This dataset consists of 48283 different accidents in different parts of UK and explains about their casualties. All accidents were unvalidated and occurred in 2022.

Some Terms About Casualties

- Accident Reference: a unique value for each accident
- Accident Index: year of accident + Accident Reference
- Vehicle Reference: unique value for each vehicle in a singular accident
- Casualty Reference: unique value for each vehicle in a singular accidents
- Casualty Class: whether the casualty was a driver/rider, passenger or pedestrian
- Casualty Severity: degree of severity (slight, severe, fatal)
- Casualty Home Area Type: where did the casualty live? (urban area, small town or rural area)
- Casualty IMD Decile: In which IMD decile was the casualty located?

Analysis Techniques

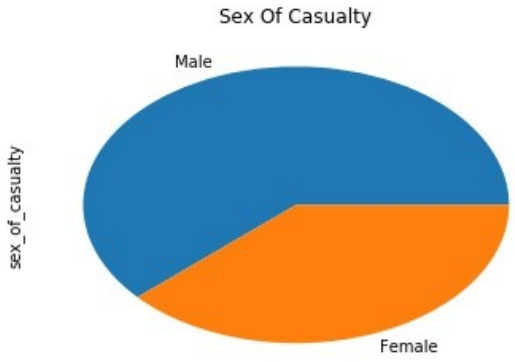
- Basic EDA
  - Descriptive Statistics
  - Checking unique values of columns
- Data Preparation
  - Nonanonymizing Data
    - Finding Guide Data From Source Website
      - Mapping to meaningful data
  - Handling Missing Values
    - Figuring Out Type of Missing Values
      - np.nan
      - None
      - ‘unkown’
    - MCAR, CAR and NCAR Analysis
      - checking if missing values are randomly occuring in the data or not
    - Checking number of missing values in each record
      - a simple programmer-written function
    - Dropping Rows with multiple missing values
    - Imputing Some Rows With Mode, Mean or Median
  - Feature Selection
    - Identification of invaluable columns which provide no information and dropping them
      - status, accident_index, accident_year
    - Identification of controversial columns and records
      - accident_reference: 34NNC2522
      - pedestrian location, movement and road maintenance worker
      - car, bus/coach passenger
  - Univariate Analysis
    - Plotting Each Column To Get Insight About Its Distribution

- Multivariate Analysis
  - Plotting Some Columns Together To See If They’re Related Or Not
    - kde plot
    - violin plot
    - heatmap
    - countplot
    - boxenplot
  - Statistical Analysis
    - chi2 independence and goodness of fit test

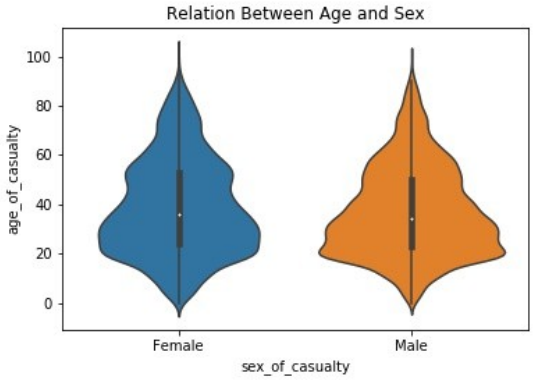
- Preprocessing Data
  - One Hot Encoding
    - converting categorical data to a numerical matrix of zeros and ones
  - Train Test Split
    - creating data for training the model and testing how well it’s learnt
  - Random Over Sampling
    - balancing data because number of records for different categories were not equal
- Training Models For Severity Prediction
  - Decision Tree Classifier
  - Random Forest Classifier
  - SVC
- Model Evaluation
  - Classification Report
    - Accuracy
    - Precision
    - Recall
    - F1 Score

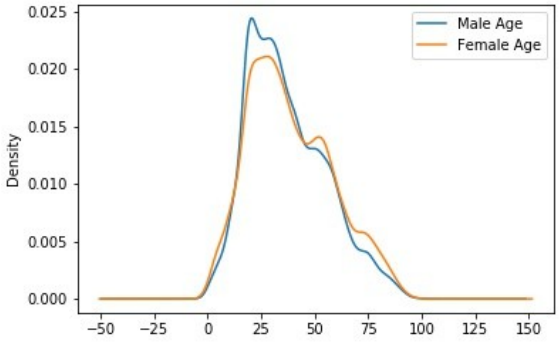
Results

- Males tend to have more accidents that females

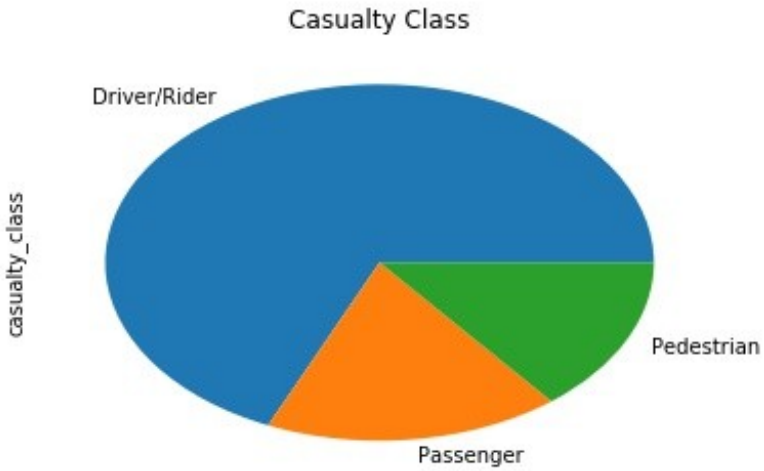


- There’s no difference in age distribution between different genders.

