Predict Death by Auto Accident

By Partha Ray

1. Introduction

- 1.1 Chosen data set
- This capstone presentation identifies features within a large dataset which significantly contribute to a road accidents in France from 2005 to 2016
- 3 million samples with 52 parameters per sample
- The said Kaggle data set can be found here: https://www.kaggle.com/ahmedlahlou/accidents-in-france-from-2005-to-2016/home
- And much appreciation to its previous contributors of analysis

1.2 Goal(Frame the Problem):

Predict Death or No Death with a measure of confidence given an arbitrary sample and identify the most relevant features

1.3 Outcome of Interest

- Dead or Gracefully Alive
  - Key Metric: Accuracy Score
- Most Relevant Features
  - Key Metric: Random Forest Feature Importances Attribute

2. Collect raw data needed for problem

3. Process the data for analysis

The Dataset consists of :

- 52 columns
- 3,553,976 rows

Dimensions in Characteristics category

- Accident ID
- Day of Accident
- Month of Accident
- Year of Accident
- Time of Accident
- Lighting Conditions
- Department
- Municipality
- Localisation(congestion level)

Dimensions in Characteristics Category (con’t.)

- Type of Intersection
- Atmospheric Conditions
- Type of Collision
- Postal Address
- GPS Coding
- Geographic Coordinates

Dimensions in Places category

- Road Category
- Road Number
- Numeric Index Route
- Alphanumeric Index Road
- Traffic Regime
- Total Traffic Lines
- Reserved Lane Existence
- Road Gradient
- HomePRNumber

Dimensions in Places category (con’t.)

- PR Distance
- Lane Structure
- Central Lane Width
- Outer Lane Width
- Surface Condition
- Infrastructure
- Situation of Accident
- School Point

Dimensions in Users category

- Vehicle Identification
- Place
- User Category
- Sex of User
- User Year of Birth
- Trip Reason
- Safety Equipment
- Location of Pedestrian
- Action of Pedestrian
- Pedestrian Group

Dimensions in Vehicle Category

- Flow Direction
- Vehicle Category

4. Explore the Data

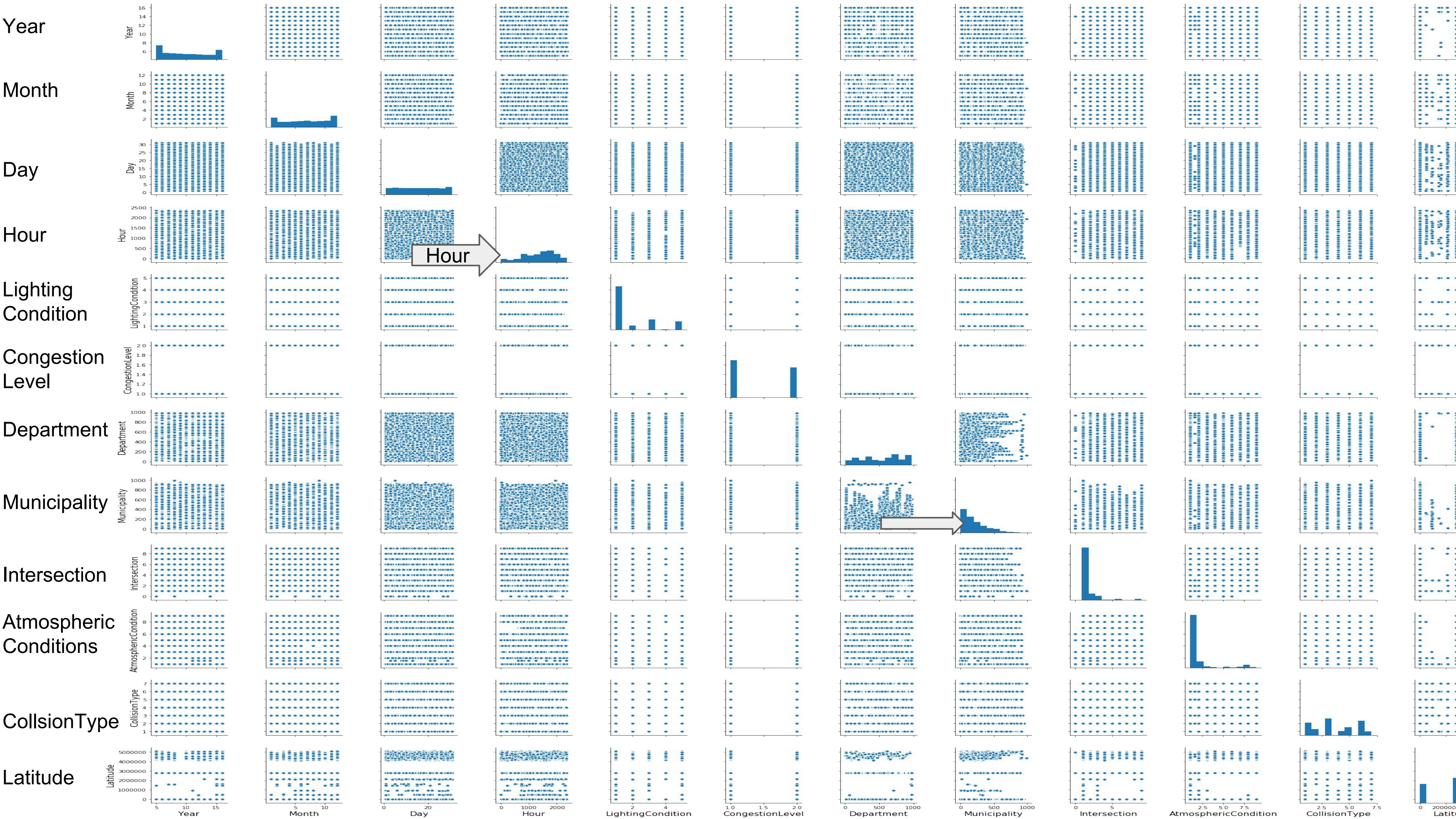
Y outcome parameter class imbalanced?

