# **The impact of COVID-19 on different categories of age, race, and gender in the US**

***Komali Reddy Konda Ram Kiran Godavarthi Apurva Ingole***

(731) 812-1312 (517) 213-5605 (551) 260-1254

kreddy@saintpeters.edu rgodavarthi@saintpeters.edu aingole@saintpeters.edu

***Dr. Gulhan Bizel Dr. Joseph W. Gilkey Jr.***

gbizel@saintpeters.edu jgilkey@saintpeters.edu

Data Science Institute

Saint Peter’s University

2641 John F. Kennedy Blvd

Jersey City, NJ USA, 07306

**ABSTRACT**

The recent pandemic Coronavirus (COVID-19) disease does not influence everybody similarly. Like the many other countries, it is exposing imbalances in the health system also in the US. Since COVID-19 discovery, the virus has spread globally, causing thousands of deaths, and having an enormous impact on our health systems. The relationship between ethnicity, age, gender, and COVID-19 is questionable. In this research, it has performed distinctive correlation models and machine learning algorithms to evaluate the strength of the relationship of these categories to the COVID-19 cases across all the states of the USA. Comparing the R values and accuracy of the models we have inferred some vital inputs like which is the most vulnerable category that is getting affected. Experiment results lead to interesting facts about the behavior of a widespread coronavirus case across the nation in different age groups, ethnicity, and gender.

**Keywords:** Coronavirus, Vulnerable, Multiple Linear Regression, OLS, Spearman’s correlation, Association, Machine Learning.

**1 INTRODUCTION**

Different Coronaviruses cause illness varying from the basic virus-like common cold & flu to progressively severe infections, like Serious Intense Respiratory Disorder (SARS) and the Middle East Respiratory Syndrome (MERS). COVID-19 being the seventh one to affect humans is caused by the SARS-COV-2 virus and is said to have a zoonotic origin (Zhang, 2020). It is known that the reported number of cases in the USA has reached an alarming number of 4 million affected cases and segments of people who are more vulnerable to it. This research tries to shed light on these segments by seeing their correlation with the number of cases in the USA and by finding the highest and lowest factors and comparing them with the help of multiple regression models.

In here, the datasets are divided into three independent factors: age, race, and gender for which our analysis would be focused on 19 states and those spread across the country through which we are delineating the overview of the stats in the USA, where, the analysis for race factor is also performed across all the states of the country. For the factor age, a detailed approach is followed by a gap of 10 years between their groups. The stats of the data in the race factor is segregated as White, Hispanic, Black, American Indian and Alaska Native, Asian, and Native Hawaiian, while “gender” is divided by male and female. The stats used here is the data that was collected from the “Census Bureau” and through Johns Hopkins University on GitHub which were merged to perform correlation and modeling.

Multiple regression models have been designed to predict MAE, MSE, and RMSE for each section of the three independent factors. Mean absolute error is a measure of the errors between paired observations expressing the same phenomenon. Mean squared error measures the average of the squares of the errors that is the average squared difference between the estimated values and the actual values and root mean square error is frequently used to measure the differences between predicted values and observed values by a model. Also, a correlation was found which was the coefficient between the section of factors and the number of cases by Spearman Correlation and resulted in r squared and adjusted r square values from the Ordinary Least Square method.

To begin with, to build and develop Machine Learning models, relevant datasets are to be acquired from known sources. This dataset will be composed of data gathered from multiple and disparate sources which are then combined in a proper format to form a dataset. Dataset formats differ according to use. For instance, a business dataset will not be the same as a clinical dataset. While a business dataset contains significant industry and business information, a clinical dataset will incorporate human services related information. It is also possible to create a dataset by collecting data via different sources viz. Python APIs. Once the dataset is ready, it must be placed in a CSV/HTML/XLSX file format.

**2 MATERIALS AND METHODS**

The main goal of this research is to identify which category of the demographics (age, ethnicity, and gender) are getting affected and are vulnerable to this communicable disease. This requires a correlation model that identifies the strong correlation measure between the demographic categories and cases. We have used Spearman’s correlation model to obtain correlation measures for the data which is acquired from the verified sources to get the best and genuine outcomes from our analysis. The data related to demographics (age, ethnicity, and gender) have been taken from the census bureau.

The data related to the COVID-19 cases for each state have originated from a variety of open sources and was ordered in the first instance through Johns Hopkins University on GitHub. Data is in .CSV format and refreshed day by day which is sourced from this upstream repository maintained by a group at Johns Hopkins University Center for Systems Science and Engineering (CSSE) who have been doing extraordinary work from an early point by collecting data from around the globe.

