**СОФИЙСКИ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“**

**ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА**

Картина, която съдържа текст, графична колекция

Описанието е генерирано автоматично

**СТАТИСТИКА И ЕМПИРИЧНИ МЕТОДИ – ПРАКТИКУМ**

**КУРСОВ ПРОЕКТ**

**Тема:**

**Изследване на зависимости между видове автомобили според двигателя и техните годишни данъци и разход на гориво**

**Изготвил:**

**Габриел Миндев, ФН 71846**

**23 Август 2021 г.**

1. **Въведение**
   1. **Цели на проекта**

Целта на проекта е да се изследва дали има зависимост между вида на двигателя на проучваните коли и съответно техния данък и разход на гориво.

* 1. **Описание на данните**

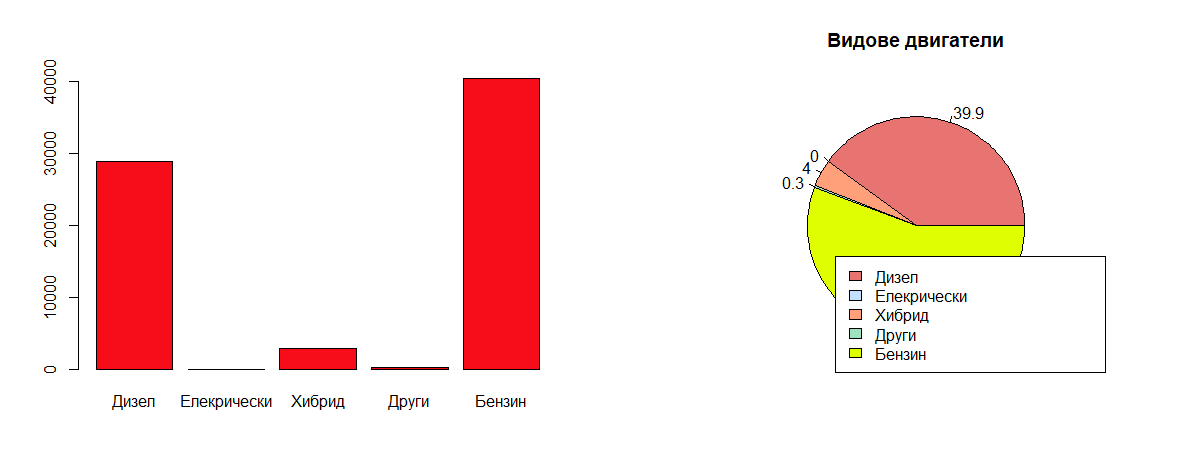
Използваните данни са взети от сайта Kaggle -линк към данните: https://www.kaggle.com/aishwaryamuthukumar/cars-dataset-audi-bmw-ford-hyundai-skoda-vw, като в случая ползваме статистическите данни за употребявани автомобили от Великобритания . Като аз разглеждам следните от тях :

* Вид на двигателя - категорийна променлива
* Годишен данък(в британски лири) - числова променлива
* Разход на гориво(мили за галон)- - числова променлива

Първо изследваме как са разпределени данните поотделно, техните локация и разсейване. След това ще видим дали категорийната променлива може да е обясняваща за всяка от числовите и ще търсим дали има връзка между вида двигател и неговия данък, както и неговия разход.

