**Introdução**

Serve o presente relatório para explicitar a análise, preparação dos dados, extração de conhecimento, obtenção de resultados e análise crítica dos mesmos para o dataset da competição de Sistemas Baseados em Similaridade e o o dataset escolhido pelo grupo.

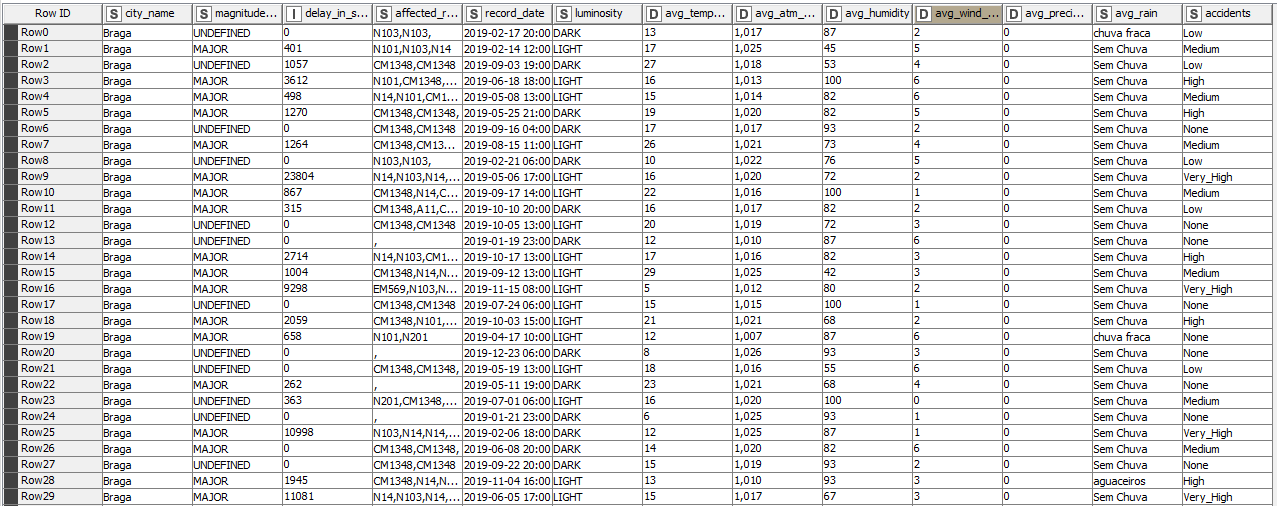
Os workflows criados visam a concessão e otimização dos modelos baseados em árvores usando um sistema de aprendizagem com supervisão de forma obter uma classificação final que no caso do dataset da competição diz respeito a uma previsão do número de incidentes rodoviários que irão acontecer numa cidade a uma dada hora e no caso do dataset escolhido pelo grupo diz respeito à previsão do vencedor de um jogo de League of Legends com dados apenas dos primeiros 10 minutos da partida.

De forma a obter os melhores resultados nos modelos criados foi necessário por em prática as técnicas aprendidas na aula de preparação, extração de dados, seleção de features e hiper-parâmetros e validação de resultados.

**Dataset competição**

Para a competição foram-nos disponibilizados dois datasets. Um training\_data.csv e um test\_data.csv.

