**对于英国交通事故数据的**

**可视化分析及相关性研究**

**李芳 10214602404**

**华东师范大学**

目录

1. 实验引言................................................................................. 3
2. 摘要............................................................................................................. 3
3. 研究背景及方法......................................................................................... 3
   1. 研究背景.............................................................................................. 3
   2. 研究思路.............................................................................................. 4
   3. 研究方法.............................................................................................. 5
4. 车祸数据处理......................................................................... 6
5. 数据整体观测............................................................................................... 6
6. 数据处理及特征化....................................................................................... 6
7. 整体数据可视化..................................................................... 9
8. 车祸严重程度占比分析............................................................................... 9
9. 事故人受伤程度占比分析........................................................................... 10
10. 事故数据相关性研究..............................................................12
11. 交通事故严重程度相关因素可视化............................................................ 12
12. 以速度为基准，其余因素相关性可视化.................................................... 14
    1. pedestrain control................................................................................. 15
    2. junction control.................................................................................... 17
    3. light conditions..................................................................................... 19
    4. carriageway hazards.............................................................................. 21
13. 事故相关性总结............................................................................................ 23
14. 后续完善................................................................................. 24
15. 总结......................................................................................... 24
16. **实验引言**
17. **摘要**

车祸已经成为当今社会公害，为城市人口死亡的四大原因之一。此次案例分析的数据为2019年英国发生的所有车祸及其相关的道路因素，包括光照条件，道路等级、道路类型，路口控制等因素。经过数据可视化呈现以及相关性因素的研究，车祸严重程度可能与多种自然因素和社会因素有关，主要利用图标模型进行探索。

关键词：车祸；数据可视化；交通事故；相关性；相关分析

1. **研究背景及方法**

2.1. 研究背景

交通事故已是世界性的严重社会问题。惨重的交通事故后果使人们不得不对交通安全状况予以高度重视，并将不断进步的科学技术应用于交通安全研究工作中，使汽车更好地造福于人类。目前，各国正在不同程度地从道路和交通工程上采取对策。

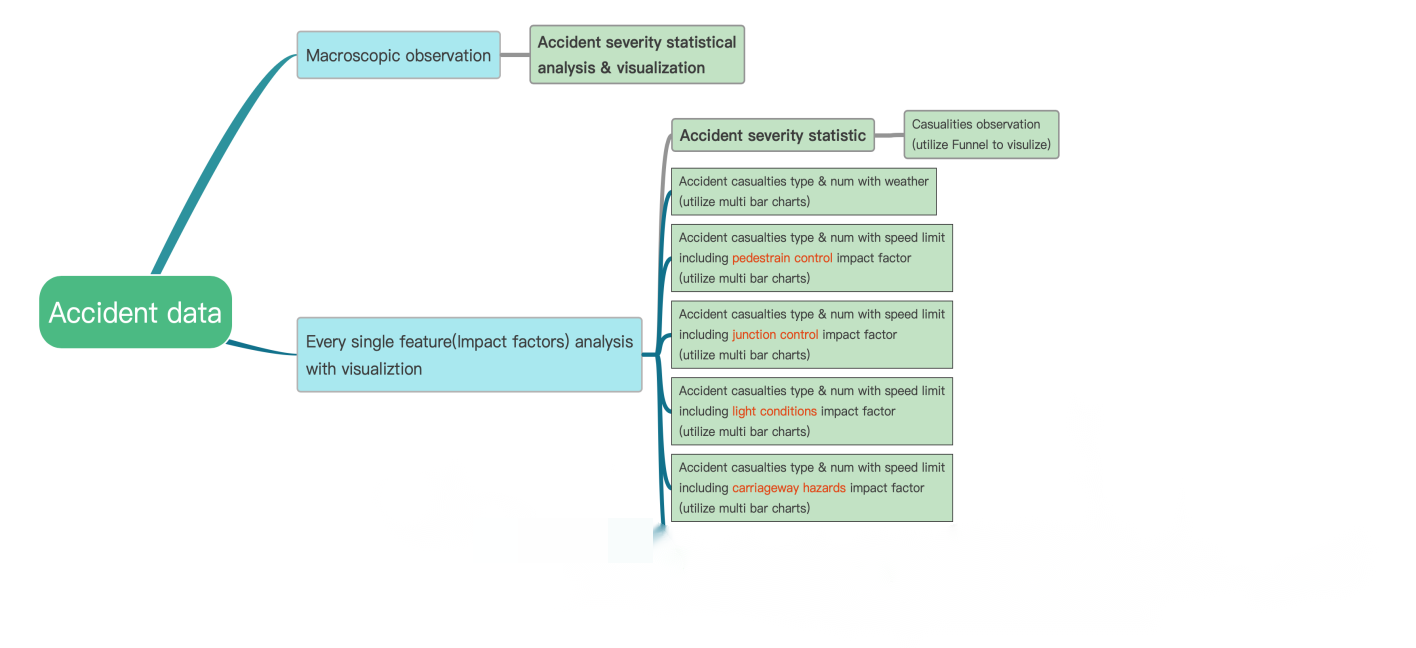
车祸往往是由司机的疏忽大意造成的。随着世界交通运输的不断发达，车辆的日益增多，交通事故也逐渐增多。道路交通事故的原因很复杂，涉及人、车、路等多方面因素。机动车驾驶人的驾驶水平、非机动车驾驶人和行人的道路交通安全意识、交通出行环境、道路交通规划的合理程度，以及道路交通标志标线是否完善等因素。本实验主要对交通事故影响因素进行了分析研究。此次案例分析的数据为2019年英国发生的所有车祸及其相关的道路因素，包括光照条件，道路等级、道路类型，路口控制等因素。经过此次实验数据可视化分析以及相关性因素的研究，希望能对交通安全的重视性有所提高，并且在事故发生前的风险规避以及事故发生后的应急处理措施有所启发，对交通资源的管理与建设能够有相关性的裨益。

2.2. 研究思路

由于实验数据的原始性和杂乱性，拿到数据集之后，首先进行数据的观测、清洗以及特征选择。

然后对于清洗完毕的实验数据可视化分析主要分为两大部分，即宏观观测以及可视化操作，其中可视化操作分为整体数据的可视化以及与车祸相关因素相关性的可视化分析。

This is processed logic and procedure of this project:



1. .数据的宏观观测

对于数据进行清洗时，就可以对表格中的原始数据进行初探包括数据的总量以及数据的种类，即车祸数据序号，车祸时间，车祸地点经纬度，警察局报警站点，以及事故发生时的各种环境、人为因素包括天气、道路拥堵状况、行人类型、限速、灯光等等

1. .数据可视化操作

对于数据的可视化操作分为两个阶段：第一阶段，整体数据的可视化观测，即根据交通事故的严重程度进行观测，然后再对事故人的伤亡情况进行可视化操作。

第二阶段，对车祸事故的相关性因素进行研究。首先摘取气候条件进行图表绘制，然后以速度为基准，分别以pedestrain control、junction control、light conditions、carriageway hazards四个因素进行分类直方图的绘制。

