1. Answer:

1.1 The number of variables with data is 19. The number of rows that contain data is 19371.

1.2 How many products are there in the data? 2149

How many organizations sell products in the data? 229

Which organization sells the most products in the data? UnitedHealthcare

1.3 How many duplicates are there? 74

What's different among the observations? They belong to different organization types or plan types or benefit types.

Drop all observations that are duplicated in all columns and report the remaining number of duplicates and lines dropped. 19297 lines remain. 74 duplicates and lines are dropped.

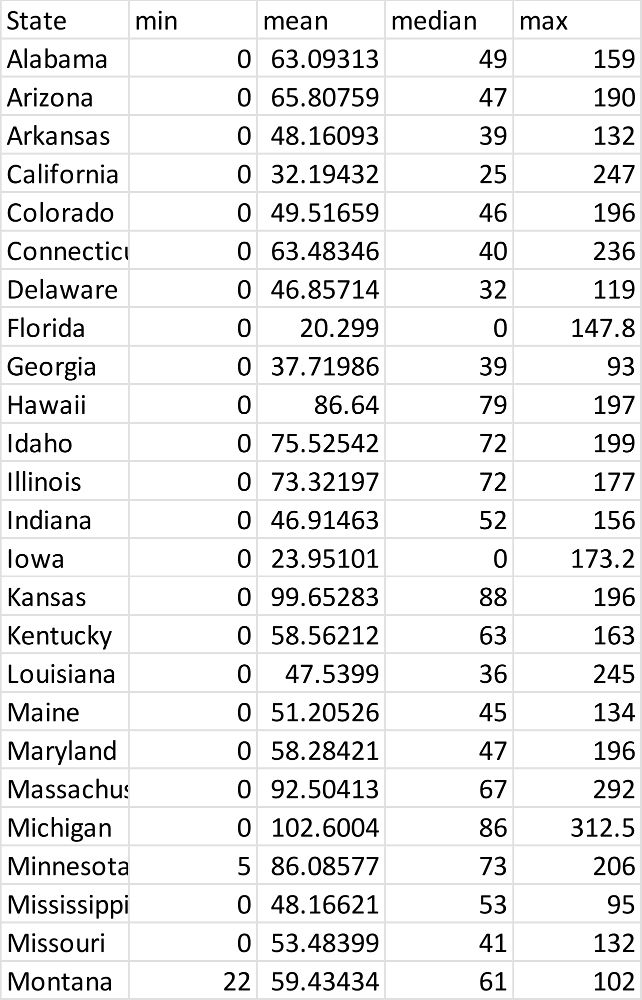
1.4 How many lines were removed? 594

How many lines were removed in total? 5396

1.5 The final dataset sees DataFrame "premium2".

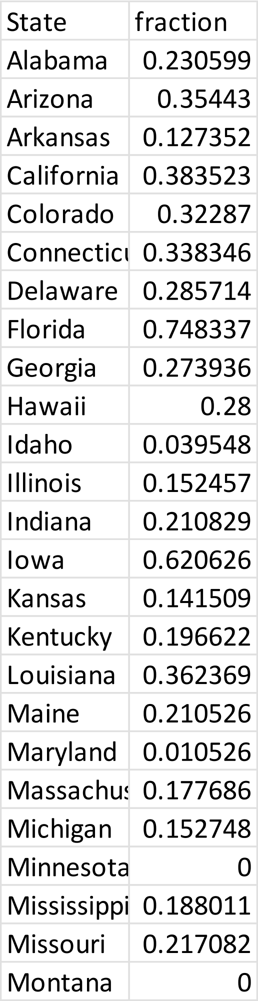
1.6 answers:

1.6.1 By state, what is the min, mean, median, max of total premium?



State	min	mean	median	max
Alabama	0	63.09313	49	159
Arizona	0	65.80759	47	190
Arkansas	0	48.16093	39	132
California	0	32.19432	25	247
Colorado	0	49.51659	46	196
Connecticu	0	63.48346	40	236
Delaware	0	46.85714	32	119
Florida	0	20.299	0	147.8
Georgia	0	37.71986	39	93
Hawaii	0	86.64	79	197
Idaho	0	75.52542	72	199
Illinois	0	73.32197	72	177
Indiana	0	46.91463	52	156
Iowa	0	23.95101	0	173.2
Kansas	0	99.65283	88	196
Kentucky	0	58.56212	63	163
Louisiana	0	47.5399	36	245
Maine	0	51.20526	45	134
Maryland	0	58.28421	47	196
Massachus	0	92.50413	67	292
Michigan	0	102.6004	86	312.5
Minnesota	5	86.08577	73	206
Mississippi	0	48.16621	53	95
Missouri	0	53.48399	41	132
Montana	22	59.43434	61	102

What fraction of plans are offered to consumers at zero total premium?



State	fraction
Alabama	0.230599
Arizona	0.35443
Arkansas	0.127352
California	0.383523
Colorado	0.32287
Connecticu	0.338346
Delaware	0.285714
Florida	0.748337
Georgia	0.273936
Hawaii	0.28
Idaho	0.039548
Illinois	0.152457
Indiana	0.210829
Iowa	0.620626
Kansas	0.141509
Kentucky	0.196622
Louisiana	0.362369
Maine	0.210526
Maryland	0.010526
Massachus	0.177686
Michigan	0.152748
Minnesota	0
Mississippi	0.188011
Missouri	0.217082
Montana	0

1.6.2 Which state has the most expensive plans on average? Michigan
and which is the cheapest? Florida
Are there states where no plan is offered at zero premium? Montana & Minnesota

1.6.3 Which plan types are the most expensive on average? PFFS
Which organization sells the most expensive plan? Vantage Health Plan, Inc.

1.7 Stored data sees ‘premium2.csv’.

2. Answers:
   2.1. How many observations are there? 2304216
   What information do the columns contain? They contain contract ID, plan ID, state, county, and number of enrollment, also the SSA State County Code and FIPS State County Code.
   2.2. Drop these and report the change. Observation changes from 2304216 to 2298189.
   2.3. how many individuals enrolled in MA plans in the country in 2018? 46300084
   What is the state with the largest enrollment? CA
   2.4. How many duplicates are there? 805
   Aggregate enrollment across the duplicated rows, which present partitions of certain states. See DataFrame ‘temp’.
   2.5. See DataFrame ‘temp’.
   2.6. Stored data sees ‘enroll2.csv’.
3. Answers:

3.1. See DataFrame ‘premium_new’.

3.2. What fraction of observations were deleted? 0.48183956966260844

3.3. What is the match rate? 0.010293561541632457

3.4. What fraction of the price data has any match? 0.9991732804232805

What is different about the unmatched price data? They have different State-county combinations.

Are their states and counties in the enrollment data? No

3.5. (extra) Can you match the remaining ones? Yes. The reason why they cannot match is because of the inconsistency in naming. If we replace "California,Los Angeles (Partial)" & "Arizona ,Pinal (Partial)" with "California,Los Angeles" & "Arizona ,Pinal", and replace "LaSalle" with "La Salle", we will match the remaining ones. For details see code 3.e.