- 3,471,825 accidents resulted in NO death
- Majority class: 3 million

- 82,151 accidents resulted in a death
- Minority class: mere 82,000

Class imbalanced -> downsample due to bigger initial dataset of 3 million

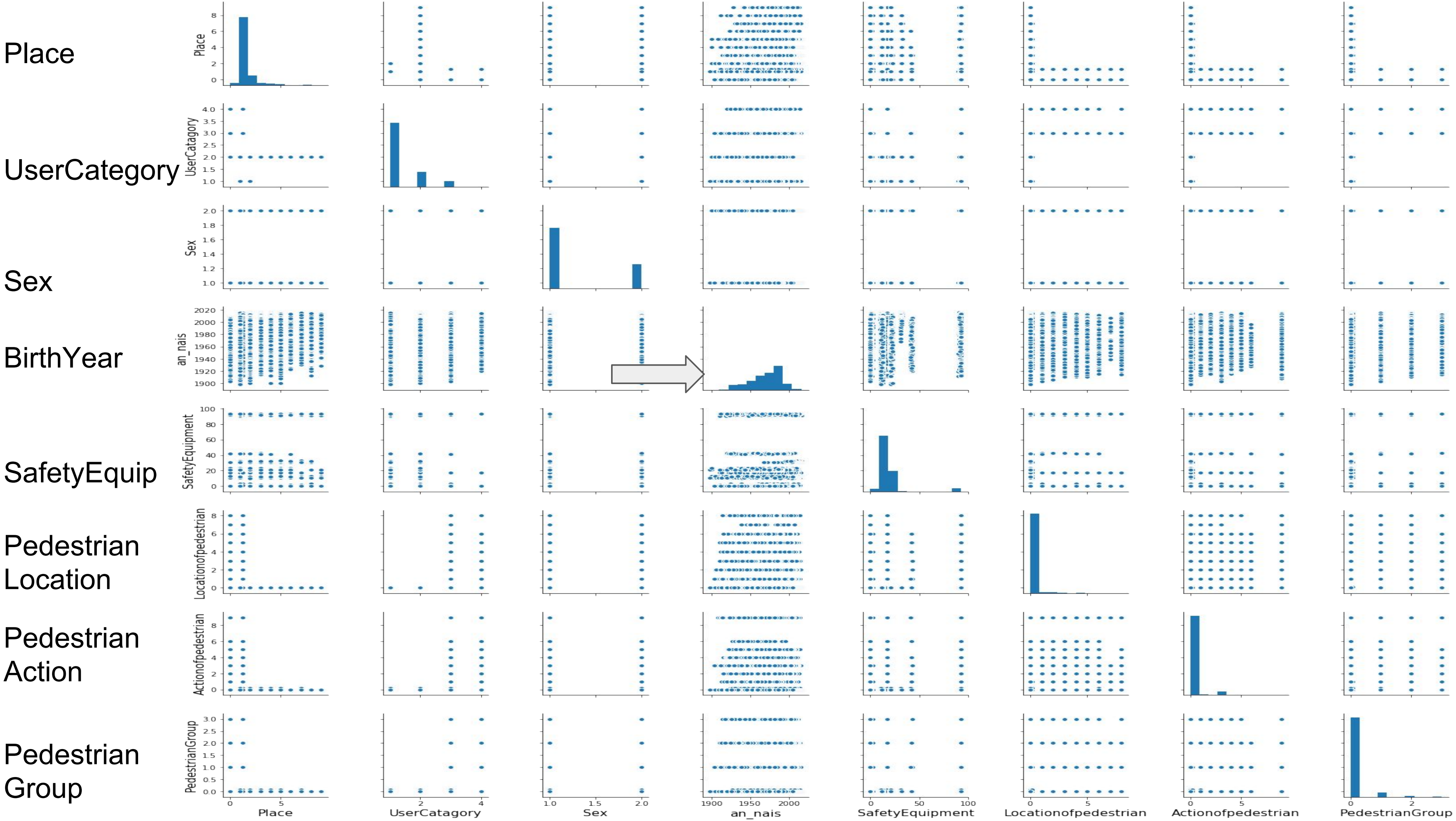
Frequencies
&
Distributions
of
Characteristics’ Features


