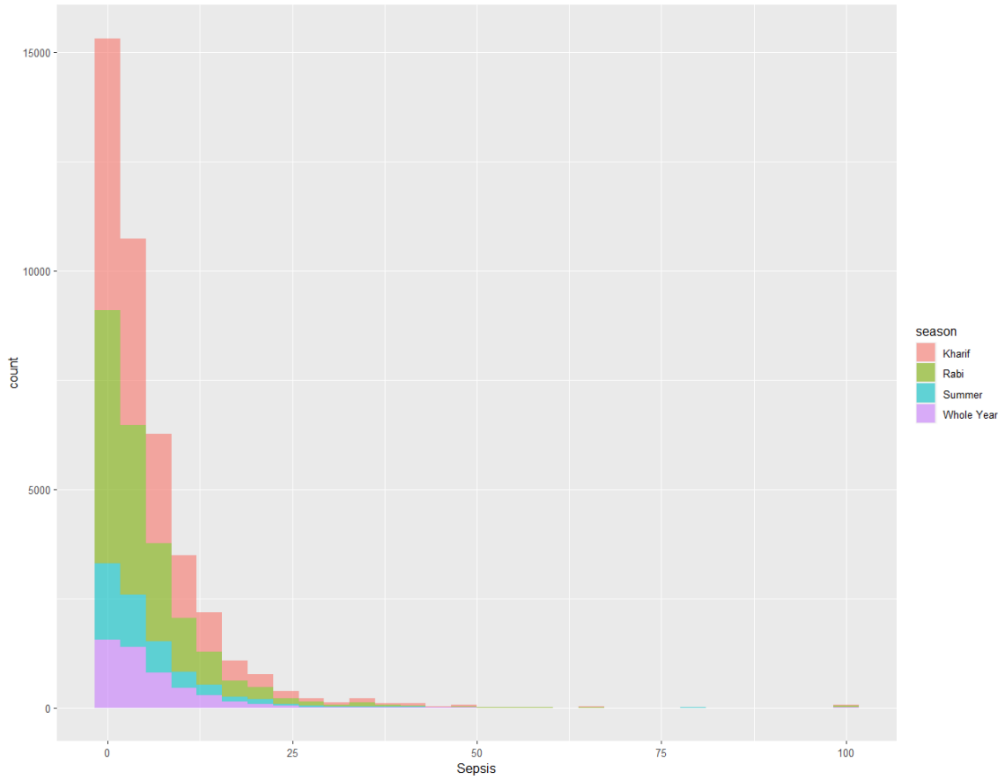
A1

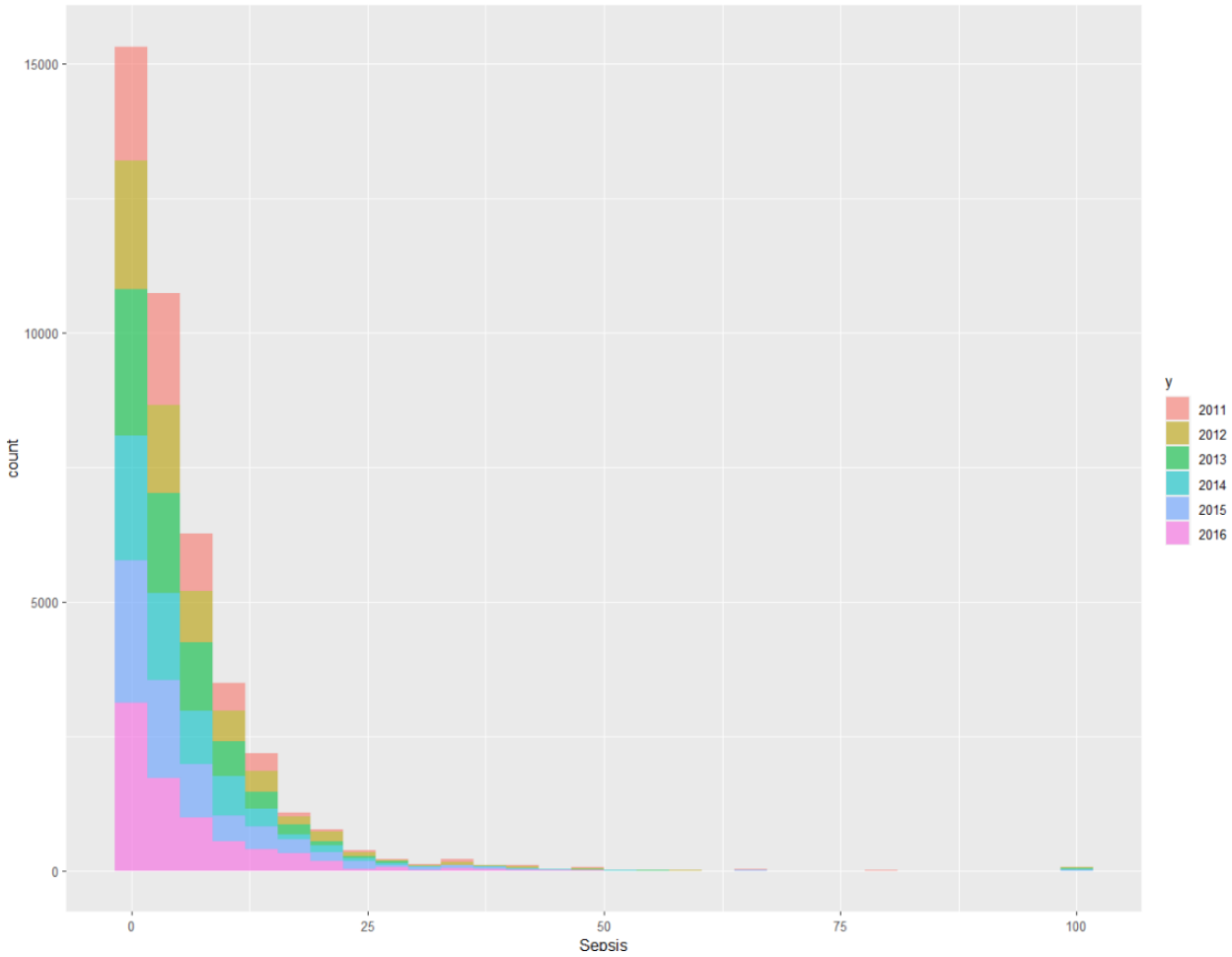
The way I made the dataset and my assumptions for the further questions are listed here

- The data set asked for was created by directly appending the required data as separate columns
- To make it easy for myself to conduct the analysis I removed all the unnecessary columns such as v1 to v39 and a few other miscellaneous columns such as Country.
- Found the mean, mode, median, standard deviation after removing duplicates but before removing outliers
- Made the histograms to find the outliers and then removed them using the mathematical relation of mean +- 3* standard deviation as this will remove the outliers without removing data that should be considered
- The models were run after removing outliers and duplicates from the variables on which the model is being run and not indiscriminately to ensure that repeating values and observations that are counted multiple times do not affect the result and skew it.
- Calculated the yield_indexes again because it was stated in the document to be year wise however the CSV file had it season wise

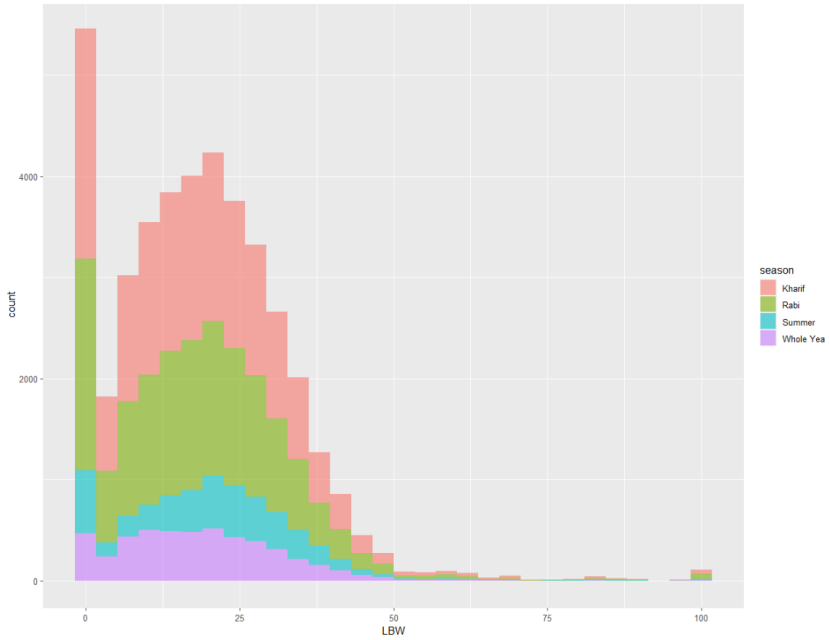
A2.