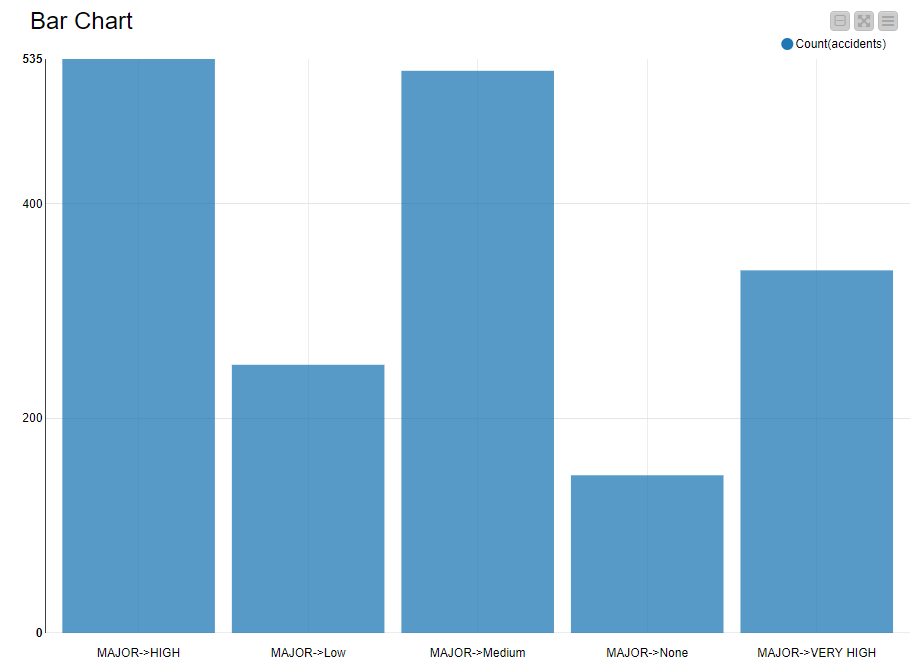
O taining set foi usado para treinar o modelo de Machine Learning criado. Neste encontra-se informação referente ao nível de incidentes rodoviários de cada registo para que treinar os modelos de aprendizagem.

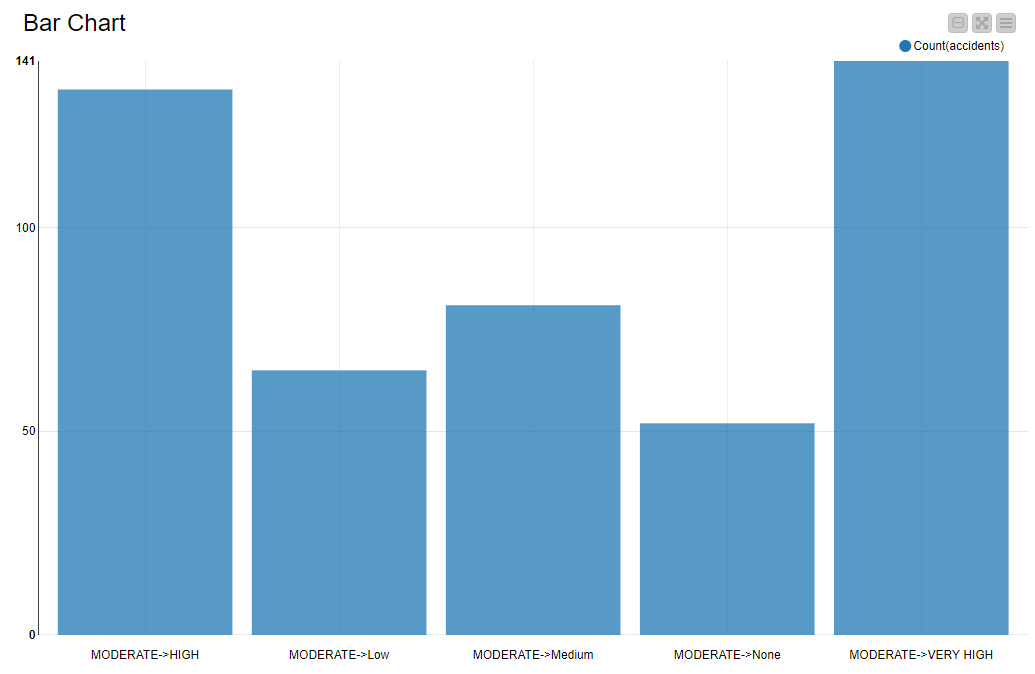


À primeira vista temos algumas features que intuitivamente parecem ter mais impacto na classificação final da intensidade dos acidentes do que outras. Começamos então isolar as essas feutures e ver graficamente o seu impacto.