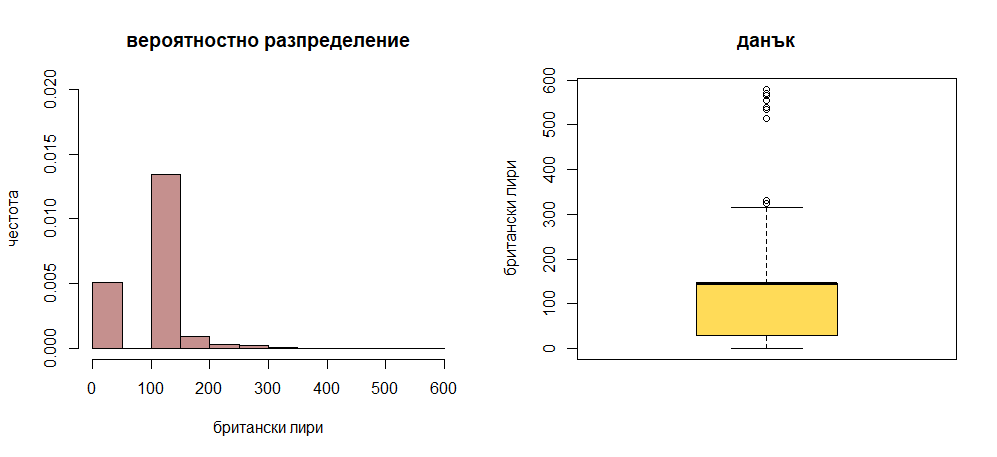
1. **Изследване на променливите поотделно**
   1. **Вид двигател - категорийна**

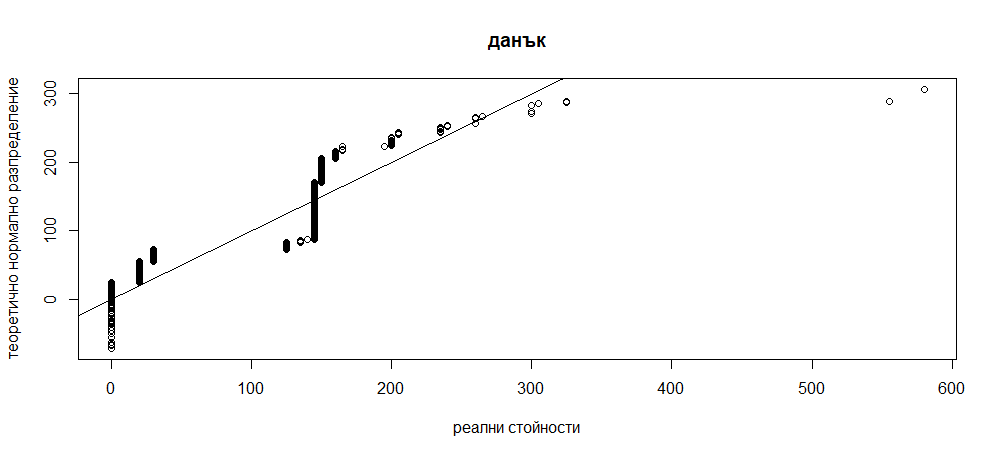
От графиката виждаме, че конвенционалните двигатели като дизела и бензина имат голямо преимущество в бройката, което е и логично, тъй като хибридните и най-вече електрическите автомобили тепърва навлизат по-масово.

****

Това разпределение до известна степен пречи на нашето изследване, тъй като голямата разлика в броя би изкривило до някаква степен резултатите.