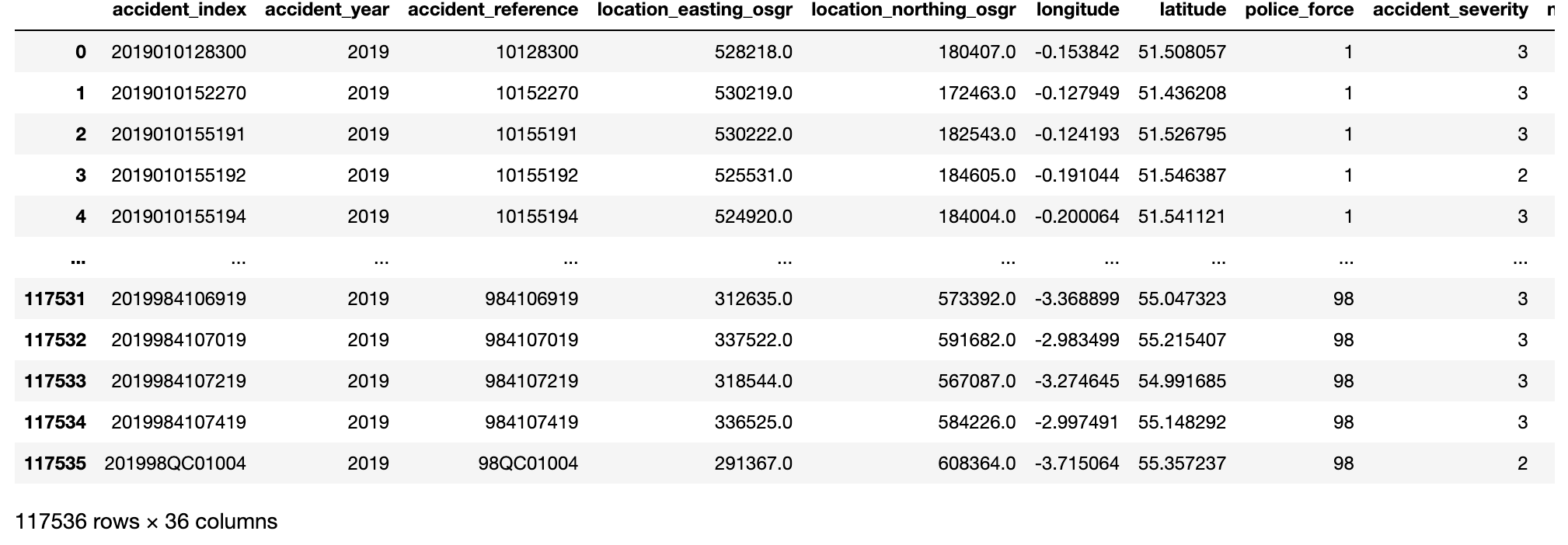
2.3.研究方法

首先，对于数据的清洗和处理，利用Numpy和Pandas库，导入csv文件的数据后，主要用info（）函数进行快速探索和透视、drop（）函数进行去空、去重复以及特征工程筛选。

其次，对于数据的可视化图表绘制，利用matplotlib、seaborn、plotly

库，摘取特定的数据，绘制饼状图、沙漏图、柱状图以及多种分类直方图进行可视化呈现。

1. **车祸数据处理**
2. **数据整体观测与审查**

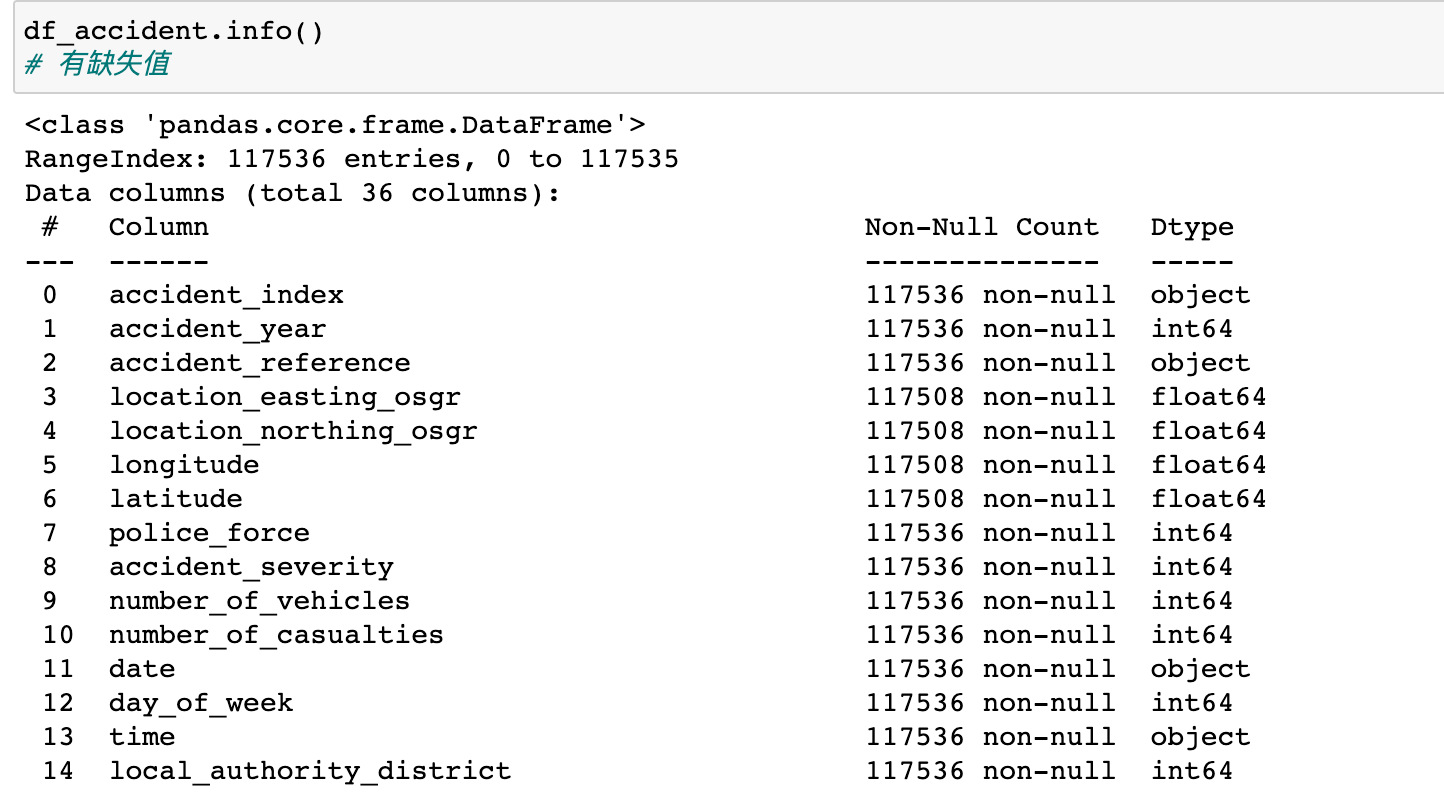


数据观测：此命名为data的表格有117563条数据量，特征有36个.

表内内容和特征含义：是关于车祸事件的描述，从各种特征的维度来描述车祸的状况。其大致特征有：车祸发生的事件号，车祸的严重程度、发生事故的交通工具的数量、发生事故的位置、经纬度、事故人的受伤程度、路段限速，以及其本事故发生时的天气和道路类型例如对人行横道是否有人为控制、人行横道是否有物理设施、第二路类、第二路号、光照条件、路表状况、车道危险、是否有警察参与维持秩序等等。