Year
Month
Day
Hour
Lighting Condition
Congestion Level
Department
Municipality
Intersection
Atmospheric Conditions
CollsionType
Latitude
Hour
Year
Month
Day
Hour
LightingCondition
CongestionLevel
Department
Municipality
Intersection
AtmosphericCondition
CollsionType
Latitude

Frequencies
&
Distributions
of
Places’ Features

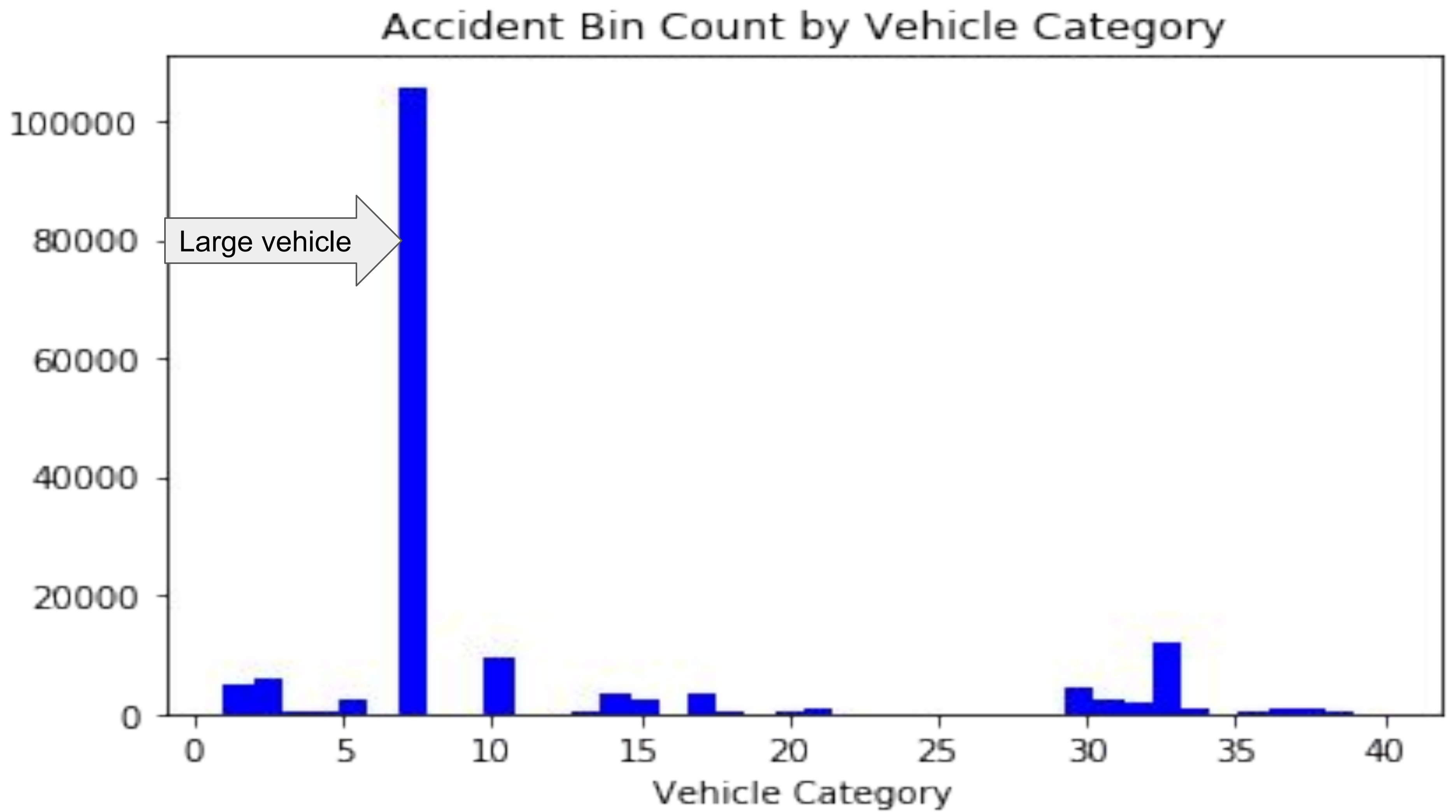

RoadCategory
TrafficType
NumofLanes
OuterLane
gradient
HomePR Number
PRDist
LaneStruc
CentralLane Width
OuterLane Width
SurfCond
Infrastructure
Situation
nearSchool
RoadCategory
TrafficType
NumberofLanes
OuterLane
prof
HomePRNumber
PRDistance
LaneStructure
CentralLaneWidth
OuterLaneWidth
SurfaceCondition
Infrastructure
SituationofAccident
SchoolPoint

Frequencies
&
Distributions
of
Users’ Features



Place
UserCategory
Sex
BirthYear
SafetyEquip
Pedestrian Location
Pedestrian Action
Pedestrian Group
Place
UserCatagory
Sex
an_nais
SafetyEquipment
Locationofpedestrian
Actionofpedestrian
PedestrianGroup

Frequency visualization
of
Vehicles’ Feature



Accident Bin Count by Vehicle Category
100000
80000
60000
40000
20000
0
0
5
10
15
20
25
30
35
40
Vehicle Category
Large vehicle

Delve into Dimension Reduction -> smaller set of Relevant Features

Drop seemingly Irrelevant Features

- AccidentID
- Department
- Municipality
- Road Number
- Numeric Index Route
- Alphanumeric Index Road
- Home PRNumber
- PRDistance
- Vehicle Identification
- Place

Intuitive dropping (con’t.)

