Group exercise 3: The spread of COVID-19 in the US

DATA5207: Data Analysis in the Social Sciences

Pooja Vijay Mahajan | SID: 510282930 | Group 34

Introduction

The first section discusses the factors that influence the spread of COVID-19. The next section focuses on cleaning and combining the data sets. This is followed by the selection of variables identifying the independent and dependent variables. In the fourth section, the descriptive analysis is done for the for your chosen variables, including analysis of their distributions as well as their relationship to the dependent variable. These are then fit into the regression model in the fifth section, the results are interpreted in the sixth section followed by a some concluding inferences in the seventh section.

1. Factors may influence the spread of COVID-19

There are several factors that may influence the spread of COVID-19. Amongst all the factors, the following factors are proved to have a solid affect on the COVID case count.

1. Mobility trends within workplaces and residential areas: People who stay at home are less vulnerable to COVID-19 since they are in protected areas whereas people who went to workplace even during the outbreak were exposed to this virus. Thus as the case count increased, more and more people stayed at home, and the number of people going to office reduced significantly.
2. Mobility trends for public places such as grocery markets, drug stores and pharmacies: Due to the fear of running out of grocery products during the lock downs, people started indulging in grocery shopping. Also they started vsisiting the pharmacies to equip themselves with all the medicines required for the treatment. Thus, as the case count increased, so did the visits to the grocery store and pharmacies.
3. Age: People belonging to old age are more prone to developing symptoms for the corona virus because of weaker immune systems. Thus as the case count increased more in areas which had more number of old age people.
4. Poverty: The living conditions of poor people could nopt allow them to buy expensive medicines, go for regular treatments and thus people living in poverty conditions influenced the spread of COVID-19.
5. Population density: Densely populated areas were difficult to control during the outbreak. If the state was densely populated, it was more than likely that it had more number of COVID-19 cases.

2. Clean and combine these three datasets

3 datasets have been used in this research work:

1. Johns Hopkins dataset (time series dataset which contains the number of COVID-19 cases)(Hopkins, 2022)
2. Google Mobility dataset (time series dataset which contains the mobility trends of various locations)(Google.com, 2021)

3. County dataset (static datset which contains the demographic, economic and climate characteristics)(COVID-19_US_County-level_Summaries/list_of_columns.md at master · JieYingWu/COVID-19_US_County-level_Summaries, 2022)

The following chunks of code cleans and combines these three datasets into one single dataset.

All the three datasets have been loaded and combined into one single data set. Data cleaning and manipulation involved renaming column, selecting columns that are necessary, and scaling the values of the remaining columns. The values have been scaled so that one variable is normalized within the range and different variations in the dataset do not affect the model outcomes. These continuous variables are standardized so that we can interpret their beta coefficients in relation to one another.

```
# load data
confirmed.cases.data <- read.csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_

# clean data
confirmed.cases.data2 <- confirmed.cases.data %>%
  dplyr::rename(state = Province_State,
                county = Admin2) %>%
  dplyr::select(-UID, -iso2, -iso3, -code3,
                -Lat, -Long_, -Combined_Key, -Country_Region) %>%
    gather(date, confirmed.cases,-state,-county,-FIPS) %>%
  dplyr::mutate(date = gsub('X', '', date),
                date = as.Date(as.character(date), "%m.%d.%y"),
                county_state = paste0(county, ', ', state)) %>%
  arrange(county_state, date) %>%
  dplyr::mutate(lag = slide(.,
                                  Var = 'confirmed.cases',
                                  NewVar = 'new',
                                  GroupVar = 'county_state',
                                  slideBy = -1)[,'new'],
                new.cases = confirmed.cases - lag)
```

```
##
## Remember to order . by county_state and the time variable before running.
```

```
##
## Lagging confirmed.cases by 1 time units.
```

```
## calculate rolling averages and remove non-states
covid.smooth.data_county <- confirmed.cases.data2 %>%
    dplyr::group_by(county_state, date) %>%
    dplyr::summarise(new.cases = sum(new.cases)) %>%
      dplyr::group_by(county_state) %>%
    dplyr::mutate(cases_14days = zoo::rollmean(new.cases,
                                               k = 14, fill = 0)) %>%
  dplyr::ungroup() %>%
  mutate() %>%
  merge(confirmed.cases.data2 %>%
          dplyr::select(county_state, date, county, state)) %>%
  filter(!state %in% c('American Samoa',
                       'Diamond Princess',
                       'Grand Princess',
                       'Guam',
```

```
                        'Northern Mariana Islands',
                        'Puerto Rico',
                        'Virgin Islands'))
```

```
## `summarise()` has grouped output by 'county_state'. You can override using the
## `.groups` argument.
```

```
#load google mobility data
google.mobility <- read.csv('Data/Global_Mobility_Report.csv') %>%
  filter(country_region == 'United States')
```

```
#load additional predictor data
county.data <- read.csv('https://raw.githubusercontent.com/JieYingWu/COVID-19_US_County-level_Summaries/
```

```
#rename equivalent columns in both datasets as county
names(google.mobility)[names(google.mobility) == 'sub_region_2'] <- 'county'
merged_df=merge(covid.smooth.data_county, google.mobility, by=c("date","county"))
```

```
#formatting county column in third dataset and selecting relevant columns
county.data.1<-county.data%>%
  filter(grepl('County', Area_Name))%>%
  mutate(Area_Name = str_remove_all(Area_Name, " County"))

names(county.data.1)[names(county.data.1) == 'Area_Name'] <- 'county'

county.data.2 <- county.data.1%>%
  dplyr::select('county','POP_ESTIMATE_2018', 'POVALL_2018','Total_age65plus')

merged_df_1=merge(merged_df, county.data.2, by="county")
```

```
#remove unwanted columns and scale the remaining columns
scaled_df <- merged_df_1%>%
  mutate_at(c("new.cases","cases_14days","grocery_and_pharmacy_percent_change_from_baseline","parks_per
  dplyr::select(-county_state, -country_region_code,-country_region, -sub_region_1, -metro_area, -iso_3
```

```
#renaming the columns

names(scaled_df)[names(scaled_df) == "new.cases"] <- "case_count"
names(scaled_df)[names(scaled_df) == "grocery_and_pharmacy_percent_change_from_baseline"] <- "grocery_v
names(scaled_df)[names(scaled_df) == "workplaces_percent_change_from_baseline"] <- "workplace_mobility"
names(scaled_df)[names(scaled_df) == "POP_ESTIMATE_2018"] <- "population_density"
names(scaled_df)[names(scaled_df) == "POVALL_2018"] <- "poverty"
names(scaled_df)[names(scaled_df) == "Total_age65plus"] <- "age65plus"
```

3. Selection of variables

After performing the descriptive analysis, the following variables have been selected as:

1. Dependent variable: case_count
2. Independent variable: population_density, poverty, workplace_mobility, grocery_visit, age65plus

This implies that factors such as crowded areas, living conditions(poverty), commuting to places such as offices, pharmacies and grocery shops all have an impact on the spread of COVID-19.

