In the [moodle page](https://moodle2.units.it/pluginfile.php/414751/mod_resource/content/2/exam_projects2021-2022.html) the professors write a list of variables that we need to understand and study for predicting the response variable. In this first table there are all the variables with a brief description, an abbreviation or an English term, that we’re going to use in R as a name of the variables. The dataset with this data is located on the [project folder on drive](https://drive.google.com/drive/folders/1BlDFj87se3mJL6vGlC-PxogRkECPPBpp).

|  |  |
| --- | --- |
| **Variable** | **Description** |
| date | Date of notification |
| state | Country of reference |
| region\_code | Code of the Region |
| region\_name | Name of the Region |
| lat | Latitude of the Region |
| long | Longitude of the Region |
| patients\_hospitalized | Patients in Hospital with symptom |
| patients\_intensive-care | Patients in Intensive Care (Y) |
| total\_patients\_hospitalized | Patients in Hospital.  patients-hospitalized + patients-intensive-care |
| home\_confinement | Home confinement |
| positive | Total amount of current positive cases  total-patients-hospitalized + home-confinement |
| positive\_variation | Variation of positive current cases  positive (curr. day) - positive (prev. day) |
| new\_positive | New amount of current positive cases  total-cases (curr. day) - total-cases (prev. day) |
| recovered | Recovered |
| death | Death (cumulated values) |
| total\_cases | Total amount of positive cases |
| tests\_performed | Tests performed |
| total\_people\_tested | Total number of people tested |
| intensive\_care\_admission | Daily admissions to intensive care |

In the following table we reported some ideas of correlation just looking to the variables and their data.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Corr.** | **Notes** | **Cumulative?** |
| date |  | Maybe there is a correlation with the period? Example: in December there are more intensive care cases. | NO |
| state |  | Same value for all the dataset “ITA” | NO |
| region\_code |  | Same value for all the dataset “8” | NO |
| region\_name |  | Same value for all the dataset “Emilia” | NO |
| lat |  | Same value for all the dataset “..” | NO |
| long |  | Same value for all the dataset “..” | NO |
| patients\_hospitalized |  | Quantitative variable | NO |
| total\_patients\_hospitalized |  | Quantitative variable | NO |
| home\_confinement |  | Quantitative variable | NO |
| positive |  | Quantitative variable | NO |
| positive\_variation |  | Quantitative variable | NO |
| new\_positive |  | Quantitative variable | NO |
| recovered |  | Quantitative variable | YES |
| death |  | Quantitative variable | YES |
| total\_cases |  | Quantitative variable | YES |
| tests\_performed |  | Quantitative variable | YES |
| total\_people\_tested |  | Quantitative variable | YES |
| intensive\_care\_admission |  | Quantitative variable | NO |

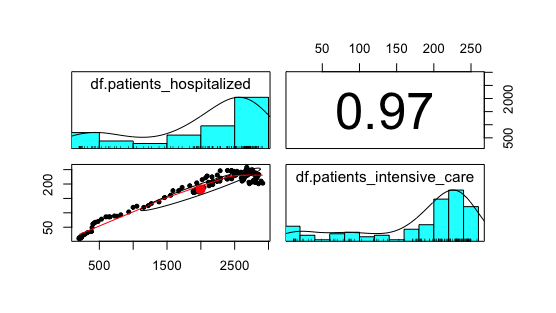
A For sure no correlation

A Maybe there is a correlation. We need to plot and see if there is a correlation

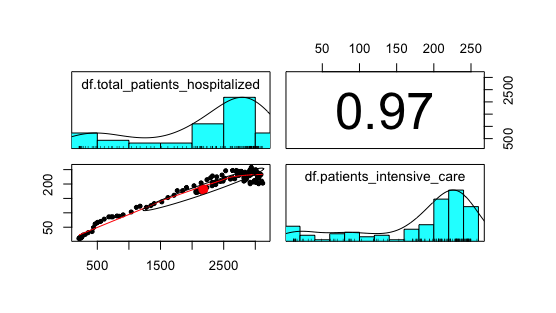
There are not NA row because we take the data from 01/10/2020 to 01/02/2021.