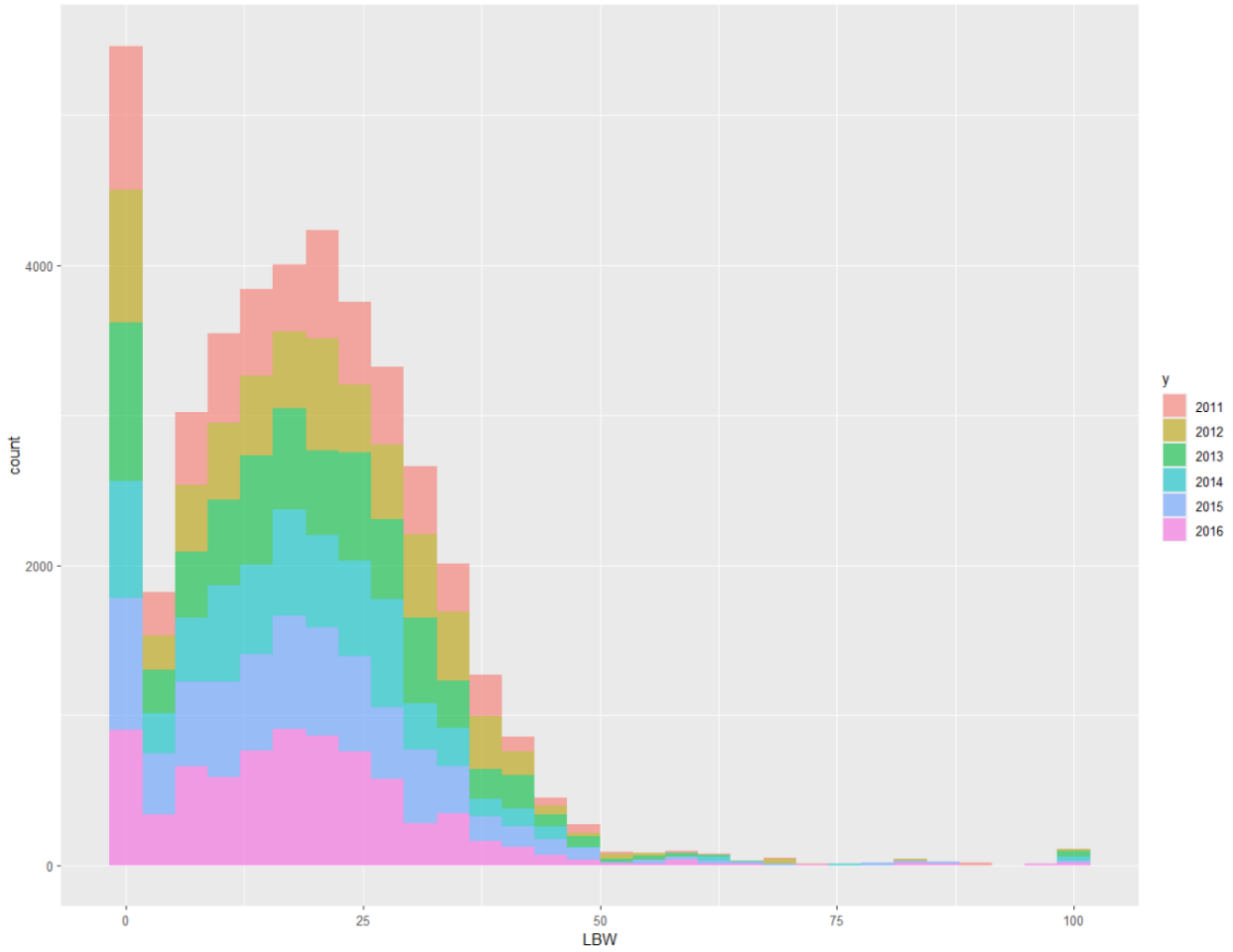
a. Sepsis



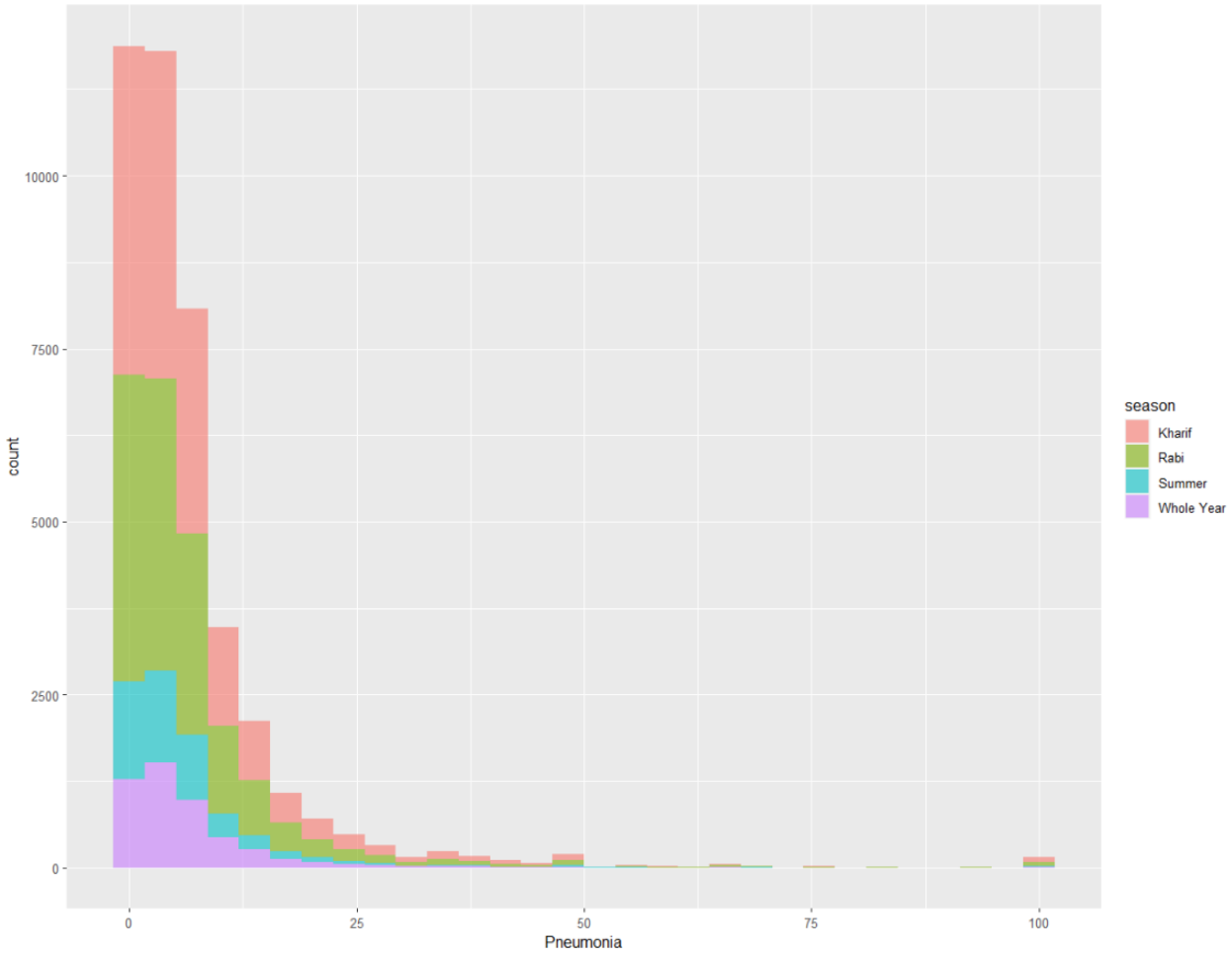

15000
10000
5000
0
count
0
25
50
75
100
Sepsis
y
2011
2012
2013
2014
2015
2016

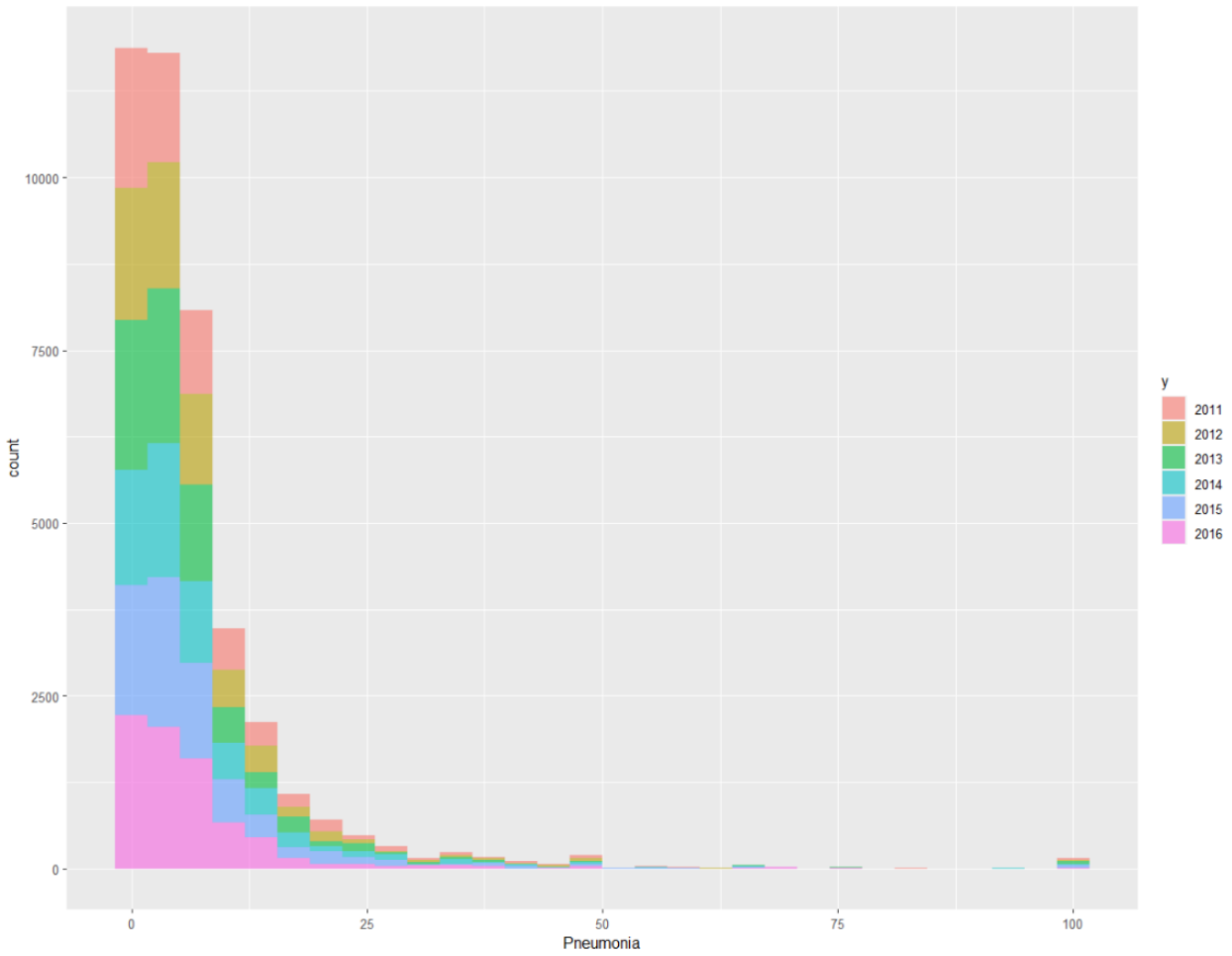
LBW


4000
