Car Crashes in South Australia

Kerimzhanov Aziz

## Problem statements:

The Department of Infrastructure and Transportation is committed to meeting the evolving planning, transportation, and infrastructure needs of South Australia. Their official website publishes a wide range of information, including data related to vehicular accidents. For this study, I download the road crashed data for the year 2020, as the 2021 data remains incomplete. The objective of this report is to provide a comprehensive EDA analysis of the data, exploring various relationships and dependencies. Specifically, the study will investigate potential associations between the age of the vehicle, the number of individuals involved in accidents, and the incidence of accidents.

## Dataset:

Link: [Road Crash Data - Road Crash Data 2020 - data.sa.gov.au](https://data.sa.gov.au/data/dataset/road-crash-data/resource/1c97e5d4-3a2f-4db7-b856-4eff7e6acef9)

This dataset consists of 3 datasets:

* 2020\_DATA\_SA\_Casualty – it stores data about casualties. It has 4877 raws and 12 columns and contains the following data: 'REPORT\_ID', 'UND\_UNIT\_NUMBER', 'CASUALTY\_NUMBER', 'Casualty Type', 'Sex', 'AGE', 'Position In Veh', 'Thrown Out', 'Injury Extent', 'Seat Belt', 'Helmet', 'Hospital'], dtype='object'.
* 2020\_DATA\_SA\_Crash – it stores data about the crashes. It has 11534 raws and 33 columns and contains the following data: 'REPORT\_ID', 'Stats Area', 'Suburb', 'Postcode', 'LGA Name', 'Total Units', 'Total Cas', 'Total Fats', 'Total SI', 'Total MI', 'Year', 'Month', 'Day', 'Time', 'Area Speed', 'Position Type', 'Horizontal Align', 'Vertical Align', 'Other Feat', 'Road Surface', 'Moisture Cond', 'Weather Cond', 'DayNight', 'Crash Type', 'Unit Resp', 'Entity Code', 'CSEF Severity', 'Traffic Ctrls', 'DUI Involved', 'Drugs Involved', 'ACCLOC\_X', 'ACCLOC\_Y', 'UNIQUE\_LOC'
* 2020\_DATA\_SA\_Units – it stores data about the units. It has 24539 raws and 18 columns and contains the following data: 'REPORT\_ID', 'Unit No', 'No Of Cas', 'Veh Reg State', 'Unit Type', 'Veh Year', 'Direction Of Travel', 'Sex', 'Age', 'Lic State', 'Licence Class', 'Licence Type', 'Towing', 'Unit Movement', 'Number Occupants', 'Postcode', 'Rollover', 'Fire'.

Explanation of the fields.

**Report\_ID** - Unique number assigned to records to separate individual crashes

**Stats Area** - A code defining whether the road crash occurred within City, Metro or Country Area

**Suburb** - The Suburb that the crash occurred in

**Postcode** - The Postcode that the crash occurred in

**Local Name** - The Local Government Area that the crash occurred in

**Total Units** - The total number of units involved in a road crash

**Total Casualties** - Total number of casualties (fatalities + treated injuries) as a result of a road crash

**Total Fatalities** - Total number of fatalities as a result of a road crash

**Total Serious Injuries** - Total number of people admitted to hospital with overnight stay as a result of a road crash

**Total Minor Injuries** - Total number of people treated by private doctor or treated at hospital but not admitted

**Year** - Year of Crash

**Month** - Month crash occurred

**Day** - The day of the week the crash occurred

**Time** - The reported time of the crash

**Area Speed** - Limit The speed limit at the time and location of the crash

**Position Type** - Identifying if a crash location was at an intersection or midblock

**Horizontal Alignment** - Defines the horizontal alignment of the road at the sight of the crash

**Vertical Alignment** - Defines the vertical alignment of the road at the sight of the crash

**Other Features** - Defines other relevant features of the crash site locations

**Surface Type** - Defines the road surface type at the crash location

**Road Moisture** - Defines the pavement surface moisture condition at the crash location

**Weather** - Defines the weather condition at the time and location of the crash

**Light Conditions** - The lighting condition at the time and location of the the crash

**Crash Type** - Defines the road crash type

**Unit Responsible** - The number of the unit determined to be responsible for the road crash

**Entity Code** - A code defining the entity deemed to be responsible for the road crash

**Crash Severity** - Defines the road crash severity (classified by the highest injury severity sustained in the crash). Decoded values:

“3: SI” = Serious Injury, “2: MI” = Minor Injury, “1: PDO” = Property Damage Only

**Traffic Control** - Defines the traffic control at the time and location of the road crash

**DUI Involved** - Involved if at least one controller in the crash recorded an illegal Blood Alcohol Concentration level

**Drugs Involved** - Involved if at least one controller in the crash tested positive for a prescribed drug (THC (cannabis), methylamphetamine (speed, ice or crystal meth) or MDMA (ecstasy))

