# Data of COVID-19 Country\_wise\_latest statistics

Data that is used in this paper is got from the Kaggle data center website and it shows statistics regarding the COVID-19 that was first case was identified in the Wuhan china Hubei Province and as 30 January 2020 it shows almost 8243 case throughout the world [1].

This Covid-19 data illustrate number of confirmed, Deaths, Recovered, active, new cases, new cases, new death, new recovered, new deaths, new recovered, deaths in 100 cases and recovered in 100 cases. It is a summary of our COVID-19 data Table 2 that shows name of the countries and its attributes. Name of each variables are defined as below in the data since it results in better and clear graph with no complexity Table 1.

TABLE I. Parameters description

|  |  |
| --- | --- |
| Parameters Description | Symbols |
| Confirmed | C |
| Deaths | D |
| Recovered | R |
| Active | A |
| New cases | NC |
| New Deaths | ND |
| New recovered | NR |

TABLE II. COVID-19 Dataset

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Country | C | D | R | A | NC | ND | NR |
| Afghanistan | 36263 | 1269 | 25198 | 9796 | 106 | 10 | 18 |
| Albania | 4880 | 144 | 2745 | 1991 | 117 | 6 | 63 |
| Algeria | 27973 | 1163 | 18837 | 7973 | 616 | 8 | 749 |
| Andorra | 907 | 52 | 803 | 52 | 10 | 0 | 0 |
| Angola | 950 | 41 | 242 | 667 | 18 | 1 | 0 |
| Antigua | 86 | 3 | 65 | 18 | 4 | 0 | 5 |
| Armenia | 37390 | 711 | 26665 | 10014 | 73 | 6 | 187 |
| Australia | 15303 | 167 | 9311 | 5825 | 368 | 6 | 137 |
| Austria | 20558 | 713 | 18246 | 1599 | 86 | 1 | 37 |
| Azerbaijan | 30446 | 423 | 23242 | 6781 | 396 | 6 | 558 |

First step in the visualization of the data is boxplot in this paper. It can illustrate graphically a large group of numerical data through its quartiles. As it is shown in the fig 2 the data are standardized, and all of the features have outliers. Plus, D100 (death in 100 case) distribution shows very similar behavior to the normal distribution.

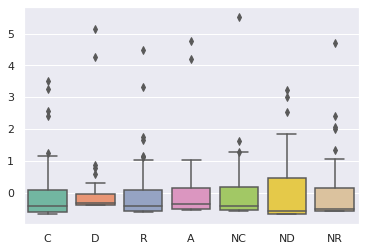


Fig 1. Box plot of COVID-19 data

According to the Fig 1 the distribution of the data is shown after centering the data. The outliers are shown in the Fig 2 for all the feature of the data. The biggest median is related to the Death (death case) in COVID-19 data and the lowest is the related to ND (New deaths) cases [1].

Covariance matrix Fig 2 is a table that illustrate the covariance between all element with itself [2]. As it is shown in Fig 2 confirm cases highly shows correlation with other variable, although it is having a negative correlation with NR (new recovered cases) and positively correlated with R (recovered cases). The highest positive correlation exists between R (Recovered cases) and C this is clearly illustrated in the pairplot Fig 3. The highest negative correlation exists between NR and D. plus NR has very low correlation with C and R and this is shown in the pairplot.

Pair plot is a very clear method for visualization of relation between many variables in a data. It is a very clear visualization that show how relationship between two variables contribute to better extracting information form the data. The high correlation between R and C is shown in pairplot by having linear relation in scatter plot. So, by increasing in the R, C would increase.

# PRINCIPLE COMPONENT ANALYSIS

Principal component analysis (PCA) is a statistical method that uses multivariate methods in order to obtain vital data out of the data set and converting them I order to reach orthogonal variables. PCA to reduce the dimension of the data meanwhile having the most important variation in the data.