2000
0
count
0
25
50
75
100
LBW
season
Kharif
Rabi
Summer
Whole Yea

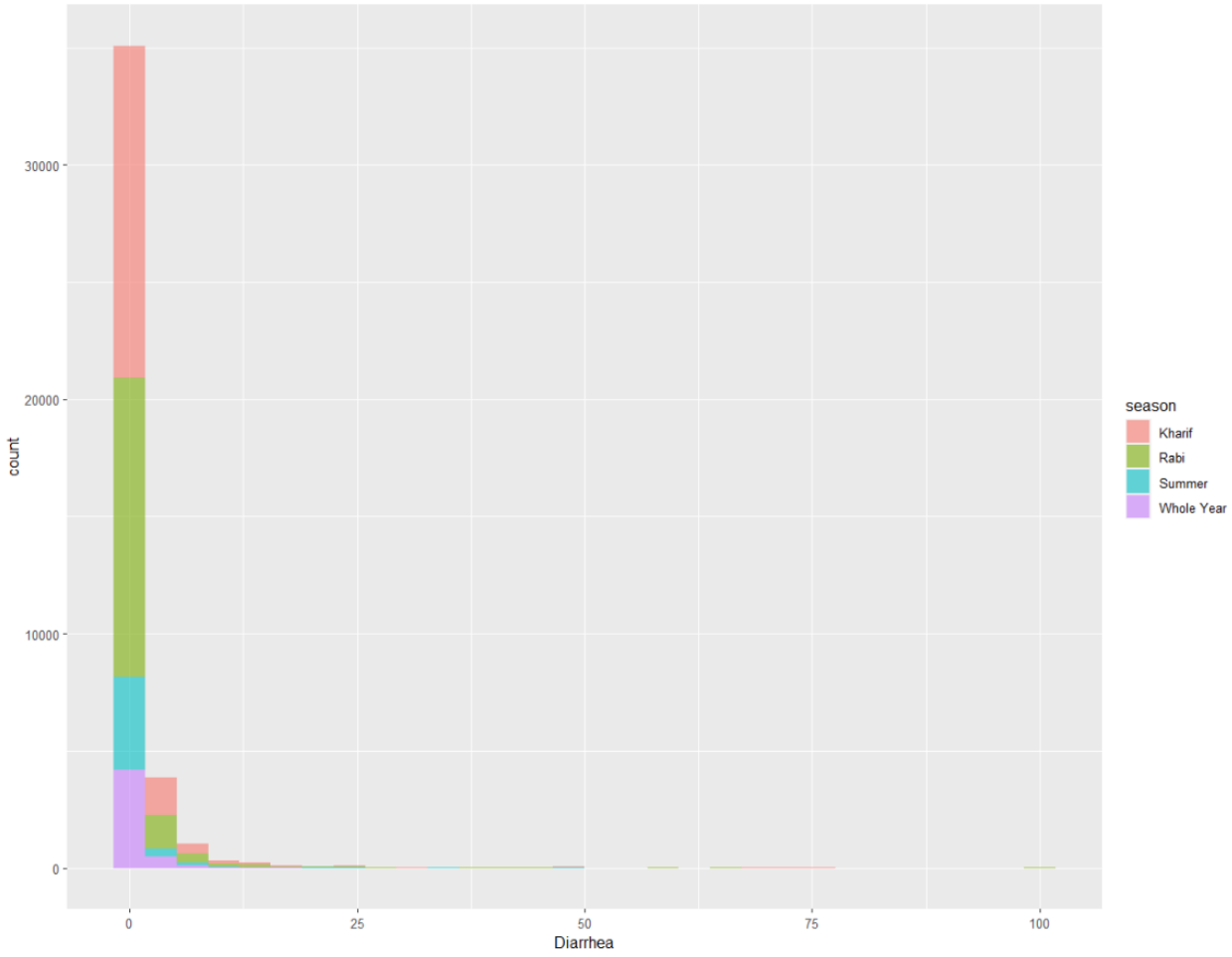

count
4000
2000
0
0
25
50
75
100
LBW
y
2011
2012
2013
2014
2015
2016

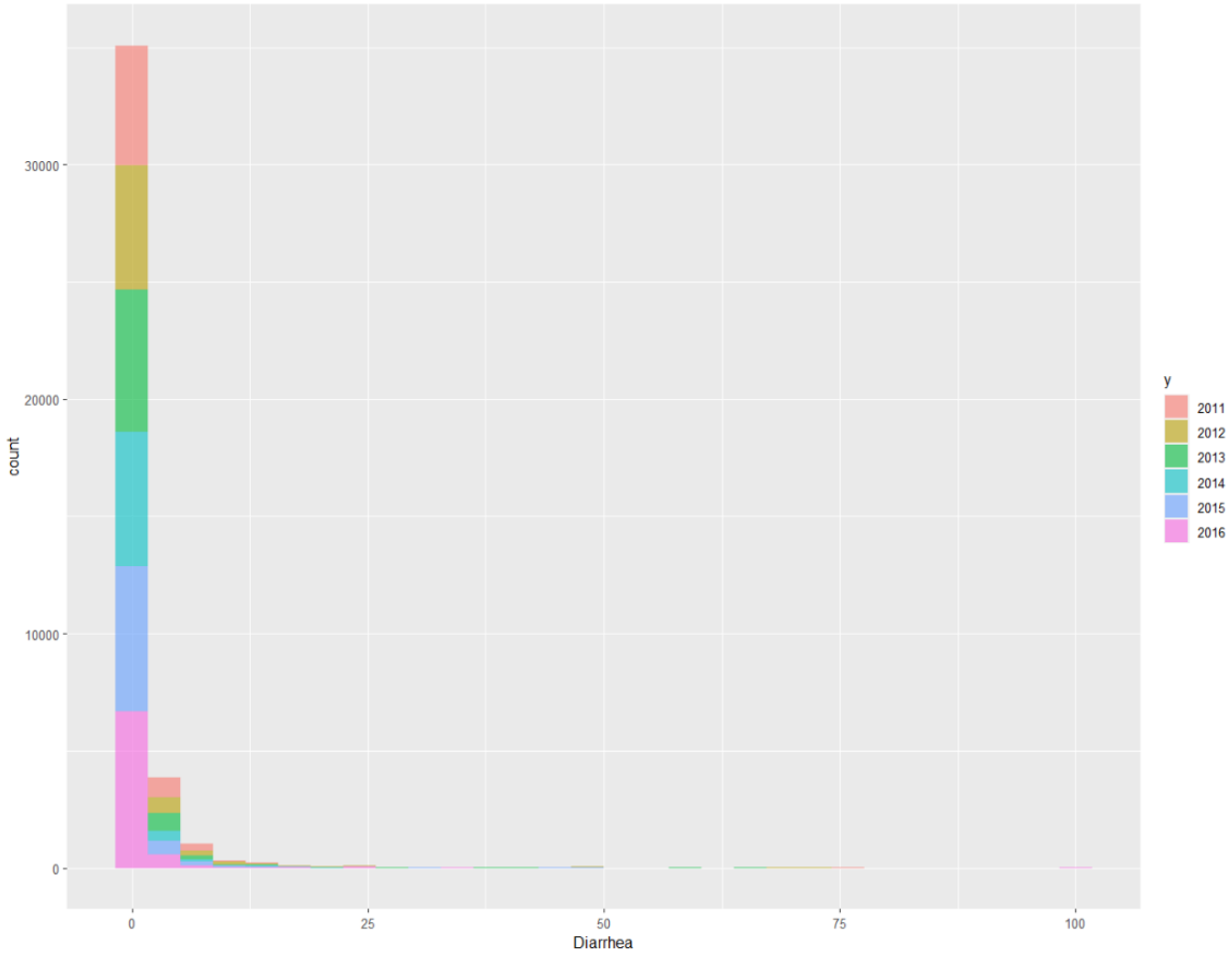