Opinion:

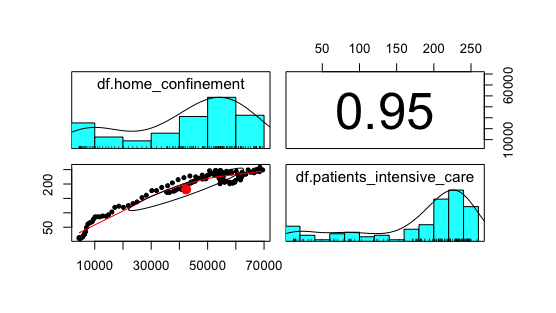
* **data**: we need to transform the date in some way in order to plot it;
* **patients\_hospitalized**: there is a strong correlation;



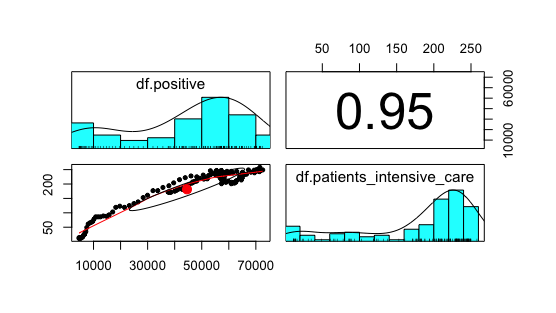
* **total\_patients\_hospitalized**: there is a strong correlation;



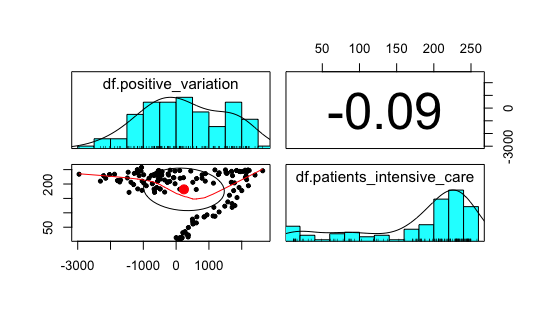
* **home\_confinement**: there is a strong correlation. (if there are more persons in confined at home there are more persons in intensive care)



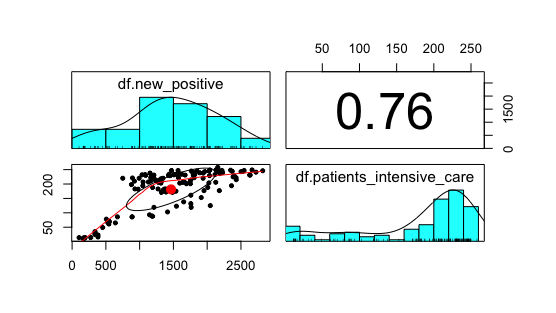
* **positive**: there is a strong correlation.



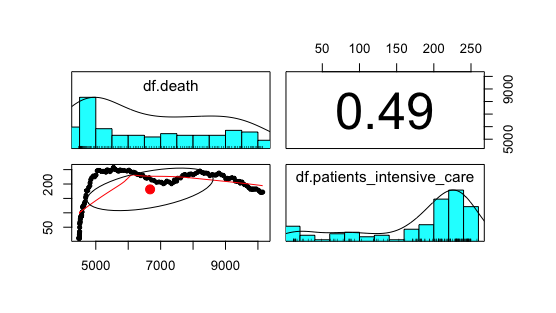
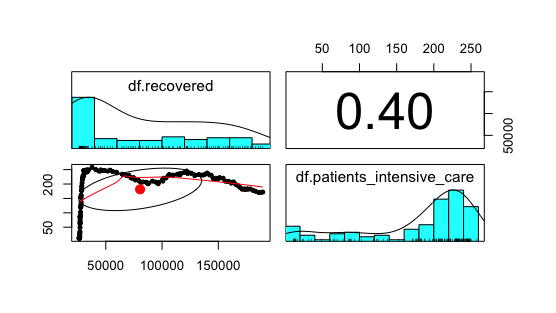
* **positive\_variation**: NEED TO LOOK TO THIS DATA MAYBE WITH THE COLOR OF THE REGION OR THE DATE.

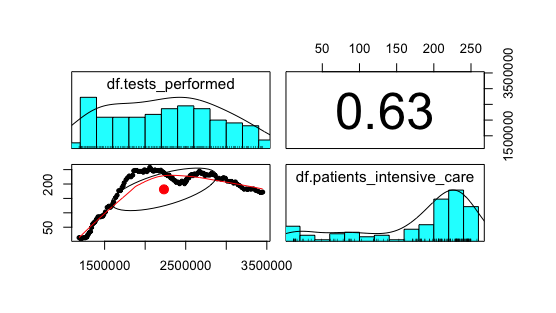
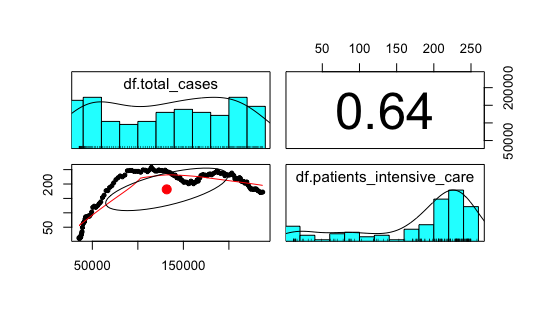