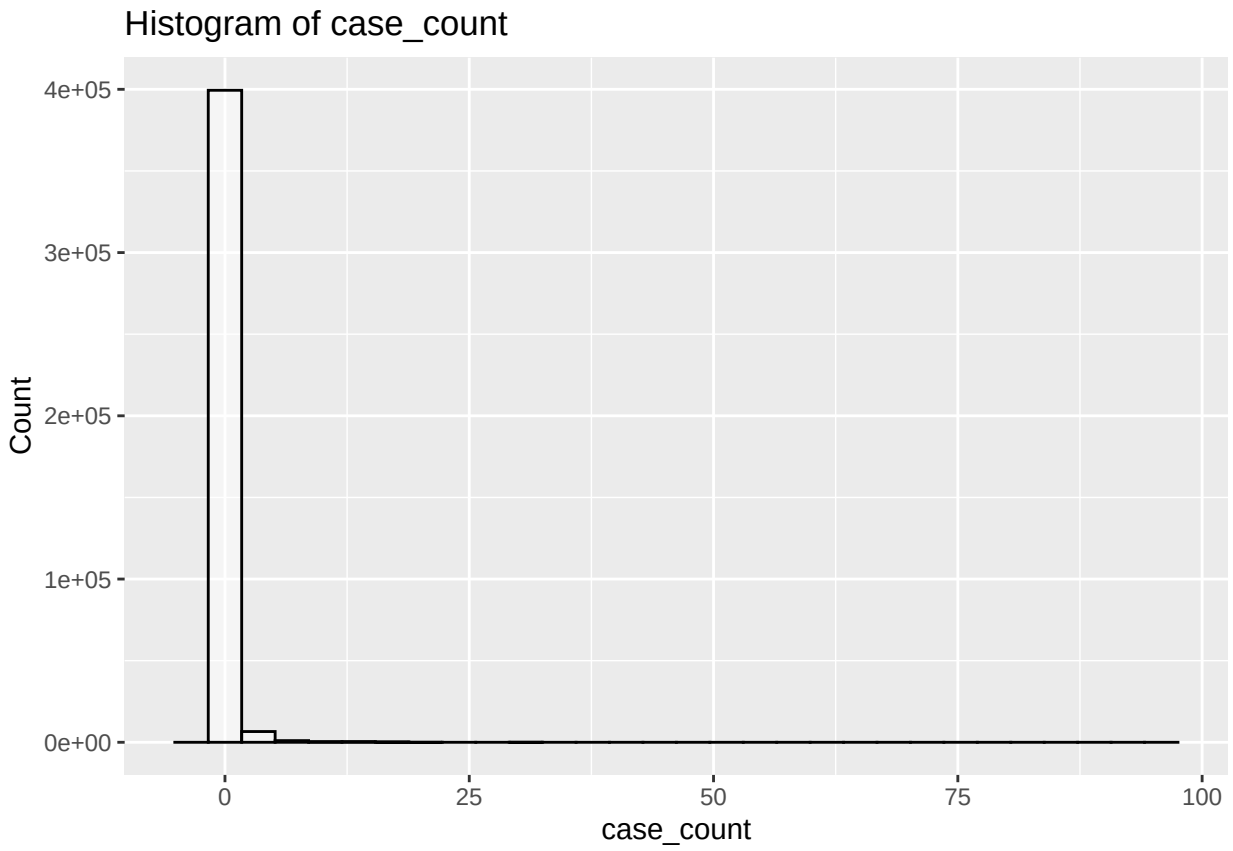
4. Descriptive Analytics

Histograms: The histograms for all the variables were plotted. It is left skewed for case_count, majority of the values for case count, population density, poverty, age65plus are centered towards mean and there are few values plummeting towards the extreme right end. The histograms for grocery_visits and work-place_mobility are bell-shaped implying majority of the values are centered towards the mean.

Scatter plot: The distribution of case count over time reflects the dates with highest number of cases and the time duration when the case count was plummeting.

