Semester 2 Combined Project

Group 6

D GUNAVANTH MAHENDRA

JASH DATTANI

JAS RATHI

PRANAVI SHAH

SHRIYA KATAKAM

**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.
* In addition to tidyverse, the stats and MASS packages have been used in the project.
* Wherever necessary, functions have been defined and highlighted in the report below.

**PART I:**

State wise 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. Total samples vs Positive

Correlation: 0.8677797

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. Negative vs Positive

Correlation: 0.8329745

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. State vs Positive

Correlation: -0.03968859

p-value: 0.001918

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.

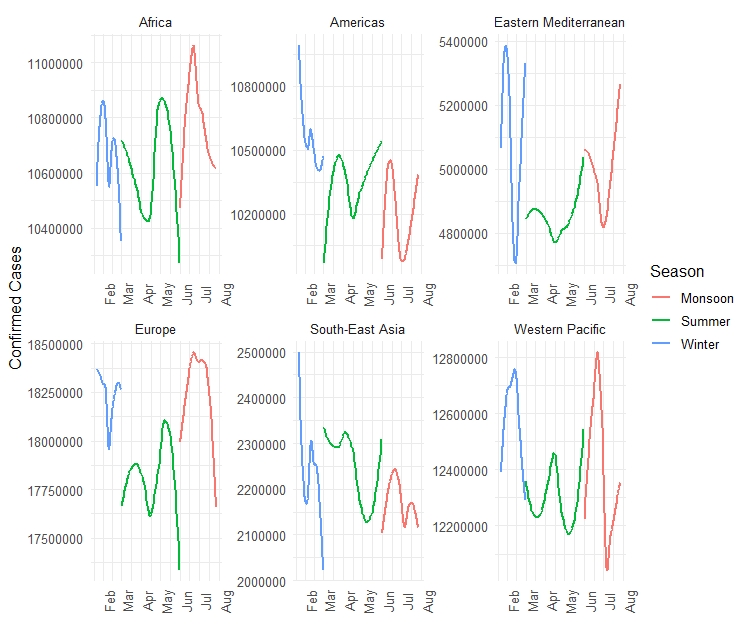
**Additional Packages Used**:

* lubridate
* 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: 22-01-2020 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.

1. **Correlation Tests:**

All tests have been performed at the 95% level.

Average Deaths vs Average Recovered:

H0: Correlation between average deaths and average recovered is 0.

H1: Correlation between average deaths and average recovered is 0.

1. Africa deaths vs Africa recovered

Correlation: 0.03529171

p-value: 0.6306

95% confidence interval: (-0.1083658 0.1775054)

Conclusion: There is insufficient evidence to reject H0 at the 95% level, hence it is reasonable to assume that the correlation is equal to zero.

1. Americas deaths vs Americas recovered

Correlation: -0.01053093

p-value: 0.8859

95% confidence interval: (-0.1534100 0.1327794)

Conclusion: There is insufficient evidence to reject H0 at the 95% level, hence it is reasonable to assume that the correlation is equal to zero.

1. Eastern Mediterranean deaths vs Eastern Mediterranean recovered

Correlation: -0.01748684

p-value: 0.8117

95% confidence interval: (-0.1601962 0.1259386)

Conclusion: There is insufficient evidence to reject H0 at the 95% level, hence it is reasonable to assume that the correlation is equal to zero.

1. Europe deaths vs Europe recovered

Correlation: 0.03857448

p-value: 0.5992

95% confidence interval: (-0.1051160 0.1806873)

Conclusion: There is insufficient evidence to reject H0 at the 95% level, hence it is reasonable to assume that the correlation is equal to zero.

1. South-East Asia deaths vs South-East Asia recovered

Correlation: 0.04004836

p-value: 0.5853

95% confidence interval: (-0.1036560 0.1821148)

Conclusion: There is insufficient evidence to reject H0 at the 95% level, hence it is reasonable to assume that the correlation is equal to zero.