1. **数据处理及特征化**
2. . info()函数处理

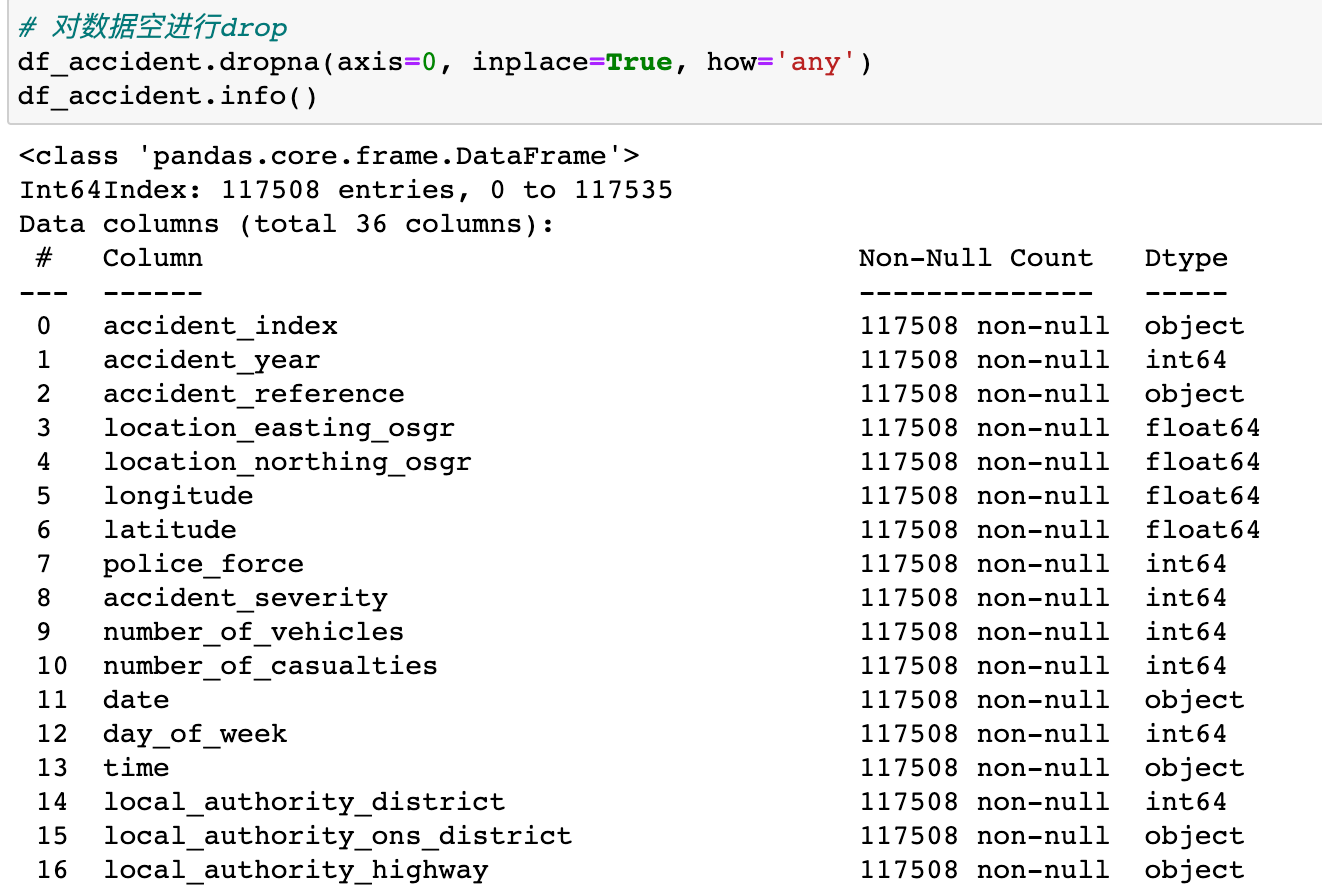
对于未经处理的实验数据，先使用Pandas库自带的info（）函数进行简单透视观测，了解大致的数据类型了解以及缺失值检测，它的速度更快，探索性更高。



可以看出，数据集部分特征中有明显的缺失值，同时发现数据中有部分-1的值，应当都作为空值清洗掉，防止后续进行可视化操作时影响图表精确程度，并且有空值的数据条不在此次主要研究的对象中，以防有记录错误或者漏除的影响。

1. .nan drop

对含有空值的数据借助dropna（）函数进行初次清洗，将详细信息有缺失的数据进行删除，然后再次利用info（）函数进行探测，看是否清理干净，确保空信息的数据不会对后续可视化分析操作造成影响。