This dataset includes the number of people infected by COVID-19 across the USA, including people among different races and ethnicity, different age groups, and gender. For the factors age, race, and gender we have a dataset for 19 states, also for the factor race, we have the dataset for 50 states in the USA. Since the data has been gathered from different sources, where it is pre-processed to analyze and retrieve accurate results. As we are using a machine learning model, splitting the dataset into two separate sets – training set and test set is mandated, and to normalize the independent factors of the dataset inside a range, feature scaling is done.

**Data Preprocessing**

Data preprocessing in Machine Learning is an initial step that helps improve the nature of data to advance the extraction of meaningful insights from the data. Data preprocessing in Machine Learning allude to the procedure of getting ready (cleaning and sorting out) the raw data to make it suitable for building and preparing Machine Learning models. In simple words, data preprocessing in Machine Learning is a data mining method that changes raw data into an understandable and readable format.

Regarding making a Machine Learning model, data preprocessing is the initial step denoting the inception of the procedure. Normally, real-world data is inadequate, conflicting, mistaken (contains errors or outliers) and regularly needs explicit trait values/patterns. This is the place data preprocessing enters the situation – it assists with cleaning, design, and arrange the raw data, consequently preparing it to work for Machine Learning models.

The significant steps in data preprocessing in Machine Learning will be explained further in the next upcoming steps.

***Identifying And Handling The Missing Values***

In data preprocessing, it is significant to recognize and effectively handle the missing values, neglecting to do this, it may reach incorrect and broken determinations and surmise from the data. This will hamper the ML research. Essentially, there are two different ways to deal with missing information.

***Dropping A Row Or Column***

In this method, it is possible to remove a specific row/column that has a null value for a feature or a row/column where more than 75% of the values are missing. However, this method is not 100% efficient, and it is recommended to use it only when the dataset has adequate samples. It must be ensured that after deleting the data, there remains no addition of bias.

***Calculating The Mean***

For features having numeric data like age, salary, year this method is more useful. Here, we can calculate the mean, median, or mode of a feature or column or row that contains a missing value and replace the result for the missing value. This technique can add fluctuation to the dataset, and any loss of data can be effectively nullified. Consequently, it yields better outcomes contrasted with the primary strategy (exclusion of rows/columns). Another method of a guess is through the deviation of neighboring values.

**Splitting The Dataset**

Every dataset of Machine Learning models must be split into two separate sets – training set and test set. The training set denotes the subset of a dataset that is used for training the machine learning model where we are already aware of the output. A test set, on the other hand, is that the subset of the dataset that is used for testing the machine learning model. Machine Learning models use test data to predict outcomes. Usually, the dataset is split into a 70:30 ratio or 80:20 ratio. This suggests that it is possible to take either 70% or 80% of the data for training the model while leaving out the rest 30% or 20% to test the model. The splitting process varies consistent with the form and size of the dataset that is considered (Mirjalili, 2019).

train\_test\_split function includes four parameters, the first two are for arrays of data. The test\_size function specifies the size of the dimensions of the test set. The test\_size may be 0.5, 0.3, or 0.2 – this specifies the dividing ratio between the training and test sets. The last parameter, “random\_state” sets the values for a random generator in order the output is usually the same.

**Feature Scaling**

Feature scaling marks the finish of the data preprocessing in Machine Learning. It is a strategy to normalize the independent factors of a dataset inside a range. At the end of the day, including scaling limits the scope of factors with the goal that you can think about them on regular grounds. The following legitimate advance in our preprocessing pipeline is proportional to our features. Before applying any scaling changes, it is imperative to part our data into a train set and a test set. If we start scaling before, the training (and test) data might end up scaled around a mean value that is not actually the mean of the train or test data and then again go past the entire motive behind why you're scaling in any case (Mirjalili, 2019).

In the dataset, it may be observed that all columns do not have the same scale. In such a scenario, if it is computed by any two values from the columns, one value will dominate the other and deliver incorrect results. Thus, it must be removed from this issue by performing a feature scaling for Machine Learning.

***Standardization***

Standardization is a change that focuses the data by expelling the mean estimation of each element and afterward scale it by separating (non-consistent) features by their standard deviation. In the wake of standardization data, the mean will be zero and the standard deviation will be one. Standardization can improve the exhibition of models (Mirjalili, 2019). For example, numerous components utilized in the objective function of a learning algorithm (for example, the RBF piece of Support Vector Machines or the l1 and l2 regularized of direct models) expect that all features are based on zero and have a difference in a similar order. On the off chance that an element has a change that is significant degrees bigger than others, it may rule the objective function and make the estimator unfit to gain from other features correctly as expected. Depending on the needs and data, sklearn provides a bunch of scalers: StandardScaler, MinMaxScaler, MaxAbsScaler, and RobustScaler. We have chosen the Standard Scaler for our data.