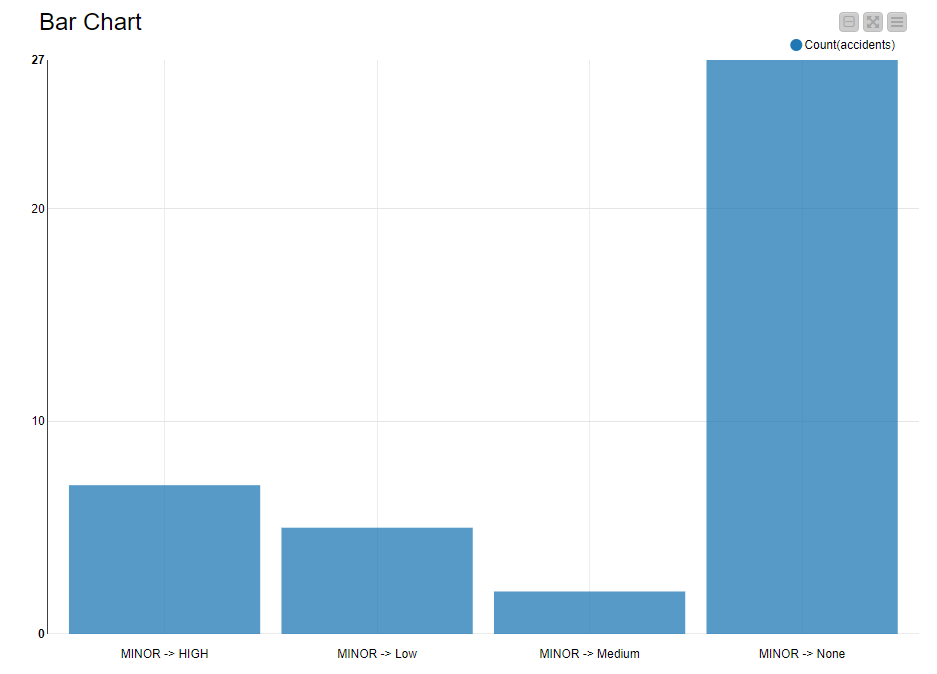
**Magnitude\_of\_delay**

Começando então pela magnitude do delay, podemos ver através dos seguintes gráficos de barras a distribuição do nível de acidentes rodoviários e a magnitude de atraso provocados pelos incidentes.

