Today's Agenda

1. Organizing Data

2. Univariate Analyses

3. Bivariate Analyses

Justin Leinaweaver (Summer 2023)

```
  [1] 101.330000 121.596000  60.798000  50.665000 101.330000 125.649200 103.356600  18.239400
  [9]  58.771400  64.851200   4.053200  70.931000  54.718200  13.233698   0.000000  60.798000
 [17]  24.319200  30.399000  24.319200  40.532000   0.000000  50.665000  44.585200  50.665000
 [25]  40.532000  34.452200  81.064000  89.170400   0.000000  14.186200 107.409800  10.133000
 [33]  10.133000  28.372400  91.197000  11.146300  81.064000  68.904400  10.133000  20.266000
 [41]   0.000000  54.718200 125.649200  30.399000   0.000000 101.330000   0.000000  83.090600
 [49]   0.000000  30.399000  50.665000  50.665000 151.995000  54.718200  24.319200  81.064000
 [57]  15.199500  50.665000  60.798000  42.558600  54.718200  91.197000   6.079800  50.665000
 [65]  48.638400  64.851200  20.266000   0.000000  22.292600  37.897420  40.532000   7.093100
 [73]  26.345800  50.665000  42.558600  68.904400  10.133000  12.159600  34.452200  70.931000
 [81]   8.106400  10.133000  28.372400  20.266000  50.665000   0.000000 176.314200  32.425600
 [89]   0.000000  30.399000   0.000000  32.425600  50.665000  33.438900  30.399000   8.106400
 [97]   7.782144  10.133000  44.585200   0.405320  52.691600   5.066500  34.452200  50.665000
[105]  16.212800  50.665000   0.000000  24.319200  20.266000  10.133000 202.660000  20.266000
[113]   0.000000  30.399000  91.197000  30.399000   4.863840  60.798000  60.798000  20.266000
[121]  10.133000  24.319200  40.532000  40.532000  40.532000  50.665000  10.133000   0.000000
[129]  50.665000   2.431920  10.133000   1.418620  40.532000  20.266000  60.798000  50.665000
[137]  81.064000  50.665000  91.197000  20.266000 121.596000  91.197000   0.000000  36.478800
[145]  30.399000 101.330000  32.506664  30.399000  20.266000  66.877800  36.478800  30.399000
[153]  42.558600  42.558600  74.984200  77.010800  34.452200   0.000000   0.000000  30.399000
[161]  64.851200  55.731500  30.399000  33.438900   0.000000   0.000000  50.665000  54.718200
[169]  10.133000 141.862000  10.133000  10.133000 125.649200  30.399000  40.532000   8.106400
```

And 1,639 more responses...

The Middle	
Mean	42
Median	32
The Range	
Minimum	0
Maximum	405
Range	405
The Variation	
Standard Deviation	42

```
  [1] 101.330000 121.596000  60.798000  50.665000 101.330000 125.649200
  [7] 103.356600  18.239400  58.771400  64.851200   4.053200  70.931000
 [13]  54.718200  13.233698   0.000000  60.798000  24.319200  30.399000
 [19]  24.319200  40.532000   0.000000  50.665000  44.585200  50.665000
 [25]  40.532000  34.452200  81.064000  89.170400   0.000000  14.186200
 [31] 107.409800  10.133000  10.133000  28.372400  91.197000  11.146300
 [37]  81.064000  68.904400  10.133000  20.266000   0.000000  54.718200
 [43] 125.649200  30.399000   0.000000 101.330000   0.000000  83.090600
 [49]   0.000000  30.399000  50.665000  50.665000 151.995000  54.718200
 [55]  24.319200  81.064000  15.199500  50.665000  60.798000  42.558600
 [61]  54.718200  91.197000   6.079800  50.665000  48.638400  64.851200
 [67]  20.266000   0.000000  22.292600  37.897420  40.532000   7.093100
 [73]  26.345800  50.665000  42.558600  68.904400  10.133000  12.159600
 [79]  34.452200  70.931000   8.106400  10.133000  28.372400  20.266000
 [85]  50.665000   0.000000 176.314200  32.425600   0.000000  30.399000
 [91]   0.000000  32.425600  50.665000  33.438900  30.399000   8.106400
 [97]   7.782144  10.133000  44.585200   0.405320  52.691600   5.066500
[103]  34.452200  50.665000  16.212800  50.665000   0.000000  24.319200
[109]  20.266000  10.133000 202.660000  20.266000   0.000000  30.399000
[115]  91.197000  30.399000   4.863840  60.798000  60.798000  20.266000
[121]  10.133000  24.319200  40.532000  40.532000  40.532000  50.665000
[127]  10.133000   0.000000  50.665000   2.431920  10.133000   1.418620
[133]  40.532000  20.266000  60.798000  50.665000  81.064000  50.665000
[139]  91.197000  20.266000 121.596000  91.197000   0.000000  36.478800
[145]  30.399000 101.330000  32.506664  30.399000  20.266000  66.877800
[151]  36.478800  30.399000  42.558600  42.558600  74.984200  77.010800
[157]  34.452200   0.000000   0.000000  30.399000  64.851200  55.731500
[163]  30.399000  33.438900   0.000000   0.000000  50.665000  54.718200
[169]  10.133000 141.862000  10.133000  10.133000 125.649200  30.399000
[175]  40.532000   8.106400 151.995000 121.596000  10.133000  60.798000
[181] 141.862000  10.133000 101.330000  89.170400  81.064000  60.798000
[187]  30.399000  70.931000   0.000000  20.266000  46.611800   0.000000
[193]  91.197000  30.399000   8.106400  34.452200  60.798000  55.731500
[199]  11.527301  36.478800  70.931000  91.197000  87.143800  64.851200
```

Defining Statistics: Level 1

Statistics is a set of tools we use to summarize data

Summarize: "give a brief statement of the main points of (something)" (Oxford Dictionary).

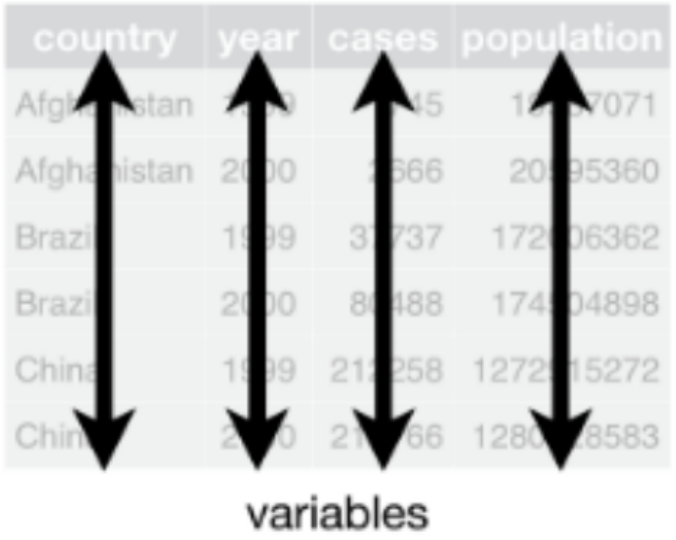
Defining Statistics: Level 2

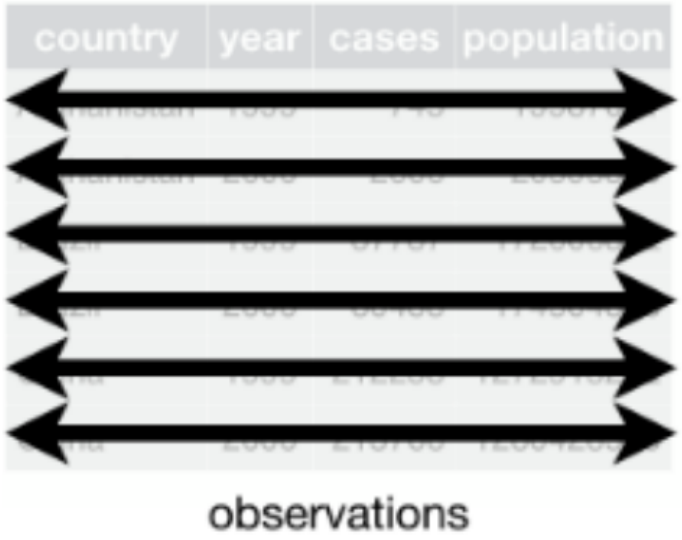
"The practice or science of collecting and analyzing numerical data in large quantities, **especially for the purpose of inferring proportions in a whole from those in a representative sample**" (Oxford Dictionary).

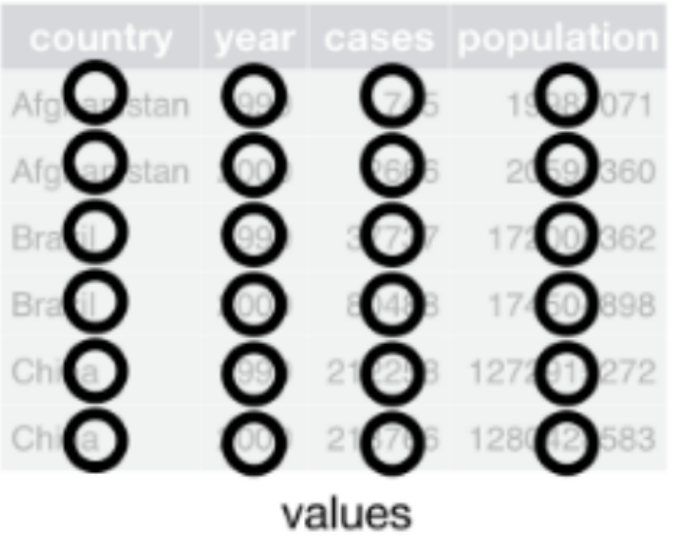
Economic Data on the US States

state	abbrev	year	gdp_millions	gdp_rate	gdp_category	min_wage	unemployment
Alabama	AL	2018	221735.5	0.05	Medium	0.00	3.9
Alaska	AK	2018	54734.1	0.06	Low	9.84	6.6
Arizona	AZ	2018	348297.1	0.06	Medium	10.50	4.8
Arkansas	AR	2018	128418.9	0.04	Low	8.50	3.7
California	CA	2018	2997732.8	0.06	High	11.00	4.2
Colorado	CO	2018	371749.6	0.06	High	10.20	3.3
Connecticut	CT	2018	275726.9	0.03	Medium	10.10	4.1
Delaware	DE	2018	73481.3	0.04	Low	8.25	3.8
Florida	FL	2018	1039236.4	0.05	High	8.25	3.6
Georgia	GA	2018	592153.4	0.05	High	5.15	3.9
Hawaii	HI	2018	93797.9	0.05	Low	10.10	2.4

"Three Rules of Tidy Data" (Wickham 2018)







Economic Data on the US States

state	abbrev	year	gdp_millions	gdp_rate	gdp_category	min_wage	unemployment
Alabama	AL	2018	221735.5	0.05	Medium	0.00	3.9
Alaska	AK	2018	54734.1	0.06	Low	9.84	6.6
Arizona	AZ	2018	348297.1	0.06	Medium	10.50	4.8
Arkansas	AR	2018	128418.9	0.04	Low	8.50	3.7
California	CA	2018	2997732.8	0.06	High	11.00	4.2
Colorado	CO	2018	371749.6	0.06	High	10.20	3.3
Connecticut	CT	2018	275726.9	0.03	Medium	10.10	4.1
Delaware	DE	2018	73481.3	0.04	Low	8.25	3.8
Florida	FL	2018	1039236.4	0.05	High	8.25	3.6
Georgia	GA	2018	592153.4	0.05	High	5.15	3.9
Hawaii	HI	2018	93797.9	0.05	Low	10.10	2.4

Univariate Analyses

Variable type determines analysis tool

- Categorical Variables
  - Nominal
  - Ordinal
- Numerical Variables
  - Interval
  - Ratio

Univariate Analyses

Variable type determines analysis tool

- Categorical Variables
  - Nominal (named items)
  - Ordinal (items in order)
- Numerical Variables
  - Interval (Differences matter)
  - Ratio (Zero matters)

Analyzing Categorical Variables: Tables and Bar Plots

Nominal (e.g. gender, ethnicity)

P
income_tax
Income Tax
No Income Tax
Income Tax
Income Tax
Income Tax
Income Tax
Income Tax
Income Tax
No Income Tax
Income Tax
Income Tax
Income Tax
Income Tax

Ordinal (e.g. rating scales)

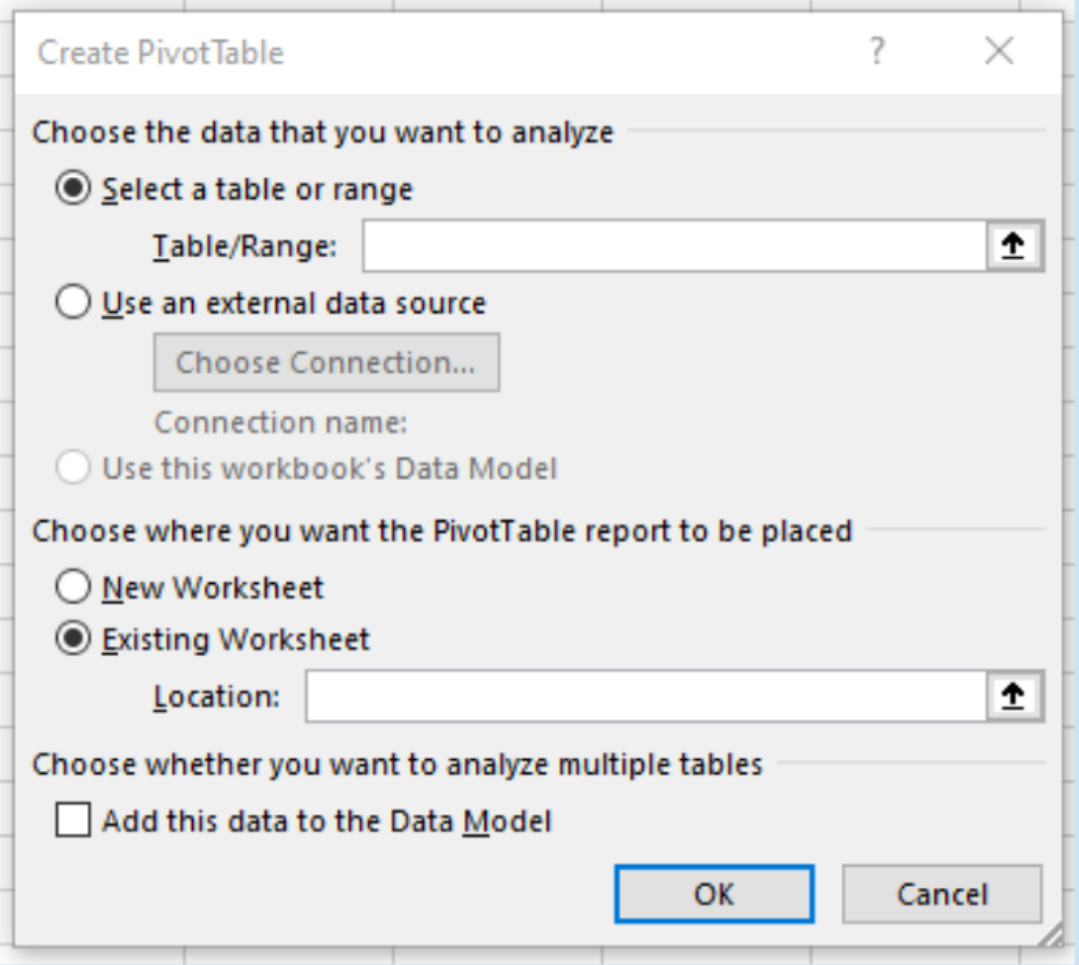
F
gdp_category
Medium
Low
Medium
Low
High
High
Medium
Low
High
High
Low
Low

Univariate Analyses: Categorical Variables

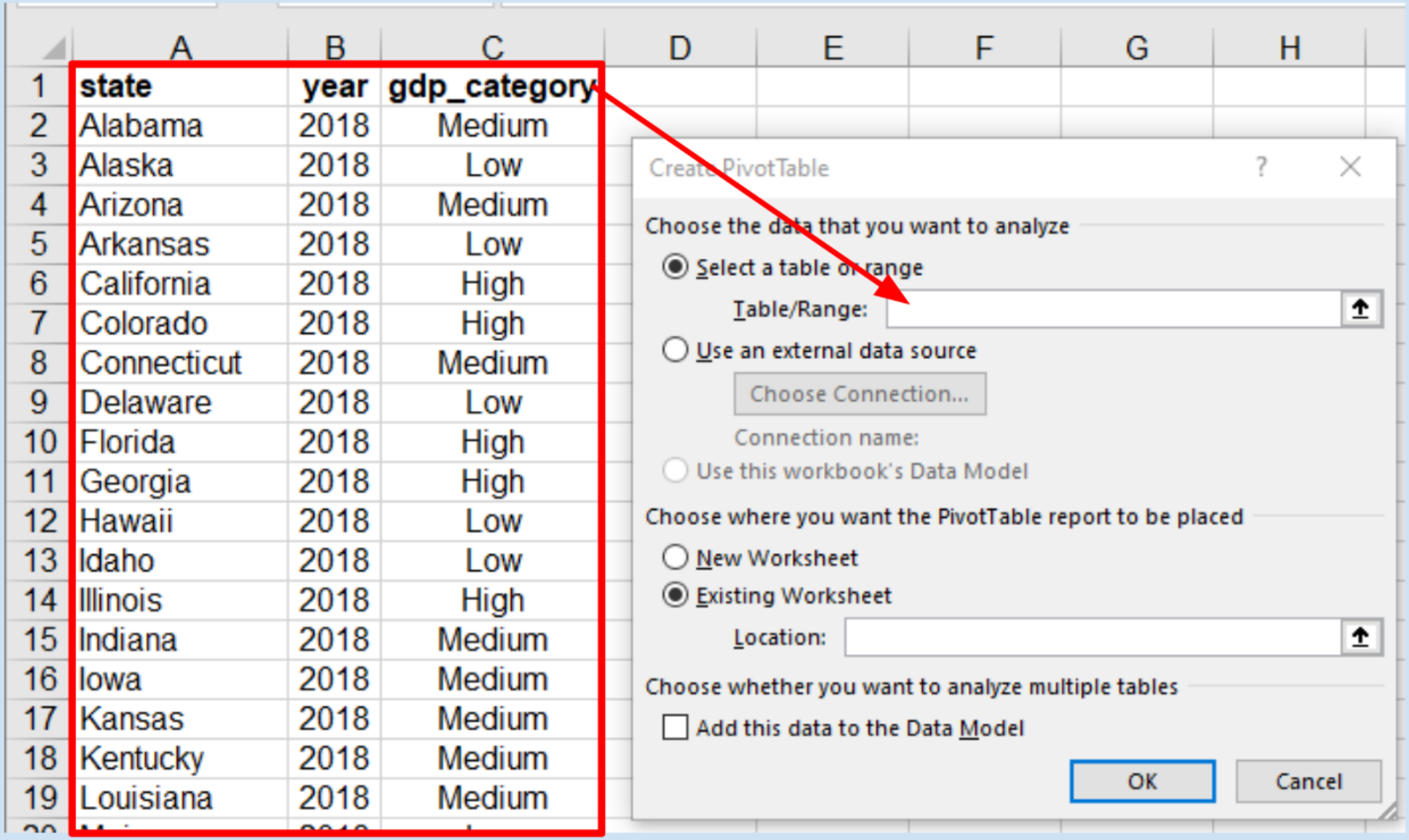
	A	B	C
1	**state**	**year**	**gdp_category**
2	Alabama	2018	Medium
3	Alaska	2018	Low
4	Arizona	2018	Medium
5	Arkansas	2018	Low
6	California	2018	High
7	Colorado	2018	High
8	Connecticut	2018	Medium
9	Delaware	2018	Low
10	Florida	2018	High
11	Georgia	2018	High
12	Hawaii	2018	Low
13	Idaho	2018	Low
14	Illinois	2018	High
15	Indiana	2018	Medium
16	Iowa	2018	Medium
17	Kansas	2018	Medium
18	Kentucky	2018	Medium
19	Louisiana	2018	Medium

Univariate Analyses: Categorical Variables

	A	B	C
1	**state**	**year**	**gdp_category**
2	Alabama	2018	Medium
3	Alaska	2018	Low
4	Arizona	2018	Medium
5	Arkansas	2018	Low
6	California	2018	High
7	Colorado	2018	High
8	Connecticut	2018	Medium
9	Delaware	2018	Low
10	Florida	2018	High
11	Georgia	2018	High
12	Hawaii	2018	Low
13	Idaho	2018	Low
14	Illinois	2018	High
15	Indiana	2018	Medium
16	Iowa	2018	Medium
17	Kansas	2018	Medium
18	Kentucky	2018	Medium
19	Louisiana	2018	Medium