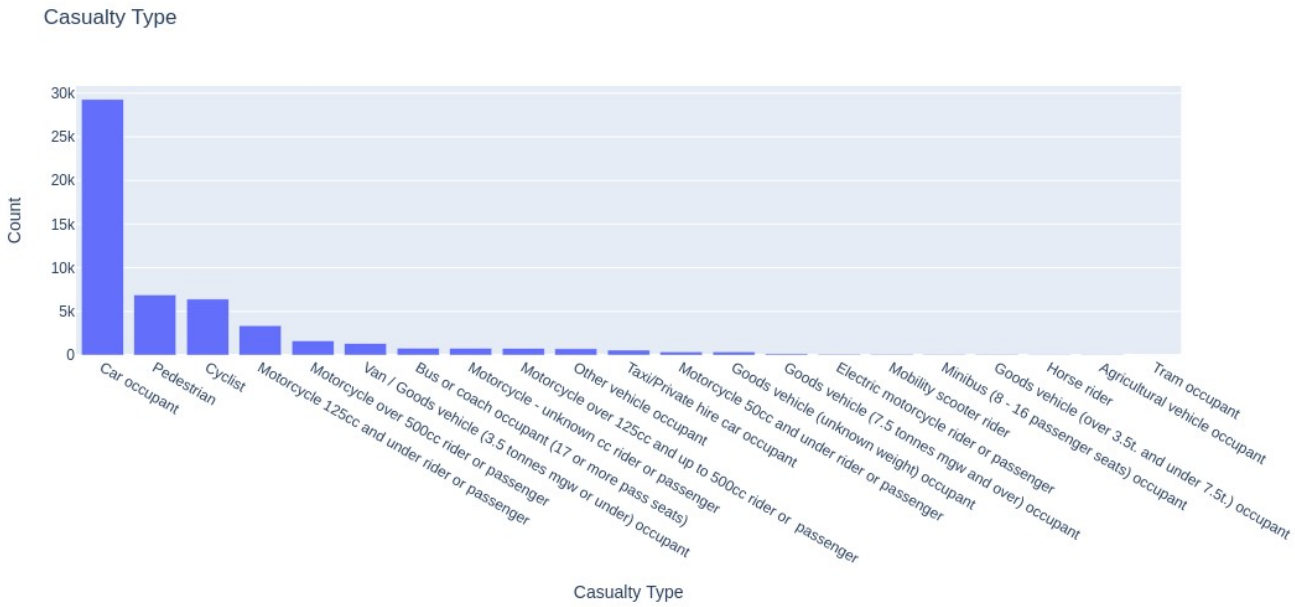




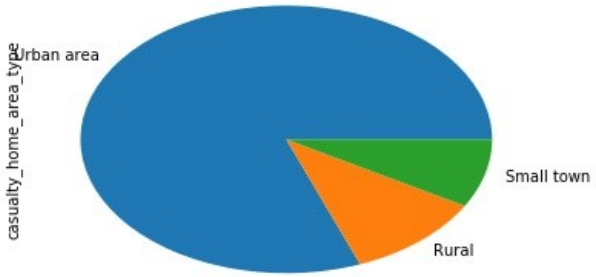
- Drivers have more accidents than passengers and pedestrians



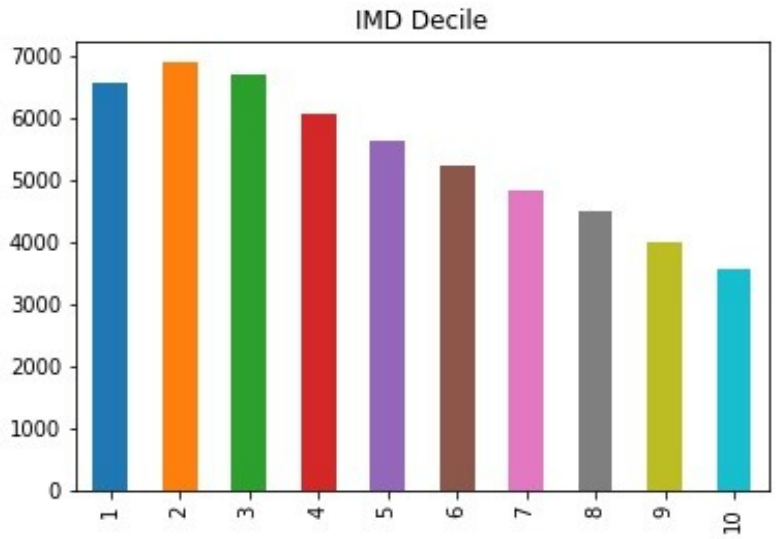
- Most of accidents have occurred on cars