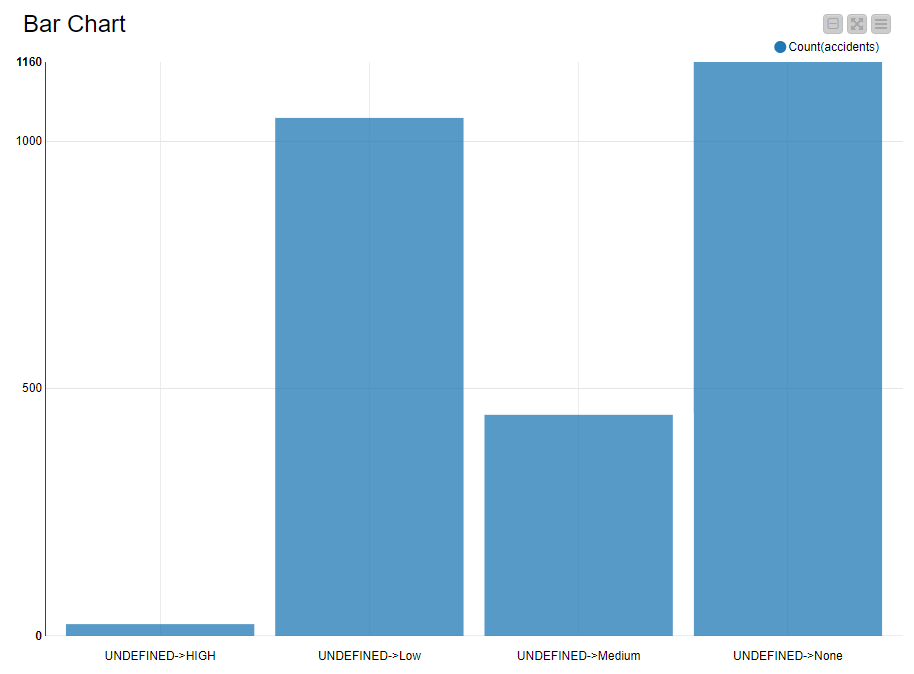


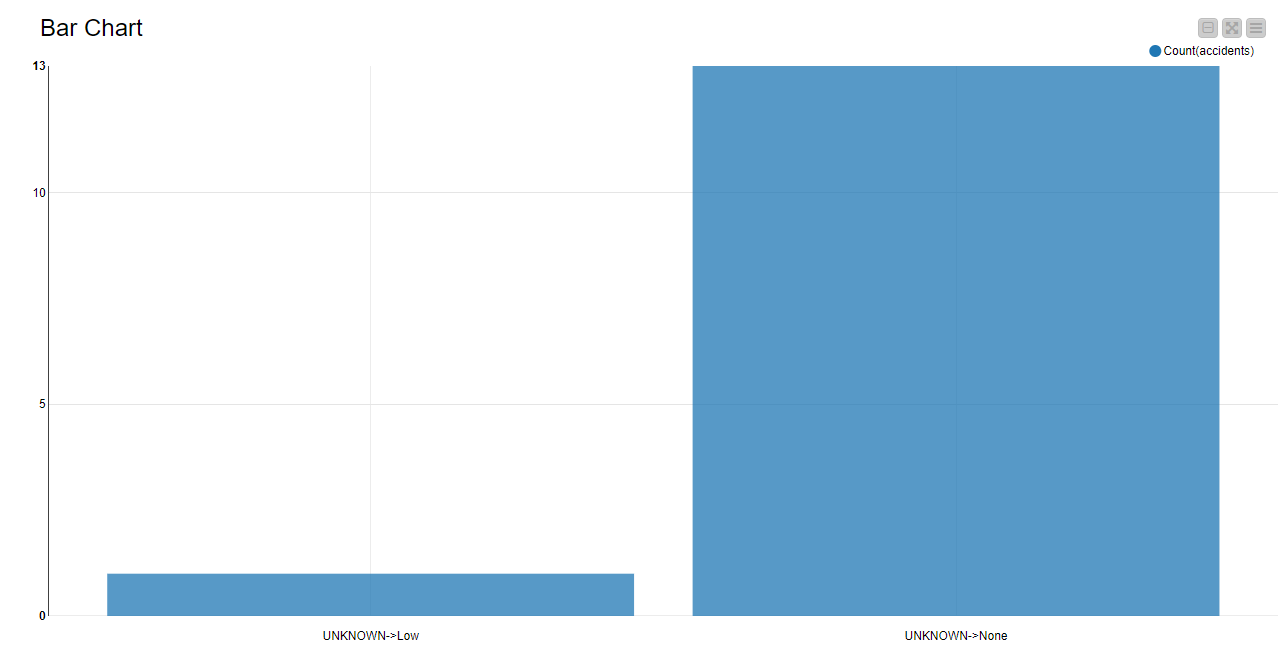


Como podemos ver quando existe uma magnitude alta ou moderada o delay o nível de acidentes tende a variar entre o “Very\_High”, “High” e “Medium”. Sendo que também existe uma percentagem considerável de “Low”.



Quando a magnitude é menor temos um maior nível de incidentes “None”.





