فهرست

درباره‌ی دیتاست..... ۲
هدف پروژه ۶
پیش پردازش داده‌ها ۶
Feature Engineering..... ۷
Sampling ۷
Clustering ۷
Association Rules..... ۹
نتیجه‌گیری ۱۲
Visualization ۱۲
منابع ۱۸

US-Accidents: A Countrywide Traffic Accident Dataset بررسی تصادفات امریکا با استفاده از دیتاست

درباره‌ی دیتاست

این دیتاست یک مجموعه داده تصادفات رانندگی در سراسر کشور آمریکا است که ۴۹ ایالت ایالات متحده را پوشش می‌دهد. داده‌ها به طور مداوم از فوریه ۲۰۱۶ با استفاده از چندین ارائه‌دهنده داده، از جمله دو API که داده‌های جریان رویداد ترافیک را ارائه می‌دهند، جمع‌آوری می‌شود. این API ها رویدادهای ترافیکی را که توسط نهادهای مختلفی مانند وزارت حمل و نقل ایالات متحده، سازمان‌های اجرای قانون، دوربین‌های ترافیکی و حسگرهای ترافیک در شبکه های جاده‌ای ضبط شده است، پخش می‌کنند. در حال حاضر حدود ۴,۲ میلیون پرونده ثبت تصادف در این مجموعه داده وجود دارد.

جدول زیر ویژگی‌های داده‌ها را توصیف می‌کند:

#	Attribute	Description	Nullable
1	ID	This is a unique identifier of the accident record.	No
2	Source	Indicates source of the accident report (i.e. the API which reported the accident.).	No
3	TMC	A traffic accident may have a Traffic Message Channel (TMC) code which provides more detailed description of the event.	Yes
4	Severity	Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay).	No
5	Start_Time	Shows start time of the accident in local time zone.	No

#	Attribute	Description	Nullable
6	End_Time	Shows end time of the accident in local time zone. End time here refers to when the impact of accident on traffic flow was dismissed.	No
7	Start_Lat	Shows latitude in GPS coordinate of the start point.	No
8	Start_Lng	Shows longitude in GPS coordinate of the start point.	No
9	End_Lat	Shows latitude in GPS coordinate of the end point.	Yes
10	End_Lng	Shows longitude in GPS coordinate of the end point.	Yes
11	Distance(mi)	The length of the road extent affected by the accident.	No
12	Description	Shows natural language description of the accident.	No
13	Number	Shows the street number in address field.	Yes
14	Street	Shows the street name in address field.	Yes
15	Side	Shows the relative side of the street (Right/Left) in address field.	Yes
16	City	Shows the city in address field.	Yes
17	County	Shows the county in address field.	Yes
18	State	Shows the state in address field.	Yes
19	Zipcode	Shows the zipcode in address field.	Yes
20	Country	Shows the country in address field.	Yes
21	Timezone	Shows timezone based on the location of the accident (eastern, central, etc.).	Yes

#	Attribute	Description	Nullable
22	Airport_Code	Denotes an airport-based weather station which is the closest one to location of the accident.	Yes
23	Weather_Timestamp	Shows the time-stamp of weather observation record (in local time).	Yes
24	Temperature(F)	Shows the temperature (in Fahrenheit).	Yes
25	Wind_Chill(F)	Shows the wind chill (in Fahrenheit).	Yes
26	Humidity(%)	Shows the humidity (in percentage).	Yes
27	Pressure(in)	Shows the air pressure (in inches).	Yes
28	Visibility(mi)	Shows visibility (in miles).	Yes
29	Wind_Direction	Shows wind direction.	Yes
30	Wind_Speed(mph)	Shows wind speed (in miles per hour).	Yes
31	Precipitation(in)	Shows precipitation amount in inches, if there is any.	Yes
32	Weather_Condition	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)	Yes
33	Amenity	A POI annotation which indicates presence of amenity in a nearby location.	No
34	Bump	A POI annotation which indicates presence of speed bump or hump in a nearby location.	No
35	Crossing	A POI annotation which indicates presence of crossing in a nearby location.	No
36	Give_Way	A POI annotation which indicates presence of give_way in a nearby location.	No

#	Attribute	Description	Nullable
37	Junction	A POI annotation which indicates presence of junction in a nearby location.	No
38	No_Exit	A POI annotation which indicates presence of no_exit in a nearby location.	No
39	Railway	A POI annotation which indicates presence of railway in a nearby location.	No
40	Roundabout	A POI annotation which indicates presence of roundabout in a nearby location.	No
41	Station	A POI annotation which indicates presence of station in a nearby location.	No
42	Stop	A POI annotation which indicates presence of stop in a nearby location.	No
43	Traffic_Calming	A POI annotation which indicates presence of traffic_calming in a nearby location.	No
44	Traffic_Signal	A POI annotation which indicates presence of traffic_signal in a nearby location.	No
45	Turning_Loop	A POI annotation which indicates presence of turning_loop in a nearby location.	No
46	Sunrise_Sunset	Shows the period of day (i.e. day or night) based on sunrise/sunset.	Yes
47	Civil_Twilight	Shows the period of day (i.e. day or night) based on civil twilight.	Yes
48	Nautical_Twilight	Shows the period of day (i.e. day or night) based on nautical twilight.	Yes
49	Astronomical_Twilight	Shows the period of day (i.e. day or night) based on astronomical twilight.	Yes

در جدول زیر نیز می‌توان جزئیات دیتاست تصادفات را به صورت خلاصه مشاهده کرد:

Table 4: US-Accidents: details as of March 2019.