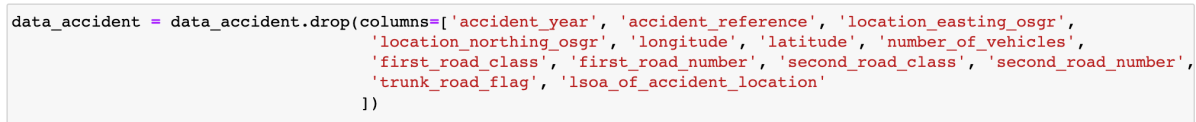
1. .clean duplicated value

根据数据的事故索引进行去重复值处理，避免有重复值存在影响后续可视化分析

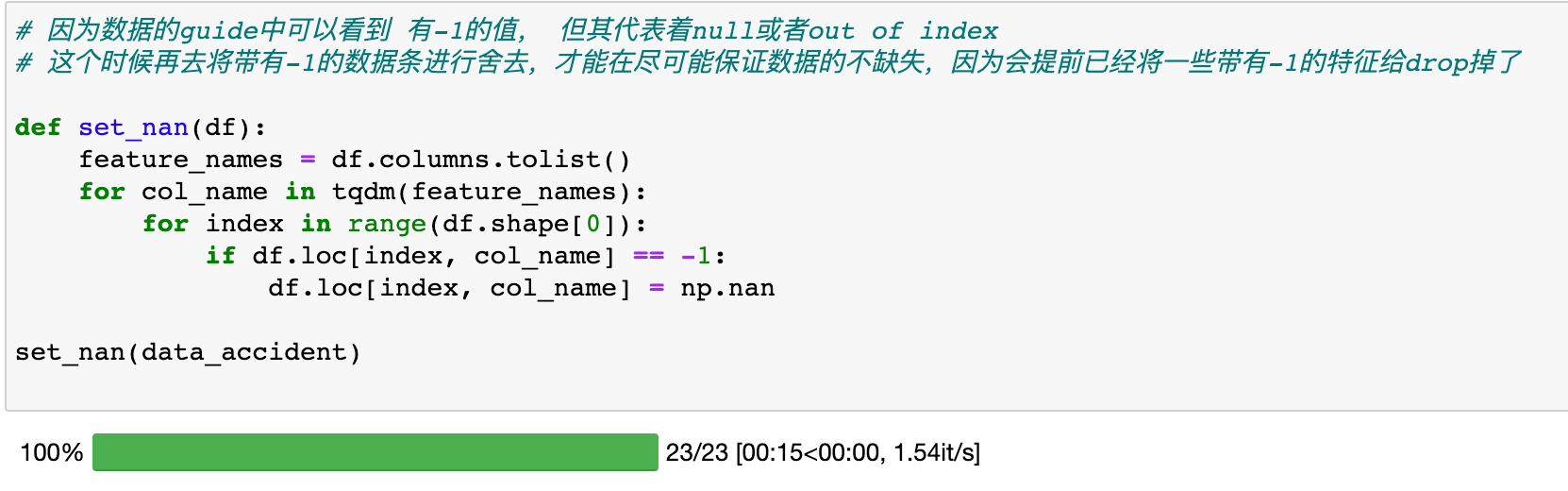


1. .feature engineering

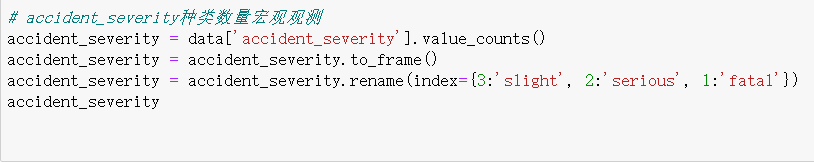
特征选择，删除掉一些无用特征



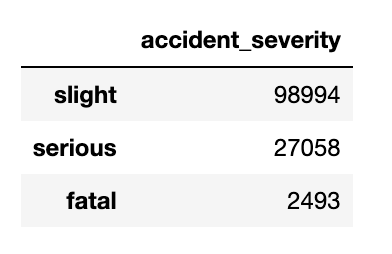
将特征中为-1的值替换成np.nan，然后使用dropnan去进行删除



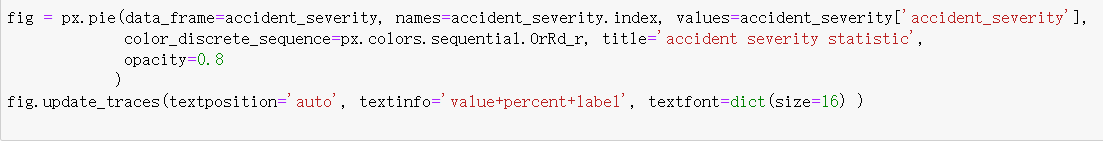
1. **整体车祸事故数据可视化**
2. **车祸严重程度占比分析**



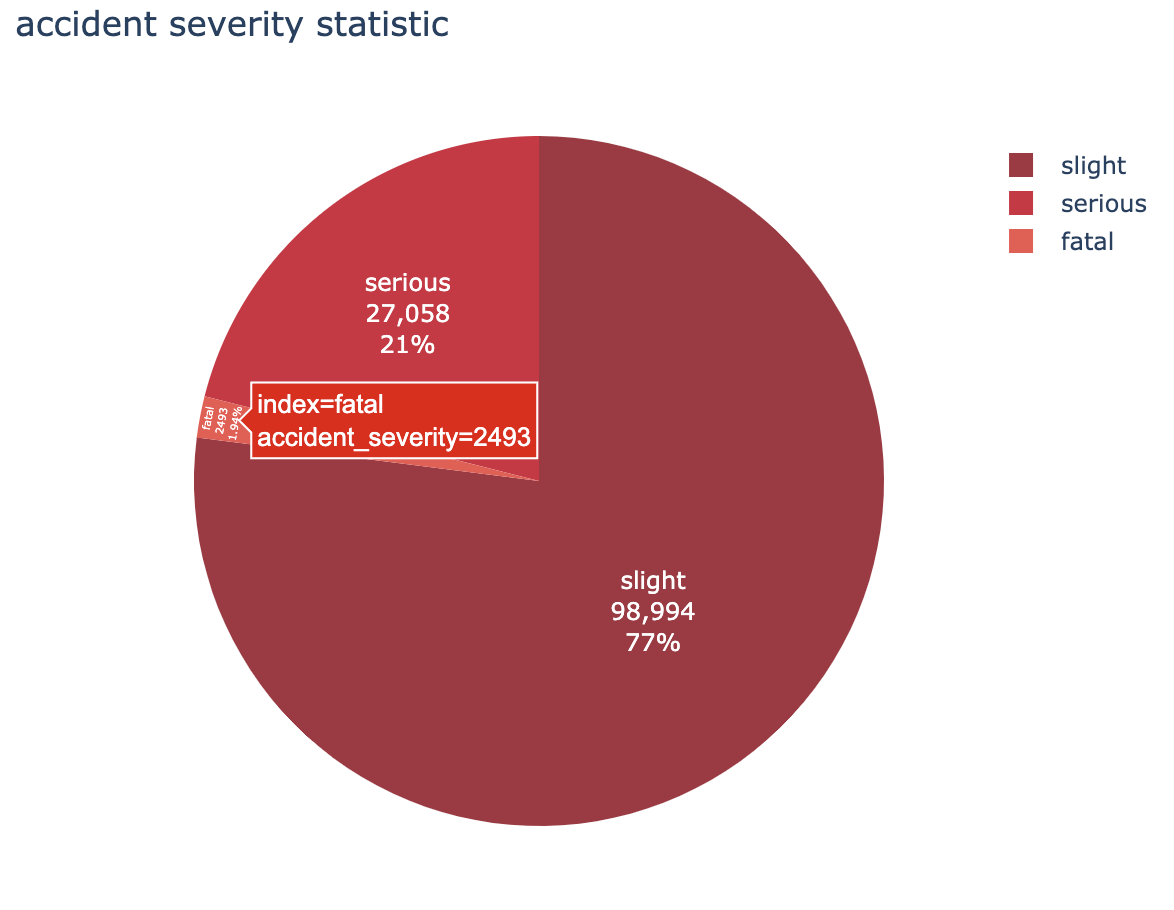
通过数据过滤的方式取出特征名称为’accident\_severity’这一列的数据，对此列的数据进行类别和其对应的数量进行统计。此特征中有：1，2，3三个类别，通过’guide.csv’中的特征解释做研究，然后对其三个类别进行名称更改为：fatal，serious，slight。更改后的数据如下：



对accident\_severity 统计表格进行可视化：



绘制出的饼状图如下：

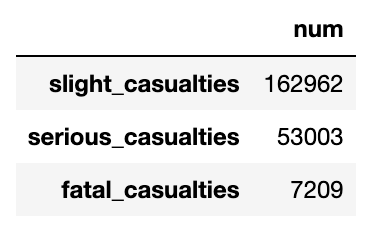


分析出：轻微车祸占到全部车祸数据的77%，严重车祸占比21%，致命性重大车祸占比1.94%。

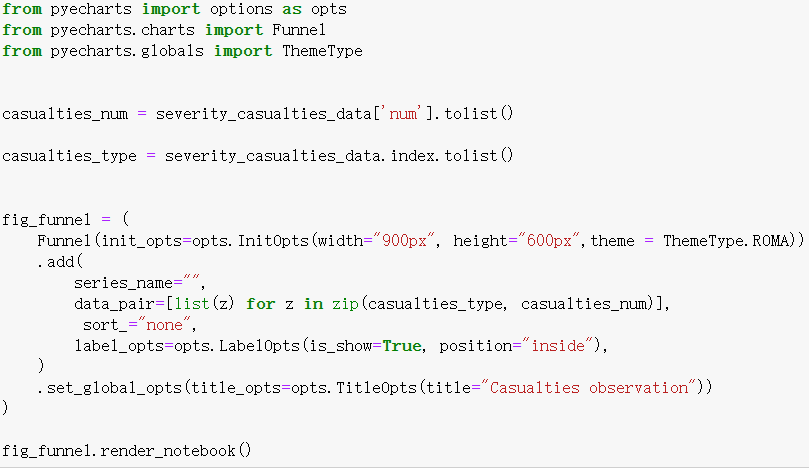
1. **事故人受伤程度占比分析**



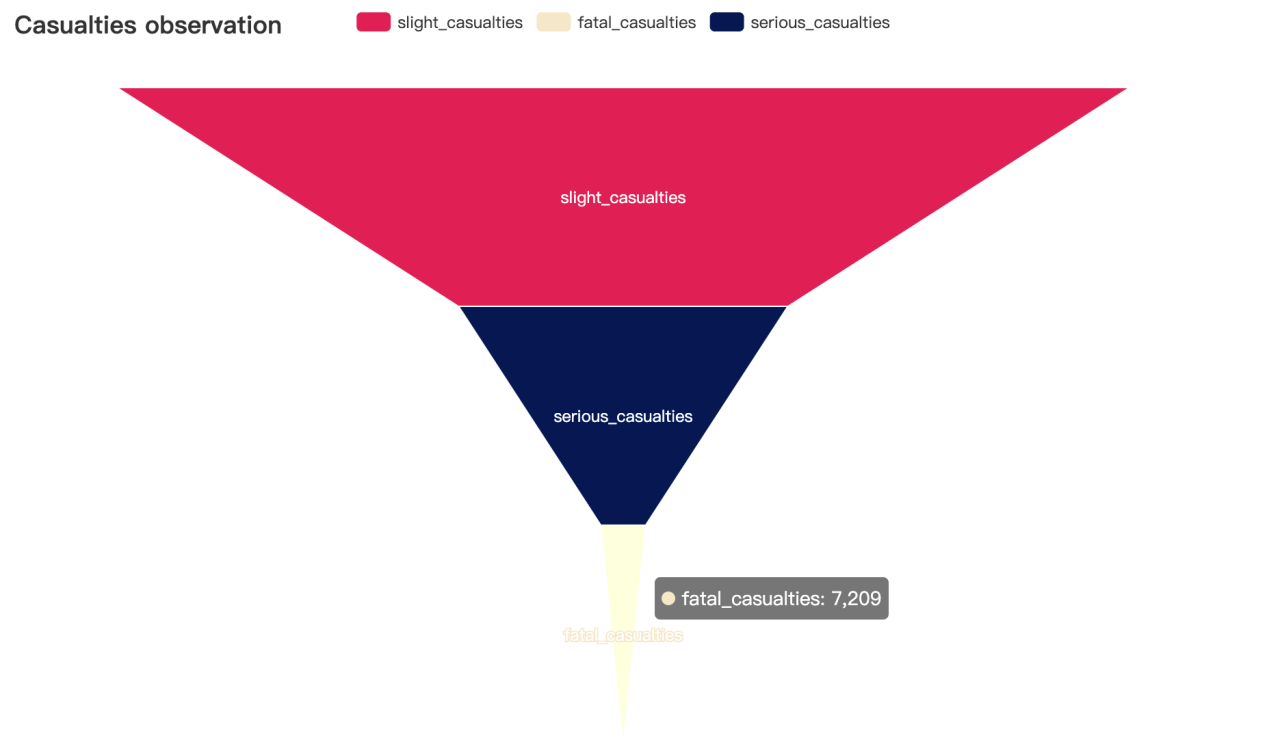
数据表格中的名为‘accident\_severity’这一特征中，数据的类别有：1，2，3，分别对应着轻伤-’slight\_casualties’, 重伤-’serious\_casualties’, 死亡-‘fata\_casualties’，进行名称更改。同时，将这些类别的数据分别取出，并且过滤出特征’number\_of\_casualties’，对这些各个类别的伤害程度的类别和数量进行统计：



