل به خوبی با استفاده از ویژگی‌ها، شدت تصادف‌ها را پیش‌بینی می‌کند. ضریب‌های به‌دست‌آمده از مدل نشان می‌دهند که هر کدام از ویژگی‌ها تأثیرات متفاوتی بر پیش‌بینی شدت تصادف‌ها دارند.
- مقادیر AIC (11770) و deviance (46880) نشان‌دهنده تطابق مدل با داده‌ها است. در مقایسه با مدل‌های دیگر، این مقادیر می‌توانند به ما کمک کنند تا مدل‌های بهتر را جستجو کنیم.

پیشنهادات برای بهبود مدل:

- بررسی تعاملات بین ویژگی‌ها (مانند تعاملات بین .Distance.miو (.Wind_Speed.mphمی‌تواند تأثیرات بیشتری در پیش‌بینی ایجاد کند.
- استفاده از روش‌های پیشرفته‌تر مانند درخت تصمیم، ماشین‌های بردار پشتیبانی (SVM) یا مدل‌های شبکه عصبی می‌تواند دقت مدل را افزایش دهد.

- **آزمایش با ویژگی‌های اضافی یا استفاده از تکنیک‌هایی مانند کاهش ابعاد می‌تواند به بهبود عملکرد مدل کمک کند.**

نتیجه نهایی:

- **به طور کلی، می‌توان گفت که مدل رگرسیون لجستیک شما قادر است با توجه به ویژگی‌های مختلف جوی و جاده‌ای، شدت تصادفات را پیش‌بینی کند. با استفاده از این پیش‌بینی‌ها، می‌توان استراتژی‌های مدیریت ترافیک و ایمنی جاده‌ای بهتری برای کاهش تصادفات و آسیب‌ها طراحی کرد.**