NCT pass rate analysis and forecast for given car using Data mining methods

Table of Contents

[Introduction 2](#_Toc510705373)

[1. Problem Statement 2](#_Toc510705374)

[2. Data Description 2](#_Toc510705375)

[3. Data Preparation 4](#_Toc510705376)

[4. Data mining model construction 4](#_Toc510705377)

# Introduction

Project was made for Data and web mining module. The goal is to predict result of car test – NCT in Ireland.

# Problem Statement

# Data Description

Dataset is a dataset provided by NCT test center. It is publicly available information and can be found on <https://data.gov.ie/dataset/2016-make-model-year-failures-at-nct> .

After simple analysis in Python we can say that:

It contains 8089 rows and 37 columns. We are able to see all model and make of car of all cars who made NCT test in Ireland in 2016. In columns we have attributes:

VehicleMake

VehicleModel

YearOfBirth

Total

PASS

PASS %

FAIL

FAIL %

Vehicle and Safety Equipment

Vehicle and Safety Equipment %

Lighting and Electrical

Lighting and Electrical %

Steering and Suspension

Steering and Suspension %

Braking Equipment

Braking Equipment %

Wheels and Tyres

Wheels and Tyres %

Engine, Noise and Exhaust

Engine, Noise and Exhaust %

Chassis and Body

Chassis and Body %

Side Slip Test

Side Slip Test %

Suspension Test

Suspension Test %

Light test

Light test %

Brake Test

Brake Test %

Emmissions

Emmissions %

OTHER

OTHER %

Incompletable

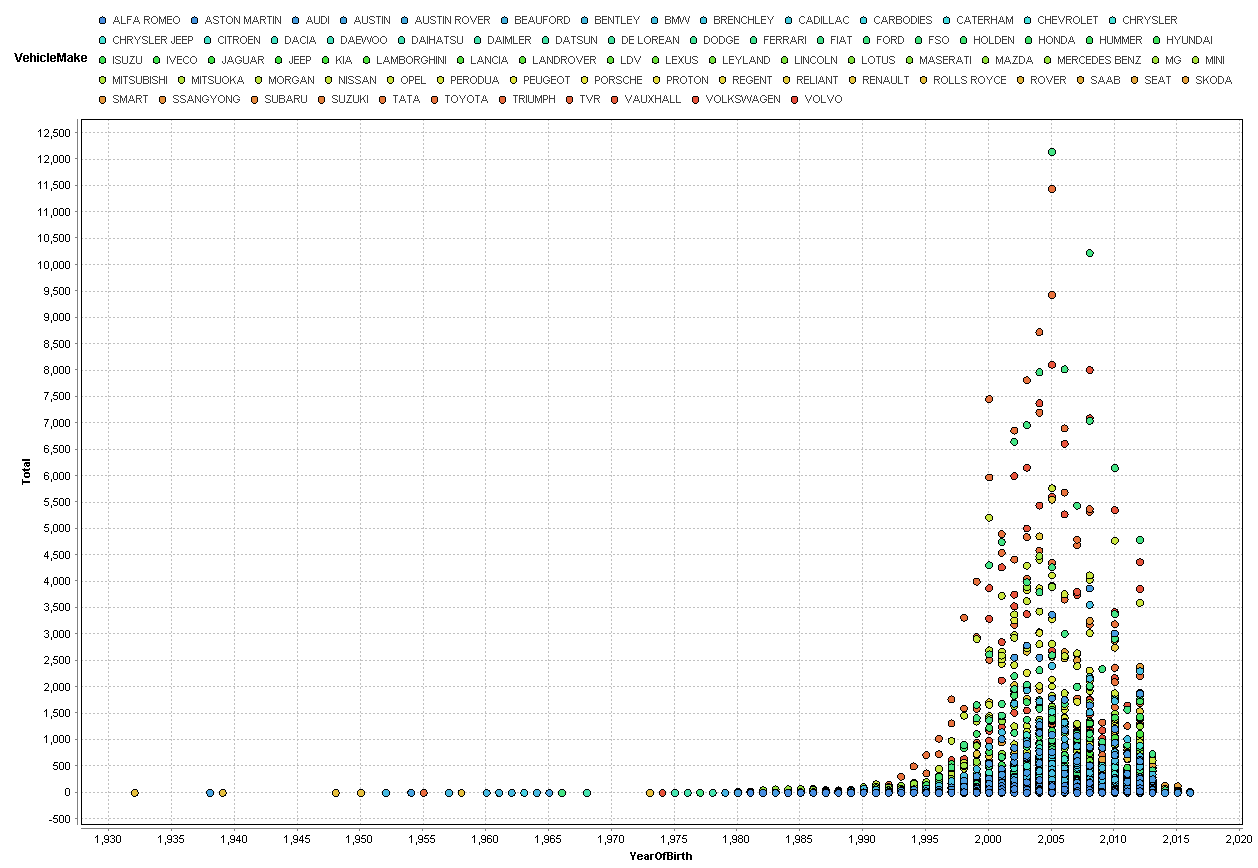
Incompletable %

Country

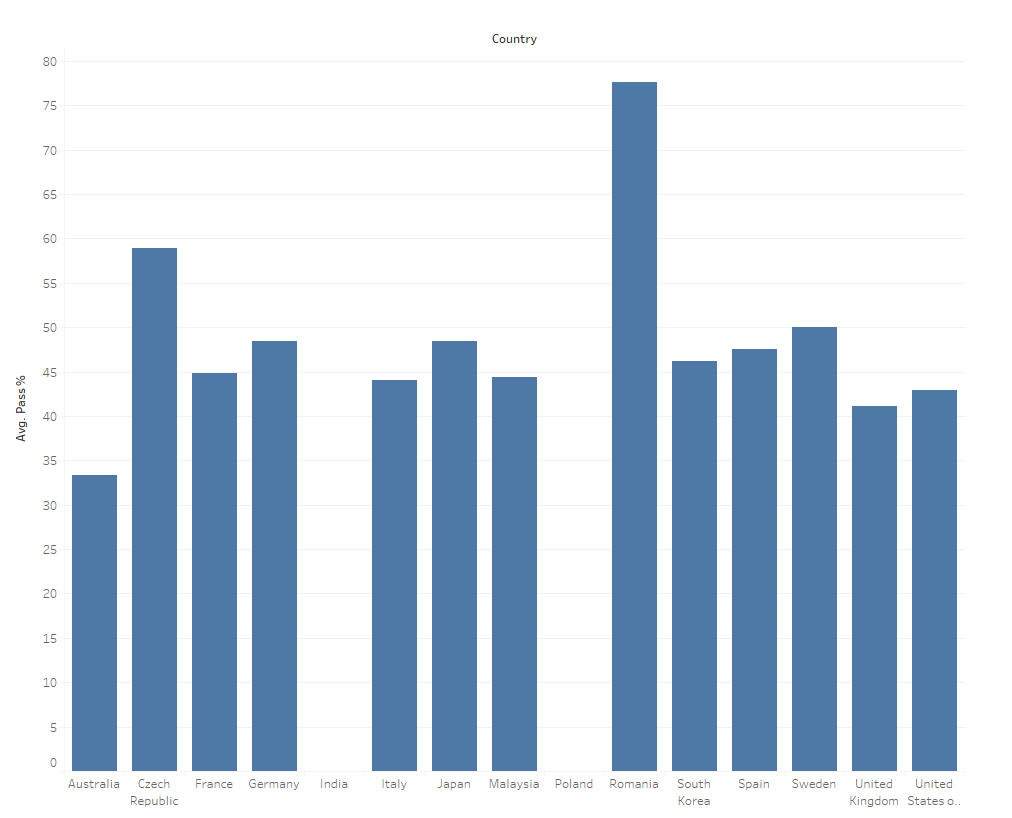
Not all attributes are integer but most are numerical. Some like Make and Model and Country are strings.