Total Attributes	45
Traffic Attributes (10)	id, source, TMC [33], severity, start_time, end_time, start_point, end_point, distance, and description
Address Attributes (8)	number, street, side (left/right), city, county, state, zip-code, country
Weather Attributes (10)	time, temperature, wind_chill, humidity, pressure, visibility, wind_direction, wind_speed, precipitation, and condition (e.g., rain, snow, etc.)
Period-of-Day (4)	Sunrise/Sunset, Civil Twilight, Nautical Twilight, and Astronomical Twilight
Total Accidents	2,243,939
# MapQuest Accidents	1,702,565 (75.9%)
# Bing Accidents	516,762 (23%)
# Reported by Both	24,612 (1.1%)
Top States	California (485K), Texas (238K), Florida (177K), North Carolina (109K), New York (106K)

هدف پروژه

هدف از انجام این پروژه که ایده‌ی آن برگرفته از مفاهیم داخل منبع ۲ است، پیاده‌سازی Association Rules و استخراج قوانین از درون دیتاست است.

پیش پردازش داده‌ها

این مرحله از حیاتی‌ترین و متناسب با آن زمان‌برترین بخش‌های پروژه است.

ابتدا به بررسی missing valueها پرداخته شد. به همین منظور تعداد کل داده‌های null در ستون‌ها و سطرهای مختلف به صورت مرتب شده محاسبه شدند. در ادامه سیاست‌های مختلفی برای حذف یا جایگزینی داده‌های از دست رفته متناسب با شرایط اتخاذ شد:

۱. ستون‌هایی از داده‌ها که تعداد missing value آن‌ها بسیار زیاد بود و یا اطلاعات مشابه در ستون‌های دیگر وجود داشت حذف شدند.
۲. برخی از مقادیر missing ستون بارندگی یا precipitation با استفاده از اطلاعاتی که از weather condition یا شرایط آب و هوایی داریم تخمین زده شدند.
۳. سایر مقادیر با استفاده از imputation و روش iterative mice مقدار دهی شده‌اند. با توجه به حجم بسیار زیاد داده‌ها و دریافت ارور حافظه، این کار در چندین مرحله انجام شده است. به این صورت که ابتدا مقادیر موجود در df_impute، impute شدند. سپس با کمک df_impute2 مقادیر precipitation به دست آمده‌اند. لازم به ذکر است به علت مقادیر منحصر به فرد زیادی که در ستون Weather_Condition وجود دارند، ابتدا از دستور Feature selection استفاده شده و خروجی‌های آن را در برآورد precipitation استفاده شده‌اند.

Feature Engineering

برخی از ستون‌های موجود در دیتاست به تنهایی آورده‌ی مناسبی ندارند. در این حالت متغیرهای کاربردی را از درون این متغیرها استخراج می‌کنیم. برای مثال از ستون‌هایی که حاوی زمان و تاریخ هستند (مانند Start_Time و End_Time) ماه، سال، ساعت، دقیقه، ثانیه و مدت زمانی که آن رخداد طول کشیده است را به عنوان متغیرهای جدید به دیتافریم خود اضافه می‌کنیم.

برای اعمال association rules در نهایت باید داده‌ها را از حالت عددی خارج کنیم. بنابراین برای مقادیری مانند Time_diff و Time از labelهای مناسب استفاده شده است.

Sampling

برای پردازش داده‌های این دیتاست، به دلیل تعداد رکوردهای بالای آن به سیستم‌های پیشرفته نیاز است. به همین علت با استفاده از نمونه‌گیری رندم از داده‌ها سعی می‌کنیم این چالش را تا حدودی برطرف کنیم. البته شایان ذکر است در ادامه‌ی پروژه نیز قسمت‌هایی وجود دارد که به دلیل نیاز به ظرفیت پردازشی بالا، راهکارهایی برای کاهش ظرفیت موردنیاز انجام شده است. بدیهی است که کار با سیستم‌های پیشرفته‌تر، نتایج بهتری به ما خواهد داد.

Clustering

در این مرحله به منظور dimension reduction یا کاهش ابعاد و عدم امکان استفاده از داده‌های پیوسته در association rules از clustering برای داده‌های مربوط به آب و هوا استفاده شده است که در دیتافریم

df_weather اسامی آن‌ها قابل مشاهده است. silhouette score یکی از ابزارهای به کار رفته در این بخش بوده است.