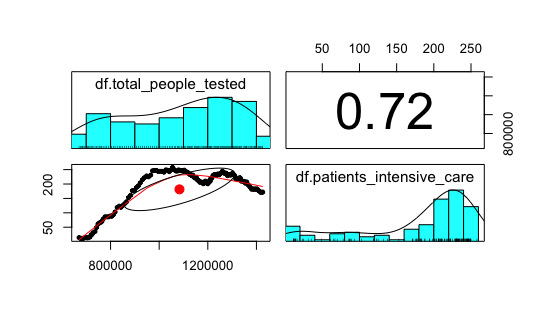
* **new\_positive**: NEED TO LOOK TO THIS DATA MAYBE WITH THE COLOR OF THE REGION OR THE DATE.



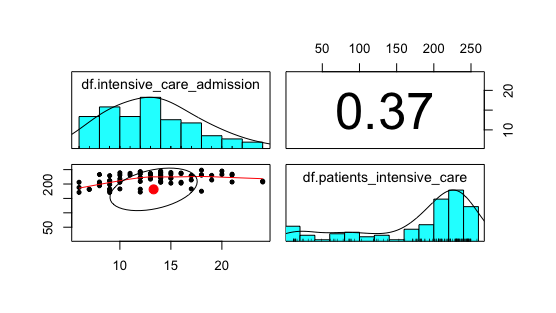
* **recovered, death, total cases, tests performed, total people tested**: these variables are cumulative. Does it make sense to use these variables as covariates?







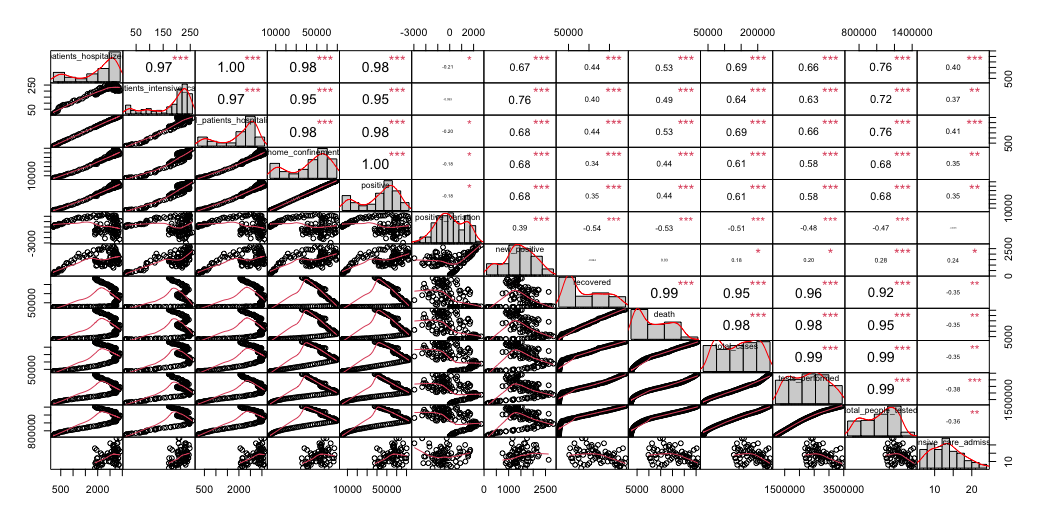
* **intensive care admission**: there is a correlation with the intensive care admission but maybe is not too relevant. We need to train a model with this variable and without this variable



Recap of the results:

|  |  |  |
| --- | --- | --- |
| **Variables** | **Correlation** | **Pearson** |
| patients\_hospitalized | YES | 0.97 |
| total\_patients\_hospitalized | YES | 0.97 |
| home\_confinement | YES | 0.95 |
| Positive | YES | 0.95 |
| positive\_variation | NO | -0.9 |
| new\_positive | YES | 0.76 |
| Recovered | need cumulative covariant? | 0.40 |
| Death | need cumulative covariant? | 0.49 |
| total\_cases | need cumulative covariant? | 0.64 |
| tests\_performed | need cumulative covariant? | 0.63 |
| total\_people\_tested | need cumulative covariant? | 0.72 |
| intensive\_care\_admission | IDK (is it relevant)? | 0.37 |

List of all possible correlation (numeric)



Possible methods that we can use:

* Linear Models
* Generalized Linear Models
* Regression Tree (Bagging, Boosting, RF)

<https://kevintshoemaker.github.io/NRES-746/RandomForests.html>

* Nonlinear Regression (GAM)

Ideas:

* We can compare GLM and GAM like the professor did in lab 6.
* We can use an approach top down like Stefanucci told us.