194.045 Data Stewardship (UE 2,0) 2020S

Exercise 1
Aleksandar Sibincic (0727895)

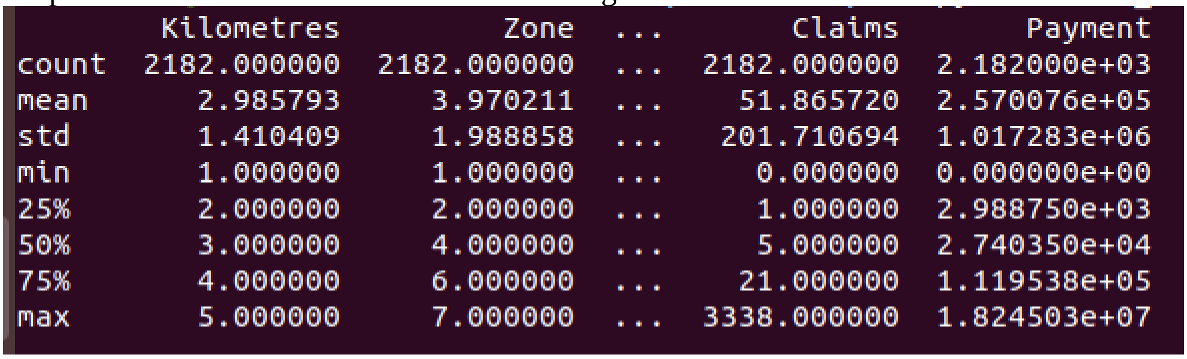
Experiment

For this exercise I chose the experiment from the first exercise of the course *184.702 Machine Learning (VU 3,0)*. The dataset used here is from Swedish Motor Insurance . This data describes third party automobile insurance claims for the year 1977 in Sweden. The goal here was to analyze the data and to perform simple linear regression for the 2 points of interest in this dataset: number of claims (the frequency) and sum of payments (the severity). In order to do that, we had to split the dataset into training and test set, apply predictions in the test set according to a training and to come up with measures for our predictions. The complete code in Python for this experiment is uploaded to Tuwel as file **smi_regression.py**

Flow

Analyzing data set

Our data set consists of the 9 dimensions(columns) and 2182 samples (rows).
The output of the function that describes our set is given as follows:

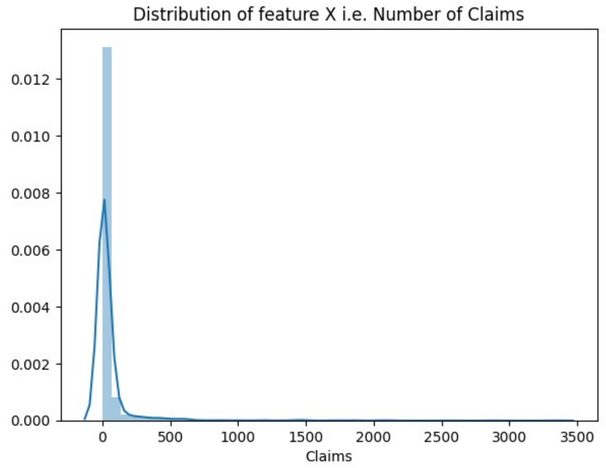


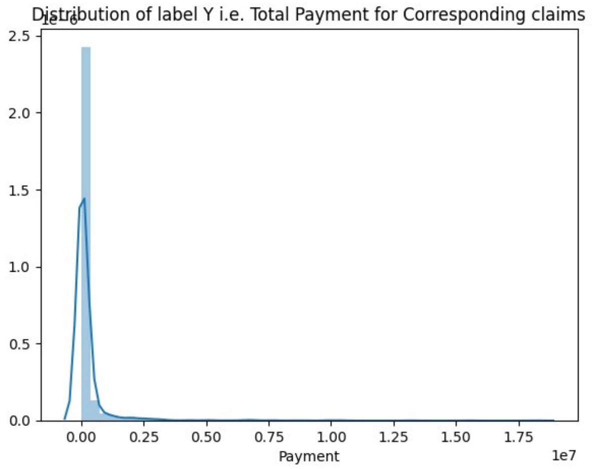
```
        Kilometres         Zone  ...       Claims       Payment
count  2182.000000  2182.000000  ...  2182.000000  2.182000e+03
mean      2.985793     3.970211  ...    51.865720  2.570076e+05
std       1.410409     1.988858  ...   201.710694  1.017283e+06
min       1.000000     1.000000  ...     0.000000  0.000000e+00
25%       2.000000     2.000000  ...     1.000000  2.988750e+03
50%       3.000000     4.000000  ...     5.000000  2.740350e+04
75%       4.000000     6.000000  ...    21.000000  1.119538e+05
max       5.000000     7.000000  ...  3338.000000  1.824503e+07
```

After that we search for missing values. This data set is clean and has no missing values so data imputation as preprocessing step is not needed.