通过Funnel漏斗图的方式对其进行可视化：



得出的沙漏图如下：

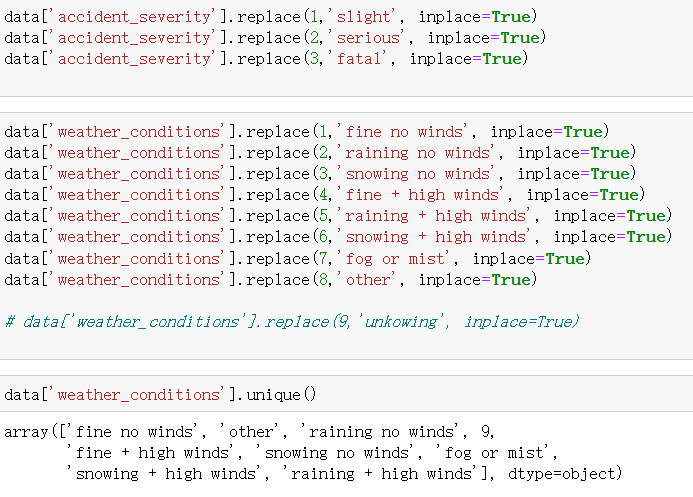


从此图中可以分析出：死亡人数虽占比数量最少，但是7209的死亡数量仍然需要深度关注，并且后期重点分析造成死亡的车祸的各类原因。其次，重伤人数和轻伤人数分别为：53003和162962。所以交通安全仍应引起注意，即使是轻型事故也可能会引发人员命危和伤亡。

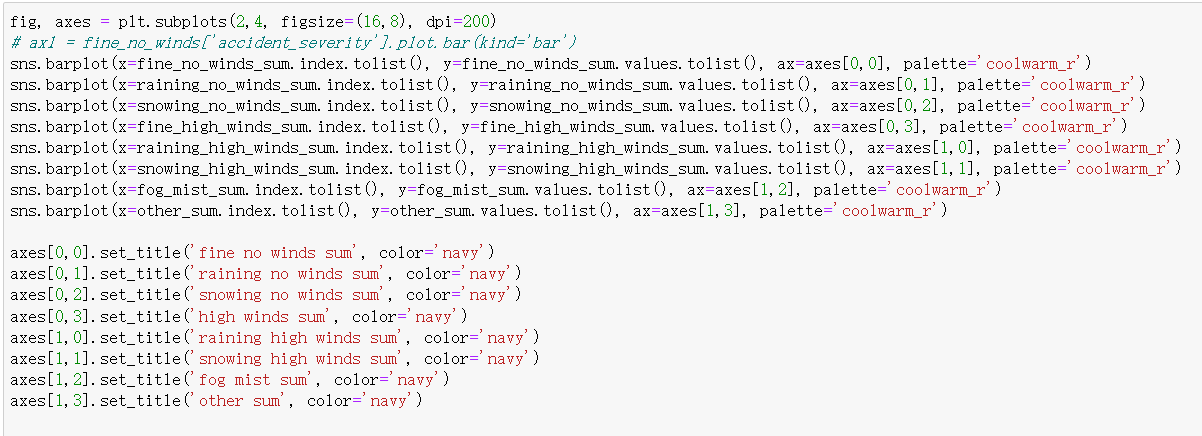
1. **事故数据相关性研究**
2. **车祸交通事故严重程度相关因素可视化**

Establish the relationship model between accident severity and outer impact factor weather

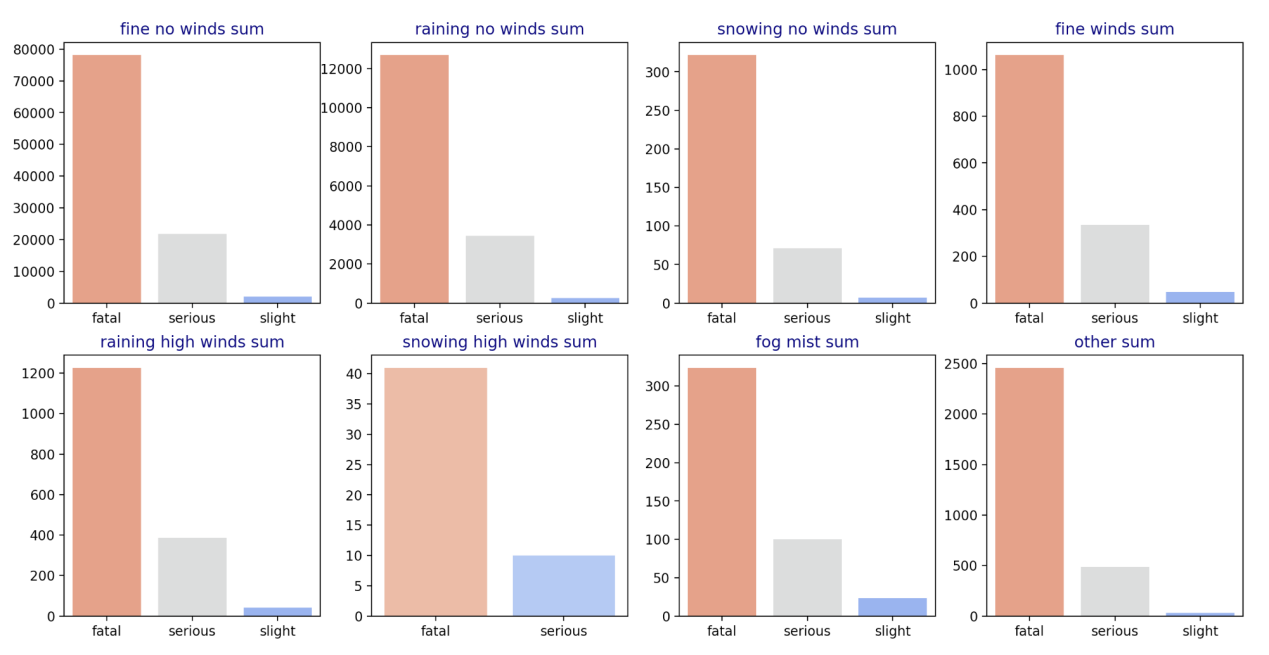
特征‘weather\_condition’中的数据类别分别为1、2、3、4、5、6、7、8，分别代指为‘fine no winds’,’raining no winds’, ‘snowing no winds’, ‘fine + high winds’, ‘raining + high winds’, ‘snowing + high winds’, ‘fog or mist’, ‘other’，并且对这些类别名称进行更改。



同时，将数据按照这八个类别进行分类，对每个类别的数据中的‘accident\_severity’进行分类统计，并进行可视化：



得出的柱状图如下：



分析得知：排除正常天气发生的事故造成的死亡（或由其他原因造成），雨天、大风的天气造成的车祸死亡和严重受伤的数量都是最多的。其次，在其他类别的天气中，车祸造成死亡的数据量高达2500，可见这类别的天气中也包含了一些相对恶劣的天气。