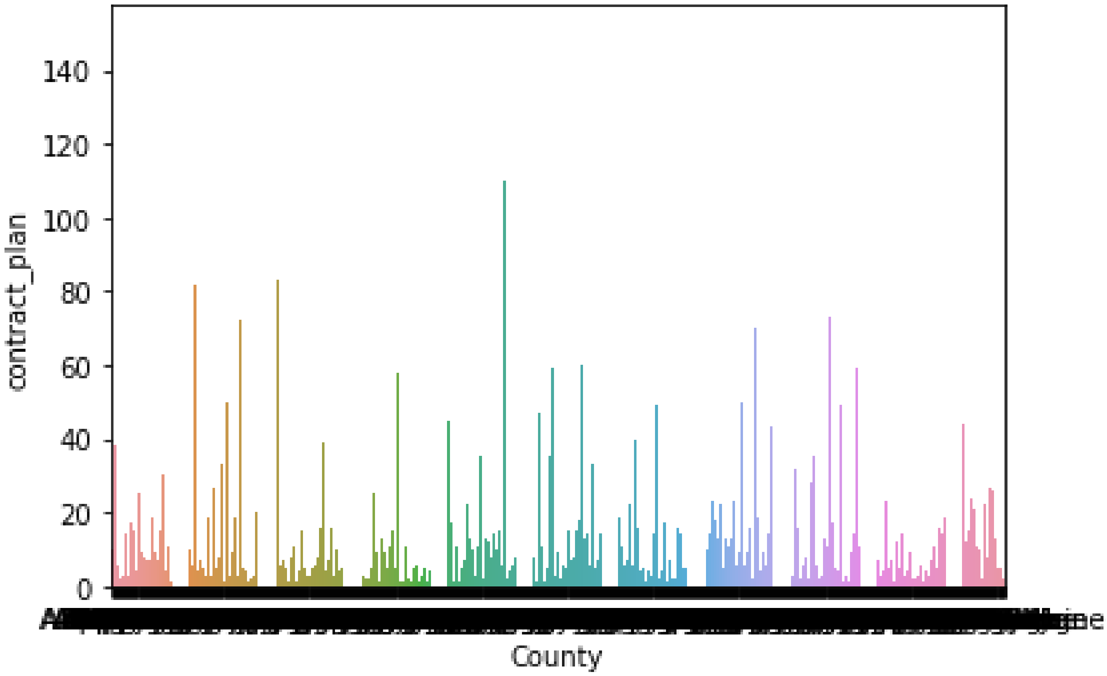
3.6. What percentage of the missing are recovered that way? 0.98969784

3.7. See code 3.g.

3.8. Plots

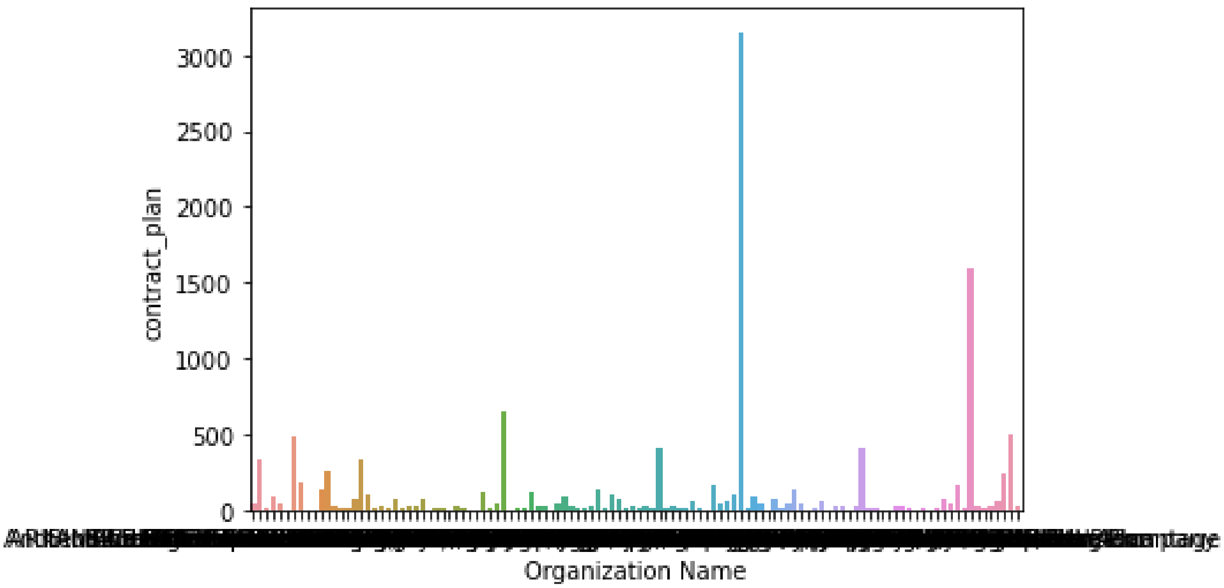
3.8.1. Plot the histogram of contract-plans per county.