Pneumonia


10000
7500
5000
2500
0
count
0
25
50
75
100
Pneumonia
season
Kharif
Rabi
Summer
Whole Year

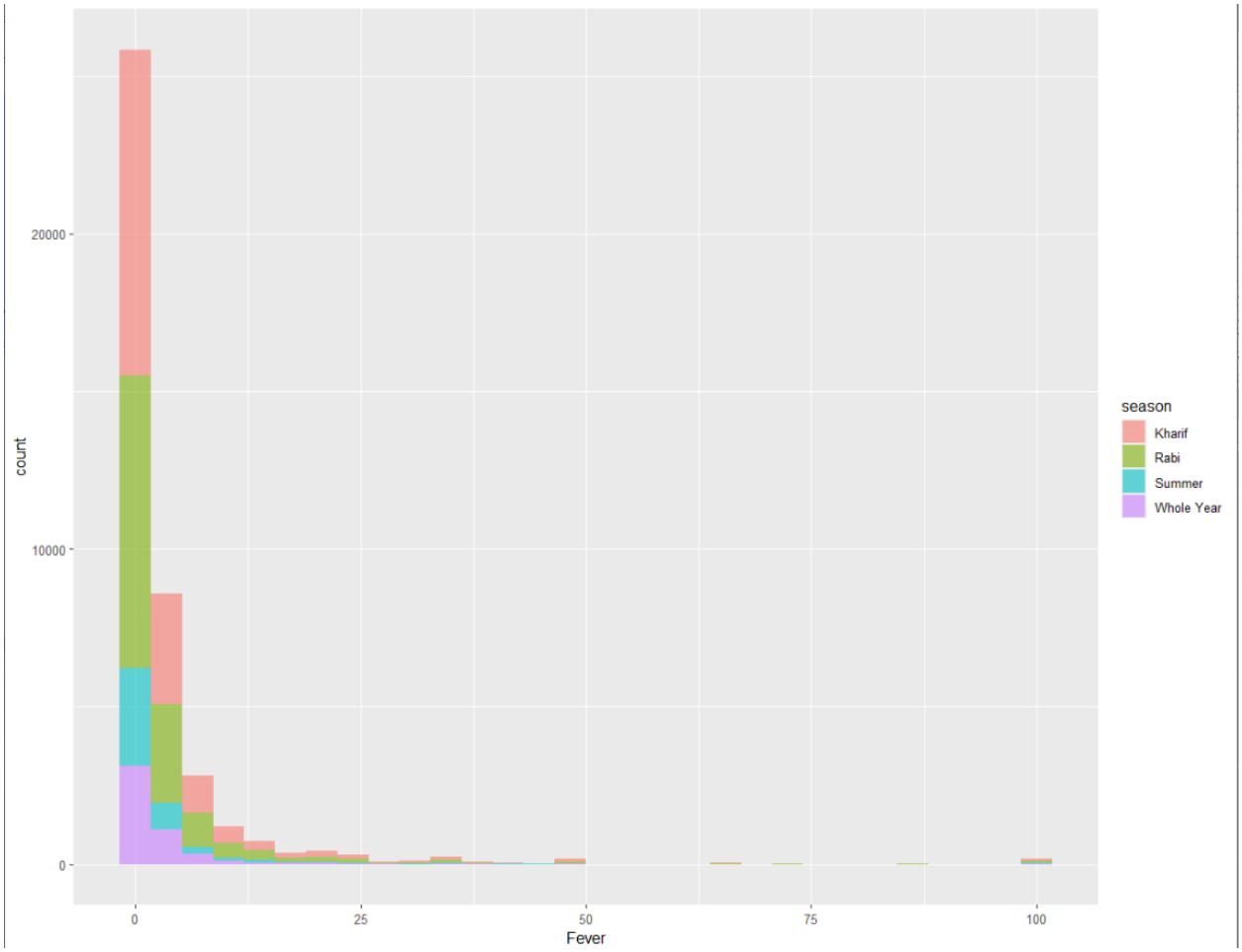

10000
7500
5000
2500
0
count
0
25
50
75
100
Pneumonia
y
2011
2012
2013
2014
2015
2016