1. **以速度为基准，其余因素相关性可视化**

建立以限速为基准的分类绘图模型，继而对剩余的四种相关因素进行套用。主要操作大同小异，先对不同相关因素进行中文的代替处理，方便后续的建轴和情况分明，然后进行分类柱状图的绘制

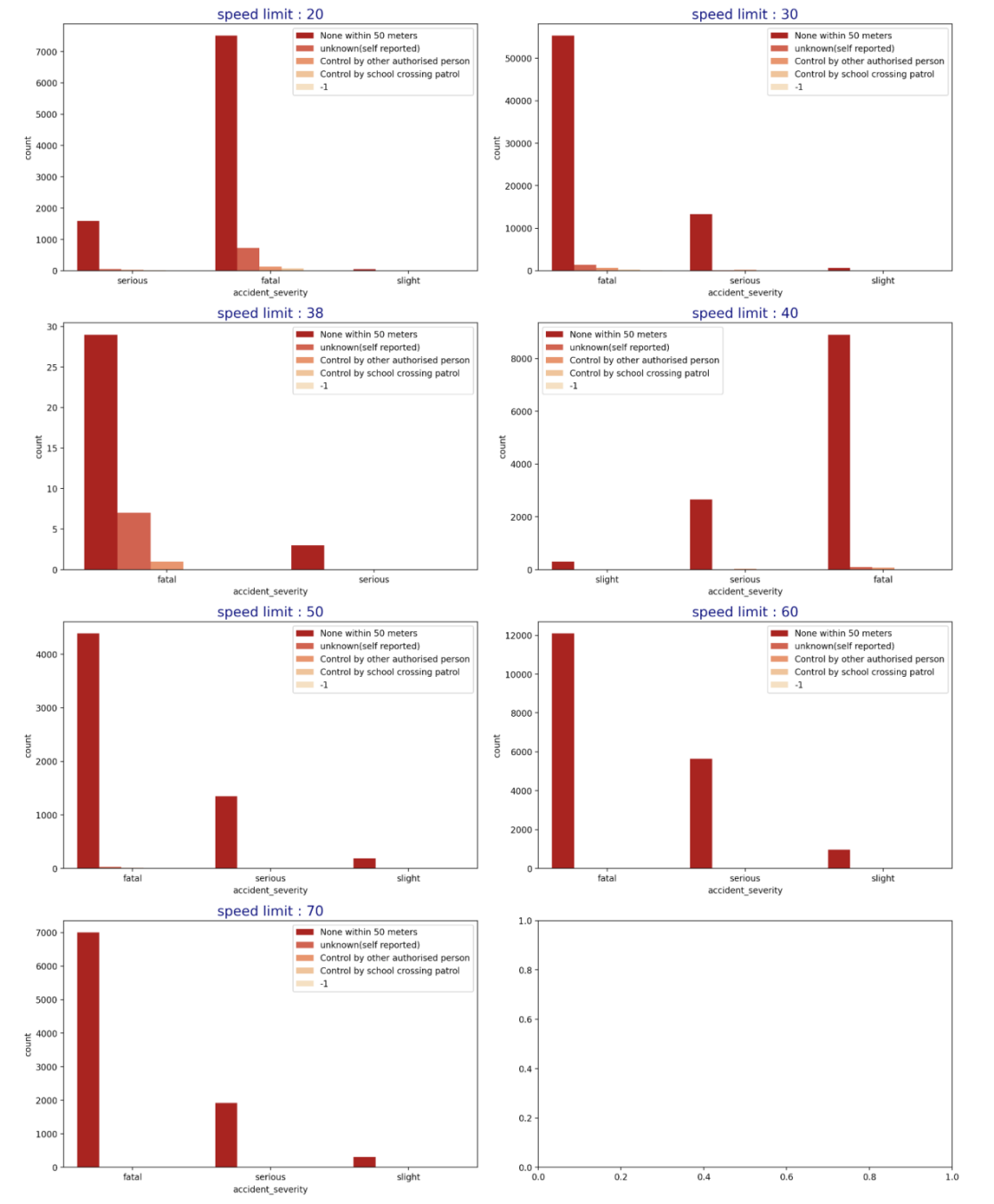




* 1. Establish the relationship model between speed limit and accident severity including impact factor pedestrain control

将总数据中的特征’pedestrain\_crossing\_human\_control’中的类别：0、1、2和9进行命名更改，分别更改为：‘None within 50 meters’、‘Control by school crossing patrol’、‘Control by other authorised person’、‘unknown(self reported)’。

将特征‘speed\_limit’作为数据分类的指标，此特征类的数据种类有：20、30、38、40、50、60和70，按照限速设置对总数据进行划分七部分子数据。同时，各类限速指标和车祸伤亡建模，并可视化，同时将‘pedestrain\_crossing\_human\_control’特征中各类别加入进去：