***Standard Scaler***

Sklearn, its principle scaler, the StandardScaler, utilizes an exact meaning of standardization to standardize data. It focuses the data by utilizing the following formula, where u is the mean and s is to be a standard deviation.

x\_scaled = (x — u) / s (Mirjalili, 2019).

For the dataset, the standardization strategy is being utilized. To do so, StandardScaler class of the sci-kit-learn library is imported by using the following line of code: from sklearn. preprocessing import StandardScaler

The next step is to create the object of StandardScaler class for independent variables. After, it fits and transforms the training dataset using the following code:

st\_x= StandardScaler()

x\_train= st\_x.fit\_transform(x\_train) (Mirjalili, 2019).

For the test dataset, it can be directly applied to the transform () function (no need to use the fit\_transform () function because it is already done in the training set). The code will be as follows:

x\_test= st\_x. transform(x\_test)(Mirjalili, 2019).

The scaled output values for x\_train and x\_test for the test dataset will show as:

All the variables in the output are scaled between the values -1 and 1.

**3 STATISTICAL MODELS**

A statistical model is generally determined as a numerical connection between at least one or more random factors. Insights include data collection, understanding, and approval. Statistical analysis is the procedure of applying several statistical operations to evaluate the data and apply Statistical analysis. Quantitative data include distinct information like studies and observational data. It is likewise called a descriptive analysis. It incorporates different tools to perform statistical data analysis, for example, SAS (Statistical Analysis System), SPSS (Statistical Package for the Social Sciences), Stat soft, and more.

Correlation is one of the most broadly utilized statistical concepts. Correlation is a statistical measure that depicts the relationship between irregular factors. Linear regression models express the relationship between two variables by fitting a line to the observed data and allows to predict a change in a dependent variable as the independent variables change. Ordinary Least Squares is a type of linear least squares method used to find unknown parameters in a linear regression model. All these three models will be explained in detail.

**Correlation**

Variables within a dataset are often associated with many reasons. We are assuming that every category from age, race, and gender has a strong correlation to the number of reporting COVID-19 cases. It could be helpful while doing data analysis and modeling to understand the relationships between variables. The statistical relationship between the two variables is brought up as their correlation. It's a measure of association between two variables. A positive correlation could be, meaning both variable’s move within the same direction, or negative, meaning that when one variable’s value increases, the other variables’ values decrease. Correlation may be neutral or zero, meaning that there is no relation between the variables (Brownlee, 2020).

* **Positive Correlation:** both variables change in a similar direction.
* **Neutral Correlation:** No relationship within the change of the variables.
* **Negative Correlation:** variables change in opposite directions.

There are different approaches in statistics to search out the association between the variables. The most popular correlation methods are Pearson's correlation and Spearman's correlation which will be explained further below.

***Pearson's Correlation***

In 1904 Spearman adopted Person’s correlation coefficient as a measure of the strength of the relationship between two variables that can’t be measured quantitatively (Hauke, 2011). Pearson's relationship coefficient is determined as the covariance of the two factors separated by the result of the standard deviation of each data sample. It is the normalization of the covariance between the two variables to give an interpretable score.

Pearson's correlation coefficient = covariance (X, Y) /(stdy(X) \* stdv(Y)) (Brownlee, 2018).

● The use of mean and standard deviation in the calculation suggests the need for the two data samples to have a Gaussian or Gaussian-like distribution (Brownlee, 18).

● The result of the calculation, the correlation coefficient can be interpreted to understand the relationship.

The coefficient restores a value between -1 and 1 that speaks to the furthest reaches of correlation from a full negative correlation to a full positive correlation. An estimation of zero means no correlation. The value must be interpreted, where frequently a number beneath -0.5 or above 0.5 shows a prominent connection, and values underneath those numbers recommend a less outstanding correlation (Schober, 2018). A t-test is accessible to test the null hypothesis that the correlation coefficient is zero. Note that the P-value obtained from the test gives no data on how strongly the 2 factors are connected. With huge datasets, little relationship coefficients can be "statistically significant." Therefore, a statistically huge correlation must not be mistaken for a clinically applicable relationship.