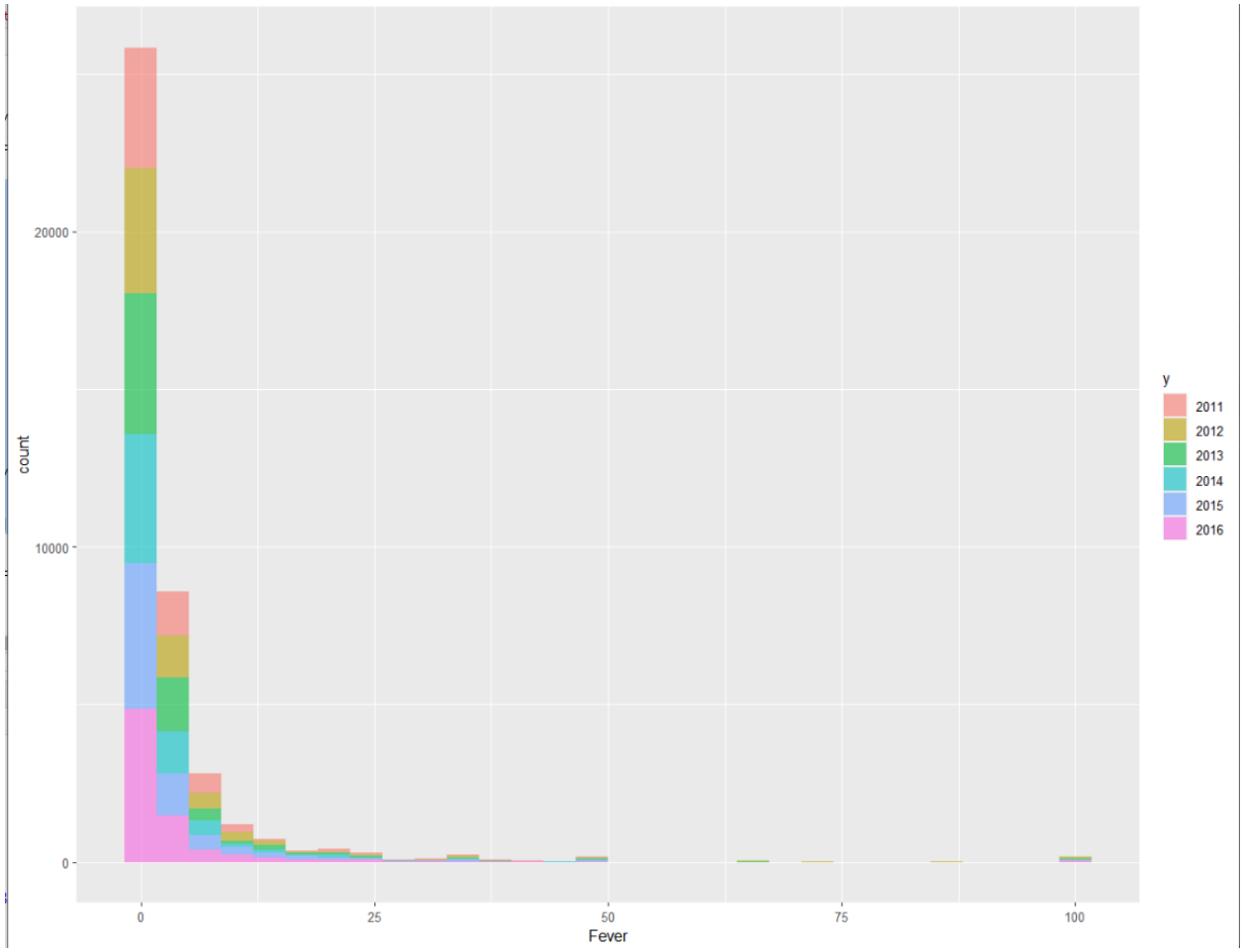
Diarrhea


30000
20000
10000
0
count
0
25
50
75
100
Diarrhea
season
Kharif
Rabi
Summer
Whole Year

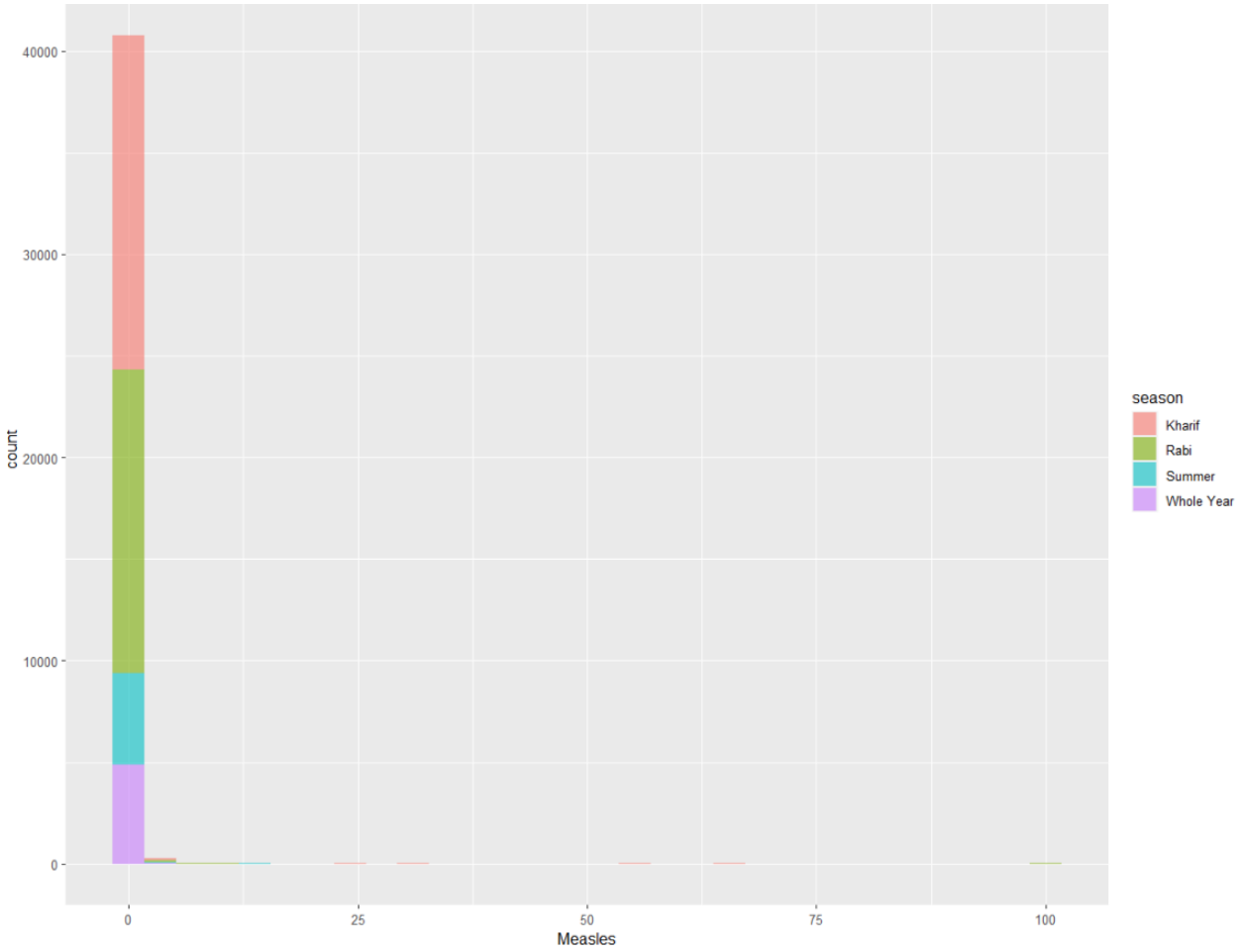

30000
20000
10000
0
count
0
25
50
75
100
Diarrhea
y
2011
2012
2013
2014
2015
2016