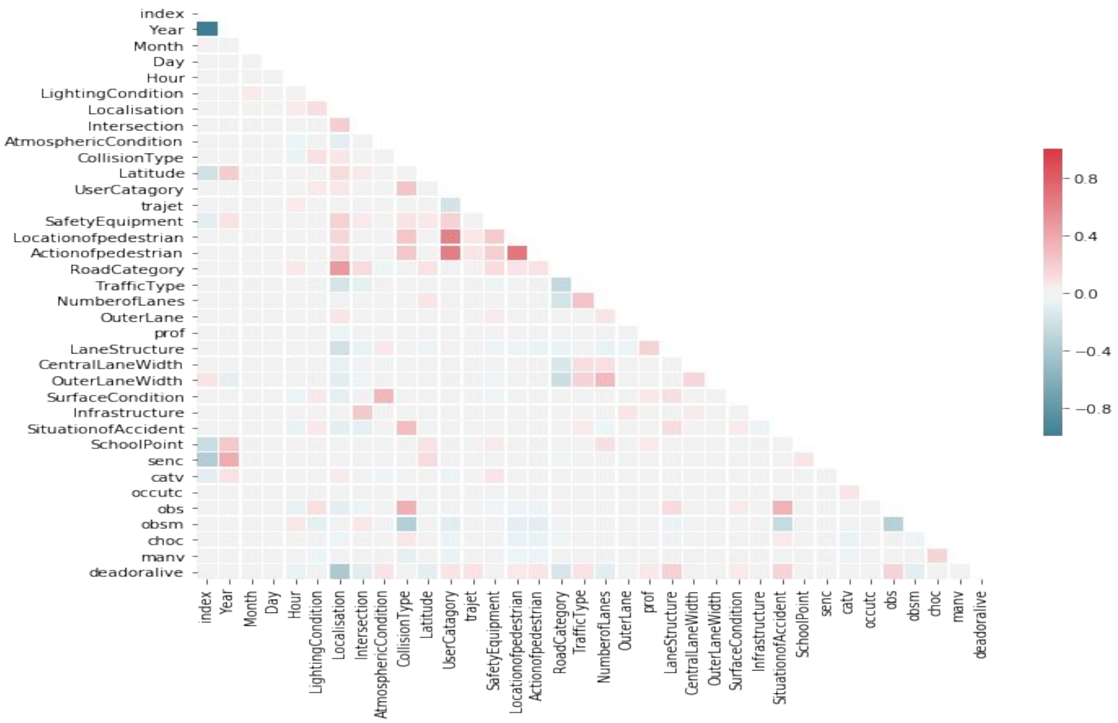
- Sex
- User Year of Birth
- Trip Reason
- Pedestrian Group
- Flow Direction

How big is the data set now?

- Columns: 37
- Rows: 164302

Normalize dataset -> machine learning models can better learn & utilize

Visualize correlations among features



- Very low , non existent correlations
- Action of pedestrian, pedestrian group, location of pedestrian are the exceptions

PCA works best when:

- Features normally distributed
- Linear relationship among features