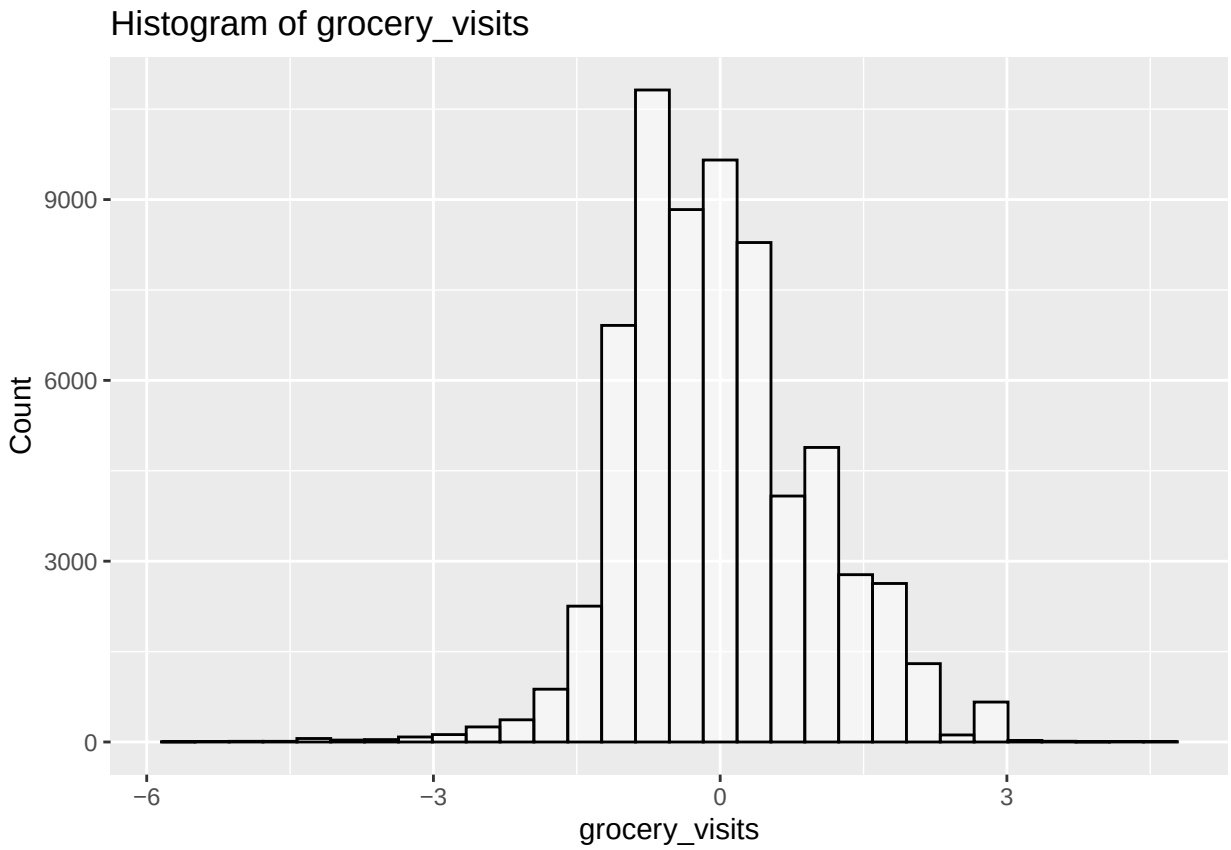
```
ggplot(scaled_df, aes(x=case_count)) +
        geom_histogram( colour="black", fill="white", alpha=0.5, xlab=NULL)+
        labs(title="Histogram of case_count",x="case_count", y = "Count")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



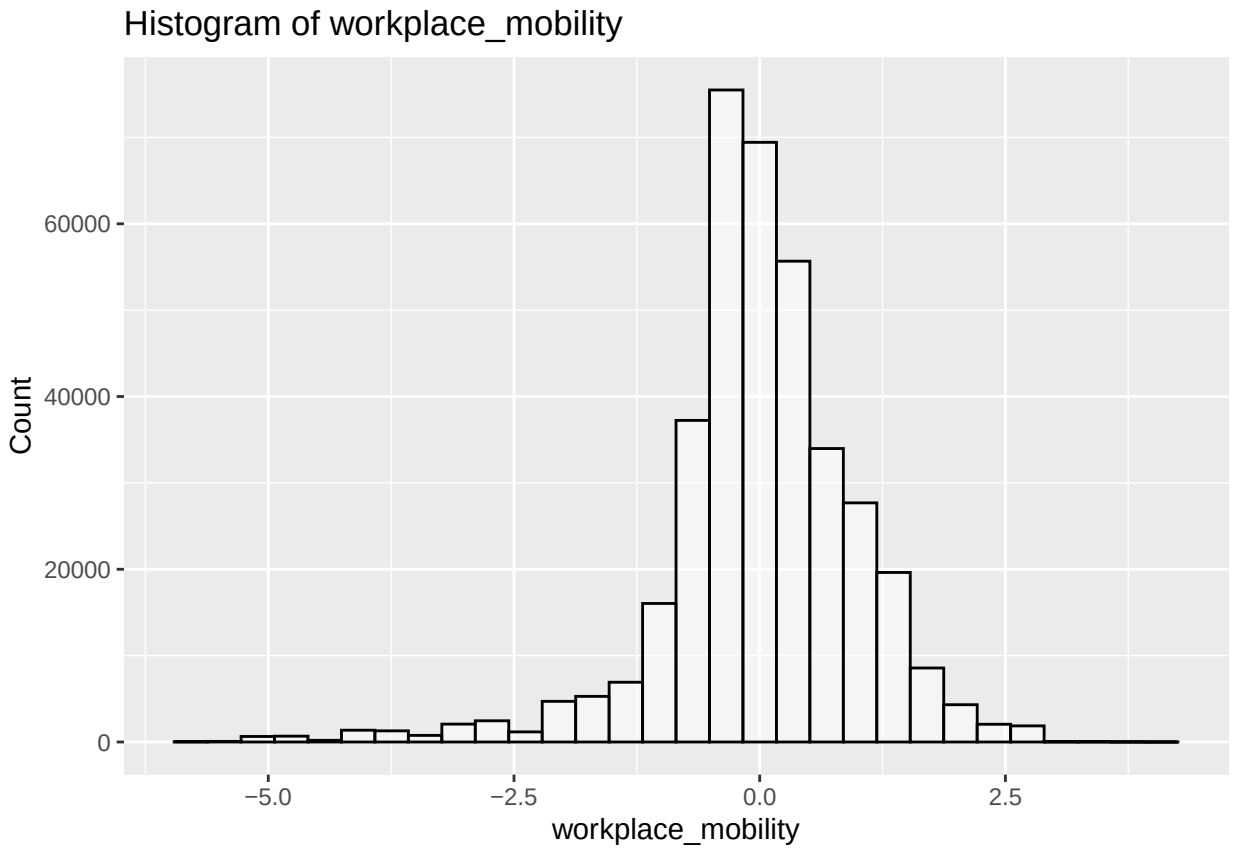

```
#plotting histogram for grocery_visits
ggplot(scaled_df, aes(x=grocery_visits)) +
        geom_histogram( colour="black", fill="white", alpha=0.5, xlab=NULL)+
        labs(title="Histogram of grocery_visits",x="grocery_visits", y = "Count")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