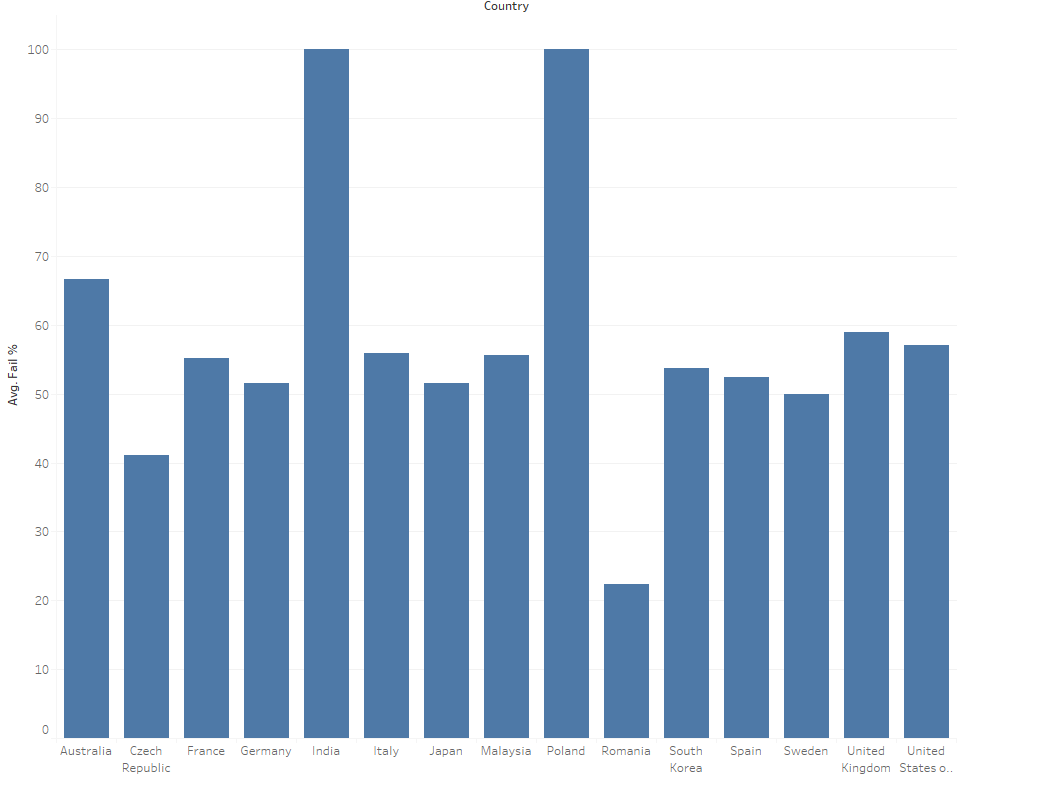
Rapidminer studio have amazing tool for rapid visual representation of data. We made simple charts to understand distribution of data in our dataset.