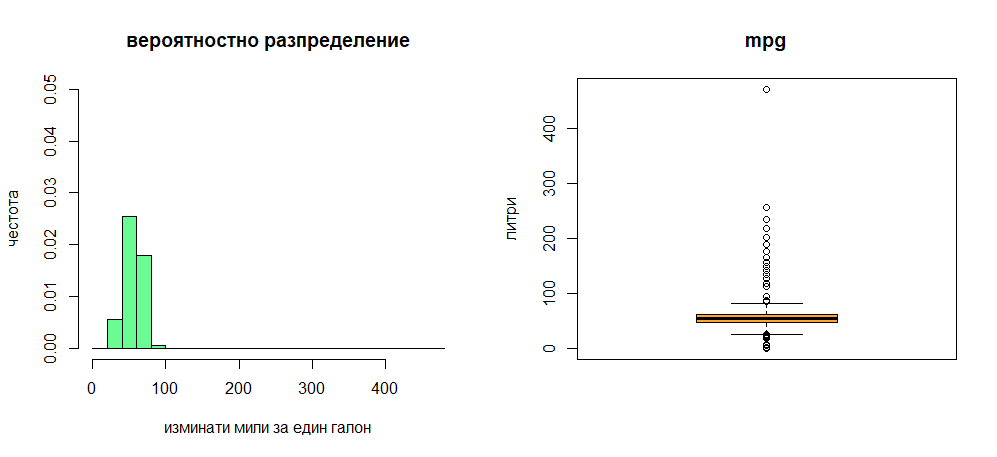
* 1. **Годишен данък - числова**

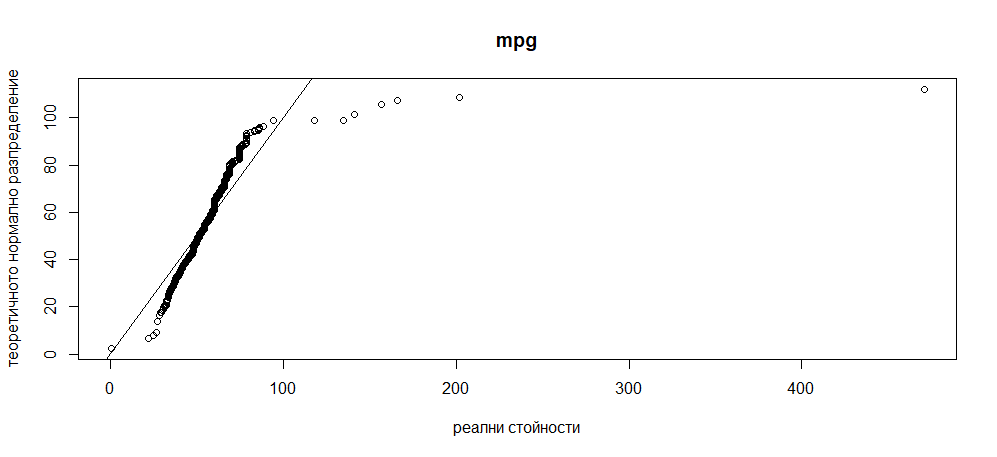
****

От графиките можем да видим, че данъците варират от 0 лири до 500-600 лири, където имаме съответно outlier-и. Но като средна стойност данъците се държат около 150 лири. Също така от тях можем да заключим, че разпределението не е нормално, което се потвърждава и от теста на Kolmogorov-Smirnov, който показва p-value< 2.2e-16

Използваме непараметрични оценки за намиране на локацията и дисперсията, т.е. средно аритметично и стандартно отклонение.

* локация= round(median(tax), 3) = 145
* дисперсия=round(mad(tax), 3)= 7.413
  1. **Разход на гориво (мили за галон) - числова**