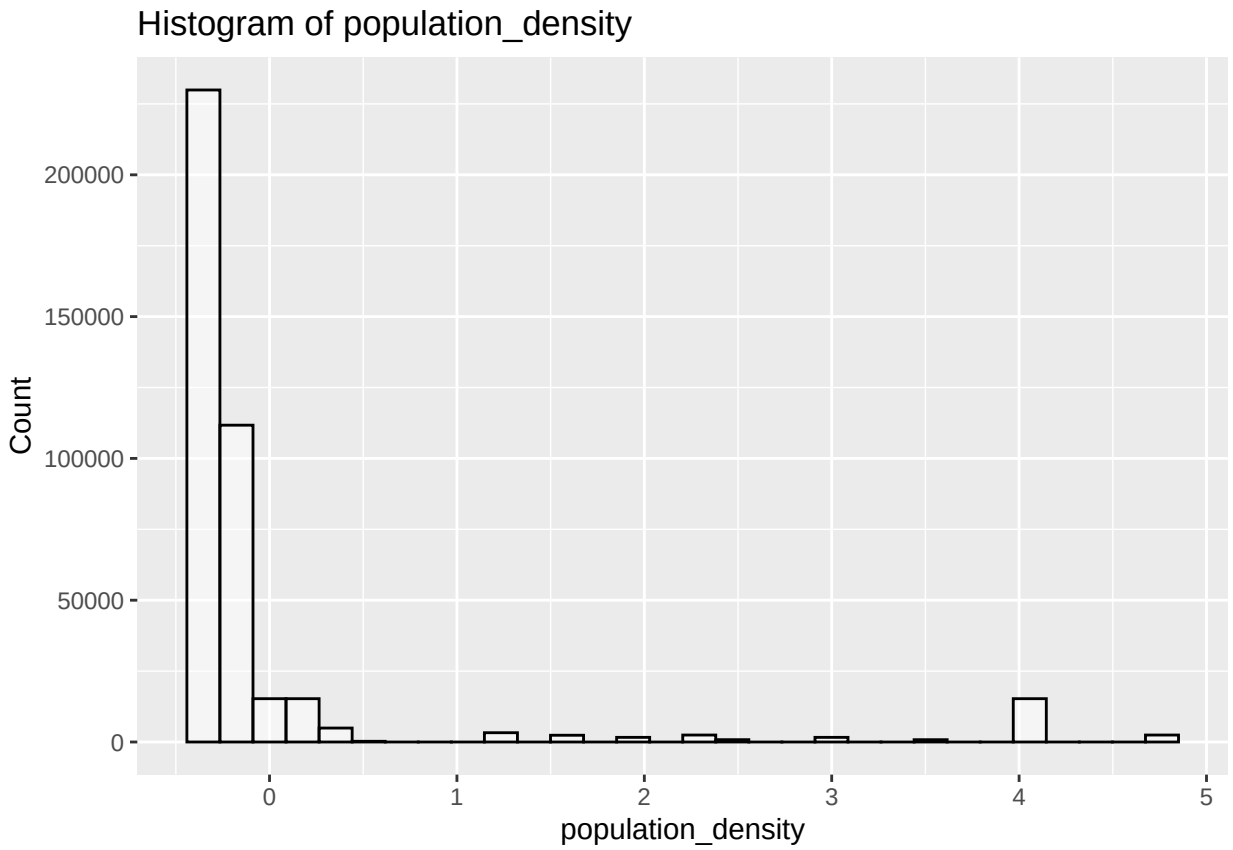

```
#plotting histogram for workplace_mobility
ggplot(scaled_df, aes(x=workplace_mobility)) +
         geom_histogram( colour="black", fill="white", alpha=0.5, xlab=NULL)+
         labs(title="Histogram of workplace_mobility",x="workplace_mobility", y = "Count")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



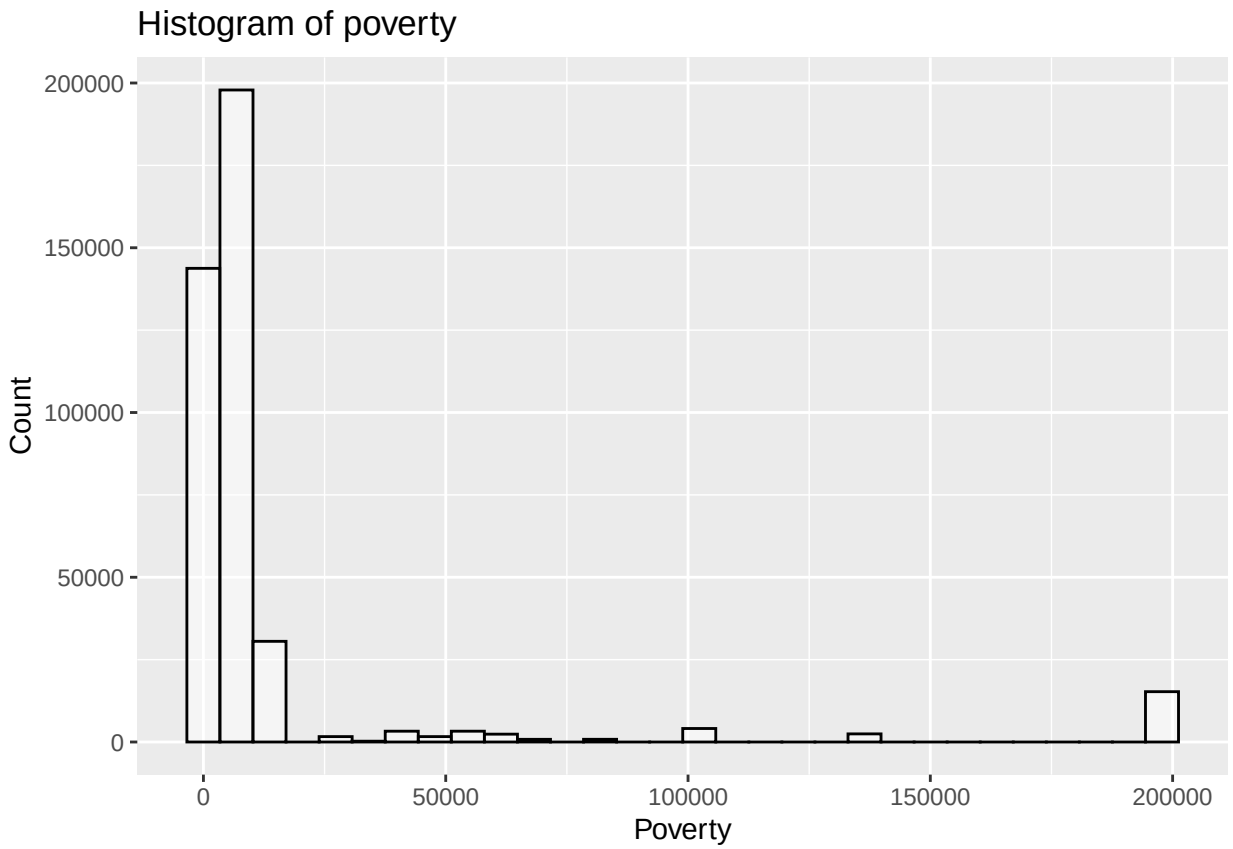

```
#plotting histogram for population_density
ggplot(scaled_df, aes(x=population_density)) +
         geom_histogram( colour="black", fill="white", alpha=0.5, xlab=NULL)+
         labs(title="Histogram of population_density",x="population_density", y = "Count")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



