SEMESTER 2 COMBINED PROJECT

GROUP 6

D GUNAVANTH MAHENDRA

SHRIYA KATAKAM

JASH DATTANI

PRANAVI SHAH

JAS RATHI

**Preface:**

* The given data was based on the Covid-19 Pandemic.
* Part I involves state wise Indian data from the 1st of April 2020 to the 8th of December 2020.
* Part II has worldwide data on the number of confirmed, active, recovered cases and deaths.
* Part III looked at how one would price insurance policies based on given data and charges for a person by analysing factors such as age, BMI, smoking status, and region.

**Basic methodology followed:**

* The first step followed in all the parts was cleaning of data, which has been explained in the detailed report below.
* Data cleaning was followed by analysis of the data and making required transformations wherever necessary.
* Code was then written in R to meet the objectives set out in the project description.
* Wherever necessary, assumptions were made and highlighted in detailed analysis below.
* The tidyverse package has been used extensively throughout the project. Dplyr and GGplot2 are the major tidyverse packages that have been used.
* Wherever necessary, functions have been defined and highlighted in the report below.

**PART I:**

**Statewise testing data has been used here to ascertain the effects of state wise sample testing.**

**The data has been cleaned using the following steps:**

* **All columns converted to their appropriate types:**

**(Date: date, Total Samples: Numeric, Negative: Numeric, Positive: Numeric)**

* **Added in a state\_factor column that considers each state as a separate factor.**
* **Assumption is made here that for missing datapoints, the Total Samples add up to the sum of Positive and Negative samples with no samples missing.**
* **Added in missing values for Positive and Negative columns where possible using Total Samples – Positive (for Negative), and Total Samples – Negative (for Positive). Ignored entries where both values were missing.**

1. **Correlation Tests:**

**For all the tests:**

* **Ran t-tests at the 95% confidence level.**
* **All columns vs Positive**
* **H0: True correlation is equal to zero.**
* **H1: True correlation is not equal to zero.**

1. **Date vs Positive:**

**Correlation: 0.3657452**

**p-value: 2.2e-16**

**Conclusion: There is sufficient evidence to reject H0 at the 95% level, hence it is reasonable to assume that the correlation is not equal to zero.**

1. **Sum of cases with respect to each state :-**

|  |  |
| --- | --- |
| **State** | **Cases** |
| **Andaman and Nicobar Islands** | **474804** |
| **Andhra Pradesh** | **67644725** |
| **Arunachal Pradesh** | **2237299** |
| **Assam** | **2065991** |
| **Bihar** | **1859345** |
| **Chandigarh** | **1427953** |
| **Chhattisgarh** | **467857** |
| **Dadra and Nagar Haveli and Daman and Diu** | **224199** |
| **Delhi** | **6848173** |
| **Goa** | **266181** |
| **Gujarat** | **8009517** |
| **Haryana** | **19221575** |
| **Himachal Pradesh** | **2447740** |
| **Jammu and Kashmir** | **9869319** |
| **Jharkhand** | **9922526** |
| **Karnataka** | **4701197** |
| **Kerala** | **36405434** |
| **Ladakh** | **582004** |
| **Madhya Pradesh** | **22364678** |
| **Maharashtra** | **96902754** |
| **Manipur** | **101501** |
| **Meghalaya** | **704121** |
| **Mizoram** | **19785** |
| **Nagaland** | **90682** |
| **Odisha** | **2214458** |
| **Puducherry** | **3661073** |
| **Punjab** | **960287** |
| **Rajasthan** | **21231125** |
| **Sikkim** | **19642** |
| **Tamil Nadu** | **12772604** |
| **Telangana** | **3855373** |
| **Tripura** | **3021038** |
| **Uttar Pradesh** | **2743971** |
| **Uttarakhand** | **7962420** |
| **West Bengal** | **3487431** |

1. **Calculated the average daily positive and negative cases by considering the total number of samples, positive and negative tests on each day for all states.**

**Took average using formula:**

**Sum of positive (negative) tests / Number of positive (negative) tests observations**

1. **Average daily positive cases = 1415829**
2. **Average daily negative cases = 16058298**

**PART II:**

**Functions Defined**:

* colfunc (number): Generates specified number of colours between 2 colours in a spectrum. Used for adding colour to plots.
* date\_generator (date, nday): Takes an initial date and returns a vector that contains date plus the next number of days (nday) specified.

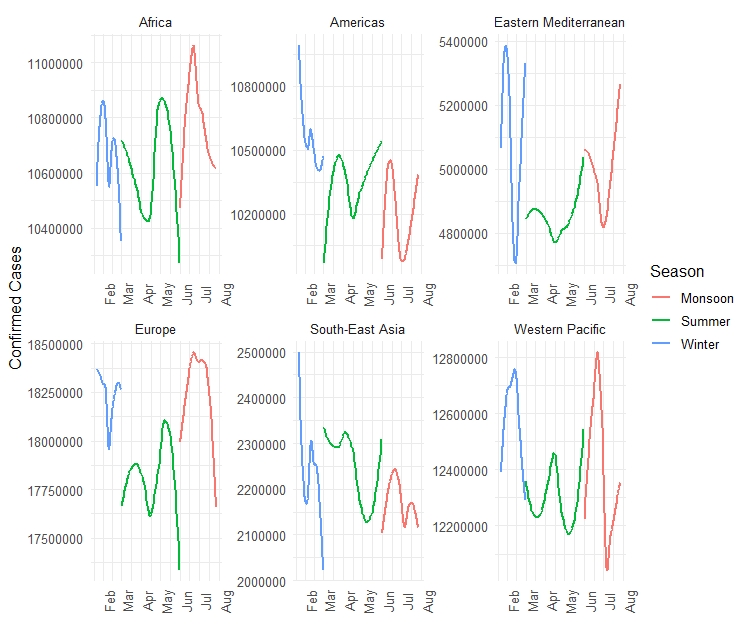
**Additional Packages Used**:

* lubridate
* MASS
* Scales

Data Cleaning:

* Took a country and date wise sum of Confirmed, Deaths, Recovered, and Active columns and added in the WHO Region column.

1. **Seasonality Analysis:**
   * Grouped data by region and date.
   * Defined seasons as follows:
     + Winter: 2020-01-22 to 2020-02-29
     + Summer: 2020-03-01 to 2020-05-31
     + Monsoon: 2020-06-01 to 2020-07-27
   * Seasons have been used to group the data based on set of dates to provide a clearer picture of seasonality trends. Common seasonality has been taken for the entire data set.



* A closer look at the graphs shows us that on average, the cases all over the world reduced during the months between February 2020 and April 2020. However, we see a 2nd wave of infections come in towards the beginning of May which also wanes out as we approach July.
* Within seasons we see marginal spikes in cases mainly in the summer and monsoon which is in line with our observation above. In the winter months, the cases see fluctuations albeit, the overall effect was one of decreasing cases.