Original data can be converted to uncorrelated data by using PCA. Generally, a data that has a dimension of the I\*J should be considered in advanced and then we go start PCA analysis. In this state minimum (I,J-1) would be the number of PCA. In advanced culmination of teach column would be zero which mean it would be zero centered. In the following picture the covariance matrix of zero is shown Fig 2.

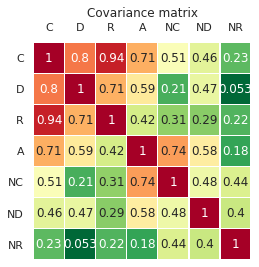


Fig. 2. Covariance matrix of zero centered data.

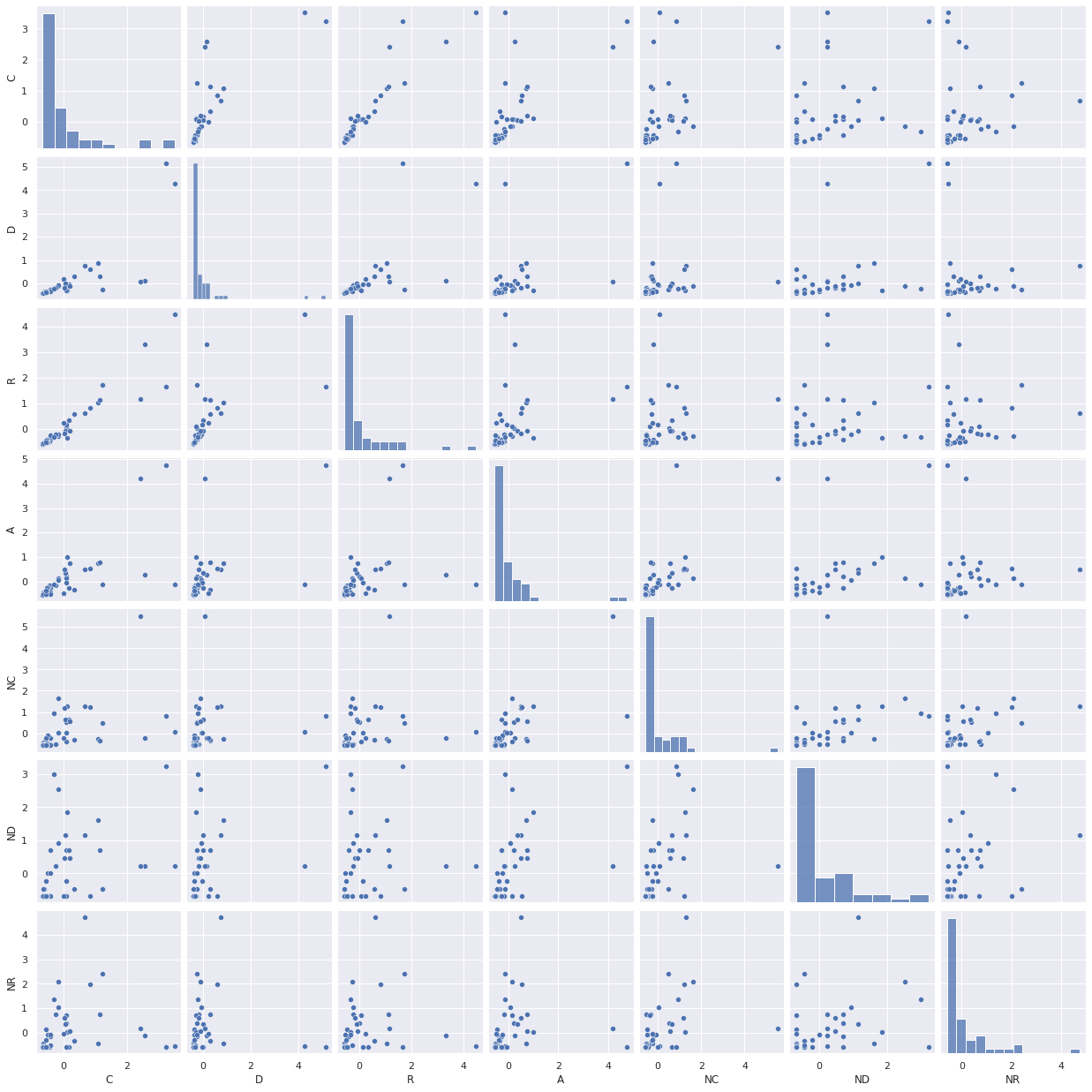


Fig 3. pair plot

By having the pareto diagram Fig, the total of the explained variance can be reached. According to the Fig 4, most of the variation in the data is because of the first three component, it means according to the scree and pareto diagram and .so approximately 93 percent of the variance can be explained only by three first component. In order to obtain the important information that represent a data matrix always the most challenging part is to decide how many components should be considered. A useful method for this purpose is explained variance for each component that can be calculate through this formula:



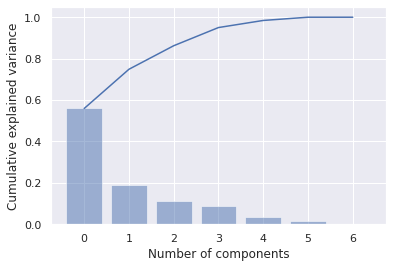
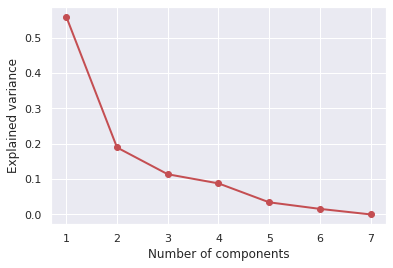


Fig 4 pareto diagram

Fig 5. Explained variance plot