- Correlation among features weak -> moderate

LAST TWO ASSUMPTIONS ARE NOT MET MOST LIKELY DUE TO HEAVY CATEGORICAL PRESENCE -> STATSMODEL OLS FUNCTION -> FURTHER DIMENSION REDUCTION

Statsmodel ols function -> features to drop with > .05 pvalue:

- Month 1.108377e-01
- Day 4.787969e-01
- Occutc 1.443713e-01
- choc 6.286442e-01

How big is the data set now:

- Columns: 32
- Rows: 164302

Even further Dimension Reduction -> Random Forest Feature Importances :

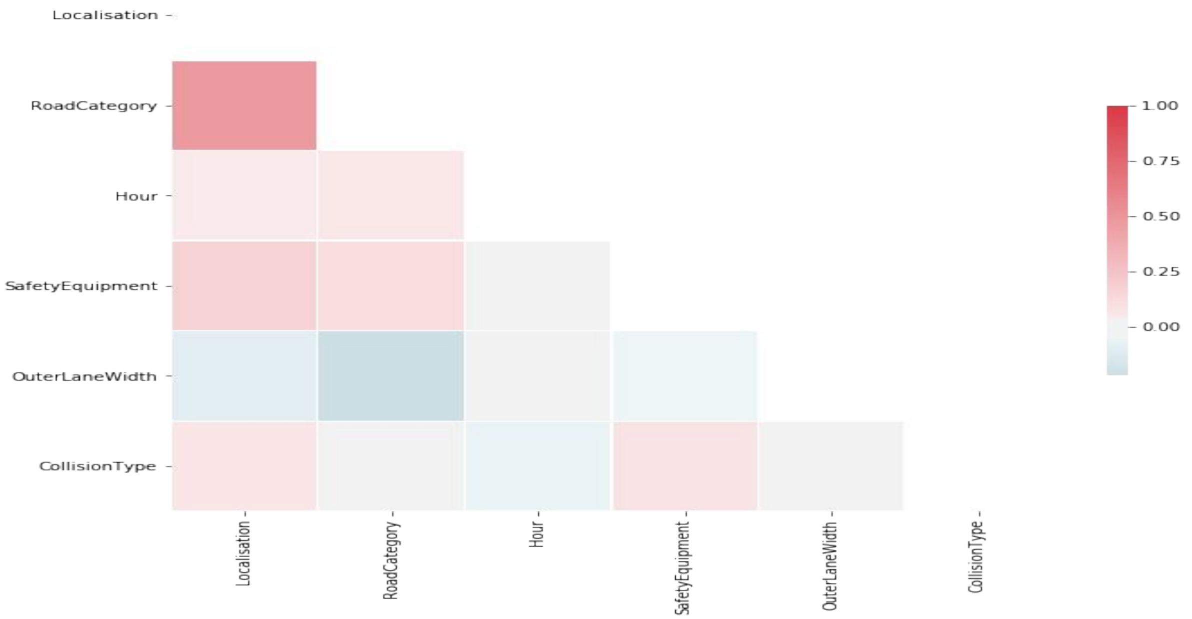
Extract most relevant features:

- Localisation : .082
- Road category: .073
- Hour : .072
- Safety Equipment: .063
- OuterLaneWidth: .060
- CollisionType: .049

How big is the dataset at this point:

- Columns: 6
- Rows: 164302

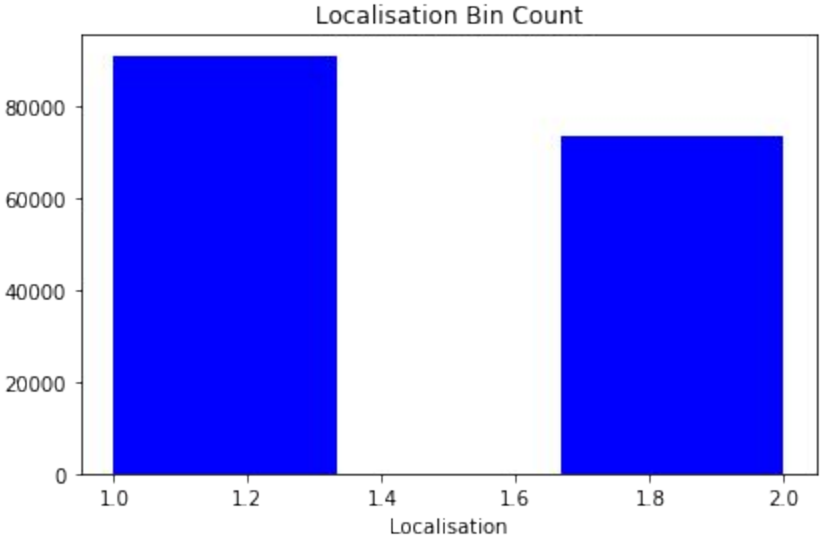
Visualize correlations among dimension again:



- No further dimension reduction via PCA because non linear relationship among features

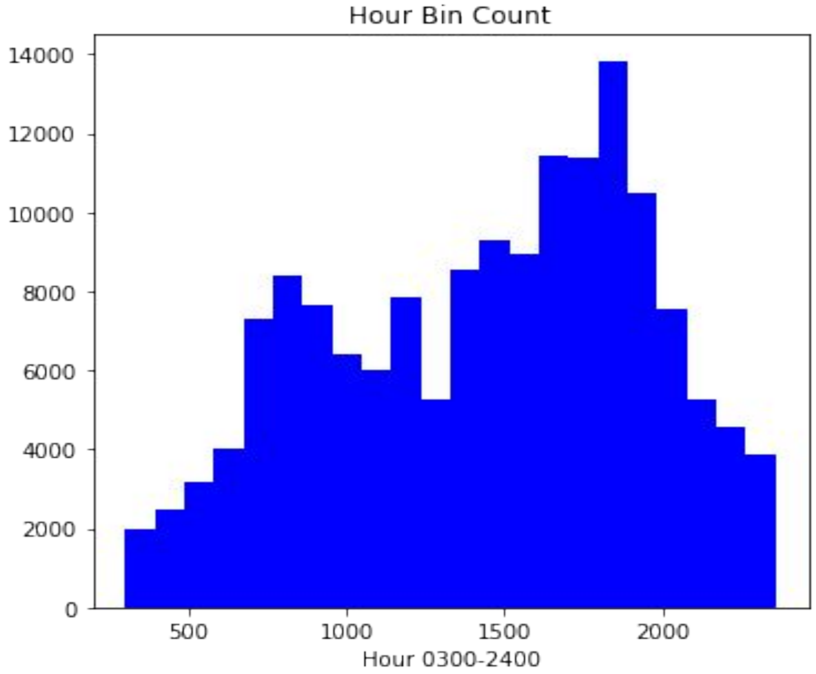
Look more Closely at the Distributions & Frequencies of the relevant features:

Localisation



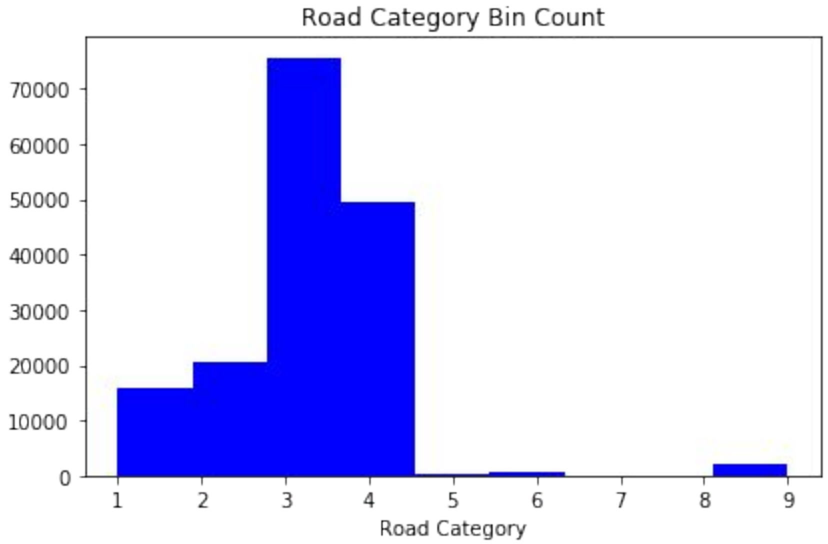
- First bar -> less congested traffic
- Second bar-> heavy traffic congestion
- Counter intuitive -> heavy traffic congestion resulted lower accident count

Hour



- Bimodial distribution
  - 8am : accident count 8000
  - 7pm: accident count 13000
- Indicative of two normal gaussian distributions

Road Category



- Bars increasing -> roads become more local from highway to parkinglot
- Accident count highest for departmental roads

5. Perform In-depth analysis (Modeling, Training, Validating, Testing)

6. Communicate the results of the analysis (data product)

Comparing Accuracy Scores on different models

	acc_sc_val	**acc_sc_test**
NBC	.694	.694
KNN_C	.741	.747
DTC	.733	.735
RFC	**.779**	**.785**
GBC	.761	.762

- All ml classification model accuracy scores’ has higher/same test scores than validation scores -> no overfitting due to low variance

Best Performing Model: RFC Classifier

PROS:

- Ensemble model -> low variance -> avoids overfitting
  - Test set acc sc > validation set acc sc
- Low bias -> avoids underfitting
  - Relatively high validation & test set acc scores
- Most significant features
  - RFC feature importances attribute

CONS:

- Low memory footprint
- Higher relative training time due to higher relative accuracy