There is a problem with the question itself. Instead of plotting a histogram, we should plot a bar chart. The x-axis of histogram cannot be categories such as ‘per county’.

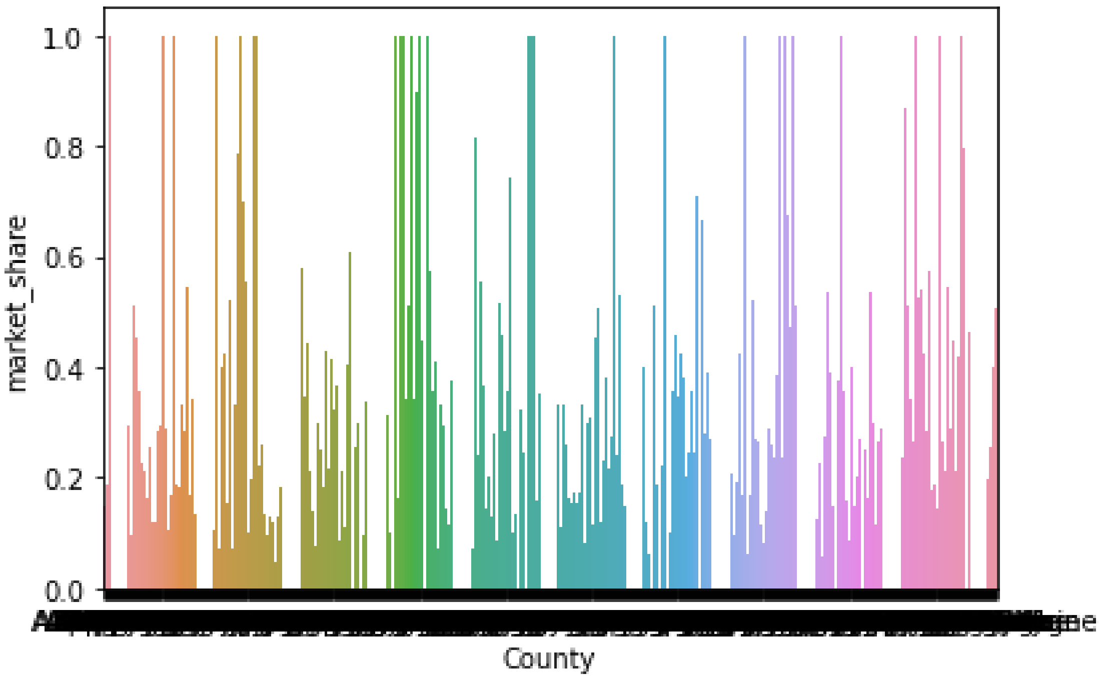


3.8.2. Plot the histogram of contract-plans per organization.

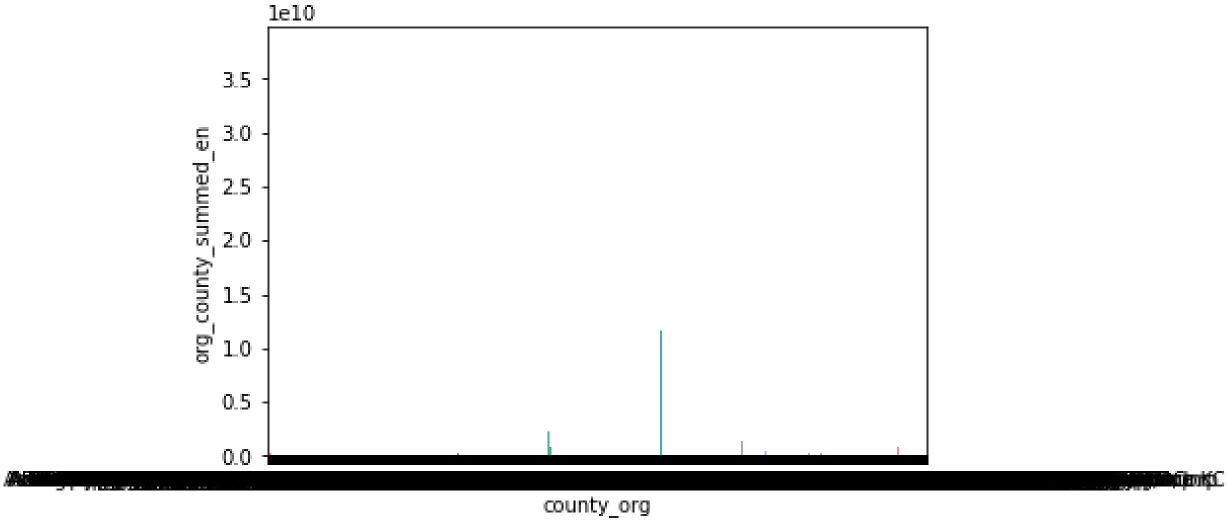
Same problem as the previous question.