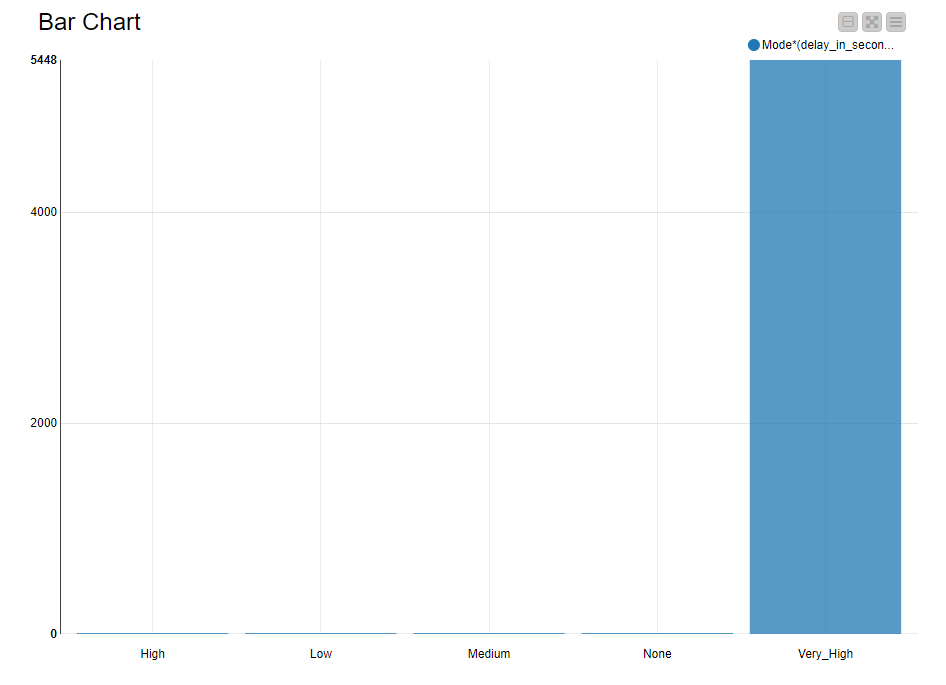
Quando a magnitude se apresenta indefinida ou desconhecida temos muito provavelmente um nível de incidentes baixo ou até inexistente.

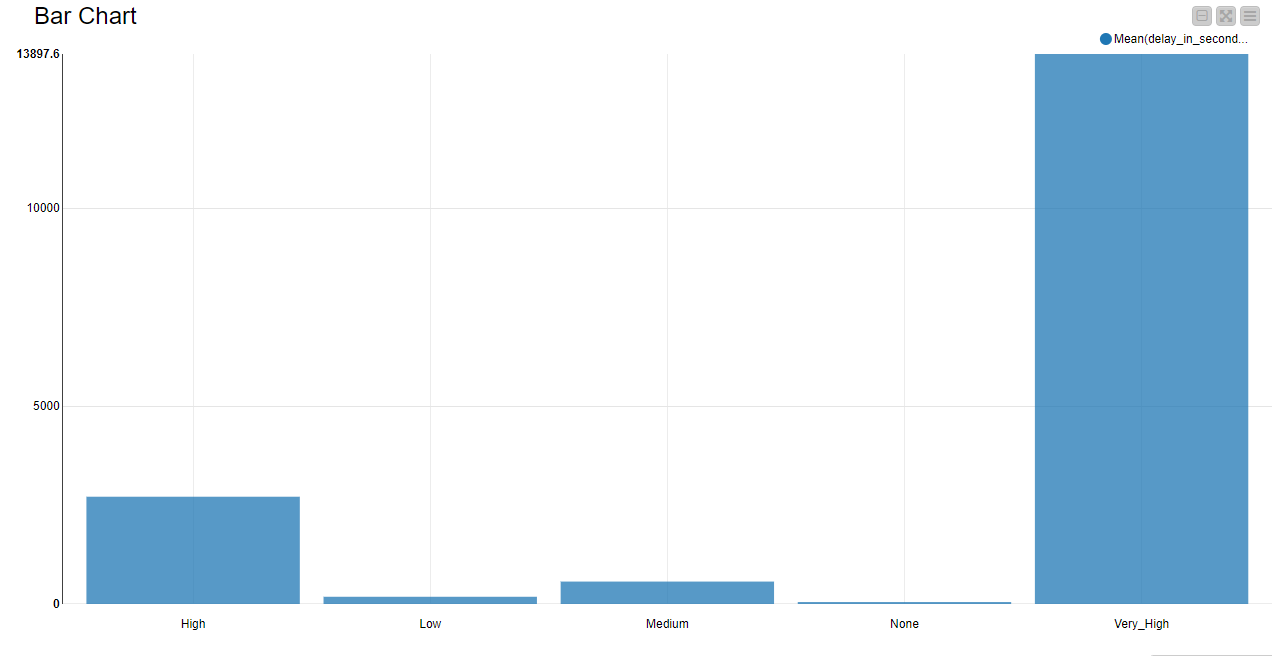
Como podemos ver a magnitude de atraso permite algumas previsões de forma intuitiva para prever o nível de incidentes. O que torna esta feature, à partida, uma forte candidata para a escolha do modelo final.