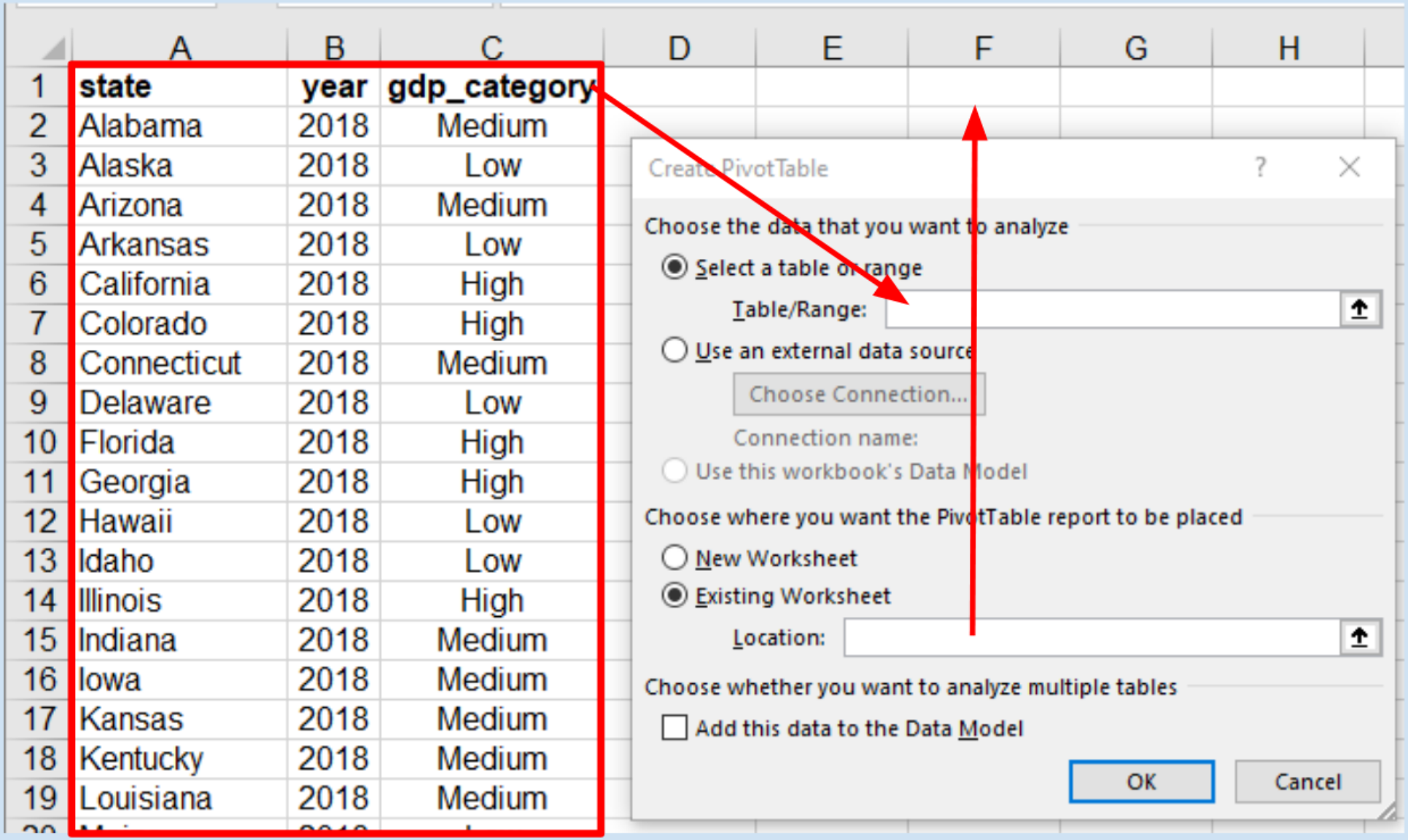


Univariate Analyses: Categorical Variables

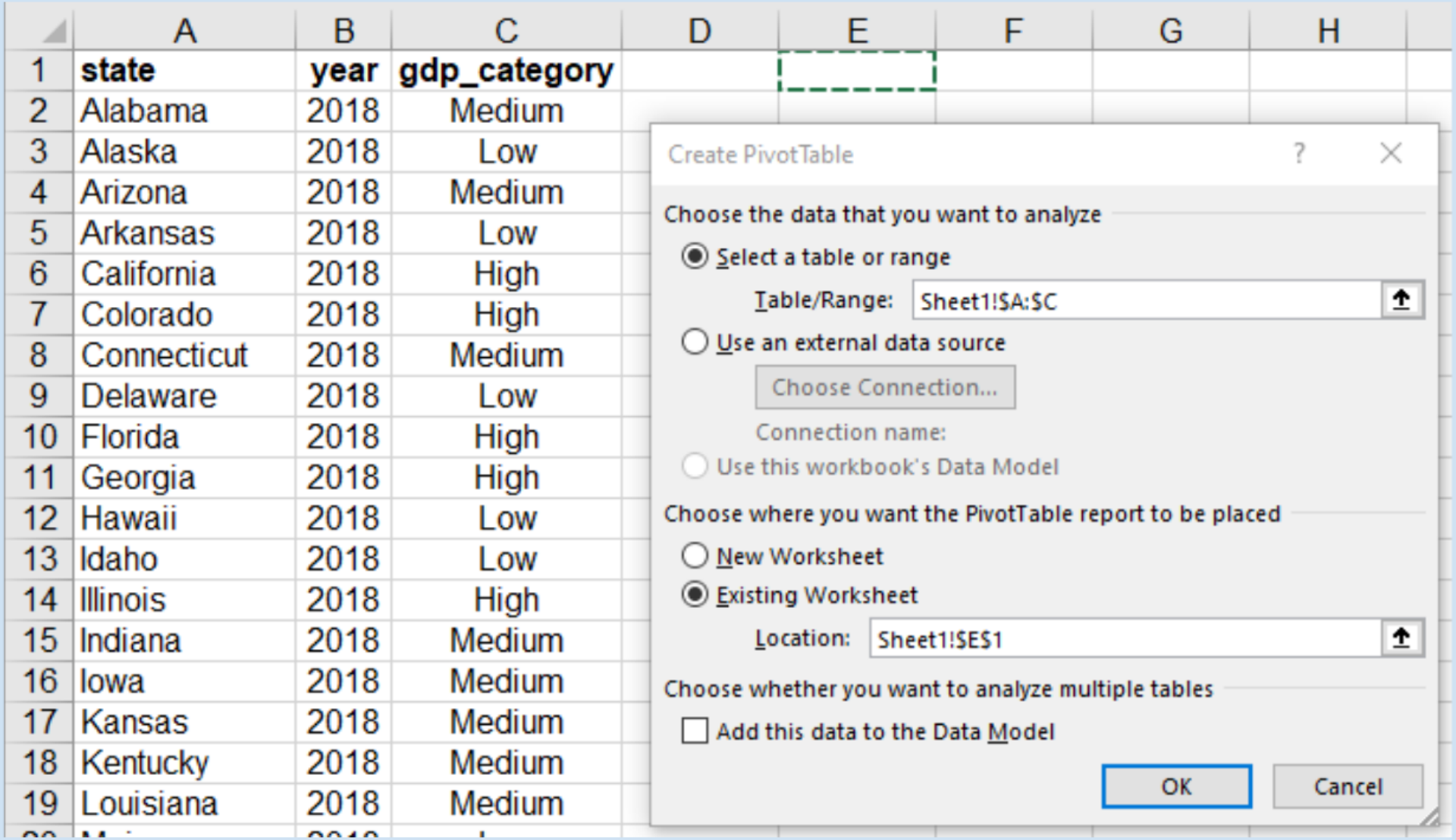


	A	B	C
1	**state**	**year**	**gdp_category**
2	Alabama	2018	Medium
3	Alaska	2018	Low
4	Arizona	2018	Medium
5	Arkansas	2018	Low
6	California	2018	High
7	Colorado	2018	High
8	Connecticut	2018	Medium
9	Delaware	2018	Low
10	Florida	2018	High
11	Georgia	2018	High
12	Hawaii	2018	Low
13	Idaho	2018	Low
14	Illinois	2018	High
15	Indiana	2018	Medium
16	Iowa	2018	Medium
17	Kansas	2018	Medium
18	Kentucky	2018	Medium
19	Louisiana	2018	Medium

Univariate Analyses: Categorical Variables



Univariate Analyses: Categorical Variables



	A	B	C
1	**state**	**year**	**gdp_category**
2	Alabama	2018	Medium
3	Alaska	2018	Low
4	Arizona	2018	Medium
5	Arkansas	2018	Low
6	California	2018	High
7	Colorado	2018	High
8	Connecticut	2018	Medium
9	Delaware	2018	Low
10	Florida	2018	High
11	Georgia	2018	High
12	Hawaii	2018	Low
13	Idaho	2018	Low
14	Illinois	2018	High
15	Indiana	2018	Medium
16	Iowa	2018	Medium
17	Kansas	2018	Medium
18	Kentucky	2018	Medium
19	Louisiana	2018	Medium

Univariate Analyses: Categorical Variables

	A	B	C	D	E	F	G
1	**state**	**year**	**gdp_category**				
2	Alabama	2018	Medium				
3	Alaska	2018	Low				
4	Arizona	2018	Medium				
5	Arkansas	2018	Low				
6	California	2018	High				
7	Colorado	2018	High				
8	Connecticut	2018	Medium				
9	Delaware	2018	Low				
10	Florida	2018	High				
11	Georgia	2018	High				
12	Hawaii	2018	Low				
13	Idaho	2018	Low				
14	Illinois	2018	High				
15	Indiana	2018	Medium				
16	Iowa	2018	Medium				
17	Kansas	2018	Medium				
18	Kentucky	2018	Medium				
19	Louisiana	2018	Medium				
20	Maine	2018	Low				

PivotTable1

To build a report, choose fields from the PivotTable Field List

PivotTable Fields

Choose fields to add to report:

Search

- state
- year
- gdp_category

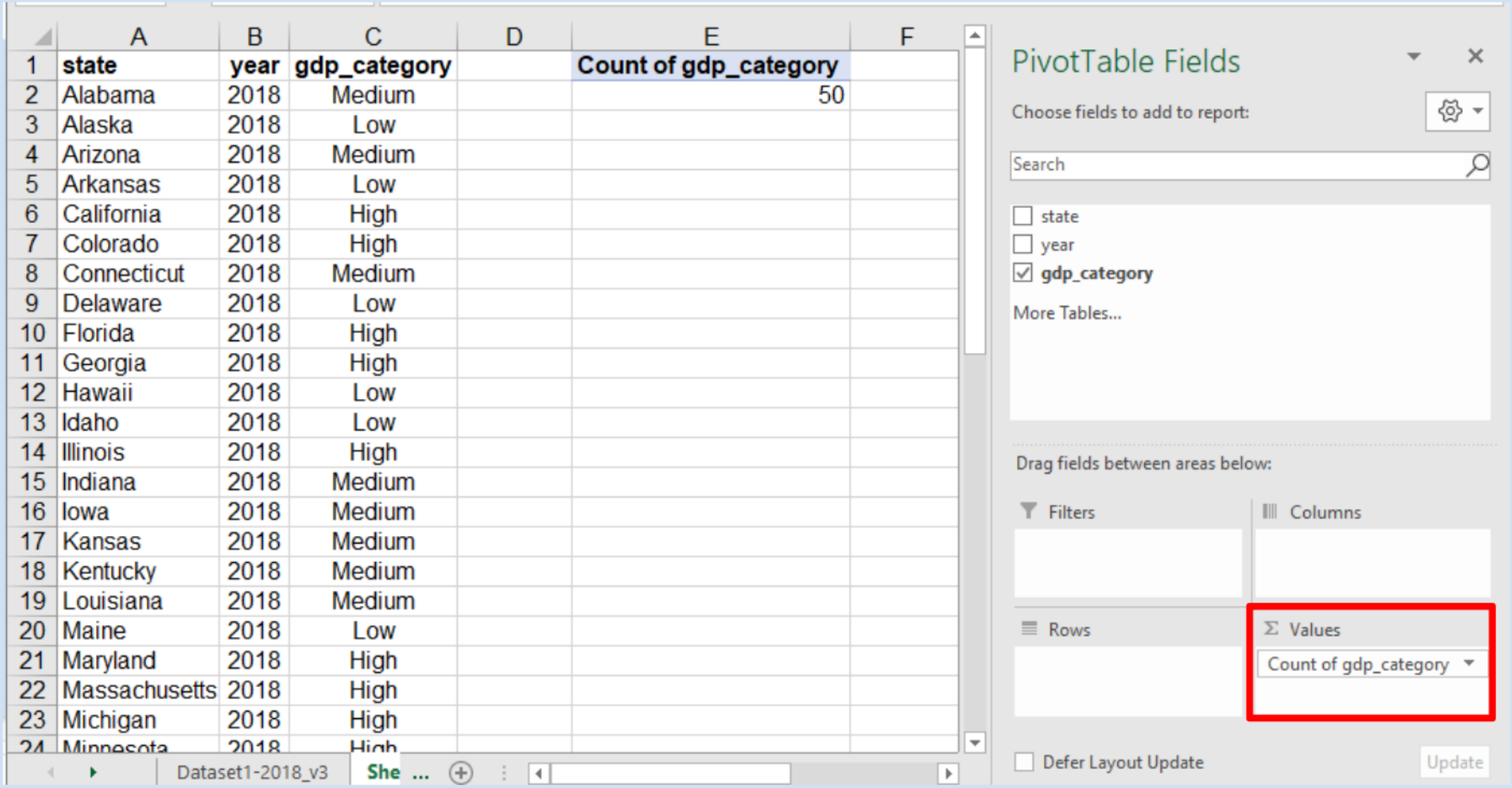
More Tables...

Drag fields between areas below:

Filters | Columns

Rows | Values

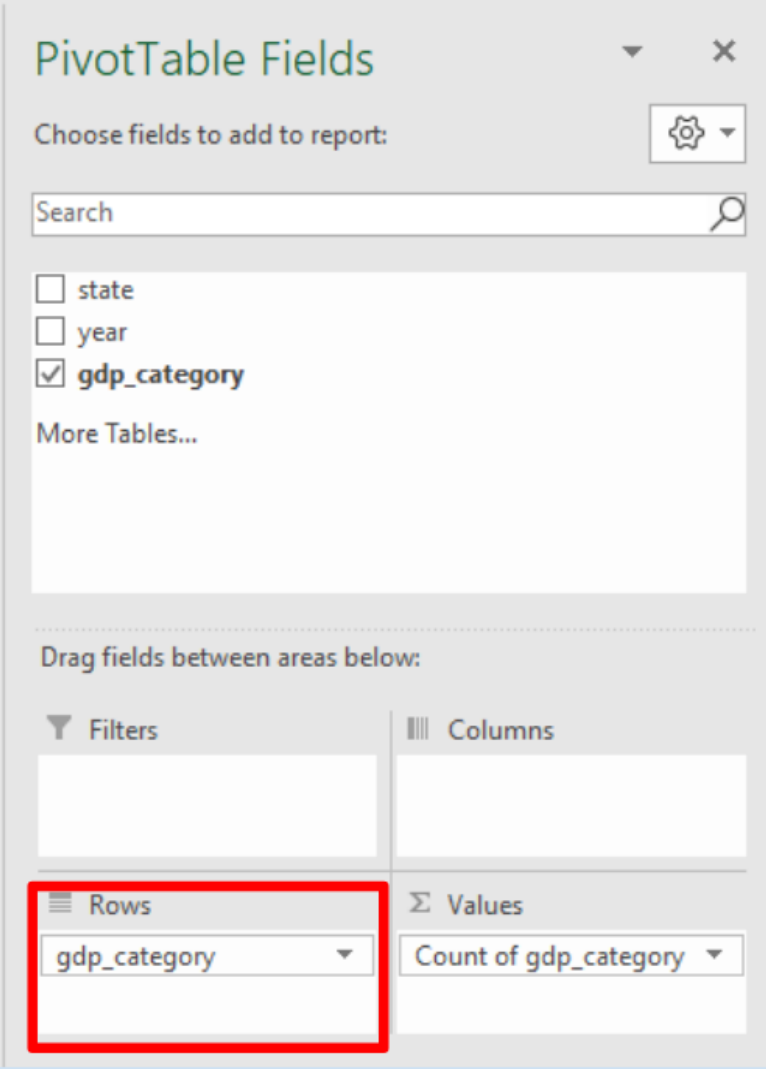
Univariate Analyses: Categorical Variables

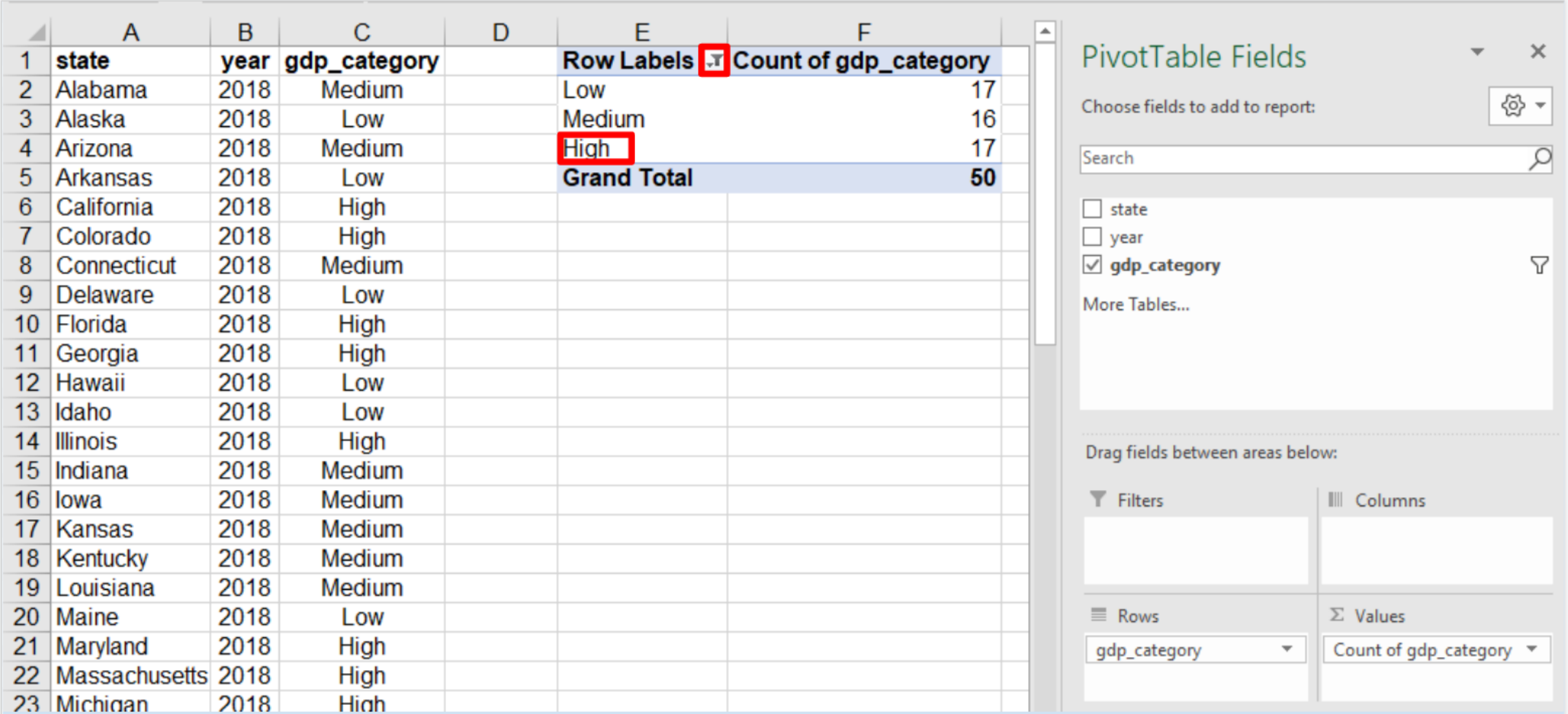


	A	B	C	D	E	F
1	**state**	**year**	**gdp_category**		**Count of gdp_category**	
2	Alabama	2018	Medium		50	
3	Alaska	2018	Low			
4	Arizona	2018	Medium			
5	Arkansas	2018	Low			
6	California	2018	High			
7	Colorado	2018	High			
8	Connecticut	2018	Medium			
9	Delaware	2018	Low			
10	Florida	2018	High			
11	Georgia	2018	High			
12	Hawaii	2018	Low			
13	Idaho	2018	Low			
14	Illinois	2018	High			
15	Indiana	2018	Medium			
16	Iowa	2018	Medium			
17	Kansas	2018	Medium			
18	Kentucky	2018	Medium			
19	Louisiana	2018	Medium			
20	Maine	2018	Low			
21	Maryland	2018	High			
22	Massachusetts	2018	High			
23	Michigan	2018	High			
24	Minnesota	2018	High			

Univariate Analyses: Categorical Variables

	A	B	C	D	E	F
1	**state**	**year**	**gdp_category**		**Row Labels**	**Count of gdp_category**
2	Alabama	2018	Medium		High	17
3	Alaska	2018	Low		Low	17
4	Arizona	2018	Medium		Medium	16
5	Arkansas	2018	Low		(blank)	
6	California	2018	High		**Grand Total**	**50**
7	Colorado	2018	High			
8	Connecticut	2018	Medium			
9	Delaware	2018	Low			
10	Florida	2018	High			
11	Georgia	2018	High			
12	Hawaii	2018	Low			
13	Idaho	2018	Low			
14	Illinois	2018	High			
15	Indiana	2018	Medium			
16	Iowa	2018	Medium			
17	Kansas	2018	Medium			
18	Kentucky	2018	Medium			
19	Louisiana	2018	Medium			
20	Maine	2018	Low			
21	Maryland	2018	High			
22	Massachusetts	2018	High			
23	Michigan	2018	High			