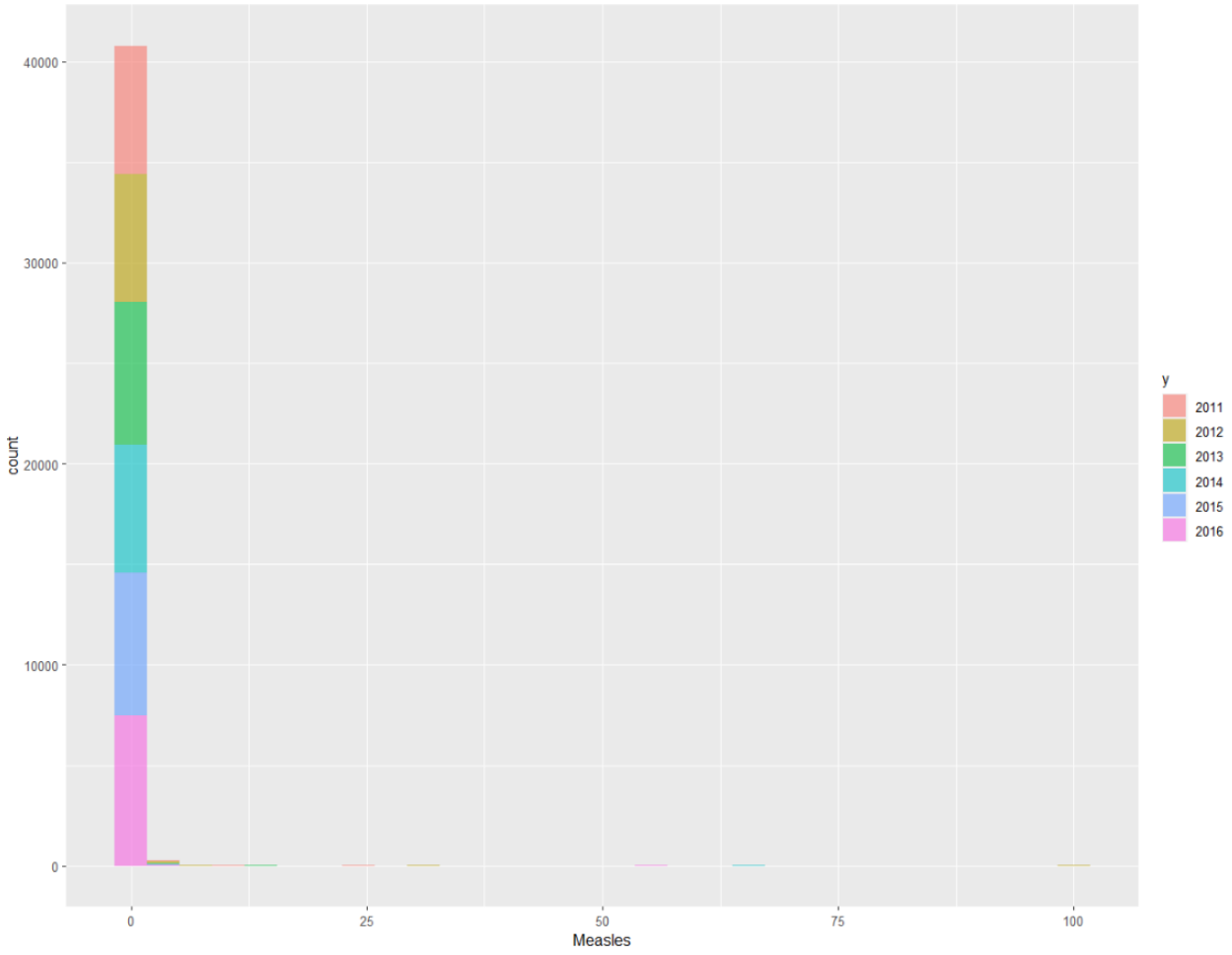
Fever


20000
10000
0
count
0
25
50
75
100
Fever
season
Kharif
Rabi
Summer
Whole Year


20000
10000
0
count
0
25
50
75
100
Fever
y
2011
2012
2013
2014
2015
2016

Measles


40000
30000
20000
10000
0
count
0
25
50
75
100
Measles
season
Kharif
Rabi
Summer
Whole Year



A2

A.. Sepsis

i.

```
> cor(unique_raw_sepsis$Sepsis,unique_raw_sepsis$gdp)
[1] 0.1119681
> cor(unique_raw_sepsis$Sepsis,unique_raw_sepsis$tap)
[1] -0.07472496
> cor(unique_raw_sepsis$Sepsis,unique_raw_sepsis$beds)
[1] 0.1077667
```

Ii

```
> cor(cropcash$Sepsis,cropcash$yield_index)
[1] 0.06581904
> cor(croppulse$Sepsis,croppulse$yield_index)
[1] -0.02525077
> cor(cropcereal$Sepsis,cropcereal$yield_index)
[1] 0.03681738
> cor(crophorti$Sepsis,crophorti$yield_index)
[1] -0.01836468
> cor(cropoil$Sepsis,cropoil$yield_index)
[1] 0.0166371
> cor(cropcoarse$Sepsis,cropcoarse$yield_index)
[1] 0.00104871
> cor(cropcash$Sepsis,cropcash$rate,use="complete.obs")
[1] 0.002869064
> cor(croppulse$Sepsis,croppulse$rate,use="complete.obs")
[1] -0.00827013
> cor(cropcereal$Sepsis,cropcereal$rate,use="complete.obs")
[1] -0.04322133
> cor(crophorti$Sepsis,crophorti$rate,use="complete.obs")
[1] -0.008203282
> cor(cropoil$Sepsis,cropoil$rate,use="complete.obs")
[1] -0.04212827
> cor(cropcoarse$Sepsis,cropcoarse$rate,use="complete.obs")
[1] -0.04158607
```

LBW

```
> cor(unique_raw_LBW$LBW,unique_raw_LBW$gdp)
[1] 0.2233036
> cor(unique_raw_LBW$LBW,unique_raw_LBW$tap)
[1] 0.1414683
> cor(unique_raw_LBW$LBW,unique_raw_LBW$beds)
[1] 0.04110667
```

```
> cor(cropcash$LBW,cropcash$yield_index)
[1] -0.1431475
> cor(croppulse$LBW,croppulse$yield_index)
[1] -0.09543898
> cor(cropcereal$LBW,cropcereal$yield_index)
[1] -0.1509078
> cor(crophorti$LBW,crophorti$yield_index)
[1] -0.09051156
> cor(cropoil$LBW,cropoil$yield_index)
[1] -0.04602501
> cor(cropcoarse$LBW,cropcoarse$yield_index)
[1] -0.1625562
> cor(cropcash$LBW,cropcash$rate,use="complete.obs")
[1] -0.003324827
> cor(croppulse$LBW,croppulse$rate,use="complete.obs")
[1] -9.850148e-05
> cor(cropcereal$LBW,cropcereal$rate,use="complete.obs")
[1] 0.004360697
> cor(crophorti$LBW,crophorti$rate,use="complete.obs")
[1] 0.02750287
> cor(cropoil$LBW,cropoil$rate,use="complete.obs")
[1] 0.04337418
> cor(cropcoarse$LBW,cropcoarse$rate,use="complete.obs")
[1] 0.04337534
>
```