در ادامه برای تحلیل تعداد کلاسترها برای مثال سه حالت k=3,4,6 را می‌بینیم و در نهایت k=4 انتخاب می‌شود.

	Temperature(F)	Wind_Chill(F)	Humidity(%)	Pressure(in)	Visibility(mi)	Wind_Speed(mph)	Precipitation(in)
0	76.38276	76.24171	42.09599	29.63434	10.04309	8.87879	0.14126
1	37.08000	31.25003	71.65058	29.72438	8.57695	8.81363	0.12894
2	65.34578	64.83711	81.16481	29.81528	8.64554	7.28395	0.11942

K=3

با توجه به تصویر بالا خوشه‌ی اول مربوط به دمای بالا، wind chill بالا، رطوبت و فشار متوسط، دید، سرعت و بارندگی بالاست. خوشه‌ی دوم مربوط به دما، دمای باد و دید کم، رطوبت، فشار و بارندگی متوسط و سرعت باد بالاست. خوشه‌ی سوم دارای رطوبت و فشار بالا، دما و دمای باد متوسط و دید، سرعت باد و بارندگی کم است. دو تصویر پایین نیز به همین ترتیب قایل تفسیر هستند. در نهایت به این نتیجه می‌رسیم داشتن ۴ خوشه علاوه بر این که بازه‌بندی مناسبی از لحاظ عددی برای داده‌ها ارائه می‌دهد، مقیاس خوبی برای مقایسه نیز فراهم می‌کند.

	Temperature(F)	Wind_Chill(F)	Humidity(%)	Pressure(in)	Visibility(mi)	Wind_Speed(mph)	Precipitation(in)
0	82.17081	82.60128	43.69858	29.63962	9.99515	8.85788	0.14072
1	34.71079	28.48872	78.57603	29.71235	8.07534	8.64709	0.12106
2	55.93786	53.48226	45.00065	29.69412	10.13451	8.93287	0.14491
3	65.68061	65.22640	82.61886	29.81762	8.56673	7.21602	0.11857

K=4

۰: داده‌هایی با دما، دمای باد بالا، دید، سرعت باد و بارندگی نسبتا بالا، رطوبت و فشار پایین

۱: داده‌هایی با دما، سرعت باد، دید و بارندگی کم، رطوبت و سرعت باد نسبتا بالا و فشار متوسط

۲: داده‌هایی با دید، سرعت باد و بارندگی بالا، رطوبت کم، دما، دمای باد و فشار متوسط

۳: داده‌هایی با رطوبت و فشار بالا، سرعت باد و بارندگی کم، دید متوسط، دما و دمای باد نسبتا بالا

	Temperature(F)	Wind_Chill(F)	Humidity(%)	Pressure(in)	Visibility(mi)	Wind_Speed(mph)	Precipitation(in)
0	82.63064	82.90157	28.83427	29.43285	10.12190	8.84677	0.13628
1	55.26198	52.74517	46.02325	29.70221	10.12519	8.88782	0.14497
2	79.36619	79.74376	58.48910	29.81570	9.88064	8.77269	0.14247
3	68.75695	68.63151	83.81728	29.82347	8.66272	7.06595	0.12083
4	27.64020	19.43581	72.28993	29.70419	8.30733	9.73330	0.12930
5	48.91875	46.29748	86.08475	29.76167	7.79032	7.19638	0.10840

K=6

پس از استفاده از clustering، داده‌های زیر نیز از دیتاست حذف می‌شوند:

(df_final = df.drop(['Temperature(F)','Wind_Chill(F)','Humidity(%)','Pressure(in)

','’Visibility(mi)','Wind_Speed(mph)','Precipitation(in)'], axis=1)

در نهایت با استفاده از سایر فیچرهای باقی‌مانده می‌توان از association rules استفاده کرد.

Association Rules

در این قسمت با آزمون و خطا مقادیر support و sample را در ماکسیمم حالتی که سیستم اجازه‌ی پردازش آن را می‌دهد محاسبه می‌کنیم. قبل از انجام این کار بر روی دیتاست از دستور get dummies استفاده می‌کنیم تا مقادیر به فرمت دلخواه برای association rules در بیایند.

لیستی از قوانین به دست آمده در فایل اکسل Rules_1 که در پیوست آمده‌اند قابل مشاهده است. در ادامه به بررسی برخی از آن‌ها می‌پردازیم. لازم به ذکر است که در بیان این قوانین ابتدا قوانین بر اساس confidence از بزرگ به کوچک مرتب می‌شوند. در انتها نیز قوانین با leverage منفی را رد می‌کنیم. مقادیر lift بالا قوانین بدیهی هستند. عمدتا liftهای کمتر از ۱ نیز قابل پذیرش نیستند.