***Spearman’s Correlation***

Spearman's rank correlation is a nonparametric (distribution-free) rank measurement proposed as a proportion of the strength of the relationship between two variables. It is a proportion of a monotone affiliation that is utilized when the dispersion of information makes Pearson's correlation coefficient unwanted or misleading (Hauke, 2011). As with the Pearson correlation coefficient, the scores are between -1 and 1 for consummately adversely associated variables and impeccably emphatically corresponded separately. Rather than computing the coefficient utilizing covariance and standard deviations on the samples themselves, these measurements are determined from the overall rank of values on each sample. This is a typical methodology utilized in non-parametric statistics, e.g. statistical methods where we do not assume a distribution of the data such as Gaussian (Brownlee, 2020).

Spearman's correlation coefficient = covariance (rank(X), rank(Y)) / (stdv(rank(X)) (Brownlee, 2018).

In uncertain situations of the distribution and possible relationships between two variables, the Spearman correlation coefficient is a good tool to use. Spearman’s correlation has given the best results when performed for our analysis. It gave the features that are strongly related to the COVID-19 cases along with the p-value. The p (or likelihood) value obtained from the correlation is a proportion of how likely or plausible it is that any observed connection is because of possibility. P-values run between 0 (0%) and 1 (100%). A p-value near 1 proposes no relationship other than because of possibility and that your null hypothesis assumption is correct. If your p-value is near 0, the observed association is probably not going to be because of possibility and there is an extremely high likelihood that your null hypothesis isn't right. For this situation, you should acknowledge the Alternative (H1) hypothesis that there is an association between the data sets.

The data we have used for correlation analysis were not normalized, hence we have used Spearman’s correlation which gave the best results when compared to Pearson’s correlation method.

**Linear Regression Model**

Linear regression is perhaps one of the standards and well-known algorithms in statistics and machine learning. Machine learning, even more explicitly the field of prescient demonstrating, is essentially worried about limiting the mistake of a model or making the most precise expectations conceivable, to the detriment of reasonableness. In applied machine learning, we will obtain, reuse, and make calculations from various fields, including insights and use them towards these ends (Brownlee, 2016).

In that capacity, linear regression was created in the field of statistics and is read as a model for understanding the relationship among input and output numerical variables, however, has been borrowed by machine learning. It is both a statistical algorithm and a machine learning algorithm (Brownlee, 2016). By fitting a linear equation to observed data linear regression aims to model the association between two variables. We consider one variable as explanatory and the other as a dependent variable. An important numerical proportion of the relationship between two factors is the correlation coefficient, which is an incentive between -1 and 1 showing the strength of the relationship of the observed data for the two variables.

**Ordinary Least Squares Method (OLS)**

OLS is a computationally modest, simple to-clarify quotient of determination value that depends on simple statistics. It is the most used method in Statistical Learning which is also known as the curve fitting method. It is additionally the most seasoned, going back to the eighteenth century and crafted by Carl Friedrich Gauss and Adrien-Marie Legendre. It is additionally one of the simpler and progressively natural methods to comprehend, and it gives a decent premise to learning further developed ideas and strategies. Following that, OLS is an estimator where the estimations of m and c (from the above condition) are picked to limit the whole of the squares of the contrasts between the watched subordinate variable and the anticipated ward variable. That is the reason it's called ordinary least squares. Likewise, it ought to be noticed that when the whole of the squares of the distinctions is least, the misfortune is additionally least—thus, the forecast is better. OLS Regression gives us the outcomes as far as R-value and other statistical terms. Linearity can also be assessed by the least square method (Mukhopadhyay, 2018).

OLS methods give these Statistical outputs in terms of a measure of how the data is fit to the linear regression line using R squared, Adjusted R squared, also few standard error values are interpreted using t-static, F-statistic, and Akaike Information Criterion, Bayesian Information Criterion gives the measure of the quality of the statistical model for the data used.

* **R-squared** is also called the coefficient of determination. It's a statistical proportion of how well the regression line fits the data.
* **Adjusted R**-**squared** modifies the measurements dependent on the independent variables present.
* The ratio of deviation of the assessed estimation of a boundary from its hypothesized to its standard error is called **t**-**statistic.**
* **F-statistic** is determined as the proportion of mean squared error of the model and mean squared error of residuals.
* AIC stands for the **Akaike Information Criterion**,which estimates the relative quality of statistical models for a given dataset.
* BIC represents the Bayesian Information Criterion, which is utilized as a standard for model choice among a limited arrangement of models. BIC resembles AIC; be that as it may, it includes a higher penalty for models with more parameters.

The approach has been made for the OLS model in this research to have a better idea of how well the data is the best fit for the linear regression line.