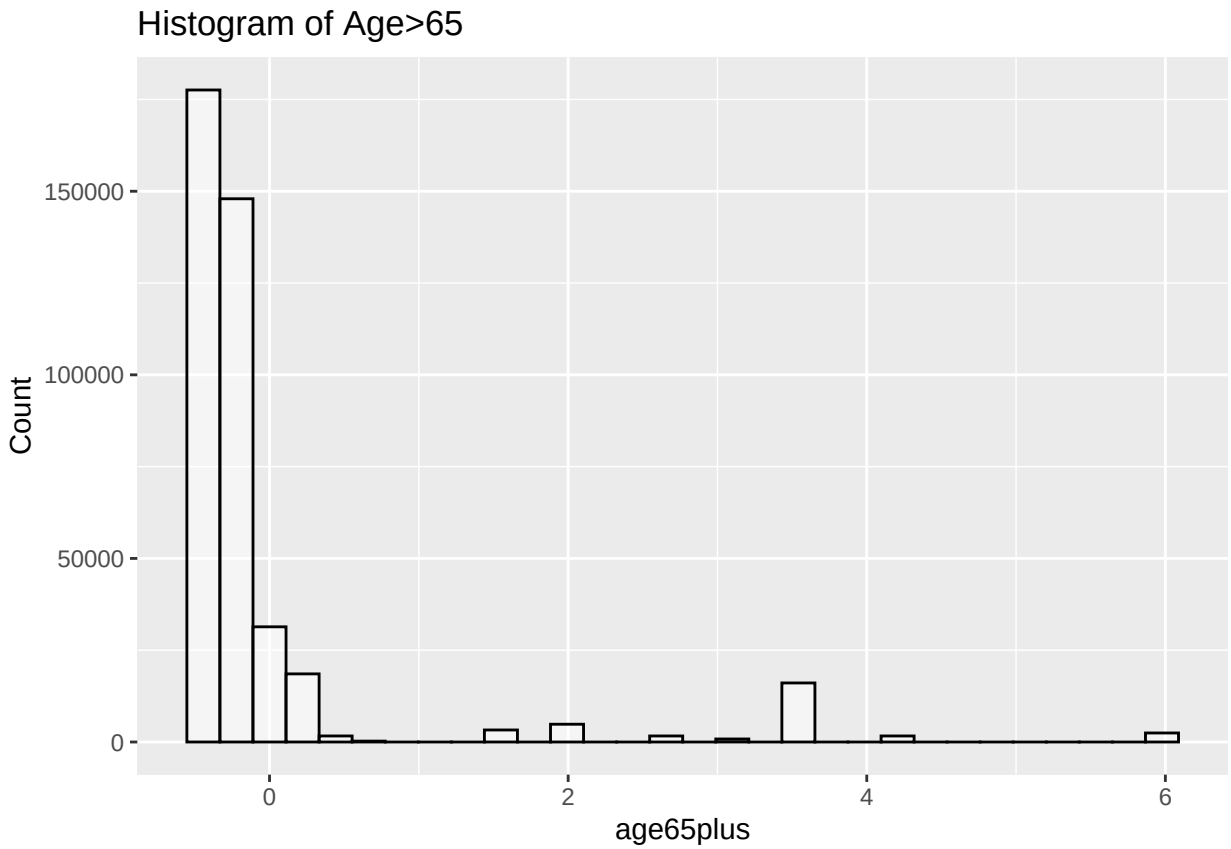

```
#plotting histogram for poverty
ggplot(scaled_df, aes(x=poverty)) +
         geom_histogram( colour="black", fill="white", alpha=0.5, xlab=NULL)+
         labs(title="Histogram of poverty",x="Poverty", y = "Count")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#plotting histogram for age65plus
ggplot(scaled_df, aes(x=age65plus)) +
         geom_histogram( colour="black", fill="white", alpha=0.5, xlab=NULL)+
         labs(title="Histogram of Age>65",x="age65plus", y = "Count")
```

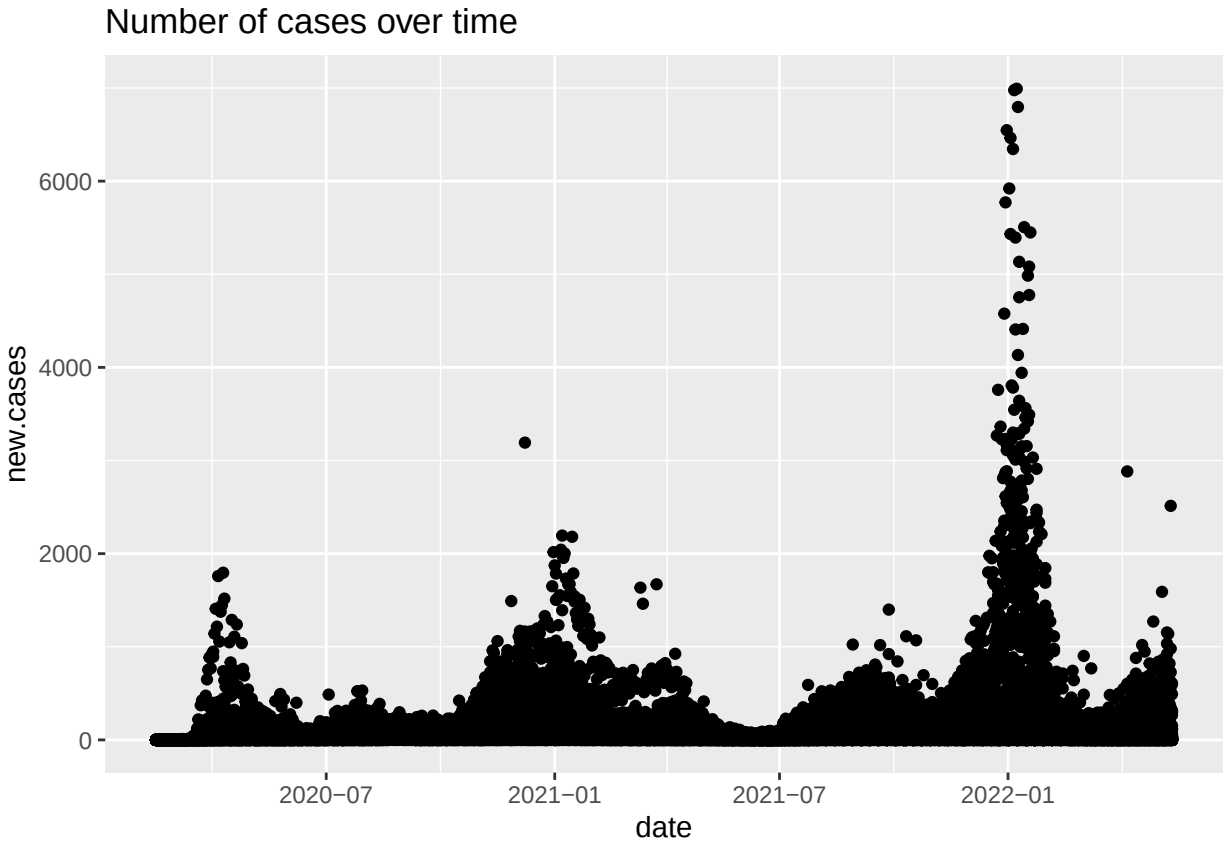
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Since the case count is the dependent variable, it is important to this analyse the variable over time. As it

can be observed from the plot below, the case count has increased in January 2021, reduced in July 2021 and increased again in January 2022.