**Delay\_in\_secs**

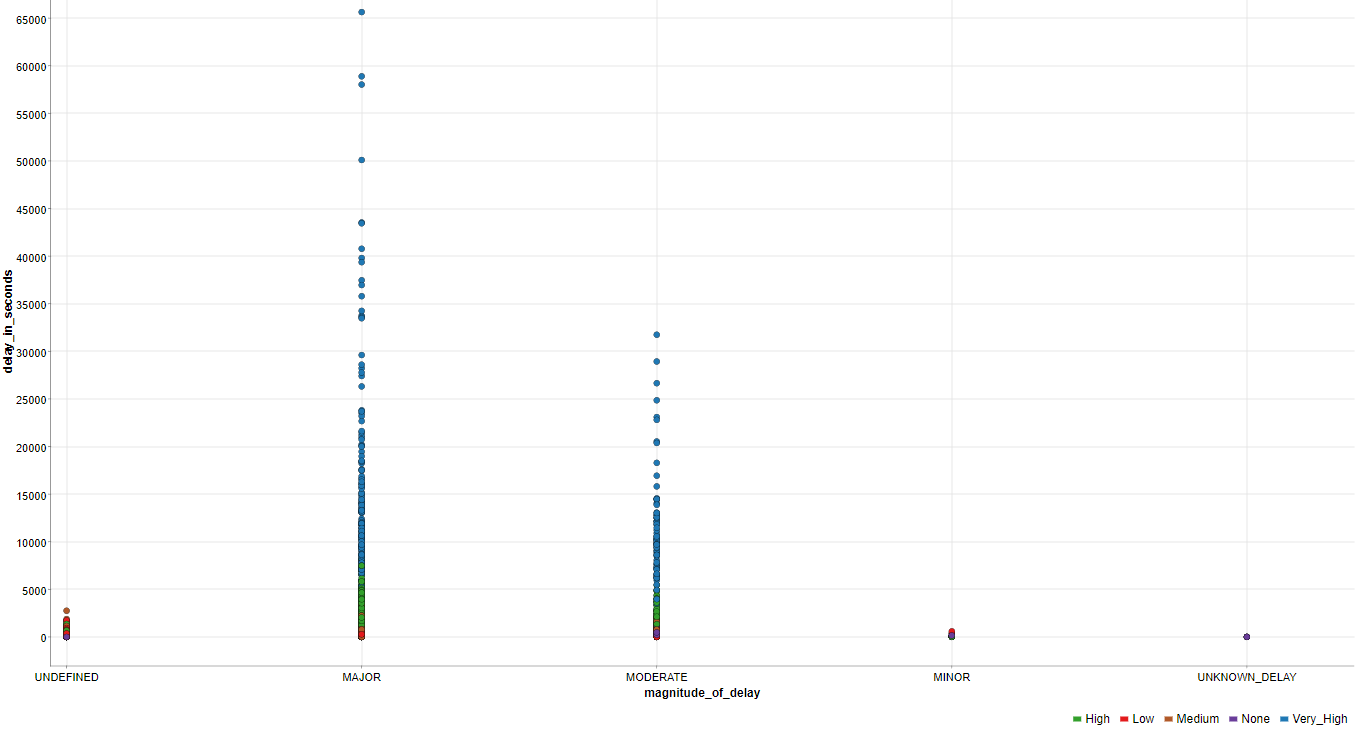
O dataset apresenta ainda o delay em segundos, provocado pelos incidentes que se verificam.

Começamos então por ver a média e moda de delay para cada nível de incidentes.





Destes valores podemos afirmar que o delay\_in\_secs pode ser usada como um forte indicativo para níveis de incidentes “Very\_High”.