The data from the GitHub and Census bureau had mixed types of data that are categorical and numerical. We have focused on recognizing the correlation between them using Spearman’s correlation and OLS model for best linear fit identification.

**4 TOOLS**

Once the data are gathered, preprocessed, to interpret the results for the analysis python programming language is used. Python is the most widely utilized and the most favored library by Data Scientists around the globe. There are different python packages which we have used for the analysis. The predefined Python libraries can perform explicit data preprocessing occupations. The three core Python libraries used for this data preprocessing in Machine Learning are NumPy, Pandas, Matplotlib, and SciPy as explained below:

* **NumPy** – NumPy is the basic bundle for the logical count in Python. Henceforth, it is utilized for embedding any kind of mathematical operation in the code. Utilizing NumPy, you can likewise include enormous multi-dimensional exhibits and matrices in your code.
* **Pandas** – Pandas is a superb open-source Python library for data manipulation and analysis. It is broadly utilized for bringing in and overseeing datasets. It packs in elite, simple to-utilize data structures and data analysis tools for Python.
* **Matplotlib** – Matplotlib is a Python 2D plotting library that is utilized to plot any sort of charts in Python. It can convey distribution quality figures in various printed copy designs and intelligent situations across platforms (IPython shells, Jupyter notebook, web application servers, and so on.).
* **SciPy** -SciPy is a free and open-source Python library utilized for scientific computing and specialized processing. We have used SciPy to the library for performing the statistical models, to find the correlations.

We have used NumPy to do various mathematical operations and statistical models on the dataset and also to find the data types of the values in the dataset, which helped us to optimize the code. Pandas is used to read the dataset and perform some manipulations and analysis on the data, it is built on the NumPy. Matplotlib is a mathematical extension to NumPy, it is used to plot charts in a linear regression model to have an idea on the behavior of values in the data. Scipy is used to import spearman and Pearson correlations, the SciPy is built on the NumPy, it also makes use of Matplotlib.

**5 RESULTS**

To extract the most affected groups amongst Age, Race, and Gender in relation with COVID-19 cases, we have performed Spearman’s correlation approach, followed by the Linear Regression model and OLS method.

**Effect Of COVID-19 On Different Categories Of Age, Race, And Gender Using Spearman’s Correlation**

After finding the correlation between the cases and different Age groups, races, and gender by using Spearman’s Correlation the following results are obtained.

From the Table 5(a) for 19 states after finding the correlation between cases and different age groups using Spearman’s correlation, it has been observed that people of age group between 40-49 years, 50-59 years, and 60-69 are highly affected with the correlation values R (correlation coefficient) of 0.937, 0.935, and 0.933 respectively among different age groups.

|  |  |  |
| --- | --- | --- |
| **Age** | **Correlation (r)** | **P-values** |
| 0-19 years | 0.665 | 0.0018 |
| 20-29 years | 0.851 | 3.88e-06 |
| 30-39 years | 0.874 | 1.029e-06 |
| 40-49 years | 0.937 | 3.543e-09 |
| 50-59 years | 0.935 | 4.446e-09 |
| 60-69 years | 0.933 | 5.543e-09 |
| 70-79 years | 0.840 | 6.688e-06 |
| 80+ years | 0.698 | 0.0008 |

**Table 5(a)** Correlation for different age groups for 19 states.

These correlation results for different age groups describe that they show a strong positive correlation with cases by seeing that their r values are greater than 0.8, except for age groups 0-19 and 80+ years as they have below 0.8 while they still show a positive correlation. Also, the p-values are close to zero for all the age groups, with this evidence we accept alternative (H1) hypothesis that there is an association between the age groups and COVID-19 cases.

As shown in the Table 5(b) for 19 states amongst different races, Whites show strong relationship with the number of cases when compared to other groups, with a correlation value of 0.798 and also observed that their p-value is close to zero, so we accept alternative (H1) hypothesis that there is an association between the white race and COVID-19 cases. An inverse correlation among the American Indian and Alaska Native race having a negative value of -0.180.

|  |  |  |
| --- | --- | --- |
| **Race** | **Correlation (r)** | **P-value** |
| White | 0.798 | 4.199e-05 |
| Black or an African American | 0.561 | 0.047 |
| Hispanic | 0.460 | 0.125 |
| American Indian and Alaska Native | -0.180 | 0.4606 |
| Asia | 0.088 | 0.7210 |

**Table 5(b)** Correlation for Race for 19 states.

Except for the American Indian and Alaska Native and Asian races, the correlation results for all the races shows a positive correlation whereas, American Indian and Alaska Native shows negative and Asian describe almost neutral correlation.