The eigenvectors matrix is calculated according to the Fig 6 as below:

According to the component that is calculated above it is obtain that Confirmed, Deaths and Recovered are respectively, contribute to first component. Plus, the second component is a contrast between Deaths

and New Deaths and on the other hand between and .

Finally, third principal component, there is a contrast of attribute of Deaths and form the other hand New cases .

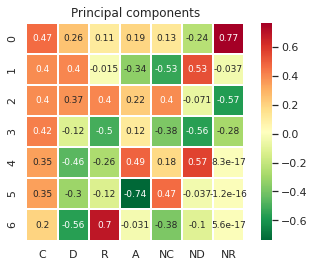


Fig 6. Covariance matrix of zero

Scatter plot of PC2 coefficient vs PC1 coefficients illustrated in the figure below Fig 7.

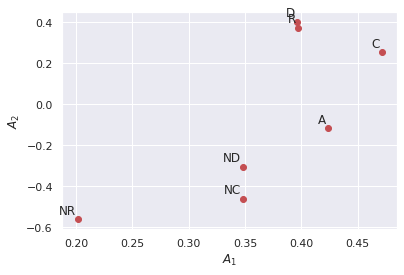


Fig 7. Scatter plot of PC2 coefficients vs. PC1Coefficients.

This scatter plot would illustrate the variables that have the same involvement within PCs. As can be shown in the Fig 7.

Variables NR (New recovered) is located on the bottom left and other variables are located on the other side of the plot. So, this is obtained according to the value of the PCs coefficients. D, R that are located on the top of the plot achieved the same involvement within the principal component and also this is the case with the ND and NC achieved.

Here is the illustration of the scatter plot of the PC2 score compare with PC1 for each observations Fig 8.



Fig 8. Scatter plot of PC2 score vs. PC1

score.