Relacionando agora a magnitude de delay com o delay em segundos e o nível de incidentes podemos ver que há uma relação direta entre o tempo de delay e o nível de incidentes. Sendo que quanto maior for o delay maior a probabilidade de um nível de incidentes alto. Também podemos retirar que quanto maior for a magnitude do delay maior será a probabilidade de temos um grande delay em segundos.

**affected\_roads**

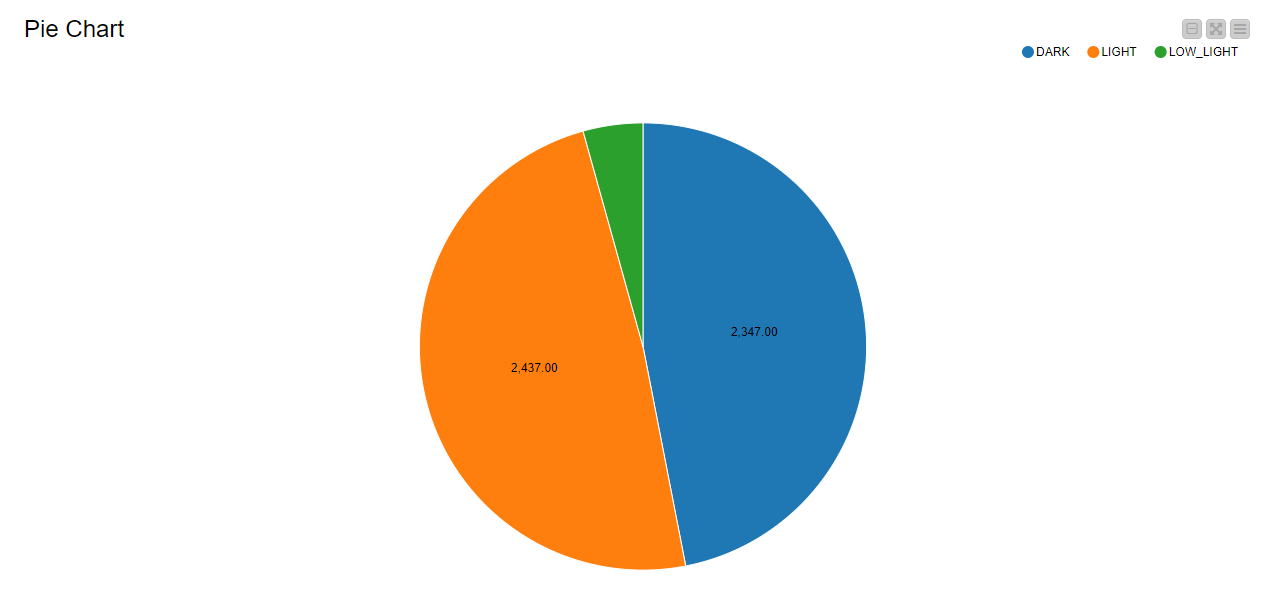
Uma das features disponíveis no dataset mais interessantes é a affected\_roads. No entanto da forma como os dados são fornecidos não é possível extrair informação relevante. Só após termos feito tratamento de dados, como mais à frente no relatório vai ser descrito, é que foi possível extrair informação relevante para o modelo.

**Record\_date**

Tal como a affexted\_roads esta feature embora bastante relevante numa análise inicial não tem grande utilidade. Só após termos feito o tratamento de dados é que foi possível extrair informações relevantes para o modelo.

**Luminosity**

Uma das preocupações que tivemos ao analisar esta feature foi a distribuição de dados. Mas como podemos ver através do seguinte pie chart, temos uma percentagem de quase 50% de incidentes entre a luminosidade “Dark” e “Light”. Sendo que os únicos dados não muito relevantes são os da “Low\_Light” devido aos poucos registos.



Após a análise do nível de incidentes nas diferentes condições de luminosidade é possível concluir que os registos com uma luminosidade “Light” se encontram bem distribuídos entre os diferentes níveis de incidentes, já para uma luminosidade “Dark” uma maior número de registos com níveis de intensidade “Low” ou “None”.