1. Western Pacific deaths vs Western Pacific recovered

Correlation: 0.0110278

p-value: 0.8806

95% confidence interval: (-0.1322912 0.1538952)

Conclusion: There is insufficient evidence to reject H0 at the 95% level, hence it is reasonable to assume that the correlation is equal to zero.

Average Active to Average Deaths:

H0: Correlation between average active and average deaths is 0.

H1: Correlation between average active and average deaths is 0.

1. Africa active vs Africa deaths

Correlation: -0.03192917

p-value: 0.6636

95% confidence interval: (-0.1742432 , 0.1116914)

Conclusion: There is insufficient evidence to reject H0 at the 95% level, hence it is reasonable to assume that the correlation is equal to zero.

1. Americas active vs Americas deaths

Correlation: 0.01755777

p-value: 0.811

95% confidence interval: (-0.1258687, 0.1602653)

Conclusion: There is insufficient evidence to reject H0 at the 95% level, hence it is reasonable to assume that the correlation is equal to zero.

1. Eastern Mediterranean active vs Eastern Mediterranean deaths

Correlation: -0.01476824

p-value: 0.8406

95% confidence interval: (-0.1575455, 0.1286138)

Conclusion: There is insufficient evidence to reject H0 at the 95% level, hence it is reasonable to assume that the correlation is equal to zero.

1. Europe active vs Europe deaths

Correlation: 0.06424406

p-value: 0.3811

95% confidence interval: (-0.07959801 , 0.20546526)

Conclusion: There is insufficient evidence to reject H0 at the 95% level, hence it is reasonable to assume that the correlation is equal to zero.

1. South-East Asia active vs South-East Asia deaths

Correlation: 0.1093876

p-value: 0.1351

95% confidence interval: (-0.03425891 0.24860608)

Conclusion: There is insufficient evidence to reject H0 at the 95% level, hence it is reasonable to assume that the correlation is equal to zero.

1. Western Pacific active vs Western Pacific deaths

Correlation: 0.01306756

p-value: 0.8587

95% confidence interval: (-0.1558863, 0.1302863)

Conclusion: There is insufficient evidence to reject H0 at the 95% level, hence it is reasonable to assume that the correlation is equal to zero.

1. **Hypothesis Test:**

Region: South-East Asia

Date: 01-06-2020

Previous Period: 22-01-2020 to 31-05-2020

Mean Recovery Rate for previous period: 0.3823964

H0: Mean Recovery Rate on 01-06-2020 is equal to mean deaths in previous period.

H1: Mean Recovery Rate on 01-06-2020 is not equal to mean deaths in previous period.

Performed t-test and got a p-value of 0.9168.

There is insufficient evidence to reject H0 at the 95% level, hence it is reasonable to assume that the mean recovery rate on 01-06-2020 is equal to the mean deaths of the previous period.

1. **Hypothesis Test:**

Region: South-East Asia

H0: Mean deaths in March 2020 is equal to mean deaths in June 2020.