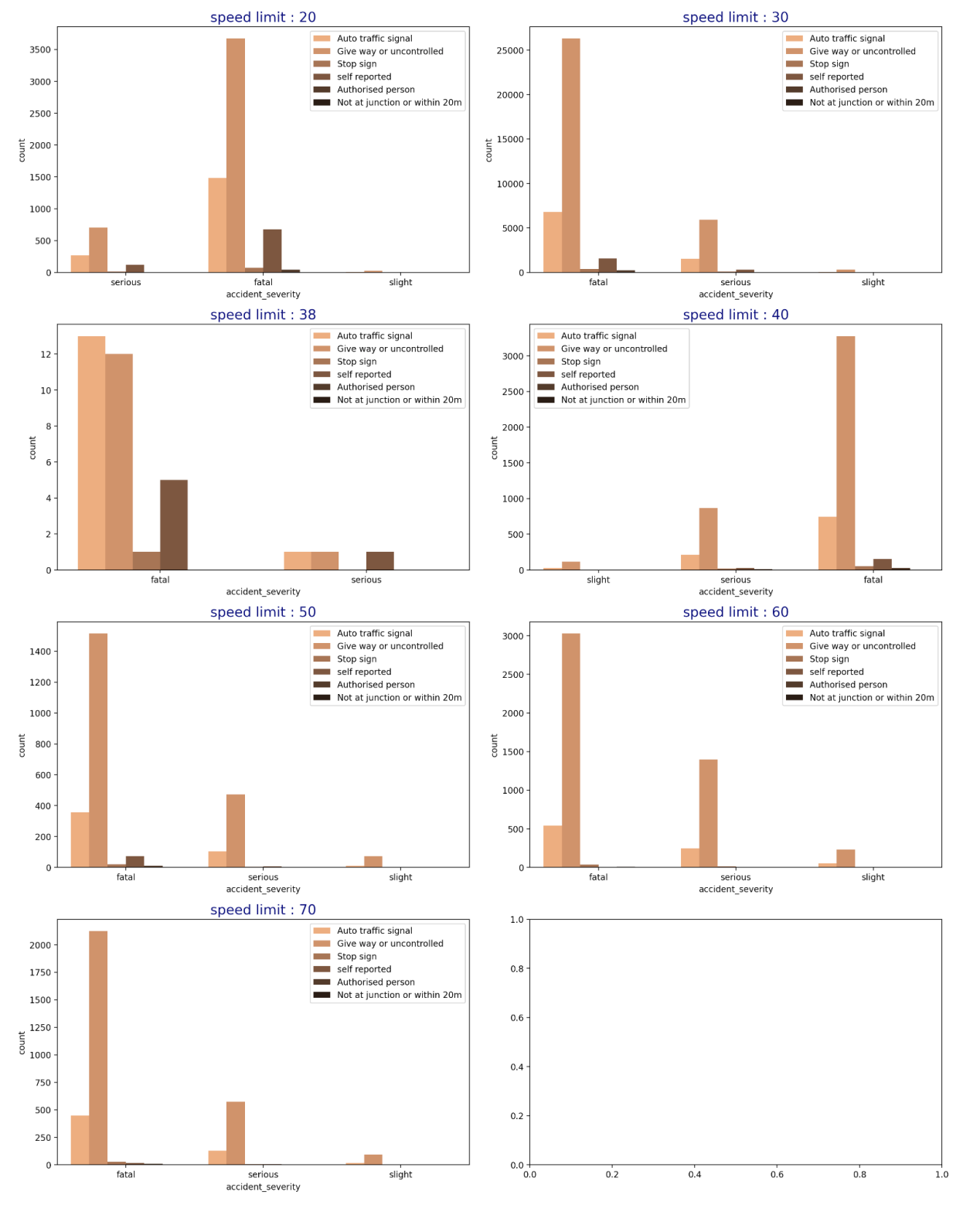


总结分析：在七个速度限制的分析柱状图中，都明显的可以看到，无论是致死、严重还是轻微的车祸中，在交叉路口外五十米内没有设置任何的任性道提醒标识是都是造成车祸的巨大祸端。相反，我们可以看到在‘Control by other authorised person’和‘Control by school crossing patrol’这两个影响因子限制下，车速限制从20-70的图表中都显示出有很少量的车祸发生。以此可见，在人行道的远处设置标识来以告知前方有行人通过，从而使得车主在远处就降低车速是降低车祸发生概率的一个最大的举措。

* 1. Establish the relationship model between speed limit and accident severity including impact factor junction control

将总数据中的特征’junction\_control’中的类别：0、1、2、3、4和9进行命名更改，分别更改为：‘Not at junction or within 20 meters’、‘Authorised person’、‘Stop sign’、‘Give way or uncontrolled’和‘unknown(self reported)’。

将特征‘speed\_limit’作为数据分类的指标，此特征类的数据种类有：20、30、38、40、50、60和70，按照限速设置对总数据进行划分七部分子数据。同时，各类限速指标和车祸伤亡建模，并可视化，同时将‘junction control’特征中各类别加入进去：

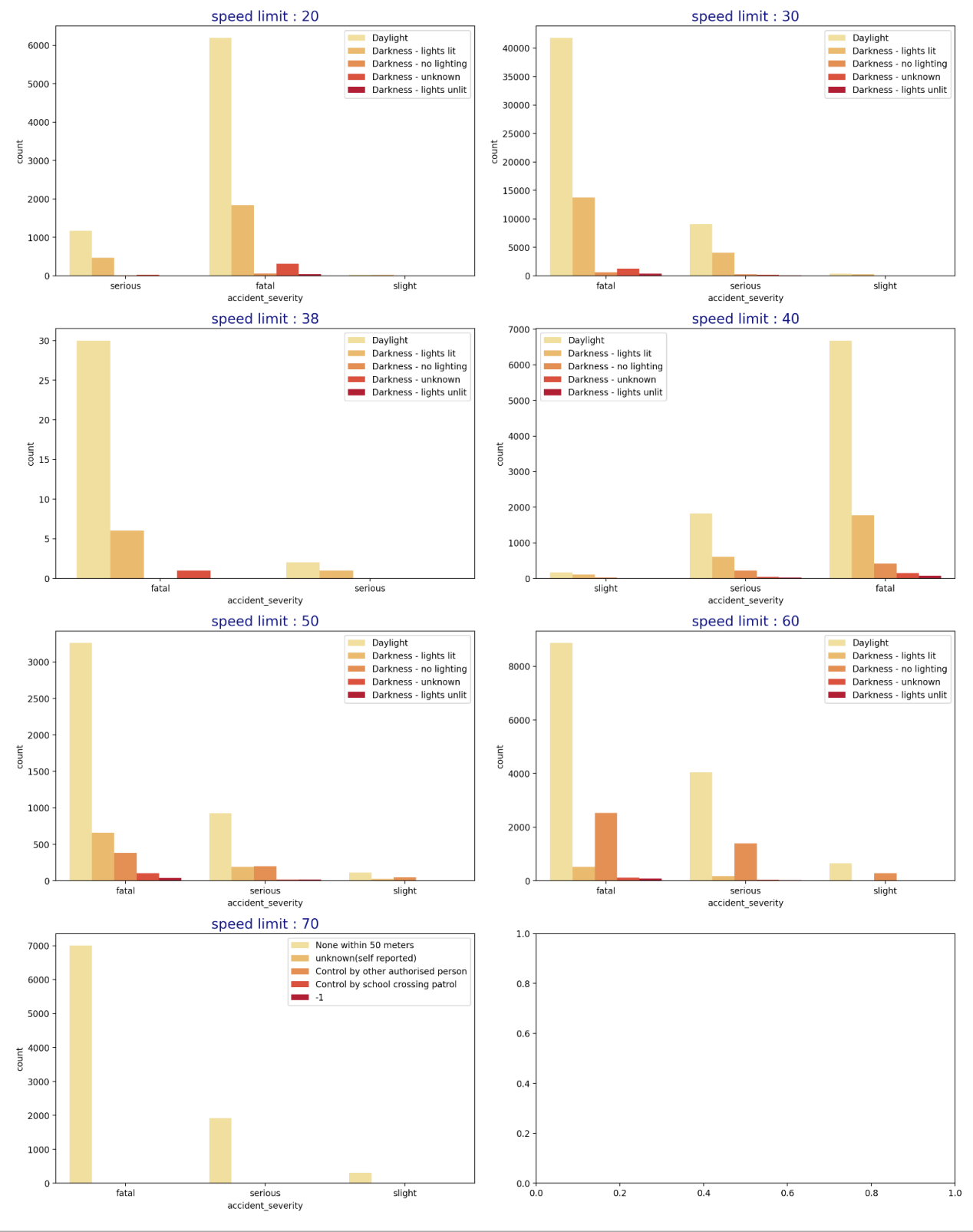


在速度限制从20-70的图表中，观测车祸导致死亡的这一统计，可以发现，当有车辆堵塞时，’Give way or uncontrolled’的这一举措指标的影响是最大的，也是导致车祸死亡量最具有影响的一个因子，可见‘快车先行’和‘无任何的管控车辆的措施’是造成车祸的最大的因素；其次，‘Auto traffic signal’对于车辆堵塞的情况，在各类速度限制的前提下，也会导致大量的车祸。相反，‘Stop sign’和‘Authorised person’在车辆堵塞的情况下，会大大降低车祸发生率。可见，明显的有明显的停车标识和在政府相关人员管控的情况下，在发生车辆堵塞时，车祸率会降低。

* 1. Establish the relationship model between speed limit and accident severity including impact factor light conditions

将总数据中的特征’light conditions’中的类别：1、4、5、6和7进行命名更改，分别更改为：‘Daylight’、‘Darkness - light lit’、‘Darkness - lights unlit、‘Darkness - no lighting’和‘Darkness - unknown(self reported)’。

将特征‘speed\_limit’作为数据分类的指标，此特征类的数据种类有：20、30、38、40、50、60和70，按照限速设置对总数据进行划分七部分子数据。同时，各类限速指标和车祸伤亡建模，并可视化，同时将‘light conditions’特征中各类别加入进去：