۱. در قوانین، به صورت عمده شدت تصادفات ۲ دیده می‌شود. این امر باعث ایجاد شک در مورد این موضوع می‌شود که تصادفات با شدت ۲ بیشترین تعداد دارند. برای اثبات این موضوع به صورت زیر عمل شد:

```
>>> df.groupby(['Severity']).count()
          Unnamed: 0    Side    City  ...  labels  Time_diff  time_intervals
Severity                              ...
1               9180    9180    9180  ...    9180       9180            9180
2             722676  722676  722676  ...  722676     722676          722676
3             283348  283348  283348  ...  283348     283348          283348
4              34796   34796   34796  ...   34796      34796           34796

[4 rows x 30 columns]
```

که همان طور که دیده می‌شود تصادفات با شدت ۲ با تعداد ۷۲۲۶۷۶ بیشترین سهم تصادفات را دارند.

۲. یکی از مواردی که قابل توجه است شباهت بسیار زیاد بین ستون‌های Sunrise_Sunset، Civil_Twilight، Nautical_Twilight، Astronomical_Twilight است. در نتیجه در قوانین به دست آمده‌ای که شامل این فیچرها هستند شباهت بسیاری می‌بینیم. موضوع قابل توجه رخ دادن تصادفات در سمت راست خیابان است. به صورتی که با ترکیب مختلفی از ستون‌های مربوط به گرگ و میش، قوانینی که به سمت راست خیابان اشاره می‌کنند توجه را به خود جلب می‌کنند. برای مثال می‌توان به قوانین زیر اشاره کرد:

304: frozenset({'Astronomical_Twilight_Day', 'Civil_Twilight_Day'}) → frozenset({'Nautical_Twilight_Day', 'Side_R'})

439: frozenset({'Astronomical_Twilight_Day', 'Civil_Twilight_Day', 'Sunrise_Sunset_Day'}) →frozenset({'Nautical_Twilight_Day', 'Side_R'})

177: frozenset({'Severity_2','Civil_Twilight_Day'})→frozenset({'Nautical_Twilight_Day', 'Side_R'})

اگر بعضی از فیچرها مانند شهر و شرایط گرگ و میش حذف شوند، با نمونه‌ی جدید حاصله با تعداد رکورد بیشتر قوانین موجود در فایل اکسل Rules_3 به دست می‌آیند. لازم به ذکر است برای پیدا کردن حالت بهینه لازم بود تا در چند مرحله فیچرها حذف شوند تا با کمک Apriori قوانین جدید استخراج شوند. این چند مرحله‌ای بودن با مشخص کردن stageها به صورت، ۱ ۲ و ۳ در کد قابل مشاهده است. بعضی از این قوانین پس از حذف قوانین با leverageهای منفی و liftهای کمتر از ۱ به صورت زیر هستند:

22 frozenset({'Severity_3'}) → frozenset({'Side_R'})

0 frozenset({'Traffic_Signal'}) → frozenset({'Severity_2'})

67 frozenset({'Astronomical_Twilight_Day','Traffic_Signal'}→ frozenset({'Severity_2'})

160 frozenset({'time_intervals_15-18', 'Astronomical_Twilight_Day'}) → frozenset({'Side_R'})

42 frozenset({'time_intervals_15-18'}) → frozenset({'Side_R'})

33 frozenset({'year_2020'}) → frozenset({'Side_R'})

36 frozenset({'labels_2'}) → frozenset({'Side_R'})

157 frozenset({'Astronomical_Twilight_Day','time_intervals_12-15'})→ frozenset({'Side_R'})

41 frozenset({'time_intervals_12-15'}) → frozenset({'Side_R'})

142 frozenset({'labels_2','Astronomical_Twilight_Day'})→ frozenset({'Side_R'})

111 frozenset({'Time_diff_high', 'Astronomical_Twilight_Day'})→ frozenset({'Severity_2'})

17 frozenset({'Time_diff_high'}) → frozenset({'Severity_2'})

7 frozenset({'Weather_Condition_Fair'}) → frozenset({'Severity_2'})

119 frozenset({'Astronomical_Twilight_Day', 'time_intervals_6-9'})→ frozenset({'Severity_2'})

21 frozenset({'time_intervals_6-9'})→ frozenset({'Severity_2'})

103 frozenset({'year_2019','Astronomical_Twilight_Day'}→ frozenset({'Severity_2'})

94 frozenset({'Astronomical_Twilight_Day', 'Weather_Condition_Clear'})→ frozenset({'Severity_2'})

9 frozenset({'Astronomical_Twilight_Day'})→ frozenset({'Severity_2'})

13 frozenset({'labels_0'})→ frozenset({'Severity_2'})

16 frozenset({'labels_3'})→ frozenset({'Severity_2'})

6 frozenset({'Weather_Condition_Clear'})→ frozenset({'Severity_2'})

121 frozenset({'year_2019', 'Severity_2'})→ frozenset({'Time_diff_high'})

63 frozenset({'year_2018'})→ frozenset({'Time_diff_low'})

197 frozenset({'Time_diff_high''Astronomical_Twilight_Day'})→ frozenset({'Severity_2', 'Side_R'})

تفسیر قوانین استخراج شده تا حد زیادی به هدف و مخاطب قوانین بستگی دارد. بسیاری از قوانین بالا نشان می‌دهند که شرایط مختلفی باعث رخ دادن تصادف در سمت راست خیابان می‌شود که می‌تواند دلایل متفاوتی داشته باشد. از جمله تمایل رانندگان برای منحرف کردن فرمان به سمت راست خیابان، وجود علائمی

که حواس راننده را پرت می‌کنند و ... . بنابراین آگاه کردن رانندگان در این مورد می‌تواند در کاهش نرخ تصادفات موثر واقع شود.