Pneumonia

```
> cor(unique_raw_Pneumonia$Pneumonia,unique_raw_Pneumonia$gdp)
[1] -0.2429091
> cor(unique_raw_Pneumonia$Pneumonia,unique_raw_Pneumonia$tap)
[1] NA
> cor(unique_raw_Pneumonia$Pneumonia,unique_raw_Pneumonia$beds)
[1] -0.1755915
>
```

```
> cor(cropcash$Pneumonia,cropcash$yield_index)
[1] -0.06154182
> cor(croppulse$Pneumonia,croppulse$yield_index)
[1] -0.002307674
> cor(cropcereal$Pneumonia,cropcereal$yield_index)
[1] -0.07419651
> cor(crophorti$Pneumonia,crophorti$yield_index)
[1] -0.03273208
> cor(cropoil$Pneumonia,cropoil$yield_index)
[1] -0.06969377
> cor(cropcoarse$Pneumonia,cropcoarse$yield_index)
[1] -0.06027778
> cor(cropcash$Pneumonia,cropcash$rate,use="complete.obs")
[1] -0.003027755
> cor(croppulse$Pneumonia,croppulse$rate,use="complete.obs")
[1] 0.01856315
> cor(cropcereal$Pneumonia,cropcereal$rate,use="complete.obs")
[1] -0.0142572
> cor(crophorti$Pneumonia,crophorti$rate,use="complete.obs")
[1] 0.003857308
> cor(cropoil$Pneumonia,cropoil$rate,use="complete.obs")
[1] -0.02367379
> cor(cropcoarse$Pneumonia,cropcoarse$rate,use="complete.obs")
[1] -0.01380632
```

Diarrhea

```
> cor(unique_raw_Diarrhea$Diarrhea,unique_raw_Diarrhea$gdp)
[1] -0.1193406
> cor(raw_Diarrhea$Diarrhea,raw_Diarrhea$tap)
[1] -0.08330194
> cor(unique_raw_Diarrhea$Diarrhea,unique_raw_Diarrhea$beds)
[1] -0.07836866
```

```
> cor(cropcash$Diarrhea,cropcash$yield_index)
[1] 0.01208039
> cor(croppulse$Diarrhea,croppulse$yield_index)
[1] 0.02330668
> cor(cropcereal$Diarrhea,cropcereal$yield_index)
[1] -0.0127073
> cor(crophorti$Diarrhea,crophorti$yield_index)
[1] -0.008068404
> cor(cropoil$Diarrhea,cropoil$yield_index)
[1] -0.01800715
> cor(cropcoarse$Diarrhea,cropcoarse$yield_index)
[1] 0.007769893
> cor(cropcash$Diarrhea,cropcash$rate,use="complete.obs")
[1] 0.0294801
> cor(croppulse$Diarrhea,croppulse$rate,use="complete.obs")
[1] -0.009927684
> cor(cropcereal$Diarrhea,cropcereal$rate,use="complete.obs")
[1] 0.0005929961
> cor(crophorti$Diarrhea,crophorti$rate,use="complete.obs")
[1] -0.008565139
> cor(cropoil$Diarrhea,cropoil$rate,use="complete.obs")
[1] -0.003906181
> cor(cropcoarse$Diarrhea,cropcoarse$rate,use="complete.obs")
[1] -0.01752626
```

Fever

```
> cor(unique_raw_Fever$Fever,unique_raw_Fever$gdp)
[1] -0.2125461
> cor(raw_Fever$Fever,raw_Fever$tap)
[1] -0.1465661
> cor(unique_raw_Fever$Fever,unique_raw_Fever$beds)
[1] -0.1150073
```

```
>
> cor(cropcash$Fever,cropcash$yield_index)
[1] 0.01098647
> cor(croppulse$Fever,croppulse$yield_index)
[1] 0.01931528
> cor(cropcereal$Fever,cropcereal$yield_index)
[1] -0.0374173
> cor(crophorti$Fever,crophorti$yield_index)
[1] 0.01441371
> cor(cropoil$Fever,cropoil$yield_index)
[1] -0.05983226
> cor(cropcoarse$Fever,cropcoarse$yield_index)
[1] 0.03586155
> cor(cropcash$Fever,cropcash$rate,use="complete.obs")
[1] 0.05869889
> cor(croppulse$Fever,croppulse$rate,use="complete.obs")
[1] -0.00392629
> cor(cropcereal$Fever,cropcereal$rate,use="complete.obs")
[1] -0.02525541
> cor(crophorti$Fever,crophorti$rate,use="complete.obs")
[1] -0.01607291
> cor(cropoil$Fever,cropoil$rate,use="complete.obs")
[1] -0.004056805
> cor(cropcoarse$Fever,cropcoarse$rate,use="complete.obs")
[1] -0.01826271
```

Measles

```
> cor(unique_raw_Measles$Measles,unique_raw_Measles$gdp)
[1] -0.01006789
> cor(raw_Measles$Measles,raw_Measles$tap)
[1] -0.02987639
> cor(unique_raw_Measles$Measles,unique_raw_Measles$beds)
[1] -0.02616381
```

```
> cor(cropcash$Measles,cropcash$yield_index)
[1] 0.00623137
> cor(croppulse$Measles,croppulse$yield_index)
[1] 0.01586359
> cor(cropcereal$Measles,cropcereal$yield_index)
[1] 0.0004121292
> cor(crophorti$Measles,crophorti$yield_index)
[1] 0.03860467
> cor(cropoil$Measles,cropoil$yield_index)
[1] -0.004670847
> cor(cropcoarse$Measles,cropcoarse$yield_index)
[1] 0.01538848
> cor(cropcash$Measles,cropcash$rate,use="complete.obs")
[1] 0.006092321
> cor(croppulse$Measles,croppulse$rate,use="complete.obs")
[1] -0.00643816
> cor(cropcereal$Measles,cropcereal$rate,use="complete.obs")
[1] -0.003862562
> cor(crophorti$Measles,crophorti$rate,use="complete.obs")
[1] -0.004048036
> cor(cropoil$Measles,cropoil$rate,use="complete.obs")
[1] -0.005537327
> cor(cropcoarse$Measles,cropcoarse$rate,use="complete.obs")
[1] -0.02478619
```

A3

All of the models were run using LBW as our health indicator

A)

Model 1

Dependent Variables	OLS estimates
Intercept	1.369e+01
GDP	1.773e-07
Beds	-5.655e-05
Tap	2.147e-02
N = 3559	R^2 = 0.1102

B)

Model 2

Dependent Variables	OLS estimates of coefficient

Intercept	1.539e+01
GDP	1.679e-07
Beds	-5.187e-05
Tap	1.739e-02
Yield Index for Cash	-5.157e-02
N = 3091	R^2 = 0.125

Model 3

Dependent Variables	OLS estimates of coefficient
Intercept	1.540e+01
GDP	1.751e-07
Beds	-5.886e-05
Tap	2.894e-02
Yield Index for Pulses	-1.838e+00
N = 3434	R^2 = 0.1253

Model 4

Dependent Variables	OLS estimates of coefficient
Intercept	1.795e+01
GDP	1.737e-07
Beds	-5.192e-05
Tap	4.004e-02
Yield Index for Cereal	-2.014e+00
N = 3546	R^2 = 0.1427

Model 5

Dependent Variables	OLS estimates of coefficient
Intercept	1.475e+01
GDP	1.691e-07
Beds	-5.308e-05
Tap	2.837e-02
Yield Index for Horticulture	-1.375e-01
N = 3334	R^2 = 0.1125