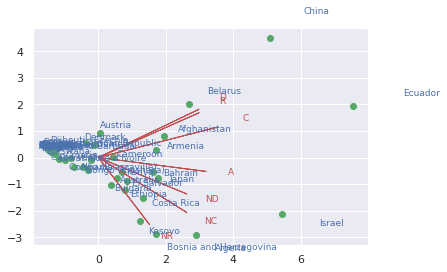


Fig 9 Biplot

PCA involve other very important and useful plot called biplot. It shows principal component and scores for variables and observations, respectively. com Fig 9. We have vectors and point in this plot that they define the countries that are our observations and variables, respectively. The axis of the biplot are variables and principal component are, respectively.

It is shown that the first principal components

Of course, here we have the third principal component here but for the analyzing the principal component in the following we just analyze with two first principal component because analyzing a 3D plot is very sophisticated.

As it is shown in the picture first principal component has 7 positive coefficients that all vectors are directed to the right-hand side of the plot, respectively. The second principal component that is illustrated through vertical axis, 3 positive coefficients with the variables confirmed, recovered and death items and 4 negative coefficients for the variables Active, new cases, new death and new recovered.

As it is shown in the fig 8 the variables confirmed that has a narrow angle such and long direction with the component 2. Also, new recovered has the narrowest angle with the first principal component [3].

# Regression Analysis

Regression analysis is the prediction of the dependent variable according to the value on independent variables [4]. It can let you know which factors are more important or which are not that can be ignored [5]. It also explains the effect of the variance in an independent variable on dependent one [6,7].

Here we have real state prediction Datasets that is given from where x is the age of the persons and Y is the charges for the insurance related to each person. There is column in our dataset that are not considered such as age etc. [8].

In this dataset, house price of unit area depends on the number of convenience stores, so the price for convenience stores is the dependent variable and convenience stores is independent variables within the regression analysis model. Furthermore, at beginning it is expected to have increase in the house price of unit area by the increase in the convenience. So, a positive relationship exists between convenience stores and house price of unit area variables Fig 10.

According linear regression model the least square regression line is given:

The value of the explain the value of the for x=0 and it define the real state prediction price without considering the number of the convenience store. And the would give an insurance price by changing the X and explained that by each 10 increase in number of convenience stores; the price of each unit of area would increase by approximately 3 $.

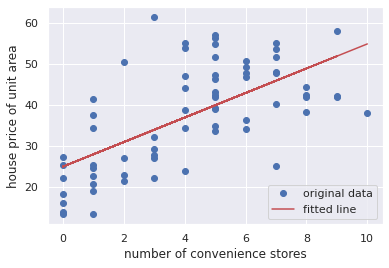


Fig 10 Scatter plot of regression model

The value of explain that the number of convenience stores and the price of each unit area are positively related. Plus, the value of of the total changes in the price of each unit area is explained via number of convenient stores.

The Anova table of the linear regression is illustrated below. It indicates the mean squares, mean squares and degree freedom for error and regression. Also, the value of the Anova F statistic is shown in the table 3.

## Hypothesis test

Testing significant of linear regression for our data is illustrated here in table 3. It is supposed to test the significance of having ANOVA with α = 0.05.

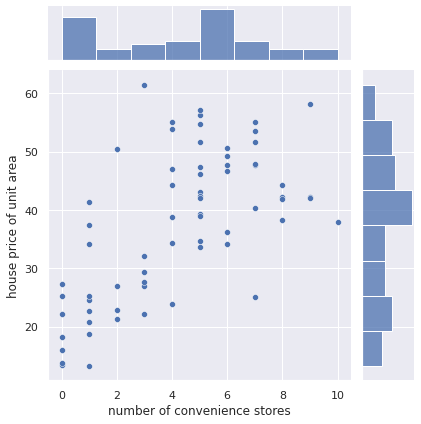
The hypothesis is if there is significant linear relationship or if there is a significant linear relationship [9].

The value of the test statistic is provided for 44.499. Since , the null hypothesis will be rejected. Therefore, we include that the number of convenience stores is linearly related to the price of the unit are.

Also parameter estimation are retrieved in the Fig 11.