**X** - The X coordinate of the crash when located

**Y** - The Y coordinate of the crash when located

**Unique Loc** - A concatenation of the X and Y coordinates for the purpose of grouping crashes

**Unit Number** - A number assigned to a unit involved in a road crash

**Number of Casualties** - The total number of casualties for a unit involved in a road crash

**Veh Reg** - State The state or Territory the unit is registered in

**Unit type** - A code defining the type of unit involved in a road crash

**Vehicle** - Year The year of manufacture of a vehicle

**Direction of** **Travel** - A code defining the direction of travel of a unit immediately prior to the road crash

**Unit sex** - A code defining the sex of the controller of the unit involved in a road crash

**Unit Age** -The age of the controller of the unit involved in a road crash

**Lic State** - A code defining the state of issue of the licence of the driver/rider

**Licence Class** - A code defining the class of the licence held by the driver/rider

**Licence Type** - A code defining the level of licence held by the driver/rider

**Towing** - A code defining if the unit was towing another unit

**Unit Movement** - A code defining the movement of a unit immediately prior to the road crash

**Number** **of occupants** - A code defining the number of occupants in the unit at the time of the crash

**Postcode** - The residential postcode of the unit in a road crash

**Rollover** - Subsequent rollover of vehicle

**Fire** - Vehicle caught fire at some stage in the crash

**Casualty number** - A number assigned to a casualty within a unit involved in a road crash

**Casualty Type** - A code defining the entity injured in a unit in a road crash

**Casualty Sex** - A code defining the sex of a casualty injured in a road crash

**Casualty Age** - The age of a casualty injured in a road crash

**Position in vehicle** - A description of the position of a casualty in or on a unit involved in a road crash

**Thrown Out** - An indication of whether the casualty is thrown out of a unit as a result of a road crash

**Injury Extent** - A description of the severity of the injury sustained as a result of a road crash

**Seat belt** - An indicator of the casualty's seatbelt usage in a road crash

**Helmet** - An indicator of the casualty's helmet usage in a road crash

**Hospital** - The name of the hospital where a casualty is treated or

admitted

## Exploratory Data Analysis (EDA):

Exploratory data analysis helps us understand how the characteristics of the data set vary for vehicle insurance fraud cases. We first begin by understanding individual characteristics (univariate analysis) and then conduct bivariate and multivariate analyses.

**Univariate Analysis:**

- 1 casualty encounters the most

- total units involved the most is 2

- In most accidents, the number of fatalities is 0

- In most crashes, total number of people who got serious injuries is equal 0

- In most crashes, total number of people who got minor injuries is equal 1

- Most accidents occurred at the beginning and end of the year

- Most of the accidents occurred in the middle of the week.

- Most of the accidents occurred on the highway with a maximum speed of 60 km/h.

- Most accidents occurred on the sealed road

- Most accidents occurred in a non-alcoholic state

- Most accidents involved 1 unit

- Most accidents involved 1 injured people

- Most of the accidents involved cars between 2007 and 2018

- Most of the accidents involved men

- Most of the accidents involved people in their 20s

- There were no passengers in most accidents

- In most accidents, the cars were not overturned  
 - Most of the accidents occurred in the metropolitan

- Most accidents occurred at intersections, undivided roads, and T-junctions

**Bivariate Analysis:**

- Most of the fatalities occurred on the 100-110 km/h highway

- Most of the injured people were on the sealed road surface

- The number of fatalities was higher at night than during the day

- The number of injured men is higher than the number of injured women.

- In most cases, the number of injured people both in the rain and in the day is 1

- Least of all the units where the car was rolled over

- In most cases, there were right angle and rear end crashes

- There were 0 serious injuries in most rear-end accidents

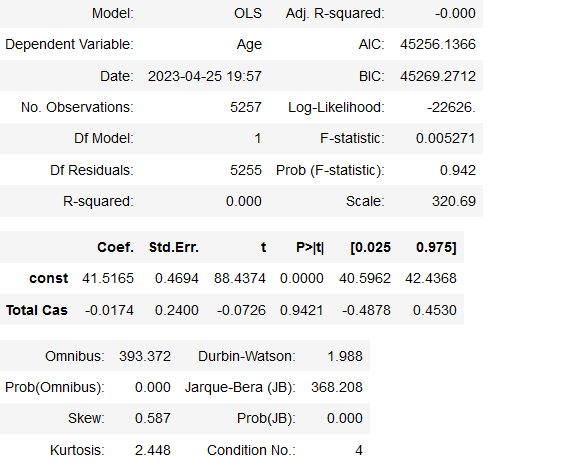
- In most cases, people were minor injured during the day

- In most cases, there were rear end crashes during the day

## Linear regression

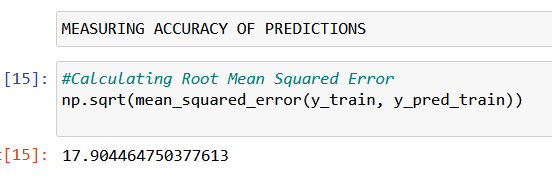
**Simple linear regression**

In the first part, I conducted a simple linear regression analysis using the 'Age' and 'Total Cas' columns. In summary, the following table was obtained:

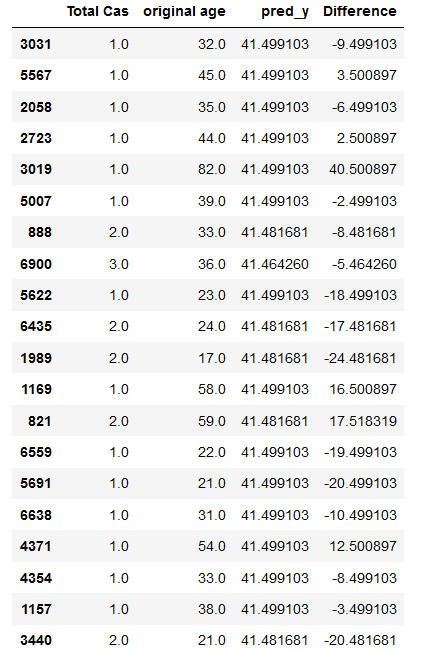


Here, the R-squared value is equal to 0, indicating that the model explains 0% of the variation in 'Age'. The p-value is 0.9421, suggesting that there is no significant relationship between the 'Age' and 'Total Cas' columns. To measure the accuracy of the predictions, Root Mean Squared Error (RMSE) was used, which is equal to 1.003.

To measure the accuracy of the predictions, Root Mean Squared Error (RMSE) was used, which is equal to 1.003.



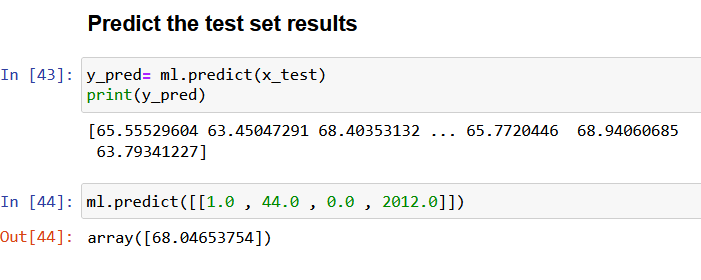
In conclusion, based on the results of this analysis, there is no significant linear relationship between 'Age' and 'Total Cas', and the model does not explain any of the variation in 'Age'. The RMSE value suggests that the predictions based on this model may not be very accurate

The difference:  


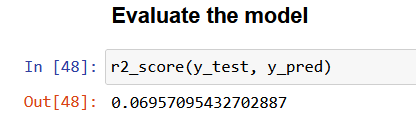
**Multiple Linear regression**

In the multiple linear regression, it was used a dataframe that contained following columns: After splitting the dataset and training the model on the training set, there is such prediction:

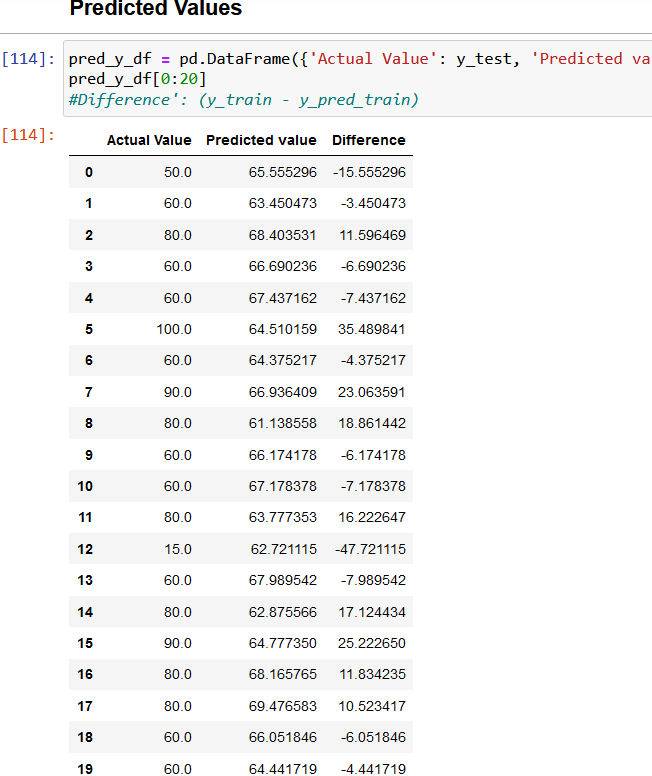
'Total Units', 'Age', 'Total Fats', 'Veh Year', 'Area Speed'



To evaluate the accuracy, it was used r2\_score that is equal to 0.069.



The difference:



The model needs to be improved and is not so close to actual.

## Conclusion

In summary, the EDA analysis conducted on the dataset revealed some interesting insights. The dataset contained a mix of both categorical and numerical variables, with most of the variables showing a moderate to weak correlation

The simple linear regression analysis showed that there was no significant linear relationship between 'Age' and 'Total Cas' columns. The R-squared value of 0 and the p-value of 0.9421 suggested that the model did not explain any of the variation in 'Age', and there was no significant relationship between the two columns

The multiple linear regression analysis showed that the 'Total Units', 'Age', 'Total Fats', 'Veh Year' columns slightly predicted the 'Area Speed' column. However, the model did not fit the data very well, as indicated by the RMSE value of 6.29. This suggests that the model needs to be improved for more accurate predictions.