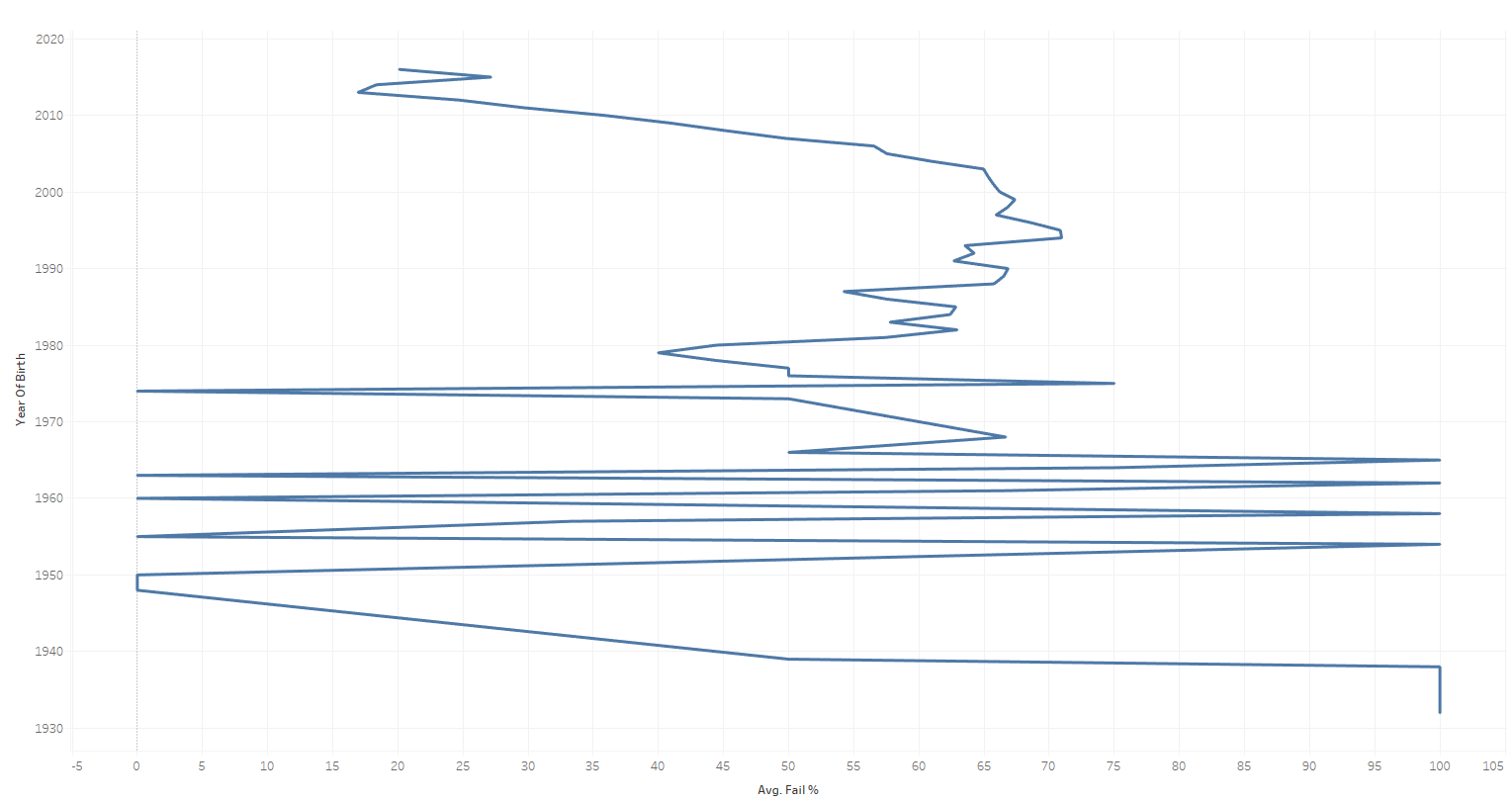
For example for Year of Birth:



Also tableau was helpful in this part of work:







# Data Preparation

Data has been checked for all missing values, duplicates and outliers.

There are no NAN values in our dataset. All duplicates has been carefully checked and are normal like for example in total column we have number of cars of this make and model which took part in test in 2016. It is normal that the numbers will have duplicates. Not formal will be in Model column but that has been checked too.

A column with YearofBirth has been check for entries. It should contain only year of production and all data seems to be correct. No outliers have been found.

There was a not common entry in column with Coutries. ‘c’. It was Audi, so we change that to Germany.

Dataset after every attribute column contains this attribute percentage of failing that particular test. That’s ratios will be very useful in our model.

# Data mining model construction

## Classification rules

Firstly we focused on classification rules for our model.