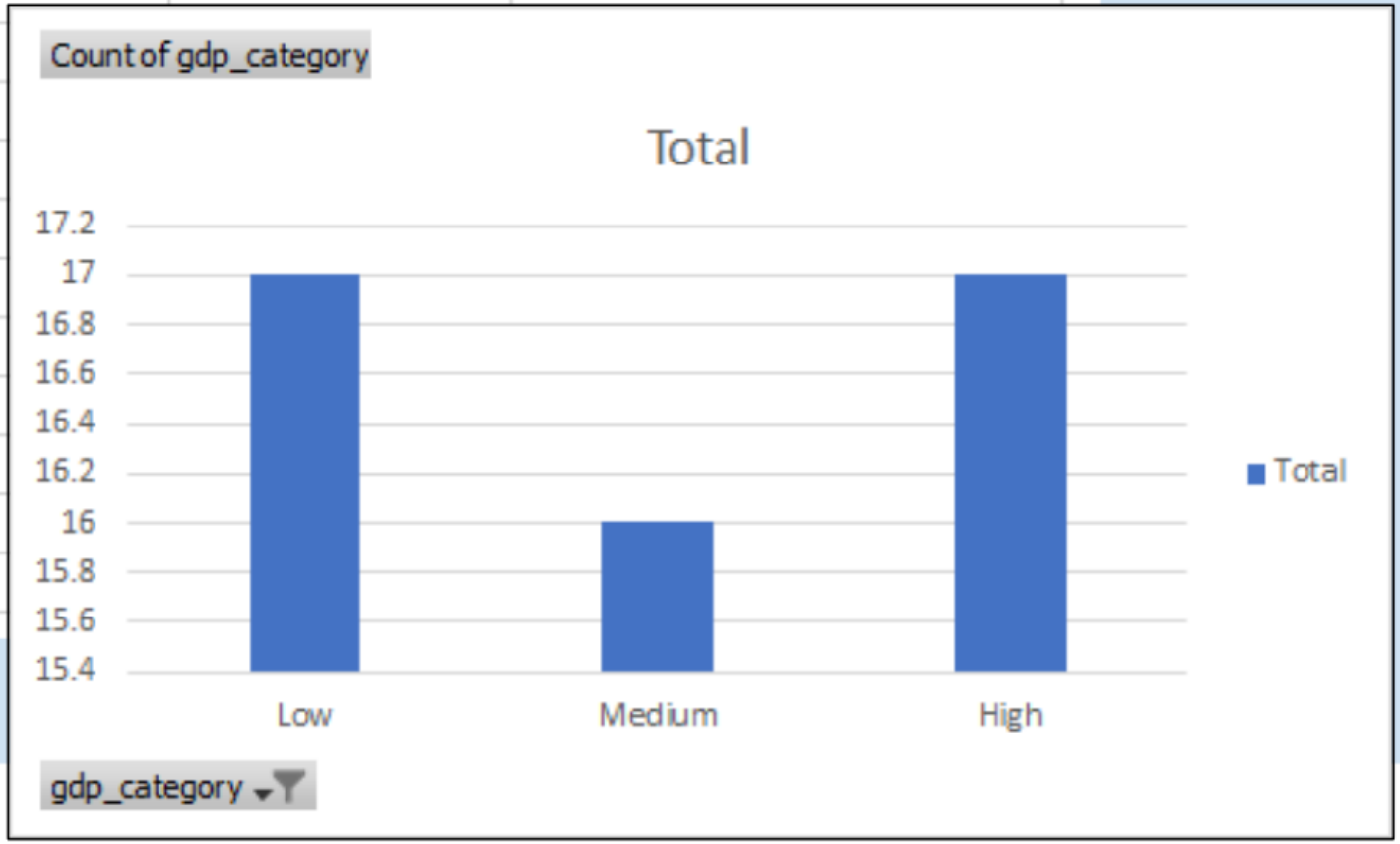


	A	B	C	D	E	F
1	state	year	gdp_category		Row Labels	Count of gdp_category
2	Alabama	2018	Medium		Low	17
3	Alaska	2018	Low		Medium	16
4	Arizona	2018	Medium		High	17
5	Arkansas	2018	Low		Grand Total	50
6	California	2018	High			
7	Colorado	2018	High			
8	Connecticut	2018	Medium			
9	Delaware	2018	Low			
10	Florida	2018	High			
11	Georgia	2018	High			
12	Hawaii	2018	Low			
13	Idaho	2018	Low			
14	Illinois	2018	High			
15	Indiana	2018	Medium			
16	Iowa	2018	Medium			
17	Kansas	2018	Medium			
18	Kentucky	2018	Medium			
19	Louisiana	2018	Medium			
20	Maine	2018	Low			
21	Maryland	2018	High			
22	Massachusetts	2018	High			
23	Michigan	2018	High			

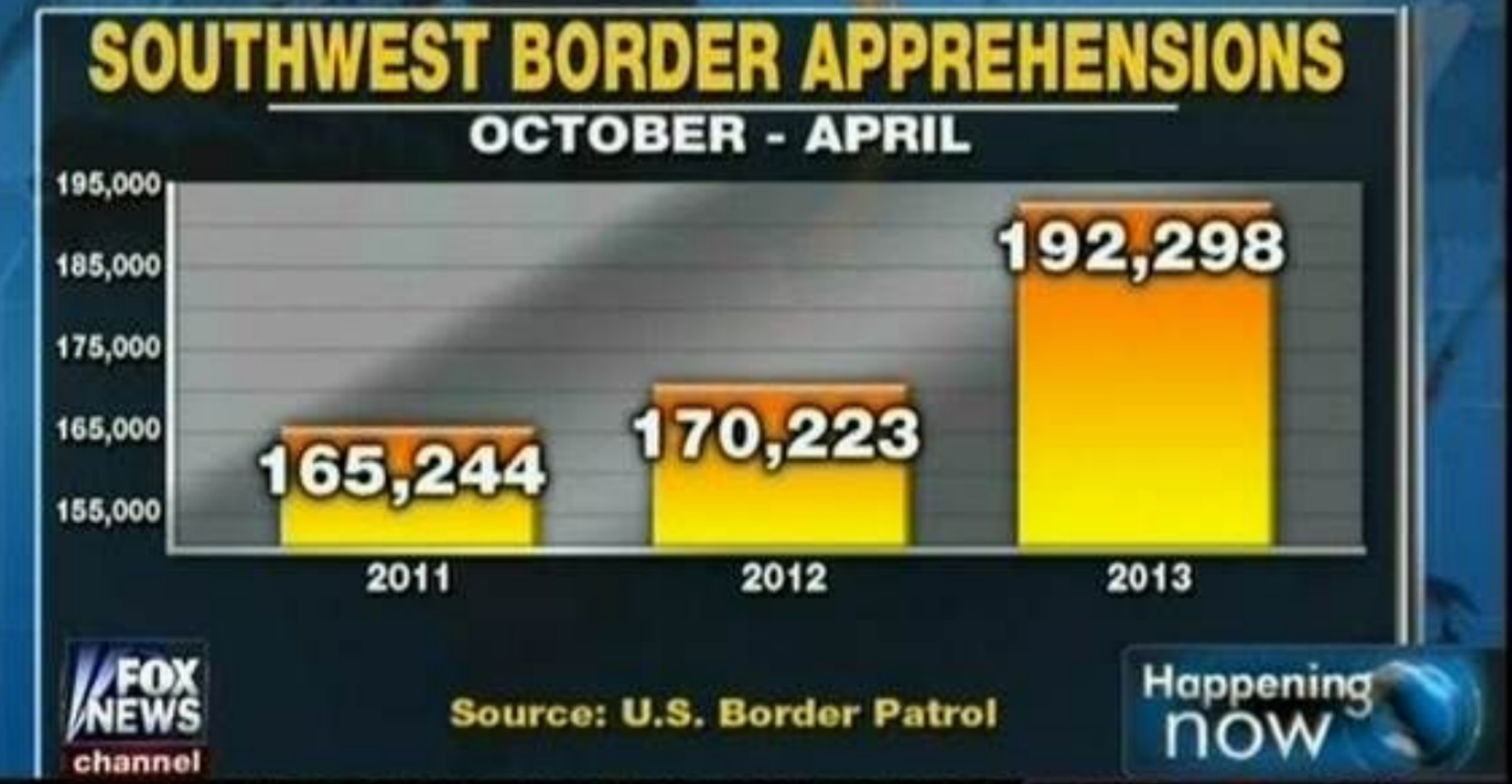
1. Remove 'blank' rows using the drop-down beside 'Row Labels'
2. Right-click 'High' and 'Move' to 'End'

	state	year	gdp_category
1	Alabama	2018	Medium
2	Alaska	2018	Low
3	Arizona	2018	Medium
4	Arkansas	2018	Low
5	California	2018	High
6	Colorado	2018	High
7	Connecticut	2018	Medium
8	Delaware	2018	Low
9	Florida	2018	High
10	Georgia	2018	High
11	Hawaii	2018	Low
12	Idaho	2018	Low
13	Illinois	2018	High
14	Indiana	2018	Medium
15	Iowa	2018	Medium
16	Kansas	2018	Medium

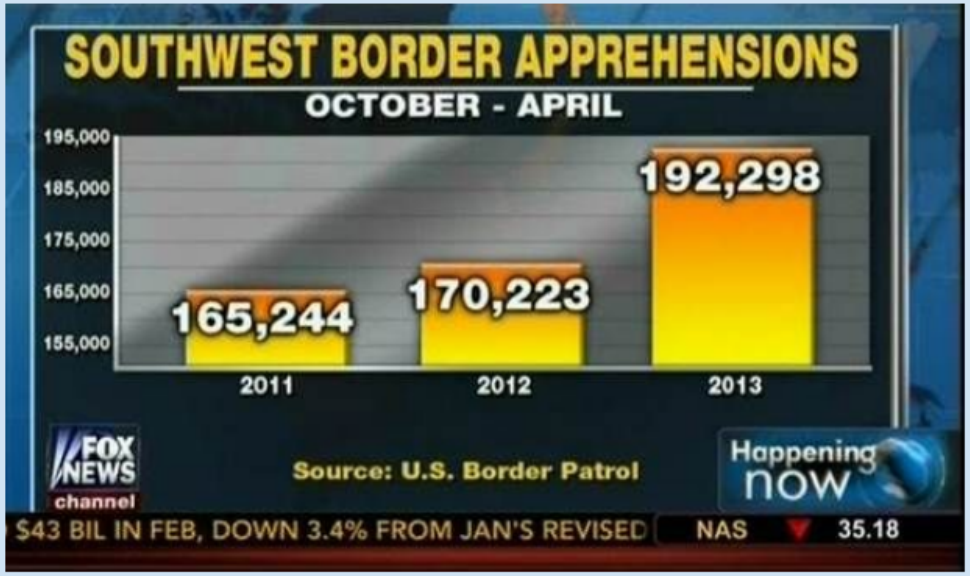
Row Labels	Count of gdp_category
Low	17
Medium	16
High	17
Grand Total	**50**

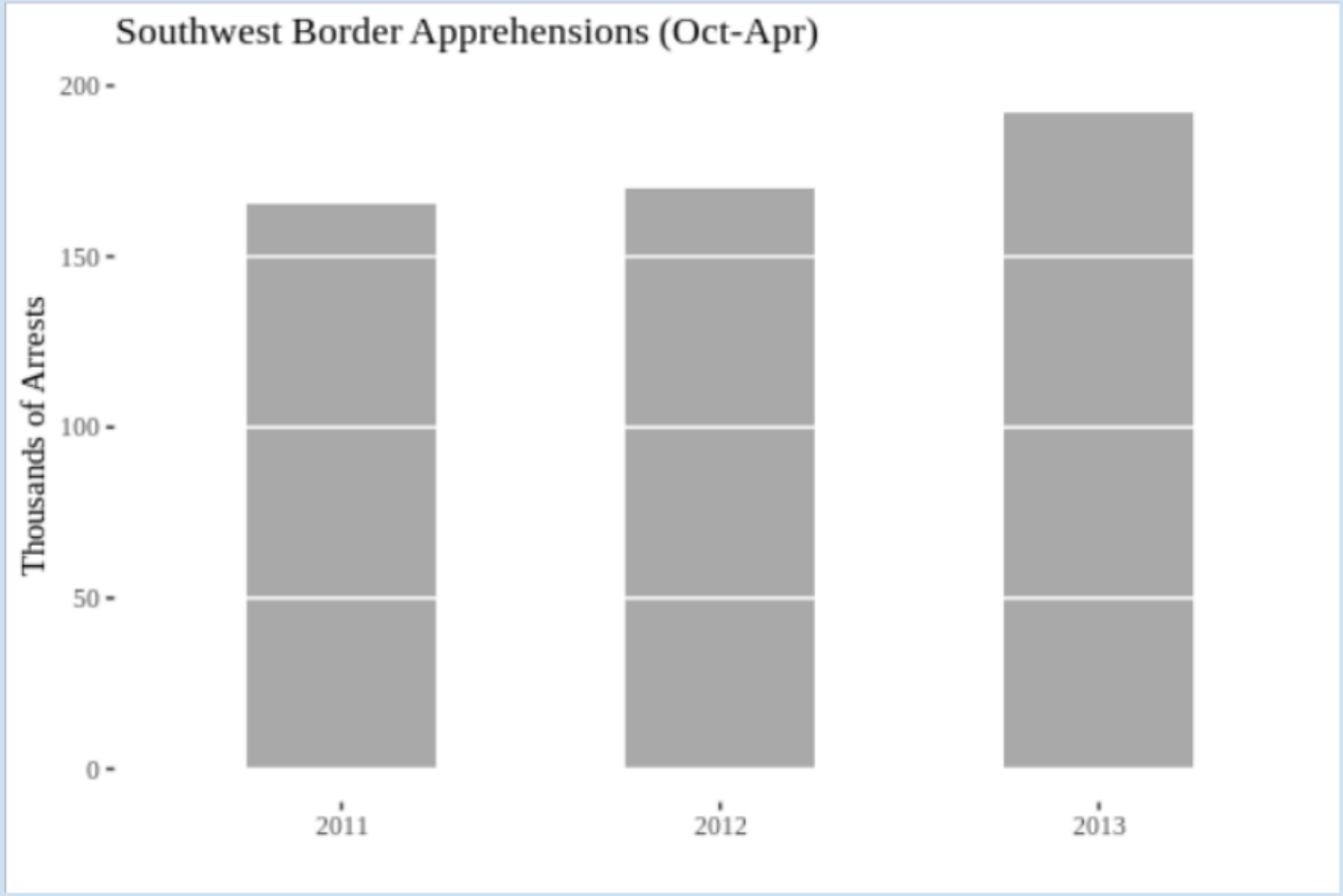


1. Highlight the pivot table
2. 'Insert' a 2-D Column Chart (aka a bar plot)


SOUTHWEST BORDER APPREHENSIONS
OCTOBER - APRIL
195,000
185,000
175,000
165,000
155,000
165,244
170,223
192,298
2011
2012
2013
FOX NEWS channel
Source: U.S. Border Patrol
Happening now

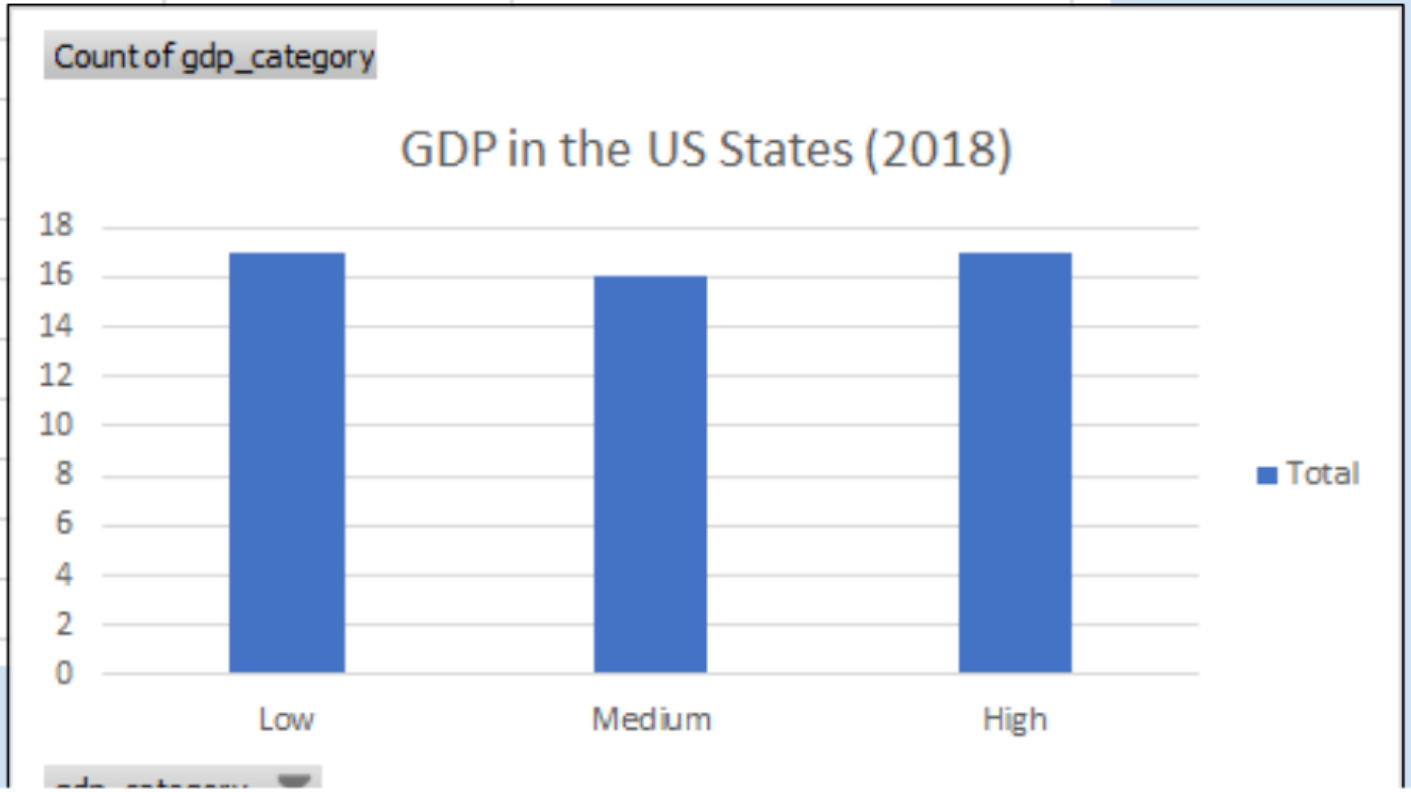
Start all bar plots (y-axis) at zero





	state	year	gdp_category
1			
2	Alabama	2018	Medium
3	Alaska	2018	Low
4	Arizona	2018	Medium
5	Arkansas	2018	Low
6	California	2018	High
7	Colorado	2018	High
8	Connecticut	2018	Medium
9	Delaware	2018	Low
10	Florida	2018	High
11	Georgia	2018	High
12	Hawaii	2018	Low
13	Idaho	2018	Low
14	Illinois	2018	High
15	Indiana	2018	Medium
16	Iowa	2018	Medium

Row Labels	Count of gdp_category
Low	17
Medium	16
High	17
Grand Total	**50**



1. Right-click the y-axis and 'Format the Axis' to start at 0
2. Add an informative title

Professional Visualizations

1. Informative titles
2. Figure labels
3. Clean axis labels
4. Source info
5. Minimal clutter / no chart junk

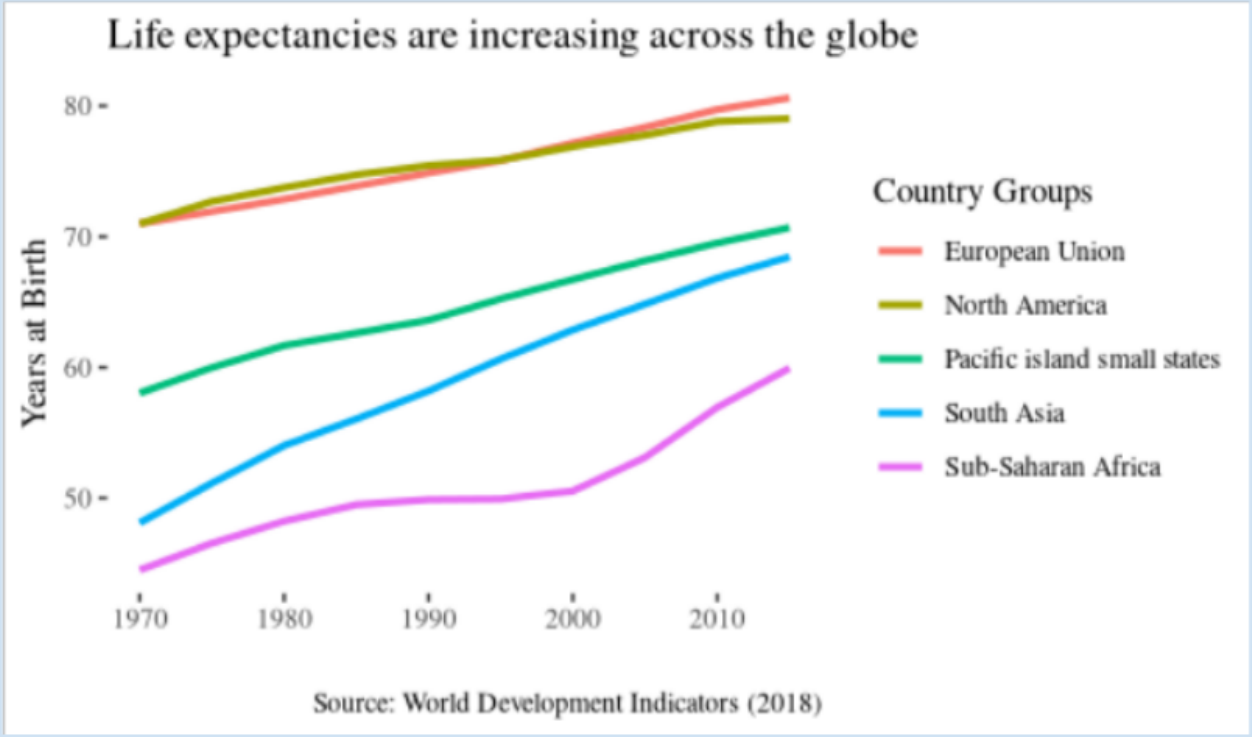
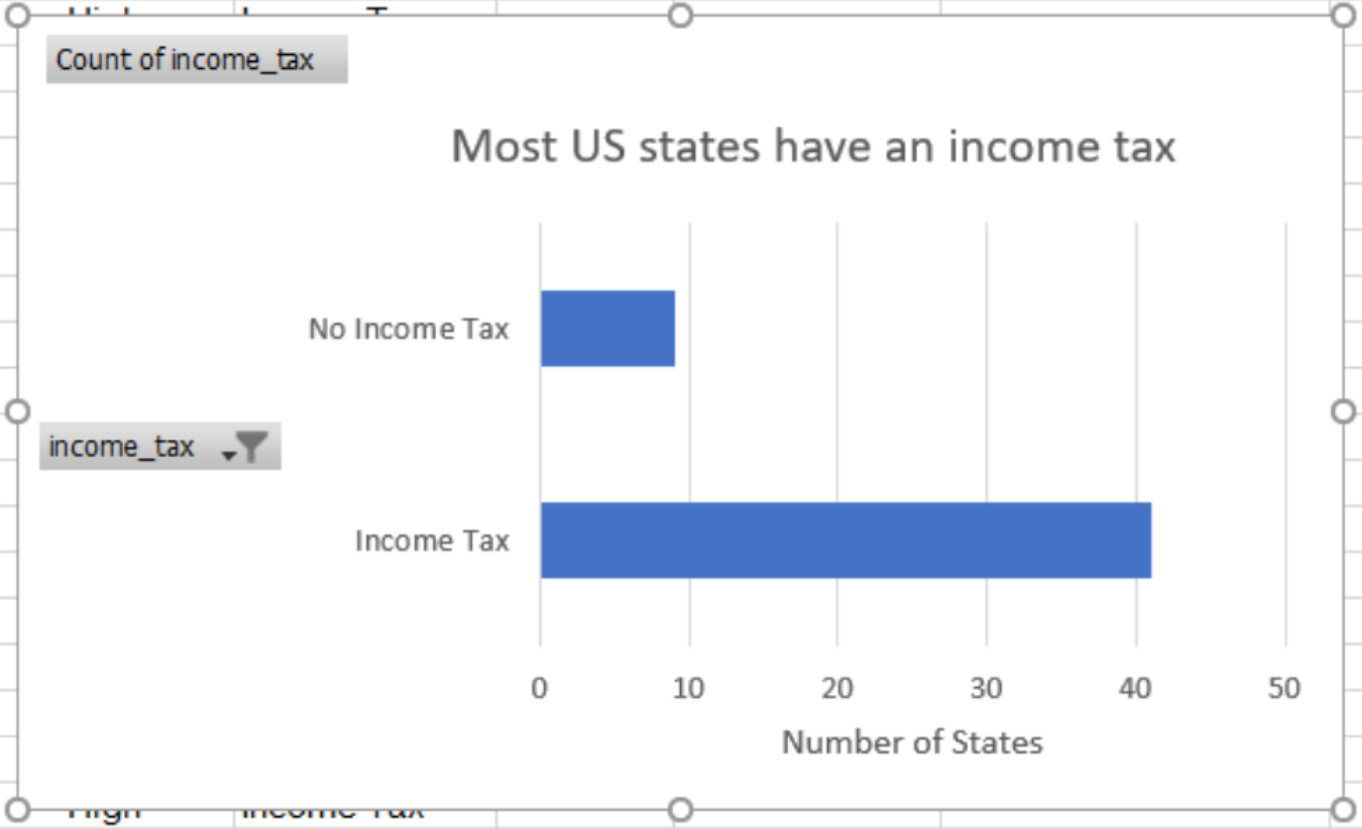


Figure 1. Across time data taken from the World Bank's WDI database makes clear that although life expectancies differ across the globe they are increasing across time in every one of these regions.