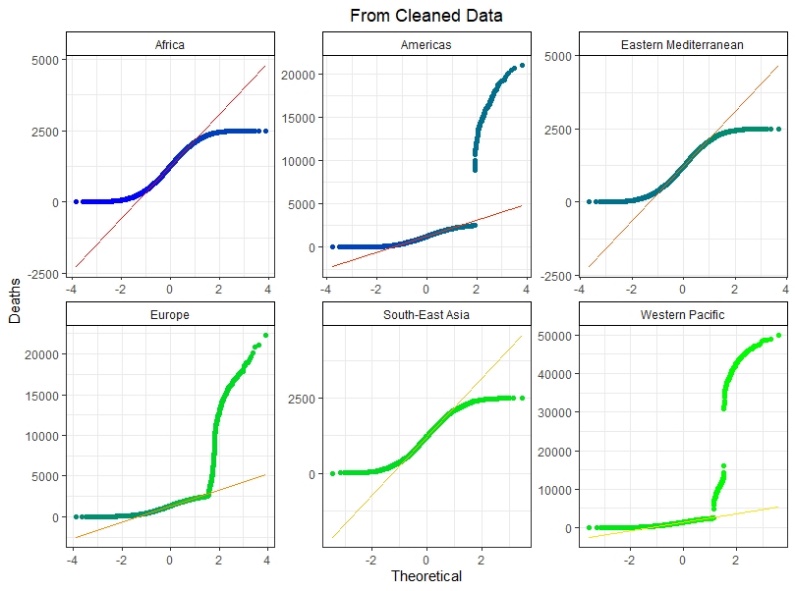
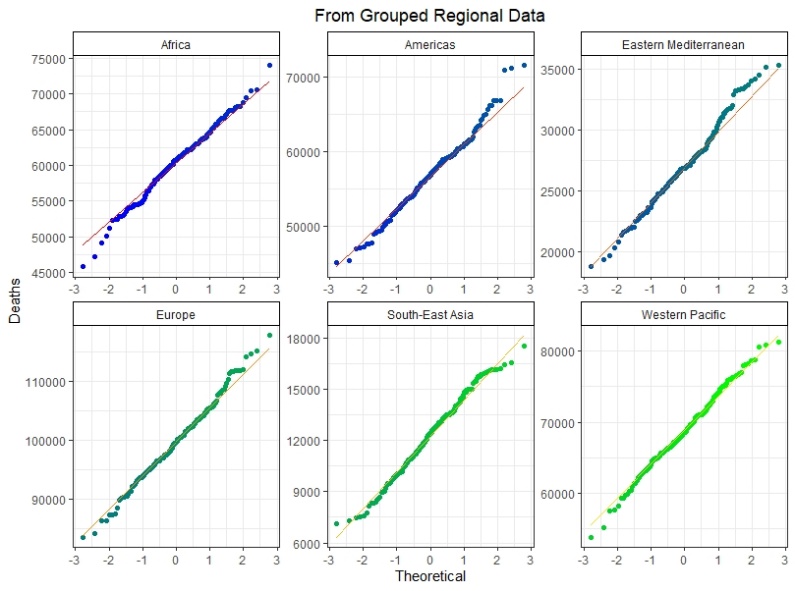
H1: Mean deaths in March 2020 is greater than mean deaths in June 2020.

Performed t-test and got a p-value of 0.7768.

There is insufficient evidence to reject H0 at the 95% level, hence it is reasonable to assume that the mean deaths in March 2020 is greater than the mean deaths in June 2020.

1. **Linear Model:**

From an initial analysis we see that the data grouped by region is more suitable for our model than the initial cleaned data.



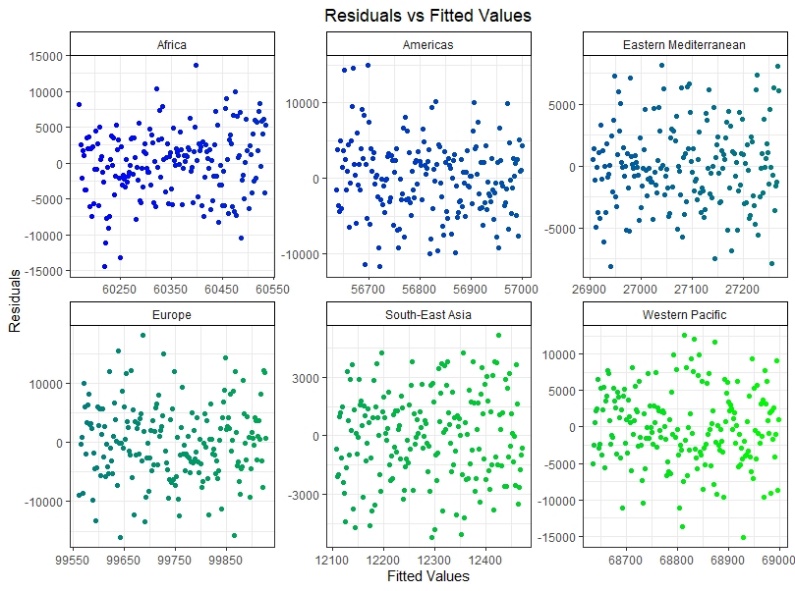
We can see from the above QQ-plots that the data grouped by region is more predictable and follows a Normal Distribution more closely.