TABLE III. Anova table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | df | Sum\_sq | Mean\_sq | F | PR(>F) |
| X | 1 | 71834.0034 | 71834.0034 | 44.499 | 6.594221e-09 |
| Residuals | 65 | 6390.58258 | 98.316655 | NaN | NaN |
| Alpha | 0.05 | 3.99 |  |  |  |



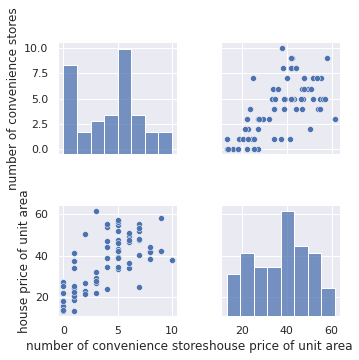


Fig 11. Retrieve the parameter estimates

## Residual analysis

Residual analysis is utilized in order to assess suitability of a linear regression model through checking the residuals that can be analyze by its plot. Residual the difference between independent and dependent value [9]. It is given in the following formula:

The residual plot for the prediction real state price indicate that point is distributed randomly around the zero. So linear regression model is appropriated.

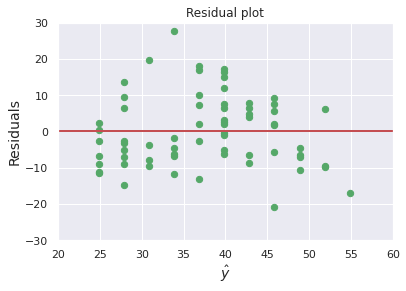


Fig 12 Residual plot

# Conclusion

In this paper we applied principal component analysis for the COVID-19 Dataset and the effect of the principal components on the data determined by analyzing the results that is obtained through python. Firstly, PCA is applied on the dataset that is provided by Kaggle and it is realized that approximately 93 percent of the variance can be explained only by three first components. But we choose to work on two components because of having problem in plotting 3D graph and analyzing. Secondly, we applied regression analysis in order to predict the dependent variable according to the value on independent variables. According to the result it is shown that, the value of of the total changes in the price of each unit area is explained via number of convenient stores. Furthermore, a hypothesis test is applied to very that the dependent (number of convenience stores) and independent (house price of unit area) are related. Finally, Residual analysis is utilized in order to assess suitability of a linear regression model through checking the residuals that can be analyzed by its plot and concluded that the residual plot for the prediction of house price of unit area indicate that point is distributed randomly around the zero, so linear regression model is appropriated

##### References

1. <https://www.kaggle.com/imdevskp/corona-virus-report>
2. <https://en.wikipedia.org/wiki/Box_plot#:~:text=In%20descriptive%20statistics%2C%20a%20box,whisker%20plot%20and%20box%2Dand%2D>
3. https://www.statisticshowto.com/correlation-matrix/.
4. <https://moodle.concordia.ca/moodle/pluginfile.php/4552067/mod_resource/content/1/W8INSE6220.pdf>.
5. <https://dss.princeton.edu/online_help/analysis/regression_intro.htm#:~:text=Regression%20analysis%20is%20used%20when,a%20number%20of%20independent%20variables.&text=Independent%20variables%20with%20more%20than,that%20have%20only%20two%20levels>.
6. <https://stagewww.surveygizmo.com/resources/blog/regression-analysis/>
7. <https://moodle.concordia.ca/moodle/pluginfile.php/4558964/mod_resource/content/1/W9INSE6220.pdf>
8. <https://www.kaggle.com/quantbruce/real-estate-price-prediction>
9. https://moodle.concordia.ca/moodle/pluginfile.php/4552068/mod\_resource/content/1/A2INSE6220sol.pdf
10. <https://www.tutorialspoint.com/statistics/residual_analysis.htm#:~:text=Residual%20analysis%20is%20used%20to,examining%20the%20residual%20plot%20graphs>.

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**