Practice Univariate Analysis of a Categorical Variable

1. Pivot table: Count the levels of income_tax
2. Insert a horizontal bar plot and polish the visualization

	B	C	D	E	F	G
1	year	gdp_category	income_tax		Row Labels	Count of income_tax
2	2018	Medium	Income Tax		Income Tax	41
3	2018	Low	No Income Tax		No Income Tax	9
4	2018	Medium	Income Tax		Grand Total	50
5	2018	Low	Income Tax			
6	2018					
7	2018					
8	2018					
9	2018					
10	2018					
11	2018					
12	2018					
13	2018					
14	2018					
15	2018					
16	2018					
17	2018					
18	2018					
19	2018					
20	2018					
21	2018					
22	2018					
23	2018					
24	2018	High	Income Tax			



PivotChart Fields

Choose fields to add to report:

Search

- state
- year
- gdp_category
- **income_tax**

Drag fields between areas below:

Filters

Legend (Series)

Axis (Categories)

income_tax

Values

Count of income_tax

Analyzing Categorical Variables: Tables and Bar Plots

Nominal (e.g. gender, ethnicity)

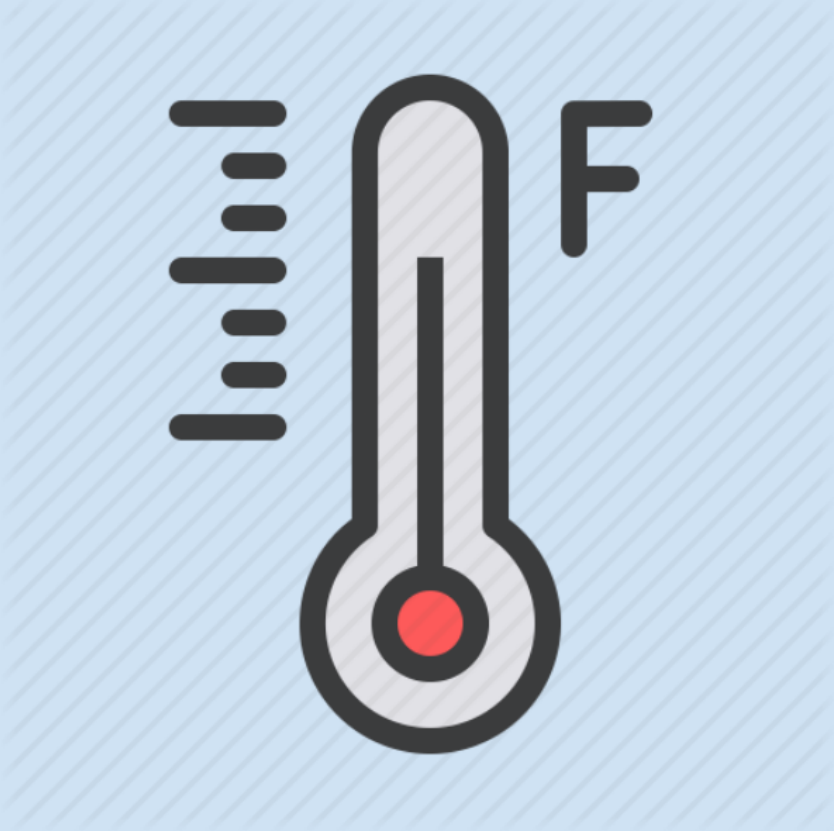
P
income_tax
Income Tax
No Income Tax
Income Tax
Income Tax
Income Tax
Income Tax
Income Tax
Income Tax
No Income Tax
Income Tax
Income Tax
Income Tax
Income Tax

Ordinal (e.g. rating scales)

F
gdp_category
Medium
Low
Medium
Low
High
High
Medium
Low
High
High
Low
Low

Analyzing Numerical Variables

Interval
(e.g. temperature)



Ratio
(e.g. zero matters)

G	H	I
min_wage	**unemployment**	**population**
0	3.9	4887.681
9.84	6.6	735.139
10.5	4.8	7158.024
8.5	3.7	3009.733
11	4.2	39461.588
10.2	3.3	5691.287
10.1	4.1	3571.52
8.25	3.8	965.479
8.25	3.6	21244.317
5.15	3.9	10511.131

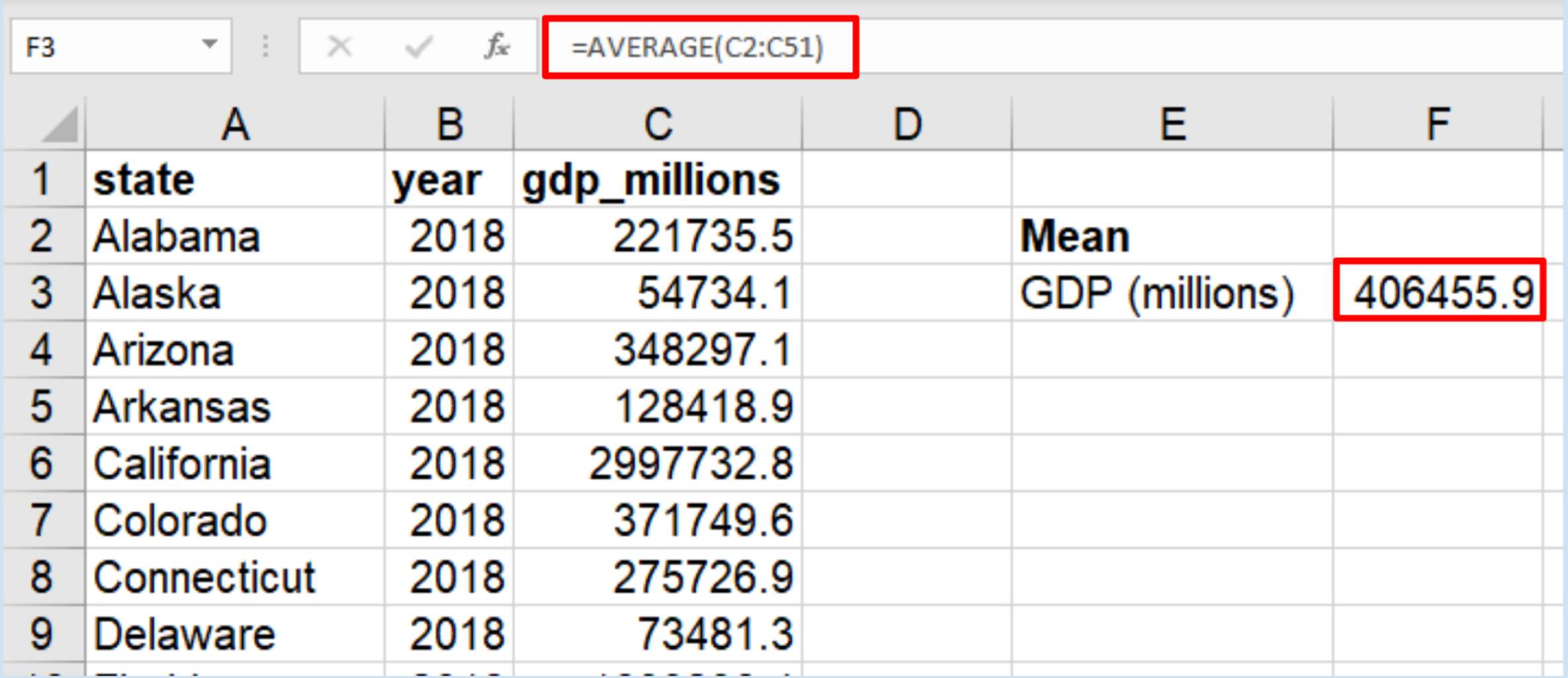
Analyzing Numerical Variables (Johnson 2012)

- Measures of Central Tendency
  - Mean
  - Median

- Deviations from Central Tendency
  - Standard deviation

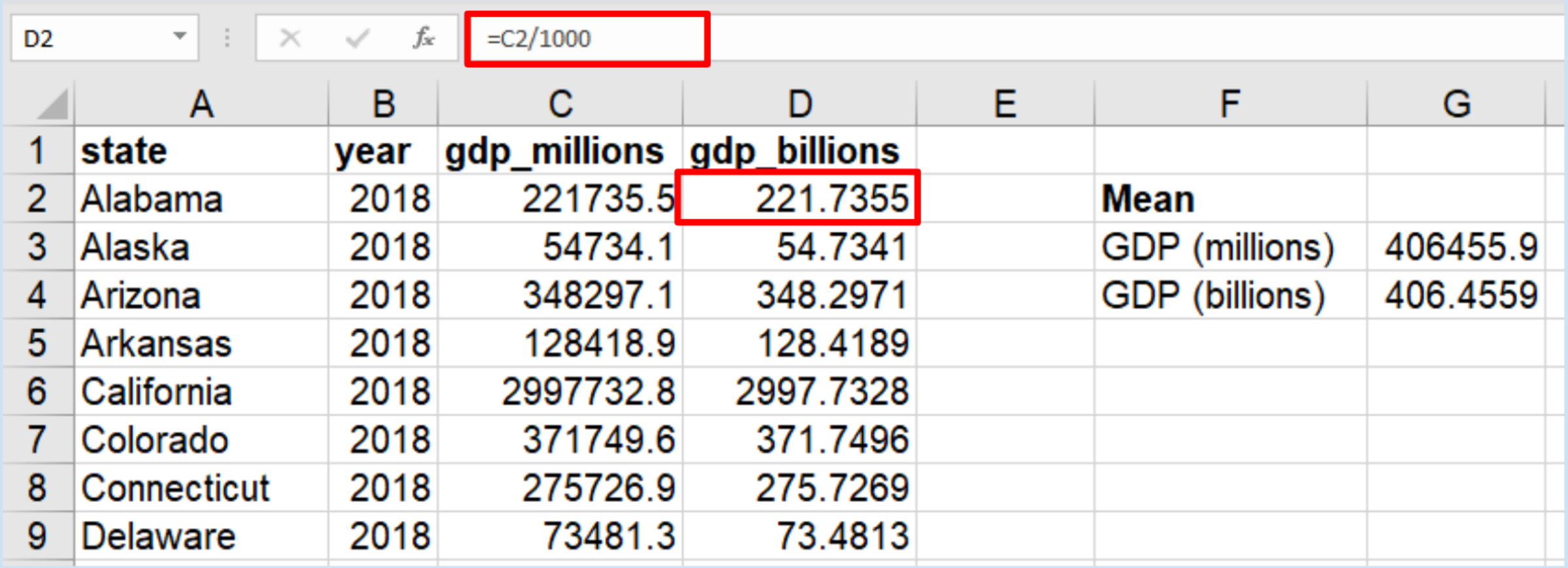
- Measures of Variability
  - Range = Maximum - Minimum
  - IQR = 75th - 25th percentile

Descriptive Statistics in Excel: Using Functions



	A	B	C	D	E	F
1	**state**	**year**	**gdp_millions**			
2	Alabama	2018	221735.5		**Mean**	
3	Alaska	2018	54734.1		GDP (millions)	406455.9
4	Arizona	2018	348297.1			
5	Arkansas	2018	128418.9			
6	California	2018	2997732.8			
7	Colorado	2018	371749.6			
8	Connecticut	2018	275726.9			
9	Delaware	2018	73481.3			

Descriptive Statistics in Excel: Using Functions



	A	B	C	D	E	F	G
1	**state**	**year**	**gdp_millions**	**gdp_billions**			
2	Alabama	2018	221735.5	221.7355		**Mean**	
3	Alaska	2018	54734.1	54.7341		GDP (millions)	406455.9
4	Arizona	2018	348297.1	348.2971		GDP (billions)	406.4559
5	Arkansas	2018	128418.9	128.4189			
6	California	2018	2997732.8	2997.7328			
7	Colorado	2018	371749.6	371.7496			
8	Connecticut	2018	275726.9	275.7269			
9	Delaware	2018	73481.3	73.4813			

Statistic	Function
Mean	= AVERAGE
Median	= MEDIAN
Standard Deviation	= STDEV.S
Minimum	= MIN
Maximum	= MAX
25th Percentile	= QUARTILE.EXC(, 1)
75th Percentile	= QUARTILE.EXC(, 3)

Practice: Descriptive stats for GDP (billions)

Variable	gdp_billions
Mean	406.5
Median	236.9
StdDev	526.7
Minimum	33.3
Maximum	2997.7
25th Percentile	95
75th Percentile	531.4

Univariate Analyses: Numerical Variables

- min_wage
- unemployment
- population
- bachelors
- homeowner_rate
- manufacturing

Statistic	Function
Mean	= AVERAGE
Median	= MEDIAN
Standard Deviation	= STDEV.S
Minimum	= MIN
Maximum	= MAX
25th Percentile	= QUARTILE.EXC(, 1)
75th Percentile	= QUARTILE.EXC(, 3)

Practice: Descriptive Statistics

Variable	Mean	StdDev	Min	pct25	Median	pct75	Max
bachelors	31.6	5.3	21.3	28	30.9	34.8	44.5
homeowner_rate	66.7	4.7	51.0	65	67.3	69.8	74.7
manufacturing	253.3	256.5	9.8	59	167.6	343.4	1325.4
min_wage	7.7	3.0	0.0	7	8.2	10.0	11.5
population	6519.7	7355.8	577.6	1835	4560.4	7432.4	39461.6
unemployment	3.8	0.8	2.4	3	3.8	4.2	6.6

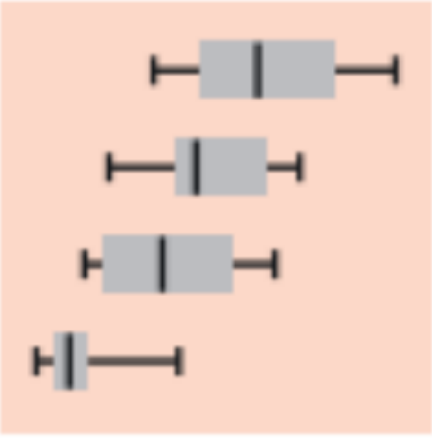
Variable	Mean	StdDev	Min	pct25	Median	pct75	Max
bachelors	31.6	5.3	21.3	28	30.9	34.8	44.5
homeowner_rate	66.7	4.7	51.0	65	67.3	69.8	74.7
manufacturing	253.3	256.5	9.8	59	167.6	343.4	1325.4
min_wage	7.7	3.0	0.0	7	8.2	10.0	11.5
population	6519.7	7355.8	577.6	1835	4560.4	7432.4	39461.6
unemployment	3.8	0.8	2.4	3	3.8	4.2	6.6

Identify the **TWO** variables with the **LEAST** variation across the states.

Identify the **TWO** variables with the **MOST** variation across the states.

Visualizing Numerical Variables

Boxplot



Summarise multiple distributions by showing the median (centre) and range of the data

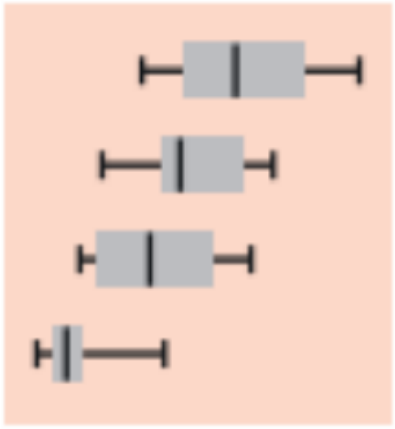
Histogram

The standard way to show a statistical distribution - keep the gaps between columns small to highlight the 'shape' of the data.

Moving from descriptive stats to a box plot to a visualization of the entire distribution.