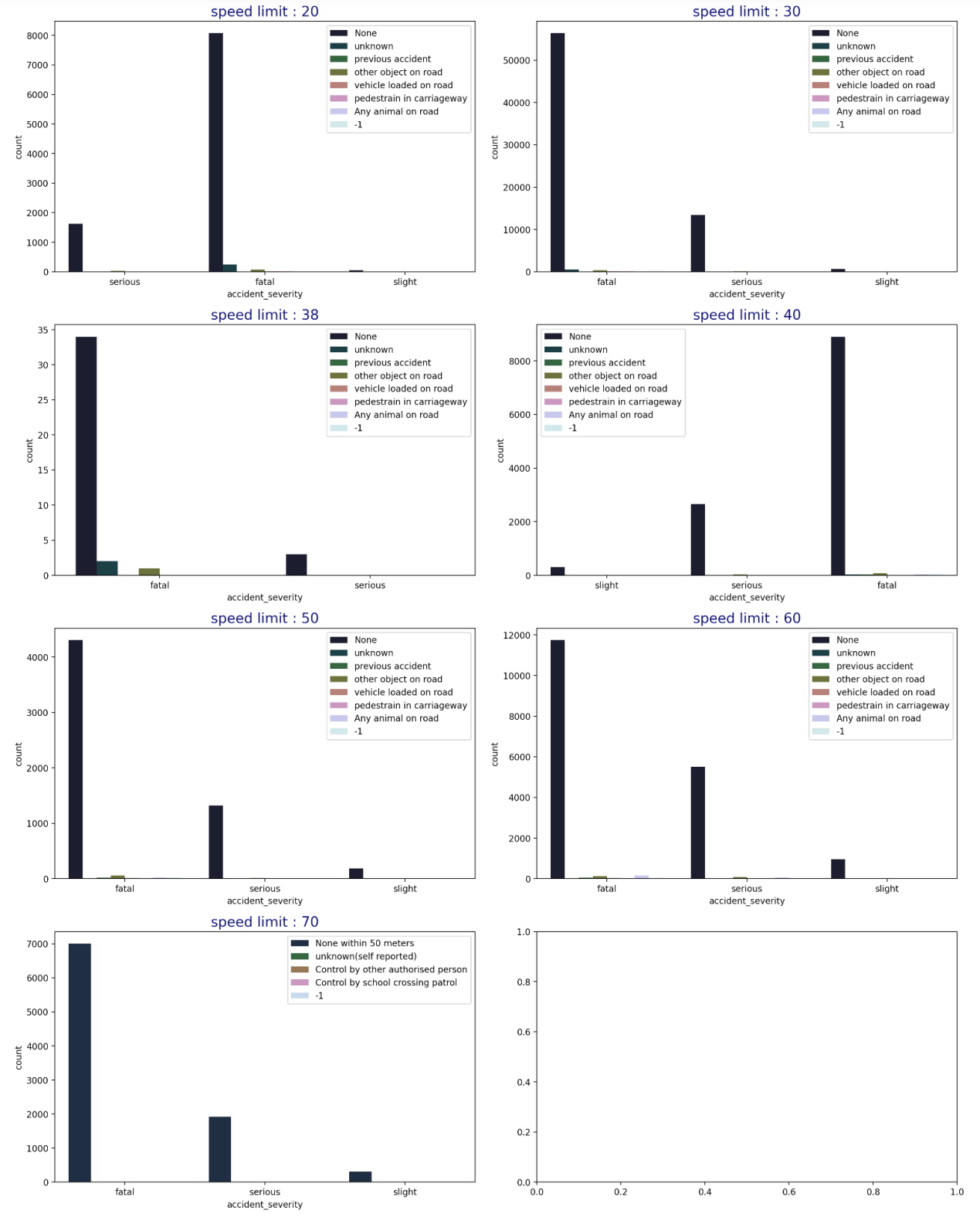


排除掉白天Daylight的限制条件，在夜晚Darkness的限制条件中，可以发现只有20-60的速度限制中，会有车祸的发生。其中，‘Lights lit’相较于‘no lighting’和‘unlit’的限制，会有更大的车祸事故发生率，可能在夜晚中有足够的灯光，会给司机带来更大的安全假象，从而放松注意力，导致车祸的发生。但在速度限制为60的图标中，’no lighting’的条件下，车祸的事故率相较于’lights lit’和’lights unlit’的情况下会更高。

* 1. Establish the relationship model between speed limit and accident severity including impact factor carriageway hazards

将总数据中的特征’carriageway\_hazards’中的类别：0、1、2、3、4、5、6、7和9进行命名更改，分别更改为：‘None’、‘vehicle loaded on road’、‘previous accident’、‘dog on road’、‘other animal on road’、‘pedestarain in carriageway’、‘Any animal on road’和‘unknown(self reported)’。

将特征‘speed\_limit’作为数据分类的指标，此特征类的数据种类有：20、30、38、40、50、60和70，按照限速设置对总数据进行划分七部分子数据。同时，各类限速指标和车祸伤亡建模，并可视化，同时将‘carriageway\_hazards’特征中各类别加入进去：



由此上图标分析：排除到路面无任何障碍物的情况下，在车速限制为20、30、38、50和60的情况下，前方发生事故导致后面车辆发生事故的数量较大。所以，当前方有车辆发生事故的同时，需要及时的将车辆挪至路边，不影响后方车辆的正常行驶和判断。

1. **事故相关性总结**

不论是从‘accident\_severity’还是‘accident\_casualty’特征中的数据分类统计中，都可以发现轻微车祸和车祸经历者受轻伤占大比重，但这并不代表着我们的交通管制和交通执法存在着一些问题。因为我们仍然发现会21%的严重交通事故的发生和1.94%的致命性重大交通事故的发生，这也分别对应着27058例车祸受伤者和2493例的车祸导致死亡者，所以我们仍然要关注这些小比重的伤亡。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。

从对‘weather’的分析过程中，这一特征实质上对车祸的影响权重并非是具有影响力的特征影响因子。在加上了各种‘speed\_limit’的特征限制之后（排除车主因超速违规之类导致的车祸对后期各个特征分析造成影响），对特征‘pedestrain\_crossing\_human\_control’、‘junction\_control’和’carriageway\_hazards’的进行探索性数据分析之后，可以发现对于人行道管控当中，在车辆行驶主干道路上的人行道50m外设置‘前方有行人’的道路标识是后期管控措施进行升级的重要举措，这会大大降低各种级别的车祸。同时，在易于发生堵塞的路段，也应当在远处设置好‘stop’提醒的标识，以让车主慢下车速停止往前行驶或换路行驶，并且在发生堵塞的地方，及时地分派出交通管控人员及时的进行疏散。因为在分析’carriageway\_hazards’之后，前方车祸的发生导致后方车辆继发性地发生二次车祸的概率是最大的，所以交通管控人员的及时派发会对降低车祸的发生率起到重要的影响。

1. **后续完善**
2. 由于表中仍有部分特征值未被用于可视化的图表绘制，以及一些相关因素还没有加入相关性图标的绘制。比如，只绘制了车祸严重程度的饼状图和事故人伤亡程度的沙漏图，一些其他的整体因素没有呈现如交通工具、季节时间还有地区的占比情况；相关性因素中，只考虑了天气的八种状况，还可以考虑其他特征值的不同状况对于车祸严重程度的影响；只以限速为基准，分析其他因素对于车祸严重程度的影响，还可以以发生时间、地点等等为基准，分析其他相关因素对于人员伤亡程度的影响等等。
3. 还可以引入相关性研究的一些模型，与绘制的图表结合起来，利用机器学习的一些知识，进一步对于相关性进行探究，同时增加模型的预测及可靠性的验证，以便于以后可以用于其他数据的相关性研究。
4. 增加一些图表的类型，让可视化成果更加个性化、多样化，有利于实验成果的丰富，让数据可视化的特点更加突出。
5. **总结**

通过此次导论课的期末实验作业，我进一步掌握了数据处理与可视化呈现的知识，并且自学了一些图表的绘制。不论是实验本身还是我的专业技能都有一定的提升。

通过这次实验数据的分析，我们仍要重视交通安全与生命安全，虽然轻型交通事故与轻度受伤占大多数，但是致命性的交通事故数量仍然众多，会对社会生命及财产安全造成威胁。同时极端恶劣天气下，我们尤其要注意交通安全，注意好道路限速条件，对于没有交警，没有路牌的交叉路口和人行横道注意观察，对于车道的突发状况注意留意；而交通资源的平均分配在事故发生前的风险规避和事故后的应急处理可以很大程度上削减交通事故的发生率和严重程度，对于我国的交通管理也有一定的借鉴意义。