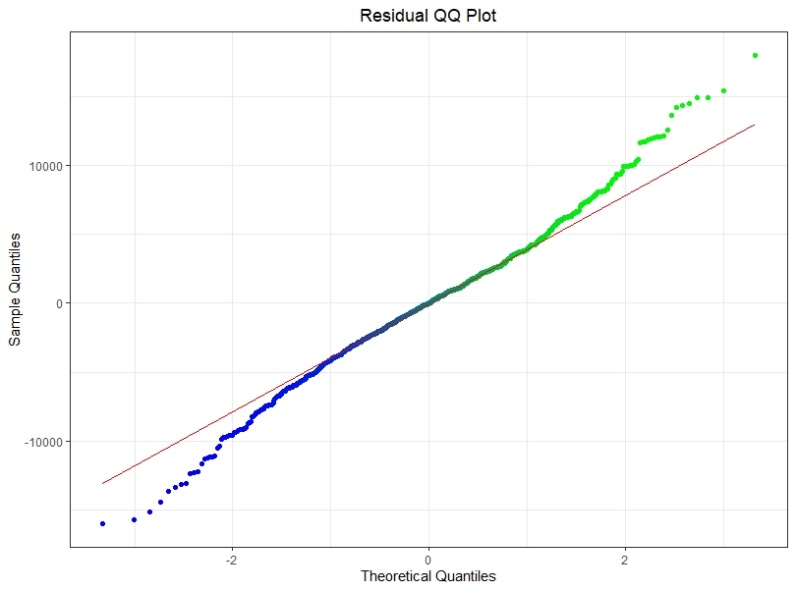
Model: Model considered is a linear regression model and requires Date and Region as input to make predictions.

R-squared = 97.5%

Predicted Deaths in August by Region:

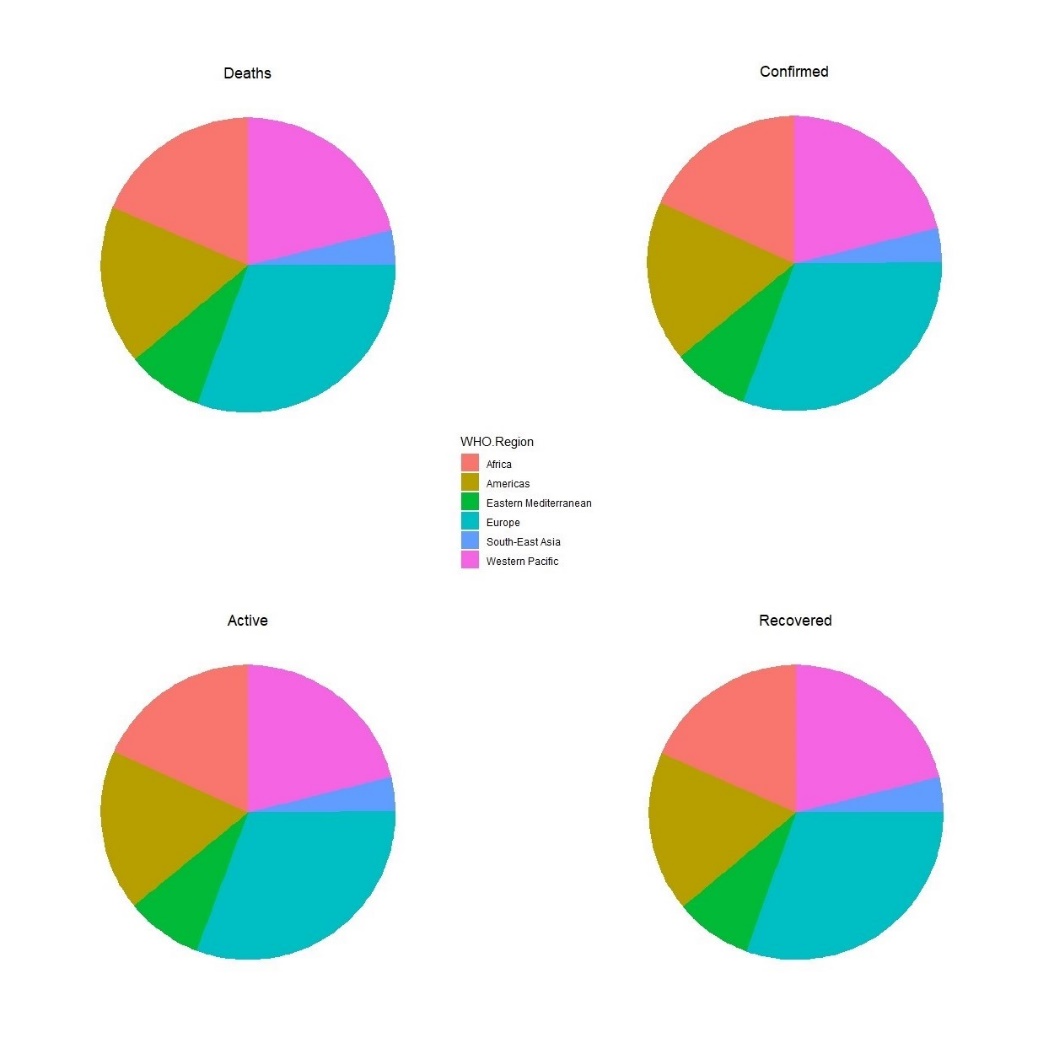
1. Africa: 1817149
2. Americas: 1711113
3. Eastern Mediterranean: 819171
4. Europe: 2998948
5. South-East Asia: 375274
6. Western Pacific: 2071064