****

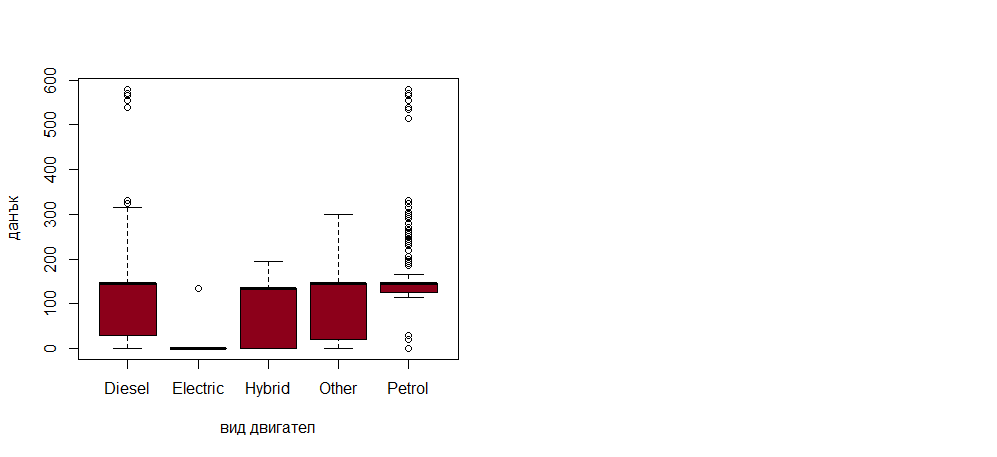
****

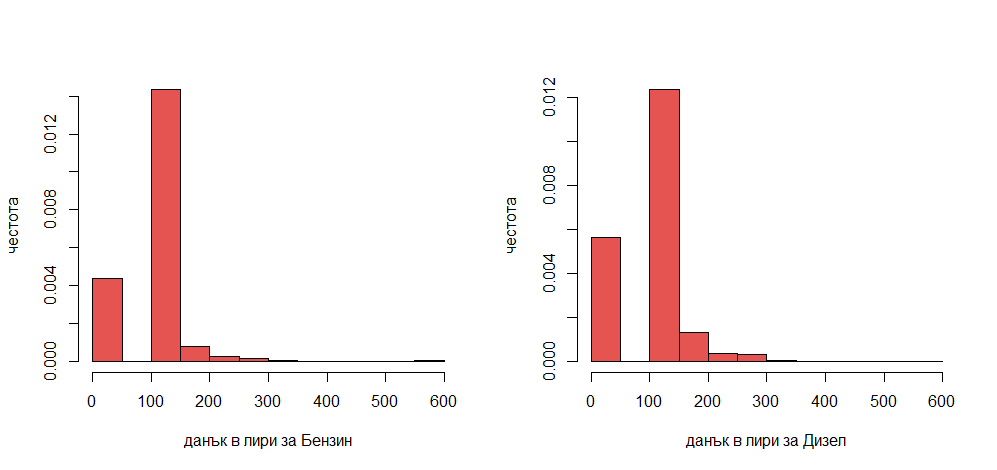
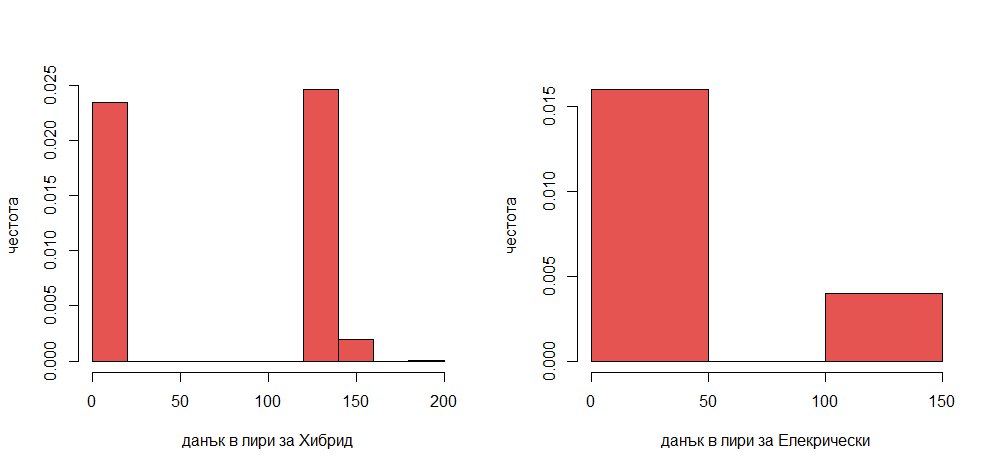
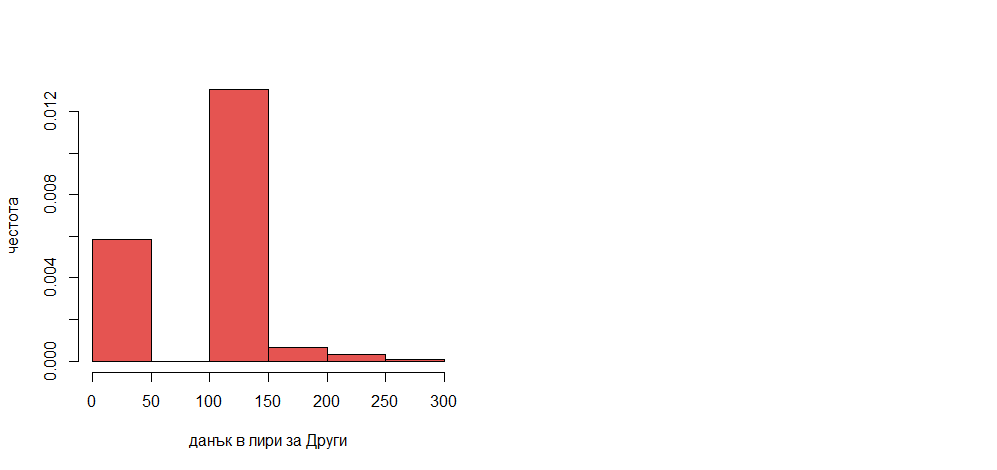
При тези графики тук също виждаме, че нямаме нормално разпределение поради наличието на доста outlier-и , което показва и теста за нормалност, който пак показва p-value<2.2e-16