From Table 5(c) it is also seen that from the correlation of different races with a number of cases across all 50 states, Whites have the highest association with cases than any other groups, with a correlation value of 0.730 and p-value which is close to zero, where we accept alternative (H1) hypothesis that there is an association between the white race and COVID-19 cases. Also, a Negative correlation is observed in the American Indian and Alaska Native and Native Hawaiian and Other Pacific Islander with the r values of -0.008 and -0.113, respectively.

|  |  |  |
| --- | --- | --- |
| **Race** | **Correlation (r)** | **P-value** |
| White | 0.730 | 5.949e-09 |
| Black or an African American | 0.454 | 0.00136 |
| Hispanic | 0.655 | 5.777e-07 |
| American Indian and Alaska Native | -0.008 | 0.9580 |
| Asian | 0.216 | 0.144 |
| Native Hawaiian and Other Pacific Islander | -0.113 | 0.448 |

**Table 5(c)** Correlation for Race for 50 states.

These results indicate that except for inversely correlated races, the rest of them show a positive correlation. Whereas Asians are weakly correlated as they have a lower r value with 0.216.

From Table 5(d), In the factor gender for 19 states, it is seen that the correlation values are almost similar among both genders, while males have slightly higher correlation values with 0.858 when compared with females who have 0.851.

|  |  |  |
| --- | --- | --- |
| **Gender** | **Correlation (r)** | **P-value** |
| Male | 0.858 | 2.647e-06 |
| Female | 0.851 | 3.899e-06 |

**Table 5(d)** Correlation for Gender for 19 states.

It is also observed that their p-values are close to zero, where we accept the alternative (H1) hypothesis that there is an association between the male, female with COVID-19 cases separately.

**Linear Regression Model Using Machine Learning**

After evaluating the train test data in the linear regression model, we have obtained the MAE (Mean Absolute Error), MSE (Mean Square Error), and RMSE (Root Mean Square Error) values for each model.

From Table 5(e), for the 19 states among different age groups, it is observed that RMSE values are as low as for the age groups 50-59 years and 40-49 years values with 0.214 and 0.219 respectively and as high as 1.002 for 0-19 years of age group.

|  |  |  |  |
| --- | --- | --- | --- |
| **Age** | **MAE** | **MSE** | **RMSE** |
| 0-19 years | 0.872 | 1.004 | 1.002 |
| 20-29 years | 0.466 | 0.411 | 0.641 |
| 30-39 years | 0.489 | 0.515 | 0.718 |
| 40-49 years | 0.147 | 0.048 | 0.219 |
| 50-59 years | 0.185 | 0.046 | 0.214 |
| 60-69 years | 0.326 | 0.197 | 0.443 |
| 70-79 years | 0.477 | 0.279 | 0.529 |

**Table** **5(e)** Error-values of age groups for 19 states.

The RMSE values greater than 0.2 are said to be that the model can relatively predict the data accurately, whereas we have the RMSE values greater than 0.2 for all the age groups which shows that each model for the individual age group has predicted the data accurately.

From Table 5(f) and Table5(g), It is seen in the race among the 19 states and 50 states data Asian & American Indian and Alaska Native are having low RMSE values with 0.514 and 0.819 individually among their categories.

|  |  |  |  |
| --- | --- | --- | --- |
| **Race** | **MAE** | **MSE** | **RMSE** |
| White | 0.643 | 0.460 | 0.678 |
| Black or African American | 0.560 | 0.439 | 0.662 |
| Hispanic | 0.610 | 0.433 | 0.658 |
| American Indian and Alaska Native | 0.556 | 0.461 | 0.679 |
| Asian | 0.451 | 0.264 | 0.514 |

**Table** **5(f)** Error-values for Race for 19 states.

|  |  |  |  |
| --- | --- | --- | --- |
| **Race** | **MAE** | **MSE** | **RMSE** |
| White | 0.516 | 0.744 | 0.862 |
| Black or African American | 0.574 | 0.932 | 0.967 |
| Hispanic | 0.595 | 0.999 | 1.000 |
| American Indian and Alaska Native | 0.735 | 0.671 | 0.819 |
| Asian | 0.595 | 0.929 | 0.964 |
| Native Hawaiian and Other Pacific Islander | 0.716 | 0.968 | 0.984 |

**Table** **5(g)** Error-values for Race for 50 states.

These RMSE results for races of both 19 states and 50 states show that the respective models for each race have predicted the data accurately.