Residual Analysis:



We can clearly see that there is no apparent pattern in any of the graphs above. Moreover, on analysis of the residual QQ Plot we see that the residuals approximately follow a normal distribution. Hence, we can conclude that the model is a good fit.

1. Pie Charts:



1. Analysis of variance for number of cases in May 2020 to August 2020:

* A similar model to the one in part 5. was used to predict the number of cases for August 2020.
* Data has been filtered for May 2020 from the regional dataset and ANOVA has been run on May 2020 vs August 2020.

Results of ANOVA:

|  |  |  |  |
| --- | --- | --- | --- |
| Source of Variation | Degrees of Freedom | Sum of Squares | Mean Sum of Squares |
| Regression | 1 | 4.7207e+15 | 4.7207e+15 |
| Residual | 178 | 1.5036e+14 | 8.4474e+11 |
| Total | 179 | 4.87106e+15 |  |
| p - value | | | 2.2e-16 ≈ 0 |

* Using the ANOVA table, we carry out a hypothesis test to check if there exists a linear relationship between the cases in May 2020 and August 2020.

H0: No linear relationship exists.

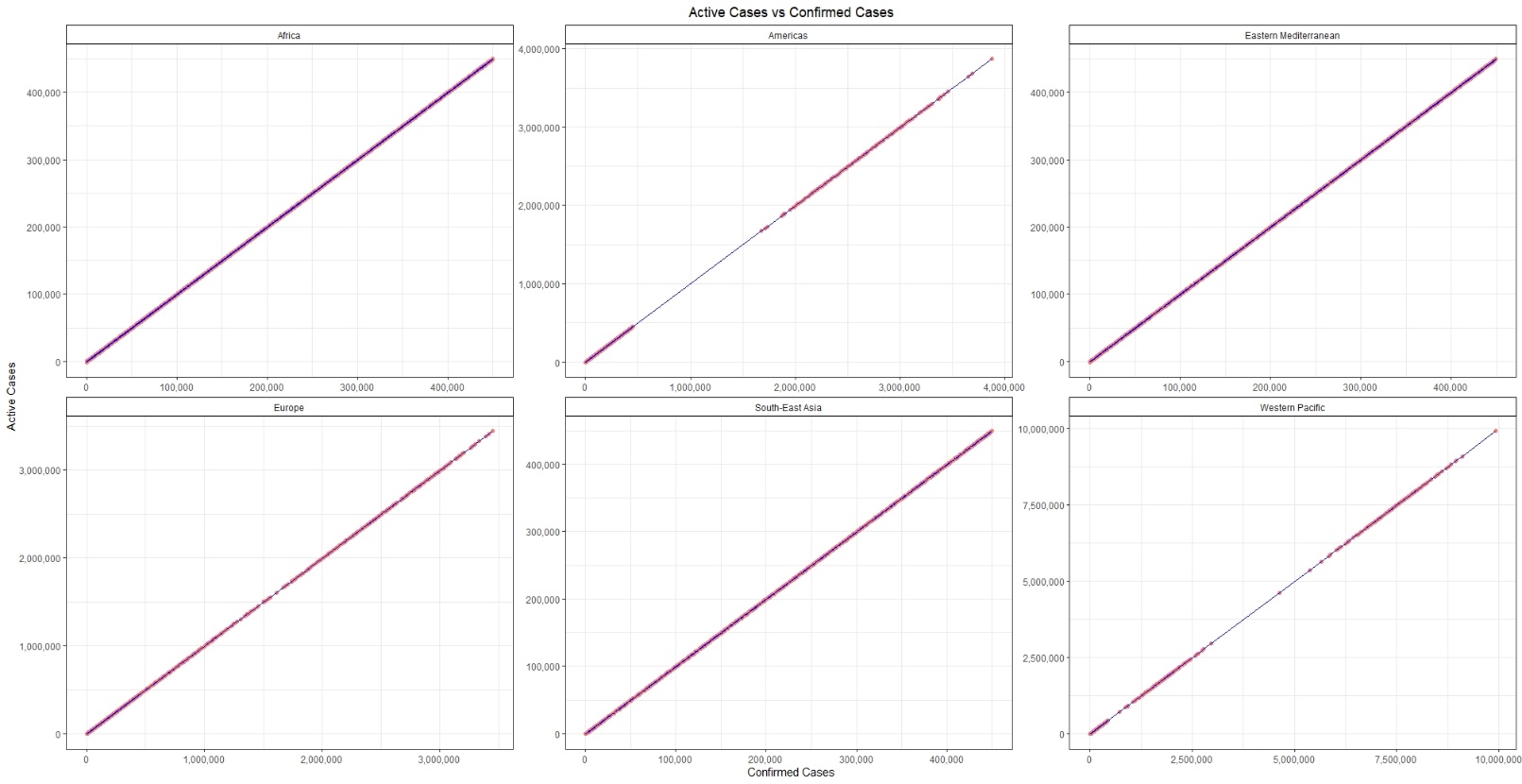
H1: Linear relationship exists.

We get a p-value that is close to 0. Hence, there is sufficient evidence to reject H0 at the 95% level and it is reasonable to assume that cases in May 2020 and August 2020 are linearly related.

1. Correlation between Confirmed and Active cases:

* Correlation between confirmed and active cases has been calculated for each region. The results are as follows:

|  |  |
| --- | --- |
| Region | Correlation (Confirmed vs Active) |
| Africa | 0.9999949 |
| Americas | 0.9999954 |
| Eastern Mediterranean | 0.9999964 |
| Europe | 0.9999963 |
| South-East Asia | 0.9999959 |
| Western Pacific | 0.9999975 |



* From the values and graphs we can see that there is strong positive correlation between confirmed and active cases for each region.

**PART III:**

Analysis of insurance data and preparing a suitable pricing model.

The colfunc functions from part II have been used in plotting of graphs in part 3 as well.

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

* All columns converted to their appropriate types:

(Age: numeric, sex: factor, BMI: numeric, children: numeric, smoker: factor,

region: factor)

* Wherever necessary, appropriate conversions of the above columns have been done to aid in the analysis.