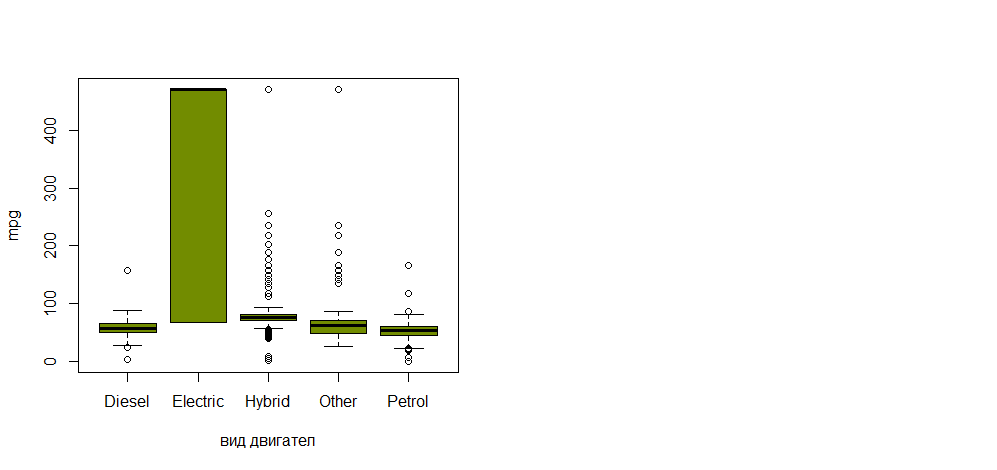
За локацията и дисперсията намираме, че са:

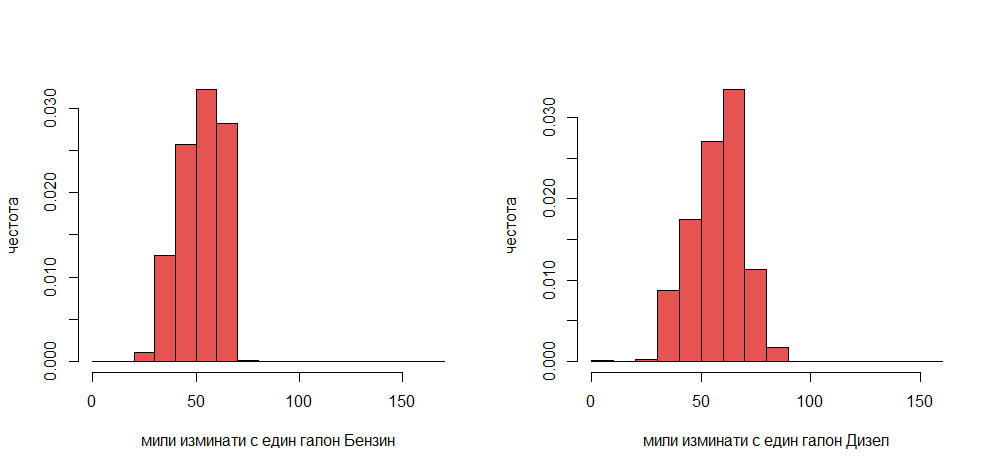
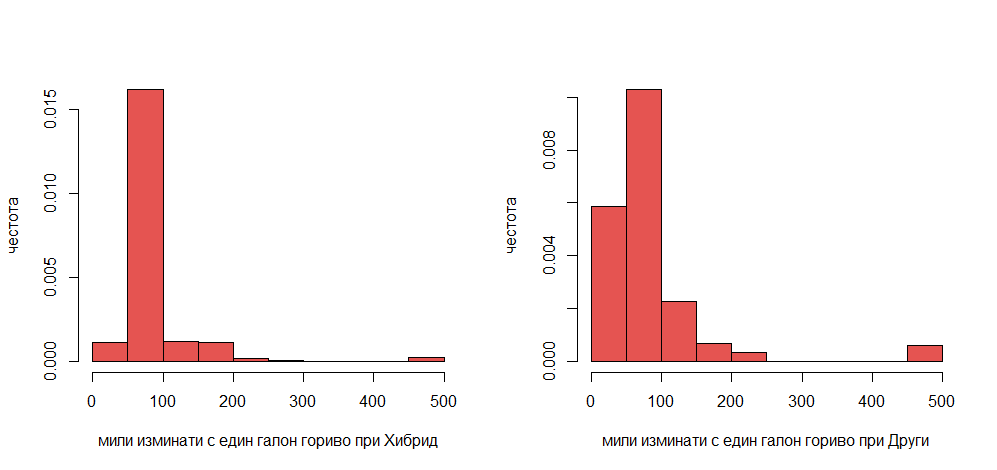
* локация = round(median(mpg), 3) = 55.4
* дисперсия = round(mad(mpg), 3) = 11.119