Visualizing Numerical Variables

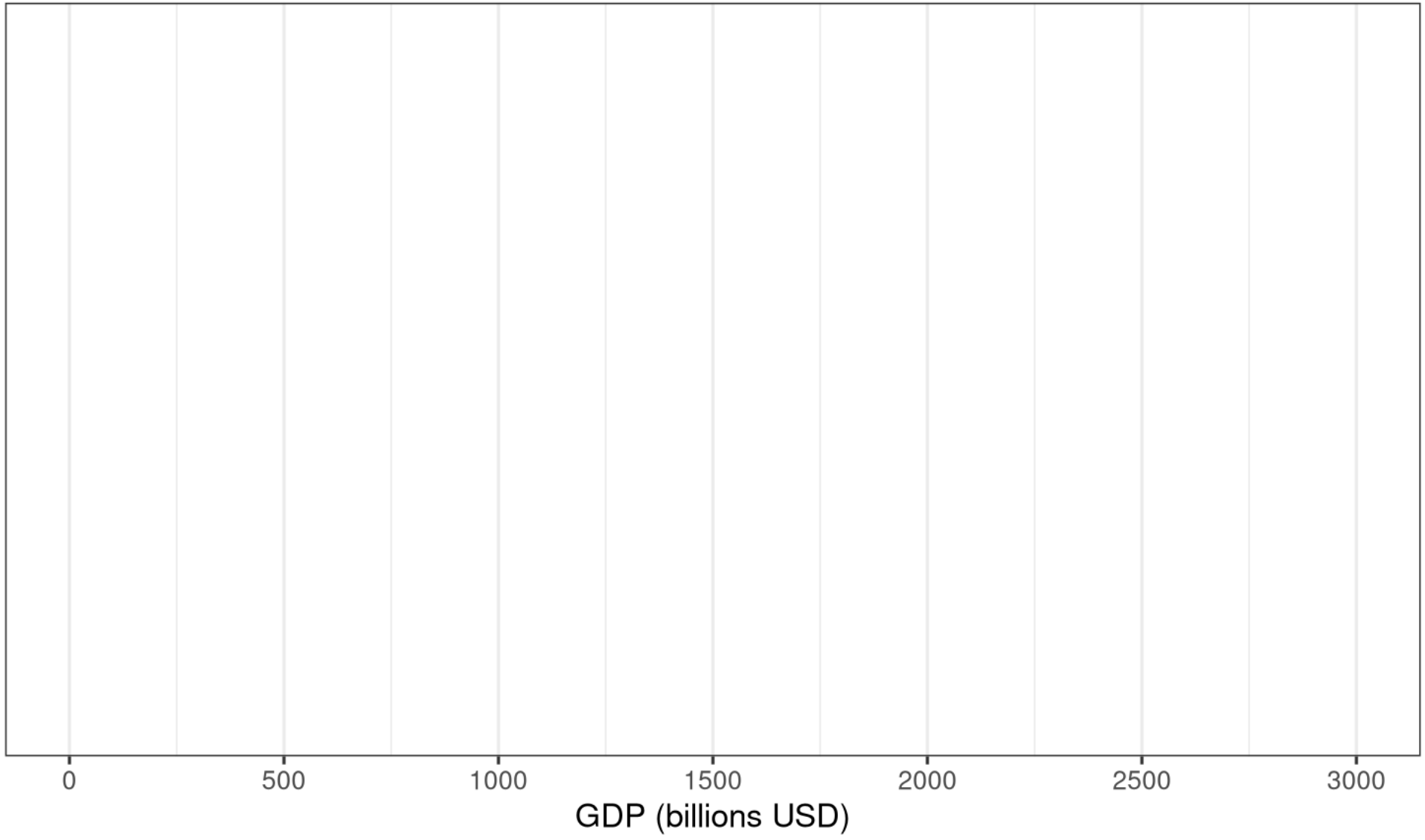
Boxplot



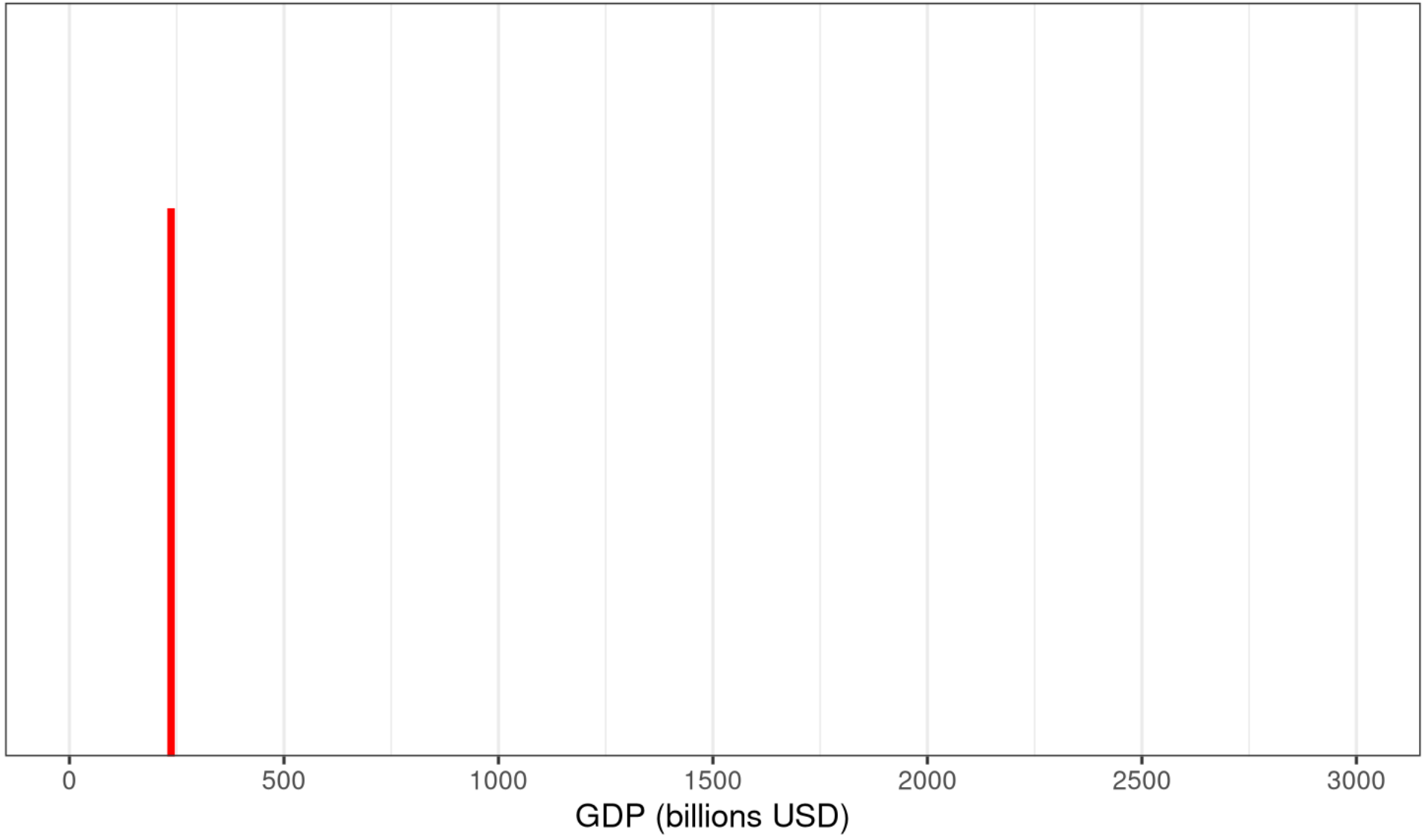
Summarise multiple distributions by showing the median (centre) and range of the data

Use **ONLY** the descriptive statistics for **GDP (billions)** to draw a boxplot **by hand**.

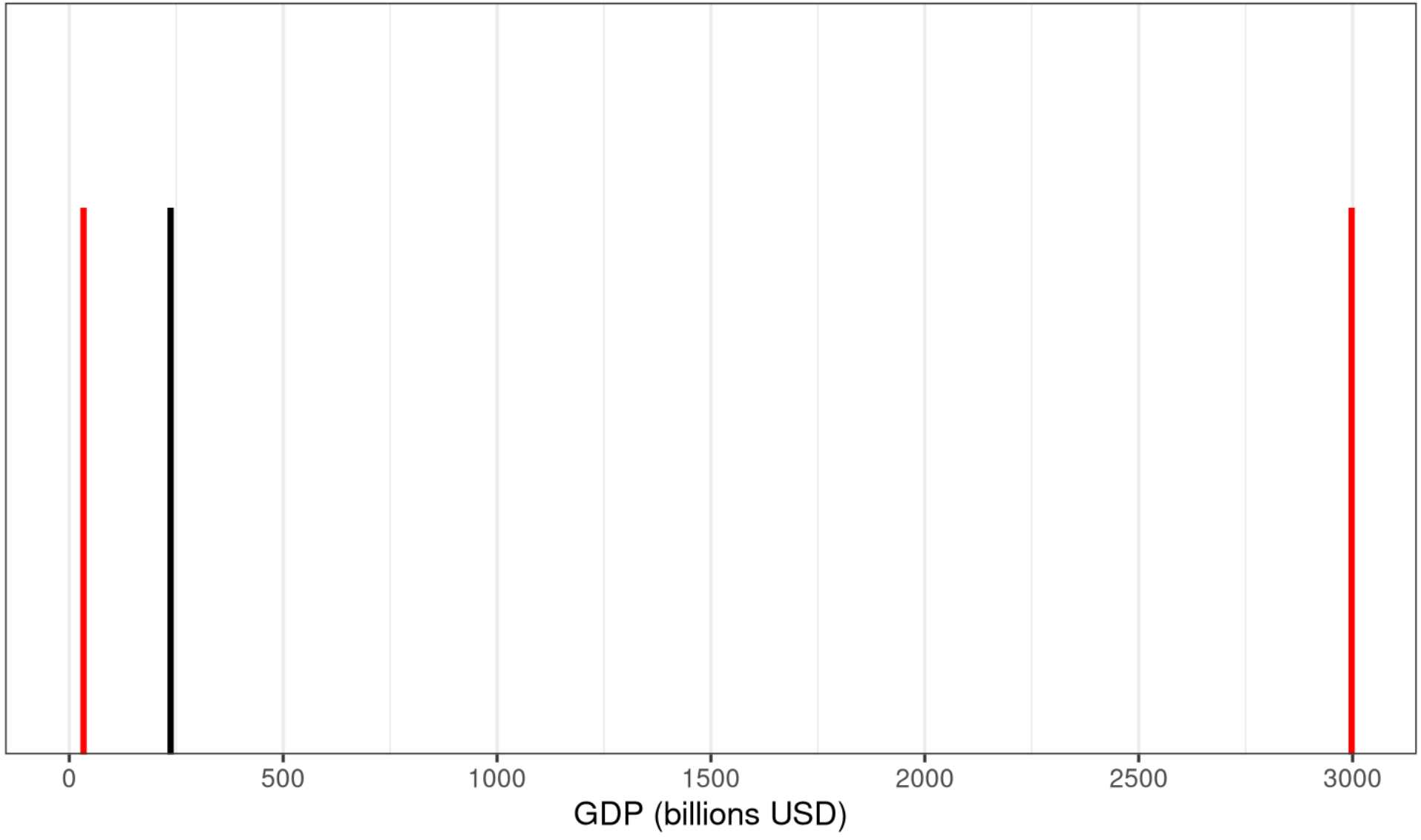
Set the X-Axis Range



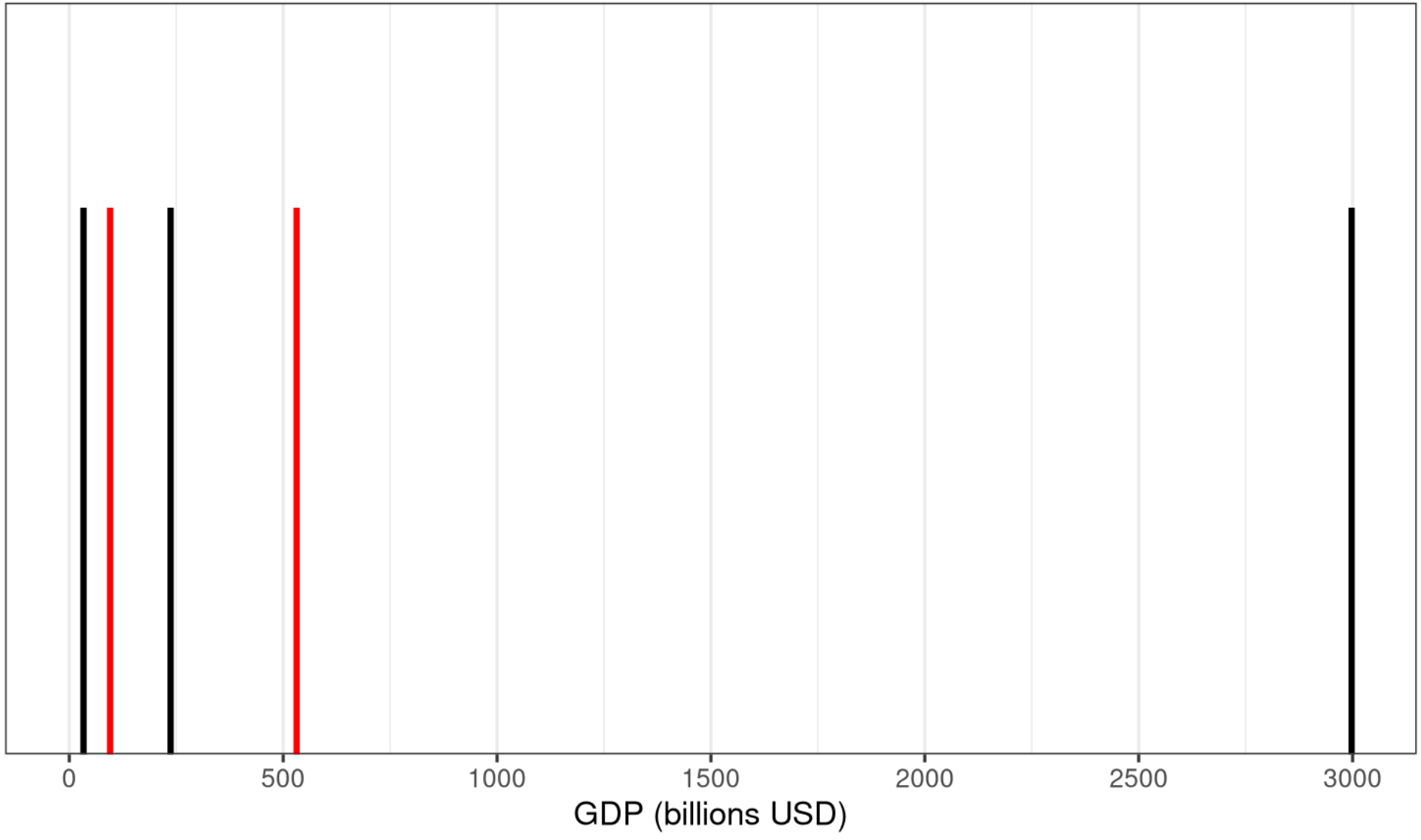
Mark the median (50th percentile)

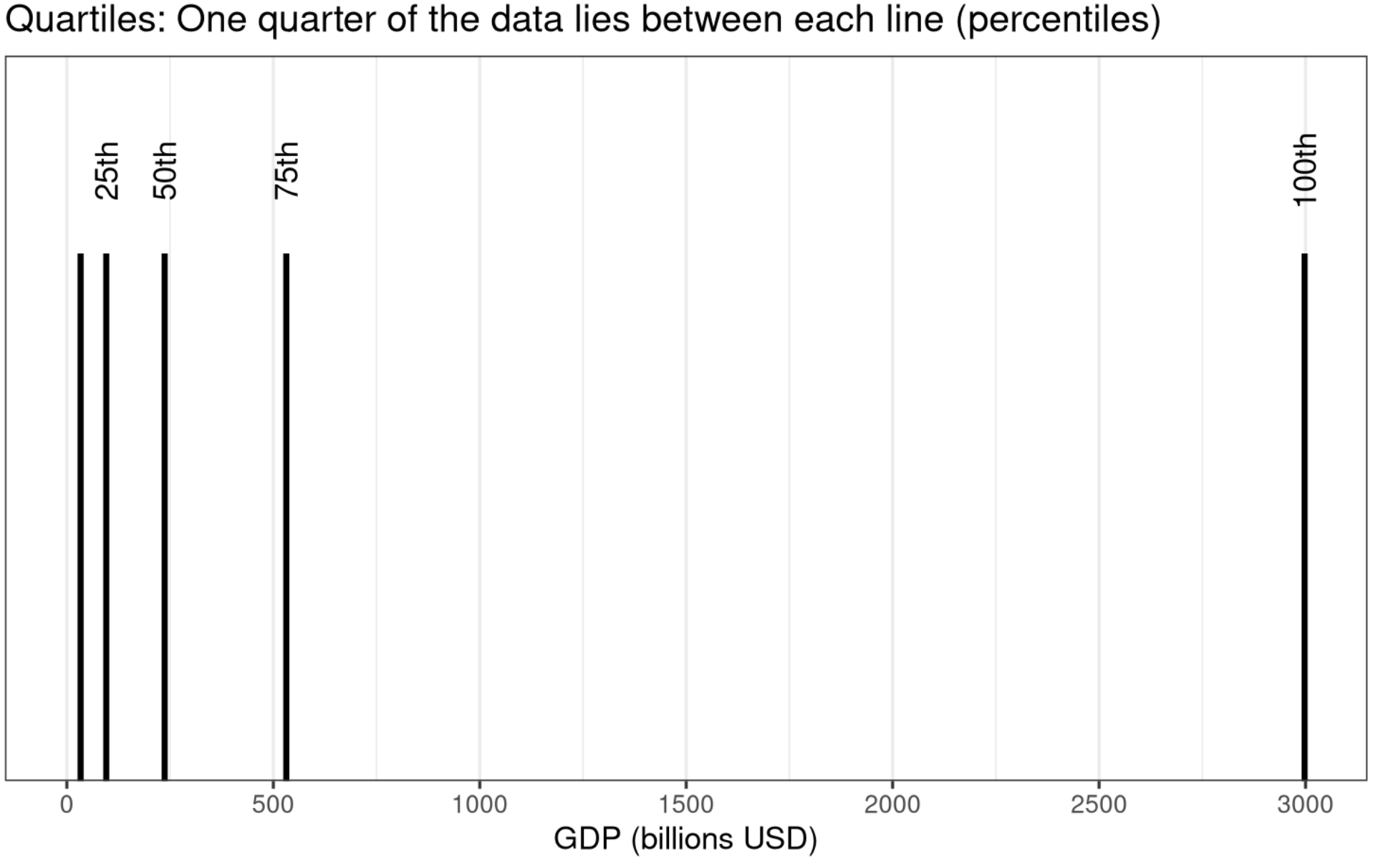


Add the minimum and maximum values

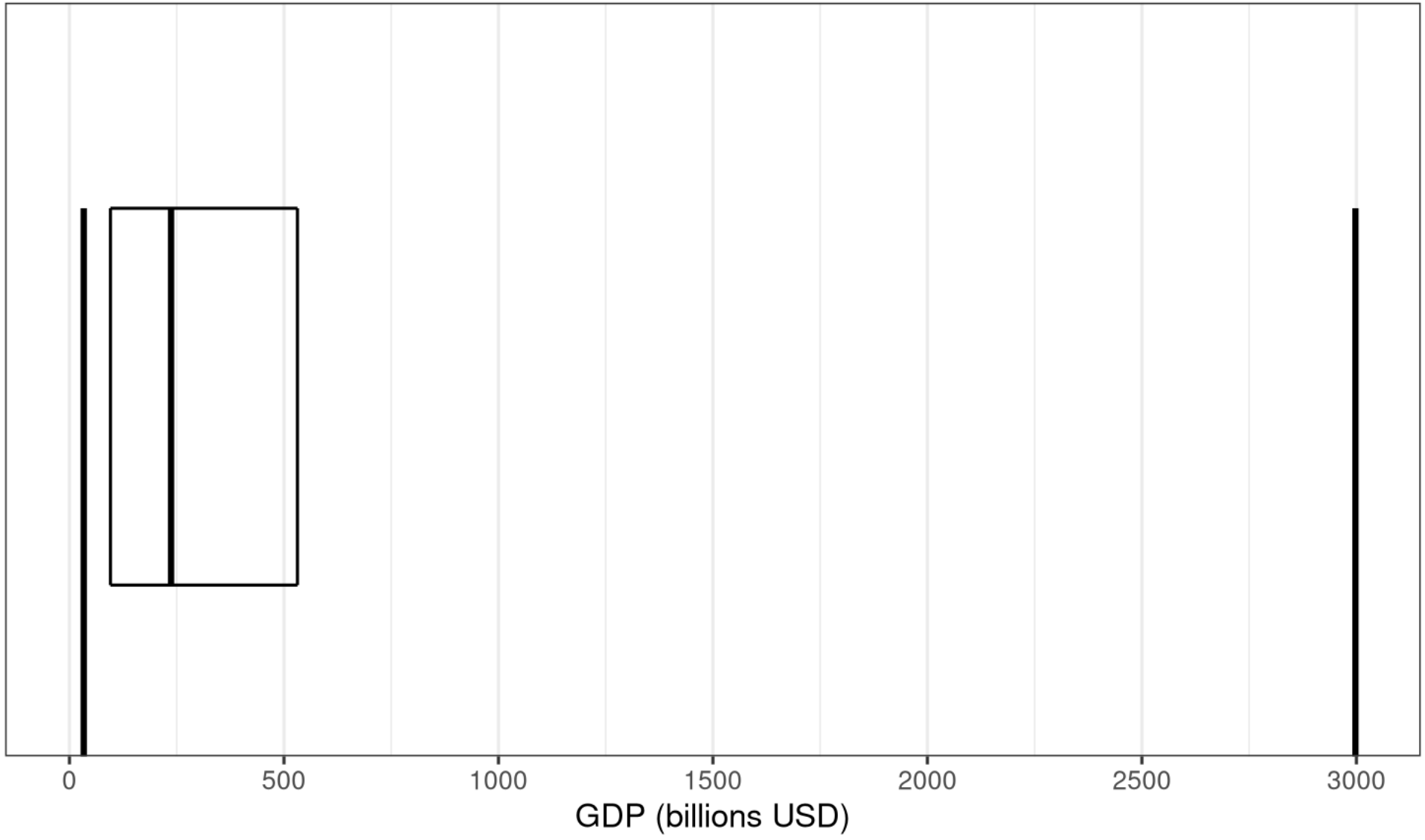


Add the 25th and 75th percentiles

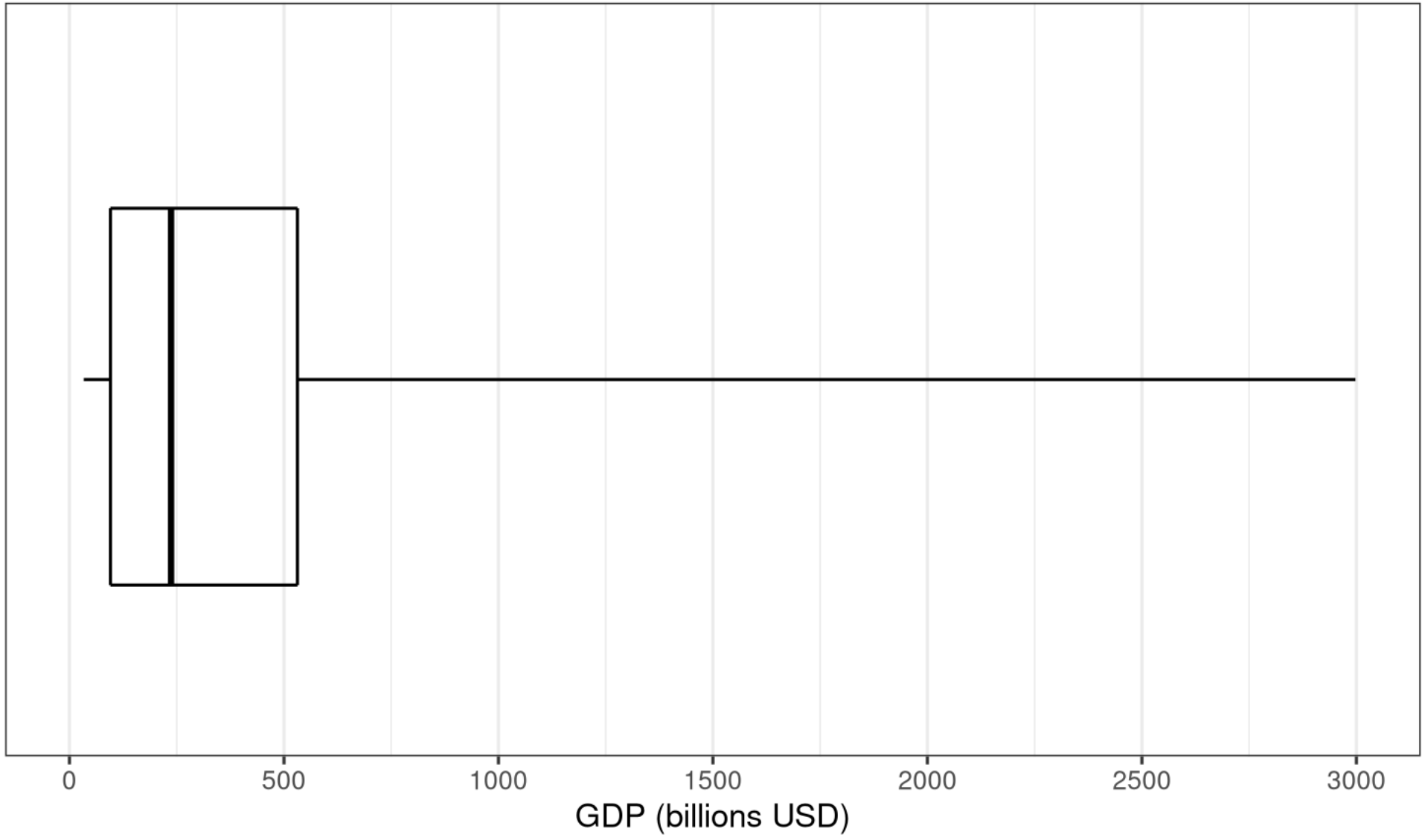



Quartiles: One quarter of the data lies between each line (percentiles)
25th
50th
75th
100th
0
500
1000
1500
2000
2500
3000
GDP (billions USD)