1. Region specific sales targeting:

* Considered smoker as factor (1: No, 2: Yes).
* Took region wise mean of age and BMI, and sum of smokers and charges.
* Sorted the data in descending order of cases.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Region** | **Mean Age** | **Mean BMI** | **Total Smokers** | **Total Charges** |
| **southeast** | **38.939** | **33.355** | **455** | **5363689.763** |
| **northeast** | **39.268** | **29.173** | **391** | **4343668.583** |
| **northwest** | **39.196** | **29.199** | **383** | **4035711.997** |
| **southwest** | **39.455** | **30.596** | **383** | **4012754.648** |

Conclusion:

From the data we see that the southeast contributes the most to the total charges collected, has the highest mean BMI and highest number of smokers. Similarly, we could target regions based on the charges collected historically, number of smokers and average BMI. Given the data it would be reasonable to target the southeast.

1. Checking for charges based on smoking status:

* Data was filtered into smokers and non-smokers.
* The mean age and BMI and sum of charges was taken for smokers and non-smokers.
* Initial analysis was done, and it was found that non-smokers are responsible for more charges paid, but this could also be down to the greater number of non-smokers in the data set.

|  |  |  |  |
| --- | --- | --- | --- |
| Smoking Status | Mean Age | Mean BMI | Total Charges Paid |
| No | 39.4 | 30.7 | Rs. 8974061 |
| Yes | 38.5 | 30.7 | Rs. 8781764 |

* We perform a Hypothesis test to check if smokers should be charged more.

Number of smokers: 274

Number of non-smokers: 1064

Hypothesis test:

H0: Mean of smoker’s charges is equal to mean of non-smoker’s charges.

H1: Mean of smoker’s charges is greater than mean of non-smoker’s charges.

Test Results:

p-value: 0. 2.2e-16 ≈ 0

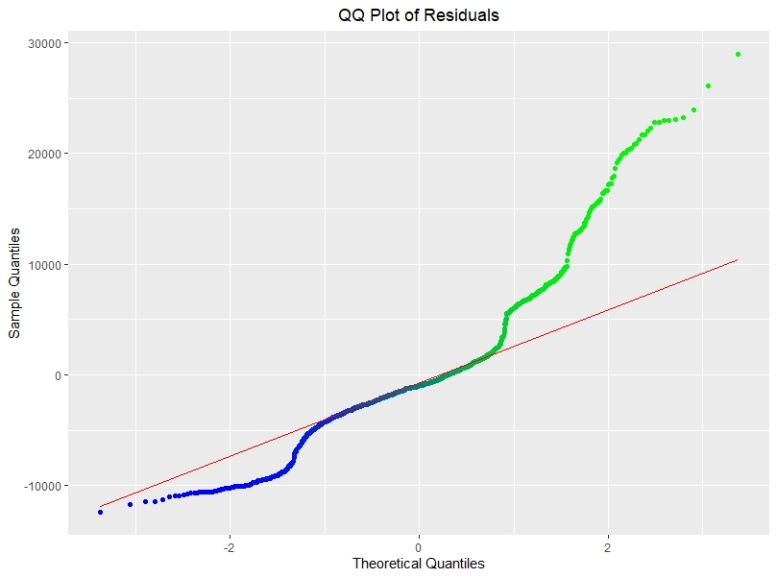
95% confidence interval: (22426.4, ∞)

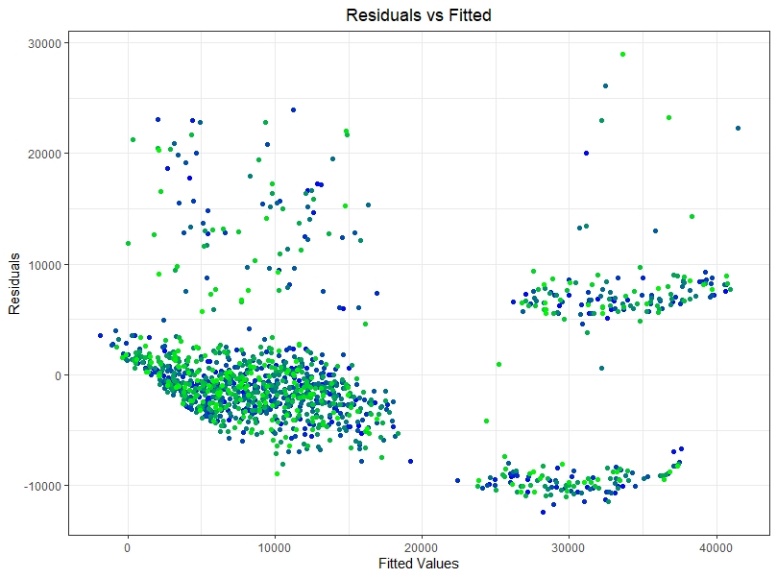
Conclusion: There is sufficient evidence to reject H0 at the 95% level, hence it is reasonable to assume that smokers are charged more than non-smokers.