3.8.3. Plot the histogram of HHIs per county.

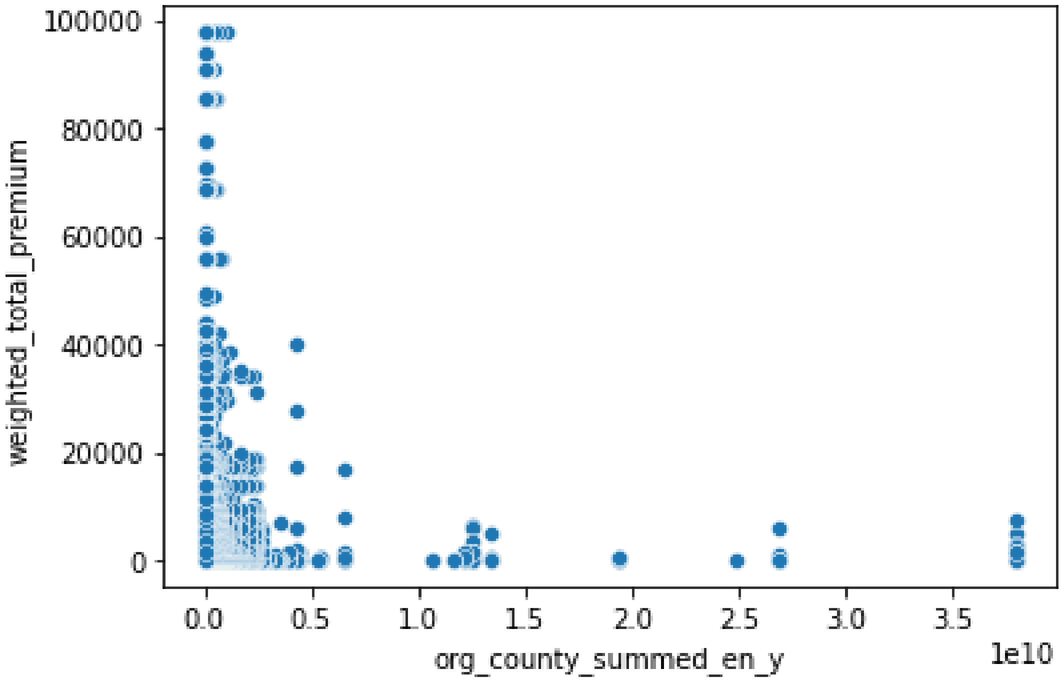


3.8.4.Plot the histogram of HHI per firm-county.



On average, how many firms account for 90% or more of the market share by county? 83

Create a scatter plot (binned if possible) of firm-county HHI and the average premium. What is the correlation between the two?



The higher the HHI is, the lower average premium it will have.

(extra) Explain why we cannot infer a causal link between market concentration and prices from this data alone. How would you proceed if you were trying to study their interconnection?

There are some bias that will stop us from inferring a causal relationship.

For example: Omitted variable bias: Other factors will also influence the premium such as ages, health conditions...

Simultaneous endogeneity: The premium will affect the companies concentration on the market and the market concentration can also be a factor in consideration while pricing.

We can add more variables that are related to the premium. We can also add IV.