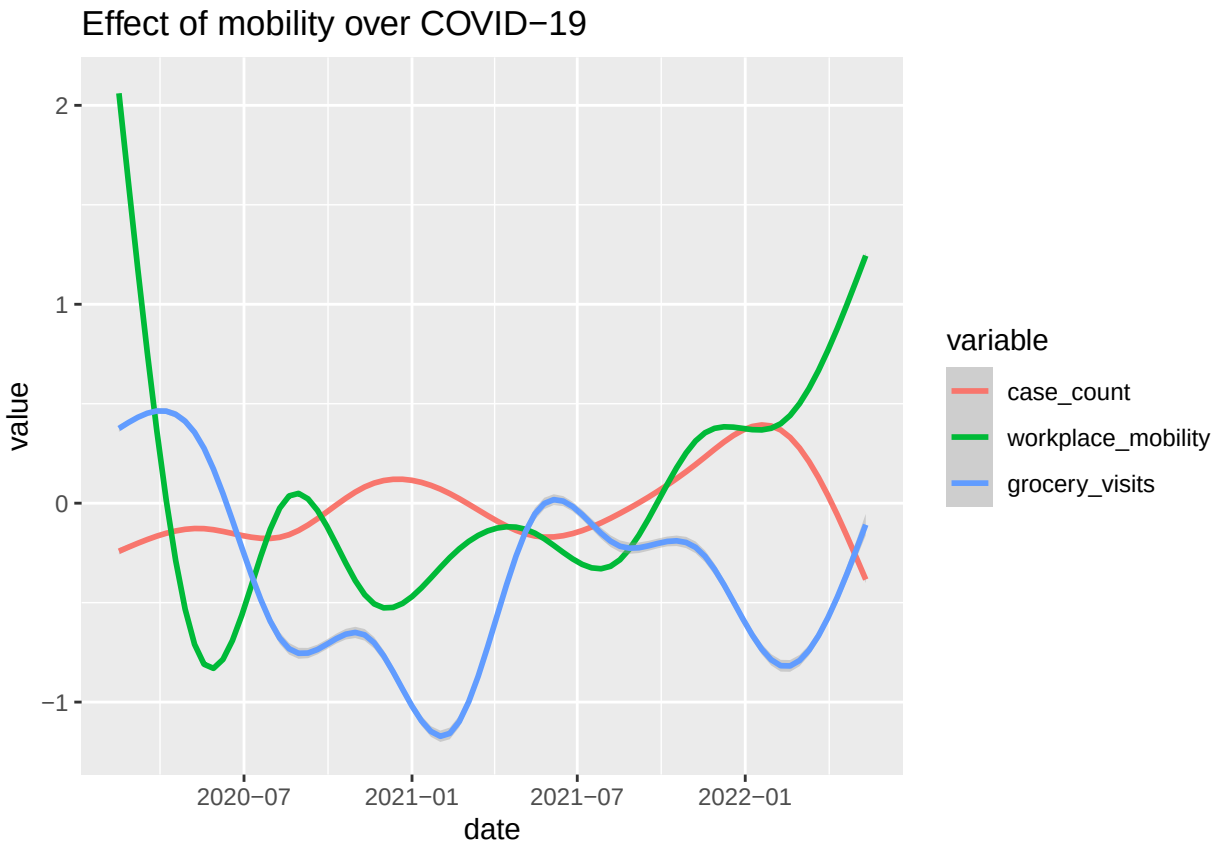
```
p1<-ggplot(merged_df, aes(x=date, y=new.cases))
p1+geom_point()+ylim(-1,7000)+ggtitle('Number of cases over time')
```



The variables were melted and overlapped to analyse them concurrently. The plot below describes the relationship between case count and mobility of people in workplaces and grocery stores and pharmacies. It can be observed that the case count and the predictors moved in opposite directions. As the case count increased, the mobility towards crowded areas such as offices, grocery shops decreased.

```
#analyzing correlation between 'case_count','workplace_mobility','grocery_visits'
col1 = c('case_count','workplace_mobility','grocery_visits')
melted_df <- melt( scaled_df, measure.vars=col1, value.names="Values", varialbe.name="varialbe" )
plot <- ggplot(melted_df, aes(x=date, y=value, color=variable)) + stat_smooth(span=0.05)
plot+ggtitle('Effect of mobility over COVID-19')
```

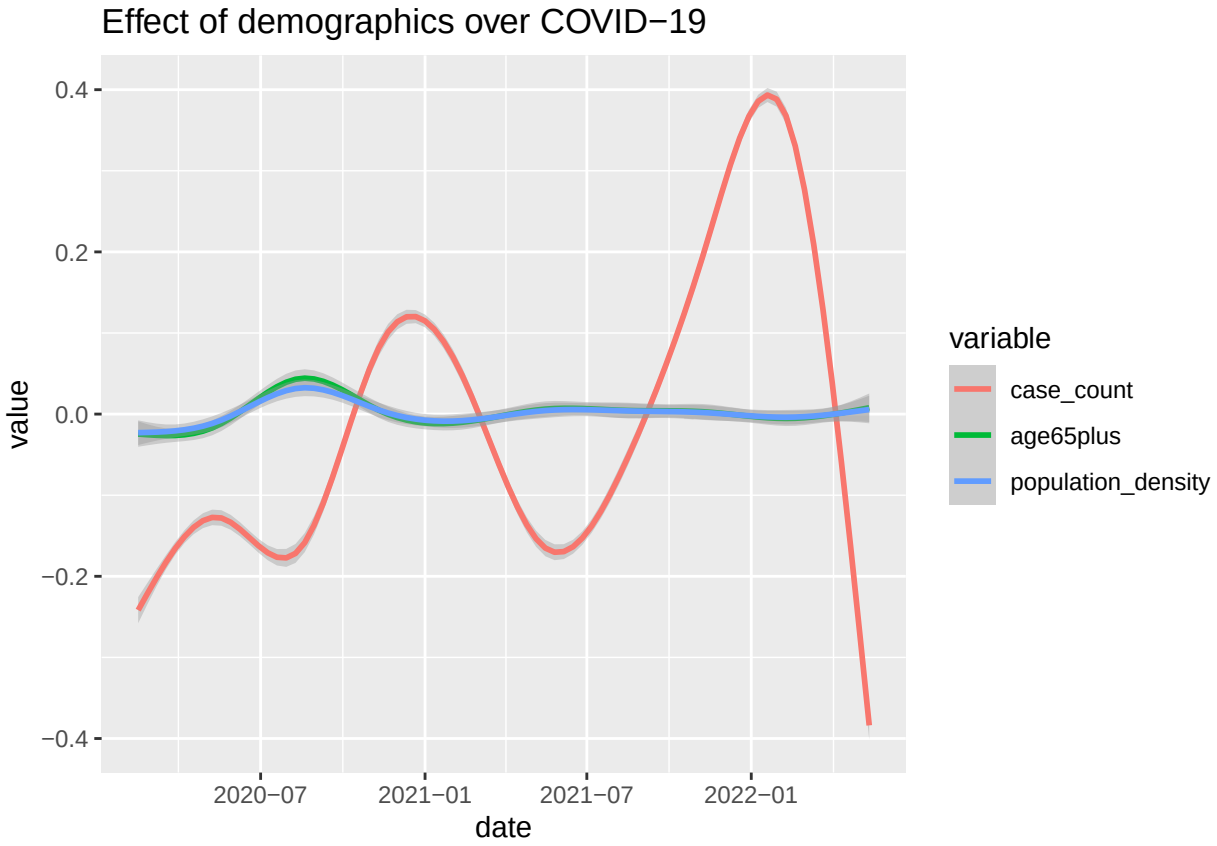
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