It is advisable to charge smokers more.

1. Linear Model:

* Response Variable: Charges
* Predictor Variables: age, bmi, smoker
* Coefficient of Determination = 0.7469





* After analysing the plots above, we see that the residuals do not follow a normal distribution and the variance of residuals for lower fitted values is less than that of higher values. Hence, we can conclude that the model is not a good fit.

1. Correlation Tests:

All tests have been performed at the 95% level.

H0: True correlation is equal to zero.

H1: True correlation is not equal to zero.

1. Charges VS Age

Correlation: 0.2990082

p-value: 0. 2.2e-16 ≈ 0

Confidence Interval: (0.25, 0.35)

Conclusion: As the p - value is less than 0.05 we have enough evidence to reject H0 and conclude that there is correlation between Charges and age, and it is not equal to zero. Though as the value of correlation is only 0.3 there is only a weak correlation

1. Charges VS BMI

Correlation: 0.198341

p-value: 2.459e-13 ≈ 0

Confidence Interval: (0.15, 0.25)

Conclusion: As the p - value is less than 0.05 we have enough evidence to reject H0 and conclude that there is correlation between Charges and BMI, and it is not equal to zero. Though as the value of correlation is only 0.2 there is only a weak correlation.

1. Charges VS Children

Correlation: 0.06799823

p-value: 0.01285

Confidence Interval: (0.014, 0.121)

Conclusion: As the p - value is less than 0.05 we have enough evidence to reject H0 and conclude that there is correlation between Charges and Children, and it is not equal to zero. Though, as the value of correlation is 0.06 there is a very weak correlation.

1. Charges VS Smoker

Correlation: 0.7872514

p-value: 2.2e-16 ≈ 0

Confidence Interval: (0.76, 0.80)

Conclusion: As the p - value is less than 0.05 we have enough evidence to reject H0 and conclude that there is correlation between Charges and smoker, and it is not equal to zero. As the value of correlation is 0.8 there is a very strong correlation.

1. Charges VS Region

Correlation: -0.006208235

p-value: 0.8205

Confidence Interval: (-0.059, 0.047)

Conclusion: As the p - value is greater than 0.05 we do not have enough evidence to reject H0 and conclude that there is no correlation between Charges and region, and it is equal to zero.

1. Pricing Model:

* GLM was created using the model in part 3. as a base.
* Prepared the data for the model by grouping across the age, sex, smoker, and region columns and used log charges as response variable to increase predictability and normality of the data.

Model: Model considered is a GLM and requires Region, Age, Sex, BMI, and Smoking Status as inputs to make predictions.

Sample Predictions:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Region** | **Smoking Status** | **Sex** | **Age** | **BMI** | **Estimated Charge** |
| southeast | yes | female | 50 | 30 | 32,148.76 |
| northeast | yes | male | 35 | 23 | 22,571.84 |
| northwest | no | male | 40 | 27 | 6,273.17 |
| southwest | no | female | 19 | 37 | 3,932.39 |

Regional Effects:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Region** | **Smoking Status** | **Sex** | **Age** | **BMI** | **Estimated Charge** |
| southeast | yes | male | 35 | 25 | 24,903.38 |
| northeast | yes | male | 35 | 25 | 24,241.15 |
| northwest | yes | male | 35 | 25 | 24,100.34 |
| southwest | yes | male | 35 | 25 | 24,019.89 |