1. **Изследване на взаимодействия между променливите**
   1. **Категорийни vs Числови**

**3.1.1 Вид двигател и данък**

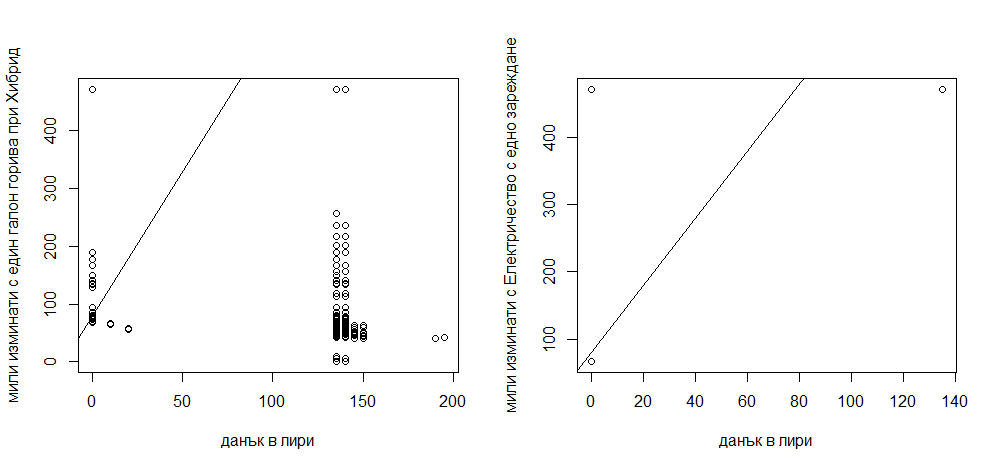
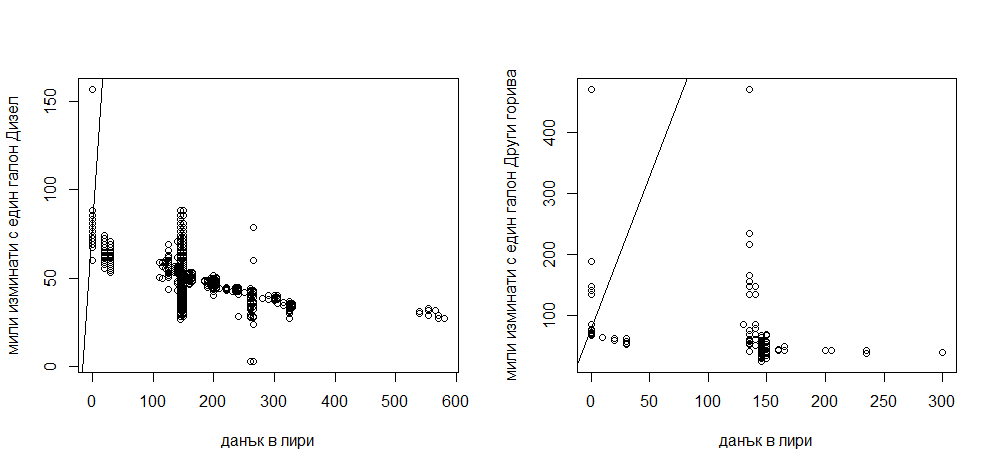
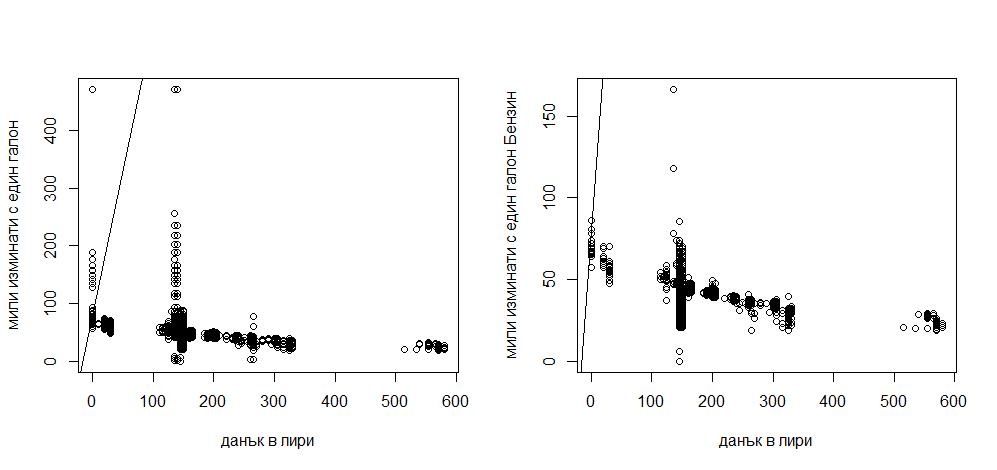
От графиката забелязваме, че електрическите автомобили са изцяло освободени от данък с изключение на един outlier за разлика от бензиновите например ,които изцяло са над 100-те лири с малки изключения. Като можем да заключим, че конвенционалните автомобили са определено с по-висок данък. Ако разгледаме данните, като разделим видовете двигатели в отделни променливи от хистограмата изглежда, че те не са с нормално разпределение.

**3.1.2 Вид двигател и разход на гориво**

 От графиката забелязваме, че електрическите и хибридните автомобили изминават по-дълго разстояние като виждаме от това, че първия и третия им квартил е над третия на останалите.

От съответните хистограми по вид двигател виждаме, че отново нямаме нормално разпределение.

* 1. **.Числови vs Числови**

Ще търсим дали може да има връзка между големината на данъка и разхода на гориво

В първата графика разглеждаме данъка и разхода съвкупно за всички видове двигатели. Ако разгледаме корелацията при съвкупното разглеждане тя е 0.579, което говори за средна връзка. При пресмятането ѝ използваме непараметрична оценка метода на Spearman, защото данните не са разпределени нормално.

При разглеждането отделно по видове двигатели намираме, че:

-При бензиновите връзката е 0.47, което също е сравнително средна

- При дизеловите връзката е 0.664, което отива към сравнително силна

-При автомобилите с друг вид гориво връзката е 0.672, което също е сравнително силна

-При хибридните връзката е 0.26, което е слаба

-При електрическите връзката е 0.408, което клони към слаба.

1. **Заключение**

Предвид разгледаните визуализации и получените резултати можем да обобщим, че като цяло данъка и разхода на гориво се влияят от вида двигател като има тенденция електрическите и хибридните автомобили да са с по-нисък и даже нулев данък като същевременно с това са и с по-нисък разход от конвенционалните автомобили.