The following plot addresses the variation of population density and specifically people with age greater than 65 years. It can be observed that during July 2020, the case count was low and the age and population density was high. However in January 2021 (within 6 months), things turned upside down. The case count increased and the population count and the number of people with age greater than 65 years decreased.

```
#analyzing correlation between 'case_count','age65plus', 'population_density'
col1 = c('case_count','age65plus', 'population_density')
melted_df <- melt( scaled_df, measure.vars=col1, value.names="Values", varialbe.name="varialbe" )
plot <- ggplot(melted_df, aes(x=date, y=value, color=variable)) + stat_smooth(span=0.05)
plot+ggtitle('Effect of demographics over COVID-19')
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

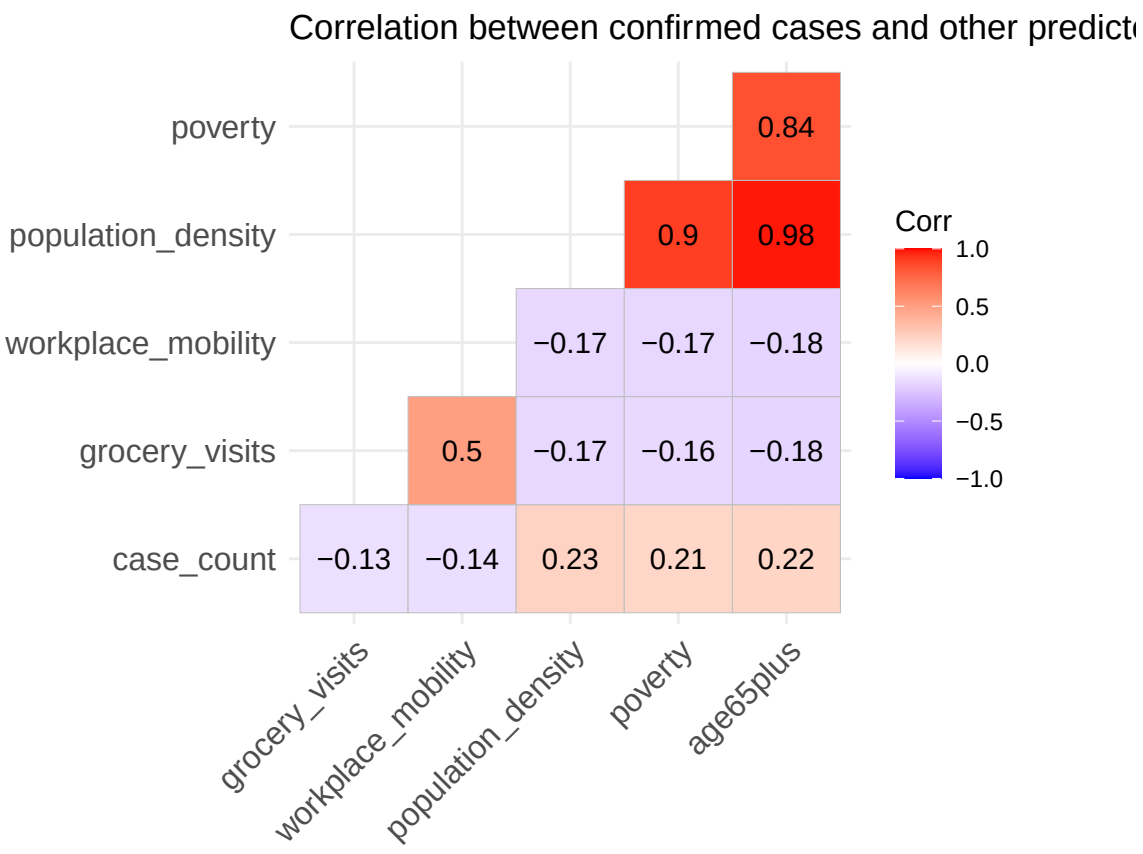


A correlation matrix was created to illustrate the strength of relationship between the independent variables and dependent variable.The pearson coefficient was applied in order to determine the statistical significance between the variables. The reason behind this is that the past research work has proved pearson coefficient to be reliable accurate for tracing correlation (Benesty, 2009). 0.5 is the sig.level (significance value) so that the model has a metric to gauge trust for correlation identification and it's statistical significance.

From the correlation matrix it can be observed that:

1. There is a strong positive correlation between age, population density and age65plus. This implies that as population increases, poverty increases and so do the number of people with age greater than 65.
2. There is a negative correlation between mobility trends in workplace and grocery shops when compared with variables such as population density, age65plus and poverty. For instance, With reduced mobility, there is increase in population density (crowded areas in offices and shops). However, there is a positive correlation between mobility in workplace and mobility in grocery shops.
3. There is a negative correlation between case count and mobility in workplace and grocery shops. This is explained by the fact that as the number of COVID-19 cases increased, people started working from home instead of offices and made less frequent visits to public places such as shopping marts, pharmacies etc.
4. There is positive correlation between case count and population_density, age65plus, poverty. This implies that there were more COVID-19 cases where the population density was higher. Old age people were most vulnerable to this virus. The living conditions of poor people made them more susceptible to COVID-19.