Replace IQR with a box

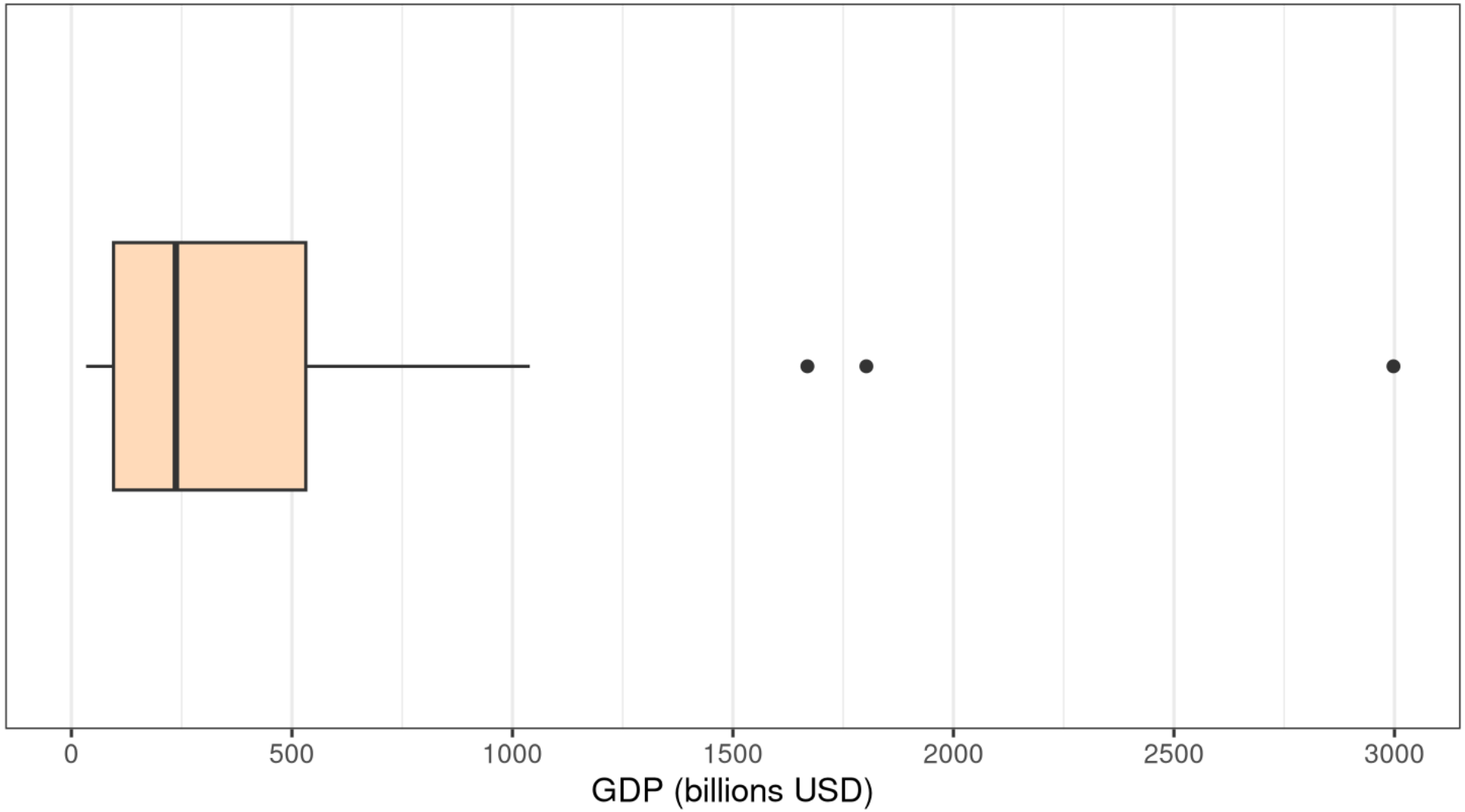


Replace min and max with 'whiskers'



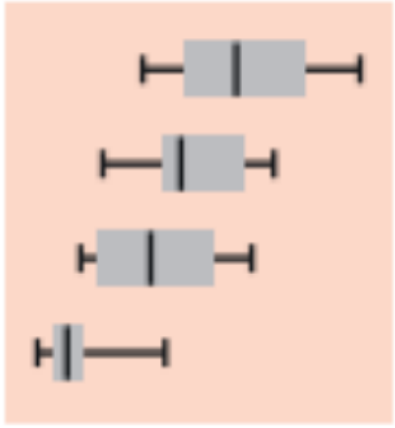
Technically the whiskers should only extend 1.5x the IQR

(With outliers represented as points)



Visualizing Numerical Variables

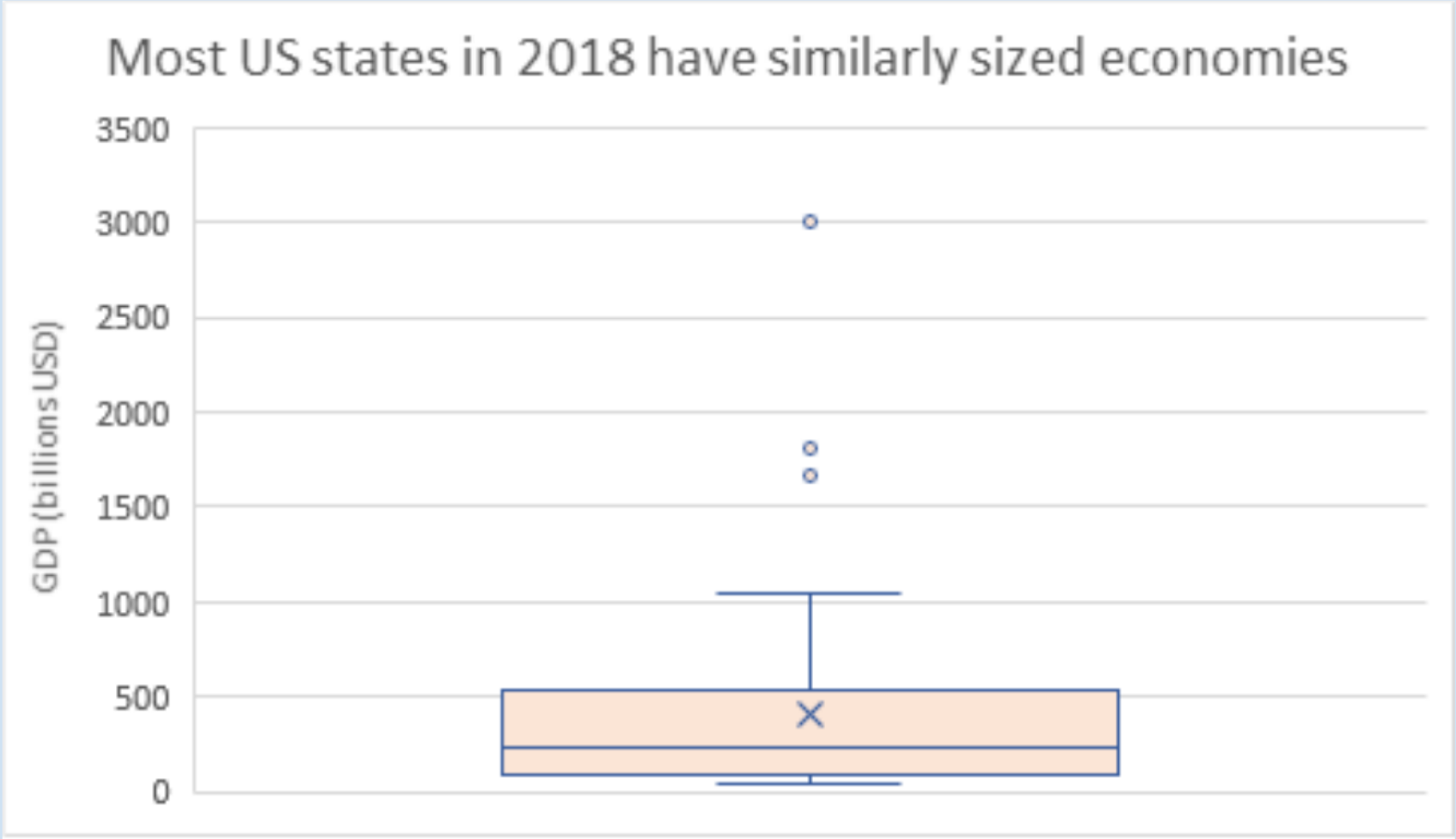
Boxplot



Summarise multiple distributions by showing the median (centre) and range of the data

Use **Excel** to make a **boxplot** of **GDP**

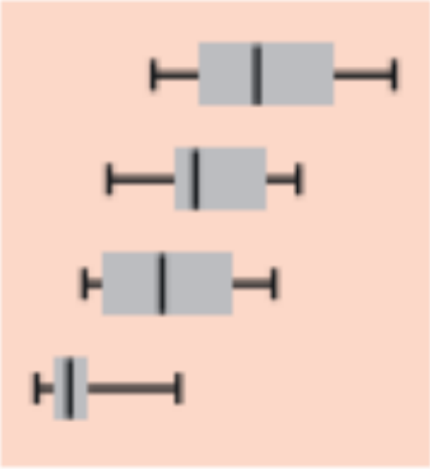
Univariate Visualizations of GDP (billions)



Notes: Changed fill color, added an axis label and increased font size.

Visualizing Numerical Variables

Boxplot



Summarise multiple distributions by showing the median (centre) and range of the data

Histogram

The standard way to show a statistical distribution - keep the gaps between columns small to highlight the 'shape' of the data.

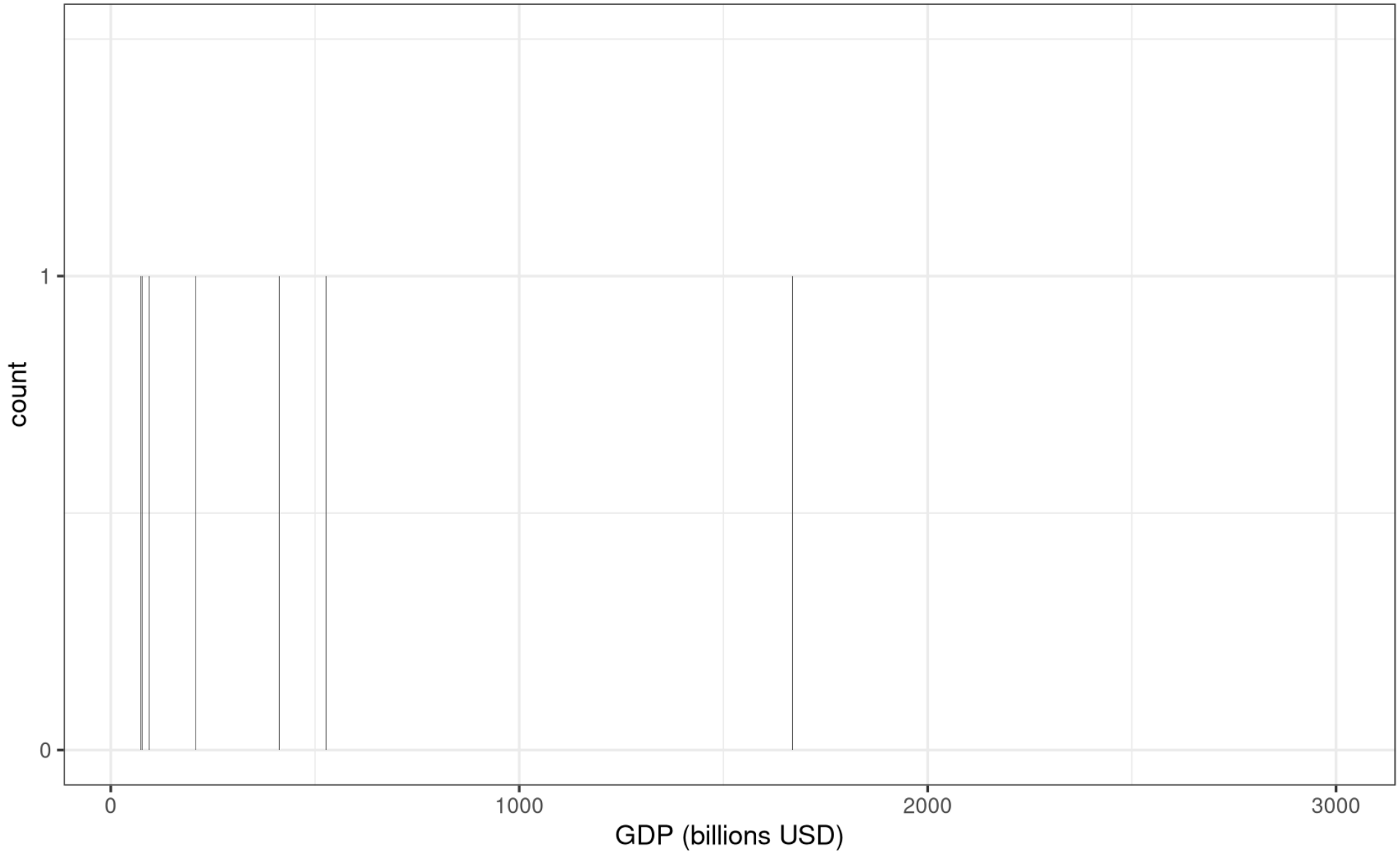
Converting a box plot to a histogram
(visualizing the entire distribution)

Building a Histogram

GDP	33.3	39.1	50.3	52	54.7	56.1	60.6	64.9	73.5	77.1	77.4	84.5	93.8	100.3	114.8	124	128.4	168.3	169.3
n	1.0	1.0	1.0	1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1	1.0	1.0	1.0

GDP	178.1	189.7	202.6	208.1	221.7	233.9	239.8	257.3	275.7	318.9	336.3	348.3	364.1	366.8	368.9	371.7
n	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

GDP	412.6	527.1	532.9	563.7	565.8	569.5	592.2	622	675.9	783.2	865.3	1039.2	1668.9	1802.5	2997.7
n	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1	1.0	1.0	1.0	1.0	1.0	1.0	1.0


count
1
0
0
1000
2000
3000
GDP (billions USD)

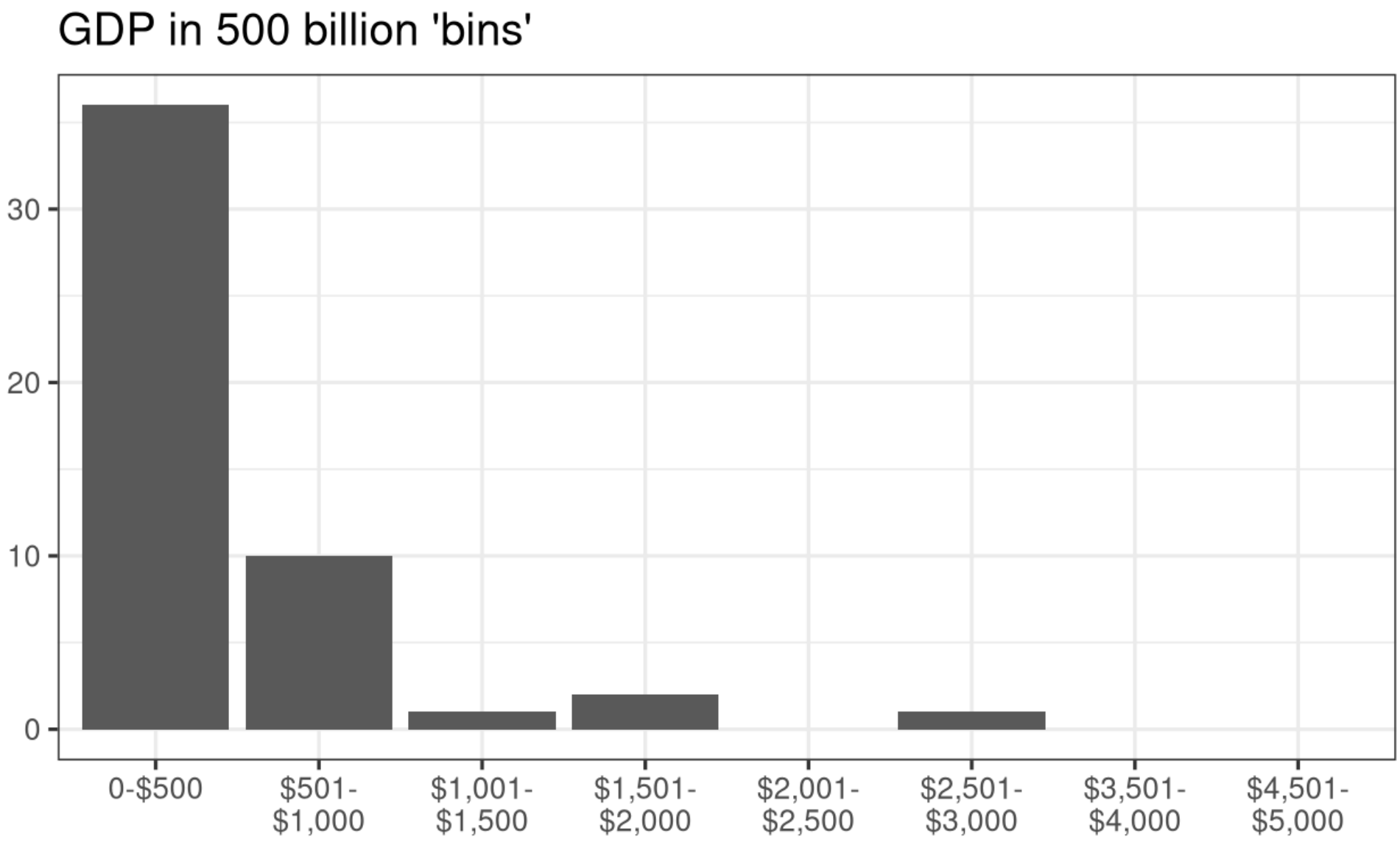
Building a Histogram

GDP	33.3	39.1	50.3	52	54.7	56.1	60.6	64.9	73.5	77.1	77.4	84.5	93.8	100.3	114.8	124	128.4	168.3	169.3
n	1.0	1.0	1.0	1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1	1.0	1.0	1.0

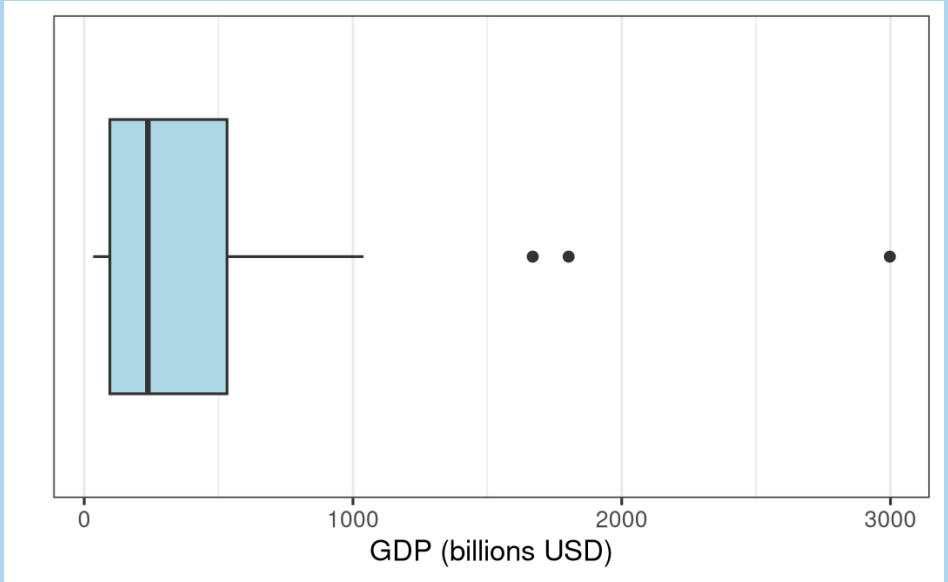
GDP	178.1	189.7	202.6	208.1	221.7	233.9	239.8	257.3	275.7	318.9	336.3	348.3	364.1	366.8	368.9	371.7
n	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

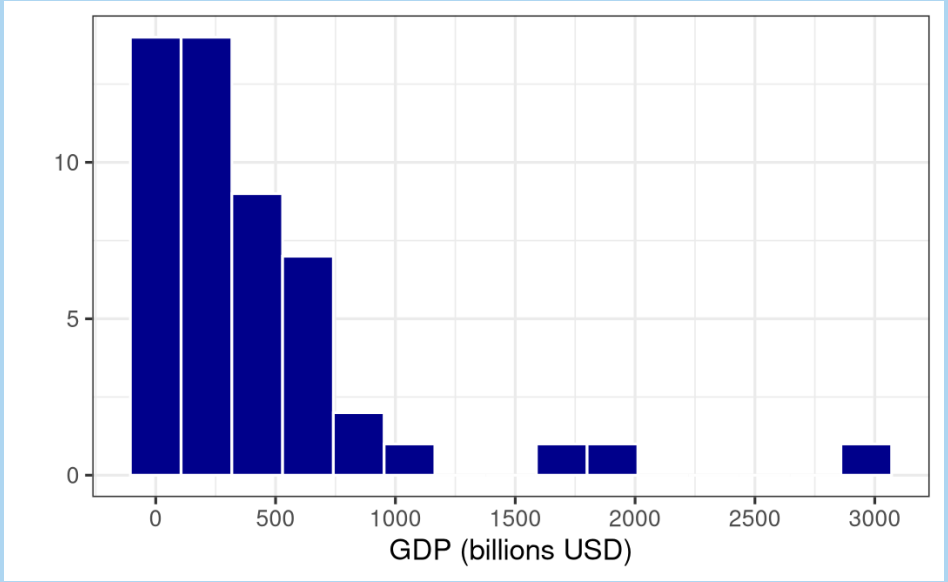
GDP	412.6	527.1	532.9	563.7	565.8	569.5	592.2	622	675.9	783.2	865.3	1039.2	1668.9	1802.5	2997.7
n	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Create "bins" of $500 (e.g. 0-500, 501-1,000, etc.), remake the table and make a new bar chart by hand.