عمده‌ی تصادفاتی که مدت زمان نسبتا طولانی‌ای ادامه داشته‌اند شدت تصادف آن‌ها ۲ بوده است. بنابراین طولانی بودن تصادفات لزوما به معنای سهمگین بودن آن‌ها نیست.

تصادفاتی که در بازه‌ی ۶-۹ صبح رخ می‌دهند عمدتا شدت ۲ دارند و در رده‌ی تصادفات سنگین محسوب نمی‌شوند. بنابراین گزاره‌ی خواب آلودگی صبح زود با تصادفات سنگین ارتباط دارد با این قانون قابل رد است.

تصادفاتی که شرایط آب و هوایی آن‌ها متعلق به label 0 و label 3 هستند (که در قسمت clustering در مورد ویژگی‌های آن‌ها صحبت شد) عمدتا با شدت ۲ اتفاق می‌افتند.

تصادفات سال ۲۰۱۹ عمدتا طولانی بوده‌اند.

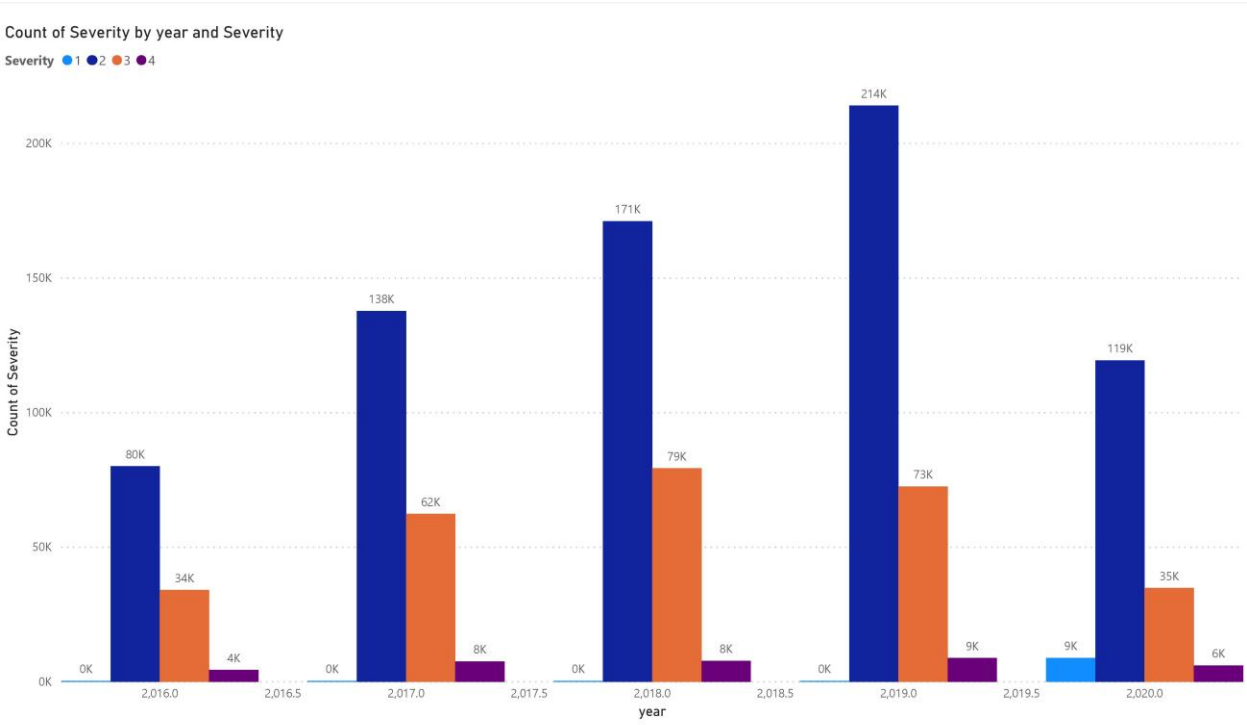
تصادفات سال ۲۰۱۸ عمدتا کوتاه بوده‌اند.

نتیجه‌گیری

در نهایت می‌توان گفت، با استفاده از سیستم‌هایی با ظرفیت بالاتر، انتظار می‌رود با اضافه کردن سایر فیچرها مانند city، street و ...قوانین بیشتر و جالبی از داخل داده‌های این پروژه به دست آورد. همچنین انجام متن کاوی بر روی ستون description اطلاعات با ارزشی چه برای association rules و چه سایر اهداف داده‌کاوی در اختیار مخاطبان قرار خواهد داد.