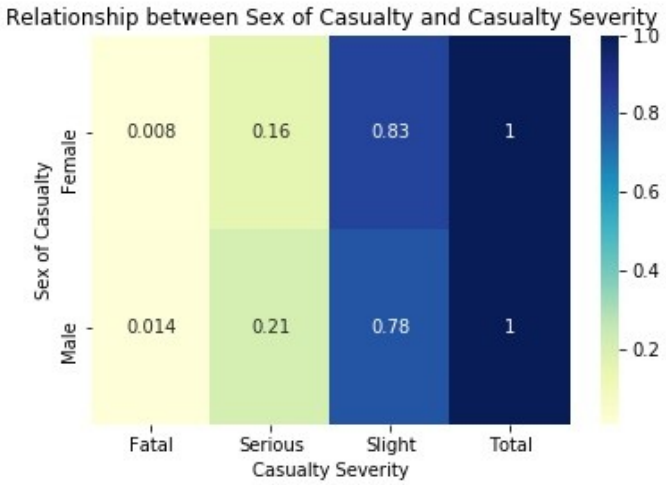
- Most of accidents occur in urban areas



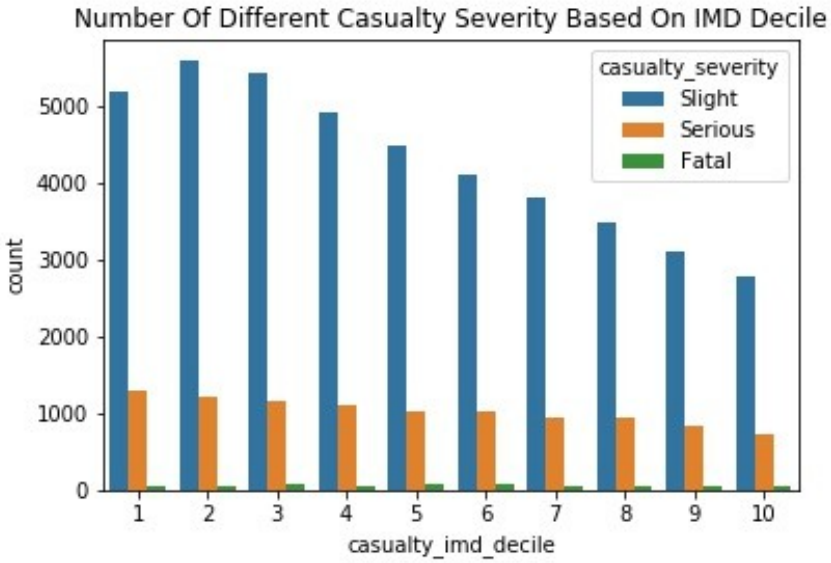
- Higher imd deciles have less accidents



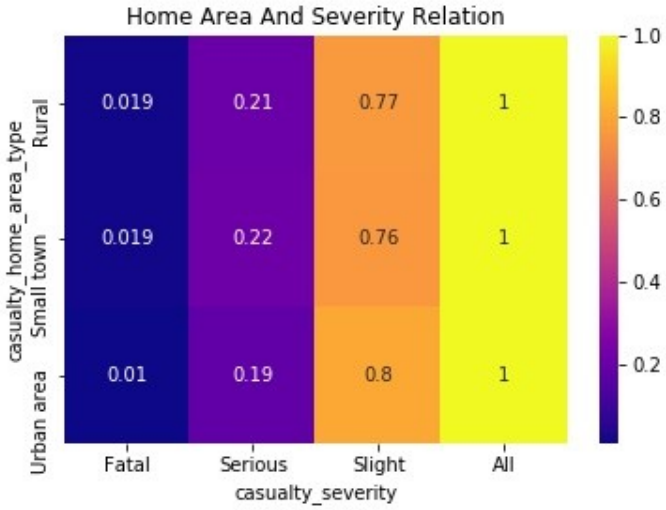
- Men tend to have more severe casualties



- People from lower imd deciles tend to have more severe casualties



- People from less developed areas have more severe accidents



Discussion and Recommendations

- Since less developed areas have more rates of fatal and serious casualties it’s recommended to put signs in roads and train people to have safer driving experience.
- Since teens have more rates of accidents and casualties It’s better to have them practice more and be more careful.
- Constructing Healthcare centers is a good idea to prevent more severe casualties
- Further investment in lower imd decile regions could prevent fatal casualties.
- Enhancing Education for teens and lower imd decile people should be one the most important things to consider