GDP in 500 billion 'bins'
30
20
10
0
0-$500
$501-
$1,000
$1,001-
$1,500
$1,501-
$2,000
$2,001-
$2,500
$2,501-
$3,000
$3,501-
$4,000
$4,501-
$5,000

Variable	gdp_billions
Mean	406.5
StdDev	526.7
Minimum	33.3
25th Percentile	95
Median	236.9
75th Percentile	531.4
Maximum	2997.7

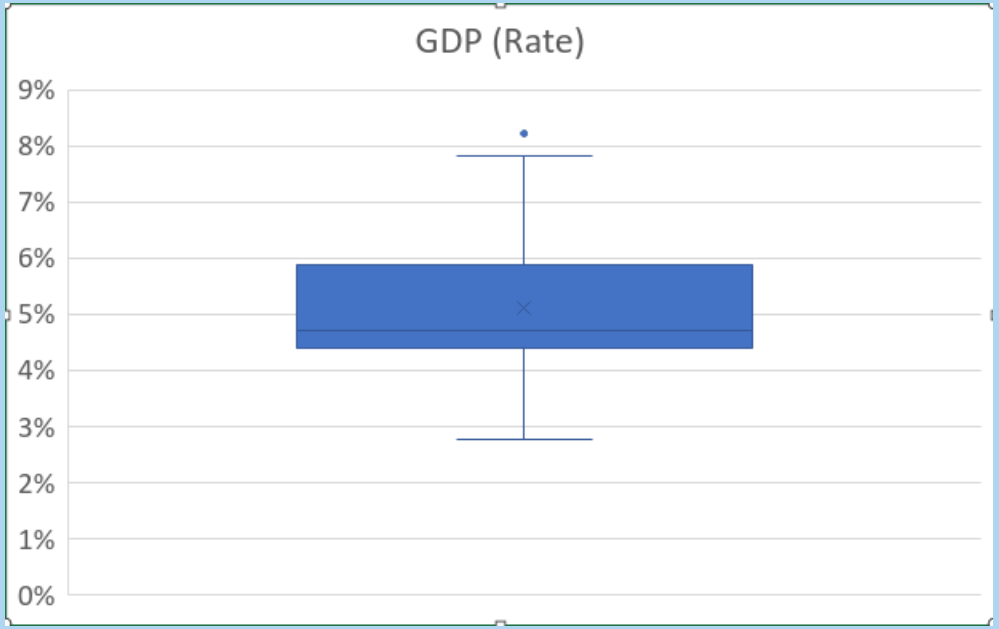


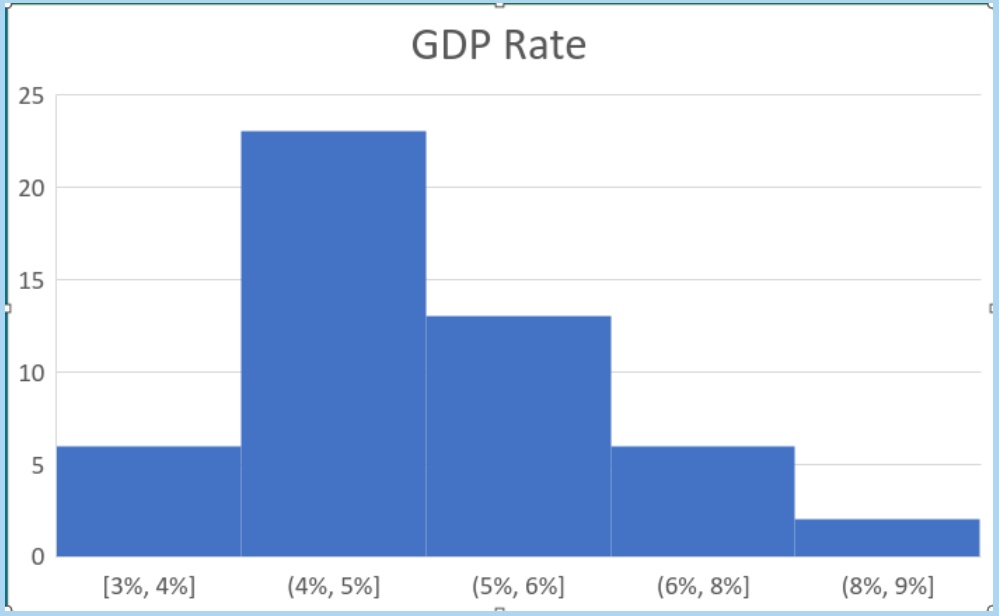


Univariate Analyses: GDP (rate)

1. Use Excel to calculate the descriptive stats
    - Mean, StdDev, Minimum, 25th pct, Median, 75th pct, Maximum
2. Use Excel to make a box plot
3. Use Excel to make a histogram

Variable	gdp_rate
Mean	0.051
StdDev	0.012
Minimum	0.028
25th Percentile	0
Median	0.047
75th Percentile	0.058
Maximum	0.082





Univariate Analyses

1. Descriptive Statistics
2. Bar plots (categorical variables)
3. Box plots (numeric variables)
4. Histograms (numeric variables)

Bivariate Analyses

Analyzing pairs of variables for associations / relationships

Remember: Variable type dictates tool selection

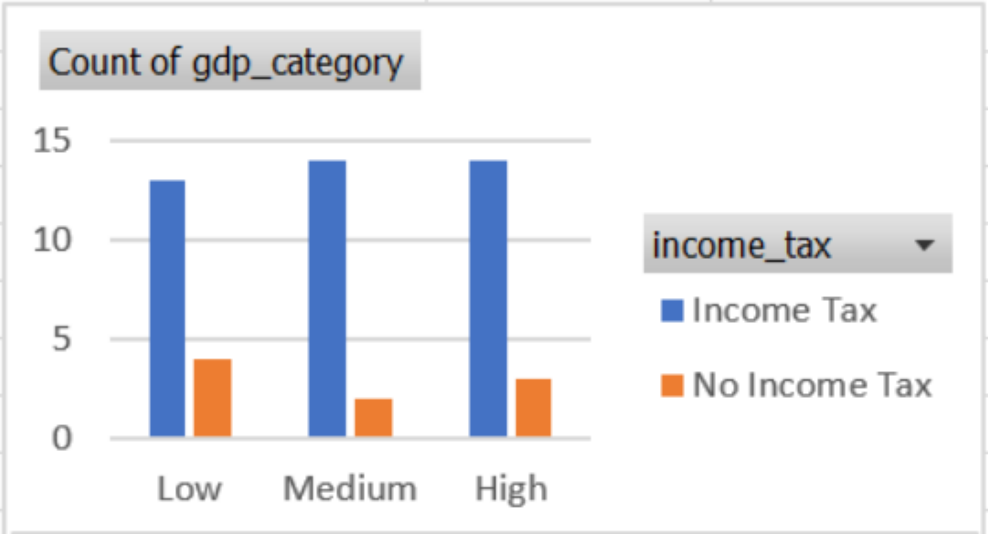
Bivariate Visualizations: Categorical x Categorical

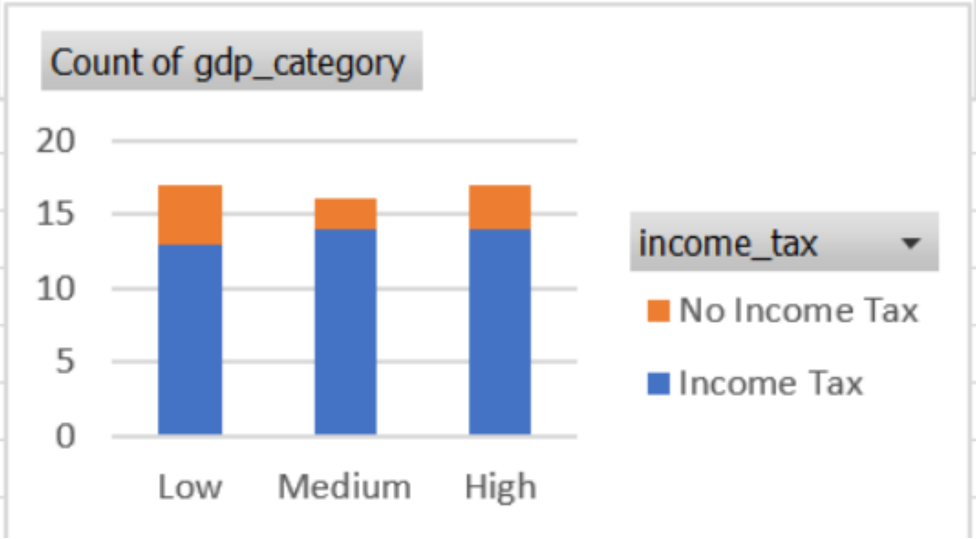
1. Pivot table: Count gdp_category levels (rows) by the levels of income_tax (columns)

2. Plot 1: Insert a clustered column bar chart

3. Plot 2: Insert a stacked column bar chart

Count of gdp_category	Column Lab	
Row Labels	Income Tax	No Income Tax
Low	13	4
Medium	14	2
High	14	3





PivotTable Fields

Choose fields to add to report:

Search

- state
- year
- gdp_category
- income_tax

More Tables...

Drag fields between areas below:

Filters	Columns
	income_tax
Rows	**Values**
gdp_category	Count of gdp...

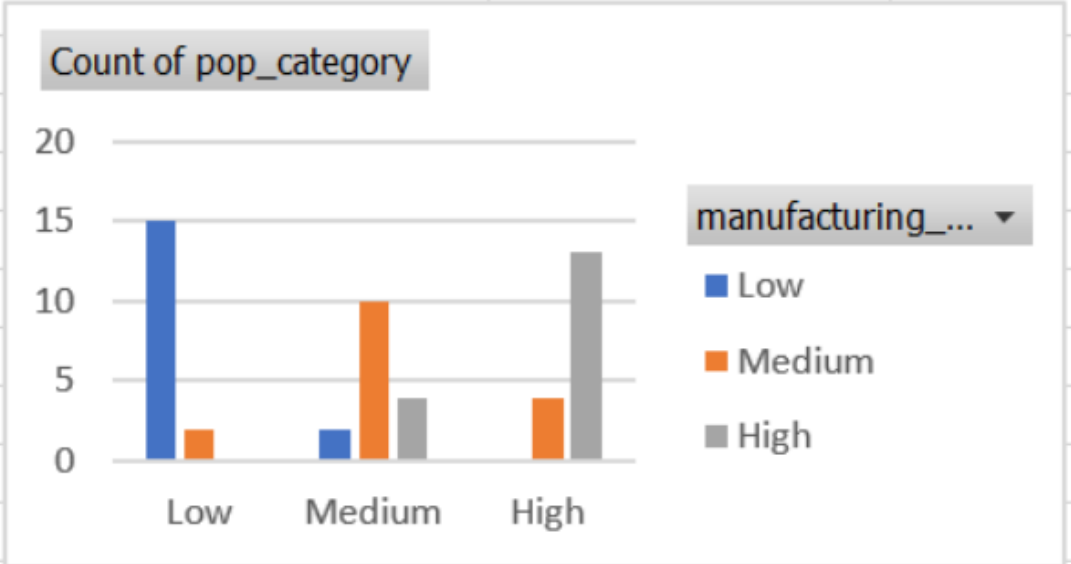
Bivariate Visualizations: Categorical x Categorical

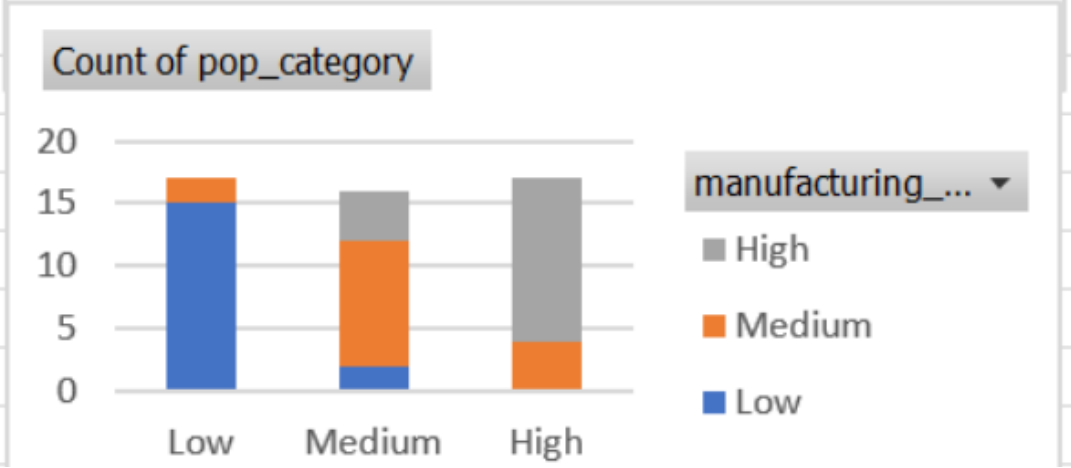
1. Pivot table: Count pop_category (rows) levels by the manufacturing_category (columns)

2. Plot 1: Insert a clustered column bar chart

3. Plot 2: Insert a stacked column bar chart

Count of pop_category	Column Labels		
Row Labels	Low	Medium	High
Low	15	2	
Medium	2	10	4
High		4	13





PivotTable Fields

Choose fields to add to report:

Search

- [ ] state
- [ ] year
- [x] **pop_category**
- [x] **manufacturing_category**

More Tables...

Drag fields between areas below:

Filters	Columns
	manufacturing_categ...
Rows	**Values**
pop_category	Count of pop_category

Bivariate Analyses: Numerical x Categorical

Example:

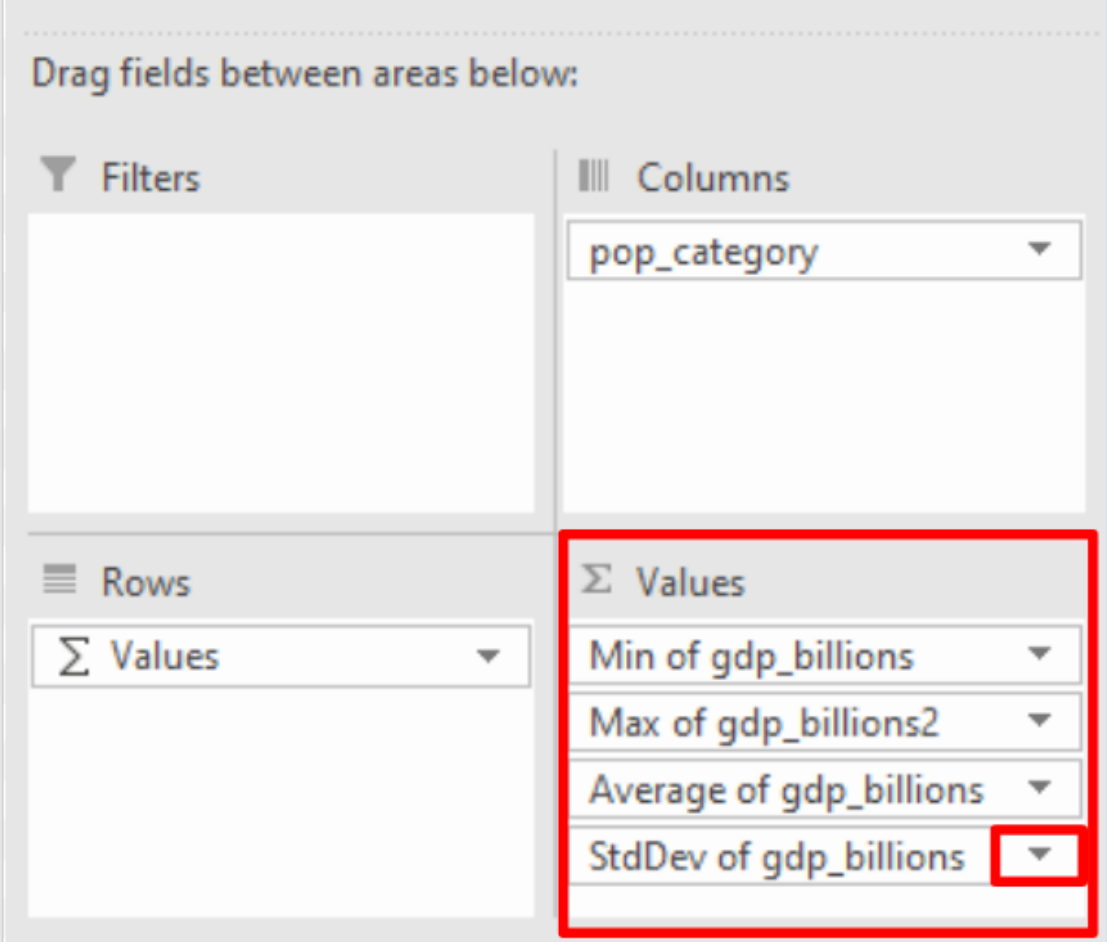
Analyze the size of state economies by the size of their populations

Bivariate Analyses: Numerical x Categorical

1. Convert GDP into billions of USD

2. Pivot table: Calculate values for GDP (billions) by the levels of pop_category

    - Minimum
    - Maximum
    - Mean
    - Std Deviation



1. In the 'Values' window, add gdp_billions four times

2. Edit the 'Value Field Settings' to select each descriptive statistic

	Population Size		
GDP (billions)	**Low**	**Medium**	**High**
Minimum	33.3	128.4	348.3
Maximum	168.3	412.6	2997.7
Mean	77.9	257.1	875.6
Std Deviation	34.4	82.8	686.3

Bivariate Analyses: Numerical x Categorical

Box Plots

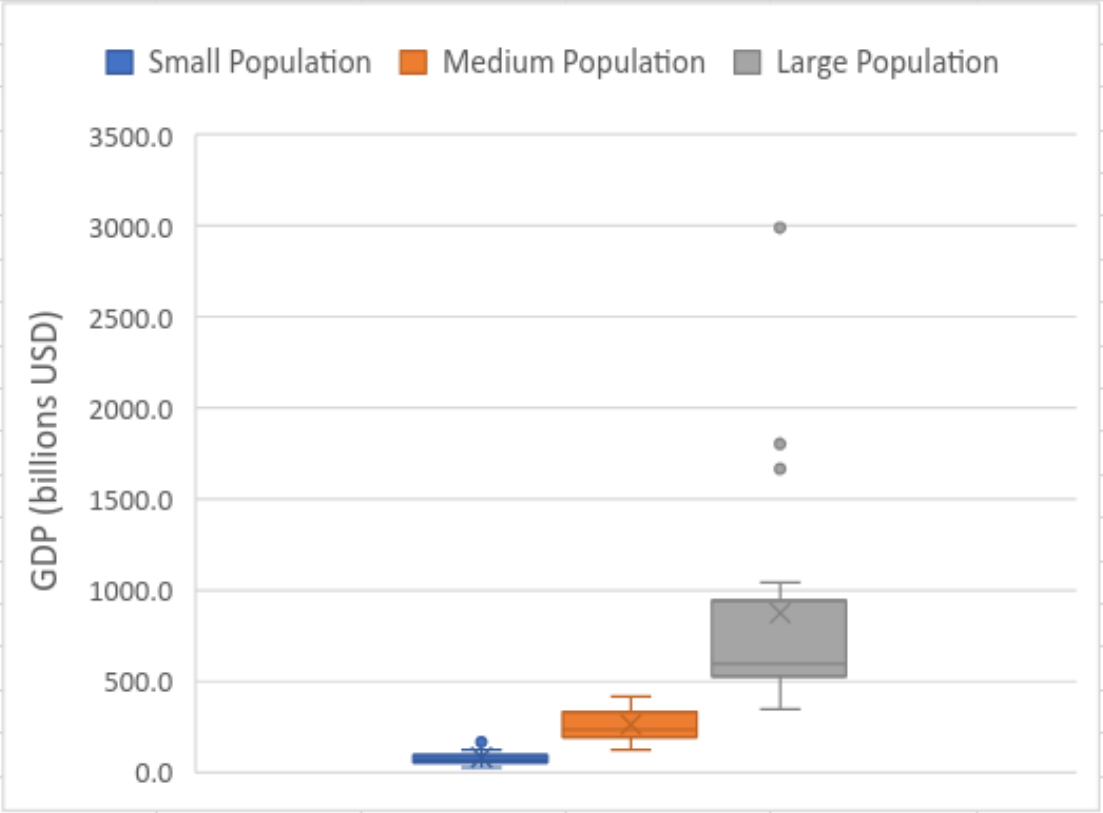
1. Sort the data by the categorical variable