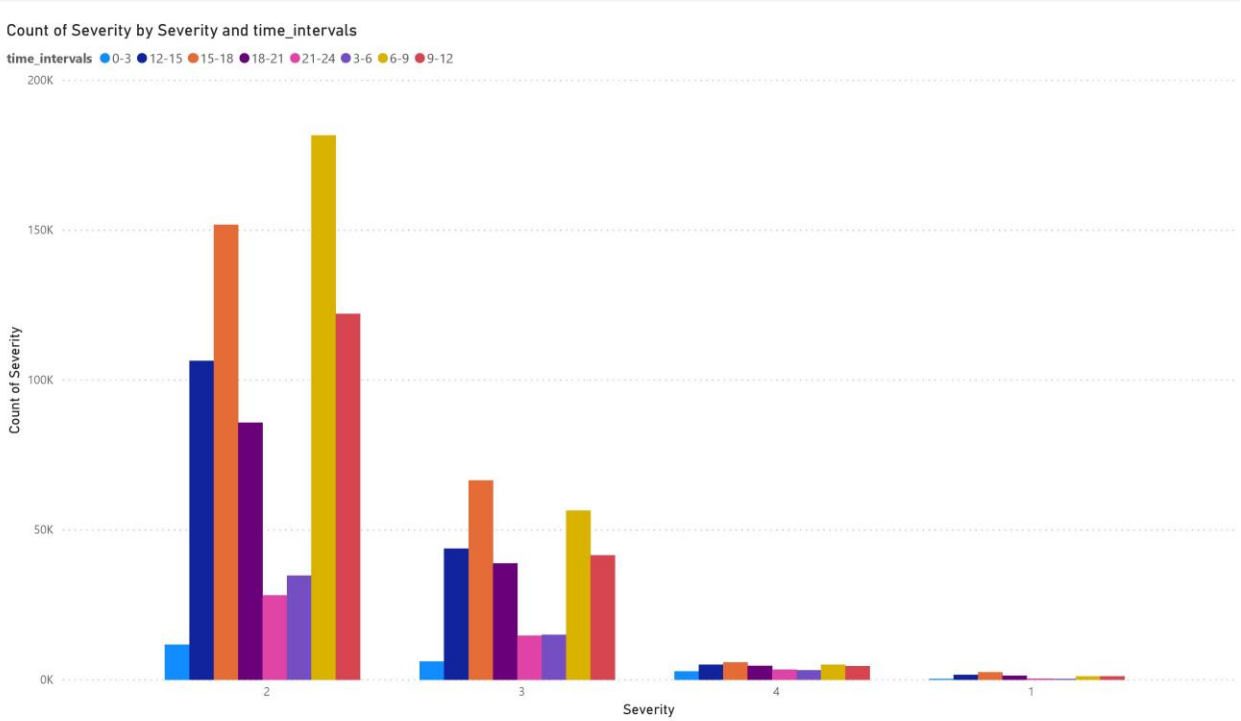
Visualization

در ادامه می‌توان با استفاده از مصورسازی، به اطلاعات جالبی در مورد داده‌ها دست پیدا کرد. به این منظور نمودارهای منتخب که با استفاده از Power BI رسم شده‌اند ارائه می‌شود.