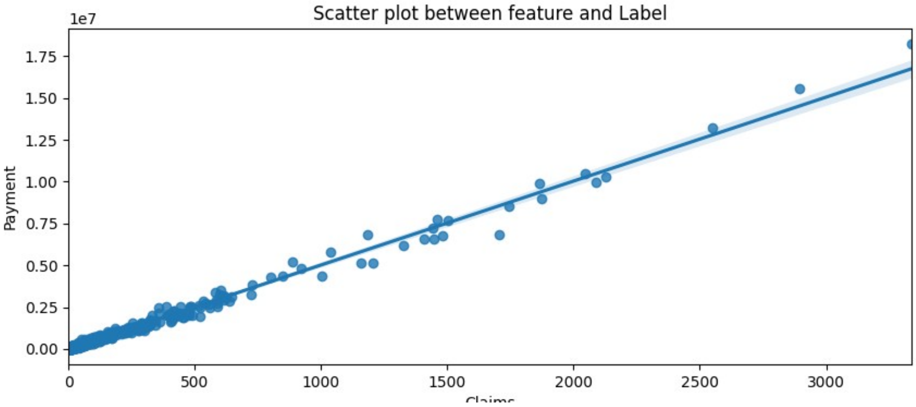
Analyzing our points of interests

As already said, the goal here is to predict sum of total payments based on the number of claims. Therefore, we will need only data from the last two columns. By visualizing those data we get the following output:





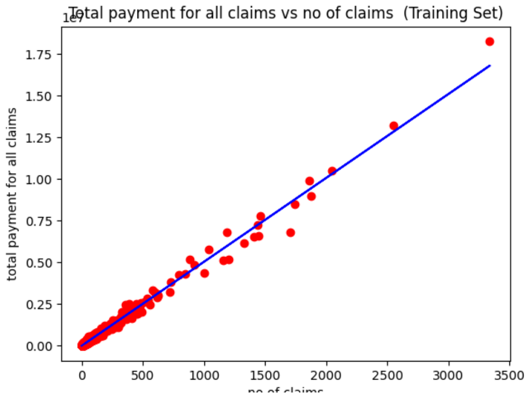
As we can clearly see the distributions have approximately the same shape which indicates that there is a strong relationship the number of claims and sum of payments.

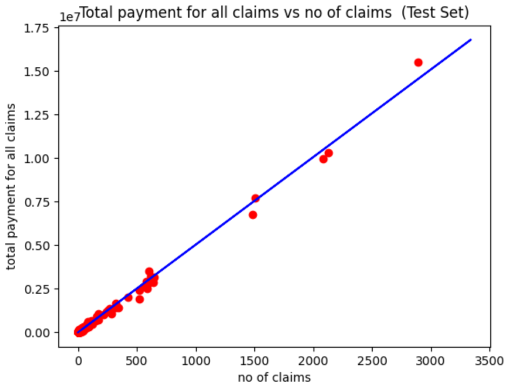


Also, they fit approximately the same regression line.

Splitting the data set

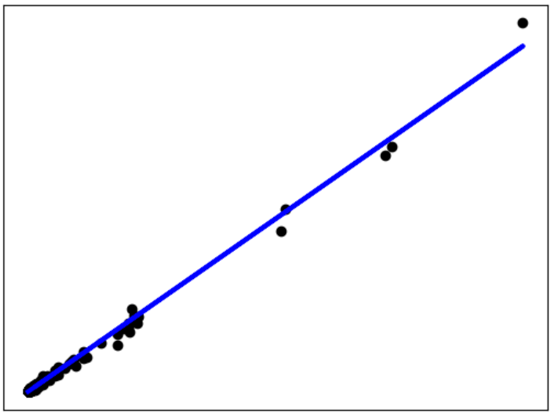
In this step we split our set into training set and test set. Here, we chose 80% for the training set and 20% for the test set. We also save those datasets. By visualizing those 2 sets, we get following:





Predicting the Test set result

Using simple linear regression, we were able to predict the sum in the test set. We save our predictions in csv file. We can see that the data don’t lie far from the regression line.



Measuring predictions

In order to measure our predictions, we use Root Mean Square Deviation. And our output is:
RMSE: 94170.81