Model 6

Dependent Variables	OLS estimates of coefficient
Intercept	1.467e+01
GDP	1.838e-07
Beds	-5.928e-05
Tap	2.817e-02
Yield Index for Oilseeds	-1.045e+00
N = 3389	R^2 = 0.1255

Model 7

Dependent Variables	OLS estimates of coefficient
Intercept	1.699e+01
GDP	1.501e-07
Beds	-4.704e-05
Tap	5.868e-02
Yield Index for Coarse Cereals	-2.821e+00

N = 2951	R^2 = 0.1441

Model 8

Dependent Variables	OLS estimates of coefficient
Intercept	1.541e+01
GDP	1.709e-07
Beds	-5.444e-05
Tap	3.139e-02
Yield Index for Cash	-5.438e-02
Yield Index for Pulse	-2.012e+00
Yield Index for Cereals	-9.591e-01
Yield Index for Coarse Cereals	-1.735e+00
Yield Index for Horticulture	-1.829e-01
Yield Index for Oilseeds	-1.294e+00
N = 19745	R^2 =

rate = Growth Rate

Model 9

Dependent Variables	OLS estimates of coefficient
Intercept	1.459e+01
GDP	1.623e-07
Beds	-5.417e-05
Tap	1.633e-02
Yield Index rate for Cash	-2.291e-02
N = 3091	R^2 = 0.1101

Model 10

Dependent Variables	OLS estimates of coefficient
Intercept	1.423e+01
GDP	1.660e-07
Beds	-5.348e-05
Tap	1.737e-02
Yield Index rate for Pulses	-7.852e-02
N = 3434	R^2 = 0.1157

Model 11

Dependent Variables	OLS estimates of coefficient
Intercept	1.418e+01
GDP	1.660e-07
Beds	-5.282e-05
Tap	1.579e-02
Yield Index rate for Cereals	-1.669e-01
N = 3546	R^2 = 0.1095

Model 14

Dependent Variables	OLS estimates of coefficient
Intercept	1.496e+01
GDP	1.438e-07
Beds	-5.004e-05
Tap	3.583e-02
Yield Index rate for Coarse Cereals	4.121e-01

N = 2951	R^2 = 0.1094

Model 12

Dependent Variables	OLS estimates of coefficient
Intercept	1.427e+01
GDP	1.623e-07
Beds	-5.181e-05
Tap	2.352e-02
Yield Index rate for Horticulture	-2.910e-03
N = 3334	R^2 = 0.09793

Model 13

Dependent Variables	OLS estimates of coefficient
Intercept	1.429e+01
GDP	1.649e-07
Beds	-5.347e-05
Tap	1.572e-02
Yield Index rate for Oilseeds	1.206e-01
N = 3389	R^2 = 0.1137

E)

Model 15

Dependent Variables	OLS estimates of coefficient
Intercept	1.439e+01
GDP	1.612e-07
Beds	-5.265e-05
Tap	2.039e-02

Yield Index rate for Cash	-2.265e-02
Yield Index rate for Pulse	-8.472e-02
Yield Index rate for Cereals	-1.904e-01
Yield Index rate for Coarse Cereals	4.465e-01
Yield Index rate for Horticulture	-2.110e-03
Yield Index rate for Oilseeds	1.120e-01
N = 19745	R^2 =0.1101

F)

Model 16

Dependent Variables	OLS estimates of coefficient
Intercept	-69.0046
log(GDP)	9.1750
log(Beds)	-5.57135
log(Tap)	0.53670
log(Yield Index Horticulture)	-1.57788
N = 3036	R^2 = 0.1567

Model 17

Dependent Variables	OLS estimates of coefficient
Intercept	-73.1105
log(GDP)	9.1456
log(Beds)	-6.0845
log(Tap)	0.7431
log(Yield Index Cereals)	-3.7133
N = 3470	R^2 =0.1513

Dependent Variables	OLS estimates of coefficient
Intercept	-69.2988
log(GDP)	8.8803
log(Beds)	-6.1400
log(Tap)	0.6062
log(Yield Index Horticulture)	-0.8981
N = 3260	R^2 =0.1388

Dependent Variables	OLS estimates of coefficient
Intercept	-72.2397
log(GDP)	9.2568
log(Beds)	-6.6453
log(Tap)	0.6834
log(Yield Index Pulse)	-1.4477
N = 3370	R^2 =0.1322

Dependent Variables	OLS estimates of coefficient
Intercept	-81.0625
log(GDP)	9.9539
log(Beds)	-6.9252
log(Tap)	0.7473
log(Yield Index Oil)	-2.3568
N = 3329	R^2 =0.1501

Dependent Variables	OLS estimates of coefficient
Intercept	-72.4804
log(GDP)	8.8602
log(Beds)	-6.0103
log(Tap)	1.0328
log(Yield Index Coarse Cereals)	-3.2318
N = 2894	R^2 =0.1654

G)

Dependent Variables	OLS estimates of coefficient
Intercept	-73.23660
log(GDP)	9.26520
log(Beds)	-6.47390
log(Tap)	0.72081
log(Yield Index for Cash)	-0.49494
log(Yield Index for Pulse)	-0.72430

log(Yield Index for Cereals)	-1.70435
log(Yield Index for Coarse Cereals)	-3.04307
log(Yield Index for Horticulture)	-0.64058
log(Yield Index for Oilseeds)	-2.08302
N = 19359	R^2 = 0.1427

Ans 4

We know that there exists a relation between the goodness of fit(R^2) and correlation coefficient(r) which is usually r= root(R^2) however in this case since we have multiple independent variables the conclusion that we reach is that theoretically the sums of the squares of these correlation coefficients should be somewhat similar to the goodness of fit since each variable will be explaining some x% of the variation in the dependent variable

Let us take model 2, if we sum up the correlation coefficients we get in the manner described above it will be 0.5 , this is quite different from our goodness of fit and which implies that the model is not good at explaining the relationship that we have proposed, this can be due to a few reasons chief among them that the independent variables in question such as gdp, beds, tap are not truly independent variables , they have a high correlation coefficient which makes sense as well since a state and by extension a district having higher gdp would be able to buy more taps and beds which would reduce the health indicator's value.

Even if in some case we see that either the sum of the correlation coefficients or their squares are close to the goodness of fit it does not mean that the model is accurate it might just be a sampling bias, this can be seen even more clearly in part C where all the yield indexes were taken together and it is quite apparent that the proposed relation does not hold. So it is important to critically analyse the results we get from the lens of reality to find out what the model might have missed and how we can rectify it.

Ans 5

The problem with including yield indexes for all the crop categories in one model is that we are unable to accurately find out how a change in one of the variables individually affects the dependent variable ie the health indicator in this case, since different crop categories may vary in their yields across districts and even for the same district there can be multiple crop categories , taking all the indexes in one model distorts the analysis as each crop category may not be perfectly independent from each other.

Another factor that distorts the analysis is that different crop categories may only grow in certain districts which would affect the results for the districts where they do not grow.

Ans 6

Yes, the relation between yield growth rates and health indicators across crop categories is quite similar to an extent, the OLS slope estimators for Oilseeds, Cereals, Coarse Cereals share the same magnitude as well as the estimators for Pulses and Cash crops with horticulture being the only odd one out and all of them are quite close to each other in terms of their magnitudes differing at most by a power of 2.

This also kind of lines up with real life circumstances as well since infant low birth weight is partially caused by poor prenatal nutrition of the mother and since oilseeds and cereals in general are quite lean foods it makes sense that all of them will share some similarity in how they affect the health indicator. Cash crops and pulses share similarities in their effect on the health indicator as pulses and cash crops are both more expensive to cultivate and buy which would indicate that wealthy families would be able to afford these and hence they share some similarity in their effect on infant mortality due to low birth weight

Another factor that can possibly explain these trends is that cereals are