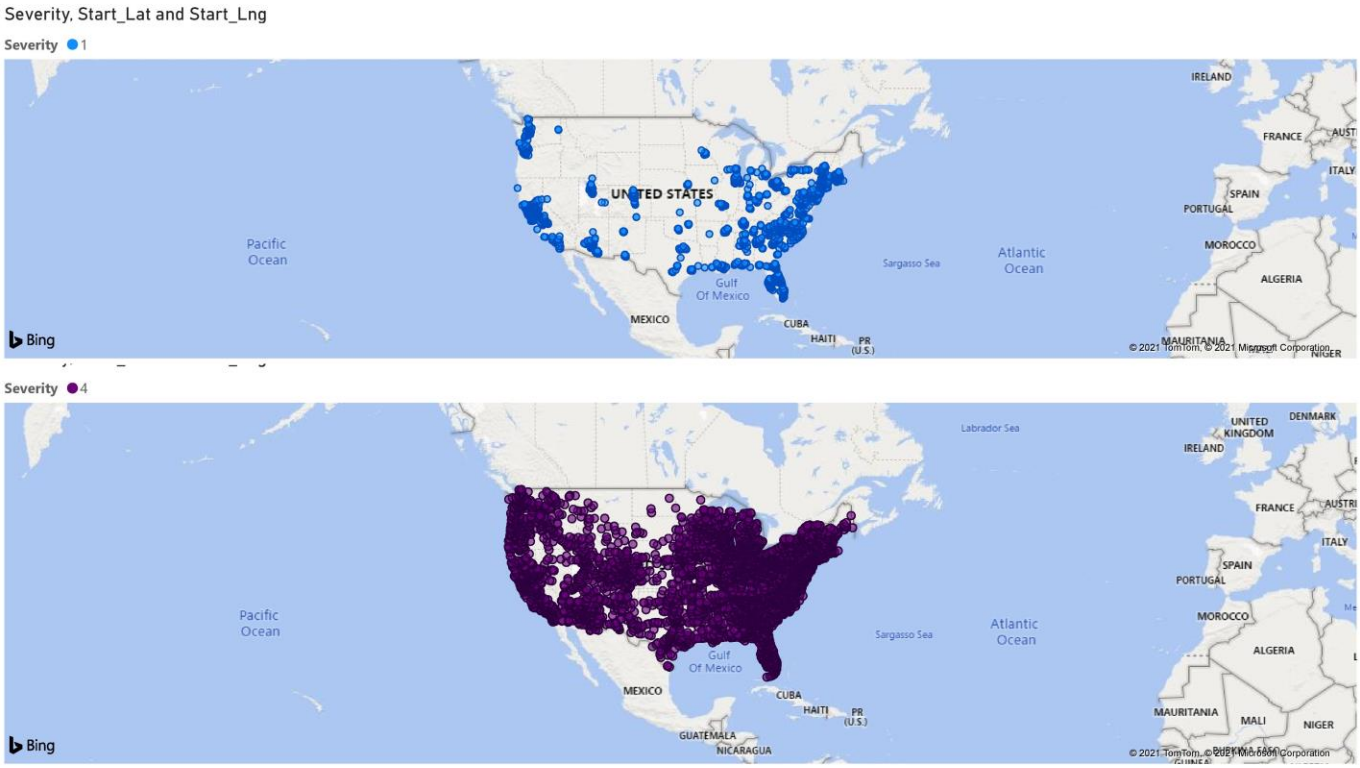
نمودار شدت تصادفات بر حسب سال

همان طور که دیده می‌شود طی سال‌ها تصادفات عمدتا سیر صعودی داشته‌اند و بیشترین شدت تصادفات نیز ۲ بوده است. اما کاهشی در روند تصادفات در سال ۲۰۲۰ دیده می‌شود که علت آن می‌تواند کم‌تر شدن ترددها به سبب قرنطینه و محدودیت‌ها باشد.