From Table 5(h), for the factor gender in 19 states, we can say that the RMSE value for males is less in the gender with a value of 0.341 when compared to women who are having a value of 0.348.

|  |  |  |  |
| --- | --- | --- | --- |
| **Gender** | **MAE** | **MSE** | **RMSE** |
| Male | 0.207 | 0.116 | 0.341 |
| Female | 0.174 | 0.121 | 0.348 |

**Table** **5(h)** Error-values for Gender for 19 states.

Which still shows us that the models for both male and female predicted the data accurately.

**Outcomes Of Statistical Model For Different Categories Of Age, Race, And Gender Using Ordinary Least Squared Method**

From the Ordinary Least Square (OLS) Method R-squared and adjusted R-squared values were obtained which are unknown parameters in a linear regression model. From Table 5(i), for 19 states among different age groups, it is observed that R-squared values were high for age groups 60-69 years and for the people who are 50-59 years in age with 0.869 and 0.850, respectively.

|  |  |  |
| --- | --- | --- |
| **Age** | **R-squared** | **Adjusted R-squared** |
| 0-19 years | 0.618 | 0.596 |
| 20-29 years | 0.681 | 0.664 |
| 30-39 years | 0.774 | 0.761 |
| 40-49 years | 0.792 | 0.781 |
| 50-59 years | 0.850 | 0.842 |
| 60-69 years | 0.869 | 0.862 |
| 70-79 years | 0.802 | 0.791 |
| 80+ years | 0.633 | 0.612 |

**Table** **5(i)** R squared, and Adjusted R squared values for age across 19 states.

These R-squared values for different ages indicate that they all are better fit to their models and the age groups having R-squared values greater than 0.8 have a smaller difference between the observed data and the fitted values.

Table 5(j) and Table 5(k) show that in the race among the 19 states and 50 states data, White has the highest R-squared value of 0.937 and 0.882, respectively. Which tells us that they have a smaller difference between the observed data and the fitted values.

|  |  |  |
| --- | --- | --- |
| **Race** | **R-Squared** | **Adjusted R-Squared** |
| White | 0.937 | 0.934 |
| Black or African American | 0.842 | 0.833 |
| Hispanic | 0.706 | 0.690 |
| American Indian and Alaska Native | 0.257 | 0.216 |
| Asian | 0.531 | 0.505 |

**Table** **5(j)** R squared, and Adjusted R-squared values for Race 19 states.

|  |  |  |
| --- | --- | --- |
| **Race** | **R-Squared** | **Adjusted R-Squared** |
| White | 0.882 | 0.880 |
| Black or African American | 0.581 | 0.572 |
| Hispanic | 0.770 | 0.765 |
| American Indian and Alaska Native | 0.219 | 0.202 |
| Asian | 0.334 | 0.320 |
| Native Hawaiian and Other Pacific Islander | 0.138 | 0.120 |

**Table** **5(k)** R squared, and Adjusted R squared values for Race for 50 states.

All the races show a better fit for their respective models but the American Indian and Alaska Native and Native Hawaiian and Other Pacific Islander races have noisy, high-variability data that can have a significant trend having low R-squared values.

From the above results, it can be inferred that the Whites among different races are vulnerable to get affected with COVID-19, age groups of 40-49 years, 50-59 years, and 60-69 years, and male in gender are showing a strong relationship with the COVID-19 cases while using Spearman’s Correlation. Since the RMSE values are good for all models, it can be described that the linear regression model is performed better for every model and relatively predicts the data accurately. The R-Squared values which are generated by the OLS method explain that the Whites among different races, age groups of 50-59 years, and 60-69 years are having a smaller difference between the observed data and the fitted values and hence show a strong association with the COVID-19 cases.

**6 DISCUSSION**

This research succeeded in implementing the Correlation, linear regression models, and recognized how people of different ethnicity/races, ages, and gender are correlated with the COVID-19 cases from the contemporary data which we have.

Our initial analysis from different research papers has shown that White among the different races, people aged between 18-29 years and 50–64 years across all the age groups are more vulnerable to get infected with COVID-19 (CDC COVID Data Tracker). In the factor, gender males are more likely affected by COVID-19 (Klein, 2020). Whereas, our findings from the correlation and linear regression with the cases and different races have shown that the whites are strongly correlated when compared to any other races with the correlation coefficient of 0.798 which is significant with our initial analysis. For finding the association within the cases and different age groups, we have considered the age groups in 10 years of difference in age. While considering this approach it is observed that the age groups of 40-49 years, 50-59 years, and 60-69 years are having a strong correlation with correlation coefficients of 0.937, 0.935, and 0.933 which show a stronger correlation with cases among the other different age groups, which is also supported by our initial analysis. Finally, considering the correlation between gender and cases, we can understand that they are equally correlated with the cases, but the male population is more associated with the cases than the female population by having correlation coefficients of 0.858 and 0.851 respectively, which is again a significant result with our initial analysis.