2. Make new columns for each level of the categories

3. Highlight new columns and insert box plot

	B	C	D	E	F	G	H	I
1								
2		year	gdp_millions	gdp_billions	pop_category	Small Population	Medium Population	Large Population
3		2018	348297.1	348.3	High	54.7	221.7	348.3
4		2018	2997732.8	2997.7	High	73.5	128.4	2997.7
5		2018	1039236.4	1039.2	High	93.8	371.7	1039.2
6		2018	592153.4	592.2	High	77.1	275.7	592.2
7		2018	865310.4	865.3	High	168.3	189.7	865.3
8		2018	366800.5	366.8	High	64.9	208.1	366.8
9	ts	2018	569488	569.5	High	114.8	257.3	569.5
10		2018	527095.8	527.1	High	50.3	412.6	527.1
11		2018	622002.8	622.0	High	124.0	368.9	622.0
12		2018	1668866.2	1668.9	High	84.5	318.9	1668.9
13	a	2018	563690.5	563.7	High	100.3	169.3	563.7
14		2018	675905.2	675.9	High	56.1	202.6	675.9
15		2018	783167.8	783.2	High	60.6	239.8	783.2
16		2018	364104.9	364.1	High	52.0	233.9	364.1
17		2018	1802511.2	1802.5	High	33.3	178.1	1802.5
18		2018	532892.5	532.9	High	77.4	336.3	532.9
19		2018	565831	565.8	High	39.1		565.8
20		2018	54734.1	54.7	Low			
21		2018	73481.3	73.5	Low			

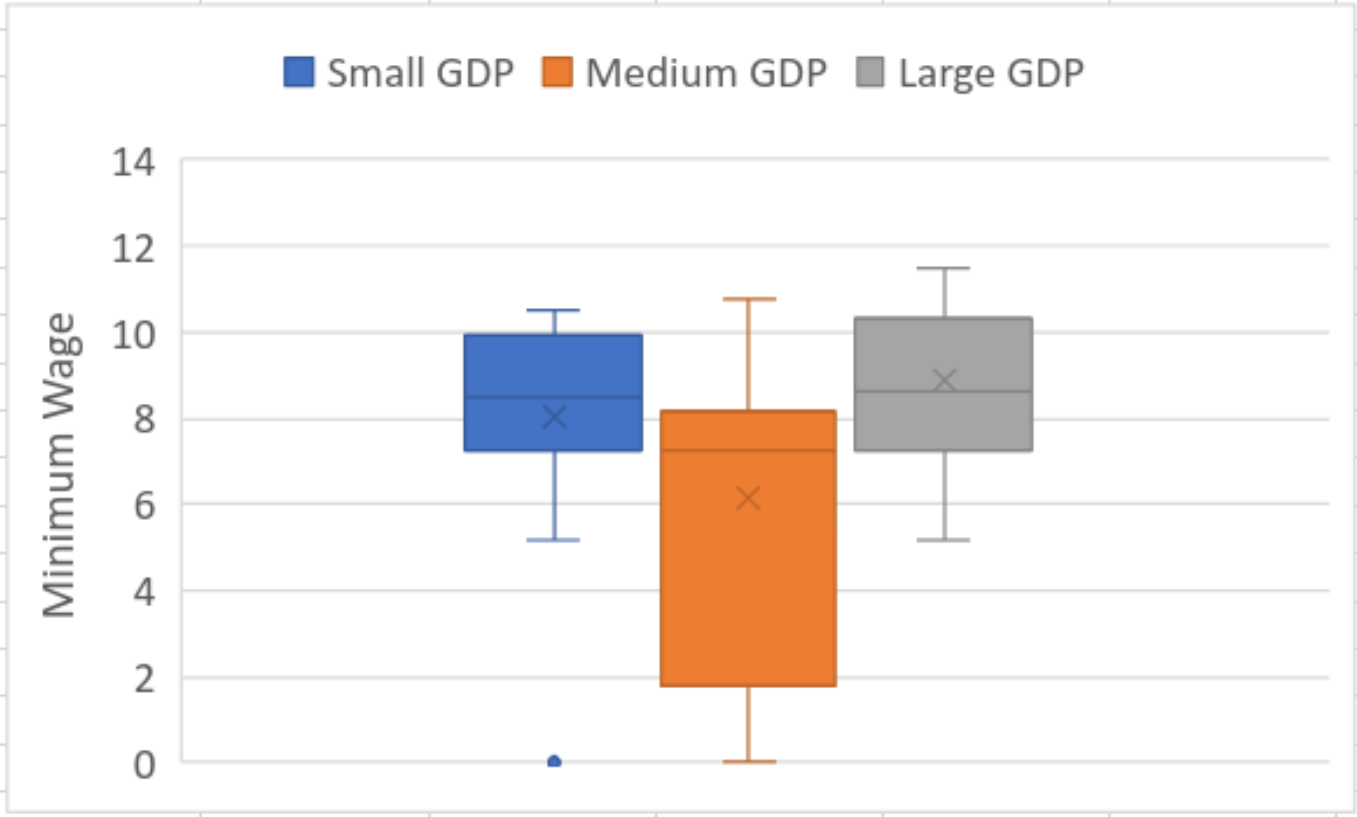
Small Population	Medium Population	Large Population
54.7	221.7	348.3
73.5	128.4	2997.7
93.8	371.7	1039.2
77.1	275.7	592.2
168.3	189.7	865.3
64.9	208.1	366.8
114.8	257.3	569.5
50.3	412.6	527.1
124.0	368.9	622.0
84.5	318.9	1668.9
100.3	169.3	563.7
56.1	202.6	675.9
60.6	239.8	783.2
52.0	233.9	364.1
33.3	178.1	1802.5
77.4	336.3	532.9
39.1		565.8



Bivariate Analyses: Numerical x Categorical

Do states with bigger economies (gdp_category) pay higher minimum wages?

F	G	H	I	J	K	L	M	N
Small GDP	**Medium GDP**	**Large GDP**						
9.84	0	11						
8.5	10.5	10.2						
8.25	10.1	8.25						
10.1	7.25	5.15						
7.25	7.25	8.25						
10	7.25	10.1						
0	7.25	11						
8.3	0	9.25						
9	7.85	9.65						
7.25	8.25	8.6						
7.5	7.25	10.4						
7.25	10.75	7.25						
10.1	0	8.3						
8.85	0	7.25						
10.5	7.25	7.25						
8.75	7.25	7.25						
5.15		11.5						



Bivariate Analyses: Numerical x Numerical

Do states with bigger populations have higher levels of employment in manufacturing?

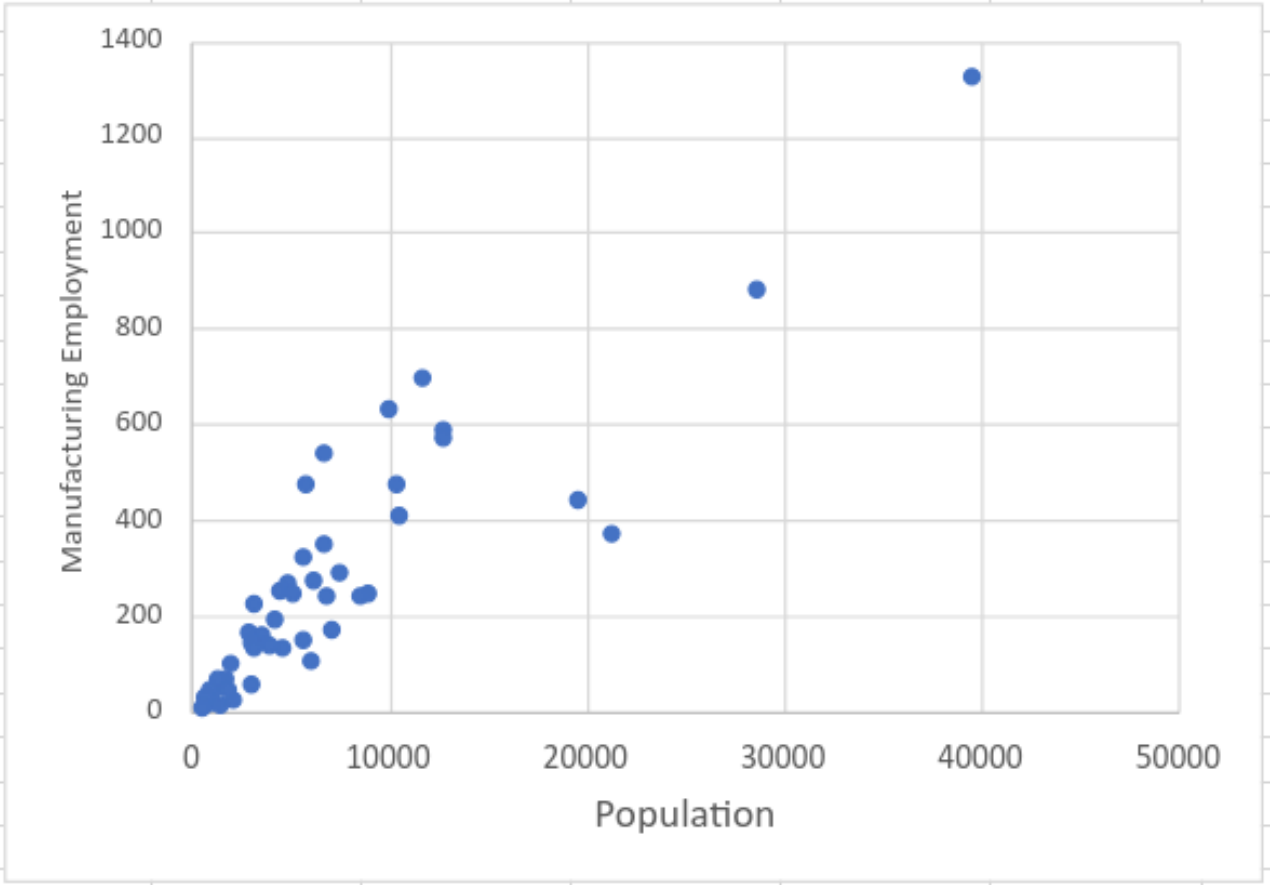
Bivariate Visualizations: Numerical x Numerical

Make a scatter plot of manufacturing employment and population

1. Highlight the two columns
2. Insert a scatter plot

Do states with bigger populations have higher levels of employment in manufacturing?

	A	B	C
1	state	population	manufacturing
2	Alabama	4887.681	267.1
3	Alaska	735.139	12.5
4	Arizona	7158.024	170.1
5	Arkansas	3009.733	160.5
6	California	39461.588	1325.4
7	Colorado	5691.287	147.6
8	Connecticut	3571.52	160.3
9	Delaware	965.479	27.1
10	Florida	21244.317	372
11	Georgia	10511.131	408
12	Hawaii	1420.593	14.2
13	Idaho	1750.536	68.2
14	Illinois	12723.071	588.3
15	Indiana	6695.497	542
16	Iowa	3148.618	223
17	Kansas	2911.359	165.1
18	Kentucky	4461.153	252.1
19	Louisiana	4659.69	134.9
20	Maine	1339.057	52
21	Maryland	6035.802	108.3
22	Massachusetts	6882.635	244.1

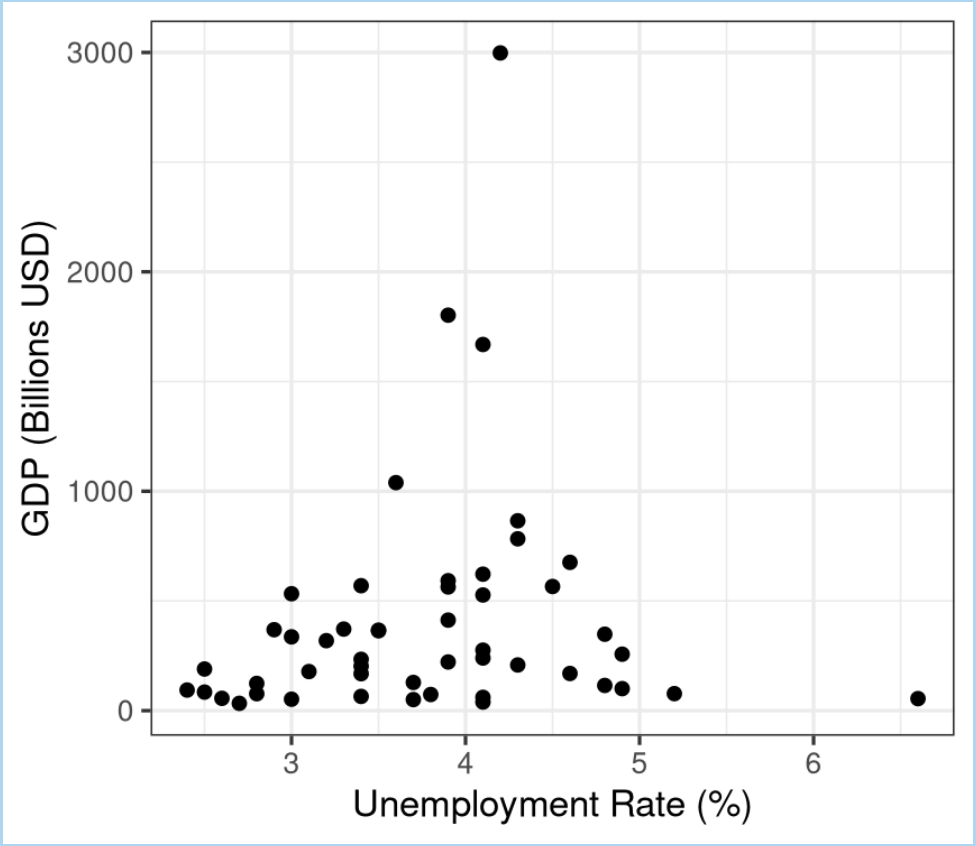


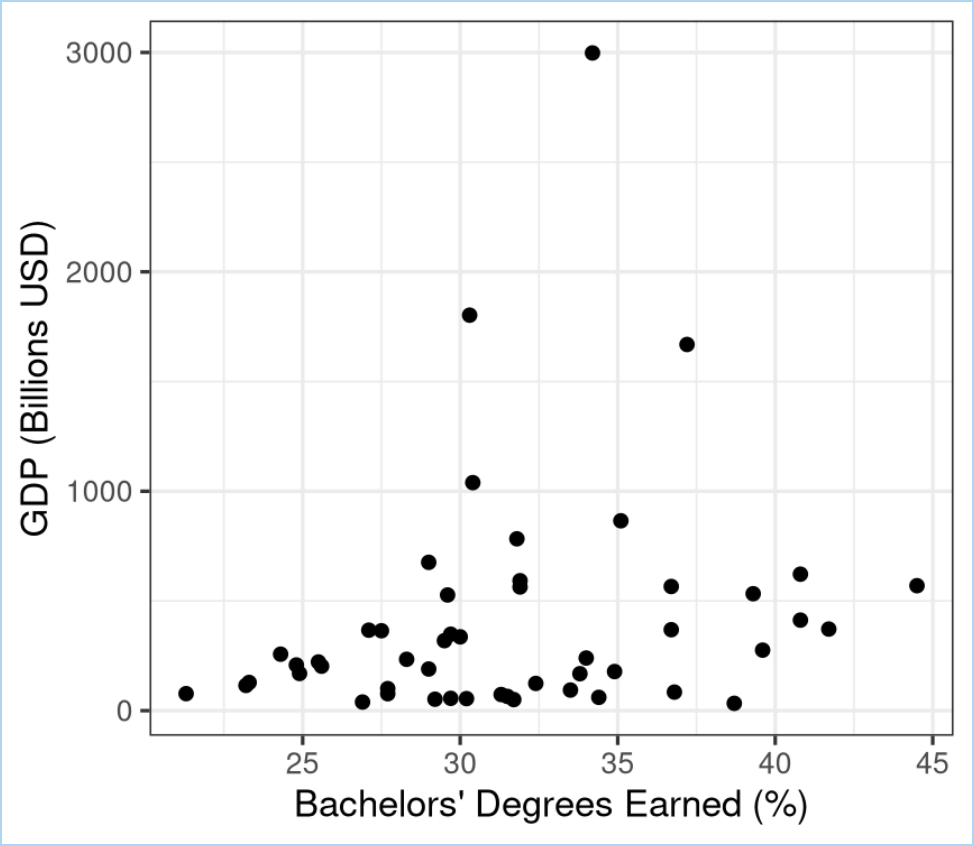
Bivariate Analyses: Numerical x Numerical

Make and analyze the following four scatter plots:

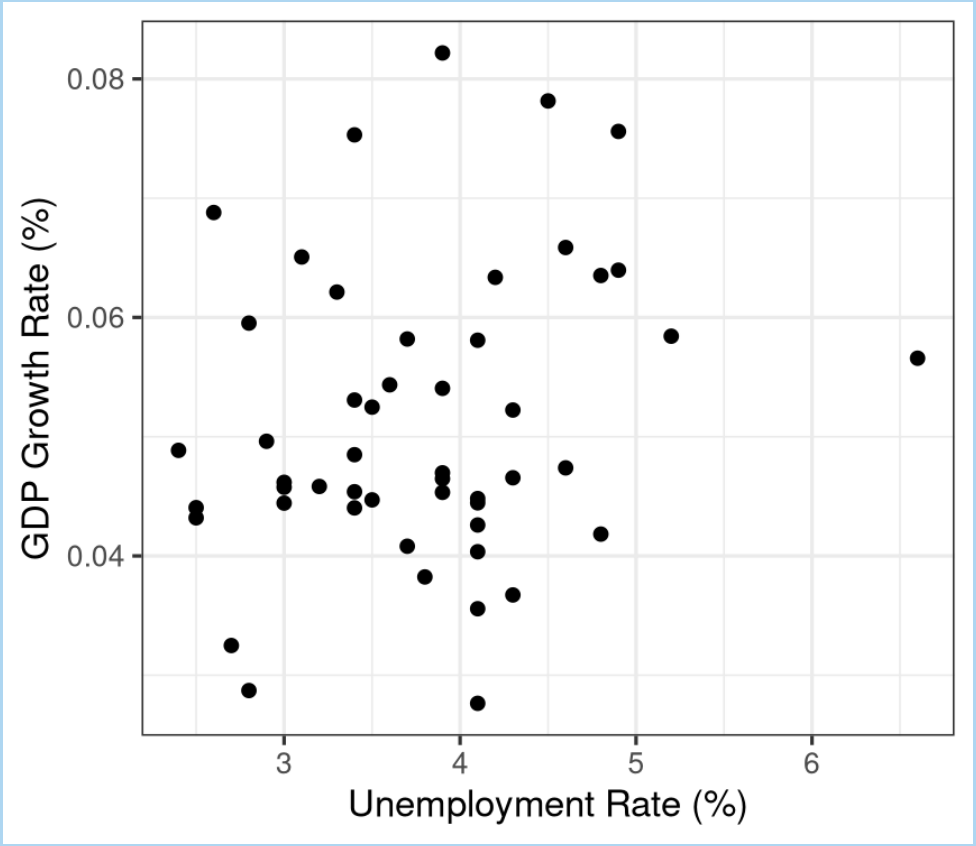
- GDP (billions) x Unemployment
- GDP (billions) x Bachelors' Degrees
- GDP (rate) x Unemployment
- GDP (rate) x Bachelors' Degrees

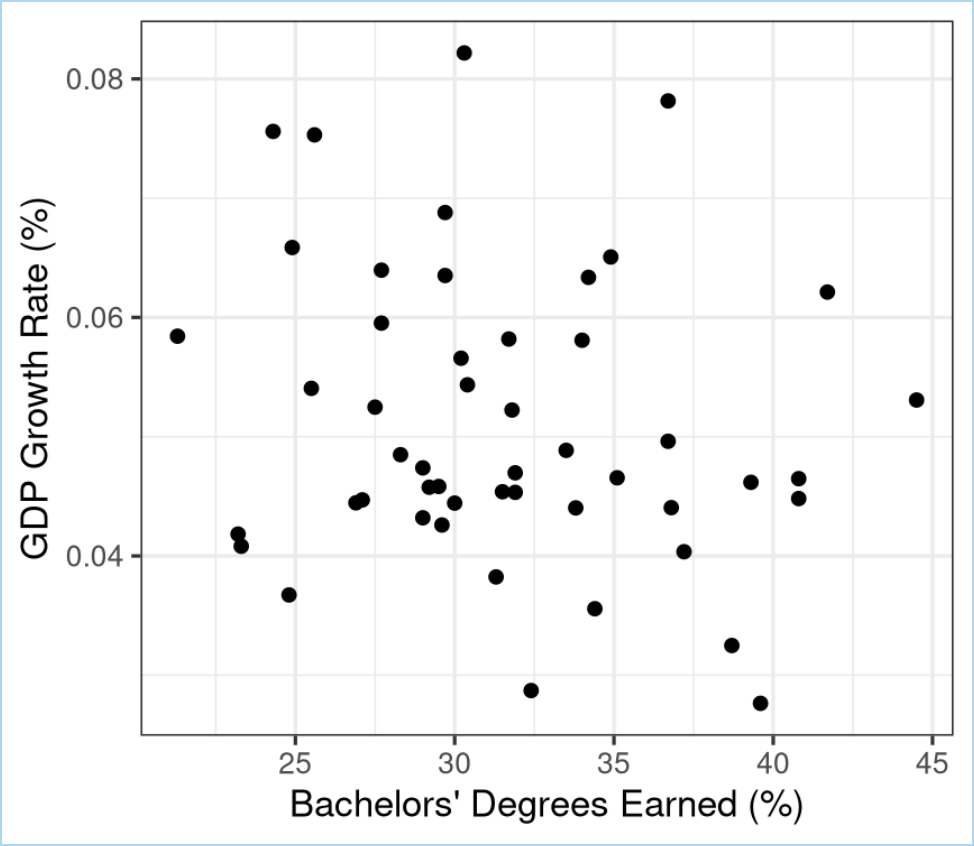
Explaining GDP (Billions USD)





Explaining GDP (Rate)





Session 1: Descriptive Statistics and Visualizations

3) Matching our Tools to the Type of Variable

- **Categorical**: Tables of counts and bar plots
- **Numerical**: Descriptive stats, box plots and histograms
- **Cat x Cat**: Pivot tables, side-by-side and stacked bar plots
- **Cat x Num**: Pivot tables and box plots
- **Num x Num**: Scatter plots

For Tomorrow

1. Read Wilson, Keating and Beal-Hodges ch 2-5

2. For any statistical concepts from today that remain unclear, re-read the relevant sections of Johnson (2012) and be ready to ask about it tomorrow.

3. Make sure you are comfortable using Excel to perform the analyses we explored today.