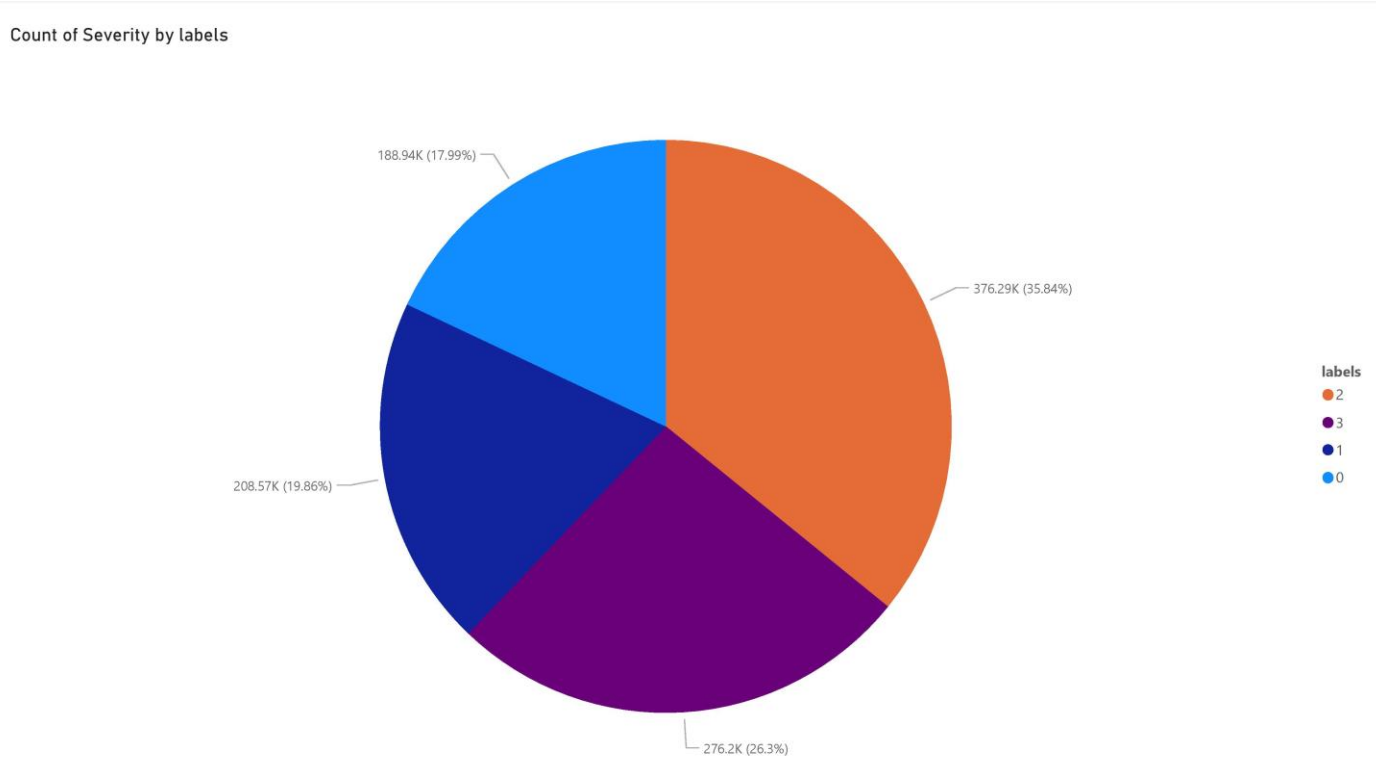


نمودار شدت تصادفات در ساعات مختلف شبانه‌روز

با استفاده از نمودار بالا می‌توان دید در هر بازه‌ی زمانی بیشترین شدت تصادفات به چه صورت است.

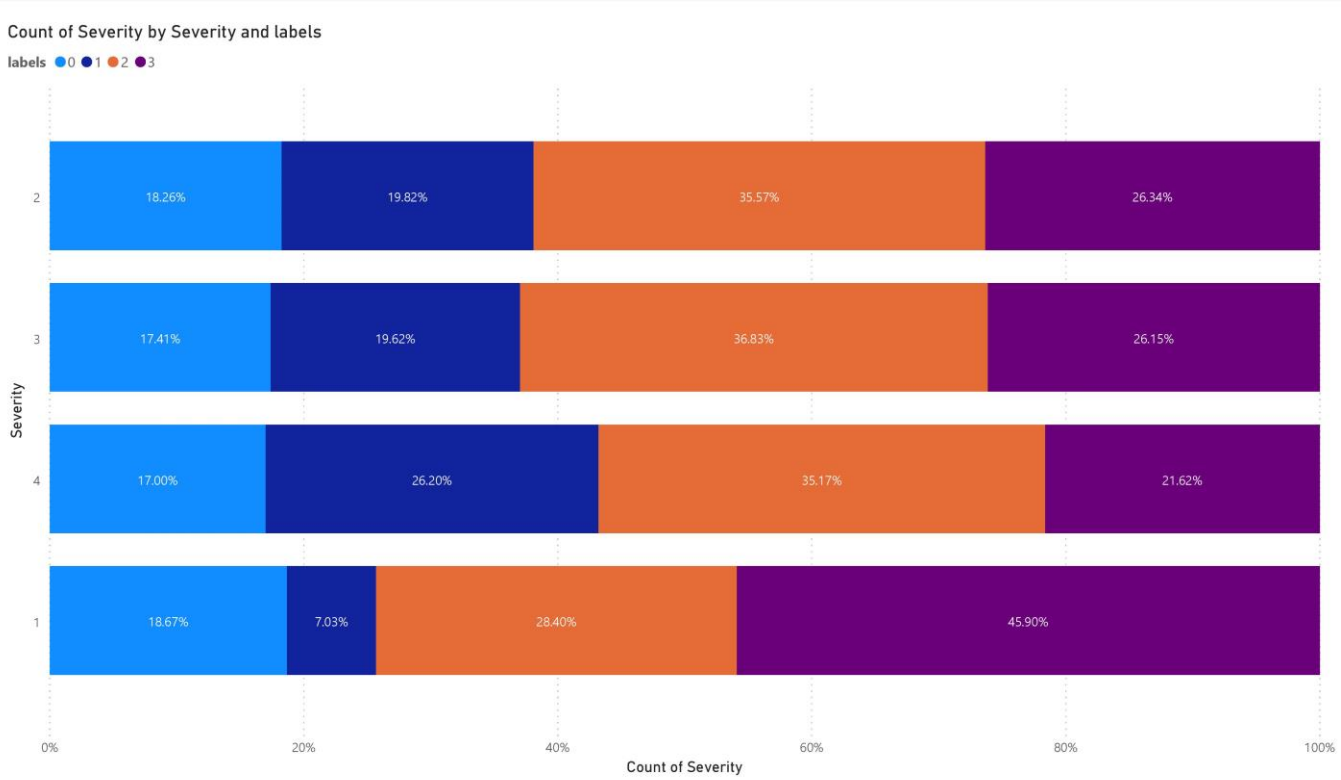


توزیع شدت تصادفات ۱ و ۴ در آمریکا



توزیع تصادفات بر حسب آب و هوا

بیشترین تصادفات در شرایط آب و هوایی که در label 2 آن را مشخص کرده‌ایم اتفاق می‌افتند.



توزیع شدت تصادفات بر حسب آب و هوا

هر کدام از نوارهای افقی نمودار بالا نماینده‌ی یک شدت تصادف از ۱ تا ۴ هستند. بخش‌بندی‌های موجود در هر نوار نشانگر سهم انواع آب و هواهای مشخص شده با labelهای ۰ تا ۳ هستند.

منابع

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, arXiv preprint arXiv:1906.05409 (2019).

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. “Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights.” In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.