A linear regression model is performed using machine learning algorithms to obtain the MAE (Mean Absolute Error), MSE (Mean squared Error), RMSE (Root Mean Squared Error) values. The RMSE value indicates how close the observed data values are to the model’s predicted values, RMSE values greater than 0.2 are preferable. From models of the race category, we have seen the highest RMSE values in Hispanic and Black or African American with 1.00 and 0.967 individually, while the remaining races have values greater than 0.8. Considering the different age categories, the RMSE value is high for the 0-19 years age group with 1.002 while for the other groups, it is greater than 0.2 which means that the predicted values are accurate to the observed values. Finally, for gender, the RMSE value for males is 0.341, and females 0.348 which indicates the predicted values as accurate to the observed values.

The OLS (Ordinary least square method) is used to find the unknown parameters in a linear regression model. Where we consider R-squared which is also called a coefficient of determination it is a statistical measure of how well the regression line fits the data and adjusted R-squared adjusts the statistics based on the number of independent variables present. An R-squared value of 1.0 indicates that the data perfectly fit the linear model. Any R-squared value less than 1.0 indicates that at least some variability in the data cannot be accounted for by the model (e.g., an R-squared of 0.5 indicates that 50% of the variability in the outcome data cannot be explained by the model.) (Hamilton, 2015). It is seen that every model in the Age and Gender factors has the R-squared values above 0.6 which indicates that the data are a better fit for their respective models. Whereas for particular age groups of 50-59 years and 60-69 years, high R-squared values are observed, which indicates they are strongly associated with cases and also it is supported by our initial analysis. For the Race category, the R-squared value is as low as 0.257 for the American Indian and the Alaska Native among the 19 states and the races across all states in the US have the R- squared value for the Native Hawaiian and Other Pacific Islander is 0.138 and 0.219 for the American Indian and Alaska Native and below 0.5 for the Asian which shows that even noisy, high-variability data can have significant trends. Despite having low R-squared values, the trend indicates that the predictor variable still provides information about the response even though data points fall further from the regression line. Whereas the Whites are showing a high R- squared value, which means they are highly affected with cases than any other age races and supported by our initial analysis.

In conclusion to this research, it can be inferred that there are divisions of independent variables from age, race, and gender that have a higher chance of contracting with COVID-19. Overlap of these high correlation divisions like age (40-49 years), race (White), and gender (Male) from its variables which can make them more susceptible to COVID-19. Such information can be used to take precautions for these vulnerable groups, which may result in fewer affected cases.

**FUTURE WORK**

To make our findings more accurate we would like to include more states to our analysis whenever the data will be available and also more factors like underlying conditions, obesity, population density, travel history, lockdown restrictions, and other relevant factors so we can predict the correlation between the population and number of cases more accurately.

**REFERENCES**

**Articles**

Zhang, Tao, et al. “Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak.” Current Biology, Cell Press, 19 Mar. 2020.

Schober, Patrick, et al. “Correlation Coefficients: Appropriate Use and Interpretation.” Anesthesia and Analgesia, U.S. National Library of Medicine, May 2018.

Klein, S. L., Dhakal, S., Ursin, R. L., Deshpande, S., Sandberg, K., & Mauvais-Jarvis, F. (2020). Biological sex impacts COVID-19 outcomes. PLoS pathogens, 16(6), e1008570.

Hamilton, D F, et al. “Interpreting Regression Models in Clinical Outcome Studies.” Bone & Joint Research, British Editorial Society of Bone and Joint Surgery, Sept. 2015.

**Books**

Mirjalili, V., Raschka, S. (2019). Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-learn, and TensorFlow 2, 3rd Edition. United Kingdom: Packt Publishing, 54-57.

Mukhopadhyay, S. (2018). Advanced Data Analytics Using Python: With Machine Learning, Deep Learning and NLP Examples. United States: Apress, 68-69.

Statistical Methods for Machine Learning: Discover how to Transform Data into Knowledge with Python. Brownlee, J. (2018). (n.p.): Machine Learning Mastery.

Brownlee, J. (2016). Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch. United States: Jason Brownlee, 34-35.

**Periodicals**

Hauke, Jan, and Tomasz Kossowski. “Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data.” Sciendo, Sciendo, 1 June 2011.

“CDC COVID Data Tracker.” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, www.cdc.gov/covid-data-tracker/index.html#demographics.