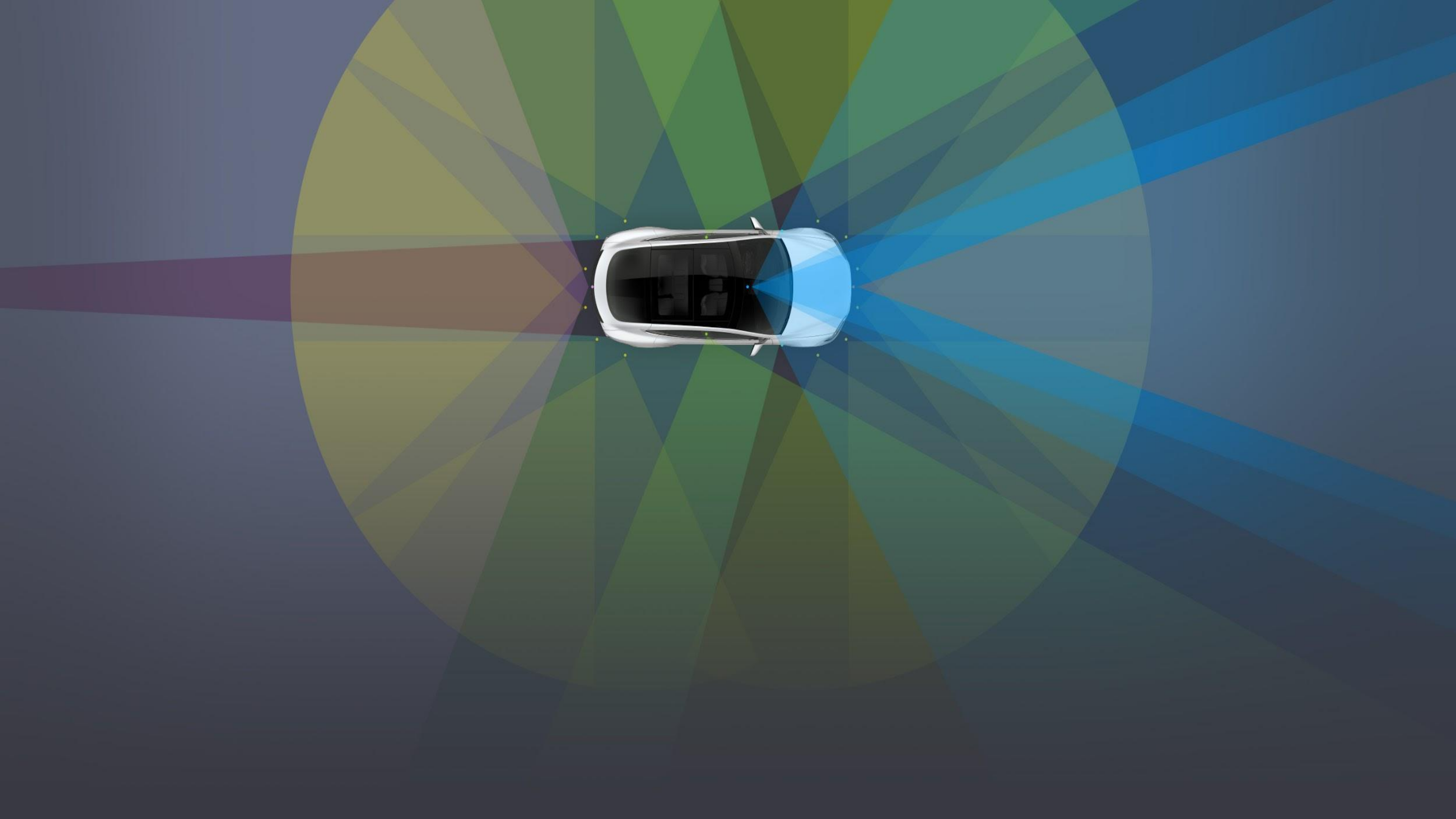
Ranking of said Relevant Features:

- Hour: .368
- OuterLaneWidth: .208
- SafetyEquipment: .117
- RoadCategory: .111
- Localisation: .105
- CollisionType: .092

Practical Uses of results(data product):

- Government uses data product & most relevant features as input for future model for re-engineering of traffic infrastructure like San Diego's smart highway
  - Lower car insurance premium for drivers
- Tesla uses data product as input for its model used in real time telemetry for risk assessment!
- Government enforces limiting registering cars to which meet certain quality standards
  - Insurance companies will raise premiums for those with cars that do not meet said standards
- Insurance companies and car dealerships will realize a short lived initial profit
- **LAST BUT NOT LEAST HUMAN LIFE IS SO VALUABLE IT IS INVALUABLE!!**


Telecom
20.0°
20.0°



Link to the corresponding Jupyter Notebook:

https://github.com/pman117/thinkful-data-science/blob/master/supervised_learning_capstone/DS_U3_Supervised_Learning_Capstone.ipynb