```r
#creating a correlation matrix
mat=data.frame(scaled_df$case_count,scaled_df$grocery_visits, scaled_df$workplace_mobility, scaled_df$p

#renaming columns
names(mat)[names(mat) == "scaled_df.case_count"] <- "case_count"
names(mat)[names(mat) == "scaled_df.grocery_visits"] <- "grocery_visits"
names(mat)[names(mat) == "scaled_df.workplace_mobility"] <- "workplace_mobility"
names(mat)[names(mat) == "scaled_df.population_density"] <- "population_density"
names(mat)[names(mat) == "scaled_df.poverty"] <- "poverty"
names(mat)[names(mat) == "scaled_df.age65plus"] <- "age65plus"

#ommitting null values
mat<- na.omit(mat)

#correlation matrix
correlation_matrix = cor(mat)

#using pearson coefficient
pearson_corr_mat <- cor_pmat(mat, use = "complete", method = "pearson")

#plotting the correlation matrix
ggcorrplot(correlation_matrix, type = "lower", title='Correlation between confirmed cases and other pre
```



5. Fitting a regression model

Now that the correlation strength was identified between the independent and dependent variables, it was fit into linear regression model. This model was chosen because the variables exhibited a strong relationship in their descriptive analytics and thus, linear regression is best option to deal with such cases.

```
#Fitting a regression model
model.1<-lm(case_count~ grocery_visits+workplace_mobility+age65plus+population_density+poverty, data=sca
summary(model.1)
```

```
##
## Call:
## lm(formula = case_count ~ grocery_visits + workplace_mobility +
##     age65plus + population_density + poverty, data = scaled_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.774 -0.399 -0.139  0.049 95.960
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          4.347e-02  1.212e-02   3.588 0.000334 ***
## grocery_visits      -8.651e-02  8.025e-03 -10.780  < 2e-16 ***
## workplace_mobility -8.882e-02  5.310e-03 -16.728  < 2e-16 ***
## age65plus           -4.856e-02  3.335e-02  -1.456 0.145469
## population_density   2.995e-01  4.967e-02   6.030 1.65e-09 ***
## poverty             -8.013e-08  5.215e-07  -0.154 0.877872
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.628 on 55175 degrees of freedom
##   (352895 observations deleted due to missingness)
## Multiple R-squared:  0.06445,    Adjusted R-squared:  0.06437
## F-statistic: 760.2 on 5 and 55175 DF,  p-value: < 2.2e-16
```

6. Interpreting the results

The linear regression model was used to evaluate the impact of predictors on number of confirmed cases. This is because the predictors seems to have a correlation with the outcome variable. As per the summary, it can be concluded that:

1. In the values of residuals, the value of median is close to zero implying that the residuals are symmetric in nature giving balanced predictions at the extremities of the data set.
2. The standard error coefficient is 4.347e-02. It means that if all the predictors were set to their baseline value or zero, the value of case count would be 4.347e-02. This implies that if people did not form crowds at grocery shops or they did not go to workplace and stayed at home, the case count would be negligible.
3. The variable 'grocery_visits' , workplace_mobility,age65plus has a negative correlation with a very small p-value implying greater statistical significance.
4. The residual standard error is 1.62.This implies that the predicted values are 1.62 away from the actual value. The adjusted R squared is 0.064 implying that the predictors account for just 6.4% of variation

in the depend_ent variable, case_count. This could be improved by selecting better predictors in order to facilitate better variation in the case_count. The F-statistic has a high value and thevalue of p-value is negligible. All these inferences conclude that the relationship outcome delivered by the model doesnt imply strong correlation between the independent and dependent variables.

7. Conclusion

The COVID-19 pandemic has made a huge impact on all beings and proved to be disruptive to social groups which were most vulnerable eg. infants, old age population, crowded groups at public places etc. This has had a huge impact on the lifestyle of people in such a way that people have started working from home rather than going to office, there are more frequent visits to the pharmacy now more than ever, people indulged in panic-shopping and so on and so forth. This research intended to study the county-level data from the US fit a model to estimate the predictors for the spread of the corona virus. It was evident that there were more cases in areas which were densely populated, or had people living in poor conditions or people with age greater than 65. The count count also had an impact by the mobility trends within workplace, grocery shops, pharmacies. The linear regression model partially supported the null hypotheis and it's performance can be improved by selection of different variables and rigorous testing.

8. References

Abu-Rayash, A. and Dincer, I., 2020. Analysis of mobility trends during the COVID-19 coronavirus pandemic: Exploring the impacts on global aviation and travel in selected cities. Energy research & social science, 68, p.101693.

Benesty, J., Chen, J., Huang, Y. and Cohen, I., 2009. Pearson correlation coefficient. In Noise reduction in speech processing (pp. 1-4). Springer, Berlin, Heidelberg.

Galanti, T., Guidetti, G., Mazzei, E., Zappalà, S. and Toscano, F., 2021. Work from home during the COVID-19 outbreak: The impact on employees' remote work productivity, engagement, and stress. Journal of occupational and environmental medicine, 63(7), p.e426. GitHub. 2022. COVID-19_US_County-level_Summaries/list_of_columns.md at master · JieYingWu/COVID-19_US_County-level_Summaries. [online] Available at: https://github.com/JieYingWu/COVID-19_US_County-level_Summaries/blob/master/data/list_of_columns.md [Accessed 30 May 2022].

JieYingWu/COVID-19_US_County-level_Summaries, 2022) Google.com. 2021. [online] Available at: https://www.google.com/covid19/mobility/data_documentation.html?hl=en [Accessed 26 May 2022].

Gupta, S. and Jawanda, M.K., 2020. The impacts of COVID-19 on children. Acta Paediatr, 109(11), pp.2181-2183. Hopkins, J., 2022. CSSEGISandData - Overview. [online] GitHub. Available at: https://github.com/CSSEGISandData [Accessed 20 May 2022].

Kadi, N. and Khelfaoui, M., 2020. Population density, a factor in the spread of COVID-19 in Algeria: statistical study. Bulletin of the National Research Centre, 44(1), pp.1-7. Kang, S.J. and Jung, S.I., 2020. Age-related morbidity and mortality among patients with COVID-19. Infection & chemotherapy, 52(2), p.154.

Kuy, S., Tsai, R., Bhatt, J., Chu, Q.D., Gandhi, P., Gupta, R., Gupta, R., Hole, M.K., Hsu, B.S., Hughes, L.S. and Jarvis, L., 2020. Focusing on vulnerable populations during COVID-19. Academic Medicine.

Natividade, M.D.S., Bernardes, K., Pereira, M., Miranda, S.S., Bertoldo, J., Teixeira, M.D.G., Livramento, H.L. and Aragão, E., 2020. Social distancing and living conditions in the pandemic COVID-19 in Salvador-Bahia, Brazil. Ciência & Saúde Coletiva, 25, pp.3385-3392.

Sy, K.T.L., White, L.F. and Nichols, B.E., 2021. Population density and basic reproductive number of COVID-19 across United States counties. PloS one, 16(4), p.e0249271.